Encyclopedia of ASTRONOMY and ASTROPHYSICS

VOLUME 1

Institute of Physics PUBLISHING

Öpik (Opik), Ernst Julius (1893–1985)

Born in Estonia, Öpik studied at Moscow University, and helped establish Turkestan University in Tashkent, becoming the Astronomer (director) at Tartu Observatory in Estonia. He fled the Red Army by horse cart during the Second World War and went to Armagh Observatory (Northern Ireland) in 1948. His wide-ranging interests are reflected in his discoveries and theories. These include the discovery of degenerate stars, e.g. white dwarfs, in his calculation of the density of o^2 Eridani (1915). He calculated the distance of M31 as 450 000 parsecs from the Sun (1922). He computed by hand evolutionary models of main-sequence stars into giants (1938) over a decade earlier than the computer computations of HOYLE and SCHWARZSCHILD. He predicted the density of craters on the surface of Mars, which was confirmed 15 years later by planetary probes. He put forward an unproven theory of the Ice Ages based on a calculation of changes in the convection in the internal structure of the Sun rather than MILANKOVITCH cycles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ångström, Anders Jonas (1814–74)

Physicist, born in Lödgö, Sweden. He was keeper of the observatory and professor at Uppsala where he studied heat, magnetism and optics and examined the spectra of the Sun and auroras. His name is commemorated with the *angstrom unit*, 10^{-10} m, which is used for measuring wavelengths of light and x-rays, and the separation of atoms in molecules and crystals.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

61 Cygni

The star 61 Cygni is important because of its large proper motion, first measured at the Palermo Observatory by Giuseppe Piazzi (1746–1826). The star became popularly known as 'Piazzi's flying star'. Its measured annual proper motion of 5.23″ is still the seventh largest known, and is the largest for a naked-eye star (apparent magnitude 4.8).

This star was also the first to have its annual parallax reliably determined, by Friedrich Bessel (1784–1846). Bessel's measurement of 0.314″, made at Königsberg with Fraunhofer's 6.25 inch heliometer in 1838, compares quite well with the modern value of 0.286″. Determinations of stellar parallax made in the same year by other observers were far less accurate. It is the eleventh closest star, at a distance of 11.4 light-years.

61 Cygni is also a well-known binary system, with a period of 653.3 years. Its components are currently separated by 30.3″ at position angle 150°. The primary star 61 Cyg A is an orange dwarf, spectral type K5V, of apparent magnitude 5.20 and absolute magnitude 7.5. Its companion 61 Cyg B is also an orange dwarf, of spectral type K7V, with an apparent magnitude of 6.05 and absolute magnitude 8.3. The system is reputed to be the most extensively observed double star, some thousands of visual observations being supplemented by more than 34 000 photographic plates. Precise astrometric measurements of these plates have indicated that the system has at least one invisible component, believed to be a planet of similar mass to Jupiter, and perhaps as many as three, with orbital periods of between 5 and 12 years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abastumani Astrophysical Observatory

The Abastumani Astrophysical Observatory (AbAO)—longitude, 42.83; latitude, 41.81 degrees—was founded in 1932. It is located in Abastumani, in the south-west part of the Republic of Georgia, 250 km west of the capital city Tbilisi, on the top of Mount Kanobili at 1700 m.

AbAO's primary mission is to enable astronomers of the former Soviet Union to carry out high-quality observations. The average number of clear nights is 130 per year with 25% of seeings smaller than one arcsec. At present, about 100 staff members work at the Observatory in six departments and four laboratories. The Observatory's main facilities are 125 cm Ritchey–Chrétien and 70 cm meniscus telescopes.

The major research areas are accretion disks and pulsars astrophysics, solar system cosmogony, objective prism spectroscopy, low-amplitude short-period variables, AGNs variability, solar physics, solar–terrestrial phenomena and atmospheric physics.

For further information see
http://gamma.bu.edu/webt/abastumani.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abbe, Ernst (1840–1905)

Born in Eisenach, Grand Duchy of Saxe-Weimar-Eisenach (now Germany), Abbe became director of the observatory at Jena and research director of the CARL ZEISS optical works in Jena. He discovered the *Abbe sine condition*, which describes a lens that will form an image, without defects of coma and spherical aberration. His mathematical treatment founded the present-day science of optics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abbott, Francis (1799–1883)

Watchmaker, born in Derby, England, convict, transported to Tasmania in 1845. He made astronomical observations at the Rossbank Observatory after the end of his sentence until its closure in 1854 and in his private laboratory in Hobart.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abell Clusters

Abell clusters are the most conspicuous groupings of galaxies identified by George Abell on the plates of the first photographic survey made with the SCHMIDT TELESCOPE at Mount Palomar in the 1950s. Sometimes, the term Abell clusters is used as a synonym of nearby, optically selected galaxy clusters.

George Abell constructed a catalogue containing 2712 of the richest such groupings in the northern sky, which was later extended to the southern sky. It is no exaggeration to say that the total sample of 4076 cluster candidates over the whole sky has revolutionized the study of the large-scale structure in the universe. The Abell catalogue has formed the basis for the first quantitative studies of the densest component of the large-scale structure in the local universe. In recent years, the definition of samples of candidate clusters from wide-field survey plates has been repeated with automatic plate-scanning machines. This showed objectively that the subjective factor in Abell's visual selection is quite small, at least for the richer and more nearby clusters.

The reality of the cluster candidates in Abell's catalogue has been the subject of some debate, until spectroscopic observations of large numbers of galaxies in the directions of the Abell clusters showed convincingly that only a small fraction of the rich clusters are the result of chance superpositions. That is, a very large fraction of the rich cluster candidates in the catalogue made by Abell (or, including the southern clusters, by Abell, Corwin and Olowin) represent compact, localized peaks in the spatial distribution of galaxies, mostly with redshifts less than 0.2, and held together by gravity.

Already in the 1930s, Fritz ZWICKY had concluded that the luminous matter (i.e. the galaxies) in clusters represents only about 10% of the total cluster mass, most of which can therefore be detected only through its gravitation. This has led to estimates of the total mass (both visible and dark matter) by various means. The most common of those are the velocities of the galaxies in the cluster, the amount and temperature of the hot (x-ray-emitting) gas and the distortion of the images of galaxies at distances well beyond that of the cluster by GRAVITATIONAL LENSING.

For a long time, several of the better-known Abell clusters, like those in the COMA BERENICES and PERSEUS constellations, have shaped our vision of the class of rich, populous, clusters. In this schematic view, rich clusters are smooth, round and virialized structures. This idealized picture coexisted with the knowledge that there are significant variations in the various properties of the Abell clusters. This has led to many studies of those properties, and of correlations between them, as well as to several attempts to describe the formation and evolution of rich clusters. It is now realized that clusters are still forming and evolving at the present epoch.

Among the cluster properties that can be studied, and for which theoretical predictions have been made are the 3D shape (or rather, the axial ratios of the galaxy distribution), the composition of the galaxy population (i.e. the fractions of galaxies of different morphological types), the distributions of the luminosities of the galaxies, the detailed dynamics of the various galaxy classes, dynamical substructure and segregation and the fraction of the total mass consisting of baryonic (i.e. 'ordinary' nucleonic) matter. An important recent development is the search for, and study of, galaxy clusters at very large distances (i.e. at high redshifts), which are the 'forebears' of the local rich clusters in the Abell catalogue. For those younger clusters at high redshifts, the Abell clusters serve as a local, present-day, reference population.

Abell clusters as a subset of the total cluster population

When searching for cluster candidates on the Palomar Sky Survey plates, Abell had no information about distances (or redshifts) of the galaxies. Therefore he used the distribution of the galaxies in apparent magnitude to select those peaks in the projected galaxy distribution that are most likely to correspond to a spatially compact structure. Taking the magnitude of the 10th brightest galaxy as an approximate 'standard candle', a redshift was estimated; this yields the angle subtended by a fixed linear size of $1.5h^{-1}$ Mpc at the distance of the cluster (where h is the value of the present Hubble parameter, expressed in units of 100 km s^{-1} Mpc^{-1}). In a circular aperture with radius equal to that angle, the number of galaxies with a magnitude not more than two magnitudes fainter than the third-brightest galaxy was counted. Finally, the number of unrelated galaxies in the aperture (and down to the same magnitude limit) was estimated from the galaxy density in background fields without obvious cluster candidates.

The corrected number of galaxies (the richness count, i.e. the estimated number of members in the aperture above the magnitude limit) was found to have an uncertainty of about 17. Therefore, only clusters with a corrected galaxy count of at least 50 were considered by Abell to have been sampled in an unbiased fashion out to redshifts of 0.1–0.2. In Abell's original ('northern') catalogue, 1682 of the 2712 cluster candidates have a count of at least 50. The lower limit in richness count must be applied if one uses the Abell catalogue for statistical purposes. Clearly, many less rich clusters exist but at larger distances–redshifts their contrast with respect to the field is too low to allow a robust definition of a statistically reliable sample.

In recent years, an extensive redshift survey (the ESO Nearby Abell Cluster Survey) has been made of close to 6000 galaxies in about 100 cluster candidates (mostly from the southern part of the Abell, Corwin and Olowin catalogue) with a richness count of at least 50 and estimated redshifts less than 0.1 (see GALAXY REDSHIFT SURVEYS). The contamination in these redshift surveys by galaxies that do not belong to the main cluster is far from negligible, i.e. about 25%. However, the majority of the redshift surveys contains a spatially compact cluster to which at least 50% of the galaxies with measured redshifts

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

belong. Only about 10% of the candidate clusters appear to be a superposition of two almost equally rich (but relatively poor) systems at different redshifts along the same line of sight.

For the spatially compact systems, the velocity dispersion shows a global correlation with richness count (clusters with higher richness counts on average have larger velocity dispersions), but the correlation is very broad (at least a factor of 2 in both quantities). The uncertainty in the visually estimated richness counts might be thought to be responsible for this, but the width of the relation does not decrease if one uses richness counts based on machine scanning instead of the original ones.

For a sample of about 150 Abell clusters with redshifts less than 0.15, cluster masses were calculated from the relative velocities and positions of the galaxies, assuming that the virial theorem holds in the central regions of the clusters. The cluster masses correlate fairly well with the velocity dispersions, but the mass distributions in the various intervals of richness counts appear to have considerable overlap. Therefore, application of a limit in richness count to a sample of Abell clusters (which is necessary for practical reasons) induces quite a diffuse limit in mass.

The clusters, or rather cluster candidates, in Abell's catalogue with richness counts of at least 50 are therefore a subset of all clusters in the mass range from about 4×10^{13} to $2 \times 10^{15} M_{\odot}$. However, for clusters with a velocity dispersion of at least 800 km s^{-1}, essentially all richness counts are larger than 50. In other words, all clusters with a velocity dispersion of at least 800 km s^{-1} are contained in the sample with a limiting count of 50, and the estimate of their space density is unbiased. Clusters with apparent velocity dispersions greater than about 1200 km s^{-1} turn out either to be superpositions or to have lots of dynamical substructure.

With the advent of all-sky x-ray surveys like those from the EINSTEIN (HEAO-2) and ROSAT missions, it has become possible to construct complete samples of clusters for which the x-ray flux from the hot gas in the potential well of the cluster is larger than a threshold value. This produces cluster catalogues that are fundamentally different from, and thus complementary to, the Abell catalogue, although there is quite some overlap. The mass of the x-ray gas is generally at least as large as the mass of the cluster galaxies, but the combined mass of these two baryonic components is typically only 10–15% of the total mass. When the total mass of a cluster can be estimated both from the kinematics of the galaxies and from the x-ray temperature and brightness, the two estimates in general agree reasonably well.

Properties of the galaxy population in Abell clusters

In the past, several schemes have been proposed for the classification of Abell clusters. All of them summarize in one way or another the distribution of the cluster galaxies in position, magnitude or morphological type, or any combination of those. The projected distribution of the galaxies has many forms and ranges between the following extremes. There may be a central concentration of bright galaxies, generally of early type, i.e. ellipticals, and frequently one of them is a cD galaxy, i.e. a giant elliptical surrounded by an extended envelope (see ELLIPTICAL GALAXIES). At the other extreme there are clusters that do not have a clear central concentration.

In some clusters, the galaxy distribution is quite smooth, and in general those clusters contain relatively few spiral galaxies. When the fraction of spiral galaxies is large, the galaxy distribution is in general less regular. The relative fractions of early- and late-type galaxies are correlated with richness count, and this is a manifestation of the morphology–density relation. The latter shows a clear correlation between the relative fractions of ellipticals, lenticulars (S0s) and spiral galaxies, and the (local) projected galaxy density (and therefore radial distance). The S0s may contribute up to 50% in the center, with ellipticals not far behind and spiral galaxies about 10%. In the outer parts, ellipticals are almost absent while spiral galaxies may represent up to 60%. Note that these are global values: individual clusters show a considerable spread around these.

Even though in a sizeable fraction of the Abell clusters the galaxy distribution is not very regular or circularly symmetric, one can always derive the azimuthally averaged projected number density profile $\Sigma(R)$, in which R is the projected distance from the cluster center, i.e. the shortest distance between the line of sight through a galaxy and the cluster center. Several expressions have been proposed for the mathematical description of $\Sigma(R)$, all of which have three parameters. Those are the central value of $\Sigma(R)$, i.e. $\Sigma(R = 0)$, a characteristic length R_c (the distance at which $\Sigma(R)$ has decreased by a given factor, say 2) and a measure of the decrease of $\Sigma(R)$ in the outer parts (generally the logarithmic slope α of $\Sigma(R)$).

Recently, $\Sigma(R)$ has been derived for galaxies of different morphological types in about 70 rich Abell clusters. In individual clusters, the number of galaxies of a particular type is generally not sufficient to allow an accurate estimate of the three parameters of $\Sigma(R)$. By properly combining data for many clusters one can compare the representations of $\Sigma(R)$ for ellipticals, S0s, spirals and galaxies with emission lines (mostly very 'late' spirals, such as Sc and Sd, with ionized gas in their interstellar medium). In other words, by sacrificing the detailed properties of individual clusters, one obtains a picture of an average rich Abell cluster.

There appears to be a clear correlation between galaxy type and $\Sigma(R)$: the characteristic length R_c increases markedly from early to late galaxy type (from about 0.1 to 0.5 Mpc). This shows that ellipticals are indeed much more centrally concentrated than spirals, while the emission-line galaxies form the most extended population. These differences must be accompanied by differences in the kinematics of galaxies of the various types, because all

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

galaxy classes move in the same cluster potential, which is mostly determined by the dark matter.

Such kinematical differences are indeed observed: the ellipticals and S0s show the smallest dispersion of the line-of-sight component of their velocities, and this dispersion varies little with projected distance from the center. Spirals, and in particular emission-line galaxies, have a larger velocity dispersion (by as much as 20–30%) which decreases markedly towards larger projected distances. Actually, the kinematics of the emission-line galaxies indicates that they have not yet traversed the dense central cores, which is probably the reason why they have not yet lost their line-emitting gas in encounters with other galaxies.

Combining the projected galaxy distributions with the kinematics one may estimate the distribution of the total (visible plus dark) mass via the Jeans equation of stellar dynamics. By comparing the distribution of the dark matter with that of the luminosity of the galaxies, one can in principle study the variation of the so-called mass-to-light ratio with distance from the cluster center. This may give clues about details of the formation process, such as the effects of galaxy encounters, the role of the dark matter haloes of the galaxies, etc.

Abell clusters as cosmological probes

Several observational properties of Abell clusters have been used to constrain the theories of formation of large-scale structure in the universe and the parameters in those theories (see also UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION). These properties include the spatial distribution of Abell clusters, their shapes and their masses. In different ways, these all carry information on the way in which the largest well-developed structures in the universe have formed through the growth of the initial fluctuations in the matter density.

The spatial distribution of Abell clusters has been analyzed through the two-point correlation function $\xi(r)$, i.e. the fraction of cluster pairs with a certain separation, in excess of the expected number of pairs for a random distribution, which has been derived for clusters of various richness counts. In general, the correlation function is found to have a power-law form: $\xi(r) = (r/r_0)^{-\gamma}$; the exponent γ (about 2) does not appear to depend on the limiting richness count, but the value of the correlation length r_0 does, and is larger for the richer clusters (with a characteristic value of about 20 Mpc). In principle, these data allow one to derive the value of the cosmological density as well as the amplitude of the fluctuation spectrum.

Another aspect of the distribution of rich Abell clusters is that they are generally located in the vertices where the sheets and filaments in the general galaxy distribution come together. Therefore, the distribution of rich clusters has sometimes been compared with the distribution of the vertices in so-called Voronoi tesselations, which are geometric partitionings of space.

The shapes of Abell clusters have been derived from the projected distributions of galaxies. Using the galaxy positions irrespective of galaxy type, one can calculate the apparent ellipticity of a cluster. In general, the richer clusters are less elongated than the less rich ones. The apparent ellipticities for a cluster sample of about 100 northern Abell clusters suggest that the elongated clusters are prolate (cigar like) rather than oblate. Comparison of these data with the results of numerical N-body calculations can constrain the theories of structure formation.

The full distribution of the masses of a volume-limited sample of Abell clusters (i.e. its shape *and* normalization) can also give information for cosmological structure formation theory. As the sample of Abell clusters with a limiting richness count of 50 has a rather badly defined completeness at smaller masses, one must restrict the comparison between observations and predictions to the most massive clusters for which the Abell catalog is complete.

It is far from trivial to derive independent information for the several parameters in the formation theories that influence the properties of the most massive structures. Yet, there seems to be general agreement that the latter are more naturally understood in a universe in which the matter density is considerably smaller than the critical density.

Bibliography

Book:

Giurin G and Mezzetti M (ed) 1999 *Observational Cosmology: The Development of Galaxy Systems (Astron. Society of the Pacific Conf. Ser. 176)*

Journal articles:

Abell G O 1958 The distribution of rich clusters of galaxies *Astrophys. J. Suppl.* **3** 211–88

Abell G O, Corwin H G and Olowin R P 1989 A catalog of rich clusters of galaxies *Astrophys. J. Suppl.* **70** 1–138

Reviews:

Bahcall N A 1977 Clusters of galaxies *Ann. Rev. Astron. Astrophys.* **15** 505

Bahcall N A 1988 Large-scale structure in the Universe indicated by galaxy clusters *Ann. Rev. Astron. Astrophys.* **26** 631

Sarazin C L 1986 X-ray emission from clusters of galaxies *Rev. Mod. Phys.* **58** 1

Peter Katgert

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aberration

(1) The apparent displacement of a star from its mean position on the celestial sphere due to the velocity of the Earth in its orbit around the Sun. The phenomenon was discovered in 1729 by James Bradley (1693–1762) who was, in fact, trying to measure stellar parallax. The displacement is caused by the combination of the velocity of the Earth and the velocity of light approaching from the source. If the Earth were stationary, light from a star would arrive from the true direction of this source, but the motion of the Earth causes the light to appear to be approaching from a point which is slightly displaced in the direction of the Earth's motion. In the course of a year, as the Earth travels round the Sun, a star will trace out a small ellipse in the sky about its mean position. The maximum radius of this ellipse (in radians) is equal to the ratio of the speed of the Earth to the speed of light (30 km s^{-1}:300 000 km s^{-1}), that is about 20.5 seconds of arc. The eccentricity of the ellipse depends on the celestial latitude of the star (the figure becomes a circle at the pole of the ecliptic and a straight line on the ecliptic).

The displacement due to aberration is much greater than that due to parallax (the annual parallax of the *nearest* star is 0.76 seconds of arc) and this must be allowed for before the parallax can be determined for a star. A similar, though smaller, aberration effect occurs due to the speed of rotation of the Earth on its axis. This is known as diurnal aberration.

(2) In optical systems, such as lenses and curved mirrors, aberration refers to the inability of the system to produce a perfect image. Unlike a plane mirror, which does not create aberrations, a lens or curved mirror is an imperfect image producer, becoming ideal only for rays passing through (or reflecting from) its center parallel to the optical axis (a line through the center, perpendicular to the lens surfaces). The main aberrations are chromatic, spherical, coma and astigmatism.

See also: astigmatism, atmospheric refraction, chromatic aberration, coma, scintillation, spherical aberration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abetti, Antonio (1846–1928) and Abetti, Giorgio (1882–1982)

Antonio was born in San Pietro di Gorizia, Italy. A civil engineer, he turned to astronomy and became director of the observatory in Arcetri and professor of astronomy at the University of Florence. His main interest was positional astronomy, observation of minor planets, comets and star occultations. In 1874 he observed the transit of Venus across the Sun's disk through a spectroscope. His son, Giorgio Abetti, was a solar physicist, born in Padova, Italy. He succeeded his father as director of the Arcetri Observatory. In the United States at Mount Wilson Observatory, he took up solar physics and, on his return to Italy, established a spectroheliograph in Arcetri, about 180 degrees in longitude away from Mount Wilson, to give nearly 24 h coverage of solar phenomena. Discovered the radial motion of gases in sunspots (*Evershed–Abetti* effect). He was well known as a popularizer of astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Absolute Zero

The lowest possible temperature; according to the kinetic theory of gases, this is the temperature at which all motion of atoms and molecules ceases. This temperature is equivalent to $-273.16\,^{\circ}\mathrm{C}$, and defines the zero of the kelvin, or absolute, temperature scale.

See also: black-body radiation, temperature scales.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Absorption Spectrum

A pattern of dark lines or bands superimposed on a continuous spectrum. When a continuous spectrum of radiation (a broad range of wavelengths) passes through a material medium (for example, a cool, low-pressure gas), selective absorption occurs at certain particular wavelengths. This gives rise to a series of dips in intensity (absorption lines) which, in the visible region of the spectrum, appear as dark lines against the bright background of the 'rainbow' band of colors that comprises the continuous spectrum.

Photons of electromagnetic radiation may be absorbed through radiative excitation, a process that occurs when an electron in one of the lower energy levels of an atom or ion absorbs a photon with energy precisely equal to the difference in energy between that level and one of the higher permitted levels and, as a result, jumps (makes an 'upward transition') from the lower to the higher level. Because the energy of a photon is inversely proportional to wavelength, the larger the energy gap, the shorter the wavelength of the radiation that is absorbed when an electron makes a particular transition. An electron does not normally remain in an excited level for long (typically 10^{-8} s). When it drops down again, it emits a photon or photons that may or may not have the same wavelength as the one that was originally absorbed (depending on whether it drops directly to its original level or descends in a series of smaller steps). If it drops in a series of steps, each of which corresponds to the emission of a photon of lower energy (and longer wavelength) than the one that was originally absorbed, the total number of photons of the original energy will be reduced and the spectrum will be depleted at the input wavelength. Furthermore, although all the input photons were traveling along essentially the same direction (from the source to the observer), the re-emitted photons travel away in random directions. Consequently, far fewer photons at the absorption wavelength reach the observer than photons of other wavelengths. The resulting absorption line is darker than the adjacent part of the continuous spectrum but, because some photons of that wavelength do reach the observer, not totally black.

Because each permitted transition corresponds to absorbing light of one particular wavelength, atoms or ions of a particular element produce absorption lines at a number of different wavelengths, each chemical element having its own distinctive 'fingerprint' pattern of absorption lines. The shortest-wavelength lines correspond to the largest energy gaps (i.e. to transitions from the lowest, or 'ground', level of the atom or ion). If an electron absorbs a photon with energy in excess of the ionization energy (or 'ionization potential'), it will be removed from the atom. The absorption spectrum of a particular species of atom consists of several series of lines, corresponding to the various permitted transitions, the short-wavelength limit of the series corresponding to the photon energies above which ionization takes place. Because ionizing photons can, in principle, have any value of energy above the ionization energy, absorption can take place over a continuous range of wavelengths, shorter than the series limit. Absorption of this kind is called an absorption continuum.

The prominence ('strength') of any particular line depends on the number of atoms of the appropriate chemical element that have electrons residing in the energy level from which the relevant upward transition takes place (the degree of excitation). That, in turn, depends on the abundance of the particular chemical element (the relative proportion of that element in the absorbing substance) and on a number of other factors, in particular the temperature (the higher the temperature, the greater the proportion of electrons in excited states).

In addition to producing absorption through electronic transitions (like atoms and ions), molecules may also absorb (or emit) radiation by changing their states of vibration (their constituent atoms vibrate relative to each other) or rotation (a molecule, having a physical shape, can rotate about a particular axis). Molecular absorption spectra are complex, their various lines often merging into broader bands.

See also: atom, continuous spectrum, electromagnetic radiation, element (chemical), emission spectrum, energy, ion, ionization, molecule, photon, quantum theory, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Abul Wafa Mohammad al-Buzjani (940–97)

Born in Buzjan, Nishapur (now Iran), Abul Wafa flourished as a mathematician and astronomer at the observatory at Sharaf al Daula, where he was the first to build a wall quadrant (over 6 m long) to observe star positions. Best known for trigonometry, he contributed to the theory of the Moon, in which he first used the tangent function, and introduced the secant and cosecant. The trigonometric tables that he calculated were 100 000 times more accurate than Ptolemy's.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Acceleration

Rate of change of velocity. This is expressed in units of meters per second per second (m s^{-2}), or other equivalent units. Thus, an acceleration of 10 m s^{-2} would imply an increase in velocity of 10 meters per second in one second.

See also: energy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Accretion

The process by which a celestial body increases its mass by aggregating smaller objects which collide with it. Several types of object grow by accretion. In binary stars in which mass transfer is taking place, one member grows at the expense of the other; black holes, including supermassive black holes believed to be present in active galactic nuclei, also increase their mass by accretion. In both these cases matter is accumulated via an *accretion disk* orbiting the accreting object.

The process is particularly important in the formation of planets. Dust grains in a protoplanetary disk (or proplyd) around a young star collide and coalesce, gradually building up larger objects, which in turn collide and merge. Objects accreted in this initial phase of planet formation, while their sizes are in the range from millimeters to tens of kilometers, are termed planetesimals. Once an object's size is measurable in kilometers its own gravity becomes a factor in accretion. Collisional mergers between planetesimals allow larger bodies to be built up, and these are termed protoplanets, objects of a thousand kilometers or more. The subsequent evolution of protoplanets proceeds by mutual collisions of increasing violence, and the accretion of fragments produced in these collisions. The bodies left orbiting the star when this final phase of bombardment is over are termed planets.

See also: accretion disks, black hole, solar system: formation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Accretion Disks

By accretion one understands in astrophysics the accumulation of matter onto a heavy object, under the influence of its pull of gravity. The object can be, for example, a star, a black hole or a neutron star. Most of the objects of these types are quite isolated, without much gas around that could be pulled in by their gravity. Accordingly, the vast majority of stars do not show any evidence of accretion, while most neutron stars and black holes in our galaxy are unobservable for this reason.

While somewhat exotic compared with normal stars or galaxies, accretion disks attract attention because they are involved in a wide range of highly visible phenomena. These range from the formation of our planetary system and the rings of Saturn to the x-ray binaries in our galaxy and the collimated jets from protostars and accreting black holes.

The observable effects of accretion can range from modest to quite dramatic. This depends on the amount of mass accreted per unit time, but even more on the depth of the potential well of the gravitational field of the accreter. The simplest model of accretion would be a free particle falling radially to the accreter from a large distance. If it starts from rest at infinity, it arrives at the surface R of an accreter of mass M with the escape speed $v_e = (GM/R)^{1/2}$. We can also write this as $v_e/c = (r_g/2R)^{1/2}$, where $r_g = 2GM/c^2 = 2.8M/M_\odot$ km is the Schwarzschild radius of the mass M. For a neutron star with $M = 1.4M_\odot$ and $R = 10$ km, for example, the particle would arrive at the surface with some 30% of the speed of light. Accretion onto compact objects, in which M/R is large, is therefore accompanied by the release of large amounts of energy.

A free particle attracted to a compact object from a large distance has only a small chance of hitting it. If its initial motion is not directed very precisely to the object, it will just make an orbit around and return to the same large distance. In other words the cross section for accretion by free particles is small. The same effect plays a role when we consider the more realistic case of accretion of a gas and is then called the 'angular momentum problem', discussed below.

The accreting gas might, for example, be provided by a protostellar cloud, in the case of a growing protostar, or by a companion star in a binary. If the accreting star is a neutron star or black hole, such a binary is called an X-RAY BINARY; if it is a white dwarf, the binary is called a CATACLYSMIC BINARY. Finally, accretion of some form of interstellar matter onto a massive black hole in the nucleus of a galaxy produces an ACTIVE GALACTIC NUCLEUS (AGN).

Figure 1 shows an artist's impression of an accretion disk fed by mass overflow from the secondary in a binary. Images like this cannot be made with telescopes. Almost all disks are so small, as seen from Earth, that their geometry can be inferred only by indirect means. Disks that can be observed directly are our planetary system, the rings of Saturn and spiral galaxies. None of these, however, is typical for gaseous accretion disks. Our solar system is only the solid residue of an accretion disk that existed during its formation. In the case of Saturn's rings, the particles making up the rings also behave more like free particles than like a gas. The disks in spiral galaxies are much more complicated than gaseous disks. Also, galactic disks are not old enough to have accreted much onto the galactic core since their formation.

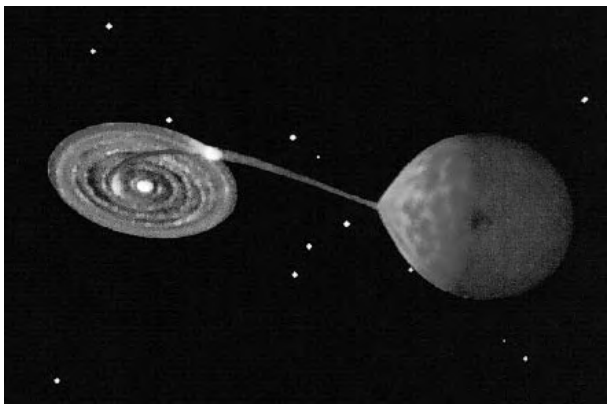


Figure 1. Artist's impression of a binary system consisting of a low-mass main sequence star (red) and a white dwarf. The main sequence star fills its Roche lobe and transfers mass to the white dwarf. Because of the orbital motion of the stars around each other, the mass-transferring stream of gas is not directed straight to the white dwarf and accumulates in an orbiting accretion disk. The impact of the stream on the disk edge causes a bright hot spot. (Copyright Dana Berry and Keith Horne, Space Telescope Science Institute.)

The best prospect for directly observing accretion disks is in protostars. They are the largest disks in angular diameter, as seen from Earth, and with currently developing high-resolution infrared and mm-wave imaging detailed observations of their structure will be possible (see INFRARED ASTRONOMY). Large 'proplyds' (protoplanetary disks) have been imaged in the optical by the Hubble Space Telescope; an example is shown in figure 4 of the article on YOUNG STELLAR OBJECTS.

In the case of binary systems like figure 1, information on the size, thickness and temperature of the disk is provided by eclipses. If the binary is oriented such that the Earth happens to be near its orbital plane, the disk and the mass-providing secondary star eclipse each other regularly as the components orbit around each other. The shape of the light curve in different wavelengths can then be used to infer the properties of the disk, such as its size and temperature (figure 2; see also BINARY STARS).

Accretion temperatures

A characteristic temperature in accretion problems is the virial temperature,

$$T_v = |\Phi|/\mathcal{R} \tag{1}$$

where $\mathcal{R} = 8.314 \times 10^7$ erg g^{-1} K^{-1} mol^{-1} is the molar gas constant and Φ the depth of the potential well. At

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

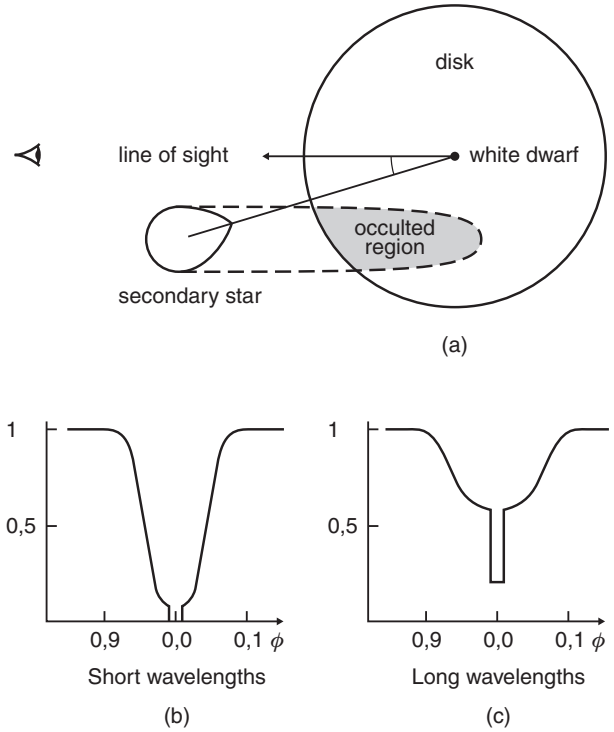


Figure 2. Light curve of an eclipsing binary with an accretion disk around the primary star. The accretion disk widens the eclipse of the primary star. Details of the light curve can be used to infer size and temperature of the disk. (Figure modified after Frank, King and Raine 1992 *Accretion Power in Astrophysics* (Cambridge: Cambridge University Press).)

a distance r from a point mass in Newtonian gravity, $\Phi = -GM/r$. If all the kinetic energy gained by dropping into the potential well were dissipated into internal energy (heat), the gas would reach a temperature of the order T_v. For a neutron star, $T_v \sim 2\times10^{12}$ K, corresponding to a mean thermal energy per proton of the order 1 GeV. The radiation actually observed from accreting systems indicates much lower temperatures. This is because a thermal plasma at this temperature radiates energy at an enormous rate, much higher than can be supplied by the infalling gas. Instead of T_v, the gas settles at a lower temperature, such that the dissipated energy is roughly balanced by energy lost by radiation.

A second characteristic temperature is obtained by assuming that the radiation is emitted as if by a black body from the surface of the accreting object. Equating the luminosity $L = 4\pi r^2\sigma T^4$ of a black body of radius R and temperature T_B to the energy generated by accretion at a rate $\dot{M}$ (mass per unit time) yields

$$T_B = \left(\frac{GM\dot{M}}{4\pi R^3\sigma}\right)^{1/4} \tag{2}$$

where $\sigma = 5.6692\times10^{-5}$ erg cm^{-2} K^{-4} s^{-1} is the constant in Stefan–Boltzmann's law. Actual temperatures differ from this, since the emitted spectrum differs from that of a black body, the emitting surface is not equal to that of the accreting object or, in the case of black holes, much of the accretion energy is swallowed by the hole rather than being radiated. Still, T_B often provides a good order of magnitude for the observed temperatures. The reason is that in these cases the accreting gas is optically thick, so that by repeated photon emission and absorption processes the accretion energy is reprocessed into something like a black body spectrum. For a cataclysmic variables, protostars and AGN, $T_B \sim 10^4$–10^5 K, for x-ray binaries around 10^7 K, corresponding to photon energies of 1–10 eV and 1 keV, respectively. Optically thin accretion processes can also occur, however (see below under 'Radiation processes, two-temperature accretion'), producing photons of much higher energies.

Radiation pressure and the Eddington limit

The photons released by an accreting object exert a force on the accreting gas. By scattering (or by absorption and reemission) on an atom, ion or electron, the outward direction of the photon is changed into a more random direction. The outward momentum of the photons is thereby transferred to the gas: the radiation exerts an outward force. If $F = L/4\pi r^2$ is the radiative energy flux at distance r, and κ the opacity of the gas (scattering plus absorption), the acceleration due to this force is $g_{rad} = F\kappa/c$. This force is just balanced by the inward acceleration of gravity $g = GM/r^2$ when the luminosity L has a value called the Eddington limit or Eddington luminosity:

$$L_E = 4\pi GMc/\kappa. \tag{3}$$

A steady photon source, bound by gravity, cannot have a luminosity significantly exceeding this limit. At a larger luminosity, the atmosphere of the source is blown off by radiation (this happens, for example in a NOVA explosion). The value of the Eddington luminosity depends on the opacity of the gas, and thereby on its state of ionization. It depends on the mass of the source but not its size. Close to the compact object in an x-ray binary the gas is nearly fully ionized, usually of normal stellar composition, and electron scattering the dominant radiation process, with opacity $\kappa \sim 0.3$ cm^2 g^{-1}. The Eddington limit is then

$$L_E \approx 1.5\times10^{38} M/M_\odot \text{ erg s}^{-1}. \tag{4}$$

For a neutron star of $1.4M_\odot$, this is about 50 000 times the luminosity of the Sun.

If this luminosity is generated by accretion, equating it to the accretion energy defines the Eddington accretion rate

$$\dot{M}_E = 4\pi cR/\kappa \tag{5}$$

or about $1.5\times10^{-8} M_\odot$ yr^{-1} for a neutron star of 10 km radius. A neutron star cannot accrete much more than this. The radiation pressure building up would prevent further accretion, and the gas would accumulate in an extended atmosphere around the star instead of settling

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

onto the neutron star surface. (This assumes that radiation is only by photons: at sufficiently high temperature and density energy losses by neutrinos become important.) The accretion rate onto a black hole can be much higher than $\dot{M}_{\rm E}$. An atmosphere surrounding a hole is not supported at its base, but flows in through the horizon. In the process it takes with it the photons trapped in the gas. Depending on the density of such an atmosphere, the accretion rate can be arbitrarily high. The luminosity as seen from Earth does not become much bigger than $L_{\rm E}$, however.

Disks

Mass transfer in a binary

As an example for the formation of an accretion disk, consider a binary in which one of the stars is large enough to fill its Roche lobe (see CLOSE BINARY STARS), the other being a compact star (white dwarf, neutron star or black hole). The Roche-lobe-filling star is called the secondary, since it is the less massive star in such binaries. Mass flows off the secondary at the inner Lagrange point L_1. The orbit of a hypothetical free particle from this point is shown in figure 3. Except directly at L_1, the orbital velocity is very high compared with the sound speed of the gas. Consequently, its path is almost ballistic, i.e. close to that of a free particle. Since gas clouds cannot flow through each other, however, the ballistic flow ceases at the first intersection point of the orbit. The supersonic relative motion of the two parts of the gas stream is dissipated here through shock waves, heating the gas and deflecting the stream.

The further evolution of the stream depends on additional physics such as the rate at which the gas can cool. An example of what the stream might look like after a few orbital periods of the binary is shown in figure 4. The orbiting gas has accumulated into a ring; newly arriving gas is incorporated into the ring through a system of shock waves. The strongly non-circular motion of the gas has settled into a more quiet circular orbit; most of the energy dissipated in the process has been radiated away. Leaving aside perturbations by the impacting stream, and in the absence of viscous forces, the gas can orbit indefinitely on such circular orbits.

The 'angular momentum problem'

In the process of settling onto a circular orbit, a great deal of energy is dissipated, but the angular momentum of the gas around the primary has not changed (neglecting corrections due to the gravitational pull of the secondary). Since a circular orbit has the lowest energy for a given angular momentum, the gas can only sink further into the gravitational potential and accrete onto the primary, if it can lose some angular momentum. Finding the process by which this is done in real systems is called the angular momentum problem. We have illustrated it here with the example of mass transfer in a binary, but the same problem arises for the formation of stars from interstellar clouds or the accretion of gas onto the massive black holes in AGN. In these cases, the initial angular momentum due to random motion of the gas clouds is many orders of magnitude larger than can be accommodated by the accreting object. Rather than accreting directly, the gas forms a disk, acting like a temporary 'parking orbit'. The orientation of the disk is given by the direction of the total angular momentum vector of the clouds, while the time it takes the gas to spiral in through the disk depends on the processes solving the angular momentum problem.

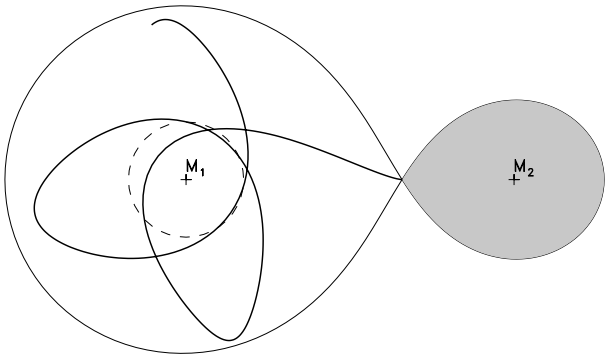


Figure 3. Binary showing Roche lobes, the lobe-filling secondary star (M_2, hatched) and the orbit of a particle released from the inner Lagrange point. The 'rosette' shape of the orbits around the primary (M_1) is caused by the orbital motion of the binary.

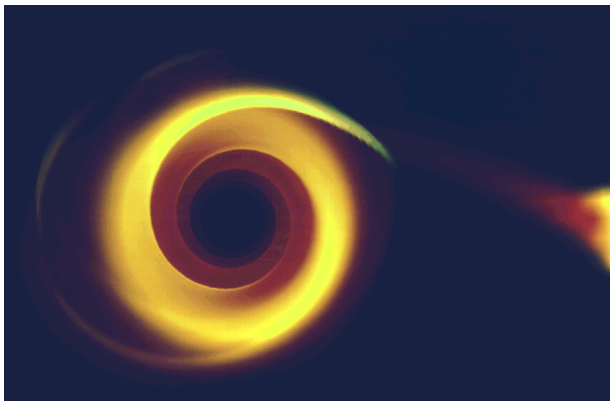


Figure 4. Hydrodynamic simulation of mass transfer in a binary, showing the accumulation of gas in a ring. An additional viscous process has to be added to make the gas accrete onto the primary star (at the center of the ring).

Viscous disks

In many cases (dwarf novae and x-ray transients; see X-RAY BINARY STARS) the accretion onto the compact object is episodic, in the form of outbursts. The decay of such an outburst gives some information on the accretion time, that is, the time it takes the gas to spiral in from the secondary star to the accreting primary. For dwarf novae, for example, this time is a few days, showing that the angular momentum transport process allows accretion in a few days or less of the mass that is involved in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

producing the outburst. Though actual numbers are model dependent, these observations indicate angular momentum transport 12–15 orders of magnitude more efficient than expected from the natural viscosity of the gas. The process responsible for the angular momentum transport in disks is not known with certainty. In view of this uncertainty, the theory of accretion disks makes an arbitrary assumption about the angular momentum transport process. It is assumed that the disk behaves, in effect, like a fluid with a very high viscosity.

Such an enhanced viscosity does indeed solve the angular momentum problem. As a thought experiment, start with a ring of gas orbiting at some distance from the source of the gravitational potential. Viscous stress between neighboring orbits rotating at slightly different velocity makes the ring spread both outward and inward, forming a disk. This spreading has a remarkable property: if sufficient time is available, almost all the mass accretes onto the central object. A small amount of mass in the the outer parts of the disk expands indefinitely, carrying away almost all the angular momentum of the original ring. The time for spreading to a given distance is inversely proportional to the viscosity.

A characteristic velocity in the disk is the isothermal sound speed $c_s = (\mathcal{R}T)^{1/2}$, where T is the temperature at the midplane of the disk (the surfaces are cooler). A characteristic frequency is the orbital frequency Ω. The unknown disk viscosity ν can then be measured in terms of a dimensionless viscosity α, defined by

$$\nu = \alpha c_s^2/\Omega. \tag{6}$$

Where observational indications, such as the decay time of an outburst, are available, they indicate viscosities in the range $\alpha = 10^{-3}$–1. Such values are then used when making theoretical estimates of the structure of accretion disks. The assumption that the angular momentum transport processes can be represented by equation (6), with α something of order unity, is called the α-viscosity assumption.

Thin disks

Accretion disks tend to be cool because the accretion times are long and sufficient time is available to radiate away the dissipated gravitational energy. If they are cool, the effects of gas pressure are small. To see this consider the equation of motion for an ideal gas in a gravitational potential of a point mass M, $\Phi = -GM/r$:

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\frac{1}{\rho}\boldsymbol{\nabla} p - \nabla\Phi \tag{7}$$

where $\boldsymbol{v}$ is the velocity, ρ the mass density and p the gas pressure. At a distance r_0, a characteristic time scale t_0 is the inverse of the Kepler frequency, $t_0 = \Omega^{-1} = (r_0^3/GM)^{1/2}$. Introducing dimensionless quantities $t' = t/t_0$, $r' = r/r_0$, $\boldsymbol{\nabla}' = r_0\boldsymbol{\nabla}$, $\boldsymbol{v}' = \boldsymbol{v}/\Omega r_0$, the equation of motion for a gas of temperature T becomes

$$\frac{\mathrm{d}\boldsymbol{v}'}{\mathrm{d}t'} = -\frac{T}{T_v}\boldsymbol{\nabla}' \ln p - \frac{\hat{\boldsymbol{r}}}{r'^2} \tag{8}$$

where $\hat{\boldsymbol{r}}$ is a unit vector in the radial direction and T_v the virial temperature introduced above. For processes taking place on a length scale r_0 and time scale t_0 the terms in this equation are of order unity, except the pressure term which involves the factor T/T_v. T/T_v is of order unity if all the dissipated energy stays in the gas, but, when cooling is effective, it can be very small. The pressure term is then small, and the gas flows ballistically in the potential Φ.

A disk with $T/T_v \ll 1$ rotates approximately on circular Kepler orbits. The orbital motion is supersonic, with Mach number $\mathcal{M} = \Omega r/c_s \sim (T_v/T)^{1/2}$. The thickness of the disk is found by considering the vertical distribution of gas at some distance r_0, assuming it to be in hydrostatic equilibrium in a frame rotating with the Keplerian rate $\Omega(r_0)$. An isothermal gas is then distributed with height z above the orbital plane as $\rho \sim \exp[-\frac{1}{2}(z/H)^2]$, where H is the scale height, $H = r_0(T/T_v)^{1/2} = c_s/\Omega$. If radiative losses are small, the disk is hot, and the aspect ratio H/r_0 is of order unity. Efficiently radiating disks on the other hand are cool and geometrically thin, $H/r \ll 1$.

Under the assumption $H/r \ll 1$ the hydrodynamic equations for an axisymmetric, plane, viscous disk are simple. To lowest order, the radial equation of motion reduces to $v_\phi = \Omega r$. The azimuthal equation of motion can be combined with the continuity equation into a single equation governing the surface density $\Sigma = \int_{-\infty}^{\infty} \rho \, \mathrm{d}z$:

$$\frac{\partial \Sigma}{\partial t} = \frac{3}{r}\frac{\partial}{\partial r}\left[r^{1/2}\frac{\partial}{\partial r}(\nu\Sigma r^{1/2})\right]. \tag{9}$$

As expected, this equation is of the diffusion type. It shows that, for a cool disk, all physical factors influencing the evolution of the disk enter only through the viscosity ν. This viscosity, of course, also contains most of the assumptions about unknown processes. Usually, additional equations are needed to determine how radiative cooling affects the temperature structure of the disk, on which ν depends.

For steady accretion, and not too close to the central object, the accretion rate is related to the viscosity by

$$\dot{M} \approx 3\pi\nu\Sigma. \tag{10}$$

A high viscosity implies a low surface density, since the accreting mass spends little time in the disk. The characteristic time t_a for gas orbiting at a distance r to accrete is the viscous time scale,

$$t_a = r^2/\nu \approx \frac{r^2}{H^2}\frac{1}{\alpha\Omega}. \tag{11}$$

For cool disks, this is long compared with the orbital time scale. On the other hand, a long time scale implies that there is plenty of time for the disk to cool by radiation, so there is a certain circularity in the mechanism that fixes the disk temperature. In a given situation (central mass, distance), the disk temperature is determined by the mass accretion rate and the dominant cooling process. In some cases, disks can exist in either of two stable steadily

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

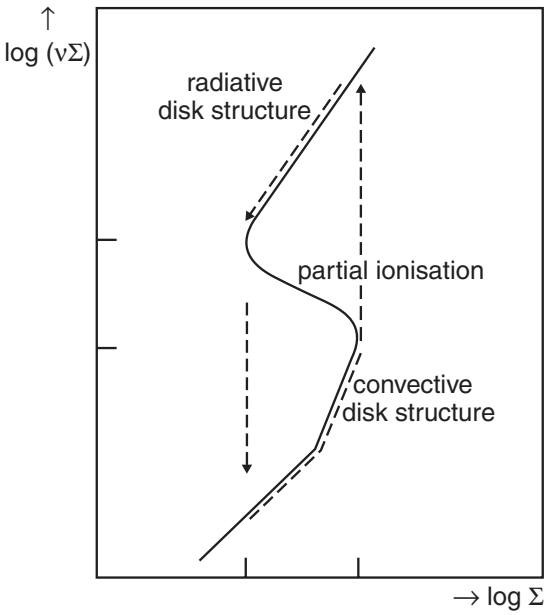


Figure 5. Dependence of the viscous stress $\nu\Sigma$ on surface mass density Σ in a disk where hydrogen is partially ionized. If the mean accretion rate requires a viscous stress in the range where the slope of the curve is negative, this part of the disk is unstable and executes a limit cycle as sketched by the arrows.

accreting states, a cool one with long accretion times and a hot one with short accretion times. In other cases, disk models are found to be unstable for certain mass transfer rates, oscillating between states of high and low accretion. The details of the cooling processes determine when such multiple accretion states exist (see 'Radiation processes, two-temperature accretion' below).

Disk instabilities

The vertical structure of a disk is determined by the need to transport the viscously dissipated energy to the radiating surfaces. For a given surface density Σ, this structure determines the temperature and viscosity of the disk. In conditions where the energy transport process changes rapidly with temperature, the viscous stress $\nu\Sigma$ can depend non-monotically on the surface mass density, in the form of an 'S-curve', as illustrated in figure 5. If the mean accretion rate which is imposed (for example, by the mass-transfering secondary star) lies in the range where Σ decreases with increasing $\nu\Sigma$ (cf equation (10)), the disk is unstable. Instead of accreting steadily, it executes a limit cycle alternating between states of high and low accretion. Usually, conditions like those in figure 5 occur only in limited regions of the disk, for example in the region of partial ionization of hydrogen. If this region contains enough mass, however, the limit cycle will affect the entire disk. This is believed to be the cause of the dwarf nova outbursts in cataclysmic variables.

Angular momentum transport processes

Historically, the idea that some form of hydrodynamic turbulence exists in disks has played a significant role. It was part already of the ideas of Kant and of Laplace, who proposed that the solar system was formed from an initially gaseous disk-like cloud. If $\nu_{\rm m}$ is the microscopic viscosity of the gas, the Reynolds number $r^2\Omega/\nu_{\rm m}$ in an accretion disk is very large, a situation called 'fiercely turbulent' in fluid mechanics (where high Reynolds numbers and turbulence are considered equivalent). The assumption is that the shear flow in a disk would be unstable and develop into turbulence, as in laboratory shear flows. This has been questioned in the astrophysical community, on account of the fact that the flow in a cool disk is close to Kepler orbits, which are very stable. At the time of writing, the issue is still controversial, but it seems quite possible that a Keplerian gas flow does not, in fact, produce hydrodynamic turbulence. A closely analogous laboratory case is the rotating Couette flow. Significantly, this experiment has not shown turbulence for the conditions corresponding to an accretion disk, in which the rotation frequency decreases with distance but the specific angular momentum increases (although turbulence abounds for other parameter values). Attempts to demonstrate turbulence for this case by analytical studies or numerical simulations of the experiment have also been negative.

Among the processes known to work are spiral shock waves. In a large disk (in the sense that the outer radius is much larger than the inner radius), a small compressive perturbation propagating inward steepens into a weak shock wave. Dissipation in the shock damps the wave, but since it also propagates into a flow of increasing velocity, the interaction between wave and mean flow though the shock increases its amplitude. In the absence of other processes damping the wave, a shock of finite strength develops by the balance between shock dissipation and energy extraction from the mean flow. The process produces a modest amount of angular momentum transport, corresponding to $\alpha = 0.01(H/r)^{3/2}$. For protostellar disks, this is in the range of the values inferred from observations, but it is too low for cataclysmic variables and x-ray binaries.

If the disk is massive enough that self-gravity is important, angular momentum transport by gravitational instability is possible. The intrinsic growth rate of a gravitational instability is of the order $(2\pi G\rho)^{1/2}$. If this is larger than the shear rate $\sim\Omega$, instability can grow in a disk; if it is smaller, a prospective perturbation is sheared apart before it can grow. In terms of a characteristic disk mass $M_{\rm d} = 2\pi r^2\rho H$, this condition can be written as $M_{\rm d} > MH/r$. Such disks are called self-gravitating. Self-gravity is more important in cool disks, with small aspect ratio H/r. With the viscosities assumed, the masses of disks in observed systems can be estimated. For x-ray binaries and cataclysmic variables they turn out to be quite small, less than $10^{-8}M_\odot$, and self-gravity is unimportant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In the outer parts of disks in AGN, and in particular in protostellar disks, on the other hand, self-gravity can be important (see YOUNG STELLAR OBJECTS). In this case, consider a disk cooling by radiation so that H/r decreases with time. As the threshold for instability is reached, non-axisymmetric irregularities with length scales of the order H are formed, which exert forces on each other. The shear flow acting on these forces dissipates energy, which in turn heats the disk. A balance is reached at a temperature just above the threshold for gravitational instability, and angular momentum is transported by the gravitational and pressure forces between the perturbations. This process is likely to be important in young protostellar disks, with inferred masses of the order of a few per cent of a solar mass (see STAR FORMATION).

The currently favorite process for non-self-gravitating disks relies on the fact that most disks are partially or fully ionized and hence support magnetic fields (see MAGNETOHYDRODYNAMICS). The possibility that the actual angular momentum transport in disks is done by some form of small-scale magnetic field has already been proposed by Shakura and Sunyaev together with their introduction of the α-viscosity assumption. It can be shown that an initially weak magnetic field in an accretion disk is unstable, its energy density growing by extracting energy from the shear flow. The mathematics of the process was computed around 1960 by Velikhov and by Chandrasekhar, its physical interpretation given by Fricke in 1972 and its importance for accretion disks finally realized by Balbus and Hawley in 1992. Numerous three-dimensional numerical simulations since then have shown how this instability gives rise to the magnetic turbulence postulated earlier. For sufficiently ionized disks the process yields an effective viscosity $\alpha \sim 0.1$. It is a small-scale magnetic field, with length scales in the radial direction of the order of the disk thickness H, and appears to behave roughly as expected from a viscous process. The magnetic field appears to break the dynamic constraints that prevent a Keplerian flow from becoming turbulent by purely hydrodynamic means. In some respects the process resembles the hydrodynamic turbulence proposed earlier, but differs in important aspects as well. The angular momentum transport, for example, is dominated by the magnetic (Maxwell) stresses rather than by fluid motions.

Planets inside disks

Planets grow from a protostellar disk at the same time as the star is formed (see SOLAR SYSTEM: FORMATION). A planet like Jupiter is massive enough to significantly affect the dynamics of this disk. By its tidal effect it clears a gap, a region of low gas density around its orbit, as shown in figure 6. The tidal force excites waves in the disk which propagate away from the planet. Through the tidal forces, the planet attracts gas from the sides of the gap which accretes onto it in the form of two streams. On the other hand, the disk also exerts a gravitational torque on the planet. By this force, the planet's angular momentum changes. It moves to a different orbit, usually closer to the star. The planets in our solar system are therefore probably not at the distances where their formation started. An extreme example of this 'drifting planet' process may be the massive planets observed in very tight orbits around some nearby stars (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS).

Radiation processes, two-temperature accretion

The radiation produced by a disk depends on its optical thickness τ. The energy released in the disk by viscous dissipation, per unit surface area and time, at a distance r from the central mass M is of the order $W = GM\dot{M}/r^3$. This heat diffuses out by radiation. If τ (more precisely, the optical thickness for absorption processes) is large, the radiation emitted at the surface is roughly a black-body spectrum. The approximate surface temperature T_s of a cool optically thick disk is then given by equation (2), with R replaced by the distance r.

If the accretion is steady, so that $\dot{M}$ is independent of r, the predicted temperature varies as $r^{-3/4}$. In the case of protostellar disks and disks in cataclysmic variables, these temperatures are similar to those of normal stellar surfaces, and the same radiation processes determine the details of the emitted spectrum. In accreting black holes and neutron stars, the predicted temperatures are of the order 1 keV (10^7 K). Indeed, the x-ray spectra of these objects often show a component that agrees with this prediction. Usually, however, there is an additional hard component of much more energetic photons, around 100 keV. It often dominates the radiated energy. In these binaries there is an additional component in the accreting plasma that behaves differently from an optically thick disk. The precise nature of this plasma is one of the classical problems of x-ray astronomy, which at the time of writing has not been solved entirely. From the observed spectrum, it appears to be a thermal plasma of modest optical thickness ($\tau \sim 1$) and a temperature around 100 keV, much hotter than the cool disk. The main process producing hard photons under such conditions is inverse Compton scattering of soft photons by hot electrons.

In the neighborhood of black holes and neutron stars accretion is possible not only via a cool optically thick disk. Theory also predicts the possibility of accretion via a hot state (cf discussion above under 'Thin disks'). In this state, the protons are near the virial temperature, while the electrons are much cooler. The accretion flow is geometrically thick ($H/r \sim 1$) but optically thin. Under these conditions, the Coulomb interactions between ions and electrons can be slow compared with the accretion time (depending on accretion rate and viscosity parameter). Since the electrons radiate much more effectively, and are inefficiently heated by the ions, they remain much cooler than the ions: the accretion plasma is not in thermal equilibrium. The Coulomb interaction rate decreases with increasing electron temperature while radiation losses by inverse Compton and synchrotron radiation increase. In a flow of modest optical depth, the inverse Compton losses increase exponentially above

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

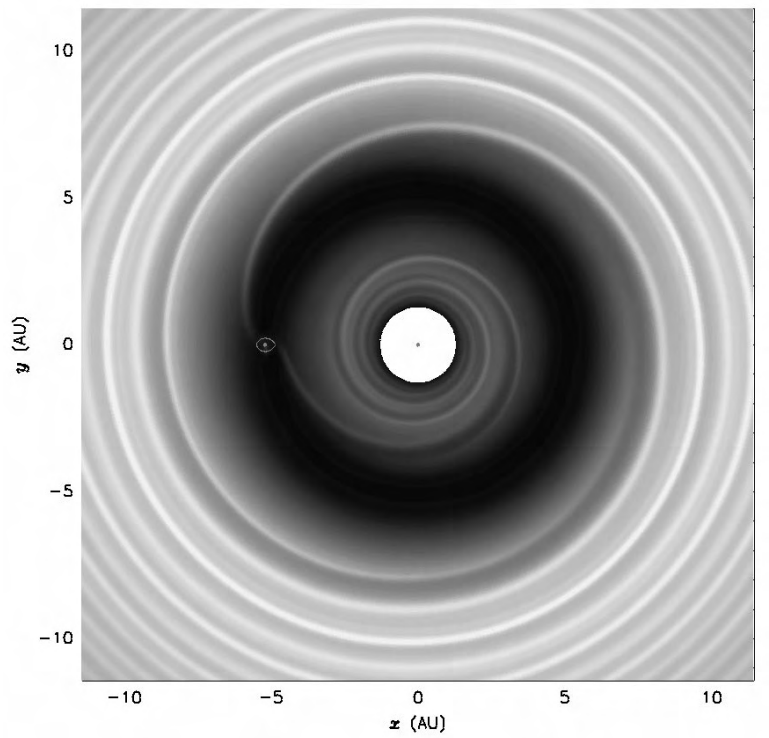


Figure 6. Numerical simulation of a growing gaseous planet inside the protoplanetary disk. The planet has cleared a gap in the disk and accretes gas from the sides of this gap. (W Kley, Institute for Theoretical Physics, University of Jena.)

about 100 keV, so that this a somewhat natural temperature to expect for the emitted radiation. The possibility of such a two-temperature accretion flow to explain the hard spectra of x-ray binaries has already been proposed in the beginning of the x-ray astronomy era. The geometry of such a flow and the nature of its interaction with the cool disk component are still uncertain, however.

Jets from disks

Most systems with accretion disks appear able to produce strongly collimated outflows called jets, at least at some periods in their existence (see ASTROPHYSICAL JETS). Jets with relativistic flow speeds are known from accreting black holes, both the stellar mass holes in x-ray binaries and the massive holes in AGN. Jets at more modest speeds are produced by protostars. They are also known from at least one accreting neutron star (Cir X-1) and an accreting white dwarf (R Aqr). The connection between disks and jets is thus suggestive, but at the same time puzzling. Although jets always seem to be associated with accreting systems in which there is direct or indirect evidence for a disk, not all systems with disks produce jets, or not all the time. A good example are the HERBIG–HARO OBJECTS produced by jets from protostars, which clearly demonstrate that the jets in this case are episodic or highly variable in time (see also PRE-MAIN SEQUENCE STARS).

Summary of types of disk

On a stellar scale, disks are produced in binaries by overflow of gas from one of the components onto the other. The sizes of these disks are as large as normal stars, from a few tenths to a few solar radii. If the primary (the mass-receiving star) is a white dwarf, such as in cataclysmic variables, the inner parts of the disk radiate mostly in the UV, the outer parts in visible light. The mass transfer through these disks is often unstable, causing dwarf nova outbursts. If the primary is a neutron star or stellar-mass black hole, the inner disk radiates in x-rays. Mass transfer in these disks is also unstable at low transfer rates, causing the so-called soft x-ray transients. They sometimes produce jets at relativistic speeds. Disks around stars in the process of formation, the protostellar disks (also called PROTOPLANETARY DISKS), are much larger (around 100 AU). Their inner regions radiate in the visible, the outer parts at infrared to radio wavengths. They are probably also unstable, causing the FU Ori outbursts, and are associated with jets and Herbig–Haro objects. Disks in AGN rotate around massive ($(10^6$–$10^9)M_\odot$) black holes at the centers of active galaxies. Their central regions radiate both in the UV and in x-rays and often produce the relativistic jets seen in double-lobed radio sources. Their sizes are somewhat uncertain, but probably of the order of a parsec or larger.

Disks are also seen around BE STARS. The mass in these disks is believed to be slowly expelled by the rapidly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rotating star. Some of their properties are similar to those of accretion disks, but with mass drifting away rather than accreting onto the star. They are sometimes called excretion or decretion disks.

Bibliography

Frank J, King A R and Raine D J 1992 *Accretion Power in Astrophysics* 2nd edn (Cambridge University Press Astrophysics Series vol 21) (Cambridge: Cambridge University Press)

Pringle J E 1981 Accretion disks in astrophysics *Ann. Rev. Astron. Astrophys.* **19** 137–62

Spruit H C 2000 Accretion disks http://arXiv.org/abs/astro-ph/0003144

H Spruit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ace (Advanced Composition Explorer)

NASA mission to study the solar corona, interplanetary medium, solar wind and cosmic rays. Launched August 1997 and placed in a halo orbit at the L1 Lagrangian point between Earth and Sun. It carries six high-resolution spectrometers and three monitoring instruments with a collecting power 10 to 1000 times greater than past or planned experiments. Its mission is to sample low-energy particles of solar origin and high-energy galactic particles. (See also EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Achernar

The star α Eridani, at the mouth of the great river depicted in ancient star maps. It is a bluish-white main-sequence star, spectral type B3Vp, and at apparent magnitude 0.45 is the ninth brightest star in the heavens. It is at a distance of 142 light-years, and has a parallax of 0.023″. Its absolute magnitude is −2.7.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Achilles

The first Trojan asteroid to be discovered, by Max Wolf in 1906, designated (588) Achilles. It belongs to the largest group of Trojans, orbiting ahead of Jupiter around the L_4 Lagrangian point. It has an estimated diameter of 147 km, and orbits the Sun at a mean distance of 5.17 AU (774 million km) in a period of 11.78 years. The inclination is 10° and the eccentricity 0.15. Like the majority of Trojans, Achilles is D-type, with a reddish reflectance spectrum indicating a carbon-rich surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Achondrite

One of the two main divisions of stony meteorites, the other being *chondrites*. Unlike the chondrites, the achondrites lack chondrules—millimeter-sized grains of silicate—and have a coarser crystalline structure than the chondrites. They contain almost no nickel–iron or sulfide. Achondrites are classified principally by their calcium content. The main calcium-rich (about 5%) varieties are the *eucrites* and *howardites*; the main calcium-poor (1% or less) types are the *aubrites*, *diogenites* and *ureilites*. Howardites, eucrites and diogenites are known collectively as *HED meteorites* or *basaltic achondrites*. The basaltic achondrites have all the characteristics of igneous terrestrial rocks, such as basalt, that were once molten, and are therefore inferred to be fragments from the crust of a parent body that underwent at least partial differentiation; the prime candidate is the Asteroid (4) Vesta. The largest known single achondritic mass is a 1 tonne fragment of the Norton County (Kansas, 1948) meteorite. Into this category also fall all known lunar meteorites, whose composition is consistent with them having been ejected from the surface of the Moon by impact, and all of the SNC meteorites, believed to have been ejected from the surface of Mars.

Compare: Chondrites.

See also: SNC meteorite, Vesta.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Acousto-Optic Spectrographs

The acousto-optic spectrograph (AOS) has played an interesting niche role in radio astronomical observation. By using the interaction between a laser beam and an ultrasonic vibration generated from the incoming radio signal, the AOS makes possible sensitive, wide-band observations of radio phenomena which vary in both frequency and time. This has been especially true in observations of solar radio emissions and in wide-band spectral line observations as found at millimeter wavelengths. Extensions of the technique have been applied to radio antenna arrays as an alternative signal processing technique to the use of high-speed digital electronics. However, the inherently analog nature of the instrument along with the enormous advances in digital technology have reduced progressively those areas of application where the AOS has a distinct advantage over purely digital approaches. Nevertheless, the simplicity, compactness and a capacity to process a thousand or more resolution channels over a radio bandwidth of a gigahertz or more continue to make the AOS a useful instrument in MILLIMETER and INFRARED ASTRONOMY for both ground-based and space-based observations.

The core component of the AOS is the ultrasonic delay line in which a piezoelectric transducer bonded to an acoustic transmission medium converts the applied electrical signal into a travelling ultrasonic acoustic wave in the form of progressive wavefronts and troughs of increased and decreased material density. In a typical crystal medium where the acoustic velocity is of order 3 mm μs^{-1}, it is quite feasible to achieve (along a 30 mm length of medium) a 10 μs sample of the incoming signal. Through polished faces to the medium, this can be probed with an applied laser beam which has been expanded and collimated. The density variations are, in effect, an optical phase grating and cause scattering of the light. This is the acousto-optic interaction in which light is scattered at an angle from the applied light. The sine of this angle is directly proportional to the frequency of the applied radio signal. In the limit of small angles and small scattered light, energy at the various frequencies within the applied radio signal will produce a proportional amount of optical energy at angles proportional to the frequency. A lens is used to focus the scattered light onto an integrating, one-dimensional solid-state photodiode array for read-out to a computer or a display or for recording on a moving piece of film. That is, the one-dimensional distribution of light intensity is an optical representation of the power spectrum of the applied radio signal. The limit to resolution of two closely spaced spectral features is then the inverse of the interaction length which, in this example of a 10 μs system, would be of order 100 kHz.

The technique was first described by Lambert [1] in 1962 and first applied to RADIO ASTRONOMY in 1966 at CSIRO in Australia by Cole [2] using a water acousto-optic medium, with a 1 MHz bandwidth and by recording through a slit onto continuously moving film. It was reported to the 1966 URSI conference in Munich. An improved instrument using a fused silica medium giving 800 resolution cells over a 100 MHz bandwidth was reported in 1973 [3] which produced radio spectra of solar burst activity and led to various instruments being built for solar and Jupiter time-resolved radio emission spectroscopy. These were at observatories in the Soviet Union, Finland, China, Japan and Australia.

Progressive improvements to the AOS included better acoustic impedance matching between the transducer and medium in order to increase the fractional bandwidth of the interaction and the use of solid-crystal acousto-optic media such as lithium niobate, lithium tantalate and lead molybdate to improve the efficiency and spectral range over which the spectrograph would work.

However, it was the development of the one-dimensional solid-state photodiode array and the micro-computer which provided the improvements needed for accurate astronomical spectral line observations. By integrating the light onto the photodiodes and by use of calibration, it became possible to use the computer to integrate and subtract background in order to observe weak spectral lines. Although this was first demonstrated by adding a photodiode array to the existing solar instrument [4], the first serious AOS spectral line applications were carried out with the Parkes radio telescope in 1977 [5].

The clear advantages of the AOS in simplicity, spectral range and number of channels were apparent and resulted in a number of observatories in Japan, France, Germany and Australia constructing AOSs. As the bandwidth and channel capacity of the transducers and available photodiode arrays grew, the instruments became focused more and more on the challenges of millimeter wavelength spectral line observations.

However, other, unique results were obtained with the AOS. The capacity to obtain high-time-resolution and frequency-resolution spectra over wide bandwidths revealed very clearly the structure of interplanetary SCINTILLATION phenomena previously only interpreted from observations taken at widely spaced, discrete frequencies [6]. Similarly, much detail was revealed in real-time and very wide fractional bandwidth observations of PULSAR SIGNALS [7]. Such observations revealed in a very graphic way the complex phenomena of pulsar signals including interstellar dispersion, Faraday rotation and scintillation.

Extension of the technique to the processing of RADIO INTERFEROMETER data becomes possible by a number of techniques which would include construction of a series of acoustic beams displaced in an analogous manner to the radio antennas along a one-dimensional antenna array. In this case the optics can display a two-dimensional image of the spectrum as before but, orthogonally, radio spatial image data. However, although demonstrated in principle, the inherently analog nature of the technique has not been able to achieve the extremely high dynamic ranges now required in radio astronomical imaging and able to be achieved with digital processing techniques.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The usefulness of the AOS therefore becomes apparent as observations move into the sub-millimeter and infrared spectral ranges where the spectral details require very wide spectral coverage combined with a large number of resolution channels. Digital circuitry does not currently operate at such speeds and any digital solution becomes bulky and consumes significant power. Both of these commodities are scarce on the spacecraft needed to take the observing instrument above the Earth's absorbing atmosphere.

The AOS therefore continues to form a useful niche technique within radio astronomical observation. It provides access to a range of frequencies, resolution and channel number not easily possible by other techniques. However, the inexorable progress of digital techniques has restricted progressively the area of its unique advantage to the more exotic fields of space, millimeter and infrared astronomy.

Bibliography

[1] Lambert L 1962 Wide-band, instantaneous spectrum analyzers employing delay-line light modulators *IRE Nat. Conv. Rec.* **10** (part 6) 69–78
[2] Cole T W 1968 Electro-optic spectrograph for radio astronomy *Opt. Technol.* **1** 31–5
[3] Cole T W 1973 An electrooptical radio spectrograph *Proc. IEEE* **61** (9) 1321–3
[4] Cole T W 1975 Electro-optical processing in radio astronomy *Opt. Acta* **22** (2) 83–92
[5] Cole T W and Milne D K 1977 Acousto-optical radio spectrograph for spectral integration *Proc. ASA* **3** (2) 108–11
[6] Cole T W and Slee O B 1980 Spectra of interplanetary scintillation *Nature* **285** (5760) 93–5
[7] Cole T W 1980 The AOS: widening the radio sky *Sky Telesc.* **60** (2) 108–10

Trevor Cole

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galactic Nuclei: Variability

Variability studies have been essential in understanding the physics of ACTIVE GALACTIC NUCLEI, AGN, because the size of the region emitting the enormous amount of energy characteristic of AGN is too small to be resolved even with existing or planned optical/infrared (IR) INTERFEROMETERS.

AGN emit a continuum radiation coming from the vicinity of the BLACK HOLE. They also emit intense and broad emission lines originating in fast-moving gas clouds located in a small region around the continuum source: the broad line region or BLR.

The flux level of the continuum and the intensity of the spectral emission lines emitted by AGN undergo variations on timescales ranging from hours to years depending on the energy range and on the intrinsic luminosity of the AGN.

When the flux emitted by a source of energy varies significantly (say by a factor of 2) with a timescale t this flux variation sets a limit to the size R of the emitting region such that $R \leq ct$. In AGN the combination of the enormous amount of energy emitted and the very small size of the emission regions set by the short variability timescale is not compatible with the energy being produced by normal stellar processes. The most popular model for the energy production in AGN and the basis of the current paradigm is that AGN/quasars are powered by accretion of matter onto a SUPERMASSIVE BLACK HOLE.

At higher photon energies, more rapid and higher amplitude variability is observed, e.g. the x-rays vary more than the optical (figure 1). The variability timescales tend to be longer in intrinsically brighter objects. The timescales and the correlations and delays between the variations of the continuum flux and of the various spectral line intensities provide information on the nature of the physical processes emitting the lines and the continuum, and on the location of the regions from which they originate.

The term AGN is used here regardless of redshift and luminosity and therefore encompasses the terms Seyfert 1 and QSO or quasar, which are often used to designate medium- and high-luminosity AGN separately. In general, there are more data and of higher quality available for Seyfert 1s than for quasars. The reason is that they are numerous and quite a few are located within 100 Mpc of the Earth and are easily observable with the current instrumentation. In contrast, quasars are rare and even the closest ones are at such large distances from us that the flux received on Earth is much less than that from Seyfert 1s (there is one exception, 3C 273). In summary, the body of knowledge of AGN variability is dominated by the observations of Seyfert 1s, which have led to physical models which satisfactorily explain the observed phenomena. It remains to be investigated how these models apply to much brighter and to much fainter AGN.

The AGN that are the subject of this article are those in which the central optical, UV, and x-ray source is viewed directly and where the variations are not amplified by relativistic effects (see ACTIVE GALAXIES: OVERVIEW; UNIFIED MODEL). For nearly a dozen Seyfert 1s, and especially for three of them (NGC 4151, NGC 5548 and NGC 7469), and for the quasar 3C 273, large international collaborations have organized long coherent campaigns of observations, some comprising simultaneous observations in the optical, UV and x-ray ranges, producing large amounts of data of uniform quality and gathered with optimum time sampling. A number of Seyfert 1s (∼30) have been monitored less extensively but still sufficiently to establish that the results from the best studied Seyfert 1s are representative.

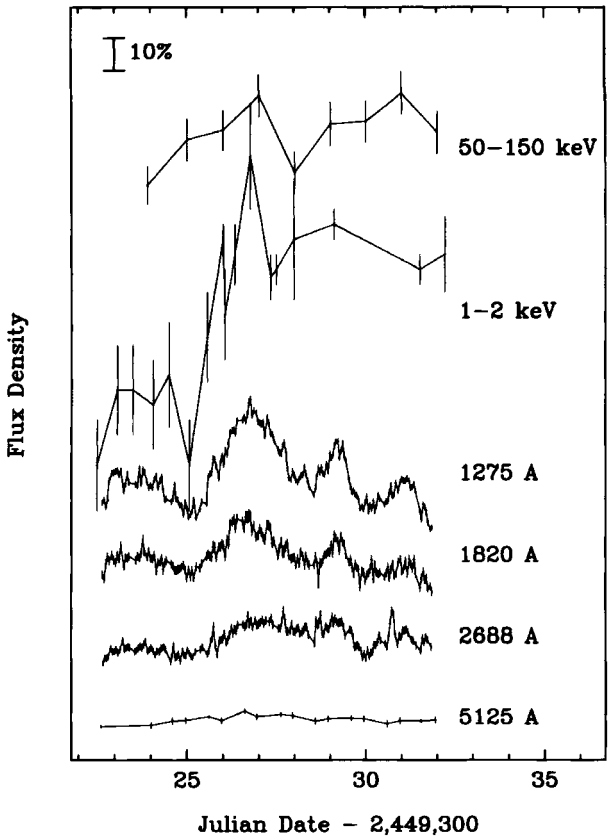


Figure 1. Variations of the continuum flux at different energies during a 10 d intensive monitoring of the AGN NGC 4151. The light curves are shifted vertically for convenience and a 10 % change is indicated by the bar. (From R A Edelson *et al* 1996 *Astrophys. J.* **470** 364)

Variability of the continuum flux

The electromagnetic spectrum (continuum spectrum) of radio-quiet AGN, after subtraction of the stellar continuum emitted by the host galaxy, extends from ∼1 mm to ∼100 keV with a prominent broad peak in the UV–extreme UV (EUV) range and a secondary peak in the IR (see ACTIVE GALAXIES: OBSERVATIONS). The x-ray emission is dominated by a power law component (1–10 keV; $\alpha \simeq 0.9$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where $f_\nu \propto \nu^{-\alpha}$), which steepens in the hard x-ray range (>50 keV), possibly with a cutoff around 100 keV.

It is recalled here that in the theory of accretion disks, the accreted matter spirals in towards the center under the gravitational attraction of the black hole. The matter, however, does not flow freely. It has an internal stress (very likely of magnetic origin) whose dissipation produces heat, and also sets the timescale of the propagation of the accreted matter towards the center. Astronomers talk about the disk being heated by viscosity and about viscous timescales, and theoreticians can calculate the energy produced by viscosity, the resulting spectrum emitted by the disk, and the timescales on which the radiation from the disk varies in response to a varying accretion rate (ACCRETION DISK). The observed spectral shape of the continuum of AGN from 1200–5000 Å (or 10–2.5 eV) is roughly consistent with thermal emission from an accretion disk. However, the variations of the UV continuum flux occur on timescales of days, weeks or months which are shorter by many orders of magnitude than the timescales expected from variations in the accretion rate. Remarkably, in NGC 4151 and NGC 5548, the medium energy x-ray flux (2–10 keV) and the optical/UV flux vary simultaneously (figure 1) with the high-energy variations being much larger than the low-energy variations.

Taken together, these observations suggest a model where the surface of the disk is irradiated by the central variable x-ray source.

The x-ray source variability itself is believed to be due to explosive reconnections of magnetic-field lines permeating the accretion disk and the corona (the gas of hot electrons surrounding the central parts of the disk) through a phenomenon reminiscent of the solar flares in the solar corona. A significant fraction of the accretion energy is spent in this way in heating the corona.

The hot electrons transfer some of their energy to ambient UV photons (emitted by the disk) and soft x-ray photons and produce medium and hard energy x-ray photons via inverse Compton emission. This 'Comptonization and reprocessing model' explains the simple correlation among x-rays and UV emission on short timescales in NGC 4151 and NGC 5548, and also explains the average slope of the x-ray spectrum, the high-energy cutoff and other features of AGN x-ray spectra.

But life is not so simple and the good correlation x-rays vs UV flux observed in NGC 4151 and NGC 5548, might have been a lucky circumstance in the intricate interplay between the hot electron gas, the UV photons, and the x-ray photons ('which is the chicken and which is the egg?'). In fact, in a third Seyfert 1 nucleus extensively observed simultaneously in the UV and x-ray, NGC 7469, the UV variations sometimes lag behind the x-ray variations and sometimes are simultaneous with the x-ray variations. This case places some serious doubt on the reprocessing picture provided earlier. A close look at the data, however, reveals a striking similarity of the variations in the two energy ranges, which suggests a basic link between the emission in the two energy ranges such as generally provided by Comptonization.

The most extreme soft x-ray variability occurs in narrow-line Seyfert 1 galaxies (NLS1), a subset of AGN with very intense soft x-ray emission and very steep soft x-ray spectra, narrow optical emission lines and prominent optical Fe II emission. To explain the absence of broad lines in NLS1, it has been proposed that the intense soft x-rays could blow away the inner broad-line region (BLR) or ionize it to states currently undetectable. The steep soft x-ray spectrum of NLS1 may indicate a high accretion rate or a small black hole mass.

Turning now to the infrared range, variations, when detected at all, are of smaller amplitude and longer timescale than in the optical, and are consistent with being the delayed response of dust around the central source to the long-term variations of the UV and x-ray continuum flux.

The emission line variations

A method to estimate the black hole mass

Monitoring of the flux variations of the continuum and the spectral line intensities of several Seyfert 1s have shown that the intensities of the spectral lines mimic the variations of the optical/UV continuum but with a certain time delay τ. This delay is interpreted as the time needed for the light from the central continuum source located in the immediate vicinity of the black hole to reach the gas clouds emitting the spectral lines (and located around the central continuum source) and increase or decrease the emissivity of the gas (figure 2).

The time delay τ gives a measure of the distance $d \sim c\tau$ of the line-emitting gas clouds from the central continuum source. This method can be expanded to map the line emission region and it is known as 'reverberation mapping'.

A rough estimate of the mass of the central black hole can then be calculated from the variations of the variable component of an emission line by combining the component width (which gives the virial velocity of the gas clouds) and τ (which is the delay by which the component variations follow the continuum variations). The assumption that the clouds' motions are orbital motions in the gravity field of the black hole is strongly supported by the fact that the blue and red wing of any line or line component vary together (i.e. one wing does not vary before or after the other one, to first order; see below).

This method of determining the mass of the central black hole has been applied to more than a dozen AGN and yields values in the ranges a few $10^6 M_\odot$ to a few $10^8 M_\odot$.

The velocity field of the gas

One of the crucial elements contributed by variability studies is that the broadest emission lines vary with the largest amplitude and on the shortest timescale, indicating that the fastest moving gas is closest to the center.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

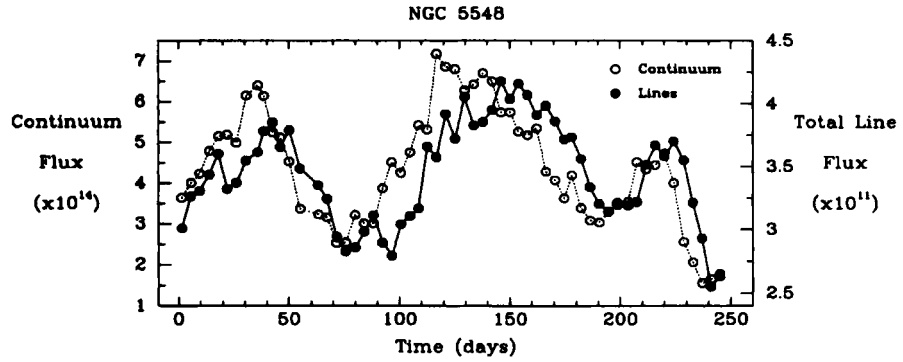


Figure 2. The UV continuum flux (open symbols; left vertical scale in erg s^{-1} cm^{-2} A^{-1}) and the emission line intensities (here the sum of all the lines; dots; right vertical scale in erg s^{-1} cm^{-2}) as a function of time for the AGN at the center of the galaxy NGC 5548. The delay of ∼15 d between the two light curves means that the lines are formed at a distance of ∼15 light d or 5×10^{16} cm. (From D Maoz 1997 *Emission Lines in Active Galaxies: New Methods and Techniques. ASP Conf. Ser.* **113** ed B M Peterson, F-Z Cheng and A S Wilson (San Francisco, CA: Astronomical Society of the Pacific) p 138.)

The detailed variations in line intensity and profile provide the answer to the question which is critical to the determination of the mass, as to whether the clouds are gravitationally bound to the central black hole: The line profile variations do not show the systematic effect of one wing responding before the other that would be expected for radial flows: The blue wing would be observed to vary first in the case of outward flows and spherical winds, and the red wing would be observed to vary first in the case of spherical accretion[1]. This, and the fact that the fastest moving gas is closest to the center, strongly suggest that the high ionization clouds are in orbital motions (chaotic or rotation motions) in the gravitational field of the black hole, with only minor components of net infall or outflow. Chaotic motions are ruled out because the gas clouds would be destroyed when crossing the accretion disk.

The most detailed result on the velocity field in the BLR comes from an intensive monitoring of the C IV line profile during 'a month in the life of NGC 4151', with nearly complete 1 d sampling over a 35 d interval. The analysis of the variations in each interval of velocity in the C IV line profile shows the variations to be consistent with gravitational motions around a mass of $\sim 10^7 M_\odot$. This value agrees with that resulting from observations of the entire UV spectrum of NGC 4151 at other epochs, which show a linear relation between the response time of a line and the inverse square of the line width. A similar study of NGC5548 yields a mass of $(5 \pm 3) \times 10^7 M_\odot$. Note that in both NGC 5548 and NGC 4151, there is possibly a weak signature of an infall component. Other AGN give consistent results (albeit less detailed) and it is tentatively concluded that the line width is dominated by rotation. This forms the basis of the estimate of the black hole mass from the response of the emission line intensities to the continuum flux variations. The uncertainties attached to this method are difficult to evaluate and are probably less than a factor 10.

[1] Take the example of an expanding sphere of gas whose emissivity varies under the influence of the varying central continuum. A signal from the central source—for example, an increase of flux—reaches all the points of the sphere simultaneously, but the first signal which we receive from the sphere originates from the part closest to us and which is in motion towards us. We will, therefore, first observe an increase of the short-wavelength part, or blue wing, of the emission line.

The 'disk and wind' model

In AGN spectra there are low-ionization lines, such as the Balmer lines and lines of Fe II and Mg II, and high-ionization lines (the strongest are lines of C IV, N V, He II, Si IV). Interpretation of the observations show that the low-ionization lines come from a flat aggregate of very dense clouds ($N_e \geq 10^{11}$ cm^{-3} and total optical thickness of 10^{24} cm^{-2}) which is identified with the relatively cool part (surface temperature of 5000 K) of the accretion disk.

The high-ionization lines, in contrast, come from a comparatively more dilute medium, above and below the disk, the lines of the most highly ionized elements being the broadest. This indicates that the degree of ionization is correlated with the gas velocity (figure 3).

Remarkably, the high-ionization lines are not exactly at the redshift of the low-ionization lines (emitted by the accretion disk), which themselves are at the host galaxy redshift. They are blueshifted (by 0 to ∼1500 km s^{-1}) with respect to the disk and the host galaxy, indicating that the highly ionized gas is flowing out towards us. (By symmetry, a flow is probably moving away from us, on the other side of the disk, but is hidden from view because the disk is opaque). The combination of rotation and outflow from the disk suggests and is consistent with a model where the high ionization lines are emitted by gas clouds emerging from the disk, still retaining a large part of the angular momentum that the gas had in the disk.

Such motions are predicted by the magnetohydrodynamical model of accretion disks in which magnetically accelerated winds are formed from filaments pulled from the originally dense low-ionized material of the accretion disk and subjected to the intense central ionizing field. This

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

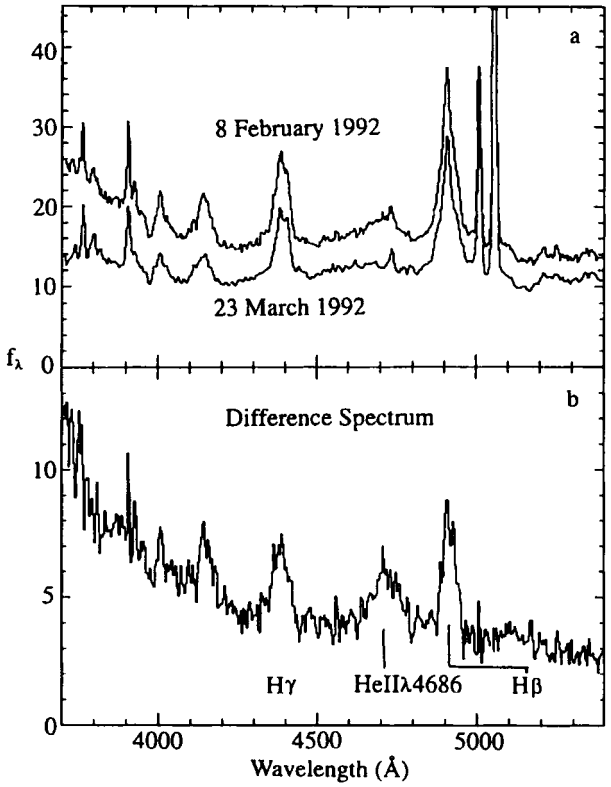


Figure 3. The variations over 44 d of the optical spectrum of the active nucleus of NGC 3783. The difference between the two spectra illustrates the large width and variability of the high ionization line of He II as compared to the other lines. The continuum spectrum is steeper at high state, a general property of AGN. Same unit of flux on both panels, 10^{-15} erg s^{-1} cm^{-2} A^{-1}. (From G M Stirpe *et al* 1994 *Astrophys.J.* **425** 609)

forms a highly ionized wind above and below the disk. The physics of these phenomena, however, is complicated and many features of this promising model remain unspecified at present.

Current problems and future developments

There remain a number of questions and exciting avenues of investigation.

First, the theoretical investigation of the magneto-hydrodynamical effects which, together with the gravity field of the black hole, dominate the physics in the black hole surroundings needs to be vigorously pursued.

Second, the variability phenomenon has been best observed and is best understood in the case of the medium luminosity AGN (total luminosity $L \sim 10^{42}$–10^{45} erg s^{-1}), but variability data on powerful quasars (up to 10^{49} erg s^{-1}) and on very low luminosity AGN (10^{39} erg s^{-1}) are still very few. Fortunately, the relevant observations are under way. They will make it possible to investigate the entire parameter space defined by the black hole mass and the accretion rate and to understand the specific characteristics of the accretion process in the different regions of this parameter space (e.g. advection, accretion of single stars, super-Eddington accretion rate).

Finally, although this article is focused on the directly observable variability of AGN, i.e. occuring on timescales of hours to years, it is pointed out that other variability phenomena must occur on longer timescales, say 10^2, 10^6 or 10^9 yr. Particularly relevant to this is the recent discovery (from measurements of the velocities of stars in the nucleus of nearby galaxies) that black holes are likely to be present in the centers of non-active galaxies, including our own Galaxy, with masses roughly in the same range of values as those determined from AGN investigations. A number of questions then naturally arise such as to whether AGN could be recurrent phenomena with dormant phases, and whether they can be fuelled by gas produced through the evolution of the stars in the host galaxy or whether the accreted material always come from outside the host galaxy.

Bibliography

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

An assessment of the current state of the field with numerous references is given by:

Ulrich M-H, Maraschi L and Urry C M 1997 Variability of active galactic nuclei *Ann. Rev. Astron. Astrophys.* **35** 445–502

Marie-Helene Ulrich

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galactic Nucleus

The central region of an active galaxy within which exceptionally large amounts of energy are being generated. An active galactic nucleus (AGN) emits strongly over a wide range of wavelengths, from x-ray to radio, the most powerful examples radiating a thousand times as much energy as the galaxies within which they are embedded. Typical luminosities range from about 10^{37} to 10^{39} or even 10^{40} watts. The spectrum of the radiation emitted by an AGN is markedly different from that of an ordinary galaxy (which shines with the combined light of its constituent stars and nebulae) and implies that some mechanism other than stellar radiation is needed to explain its copious output of energy. Many AGNs appear to be ejecting jets of highly energetic particles. In some cases one jet is seen, in others two oppositely directed jets are visible.

Many AGNs vary in brightness by substantial amounts over timescales as short as years, months, days, or even hours. Because a source of light cannot vary in brightness on a timescale shorter than the time taken for light to cross the diameter of the source, these short-term variations imply that the energy sources in AGNs must be very compact. For example, if an AGN varies in one year, its energy source must be no more than a light-year across; if it varies in one day, its energy source must be less than a light-day across. Consequently, some AGNs appear to be radiating up to a thousand times as much energy as a conventional galaxy from a region of space comparable in size to the solar system.

A widely supported hypothesis suggests that each and every AGN contains a supermassive black hole, with a mass of between ten million and several billion solar masses that is accreting matter from its surroundings. Because all galaxies are rotating, infalling matter will preserve its angular momentum (rotational motion) and will form a rapidly spinning disk of gas (an accretion disk) around the central black hole. Kinetic energy provided by infalling matter as it ploughs into the disk, and frictional effects within the disk, raise the temperature of the disk to very high values, causing it to radiate copious amounts of energetic radiation. The gravitational energy released by infalling matter is more than adequate to account for the power outputs of all types of active galactic nuclei. Because a black hole is so small compared with the size of a galaxy, and most of the energy is radiated from the inner regions of the accretion disk, this model neatly accounts for the very compact nature of the energy sources in AGNs.

By a process that is still not fully understood, the central 'engine' accelerates streams of charged particles to relativistic speeds (large fractions of the speed of light). The inner part of the disk is believed to form a nozzle that confines the outward flow of energetic particles into two oppositely directed jets that shoot out perpendicular to the plane of the disk.

Many astronomers believe that the same basic model (the 'unified model') can account for all kinds of active galaxy (Seyfert galaxies, radio galaxies, quasars, BL Lacs and blazars). According to this model, the supermassive black hole and its inner accretion disk is surrounded by a thick dusty torus ('donut') of matter, and the type of active galaxy that is seen depends on the orientation of the torus and jets relative to the observer's line of sight. The jet radiates most strongly along its axis. The view of an observer, who is looking along, or very close to, this axis, will be dominated by the head-on jet, and a violently variable source with no spectral lines (a BL Lac or blazar) will be seen. Looking at a modest angle to the jet, an observer will see an unobscured compact source (a quasar) inside the torus. From a viewpoint closer to the plane of the torus, the central 'engine' is hidden from view, and we see only the jets and lobes (extended radio-emitting clouds) of a radio galaxy. A similar picture applies to Seyfert galaxies, a type II Seyfert being seen when the central engine and the gas clouds in its vicinity are hidden, and a type I Seyfert when the observer's line of sight looks over the rim of the torus towards the innermost regions of the AGN.

There is strong observational evidence to support the suggestion that active galactic nuclei contain supermassive black holes. In particular, scans across the centers of a significant number of active galaxies have shown that the intensity of starlight rises sharply in a central brightness 'spike' in a way that is consistent with stars being bunched together in the gravitational field of a highly massive compact object. Furthermore, where it has been possible to measure the orbital velocities of gas clouds in the innermost cores of active galaxies, the results indicate that these clouds are moving in the gravitational fields of extremely massive compact objects. For example, measurements of the intensity of light and the rotational velocities in the core of the giant elliptical galaxy M87 (in the Virgo cluster) are consistent with the presence of a supermassive black hole of some three billion solar masses. Furthermore, high-resolution images obtained by the Hubble Space Telescope have revealed flattened disks of gas and dust in the cores of a number of active galaxies.

See also: accretion, active galaxies, astrophysical jets, black hole, blazar, BL Lacertae objects, quasar, radio galaxy, Seyfert galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galaxies: Observations

'Active galaxy' is a general term which refers to any galaxy that produces significant emission in addition to that from its constituent stars, stellar remnants and interstellar medium. The characterization of such galaxies as 'active' is because the unusual emission characteristics are generally assumed to be associated with high-energy, eruptive phenomena. The earliest consistent user of the word 'activity' as applied to the nuclei of galaxies seems to have been V A Ambartsumian.

In most cases, this non-stellar emission appears to originate in the very center, or nucleus, of the galaxy, and these sources are known generically as 'active galactic nuclei' (AGN). A galaxy that harbors such a source is known as the AGN host. Historically, some other types of sources, such as STARBURST GALAXIES, that have some phenomenological similarities to AGN (e.g. strong near-UV emission) have also been known as active galaxies, though in these cases the activity might not be confined to the nuclear regions.

AGN are the most luminous long-lived sources in the universe. They emit strong radiation over the entire observable wavelength range, from x-rays and γ-rays through long-wavelength radio. A complete picture of the emission of an AGN can be obtained only by observing it at many wavelengths, preferably simultaneously over the entire spectrum because they are variable sources. There are a number of different classes of AGN (see ACTIVE GALAXIES: OVERVIEW); SEYFERT GALAXIES constitute the low-luminosity end of the AGN phenomenon, i.e. their nuclear luminosities are roughly comparable with the total starlight from the host galaxy. Higher-luminosity AGN are known as 'quasars' or 'quasi-stellar objects' (QSOs) (see QUASISTELLAR OBJECTS: OVERVIEW) with the former being reserved (originally) for the stronger radio sources, and the latter weaker; in these cases the optical light from the AGN can exceed, sometimes by factors of more than 100, the stellar light from the host galaxy at all wavelengths. There are a wide variety of subclasses within these groups, based on the relative prominence of UV–optical emission lines, radio properties and polarization (see POLARIZATION IN ACTIVE GALAXIES). These will be mentioned below as necessary.

This section will discuss the current status of observations over the electromagnetic spectrum. We first discuss the continuous emission of AGN in terms of their 'spectral energy distributions' (SEDs), i.e. amount of energy emitted in various wavebands. We then focus on the prominent emission lines that are detected in the UV–optical (and, more recently, x-ray) spectra of most AGN and follow this with a brief description of absorption features in AGN spectra. Spatially resolved structures in AGN are also discussed.

Spectral energy distributions

The SEDs of normal stars (and galaxies) are well approximated as blackbodies in the temperature range 2000–50 000 K, and thus their emission is strongly concentrated in the ultraviolet (UV) through near-infrared (IR) parts of the spectrum. In contrast, AGN emit comparable energy (per unit logarithmic bandwidth) over most of the observable spectrum, as can be seen in figure 1. The one exception is the radio region, in which 'radio-loud' (RL) AGN are some 3 orders of magnitude brighter than their 'radio-quiet' (RQ) counterparts. These RL AGN constitute ~10% of the population. Figure 1 shows the low-redshift median SED for RL and RQ objects from a sample of more than 40 AGN. It is plotted as $\log(\nu L_\nu)$ versus $\log \nu$, which shows the energy output in each waveband and also highlights the important structure in the SED.

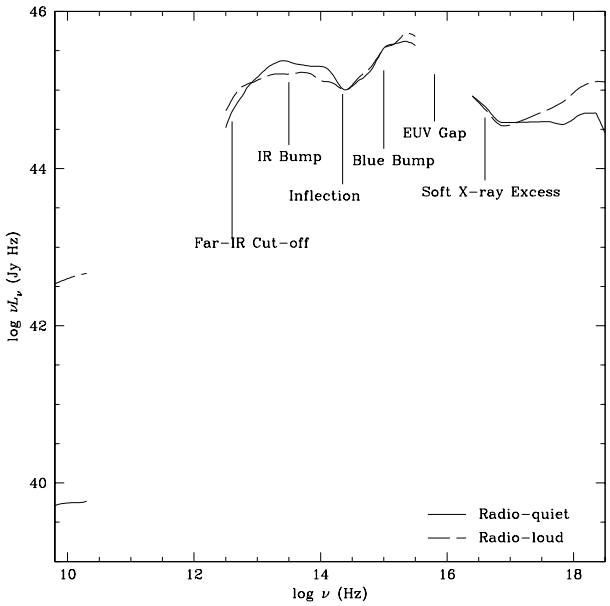


Figure 1. The median radio–soft x-ray spectral energy distribution for radio-loud (dashed) and radio-quiet (solid) low-redshift quasars. Data from Elvis *et al* 1994 *Astrophys. J. Suppl.* **95** 1.

Our knowledge of these SEDs is limited by current observational technology to fairly low-redshift, bright sources and includes (as is clear in figure 1) important gaps in various wavebands.

1. The absence of observations at high energies (γ-rays) is due to current technical limitations; higher-energy detectors are not sensitive enough to detect most AGN. Those that have been detected so far are generally core-dominated (CD) RL AGN, i.e. those whose emission is boosted by a relativistic flow along our line of sight (see ACTIVE GALAXIES: UNIFIED MODEL for more detail).
2. The EUV gap, between the far-UV and soft x-ray regions of the spectrum (i.e. between 912 Å and about 0.1 keV, spanning a factor of about 10 in photon energy), is due to the large opacity of the interstellar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

medium in our own Galaxy to hydrogen-ionizing photons. Photons in this energy range are especially important because these drive the strong emission lines seen throughout the UV and optical spectra of AGN.

3. The gap between the submillimeter and short-wavelength radio is again due to technological limits. New detectors and satellites have recently reduced these limitations considerably and revealed sharp far-IR spectral breaks for all but the CD RL AGNs.

The most prominent feature of AGN SEDs is the strong peak in the UV spectrum. The feature, which is often known as the 'big blue bump' (BBB), begins in the near-IR and may extend all the way to soft x-ray energies. It is observed with a combination of low-resolution optical and UV spectroscopy and/or multicolor, optical photometry. It can be plausibly identified as arising in an accretion disk (AD) containing material with a wide range of temperatures and orbiting around a supermassive black hole. The BBB peaks in the extreme UV, which is consistent with the expected emission from an AD around a $\sim 10^8 M_\odot$ black hole that is accreting material at the Eddington rate.

The IR emission from RQ AGN is mainly observed via multicolor photometry, although IR spectroscopy is a rapidly expanding field. The IR spectrum is thought to be dominated by thermal emission from dust in the host galaxy or a molecular torus that surrounds the AGN. The most compelling evidence for this interpretation is as follows.

1. There is a local minimum in AGN SEDs at a wavelength of about 1 μm. This probably represents emission from the highest-temperature (about 1500 K) grains that can exist in the vicinity of an AGN; at higher temperatures, i.e. smaller distances from the central source, dust will sublimate.
2. In some AGN, the infrared continuum has been observed to vary, apparently in response to earlier, similar continuum variations in the UV–optical spectrum (see ACTIVE GALACTIC NUCLEI: VARIABILITY). The light-travel time inferred is consistent with the maximum distance at which grains exposed to the AGN radiation field should sublimate.
3. The sharp break in the far-IR–submillimeter region of the SED in all but CD RL AGNs is generally too steep to be ascribable to optically thick synchrotron emission or to single-temperature blackbody emission, but is consistent with thermal emission from grains (depending somewhat on the unknown grain emissivity law).

In RL AGN, the IR is probably a mix of isotropic thermal and beamed non-thermal emission. The beamed emission dominates in the CD RL AGN as it is boosted by relativistic outflow along our line of sight. Strong evidence for this interpretation is provided by the coordinated variability observed in several of these sources, most notably 3C 273, which is the brightest nearby QSO and consequently has been extensively monitored. Preliminary results in the mid- and far-IR from the ESA satellite *ISO* show the progression from non-thermal to thermal emission in a small sample of CD and lobe-dominated RL AGN and RQ AGN, supporting this general picture.

While the UV–optical–IR emission from AGN appears to be predominantly thermal in origin, both the high- and low-energy extremes of the observable SEDs are thought to be produced by non-thermal processes. X-ray observations of AGN are largely confined to very low-resolution spectroscopy which yields the global continuum shape. The principal feature of RQ AGN hard x-ray spectra is a power-law continuum ($F_\nu \propto \nu^{-\alpha}$, with $\alpha \approx 0.9$) over the range $\sim$1–100 keV. The spectrum generally turns over at energies $\gtrsim$100 keV and is thought to arise through repeated Compton scattering of thermal or non-thermal photons.

RL AGN show a harder x-ray spectrum than their RQ counterparts (Figure 1). The relative strength of this hard component increases with the core dominance of the radio source, implying that the emission is beamed similarly to the radio and supporting an origin in Compton up-scattering of the radio synchrotron photons. CD RL AGN are, to date, the only AGN detected in the high-energy γ-rays, implying that beaming is a key factor. A few very high-energy observations have been made, including ground-based detection of Cerenkov showers that occur when TeV energy photons hit the Earth's atmosphere. So far, the AGN detected by this method are all BL Lac objects (see BL LACERTAE OBJECTS). Similar to CD RL AGN, the observed emission from these sources is apparently dominated by a relativistic beam or jet (see ASTROPHYSICAL JETS) directed towards the observer.

At energies around 10 keV and above, RQ AGN spectra also have a 'hard tail' above the x-ray power law which is stronger in lower-luminosity sources. This is thought to arise through 'Compton reflection' (a combination of inelastic scattering of higher-energy photons and photoelectric absorption of lower-energy photons) off relatively cooler material. This feature is strong enough that the cooler material must cover about half of the sky as seen from the x-ray source itself, which has led to speculation that the reflecting body must be the AD itself, with the x-ray source located somewhere above it. An example of such a spectrum is shown in figure 2.

Spectral variability

One of the most remarkable characteristics of AGN is their strong flux variability. Indeed, the very existence of intraday continuum variability puts a severe upper limit (set by light travel time and source coherence arguments) on the size of the continuum source, and thus provides one of the strongest arguments for identifying an AD around a supermassive black hole as the source of continuum radiation. At a fundamental level, the cause of continuum

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

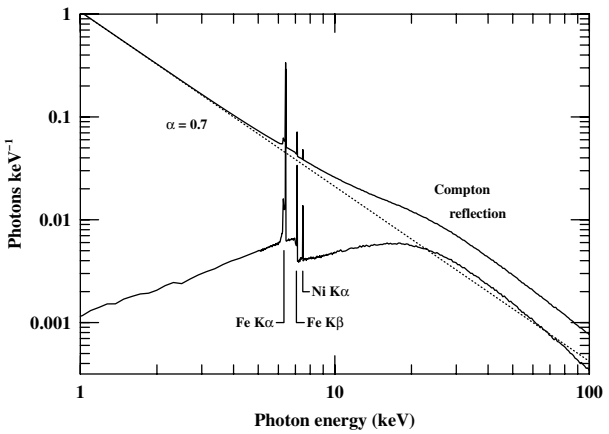


Figure 2. An x-ray reflection model with an input power spectrum $F_\nu \propto \nu^{-0.7}$ (dotted line) irradiates a slab of 'cold' ($T < 10^6$ K) gas in the limit of zero height on the disk axis. The lower spectrum shows the reflected spectrum, in which fluorescence lines appear prominently. At high energies, the reflected spectrum turns down as a result of Compton down-scattering within the slab. Data courtesy of I George, figure from Peterson (1997).

variability is not understood, although AD instabilities are sometimes invoked as a driving mechanism.

In general, the most rapid and highest-amplitude variations are seen at the highest photon energies. X-ray flux variations have been detected in Seyfert galaxies on time scales as short as several minutes. For phenomenological purposes, we can characterize the strength of the variations as a function of temporal frequency f by their 'power density spectra', which are conventionally modeled as a power law, i.e. $P(f) \propto f^{-\alpha}$ where α is typically in the range 1–2 for most AGN. The case $\alpha = 0$ is 'white noise', i.e. the amplitude of variation is independent of time scale, and $\alpha = 1$ corresponds to variations that can be described as a random walk. Larger values of α correspond to increased amplitudes of variability over longer time scales (lower frequencies f). In some cases, e.g. NGC 4151 and 3C 390.3, the most prominent variations seem to occur quasi-simultaneously (i.e. simultaneous to within less than a day) from x-rays (or even γ-rays for NGC 4151) through to optical wavebands. The smaller-amplitude structures in the light curves differ across the spectrum, with lower amplitudes and less structure seen at lower photon energies. The near-simultaneity of the variations requires that the variations are driven by radiation rather than, for example, by propagation of disturbances through the AD on the much-longer viscous time scale. These observations, along with detection of the x-ray hard tail and the Fe Kα emission line, point to 'reprocessing' models, in which a variable x-ray source illuminates the AD from above, with the absorbed x-ray energy driving variations observed at lower photon energies.

In at least one case (the Seyfert galaxy NGC 7469), however, the x-ray variations do not correlate well with the UV–optical variations; variations of similar fractional amplitude are seen in both the x-ray and UV–optical regions, but there is no clear causal relationship between them.

It is in this same object, however, that a time delay between UV and optical variations has been detected. This is the only Seyfert galaxy that has been monitored sufficiently well for such wavelength-dependent continuum time delays to be detected. Relative to the variations in the shortest-wavelength UV, variations at longer wavelengths are delayed by an amount $\tau \propto \lambda^{4/3}$, which is precisely the wavelength dependence expected if the variations arise in a thin AD that is irradiated by a variable x-ray source. Of course, this interpretation fails utterly to account for the observed lack of correlation between the UV–optical and x-ray variations. At this time, NGC 7469 seems to present serious difficulties for reprocessing models, although no other explanation of the coupling (or lack of coupling) between the x-ray and UV–optical variations has been forthcoming.

Extreme flux variations are observed in BL Lac objects as well. In this case, however, the fundamental origin of the variations is probably related to the propagation of shocks through the relativistic jet. Unlike the Seyferts, while there is a close correspondence between the variations in different bands, a completely consistent relationship between time delay and wavelength has not yet emerged; the same source can apparently show different behavior at different times, and the reason for this is not understood.

Emission lines

Basic properties of UV–optical lines

Strong broad emission lines are a defining characteristics of all types of AGN, except BL Lacertae objects in which the, presumably present, emission lines are swamped by the highly beamed continuum emission from the jet. In the UV–optical–IR part of the spectrum, the most prominent emission lines (those with equivalent widths[1] larger than some tens of Å) are usually Ly$\alpha\lambda$ 1215, C IV λ 1549, C III] λ 1909, Mg II λ 2798, Hβ λ 4861, Hα λ 6563 and Pα λ 1.88μm (figure 3). In most Seyferts and quasars, the observed emission arises in two physically distinct regions, a spatially compact (10^{15}–10^{17} cm) 'broad-line region' (BLR) with relatively high particle densities ($n_e \approx 10^{11}$ cm^{-3}) and large velocity dispersions ($v_{FWHM} \approx$ 1000–10 000 km s^{-1}), and a spatially extended (100 pc–1 kpc) 'narrow-line region' (NLR) with relatively low particle densities ($n_e \approx 10^3$ cm^{-3}) and smaller velocity dispersions ($v_{FWHM} < 1000$ km s^{-1}). Certain nebular 'forbidden' lines, such as [O III] $\lambda\lambda$ 4959, 5007, are prominent features of narrow-line spectra but are not observed in broad-line spectra. In the low-density NIR, both the Balmer lines and the [O III] are in the low-density limit in which the emissivity is proportional to n_e^2. However, in the denser,

[1] The equivalent width of an emission line can be thought of as the width (in wavelength units) of underlying continuum that would be required to produce the same emission-line flux.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

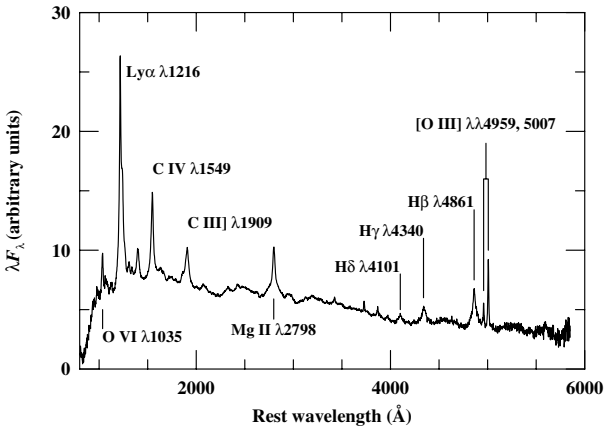


Figure 3. A mean optical–UV QSO spectrum formed by averaging spectra of over 1000 QSOs from the Large Bright Quasar Survey, with prominent emission lines marked. Data courtesy of S Morris and K T Korista.

but much lower-mass BLR, the Balmer lines are still in the low-density limit, but the [O III] lines are in the high-density limit in which the emissivity is proportional only to n_e. The [O III] lines from the broad-line region are thus very weak relative to the Balmer lines, and the forbidden lines are often referred to as being 'collisionally suppressed'.

Seyfert galaxies have traditionally been divided into two spectroscopic subclasses: type 1 Seyferts have both broad-line and narrow-line emission, whereas type 2 Seyferts have only narrow lines. In some and possibly all cases, the type 2 objects are those whose BLRs are obscured from our direct vision by dust. Spectra of quasars are always similar to those of type 1 Seyferts. The question of whether or not there are any *bona fide* 'type 2 quasars' remains open. Recent polarimetry has revealed broad lines in some ultraluminous infrared galaxies: spectroscopically these could sometimes be classified as the elusive type 2 quasars, but fundamentally these are still obscured type 1 quasars. Beyond semantics, the question is whether or not unified models explain all type 2 objects.

Emission-line variability

The broad components of AGN emission lines vary in flux in response to changes in the continuum flux, with time delays that are assumed to be due to light travel-time effects within the BLR. Indeed, measurement of the BLR size relies on measurement of the mean time delays (or 'lags') between continuum and emission-line variations. This process, known as 'reverberation mapping', has been carried out successfully for more than two dozen AGN. The BLR sizes in Seyfert galaxies are in the range of a few to around 100 light-days (i.e. $r_{\rm BLR} \approx 10^{15-17}$ cm), and scale with luminosity approximately as $r_{\rm BLR} \propto L^{1/2}$, consistent with the simplest theory. Different lines in a given AGN respond with different time delays, usually in the sense that higher-ionization lines respond faster than lower-ionization lines; this indicates that the BLR has a radially stratified ionization structure and that the maximum variability response of different lines occurs at different radii. The range of time responses is such that the outer edge of the BLR is at least 10 times larger than the inner edge.

In principle, reverberation mapping should be able to determine the velocity field of the BLR because the broad lines are resolved in line-of-sight (Doppler) velocity. For example, if the BLR clouds were traveling radially outwards from the central source, then an external observer would see the short-wavelength (relatively blueshifted) gas on the near side of the nucleus respond to continuum variations before seeing the long-wavelength (relatively redshifted) gas on the far side of the nucleus. The absence of such a strong signature indicates that the predominant gas motions are not radial. However, current observations are not extensive enough to determine whether the BLR gas is confined to a disk or has some different geometry.

Regardless of the details, if it is accepted that the BLR gas motions are primarily orbital around the central source, the mass of the central source ($M_{\rm black\,hole}$) can be measured by combining measurements of the size of the region with emission-line width, i.e. to within some geometry-dependent factor of order unity, $M_{\rm black\,hole} \approx r_{\rm BLR}\, v^2_{\rm FWHM}$. For the two dozen or so low-luminosity AGN for which this has been done, masses in the range (10^6–10^8)$M_\odot$ have been inferred.

The narrow emission lines generally do not vary on time scales of years or less because the longer light-travel and recombination times tend to smear out the effects of any continuum variations. Reports of narrow-line variability are rare. Probably the best case for actual narrow-line variability is the case of the broad-line radio galaxy 3C 390.3, which seems to have an abnormally compact, high-density NLR.

Narrow-line widths seem to correlate well with the central masses of the AGN host galaxy on scales comparable with the size of the NLR (i.e. the interior galactic bulge mass), although there is also clear evidence that part of the narrow-line widths and asymmetries is due to interaction with jets.

The x-ray iron (Fe) Kα line

In the low-resolution spectra typically obtained in the x-ray region by satellites such as ASCA (Japan), SAX (Italy) and RXTE (USA), a fluorescent Fe Kα emission line is commonly observed. This emission line tends to be stronger in lower-luminosity sources. The energy at which the line is observed crudely indicates the ionization state of the emitting material. A wide range of behavior is seen from unresolved, cool lines (i.e. from low-ionization material), consistent with an origin in absorbing material along our line of sight through the quasar, e.g. the molecular torus, to hot, resolved lines with widths, $v_{\rm FWHM} \gtrsim 10\,000$ km s^{-1}, which imply material much closer to the central AGN than the UV and optical lines discussed above. In a few of these, most notably MCG-6-30-15,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

complex line profiles are observed which suggest emission from very close to the central AGN. This interpretation is very attractive as the line then provides our only direct observation of the inner AD and is extremely important for diagnosing and constraining both AD and central-source models.

X-ray satellites, such as XMM (ESA) and the Chandra X-Ray Observatory (NASA), both launched in 1999, will be able to observe Fe Kα emission in a larger fraction of the AGN population. Their gratings will also provide extremely high spectral resolution ($E/\Delta E > 100$–1000), allowing more detailed profile studies as well as the detection of more and weaker lines in bright, nearby AGN.

Spatially resolved structures in AGN

The narrow-line region

Whereas the size and structure of the BLR must be inferred indirectly by processes such as reverberation mapping, the NLR is sufficiently extended that, in the nearer AGN, it can be spatially resolved with the Hubble Space Telescope or, in some cases, even with ground-based observations. The NLR is typically approximately axisymmetric, with its long axis coinciding with the radio axis in those sources in which extended radio emission is detected. In some cases, there appear to be shock fronts at the interface between the radio-emitting plasma and the interstellar medium in the host galaxy. While the narrow-line emission is apparently driven primarily by photoionization by the central source, strong emission from post-shock cooling regions is also detected.

In some cases, narrow-line emission seems confined to wedge-shaped 'ionization cones' that emanate from the nucleus. The opening angles of these cones are typically 30°–100°. These cones are thought to be a result of anisotropy in the AGN radiation field that is introduced on much smaller scales by an obscuring torus that surrounds the central source.

AGN host galaxies

In the present-day universe, AGN constitute fewer than 5–10% of all bright galaxies. Why are some galaxies 'active', whereas others are not? Does the AGN phenomenon represent a transient phase in galaxy evolution (noting that there were far more AGN in the distant past than there are today? Attempts to address these problems have led to studies of the 'host galaxies' of AGN. This is often a difficult undertaking, since AGN are typically at large distances, so the host galaxies are faint and have small angular sizes, and, in the case of quasars in particular, much of the host-galaxy light is simply lost in the glare of the quasar.

Carl Seyfert's original list of high-surface-brightness emission-line galaxies was composed almost entirely of spiral galaxies. Over time, the definition of 'AGN' has become primarily spectroscopic (based on the presence of strong emission lines, except in the case of BL Lac objects). High-luminosity AGN are found in both elliptical and spiral galaxies, although at low redshift RL AGN are found preferentially in ellipticals.

Until relatively recently, it was generally believed that there is a correlation between AGN and host-galaxy luminosity: more luminous AGN reside in more luminous galaxies. There are, however, strong selection effects that can lead to such a conclusion. Recent observations of quasars with the Hubble Space Telescope (to aid detection of faint host galaxies with high-luminosity AGN) and more through survey work on nearby galaxies (to find weak AGN in luminous hosts) have not supported the existence of a strong correlation between AGN and host-galaxy luminosities. Faint AGN are found in both low- and high-luminosity host galaxies, whereas high-luminosity AGN are found only in more luminous hosts.

As a result in a large part of observations with the Hubble Space Telescope, it now seems that virtually *all* galaxies harbor supermassive black holes in their nuclei (see SUPERMASSIVE BLACK HOLES IN AGN). Identification of supermassive objects requires measurement of high velocity dispersions on small spatial scales (thus requiring the high angular resolution obtainable with space-based imaging), leading to virial mass measurements. The most interesting question now seems to be shifting from 'do supermassive black holes exist?' to 'why are some supermassive black holes accreting mass (AGN) whereas others are not (normal galaxies)?' Whether or not the nucleus is fueled may be related to how easy it is to drive gas into the galaxy nucleus, and this may in turn depend on host-galaxy characteristics.

Extended x-ray emission

The high (for the x-ray) spatial resolution of the High Resolution Imager on the ROSAT x-ray satellite (~5″) facilitated imaging of bright, nearby Seyfert galaxies such as NGC 1068 and NGC 4151. X-ray imaging shows that a significant fraction (up to ~50%) of the x-ray emission can originate in an extended region around the central AGN. In a few sources, this x-ray emission is spatially correlated with the NLR. The origin of the extended x-ray emission could be due to electron scattering of the central x-ray emission or to x-ray emission from a hot, outflowing wind. The Chandra X-Ray Observatory (NASA) with its unprecedented spatial resolution (~0.5″) will allow many more of these spatial studies to be made.

Megamaser emission

Among the most exciting observations of AGN in recent years are those of water maser emission in the radio regime. Radio observations, because of the long wavelengths and correspondingly larger telescopes that can be built or synthesized[2], are currently unsurpassed

[2] The required surface accuracy of a telescope is proportional to the wavelength at which it is to be used. The relatively crude surface accuracy required for a radio telescope means that large collecting surfaces are comparatively easy and cheap to fabricate and maintain. Furthermore, the low frequencies involved make it relatively easy to correlate signals from individual radio receivers separated by large distances and thus attain higher angular resolution through interferometry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the spatial resolution they can obtain, typically down to milliarcsecond (mas) levels. Recent observations of water maser emission in the low-luminosity Seyfert 1.9 galaxy NGC 4258 has shown several components whose velocities accurately follow a Keplerian rotation curve. Combined with their angular positions, these maser sources indicate a warped, inclined, geometrically thin, sub-parsec-scale, molecular disk, and lead to an estimate of $(3.5 \pm 0.1)\times10^7 M_\odot$ for the mass of the central black hole, the first direct and precise such measurement. Maser observations of other sources have also provided information on similar scales, but so far none as detailed as NGC 4258.

Absorption lines

Optical and UV

The optical and UV spectra of AGN contain absorption features due to material along the line of sight between us and the optical–UV continuum source both within the AGN itself and in intervening space.

The intervening absorbers, whose redshifts are always lower than those of the background QSO, are most numerous in high-redshift QSOs with their correspondingly longer line of sight. They are dominated by lines from the fundamental transition of neutral hydrogen, Lyα λ 1215. Known as the 'Lyman α forest', these lines provide a wealth of information concerning the distribution of cold material in the universe. The range of column densities in these absorbing systems is wide, $N_{\rm H} \approx 10^{13}$–$10^{21}$ cm^{-2}, so that metal features, Mg II λ 2798, C IV λ 1549 and Lyman limit breaks are observed in a subset. At high redshift, where only the brightest sources, mostly QSOs, are visible to us, the Lyman α forest provides our only view of the cold material and so is a very important window on the early universe for cosmological studies (see QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES).

The most interesting class of absorbers, in terms of QSO studies, is that of associated–intrinsic absorbers. These are high column density, metal line systems which occur in more than 50% of all AGN and whose redshift is close to and sometimes even slightly in excess of that of the background QSO. In UV–optical spectra with sufficiently high resolution ($\lesssim$5 Å), the lines are often visible within the broad emission-line profiles. This makes them difficult to study as neither their profiles, which often include multiple components, nor those of the emission lines they absorb are known. High-resolution and signal-to-noise data are required and, even then, the uncertainties tend to be high. The absorption is thought to originate in material close to the nucleus of the QSO and moving relative to it, usually in an outward direction. They show a range of ionization from low, where Mg II is the dominant line, to high, where C IV or even O VI λ 1034 predominate. These high-ionization absorbers are thought to also be responsible for the x-ray warm absorbers discussed below.

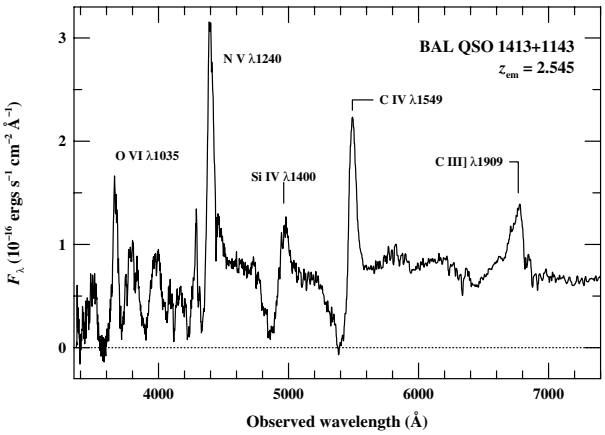


Figure 4. An example of a QSO spectrum showing BALs. Data courtesy of C B Foltz, figure from Peterson (1997).

X-ray warm absorbers

The soft x-ray spectral region is strongly affected by atomic absorption due to any material along the line of sight between us and the x-ray source within the AGN itself. Our Galaxy has a significant column density of neutral (cold) material whose signature is present in AGN soft x-ray spectra (in the range 0.1–5 keV). The host galaxy of the AGN produces a similar signature. In addition to this cold absorption, absorption by ionized (warm) material is often observed in the soft x-ray spectra. This is generally evidenced by the presence of strong absorption at the edge(s) of highly-ionized oxygen (O VII or O VIII) or, in lower-resolution data, a 'leakage' of soft photons above the expectations of cold absorption since the low-energy opacity decreases as the ionization increases.

Indications are that all or nearly all QSOs showing x-ray warm absorption also have associated absorption lines in the UV which include high-ionization lines such as O VI and N V λ 1240. Recent work has shown that high-column-density, outflowing material can produce both the x-ray and the UV features of such absorbers. The resulting combination of UV and x-ray constraints on the absorbing material has provided us with unprecedentedly strong diagnostics for the absorber. It appears to be high-column-density ($N_{\rm H} \approx 10^{20-22}$ cm^{-2}) material, outflowing from the central regions at velocities $\sim$100–1000 km s^{-1}, located between the BLR and NLR at $\approx10^{17}$ cm and often including several distinct components within an individual system.

Broad absorption lines

The most dramatic absorption features seen in the UV–optical spectra of QSOs are the broad absorption lines (BALs) (figure 4). These are high-column-density ($N_{\rm H} \approx 10^{20-23}$ cm^{-2}) intrinsic absorbers outflowing from the QSO nucleus. They have a range of ionization (hence C IV or Mg II BALs) and a wide range of velocities sometimes approaching $\sim$10% of the speed of light. The absorption generally contains a great deal of structure although a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

small subset have smooth P-Cygni-type profiles indicating a smooth, spherical outflow. Since they absorb out so much of the QSO continuum and line emission, it is very hard to study them quantitatively. High-resolution and high-signal-to-noise spectra are essential along with detailed modeling of both emission and absorption features in order to build up information on the column densities as a function of velocity for this material. Interpretation is generally in terms of a turbulent, outflowing wind close to the QSO nucleus and covering a significant fraction of the central continuum source and the BLR. These sources are universally weak x-ray emitters, mostly undetected down to limits well below those of other QSOs. The strong absorption seen in those few detected to date leads to an interpretation in terms of absorption of the x-ray emission by the same, outflowing wind and suggests that the column densities is at the high end of the quoted range.

Bibliography

Blandford R D, Netzer H and Woltjer R 1990 *Active Galactic Nuclei* (Berlin: Springer)

Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Mill Valley: University Science Books)

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

Belinda J Wilkes and Bradley M Peterson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galaxies: Observing their Variability by Amateurs

Amateurs began to take an interest in observing the brighter active galactic nuclei (AGN) relatively soon after their discovery in the late 1960s, although a small number were observed as if variable stars prior to this. Leading amateurs have thus been observing the brighter AGNs for more than three decades, monitoring changes in their brightness. In addition, many recreational observers have tracked down these faint specks visually, especially 3C273 in Virgo, and marveled at the view, although less at their appearances than their extraordinary nature.

The first formal coordinated program for amateurs to monitor AGNs began in the UK in the late 1970s with the British Astronomical Association's Variable Star Section adding 3C273, NGC 4151 and Markarian 421 to its program. These were monitored visually as if faint variable stars. As telescopic apertures increased over the 1980s, more AGNs were added to the program. The Astronomer organization took up an interest and added this class of objects to its program, which remains active—observations are still being published in *The Astronomer* magazine monthly.

The potential for serious photographic work was realized in the late 1980s, the technique enabling the addition of several fainter examples that were too faint for most visual observers. This was a limited success, the difficulties of photographic patrolling limiting the popularity. It was not until the advent of the CHARGE-COUPLED DEVICE (CCD) and its availability to the amateur that the next boost to AGN observation was given. Of professional interest in amateur observations, there is no doubt—support and encouragement having been given by several international astronomers, CCDS, FILTERS AND CCD PHOTOMETRY FOR AMATEURS.

The broad aims of amateur AGN observations are threefold:

- to monitor apparent changes in brightness of accessible AGNs (see figure 1);
- to measure such changes and build up as accurate a light curve in visible wavelengths as possible;
- to alert professional astronomers to unusual activity, and anticipated activity of interest, and provide routine observations on a regular basis.

In order to monitor as many AGNs as often as possible, a coordinated program has been set up. There are currently over 20 AGNs available to the well-equipped amateur (some are listed in table 1). Although most are relatively faint, several are accessible to modest telescopes visually and all should be fairly straightforward targets for CCD users.

Observers are encouraged to make observations as regularly as skies and conditions allow. Almost all AGNs on the program are plotted in Uranometria 2000, although some have alternative designations from their more popular terms. However, use of coordinates should allow identification without difficulty.

Once located, the AGN's brightness can be estimated or measured, visually, photographically or with a CCD. Each has its merits and problems. In favor of a visual approach is the ability to observe several objects in the session; the results are rapidly obtained and the equipment is relatively inexpensive, although realistically a larger aperture is needed for such faint objects. However, the telescope need not be equatorially mounted. Set against these advantages are the problems of finding relatively faint objects from less than ideal sites, with problematic sky conditions, due to light pollution, moonshine, high cirrus and so forth. Experience in variable star work is essential. Inevitably, some observer bias occurs, although this can be compensated for, as with other variable star observations. Even if the AGN itself

Table 1

Object	*Right ascension*	*Declination*	*Constellation*	*Type*	*Magnitude or approximate range*	*Uranometria 2000, chart number*
3C66A	02h 22m 39s	+43 n 05 16	AND	BL LAC	14–16.3	62
3C120 (BW TAU)	04h 33m	05 n 21	TAU	BL LAC	13.7–14.6	178
OJ287	08h 54m 48.9s	20 n 06 32	CNC	BL LAC	12.4–16	142
Markarian 421	11h 04m 27s	38 n 12 32	UMA	BL LAC	13.6–14	106
NGC 4151	12h 10m	39 n 24	CVN	Seyfert	11.1	74
Markarian 205	12h 21m 44s	75 n 18 37	DRA	BL LAC	14.5	9
W Comae	12h 21m 31.6s	28 n 13 58	COM	BL LAC	11.5–16	148
3C273	12h 29m 6.8s	02 n 03 m 07 s	VIR	Quasar	12.2	238
3C279	12h 56m 11s	05 s 47 21	VIR	Quasar	17.7	284
BL Lacertae	22h 02m	42 n 16	LAC	BL LAC	12.5–15	87

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

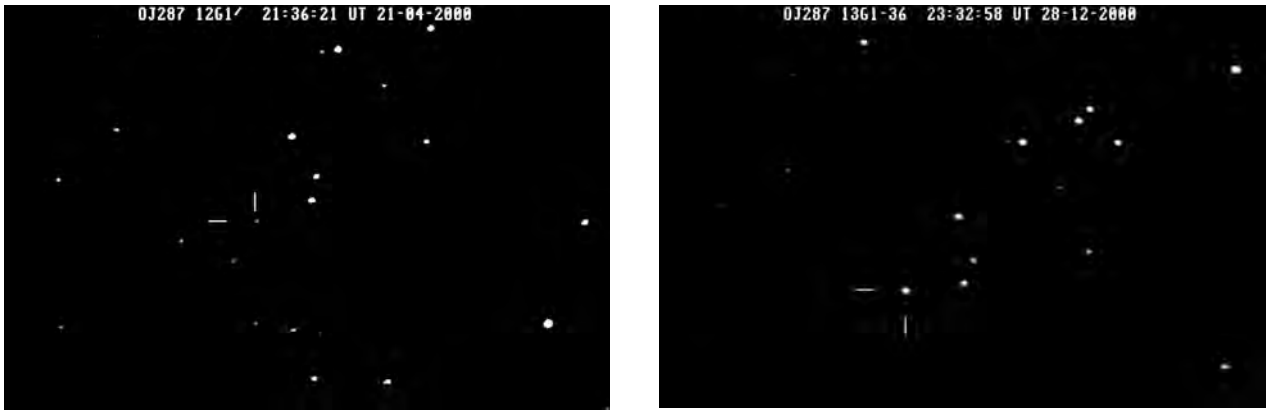


Figure 1. The dramatic rise in brightness of the very active BL Lacertae object OJ287 in Cancer is demonstrated in this pair of images. The image of 21 April 2000 shows the AGN to be at magnitude 16, that of 28 December 2000 at 14.5. Both images were obtained with a Meade LX 200, 250 mm *f*6.3 SCT and Starlight Xpress CCD camera; the April image was exposed for 80 s, that in December for 160 s. Neither was filtered.

poses no great problem, the fainter comparison stars may well be difficult. While the numbers of AGNs available are limited, there are still over a dozen to be observed by the owners of a 16 in telescope or larger.

Photography would seem to reduce some of these barriers. Fainter magnitudes can be reached so more active galaxies become available and the comparisons are more easily accessed. Personal bias is reduced and the results are easily displayed and reproduced. However, an equatorial or driven telescope becomes essential, raising the financial commitment. For the accurate results which the time and financial outlay deserve, standardization and formal measurement become necessary. This would require the observer to use a standard film (Kodak 103aB was recommended in the past, but it is difficult to obtain and other emulsions need to be tested). Also, filtered images are desirable, and for accurate magnitude estimates microdensitometric measures are ideal. Add to this the laborious process of photography itself and dark room work, and it is perhaps understandable that the photographic programs of the late 1980s never really came to fruition. The chance of discovering a nova or supernova is often adequate incentive to be involved in patrol work, but monitoring of an existing object year in, year out, is perhaps less enticing.

The CCD revolution seems be the answer to this. Images made with CCDs can go very deep, even with modest apertures. Many AGNs are available to owners of 10 in telescopes equipped with a CCD. Finding objects with smaller CCD chips was once notoriously difficult but modern GoTo facilities of many modern telescopes mean that the field can be confidently found and then the AGN identified from charts and imaged really quite rapidly and a measurement made.

Computer software is becoming more sophisticated and, with care, photometry accurate to around 0.1 magnitudes should be possible. The data are easily stored on disk and retrieved at leisure. Rapid comparison with a master image is possible to show any significant changes in brightness. This happy state of affairs is tempered by the need to spend a fair amount of money on equipment. In addition to a driven telescope, there is the CCD camera and computer plus necessary software. So to achieve the first meaningful CCD image of an AGN may cost the observer several thousand dollars.

Also, not all CCDs have the same sensitivities. Most have a red bias, which distorts the results, but different cameras have different profiles, so, for photometry to be really useful to the professionals, it is increasingly clear that filtration is needed (see CCDS, FILTERS AND CCD PHOTOMETRY FOR AMATEURS, FILTERS). It may ultimately be acceptable to use just B and V filters, even one of these. However, simple monitoring without photometry does not need filtered images and this can still be useful even if no filters are available to the observer. At present the majority of observations remain visual, with some CCD imaging, but, with more CCD users looking for serious projects, the emphasis may change.

AGNs offer a wonderful challenge to the serious amateur observer who wishes to contribute in their small way to an area of science in which the professionals welcome help, if provided carefully (see also ACTIVE GALAXIES: OBSERVATIONS). While not giving much in the way of spectacular views, AGNs feed the imagination in a way that is equally satisfying to the enquiring mind. The techniques are improving, the quality of results will become more refined and the study of the deepest of deep-sky objects is about to take off.

Nick Hewitt

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Active Galaxies: Overview

Defining an 'ACTIVE GALACTIC NUCLEUS' (AGN) can be difficult. The last two words are relatively simple, as an AGN is an energetic event which is *never* found anywhere in the universe except within the central parsecs of a galaxy. This is not surprising, because any dynamical friction should soon force any very massive compact object to settle down into the center of the galactic gravitational potential.

A GALACTIC NUCLEUS is said to be 'active' if it generates a 'substantial' (or detectable) amount of energy by processes different from those found anywhere else in galaxies. That is, the energy of the 'activity' is not produced by evolution of stars and interstellar matter. This definition is clean because it is theoretical. The actual observational task of identifying AGN can be messy, especially when the nonstellar activity is weak compared with the background of normal stellar processes, including birth and death. Classifying AGN meaningfully is harder. Understanding their physical nature is harder still.

The problems are further complicated because the nonstellar activity releases energy in many forms, and at photon wavelengths covering nearly the entire electromagnetic spectrum. Since observing instruments and methods differ markedly from one waveband to another, many overlapping wavelength-specific classification schemes and models have been proposed.

Not surprisingly, astronomers became aware of the more spectacular and luminous examples of AGN first. These nuclei were sufficiently luminous that they could outshine the entire galaxy in which they resided. They were also very distant, and it was often not possible, before very sensitive imaging had detected the surrounding galaxy, to distinguish their optical appearance from that of a star. This led to their designation as 'quasistellar' objects. Some of the first of these were also powerful radio sources, and the term was soon abbreviated to 'quasar', which now generally refers to all high-luminosity AGN (see QUASISTELLAR OBJECTS: OVERVIEW). Their unique nature became more obvious when studying their spectra, in almost any wavelength region.

These unique properties have now been found in active nuclei spanning an extraordinary range in luminosity[1]. There are only surprisingly subtle spectroscopic differences between a quasar with a luminosity of 10^{48} erg s^{-1} and what had previously been identified as a 'Seyfert galaxy nucleus' of luminosity 10^{42} erg s^{-1}. The evidence is overwhelming that these two astronomical categories must have a single common physical explanation. Thus today, hardly any distinction is made between 'quasar' and 'Seyfert 1 nucleus', as their main difference is (somewhat arbitrarily) in their luminosity. It follows that this overview of AGN should be supplemented by reading the accompanying chapters on quasars in this Encyclopedia.

[1] The vast range of AGN luminosities, much larger than what is seen in stars for example, is one of their outstanding features.

Understanding The Central Engine in AGN

There are several key ideas which are being used to organize and make sense of our large and growing observational information about AGN.

Unified Models

Since the earliest AGN studies, it was recognized that they can show strong observational differences in spite of many other similarities. Two characteristics seem to have a sufficiently bimodal distribution to justify using them to divide AGN into two pairs of classes: strong radio emission (which distinguishes radio-loud (RL) from radio-weak (RW) AGN) and broad permitted line emission (which distinguishes type 1 Seyfert nuclei from type 2). A more extensive discussion of these ideas is given in ACTIVE GALAXIES: UNIFIED MODEL and ACTIVE GALAXIES: VARIABILITY.

The majority of quasars are only moderately stronger radio sources than normal galaxies. (They are sometimes referred to as radio-quiet (RQ) AGN, although strictly speaking they do emit some radio flux so that RW is a more accurate description.) An important minority of quasars are well known for the beams of high-energy particles they accelerate. The radio emission from these jets is nearly always bipolar—a key indicator that the central engine can have axial symmetry. In many models, this jet–lobe axis is the spin axis of the central massive black hole. In the most dramatic subsample of RL AGN, the 'blazars' (loosely named after the prototype object BL Lacertae), the highly polarized, violently variable, steep red continuum was identified as nonthermal synchrotron emission. Blazar is an informal term used to encompass the prototypical nonthermal BL Lacertae objects (with very weak emission lines) as well as the optically violently variable quasars. Interferometric maps showed that the core-dominated radio emission generally had a one-sided jet, which often appears to be expanding at apparent transverse speeds several times larger than c. The most extreme of these objects appear to emit a substantial fraction of their energy in the gamma rays, and possibly even up to TeV energies. All of these remarkable (and manifestly nonstellar) blazar properties were soon understood as the result of viewing a relativistic jet (of electrons and either protons or positrons), which is emerging from the active nucleus close to our line of sight. This nearly 'pole-on' viewing angle produces several relativistic effects such as Doppler boosting of the 'beamed' synchrotron emission and a compression of apparent timescales which accounts for 'superluminal motion' without violating special relativity. The one-sided nature of the innermost jets is usually attributed to 'Doppler favoritism'. If there is currently an outflow in the opposite direction (which is predicted by some theoretical models, but not yet confirmed), it is invisible because Doppler de-boosting makes it thousands of times fainter.

An immediate consequence of the beaming explanation of core-dominated RL AGN is that more AGN with radio jets must be pointing away from Earth. These must be detectable by their isotropic emission (lines or extended radio lobes), as either lobe-dominated RL AGN, or perhaps

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

'radiogalaxies' (whose optical properties can be Sefyert 2s or even low-ionization emission line regions (LINERs)). The (unbeamed) emission lines may or may not be detectable but are harder to detect over the glare of the amplified nonthermal jet continuum.

These studies proved that the central engine in AGN (the RL ones, at least) have a strong bipolar symmetry which has been preserved over millions of years or more in the largest double-lobe radio sources. They also proved that the observational appearance of RL AGN can depend dramatically on our viewing angle. However, it is not yet certain how completely these same insights may be transferred to the majority of AGN, which are RW.

Broad permitted emission lines are usually taken as one of the defining characteristics of AGN. They arise from high-density photoionized gas. Their high ionization state and large velocity range both indicate that they arise in a 'broad line region' (BLR) relatively near the central engine. However, for every broad-emission-line AGN of a given luminosity (classified as a 'Seyfert 1'), there is at least one galactic nucleus which might otherwise be classified as a Seyfert, but which lacks broad line emissions and is classified as a 'Seyfert 2'. In the Seyfert 2s it is difficult or impossible to view the central engine, and its surrounding BLR, directly. In a minority of cases, strong evidence for the central engine is present, in the form of either a highly absorbed ($N_H \sim 10^{23}$ cm^{-2}) hard-x-ray continuum source or broad Balmer emission line wings seen in the spectrum of linearly polarized light.

Less direct evidence is the detection of fluorescent emission from the x-ray Kα line at 6.4 keV. In some Seyfert 2s, the very large equivalent width of this line (about an order of magnitude larger than in unobscured Seyfert 1s) is interpreted as being powered by a nuclear hard-x-ray source, virtually none of which is visible along our direct line of sight to the nucleus. A similar argument has been made from UV spectroscopy: adding up all the ionizing photons inferred from the UV continuum seems to be inadequate to account for all the Seyfert emission line flux detected, and especially for the infrared continuum also associated with the active nucleus. Assuming that the 2–10 μm continuum in Seyfert 1s and 2s is AGN energy that has been re-radiated by dust grains (see 'Reprocessing' section below), even more unseen nonstellar power is implied in Seyfert 2s. This 'energy deficit' argues for a high-energy power source which is present but not directly observable. However, these arguments are not conclusive. The energy accounting has many large uncertainties, and there are other plausible explanations of the Kα emission line that do not require the existence of a buried AGN x-ray source. Although it is difficult to obtain the very sensitive observations required for these tests, many thorough observations have failed to reveal any direct evidence of nonstellar continuum from a central engine. The more extended indications which are seen in these Seyfert 2s—narrow emission lines and radio sources—could have been produced by an AGN which 'turned off' centuries ago.

This raises an empirical question of whether most Seyfert nuclei can turn on and off over intervals of centuries. Although none has been observed for that long[2], very few complete Seyfert 2 $\leftrightarrow$ Seyfert 1 transitions have ever been seen, in decades of observing. Nonetheless, the unification assumption is that all of these Seyfert 2s harbor a currently active central engine, which is just extremely obscured.

Even if unification is generally correct, it simply requires optically thick gas, with a very large associated dust extinction, between us and the nuclei of Seyfert 2 galaxies. All we know is that this obscuring region must be large enough to cover up the broad emission line region, as viewed by a significant portion of the lines of sight to the central engine[3]. Similarly, dust grains must intercept a substantial portion of the central continuum and re-radiate it to produce the strong near- to mid-IR emission characteristic of both Seyfert 1 and 2 nuclei.

A popular special version of unification proposes that this absorbing matter is strongly nonspherical, and is distributed in a fat torus (of characteristic diameter parsecs to tens of parsecs) which is co-axial with the central engine. In the fat torus model the absence of obscuring material above the poles of the torus allows photons from the nuclear central engine to scatter into a clear line of sight to the Earth. The result would be a scattered nuclear spectrum—broad emission lines and nonstellar continuum—which we observe only in polarized light. Since the torus is supposed to be aligned with the axis of the central engine, the polarization E vector should be parallel to the radio axis. The latter is difficult to measure but is approximately parallel to the polarization in some of the best-measured Seyfert 2s. Further weak evidence for a preferred 'escape axis' for ionizing AGN photons is the loose tendency for extended ionized emission lines in Seyfert 2s to come from bipolar 'cones'.

In many Seyfert 2s, however, there is evidence (e.g. from HST imaging) that much obscuration instead occurs in galactic dust lanes. These are observed to be hundreds of parsecs in extent and not related to the central engine at all, except that they happen to block our view of it. It is still not clear how many, if any, Seyfert 1s and 2s really harbor complete fat tori wrapped around their central engines.

Thermal emission from an accretion flow

The enormous energy outputs from very small volumes that are observed in luminous quasars are believed to be generated by black hole accretion. To avoid greatly exceeding the Eddington limit, the accreting black holes must have masses of several 10^6–10^9 Suns. However, the conversion of the gravitational potential energy of the inflowing matter into radiation is not at all efficient if there is a purely radial infall. Also, general considerations of

[2] However, NGC 1068 = Messier 71 might, in a bright Seyfert 1 state, have been marginally visible to the naked eye.

[3] Since the space density of Seyfert 2s is somewhat higher than that of Seyfert 1s, this covering fraction of obscuration must be larger than 50%

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the angular momentum of gas deep in a galactic nucleus strongly suggest that the accretion flow must end up rotating rapidly as it nears the black hole. Then it can only accrete and power the AGN if it experiences friction. Although not yet well understood, this crucial process transfers angular momentum outward, allowing the black hole to be fueled. This accretion is accompanied by a large release of radiative energy, which powers the AGN. The fact that angular momentum is important in the accretion flow is supported by the observation that the central engine is bipolar. Its symmetry axis is naturally identified with the spin axis of the accretion flow and is the direction of the most likely outflows. There is also observational evidence that we observe directly the radiation emitted thermally from this accretion flow, as it spirals into the black hole.

One of the defining characteristics of the spectrum of the central AGN continuum source is a broad hump which appears to peak somewhere in the UV. This inflection, referred to as the 'ultraviolet excess' or 'big blue bump', often contains the largest portion of the total AGN luminosity.

The main arguments in favor of a thermal origin of this blue–UV continuum are as follows.

- *Its spectral shape*. The broad peak in the UV can be fitted by a sum of thermally emitting regions spanning a range of temperatures, centered on ∼30 000 K. In a simple optically thick accretion disk, such as would be expected for most luminous AGN, the thermal temperature of the accretion flow is proportional to $r^{-3/4}$. With the inclusion of relativistic corrections, the sum of thermal emission from rings of varying radii adds up to a spectrum with a broad peak which is flat in flux density units (F_ν = constant) and has a high-frequency turn-over. Models which treat the disk as a sum of blackbodies provide remarkably good fits to all the optical–UV spectral energy distributions observed in AGN.
- *Its peak frequency*. The effective temperature in an accretion flow is approximately determined by the balance between the gravitational potential energy released as a result of viscous accretion and the radiation emitted[4]. In a thin disk this condition is $T_{\rm accretion} = (3GM_{\rm bh}\dot{M}/8\pi r^3)^{1/4} = (30\,000\,{\rm K})\,(M_8)^{1/4}(L/L_{\rm Eddington})^{1/2}$. The luminosities and peak frequencies observed in the continuum spectra of Seyfert 1 nuclei and quasars are consistent with the equilibrium temperatures predicted for disk accretion. The required respective mass infall rates are ∼ several per cent and ∼ tens of per cent of the Eddington limit for the inferred black hole masses[5].

Even for the highest plausible values of viscosity (corresponding to the most rapid inflow), the accretion flow is still expected to be very opaque to electron scattering and is probably effectively optically thick ($\tau_{\rm effective} = (\tau_{\rm es}\tau_{\rm ff})^{1/2} > 1$). The best observational evidence for this is the high inferred emissivity of the gas emitting the UV continuum. It is true that optically thin thermal models can also fit the observed optical–UV continuum. The characteristic temperatures of the bremstrahlung emission would have to be 100 000 K or higher, because free–free emission peaks at around $\nu_{\rm thin} = kT/h$, rather than the higher value for a Planck function: $\nu_{\rm thick} = 3kT/h$. EUV emission lines that might be expected from such a luminous plasma have not yet been seen. The most serious difficulty with the free–free hypothesis is explaining how the very large continuum luminosity of a quasar is emitted from a small area. For the UV brightness of a luminous quasar such as 3C 273 to vary on timescales of months requires a small thermally emitting region with high emissivity, so high, in fact, that the emitter must have an optical depth exceeding 1. At this point the emitted spectrum will resemble optically thick emission more than an optically thin $L_\nu \propto \exp(-h\nu/kT)$ spectrum.

The principal limitations or incompletenesses of the thin accretion disk model are the weakness of any observed spectral feature at the Lyman limit, the low amount of linear polarization, and multiwavelength monitoring evidence for reprocessing, discussed in the next section. The spectrum emerging from an accretion disk at a given radius depends on the vertical structure of the disk 'atmosphere'. A ring of an accretion disk differs radically from the atmosphere of any star, but it has recently become computationally feasible to make realistic non-LTE models of the emergent spectrum. Many UV-emitting rings of the disk have low enough temperatures that they should still contain some fraction of neutral hydrogen. This should lead to a sharp increase in the bound–free opacity across the Lyman edge. Some models predict that this will lead to a spectral jump, with the flux above the Lyman limit either higher or lower than the flux below it. These discontinuities are not in general observed. Opacity in the accretion disk is probably dominated by electron scattering. A pure electron scattering atmosphere of infinite optical depth viewed at an average inclination of $\cos i = 0.5$ should emit a thermal continuum which is linearly polarized by almost 3%. The position angle of the electric vector should be parallel to the projected major axis of the disk—perpendicular to the disk spin axis. However, observations show that most AGN polarizations are under 0.5% in the UV[6]. Also, in those cases where a radio 'axis' can be defined, it is not generally perpendicular to the E vector of the polarized flux. The greatest uncertainties in this test are the assumptions that the disk surface is perfectly flat and that magnetic fields in the disk are small enough to avoid causing appreciable Faraday rotation.

One deficiency of the simple thin accretion disk models that may be related to these observational conflicts

[4] In fact the viscous torques transport some of this potential energy outward, but most of it should eventually emerge as photons.

[5] Above the Eddington accretion rate, the disk is expected to thicken somewhat, but its overall observational properties may not change very greatly.

[6] However, the linear polarization appears, mysteriously, to increase dramatically, at rest wavelengths shorter than ∼800 Å.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is the likely presence of a hot optically thin corona above the photosphere of the disk surface. The fast electrons in this low-density gas can Comptonize the thermal photons escaping from the dense disk, modifying their spectrum and polarization.

Reprocessing

In the simplest accretion disk models, all of the thermal energy emitted by the disk was derived directly from gravitational potential energy, which is transported outwards by viscous torques. We know, however, that AGN also emit up to 10% or 20% of their luminosity at x-ray frequencies. These photon energies are too high to be readily explained by the thermal accretion disk paradigm, and their flux near the center is very high. Depending on the geometry, some of these x-rays could be re-absorbed by the optically thick accreting gas before they manage to escape. Some absorption of x-rays is probably required to explain the 'Compton reflection' hump observed around 10–30 keV (see discussion in ACTIVE GALAXIES: OBSERVATIONS). Reprocessing of x-rays is presumed to be the origin of the strong Fe Kα emission line seen in many AGN.

Reprocessing of photons at radii much larger than where they originated tends to degrade the energies of the photons that escape. In the simplest case, where the equilibrium blackbody temperature of the reprocessing matter drops as $T_{\rm equil} \propto L^{1/4} r^{-1/2}$ (where L is the luminosity of the central point source), the resulting thermal continuum spectrum tends to steepen to a slope of $L_\nu \propto \nu^{-1}$. Some reprocessing must occur; the main questions are: where and how much?

The first question is easy to answer assuming the absorbed photons are thermalized: the radius of the reprocessing determines $T_{\rm equil}$ of the emerging radiation, which should peak somewhere around a few $h\nu/kT_{\rm equil}$. The second question can be answered if we can measure the fraction of the total amount of energy which has been absorbed. If we assume that the reprocessing matter (gas, and dust if $T_{\rm equil}$ is under 1000 K) is opaque to incoming high-energy photons, then the energy budget can tell us its total sky covering fraction, i.e. the fraction of the sky as seen by the central continuum 'point source' which is covered by absorbers. By definition, the reprocessed luminosity cannot exceed the luminosity of the original high-energy continuum source in the center. It is actually unlikely that their luminosities would even be equal, because that would imply that 100% of the lines of sight to the center are blocked, in which case we should not be able to see it directly at all.

The best observational proof of reprocessing is to show that time variability in the primary waveband is then duplicated by the same flux changes at the reprocessed wavelength, except with a lag. The delay time in this 'reverberation' should equal the amount of time it takes light to travel across a region whose radius is given by the value of $T_{\rm equil}$ expected for the wavelength of the reprocessed continuum.

The best evidence for reprocessing is for the near-infrared continuum, which caries a substantial fraction (20% or more) of the total energy budget of most AGN. In several cases, simultaneous monitoring of the near-infrared and optical or UV continua has shown the expected correlated brightening and fading. The delay of the infrared light curve with respect to the light curve at shorter wavelengths is consistent with thermal re-emission by hot dust grains within a light year (or even light month) of the central powerhouse. The response is mostly from the hottest grains, surviving at just below their sublimation temperature of around 1000–1500 K. The distance from the central engine at which dust equilibrates to this temperature is proportional to the square root of the central luminosity, explaining the range of expected dust reverberation radii.

Less decisive evidence suggests that some of the optical continuum, and possibly even the UV continuum, in some Seyfert 1 nuclei, may have a measurable contribution from reprocessed x-rays. This is in several well-sampled x-ray and optical–UV light curves where a correlation and lag were detected. The lags are so small that they imply that the influence of the high-energy photons propagates outward to the low-energy emitting region at very high speed. The lags are consistent with this speed being c, as it should be if the influence is carried by photons which are then reprocessed. However, the variability of the x-rays (in other galaxies, as well as in some of these Seyfert 1s at other epochs) is often not correlated with changes at other wavelengths, casting doubt on whether much of the optical–UV light in most AGN is actually reprocessed from the x-rays.

Influence of host galaxies

Galaxies with Seyfert nuclei account for only a fraction of all galaxies. This estimation is complicated by the difficulty of identifying AGN with very low nonstellar luminosities. For example, many galaxies have weak emission line spectra with line ratios intermediate between those of H II regions and Seyfert nuclei, which are referred to as 'LINER' galaxies. Although these LINER spectra can be explained by fast shocks, they might still be included as having active nuclei (and the case is getting stronger for at least some of the members of this heterogeneous class). If so, then it is even possible that most galaxies are in some sense 'active'. The most definitive evidence of a truly nonstellar nucleus is probably a compact central source of hard x-rays, preferably variable. However, this would be obscured by gas with a column density of 10^{25} cm^{-2}, which is optically thick to Thomson electron scattering of hard x-rays. It may not be entirely surprising that, outside of the nucleus, the 'host galaxies' themselves hardly appear different from galaxies lacking obvious AGN. A very powerful AGN can have some limited effects on the interstellar medium of its galaxy, such as ionizing it, shocking and compressing it, and in some special cases even inducing stars to form, near an outflowing jet for example.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A deeper mystery is what properties of the galaxy cause the AGN in the first place. The two most obvious ones are the origin of a massive black hole and the ability to provide enough fuel close into the nucleus to 'feed' it. The formation of a massive black hole probably requires the dynamical collapse of a dense star cluster in the dense center of a galaxy. As the first dense structures to come together in the young universe are thought to be proto-galactic bulges, there is some weak theoretical motivation for associating massive black holes with modern-day bulges. The dramatically higher (by orders of magnitude) proper space density of luminous AGN in the early universe requires that many seemingly 'normal' galaxies today must once have harbored an AGN central engine. If it was a black hole, it should still be present, and may be detectable by its gravitational acceleration of stars and gas within a few hundred parsecs. (The gravity of a black hole of mass M_{bh} dominates out to a radius of roughly GM_{bh}/σ^2 where σ is the characteristic velocity dispersion in the stellar system.) This region is small enough that it can only be resolved optically in the nearest galaxies with the best spectrographs (Hubble Space Telescope, or ground-based telescopes with excellent seeing). The kinematic search for nearby massive black holes (using Doppler spectroscopy to measure orbital velocities around the nucleus) is still in its early phases, with less than a few dozen detections (see also SUPERMASSIVE BLACK HOLES IN AGN). These provide a tentative suggestion that the central black hole mass might correlate linearly with the total galactic bulge mass, with a normalization constant of about half of a per cent. However, this claimed correlation may have enormous scatter or may not even prove significant when an unbiased sample of galaxies is studied intensively and upper limits to M_{bh} are all included.

In the near future, the set of galaxies with reliable central black hole masses has unfortunately very little intersection with the set of active nuclei for which mass estimates can be attempted. This is because the latter is measured from nonstellar emission, while the former comes from nonactive emission. If one of these dominates and is easily studied, the other will not be well measured. Nonetheless, the current evidence on black holes, mostly in nonactive galaxies, suggests that the key potential ingredient for the central engine—the massive black hole—may be present in many or even most galaxies. The fact that most of them today are not Seyferts must then be attributable to a lack of fuel.

Even assuming the most efficient known models, black hole accretion disks, the energetics of the most powerful quasars require tens of solar masses per year to be supplied to the central parsec of the galaxy. To reach such a small radius, the fuel—probably gas—must lose essentially all of its orbital angular momentum. On theoretical grounds, the required torques may be provided by distortions and asymmetries in the gravitational potential of the host galaxy, such as bars (see GALAXIES: CLASSIFICATION; FORMATION)[7].

Unfortunately, it has been difficult to show much association between the presence of close companions and an active nucleus. Also there is no correlation between bars and Seyfert activity. One possibility is that bars and/or encounters only enhance nuclear activity with a great delay (of perhaps a billion years), since it could take that many orbits for the gas to eventually fall down into the nucleus. One possible indication of this is the finding that Seyfert galaxies are more likely to show rings, which may be long-lived relics of bars that have already dissipated. A possible connection between Seyfert activity and earlier star-formation activity is the observation that the disks of Seyfert host galaxies tend to have higher surface brightnesses than those of non-Seyfert galaxies.

Future prospects

Understanding AGN is a seriously data-limited problem. Its typical variability proves that the central engine is not in a steady state. We can turn this into an asset in trying to understand it, but only if we can obtain extensive, nearly continuous time histories of the emission from individual AGN. Beyond limited time sampling, the two most serious observational limitations have been the following.

- *Inadequate observations of many of the wavelength regions where AGN emit*. AGN are the most broad-band emitters in the universe, and many of the wavelengths at which they emit are blocked by the Earth's atmosphere, or more fundamentally by H I absorption in the intergalactic medium (shortward of a rest wavelength of 912 Å).
- *Inadequate spatial resolution*. The largest relevant scales for most of the interesting and extreme AGN phenomena are of order 10 pc or even less.

Fortunately both of these limitations are rapidly easing with the advent of new detectors and observatories. The critical AGN observations during the next decade will benefit from new and planned space observatories and new ground-based techniques such as interferometry. Also, the ultimate in spatial resolution is becoming available by using intensive monitoring to detect light-travel delays, as in reverberation mapping.

Matthew Malkan

[7] Two controversial observational claims may support applying this idea to star formation. One is that spiral galaxies with bars are more likely to have higher rates of recent star formation than unbarred spirals. Another is that close galaxy encounters, which also apply an $n = 2$ mode perturbation to the galactic potential, are also associated with elevated rates of star formation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galaxies: Unified Model

Objects known as 'ACTIVE GALACTIC NUCLEI' (AGN) can be grouped into a wide variety of phenomenological subclasses that have many overlapping or similar properties. The first known 'active galaxies' were the high-surface-brightness spiral galaxies identified by Carl Seyfert in 1943. Almost two decades later, the first radio surveys of the sky led to the identification of other objects (radio galaxies, such as Centaurus A, and quasars, such as 3C 48) that we now also think of as active galaxies, though the similarities between the radio sources and Seyfert's spirals were not obvious at the time of discovery. Early failure to recognize the underlying similarities of these objects was largely because the different identification techniques used isolated extreme members of the active-galaxy population: radio-source identifications tended to yield relatively high-luminosity, high-redshift quasars (short for 'quasistellar radio sources'), often members of the 'blazar' class (see BL LACERTAE OBJECTS), in which beamed emission from a relativistic jet dominates most of the observed spectrum. On the other hand, identification of bright galactic cores led to identification of the spatially common, relatively low-luminosity radio-quiet Seyfert nuclei (see ACTIVE GALAXIES: OBSERVATIONS).

It was only in the late 1960s and 1970s that optical, color-based detection criteria (either multiple-color broad-band photometry or low-dispersion spectroscopy) began to yield large enough samples of both high- and low-luminosity AGN that the fundamental similarities began to become clear. Indeed, color-based selection of AGN led to the discovery that most high-luminosity AGN are not of the 'radio-loud' quasar variety, but are 'radio-quiet' sources that came to be known as 'QUASISTELLAR OBJECTS' (QSOs). Despite the emerging similarities between Seyfert galaxies and QSOs, important differences among types of AGN were also identified, leading to a broad and fairly complex AGN taxonomy (see ACTIVE GALAXIES: OVERVIEW).

Comparison of various types of AGN should help us understand which phenomena are fundamental to nuclear activity and which are somehow incidental or secondary. In some cases, as we will see below, differences among various types can be used to infer source structure on scales too small to resolve directly. Some of the important questions that arise in this context include the following. Why are only ∼10% of active galaxies radio-loud? Why are broad emission lines weak or absent in the UV–optical spectra of BL Lac objects? Why do some Seyfert galaxies have no broad emission lines?

Unification

Principles

In attempting to explain the differences among various classes of AGN, our goal is to explain the widest variety of phenomena using the simplest model that is consistent with the data. A simple hypothesis to explain much of AGN phenomenology is that the differences among various types of AGN arise from orientation dependence; simply put, we try to 'unify' various AGN types in terms of a single basic source structure whose appearance to the observer depends strongly on viewing angle. There is indeed abundant direct evidence that AGN have axisymmetric structure and thus radiate anisotropically; the observable properties of a particular source will thus depend on the location of the observer. It is therefore postulated in such 'AGN unification' schemes that the different appearance of one class compared with a second is the result of viewing the same type of object at a different angle. Properties that may depend on viewing angle and may so contribute to intrinsic anisotropy in these sources include absorption by dust or optically thick gas in any non-spherically symmetric distribution and relativistic motion which leads to Doppler boosting of the emission which peaks in the direction of this motion. Unification schemes incorporating one or both of these intrinsic source properties seek to explain at least part of the wide diversity we observe in AGN in terms of these differences in viewing angle. There are several unification schemes seeking to unify two or more classes of AGN that have met with some degree of success; several of these are outlined in table 1 and will be described below.

Table 1. Possible simple unifications.

Radio properties	Orientation	
	Face on	Edge on
Radio quiet	Seyfert 1	Seyfert 2
	QSO	ULIRG?
Radio loud	BLRG	NLRG
	Core dominated	Lobe dominated
	BL Lac	FR I
	Quasar–OVV	FR II

An obvious key element in any unification scheme is proof of its plausibility. If we hypothesize, for example, that some specific type of AGN (say, BL Lacertae objects) are members of some particular broad class or 'parent population' (say, Fanaroff–Riley class I radio sources[1]), then the space density of the subclass must be consistent with the space density of the parent population times the probability of observing a member of the parent population at a suitable orientation. Determining the space density of the parent population is often a difficult task in itself and must rely on source identifications based on some property which is supposed to be isotropic. Hard x-ray emission, for example, is thought to be emitted approximately isotropically by non-blazar AGN, and in principle hard x-ray emission may be the best way to

[1] Fanaroff and Riley showed that extended radio sources can be divided into two luminosity classes that have different morphologies. Class I (FR I) are less luminous sources, $L_\nu(1.4\,\mathrm{GHz}) \lesssim 10^{32}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}$, that are brightest in the center with decreasing surface brightness towards the edges. Class II (FR II) sources are the more luminous sources, which are limb-brightened or have non-central regions of enhanced emission.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

find a homogeneous sample of AGN. Indeed hard x-rays are probably the *sine qua non* of an active nucleus. In practice, however, relatively low hard x-ray sensitivities have limited detection of AGN to those that are relatively nearby and thus apparently bright. For radio sources (which includes blazars), radiation from the optically thin extended lobes is assumed to be emitted isotropically and in principle can be used to identify objects with intrinsically similar luminosities.

The basic model

The current AGN paradigm is built around an axisymmetric central engine that consists of an accretion disk surrounding a supermassive ($\gtrsim 10^6 M_\odot$) black hole (see SUPERMASSIVE BLACK HOLES IN AGN). The UV–optical continuum emission is supposed to arise primarily in the accretion disk. Bi-directional relativistic jets emerge from this system along the disk axis, emitting Doppler-boosted radiation via synchrotron and inverse Compton mechanisms. Both the jets and the accretion disk structure are thought to contribute to the x-ray emission. The broad emission lines (whose width can be characterized in terms of the full width at half-maximum, FWHM $\approx$ 1500–10 000 km s^{-1}) that are so prominent in the UV–optical spectra of AGN are produced in relatively dense (electron densities $n_e \approx 10^{11}$ cm^{-3}) gas clouds at distances from the black hole of several hundred to several thousand gravitational radii ($R_{grav} = GM/c^2$, where M is the mass of the central black hole, G is the gravitational constant, and c is the speed of light). On parsec scales, this entire system is embedded in a dusty torus that is opaque over most of the electromagnetic spectrum; this torus plays a key role in AGN unification models since it shields both the accretion disk, broad-line region, and inner jet structure from the direct view of external observers in the torus plane. The torus absorbs radiation from the central source and re-emits this energy in the infrared. Narrow emission lines (FWHM $\lesssim$ 300–800 km s^{-1}) arise in low-density clouds at the torus scale and beyond, mostly driven by radiation from the central source, and therefore largely constrained to lie along the system axis.

AGN unification schemes

Seyfert types

Probably the best known unification scheme is that between the two types of SEYFERT GALAXIES. In the mid-1970s, Khachikian and Weedman found that Seyfert galaxies fell into two spectroscopic classes, those with both narrow and broad emission lines (type 1 or Sy1) and those with narrow lines only (type 2 or Sy2). The narrow-line spectra of types 1 and 2 are statistically indistinguishable from one another, so Sy2s seem to be Sy1s without the broad lines. Moreover, Sy2s are typically less luminous than Sy1s by about 1 magnitude in the optical part of the spectrum. This led to unification hypotheses in which Sy2s are intrinsically Sy1s whose continuum and broad-line emission is attenuated in the direction of the observer. While simple and attractive, this hypothesis had two major difficulties: first, why is the AGN continuum source extinguished by only a factor of a few, but the broad lines by an apparently larger factor? Second, the shape of the Sy2 continuum does not look like a dust-attenuated Sy1 AGN continuum, i.e. Sy2 continua do not appear to be highly reddened by passage through a medium in which the absorption or scattering cross-section decreases with increasing wavelength.

A breakthrough occurred in the early 1980s when Antonucci and Miller found that the polarization spectra of some Sy2 galaxies such as NGC 1068 (see the articles on NGC 1068 and POLARIZATION IN ACTIVE GALAXIES) contained broad emission lines like those seen in Sy1 spectra. At least some Sy2 galaxies contain broad emission lines, but with their strength greatly reduced such that they are dominated by the continuum and narrow lines except when viewed in polarized light. Since the most common cause of polarization is scattering of light by either dust or electrons, this observation led to an interpretation of Sy2 galaxies as edge-on Sy1 galaxies where optically thick material in a flattened, disk-like geometry obscures our direct view of the broad emission-line region, as described above and illustrated in figure 1. The broad lines are visible in polarized light when they are scattered into our line of sight by dust or electrons above and/or below this material. In the case of NGC 1068, the AGN continuum light is $\sim$16% polarized, independent of wavelength through the UV–optical spectrum, which argues that the scattering mechanism is electrons rather than dust, since the Thomson scattering cross-section is wavelength independent while dust scattering is strongly wavelength dependent. In other polarized Sy2s, however, there is evidence that dust scattering also plays a role. The polarization spectra of these Sy2s also show that the emission lines are more strongly polarized than the continuum. There is apparently a second unpolarized component to the featureless AGN continuum that constitutes some 60–90% of the total continuum and has approximately the same spectral shape as the polarized component. The origin of the unpolarized component is unclear, although it may arise from free–free emission in the scattering region itself. In any case, the existence of this component explains why the non-stellar continuum in Sy2s is not as thoroughly suppressed as the broad lines.

Predictions of the relative numbers of each class provide an important test but are very sensitive to assumptions such as the opening angle of the torus and the angular dependence of the luminosity. Within these limitations the relative numbers are consistent with expectations and indicate torus semi-opening angles of $\sim 60^\circ$. As more Sy2 galaxies and lower-luminosity AGN, such as LINERs (low-ionization nuclei emission region galaxies), are observed in polarized light, more edge-on Sy1 galaxies are discovered. It is not yet clear, however, whether all AGN, or even all Seyfert galaxies, fit into this scheme. Current studies show a detection rate for polarized broad lines in Sy2 galaxies of $\sim$50%, implying that many Sy2 galaxies either do not possess a scattering region or that they have no broad-line region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

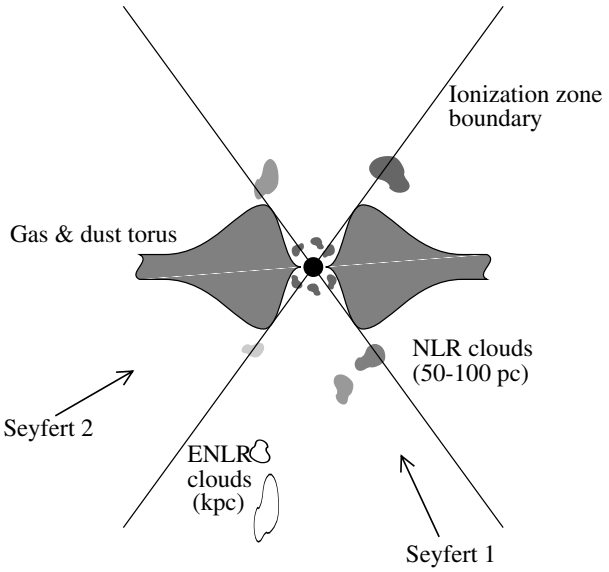


Figure 1. The conceptual scheme for unification of Sy1 and Sy2 galaxies (not to scale) with an optically thick torus of cool material surrounding the continuum source (central circle) and the broad emission-line region (BLR, small 'clouds' surrounding continuum source). From Peterson (1997), originally from R W Pogge.

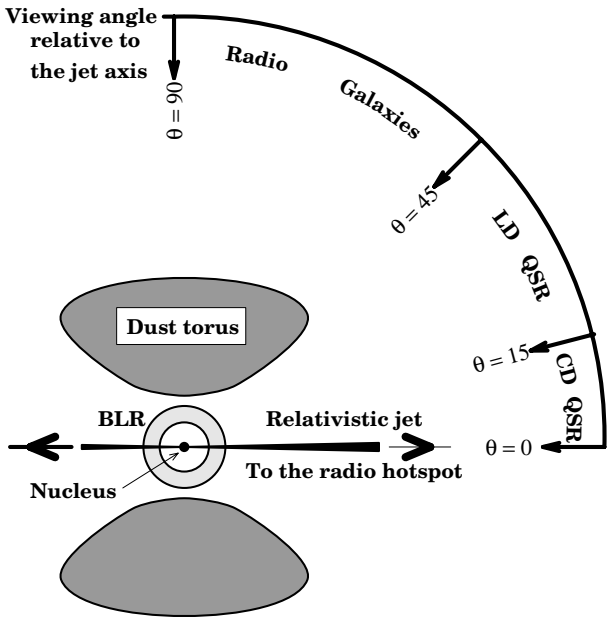


Figure 2. A scheme for unification of strong-line (BLRG, NLRG and quasars (QSR)) radio sources. Compare with figure 1 (rotated through 90°). Here the BLR is shown as a shaded shell around the continuum source (circular spot). From Kembhavi and Narlikar (1999), originally from R Athreya.

X-ray observations of AGN also generally support this unification picture. The x-ray luminosities of Sy2 galaxies are lower than those of Sy1 galaxies. Moreover, Sy2 galaxies have significantly larger amounts of soft x-ray absorption, corresponding to equivalent columns of neutral hydrogen, $N_{\rm H} \approx 10^{22-23}$ cm^{-2}, in comparison with $N_{\rm H} \approx 10^{20-21}$ cm^{-2} for Sy1 galaxies. This is consistent with our line of sight through the AGN being through an optically thick torus of cold material.

A similar unification scheme probably explains the difference between 'broad-line radio galaxies' (BLRGs), which might be described to first-order accuracy as radio-loud Sy1, and 'narrow-line radio galaxies' (NLRGs), which have Sy2-like optical spectra. A schematic of the geometry to unify the strong-lined radio galaxies is shown in figure 2.

A remaining puzzle is the apparent lack of an equivalent bifurcation in the higher-luminosity QSOs; there is no generally accepted class of 'QSO 2s' which would correspond to the low-luminosity Sy2 galaxies. The most likely candidates for dust-obscured QSOs are near IR-selected AGN, some of which do show broad lines in polarized light, or the cooler, ultraluminous IR galaxies (ULIRGs) whose emission may be predominantly due to reprocessing by cool dust. At the present time, it is clear that some ULIRGs harbor unseen AGN, but others are powered by starbursts, not by *bone fide* AGN.

Core- and lobe-dominated quasars and radio galaxies

As noted earlier, only about 10% of AGN are strong radio sources. Within this subset, there is a wide range of radio morphological structures (see also RADIO EMISSION FROM BINARY STARS) from classical double-radio sources on scales as large as megaparsecs to small, core-dominated sources with little or no extended structure. The core-dominated sources commonly show superluminal motion indicating relativistic outflow velocities close to our line of sight (see SUPERLUMINAL MOTION). Classical double sources often have one-sided, highly collimated jet structures extending from their compact core towards one of the extended lobes. The one-sided nature of the jets can be understood in a relativistic scenario whereby an approaching jet and an oppositely directed receding jet are respectively Doppler boosted and dimmed by relativistic effects. Thus both of these properties are consistent with relativistic jets originating in the core of a radio source and feeding the outer lobes. This suggests that the main variable governing the observed morphology of a radio AGN is its orientation to our line of sight. In core-dominated sources we are looking down the relativistic jet, while in double sources the jet is projected onto the plane of the sky.

A prime test of this unification scheme is to compare the relative numbers of core- and lobe-dominated sources and the distribution of projected sizes for radio AGN with the model predictions. The jet is highly collimated so that the main uncertainty is the amount of boosting due to the relativistic motion in the jet, determined by the bulk Lorentz factor $\gamma = (1 - V^2/c^2)^{-1/2}$, where V is the outflow velocity and c the velocity of light. It is thus difficult to match core- and lobe-dominated objects in intrinsic luminosity and so ensure that relativistic beaming effects

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are not a factor in any comparative studies. In general, these studies show that there are too many core-dominated sources compared with the lobe-dominated ones and that the former are too large when de-projected based on the observed Lorentz factor. There is also a lack of two-sided jets, i.e. sources aligned with the plane of the sky. Many of these inconsistencies can be solved if the highly luminous FR II radio galaxies (BLRGs and NLRGs) are included in the unification. In this case, the lobe-dominated, radio-loud QSOs are viewed at an angle intermediate between those of the radio galaxies, and the QSOs and the larger size of the radio galaxies matches that of the de-projected core-dominated sources. However this solution introduces a new problem, namely that the narrow [O III] $\lambda\lambda$ 4959, 5007 line emission is significantly (by a factor of 4–10) weaker in radio galaxies than in quasars, which requires that these lines also be emitted anisotropically for unification to work. The similar strength of the lower ionization [O II] λ 3727 line in the two classes supports this scenario since it originates at larger distances from the AGN core and so is expected to be emitted more isotropically.

Within the context of this orientation picture, the inclination of a radio source is related to the ratio of core to lobe luminosity, largely as a result of the relativistic boosting of the core luminosity. This parameter, known as core dominance R, can then be used to study the dependence of other parameters on inclination. For example the widths of the broad Hβ and C IV emission lines indicate that face-on (core-dominated) objects tend to have relatively narrow broad lines (FWHM $\approx$ 1000–5000 km s^{-1} while edge-on (lobe-dominated) have a wider range of line width (FWHM $\approx$ 1000–10000 km s^{-1}). This result suggests that the orientations of the broad emission-line region and the obscuring torus are similar.

Blazars and radio sources

With the recognition that blazar emission originates in relativistic jets towards the observer, radio galaxies with jets in the plane of the sky presented themselves as candidates for 'misaligned' blazars. Blazars themselves fall into two distinct categories, 'BL Lacs' and 'optically violent variables' (OVVs), and these seem to be drawn from different parent distributions of radio galaxies. The primary distinction between OVVs and BL Lacs is that OVVs have strong UV and optical emission lines so their UV–optical spectra resemble those of non-beamed AGN, whereas in BL Lacs other spectral features are swamped by the emission from the jet. OVVs also have higher luminosities and higher inferred Lorentz factors than BL Lacs.

It is possible to make a fairly convincing argument that BL Lacs reside in FR I radio galaxies. By disregarding the spatially unresolved blazar core emission and considering only the luminosity of the 'extended' radio source (presumably the 'lobes' seen along their axis), it is found that FR I galaxies and BL Lacs have comparable radio luminosity, supporting the notion that they are drawn from a single population. The luminosity functions and space densities are also consistent with unification of these two types of object. It should be mentioned at least in passing that BL Lacs themselves have been sometimes divided into radio-selected and x-ray-selected subcategories, and on average sources in these categories tend to have spectral energy distributions that peak, not surprisingly, at low and high frequencies, respectively. Recent work suggests that these are not, in fact, distinct classes, but rather represent the extremes of a continuous distribution in peak wavelength for the entire class of BL Lacs.

The more luminous FR II radio galaxies afford a plausible parent population for OVVs. Compared with the FR I–BL Lac unification scheme, this is more difficult to demonstrate, at least in part because the FR II population is relatively less homogeneous.

'Grand unification': radio-loud and radio-quiet AGN

One of the longest-standing questions in AGN astrophysics is why, despite being optically similar, a small minority ($\lesssim$10%) of AGN are radio loud while the remainder are radio quiet? Any unification scheme that can explain the difference between these two major classes of AGN would go a long way towards explaining the AGN phenomenon in general, and is therefore sometimes referred to as 'grand unification'. There is a clear dichotomy between the radio-loud and radio-quiet sources; for a given IR through soft x-ray luminosity, the radio-loud AGNs are 2–3 orders of magnitude more luminous in the radio than the radio-quiet objects. There are very few objects of intermediate radio luminosity so the distribution of radio power relative to power at shorter wavelengths is clearly bimodal. In other words, the evidence clearly points to two distinct AGN populations defined by relative radio luminosity.

An early suggestion advanced by Scheuer and Readhead was that the radio-loud and radio-quiet objects constitute a single population in which the radio-louds represent the sources with axes oriented towards the observer. This failed, however, to account for the lack of extended radio emission in the radio quiets.

Radio-quiet and radio-loud AGN differ in other respects. Radio-loud AGN are, on average, about a factor of 3 brighter in the mid-energy x-rays (0.2–3.5 keV) with somewhat flatter spectral slopes (i.e. relatively more high-energy photons). This is thought to be due to the presence of beamed X-ray emission related to the radio-emitting jet and provides evidence for the unification of core- and lobe-dominated radio-loud AGN discussed above. However, the difference in x-ray spectral slope would be hard to account for in a grand unification scenario.

The emission-line spectra of the two types of object also show differences, at least in a statistical sense (e.g. stronger optical Fe II blends, weaker narrow lines and narrower broad lines in the radio-quiet objects).

Until relatively recently it was believed that broad relatively blueshifted absorption-line features that are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

indicative of mass outflows are only found in radio-quiet AGN; it is now clear that these also occur in radio-loud AGN, although at the present time it appears that the flows in radio-loud objects have relatively lower terminal velocities. The outflowing material responsible for the absorption is generally thought to be accelerated away from the surface of the torus and so expected to be present at certain orientations in both radio-loud and radio-quiet AGN.

No satisfactory explanation for the differences between radio-loud and radio-quiet sources has yet emerged. Although hard to understand, it was for a long time thought that the nature of the host galaxy determined the radio class with radio-quiet AGN in spiral hosts and radio-loud AGN in ellipticals. However, recent observations with the Hubble Space Telescope have shown that at high luminosity nearly all AGN lie in elliptical galaxies. It is now widely suspected that the angular momentum of the central black hole might be the determining factor, since the orbital motions of charged particles around the black hole might determine whether or not electromagnetic effects will be sufficiently powerful to generate and collimate jets.

Bibliography

Antonucci R 1993 *Ann. Rev. Astron. Astrophys.* **31** 473

Goodrich R W 1997 *Emission Lines in Active Galaxies: New Methods and Techniques (Astronomical Society of the Pacific Conference Series vol 113)* ed B M Peterson, F-Z Cheng and A S Wilson, pp 445–52

Kembhavi A K and Narlikar J V 1999 *Quasars and Active Galactic Nuclei: an Introduction* (Cambridge: Cambridge University Press)

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

Urry C M and Padovani P 1995 *Publ. Astron. Soc. Pac.* **107** 803

Bradley M Peterson and Belinda J Wilkes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Galaxy

A galaxy that has an exceptionally high luminosity and which radiates large amounts of non-stellar radiation. A typical active galaxy will exhibit all or most of the following properties: an overall luminosity in excess of 10^{37} watts; a much higher output of x-ray, ultraviolet, infrared and radio radiation than a normal galaxy; a highly luminous and compact central core, known as an active galactic nucleus (AGN), which, in many cases, varies rapidly in brightness; narrow jets of radiating matter emerging from the central core; large-scale clouds of radio-emitting material; and, in some cases, broad emission lines in its spectrum. Much of the energy radiated by an active galaxy is emitted by charged particles moving at high speeds in magnetic fields (synchrotron radiation) rather than by stars

Active galaxies are classified according to their appearance, luminosity and spectra into the following principal types: Seyfert galaxies, radio galaxies, quasars, BL Lacertae objects (BL Lacs) and blazars.

See also: active galactic nucleus, astrophysical jets, blazar, BL Lacertae objects, galaxy, quasar, radio galaxy, Seyfert galaxies, synchrotron radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Optics

A system that enables the shape and relative positions of the principal optical elements of a telescope to be adjusted to compensate for optical errors induced by a variety of possible sources of deformation in the system. Sources of deformation include: flexure of the primary mirror or telescope structure as the orientation of the telescope changes, thermal effects induced by changing temperatures, buffeting by wind or manufacturing errors in the shape of the mirrors. An active optics system monitors the quality of the image of a reference star, calculates the optical errors (including, for example, coma and astigmatism) and adapts the shape and relative positions of primary and secondary mirrors to counteract these errors. The shape of the primary mirror is controlled by a system of actuators that applies forces to the rear of the mirror. The primary mirror in such a system has to be thin relative to its diameter in order to respond effectively to these forces. For example, the primary mirror in each of the four telescopes that comprise the European Southern Observatory's Very Large Telescope has a diameter of 8.2 m, a weight of 23 tonnes and a thickness of only 175 mm, the shape of each mirror being controlled by 150 actuators.

See also: adaptive optics, astigmatism, coma, mirror.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Active Region

A localized region of the Sun's surface and atmosphere that displays most or all of the following phenomena: sunspots, faculae, plages, filaments, prominences, flares and bright condensations in the corona. The underlying feature of a typical active region is a bipolar magnetic region, an area of concentrated magnetic fields that, although often complex in structure, contains approximately equal quantities of positive and negative magnetic flux. The overall diameter of an active region, within which the general magnetic field strength is about 0.01 T, may be several hundred thousand kilometers; localized fields of up to 0.4 T occur within sunspots.

An active region begins to form when magnetic flux tubes (bundles of magnetic field lines) erupt through the photosphere from below to create loop structures that have positive magnetic polarity where they emerge from the solar surface, and negative polarity where they re-enter the solar surface. Typically, an active region will grow to its maximum extent in about 10 days and may persist for several months, declining slowly as its underlying magnetic fields dissipate. The sunspots that form within an active region usually disappear within 2–4 weeks.

See also: chromosphere, corona, faculae, filament, flare, plage, prominence, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Adams, John Couch (1819–92)

Born in Laneast, Cornwall, England, he became fellow and mathematical tutor at Cambridge. He developed a procedure for numerical integration of differential equations and, inspired by MARY SOMERVILLE, he deduced mathematically the existence and location of the planet Neptune from its perturbations on the planet Uranus. Neptune was discovered by JOHANN GALLE in Berlin, in September 1846, using URBAIN LEVERRIER's independently computed position. Adams' claim to priority was advanced shortly thereafter, and it transpired that, with a letter of introduction from JAMES CHALLIS in Cambridge, Adams had applied to GEORGE AIRY, the Astronomer Royal, for some kind of assistance, though he failed to secure an interview with Airy, and nothing further happened. This seems to have been because of Adams' youth and unforceful personality, and Airy's unapproachability. Challis himself made a start to a somewhat half-hearted search for Neptune on the basis of Adams' calculations, but was overtaken by Galle. Adams was appointed professor of astronomy at Cambridge in 1858, and was director of the Cambridge Observatory from 1861.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Adams, Walter Sydney (1876–1956)

Born in Antioch, Syria, to American missionary parents, he worked under GEORGE HALE at Yerkes Observatory at the University of Chicago, accompanying Hale to California to set up the Mount Wilson Observatory, and becoming its director on Hale's retirement. He helped design the 200 in telescope for Mount Palomar Observatory. His method of spectroscopic parallaxes, a technique using spectra to give indications of stellar luminosities, made it possible to measure the distances to more distant stars than those whose parallax could be measured by the trigonometric method.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Adaptive Optics

A system that enables the optical surfaces of telescope mirrors to be continuously modified to compensate for distortions in incoming light waves as they pass through the Earth's atmosphere. Light waves from a distant source, such as a star, arrive at the top of the atmosphere as straight parallel wavefronts (rather like waves advancing across the surface of the ocean) but are then distorted by the atmosphere and arrive as 'corrugated' wavefronts at the primary mirror of a telescope. Because all parts of such a wavefront cannot be focused at the same point, the image quality is degraded, so limiting, quite severely, the resolutions that can be achieved by large ground-based telescopes. The adaptive optics technique senses and counteracts these wavefront distortions.

Wavefront distortions are sensed by monitoring a suitable bright star, if there happens to be one in the field of view of the telescope, or by monitoring an artificial 'star' generated by shining a powerful sodium laser beam into the upper atmosphere. The laser beam stimulates the emission of light by sodium atoms in a layer that exists at an altitude of about 100 km, so generating a star-like point of light that may be positioned conveniently close to the object that is being studied by the telescope. Because light from this artificial 'star' travels down to the telescope along (almost) the same path as the light from the object of interest, it suffers the same wavefront distortions. A wavefront sensor measures the distortions, and a real-time control computer instructs a set of actuators located behind a thin flexible mirror to alter the mirror shape and thereby cancel out the wavefront distortions. Because the distortions induced by the atmosphere are continually changing, the system has to sense and respond to them as rapidly as possible—tens or even hundreds of times per second.

In principle, adaptive optics systems ought to be able to allow telescopes to achieve resolutions approaching their theoretical limits. A system of this kind has been in operation on the European Southern Observatory's 3.6 m New Technology Telescope since 1992. Adaptive optics systems on the Very Large Telescope (an array of four 8.2 m telescopes) are expected to enable this instrument to achieve its theoretical (diffraction limited) resolution at near-infrared wavelengths.

See also: active optics, instrumentation: adaptive optics, mirror, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Adelard of Bath (1075–1160)

Born in Bath, England, a teacher and translator of scientific work from the Arabic, Adelard traveled in France, Italy, Syria and Palestine. On returning to Bath, he translated Euclid's *Elements* from Arabic, and his translation became for centuries the chief geometry textbook in the West. He also translated the *zij* (astronomical table) of Muhammed ibn Musa al-Khwarazmi. He wrote works on the abacus and on the astrolabe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aepinus, Franz Ulrich Theodosius (1724–1802)

Mathematician, born in Rostock, Mecklenberg-Schwerin (now Germany), he was director of the Observatory in Berlin and a professor at St Petersburg. His book *An Attempt at a Theory of Electricity and Magnetism* was the first work to apply mathematics to such a theory, and he demonstrated the effects of parallax in the transit of a planet (Venus) across the Sun's disk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aerolite

An obsolete name for a stony meteorite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aglaonike (*c*. 200 BC)

A natural philosopher of ancient Greece, she is referred to as a sorceress who could make the Moon disappear at will. She might therefore be regarded as the first recorded woman astronomer in that she seems to have mastered the art of predicting lunar eclipses, presumably by knowledge of the Metonic cycle, by which lunar eclipses repeat themselves every 19 years. The cycle was discovered in the fifth century BC by Babylonian astronomers and named for METOS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Air Force Maui Optical Station

Located on the island of Maui, Hawaii. Instruments include visible and infrared sensors and the 3.67 m (145 in) Advanced Electro-Optical System (AEOS) telescope, which is fitted with adaptive optics. This is the US Air Force's largest and most advanced telescope system. The telescope is used to track and image Earth-orbiting satellites and astronomical objects such as asteroids.

The AEOS telescope was made available to the civilian astronomical community on 1 September 1999, under a pilot program agreed by the Air Force Office of Scientific Research and the National Science Foundation's Division of Astronomical Sciences. The agreement will make available 100 observing nights over a 2 year period to determine the level of interest in the AEOS facility from the US astronomical community.

For further information see http://ulua.mhpcc.af.mil/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Airglow

The airglow consists of the non-thermal radiation emitted by EARTH'S ATMOSPHERE. The thermal emission of Earth, heated by sunlight and subsequently radiating to space, takes place in the infrared portion of the spectrum, at wavelengths longer than 7 μm. In its simplest form, the airglow arises from de-excitation of molecules, atoms or ions excited by solar UV photons or energetic particles, and the resulting emitted photons are generally, although not exclusively, located in the visible and UV ranges of the spectrum. The existence of a night sky glow, long after sunset, shows that indirect mechanisms play an important role in producing airglow. In this case, the initial energy, brought by solar photons, is stored in the high atmosphere under the form of radicals or molecules able to recombine through exothermic reactions during the night and produce atoms or molecules in their excited states, with subsequent photonic emissions. The airglow does not include the radiation produced by lightning or meteor trains. Although the distinction is more difficult to establish, auroral phenomena, occurring at high latitudes and resulting from localized, intense precipitation of particles whose origins are outside the atmosphere, are not considered as being airglow. In contrast to AURORAS, and in spite of the fact that the general mechanism of emission (excitation–de-excitation) is of the same nature, the airglow occurs at any time and any latitude, in a much less sporadic way.

Historically, the existence of a terrestrial component of the light in the night sky was recognized in the years around 1900. Yntenna, who first established by photometric methods the existence of airglow, remarked on the large variability of this phenomenon from night to night and showed that the star light scattered by atmospheric molecules was insufficient to explain the night sky light. Spectroscopic techniques showed that the green line of atomic oxygen at 558 nm was present over the whole sky at all times and confirmed the existence of a 'permanent aurora'. Rayleigh (JOHN WILLIAM STRUTT) showed in the 1920s that the geographical pattern of the oxygen green line differed from that of the aurora, and that the N_2^+ bands seen in auroras were absent from the night sky. Rayleigh first expressed the brightness of the green line in terms of the number of atomic transitions per second in a column of unit cross-sectional area along the line of sight. He gave his name to the photometric unit universally used today for airglow studies: one rayleigh (or 1 R) represents an omnidirectional emission rate of 10^6 photons per square centimeter per second. The oxygen red lines at 630–636 nm were predicted and observed around 1930, as well as the D line of sodium.

Although ground-based spectrophotometric instruments provide a convenient way to monitor the spatial and temporal variations of the airglow, they are not appropriate for the retrieval of the emission vertical profiles. One possibility is to use rocket-borne zenith photometers flown through the airglow layers. Differentiation of the signal with respect to altitude allows the retrieval of the local volume emission rate as a function of altitude. Another way to proceed is to perform airglow measurements from an orbital platform in limb viewing geometry. This technique has several advantages. Because the signal is integrated along a tangential path, the intensity of the measured signal is greatly enhanced. It makes weak features, like the airglow continuum emission, easier to detect and characterize. Second, the tangential view allows the separation of the background galactic signal from the atmospheric glow because the galactic signal is directly provided by measurements performed at very high altitudes, where the atmospheric density is virtually zero.

The vertical profiles of several identified night airglow lines, as observed from the space shuttle STS-37 mission orbiting at 267 km altitude, are plotted in figure 1. The well known green line of oxygen at 557.7 nm, produced by the $O(^1S) \rightarrow O(^1D)$ transition, covers a relatively narrow altitude region and strongly peaks around 100 km altitude. Some emission is also recorded in the red line of oxygen at 630 nm, due to the $O(^1D) \rightarrow O(^3P)$ transition. The signal at 762 nm results from radiative de-excitation of an excited state of O_2. The case of atomic and molecular oxygen lines will be treated in a more detailed way in the following section. The D line of Na at 589 nm, due to $Na(^2P) \rightarrow Na(^2S)$, peaks around 90 km, similarly to the OH (6–0) Meinel band at 527 nm. The OH emission is produced by the reaction cycle

$$H + O_3 \rightarrow OH^* + O_2^*$$

where OH^* is a rotationally excited level of OH, with a vibrational quantum number ν less than 9 (no bands originating from $\nu > 9$ are detected) and a recycling mechanism of the type

$$OH + O \rightarrow H + O_2.$$

The 527 nm band is produced by the $\nu = 6 \rightarrow 0$ de-excitation of OH*. Note that the strong peak at low altitude in the 630 nm emission is due to the (9–3) band of OH. Because the chemistry of sodium also involves ozone (O_3) through the cycle

$$Na + O_3 \rightarrow NaO + O_2$$
$$NaO + O \rightarrow Na + O_2$$

a coupling between the airglow OH intensities and the Na density (derived either from lidar measurements or from airglow measurements) may be expected, and is definitely exhibited by measurements.

There is also a bright emission that extends from 400 to 600 nm. It is due to O_2 in the near ultraviolet and blue ranges, up to 480 nm, and NO_2 at longer wavelengths. The O_2 band emission mainly consists of Herzberg I, Herzberg II and Chamberlain bands, as detailed in the next section. The NO_2 diffuse feature may be attributed to the radiative recombination of NO and O, and is a true continuum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

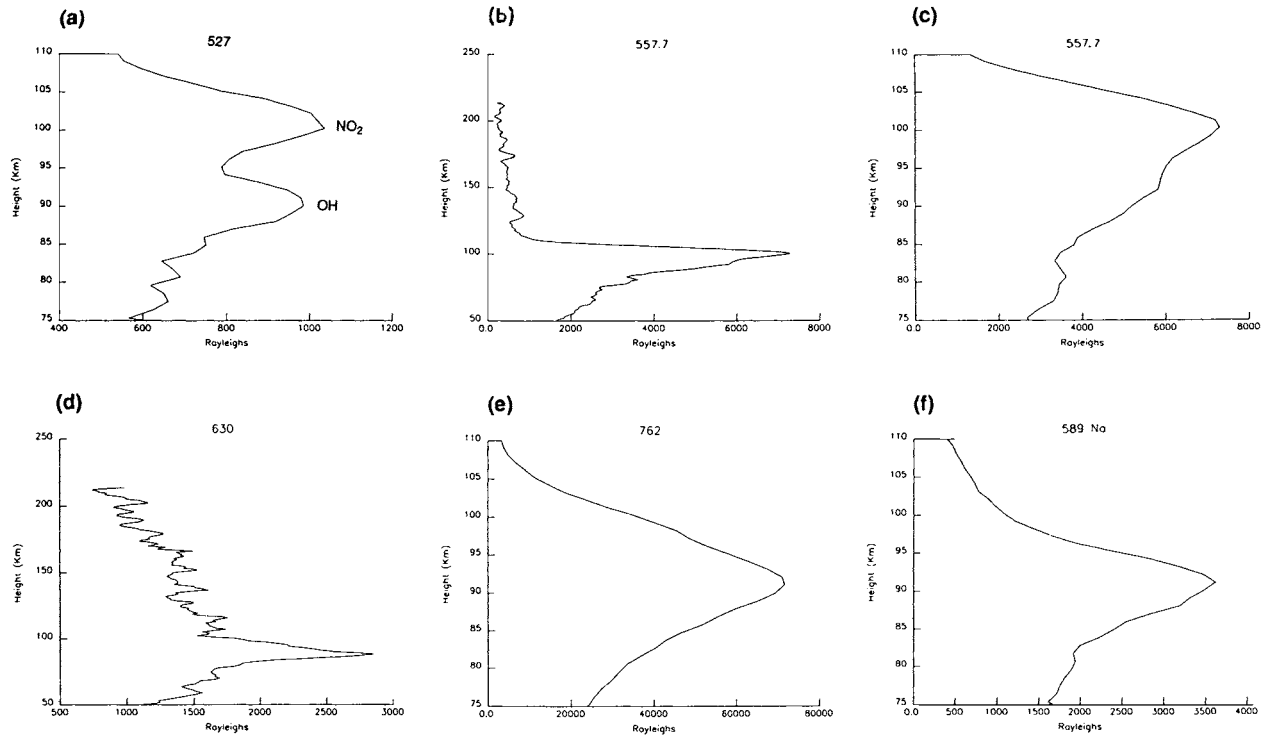


Figure 1. Altitude profiles of Earth's airglow in six wavelength ranges as seen from the space shuttle on the STS-37 mission. (From Mende *et al* 1993 *J. Geophys. Res.* **98** 19 117.)

A profile taken at 527 nm is shown in figure 1. The marked airglow peak between 95 and 105 km altitude is in good agreement with microwave measurements of thermospheric NO made during the same solar cycle, suggesting a maximum concentration region between 100 and 110 km.

The study of upper atmospheric airglow (dayglow and nightglow) emissions, either from ground-based spectrometry or by using spectrometers as well as imagers on board satellites like the Dynamics Explorer and Space Shuttle, has seen very substantial progress in the last 20 years. These measurements have allowed the characterization of thermospheric and ionospheric processes, in fields of both aeronomy and dynamics, with the help of laboratory and theoretical advances. The interpretation of airglow measurements is extremely complicated, involving the detailed characteristics of the excitation fluxes (solar UV/EUV, particles), interaction mechanisms (direct excitation of molecules, atoms or ions, indirect excitation through dissociation and ionization processes, as well as chemical reactions), chemical cycles, radiative transfer of airglow light. Oxygen emission lines in the nightglow are now detailed.

Excitation of oxygen emissions in the nightglow of Earth

The red doublet of oxygen at 630 nm and 636.4 nm, resulting from the $O(^1D) \rightarrow O(^3P)$ transition, that is from the first excited level to the ground state, with transition probabilities of 0.0069 and 0.0022 s^{-1} respectively, is observed during several hours after sunset. Since the average emission time of $O(^1D)$ is equal to the inverse of the transition probability, that is a few minutes, direct excitation of oxygen atoms by solar UV photons followed by de-excitation cannot explain the persistence of the red glow for a long time after sunset. The source of $O(^1D)$ during night time is dissociative recombination of O_2^+ ions

$$O_2^+ + e^- \rightarrow O + O(^1D).$$

The O_2^+ ions result from the night time decay of ionization which is controlled by charge exchange reactions that converts O^+ ions into O_2^+ ions.

The vertical distribution of the red line emission is complex and may be shown to evolve with time in a different way at different altitudes. Below 200 km, the red line emission is weak because O^* is deactivated (or 'quenched') by collisions with air molecules in a time smaller than the emission time. Above 200 km, the combined effects of the collisional quenching rate, which decreases with altitude, and the vertical variation of the electronic density, which increases with altitude, results in a strong increase, for increasing altitudes, of the red line glow. An important application of the measurement of thermospheric nightglow emissions, like the oxygen red line, is the inference of winds and temperatures from high-resolution spectroscopy. By using Fabry–Perot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and Michelson interferometers, the thermospheric neutral wind at the 250 km altitude level has been retrieved from the O(^{1}D) 630 nm emission line by measuring the associated Doppler shift, which yields the line-of-sight wind vector. Neutral temperatures may be inferred from the width of the line. Doppler spectroscopy of airglow lines is a powerful means to study the dynamics of the thermosphere, strongly controlled by thermal tides and general wave activity.

The green line of oxygen at 558 nm, which results from the O(^{1}S)→O(^{1}D) transition between the second and the first excited states, is known to arise from the 100 km neighborhood, that is definitely lower atmospheric levels than the red lines. Although a part of the green line may arise from dissociative recombination of O_2^+ in the F region

$$O_2^+ + e^- \rightarrow O^* + O^*$$

another mechanism must be invoked to explain the low emission altitude of this line. Sydney Chapman proposed a mechanism of 'photochemical excitation'

$$O + O + O \rightarrow O_2 + O(^1S)$$

where one of the O atoms is excited by some of the energy liberated in the association of two O atoms. This process results in a maximum O concentration around 100 km, which is consistent with the location of the green line emission. De-excitation of O(^{1}S) into O(^{1}D) occurs through quenching by ambient air molecules (mainly molecular nitrogen).

The so-called Chapman process, although it may participate in the green line emission, is now thought to be of secondary importance. At the beginning of the 1960s, Barth and Hildebrandt found in the laboratory that the emission produced by the previous process is much too small to account for the observed glow. This result stimulated Barth to put forward, in 1964, the following two-step mechanism:

$$O + O + M \rightarrow O_2^* + M$$

where M is an air molecule and O_2^* an excited state of the O_2 molecule, followed by the excitation transfer process

$$O_2^* + O(^3P) \rightarrow O_2 + O(^1S).$$

The net effect of these two reactions, obtained by adding them, is similar to the effect of the single Chapman reaction. The other loss processes are deactivation through quenching by collision with an air molecule

$$O_2^* + M \rightarrow O_2 + M$$

or radiation of a photon through radiative de-excitation

$$O_2^* \rightarrow O_2 + h\nu.$$

Therefore, the Barth process may give rise simultaneously to both O(^{1}S) (green line) and O_2 (762 nm line) emissions. Comparison of the two corresponding profiles in figure 1 may be shown to be consistent with the two-step scheme, despite the 10 km difference between the peak altitudes.

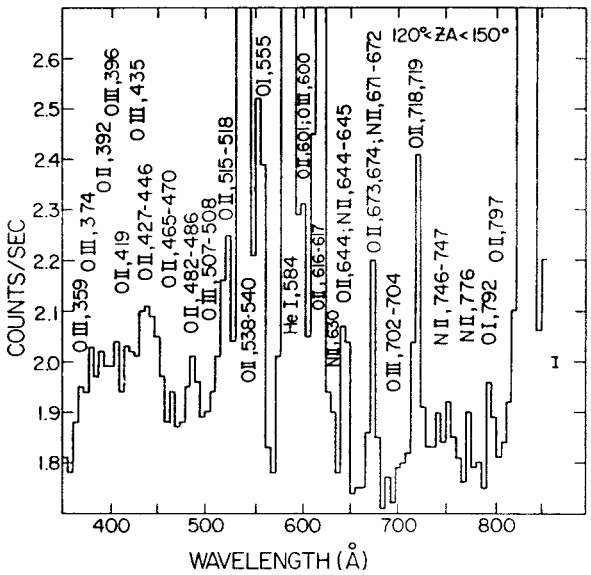


Figure 2. Nadir viewed UV dayglow spectrum between 350 and 950 Å acquired by the STP78-1 satellite from 600 km between 50°N and 50°S latitudes. (From Chakrabarti *et al* 1983 *J. Geophys. Res.* **88** 4898.)

Dayglow and twilight emissions

Airglow emissions are also observed during the day, and they mainly reflect the diurnal variation of the concentration of emitting species. Twilight spectra may be observed at the transition between the dark and fully illuminated atmosphere, as the shadow height moves vertically over the full range of emissive layers. In the day time, several lines and bands are very bright, mainly due to higher production rates. Indeed, dayglow and twilight emissions mainly result from direct excitation by solar ultraviolet photons, photoelectrons produced through ionization processes in the ionosphere, and by resonance scattering or fluorescence of solar radiation. Such energetic phenomena result in a complex ultraviolet dayglow spectrum, as shown in figure 2. These measurements were made from an Earth-orbiting satellite (STP78-1), since ultraviolet photons are prevented from penetrating down to the ground due to the screening effect of the atmosphere. Even at near-UV or visible wavelengths, dayglow are difficult to acquire with ground-based instruments because of the high level of the background scattered solar light.

The short-wavelength ultraviolet spectrum is dominated by singly ionized and neutral lines of oxygen (O I, O II) and nitrogen (N I, N II). These lines are produced by ultraviolet photons and photoelectron impact. Short-wavelength lines are produced by de-excitation from high-energy upper states, which cannot be excited by chemical or recombination reactions. Several atomic and ionic lines in the dayglow spectrum are also lines that appear in the solar spectrum, like the He I line at 584 Å. In the case of helium, this phenomenon is due to resonant scattering of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the He solar emission line. Similar features are observed at 834 Å (O II triplet), 1216 Å and 1025 Å (H I, Lyα and Lyβ), 1302–1306 Å (O I triplet) and others wavelengths. They are due to resonant scattering or fluorescence.

Molecular oxygen emissions in the airglows of Earth, Venus and Mars

Whereas the case of the O(^{1}S) state is well understood, the detailed process of O_2 emissions is still not well known. In addition to the discrete 762 nm O_2 band, there is a near-UV–blue extended emission attributed to Herzberg and Chamberlain bands. In previous reactions, the excited bound state O_2^* of O_2 can be $A\,^3\Sigma_u^+$, $A'\,^3\Delta_u$ or $c\,^1\Sigma_u^-$ (figure 3). The main precursor of O(^{1}S) could be $O_2(c\,^1\Sigma_u^-)$, but it is still not certain what the nature of the precursor is, and that is still an active field of investigation today. In Earth's atmosphere, the Herzberg I system is the major cause of the blue nightglow continuum, with an average emission of 400 R. From spaceborne observations of the Venus airglow by the Soviet Venera probes, similar oxygen airglows were found, although in quite different amounts. The Venus molecular oxygen nightglow shows a strong pre-eminence of the Herzberg II system (2700 R). Such large differences are due to the different chemical compositions of the atmospheric environment for Earth and Venus (see VENUS: ATMOSPHERE). In the luminous layers, the ambient air (N_2+O_2 for Earth, CO_2 for Venus) density is nearly 100 times larger on Venus than on Earth, the reverse situation occurring for O (three times less on Venus than on Earth). Because the collisional quenching of $O_2(A\,^3\Sigma_u^+)$ is mainly realized by air molecules, whereas $O_2(c\,^1\Sigma_u^-)$ is preferentially quenched by collisions with O, Herzberg I bands are weak on Venus and Herzberg II bands are strongly enhanced with respect to the case of Earth (2700 R on Venus versus 100 R on Earth). Another interesting feature is the extreme weakness of the oxygen green line in Venus nightglow (5 R versus 120 R on Earth), which was an important element in favoring the Barth mechanism over to the Chapman process.

As seen in figure 3, a glow of the infrared atmospheric system of $a\,^1\Delta_g \rightarrow X\,^3\Sigma_g^-$ is expected in the infrared, at 1.27 μm. This glow, first detected on Venus by the French spectroscopist Pierre Connes and his collaborators at the end of the 1970s is quite intense (>1 MR, that is 10^6 R). Such a strong emission is also measured on Earth. Due to the long radiative lifetime of $O_2(a\,^1\Delta_g)$ (1 h), the analysis is complicated by transport across the terminator. Photolysis of O_3 under sunlit conditions according to the following reaction

$$O_3 + h\nu \rightarrow O_2(a\,^1\Delta_g) + O(^1D)$$

constitutes part of the source in the dayglows of Earth and Mars, but it is negligible on Venus. The major source in the nightglows of all the terrestrial planets, the dayglow of Venus and a part of the dayglows of Earth and Mars, is the recombination of O atoms produced in the photolysis of O_2 on Earth and CO_2 on Venus and Mars (see MARS: ATMOSPHERE).

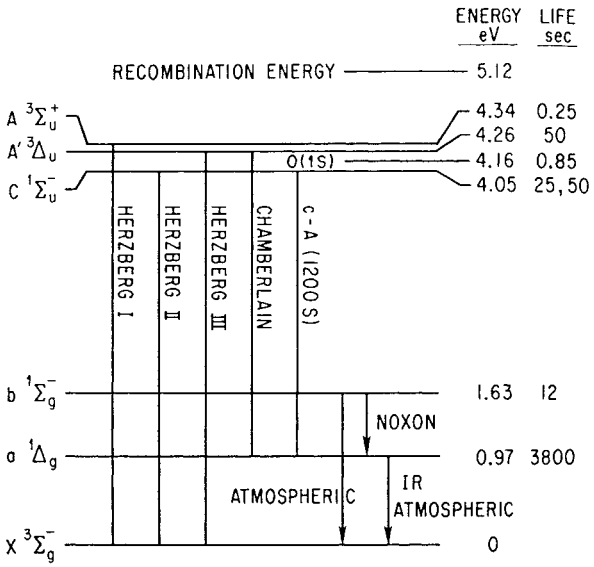


Figure 3. The six lowest levels of molecular oxygen and the corresponding band systems (from Krasnopolsky 1983).

Connes measured individual rotational lines of this band in the Martian spectrum with the 5 m telescope on Mount Palomar in 1973 and 1975. Rotational temperatures and airglow intensities were derived from the spectra. In each spectrum, equivalent widths of about 30 lines were measured to be typically 0.005–0.01 nm. By ratioing the equivalent width to the line strength, and by examining the dependence of this ratio on the energy of the upper rotational state of the transition, a rotational temperature of 200 K was deduced, which corresponds to the real temperature averaged over the airglow altitude profile. Typical airglow intensities of 3 to 30 MR, depending on latitude, were found. Assuming the quenching rate coefficient by air molecules (CO_2 on Mars) is known, it is possible to infer the ozone amount from 1.27 μm airglow measurements, although in a rather inaccurate way.

Bibliography

Barth C A, Stewart A I F, Bougher S W, Hunten D M, Bauer S J and Nagy A F 1992 Aeronomy of the current Martian atmosphere *Mars* ed H H Kieffer, B M Jakosky, C W Snyder and M S Matthews (Tucson, AZ: The University of Arizona Press) pp 1054–89

Chamberlain J W 1995 *Physics or the Aurora and Airglow (Classics in Geophysics volume 1)* (Washington, DC: American Geophysical Union)

Fox J L 1997 Airglow *Encyclopaedia of Planetary Sciences* ed J H Shirley and R W Fairbridge (London: Chapman and Hall)

Krasnopolsky V A 1983 Venus spectroscopy in the 3000–8000 Å region by Veneras 9 and 10 *Venus* ed D M Hunten, L Colin, T M Donahue and V I Moroz (Tucson, AZ: The University of Arizona Press) pp 459–83

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rees M H 1989 *Physics and Chemistry of the Upper Atmosphere* (Cambridge: Cambridge University Press)
Solomon S C 1991 Optical aeronomy *Rev. Geophys. Suppl.* 1089–109

Eric Chassefiére

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Airy Disk

The bright spot at the center of the diffraction pattern that is formed when a point source, such as a star, is imaged at the focus of a telescope. Light waves from a distant star arrive at the Earth as a series of parallel wavefronts (rather like waves rolling onto a beach). When these wavefronts are interrupted at the edge of a telescope's aperture (the rim of the objective or primary mirror), and are brought to a focus, interference effects between different parts of each wavefront result in the formation of a diffraction pattern. For a point source, the resulting image (assuming a perfect optical system) consists of a central spot of light surrounded by a series of light and dark fringes, or rings. According to theory, 84% of the light energy is concentrated into the central spot (the Airy disk), the diameter of which depends on the aperture of the telescope and the wavelength of the light. The radius of the first dark minimum in light intensity, that lies immediately outside the Airy disk, is given, in radians, by $1.22\lambda/D$, where λ denotes wavelength and D the aperture of the telescope. The central spot is known as the Airy disk because this type of diffraction pattern was first investigated by Sir George Biddell Airy (1801–92), the seventh Astronomer Royal.

See also: aperture, Rayleigh limit, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Airy, Sir George Biddell (1801–92)

A brilliant Cambridge mathematician (Senior Wrangler 1823, i.e. leader of the graduating mathematics class), Airy became the seventh Astronomer Royal in 1835 after a brief period as Lucasian Professor at Cambridge. His output was prodigious, and he published nearly 400 scientific papers and 150 reports on various scientific issues, such as the gauge of railways, spectacles to correct astigmatism, and methods to correct for compass readings in ships made of iron. His work on optics is recognized by the use of the term *Airy disc* for the resolution element of a telescope due to diffraction at its aperture, which he studied. As Astronomer Royal, he saw that the Royal Observatory at Greenwich was re-equipped with modern instruments and that its work was carried out punctiliously by its many human 'calculators' and observers, insisting that everything was carried out in the best way, namely his way. This regime was effective in raising standards. He himself determined the density of the Earth, the mass of the planet Jupiter and its rotation period. He calculated the orbits of comets and cataloged stars. He made numerous contributions to the prediction of the motion of the Moon, and analysed transits of the planet Venus and eclipses, including the eclipse predicted by THALES. As a result of the accuracy of the observations made under Airy at Greenwich, and his practical exploitation of the railway telegraph to distribute it, Greenwich Mean Time was established in 1880 as the official time service throughout Britain, and afterwards became the basis for the timekeeping of the world.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aitken, Robert Grant (1864–1951)

Astronomer, born in California, Aitken became director of the Lick Observatory. He made systematic visual surveys of binary stars, measuring their positions visually. His massive *New General Catalogue of Double Stars* (1932) contained data on 17 000 stars that allowed orbit determinations of many, from which it was possible to calculate data on the masses of the stars' components. He also measured positions of comets and planetary satellites and computed orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Akebono (EXOS-D)

Japanese satellite launched in 1989 to study Earth's auroras. Named from the Japanese for the 'rising Sun'. Auroral image capability was lost in early 1995, but other instruments continue to operate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Battani, Abu Abdullah [known as Albategnius] (*c*. 868–*c*. 929)

Born in Harran, in present-day Syria, al-Battani spent most of his life in Raqqa, situated on the bank of the Euphrates, where he was educated and flourished as a scholar and a Sabean which means 'worshipper of the stars'. His best-known achievement was the determination of the solar year as being 365 days, 5 hours, 46 minutes and 24 seconds, an extremely accurate value that was used in the Gregorian Calendar reform of the Julian Calendar. Using trigonometrical rather than geometrical methods, he also determined the true and mean orbit of the Sun, proving the variation of the apparent angular diameter of the Sun (an indication of the variable distance between Sun and Earth) and the possibility of annular eclipses. He wrote a number of books on astronomy and trigonometry, his most famous book being his astronomical treatise with tables (a *zij*), which was translated into Latin in the twelfth century as *De Scienta Stellarum—De Numeris Stellarum et Motibus*. His treatise on astronomy was extremely influential in Europe till the Renaissance, with translations available for centuries in several languages.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Albedo

A measure of the reflectivity of a material or object. For bodies in the solar system, albedo is the proportion of sunlight falling on them that is reflected away. It is measured on a scale from 0 (a perfectly absorbing black surface) to 1 (a perfectly reflecting white surface).

Albedo may be defined in various ways. The most common is *geometrical albedo*, sometimes called *physical albedo*. This is the ratio of the amount of light reflected from an object, as viewed from the direction of the Sun, to the amount that would be reflected by a perfectly diffusely reflecting sphere (which by definition has an albedo of 1) of the same size and at the same distance as the object. As well as applying to visible light, geometrical albedo may be defined for radiation in any specified wavelength range; *bolometric geometrical albedo* applies to all wavelengths. *Bond albedo*, also known as *spherical albedo*, is defined similarly but for radiation of all wavelengths reflected in all directions; it is named after George Phillips Bond. Another measure is *hemispherical albedo*: the ratio of scattered to incident light as a function of the angle of incidence, and again the body is assumed to be a perfectly diffusely reflecting sphere.

Of the major planets, Mercury has the lowest geometrical albedo, at 0.11 (comparable to the Moon's 0.12), while Venus, by virtue of its blanket of highly reflective clouds, has the highest, 0.65. The lowest albedo in the solar system seems to be possessed by the particles that make up the rings of Neptune, which probably have a value close to the theoretical lower limit of zero. Asteroids which originally formed between about 2.5 and 3.5 AU from the Sun, where it was cool enough for dark carbonaceous compounds to condense from the solar nebula, but not water ice, have albedo values as low as 0.02. The highest measured albedo is that of Saturn's satellite Enceladus, to which some sources assign a value of 1.0.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Albireo

The star β Cygni, said to represent the eye of the swan in the ancient constellation figure. It is seen as a single star of apparent magnitude about 2.9 with the unaided eye, but with good binoculars or a small telescope it appears as a beautiful double, comprising an orange star (β^1 Cyg) of apparent magnitude 3.05, spectral type K3II, with a bluish-white companion (β^2 Cyg) of apparent magnitude 5.17 and spectral type B8V. Their separation is 34.4″ at position angle 54°. The system is 380 light-years distant, having a parallax of 0.008″. The absolute magnitudes of the two stars are β^1 Cyg −2.3, β^2 Cyg −0.2.

With more powerful instruments β^1 Cyg can be further resolved: it is a binary system with components of apparent magnitude 3.4 and 5.5, separated by 0.39″ at position angle 152°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Biruni, Abu Raihan (973–1048)

Born in Kheva near Ural, present-day Uzbekhistan, al-Biruni was a polymath and traveler (to India), making contributions in mathematics, geography and geology, natural history, calendars and astronomy. His book *Qanun-i Masoodi*, which he dedicated to his patron Sultan Masood, discusses astronomy, trigonometry, solar, lunar and planetary motions, including the question whether the Earth rotates or not. He undertook experiments (observations) related to astronomical phenomena, for example eclipses and the dates of the equinoxes, and determined that, compared with the speed of sound, the speed of light is immense. He recognized the Milky Way as 'a collection of countless fragments of the nature of nebulous stars'. Al-Biruni's quest for scientific knowledge is epitomized by his statement that the phrase *Allah is omniscient* does not justify ignorance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Albumazar [Abu Mashar al-Balkhi; Abu-Mashar; Abu Maschar] (787–885)

Astronomer and astrologer, born in Balkh, Afghanistan, flourished in Baghdad. An eleventh century Latin translation of his *Introduction to Astrology* introduced the West to Aristotelian physics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aldebaran

The star α Tauri, a red giant of spectral type K5III, apparent magnitude 0.87. Its distance is 65 light-years, its parallax 0.050″. The absolute magnitude is −0.6. Aldebaran appears at the eastward corner of the striking triangular open cluster the Hyades, although it is not actually a member of the cluster which is much more distant (about 100 light-years). The name derives from the Arabic *Al Dabaran*, 'the Follower'—presumably of the Pleiades.

Aldebaran was one of the 'Royal Stars' or 'Guardians of the Sky' of the Persian astronomer/astrologers *c*. 3000 BC. Although not the brightest stars these four were carefully chosen, apparently to mark the seasons, as they are approximately 6 h apart in right ascension. Aldebaran (ancient Persian name Tascheter) was prominent in the evening sky in March and was associated with the vernal equinox; the others were Regulus (summer solstice), Antares (autumnal equinox) and Fomalhaut (winter solstice).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aldrin, Edwin Eugene, Jr ['Buzz'] (1930–)

Astronaut, born in Montclair, NJ, trained as a fighter pilot (Korean War), he space-walked during the Gemini 12 mission, and was pilot of the lunar module of Apollo 11, the first manned lunar landing. Aldrin followed NEIL ARMSTRONG as the second man to walk on the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alexis (Array of Low Energy X-ray Imaging Sensors)

US Department of Energy satellite launched under the US Air Force Space Test Programme in April 1993. Recovered after the loss of a solar panel during launch. Designed to monitor soft x-ray sources.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Farghani, Abu'l-Abbas Ahmad ibn Muhammad ibn Kathir (*c*. 860)

Born in Farghana, Transoxiana (northern Iran), he wrote *A Book on Celestial Motion and Science of the Stars* (*Kitab fi al-Harakat al-Samawiya wa Jawami Ilm al-Nujum*), which was translated several times into Latin and Hebrew starting in the twelfth century. The book spread knowledge of Ptolemaic astronomy in Europe, at least until this role was taken over by Sacrobosco's *Sphere* on which it was based; DANTE, for example, used it as a source of astronomical knowledge. He determined the diameter of the Earth to be 6 500 miles and. found the distances and diameters of the planets. He also wrote a summary of the *Almagest* and a *Book on the Construction of Sun-dials*. He was additionally an engineer, but not a very good one, judging by the story that he built the beginning of an irrigation canal from the Tigris deeper than the rest of its length, so that not enough water would run through to its other end. The explanation given for Al-Farghani's mistake is that, a theoretician rather than a practical engineer, he had never successfully made anything.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alfonso X [Alfonso X of Castile; Alfonso the Astronomer; Alfonso the Wise] (1221–84)

King of León and Castile (from 1252), born in Burgos, Spain, founded the Castilian national culture. He commissioned the first general history of Spain to be composed in Castilian, a translation of the Old Testament, a great code of laws (*Siete Partidas*) and a planetary table. The planetary tables were a development of the work of Azarchiel (the *Toledan Tables*). These included data on astronomical phenomena, and calculations of the future positions of the planets according to Azarchiel's re-working of the Ptolemaic theory. It was Alfonso who financed and directed a team of astronomers led by Ishak ben Said, called Hasan, and Jehuda ben Mose Cohen to make the calculations, which were finished just as Alfonso became king, and published as the *Alfonsine Tables*. They survived as the standard work of astronomical prediction in Europe in manuscript and printed form for 300 years. Alfonso remarked during this time, apparently because of the arbitrariness and complexity of the Ptolomaic theory of the planets, that he wished that he had been around at the time of the creation since he could have given the Almighty some good advice. The major discrepancies which built up with time between the calculations in the *Alfonsine Tables* and reality was the inspiration to TYCHO BRAHE in 1563 at the age of 16 to begin making systematic astronomical observations, a motivation stimulated into action by his startling observations of the New Star of 1572.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alfvén, Hannes Olof Gösta (1908–95)

Physicist, born in Norrkorping, Sweden, Nobel prizewinner (1970), developed plasma physics, and applied theory to a wide range of physical and astrophysical phenomena including charged particle beams in accelerators, and interplanetary and magnetospheric physics. He founded magnetohydrodynamics, the branch of physics which shows how magnetic and electrical phenomena act on gases that are hot and in motion. His theory is used in rocket propulsion, aerodynamic braking of space vehicles, explanations of solar–terrestrial 'weather', the formation of comet tails, the formation of the solar system, the dynamics of plasmas in our Galaxy, and the fundamental nature of the universe itself. In 1937 he postulated the galactic magnetic field as the 'bottle' which contained cosmic rays, and in 1950 identified non-thermal radiation from astronomical sources as synchrotron radiation, which is produced by fast-moving electrons in the presence of magnetic fields: this is essentially the theoretical basis for most of radioastronomy. Alfvén held positions in Uppsala, Stockholm and California. His ideas were often original and therefore regarded as unorthodox: his career was marked by disputes over publications and controversy with established leaders. He was active in the worldwide nuclear disarmament movement. *Alfvén waves*, named after him, are electromagnetic waves that propagate through a highly conducting medium, such as the ionized gas of the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Algol

Algol, also known as β Persei, is the prototype of a class of ECLIPSING BINARY STARS that undergo dramatic brightness changes during the eclipses. Like other Algols, Algol itself consists of a hot, unevolved primary and a cool, evolved secondary. It is this large difference in temperatures between the two stars that gives rise to the dramatic eclipses. When the hot primary is eclipsed by the cool secondary, the overall brightness of the system drops from its out-of-eclipse brightness of 2.1 magnitudes to about 3.4 magnitudes. These eclipses, repeating every 2.87 days, are easily seen with the naked eye, and it is obvious that the ancients observed them given that Algol is invariably associated with something evil or demonic. The name Algol comes from the Arabic *Ra's Al Ghul* meaning 'the head of the demon.' The Greeks saw Algol as the blinking eye of Medusa, whose severed head was held aloft by the hero Perseus. The modern discovery of the eclipses was made by MONTANARI in 1670 and their proper explanation was made by GOODRICKE in 1782.

A triple system

Algol has been studied extensively from the radio to the x-ray end of the spectrum with techniques ranging from photometry and spectroscopy to interferometry. Algol is actually a triple system, with the eclipsing pair known as AB and the tertiary component Algol C orbiting them with a period of 1.86 years. Algol A, the hot primary, is a B8 V star and Algol B is a K subgiant, while Algol C is a late A or early F star.

Close binary evolution

Algol has played an important role in our understanding of the evolution of CLOSE BINARY STAR systems. The Algol Paradox arose when researchers discovered that the less massive secondary was more evolved than the primary. Since high-mass stars evolve more rapidly than low-mass stars, Algol presented a puzzling situation. Through the work of Hoyle, Morton and others, the paradox was resolved when they showed that the now less massive secondary had once been the more massive one. When it evolved to core hydrogen exhaustion and expanded, it reached the inner critical surface, also known as the ROCHE LOBE. Having the more massive component of a binary in contact with the Roche lobe leads to an instability that results in rapid mass transfer to the other component. The mass transfer is so extensive that it reverses the ratio of the masses, leaving the system in a semi-detached configuration (one star filling its Roche lobe while the other is smaller than the lobe).

Evolutionary studies by Sarna indicate that the system initially consisted of $2.8M_\odot$ and $2.5M_\odot$ stars. The $2.8M_\odot$ star evolved, reached its Roche lobe, transferred mass to the other star and now has a mass of $0.8M_\odot$. The initially $2.5M_\odot$ star now has a mass of $3.7M_\odot$, meaning that $0.8M_\odot$ was lost from the system.

Current work

Current work on Algol is focused primarily on studying the nature of the circumstellar matter and its interaction with the two stars. Doppler tomography and polarimetry show promise as tools to uncover the nature of the circumstellar matter.

Bibliography

Kim H 1989 BV light curve analysis of Algol *Astrophys J.* **342** 1061–7

Liou J-C and Wilson R E 1993 Quantitative modeling and impersonal fitting of Algol polarization curves *Astrophys. J.* **413** 670–9

Pen X, Shao M and Colavita M 1993 High angular resolution measurements of Algol *Astrophys. J.* **413** L129–L131

Sarna M 1993 The evolutionary status of Beta Per *Mon. Not. R. Astron. Soc.* **262** 534–42

Terrell D 1992 Demon variables *Astronomy* **20** 34–9

Dirk Terrell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Algonquin Radio Observatory

A Canadian radioastronomy observatory in Ontario. The main instrument is a 46 m (150 ft) fully steerable dish which began operation in May 1966. The telescope was operated by the National Research Council of Canada until 1989. It is now jointly supported by the Space Geodynamics Laboratory of the Center for Research in Earth and Space Technology at York University in Toronto, and the Federal Government's Geodetic Survey Division in Ottawa.

Today it is mainly used in international VLBI programs to maintain the International Celestial Reference Frame. By measuring the precise celestial locations of about 100 quasars, geophysicists are able to study tectonic movements on the Earth, and measure variables such as the planet's rates of precession, nutation and rotation.

It is still used occasionally for astrophysical research. Recent examples include imaging of supernova remnants and precise determination of pulsar rotation.

For further information see
http://www.sgl.crestech.ca/cglbi/stations/algo.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Khwarizmi [al-Khawarizmi; al'Khwarizmi], Abu Abd-Allah [Abdullah; Ja'far] Mohammad [Muhammad] ibn Musa [Bin Musa] (*c*. 800–*c*. 850)

Born at Khawarizm (Kheva), south of the Aral Sea, he flourished in Baghdad from 813 to 833. He was an astronomer and geographer but is best known as a mathematician. The word *algebra* was derived from his book *Al-Jabr wa-al-Muqabilah*. He brought into mathematics the use of zero and the rest of the Indian system of numerals (now known as 'Arabic numerals'), and developed the decimal system. His name survives in English in the word *algorithm*. The development of astronomical tables by him was a significant contribution to the science of astronomy, on which he also wrote a book derived from an earlier Sanskrit astronomical work, brought to Baghdad around AD 770. The contribution of al-Khwarizmi to geography is also outstanding, in that not only did he revise PTOLEMY's views on geography, but also corrected them in detail as well as his map of the world. His other contributions include original work related to clocks, sundials and astrolabes. His books were widely translated, and a twelfth century Latin translation of *Al Jabr* introduced this new science to the West.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Allegheny Observatory

At the Alleghany Observatory, University of Pittsburgh, the Thaw 0.76 m $f/18$ red-light refractor and Multichannel Astrometric Photometer is a major contributor of accurate parallaxes, stellar masses and the astrometic search for extra solar planets. Founded in 1869, the Observatory established the standard time service, often called Rail-Road Time.

Previous directors of note include Samuel Pierpont Langley, who began his experiments with the wing and invented the bolometer here; James Keeler, who showed the rings of Saturn to be made up of particles; John Brashear, known for his exceptional astronomical optics; and Frank Schlesinger the father of photographic astrometry.

For further information see
http://www.pitt.edu/~aobsvrty/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Allen, Clabon ['Clay'] Walter (1904–87)

Australian solar physicist, author of an influential photometric atlas of the solar spectrum, theoretical analyses of the Sun's atmosphere, and *Astrophysical Quantities*, a compendium of astronomical data.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Allen, David (1946–94)

English-born infrared observational astronomer at the Anglo-Australian Observatory, Sydney. Made and used instruments on a wide range of astronomical objects; first person to detect surface features on Venus, from infrared emission of ground beneath the clouds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Allende Meteorite

The largest known fall of a carbonaceous chondrite. On 8 February 1969 an exceptionally brilliant fireball was seen in the skies of northern Mexico, followed by a series of sonic booms. The next day a meteorite was found in the village of Pueblito de Allende. This was the first of hundreds of fragments, totalling 2 tonnes, subsequently collected from the surrounding area. The elliptical strewnfield—the name given to the area in which the fragments of a meteorite fall—measured about 7 × 50 km; the largest single fragment (110 g) was found at its far tip. It has been estimated that the parent body weighed as much as 30 tonnes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Almanac

An annual publication that tabulates daily positional and related data for the Sun, Moon, planets and other celestial bodies, and lists events and phenomena such as lunar phases, times of sunrise and sunset and eclipses. Almanacs of this kind are used primarily by astronomers and navigators.

See also: ephemerides.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alpha Centauri

The star α Centauri, also known (especially to navigators) as 'Rigel Kentaurus' or 'Rigil Kent'. With an apparent visual magnitude of -0.28 it is the third brightest star in the sky. A telescope, however, shows it to be a binary system, the components being a yellow G2V star, α Cen A, of magnitude -0.01, and an orange K1V star, α Cen B, of magnitude 1.35. Their separation is currently 19.1″ at position angle 215°; the period is 81.2 years. The absolute magnitudes of the component stars are 4.3 for α Cen A and 5.7 for α Cen B. There is also a much more distant third member of the multiple star system, Proxima Centauri, of apparent magnitude 11.01; it is separated from the main pair by 2.2° at position angle 213°. Proxima (α Cen C) is so named as it is believed to be the closest star to the Sun, at a distance of 4.22 light-years, parallax 0.772″. The parallax of the main binary pair is 0.742″, equivalent to a distance of 4.39 light-years. Proxima is a red dwarf of very low luminosity (absolute magnitude 15.4, spectral type M5Ve). It is a noted flare star, which often (but unpredictably) brightens briefly by about one magnitude.

See also: Proxima Centauri.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alpha Centauri A+B

Alpha Centauri A and Alpha Centauri B comprise the closest BINARY SYSTEM to Earth. A third star, PROXIMA CENTAURI, is thought to be in orbit about the center of mass of Alpha Centauri A and B. Other common names for this binary include Rigel Kentaurus and Toliman.

The word Centauri translates to the centaur, a mythological creature that was half human and half horse. It is believed that this constellation was named for Chiron, an unusually wise and gentle creature who was a tutor to Heracles. When Heracles accidentally wounded the immortal Chiron with a poisoned arrow, Zeus graciously allowed him to die, giving him a place among the stars.

Alpha Centauri A and B are located in the southern sky at right ascension 14 hours 39 minutes 36 seconds and declination -60°, 50 minutes and 07 seconds at a distance of 4.4 light-years. Alpha Centauri A, with spectral type G2, has apparent and absolute magnitudes of 0.0 and 4.4 respectively. Alpha Centauri B, spectral type K1, has apparent and absolute magnitudes of 1.3 and 5.7 respectively. Note that the difference between absolute and apparent magnitudes for each star is the same. This is a direct result of the fact that they are at the same distance.

The supposed third star in the system, Proxima Centauri, is much dimmer. It has a spectral type of M5 and apparent and absolute magnitudes of 11.1 and 15.5 respectively. This star is believed to be orbiting about the center of mass for Alpha Centauri A and B along a hyperbolic orbit. It has been suggested that this star system could be extended to include as many as 12 stars.

Since this binary system is the closest to Earth, and may have additional members of non-trivial influence, it provides an excellent environment for studying gravitationally bound stellar systems that orbit about a common center of mass. Accurate measurements of the size, mass and orbital parameters for these stars are necessary in order to fine-tune methods for determining the same parameters for more distant multiple stellar systems.

Robert Mohr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alpha Particle

Term that is sometimes used to describe a helium nucleus, a positively charged particle that consists of two protons and two neutrons, bound together. Alpha particles, which were discovered by Ernest Rutherford (1871–1937) in 1898, are emitted by atomic nuclei that are undergoing alpha radioactivity. During this process, an unstable heavy nucleus spontaneously emits an alpha particle and transmutes (changes) into a nucleus of a different (lighter) element. Studies of the way in which alpha particles were scattered by atoms in thin sheets of mica and gold led Rutherford, in 1910, to the discovery of the atomic nucleus.

Alpha particles are found in the solar wind and are produced in collision events in high-energy particle accelerators. An important nuclear reaction, called the triple-alpha reaction, which welds helium nuclei (alpha particles) together to form carbon nuclei, occurs in red giant stars.

See also: atom, cosmic rays, neutron, nucleosynthesis, proton, radioactivity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alpher, Ralph Asher (1921–)

Physicist, born in Washington, DC. With HANS BETHE and GEORGE GAMOW, he proposed in 1948 the 'alpha-beta-gamma' theory of the creation of the elements by successive addition of nucleons to hydrogen to build up the periodic table (the *alpha* part of the name of the theory referred to Alpher). This idea, subsequently shown by MARGARET BURBIDGE, Geoffery Burbidge, WILLIAM FOWLER and FRED HOYLE to be incorrect, became part of the 'big bang' model of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Quhi, Abu Sahl Wayjan ibn Rustam (*c*. 940–*c*. 1000)

Persian geometer and astronomer born in Tabaristan (now Mazanderan, Iran), built and directed an observatory in Shiraz to study the planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Sufi, Abu al-Rahman (903–86)

Persian astronomer, published works on astronomical instruments. Revised PTOLEMY's star catalog as the *Book of the Constellations of the Fixed Stars*, with improved magnitudes, stars named by reference to constellation figures (much copied, for example by ALFONSO X as *Libros del Saber de Astronomìa*), and nebulae, including the Andromeda Galaxy (M31). However, the work was not entirely original since not all the star positions were re-observed. Al-Sufi also mentions a 'nebulous star' north of delta Velorum, which is probably the open cluster IC 2391, and an asterism in Vulpecula (Brocchi's Cluster, Collinder 399, also known as the Coat-hanger Cluster).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Altazimuth Mounting

A telescope mounting that has axes in the horizontal and vertical planes to enable the telescope to be swivelled in altitude (perpendicular to the horizon) and azimuth (parallel to the horizon). This type of mounting is simpler, cheaper to construct, and less prone to flexure and changes of balance than the equatorial mounting. However, because (except for observers located at the north and south poles of the Earth) both the altitude and the azimuth of a celestial object change continuously as a result of its diurnal motion across the sky, the telescope must be driven around both axes at varying speeds. Because of the difficulties involved in achieving this before the development of computer control techniques, until recently the overwhelming majority of optical astronomical telescopes (apart from some small instruments and specialist amateur telescopes such as the Dobsonian) were attached to equatorial mountings. In recent years, altazimuth mountings in which the drive rates for the two axes are computer controlled have become the only way to construct large modern optical telescopes such as the two 10 m Keck telescopes (the largest individual optical telescopes), the component telescopes of the Very Large Telescope (VLT) and the 8 m Gemini instruments and are economical for all optical telescopes over, say, 2 m aperture. Altazimuth mountings are also widely used for large radio dishes.

See also: altitude, azimuth, celestial coordinates, diurnal motion, Dobsonian telescope, equatorial mounting, Nasmyth telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Altitude

The angle between the horizon and a star, measured in a direction perpendicular to the plane of the horizon. Together with a value of azimuth, this specifies the position in the sky of a celestial body at a particular instant, as seen by an observer at a particular point on the Earth's surface.

Due to the rotation of the Earth, the altitude of a star continually changes. For example, an observer in the northern hemisphere will see a star rise in the east, reach its maximum altitude when due south (culmination), and set in the west. Observers located at the north or south poles, however, will see stars move parallel to the horizon, so maintaining constant altitude.

Altitude also refers to the height of an object (e.g. an artificial satellite) above mean sea level.

See also: azimuth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Alvarez, Luis Walter (1911–88)

Physicist and astronomer, born in San Francisco, CA, professor at the University of California, Nobel prizewinner (1968) for his discoveries in particle physics. Used cosmic rays to 'x-ray' the pyramids of Egypt, finding in particular that the tombs in the Great Pyramid at Giza had no hidden rooms. Alvarez (and his son) discovered globally distributed iridium at the Cretaceous/Tertiary boundary in the Earth's geological record (which, described in terms of geological strata, is a widespread discontinuity in rock types and directions, and can be interpreted as a wholesale change in the Earth's climate). He proposed that the change of climate and the disappearance of the dinosaurs at that time, some 65 million years ago, resulted from a cataclysmic collision of an asteroid with the Earth, asteroids being relatively rich in iridium. The crater of the collision has since been identified in the Yucatan peninsula of north-eastern Mexico.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

al-Zarqali [known as Azarchiel, Azaquiel, Azachel or As Sarkali] (around 1080)

Spanish Islamic astronomer and instrument-maker (in Andalusia). Developed new theories of solar motion and edited the *Toledan Tables*, based in part on Indian methods of calculation of the positions of the planets. These tables influenced the *Alfonsine Tables* (see ALFONSO X).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

AM Herculis

AM Herculis (AM Her) is the prototype and brightest of a class of short-period CATACLYSMIC BINARIES (CB) in which a strongly magnetic WHITE DWARF accretes gas from a Roche-lobe filling sub-solar mass companion. The magnetic field is sufficiently strong ($B_1 \gtrsim 20$ MG; MG $= 10^6$ gauss) to affect qualitatively the most fundamental characteristics of the binary: preventing the formation of an accretion disk and forcing mass transfer to occur entirely through one or more gas streams; channeling flow near the white dwarf into a near-radial funnel which liberates its accretion energy in a shock above the magnetic pole; and synchronizing the rotation of the white dwarf to the orbital motion of the companion, such that all basic phenomena are orchestrated on a common period $P_{\rm spin} = P_{\rm orb}$.

Because the gas arrives at the white dwarf surface under virtual free-fall, the post-shock plasma temperature is high and AM Her-type binaries are strong x-ray emitters. Most of the current catalog of ∼50 such systems (see X-RAY BINARY STARS) results from all-sky satellite surveys for x-rays. AM Her itself was cataloged as a VARIABLE STAR a half-century before its identification with the x-ray source 3U 1809+50 attracted the attention of observers. With typical periods in the range $1.5 \lesssim P_{\rm orb} \lesssim 4$ h and stellar masses $M_1 \sim 0.6 M_\odot$, $M_2 \sim 0.2 M_\odot$, an entire AM Her-type binary would commonly fit within the disk of the Sun. The systems (figure 1) are therefore compact laboratories for studying the physics of high-temperature, magnetized PLASMAS as well as the magnetic interactions between stars.

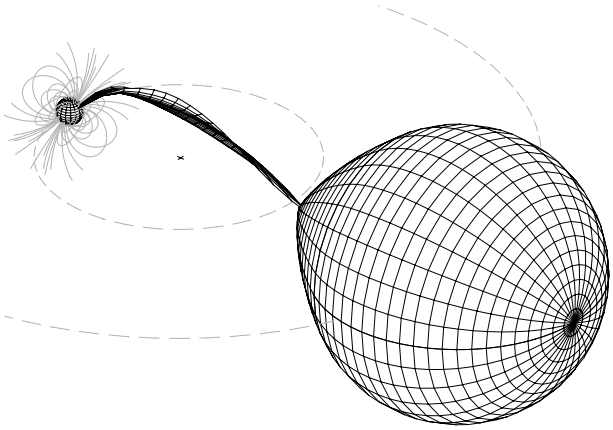


Figure 1. Sketch of the cataclysmic binary AM Her as seen from the Earth for a binary phase $\varphi \sim 0.1$. Gas flows through the inner Lagrangian point of the low-mass companion star in a largely free-falling stream onto an accretion shock located near the magnetic pole of the white dwarf. From here, the emission of x-rays and circularly polarized optical/infrared cyclotron emission dominates the light output of the system. The binary orbits in a counter-clockwise fashion around the center of mass (cross) in this figure.

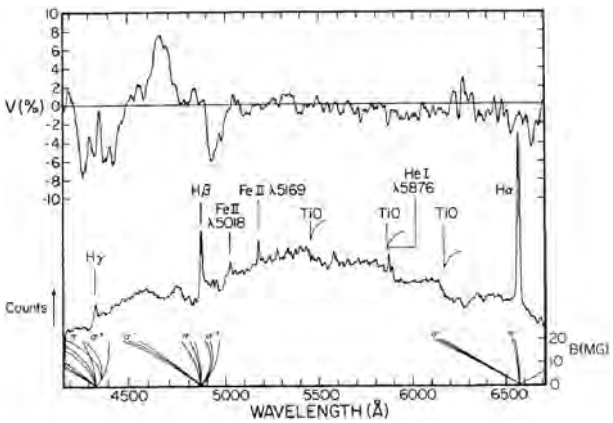


Figure 2. Total flux (bottom) and circular polarization (top) spectra of AM Her obtained during a lapse in accretion. Evident are the polarized, Zeeman-split absorption lines of Hα, Hβ, and Hγ in a photospheric field of 10–20 MG. The TiO features originate on the cool companion star.

The accretion spot

Traditionally, the defining characteristic of an AM Her-type binary has been the strong optical circular polarization (∼5%–30%) of the cyclotron emission radiated in the post-shock flow just above the white dwarf surface. Cyclotron emission is the radiation of high-speed (but non-relativistic) electrons spiraling around magnetic field lines, and one of the very few astrophysical processes which gives rise to a tendency for the electric vector of the emitted radiation to oscillate in a pattern with a net circular component, as opposed to purely linear behavior (linear polarization) or no preferential pattern (unpolarized). The property is so unique to the magnetic synchronized CBs as to earn them the designation *polars*.

The emission spectrum of an electron gyrating in a magnetic field occurs at multiples of the Larmor angular frequency

$$\omega_{\rm L} = \frac{eB}{m_e c}.$$

By analogy with isolated white dwarfs, the strong optical circular polarization of AM Her was initially interpreted to occur near the cyclotron fundamental $\omega_{\rm c} = \omega_{\rm L}$, hence to imply a magnetic field strength $B_1 \sim 200$ MG. However, spectropolarimetric observations during a lapse in ACCRETION (figure 2) revealed the Zeeman signature of a magnetic field of only 10–20 MG in the absorption lines of the hydrogen Balmer series. Since the secondary star was also evident in this low state, the Zeeman spectrum was assigned to the disk-averaged photosphere of the magnetic white dwarf and the high-state polarized light attributed to cyclotron emission in harmonic numbers $5 \lesssim m \lesssim 10$. Subsequently, individual thermally broadened cyclotron harmonics were recognized in high-state spectra, and a polar field strength of 14 MG was inferred[1].

The accretion stream impacts the white dwarf with a

[1] The field pattern is not that of a centered dipole (see table 1).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

velocity

$$v \sim \sqrt{2GM_1/R_1} \sim 3000 \text{ km s}^{-1}.$$

In addition to cyclotron radiation, the post-shock flow cools by x-ray bremsstrahlung (free-free emission) with a characteristic temperature $kT_{\rm brem} \sim$ 20–30 keV. The ∼50% of each of these components which is directed downward is absorbed and re-radiated by the white dwarf, giving rise to a third, 'blackbody', component which was predicted by the original models to have a luminosity $L_{BB} \sim L_{\rm cyc} + L_{\rm brem}$, and temperature $kT_{BB} \sim 20$ eV (in the soft x-rays). These ideas were very broadly consistent with early observations. However, a 'reversed' x-ray mode of AM Her was later observed in which the hard and soft x-rays were each strongly modulated on the binary period, but *antiphased* with each other. It was quickly realized that this mode signifies the temporary activation of gas flow onto the opposite magnetic pole. Here, for whatever reason, accretion apparently occurs primarily in the form of dense clumps each of mass $\sim 10^{16}$ g that deposit their kinetic energy deep in the atmosphere of the white dwarf, heating the surrounding photosphere from below to $\sim 10^5$ K. Direct evidence for 'blobby' accretion exists in the form of flickering in the optical and x-ray light curves even in the 'normal' x-ray mode; hence clumps must be more or less important in all flows, a trait which distorts the energy balance from that predicted by the simple model[2].

For the stellar parameters listed in table 1, the observed total high-state luminosity of AM Her, $L \sim 5 \times 10^{32}$ erg s^{-1}, implies an accretion rate

$$\dot{M} = \frac{LR_1}{GM_1} \approx 1 \times 10^{-10}\, M_\odot \text{ yr}^{-1}$$

with the low-state mass-transfer rate down by an order of magnitude or more. The emission is not likely to be uniform over the stream footpoint; a high-density core will be relatively compressed and radiate primarily in the hard x-rays, while a more rarefied sheath will have a greater stand-off shock height and favor cyclotron cooling. Because of the importance of irradiation from above and sub-photospheric heating, the soft x-rays are expected to arise from still larger regions. These ideas are confirmed by a variety of observations which reveal that the spot area as a fraction of the white dwarf surface varies from $f < 10^{-4}$ for the bremsstrahlung emission to $f \sim 0.1$ for the heated polar cap in the near-UV.

The gas stream

In a ROCHE-LOBE filling accretion binary, mass transfer occurs through the inner Lagrangian point $L1$, where the net gravitational force in the co-rotating coordinate frame is zero. Launched by their modest thermal velocity, atoms are quickly ionized by the intense x-ray/UV radiation and accelerate toward the white dwarf under the combined influence of the increasing gravitational field and Coriolis effects. The result is to divert the stream in the prograde direction. In a non-magnetic CB, the gas then loops around the white dwarf, collides with the upstream flow, and diffuses into an ACCRETION DISK surrounding the compact star. In the AM Her-type systems, the magnetic field of the white dwarf begins to dominate the kinematics of the ionized stream at roughly the Alfvén radius, given by

$$\frac{B^2}{8\pi} \sim \frac{\rho v^2}{2}.$$

This transition from ballistic to magnetically controlled regions of the flow occurs at a distance of ~ 5–$10R_1$ above the white dwarf surface. Downstream of the 'coupling region', the plasma clumps are more or less fully threaded by the rapidly increasing magnetic field and approximately follow dipolar trajectories. In general, this includes deflecting the flow out of the orbital plane (see figure 1).

Because the original stream has a non-negligible cross section and coupling occurs over a significant linear distance, different portions of the stream eventually trace different field lines to the surface and the resulting accretion footpoint better resembles an arc than a circular spot. Indeed, gas is already traveling at a speed of ∼1000 km s^{-1} when it encounters the magnetosphere of the white dwarf, so threading is unlikely to be a casual process. When AM Her is in its two-pole mode, the flow apparently branches from the coupling region into a second stream which impacts the magnetic star at a site roughly diametrically opposite the main pole. A more familiar example of the same physics is the twin auroral zones produced by the capture of solar flare particles by the Earth's magnetosphere.

The radiatively-ionized gas stream is also a copious emitter of recombination radiation. Due to the high-velocity, collimated flow, emission lines are broad and highly variable in radial velocity as well as in profile throughout the orbit. In contrast to a typical astrophysical nebula, the line flux from the stream shows $F_{H\gamma} \sim F_{H\beta} \sim F_{H\alpha}$, indicating high optical depths in the dense flow ($N_e \gtrsim 10^{13}$ cm^{-3}).

Table 1. Properties of AM Her

Apparent magnitude, m_V	12–15.8
Period, P	3.09 h
White dwarf mass, M_1	$\sim 0.45 M_\odot$
Companion star mass, M_2	$\sim 0.23 M_\odot$
Companion star spectral type	M4–5
Distance, D	90 pc
Field strength, B_1 (dipolar)	18 MG
Dipole offset, $\Delta R_1/R_1$	0.17
Inclination, i	50°
Dipole colatitude, β	55°
Dipole longitude, ψ	30°

[2] This stochastic flickering should be distinguished from the quasi-periodic oscillations (QPOs) observed in optical light and attributed to oscillations of the shock front.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The stellar components

Cyclotron opacities at high harmonic number are strongly direction dependent and polarization sensitive. For viewpoints near the local magnetic field direction, the emission is comparatively faint and circularly polarized with a sense (+/−) that reflects the field polarity. Perpendicular lines of sight measure strong linear polarization and a source enhanced by the effects of 'cyclotron beaming'. Because the accretion funnel and resulting shock are affixed to roughly radial magnetic field lines on the co-rotating white dwarf, the variations of cyclotron flux and polarization depend on the detailed geometry of the binary: the inclination i, the co-latitude of the magnetic pole β, and the longitude of the field axis relative to the stellar line of centers ψ. When phase-resolved observations of the polarized cyclotron light are combined with radial-velocity studies of the funnel emission lines and stellar absorption lines, these quantities can be more or less uniquely determined and remarkably detailed pictures of the systems constructed. Moreover, comparison of magnetic field strengths in the opposing accretion regions of two-pole systems can be used to deduce the morphology of the field pattern on the white dwarf (e.g. dipole, offset dipole, quadrupole). The entries in table 1 and the scale drawing in figure 1 result from such analyses of AM Her.

In virtually all CBs, the mass of the secondary is sufficiently low that the star's main-sequence lifetime exceeds the age of the Galaxy and it is assumed to be largely unevolved. When the star can be observed directly, such as in a low-accretion state, the measured spectral type is generally consistent with the inferred mass and radius. However, for $M_2 < M_1$, contact between the donor star and its Roche surface can be maintained in the face of mass transfer only if angular momentum is removed from the system. This implies that the orbit must decay with time—and the period decrease—at least until the secondary becomes fully degenerate. Current ideas identify the angular momentum loss mechanism with gravitational radiation plus, for the higher mass-transfer rates of objects with $P \gtrsim 3$ h, magnetic braking by a solar-type wind swept outward along the magnetic field lines of the tidally-locked magnetic companion star. It has been argued that this situation is modified when a strongly magnetic, synchronized primary star is present, since a large fraction of the field lines of the secondary would then terminate on the white dwarf. The reduced number of 'open' lines from the secondary star would reduce the effectiveness of the magnetic braking mechanism. The synchronized systems—including AM Her—do indeed appear to have accretion rates that are systematically lower than non-magnetic CBs at the same orbital period, so it is possible that the magnetic field even affects the period evolution of the binary. At the same time, it is not clear that even two decades of observation is a sufficiently long baseline to evaluate the secular mass-transfer rate of a close binary which has evolutionary timescales ranging from $\sim 10^3$–10^4 yr (for nova eruptions) to nearly 10^9 yr (for orbital period decay). The cause of the low-accretion states, which include an abrupt shut-off in the x-ray flux of AM Her by a factor 7 in just 40 min and can persist for many months, is also unknown.

Spin/orbit synchronism

The primary star of AM Her is subject to an accretion torque

$$N_{\rm acc} \sim \frac{2\pi}{P} R_{L1}^2 \dot{M} \sim 5 \times 10^{33} \text{ dyn cm}$$

where R_{L1} is the distance of the $L1$ point from the white dwarf. Left alone, this would spin up the white dwarf on a timescale $\tau \sim 10^8$ yr. Thus, the accretion torque must be balanced by an offsetting torque which maintains the spin period of the white dwarf equal to the orbital period of the companion. Because any magnetic interaction must be weaker when the component separation is larger (binary younger), it is often felt that an additional, dissipative, torque must act to slow the white dwarf spin rate and initially synchronize the motions. Among the proposed possibilities are magnetostatic interactions with either intrinsically magnetic or nonmagnetic secondary stars, and magnetohydrodynamic effects in which electrical currents are driven through the rarefied plasma medium of the binary. While it is probably not possible to be conclusive in any individual case, empirical clues as to the nature of the interaction(s) are potentially available in the distribution of dipole longitudes ψ for the equilibrium configurations and in possible drifts or slow angular oscillations of the dipole within the binaries. Real-time observations of re-synchronizing magnetic NOVA systems which have been decoupled by nova eruptions also hold promise for elucidating this unique aspect of the physics of AM Her-type systems.

Bibliography

An inkling of the excitement (and confusion) which surrounded the early work on AM Her can be gained from four successive papers in the 15 March 1977 issue of *Astrophys. J.*, including especially: Tapia S 1977 Discovery of a magnetic compact star in the AM Herculis/3U 1809+50 system *Astrophys. J.* **212** 125.

Much of the initial physics and the model for a synchronized system was presented by Stockman H S, Schmidt G D, Angel J R P, Liebert J, Tapia S and Beaver E A 1977 New observations and a slow rotator model of the x-ray binary AM Herculis *Astrophys. J.* **217** 815.

A comprehensive summary of modern research on magnetic as well as non-magnetic CBs is compiled by Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press).

Gary Schmidt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Amateur Astronomical Societies

Amateur astronomy has a long and distinguished history. For example, William Herschel, discoverer of the planet Uranus and perhaps the greatest of all observers, began his career as an amateur; in 1845 the third Earl of Rosse built what was then much the most powerful telescope in the world—the Birr Castle 72 in reflector—and used it to discover the spiral forms of the objects we now know to be galaxies. Modern amateurs carry out work which professionals have no time to do, are disinclined to do or are not equipped to do (see AMATEUR–PROFESSIONAL COLLABORATION IN ASTRONOMY). However, an observer working on his own can produce relatively little of value—and this is where amateur astronomical societies play such a major role.

It is important to note that to a considerable degree amateur and professional societies show considerable overlap. For example, in Britain the ROYAL ASTRONOMICAL SOCIETY, the main professional organization, has many amateur members, and in fairly recent times some amateurs have even served as presidents (for example, W H Steavenson, a medical doctor). The BRITISH ASTRONOMICAL ASSOCIATION (BAA) is mainly amateur but also includes many professionals, and past presidents have included several Astronomers Royal. The same situation applies in most other countries. Even the INTERNATIONAL ASTRONOMICAL UNION, the controlling body of world astronomy, has a few amateur members.

Britain has always been prominent in astronomical research, so it may be useful to give a brief account of how the various societies have developed there.

The Royal Astronomical Society, of course, dates back to the 1820s, and one of the earliest presidents was Sir William Hershel (see HERSCHEL FAMILY). There were restrictions on entry; in particular, no women were allowed to become fellows until much later. There was a clear need for a national amateur society, and in the 1880s came the Liverpool Astronomical Society, which mushroomed into an international organization. Then, suddenly, it collapsed, partly because of financial problems and partly because of poor administration. It was then that the BAA was formed; its first meeting was held in London in 1890.

From the outset, the BAA set out to be an observational society, and different sections were formed, each controlled by an experienced director. For example, one section was devoted entirely to the observation of the planet Jupiter, and regular reports were issued, both in the Association's monthly *Journal* and in the specialized *Memoirs* (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS). It is probably true to say that these publications represent the best and most continuous series of observations of Jupiter available, and the work continues today, although it is true that the space-probe results have to some extent superseded any Earth-based observational program. Other sections dealt with the Sun, the Moon, Mars, comets, variable stars and so on.

Initially, the BAA was frowned upon by the Royal Astronomical Society, but it soon became clear that there would be no conflict of interest, and for many years now the two societies have been based in the same building at Burlington House in Piccadilly. Various branches of the BAA were established, although one by one they decided to break away and become totally independent. Only the New South Wales Branch in Australia remains today.

There were no restrictions on entry, and there have been some very young members[1]. However, at one point it was felt that there could be a need a for society catering entirely for young people. A proposal for a junior section was put to the BAA Council but was not accepted. This led to the formation of the Junior Astronomical Society (JAS), and it was assumed that young enthusiasts would join the JAS and then move on to the BAA. This did happen initially, but in the 1960s the JAS decided to become a parallel organization; it is now known as the Society for Popular Astronomy.

There can be no doubt that the BAA has been an extremely successful organization, and for many years it was in a class of its own. To some extent this is still true, but it was centered mainly in London, and there was a need for local societies also. By now almost every major city and town in the British Isles has its own society, and many of these are affiliated to the BAA.

One great advantage of a local society is that it can set up an observatory. Public observatories in Britain are rare, and local societies have been invaluable in filling this gap. Some are extremely well equipped and capable of carrying out work of professional standard. Moreover, many young enthusiasts who begin by joining their nearest society go on to become professional astronomers. School societies flourish, although it is true that they depend on the presence of a least one dedicated member of staff. A few schools even have their own radio astronomy observatories (Taunton School in Somerset, for example).

There are many amateur societies in Scotland and also in Ireland. Active societies also exist in the Channel Islands and the Isle of Man. All these various societies are linked and work together to their mutual benefit. Moreover, many of them publish their own regular journals, and their activities and results are reported in national periodicals such as *Modern Astronomer* and *Astronomy Now*.

The situation in the United States is much better from the public observatory point of view, and amateur societies are very numerous indeed. Of particular importance is the AMERICAN ASSOCIATION OF VARIABLE STAR OBSERVERS (AAVSO), which is the largest organization of its kind anywhere in the world and produces a vast number of important observations. Predictably it is made up of both amateurs and professionals, but for sheer visual work it is the amateurs who are the most prolific, especially now that they can use electronic devices such as CCDs and

[1] If I may add a personal note: I was elected in 1934, at the age of 11, when the President was the then Astronomer Royal, Sir Harold Spencer Jones. I became President exactly 50 years later.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can produce results which will satisfy any professional researcher. To list all the other major societies would take many pages, but particular mention should be made of the ASTRONOMICAL SOCIETY OF THE PACIFIC, which has many amateur members.

Canada has two main societies: the Canadian Astronomical Society (mainly professional) and the ROYAL ASTRONOMICAL SOCIETY OF CANADA (part professional, part amateur). The same sort of situation applies in Australia and in New Zealand. Amateur societies flourish in almost all European countries and in Russia. Japan, of course, has a great tradition of amateur observational work, and here too there are both national and local organizations.

Naturally, all societies—national and local—have their regular meetings, and these are invaluable, because it is often possible to invite well-known speakers who will be able to give advice and encouragement.

Financial considerations are always very much to the fore. Subscription rates are kept as low as possible, but bills have to be paid, particularly if premises of any kind are involved. Local authorities will often help, though others are much less cooperative (in Britain, one society was recently forced to abandon its observatory because the council decided to increase the ground rent to an impossible amount). However, even a society which has no headquarters or observatory will almost certainly have members who possess telescopes of their own and will make them available to newcomers. By now there are also some ambitious societies which are setting up their own planetaria—and surely a planetarium is the best possible educational aid. There are plenty of these in the United States and a few in Britain, notably the planetarium now being erected near Chichester, in Sussex, by the South Downs Astronomical Society.

There has been one setback in recent years. It was proposed to form an international amateur organization to correlate the work of societies all over the world, and during the 1960s, at a General Assembly of the International Astronomical Union, such an organization was actually set up: the International Union of Amateur Astronomers (IUAA). It has issued some publications and has held several meetings in different countries, but it has to be admitted that it has never been a real success, again for reasons of financial constraint. In the future it is hoped that it will be extended and that it will play a really useful role, but it is difficult to organize, and much remains to be done.

Meanwhile, amateur astronomical societies exist in virtually all countries. Their value cannot be overestimated, and today, in the Space Age, their role is even more important than it has been in past years.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Amateur–Professional Collaboration in Astronomy

It is probably true to say that, in most branches of science, the day of the amateur is over. No modern amateur can expect to undertake important research work in physics, for example, where elaborate and expensive equipment is essential; neither can he expect to make a major theoretical discovery, because a very rigorous training is needed. True, there is still scope for some good amateur work in geology and paleontology, but it is with astronomy that the amateur really comes into his own.

A century ago, there were various branches of astronomy in which the amateur reigned supreme. The best observations of the Moon and the surfaces of the planets were made by amateurs, and this was still true until not long before the Second World War. It is also worth remembering that in 1845 the 3RD EARL OF ROSSE, with no expert help, built what was then much the most powerful telescope ever constructed—the 72 in reflector at Birr Castle, in Ireland—and used it to discover the forms of the objects then called spiral nebulae, and now known to be external galaxies. (It is pleasing to note that, after a long period of neglect, the Birr telescope has now been restored to full use, even though it is of course by no means the equal of a modern telescope of the same aperture.)

Around 1900, photography took over from the eye for most branches of astronomy. Today photography is itself being superseded by electronic aids, such as CHARGE-COUPLED DEVICES (CCDs), which are far more sensitive. No longer does the astronomer have to spend hours in a cold, dark dome, making sure that his telescope is still firmly fixed on its target object; he need not be in the dome at all, or even in the same country. Everything comes through on a television screen. This has obvious advantages, quite apart from comfort, but it does mean that the professional astronomer seldom or never actually looks through a telescope, and he has no need to have a practical knowledge of the sky. It is almost certainly true to say that few professional astronomers can recognize more than half a dozen constellations[1]. Some amateurs, on the other hand, have a truly encyclopedic knowledge of the sky and can identify many hundreds of stars with no trouble at all.

Amateur equipment has changed markedly in the recent past. Fifty years ago the average amateur used a telescope of moderate aperture—say up to 18 in for a REFLECTOR—and was more concerned with visual work than with photography. Today, the well-equipped amateur has become skilled in electronics. Using a CCD with, say, a 15 in telescope, a modern amateur can produce results which would match those of major professional observatories in the immediate post-war era.

[1] I well remember one well-known professional astronomer telephoning me to say that he had just found a nova, and was about to notify the Royal Observatory. In fact, he had made a completely independent discovery of the planet Saturn.

There are, of course, some amateurs who do not use telescopes at all. Naked-eye work can be useful, particularly with regard to meteors and aurorae (see the article on NAKED-EYE ASTRONOMY). However, binoculars can be very valuable indeed (see BINOCULAR ASTRONOMY), notably in the hands of those amateurs who spend their time in searching for comets and novae. George Alcock, an English schoolmaster, has now discovered four comets and three novae, all with the aid of powerful binoculars. He 'sweeps' the sky, and his knowledge of the star patterns is so good that he can recognize a newcomer at a glance. He is not alone in this; there have been many recent comet discoveries by amateurs—including the two most recent 'great' comets, Hyakutake of 1996 and Hale–Bopp of 1997.

One of the outstanding amateurs of the present day is an Australian clergyman, the Rev. Robert Evans. He uses a portable telescope and scans external galaxies to search for supernovae. So far he has found more than 20. A supernova is a colossal stellar outburst, marking the death of a star; in our own Galaxy no naked-eye supernovae have been seen since the year 1604, but they do occur in the outer systems, many millions of light-years away. Nobody can tell when a supernova will burst forth, and it is important to obtain observations as early as possible, so that the advance warnings given by amateurs such as Evans are of vital significance. Searches of this kind are now being carried on by many amateurs, with great success. Some observers, such as Evans, work visually; others use photography or CCD equipment. It is hardly necessary to add that contributions of this kind are very warmly welcomed by professional astronomers.

Ordinary novae can also be identified by amateurs, with their knowledge of the sky, and in recent years several minor planets or asteroids have also been found in the same way. Clearly it is essential to give prompt notification of a discovery. There are centers in many countries, but every confirmed discovery is announced by the telegram service of the INTERNATIONAL ASTRONOMICAL UNION (IAU), which has world-wide coverage. It is also true, of course, that many alleged 'discoveries' turn out to be false alarms, and every effort is made to filter these out before they waste valuable observing time.

Variable-star work has always been a major part of amateur observational astronomy. VARIABLE STARS are of many kinds. Some, such as the Cepheids, are as regular as clockwork; others, such as the Mira stars, differ in behavior from one cycle to another; yet others are completely irregular, so that one can never tell what they will do next. There are so many variable stars that professional astronomers cannot hope to keep track of them all, and great reliance is placed on amateur work. Much of this is purely visual. The procedure is to compare the variable with nearby stars which do not fluctuate. With a little practice it is not difficult to make estimates correct to one-tenth of a magnitude, and for most purposes this is good enough. Using photometers, it is possible to obtain estimates which are far more accurate that this, and it is true that most modern amateurs who are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

seriously concerned with variable-star research do have photometric equipment. A really large telescope is not essential; a reflector of, say, 6–12 in aperture will give its owner more work than he can possible handle. The results are correlated by organizations such as the Variable Star Section of the BRITISH ASTRONOMICAL ASSOCIATION and the AMERICAN ASSOCIATION OF VARIABLE STAR OBSERVERS.

There are some variables which are within naked-eye range; such is Gamma Cassiopeiae, in the W of Cassiopeia, which is an unstable 'shell' star. It is usually just below the second magnitude, but in the late 1930s brightened up abruptly by more than half a magnitude, subsequently sinking to below magnitude 3 before returning to its normal state. The early observations of the 1930s outburst were amateur, and one never knows when the same will happen again. Many dozens of variables can be followed with binoculars, and accurate light curves can be drawn up.

So far as the Moon is concerned, amateur work led the way for many years. Before the possibility of space missions, most professionals were concerned only with lunar movements, and amateur maps were the best of the time. There was also the chance of making really interesting discoveries, such as formations in the libration areas, near the edge of the Earth-turned hemisphere. When the Russians sent up their first circumlunar probe, Lunik 3, in 1959, they obtained photographs of the far side, which is never visible from Earth because it is always turned away from us; amateur maps were used to correlate the Lunik 3 results with the Earth-turned regions. However, the situation has changed now. Space missions have enabled us to obtain detailed charts of the entire Moon, and amateur cartography has lost its value, but there is still work to be done. Although the Moon is more or less inert, there are occasional very brief, obscure outbreaks which are known as TLP or transient lunar phenomena. Presumably these are due to mild release of gases from below the crust, and they have been seen so often, and by so many reliable observers (including a few professionals), that it is difficult to question their reality. Confirmation of their reality was obtained in 1999 by A Dolfus using visual and polarimetric measurements. Dedicated amateurs are studying them, and the results are of great interest.

Occultations of stars by the Moon are still useful, and amateurs are expert at timing them. There are also occultations of star by asteroids, and because the tracks are so narrow the amateur has a great advantage; he can take a suitable telescope to just the right place at just the right time, which the professional cannot do. Important measurements of asteroid positions and dimensions have been made in this way.

Amateur solar work is also becoming of increasing importance. Casual tracking of sunspots is interesting, but for real solar research a considerable amount of complex equipment is needed. H-alpha filters are now common, and there are a few excellent spectrohelioscopes; indeed one leading British amateur solar researcher, Commander Henry Hatfield, has gone so far as to build his entire house round his spectrohelioscope. Amateur observations of flares and associated phenomena are used to supplement the work carried out at major professional observations.

Now that we have the amazingly detailed views of the planets from the various space missions, it might be thought that amateur planetary observation has lost all its value. This is most emphatically not the case. For example, at the present time (2000) Saturn is not being monitored by any probe. Major outbreaks do occur there from time to time, and many of these are first seen by amateurs; for example, the brilliant white spot of 1933 was discovered by an amateur, W T HAY, probably best remembered by the general public as Will Hay, the stage and screen comedian. The latest white spot was also first seen by an amateur, Wilber; only later was it imaged by the Hubble Space Telescope. It is very likely that the next violent outbreak on Saturn, whenever it occurs, will also be an amateur discovery, and early notification means that professional equipment can be used as soon as possible. Jupiter is another favorite target for amateurs, because the surface features are always changing; on the Giant Planet there is never any lack of something new to see.

For Jupiter, and perhaps Saturn, a modest telescope can be used with success. Mars is much more of a problem, because a high magnification is necessary, but when the planet is near opposition the amateur with, say, a 12 in or 15 in telescope can help in monitoring cloud phenomena and the behavior of the polar caps—particularly if he has access to electronic equipment.

Even with remote Pluto there is scope. Very occasionally Pluto occults a star, and when this last happened amateur measurements were very valuable; pre-occultation fading of the star confirmed that Pluto does have an extensive if highly tenuous atmosphere.

In astronomy, as in all other sciences, cooperation is of paramount important. There is no point in making an observation unless other workers hear about it. The controlling body of world astronomy is the IAU, which really is international—and remained so even during the worst part of the Cold War. There are some amateur members of the IAU, and there have even been amateur officials. General Assemblies are held every 3 years, in different countries; that of 2000 will be in England (Manchester), following 1997 (Holland) and 1994 (Argentina). During the 1960s an attempt was made to form an analogous amateur organization, the International Union of Amateur Astronomers. It does exist, but it would be idle to pretend that it has been as successful as had been hoped, and better organization is needed. Meanwhile, all countries have their national societies—in Britain, for example, the British Astronomical Association, which is mainly amateur and has an observational record second to none. Most major towns and cities of the world have their own societies, and there is a pleasing amount of cooperation, both national and international (see also AMATEUR ASTRONOMICAL SOCIETIES).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

All in all, it is fair to say that amateur astronomers undertake work which the professional does not wish to do, does not have the time to do or genuinely cannot do. It is also true that amateur work is valued and warmly welcomed and that the amateur astronomer is every bit as important now as he has been in the past.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ambartsumian [Ambarzumjan], Viktor Amazaspovich (1908–96)

Astrophysicist, born in Armenia, professor at the University of Leningrad and at Erevan, founded the Byurakan Astronomical Observatory. He applied an original and creative scientific imagination to a range of observational phenomena in astronomy. He suggested that T Tauri stars are very young and that loose stellar associations are dissociatiating. He made the connection between the various energetic phenomena in galaxies (jets, high-speed motions, bright ultraviolet luminosities, etc) as pointing to energetic sources at the centers of galaxies, now identified as massive black holes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

American Association of Variable Star Observers

The largest organization of variable star observers in the world, with over 1000 members in more than 40 countries. Its database contains over 9 million observations of variable stars. The organization's headquarters are in Cambridge, Massachusetts.

It was founded in 1911 by amateur observer William Tyler Olcott. Inspired by a talk at the 1909 meeting of the American Association for the Advancement of Science, Olcott began to send his observations of variations in stellar brightness to Edward Pickering at Harvard College Observatory.

For further information see http://www.aavso.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

American Astronomical Society (AAS)

Founded in 1899, the AAS is a non-profit scientific society created to promote the advancement of astronomy and closely related branches of science. Its membership consists primarily of professional researchers in the astronomical sciences, but also includes educators, students and others interested in the advancement of astronomical research. About 85% of the membership is drawn from North America. The AAS operates in five major areas: Publications, Meetings, Education, Employment and Public Policy. Its headquarters are in Washington, DC.

The AAS research journals, *The Astrophysical Journal* and *The Astronomical Journal*, appear in both print and electronic formats. *The Bulletin of the American Astronomical Society* reports the latest institutional developments and documents the content of the meetings held by the AAS and its five Divisions. AAS meetings, held twice a year, provide a forum for the presentation of scientific results.

For further information see
http://www.aas.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anaxagoras of Clazomenae (*c*. 500–*c*. 428 BC)

Born in Clazomenae, now Turkey, an Athenian philosopher who correctly explained the phases of the Moon and eclipses of the Moon and the Sun. He believed, like the atomists, that the heavens and the Earth were brought into existence by the same processes and composed of the same particles, but also recognized that further agencies were required to create their motions, identifying that agency as 'mind' (*nous*). He was impeached for impiety, for claiming that the Sun was a red-hot stone, ignited by its motion, and larger than the Peloponnesus, but not a god, and that the Moon was Earth that reflected the Sun's light. He was banished to Ionia.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anaximander of Miletus (*c*. 611–*c*. 547 BC)

Greek philosopher, born in Miletus, pupil of THALES. Believed that the Earth was a cylinder with a diameter three times its height, unsupported, at the center of the universe. We live on the top end of the cylinder. His theory of the universe was that objects are formed from a vortex process by which light objects were flung out to their periphery. This separated hot and cold, moist and dry, and Earth and ether. He thought the Moon and Sun and stars were self-luminous wheels of fire. He is credited with producing the first map of the Earth, and with being the first to speculate on the relative distances of the heavenly bodies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anaximenes of Miletus (*c*. 585–*c*. 525 BC)

Greek philosopher, born in Miletus, pupil of ANAXIMANDER. The first Greek to distinguish clearly between planets and stars, Anaximenes conceived the idea that the stars were fixed like nails on a crystalline sphere, which he apparently believed was a hemisphere and rotated like a hat on a head. He believed the primary substance of the universe was air, which could form the other elements of water, earth and fire by compression and rarefaction. For example, the Earth was made from dense air, which, vaporized, made the Sun, Moon and planets. Credited with the invention of the gnomon, the central shadow-casting stem of a sundial; he certainly appears to have had an interest in sundials and exhibited at Sparta the sundial called Sciotherion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anderson, Carl David (1905–91)

Physicist, born in New York City, NY, Nobel prizewinner (1936, shared with VICTOR HESS) for his discovery of the positron (anti-electron) while studying cosmic rays at the California Institute of Technology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Andromeda

(abbrev. And, gen. Andromedae; area 722 sq. deg.) A northern constellation that lies between Perseus and Pegasus, and culminates at midnight in early October. It is named after the daughter of King Cepheus and Queen Cassiopeia in Greek mythology, who was rescued by Perseus from being sacrificed to the sea monster Cetus, and is usually shown on early celestial charts as a chained maiden. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A rather inconspicuous constellation, the brightest stars in Andromeda are α Andromedae (Alpheratz or Sirrah), magnitude 2.1, β Andromedae (Mirach), also magnitude 2.1, and γ Andromedae (Almach), a fine orange (K3) and bluish-white (B8) double star, magnitudes 2.3 and 5.0, separation $10''$. All other stars are of the third magnitude or fainter. α Andromedae, which marks the north-east corner of the Square of Pegasus, was originally designated δ Pegasi. α, β and γ Andromedae form a slightly curved line of stars, pointing diagonally away from the Square.

To the north-west of β Andromedae are two fourth-magnitude stars, μ and ν Andromedae, which are at almost a right angle to the line joining α, β and γ. About 1° west of ν lies the Andromeda Galaxy (M31, NGC 224), the largest spiral galaxy in the Local Group, which, at a distance of 2.7 million light-years, is just visible to the unaided eye as a faint misty patch.

Other interesting objects in Andromeda include the Mira-type variable stars R Andromedae (range 5.8–14.9, period about 409 days) and W Andromedae (range 6.7–14.6, period about 396 days), NGC 752, an open cluster of several dozen stars fainter than eighth magnitude, and NGC 7662, a ninth-magnitude planetary nebula.

See also: Andromeda Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Andromeda Galaxy

The Andromeda galaxy is the closest SPIRAL GALAXY to the MILKY WAY, just visible to the naked eye on a dark night as a faint smudge of light in the constellation Andromeda. The earliest records of the Andromeda nebula, as it is still often referred to, date back to AD 964, to the 'Book of the Fixed Stars' published by the Persian astronomer AL-SÛFI. The first European to officially note the Andromeda nebula was Simon Mayer, a German mathematician and astronomer, who was not aware of Al-Sûfi's earlier work. The object is number 31 in the famous MESSIER CATALOG, which dates back to the 18th century, and is therefore often referred to as M31. Another alias for the Andromeda galaxy is NGC 224, since it is number 224 in the NEW GENERAL CATALOG (NGC) compiled by DREYER in 1888. Modern measurements based on Cepheid variables place M31 at a distance of 740 kpc or about 2.4 million light-years. This makes it the closest spiral galaxy to the Milky Way. It is the dominant member of the LOCAL GROUP, and as such it has and continues to fill an important role in studies of galaxy structure, evolution and dynamics, stellar populations, star formation and interstellar medium. Crucial historical developments include HUBBLE's work in the 1920s on the distance to M31, which proved that galaxies outside our own Milky Way exist, and BAADE's work in the 1940s on stellar populations which led to the concepts of old (population II) and young (population I) stars. Many small companion galaxies believed to be associated with M31 have been identified, some only in the past year. They are M32 (low-luminosity elliptical), NGC 147, 185 and 205 (dwarf ellipticals), IC 10 (dwarf irregular), LGS 3 (transition object between dwarf irregular and dwarf spheroidal) and And I, II, III, V and VI (dwarf spheroidals). It is likely that more faint companions remain to be discovered. Figure 1 shows an optical image of the Andromeda galaxy.

Figure 1. Optical image of the Andromeda galaxy and two of its nearby elliptical companions, M32 (below center) and NGC205 (upper right corner). The image measures 100′ on a side, corresponding to 21.5 kpc. Based on photographic data obtained using the Oschin Schmidt Telescope on Palomar Mountain. The Palomar Observatory Sky Survey was funded by the National Geographic Society. The Oschin Schmidt Telescope is operated by the California Institute of Technology and Palomar Observatory. The plates were processed into the present compressed digital format with their permission. The Digitized Sky Survey was produced at the Space Telescope Science Institute (ST ScI) under US Government grant NAG W-2166. Copyright © 1994, Association of Universities for Research in Astronomy, Inc. All rights reserved.

Morphology, mass and stellar content

The Andromeda galaxy is a large, early-type spiral of HUBBLE CLASSIFICATION Sb, luminosity class I–II. It has a prominent central bulge of stars with an effective radius of 2 kpc, which extends smoothly into an extended flattened spheroidal component of old stars. As in most spiral galaxies, in addition to the bulge or spheroidal component, there is a thin disk of stars and gas with superposed spiral arm structure. The combined light from the disk stars declines exponentially with distance from the center, as is common for spiral galaxies. The disk scale length is about 5–6 kpc, and M31 is some 50% larger than our Milky Way Galaxy and about twice as luminous. The disk becomes bluer at larger radial distance from the center, and this could be caused by a relatively larger fraction of younger stars there, or a lower abundance of heavy elements. Abundance determinations of oxygen, nitrogen and sulfur in the interstellar medium of M31 do indeed show an overall decrease in these elements compared with hydrogen with increasing radial distance.

The SPIRAL ARM structure of M31 has been difficult to determine, owing to its orientation on the sky. Spiral arm segments show up clearly in the distributions of young massive stars and various interstellar medium tracers such as ionized, neutral and molecular hydrogen gas, and dust, but it has not been possible to link the spiral arm segments into an unambiguous grand design spiral structure. A further complication is that the interaction with the close small elliptical companion M32 seems to have distorted the spiral arms as well. The outer stellar and gaseous disks, beyond about 20 kpc, bend out of the principal plane, producing a phenomenon known as warping. Given its bulge size and luminosity, there seems little doubt that M31 is an earlier-type spiral than our Milky Way. For example, the total number of GLOBULAR CLUSTERS, which has been shown to be related to the bulge or spheroid luminosity, is probably between 400 and 500 in M31, almost a factor 3 larger than for the Milky Way. The velocity dispersion of the stars in the bulge of M31 is about 155 km s^{-1}, while for the Milky Way this is 130 km s^{-1}.

From its H I rotational velocity out to 30 kpc radius, one can infer that M31 has a mass of at least 3×10^{11} solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

masses, but the total mass will probably be much higher depending on the unknown radial extent of a DARK MATTER halo, a common component in spiral galaxies inferred from the high rotational velocities of gas clouds in the far outer disks. The timing argument for the Local Group implies a mass of a few times 10^{12} solar masses for M31. Given these figures, dark matter must strongly dominate over the mass implied by the visible stars and gas.

The stellar populations in M31 are discussed in the article M31: THE OLD STELLAR POPULATIONS, so a brief summary will suffice here. The bulge and spheroid population is old, although the considerable spread in metal abundances of the stars in the bulge and also the outer halo implies a more complicated formation scenario than a very rapid initial collapse. There is no current star formation in the inner bulge, contrary to the situation in the Milky Way. The properties of the globular clusters of M31 are similar in many ways to those of the Milky Way, i.e. the globular clusters are old, show a range of metal abundances and the same overall correlations between velocity dispersion and luminosity, central and average surface brightness as the Galactic globular clusters. A small number of young massive clusters is found in the disk, at about 10 kpc from the center. Such clusters are common in late-type galaxies, such as M33 and the Magellanic Clouds, but apparently rare in early-type spirals. These could be the progenitors to (low-mass) globular clusters. Interestingly, these objects are mostly concentrated in the tidally disturbed SW spiral arm, and they may be similar to young blue clusters observed in interacting galaxies.

One of the advantages of the close distance of M31 to us is that it enables unique experiments. One of these is the recent attempt to detect MICROLENSING towards M31 bulge stars to constrain the nature of the dark matter. While M31 is too distant to resolve most of its light in individual stars, microlensing can in principle still be detected as a brightening of individual pixels on the detector, if one star of the many contributing to the brightness of that pixel is being lensed by a dark object in M31's halo. Several of these studies are now underway, thereby extending the ongoing microlensing observations of the Galactic bulge and of the Magellanic Clouds.

The central region

The Andromeda galaxy does not contain a particularly active nucleus at its center, but there does appear to be a BLACK HOLE of about 3×10^7 solar masses there. The evidence for this comes from measurements of the stellar velocities close to the center both with the HUBBLE SPACE TELESCOPE and from the ground. The morphology of the central region is quite surprising, with two peaks in the brightness distribution separated by 0.5 arcsec. The currently favored interpretation is that this apparent double nucleus is actually caused by a single eccentric disk of stars orbiting the black hole. The brighter of the two 'nuclei' would then be caused by stars piling up near the apocenter of the eccentric disk, where their speeds are low, while the fainter 'nucleus' would be due to the increase in density towards the center.

The lack of strong nuclear activity associated with the black hole could be due to the general lack of interstellar medium in the M31 bulge. No H I or CO has been detected within 1 kpc from the center, although there are some dust clouds seen in projection against the bright stellar continuum. A faint diffuse 'spiral disk' is seen in ionized gas, with a radial extent of about 500 pc, but it only contains a small amount of mass, perhaps as little as 1500 solar masses total. This gas possibly originates from mass lost by low-mass evolving stars in the form of stellar winds. The lack of activity also manifests itself in the weak level of radio continuum emission, about a factor of 10 less than for the Galactic center.

Star formation and interstellar medium

As the closest nearby spiral, M31 has been a prime target for exploring the properties of various gaseous phases of the interstellar medium and massive stars, and the connections between them. The spatial distribution of all these components, be they neutral or ionized gas, dust or OB stars, is very similar: there is a general lack of gas and young stars in the inner several kpc, some gas and star formation in spiral arms 3–5 kpc from the center, strong concentrations in an annulus between 8 and 12 kpc and modest concentrations in arms out to 15 kpc. Beyond this, the H I disk continues to show spiral structure out to 25 kpc or so, with weak ongoing STAR FORMATION in selected regions. Star formation rates can be estimated from the distribution of ionized gas (H II regions and diffuse ionized gas), which requires hot, luminous, hence massive and young, stars to ionize it, far-infrared emission radio continuum emission and UV light. In addition, the proximity of M31 allows detection of individual OB stars and of evolved tracers of young massive stars such as WOLF–RAYET STARS and LUMINOUS BLUE VARIABLES. All these tracers indicate that the overall level of star formation activity in M31 is quite low, about $\approx$ 0.3–0.5 solar masses yr^{-1}, compared with the Milky Way and other spirals, such as M33. In addition, M31's largest star-forming complexes appear modest compared with those found in other galaxies.

The mass of atomic hydrogen gas is about 4×10^9 solar masses, while that of molecular hydrogen is more than a factor of 10 lower; the uncertainty in the conversion factor of the CO line flux to molecular gas mass makes the latter estimate rather inaccurate, but the low molecular hydrogen surface density is consistent with the overall low rate of star formation in M31. There is some evidence for the presence of very cold molecular gas in M31, which might have further implications for the total molecular gas mass. M31 was one of the first external spirals where a widespread distribution of ionized gas outside of traditional H II regions was detected. This diffuse ionized gas, also referred to as warm ionized medium, is the dominant component, in terms of mass and spatial extent, of ionized gas in galaxies. It requires a large amount of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

energy to remain ionized, and its widespread distribution implies that the morphology of the interstellar medium has to be very porous, or else the ultraviolet photons from OB stars could not travel far from their origins before being absorbed.

Information on COSMIC RAYS and the magnetic field distribution in M31 has been obtained from detailed radio continuum observations at several frequencies. The radio continuum is strongly dominated by non-thermal synchrotron emission, stemming from highly relativistic electrons spiraling in a magnetic field. Information on field strength and direction can be inferred from the intensity of the emission, the polarization vectors and the rotation measures observed towards extragalactic background sources. The large-scale magnetic field extends from 5 kpc interior to the star formation annulus to as far as 25 kpc from the center. The main field organization is toroidal, aligned along the annulus, with some evidence for a poloidal component extending into the halo. The structure of the field appears to be that expected for an even-mode DYNAMO.

Outlook for the future

There is no question that the Andromeda galaxy will continue to play a central role in future investigations of the morphology and evolution of galaxies, just as it has in the past. With the next generation of telescopes, on the ground and in space, we can expect new results in many areas, such as microlensing studies, high-resolution maps of all the interstellar medium components, abundance measurements in individual stars and further unraveling of the star formation history of M31's disk and bulge.

Bibliography

An extensive overview of the characteristics of the Andromeda galaxy, providing references to all relevant papers from before 1992, is given in the monograph

Hodge P W 1992 *The Andromeda Galaxy* (Dordrecht: Kluwer)

Photographs indicating the positions of open clusters, globular clusters, OB associations and dust clouds in the Andromeda Galaxy are presented in

Hodge P W 1981 *Atlas of the Andromeda Galaxy* (Seattle, WA: University of Washington Press)

A good impression of the morphology of H II regions and diffuse ionized gas can be obtained from

Walterbos R A M and Braun R 1992 *Astron. Astrophys. Suppl.* **92** 625

A complete CO map is presented by

Dame T M *et al* 1993 *Astrophys. J.* **418** 730

The most recent information on the magnetic field is discussed by

Han J L *et al* 1998 *Astron. Astrophys.* **335** 1117

The properties of the globular cluster system are discussed in

Djorgovski S G *et al* 1997 *Astrophys. J.* **474** L19

while the most recent compilation of globular clusters is presented by

Barmby P *et al* 2000 *Astron. J.* **119** 727

The properties of the central black hole and stellar dynamics are discussed by

Kormendy J *et al* 1999 *Astrophys. J.* **522** 772
Statler T S *et al* 1999 *Astron. J.* **117** 894

René Walterbos

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anglo-Australian Observatory

The Anglo-Australian Observatory (AAO) operates two optical telescopes at Siding Spring Mountain, the 3.9 m Anglo-Australian telescope and the (1.2 m) 47 in UK Schmidt Telescope. These telescopes are located beside the spectacular Warrumbungle National Park outside Coonabarabran in New South Wales, 450 km northwest of Sydney.

The Anglo-Australian Telescope (AAT) was established after representations had been made to the Australian and British governments in the mid-1960s. The agreement was secured in an act of the Australian Parliament, The Anglo-Australian Telescope Agreement Act, 1970. Each government contributes equally in return for equal use of the telescopes. The AAT was inaugurated in 1974.

The UK Schmidt Telescope (UKST) is a special purpose survey telescope. Its initial task was the first detailed photographic survey in blue light of the southern skies. Other major projects have since been undertaken, and more are in progress. The UKST recently began an Hα survey of the southern Milky Way and Magellanic Clouds.

The AAO spends 15% of its budget each year on new instruments and associated software and detectors.

Over the past decade, the AAO has pioneered the use of optical fibres in astronomy. The recently completed 2dF uses optical fibers to enable 400 objects to be analysed at once. The 2dF is being used in a very ambitious project to map a very large volume of the universe. The 2dF Redshift Survey aims to measure the redshifts of 250 000 galaxies over two years. In parallel, there is a similar redshift survey of 30 000 quasars.

The AAO has recently begun building a fiber positioner similar to the 2dF for the EUROPEAN SOUTHERN OBSERVATORY's Very Large Telescope (VLT).

David Malin, an astronomer at the AAO, has developed techniques to make astronomical color photographs from plates taken in three separate colors. These beautiful images have earned recognition as among the finest in the world.

For further information see
http://www.aao.gov.au.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Angstrom

A unit of length frequently used to describe the wavelength of light. It is named after the Swedish physicist Anders Jonas Ångström (1814–74), a founder of spectroscopy. It is equal to 10^{-10} meters (i.e. one ten thousand millionth of a meter) and is denoted by the symbol Å. Most of the visible spectrum lies in the wavelength range 3900 to 7500 Å.

See also: light, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Angular Momentum

A property of rotating bodies or systems of masses which is dependent upon the distribution and velocities of the masses about the axis of rotation. The angular momentum of a single particle of mass, m, moving in a circular orbit of radius, R, at velocity, v, is given by mvR. Angular momentum is a conserved quantity, in other words, the total angular momentum of a system is constant. Thus if a large slowly rotating gas cloud contracts, its rotational velocity must increase to conserve angular momentum; as it contracts, therefore, it spins more rapidly. Appropriate SI units for angular momentum are kilogram meters squared per second (kg m^2 s^{-1}). The conservation of angular momentum is of central importance to such questions as star formation, the origin of planetary systems, etc.

See also: meridian, oblateness.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anomalous Cosmic Rays

The anomalous cosmic rays are an ionic component of the cosmic-ray flux which is observed, in the heliosphere, to be significant at energies ranging from 20 MeV to more than 500 MeV. They consist of a number of ion species ranging from hydrogen to beyond neon, but anomalous carbon is significantly underabundant. The highest kinetic energy reported for this component is more than 1.6 GeV. Observations from detectors on board heliospheric spacecraft show that their intensity continues to increase out to the maximum distance reached by current spacecraft, indicating that they are accelerated in the outer solar system. They were first seen in the early 1970s and were initially called anomalous cosmic rays because of their unusual composition and because their spectrum was unlike that of the already well-known galactic or solar cosmic rays. This name has stuck, even though we now know much more about the nature and origin of these particles.

In 1974 it was suggested that the anomalous component was the result of the acceleration, in the heliosphere, of recently ionized interstellar neutral atoms (see also COSMIC RAYS: PROPAGATION IN THE HELIOSPHERE). The neutral atoms stream into the solar system from the interstellar medium and are then ionized by solar ultraviolet radiation or by charge exchange with the SOLAR WIND. This hypothesis explained very nicely the unusual composition of the anomalous component, since only initially neutral interstellar particles can enter the heliosphere to participate in the process. For example, carbon ions, which are nearly fully ionized in the interstellar medium and hence are mostly prevented by the interplanetary magnetic field from entering the solar system, are notably depleted in the anomalous cosmic rays. More-recent observations have established that the bulk of the anomalous cosmic rays are indeed singly charged, as expected in this picture.

Later, in 1981, it was pointed out that many features of the anomalous component could be explained if the major part of the acceleration of the newly ionized particles occurs at the termination shock of the solar wind, by the mechanism of diffusive shock acceleration. Subsequent sophisticated numerical simulations carried out in the mid-1980s clearly showed that energetic, singly charged particles could be accelerated very efficiently at the termination shock. The essential observed features of the spectrum and spatial gradients could be explained very naturally in terms of this picture. This picture of acceleration of newly ionized interstellar neutrals and subsequent acceleration to high energies at the termination shock of the solar wind is now the standard paradigm (see also SOLAR WIND: ENERGETIC PARTICLES). This paradigm is illustrated schematically in figure 1, which shows the incoming neutral atoms, their ionization and the subsequent acceleration at the termination shock of the solar wind. A typical observed anomalous cosmic-ray spectrum (with the solar and galactic contributions removed) is shown in figure 2(*a*).

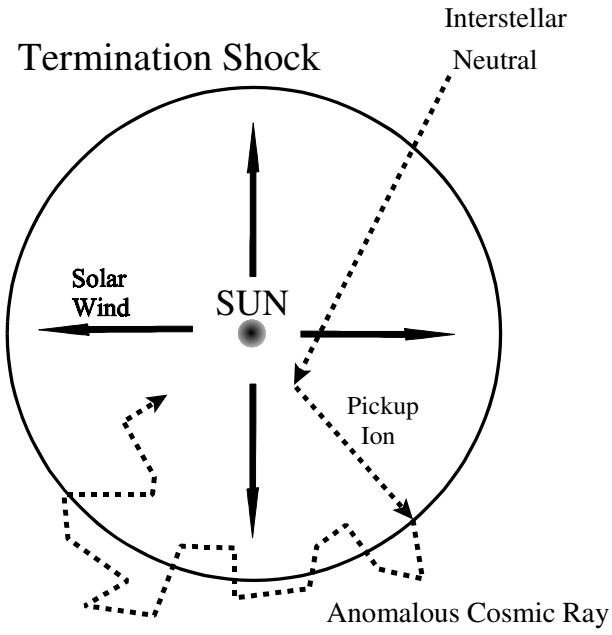


Figure 1. Schematic illustration of the heliosphere and the acceleration of the anomalous cosmic rays. Shown are the incoming interstellar neutral, the ionization to form a pickup ion and the subsequent acceleration and propagation of the anomalous cosmic ray.

Acceleration

The acceleration of the anomalous cosmic rays from low energies to the typical observed energies of ≈200–300 MeV must occur in less than a few years, because otherwise further electron loss would occur owing to interactions with ambient gas, and they would not be singly charged. In fact, independent of the acceleration site or model, the observed fact that anomalous cosmic-ray oxygen particles at ≈160 MeV are still mainly singly charged implies that the total path length they have traversed since being accelerated is less than 0.2 pc. Given that they cannot travel in a straight line because of the ambient, irregular magnetic fields, this fact requires that the anomalous cosmic rays must be locally accelerated. At their observed speed, this implies that the time since acceleration is less than a few years. Diffusive shock acceleration at the quasi-perpendicular termination shock of the solar wind, which accomplishes the acceleration to ≈200 MeV in about a half a year, is the only process known at present which can accomplish the acceleration in the allowed time. In particular, the commonly invoked second-order Fermi acceleration mechanism, involving fluctuating electric fields associated with waves, is far too slow.

Diffusive shock acceleration is contained within the Parker cosmic-ray transport equation (see equation (1) of the main article on COSMIC RAYS), if one allows the wind flow to have a compressive discontinuity corresponding to the shock. This compressive discontinuity is the site of the acceleration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

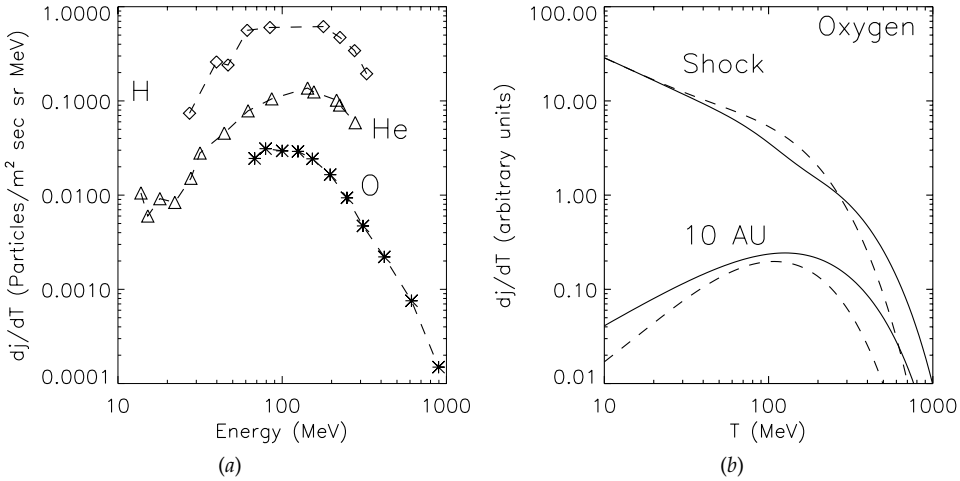


Figure 2. (*a*) Energy spectra of anomalous oxygen, helium and hydrogen observed near 21 AU in 1985 (Cummings and Stone 1987). (*b*) Plot of the computed energy spectra of anomalous oxygen near the heliospheric equator, for two radii, and for the two signs of the heliospheric magnetic field. The full curves correspond to the case where the northern heliospheric magnetic field is directed outward, corresponding to the present sunspot minimum. The broken curves are the same for the 1986 sunspot minimum.

The acceleration process at a quasi-perpendicular shock may be shown to be more closely related to the particle drift in the ambient electric field associated with the motion of the fluid than to the compression of the fluid at the shock, although mathematically it is contained in the compression term of the transport equation. It may be shown that the energy ΔT gained by a particle having electric charge Ze, at a quasi-perpendicular shock, is approximately the electrostatic potential energy $\Delta\phi$ gained in drifting along the shock face: $\Delta T \approx Ze\,\Delta\phi$. We may write for the solar wind

$$(\Delta\phi)_{\max} = B_r r^2 \Omega_\odot / c \approx 240\ \text{MeV}/Z \qquad (1)$$

where c is the speed of light and the numerical value results from using a radial magnetic field of 3.5 γ at a radius of 1 AU and a solar rotational angular velocity $\Omega_\odot = 2.9 \times 10^{-6}$. It follows, then, that for particles having charge Z, there is in general a characteristic energy $T_c \approx Ze(\Delta\phi)_{\max}$ above which the spectrum begins to decrease rapidly. Because of this, any multiply charged particles should be accelerated to higher energies than singly charged ones. Particles will 'drift off' of the shock before gaining more energy. Singly charged particles accelerated at the termination shock would then have an energy spectrum at the termination shock which is roughly a power law and which exhibits a decrease above an energy between 200 and 300 MeV. This behavior is seen in figure 2(*a*). This energy limit will not necessarily produce a sharp cutoff in the energy spectrum because the diffusion will cause some particles to gain more energy and some less.

Multiply charged particles can be produced by further interactions with ambient material. These then could gain much more energy, depending on their charge, and will therefore become dominant at higher energies, beyond the characteristic energy for singly charged particles. Observations in the past few years have observed a transition to multiply charged anomalous cosmic rays at energies greater than some 240 MeV.

Results of model calculations

The scenario discussed above has been incorporated into both two- and three-dimensional numerical simulations of transport and acceleration in the heliosphere, which solve the transport equation for realistic parameters.

We note that the general heliospheric magnetic field (see article on the SOLAR MAGNETIC FIELD) changes sign every 11 yr, during sunspot maximum. The motions of energetic charged particles depend to a significant degree on the sign of the magnetic field. In particular, the drift velocity, given for an isotropic distribution of particles of velocity w and momentum p in an ambient magnetic field $\boldsymbol{B}(\boldsymbol{r})$, is given by the expression

$$\boldsymbol{V}_d = \frac{pcw}{3Ze} \nabla \times \frac{\boldsymbol{B}}{B^2}. \qquad (2)$$

Clearly, $\boldsymbol{V}_d$ reverses direction with the change in sign of the magnetic field. During the current period of sunspot minimum, the northern magnetic field of the Sun is directed outward from the Sun, whereas prior to the last maximum it was directed inward. The present models are most accurate during sunspot minimum, when the interplanetary current sheet is closest to being equatorial, and so we basically discuss two heliospheres, one near a sunspot minimum when the field is outward (current) and one near the next sunspot minimum, when the field is inward.

First consider the model results for the case where low-energy particles are injected uniformly at all latitudes at the shock. The computed and observed spectra are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

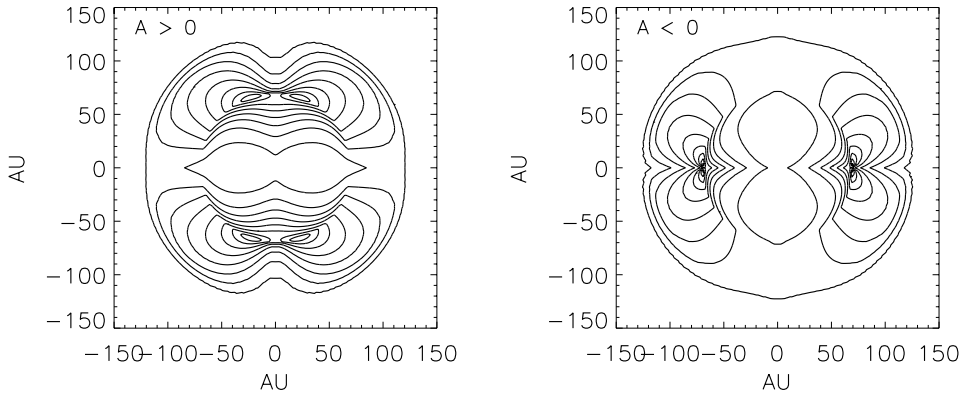


Figure 3. Equal-intensity contours computed for 8 MeV/nucleon anomalous oxygen, as a function of position in a meridional plane, as computed for the model parameters discussed in the text. The left panel is for the present sunspot minimum and the right is for the 1986 minimum.

compared in figure 2, for data taken in 1985 (northern magnetic field inward). The computed energy spectra of anomalous oxygen are clearly in good agreement with the observed spectra. This is also true of other spectra, as well.

Figure 3 illustrates contour plots of the intensity of the modeled anomalous oxygen as a function of heliocentric radius and polar angle, in a solar meridional plane. We see that the intensity increases with radius, much as does that of the galactic cosmic rays, out to the termination shock (at a radius of 60 AU). Beyond, the intensity decreases out to the outer boundary of the heliosphere. Along the shock, the **maximum** intensity occurs at a latitude which shifts as the sign of the magnetic field changes. If the northern interplanetary magnetic field is outward ($A > 0$), the particles drift toward the pole along the shock face and then inward and down from the poles to the current sheet, and the intensity maximum is near the poles. For $A < 0$, the drifts are in the opposite direction, and the maximum shifts to the equator.

There have been numerous measurements of the radial and latitudinal gradients of the anomalous component. It is a general feature of the simulations of anomalous cosmic rays and galactic cosmic rays that, for identical parameters, the latitudinal gradients of the anomalous cosmic rays are significantly larger than those for galactic particles. This is in large part due to the large latitudinal gradient of anomalous cosmic rays induced by drift during the acceleration at the shock. This is also true in the models.

A robust prediction of the models is that the latitudinal gradient near the current sheet, near sunspot minimum, should change sign in alternate sunspot minima. In particular, during the last sunspot minimum, near 1975, the intensity of both galactic cosmic rays and the anomalous component should **increase** away from the current sheet, whereas during the 1986 sunspot minimum the cosmic rays should **decrease** away from the current sheet. Near sunspot minimum, when the current sheet is nearly flat, the effects of these drifts are expected to be the most important. In 1977, PIONEER observed a positive gradient of anomalous helium away from the current sheet, as predicted. Observations carried out during the last sunspot minimum during a period in 1984 and 1985, when the current sheet tilt went below the latitude of the VOYAGER 1 spacecraft, found that the sign of the latitudinal gradient changed from being positive to negative, again as predicted by the theory.

Pickup ions

The above discussion has summarized the present status of the acceleration of anomalous cosmic rays to the highest energies observed, in excess of 1 GeV. Diffusive acceleration at the termination shock of the solar wind seems to be the most likely mechanism for this acceleration.

Not yet fully resolved is the question of how previously unaccelerated pickup ions are accelerated to energies large enough to be further accelerated at the termination shock. This 'injection problem' stems from the fact the the non-accelerated pickup ions have speeds that are less than or equal to the convection speed of the solar wind and hence are not mobile enough to encounter the nearly perpendicular shock the many times necessary to gain significant energy. It has been shown that propagating shocks more readily accelerate low-energy pickup ions than shocks that stand in the solar wind. Consequently, they are more natural injectors of particles and may represent the first stage in a two-step process in accelerating pickup ions to anomalous cosmic-ray energies.

Acknowledgments

This work was supported by NASA/JPL as part of an interdisciplinary investigation on the Ulysses mission, JPL contract number 960843. It was also supported in part by the National Science Foundation under grant ATM 9616547 and by the National Aeronautics and Space Administration under grants NAG 2251 and NAGW 1931.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Cummings A C and Stone E C 1987 Energy spectra of anomalous cosmic-ray oxygen during 1977–1987m *Proc. 20th Int. Cosmic Ray Conf. (Moscow)* vol 3, p 421

Jokipii J R and Giacalone G 1998 The theory of anomalous cosmic rays *Cosmic Rays in the Heliosphere* (Dordrecht: Kluwer)

Jokipii J R and McDonald F B 1995 The quest for the limits of the heliosphere *Sci. Am.* (April)

Klecker B The anomalous component of cosmic rays in the 3-D heliosphere *Space Sci. Rev.* **72** 419

J R Jokipii

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anomaly

For a particle moving in a conic orbit, the true anomaly is the angle between the point of closest approach to a focus, the focus, and the position of the particle. In the case of a planet moving around the Sun in an elliptical orbit, its true anomaly at any particular instant is the angle between its perihelion position, the Sun (which lies at one focus of the ellipse), and the position of the planet at that instant.

For a body moving around a massive body in an elliptical orbit, the mean anomaly at a particular instant is the angle through which a particle moving at uniform angular speed on a circle of area equal to the ellipse (and with an orbital period equal to that of the real body in its elliptical orbit) will have moved in the time which has elapsed since the real body last passed its closest approach to the massive body. For an ellipse, mean anomaly equals true anomaly at opposite ends of the major axis; for a circular orbit, mean and true anomaly are identical at all times. The concept of mean anomaly is useful in the calculation of the time taken for a body to move along a segment of its orbit.

See also: barycenter, ellipse, inclination, node, sidereal period, synodic period.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Antapex

The point on the celestial sphere directly opposite the solar apex (the point towards which the Sun is moving). The antapex is located in the constellation of Columba, at about right ascension 6 h and declination -30°.

See also: apex, celestial sphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Antares

The star α Scorpii, said to represent the heart of the scorpion. The name derives from 'rival to Mars', that planet being of similar brightness and color at favorable oppositions. It is a striking red supergiant star, spectral type M1Ib, apparent magnitude 1.06. It is 604 light-years distant (parallax 0.005″), its apparent brightness being due to its great intrinsic luminosity (absolute magnitude −5.3). Antares is a slow, irregular, pulsating variable, its magnitude varying by ±0.14. There is a binary companion of magnitude 5.3 at position angle 274°, separation 2.9″; the binary period is estimated to be of the order of 1000 years.

Antares was known as Satevis by Persian observers *c*. 3000 BC and was one of their four 'Royal Stars'. It marked the autumnal equinox, being prominent during September evenings.

See also: Aldebaran.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anthelme, Voituret (*c.* 1618–83)

Born in Chatenay-Vaudin, France, Carthusian monk; discovered the nova variable star R Vulpecula in 1672.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anthropic Principle

Early humans started with the 'anthropocentric' view that we are located at the center of the universe. Phenomena tended to be interpreted anthropomorphically and even the laws of nature seemed to be human-centered in the sense that they could be exploited for our own purposes. However, this perspective was soon shattered: first by NICOLAUS COPERNICUS' revelation that the Earth moves around the Sun and then by NEWTON's '*Principia*' which led to the 'mechanistic' view of the world (see also COSMOLOGY: A BRIEF HISTORY). Mechanism says that the universe is a giant machine in which the laws of nature operate completely oblivious to the presence of humanity. In the last 40 yr or so, however, there has developed a reaction to mechanism which is called the 'anthropic' view (Barrow and Tipler 1986). This says that, in some respects, the universe has to be the way it is because otherwise it could not produce life and we would not be here speculating about it. Although the term 'anthropic' derives from the Greek word for 'man', it should be stressed that this is really a misnomer since most of the arguments pertain to life in general rather than humans in particular.

To start with a very simple illustration of an anthropic argument, consider the question, why is the universe as big as it is? The mechanistic answer is that, at any particular time, the size of the observable universe is the distance travelled by light since the big bang (see BIG BANG THEORY). Since the universe's present age is about 10^{10} yr, its present size is about 10^{10} light-years. Inherent in this straightforward answer is the belief that there is no compelling reason the universe has the size it does; it just happens to be 10^{10} yr old. There is, however, another answer to this question, one which Robert Dicke first gave (Dicke 1961). His argument runs as follows. In order for life to exist, there must be carbon or at least some form of chemistry. Now carbon is produced by cooking inside stars and this process takes about 10^{10} yr. Only after this time can the star explode as a SUPERNOVA, scattering the newly baked elements throughout space, where they may eventually become part of life-evolving planets. On the other hand, the universe cannot be much older than 10^{10} yr, else all the material would have been processed into stellar remnants such as WHITE DWARFS, NEUTRON STARS and BLACK HOLES. Since all the forms of life we can envisage require the existence of stars, this suggests that life can only exist when the universe is aged about 10^{10} yr.

This startling conclusion turns the mechanistic answer on its head. The very hugeness of the universe which seems at first to point to humanity's insignificance is actually a consequence of its existence. This is not to say that the universe itself could not exist with a different size, only that we could not be aware of it when its size was different. Of course, it could just be a coincidence that the age of the universe happens to be about the time required to produce intelligent life. Also, Freeman Dyson has argued that life could persist well after 10^{10} yr (Dyson 1979). However, this argument does at least give a taste for the type of reasoning the anthropic principle entails. Indeed, the evidence for the anthropic principle rests almost entirely on the large number of numerical 'coincidences' in physics which seem to be prerequisites for the emergence of life and which would otherwise have to be regarded as purely fortuitous.

Scales of structure

Most anthropic arguments involve dimensionless combinations of constants of physics, such as the speed of light (c), Planck's constant ($\hbar$), the gravitational constant (G), the charge of the electron (e) and the masses of elementary particles such as the proton (m_p) and the electron (m_e). For example, the electric 'fine structure' constant, $\alpha = e^2/\hbar c = 1/137$, determines the strength of the electric interaction and plays a crucial role in any situation where electromagnetism is important. Likewise the gravitational fine structure constant, $\alpha_G = Gm_p^2/\hbar c = 6 \times 10^{-39}$, determines the strength of the gravitational interaction and plays an important role in determining the structure of very large objects (such as stars). The fact that α_G is so much smaller than α reflects the fact that the gravitational force between two protons is so much smaller than the electric force between them. Gravity dominates the structure of large bodies only because these tend to be electrically neutral, so that the electric forces cancel out.

What is remarkable is that straightforward physics shows that, to an order of magnitude, α and α_G determine the mass and the size of nearly every naturally occurring object in the universe (Carr and Rees 1979). This is illustrated in figure 1, from which one can read off the scales associated with the universe, galaxies, stars, planets, asteroids, exploding black holes, humans, atoms and protons. Also shown are the Planck scales, the fundamental mass and length units associated with QUANTUM GRAVITY effects. The scales are given in grams and centimeters but also as powers of α and α_G times the proton mass and the size of a hydrogen atom. For example, all stars have a mass of roughly $\alpha_G^{-3/2} \sim 10^{60}$ times the proton mass; the largest planets (such as Jupiter) have a mass which is smaller than this by a factor $\alpha^{3/2}$; and the mass of humans—if we assume that they must live on a planet with a suitable temperature and a life-supporting atmosphere and that they must not shatter whenever they fall down—is of order $(\alpha/\alpha_G)^{3/4}$ times the proton mass. The characteristic mass of a galaxy is less understood but, on one possible argument, can be written as $\alpha_G^{-2}\alpha^4$ times the proton mass. One can express the sizes of these objects similarly. It should be stressed that the dependences on α and α_G in figure 1 are all consequences of straightforward physics. No anthropic argument has been introduced, except—as we shall see shortly—in deriving the scale of the universe itself.

The weak and strong anthropic principles

We now return to Dicke's assertion that the age of the universe (t_0) must be roughly the nuclear-burning time of a star. We can make this argument more specific

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

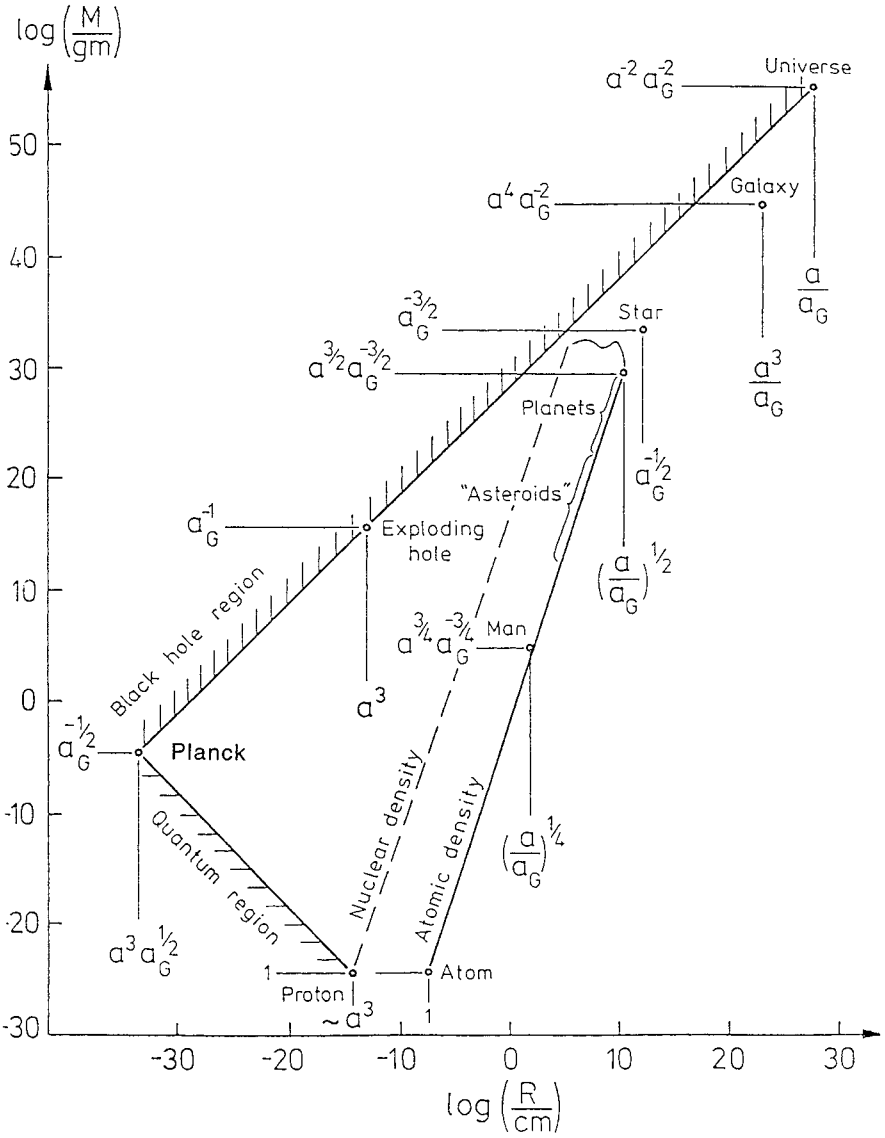


Figure 1. The mass and length scales of various structures in the universe. Also shown are the dependences of these scales on α and α_G in units of the proton mass and the size of a hydrogen atom.

because physics predicts that the lifetime of a star is around $\alpha_G^{-1} \sim 10^{40}$ times the time $t_p = \hbar/m_p c^2 \sim 10^{-23}$ s required for light to traverse the proton. One can infer that the ratio of the size of the universe to the size of an atom must be comparable with the ratio of α to α_G. Thus the anthropic principle explains why these two ratios must have the same huge value of 10^{36}. One can also infer a simple expression for the number of protons in the universe. This is because the big bang model implies that the mass of the observable universe is $\sim \rho_0 c^3 t_0^3$ where ρ_0 is its present density. Since the model also requires $\rho_0 \sim 1/Gt_0^2$, this mass is $c^3 t_0/G \sim \alpha_G^{-2} m_p$, so the number of protons in the universe is of order $\alpha_G^{-2} \sim 10^{80}$. (Note that the number of stars in the observable universe must then be around $\alpha_G^{-1/2} \sim 10^{20}$, a result we invoke later.) The existence of these 'large-number' coincidences has been appreciated for some time but, without resort to the anthropic principle, or alternatively the sort of variable-G model invoked by Dirac (1937), they have to be regarded as fortuitous.

Dicke's argument helps us to understand why the above 'large-number' coincidences prevail but it does not tell us why α_G^{-1} has its particular immense value. His argument is therefore an example of what is termed the weak anthropic principle. This accepts the constants of nature as given and then shows that our existence imposes a selection effect on when we observe the universe. As such it is no more than a logical necessity: saying that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

we have to exist at a certain time is no more surprising than saying we have to exist in a certain place (e.g. close to a star). It might be surprising to find what the selection effects are but their existence is not surprising in principle. In fact, most physicists would agree with the weak version of the anthropic principle.

The problem comes when we consider whether the constants of nature such as G are themselves determined by the requirement that life should arise, a notion which is sometimes referred to as the 'strong anthropic principle'. (Rather confusingly, this phrase is also used to describe the rather different notion that the universe *must* have properties such that life can arise (Barrow and Tipler 1986).) That the weak principle may not be the whole story is also suggested by the fact that all the scales shown in figure 1 are relative. If the fine structure constants differed from what we observe them to be, all the scales would change but the basic relationships between them would be the same. For example, one could envisage a hypothetical universe in which all microphysical laws were unchanged, but G was (say) 10^6 times stronger. Planetary and stellar masses ($\sim\alpha_{\rm G}^{-3/2}$) would then be lowered by 10^9 but hydrogen-burning main-sequence stars would still exist, albeit with lifetimes ($\sim\alpha_{\rm G}^{-1}$) of 10^4 yr rather than 10^{10} yr. Moreover, Dicke's argument would still apply: a hypothetical observer looking at the universe when $t_0 \sim t_{\rm MS}$ would find the number of particles in the universe 10^{12} times lower than in ours but they would still find the 'large-number' coincidences described above. If one fixed $\alpha_{\rm G}$ but allowed α to change, the effects would be less extreme, but still very noticeable.

What are the arguments against the 'cognizability' of this kind of small-scale speeded-up universe? One rather loose constraint on $\alpha_{\rm G}$ comes from biological considerations. We have seen that the number of stars in the observable universe is of order $\alpha_{\rm G}^{-1/2}$. If we regard stars—or at least their associated solar systems—as potential sites for life, this is also the number of places where life may have arisen. Of course, this is not a sufficient condition for life because there are a whole set of extra conditions, each of which may be very improbable. For example, we need the star to have a planet, we need the planet to be at a suitable distance from the star, it needs to have a suitable atmosphere and chemistry, and there must be the appropriate conditions for the first self-replicating cells to arise. Clearly therefore the overall probability (P) of life arising at any particular site must be very small. Now if we want to insist that there is life somewhere in the universe, we need the number of sites for life times the probability P to exceed 1. This implies that $\alpha_{\rm G}$ must be less than P^2. For example, if P were 10^{-15}, one would need $\alpha_{\rm G} < 10^{-30}$.

This is not a very precise argument but it does give a qualitative reason why $\alpha_{\rm G}$ needs to be small. There are, however, more specific anthropic arguments that pin down $\alpha_{\rm G}$ more narrowly. The first example of such an argument was given by Brandon Carter and relates to the existence of stars with convective envelopes (Carter 1974). Such a star arises if the heat generated in its core by nuclear reactions is transported to the surface primarily by way of large-scale motions of the stellar material itself. This tends to be the case for small-enough stars (red dwarfs). By contrast, larger stars (blue giants) tend to be 'radiative' in the sense that the heat gets out primarily via the flow of radiation. The dividing line between the two types is some critical mass which can be shown to be around $\alpha_{\rm G}^{-2}\alpha^{10}m_{\rm p}$. This critical mass happens to lie in the mass range $\sim\alpha_{\rm G}^{-3/2}m_{\rm p}$ in which stars actually exist only because of the remarkable coincidence $\alpha_{\rm G} \sim \alpha^{20}$. Were G (and hence $\alpha_{\rm G}$) slightly larger, all stars would be blue giants; were it slightly smaller, all stars would be red dwarfs. The convective condition does not pin down the actual values of α and $\alpha_{\rm G}$ but it does specify a scaling law between them and it explains why $\alpha_{\rm G}$ is so much smaller than α.

Carter ascribed anthropic significance to this relationship on the basis that the formation of planetary systems may be associated with convective stars. This suggestion was based on the observational feature that red dwarfs have much less angular momentum than blue giants and a loss of angular momentum could be a consequence of planet formation. This argument is no longer compelling because there are other ways of losing angular momentum. A better argument might be that only convective stars generate winds in their early phase intense enough to blow away the gaseous envelope of nearby planets, thereby facilitating the formation of solid planets with non-hydrogen atmospheres. In any case, Carter infers that no planets, and hence no life, would form if $\alpha_{\rm G}$ were much larger that α^{20}. If it were much smaller, all stars would be chemically homogeneous owing to convective mixing and one would not get the 'onion-skin' shell structure which characterizes pre-supernova models.

The weak and strong forces

On scales smaller than atoms two more fundamental forces of nature come into play: the strong and the weak force. Like gravity and electromagnetism, the strength of these forces can be described by dimensionless fine structure constants. The weak force has a coupling constant $\alpha_{\rm W} = (gm_{\rm e}^2c/\hbar^3) \sim 10^{-10}$ where $g \sim 10^{-49}$ erg cm^3 is the Fermi constant. Thus its interaction strength is midway between those of gravity and electricity. The dimensionless constant describing the strong force is denoted by $\alpha_{\rm S}$ and has a value of order 10. Although the strong and weak forces are many orders of magnitude stronger than the gravitational force, they are both short range, becoming negligible at distances of 10^{-13} cm and 10^{-15} cm, respectively. For this reason they do not play an important role in determining the structure of objects larger than atoms.

It turns out (Barrow and Tipler 1986) that many features of chemistry are sensitive to the value of $\alpha_{\rm S}$. For example, if $\alpha_{\rm S}$ were increased by 2%, all the protons in the universe would combine at big bang nucleosynthesis into diprotons (nuclei consisting of two protons). In this

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

case, there would be no hydrogen and hence no hydrogen-burning stars. Since stars would then have a much reduced main-sequence time, there might not be time for life to arise. If α_S were increased by 10%, the situation would be even worse because everything would go into nuclei of unlimited size and there would be no interesting chemistry. The lack of chemistry would also apply if α_S was decreased by 5% because all deuterons would then be unbound and one could only have hydrogen.

The most sensitive constraint on the value of α_S is associated with what is called the 'triple-alpha' reaction. The way a star makes carbon is by first combining two alpha particles to make a beryllium nucleus and then adding another alpha particle to form a carbon nucleus ($^4\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^8\mathrm{Be}$, $^8\mathrm{Be} + {}^4\mathrm{He} \rightarrow {}^{12}\mathrm{C}$). The trouble is that beryllium is unstable (otherwise the 'helium flash' in giants would lead to a catastrophic explosion) and it used to be thought that it would decay before the extra alpha particle could combine with it. For many years, therefore, it was difficult to understand why there is any carbon in the universe. Then Fred HOYLE (Hoyle 1953) realized that there must be a resonance (i.e. an enhanced interaction rate) in the second step which allows the carbon to form before the beryllium disappears, i.e. ^{12}C must have a state with energy just above the sum of the energies of ^{8}Be and ^{4}He. There is, however, no similar favourably placed resonance in ^{16}O; otherwise almost all the carbon would be transmuted into oxygen. Once the suggestion was made, the resonance was looked for in the laboratory and rapidly found, so this might be regarded as the first confirmed anthropic prediction. Indeed, the fine tuning required is so precise that Hoyle concluded that the universe has to be a 'put-up job'.

The value of α_W is also involved in an interesting anthropic constraint involving supernovae (Carr and Rees 1979). We have seen that supernovae are essential for life because they spread heavy elements throughout the universe. However, it is still uncertain *why* a star explodes after burning its nuclear fuel. One explanation is that the core of the star becomes very hot when it collapses and this generates a lot of neutrinos which then blow off the envelope as a result of weak interactions. For this model to work, one requires the timescale on which neutrinos interact with nuclei in the envelope to be comparable with the dynamical timescale. If it were much longer, the envelope would be essentially transparent to the neutrinos; if it were much shorter, the neutrinos would be trapped in the core, and could not escape to deposit their momentum in the less tightly bound surrounding layers. The two timescales are comparable provided that $\alpha_G \sim \alpha_W^4$. We know that this relationship holds numerically but the anthropic principle explains why it must hold. So if we accept that α_G is determined anthropically, we must also accept that α_W is so determined.

Rather remarkably the relation $\alpha_G \sim \alpha_W^4$ also plays a crucial role in big bang nucleosynthesis calculations. The prediction that 24% of the mass of the universe is turned into helium when it was just a few minutes old is one of the great triumphs of the big bang picture. However, the helium production depends sensitively on the neutron-to-proton ratio when the weak interactions become slower than the cosmological expansion rate at a temperature of 10^{10} K and the only reason one obtains an interesting helium yield is because the 'freeze-out' temperature is comparable with the neutron–proton mass difference. It turns out that the condition for this is $\alpha_G \sim \alpha_W^4$, precisely the condition required for supernovae. It is not clear to what extent this coincidence can be interpreted anthropically. If α_W were slightly smaller than $\alpha_G^{1/4}$, the entire universe would have burned to helium and the lifetime of a helium star might not be long enough to permit the evolution of life. If α_W were slightly larger than $\alpha_G^{1/4}$, there would have been no helium production at all, although it is not clear that this has any anthropic significance.

Cosmological anthropic constraints

Several more qualitative anthropic constraints derive from cosmological considerations. For example, one can presumably exclude a universe with zero matter density (i.e. pure radiation) since there would then be no baryonic content at all. There are also simple anthropic reasons for why the total density parameter Ω lies within an order of magnitude of 1. (This is the density in units of the critical value which divides models which expand forever from those which recollapse.) If Ω were much larger than 1, the universe would recollapse on a timescale much less than the main-sequence time of a star. On the other hand, if Ω were much smaller than 1, density fluctuations would stop growing before galaxies could bind. This argument requires that Ω be in the range 0.01–100.

It is sometimes claimed that the inflationary scenario requires that Ω equals 1 to much greater precision than this, in which case the anthropic explanation may no longer seem relevant. However, even the inflationary scenario only works if one chooses the vacuum potential so that one has a sufficient number of expansion e-folds and density fluctuations of the right amplitude. The form of this potential may therefore itself be constrained by anthropic arguments. In fact, some theorists now advocate 'open inflationary' scenarios with $\Omega < 1$. In this case, in order to specify the present value of Ω, one needs to invoke what Alex Vilenkin terms the 'mediocrity principle' (the idea that we are most likely to live in a universe with the largest number of galaxies) (Vilenkin 1995). However, this is really just the anthropic principle by another name.

Another important cosmological parameter is the amplitude of the primordial density fluctuations required to explain the currently observed large-scale structure. When these fluctuations came inside the horizon, they must have had an amplitude ε of order 10^{-5}. In this case, galaxies formed when the universe was 10^9 yr old, while clusters of galaxies are separating out at the present epoch. What would happen if the value of ε were different for some reason? If ε were below 10^{-6}, baryons would condense only at very late times when they would be very

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

diffuse. Cooling would then be ineffective, so gas would never condense into stars. On the other hand, if ε were substantially larger than 10^{-3}, baryons would condense early into tightly bound systems that would trap radiation and collapse to supermassive black holes. The evolution in such a universe would be hard to compute but certainly it would contain nothing like galaxies. The anthropically allowed range of ε therefore seems to lie between 10^{-3} and 10^{-6} (Tegmark and Rees 1998).

A third important cosmological parameter is the photon-to-baryon ratio $S \sim 10^9$ (i.e. the ratio of photon density in the microwave background to the average proton density). In the standard big bang model, the background radiation dominated the density of the universe until a time $\sim 10^4$ yr and the formation of galaxies cannot occur until after this. However, for a general value of S this time would be $\sim S^2 \alpha_{\rm G}^{-1/2} t_{\rm p}$ and the age of the universe given by the Dicke argument would then only exceed this for $S < \alpha_{\rm G}^{-1/4}$. Thus we have an upper limit on S of about 10^{10}. A lower limit on S is obtained if one requires that the universe be radiation dominated at cosmological nucleosynthesis, else all the hydrogen would go into helium, and this implies $S > (m_{\rm p}/m_{\rm e})^{4/3} (\alpha_{\rm w}^4/\alpha_{\rm G})^{1/6} \sim 10^4$. Thus anthropic arguments require S to lie in the range 10^4–10^{10}. Of course, in the standard picture the value of S results from a small fractional excess of particles over antiparticles imprinted at around $\sim 10^{-34}$ s after the big bang because of violation of baryon conservation occurring at the grand unified epoch. However, in most GRAND UNIFIED THEORY (GUT) models S is predicted to be of the form α^{-n} where n is an integer, so the anthropic constraint $S < \alpha_{\rm G}^{-1/4}$ merely translates into the constraint $\alpha_{\rm G} < \alpha^{4n}$. If $n = 5$, this just gives the convective star condition (Carr 1991).

Finally we must consider anthropic constraints on the value of the cosmological constant Λ. This is often taken to be zero but this is neither theoretically nor observationally inevitable. The effect of a positive Λ is to cause the universe to embark on exponential expansion when the density of ordinary matter has dropped sufficiently low. The growth of density perturbations is then quenched, so bound systems which have not formed by then never will. For a given value of ε, anthropic considerations therefore set an upper limit to Λ (Efstathiou 1995). Interestingly, this limit is only a factor of 10 higher than value of Λ which some cosmologists favor for our present universe. In the perspective of the 120 powers of 10 uncertainty in the vacuum energy density which particle theorists predict, this is a very severe constraint. Of course, if more precise cosmological observations show that the vacuum energy density is much less than critical density, then one might suspect that Λ is very small, perhaps exactly zero, for some fundamental physical reason.

Interpretations of the anthropic principle

Whether one regards anthropic arguments as coming under the heading of physics or metaphysics depends on how one interprets them. Various sorts of explanation have been suggested and these are illustrated in figure 2. The first possibility is that the anthropic coincidences reflect the existence of a 'beneficent being' who created the universe with the specific intention of producing us (figure 2(*a*)). Such an interpretation is logically possible and may have appeal to theologians, but it is probably unpalatable to most physicists.

Another possibility, proposed by John Wheeler (Wheeler 1977), is that the universe does not properly exist until consciousness has arisen. This is based on the notion that the universe is described by a quantum mechanical wavefunction and that consciousness is required to collapse this wavefunction. Once the universe has evolved consciousness, one might think of it as reflecting back on its big bang origin, thereby forming a closed circuit which brings the world into existence (figure 2(*b*)). Even if consciousness really does collapse the wavefunction (which is far from certain), this explanation would also have to be regarded as metaphysical.

The third possibility is that there is not just one universe but lots of them, all with different randomly distributed coupling constants. In this case, we necessarily reside in one of the small fractions which satisfy the anthropic constraints. With this interpretation, the fact that the constants have the values required for life just becomes an aspect of the weak anthropic principle, for conscious observers need not only be at special points in time and space but also in special universes. Invoking lots of extra universes might seem rather uneconomical but there are several physical contexts in which it makes sense.

One possibility is connected with the 'Many Worlds' interpretation of quantum mechanics originated by Hugh Everett (Everett 1957). The idea here is that the universe splits every time the wavefunction collapses as a result of an observation. This means that the universe is continuously branching into a huge number of parallel worlds and this gives rise to the famous Schrödinger cat paradox. David Deutsch argues that all these worlds are 'real' in some sense (Deutsch 1991). However, they all have the same coupling constants in the Everett picture, so it is a considerable extrapolation to argue that there are also worlds in which the constants differ.

A more plausible context for the 'many universes' scenario arises if the universe is closed. In this case, one could envisage it undergoing cycles of expansion and recollapse, with new values of the coupling constants being generated at every bounce (figure 2(*c*)). During most cycles the constants would not allow life to arise but occasionally the appropriate values would occur and so the universe would become aware of itself. However unlikely the constants are to have the correct values, with an infinite number of cycles it is bound to happen sometimes.

Another 'many universes' scenario arises in the context of the inflationary picture. This proposes that at very early times quantum fluctuations cause tiny regions to undergo an exponential expansion phase (see INFLATION).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

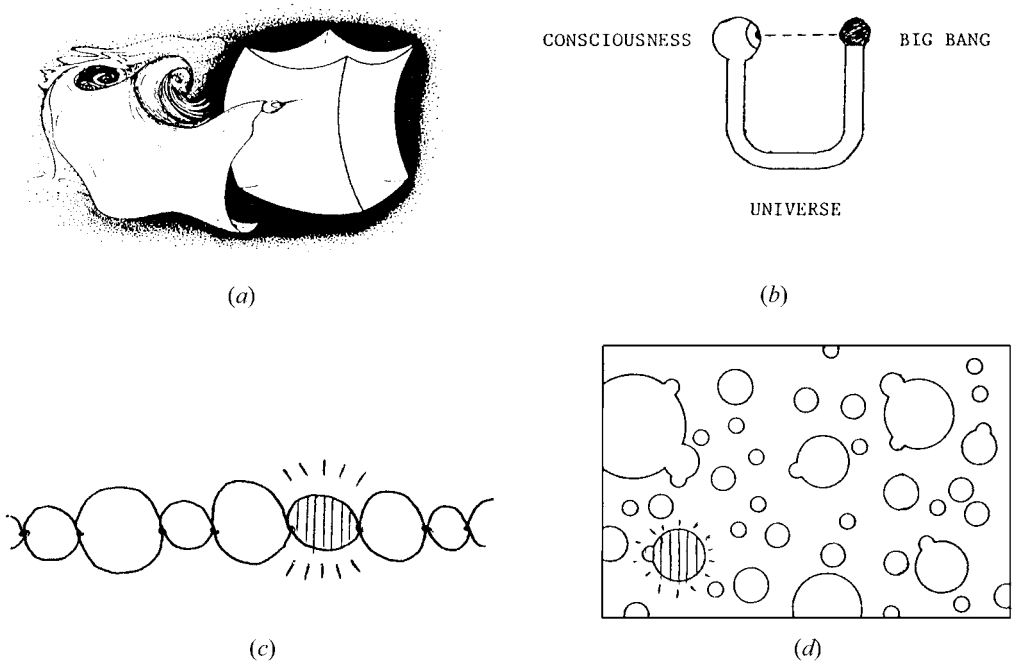


Figure 2. Different interpretations of the anthropic principle. (*a*) God creates the universe; (*b*) consciousness creates the universe; (*c*) cyclic universe; (*d*) inflationary universe.

Each region becomes a 'bubble' and our entire visible universe is then contained within one of these. In principle, as stressed by Andrei Linde (Linde 1985), there could be different values for the coupling constants within each bubble (figure 2(*d*)). In this case, only a small fraction of them would develop consciousness but our universe would necessarily be among that fraction. Both the cyclic and the inflationary models invoke a large—and possibly infinite—number of universes. The only difference is that the first invokes an infinity in time, while the second invokes an infinity in space.

Concluding assessment

The possibility of life as we know it evolving in the universe is contingent on the values of a few basic physical constants—and is in some respects remarkably sensitive to their numerical values. On the other hand, from a physical point of view, the anthropic 'explanation' of the various coincidences in nature can be criticized on a number of grounds. (1) The anthropic arguments are mainly *post hoc*—apart from the triple-alpha resonance, nobody has ever based a successful prediction on the anthropic principle. (2) The anthropic arguments may be unduly anthropocentric in that we have assumed conditions which are specifically associated with human-type life. (3) The anthropic principle does not predict exact values for the constants but only order-of-magnitude relationships between them, so it is not a complete explanation.

The last two objections might be met within the 'many universes' interpretation if one envisages the universes as occupying an abstract multidimensional space of coupling constants (with one dimension for each α). The anthropic principle could then be interpreted as saying that life can only exist in small regions of that space. Lifeforms similar to our own will be possible in some regions; other lifeforms may be possible elsewhere, but life will not be possible everywhere (see also EXOBIOLOGY, LIFE ON OTHER WORLDS). One may not have the same anthropic relation in every life-supporting region of the space but one will have some relation.

The other attraction of the 'many universes' explanation, as we have seen, is that it may be possible to give it a reasonable physical basis (Leslie 1989). Of course, it is not clear that these physical ideas will ever be testable. For example, we will probably never know whether the universe can undergo cycles; even if we can prove that it will recollapse, we cannot be certain that it will bounce into another expansion phase. In the inflation picture the situation is no better because we will never be able to contact the other bubbles. Both pictures are therefore untestable and, in this sense, one might still regard the anthropic principle as being metaphysical.

Perhaps the least radical explanation of the anthropic coincidences is that they may turn out to be a consequence of some unified theory of particle physics. Such theories do, after all, set out to relate the different coupling constants, so it is not inconceivable that they would predict the sort of connections between the different α's discussed above. However, as far as we know, the relationships discussed above are *not* predicted by any unified theory. Even if they were, it would still be remarkable that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

theory should yield exactly the coincidences required for life.

Finally it should be cautioned that many of the anthropic arguments merely relate to the presence of complexity—the anthropic coincidences could be equally regarded as prerequisites for inanimate objects such as motor cars—so the emphasis on life may be misplaced. In this context, Lee Smolin has conjectured that the constants governing our universe are determined not by anthropic considerations but by a kind of 'natural selection' among universes (Smolin 1992). He postulates that new universes sprout whenever black holes form and that the new constants, while differing slightly from their values in the parent universe, retain some memory of their parentage. This idea is exceedingly speculative and involves physics far beyond anything we can be confident of. However, Smolin's idea at least has the virtue of being testable: it predicts that the properties of our universe should be optimal for black hole formation, so that any tinkering with the constants would reduce the propensity of black holes to form. Conventional astrophysics can at least start to address this question.

Bibliography

Barrow J D and Tipler F J 1986 *The Anthropic Cosmological Principle* (Oxford: Oxford University Press)
Carr B J 1991 *J. Brit. Interplan. Soc.* **44** 63
Carr B J and Rees M J 1979 *Nature* **278** 605
Carter B 1974 *Confrontation of Cosmological Models with Observations* ed M S Longair (Dordrecht: Reidel) p 291
Deutsch D 1991 *Phys. Rev.* D **44** 3197
Dicke R H 1961 *Nature* **192** 440
Dirac P A M 1937 *Nature* **139** 323
Dyson F 1979 *Rev. Mod. Phys.* **51** 447
Efstathiou G 1995 *Mon. Not. R. Astron. Soc.* **274** L73
Everett H 1957 *Rev. Mod. Phys.* **29** 454
Hoyle F 1953 *Astrophys. J.* **118** 513
Leslie J 1989 *Universes* (London: Routledge)
Linde A 1985 *Rep. Prog. Phys.* **47** 925
Smolin L 1992 *Class. Quantum Grav.* **9** 173
Tegmark M and Rees M J 1998 *Astrophys. J.* **499** 526
Vilenkin A 1995 *Phys. Rev. Lett.* **74** 846
Wheeler J 1977 *Foundational Problems in the Special Sciences* ed Butts and Hintikka (Dordrecht: Reidel) p 3

Bernard Carr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Antimatter

Matter composed of elementary particles that have the same masses as their ordinary matter equivalents but which have opposite, or 'mirror image', values of other quantities, such as charge. For example, an anti-electron (or 'positron') has the same mass as an electron but has positive, rather than negative, charge; an anti-proton has the same mass as a proton but has a negative, rather than a positive, charge.

Antiparticles are found naturally among cosmic rays and can be produced in high-energy particle accelerators. Because of the equivalence of mass and energy, high-energy gamma-ray photons can transform into particle–antiparticle pairs (for example, an electron and a positron), the simultaneous production of a particle and an antiparticle being required to satisfy conservation laws, such as the conservation of charge (because a photon has zero charge, it has to transform, for example, into a positive positron and a negative electron, so ensuring that the net charge remains zero). When a particle and its antiparticle collide, they annihilate each other, transforming their mass into energy, normally in the form of gamma-ray photons.

Any antiparticles that are produced in our locality are quickly annihilated through colliding with particles of ordinary matter. While in principle it is possible for antimatter to exist in bulk (for example, stars composed of atoms that consist of anti-protons, anti-neutrons and positrons), and the possibility cannot be excluded, there is no evidence to suggest that this is so. All the bulk matter in our part of the universe is composed of conventional matter, and it is widely believed that the universe itself is dominated by matter rather than antimatter.

See also: cosmic rays, electron, elementary particles, matter, photon, proton, positron, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Antlia

(the Air Pump; abbrev. Ant, gen. Antliae; area 239 sq. deg.) A southern constellation that lies between Hydra and Vela, and culminates at midnight in late February. It was named Antlia Pneumatica by the French astronomer Nicholas L de Lacaille (1713–62), who charted the southern sky in 1751–2. The IAU adopted the shortened form in 1922.

A small, inconspicuous constellation, overshadowed by its neighbors Centaurus and Vela, the brightest star in Antlia is of magnitude 4.3. Interesting objects include ζ Antliae, a wide double star with pale yellow (A9 and A1) components, magnitudes 5.9 and 6.2, the latter of which is itself double, having a pale yellow (A1) component, magnitude 7.0, separation 8″, S Antliae, an eclipsing binary (range 6.4–6.9, period 0.65 day), and NGC 2997, a tenth-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Antoniadi, Eugenios (1870–1944)

Turkish-born astronomer. Starting with an interest in historical astronomy, he met CAMILLE FLAMMARION who inspired him to observe the planets. He worked in France at the Juvisy and Meudon observatories and made the best maps of Mars before the Space Age.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apache Point Observatory

The principal projects at Apache Point Observatory at Sunspot, New Mexico, USA, are the 3.5 m telescope, the Sloan Digital Sky Survey and New Mexico State University's 1.0 m telescope. The 3.5 m telescope construction incorporates many innovations: compact and lightweight design, thermal control, multiple instruments and remote observing. The SLOAN DIGITAL SKY SURVEY is generating a three-dimensional map of a large volume of the northern night sky using an array of sensitive instruments coupled to a 2.5 m telescope.

The Observatory is owned by the Astrophysical Research Consortium consisting of the University of Chicago, Institute for Advanced Study, Johns Hopkins University, New Mexico State University, Princeton University, University of Washington and Washington State University. The Sloan Digital Sky Survey is a joint project of the University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, the Johns Hopkins University, the Max Planck Institute for Astronomy, Princeton University, the United States Naval Observatory and the University of Washington.

For further information see
http://www.apo.nmsu.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apastron

The farthest point in the relative orbit of one component of a binary system from the other component. The term is also used for the most distant position of a planet orbiting a star other than the Sun.

See also: aphelion, binary system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aperture

The clear diameter of the objective of a refracting telescope or the primary mirror of a reflecting telescope. A telescope is characterized by its aperture; for example, a telescope with an aperture of 2 m would be referred to as a '2 m' telescope.

See also: light-gathering power, mirror, objective lens, resolving power, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aperture Synthesis

A technique, also known as Earth-rotation synthesis, that enables the outputs of two or more radio telescopes to be combined to give the same resolution as a single radio dish of aperture equivalent to the maximum separation of the dishes.

A single observation with a conventional radio interferometer produces high resolution only in a direction parallel to the line joining the two antennae, or dishes. If two radio telescopes (A and B) are set up at opposite ends of a track and continue to observe a particular radio source for 12 h then, during that time, the rotation of the Earth will change the orientation of the baseline relative to the source. In particular, if the radio telescopes were located at the north pole, the Earth's rotation would cause A to trace out a semicircle around B, so effectively tracing out a half-ring strip of a 'dish' centered on B. The data received by the dishes are stored and the other half of the ring can be constructed mathematically. Dish A is then moved a little closer to B to enable another 'ring' to be filled in and the process repeated until the entire 'dish' has been synthesized. The mass of accumulated data is then converted into an image of the source equivalent to that which would be obtained by a single dish of aperture equal to the maximum separation of the two dishes. In practice, the resolution achieved along one diameter of the 'dish' (north–south, say) will be the same as that at right angles to this direction (east–west) only if the declination of the source is 90°; resolution along the north–south direction decreases for sources closer to the celestial equator.

A synthesis array with only two dishes would take a very long time to build up the effect of a single large dish. The process may be speeded up by using more than two dishes, some of which may be fixed and others movable and, in some cases, the dishes making up the array can be set out along more than one track. The largest complete Earth-rotation synthesis system is the Very Large Array (VLA), located in the New Mexican desert close to Socorro. The VLA, which comprises 27 dishes, each 25 m in diameter, set out along a Y-shaped track, can simulate a single dish 36 km in diameter. Usually, the outputs from the various dishes are linked together by cables, but the Multi-Element Radio-Linked Interferometer Network (MERLIN), in the UK, uses radio links to connect dishes separated by distances of up to 230 km. Because it links a number of widely separated individual dishes it cannot simulate, and attain the sensitivity of, a complete dish. Nevertheless, it can detect details as small as 0.01 arcsec at its shortest operating wavelengths.

See also: radio interferometer, radio telescope, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apex

The direction of motion of the Sun relative to the stars in the local part of the Galaxy is known as the solar apex, or the apex of the Sun's way. Observations of the proper motions of stars indicate that they show a tendency to diverge from a point in the constellation Hercules, and to converge towards a point in the constellation Columba. These observations indicate that the Sun is moving at a velocity of about 19.5 km s^{-1}, towards a point at right ascension (RA) 18 h and declination (dec) $+30^{\circ}$ in Hercules and away from a point at RA 6 h and dec -30° in Columba (the solar antapex).

Apex can also refer to the point on the celestial sphere towards which the Earth appears to be moving as a result of its orbital motion around the Sun.

See also: antapex.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aphelion

The point in the elliptical orbit of a planet or other object around the Sun at which it is farthest from the Sun. The Earth, for example, reaches aphelion in July, when it is just over 152 million km from the Sun.

Compare: perihelion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aphrodite Terra

The largest of the three main upland regions of Venus (the other two being Ishtar Terra and Lada Terra), extending almost halfway around the planet's equator and centered at 5.8 °N, 104.8 °E. Its overall length is 9999 km. It takes its name from the Greek goddess of love. The western end of Aphrodite is formed by two distinct units of elevated terrain: Ovda Regio at the far west, and Thetis Regio to its east, both rising to about 4 km about the planet's mean surface level. At the eastern end lies a similar region, Alta Regio, and beyond it the less elevated Ulfrun Regio. Like Ishtar and Lada Terrae, these regiones in Aphrodite are characterized by the type of terrain termed tesserae—areas of intersecting troughs and ridges. Between Thetis and Alta is a major complex canyon system, the most prominent parts of which are named Diana Chasma and Dali Chasma. These chasmata are steep-sided with rims raised above their surrounds, with a maximum difference, in the case of Dali Chasma, of 6 km between rim and canyon floor. Another, almost circular canyon, Artemis Chasma, lies to the south of Thetis; Ganis Chasma extends northward from Alta. In Alta Regio are situated a number of prominent volcanic peaks, the highest of which, Maat Mons, is over 8 km high. A global analysis of wrinkle ridges—compressive-stress features common on Venus's plains—shows that they tend to be radially aligned on Aphrodite Terra, suggesting a common origin in the event that uplifted Aphrodite.

See also: Venus: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apian [Bienewitz or Bennewitz], Peter (1495-1552)

Cartographer, born in Leisnig, Saxony, became professor of mathematics at Ingolstadt, where he founded a press that published maps. He made astronomical instruments, including sundials, quadrants and armillary spheres, and other instruments that would solve spherical trigonometry. His world map is the oldest that uses the name America. His major work (1540), the *Astronomicum Caesareum*, is a sumptuous atlas of the skies with revolving paper models for computation. Apian made clear JOHANN WERNER's proposal for the use of the Moon's position to determine longitude, so that sailors could understand its potential, and thus influenced the establishment of the Royal Observatory at Greenwich, and the Paris Observatory. Observed five comets, including Halley's, and noticed that their tails all pointed away from the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apoapsis

The point in an elliptical orbit at which the orbiting body is closest to the body it is orbiting. The prefix 'apo-' or 'ap-' may be attached to various words or roots depending on the body being orbited: for example, apastron for an orbit round a star; apojove for an orbit around Jupiter.

See also: aphelion, apogee.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apogee

The point in its orbit around the Earth at which the Moon or an orbiting spacecraft is farthest from the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apollo

US programme to land men on the moon. Included 11 manned missions, October 1968–December 1972, with three missions restricted to a lunar flyby or orbital survey (Apollos 8, 10 and 13), and six landings (Apollos 11, 12, 14, 15, 16 and 17). Returned 385 kg of lunar soil and rock samples which provided evidence that the Moon was about the same age as the Earth and probably originated from material derived from Earth during a gigantic impact event. Astronauts deployed a wide variety of surface experiments including seismometers, laser reflectors for refining the Earth–Moon distance, heat flow experiments, magnetometers, ion detectors and solar wind spectrometers. Orbital surveys were conducted from the command modules, most notably on Apollos 15–17, which carried mapping and panoramic cameras, and various spectrometers. Data on the distribution of lunar gravity variations were obtained by studies of two subsatellites deployed from Apollos 15 and 16.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apollonius [Apollonios] of Perga (200–100 BC)

Greek geometer from Perga, now in Turkey, who proposed that the planets revolved around the Sun, and the Sun revolves around the Earth. Apollonius left no astronomical works but is believed to be the inventor of the geometric system of epicycles and eccentric circles, by which perfect circular motions were combined into the actual elliptical orbits of the planets. His work was developed by HIPPARCHUS and PTOLEMY. Apollonius, known as 'The Great Geometer', also wrote a monumental treatise on conic sections entitled *On Conics* in which the terms *parabola*, *ellipse* and *hyperbola* were first used. In a book of optics called *On the Burning Mirror*, he showed that spherical mirrors do not focus parallel rays of light and discussed instead the use of parabolic mirrors.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apparent Solar Time

Time reckoned by the apparent position of the Sun in the sky. The local apparent solar time (or local apparent time) is defined to be the local hour angle of the Sun (the angle between the observer's meridian and the Sun, measured westwards from the meridian) plus 12 hours. Thus, when the Sun is crossing the meridian at noon, its hour angle is zero, and the local apparent solar time is 0 h + 12 h = 12 hours; when the hour angle of the Sun has increased to 6 hours, the local apparent solar time is 6 h+12 h = 18 h, and so on. Apparent solar time is the time that is displayed on a sundial. Compared to another, regular clock such as a pendulum, it is fast or slow by up to 15 min.

See also: day, hour angle, mean solar time, meridian, sidereal time, Universal Time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apparition

The appearance of a celestial body at a time when it is well placed for observation. The term is used especially for objects in the solar system whose orbits are such that they are unobservable for periods of time: for the planets Mercury and Venus, whose morning and evening apparitions occur when they are at greatest elongation (their maximum angular separation from the Sun), and for long-period comets when their orbits bring them into the inner solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Appleton, Edward Victor (1892–1965)

Radiophysicist, born in Bradford, Yorkshire, England, Nobel prizewinner 1947, professor at London, Cambridge and Edinburgh. Discovered the existence of a layer of electrically charged particles about 200 km high in the upper atmosphere (named the *Appleton layer* after him) which, with the Heaviside layer, channels propagating radio waves between distant stations. The layer is affected by the position of the Sun, the sunspot cycle and the solar–terrestrial 'weather'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Appulse

An apparently close approach between two objects as seen on the celestial sphere, as their motions bring them close to the same line of sight of an observer on the Earth. Examples are when the Moon or a planet narrowly misses occulting a star.

See also: occultation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Apus

(the Bird of Paradise; abbrev. Aps, gen. Apodis; area 206 sq. deg.) A southern constellation that lies between Triangulum Australis and the south celestial pole, and culminates at midnight in late May. It was first shown on Petrus Planicus's celestial globe of *c*. 1598 as Apis Indica (the Indian Bird), though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, the brightest stars in Apus are α Apodis, magnitude 3.8, and β apodis, magnitude 3.9. There are no other stars brighter than fourth magnitude. Interesting objects include δ Apodis, a wide double star with red (M5) and orange (K3) components, magnitudes 4.7 and 5.3, separation 103″, and NGC 6101, a ninth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aquarius

(the Water-bearer; abbrev. Aqr, gen. Aquarii; area 980 sq. deg.) A southern zodiacal constellation that lies between Pegasus and Piscis Austrinus, and culminates at midnight in late August. Its origin dates back to Babylonian times and it is said to represent Ganymede in Greek mythology, who was snatched up by Aquila (the eagle) on Zeus's behest, to become cup-bearer to the gods. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest* and he included α Piscis Austrini (Fomalhaut) in the constellation.

A large but rather inconspicuous constellation, the brightest stars in Aquarius are α Aquarii (Sadalmelik) and β Aquarii (Sadalsuud), both magnitude 2.9. All other stars are of the third magnitude or fainter. Interesting stars in Aquarius include several doubles, notably ζ Aquarii, a close binary with yellow (F3 and F6) components, magnitudes 4.4. and 4.6, separation 2″, period 856 years, and the Mira-type variable R Aquarii (range 5.8–12.4, period about 387 days). Other interesting objects include M2 (NGC 7089), a seventh-magnitude globular cluster, and two planetary nebulae, NGC 7293 (the Helix Nebula), also seventh magnitude and one of the closest planetary nebulae, with an apparent size of 12.8′, and NGC 7009 (the Saturn Nebula), which is eighth magnitude.

Three major meteor showers appear to radiate from the constellation: the Eta Aquarids, Delta Aquarids and Iota Aquarids.

See also: Delta Aquarids, Eta Aquarids, Helix Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aquila

(the Eagle; abbrev. Aql, gen. Aquilae; area 652 sq. deg.) An equatorial constellation that lies between Sagitta and Sagittarius, and culminates at midnight in mid-July. Its origin dates back to Babylonian times and it is said to represent the eagle of Zeus in Greek mythology, which carried the thunderbolts that Zeus hurled at his enemies and which snatched up Ganymede to become cup-bearer to the gods. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest* and he placed the southern ones in the now obsolete constellation of Antinous.

The brightest star in Aquila, α Aquilae (Altair), magnitude 0.8, is one of the stars that, together with α Lyrae (Vega) and α Cygni (Deneb), makes up the asterism of the Summer Triangle. Two fainter stars, γ Aquilae (Tarazed), magnitude 2.7, and β Aquilae (Alshain), magnitude 3.7, lie about 2° north-west and 2.5° south-east respectively of Altair, and together with it form an unmistakable configuration. There are seven other stars of magnitude 4.0 or brighter, including η Aquilae, one of the brightest Cepheid variables (range 3.5–4.4, period 7.18 days). Another bright variable is the Mira-type star R Aquilae (range 5.5–12.0, period about 284 days). The constellation also contains Van Biesbroeck's star, an eighteenth-magnitude red dwarf which, with an absolute magnitude of 19.3, is one of the intrinsically faintest stars known.

The Milky Way passes through Aquila and the constellation contains a number of star clusters and planetary nebulae, the brightest of which is NGC 6709, an open cluster of about 40 stars fainter than ninth magnitude.

See also: Altair

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ara

(the Altar; abbrev. Ara, gen. Arae; area 237 sq. deg.) A southern constellation that lies between Scorpius and Apus, and culminates at midnight in early June. It is said to represent the altar upon which in Greek mythology the gods swore allegiance before their battle against the Titans, and that upon which Chiron, the centaur, is about to sacrifice Lupus, the wolf (which was a neighboring constellation before the introduction of Norma in the mid-eighteenth century). Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small, rather inconspicuous constellation, the brightest stars in Ara are α and β Arae, both magnitude 2.8. There are five other stars brighter than fourth magnitude. The Milky Way passes through Ara and the constellation contains a number of star clusters, including NGC 6193, a fifth-magnitude open cluster, and NGC 6397, one of the closest globular clusters, which is just visible to the unaided eye as a faint misty patch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arago, Dominique François Jean (1786–1853)

Scientist and statesman, born in Estagel, France. Director of the Paris Observatory, where he invited URBAIN LE VERRIER to determine why Uranus was deviating from its predicted orbit, a study which led to the discovery of Neptune. His scientific achievements were mainly in the fields of astronomy, magnetism and optics. He showed that the light changes polarization when passing through quartz and performed experiments that failed to detect the motion of the Earth with respect to the ether. He took part in the July Revolution (1830), and in 1848 was a member of the provisional government, and was effectively the Head of State for 46 days (perhaps only the second astronomer who has had this distinction, counting ALFONSO X as the first). Working with JEAN-BAPTISTE BIOT, Arago defined the meridian of Paris, used by French sailors until supplanted by the 'prime' meridian of Greenwich in 1884. He made measurements along it of the circumference of the Earth as a preliminary to the determination of the meter, in order to standardize the unit of length. A 'virtual monument' to Arago by the Dutch sculptor Jan Dibbets runs through Paris in the form of 135 brass circles engraved with his name and located along the meridian line.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aratus [Aratos] (*c*. 315–*c*. 245 BC)

Greek poet and naturalist, author of *The Phenomena*, an account of the constellations and their positions, and of weather signs in heavenly bodies and other natural phenomena. It is based on work by the astronomer EUDOXUS and on a lost treatise of the fourth century BC, perhaps by Theophrastus. Aratus' work is the earliest surviving account of the constellations still in use at the present day.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arcetri Astrophysical Observatory

The Arcetri Astrophysical Observatory, a government research institute founded in 1972, is located close to the villa where Galileo spent the last 11 years of his life. Under the directorship of Giorgio Abetti (1921–53) it became the growth point of Italian astrophysics with emphasis on solar physics; a tradition continued by his successor Guglielmo Righini (1953–78). Since 1978 the activities have also included: interstellar medium and star formation, extragalatic and high-energy astrophysics, astrophysical techniques and large telescopes.

The Observatory collaborates with the National Research Council in the operation of the 0.5 m Infrared Telescope on the Gornergrat (Swiss Alps) and is the Italian focal point in the collaboration between institutions in Germany, Italy and the United States for the construction of a 2×8.4 m binocular telescope on Mount Graham (Arizona).

Arcetri's staff includes about 45 tenured astronomers and a large number of visiting scientists.

For further information see
http://www.arcetri.astro.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Archelaus (fifth century BC)

Greek philosopher and cosmologist. Pupil of ANAXAGORAS and teacher of Socrates, Archelaus believed that air is the primitive substance, which formed water at the center of the universe. Part condensed to Earth, pieces of which broke off to become the stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Archimedes of Syracuse (287–212 BC)

Greek scientists and mathematician, born in Syracuse. Remembered for *Archimedes' screw*, used for raising water, and his cry of *eureka* ('I have found it') when he discovered, according to legend while in the bath, the principle of the hydrostatic upthrust on a floating body (*Archimedes' principle*). Discovered formulae for the areas and volumes of spheres, cylinders and a variety of figures. Most of his original astronomical work is lost. In a curious book he attempted to calculate an upper limit to the number of grains of sand which would fill the universe. He was killed, ridiculously, at the siege of Syracuse by a Roman soldier whose challenge he ignored while immersed in a mathematical problem drawn in the sand.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arcturus

The star α Boötis, a golden yellow giant star (spectral type K2IIIp), with an apparent magnitude of -0.05 it is the fourth brightest star in the sky. Its distance is currently 37 light-years (parallax $0.089''$), but it has been calculated that its high velocity through space will carry it past the solar system in a few thousand years. It will be a spectacular visitor, being about 21 times the Sun's diameter and 98 times its luminosity, having an absolute magnitude of -0.3.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arecibo Observatory

Located in the limestone hills of northwestern Puerto Rico, Arecibo Observatory is operated by the National Astronomy and Ionosphere Center, managed by Cornell University for the National Science Foundation.

Conceived by William E Gordon, a professor of electrical engineering at Cornell University, the radio-radar telescope was completed in 1963. With its 305 m (1000 ft) diameter dish, which was built into a hillside, Arecibo has the largest collecting area (8 hectares or 18 acres) of any radio antenna in the world.

The Arecibo system operates at frequencies between 50 MHz and 10 GHz (wavelengths between 6 m and 3 cm). With the aid of highly sensitive receivers which are cooled to about 20 K it is able to study extremely weak radio-emitting objects.

Arecibo's spherical reflector requires a 29 m (96 ft) line feed to collect the partially focused radiation. A Gregorian feed system, which uses two mirrors to focus radio waves to a point, was installed in 1997. At the same time, a ground screen and powerful new radar transmitter were installed.

A number of major discoveries have been made with the telescope. They include the first discovery of planets around a pulsar (B1257+12); the first discovery of a pulsar in a binary system; and the discovery of millisecond pulsars. During the 1980s it was used to map the distribution of galaxies in the universe.

Since the first radar transmitter was installed in 1974, Arecibo has been used for solar system exploration, including detailed mapping of the surfaces of the Moon, Venus and a handful of near-Earth asteroids. In the 1960s it enabled astronomers to determine the rotational period of Mercury, and more recently it has been used to search for ice in polar craters on both Mercury and the Moon. The Arecibo radar is also the world's most sensitive instrument for detecting meteors.

During the dedication of a major upgrade to the telescope on 16 November 1974, Arecibo sent the first radio message to extraterrestrials. The transmission contained representations of the fundamental chemicals of life, the formula for DNA, a crude diagram of our solar system and simple pictures of a human being and the Arecibo telescope.

The telescope is still used for SETI research today. One project, known as Phoenix, aims the telescope at specific stars. Another, called Serendip, collects data on certain likely frequencies during all the telescope's other operations, and distributes the data to thousands of volunteers for processing on personal computers.

For further information see
http://www.naic.edu/open.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Argelander, Friedrich Wilhelm August (1799–1875)

Astronomer, born in Memel, East Prussia. Director of the Bonn Observatory, where he organized a survey of the position of all 324 198 stars of the northern hemisphere above the ninth magnitude, published as the star charts and catalogs of the *Bonner Durchmusterung* (BD). His assistant, Eduard Schönfeld (1828–91), made the extension (BDE) into the southern sky. It is staggering to think about this pre-information technology project to observe half a million stars twice each by eye, their brightnesses and positions calculated, charted and printed by hand.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Argo Navis

(the Ship) a large southern constellation representing the ship *Argo* of Jason and the Argonauts in Greek mythology whose brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*. It was divided up by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, into the constellations Carina (the Keel), Vela (the Sails), Puppis (the Poop or Stern) and Pyxis (the Compass), the stars of which all share a single series of Bayer letters. The brightest star of Argo Navis is now α Carinae (Canopus or Suhel), magnitude -1.9.

See also: Carina, Puppis, Pyxis, Stellar nomenclature, Vela.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ariel

A mid-sized satellite of Uranus, discovered by William Lassell in 1851. Its diameter is 1160 km and it orbits at a distance of 191 000 km. Of Uranus's five mid-sized satellites, Ariel has the brightest surface, and has been the most geologically active (Miranda excepted). An older cratered surface is criss-crossed by spectacular deep, steep-sided canyons whose broad floors themselves show signs of further activity, with volcanic flooding and sinuous, winding troughs reminiscent of the Moon's sinuous rilles. (Such low-temperature 'cryovolcanism' is possible because a water–ammonia eutectic mixture can behave as a viscous lava.) The longest of these canyons is the 622 km long system known as Kachina Chasmata. Such activity must have been caused by tidal heating. At present Ariel lacks any orbital resonance with neighboring satellites that could cause tidal heating, but this may have been different in the past. When the molten icy interior eventually solidified, it would have expanded, splitting the crust to form the canyons.

See also: Uranus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ariel

Series of six UK-led scientific satellites launched 1962–79. The first four studied the ionosphere, while the last two were x-ray astronomy satellites. Ariel 3 was the first all-British satellite. The UK–US Ariel 5, launched in October 1974, operated for more than five years and discovered dozens of new x-ray sources, including a number of transient sources such as Cygnus X-1. Ariel 6 suffered technical problems and produced few results.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aries

(the Ram; abbrev. Ari, gen. Arietis; area 441 sq. deg.) A northern zodiacal constellation that lies between Taurus and Pisces, and culminates at midnight in late October. It represents the ram in Greek mythology whose golden fleece was the quest of Jason and the Argonauts. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*. In Ptolemy's day the Sun was in Aries at the vernal (spring) equinox and although as a result of precession this position now lies in the neighboring constellation of Pisces, it is still sometimes called the first point of Aries.

A rather inconspicuous constellation, the brightest stars in Aries are α Arietis (Hamal), magnitude 2.0, and β (Sheratan), magnitude 2.6. γ Arietis (Mesarthim) is a fine visual binary with white (B9) and pale yellow (A1) components, both of magnitude 4.6, separation $8''$. It was one of the first double stars to be discovered, being found serendipitously by the English physicist Robert Hooke (1635–1703) while observing the comet of 1664. α, β and γ Arietis form a narrow triangle, which is the most easily recognized part of the constellation.

Variable stars in Aries include the Mira-type stars U Arietis (range 7.2–15.2, period about 371 days) and R Arietis (range 7.4–13.7, period about 187 days). The brightest galaxy is NGC 772, a tenth-magnitude spiral.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aristarchus of Samos (310–230 BC)

Greek philosopher, born on the island of Samos. According to ARCHIMEDES, Aristarchus hypothesized that the fixed stars are stationary, and that the Earth orbits the Sun, which is at the center of the universe. Since the stars did not show motions that reflected the motion of the Earth, the stars must be at great distances. The Babylonian astronomer SELEUCUS was the only astronomer who took up this concept until COPERNICUS in 1543, who knew something of Aristarchus' theory (*see* THOMAS KUHN). Aristarchus also used geometric methods to measure the size of the Sun, Moon and Earth. His methods were correct in principle and inaccurate only by an unprecedentedly small one order of magnitude, due to failures in the instruments and assumptions rather than the reasoning. According to Plutarch, Aristarchus's contemporaries rejected 'with abhorrence the view of those who have brought to rest the things which move and set in motion the things which by their nature and position are unmoved, such a supposition being contrary to the hypotheses of mathematics'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aristillus (*c*. 200 BC)

Greek philosopher who worked in the Library at Alexandria and who with TIMOCHARIS, prepared the first star catalog.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aristotle of Stagira (384–322 BC)

Aristotle was a Greek philosopher and scientist, born in Stagira. Aristotle's lectures were compiled into 150 volumes, including *Physics*, *Metaphysics* and *De Caelo et Mundo* (*On the Heavens and Earth*), in which he accepted the heavenly spheres of EUDOXUS, thought the Earth to be spherical and imagined a perfectly spherical, unchanging universe of spheres centred on the Earth and carrying the planets. Aristotle followed Eudoxus in suggesting that there were four elements: earth (solid), fire (energy), water (liquid) and air (gas). Each element could be hot, wet, dry or cold. Aristotle added a fifth element called ether, which he believed to be the main constituent of the celestial bodies. This divine element, he believed, is uncompounded, ungenerated, eternal, unalterable and neither heavy nor light. It can be found in its purest form in the celestial regions, but becomes adulterated in the area below the Moon. Aristotle's view of the universe was hierarchical, and he made a sharp distinction between the eternal and immutable heavens and the sublunar, imperfect world of change (hence the subsidiary meaning of the word 'sublunary'). The motion of celestial objects was therefore always natural, circular and continuous. In medieval times, Aristotle's cosmology was codified and systematized to become the dominant world view (see MEDIEVAL COSMOLOGY).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Armagh Observatory

The Armagh Observatory is a modern astronomical research institute with a rich heritage. Founded in 1790 by Archbishop Richard Robinson, its mission is to advance the knowledge and understanding of astronomy and related sciences. Around 20 astronomers carry out multi-disciplinary research on problems concerning the Sun and other stars, star formation, comets and asteroids, extra-solar planetary systems and solar system–terrestrial relationships.

The Observatory is supported by the Department of Education Northern Ireland and the UK PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL. It is situated close to the center of the city of Armagh in Northern Ireland in grounds that include the Armagh Planetarium and the Armagh Astropark.

For further information see http://www.arm.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Armillary Sphere

A device of ancient origin that was used to measure or describe celestial positions. In essence, it is a model of the celestial sphere consisting of a set of rings, each of which represents a great circle such as the celestial equator or ecliptic, that revolves within a framework of fixed circles that represent, for example, the observer's horizon and meridian. When used for observation, it included a pivoted ring on which open sights were mounted, the position of the object of interest being read off from graduated scales on the various rings.

Armillary spheres were used for positional measurements in ancient and medieval times by observers such as Hipparchus, Ptolemy and Tycho Brahe and were widely used from the Middle Ages onwards for the teaching of astronomy.

See also: celestial sphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Armstrong, Neil Alden (1930–)

Astronaut, born in Auglaize County, OH, trained as a pilot (Korean War). He flew on Gemini 8 and in 1969, with Michael Collins and EDWIN ALDRIN, took Apollo 11 to the Moon. On 21 July 1969, at 02:56 GMT, Armstrong became the first man to walk on the Moon, with the famous statement, 'That's one small step for a man, one giant leap for mankind'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arrhenius, Svante August (1859–1927)

Swedish chemist, Nobel prizewinner (1903). In *Worlds in the Making* (1908), he suggested that life on Earth had begun when space-traveling spores reached Earth on a meteorite—the panspermia hypothesis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Art and Literature

The mind-expanding properties of astronomy affect writers and artists as much as (or possibly more than) people in general, and there is a long tradition of the use of astronomical references in literature and art.

Astronomy and art

Among the intents of fine art (painting, sculpture and objects which are made to be admired, rather than to be useful) is to comment by metaphor or analogy on human activities. Such activities may be set against or associated with astronomical phenomena, either for their own interest or as a comment.

In paintings of the crucifixion, for example, are often shown images of eclipses. These are literal representations of the sudden darkness that the scriptures say occurred at the death of Jesus. They are also metaphors for the extinguishing of an important person, and examples of the 'pathetic fallacy' in which the natural world is identified as reacting to the human activities which happen in it.

Likewise, the Florentine painter Giotto de Bondonel was commissioned in 1303 to decorate the interior of a chapel belonging to a wealthy Paduan merchant, Enrico Scrovegni, and he painted 38 religious scenes, one of which was *The Adoration of the Magi* (see also http://www.mystudios.com/gallery/giotto/17a.html). The worship of the infant Jesus in a stable was set under the Star of Bethlehem, a symbol of change and a portentous celestial mark of the birth. Giotto depicted it as a blazing comet. He was inspired by Halley's Comet, which had returned to perihelion in 1301 and which the painter had undoubtedly seen. It was fitting that the spacecraft sent into the nucleus of Halley's Comet during the return of 1986 should have been named GIOTTO.

On a more individual level, Vincent van Gogh painted *Starry Night* (1889, Museum of Modern Art, New York City; see also http://sunsite.org.uk/cgfa/gogh/gogh3.htm). It shows a Provençal village with a church and houses with lit windows in a landscape of hills, olive and cypress trees, shown at night below a starry sky, with crescent moon and swirling Milky Way. It appears that the skyscape is a free rendering of the night sky at the time of its painting. Whatever its natural realism, the cosmic background of dizzying lights, overarching the human landscape below, represents a disturbing cosmic background to the smaller human events below, including, one must suppose, Vincent's developing madness, culminating in his suicide the following year.

The same Milky Way runs across the night sky in Albrecht Altdorfer's *Flight into Egypt* (1609, in the Pinakothek at Munich). Again, it is a cosmic background to the events, shown small, that are taking place on Earth below as the Holy Family travels on donkey, fleeing from the massacre of the infants by Herod. Altdorfer's is the first representation of the Milky Way as a mass of individual stars.

Astronomical instruments are often shown in portraits. They are interesting decorations that reflect the interests, attainments and education of the sitter(s) and may or may not be directly associated with the subject. *The Ambassadors* (1533, National Gallery, London; see http://www.nationalgallery.org.uk/collection/w_wing/index.html) is a full-length portrait of two wealthy, educated and powerful young men, Jean de Dinteville, the French Ambassador to England, and Georges de Selve, Bishop of Lavaur. The artist, Hans Holbein the Younger (1497/8–1543), has depicted the two men standing either side of a table bearing a collection of globes, books and astronomical and musical instruments. A sundial on the table is identical to that featured in Holbein's portrait of the astronomer *Nicholas Kratzer* of 1528 (Musée du Louvre, Paris; see http://www.louvre.fr img/photos/collec/peint/grande/inc1343.jpg). The plausible guess is that the sundial was designed or made by Kratzer, and its image was re-used and incorporated by Holbein into *The Ambassadors* as flattery.

Astronomical references in the illustrated books of the English romantic poet William Blake are not favorable to science. *God as an Architect*, an illustration from *The Ancient of Days*, depicts God as an old man, kneeling on stormy clouds against a fiery sun, coldly measuring the cosmos with dividers (see http://sunsite.org.uk/cgfa/blake/index.html). Blake depicted *Newton* (see http://www.cube.icestorm.net/peirson/phil/blake/blake2.html) as a misguided Greek-god-like hero whose gaze was directed only downwards at sterile geometrical diagrams that he also measures with dividers. It is curious that this scornful depiction of Newton is often seen displayed as a poster on the walls of university physics departments.

By contrast to Blake's pessimism, his contemporary and countryman, Samuel Palmer, shows bucolic scenes in an idealized England (e.g. worshippers *Coming from Evening Church*, 1830, *A Hilly Scene* circa 1826–1828, Tate Gallery, London; see http://www.tate.org.uk/coll/cchtm/n03697_c.htm and http://www.tate.org.uk/coll/cchtm/n05805_c.htm) under a benign, evening sky with glowing moons, and the planet Venus.

Astronomy and literature

Just as artists have used astronomical themes as a backdrop to human affairs, so too have writers. Astronomical references abound in the writings of William Shakespeare (1564–1616). His plays contain numerous speeches for and against astrology, views against being expressed mostly by the more pushy, rebellious characters. Shakespeare was born after Copernicus published his theory of the solar system in 1543 and was a contemporary of enthusiastic English

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

proponents of the new view of the Universe, such as THOMAS DIGGES (died 1595), WILLIAM GILBERT (1540–1603) and JOHN DEE (1527–1608). Nevertheless, no trace of this controversy is found in Shakespeare's works, save in disputes about the validity of astrology.

Thomas Hardy (1840–1928) was a poet and prolific novelist of the Wessex countryside. He described his novel *Two on a Tower* as a 'slightly-built romance' that 'was the outcome of a wish to set the emotional history of two infinitesimal lives against the stupendous background of the stellar universe, and to impart to readers the sentiment that of these contrasting magnitudes the smaller might be the greater to them as men' (preface of edition of 1895). Peter Ackroyd (1949–), an English novelist, known for his handling of the relativity of time in his work, also set his novel *First Light* in a west-country landscape which shows a palimpsest of time in its tumuli, folk memories and modern observatory.

It is possible in literature to make explicit comment on quite abstract astronomical ideas. The size of the universe is one of astronomy's imaginative pulls, and the very word 'astronomical' has gathered a subsidiary meaning of 'immense' because the quantities in astronomy involved are so vast. Sometimes the reaction to the distances of stars is awe and fear, as in *Pensées* by Blaise Pascal (1623–1962):

> The eternal silence of those infinite spaces strikes me with terror.

Because of the long times for the travel of light from a star to us, the image of the sky that we perceive at this moment originated long ago and may show the memory of a star that has since reached the end of its life and no longer exists. The *Ode to Charles Sumner* by Henry Longfellow (1807–1882) (1979 *Complete Works* (Oxford: Oxford University Press) p 326) illustrates this idea:

> Were a star quenched on high,
> For ages would its light,
> Still travelling downward from the sky,
> Shine on our mortal sight.
>
> So when a great man dies,
> For years beyond our ken
> The light he leaves behind him lies
> Upon the paths of man.

American poet Robert Frost (1874–1963) had a lifetime interest in astronomy, which started from the age of 15 when he installed in the upper room of his parents' house a small telescope, bought with money from selling magazine subscriptions. In a speech two months before his death, he gave his philosophical reaction to astronomy: 'How stirring it is, the Sun and everything. Take a telescope and look as far as you will. How much of the universe was wasted just to produce puny us. It's wonderful … fine'. In *A Star in a Stone Boat* (Frost R 1962 (Harmondsworth: Penguin) p 116) he described how celestial material was part of the fabric of the Earth and its contents.

> Never tell me that not one star of all
> That slip from heaven at night and softly fall
> Has been picked up with stones to build a wall.

Primo Levi (1919–1987) was an Italian Jewish poet and writer. Trained as an industrial chemist, he had a dramatic life as a Piedmont partisan, and was a witness, participant and survivor from the concentration camp at Auschwitz. He committed suicide, apparently in despair at the human condition. His poetry *In the Beginning* from 1998 *Collected Poems: Shema* transl R Feldman and B Swann (London: Faber and Faber) p 27) describes the same thought as Frost, in relation to the big bang:

> From that one spasm everything was born:
> The same abyss that enfolds and challenges us,
> The same time that spawns and defeats us,
> Everything anyone has ever thought,
> The eyes of a woman we have loved,
> Suns by the thousands
> And this hand that writes.

However again, just as in fine art, not all references in literature to astronomy are favorable, and Newton is in the firing line. At a drunken dinner on 28 December 1817, attended by four literati—Benjamin Haydon, William Wordsworth, Charles Lamb and John Keats—Lamb abused Haydon for putting a bust of mathematical scientist Isaac Newton into a picture that he had just painted. Lamb derided Newton as 'a fellow who believed in nothing unless it was as clear as three sides of a triangle'. Lamb proposed a toast to 'Newton's health and confusion to mathematics'. Keats agreed, and, with Wordsworth, lamented the destruction of beauty by science. This derision of science appears in Keats' and Wordsworth's poetry. No scientist would agree with the sentiment expressed by Keats (*Lamia Part II* lines 229–38 in Garrod H W (ed) 1956 *The Poetical Works of John Keats* (Oxford: University Press) p 176):

> Do not all charms fly
> At the mere touch of cold philosophy?
> There was an awful rainbow once in heaven:
> We know her woof, her texture; she is given
> In the dull catalogue of common things.
> Philosophy will clip an Angel's wings,
> Conquer all mysteries by rule and line,
> Empty the haunted air, and gnomed mine—
> Unweave a rainbow…

Science fiction

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

Astronomy is associated with a distinct literary genre, namely science fiction. 'Science fiction is that class of prose narrative treating of a situation that could not arise in the world we know, but which is hypothesised on the basis of some innovation in science or technology, or pseudo-technology, whether human or extra-terrestrial in origin' (Amis 1960).

Science fiction has a long and honorable history. It goes back to the 2nd century AD, since the first true novel about an expedition to the Moon was written at that time by a Greek satirist, Lucian of Samosata. He called it the *True History*, because it was made up of nothing but lies from beginning to end. It describes things that he had 'neither seen nor suffered nor learned from another, things which are not and never could have been; therefore my readers should by no means believe them'. Lucian combined a cool brain with a gift for fluent, easy writing, and he possessed a strong sense of humor.

His travelers are sailors, who were caught up in a water spout as they passed through the Pillars of Hercules (our Straits of Gibraltar), and were hurled upward so violently that after seven days and seven nights they landed on the Moon. They were arrested by lunar warriors and imprisoned by the King of the Moon. Associating the Aristotelian perception of celestial bodies as more perfect than the Earth with the Moon, Lucian described the Moon-men as far more advanced than the people of Earth were. Anything unclean or impure was abhorrent to them. Sex was either unknown or ignored, and when a Moon-man died he merely dissolved into smoke, so that no remains should be left for burial. This convention that celestial beings are more pure than we are survives to this day in beliefs about extraterrestrial visitors in UFOs.

Johann KEPLER's *Somnium (Dream)* was very different. In 1593, as a 22 yr old student at the University of Tübingen, Kepler wrote a dissertation, from a Copernican standpoint, on how the heavens would appear to an observer on the surface of the Moon. Debate on this radical thesis was suppressed. *Somnium*, written by 1609 and published in 1634, expands the thesis into a work of imagination. It was, and was meant to be, educational as well as entertaining. There were 223 notes explaining its references. In its emphasis on what to Kepler was science rather than on the adventure story, the work was the first science fiction story (hard core, in which the science is the most important feature, as opposed to science fantasy). Kepler's hero, an Icelander named Duracotus, is carried to the Moon by a demon. Kepler knew that the Earth's atmosphere does not extend all the way to the Moon, and that there must be a neutral point where the gravitational pulls of the Earth and Moon balance. Kepler also knew that the Moon always keeps the same face turned toward the Earth, and explains that conditions on the two hemispheres are quite different. He describes a lunar eclipse as seen from the Moon (i.e. an eclipse of the Sun by the Earth) and what a solar eclipse looks like to a moon-dweller (i.e. the transit of the umbral shadow across the face of the Earth).

Jules Verne's *From the Earth to the Moon* appeared in 1865 and was followed later by the sequel, *Round the Moon*. Verne made a conscious effort to keep to the facts as he knew them, and on the whole he succeeded remarkably well, even though his basic method of travel was wrong. His travelers were fired to the Moon from the barrel of a huge cannon, and set off at a speed of 7 miles^{-1}. Numerically Verne was correct; 7 miles s^{-1} is the Earth's escape velocity, and a projectile launched at this speed would never return. Unfortunately, the projectile would at once be destroyed by friction against the atmosphere, and in any case the shock of departure would certainly reduce any travelers to jelly, but at least Verne made it all sound plausible. His description of the Moon was based on the best information available, and his plot is ingenious; the projectile encounters a minor satellite of the Earth, and its path is changed so that instead of reaching the Moon the travelers are boomeranged back to the Earth.

H G Wells' *The War of the Worlds* (1898) describes how Earth is invaded by grotesque monsters from Mars, fleeing the desiccation of their planet. They cause devastation until they are destroyed by terrestrial bacteria, against which they have no immunity. In 1938 a radio broadcast of the novel, produced by Orson Welles, caused widespread panic in parts of the USA, listeners mistaking it for a real news bulletin.

A novel of very different kind was written by KONSTANTIN EDUARDOVICH TSIOLKOVSKY, born in 1857 in Ijevsk, a remote village in Russia. He realized that the only practicable method for space travel is the rocket, which depends on the principle of reaction and can work in the vacuum of space. His only novel, *Beyond the Planet Earth*, was probably complete by 1895, although it did not appear in print until 1920 and its English translation was delayed until 1960. As a story, and as a literary effort, it can only be described as atrocious, but as a scientific forecast it was years ahead of its time. In his novel, the travelers used a liquid-fuel rocket motor instead of solids such as gunpowder; he described the 'step' principle of mounting one rocket on top of another (what we would call a multistage rocket). He also gave a perfectly accurate description of the causes and effects of weightlessness, or zero gravity. He was equally aware of the many problems involved in long space journeys, and he proposed to take along various types of green plants to remove excess carbon dioxide from the atmosphere inside the space-craft and replace it with free oxygen.

During the 1930s large numbers of low-cost science fiction (and detective story) magazines appeared. This generated the name 'pulp fiction', now a synonym for story-telling without sophistication. Nevertheless, some young and now established writers started in this market.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

By contrast to the bulk of pulp fiction, among the recent authors whose books contain a great deal of sound science, are the space travel visionary Sir ARTHUR C CLARKE, author of *2001: a Space Odyssey*, filmed by Stanley Kubrick. The astronomer FRED HOYLE is also an established writer, whose first and best known work of science fiction, *The Black Cloud*, includes mathematical formulae.

'Space art'

Like science fiction, 'space art' is intended mostly to convey thoughts about space and astronomy, rather than demonstrate the finer points of an art-form standing in its own right. Space art is a form of scientific illustration, akin to botanical illustration. Since the subject matter is not readily seen directly, space artists extrapolate what is known about a subject into an illustration, much as an illustrator would do when reconstructing a dinosaur from fossils or a historical site from an archaeological dig. Space artists might show what a far-off planet, double-star system or galaxy would look like if we were there in person, or how an activity in space might be implemented, such as mining an asteroid or colonizing Mars.

Space art may also incorporate fantasy elements into its genre, placing fish, flowers, etc, into a space scene for dramatic effect. An early example of this form of space art was *The Sower of Systems* (1902, The Watts Gallery, Compton; see also http://www.webmagick.co.uk/prcoll/artists/watts/index.html), a representation of the chaos prior to the formation of planets by George Frederick Watts (1817–1904). An expressionist image of the Laplace nebular hypothesis is combined with the vague form of a striding, grey-cloaked figure, as if sowing a field.

The pioneer space artist in the modern school was CHESLEY BONESTELL, who illustrated books by space-travel experts such as WERNHER VON BRAUN and Willy Ley. Modern space art is distinguished by a naturalistic method of illustration, creating detailed landscapes reminiscent of those of the American Romantic landscape painters. The subject matter and the material available both favor digital methods of work, but more traditional techniques are also common. About 120 space artists form the International Association of Astronomical Artists (http://www.iaaa.org/), founded in 1982. See also http://www.spaceart.org/index.html. Modern practitioners include David Hardy and Paul Doherty (UK), Lynette Cooke (USA), and cosmonauts Vladmir Dzhanibekov and Alexei Leonov, with Georgii Poplavski and Andrei Sokolov (Russia).

There is endless scope for space art. In the future it will be fascinating to look back at some of the speculations of the 20th century; for example, will a lunar base really be like the graceful, domed structure visualized by Bonestell? We wait and see.

Bibliography

Allen R H 1963 *Star Names and their Meanings* (New York: Dover)

Bush D 1950 *Science and English Poetry* (New York: Oxford University Press)

Byard M M 1977 Poetic responses to the Copernican revolution in *Sci. Am.* **236**(6) 120

Clark C 1921 *Astronomy in the Poets* (Bournemouth: Sydenham & Co)

Fauvel J, Flood R, Shortland M and Wilson R (ed) 1988 *Let Newton Be!* (Oxford: Oxford University Press)

Heath-Stubbs Jean and Salman P (ed) 1984 *Poems of Science* (Harmondsworth: Penguin)

Knuijt J K 1989 *Poetry of the Heavens* (Hortonville: Mira)

Long C H 1963 *The Myths of Creation* (New York: Collier)

Maclagan D 1977 *Creation Myths* (London: Thames and Hudson)

Marschall A 1983 Modern poetry and astronomy *Mercury* (March/April) p 41

Meadows A J 1969 *The High Firmament* (Leicester: Leicester University Press)

Nicolson M 1956 *Science and the Imagination* (Ithaca: Cornell University Press)

Rothenberg J (ed) 1969 *Technicians of the Sacred* (New York: Doubleday Anchor)

Tomlinson G and Trainque D 1984 *Anthology of Astronomical Poetry (TIPS Booklet 12)* (Great Lakes: Planetarium Association)

Astronomy (comets) and art:

Olson R M 1985 *Fire and Ice* (New York: Walker)

Space art:

The Guild of Natural Science Illustrators 2000 *The Guild Handbook of Scientific Illustration* 2nd edn (New York: Wiley)

Science fiction:

Amis K 1960 *New Maps of Hell* (London: Gollanz)

Lear J 1965 *Kepler's Dream* (Berkeley, CA: University of California Press) (with the text and notes of *Somnium, Sive Astronomia Lunaris*, transl P F Kirkwood)

Paul Murdin

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

Artificial Satellite

A man-made object placed in orbit round the Earth or some other celestial body. The first artificial Earth satellite was Sputnik 1, launched by the then Soviet Union, on 4 October 1957. Spherical in shape, and with a mass of 84 kg, it entered an orbit with a perigee altitude of 229 km, an apogee altitude of 947 km and a period of 96 min. Artificial satellites are used for a wide variety of roles, including astronomical observation, study of the Earth's magnetosphere and space environment, monitoring the atmosphere, weather, oceans and surface of the Earth, geodesy (study of the shape and gravitational field of the Earth), communications and military reconnaissance.

All satellites follow elliptical (or circular) orbits around their parent body. The shape and orientation (i.e. the orbital elements) of a satellite's orbit undergo changes under the action of various perturbing forces. For example, the tenuous outermost layers of the Earth's atmosphere exert a small but finite frictional drag on the motion of a satellite that causes its perigee altitude to decrease (so tending to make the orbit more nearly circular) and lead eventually to its spiralling inwards and burning up in the denser atmospheric layers. Drag forces of this kind caused Sputnik 1 to re-enter the atmosphere and burn up 92 days after it had been launched. Gravitational perturbations exerted, for example, by the Earth's equatorial bulge, or by the Moon and Sun, cause the planes of satellite orbits to precess (slowly rotate around the Earth). If the orbital inclination (the tilt of the orbital plane relative to the Earth's equator) is carefully chosen, the orbit can be made to precess at a rate that ensures that its orientation relative to direction of the Sun remains the same while the Earth revolves around the Sun. An orbit of this kind is called Sun-synchronous. If a satellite is placed in a circular orbit at an altitude of just less than 36 000 km above the Earth's equator, its orbital period will be exactly equal to the rotation period of the Earth. Because the satellite then remains continuously above the same point on the equator (and remains stationary in the sky when viewed from the Earth's surface), it is said to be geostationary. The geostationary orbit is extensively used by communications and meteorological satellites.

See also: apogee, magnetosphere of Earth, orbit, orbital elements, perigee.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ASCA (Advanced Satellite for Cosmology and Astrophysics/Astro-D)

Japanese–US x-ray satellite, launched February 1993, reentered over the Pacific Ocean on 2 March 2001. During its eight-year lifetime, it provided broad-band imaging over the 1–12 keV energy range. It carried four US conical grazing telescopes and the first two CCD-based imaging spectrometers to be used on an x-ray observatory. ASCA found evidence for the creation of cosmic rays during supernovae, and played a major role in studies of x-ray binaries and gamma ray bursters. Also known as Asuka, which means 'flying bird'. (See also ASTRO.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ashen Light

An apparent slight brightening of the night side of Venus occasionally seen when the planet is observed as a thin crescent, near inferior conjunction, when it is nearest the Earth. The ashen light may first have been observed by Giambattista Riccioli in 1643, though it is not absolutely clear from his description that this is what he recorded as having seen; the first unequivocal description was by William Derham in 1715. The ashen light, like other observational phenomena on Venus long reported by amateur observers, such as cusp caps and the Schröter effect, has been a controversial subject: it has never been satisfactorily imaged, for example. It has been likened in appearance to earthshine (in which the portion of the crescent Moon's disk in shadow is illuminated by sunlight reflected off the Earth), though it is much more feeble, and of course it cannot have a comparable cause.

There are two possible physical explanations for the ashen light. First, it may be similar to the Earth's airglow, a faint but persistent illumination of the Earth's upper atmosphere. This is caused by the recombination of atoms and molecules that have been ionized by solar ultraviolet radiation (the atoms and molecules recombine with the electrons that were knocked off them in the ionization process). On recombination, the atoms and molecules emit a weak radiation at optical wavelengths. If the process is operating in Venus's upper atmosphere it may give rise to a stronger airglow, as the flux of solar radiation is stronger closer to the Sun. Alternatively, the ashen light may be the result of the refraction of sunlight in Venus's thick atmosphere. In theory, the atmosphere is so dense that 'super-refraction' can occur: light could be refracted all the way round the planet. There are those who maintain that it is a wholly subjective phenomenon, that the observer 'sees' the whole disk as illuminated when only the crescent is visible, particularly when cusp caps (brightenings of the cusps, the tips of the crescent) are pronounced.

See also: airglow, Venus: atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Association of Universities for Research in Astronomy

The Association of Universities for Research in Astronomy, Inc. (AURA) is a consortium of educational and other non-profit institutions that operates world-class astronomical observatories. Its members are 29 US institutions and five international affiliates. For the National Science Foundation, the AURA manage the National Optical Astronomy Observatories (NOAO) and THE GEMINI OBSERVATORY. For NASA, they manage the SPACE TELESCOPE SCIENCE INSTITUTE (Baltimore, MD).

NOAO operates ground-based observatories for night-time astronomy at KITT PEAK NATIONAL OBSERVATORY (Arizona) and CERRO TOLOLO INTER-AMERICAN OBSERVATORY (Chile); and for solar research, at the NATIONAL SOLAR OBSERVATORY (New Mexico) and at Kitt Peak.

Gemini is an international project among the US, UK, Canada, Chile, Australia, Brazil and Argentina to provide two 8 m telescopes (Hawaii and Chile).

STScI carries out the scientific mission of the Hubble Space Telescope—the most powerful optical/ultraviolet observatory in space.

For further information see http://www.aura-astronomy.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Asterism

A readily recognizable group or arrangement of (usually bright) stars, which are not necessarily members of a single constellation. Well-known examples are the Plough (part of the constellation Ursa Major), and the False Cross, the Summer Triangle and the Square of Pegasus, all of which comprise stars from more than one constellation. The term is also occasionally used to denote a close group of faint stars which appear to be, but are not, members of a cluster.

See also: False Cross, Plough, Square of Pegasus, Summer Triangle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Asteroid Belt

The region of the solar system, between the orbits of Mars and Jupiter, where the greatest numbers of asteroids are found, is known as the *main belt*. Asteroids that orbit in this region are known as *main-belt asteroids*. The belt does not have well-defined limits. In terms of the semi-major axes of asteroid orbits, by far the greatest concentration lies between about 2.1 and 3.25 AU. These two distances correspond to two of the main Kirkwood gaps, respectively those defined by 4:1 and 2:1 orbital resonances with Jupiter. Orbiting within the 4:1 gap are the Hungaria group, in orbits peaking at around 1.95 AU, into which asteroids of various types have been gravitationally perturbed. The inner margin of this group is sometimes taken to define the main belt's inner edge, which is not circular but defined by the markedly eccentric orbit of Mars. Outside the 2:1 resonance are found the Cybele group, spread out as far as the 5:3 resonance and peaking at 3.4 AU, and the Hilda group, clustered around the 3:2 resonance at 4 AU. The main belt's outer edge is sometimes said to be marked by lone (279) Thule, named after the remote northern land of Norse legend, which orbits at 4.27 AU, a distance corresponding almost exactly to the 4:3 resonance. Within the main belt proper are several Hirayama families of asteroids, most notably the Flora, Phocaea, Koronis, Eos and Themis groups, separated by other Kirkwood gaps. These gaps are less favored regions rather than true gaps; the main ones correspond to resonances of 7:2, 3:1, 5:2 and 7:3.

In 2002, the existence of a new family of asteroids was confirmed. It is believed to be the youngest at under 6 million years old. David Nesvorny and co-workers at the Southwest Research Institute in Boulder, Colorado, recognized the 39 or so asteroids as belonging to the same family by studying their orbits. Projecting the current orbits backward in time, the researchers found that they converge about 5.8 million years ago—so this must be when they formed from the break-up of a larger body. Of the 39, two are big asteroids. One, called (832) Karin, after which the whole cluster is named, is about 19 km across; the other, called (4507) 1990, measures about 14 km. The team estimates that the parent body was at least 24.5 km across. The size of the parent body and of the fragments it produced can provide a test of theoretical models of how impacts shatter rocky bodies. The relatively fresh surfaces of the two larger fragments can also reveal what asteroids are made of, and how rapidly weathering in space pockmarks them with small craters. The Karin collision might also have been responsible for a recently discovered band of asteroidal dust.

See also: HIRAYAMA FAMILY, KIRKWOOD GAPS.

Bibliography

Nesvorny D, Bottke, W F Jr, Dones L and Levison H F 2002 The recent breakup of an asteroid in the main-belt region *Nature* **417** 720–722

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Asteroid Discovery in History

The discovery of five of the major planets (other than Earth) in the solar system is lost to antiquity. With the acceptance of Copernicus' heliocentric model for the solar system an apparent anomaly in the spacing of the planets was noted. JOHANNES KEPLER believed that the gap between Mars and Jupiter was too large in proportion to the gaps between other pairs of neighboring planets. He concluded that an unknown planet orbited the Sun between Mars and Jupiter.

In 1766 an empirical formula for the distances of the planets from the sun was developed by Titius von Wittenburg. The formula was popularized by Johann Bode and is known today as BODE'S LAW or the Titius–Bode law. The formula is of the form:

$$A = 0.4 + 0.3 \times 2^n$$

where A is the distance from the sun (in units of the Earth's distance) and $n = -\infty, 0, 1, 2, \ldots$. The values of A determined by this formula, the corresponding planet and actual distances are shown in table 1.

Titius attributed the gap at $n = 3$ to undiscovered moons of Mars. Like Kepler, Bode argued that this gap was occupied by a major planet. The discovery of Uranus in 1781 at a distance corresponding very closely to the $n = 6$ case bolstered the belief in a missing planet. (It should be noted that Neptune is a very poor fit for the $n = 7$ case and that the 'Law' appears to be nothing more than a numerical curiosity.)

Table 1. The Titius–Bode law.

n	A	Planet	Actual A
$-\infty$	0.4	Mercury	0.39
0	0.7	Venus	0.72
1	1.0	Earth	1.00
2	1.6	Mars	1.52
3	2.8		
4	5.2	Jupiter	5.20
5	10.0	Saturn	9.54
(6	19.6	Uranus	19.18)
(7	38.8	Neptune	30.06)

The celestial police

One of the proponents of Bode's law was BARON FRANZ VON ZACH, who went as far as predicting orbital elements for the 'missing planet' and undertook a solo search to find it. After several years of fruitless searching he gave up. In September 1800, von Zach held a meeting at JOHANN SHRÖTER's observatory in Lilienthal to organize a concerted search for the presumed planet. The search was to be conducted within about 4º of the ecliptic, which was divided up into 24 sections, each section to be searched by a different astronomer. On 1 January 1801, before these celestial police could get to work, a Sicilian monk, Father GUISEPPE PIAZZI, discovered a moving stellar object in the course of constructing a star catalog. Piazzi observed the object for 6 weeks before being interrupted by illness. Although Piazzi wrote letters to various astronomers, including Bode, announcing the discovery, by the time the letters were received the new object was no longer observable. Using Piazzi's observations, various orbit computers attempted to predict where the object would be the following year, but the predictions covered 5º of sky. Prompted by Piazzi's discovery, the German mathematician CARL FRIEDRICH GAUSS developed a new method of general orbit computation. Gauss' work enabled the new object, which received the name Ceres Ferdinandea (later thankfully shortened to just CERES), to be recovered within 0.5º of his prediction independently by von Zach and HEINRICH OLBERS.

The visual era

On 28 March 1802 Olbers was checking positions of stars in the vicinity of Ceres when he discovered PALLAS. On 1 September 1804, KARL HARDING discovered the third asteroid, JUNO. Olbers noted that Juno had been discovered near to one of intersection points of the orbits of Ceres and Pallas and, believing that the asteroids were the fragments of a single object, concentrated his searching near these points. On 29 March 1807 his diligence paid off with the discovery of VESTA.

In 1845, Karl Hencke, a German postmaster, discovered the fifth asteroid, Astraea, reportedly after searching for 15 years. Hencke discovered Hebe in 1847 and other astronomers quickly began their own searches. Various observing techniques were used. Brighter asteroids could be detected by their absence on published star charts. For asteroids fainter than were shown on charts, observers had to resort to drawing their own charts and comparing drawings of the same star field at different times. By 1850 13 asteroids had been discovered. The tradition quickly developed that newly discovered objects were named immediately on (or very soon after) announcement of the discovery. The early discoveries were also allocated symbols, though this practice died out by 1855. Numbers, assigned in the order of discovery, were introduced in 1851.

The fact that numbers and names were assigned immediately on publication of the discovery announcement often led to confusion. The rapid growth in the number of discoveries (100 by 1868, 200 by 1879 and 300 by 1890) overwhelmed the capabilities of the orbit computers trying to keep track of all the objects, and as a result many objects were lost. In 1856 Hermann Goldschmidt discovered (41) Daphne. The object was followed for only four days. Goldschmidt thought he had recovered (41) the following year, but it turned out to be a new object, which became (56) Melete. This latter object was observed for only two weeks and was lost until Goldschmidt made an exhaustive search for it in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1861. Karl Luther accidentally found (41) again in 1862 but it was mistaken for a new object and for a while was numbered (74).

As more asteroids were discovered, it became clear that the majority of the objects were to be found in a belt, the Main Belt, that stretched from 2.2 to 3.28 AU from the Sun (see ASTEROID BELT). The first object discovered lying outside the Main Belt was (65) Cybele in 1861. In 1875, (153) Hilda was found at a mean distance of almost 4 AU. This was the prototype of the Hildas, objects that are in the 3:2 mean-motion resonance with Jupiter and are protected from having close approaches to that planet despite their orbits approaching quite closely.

The photographic era

Until 1892 all observing had been performed visually, using transit circles or micrometers. The introduction of photographic searches in late 1891 by MAX WOLF in Heidelberg, Germany, and shortly thereafter by Auguste Charlois in Nice, France, meant not only that observers could detect fainter objects than could be observed visually but that the photographic plate was a permanent record of an observation. Visual discoveries had been made by noting changes in star fields over the course of several hours as objects moved through the field. With long-exposure photographs, taken with the telescope tracked at sidereal rate, moving objects would betray their presence by producing short trails on the plate. The problem with photographic observation was that, at least initially, accurate positions for objects on the plates were often not measured, primarily because this was a time-consuming process requiring specialized measuring equipment. In general, approximate positions, which could be derived quickly, were obtained. Accurate follow-up observations would then be obtained by visual observers using micrometers. However, the visual observers had difficulty following the fainter photographic discoveries, and an increasing number of newly discovered objects were quickly lost.

To cope with this growing problem provisional designations were introduced. Newly discovered objects would first receive a provisional designation, a permanent number being assigned only when it was felt that the orbit of the new object was secure. Intially, the provisional designation consisted of the year and a single letter. Thus the first object so designated was 1892 A, the second 1892 B, etc. The letter I was omitted so this scheme could cope with 25 discoveries per year. The year 1892 ended with 1892 V. The letter sequence was started afresh in 1893, but the designation 1893 Z was reached in May of that year. A new series of double letters, beginning with AA, AB, … , AZ, BA, BB, etc, was instituted. This double-letter sequence was not started afresh each year; the last object discovered in 1893 was 1893 AP, the first in 1894 was 1894 AQ. The double-letter sequence was exhausted with 1907 ZZ. Rather than start a triple-letter sequence, the double-letter sequence was simply restarted with 1907 AA. The sequence was again exhausted with 1916 ZZ and a fresh start was made with 1916 AA. In 1925 a new system for provisional designations was introduced. This used a double-letter sequence and was tied to the calendar date of discovery. The first letter reflected the half-month of the discovery of the object: A represented discoveries made between 1 and 15 January, B between 16 and 31 January, C between 1 and 15 February, … , I was omitted, … , Y between 16 and 31 December. The second letter indicated the order within each half month (again I was omitted). If there were more than 25 discoveries in a half month the second letter sequence was recycled and a number was added to the designation. Thus the sequence of designations for objects discovered in the first half of January 1998 would be 1998 AA, 1998 AB, … , 1998 AZ, 1998 AA_1, 1998 AB_1, … , 1998 AZ_1, 1998 AA_2, etc.

A substantial search program had been undertaken at Heidelberg since 1891 and the primarily visual program of Johann Palisa at Vienna had been operating since 1881. Between 1874 and 1880 Palisa discovered 28 asteroids at Pola, which was then in Austria. After moving to Vienna he discovered a further 55 before he abandoned his searching in order to dedicate himself to follow-up of other people's discoveries. He resumed his searching for new objects in 1905 and discovered a further 38 by 1923. Further substantial programs were begun at Johannesburg and Simeis in 1911, Uccle (Belgium) in 1925, Flagstaff (USA) in 1929 (a search that produced the discovery of Pluto) and Turku (Finland) in 1935.

New classes of asteroid continued to be discovered: 1898, (433) EROS, the first object with a perihelion distances less than 1.3 AU; 1903, (588) ACHILLES, the first object in 1:1 resonance with a major planet (Jupiter Trojans); 1920, (944) Hidalgo, an asteroidal object with a cometary orbit; 1932, (1862) Apollo, the first asteroid with a perihelion distance less than 1 AU (see NEAR-EARTH ASTEROIDS).

The task of cataloging the asteroids had been undertaken in Germany. The principal journal for publishing discovery, positional and orbital information was the *Astronomische Nachrichten* (AN). The Astronomishes Rechen-Institut, Berlin-Dahlem, began publishing a series of rapid-alert circulars, the *RI Circulars*, in 1926. Annual collections of minor planet ephemerides were published, first in the *Jahrbuch* of the Rechen-Institut, then from 1917 separately as the *Kleine Planeten*. However, the operation of these entities was severely compromised by World War II.

In 1947 the International Astronomical Union divided up the task of cataloging the minor planets. The responsibility for publishing annual volumes of ephemerides for the numbered objects was assigned to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the Institute of Theoretical Astronomy in Leningrad, USSR (now St Petersburg, Russia). The Minor Planet Center, set up at the University of Cincinnati, was charged with tracking newly discovered objects and assigning provisional and permanent designations. The Center began publication of the *Minor Planet Circulars*, a replacement for the defunct *RI Circulars*. An early priority was the rediscovery of those numbered objects that had become lost. Of the 1564 numbered objects in 1945, the orbits of some 20% were considered to be in bad shape. New survey programs at Indiana's Goethe Link Observatory, the McDonald Observatory and the Crimean Astrophysical Observatory addressed the problem of the lost objects and helped to recover many objects. On the computational front, Paul Herget, first director of the Minor Planet Center, used the recently developed electronic computer to perform orbit computations and improvements.

Prior to 1973, all the known near-Earth objects (NEOs) had been discovered as byproducts of other astronomical programs. Two such objects had been found in 1948 in the course of the Lick Proper Motion Survey and five were found at Palomar between 1949 and 1954 during the Palomar Sky Survey. In 1973, Eleanor Helin and Eugene Shoemaker began the Palomar Planet-Crossing Asteroid Survey, the first survey whose primary goal was the discovery of NEOs.

More new classes of asteroid were found: 1976, (2062) Aten, the first asteroid with an orbit smaller than the Earth's orbit; 1977, (2060) CHIRON, the first Centaur with an orbit between the orbits of the Jovian planets. In 1978, when the Minor Planet Center moved from Cincinnati to the Smithsonian Astrophysical Obervatory in Cambridge, Massachusetts, 4390 *Minor Planet Circulars* had been issued and there were 2060 numbered objects, of which 20 or so were considered lost.

The CCD era

Just as with the change from visual to photographic observation, the change from photographic to CCD observing has meant that observers can reach fainter objects more rapidly than was possible before. One drawback that has only recently been addressed is that a CCD's field of view is much smaller than the coverage afforded by a photographic plate. A plate taken with a large Schmidt typically covers an area 6º by 6º, whereas early CCDs might cover only 6 arcmin by 6 arcmin. Modern large-format CCDs and CCD arrays are alleviating this coverage problem.

The first systematic application of a CCD to asteroid searching was by the Spacewatch project in 1981. Using the venerable 0.91 m reflector at the Steward Observatory, Kitt Peak, Spacewatch has discovered over 180 NEOs and 25 000 Main-Belt asteroids. Most of the current professional CCD searches for asteroids are optimized for discovering NEOs but sweep up many routine objects in the course of their searches. The LINEAR program (Lincoln Laboratory Near-Earth Asteroid Research team) has discovered 153 NEOs and more than 19 000 Main-Belt asteroids. Other active programs include NEAT (Near-Earth Asteroid Tracking), LONEOS (Lowell Observatory Near-Earth Object Search) and ODAS (OCA–DLR Asteroid Survey). Other professional programs, such as the Czech programs at Klet and Ondrejov, concentrate on follow-up.

New classes of asteroid continue to be discovered. The first TRANS-NEPTUNIAN OBJECTS, after Pluto, were discovered in 1992 and 1993. Some of these objects have orbits that lie well beyond the orbit of Neptune and are essentially circular, while others have orbits that are in mean-motion resonances with Neptune (principally the 2:3 resonance, as for Pluto). Although the orbits of these latter objects cross the orbit of Neptune, the resonance prevents close encounters with Neptune. The extreme faintness of the majority of these objects has made follow-up difficult and many of the objects discovered must already be classed as lost.

The availability of affordable CCD cameras to the amateur market in the early 1990s led to a resurgence of amateur activity in the field of asteroid discovery. Previously many amateurs, although possessing suitable telescopes, had been unable to observe asteroids astrometrically because they did not have the facilities for measuring photographic plates or films. CCDs and the emergence of computer programs to process CCD images have made it feasible for many amateurs to observe and discover asteroids. Other amateurs have become dedicated follow-up observers, diligently tracking down discoveries of potential NEOs by the professional surveys. Japanese amateurs have been very active since the 1970s, first photographically and then using CCDs. Major amateur activity is also found in Italy, Australia and the US.

At time of writing there are 10 986 numbered asteroids, over 164 000 provisional designations have been assigned and over 35 000 *Minor Planet Circulars* have been issued. Only one numbered asteroid remains lost. There are over 12 000 unnumbered objects with good orbits that will be eligible for numbering in the near future. More than 25 000 other objects, observed only in their discovery apparition, have orbits of varying quality and await recovery or rediscovery.

Bibliography

Carusi A, Gehrels T, Helin E F, Marsden B G, Russell K S, Shoemaker C S, Shoemaker E M and Steel D I 1994 Near-earth objects: present search programs *Hazards Due to Comets and Asteroids* ed T Gehrels (Tucso, AZ: University of Arizona Press) pp 127–47

Cunningham C J 1988 *Introduction to Asteroids* (Richmond, VA: Willmann-Bell)

Marsden B G 1980 *The Minor Planet Center Celestial Mech.* **22** 63–71

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Marsden B G 1994 Asteroid and comet surveys *Astronomy from Wide-Field Imaging* ed H T MacGillivray, E B Thomson, B M Lasker, I N Reid, D F Malin, R M West and H Lorenz (Dordrecht: Kluwer) pp 385–400

Minor Planet Center
http://cfa-www.harvard.edu/iau/mpc.html and the links therein

Gareth V Williams

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Asteroids

Asteroids are small bodies which orbit around the Sun and which show no cometary activity nor have the potential for it. The word 'small' is intended to exclude planets from the definition. All known asteroids have diameters less than 1000 km down to a few centimeters (METEOROIDS). The largest known asteroid is 1 CERES, discovered on the 1 January 1801 by Giuseppe Piazzi in Palermo, Italy with a diameter of 950 km. Asteroids are also called minor planets. The newly discovered minor body known as 2001 KX76, whose diameter makes it the largest-known minor body in the solar system (its diameter, which is not very well determined, lies in the region of 1200 km, possibly even 1400 km) is a TRANSNEPTUNIAN OBJECT (TNO). Most asteroids orbit within a confined region between Mars and Jupiter, called the main ASTEROID BELT.

Out of about 30 000 known objects, only 8500 asteroids have precisely determined orbits. Each of these asteroids has a definitive identification, consisting of a number that indicates its order of entry and a name proposed by the discoverer (e.g. '3752 Camillo' identifies the 3752nd asteroid discovered). Only preliminary orbits are available for the other 21 500; each is marked by its year of discovery and two letters indicating the date of first observation as temporary identification (e.g. '1998 BC' identifies the third asteroid (C) discovered during the second half of January (B) 1998). A thousand asteroids have been detected by the INFRARED ASTRONOMY SATELLITE (IRAS) which provided a radiometric diameter and the albedo determination of most of them. About 700 asteroids have the rotational period determined. Less than 1000 asteroids are larger than 30 km across, and of these, about 200 asteroids are larger than 100 km. Extrapolating the asteroid size distribution down to 1 km, an estimation of about one million asteroids whose diameter is larger than or equal to 1 km is obtained. The global mass of the population has been estimated to be of the order of one-thousandth of the Earth's mass.

Asteroids were considered as uninteresting 'vermin of the skies' until it was understood that they store a rich variety of information. In fact, asteroids are believed to be the remnants, either fragments or 'survivors' of the swarm of PLANETESIMALS from which the terrestrial planets were formed: the records of primordial chemical and physical processes can still be found frozen in the actual characters of asteroids. The study of these objects provides information on the nature of these primordial swarms even if differentiation and/or internal activity affected in a different way their evolution. Ground-based and *in situ* observations of asteroids show a population of solid bodies exhibiting a wide variety of chemical and physical properties: one of the most interesting characteristics of asteroids is this variety. Understanding the origin of the diversities requires the study of a large number of bodies. The information collected *in situ* on the various space mission target asteroids represents a powerful test to calibrate the techniques used in determining the chemical and physical properties of asteroids from the ground-based observations which are and will remain the main source of asteroid data.

The current knowledge on the asteroid population has been obtained mainly from the ground-based observations. Recent observations taken by the 10 m Keck II telescope on Mauna Kea, for example, reveal the presence of a small satellite orbiting around asteroid 87 Sylvia. The projected separation between primary and secondary was 0.59" in H-band images obtained on 18 February 2001 using the adaptive optics system on the 10 m W. M. Keck II telescope. Little relative motion was apparent in images obtained an hour apart (upper limit 0.02"). The brightness ratio was measured to be 420±70, implying about a 1:20 ratio of sizes. This implies that the satellite is around 7 km wide, in contrast to Sylvia, which is around 130 km wide.

Few large asteroids are included in the observational programme of the HUBBLE SPACE TELESCOPE (HST) and the first results concerning 4 VESTA are of high interest. The Infrared Satellite Observatory (ISO) observed a few tens of asteroids, allowing astronomers to obtain spectral data up to 45 μm. The National Aeronautics and Space Administration (NASA) GALILLEO MISSION approached for the first time two asteroids, 951 GASPRA (October 1992) and 243 IDA (August 1993), offering to the planetary science community the opportunity to have a close look at two S-type main-belt asteroids. The 253 Mathilde flyby (June 1997, Near Earth Asteroid Rendezvous (NEAR) mission) provided us with the first *in situ* data for a C-type asteroid.

Main-belt asteroids

The asteroid main belt occupies a doughnut-like region around the Sun which contains the orbits of most asteroids, characterized by semimajor axes (a) ranging between 2.1 AU and 3.3 AU, eccentricities (e) ranging between 0.01 and 0.3 and inclinations (i) with respect to the plane of the Earth's orbit (the ecliptic) between 0° and 35° (with an average value of 10°).

The structure of the asteroid belt is determined by gravitational interactions with both Mars and Jupiter. Because of its larger mass, Jupiter's effects are dominant and cause a depletion of objects in correspondence to the resonances, i.e. the regions where the orbital period of a body would be some exact integer ratio of Jupiter's orbital period (an asteroid 2.5 AU from the Sun is said to be at the 3:1 'resonance': that is, the body would complete exactly three orbits around the Sun for every one that Jupiter completes). Those breaks in the uniformity of the asteroid belt are called KIRKWOOD GAPS from the name of the American astronomer who identified them.

Removing the effects of secular planetary perturbations it is possible to obtain the description of an asteroid orbit in terms of its proper elements. Some main-belt asteroids have very close proper semimajor axes, proper

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

eccentricities and proper inclinations: they form clusters in the space of proper elements, and generally all the members of a cluster belong to the same compositional type. In 1918, the Japanese astronomer Hirayama called these clusters 'asteroid families' (see HIRAYAMA FAMILY). Analysing the 950 known asteroids, Hirayama named the identified families from the largest family member: Themis, Eos, Koronis and Flora. Since then, several other families and clans (a cluster of asteroids is called 'clan' when unequivocal family member definition and/or separation from other background objects is impossible) have been identified. Asteroid families are interpreted as the result of the catastrophic disruption of a parent body due to a collision with another smaller asteroid. If so, asteroid families represent a natural laboratory where high velocity (v~5 km s^{-1}) impacts were produced at a scale which is impossible to reproduce in our laboratories. The size distribution of family members (fragments of the parent body) is a picture of the outcomes of a collision; the chemical composition of the fragments gives an indication of the internal structure and composition of the parent body. The study of asteroid families provides information on the collisional history of the asteroid population since its formation.

The discovery of a new family of asteroids was confirmed in 2002. It is believed to be the youngest at under 6 million years old. David Nesvorny and co-workers at the Southwest Research Institute in Boulder, Colorado, recognized the 39 or so asteroids as belonging to the same family by studying their orbits. Projecting the current orbits backward in time, the researchers found that they converge about 5.8 million years ago—so this must be when they formed from the break-up of a larger body. Of the 39, two are big asteroids. One, called (832) Karin, after which the whole cluster is named, is about 19 km across; the other, called (4507) 1990, measures about 14 km. The team estimates that the parent body was at least 24.5 km across. The size of the parent body and of the fragments it produced can provide a test of theoretical models of how impacts shatter rocky bodies. The relatively fresh surfaces of the two larger fragments can also reveal what asteroids are made of, and how rapidly weathering in space pockmarks them with small craters. The Karin collision might also have been responsible for a recently discovered band of asteroidal dust.

The rotation rate versus diameter distribution of the main-belt asteroids shows that the objects with a diameter of about 100 km have rotation periods longer than those of both larger and smaller asteroids. In the hypothesis of a collisional evolution of the asteroids from a planetesimal swarm, this diameter value has been interpreted as the size where the separation between primordial larger asteroids and their smaller collisional products occurs. The spin rate distribution of asteroids with diameter larger than 50 km is unimodal and reflects the rotational behavior acquired during the primordial phases of the asteroid accretion process, while the distribution of the asteroids with D<50 km is the superimposition of three components, the 'primordial', one of fast rotators and one of very slow rotators. Disruptive collisions allow small fragments to gain angular momentum forming the fast rotator group, while a slow-down mechanism which originated the slow rotators, including 253 Mathilde, has to be further investigated.

The shapes of 50 extensively observed large asteroids have been evaluated with the Fourier analysis of the complete set of light curves and radar measurements have been used to determine the shapes of small objects. These data indicate that larger asteroids are more regular than the smaller ones. This result indicates again that the larger bodies are the remnants of the original population, which preserved in their shape the equilibrium figures, while the smaller generally irregular objects are the records of disruptive collisions.

The direction of the spin axis (pole orientation) has been determined (within an accuracy of 15% at the best) for 50 asteroids. Even if the available sample is too small to draw firm conclusions, the distribution of these pole directions seems bimodal: again a possible indication of the superimposition of two different groups of asteroids. The precise determination of the pole direction of a significative number of asteroids will lead to a better understanding of the history of the population. In fact, even though most of the spin axis inclinations of the primordial asteroids were distributed around a single value as a consequence of the accretion process, the collisions will reorient randomly the spin axes, whose distribution will be spread and flattened.

The asteroid population spreads between the orbits of Mars and Jupiter. In this zone lies the boundary separating the inner solar system (where the water in the protosolar nebula vaporized) from the outer regions (where water condensation occurred). This zone, spanning between 2.6 AU and 3.1 AU, is known as the aqueous alteration zone where spectral signatures of electron transitions, possible only in presence of liquid water, have been detected in numerous C-type asteroid spectra. This is strong evidence for the presence, sometimes in the past, of liquid water on these asteroids. The amount of water present in an asteroid is questioned but its detection would be relevant to understanding the relationships between asteroids and COMETS.

Asteroid composition

Chemical composition and mineralogy of asteroid surfaces can be determined by spectroscopy or, even if with lower precision, by multifilter spectrophotometry. Analysing with sophisticated clustering codes all the available asteroid spectra a taxonomic system has been obtained. The principal recognized classes are designated, in order of their observed relative abundance, by the letters: S, C, M, D, F, P, V, G, E, B, A (initials of the words used to roughly describe an asteroid or its spectrum, i.e. stony, carbonaceous, metallic, …). Most of the asteroids belong to the low-albedo C-type, characterized by relatively neutral spectra, exhibiting an absorption in

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

the ultraviolet wavelengths and sometimes the signatures of aqueous alteration. Their spectra match the spectra of the carbonaceous chondritic meteorites, a strong indication of similar composition (see CHONDRITES). A few asteroids (B-, G- and F-type) show C-like spectra with minor differences at the shorter wavelengths, which may reflect degrees of metamorphism of C-like material. D-type asteroids, characterizing the Trojan group, have low albedo and are redder than the C-type ones. The low-albedo P-type asteroids, predominant at 3.6 AU<a<4.9 AU, represent a transition group between C-type and D-type asteroids, their spectra are redder than C's spectra but flatter than D's spectra. The inner-belt asteroids belong to the higher-albedo S-type, whose reddish spectra are characterized in the ultraviolet by a deep absorption charge transfer band, and in the near-infrared side by the absorption band(s) of pyroxene and olivine (magnesium- and/or iron-bearing minerals: $[Mg, Fe]_2SiO_4$). These asteroids, which belong to the most abundant asteroid type, have no spectral analog in the METEORITES, even if their bulk mineralogical composition seems to fit the mineralogy of both the stony iron and the ordinary chondritic meteorites, which are the most common meteorites in the natural museum collections. One possible answer to this puzzle (known as the ordinary chondrites conundrum) may be 'space weathering'. Catastrophic collisions, solar radiation and micrometeoroid impacts could have changed an ordinary chondrite asteroid's surface enough to alter its spectral characteristics. This hypothesis seems to be confirmed by a spectroscopic study of several near-Earth asteroids which shows a continuous trend from the ordinary chondrite-type spectrum to the common S-type spectrum. This behavior can also explain the slight differences within the spectra of S-type asteroids. Some moderate-albedo asteroids have slightly reddish spectra and very high radar reflectivity; they are almost certainly made out of pure metal, probably the nickel–iron alloy which characterizes the metallic meteorites. A small number of high-albedo asteroids exhibit spectra typical of highly differentiated mineral assemblages: the surface of A-type asteroids probably contains rocks and regolith (regular lithology, indicating the dusty surface layer) formed by pure olivine; the V-type asteroids are probably all members of the Vesta family, they represent crust fragments of 4 Vesta, which is the only known intact differentiated asteroid having a basaltic lava surface. Some very-high-albedo objects have a neutral spectra resembling those of enstatite meteorites (E-type asteroids).

One of the most significant results of the asteroid taxonomy is that it reveals a distinct structure in the composition of asteroids, which vary with distance from the Sun. E-type objects are present at the inner edge of the belt and S-type asteroids populate the inner one-third of the main belt, while the C-type fill the rest of the belt. P-type predominate in the outer regions and D-type characterize the asteroids in Jupiter's orbit.

Interpreting the variation of the asteroid taxonomic types (somehow indicative of their composition) across the belt in terms of the formation processes, a few possible evolutive physical–chemical trends of the asteroid population can be sketched. All planetesimals were formed by the solid component of the PROTOPLANETARY NEBULA accreting to form a material similar to the one characterizing the D-type asteroids. This pristine material is still present in the outer belt and among the Trojans, and produced different end members (C and B, M and E, V and A-types) characterized by the variation (in space and in time) of environmental conditions in the nebula, by the internal differentiation and/or by the collisions. In figure 1, four evolutionary trends are outlined by the arrows. A first trend links the D objects to the B ones including the C-type asteroids. This sequence can be interpreted in terms of diminishing volatile content (from D to B): these asteroids seem to be composed of mineral assemblages accreted from material condensed at low temperature which experienced different thermal heating owing to the differing heliocentric distance (increasing from B to D). A second trend, connecting the primitive D-type asteroids to the G, M and E objects, can be interpreted as a progressive evolution of the pristine materials (D) toward the enstatite achondrites (E), passing through an ultraprimitive, high-carbon, low metamorphic grade C-type mineral assemblage (G) and a reduction silicate similar to the enstatite chondrites (M).

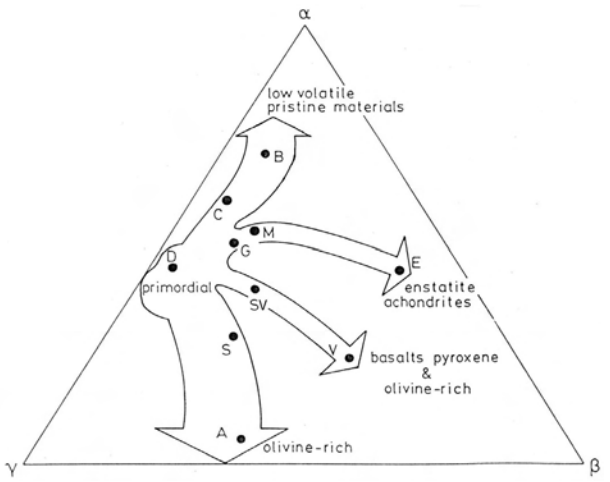


Figure 1. The asteroid compositional classes have been defined on the basis of their reflectance spectra and albedo. In this ternary diagram, the variables are grouped so that at each point $\alpha+\beta+\gamma=1$. The vertices of the ternary diagram are characterized as follows: α represents the shorter wavelength (U, B, V) regions of the asteroid reflectance spectra, β represents the longer wavelength (I, IR) regions of the asteroid reflectance spectra, and γ accounts for the asteroid albedo. The principal compositional classes are shown in the ternary diagram. Four asteroid evolutive trends are outlined by the arrows: from the 'primitive' D asteroids each trend goes towards an end member. The arrows indicate a decrease in heliocentric distance too.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The E-type asteroids may be composed by minerals formed when M-type bodies are differentiated. This trend describes the evolution of asteroids whose material appears to be a higher-temperature agglomeration of nebular condensate. The third and fourth trends ending at V-type and A-type respectively seem to represent lines of increasing differentiation, starting from the pristine D-type material and including the S-type objects. The metal-rich silicate surface material, characterizing these trends, may be the result of magmatic differentiation produced inside a large body (e.g. Vesta) or due to successive heating events such as large impacts. The differences of these evolution trends toward pyroxene–olivine basalts (V) and pure olivine (A) end members can take into account different collisional histories of the asteroids with respect to their hypothetical differentiation sequence.

The distant asteroids

Outside the main belt, toward the external region of the solar system, a few groups of asteroids behave in a completely different way to the main-belt ones: in fact, they reside within the main resonances with Jupiter. At 3.9 AU (resonance 3:2) there is the group of Hilda; the asteroid Thule falls in the 4:3 resonance (at 4.3 AU).

An especially interesting group, called the TROJAN ASTEROIDS, resides at the 1:1 resonance. These asteroids have the same heliocentric distance and orbital period as Jupiter does: they orbit near the two equilibrium position, known as Jupiter Lagrange points, that lie about 60° east and west of Jupiter within its orbital plane. About 200 Trojans are currently known; several thousand, larger than 15 km, are supposed to exist. While classified as asteroids on the basis of their lack of cometary activity, the physical nature of Trojans is uncertain: their optical spectra show red continua with no evidence of absorption or emission features (taxonomic type D): their albedos are very low (2–3%). Both of these characteristics are comparable with those of the few observed cometary nuclei and may be evidence for common organic compounds on the two classes of objects. As in the case of comets, the dark reddish surfaces of the Trojans could be thin rubble mantles, shielding buried ice whose sublimation rate (if exposed, e.g. by a collision) is too small to sustain an observable coma.

The Jupiter orbit marks the limit between the rocky worlds (Earth-like planets, asteroids, dust) and the ice worlds (the giant planets' satellite systems, comets, rings); in these regions it would be possible to find the transition element between asteroid and comets, which are probably the opposite end members of the same population of primeval planetesimals. In fact, a few large objects, called CENTAURS, have been recently discovered in orbits crossing the orbits of gas-giant planets. 2060 CHIRON, first Centaur discovered, is the only member of this group to display cometary activity (its comet designation is P/Koval). The Centaurs are probably a vast population of unstable, outer-planet-crossing bodies; their total number has been estimated to be 10^5–10^6 objects down to km size. All the known Centaurs are dynamically similar, but their surfaces show dramatic spectral differences: some have a completely neutral spectrum (solar like), while other exhibit the reddest spectra of solar system bodies. Are they bald comets, as well as TRANS-NEPTUNIAN OBJECTS?

The near-Earth asteroids

The NEAR-EARTH ASTEROIDS (NEAs) belong to three different groups of generally small objects residing within the inner solar system. About 500 NEAs are known and their discovery rate is continuously increasing. Of the NEAs, 5.8% belong to the Atens group characterized by a semi-major axis $a<1$ AU and an aphelion distance $Q>0.983$ AU, 48.5% belong to the Apollo group ($a \geqq 1$ AU and a perihelion distance $q \leqq 1.017$ AU) and 45.7% belong to the Amor group ($1.017 \text{ AU} < q \leqq 1.3$ AU).

The study of the physical and the dynamical properties of these objects is necessary for understanding their histories and relations with comets, main-belt asteroids and meteorites.

Understanding the dynamical evolution of NEAs is important because (i) they can have close approaches with Earth, representing a long-term danger for the biosphere and the human species, (ii) they represent the larger members of meteoroids and small interplanetary objects and (iii) physical remote-sensing observations of them have recently revealed a number of surprising features, such as high metal abundance, binary structures and tumbling rotational states which could be related to their asteroidal or cometary origin.

The general idea on NEA origin is that these objects are efficiently removed from other regions of the solar system by collisions and subsequent gravitational interactions with the planets on time scales of 10^6–10^8 yr. Since the NEAs have unstable orbits, a continuous supply of new objects is needed. Two main mechanisms have been commonly proposed for supplying the Earth-approaching population: the first assumes that NEAs are asteroidal fragments coming from the main belt, while the second supposes that a possible source for NEAs are dormant or extinct comet nuclei. NEAs of both asteroidal and cometary origin are widely believed to be continuously injected into Earth-crossing orbits through a few different resonant channels, which collect fragments randomly ejected from main-belt asteroids as a consequence of energetic mutual asteroid collisions. These objects undergo a fairly complex orbital evolution process, driven by mean motion and secular resonances, by non-resonant secular perturbations and by a sequence of close encounters with planets. Eventually, they mostly fall into the Sun or are ejected from the solar system after a Jovian encounter; only a few per cent collide with terrestrial planets. To maintain the balance between the mean rate of asteroid loss in Earth-approaching orbits and the rate of their replenishment, source rates must yield a few hundreds of new objects larger than 1 km in diameter per 10^6 yr.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

So far relatively few spectroscopic observations of NEAs have been obtained because of their small sizes, faint apparent magnitudes and limited intervals of visibility. Our current knowledge of the nature of NEAs suggests that this population is characterized by a variety of surface mineralogies and includes almost all taxonomic types found in the main belt. Even if our sampling of NEAs for physical observations is strongly biased, no correlation seems to exist between taxonomic type and dimension or rotational period: both the largest and the smallest NEAs are S-type asteroids; the few C-types span a considerable range in diameter while the rare taxonomic types have a range of rotation rates as do the abundant S-types. On the basis of these data it seems that the NEA population has approximately the same proportion of S-type and C-type asteroids as the main belt. This similarity raises the question about where the dead comets, which are not expected to be a major component of the main belt, are. Several NEAs classified as C-types are apparently formed by primitive material similar to hydrous CI and CM carbonaceous chondrite meteorites. A spectroscopic survey of NEAs in the 0.45–0.9 μm range found that out of 35 objects six have spectra similar to those of ordinary chondrite meteorites, while 29 have spectral properties spanning the range between the domains of ordinary chondrite meteorites and the most common S-type main-belt asteroids. This range could arise through a diversity of mineralogies and regolith particle sizes, as well as through a time-dependent surface weathering process. In this case, if a time-dependent weathering process is active (as proposed to explain the surface variations measured in Galileo spacecraft images of 243 Ida), asteroids closely resembling ordinary chondrite meteorites would be those with the youngest surfaces.

Observations have suggested binary asteroids are common in Earth-crossing orbits. Using the world's two most powerful astronomical radar telescopes, astronomers estimate that about 16% of near-Earth asteroids larger than 200 meters (219 yards) across are likely to be binary systems. Near-Earth asteroids may become binaries when the planets' much larger gravities pull on them, distorting and sometimes breaking off a satellite. Theoretical and modelling results show that binary asteroids most likely form when the asteroids make a close encounter with the inner planets Earth or Mars, sometimes just 10 000 miles from a planet's surface.

Two asteroids made the news during 2002 because of their predicted close approach to the Earth. On 9 July 2002, MIT astronomers discovered 2002 NT7, a 2 km-wide space rock in a curious orbit. Unlike most asteroids, which circle the Sun in the plane of the planets, 2002 NT7 follows a path that is tilted 42 degrees. It spends most of its time far above or below the rest of the solar system. Every 2.29 years, however, the asteroid plunges through the inner solar system not far from Earth's orbit. After a week of follow-up observations, researchers did some calculations. There was a chance, they concluded, that 2002 NT7 might hit our planet on 1 February 2019. This made headline news, worrying some members of the public. What was not reported was that the odds of impact are 1-in-250 000. On 14 July astronomers discovered 2002 NY40. This asteroid measures about 800 meters across, and follows an orbit that ranges from the asteroid belt to the inner solar system. Its closest approach was on 18 August 2002, passing only 1.3 times farther away than the Moon, and being visible through binoculars.

Asteroid flyby

The Galileo mission on its way toward Jupiter explored the asteroids 951 Gaspra and 243 Ida, discovering around the latter the first asteroid satellite, named Dactyl; all these three are small S-type objects. 243 Ida is a member of the Koronis family. The Galileo flybys have provided images of these asteroids with a spatial resolution never obtained in Earth-based observations. The observations have shown that both 951 Gaspra and 243 Ida are of elongated shape and quite irregular, while Dactyl has a rather smooth, regular shape, which contrasts with the angularity of other known small bodies (951 Gaspra, 243 Ida, comet P/Halley nucleus, 4179 Toutatis and 4769 Castalia), but presents some aspects with comparable limb roughness with the Martian moons (PHOBOS AND DEIMOS). One prominent surface feature concerns the system of linear striae (depressions of uniform width as long as few km and few hundred meters in width with depths of probably no more than few tens of meters), which are similar to the grooves observed on Phobos and could be the signature of their origin due to a near- catastrophic collision. The most significant results were the determination of the crater size distribution on the object surfaces. From the crater frequency distribution, the age of 243 Ida was determined to be about 1 billion yr while that of 951 Gaspra was only 200 million yr. Both asteroids are covered by a layer of regolith with significant albedo and color variations.

The sequence of images where Dactyl is visible allowed us to determine its orbit as a function of 243 Ida's mass. The combination of this value with the estimation of 243 Ida's volume yields a bulk density of 2.6 ± 0.5 g cm^{-3} for 243 Ida, which is consistent with a bulk chondritic composition.

On June 27, 1997, NEAR passed close (about 1200 km) to the main-belt asteroid 253 Mathilde, providing the first flyby of a C-type asteroid.

Several ground-based reconnaissance studies were performed to characterize 253 Mathilde. A rotational period of 17.4 days has been determined: 253 Mathilde is an exceptionally slow and tumbling rotator. Spectral observations of 253 Mathilde showed its neutral reflective properties to be highly homogeneous over all the explored surface and its spectrum results comparable with those of carbonaceous and black ordinary chondrites. Whole-disk NEAR color observations between 400 and 1100 nm agree closely with these telescopic spectral results. NEAR found 253 Mathilde to be uniformly

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

bland in both albedo and color (more bland in color variations than Phobos). 253 Mathilde has a geometric albedo of 0.041 ± 0.003 at 0.55 μm. NEAR discovered five craters of diameter 19–33 km, comparable with the 26.5 km mean radius. The flyby yielded also a direct determination of the mass and a density estimate of 1.3 ± 0.2 g cm^{-3}. Comparison of this density with that of carbonaceous chondrites suggests a porosity of ~50% for 253 Mathilde.

In February 2000 the main goal of the mission, the rendezvous with the highly elongated S-type Amor group asteroid 433 EROS, was reached by means of a series of rendezvous maneuvers that slowed the spacecraft to a velocity relative to the asteroid of only 5 m s^{-1} before the insertion into orbit.

A set of detailed and precise data on three members of the S-type (951 Gaspra, 243 Ida and Dactyl), the second most abundant asteroid compositional type, is already available and within the next year new information on another peculiar S-type asteroid (433 Eros) will be obtained. On the basis of the analysis of these data it will be probably possible to determine whether S asteroids are primitive mineral assemblages, little changed since they first formed like the ordinary condrites, or material like basaltic and stony-iron meteorites that have undergone extensive melting and evolution. The first flyby of a C-type (253 Mathilde) gave an indication of a very low density of this body and of a highly homogeneous surface color. The collisions on 253 Mathilde have strongly affected its surface morphology. Are these characters common to the whole C-class?

On Friday 2 November 2002, NASA's Stardust spacecraft was tested in a flyby of asteroid Annefrank, discovering that the asteroid is about twice the size anticipated, but with a dimmer surface. Stardust flew within about 3300 km (2050 mi) of the asteroid as a rehearsal for the spacecraft's encounter with its primary target, comet Wild 2, in January 2004. The camera's resolution was sufficient to show that Annefrank is about 8 km (5 mi) in length, twice the predicted size from Earth-based observations. The surface reflects about 0.1 to 0.2 percent of sunlight, slightly less than anticipated. A few craters that are hundreds of meters across can be seen. The straight edge in the right side of the image may be an artifact of processing (figure 2).

Figure 2. Asteroid Annefrank imaged by Stardust. Credit: NASA/JPL.

Collisional evolution

After the formation of Jupiter and Mars, the planetesimal swarms orbiting between these two planets experienced strong perturbations and were widely depleted and the remnants were 'frozen' in much more stable orbits. Long-term perturbations could have produced mutual collisions, which, together with the space weathering, have been the only processes modifying the original planetesimal population.

Although the volume of space where asteroids orbit is large, and the asteroidal cross sections are small, the probability that asteroid collisions do occur in a time frame compatible with the solar system age is quite high (of the order of one collision per billion yr). Moreover, the fact that the rotation periods of asteroids follow a Maxwellian distribution may be taken as an indication that collisions were sufficiently frequent within the asteroid population to produce a collisionally relaxed state from a more massive primordial population. Many other observable characters of asteroids are presumably the effects of collisional evolution: the shape and the size distributions, the physical and chemical characteristics of families, the rotational properties.

Unfortunately the outcomes of a collision between two asteroids, depending mainly on their sizes and the relative velocity, are poorly known. Laboratory simulation of high-velocity (few km s^{-1}, of the order of the sound speed in rocky materials) impacts of a small projectile (~10^{-3} kg) on a larger target (~10 kg) cannot take into account the gravitational regime which dominates the asteroid collisions, even if scale laws are applied. What is the proportion of catastrophic disrupting collisions (where the majority of the mass of a body is dispersed) to the fragmenting ones (where the body is broken up, but the fragments re-accumulate in a rubble pile configuration)? How does the formation of an impact crater affect the chemistry of an original population asteroid? What are the effects of a collision on a rubble pile object? What is the mass loss (ejection out of the solar system or insertion in planet- or Sun-crossing orbits) induced by the collisional evolution process? The answers to these questions will allow us to understand the role of mutual collisions in shaping the present asteroid population.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

On the basis of what we believe today and of the information provided by the observations, we can only draw a preliminary scenario for the processes which governed the formation and the evolution of planetesimal swarms located between 2.3 AU and 5 AU from the Sun.

The break-up of very large bodies, such as Ceres, PALLAS and Vesta, seems to have been an improbable event and the fact that some of them preserve most of their ancient crust suggests that even large impacts had limited modifying effects on the larger asteroids. These considerations imply that, if the mass of the original asteroid population was larger, the potential impactors belonged to the rather small bodies. The smallest remnants in the actual asteroid population are probably fragments of larger bodies. The intermediate-size asteroids may have undergone repeated collisions, able to partially modify their composition (formation of shock mineral assemblages, thermal processing, etc) and/or their physical properties (rubble piles, rotation period, spin axis orientation, etc), but still retaining some primordial character.

If collisions have been rare events, the scenario depicted by the observations is still the primitive one, and we are looking at a relatively pristine remnant of the primordial planetesimals. If collisions affected the original swarm all along the solar system lifetime, only the larger asteroids may preserve intact some original characters, while most asteroids reflect the effects of a thorough reprocessing, and the search for the tracks of the original imprints will have to be much more careful.

Bibliography

Margot J L, *et al.* 2002 Binary Asteroids in the Near-Earth Object Population *Science* **296 (5572)** 1445–1448

Nesvorny D, Bottke W F Jr, Dones L and Levison H F 2002 The recent breakup of an asteroid in the main-belt region *Nature* **720–722**

Marcello Fulchignoni

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Astigmatism

An optical aberration whereby a lens (or mirror) has different focal lengths in planes that are perpendicular to each other. In general, if light rays from a point source enter a lens at an angle to its optical axis (the line passing perpendicularly through the center of the lens), the focal length for rays that lie in the plane defined by the direction of the source and the optical axis (the 'tangential' plane) will differ from the focal length for rays that lie in the plane perpendicular to this (the 'sagittal' plane). Moving outwards from the focus for tangential rays to the focus for sagittal rays, the shape of the image varies from a short line perpendicular to the tangential plane, through a circle (the circle of least confusion and position of best focus), to a short line perpendicular to the sagittal plane.

An optical system that is configured to remove astigmatism completely will suffer from field curvature (it will form sharply focused images on a surface that is curved rather than flat).

See also: focal length, focus, spherical aberration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrid

Two Swedish microsatellites designed to investigate the near-Earth plasma environment. Astrid-1 was launched in January 1995 and carried a neutral particle imager, an electron spectrometer and a miniature ultraviolet imaging system. It ceased to operate in March 1995. Astrid-2 was launched in December 1998 to measure particle distribution, electric and magnetic field variations in the auroral region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astro

Set of three ultraviolet telescopes flown on the Space Shuttle in December 1990 (STS-35) and March 1995 (STS-67). Included the 0.5 m Wisconsin ultraviolet photo-polarimeter experiment, a 38 cm ultraviolet imaging telescope and the 90 cm Hopkins ultraviolet telescope. Astro-1 also carried a broad-band x-ray telescope. Astro-1 observed some 135 targets, including Comet Austin. Astro-2 discovered helium gas in the intergalactic medium by observing a distant quasar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astro

Series of Japanese high-energy astronomy satellites. They include TENMA (Astro-B), GINGA (Astro-C), ASCA (Astro-D), Astro-E and IRIS (Astro-F).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astro Space Center, Moscow

Established in 1967 as part of the Institute for Space Research of the Russian Academy of Sciences. The Astro Space Center conducts fundamental research in astrophysics including cosmology, structure and evolution of astronomical objects, and the interstellar and interplanetary medium.

The center is particularly involved with developments in radioastronomy. It operates the Pushschino radioastronomy station and has an agreement to build a 70 m radio telescope on the Suffa Plateau in Uzbekistan.

The center is also developing the RadioAstron satellite, a 10 m radio telescope which will be inserted into a high elliptical orbit in order to make VLBI observations in conjunction with radio telescopes on the ground. It is part of the Spectrum series of spacecraft.

The center's Submillimetron project is under study as an astronomical payload on the Russian segment of the International Space Station. It would use a cryogenic telescope for studies at submillimeter and far-infrared wavelengths.

For further information see
http://www.asc.rssi.ru/default.asp.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrochemistry

The study of the chemistry of, and chemical processes operating in, astronomical sources and phenomena. As far as interstellar material is concerned much of the chemistry involved is organic chemistry of light molecules (i.e. molecules containing one or more carbon atoms, possible other atoms, and in total less than, say, 50).

See also: cosmic abundance of elements, heavy elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrodynamics

The study of the motion of bodies in gravitational fields, with particular reference to the motions of artificial satellites and space probes.

See also: celestial mechanics, rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrolabe

The astrolabe (more exactly, the planispheric astrolabe, to distinguish it from less important variants) was an observing and calculating instrument that had its roots in antiquity and reached maturity in the Islamic world. Christian monasteries on the slopes of the Pyrenees made early contact with ISLAMIC ASTRONOMY, and late in the tenth century Gerbert of Aurillac (later Pope Sylvester II) may have brought an understanding of the astrolabe from Spain to the cathedral school of Rheims. Certainly the instrument itself had appeared in the West by about 1025. Its subsequent availability in the Latin Middle Ages (see DIVIDED INSTRUMENTS) did much to encourage a quantitative science of astronomy.

The basis of the instrument is a heavy disk of brass that can be suspended from a ring so as to hang vertically. On the back of the instrument is an observing bar that pivots about a central pin; the observer measures the altitude of a celestial body by sighting along the bar and reading the altitude from a scale in degrees engraved around the circumference.

The front of the instrument is for calculation. To allow the instrument to be used at different latitudes, a set of alternative disks was supplied, and the user arranged them so that the most suitable latitude disk was face up within the recess provided. On this disk were engraved circles associated with a plane representation of the celestial sphere. If the celestial sphere is imagined as projected from the South Celestial Pole onto the plane of the equator (represented by the latitude disk), then the circle that is the celestial equator projects into itself, and the North Celestial Pole into the center of this circle. Points on the northern hemisphere project into points of the plane that are internal to the celestial equator, and those on the southern to points external. Because the brass astrolabe was necessarily of finite size and northern observers had little interest in skies south of the Tropic of Capricorn, the projection of this tropic was made the perimeter of the latitude disk.

The observer's zenith projected into a point, and circles of equal altitude in the sky as seen from his position projected into circles enclosing (but not centered on) this point, while arcs of equal azimuth projected into arcs passing through this point.

Thus far the projection incorporated the principal fixed circles in the sky and the coordinate circles or arcs associated with observations by the observer. To represent the rotating heavens, and in particular the ecliptic path of the Sun and the positions of the stars, the instrument maker provided a further disk (the *rete*) superimposed on the latitude disk and rotating about the central pivot (thus matching the rotation of the sky about the North Celestial Pole). To avoid obscuring the latitude disk, the maker cut away as much of the rete as possible, leaving little more than a network of labelled star-pointers together with a circle that represented the ecliptic.

The sky rotates with one degree of freedom, and therefore a single observation (the ambiguity between rising and setting presenting little difficulty in practice) is sufficient to determine the correct position of the rete at any given time. So, for example, the observer might first observe the altitude of Sirius, and then, having retreated to the warmth of his study, he would rotate the rete until the point representing Sirius was correctly positioned above the relevant circle of altitude. All the other stars would then be correctly positioned likewise, and he could determine, for example, which were now above the horizon and which below.

On Western astrolabes the back of the instrument contained not only the circular scale in degrees for measuring altitudes, but a second circular scale with the days of the year, placed so that the appropriate position of the Sun on the ecliptic was aligned with the calendar date. Knowing in this way the current position of the Sun on the ecliptic, the user could imagine an appropriate mark on the representation of the ecliptic. A brass bar located on top of the rete and also rotating about the central pivot could then be placed across this mark, so as to indicate the time on a 24 h scale located outside the perimeter of the latitude disk, and in this way the instrument became a clock that told the time both night and day. Thus the night-time observation of Sirius that had allowed the rete to be correctly positioned also gave the position of the Sun and hence the time of the observation. In the daytime, observation of the Sun itself gave the same information. Similarly, by rotating the rete the user could establish at what time a particular event—the rising of Sirius, for example—would occur.

As an observing instrument accurate to a small number of degrees, the astrolabe could also be used to measure the altitudes of the planets, and this fostered a revival of planetary astronomy in the Latin Middle Ages that was to culminate in the *De Revolutionibus* of COPERNICUS. However, its primary use was in astrology, especially as a guide to medical treatment.

Bibliography

Evans J 1998 *The History and Practice of Ancient Astronomy* (New York: Oxford University Press) pp 141–61

Hoskin M (ed) 1997 *The Cambridge Illustrated History of Astronomy* (Cambridge: Cambridge University Press) pp 64–7

North J D 1974 'The astrolabe' *Sci. Am.* **230** 96–106

Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrology

Astrology is the theory that the planets, the Sun and the Moon, as well as the 12 'zodiacal signs', combine in various, ever-changing configurations with respect to each other and the local horizon to influence 'sublunar' events.

In astrology, as in classical astronomy before the acceptance of the Copernican hypothesis, it was assumed that the Earth was positioned in the center of a finite universe, and that the Earth's center was the center of the spheres of the celestial bodies—in upward order from the Earth, the Moon, Mercury, Venus, the Sun, Mars, Jupiter, Saturn, the 'fixed' stars, the ecliptic and, often, the 'prime mover'. The motion of these 10 spheres was, according to ARISTOTLE, circular and uniform in accordance with the nature of the element, ether, of which they consisted. The motions of the four elements that constitute the sublunar world were linear with respect to the center of the Earth, the elements earth and water moving 'down' towards the center, and the elements air and fire moving 'up' away from the center. The circular motions of the celestial spheres disturb the linear motions of the four sublunar elements, thereby causing change in and on the Earth and in the surrounding layers of water, air and fire. The resulting combinations of elements form the basic materials out of which terrestrial and atmospheric bodies are shaped, which are then continuously further affected by the motions of the celestial bodies and by subsidiary interactions of sublunar bodies. Thus the UNIVERSE is a vast but finite machine powered by the natural motions of its five constituent elements, among which the ETHER with its regular circular motion has the major effect, and astrology is the science that investigates the operations of this universal machine.

Celestial omens

In ancient Mesopotamia and in the many cultures influenced by it many celestial phenomena—eclipses, conjunctions of the Moon and the planets with each other and with the stars, a large number of phenomena due to the distortion of light as it passes through the atmosphere, and other appearances—were regarded as messages sent by the gods to warn the rulers of men or collective groups of men of impending disaster or good fortune. The interpretation of these celestial omens was also a science by which men of great learning strove to provide useful advice to kings and other government officials. The experience of these learned readers of omens led to the recognition of the periodic behavior of the Sun, the Moon and the planets, and ultimately to the development of mathematical methods of predicting the phenomena themselves (which, because of their predictability, should then have been, but were not, no longer regarded as messages sent by angry or well-pleased gods). It also led to the development of some techniques of interpretation of celestial omens that were carried over into astrology, including that of predicting the fate of a person from the positions of the planets at his or her birth or computed conception. In practice, these protohoroscopes of the last four centuries BC were not a part of astrology since the phenomena were ominous rather than effective and since the elaborate geometry used in astrology was not a factor in their interpretation.

The thema or horoscope

The basic tool used to interpret the influences of the celestial bodies at any given moment—say, that of a native's birth—was the diagram called the thema or horoscope. This diagram represents the ECLIPTIC, divided into 12 zodiacal signs of 30° each, as a circle or a quadrilateral with the four cardinal points marked on it—the ascendant on the eastern horizon, the descendant on the western horizon, the midheaven to the south at a point computed by using the local oblique ascensions, and the antimidheaven 180° opposite it to the north. Between and at these cardinal points are located the cusps of the 12 astrological places. The zodiacal signs are divided into various subdivisions: decans of 10° each, twelfths of 2°30′ each, and terms of varying length, usually five to a zodiacal sign. The longitudes of the planets are noted in the thema, which locations reveal their configurations (aspects) with each other—conjunction, sextile (to 60°), quartile (to 90°), trine (to 120°) and opposition (to 180°)—which may be computed by various mathematical formulae so as to be measured in oblique or right ascensions or so as to accommodate the planets' latitudes. Moreover, various lots (defined as the distance between two real or imaginary celestial bodies measured off from a third) and the prorogator (computed in various ways) are noted on the thema, as may also be such fixed points as the exaltations (points of maximum effectiveness) of the planets and the four triplicities (Aries, Leo, Sagittarius; Taurus, Virgo, Capricorn; Gemini, Libra, Aquarius; Cancer, Scorpion, Pisces).

Domains of influence

Each zodiacal sign is classified according to numerous categories: masculine–feminine; odd–even; animal–human; village–forest; etc. Moreover, there are zodiacal melothesias, assigning different parts of the human body to the zodiacal signs; topothesias, doing the same with types of terrain on the surface of the Earth; lists of the animals, the professions, the metals, the jewels, the colors, etc of the zodiacal signs. Each of the signs has especial influence over the categories assigned to it. The triplicities, being four in number, are each associated with a cardinal direction and an element.

The 12 astrological places each have dominion over a certain aspect of the native's life (bodily form; inherited wealth; siblings; parents; children; illnesses; marriage; manner of death; travels; profession; gain; loss) or of an enterprise undertaken. They are influential in their domains in accordance with the zodiacal signs they overlap and the presence within them or aspects to them of the planets and the lots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The seven planets also (nine in India, where the ascending and descending nodes of the Moon are included) have each their own domains among physical types of humans and animals as well as psychological types, and each is related to types of terrain, of animals, of plants, of minerals, etc, and to an element and its humors (fire–hot and dry; water–cold and humid; air–hot and humid; earth–cold and dry). They are also classified as male or female, and as benefic, malefic or neuter. As the seven planets revolve through the zodiacal signs, and they in turn revolve about the stationary Earth, the influences they project into the sublunar sphere are constantly changing, creating ever-new events and transformations on Earth. These can, according to astrological theory, all be predicted on the assumption (alleged to possess some empirical foundations) that the accepted domains of the celestial bodies' influences are indeed correct. Historically, the system did not remain static; it was simply not complex enough to differentiate between the apparently limitless variations of human experience. In order to accommodate these variations, astrologers—especially those of India and Islam, who were duly followed in medieval Europe and in Byzantium—devised many more subdivisions of zodiacal signs in order to multiply the potential number of the effects of a planet, invented elaborate methods for assigning a numerical weight to the influence of each planet and expanded the number of effective bodies by introducing new lots.

Types of astrology

The paradigmatic type of astrology is genethlialogy, the science of interpreting the themata of nativities or of computed times of conception, probably invented in the 1st century BC in Ptolemaic Egypt. Such predictions can be and were also applied to animals and even plants. Although in theory the predictions are unique for each moment at a given locality, many astrologers wisely advised neophytes to learn something about the social and economic status of the native's family before attempting to describe his or her future life. A branch of genethlialogy is continuous horoscopy, which allows for a life-long usefulness of the astrologer. Since the planets and the ecliptic do not cease their revolutions, one can find their changing influence on the original life-pattern determined from the birth horoscope by casting a new horoscope for every anniversary of the birth, or for the beginning of every month or even day of the native's life, and one can also compute the time when each planet transits to another zodiacal sign or astrological place, thereby altering its effects, and one can watch the imaginary motion, in oblique ascensions, of the prorogator through different planets' terms and past different planets' bodies and aspects.

Catarchic astrology, invented simultaneously with the genethlialogical version, allows one to predict the course of some human undertaking from the horoscope of the moment of its inception. There are two major subdivisions of catarchic astrology, relating to marriage and to warfare. Some aspects of medical astrology are also catarchic, although other aspects depend on genethlialogy.

These two basic forms of astrology were transmitted from Greco-Roman Egypt to India in the 2nd century AD. In India interogational astrology was developed from catarchic astrology. In this form the astrologer answers specific questions posed by his client on the basis of the thema of the moment the question was asked. This clearly raises serious questions about the general theory of astrology; for either the client is exercising free will in determining when he or she will ask the question, so that there ceases to be a purely physical connection between the celestial bodies and the terrestrial event, or else his or her choice is determined by the stars and therefore could, in theory, be predicted by astrology and answered without recourse to a new thema.

These three types of astrology were transmitted from India to Sasanian Iran in the 3rd–5th centuries AD, as were the first two from the Roman Empire. The Sasanians added to them a fourth major type, historical astrology, which uses forms of continuous horoscopy (e.g. the horoscopes of the commencements of all or of selected years) as well as conjunctions of the two furthest of the then known planets, Saturn and Jupiter, or conjunctions of the two malefic planets, Saturn and Mars, in Cancer. Practitioners of historical astrology also cast the horoscopes of the coronations of kings and other historical beginnings, in these cases following the rules of catarchic astrology.

The question of astrology's validity

Clearly, whatever claims were made for an empirical basis, this whole theory is founded on arbitrary assumptions concerning the relationships between the individual zodiacal signs and planets and the physical parts and psychological aspects of sublunar objects and beings, as well as on a subsequently disproved theory of the COSMOS. In antiquity the arguments against astrology's validity were largely based on practical problems: how to explain the often great differences between the physical and mental attributes of twins and their separate lives; how to explain the simultaneous deaths of large numbers of people of different ages in a battle or natural catastrophe; how to be sure of the exact moment of birth and of the local ascendant at that moment. The Christian Church, and later Islam, frequently condemned astrology because it denies man or woman free will and so deprives him or her of the capacity to gain salvation through his or her own efforts and because it makes God irrelevant to the universe, which is an eternal and automatic machine. None of these arguments seriously damaged the credibility of astrology among many, nor has the collapse of the geocentric theory, although that has discouraged many scientifically minded persons from granting it any credence. Like several other current sciences that in fact continuously demonstrate their inability to predict the future accurately, astrology is still able to attract and to retain the patronage of innumerable clients.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Bouché-Leclercq A 1899 *L'astrologie Grecque* (Paris)
Koch-Westenholz U 1995 *Mesopotamian Astrology* (Copenhagen)
Pingree D 1978 *The Yavanajātaka of Sphujidhvaja* 2 vols (Cambridge)
Pingree D 1997 *From Astral Omens to Astrology, From Babylon to Bīkāner* (Rome)

David Pingree

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrometric Surveys

The meaning of astrometric surveys

Applied to astronomy, a survey has been defined as 'an examination for purposes of ascertaining the quality or quantity of something. Depending on the nature of the survey, this may mean looking at a large number of objects in an attempt to discover those meeting specific criteria.' (Warren 1992).

What is not specified in the above definition is whether the survey of objects meeting the established criteria must be the result of an observational program or a search in the literature. Although the term survey in astronomy is quite often applied to observational programs, some recent astrometric surveys have made use of the material published in the available literature. What is more, the new generation of astrometric surveys may even be carried out over the web (see 'The new millennium', below). Therefore, when speaking about surveys, we may refer to observational astrometric surveys, bibliographic astrometric surveys and the coming computer astrometric surveys.

When can a survey be called astrometric? In other words, what are the main characteristics of an astrometric survey? To answer these questions, the first point to consider is what is astrometry? Astrometry is that part of astronomy dealing with the positions (absolute or relative), motions and trigonometric distances of celestial objects (see also ASTRONOMY: TELESCOPES AND TECHNIQUES). Each of these parameters has its own importance in the study of our stellar environment.

What should be the accuracy of a given astrometric parameter? It is important to define the level of accuracy because that value will, to some extent, determine the quality of the survey. For example, no one would attempt to carry out an astrometric survey today with the accuracy level of the Durchmusterungs (Bonner, Córdoba or Cape) since such a level would be extremely low for today's standards; however, the importance of these surveys at the time they were published is undeniable.

What is an astrometric survey? Putting all the above considerations together, an astrometric survey can be defined as a search oriented to find those celestial objects meeting a given accuracy in their absolute or relative positions, motions, or trigonometric distances. Some of these surveys have been or are being conducted with ground-based meridian circles, ASTROLABES or astrographic telescopes, while others—already completed or planned for the future—have used or will be using space-based instruments.

Brief historical account of astrometric surveys

The compilation of catalogs including positions and PROPER MOTIONS is among the best-known activities in astrometry. In the early beginnings of astronomy, providing positions of celestial bodies was a real necessity for many of the ancient cultures. With roots dating back to the pre-Greek civilization, astrometry provided the information for computing solar and lunar eclipses and, more importantly, for the determination of time.

Thanks to the positions obtained by his predecessors and his own observations, HIPPARCHUS was able to identify the precession of the Earth's axis of rotation. It was also thanks to old and new astrometric catalogs that Edmund HALLEY introduced—in 1718—the proper motion of the stars. Some years later, in 1802, HERSCHEL noted the relative orbital motion for the two components of Castor, marking the beginning of double-star astrometry. The next landmark appeared in 1838, when Friedrich Wilhelm BESSEL was able to determine the PARALLAX of 61 Cygni. The history of astrometric catalogs—taking the detector used in the observational process as a parameter—could be divided in three different periods: visual (with or without an instrument), photographic and CCD based (Earth- or space-based instruments). With regard to the epochs in which the different detectors were employed, it could be said that, in general, the unaided human eye was used as a detector from the very beginning of astronomy until the year 1610, when GALILEO GALILEI started to use the telescope during his observations. After 1610, meridian circles or some other types of optical instruments were used to aid the human eye in most of the astrometric surveys. This approach lasted for over 300 years.

One of the first catalogs ever compiled was prepared by Hipparchus. It was completed around 129 BC (there has been a debate concerning the number of objects listed by Hipparchus: some historians have agreed that there were 1080 stars while others consider that there were only 850 stars). Hipparchus' catalog was included in Ptolemy's Almagest, the biggest compendium of astronomy of that time, which was in use for about 14 centuries. Previous compilations by Greek and Chinese astronomers date back to roughly 300 BC and even to 360 BC (Eichhorn 1992). During the Middle Ages the most important compilation worth mentioning—although it may not be considered a catalog in the current sense—is called the Alfonsine Tables, compiled under the reign of Alfonso the Wise and first published—probably—after 1277. Actually, the real purpose of the Tables was to offer a means to calculate the positions of the planets; one of the main users of the Alfonsine Tables was Copernicus (Gingerich 1992). Another catalog compiled in the first half of the 15th century was that of ULEGH BEG.

TYCHO BRAHE is among the first astronomers who—using his mural quadrant—started to make accurate systematic observations during the pre-telescope years. His observations include the supernova of 1572, the comet of 1577 and a large collection of positions for Mars. After Brahe's death in 1601, these positions were used by his

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

assistant JOHANNES KEPLER and were a key element in the derivations of Kepler's laws.

With the introduction of the telescope as an astronomical instrument, many zone (independent) catalogs started to be observed by means of meridian circles. Most of the observational surveys of the early to mid-1800s were limited to some specific areas and there was no agreement on the 'system' that should be used (the convenience of having a system in the modern sense of the term to which positions have to be referred was proposed in the late 1800s). One of the first attempts to carry out a comprehensive survey was started by ARGELANDER with the observation of the zone from +80° to –2°. The zones from –2° to the south pole were added later. This first survey—known as the Durchmusterungs—was made in three parts: the Bonner Durchmusterung, the Córdoba Durchmusterung and the Cape Photographic Durchmusterung.

The most important work of the early years of the 20th century was the General Catalogue (GC) (Boss 1937), a preliminary version of which was published by Lewis Boss in 1910. This version was called the Preliminary General Catalogue (PGC), and it was later completed with the publication of the GC in 1937, which included over 33 000 stars. What is important to highlight about the PCG/GC project is Lewis Boss' ultimate aim: the construction of a fundamental system. This goal was finally accomplished by his son, Benjamin Boss. In this sense, the N30 catalog compiled by Morgan (1952) is also worth mentioning.

For a detailed historical account of astrometric catalogs see Eichhorn (1974, pp 101–278). Also see Jaschek (1989) for a discussion of the different types of catalogs.

When the photographic plate was finally accepted by the astronomical community as a detector (around 1880), the photographic astrometric surveys started to develop. Curiously, the first observing program with the newly introduced astrographic telescope (the accompanying instrument to the photographic plate) was the observation of the biggest international project of the last years of the 19th century: the Astrographic Catalogue (AC). This program received a grat deal of criticism, as many considered that it was too early in the development of astrographic telescopes and photographic plates to undertake such a large project using a rather small (2° × 2°) field of view. However, from a current-day perspective, the launch of the AC can be regarded as a positive initiative. Thanks to that decision, astrometry now has a marvellous set of over four million first-epoch positions (nowadays 100 years old) for high-accuracy proper motion determinations.

Although the AC—in itself—should be considered an observational survey, it was never used as such until very recently. In this sense, two major catalogs have been produced using the published coordinates of the AC: the 4 Million Catalogue (Gulyaev and Nesterov 1992) and the AC2000 (Urban *et al* 1998). The former combined the AC with the Hubble Space Telescope Guide Star Catalogue (GSC) in order to compute proper motions; both sets of positions (AC and GSC) were reduced into the PPM system. In addition to positions and proper motions, the 4 Million Catalogue also provides cross identifications with other surveys such as the Durchmusterungs, PPM, SAO and VARIABLE STARS. The AC2000 contains over four and a half million positions (and magnitudes) with an average epoch of 1907. In order to reduce the published x, y coordinates of the AC, Corbin and Urban (1990) first compiled the ACRS to be used as the reference frame. The positions of the AC2000 are on the Hipparcos/ICRS system. A more recent reanalysis of the AC2000 has yielded the AC2000.2 (Urban *et al* 2001).

Apart from the 4 Million and the AC2000, there are some other surveys that have made use of the AC. Of these, the PPM was the first.

For a detailed and enthralling history of the Astrographic Catalogue see Eichhorn (1974, pp 279–323).

While the observation of the AC was underway, other photographic—zonal—surveys were also in progress: the Yale zones and the AGK2 (the AGK1 was observed with meridian circles). In total, Yale published 30 catalogs with epochs ranging from 1928 to 1942. The first volume of the Yale zones to appear included over 5000 equatorial stars (Schlesinger *et al* 1926); the last one (the zone from –60° to –70°) suffered a number of delays and it was not published until 1983 (Fallon and Hoffleit 1983). The last one of the AGK series was the AGK3 (Dieckvoss 1971). Because of the development of artificial satellites and the urgent necessity of tracking them, by the mid-1960s, the need was felt to have a general comprehensive compilation embracing the existing catalogs into a single one. The realization of such a project was undertaken by the Smithsonian Astrophysical Observatory, and in 1966 the 'Star Catalogue' subtitled 'Positions and Proper Motions of 257,997 Stars for the Epoch and Equinox of 1950' became available to the astronomical community. The catalog, known as SAO (see Haramundanis 1966 for details of the construction of the catalog), covers from +90° to –90° and it was for many years the standard reference frame for the reduction of most of the photographic plates. Thousands of positions of asteroids, comets, major planets and natural satellites were obtained using the SAO as a primary reference catalog for stellar positions and proper motions. However, as is well known, time—through the proper motions—is always responsible for the degradation of any astrometric catalog. The SAO was no exception, and, after being in use for over 25 years, it was replaced by the PPM (Roeser and Bastian 1993). The PPM—also covering the entire sky—lists over 375 000 positions and proper

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

motions in the J2000/FK5 system. Typical values of the r.m.s. errors in the northern hemisphere are 0.27 arcsec in the positions at epoch 1990 and 0.42 arsec century^{-1} in the proper motions, while the corresponding figures in the southern sky are much better (Roeser *et al* 1994). The main improvement of the PPM over the SAO was made through the inclusion of the century-old AC data. After its publication, the PPM replaced the SAO as the reference frame for most of the photographic astrometry programs. Because of the discontinuation of photographic plates and the advent of CCD detectors, the PPM has now been replaced by catalogs of much higher density.

While some astronomers were engaged in the observation and reduction of large catalogs of positions, others focused their attention on the search of stars with high proper motions. According to Luyten (1963), this type of project benefited greatly from the introduction of the blink microscope. Several surveys for stars with large proper motions were carried out by Luyten (1970), resulting in the detection of a large number of objects. Subsets of these surveys and/or a revision or reorganization of them were published as the Luyten Half Second (LHS), Luyten Two Tenths (LTT) and the New Luyten Two Tenths (NLTT) surveys. Most of the results were published by the University of Minnesota, Minneapolis (USA). Among the large proper motions surveys, the Lowell Proper Motion Survey (Giclas *et al* 1971) is one of the most important, as it includes finding charts for each star.

In order to present a more complete historical account of astrometric surveys, a reference should be made to the Hipparcos mission. The Hipparcos satellite—funded and operated by ESA—was the first space experiment dedicated to astrometry. Its primary goal was the determination of accurate positions, proper motions and parallaxes of 118 218 stars. The accuracies of the astrometric parameters are roughly 1 mas for the positions and parallaxes and 1 mas yr^{-1} for the proper motions for stars as faint as magnitude 10. The positions and proper motions of the Hipparcos Catalogue (ESA 1997a) are the optical realization of the ICRS and they constitute the most accurate large astrometric catalog ever observed. The importance of Hipparcos is that it has for the first time enabled astronomers to calibrate accurately the luminosities and masses of many nearby stars and to study the kinematic structure of the Galaxy in the solar neighborhood. The impact that Hipparcos has had in astronomy can be appreciated in the proceedings of the Hipparcos Venice '97 meeting (ESA 1997b) which contains nearly 200 research papers based on Hipparcos' data.

As part of the Hipparcos mission, the 'stars mappers' of the satellite were used for the construction of a less accurate catalog: the Tycho Catalogue. A re-analysis of the original data of this experiment yielded the Tycho-2 Catalogue (Høg *et al* 2000) which has become one of the standard compilations in the astrometric reduction of the CCD frames (mainly in the meridian circle observations). More details about HIPPARCOS and the HIPPARCOS AND TYCHO CATALOGS are given in other articles.

In the last 14 years, the history of astrometry can be said to have been signaled by two main events: the first one was the launch of the above-mentioned Hipparcos mission in 1989; the second one was the establishment—on 1 January 1998—of the International Celestial Reference System (ICRS) as the new reference system, thus bringing to an end the regime of the FK series. With this major change, and for the first time, astrometry's reference frame—the International Celestial Reference Frame (ICRF)—to which positions have to be referred, is based on 212 extragalactic RADIO SOURCES observed with very-long-baseline interferometry (VLBI) techniques (see Ma *et al* 1998). An important review concerning the adoption and implications of this change has been made by Feissel and Mignard (1998).

Today, after significant achievements through its history and thanks to the use of a new generation of extremely high-quality Earth- and space-based instruments, astrometry is undergoing new and profound changes. From the multiple-fiber-fed instruments on Earth to the most advanced and sophisticated satellites, there is a common need for better positions of celestial objects and astrometry, once again, has to provide accurate positions, proper motions and stellar parallaxes. The fact that some of the coming space missions need to rely on ground-based astrometric observations for their success illustrates the importance of the new astrometry.

In order to meet the requirements imposed by today's astronomy, new astrometric surveys oriented to providing the classical parameters are underway. Because of the increasing power in both computer and detector development, the current projects are growing in size (more than 500 million entries in general) and accuracy (a few mas). Some of these new surveys are based on new observations, while many others are making use of the extensive material provided by the databases, thus bringing up a new way of doing astronomy in general: the virtual observatories (see also ASTROMETRY: HISTORY).

Major achievements

Over the last few years there has been outstanding progress in most of the astrometric projects. These can be summarized as follows.

Positions and proper motions

Since the CCDs became the standard detectors in almost every astronomy project, there has been a real need for catalogs of positions (and, ideally, proper motions as well) containing a large number of stars. In some sense it

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

could be said that astronomy is—again—experiencing a similar situation to the one that occurred in the mid-1960s when the SAO was published (then replaced by the PPM). Because of the low number of objects they include (by today's standard), these two catalogs are now inappropriate for the astrometric reduction of most CCD frames.

In order to overcome the problem of having too few stars per square arcminute, it was necessary to compile much larger catalogs. The first step in this sense was taken by the Hubble Space Telescope Guide Star Catalogue (GSC) by Lasker *et al* (1990), which now has been superseded by the GSC II in both quantity and quality.

Another major compilation worth mentioning is the USNO-A2.0 (Monet *et al* 1998), which is the standard—at least for the moment—for the astrometric reduction of most of the CCD frames. A subset—called USNO-SA2.0—containing about fifty million stars is mostly used in the reduction of minor planet and comet CCD frames.

As usual, after a catalog is published, its contents are analysed and externally compared with other compilations for a final evaluation of quality. The USNO-A2.0, as well as the GSC II, is no exception. A very detailed analysis of the USNO-A2.0 has been performed by Assafin *et al* (2001). In this paper, the authors have compared some of the positions of USNO A2.0 against the ones observed with the Valinhos CCD Meridian Circle (Viateau *et al* 1999). An extension of the USNO-A2.0 catalog, the USNO-B, is currently in progress (see Monet 2000). It will include proper motions and star/galaxy classifications.

One step above the USNO-A2.0 (at least from the quantity point of view) is the GSC II. When finished it will have close to 109 entries with a typical error of 0.3 arcsec (the worst case is between 0.35 and 0.65 arcsec for stars fainter than 16th magnitude near the edge of the plate). A preliminary version of this compilation, known as the GSC 2.2, is ready for use, listing more than 430 million stars. The positions of the GSC II, as well as the GSC 2.2, are in the system of the ICRF. The mail goal of the GSC 2.2 is to support the opration of both GEMINI and VLT telescopes.

Momany *et al* (2001) have performed an external comparison of the GSC 2.2. As part of the ESO Imaging Survey (EIS) these authors have compared the positions of GSC 2.2 with those reported by others in the area of some open clusters.

Probably the most recent project to have accomplished part of its goals is the USNO CCD Astrograph Catalogue (UCA). A preliminary catalog—the UCAC1—in the ICRS contains over 27 million stars with a positional precision of less than 20 mas for objects in the magnitude range 9–14. Preliminary proper motions from a variety of sources are also given. The UCAC1 will be superseded by the UCAC2—covering from –90° to +30°—which will be available by mid-2002. When finished, the UCA project will provide the first astrometric catalog entirely observed with an Earth-based telescope using a CCD as a detector.

In order to improve the quality of the proper motions to be included in the UCAC project, the AGK2 plates (more than 70 years old at this moment) are being re-measured at USNO and will then be combined with modern positions (see Winter and Holdenried 2001). Various progress reports on the UCAC project can be found in Zacharias *et al* (2000a, b, 2001).

Among the projects oriented towards the determination of absolute proper motions, two programs have provided a large number of data: the Northern Proper Motion Program (NPM) developed at Lick Observatory (Klemola *et al* 1987) and the Southern Proper Motion Program (SPM), which is being developed by a Yale University (USA)–National University of San Juan (Argentina) consortium. The first batch of the NPM—called NPM1—was released in 1993. It contains positions, proper motions and two-color photometry for 149 000 stars outside the Milky Way. The NPM2 (Jones *et al* 2000), the second batch of the series, will contain around 300 000 stars located in the Milky Way. Since reference galaxies are 'absent' in the Milky Way sky, the Hipparcos Catalogue is being used to link the NPM2 proper motions to the (absolute) system defined by NPM1. The NPM2 will be released in 2003.

The first-epoch observations of the SPM were started in 1965 and lasted until 1974. The second epoch, using photographic plates, began in 1987 and was stopped in 1993 owing to the lack of photographic plates. Now the observation of the second epoch will be resumed by using two CCD cameras (van Altena 2002). Two major results from the SPM have been released: the SPM 1.0 (Platais *et al* 1998) and the SPM 2.0 (for a detailed analysis of the SPM 2.0 proper motion system and its relation with Hipparcos see Zhu 2001). Also, as part of the SPM project, a catalog of positions for approximately 30 million stars has been announced (Platais *et al* 2001b).

Results from another major digitization process of photographic plates have recently been published. The astrometric properties of the SuperCOSMOS survey have been analysed by Hambly *et al* (2001). By comparing the first release of data—the south galactic polar cap—with external data sets, they have concluded that the positions (in the ICRF via the Tycho 2) are accurate to better than ±0.2 arcsec at magnitudes J, $R \sim 19$, 18 while the proper motion measurements are shown to be accurate to typically ±10 mas yr $^{-1}$ at the same magnitude level.

Open-cluster astrometry has also received very close attention. A number of articles on the subject have been published in the past few years. Some of them have

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

made use of the extensive databases available while others have relied on new observations. In both cases—database analysis or new observations—the outcome has been the determination of positions, proper motions and magnitudes for a large number of stars. An important survey, which is being conducted with the WIYN telescope, is the WIYN Open Cluster Study (WOCS). Results from this survey have been published by Platais *et al* (2001a, c).

There is a significant number of contributions to the open-cluster studies that have used databases and highly accurate astrometric catalogs. Among them are Baumgardt *et al* (2000); Dias *et al* (2001), Sanner and Geffert (2001) and Masden *et al* (2002).

WEBDA is an updated collection of open-cluster data maintained by Jean-Claude Mermilliod at the Institute of Astronomy of the University of Lausanne (Switzerland). The database has very simple features that ensure easy retrieval of data and offers astrometric, photometric and spectroscopic information. Miscellaneous data (such as membership probabilities) and bibliographic references are also included.

Double stars

The most important compilation of DOUBLE STARS is the Washington Double Star (WDS) catalog. The main advantage of the compilation is that it is updated on a daily basis. A series of four CDs has been recently published by the US Naval Observatory. Volume I (Mason *et al* 2001) contains the basic astrometric data (coordinates for J2000, position angle, separation, etc,) for over 84 000 systems. The second volume (Hartkopf *et al* 2001) is a collection of 1465 orbits. Volume III (Hartkopf *et al* 2001) is actually the Third Catalogue of Interferometric Measurements of Binary Stars. Volume IV (Worley *et al* 2001)—the last of the series—is a collection of magnitude differences for over 10 000 systems.

Since the early 1980s, CHARA has been the leading survey group using speckle techniques to measure the separation and position angles of double stars (basic information for stellar mass determinations). To date, 19 catalogs with position angles and separations for approximately 10 000 stars have been published (Mason *et al* 1998). The group has finished the construction of an interferometric array of six 1 m telescopes to resolve binaries with angular separations down to 0.2 mas. The CHARA array will be fully operational sometime in 2002. Progress reports on this array can be found in McAlister (2001), Turner *et al* (2001), Sturmann *et al* (2001) and ten Brummelaar *et al* (2001).

Other important surveys—using speckle techniques—have continued the observation of double stars. Recent reports of the Rochester Institute of Technology (RIT) survey can be found in Horch *et al* (2000) and Horch *et al* (2001b). Another important result from this group is related with the magnitude differences for about 130 pairs (Horch *et al* 2001a). A summary of the last 4 years of activity has been presented by Robinson *et al* (2001) while the new camera—called RYTSI—designed by the group is described in Horch *et al* (2001c). Results from the Pic du Midi observatory group, obtained with the Pupil Interferometry Speckle and Coronograph (PISCO) system (Prieur *et al* 1998), have been published by Scardia *et al* (2000) and Prieur *et al* (2001). On the other hand, angular separations, position angles as well as magnitude differences for 606 systems—using a 'regular' CCD camera—have been reported by Gili and Bonneau (2001) and by Lampens *et al* (2001).

Mason *et al* (2000) have re-analysed the information contained in Tycho-2 in attempt to provide basic data for double stars. Another revision and re-reduction of the Tycho-2 by Fabricius *et al* (2002) has resulted in the discovery of over 13 000 new doubles mostly with separation between 0.3 and 1 arcsec. A proper combination with WDS data has permitted the compilation of the Tycho Double Star Catalogue (TDSC) which has more than 100 000 entries. The TDSC is available from the CDS.

Parallaxes

Within the astrometry context, the word parallax is almost always associated with trigonometric parallaxes; that is, a direct way of determing stellar distances. The method is limited to distances less than 100 pc.

The history of the parallax determinations—like that of the astrometric catalogs—is very rich. It is said, for example, that the first attempt to get the parallax of a star was made by Tycho Brahe. Since he could not get any reasonable value he concluded that either (a) the Earth was at rest in the center of the universe or (b) the stars were too far away to make it possible to measure their parallaxes. Since Brahe did not believe the stars could possible be so far away, he chose the first option. This result allowed him to confirm that Copernicus was absolutely wrong in his heliocentric theory of the solar system.

As already mentioned, it was Bessel who—in 1838—detemined the first triginometric parallax of 61 Cygni. By 1890 there were about 30 stars whose parallaxes were known to an accuracy of ±0.05 arcsec (Hoffleit 1992).

By the end of the 19th century—with the adoption of the photographic plate as a detector— the determination of trigonometric parallaxes became a standard research program at some of the most important observatories. In this sense, in 1903, Frank Schlesinger—known as 'the father of modern astrometry' (Hoffleit 1992)—started at the Yerkes Observatory one of the outstanding projects of those days (see Schlesinger 1910). Other institutions such as McCormick, Allegheny, Sproul, Greenwich, Cape, Boschha and Stockholm soon initiated similar projects while Lick Observatory and the USNO Flagstaff Station

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

started later using their reflecting telescope, and both extended the determinations to fainter stars.

Most of the work on the determination of trigonometric parallaxes during the photographic era was summarized in the different editions of the General Catalogue of Trigonometric Parallaxes—better known as the Yale Parallax Catalogue (YPC). The first edition of the series was prepared by Schlesinger (1924) and Schlesinger and Jenkins (1935). The third edition was published by Jenkins (1952), followed by her 1963 supplement. The fourth edition of the YPC, the last one of the series, was prepared by van Altena *et al* (1995). It lists trigonometric parallaxes for 8112 stars including photographic as well as CCD determinations.

Once the CCD became available for scientific use, different programs aimed at the determination of trigonometric parallaxes were started. Two of the first reports using CCD techniques are those by Monet (1981) and Monet and Dahn (1983). Fortunately, trigonometric parallax determinations were also made in the southern hemisphere during the early stage of the CCD technique era (see Anguita and Ruiz 1988).

Today, because of the current need of astronomy in surveying and cataloging the solar neighborhood, nearby stars have gained a new interest and so new distances are being determined. From detailed studies of single stars to hemispheric surveys, new results are published almost every year. Among the single-star results we can mention new insights about Proxima Centauri (our nearest neighbor, discovered by R. Innes in 1915) and Barnard's Star by Benedict *et al* (1999); even some evidence for a companion to Proxima Centauri has been mentioned (see Schultz *et al* 1998). Also in the single-star category, it is worth mentioning that a field star that may be the nearest M9 dwarf to the Sun has been identified (Deacon and Hambl 2001). Probably one of the most important results (mainly for the consequences it will have) is the determination of the absolute parallax and relative proper motion of RR Lyrae, a distance scale calibrator, by Benedict *et al* (2002).

The different searches for nearby stars have also reported new results. Among them a survey aimed at the identification of brown dwarfs within 8 pc of the Sun (Hinz *et al* 2002) has revealed no new stellar companions at wide separations; this program is making use of already existing first-epoch images which are combined with new J-band observations.

A search oriented to finding the youngest and nearest (d<60 pc) stars to Earth (using x-ray information as well as astrometric data from Hipparcos and Tycho) is under way by Song *et al* (2001). The final goal of the project is the detection of giant planets. The observations are being carried out from Lick observatory for the northen targets and from Siding Spring Observatory in the south.

A detailed analysis of the Automatic Plate Measuring (APM) measurements of the UK Schmidt Telescope plates by Scholz *et al* (2002) has resulted in the discovery of previously unknown bright ($12<R<15$) high proper motion stars; among them, there are some nearby (d<25 pc) stars.

In the southern hemisphere the Research Consortium on Nearby Stars (RECONS) is engaged in the discovery of new solar neighbors. A large parallax determination program being carried out at CTIO (CTIO Parallax Investigation, CTIOPI) has revealed several nearby stars; one of them is located at 5.5 pc (see Henry *et al* 2001, Jao *et al* 2001). An astrometric survey of companions to nearby stars, with observing telescopes to be located in both hemispheres, is under consideration (Kaplan *et al* 2001).

While in the past most of the parallaxes were determined by ground-based telescopes, the largest number of today's parallaxes are being determined by space satellites. Hipparcos provided almost 120 000 determinations at the accuracy level of 1 mas and SIM and GAIA are expected to furnish a much larger body of data (including parallaxes). In this sense, GAIA for instance will provide astrometric data at the 10 mas level accuracy at V=15 (Torra *et al* 2001) and accurate positions and radial velocities for as much as 1% of the stellar population of the Galaxy (Perryman *et al* 2001). It is obvious that the future of stellar parallaxes—and astrometry in general—looks very promising.

Other projects

Now that CCDs have became the standard detectors for astronomy, it is probably time to change the old classification of astronomical instruments. The most striking case is with meridian circles (MCs). From their first use in astronomy until some years ago, stars were observed by eye (one at a time) when they crossed the meridian. The next step in the development of observing techniques was to replace the human eye by photomultiplier tubes. Now that CCDs have been attached to most of the MCs around the world, the observing procedure has been replaced by the drift scanning mode. This approach allows one to obtain strips from a few minutes up to some hours long in right ascension (Muiños 2002) by some 10–20 arcmin (depending on the size and type of CCD chip) in declination. This data set could easily be compared with that from a normal astrographic telescope. The observations are then reduced using any reference catalog.

Also, some MCs have been engaged in the observation of objects that were traditionally observed by astrograph-type telescopes: minor planets (probably this could be interpreted as a 'return to the roots' since Ceres was discovered by Piazzi observing with an MC). Among these observations we can mention the ones included in

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

the different Carlsberg Meridian Circle Catalogues or those observed by the San Fernando Automatic Meridian Circle—located in San Juan, Argentina, since 1996. Probably the most comprehensive report in this sense is the one by Stone (2000) which includes more than 56 000 positions (in the ICRF) for 1940 asteroids.

The Bordeaux Automated Meridian Circle has just released the M2000 catalog (Rapaport *et al* 2001), which lists over two million stars down to the magnitude limit 16.3. The declination band observed is coincident with the Carte du Ciel observed at Bordeaux. The median internal standard error in position is on the order of 35 mas, while an external comparison with Hipparcos and Tycho-2 catalogs yields an estimate better than 40 mas.

The San Fernando Automated Meridian Circle (equipped with a CCD since December 1999) is engaged in the observation of part of the southern sky to a limit around 16th *V* magnitude. The project is to observe from +3° to –60°. Until now approximately 30% has been completed. The strips (ranging from 20 min up to 3 h long in right ascension by 18 arcmin in declination) are being reduced using Tycho-2 as a reference frame (Muiños and Mallamaci 2002). The observations are expected to be finished sometime during 2005.

The MC of the Abrahão de Moraes Observatory (near Valinhos, Brazil) has been involved in miscellaneous projects. Among them, it is important to mention the extension of the ICRF down to the 15th *V* magnitude (Camargo *et al* 2001). This project includes more than 44 000 stars located around some extragalactic radio sources. This instrument has also been used in the determination of positions for proper motion determination in southern star-forming regions (Teixeira *et al* 2000).

Another important catalog completed during 2000 was prepared at the Nikolaev Astronomical Observatory. The observations were made with the axial meridian circle (AMC). The catalog—available from the web—lists over 14 000 stars and was reduced using USNO-A2.0 as a reference frame. The AMC was put in operation in 1995 (see Pinigin *et al* 1995) and since then it has been involved in observations for different projects (see Kovalchuk *et al* 1997, and Tang *et al* 2000).

Astrolabes have also taken part in many surveys. Over the last 20–25 years a number of improvements in the instruments have been introduced, one example being the Photoelectric Astrolabe II (PA II) developed in China (Luo and Li 1986). Recent results obtained with this new instrument have been published by Manrique *et al* (1999). Other smaller astrometric compilations—using photoelectric astrolabes—have been reported by Hui and Rui (2002) and Martin *et al* (1999). Other surveys, although astrometric in their scope, were actually designed and oriented to improve astrometric parameters of astrophysical objects. López and Girard (1990), for example, improved the positions of over 1000 variable stars. For this project, the authors measured the first-epoch plates of the SPM.

Another project aimed at the improvement of positions and proper motions for variable stars is the Astrometric Catalogue of Variable Stars (ACVS) by Gulyaev and Ashimbaeva (1997). For this study, the authors took the stars listed in the GCVS and the NSV and checked them against those in the 4 Million Catalogue (Gulyaev and Nesterov 1992). Using the finding charts recommended by the GCVS and NSV approximately 21 500 positive cross-identifications were made.

The SDSS (Gunn *et al* 1998) is a large-scale project designed to produce optical photometry and positions for stars down to magnitude 22 in approximately one-quarter of the sky centered on the north galactic cap. In addition, spectra of about one million galaxies, 100 000 quasars and 50 000 stars will be obtained. This 'non-astrometric' survey should produce an enormous quantity of high-quality positions, which, when compared with the first epoch of the POSS, will also result in proper motions. Observations for this project started in 1999 and are projected to last for 5 years.

Besides the optical wavelength projects mentioned in the previous paragraph, some radio-wavelength-based proposals are also making an important contribution to astrometry. Among them, one such proposal, very important because of its global scope, is VERA (VLBI Exploration of Radio Astrometry), which is sponsored by the National Astronomical Observatory of Japan and a group of Japanese universities. The project is aimed at the determination of positions, parallaxes and proper motions of maser sources in the whole Galaxy (see Honma *et al* 2000).

The new millennium

Astronomy in general and astrometry in particular are being favored with the large databases currently in use and those to come in the very near future. This new way of doing astrometry or extracting data for re-analysis includes the use of facilities such as Aladin Sky Atlas (CDS), Astrovirtel, Astrogrid (the UK Virtual Observatory), the SuperCOSMOS Sky Survey or the announced Astrophysical Virtual Observatory (AVO) and the National Virtual Observatory (NVO), among others. The enormous power of these tools—when fully available for public and general use—will allow the handling of different very large databases. Some of the facilities—such as VizieR or SIMBAD—also permit the visualization of all the data (coordinates, magnitude, proper motion, etc.) stored for a given object.

A number of sources will drastically change the way the new astrometric surveys will be performed in the years to come. Among them are the already existing means to extract images (frames) from the digitization of the major photographic surveys (POSS I, POSS II, etc.)—

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

known as Digitized Sky Survey (DSS)—added to the new 'bank of images' such as EIS, the MISAO project, or the Maria Mitchel Observatory (MMO). Another alternative place for images obtained from the digitization of major photographic surveys is the USNO Flagstaff Station (see Levine and Monet 2000). With all this material at the reach of our fingertips, through the web, it will not always be necessary really to have telescopes (big or small) to undertake outstanding astrometrically-oriented research programs. A number of projects, making use of these new possibilities, are currently underway.

Today the DSS has proved to be one of the most extensive bank of images in use. From the confirmation of supernovas (Evans *et al* 2001) to the identification of millions of galaxies (Paturel *et al* 2000), the DSS allows one to undertake almost any project. As an astrometric example of this possibility, we can mention the improvement of positions of ICRF sources (Da Silva Neto *et al* 2000) and almost three thousand objects listed in the Second Byurakan Survey of Blue Compact Galaxies (Bicay *et al* 2000). Also, the combination of DSS frames of different epochs or the blinking between DSS images with those from other surveys (such as the ones available in the MISAO image bank or those from SkyMorph) allows the identification of high proper motion stars (see Scholz *et al* 2001). Another possible way of locating 'missing' high proper motion stars is the combination of different surveys; a project using this approach is being developed by Gould and Salim (2001).

However, not only stellar objects can be included in the search of database compilations; non-stellar objects—such as minor planets and comets—are also being considered and a number of precovery positions have been submitted to the MPC, thus improving the determination of orbital parameters (for a detailed project in this sense see Boattini *et al* 2001). In addition, those frames obtained by some of the most important minor planets (NEOs in particular) searches, such as LINEAR, NEAT and LONEOS, are also appropriate for use in, for example, proper motion determinations. These frames are easily accessible through the SkyMorph facilities.

There are also a number of research programs making use of the already existing high-quality astrometric catalogs mainly Hipparcos and Tycho-2. These databases have been used in—among other projects—the study of proper motions of open clusters (Dias *et al* 2001) and in a star-forming complex (Belikov *et al* 2002). In some other cases, the high accuracy of these compilations (mainly Tycho-2) has allowed the determination of 'astrometric radial velocities' for stellar members of open clusters (see Masden *et al* 2002). The articles, as well as the web addresses, mentioned in this section are only a small (and far from complete) example of what could be done by using the material available in the major data centers. However, the potentialities of these new tools have not yet been exploited in all of their capabilities; once this happens, there will be a big impact in the way astrometry will carry out the new astrometric surveys from that moment on.

These new possibilities, plus the even larger body of data to be furnished by the coming space missions (GAIA, SIM, etc.) planned for the years ahead, will change the way the astrometric surveys of the near future will be done. They will be much easier and—at the same time—much more complicated to develop: easy because simply with the access to the web it will be possible to extract many megabytes of information; complicated, because it will require not only larger computer facilities to handle that amount of data properly but also (and probably more important), new concepts in computer technology (see Longo *et al* 2001). When this happens, astrometry will have the opportunity to grow and to expand into new and unexpected domains.

However, of course, not only astrometry will change; in fact, the most significant change will be in the way we see and understand the universe around us.

Appendix

Summary of Acronyms

A large number of acronyms are currently being used to refer to the astrometric surveys, instruments and institutions engaged in astrometric projects. Below is a summary of acronyms mentioned in the present work, and others found elsewhere in the current relevant literature.

2MASS	Two-Micron All Sky Survey
AC	Astrographic Catalogue
ACRS	Astrographic Catalogue Reference Stars
ACT	Astrographic Catalogue and Tycho
ACVS	Astrometric Catalogue of Variable Stars
AGK	Astronomisches Gesselschaft Katalog
APS	Automated Plate Scanner
AVO	Astrophysical Virtual Observatory
BD	Bonner Durchmusterung
CAMC	Carlsberg Automatic Meridian Circle
CCD	Charge-coupled device
CDS	Centre de données astronomiques de Strasbourg
CHARA	Center for High Angular Resolution Astronomy
CoD	Córdoba Durchmusterung
CPC2	Second Cape Photographic Catalogue
CPD	Cape Photographic
CTIO	Cerro Tololo Inter-American Observatory
CTIOPI	CTIO Parallax Investigation
DENIS	Deep Near-Infrared Survey
DSS	Digitized Sky Survey
EIS	ESO Imaging Survey
ESA	European Space Agency
ESO	Europen Southern Observatory

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

FAME	Full-Sky Astrometric Mapping Explorer
FK	Fundamental Katalogs
FOCAT-S	Photographic Catalogue—South
GAIA	Global Astrometric Interferometer for Astrophysics
GC	General Catalogue
GCVS	General Catalogue of Variable Stars
GSC	Guide Star Catalogue
HDEC	Henry Draper Extension Charts
HDF	Hubble Deep Field
HIPPARCOS	High Precision Parallax Collecting Satellite
HST	Hubble Space Telescope
IAU	International Astronomical Union
ICRF	International Celestial Reference Frame
ICRS	International Celestial Reference System
LHS	Luyten Half Second
LINEAR	Lincoln Near Earth Asteroid Research
LONEOS	Lowell Observatory NEO Search
LTT	Luyten Two Tenths
MISAO	Multitudinous Image-Based Sky-Survey and Accumulative Observations
NASA	National Aeronautics and Space Administration
NEA	Near-Earth asteroid
NEAT	Near-Earth asteroid tracking
NEO	Near-Earth object
NLTT	New Luyten Two Tenths
NOAO	National Optical Astronomical Observatories
NPM	Northern Proper Motion Survey
NSV	New Catalogue of Suspected Variable Stars
NTT	New Technology Telescope
NVO	National Virtual Observatory
OAFA	Observatorio Astronómico Félix Aguilar
PGC	Preliminary General Catalogue
PISCO	Pupil Interferometry Speckle and Coronograph
POSS	Palomar Observatory Sky Survey
PPM	Positions and proper motions
RECONS	Research Consortium on Nearby Stars
RIT	Rochester Institute of Technology
ROA	Real Instituto y Observatorio de la Armada
SAOC	Smithsonian Astrophysical Observatory Catalogue
SDSS	Sloan Digital Sky Survey
SIM	Space Interferometry Mission
SIMBAD	Set of Identifications, Measurements Bibliography for Astronomical Data
SPM	Southern Proper Motion Survey
SRS	Southern Reference System
TAC	Twin Astrograph Catalogue
TPMC	Tokyo Photoelectric Meridian Circle
TRC	Tycho Reference Catalogue
UCAC-S	United States Naval Observatory CCD Astrograph Catalogue—South
UKST	United Kingdom Schmidt Telescope
USNO	United States Naval Observatory
VERA	VLBI Exploration of Radio Astrometry
VLBI	very-long-baseline-interferometry
VLT	very large telescope
WEBDA	WEB Database of Open Clusters
WDS	Washington Double Star Catalogue
WIYN	Wisconsin Indiana Yale NOAO
WOCS	WIYN Open Cluster Studies
YPC	Yale Parallax Catalogue

Bibliography

Anguita C and Ruiz M T 1988 Southern CCD Parallax Program *Astron. Soc. Pac. Conf. Ser.* **1** 344–5

Assafin M, *et al* 2001 Investigation of USNO-A2.0 Catalogue positions *Astrophys. J.* **552** 380–5

Baumgardt H, Dettbarn C and Wielen R 2000 Absolute proper motions of open clusters. I. Observational data *Astron. Astrophys. Suppl. Ser.* **146** 251–8

Belikov A N, *et al* 2002 Study of the Per OB2 star-forming complex. I. The Compiled Catalogue of kinematic and photometric data *Astron. Astrophys.* **348** 145–54

Benedict G F, *et al* 1999 Interferometric astrometry of Proxima Centauri and Barnard's Star using Hubble Space Telescope Fine Guidance Sensor 3: detection limits for substellar companions *Astron. J.* **118** 1086–100

—— 2002 Astrometry with the Hubble Space Telescope: a parallax of the fundamental distance calibrator RR Lyrae *Astron. J.* **123** 474–84

Bicay M D, *et al* 2000 Accurate optical positions for 2978 objects from the Second Byurakan Survey (SBS) with the Digitized Sky Survey *Astron. Astrophys. Suppl. Ser.* **147** 169–73

Boattini A, *et al* 2001 The Arcetri NEO Precovery Program *Astron. Astrophys.* **375** 293–307

Boss B 1937 *General Catalogue of 33342 Stars for the Epoch 1950* (Carnegie Institution of Washington Publ. 486)

Camargo J, *et al* 2001 Extension of the ICRF for selected areas down to the 15th magnitude *Astron. Astrophys.* **375** 308–18

Corbin T and Urban S 1990 *Faint reference stars IAU Symposium 141, Inertial Coordinate System on the Sky* ed J Lieske and V Abalakin (Dordrecht: Kluwer) pp 433–42

Da Silva Neto D, *et al* 2000 Optical positions for a sample of ICRF sources *Astron. J.* **119** 1470–79

Deacon N and Hambl N 2001 The trigonometric parallax of DENIS-P J104814.7-395606.1 *Astron. Astrophys.* **380** 148–50

Dias W, Lépine J and Alesi B 2001 Proper motions of open clusters within 1 kpc based on the TYCHO2 Catalogue *Astron. Astrophys.* **367** 441–7

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Dieckvoss W 1971 The AGK3, a basis for general (northern) reference catalog of positions and proper motions *Conference of Photographic Astrometric Technique* ed H Eichhorn (Washington, DC: NASA) pp 161–7

Eichhorn H 1974 *Astronomy of Star Positions* (New York: Ungar)

—— 1992 Star catalogs, historic *The Astronomy and Astrophysics Encyclopedia* ed S P Maran (Cambridge: Cambridge University Press) pp 661–3

ESA 1997a *The Hipparcos and Tycho Catalogues* SP-1200

—— 1997b *Hipparcos Venice '97* SP-402

Evans R, *et al* 2001 *Supernova 2001du in NGC 1365 IAU Circular 7690*

Fabricius C, *et al* 2002 The Tycho double star catalog *Astron. Astrophys.* **384** 180–9

Fallon F and Hoffleit D 1983 Catalogue of the positions and proper-motions of stars between declinations –60° and –70° *Trans. Yale Univ. Obs.* **32** (2)

Feissel M and Mignard F 1998 The adoption of ICRS on 1 January 1998: meaning and consequences *Astron. Astrophys.* **331** L33–6

Giclas H, Burham R Jr and Thomas N 1971 *Lowell Proper Motion Survey (Northern Hemisphere). The G Numbered Stars* (Flagstaff, AZ: Lowell Observatory) pp 1–175

Gili R and Bonneau D 2001 CCD measurements of visual double stars made with the 74 cm and 50 cm refractors of the Nice Observatory (2nd series) *Astron. Astrophys.* **378** 954–7

Gingerich O 1992 *The Great Copernicus Chase and Other Adventures in Astronomical History* (Massachusetts: Sky Publishing Corporation and Cambridge University Press) pp 57–62

Gould A and Salim S 2001 Luyten High Proper Motion Stars in 2MASS and SDSS *Bull. Am. Astron. Soc.* **33** 1494

Gulyaev A and Ashimbaeva N 1997 ACVS: an astrometric catalog of variable stars *Astron. Rep.* **41** 215–8

Gulyaev A and Nesterov V 1992 *On the 4 Million Star Catalogue* (in Russian) (Moscow: Moscow University Press) pp 1–7

Gunn J, *et al* 1998 The Sloan Digital Sky Survey Photometric Camera *Astron. J.* **116** 3040–81

Hambly N C, *et al* 2001 The SuperCOSMOS Sky Survey—III. Astrometry *Mon. Not. R. Astron. Soc.* **346** 1315–27

Haramundanis K 1966 Experience of the Smithsonian Astrophysical Observatory in the construction and use of star catalogues *Astron. J.* **72** 588–96

Hartkopf W, Mason B and Worley C 2001a The 2001 US Naval Observatory Double Star CD-ROM. II. The Fifth Catalogue of Orbits of Visual Binary Stars *Astron. J.* **122** 3472–9

Hartkopf W, McAlister H and Mason B 2001b The 2001 US Naval Observatory Double Star CD-ROM. III. The Third Catalogue of Interferometric Measurements of Binary Stars *Astron. J.* **122** 3480–1

Henry T J, *et al* 2001 RECONS discoveries of new nearby stars *Bull. Am. Astron. Soc.* **33** 1495

Hinz J L, *et al* 2002 A near-infrared, wide-field, proper-motion search for brown dwarfs *Astron. J.* **123** 2027–32

Hoffleit D 1992 Astronomy at Yale 1701–1968 *Mem. Connecticut Acad. Arts Sci.* **32** 115

Høg E *et al* 2000 The Tycho-2 catalogue of the 2.5 million brightest stars *Astron. Astrophys.* **355** L27–30

Honma M, Kawaguchi N and Sasao T 2000 Science with VERA: VLBI exploration of radio astrometry *Proc. SPIE* **4015** 624–31

Horch E, Franz O and Ninkov Z 2000 CCD speckle observations of binary stars from the southern hemisphere. II. Measures from the Lowell–Tololo Telescope during 1999 *Astron. J.* **120** 2638–48

Horch E, Ninkov Z and Franz O 2001a CCD speckle observations of binary stars from the southern hemisphere. III. Differential photometry *Astron. J.* **121** 1583–96

Horch E, *et al* 2001b Speckle interferometry of southern stars. II. Measures from the CASLEO 2.15 meter telescope, 1995 1996 *Astron. J.* **121** 1597–606

—— 2001c RYTSI: a new way to do speckle imaging *Bull. Am. Astron. Soc.* **33** 790–1

Hui H and Rui W 2002 Optical positions of 55 radio stars from astrolabe observations from the Yunnan Observatory *Astron. Astrophys.* **383** 1062–6

Jaschek C 1989 *Data in Astronomy* (Cambridge: Cambridge University Press) pp 52–4

Jao W *et al* 2001 The RECONS effort to characterize the highest proper motion stars *Bull. Am. Astron. Soc.* **33** 1494

Jenkins L 1952 *General Catalogue of Trigonometric Stellar Parallaxes* (New Haven, CT: Yale University Observatory)

Jones B, Hanson R and Klemola A 2000 Lick Northern Proper Motion Program: NPM2 *Bull. Am. Astron. Soc.* **32** 759

Kaplan I, Koerner D and Henry T 2001 An astrometric survey of companions to nearby stars *Bull. Am. Astron. Soc.* **33** 1495

Klemola A, Jones B and Hanson R 1987 Lick Northern Proper Motions Program. I. Goals, organizations, and methods *Astron. J.* **94** 501–15

Kovalchuk A, *et al* 1997 First steps to re-observation of the HIPPARCOS/TYCHO star by ground-based automatic AMCs *ESA SP* **402** pp 139–41

Lampens P *et al* 2001 CCD photometry and astrometry for visual double and multiple stars of the HIPPARCOS catalogue III. CCD photometry and differential astrometry for 253 southern 'intermediate' systems *Astron. Astrophys.* **374** 132–50

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Lasker B, *et al* 1990 The Guide Star Catalogue. I. Astronomical Foundation and Image Processing *Astron. J.* **99** 2019–58

Levine S and Monet D 2000 The USNO PMM Integrated Image and Catalogue Arch. *Bull. Am. Astron. Soc.* **32** 1596

Longo G, *et al* 2001 Advanced data mining tools for exploring large astronomical databases *Proc. SPIE* **4477** 61–75

López C and Girard T 1990 Improved positions of variable and suspected variable stars south of –67 degrees *Publ. Astron. Soc Pac.* **102** 1018–38

Luo D and Li D 1986 Photoelectric astrolabe and astrolabe star catalogues *IAU Symposium 109, Astrometric Techniques* ed H Eichhorn and R Leacock (Dordrecht: Reidel) pp 375–7

Luyten W 1963 Proper motion surveys *Basic Astronomical Data* ed K. Aa. Strand (Chicago, IL: Chicago University Press) pp 46–54

—— 1970 The Bruce and Palomar Schmidt Proper Morions Survey *IAU Colloquium 7, Proper Motions* ed W Luyten (Minneapolis, MN: University of Minnesota) pp 59–63

Ma C, *et al* 1998 The international celestial reference frame as realized by very long baseline interferometry *Astron. J.* **116** 516–46

Manrique W, *et al* 1999 Second San Juan Photoelectric Astrolabe Catalogue *Astron. Astrophys. Suppl. Ser.* **136** 1–5

Martin V, *et al* 1999 Absolute declinations with the photoelectric astrolabe at Calern Observatory (OCA) *Astron. Astrophys. Suppl. Ser.* **137** 269–72

Masden S, Dravins D and Lindegren L 2002 Astrometric radial velocities. III. Hipparcos measurements of nearby star clusters and associations *Astron. Astrophys.* **381** 446–63

Mason B, *et al* 1998 ICCD Speckle observations of binary stars. XIX. An astrometric/spectroscopic survey of O stars *Astron. J.* **115** 821–47

Mason B, *et al* 2000 Double stars in the Tycho-2 Catalogue *Astron. J.* **120** 3244–9

Mason B, *et al* 2001 The 2001 US Naval Observatory Double Star CD-ROM. I. The Washington Double Star Catalogue *Astron. J.* **122** 3466–71

McAlister H 2001 Science with the CHARA array *Bull. Am. Astron. Soc.* **33** 859

Momany Y, *et al* 2001 ESO imaging survey Pre-FLAMES survey: observations of selected stellar fields *Astron. Astrophys.* **379** 436–52

Monet D 1981 Precision astrometry and the Kitt Peak National Observatory KPNO 4-m charge-coupled device CCD parallax program *Proc. SPIE* **290** 208–14

Monet D G 2000 Progress report on the USNO-B Catalogue *Bull. Am. Astron. Soc.* **32** 1596

Monet D and Dahn C 1983 CCD astrometry. I—preliminary results from the KPNO 4-m CCD parallax program *Astron. J.* **88** 1489–507

Monet, D G *et al* 1998 *USNO-A2.0* (Washington, DC: US Naval Observatory)

Morgan H R 1952 Catalogue of 5268 standard stars, based on the normal system N30. *Astron. Pap. Am. Ephemer. and Naut. Alm.* **13** (3) 190–231

Muiños J L 2002 Private communication

Muiños J L and Mallamaci C C 2002 Private communication

Paturel G, *et al* 2000 An image database. III. Automatic extraction for millions of galaxies *Astron. Astrophys.* **146** 19–29

Perryman M A, *et al* 2001 GAIA: composition, formation and evolution of the Galaxy *Astron. Astrophys.* **369** 339–63

Platais I, *et al* 1998 The Southern Proper Motion Program. II. A catalog at the south galactic pole *Astron. J.* **116** 2556–64

Platais I, Kozhurina-Platais V and Girard T 2001a WIYN Open Cluster Study: astrometric membership of the a Car open cluster *Bull. Am. Astron. Soc.* **33** 1387

Platais I, *et al* 2001b Preliminary SPM catalog of positions in the southern hemisphere *Bull. Am. Astron. Soc.* **33** 1189

—— 2001c WIYN Open Cluster Study. VII. NGC 2451A and the Hipparcos distance scale *Astron. J.* **122** 1486–99

Pinigin G, *et al* 1995 The axial meridian circle of the Nikolaev Astronomical Observatory *Astron. Astrophys. Trans. 8* 161–3.

Prieur J.-L, *et al* 1998 The 'Pisco' speckle camera at Pic du Midi Observatory *Exp. Astron.* **8** 297–315

—— 2001 Speckle observations of binary systems measured by Hipparcos *Astron. Astrophys.* **367** 865–75

Rapaport M, *et al* 2001 M2000: an astrometric catalog in the Bordeaux Carte du Ciel zone +11degr <= delta <= +18degr *Astron. Astrophys.* **376** 325–32

Robinson S, *et al* 2001 Four years of speckle observing at the WIYN observatory *Bull. Am. Astron. Soc.* **33** 849

Roeser S and Bastian U 1993 The final PPM Star Catalogue for both hemispheres *Bull. Inform.* CDS 421–16

Roeser S, Bastian U and Kuzmin A 1994 PPM Star Catalogue: the 90,000 Stars Supplement *Astron. Astrophys. Suppl. Ser.* **105** 301–3

Sanner J and Geffert M 2001 The IMF of open star clusters with Tycho-2 *Astron. Astrophys.* **370** 87–99

Scardia M, *et al* 2000 Speckle observations of double stars with PISCO at Pic du Midi: measurements in 1998 *Astron. Astrophys.* **131** 561–9

Schlesinger F 1910 Photographic determinations of stellar parallax made with the Yerkes Refractor *Astrophys. J.* **32** 372–87

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

—— 1924 *General Catalogue of Stellar Parallaxes* (New Haven, CT: Yale University Observatory)
Schlesinger F and Jenkins L 1935 *General Catalogue of Stellar Parallaxes*, 2nd ed (New Haven, CT: Yale University Observatory)
Schlesinger F, *et al* 1926 Catalogue of 5833 stars –2° to +1° *Trans. Yale Univ. Obs.* **5**
Scholz R-D, Meusinger H and Jahreiss H 2001 Search for nearby stars among proper motion stars selected by optical-to-infrared photometry. I. Discovery of LHS 2090 at spectroscopy distance of $d \sim 6$ pc. *Astron. Astrophys.* **374** L12–5
Scholz R-D, *et al* 2002 New nearby stars among bright APM high proper motion stars *Mon. Not. R. Astron. Soc.* **329** 109–14
Schultz A B, *et al* 1998 A possible companion to Proxima Centauri *Astron. J.* **115** 345
Song I, *et al* 2001 Searching for young nearby stars *Bull. Am. Astron. Soc.* **33** 914
Stone R C 2000 Accurate FASTT positions and magnitudes of asteroids: 1997–1999 observations *Astron. J.* **120** 2708–20
Sturmann L, *et al* 2001 Testing the telescopes of the CHARA array *Bull. Am. Astron. Soc.* **33** 876
Tang Z, *et al* 2000 Determination of optical positions for extragalactic radio sources under the collaboration between SHAO and NAO *Proc. IAU Colloquium 180* (Washington, DC) pp 57–60
Teixeira R, *et al* 2000 Proper motions of pre-main sequence stars in southern star-forming regions *Astron. Astrophys.* **361** 1143–51
ten Brummelaar T *et al* 2001 Commissioning observations for the CHARA Array *Bull. Am. Astron. Soc.* **33** 877
Torra J, *et al* 2001 The GAIA mission: a stereoscopic and kinematic census of our Galaxy *ESA SP-459* pp 645–8
Turner N H, *et al* 2001 The CHARA array control system *Bull. Am. Astron. Soc.* **33** 876
Urban S, Corbin T and Wycoff G 1998 The ACT Reference Catalogue *Astron. J.* **115** 2161–66
Urban S, *et al* 2001 *Bull. Am. Astron. Soc.* **33** 1494
van Altena W 2002 Private communication
van Altena W, Lee J and Hoffleit D 1995 *The General Catalogue of Trogonometric Parallaxes* 4th ed (New Haven, CT: Yale University Observatory)
Viateau B, *et al* 1999 The Bordeaux and Valinhos CCD meridian circles *Astron. Astrophys. Suppl. Ser.* **134** 173–186
Warren W 1992 Star catalogs and surveys *The Astronomy and Astrophysics Encyclopedia* ed S P Maran (Cambridge: Cambridge University Press) pp 655–61
Winter L and Holdenried E 2001 Measuring the AGK2 Survey with the Modified StarScan measuring machine at the USNO *Bull. Am. Astron. Soc.* **32** 1596
Worley C, Mason B and Wycoff G 2001 The 2001 US Naval Observatory Double Star CD-ROM. IV. The Photometric Magnitude Difference Catalogue *Astron. J.* **122** 3482–4
Zacharias N, *et al* 2000a Comparing Tycho-2 astrometry with UCAC1 *Astron. J.* **120** 1148–52
—— 2000b The First US Naval Observatory CCD Astrograph Catalogue *Astron. J.* **120** 2131–47
——2001 UCAC2: a new high precision catalog of positions and proper motions *Bull. Am. Astron. Soc.* **33** 1495
Zhu Z 2001 Hipparcos proper-motion system with respect to SPM 2.0 *Publ. Astron. Soc. Jpn.* **53** L33–6

See the appendix for a list of the most common acronyms used in the astrometry related projects

http://www.eso.org/eis/
http://www.astro.yale.edu/spm/spm2cat/spm2.html
http://obswww.unige.ch/webda/
http://ad.usno.navy.mil/wds/
http://www.chara.gsu.edu/CHARA/index.html
http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/384/180
http://mypage.goplay.com/noam/amc1b.zip
http://veraserver.mtk.nao.ac.jp/index.html
http://aladin.u-strasbg.fr
http://www.stecf.org/astrovirtel/
http://www.astrogrid.ac.uk/
http://www.roe.ac.uk/cosmos/scosmos.html
http://www.eso.org/projects/avo/intro.html
http://www.srl.caltech.edu/nvo/
http://vizier.u-strasbg.fr/
http://simbad.u-strasbg.fr/Simbad
http://stdatu.stsci.edu/dss/
http://www.aerith.net/misao/index.html
http:// www.mmo.org
http://www.nofs.navy.mil/data/fchpix/
http://skys.gsfc.nasa.gov/skymorph/skymorph.html
http://www.ll.mit.edu/LINEAR/
http://neat.jpl.nasa.gov/
http://www.lowell.edu/users/elgb/loneos_disc.html

According to J McDowell 2002 (*Sky Teles.* **103** (5), 28), NASA managers have recently cancelled the FAME mission.

Carlos E. López

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Astrometry

The branch of astronomy dealing with the precise measurement of the position of astronomical bodies. Many problems of astronomy require accurate observations of stellar positions. Positions of the brighter stars can be measured very accurately in the equatorial system. Fainter stars are measured using photographic plates or some other imaging device (e.g. a charge-coupled device, or CCD) with respect to the brighter stars. The entire stellar group can then be referred to the positions of known external galaxies, which are far enough away to define an essentially fixed system. Over relatively short periods of time both the bright and faint stars are affected by galactic rotation and by their own motions through the Galaxy. The most precise stellar positions to date have been obtained by the Hippacos satellite.

See also: astronomy, astrophysics, Hipparcos, universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrometry with Hubble Space Telescope

In 1990 NASA launched the HUBBLE SPACE TELESCOPE. In addition to cameras and spectrographs usable from the far ultraviolet to the near-infrared, the observatory contains three white-light INTERFEROMETERS. As part of engineering and science support their primary task was telescope guiding; to position and hold science targets within the science instrument apertures with tolerances approaching 0.1″, equivalent to 100 milliseconds of arc (100 mas). Pointing and tracking requires two such fine guidance sensors (FGS), leaving the third free for ASTROMETRY.

The design goal for astrometry with an FGS was 3 mas precision over the entire field of regard. When designed and built in the early 1980s, The FGS represented an order of magnitude improvement over existing ground-based techniques. HST launch delays provided sufficient time for ground-based techniques to equal and surpass this goal. Thus, the post-launch goal was redefined. This goal, 1 mas precision small-field astrometry, has been achieved, but not without significant challenges.

An HST FGS will remain a competitive astrometric tool for faint targets in crowded fields and for faint small-separation binaries, until the advent of large-aperture ground-based and longer-baseline space-based interferometers (see also ASTROMETRY: TELESCOPES AND TECHNIQUES).

This article describes the science that can be done with the FGS and presents some recent science results. We outline how these data are modeled, acquired and calibrated. We next show how the astrometer, FGS 3, works as an interferometer. Finally, we present a guide to the literature that provides additional detail for each item discussed in this article.

Astrometry with 1 mas precision is obtainable with the HST wide field planetary camera. The techniques are similar to those used in ground-based CCD astrometry and should be discussed separately.

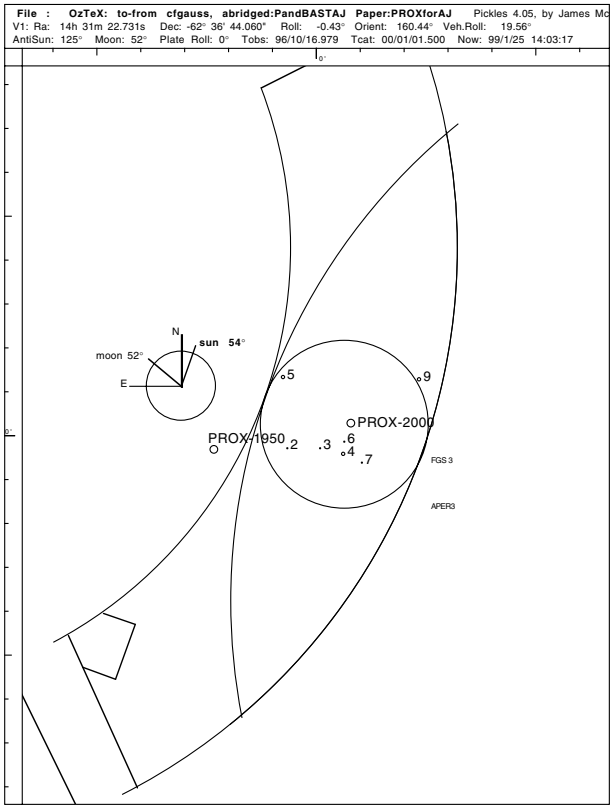


Figure 1. Field of regard of FGS 3 on the sky at the location of the planet search target, Proxima Centauri, on 16 October 1996. The position of Proxima Centauri at two epochs is plotted. Also shown are the reference stars, relative to which parallax and proper motion are obtained. Tick marks are separated by 1 arcminute.

The choice of science targets

The choice of HST astrometry science targets is rightly determined by what can be done from the ground. The unique capabilities of HST must remain reserved for projects demanding them. In this section we summarize the strengths of HST astrometry, pointing out a few examples of the kinds of objects for which HST and the FGS are ideally suited. Since the FGS interferometer offers two observational modes, fringe scanning and fringe tracking, we divide the remainder of this article by mode at each step.

Fringe scanning: using HST for binary star astrometry

Four characteristics of a high-resolution (<1″) observing technique must be considered when attempting to resolve a target BINARY STAR: the source brightness limit, the resolving capability, the astrometric accuracy and the magnitude difference between the components that can be observed. Although all four parameters are interdependent, we provide current limits for each characteristic separately in order to illustrate the efficacy of HST FGS 3 as an astrometric instrument.

For binary stars HST FGS 3 observations provide results currently unachievable with ground-based techniques. HST FGS 3 is accurate to 1–3 mas at a resolution limit of 15 mas for magnitude differences less than 2 mag. It can effectively observe targets to at least $V \sim 15$. At component separations greater than 200 mas, it can bridge brightness differences of at least $\Delta V = 4$ mag, as demonstrated by the detection of Gliese 623B.

Future developments in ground-based interferometers (see INTERFEROMETRY: GROUND) should improve the accuracy of astrometric measurements and provide better resolution, but will still be limited to relatively bright targets and moderate magnitude differences. HST FGS 3 is the only high-resolution instrument currently available that (a) can provide high-precision astrometry for very close binaries, (b) allows relatively faint targets to be observed, and (c) can bridge at least moderately large magnitude differences between the components in a binary system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fringe tracking: using HST to obtain parallaxes

In the absence of distance, astrophysics lacks precision. Our demonstrated measurement precision for PARALLAXES is 0.5 mas or better, given six epochs of observation at maximum parallax factor. Often a parallax target is unmeasurable because reference stars are not located within the field of view of the measuring device; this is especially true for small-format (typically several square arcmin) CCD cameras. An FGS provides a large field of regard (figure 1), shaped somewhat like a pickle. This shape comes from the the pick-off mirror (figure 5) in the HST focal plane. Parallax observations are generally spaced by six months, during which time, due to HST solar array illumination constraints, the FGS field of view rotates by 180°. The lens-shaped region (figure 6) in common to the two extreme orientations provides a relatively large field with short axis = 3.5′ and long axis = 14′.

FGS 3 has a large dynamic range, able to obtain fringe tracking position measures for stars in the magnitude range $4 \leq V \leq 17$. This large dynamic range is provided by a neutral density filter that reduces the magnitude of bright stars by 5 mag. The unfiltered range is $8.5 \leq V \leq 17$.

HST can obtain precise parallaxes for binaries whose components are separated $0.03'' < \rho < 1.0''$. Nearly simultaneous fringe scanning and tracking measurements provide the component positions relative to reference stars, essential in determining the center of mass of a system, and thus individual component masses.

Lastly there is timeliness of result. One no longer need wait three to six years for the parallax of astrophysically interesting objects. The distance to a sufficiently interesting and important object can be obtained on the same time scale as other astrophysical information. If an object or class of objects is interesting now, a theory can be tested now.

Science results

Science targets have included nearby stars used as probes for an extrasolar planet search; stars whose astrophysics would be greatly aided by accurate distances; low-mass binary stars to define the mass–luminosity relation for the lower main sequence; members of the Hyades cluster, key to the distance scale; extragalactic objects (QSO) to provide an inertial reference frame for Hipparcos; and the epitome of crowded fields, an extragalactic cluster, R136 in the Large Magellanic Cloud.

Once again we divide by technique, discussing representative results from first fringe scanning then fringe tracking.

Fringe scanning and binary stars

Fringe scanning science depends on deconvolution and the assumption that all objects in the field are point sources. For a recent science result we turn to the low-mass binary Wolf 1062 = Gl 748, observed in support of a lower main sequence mass–luminosity project. Figure 2 presents fringe scans of this binary along the two orthogonal axes of FGS 3. Since each observed fringe is a linear superposition of two fringes (one for each component star in the binary), it is modeled with two identical single star fringes. Their placement along the axis and relative amplitudes provide separation and brightness differences. Relative separation along the two axes provides position angle information. Figure 3 shows all the measured separations and position angles on the best-fit orbit. The mean absolute difference between the observed and computed separations is only 1.1 mas, and only 0.77° in the position angle. Residual vectors for all data points are drawn, but are smaller than the points. Perhaps it is even more impressive to realize that the box illustrated is only 0.6″ in size, so that good seeing from the ground would result in a stellar image the size of the entire figure.

A combination of fringe scanning and fringe tracking observations for the low-mass binary L722-22, has yielded a relative parallax $\pi = 0.1656'' \pm 0.0008''$ and component masses of 0.179 and 0.112 $M_{\odot}$ with formal random errors for the mass as low as 1.5%.

Fringe tracking: positions in a reference frame

Fringe tracking science is primarily that of relating the position of a target to positions of stars defining a reference frame. Here we discuss a Hyades parallax program and a fundamental astrometry project involving HIPPARCOS.

Trigonometric parallax observations were obtained for seven Hyades members in six fields of view. These have been analysed along with their proper motions to determine the distance to the cluster. Formal uncertainties on individual parallaxes average 1 mas. This relatively large error is due both to poor spatial distribution and the small number of reference stars in each field. Knowledge of the convergent point and mean proper motion of the Hyades is critical to the derivation of the distance to the center of the cluster. Depending on the choice of the proper-motion system, the derived cluster center distance varies by 9%. Therefore, a full utilization of the HST FGS parallaxes awaits the establishment of an accurate and consistent proper-motion system.

Observations of separations of Hipparcos stars from extragalactic objects have been made to determine the rotation of the Hipparcos instrumental system with respect to the ICRS-VLBI reference frame. A determination from 78 observations yields accuracies of the order of 2 mas rms in the coordinate rotational offsets near the mean epoch of the HST observations and 2 mas yr^{-1} in the coordinate rotation rates. The main contributing sources of error are HST measurement errors and proper motion errors introduced by the three year time difference between the mean Hipparcos and HST observational epochs.

Data modeling

Our two examples are a binary star for fringe scanning and a parallax field for fringe tracking.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

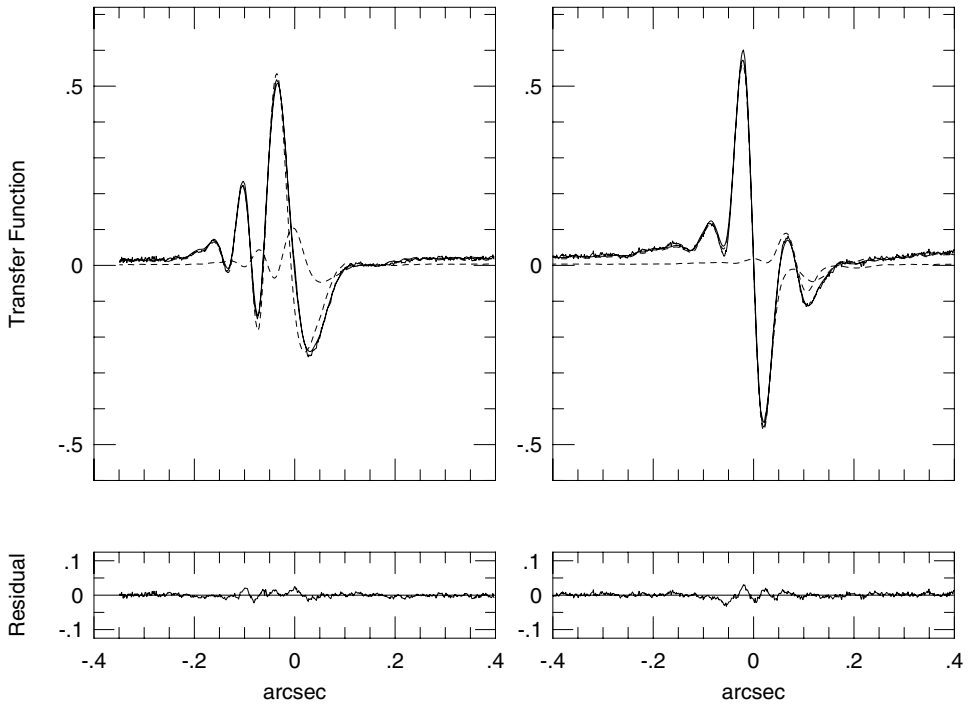


Figure 2. Analysis of the 1995.5712 observation of Wolf 1062. The computed (smoothed full) curves in X (left) and Y (right) represent best-fitting linear superpositions of two single-star fringes (dashed curves). The relative displacements of the single-star curves are $0.0329'' \pm 0.0003''$ and $0.0862'' \pm 0.0002''$ along FGS X and Y respectively. They yield separation and position angle, while the relative amplitudes of the single-star curves yield a magnitude difference for the binary components. The systematic trends in the curve-fitting residuals (bottom panels) are due to imperfect single-star calibrations. Their effect upon the astrometry and photometry of the Wolf 1062 components is negligible.

Fringe scanning

We consider a binary star as the simplest object requiring fringe scanning.

We assume that the fringe produced by a binary star is a linear superposition of the fringes produced by two single stars. The interferometer response to an actual binary star is shown in figure 2. If $F(x)$ is the fringe produced by a single star on the x-axis, a binary star should be described by

$$D(x) = A \times F(x + z_x) + B \times F(x + z_x + S_x) \qquad (1)$$

where A and B are the relative intensities of the two components (constraining $A + B = 1$). The zero point offset, z_x, and a component separation, S_x, complete the model, there being a similar expression for the fringe produced along the y-axis. S_x and S_y yield the binary separation, and, once transformed to equatorial coordinates, the position angle. We obtain the magnitude difference from

$$\Delta m = -2.5 \times \log(A/B). \qquad (2)$$

If a binary is observed in fringe tracking mode, the FGS will lock on an erroneous zero-crossing position which is generated by two closely overlapping s-curves. A fringe scanning mode observation gives relative positions of the two components. One can determine the relative positions with an accuracy and precision of 1 mas, once the fringes from the two stars are deconvolved.

Fringe tracking

We consider a typical parallax target and associated reference stars as an example of fringe tracking. The primary science target is ideally surrounded by five to ten other stars used as a reference frame relative to which we determine position and motion.

Fringe tracking astrometry is generally a two-step process. We first determine the characteristics of the reference frame. One of the epochs of observation is chosen as the constraint plate. From these data we determine the scale and rotation relative to the constraint plate for each observation set within a single orbit. Since for many of our targets the observation sets span over two years, we also include the effects of reference star parallax (π) and proper motion (μ)

$$\xi = ax + by + c - (P_x * \pi + \mu_x * t) \qquad (3)$$

$$\eta = dx + ey + f - (P_y * \pi + \mu_y * t). \qquad (4)$$

The orientation relative to the celestial sphere is obtained from ground-based astrometry. Uncertainties in the field orientations are generally $0.03^\circ < \epsilon_\theta < 0.09^\circ$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

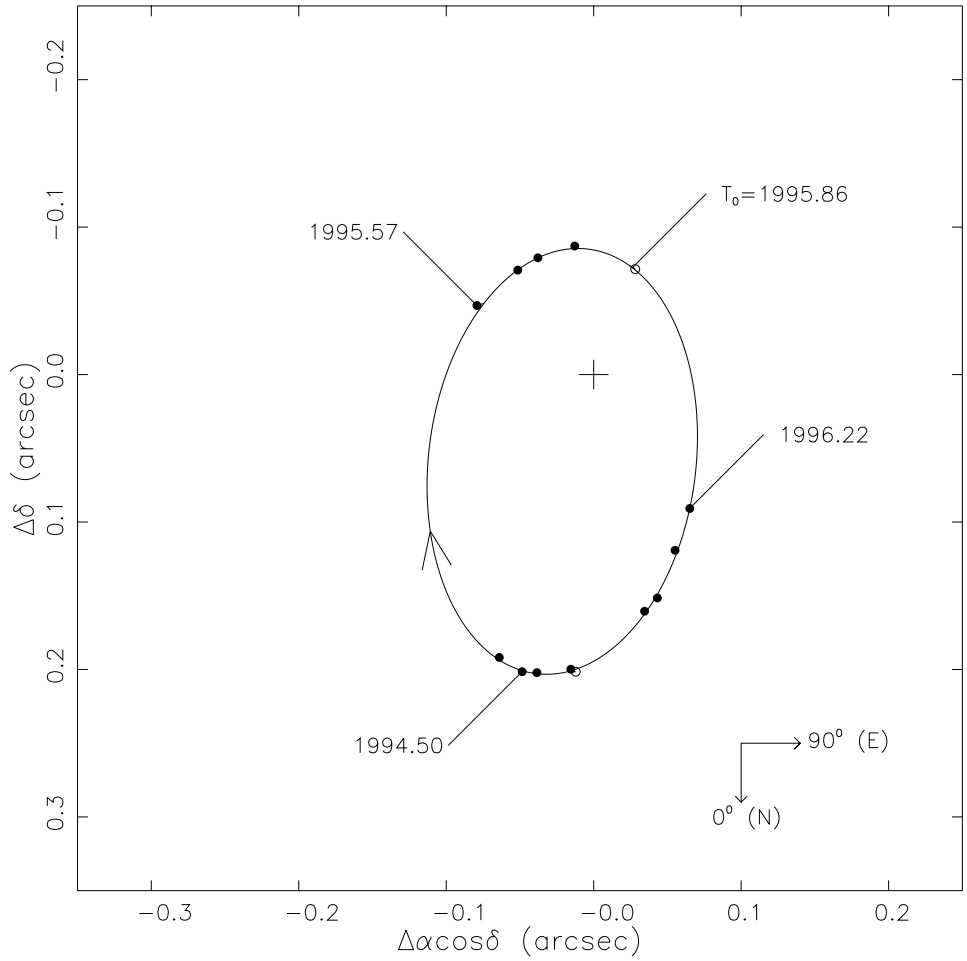


Figure 3. HST FGS3 measurements of Wolf 1062AB and the orbital path predicted by the derived elements are shown. Three epochs of observation and the time of periastron are labelled. Residual vectors for all data points are plotted, but are smaller than the points themselves.

We obtain the parallax factors P_x and P_y from a JPL Earth orbit predictor. Finally, for a rich-enough reference field ($n >$ four stars) we constrain $\Sigma\mu = 0$ and $\Sigma\pi = 0$ for the entire reference frame.

The second step consists of applying the plate constants to the measurements of the science target. Plate coefficients $a, b, ..., f$ are applied as constants, while we solve for the science target π and μ in the above equations.

The astrometer

The design

FGS 3 is an interferometer. Interference takes place in a prism that has been sliced in half, had a quarter-wave retarding coating applied, and then reassembled. Figure 4 shows one of these Koester prisms. Most of the FGS consists of supporting optics used to feed the Koester prisms (figure 5). In particular the star selectors walk the 5″ instantaneous field of view throughout the interferometer field of regard shown in figure 1. The output of each face (face A and face B) is measured by a PMT. These signals are combined

$$S = \frac{A - B}{A + B} \tag{5}$$

to form a signal, S, that is zero for waves exactly vertically incident on the Koester prism's front face. Tilting the wavefront back and forth (equivalent to pointing the telescope slightly off, then on target, then slightly off to the other side) generates the fringe pattern seen in figure 2. A perfect instrument would generate a perfectly symmetric fringe pattern. The significant spherical aberration of the as-built HST primary mirror, in the presence of internal FGS misalignments, produces a signature in the fringe which mimics coma. Coma causes decreased modulation and multiple peaks and valleys in a fringe. FGS 3 produces the least complicated fringes over the widest area within the field of regard.

A replacement FGS installed in 1997 contains an articulated fold-flat mirror that removes most of the internal misalignments. This FGS (FGS 1r) produces

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

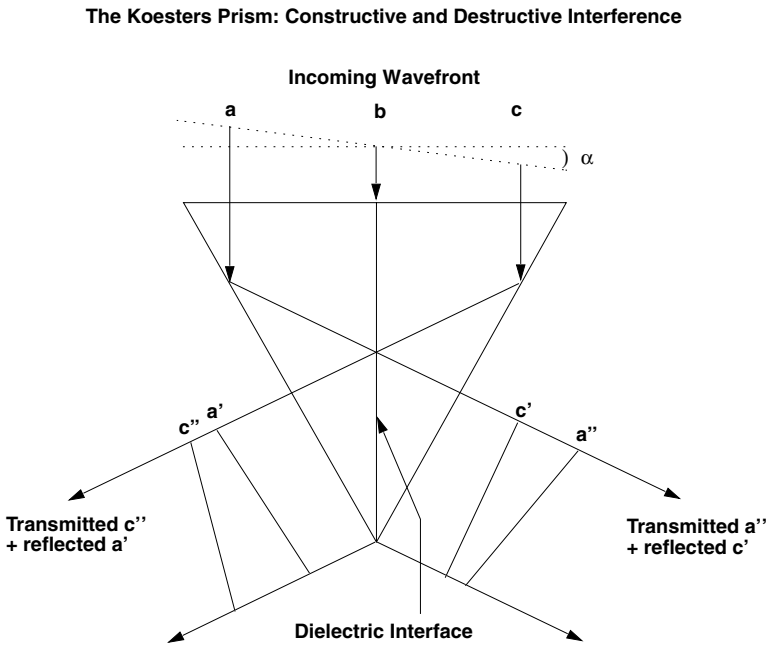


Figure 4. The heart of the FGS interferometer, the Koester prism.

nearly perfect fringes and should at a minimum yield superior binary star results. FGS 3 will remain in service until FGS 1r is fully calibrated in 2000.

FGS operating modes

We discuss strategies for obtaining the highest quality data possible. Our goal, 1 mas precision small-field astrometry, has been achieved, but not without significant challenges. These included a mechanically noisy on-orbit environment, the self-calibration of FGS 3, and significant temporal changes in our instrument. Solutions include a denser set of drift check stars for each science observation, fine-tuning exposure times, overlapping field observations and analyses for calibration, and a continuing series of trend-monitoring observations.

The single greatest contributor to data quality is to treat all data aquired in the same orbit as a unit observation, e.g. a 'plate'.

Fringe scanning

The target is placed in the center of the 'pickle' and the star selectors are commanded to move the instantaneous field of view across the target star image. This action produces a fringe. In practice 10–30 scans are obtained in a reciprocating pattern forwards and back across the star. The final fringe results from a reversal, shift and add process. We have evidence that over the span of an orbit the positions reported by FGS 1 and FGS 2 for the guide stars change. This results in a drift-like motion of FGS 3, the astrometer. Drift can exceed 30 mas over the span of 36 min. The x and y drift rates are generally dissimilar and the drift is not of constant rate. However, since a single scan across a science target requires about 1 min, the drift per scan is reduced to less than 1 mas. The reciprocating data acquisition strategy is very nearly self-compensating for drift.

Fringe tracking

For fringe tracking, onboard electronics locate the zero crossing between the highest positive and lowest negative fringe peak (see figure 2). The position of this zero crossing is determined at a 40 Hz rate during an observation time ranging between $10 < t < 300$ s. The median of >2400 zero crossings provides a robust position estimate. The star selectors are used to move the instantaneous field of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

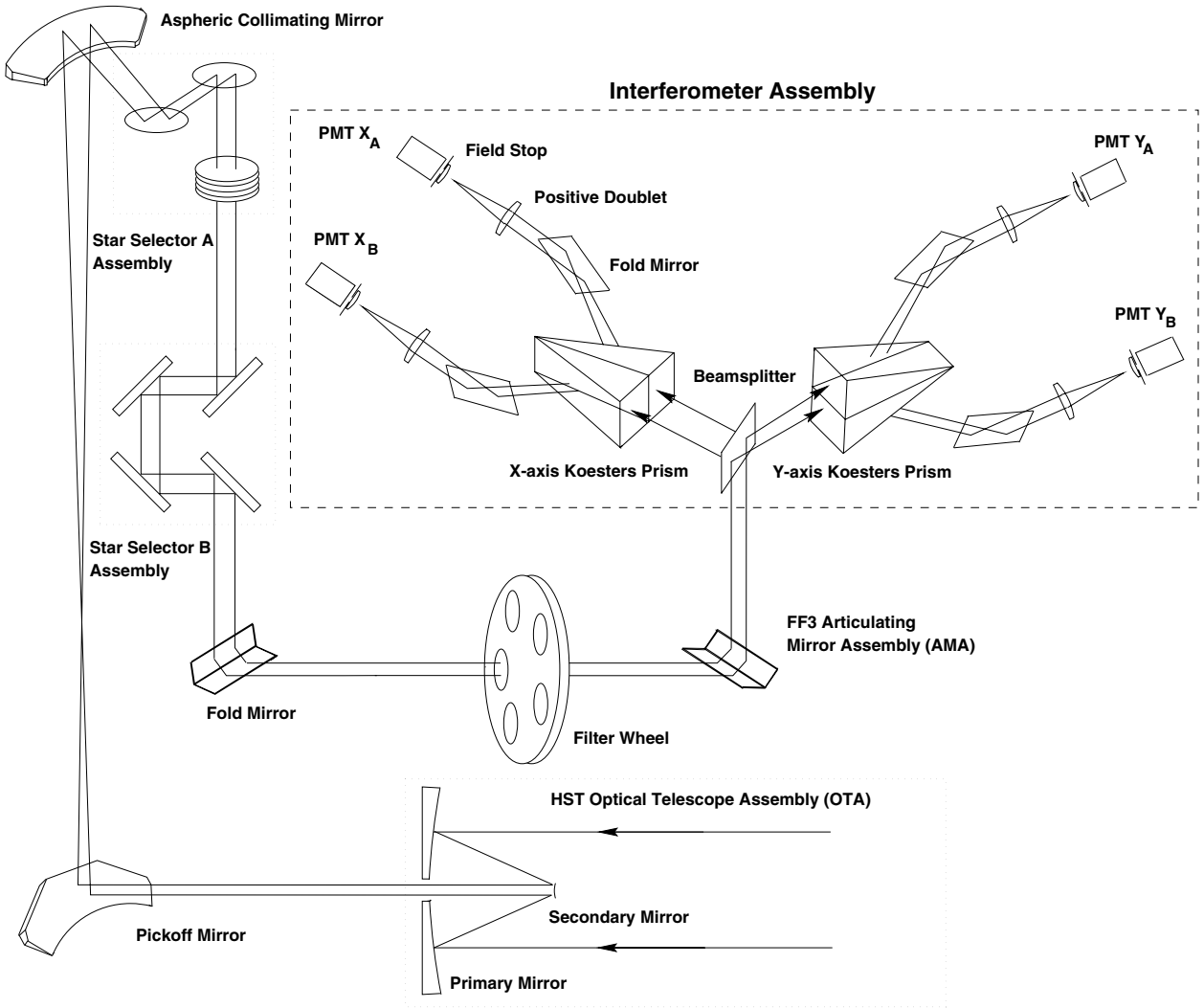


Figure 5. The optical layout of a fine guidance sensor.

view from one star to the next in the FGS field of regard shown in figure 1.

Drift is correctable, but imposes additional overhead, reducing the time available within an orbit to observe the science target. An astrometric observation set must contain visits to one or more reference stars, multiple times during each observation sequence. Presuming no motion intrinsic to all reference and science target stars over a span of 40 min, one determines drift and corrects the reference frame and target star for this drift. As a result we reduce the error budget contribution from drift to less than 1 mas. In figure 1, the science target might be observed three times during an observing session (single orbit) and each reference star twice.

Astrometric calibration

The calibration of the fringe scanning mode has as its goal the determination of separation and magnitude differences between components of binary stars. As discussed earlier, all fringe scanning targets are observed near the pickle center. In contrast fringe tracking takes place over the entire field of regard. Calibration requirements for these two modes of operation are quite different.

Fringe scanning calibration

The most fundamental calibration for any astrometric instrument is to determine the scale in units of the celestial sphere. The FGS scale comes from measurements of several calibration binary stars frequently observed by speckle interferometry with 4 m class telescopes. On the basis of these speckle measures, extremely well-determined orbital elements have been obtained. These binary star orbits yield, for any date of HST FGS observation, accurate angular separations of the binary, and hence the scale.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

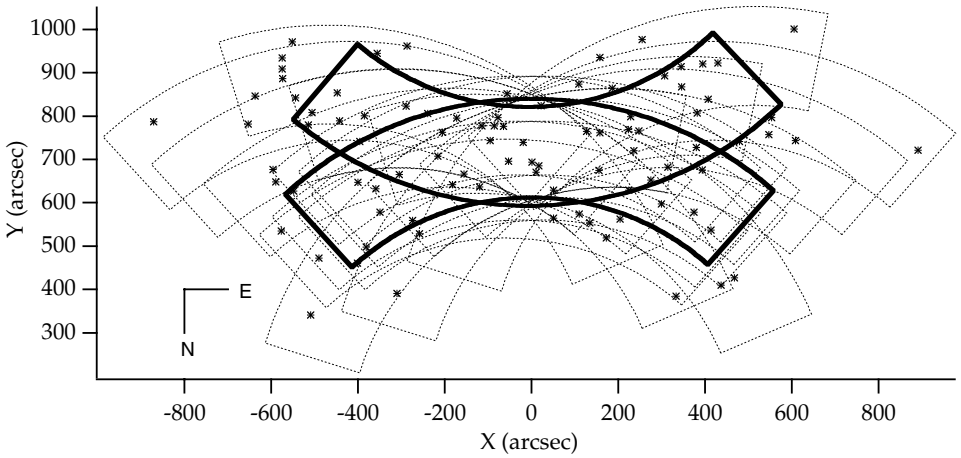


Figure 6. The fringe tracking self-calibration of FGS 3. Stars in the cluster M35 were observed at 19 different pointings and rolls of HST. The pickle-shaped regions denote the total FGS field of regard. The black dots are stars within the cluster M35. The lens-shaped area in the middle shows the field of view for parallax studies.

Secondary calibration consists of the development of a library of suitable single star fringe templates. This library is required since fringe morphology weakly depends on star color. From a library of templates the required number (one for each component of a binary or multiple star system) are chosen and fitted to the binary star fringe using the least-squares algorithm, GaussFit, that allows for errors in both independent and dependent variables.

Fringe tracking calibration

The optical telescope assembly (OTA) of the HST is an aplanatic CASSEGRAIN TELESCOPE of Ritchey–Chrètien design. Our initial problem included mapping and removing optical distortions whose effect on measured positions exceeded 0.5″. There was no existing star field with cataloged 1 mas precision astrometry, our desired performance goal, to use as a fiducial grid. Our solution was to use FGS 3 to calibrate itself. As a result of this activity distortions are reduced to better than 2 mas over much of the FGS 3 field of regard (figure 1). This model is called the optical field angle distortion (OFAD) calibration.

To describe these distortions we have adopted a pre-launch functional form originally developed by the instrument builder, the Perkin-Elmer Corporation. These distortions can be described (and modeled to the level of one millisecond of arc) by the two-dimensional fifth-order polynomials

$$\begin{aligned} x' = {} & a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{02}y^2 + a_{11}xy + a_{30}x(x^2+y^2) \\ & + a_{21}x(x^2-y^2) + a_{12}y(y^2-x^2) + a_{03}y(y^2+x^2) \\ & + a_{50}x(x^2+y^2)^2 + a_{41}y(y^2+x^2)^2 + a_{32}x(x^4-y^4) \\ & + a_{23}y(y^4-x^4) + a_{14}x(x^2-y^2)^2 + a_{05}y(y^2-x^2)^2 \\ y' = {} & b_{00} + b_{10}x + b_{01}y + b_{20}x^2 + b_{02}y^2 + b_{11}xy \\ & + b_{30}x(x^2+y^2) + b_{21}x(x^2-y^2) + b_{12}y((y^2-x^2) \\ & + b_{03}y(y^2+x^2) + b_{50}x(x^2+y^2)^2 + b_{41}y(y^2+x^2)^2 \\ & + b_{32}x((x^4-y^4) + b_{23}y(y^4-x^4) + b_{14}x(x^2-y^2)^2 \\ & + b_{05}y(y^2-x^2)^2 \end{aligned} \tag{6}$$

where x, y are the observed positions within the FGS field of view, x', y' are the corrected positions, and the numerical values of the coefficients a_{ij} and b_{ij} are determined by calibration. Ray-traces were used for the initial estimation of the OFAD. Gravity release, outgassing of graphite-epoxy structures within the FGS, and post-launch adjustment of the HST secondary mirror required that the final determination of the OFAD coefficients a_{ij} and b_{ij} be made by an on-orbit calibration.

M35 was chosen as the calibration field. Since the ground-based positions of our target calibration stars were known only to 23 milliseconds of arc, the positions of the stars were estimated simultaneously with the distortion parameters. This was accomplished during a marathon calibration, executed on 10 January 1993 in FGS 3. The entire 19 orbit sequence is shown in figure 6. GaussFit was used to simultaneously estimate the relative star positions, the pointing and roll of the telescope during each orbit (by quaternions), the magnification of the telescope, the OFAD polynomial coefficients, and four parameters that describe the star selector optics inside the FGS.

Other fringe tracking calibrations

Because each FGS contains refractive elements (Bradley *et al* 1991), it is possible that the position measured for a star could depend on its intrinsic color. Changes in position would depend on star color, but the direction of shift is expected to be constant, relative to the FGS axes. This lateral color shift would be unimportant, as long as target and reference stars have similar color. However, this is certainly not the case for many of the science targets (e.g. very red stars such as Proxima Centauri and Barnard's star).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finally, to provide the large dynamic range, a neutral-density filter can be placed in front of the Koester prism. As a consequence there will be a small but calibratable shift in position (due to the filter wedge) when comparing the positions of the bright star to the faint reference frame. The shift is constant in direction (relative to FGS 3) and size, since the filter does not rotate within its holder.

Maintaining the fringe tracking calibration

The FGS 3 graphite-epoxy optical bench was predicted to outgas for a period of time after the launch of HST. The outgassing was predicted to change the relative positions of optical components on the optical bench. The result of whatever changes were taking place was a change in scale. The amount of scale change was far too large to be due to true magnification changes in the HST optical assembly. Two of the parameters that descibe the star selector optics cause a scale-like change, if they are allowed to vary with time.

The solution was to revisit the M35 calibration field periodically to monitor these scale-like changes and other slowly varying nonlinearities. Revisits will be required as long as it is desirable to do 1 mas precision astrometry with any FGS. The result of this activity judges the validity of the current OFAD coefficients and provides warning for the need for recalibration. With these data we remove the slowly varying component of the OFAD, so that uncorrected distortions remain below 2 mas for center of FGS 3. The character of these changes is generally monotonic with abrupt jumps in conjunction with HST servicing missions.

Bibliography

Here are pointers to a number of key papers in the development of HST interferometric astrometry, and to some recent results.

A detailed description of the FGS interferometer can be found in:

Bradley A, Abramowicz-Reed L, Story D, Benedict G and Jefferys W 1991 The flight hardware and ground system for Hubble Space Telescope astrometry *Publ. Astron. Soc. Pacific* **103** 317

Lupie O and Nelan E P 1998 *FGS Instrument Handbook* version 7 (STScI)

Our choice of astrometer (there being three from which to choose) is documented in:

Benedict G F *et al* 1992 Astrometric performance characteristics of the Hubble Space Telescope fine guidance sensors *Publ. Astron. Soc. Pacific* **104** 958

Our primary calibration tool is GaussFit, a least-squares and robust estimation algorithm described in:

Jefferys W, Fitzpatrick J and McArthur B 1988 *Celest. Mech.* **41** 39

Calibration and observing strategies for fringe scanning are discussed in:

Franz O G *et al* 1991 Binary star observations with the Hubble Space Telescope Fine Guidance Sensors. I - ADS 11300 *Astrophys. J.* **377** L17

Hershey J L 1992 A deconvolution technique for Hubble Space Telescope FGS fringe analysis *Publ. Astron. Soc. Pacific* **104** 592

Introductions to deriving and maintaining the fringe tracking calibration and observing strategies include:

Benedict G F *et al* 1994 Astrometry with Hubble Space Telescope Fine Guidance Sensor number 3: Position-mode stability and precision *Publ. Astron. Soc. Pacific* **106** 327

Benedict G F *et al* 1998 Working with a space-based optical interferometer: HST Fine Guidance Sensor 3 small-field astrometry *Proc. SPIE* **3350** 229

McArthur B, Benedict G F, Jefferys W H and Nelan E 1997 Maintaining the FGS3 Optical Field Angle Distortion Calibration *The 1997 HST Calibration Workshop with a New Generation of Instruments* ed S Casertano, R Jedrzejewski, C D Keyes and Mark Stevens (Baltimore, MD: Space Telescope Science Institute) p 472

Recent astrometric results naturally divide into fringe scanning and fringe tracking. For fringe scanning:

Franz O *et al* 1994 Astrometric Companions Detected at Visible Wavelengths with the Hubble Space Telescope Fine Guidance Sensors *American Astronomical Soc. Meeting* **185** 85.24

discusses the first results for large Δm.

The first definitive binary orbit determined with the Hubble Space Telescope Fine Guidance Sensors was for Wolf 1062 = Gliese 748

Franz O G *et al* 1998 *Astrophys. J.* **116** 1432

From the beginning, stellar interferometers have explored stellar diameters. The FGS has determined interferometric angular diameters of Mira variables:

Lattanzi M G, Munari U, Whitelock P A and Feast M W 1997 *Astrophys. J.* **485** 328

FGS fringe scanning has contributed to the optical mass–luminosity relation at the lower main sequence (0.08 to 0.20 $M_{\odot}$), discussed in:

Henry T J, Franz O G, Wasserman L H, Benedict G F, Shelus P J, Ianna P A, Kirkpatrick J D and McCarthy D W 1999 *Astrophys. J.* **512** 864

Recent fringe tracking results include a report on the search for planets around Proxima Centauri

Benedict G F *et al* 1999 *Astron. J.* **118** 1086

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parallaxes of two high-velocity stars:

Macconnell D J, Osborn W H and Miller R J 1997 *Astron. J.* **114** 1268

And the distance to the Hyades Cluster based on FGS parallaxes:

Van Altena W F *et al* 1997 *Astrophys. J.* **486** L123

An FGS program to link the Hipparcos reference frame to an extragalactic reference system is detailed in:

Hemenway *et al* 1997 *Astron. J.* **114** 2796

Finally a combination of fringe tracking and scanning yielded a parallax and component masses for the low-mass binary L722-22:

Hershey J L and Taff L G 1998 *Astron. J.* **116** 1440

A snapshot, circa 1994, of the scientific capabilities of the FGS can be found in:

Hog E and Seidelmann P K (eds) 1995 *Astronomical and Astrophysical Objectives of sub-milliarcsecond optical astrometry (Proc. 166th Symp. of the International Astronomical Union)* (Drodrecht: Kluwer)

Papers include:

Benedict G F *et al* Hubble Space Telescope: a generator of submilliarcsecond precision parallaxes p 89
Lattanzi M G *et al* Fringe interferometry in space: the fine structure of R136A with the astrometer guidance sensor aboard HST p 95

George F Benedict

, McArthur B, Chappell D W, Nelan E, Jefferys W H, Van Altena W, Lee J, Cornell D, Shelus P J, Hemenway P D, Franz O G, Wasserman L H, Duncombe R L, Story D, Whipple A and Fredrick L W

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrometry: History

Although the term 'astrometry' was not coined until the 19th century, the business of measuring the position and angular separation of the stars and planets constitutes one of the most ancient occupations of the astronomer. By the time of HIPPARCHUS (*fl* 126 BC) in classical Greece, the measurement of astronomical positions with relation to the 360° circle was already established, while the ancient Babylonians of 1750 BC left non-instrumental observations of the western and eastern elongations of Venus with relation to the Sun, which might be correctly considered as astrometric. From Greek times to the present day, astronomers involved in astrometric work have used instruments employing the 360° scale, and one can follow this technical instrumental development in the article DIVIDED INSTRUMENTS.

By its very nature, astrometry entails establishing and refining physical coordinates in the sky, against which the movements of bodies can be quantified. Hipparchus, Claudius PTOLEMY and other Greeks recognized that the ecliptic provided a natural and fairly obvious fundamental astronomical coordinate system, as the mid-summer and mid-winter altitudes of the Sun produced the solstices, equinoxes and zodiac poles. These zodiac coordinates, moreover, could be related to a different set of coordinates, based on the terrestrial equator (or equatorial) and terrestrial poles; in the time of Ptolemy these were 23º 51' 20" apart and are now 23º 26' 21". The geometrical points in the sky where the ecliptic and the equatorial planes intersect, the equinoxes, form a coordinate of natural origin, and the spring intersection in March was used by Ptolemy and other classical astronomers to delineate the beginning of the stellar day. Over two millennia ago, this point resided in the constellation Aries, giving rise to the term 'FIRST POINT OF ARIES', which is still used today as the zero point in celestial longitude (or right ascension) on astronomical charts. Even in antiquity, however, it was recognized that this first point is not fixed (as, indeed, are few points in the sky, given enough time), but moves back upon itself, or 'precesses'. Ptolemy quantified this precession as 1º every century, although he obtained this amount by rounding up the figures for the first point, which he derived from observations covering about 400 yr up to AD 150. Employing greatly superior instruments, and using them to constantly refine and re-define his figure between 1675 and 1715, JOHN FLAMSTEED in England obtained 1º every 72 yr, or 50" yr^{-1} for the precessional constant, which compares very favorably with the 50.290 966" yr^{-1} accepted today.

The establishment of the first point of Aries, and the quantification of its albeit errant motion around the sky, is one of the fundamental achievements of astrometry, and still lies at the heart of stellar positional measurement, requiring charts and tables to be constructed to a particular 'epoch' in the precessional cycle, such as AD 1900 or AD 2000.

The practical measurement of the exact position of this slowly moving piece of empty space in the sky has been a test of ingenuity over the centuries, for inevitably it is occupied by no specific star. In the 2nd century AD Ptolemy in his *Almagest* described a technique employing the new moon as an intermediary. On or around the day of the spring equinox, in such years as there was a convenient new moon at the equinox, Ptolemy measured the angular distance of the setting Sun from the Moon. Once the Sun was below the horizon, he recommended that the position of the Moon be measured from the adjacent stars when they became visible. From these star positions, it was possible to establish the position of the Sun and determine thereby (from a knowledge of the stars in that region of the sky gained some months before) which bright star was closest to the first point of Aries.

Recognizing that the motion of the Moon could itself be a source of error, during the hour or so that the Sun was declining below the horizon, TYCHO BRAHE in Denmark from 1582 repeated Ptolemy's technique using the much slower-moving planet Venus as an intermediary, and computed for that epoch that the first point of Aries stood 26º 0'" 30" removed from the star α Arietis, which was within 15" of the same value calculated for 1585 by Dreyer in the 1890s. The success of Tycho's astrometric work, and the superior accuracies he achieved when compared with the medieval Arabs such as Ulugh Beigh and Al Sufi, and the ancient Greek astronomers who had been observing a millennium earlier, depended on improvements in technique. These included fundamental innovations in instrument design, sophisticated cross-checking procedures and a growing awareness of atmospheric refraction.

The massive strides that astrometry has made over the last four centuries have followed on from an increasing awareness of the factors recognized as important by Tycho: instrument innovation, refined techniques, and an alertness to potential sources of error. However, astrometry's meteoric development since 1580 has only been in response to the intellectual problems posed in astronomy and cosmology, the first of which was the Copernican hypothesis. It was recognized in the 16th century that if the Earth did indeed move in space, then the angular sizes of the constellation figures should display a slight seasonal variation, as the earth-based observer approached or receded from the June and December positions. Tycho himself attempted to detect such seasonal variations. With the development of the concept of the infinite universe in the 17th century, however, this Copernican stellar parallax was seen as most likely to be detected in the six-monthly angular displacement of first-magnitude stars, such as Sirius,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

working on the assumption that, if all stars were of similar intrinsic brightness, those of the first magnitude were the closest to Earth, and hence should display the largest parallaxes.

The development in the 17th and 18th centuries of superior scale graduations on astronomical quadrants, telescopic sights, micrometers and verniers, and of similar instrumental techniques, was strongly driven by the urge of astronomers to detect this elusive stellar parallax. As the astrometric capacity of the age increased from Tycho Brahe's 1' to Hevelius's 25", and then to Flamsteed's 10", and still no parallax was found, astronomers across Europe realized two things: firstly, the three-dimensional stellar universe must be unbelievably vast and, secondly, any hope of successfully fathoming it must reside in an ongoing improvement in the construction of precision instruments, combined with a parallel development in the refined techniques of their usage. This recognition, moreover, has been the creed of the astrometer for the last four centuries from Tycho's Uraniborg to the European Space Agency's Hipparcos space satellite.

The search for a stellar parallax, while not successfully accomplished until 1838, created such an urgency and an inspiration to astronomers that numerous other discoveries and lines of inquiry were born out of it on the way. It was CHRISTIAAN HUYGENS's recognition in the 1650s that graduated scales were insufficiently refined to detect a parallax, indeed, that led to the beginning of stellar photometry. By attempting to compare the brightness of Sirius with the light emitted from a pinhole-sized area of the solar surface (assuming that the Sun, Sirius and all stars possessed the same intrinsic luminosity), Huygens calculated that Sirius was 27 664 astronomical units distant from the solar system. Huygens's technique encouraged the photo-astrometric studies of Sir John Herschel in 1836, where he calculated that Vega was 40 times brighter than the Sun, and Arcturus 200, and hence correspondingly remote from the solar system. Then in the wake of the successful application of spectroscopy and photography to astrometry in the 1860s, photometry began its independent development as a fundamental tool of astrophysics.

The quest for the stellar parallax was also to lead ROBERT HOOKE (1669), John Flamsteed (1698–1702) and especially JAMES BRADLEY (1727–1728) to the invention and development of the zenith sector (a long-focus telescope hung to point exactly at the zenith) as a way whereby the stellar parallax, now recognised as being no more than a few seconds of arc in extent, might be detected in the atmospheric-refraction-free sky of the zenith. Both Hooke and Flamsteed came up with suspiciously large parallaxes: Hooke made that of the star γ Draconis to be around 27" and Flamsteed that of Polaris 40", although the superior instrumentation and observing procedures of Bradley led to his recognition in 1728 that a 'New Found Motion' of the fixed stars, which was itself the product of the Earth's rotation about the Sun, had an annual displacement of 20.00"–20.25". This was the aberration. Its detection and quantification were a triumph of the refined astrometry of the 18th century and, while still not revealing a parallax, none the less it provided physical proof of the Earth's motion in space. Bradley claimed that his zenith sector was capable of detecting an angle of 0.5", and in 1748 he used the tiny nodding motion of γ Draconis to detect a small periodic change in the inclination of the earth's axis of rotation with regard to the ecliptic.

One of the assumptions implicit within the concept of the infinite universe was the possible independent motion of the stars themselves, especially in the wake of Sir ISAAC NEWTON's theory of universal gravitation after 1687. From time immemorial it had been assumed that the stars were fixed in their places, although in 1718 EDMOND HALLEY published a paper which showed that the first-magnitude stars Aldebaran (Palilicium), Sirius, and Arcturus had displayed an independent or 'proper motion' *vis-à-vis* their adjacent, dimmer stars. This had been the product of an ingenious historical investigation, in which Halley had detected a systematic displacement in latitude of these stars, as recorded by Hipparchus, Ptolemy, Tycho and contemporary astronomers. Indeed, he calculated that Sirius had changed its longitudinal position by 2' in the 120 yr since Tycho. The implication was not only that these stars might be moving in space but that our observing position in the solar system might be moving as well. William Herschel in 1783 was the first to use proper motions to derive the solar apex, the point in the sky towards which the solar system is moving (see HERSCHEL FAMILY).

It says something about the rapid advances in astrometry and its instrumentation that no fewer than three stellar parallaxes were found between 1838 and 1840, by astronomers using three different techniques of observation. None of these parallaxes, moreover, would have been measured but for the discovery and increasingly accurate quantification of stellar proper motions, for it was the large proper motions displayed by 61 Cygni (5".2 yr^{-1}) and α Centauri (3".7) that suggested their relative closeness to the solar system and hence their suitability for parallax measurement. Although displaying a much smaller (yet clear) proper motion, Vega (α Lyrae) was also considered as a suitable candidate for parallax measurement. These stars were measured respectively by FRIEDRICH WILHELM BESSEL at Königsburg, using a heliometer, by THOMAS HENDERSON at the Cape of Good Hope, using a mural circle, and by Friedrich Georg WILHELM STRUVE at Dorpat, with a filar micrometer on the 24 cm Fraunhofer refractor. Although their figures were revised somewhat by later observers, using larger runs of observations, their parallaxes remain

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

substantially correct: 61 Cygni, 0".3483; α Centauri, 0".913; Vega, 0".261. In many respects, these parallaxes represented a high-water mark in astrometry, whereby a long-standing astronomical question had been answered by the combined application of higher mathematical theory and superlatively accurate observations made with a new breed of instrument.

The early and middle decades of the 19th century saw the coming of age of manual astrometry (in which the astronomer worked with eyes, hands and ears, without any electromagnetic or photographic aids). This astronomy was directed not only to the definitive establishment of a series of astronomical constants but to the provision of a continuously updated and refined body of data from which original investigations in celestial mechanics could proceed. They came to depend, moreover, on the integrated use of Anglo-German techniques in instrumentation and French mathematical analysis.

The study of binary and triple stars, which had been largely pioneered by William Herschel in 1802, showed a way in which Newtonian gravitation, which hitherto had only been shown to operate in the solar system, was manifestly universal. Bessel, Struve, Sir James South and others across northern Europe had tried to quantify the orbital elements of a binary system, as a way of extracting the gravitational elements of ξ Ursae Majoris and other star systems, and it was Felix Savary, Johann Encke and John Herschel in 1830–1832 who succeeded. Binary star work demanded the highest levels of instrumental accuracy, as equatorially mounted refractors, with achromatic optics by Josef Fraunhofer, Robert Aglaé Cauchoix and others, incorporating micrometer or heliometer functions of the greatest delicacy, made reliable astrometric data available by 1830 that would have been unimaginable in 1780. By 1833, John Herschel was publishing the orbital axial components of six double stars, to four decimal places of a single second of arc, and while many of the hundredths and thousandths of second of arc were subsequently modified, their tenths, as originally measured, rarely needed correction. The refined astrometric investigations of the mid-19th century revealed the presence of many systematic stellar irregularities from which a rich harvest of physical discoveries emerged.

By 1840, all the great observatories of Europe—Königsburg, Dorpat, Berlin, Bonn, Greenwich and others—were primarily concerned with astrometric work, which by that date represented the supreme pinnacle of astronomical achievement. On the one hand, their energies were devoted to the equatorial-refractor-based work of binary star astronomy, although the mural and transit circle also facilitated the routine measurement of meridian positions to within a fraction of a single second of arc. As it was becoming *de rigueur*, in the wake of the examples of Bessel, Struve, Encke, Friedrich Wilhelm Augustus Argelander and George Biddell Airy, for an observatory to publish its work as quickly as possible, a vast body of superlative astrometric data was rapidly becoming available across Europe and America.

It was from this published astrometric archive of meridional data that both John Couch Adams and Jean Joseph Urbain Le Verrier made their spectacular independent discoveries of Neptune in 1846. It was also from relatively long runs of high-quality astrometric data that Bessel was able to quantify a periodic wobble in the position of Sirius from which he was able to compute the presence of Sirius B in 1844. Moreover, both of these discoveries, extracted from existing astrometric data, were followed by telescopic confirmation, when Johann Galle and Heinrich d'Arrest observed Neptune from Berlin on 23 September 1846 and when Alvan Clark saw Sirius B with a new 18 in object glass in Boston in 1862.

The invention of photography in 1839 opened up enormous potential for astrometry. In addition to its capacity for much greater accuracy, it liberated the astronomer from the eyepiece of a telescope, to permit the more comfortable, relaxed measurement of astrometrical positions from photographic plates. However, it was not until after 1881, with the establishment of the *Carte du Ciel*, that photographic astrometry really showed what it could do.

As astrophysics has come to dominate astronomy in the 20th century, however, the fundamental role of astrometry within the science has declined, but the development of radio interferometry in the 1950s introduced a new and important form of astrometry, as radio dishes many miles or even continents apart were used to locate the exact positions of radio sources in the sky to within seconds of arc.

The development of the European Space Agency's HIPPARCOS project in 1989, however, has brought optical astrometry back to a central role once again. The Hipparcos space satellite has enabled the production of excellent astrometric measurements down to 0.002" on a virtually industrial basis without any of the atmospheric problems associated with ground-based observatories. Although astrophysics still contributes much of the driving force of modern astronomical research, critically accurate astrometric data of the kind made possible by Hipparcos are essential, especially as a way of understanding the physical processes going on within our own star system.

Satellite astrometry, indeed, is clearly seen as the way forward when it comes to the most critical accuracies. The forthcoming NASA, FAME (Full Sky Astrometric Mapping Explorer), due to be launched around 2004, will map an incredible 40 million star positions, with an increase in accuracy by a factor of two to ten (depending on magnitude) on previous satellites, during its expected 5 yr of operation. The ESA GAIA (European Space Agency Global Astrometric

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interferometer for Astrophysics), moreover, will also have a 5-yr operational life, in which it is expected that it will map 500 times more objects than Hipparcos, 100 times more accurately.

It could be argued, moreover, that astrometry in its photographic, photo-electric, radio and satellite departments still lies at the heart of astronomical research, for astrophysics itself hinges upon the precise measurement and interpretation of various forms of energy from space. For it is the very act of using instruments to measure and quantify the energy that reaches the Earth from space that makes astronomy by far the oldest of the exact sciences.

Bibliography

Chapman A 1995 *Dividing the Circle: A History of Critical Angular Measurement in Astronomy, 1500–1850* 2nd edn (Chichester: Praxis-Wiley)

Grant R 1852 History of Physical Astronomy (London: Bohn)

Herschel J F W 1833 *A Treatise on Astronomy* (London: Longmans)

Herschel J F W 1849 *Outlines of Astronomy* (London: Longmans)

Hoskin M (ed) 1997 *The Cambridge Illustrated History of Astronomy* (Cambridge: Cambridge University Press)

King H C 1955 *The History of the Telescope* (London) (a modern reprint edition is available)

North J D 1994 *Fontana History of Astronomy* (London: Fontana)

Pannekoek A 1961 *A History of Astronomy* (London: Allen & Unwin)

Allan Chapman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrometry: Telescopes and Techniques

The goal of astrometry is to determine positions and motions of celestial objects such as stars and planets, as well as artificial satellites. Astrometry is the oldest branch of astronomy and until the 19th century all of professional astronomy was astrometry (see ASTROMETRY: HISTORY). One of the main astrometric objectives is to measure angular separations of stars and other celestial objects relative to reference stars which have known positions. This can be achieved by taking images through a telescope, similar to a camera, be it on photographic emulsion (see DETECTORS: PHOTOGRAPHY) or using electronic detectors such as a CHARGE-COUPLED DEVICE (CCD) as used in most cameras today. Reference star positions are established with methods of absolute astrometry . A catalog of such positions and motions defines a reference system. Most astrometric measurements are made in two dimensions, on the celestial sphere, similar to geographic latitude and longitude. The third dimension, the distance of celestial objects in space, is much harder to measure and one of the biggest astronomical challenges. Only distances within our solar system can be measured with great precision using radar and laser ranging techniques.

Very precise angular measurements on the order of 1 milliarcsecond (mas) can be achieved today by ground-based optical telescopes, astrometric satellites and radio telescopes. An angle of 1 mas is the 1000th part of an arcsecond, which in turn is the 3600th part of a degree. The Moon in the sky measures about half a degree or 1800 arcsec across. A pair of candles 10 cm apart are separated by 1 mas when seen from a distance of 20 000 km.

Not only the Sun, Moon and planets are moving in the sky. If you look closely enough, the 'fixed' stars move as well, slowly circling the center of our Galaxy, the Milky Way. Even galaxies move within clusters of galaxies. Thus precise astrometric positions are time dependent. Without knowing the motions, no good prediction of positions for the future can be made.

A dramatic increase in precision has been achieved in the last two decades, after centuries of slow improvements. In 1980 an absolute position good to 100 mas was state of the art, since 1995 it has been about 1 mas and space missions currently being designed promise 0.05–0.005 mas. This amazing progress is technology driven, taking advantage of new techniques and hardware.

Astrographs and the tangential plane concept

One of the basic astrometric techniques is the use of a camera (astrograph) to map a section of the celestial sphere onto an image surface. A camera here can also be a telephoto lens, a dedicated lens telescope (refractor) or mirror telescope (reflector). The image surface is ideally a plane, orthogonal to the optical axis of the camera, tangential to the sphere. The camera (figure 1) is aimed at the position A in the sky and a star at S is mapped onto S' in the tangential plane. The focal length, f, determines the scale of this mapping:

$$\text{scale (arcsec mm}^{-1}) = 206\,265/[f\ (\text{mm})].$$

The basic idea is to measure rectangular x, y coordinates in the tangential plane, then use the known position of some reference stars in the field to determine the mapping parameters and finally to calculate the spherical coordinates of the other measured field stars. Technical details are explained below.

Wide field

A telescope with a high-quality, wide FIELD OF VIEW of several degrees requires a complex optical design, which cannot be achieved with a two-lens refractor or a two-mirror telescope. Multilens astrographs are the classical telescopes in this category, similar to telephoto lenses. For larger apertures (about 0.7 m or more), catadioptic systems are often used, which are reflecting telescopes with additional lenses. Examples are the SCHMIDT TELESCOPE and large reflectors with 'field correctors'.

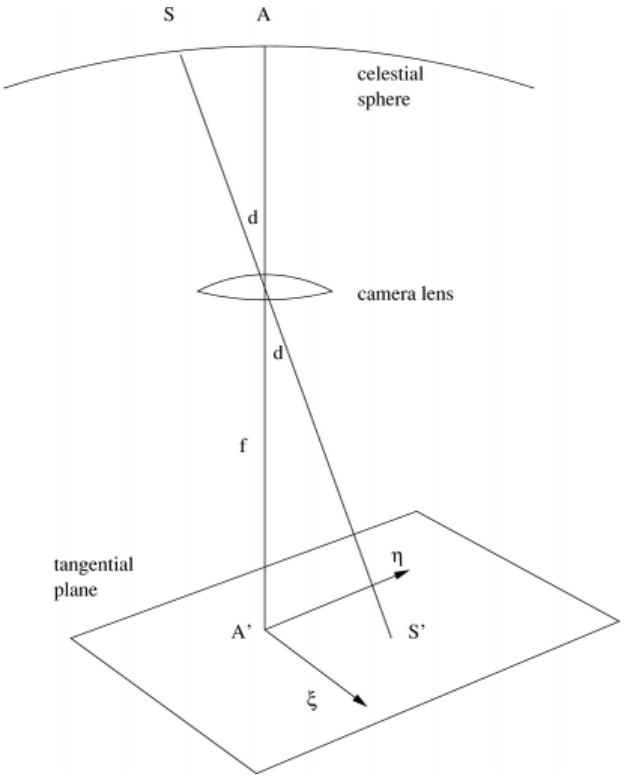


Figure 1. The tangential plane concept.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

These wide-field telescopes are used in astrometry mainly to densify existing reference systems. The positions of many fainter stars are determined with respect to a few bright reference stars with known positions. The large field of view allows for simultaneous observation of a sufficiently large number of reference stars in any given field. The most important examples of astrographs are the Astrographic Catalog (AC) astrographs, built around 1900 with 3.7 m focal length, the Astronomische Gesellschaft Katalog (AGK) series 2 and 3 instruments of 2 m focal length, the five-lens 1975 design of the Hamburg astrograph, the US NAVAL OBSERVATORY (USNO) twin astrograph with its 1990 designed red-lens, and the Lick and Yale 0.5 m aperture astrographs.

The Lick and Yale astrographs use distant galaxies as fixed benchmarks and a set of exposures taken about 20–30 yr apart (epoch difference) to determine absolute PROPER MOTIONS of stars in the Northern and Southern Proper Motion (NPM, SPM) programs respectively. The NPM is complete, while second-epoch observations of SPM in Argentina are in progress.

The USNO CCD Astrograph Catalog (UCAC) project started in 1998 at CERRO TOLOLO INTER-AMERICAN OBSERVATORY and will eventually cover all sky to about 16th magnitude. Positions of stars in the 10–14 mag range are accurate to about 20 mas. Proper motions will be obtained by utilizing new measures of the early-epoch NPM and SPM data.

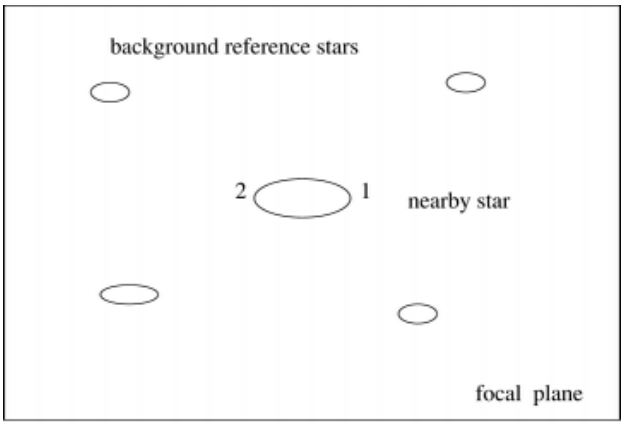
Narrow field

Long focal length (≈ 10–15 m) refractors and large reflectors provide a field of view of several arcminutes to 1º at high magnification to resolve tiny displacements of stellar images. These telescopes are mainly used in astrometry for the determination of trigonometric PARALLAXES, which is the most direct, fundamental technique to determine distances outside the solar system. Small shifts of position of a relatively nearby target star are measured relative to distant stars, which are selected by spectral and photometric properties. The tangential plane concept is used here as well; however, no precise (absolute) reference star positions are needed. Observations of a given field are obtained over several years and then mapped onto each other, assuming the mean of the background stars fixed.

Figure 2 explains the principle of trigonometric parallax observations. While the Earth circles the Sun the position of a nearby star changes slightly (order 0.5 arcsec or less). The largest shifts are observed from positions 1 and 2, about half a year apart. On the focal plane the positions of a nearby target star and some distant background stars are plotted over the course of a year (neglecting proper motions). A major problem is to estimate the distances (parallaxes) of the background stars. Ideally they should be infinitely far away, thus being at a fixed position, the center of the ellipses shown in figure 2. In that case the observed position change of the target star with respect to the background stars would directly give its true, absolute parallax. In reality, the measurement reveals the relative parallax with respect to the mean parallax of the background stars.

Examples are the refractors of the YERKES OBSERVATORY and ALLEGHENY OBSERVATORY as well as the US Naval Observatory 1.55 m Strand reflector, the most precise and productive instrument in this category. Results of all significant trigonometric parallax and associated proper motion observations are compiled in catalogs issued by the Yale University.

In order to obtain high-precision positions of faint objects such as compact, extragalactic radio reference frame sources, a multistep procedure is adopted. A wide-



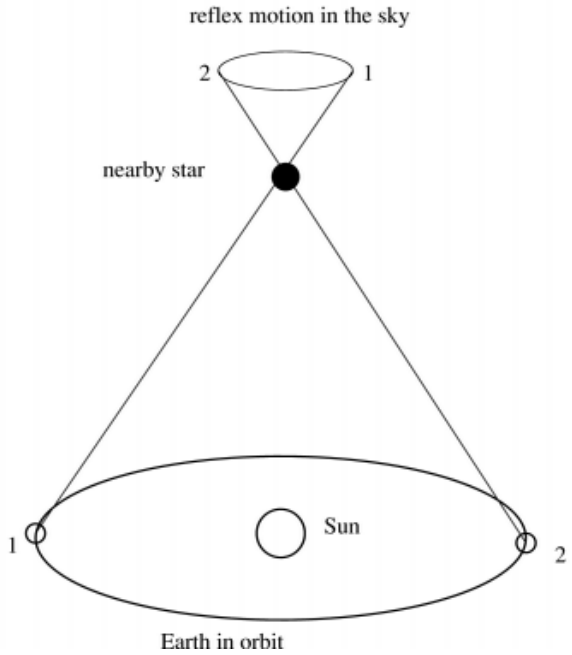


Figure 2. Principle of trigonometric parallaxes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

field telescope is used to obtain positions of secondary reference stars in the system of primary reference stars. Long-focus instruments continue towards fainter limiting magnitudes and densify the net of reference stars.

Detectors

There are mainly three types of detectors important in optical astrometric programs today: the CCD, the grid modulation photometer and photographic emulsions.

A CCD detector at a telescope is the most important detector, widely used in astronomical data acquisitions. The high quantum efficiency and geometric stability make it a very suitable device for astrometry. The only major problem with a CCD is its relatively small size as compared with a photographic plate. A typical CCD today measures 30–50 mm on a side, while a typical photographic plate can be 100 times larger in area. A mosaic of CCDs can cover a large area in a focal plane; however, gaps remain between chips and, more importantly for precise astrometry, additional parameters in the mapping model are needed for each single CCD chip. This can be a serious problem when the reference star density is low.

The grid modulation technique is applied in the Multichannel Astrometric Photometer (MAP) of the Allegheny Observatory. It measures relative positions of stars in the focal plane with the help of a moving ronchi ruling. This instrument is very precise (3 mas precision per night on a bright target); however, only one target source at a time can be measured with respect to several reference stars. The grid modulation principle is also used in the HIPPARCOS space mission and is explained in figure 3. Photometric measurements with high temporal resolution are made behind a grid in the focal plane. Stars (or the grid) are moving and the phase difference between the signals from two stars gives their precise relative position in fractions of the grid constant. *A priori* knowledge is required for the integer number of grid steps the stars are separated.

Photographic emulsions, on either film or glass plates, require a measuring machine to extract the x, y information of stellar image positions. Photographic emulsions are no longer used in any major current astrometric program. However, the data stored in plate archives are still valuable and their extraction is by no means completed yet.

Measuring machines

In the past humans manually operated one- and two-axis plate measuring machines, centering stellar images on microscope cross-hairs and reading circles. Image dissector techniques can automatically center a stellar image utilizing the symmetry property of the image profile. A complete digitization (scanning) of photographic plates is performed on modern machines. In this case a photometric measurement (transparency or density) is recorded as a function of the x, y position of the scanning detector. This can be achieved by using a microdensitometer, a CCD or a line-scanning device. In all cases an exact measure of the location of the detector with respect to the plate along two axes is critical and most machines use laser metrology or high-precision encoders providing a repeatability of about 0.1 μm. The accuracy is limited by systematic errors in the measuring process and the geometric stability of the photographic media which is on the order of 0.4–1.0 μm, depending on the type of emulsion.

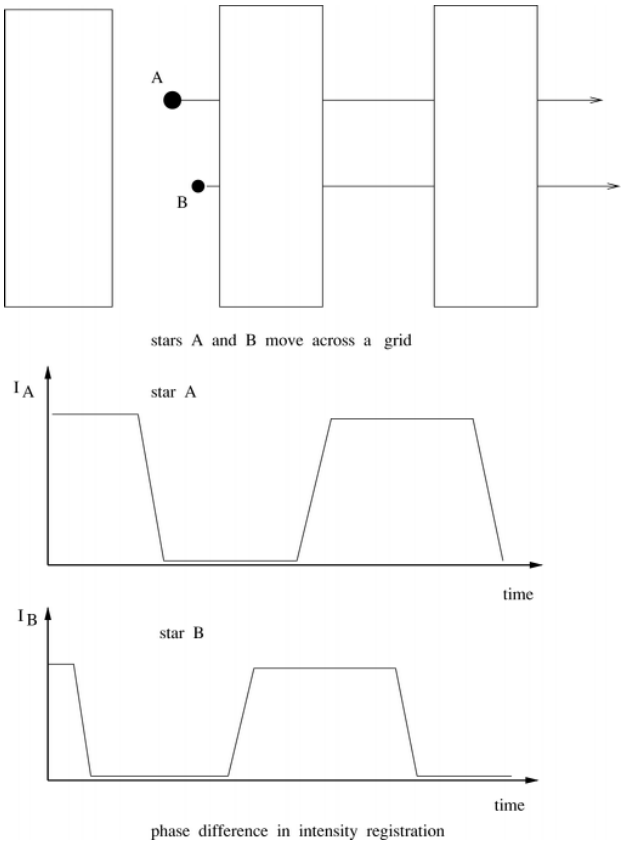


Figure 3. Principle of the grid modulation technique.

Examples are PDS microddensitometer machines, which were commercially available, the GALAXY machine of the former ROYAL OBSERVATORY, GREENWICH the APM at Cambridge, UK, the APS at Minneapolis, the MAMA at OBSERVATOIRE DE PARIS, the PMM of the US Naval Observatory, the HAM1 at Hamburger Sternwarte and the SuperCosmos of the ROYAL OBSERVATORY, EDINBURGH.

Transit circle and absolute astrometry

A transit circle (TC) is a telescope mounted on a fixed east–west axis, so it can move only along the local meridian. The meridian is the great circle through the celestial poles and the zenith, perpendicular to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

horizon (figure 4). The TC makes two independent observations: the time when a target object crosses the meridian (transit) and an angle, the zenith distance, which is read off circles attached to the east–west axis of the telescope. The zenith direction is determined in practice by looking in the opposite direction (nadir), down into a mercury mirror surface which represents the local horizon plane due to the Earth's gravity. Collimated fiducial marks located some distance towards the north and south are observed to check on the stability of the instrumental azimuth.

In past centuries transit observations were made by the 'eye and ear' method, watching a star crossing the micrometer wire through the eyepiece of the telescope and listening to time signals to estimate the fraction of a second. Modern instruments use photoelectric detectors and follow the star for some time in the focal plane to obtain multiple measures for a single transit.

A TC can perform relative observations; however, it is also capable of absolute (fundamental) astrometry, meaning that it can determine positions of celestial objects without *a priori* knowledge of other star positions. Thus TC telescopes provided the primary reference star catalogs (fundamental catalogs) throughout centuries. The most important are the FK3, FK4 and FK5 series of this century, published by the Astronomisches Recheninstitut in Heidelberg, Germany. Recently, fundamental work with TC telescopes has been superseded by more accurate techniques such as the Hipparcos space mission and radio VLBI observations (see below).

TC observations are affected by a variety of issues. Instrumental imperfections such as flexure of the telescope tube can be determined by appropriate modelling of a series of observations. Critical is the stability of the instrument to repeatably follow the same pattern over a given interval of time. Polar motion, the wandering of the Earth's axis with respect to the surface of the Earth, and other changes of the local coordinate system enter TC observations and need to be resolved. Furthermore, the rotation axis of the Earth is not fixed in inertial space. A coordinate system related to the equator of the Earth is moving with respect to the stars (precession and nutation). The finite speed of light causes astronomical aberration, an angular shift of the observed trajectory with respect to the true position. All those effects enter absolute TC observations to the full extent, while differential (relative) observations have to deal only with variations of those effects over the field of view.

Of particular concern for TC observations is the refraction in the Earth's atmosphere. Observations of both the lower and upper transit of stars near the pole help to determine a refraction model. Nevertheless, refraction is changing with atmospheric conditions such as temperature. The turbulence of the atmosphere ultimately limits absolute TC observations by unmodeled systematic contributions and random noise. Examples of TC instruments are the CARLSBERG MERIDIAN TELESCOPE and the meridian circle at the OBSERVATOIRE DES SCIENCES DE L'UNIVERS DE BORDEAUX and the 6 and 7 in telescopes of the US Naval Observatory.

Scanning TDI technique

A recent innovation in astrometry is the use of CCDs in the focal plane of TC-type telescopes. The CCD can be operated in a time-delayed integration (TDI) mode, where the charge transfer for the CCD readout is synchronized with the motion of the star across the chip. However, this works well only for stars moving along a great circle (equator) and for a small width of a strip being scanned. In all other cases stars in different locations on the CCD move along significantly non-linear paths, resulting in image smearing.

A pioneering instrument of that type is the Flagstaff Astrometric Scanning Transit Telescope (FASTT). The Carlsberg Meridian Telescope (CMC, La Palma and Cambridge, UK) and the Bordeaux TC have utilized the TDI scanning technique for zonal catalogs. The imaging part of the SLOAN DIGITAL SKY SURVEY (SDSS) project also uses the same technique. However, the SDSS telescope can follow arbitrary great circles on the sky by actively steering the telescope and rotating the detector plane at the same time.

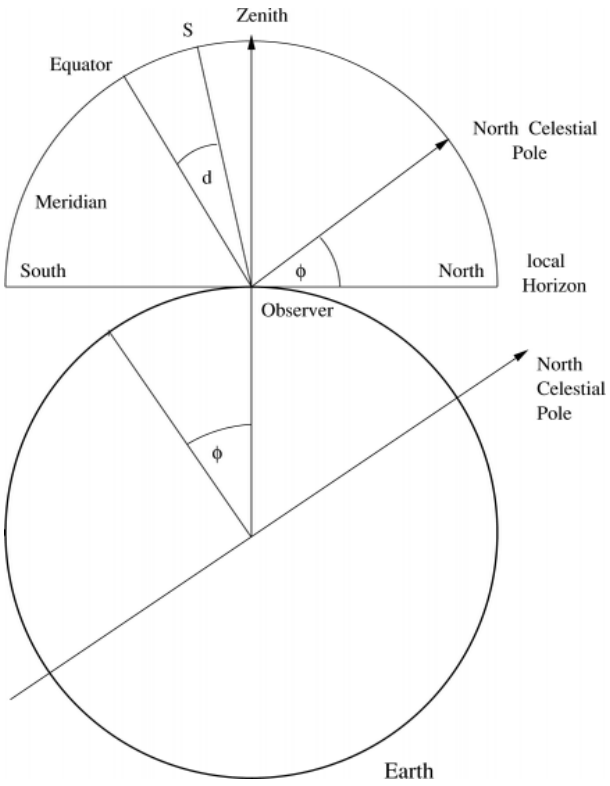


Figure 4. Celestial sphere for an observer on Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrolabe

An ASTROLABE is an example of an equal altitude instrument, measuring the time of transit of a star through a small circle of fixed altitude or zenith distance (often 30º or 45º). This is achieved by a mirror–prism configuration in front of a horizontally mounted telescope. Direct and reflected images of a target star are mapped into the focal plane. The apparent motions of both images are in opposite direction, and they cross each other when the star reaches the fixed zenith distance the instrument is designed for. Star positions can be deduced from observations of crossings toward the east and west.

There are two main advantages of an astrolabe over a TC. All stars are observed at a fixed zenith distance, simplifying the corrections for refraction, and the instrumental setup is more stable, avoiding flexure problems. The main disadvantage of an astrolabe is its restricted sky coverage in declination, δ, of $\phi - z < \delta < \phi + z$, with ϕ being the geographical latitude and z the fixed zenith distance of the instrument. As with the TC, the astrolabe's capability to produce a fundamental catalog and monitor the Earth's rotation is now only of historical interest owing to superior accuracies from VLBI and space astrometric techniques.

Photographic zenith telescope (PZT)

A PZT is a vertically mounted telescope with a lens on a rotary stage on top and a mirror at the bottom. With a PZT the relative position of the local zenith with respect to the stars is observed. For decades the PZT was utilized for Earth orientation determination. Since VLBI techniques are now much more accurate, the PZTs are obsolete.

Interferometry

Interferometry is the most accurate astrometric technique currently known, utilizing the wave nature of coherent light (or electromagnetic radiation in general) for angular measurements. Widely used is the Michelson interferometer. Its basic principle is illustrated in figure 5. Two fixed telescopes (sidereostats) A and B observe the same star. The beam from at least one of the telescopes needs to pass through a delay line of variable length, in order to observe the same incoming wave front simultaneously (coherent) and thus obtain an interference pattern (fringes) at the beam combiner. Depending on the position of the star relative to the baseline (angle γ), the delay (d) varies and has to be adjusted accordingly. The basic interferometric measurement is the delay length as a function of time, while keeping the fringe contrast (visibility) at a maximum (see INTERFEROMETRY):

$$\cos\gamma = d/b.$$

With the known baseline b this gives a measure of γ. Observing several stars many times in a night allows us to solve for the positions of the stars and Earth orientation parameters. *A priori* information on the approximate star positions is required to steer the delay line and to obtain fringes. The interferometer then determines d to a fraction of the wavelength, while the baseline length is monitored with additional metrology. An accuracy of a few mas can be obtained with this technique globally and relative measures (double stars, stellar limb darkening) are at least 10 times more accurate. As with transit circles, the properties of the atmosphere are currently limiting the accuracy.

The Mark I (MOUNT WILSON OBSERVATORY) was the first phase-coherent stellar interferometer, followed by Mark II and Mark III (around 1990). Examples of operational optical interferometers are the COAST (Cambridge, UK), SUSI (Australia), GI2T (France), PTI (PALOMAR OBSERVATORY) and the Navy Prototype Optical Interferometer (NPOI). Only COAST and NPOI have three or more sidereostats and regularly can perform imaging. The CHARA instrument has started operation on Mt Wilson. The SPACE INTERFEROMETRY MISSION (SIM) project (approved by NASA for launch in 2009) is based on the same technique.

Similar techniques are used at radio frequencies. The signals I_A and I_B from two radio telescopes are fed into a correlator. The correlator product, C,

$$C = \int_{t_1}^{t_2} I_A(t) I_B(t+\tau)\,dt$$

is at maximum when $\tau = d/c$ with c being the speed of light and d the delay as in figure 5.

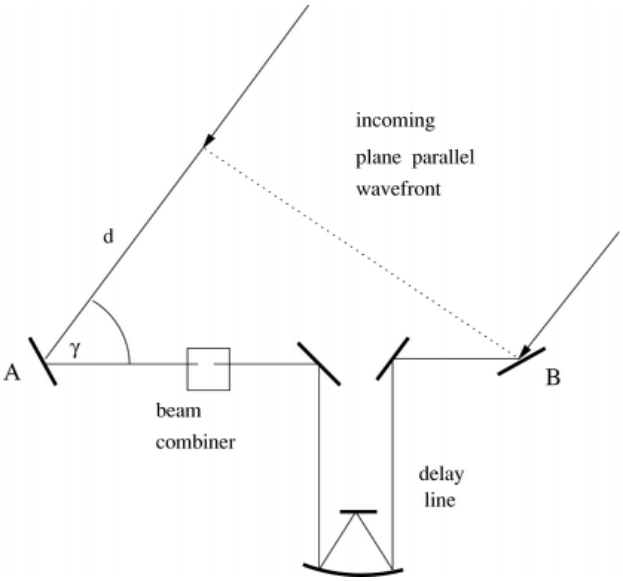


Figure 5. Principle of a Michelson interferometer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Very long baseline interferometry (VLBI) is the most accurate radio astrometric technique. Here the individual telescopes are usually far apart on different continents and no longer directly connected. Instead, the signal at each station is modulated with a very precise frequency standard (atomic clock, maser), digitized and written to special tapes. The tapes from all stations are later processed at a correlator center and delays and delay rates are extracted as basic information.

The resolving power, R, of an interferometer no longer equals the resolving power of its individual apertures; instead the Rayleigh criterion

$$R\ (\text{rad}) = 1.22\lambda/b$$

holds for the separation of two linked telescopes with a baseline length b. This is particularly important at radio wavelengths, λ, which are very long (millimeter to meter) as compared with optical wavelengths.

Optical interferometers as well as VLBI have the capability for absolute astrometry. VLBI currently has the highest accuracy of all wide-field astrometric techniques. Earth orientation parameters are being routinely obtained on the sub-mas level and the International Celestial Reference System (ICRS) is defined by the positions of 212 extragalactic, compact radio sources observed with VLBI.

Speckle interferometry

Speckle interferometry is a technique to overcome the seeing limitation of ground-based, optical–IR telescopes in very small field applications such as double-star observations. If the aperture of a telescope is large as compared with the Fried parameter ($r_0 \approx 0.15$ m, see ADAPTIVE OPTICS), a large number of individual speckles are seen within the seeing disk when a star is imaged at high magnification. Taking individual images with short integration times (≈1/30 s) allows one to resolve the speckles. The size of the speckles is determined by the aperture of the telescope. The autocorrelation, A,

$$A(u) = \int I(x+u)I(x)\,\mathrm{d}x$$

over the image area reveals the separation and orientation angle (with a 180º ambiguity) of a double star with a resolution similar to the theoretical resolving power of the telescope without the influence of the atmosphere. A more complex structure in the object can be analyzed with Fourier transformation techniques. Speckle interferometry works only within the isoplanatic patch and thus is limited to a field of view of about 5 arcsec.

Space astrometry

The HUBBLE SPACE TELESCOPE (HST) has two instruments for astrometric observations, the Fine Guidance Sensors (FGS) and the Wide Field and Planetary Camera (WFPC). The FGS allow a two-dimensional relative position determination of objects within a ≈13' by 4' annular field of view, observing one target at a time (see ASTROMETRY WITH HUBBLE SPACE TELESCOPE).

The second Wide Field and Planetary Camera (WFPC2) of HST uses four CCD detectors for direct imaging in bandpasses from UV to IR. Astrometric information can be obtained following the tangential plane concept as described in the first section. Optical field angle distortions have been determined with dense calibration fields (globular cluster), reducing the number of parameters required in the mapping model for user applications. The field of view is only about 75" on the three Wide Field CCD chips and about 34" on the Planetary Camera CCD. Correcting optics is integrated in the second WFPC2, which was installed in December 1993. The instrument is suitable for astrometry of very faint sources (to ≈26 mag); however, the images are undersampled and dithering techniques are used for high-precision centering.

Hipparcos, the High Precision Parallax Collecting Satellite, was the first space mission dedicated to astrometry (see HIPPARCOS). Launched in 1989, it observed the entire sky for 3.5 yr, producing a catalog of 118 000 stars with absolute positions, proper motions and trigonometric parallaxes on the 1 mas accuracy level. With a split mirror, two fields in the sky separated by ≈58º were mapped into the same focal plane. The satellite continuously rotated and light from stars crossing the fields was modulated by a well calibrated grid. Using an input catalog, selected spots in the focal plane were picked up by the photoelectric detector. The raw data comprise a steady stream of light intensity measures as a function of time. Only one dimension (along the scan) was observed with high precision. The rotation axis of the satellite slowly precessed to allow for a global coverage of the sky. Each point in the sky was scanned about 120 times with a variety of scan directions in order to resolve both coordinates. Abscissa data, angles measured along the scan direction, of all great circle scans were later combined in a single large reduction (sphere reconstruction) in order to simultaneously determine the astrometric parameters of stars as well as instrumental parameters. As a result, a rigid, self–consistent, instrumental coordinate system was established.

The external orientation of the axes (three position and three proper motion parameters) were provided in combination with ground–based observations, linking the Hipparcos frame to the ICRS. All stars to about 7th magnitude were observed in two bandpasses, similar to B and V, as well as most stars to magnitude 8.5 and some selected to 12.5 mag.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The star mapper data of the Hipparcos mission were utilized to survey all stars to about 11th magnitude in the Tycho project. A re-evaluation of the Tycho data, completed in 1999, produced the Tycho-2 catalog of 2.5 million stars (photometry and position), supplemented by proper motions in connection with ground-based catalogs. The accuracy of the Tycho positions is on the order of 10–70 mas depending on magnitude (see TYCHO STAR CATALOGS).

In 1999, NASA selected the Full-Sky Astrometric Mapping Explorer (FAME) space mission for a scheduled launch in 2004. Similar to Hipparcos, this satellite will scan the entire sky in a 5 yr mission to obtain positions, proper motions and parallaxes of 40 million stars with an accuracy of 0.5 mas–50 μas, depending on magnitude. In contrast to Hipparcos, FAME utilizes an array of CCDs in scanning TDI mode (see above).

A global view of astrometric techniques

More than in many other disciplines of physics and astronomy, a thorough investigation of measuring errors is an essential part of astrometry. The precision of an observation is an internal, formal error, often derived from random error statistics of repeated observations. The accuracy of an observation refers by definition to its true error, which is almost never known precisely. The accuracy is often estimated from external comparisons with results obtained with an independent set of data or technique. The differences between the values of precision and accuracy estimates are caused by systematic errors, and a large fraction of all astrometric effort is to detect and reduce systematic errors.

This starts with the design of astrometric telescopes. One common theme in many astrometric instruments is the built-in feature of reversal. A TC can be lifted out of its mount and put in reversed. An astrograph usually can be shifted from the east side of the pier to the west, in order to take the same field in the sky 'upside down'. A plate measuring machine can measure plates in several orientations. Another feature of astrometric importance is cross-connection and redundancy. Individual TC programs largely overlap in the sky coverage with each other. Hipparcos observed two fields of view widely separated simultaneously, and VLBI uses multiple telescopes and baselines for simultaneous observations of the same source. Combined with appropriate reduction methods this allows one to detect and correct some systematic errors caused by part of the instrumentation.

However, this can only be achieved with stable instruments. Properties of an astrometric instrument must not change with time or they must be monitored and accounted for. Often a non-astrometric telescope such as the Schmidt telescope is used to obtain astrometric results, which then become compromised. Often this is the only available method to obtain some results at all, e.g. positions of very faint stars all over the sky. Because of the revolution in computer technology, it is nowadays cheaper to control and correct instrumental defects afterwards than to start with a rigid nicely behaving instrument in the first place. Examples are the point spread function (PSF) and geometric mapping properties of modern telescope designs, from ground-based wide-field survey telescopes to the Next Generation Space Telescope (NGST) as compared with the PSF of an astrograph.

For an astrometric instrument usually its resolution (Rayleigh criterion) is not the limit of precision. A stellar image profile of about 1 arcsec diameter might very well be centered to 1/100 arcsec precision or better.

Space- versus ground-based techniques

Basically there are three things different in space astrometry as compared with ground based.

1. *Atmosphere*. The lack of the Earth's atmosphere allows for sharper images, not disturbed by a turbulent medium. The higher resolving power directly leads to a better signal-to-noise ratio and to fainter limiting magnitudes. The lack of turbulence also means a significant reduction in random noise for the measurement of angular separations. The atmospheric random error is roughly

$$\sigma_{\mathrm{atm}} \approx k t^{-1/2} F^{1/3} (\cos z)^{-1}$$

with exposure time t in seconds, field of view F in arcminutes, zenith distance z and the parameter k in the range 10–70 mas, depending on atmospheric conditions. The availability of a larger frequency spectrum (including UV and IR) for space-based observations so far has not been of significant importance to astrometry.
2. *Gravity*. Near-zero gravity for space-based instruments directly leads to greater stability and independence of some errors from the pointing direction. However, thermal effects (satellite in sunlight versus shadow) cause other problems and temperature control to 1/100 K levels is often a requirement for astrometric space missions. The gravitational gradient of Earth-based instruments provides a convenient method to define one axis of a coordinate system. This is no longer available with space missions.
3. *Orientation*. Independence of the Earth means no problems with Earth rotation and irregularities of local terrestial coordinate systems. It also means no accessible definition of north or a similar external orientation. The equator has no meaning for space missions and the plane of the ecliptic only weakly enters in light aberration terms.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tangential plane reduction technique

Refering back to figure 1, let ξ, η be the rectangular Cartesian coordinates defined as zero point at the intersection of the optical axis with the tangential plane (A'), ξ pointing towards east and η north. A gnomonic projection gives the tangential coordinates as

$$\xi = \frac{\cos\delta \sin(\alpha - \alpha_0)}{\sin\delta_0 \sin\delta + \cos\delta_0 \cos\delta \cos(\alpha - \alpha_0)}$$

$$\eta = \frac{\cos\delta_0 \sin\delta - \sin\delta_0 \cos\delta \cos(\alpha - \alpha_0)}{\sin\delta_0 \sin\delta + \cos\delta_0 \cos\delta \cos(\alpha - \alpha_0)}$$

with α, δ being the right ascension and the declination of S and α_0, δ_0 those of A. The unit for ξ, η is radians at this point.

Measured coordinates *x*, *y* if exactly parallel to ξ, η are then

$$x = f\xi \qquad y = f\eta$$

with *f* being the focal length and *x*, *y* in the same units as *f*, e.g. millimeters. Equation (8) actually is the simplest physical model between the purely mathematical ξ, η and the real-world observables *x*, *y*. An example of a more realistic plate model is

$$\xi = P_1 x + P_2 y + P_3 + P_5 x + P_6 y + P_7 x^2 + P_8 xy + P_9 xr^2$$

$$\eta = -P_2 x + P_1 y + P_4 + P_6 x - P_5 y + P_7 xy + P_8 y^2 + P_9 yr^2$$

with P_1 to P_9 being parameters to be estimated in an adjustment process using reference stars with known positions (ξ, η), and $r^2 = x^2 + y^2$. Parameters 1 to 4 represent an orthogonal model, which is usually the simplest practical model. It allows for a zero point, scale and rotation (alignment) between the *x*, *y* and ξ, η. Including the P_5 and P_6 terms gives a full linear model. Parameters 7 and 8 represent a tilt of the focal plane with respect to the ideal tangential plane, and P_9 is the third-order optical distortion term, significant in most real telescopes. Once the parameters are estimated, a model function such as equation (9) gives ξ, η coordinates from measured *x*, *y* of any star with previously unknown position. Its spherical coordinates, α, δ are then obtained by solving

$$\cot\delta \sin(\alpha - \alpha_0) = \frac{\xi}{\sin\delta_0 + \eta \cos\delta_0}$$

$$\cot\delta \cos(\alpha - \alpha_0) = \frac{\cos\delta_0 - \eta \sin\delta_0}{\sin\delta_0 + \eta \cos\delta_0}$$

Advanced topics and reduction methods

Systematic errors in positions depending on magnitude (called the magnitude equation) are a common problem in many traditional astrometric techniques, particularly in photographic astrometry. Except for external comparisons these systematic errors can be determined by observing images of the same star with different exposure levels. Obviously this can be achieved by taking the same fields with different exposure times. Another option is to use a DIFFRACTION GRATING in front of the aperture (near the entrance pupil). Besides the zero-order central image, the grating produces much fainter, higher order images spaced symmetrically around the central image. Those images are spectra in principle; however, depending on the grating constant and bandpass they can be made almost circular. Instrumental magnitude equations can also be reduced by combining data taken with the instrument in normal and reversed positions. However, guiding error induced magnitude equations cannot be reduced with this technique. With linear detectors such as CCDs, magnitude equations in principle are no longer expected. Nevertheless, with non-perfect charge transfer efficiencies a coma-like systematic error can be significant in CCD data.

Block adjustment (BA) techniques in astrograph-type observing can achieve wide-angle (up to global) astrometric results using only differential narrow-field observations. Overlapping fields are required. Instead of solving for each individual field separately, a reduction scheme is used to solve for all plate parameters and all star positions simultaneously utilizing the fact that a given star has a unique position on all photographs or CCD frames it is on.

This technique was introduced in astronomy in 1960; however, it was not exploited until the 1970s because of a lack of computer power. The method is very sensitive to any systematic errors not modeled in the data and if not applied properly can produce worse results than a traditional reduction.

The sphere reconstruction process of the Hipparcos project actually is very similar to the BA. Because of the larger number of observations per star and the simultaneous observation of two widely separated fields the Hipparcos data are better conditioned for a global solution than overlapping plates.

State of the art accuracies and limitations

Currently the most accurate positions are obtained from VLBI radio observations for a couple of hundred compact, extragalactic sources. Positions for some sources are believed to be accurate to 0.1 mas. Source structure is a serious problem as well as atmospheric and ionospheric properties and the stability of ground-based stations (modeling of short-period effects). VLBI is also a very expensive technique to maintain and mainly is supported for Earth orientation work and geodesy, good astrometry being obtained as an aside.

Interferometric techniques in principle have a better potential for the highest positional accuracy than other methods. However, observing is slow, one object at a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

time. The highest angular resolutions and most accurate astrometry of close double stars are obtained by optical interferometers (NPOI, 70 μas).

Parallaxes on the 0.5 mas level are achieved for well-observed stars by traditional ground-based long-focus telescopes equipped with CCDs, perfoming as well as the best space-based instruments, with the FGS of HST (0.4 mas). However, the ground-based technique is at its theoretical limit set by the atmosphere and next-generation space missions are likely to improve on this by a factor of 10.

The Hipparcos catalog is currently the most accurate realization of an optical reference frame, giving positions of 118 000 stars on the 1 mas level. With a proper motion error of about 1 mas yr^{-1} and a mean epoch of 1991.25 the position errors for epoch 2000 already are of the order of 10 mas. The most accurate positions for the majority of the 9–12 mag stars are currently provided by the Tycho-2 catalog (20–70 mas). Ground-based wide-field astrographs are already capable of improving on this, getting 20 mas down to 14th magnitude (UCAC project). The limit here again is the atmosphere in combination with the available integration time owing to scheduled project duration and the limited field of view. A successful FAME space mission is likely to make future astrograph-type astrometric surveys obsolete, down to its limiting magnitude of 15. Going to fainter magnitudes will remain the domain of ground-based astrometric telescopes for at least a decade owing to their cost efficiency. The USNO A2.0 and the GSC-2 (Space Telescope Science Institute) are currently providing the only all-sky positions (approximately 0.2 arcsec) of faint stars down to 20th magnitude. These are derived from Schmidt survey plates, measured with the PMM. For even fainter stars only small selected areas have so far been observed.

Space-based interferometers such as SIM will push the astrometrical accuracy into the few microarcsecond range; however, only for few selected sources. Ultimately the achievable positional accuracy will depend on the signal-to-noise ratio, and thus the number of collected photons, and the success in controlling systematic errors, which will remain the most important technical challenge for astrometric projects.

Obtaining high-precision proper motions is problematic right now and utilizing decade- or century-old photographic data to their full capacity is an active task in current astrometry. This is likely to change in the future, when all magnitudes of old-epoch data will be covered by a successful next-generation astrometric space mission.

Bibliography

Good starting points for literature on astrometry are the following articles in the *Annual Review of Astronomy and Astrophysics*

Johnston K J and de Vegt Chr 1999 Reference frames in astronomy *Ann. Rev. Astron. Astrophys.* **37** pp 97–125

Monet D 1988 Recent advances in optical astrometry *Ann. Rev. Astron. Astrophys.* **26** pp 413–40

van Altena W F 1983 Astrometry *Annu. Rev. Astron. Astrophys.* **21** 131–64

For a recent, comprehensive overview textbook see

Kovalevsky J 1995 *Modern Astrometry* (New York: Springer)

Rigorous definitions of astrometric concepts and a detailed account of photographic astrometry and star catalogs in general can be found in

Eichhorn H 1974 *Astronomy of Star Positions* (New York: Ungar)

For a detailed account of classical instruments (TC, astrolabe, PZT) see

Mueller I I 1977 *Spherical and Practical Astronomy as Applied to Geodesy* (New York: Ungar)

For trigonometric parallaxes see

van Altena W F, Lee T J and Hoffleit E D 1995 *The General Catalogue of Trigonometric Parallaxes* 4th edn (Yale, CT: Yale University Observatory)

A comprehensive account on astrometric techniques is given in

Eichhorn H K and Leacock R J (ed) 1986 *Proc. IAU Symp. 109* (Dordrecht: Reidel) pp 1–838

A summary of the state of the art in the mid-1990s is given in

Hoeg E and Seidelmann K P (ed) 1995 *Proc. IAU Symp. 166* (Dordrecht: Reidel)

For the most recent progress in astrometry see

Johnston K *et al* 2000 Towards models and constants for sub-microarcsecond astrometry (*Proc. IAU Colloq. 180*) edited by K J Johnston, D D McCarthy, B J Luzum and G H Kaplan US Naval Observatory, Washington DC

Cox A N (ed) 2000 *Allen's Astrophysical Quantitities* 4th edn (New York: Springer) pp 677–87

For links to many project websites see

http://ad.usno.navy.mil/ucac/

Norbert Zacharias

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astron

French–Soviet ultraviolet astronomy satellite, launched March 1983. Carried an 80 cm telescope which operated in the range 150–350 nm. Placed in a highly elliptical orbit (1950 × 201 000 km) which limited observations to 5–6 hours per day. Observed many sources, including supernova 1987A. Ceased operations 1989.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Institute of Athens

The Astronomical Institute of Athens is the oldest research institute of modern Greece (it faces the Parthenon). The Astronomical Institute (AI) of the National Observatory of Athens (NOA) started its observational projects in 1847. The modern computer and research center are housed at the Penteli Astronomical Station with major projects and international collaborations focused on extragalactic astronomy.

Currently the 1.2 m telescope at Kryoneri (Peloponnese) is the observational facility of the AI. The AI's most prominent recent project is Aristarchos, the new 2.3 m Ritchey–Chrétien telescope build by Zeiss to be installed at Roania (2340 m altitude).

For further information see
http://www.astro.noa.gr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Observing Notebooks for Amateurs

An observational astronomer must, above all, be systematic. A faulty observation is worse than useless, it may easily confuse any subsequent analyses. Keeping careful, accurate records is essential. A story may here be relevant, although it may be apocryphal; it has been quoted time and time again. In 1781 William Herschel discovered the planet Uranus, and looking back on old records it was found that Uranus had been seen on several occasions, although it had always been mistaken as a star. According to the story, the French astronomer Le Monnier (see LE MONNIER) saw it no less than six times, and had he compared his observations he could hardly have failed to make the discovery even though Uranus was at that time near one of its stationary points. Le Monnier failed to check, and one of his observations was subsequently found scrawled on the back of a paper bag which had once contained hair perfume.

In any observation—visual, photographic or electronic—a few necessary notes must be attached. If any of these is omitted, the observation promptly loses much or the whole of its value. Basically, the main requirements are

name of observer
date
time, GMT (never use variants such as summer time in Britain, or time zones in other countries)
equipment used (for a telescope, type, aperture, magnification; for a camera, type, film, exposure time)
seeing conditions
plus any other data which seem to be relevant.

Of course, the type of notebook used depends entirely on the nature of the observations which are being made. The 'old-fashioned' planetary observer would keep a separate book for each object—one for Jupiter, one for Mars and so on. When making a sketch of a planet, or a lunar feature, much depends on the artistic skill of the observer. Some are able to make 'finished' drawings when actually at the telescope; others make preliminary sketches, go to a desk in order to make final copies and then re-check at the telescope. There can sometimes be a temptation to postpone making the final copies until the next day. This must always be resisted; errors are virtually certain to creep in.

Planetary observers often use a scale of seeing devised by E M ANTONIADI long ago. There are five grades, from 1 (perfect conditions of clarity and steadiness) down to 5 (conditions so poor that there would be no point in observing at all except to follow some phenomenon which will not recur). In practice, Antoniadi 1 is very seldom found, at least from countries such as Britain. When making a drawing of Jupiter or Mars, it is helpful to use a prepared disk to allow for the flattening in the case of Jupiter, or the phase in the case of Mars; this is not so easy with Saturn, where freehand sketching is really the only answer.

Photographs should be very carefully filed, with dates and times added as well as other necessary data. The same applies to electronic records. Many amateurs now hunt systematically for novae, comets, supernovae in external galaxies (see AMATEUR–PROFESSIONAL COLLABORATION IN ASTRONOMY), and with these records it is the time which is all-important, particularly if a discovery is made and the observer is anxious to claim priority.

All in all, the method of keeping an astronomical notebook is a matter of common sense. Never record anything of which you are not absolutely certain; if there is a trace of doubt, note it by the side of the observation. Remember, too, that it is only too easy to 'see' what one half expects to see. This was shown very clearly with observations of Mars before and after the first images were sent back by the space probes. Until then, many perfectly honest observers had shown an intricate network of canals. As soon as it became clear that the canals did not exist, they vanished from the observers' notebooks.

One extra point is worth making here. No observation is of much real use if it is left lying fallow in a notebook; if needed for analysis, it must be sent away to an organization such as the BRITISH ASTRONOMICAL ASSOCIATION or the AMERICAN ASSOCIATION OF VARIABLE STAR OBSERVERS. However, to send away one's original notebook is the height of folly; it will probably be mislaid somewhere or other, and perhaps lost in the post. Before sending away an observation, either a drawing or a written record, make absolutely certain that you have a perfect copy. True, this takes time, but it is well worth while.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Photography by Amateurs with Telescopes

The modern amateur astronomer is capable of taking beautiful pictures of the night sky, often with surprisingly humble equipment. One of the best ways to get started is to place a single-lens reflex (SLR) camera on a sturdy tripod, open the lens aperture as wide as possible and take a 10 min exposure (see WIDE-FIELD ASTROPHOTOGRAPHY). This will record the stars as trails, because of the rotation of the Earth, and is a good way to show the different colors of the stars. Shorter exposures (typically 1–2 min) will record the stars as points and are a good way of showing the different constellations. Transient events such as aurorae and bright comets can also be photographed in this way.

Unfortunately, modern SLR cameras are not really suitable for astrophotography as they rely on batteries to keep the shutter open. Astrophotographic exposures can be quite long (up to 1 h) and these will drain the camera's batteries very quickly. The best type of camera for this work is those which were manufactured in the late 1970s, such as Nikon, Pentax and Canon. These are rugged, manual cameras and can be purchased for reasonable prices from most second-hand camera shops.

The choice of lenses is also important. Wide-angle to standard lenses are excellent for producing constellation and Milky Way pictures. Small telephoto lenses (135 mm) are good for pictures of larger star clusters and deep sky objects as are long focal length telephotos (200–500 mm), but the latter can be used for eclipses and comets.

For long exposures it is necessary to place the camera on a driven equatorial mount. In this way the camera can track the motion of the sky and produce pinpoint images of the stars. Once the mount is accurately polar aligned it is possible to take exposures up to 30 min without having to make any drive corrections. A useful piece of equipment for this type of astrophotography is a 'ball-and-socket' joint that fits between the camera and mount and allows the camera to point in virtually any direction. Another piece of very useful equipment is a locking cable release to hold the shutter open for extended periods. Closing down the lens aperture by one f/stop (i.e. $f/2$ to $f/2.8$) dramatically improves the quality of star images, particularly around the edges of the frame.

Attaching the camera to a modern telescope allows detailed photography but means that the drive system has to be more accurate. The camera is usually attached by placing an adapter between the rear cell of the telescope focuser and the camera body. Good photographic targets include the Moon (see MOON, MOON MAPS) and this is probably the easiest form of telescopic photography as, because of the Moon's brightness, exposure times can be quite short. Capturing a sequence of the Moon's changing phases can be rewarding as can the eerie spectacle of a total lunar eclipse.

For all types of astronomical photography it is a good idea to bracket the exposures. This means that a series of varying exposure times around the likely correct exposure time are used. By doing this, a correct exposure will usually be attained, and, by noting the exposure time that gives the best result, a more educated guess can be made the next time a similar object is photographed under similar conditions.

The Sun (see SOHO (SOLAR AND HELIOSPHERIC OBSERVATORY) is also a good photographic target but great care must be taken when observing the Sun as even glancing at the Sun in the sky can damage the eyes. Care must also be taken with equipment as the heat from the Sun can damage lenses if they are pointed directly at the Sun for too long. The easiest and safest method of solar photography is to project an image of the Sun onto a piece of white card and photograph the image directly on the card. In this way it is possible to photograph sunspots on a daily basis. Full-aperture and off-axis solar filters give good views of sunspots. Advanced amateurs also use hydrogen alpha filters to photograph the Sun at a very specific wavelength, chosen to show the dramatic prominences that are visible around the edge of the Sun. Although expensive, these filters produce stunning results, especially when the Sun is in an active phase. The large amount of light given off by the Sun means that photographic exposures can be kept short.

To photograph the planets, which have very small angular diameters, an option is to use eyepiece projection. This involves placing an eyepiece between the telescope and camera to project a highly magnified image of the planet onto the film. Although capable of producing good results it is usually quite difficult because of the unsteadiness of the Earth' s atmosphere (poor seeing) and the difficulty of focusing on such a faint image.

Deep sky photography of galaxies and nebulae (see figures 1 and 2) is rewarding but requires very good tracking from a polar-aligned equatorial drive. Although most modern telescopes are equipped with accurate drives during longer exposures it is necessary to make minor corrections to the tracking of the telescope. Traditionally, this has been done in one of two ways. A separate telescope can be attached to the mount and used as a guidescope to monitor any deviations from perfect tracking performance. This method works well as bright stars which are suitable for guiding are usually easy to find. The guidescope needs to be fixed rigidly otherwise flexure can occur which will degrade the guiding. The second method involves the use of an off-axis guider, attached between the telescope and camera, which diverts the light from a guidestar using a small prism or mirror. These work well as the guiding is carried out through the main optics of the telescope, so no flexure

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Figure 1. The Pleiades, a young star cluster surrounded by a reflection nebula, photographed by Nik Szymanek.

Figure 2. The Lagoon Nebula, M8, a diffuse emission nebula in Sagittarius, photographed by Nik Szymanek.

occurs, but can be difficult in practice as the choice of suitable guidestars is severely limited. A modern option is to use a charge-coupled device autoguider (see CCDS, FILTERS AND CCD PHOTOMETRY FOR AMATEURS) with the guidescope as this will deliver the best tracking of all and alleviates the need for an observer to peer through a guiding telescope during long exposures.

Focusing on deep sky objects is quite demanding and is best achieved by using a moderately bright star and racking the focus back and forth until good focus is attained. If the camera used has the facility to change the focusing screen, a good option is to install a high-transmission screen such as those manufactured by Beattie. These feature very bright views and minimal clutter as no 'daytime' focusing aids (split image) are necessary.

Traditionally, fast films have been used for deep sky astrophotography (generally 400–1600 ISO), but one of the best films to emerge for this kind of photography is Kodak EliteChrome 200 ISO slide film. Several characteristics, such as fine grain and good red sensitivity, make it an excellent performer. Also, the digital manipulation of astrophotographs is very popular among modern astrophotographers. Film can be scanned and digitally enhanced to remove light pollution, aircraft and satellite trails, as well as contrast and color enhanced in what is effectively a digital darkroom.

Nik Szymanek

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Astronomical Societies and Publications

The first publicly sponsored, modern observatories came into being to serve the needs of navigation and time-keeping, and their first publications were, correspondingly, in the form of almanacs, ephemerides and the like. Early arrivals on the scene included *La Connoissance des Temps* (Paris Observatory, 1679) and *Nautical Almanacs* (Greenwich Observatory, 1767; San Fernando Observatory (Spain), 1792; US Naval Observatory, 1855). Producing calendric and navigational data appropriate to specific locations and languages was also a major function of the far more numerous observatories of the 19th century.

Professional societies (and publications) serve to validate a scientific community and its shared store of knowledge, customs and experiences. The first scientific society, however—the Royal Society of London, founded in 1660—was largely amateur. Its *Philosophical Transactions*, begun in 1665, for the first time facilitated the rapid dissemination of new scientific discoveries as well as providing a forum for their critique.

The first specifically astronomical society was the Astronomical Society of London (soon to become the ROYAL ASTRONOMICAL SOCIETY (RAS)), founded in 1820 by John Herschel and others (see HERSCHEL FAMILY) to supply a need that was not being satisfied by the Royal Society. Jewish names appear in the early membership lists, long before Jews were admitted to English universities, and a woman, Caroline Herschel, received a medal in 1828; the second was Vera Rubin, in 1996. Publication of the Society's *Memoirs* began in 1821, and they at first consisted of papers that had been presented in brief form at the monthly meetings. Soon it was necessary for the editors to point out that thanking a speaker for his presentation did not necessarily reflect unqualified approval or guarantee space in the *Memoirs*. A small committee reviewed contributions from time to time and recommended to the Council of the Society which ones should be published. Thus the *Memoirs* can claim to be both the first specifically astronomical publication and the first such publication to be refereed.

Memoirs, however, made no provision for less substantial items such as advance notices of meetings, short notes on astronomical phenomena and reports from observatories, and so in 1827 the Society instituted its *Monthly Notices*. This developed and eventually increased its frequency; it now appears thrice monthly and is a leading journal in the field. *Memoirs* became a vehicle for the publication of very long papers and compilations of data, and in 1977 it self-destructed in favor of microfiches in *Monthly Notices*. Early issues of both journals had occasional items in French but were monoglot for most of their histories.

The other astronomical publication to appear in the 1820s, *Astronomische Nachrichten*, was the fruit of the enthusiasm of a single man, Heinrich Christian Schumacher, director of the observatory at Altona from 1823 until his death in 1850. *Astronomische Nachrichten* continued to be edited at Altona, first by C A Peterson and then, from 1855, by C A F Peters. In 1872 the Altona observatory was moved to Kiel, and the journal was edited from there until the 1930s, when it was moved to Berlin. Schumacher offered to publish articles in 'any European language'; contributions in French and English were common from the start, and today nearly all the articles are in English.

Another publication that was the creation of a single person appeared in the USA: *Astronomical Journal*, founded by BENJAMIN APTHORP GOULD in 1849 in deliberate imitation of the German periodical. Its pages were open to professionals and amateurs alike, worldwide. Publication was interrupted by the Civil War, but Gould resumed it in 1886. *Astronomical Journal* is now published by the AMERICAN ASTRONOMICAL SOCIETY.

An oddity among astronomical publications is *The Observatory*, founded in 1877 and to this day owned collectively by its current editors. In addition to short research papers, it carries meeting reports, reviews and light-hearted items.

France produced a wide range of journals, beginning in 1884 with *Bulletin Astronomique*, published at the Paris Observatory but not limited to work done there. The Société Astronomique de France, and its *Bulletin*, were founded in 1887 (see SOCIÉTÉ FRANÇAISE D'ASTRONOMIE ET D'ASTROPHYSIQUE), and the *Journal des Observateurs* was founded in 1915. In 1969 this and the *Bulletin Astronomique* were merged into *Astronomy and Astrophysics*, together with *Annales d'Astrophysique*, *Zeitschrift für Astrophysik*, and the *Bulletin of the Astronomical Institute of the Netherlands*, which had been founded in 1921 by five Dutch observatories and institutes that had previously produced their own publication s

In 1863, German astronomers founded a society, ASTRONOMISCHE GESELLSCHAFT, in the context of the gradual unification of the country; it began its own publications soon after, and late in the century assumed responsibility for *Astronomische Nachrichten*. Other national societies founded in the 19th century were Società Astronomica Italiana (whose *Memorie* appeared from 1872), Société Royale Belge d'Astronomie (*Ciel et Terre* from 1880, *Bulletin* from 1895), the Royal Astronomical Society of Canada (founded in 1868) and the American Astronomical Society (1899).

Periodical literature in the late 19th and 20th centuries

The rise of 'astrophysics' in the later decades of the 19th century led to the involvement in the 'new' astronomy of scientists from a wide range of disciplines, and this sometimes led to strain between them and the 'old' (positional) astronomers. In Britain the RAS and its publications managed to absorb the new topics. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

American community partially fractured with the founding of the *Astrophysical Journal* (*ApJ*) in 1895 by George Ellery Hale and his associates. At first it mostly published spectroscopic results from Yerkes and, soon after, Mount Wilson Observatories, with some resistance from astronomers at other locations. It eventually spread over the entire country (and later the world) and embraced all aspects of astronomy, although there have been times when the editors thought it necessary to say explicitly that they would welcome papers in radio astronomy and astronomical instrumentation, invitations that not all found convincing. *ApJ*, originally owned by Yerkes Observatory, is now published by University of Chicago Press, but with management provided by the American Astronomical Society. It has been absolutely monoglot from the beginning.

The German and French communities also fragmented as they grew, with the founding of *Zeitschrift für Astrophysik* in 1930 and *Annales d'Astrophysique* in 1938, both originally monoglot but with numerous papers in English by the time they were absorbed into *Astronomy and Astrophysics* in 1969. No other national communities seem to have been large enough to support two professional publications, and astrophysicists elsewhere turned to academy proceedings and journals of physics in their home countries or, increasingly, to the *ApJ*.

Further fragmentation of the astronomical literature is a relatively late development, except for specialized publication in meteoritics (from 1935 onward in the USA and 1941 in the USSR) and selenography, both tied to specialized societies. Publications focusing on the Sun and solar system, and parallel divisions of the American Astronomical Society and some other professional societies, appeared after the Second World War.

In the second half of the 20th century there have been two main drivers of the establishment of new societies and publications. In the more advanced of the developing countries, and in those resulting from the dissolution of the USSR, Yugoslavia and Czechoslovakia, astronomers have often started their own, typically quite small, societies. Countries with still smaller populations of professional astronomers usually adhere to the INTERNATIONAL ASTRONOMICAL UNION via a national committee under the auspices of the national academy of sciences.

New journals have come largely from commercial publishers, sometimes with encouragement from members of the astronomical community who felt inadequately served by the older publications. None of these is currently a serious competitor, in the sense of being both large and highly regarded. In fact, as we shall see, the dominant trend has been towards a further concentration of important papers into the largest refereed journals, all of which are now published almost exclusively in English. The extent to which electronic publishing will modify current patterns remains to be seen.

Already in the 19th century, some astronomers perceived that the literature was becoming too voluminous for everyone to scan it all. An attempt at a *Bibliographica Astronomica* lasted only a year or two (around 1895). More successful was the *Astronomischer Jahresbericht*, begun by the Astronomische Gesellschaft in 1899, with the abstracts initially all in German (but article titles normally in the original language). It survived both world wars (although with incomplete coverage of the literature during the second) and in 1969 was transferred to a commercial publisher and English language text, as *Astronomy and Astrophysics Abstracts*.

Amateurs in astronomy

From its foundation, the Royal Astronomical Society welcomed both full-time and part-time, paid and volunteer, professional and amateur astronomers. However, as its activities and publications focused increasingly on the interests and products of the full-time, paid, professional sector, amateur astronomical societies inevitably came into being. Of those that survive, Leeds (1859) was the first and Liverpool (1881) the second, followed by the broader-based British Astronomical Association in 1890. The Association's *Journal* began publication the same year and still flourishes.

In the USA, amateur societies came first, Chicago (1863) and Baltimore (1881) pre-dating the founding in 1899 of the American Astronomical Society. The Astronomical Society of the Pacific and its *Publications* date from 1889. A 'pro/am' organization from its foundation (in connection with the coordination of observations of a solar eclipse visible from California), it has remained so, although eventually separate publication series became necessary, the *Publications* serving the paid professional and a series of *Leaflets* (1925–71) and, more recently, the journal *Mercury* being targeted at amateurs.

The AMERICAN ASSOCIATION OF VARIABLE STAR OBSERVERS, founded in 1911, is an amateur-dominated joint organization whose contributions to astronomical research are approached only by the somewhat similar Association Française des Observateurs d'Étoiles Variables (1921) and, in its day, by the Observing Corporation of the Society of Amateur Astronomers of Moscow, whose *Bulletin* (1925–1937) was later incorporated as one of the publications of the Academy of Sciences of the USSR.

International collaboration

International collaboration in astronomy began in the 18th century with cooperation over expeditions to observe transits of Venus, and in 1800 at Lilienthal there was organized a corps of 24 'celestial police' to monitor the ecliptic in search of a planet between Mars and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jupiter as predicted by Bode's law. The projects that resulted in the massive *Bonner Durchmusterung* (of stars), begun in 1859, and its southern extension, the *Cordoba Durchmusterung*, were largely parochial, but led to an agreement among 17 observatories to provide very accurate positions for the *Durchmusterung* stars down to ninth magnitude. However, work proceeded slowly and was overtaken by the advent of dry-emulsion photographic plates.

Successful photography of astronomical objects at many observatories led the director at Paris, Admiral E B Mouchez, with support from David Gill at the Cape of Good Hope, to attempt from 1885 the coordination of an international *Carte du Ciel*, a photographically based map of the sky and accompanying catalog of positions and magnitudes down to 11^m. The Permanent International Committee and its *Bulletin* (1892–1915) lasted until the onset of the First World War, holding regular conferences, encouraging progress among the 20 or so observatories and arranging for publication of portions of the *Carte* as material became available. The work was actually completed only in 1964, but the earlier plates are a valuable resource for determining proper motions over a long period.

Initially successful, but also a victim of the First World War, was the International Union for Co-operation in Solar Research (*Transactions* published from 1906 to 1914), whose prime mover was Hale. The Union attempted, for example, the standardization of measurements of solar rotation, so that real variations with latitude and with time could be identified. Last on the pre-war scene was Jacobus Kapteyn's scheme for Selected Areas (so that measurements of stellar brightness, color, proper motion, radial velocity and so forth could be concentrated in parts of the sky that would best reveal the large-scale structure of the Galaxy).

In 1919 representatives of 16 of the victorious nations met in Brussels to form the International Research Council (now International Council of Scientific Unions or ICSU), for which the cooperation of neutral nations was invited. The organizational meeting for the International Astronomical Union took place at the same time. An additional 16 nations were declared eligible for IRC membership over the next few years, an important constraint on the IAU, since no country could join it (or, initially, even send people to its meetings) without acquiring IRC membership first. The first official General Assembly of the IAU was held in Rome in 1922 and nine countries (Belgium, Canada, France, Great Britain, Greece, Italy, Japan, Mexico and the United States) participated. By the second Assembly the number had grown to 22. China and the Soviet Union were important additions between the world wars, but Germany, Austria, Hungary, Bulgaria and Turkey were still unwelcome until after 1945. The IAU eventually assumed responsibility for both the *Carte du Ciel* and the Selected Areas, although with some difficulty, since both had considerable German involvement. After the Second World War it took charge of the collection and dissemination (via telegram and, more recently, e-mail) of rapidly breaking astronomical news.

The IAU membership of some 150 astronomers at its foundation has now grown to about 8000 in 81 countries. From its beginning the IAU had its own series of publications, at first simply the reports of the triennial General Assemblies, but now including the proceedings of the specialized symposia and colloquia sponsored or endorsed by the IAU. The Union has become increasingly active in promoting astronomical education and research in developing countries, and it remains unique among the international scientific unions in having individual members as well as adhering countries.

Patterns of publication

A leading 19th century astronomer might well publish half or more of his papers in non-astronomical journals, and nearly all produced books recording their more extended efforts. A century later the astronomer or astrophysicist is likely to publish all his or her non-conference papers in one or another of the very few specialized, prestigious journals, and to produce monographs only of the textbook or trade book variety, if at all.

Changes over the first half of the 20th century are revealed by an analysis of the volumes of the *Astronomischer Jahresbericht* for 1900, 1925 and 1955 (these last dates being chosen to avoid distortion by the world wars). The first and most obvious trend is the growth of the field: not only do the numbers of papers increase, but their mean length has grown enormously. The second is the gradual shift between observation and theory, a shift that has doubtless continued in recent decades, although many of today's 'observations' (like the abundance of an element in a star derived from its spectrum) would have counted as theory in 1925 (and witchcraft in 1900). Third, the heyday of observatory publications came between the wars, when more than 200 existed, and they contained about one-sixth of the papers, compared with one-tenth in 1900 and in 1955; that fraction has now fallen much lower. The fourth trend is in the changes in the balance between nations, most notably in the increase in the contributions from the United States. Finally, and still more obviously if one looks at more recent times, is the gradual concentration of 'publishing power' in a few editorial hands. In 1997, 5272 pages appeared in the *Astronomical Journal*, 9136 in *Monthly Notices*, 17 141 in *Astronomy and Astrophysics* and its *Supplement* series and 22 948 in *Astrophysical Journal*, its *Letters* and *Supplements*. Nothing else (except perhaps conference proceedings and abstracts) remotely competes in size and level of quality as perceived by the astronomical community.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Blaauw A 1994 *History of the International Astronomical Union* (Dordrecht: Kluwer)

DeVorkin D H (ed) 1999 *The First Century of the American Astronomical Society* (New York: Institute of Physics)

Dreyer J L E and Turner H H (ed) 1923 *History of the Royal Astronomical Society 1820–1920* (London: Royal Astronomical Society)

Heck A 2001 *StarGuides* (Dordrecht: Kluwer); see also http://cdsweb.u-strasbg.fr/~heck/sgpres.htm

Lola Bausch J A 1983 1994 *Union List of Astronomy Serials* (Yerkes Observatory) (later editions in electronic form)

Virginia Trimble

Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Society of Australia

The Astronomical Society of Australia (ASA) was formed in 1966. It has a membership of professional astronomers and has as its objectives 'the promotion and extension of knowledge of Astronomy and related branches of science'. The business of the ASA is conducted by an elected Council. It is represented on a number of astronomical bodies such as the National Committee for Astronomy. The Society publishes an international journal, *Publications of the Astronomical Society of Australia*, and it holds an annual four day scientific meeting.

For further information see http://www.atnf.csiro.au/asa_www/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Society of the Pacific

The Astronomical Society of the Pacific (ASP) has worked since 1890 to advance astronomy, and to explain the universe to students, teachers and the public. It is the world's largest general astronomy organization. Its annual meetings bring together professional and amateur astronomers, historians, teachers, students and the public. It publishes a respected professional journal *Publications of the ASP*, the ASP Conference Series containing the proceedings of over 200 astronomy conferences, and the non-technical magazine *Mercury*. It is also the publisher for the INTERNATIONAL ASTRONOMICAL UNION. Its quarterly teachers' newsletter *The Universe in the Classroom* reaches thousands of classrooms, and is translated into a dozen languages. It also produces a wide range of high-quality, low-cost teaching resources, including books and slides. It co-manages the education and public outreach program for NASA's SOFIA airborne observatory. Its exemplary Project ASTRO develops productive long-term partnerships between astronomers and teachers across North America.

For further information see
http://www.aspsky.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Spectroscopy

As a passive science, astronomy extracts most of its information from the incoming RADIATION out of the heavens. Such an analysis would be all too simple were it not for the quantum nature of matter and radiation. Electrons are bound in quantized orbits around their nuclei and must emit radiation at discrete wavelengths which are defined precisely by the transitions between discrete energy states. The resulting spectral lines are extremely important tools for identifying and characterizing the emitting source.

Spectral signatures

The first concept of spectroscopy can be traced to Newton, who first noted that the dispersion of sunlight into its different colors by a prism reflected the properties of sunlight itself. It is furthermore remarkable that some 200 yr later, in the late 1800s, when Balmer recognized the hydrogen series in the absorption lines of the solar spectrum first seen by Fraunhofer, it pre-dated the discovery of the electron, the quantum theory or the Bohr atom. This hint from the heavens, that matter will absorb light in a very well-behaved manner, with a SPECTRUM which can be modeled as the hydrogen series and which can be applied to other atomic species, motivated the development of modern physics. Today, we know of course that the quantized energy states of an electron orbiting an atom depend on the mass of the nucleus and the nuclear charge. Each isotope of each atomic species therefore has a distinctive set of electronic levels. Hence the resulting transitions form a unique and unambiguous signature for identification.

Measuring large-scale motions

Perhaps the most important feature of the atomic spectra is the rest frequency or wavelength of the transitions. These are very precisely known, either through quantum mechanical calculations or laboratory measurements, to typical accuracies of better than one part in a billion. This high accuracy allows the measurement of any shift in the rest wavelength due to the motion of the emitting source. The Doppler shifts, measured through precise spectroscopy, have been the cornerstone of modern astronomy. Such diverse phenomena as binary stellar orbital motions, rotation curves of galaxies and radial motions around massive black holes represent a few examples of the important kinematic motions revealed through spectroscopy. Perhaps the most important measurement of all was the recognition of the Hubble flow of galaxies. The systematic REDSHIFT of the spectra of galaxies, the recognition of their cosmological nature and the ever-increasing magnitude of motions with distance are the foundation of the model of the expanding universe. This result has the enormous implication that the light which we observe at different redshifts comes from different distances. Because of the finite speed of light, this means that we have the means to examine the history of the universe as a function of time, by sampling only in our own epoch. In the past decade, the redshifted Doppler motions of galaxies have been used to map the spatial distribution of matter on very large scales, suggesting the formation of structures in the very early universe. The study of galaxies and QUASARS at very high redshifts has become one of the fundamental problems of modern astrophysics.

Measuring local motions

Because of the discrete nature of the electronic energy states, the ensuing transitions not only have precisely determined transition frequency but also intrinsically narrow line widths, typically dominated by Doppler motions within the emitting systems. When spectrometers became powerful enough and the detectors became sensitive enough, line profiles began to be resolved. The measured spectral profiles were found to carry important information on the local kinematics. For example, extraordinarily broad line profiles seen toward SEYFERT GALAXIES were interpreted as outflowing material from the nuclear regions, since from the magnitude of the motions they do not appear to be gravitationally bound. P Cygni line profiles in stellar envelopes, with redshifted emission and blueshifted absorption, were taken to be evidence of outflowing winds. This is because of the idea that the material in front of the star can be identified by its absorption against the star while its spectra can be seen to be blueshifted by its outflowing motions towards us. By the same argument, inverse P Cygni profiles, where the motions of the emission and absorption spectra are reversed, have been interpreted as contraction. In yet other sources, asymmetric line profiles with substantial enhancements of one side of the line profile were kinematic evidence for shocked gas. This became even more convincing when combined with excitation arguments. In nearby systems, the observed Doppler motions in the radial direction have been found to be consistent with detectable PROPER MOTIONS across the line of sight. This is consistent with our expectation that the Earth is not an unique vantage point. After systematic motions are removed, the spectra often still exhibit nonthermal motions. That is to say, the line profiles are broader than can be explained by thermal motions alone. This has led to the studies of turbulence, especially with regard to the role of magnetic fields.

Measuring excitation

The various energy levels within the atom will be populated according to the ambient conditions such as temperature, particle density and radiation field. By sampling an appropriate set of transitions, we can construct the population distribution and thereby measure the excitation of the emitting source. In particular, it is easy to see that successively higher ionization states will require higher excitation, since the electrons will see a less shielded nucleus (see EXCITATION AND IONIZATION). Hence, if the detected radiation is dominated by lines from highly ionized atoms, we can deduce that the emitting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

region must be ionized by energetic photons. The relative strengths of the various atomic lines are therefore a means to identify the spectral type of a star, in addition to its colors (see also CLASSIFICATION OF STELLAR SPECTRA). The relative line intensities also determine the hardness of the radiation. This has allowed us to distinguish between normal galactic nuclei, whose ionization can be easily explained by the radiation from young stars, and LINERS and Seyferts whose radiation field must be progressively harder. Another example of the power of spectroscopy was the identification of HERBIG–HARO objects as regions of shock excitation. Whereas these sources were once thought to be sites of young stars, spectroscopy showed otherwise. Together with proper motions and Doppler velocity measurements and line profiles, the relative intensities of the various ionized lines were consistent with shock models.

Measuring abundances

The detection of numerous atomic transitions allows the assessment of the abundances of the elements. With improved sensitivity, the abundances can be measured not only in the Sun and other stars but also in clouds and in external galaxies. Such data are crucial for checking NUCLEOSYNTHESIS models, for understanding the metal abundances as a function of time and as a gauge on evolution. For example, POPULATION I and POPULATION II stars can be distinguished based on their metallicity. Whether HIGH-VELOCITY CLOUDS in the Galaxy are primordial in nature or recycled from the disk of the Galaxy can also be determined by measuring the relative abundances of metals. Detailed measurements of the various elements in SUPERNOVA ejecta can also be used to classify the type of supernova event and also to check on elemental production mechanisms.

Important atomic lines

There are a number of atomic lines which are of particular interest. These include the Balmer lines in the optical, the Brackett lines in the near infrared, the radio recombination lines, the 21 cm atomic hydrogen line and the ionized carbon line in the far infrared. In general, the spectral lines from ionized hydrogen are important for tracing H II REGIONS. In particular, the infrared and radio lines are useful for studying deeply embedded sources.

Optical lines

The Lyman series of hydrogen, which consists of the recombination lines which end in the $n = 1$ state (where n is the total quantum number for Bohr–Sommerfeld orbits), is found at ultraviolet wavelengths. The Balmer series, the lines of which end in the $n = 2$ state, is found in the optical. The hydrogen lines together with lines from other elements have been used to classify the spectral types of stars. The temperature in the photosphere, as inferred from the type of lines which are excited, is an important part of the HERTZSPRUNG–RUSSELL DIAGRAM for stellar evolution. The Hα line at 6563 Å, which is the $n = 3$–2 Balmer line, has been one of the most useful lines in optical work, tracing regions which have been photoionized by a nearby star. This is particularly important in external galaxies because the nuclear regions are often highly ionized by a concentration of massive stars or an active nucleus powered by a BLACK HOLE. The bright nuclear emission in the Hα line can be used to measure the ionizing luminosity, the absolute motion of the gas from its redshift, which in turn yields a distance via the Hubble relation, and perhaps most importantly the nature of the nucleus from the observed line widths. The advent of SPECTROGRAPHS with high resolving power allows the astronomer to distinguish between narrow line widths consistent with galactic rotation and much wider line widths which may represent outflowing winds from the nuclear region. In the case of active nuclei powered by nonthermal processes, the radiation has a harder spectrum than ordinary starlight. Hence, lines from higher excitation states of hydrogen or rarer elements become detectable and even stronger sometimes. In the past decade, optical spectroscopy has made great progress because of the construction of large-aperture instruments such as the Keck telescopes and the launch of the Hubble telescope. The larger collecting area has meant that ever higher-redshift systems are being detected. The Hubble telescope with its superb imaging capability above the atmosphere has meant that spectroscopy can be done at under 0″.1 resolution. Some of the best evidence for the existence of massive black holes in the nuclei of galaxies has come from the Hubble measurements of very large rotational motions in the central few tenths of an arcsecond (see figure 1). For stellar work, one of the most important advances has been the development of spectrographs with tremendous resolving powers. Eschelle spectrographs are now able to resolve a line with a resolution of better than 100 000. By fitting many spectral lines at once, Doppler motions on the order of a few meters per second can be detected. This has allowed the development of the study of extrasolar planets, where planets in orbit around nearby stars are inferred from the reciprocal motions of the stars due to the gravitational pull from a secondary body. Some 20 planets have now been inferred from the Doppler motions of these wobbling stars (see EXOPLANETS).

Ultraviolet lines

Because the Lyman series lies in the ultraviolet, little progress was made until space-borne telescopes became available. With the Goddard High-Resolution Spectrograph on the Hubble Telescope, absorption lines can be measured for many atomic and molecular species. This has allowed the measurement of elemental abundances towards Galactic and extragalactic sources. In particular, metal abundances are important for identifying whether material is primordial in nature or has already been processed through star formation. Elemental abundances in interstellar clouds as measured in the gas phase can be compared with cosmic abundances in order to deduce the fraction of various elements which are locked up in dust

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

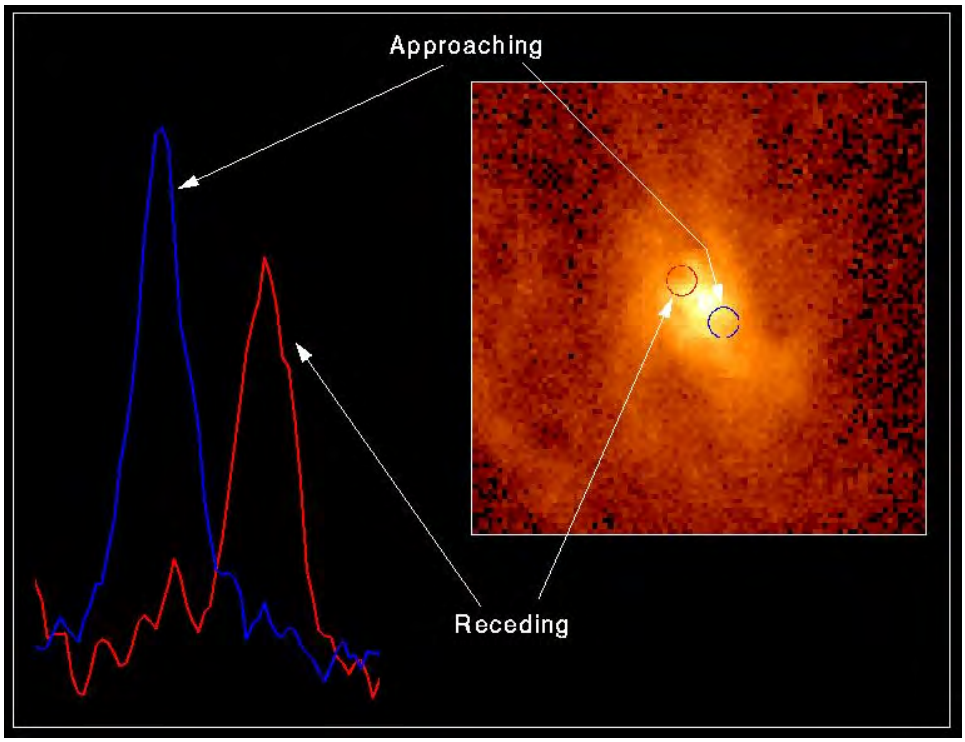


Figure 1. An image of the ionized disk surrounding the nucleus of M87 from the Hubble Space Telescope. Shown are two spectra of the O III emission line at 0″.25 offset on either side of the nucleus. The two spectra are redshifted and blueshifted by about 500 km s^{-1} over a size scale of about 35 pc. The implied enclosed binding mass is about 2 billion solar masses. Such motions, if bound, constitute some of the best evidence for massive black holes in the centers of galaxies.

grains. The measurement of the Lyman lines from deuterium can be compared with the corresponding hydrogen lines in order to estimate the deuterium fraction. This is of course a check on the synthesis of elements in the early universe since deuterium is not made through stellar nucleosynthesis. The deuterium fraction depends on the primordial baryon density during the first few minutes of the big bang. A higher baryon density would drive a faster rate for the formation of helium 4, which would tend to deplete the deuterium. The Ly α lines are also studied in absorption against quasars. Since there is a lot of intervening hydrogen gas on the way to distant quasars, all the Lyman absorption lines form a band of absorption on the blue side of the Lyman line which is intrinsic to the quasar system. This so-called LYMAN ALPHA FOREST has become an important tool to probe the intergalactic medium.

Infrared lines

While the Hα line and other optical lines have been very useful, they are susceptible to dust extinction effects. This is particularly bothersome for cores of molecular clouds where young stars are deeply embedded or for edge-on nuclear disks where the line-of-sight extinction can be many magnitudes in the optical wavelengths. One could compare the Hα and Hβ line intensities to estimate extinction effects. However, in practice, this is unreliable if both lines are suffering heavy extinction. Moreover, if the extinction is high enough, the Hα line may not be detectable at all. In the near infrared, the extinction due to dust is much reduced. As a result, the Brackett lines of hydrogen at 2 μm and 4 μm have been very useful for studying embedded sources both in the Galaxy and in external systems. In the mid-infrared, there are also fine-structure lines due to the spin–orbit interactions of the electrons, where there is an energy difference depending on whether the spin of the electron is aligned with the orbital angular momentum vector of the electron. For example, a particularly useful transition is due to the $^2P_{3/2}$–$^2P_{1/2}$ line for Ne II at 12.8 μm. This line has been successfully used to study H II regions, planetary nebulae, extragalactic nuclei and most famously the ionized gas streamers around the Galactic Center (see figure 2). Not only is the ionized gas detected to be in the form of streamers which orbit the Galactic Center but the kinematics is determined from the Ne II line profiles. Orbits fitted to the motions of the ionized gas suggest a central massive object with about three million solar masses.

With the development of more and more sensitive detectors as well as space-borne telescopes, the near-infrared window is becoming a very important diagnostic for external galaxies. The Infrared Space Observatory was able to detect and measure many atomic lines in the near

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

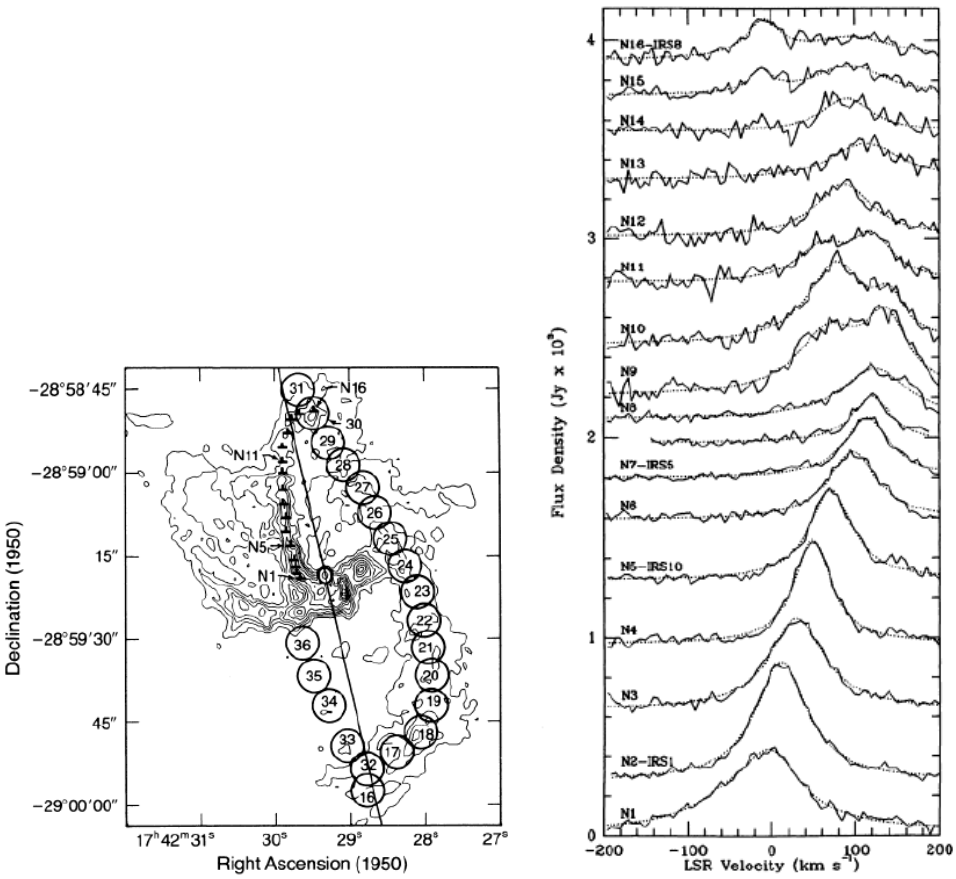


Figure 2. The left image is a picture of the ionized streamers detected in radio continuum toward the Galactic Center region. Overlaid are the positions where the Ne II 12.8 μm fine structure line emission has been measured. A number of ionized gas streamers can be seen to be in orbit around the central object known as SgrA*, a point-like radio continuum source. The right image is a display of the spectral profiles measured in the northern ionized arm. The kinematics of the Ne II emission has been fitted to orbital motions around SgrA*. This resulted in an estimate of $3 \times 10^6 M_\odot$ for the central mass if the motions are bound. These results are some of the best evidence for a massive black hole in the center of the Galaxy.

and mid-infrared toward many systems. These provide information on the chemical abundances and excitation (see figure 3) as well as on motions for these galaxies. Such information has been very useful for understanding whether the infrared-luminous galaxies can be driven by star formation processes alone.

Neutral atomic hydrogen

As 90% of the universe is in the form of hydrogen, it is clearly the most important atom. While the original identification of the Balmer series was necessarily in the optical, the Lyman series is in the ultraviolet, while the Paschen, Brackett and later series are in the infrared. All of these lines require a good deal of energy to excite, making them good tracers in stellar environments and in surrounding ionized regions. For the lower excitation regions, there is the 21 cm spin-flip line. This line originates in the interaction between the spins of the electron and the proton. The hydrogen atom has slightly more energy when the spins are parallel versus anti-parallel. The resulting photon on de-excitation comes out at a wavelength of 21 cm. The required excitation is so small that this line can be detected throughout the Galaxy and easily in external galaxies. Moreover, unlike optical lines, the 21 cm line is impervious to extinction effects. Hence the kinematic information from the line profiles can be traced throughout the entire Galaxy even though we are viewing the system edge on and within the disk. One of the first 21 cm results was the identification of a number of SPIRAL ARMS within the Galaxy, heretofore unseen. This 21 cm line has become a most important tool for studying galactic dynamics, interacting galaxies, cluster environment and redshifts in the nearby universe. One of the important features of the hydrogen line is that it can exist where stellar matter does not. Hence far out in the Galaxy, where matter is no longer dense enough to make stars, we can still trace the INTERSTELLAR MATTER. Studies at 21 cm have shown that often galaxies are several times larger in atomic hydrogen than their optical extents.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

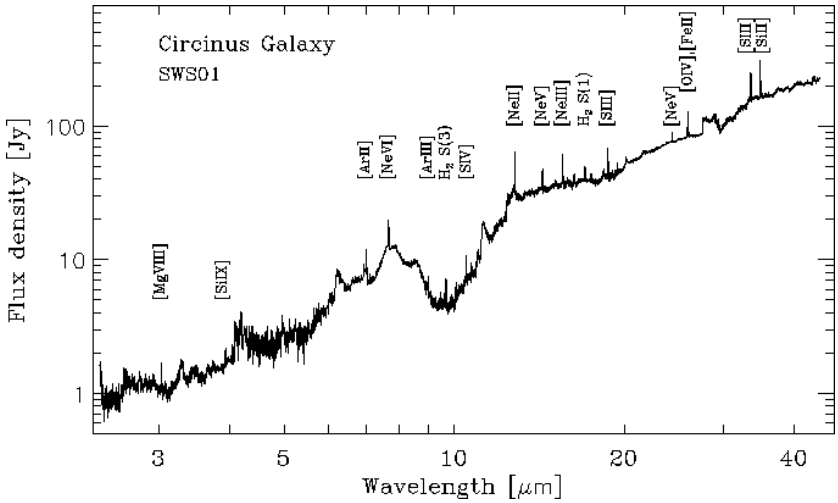


Figure 3. An infrared spectrum of the Circinus galaxy from the Infrared Space Observatory. A number of spectral features from atomic ions can be seen. The comparison of the relative intensities of these ionic features tells us the excitation state of the gas.

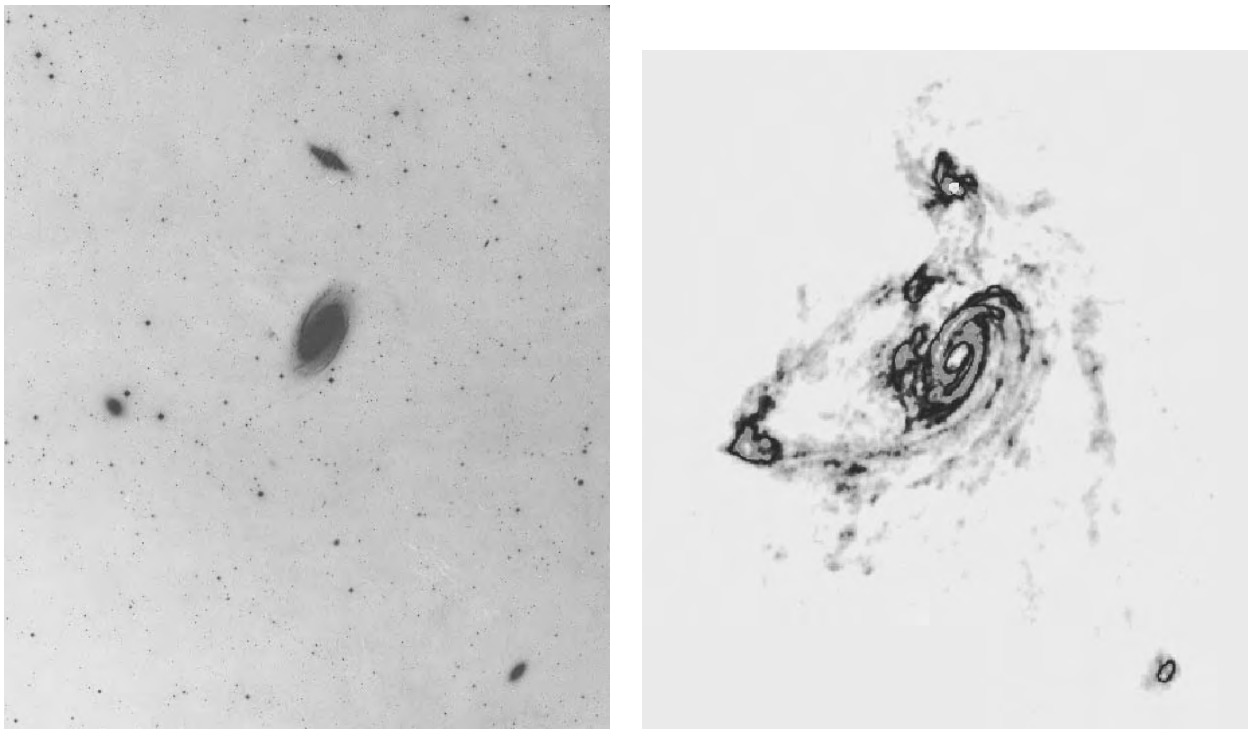


Figure 4. The left panel shows an optical picture of the M81–M82–NGC2077 group. The right panel shows an atomic hydrogen 21 cm line image of the same group of galaxies. The atomic hydrogen clearly shows a lot of material which is not seen in the optical and in particular shows that interactions among the three members of the group have resulted in tidal distortion of the gaseous disks of these galaxies.

There are two tremendously important effects which have been found. First, in SPIRAL GALAXIES, where we can see the rotational motions of the stars and the gas, it has been found that the rotation curve of the galaxy, when traced in atomic hydrogen, appears to be flat as far out as we can detect the hydrogen. The implication is that there is a lot of matter which is enclosed in the outer parts of the galaxy. Otherwise, the rotation curve will fall with increasing radius because of the decreasing gravitational effects of the enclosed mass with increasing distance. This is one of the key results suggesting the existence of DARK MATTER. If sufficient dark matter is there, it may have implications

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for the question of whether the universe is open or closed, i.e. whether there is enough matter to bound the Hubble expansion of the universe. The second important result has been the detection of peculiar motions and structures in the outer parts of galaxies which can be attributed to interactions with nearby systems (see figure 4). Such interactions are typically not seen in the optical which traces the inner parts of the galaxy. This has implications for the dynamic nature of galaxy formation and evolution, as well as for the formation and fueling of massive black holes especially in interacting systems.

Radio and far-infrared lines

For hydrogen, the higher recombination lines, those ending in the higher n states, occur in the radio wavelengths. The main advantage of these lines is the absence of dust extinction at radio wavelengths, making the intensity of these lines a very accurate measure of the total ionizing luminosity. The main drawback for radio studies, the coarse angular resolution, was overcome by the development of RADIO INTERFEROMETERS. While the recombination lines at centimeter wavelengths can suffer from nonthermal excitation effects, the lines in the millimeter wavelengths in general are free from such effects and have line strengths nearly equal to the radio continuum emission from the H II regions. In the far infrared, there are a number of important lines. One of the interesting lines is the C II $^2P_{3/2}$–$^2P_{1/2}$ fine-structure line at 158 μm. This line is excited by collisions with H I and electrons in photodissociation and photon-dominated regions. It is an important coolant and has been estimated to carry as much as 0.5% of the total infrared luminosity of a galaxy. This C II line can therefore be an important tool for studying high-redshift systems where other lines may be too faint to detect.

Molecular lines

While a few molecules such as CN were detected via optical absorption lines, the floodgates to molecular astronomy did not open until the late 1960s. This was partially due to the necessary wait for the development of detector technology. However, there was also great skepticism that molecules could exist at all in the interstellar medium. This pessimism was based in part on the low gas density seen associated with atomic hydrogen clouds which implied that molecules would be dissociated by the ambient radiation fields. In hindsight, we can now assert that since stars form from gas, the gas must pass through a dense phase which is substantially denser than the atomic phase. Such a dense phase would of course provide shielding against interstellar radiation and would allow the formation of molecules. However, back in the 1960s, none of the theoretical underpinnings on STAR FORMATION, INTERSTELLAR CHEMISTRY or INTERSTELLAR MOLECULAR CLOUD evolution was in place. Persistence by a few radio astronomers led to the discovery of OH, H_2O, NH_3, CO and many other molecules.

Molecules

While atomic transitions typically fall within the optical wavelengths or the near infrared, molecular transitions are typically at much longer wavelengths. This is due to the fact that, for molecules, rotational and vibrational motions of the atoms within the molecule are also possible. These provide new sets of quantized energy states. Transitions between rotational states typically fall into the radio wavelengths, and vibrational transitions are typically in the infrared. While atomic transitions are characterized by temperatures on the order of thousands of kelvins, molecular transitions require temperatures of tens to hundreds of kelvins. The molecules are therefore good probes for the bulk of the universe which is cool with temperatures near 10 K. Since the early detection of the key molecules of H_2O and NH_3 in the late 1960s, and of CO in the early 1970s, over 110 molecular species have now been identified in the interstellar medium. The availability of molecular transitions with different excitation requirements allows the study of the interstellar medium under a wide combination of temperatures and densities. In particular, transitions with high dipole moments have been used successfully to study the very dense regions of molecular clouds which are forming stars. With the tremendous improvement of detector technology at radio wavelengths, and the construction of large-aperture telescopes, molecules are now easily detected in external galaxies. Perhaps the most important results have been the discovery and realization that molecular clouds are a dominant constituent of the interstellar medium, that molecular lines along with dust are important coolants for the gas phase and that molecules can trace the mass, the temperature, the density and the motions of the interstellar medium.

Instrumentation developments

Progress in the field of spectroscopy of atoms and molecules in astronomy has come about through a number of technological innovations. Grating spectroscopy with finer and finer resolution has been developed. Whereas long-slit spectroscopy can obtain spectra for a slice of the sky, the trend today is towards multifiber spectroscopy. In this scheme, hundreds of optical fibers are positioned at once on predetermined target sources in the focal plane. The fibers are then fed to a spectrometer in order to measure the spectra for hundreds of sources simultaneously. At radio wavelengths, interferometry with spectroscopy is even more powerful in that the entire primary beam of one of the telescope elements of the array is imaged at once. With the development of powerful digital CORRELATORS, large spectral bandwidths can be analyzed for every pixel in the primary beam. Compensation for atmospheric effects is also making good progress. Space-borne platforms and high-flying aircraft and balloons have opened up new spectral windows. ADAPTIVE OPTICS at optical wavelengths and self-calibration techniques at radio wavelengths have successfully compensated for atmospheric SEEING effects.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This has meant that spectroscopy of fainter sources is possible. Of course, detector technology has improved for all the wavelengths, while larger and larger telescopes are being built. These efforts have made tremendous improvements in sensitivity.

Bibliography

To obtain a deeper level of details on this topic, the readers are referred to the *Annual Review of Astronomy and Astrophysics*. Some articles of interest are

Savage B D and Semback K R 1996 *Ann. Rev. Astron. Astrophys.* **34** 279

on interstellar abundances and

Rauch M 1998 *Ann. Rev. Astron. Astrophys.* **36** 267

on the Lyman α forest in QSOs. There are many books on atomic spectroscopy:

White H E 1934 *Introduction to Atomic Spectra* (New York: McGraw-Hill)

Paul Ho

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical Unit

The semi-major axis of the elliptical orbit of the Earth. The astronomical unit (abbreviation: AU), is also described as the 'mean' distance between the Earth and the Sun, where 'mean' in this context refers to the average of the maximum (aphelion) and minimum (perihelion) distances between the Earth and the Sun. The currently accepted value for the astronomical unit is 149 597 870 km.

An alternative definition is the radius of the circular orbit in which a body of negligible mass, if free of perturbations, would revolve round the Sun in a period of $2\pi/k$ days, where k is the Gaussian gravitational constant (=0.017 202 098 95).

See also: ellipse, orbit, orbital elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomische Gesellschaft

The AG was founded in 1863 in Leipzig as an international society dedicated to the 'advancement of science by means of supporting projects which require systematic cooperations of many people'.

The primary activities of the AG are the organization of scientific meetings and conferences, the promotion of young astronomers, awarding of medals and prizes in recognition of outstanding scientific works, awarding of honorary membership, the publication of scientific literature, the fostering of public relations, and the promotion of astronomical education at schools and of the general public.

At present the AG has more than 800 members.

For further information see
http://www.astro.uni-jena.de/Astron_Ges/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomische Nederlands Satelliet (ANS)

First Dutch satellite, launched in August 1974. Operated until 1977. Carried x-ray and ultraviolet instruments. Discovered x-rays from cool dwarf (UV Ceti) stars and the first x-ray burster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomisches Institut der Universität Würzburg

Part of the ancient university of Würzburg in Bavaria, southern Germany. It specialized in stellar dynamics, star formation studies and solar physics. It is in the process of reorganization.

For further information see http://www.astro.uni-wuerzburg.de/homepage.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomy

The science of the universe, its constituent bodies and phenomena. Arguably the oldest of the sciences, its beginnings go back to the dawn of recorded history, to the time when mankind first began seriously to study the apparent motions of the Sun, Moon and planets and to identify patterns and cycles in their behavior. Astronomy found practical application in providing the basis for calendars, timekeeping and navigation, but is now essentially a pure science.

Astronomy is primarily an observational science rather than an experimental one. Whereas in physics or chemistry, for example, it is possible to set up experiments under controlled conditions, to change those conditions and measure the outcomes, in astronomy, apart from using satellites and spacecraft to explore our immediate locality, it is not possible to carry out experiments on stars and galaxies. Instead, astronomers mainly rely on collecting and interpreting information that arrives from astronomical sources in the form of electromagnetic radiation and particles in order to formulate and test hypotheses and theories. Ground-based astronomy is hampered by the atmosphere, which—even under ideal conditions—transmits to ground level only a small fraction of the complete range of electromagnetic radiations (the electromagnetic spectrum). Until the middle of the twentieth century, astronomy was carried out almost exclusively at visible (optical) wavelengths. Since then, ground-based astronomers have extended their observations into the radio and infrared regions of the spectrum (only limited quantities of infrared radiation reach ground level) and to the detection of particle radiations (neutrinos and cosmic rays). During the past four decades or so, with advances in space technology, astronomers have been able to explore planets and the interplanetary medium directly with the aid of spacecraft and, most significantly, to place instruments in orbit above the atmosphere, thereby gaining access to virtually the entire electromagnetic spectrum from gamma rays to radio waves.

In a few cases, astronomers can handle samples of cosmic material—neutrinos, meteorites, interstellar particles, Moon material, soon (possibly) dark matter astroparticles. Gravitational waves are also in the future of astronomy.

The fundamental types of observation that astronomers can carry out include astrometry (the measurement of positions and motions), photometry (the measurement of brightness and brightness changes), imaging and spectroscopy (the detailed analysis of the spectrum of radiation arriving from an astronomical source).

See also: astrology, astrometry, astrophysics, electromagnetic radiation, electromagnetic spectrum, gamma-ray astronomy, infrared astronomy, optical astronomy, neutrino astronomy, radioastronomy, ultraviolet astronomy, x-ray astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomy and Astrophysics in India

Jayant Narlikar

The growth in astronomy and astrophysics (A&A) in India has been mostly since the country achieved independence in 1947. The present work is carried out in a few select research institutes and in some university departments. The Astronomical Society of India has around 300 working A&A scientists as members, with another 50–60 graduate students.

Facilities

In the institutional sector, the major facilities include the Vainu Bappu Optical Telescope of 2.3 m diameter and a 1 m telescope at Kavalur in South India, under the management of the Indian Institute of Astrophysics (IIA) at Bangalore, a 1 m telescope at the Uttar Pradesh State Observatory (UPSO) at Naini Tal in the foothills of the Himalayas and a 1 m infrared telescope at Gurushikhar in southern Rajasthan under the management of the Physical Research Laboratory (PRL), Ahmedabad. Solar telescopes exist in Kodaikanal in southern India, at the UPSO in Naini Tal and in Udaipur, the first being managed by the IIA and the third by PRL.

In the 1960s radio astronomy entered in a big way with a 550 m parabolic cylindrical antenna being set up on a north–south hill slope near the southern hill resort at Ooty, the axis of the cylinder being parallel to that of the Earth. This telescope is operated by the TATA INSTITUTE OF FUNDAMENTAL RESEARCH (TIFR), in Mumbai, which now has just completed an even more ambitious project known as the Giant Metrewave Radio Telescope (GMRT) at Khodad, about 90 km from Pune. Controlled and operated by its NATIONAL CENTRE FOR RADIO ASTROPHYSICS (NCRA) at Pune, this system has 30 antennas, each of 45 m diameter, spread over a Y- shaped array of arms about 14 km long. It works optimally at meter wavelengths. At lower frequencies (34.5 MHz), there is a T-shaped radio array at Gauribidnur near Bangalore operated jointly by the IIA and the Raman Research Institute (RRI), Bangalore. Indian astronomers have erected a similar array in the island of Mauritius. The RRI also has a 10 m dish for millimeter wave astronomy at its headquarters in Bangalore.

Space astronomy in India began with cosmic ray research under H J Bhabha at TIFR and later under V A Sarabhai at PRL. A national balloon facility at Hyderabad caters to the various balloon borne experiments. This activity blossomed into a fully fledged space program which now routinely launches satellites and prepares payloads of a commercial as well as a scientific nature. Its headquarters are at the Indian Space Research Organization at Bangalore.

In the university sector, the only significant working facility is the 1.2 m telescope at the Japal Rangapur Observatory under the control of Osmania University, Hyderabad. A major resource facility for the universities was, however, set up in 1988 by the University Grants Commission (UGC), New Delhi, at Pune. Known as the Inter-University Centre for Astronomy and Astrophysics (IUCAA), this center provides an excellent resource library in A&A, an advanced computer centre with A&A-related software, a data centre and an instrumentation laboratory. These facilities can be used by visiting academics and students from universities with their travel funding met by grants provided by the UGC at IUCAA. The IUCAA also assists university academics in using observing facilities in India and abroad under the various guest observing programs.

There are three new optical telescopes under various stages of construction. The IUCAA will have a 2 m telescope near Giravali, about 70 km from Pune, whereas the IIA has recently built a 2 m telescope at Han Le, at a height of 4000 m near Ladakh in the northern Himalayas. The UPSO and TIFR are jointly planning to set up a 3 m telescope in Devasthal, near Naini Tal.

Research areas

INDIAN ASTRONOMY has traditionally been working in stellar astronomy and spectroscopy, solar and planetary astronomy. However, in the last three decades of the 20th century it has expanded its range to include solar physics including solar oscillations, pulsar astronomy and modelling, problems in cosmic ray and high-energy astrophysics, gravitational lensing, extragalactic astronomy and cosmology. The Ooty Telescope was responsible for a major survey of angular sizes of radio sources through the lunar occultation method. Theoretical cosmology has ranged from abstract mathetical models in general relativity (for India has always had a strong school in general relativity) to fitting models to observations of discrete extragalactic source populations. Thanks to a less rigid outlook towards alternatives to mainstream ideas, these have flourished in India more than they have in the West.

Indian astronomy has made its presence felt internationally. The late Vainu BAPPU was President of the International Astronomical Union, while several Indian astronomers have occupied positions of Presidents of various IAU Commissions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysical Bounds on Particle Properties

Ever since NEWTON proposed that the Moon on its orbit follows the same laws of motion as an apple falling from a tree, the heavens have been a favorite laboratory for testing the fundamental laws of physics, notably Newton's and EINSTEIN's theories of gravity. More recently, astrophysics and cosmology have become crucial testing grounds for the microcosm of elementary particles. This area of scientific inquiry is part of a discipline often called astroparticle physics or particle astrophysics. There are a few important cases where unexplained astrophysical phenomena can be attributed to new elementary particles or new properties of known particles; the solar neutrino deficit and the dark matter problem are cases in point. Here, another widespread method will be discussed, where one uses established properties of stars or even the universe at large to constrain possible modifications caused by novel properties of elementary particles.

Neutrino masses

One of the earliest astrophysical particle limits is the cosmological bound on the masses of the three known NEUTRINOS, the electron, muon and tau neutrino ν_e, ν_μ and ν_τ. The main idea is that these particles, in spite of their weak interactions, must have been thermally produced in the hot and dense early universe and thus form a 'cosmic neutrino sea', much like the 3 K cosmic microwave background radiation. A detailed calculation reveals a present-day density of neutrinos plus antineutrinos of about 113 cm^{-3} for each of the three 'flavors' ν_e, ν_μ and ν_τ, to be compared with about 400 cm^{-3} microwave photons. All of the neutrino masses must obey $m_\nu \lesssim 40$ eV or else their gravitational effect would have slowed down the expansion of the universe below its observed rate. It was the discovery of ν_μ in 1962, together with the 1963 discovery of the cosmic microwave background, that inspired S S Gershtein and Ya B ZELDOVICH to advance this argument in 1966.

The direct experimental m_ν limits are much less restrictive than the cosmological bound except for ν_e (figure 1), and it is not known if neutrinos have masses at all. However, much evidence for 'neutrino flavor oscillations' has accrued from measurements of atmospheric and solar neutrinos, implying that they do have masses, and that the mass differences between them are extremely tiny, smaller than about 0.1 eV. In any case, neutrino masses appear to be much smaller than those of the other fundamental fermions, the electron and its two heavier brothers (muon and tau) as well as the six quarks (figure 1). The origin of this mass spectrum remains a mystery of elementary particle physics.

The cosmological m_ν limit represents a standard argument to constrain the properties of any theoretically proposed new particles such as monopoles, axions or supersymmetric particles. Their contribution to the cosmic mass density must be compatible with the age

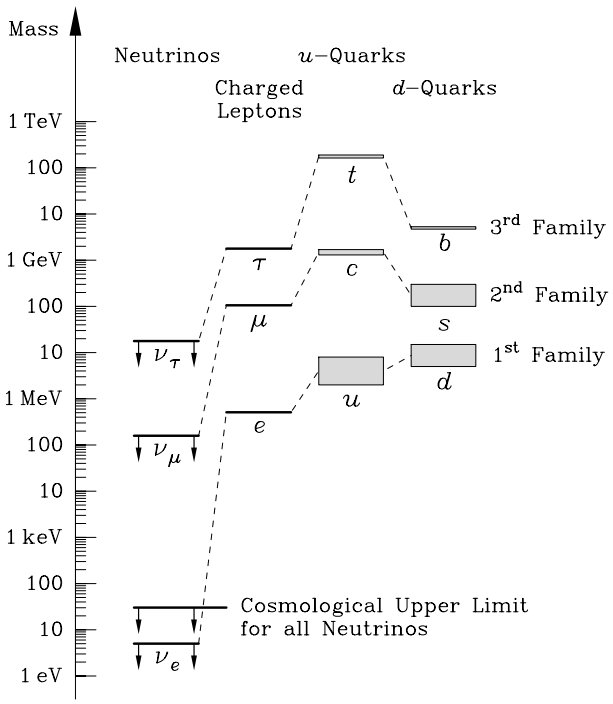


Figure 1. Mass spectrum of the elementary fermions. The vertical box size for the quarks indicates the mass uncertainty. For the neutrinos, the experimental upper limits are shown as well as the cosmological bound.

and expansion rate of the present-day universe. Usually, their relic density, the particle density surviving from the early universe, is more complicated to evaluate than for neutrinos where it simply scales with the density of cosmic microwave photons. In general, one has to worry about the detailed primordial production mechanism and annihilation rate of the new particles.

A completely different astrophysical method to constrain neutrino masses relies on the time-of-flight from a pulsed source. Neutrinos of mass m_ν which are emitted from a source at distance D at time $t = 0$, but with different energies E_ν, arrive at

$$t = \frac{D}{c\sqrt{1 - (m_\nu c^2/E_\nu)^2}} \qquad (1)$$

where $E_\nu \geq m_\nu c^2$ is the total energy including the rest mass and c is the speed of light. They are delayed because they move slower than c. Measuring their arrival times and energies allows one, in principle, to extract m_ν. A type II supernova (SN), the collapse of a massive star, produces a huge neutrino burst, lasting for a few seconds. Such an event has been measured once on 23 February 1987 when the electron antineutrinos from SN 1987A in the LARGE MAGELLANIC CLOUD ($D \approx 50$ kpc $= 163\,000$ light years) were registered in three neutrino detectors that were operational at that time (figure 2). The signal is compatible with the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

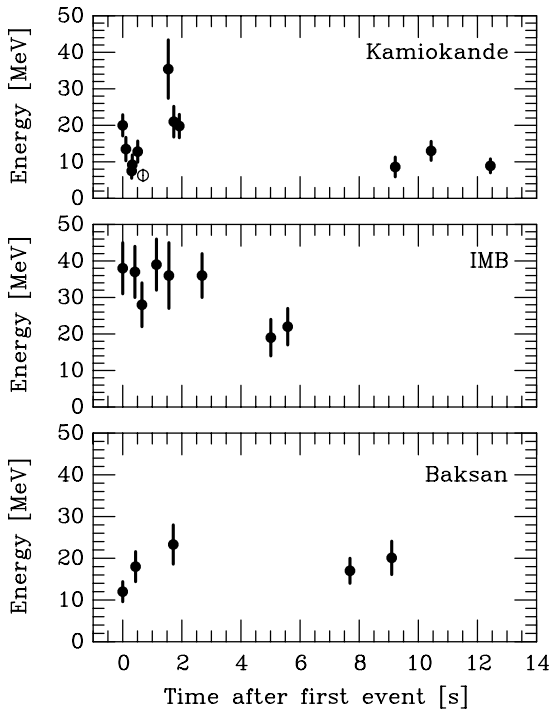


Figure 2. Neutrino signal from supernova 1987A at the Kamiokande, IMB and Baksan detectors. The clocks have unknown relative offsets; in each case the first event was shifted to $t = 0$. The Kamiokande event marked with an open circle is attributed to background.

expected intrinsic burst duration, with no evidence for time-of-flight dispersion, leading to a limit $m_{\nu_e} \lesssim 20$ eV.

Supernovas in or near our own galaxy are rare, perhaps a few per century. However, if one were lucky to observe another neutrino burst with one of the much larger detectors available today, one could improve the m_ν limits, even on ν_μ and ν_τ which are more difficult to detect. With the upcoming large 'neutrino telescopes' one may be able to measure neutrino pulses from more speculative sources such as gamma-ray bursts, with a potentially much better m_ν sensitivity.

Big-bang nucleosynthesis

Another standard cosmological argument is based on the expansion rate of the early universe at the epoch of primordial NUCLEOSYNTHESIS, about 3 min after the big bang. The radiation density at that time determines the expansion rate, which in turn influences the abundance of primordial helium which is observed to be 22–25% by mass, the remainder being hydrogen and traces of deuterium, helium-3 and lithium-7. Within their uncertainties, the observed primordial light-element abundances imply a radiation density corresponding to two to four families of neutrinos, nicely bracketing the three known flavors ν_e, ν_μ and ν_τ. These are the only possible families because additional ones would contribute to the decay width of the Z^0 gauge boson, which has been precisely measured since about 1990.

However, the big-bang nucleosynthesis method of counting neutrino families can be applied to more exotic particles that are not covered by the Z^0 decay-width argument. Any form of radiation which has been thermally excited in the early universe contributes to the expansion rate. Therefore, the interaction strength and masses of any proposed new particles must pass the test that they do not lead to excessive cosmic mass-energy density at the nucleosynthesis epoch.

Photon mass

In contrast with neutrinos which are expected to have masses, photons should be strictly massless, a theoretical expectation that requires experimental verification. Time-of-flight limits from the radio signal of pulsars are not much better than about 10^{-11} eV, the plasma frequency of the interstellar medium which mimics a photon mass. Laboratory tests of Coulomb's law give $m_\gamma \lesssim 10^{-14}$ eV; a non-vanishing m_γ would modify the inverse-square behavior. Likewise, the spatial variation of higher-multipole terms of large-scale magnetic fields give 0.6×10^{-15} eV and 0.8×10^{-15} eV from Jupiter's and Earth's magnetic fields respectively. The most restrictive limit of 10^{-27} eV is based on the magneto-gravitational equilibrium of the gas in the SMALL MAGELLANIC CLOUD which requires that the range of the interaction exceeds the characteristic field scale of about 3 kpc. This limit is surprisingly close to 10^{-33} eV where the photon Compton wavelength would exceed the radius of the observable universe and thus would cease to have any observable consequences.

Stellar energy loss

A particularly successful astrophysical argument to constrain the properties of low-mass, weakly interacting particles, notably neutrinos, goes back to J Bernstein, M Ruderman and G Feinberg (1963). They argued that a putative neutrino magnetic dipole moment would lead to an unacceptably large energy loss of stars by virtue of the plasma process $\gamma \rightarrow \nu\bar{\nu}$ which is kinematically allowed because the plasma frequency in the interior of a star gives photons effectively a mass. This energy-loss channel would add to the emission of neutrinos from the interior of stars by standard processes.

A non-standard energy-loss channel, caused by the emission of neutrinos with non-standard couplings or the emission of hypothetical particles such as axions, could cause a variety of observable effects. Stellar remnants (WHITE DWARFS, NEUTRON STARS) would cool faster. Normal stars would consume their nuclear fuel faster and thus burn out earlier. The helium-burning lifetime of horizontal-branch stars as measured by number counts in globular clusters has been a particularly precise measure of novel energy-loss channels. Likewise, a possible delay of helium ignition by particle cooling of red-giant cores can be well constrained by the observed brightness limit of the tip

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the red-giant branch in globular clusters. Even our Sun provides interesting constraints. An increased burning rate would make it more evolved than corresponds to its age, causing observable shifts of its precisely measured p-mode oscillation frequencies, a field of research known as helioseismology.

Arguments of this sort have been applied to a great variety of cases; perhaps the most useful limit is the constraint $\mu_\nu \lesssim 3 \times 10^{-12}\mu_B$ on a magnetic dipole moment of any neutrino. Here, $\mu_B = e\hbar/2m_e$, with e the electron charge and m_e the electron mass, is the Bohr magneton, corresponding to the electron's magnetic moment, which is usually employed as a measure for possible neutrino magnetic moments. Laboratory experiments give much weaker limits.

Likewise, one can derive a limit on a hypothetical neutrino electric charge, $e_\nu \lesssim 2 \times 10^{-14}e$, applicable to all flavors. For the measured electron antineutrinos from SN 1987A, the galactic magnetic field would have caused an energy-dependent deflection and thus a time-of-flight signal dispersion; the resulting limit is $e_{\nu_e} \leq 3 \times 10^{-17}e$. A laboratory limit, based on charge conservation in β decay, is $e_{\nu_e} \lesssim 3 \times 10^{-21}e$.

If neutrinos have masses they may decay; one possible channel is by the emission of a photon. The absence of anomalous x- or gamma-ray fluxes from the Sun or in conjunction with the SN 1987A neutrino burst as well as limits on a spectral deformation of the black-body nature of the cosmic microwave background constrain this radiative decay channel. However, the matrix element of this process depends on an electromagnetic transition moment which is subject to the above stellar energy-loss constraint; it is more restrictive for $m_\nu \lesssim 1$ eV than the direct limits on decay photons.

Supernova 1987A

The SUPERNOVA 1987A neutrino signal (figure 2) allowed for a particularly useful application of the stellar-energy loss argument. The nascent neutron star that has formed after collapse cools by neutrino emission over several seconds, a time scale which is set by the neutrino diffusion rate; a proto neutron star is so hot and dense that even neutrinos cannot freely escape. A competing cooling channel by more weakly interacting particles which are not trapped would shorten and weaken the neutrino signal because the total amount of energy, the gravitational binding energy of the neutron star, is fixed.

This simple argument has been applied to many cases, but perhaps the most interesting one is that of axions, hypothetical low-mass particles that were originally proposed in the context of quantum chromodynamics, the theory of the strong interaction among quarks. It was later recognized that axions are also a good particle candidate for the dark matter of the universe. Axion models are essentially characterized by one free parameter which can be chosen to be their mass m_a. The requirement that axions do not carry away the bulk of the SN 1987A energy translates into a limit $m_a \lesssim 10^{-2}$ eV. On the other hand, the requirement that their cosmic mass density is not too large leads to $m_a \gtrsim 10^{-5}$ eV, in contrast with the above case of neutrinos where cosmology provides an upper mass limit. In the remaining mass range, axions could well exist and provide most or all of the cosmic dark matter.

Strange stars

The fundamental constituents of nuclear matter are quarks, not protons and neutrons, and one may well wonder if neutron stars do not undergo a phase transition to a medium consisting of free quarks. Even the strange quark is relatively light (figure 1) so that one may speculate that 'neutron stars', the remnants of massive stars after a type II supernova collapse, are actually 'strange stars', consisting of a strongly interacting medium of up (u), down (d) and strange (s) quarks. It is surprisingly difficult to distinguish the two cases observationally. It is thought that strange stars can rotate faster before breaking apart so that sub-millisecond pulsars, if they were ever detected, could perhaps resolve the question of the ground state of nuclear matter.

In summary, astrophysics and cosmology provide important constraints on the properties and interactions of elementary particles which are often complementary to what can be measured in laboratory experiments.

Bibliography

Caso C *et al* 1998 Review of Particle Physics *Euro. Phys. J.* **C3** 1 http://pdg.lbl.gov/

Kolb E W and Turner M S 1990 *The Early Universe* (Redwood City, CA: Addison Wesley)

Raffelt G G 1996 *Stars as Laboratories for Fundamental Physics: The Astrophysics of Neutrinos, Axions, and Other Weakly Interacting Particles* (Chicago: University of Chicago Press)

Raffelt G G 1999 Particle physics from stars *Ann. Rev. Nuc. Particle Phys.* **49**

Will C M 1993 *Theory and Experiment in Gravitational Physics* revised edn (Cambridge: Cambridge University Press)

Georg G Raffelt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysical Institute, Potsdam

Built upon a tradition of almost 300 years, the Astrophysical Institute Potsdam (AIP) is in an historical sense the successor of one of the oldest astronomical observatories in Germany. It is the first institute in the world which incorporated the term 'astrophysical' in its name, and is connected with distinguished scientists such as Karl Schwarzschild and Albert Einstein. The AIP constitutes one of the four centers of astrophysical research in Germany and is the largest astronomical institute in the new states of Germany. The cooperation with the University of Potsdam is of utmost importance.

The bodies responsible for AIP are the Ministry for Science, Research and Culture of the State of Brandenburg and the Federal Ministry for Education, Science, Research and Technology. The AIP is an institute of the 'Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz'.

The AIP concentrates its efforts on two main avenues of research: cosmic magnetic fields, solar and stellar activity (Bereich I); and extragalactic astrophysics and cosmology (Bereich II).

Within a German consortium the AIP is part of the international 'LARGE BINOCULAR TELESCOPE' (LBT) to be built in Arizona by 2004 and contributes to the operation of the German solar telescopes on Tenerife.

The main site of the institute is the Sternwarte Babelsberg. The Solar Observatory Einstein Tower at the Telegrafenberg in Potsdam and the Observatory for Solar Radio Astronomy in Tremsdorf near Potsdam are branches of the institute.

For further information see
http://www.aip.de/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysical Jets

There are many circumstances in astronomy when gas is found to circulate around a central gravitating mass. This happens in BINARY STARS, where one component is a compact object—white dwarf, neutron star or black hole—and mass is transferred from the normal stellar companion. It happens in the nuclei of galaxies, where the gravitating object is a massive black hole and it also happens during the earliest phases of evolution of a protostar. These gas flows are called ACCRETION DISKS and they are frequently accompanied by rapid outflows, launched in antiparallel directions, roughly perpendicular to the disks. The outflows are usually called jets and this article describes some of what we know about them.

The first example of an astrophysical jet was discovered by HEBER CURTIS in 1918 using the Lick Observatory in California. He observed the galaxy known as M87 in the Virgo cluster of galaxies and, when he inspected the nucleus, he found a 'curious straight ray... apparently connected with the nucleus by a thin line of matter'. This was a remarkable observation (by a remarkable astronomer), and the jet in M87 is still one of the most carefully studied of this class of objects (figure 1).

What we observe directly in M87 is a fairly straight feature, some 2 kpc in length. It can be seen from long radio wavelengths to high-energy x-rays and is found to exhibit about eight regions of high intensity, known as 'knots', along its length. The jet can be traced down to a scale size of order 0.01 pc which is smaller than 10^{-5} times its total length and less than 100 times the radius of the black hole that has been found to lie in the nucleus of M87, and whose mass is measured to be about 3 billion solar masses (see M87: THE NUCLEUS AND JET.

We know that the gas is flowing away from the black hole in M87 because we can use very long baseline interferometry to measure its speed. What is found is that some parts of the jet appear to be moving outward with a speed that appears to be as much as six times faster than the speed of light. This is actually quite a common phenomenon called superluminal expansion. It does not violate the special theory of relativity (which stipulates that all material motion occurs slower than the speed of light). Instead, superluminal expansion is an illusion which only requires that the gas in the jet moves towards us with a speed close to that of light. For this reason, jets like that in M87 are often styled relativistic jets. It is quite natural that these jets move with speeds close to that of light; after all this is the escape velocity from the central black hole.

The jet in M87 exhibits another common property of these objects in that its antiparallel counter jet, which is expected to be present, has not yet been detected. This is an illustration of an effect called ABERRATION. When a source of radiation moves at near light speed, it will appear to beam its emission along its direction of motion. Consequently, if the source happens to be approaching us, then it will appear to be very bright, whereas, if the source moves away from us, it will appear to be faint. This explains why we are only able to detect one jet in M87 and why a high proportion of the brightest sources that we see exhibit SUPERLUMINAL MOTION.

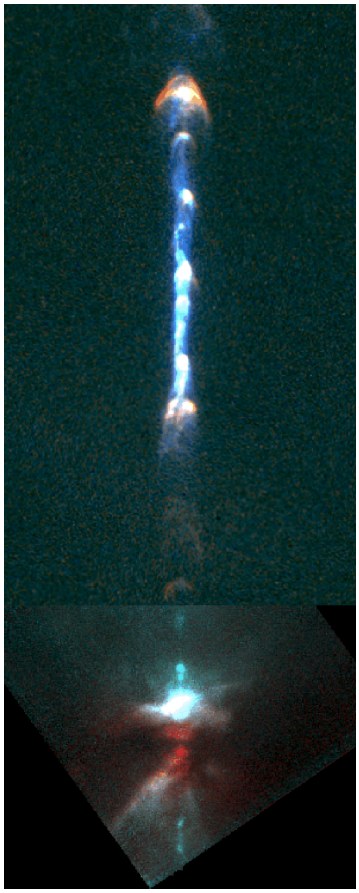


Figure 2. HH 111. Protostellar jets emerging from a young star—one member of a triple system observed using Hubble Space Telescope. The jets are over 10 light years in length. The bright knots are probably associated with shock waves formed when faster moving gas runs into slower gas.

Let us now turn to the knots. These are believed to be shock waves formed in the outflowing jets when, for example, the jet speed at the source increases by an amount greater than the speed of sound in the jet and therefore faster than the speed with which information can be transmitted by the moving jet fluid. A strong shock discontinuity must eventually form in the flow and this turns out to be an excellent site for accelerating high-energy, relativistic electrons with energies up to 100 TeV. These electrons radiate by spiralling in the magnetic field that permeates the jet—a process known as SYNCHROTRON RADIATION—and this is thought to be responsible for the entire radio through x-ray emission.

M87's jet is relatively weak. We know so much about it because it is so close. In the past, it was almost surely supplied with gaseous fuel at a far higher rate than at

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

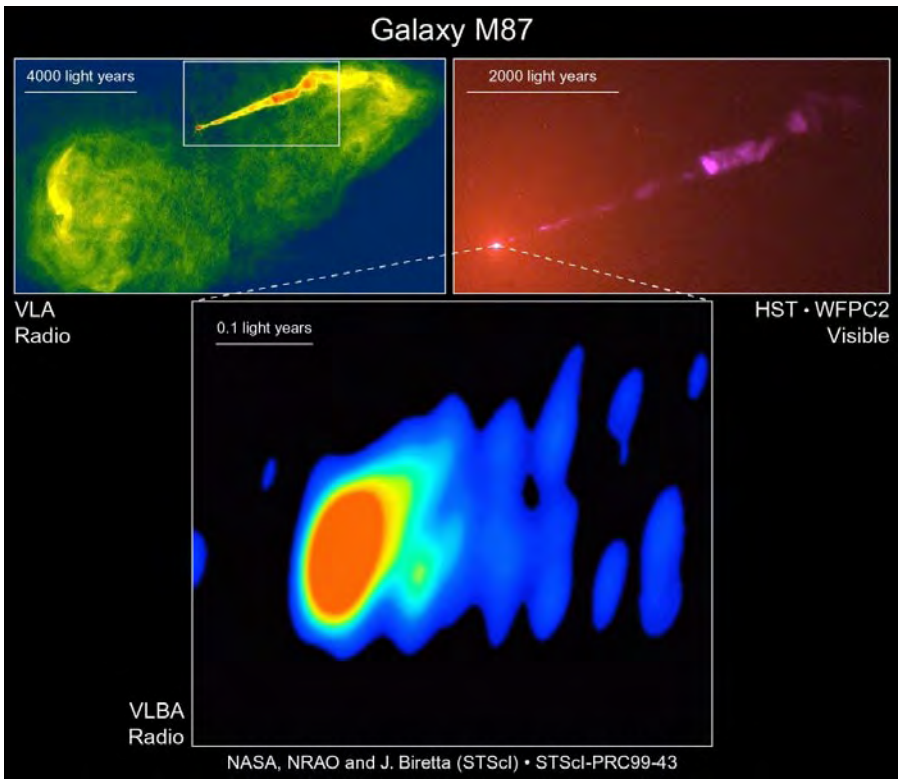


Figure 1. Radio and optical maps of the jet observed emerging from the nucleus of the nearby elliptical galaxy M87. The jet can be traced down to a distance smaller than 100 times the size of the central black hole.

present. It would then have been classified as a QUASAR. Quasars are the hyperactive nuclei of very distant galaxies that can outshine their stellar hosts at optical wavelengths. They emit prodigiously throughout the electromagnetic spectrum. Quasars often exhibit giant, double, radio sources—two 'lobes' of intense radio emission located on opposite sides of the galaxy. Historically, this was the way in which quasars were originally identified. For a long while there was a mystery as to how these radio lobes could be formed and sustained. With the discovery of relativistic jets, the mystery was solved. The source of their power is the central black hole and the accretion disk that surrounds it. This creates relativistic particles and magnetic field which flow outward, along two antiparallel directions to replenish the radio lobes and to propel them away from the galaxy, rather like the exhausts that are associated with automobiles and jet aircraft. In some sense, the jets carry away the waste products of accretion onto a black hole. We know of thousands of these double radio sources and have documented their properties so that we are beginning to understand how they fit into an evolutionary sequence.

The most powerful relativistic jets, that are beamed almost directly at us, are sometimes called BLAZARS and these are commonly observed as γ-ray sources. Photons, with energies as high as several GeV, are formed when an x-ray photon is scattered by a relativistic electron—a process known as inverse Compton scattering. (In a few local examples, it has been possible to detect very-high-energy γ-rays with TeV energies that are probably produced close to the central black hole.) The surprisingly large intensity of these γ-ray jets tells us that, in many instances, they are transporting away a large fraction of the total power generated by the central black hole engines.

Not all jets are associated with distant galaxies. There are many stellar-sized black holes that have been found in our Galaxy and in nearby galaxies. When these have stellar companions that can lose gas, the gas will be attracted by the black hole and can also form an accretion disk. An early example was the source SS433 where two precessing antiparallel jets moving with speed just over a quarter the speed of light have been found. In some instances, these sources also produce relativistic jets, small-scale versions of the extragalactic radio sources. They exhibit giant outbursts, superluminal expansion and beaming and, as a result, are sometimes called 'microquasars'. These are particularly good to study because astronomers can observe many cycles of variation in the time it takes a quasar to change.

Jets have also been found associated with newly formed stars. Here, again there is an orbiting disk made of the gas that has not yet settled onto the star (or condensed into planets), These 'YOUNG STELLAR OBJECTS' are observed primarily in optical and infrared emission lines and the jet moves with speeds of several hundred

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

km per second which is comparable with the escape velocity from the central protostar. These jets sometimes also exhibit moving shock waves known as Herbig–Haro objects (figure 2), which propagate along the length of the jets. (As the jet speeds are much less than the speed of light, there is neither superluminal motion nor aberration in this case.) It is almost likely that similar basic physical processes are at work here too. These 'protostellar jets' are particularly good to study because the flow velocities can be measured accurately using spectral lines.

The nature of the mechanism for the formation and collimation of jets is still a matter of controversy. Many astronomers believe that strong magnetic fields are generated in the accretion disk and that these are responsible for launching and focusing the outflow. Most of the jet power probably derives from close to the central object, black hole, neutron star, white dwarf or protostar, also through the agency of magnetic field. Computer simulations of magnetized gas flows around central gravitating bodies are starting to give some insight into how and why jets are such a frequent companion to accretion disks.

R D Blandford

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysics

That branch of astronomy concerned with the physics of, and physical processes operating in, stars and other celestial bodies. It may be argued that the subject of astrophysics really began with the development of spectroscopy in the nineteenth century. Today astrophysics has transformed cosmology, the study of the evolution of the universe, from an almost purely speculative activity to a modern science capable of predictions that can be tested.

See also: astrometry, astronomy, universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysics to 1900

In practice, there is no clear dividing line between astrophysics and astronomy. The former is usually defined as the study of the physical properties of celestial objects, whereas the latter is concerned with their positions and motions. However, the two types of measurement are often linked. For example, stellar mass—a physical property—is typically measured by observing the orbits of two stars round each other. However, it was evident from the start that the equipment used and knowledge required for the two types of activity were different. During the latter part of the 19th century, as the subject began to develop rapidly, it was therefore recognized as a distinct field of study, although it took some time before 'astrophysics' became the accepted name. At the time, astrophysical measurements were as often described in terms of 'celestial chemistry' as 'celestial physics'. Because such observations differed from traditional positional measurements, astrophysics was also, on occasion, called the 'new astronomy'. (To add to the confusion, studies of motion were sometimes referred to as 'physical astronomy'.) These contemporary terms will be ignored here: relevant work will simply be called 'astrophysical'.

Early astrophysical observations

Some physical properties of celestial bodies can be determined by direct observation through a telescope. From this viewpoint, observations of astrophysical significance date back to the invention of the telescope in the early 17th century. Prior to this there had been speculation—for example, that stars were bodies like the Sun—but little by way of firm evidence. Several of GALILEO's observations in 1610, with his newly invented astronomical telescope, proved to have physical implications. One was his discovery that lunar topography bears some resemblance to the Earth's (the first step towards the lunar geology of the space age). The most important of his observations for subsequent astrophysics was his discovery of SUNSPOTS. Apart from recording the complex, changing nature of sunspots, Galileo also used them to show that the Sun, like the Earth, was rotating on its axis. Later in the 17th century, another of Galileo's discoveries—the moons of Jupiter (see JUPITER: SATELLITES)—led to the first determination of the velocity of light. RØMER, working at the newly established Paris Observatory, found that the times at which the moons appeared and disappeared behind Jupiter seemed to vary systematically. He attributed this to changes in the Earth–Jupiter distance, together with a finite speed of light. Rømer's measurements provide an early example of the fact that, although the study of astrophysics mostly entails astronomers relying on developments in physics, they have sometimes been able to contribute something in return.

In the latter part of the 17th century, NEWTON examined the way in which sunlight, after passing through a prism, was spread out into its constituent colours. Such spectroscopic observation was to become the most significant part of astrophysics, but only after some delay. In 1800, W Herschel in England repeated Newton's experiment (see HERSCHEL FAMILY), but with the new aim of measuring the heating power of different parts of the solar spectrum. The thermometer he used recorded the existence of considerable heating beyond the red end of the visible spectrum. This was the first identification of infrared radiation (although, as is often the case with important advances, Herschel interpreted it differently). It was paralleled in the following year by the first detection of ultraviolet radiation. Ritter in Germany, inspired by Herschel's work, investigated the action of the solar spectrum on silver chloride. He found that maximum breakdown of the chemical occurred beyond the violet end of the visible spectrum. (This was an early step on the way to photography.) A year later still, in 1802, Wollaston, who was also studying the solar spectrum back in England, found that it was crossed by a few dark lines. He thought that these simply represented demarcation points where the various spectral colours began and ended. It was left to Fraunhofer in Germany, a decade later, to discover that the whole of the solar spectrum was intersected by dark lines. He detected nearly 600 of these; so, hardly surprisingly, they came to be referred to as 'FRAUNHOFER' LINES.

The solar spectrum

As this rush of discoveries suggests, it was an interest in the Sun, and, more especially, in the solar spectrum, that primarily triggered the 19th century growth of astrophysics. There were various reasons for this interest. One was that the Sun was an excellent light source. As compared with other light sources, whether celestial or in the laboratory, the Sun produced copious amounts of light. Consequently, even highly inefficient optical equipment could yield results. Another reason was the particular interest that early 19th century scientists had in color phenomena. The solar spectrum provided the easiest way of studying these.

Laboratory investigations showed that different substances produced different sets of spectral lines. So there was a tendency from early on to suppose that the Fraunhofer lines could be related to chemical substances on the Earth. However, this supposition was not fully confirmed until the second half of the 19th century. There were many problems to resolve first. For example, it was found that some Fraunhofer lines did not originate on the Sun at all, but were impressed on the sunlight during its passage through the Earth's atmosphere. More fundamentally, laboratory spectra consisted of bright, not dark, lines. How were the two connected? In 1849, FOUCAULT finally managed to show the identity of lines appearing in emission and absorption. To do this, he experimented with the pair of lines in the solar spectrum that Fraunhofer had labeled 'D', because they were particularly intense. However, these same lines were also responsible for retarding attempts to interpret

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spectra. The difficulty was that the D lines appeared to be ubiquitous in the laboratory. Whatever substance was being studied, they always seemed to appear. It was only in the 1850s that the widespread contamination of laboratory chemicals by salt—leading to the appearance of the sodium D lines—was fully recognized.

All this clearing of the ground had to occur before the solar spectrum could be properly interpreted. The final step in this process can be seen as the key event in the foundation of astrophysics. It was taken by KIRCHHOFF, working in conjunction with his Heidelberg colleague, Bunsen, in the period around 1860. Their studies actually began in the laboratory with the invention of the Bunsen burner. Bunsen intended to apply the non-luminous flame this produced to studying the colours emitted by different elements when they were burnt. For this purpose, he proposed to use a set of color filters. Kirchhoff suggested Bunsen should use a prism instead, and this led to a comparison with the solar spectrum.

Kirchhoff established two basic principles for the analysis of spectra. The first was that incandescent solids or liquids give continuous spectra, but gases produce lines or bands (the positions of which depend on the particular gas present). The second was that, when a source of continuous spectrum is viewed through a cooler gas, the wavelengths at which absorption occurs correspond to those at which emission occurs when the gas is heated. The immediate implications for the Sun were similarly twofold. Firstly, the body of the Sun must consist of hot dense material to produce the observed continuous spectrum. Secondly, the absorption lines produced in the cooler solar atmosphere can be used to study the chemical composition of the Sun.

Although it was a study of the atmospheric lines that dominated subsequent 19th century astrophysics, Kirchhoff's conclusion regarding the body of the Sun had an immediate impact on contemporary thought. The picture of the Sun widely held in the first half of the 19th century saw the sunspots as holes in a luminous cloud, through which a colder interior could be viewed. Indeed, W Herschel speculated that the solar interior, visualized as having a solid surface, might even be inhabited. Kirchhoff's spectrum analysis showed that such a picture was impossible.

Other solar studies

By this time, another problem concerning the Sun had arisen. The first half of the 19th century saw the first significant attempts to measure the amount of heat reaching us from the Sun. Such measurements of what came to be called the 'SOLAR CONSTANT' provided a quantitative estimate of the vast amount of energy being emitted by the Sun. At about the same time, the idea of the conservation of energy began to establish itself. So the question arose—where does the Sun obtain its energy? It was clear that chemical sources would only keep the Sun burning for a very short time, so some physical source was necessary. The solution that came to be widely accepted during the second half of the 19th century was put forward by HELMHOLTZ in 1854. He suggested that the Sun gained its heat from gravitational contraction. The rate required was well below the observational limit, and the process was capable of sustaining the Sun's output for several million years. This lifetime seemed reasonable to physical scientists, but it came under increasing attack from geologists and biologists. They thought—correctly, as it turned out—that the evolution of the Earth's surface and its inhabitants required a much longer span of time.

One further characteristic of the Sun became the focus of attention in the mid-19th century. This was its changing activity. In the early 1840s, SCHWABE, a German amateur astronomer, found from his counts of sunspots that the number visible on the Sun's disc varied periodically with time. His conclusion was publicized a few years later by HUMBOLDT, at a time when a number of studies were being made of fluctuations in the Earth's magnetic field. In 1852, SABINE in England pointed out that these magnetic fluctuations varied with a period that coincided with the sunspot cycle. This conclusion was soon confirmed and extended by others and led to a continuing interest in solar–terrestrial relations throughout the second half of the 19th century.

The major features of the Sun's atmosphere had already been discerned before Kirchhoff became interested in the solar spectrum. Since the solar atmosphere was too faint to be seen in the presence of full sunlight, its characteristics had had to be examined during solar ECLIPSES. Expeditions to eclipses in 1842 and 1851 had shown that the Sun's surface was surmounted by a CHROMOSPHERE (so called because of its red color). Above this stretched a much thinner CORONA. Extending into this from the chromosphere were large 'flames' (later called 'prominences'). Kirchhoff's explanation of the solar spectrum now opened the way for a more detailed examination of the atmospheric structure.

The first astrophysical advance was made in 1868. In that year, a solar eclipse was visible from India, and one of the observers—the Frenchman, JANSSEN—looked at the spectra of prominences. He and others found bright lines, showing that prominences were clouds of gas. Janssen concluded that the lines were so bright that they should be visible out of eclipse. He subsequently scanned the solar limb and showed that this was, indeed, the case. At the same time, LOCKYER in England made spectroscopic observations of the Sun and observed the same lines. This meant that both the chromosphere and the prominences could be mapped at the solar limb, without requiring an eclipse. Because the cornona was so faint, it continued to be observable only at eclipses throughout the 19th century. At the 1869 eclipse, the first bright coronal line was found. It was now evident that the emission lines of both the chromosphere and the corona differed significantly from the absorption spectrum of the Sun. From Kirchhoff's work, this meant that neither could be the source of the Fraunhofer lines. At the 1870 eclipse, the American astronomer, YOUNG, showed that the absorption

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

lines were actually produced in a restricted region—which was labeled the 'reversing layer'—between the chromosphere and the apparent surface of the Sun.

The explanation of the bright solar surface also developed further in the 1860s. At the end of that decade, Andrews in England showed that gases at a high enough temperature could not be liquified. This confirmed what was already suspected: that the body of the Sun must consist of gas under high pressure, rather than of liquid as Kirchhoff had supposed. This raised a new question—why did the Sun appear to have a clearly defined surface (labeled the photosphere) at all? It was generally agreed that the photosphere must represent the level in the Sun where the gases became sufficiently cool for refractory materials to condense out. The renewed belief in a cloudy solar surface remained in vogue until the early 20th century.

Thus, within the space of a few years, spectroscopic observations had provided a reasonably clear picture of the Sun's atmosphere. The next few decades were spent in sorting out some of the uncertainties in this picture. One related to chemical composition. Following Kirchhoff, a number of observers compared terrestrial spectra with the solar spectrum in an attempt to determine the chemical composition of the Sun. It was found that almost all the elements that could be detected were metals (although hydrogen was noted early on). However, it proved impossible to identify one very bright line in the chromospheric spectrum. Lockyer suggested that this was produced by a new element, consequently called 'helium', which had not yet been found on Earth. This suggestion was strongly contested, but Lockyer was vindicated in the 1890s when Ramsay discovered the gas, via its spectrum, in terrestrial rocks. It was similarly found to be impossible to identify the coronal lines. The parallel suggestion was therefore made that they were due to a new element, labeled 'coronium'. In this case, no corresponding element was ever found. (It was not until near the middle of the 20th century that the coronal lines were shown to be due to known elements in highly ionized states.)

Stellar astrophysics

In the early days of astrophysics, the Sun was such a popular target for observation that the whole subject was sometimes simply called 'solar physics'. This overlooked the fact that Kirchhoff's findings were also immediately applied to observations of stellar spectra. It was soon apparent that stellar spectra could differ greatly from each other. So, together with investigating stellar chemical composition, it was also necessary to explore the possibility of classifying stars into groups. This was pursued in the 1860s primarily by the Jesuit astronomer, SECCHI. After examining a few thousand stars, he came to the conclusion that they could all be placed into four groups, depending on the appearance of their spectra. It was noted that these spectral types were correlated with the colors of the stars—white, yellow or red. During the same period, HUGGINS in England used higher-dispersion spectra to look for chemical constituents. Despite the differences between the spectra, he found that many lines could be identified with elements found on the Earth and the Sun.

In these early days, stellar spectra—like the solar spectrum—were observed visually. Whereas the Sun produced adequate light, this was not true of stars. The obvious answer was to use photography, but the 'wet plate' process in vogue in the 1860s was insufficiently sensitive. It was not until the end of the 1870s, when the 'dry plate' process became available, that the photography of stellar spectra became feasible. New discoveries then followed quickly. An important one derived from the fact that the photographic plates were more sensitive than the human eye in the ultraviolet, with the result that a number of new lines were detected in that region. Huggins in England and VOGEL in Germany found lines there which they believed extended the sequence of hydrogen lines found in the visible. A study of the wavelengths of these lines led the Swiss, BALMER, to announce the numerical formula for the wavelengths of hydrogen lines now called after him. His success encouraged the idea that spectral lines are produced by related sequences of atomic vibrations. Photographic recording of stellar spectra also allowed the first reasonably accurate radial velocities to be measured. This led to the discovery of spectroscopic binaries: the first found—Mizar—was detected by Miss Maury at Harvard towards the end of the 1880s.

Stars and nebulae

Photography also ultimately helped with another type of astrophysical measurement—stellar brightness. During the 1880s, E C PICKERING at Harvard and Pritchard at Oxford produced photometric catalogs of the brighter stars in the Northern hemisphere based on visual estimates of brightness. They were extended to Southern hemisphere stars in the early 1890s by Bailey observing in Peru. It was generally believed by this time that photographic photometry would be the way forward in the future. Unfortunately, it proved difficult to reach agreement on how to reduce the photographic images to equivalent magnitudes (and it was only slowly realized that photographic and visual magnitudes could differ considerably, depending on the color of the star). Consequently, photographic photometry had only contributed to a minor extent by the end of the 19th century.

By the latter years of the century, photography—particularly the Harvard survey work—was proving a valuable way of picking up variable stars. Interest in the visual observation of variables, originally stressed by Argelander in Germany, had grown from the mid-century onwards. The system of designating variable stars via letters of the alphabet was already in use by the 1880s, and exemplars of the main types of variable were under investigation by the end of the century. The ones that attracted most attention were the NOVAE (supernovae were not distinguished as a separate category until the 20th

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

century). Huggins, in 1866, was the first to examine the spectrum of a nova. He found that it consisted of an ordinary stellar absorption spectrum with superimposed bright lines. Both this and later novae spectra were therefore interpreted as indicating the explosion of a cloud of gas from a stellar body.

During the 1860s, Huggins undertook a survey of the spectra of bright nebulae. (In the 19th century, a 'nebula' meant any extended patch of light, as compared with stars, which were point sources.) Ever since W Herschel's time, there had been a debate concerning the nature of nebulae. Were they some kind of fluid, or stars clustered so close together that they could not be resolved? The first object Huggins examined was a PLANETARY NEBULA. It showed a bright line spectrum, which was immediately interpreted as meaning that nebulae were gas clouds. The problem was that some of the spectral lines could not be linked to any known terrestrial elements. It was therefore suggested, by analogy with helium, that a new element—'nebulium'—was present. Although the idea remained in play for the rest of the century, nebulium, like coronium, ultimately proved to be spurious. The bright lines were shown in the 20th century to be due to oxygen in a rarified state.

As more spectra of nebulae were observed, the initial conclusion concerning their nature became more questionable. Planetary nebulae and what were called 'irregular' nebulae (of which the Orion nebula was a prime example) always gave emission lines. However, other nebulae (a key example being the Andromeda nebula) gave spectra that appeared to be continuous, without either dark or bright lines. The interpretation of this kind of spectrum was a matter of dispute for the rest of the century. The first hint of the way forward did not come until 1899, when SCHEINER in Germany found faint traces of absorption lines on a long-exposure photograph of the spectrum of the Andromeda nebula. Direct photography had by this time provided some further information on this latter type of nebula. Lord ROSSE, observing in Ireland in the 1840s, had discerned visually that a few nebulae seemed to have a spiral structure. When sensitive photographic plates became available towards the end of the century, it was shown—particularly by KEELER in the United States—that many more spiral nebulae existed. It was natural to suppose that the unusual spectrum and the unusual shape of such nebulae were linked.

At the end of the 18th century, the French mathematician, LAPLACE, had suggested a way in which the solar system might have formed. He envisaged a contracting nebula, which spun more rapidly as it shrank. The spin rate finally reached a point where material was ejected in rings round the equator. This material subsequently condensed into planets, with the final central condensation forming the Sun. The picture painted by this 'nebular hypothesis' suggested that the contracting material would look very like a spiral nebula. So it came to be thought that such nebulae were simply nearby planetary systems in the process of forming.

In 1885, the confusion was added to by the appearance of a new star—actually a supernova—in the Andromeda nebula. This was naturally equated to other new stars—ordinary novae—observed in the Milky Way during the 19th century. The corresponding misestimate of distance suggested that the Andromeda nebula was part of our own Galaxy. Another contributory factor was the distribution of the spiral nebulae. It became apparent that these did not occur near the plane of the Milky Way (the so-called 'zone of avoidance'). Although the cause was not known (subsequently found to be due to obscuring dust in the Milky Way), this distribution suggested that the nebulae were linked to our Galaxy. At the beginning of the century, the idea that many nebulae were clusters of stars, some equal to our own Milky Way in extent, had been widely debated. By the end of the century, most believed the universe to be dominated by the Milky Way.

Stellar evolution

By this time, a lively interest was developing in stellar evolution. It was inevitable, once different categories of stars had been established, to ask how the groupings related to each other. The contemporary concern with evolution in geology and biology naturally encouraged a search for relationships that were evolutionary in nature. The first widely held view, developed particularly by Zöllner and Vogel in Germany, was that stars formed a single evolutionary sequence. They were born hot and white in color. As they dissipated their heat, they became successively yellow and then red. There was a problem here. The only physical mechanism known to drive change was contraction. Both Lane in the USA and Ritter in Germany showed that a contracting sphere would actually heat up, rather than cool. However, their results assumed that a star was a sphere of perfect gas. No-one at the time thought that such a model could represent a real star.

By the end of the century, evidence was beginning to accumulate that stars of the same color could differ appreciably in other characteristics, such as size. The most radical revision to the 19th century picture of stellar evolution—Lockyer's 'meteoritic hypothesis', proposed in the 1880s—assumed that this was the case. Lockyer suggested that stars formed from nebulae that consisted of clouds of meteorites (popular objects of study in the latter part of the century). Contraction of the nebula led to an increased rate of collision between its constituent meteorites, and so raised its temperature. The nebula, now visible as a star, gradually changed its color from red, through yellow, to white. By the time this latter point had been reached, all the meteorites had been vaporized. The white star therefore now cooled down, becoming smaller as it went back through the sequence from yellow to red. Unlike other schemes, Lockyer's hypothesis thus supposed that red stars fell into two categories—either very large, or very small.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interpreting spectra

Lockyer based his arguments on a comparison of the spectra of meteorites in the laboratory with spectra from a wide range of stars. Some of this work came under attack, and his hypothesis was not widely supported. Nevertheless, detailed studies of stellar spectra became increasingly a topic of interest. In the mid-1880s, E C Pickering at Harvard initiated a large-scale investigation of stellar spectroscopy and photometry. The first results appeared in the 1890s as the Draper memorial catalog of stellar spectra. (Draper was a pioneering US astrophysicist, whose family provided money for the project subsequent to his death.) The catalog introduced the alphabetical classification of stellar spectra still in use today. The spectra used for classification purposes at Harvard were too small scale for detailed study, but Miss Maury examined a subset of the stars in greater detail. She found that stars of the same color could still have subtle differences in the appearance of their spectra. She speculated that these might have implications for stellar evolution. (Indeed, her work led onto Hertzsprung's distinction between giant and dwarf stars in the early years of the 20th century.)

The basic problem was that the theoretical understanding of spectra remained fairly primitive throughout the 19th century. The difficulty was that, not only were the spectra often complex in themselves, but, even worse, they could change according to the light source. For example, the laboratory spectrum of an element produced with an electric arc could have some lines different from a spectrum of the same element produced with an electric spark. One proposed explanation was that elements could be broken down to an extent that varied according to the conditions to which they were subjected. This suggestion met considerable opposition, not least because it had been a basic tenet of 19th century science that atoms were indestructible.

Unfortunately, Lockyer, who was the strongest supporter of this argument, initially held a mistaken idea of what was happening. It had been noted early on that laboratory spectra from different elements often had some lines coincident in position. Locker called them 'basic' lines and believed that they revealed the presence of a common breakdown product from the initial elements. As higher-dispersion spectroscopes became available, it was found that these basic lines were simply close to each other in position, rather than coincident. Consequently, they possessed no fundamental significance. Towards the end of the century, Lockyer was influenced by his assistant, A FOWLER, to alter his ideas about the breakdown of elements. Attention was now called to another group of lines—labeled 'enhanced' lines—which varied from source to source. These lines were, indeed, associated with what would now be called ionized atoms. However, Lockyer's contemporaries were mostly so confused by his reinterpretation that the new idea received little initial attention. By this time—the end of the century—evidence was beginning to accumulate that atoms could be broken down. J J Thomson's work on the electron is a key example of this. Detailed understanding of astronomical spectra had to await the development of quantum theory in the 20th century.

In trying to interpret spectra, the Sun remained the prime source of comparison data throughout the 19th century. For example, the spectra of sunspots were found to have bands similar to those found in the spectra of red stars. Since sunspots were believed to be cooler than the Sun as a whole, this fitted in well with expectations regarding stellar temperatures. The problem lay in actually putting a figure to the temperatures involved. A white star might be hotter than a red star, but how much hotter? Again, it was hoped to answer this question first for the Sun. Ways of linking the temperature of a body to the radiation it emitted were limited in the mid-19th century. They mostly depended either on Newton's law or on Dulong and Petit's law, but these gave widely differing results. For example, one estimate in 1860 suggested the solar surface was at some 7 000 000 °C, whereas another in 1872 suggested 1400 °C. Of the various proposed revisions, the one that ultimately proved most useful was put forward by Stefan in 1879. His suggestion of a dependence on the fourth power of the temperature had no theoretical backing until the development of quantum theory. It was therefore treated with some reserve in the 19th century.

The development of astrophysics

In 1859, a solar FLARE was observed for the first time. It was intense enough to be visible in white light against the Sun's disc and was accompanied by a major magnetic storm and widespread AURORAE on Earth. The difficulty was that such events were normally invisible, since instrumental limitations meant that most solar observing was done at limb. This changed in the 1890s, when HALE in the USA (and also DESLANDRES in France) introduced the idea of viewing the entire Sun in the light of a single line, so allowing solar activity to be scanned across the disc. This development well illustrates the continuing link between new astrophysical research and better instrumentation. By the end of the century, not only was an increasing diversity of equipment being hung on the end of telescopes: the telescopes, themselves, were undergoing change. Thus, the fact that photographic plates were sensitive to a different region of the spectrum from the human eye meant that refracting telescopes for photographic use had to be specially designed and adjusted. Furthermore, the need for the maximum possible light grasp led to a growing interest in the use of reflecting telescopes for astrophysical research.

In its early years, many astronomers regarded astrophysics as an inferior adjunct to classical astronomy. It was less rigorous, not requiring the grounding in mathematics essential for classical astronomy. Despite attempts to link solar activity with terrestrial weather, it had less practical outcome than classical astronomy, which was the traditional basis for terrestrial measurement of time and position. Above all, it had no satisfactory theory to integrate it all together. Hardly surprisingly, many of the pioneer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

astrophysicists were amateurs, or came from other disciplines. Yet, within a few decades, astrophysics had become a subject in its own right: by the 1890s a major journal—the *Astrophysical Journal*—was already devoted to it (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS). Significantly, this was launched in the United States. North America had played only a minor role in the development of classical astronomy. As the new century dawned, it was widely accepted in the astronomical world that it would play a much greater part in the development of astrophysics.

Bibliography

Gingerich O (ed) 1984 *Astrophysics and Twentieth-Century Astronomy to 1950* (Cambridge: Cambridge University Press)

Hearnshaw J B 1986 *The Analysis of Starlight* (Cambridge: Cambridge University Press)

Jones B Z and Boyd L G 1971 *The Harvard College Observatory* (Cambridge, MA: Belkap)

Meadows A J 1970 *Early Solar Physics* (Oxford: Pergamon)

Meadows A J 1972 *Science and Controversy* (London: Macmillan) p 13

Jack Meadows

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Atacama Large Millimeter Array

The Atacama Large Millimeter Array (ALMA) will be the largest aperture synthesis telescope, operating at millimeter and submillimeter wavelengths, ever constructed. It is to be built on the plateau close to Cerro Chajnantor in the eastern Atacama desert of northern Chile, at an altitude of 5000 m. The project is being developed by an international consortium, which currently includes the US and nine European countries, and should be ready for scientific use in 2009.

The project is the natural successor to existing millimeter-wave interferometers at IRAM, OVRO, BIMA and Nobeyama. It will have 64 12 m diameter antennas, and will operate at wavelengths from 5–0.35 mm, allowing it to map at high resolution the molecular gas and dust in nearby star-forming regions, and star-forming galaxies at high redshift.

For further information see
http://www.mrao.cam.ac.uk/~jsr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Atmospheric Pressure

The pressure (i.e. the force per unit area) exerted on the surface of a planet by the weight of gas contained in a column extending vertically upwards to the limit of the atmosphere. At the surface of the Earth, the average value of atmospheric pressure, in SI units, is 1.013×10^5 pascal (Pa), where one pascal is equivalent to a force of one newton per square meter (N m^{-2}). A further unit of pressure is the bar, which is equivalent to 10^5 pascal, a common submultiple of which is the millibar (mb); 1 millibar $=$ 0.001 bar. A pressure of 1.013×10^5 Pa is equivalent to 1.103 bar, or 1013 millibars. The average value of atmospheric pressure at the surface of the Earth (1013 mbar) is also referred to as 1 atmosphere. By way of comparison, the atmospheric pressures at the surfaces of Venus and Mars are, respectively, 90 and 0.006 atmospheres.

See also: density, kinetic theory of gases.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Atmospheric refraction

Rays of light entering the Earth's atmosphere (i.e. passing from the vacuum of space to the medium of air) are bent, or refracted. As a result of this, the apparent positions of the stars are displaced by a small amount towards the zenith. In other words, the effect of atmospheric refraction is to increase the apparent altitudes of stars. At relatively small values of zenith distance, the effect is small, and is proportional to the tangent of the zenith distance. However, at large zenith distances, the effects are quite appreciable. At the horizon (zenith distance = 90°) refraction increases apparent altitudes by about 35 arc minutes. This means, for example, that at the moment that the lower edge of the Sun appears to touch the horizon at sunset, the whole disk of the Sun is actually below the horizon.

See also: aberration, scintillation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Atom

The smallest particle of a chemical element that retains the properties of that element. Each atom consists of a compact nucleus, in which all but a tiny fraction of its total mass resides, surrounded by a cloud of electrons—lightweight particles with negative electrical charge. A typical atomic nucleus has a radius of about 10^{-14} m; the overall radius of an atom (i.e. the radius of its cloud of electrons) is about 10 000 times larger (about 10^{-10} m). The nucleus consists of a number of protons (massive particles with positive electrical charge) together with a number of neutrons (particles of similar mass to the proton but with zero electrical charge), all of these particles being bound together by the strong nuclear interaction. The surrounding cloud of negatively charged electrons is held around the positively charged nucleus by the electromagnetic force. A complete (neutral) atom contains the same number of protons as electrons and, because it has equal numbers of positively and negatively charged particles, has zero net charge. An atom that has lost one or more of its electrons and which, therefore, has a net positive charge, is called a positive ion. An atom that has acquired one or more extra electrons and which, therefore, has a net negative charge, is called a negative ion.

Each element is assigned a chemical symbol that takes the form of a letter or a pair of letters; for example, hydrogen is denoted by H, helium by He, and so on. An atom is characterized by its atomic number (denoted by the symbol Z), which is the number of protons in its nucleus, and by its mass number (A), which is the total number of protons and neutrons contained in its nucleus. Conventionally, an element is denoted by its chemical symbol, preceded by mass number as a superscript and its atomic number as a subscript. Thus, an element, X, with atomic number Z and mass number A, would be written as ${}^{A}_{Z}\mathrm{X}$. For example, a normal hydrogen atom with one proton and no neutrons in its nucleus ($Z = 1$; $A = 1$) would be denoted by ${}^{1}_{1}\mathrm{H}$; a helium nucleus with two protons ($Z = 2$) and two neutrons ($A = 2 + 2 = 4$) would be denoted by ${}^{4}_{2}\mathrm{He}$, and so on. Atomic number defines the chemical element of which a particular atom is an example. Thus, for example, every hydrogen atom has $Z = 1$, every helium atom has $Z = 2$, and so on. Nuclei of the same chemical element may, however, contain different numbers of neutrons. Atoms that have the same number of protons but different numbers of neutrons, in their nuclei, and which, therefore, have the same atomic number but different mass numbers, are called isotopes. For example, 'heavy hydrogen', or deuterium, denoted by ${}^{2}_{1}\mathrm{H}$ (or ${}^{2}\mathrm{D}$) is an isotope of hydrogen that has one proton and one neutron, rather than just a single proton, in its nucleus. Helium exists in two forms, helium-4 (${}^{4}_{2}\mathrm{He}$) and helium-3 (${}^{3}_{2}\mathrm{He}$), a lighter isotope with two protons but only one neutron in its nucleus. The average of the various naturally occurring isotopes of a particular element, weighted according to their relative abundances, is the atomic weight of that element.

Although quantum mechanics has shown that electrons in atoms cannot be regarded strictly as points localized in space and that the electron clouds surrounding a nucleus represent regions where the probability of finding electrons is highest, the Bohr model of the hydrogen atom, which was proposed in 1913 by the Danish physicist Neils Bohr (1885–1962), continues to provide a good basis for describing the way in which atoms emit and absorb radiation. According to this model, a hydrogen atom consists of a single proton around which a single electron travels in one of a number of permitted orbits, each of which corresponds to a different energy level. If the electron drops down from a higher level to a lower one (makes a 'downwards transition'), it emits a photon (a quantity of radiant energy) with an energy equal to the difference in energy between the upper and lower levels. The wavelength of the emitted radiation is inversely proportional to the energy of the photon (the greater the difference in energy levels, the shorter the wavelength of the emitted radiation). Conversely, if an electron absorbs a quantity of energy equal in magnitude to the energy difference between a lower and a higher level, it will jump up to the higher level (makes an 'upward transition'). The emission and absorption of radiation that occurs when electrons make downward and upward transitions between the various energy levels of hydrogen atoms gives rise to the series of emission or absorption lines that constitutes the hydrogen spectrum. With more massive atoms, which have numerous electrons arranged in concentric shells, the set of energy levels and permitted transitions is more complex, but the same general principle applies: downward transitions give rise to the emission of radiation and upward transitions occur when energy is absorbed.

See also: absorption spectrum, electron, elementary particle, emission spectrum, fundamental forces, hydrogen spectrum, ion, isotope, molecule, neutron, Pauli exclusion principle, photon, proton, quantum mechanics, subatomic particles, wave–particle duality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aureole

Two Soviet–French satellites launched from Plesetsk in 1971 and 1973 to study Earth's upper atmosphere and auroras.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Auriga

(the Charioteer; abbrev. Aur, gen. Aurigae; area 657 sq. deg.) A northern constellation that lies between Perseus and Gemini, and culminates at midnight in early December. It is usually said to represent Erichthonius, legendary king of Athens and inventor of the four-horse chariot who was raised by the goddess Athene, and is shown on early celestial charts as a charioteer supporting a goat and its two kids on his left arm. Its brightest star, α Aurigae (Capella), is known as the 'little she-goat', and two of its other naked-eye stars, η Aurigae and ζ Aurigae, as 'the kids'. The brighter stars of Auriga were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large, conspicuous constellation, Auriga is easily recognized by virtue of Capella, which at magnitude 0.1 is the sixth brightest star in the sky and from northern temperate latitudes is almost immediately overhead on winter evenings. Other bright stars include β Aurigae (Menkalinan), an eclipsing binary (range 1.9–2.0, period 3.96 days), ε Aurigae, a remarkable eclipsing binary (range 3.0–3.8, period 27.1 years—the longest of any such star) and ζ Aurigae, yet another eclipsing binary (range 3.7–4.0, period 972 days). There are six other stars of magnitude 4.0 or brighter. The star Elnath, magnitude 1.7, was originally designated γ Aurigae, but has been reallocated to the neighboring constellation of Taurus (as β Tauri).

The Milky Way passes through the south-western part of Auriga and the constellation contains several open star clusters, including M36 (NGC 1960), which contains about 60 stars between ninth and fourteenth magnitude, M37 (NGC 2099), which contains about 150 stars between magnitudes 9 and 12.5, and M38 (NGC 1912), which has more than 100 stars fainter than eighth magnitude.

See also: Capella; Epsilon Aurigae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aurora

An aurora is an extended source of light of different forms and colors observable at high latitudes and sometimes at mid-latitudes. Their brightness may reach the intensity of the full moonlight. The auroral oval (one in each hemisphere) is the location where most of the auroras are observable. It is a belt surrounding each of Earth's magnetic poles with a maximum latitudinal extent at magnetic midnight. The polar cap auroras (inside the oval) and daytime auroras are usually weak and diffuse while in the oval, they are more intense and variable in time. In both cases, auroras are generated by particles of solar origin interacting with the atoms and molecules of the Earth's upper atmosphere through direct collisions and chemical processes. However, in the case of the auroral oval, the particles are accelerated by some complex magnetospheric processes leading to phenomena of brighter intensity, of various forms, both characters being highly time dependent. The auroras are usually observed between 80 and 300 km high depending on the characteristics of the precipitating particles (electron and proton energies) and atmospheric composition, leading to an auroral spectrum made of many lines from atomic and molecular neutral or ionized species. The occurrence and intensity of auroras follow the solar activity; for example the typical 11-year solar cycle is found in the observations. The aurora's characteristics represent the last stage in a series of processes starting in the solar atmosphere and taking into account the Earth's magnetosphere and atmospheric physics.

Brief historical background

When observing our environment, different phenomena are revealed. There are regular phenomena such as sunrise and sunset, movements of stars and planets. There are also irregularly occurring phenomena, either of astronomical or geophysical nature, such as rainbows, comets, halos and auroras. Before becoming predictable, their occurrence was associated with some important events. This is why these phenomena were recorded for divinatory purposes and also likely for curiosity.

In the first few centuries AD some auroras were recorded by the Chinese, Greek and Romans. However, people living in the high-latitude regions were more accustomed to their occurrence and they named them the Northern Lights.

The first modern observations of auroras started in the second half of the sixteenth century with Tycho Brahe from his observatory at Uraniborg. Sometime later the occurrence of auroras decreased, especially at mid-latitudes. P Gassendi described several auroras, observed in the early seventeenth century, and suggested the name 'aurora borealis'.

After the discovery of SUNSPOTS in 1609, the Sun was regularly observed. Sunspot occurrence experienced a significant decrease from 1645 to 1715, the 'Maunder minimum', during which few auroras occurred. Solar activity resumed in the early eighteenth century and some intense auroras were observed at mid-latitudes (1716 and 1726). E Halley and later J J Dortou de Mairan were asked to study their origin by the Royal Society (London) and the Académie Royale des Sciences (Paris) respectively.

Various hypotheses had previously been made to explain the occurrence of auroras. Mairan suggested in 1733 that the auroral light was generated by a solar fluid impinging upon the EARTH'S ATMOSPHERE. Even though approximate, this was the first hypothesis based on a relationship between the Sun and the Earth. In the second edition of his report, Mairan felt there was a link between sunspots and auroras as well as their variations in seasonal occurrence. The relationship between the occurrence of auroras and magnetic perturbations was experimentally demonstrated by P Hiorter in 1741 by observing the tiny perturbations experienced by a compass needle (provided by A Celsius) at the same time. Mairan suspected that auroras could occurr in the southern hemisphere and addressed a note to Don Ulloa, a Spanish naval officer, who confirmed observations of a polar light when rounding Cape Horn in 1745. J Cook exploring the high southern latitudes confirmed the occurrence of the 'aurora australis' in 1770. In the second half of the eighteenth century, W Wargentin (1717–83) concluded from various observations that auroras are observable from different places at the same time and extend horizontally as a belt surrounding the North Pole.

After the discovery of the electron, K Birkeland proposed in 1896 that auroras were produced by electrons ejected from the Sun and guided by the Earth's magnetic field to the high-latitude regions. He supported this idea by laboratory experiments. However, due to electrostatic repulsion, it was necessary to consider a PLASMA ejected by the Sun rather than electrons alone. S Chapman and V Ferraro showed that a moving plasma will interact with the Earth's magnetic field, confining it in a cavity. This was the birth of magnetospheric physics.

Several expeditions at the end of the nineteenth century localized the auroral zone with accuracy. Auroras were the subject of many observations (forms, spectrum, etc) and theoretical developments.

Main properties

The modern means of auroral observations are spectrometers which analyse their spectrum, photometers which study the behavior of one or several emissions lines, cameras which record the aurora forms as a function of time, and radar. These instruments can be used on the ground or from orbiting platforms where they provide a global survey of auroras. However, before the instrumental study of auroras, observations were made visually. In order to allow comparisons of observations made by different authors, a nomenclature to describe auroral characteristics (intensity, form and color) was adopted internationally. It is still in use.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Intensity

The line of the greatest intensity is due to the emission of atomic oxygen at 557.7 nm frequently referred to in the scientific literature as the 'green line'. Auroral brightness is scaled to the green line intensity through the International Brightness Coefficient (IBC) as shown in table 1.

Table 1. The unit used is the rayleigh (R) which corresponds to 10^6 photons emitted in 4π steradians and received per cm^2 and per second. This scale was proposed by Seaton in 1954 and formalized by Hunten in 1955.

IBC	Intensity (kR)	Equivalent source
I	1	Milky Way
II	10	Thin moonlit cirrus clouds
III	100	Moonlit cumulus clouds
IV	1000	Full moonlight

For the purpose of comparison, equivalent brightness sources are indicated. For the modern measurements, auroral emission intensities are given in rayleigh. However, when using old records for long-term studies, the reference to the equivalent source intensity is useful despite its relative inaccuracy.

Forms and colors

Visual observations and images recorded by cameras reveal several typical forms as shown for example in figure 1. They are accurately described as arcs, bands, rayed or not, diffuse or pulsating surface, corona and flaming. The description of auroral forms is based on an international nomenclature shown in table 2.

Auroral colors are usually green, red or blue depending on the atmospheric species involved in the excitation processes. Auroras of intensity IBC I or II appear colorless. Yellow auroras are due to the blending of red and green auroras.

Auroral intensity is a function of time. The change may be either slow, as for a pulsating surface (0.1 Hz), or reach 10 Hz for certain homogeneous forms. Changes in time may be due to traveling forms as for the flaming-type auroras which appear to move upwards. Apparent horizontal movements are found in arc or band auroras with velocities as great as 100 km s^{-1}. This apparent velocity generally increases with the amplitude of the magnetic storm. When making visual observations of such events, the rapidity of the changes with time is certainly the most striking and spectacular feature of the auroral phenomenon.

Auroral events are due to precipitations of particles in the Earth's upper atmosphere. Their patterns are complex and this is why the description of an aurora needs to be accurately defined. Based on that, it has been possible to classify the auroral events and to obtain information about their generation processes.

Location

When observations are made on the ground as a function of latitude, a mean frequency of occurrence can be determined. It allows definition of the region of maximum occurrence now known as the 'auroral zone'. The auroral zone and auroral oval are two different concepts which will be explained later. Pictures of the auroral oval were obtained from orbit by the DYNAMICS EXPLORER 1 spacecraft (figure 2). The auroral oval center is shifted by about 3° from the geomagnetic pole towards the dark hemisphere. Its thickness is a maximum at magnetic midnight. The auroral oval lies around 70° magnetic latitude and has a mean width of about 5° in quiet conditions. When auroras are observed at the same time from two magnetically conjugate locations (one in the northern and the other in the southern auroral ovals), they develop similar patterns.

The occurrence of auroras is greater in the auroral oval than in its central part. This latter region is named the POLAR CAP and auroras have a weaker intensity and a smaller rate of change with time than in the auroral oval. Sometimes extraordinary auroral displays may be seen inside. For example, an arc observed near the zenith remains co-aligned with the Sun's direction. This was discovered by D Mawson in 1913 at the Dumont d'Urville station (Antarctica). This arc may extend up to both inner edges of the auroral oval. It was named the theta aurora as it looks like the Greek letter Θ.

The above auroras are observed at night. Another type of aurora is observable in the daytime. They are usually weak and diffuse.

Height and spatial extent

Heights were first measured using parallactic methods from the ground. Today, measurements from rockets and spacecraft are commonly used. Most auroral occurrence takes place in the altitude range 80 to 120 km. However, the lower border of very intense auroras may decrease to 70 km. Their vertical extent depends on the generation processes and may reach an altitude of 300 km for certain aurora. The intensity distribution is not homogeneous along the vertical; its vertical structure is made of elements having a thickness of hundreds of meters to some kilometers.

Auroras remain in a latitude strip, the thickness of which is function of the magnetic time (defined as the solar local time, but counted from the meridian containing the magnetic and geographic poles) as seen in figure 2. Within the auroral oval, the quiet arcs are mostly oriented in longitude along the oval and may be 500 km long.

Occurrence periodicity

For a given ground-based station within the auroral zone, the yearly mean occurrence follows the 11-year SOLAR CYCLE, with, however, a time lag of one to two years.

When using several years of data sorted by monthly mean, a semi-annual variation is shown with maximum auroral occurrence near the equinoxes. Similarly, the 27 day solar rotation can be pointed out. The occurrence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. Auroral display observed at Dumont d'Urville (Antarctica). An all-sky camera of the Cassegrain type is also shown.

Table 2. Common auroral forms from Vallance-Jones (1974).

Homogeneous arc (HA)	Regular form in height, direction and intensity
Homogeneous band (HB)	Similar to HA, but variable along length
Rays arc or band (RA or RB)	Narrow with luminosity quasi vertically structured
Diffuse surface (DS)	Uniform emitting area
Pulsating surface (PS)	As DS, with intensity variable in time (several seconds)
Pulsating arc (PA)	As PS for an arc
Corona (C)	Ray system viewed parallel to the magnetic field lines
Flaming (F)	Forms of variable intensity moving towards zenith

of polar cap and daytime auroras does not strictly follow these rules, especially for the most intense events.

Relationship with magnetic activity

The Sun is a variable star. It experiences changes more or less cyclically, with some others of a more chaotic nature. Its variable activity induces fluctuations in the SOLAR WIND characteristics (velocity, density, ion and electron temperatures and magnetic field) of greater or lesser importance. It results in small fluctuations in the Earth's magnetic field which are detected by magnetometers on the ground. Their amplitudes are used to define indices of magnetic activity. These events are called magnetic storms. Auroral occurrence and intensity have been studied as a function of magnetic disturbance and the corresponding auroral oval has been determined. From a typical shape (figure 2), with increasing magnetic activity, the oval shifts slightly equatorward and extends polarward and equatorward. This explains why some auroras may be observable at mid-latitudes.

Magnetic activity follows the 11-year solar cycle pattern. However, given the chaotic nature of the Sun, strong magnetic events may occur even during low solar activity.

Electron and proton auroras

Auroras result mainly from collisions between atmospheric species and particles, electrons or protons. Most of the properties discussed above are those of electron auroras.

Differences between electron and proton auroras arise from their different mode of propagation in the upper atmosphere. Due to the proton mass, its radius of gyration is much greater (200 m) than the electron one. Furthermore, protons moving around the magnetic field may collide with some species and capture an electron. This results in an atom of hydrogen which moves in a straight line, free from the influence of the Earth's magnetic field. Another collision may later restore the initial proton. Consequently, proton auroras are of weak intensity (maximum IBC I) and diffuse. The height of occurrence is around 100 km. The proton auroral oval is shifted slightly equatorward with respect to the electron auroral oval.

The auroral spectrum and emission mechanisms

Most of the auroral displays are green due to the $O(^1S)$ emission line at 557.7 nm. However, many other lines from either atoms or molecular or ionized species exist. Figure 3 shows a spectrum between 450 and 750 nm. It

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

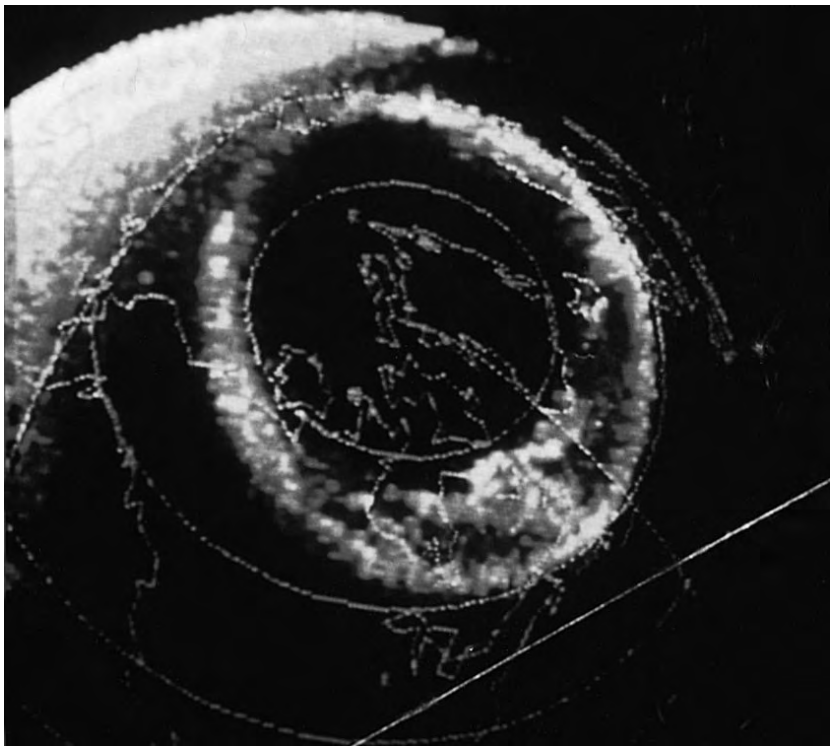


Figure 2. The auroral oval observed by the Dynamics Explorer 1 satellite (Frank *et al* 1982).

is a superposition of many lines from: O and N; O_2, N_2, NO and OH; O^+, N^+; and O_2^+, N_2^+. The blue color originates from the N_2^+ band at 391.4 nm. Figure 3 also shows the Balmer lines emitted during the proton events. The auroral spectrum arises from collisions between electrons/protons with the species in the Earth's atmosphere, and photochemical reactions:

(a) Collisions with electrons. Primary electrons from the magnetosphere are impinging upon the atoms (A) and molecules (M) of the terrestrial atmosphere and set these species in metastable states (*), such as

$$A + e \rightarrow A^* + e$$
$$M + e \rightarrow M^* + e$$

with similar reactions leading to ions. The electrons on the left-hand side are the primary electrons and those on the right-hand side are the secondary electrons which are also able to collide with other species and to generate excited molecular and atomic states until they are fully thermalized. The excited states (*) are deactivated either by radiating photons or by collisions with other species before light emission (quenching). As the density of the neutral species is increasing downward, quenching explains why no auroras are observed at low altitude (typically below 70 km).

(b) Indirect excitation mechanisms. The two most important mechanisms are:

(1) Dissociative recombination; a typical example is

$$O_2^+ + e \rightarrow O(^1D \text{ or } ^1S) + O.$$

(2) Energy transfer and quenching; a typical example is

$$O(^1D) + N_2 \rightarrow O(^3P) + N_2\dagger.$$

The symbol † denotes either an excited state or the ground state. For the first case there is an energy transfer, and for the second case there is a collisional deactivation. The efficiency of the quenching is a function of altitude and the lifetime of the excited state of the colliding species.

(c) Collisions with protons. Reactions involve protons and neutral hydrogen colliding with molecules. During these collisions, neutral hydrogen may lose its electron, ionize the molecule or transfer its energy to it. As for the proton, it may also ionize the molecule, transfer its energy or capture an electron.

To calculate the various emission line intensities, it is necessary to dispose of the energy spectrum of the primary electrons. This is generally measured at high altitudes (typically 300 km). The calculation of the auroral spectrum is based on the energy spectrum of the primary electrons, the cross section of the atoms and molecules,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

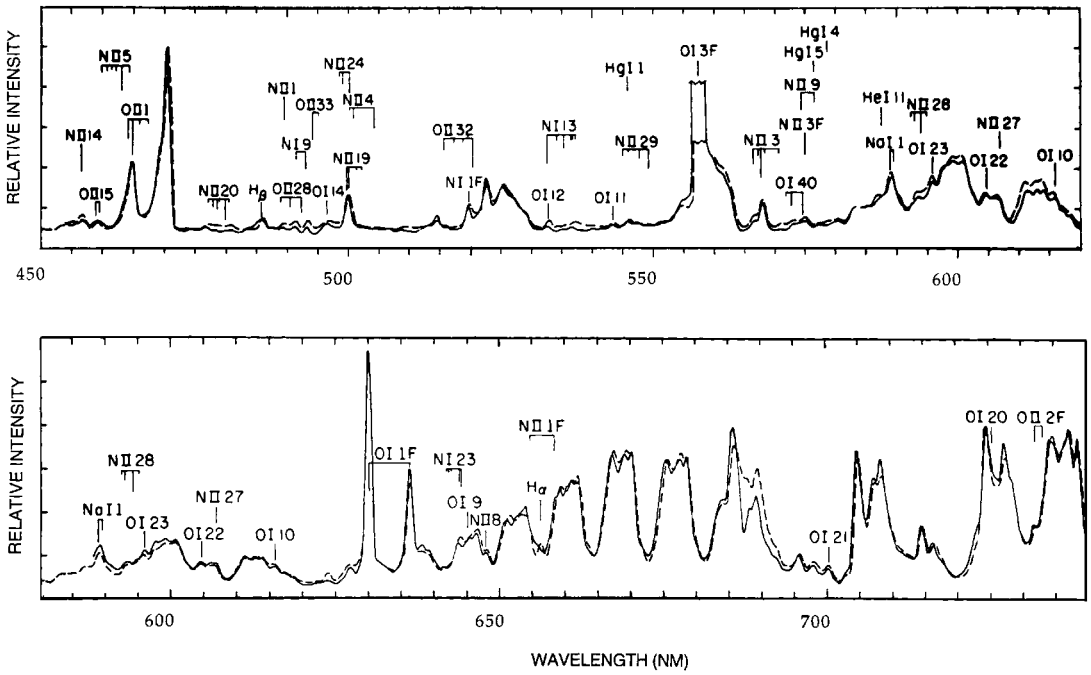


Figure 3. An auroral spectrum from 450 to 740 nm at 1 nm resolution. The observations (solid curve) are compared with the synthetic spectrum of Gattinger and Vallance Jones (1974). The green line intensity is in the range of an IBC II to III.

and the processes occurring as the electrons penetrate the atmosphere in a cascade of successive collisions. The problem is rather difficult and also suffers from the uncertainties affecting the cross sections.

Visual observations, as well as instruments on board rockets and spacecraft, have shown the dependence of the line emission intensity on altitude. Let us consider the case of the atomic oxygen emission at 630 nm. The mechanism of emission (dissociative recombination and excitation by collision), the vertical distribution of the neutral species and the energy spectrum of the precipitating electrons allow us to calculate the production rate of the excited state (^{1}D) as a function of altitude. Taking into account the deactivation of the excited oxygen atoms by collisions provides the emission rate profile with altitude. In the case of this emission, the 110 s lifetime of the ^{1}D excited state suppresses all emission at low altitude. This is why the aurora observed at 630 nm will be found around an altitude of 250 km. The same process allows us to calculate the altitude profile of the green line atomic oxygen ^{1}S excited state. Its lifetime, being 0.7 s, avoids the low-altitude deactivation and this emission will observable as low as 80 km depending on the intensity of electron precipitation.

For a given emission line, the stronger the electron precipitation is, the lower the aurora will be.

Aurora and the Earth's magnetosphere and other related phenomena

It is suggested that the reader consults the sections dedicated to the Earth's magnetosphere in order to have in mind its representation as well as the terminology (see MAGNETOSPHERE OF EARTH articles).

As the Sun is a variable star, the solar wind presents different regimes which affect the state of the Earth's magnetosphere. The interplanetary magnetic field (IMF) experiences changes with time. In the quiet condition, the IMF has a northward component (i.e. parallel to the terrestrial magnetic field) which turns southward during magnetic storms.

To investigate the mechanism of auroral emission, the spectrum of the precipitating particles has been measured during auroral display. For daytime and polar cap auroras, it was found to be of the order of the energy of the solar wind. Consequently, daytime auroras are believed to be generated by solar wind particles entering the upper atmosphere via the northern and southern dayside cusps where a neutral point exists in magnetically quiet conditions (northward IMF). Polar cap auroras are generated by solar wind particles from the magnetotail.

Within the auroral oval, particle energy is of a factor of 50 or more greater than in the solar wind. Consequently, the fundamental question is the acceleration mechanism of the particles. This is still being debated and cannot yet be considered as fully understood. However, reconnection between the Earth's magnetic field and the IMF is the most probable mechanism. A reconnection occurs when the IMF has a southward component (antiparallel to the terrestrial magnetic field). In this case, it generates a neutral sheath where the plasma jets are injected upstream and downstream from that region. It results in a thinning

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the plasma sheet and an inward movement of the particles. The shrinking of the plasma sheet leads to an increase in the energy of the particles. Several processes are proposed. The most likely is based on adiabatic compression of a certain quantity of plasma being confined in a compressed magnetic flux tube as it is convected forward until it is finally precipitated into the Earth's upper atmosphere just as the system becomes unstable.

The auroral oval is then the intersection of the plasma sheet with the upper atmosphere in the 100–200 km altitude range. The auroral oval is fixed with respect to the Sun, but a ground-based station at auroral latitudes rotates below, around the geographic pole. As a function of time, it may be in the polar cap, inside or outside the oval, experiencing different auroral occurrence regimes. Their mean defines the auroral zone as seen from the ground.

When the magnetosphere is moderately disturbed due to small IMF changes, the corresponding events are called magnetic substorms, which are made up of four distinct phases (quiet, growth, expansion and recovery) having specific characteristics such as typical changes in the Earth's magnetic field. Major solar events such as FLARES generate magnetic storms which are analysed as a series of successive substorms.

During a magnetic storm, the moving particles generate a magnetic field which combines with that of the Earth. The variable characteristics of the precipitating particles (velocity, direction and number density) and the collisions with the Earth's atmospheric species are the cause of the complexity of the magnetic field disturbance. Consequently, auroras constitute a visual illustration of the perturbations of the Earth's magnetic field. In certain cases, the field pattern is simple, as for example the rays which appear as nearly vertical columns of light showing the high inclination of the field line at the latitude of the auroral zone.

Associated with auroras are many other phenomena such as the auroral electrojet circulating in the E-region (around 100 km altitude) which is heated by Joule dissipation and a ring current at low latitudes. Also, x-rays and radio waves are generated in the range of 40 to 500 MHz. During auroras, electric fields are created, acting on the movement of the electrons, protons and ions. Furthermore, all these species are moving in a neutral atmosphere which is driven by collisions with the ions. Consequently, dynamics are strongly involved in the auroral events as illustrated by the development of vertical neutral winds during the substorms.

The use of auroral observations

The link between auroral occurrence and solar activity has been statistically established from observations. This is why the occurrence of auroras has been used to reconstruct the solar activity before the first instrumental observations. The long periods in the solar activity found in the ^{14}C variations were confirmed by the occurrence of auroras. More directly, auroral occurrence represents the ultimate stage in the various phenomena taking place from the solar wind up to the Earth's atmosphere. Understanding the detailed auroral characteristics (occurrence, forms, spectrum) ensures that the entire physical process is understood. The same applies to auroras observed in the atmosphere of Jupiter and Saturn.

Auroral data show how important in science are a long series of carefully made observations.

Bibliography

Bone N 1997 *The Aurora* (New York: Wiley)

Brekke A 1997 *Physics of the Upper Polar Atmosphere* (New York: Wiley)

Brekke A and Egeland A 1983 *The Northern Light* (Berlin: Springer)

Eather R H 1980 *Majestic Lights* (Washington, DC: American Geophysical Union)

Frank L, Craven J D, Burch J L and Wimingham J D 1982 Polar views of the Earth's aurora with Dynamics Explorer *Geophys. Res. Lett.* **9** 1001–4

Gattinger R L and Vallance Jones A 1974 Quantitative spectroscopy of the aurora. II : The spectrum of medium intensity aurora between 4500 and 8900 Å *Can. J. Phys.* **52** 2343–56

Vallance Jones A 1974 *Aurora* (Dordrecht: Reidel)

Gérard Thuillier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Aurorae: Observing by Amateurs

Amateur scientists have had a long interest in the sky, so there has been an amateur involvement with observing the aurora for as far back as records exist. In the modern era, amateurs have been credited with making significant contributions to investigations of the aurora since advancements in many sciences began to escalate during the 18th century. Because the aurora is a large-scale, global phenomenon, the value of international collaborations involving both professionals and amateurs was recognized early. One of the most successful demonstrations of such cooperation occurred during the International Geophysical Year (1957/58).

With the advent of the space age, investigators have come to rely more on satellites and sophisticated ground-based observatory equipment to monitor the aurora, but some value is still placed on visual reports to supplement these. An important purpose of the visual observation is in continuing the written records of auroral activity that have been preserved by several civilizations since Greek times.

A minimum of equipment is required to make useful visual observations of the aurora, and the only optical instrument needed is the unaided eye. To a certain extent, it is possible to forecast the occurrence of auroral displays based on recent events on the Sun, so that observing efforts can be planned to coincide with the nights when a display is likeliest to occur. Magnetometers and short-wave radios can also be helpful in alerting the observer to the presence of auroral activity (although there is not always a visual counterpart). You will see auroral displays most often, however, if you persevere with making regular visual inspections of the poleward sky throughout the night, between sunset and dawn, whenever weather and moonlight permit.

Amateur observers living in cities or where there is frequent cloud cover may not be able to observe the aurora visually very often, but they can nonetheless submit an observation that auroral activity was detected by a magnetometer or short-wave radio. In this case, the observation is not confined to the night-time hours as for the purely visual observer.

No special equipment is needed for basic aurual observations. It can be useful, however, to have an instrument, such as a simple theodolite or alidade, to accurately determine the bearings and heights of features in the aurora. As an adjunct to the visual record, amateur observers sometimes detect subvisual details in auroral displays with photography or CHARGE-COUPLED DEVICE (CCD) imaging.

Photography is a popular way of recording an aurora and is the ideal means by which to share the spectacle of a display after the event (figures 1 and 2). All that is needed is any camera that can make time exposures of a few seconds to a few minutes, with the lens open to its widest aperture—many photographers choose fast color film (such as ISO 400), and a fast, wide-angle lens (such as $f/3$ focal ratio and 28 mm focal length). Photographs also make a permanent record of the position and structure of a display, which may be useful for triangulation work. For those interested in applying other techniques to expand on their observations, videography, CCD imaging and spectroscopy may also be employed.

As a means of identifying the presence of, or forecasting the likelihood of, auroral activity, observers with an interest in electronics may use magnetometers or short-wave radios to detect disturbances in the geomagnetic field or the ionosphere.

Contributors are encouraged to submit observations of aurorae in a format recommended by international agreement, which increases the value of a sighting and also the observer's appreciation of the phenomenon. A record includes the observer's location and observing

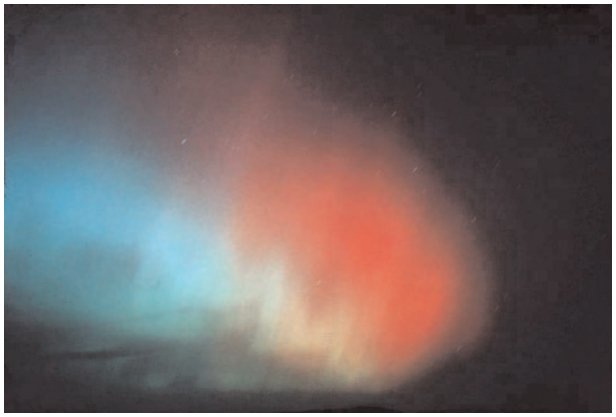


Figure 1. An aurora seen over Hobart, Tasmania, on 23 October 1981. Photographed by Martin George, Launceston Planetarium, Tasmania.

Figure 2. An aurora seen over northern Tasmania on 31 March 2001. Photographed by Karenne Barnes, Astronomical Society of Tasmania Inc.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

conditions, the date and time that a display was seen, plus details about brightness, color, form, activity, directions and elevations of features, and changes that occur. This information can be recorded as notes or diagrams, reported to varying degrees of detail according to the individual observer's preference or experience.

Any photographs or other types of observation of the aurora should also be accompanied by a record of at least location, date and time, plus relevant technical information, to give them scientific value. This also increases the interest and usefulness of the record to other observers.

To make the best use of observations, results can be contributed to one of several well-established organizations that collect, analyse and archive the reports for future use.

Initial analysis of reports usually includes a comparison with the forecasts and a correlation of the auroral event with activity on the Sun and in the Earth's geomagnetic field. For the longer term, it is interesting to determine the frequency of displays seen from particular locations and their frequency under certain conditions or at particular times of the year or the sunspot cycle.

An aurora reported by a number of observers from different locations allows that aurora's shape, size, position, development and duration to be described. Results shared between regional and international organizations make it possible to chart an entire auroral event, across the daylight boundaries and between hemispheres, so that the global pattern of auroral activity can be studied. The visual observations can ultimately be compared with images and other data about the auroral oval obtained from satellites and professional observatories.

The British Astronomical Association (BAA) collects observations of the aurora through its Aurora Section and has a network of observers who include members of the BAA or other astronomical societies, individual observers, professional meteorologists and ships' officers at sea. The BAA collects and publishes observations, mainly of the mid-latitude storm aurora, from the British Isles, Europe, Iceland, Canada and the United States of America; they are also published by the British Meteorological Office. In the southern hemisphere observations are coordinated by the Royal Astronomical Society of New Zealand (RASNZ), which receives and publishes observations from New Zealand and Australia. There is close cooperation and exchange between the BAA and the RASNZ.

Reports coordinated by the BAA and the RASNZ are then archived at Aberdeen University, Scotland, where they are available to researchers. There are similar archives in such countries as China, Korea and Japan.

Radio observations of the aurora are coordinated by the Radio Society of Great Britain and the American Radio Relay League in the USA.

Professionals draw on the data contributed by amateurs; for example, amateurs' results have helped to reveal processes in the thermosphere. Amateur observers are also important contributors to ongoing investigations into the complex correlation between the aurora and noctilucent clouds.

Auroral processes and the complex interactions that cause them are not yet fully understood, so the aurora is still under vigorous investigation. Since scientists no longer record the visual aurora, the only source of such observations is the amateur astronomical community. Records of the visual aurora have been kept for hundreds of years, and in the past they have provided valuable insights into solar activity and the Earth's upper atmosphere—there are extensive archives yet to be exploited that may be of similar importance in new studies in the future. Some of the areas that may make use of records of the visual aurora include geomagnetism, solar and interplanetary magnetism, terrestrial weather, the upper atmosphere and space weather.

See also: RADIO ASTRONOMY FOR AMATEURS and AMATEUR OBSERVATIONS OF THE SUN.

Bibliography

Bone N 1996 *The Aurora: Sun–Earth Interactions* 2nd ed (New York: Wiley)

Davis N 1992 *The Aurora Watcher's Handbook* (University of Alaska)

Further information is available at

American Radio Relay League webpage
http://www.arrl.org/

British Astronomical Association Aurora Section webpage
http://www.baa-aurora.fsnet.co.uk

Radio Society of Great Britain 'Auroral propagation' on the *The Radio Propagation Page* webpage
http://www.keele.ac.uk/depts/por/psc.htm#au

Royal Astronomical Society of New Zealand Aurora and Solar Section webpage
http://homepages.ihug.co.nz/~Sbevans/Aurora_Sun/aursun.htm

Karenne Barnes

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Auroral Kilometric Radiation

The Earth is a powerful natural radio source that emits a bright emission at kilometric wavelengths: the auroral kilometric radiation (AKR). As indicated by its name, the source regions of the AKR are at high latitudes, at altitudes ranging from typically 0.5 to 2 Earth radii. Its maximum frequency (of the order of 700 kHz) is smaller than the ionospheric cut-off frequency and therefore this radiation cannot be observed from the ground. It was thus discovered and studied after the beginning of the space era (Benedictov *et al* 1965).

The AKR is organized in radio storms lasting a few 10 min and having power as strong as a few megawatt. This sporadic activity is correlated with periods of intense auroral particle precipitations during which energetic electrons with typical energies of a few keV are observed flowing down along the auroral magnetic field lines and forming the aurora borealis (see AURORA). These precipitating electron populations result from a complex chain of physical processes linked to the major disruptions of the magnetospheric tail known as magnetospheric substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS). The AKR is thus directly related to the global magnetospheric activity and can even be considered as an indicator of this activity. Concerning its intrinsic properties, the AKR is characterized by its extreme brightness temperature (up to 10^{20} K) which is completely unexplainable by classical radiating mechanisms (as gyrosynchrotron radiation). Coherent processes, related to the development of plasma microinstabilities, must then be invoked for explaining its generation. In addition to this extreme brightness, a good theory of the AKR must also describe the polarization of the emission and its strong temporal and spectral organization.

The AKR is not by itself a particularly powerful phenomenon. It is even neglectable in the global energy budget of magnetospheric activity. However, it presents some astrophysical interest since it belongs to the wide class of radio emissions directly related to the development of acceleration mechanisms in the collisionless plasmas that surround most astrophysical objects (see COLLISIONAL PROCESSES IN ASTROPHYSICAL PLASMAS). Radio emissions presenting characteristics comparable with the AKR have indeed been observed coming from astrophysical objects as different as the other magnetized planets of the solar system, the Sun, some stars and binary systems (see MAGNETOSPHERES: JUPITER, RADIO EMISSIONS, RADIO EMISSION FROM BINARY STARS). As its sources can be directly explored by spacecraft, the AKR thus constitutes a very interesting case of an astrophysical phenomenon for which theoretical models can be checked using *in situ* measurements. Most of the above-mentioned radio emissions probably originate from the very site of particle acceleration and their analysis can serve as a diagnostic tool for the study of the energy release in plasma environments. In this context, the understanding of the generation mechanism of the AKR is crucial. The observations reported below concern this particularly important aspect of the problem.

Physical conditions within the sources of AKR

An example of the precise wave observations performed near and inside the source of the AKR is given in figure 1. These measurements have been obtained by a Swedish spacecraft—Viking—specially designed for the analysis of the acceleration and radiating mechanisms taking place in the auroral zones of the MAGNETOSPHERE OF EARTH. When it made this observation, Viking was at an altitude of the order of 10 000 km, at a magnetic latitude close to 70° North, around midnight local time, and crossed a region magnetically connected to a discrete auroral arc. In the figure, a time–frequency diagram of the electric component of the electromagnetic waves is presented. The intensity is given by a color code (from blue to red and white as the intensity increases from less than 10^{-8} V m^{-1} Hz$^{-1/2}$ to more than 10^{-5} V m^{-1} Hz$^{-1/2}$). The frequency range is 10 kHz–500 kHz and 20 min of data are shown. A white line indicates the electron gyrofrequency (f_{ce}) measured by the spacecraft. The AKR is observed above this characteristic frequency (here around 150 kHz), forming V-shaped emissions. Its lower frequencies which correspond to a clear cut-off has a minimum around 22.39 UT. At that time, the AKR frequencies extend down to f_{ce}. This region of minimum frequencies is also the region of maximum level of the waves (white intensity code). This is the source of the AKR. The V shape of the AKR dynamic spectra is easily understood if one assumes that the waves are generated at f_{ce} (thus at increasing frequencies as the altitude decreases) in a source elongated along the magnetic field, at oblique angles from the magnetic field lines and in the upward direction. Far from the source (for example at 22.35 or 22.45), only waves generated well below the spacecraft (thus at high frequencies) are detected. As the spacecraft approaches the source, waves generated closer and closer to the spacecraft and thus with lower and lower frequencies are detected, until the spacecraft crosses the source which corresponds to a maximum intensity and minimum frequencies. On this spectrogram, one can also observe a systematic modulation of the AKR, in the whole frequency range, at a period of 20 s. This modulation is linked to the spin of the spacecraft (the electric antenna are then alternately parallel and perpendicular to the magnetic field) and is an indication of the strong polarization of the waves. This can be used to identify the mode of propagation of the AKR. One finds that it is the so-called X mode, an electromagnetic supraluminous mode that has a right-handed circular polarization (the electric component of the waves rotates in a plane perpendicular to the magnetic field in the same sense that the electrons).

In the lowermost part of the spectogram, a broadband emission (the auroral hiss) is observed. The upper-frequency cut-off of this emission (yellow to green transition on the spectrogram) is the local plasma frequency. It is of the order of 30 kHz indicating that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

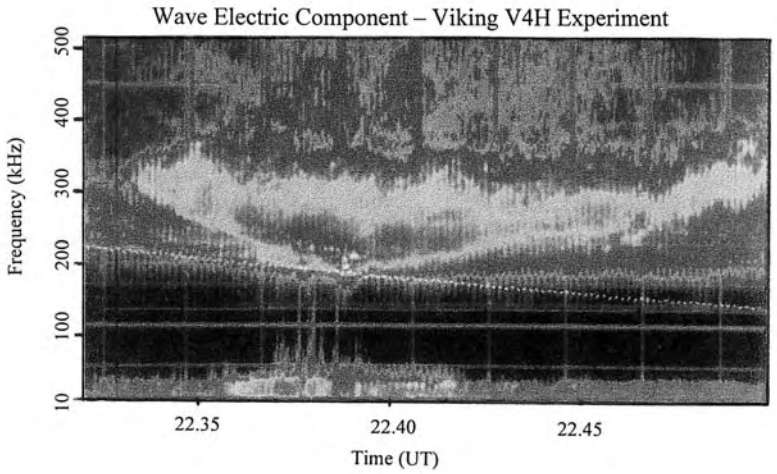


Figure 1. This figure is reproduced as Color Plate ??.

plasma density is typically about 10 cm^{-3}. Just around 22.39, during the crossing of the AKR source, this cut-off disappears. It is in fact below the frequency bandwidth of the analyser (10 kHz). The source of AKR is then a region of significant plasma depletion, the typical density varying from typically 10 cm^{-3} to less than 2 cm^{-3} over a short distance (a few km). The sources can then be seen as 'curtains' of depleted plasma, with rather large vertical and East–West extensions but very thin (less than 50 km) in the North–South direction.

Particle measurements are important for an understanding of the generation mechanism. In figure 2, time–energy spectrograms for the electrons (upper panel) and the ions (lower panel) are presented. The energy goes from 40 eV to 40 keV and the intensity of the flux is given by a color code. Because of the spin of the spacecraft, the particle fluxes are measured at variable angles with respect to the magnetic field (sawtooth below the proton fluxes). 0° corresponds to the direction parallel to the magnetic field, oriented towards the Earth (precipitating particles). For the data presented in figure 2, we know from wave measurements (not shown here) that a source of AKR has been crossed between the two arrows. This time interval corresponds to the detection of ion beams that propagate upward with typical energies of the order of 4 keV. Simultaneously, the maximum energy of the electrons decreases from 8–10 keV (before the first arrow and after the second one) to typically 4 keV (between arrows). The most simple explanation of these observations is that, when entering the AKR source, the spacecraft also enters a region of particle accelerations that corresponds to the existence of a potential drop along the magnetic field. Just before or after the source crossing, this potential drop is entirely above the spacecraft so that precipitating electrons of maximum energy are detected. Inside the source, a part of the potential drop is now below the spacecraft; the energy of

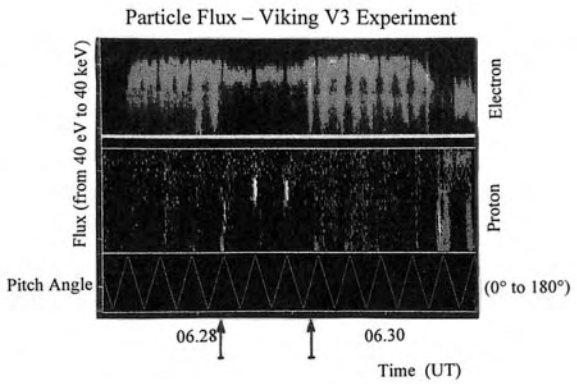


Figure 2.

the electrons decreases but energized ions coming from below are observed. This constitutes one of the fundamental observations concerning the AKR: the sources of this radiation are also the sites of the particle acceleration (Louarn *et al* 1990). This was recently confirmed by observations of the FAST spacecraft (Ergun *et al* 1998).

A more precise study of the fluxes of particles can help us to understand the type of distribution functions that are responsible for the wave generation. These distributions are characterized by expanded loss cones (large gap in the flux of upward electrons) and an accumulation of electrons at rather large perpendicular and small parallel velocities. The formation of such distributions can be explained by the creation of the parallel electric field responsible for the particle acceleration. This also corresponds to the evacuation of the low-energy plasma and to the formation of the region of density depletion. These two characteristics of the distribution functions, accumulation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of electrons at high perpendicular velocities and low densities, are of fundamental importance for the most plausible generation mechanism of the AKR: the cyclotron maser instability.

The cyclotron maser theory

Many theories have been proposed in the past to explain the generation of the AKR (linear mode conversion of electrostatic upper hybrid waves, non-linear beating of electrostatic waves, radiation from collapsing cavitons). It is now widely accepted that a particular microscopic instability—the cyclotron maser instability (CMI; first proposed by Wu and Lee (1979))—adequately describes the phenomenon. This theory is based on a resonant interaction between the electric field of the waves and the gyrating electrons and describes a transfer of angular momentum from the electron population to the wave electromagnetic field. The success of this theory is explained by the fact that it takes into account some of the most important experimental results presented before. The angular momentum transfer is particularly important if the electric field follows the electron motion, i.e. if the waves present a right-handed circular polarization (which is the case of the X-mode waves). Moreover, this transfer is to the advantage of the waves if the plasma presents an excess of angular momentum with respect to the thermal equilibrium. Such an excess can correspond to the existence of positive $\partial f/\partial P_\perp$ slopes in some region of the momentum phase space (where $P_\perp$ is the component of the particle momentum perpendicular to the magnetic field) which is precisely what is observed inside the source regions of the AKR. This mechanism of generation is then just a particular example of wave excitation by an out-of-equilibrium plasma characterized by an inversion of population in 'perpendicular' energy.

This mechanism can be described in a more mathematical way using the Maxwell–Vlassov system of equations. If one retains only the most important interaction between the waves and the electrons (pure cyclotronic motion and interaction with a homogeneous electric field that rotates in the right sense at a frequency close to the electron gyrofrequency), the system can be written as

$$\frac{\partial \boldsymbol{E}_1}{\partial t} = -\frac{\boldsymbol{J}_1}{\varepsilon_0}$$

$$\left(\frac{\partial}{\partial t} + \frac{\omega_c}{\Gamma}\frac{\partial}{\partial \theta}\right) f_1 = \frac{e}{\sqrt{2}}(E_{1+}\,e^{-i\theta} + E_{1-}\,e^{+i\theta})\frac{\partial f_0}{\partial P_\perp} + eE_\parallel \frac{\partial f_0}{\partial P_\parallel}. \quad (1)$$

Here f_0 is the background distribution function, E_1, J_1, f_1 are the perturbed electric field, current and distribution function, respectively, the subscripts $+$, $-$ and $\parallel$ refer to left-, right-handed and parallel components, θ is the Larmor angle of the electrons, Γ their Lorentz factor $((1-v^2/c^2)^{-1/2})$ and ω_c the electronic gyrofrequency. From a temporal Fourier transform of equation (1), a simplified dispersion equation is obtained:

$$\omega = \pi\omega_P^2 \iint \frac{dP_\parallel P_\perp^2 dP_\perp}{\Gamma} \frac{1}{\omega - \omega_c/\Gamma}\frac{\partial f_0}{\partial P_\perp}. \quad (2)$$

The growth rate of the waves is then easily calculated. It is proportional to to integral of $\partial f_0/\partial P_\perp$ along the resonant curve defined by $\Gamma\omega = \omega_c$. This means that the maximum transfer of momentum occurs if the wave frequency exactly matches the gyrofrequency of the electrons. A slightly more complex calculation gives the exact resonant condition: $\omega - k_\parallel v_\parallel - \omega_c/\Gamma = 0$. For $\partial f_0/\partial P_\perp > 0$ along the resonant curve, the absorption coefficients can become negative, thus leading to an instability: the so-called CMI. The exact and complete calculation makes the dispersion equation more complex but does not change its physical meaning.

Two physical processes could be responsible for the formation of $\partial f_0/\partial P_\perp$ non-thermal features in the distribution function: (1) the ionospheric absorption of the electrons which produces loss cone distributions functions and (2) the adiabatic evolution of electron populations accelerated downward along the magnetic field lines. Because of the mirror force, their perpendicular energy indeed gradually increases at the expense of their parallel energy such that a local excess in perpendicular momentum can be formed. The presence of an accelerating electric field also favors the occurrence of the maser instability in another way: it evacuates the cold electrons, a plasma component that always contributes to stabilize the CMI. For the X modes to be unstable, it is indeed necessary that, roughly speaking, $(\omega_{pc}/\omega_c)^2 = W/2mc^2$, where ω_{pc} is the cold plasma frequency and W the typical energy of the electrons—with rest mass m—for which $\partial f_0/\partial P_\perp$ (Melrose 1986).

Although *in situ* observations provide strong evidence that the AKR is generated by the CMI, the growth rates directly determined with the measured distribution functions are not necessarily very meaningful. This is because of the sensitivity of the instability to the various inhomogeneities existing in the plasma and/or its fast saturation by non-linear processes. Among the various gradients present in the magnetospheric medium, two of them probably affect the AKR generation: (1) the inhomogeneity of the magnetic field that induces slight modifications of the gyrofrequency with the altitude such that the resonant condition cannot be fulfilled over a very large distance and (2) the perpendicular inhomogeneity linked to the existence of the sharp density gradients that limit the AKR sources (Le Quéau *et al* 1985, Louarn and Le Quéau, 1996). These effects could have some importance in the determination of the macroscopic properties of the radiation (shape of the emitting lobes, in particular). Concerning the non-linear effects, the growth of the waves can saturate either by quasilinear relaxation or by a trapping-type saturation process, which is still a matter of debate. The non-linear saturation and the inhomogeneity effects could be important for explaining the formation of some of the fine structures observed in the AKR dynamic spectra.

Conclusion

The AKR can be considered as a relatively well-understood magnetospheric phenomena. The sources of this radiation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

have been explored and a theory—the CMI—gives a good qualitative explanation of its generation. However, there is no complete theory that can explain some quantitative aspects of the radiations. For example, starting with a given unstable distribution function, the prediction of the final intensity of the waves and of their spectral characteristics is still an open question. This is an important issue since it is only with solutions to this type of problem that radio waves could be used as precise diagnostic tools.

Bibliography

Benedictov F A, Getmansev G G, Sazonov T A and Trosov A F 1965 *Kosm. Issled.* **3** 614

Ergun *et al* 1998 FAST satellite wave observations in the AKR source region *J. Geophys. Res.*

Le Quéau D, Pellat R and Roux A 1985 The maser synchrotron instability in an inhomogeneous medium: application to the generation of the auroral kilometric radiation *Ann. Geophys.* **3** 273–92

Louarn P, Roux A, de Feraudy H, Le Quéau D, André M and Matson L 1990 Trapped electrons as a free energy source for the auroral kilometric radiation *J. Geophys. Res.* **95** (A5) 5983–95

Louarn P and Le Quéau D 1996 Generation of the auroral kilometric radiation in plasma cavities: II. The cyclotron maser instability in small size sources *Planet. Space Sci.* **44** 211–24

Melrose D B 1986 *Instabilities in Space and Laboratory Plasmas* (Cambrige: Cambridge University Press) pp 194–203

Wu C S and Lee L S 1979 A theory of terrestrial kilometric radiation *Astrophys. J.* **230** 621

Dominique Le Quéau and Philippe Louarn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Australia Telescope National Facility

The Australia Telescope National Facility (ATNF), created in 1989, operates the Australia Telescope Compact Array, the only radiosynthesis array in the southern hemisphere. The array, opened in 1988, consists of six 22 m diameter antennas, which together simulate a single antenna 6 km in diameter. It is located near the town of Narrabri, about 580 km northwest of Sydney. The array has high resolving power, wide bandwidth, good spectral-line capability and the ability to accurately measure polarization to low levels. New techniques employed with the instrument include the 'mosaicing' of large fields and 'pulsar gating'. It has been used to make the first three-dimensional image of Jupiter's magnetic field, and has followed the radio evolution of the remnant of Supernova 1987A, the most important supernova of modern times.

The ATNF also operates the 64 m diameter Parkes radio telescope (opened in 1961). The premier instrument is a cooled 21 cm multibeam receiver which is located at the prime focus of the telescope. This instrument was installed in 1997, and allows deep observations to be made of large areas of sky. The feed cluster is an arrangement of 13 beams placed on a hexagonal grid. This receiver has been used recently to find hundreds of new galaxies, previously hidden by obscuration by galactic dust, and to find hundreds of new pulsars in the Milky Way. It is also being used to conduct the H I Parkes All-Sky Survey (HIPASS), the first wide-angle 'blind' survey aimed at surveying the nearby universe in the 21 cm transition of neutral hydrogen.

The ATNF's third telescope is a 22 m diameter antenna used for millimeter-wave observing. It is located near the town of Coonabarabran, half-way between the other two observatories.

Antennas from all three observatories are used together and with those of other institutions for very long baseline interferometry experiments.

Part of CSIRO, Australia's largest national research organization, the ATNF has 135 staff and an annual operating budget of A$13 million, most of which is direct funding from CSIRO.

The objectives of the ATNF are to: operate and develop the Australia Telescope as a national research facility for use by Australian and international researchers; to exploit the telescope's unique southern location and technological advantages to maintain its position as a world-class radioastronomy observatory; and to further the advancement of knowledge.

For further information see
http://www.atnf.csiro.au.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Australian Astronomy

Astronomy has been one of the principal sciences in Australia since the first European settlement over 200 yr ago. In the modern era, the access to the rich southern skies and a tradition of innovation in instrumentation sustain an active and diverse research program.

Early history

The main objective of Captain James Cook's voyage in 1769, during which he discovered and charted the east coast of Australia, was to observe the transit of Venus from the South Seas. The exact time at which Venus transited the disk of the Sun provided a more accurate value of the distance from the Earth to the Sun, and indeed the size of the entire solar system, as first suggested by Kepler. When the first European settlement was made at Sydney Cove in 1788, the importance of navigation, surveying and time-keeping was the reason for setting up an observatory. Later, official State observatories were founded, the first being Sydney Observatory in 1855. It provided a time service for ships in Sydney Harbor and contributed to the astronomical research of the day, particularly stellar positional astronomy. It finally produced 53 volumes of the Cart du Ciel (International Astrographic Catalogue) project.

State observatories were also set up in Perth and Melbourne—the latter becoming well known because of its acquisition of the Great Melbourne Telescope. This was a 48 in (1.2 m) telescope and at its erection in 1869 was the largest fully steerable telescope in the world. Some of its early photographs of the Moon were hailed as the best to date, and visual observations of southern nebulae were made which greatly improved on those made by Sir John Herschel in his pioneering work from the Cape of Good Hope (see HERSCHEL FAMILY). Nevertheless, problems with the speculum mirror and wind-shake due to the use of a roll-off building, among other things, led to the telescope falling into disuse by the end of the century. The problems with the Great Melbourne Telescope set back the building of large telescopes by many years, and the next one was the Mount Wilson 60 in (1.5 m) built in California in 1908 and arguably the first of the modern telescopes. The Great Melbourne Telescope was sold to Mount Stromlo Observatory and in recent years (with new optics) has been used in the MACHO search for gravitational lensing by stars. Today, of the former colonial observatories, only Perth Observatory retains a research role.

Post-war development

The modern era of astronomy in Australia began immediately after World War II. The impact of wartime technical developments was most noticeable in the new field of RADIO ASTRONOMY, where the expertise in radar systems and the surplus equipment available formed an ideal base. In Australia the work was carried out at the Radiophysics Laboratory of the CSIR (Council for Scientific and Industrial Research, later CSIRO). At this time little was known about radio emission from outside the Earth: Karl Jansky had discovered radio waves from the Milky Way Galaxy in 1932 while working on sources of noise in radio transmissions, and another American, Grote Reber, had built a 31 ft dish antenna in his own backyard and made the first broad-scale maps of the Galactic emission. Finally, a wartime investigation of suspected radar jamming had been traced to unusually intense radio emission from the Sun.

It was in this climate that the radio astronomy groups in the United Kingdom, the Netherlands and Australia began their work. The radiophysics work was carried out at a number of field stations in the Sydney area, and discoveries came quickly: for example the correlation of solar emission with sunspot number and the demonstration of an effective temperature of about one million kelvins, far above the optical surface temperature of about 6000 K. Perhaps the best-known work was carried out with a cliff-top INTERFEROMETER, which was used to discover a number of isolated radio sources and to show that they had very small angular size. The identification of several of these sources with known optical objects, including the Crab Nebula and external galaxies in the constellations of Virgo and Centaurus, showed that radio astronomy would have much to teach us about distant objects. Development of interferometry continued, as a way to overcome the coarse angular resolution of early radio telescopes. The early 'radio star' theory of the nature of the discrete radio sources was shown to be incorrect because the majority of radio sources away from the plane of the Milky Way Galaxy had angular sizes greater than 1 arcmin. They became known as RADIO GALAXIES, in recognition of a radio power output thousands or even millions of times greater than that of a normal galaxy. Surveys of such sources showed an excess of faint sources, but less pronounced than in the early surveys carried out at Cambridge, which were shown to be affected by 'confusion' of multiple sources within the instrumental beamwidth. Later counts of these distant extragalactic sources, from both Cambridge and Sydney, were in agreement and played a pivotal role in cosmological controversies, in particular the demise of the STEADY-STATE THEORY.

Wartime developments also had an impact on OPTICAL ASTRONOMY, although this was already one of the oldest-established sciences overseas. The Commonwealth Solar Observatory at Mount Stromlo near Canberra had been established in 1924 and expanded during the war to manufacture gunsights and other optical munitions. After the war the research field was changed to stellar astronomy, and the name was changed to the Commonwealth Observatory and later to Mount Stromlo Observatory. Modernization of the instrumentation, for example the introduction of photoelectric techniques, proceeded in parallel with numerous research programs. The MAGELLANIC CLOUDS have long been much studied from Mount Stromlo: these near-neighbor galaxies have the great advantage of being at a known distance, yet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

being close enough that star clusters and the brighter stars can be discerned individually. In 1955 the 74 in (1.88 m) telescope was commissioned at Mount Stromlo and, together with a similar telescope in Pretoria, was the largest optical instrument in the southern hemisphere for some 20 yr. Prompted by the growth of the nearby city of Canberra, a site at Siding Spring Mountain near the town of Coonabarabran in New South Wales was selected for future telescopes, and the name changed to MOUNT STROMLO AND SIDING SPRING OBSERVATORIES (MSSSO). The observatory has been part of the Australian National University for many years, and is now the Research School of Astronomy and Astrophysics.

The era of the CSIRO radiophysics fieldstations drew to a close in the early 1960s, with the decision that the increasing sophistication of radio astronomy required the resources to be concentrated into a much smaller number of major instruments. These were the 210 ft (64 m) fully steerable paraboloid at Parkes and the Culgoora Radioheliograph. The Parkes telescope was completed in 1961, when the only other large dish telescope was the Jodrell Bank (UK) 250 ft (76 m). A notable early result from the Parkes telescope was the lunar occultation observation which pinpointed the position of 3C 273, allowing identification with the star-like object which first revealed the great optical redshifts of quasars. The telescope was soon well known for its high-frequency radio surveys, which resulted in the discovery of many quasars, and also for its support during the 1960s and 1970s of the NASA Apollo missions to the Moon. Many other projects have occupied the telescope over the years, covering both spectral line and continuum, Galactic (including pulsars) and extragalactic sources, polarization studies etc.

The Culgoora Radioheliograph consisted of 96 simple dishes of 13 m diameter, arranged in a circle of diameter 3 km on a site near Narrabri, New South Wales. It was completed in 1967. In this era before the use of computers to create images from raw data, ingenious hardware was used to enable a two-dimensional image of the Sun to be synthesized every second. Separate images in the two polarizations revealed further information about plasma emission processes on and above the Sun's disk. The original operating frequency of 80 MHz was later augmented with capabilities at 43, 160 and 327 MHz. The radioheliograph was closed down in 1984, and the site used for the Australia Telescope Compact Array.

The 1960s also saw the construction of the 1 mile Molonglo Mills Cross radio telescope, a transit instrument designed for rapidly surveying the sky at 408 MHz. Operated by the University of Sydney, it enjoyed early success in finding many pulsars, soon after this class of object was discovered at Cambridge. The Molonglo Cross went on to produce a survey of the southern sky, as well as numerous studies of particular classes of object. In 1978 operation as a cross was halted, and the east–west arm was converted to aperture synthesis at 843 MHz, as the MOLONGLO OBSERVATORY SYNTHESIS TELESCOPE (MOST). In this form it has improved sensitivity and image-forming capability. The field of view has since been further enlarged to nearly 2°, and the telescope is carrying out the 'SUMSS' survey of the entire southern sky with a resolution and sensitivity similar to those of the NVSS northern (and equatorial) survey. Another radio telescope project of the 1960s was the Fleurs Synthesis Telescope, converted from an earlier CSIRO grating cross instrument to aperture synthesis operation by the University of Sydney. It provided the highest angular resolution radio imaging in the south for many years, but was superseded by the Australia Telescope and has since been closed down.

A novel instrument of the 1960s was the Narrabri Stellar Intensity Interferometer (NSII). The principle of intensity interferometry had been developed and tested in the context of radio astronomy, and had the advantage of being unaffected by observation through a turbulent medium. Applied to optical observations, this enabled the NSII to observe the angular diameters of stars down to 0.0004 seconds of arc. The instrument was operated by the University of Sydney and comprised two large (6.5 m) REFLECTORS on a circular rail track of diameter 188 m. It successfully observed the sizes of the 32 stars within its sensitivity limit, thereby defining a true temperature scale for hot stars. The instrument was then closed down and work commenced on the Sydney University Stellar Interferometer (SUSI), a 'modern Michelson' instrument using interference of two beams of light from widely separated siderostats instead of correlating the radiofrequency fluctuations as in the NSII. SUSI is located on the same site as the Australia Telescope Compact Array and when fully commissioned will have baselines up to 640 m, giving the highest angular resolution of any of the current optical interferometer projects. As well as extending stellar diameter and hence effective temperature measurements to a very much greater range of stars, a key program will be to observe directly the radial pulsation of Cepheid variables, thereby improving the use of these stars as distance indicators.

The National Facility instruments

It was clear by the 1960s that a larger general-purpose optical telescope was needed, both to extend the existing stellar and galactic work and to pursue the optical identification of radio sources. Astronomers in the UK also felt the need for a larger instrument, and an agreement was reached between the Australian and UK governments which resulted in the construction of the 150 in (3.9 m) Anglo-Australian Telescope (AAT) on the existing Siding Springs site. Completed in 1974, the AAT incorporated innovations in servo drive systems, mirror support and especially notably in computer control. The resulting pointing precision of a few arcseconds was unprecedented at the time and allowed any observer to readily locate even faint objects. The AAT is operated by the ANGLO-AUSTRALIAN OBSERVATORY (AAO), an organization which was set up for this purpose. An early success with the AAT was the detection of optical pulses from the Vela pulsar, at the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

time the faintest object ever recorded (24th magnitude). Although the Siding Spring site is not as good as the best in Hawaii or Chile, the AAT is a highly productive telescope, largely because of the strong technical support and continued investment in new instrumentation. In the current era of 8–10 m class telescopes, the AAT continues to offer general-purpose imaging and spectroscopy but has been moved towards wide-field survey science. Especially notable is the major two-degree field (2dF) instrument, which provides simultaneous fiber-based spectroscopy of 400 objects over a 2° field size. The major program for this instrument is a redshift survey of 250 000 galaxies, to greatly improve the knowledge of large-scale structure.

By the late 1970s it was clear that Australia needed increased capability for high-angular-resolution high sensitivity radio imaging and spectroscopy. This need was met by the construction of the Australia Telescope (AT) array, opened in 1988. The main component is the AT Compact Array at Culgoora. It is an aperture synthesis telescope of six 22 m antennas, movable to different stations along a 6 km east–west baseline. The array is at present being upgraded to operate at wavelengths as short as 3 mm, by the addition of new receivers, improved antenna surfaces, and a north spur track. The AT is operated as a national facility by the AUSTRALIA TELESCOPE NATIONAL FACILITY (ATNF), a division of CSIRO which was split from the parent radiophysics division for this purpose. There is also a similar 22 m antenna at Mopra near Coonabarabran, used for VLBI or for stand-alone programs. Australian industry gained a valuable spin-off from the AT project: the company which constructed the antennas has gone on to win contracts for satellite ground station antennas in Australia and South-East Asia. A well-known earlier spin-off, also from radio astronomy, was the Interscan aircraft microwave landing system.

The ATNF also operates the Parkes 64 m dish telescope, which has been upgraded over the years in order to remain competitive. Recently, a multibeam system with 13 simultaneous feeds at 21 cm has commenced operation and is being used for an H I survey of the entire southern sky out to a radial velocity of 12 700 km s^{-1}.

Astronomy at diverse institutions

Other radio astronomy facilities include the University of Tasmania's 26 m paraboloid located at Mt Pleasant near Hobart and ex-Telstra 30 m antenna at Ceduna in South Australia. Both antennas are used primarily for VLBI, in combination with the AT Compact Array and the telescopes at Parkes and Mopra. This university also operates several optical and cosmic ray telescopes. A number of other university groups are active in astronomy, either using national facilities (e.g. Universities of Melbourne and Western Sydney and Macquarie University) or with their own instruments (e.g. the University of Adelaide work on high-energy cosmic rays, the University of Western Australia's development of gravitational wave detectors and the Australian Defense Force Academy's balloon-borne telescopes). Australian astronomers also make considerable use of overseas facilities and satellite observatories including the HST, and strong international collaboration sees some 350 overseas astronomers visiting Australia each year.

The major new optical telescope of the 1980s is the 2.3 m telescope built by MSSSO at the Siding Spring site. Like virtually all modern large telescopes, it uses an alt–az mounting to reduce the expense of the enclosure building. An early notable result obtained with this telescope was the demonstration that the Small Magellanic Cloud is split along the line of sight into two fragments, probably as a result of a collision with the Large Magellanic Cloud some 300 million yr ago. Another major facility at Siding Spring Mountain is the 1.2 m Schmidt telescope, originally a purely UK operation, but merged with the AAO since 1988. This telescope is well known for its high-quality photographic surveys of the southern sky. As well as continuing large surveys and smaller photographic projects, the telescope is also being used for multifiber spectroscopy within the 6° field.

Future opportunities

Australian astronomers have recognized the need for access to an 8 m class optical–INFRARED TELESCOPE in order to remain competitive in the current environment. As a result Australia has joined the international Gemini project as a partner at the 5% level. This will give Australian observers access to both the northern and southern hemispheres, with a variety of high-quality optical and IR instrumentation. However, with an allocation equivalent to 10% of a single telescope, as compared with 50% of the AAT during the era in which 4 m class telescopes led the field, access to 8 m time will be severely limited. Thus it will be vital that the AAT be maintained with a high standard of competitive instrumentation.

Several other development initiatives are worthy of note. Australian astronomers are active in the development of Antarctica as an observing site. The low temperatures and minimal atmospheric water vapor give exceptionally good conditions for infrared observations, including wavelengths that are not normally observable from the ground. There is also the possibility that the high Antarctic plateau may have the best SEEING of any ground-based site. The Joint Australian Center for Astrophysical Research in Antarctica (JACARA), based at the University of New South Wales, is already operating a 60 cm infrared telescope at the South Pole. It is planned to move the telescope at a later date to one of the probably superior (but less accessible) sites on the high plateau.

In radio astronomy, Australia is actively participating in the planning and development of the square kilometer array, an international project aimed at a cm wavelength radio telescope with sensitivity 1–2 orders of magnitude beyond any current instrument. One of the primary astrophysical goals would be the detection and study of the 21 cm emission from neutral hydrogen at high redshifts, in protogalaxies and star formation regions. Subsections of the array would extend over hundreds of km, and one

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the best prospects for an interference-free site of the required size is in the western desert areas of Australia.

Organization

In 1989 a report into astronomy in Australia noted a weakness in theoretical astrophysics. This area was strengthened at the start of the 1990s by the establishment of the Research Center for Theoretical Astrophysics (RCfTA) at the University of Sydney, and later the Australian National University Astrophysical Theory Center in Canberra. Both groups are able to take advantage of the mutual benefits derived from having theoreticians working in close contact with observers. The RCfTA has been created with the financial support of the Australian Research Council (ARC), which disburses government funds in a wide range of scientific and technological areas. Through its grant schemes the ARC also supports much of the observational research work carried out in Australian universities.

The professional astronomical community in Australia consists of some 200 positions, with about 30 of those being fixed-term appointments. There are about 75 PhD students in the various universities nationwide, with about 15 graduating each year. The main association for professional and student astronomers is the ASTRONOMICAL SOCIETY OF AUSTRALIA (ASA), with around 300 members. The principal activities of the ASA comprise organization of annual scientific meetings, management of the publications of the ASA and representation of astronomers' interests to government and the community at large. Official liaison with government is provided by the National Committee for Astronomy (NCA), which operates under the auspices of the Australian Academy of Science. The NCA manages Australia's relations with the INTERNATIONAL ASTRONOMICAL UNION (IAU), and Australia currently has 190 IAU members.

Bibliography

Bhathal R and White G 1991 *Under the Southern Cross: A Brief History of Astronomy in Australia* (Sydney: Kangaroo)

Gascoigne S C B, Proust K M and Robins M O 1990 *The Creation of the Anglo-Australian Observatory* (Cambridge: Cambridge University Press)

Haynes R F, Haynes R, Malin D F and McGee R X 1996 *Explorers of the Southern Sky: A History of Australian Astronomy* (Cambridge: Cambridge University Press)

Morganti R and Couch W J (ed) 1999 *Looking Deep in the Southern Sky* (Berlin: Springer)

Robertson P 1992 *Beyond Southern Skies: Radio Astronomy and the Parkes Telescope* (Cambridge: Cambridge University Press)

J Gordon Robertson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Auzout, Adrien (1622–91)

French astronomer, independently invented the micrometer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Axis (pl: axes)

A reference line in space or drawn through a body. Usually this is a line about which a body has some degree of symmetry, for example, a line passing through the center of a sphere or which runs centrally through the length of a cylinder. For an ellipse the major axis is a line passing through the center and crossing the greatest diameter of the figure.

The axis of rotation is a real or imaginary line passing through a body, around which that body rotates.

The axial period is the period of time taken for a body to complete one rotation on its axis. The period is usually referred to rotation relative to the stars, in other words, the sidereal rotation period. Thus the axial period of the Earth is 23 h 56 m, although the mean solar day consists of 24 h.

See also: astrodynamics, orbit, revolution, rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Azimuth

The angle, measured parallel to the horizon in a clockwise direction from north, between the meridian and a celestial body. It is the angle between the north point of the horizon and a point on the horizon vertically below the celestial body.

Azimuth may take any value between 0° and 360°. In some textbooks, azimuth is defined as being measured clockwise from north from 0° to 180° if the object is east of the meridian, and anticlockwise from north from 0° to 180° if the object is west of the meridian. Together with a value of altitude, this specifies the position in the sky of a celestial body as seen at a particular instant from a particular point on the Earth's surface.

For non-astronomical purposes, azimuth (or bearing) is generally measured clockwise from due north.

See also: altitude, culmination, horizon, meridian, zenith distance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baade, Wilhelm Heinrich Walter (1893–1960)

Astronomer, born in Westphalia, Germany, worked at the Hamburg Observatory and moved to California in 1931. He and FRITZ ZWICKY proposed in 1934 that supernovae could produce cosmic rays and neutron stars. Identifying it as the supernova of 1054, Baade made extensive studies of the Crab Nebula. He studied its polarization (which led to identification of the light from the nebula as synchrotron radiation), its filaments (the out-rushing body of the exploding star) and the nebula's central star, now known to be the pulsar remnant of the supernova explosion. During the Second World War blackouts of the Los Angeles area, Baade, confined as an alien to the astronomer's paradise of an otherwise little used Mount Wilson, used the 100 in telescope to resolve stars in the central region of the Andromeda Galaxy for the first time. This led to his definition of two stellar populations (Populations I and II), to the realization that there were two kinds of Cepheid variable stars, and from there to a doubling of the assumed scale of the universe. With RUDOLPH MINKOWSKI, identified and took spectra of optical counterparts of many of the first-discovered radio sources, including Cygnus A and Cassiopeia A.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Babbage, Charles (1792–1871)

Polymath, eleventh Lucasian Professor at Cambridge. Known not only for work on astronomy but also in cryptanalysis, probability, geophysics, altimetry, ophthalmoscopy, statistical linguistics, meteorology, actuarial science, lighthouse technology, the use of tree rings as historic climatic records, magnetism, biology, geology, religion, submarines, politics, economics and mathematics. Most important for astronomy, he developed calculating engines (the precursors of modern computers), for the development of which he was unable to obtain funding from the government after Airy's assessment that they would be worthless. One of the founders of the [Royal] Astronomical Society in 1820 and awarded its first Gold Medal.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Babcock, Harold Delos (1882–1968) and Babcock, Horace Welcome (1912–)

Harold was a physicist, born in Edgerton, WI. Became one of the first staff members of the Mount Wilson Observatory, working on precise laboratory studies of atomic spectra, improved the precision of the wavelengths of some 22 000 lines in the solar spectrum, and with Horace, measured the distribution of magnetic fields over the solar surface. He measured the magnetic field of the star 78 Virginis, invented the solar magnetograph, and ruled excellent large gratings, including those used in the coudé spectrographs of the 100 and 200 in telescopes. His son Horace, born in California, also became a staff member of the Mount Wilson Observatory. He not only worked with his father on observational astronomy but also built astronomical instruments, including a ruling engine that produced excellent diffraction gratings, the solar magnetograph, and microphotometers, automatic guiders and exposure meters for the 100 and 200 in telescopes. By combining his polarizing analyser with the spectrograph he discovered magnetic fields in other stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Babylonian Astronomy

The civilizations of Mesopotamia nurtured a rich astronomical tradition for about two thousand years, from the early 2nd millennium BC—if not before—until the 1st century AD. Incited by the belief that events seen in the sky were portents signifying future mundane conditions, Babylonians and Assyrians evolved an intricate practice of omen watching, from which they learned the varieties of celestial phenomena and devised rules for predicting their occurrence. From the 7th century BC on, records were kept at Babylon of observations of the Moon, the five visible planets and the Sun. The Babylonians discovered not only the periodic nature of many of the phenomena that they observed, but also that certain phenomena involved several periodic components, which they succeeded in isolating, measuring and finally mimicking by means of arithmetical sequences. Thus arose a mathematical astronomy that operated on quite different principles from the geometrical models of Greek and MEDIEVAL ASTRONOMY. Notwithstanding this disparity of methods, the influence of Babylonian astronomy on its Greek counterpart was deep, and some elements of Babylonian origin persisted in living astronomical traditions across Eurasia through the Middle Ages and into modern times.

Sources

Almost all that we know about Babylonian astronomy is derived from contemporary documents written in the Akkadian language and cuneiform script on clay tablets. Cuneiform tablets were recovered in abundance in excavations in the mid-19th century, and the decipherment of the script followed soon after. The texts that have to do with astronomy number in the thousands. A large and growing fraction of them are available in editions and translations into modern languages, but there are still numerous unpublished tablets, especially among the omen texts and observational records.

The tablets held in modern museums and other collections were only in part obtained by scientific excavations that provided a record of their provenance, the rest having been purchased from dealers. Nevertheless, most of the astronomical tablets can be traced to three sites: the royal library and archive of the Assyrian capital Nineveh, the most important source for texts pertaining to celestial omens, chiefly dating to the first half of the 7th century BC, tablets from uncertain locations in Babylon, almost the sole source for observational records, and also rich in mathematical astronomy, and a temple archive from the southern Babylonian city Uruk, which is the source for the remainder of the mathematical texts.

Before the modern decipherment of cuneiform, practically the only information available about astronomy in Mesopotamia consisted of scattered statements in classical Greek and Latin authors. These range from patently absurd claims of thousands of years of observations to specific technical data that have been triumphantly confirmed from the Babylonian tablets. The diffusion of Babylonian methods and concepts can now be traced more fully in Greek and Egyptian astronomical papyrus fragments from Roman Egypt and in certain Indian astronomical texts.

Astronomy in Mesopotamian omen lore

Celestial omens constituted a large part of the vast repertoire of events regarded by the Mesopotamians as ominous. The belief in omens was essentially religious: the gods, it was supposed, gave humans notice of their mood and intentions through visible signs, which it was the business of expert diviners to interpret. Omen texts, typically expressed as conditional sentences linking a portent to a predicted outcome (e.g. 'If the moon at its appearance glides through clouds: the flood will come'), were collected in compendia such as the series of celestial omens *Enūma Anu Enlil* ('when the gods Anu and Enlil', from the opening words), which comprised some 70 tablets. *Enūma Anu Enlil* survives in fragmentary copies from the 7th century BC and later, but some celestial omens both contained in this collection and preserved outside it date from the first half of the 2nd millennium BC, and a small number refer to historical events of the 3rd millennium.

The omens of *Enūma Anu Enlil* are, in the broadest sense, 'sky omens', encompassing meteorological phenomena as well as sightings of the Moon, the Sun, the five planets and fixed stars. Even in the observation of a heavenly body, color and brightness, halos and attendant weather are treated on an equal basis with times, dates and positions. It would be misleading to call the observational component of this divination astronomy, but much astronomy is latent in it. Thus the aspects of a lunar eclipse that helped to determine its ominous import—generally the death of a king—included the date and rough time of day, the directions from which the Moon's disk was obscured at the beginning and end of the eclipse and occasionally also the constellation within which the eclipse took place.

The archives of the Assyrian kings at Nineveh, now in the British Museum, preserved numerous letters and reports addressed to the kings Esarhaddon (680–669 BC) and Ashurbanipal (668-627 BC) by a network of omen watchers in Nineveh and other Assyrian and Babylonian cities. In this fascinating correspondence we can see that a vigilant watch on the Moon, Sun and planets was a matter of the highest political importance, affecting decisions concerned with all the king's actions. The scholars on whom the king depended had the best motives for seeking to know in advance when ominous phenomena could be expected or ruled out: colleagues were always at the ready to point out to the king a mistaken or late sighting.

Forecasts, which are frequent in the royal correspondence, chiefly concern lunar and solar ECLIPSES and the first visibilities of planets. Typical forecasts are either confidently negative ('an eclipse will not occur this month', or 'a specific planet's appearance is not imminent') or conditionally affirmative ('if an eclipse will occur this month, we should watch for it at a particular time of day or night').

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In addition to divinatory methods, which can hardly have been reliable, the omen watchers seem to have depended on knowledge of the approximate possible intervals between eclipses and between successive appearances and disappearances of the planets. Whether they also made use of longer periodicities is not known.

MUL.APIN ('Plough Star') is the most important of a handful of texts from the 2nd and early 1st millennium BC that are concerned primarily with genuine astronomical topics. The longest part of this compilation consists of lists of constellations and stars, bearing more or less pictorial names such as the 'wagon', the 'frond of Eru' and the 'hired man'. Stars that rise and set are grouped in three 'paths', apparently according to the parts of the eastern horizon that they rise within, but no further attempt is made to describe their arrangements or locations. Astronomical tests are given for determining when an intercalary month should be inserted in the lunar calendar to keep it in line with the seasons. *MUL.APIN* contains little lunar astronomy, and its treatment of the planets is limited to rough estimates of the intervals of time between appearances and disappearances.

A second text of great interest is the fourteenth tablet of *Enūma Anu Enlil*, which, unlike the rest of the compilation, is not a collection of omen texts, but a set of tables of numbers. One of these tables represents the varying duration of nocturnal lunar visibility through the course of a synodic month on the assumption that the length of night is at its maximum, i.e. close to winter SOLSTICE, and a second table represents, month by month through a schematic solar year, the daily change in the duration of nocturnal lunar visibility. Except at the beginning and end of the first table, where a curious pattern of constant second differences is used, the numbers increase to a maximum and thereafter decrease again by constant differences, a kind of sequence that is called by modern authors (from its appearance when graphed) a linear zigzag function. Simple zigzag functions consisting of small whole numbers are an obvious means of describing a periodically varying quantity, and the oldest known example is a crude table representing the varying length of daylight through the year in a second millennium BC papyrus from Egypt, a land not noted for its mathematical astronomy. What is remarkable here is that by about 1000 BC, Mesopotamian astronomers were attempting to isolate, and present mathematical models for, two independent periodic elements in a phenomenon that oscillates in a complex way.

Diaries and Goal Year Texts

The extent to which omen watchers maintained records of what they saw in the heavens is not known. In at least one city, Babylon, a group of scribes who may or may not have been omen watchers had already begun by the middle of the 7th century to record a remarkable variety of dated phenomena in tablets that we now call Diaries. Many hundreds of fragments of Diary tablets from Babylon survive, pertaining to years ranging from 652 to 61 BC, with the bulk of them coming from the last three centuries BC. Typically a Diary tablet covered half a lunar year, that is, six or seven lunar months. For each month, night by night and day by day, the Diary records observations of the heavenly bodies and weather, and each month ends with records of commodity prices, the level of the Euphrates River, and current events.

The most conspicuous object in the Diaries is the Moon. Every month, six time intervals (the 'Lunar Six') were measured in units called *uš* or 'time-degrees', such that 360 time-degrees equal one (mean) day from sunset to sunset:

1. sunset to moonset on the evening of first lunar visibility;
2. moonset to sunrise on the morning before full Moon;
3. moonrise to sunset on the evening before full Moon;
4. sunrise to moonset on the morning after full Moon;
5. sunset to moonrise on the evening after full Moon;
6. moonrise to sunrise on the morning of last lunar visibility.

This is an extension of the interest in lunar nocturnal visibility evinced in *Enūma Anu Enlil* tablet 14. Also recorded with great frequency in the Diaries are the Moon's passages by about 30 bright stars close to the ECLIPTIC, known as Normal Stars. The circumstances of lunar and solar eclipses are given in detail.

For the planets, the Diaries record passages by the Normal Stars and close approaches to each other, and also certain events in their synodic cycles (see SYNODIC PERIOD. The reported synodic phenomena of Saturn, Jupiter and Mars are their first visibilities and disappearances, stations, and acronychal (sunset) risings. First visibilities and disappearances as morning and evening star are reported for Venus and Mercury. The locations of the planets at their synodic phenomena are specified, if at all, only roughly in relation to a constellation, or, in later Diaries, by the zodiacal sign occupied by the planet. The transition from constellations to conventional zodiacal signs covering equal twelfths of the ecliptic seems to have taken place about the middle of the 5th century BC.

Many entries in the Diaries are not observations. Predicted or corrected data are recorded where weather delayed or obstructed observation, often accompanied by the remark, 'I did not watch'. The sources of many of these predictions have not been established beyond dispute. One important method of forecasting planetary and lunar phenomena was certainly the application of the 'Goal Year periods'. These were periodicities ranging from as few as 8 yr for Venus to as many as 83 yr for Jupiter, after which an approximate repetition of phenomena (synodic events or Normal Star passages) takes place. The planetary periods are all whole numbers of solar years. The important periodicity for the moon was the so-called 'SAROS', an eclipse period comprising 223 synodic months (about 18 yr) that also contains close to a whole number of periods of lunar anomaly and lunar latitude. In tablets called Goal Year Texts, observation reports extracted from the Diaries one Goal Year back are collected to provide the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

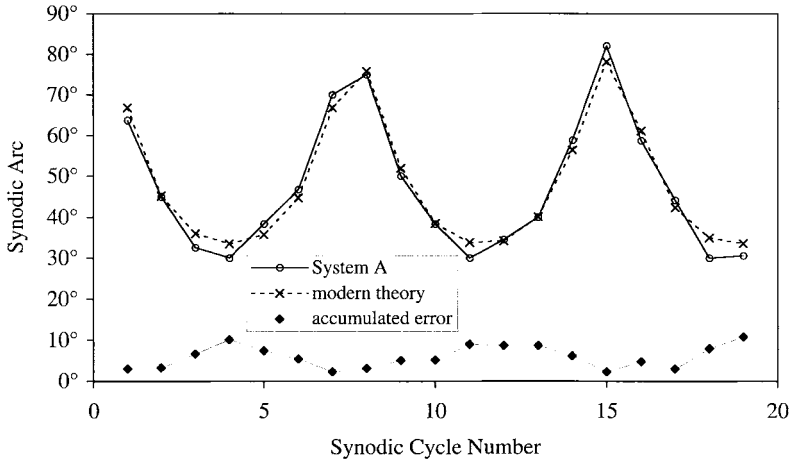


Figure 1. Predictions of longitudes of Mars' first stationary point by system A.

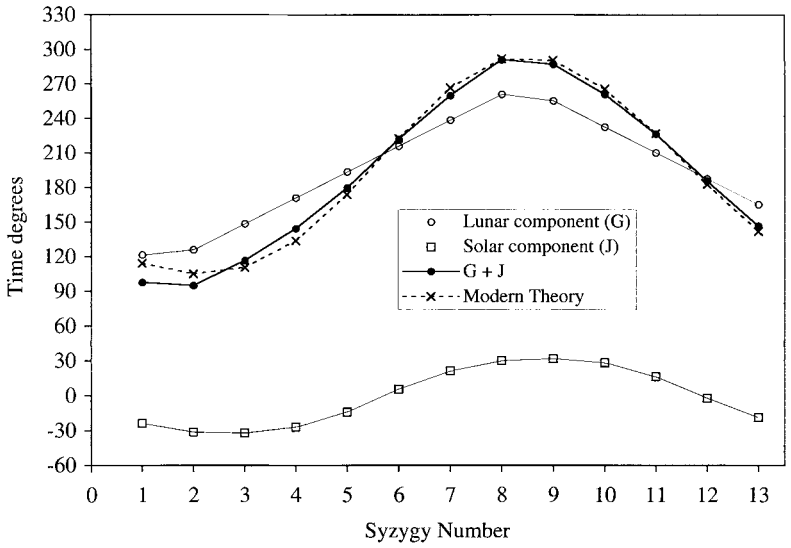


Figure 2. Predictions of lengths of synodic months by system B.

basis for predictions for an approaching year. To make the actual predictions, corrections would have to be applied to take account of the inexactness of the periodicities. Tablets known as Almanacs and Normal Star Almanacs, containing predictions of certain categories of Diary-type phenomena for a specific year, are likely to be the products of this process.

Mathematical astronomy

The primary goal of Babylonian mathematical astronomy was to calculate certain of the phenomena that the Diaries recorded. These included the dates and longitudes of the synodic phenomena of the planets, the 'Lunar Six' time intervals and eclipses. The predictions were achieved by means of sequences of numbers generated arithmetically according to rules such that each sequence would repeat with exactly the periodicity required for a particular component of the motion of a heavenly body. The predictive models fall into two categories. Models in which linear zigzag functions are prominent among the sequences are referred to as system B models. Those in which the numbers follow 'step functions' depending numerically on the location of the heavenly body within partitions of the ecliptic are called system A models. System A and system B models were used concurrently at both Babylon and Uruk.

The planetary models work with 'synodic arcs', the numbers of degrees by which a planet progresses through the ecliptic from one occurrence of a synodic phenomenon to the next occurrence of the same phenomenon, and 'synodic times', the numbers of months and days from one event to the next. The actual longitudes (degrees within zodiacal signs) and dates of phenomena are obtained as the running total of the synodic arcs and times, which are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

generated as periodic sequences. The periodicity can be expressed by the equation

$$\Pi \text{ occurrences} = Z \text{ revolutions of the ecliptic} = Y \text{ yr}$$

where Π, Z and Y are whole numbers, and Π and Z are relatively prime. Thus in a table of planetary phenomena predicted according to a Babylonian model, after Π lines the longitude will return exactly to its original value.

System A planetary models employ rules equivalent to the following assumptions: (1) the ecliptic is divided into two or more zones with fixed boundaries, (2) the Π occurrences of a phenomenon within a full period of Y yr are distributed along the ecliptic such that they are spaced at equal intervals inside each of the zones and (3) the progress from one occurrence to the next is Z intervals. The Babylonian base 60 place-value notation for numbers introduces some nonastronomical restrictions on the model, since one always wants the calculations to lead to numbers that terminate after a small number of fractional places, usually two. Nevertheless, the sequence of synodic arcs can be shaped by a suitable choice of parameters so as to reproduce successfully the actual behavior of a planet, as is illustrated by figure 1, in which a run of synodic arcs between first stationary points of Mars calculated by the system A model is compared with the same intervals computed by modern theory. System B, according to which the synodic arcs and times are zigzag functions, is a more restrictive model, and less successful in reproducing the phenomena.

A complete table of computed phenomena for a planet might contain columns giving the dates and longitudes of the full cycle of synodic phenomena, for example Mars' first visibility, first station acronychal rising, second station, and disappearance. Longitudes on intermediate dates could be obtained by extrapolation from or interpolation between the tabulated date–longitude pairs.

The Babylonian lunar models are a more remarkable achievement than the planetary models. Eclipses and lunar visibility times are affected by several independent periodic components, which it cannot have been easy to isolate and measure. In the system A and system B models, at least four periodicities are present: the year (apparently no distinction was drawn between sidereal, tropical, and anomalistic years), the synodic month and the periods of lunar anomaly and latitude. The calculation of the precise time of conjunction may serve to illustrate the approach. The interval between the preceding and current conjunction is assumed to be 29 days plus two variable components (expressed in time-degrees), called J and G. G varies with the Moon's anomalistic period. In system B it is a linear zigzag function, while in system A it is a zigzag function 'snubbed' into a sinusoidal shape at its extrema by second-order arithmetical sequences. J has the solar year for its period. It is modelled in system A by a step function dependent on the Sun's longitude, whereas in system B it is a smooth sinusoidal sequence obtained as the running total of a linear zigzag function. Figure 2 shows a run of values of G, J and their sum from a system B tablet. The errors in the calculated times, which can be as large as 2 h, are due partly to the use of fairly simple arithmetical functions and partly to the inherent inaccuracy of representing the true synodic month as the sum of two independent periodic functions. The periodicities, especially those of the system B lunar theory, are very accurate.

The last tablets of Mesopotamian astronomy are almanacs from the AD 70s. By that date, however, the methods of Babylonian mathematical astronomy had spread to the Greek-speaking world, where they provided the computational foundation for Greco-Roman ASTROLOGY until they were superseded by Ptolemy's tables in late antiquity. HIPPARCHUS and PTOLEMY used Babylonian observation reports and the periodicities of the system B lunar theory, and their work may have been influenced less directly by other aspects of the mathematical models.

Bibliography

Neugebauer O 1955 *Astronomical Cuneiform Texts* (London: Lund Humphries)

Neugebauer O 1975 *A History of Ancient Mathematical Astronomy* (Berlin: Springer)

Rochberg F 1998 *Babylonian Horoscopes* (Philadelphia, PA: American Philosophical Society)

Sachs A and Hunger H 1988ff *Astronomical Diaries and Related Texts from Babylonia* (Vienna: Österreichische Akademie der Wissenschaften)

1987ff *State Archives of Assyria* (Helsinki: Helsinki University Press)

Alexander Jones

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Backlund, Jöns Oskar (1846–1916)

Born in Sweden, worked in Russia, at the Dorpat (now Tartu) Observatory, became director of Pulkovo Observatory, and the Imperial Academy of Sciences in Petrograd. He employed a large staff to compute the orbit of Comet Encke, eventually concluding that the comet's motion was affected by non-gravitational forces (now known to be light pressure and magnetic forces from the Sun).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baffin, William (c 1584–1622)

Navigator, probably born in London, pilot on several expeditions in search of the Northwest Passage. He seems to have been the first person to determine a degree of longitude at sea by lunar observation. *Baffin Island* is named after him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bailey, Solon Irving (1854–1931)

American astronomer, born New Hampshire. Discovered the variable stars in globular clusters now known as RR Lyrae variables, and divided them into classes a, b and c, according to the shapes of their light curves, now known to be an indication whether the star is pulsating in the fundamental mode or an overtone.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baillaud, Édouard Benjamin (1848–1934)

French astronomer, became director of Toulouse Observatory and Paris Observatory. He worked in celestial mechanics, especially on the motions of the satellites of Saturn, and inaugurated the telegraphy of time signals from the Paris Observatory via the Eiffel Tower. First president of the International Astronomical Union.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bailly, Jean Sylvain (1736–93)

Astronomer and politician, born in Paris, studied Halley's comet and the satellites of Jupiter, wrote *Histoire de l'Astronomie*. Guillotined as the result of an incident when, as Mayor of Paris during the Revolution in 1789, he let the National Guard fire on republican crowds. Unfortunately not the only astronomer to be executed—see ARCHIMEDES, GIORDANO BRUNO, HYPATIA OF ALEXANDRIA, ULUGH BEG, who represent nearly 1% of the astronomers whose biographies are in this encyclopedia; perhaps astronomy is a riskier profession than one would think!

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baily, Francis (1774-1844)

Stockbroker and astronomer, born in Newbury, Berkshire, England. During an eclipse of the Sun in 1836, he identified the phenomenon known as *Baily's beads*, due to sunlight at the edge of the Sun shining down lunar valleys.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baily's Beads

Brilliant points of light seen at the edge of the Moon's disk during a total eclipse of the Sun, just at the onset and termination of totality. They are due to sunlight passing through valleys and indentations at the edge of the Moon's visible disk. The English astronomer Francis Baily (1774–1844) drew attention to the phenomenon during the solar eclipse of 1836.

See also: eclipse.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ball, Sir Robert Stawell (1840–1913)

Irish astronomer, Astronomer Royal for Ireland, popularizer of astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Balmer, Johann Jakob (1825–98)

German phyicist, explained the optical emission of the visible hydrogen lines with the *Balmer formula*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Banneker, Benjamin (1731–1806)

The 'Sable Astronomer', first black American scientist, born in Ellicott's Mills, MD of a family of former slaves. As a young man, Banneker borrowed a pocket watch, took it apart, made a drawing of each component, and reassembled it; after which he made out of wood a working, enlarged replica which lasted over 50 years. At age 58, he began the study of astronomy and was soon predicting eclipses. He compiled this information, including other astronomical and tide calculations, weather predictions, proverbs, poems and essays contributed by himself and others into *Benjamin Banneker's Almanac* (1792–1802). This almanac was cited by opponents of slavery (including Thomas Jefferson) as evidence of African-Americans' abilities.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bappu, Manali Kallat Vainu (1927–82)

Born in Hyderabad, India, of an astronomer father, and became director of the Kodaikanal Observatory. Following a chance meeting with SIR HAROLD SPENCER JONES and HARLOW SHAPLEY, enrolled in Harvard, was co-discoverer of comet Bappu–Bok–Newkirk. Working with OLIN WILSON at the Mount Wilson Observatory, Bappu discovered that for stars of spectral types later than G0 there is a relationship between their luminosities and the characteristics of the H and K emission lying in the absorption lines of singly ionized calcium. The reason was that the emission is characteristic of the surrounding atmosphere of the star, which is related to its size and thus to its luminosity. This phenomenon, is known as the *Wilson–Bappu effect*, and is a method to determine the absolute magnitudes, and therefore the distances, of late-type stars. He returned to India where he set up a modern astronomical observatory at Nain Tal in Uttar Pradesh. The Vainu Bappu Observatory is the main optical observatory of the Indian Institute of Astrophysics, located in the state of Tamil Nadu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barlow Lens

An additional lens that increases the effective focal length and magnification of a telescope. It is a negative diverging lens (either concave on both sides or, more usually, 'plano-concave'—flat on one side and concave on the other) that is placed in the converging cone of light a short distance in front of the focal plane of the objective or primary mirror. By decreasing the angle at which the light rays are converging, it increases the effective focal length of the telescope, so enabling a given eyepiece to achieve a higher magnification. The amplification factor (the ratio of the effective focal length of the objective with the Barlow lens in place to the focal length of the objective alone) of a Barlow lens of focal length f_b, placed a distance d inside the focal plane of an objective, is $f_b/(f_b - d)$. For example, if a Barlow lens with a focal length of 100 mm is placed 50 mm inside the focus of the objective, the effective focal length will be increased by a factor of $100/(100 - 50) = 100/50 = 2$.

See also: focal length, focal ratio, focus, magnification, mirror, objective lens.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barlow, Peter (1776–1862)

Nautical engineer, born in Norwich, Norfolk, England, professor of mathematics in the Royal Military Academy, Woolwich, where he worked on the strength of ships' timbers, tidal engineering and ships' magnetism. His best known publication is *New Mathematical Tables*, giving the factors, squares, cubes, square and cube roots, reciprocals and hyperbolic logarithms of all numbers from 1 to 10 000, together with the powers of numbers up to 1000. The tables were widely used for navigation and astronomy. His optical experiments began about 1827, and with George Dollond he invented the *Barlow lens*, a concave achromatic lens used to magnify the image produced by a telescope, and employed first to measure close double stars (also used in photography).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barnard, Edward Emerson (1857–1923)

Astronomer, born in Nashville, Tennessee. As a poor photographer's assistant he made his own telescope and discovered several comets. During this time, a wealthy patron of astronomy awarded $200 each time a new comet was found, and Barnard's eight comets earned enough to build a house for his bride and led to his gaining an education at Vanderbilt University and joining the initial staff of the Lick Observatory. In 1895 he moved to the new Yerkes Observatory. He discovered *Barnard's Star* (the star with the largest proper motion, and one of the nearest stars), Jupiter's satellite Amalthea, and, by wide-field photography, dark clouds and globules in the Milky Way, including *Barnard's Loop* around Orion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barnard's Loop (Sh2-276)

A large emission nebula in the constellation Orion in the form of an arc centered on the Sword of Orion; the midpoint of the arc is at approximately RA $05^{h}\,56^{m}$, dec. -02°. The emission is thought to be caused by radiation pressure from the hot stars in the Sword region ionizing interstellar matter. It was discovered by E E Barnard.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barnard's Star

A noteworthy star in the inconspicuous constellation Ophiuchus, it lies about 34° south of Vega and 28° west of Altair. It is identified only by a catalog designation, e.g. BD+4°3561. Despite being very faint, apparent magnitude 9.54 and hence visible only with powerful binoculars or a small telescope, it has several claims upon our interest. With a parallax of 0.549″, at a distance of 5.9 light-years, only the triple system of Alpha Centauri is closer to us. Discovered in 1916 by the great American observer Edward Barnard (1857–1923), it soon became clear that this star has the largest known proper motion, 10.36″ per annum. This so-called 'fixed star' is moving across the sky through a distance equivalent to the Moon's diameter every 174 years.

It is a red subdwarf star, spectral type sdM4, of very low intrinsic luminosity (absolute magnitude 13.2). A study of its observed motion by Peter van de Kamp (b. 1901) suggested that it is subject to regular, small fluctuations in both right ascension and declination, suggesting that its motion is perturbed by the presence of a binary companion too faint to be separately observed. Van de Kamp estimated that this companion might well be a non-luminous body of mass about 1.6 times that of Jupiter, i.e. a planet. This was the first well-documented case of this kind but has not been confirmed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barred Spiral Galaxy

A galaxy in which the spiral arms emerge from the ends of a bar or elongated ellipsoid of stars that straddles the nucleus rather than from the nucleus itself. Barred spiral galaxies, denoted by 'SB' in the Hubble classification scheme, are classified according to the size of the central bulge, the tightness of the spiral pattern and the degree of clumpiness in the arms.

The central bulge of our own galaxy, the Milky Way, appears to be an elongated ellipsoid that is seen almost end-on from the present location of the Sun. Consequently, there are grounds for classifying it as a barred spiral.

See also: galaxy, Hubble classification, Milky Way Galaxy, spiral arms, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barringer, Daniel Moreau (1860–1929)

A Philadelphia mining engineer who, in 1903–5, investigated the 'Coon Butte' crater in Arizona, in pursuit of the theory that it was of meteoritic origin. Before he realized that the impact was oblique and the meteor was off-center, he unsuccessfully searched for a buried mass of nickel and rare-earths by drilling in the crater bottom. He collected meteoric iron fragments, which he noticed were intermixed with terrestrial rock fragments, and used this as evidence of his theory that the crater had been made by a meteor impact. He identified geological evidence of the shock that had excavated the crater, such as pulverized silica and a type of quartz glass, later called lechatelierite. Barringer's scientific conclusion about the origin of the crater was eventually shown by FOREST RAY MOULTON, HENRY NORRIS RUSSELL and HARLOW SHAPLEY to be correct. However, as a result of his theory being proved, it was concluded that what he had thought was a valuable lode of nickel was not only off-center, but smaller than he had at first estimated and would have vaporized on impact. The directors of the mining company which supported his work withdrew funding and Barringer died of a heart attack within weeks. The crater is now referred to as the *Barringer Meteor Crater*, the first terrestrial crater recognized to be of meteoritic origin.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Barycenter

The center of mass of the Earth–Moon system or any other set of bodies (e.g. the solar system). If two bodies move under the influence of their mutual gravitational attractions, they will describe orbits around a point between them, the center of mass of the two bodies. The relative distances of the two bodies from this point will be inversely proportional to their masses. Thus, for example, if one body is twice the mass of the other, the center of mass will lie one-third of the way from the more massive body to the less massive one (i.e. their distances are in ratio 1:2).

In the case of the Earth and Moon, the ratio of masses is approximately 81:1. The barycenter is therefore located at a distance from the Earth's center equal to 1/82 of the distance from the Earth to the Moon, and on the line joining the centers of Earth and Moon, in other words at a distance of about 4700 km from the center of the Earth. The point is, in fact, located *within* the Earth. The distance between the centers of the Earth and the Moon varies rather widely due to the combined gravity of the Earth, the Sun and the planets. For example, during the period 1969–2000, apogee (the greatest distance) varies from 404 063 to 406 711 km. Perigee (the least distance) varies from 356 517 to 370 354 km. This has a proportionate effect on the distances of the centers of the Earth and the Moon from the barycenter.

See also: anomaly.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baryogenesis (baryon genesis or matter/antimatter asymmetry)

Associated with every type of particle there is an ANTIMATTER particle with the opposite quantum numbers. For example, the positron (charge +1) is the antiparticle of the electron (charge −1), and the antiproton (charge −1, baryon number −1) is the antiparticle of the proton (charge +1, baryon number +1). Particles can annihilate with their antiparticles into radiation, and particle–antiparticle pairs can be created from the vacuum.

Observations seem to suggest that the universe consists of matter, with no appreciable amounts of antimatter. Domains of antimatter galaxies and antimatter galaxy clusters, if they exist, must be widely separated. The simplest explanation is that there is an antimatter asymmetry in the universe: today the universe consists solely of particles with no large antimatter systems.

The matter/antimatter asymmetry is usually expressed in terms of a dimensionless number known as the 'baryon number of the universe'. If a baryon number of +1 is assigned to BARYONS (neutrons, protons and the more massive baryon resonances) and a baryon number of −1 assigned to antibaryons, then because of the absence of large amounts of antimatter, the present baryon number density (baryon number density minus the antibaryon number density) can be expressed in terms of the total mass density of the universe. The baryon number is the ratio of the baryon number density to the entropy density of the universe. With an uncertainty of a factor of three to five, the baryon number of the universe is about 10^{-10}.

So long as baryon number is conserved in every individual reaction and the expansion of the universe is isentropic (entropy conserving), the baryon number of the universe is conserved over the history of the universe.

Although there is no antimatter in the universe today, if the temperature of the universe was ever larger than about 10^{13} K, and thus its thermal energy exceeded the rest-energy of the proton, then antimatter would have existed in great abundance. Earlier than about a millisecond after the BIG BANG the universe should have been hot enough for the radiation to create particle–antiparticle pairs. If the baryon number is conserved, then at sufficiently high temperature a baryon number of 10^{-10} implies that there were roughly equal numbers of baryons and antibaryons in the universe, but for every 10 billion antibaryons there were 10 billion and one baryons. As the universe cooled below the temperature required to produce proton–antiproton pairs from the radiation, all of the antibaryons annihilated, leaving behind the excess baryon number of 1 part in 10 billion. While such a tiny asymmetry between the number of baryons and the number of antibaryons might seem insignificant, without the asymmetry all of the baryons and antibaryons would have annihilated, resulting in a universe without the matter (or antimatter) required to form galaxies, stars and planets.

While the present baryon number of the universe may simply reflect an initial condition of the big bang, it seems more natural that a universe with zero baryon number, as would result if the universe underwent a period of 'INFLATION', evolved into a universe with a small baryon number.

In 1967, the Russian physicist Andrei Sakharov outlined the three 'ingredients' necessary to generate a baryon asymmetry from initially symmetric conditions. The first requirement is that some particle reactions must violate baryon number. The second condition is a violation of CP symmetry (charge conjugation times parity). Finally, baryogenesis requires a departure from local thermodynamic equilibrium.

These three Sakharov ingredients may have been present in the early universe. The expansion of the universe can lead to nonequilibrium conditions. CP violation was discovered in 1964 by James Cronin and Val Fitch in reactions involving particles known as kaons. While violation of baryon number has never been observed, it is predicted to occur in several theories. The different possibilities for baryon-number violation have resulted in the development of different scenarios for baryogenesis.

In theories of the unification of the strong weak, and electromagnetic interactions into a GRAND UNIFIED THEORY (GUT), baryon number is predicted to be violated by the decay of supermassive particles of mass as large as 10^{16} times the mass of the proton. In the simplest versions of GUT baryogenesis, the supermassive particles were present in the thermal radiation, and as the temperature of the universe fell below the equivalent mass of the supermassive particle, they decayed preferentially into baryons. In the GUT scenario, baryogenesis occurred 10^{-35} seconds after the bang.

Unification of the electromagnetic and weak interactions involve a breaking of the electroweak symmetry where the Higgs field acquires a vacuum expectation value. Although broken today, at sufficiently high temperatures the symmetry should be restored. A phase transition associated with the breaking of electroweak symmetry is expected as the temperature of the universe fell below about 10^{15} K. Nonperturbative baryon-number violating processes are predicted to operate above the phase transition. A scenario of electroweak baryogenesis has been developed where the baryon asymmetry is generated around the time of the electroweak phase transition, around 10^{-12} seconds after the bang.

SUPERSYMMETRY theories have bosonic partners for the quarks and leptons. These particles, known as squarks and sleptons, are bosons, and may develop vacuum expectation values. Any change in the vacuum expectation value of a squark field results in a violation of baryon number. The age of the universe during baryogenesis is not specified in this scenario.

Other baryogenesis scenarios involve variations or combinations of the three scenarios above. Although the exact mechanism of baryogenesis is still a matter of intense speculation, the three Sakharov ingredients are expected to be present in the early big bang: CP violation has been

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observed in nature, the expansion of the universe naturally leads to departures from equilibrium, and baryon-number violation is expected to occur in a variety of ways.

Although the details of baryogenesis are unknown, laboratory experiments may help to unravel the mystery. Study of CP violation involving bottom quarks may lead to a deeper understanding of this ingredient. Knowledge of the electroweak Higgs mass is important for electroweak baryogenesis. Discovery of low-energy supersymmetry would have important implications for electroweak baryogenesis, as well as scenarios involving squark expectation values. Finally, observation of proton decay, or other direct evidence for grand unification, would be important for GUT baryogenesis scenarios.

Edward Kolb

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Baryons

Subatomic particles that are composed of three quarks and which are acted on by the strong nuclear force. This class of particle includes the nucleons (protons and neutrons), which are the constituents of atomic nuclei, and a variety of heavier, but short-lived, particles called hyperons.

Because the mass of a nucleon is nearly 2000 times greater than the mass of an electron, and all atoms and molecules contain at least as many nucleons as electrons, baryons account for more than 99.5% of the total mass of all the chemical elements in the universe. Matter that is composed of baryons is known as 'baryonic matter'. If the mean density of the universe is sufficiently high to ensure that it is 'closed' (a closed universe eventually ceases to expand and then collapses) or 'flat' (a flat universe is just, but only just, capable of expanding forever), at least 90% of it must exist in the form of dark, non-baryonic, matter, the nature of which is not yet known.

See also: atom, electron, fundamental forces, hadrons, neutron, proton, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bayer, Johann (1572–1625)

Lawyer and astronomer, born in Rhain, Germany. His catalog and atlas *Uranometria* (1603) depicts the positions of double the thousand stars in Tycho Brahe's catalog on which it is based, designating the brighter stars by the letters of the Greek alphabet, usually in brightness order, followed by Latin letters. This was instead of Arabic proper names or by AL-SUFI's awkward descriptions relative to the constellation figures, such as 'the star at the tip of the lesser bear's tail'. Bayer's lettering system is still used. Bayer included 12 new southern asterisms to fill in the region of the sky unseen by classical astronomers like PTOLEMY. They were: Apus, Chamaeleon, Dorado, Grus, Hydrus, Indus, Musca (Bayer called it Apis), Pavo, Phoenix, Triangulum Australe, Tucana and Volans, all still in use.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Be Stars

A Be star (pronounced 'bee-ee' star) is a non-supergiant B-type star whose spectrum displays or has displayed one or more Balmer lines in emission and Be is the notation for the spectral classification of such a star (see also CLASSIFICATION OF STELLAR SPECTRA). 'Classical' Be stars are believed to have acquired the circumstellar (CS) material that produces the Balmer emission through ejection of matter from their photospheres instead of from mass ACCRETION, and therefore are regarded as a separate subgroup distinct from the Herbig Be objects and interacting BINARY STARS. For the early B stars in our region of the galaxy, the fraction of Be/B stars is about 20%. This ratio drops to about 10–15% for late B stars and early B stars in the interior of our galaxy. The percentage of Be stars appears to be larger in the MAGELLANIC CLOUDS (25–45%). As a group Be stars are rapidly rotating. Spectral analysis, and more recently interferometric imaging, suggests that the CS material is in the form of a disk positioned about the equator of the star (figure 1). The Be-shell stars are classical Be stars that also exhibit narrow absorption lines, usually from moderately ionized species such as Fe II, as well as central narrow components to the Balmer emission lines that extend below the continuum. The current consensus is that the Be-shell stars are classical Be stars viewed nearly equator on.

The first Be star to be discovered was γ Cas in 1866 by A. Secchi, although this object has not endured as a prototype of the class. Based on the tendency of Be stars to display diffuse, rotationally broadened lines, in 1931 O Struve hypothesized that Be stars are rotating just short of their critical, or break-up, velocities, thus making it favorable for material at the star's equator to slowly dribble off into a CS disk. Although the disk geometry prevails, this model is no longer believed as there is no evidence that any Be star rotates faster than $0.85V_{\rm crit}$. Rapid rotation may play some role in the mass loss process, however. Spectroscopic observations of Be stars since 1980 have revealed that material is ejected from the photosphere, sometimes in a violent manner. An individual star can indeed oscillate between showing Balmer emission and apparently a pure photospheric spectrum. It is now clear that one must consider the star and disk as a whole in order to gain an understanding of what is called the 'Be phenomenon'.

The disk

The hallmark of a Be star is its CS disk that produces the strong Balmer line emission. A simple model for a Be disk is shown in figure 1. In first order, it is the kinematical motion of the CS gas in the disk that broadens the emission lines and gives the generic double-peaked line profile shown in figure 1. The arrows in the disk show the motion of the material and the gray tones map into the emission line profile. Material in the inner disk, where the radial motions toward–away from us are the largest, contribute most of the flux in the wings of the emission profile.

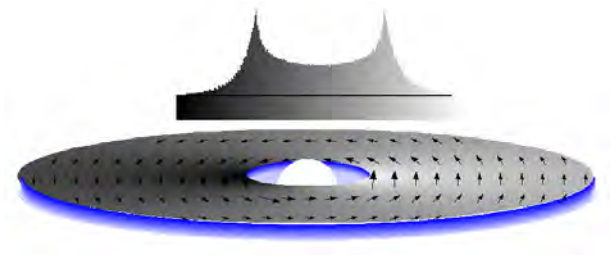


Figure 1. A generic representation of a Be star. A typical Balmer emission feature that the disk would produce is shown above the star. Illustration courtesy of S Owocki.

INTERFEROMETRY and IR photometry (see STELLAR PHOTOMETRY) suggest that the radial extent of the disk is $>20R_*$ and that the material displays Keplerian motion. The vertical extent of the inner disk is calculated from observation and theory in which the disk is assumed to be in hydrostatic balance (justified as disks can be stable for long periods of time). The inner disk appears to have a vertical height of $\sim 1R_*$ (Be-shell stars can show deep shell lines that have residual intensities of <10%), but at large radii the disk can flare to several stellar radii (not shown in figure 1). It is currently unknown whether there is a gap between the disk and the star as depicted in figure 1. Fitting of theoretical disk models to Balmer and Paschen emission lines suggests that the density, N_0, at the inner edge of the disk varies from 10^{12} to 10^{14} cm^{-3} for optically thin/thick disks. The density distribution in the disk is usually assumed to follow the formula: $N_{\rm r} = N_0(r/R_*)^{-m} \exp[-\frac{1}{2}(z/h)^2]$, where z and h are the vertical distance above the disk plane and the scale height, and $2 < m < 3.5$. Temperatures in the equatorial regions of B0–B5e stars range from 8000 to 5000 K and are 2–3 times higher at the upper–lower edges of the disk.

The profiles of emission lines in Be stars assume a variety of shapes both symmetric and asymmetric. When a star shows symmetrical emission lines spectral variablity is minimal, but long-term cyclical variability is commonplace in stars that display asymmetrical emission features. The star's inclination, i, to our line of sight has the greatest influence on the line profile, but the optical depth, τ, in the disk and the overall kinematic motion of the disk material are also important. This allows one to probe the density structure and kinematics in the disk by fitting theoretical profiles to observations of emission lines from ions of widely different abundances (e.g. Balmer and Fe II lines). Some representative Hα profiles of Be stars viewed at pole-on, intermediate and equator-on inclinations are shown in figure 2. The term pole on usually refers to objects with $i < 30°$; for equator-on stars $i > 80°$. Keplerian motion dominates the broadening of the line profile. The kinematical velocity width of the emission line depends on the projected Keplerian velocity, $v_{\rm K}$ (~400–500 km s^{-1}), of the material at the inner edge of the disk, $\Delta v_{\rm kin} = v_{\rm K} \sin i$. High-resolution spectra of Be stars seen at low–intermediate inclinations often reveal inflections,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

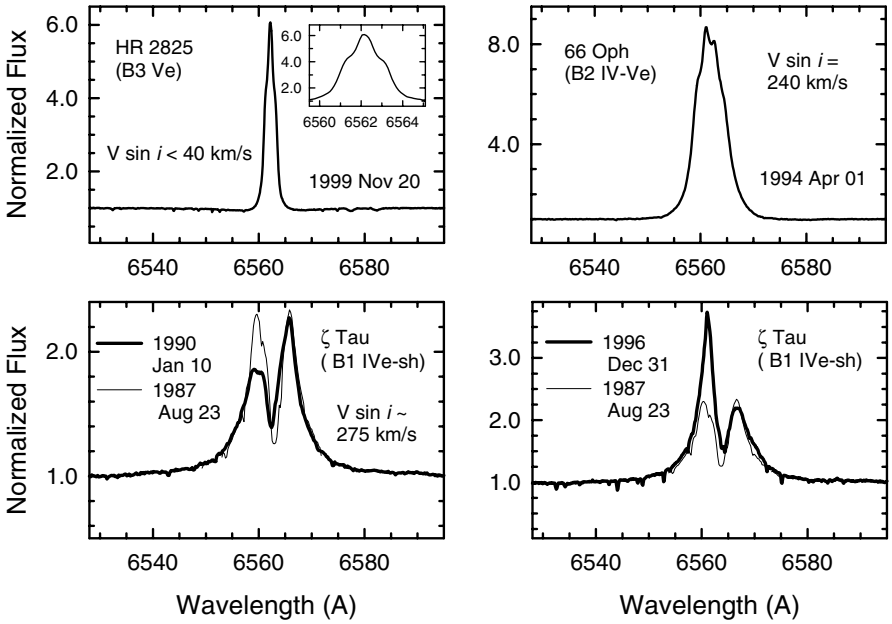


Figure 2. Representative Hα profiles observed in Be stars. These CCD spectra are from the Coudé Feed Telescope at Kitt Peak National Observatory.

or shoulders to the Hα or Hβ emission lines (cf. inset to the profile in the upper-left corner of figure 2). This so-called winebottle appearance is a result of non-coherent scattering in an optically thick disk. Photons scattered into the wings of the line deep in the disk readily escape from the star. In the wings of an optically thick emission line one is seeing radiation from deeper in the disk where the source function is larger. For Be stars seen at large inclination angles, such as the Be-shell stars, the absorption of the stars's light as it passes through the disk material produces a deep central reversal in the emission feature.

The Balmer emission lines seen in Be-shell stars frequently undergo long-term (~2–10 yr) periodic or quasi-periodic oscillations in the relative intensities of the violet-shifted (I_V) and red-shifted (I_R) components to the line. These are termed V/R variations, where V and R refer to the emission above the local continuum: $V = I_V - I_{\rm cont}$ and $R = I_R - I_{\rm cont}$. It is currently thought that the V/R oscillations are caused from the precession of a density enhancement or wave in a quasi-Keplerian disk. These global disk oscillations are believed to also cause the long-term profile variations seen in the Be stars with asymmetrical emission lines. A typical example of a long-term V/R variation in Hα in a Be-shell star is shown in figure 2. Variability in the peak intensity of the emission from cycle to cycle is typical.

The wind

The mass flux from the stellar winds in Be stars is generally larger than in non-emission B stars; however, there appears to be a prominent dependence on the inclination of the star to our line of sight. Information on the winds is obtained from observations of the UV resonance lines (e.g. N V λ 1238.8, 1242.8, C IV λ 1548.2, 1550.8, Si IV λ 1393.8, 1402.8, Si III λ 1206.5, and Al III λ 1854.7, 1862.8). N V tends to be observed in only the earliest Be stars. The wind absorption seen in stars viewed equator on or at intermediate inclinations is typically strong and frequently contains discrete absorption components (DACs), while pole-on objects usually show either no or very weak wind features. DACs are discernible absorption features superimposed upon apparently smooth underlying wind lines. These relatively narrow (usually $\Delta V < 200$ km s^{-1}) features are thought to be formed as a result of shocks in the wind. Some typical C IV wind features are displayed in figures 3 and 4. The maximum velocity of the wind absorption is in the range (0.4–2.0$V_{\rm esc}$) and appears to increase toward earlier spectral types.

Although the UV wind features are very prominent, they are actually formed in a region of much lower density than the optical–IR emission lines. The mass loss rates (~$10^{-10} M_\odot$ yr^{-1}) computed from the UV wind lines are about 50–100 times smaller than those determined from IR photometry and fitting of the optical–IR emission lines. It is currently believed that the high-velocity, low-density stellar wind flows out of the equatorial plane toward intermediate latitudes and has very little influence on the more massive Keplerian disk.

Variability is commonplace in the winds of Be stars, both short and long term. Examples are shown in figure 3 and figure 4 for winds with and without DACs. The upper panel in each figure shows a Be star that displays daily, cyclical wind variations, while the lower panel illustrates typical long-term (yearly) quasi-cyclical variations. Most

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

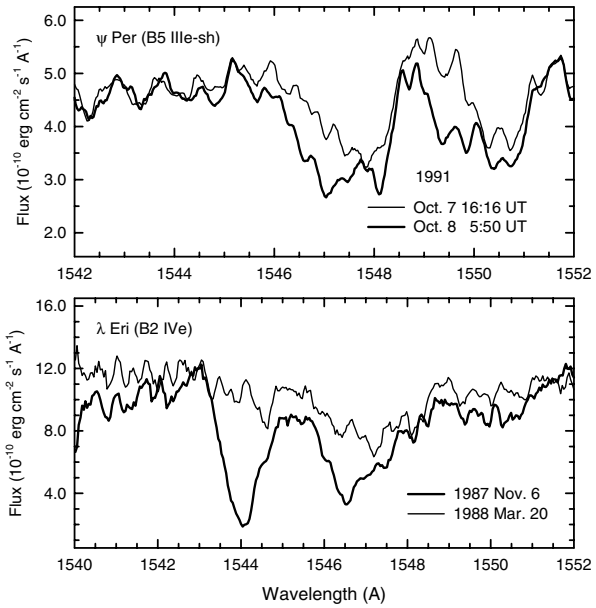


Figure 3. Typical wind variability in Be stars that do not show highly structured winds. Upper panel: a Be star that displays short-term daily cyclic variability. Lower panel: an object in which the wind varies on a time scale of ∼year.

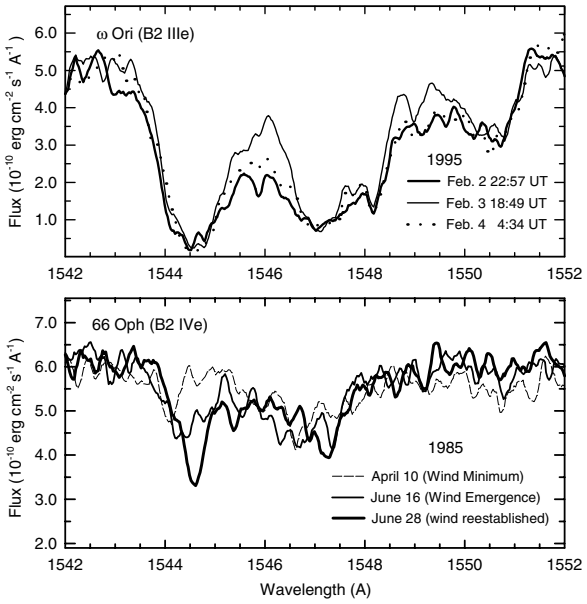


Figure 4. As in figure 3 but for Be stars that display prominent DACs in the wind.

of the wind variability (especially long term) is usually in the DACs that are probably caused by fast-moving localized flows colliding with the exterior wind material. Since short-term variability patterns (red–blue movement) often resemble those observed in O star winds, the DACs may be formed in co-rotating interaction regions. Long-term variability, however, frequently involves DACs moving from blue to red as the wind strengthens.

The photosphere

The current consensus is that the material in the CS disk of a classical Be star originates on the photosphere. Observation of μ Cen, a Be star that has recently been in the process of re-building its CS disk that it lost in the late 1970s, suggests that mass is lost from a localized site of unknown size and the mass loss mechanism must produce a rapid injection of material into the disk. Initially the CS material resides on one hemisphere of the star, but it circularizes into a Keplerian disk in just a few days. Two hypotheses that are currently being discussed are mass loss induced by non-radial pulsations (NRPs) (see NON-RADIAL STELLAR PULSATIONS) and localized magnetic fields.

Since 1980, spectroscopic data of high resolution and S/N from electronic detectors have revealed evidence for NRPs as well as transient micro-emission activity associated with mass ejection from the photosphere. Evidence for NRP comes from observations of weak (∼1–2% $I_{\rm cont}$) absorption components, often called 'bumps', that slowly move from blue to red across the photospheric line profile cyclically, usually with a period less than the star's expected rotation period of ∼1 day. An example of this behavior is shown in figure 5. The photospheric velocity field is given by $V(r,\theta,\phi) = A(1, k\,\partial/\partial\theta, k/\sin\theta\,\partial/\partial\phi)Y_l^m(\theta,\phi)\,e^{i\sigma t}$, where k is the ratio of horizontal to vertical velocity amplitude, Y_l^m is a spherical harmonic function and σ is the NRP frequency in the co-rotating frame. l and m are the pulsation modes. There are m meridians passing through the poles perpendicular to the equator. l is the total number of meridians on the surface and $l - |m|$ is the number of nodes between the poles. Be stars typically show low-order modes with $l = |m|$, or sectorial modes that resemble the sections of an orange. Usually the wave motion is retrograde and beats backwards against the star's rotation with a period of $mP_{\rm obsd}$. For these low-order (g) modes, horizontal motions dominate, as $k \sim \sigma^{-2}$, and the wave crests are hotter than the troughs. The patches on the star containing the crests will contribute more to the integrated line profile than will the trough regions. Consider an $l = 2$ mode in which there are two crests and troughs. The star should be brightest as the absorption 'bump' crosses the center of the line. When the crest is near the stellar limb the star will be fainter and the line should show its greatest distortion. Both effects are observed in many Be stars.

Several types of transient behavior associated with apparently localized mass ejection have been identified in Be stars. These transients are most often observed in the He I lines of Be stars with meager CS disks. Emission flares can abruptly appear and wane in <1 h. Sometimes 2–3 emission components at discrete velocities display ordered variations over the course of a few hours. Weak, central absorption features flanked by weak emission can appear–disappear in <1 h. Transient, high-velocity absorption

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

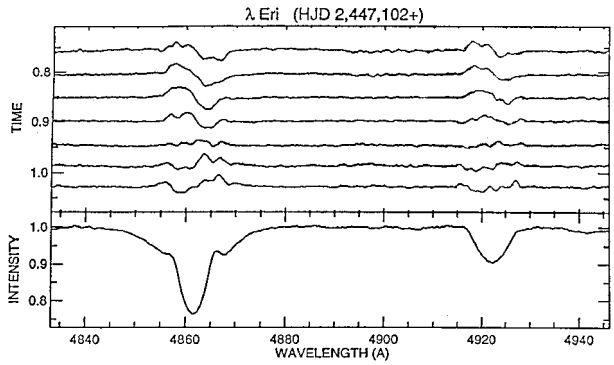


Figure 5. Example of blue–red traveling 'bumps' in photospheric line profiles that suggest the presence of NRPs in a low-order mode. The series of profiles in the upper panel are differences from the mean Hβ and He I λ 4922 profiles displayed in the lower panel. Illustration courtesy of D Gies.

components can appear on the violet side of the line in less than 15 min and fade in 1–2 days. All these short-term phenomena could result from magnetic fields, but so far no magnetic field has been detected in a classical Be star. NRP could be the cause of the transient activity. Some of the most exciting research on Be stars to come will focus on the details of the mass ejection from the photosphere, the process of building the disk and the behavior of the disk as it receives material from a localized site on the star.

Bibliography

Information on contemporary research on Be stars can be found in the proceedings from IAU Symposium No 162

Balona L, Henrichs H F and Le Contel M (ed) 1994 *Pulsation, Rotation, and Mass Loss in Early-Type Stars* (Dordrecht: Kluwer)

and IAU Colloquium No. 175,

Smith M A, Henrichs H F and Fabregat J (eds) 2000 *The Be Phenomenon in Early-Type Stars* (San Francisco, CA: Astronomical Society of the Pacific)

A good collection of Hα and Fe II emission line profiles and summary of important physical processes that produce them is

Hanuschik R W, Hummel W, Sutorius E, Dietle O and Thimm G 1996 Atlas of high-resolution emission and shell lines in be stars—line profiles and short-term variability *Astron. Astrophys. Suppl. Ser.* **116** 309–58

A good treatise on wind parameters is

Prinja R K 1989 Ultraviolet observations of stellar winds in be and 'normal' B non-supergiant stars *Mon. Not. R. Astron. Soc.* **241** 721–52

An excellent summary of the modeling of NRP is

Gies D R 1991 Nonradial pulsation and line profile variability in early-type stars *Rapid Variability of OB-Stars: Nature and Diagnostic Value* ed D Baade (Garching bei München: ESO) pp 229–43

Geraldine J Peters

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Becklin–Neugebauer Object (BN Object)

One of the most powerful infrared sources known, located in the Kleinmann–Low Nebula behind the Orion Nebula, and discovered by Eric Becklin and Gerry Neugebauer in 1967. Its position is RA $05^{h}\ 35.2^{m}$, dec. $-05^{\circ}\,22'$. It is thought to be a very young and massive B star embedded in a dense expanding dust envelope which absorbs the star's visible light and re-emits it in the infrared.

See also: Kleinmann–Low Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Beer, Wilhelm (1797–1850)

Berlin banker and amateur astronomer, collaborated with JOHANN MÄDLER in a map of the Moon, in its organization following closely the plan laid out by the less well known selenographer W G LOHRMANN. Suggested with Mädler in 1837 the advantages of an observatory on the Moon (continuous observation over a 14 day lunar night, for example: they did not know that there is no lunar atmosphere so that stars are visible during the lunar day too).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Beijing Astronomical Observatory

The Beijing Astronomical Observatory (BAO) is primarily engaged in astrophysical research. The modern Beijing Observatory was built by the Chinese Academy of Sciences in 1958. Its headquarters are located in Beijing, and there are five observing sites: the Xinglong, Miyun, Huairou, Shahe and Tianjin stations.

Instruments include a 2.16 m (84 in) optical telescope, the largest in the Far East; a multichannel solar magnetic telescope; a meter-wave aperture synthesis radio telescope; and a 1.26 m (50 in) infrared telescope. BAO is also constructing a Large Area Multi-Object Spectroscopic Telescope. In 1991, an observing site in the southern hemisphere was set up in collaboration with the San Juan University of Argentina.

The main research programs include stellar physics, extragalactic studies, solar physics, radioastronomy, an atomic time and frequency service, solar activity prediction, artificial satellite observation and the history of astronomy. A number of discoveries have been made during searches for supernovae, quasars, active galactic nuclei and near-Earth asteroids.

For further information see
http://www.bao.ac.cn/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bell, Susan Jocelyn [name following her marriage Burnell, or Bell-Burnell] (1943–)

Radioastronomer, born in York, England. She studied at Glasgow and Cambridge, where in 1967 she was a research student working with ANTHONY HEWISH on a radioastronomy project on interplanetary scintillation. She noticed an unusually regular signal, shown to be bursts of radio energy at a constant interval of just over a second—the first pulsar. Hewish was awarded the Nobel prize for this discovery, circumstances which led to Bell's adoption as a feminist icon. Bell became professor at the Open University.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bennett, Willard Harrison (1903–87)

Physicist, inventor, born in Findlay, OH, invented the radiofrequency mass spectrometer, a device that measures the mass of ions. This instrument has been much used by space scientists to identify the composition of distant places in the solar system, ever since it first went into space in 1957 aboard the Russian satellite Sputnik.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

BeppoSAX (Satellite for X-ray Astronomy)

Italian–Dutch gamma ray and x-ray astronomy satellite launched in April 1996. Named in honour of Italian physicist Giuseppe Occhialini, whose nickname was 'Beppo'. Used for spectroscopic and time variability studies of x-ray sources in the energy band 1–200 keV, including an all-sky monitoring investigation of transients in the 2–30 keV energy range. In 1997, played a major role in identification of gamma ray bursters by pinpointing x-ray emissions from these events, so enabling confirmation of their cosmological distances. The board of administrators of the Italian Space Agency has approved the continuation of BeppoSAX operations until at least 30 April 2002.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bessel, Friedrich Wilhelm (1784–1846)

Astronomer, geodesist, mathematician, born in Minden, Westphalia, Germany. At the age of 26, became director of Frederick William III of Prussia's new Konigsberg Observatory, where he remained for the rest of his life. Published (1838) the first recognizedly accurate stellar parallax (of 61 Cygni), the earliest accurate method for measuring stellar distances (10 light years for 61 Cygni) (but see WILHELM STRUVE's and THOMAS HENDERSON's claims to priority). Discovered the orbital deflections of Sirius and Procyon due to their then unseen white dwarf companions. These were results of a monumental astrometric program to determine the positions and proper motions of stars—it could be called the beginning of modern astrometry. *Bessel functions*, originally encountered by Daniel Bernoulli, were systematically investigated by Bessel and named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Beta Pictoris

A white-dwarf star of apparent magnitude 3.85, found by the IRAS survey in 1983 to be a very strong emitter of infrared radiation. Following this discovery, extensive studies with large optical telescopes revealed the presence of a disk of dust surrounding the star—possible raw material for the formation of a planetary system. This is particularly interesting as it is a star similar in size to the Sun (1.4 solar diameters), although rather hotter (spectral type A3V). Its luminosity (absolute magnitude 2.4) is about eight times that of the Sun. It is 63 light-years distant, its parallax is 0.052″.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Beta Regio

A prominent upland region of Venus, situated to the north of the planet's equator and centered at 25.3 °N, 77.2 °W. In its greatest dimension it measures 2869 km. It is one of only a handful of Venusian features not to bear a female name. The first features that showed up in early radar surveys of the planet were given the names of letters of the Greek alphabet. When the convention of assigning female names was adopted, a few names that had become well established were allowed to survive. Beta Regio is dominated by three features: two large volcanic calderas, Rhea Mons, and, to its south, Theia Mons, and a prominent rift known as Devana Chasma. This rift passes through both calderas and continues southwards, with a total length of 1616 km. The whole region is considered to be one of the youngest features of the planet, and probably sits atop a local upwelling of the underlying mantle. The lava flows here are some of the most recent on Venus, and it may be that the region is currently volcanically active.

See also: Venus: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Betelgeuse

The star α Orionis, marking the right shoulder of the Hunter in the ancient constellation figure. It is a red giant of spectral type M2Ib, apparent magnitude 0.45. At a distance of 427 light-years (parallax 0.008″), its apparent brightness arises from its high intrinsic luminosity (absolute magnitude (5.1) which is about 14 000 times that of the Sun. It was the first star to have its diameter reliably determined, by A A Michelson (1852–1931) at the Mount Wilson Observatory, using a 20 foot beam interferometer attached to the 100 inch Hooker reflector. The observed angular diameter of 0.047″ is equivalent to an actual diameter more than double the size of the orbit of Mars. This implies a volume 160 million times that of the Sun, yet its mass is estimated at only 20 solar masses: it is thus a star of exceptionally low density.

Betelgeuse is a semi-regular pulsating variable, its magnitude varying by ± 0.15 over a period averaging about 6.4 years. It is also a radio source and an emitter of strong infrared radiation. Observations by the Infrared Astronomical Satellite (IRAS) indicated that three concentric shells, ejected from the surface of the star in the past, are the sources of the long-wavelength radiation. There are several very faint optical companions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bethe, Hans Albrecht (1906–)

Physicist, born in Strasburg, Germany (present-day France), Nobel prizewinner (1967). As professor of physics at Cornell University he worked out the nuclear reactions occurring in the Sun. He worked on the development of the atomic bomb in the Manhatten project, and, after the second World War, pursued again research on stellar nuclear energy sources and the origin of the chemical elements in the alpha-*beta*-gamma theory (Bethe's is the second Greek letter of the theory's name, *see* RALPH ALPHER, GEORGE GAMOW).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bevis, John (1695–1771)

Doctor and amateur astronomer, born in Old Sarum, Wiltshire, England. Bevis was the only person ever to have telescopically observed (through patchy cloud from Greenwich) the occultation of one planet by another, namely the occultation of Mercury by Venus on 28 May 1737 (the next chance, in theory, is an occultation of Neptune by Mercury in 2067). URBAIN LEVERRIER used Bevis's observations to check his calculations of the motions of the two planets. Bevis erected an observatory in Stoke Newington with which he observed star positions and created a star atlas called *Uranographia Britannica*. It was completed in 1750, but the printer went bankrupt and the printing was never distributed. Less than 20 proof copies survive, and the complementary catalog was never published. Bevis discovered the Crab Nebula, apparently in 1731, and illustrated it on the atlas. Messier had access to a copy of this atlas, referring to it as the *English Atlas* several times, e.g. in the descriptions for the objects M1 (the Crab Nebula, which Messier independently discovered but acknowledging the earlier discovery by Bevis), M11, M13, M22, M31 and M35.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Biela, Wilhelm von (1782–1856)

Austrian army officer, astronomer. Observed in 1826 the periodic comet named after him, and realized that it was identical to the comet seen in 1772 and 1805. The comet broke up in the 1850s and was not seen after the 1852 apparition. Its debris became responsible for the Andromedid (or *Bielid*) shower of meteors seen annually in November in the second half of the nineteenth century, until their orbit shifted out of the annual path of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bielids

A periodic meteor shower that produced November meteor storms after the break-up of Comet 3D/Biela in the 1850s. The shower, also known as the *Andromedids*, had its radiant in the constellation Andromeda. The comet, which had a period of 6.6 years, was last seen in 1852. In 1872, when it should have made a favorable return, there occurred instead a meteor storm peaking at 6000 per hour. The shower returned again in 1885 with a spectacular display touching 75 000 meteors per hour. Weaker activity was seen in 1899 and 1904, after which gravitational perturbations by Jupiter shifted the Bielid stream away from the Earth's orbit. It may return again early in the twenty-second century.

See also: Comet Biela.

Andromedids *see* Bielids

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Biermann, Ludwig Franz Benedikt (1907–86)

German astrophysicist, became director of the astrophysics section of the Max-Planck-Institut für Physik und Astrophysik, first at Göttingen and later in Munich. Biermann applied atomic physics and plasma hydrodynamics to problems in stellar interiors, the solar chromosphere and corona, the Galaxy, the solar system and comet tails (which led him to predict the solar wind).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Big Bang Theory

The theory which asserts that the universe originated a finite time ago by expanding from an infinitely compressed state. According to this model, space, time and matter originated together, and the universe has been expanding ever since. Key stages in the history of the Big Bang universe are summarized below.

According to the standard Big Bang model, the very early universe consisted of an extremely high-temperature mixture of radiation (photons) and particles. In accordance with the equivalence of mass and energy inherent in Einstein's special theory of relativity, collisions between high-energy photons would have transformed radiation energy into particles of matter. To be consistent with the conservation laws of particle physics, a collision between sufficiently energetic photons creates a particle–antiparticle pair (e.g. an electron and a positron, or a quark and an antiquark). Conversely, a collision between a particle and its antiparticle results in the annihilation of the particle–antiparticle pair and their conversion into photons.

As the universe expanded and cooled, photon energies quickly declined below the thresholds at which they could continue to create particle–antiparticle pairs. Thereafter, particles and antiparticles rapidly collided and annihilated each other. Baryons and antibaryons (particles such as protons, neutrons and the quarks of which they are composed) underwent mutual annihilation about one millionth of a second after the initial event, when the temperature dropped below 10^{13} K. Less-massive electrons and positrons experienced a similar event a few seconds later, when the temperature had dropped to about 5×10^{9} K. Had there been an exact equality between the numbers of particles and antiparticles, virtually no matter would have remained to form stars and galaxies. Because the destruction of each particle–antiparticle pair produced a pair of photons, and the ratio of photons to baryons in the present-day universe is nearly a billion to one, it follows that the excess of particles over antiparticles in the early universe must also have been about one in a billion.

About 100 seconds after the beginning of time, when the temperature had dropped below 10^{9} K, nuclear reactions combined some of the protons and virtually all the neutrons into nuclei of helium, together with small quantities of other light elements and isotopes such as deuterium, helium-3 and lithium. At the end of this phase there were about 11 hydrogen nuclei (protons) for every helium nucleus. Because helium nuclei are four times heavier than hydrogen nuclei, this corresponds to a ratio, by mass, of hydrogen to helium of about 73:27. The observed relative abundances of the lightest elements very closely matches the predictions of the Big Bang theory.

Thereafter, the expanding universe consisted of a mixture of atomic nuclei, electrons and photons in thermal equilibrium (frequent collisions between photons and particles ensured that matter and radiation shared a common temperature). A few hundred thousand years after the initial event, when the temperature everywhere had declined to a few thousand kelvin, nuclei were able to capture electrons to form complete neutral atoms. This process mopped up the free electrons that had been responsible for scattering photons and, thereby, making space opaque to electromagnetic radiation. The close connection between matter and radiation was broken (this event is called the 'decoupling' of matter and radiation), space became essentially transparent, and the primordial radiation content of the universe was free thereafter to spread ever more thinly through the expanding volume of space. The dilute, redshifted remnant of the primordial radiation is detectable today as the cosmic microwave background radiation.

Some time later, probably within the first billion years after the decoupling, clumps of matter aggregated together to form galaxies, clusters, superclusters and the other large-scale features of the universe. How, when, and in what order these various structures appeared is a matter of debate.

The Big Bang theory is consistent with the observed expansion of the universe (provided that the age of the universe exceeds that of the oldest stars), the relative abundance of the lighter chemical elements, and the existence and properties of the cosmic microwave background radiation. Nevertheless, although certain other issues have been addressed by the 'inflationary' hypothesis, many questions remain unresolved.

See also: cosmic microwave background, cosmological model, cosmology, expanding universe, inflationary universe, primeval fireball.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Big Bear Solar Observatory

Big Bear Solar Observatory (BBSO) is located at the end of a causeway in a mountain lake more than 2 km above sea level. The site has more than 300 sunny days a year and a natural inversion caused by the lake which makes for very clean images. BBSO is the only university observatory in the US making high-resolution observations of the Sun. Its daily images are posted at http://www.bbso.njit.edu. It specializes in the study of the Sun's magnetic fields and their evolving nature. The BBSO uses daily images to forecast solar magnetic storms which can impact satellites, communication and the power grid. BBSO is operated by New Jersey Institute of Technology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Big Dipper

Alternative name, used particularly in the USA, for the asterism in Ursa Major more usually known as the Plough.

See also: Plough.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary Stars in Globular Clusters

Globular clusters have long been known to be among the richest stellar groupings within our Galaxy, but for many years they were believed to be largely devoid of the most minimal stellar group: binary stars (see BINARY STARS: OVERVIEW). For many years, the only evidence that any binaries existed in these clusters came from the presence of BLUE STRAGGLERS—stars that appear to be significantly younger than the vast majority of the cluster stars. The nature of blue stragglers remains somewhat unclear, but an early and persistent explanation for their existence is that mass transfer between stars in binary systems prolonged the lifetimes of the mass-gaining stars. Other mechanisms involving single stars can also conceivably explain the existence of blue stragglers, so for many years they offered only weak circumstantial evidence that binary stars reside in globular clusters.

From the early 1950s, when blue stragglers were first discovered, until the late 1980s, more direct evidence for the existence of binaries in globular clusters remained elusive. For example, some of the earliest comprehensive studies of the RADIAL VELOCITIES of the RED GIANT STARS in clusters, in the 1970s, revealed that very few if any of these stars showed the long-term velocity variations expected for SPECTROSCOPIC BINARY STARS. By the mid 1980s it was generally accepted that the binary frequency—the fraction of stars that appear single but that are actually binary stars—of globular clusters was probably under 10%, and certainly much lower than the 50% frequency observed among stars near the Sun.

The lack of binaries in globular clusters seemed consistent with some expectations of that time. Since the stars in globular clusters are closely packed together, they frequently pass quite close to one another. Simple models of the interactions of binaries and single stars in these environments suggested that binaries with initially wide separations (e.g. similar to the Earth–Sun separation or larger) would be disrupted. It was also believed that the low abundance of heavy elements in most globular clusters might somehow inhibit the formation of binary stars.

However, from the standpoint of the stability of globular clusters, the complete lack of binaries became an increasingly perplexing problem. The gravitational interactions of stars within clusters tends to drain energy systematically from the cluster core, leading ultimately to a core collapse to an infinite mass density in a finite time. Since no clusters show convincing evidence of having collapsed in such a dramatic manner, it became clear that something must eventually stop the core collapse process. The easiest way to do so would be to form binary stars in the cluster cores during the late stages of the collapse. The formation of even a few 'hard' binaries—binary stars with very short orbital periods—could entirely halt core collapse in even the most massive clusters. The problem was finding evidence that these binaries actually existed.

When the first X-RAY TELESCOPES were lofted into Earth orbit, one surprising result was the discovery of bright sources seemingly associated with globular clusters. With time, it became clear that these x-ray sources represent short-period binary systems in which mass is transferred from a normal star onto a companion NEUTRON STAR. The frequency of such x-ray binaries is much higher in globular clusters than in the rest of the Galaxy, so they form preferentially in the clusters, perhaps as a result of core collapse. However, many clusters that should have gone through core collapse do not contain x-ray binaries, so other types of binaries must exist in these objects to have halted their collapse.

One of the more unusual types of binaries in globular clusters was discovered with radio telescopes. In 1987, the first PULSAR within a globular cluster (Messier 28, or

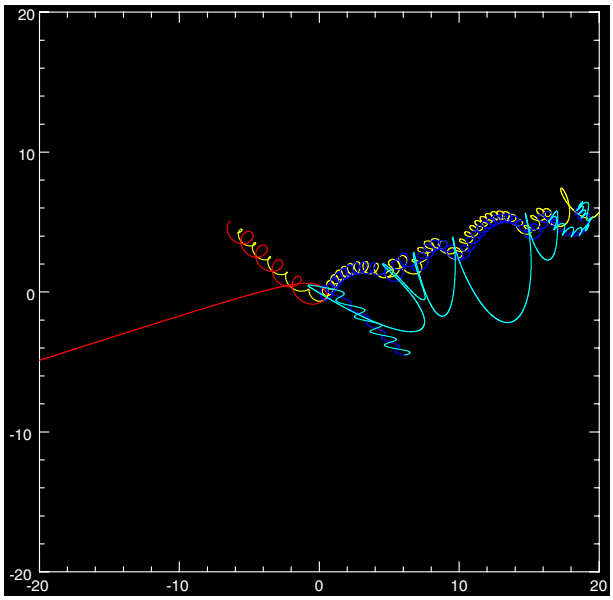


Figure 2. A color plot of the collision of two binary stars in a globular cluster. The lines show the trajectories of the four stars. Initially, two binaries enter the region, one from the upper left (denoted by the red and yellow lines spiraling around one another) and the other from the lower right (as shown by the dark and light blue lines). As the binaries collide, the stars interact strongly. The first effect is that the star denoted by the red line is rapidly ejected into the surrounding cluster as it shoots off the picture to the left. Left behind is a triple system comprised of the stars traced by the yellow, dark blue and light blue lines. However, this triple is not entirely stable as it swaps stars between an inner tight binary system with a more loosely bound star traveling in a larger orbit. In many cases, one of the stars of the triple is eventually lost from the system, leaving one short-period binary and two free stars where there had originally been two intermediate-period binary stars. Such collisions are common in globular clusters, and may provide a way to form contact binaries even in low-density clusters. The details of any given binary–binary collision depend sensitively on the initial configuration of the binaries, so it is still impossible to model any specific interaction within an actual globular cluster. However, the statistical effects of collisions such as the one shown here can be estimated and applied to clusters. (Courtesy E S Phinney and S Sigurdsson.) **This figure is reproduced as Color Plate 5.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

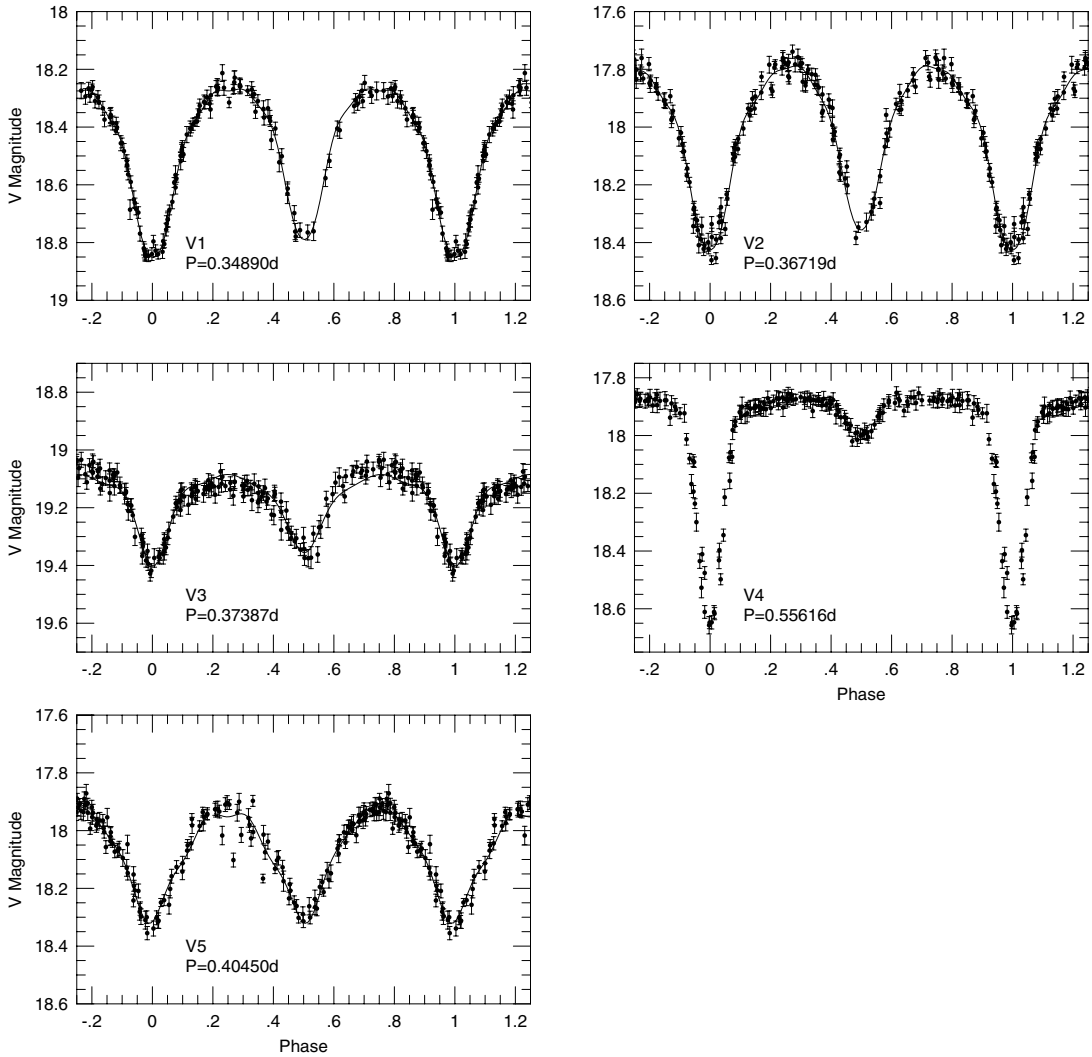


Figure 1. Light curves for five eclipsing binary stars in the globular cluster M 71. Four of these light curves indicate that the binary stars are contact systems in which the two stellar components are in physical contact as they orbit about their center of mass. The star labeled V4 exhibits nearly flat portions in its light curve between the obvious eclipses. This indicates that the binary is a detached system in which the stars are physically separated. All of these binaries have orbital periods of less than one day, while the individual stars are comparable in size to the Sun.

simply M 28) was discovered. These rotating neutron stars exhibit ultra-regular brightness variations on time-scales ranging from 0.002–10 s. Because these natural clocks are so precise, the modulation of their periods can reveal clearly when pulsars are part of a binary system. At least 11 such binary pulsars with normal stellar companions are now known to exist in globular clusters, generally with orbital periods of less than 3 days. The existence of very short pulsation periods of many *single* pulsars found in clusters offers additional indirect evidence of binaries: mass transfer within close binary systems can effectively 'spin up' the pulsars to their observed ultra-fast rotation rates. Subsequent interactions with other single or binary stars in the clusters can disrupt the binaries, leaving the single, short-period pulsar behind. In fact, this seems to be the *only* viable way to produce short-period pulsars in globular clusters. The natural tendency for isolated pulsars to spin more slowly with time would suggest that no short-period pulsars should exist among the ancient stars of a globular clusters.

Some of the first searches in the optical centered on the detection of cataclysmic binaries (CBs) in globular clusters (see CATACLYSMIC BINARIES: CLASSICAL AND RECURRENT NOVAE). CBs are closely related to the bright x-ray sources, except that the compact companion is a WHITE DWARF rather than a neutron star. NOVAE—CBs that undergo particularly violent eruptions on timescales of many years—had been observed near the cores of two globular clusters, but the initial searches for less luminous CBs were frustratingly inconclusive. Only in recent years have ground- and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

space-based observations successfully identified normal CBs in many clusters using optical and x-ray telescopes.

Prior to the discovery of numerous CBs in clusters, examples of more mundane binaries began to reveal themselves in globular clusters. Spectroscopic surveys of red giants started to identify stars that appeared to vary in a way that would be expected for normal binary stars. These spectroscopic binaries have orbital periods ranging from many months to 10 years or longer and demonstrated that relatively wide binaries could survive in many cluster environments. Studies of the changes in the brightness of stars in globular clusters also revealed that some of these are ECLIPSING BINARY STARS. Such systems contain stars in close proximity that, because of the fortuitous alignment of their orbital plane with the Earth, exhibit periodic eclipses as the stars alternately pass in front of one another. Light curves for five such binaries in the metal-rich globular cluster M 71 are shown in figure 1. Both spectroscopic and eclipsing binaries provide indisputable evidence that binary stars do exist in globular clusters. Surprisingly, the numbers of both spectroscopic and eclipsing binaries in clusters imply that, rather than being deficient of binaries, globular clusters may have binary frequencies in the range 25–60%. Such values would be considered nearly normal for stars located in the vicinity of the Sun.

The most recent surveys of many clusters have largely confirmed this result. Indeed, no clusters studied sufficiently carefully have failed to reveal either spectroscopic or eclipsing binaries or blue stragglers. Though the latter may form via stellar collisions or as a result of complex interactions between wide binary stars, these observations have strongly confirmed that binary stars are *not* rare in globular clusters, simply hard to find!

Blue stragglers re-entered the subject of cluster binary stars as it became clear that 5–10% of these stars are eclipsing systems. Many of these binary blue stragglers turn out to be contact binaries, systems in which the two orbiting stars are in physical contact. Because the stars significantly distort each other, they are not spherical in shape and are always either entering or leaving eclipse; thus, their light curves are never constant (four of the five stars whose light curves are shown in figure 1 are contact binaries). The components of such contact binaries cannot avoid a future fate in which the two stars coalesce into a single, more massive star. The resulting merged object would become a blue straggler. Especially in the lower-density globular clusters, this appears to be the most likely way to form blue stragglers (though see below for another possible mechanism that may operate in low-density clusters).

Now that binary stars are routinely detected in globular clusters, new questions have been raised. For example, is the mixture of binary star types and orbital periods similar to that observed near the Sun? Certainly, clusters contain far more x-ray binaries and a large proportion of short-period binary pulsars than among field stars throughout the Galaxy. It is also clear that the longest-period binaries in clusters are preferentially destroyed. Short-period systems can be created preferentially in dense clusters, either through star–star captures (this is rare except in the densest environments) or via 'collisions' of wide binaries (more common; see figure 2 for an example of what this may look like). Different clusters appear to affect their binary star populations in different ways, but we lack a good understanding of why this is the case or even if the binary star populations depend on global globular cluster properties such as total mass, or chemical abundance. The origin(s) of cluster binaries is still also under debate. There is clear evidence that some binaries are 'primordial'—that is, they were formed when the cluster itself was formed. But dynamical processes in some clusters—particularly the densest clusters—clearly can form binaries via a variety of mechanisms that have only begun to be explored in detail.

Bibliography

Bailyn C 1995 Blue stragglers and other stellar anomalies: implications for the dynamics of globular clusters *Ann. Rev. Astron. Astrophys.* **33** 133

Milone E F and Mermilliod J-C (ed) 1996 *The Origins, Evolution, and Destinies of Binary Stars in Clusters (Astron. Soc. Pacific Conf. Ser. 90)* (contains a variety of technical articles on cluster binaries)

Mario L Mateo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary Stars in Planetary Nebulae

The classical view of PLANETARY NEBULAE (PNe) is that they are ejected from all single stars of low to intermediate mass at the end of the asymptotic-giant-branch stage. However, observations made over the past two decades have made it increasingly clear that binary stars actually play a significant, and possibly dominant, role in the PN ejection process.

Common envelopes and UU Sagittae

Two key events in this subject occurred almost simultaneously in the mid-1970s, and provide a beautiful example of the interplay of theory and observation in astrophysics.

One of these events was an apparent theoretical resolution of the puzzle of the origin of extremely close binaries containing WHITE-DWARF (WD) components. These systems include the short-period cataclysmic binaries and close but detached 'pre-cataclysmic' main-sequence–WD pairs such as V471 Tauri (an eclipsing binary in the Hyades cluster, containing a hot WD and a K-type main-sequence companion, with an orbital period of only 12.5 h). WDs are the cores of RED GIANT STARS, yet for these extremely close binaries the entire system is much smaller than the red-giant progenitor of the WD. It was realized that, in fact, the companion main-sequence star must recently have been *inside* the red giant. This insight led to the concept of 'common-envelope' (CE) interactions during the evolution of binary stars.

The CE scenario starts with a binary that is wide enough for the more massive component to grow to red-giant dimensions before any interaction with its companion occurs. Only then does the red giant swallow its companion, which encounters frictional drag as it orbits within the low-density envelope of the red giant. The orbit spirals down, transferring its energy to the CE. Eventually, the CE is ejected, leaving a much closer binary containing the main-sequence star and the remnant core of the red giant. The core then evolves to high surface temperature, pouring out ultraviolet radiation that ionizes the ejected envelope and making it visible as a PN. Next, after the PN has dissipated into space, the core enters the WD cooling sequence, producing a V471-Tau-like binary. Eventually, as the orbit contracts owing to gravitational radiation or other mechanisms, the main-sequence star will start transferring mass to the WD, transforming the system into a cataclysmic or nova-like variable. It is now generally believed that the CE mechanism is the dominant way in which cataclysmic binaries are made. Moreover, the system may eventually undergo a *second* CE interaction, when the main-sequence star attempts to become a red giant itself, leading to a very close, double WD binary. Such systems may be the progenitors of type Ia supernovae.

This theoretical scenario predicts the existence of very close binaries inside PNe, which indeed would be the 'smoking gun' showing that the CE interaction actually does occur in the real universe. The second important event in the 1970s was the discovery of a close-binary PN nucleus (PNN), found in the same year that the first theoretical papers predicting its existence were published. This was UU Sagittae, the central star of the planetary nebula Abell 63. It is an eclipsing binary, first identified in 1932, although the realization that it is the nucleus of a faint PN did not occur until 1975. The orbital period was found in 1976 to be an astonishingly short 11.2 h, implying that the hot nucleus orbits only about one stellar diameter away from its main-sequence companion.

Known close-binary nuclei

The discovery of UU Sge sparked systematic surveys of PNNi in order to search for binarity. Most of the searches have been done photometrically, i.e. by seeking periodic light variations of the central stars. Four PNNi—UU Sge, V477 Lyr, BE UMa, and the nucleus of SuWt 2—have been found to be eclipsing binaries, with periods of about half a day up to 2.4 days. Eight additional systems do not eclipse, but are still variable in light because of the strong heating effects that arise on one hemisphere of a cool main-sequence star that orbits closely around an extremely hot companion. The orbital periods of these systems range from a remarkable 2.7 h (one of the shortest periods known for a detached binary) to 2.9 days. In addition, the central star of NGC 2346 is a 16 day spectroscopic binary, and there are three known 'Abell 35-type' PNNi whose composite spectra clearly show that two stars—a hot nucleus, and a cool, rapidly rotating companion—must be present.

These known binary PNNi constitute roughly 10% of randomly selected central stars. If we take into account the fact that the main discovery technique for these binaries—photometric variability—is ineffective for binaries with periods of more than several days (since then the heating effects do not occur), we see that the true fraction of binaries among PN central stars could be much higher, if there is a significant population of binaries with longer periods.

Morphologies of nebulae with binary nuclei

It is generally found that PNe known to have been ejected from close binaries are axisymmetric, showing either elliptical or more pronounced butterfly or bipolar shapes. These non-spherical morphologies appear consistent with the CE scenario, if one imagines that the CE interaction will eject matter predominantly into the orbital plane of the binary. Later on, when the hot component of the nucleus develops a fast, spherical, low-density wind (typical of all hot stars), the wind will attempt to plow out into the slower-moving ejected envelope. Along the polar directions, the wind will break out first, producing balloon-shaped bubbles; however, in the orbital plane, where the previously ejected envelope is much denser, the fast wind will be decelerated in a 'snowplow' fashion.

There is space here only for a few images of PNe known to have binary nuclei. Figure 1 shows a ground-based image of Abell 41. It indeed has the general morphology predicted above: the blobs on either side of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

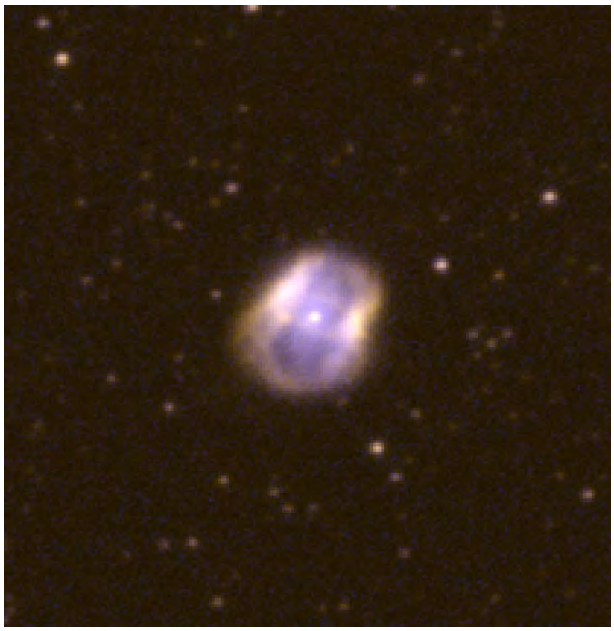


Figure 1. Ground-based CCD image of Abell 41, a planetary nebula whose central star is an extremely close binary with a period of 2.7 h. The image is the sum of [O III] and Hα, obtained with the Cerro Tololo 0.9 m telescope. Note the axisymmetric structure, with blobs marking the orbital plane of the binary, and wind-blown balloons at the poles.

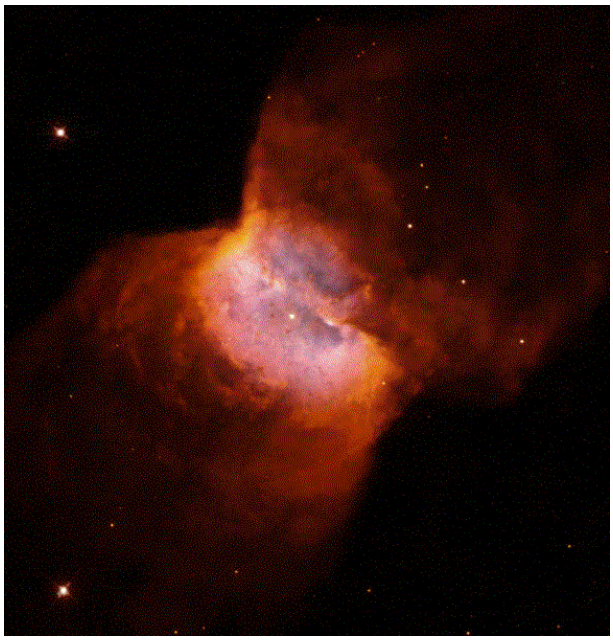


Figure 2. Hubble Space Telescope image of NGC 2346, whose nucleus is a 16 day binary. This is a classical 'butterfly'-shaped planetary nebula, with a dense equatorial ring and wind-blown bubbles perpendicular to the orbital plane. *NASA/STScI Hubble Heritage image.*

the central star mark the orbital plane of the binary, where the gas density of the ejected CE is high and the subsequent fast wind has been unable to escape. Perpendicular to the orbital plane, however, the fast wind has succeeded in inflating two 'bubbles', because the density of the ejected envelope is much lower there.

Figure 2 shows an extreme case of a 'butterfly' PN, NGC 2346, obtained with the HUBBLE SPACE TELESCOPE. Here again the binary PNN is surrounded by a dense ring, ejected as a consequence of the CE interaction; note the pinched waist around the central star. The fast wind has broken out perpendicularly to the orbital plane, producing the spectacular bipolar shape.

In some cases, only the equatorial ring is seen. Figure 3 is an example, showing an almost edge-on ring around the central star of SuWt 2, a 2.45 day eclipsing binary. This 'wedding-ring' PN has an extremely high 'density contrast' between its equatorial and polar regions. In fact, if one is in a poetic mood, one can think of the wedding ring as symbolizing the mystical union of the two stars. . .until one realizes that cataclysmic events and an eventual blow-up loom in the future for this couple.

Are most planetary nebulae ejected from binaries?

As described above, PNe known to contain close-binary nuclei are virtually all axisymmetric, with varying degrees of density contrasts between orbital plane and poles. It is at present an unsolved question whether the opposite is true, i.e. whether *all* non-spherical PNe are ejected from binaries. This is a crucial issue, since the large majority of PNe are in fact observed to be highly non-spherical, leading to the suggestion that most PNe are ejected from binary stars. There may be other mechanisms (e.g. rotation of the red-giant progenitor, magnetic fields, pulsation, convection) that could produce a non-spherical ejection from a single, spherical star. However, on closer examination, the presence of a reservoir of angular momentum generally seems to be required (for example, to spin up the star to sufficient rotational velocity to make the PN ejection non-spherical or to generate a sufficient magnetic field to influence the ejection magnetically), and a companion object is the most obvious such reservoir.

Moreover, it may be that a companion of only modest mass (such as a brown dwarf or even a massive planet) may be sufficient to cause a non-spherical PN ejection, either through the CE interaction or through tidal forces without needing the actual stellar cannibalism.

If we combine the above considerations with the already high fraction of directly detected binary central stars, which were found using an observational technique that is only sensitive to the shorter-period binaries, we are tempted to conclude that binary interactions have a major, and possibly dominant, influence on the ejection and shaping of planetary nebulae.

This conclusion, however, has not yet been tested with direct observations (e.g. a large-scale search for radial-velocity variations on timescales of weeks to months). At present we are left with the words of the poet Robert FROST:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3. Ground-based [O III]+Hα image of SuWt 2, obtained with the Cerro Tololo 0.9 m telescope. This PN appears to be a ring with a very low ratio of height to radius. It is viewed close to edge on, and the central star is an eclipsing binary, showing that the ring was ejected into the orbital plane.

We dance 'round in a ring and suppose,
But the Secret sits in the middle and knows.

Bibliography

A review article on CE interactions is

Iben I and Livio M 1993 Common envelopes in binary star evolution *Publ. Astron. Soc. Pac.* **105** 1373–406

A comprehensive theoretical study of the evolution of binary planetary nuclei is given by

Iben I and Tutukov A V 1993 Formation and evolution of binary planetary nebula nuclei and related objects *Astrophys. J.* **418** 343–67

Theoretical predictions of the orbital periods of binary PNNi have been made by

Yungelson L R, Tutukov A V and Livio M 1993 The formation of binary and single nuclei of planetary nebulae *Astrophys. J.* **418** 794–803

The discovery and early studies of UU Sagittae are recounted by

Bond H E, Liller W and Mannery E J 1978 UU Sagittae: eclipsing nucleus of the planetary nebula Abell 63 *Astrophys. J.* **223** 252–9

Key articles on planetary nebulae ejected from close binaries include

Bond H E and Livio M 1990 Morphologies of planetary nebulae ejected by close-binary nuclei *Astrophys. J.* **355** 568–76

Pollacco D L and Bell S A 1997 Imaging and spectroscopy of ejected common envelopes *Mon. Not. R. Astron. Soc.* **284** 32–44

Soker N 1998 Binary progenitor models for bipolar planetary nebulae *Astrophys. J.* **496** 833–41

Howard E Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary Stars: Detached

Detached binaries are binary star systems where both stars are smaller than their ROCHE LOBES. These systems are useful in determining the basic properties of stars such as mass and radius since in most cases there has been no large-scale mass transfer between the two stars.

Basic stellar data

If a detached system is a double-lined spectroscopic binary with the orbit oriented nearly edge-on so that eclipses occur, then the fundamental properties of the two stars can be determined through careful analysis of the radial velocities and light curve in one or more photometric passbands (see also ECLIPSING BINARY STARS). Although it is possible in principle to measure the mass of a single star via the gravitational REDSHIFT, the effect in non-degenerate stars is so small as to be immeasurable. Thus practically all of our knowledge of stellar masses has come from studies of binary stars, and particularly detached binaries for stars of solar mass and above.

The radius of a single star can be measured directly using interferometric techniques, but to date this has only been possible for a handful of nearby supergiants such as BETELGEUSE. If the distance to a star is known, then the radius (R) can also be determined indirectly with the Boltzmann equation since the luminosity (L) can be estimated:

$$L = 4\pi R^2 \sigma T^4$$

where the effective temperature (T) of the star is determined from the spectrum and σ is the Stefan–Boltzmann constant. This method is limited by its requirement for an accurate distance to the star and by the numerous assumptions and calibrations involved. Therefore the majority of our knowledge about stellar radii has come from the analysis of observations of binary stars. Detached binaries whose components are much smaller than their Roche lobes are good approximations to single stars because there is little physical interaction between the stars and they are not subject to the evolutionary complications arising from mass transfer as in the semidetached systems like ALGOL.

Double-lined spectroscopic binaries that eclipse provide masses and radii directly. The spectroscopic data provide the mass information, although in the absence of eclipses the information is incomplete. Because the orbital inclination and the semi-major axis of the orbit are perfectly correlated, only the minimum masses can be determined. If the system has eclipses, the inclination can be determined and thus the masses can be directly measured.

Data compilations

Popper (1980) compiled the then available data on the masses, luminosities and radii of stars based on observations of mostly detached systems. Popper noted the paucity of data on stars of spectral type B5 and earlier. Hilditch and Bell (1987) published data on 31 such systems with 16 of them being detached binaries. Harmanec (1988) collected observations published since 1980 and also provided approximation formulae for the basic stellar parameters based on the observations.

Bibliography

Harmanec P 1988 Stellar masses and radii based on modern binary data *Bull. Astron. Inst. Czech.* **39** 329–45

Hilditch R W and Bell S A 1987 On OB-type close binary stars *Mon. Not. R. Astron. Soc.* **229** 529–38

Popper D 1980 Stellar masses *Ann. Rev. Astron. Astrophys.* **18** 115–64

Dirk Terrell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary Stars: Overview

Binary stars are pairs of stars in a gravitationally bound, periodic orbit around their common center of mass. Half or more of all stars are parts of binary systems, and this can reach 90% where the census is very complete. In most cases, the two stars of a binary formed together, out of a single gas cloud, and so have the same age and the same initial chemical composition. A few systems arise from gravitational capture processes between previously single stars or via star exchange between a pair and a single star or two pairs, most often in environments like the centers of GLOBULAR CLUSTERS, where stars are very close together.

Binary stars are of interest and importance for several reasons. They provide the only accurate measurements of stellar masses, and the mass of a star turns out to be the primary determinant of its life cycle (see STELLAR EVOLUTION). Second, a variety of astronomical processes and phenomena occur only in binary systems, especially ones where the stars interact at some time in their lives (called close binaries; the opposite is 'wide'). Third, possible orbits for stable, habitable planets around a star are greatly restricted when the star is part of a binary (see THREE-BODY PROBLEM). About one binary in five is gravitationally bound to one or more additional stars in a MULTIPLE STAR system. Generally, the third and additional stars are far enough from a close binary that they do not affect its evolution very much.

Classification

Binary systems can be classified in terms of (a) how they were discovered and studied, (b) their current physical state or (c) their status within an evolutionary scenario. A given system can normally be classified in all three ways.

Classification by method of discovery

The classes by methods of discovery or study are:

(a) Common proper motion (CPM) pairs (where the stars are seen as separate dots of light and move across the sky together, but the orbit period is too long for orbital motion to have been seen).
(b) VISUAL BINARIES (where two dots of light are resolved and seen to orbit each other).
(c) Astrometric binaries (where a single image is seen, but it is elongated and the orientation changes with the orbit period, or a single image moves periodically in the sky).
(d) Spectrum binaries (where the spectrogram of a single image shows two sets of lines characteristic of two stars of different spectral types).
(e) SPECTROSCOPIC BINARIES (SBs, where the spectrogram of a single image shows one or two sets of lines whose radial velocity varies periodically; these are called single-lined and double-lined SBs respectively).
(f) ECLIPSING BINARIES (where one star passes in front of the other from our vantage point, so that the total light we see coming from the system is temporarily dimmed; where the two stars are very close together, gravitational distortions of their shapes or illumination of one by the other can mimic eclipses; these are called ellipsoidal binaries).

Many systems can be studied by more than one of these methods. ALGOL, for instance, is both a spectroscopic and an eclipsing binary with a period of about 3 days, while 70 Ophiuchi is both a visual and a spectroscopic binary, with a period of 88 years.

On average, eclipsing and spectroscopic binaries will have the shortest orbit periods and the smallest orbit sizes (semi-major axes), while visual binaries and CPM pairs have the longest periods and widest separations between the components. The ranges (when both stars are normal main sequence stars) is from several hours up to several million years in period and from twice the radius of the Sun up to about 0.1 PARSEC in separation. Binaries with WHITE DWARF, NEUTRON STAR or BLACK HOLE components can have still shorter periods, down to a few minutes, though the systems will be short-lived. Widely separated pairs, where the stars never interact or influence each other's evolution are also called DOUBLE STARS.

Two objects of masses M_1 and M_2, bound together by gravity can be described by a modification of KEPLER'S THIRD LAW of motion, which says

$$(M_1 + M_2)\,P^2 = a^3$$

where M_1 and M_2 are measured in solar masses ($1M_\odot = 2\times10^{30}$ kg), P is the orbit period in years, and a is the semi-major axis of the relative orbit in ASTRONOMICAL UNITS. Thus, we can find the sum of the masses for any pair where P and a are measurable, and the individual masses as well whenever the ratio can be obtained from the ratio of the stars' distances from their common center of mass, or the ratio of their speeds around it.

Classification by current physical state

Binaries are also classified in terms of the masses and evolutionary states of the two stars and the extent to which each fills its Roche lobe (see figure 1). Detached systems are ones where neither star fills its lobe; semi-detached systems have one lobe-filling star (often a RED GIANT), and material spilling through the point L_1 onto the other star or into a disk orbiting around it. In contact binaries, both stars fill their Roche lobes (also called the inner Lagrangian surface). If material has overflowed the lobes and surrounds both stars, we speak of a common

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

envelope binary. Gas that reaches the point L_2 is free to leave the system completely, often carrying large ANGULAR MOMENTUM with it, so that the stars spiral together. Material can be transferred from one star to the other even in a detached system if one has a strong stellar wind.

These classes are somewhat correlated with those of the previous subsection. A contact system, for instance, will almost certainly be seen to eclipse, while a detached system with two main sequence stars is unlikely to, and a detached system with two white dwarfs almost certainly will not.

Binary systems can include stars of all possible masses from 0.085 to 100 or more solar masses and stars at all evolutionary stages, from PRE-MAIN-SEQUENCE STARS, through the main sequence, red giants and supergiants, and later phases, up to white dwarfs, neutron stars and black holes. Not all combinations are, however, possible. For instance a $30M_\odot$ star will have completed its main sequence lifetime before a $1M_\odot$ star has begun core hydrogen burning, so you will never see a pair consisting of $30 + 1M_\odot$ with both stars on the main sequence. Pairs where both stars are in the same, short-lived evolutionary phase (like binary Cepheid variables) are naturally rather rare.

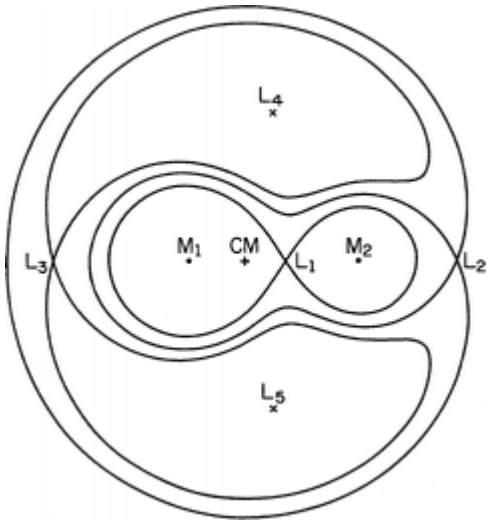


Figure 1. The Roche geometry, showing equipotential surfaces in a coordinate system that rotates with the orbiting star pair. When neither M_1 nor M_2 fills the innermost contour shown (Roche lobe), the system is non-interacting. When one star has evolved and expanded enough to fill its Roche lobe, material flows through L_1 (the first Lagrangian point) onto the other star. When both lobes are filled, the contact system is likely to merge. Common envelope binaries over-fill both lobes, and material that reaches are far out as L_2 can leave the system, typically carrying more then its fair share of angular momentum. (From Webbink R F 1991 *The Astronomy and Astrophysics Encyclopedia* ed S P Maran (New York: van Nostrand) p 81.)

Classification by evolutionary phase.
Assigning binary systems to their place in a scenario for how systems evolve is the most uncertain of the methods, but also the one that provides the most physical understanding. These classes are often named for a prototype, for instance Beta Lyrae stars (in which mass transfer from the more massive, lobe-filling star to its companion has just started and is happening rapidly), Algol variables (in which material is being transferred more slowly from a depleted red giant to its now more-massive companion) and V 471 Tauri stars (detached pairs of white dwarf + red dwarf, expected to evolve into cataclysmic variables).

Quite often, the prototype is not very typical of its class. This is not really surprising. The first example of a class to be found is likely to be unusually bright or otherwise conspicuous. Algol, for instance, is part of a triple system; and V 471 Tauri probably has a bit of mass transfer going on (so it is bright for its type). At least 50 classes of this sort have been named in the astronomical literature, but many would be recognized only by a few specialists.

History

Relatively close star pairs, resolvable with the naked eye like Castor and Pollux, were known to the ancients. Most of these are optical doubles, that is, stars at different distances from us, seen by chance in about the same direction, and not physically bound together. With the advent of the telescope, pairs closer than the naked eye separation limit (about 30") began turning up, starting with Mizar (the middle star in the handle of the Big Dipper) observed by the Italian astronomer Benedetto Castelli in 1617.

William Herschel (see HERSCHEL FAMILY) began systematic cataloguing of double stars in about 1775 and identified about 700 over the next decade. The 1827 catalog of WILHELM STRUVE include 3110 pairs. Most astronomers of the time thought that all these were optical doubles because they believed that all stars had about the same intrinsic brightness and, therefore, that a pair with very different apparent brightnesses must be at different distances. The English polymath, John Michell, had, as early as 1767, suggested on statistical grounds that many of them must be physically bound pairs, because there were too many to be just change alignments.

This hypothesis was confirmed in 1804 by William Herschel, who had been attempting to use the 5" pair Castor A and B to get the PARALLAX of the brighter, and so presumably closer, star. He found, instead, orbital motion and described himself as like Saul, who went out to seek his father's asses and found a kingdom. That the orbits are elliptical, as required by Newtonian gravity, was established for Xi Ursae Majoris by Felix Savary in 1827 (figure 2).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

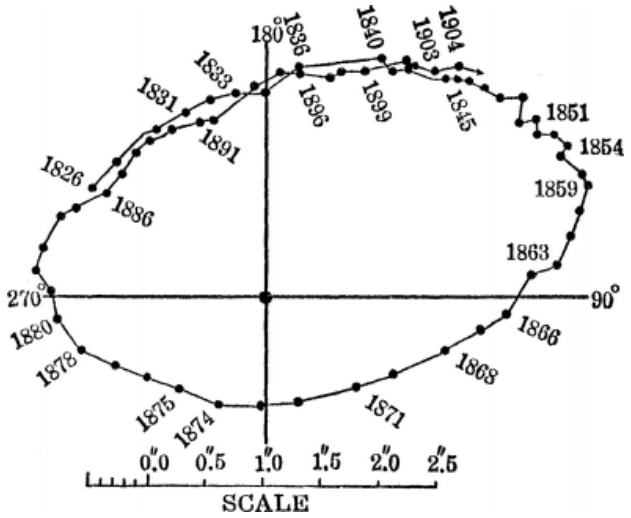


Figure 2. The orbit of Xi Ursae Majoris, the first system for which ellipticity of the orbit was established, by F Savary in 1827. The stars are now back in the same relative position as they had in about 1880. (From Russell H N, Dugan R S and Stewart J Q 1926 *Astronomy* (Boston: Ginn) p 682.)

Herschel's son John continued the cataloguing of double stars, bringing the total to more than 10 000. The modern inventory exceeds 100 000, but only about 1000 of these have had their orbits fully observed, and the number with accurate enough orbital elements for the measured masses to be useful is only a few hundred. It is necessary to know the distance to a visual system as well as the orbit parameters to get meaningful masses.

The first astrometric binaries were Sirius and Procyon, whose sinusoidal motion through the sky was traced out by Friedrich Wilhelm Bessel in the decade leading up to 1844. He described them as having invisible companions of roughly their own mass. They became visual binaries with the discovery of their faint (white dwarf) companions by Alvan G Clark in 1862 and Schaeberle in 1895. Many other astrometric binaries have since been promoted to visual systems, when improved observing techniques have allowed the resolution of smaller separations and fainter stars, often by methods of OPTICAL INTERFEROMETRY.

Algol is the eye of Medusa in the constellation Perseus, and the name means 'the ghoul' or 'the demon'. This suggests that its variable brightness was known to the medieval Arab astronomers who assigned the name. But the first documented report of its occasional fadings came from Montanari in 1669. John Goodricke established in 1783 that the variation is periodic, with the star declining to about one-third of its normal brightness and recovering during about 10 h out of its 68 h period. He suggested the cause might be dark spots on a rotating surface (like SUNSPOTS, but much bigger and darker) or partial eclipse of the star by another body. At the time only eclipses of the Sun, Moon and satellites of Jupiter had been seen, and the idea did not catch on.

The first spectroscopic binary was, again, Mizar (the brighter component of the pair resolved by Castelli). Antonia C Maury of Harvard College Observatory recognized that one of its absorption lines, the K feature of ionized calcium, was sometimes split into two components, with variable separation. This was announced in 1889 by the observatory director, William C Pickering, who correctly interpreted the changing wavelengths of the line as a Doppler shift due to the velocities of two stars of about the same brightness and temperature orbiting each other with a period of 665 days.

H C Vogel, in the same year, reported periodic wavelength shifts in the spectrum of Algol, showing it to be a single-lined spectroscopic binary. The cooler, red giant star that causes the eclipses is faint enough that the techniques of the time could not detect its absorption lines in the combined spectrum. In light of these results for Mizar and Algol, the existence of eclipsing binaries was quickly also recognized by the astronomical community.

Well over 1000 spectroscopic binaries (the majority SB1s, though this is partly a matter of how hard you work) have been studied at some level. Because the Doppler shifts we see are projections of the real orbital velocity along our line of sight, the orientation of the system must be known in order to get the stellar masses. Eclipsing systems are, of necessity, seen nearly edge-on, and the details of their light curves reveal the precise orientation. Thus binaries seen as both spectroscopic (SB2) and eclipsing systems are our main source of masses for systems that are too far away from us to be resolved as visual binaries and to have accurate distance determinations.

Progress in understanding the evolution of binary systems necessarily awaited an adequate theory of stellar evolution for single stars. This was achieved in the decade following 1939, when it became generally recognized that red giants represent an evolved phase, in which fuel for hydrogen fusion has been exhausted at the center of the star. The main sequence lifetime of a star then has to be proportional to its fuel supply (mass) divided by its consumption rate (luminosity), leading to an approximate relation, lifetime = 10^{10} yr $(M/M_{\odot})^{-2.5}$ (because massive stars are much brighter than little ones). From this soon arose the so-called Algol paradox.

Codified by Otto Struve in 1953, the paradox consisted in the fact that we see large numbers of binary systems, mostly eclipsing ones, like Algol, in which the more evolved (red giant) star is the less massive one. The seeds of a solution were already present in a 1941 paper by Gerard P Kuiper, in which he pointed out the possibility for mass transfer between stars inherent in the Roche geometry. John Crawford, in 1955, suggested that such transfer could result in reversing the initial mass ratio of a close binary, so that the original secondary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

could become the more massive star while it was still on the main sequence and receiving contributions from the original primary, evolved to a red giant configuration. More quantitative study appeared in the 1960 PhD dissertation of Donald C Morton, followed by a flowering of relevant work in three groups centered around Miroslav Plaveç (Prague), Rudolph Kippenhahn (Munich) and Bohdan Paczyński (Warsaw).

More recently, mass transfer in close binaries has been used to account for a wide variety of other unexpected combinations of properties, including BLUE STRAGGLERS, barium stars without technetium and gamma-ray bursters.

Binary statistics

Numbers, especially stellar masses, derived from study of individual binary systems help to guide our theories of stellar structure and evolution. In contrast, statistical information on binary populations can provide light on the processes of STAR FORMATION and the evolution of the binary systems themselves.

Gerard Kuiper, the first to publish on statistical issues in 1935, had data on a few dozen visual binaries and CPM pairs and a double handful of spectroscopic, eclipsing systems, with a large gap in orbit period (or semi-major axis) between the two groups. He nevertheless bravely drew a dotted line between the two peaks and suggested that the real distribution in period or separation was fairly flat, meaning roughly equal numbers of systems per logarithmic interval, over the entire possible range, from the sum of the radii of the two stars to about 0.1 pc, beyond which the stars will be pulled apart by the tidal force of the galaxy and of other passing stars. Subsequent work has, in general, validated his monumental extrapolation (figure 3).

Astrometric binaries and optical interferometry have done a great deal to bridge Kuiper's period gap. Not surprisingly, the harder you look, the more companions you find, both very faint ones and ones at awkward separations. The period distribution in modern data is very broad, with a gentle peak at a few years, and wings extending to a few hours on one side and a few million years on the other. Naturally, small separations and periods disappear when the primary becomes a red giant, and large separations and long orbit periods cannot survive in dense clusters of stars.

Kuiper also looked at the distribution of binary mass ratios, $N(M_2/M_1)$, in the hope of deciding how the systems formed. He expected that formation by the fission of a single, rotating cloud would lead to components of nearly equal mass (M_2/M_1 close to 1), while separate formation followed by capture would lead to the average primary having a companion a good deal less massive than itself, since low-mass stars are much commoner than high-mass stars. He focused primarily on single-lined spectroscopic binaries, on the grounds that SB2s and visual systems could be detected only when the stars were of nearly equal luminosity, hence nearly equal mass, and so were biased samples. He found a distribution $N(M_2/M_1)$ that indeed rose toward small values. But his sample was clearly biased in the opposite sense, against systems with mass ratio near unity.

Modern data say that the mass distribution for primaries of close binaries that have not yet experienced any mass transfer is essentially the normal initial mass function (IMF). The situation is less clear for the secondaries, there being still no large truly unbiased samples. Some samples show $N(M_2/M_1)$ rising monotonically to the smallest values to which the data are complete; others show a second peak near $M_2/M_1 = 1$. In any case the processes that might compete to form close binaries now include several additional mechanisms (fragmentation, hierarchical clustering, photodisintegration etc). Thus the functional form of the mass ratio distribution is no longer a clean test between two possible scenarios.

Are the distributions of binary period and mass ratio different in different STELLAR POPULATIONS, and is the total fraction of binaries among the stars also variable? Probably, but our data are very limited. The faint, low-mass systems that make up most of the local sample are difficult to study at large distances and in crowded regions of the sky. It was thought at one time that there

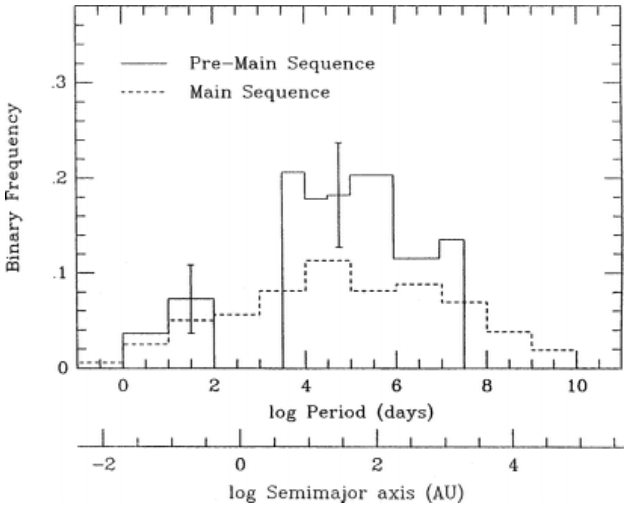


Figure 3. Distributions of semi-major axes and periods for solar-type main-sequence binaries (broken line) and pre-main sequence systems (full line—the regimes of log P = 2–3.5 and greater than 7.5 have not yet been properly surveyed). The distributions are very broad, ranging from less than a day (contact systems) to millions of years (systems at risk of tidal disruption), and peak at periods of several years to decades and separations of 10–100 astronomical units, the regime best explored with speckle interferometry and not available to Kuiper. (From Mathieu R D 1994 *Ann. Rev. Astron. Astrophysics* **32** 469.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

were almost no binaries among the old, POPULATION II stars that make up the halo and globular clusters of our Milky Way. This was due to two kinds of selection effects. First, short-lived systems are long since gone, and, second, longer-lived ones are exceedingly difficult to detect. With modern photometry, one finds that the incidence of, for instance, W Ursae Majoris (low-mass, contact) binaries is about the same in many globular clusters as it is in the solar neighborhood, one for every couple of hundred main sequence stars.

Statistics of evolved systems test our understanding of the processes of mass and angular momentum transfer and loss from the systems and the lifetimes of the survivors. Interesting issues include (a) the cause of the gap in the distribution of periods of cataclysmic variables between the 2 and 3 h, (b) the relative frequency of single and binary WOLF–RAYET STARS, (c) the numbers of binary and millisecond PULSARS relative to the number of low-mass X-RAY BINARIES that are supposed to be their ancestors, and (d) the numbers of x-ray binaries with black hole versus neutron star primaries.

The vast majority of both visual and spectroscopic eclipsing binaries with well-determined orbits contain very ordinary kinds of stars, mostly main sequence and slightly evolved ones with masses between about 0.5 and $5M_{\odot}$. We badly need:

(a) Good masses for very faint (low-mass) stars of both high and low metallicity, to trace out the boundary between true stars and brown dwarfs.

(b) More data on stars of more than about $30M_{\odot}$ to find out the maximum possible mass for a stable star under various conditions and how fast mass is lost early in the life of the star.

(c) Good orbits for systems containing stars in short-lived, but interesting, evolutionary phases, like LUMINOUS BLUE VARIABLES, Wolf–Rayet stars, Cepheid variables, RR LYRAE STARS, blue stragglers and many others.

Even then, serious thought and detailed modeling will be required, because in many cases we cannot be sure that the examples in binaries with periods short enough to be well observed are physically the same as seemingly similar stars seen in other contexts. In some cases, we are sure they are not, because the presence of a companion is likely to enhance mass loss, drive rotationally induced mixing, and make other changes over the life of any stars.

Data on binary populations in other galaxies are quite limited, but are generally consistent with these being similar to the Milky Way population. For instance, x-ray binaries have been found in all the larger galaxies of the LOCAL GROUP, and novae have been identified in galaxies as far away as the VIRGO CLUSTER, both in roughly the expected numbers. Good orbits for extragalactic binaries will require both the sensitivity of very large telescopes and the angular resolution possible only with adaptive optics or interferometry or from space.

Inventories

Many phenomena of stellar astronomy can be categorized by whether they occur only in interacting binary systems, preferentially in binaries, or reluctantly in binaries, and why. Most of the star types categorized this way below can be found in these volumes:

(a) Type I: Found only in binaries (roughly in evolutionary sequence).
- stars with well-determined masses;
- W Ursae Majoris variables;
- Beta Lyrae, W Serpentis and Algol variables;
- Zeta Aurigae stars;
- RS Canum Venaticorum stars;
- cataclysmic variables (including V 471 Tauri stars, novae, dwarf novae of several types, recurrent novae, symbiotic stars, nova-like variables and the magnetic DQ Herculis and AM Herculis variables);
- AM Canum Venaticorum stars;
- accretion-induced collapse (of white dwarfs to neutron stars or of neutron stars to black holes);
- x-ray sources (bright, compact, and not associated with rotation-powered pulsars; accretors can be neutron stars or black holes, donors of high or low mass).

(b) Type II: Types for which some or all of the promising models invoke binary systems, but where duplicity cannot always be demonstrated observationally.
- blue stragglers;
- barium stars (type stars with no technetium);
- dwarf carbon stars;
- anomalous (dwarf) Cepheid stars and SX Phoenices stars;
- millisecond pulsars with weak magnetic fields (some are definitely binaries with white dwarf or neutron star companions);
- type Ia supernovae;
- Thorne–Zytkow objects (red giants with neutron star cores; none actually known);
- type Ib and Ic supernovae (apparently from massive progenitors, but with their hydrogen-rich envelopes entirely stripped off);
- gamma-ray bursters.

(c) Type III: Favored in close binaries through effects of mass and angular momentum loss.
- Wolf–Rayet stars;
- Of and OBN stars;
- R Coronae Borealis and FG Sagittae stars (most are not known to have companions).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(d) Type IV: Inhibited in close binaries by gravitational effects.
    habitable planets;
    OBC stars;
    extended envelopes of massive stars needed for standard type II supernovae.
(e) Type V: Favored or inhibited in close binaries through the effects of synchronization of rotation to orbit period.
    BY Draconis and UV Ceti type flare stars;
    dynamo magnetic fields, stellar chromospheres and coronae, x-ray and radio emission star spots, and other indicators of stellar activity;
    Be stars;
    Delta Scuti stars;
    OB star winds;
    surface chemical peculiarities (due to diffusion) of types Am, Ap, Bp etc;
    surface chemical peculiarities (due to mixing and meridional circulation).

Update (24 December 2001)

Formation and statistical issues

Fragmentation continues to be regarded as the most important process in the formation of close binaries. An important observational discovery is that, among the very youngest stellar populations, triples are just about as common as binaries. The fate of a large fraction of these is to lose the smallest component (which then becomes an isolated BROWN DWARF or EXOPLANET), leaving the other two stars as a somewhat closer pair than they were before.

The number of binary and millisecond PULSARS, supposedly recycled in X-RAY BINARIES, continues to be much larger than the suggested progenitors can account for. Accretion-induced collapse, in which a white dwarf component is driven above the Chandrasekhar limit and so collapses to a NEUTRON STAR with rapid rotation and weak magnetic field without ever going through an X-ray binary phase, appears to be the most likely solution to this statistical mismatch.

Evolutionary issues

Early calculations of the evolution of close binaries were normally carried out under the assumption that the total mass and angular momentum of the system remained constant throughout. It has been clear for some time that this could not be true during late evolutionary phases (when single stars shed both gas and rotation in any case). More careful comparisons of observed properties of stars in binaries with such calculations show that the assumption of constant mass and angular momentum ('conservative' evolution) has failed even by the Algol phase.

The composition of the surface of the main sequence component of the black hole X-ray binary GRO 1655-40 shows evidence of accretion of material processed in nuclear reactions in the evolution and explosion of the progenitor star of the black hole. This shows that core-collapse type SUPERNOVAE can leave BLACK HOLES, as well as neutron stars, as their remnants.

Inventory issues

At least two kinds of binary star systems are now suspected of being the immediate progenitors of Type Ia supernovae. First are pairs consisting of a WHITE DWARF plus a hot subdwarf, with total mass exceeding the Chadrasekhar limit and with orbital periods short enough that the stars can spiral together in less than the age of the universe. One of these has now been found. The other is the class of recurrent novae, the subset of CATACLYSMIC VARIABLES whose white dwarf components have masses just a little smaller than the Chandrasekhar limit. It has been shown that the explosions in these, which we see as novae, do not expel as much material as was accreted to cause the explosion. Thus the mass of the white dwarf increases until the star begins to collapse and then burns most of its mass of carbon and oxygen violently to make the supernova.

The progenitors of the gamma ray bursters similarly now seem to include two sorts of object. One is binary pairs of neutron star + neutron star or neutron star + black hole. These probably make the bursts of gamma rays that last at most a few seconds. The other progenitor class is massive, very rapidly rotating (and perhaps binary) stars, whose cores collapse to rapidly rotating black holes with strong magnetic fields attached to their horizons or disks around them. These probably make the bursts that last many seconds to minutes, which are the sort for which X-ray, optical, and radio counterparts have now been found.

Bibliography

Appenzeller I (ed) 1997 Report of Commission 42, close binaries *IAU Trans.* **23A**

(and the corresponding reports in earlier and later volumes of *IAU Trans.* A)

Batten H 1973 *Binary and Multiple Systems of Stars* (Oxford: Pergamon)

Eggleton P P and Pringle J E (eds) 1983 *Interacting Binaries* (Dordrecht: Kluwer)

Kopal Z 1959 *Close Binary Systems* (New York: Wiley)

Paczynski B 1971 Evolution in close binary systems *Ann. Rev. Astron. Astrophys.* **10** 183

Sahade J, McCluskey G and Kondo Y (eds) 1993 *The Realm of Interacting Binaries* (Dordrecht: Kluwer)

Trimble V 1983 A field guide to the close binaries *Nature* **303** 137

Trimble V and Aschenbach N 2000 Astrophysics in 1999 *Pub. Astron. Soc. Pacific* **112** 434–503

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trimble V and Aschenbach N 2001 Astrophysics in 2000 *Pub. Astron. Soc. Pacific* **113** 1025–1114
Trimble V and Aschenbach N 2000 Astrophysics in 2002 *Pub. Astron. Soc. Pacific* **114** in press

Virginia Trimble

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary System

As the name implies, a binary system normally consists of two stars, which are close together and therefore bound by a common gravitational field. Many binary systems have three or more components, however, and are often termed *multiple stars*. In many such cases there are two main components, additional members of the system being much smaller and often at considerable distance from the main pair. The majority of *double stars* are believed to be binaries, but observation over a period of years is needed to confirm their binary nature. A system whose orbital movement can be observed telescopically is termed a *visual binary*.

The stars comprising a binary system are each in orbital motion around their common centre of mass. Their orbits are similar in form but different in size, that of the less massive component being the larger. It is convenient for observational purposes, however, to study the apparent motion of the fainter star around the brighter; this is termed the 'relative orbit'. It is similar in shape to the actual orbits of both stars but larger than either, its major axis being equal to the sum of the major axes of the component orbits.

The orbital periods of visual binaries are very long (typically between 20 and 700 years), so determination of their relative orbits is a slow process requiring careful observation of the *separation* and *position angle* of the pair, repeated over many decades. A few binaries of very short period are known, for example the stars BD 8° 4352 (1.7 years) and δ Equulei (5.7 years), and some of exceptionally long period nevertheless have well-determined orbital elements, for example ε Sculptoris (period 1192 years) and 38 Geminorum (3190 years).

Close binary pairs, whose separation may be as little as one diameter of the primary star, cannot be observed by visual methods but their nature can be detected by other means. *Spectroscopic binaries* reveal their nature by periodic fluctuations in the Doppler shift of lines in their spectra; from these the varying *radial velocity* of the system can be determined and, over a period of time, the relative masses and orbit deduced. The first spectroscopic binary to be identified was the brighter component of the visual binary *Mizar*, which was shown from studies of line-doubling in its spectrum to be itself a binary of 20.5 days period, by Anna Maury (1866–1952) at the Harvard College Observatory in 1889.

Eclipsing binaries are close binary systems for which the Earth happens to lie in or close to the orbital plane of the system. The overall brightness of the system varies in a very regular manner, due to the components periodically eclipsing each other. The form of the *light curve* repeats very accurately, but its exact shape is specific to the particular star system; the relative size, mass and brightness, and the separation of the components, can be deduced from it. A well-known example is *Algol*, which has a period of 2.87 days.

Astrometric binaries are very close binary systems, usually comprising a fairly bright star with a much fainter companion. The existence of the latter is revealed only by a slight 'wave' in the *proper motion* of the bright star, determined from long-term astrometric measurements. The first to be discovered and best-known example is *Sirius* (apparent magnitude −1.44), whose eighth magnitude companion was discovered by Alvan Graham Clark (1832–1897) in 1862. It has a period of 50.09 years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binocular Astronomy

It is often supposed that anyone who means to take a serious interest in astronomy must acquire a large and expensive telescope. Nothing could be further from the truth. A great deal can be done with the naked eye alone, as is described in NAKED-EYE ASTRONOMY. It is true that a telescope costs a considerable amount of money, and the small, cheap telescopes on sale in camera shops are most unsatisfactory; the most that can be said for them is that they are slightly better than nothing at all. Luckily there is an alternative: obtain binoculars.

A pair of binoculars is essentially nothing more than two small refracting telescopes joined together. The main advantage is that the observer can use both eyes instead of only one. The FIELD OF VIEW will be much wider than that of a small astronomical telescope, and binoculars are handier, as they can easily be carried around. Moreover they give an erect image, which means they can equally well be used for activities such as bird-watching or looking at ships out to sea. The main disadvantage is that they have comparatively low magnification. Zoom binoculars can be obtained, but in general they are not satisfactory for astronomical observing, and in any case the maximum magnification is not likely to be much more than $\times 20$.

Binoculars are classified according to their magnifications and the diameters of their objective lenses (always given in millimeters). Thus 7×50 indicates a magnification of 7 with each objective lens 50 mm across; 12×50 gives a magnification of 12, again with 50 mm objective lens, and so on. At the outset, it is very important to make a careful decision about just which type is to be bought.

As with telescopes, the larger the aperture the greater the light-grasp, but there are hazards here. With increased power, the field of view becomes smaller, and the binoculars become heavier. If only one pair is to be obtained, then something around 7×50 is a good choice. The magnification is adequate, the field is pleasingly large, and the binoculars are light enough to be hand-held without any awkward shake. Once the magnification exceeds around $\times 12$, hand-holding is a major problem, and the only solution is to use some form of mounting.

The following notes may be useful as a general guide:

3×20: These binoculars are very light and can be slipped into one's pocket. They have their uses, but the low magnification means that they are decidedly limited, and unless they can be bought very cheaply, they are probably not to be recommended.

7×50: Binoculars of this kind are very common. They will show considerable detail on the MOON, for example, and almost endless superb star fields. For looking at some types of object, such as loose STAR CLUSTERS of the Pleiades type, they are ideal because of their large field of view.

8.5×50: Also very suitable. The light grasp is adequate, and the binoculars can be comfortably hand-held. The increased power means that more detail becomes available, though the field of view is noticeably reduced.

11×80: These too are common, though they are not cheap. The field of view is reasonably large, and the light grasp is excellent, though they are rather too heavy to be held steadily.

12×50: We are now reaching the limit of comfortable hand-holding. One solution is to sit down and jam one's elbows into one's body. This works fairly well, but it is not ideal.

With a magnification of over 12, a mounting is really essential. The obvious solution is a tripod. Most photographic shops sell binocular mounts which can be screwed into the top of an ordinary camera tripod, and this works well, though it is often necessary to perch the tripod on top of a platform or a solid table; otherwise you will have to be something of a contortionist when trying to observe an object high in the sky (incidentally, the same is true of small REFRACTING TELESCOPES, unless a right-angle prism is used).

Making a mount for a pair of binoculars is fairly simple for anyone who is reasonably good at carpentry, but the main essential, steadiness, must always be borne in mind. What is needed is the faculty of moving the binoculars freely in any direction, in ALTITUDE and in AZIMUTH. Another solution is to buy or make a 'neck mounting', in the form of a harness which can be attached to the observer's chest. This is effective, and if well made is not nearly so uncomfortable as might be thought.

It is of course possible to be more ambitious, and use a really powerful pair of binoculars. Some observers of VARIABLE STARS and COMET hunters often have binoculars with objective lenses up to six inches across. The cost is very high indeed, and the mounting is all important, so that binoculars with objective lenses of this kind will certainly not be found except in the hands of very serious researchers.

In short: if only one pair of binoculars is to be obtained, it is wise to choose a magnification of between 7 and 10, with an aperture of around 40–60 mm.

Selection is the first essential. One advantage of binoculars is that they can be tested on the spot. Check to make sure that both of the telescopes can be brought into sharp focus to produce a clear-cut image; if this cannot be done, reject the binoculars immediately. Check also for any trace of false color; this will indicate poor lens quality, and again the binoculars should be rejected. If possible, test upon a bright star, and ensure that the star can be reduced to a colorless pin-point of light (assuming, of course, that the star is a white one!). One more point is worth making here, though it ought to be obvious enough. As soon as you take the binoculars, loop the safety cord around your neck. If you fail to do so, then it will only be a matter of time before the binoculars are dropped, and a rapid descent to the ground will be no good to them at all.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Next, a word of warning: never use binoculars to look directly at the Sun, even with a dark filter over the eyepiece. If you do so, then you will focus the light and heat on your eyes, with tragic results. It is true that Mylar filters over the objective lenses can be used, and it is also possible to project the Sun's image onto a screen in the same way as is done for an ordinary telescope, but the results are never good, and in principle it is fair to say that the Sun should be avoided completely with binoculars (see also SUN: SAFE OBSERVATION).

The MOON is a different matter, because it sends us virtually no heat. With a low power and a full moon you may dazzle yourself, but that is all. In any case, full moon is the very worst time to observe, because there are virtually no shadows on the lunar surface—the sunlight is coming straight down, so to speak, and little can be seen apart from the broad, dark plains still known as seas. The best views are obtained when the Moon is in the crescent, half or GIBBOUS (three-quarter) stage. Look along the TERMINATOR—that is to say the boundary between the sunlit and the night hemispheres—and the craters and mountains will be beautifully displayed. The view alters from night to night, or even from hour to hour, as the angle of illumination changes. A crater which is prominent when on the terminator, with its floor wholly or partially filled with shadow, may be hard to identify when the Sun is high above it. Near full, the scene is dominated by the bright streaks or rays which are associated with a few of the craters, notably TYCHO (CRATER) in the Moon's southern uplands. It does not take long to learn one's way around the lunar surface, and binoculars are ideal for this purpose.

The situation is not so good from the point of view of the planetary enthusiast because of the relatively low magnification. MARS, for example, will appear as nothing more than a small, red disk. VENUS, which is closer to the Sun than we are, shows phases similar to those of the Moon, and during the crescent stage, either in the west after sunset, or in the east before dawn, the changing phase is easy to follow even with a magnification of no more than 7. Venus is so bright that it can often be seen even when the Sun is well above the horizon; remember, however, not to sweep around for it—you might look at the Sun by mistake. Similar restrictions apply when searching for the elusive planet MERCURY. Binoculars often show it, either soon after sunset or soon before sunrise. Again, never sweep for it with binoculars unless the Sun is completely below the horizon.

JUPITER is more interesting, because binoculars will show all four of the large satellites: IO, EUROPA, GANYMEDE and CALLISTO. They look like small stars close to Jupiter, and their movements from night to night are easy to follow. Saturn's rings (see SATURN: RINGS) are below binocular range, though with a magnification of 10 or more it is possible to make out that there is something unusual about the planet's shape. The two other planets, URANUS and NEPTUNE are available to binoculars, but they look like stars, and you will need a star map in order to locate them.

Then of course, there are comets. I have already referred to those serious comet hunters who scout the skies regularly and have made many discoveries. A really sound knowledge of star patterns is essential for this (real experts can memorize the positions and magnitudes of many thousands of stars so that a newcomer—either a comet or a nova—can be recognized at once). A bright comet is an ideal binocular object, and in recent years we have seen two: COMET HYAKUTAKE in 1996 and COMET HALE-BOPP in 1997. These were both superb, with long tails and gleaming heads. We cannot tell when the next brilliant comet will appear, but there are plenty of faint ones, and the binocular owner armed with a good star map will be able to identify them.

It is wrong to say that all stars look white. Some of them are colorless, but others are blue, yellow orange or red. Only really brilliant stars show obvious color as seen with the naked eye and these are mainly orange or red; for instance nobody can overlook the lovely warm glow of BETELGEUSE in ORION which contrasts so sharply with the pure white of Bigel in the Hunter's foot. With fainter stars the colors do not show up, and it is here that binoculars come into their own. Consider μCephei, in the far north of the sky, not too far from the Pole Star. It is so dim that it is very difficult to see without optical aid; turn binoculars toward it, however, and you will see why it has been nick-named the 'Garnet star'—it has even been likened to a glowing coal. This is only one of the many cases of stars which reveal their vivid hues when observed with binoculars.

Binary stars are also on view (see BINARY STARS—OVERVIEW). One of the best examples is ALBIREO, in Cygnus (the Swan). With the naked eye it looks commonplace enough, but binoculars show a golden yellow primary together with a vivid blue companion. Many binary stars are within binocular range, and there are also variable stars, many of which are orange or red. These are stars which fluctuate in brightness over short periods; binoculars make it possible to estimate their magnitudes, and to make really useful observations.

Star clusters are glorious; look for example in the Pleiades (or Seven Sisters) in Taurus, 'The Beehive' in the Crab and (from the southern part of the world) the Jewel Box cluster in the Southern Cross. Binoculars show that the Milky Way (see MILKY WAY GALAXY), which stretches across the sky from horizon to horizon, is made up of myriads of faint stars which appear so close together that they might be thought to be in danger of collision—though in fact we are dealing with nothing more than a line of sight effect, and stellar collisions very seldom occur.

GASEOUS NEBULAE, in which stars are being formed, are also to be seen; look at the Great Nebulae in Orion, close to the three stars of the Hunter's belt. Powerful binoculars even show the four stars of the 'Trapezium', which are silhouetted against the nebulosity and are responsible for making it shine. Then there are a few galaxies, vast external star systems so remote that light takes millions of years to reach us. The Great Spiral in ANDROMEDA, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

nearest of the really large galaxies, is an easy binocular object, though it is true that photographs taken with adequate telescopes are needed to bring out its spiral form.

Binoculars are comparatively cheap, they are portable and they are easy to use. They can provide endless enjoyment, and can be also be used for serious scientific research. There can be absolutely no doubt that a pair of binoculars is much more valuable astronomically than a very small telescope.

Bibliography

Robinson L 1990 '*Outdoor Optics*' (Lyons and Burford)

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Biot, Jean-Baptiste (1774–1862)

Physicist and astronomer, born in Paris, France, professor at the Collège de France. His most important work was on the polarization of light (which he correctly suggested was due to asymmetric molecules). With Joseph Louis Gay-Lussac he made in 1804 an ascent in a balloon to study a variety of phenomena at high altitude. Biot headed a commission from the Academy of Sciences to study the meteor shower at L'Aigle in 1803, from which two or three thousand pieces were recovered. In a turning point, the Academy, for many years stubbornly ridiculing the idea that meteorites fell from space, accepted Biot's report that the L'Aigle shower was extraterrestrial.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Birkeland, Kristian (1868–1917)

Birkeland was a Norwegian physicist, born in Oslo. In 1900, he identified and then simulated the charged electron–magnetic flux tube connection between the Sun and Earth that produces the aurora. He studied the zodiacal light during expeditions to the Sudan and Egypt. Birkeland committed suicide in a depression associated with the rejection of his auroral theories by his contemporary established scientists, but was later fully recognized as a founder of auroral studies (his portrait is on a Norwegian banknote). *Birkeland currents* are electric currents linking the Earth's ionosphere with more distant regions, flowing along magnetic field lines.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

BiSON Network

The telescopes of the Birmingham Solar Oscillation Network (BiSON) are located in East and West Australia, South Africa, Chile, California and Tenerife. They permit continuous monitoring of the radial oscillations of the Sun using very narrow absorption lines of potassium in the solar atmosphere. Absorption cells of potassium vapor are used in the optical path of the telescopes to measure the precise wavelength of solar features and from this the Doppler motion of the Sun's surface can be detected to a precision of 1 cm s^{-1}. These velocities are analysed to detect oscillations in the Sun which give quantative clues to the temperature and density of the Sun's interior.

For further information see
http://bison.ph.bham.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

BL Lacertae

BL Lacertae (BL Lac or S4 2200+420) was known to be variable in the optical from as early as 1929, and because of its stellar appearance it was originally thought to be a VARIABLE STAR. It is located in the constellation Lacerta, the lizard (RA (J2000) = 22h 02′ 43.29″, DEC (J2000) = +42° 16′ 39.98″). Observations in the late 1960s showed it to have highly variable radio emission as well.

BL Lac was found to be an extragalactic object with the measurement of its redshift ($z = 0.069$, based on the detection of very weak emission lines) in the early 1970s. During the same period it was also understood to be an unusual type of ACTIVE GALACTIC NUCLEUS (AGN). Thus, BL Lacertae is the prototype of the class of AGN known as BL LACERTAE OBJECTS (also called BL Lacs). Collectively with some quasars such as 3C273 and 3C279, they are known as 'BLAZARS'. These objects are characterized as having bright nuclei with strongly polarized optical emission and large variability at all wavelengths. Their nonthermal radio–gamma-ray continua are thought to be emitted by a relativistic jet oriented close to the line of sight. BL Lac objects make up a small subset of AGNs, with about 350 known at present.

BL Lac is an important member of the class because it is relatively nearby, such that details can be well studied. The galaxy surrounding the active nucleus, or host galaxy, is a giant ELLIPTICAL GALAXY, typical of the host galaxies around other BL Lac objects.

Radio observations show BL Lac to exhibit SUPERLUMINAL MOTION, indicative of material being ejected at relativistic velocities from the nucleus. Recently, observations by the Compton Gamma-Ray Observatory showed BL Lac to be a strong gamma-ray source, together with many other blazars. The rapid variability timescales and high luminosities observed at these energies indicate that the gamma-rays are produced in a very compact region of the jet.

The nuclear emission in BL Lac is nonthermal and produces a continuum spectrum consisting of two components. The radio-to-ultraviolet spectrum is produced by synchrotron emission with peak power in the infrared–optical region while the x-ray-to-gamma-ray spectrum is produced by inverse Compton emission. (However, there are other possible explanations for this emission which are a subject of current debate.) These two components are present in all BL Lac objects.

BL Lac objects, as a class, are characterized by optical spectra which are featureless or which have extremely weak lines (less than 5 Å equivalent width). This property makes the determination of their distances difficult. BL Lac itself is no exception. Recently, however, strong, broad and variable emission lines (e.g. H-alpha, equivalent width = 7.3 Å, luminosity $\sim 2 \times 10^{41}$ erg cm^{-2} s^{-1}) appeared in the spectrum of BL Lac. Current thinking is that the lines appear when the variable continuum emission is low. Furthermore, the presence of emission lines indicates the presence of a radiation field external to the jet, which may play an important role in the jet energetics (providing seed photons for the inverse Compton emission responsible for the gamma-rays).

Bibliography

Catanese M *et al* 1997 Detection of gamma rays with $E > 100$ MeV from BL Lacertae *Astrophys. J.* **480** 562–7

Pesce J E, Falomo R and Treves A 1995 Environmental properties of BL Lac objects *Astron. J.* **110** 1554–63

Sambruna R M, Ghisellini G, Hooper E, Kollgaard R I, Pesce J E and Urry C M 1999 ASCA and contemporaneous ground-based observations of the BL Lacertae objects 1749+096 and 2200+420 (BL Lac) *Astrophys. J.* **515** 140–52

Urry C M and Padovani P 1995 Unified schemes for radio-loud active galactic nuclei *Publ. Astron. Soc. Pacific* **107** 803–45

Vermeulen R C, Ogle P M, Tran H D, Browne I W A, Cohen M H, Readhead A C S, Taylor G B and Goodrich R W 1995 When is BL Lac not a BL Lac? *Astrophys. J. Lett.* **452** L5–8

Joseph E Pesce

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

BL Lacertae Objects

BL Lacertae objects are an enigmatic group of ACTIVE GALACTIC NUCLEI (AGN) characterized by compact, flat-spectrum radio sources, nearly featureless and polarized optical continua, very luminous high-energy radiation (e.g. some BL Lacs are the only celestial detections at TeV energies; Macomb *et al* 1995) and rapid variability in all observed wavebands. In recent years, BL Lacs have been grouped with some quasars into the 'BLAZAR' class, which includes BL Lacs, flat-spectrum radio quasars, highly polarized quasars and optically, violently variable quasars. It is the absence of high equivalent width (typically $W_\lambda \leq$ 5 Å) emission lines which separates BL Lacs from the other blazar subclasses. A significant dilution of stellar absorption lines by a non-thermal continuum (typically a reduction of the Ca II 4000 Å 'break' by a factor of $\geq$1.5) separates BL Lacs from other low-power radio galaxies.

The absence of high-contrast emission and/or absorption lines also means that some BL Lacs have unknown REDSHIFTS, causing large uncertainties in their intrinsic properties. In fact, complete redshifts are not yet available for any flux-limited sample of these objects. Adding to the enigma of BL Lacs is that the class is often defined using operational 'negatives'; e.g., a BL Lac is an AGN which lacks strong emission or absorption lines. Defining membership in the BL Lac class using the absence of spectral lines can make membership dependent on the quality of the observational material in hand so that some quasars have been misidentified as BL Lacs on the basis of either poor-quality optical spectra or inadequate wavelength coverage (Antonucci and Ulvestad 1985). Also, relative variability between continuum and line-emitting regions can mean that an object might meet the 'featureless' criterion required for BL Lacs at one time but not at others (e.g. PKS 0521-365 and BL Lac itself, both of which have had high equivalent-width Hα emission appear in their spectra on some occasions; Stickel and Kuhr 1993; Vermeulen *et al* 1995). This leads to additional uncertainty in defining samples of BL Lacs for statistical studies.

This class of AGN is named for the first known example, BL LACERTAE, which, as the naming convention indicates, was discovered originally as a variable star. BL Lac was recognized to be a luminous radio source by Schmitt (1968), shown to have faint 'fuzz' surrounding its point source consistent with an ELLIPTICAL GALAXY at low redshift by Adams (1974) and proven to be extragalactic from the redshift of its 'fuzz' by Oke and Gunn (1974). Although one other 'variable star' was shown to be a BL Lac object (AP Libra), the early years of BL Lac research were dominated by studies of objects discovered as compact, flat-spectrum and time-variable radio sources.

Radio-loud BL Lacs (RBLs)

Approximately 75 BL Lacs have been discovered owing to their radio emission, including 37 in the 1 Jy sample of the radio-brightest BL Lacs in the sky (Stickel *et al* 1991). New large-angle sky surveys conducted with the VERY LARGE ARRAY are likely to contain several hundred more BL Lacs. No radio-quiet BL Lac Object has yet been discovered (Stocke *et al* 1990). Very-long-baseline interferometry of radio-selected BL Lacs (RBLs) has discovered that a large fraction (11 of 12 BL Lacs investigated by 1994; Gabuzda *et al* 1994) exhibit apparent SUPERLUMINAL MOTION, although at systematically lower velocities than for superluminal quasars (Cohen 1989). That superluminal motion should be observed in at least a few BL Lac objects is a natural prediction of the Blandford and Rees (1978) 'beaming model' for BL Lacs, in which the highly variable continuum emission from a relativistically moving 'jet' of particles ejected from an AGN (Begelman *et al* 1984) is 'beamed' in the direction of jet motion owing to the relativistic aberration of starlight effect. Lorentz factors γ as high as 40 have been inferred from this superluminal motion (see figure 1), although more typical values are $\gamma \sim$ 5–10 (Gabuzda *et al* 1994).

The combination of aberration and Doppler effect due to relativistically moving emission regions can brighten the observed jet luminosity $L_{\rm obs} = \delta^p L_0$ enormously, where

$$\delta = [\gamma(1 - \beta\cos\theta)]^{-1}. \qquad (1)$$

δ is termed the 'Doppler factor', θ is the 'viewing angle', the angle between the line of sight and the direction of emitting-particle motion, L_0 is the unbeamed, intrinsic jet luminosity, $\beta = v/c$ and $\gamma = (1 - \beta^2)^{-1/2}$. The exponent p is, in most cases, >1 but its exact value depends on the physical structure of the emitting region (see appendix B of Urry and Padovani, 1995 for details); typically $p = 2 + \alpha$ or $3 + \alpha$ in the cases of a continuous jet or discrete emitting blobs respectively, whose rest frame emission is isotropic with a power-law spectrum of the form $f(\nu) \sim \nu^{-\alpha}$. The Doppler factor is a strong function of viewing angle, especially for $\gamma > 5$ (see figure 2), so that, for typical values of the exponent p, relativistic beaming can boost the source output by a factor of several hundred. Doppler factors for BL Lacs have been inferred both from the observed luminosity functions and from details of the radio structure of BL Lacs compared with the same quantities for Fanaroff and Riley (1974) type 1 (FR-1) radio galaxies, the suspected 'parent population' of BL Lacs (see 'Parent populations' below). Typical values obtained are $\gamma \sim$ 5–10, although much larger values cannot be conclusively ruled out in some cases.

Thus, an AGN viewed from a direction nearly along the extension of its jet will have a spectral energy distribution (SED) dominated by this 'beamed' continuum radiation. There is overwhelming evidence that the emission mechanism in jets is incoherent synchrotron emission at lower energies and inverse-Compton emission at higher energies (the exact demarcation between emission mechanisms varies from object to object but is typically 0.1–10 keV; see Urry and Padovani 1995; Ulrich *et al* 1997; figure 5). Also, because this emission emerges from a very tiny region whose viewing angle can

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The properties of subclasses of BL Lacertae objects.

I. X-ray loud BL Lacs (XBLs)

- Low redshifts (≤ 0.6) and luminosities
- FR-1 (Fanaroff & Riley type 1) radio morphologies and power levels
- Weak optical absorption lines
- Weakly variable optically ($\delta m \leq 1$ mag)
- Weakly optically polarized ($\leq 5\%$)
- Many have 'preferred' position angles of polarization
- Relatively weak high energy (inverse-Compton) emission
- $\langle V/V_{\rm max}\rangle < 0.5$
- Most XBLs are HBLs (see below)

II. Radio loud BL Lacs (RBLs)

- Mixture of high and low-redshift and luminosity objects
- Mixture of FR-1 AND FR-2 ('classical doubles') radio morphologies and power levels
- Many possess weak optical emission lines
- Some highly variable optically ($\delta m \geq 1$ mag)
- Some possess high ($\geq 10\%$) and highly variable optical polarization
- Relatively strong inverse-Compton emission
- $\langle V/V_{\rm max}\rangle > 0.5$
- Most RBLs are LBLs (see below)

III. Low-energy peaked BL Lacs (LBLs)

- Spectral energy distribution peaks in the IR

IV. High-energy peaked BL Lacs (HBLs)

- Spectral energy distribution peaks in the UV/x-ray

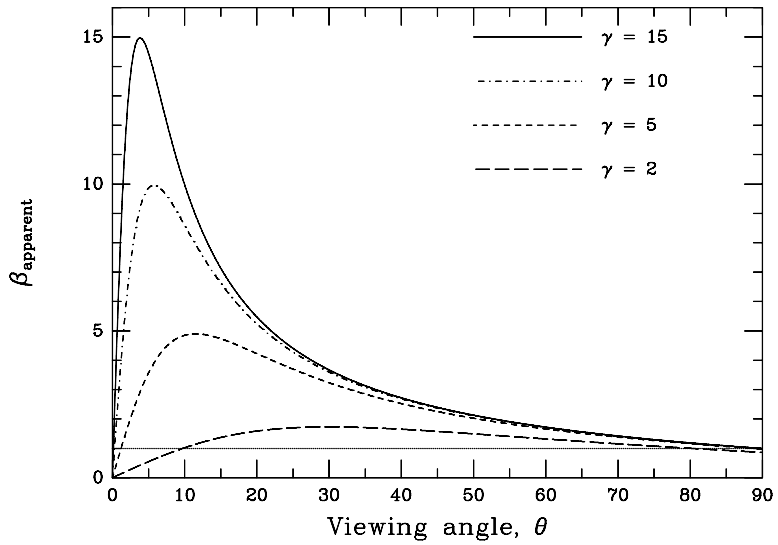


Figure 1. The apparent velocity relative to the speed of light $\beta_{\rm app}$ versus angle to the line of sight θ for an object approaching at relativistic speed. Different curves correspond to different Lorentz factors γ. Note that $\beta_{\rm app}$ is essentially independent of γ at large angles. From Urry and Padovani (1995), copyright Astronomical Society of the Pacific, reproduced with permission from the authors.

change rapidly, the BL Lac SED can be highly variable on short timescales ($\ll$1 day; Doxsey *et al* 1983; Moore *et al* 1985), highly polarized owing to a small-scale ordering of magnetic fields (see compilation in Angel and Stockman 1980) as well as overwhelm any isotropic radiation from the AGN nuclear or near-nuclear region (e.g. broad or narrow emission line clouds in the optical–UV, accretion disk emission in the UV or high-energy coronal emission in the hard x-ray). Thus, in the 'beaming model', the BL Lac SED is dominated in all observed wavebands by beamed jet emission.

The very rapid variability of BL Lac (and other blazar) radio emission sets a maximum size based on light-travel arguments (i.e. a size smaller than the distance traversed by light during the variability timescale) and thus a minimum photon density for the synchrotron-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

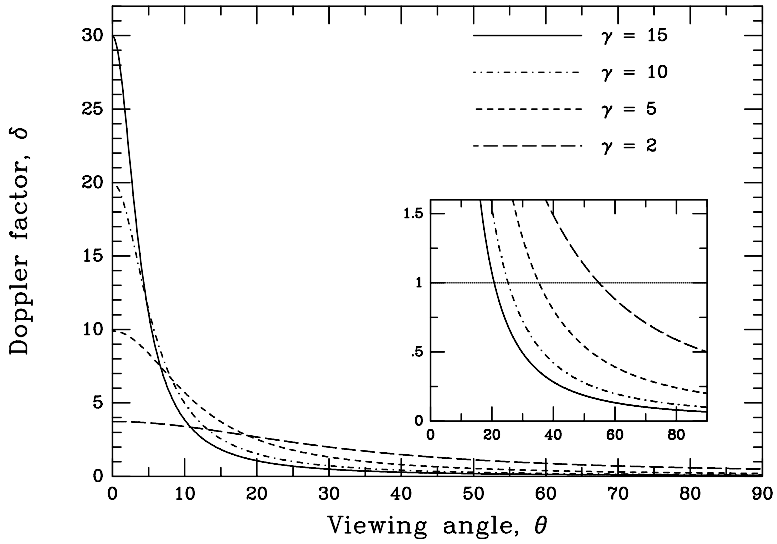


Figure 2. The dependence of the Doppler factor δ on the angle to the line of sight θ. Different curves correspond to different Lorentz factors γ. The expanded scale on the inset shows the angles for which δ=1. From Urry and Padovani (1995), copyright Astronomical Society of the Pacific, reproduced with permission from the authors.

emitting region which would create far more x-ray photons by the inverse-Compton process than is observed. This 'Compton catastrophe' (Hoyle *et al* 1966) leads to the conclusion that the blazar emission regions must be moving relativistically so the light we observe is not emitted isotropically and the variabililty timescales appear shortened. The ratio of predicted to observed inverse-Compton emission sets a lower limit on the Doppler factor, δ (cf equation (1)), which can explain this deficiency (Jones *et al* 1974). A similar argument based on γ-ray variability of blazars leads to $\delta > 6$ (Maraschi *et al* 1994), while the Compton catastrophe requires similar or even larger values of γ. Multiwavelength monitoring of BL Lac objects and blazars has been instrumental in proving that the high-energy (MeV to TeV) emission mechanism is inverse-Compton scattering within a relativistically moving jet although the location of the seed photons for this process is still debatable (Ulrich *et al* 1997). This type of observation has the potential of probing jet physics not decipherable otherwise.

X-ray loud BL Lacs (XBLs)

Because of the advent of satellite x-ray observatories, ~200 new BL Lacs have been discovered owing to their substantial high-energy luminosities, including 43 BL Lacs discovered in the Einstein Extended Medium Sensitivity Survey (EMSS; Morris *et al* 1991; Rector *et al* 2000). The vast ROSAT databases (all-sky survey and multitude of pointed observations) contain several hundred more BL Lacs that are just starting to become identified (e.g. Bade *et al* 1998). These new samples are beginning to show that the XBLs and RBLs are probably the extremes of a single population of BL Lacs (e.g. Laurent-Muelheisen *et al* 1998). However, there is also considerable evidence against this unification because, while XBLs share many of the qualities of RBLs described above, there are also significant differences between these two types including the following (see table 1 summary).

(1) Significantly higher redshifts have been measured for RBLs (e.g. $\langle z \rangle \approx 0.6$ for the 1 Jy sample; Rector and Stocke 2000; $\langle z \rangle \approx 0.3$ for the EMSS; Rector *et al* 2000). Not suprisingly, many high radio power objects are found in RBL samples but none is found among XBLs. Also, for complete samples, $\langle V/V_{\rm max} \rangle = 0.61 \pm 0.05$ for RBLs (1 Jy sample; Stickel *et al* 1991) and $\langle V/V_{\rm max} \rangle = 0.40 \pm 0.05$ for XBLs (EMSS sample; Rector *et al* 2000). Because the $\langle V/V_{\rm max} \rangle$ test allows a simple assessment of cosmological evolution using complete samples, the inference from these mean values is that RBLs were numerous and/or more luminous in the past than now and XBLs less so. For reference, $\langle V/V_{\rm max} \rangle = 0.5 \pm (12N)^{-1/2}$ is the expectation value for a complete, flux-limited sample of objects that shows no evolution (where N is the number of objects in the sample).

(2) Most BL Lacs show compact, distorted radio morphologies consistent with a highly beamed object. However, some RBLs exhibit extended radio morphologies of high-power 'classical doubles' (Fanaroff and Riley 1974 type 2 sources; FR-2s), commensurate with the radio power of their extended (and so presumably 'unbeamed') radio power levels (Kollgaard *et al* 1992; Rector and Stocke 2000). In contrast, XBLs show extended radio morphologies and power levels of low-power (FR-1) radio sources. Figure 3 shows examples of RBL and XBL radio structure.

(3) The optical spectra of RBLs and XBLs are systematically different in detail. While both are nearly 'featureless', XBLs often show weak absorption lines

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

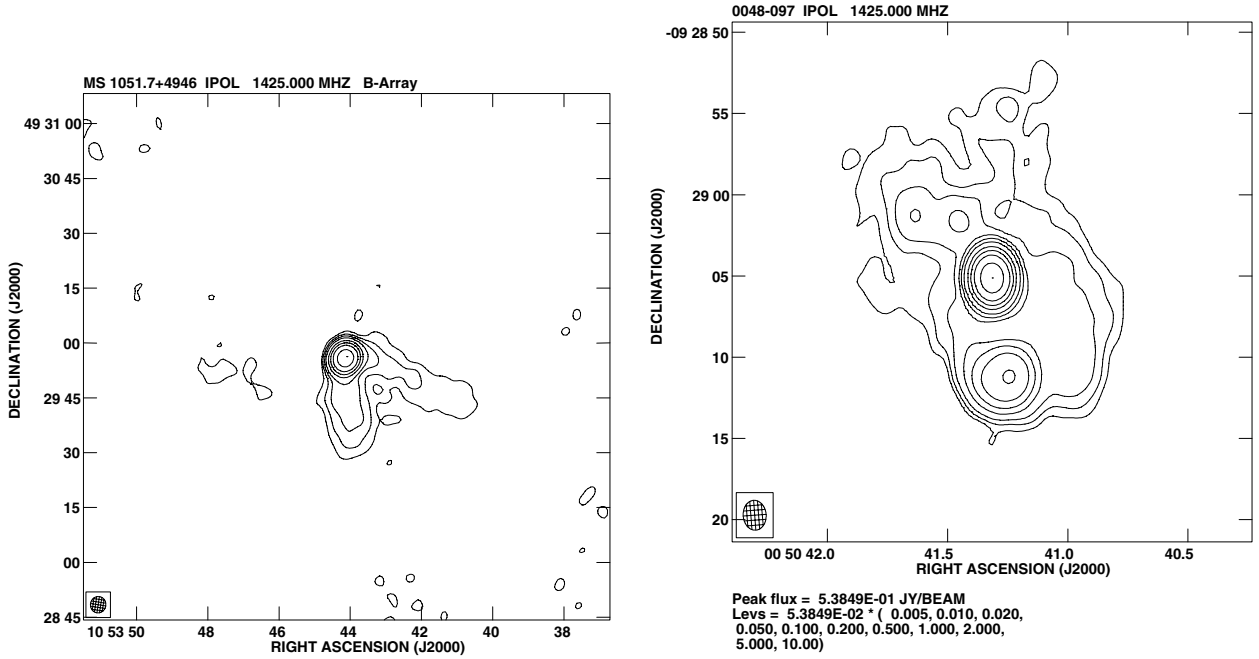


Figure 3. VLA radio maps for representative BL Lac objects. MS 1051.7+4946 (an XBL shown at left; Rector *et al* 1999) has an extended radio power (10^{24} W Hz^{-1}) and morphology typical of FR-1s (in this case the morphology of a head-tail radio galaxy). In contrast, some RBLs have extended structures which are much too luminous to be FR-1s and extended radio morphologies which resemble bent classical doubles (e.g. the RBL 1 Jy 0048-097, shown at right, has an extended radio power of $10^{26.2}$ W Hz^{-1}; Rector and Stocke 2000).

in their spectra (Morris *et al* 1991) consistent with an underlying older population of stars, as expected from deep optical imaging which reveals giant elliptical galaxies surrounding the BL Lac nucleus (Wurtz *et al* 1996; Urry *et al* 1999). While the closer RBLs are also in giant ellipticals (Wurtz *et al* 1996), the nature of the host galaxies of $z \geq 0.7$ RBLs is not known. It is the spectra of these more distant RBLs that often show weak emission lines in their spectra, unlike the XBLs. As with item (2), the weak emission lines in RBLs suggest a closer relationship between some RBLs and other blazar classes (e.g. the flat spectrum radio quasars) than do the XBLs and lower-z RBLs. Examples of high signal-to-noise ratio optical spectra of BL Lacs are shown in figure 4.

(4) While almost all BL Lac objects have overall SEDs which are very simple compared with quasars in exhibiting an overall parabolic shape with no signs of dust or accretion disk emission, some BL Lacs have SEDs which peak in the infrared (so-called 'low-energy-peaked' BL Lacs or LBLs) while others peak in the UV and x-ray ('high-energy-peaked' BL Lacs or HBLs; Sambruna *et al* 1996; Padovani and Giommi 1995). Most RBLs are LBLs and most XBLs are HBLs. Further, RBLs exhibit broadband SEDs which more closely match other blazar classes than XBLs (e.g. Brinkmann *et al* 1996). Examples of LBL and HBL SEDs are shown in figure 5.

(5) While RBLs have polarized flux fractions as high as other blazars, XBLs are more modestly polarized (Jannuzi *et al* 1994). Moreover, RBL polarizations vary considerably in both percentage and position angle, sometimes on very short timescales (<1 day; Moore *et al* 1985). Quite surprisingly, many XBL polarization vectors are quasi-stationary over long time periods (months to years; Jannuzi *et al* 1994).

The above similarities and differences between these types of BL Lacs are not easily explicable by means of a single, easily understood model. Several items, particularly the XBL trait of stable polarization position angles, strongly suggest a viewing-angle difference wherein XBLs are viewed further off the jet outflow axis than RBLs. However, the difference between the SED peaks of HBLs and LBLs is much too large to be determined by viewing angle differences alone if $\gamma \approx 3$–5 for BL Lacs. Further, since the emission line and extended radio luminosities, redshift distributions and evolutionary trends for these objects should be independent of viewing angle, differences in these observables between RBLs and XBLs are not easily explained with a simple viewing-angle hypothesis either.

For these and other reasons, models modifying the viewing-angle hypothesis have been proposed. Padovani and Giommi (1995) have suggested that XBLs simply possess a different high-energy cutoff in their electron energy distributions to RBLs. However, this model ignores orientation, which, as discussed above, is clearly important in BL Lac objects. Georganopolous and Marscher (1998) have devised a two-parameter model in which the observed properties of BL Lacs can be explained

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

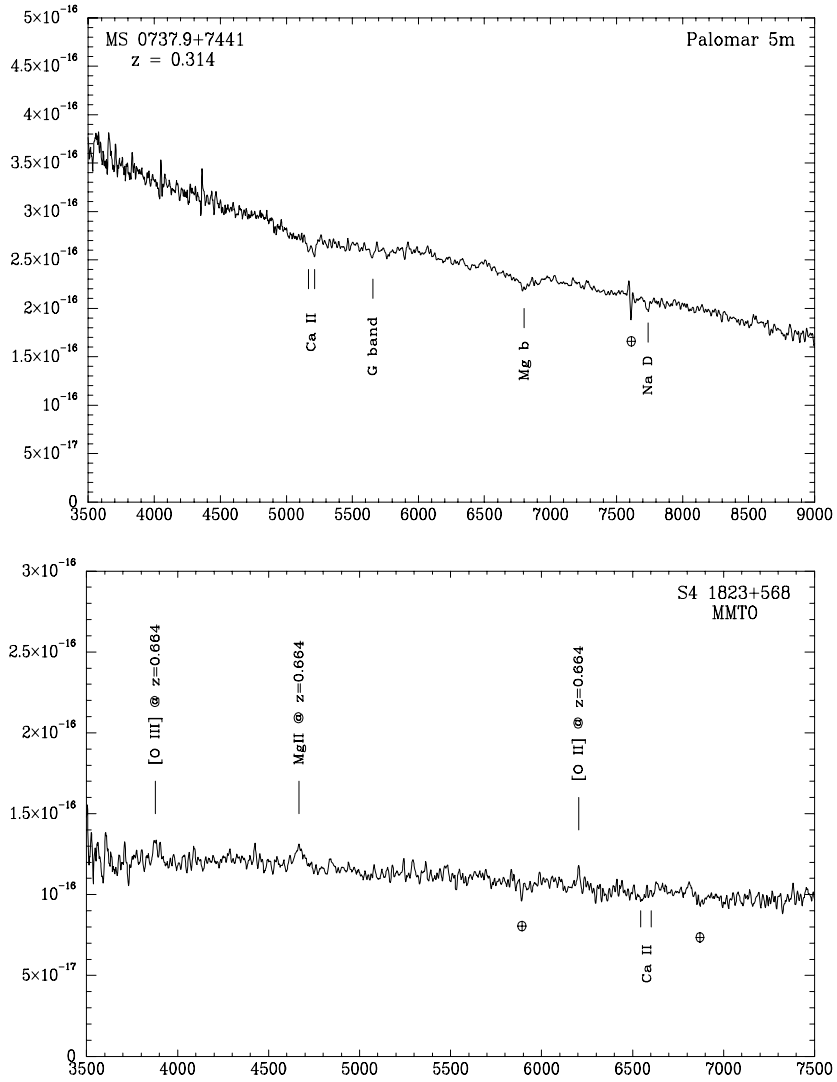


Figure 4. Representative optical spectra of BL Lac objects from the Multiple Mirror Telescope Observatory (MMTO) and from the Palomar 5 m. The spectrum of the XBL MS 0737.9+7441 (shown at top; Morris *et al* 1991) has only weak absorption features, diluted by a luminous nonthermal continuum. The spectrum of the RBL S4 1823+568 (shown at bottom; Rector and Stocke 2000) exhibits weak but moderately luminous emission lines. A few XBL and RBL spectra remain completely featureless despite heroic efforts to detect either emission or absorption lines.

by both the orientation and the electron kinetic luminosity of the jet. Their model can fit much of the data described above; however, it too is not without problems (e.g. Giommi *et al* 1999).

'Parent populations'

Given that BL Lac objects are thought to be dominated by relativistically-beamed and Doppler-boosted radiation from an outwardly moving jet seen nearly on axis, one might ask, 'what would such an object look like well away from the jet outflow axis, where the effects of beaming are negligible?' Such objects are termed the BL Lac 'parent population'. Clues concerning the parent population of BL Lacs primarily come from emission thought be isotropic (i.e. extended optical light from the 'host' galaxy, emission-line regions or extended radio emission) or from the BL Lac space density and redshift evolution.

The first identification of the 'parents' of BL Lacs came from the observed local space density of BL Lacs (Schwartz and Ku 1983). BL Lacs are simply too numerous in the local universe to be any other type of AGN than 'beamed' low-power radio galaxies (FR-1s). That is, BL Lacs at $L_x > 10^{44}$ erg s^{-1} have space densities of 10^{-7} Gpc^{-3} (Morris *et al* 1991) while FR-1 radio galaxies have space densities of 10^{-5} Gpc^{-3}, easily providing enough 'parents' for BL Lacs. Higher power FR-2 radio sources and Seyfert galaxies are actually less numerous than BL Lacs in the nearby Universe, ruling them out as parents. Urry

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

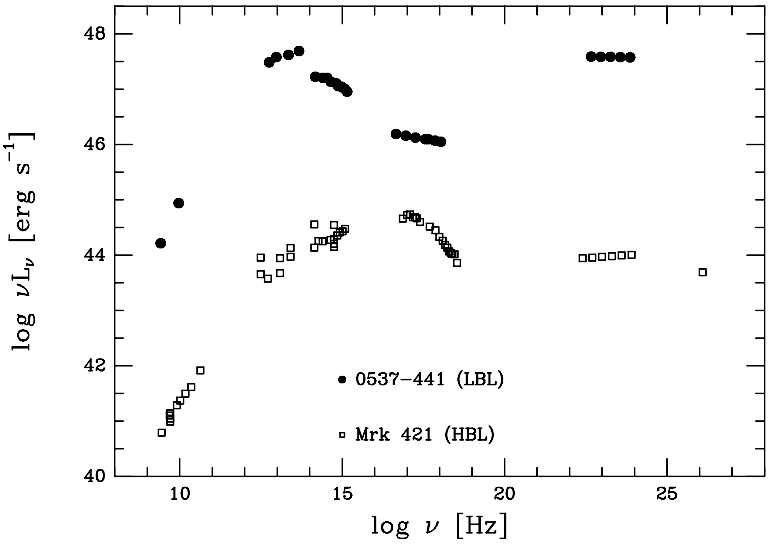


Figure 5. The SEDs of a typical RBL (PKS 0537-441) and a typical XBL (Mrk 421). RBL SEDs peak in the infrared–optical region (and are so-called 'low-energy-peaked' BL Lacs or LBL^{-1}) with extremely luminous inverse-Compton emission at high energy while XBL SEDs peak in the UV–x-ray region (and are 'high-energy-peaked' BL Lacs or HBLs) and have relatively weaker inverse-Compton emission. From Urry & Padovani (1995), copyright Astronomical Society of the Pacific, reproduced with permission from the authors.

and Shafer (1984) extended these studies significantly by constructing theoretical luminosity functions for 'beamed' objects, which showed that the luminosity function of a beamed population is flatter than that of the parent population. Subsequent papers by Urry and collaborators (see summary in Urry and Padovani 1995) used these theoretical luminosity functions to build a plausibility argument that BL Lacs are beamed FR-1 radio galaxies. Their work also suggests that only modest values of $\gamma \sim 3$–5 are required to 'boost' the luminosity function of FR-1s into the observed BL Lac luminosity function.

Very similar ranges of γ are also obtained from studies of BL Lac radio emission from 'core dominance' values (defined as the ratio of core to extended flux). BL Lacs are observed to have much higher core dominance values than FR-1s (1–30 for BL Lacs compared with 0.01–1 for FR-1s; see Wardle *et al* 1984, who first applied this method), which implies a Doppler factor $\delta \sim 30$ and thus viewing angles of a few degrees for Lorentz factors of $\gamma \sim 5$–10. However, some high-z RBLs have much higher core dominance values ($\gg$1000; Rector and Stocke 2000) and so are not easily accommodated by this picture. Further, while all XBLs have extended, and thus unbeamed, radio power levels consistent with being FR-1s, only half of the 1 Jy RBL sample are even plausibly FR-1s on the basis of their extended radio power levels; the other 50% of RBLs are too luminous to be FR-1s (Kollgaard *et al* 1992; Rector and Stocke 2000). Since FR-1 radio galaxies also lack luminous emission lines in their optical spectra, a combination of extended radio power levels and emission line properties suggests that only half (primarily the lower-z half) of the 1 Jy RBLs are plausibly FR-1s. Also, $\langle V/V_{\rm max}\rangle = 0.42 \pm 0.05$ for FR-1s (Urry and Padovani 1995) matches XBLs but not RBLs.

Optical imaging of host galaxies of BL Lacs (e.g. Wurtz *et al* 1996; Urry *et al* 1999) show convincingly that almost all BL Lacs have giant elliptical galaxies as 'hosts', which is generally consistent with the beamed FR-1 hypothesis. However, these BL Lac hosts may be a bit too faint as a class (Wurtz *et al* 1996) and in environments too poor (Wurtz et al. 1997) to match FR-1s precisely (FR-1s are often found in very rich clusters whereas BL Lacs are not). This slight inconsistency might be remedied if, for some reason, BL Lacs that are associated with the brightest cluster galaxies in rich clusters were being overlooked (i.e. misclassified as something else).

In summary, XBLs have characteristics entirely consistent with being beamed FR-1s, but only about half of all RBLs are consistent with being beamed FR-1s. The remaining, primarily high-z, RBLs are more consistent with being beamed FR-2s.

Are some BL Lacs gravitationally lensed quasars?

A continuing annoyance, and a contentious topic, of BL Lac research is whether or not gravitational 'MICROLENSING' plays a part in the appearance of some BL Lac objects. Microlensing is caused by GRAVITATIONAL LENSING of a background source by a foreground star or group of stars, creating multiple images that are microarcseconds in separation. As such, microlensing preferentially amplifies the brightness of smaller regions (down to the size of the lens itself) compared with more extended regions. Microlensing is suspected in the case of several BL Lac objects (see, e.g., Schneider *et al* 1992, p 433) because it can create one of the basic observational signatures of BL Lac objects: the very-low-contrast emission lines. In this scenario, the non-thermal continuum emission region

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a relatively normal background quasar is microlensed and amplified by foreground stars (e.g. in an intervening galaxy), thereby 'swamping' the emission lines and creating the appearance of a BL Lac.

It is now generally accepted that most, if not all, BL Lacs, are not microlensed background quasars (a hypothesis proposed by Ostriker and Vietri 1985, 1990). Still, over the years there have been several individual cases identified that are suspicious in this regard, of which AO 0235+164 is a famous example. This BL Lac is unusual in having two foreground Mg II absorption lines in its spectrum which match the redshifts of galaxies close on the sky to the BL Lac (e.g. Yanny *et al* 1989). The fast time variability of the H I 21 cm absorption line associated with one of these Mg II absorbers (Wolfe *et al* 1978) has not been explained satisfactorily using standard interpretations but is explicable using gravitational lensing scenarios. Currently there is only one confirmed case of a lensed BL Lac, the 'smallest Einstein ring' source B2 0218+357, which shows a compact ring from macrolensing. However, to this date no individual BL Lac has been proven to be due to a microlensing event. However, circumstantial evidence continues to arise; e.g., as Stocke and Rector (1997) recently noted, the 1 Jy RBL sample has a large excess of large equivalent width foreground Mg II absorbers as compared with quasars, a correlation which, if real, can be accounted for easily only in a microlensing scenario.

Conclusions

While BL Lacertae objects represent a small fraction of all AGN known, they are the only class of AGN that appear to have emission that is dominated at all wavebands by relativistically beamed jet emission. Therefore, BL Lacs routinely exhibit several unusual properties that are not common among AGN in general, including luminous emission at all wavelengths (from radio to γ-ray), apparent superluminal motion, high and rapidly variable continuum polarization, rapid time variability (especially at radio, x-ray and γ-ray frequencies), and nearly featureless UV, optical and near-IR spectra. Relativistic beaming is an essential ingredient in explaining the observed properties of BL Lacs. However, it has also become clear that a relativistically beamed jet viewed at different angles is insufficient by itself to explain both the diversity of BL Lac properties and the differences between BL Lacs and any proposed 'parent population'. Therefore, BL Lacs remain a fascinating area for future astrophysical research, not only because these AGN offer a unique opportunity to scrutinize our understandings of jet physics, but also because they remain the most enigmatic class of AGN over 25 yr after their discovery.

Bibliograhy

Adams T F 1974 *Astrophys. J.* **188** 463

Angel J R P and Stockman H S 1980 *Ann. Rev. Astron. Astrophys.* **8** 321

Antonucci R J and Ulvestad J S 1985 *Astrophys. J.* **294** 158

Bade N, Beckmann V, Douglass N G, Barthel P D, Engels D, Cordis L, Nass P and Voges W 1998 *Astron. Astrophys.* **334** 459

Begelman M C, Blandford R D and Rees M J 1984 *Rev. Mod. Phys.* **56** 255

Blandford R D and Rees M J 1978 *Pittsburgh Conf. on BL Lacs* ed A M Wolfe (Pittsburgh, PA: University of Pittsburgh) p 328

Brinkmann W, Siebert J, Kollgaard R I and Thomas H-C 1996 *Astron. Astrophys.* **313** 356

Cohen M H 1989 *BL Lac Objects* ed L Maraschi, T Maccacaro and M-H Ulrich (Heidelberg: Springer) p 13

Doxsey R *et al* 1983 *Astrophys. J. Lett.* **264** L43

Fanaroff B L and Riley J M 1974 *Mon. Not. R. Astron. Soc.* **167** 31P

Gabuzda D C, Mullan C M, Cawthorne T V, Wardle J F C and Roberts D H 1994 *Astrophys. J.* **435** 140

Georganopolous M and Marscher A P 1998 *Astrophys. J.* **506** 621

Giommi P, Menna M T and Padovani P 1999 *Mon. Not. R. Astron. Soc.* **310** 465

Hoyle F R J, Burbidge G R and Sargent W L W 1966 *Nature* **209** 751

Jannuzi B T, Smith P S and Elston R 1994 *Astrophys. J.* **428** 130

Jones T W, O'Dell S L and Stein W A 1974 *Astrophys. J.* **188** 353

Kollgaard R I, Wardle J F C, Roberts D H and Gabuzda D C 1992 *Astron. J.* **104** 1687

Laurent-Mueheisen S A *et al* 1998 *Astrophys. J. Suppl.* **118** 127

Macomb D J, Akerlof C W and Aller H D 1995 *Astrophys. J.* **449** L99

Maraschi L *et al* 1994 *Astrophys. J.* **435** L91

Moore R L *et al* 1985 *Astrophys. J.* **260** 415

Morris S L *et al* 1991 *Astrophys. J.* **380** 49

Oke J B and Gunn J E 1974 *Astrophys. J. Lett.* **189** L5

Ostriker J P and Vietri M 1985 *Nature* **318** 446

Ostriker J P and Vietri M 1990 *Nature* **344** 45

Padovani P and Giommi P 1995 *Astrophys. J.* **444** 567

Rector T A and Stocke J T 2000 *Astron. J.* submitted

Rector T A, Stocke J T and Perlman E S 1999 *Astrophys. J.* **516** 145

Rector T A, Stocke J T, Perlman E S, Morris S L and Gioia I M 2000 *Astron. J.* at press

Sambruna R, Maraschi L and Urry C M 1996 *Astrophys. J.* **463** 444

Schmitt J 1968 *Nature* **218** 663

Schneider P, Ehlers J and Falco E E 1992 *Gravitational Lenses* (Berlin: Springer)

Schwartz D A and Ku W H-M 1983 *Astrophys. J.* **266** 459

Stickel M and Kuhr H 1993 *Astron. Astrophys. Suppl.* **97** 483

Stickel M *et al* 1991 *Astrophys. J.* **374** 431

Stocke J T and Rector T A 1997 *Astrophys. J. Lett.* **489** L17

Stocke J T *et al* 1990 *Astrophys. J.* **348** 141

Ulrich M-H, Maraschi L and Urry C M 1997 *Ann. Rev. Astron. Astrophys.* **35** 445

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Urry C M, Falomo R, Scarpa R, Pesce J E, Treves A and Giovalisco M 1999 *Astrophys. J.* **512** 88

Urry C M and Padovani P 1995 *Publ. Astron. Soc. Pac.* **107** 803

Urry C M and Shafer R A 1984 *Astrophys. J.* **280** 569

Vermeulen R C *et al* 1995 *Astrophys. J. Lett.* **452** L5

Wardle J F C, Moore R L and Angel J R P 1984 *Astrophys. J.* **279** 93

Wolfe A M *et al* 1978 *Astrophys. J.* **222** 752

Wurtz R, Stocke J T, Ellingson E and Yee H K C 1997 *Astrophys. J.* **480** 547

Wurtz R, Stocke J T and Yee H K C 1996 *Astrophys. J. Suppl.* **103** 109

Yanny B, York D G and Gallagher J S 1989 *Astrophys. J.* **338** 735

John T Stocke and Travis A Rector

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blaauw, Adriaan (1914–)

Dutch astronomer, worked at Leiden, Yerkes and Groningen, where he built up the Kapteyn Astronomical Laboratory to concentrate on galactic structure. He himself studied the motions of star clusters and associations,'runaway stars' (ejected from a close double star when one of them explodes as a supernova), star formation and the cosmic distance scale. Blaauw led the foundation of the European Southern Observatory, of which he was Director General in the years 1970–1974. He was President of the International Astronomical Union in the years 1976–1979.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Black Drop

An optical phenomenon observed at the beginning and end of transits of Venus across the Sun's disk in which the dark silhouette of Venus, having just wholly passed onto the Sun's disk at one limb, or being just about to make contact with the opposite limb, is seen to be 'connected' to the limb by a narrow dark 'bridge'. This bridge is the so-called black drop. It was probably first observed at the transit of 1761; Mikhail Lomonsov concluded that it signified the presence of an atmosphere, and it is indeed caused by the refraction of sunlight through Venus's atmosphere. It has to be allowed for when timings are made of transits. Failure to appreciate this reduced the accuracy of early measurements of the solar parallax (from which the Earth–Sun distance is derived) made by timing transits of Venus from different locations on the Earth's surface. However, there may be a subjective aspect to the appearance of the black drop, for it has occasionally been reported at transits of Mercury, which has only the most tenuous of atmospheres.

See also: transit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Black Eye Galaxy

Popular name for the spiral galaxy, M64, in the constellation of Coma Berenices. It has unusually smooth spiral arms and a very prominent dark lane of dust outside its nucleus, its name deriving from the presence of this conspicuous dark patch. Designated as an Sb galaxy in the Hubble classification scheme, it lies at a distance of about 12 million light-years, and has a diameter of about 65 000 light-years.

See also: Hubble classification, interstellar absorption lines, Messier Catalog, spiral arms, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Black Hole

A region of space into which matter has fallen and from which no material object, light or signal of any kind can escape. In other words, within this region of space, the gravitational field is too powerful to let anything escape.

The basic concept is not new and, in a sense, dates back to the suggestion made in 1783 by John Michell and in 1798 by Pierre Laplace that there might exist in the universe massive bodies which would be quite invisible because the force of gravity at their surfaces would be too great to allow light to escape. At the surface of such a body the escape velocity would be greater than the speed of light. Current models of black holes are based on the general theory of relativity, from which it can be shown that if any given amount of matter is compressed sufficiently, so that it is contained within a critical radius known as the Schwarzschild radius, no signal can ever escape from it.

A black hole would arise if a quantity of matter were compressed inside its Schwarzschild radius, and the boundary of the black hole (known as the event horizon) would be a sphere having this radius. Nothing that falls inside can ever escape again. It is important to note that a black hole is not a solid body. It is a region of space whose radius depends only on the amount of matter that has fallen inside. What happens to that matter thereafter has no bearing on the hole's size, and any matter which subsequently falls in only serves to increase the radius of the hole.

The Schwarzschild radius is normally very small. For the Sun, it is only 3 km, and if, in some way, the Earth could be compressed within its Schwarzschild radius, the resulting black hole would be only the size of a marble.

Black holes will probably have been formed as a result of the collapse of the most massive stars as they run out of fuel at the end of their life cycles. Most stars collapse to form dense compact objects, but the most massive stars will continue to collapse until, in principle, all their matter is compressed into a point of infinite density known as a singularity. Before this happens, the shrinking star will have passed inside its Schwarzschild radius, giving rise to a black hole, and disappearing forever from view.

Black holes are important energy sources, as matter falling in towards a black hole, before crossing the event horizon, may release considerable amounts of gravitational potential energy. A rotating black hole (essentially similar to the simple stationary one described above) is even more efficient in this respect. Not surprisingly, black holes have been canvassed as possible 'explanations' of hitherto unexplained energy sources in the universe.

The most hopeful way in which black holes may be detected is by looking for their effects on matter in their vicinity. For example, if a black hole is a member of a binary system, then the visible star will be seen to follow an orbit round an invisible object. Again, if material is falling into a black hole, it will become excessively heated and will emit x-rays. There are a number of possible candidates under investigation, the first identified being Cygnus X-1, which is an x-ray source retaining a massive invisible object. The case of V404 Cygni is even more convincing.

Some black holes appear to have non-stellar origins. Large volumes of interstellar gas or large numbers of stars could collect and collapse into supermassive black holes at the centers of quasars and galactic systems that appear to be exploding. The enormous energy output of such objects could be accounted for by the collapse of billions of solar masses into a large black hole. In 1994, the Hubble Space Telescope provided evidence for the existence of such a supermassive black hole at the center of the galaxy M87. It has a mass of two or three billion Suns but is no larger than the solar system. In 1999, the Japanese x-ray satellite ASCA identified general relativistic effects in the supermassive black hole MCG-6-30-15, recording speeds up to $c/3$ in material orbiting 6 Schwarzchild radii from the black hole. This was the first time that general relativistic effects had been directly connected to a black hole candidate. In 2001, high-quality infrared spectra of the nucleus of Centaurus A obtained with the ISAAC instrument at the European Southern Observatory (ESO) Very Large Telescope (VLT) showed that there is a 200 million solar mass black hole at its center. The infrared spectra showed the characteristic shape of a rotating disk. High-speed motions of the gas in this disk were detected; this is the hallmark of a black hole. An analysis of the rotational speed of the disk leads to a determination of the total mass of the material inside the disk as more than 200 million solar masses—too much to be due to normal stars.

Another kind of non-stellar black hole has been proposed by the British astrophysicist Stephen Hawking. According to the theory, numerous tiny primordial black holes with masses less than an asteroid could have been created at the time of the Big Bang. These mini black holes lose mass over time and disappear.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Black-Body Radiation

An ideal emitter or absorber of radiation is called a black body. A perfect black body will absorb all radiation that falls on it, and will emit radiation that has a continuous spectrum determined only by the temperature of the black body. The distribution of energy radiated against wavelength follows a curve known as a black-body (or Planck) spectrum and, for a given temperature, there is a particular wavelength at which the maximum emission takes place.

According to the Wien displacement law, the wavelength at which a black body emits its maximum quantity of radiation is inversely proportional to its absolute temperature. This law may be used to give the approximate wavelength, λ_{max}, at which a star of effective temperature T_e has its peak emission; expressing wavelength in nanometers (nm):

$$\lambda_{max} = 2.89 \times 10^6 / T_e$$

where T_e is expressed in kelvin.

Stars and hot solid bodies are not perfect black bodies, but their radiation can be described in terms of black-body properties.

See also: absolute zero, temperature scales.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Black-Hole Candidates in X-ray Binaries

X-ray binaries are short-period (hours to days) interacting binaries in which one component is an (almost) normal star transferring material onto its compact companion via an ACCRETION DISK around the compact object. Matter spirals in the intense gravitational field, eventually falling onto the surface of the compact companion, which can be a white dwarf, neutron star or black hole. However, the x-ray luminosities generated in this accretion process are so high (sometimes $>10^{38}$erg s^{-1}) that a white dwarf can (usually) be excluded. Also while the presence of regular pulsations (usually $\sim$ seconds) indicates that the compact object must be a neutron star (a PULSAR), the absence of this signature does not automatically mean it must be a black hole. Only direct measurement of the compact object mass (from the normal star's orbital motion around the unseen compact object) can determine this. By simple application of Kepler's laws it is possible, with minimal assumptions, to infer the minimum mass of the compact object. If it exceeds a certain value (theory suggests this maximum to be $3.2M_\odot$, while observationally no confirmed neutron star exceeds $2M_\odot$) then it is declared to be a black-hole candidate on the grounds that a neutron star cannot then support itself without collapsing. Also if general relativity (GR) is the correct physical description of gravity, then such an object can only be a BLACK HOLE (see also GENERAL RELATIVITY TESTS: BINARY PULSARS). Determining the *maximum* mass of a neutron star accurately therefore has important repercussions for the nature of matter at nuclear densities.

If the mass-losing star is significantly *less* massive than the compact object (the so-called *low-mass x-ray binaries* or LMXBs), then these mass estimates provide a firm *lower* limit to the compact object's mass which is of great importance. Effectively the companion is being used as a test particle in orbit about the compact object, analogous to the (extremely low mass) planets orbiting our Sun. All-sky monitors on space-borne x-ray observatories have led to the discovery of soft x-ray transients (SXTs, so called because of their low x-ray temperatures), a class of object which produces rare, dramatic x-ray outbursts (typically separated by decades), followed by an extended period of x-ray inactivity (see also X-RAY BINARY STARS). It is only during this time that their mass-losing companions are visible, making them virtually the only class of LMXB for which dynamical mass studies are possible. More than two dozen SXTs are known, of which $\sim$10 have produced accurate mass determinations, the heaviest of which (V404 Cyg) is $12M_\odot$.

Dynamical mass measurements

A detailed knowledge of binary parameters for many of the high-mass x-ray binaries (HMXBs) (which have easily identified early-type, massive companions) is due to their pulsating neutron star compact objects. They are effectively *double-lined spectroscopic binaries*, in which each component produces a radial velocity curve (the neutron star pulsation period is also modulated by the Doppler effect). The amplitudes of these curves are K_X and K_2 (for the compact object and secondary star respectively), the ratio of which is their mass ratio, q. With an orbital period of P then Kepler's third law allows $f(M)$, the *mass function*, to be determined for each component:

$$f(M_X) = \frac{PK_2^3}{2\pi G} = \frac{M_X^3 \sin^3 i}{(M_X + M_2)^2} \quad (1)$$

$$f(M_2) = \frac{PK_X^3}{2\pi G} = \frac{M_2^3 \sin^3 i}{(M_X + M_2)^2} \quad (2)$$

where i is the orbital inclination. If the systems are eclipsing (which constrains i) then M_X and M_2 can be determined. Such work has produced the detailed dynamical mass measurements of neutron stars that are summarized in figure 1 together with the mass measurements from binary radio pulsars. All these masses are consistent with a mass of $(1.35 \pm 0.04)M_\odot$.

However, when HMXBs are suspected of harboring black holes (as in Cyg X-1 and LMC X-3), the mass measurement process is less straightforward. By definition, there will be no dynamic features (such as pulsations) observable from the compact object, and so these are *single-lined spectroscopic binaries* with only K_2, and hence $f(M_X)$, being measurable. Also, since $M_2 > M_X$, then M_X cannot be accurately determined without constraints on the value of M_2. Unfortunately, there is usually a wide range of uncertainty in M_2 ($\sim$12–20 $M_\odot$ for Cyg X-1) because of the (necessarily) unusual evolutionary history of the binary. Nevertheless the compact object in Cyg X-1 almost certainly is a black hole, since the available data constrain it to be $>3.8M_\odot$, which is just above the canonical maximum mass of a neutron star.

Low-mass x-ray binaries

However, when the mass-losing star is of low mass (the LMXBs), then the situation is very different. Such short-period x-ray binaries will have very faint companion stars, which leads to the major observational problem that the observed optical light will be dominated by reprocessed x-radiation from the accretion disk (or heated face of the companion star). This is why the optical spectra of LMXBs are hot, blue continua (U–B typically -1) with superposed (usually weak) broad hydrogen and helium emission lines, the velocities of which indicate that they largely arise in the inner disk region. This normally prevents the observation of any useful dynamical information associated with either component in the binary. Hence, the evidence for the nature of the compact object in most LMXBs has come indirectly, usually from the detection of x-ray bursts (as few are x-ray pulsars; see X-RAY BURSTERS) or the fast flickering first seen in Cyg X-1 (and hence frequently used as a black-hole 'indicator', although it too is also seen in NEUTRON STAR systems).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

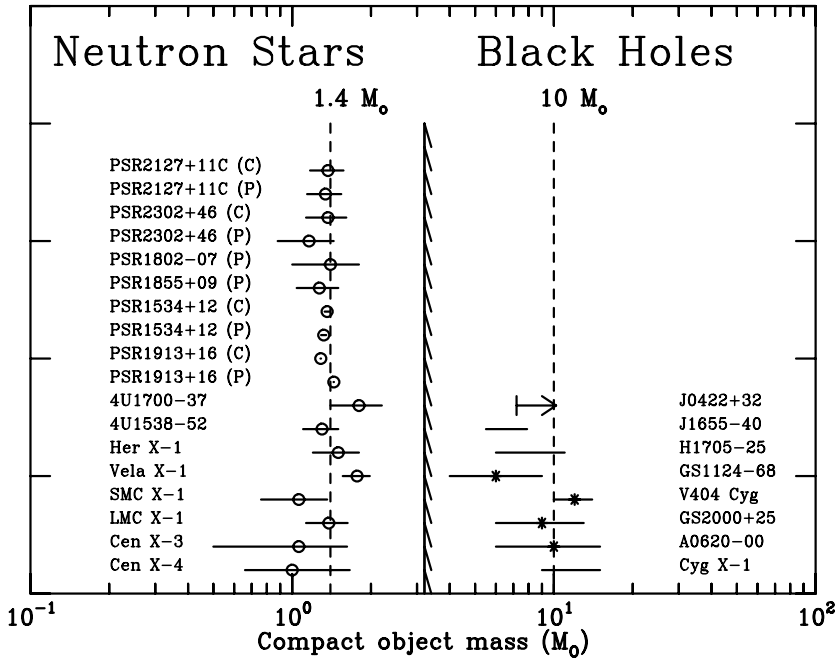


Figure 1. The distribution of neutron star and black-hole masses in binary systems.

Soft x-ray transients

It is therefore necessary to find sources where the x-ray emission switches off for some reason, so that the companion star can become visible. This is the basis of the new field of study of the SXTs (sometimes referred to as x-ray novae), which contain a remarkably high fraction (∼75%) of black-hole candidates. Only six SXTs are confirmed neutron star systems in that they display type I x-ray bursts. The observed properties of SXTs are summarized as follows.

- *X-ray transient outbursts*, reaching a typical peak x-ray luminosity of $\sim 10^{37-39}$erg s^{-1}. Erratic variability on short timescales is often present. Several new transients are found each year.
- *Fast-rise, exponential decay light curves*, with timescales ∼few days and ∼30 days respectively, recurring typically every 10–50 yr. However, linear decays are possible, especially for lower peak luminosities.
- *Variable optical counterpart* that brightens by ∼4–7^m depending on orbital period (a longer period implies a larger and hence brighter secondary star which correspondingly reduces the observed outburst amplitude). Such a large brightness increase at outburst shows that the SXTs are LMXBs, and are *not* similar to Cyg X-1 or the harder Be x-ray transients.
- Lithium has been detected in absorption in five SXTs, and may be created by spallation processes during the intense x-ray outbursts.
- Mini-outbursts sometimes occur within ∼1 yr of the main outburst.
- Superluminal radio jets analogous to those seen in QSOs and other AGN have been seen.
- Residual accretion disks seen in quiescence through the presence of strong, broad Hα emission. Usually exhibiting a double-peaked profile, this implies a disk size ∼0.4–0.7 of the primary ROCHE LOBE. Velocity studies of the emission lines have revealed small phase offsets with respect to that expected if the lines are only associated dynamically with the compact object.
- Galactic distribution of SXTs is consistent with that of population I objects, suggesting that they evolved from massive progenitor stars. There are estimated to be ∼500 in our galaxy.

Black-hole diagnostics?

X-ray spectrum

X-ray properties similar to those of Cyg X-1 are, not surprisingly, frequently used as diagnostics that lead to black-hole candidacy. In addition to a (very stable) power-law component extending to very high energies (∼MeV) which is usually attributed to Comptonization, Cyg X-1 also (occasionally) exhibits a high, soft state which has an additional (multicolor or 'disk') black-body spectrum at a temperature of $kT \sim 1$ keV. Assuming that the inner accretion disk can be represented as a set of rings of decreasing temperature with increasing radius, x-ray spectra suggest that sources suspected (on other grounds) of being black-hole candidates had inner disk radii of $\sim 3r_S$, appropriate for stellar mass black holes (where r_S is the Schwarzschild radius, and $\sim 3r_S$ is the location of the last stable orbit for matter in its vicinity). While the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

continuum spectrum is unaffected by including the effects of GR, they are important for the profiles of spectral lines formed in the inner disk. Only x-ray lines can be formed at such high temperatures, and the effects of GR are believed to have been seen in certain AGN spectra, but not yet in x-ray binaries.

The low-temperature black-body ('soft') component in SXTs has widely varying properties from source to source. Indeed, even though they are classified as SXTs, the two transients GS2023+338 and GRO J0422+32 do not show such components in their bright state (and nor do Cyg X-1 and GX339-4 when in their hard (low) states). At the other extreme, GRO J1655-40 and GRS1915+10 have the highest temperature and most luminous soft component, and are also the only SXTs to display superluminal radio ejections (see the section below on population size and distribution). This has led to the speculative suggestion that the soft component is related to the spin of the black hole, as this property is considered to be important in extragalactic radio jets.

Temporal behavior

A simultaneous analysis of the x-ray spectral and temporal variability is necessary in order to map out these source 'states'. Building on the two-state behavior of Cyg X-1, interesting correlations between their x-ray spectra and temporal variability have been found in the black-hole x-ray binaries. The presence of quasi-periodic oscillations (QPOs, as seen in neutron star systems) also reinforces their interpretation as a property of the inner accretion disk rather than the compact object. The SXT outbursts, when studied from peak to quiescence, cover a very large range in accretion rate onto the compact object, and the x-ray spectral and temporal behavior appears related to the mass accretion rate (table 1).

Furthermore, even faster QPOs (67 Hz, 184 Hz, 300 Hz) have been seen during bright phases of the transients GRS1915 + 105, XTE J1550-564 and GRO J1655-40 respectively. Such short-timescale variability must be associated with the region close to the last stable orbit of the inner accretion disk, and hence has potential importance as a property that arises very close to the black-hole primary. Various models have been proposed to account for the details of this behavior, and all include effects of GR which relate to the mass and spin of the black hole.

Quiescent flux levels

Quiescent SXT x-ray fluxes are much lower than those expected from the mass transfer rate implied by the continuing strong Hα emission, and in particular are lower than in corresponding neutron star systems. This may be due to advection-dominated accretion flows (ADAFs) in which, at low mass transfer rates, the high temperature and low density of gas in the inner disk region combine to produce a cooling time for the gas that exceeds the time to fall onto the compact object. Hence the thermal energy of the gas would be advected into the black hole and 'lost'. While similar flows would occur in neutron star systems, the thermal energy would eventually have to be radiated from the neutron star's solid surface, hence producing an apparently much higher x-ray flux from the same amount of accreted matter.

Mass measurements

Radial velocity curves

Once they reach quiescence SXTs become valuable resources for research into the nature of LMXBs. Optically fainter now by a factor of 100 or more, (they have quiescent magnitudes in the range 17–23), their emission is dominated by the companion star and, while technically challenging, their spectral type, period and radial velocity curve (whose amplitude is the K velocity) can be determined. From the latter two the mass function (equation (1)) can be calculated (table 2), and these represent the absolute minimum values for M_X, since (for all of them) $i < 90^\circ$ and $M_2 > 0$. Changing either of these parameters serves only to *increase* the implied value of M_X. However, to determine the actual value of M_X, additional constraints must be applied, so as to infer values for M_2 and i.

Rotational broadening

In short-period interacting binaries the secondary must be corotating (at velocity v_{rot}) with the primary, and since its size R_2 is given by

$$R_2/a = 0.46(1+q)^{-1/3} \tag{3}$$

then

$$v_{rot}\sin i = \frac{2\pi R_2}{P}\sin i = K_2 \times 0.46\frac{(1+q)^{2/3}}{q} \tag{4}$$

from which q can be derived if v_{rot} is measurable. Typical values are in the range 40–100 km s^{-1}.

Figure 2 shows the summed spectrum of V404 Cyg (Doppler corrected into the secondary's rest frame) and that of a template star, which clearly has much narrower absorption lines. A range of broadening velocities (together with the effects of limb-darkening) are applied to the template for comparison with the target, yielding $v_{rot}\sin i = 39 \pm 1$ km s^{-1}, from which $q = 16.7 \pm 1.4$. Note also that, while the accretion disk around the compact object might be expected to provide some velocity information, the breadth of the Hα line ($\geq$ 1000 km s^{-1}) makes this extremely difficult as the compact object's motion in such high-q systems will be very small (typically $\leq$ 30 km s^{-1}). Nevertheless such motions have been seen, but their interpretation is not straightforward as there is a small phase offset relative to the motion of the companion star.

Having determined q, the only remaining unknown is the orbital inclination i. To date, none of the SXTs is eclipsing (although GRO J1655-40 shows evidence for a grazing eclipse), and so it is the determination of i that leads to the greatest uncertainty in the final mass measurement.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. X-ray spectral and temporal properties of SXTs.

Source state	X-ray spectrum	Temporal characteristics
Low (LS)	No ultrasoft (US) component	Power-law PDS, substantial variability
Intermediate (IS)	US + steeper power law at high E	Lorentzian noise
High (HS)	US dominates (MCD model), very weak power-law component	Very little variability
Very high (VHS)	Strong US + PL component	Strong QPOs at ∼10 Hz, Lorentzian noise

Table 2. Derived parameters and dynamical mass measurements of SXTs.

Source	$f(M)$ ($M_\odot$)	q ($=M_X/M_2$)	i (deg)	M_X ($M_\odot$)	M_2 ($M_\odot$)
V404 Cyg	6.08±0.06	17±1	55±4	12±2	0.6
G2000+25	5.01±0.12	24±10	56±15	10±4	0.5
N Oph 77	4.86±0.13	>19	60±10	6±2	0.3
N Mus 91	3.01±0.15	8±2	54^{+20}_{-15}	6^{+5}_{-2}	0.8
A0620-00	2.91±0.08	15±1	37±5	10±5	0.6
J0422+32	1.21±0.06	>12	20–40	10±5	0.3
J1655-40	3.24±0.14	3.6±0.9	67±3	6.9±1	2.1
4U1543-47	0.22±0.02	—	20–40	5.0±2.5	2.5
Cen X-4	0.21±0.08	5±1	43±11	1.3±0.6	0.4

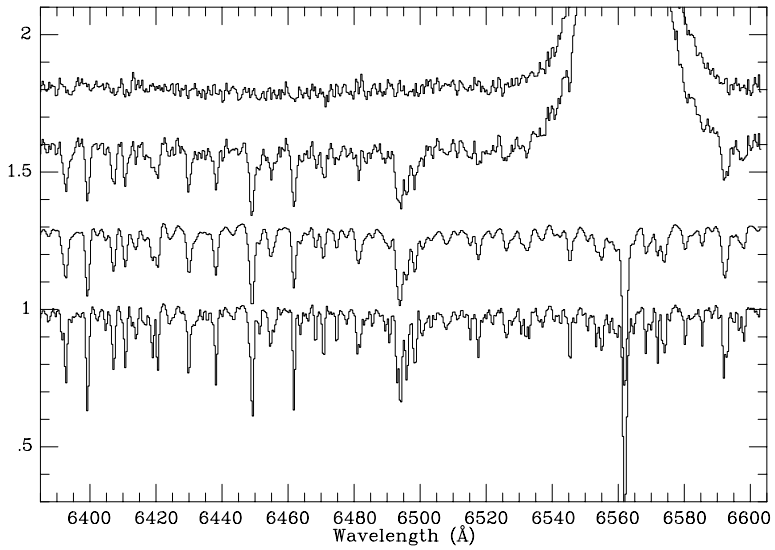


Figure 2. Determining the rotational broadening in V404 Cyg. From bottom to top: a K0IV template star; the same spectrum broadened by 39 km s^{-1}; Doppler-corrected sum of V404 Cyg (dominated by intense Hα emission from the disk); residual spectrum after subtraction of the broadened template.

Ellipsoidal modulation

The distorted shape of the secondary (the Roche lobe) produces a so-called *ellipsoidal modulation* as the projected area of the secondary varies around the orbit (figure 3). The form of the Roche lobe is theoretically well determined, and so the observed light curve depends on only two parameters, q and i (note that the modulation will disappear if i=0). Usually q is already determined (previous section), but in practice the ellipsoidal modulation is largely insensitive to q for q >5. This assumes that the secondary is actually filling the Roche lobe, which is reasonable given that mass loss from the secondary is observed via Doppler tomography even during quiescence. Results for i determined in this way are given in table 2 (and the masses in figure 1).

However, the principal uncertainty in this final stage is the problem of any other contaminating light sources, especially the accretion disk, as this will lead to i being underestimated and the compact object mass being overestimated. The disk component has already been measured in the optical (figure 2) and is typically ≤ 10%. To further reduce this effect, ellipsoidal studies

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

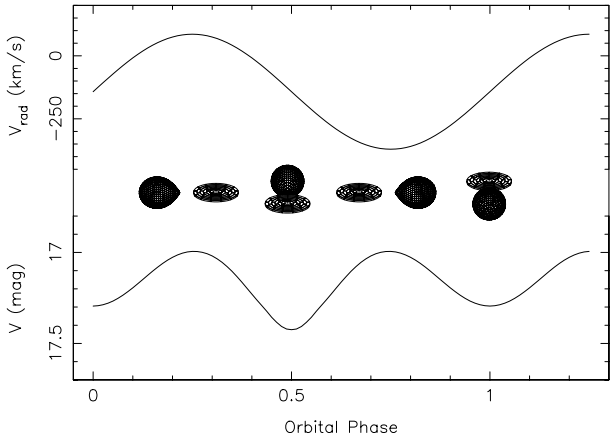


Figure 3. Schematic of the SXT (center) as a function of binary orbital phase, together with (upper) the observed radial velocity of the secondary, and (lower) the ellipsoidal light curve resulting from the distorted (Roche-lobe) shape of the secondary which produces a double-humped modulation.

are performed in the infrared whenever possible (as the disk should be blue in color). Application of this method to the neutron star SXT Cen X-4 yields a value of $1.3M_{\odot}$, in excellent accord with that expected.

The superluminal transients

The discovery, in 1994, of GRS1915 + 105 and GRO J1655-40 brought an entirely new type of behavior to this field. Although other transients were strong radio sources, VLA and VLBI observations showed that these new objects exhibited ejection events that were 'superluminal', the first time that such phenomena had been observed within the Galaxy. Furthermore, both these objects displayed continuing and extremely variable x-ray activity that is totally unlike any of the 'classical' SXTs. Combined with the extremely high interstellar extinction ($A_V \sim 26$) to GRS1915 + 105, dynamical studies of its variable, $K \sim 14$ IR counterpart are currently not possible.

GRO J1655-40 (N Sco 1994) is optically the brightest in quiescence of all the SXTs, with one of the earliest spectral types (mid-F). It also exhibits a high γ velocity which suggests that it might have been formed from a neutron star that had suffered accretion-induced collapse. A subsequent optical rebrightening was found to have begun ∼6 days before the x-rays, indicating an 'outside-in' outburst of the accretion disk. This substantial delay is explained by the inner disk needing to be re-filled before accretion onto the compact object can take place.

The brightness of GRO J1655-40 allows high-quality photometric light curves to be obtained, from which the orbital system parameters can be derived. With such an early spectral type, it has a low mass ratio ($q \sim 3$) which means that the ellipsoidal modulation is sensitive to both q and i (the latter being constrained by the observed grazing eclipse). The normal star appears to be crossing the Hertzsprung gap and about to ascend the giant branch, hence driving the much higher mass transfer rate. Temporary drops in $\dot{M}$ then return it to the transient domain.

Population size and distribution

The locations of ∼90% of SXTs within a galactic longitude range of ±80° implies that they lie within 8 kpc of the galactic centre. However, distance estimates for the current sample suggest they are all < 5 kpc away, and so only ∼10% of the transients occuring in our Galaxy are being detected (owing to a combination of interstellar absorption and sensitivity). Hence the total number of SXTs is ∼200–1000 for a recurrence time range of 10–50 yr. The narrower dispersion in z of the SXTs, compared with neutron star LMXBs, implies that their black holes are not formed by accretion-induced collapse of a neutron star (except possibly GRO J1655-40), but more likely as the end-product of evolution of a very massive progenitor.

Nature of the compact object

With the secure observation of compact objects with masses $\geq 10M_{\odot}$, their interpretation as black holes depends on a knowledge of the maximum mass of a neutron star. This is usually quoted as $3.2M_{\odot}$, but this is based on an assumption of causality (i.e. the sound speed is $<c$) and a knowledge of the density up to which the equation of state is well understood. Relaxing these can increase the limit.

Alternative suggestions for the nature of compact objects include Q stars, where the strong force confines neutrons and protons at densities below nuclear density, leading to a very different equation of state (in which the Q stands for conserved quantity, the baryon number). They can be very compact and hence consistent with the properties of neutron stars. However, the mass of V404 Cyg shows that it cannot be a Q star as it would be inconsistent with current experimental results. In which case, it must be concluded that V404 Cyg and related objects contain black holes.

Bibliography

Charles P 1998 Black holes in our galaxy: observations *Theory of Black Hole Accretion Disks* ed Abramowicz, Bjornsson and Pringle (Cambridge: Cambridge University Press) p 1

McClintock J E 1991 *Ann. NY Acad. Sci.* **647** 495

Mirabel I F and Rodríguez L F 1994 *Nature* **371** 46

Tanaka Y and Shibazaki N 1996 *Ann. Rev. Astron. Astrophys.* **34** 607

Van Paradijs J and McClintock J E 1995, *X-Ray Binaries* (Cambridge: Cambridge University Press) p 58

Philip A Charles

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blazar

The most variable class of active galaxy. This class of active galaxy includes BL Lac objects and some of the most violently variable quasars, the name 'blazar' deriving from an amalgamation of 'BL Lac' and 'quasar'. According to the unified model for active galaxies, blazar activity is caused by jets of energetic particles flowing out from an active galactic nucleus directly along, or very close to, the observer's line of sight.

See also: active galactic nucleus, active galaxy, astrophysical jets, BL Lacertae objects, quasar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blink Comparator

A device, also known as a blink microscope, that enables two photographs or images of a region of sky, obtained at different times, to be compared in such a way that any differences between them are highlighted. The light paths from the two images are arranged so that, when viewed through an eyepiece, they coincide exactly. The two images are then illuminated alternately, switching from one to the other at a frequency of once or twice per second. While features that are identical on both images appear constant and unchanging during the alternating illumination, any features that are present on one plate only appear to blink on and off. Objects that have changed in brightness fluctuate in size, while objects that have changed position appear to jump to and fro in the alternating illumination. The blink comparator, therefore, greatly facilitates the discovery and identification of objects such as variable stars, novae, supernovae, asteroids and comets. Virtual blink comparators are used in digital image analysis systems for CCD data.

See also: stereo comparator

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blinking Planetary (NGC 6826)

A planetary nebula in the constellation Cygnus, position RA 19^{h} 44.8^{m}, dec. $+50^{\circ}\,31'$. It is $\frac{1}{2}'$ in diameter and of eighth magnitude, with a tenth-magnitude central star. It gets its name because it tends to vanish when looked at directly through a small telescope, when light from the central star swamps the nebula, and reappears when viewed with averted vision.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bliss, Nathaniel (1700–64)

Born Bisley, Gloucestershire, England, fourth Astronomer Royal. Bliss directed the Royal Observatory at Greenwich for only two years and had little impact.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blue Stars at High Galactic Latitudes

When you look up into the sky at night, most of the blue stars you see are hot main-sequence stars associated with the Milky Way, which have rather recently formed from the thin layer of dust and gas in the plane of our Galaxy. Because these stars evolve rapidly, we do not see them at high galactic latitudes, where there is no interstellar material from which they can form. Hence blue stars at high galactic latitudes are rare and must have some other explanation. It was this rareness of the blue stars at high galactic latitude and the lack of a theoretical explanation for them which led Fritz Zwicky to undertake the first survey for such stars, the HZ survey, with the 0.46 m Schmidt telescope on Mt Palomar in 1938. This telescope (and its big brother, the 1.2 m Schmidt) were in fact built to carry out surveys like this one, in order to find objects which were sufficiently interesting to deserve further study with the 5 m Hale reflector.

This and subsequent surveys for blue stars at high galactic latitudes have paid off handsomely, because these stars turn out to tell us not about the early stages of stellar evolution but about the final white dwarf state and the rapid phases of stellar evolution which immediately precede it. In 1965, a new chapter in the story of blue stars at high galactic latitudes was started by Allan Sandage, who predicted that most of the high-latitude blue stars fainter than apparent magnitude 16 in existing surveys were not stars at all but (radio quiet) quasars, the most luminous known objects in the universe. Although later studies showed that this did not actually happen until apparent magnitude 18, the fact remains that the faintest blue 'stars' at high galactic latitudes are one of our most important sources of information on the structure and evolution of the universe. It is also ironic that our primary source of information on the faintest blue stars in our own Galaxy is currently a byproduct of extragalactic research.

In this article we will discuss past and current surveys of blue stars at high galactic latitudes. Deep quasar surveys, which have been recently reviewed by Hartwick and Schade (1990 *Ann. Rev. Astron. Astrophys.* **28** 437), are not included unless it was felt that they have contributed significantly to the study of faint blue stars. What we have learned from further study of the objects discovered in these surveys may be found in the articles on WHITE DWARFS, PG1159 STARS, LARGE HOT SUBDWARFS AND EXTREME HORIZONTAL BRANCH STARS, PLANETARY NEBULAE, RED GIANT STARS, HORIZONTAL BRANCH STARS, RR LYRAE STARS, BLUE STRAGGLERS and LUMINOUS BLUE VARIABLES.

The earliest surveys for blue stars at high galactic latitudes

The Humason Zwicky (HZ) Survey

The HZ survey was carried out using four sets of photographs taken with the 46 cm Palomar Schmidt which covered the Hyades star cluster and a region of about 150 square degrees near the North Galactic Pole. These were taken using emulsion–filter combinations which isolated the spectral regions 3500–5400 Å, 4900–7000 Å, 3500–5200 Å and 3500–3900 Å. The yellow and violet exposures were used to select 48 blue (color index from 0.3 to −0.1), decidedly blue (color index from −0.1 to −0.3), and very blue (color index from −0.3 to −0.6) stars down to apparent photographic magnitude 15. The blue exposures were used to eliminate variable stars. Follow-up spectroscopy at resolutions of 4–8 Å was done by M L Humason with the Mt Wilson 2.5 m reflector, and the results were published by Humason and Zwicky in 1947 (*Astrophys. J.* **105** 85).

The Luyten blue (LB) survey

The search for faint blue stars at high galactic latitudes was greatly extended by Willem J Luyten and his collaborators between 1951 and 1969 using photographic techniques similar to those of the HZ survey. A variety of instruments, including the 1.2 m Palomar Schmidt, was used to search fields all over the sky (most at high galactic latitude), resulting in the discovery of 11 444 stars brighter than apparent magnitude $B = 18$ and bluer than color index $U\text{-}V = -0.2$. Unfortunately, finding charts were not published for many of the fainter stars, and the published coordinates are only accurate to 1 minute of arc, so that much of this material is of statistical interest only. Luyten speeded up the discovery process considerably through the use of a blink microscope, a device which allowed him to view the photographic plates taken in red and blue light in rapid succession through the same eyepiece. Haro's three-image method, which is described in the next section, was also used. The results were published by the Observatory of the University of Minnesota as a series of 52 contributions and monographs by Luyten and his collaborators. Proper motions for many of the blue stars at high galactic latitude have also been published by Luyten.

The Feige survey

This survey was initiated by Jacques Feige in 1956. Red and blue plates taken with the 1.2 m Palomar Schmidt for the Palomar Observatory Sky Survey were blinked for 3600 square degrees at galactic latitudes higher than +55° and for 2400 square degrees for galactic latitudes below −45° and north of declination −15°. 114 stars with apparent photographic magnitude brighter than 14 and color indices less than $B\text{-}V = 0.2$ were identified. This survey, like the HZ survey, proved to be very important in determining the nature of the faint blue stars at high galactic latitude, because the stars were bright enough for follow-up observations to be made with the 5 m and other telescopes (*Astrophys. J.* **128** 267, *Astrophys. J.* **129** 600).

Surveys using Haro's multiple-image technique

The Palomar Haro Luyten (PHL) survey

This survey was carried out in 1958 by G Haro and W J Luyten with the 1.2 m Palomar Schmidt and used a triple exposure through three different filters and a slight offset of the telescope between the exposures to produce a triple image of each star in ultraviolet, blue and visual light on a single photographic plate. 8746 stars bluer than color

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

index $U\text{-}V = 0.1$ and brighter than apparent photographic magnitude 19 were identified in a 2400 square degree region near the South Galactic Pole (SGP). Unfortunately, as with the LB survey, the published coordinates are only given to 1 minute of arc, and finding charts were not published, so that much of this material is of statistical use only.

The Tonantzintla (Ton) survey
This survey was carried out at the same time as the PHL survey using the Schmidt telescope at the Tonantzintla Observatory. Areas of 1860 square degrees in the North Galactic Polar Cap and 1050 square degrees in the South Galactic Polar Cap were searched by B Iriarte and E Chavira using the technique described in the previous section. The survey resulted in the detection of 2008 blue stars brighter than photographic magnitude 17 (Iriarte B and Chavira E 1957 *Bol. Obs. Tonantzintla Tacubaya* (16), (17) and (18)).

The Jaidee–Lynga (JL) survey
This was based on dual-image (U and B) plates taken with the 50/65/175 cm Schmidt at Mt Stromlo (Australia). The catalog contains 296 stars in 540 square degrees at high galactic latitudes with $U\text{-}B < -0.6$ and $B < 18$ (1969 *Ark. Astron.* **5** 345).

The PB (Palomar Berger) survey
This was an extension of the PHL survey described above in 2000 square degrees near the SGP (4495 stars) and 450 square degrees near the NGP (4515 stars) (Berger J and Fringant A-M 1984 *AASS* **58** 565).

The KUV (Kiso Ultraviolet) survey
This was based on three-image UGR (ultraviolet, green, red) plates taken with the 105 cm Schmidt at the Kiso Observatory (Japan) (Noguchi *et al* 1980 *Ann. Tokyo Astron. Obs. Ser. 2* **18** 55, Kondo *et al* 1984 *Ann. Tokyo Astron. Obs. Ser. 2* **20** 130). A total of 1186 stars with ultraviolet excesses and $10.0 < G < 18.5$ were identified in a 4000 square degree area, mostly along a belt between the NGP and SGP centered on galactic longitude 180°. Follow-up spectroscopy at a resolution of 12 Å (4500–8200 Å) or 7 Å (4100–7200 Å) is being obtained with the 2.4 and 1.3 m telescopes at the Michigan–Dartmouth–MIT Observatory on Kitt Peak (see Darling G W and Wegner G 1994 *Astron. J.* **108** 2025.

The Usher (US) Survey
This survey is based on Palomar 1.2 m Schmidt plates of 206 square degrees of the sky with galactic latitude greater than 30°. The plates were taken in ultraviolet, blue and visual spectral regions and have isolated 3987 blue stellar objects which are brighter than $B = 18.8$ and bluer than F and G subdwarfs in *B-V* or *U-B*. A spectroscopic follow-up is being done between 3500 and 7000 Å at 14 Å resolution with the Kitt Peak 2.1 m spectrograph. The results to date include the detection of 67 white dwarfs, 36 hot subdwarfs and 30 B-8 and early A-type stars (Usher P D and Mitchell K J 1990 *Astrophys. J.* Suppl. Ser. **74** 885).

The Survey of Rubin, Moore and Bertiau
This survey (Rubin, Moore and Bertiau 1967 *Astron. J.* **72** 59) yielded 233 objects with apparent magnitudes between 15 and 19 and *U-V* bluer than a star of spectral type A0–A5 in a 36 square degree region of the Virgo cluster of galaxies. It is based on a single triple-image plate taken with the Palomar 1.2 m Schmidt.

KPD (Kitt Peak–Downes) survey
This was based on U and B plates taken with the 0.6 m Schmidt at Kitt Peak (Arizona) and covering 1000 square degrees of the Milky Way down to $B = 15.3$. Although this survey was not done at high galactic latitudes, it has been included because the stars selected were similar to those found in the high-latitude surveys and because it contains references to earlier color surveys at low galactic latitudes. The selection of ∼700 objects with $U\text{-}B < -0.5$ was done from visual inspection of overlapping contact films made from the plates. Spectra with 15 Å resolution covering the range 4000–6800 Å were obtained with the Mount Lemmon (Arizona) UCSD–Minnesota 1.5 m telescope (see Downes R A 1986 *Astrophys. J.* Suppl. Ser. **61** 569).

The PG Survey and its southern counterparts

The Palomar Green (PG) survey
This survey was published by R F Green, M Schmidt and J Liebert in 1986 (*Astrophys. J.* Suppl. Ser. **61** 305) and is based on ultraviolet and blue plates obtained with the 46 cm Palomar Schmidt. The plates covered an area of 10 714 square degrees above galactic latitude +30° and 2000.0 declination −10°. The plates were scanned and digitized, and the magnitudes, colors and positions of some 4 million objects then determined by computer. Comparison with photoelectric magnitudes and colors then allowed 1874 objects bluer than color index $U\text{-}B = -0.46$ and brighter than apparent magnitude $B = 17$ to be identified. The positions of all of these objects were published along with timing charts and the results of follow-up photometry and spectroscopy. The latter showed that 6% of the stars were blue horizontal-branch stars, 53% hot subdwarfs, 26% white dwarfs, 5% cataclysmic variables or other binaries and 9% extragalactic.

The Edinburgh–Cape (EC) survey
This is one of two Southern counterparts of the PG survey which are still in progress, and which will cover the 10 000 square degrees of the southern sky with $|b| > 30°$. Ultraviolet and blue photographic plates obtained with the 1.2 m UK Schmidt telescope in Australia are being analyzed by methods similar to those utilized in the PG survey, and the survey is complete to at least the limits of the PG survey, $B = 17$ and $U\text{-}B = -0.4$. Follow-up spectroscopy and photometry are being done at the South African Astronomical Observatory. The results obtained to date (zone 1) include the detection of 101 white dwarfs, 303 hot subwarfs, 75 blue horizontal-branch stars, 62 normal B and A stars, 55 extragalactic, 71 composite and 8 other.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Montreal Cambridge (MC) survey
This is the other counterpart of the PG survey which is still in progress, and it is very similar to the EC survey. It is based on doubly-exposed (U and B) photographic plates taken with the Curtis Schmidt telescope at Cerro Tololo (Chile), and should be complete to $B = 16.5$ and $U\text{-}B = -0.4$ for the 6750 square degrees south of galactic latitude -30° which were not covered by the PG survey (Demers *et al* 1986 *Astron. J.* **92** 878). Follow-up spectroscopy for 251 objects with $U\text{-}B < -0.5$ and brighter than $B = 16.5$ in an 840 square degree region at the pole produced the following results: 54 white dwarfs, 72 hot subdwarfs, 4 cataclysmic variables, 6 blue horizontal branch stars, 16 normal B and A stars, 31 extragalactic objects and 3 other emission-line objects.

Other color surveys

The survey of Braccesi et al
This survey (Braccesi *et al* 1970 *Astron. Astrophys.* **5** 264) is based on u, b, v and i plates taken with the Palomar 1.2 m Schmidt. The catalog contains 175 objects with $u\text{-}b < -0.3$ and $b < 20$ in a 42 square degree field centered at 13 h right ascension and +36° declination. Corrected values for the equatorial coordinates are given in *Astronomy and Astrophysics* (1973 *Astron. Astrophys.* **23** 159). This survey yielded the only published list of optically selected quasars for almost 10 yr. It was extended by Formiggini *et al* (1980 *Astron. Astrophys. Suppl.* **39** 129) using u, b and v plates of the central 1.72 square degrees of this region. The second catalog contains 298 objects with $u\text{-}b < -0.1$ or $u\text{-}v < +0.4$ and $18.0 < b < 20.2$.

The Survey of van den Bergh
This survey (van den Bergh 1966 *Astrophys. J.* **144** 866) yielded 31 objects with $17.5 < B < 20.5$ and $U\text{-}B < -0.5$ in a 4 square degree field in Andromeda. These objects were identified by blinking a well-matched pair of U and B plates taken with the 1.3 m Schmidt telescope of the Karl Schwarzschild Observatory (Germany).

The CCD/Transit Instrument (CTI) blue object survey
This was based on CCD photometry done with the 1.8 m transit instrument at Kitt Peak. This instrument is able to do UBVRI photometry of stars brighter than $V = 20$ in a 8.26 minutes of arc strip at declination +28°. Follow-up spectroscopy for 110 objects is reported by Kirkpatrick and McGraw 1998 *Proc. IAU Colloq.* **114**.

AAT (Anglo-Australian Telescope) survey
This survey (Boyle B J *et al* 1990 *Mon. Not. R. Astron. Soc.* **243** 1) used UK Schmidt U and J plates for 34 0.35 square degree fields at high galactic latitude. The catalog contains 1400 stars with $16.9 < B < 20.9$ and $U\text{-}B < -0.35$ and differs from its forerunners in that 12–15 Å resolution spectroscopy has been obtained with the AAT for all of the stars, and the results are included in the catalog.

Cerro El Roble (CT) survey
This survey is based on ultraviolet and blue photographic plates taken with the 0.7/1.0 m Maksutov telescope at Cerro El Roble (Chile). Blinking of these plates has produced a catalog of 373 stellar objects with $U\text{-}B < -0.5$ down to $B = 20$ in a 44 square degree region near the south galactic pole (Campusano L E and Torres C 1983 *Astron. J.* **88** 1304).

HBQS (Homogeneous Bright QSO) survey
In this survey (Cristiani *et al* 1995 *Astron. Astrophys. Suppl. Ser.* **112** 347) objects with ultraviolet excesses and $B < 19$ were identified on ESO and UK Schmidt UBVRI plates in 153 square degrees near the SGP. 285 QSOs and a comparable number of blue stellar objects were then identified from 10–30 Å resolution ESO spectra of the objects with ultraviolet excesses. This survey is to be extended to entire south galactic polar cap ($b < -60^\circ$), and has been used to make a new determination of the luminosity function of hot white dwarfs.

Objective-prism surveys

The survey of Slettebak and Stock
This was done with the Hamburg 80 cm Schmidt and covered roughly 750 square degrees centered at the NGP. The spectroscopic dispersion was 590 Å mm^{-1} at H gamma, and the limiting magnitude was between $B = 12$ and 14. The survey yielded 601 stars of spectral type F2 and earlier (1959 *Astron. Abh. Hamburg Sternwarte* **5** (5)).

The Slettebak and Brundage (SB) survey
This was done with the Schmidt telescope at Cerro Tololo (Chile) and covered 840 square degrees centered at the SGP. The spectroscopic dispersion was 580 Å mm^{-1} at H gamma, and the limiting magnitude was between $B = 14$ and 15. The survey yielded 957 stars of spectral type F0 and earlier, and the results are given in the *Astronomical Journal* (1971 *Astron. J.* **76** 338).

A G D Philip and collaborators
These researchers have identified 1108 stars of spectral type A7 and earlier on objective-prism plates of various regions at high galactic latitudes taken with the Schmidt telescopes at Tonantzintla (Mexico), Cerro Tololo (Chile) and the Warner and Swasey Observatory (Ohio). The spectra cover the blue region at dispersions ranging from 280 to 580 Å mm^{-1} at H gamma down to $B = 13$. The results are published in the Boletin de los Observatorios Tonantzintla y Tacubaya.

The FHB surveys
These high-latitude surveys were carried out with the Schmidt telescopes at Cerro Tololo (Chile) and Kitt Peak (Arizona). They employed an ingenious technique which allowed early-type stars to be identified to apparent magnitude $B = 15$ at a dispersion of 280 Å mm^{-1} at H gamma. This technique employed a blocking filter which allowed only light of the 150 Å spectral region

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

centered on Ca II H and K lines to reach the photographic plate, thus eliminating 85% of the sky background, which limits the exposures. Stars with the Ca II K line missing (FHB/AB) have effective temperatures greater than 10 000 K, and those where it is weak (FHB/A) are primarily blue horizontal branch stars at high galactic latitudes. More than 11 500 FHB/A stars have been identified to date by visual scanning of the plates (Beers *et al* 1992 *Astron. J.* **103** 267).

The LS (Luminous Stars) surveys
These were objective-prism surveys (Stephenson C B and Sanduleak N 1970 *Publ. Warner Swasey Obs.* **1** 1, Drilling J S and Bergeron L E 1995 *Proc. Astron. Soc. Pac.* **107** 846) carried out with Schmidt telescopes at Hamburg, Cleveland and Cerro Tololo, at 280–580 Å mm^{-1} at H gamma. These surveys cover the entire Milky Way plus $|b| < 30^\circ$ for $300^\circ < l < 60^\circ$. These surveys have been included because follow-up spectroscopy has yielded many of the same kinds of objects found at high latitudes and because they contain references to earlier low-latitude objective-prism surveys.

The Case Low-Dispersion Northern Sky survey
This is an objective-prism survey in progress which will cover the 8000 square degrees of the northern sky with $b > 30$ using the Case 61 cm Schmidt at Kitt Peak. The photographic spectra have a dispersion of 1360 Å mm^{-1} at H gamma and cover the spectral region from 3300 to 5350 Å down to photographic magnitude $B = 15$–18. Visual scanning of the plates has turned up 1579 blue or emission line galaxies, 1136 unresolved blue or emission line objects, 465 known or suspected blue stars, 1110 A and F stars, and 132 faint carbon and suspected late M halo giants at the time of this writing, as well as a number of probable H II regions in other galaxies. The results are being published in the *Astrophysical Journal Supplement*.

The Hamburg Quasar Survey (HQS)
This is similar to the Case survey, but using the Hamburg 80 cm Schmidt telescope at Calar Alto (Spain). The entire northern sky will be covered for $|b| > 19^\circ$ down to $B = 18.5$. The spectra cover the wavelength range 3400 to 5400 Å and have a dispersion of 1390 Å mm^{-1} at H gamma. The plates are being scanned and analyzed digitally, and this is the first objective-prism survey to utilize an automatic selection of objects rather than visual scanning (Hagen *et al* 1995 *Astron. Astrophys. Suppl. Ser.* **111** 195).

The Hamburg/ESO survey (HES)
This is the southern counterpart of the HQS survey described in the previous section, based on objective-prism plates taken with the ESO 1 m Schmidt at a dispersion of 450 Å mm^{-1} at H gamma, which reach down to $B = 17.5$. Results for the first 1000 square degrees of coverage are reported by Wisotzki *et al* (1996 *Astron. Astrophys. Suppl. Ser.* **115** 227).

The First Byurakan survey
This survey was begun by B E Markarian to search for active galactic nuclei. More than 2000 objective-prism plates were taken with the 1 m Schmidt telescope of the Byurakan Observatory (Armenia), covering the entire northern sky and that part of the southern sky with declination $> -15^\circ$ and $b > 15^\circ$. The spectra cover the range from 3400 to 6900 Å (with a gap near 5300 Å) at a dispersion of 1800 Å mm^{-1} at H gamma. Scanning of 438 of these plates by H V Abrahamian, A M Mickaelian and others has yielded 1103 blue stellar objects. Follow-up spectroscopy with the Byurakan 2.6 m telescope indicates the following composition for the sample: 56% hot subdwarfs, 16% white dwarfs, 8% main sequence OB and HBB stars, 3% cataclysmic variables and other binaries, 4% extragalactic and 13% other.

The survey of Rodgers, Roberts, and Walker
In this survey (Rodgers, Roberts and Walker 1993 *Astron. J.* **106** 591) objective-prism plates were obtained in two 70 square degree regions at $l = 90^\circ$ and 270°, $b = -45^\circ$ with the Anglo-Australian Schmidt. The spectra have a dispersion of 600 Å mm^{-1} at H gamma and 436 stars with spectral type F and earlier and $V < 16$ were classified and cataloged.

The future of surveys for blue stars at high galactic latitudes

The era of color surveys for blue stars at high galactic latitudes is probably nearing an end. The SLOAN DIGITAL SKY SURVEY (SDSS) is a good example of what lies in the future. The SDSS will employ a 2.5 m telescope and CCD detectors to obtain accurate five-color broad-band photometry covering the entire visible and near-infrared spectrum for all stars brighter than apparent magnitude $g' = 23$ in a 10 000 square degree area centered on the NGP. Photometric and positional information for both the stars and extended sources will be stored in a gigantic database which astronomers will be able to access remotely. Each astronomer will therefore be able to make his or her own survey by computer, using the SDSS database rather than the real sky. It is, in fact, already possible to do this using one of the three existing independent digitizations of the original Palomar Observatory Sky Survey. The catalog obtained by scanning and digitizing the original plates of the Palomar Observatory Sky Survey with $|b| > 19^\circ$ and $\delta > -31^\circ$ using the automatic plate scanner at the University of Minnesota is available over the internet at http://isis.spa.umn.edu.

Objective prism surveys are a different story, but the scanning and digitization of objective prism plates have already been successfully used in the Hamburg quasar surveys to make at least an initial classification of objective-prism spectra, and it is probably just a matter of time before large-scale objective-prism surveys are done with digital detectors.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Adelman S J, Upgren A R and Adelman C J (eds) 1994 *Hot Stars in the Galactic Halo* (Cambridge: Cambridge University Press)

Capaccioli M (ed) 1983 *Astronomy with Schmidt-Type Telescopes* (Dordrecht: Reidel)

Chapman J, Cannon R, Harrison S and Hidayat B (eds) 1983 *The Future Utilization of Schmidt Telescopes* (San Francisco, CA: Astronomical Society of the Pacific)

Luyten W J (ed) 1965 *1st Conf. on Faint Blue Stars* (Observatory of the University of Minnesota)

MacGillivray H T, Thompson E B, Lasker B M, Reid I N, Malin D F, West R M and Lorenz H (eds) 1994 *Astronomy from Wide-Field Imaging* (Dordrecht: Kluwer)

McLean B J, Golombek D A, Hayes J J E and Payne H E (eds) 1998 *New Horizons from Multi-Wavelength Sky Surveys* (Dordrecht: Kluwer)

Philip A G D, Hayes D S and Liebert J W (eds) 1987 *2nd Conf. on Faint Blue Stars* (Schenectady, NY: Davis)

Philip A G D, Liebert J W and Saffer R A (eds) 1994 *3rd Conf. on Faint Blue Stars* (Schenectady, NY: Davis)

Soifer B T (ed) 1993 *Sky Surveys: Protostars to Protogalaxies* (San Francisco, CA: Astronomical Society of the Pacific)

John S Drilling

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blue Stragglers

The most flagrant blue stragglers are stars that cannot be older than one or two billion years, lurking in the cores of 13 billion year-old GLOBULAR CLUSTERS. The origin(s) and evolution of these puzzling objects are the central challenges confronting researchers in this field.

ALLAN SANDAGE was the first to note the existence of blue straggler stars (BSS) when he constructed a deep color–magnitude diagram for the globular cluster M3 as part of his PhD thesis in 1952. This classic diagram (reproduced in figure 1) demonstrated beyond doubt that at least two dozen stars in the cluster lie near the hydrogen-burning main sequence, but that they are hotter and more luminous than stars at the main sequence turnoff. Only two explanations seemed possible: either these stars were considerably younger and more massive than all other cluster stars, or the theory of STELLAR EVOLUTION was seriously in error. As many lines of evidence strongly supported stellar evolution theory (then and now), the question that emerged was: how can ancient STAR CLUSTERS, devoid of star-forming gas, construct new stars billions of years after STAR FORMATION in the cluster ceased?

Ages of stars in clusters

The ages of star clusters containing hundreds to millions of members can be gauged with an accuracy approaching 15%. This is because individual stars change two of their observable properties, color and brightness, in a very predictable fashion (correlated with age) as they finish fusing their core hydrogen into helium. Such *turnoff* stars rapidly become redder and brighter as they swell in size, eventually ballooning to become red giants. Determining the age of a star cluster is thus quite simple: it's just the age of the most luminous star in the cluster which is still at the turnoff point on the hydrogen-fusing main sequence. Comparing the intrinsic brightnesses and colors of computer models of turnoff stars with direct measurements of those same quantities for real cluster stars directly yields stellar and cluster ages. This methodology leads to an age of about 13 billion years for the globular cluster M3 (and all the other Galactic globular clusters); and an upper limit of about two billion years for the brightest cluster BSS.

BSS in color–magnitude diagrams

Catalogs of BSS in clusters now contain over a thousand members. It is clear that BSS do not simply lie along the extensions of cluster main sequences, but are also found considerably redward of the zero-age main sequence. This is an important observational constraint that must be explained by any viable BSS theory.

Location

Prior to the launch of the Hubble Space Telescope, BSS were known to exist in only a handful of globular clusters. Every globular cluster whose core has been searched by Hubble in the past decade displays tens to hundreds of BSS (figure 2).

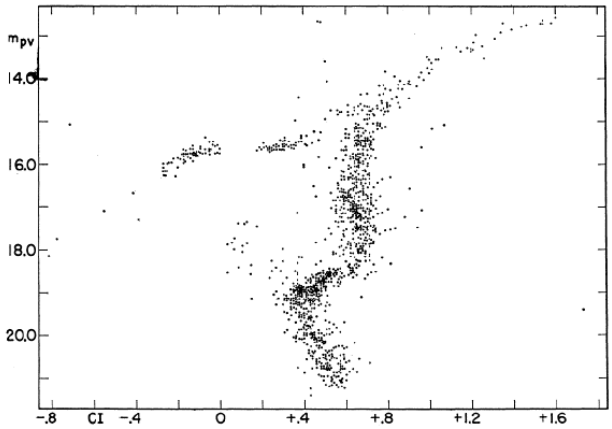


Figure 1. Reproduction of the color–magnitude diagram of M3 presented by A Sandage in 1953 *Astron. J.* **58** 61. The blue stragglers are the stars near $0 < \mathrm{CI} < 0.3$ and $17 < m < 19$.

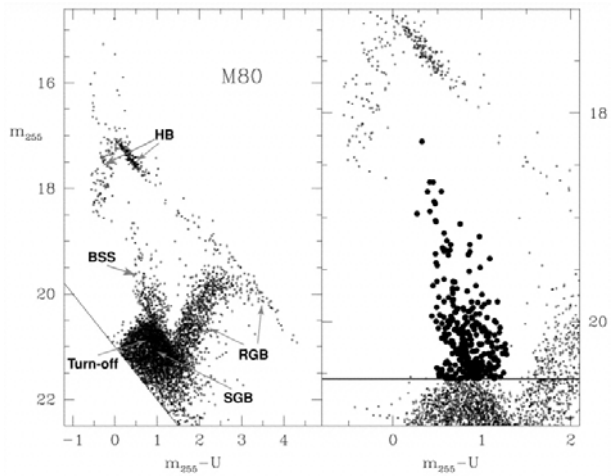


Figure 2. Color–magnitude diagram of M80 from Ferraro *et al* 1999 *Astrophys. J.* **522** 983. 'BSS' marks the blue stragglers in the left hand panel—they are the heavy dots in the magnified section on the right.

These stars are invariably more strongly centrally concentrated than cluster RED GIANTS or HORIZONTAL-BRANCH STARS (figure 3), while being virtually absent from the peripheries of almost all clusters (M3 is a notable exception).

The ubiquitous central concentration of BSS in globular clusters is probably due to equipartition of energy amongst the cluster stars. In this process the kinetic energy of all stars ($\frac{1}{2}Mv^2$) is equalized by gravitational interactions between stars in the cluster.

The more massive stars (larger M) thus acquire smaller velocities v and tend to *sink* towards cluster centers. The centrally concentrated BSS are thus expected to be significantly more massive than any other population

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

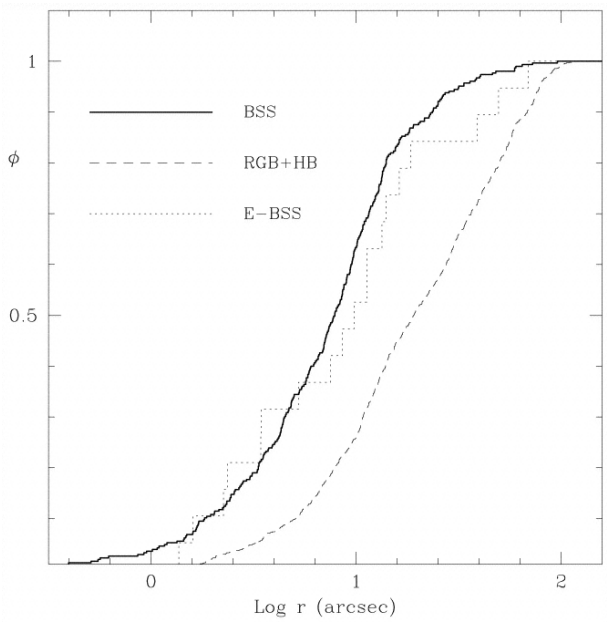


Figure 3. Cumulative radial distribution of BSS (heavy solid line) and horizontal branch and red giant branch stars (dashed line). As presented in Ferraro *et al* 1999 *Astrophys. J.* **522** 983.

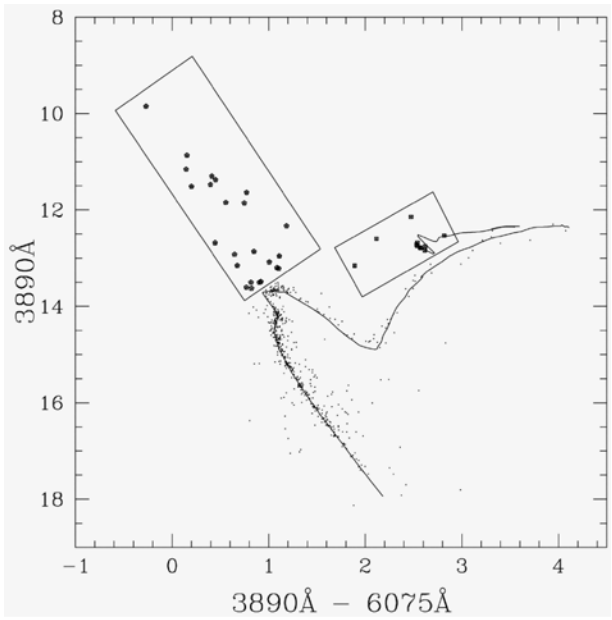


Figure 4. Color–magnitude diagram of M67. BSS are large dots in the box on the left. As presented in Deng *et al* 1999 *Astrophys. J.* **524** 824.

of cluster stars.

This is consistent with BSS's relatively high luminosities and blue colors. Strong observational confirmation has come from direct spectrographic measurement of the radius and surface gravity of the second-brightest BSS in the globular cluster 47 Tucanae. This measurement yielded a mass 1.7 times that of the Sun, or twice the mass of stars at the cluster turnoff. Other mass estimates (from an eclipsing system and from pulsation masses) are consistent with masses greater than the turnoff, but are not yet precise enough to be considered definitive.

BSS are also commonly observed in open clusters (figure 4), in the field as early-type high velocity stars, in the dwarf spheroidal galaxy companions to the Milky Way (figure 5) and in globular clusters of the Magellanic Clouds (figure 6).

Rotation

If BSS are produced by stellar mergers (including off-center collisions), they should be rapidly rotating, at least initially. Observations of a handful of BSS in old open clusters have shown most to be moderate rotators (typically 50 km s^{-1}). It has been suggested that newly formed BSS might be swollen, with large convective envelopes able to carry away angular momentum. While this may occur, much larger observational samples are needed before the oft-quoted observational result of *slow* cluster BSS rotation should be taken as seriously as it is now. In contrast, the direct measurement of $v \sin i = 155$ km s^{-1} for the second brightest BSS in the core of 47 Tuc emphasizes that at least some cluster BSS are rapid rotators.

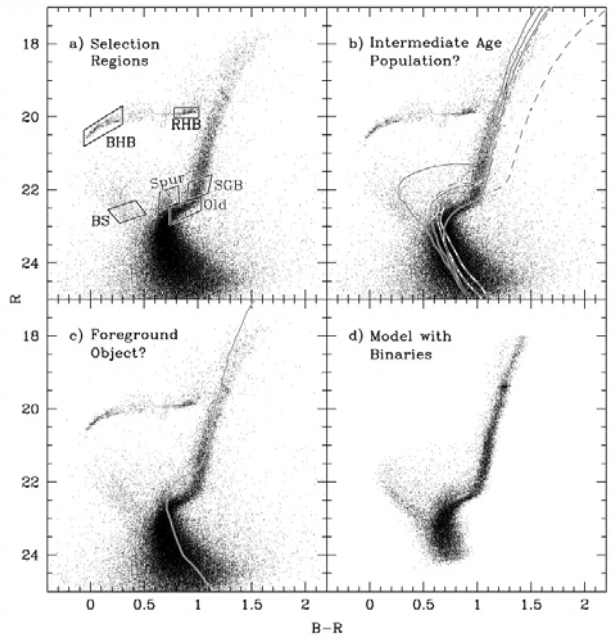


Figure 5. Color–magnitude diagrams of the inner and outer regions of the Sculptor dwarf spheroidal galaxy showing the presence of BSS. As presented in Hurley-Keller *et al* 1999 *Astrophys. J. Lett.* **523** L25.

Lithium abundances

Lithium is a fragile nucleus, destroyed by proton capture at temperatures exceeding 2 million degrees. Large-scale mixing destroys surface lithium, so its presence (or

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

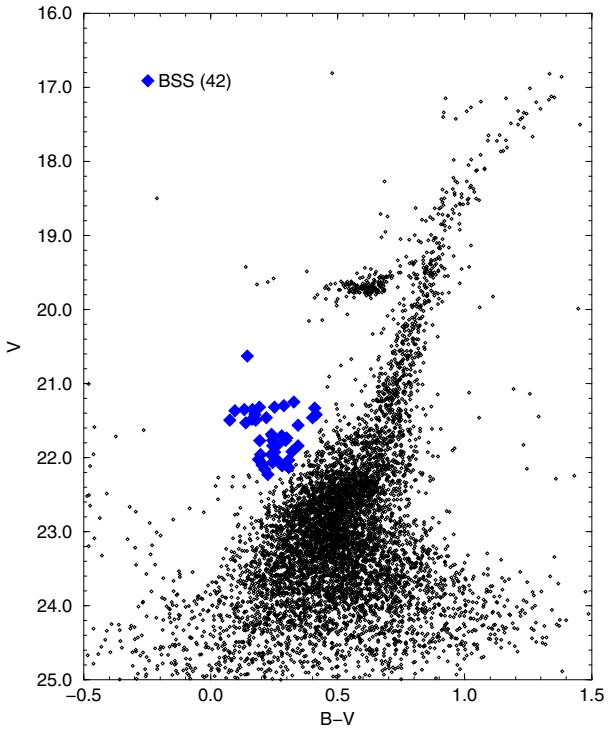


Figure 6. Color–magnitude diagram of the Small Magellanic Cloud cluster NGC 121. BSS are marked with the larger symbols. As presented in Shara *et al* 1998 *Astrophys. J.* **508** 570.

absence) in the spectrum of a star is a sensitive indicator of the temperature history of surface material.

Seven blue stragglers in the old open cluster M67 have been found to be depleted of lithium with respect to normal field main sequence stars of similar spectral types by at least a factor of 30. This is consistent with extensive mixing of internal material to the stellar surface during the production of those cluster BSS. A number of high velocity, but Population I abundance stars were also found to have low Li abundance, suggesting that they, too, are BSS.

Pulsating BSS

The low luminosity extension of the Cepheid instability strip overlaps the location of BSS in globular clusters. These stars obey period–luminosity–color relations analogous to those of classical Cepheids. The mass of a pulsating variable can be determined from the relation

$$M \sim \frac{Q^2 L^{1.5}}{P^2 T^6}$$

where Q is the star's pulsation constant (for a given mode), P is the pulsation period for that mode, L is the star's luminosity and T is its effective temperature. There is, unfortunately, no way known to unambiguously determine the pulsation mode for most pulsating variables. The strong sensitivity to adopted values of luminosity, and especially to effective temperature, can lead to significant uncertainties in the calculated masses. This can be somewhat improved by calculating the mass ratios of pulsating BSS to those of RR LYRAE STARS in the same clusters. Uncertainties in the cluster distances remain, but the resulting BSS masses are typically twice those of RR Lyrae stars. It is something of a puzzle that many BSS in the instability strip do not pulsate. Rapid rotation inhibits pulsation, so non-pulsating BSS may be rapid rotators.

BSS in contact binaries

Contact binaries (CB) have been suggested as the progenitors of at least some BSS. CB are, by far, the most common type of short-period eclipsing systems in the solar neighborhood. The components of these short-period binaries are in physical contact with their Roche equipotential lobes. The components are also in physical contact near the inner Lagrangian point located between the stars.

CB are the most common ECLIPSING BINARIES in open and globular clusters, where the advent of CCD detectors has led to the discovery of significant numbers of these stars. Surveys are still far from complete because the discovery probability plummets for systems with orbital periods longer than a day or two.

If CB give rise to BSS then one would expect the ratio of the number of BSS that are CB to the total number of BSS in a cluster to simply equal the ratio of the CB and BSS lifetimes. Unfortunately the lifetimes of CB are not well determined, so this elegantly simple comparison cannot be made with confidence.

This brief summary of the observed properties of BSS forms a set of constraints on BSS theories. Models that seek to explain the origin and nature of BSS must explain the range of environments in which they are found, their high masses, low lithium abundances, and (possibly) high rotation rates.

Blue straggler models and predictions

Ten models (and at least as many variants) have been proposed to explain the origin of blue stragglers; six involve single stars while four can operate only in binaries. Only the binary mechanisms (see below) are now considered viable. These have been summarized by Mario Livio as:

S1 Multiple bursts of star formation

S2 Prolonged lifetime due to internal mixing

S3 Main sequence stars managing to avoid mass loss

S4 Stars that accreted gas from the ISM or the intracluster medium

S5 Evolved stars that returned to the main sequence

S6 Young field stars captured by clusters

B1 Mass increase as a result of mass transfer in a close binary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

B2 Coalesced contact binaries

B3 Coalesced common envelope binaries

B4 Mergers as a result of physical stellar collisions.

Observations (especially the low lithium abundance) and theory are able to rule out the single-star mechanisms with a high degree of confidence. The binary mechanisms are all much harder to discount, and workers in the field now suggest that all four scenarios may be at work, though at different rates in different places. The thorniest problems facing the binary hypotheses (B1) and (B2) are as follows.

B1. If mass transfer in binaries produces all BSS then all BSS should be in binaries. Radial velocity studies show that while some BSS are in binaries, intensive monitoring over periods of years conclusively shows that others are not in binaries close enough to lead to mass transfer.

B2. Coalescence of a contact binary to form a BSS is potentially problematic because of the existence of BSS with masses larger than twice the host cluster turnoff mass. This might be a preferred formation route in low density clusters where no other viable mechanism (collisions in particular) works.

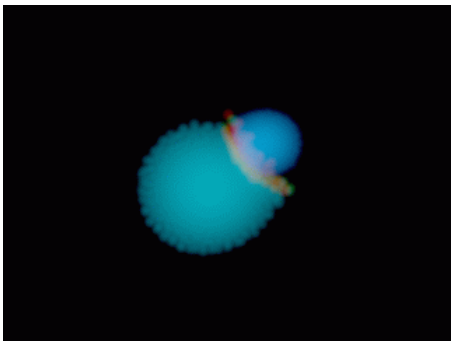
Common envelope evolution (**B3**) might be able to generate the most massive BSS by permitting the swallowing of a more distant companion or companions during the expanded envelope phase.

Long thought to be so rare that they could be ignored, physical collisions (**B4**) (and very close passages) are now known to be commonplace and very important in dense star clusters. Hundreds of physical collisions must occur in the cores of most globular clusters during a Hubble time. This is mostly due to the effect of gravitational focusing. Stars move with velocities (v) of at most 10–20 km s^{-1} in globular clusters, much less than the escape speed of a main sequence star. The effective cross section σ_{ss} of two cluster stars of masses m_1 and m_2 and radii r_1 and r_2 is

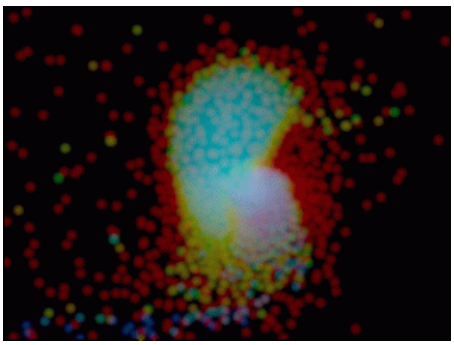
$$\sigma_{ss} = \pi(r_1 + r_2)^2 \left(1 + \frac{2G(m_1 + m_2)}{(r_1 + r_2)v^2}\right)$$

which can be 10 000 times larger than the geometrical cross sections because of the slow passages of two stars by each other. This magnification is even more remarkable for binaries. The effective cross section of a binary passing close to a single star (or another binary) is approximately the square of the semimajor axis; this can be millions of times the stars' geometrical cross sections.

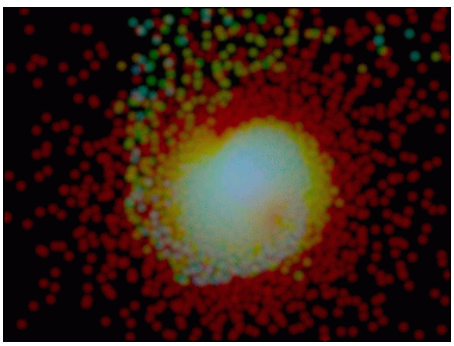
Mutually induced gravitational distortions can lead to dissipation of orbital energy and capture of an initially parabolic pair of single or binary stars into a highly elliptical orbit. Subsequent periastron passages can dissipate more energy, leading eventually to very close binaries and/or physical collisions.



(a)



(b)



(c)

Figure 7. Example of a simulation of an off-axis collision between two main sequence stars of different mass. From J E Barnes' website http://www.ifa.hawaii.edu/faculty/barnes/stars.html

Physical collisions

Even in the cores of the most dense Galactic globular clusters the interval between physical stellar collisions must be at least 10^5 years (otherwise every star near the cluster center would already have experienced a collision). It is very unlikely that the tools needed to directly image a stellar collision (far outside the Galaxy) will be developed during the next two or three decades (though gravity wave detectors might detect them indirectly in just a few years). For the present, studies of stellar collisions are of necessity confined to numerical simulations on computers.

When main sequence stars undergoing a nearly head-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on collision are of comparable density and size the initially spherical masses increasingly overlap as they compress and distort each other into nearly half-moon shapes. Temperatures and densities never become high enough to ignite significant thermonuclear burning outside the stellar cores. A few per cent of the total mass of two colliding main sequence stars is ejected perpendicular to the direction of stellar motion, but then mixing and overlap take over. Within a few dynamical times (hours) the two incoming stars are merged together into a single, new star almost as massive as the sum of its progenitors' masses.

Real collisions will almost always be off-center, involving rotating stars of unequal mass. The predicted degree of mixing that ensues in any collision is quite dependent on the density profiles, impact parameter and initial relative velocities of the colliding stars, but at least some surface mixing seems hard to avoid. Very rapidly spinning stars surrounded by massive disks have been suggested by some simulations. Other models predict that the denser of two colliding stars' cores can penetrate to the core of the less dense (and more massive) star (see figure 7). This is the fundamental reason why the apparent age (from its position in a cluster color–magnitude diagram) of a coalesced star is predicted to *not* be synchronized with the ages of all the non-coalesced stars in a cluster core. Mixing of surface and core layers during a collision may produce a coalesced star having a higher central hydrogen content than either of the stars which formed it.

In effect, the core-hydrogen-rich coalesced star has had its clock reset by the collision. Despite the uncertainties and approximations used, the detailed numerical simulations of stellar coalescence through collisions give this process a credence that the other binary mechanisms are hard pressed to match.

Cluster models with collisions

Two very recent theoretical approaches have greatly increased the realism of coalescing star simulations. In one work, simple representations of dense star cluster dynamics and detailed binary–single star encounter simulations were combined to produce cross sections and rates for a variety of collisions. The collision outcomes were calculated with smoothed-particle hydrodynamics (SPH) simulations of realistic star models, then followed in detail with a stellar evolution code. Some of the predicted distributions in theoretical cluster color–magnitude diagrams mimic observed distributions with remarkable fidelity (figure 8).

The study tested the sensitivity of the predicted color–magnitude distributions to the stellar evolution tracks, binary mass function, binary period distribution and current mass function. The first two parameters are much more important than the latter two. Fully mixed stars produce BSS distributions that start on and always remain close to the Zero Age Main Sequence. Partially mixed configurations (in accord with the SPH simulations noted above) produce color–magnitude distributions that are more spread out in color than the fully mixed distributions, in much better accord with the observations.

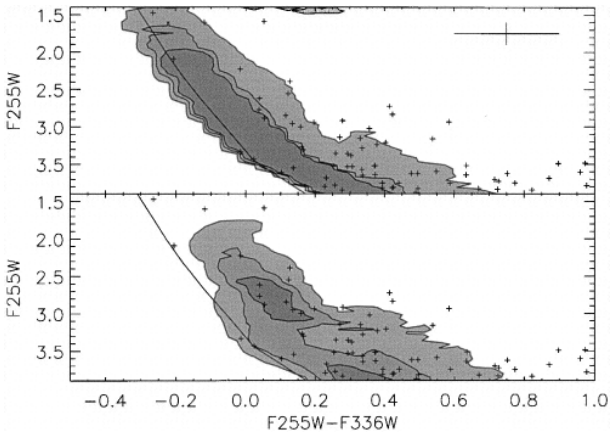


Figure 8. Theoretical color–magnitude diagrams comparing the effects of using different evolutionary tracks for the blue stragglers. As presented by Sills A and Bailyn C 1999 *Astrophys. J.* **513** 428.

Binary star cross sections are much larger than those of single stars, so binary–single collisions overwhelm single–single collisions for even a modest binary frequency. The future inclusion of binary–binary collisions and coalescences may explain the few, most luminous BSS not reasonably explained by the simulations.

A second approach uses detailed N-body simulations coupled with simple prescriptions for stellar evolution and physical collisions. The initial conditions were chosen to mimic R136, the massive and compact star cluster at the center of the 30 Doradus complex in the Large Magellanic Cloud. Simulations show that the dynamical evolution of the cluster is tightly linked to the evolution of the most massive stars. A single very massive star ($>100M_{\odot}$) is found to grow steadily through mergers with other stars in less than 3–4 Myr. This object is usually found in a binary system very close to the cluster center, and it is *rejuvenated* by each successive collision. It should appear as a BSS in the cluster color–magnitude diagram. Strong stellar winds slow the runaway process, but even large mass loss rates do not prevent the growth of very massive BSS.

The two approaches described above clearly must be merged to eventually produce a realistic scenario capable of being tested in detail against observed BSS in star clusters. Spurred largely by the Hubble Space Telescope, BSS research is now a rapidly developing field driven by the synergy between theory, simulations and observations.

Bibliography

Half a century ago the infant field of stellar evolution, coupled with state-of-the-art observations of globular clusters, was used to demonstrate the extraordinarily ancient nature of these objects. Allan Sandage was then a PhD student at Caltech, and his beautiful dataset on the cluster M3 clearly showed, for the first time, blue

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stragglers. This discovery paper (Sandage A R 1953 *Astron. J.* **58** 61) reads as well today as then.

A monograph entitled *Blue Stragglers* (ASP Conference Series 53: ed R Saffer 1992) contains a dozen excellent overviews of all aspects of the subject, as do two more recent reviews (Stryker L 1993 *Publ. Astron. Soc. Pacific* **105** 1081, Bailyn C 1994 *Science* **263** 44). A beautiful computer simulation movie of a stellar collision may be viewed online at http://www.ifa.hawaii.edu/faculty/barnes/stars.html

Recent sophisticated treatments of N-body simulations, including the effects of mergers, are reported in Portegies Zwart S *et al* 1999 *Astron. Astrophys.* **348** 117. Simulations of the detailed evolution of merged stars appear in Sills A and Bailyn C 1999 *Astrophys. J.* **513** 428. These complementary approaches provide the most convincing explanations of mergers and BSS in dense star clusters to date.

Mike Shara

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Boötes

(the Herdsman; abbrev. Boo, gen. Boötis; area 907 sq. deg.) A northern constellation that lies between Canes Venatici and Serpens Caput, and culminates at midnight in late April. The origin of the constellation figure is uncertain, though it dates back at least to ancient Greece, where it was identified with Arcas, who was conceived when his mother, Callisto, fell victim to Zeus's passion and whom Zeus placed alongside Callisto (who is represented by the neighboring constellation of Ursa Major) in the sky. The name Boötes may come from a Greek word meaning 'clamorous' or 'noisy', possibly from the herdsman's shouts to his dogs (now represented by the neighboring constellation of Canes Venatici). In some early celestial charts Boötes is shown in pursuit of a bear (the constellation of Ursa Major). The brightest stars of Boötes were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large constellation, Boötes is easily found by virtue of α Boötis (Arcturus), which at magnitude -0.1 is the fourth brightest star in the sky and lies on a continuation of the line joining the stars ε, ζ and η Ursa Majoris, in the tail of the Great Bear. Other bright stars include ε Boötis (Izar or Pulcherrima), a close binary with yellow (K0) and blue (A2) components, magnitudes 2.7 and 4.8, separation $3''$, η Boötis (Muphrid), magnitude 2.7, and γ Boötis (Seginus), magnitude 3.1. There are five other stars of magnitude 4.0 or brighter. Another interesting binary is ξ Boötis, which has yellow (G8) and orange (K4) components, magnitudes 4.8 and 7.1, separation $7''$.

There are no bright star clusters, nebulae or galaxies in Boötes.

The Quadrantid meteor shower appears to radiate from the northern part of Boötes, which contains the former constellation of Quadrans Muralis, after which it is named.

See also: Arcturus, Quadrantids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bode, Johann Elert (1747–1826)

Astronomer, born in Hamburg, Germany, director of the Berlin Observatory. He drew attention to a relation among the distances of the planets from the Sun that had been discovered by JOHANN TITIUS. The law is sometimes known as the *Titius–Bode law*, but usually just as *Bode's law*. The law starts with the sequence 0, 3, 6, 12, 24. . . , adds 4 and divides by 10 to obtain 0.4, 0.7, 1.0, 1.6. 2.8. . . , which is a good approximation for the distances from the Sun of all the planets in units of the astronomical unit. The law has entries for the then undiscovered asteroids and Uranus, but fails at Neptune. The explanation for the law (if there is one) remains unknown. Bode discovered M81 (known as Bode's nebula), M82, M53, M92 and M64, and published in 1779 *A Complete Catalogue of Hitherto Observed Nebulous Stars and Star Clusters*, which, unfortunately, contains in its 75 entries many non-existent objects and uninteresting asterisms.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bode's Law

A numerical sequence which is roughly in proportion to the distances of the major planets from the Sun. It is named after Johann Bode, who announced it in 1772; it is more properly referred to as the *Titius–Bode law*, as it was first put forward by Johann Titius six years previously. The sequence is formed by taking 0 and 3, doubling 3 to give 6, doubling 6 to give 12, and so on, then adding 4 to each term to give 4, 7, 10, 16, . . ., and finally dividing each term by 10. The formula $\frac{3}{10}(2^n - 1) + 4$, $n = 0, 1, 2, \ldots$ yields each term. The correspondence with the distances of the planets (to the nearest 0.1 AU) is given in the table.

Planet	Bode's law distance (AU)	Actual distance (AU)
Mercury	0.4	0.4
Venus	0.7	0.7
Earth	1.0	1.0
Mars	1.6	1.5
Ceres	2.8	2.8
Jupiter	5.2	5.2
Saturn	10.0	9.5
Uranus	19.6	19.2
Neptune	38.8	30.1
Pluto	77.2	39.5

Bode made his announcement before the asteroids or the planets beyond Saturn had been discovered. The relationship remained something of a curiosity until the discovery of Uranus in 1781. Once the new planet's distance from the Sun was established and seen to fit the numerical pattern, astronomers considered more seriously the apparent gap between Mars and Jupiter, and were prompted to begin searching for a missing planet. The first asteroid to be discovered, Ceres, in 1801, was by no means a major planet, but it did orbit at the predicted distance from the Sun. This predictive power gave the Titius–Bode relationship the status of a 'law', though it was transgressed by the subsequent discoveries of Neptune and Pluto at distances that broke the pattern.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bok, Bart Jan (1906–83)

Dutch astronomer, worked at Harvard and directed the Mount Stromlo Observatory (near Canberra, Australia). Studied the structure and evolution of star clusters and the Galaxy, mapping the spiral arms of the Milky Way, especially the Carina region, and the Magellanic Clouds. Bok realized that the sky quality at Mount Stromlo was deteriorating with the growth of Canberra and chose Siding Spring Mountain near Coonabarabran as the new station for the Australia's main optical observatory, eventually the site for the Anglo-Australian 150 in telescope. Bok initiated radioastronomy at Harvard and promoted it elsewhere. His investigations of interstellar gas and dust led to studies of star formation, and he became known for his work on small dark nebulae now called *Bok globules*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bolide

A bright meteor—a fireball—that produces audible sounds. Sonic effects are generated once the incoming meteoroid has penetrated to an altitude of around 30 km, and so will be heard some minutes after the fireball is seen. An initial sonic boom indicates that the meteoroid has exploded. If large enough, the resulting fragments will create their own sonic booms, and an object that undergoes multiple fragmentation may generate a continuous, thunder-like rumbling. The fragments of a meteoroid that produces a bolide may well survive their atmospheric passage to fall as meteorites. The Allende meteorite, which underwent multiple fragmentation, was preceded by a bolide.

See also: Allende meteorite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bond, William Cranch (1789–1859) and Bond, George Phillips (1825–65)

Instrument-maker and astronomer, born in Portland, ME. Massachusetts businessman, first (unpaid) director of the Harvard College Observatory, with its magnificent 15 in telescope. He identified Saturn's innermost Crêpe Ring. Independently of William Lassell, he discovered the seventh satellite of Saturn, Hyperion, with his son George Bond (the observatory's second director). George correctly realized that Saturn's rings could not be solid, as shown by JAMES CLERK MAXWELL.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bondi, Hermann (1919–)

Mathematical physicist and astronomer, born in Vienna, Austria, made his career in Britain as a mathematician. Best known as one of the originators of the steady-state theory of the universe, director-general of the European Space Research Organization, and as a scientific administrator.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bonestell, Chesley (1888–1986)

American architect (Golden Gate Bridge, and the Chrysler Tower) and astronomical illustrator, including for such motion pictures as *War of the Worlds*, *When Worlds Collide*, and *Destination Moon*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

BOOMERANG

This is an acronym for 'Balloon Observations of Millimetric Extragalactic Radiation and Geophysics', an experiment which produced an image of the farthest region of the universe observable today. This region is about 15 billion light-years away, and the structures visible in the image existed about 15 billion yr ago. They represent the first stage of the process of formation of the large-scale structures in our universe. They are visible because their presence slightly modifies the otherwise isotropic COSMIC MICROWAVE BACKGROUND (CMB). This radiation, produced in the very early universe, is observable today in the microwaves as a blackbody with a temperature of 2.73 K. The photons of the CMB, generated soon after the big bang (see BIG BANG THEORY), were continuously scattered by free electrons until the expanding universe cooled below 3000 K.

At that point, called 'recombination', about 300000 yr after the big bang and about 15 billion yr ago, the universe became neutral (hydrogen atoms formed from free electrons and protons) and transparent. Since then, the CMB photons have been free to propagate without interacting with matter anymore. They thus arrive at our telescopes carrying the imprint of all the protostructures present at the recombination epoch. Very sensitive telescopes are needed to detect the small (10–100 parts per million, ppm) perturbations in the CMB, usually called 'anisotropies', produced by these protostructures. The BOOMERANG experiment (figure 1) was flown by means of a stratospheric balloon, at an altitude of 37 km, in order to observe above the bulk of the Earth's microwave-absorbing atmosphere. It was launched from Antarctica, and remained at float from 29 December 1998 to 8 January 1999. It observed about 4% of the sky, in a high Galactic latitude region centered at RA=7.5h, dec=-45°. The experiment had 16 independent measurement channels sensitive to four wavelengths: 3.0, 2.0, 0.12 and 0.73 mm.

The detectors were 'spider web' semiconductor bolometers, cooled by a ^{3}He cryogenic system at 0.3 K to reduce their noise. The experiment used an off-axis telescope with an aluminum primary mirror, 1.3 m in diameter. The angular resolution obtained in this way was 18, 10, 14 and 13 arcmin respectively for the four bands. A subsection of the map produced at the wavelength of 2 mm is visible in figure 2. The false colors (from black to yellow) represent anisotropies of the brightness of the CMB (from –100 ppm to + 100 ppm respectively). The spots visible in the image have a typical size of 1°. This characteristic size of the fluctuations corresponds to the dimensions of the sound horizon (i.e. the distance sound waves have travelled in the ionized universe since the big bang) at the recombination. In fact, only perturbations smaller than the sound horizon at that epoch can oscillate like sound waves, thus producing significant anisotropies in the CMB.

The size of the horizon can be computed with good accuracy and compared with the angular dimension observed in the image, thus carrying out a sensitive 'angular-size versus distance' test. From the comparison it is inferred that the curvature of the universe is close to zero. According to the general relativistic cosmological models (see GENERAL THEORY OF RELATIVITY and COSMOLOGICAL MODEL), the curvature of the universe is determined by its average mass–energy density. A density larger than the critical density produces a positive curvature, a density equal to the critical one produces a flat (zero curvature) Euclidean geometry, while a density smaller that the critical one produces a negative curvature. If the curvature

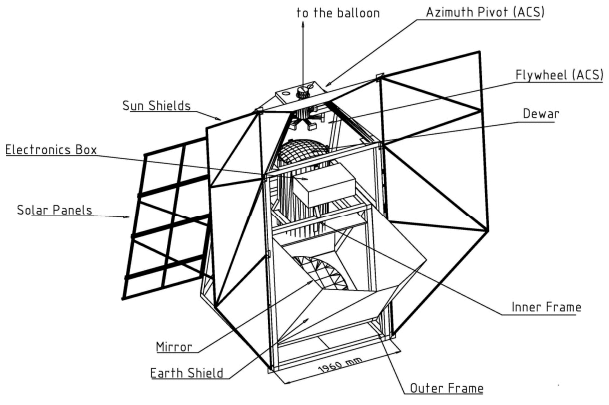


Figure 1. The BOOMERANG payload is a remotely operated microwave telescope. Its mass is about 1.5 tons and it is lifted to 37 km of altitude by a 1 million m^3 helium-filled stratospheric balloon.

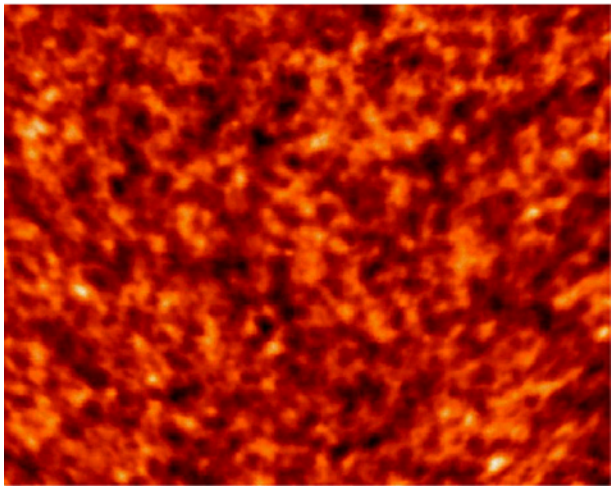


Figure 2. Map of a remote region of the universe obtained by the BOOMERANG experiment at 2 mm of wavelength. The size of this map is about 40° x 50°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

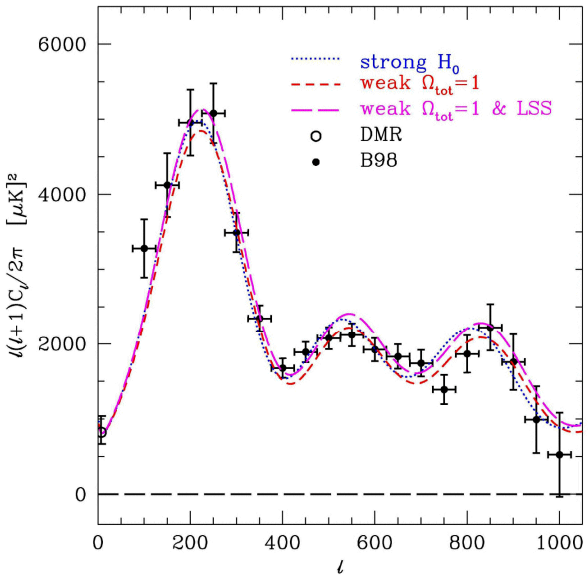


Figure 3. Angular power spectrum of the CMB measured by the BOOMERANG experiment. The quantity plotted is the intensity of the CMB temperature fluctuations for different angular scales, identified by the multipole number. A multipole 200 corresponds to an angular scale of abot 1°. A multipole 1000 corresponds to an angular scale of about 0.2°. The crosses represent the measured data with their statistical errors; the lines represent different adiabatic inflationary models fitting the data.

was positive, the spots would appear larger than observed, while if it was negative the spots would appear smaller than observed. Measuring a curvature close to zero, BOOMERANG has demonstrated that the mass–energy density of the universe is close to the critical density ($\Omega = 0.98 \pm 0.05$). The details of the size distribution of the hot and cold spots in the image (i.e. the angular power spectrum of the CMB; see figure 3) have been compared with accurate theory predictions in the framework of the adiabatic inflationary cosmology with a Hubble parameter h=0.71±0.08. In this way the values of several other cosmological parameters, such as the physical density of baryons ($\Omega_b h^2$), the tilt of the power spectrum of primordial density perturbations (n) and the energy in the form of the cosmological constant (Ω_Λ) have been measured: $\Omega_b h^2 = 0.021 \pm 0.004$; $n=0.94 \pm 0.09$, $\Omega_\Lambda=0.62 \pm 0.15$.

For further information see http://www.physics.ucsb.edu/~boomerang/

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Borelli, Giovanni Alfonso (1608–79)

Italian mathematician and physiologist, born in Naples, taught at Pisa and Messina. Suggested that comets travel in parabolic orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Boscovich, Ruggero Giuseppe (1711–87)

Mathematician, born in Ragusa, Dalmatia, Venetian territory (now Dubrovnik, Croatia), professor at the Collegium Romanum in Rome and Pavia, and director of the Brera Observatory. Using NEWTON's gravitational theories, Boscovich was the first to give a procedure to compute a planet's orbit from three observations of its position, a solution that was elaborated by CARL GAUSS, LEGENDRE and other mathematicians. He studied optics, astronomy, gravitation, meteorology and trigonometry. He wrote in 1753 a crisp and elegant little book on reasons to think that the Moon has no atmosphere, and he led an expedition to California in 1769 to observe the transit of Venus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bouguer, Pierre (1698–1758)

Mathematician and physicist, born in Le Croisic, France, he was the first to measure the density of the Earth using the deflection of a plumb line due to the attraction of a mountain in Peru, as more successfully done by the Astronomer Royal, NEVIL MASKELYNE.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Boulliau, Ismael (1605–94)

Mathematician, librarian, priest, born in Loudun, France. In 1645 he published *Astronomia Philolaica* which, on the basis of elliptical orbits for planets and Kepler's laws, suggested that the force which attracted planets to the Sun should vary inversely as the square of the distance (JOHANNES KEPLER had claimed inversely as the first power). This was the most significant advance in the theory of gravitation between Kepler and ISAAC NEWTON. It was acknowledged by Newton in his *Principia*. (*See also* ROBERT HOOKE.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bouvard, Alexis (1767–1843)

French astronomer, born a shepherd boy in Chamonix, became assistant to Laplace and eventually director of the Paris Observatory. Attempted to produce an accurate orbit of Uranus, calculating the perturbations of other planets according to corrected tables of planetary positions by JEAN DELAMBRE, published in 1792. Bouvard still could not make all the observations fit. He published his new orbit of Uranus in 1821 but noted: 'I leave to the future the task of discovering whether the difficulty of reconciling [the data] is connected with the ancient observations, or whether it depends on some foreign and unperceived cause which may have been acting upon the planet'. URBAIN LEVERRIER and JOHN ADAMS showed that the latter, Neptune, was the reason.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bowditch, Nathaniel (1773–1838)

Insurance actuary, astronomer, mathematician, born in Salem, MA. Self-taught, by age 15 he had compiled an astronomical almanac. Based on practical experience at sea, he wrote the *New American Practical Navigator*; published papers on comets and meteors, and translated PIERRE LAPLACE's *Mécanique Céleste*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bowen, Edward George ['Taffy'] (1911–91)

Radioastronomer, born in Cockett near Swansea, Wales. After war-time research on the development of radar, Bowen went to Australia where he made rain through cloud-seeding, discovered still-disputed correlations between the dates of meteor showers and the lunar phase and rainfall (which he attributed to lunar modulation of meteoritic material) and became a radioastronomer. Inspired by the Jodrell Bank reflector, he organized the construction of the Parkes Radio Telescope to design principles by Barnes Wallis (later Sir Barnes Wallis), the famous airship and aircraft designer, and played the key role in organizing the construction of the Anglo-Australian Telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bowen, Ira Sprague ['Ike'] (1898–1973)

American astronomer, worked with ROBERT A MILLIKAN, at the California Institute of Technology. In 1927 his investigation of the ultraviolet spectra of highly ionized atoms led to his identification of mysterious spectral lines of gaseous nebulae, not as an unknown element prematurely called *nebulium*, but as forbidden lines of ionized oxygen and nitrogen. Director of the Mount Wilson and Palomar Observatories, oversaw completion of the 200 in Hale and 48 in Schmidt telescopes, designing many of their instruments, including a novel spectrograph.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Boyle, Willard S (1924–)

Canadian physicist, born in Amherst, Nova Scotia, co-inventor of the charge coupled device (CCD), the heart of all modern instrumentation for optical telescopes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Boys, Charles Vernon (1855–1944)

Physicist born in Wing, Rutland, England, invented the radio-micrometer, to measure heat radiation from the Moon and planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bradley, James (1693–1762)

Astronomer, born in Sherborne, Gloucestershire, England. From 1742 third Astronomer Royal. With SAMUEL MOLYNEUX discovered the 'aberration of light' (a very large and unsuspected apparent motion of the stars that was, up to that time, an uncontrolled error in star observations). The explanation for the discovery (the addition of the Earth's orbital velocity to the velocity of the light incoming from the star being measured) provided the first observational proof of the Copernican hypothesis that the Earth moved round the Sun, and showed that the speed of light was a constant of physics. The discovery showed as a by-product that the parallax of stars is very small, i.e. their distances are immense—in the case of Gamma Draconis, the star measured, more than 400 000 times the distance of the Sun. Bradley also discovered that the inclination of the Earth's axis to the ecliptic is not constant ('nutation').

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brahmagupta (c 598–c 665)

Hindu astronomer born (possibly) in Ujjain, India, where he headed the observatory, applied advanced (for the time) algebraic methods to astronomical problem such as solar and lunar eclipses, planetary conjunctions and positions of the planets, and wrote *Brahma-sphuta-siddhanta (The Opening of the Universe)* and *Khandakhadyaka*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brashear, John Alfred (1840–1920)

Telescope maker, born in Brownsville, PA. His company became the Contraves Corp. Brashear's remains and those of his wife are interned beneath the Keeler Telescope at Allegheny Observatory in Pittsburgh (which he directed), under the words 'We have loved the stars too fondly to be fearful of the night'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brattle, Thomas (1658–1713)

Prominent Boston merchant, tutor at Harvard University, became its treasurer. Observed the comet of 1680 with Harvard's first telescope of 3.5 feet focal length—the observations were mentioned by ISAAC NEWTON in *Principia*. One of the first astronomers in America, Brattle's unique claim to a place in history is that he persuaded colonial Governor William Phips to order that reliance on spectral and intangible evidence no longer be allowed in legal processes, thus suspending the witchcraft trials.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Braun, Wernher von (1912–77)

Rocket scientist, born in Wirsitz, Germany, enthusiast of space travel from an early age, designer of the V-2 rocket used during the second World War. With 120 associates from Peenemunde, went to the United States and directed firings of captured V-2 rockets at White Sands Missile Range in New Mexico, following which he used a Jupiter-C (Juno I) to put America's first satellite, Explorer 1, into orbit in 1958. Von Braun and his team transferred to NASA when it was formed in that year, and he became Director of the Marshall Center, where he developed the Saturn V, used to put a man on the moon in 1969.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Breccia

A type of rock composed of small, shattered and crushed (and sometimes melted) angular fragments cemented together in a finer-grained material. Breccias are produced by impacts. The shock from the impact of even a meter-sized body on a planetary surface may be sufficient to compact and weld surface material into breccias. Lunar highland rock sampled by the crews of Apollo missions proved to consist of breccias, which are expected to make up the surfaces of all rocky bodies which have not been subsequently modified by volcanic or other geological activity. The fact that many chondritic and achondritic meteorites are breccias shows that they were once part of larger bodies that suffered impacts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Briggs, Henry (1561–1630)

Mathematician, born in Worley Wood, Yorkshire, England. Professor at Cambridge, London (Gresham College) and Oxford. By pointing out how much more useful logarithms would be if (unlike JOHN NAPIER's original invention) they were to base 10, he was responsible for improving astronomers' ability to calculate. Published mathematical tables (logarithms, sines, tangents, secants) as an aid to such calculations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bright Quasar 3C 273

QUASARS are the most luminous objects we know in the universe. 3C 273 is among the brightest quasars; it also happens to be one of the closest to the Earth. It is therefore bright in the night sky and hence has been the object of intense studies since its discovery in 1963.

Quasars are found in the center of some galaxies (collections of some hundred billion stars; the MILKY WAY is the galaxy in which our solar system is embedded). Quasars are very peculiar celestial objects, they outshine the galaxy that hosts them, although they are barely as large as our solar system (which is extremely small compared with the galaxy). They emit electromagnetic radiation that covers all the bands accessible to astronomical observations on the Earth and in orbit, from radio waves to gamma rays. Their energy radiation can vary in an important manner on time scales that may be as short as hundreds of seconds. The matter close to the most central region of quasars moves at speeds of several thousand kilometers per second. These properties are most likely due to the presence in the center of quasars of a black hole that may be as massive as many tens of millions of suns. This paradigm is the only one that gives some hope of understanding the general properties of these objects.

The discovery

Quasars were discovered in 1963 when it became possible to localize with precision RADIO SOURCES in the sky. 3C 273 played a major role in the discovery of quasars. 3C 273 is a bright radio source, it bears the number 273 in the third Cambridge catalog of radio sources (hence the name). It was the second source that was identified with a very peculiar 'star', a source of light that was not resolved as extended on photographic plates. The 'star', or better the source of optical light that was coincident with the radio source 3C 273, was soon seen to have peculiar emission lines. These are the signature of the presence of atoms that are excited and emit light at well defined wavelengths. Stars normally don't have emission lines, and more surprisingly it was at first impossible to recognize the type of atoms that were responsible for the emission lines in 3C 273, as the wavelengths at which the lines were observed were not among those emitted by known atoms. These very peculiar properties of the light source led to the name quasi stellar radio source that then became QSO for quasi stellar object or quasar for quasi star.

The mystery of the emission lines was solved when it was noted that one could indeed recognize the wavelength of the emission lines if one admitted that the lines are observed not at the wavelength λ_0 typical of laboratory measurements of the atoms but at longer wavelengths $\lambda = \lambda_0(1 + z)$, where z is known as the redshift, equal to 0.158 for 3C 273. Many galaxies showing the same effect were already known in 1963, but with redshifts much less than that of 3C 273. In the universe this effect is a sign that the objects are moving apart from one another (expansion) and thus are at very large distances. At a redshift of 0.158 3C 273 is thus found to be some 3 billion light years away. In 1963 3C 273 was the most distant object known (see also HIGH-REDSHIFT QUASARS).

Being so far away and still easy to find on photographic plates, 3C 273 is an object of 13th magnitude; 3C 273 is clearly a very luminous object indeed. Its light output corresponds to 10^{14} times that of the Sun.

Most quasars do not radiate substantially in the radio part of the electromagnetic spectrum, and are called radio quiet quasars. A small fraction, of which 3C 273 is a member, do, however. These quasars are called radio bright quasars. They can be discovered as described above using radio data. Radio quiet objects are found by looking for redshifted emission lines in the light of 'stars'.

It was also soon discovered that 3C 273 varies in intensity on time scales much less than 1 year. Since light sources cannot vary on times scales much less than the time the light needs to cross the source (this is due to a causality argument: in order to significantly change the brightness of an object a signal must reach all the parts of the object to convey the instruction), it was also very quickly clear that 3C 273 must be much smaller than 1 light year, a distance already much less than that separating us from the closest star in our Galaxy.

The very unusual properties of 3C 273 spurred a large set of observations using not only optical and radio telescopes but also other instruments, both on the ground and on Earth orbiting satellites. It became clear that this (and other) quasar was the source not only of visual light and radio waves, but also of infrared radiation, ultraviolet light, x-rays and gamma rays.

Deep radio, optical and x-ray images in subsequent years showed that the object is not really 'star-like' but that a cigar shaped extension is associated with the main source (see figure 1 of QUASISTELLAR OBJECTS: OVERVIEW). This 'jet' extends from very small distances to some 20″ from the 'core'. It is most prominent in the radio images. Very surprisingly, the features making up the jet (often called blobs, thus expressing our lack of understanding of these features) were seen to travel away from the core of the object at speeds seemingly larger than the speed of light (superluminal motion). This superluminal motion is in fact due to an aberration that takes place when objects are moving in a direction close to the line of sight at velocities close to (but not larger than) the speed of light. This showed that in addition to the already very unusual observed properties of quasars one had to add the presence of radio wave emitting matter moving at very high speeds along a well defined direction, close to the line of sight.

The source of energy

The only way known to us through which large quantities of energy can be released by matter in a very small volume, clearly necessary properties to understand the nature of quasars, is to consider that the matter falls into a very deep gravitational potential well created by a very massive and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

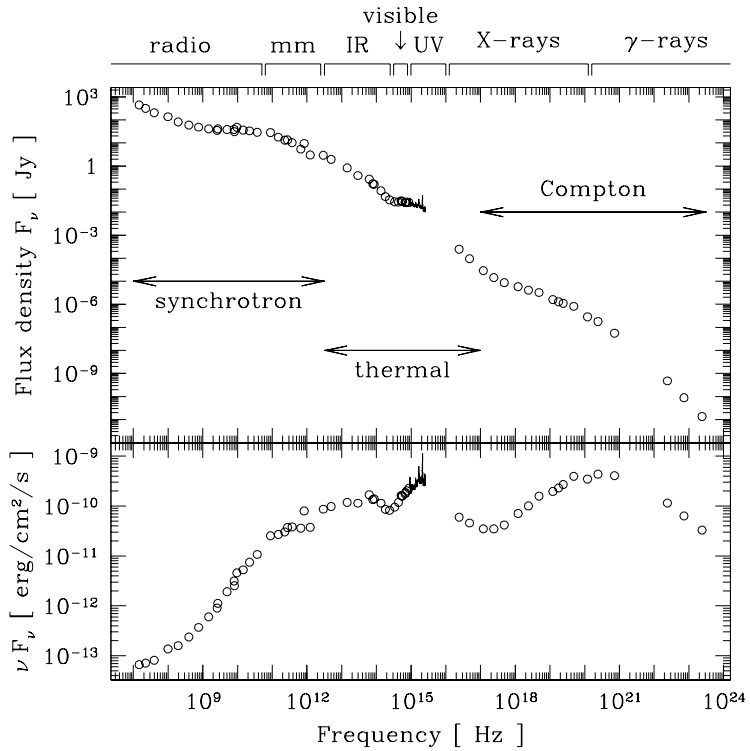


Figure 1. The observed energy per unit area per unit time and per unit frequency (flux density) from quasar 3C 273 is given in the top panel. The emission mechanisms responsible for the emission are mentioned. The bottom panel expresses the same observations per log frequency interval. This shows that we observe roughly the same energy in all regions of the electromagnetic spectrum over 10 decades of frequency.

compact object. It is thus now accepted that quasars host very massive BLACK HOLES into which matter is falling at a high rate. One can estimate the mass of the black hole in the center of 3C 273 by remembering that light exerts a pressure on matter. The gravitational pull must be larger than this pressure for the matter to fall and its gravitational energy to be released. For 3C 273, the mass of the central black hole that is implied by this reasoning is about one billion solar masses.

The mass of the black hole provides a measure of its size; in the case of 3C 373 this size is of the order of 10^{12} m. The corresponding light crossing time is 10 000 s or a few hours. This is well compatible with the fact that quasars (and in particular 3C 273) vary very rapidly.

It is also possible to estimate how much matter must fall into the black hole to explain the observed luminosity by noting that the energy released by the fall is a fraction of the in-falling mass multiplied by the square of the velocity of light. This fraction is of the order of 10%. Again in the case of 3C 273, one finds that the rate of mass in-fall is about the mass of the Sun per year.

Going beyond these simple estimates to understand the true nature of quasars has proved a very difficult endeavor, that is still nowhere close to being finished. This is due to the very unusual properties of the objects compared with the better known celestial bodies and to the very complex geometry of the sources. The latter property prevents one approximating quasars by spherical objects, the properties of which are identical when viewed from different directions. In quasars it appears that the observed properties strongly depend on the direction from which they are viewed.

The radiation of quasars

The energy radiated by quasars is gained from the gravitational binding energy of the matter that falls into the black hole. This energy is radiated by elecrons through several emission mechanisms. Electrons moving along bent paths radiate part of their energy as electromagnetic radiation. In quasars there are many ways through which the paths of electrons are bent; there are correspondingly many emission mechanisms. Magnetic fields are one way of bending electron paths. The radiation emitted by these electrons is called SYNCHROTRON RADIATION. With the parameters typical of quasars this energy appears as radio and millimetre waves. Hot electrons also move along complex paths in the thermal medium. These therefore also radiate what is called thermal radiation. In quasars thermal radiation is observed from heated dust in the infrared part of the electromagnetic spectrum and, most probably, by material heated in the process of the fall towards the black hole. This radiation appears in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

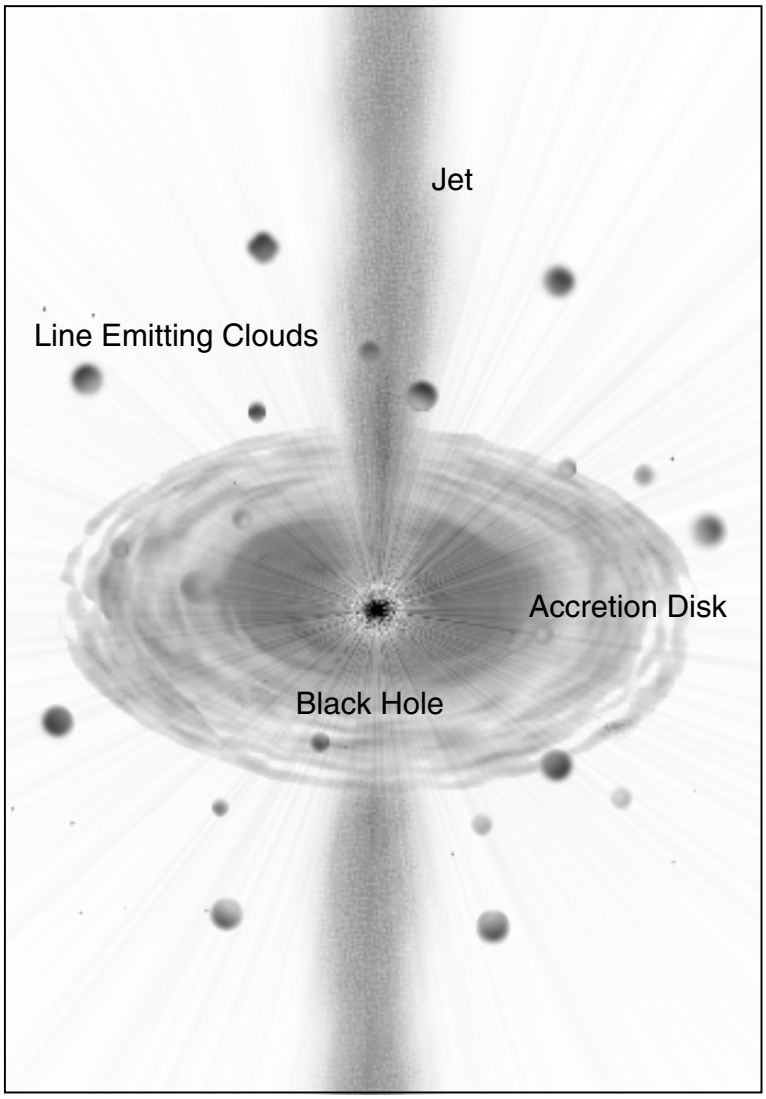


Figure 2. A popular view of the different elements making a quasar like 3C 273. The matter falls into the black hole on spiral paths in the accretion disk. The jet originates at the center of the disk. The line emitting clouds orbit the black hole.

visible and ultraviolet parts of the spectrum. A further way of bending the paths of electrons is through interaction with ambient light. These interactions are called Compton processes. Depending on the energy of the electrons the resulting emission appears in the x-ray or in the gamma ray parts of the spectrum.

We give in figure 1 the energy flux received at the Earth from the quasar 3C 273 and show which parts of this flux are caused by the mechanisms sketched above. We also give in the same figure the flux per logarithmic frequency interval (νF_ν). This representation gives the quantity of energy we receive in each region of the electromagnetic spectrum, i.e. as radio, millimeter, infrared, visible, ultraviolet, x-ray or gamma rays. It is striking to note that we observe roughly the same quantity of energy in all these regions of the spectrum. This is described here for 3C 273; it is true in the same way for other radio bright quasars and with the exception of the radio emission for the radio quiet quasars. This is undoubtedly due to the interactions that exist between the different building blocks of the quasar.

Understanding these interactions is a question of intense research. The approach that is being mostly used is based on repeated observations of the quasar in as many parts as possible of the electromagnetic spectrum. One then uses the variations observed in different parts of the spectrum to see which variations lead others in time. One thus expects to slowly understand what are the more fundamental phenomena and what are the components that are secondary. One also hopes to understand the relative organization of the regions emitting through the components described above. This will hopefully lead to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a description of the geometry of the quasar.

One possible geometry is based on the presence of an ACCRETION DISK around the black hole. In this disk the material is slowed down by viscous forces and thus slowly falls towards the black hole, releasing its gravitational energy. These disks are called accretion disks; they are known to exist around much less massive compact objects within our galaxy. Their presence is rather naturally expected also around the much more massive black holes in the center of external galaxies. The jet is then ejected perpendicular to the disk, at its center. The emission lines are emitted by small clouds that orbit the system. This geometry is illustrated in figure 2. It should be noted that this is a possible way of understanding the organization of matter in quasars and is not yet firmly established.

The future

Quasar 3C 273 was the first in which the very peculiar nature of these objects was recognized. It has been the object of intense studies in the 30 years since its discovery. Many of the observations have led to new insights and to more surprises. 3C 273 is certainly the best observed quasar and thus the best known. All the phenomena that are associated with quasars are seen in 3C 273, whether this makes it the archetype of a quasar or an exceptional object remains to be seen. What is certain is that understanding the physics of 3C 273 is still a challenge.

Bibliography

Courvoisier T J-L 1998 The bright quasar 3C 273 *Astron. Astrophys. Rev.* **9** 1–32

Robson E I 1996 *Active Galactic Nuclei* (New York: Wiley)

Thierry J-L Courvoisier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brisbane, Thomas Makdougall (1773–1860)

Soldier and astronomer, born in Largs, North Ayrshire, Scotland. After a distinguished army career, he became Governor of New South Wales (1821), initiating largely unsuccessful experiments to introduce tobacco, coffee, cotton and flax to Australia. He had built only the second observatory in Scotland at the family home of Brisbane House in 1808, and built an observatory at Parramatta in New South Wales, making the first observations of 7385 stars in the southern hemisphere since NICHOLAS LACAILLE's in 1752. On return to Scotland, built an observatory at Makerstoun and became president of the Edinburgh Astronomical Institution. He is possibly the only astronomer to have a major city named after him (Brisbane, the capital of Queensland).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

British Astronomical Association

The British Astronomical Association was founded in 1890 and has 3000 members. It is open to anyone interested in astronomy and it aims to bring together like-minded amateur astronomers, to collect their observations and organize their observational efforts under the guidance of experienced Section Directors. Papers and current astronomical observations/information are circulated by way of a journal, web page, handbook, circulars, memoirs, section publications and library. It hopes to encourage a popular interest in astronomy. Meetings are held in London and major cities throughout Great Britain. The registered office of the Association is at Burlington House, Piccadilly, London W1V 9AG.

For further information see http://www.ast.cam.ac.uk/~baa.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

British Astronomy

There has been a unified culture of astronomy in Britain, extending over the whole of the British Isles, ever since prehistoric times.

Prehistoric astronomy

The earliest surviving evidence of an interest in astronomy by the inhabitants of the British Isles lies in the alignments of megalithic monuments. Stonehenge in Wessex (2700 BC–) has alignments over the Heel Stone towards midsummer sunrise and could be, in principle, used to predict eclipses. The Temple Wood circle in Argyllshire (2nd millennium BC) appears to be a lunar observatory. The Newgrange barrow in the Boyne valley contains a unique 'roof-box' which admits the rays of the rising Sun at the midwinter solstice into the grave passage. Thus, in prehistoric times, it appears that there was a priestly clan in England, Scotland and Ireland, culturally connected in technology and intellectual knowledge, which included astronomical observation and possibly even prediction.

Archaeological evidence shows that the Roman occupation of Britain (55 BC–c AD 400) brought a knowledge of practical astronomy, including timekeeping and calendrical calculations. This knowledge was maintained, pretty much with its original theoretical bases but elaborated into sophisticated formulae, for example the calculation of Easter by Bede (672–735), a monk in Jarrow.

Medieval astronomy

Islamic astronomy came to Britain in the medieval period. JOHN OF HOLYWOOD, known as Sacrobosco (1195–1256), learnt astronomy in Britain and moved to Paris where he provided his students with a textbook on astronomy, *The Sphere*. A watered-down version of John of Seville's translation of *The Elements* by Al-Farghani, it was the long-term, best-selling, standard textbook in astronomy until the 17th century (it was published in 200 editions in the age of printing), even if it was barely adequate. Educated people, farmers and sailors, had knowledge of the practicalities of astronomical observations. The poet Geoffrey Chaucer (c 1343–1400) was author of the *Canterbury Tales*, which shows his familiarity with astrological matters, and wrote *A Treatise on the Astrolabe*, the oldest known 'technical manual' in the English language.

Renaissance astronomy

In the 16th century, Britain was, during the reign of Queen Elizabeth I, expanding its trade links overseas, taking the first steps to create an Empire. Astronomy concentrated on the practical matter of navigation, including timekeeping. It seemed less important what model of the solar system was correct than what model gave adequate predictions, so it is no surprise that William Shakespeare showed such a little interest in cosmology, although his knowledge of astronomical phenomena was detailed. His plays show no knowledge of Copernicus. They do show a (somewhat skeptical) knowledge of astrology. The astrologer JOHN DEE (1527–1608) cast horoscopes for Elizabeth, receiving the title Royal Astrologer. He claimed to have put a hex on the Spanish Armada and to have caused the bad weather that wrecked the fleet.

Up to the 17th century the teachers of astronomy in Britain were, in the main, clerics with an interest in proclaiming the glory of God and teaching the 'quadrivium' (the four classical subjects of arithmetic, harmony, geometry and astronomy, knowledge of which made a man educated in science). Copernicus' and Kepler's knowledge of the motions of the planets and Galileo's theory of dynamics were taken up by amateurs, like the group that formed around WILLIAM CRABTREE (1610–1644), a Lancashire instrument maker, and JEREMIAH HORROCKS (1619–1641), a schoolmaster. With his own theory of celestial dynamics, related to Kepler's, Horrocks predicted a transit of Venus, which he and Crabtree observed.

Astronomy in Britain became a cultural pursuit as well as a practical one and was a talking point in educated society. John Donne's poetry covers issues such as the appearance of the supernova of 1604 and the new astronomy of Galileo, which put all in doubt, canceling out the old Ptolemaic certainties. Astronomy in the universities (which meant Oxford and Cambridge) turned from the teaching of Aristotelian and Ptolemaic philosophy and astronomy to the empirical observation of new phenomena. Astronomers formulated scientific laws by induction, often expressed them geometrically and increasingly expressed mathematically.

In the 18th century the Government began to take a structural interest in astronomy. King Charles II was persuaded to found the ROYAL OBSERVATORY, GREENWICH in 1675, in order to provide astronomical predictions for the purposes of navigation, for the benefit of the Royal Navy. The predictions were based by JOHN FLAMSTEED (1646–1719), the Royal Observatory's first director, on careful and precise observations, from which increasingly accurate investigations could be made of astronomical phenomena. At the same time, from the universities, Sir Isaac NEWTON (1643–1727) formulated mathematical theories of motion and gravitation, presented in an equivalent of calculus (but expressed in traditional geometric terms). EDMOND HALLEY (1656–1742) used Newton's theory of gravity to predict the return of the comet that bears his name.

Newton and Flamsteed clashed over the running of the Royal Observatory. Flamsteed wanted the observations to be the best possible, sought grants to make the best equipment and took time to develop the observing techniques. Newton wanted observations too, but quickly in order to test his gravitational theory of the planets and the Moon. They quarreled. This fight between the institutes and the academics recurred in 20th century Britain.

Flamsteed as director of the Royal Observatory had taken the title Astronomer Royal as a sign of the state

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

position of astronomy as practiced in Greenwich but the British royal court also continued its more personal interests in astronomy, maintaining court astronomers. Sir William Herschel (1738–1822) and his sister Caroline Herschel (1750–1848), at first supporting themselves through musical performances and teaching, constructed telescopes with which they made significant discoveries—comets, nebulae, double stars, even a new planet—and both were rewarded with royal pensions (see HERSCHEL FAMILY). William was given a stipend and the title of Royal Astronomer, getting King George's telescope ready on fine nights to show discoveries to the royal court. The Queen and her ladies took a personal interest in Caroline's comet, the first discovered by a woman, although the male courtiers belittled it as a small one.

The 19th century

Three groups of astronomers had established themselves in the 19th century: amateurs, Government astronomers and university astronomers. The amateurs included the independently wealthy, for example Lord Rosse (1800–1867) and Sir John Herschel (1792–1871). They included businessmen, for example William Lassell (1799–1880), a Liverpool brewer, and James Nasmyth (1808–1890), a mechanical engineer. Both men built telescopes, and Lassell's, erected under the clear skies of Malta, foreshadowed the overseas observatories of modern times. The amateurs also included pensioners such as retired servicemen (Admiral W H Smyth (1788–1865)). The amateurs concentrated on visual observations of planets, nebulae and double stars but also used talent, time and their own resources to construct large, ingenious or novel instrumentation (such as Sir William Huggins' application of spectroscopy and Common's application of photography to astronomy). Rosse built in Ireland (then part of Great Britain) the then-largest telescope in the world, the 6 ft (1.8 m) aperture 'Leviathan of Parsonstown'.

The university astronomers, particularly at Oxford and Cambridge, but later in London, Durham, Manchester, etc, concentrated on the high intellectual content areas of astronomy, typically theoretical mathematical astronomy (at first celestial mechanics, later astrophysics). The Government astronomers were principally at Greenwich (for example Sir George Airy (1801–1892)) and later also at the ROYAL OBSERVATORY, EDINBURGH. Their work could be characterized as systematic observations by teams of astronomers and 'calculators' in programs principally of positional astronomy. They carried on a tradition started in the 18th century by Halley and Cook of overseas travel to observe astronomical phenomena not visible from Britain, such as hidden constellations, transits and eclipses, and founded the Royal Observatory at the Cape in 1825. CHARLES PIAZZI SMYTH of Edinburgh site tested the mountain of El Teide on Tenerife in the Canary Islands, setting the scene for the establishment of the La Palma observatory.

In 1820 the ROYAL ASTRONOMICAL SOCIETY was founded. Some members met at a gentlemen's dining club to discuss the latest developments. The wider society of astronomers met monthly in open session with a formal program of presentations, which were announced and summarized in the Society's *Monthly Notices*. This journal continues as the principal means of publication of professional astronomical research in Britain (with a substantial content from overseas), in modern times among the trio of the world's most important astronomical journals (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS). It retains the historical name, although it is neither monthly nor collections of notices. Reflecting its 19th century origins, the Royal Astronomical Society remains a society of both professional and amateur astronomers, numbering about 3000, but there are additional exclusively amateur astronomy societies such as the BRITISH ASTRONOMICAL ASSOCIATION.

Modern astronomy in Britain

In the 20th century astrophysics became more important in Britain than positional astronomy, and the thrust of the universities in astrophysics began to challenge astrometry at the royal observatories. Theoretical astronomy, especially cosmology, became very strong—this is the period of Sir Arthur Stanley Eddington (1882–1944), E A Milne (1896–1950) and Sir James Jeans (1877–1946). This area of strength continues, for example in gravitational physics around Steven Hawking. In mid-century a generation of young radar scientists (for example Sir Bernard Lovell (1913–), Sir Martin Ryle (1918–1984, Nobel Prize 1974)) returned to the universities (Manchester and Cambridge) from which, as young men, they had been recruited to work in the Second World War, and established radio astronomy. The radio interferometers at Cambridge made all-sky surveys, establishing the evolutionary big bang theory of the universe, demolishing the steady-state theory produced by Sir Fred Hoyle and his associates and discovering pulsars (Anthony Hewish, Nobel Prize 1974). Jodrell Bank's Lovell Telescope was for a time the largest fully steerable dish for radio astronomy. JODRELL BANK OBSERVATORY is now the center of the MERLIN interferometer spread across England. Meanwhile, space astronomy groups established themselves in the universities of London, Leicester and Birmingham, developing x-ray astronomy, solar astronomy and magnetospheric physics. The number of astronomical knighthoods signifies continuing public, establishment and royal support for astronomy.

The Royal Observatory moved out of London to less smoky, less light-polluted skies over Herstmonceux in the south of England, keeping its link with Greenwich in its new name. As celestial navigation became less important for ships, the Royal Greenwich Observatory shifted its emphasis towards scientific research. In 1965 it was transferred from the control of the Royal Navy to one of the new British Government bodies set up to fund scientific research, the Science Research Council (later the Science and Engineering Research Council, still later the PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL, which is the currently operating UK funding agency for astronomy). The Government-funded research councils

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

were given university-oriented mission statements and controlled by governing bodies with members from the universities, whose number, power and influence were growing with the increase of tertiary education in Britain.

The optical astronomers proposed a large 2.5 m aperture telescope. Because of its established position of power, it was the Royal Greenwich Observatory, under Sir Richard van der Riet Woolley, which built the Isaac Newton Telescope (1967). It did not satisfy the universities' needs. During study overseas, university astronomers had used telescopes in good climates such as California and did not like sitting under cloudy skies at night in Sussex. Also they did not easily relate to the civil service style of operation of the telescope.

A southern hemisphere facility had become important to the space astronomers with their all-sky capability. They pressed for the establishment of the ANGLO-AUSTRALIAN OBSERVATORY in Coonabarabran (1974). The EUROPEAN SPACE AGENCY was formed in 1975 and its Science Program to launch astronomical telescopes and to explore geospace and the planets has become Britain's main space astronomy program. The UK takes a high profile and active interest in ESA and its Science Program.

Infrared astronomers sought out a high site, Mauna Kea, Hawaii, for the UNITED KINGDOM INFRARED TELESCOPE (1979), operated by the Royal Observatory, Edinburgh. Later, the JAMES CLERK MAXWELL TELESCOPE for millimeter astronomy was established there. The radio astronomers pressed for an optical 'northern hemisphere observatory' to follow up their radio observations made from Britain. It was created by the move in 1984 of the Isaac Newton Telescope to La Palma in the Canary Islands. The main telescope at the observatory was completed in 1986, the 4.2 m William Herschel Telescope.

The Royal Greenwich Observatory continued to operate the telescopes in La Palma but, having lost its main telescope to an overseas site, it was moved for a second time in 1991, to the university environment of Cambridge. There was competition for resources between the two Royal Observatories (Greenwich and Edinburgh) and the universities. This led in 1998 to the closure of the Royal Observatories in Cambridge and Edinburgh and the establishment in Edinburgh of a combined UK Astronomy Technology Center to make astronomical instruments at the site still called the Royal Observatory, Edinburgh. In 1999 the first of the two 8 m Gemini Telescopes was opened on Mauna Kea, with the second soon to be established in Chile.

All Britain's current large telescopes (save MERLIN and UKIRT) are multinational, with Britain partnering Australia, Canada, the Netherlands, Spain, the USA and other countries. The UK is not however currently a member of ESO, having declined to join on grounds of expense at its foundation, and presently rueful about missing out on participation in the VLT.

In Britain today there are about 6000 amateur astronomers (judging by membership of the amateur societies, allowing for overlaps). By contrast with many continental European countries, professional astronomy has become dispersed into the universities, no longer concentrated into special institutes. There are 50 universities that teach astronomy at an undergraduate level (see EDUCATION). About half of these have significant astronomy research groups. British astronomers produce about 10% of the world's output of astronomy (about a quarter that of the USA, to whom the UK lies in second place, and about the same as France and Germany). There are some 300 permanent university astronomers, 400 post-doctoral researchers with fixed-term positions and some 450 PhD students, many of whom are independently financed. About 25% of the PhD students remain in astronomy after a decade—most transfer their skills to industry. The number of university teachers shows a steady increase, financed by the universities in response to an increasing student demand. Public interest in astronomy as a high-profile, flagship science remains strong and increasing and there is virtually daily coverage of astronomy in the UK's newspapers and broadcast MEDIA. However, the Government funding for astronomy is flat (at £95 m p.a. in 1999). Indeed there has been small but steady reduction in Government-provided purchasing power for astronomy for 15 yr, a dramatic fall in the number of astronomers employed by the Government institutes and a recent decrease in the number of doctoral and post-doctoral researchers. In spite of this, perhaps because of the competitive nature of the university environment, British astronomers are international, dynamic and flexible and, at the present time, in spite of chronic financial stress, remain successful in continuing Britain's 4000 yr old attempt to understand the universe.

Paul Murdin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brorsen, Theodore (1819–95)

Danish astronomer, observed the gegenschein.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brouwer, Dirk (1902–66)

Celestial mechanician, born in and educated in The Netherlands, director of the Yale University Observatory. He determined the mass of Titan from its influence on other Saturnian moons, and developed methods for finding orbits of comets, asteroids and planets, redetermining the astronomical constants. He was one of the first to use electronic computers for astronomical computations, computing the orbits of the first artificial satellites, obtaining increased knowledge of the figure of the Earth. He was editor of the *Astronomical Journal*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brown, Ernest William (1866–1938)

Celestial mechanician, born in England, taught mathematics at Haverford College in Pennsylvania and at Yale University. Devoted his life to calculations of the Moon's motion. His theory, which was a set of equations with an astounding 1500 terms, took account of 'the gravitational action of every particle of matter that can have a sensible effect on the Moon's motion'. He computed the constants of the terms by analysing 150 years of Greenwich observations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Brown, Robert Hanbury (1916–)

Physicist, astronomer. Educated in London, after wartime service became radioastronomer at Jodrell Bank, where he discovered that M31 is a radio source; professor at University of Sydney. Despite being assured, by colleagues who understood less quantum mechanics than he did, that it was impossible, he invented the Intensity Interferometer, which created wave-like interference in optical sensors after the incoming light had been detected. Used the interferometer, constructed at Narrabri in Australia, to determine the angular diameters of the brighter stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bruno, Giordano (1548–1600)

Philosopher, born in Nola, near Naples, Italy as Filippo, took the name Giordano upon becoming a Dominican. Was taught Aristotelian philosophy in the Dominican monastery in Naples, but became attracted to Plato and Hermes Trismegistus and came to the unwelcome attention of the Inquisition, a situation possibly exacerbated by his intemperate, disputative style. Self-exiled to France and England. Published *Cena de le Ceneri* (*The Ash Wednesday Supper*) in which he defended the heliocentric theory of Copernicus, and *De l'Infinito, Universo e Mondi* (*On Infinity, the Universe and its Worlds*) in which he argued that the universe was infinite, that it contained an infinite number of worlds, and that these are all inhabited by intelligent beings. Unwisely returning to Italy (Venice and Rome) he was arrested by the Inquisition, imprisoned and tried. In the end, he refused to recant, was declared a heretic and burned at the stake. The precise grounds for his execution are not known because his file has not been seen, but it is often maintained that Bruno was executed because of his Copernicanism and his belief in the infinity of inhabited worlds. Bruno was the first Renaissance scientist to adopt the atomist philosophy of Democritus and Lucretius.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bubble Nebula (NGC 7635)

A faint diffuse nebula in the constellation Cassiopeia, position RA $23^{h}\ 20.7^{m}$, dec. $+\ 61^{\circ}\ 12'$, appearing as a $3'$ diameter bubble surrounding a seventh-magnitude star. The bubble is a spherical shell of gas apparently shed by the central star, although it does not appear to have the characteristics of either a planetary nebula or a supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Buffon, Georges Louis Leclerc Comte de (1707–88)

French naturalist. Discovered the binomial theorem and worked on probability theory. In astronomy he suggested that the Earth might have been created by the collision of a comet with the Sun. Based on the cooling rate of iron, he calculated in *Théorie de la Terre* that the age of the Earth was 75 000 years. This estimate, so much larger than the official 6000 years, was condemned by the Catholic Church in France, which burned Buffon's books.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bulgarian Academy of Sciences, Institute of Astronomy

The Institute of Astronomy of the Bulgarian Academy of Sciences (IABG, BAS) is located in Sofia. Its activities include fundamental research in astronomy and astrophysics and teaching. The Institute was established in 1995 as a successor of the Section of Astronomy (SA) and National Astronomical Observatory 'Rozhen' (NAO). Professor Nikola Bonev was the founder (in 1958) and the first director of SA. In 1995, the first director of the newly established IABG was Professor Vladimir Shkodrov. Its total staff is 90 including 42 scientists and 34 PhD holders. Two modern observatories for optical studies belong to the IABG: the National Astronomical Observatory 'Rozhen' (long: −1 h 38 m 52 s; lat: +41 deg 43′; altitude: 1750 m) and the Observatory in Belogradchik (long: −1 h 30 m 42 s; lat: +43 deg 38′; altitude: 630).

The main telescope of NAO is a 2 m Ritchey–Chrétien Coudé made by Carl Zeiss Jena and equipped with a 1024 × 1024 Photometrics CCD camera. The Coudé spectrograph provides dispersions of 4 Å, 9 Å and 18 Å. At present, it is the biggest optical telescope in SE Europe. The other telescopes of the NAO are: 60 cm Cassegrain with a UBV photometer, 50/70 cm Schmidt (5 × 5 deg) and an 18 cm photoheliograph refractor for solar studies. The BO is equipped with a 60 cm Cassegrain photometric telescope. Research departments of the IABG include: Sun, solar system, non-stationary stars, stellar atmospheres and envelopes, chemically peculiar stars, stellar clusters and galaxies. The budget for 1999 was 400 000 DM (from the state) and 7000 DM (from the National Science Foundation). Several international organizations participate: International Astronomical Union, European Astronomical Union, The New York Academy of Sciences, etc. International agreements cover joint research projects with the astronomical institutes of Germany, France, Poland, Russia, Italy, etc.

For further information see
http://www.astro.bas.bg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Burbidge, Eleanor Margaret Peachey (1919–)

Astrophysicist, born in Devonport, Devon, England, worked in London, Yerkes Observatory, Cambridge, the California Institute of Technology, the Royal Greenwich Observatory and University of California at San Diego. Married and collaborated with Geoffrey Burbidge, a theoretical physicist. It is said that in the days when women were not permitted to observe in their own right with the 200 in telescope, it was Geoffrey who signed Margaret's applications to use it, and read astronomical journals and dozed in its control room while she worked through the night. She collaborated with her husband, FRED HOYLE, and WILLIAM FOWLER in a paper, universally known from its authors' initials as B^2FH, on the synthesis of the chemical elements beyond helium in stars. She used the spectra of galaxies to determine their rotations, masses and chemical composition, and elucidate the properties of their peculiarities and central cores (quasars).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Buridan [Buridano], John ($fl.$ fourteenth century)

Teacher at the University of Paris between about 1328 and 1358, influential, critical commentator on Aristotelian philosophy, his books were still used in universities as late as the seventeenth century.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Burnham, Sherburne Wesley (1838–1921)

Court reporter and astronomer, born in Thetford, VT. Took up astronomy as a soldier in the American Civil War. Double star observer at Yerkes Observatory. His *General Catalogue of Double Stars* (1906) listed all 13 665 double stars in the northern hemisphere, including over 1000 which he had discovered and observed himself.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Byurakan Astrophysical Observatory

Located 40 km north of Yerevan in the former Soviet Republic of Armenia. Founded in 1946 through the initiative of astronomer Victor Ambartsumian, it is now the home of the Astrophysical Observatory of the Armenian Academy of Sciences. By late 1999, the only telescope still functioning was the main 2.6 m reflecting telescope.

For further information see http://bao.sci.am/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Caelum

(the Chisel; abbrev. Cae, gen. Caeli; area 125 sq. deg.) A southern constellation that lies between Eridanus and Columba, and culminates at midnight in late November. It was named Caela Sculptoris (the Sculptor's Chisels) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2. The shortened form was adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest star in Caelum is of magnitude 4.4. Interesting objects include γ^1 Caeli, a binary with orange (K2) and white components, magnitudes 4.7 and 8.5, separation 3″, and R Caeli, a Mira-type variable star (range 6.7–13.7, period about 391 days). There are no bright star clusters, nebulae or galaxies in Caelum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cagliari Astronomical Observatory

The Cagliari Astronomical Observatory was established in 1899 as the International Latitude Astronomical Station of Carloforte, a small town on the Sardinian island of San Pietro. For about 80 years it has been one of five international stations devoted to the study of Earth's rotation and polar motion.

Observations with its zenithal telescope were only interrupted during the Second World War. Since 1978, its headquarters have been at Punta Sa Menta, a site 15 km (9 miles) from Cagliari, which has the same latitude as the Carloforte station.

Astrophysical research is mainly devoted to the study of the Diffuse Interstellar Bands in the interstellar medium, using the Coude Echelle Spectrograph at the EUROPEAN SOUTHERN OBSERVATORY in Chile. It is also the site of a satellite laser ranging station.

For further information see http://www.ca.astro.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Calar Alto Observatory

The Calar Alto Observatory, or 'Centro Astronomico Hispano-Aleman', is located at an altitude of 2168 m in the 'Sierra de los Filabres', in southern Spain. Its construction on Calar Alto mountain began in 1973. It is operated jointly by the MAX-PLANCK-INSTITUT FÜR ASTRONOMIE in Heidelberg (MPIA), Germany, and the 'Comision Nacional de Astronomia'. The MPIA provides four telescopes of diameters 3.5 m, 2.2 m, 1.23 m and a 0.8 m Schmidt telescope. The 'Observatorio Astronomico Nacional' (OAN) independently runs a 1.5 m telescope.

Telescope time is awarded semi-annually through peer-reviewed proposals. Typical time allocation for the MPIA telescopes is 40% for MPIA, 50% for other German institutes and 10% for the Spanish community. Calar Alto facilities may also be used by guest observers from any other country.

Instrumentation is state of the art and provides competitive and complementary facilities. It includes an adaptive optics system (ALFA) which employs a sodium artificial laser guide star system, various near-infrared cameras (OMEGA Prime, OMEGA Cass, Magic) with spectroscopic and polarimetric capabilities, a fiber-fed echelle spectrograph (FOCES) and focal reducers providing multi-object spectroscopy (CAFOS, MOSCA), among others.

For further information see
http://www.caha.es.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Calendar

A calendar is a system of organizing units of time for the purpose of reckoning time over extended periods. The DAY is the smallest calendrical unit of time; the measurement of fractions of a day is timekeeping. Some calendars replicate astronomical cycles according to fixed rules, others are regulated by astronomical observations. Commonly, calendars take the form of reckoning time with reference to the beginning, length and divisions of the YEAR. They are presented as a tabular arrangement of the days of each MONTH and week in a year.

The periods of the motions of the Sun and Moon and the daily rotations of the Earth have been the natural bases for development of calendars. However, some civilizations have based calendars on other solar system bodies, such as the Mayan use of Venus. Unfortunately, none of the natural resulting periods (year, lunation or day) are exact multiples of the other. Thus, calendars based on lunar months require adjustments to remain close to the year based on the apparent motion of the Sun. The exact period of the Earth's motion around the Sun is not an integral number of days, so it requires adjustments to keep the seasons, or more significantly the religious holidays, at the same times in the calendar. The seasons and some religious holidays are derived from natural body motions. Some are defined in terms of the vernal EQUINOX, which strictly follows the motion of the Earth. There is an ecclesiastical equinox defined to permit the determination of Easter without variations due to the true motion.

The tropical year is defined as the mean interval between vernal equinoxes. The current mean value of the tropical year is 365.242 190 days. The value decreases due to tidal braking by six millionths of a day per century. The true interval from a particular vernal equinox to the next may vary from this mean by several minutes. The exact length of the tropical year is dependent on the general planetary theory or numerical integration used to determine the orbital elements.

The current calendar used internationally is the Gregorian calendar. This was introduced by Pope Gregory in 1582 in some countries and as late as 1927 for others. The Gregorian calendar has 365 days per year except for leap years of 366 days. Leap years are years evenly divisible by 4, except for years that are exactly divisible by 100; these centural years are leap years only if they are exactly divisible by 400. Thus, 1800, 1900, 2100 and 2200 are not leap years, but 2000 is a leap year.

The Gregorian calendar year is 365.2425 days long compared with the current tropical year of 365.242 190 days, a difference of 0.000 310 days, or 1 day per 3000 years. The INTERNATIONAL ASTRONOMICAL UNION (IAU) has adopted the Julian year of 365.25 days as a time unit for many computations such as precession and orbital elements, to avoid the variability and uncertainty of the value of the tropical year. There have been unconfirmed statements of modifications of the Gregorian calendar by some countries to correct for this remaining discrepancy. However, there are no known corrections for long-term discrepancies, and improved ephemerides may change the magnitude of the discrepancy.

Also there are frequent radical proposals for simplifying, standardizing or improving the Gregorian calendar. The current total international use of the Gregorian calendar will require that any change must offer a worldwide benefit that justifies the change.

In many cases calendars have been based on the lunar cycle. The synodic month, the mean interval between conjunctions of the Moon and the Sun, corresponds to the cycle of lunar phases. The mean value of the synodic month is 29.530 589 days and increases due to tidal effects by two millionths of a day per century. Any lunar phase cycle can deviate from the mean by up to 7 h.

Some lunar calendars (e.g. Islamic) begin a new month based on observations of the lunar crescent after a new moon. In these cases the beginning of the month depends on weather conditions, atmospheric clarity, eye acuity and the location on Earth, since the Moon can generally only be observed about 15 h after new moon and about two degrees above the horizon. These conditions can generally be accurately predicted with about a 30 degree uncertainty in Earth longitude for actual observability. In some cases, predictions are not accepted and observations are required for the determination of the beginning of the month.

Since there are not an integral number of lunar months in a solar year, some calendars (e.g. Hebrew) have introduced intercalated months to adjust the months such that religious holidays occur in the same season each year (see also CALENDAR IN THE MIDDLE EAST AND EUROPE).

There are non-astronomical features to calendars, such as the seven day week. The origin of the seven day week is uncertain, with explanations ranging from weather cycles to Biblical and Talmudic texts. There is also an unproven assumption that the cyclic continuity of the week has been maintained without interruption from its origin in Biblical times.

Bibliography

Seidelmann P K 1991 (ed) *The Explanatory Supplement to the Astronomical Almanac* (Mill Valley, CA: University Science Books)

P Kenneth Seidelmann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Calendar in the Middle East and Europe

Calendars are the means for marking the passage of time. Their use can be civil, for the ordering of daily life, or religious, for the marking of observance days, or as a means for noting the rhythm of the seasons for agricultural purposes. Much of calendrical structure is purely arbitrary, such as the length of hours, hours in a day, days in a week and the year within an era. Yet most calendars reflect natural astronomical cycles—the apparent motions of the Sun and Moon and the daily rotation of the Earth. Crude calendars may have been used in neolithic times (see PREHISTORIC ASTRONOMY), but here we consider the development, in historical times, of calendrical ideas of the Middle East and Europe, which culminated in a system used today throughout the world.

Calendars in the Western tradition are rooted in the cyclical motions of the Sun, the Moon or both. The sidereal year measures the time required for the Sun to return to the same place in the sky with respect to the fixed stars. The tropical year is the time from one vernal EQUINOX to the next. The latter is slightly shorter because PRECESSION shifts the equinoctial points westward each year. The key lunar cycle is the synodic period, the passage of the Moon through its phases. When the two are combined as a lunisolar calendar, coordination is more complex. As neither solar nor lunar periods are precise multiples of days, early astronomers were forced to introduce intercalary days at intervals to bring the cycles into line. Otherwise, the months (or years) would slowly slide backwards through the seasons. Because the lunar and solar cycles are incommensurable, no number of synodic months will equal a given number of tropical years, even over very long periods. Nonetheless, they can be equated closely enough to produce a workable calendar.

Ancient calendars

To regulate religious practices, the early Egyptians devised a lunar calendar, inaugurated each year by the heliacal rising of Sirius, i.e. when Sirius appeared briefly in the morning twilight for the first time. For administrative use, they introduced a solar calendar early in the 3rd millennium BC. This system became the basis of our modern one. Assuming a year to have 365 days, they divided it into 12 equal months of 30 days, followed by 5 days at the end. Each month had three 10 day weeks. Being about a quarter-day short of the tropical year, the calendar's initial day slowly slipped backwards, with the new year arriving a day earlier in about 4 yr, but the discrepancy was ignored. The Egyptians also introduced the 24 h day, although the 12 night hours, 10 day hours and 2 dawn–dusk hours were of unequal length. Unequal, seasonal hours remained common throughout the West until the astronomers' equal-hour system spread during the 14th century along with mechanical clocks.

Mesopotamian civilizations developed a lunisolar calendar based on a 7 day week. This length may have symbolized the seven 'planets' (Sun, Moon and five naked-eye planets). Although Mesopotamian astronomers intercalated extra lunar months earlier, it was not until the 4th century BC that the systematic intercalation of seven lunar months over a 19 yr period appeared. This scheme, proposed by the Athenian astronomer METOS about the same time, was later known as the Metonic cycle. Observations showed that, in 19 tropical years, almost exactly 235 lunations occurred. For calculating the dates of new moons, one employed alternating months of 30 and 29 days. After 12 lunations, the total of 354 days was short of the solar year by about 11.25 days. However, these accumulated to more than 33 days after 3 yr, and an extra lunation would be inserted. Thus, 12 yr of 12 lunations, together with 7 yr of 13 lunations, would result in the cycle being brought back to the beginning. The Persian Empire standardized the system, with intercalation in the third, sixth, eighth, eleventh, fourteenth, seventeenth and nineteenth years of a cycle. The Jews and other Middle-Eastern peoples adopted this scheme.

Greek city-states employed a less accurate intercalation scheme, the *octaëteris*, devised in perhaps the 8th century BC. It assumed that eight tropical years are equal to 96 regular plus 3 intercalated lunations averaging 29.5 days. However, having no political unity, Greek cities maintained individual lunisolar calendars, with different start dates and times for intercalating months. Metos, in the 5th century, noted that the tropical year is about 365.25 days. His 19 yr cycle was spread further afield during the Hellenistic period. HIPPARCHUS, in the 2nd century BC, recognized precession and determined very accurate values for both the tropical year and the synodic period.

The Roman republican calendar was less sophisticated than the Babylonian (see BABYLONIAN ASTRONOMY). Based on an older lunar calendar, it had 12 months, some named (*Ianuarius, Februarius, Martius, Aprilis, Maius, Iunius*) and some numbered (*Quintilis, Sextilis, September, October, November, December*). As the year began in March, *Ianuarius and Februarius* were months 11 and 12, respectively. This civil calendar was a day longer than the tropical year and suffered from unsystematic intercalations. By the 1st century BC, the new year fell almost 2 months after the vernal equinox. An 8 day week was fixed by market days, although the 7 day Jewish week became more common in subject territories. Unlike other Mediterranean peoples, the Romans began the day at midnight.

The Julian reform and medieval calendars

Julius Caesar, during his sojourn in Egypt, saw the value of adopting the Egyptian solar calendar and, on the advice of the Alexandrine astronomer Sosigenes, promulgated what is now called the Julian calendar in 46 BC. In this entirely solar calendar, a year was set at 365.25 days. Instead of tacking the five extra days on the end of the year, the Julian calendar distributed the extra days across the year by alternating 31 and 30 day months. The year commenced in January, which had 31 days. Such a distribution would produce a year of 366 days, so one day was removed from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

February, which now became 29 days long. To account for the accumulation of an extra day every 4 yr, February was given 30 days during the leap year, the intercalated day usually being inserted after the 24th of the month. Only minor changes were made thereafter: in 44 BC, the month *Quintilis* was re-named *Iulius*, after the assassinated Caesar, and, in 8 BC, Augustus named *Sextilis* after himself, providing it with 31 days. This changed the order of alternation of month lengths, and robbed February of its 29th day. The Augustan adjustments to month lengths survive in modern usage. Augustus, realizing that priests had made every year a leap year since Caesar's time, suppressed all leap years until AD 8.

The Julian calendar, like its Egyptian model, had the value of simplicity. The difference between the Julian and tropical years was sufficiently small that the difference accumulated slowly. As the official calendar of the Roman Empire, the Julian scheme was in force from the Middle East to Scotland; when the Christian church became the dominant religion within the empire, the calendar's survival was assured.

The pre-exilic Jewish calendar was lunar, although little is known about it. After the Babylonian conquest of Israel, the Jewish calendar, like other Semitic calendars, became lunisolar, each year having 12 alternating months of 29 and 30 days. After changes in the 2nd and 3rd centuries AD, it followed the Metonic cycle, with intercalations made according to the Babylonian order. The year began with the autumnal equinox—before the Babylonian exile it began with the vernal equinox—with 7 day weeks, beginning at sundown on Friday. Months commenced with the first sighting of the new crescent Moon; in practice the astronomically calculated time had been used by some Jewish communities from the 3rd century. Days began at sundown, although officially they begin at 6.00 pm in modern practice, with the week commencing on the Sabbath (Friday evening). Following Babylonian tradition, each of the 24 hours in a day is divided into 1080 units based on sexagesimal reckoning. Years are counted from the beginning of the world (*anno mundi*, or 3761 BC).

The Islamic calendar had a common origin with the Jewish and Christian calendars except that it dropped the link to the tropical year. Twelve lunar months of 29 and 30 days alternate. Being strictly lunar (not lunisolar), there are no intercalated months—although an intercalated day may be added to a given year—so that religious observances, such as Ramadan, fall in different seasons over the years, cycling through all the seasons in about 33 years. Like the Jewish, the Islamic calendar has 7 days, each beginning at sundown. The actual visual observation of the new crescent Moon was the start of each lunar month. During the later middle ages, Islamic astronomers developed methods and tables to define the time of the new moon for different locations and when the Moon was obscured by cloud (see also ISLAMIC ASTRONOMY). For civil purposes, the Ottoman Empire adopted the Julian calendar in the 17th century, with Turkey moving to the Gregorian system in the 20th. Most Islamic countries now use the Gregorian calendar for civil purposes, but also employ the era dating from Mohammed's flight to Medina (AD 622).

While the Romans even before the Julian reform began their year at 1 January, only the western part of the empire followed suit. The Byzantine empire moved to 1 September. Early medieval Christians of the West could not agree on an inaugural date. In England and France, Christmas day was the new year; in other areas, 25 March, or even Easter (which was a movable date) took precedence. By the late middle ages, 1 January regained its status.

The Jewish 7 day week, adopted by early Christians, did not become legally binding in the Roman Empire until the time of Constantine (AD 321). Although the Church attempted to suppress the Roman month names—replacing *dies Saturni* with *sabbata* and *dies soli* with *domenicus*, for example—older names persisted in northern Europe. In the counting of days of the month, there was also confusion. The Roman republican method of beginning a month with the *kalends*, followed several days later by *nones* and *ides* and numbering days backwards in relation to these three fixed days, survived in some areas until the 15th century, when successive numbering of days became general.

Because the Julian calendar employed a longer tropical year than actually occurs, the date of the vernal equinox slowly moved backwards. For the Christian Church, this created a problem in setting the date for Easter. Being tied to the Jewish Passover, itself based on the lunisolar calendar, Easter was difficult to fix. The Council of Nicaea (AD 325) set 21 March as the date of the vernal equinox, and adopted the Metonic cycle. Dionysius Exiguus, in the 6th century, provided a rule for setting Easter: it would occur on the first Sunday following the first full moon after the vernal equinox. Should the full moon fall on a Sunday, Easter was deferred to the following week. These complications gave rise a calendrical literature, the *computus*, to which Bede, in the early 8th century, was a notable earlier contributor.

The common scheme was to intercalate 30 day leap months into a 19 yr cycle using the same scheme as the Jewish calendar. Over a 19 yr period, the Julian calendar accumulated 6939 days 18 h, while the lunar cycle accumulated 6940 days 18 h. To bring the two into synchronicity, a day was dropped (the *saltus lunae*), usually in the last month of the cycle. Ensuring that Easter fell on a Sunday and to know when it would occur were more complex. Bede and others introduced a perpetual calendar based on compound cycles, golden numbers and epacts. With such a scheme, tables could be drawn up well in advance to calculate Easter dates.

Agreement on an era for dating was longer in coming. Many ancient countries based their eras on regnal years or years of high officials' tenure. Some Greek cities followed olympiads. In imperial times, the Romans utilized both regnal years and the year since the foundation of Rome

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(*ab urbe condita*). The foundation of the Christian era (*anno Domini*) was suggested by Dionysius Exiguus in the 6th century. Trying to solve the Easter-reckoning issue, Dionysius adopted the Victorian period (devised by Victorinus of Aquitaine in the 5th century), which multiplied the 19 yr Metonic cycle by a 28 yr solar cycle to obtain a 532 yr cycle. He then calculated Christ's birth year, from which the first 532 yr cycle would commence. This he believed to be 753 AUC. The English, under the influence of Bede, adopted the Christian era for ecclesiastical purposes in the 7th century, but it was not generally accepted in Europe until the 11th century.

Calendar reform

By the time of Roger Bacon in the 13th century, it was clear that paschal full moons did not fall on their computed dates. Once precession was widely understood, and better tables, particularly the Alfonsine tables of the later 13th century, were available, the pressure for calendar reform grew. Despite a variety of proposals, it was only in the papal reign of Gregory XIII that a revised system, devised by Aloisio Lilio, was promulgated in 1582. The Alfonsine length of the tropical year, being just more than 10.5 min shorter than the Julian year, meant that one would need to drop a day every 134 yr. The papal commission recommended following the Julian leap-year arrangement but suppressing leap years in century years (such as 1900) that are not divisible by 400. The Gregorian reform also adjusted the system for Easter reckoning. As the date of the vernal equinox had slid backwards some 10 days from 21 March over the intervening centuries, Gregory decreed that 5 October 1582 would be followed by 15 October. Catholic countries adopted the system within the first 5 yr; it was accepted in most Protestant nations by the 18th century, such as Britain in 1752. As the Eastern Orthodox Church has never accepted the reform, eastern European countries delayed adoption until the 20th century, such as Revolutionary Russia in 1918 and Greece in 1923. Some Asian and Muslim countries also adopted the system in the late 19th and early 20th centuries.

For a brief period (1793–1805), revolutionary France instituted a calendar based upon the ancient Egyptian solar calendar to remove any religious connotations. Twelve 30 day months were followed by five days of holidays, inserted into early July; a revised leap-year scheme was modified from the Gregorian. Like the Egyptian calendar, three 10 day weeks made up each month. Months themselves were renamed to reflect seasonal weather and activities. As the beginning dates of the months were out of step with the rest of Europe, problems multiplied and Napoleon I returned the country to the Gregorian system.

Since the 19th century, a variety of calendar reform proposals have been made, particularly to solve the problem of months of unequal days. The International Fixed Calendar would provide 13 months of 28 days each, plus one unnamed day to equal 365; a leap-year day would be added each fourth year. The advantage to this scheme is that all months begin on Sunday and end on Saturday. An alternative, the World Calendar, would have four 91 day quarters, each beginning on a Sunday, with months following a 31, 30 and 30 day pattern. A world day would be added after 30 December, plus an additional world day for leap years. Both have advantages, but neither they nor others have elicited much official enthusiasm.

While many cultures over the past few thousand years have devised calendars based on astronomical phenomena, the hybrid formed from the Babylonian–Jewish tradition and the Egyptian–Roman tradition has become the dominant calendrical system world wide. While it cannot be precisely accurate, given the incommensurability of lunar and solar periods, it is entirely sufficient for political and economic needs. Calendar reform, if it does occur, is most likely to be a rationalization of the existing system rather than a wholesale change in structure.

Bibliography

Coyne G V, Hoskin M A and Pedersen O (eds) 1983 *Gregorian Reform of the Calendar: Proc. Vatican Conf. to Commemorate its 400th Anniversary 1582–1982* (Vatican City: Pontificia Academia Scientiarum)

Rickards E G 1999 *Mapping Time: The Calender and its History* (New York: Oxford University Press)

Richard A Jarrell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

California Nebula (NGC 1499)

An emission nebula in the constellation Perseus, position RA $4^{\mathrm{h}}\,00.7^{\mathrm{m}}$, dec. $+36^{\circ}\,37'$, whose shape resembles the US state of California. The nebula is large (about $2\frac{1}{2}^{\circ}$ by $\frac{2}{3}^{\circ}$), but of low surface brightness. It is illuminated by the star ξ Persei.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Callippus of Cyzicus (*c*. 370–*c*. 310 BC)

Greek philosopher, born in Cyzicus, Asia Minor (now Turkey), a younger contemporary of EUDOXUS, who added more spheres to Eudoxus' nests of spheres to account better for the non-uniformities of planetary motion. The Sun, Moon, Mercury, Venus and Mars each had five spheres while Jupiter and Saturn had four and the stars had one. This increased the accuracy of the theory while preserving the belief that the heavenly bodies moved in circles. Worked with ARISTOTLE. Determined the lengths of the seasons. Constructed a 76 year cycle (the Callippic period) as a more accurate version (four cycles) of the 19 year Metonic cycle, the length of time in which the lunar and solar calendars come back into the same synchronism and by which lunar eclipses can be predicted.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Callisto

Callisto, the outermost of the Galilean satellites (see JUPITER: SATELLITES), is approximately 38% larger in diameter than Earth's Moon (Callisto's radius is 2410 km and the radius of the Moon is 1738 km). However, Callisto's lower mass, surface gravity and higher albedo reflect its ice–rock composition. Callisto's surface is thought to consist of a mixture of ice and dust or rocky materials. Callisto's cratered terrain can be divided into irregularly shaped subunits based on ALBEDO, which vary by a factor of two between units. Several subunits are seen to be relatively smooth and lack impact craters at VOYAGER resolutions; these areas have been interpreted to be regions of possible volcanic resurfacing. However, GALILEO MISSION data which have targeted these regions show that the smoothness is due to surface degradation rather than icy volcanism. Early global tectonism has been suggested to explain the presence of parallel and radial sets of lineaments. Whether or not Callisto is differentiated could not be determined from Voyager data. Galileo images of Callisto's surface do not provide clear evidence of DIFFERENTIATION or the presence of a subsurface liquid water ocean. There is a lack of clear geological evidence of any endogenic modification of Callisto's surface.

Prior to the Galileo mission scientists believed that the surface of Callisto would be heavily cratered at all scales. However, the surface of Callisto is seen in moderate- to high-resolution Galileo images to be smooth and slightly undulate, containing a low spatial density of small (<10 km/pixel) impact CRATERS (see figure 1). The wide spatial distribution of this smooth texture suggests that the smoothing process(es) are global in extent. The dark non-ice material that appears to blanket the surface may become concentrated by the loss of volatile materials during impact events, magnetospheric bombardment and/or via sublimation. This material may also be distributed by impact events and micrometeorite bombardment resulting in the infilling of low-lying topography. Electrostatic levitation or other mechanisms might further contribute to the mobilization of Callisto's dark material.

Impact craters on Callisto exhibit a range of size and degradation states, from fresh with clearly defined ejecta blankets to craters that are barely distinguishable. Mass wasting of crater walls appears to be a predominant process in this degradation. Deposits within craters that appear to be the result of inward debris slides are prevalent. Several of these debris flows originate on the west-facing walls of geologically recent craters. This may indicate a steepening of the crater wall, perhaps by sublimation, that results in a reduction of its stability. Subsequent collapse may be triggered by seismic energy from a nearby impact, or eventual oversteepening of the wall leading to gravitational collapse. This ongoing sublimation erosion and mass wasting best explains the preferential sun-facing rim retreat and crater-wall erosion seen in Galileo images, contributing to the progressive infilling of craters on Callisto's surface.

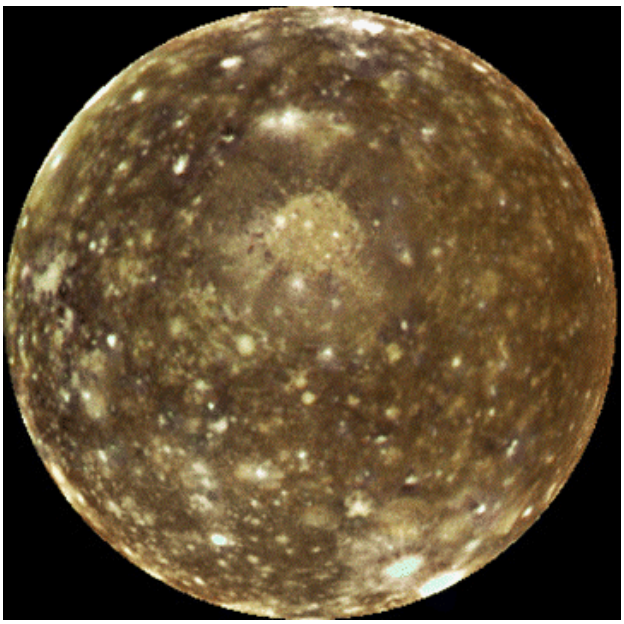


Figure 1. Callisto.

Voyager images show that Callisto's surface lacks significant topography and is heavily cratered. Deficiencies in large (greater than 60 km diameter) impact craters compared with the Moon and terrestrial planets have been identified in Voyager data. However, additional large impact structures that are highly degraded have been identified in Galileo images.

Moderate resolution (0.4–1.1 km/pixel) Galileo images show more detailed structural features within three of Callisto's multiring systems (interpreted to be large impact features) than seen in Voyager data. The youngest, Valhalla, is characterized by bright central plains surrounded by discontinuous, concentric, arcuate structures (troughs and outward-facing scarps). The Asgard multiring system consists of a bright central plains unit surrounded by several discontinuous rings consisting of troughs, but lacks outward-facing scarps. The brightness of Asgard's central plains is enhanced by the presence of a more recent impact crater, Doh, which excavated icier (brighter) subsurface materials. High-resolution images of Asgard reveal fractures within and adjacent to the innermost ring. These fractures are probably the result of the impact-related rifting of Callisto's brittle outer layer. The proximity of these fractures (seen in the innermost ring only) to the impact crater Doh may indicate that these fractures were reactivated by the more recent impact event. The well-preserved outer troughs associated with Asgard have a lower albedo than the surrounding cratered plains.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Adlinda, the oldest of these three multiring structures, lacks a bright central plains unit. However, low-albedo concentric rings, which are more resistant to degradation, are visible. The detailed morphology of Adlinda is obscured by numerous impacts, as well as ejecta from Lofn, a nearby impact crater.

The Near Infrared Mapping Spectrometer (NIMS) on the Galileo spacecraft provided up to 408-channel spectra of Callisto's surface, from 0.7 to 5.2 μm at spatial resolutions ranging from one to hundreds of kilometers per picture element. Regions with strong water-ice absorption features in NIMS spectra usually correspond to bright features (e.g. fresh impact crater deposits) in clear-filter Galileo images. Weak absorption features near 3.88, 4.05, 4.25 and 4.57 μm are interpreted to be due to the presence of an S–H bond, SO_2, CO_2 and organic compounds (e.g. tholins) respectively. In general, correlations between these compounds and surficial geology are complex and poorly understood at present. Globally, the 4.25 μm absorption is stronger on Callisto's trailing hemisphere, while the 4.05 and 4.57 μm absorptions are stronger on the leading hemisphere. In addition to providing insights into surface-material compositions and their distributions, analysis of NIMS data has revealed the presence of a tenuous (surface = 7.5×10^{-12} bar) carbon dioxide atmosphere extending approximately 100 km above the surface. Based on models of CO_2 removal rates, the estimated lifetime of Callisto's atmosphere is about 4 years. Consequently, either this atmosphere is a transient phenomenon, or some process(es) is currently supplying carbon dioxide to the atmosphere.

Like GANYMEDE, Callisto is known as an icy satellite because its relatively low mean density (1836 kg m^{-3}) requires that it is made of about equal amounts, by mass, of ice and rock-metal. Callisto is only slightly smaller than Ganymede, but differs from it in a number of striking ways. Though the mean density of Callisto and Ganymede are similar, Callisto has a larger normalized axial moment of inertia $C/(MR^2)(C/(MR^2) = 0.359 \pm 0.005$, C is the axial moment of inertia, M is the satellite's mass and R is the moon's radius, known from radio Doppler measurements by the Galileo spacecraft), implying a very different interior structure for Callisto compared with the interior of Ganymede. Unlike Ganymede, Callisto is only partially differentiated and consists of a water-ice–liquid shell less than 350 km thick surrounding an essentially undifferentiated ice–rock–metal core. The rock and ice in Callisto are not completely separated and Callisto may not have a central metallic core. The gravitational data obtained by the Galileo spacecraft in a flyby past Callisto do not rule out the existence of a small metallic core in Callisto with a radius less than about 25% of the satellite's radius. However, geophysical arguments and other properties of Callisto make the existence of a small metallic core in the moon unlikely. Consistent with the absence of a metallic core, Callisto does not have an intrinsic magnetic field. The lack of evidence of any endogenic modification of Callisto's surface is consistent with the partial differentiation of only the outer few hundred kilometers of Callisto.

Though Callisto lacks its own magnetic field, the Galileo spacecraft has detected the magnetic perturbation associated with electromagnetic induction in Callisto. Induction occurs when the satellite experiences time periodic changes in the ambient magnetic field of the Jovian magnetosphere due to its orbital motion around Jupiter, the rotation of Jupiter, and the tilt of the Jovian magnetic field (see JUPITER: MAGNETOSPHERE). The magnitude of the electromagnetic induction signal requires that there be a global layer of high electrical conductivity in close proximity to Callisto's surface. The only explanation for such a layer is that it is a subsurface liquid water ocean with a salinity similar to that of terrestrial seawater. Unlike the surface of EUROPA, Callisto's surface provides no hint that a liquid water ocean might hide beneath. The induction signal is the only indirect evidence that Callisto might have a subsurface liquid water ocean. The existence of such an internal ocean in Callisto is difficult to reconcile with the lack of endogenic modification of its surface.

Callisto is not part of the orbital dynamical Laplace resonance of the inner three Galilean satellites (IO, Europa and Ganymede) and it is therefore not tidally heated. Without such a heat source it is difficult to understand how a subsurface liquid water ocean in Callisto could be prevented from freezing solid over geological time. Subsolidus convection controlled by the viscosity of ice is expected to efficiently cool Callisto's interior. At least in the case of Europa, for which there is geological evidence of an internal liquid water ocean in the past or at present, tidal dissipation is a possible source of heat to maintain a subsurface water layer in a liquid state. Tidal heating in the distant past may also explain why Ganymede is completely internally differentiated while Callisto is only partially differentiated. After all, both these moons are comparable in size and mass, have similar amounts of radiogenic heat sources and experienced similar amounts of accretional heating on formation.

Breaking news update (30 April 2002)

A recent image from NASA's Galileo spacecraft adds evidence to a theory that Callisto may hold an underground ocean. The image shows a part of Callisto's surface directly opposite from the Valhalla basin where Callisto was punched by a major collision. The opposition point shows no effect from the impact. Points opposite major impact features on some similar-size worlds, such as Mercury and Earth's Moon, show lumpy terrain attributed to seismic shocks from the distant impacts. The new image is consistent with a 1990s'

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

model proposing that a liquid layer could be acting as a shock absorber inside Callisto.

Bibliography

Anderson J D, Schubert G, Jacobsen R A, Lau E L, Moore W B and Sjogren W L 1998 Distribution of rock, metals, and ices in Callisto *Science* **280** 1573–6

Khurana K K, Kivelson M G, Stevenson D J, Schubert G, Russell C T, Walker, R J, Joy S and Polanskey C 1998 Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto *Nature* **395** 777–80

Schubert G 1997 Inside the solid planets and moons *Phys. World* **10** 45–9

G Schubert

J Klemaszewski

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Caloris Basin

A large multiringed impact structure on Mercury, centered at 30.5 °N, 170.2 °E. Its official name is Caloris Planitia, which means 'plain of heat'—it was in this region that the highest surface temperatures were recorded by the Mariner 10 probe, which imaged about half of the structure, the other half being in shadow.

Based on the observed half, the Caloris Basin has an estimated diameter of 1340 km. The outer ring is marked by the Caloris Mountains (Montes Caloris), a formation about 100–150 km wide consisting of smooth blocky units 1 or 2 km high and up to 50 km long. Within this ring the floor of the basin is covered with wrinkle ridges and fractures, many of which follow approximately the lines of concentric arcs. Other features are associated with the impact that formed the basin. The mountain ring is punctuated by small plains and secondary craters formed by the fallback of ejecta and impact melt. To the northeast of the basin lies the Van Eyck formation, containing a series of radial features extending for some 1000 km and believed to have been formed by ejecta. At the antipodal point of the impact (the point diametrically opposite, on the other side of Mercury) lies a terrain not seen anywhere else in the solar system: a jumbled, fractured landscape consisting of hills up to 2 km high and 10 km across, and depressions that criss-cross older features. The impact sent seismic surface waves around the planet and compression waves through its core. Where these waves came to a focus at the antipodal point, the crustal rock is calculated to have been shock-fractured and forced 1 km upwards. The present landscape was formed when the upheaval subsided.

The Caloris Basin is estimated to have been formed by the impact of a 150 km object about 3.85 billion years ago, at the end of the period known as late heavy bombardment which marks the last major episode of impact cratering in the early solar system. That it was the last major impact on Mercury is testified by the below average crater density on the basin's floor.

See also: Mercury.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Caltech Submillimeter Observatory

The Caltech Submillimeter Observatory (CSO) operates a Leighton 10.4 m diameter telescope on the summit of Mauna Kea, Hawaii at an elevation of 4200 m. Construction and operations funding is from National Science Foundation grants to the California Institute of Technology. In partnership with the University of Texas and the University of Hawaii, CSO provides the international astronomical community with a state-of-the-art observing facility at submillimeter wavelengths (300–1000 microns), mainly for studies of molecular and atomic lines, and continuum emission from dust in star forming regions in the interstellar medium of the Milky Way and distant galaxies.

For further information see http://www.cso.caltech.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Camelopardalis

(formerly sometimes spelled Camelopardus; the Giraffe; abbrev. Cam, gen. Camelopardalis; area 757 sq. deg.) A northern constellation which lies between Auriga and the north celestial pole, and culminates at midnight in late December. It represents the animal upon which Rebecca rode in Canaan to become the wife of Isaac, and was introduced by the Dutch theologian and geographer Petrus Plancius, who included it on a celestial globe in 1613.

An inconspicuous constellation, the brightest star in Camelopardalis is β Camelopardalis, a wide double star with yellow (G0) and bluish-white (A5) components, magnitudes 4.0 and 7.4, separation 81″. Other interesting objects include the U Geminorum star Z Camelopardalis (range 10.0–14.5, mean period about 22 days) and NGC 1502, an open cluster of about 45 stars fainter than seventh magnitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cameras

Readers of this article are using a natural camera, the human eye. It has a simple lens but phenomenal signal processing which is a study in itself and is not the subject of this article. More relevant examples of cameras are those using photographic film as the detector. Other examples are video cameras that have a CHARGE COUPLED DEVICE (CCD) as the detector.

A telescope is a big camera, but the larger telescopes use a concave mirror for the first optical element instead of a convex lens. Whether a mirror or a lens, this optical element is called the objective. If it is a mirror, it is also called the primary mirror. A telescope system is a series of cameras, the last of which has a detector at its focus. It is sometimes only the last camera that is called the camera.

Figure 1 shows the basic optical design of all telescope systems, in its simplest form consisting of three singlet lenses. Each of these lenses functions as a major component of a telescope system, which on a modern telescope would be more complex than a singlet spherical lens.

Optical aberrations and reflection losses

The rays shown in figure 1 are accurately traced, but the layout is drawn with a larger scale in Y than in X to show the separation of rays more clearly. Obviously, the rays do not come to a sharp focus. SPHERICAL ABERRATION, SA on the drawing, is apparent at the axial focus. The outer rays focus first and are already diverging where other rays come to focus. COMA is also apparent on the drawing in the off-axis image. The ray that passes through the center of the entrance APERTURE, called the chief ray, is laterally displaced from the focal point of the outer rays. The image has the appearance of a comet with the chief ray at its head. It is also apparent that the two foci are not in the same plane, which is an ABERRATION called Petzval curvature. Another aberration is ASTIGMATISM where outer rays perpendicular to the plane of the drawing focus further from the lens, but those rays are not shown. A tutorial on aberrations can be found in the Melles Griot catalog: http://www.mellesgriot.com.

For a given FOCAL LENGTH, spherical aberration is proportional to the cube of the diameter of the aperture. Coma is proportional to the square of the aperture and directly proportional to the field angle. Astigmatism is proportional to the square of the field angle and is directly proportional to the aperture. Axial color is directly proportional to the aperture, whereas lateral color is independent of the aperture and is directly proportional to the field angle.

For a singlet lens, spherical aberration is less if the glass has a high index of refraction, N, because less curvature is needed for a given focal length. However, higher-index glasses usually have a higher dispersion that increases the CHROMATIC ABERRATION.

A single-layer antireflection coating on a lens made of high-index glass has less residual reflectivity at one wavelength, but more average loss than for a low-index glass over a wide wavelength band. Thus a high-index singlet lens is optimum for monochromatic performance, but a low-index, low-dispersion glass is better for a wide spectral region. Another consideration in choice of a glass is radiation darkening and fluorescence.

Optical layout

The first element in figure 1 is labeled the objective. It is shown as a singlet lens, even though such a simple objective has not been used since 1733. The objective is shown focusing beams of light from two stars, each beam consisting of parallel rays, onto the first focal surface. It is said to have positive power because it changes a cylindrical beam of light into a converging cone of light. The rays cross at this first focus, thus the converging cone becomes a diverging cone which is intercepted by a collimator lens which changes the diverging beam back into a parallel, collimated beam. This lens also has positive power but its function is the reverse of the OBJECTIVE LENS. The collimator lens has a shorter focal length than the objective, thus the collimated beam is smaller in diameter than the beam incident onto the objective lens. This lens also produces an image of the entrance aperture (which in figure 1 is at the objective lens). This image is called a pupil. It is also called an aperture stop, or a Lyot stop, if a black plate is located there with a hole equal to the diameter of the beam, to transmit the beam but block unwanted, stray light. The camera lens in figure 1 re-converges the beam to focus onto the detector at the same rate of convergence as at the first focus; thus it has unit magnification. If the focal length of the camera lens were tripled, the magnification, M, would be three times, thus $M = 3$.

The portion of the enormous parallel beam from a star that can enter the telescope is limited by the diameter of the entrance aperture stop, which in figure 1 is at the front surface of the objective lens. This stop is not shown; it is the mechanical structure around the lens. The ray from each star, or object point, that passes through the center of the aperture stop is called the chief ray from that object. Thus, the chief ray also passes through any image of the entrance aperture, such as the pupil between the second and third lenses. Rays that pass though the edges of a lens are called marginal rays.

In figure 1 a marginal ray from the edge of the field intersects the collimator lens far from its center resulting in such a large blur of the pupil and of the final image that is it noticeable on the drawing. If a positive lens were located at the first focus it would move the rays closer to the center of the collimator lens and would move the pupil closer to the collimator lens, but would not change the location of the final focus. Such a lens at an intermediate focus is called a field lens.

It is mostly what is inserted between the second and third lens elements in figure 1 that distinguishes one instrument from another. The inserted device is called a processor. It is sometimes coincident with the pupil but not necessarily.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

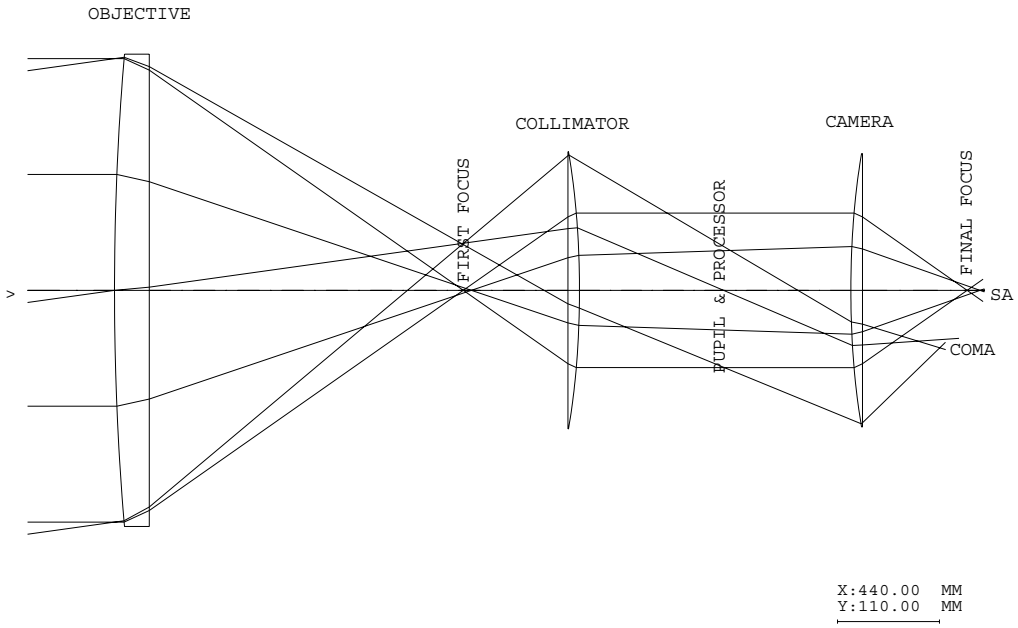


Figure 1. Basic optical layout of all telescope systems.

There are many options for the optical processor, each option producing a different telescope system. Some systems have two or more instruments in series, such as an ADAPTIVE OPTICS module followed by an imaging SPECTROGRAPH.

Some processor options

If the processor is a human eye inserted at the location of the pupil, the camera lens is not needed.

If the processor is a high-quality very flat, thin pair of windows that cause waves of light to add or subtract, the instrument is an INTERFEROMETER. If the processor is a polarizer, such as the N-shaped Wollaston prism, that causes the light waves to oscillate in a single plane, the instrument is a polarimeter (see POLARIMETRY).

If the processor is two N-shaped prisms which can be counter-rotated to increase or decrease the amount of DISPERSION to match that of the atmosphere, the processor is an atmospheric dispersion compensator, ADC.

Often, the processor is a DIFFRACTION GRATING, which changes the direction of the light by a deviation angle, and also disperses the light more or less than the mean deviation angle depending on the wavelength. Because of the deviation, the camera lens must be moved around to intercept the center of the turned beam. If the grating is on a mirror, called a reflectance grating, it would reflect the light back towards the collimator. If reflected straight back onto the collimator, that lens would also serve as the camera lens in an arrangement called a Littrow spectrograph. Usually, the tilt of the grating is chosen so that the reflected beam lies beside the collimator. The camera lens would be rotated around and positioned at that location for this instrument. Also, a field stop would be placed at the first focus in the form of a single, long slit, for a single-slit spectrograph, or by an array of short slits, one on each stellar image, for a multi-object spectrograph, MOS.

If a deformable mirror, a DM, is used instead of a reflectance grating, the system is an adaptive optics module, AOM. Some systems have an AOM and a spectrograph in series, thus there would be a second collimator and camera in the light path, in which case the final focus is the third focus.

If the processor is a simple inverted triangle-shaped prism the beam is turned through a deviation angle but the dispersion is less than from a transmission grating. Therefore, if a transmission grating is bonded to the prism such that the deviation angles cancel, then the beam is not deviated but it is still dispersed. Thus the camera lens does not need to be moved. Such a combination is called a grism, which is usually located before the first focus; thus there can be no slit, the system being called slitless spectroscopy. The surface of the prism that is not bonded to the grating can be curved, and can thus act as a lens, which in the case of the Canada France Hawaii Telescope (CFHT), is part of the prime focus corrector lens near the focus of the primary mirror. This device is sometimes called a grens.

If instead of placing a slit at the location of each stellar image, a fiber is positioned there and bent around to have its exit on the long slit at that focus the result is a fiber-fed multi-object spectrograph. An alternative is to locate an array of small, microlenses or mirrors at the first focus which segments a portion of the field and refocuses each segment along the long slit; such a device is called an image slicer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

More complex objectives

The achromatic lens was invented in 1733 by Chester Moor Hall who used it in a telescope with a 3.5 cm aperture. A positive achromat combines two types of glass in the lens: a convex lens of crown glass combined with a concave lens of flint glass. Flint glass has a higher dispersion than crown glass. A weak flint lens with a concave surface, thus with negative optical power, combined with a convex crown lens with an equal but opposite amount of chromatic aberration, results in a net positive power but with the same focal length at two different wavelengths.

The biconvex and meniscus shapes of the elements serve to reduce spherical aberration to less than that of a singlet lens. (Warning: to have low spherical aberration the achromat must be mounted in the correct orientation with object and image distances close to those for which the lens was optimized.)

Achromatic lenses suffer from secondary chromatic aberration because the focal length at intermediate wavelengths differs from that at pairs of wavelengths where it is the same. The apochromat is a lens with three different glass types. It can have the same focal length at three different wavelengths. It suffers from less secondary (or residual) chromatic aberration than an achromat but it is still noticeable with a large lens.

If in figure 1 a negative lens, which is concave, were added in front of the first focus by a distance equal to its focal length, a collimated beam would be produced, thus this lens would act as an eyepiece. This is the arrangement used in the first telescope by Galileo in 1662. The image quality was very poor.

If the negative lens were moved closer to the focus to one-half of its focal length, the beam exiting from it would no longer be collimated but would converge to focus beyond the first focus by a distance equal to one-half the focal length of the lens. Therefore, the F-number would be doubled thus the magnification would be two times in this example. However, the distance from the objective lens would not be doubled but the effective focal length, EFL, which is the F-number at the focus multiplied by the diameter of the objective lens, would be doubled. Such an arrangement is called a telephoto lens. Its advantage is that it is more compact than if the objective were simply given the longer focal length.

Reflective objectives and space telescopes

Major advantages of using MIRRORS in a telescope are that they are free of chromatic aberrations and need to be optically polished on only one side of the blank.

The telephoto equivalent in a reflecting telescope is to have a convex secondary mirror located before the prime focus of the larger concave primary mirror. Such a two-mirror system is called a classical CASSEGRAIN TELESCOPE if the primary mirror is parabolic, or a Ritchey–Chretien if the primary mirror is hyperbolic. Most large telescopes are of this type.

However, the planned Next Generation Space Telescope, NGST (figure 2) will have a third mirror added following the first focus which forms an image of the pupil in a converging beam which then comes to focus with exceptionally sharp images over a field of view of 6 arcminutes. In order for the beam from the secondary to the tertiary to have clearance to one side of the pupil, the field of view used is restricted to one side of the axis, thus a larger maximum field angle is needed. This increases aberrations, but the three-mirror telescope still has sharper images than a two-mirror telescope (Ritchey–Chretien) and also less field distortion and less curvature of the focal surface. The optical layout of the NGST is shown in figure 2. Compared with a two-mirror telescope, which has a distant exit pupil above the secondary mirror, the NGST has its final pupil much closer to the focus and is thus less telecentric than most telescopes. The purpose of the deformable mirror, DM, on NGST is to compensate for residual imperfections in the figure of the active primary mirror (see SPACE INSTRUMENTATION: NEXT GENERATION SPACE TELESCOPE).

If a space telescope is diffraction limited, then the cameras in the instruments can be smaller and have longer focal lengths than the equivalent on a large ground-based telescope whose resolution is limited by atmospheric turbulence, called SEEING. A spectrograph designed for a small ground-based telescope will work well with a large space-based telescope.

Back-surface mirrors

Large mirrors in telescopes are front-surface mirrors, but for smaller mirrors there are advantages to back-surface mirrors. A common example is a right-angle prism. In that case no metal coating is required because the reflection at the diagonal is by total internal reflection. If the angle of incidence of the light is lower than the critical angle it is necessary to have a reflective coating on the outside of the back surface, such as in a bathroom mirror. This reflective surface can be well protected against corrosion and dust. If the mirror is curved with a shorter radius on the front than on the back, it is called a Mangin mirror. The spherical aberration caused when light passes through the front is balanced by that caused by reflection at the spherical back. It is used as a primary mirror in some small telescopes.

If the glass is made thick and nearly flat but aspherized, the result is called a solid Schmidt camera.

An array of small back-surface mirrors is used in the image position sensor for the adaptive optics system for the WIYN telescope on Kitt Peak. It is described at http://claret.kpno.noao.edu/wiyn/wttm.html.

Also, an image-slicer design option for NGST uses arrays of back-surface micromirrors. An advantage is that a change of temperature changes both the curvatures of the mirrors and their separations so they stay in focus.

Eyepieces

The EYEPIECE is a reversed camera lens but in detail it has a very different task from that of the objective or of a commercial camera lens because of the location of the pupil. In the case of the eyepiece, the pupil must be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

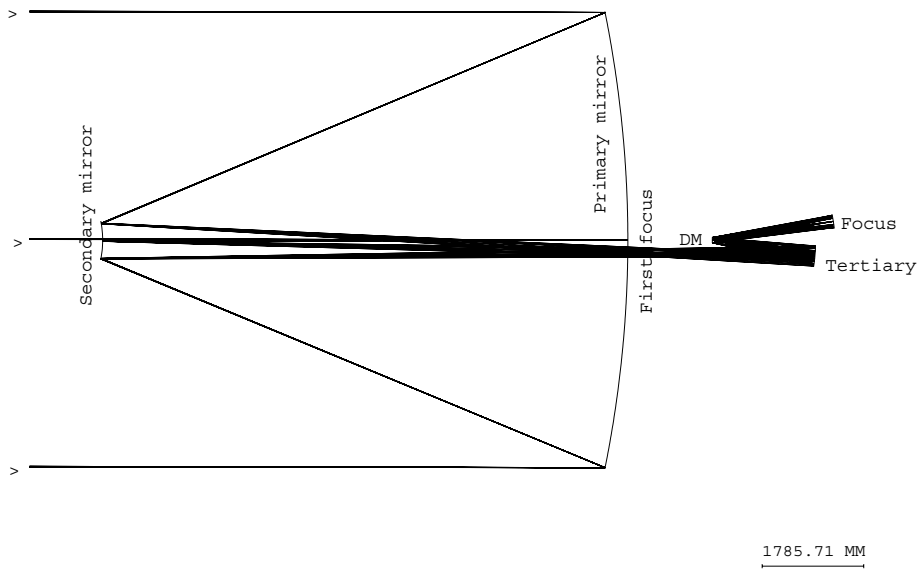


Figure 2. NGST: primary, secondary, tertiary and deformable mirrors.

external to the lens, because it must be on the eye of the beholder, which with eyelashes is 10 to 15 mm from the lens, or 18 to 20 mm if the person wears spectacles. This distance is called EYE RELIEF. Also, the diameter of the EXIT PUPIL should be small enough for all of the beam of light to enter the human eye, whose iris ranges in diameter from 2 to 5 or 8 mm depending on age. Another complication in the optical design of an eyepiece is the increased FIELD OF VIEW that is in inverse proportion to the diameter of the pupil. If the diameter of the telescope is 1 m and if it has a one-third degree field of view, the field of view of the eyepiece with a 5 mm pupil must be 200 times greater, thus 67°, requiring a complicated design with several elements. The required focal length of the eyepiece is directly proportional to the F-number of the focus being viewed, thus is twice as long at an F-20 focus as at an F-10 focus.

Modern computer designed eyepieces have up to eight elements of several glass types and up to an 82° field of view which is more than the eye can take in at once, and gives the impression of being in space. An example of a modern eyepiece, the Erfle, which has five elements, is in figure 3. A description of eyepiece designs is at the Australian web site; http://www.users.bigpond.com/PJIFL/INDEX.html.

Fiber feed

Fibers have the advantage in multi-object spectroscopy of selecting objects from anywhere in the field of view of the telescope and aligning them along the slit of a spectrograph. However, fibers suffer from focal ratio degradation, FRD. This disadvantage is greater on telescopes that produce sharp images, such as space telescopes. FRD is less of a disadvantage on ground-based telescopes where the image quality is already degraded by turbulence in the Earth's atmosphere, or seeing.

Telescope location: Earth or space

Because of atmospheric turbulence, called seeing, the image of a star at the focus of a ground-based telescope system is enlarged, thus a shorter EFL is required to match the sharpness of the image to the resolution of the detector. Thus a faster (smaller) F-number is needed than if the telescope were in space. For a large ground-based telescope, the optical elements are not only very big but also the camera is more complicated, requiring more elements to correct the optical aberrations. Figure 4 shows an outstanding spectrograph camera lens designed by Harland Epps for the Keck 10 m telescope. It has nine elements: a liquid-coupled doublet, a singlet, a liquid-coupled triplet, a doublet and finally a singlet field flattener lens near the focus. Also shown are a filter and a dewar window. The largest element has a diameter of 36 cm and is made of calcium fluoride.

By comparison, if a telescope is in space the minimum size of a stellar image can be limited only by diffraction, and thus is the same size at the focus on the detector whatever the diameter of the telescope, for a given F-number. Thus, for example, a simple spectrograph camera designed for a 1 m telescope would operate at its full efficiency if used with an 8 m space telescope. This makes more feasible the use of efficient spectrographs on giant space telescopes of the future, such as perhaps a 100 m liquid mirror telescope on the Moon.

Diffraction limit

The diffraction limit expressed as the full diameter, ϕ of the AIRY DISK, which is the diameter of the first dark diffraction

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

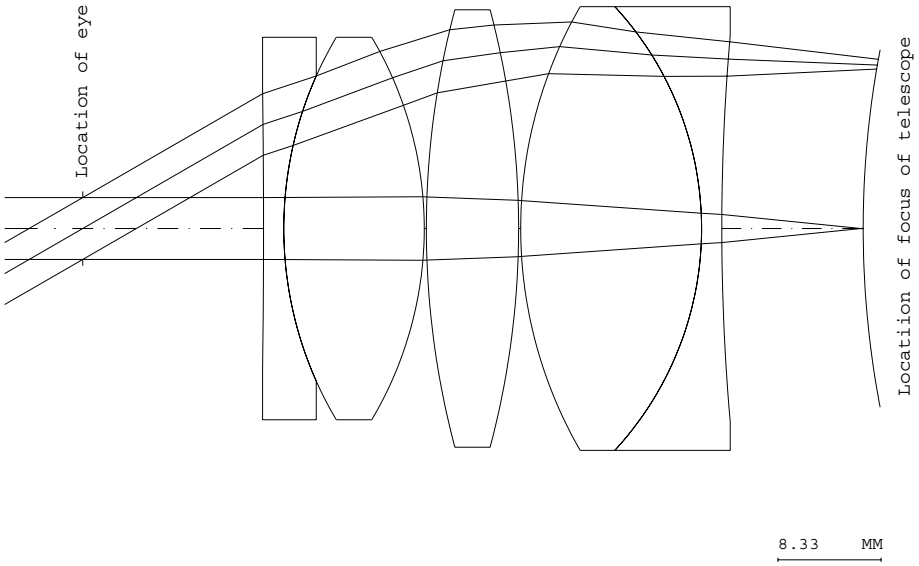


Figure 3. Erfle eyepiece, 5 mm pupil, 60° field, EFL = 25 mm.

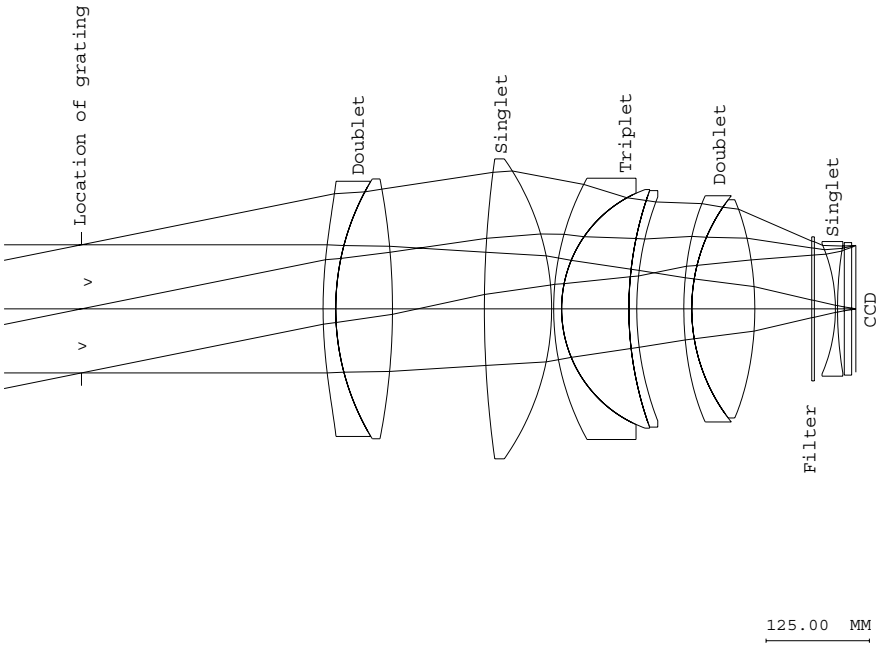


Figure 4. EPPS spectrograph camera lens for the Keck telescope.

ring, in seconds of arc, is

$$\phi = \frac{\lambda}{2D}$$

for λ in microns and D in meters. For the 8 m NGST, at $\lambda = 2\ \mu$m (near infrared) $\phi = 0.25$ arcsec. The Rayleigh limit is one-half of this amount, 0.125 arcsec, which is the half-width of the Airy disk.

Adaptive and active optics

An active optic cell can bend the mirror it supports to correct for imperfections in the shape of the mirror, such as the primary mirror. By comparison an adaptive optic is a deformable mirror, DM, located closer to (but not at) the focus where the beam is much smaller than at the primary mirror. The DM is rapidly deformable to correct aberrations caused by turbulence in the Earth's atmosphere, and it also corrects wavefront errors caused

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by imperfections in the primary mirror. The first non-military adaptive optics system, called Come-on, started operation at the European Southern Observatory in 1989 on the 3.6 m telescope. The newest, improved system is called Adonis. In 1996 the adaptive optics Bonnette, or PUEO, was commissioned on the Canada France Hawaii Telescope, CFHT. These systems are described in http://www.ls.eso.org/lasilla/Telescopes/360cat/adonis/html/aosystem.html#optic and http://cdsweb.u-strasbg.fr:2001/Instruments/Imaging/AOB/performance.html.

An adaptive optics system, like a spectrograph, has a collimator followed by a camera, usually off-axis mirrors to avoid obstruction. Instead of a grating or prism, there is a deformable mirror, DM, between the collimator and the camera. In most systems the collimator and camera mirrors are parabolic, one for the collimator and one for the camera. With only two powered mirrors, the focal surface is curved, but not sufficiently to significantly degrade the resolution of a field diameter of a few arcminutes. Because the beam between the two paraboloids is collimated, the image at the focus on the axis at the center of the field is perfect, whatever the focal lengths of the two paraboloids and whatever the distance at which they cut off the axis of the parent mirrors. The focal surfaces are tilted with respect to the axis of the paraboloids but are not tilted with respect to the axis of the telescope. Thus it is not necessary to tilt the detector when the adaptive optics system is inserted.

It is often required that the camera should have a longer focal length than the collimator to magnify the sharpened images to match the resolution of the detector. In that case, field aberrations (of the center of the field of view) are minimized if the distance of the camera mirror from the axis of the parent mirror surface is increased in proportion to the square of the magnification.

The Gemini adaptive optics system is unusual in that the module available for the adaptive optics system (on a de-rotator needed because the telescope has an alt-azimuth mounting) is further from the focus than the F-number multiplied by the diameter of the DM, thus the camera must be in the form of a reverse telephoto—it must increase the diameter of the beam from the DM before convergence. Thus the Gemini system, called Altair, has two mirrors in the camera, convex then concave, in addition to the collimator mirror, thus it is a type of three-mirror anastigmat, TMA. An advantage of having three mirrors instead of two is that the focal surface is flat, thus the images are sharper on a flat detector. It is described at http://www.dao.nrc.ca/facilities/gemini/altair/altair.html.

The wavefront from a natural guide star is analysed to generate the signal to deform the DM. A sodium laser star in the Earth's atmosphere will also be analysed. Because it is out of focus where the natural star is in focus within the central shadow (of the secondary mirror) in the laser beam, a pinhole mirror permits very efficient separation of the two beams in the Gemini design.

A system at Mount Wilson is the first to use an ultraviolet laser guide star, described at http://www.astro.uiuc.edu/projects/unisis/.

The simplest adaptive optics systems, such as for the WIYN telescope on Kitt Peak, corrects only for jitter of the image, called tip-tilt. Surprisingly, the NGST will have a simple adaptive optics system to correct for jitter of the satellite which is expected to be as much as ± 2.7 arcsec.

Glass catalogs

When a ray of light is incident in air upon a flat piece of glass, the direction of the ray is changed inside the glass. If the angle of incidence, measured with respect to the normal (perpendicular) to the glass surface is α, and the angle after refraction is β, the index of refraction, N, is defined as

$$N = \frac{\sin \alpha}{\sin \beta}.$$

The index of refraction, N, is inversely proportional to the speed of light in the glass compared with the speed in a vacuum. However, it is measured with respect to the speed in air, thus a small correction is needed to the indices listed in glass catalogs if the material is used in a vacuum.

Catalogs list about 250 different types of glass. The oldest manufacturer is Schott in Germany. A more recent supplier is Ohara in Japan (http://www.oharacorp.com/index.html).

Glasses from both manufacturers are used in telescopes. Hoya is another major supplier. Glasses are identified by their index of refraction N_d, at the yellow helium spectral line whose wavelength is 587 nm, and by the difference in N between that of the blue hydrogen line, N_f at 486 nm and the red hydrogen line, N_c at 656 nm, listed as $N_\mathrm{f} - N_\mathrm{c}$ in the catalogs. A third, very important identification is the Abbe value, which combines both the index of refraction and the dispersion,

$$V_\mathrm{d} = \frac{N_\mathrm{d} - 1}{N_\mathrm{f} - N_\mathrm{c}}.$$

Thus a low dispersion glass has a high V_d number. Each glass type has a six digit code number: the first three digits of $N_\mathrm{d} - 1$, followed by the first three digits of V_d. Each glass type is also given a name which differs depending on the manufacturer. For example, the common borosilicate crown (low dispersion) optical glass 517642 is Schott BK7, Ohara S-BSL& and Hoya BSC7. The common flint glass 620364 is Schott F2, Ohara PBM2 and Hoya EF2.

A diagram plotting N_d versus V_d, listed as an $N_\mathrm{d}/V_\mathrm{d}$ diagram, for catalog glasses is available on the web at http://www.schottglasstech.com/. The transmission of catalog glasses depends on type. It is listed for the wavelength region from 0.28 to 2.4 μm, but most of the glasses transmit over only a fraction of that region.

Special materials that transmit further into the ultraviolet and infrared are not included in these catalogs because they are not conventional glasses. Calcium fluoride which is now available in large pieces, transmits

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

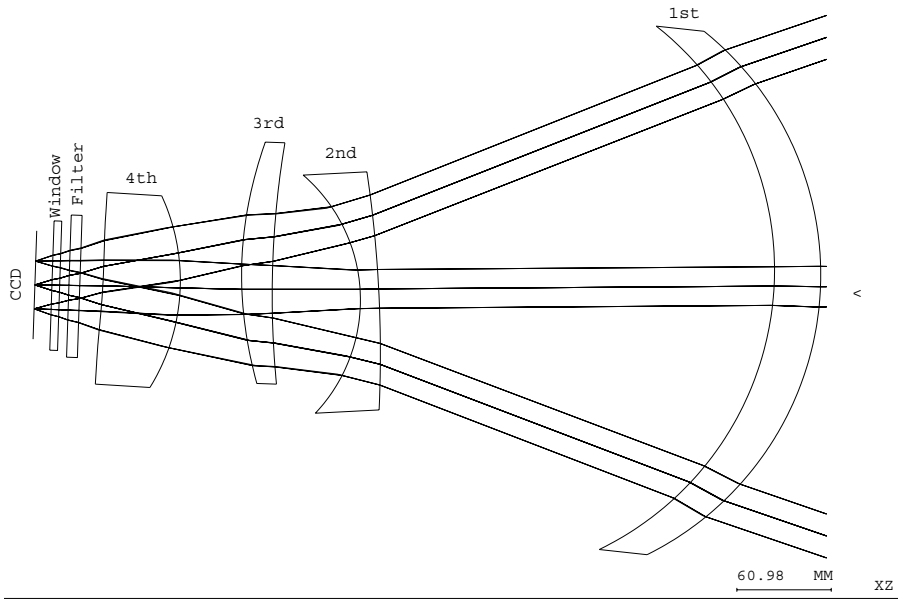


Figure 5. Four-element corrector lens for the Hickson 6 m LMT.

over a much wider wavelength region, from 0.23 to 7.8 μm. It has a very low index of refraction and dispersion: $N_d = 1.4$, $V_d = 96$. Zinc selenide has a very high index of refraction: $N_d = 2.6$, $V_d = 8$.

Correctors for liquid mirror telescopes

By rotating a thin film of mercury a parabolic surface of optical quality is formed serving as the primary mirror of a liquid mirror telescope, LMT. Because of the naturally parabolic shape, images at the center of the field of view are sharp. Several LMTs with diameters ranging from 2.4 to 3 m are operating in North America and a 6 m LMT is being constructed in Canada. Ermanno Borra at Laval, Paul Hickson at UBC and Mark Mulrooney at Cloudcroft Observatory, New Mexico, are instrumental in the development of LMTs.

The 6 m LMT being built at the University of British Columbia will have a very short focal length: 9 m thus $F/1.5$. This fast focal number is achieved without additional cost for the mirror simply by designing its mounting for a higher rate of rotation, and thus permitting a smaller enclosure. However, the short focus increases the difficulty of correcting the coma and astigmatism. The well known corrector lens of the type originated by Ray Wilson in Germany and by the late Charles Wynne in England consists of three lenses: a positive meniscus lens followed by a negative lens and ending with a positive lens near the focus. Referred to as a prime focus corrector lens, they are used on many telescopes. They have a flat focal surface but have considerable pincushion distortion that was not found to be objectionable in the days of photographic detectors. After the development of the CCD permitting much fainter star fields to be imaged it became advantageous to minimize field distortion to simplify the subtraction of background light from the sky. Such correctors have four elements instead of three.

A survey of LMTs around the world is given at http://www.astro.ubc.ca/lmt. These inexpensive telescopes are used for tracking space debris, for aeronomy and for astronomy.

For an LMT there is an additional need to control distortion because the time exposure of moving images on the CCD detector is achieved by time delay integration, TDI, also called drift scanning. With conventional CCDs the electronic readout is at constant rate along each column of pixels, thus the images should move at a constant speed and in straight lines which requires zero field distortion when observing space debris or satellites. However, when observing stars with an LMT that is not located on the equator of the Earth, the rotation of the Earth causes the images to move in a curved path and at a rate that depends on the declination of the object. Fortunately, this sidereal distortion can be removed by an asymmetrical distortion of the corrector achieved by deliberately decentering and tilting the elements of the corrector lens. The layout of such a lens is shown in figure 5 where the elements are positioned for a declination of 49°. For observation of satellites zero distortion is restored by removing the decentering and tilts of the elements. If the telescope were moved to Cloudcroft Observatory at a latitude of 31° the reduced sidereal distortion for that site could be achieved by appropriate repositioning of the elements.

An advantage of drift scanning is that variations in pixel-to-pixel sensitivity of the detector are reduced because the response is averaged over all CCD pixels in a column. Steerable telescopes are sometimes used for drift scanning by tracking along a great circle at a non-sidereal rate and rotating the detector. In this case, nonlinear

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

effects are minimized without requiring an asymmetrical corrector. However, because of the physical motion required of the camera the length of the great-circle track is limited. The observations would need to be interrupted while the camera position is set to follow different great-circle paths. By comparison, the LMT, or any telescope stopped at the zenith, can integrate continuously with a stationary detector.

E Harvey Richardson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Campanella, Tommaso [Giovan Domenico] (1568–1639)

Born in Stilo, Calabria, Italy. A Dominican, he wrote on Telesian philosophy and astrology, and cast flattering horoscopes for the influential. A great admirer of Galileo, in 1622 he published his *Apologia pro Galileo* ('*Defense of Galileo*'), defending the Copernican system and the separate paths of scripture and nature to knowledge of the Creator. He adopted an animistic, yet empirical, interpretation of the world, which influenced a number of followers. Denounced to the Inquisition, he spent nearly half his life in prison in Naples and Rome, sometimes leniently, with a salary and freedom to write, but also under the worst conditions, with torture. In 1634 the Inquisition discovered yet another plot and Campanella was implicated. He fled Rome before he could be arrested, was well received in France, but soon died in Paris.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Campani, Giuseppe (1635–1715)

Italian instrument and clock maker, born in Castel San Felice (near Spoleto), Italy, *fl*. Rome, best known for his optical instruments, primarily telescopes (for which he made the best composite eyepieces available and lenses of longer focal length than any other optician—up to 40 m), but also for microscopes. All JEAN CASSINI's discoveries were made with Campani telescopes, with which Cassini exclusively equipped the Observatory in Paris.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Campanus of Novara [Campano, Giovanni] (1220–96)

Astronomer, mathematician, born in Novara, Italy, chaplain to Pope Urban IV. Wrote *Theorica Planetarum* which describes the construction of a planetarium, the first to be described by a European. The work matches detailed calculations of the positions of the planets to the description of the motion of the mechanism.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Campbell, William Wallace (1862–1938)

Astronomer, born in Hancock County, OH. Trained as an engineer, became director of Lick Observatory, measured stellar radial velocities with the Mills photographic spectrograph (which he designed) and published them (with Joseph Moore) in a catalog (1928). From studies of the Martian atmosphere, he deduced that it could not support life. Founded the Lick southern station in Chile, discovered numerous (339) spectroscopic binaries (including Capella), and determined the motion of the solar system with respect to surrounding stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canada–France–Hawaii Telescope

The Canada–France–Hawaii Telescope (CFHT) is a 3.6 m optical-infrared telescope located on the summit of Mauna Kea on the island of Hawaii. On 20 June 1974, representatives from the National Research Council of Canada (NRC), the Centre National de la Reserche Scientifique of France (CNRS) and the University of Hawaii (UH) signed the Tripartite Agreement that culminated in the dedication of the facility on 28 September 1979. Time is allocated to the partners in proportion to their financial support, with Canada and France each receiving 42.5% of the time. The Observatory headquarters is located in the town of Waimea and houses approximately 50 people.

The telescope has interchangeable top-ends supporting prime-focus ($f/4$), Cassegrain ($f/8$ and $f/35$ IR) and Coudé foci with a large complement of imaging and spectroscopic instruments.

The CFHT has pioneered much of the methodology used to enhance image quality on ground-based telescopes. This began with dome environment control, through to the first operational tip-tilt wave-front correction system (1988), and the deployment of a facility adaptive optics system in 1996. More recently, CFHT has implemented a wide-field imaging plan. It operates CFH12K, the largest close-packed CCD mosiac camera in the world and will deploy a full square-degree CCD camera in mid-2001.

For further information see http://www.cfht.hawaii.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canadian Astronomy

Canada is big (second only to Russia in area) and sparsely populated (30 million). These facts, as trite as they are, do explain a lot about the country, even its scientific endeavors. Almost all astronomy carried out in Canada during centuries of exploration prior to 1900 was connected with surveying and time-keeping. Even the efforts by Sandford Fleming to introduce worldwide time zones in the 1880s seem appropriate in a country which spans nearly 90° in longitude. Some 20th century examples are C S Beals' identification of dozens of ancient meteorite craters in the Precambrian Canadian shield, the launching of the communications satellite Alouette I in 1962 (Canada was the third country in the world to build a satellite) and the pioneering work in long-baseline radio astronomy whereby two Canadian radio telescopes separated by thousands of kilometers were linked in 1967.

Another cliché of life in Canada is the cultural and economic dominance of its giant neighbor, the United States. Some of America's best-known astronomers of the 19th century, including SIMON NEWCOMB, were born in Canada but went south to the land of opportunity. The 'brain drain' is still a concern today. In fact the Natural Sciences and Engineering Research Council (NSERC), the government-supported granting agency, has programs to encourage bright young scientists to study and work in Canada.

Finally, it should be recognized that Canada is officially bilingual. This means that all federal services are available in both French and English. It also explains why some acronyms such as CASCA (Canadian Astronomical Society/Société canadienne d'astronomie) reflect this dual character. In practice, among institutions where astronomy is taught, the Universités de Montréal and Laval in Québec operate in French, the Universities of Moncton in New Brunswick and Laurentian in Ontario use both official languages and the rest are English language based.

Early history

Very little has been written about aboriginal astronomical traditions in Canada except for a recent book, *The Arctic Sky*, by John MacDonald, which explores the knowledge of the Inuit people of the far north. It was also in the Arctic that European astronomy first touched Canadian shores. English explorers intent on a northwest passage to the orient looked to the Sun, Moon and stars to fix their latitude and longitude. The best pre-telescopic results were achieved by Thomas James who wintered in 1632 in the bay which now bears his name. However, it was the Jesuit missionaries from France, who often spent the best part of their lives here, who really introduced astronomy in Canada. They too were concerned with geographic position and to that end observed eclipses. The most accurate of these was a lunar eclipse observation on 30 January 1646, made by Father Joseph Bressani using a small telescope at a remote mission near present-day Midland, Ontario. The missionaries at Quebec recorded positions of the bright comets of the 1660s and about the same time they, along with some lay teachers, began teaching surveying, navigation and hydrography.

The tradition of practical astronomy as the basis of mapping continued after the capture of Quebec by the British in 1759 both in the Atlantic region and in the vast western territories of the Hudson's Bay Company. Scientifically significant were the transit of Venus in 1761, observed in Newfoundland by John Winthrop from Massachusetts, and the 1769 transit observed by surveyor Samuel Holland and his assistant at two locations in Quebec and on behalf of the Royal Society by William Wales at Fort Churchill on the shores of Hudson Bay. On Wales' advice, the Company appointed its first surveyor, Philip Turnor, who mapped much of western Canada along with two younger men, Thompson and Fidler, to whom he taught the procedures of practical astronomy while they were stationed at Cumberland House, Saskatchewan.

Eastern Canada became more settled following the arrival of Loyalists from the newly independent United States. In 1789, they founded the first English-speaking institution of higher learning at Windsor, Nova Scotia, naming it King's College. Although several other universities have their origin in the early 19th century, the first to be equipped with an observatory and to offer lectures in astronomy was at Fredericton, New Brunswick, in the 1850s. Other early observatories include the Toronto Observatory (originally a magnetic observatory, opened in 1840), Charles Smallwood's private observatory near Montreal, equipped with a 18 cm Fraunhofer refractor in the late 1840s, and observatories with 15 cm refractors at Quebec and Kingston established in the 1850s.

By 1867, when four colonies formed the new Dominion of Canada, some interest in 'pure' astronomy had begun to emerge, although preoccupation with the very real needs of surveying and time-keeping continued. Reverend Bolduc at the Quebec Seminary taught astronomy and had at his disposal a 40 cm Foucault reflector, and astronomy was part of the curriculum at Queen's University, Kingston, and at King's College, Windsor, Nova Scotia. E H Ashe, of the Quebec Observatory, traveled to Iowa to study and photograph the solar eclipse of August 1869. Smallwood, in Montreal, wrote on lunar and solar phenomena. One event which ultimately had lasting significance was the formation of the Toronto Astronomical Club in 1868 by a little group of amateurs. In the 1880s, one of their number was observing spectra of the Sun and stars, some were taking photographs of the Sun and Moon and others still were observing and sketching surface features on the planets. By 1890, they had incorporated as a Society and began publishing transactions which were distributed to universities and observatories around the world. This group evolved into the ROYAL ASTRONOMICAL SOCIETY OF CANADA (RASC), whose *Journal* still remains the only

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

national publication for Canadian astronomical research and news.

One of the members of this group in the 1890s was a young physics professor at the University of Toronto (U of T), C A Chant, who had the foresight to ally astronomy at the university with physics. W F King, who had worked on the 49th parallel boundary survey in the 1870s and who rose to the position of Canada's first 'Chief Astronomer' in 1890, also recognized that research was expanding beyond the traditional fields of time and position. So it came about in the early years of the 20th century that a substantial government observatory was founded in the nation's capital, Ottawa, and a separate Astronomy Department was established at U of T. Although positional astronomy and geophysics were the backbone of the program at the Dominion Observatory (DO) when it opened in 1905, its main instrument was a 38 cm Brashear refractor and its staff for many years were mostly graduates of the U of T.

The growth of the profession 1905–1970

Astrophysical research at a national observatory was unusual and might not have occurred at Ottawa if there had been any universities in the country with suitable facilities. However, there were not, and it was recognized that the Observatory would provide a stimulus to science throughout the Dominion. It is a role which government still plays in Canadian astronomy since many smaller institutions cannot on their own afford the facilities necessary for modern research.

Solar studies and stellar spectroscopy were the main areas of research in astrophysics at the DO but it soon became apparent that the scope of a 38 cm instrument in a less-than-desirable environment was limited. The main push for improved facilities came from J S PLASKETT, who had graduated as a mature student from Toronto and had joined the staff of the DO even before it opened. By 1910 he realized 'that our output in radial velocities observations of spectroscopic binaries is likely to diminish materially. . . . The only remedy is an increase of telescope aperture'. It is a tribute to his organizational and political genius that, in spite of the First World War, the DOMINION ASTROPHYSICAL OBSERVATORY (DAO) equipped with a 1.8 m reflector opened near Victoria, British Columbia, in April 1918. At the time, it was the largest operating telescope in the world since the Mount Wilson 2.5 m was was still undergoing modifications. The DAO is still going strong—a second telescope, a 1.2 m with a Coudé spectrograph, was added in 1961 and detection equipment in both instruments has been continuously updated. Spectroscopy has always been the cornerstone of the DAO programs, with some highlights being Plaskett's work on the rotation of the Galaxy using radial velocities of O- and B-type stars, McKellar's work on interstellar absorption, Petrie's calibration of the absolute magnitudes of B stars and Batten's catalogues of spectroscopic binary orbits.

An even larger reflecting telescope (1.9 m) began operations in 1935 thanks to the persistence of Dr Chant at the U of T and the very generous private donation of the DAVID DUNLAP OBSERVATORY (DDO). It was the only major astronomical facility at a Canadian university for many years, and U of T was the only place in the country where a student could get specialized training in astronomy until the 1950s. Spectroscopy and photometry have been the basis of most of the research carried out there with HELEN HOGG's lifelong work on variable stars in globular clusters and Bolton's convincing 1972 research on Cygnus X-1 as a black hole being especially memorable.

After the Second World War, radio astronomy opened up. In Canada it was an offshoot of radar work carried out by A E Covington at the NATIONAL RESEARCH COUNCIL (NRC). He recognized the potential of some of the equipment for solar research and in 1945 began a program of monitoring the Sun at a wavelength of 10.7 cm—a program which still continues. Two large steerable radio telescopes were inaugurated in the 1960s—one being the 45 m at the ALGONQUIN RADIO OBSERVATORY (now used for geodetic measurements) and the other a 26 m dish at the DOMINION RADIO ASTROPHYSICAL OBSERVATORY (DRAO) at Penticton, British Columbia. These instruments proved very versatile in studying the interstellar medium and in identifying some of the largest molecules in space. By 1966, the DRAO also had installed two large interferometric arrays used for high-resolution imaging of discrete sources and mapping of the galactic plane.

Because of the postwar baby boom and the cold war space competition, the 1960s saw a tremendous surge in the founding of universities, some with astronomy departments or parts of departments. In fact a third of all Canadian universities were established in the 1960s. The first astronomy PhD was earned in Canada in 1953 at the U of T; 20 yr later, there were 12 awarded at seven different universities. During this period, the dissertations covered a wide range of topics but there was a preference for stellar structure and evolution as well as radial velocities applied to studies of intrinsic variables and space motions of stars in the Galaxy.

Recent developments, present status and plans for the future

By 1970, the old Dominion Observatory had outlived its usefulness, so it was closed and all government astronomical research programs were placed under the aegis of NRC. At the time this meant some programs like meteor studies which had been the life work of Peter Millman continued in Ottawa, but most of the astronomy was centered in British Columbia at the DAO and DRAO. In 1971, one of NRC's distinguished staff scientists, GERHARD HERZBERG, won the Nobel Prize in Chemistry for 'his contributions to the knowledge of electronic structure and geometry of molecules, particularly free radicals'. NRC formed the HERZBERG INSTITUTE OF ASTROPHYSICS (HIA) in his honor in 1975.

It was not only in government that a reorganization occurred. With a larger scientific community to draw on, the professional society, CASCA, was formed in 1971 and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The main Canadian universities with research programs and PhD programs in astronomy and astrophysics.

Name (and city in order east to west)	Approximate number of PhDs in last 10 yr	Telescopes ≥0.5 m (type or observatory name)	Some areas of research
Laval (Quebec)	11	1.6 m Mont Mégantic	Cosmology, galaxies, interstellar medium, white dwarfs, massive stars
Montréal	21	1.6 m Mont Mégantic	
Queen's (Kingston)	11		Solar system dynamics, star formation, interstellar medium, star clusters, galaxy structure and formation, neutrino and particle astrophysics, cosmology and general relativity
Toronto	37	0.6, 1.9 m at DDO David Dunlap observatory 0.6 m H S Hogg[a]	Stellar astrophysics, contact binary systems, variables, nebulae, interstellar matter, cosmology, galaxy clusters and evolution
York (Toronto)	6	0.6 m	Dynamics of solar system and star clusters, planetary atmospheres, nebular, supernova remnants, pulsars, active galactic nuclei
McMaster (Hamilton)	10		Stellar populations, globular cluster systems, star-forming regions, theoretical astrophysics
Waterloo/Guelph	9	Bakos Observatory	Molecular astronomy, interstellar medium, star formation, galactic structure, cosmology, gravitation, laboratory astronomy, instrumentation
Western Ontario (London)	5	1.2 m Elginfield Observatory	Stellar rotation, atmospheres, magnetic fields, surface features, interstellar matter and nebulae, star clusters, quasars
Alberta (Edmonton)	11	0.5 m Devon Observatory	Photometry and spectroscopy of binaries and variables, supernova remnants, cosmology
Calgary	11	0.5 m Rothney Observatory 1.8 m Honeycomb	Monitors asteroids and variable stars; planetary transit searches, optical and infrared photometry and spectroscopy, continuous monitoring of circumpolar objects
British Columbia (Vancouver)	19	2.7 m liquid mirror	Low-mass stars, substellar companions, astroseismology, atmospheres of hot stars, galaxy interactions
Victoria	15	0.5 m Climenhaga Observatory	Formation, structure and evolution of solar system objects, stars, galaxies, galaxy clusters, quasars and large-scale structure

[a] Helen Sawyer Hogg Telescope is located in Argentina.

by its first annual meeting had 145 members. Currently there are 381 members, 281 of whom are located in Canada. The fact that more than a quarter of the members live abroad is mainly due to emigration.

Multi-user scientific facilities and international cooperation became the only way that Canada could afford to participate in large projects at the forefront of current research. The MÉGANTIC ASTRONOMICAL OBSERVATORY, opened in 1978, is jointly run by the Université de Montréal and Université Laval. HIA is responsible for Canada's share in three international ventures—the 3.8 m CANADA–FRANCE–HAWAII TELESCOPE (CFHT) which opened in 1979 and in which Canada has a 42.5% interest and more recently the 15 m JAMES CLERK MAXWELL TELESCOPE (25%) also in Hawaii and the 8 m GEMINI OBSERVATORY telescopes (14%) in Hawaii and Chile. All three facilities have permanent staff employed by HIA but time allocation is made widely available. In addition, HIA continues its responsibility for the two Canadian observatories already alluded to—DRAO, now a major collaborator in the Canadian Galactic Plane Survey (see figure 1) and DAO, which now is home to the Canadian Astronomy Data Center, a distribution center for data from the Hubble Space Telescope and an archival center for data from CFHT. Two other recent Canadian multi-user projects are the Canadian Institute for Theoretical Astrophysics (CITA), located at the University of Toronto, and the Sudbury Neutrino Observatory (SNO), a collaborative project of several Canadian and American universities.

From 1984 onwards, financial cutbacks were the order of the day. Science budgets and astronomy funding in particular declined at the same time as international cooperative ventures used up larger proportions of what was available. Throughout the 1990s, research grants provided by NSERC in space and astronomy amounted to only about $4 million per year.

There are now about 35 Canadian institutions of higher learning where some astronomy is taught; in about half of these research is done and doctoral degrees are offered in astronomy or closely related fields (the main ones are listed in table 1). Approximately 20 PhDs are awarded each year, the most prevalent topics being

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

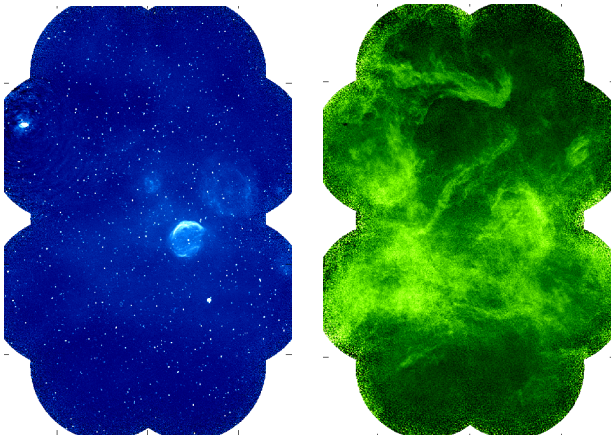


Figure 1. Images from the Canadian Galactic Plane Survey of a $6.8^\circ \times 9.7^\circ$ region in Cassiopeia. Left: 21 cm continuum; right: 21 cm line.

theoretical astrophysics, structure and evolution of normal galaxies, clusters of galaxies, cosmology, interstellar matter and the stellar environment. Greater participation in theoretical work ranging from cosmology to planetary system dynamics in recent years is partly a reflection of the establishment of CITA.

Building on technical expertise that Canadians already have in such areas as instrument and software design, a partnership among NRC, NSERC and CASCA is developing a long-range plan for astronomy in Canada until 2015. At the present time this panel is considering a wide variety of issues, including maintenance of present programs and facilities along with participation in the Next Generation Space Telescope, the Atacama Large Millimeter Array and the Square Kilometer Array operating at cm and dm wavelengths. A small Canadian satellite to study microvariability and oscillations of stars (MOST) is also in the works, with a possible launch date in 2001. Balancing various demands and aspirations is a challenging task in an environment where future funding for research is uncertain. Fortunately, economic conditions are good at present and there seems to be a political awareness that Canada's future in science and technology requires greater government support.

Bibliography

Aikman C G http://www.seameadow.com/canast.html has links to most Canadian astronomical web sites

Jarrell R A 1988 *The Cold Light of Dawn: a History of Canadian Astronomy* (Toronto: University of Toronto Press)

Peter Broughton

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canals (Martian)

Elusive linear features on Mars claimed to exist by some observers from the late nineteenth century until well into the twentieth. Although isolated streaks had been reported by earlier observers, it was Giovanni Schiaparelli who first gave linear markings prominence on a map of Mars he drew in 1877. He gave them the Italian name *canali*, which can mean 'channels' or 'canals', but was translated as the latter in English reports. The connotation of artificial construction thus became attached to them, though this was not what Schiaparelli had intended—indeed, he disapproved of the term. In the 1880s the canals remained a controversial topic, several respected astronomers claiming to have observed them, but others—notably Asaph Hall—failing to see them.

From the time of his first claimed sightings of canals in 1894, Percival Lowell began to frame the classic scenario of the canals as artificial waterways built by Martians in a last-ditch attempt to irrigate a planet whose water resources were dwindling. This idea, propounded by Lowell in books such as *Mars and Its Canals*, and depicted by him in maps of Mars showing an extensive and intricate canal network, fired the public imagination and infused literature, inspiring such stories as H G Wells's *War of the Worlds*. The power of the popular vision persisted, surviving Eugène Antoniadi's demonstration that the canals were illusory—a result of the observer struggling to resolve elusive detail. When the brain thinks it knows what the eye should see, the eye sees it. Images returned by spacecraft from the late 1960s showed no canals.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cancer

(the Crab; abbrev. Cnc, genitive Cancri; area 506 sq. deg.) A northern zodiacal constellation which lies between Gemini and Leo, and culminates at midnight in late January. It represents the crab that, in Greek mythology, Hercules crushed underfoot during his fight with the multi-headed Hydra. The name of α Cancri, its brightest star, is Acubens, which comes from the Arabic word meaning 'the claw'. The brightest stars of Cancer were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

An inconspicuous constellation, the brightest stars in Cancer are β Cancri, magnitude 3.5, δ Cancri (Asellus Australis), magnitude 3.9, and ι Cancri, a wide double star with yellow (G8) and bluish-white (A3) components, magnitudes 4.0 and 6.6, separation 30″. Another interesting star is ζ Cancri (Tegmine), a quadruple system consisting of two yellow (F8 and G5) components, magnitudes 5.4 and 6.0, separation 0.6″, period 59.7 years, with a third component, also yellow (F9), magnitude 6.2, separation 6.2″, which revolves around the others in about 1150 years, and a fourth, unseen component which is thought to revolve around the third in about 17.6 years. The constellation also contains the Mira-type variable star R Cancri (range 6.1–11.8, period about 362 days).

Midway between and slightly to the west of γ and δ Cancri is Praesepe (the Beehive or Manger; M44, NGC 2632), one of the most magnificent open star clusters in the sky, which, at a distance of about 160 pc, is just visible to the unadied eye as a faint misty patch. Another interesting open cluster is M67 (NGC 2682), which contains about 200 stars between eighth and fourteenth magnitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Candy, Michael Philip (1928–94)

Celestial mechanician, born in Bath, England. While at the Royal Greenwich Observatory became the first astronomer both to discover a comet and to calculate its orbit (Comet Candy 1960), became director of the Perth Observatory, which he supervised into measurements of minor planets and comets, including a vast number of positions for Halley's Comet in 1986.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canes Venatici

(the Hunting Dogs; abbrev. CVn, gen. Canum Venaticorum; area 465 sq. deg.) A northern constellation which lies between Ursa Major and Boötes, and culminates at midnight in early April. It was named by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* though the celestial figure had been shown in earlier charts dating back to 1533. Previously its stars had been included in Ursa Major.

A rather inconspicuous constellation, the brightest star in Canes Venatici is α Canum Venaticorum (Cor Caroli), a wide binary with bluish-white (A0) and white (F0) components, magnitudes, 2.9 and 5.6, separation 19.4″, the former of which is the prototype of a class of magnetic spectrum variable stars. There are no other stars brighter than fourth magnitude. Another interesting variable is the Mira-type star R Canum Venaticorum (range 6.5–12.9, period about 329 days).

Towards the boundary of Canes Venatici, rather more than halfway between α Canum Venaticorum and α Boötis (Arcturus), lies M3 (NGC 5272), a sixth-magnitude globular cluster. Also towards the boundary of Canes Venatici, but this time about 4° south-east of the star η Ursa Majoris (the last of the bright stars in the tail of the Great Bear) lies the eighth-magnitude Whirlpool Galaxy (M51, NGC 5194), so-called because of its noticeable spiral structure, first observed by Lord Rosse with his 1.8 m (72 in.) reflector in 1845. Three other bright galaxies in Canes Venatici are M94 (NGC 4736—eighth magnitude), M106 (NGC 4258—eighth magnitude) and M63 (NGC 5055—ninth magnitude).

See also: Cor Caroli, Whirlpool Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canis Major

(the Greater Dog; abbrev. CMa, gen. Canis Majoris; area 380 sq. deg.) A southern constellation which lies between Lepus and Puppis, and culminates at midnight in early January. It represents one of the two dogs of Orion (the Hunter), which dominates the sky to the north-west (the other dog being represented by Canis Minor). Its brightest star, α Canis Majoris (Sirius), is known as the 'dog star'. The brightest stars of Canis Major were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A small constellation, Canis Major is easily found by virtue of Sirius, which at magnitude -1.4 is the brightest star in the sky and lies on a continuation of a downward line joining the stars δ, ε and ζ Orionis, in Orion's belt. Sirius is also the brighter component of a close binary, its companion (Sirius B) being a white dwarf, magnitude 8.4, separation 4.6″, period 50 years. Other bright stars include ε Canis Majoris (Adhara), a double star with bluish-white (B2) and yellow components, magnitudes 1.5 and 7.9, separation 7.5″, δ Canis Majoris (Wezen), magnitude 1.8, and β Canis Majoris (Mirzam), magnitude 2.0. There are eight other stars of magnitude 4.0 or brighter.

Other interesting objects in Canis Major include the Algol-type variable star R Canis Majoris (range 5.7–6.3, period 1.14 days), the eclipsing binary UW Canis Majoris (range 4.8–5.3, period 4.39 days) and two open clusters, M41 (NGC 2287), which consists of about 100 stars fainter than seventh magnitude and is just visible to the unaided eye about 4° south of Sirius, and NGC 2362, which consists of about 40 stars fainter than magnitude 7.5 surrounding τ Canis Majoris (magnitude 4.4).

See also: Sirius.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canis Minor

(the Lesser Dog; abbrev. CMi, gen. Canis Minoris; area 183 sq. deg.) A northern constellation which lies between Gemini and Monoceros, and culminates at midnight in mid-January. It represents one of the two dogs of Orion (the Hunter), which dominates the sky to the west (the other dog being represented by Canis Major). Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small constellation, the brightest stars in Canis Minor are α Canis Minoris (Procyon), which at magnitude 0.4 is the eighth brightest star in the sky, and β Canis Minoris (Gomeisa), magnitude 2.9. There are no other stars brighter than fourth magnitude. Procyon, like α Canis Majoris (Sirius) to the south-west, has a white dwarf companion (Procyon B), magnitude 11.3, separation 5″, period 41 years. There are no bright star clusters, nebulae or galaxies in Canis Minor.

See also: Procyon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cannon, Annie Jump (1863–1941)

Astronomer, born in Dover, DE. Taught astronomy by her mother and at Wellesley. In 1896, she joined 'Pickering's Women,' at Harvard College Observatory to reduce data and carry out astronomical calculations. E C PICKERING had conceived a long-term project to obtain and classify stellar spectra. Begun in 1886 by Nettie Farrar and WILLIAMINA FLEMING, the first stage was to examine the spectra of more than 10 000 stars and develop a classification system containing 22 classes, modified by Antonia Maury. Annie Cannon used her own examination of bright southern hemisphere stars to divide them into spectral classes O, B, A, F, G, K, M and so on, and reclassified the previously classified stars. The resultant *Henry Draper Catalog* listed the spectral classes of nearly 400 000 stars. Cannon also published catalogs of variable stars (including 300 she herself discovered).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Canopus

The star α Carinae, named for the chief pilot of the fleet of Menelaos, who died in Egypt *c*. 1193 BC on the return from the Trojan War, the star being prominent at the time. It is the second brightest star, having an apparent magnitude of -0.62. It is a cream giant, spectral type F0Ib, at a distance of 326 light-years; its parallax is $0.010''$, and it has an absolute magnitude of -5.6.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cape York Meteorite

An iron meteorite that fell at Cape York in northern Greenland. Its fragments total 58 tonnes, making it the second largest meteorite known. The largest fragment (30.9 tonnes), known to the local people as Ahnighito, had been used by them as a source of iron before the explorer Robert Peary visited the region. He shipped Ahnighito, together with two other fragments—'the Woman' (3 tonnes) and 'the Dog' (400 kg)—to New York. Ahnighito, currently kept at the Hayden Planetarium, is the world's largest meteorite on public display. Two other fragments known by their native names (Savik 1, 3.4 tonnes, and Agpalilik, 20 tonnes) were removed to Denmark in the twentieth century.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Capella

Capella—the 'Goat Star' in the constellation Auriga, the 'Charioteer'—is the sixth brightest star in the sky, third in the Northern hemisphere. It is a well-known spectroscopic binary of yellow giants, appears to form a physical system with Capella HL, an outlying pair of RED DWARFS, and is a member of the extended Hyades MOVING GROUP (coeval with the nuclear region of the nearby young cluster). The giants of Capella are nearly twins, but differ just enough in mass to fall in distinctly separate evolutionary phases: the secondary is a lithium-rich fast-rotating Hertzsprung gap star, while the more advanced primary is a lithium-depleted, slowly rotating 'clump' giant. Capella is a linchpin of the stellar mass scale, and long has been a testbed for interferometric techniques and evolutionary models. Capella displays substantial x-ray and far-ultraviolet emissions—perhaps an exaggerated form of solar magnetic activity—making it a popular target for a long succession of space observatories. The origin of the prominent high-energy phenomena of the system and the curious differences between the two very well-characterized G-type giants continue to provoke lively debate in the stellar community.

Background

Capella (α Aurigae; HD 34029; BS 1708; Gl 194) is a nearby ($d = 12.9$ pc) bright ($V = 0.08$) multiple-star system consisting of a pair of nearly identical yellow giants (G8 III (Aa) + G1 III (Ab)) in a tight 0.7 AU, 104 day low-eccentricity orbit, circled at a great distance ($\sim 10^4$ AU) by a looser 50 AU, $\sim$ 300 yr binary of red dwarf stars (Capella HL; Gl 195; dM2 (H), dM5 (L)). The system shares the space motion of the Hyades open cluster, and O J Eggen has deemed it a member of the Hyades Moving Group (or 'Supercluster'). The spectroscopically determined masses ($2.6 M_\odot$ for the primary, $2.5 M_\odot$ for the secondary) indicate an evolutionary age of 600 million yr, similar to the cluster turn-off time of the Hyades nucleus. Figure 1 depicts the spatial arrangement of the system.

The main pair of yellow giants form a SPECTROSCOPIC BINARY STAR with radial velocity amplitudes of 26 km s^{-1} and 27 km s^{-1} (for Aa and Ab, respectively), on top of a recessional velocity of 29 km s^{-1}. The 0.″056 separation of the two giants has been cleanly resolved by a variety of interferometric techniques, beginning with Anderson's pioneering work at Mt Wilson in 1919, and extending more recently to the first true spatial maps in 1995 by the Cambridge Optical Aperture Synthesis Telescope (COAST). Spectroscopic monitoring of the radial velocity swings of the companions, accurate UV–optical–IR photometry, interferometric measurements of the orbital dimensions, narrow-band speckle reconstructions of the relative brightnesses and a highly precise PARALLAX from Hipparcos allow refinement of the stellar parameters to a degree rarely matched in binary-star astronomy. The physical dimensions of the Capella giants, and their orbit, are contrasted in figure 2 with the Sun, the solar system and the nearby dwarf-star binary α Centauri (G2 V + K1 V). α Cen shares superficial similarities with Capella, in that it is a multiple system consisting of a relatively tightly bound double (the G and K dwarfs) with a distant lighter companion (in this case the dM star Proxima Centauri). α Cen is, however, only one-third the total mass of the Capella system.

The puzzling visual spectrum

The visual spectrum of Capella was recorded by numerous observers at the close of the 19th century. W W Campbell, and independently H F Newall, discovered the periodic Doppler shifts (see DOPPLER EFFECT) that pointed to the binary nature of the star. The overall spectrum was puzzling, however: the absorption lines of one component were sharp and distinct, but those of the other were barely recognizable. At the same time, interferometric work in the 1920s indicated that the two stars were comparable in brightness. In the 1950s, O Struve and S M Kung, and later K O Wright, succeeded in measuring the secondary spectrum. Wright—who applied a cleverly designed mechanical subtraction technique to isolate the secondary spectrum from the composite—described the secondary's as 'diffuse', leading to a 'veiling' (or dilution) of the primary spectrum. The washed-out appearance of the secondary spectrum was attributed, at the time, to enhanced atmospheric turbulence or fast rotation.

In the 1960s, G Wallerstein called attention to the lithium absorption spectrum—a signature of evolutionary 'youth', completely dominated by the secondary star—and demonstrated that the fast-rotation hypothesis was correct. In the early 1970s, AM Boesgaard detected a weak Li I absorption in the primary, leading to the evolutionary scenario devised by I Iben (illustrated in the lower right portion of figure 2). Both of the Capella stars began their lives on the upper main sequence (MS) as late-B dwarfs. The slightly more massive primary evolved away from the MS first, as a hydrogen shell source, ascended the giant branch, ignited its helium core and then returned to the base of the red giant branch as a core helium burning star, losing its lithium by deep convective mixing along the way. The helium burning phase is relatively long lived, perhaps 20% of the MS lifetime; thus evolved stars tend to collect in that part of color–magnitude diagrams, leading to a 'clump.' Meanwhile, the slightly less massive secondary has only recently evolved away from the MS, at present is in the hydrogen shell burning 'first-crossing' phase, and the lithium dilution episode lies in its future. Because the horizontal portion of the shell-burning trajectory is traversed very rapidly, in perhaps only 0.1% of the MS lifetime, few stars are found there in color–magnitude diagrams, and the region has come to be called the 'Hertzsprung gap'. Thus while the Capella giants fall close to one another in the Hertzsprung–Russell diagram (i.e. in surface temperature and luminosity), the primary has covered considerably more evolutionary 'distance' than the secondary to arrive at its present position; subsidiary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

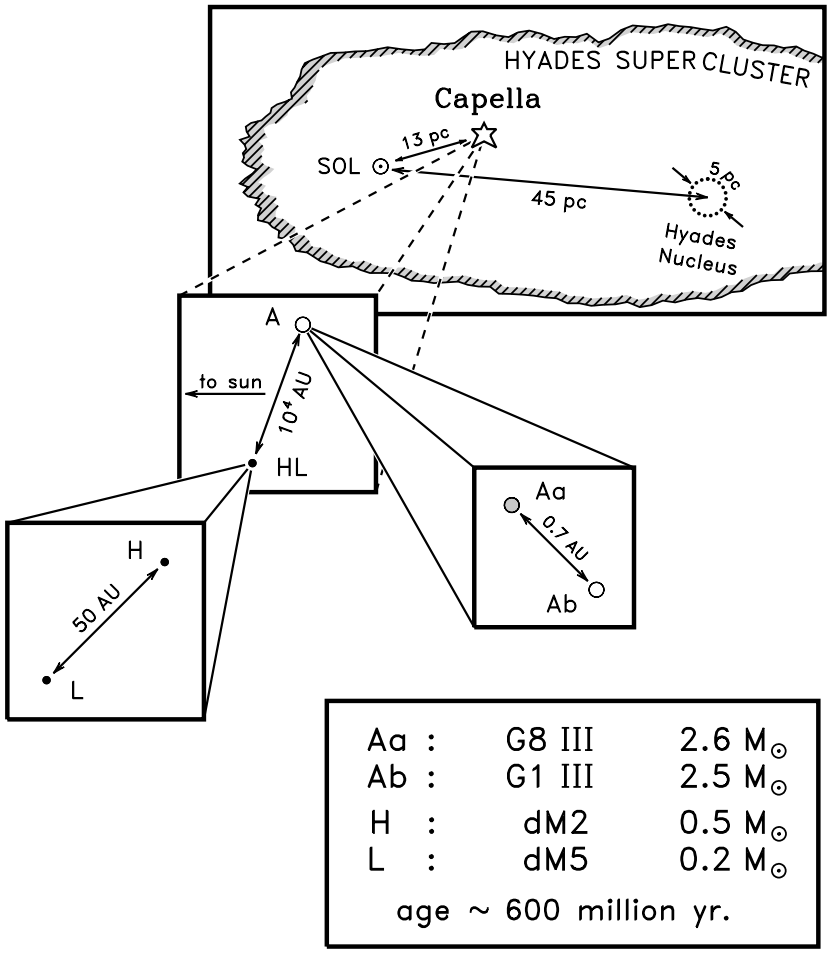


Figure 1. Cartoon of Capella: its constituent stars and relationship to the coeval Hyades cluster and the Sun. (The Sun is much older than the Hyades, and only coincidentally finds itself moving through the outskirts of the cluster.)

properties such as rotation, lithium abundance and CNO isotopic ratios tell the tale.

In the 1980s, F C Fekel and collaborators obtained high-precision rotational velocities for both components: $v_{\rm Aa} \sin i = 6$ km s^{-1}; $v_{\rm Ab} \sin i = 36$ km s^{-1}. From that time to the present, there has been steady refinement of the orbital dimensions of the Capella system using interferometry, and further improvements in the spectroscopic orbit. Very recently, the new Hipparcos parallax for Capella yields a distance (12.9 pc) about 3% closer than the previous trigonometric estimates, implying luminosities about 6% less. The radius of the primary is $13R_\odot$; that of the secondary is $9R_\odot$. The effective temperatures are 4940 K and 5700 K, respectively.

Figure 3 illustrates (simulated) visible and ultraviolet spectra of Capella at orbital phase $\phi = 0.25$ (the quadrature when Aa is receding at its maximum velocity and Ab is approaching at its maximum), and half a cycle later at $\phi = 0.75$ (Aa approaching; Ab receding). The visible spectrum of each component (middle curve, top panels) was obtained by rotationally smoothing a solar trace by the appropriate $v \sin i$; the absorption feature near 6339 Å is an unresolved close blend. The vertical dotted line marks the rest position of the center sharp absorption. For clarity, the primary spectrum was assumed to be 10% brighter than the secondary at this red wavelength (corresponding to $\Delta V \sim 0.1$ mag); the secondary spectrum is shaded. Note the washed-out appearance of the latter. When the two spectra are combined to yield the composite at the top of each panel, the sharp lines of the primary are diluted in relative flux by a factor of ∼2 and fall on the undulating background provided by the rotationally smeared secondary. No wonder it was so difficult to isolate the secondary contribution in historical noisy photographic spectra.

High-energy measurements

Capella was the first 'normal' extra-solar star detected in x-rays (serendipitously during a sounding rocket flight), and is a bright far-ultraviolet emission-line source. The high 'activity' of the system has made it a popular target for spacecraft observatories, beginning with Copernicus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

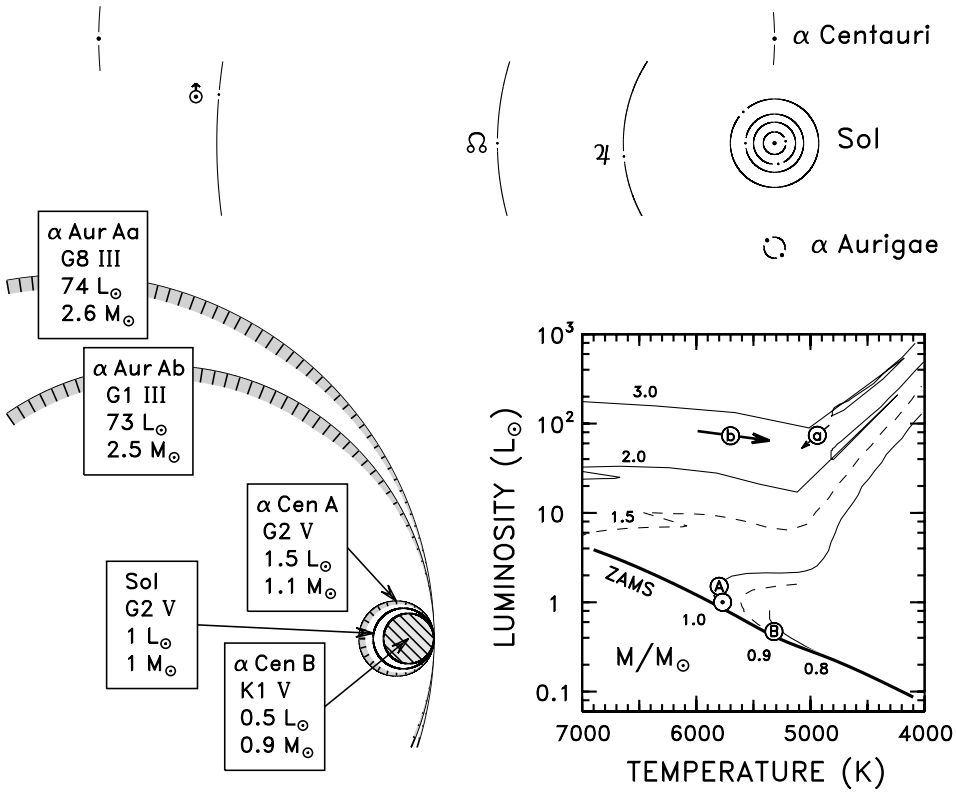


Figure 2. Dimensions of the Capella stars and their orbit, and evolutionary status. (Note: points representing the stars and planets in upper portion of the diagram are not to scale.)

in the mid-1970s, and extending in the modern era to the Japanese x-ray spectroscopy satellite ASCA and the Italian–Dutch high-energy observatory BeppoSAX.

In the early Copernicus work, A K Dupree found strongly *blueshifted* high-excitation O VI λ1032 in Capella, relative to the expected velocity of the primary star at the particular orbital phase observed. She interpreted the shifts as evidence for a massive $T \sim 10^5$ K wind emanating from the primary star. Later with the all-sky survey by HEAO-1, Capella was found to share high x-ray levels with the short-period ($P \sim$ few day) tidally locked RS Canum Venaticorum binaries, and was designated a long-period member of the class. Subsequent studies by T R Ayres and J L Linsky with the far-ultraviolet echelle spectrometer on the International Ultraviolet Explorer, beginning shortly after its launch in 1978, discovered unexpectedly that the fast-rotating G1 III secondary star was responsible for virtually all of the high-excitation emissions of the system (from species such as Si IV λ1393 and C IV λ1548, forming at $T \sim 6 \times 10^4$–10^5 K, conditions characteristic in the Sun of the subcoronal TRANSITION REGION). Only a small contribution was detected from the slowly rotating G8 III primary. (In hindsight, the earlier Copernicus result for O VI could be reinterpreted as slightly *redshifted* emission from the *secondary* star at the particular orbital phase.)

The bottom panels of figure 3 illustrate the effect for the C IV doublet based on line widths, relative shifts and continuum intensities measured in Hubble Space Telescope spectra (GHRS instrument) by B E Wood and collaborators. The vertical dotted line marks the rest position of C IV λ1548. Contrary to the visible region, the G1 secondary star completely dominates the far-ultraviolet line and (weak) continuum emission, while the G8 primary is only a small perturbation. Note also the enormous widths (∼150 km s^{-1} full width at half-intensity) of the C IV components in the secondary star (considering that the velocity scale in this depiction is 'compressed' by a factor of 4 in the lower panels compared with the visible spectra owing to the shorter wavelength but same 10 Å interval). More difficult to see in this rendition are the systematic redshifts of the C IV lines (in both stars) relative to the velocities expected at each orbital phase. Such shifts are recorded commonly in ultraviolet spectra of the Sun, in lines arising from the 10^5 K transition zone, and are thought to be caused by mass motions of material in a complex circulation system within coronal magnetic loops (see CORONAL LOOPS).

In the early 1980s, fast rotation was viewed as the key link between the coronal hyperactivity of the short-period RS CVns and long-period Capella. (Tidal synchronization in the semi-detached RS CVn binaries causes each of the component stars to rotate at the orbital period, and thus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

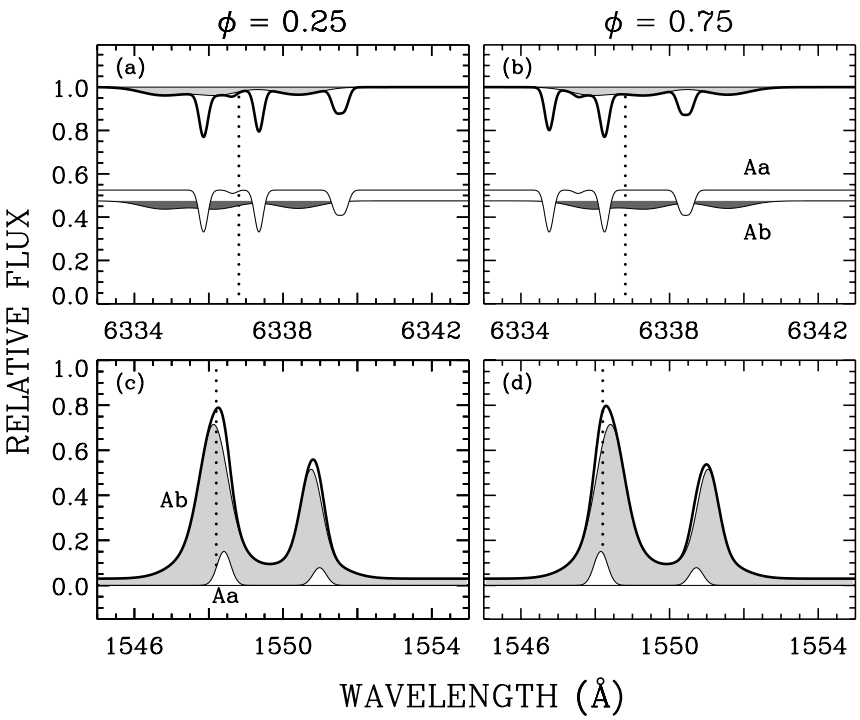


Figure 3. Simulated spectra in visible (top panels) and far-ultraviolet (bottom panels) at opposite orbital quadratures (maximum radial velocity separation).

rapidly since $P_{\rm orb}$ is short.) In reality, Capella is 'half an RS CVn', because just one of the stars rotates rapidly. The connection between fast spin and enhanced coronal activity arguably is through strong surface magnetic fields, produced in the stellar convective envelope by rotation-catalyzed 'dynamo' action.

The dichotomy between the rotation rates of the two nearly identical coeval giants was attributed to the large evolutionary separation that can develop from even a small mass difference (as described previously in Iben's scenario). Indeed, there is an apparent sharp decrease in the spins of stars just past G1 III. The cause of the rapid braking episode currently is a subject of debate (leading contenders are evolutionary redistribution of angular momentum in the stellar interior, and shedding of angular momentum by a strong stellar wind), but there is little doubt that many of the curious ultraviolet (and visual) properties of Capella can be understood if the G8 star already has passed through the rapid braking stage, while the G1 star is poised at the brink.

Current research

Recent investigations have focused on high-spectral resolution ultraviolet measurements of Capella at critical orbital phases, particularly dynamical processes revealed in global redshifts of the high-temperature lines, and the mysterious 'broad components' (extended high-velocity wings on the emission line profiles) identified by B E Wood which might be signatures of the stellar analogs of solar 'transition zone explosive events'. Another area of interest is the recognition that despite its strong far-UV emissions, the G1 secondary star (and other fast-rotating Hertzsprung gap giants) is relatively deficient in its coronal x-ray luminosity. In fact, the G1 star probably has about the same x-ray brightness as the G8 primary, despite a 10-fold advantage at C IV λ1548. The paradox is seen clearly in UV measurements by HST of the coronal forbidden line Fe XXI λ1354, which is detected in both of the giants at about equal strengths. The primary star, itself, falls in a distinct class of active clump giants, the prototype of which is β Ceti (K0 III) and which includes two of the four Hyades red giants (evolutionary cousins of the Capella stars). Another curiosity of Capella is that, despite the substantial coronal emissions (of both components), and intense scrutiny by historical and contemporary x-ray and ultraviolet missions, there are surprisingly few reports of variability and a remarkable absence of FLARES. In contrast, dramatic x-ray variability and flare activity (including nonthermal emission at radio frequencies) are trademarks of the classical RS CVn binaries.

Summary

As a nearby bright spectroscopic binary of yellow giants—one of them a rare Hertzsprung gap star—and trailing emissary of the young Hyades open cluster, Capella has drawn the attention of spectroscopists for more than a century. Astronomers have tested interferometric techniques on the 'subvisual' orbit for nearly as long.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

About 30 yr ago, the nature of the somewhat mysterious secondary star was unraveled, and explained in the context of a compelling evolutionary scenario. With the development of space observatories in the mid-1970s, Capella again found itself in the astronomical spotlight, often a prominent entry in the commissioning target list of a new mission. Interest in the system—and its message concerning the role of stellar structure on the evolution of coronal magnetic activity—remains unabated today.

Bibliography

Ayres T R and Linsky J L 1980 Outer atmospheres of cool stars. V—IUE observations of Capella—the rotation–activity connection *Astrophys. J.* **241** 279–99

Ayres T R, Schiffer F H and Linsky J L 1983 Outer atmospheres of cool stars. XIII—Capella at critical phases *Astrophys. J.* **272** 223–33

Barlow D J, Fekel F C and Scarfe C D 1993 A three-dimensional solution for the orbit of Capella *Publ. Astron. Soc. Pacific* **105** 476–86

Burnham R Jr 1978 *Burnham's Celestial Handbook* (New York: Dover)

Linsky J L, Wood B E, Judge P, Brown A, Andrulis C and Ayres T R 1995 The transition regions of Capella *Astrophys. J.* **442** 381–400

T R Ayres

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Capra, Baldassar [Baldassarre] (1580–1626)

Italian astronomer, born in Milan, studied in Padua. He attacked GALILEO over his three public lectures, delivered at the University of Padua on the new star which appeared in 1604, and crossed swords with him in an incident in which Capra plagiarized Galileo's work on the proportional compass. Galileo replied with a devastating defense, and Capra was expelled from the university. *See* SIMON MARIUS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Capricornus

(the Sea Goat; abbrev. Cap, gen. Capricorni; area 414 sq. deg.) A southern zodiacal constellation which lies between Sagittarius and Aquarius, and culminates at midnight in early August. The origin of the unusual constellation figure, which comprises the head and forelimbs of a goat and tail of a fish, dates back to Babylonian times and has also been associated in Greek mythology with Pan, who had the horns, legs and ears of a goat and turned the lower half of his body into that of a fish as he plunged into the Nile to escape the monster Typhon. The brightest stars of Capricornus were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A rather inconspicuous constellation, the brightest stars in Capricornus are δ Capricorni, (Deneb Algedi), magnitude 2.9, and β Capricorni (Dabih), a multiple system consisting of two main components, a yellow (F8, combined) primary, magnitude 3.1, which is a spectroscopic triple star, and a bluish-white (B9) secondary, magnitude 6.1, which has a magnitude 10.2 companion, separation 0.85″. There are three other stars brighter than fourth magnitude, including α^2 Capricorni (Secunda Algedi or Secunda Giedi), yellow (G8), magnitude 3.6, a naked-eye double with unrelated α^1 (Prima Algedi or Prima Giedi), yellow (G3), magnitude 4.3, separation 378″, both stars having fainter companions. Another interesting object is M30 (NGC 7099), an eighth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carbon Stars

The intrinsic properties of carbon stars

In 1868 ANGELO SECCHI , S.J., announced that he had found a fourth type of stellar spectrum. His first three types, designated by Roman numerals of course, were stars with hydrogen lines (now A and B stars), stars with metallic lines (now types F, G and K) and stars with absorption bands that diminished toward longer wavelengths (now types M and S). Secchi's type IV showed bands tailing toward shorter wavelengths. By inserting a carbon arc in his visual spectroscope he immediately recognized that the new bands were due to C_2 and other C compounds, now known to be largely CN. The difference between the oxide bands of Secchi's type III and the C bands of type IV was recognized by Fujita to be controlled by the chemistry of the very stable molecule CO. If the abundance of O exceeds that of C the remaining O, after forming CO, is available to form the oxide molecules seen in spectra of types M and S. However, in stars with C > O the O is used up in CO and the remaining C is available to form carbon molecules such as C_2, CN and CH.

The modern system of classification of carbon stars was established by Keenan (1993), who subdivided them into three sequences C-R, C-N and C-H corresponding to the old RN and CH star classifications. A numeral follows to indicate the temperature sequence, e.g. from C-N1 to C-N9. Keenan's full description of the spectra includes the strengths of the C_2, CN and Merrill–Sanford bands as well as the $^{12}C/^{13}C$ ratio and the Li line strength. An index of the s-process enhancement would be useful, but is difficult to observe except at high spectral resolutions.

The intrinsic luminosities of carbon stars can be derived either from HIPPARCOS PARALLAXES or by association of the carbon star with a stellar system whose distance has been determined. In addition, photometry from the visual to the K band (2.2 μm) is necessary. By limiting the sample to stars with parallaxes at least three times their probable error and errors less than 1.5 marcsec a list of 33 objects was assembled by Wallerstein and Knapp (1998). Most of the stars are of type C-N with only a few C-H, C-R and Mira stars represented. If the criteria for acceptance into the sample are lowered one runs into a statistical quagmire and the apparent spread in M_V and M_K increases rapidly.

There are many carbon stars in other galaxies such as the MAGELLANIC CLOUDS, where extensive surveys have been conducted. Since the distances to the clouds are now fairly well known, the luminosities of the carbon stars may be calculated. M_K values as bright as -9 are present, while the Hipparcos sample of stars mentioned above included only four stars brighter than $M_K = -8$ and none brighter than $M_K = -8.8$. The C-R stars show $M_K \simeq -5$ and $M_V \simeq +1.0$, significantly fainter than the C-N stars.

The calculation of effective temperature, $T_{\rm eff}$, depends only on measuring the star's total radiative flux and angular diameter. The latter can be done either by lunar occultation or by interferometry. As of 1998, 26 C-N stars have had their $T_{\rm eff}$ values estimated in this way. The range of values lies from 2000 to about 3300 K. The C-R and C-H stars are considerably warmer with $T_{\rm eff}$ values from 4000–5000 K, as determined by analysis of their spectral lines.

Carbon star masses are known only by indirect inference. Stellar models that lead to atmospheric carbonization by helium shell flashes and mixing can be followed for stars of roughly 1–3$M_\odot$. The upper limit may be as large as 5 or 6$M_\odot$ but the mass-function for main sequence stars ensures that such objects should be relatively rare. The spectroscopic orbits of C-H stars indicate masses near 0.8$M_\odot$ but the C-H primaries appear to have obtained their carbon-rich atmospheres by transfer from their current white dwarf companions, whose original masses are unknown, but probably were limited to the numbers suggested above. The C-H stars in the Large Magellanic Cloud, however, are very luminous and their origins are uncertain.

The evolution of carbon stars has intrigued stellar astronomers for some 50 years; i.e., ever since the significance of the $3\,^4\mathrm{He} \rightarrow {}^{12}\mathrm{C}$ reaction was pointed out independently by Öpik and Salpeter respectively in 1951 and 1952. Temperatures above about 10^8 K are needed for the triple-alpha reaction to take place, while the excess carbon eventually appears in the stellar atmosphere where the temperature ranges from 2–5 $\times$ 10^3 K. In addition, detailed abundance ratios such as the $^{12}C/^{13}C$ ratio and the enhancement of many heavy elements must be accounted for. In stars of less than about 2.0$M_\odot$ the triple-alpha reaction ignites at the tip of the red giant branch in a 'core-flash' (because its rate increases exponentially at first) but the resulting helium produced by the flash and subsequent quiescent burning seem to be confined to the stellar core. In more massive stars He burns quiescently, building up a C core with some extension to ^{16}O as ^{12}C captures an additional ^4He nucleus. Eventually, low mass stars with C cores and helium envelopes can ignite their He at the C–He boundary in a 'shell-flash'. By a series of events involving sudden convective mixing the ^{12}C reacts with protons to produce ^{13}C which, in turn, can react with helium nuclei to produce oxygen and free neutrons. The capture of the free neutrons by iron and other heavy elements then produces many of the heavy elements from rubidium to lead with specially conspicuous elements such as Sr, Y, Zr, Ba and the light rare earths. Even the unstable element Tc, with a half-life of 2 $\times$ 10^5 yr, is produced and convected to the stellar surface, where it is observed in some carbon stars and many S-type stars.

How much of this is observed on the surfaces of carbon stars? The observation of carbon stars with high spectral resolution from 400 to 2400 nm is not difficult. However, the analysis of the observed spectra is extremely complicated and uncertain due to the huge number of atomic lines and molecular bands in the spectra of carbon stars. While complete analyses are difficult, restricted analyses can be successfully carried out. For example, the ratio of $^{12}C/^{13}C$ can be determined accurately and ranges from about 100 to near 4 in carbon stars, thus showing the heterogeneity of evolutionary histories. Most of the C-N

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars show $^{12}C/^{13}C$ ratios above 30 while only a few show ratios near 4. The C/O ratios range from 1.01 to 1.76 with a mean of 1.15 ± 0.03. Nitrogen is almost never enhanced even in stars with low $^{12}C/^{13}C$, which is surprising. In fact, the sum of C+N+O is rarely enhanced in C-N stars despite the necessity of an enhancement of carbon for the star to be a carbon star.

The C-R stars are easier to analyse because of their higher $T_{\rm eff}$ values, from 4200 to 5000 K, but they are rare and few are bright. They all show enhanced ^{13}C with the $^{12}C/^{13}C$ ratio ranging from 4 to 9. Nitrogen is also enhanced, showing systematic CNO cycling after ^{12}C enhancement. None of them show heavy element enhancments. Two Cepheids with early R-type spectra have been analysed and show similar properties to the nonvariable R stars.

The C-H stars are the easiest to analyse because they are relatively hot and substantially metal-poor, so the degree of blending is greatly reduced. As mentioned above, they seem to have inherited their atmospheres from their now defunct companions. Their $^{12}C/^{13}C$ ratios are either less than 8 or more than 25. Some of the latter group may be 100 or more. All of them are metal-poor by factors of 3–50. Carbon and nitrogen are enhanced by about a factor of 10, oxygen is normal, and hence C+N+O turns out to be enhanced by a factor of 2–5. Apparently their low initial abundances of everything from C to Fe have made their nucleosynthesis processes more evident. The greatest enhancements are shown by s-process elements, i.e., Sr, Y, Zr, Ba and rare earths. The neutron source is probably the $^{13}C(\alpha, n)^{16}O$ reaction.

Circumstellar envelopes

In addition to being of intrinsic interest from the STELLAR EVOLUTION viewpoint carbon stars are important for the evolution of the interstellar medium and hence the Galaxy as a whole. Near-infrared observations out to 2.2 μm of RED GIANTS were first made in the 1960s and soon revealed many stars whose radiation beyond about 1.0 μm greatly exceeds what was expected from model atmospheres. It was soon recognized that the infrared excess had to be due to relatively cool gas and dust while its intensity could only be understood if the circumstellar matter filled a volume that was much larger than that of the star itself. Earlier observations of some red giants and supergiants had shown spectral lines of Na I and Ca II with Doppler shifts corresponding to material departing from the stellar surface at up to 20 km s^{-1}. Over the subsequent decades observations of circumstellar envelopes were extended to the middle infrared (5–30 μm), far infrared (30–200 μm), submillimeter (0.2–1.0 mm), millimeter (1–10 mm) and radio (10 mm and longer) spectral regions. Many instruments have contributed to our knowledge of circumstellar envelopes but the IRAS survey and very recently the ISO satellite have been responsible for major advances.

Circumstellar dust radiates in the near and mid-infrared wavelengths. The dust absorbs the starlight and is heated to temperatures in the range 1000–100 K near the star, diminishing at larger distances. The infrared spectra of M and C stars differ because of the different properties of silicate and carbon-compound dust. Carbon dust is opaque at visual wavelengths, hence greatly reducing the visible flux, while silicates are partially transparent (soot is black, while glass is transparent). The gas comes into thermal equilibrium with the dust by collisions, thus heating the gas. This provides sufficient excitation for many molecules to radiate at wavelengths corresponding to rotational transitions. The most common radiating molecule is CO, of course, where the (1–0), (2–1), (3–2) etc. transitions are often detected.

The envelopes around carbon stars provide a laboratory for organic, and even some inorganic, chemistry that cannot be realized on the earth. Near the star the high density together with temperatures as high as 1500 K ensure equilibrium chemistry but in the outer layers the reaction rates must be carefully considered. The star IRC+10216 seems to show the richest chemistry with a vast number of organic compounds present and even some chlorides and fluorides showing up in the inner envelope. Further out photochemical reactions become important as the interstellar radiation field with photon energies up to 10 eV penetrates the envelope; some ions are present so ion–molecule chemistry becomes important.

Infrared surveys such as IRAS have revealed that all C-N stars have infrared excesses indicating mass loss. The C-R and C-H stars do not show evidence of mass loss. Mass loss rates can be calculated from the equation of continuity: $\dot{M} = 4\pi r^2 \rho(r) V_\infty$, where the density, ρ, at the distance r must be estimated from a model and V_∞, the velocity of the moving envelope, can be estimated from the width or separation of emission lines of CO in the mm and submm regions. Mass loss can be detected readily down to $10^{-7} M_\odot$ yr^{-1} and in a few cases to almost 10^{-8}. At the other end of the scale mass loss rates as large as 10^{-5} or even a few times $10^{-5} M_\odot$ yr^{-1} have been calculated. Carbon stars with such high mass loss rates are virtually invisible in the visual region but are particularly important because of their rapid evolution and their contribution of a vast array of molecules to the interstellar medium.

Mass loss by carbon and other asymptotic giant branch stars presumably ceases when the outer hydrogen envelope has been completely ejected. The remaining core of the star then shrinks and its effective temperature rises. The last phases of mass loss may involve a much higher velocity than the usual 10–20 km s^{-1} of the gentle wind from the carbon star. This so-called 'superwind' can reach 1–200 km s^{-1} and can sweep up the material that was ejected earlier. A shock will form where the winds interact, causing the shocked gas to be greatly heated. In addition, the evolving star will shrink until its temperature reaches $1–2 \times 10^5$ K. The radiation from such a star will ionize the surrounding gas and the resulting configuration will constitute a PLANETARY NEBULA. Hence the final stages of a mass-losing carbon star may be described as 'Death and Transfiguration', the title used by G R Knapp at a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

conference held in 1989 celebrating the 50th anniversary of the dedication of the McDonald observatory.

Bibliography
An extensive list of references on recent literature on carbon stars is provided by:

Wallerstein G and Knapp G R 1998 *Ann. Rev. Astron. Astrophys.* **36** 369
Keenan P C 1993 *Publ. Astron. Soc. Pacific* **105** 905

George Wallerstein

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carbonaceous Chondrite

A meteorite having a higher carbon content than other classes of meteorite. Carbonaceous chondrites account for about 3% of all known chondrites. They are classified according to the proportion and size of the chondrules (millimeter-sized grains) they contain. One very rare subclass lacks chondrules, and so are actually achondrites. They are very similar in composition to the Sun (with the exception of volatiles), and the material from which they are formed is believed to have condensed from the solar nebula very early in the history of the solar system, making them vital pieces of evidence in the study of cosmogony. The carbon content is in the form of organic compounds, including amino acids—important biological precursors. They may therefore hold clues to the origin of life. The Allende meteorite is the largest known example, with an estimated total mass before fragmentation of 2 tonnes.

See also: Allende meteorite, chondrites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cardan [Cardano; Cardanus], Jerome [Geronimo; Hieronymus; Girolamo] (1501–76)

Lawyer, mathematician and physician, born in Pavia, Milan, Italy. Lived a life of family and professional drama worthy of a novel and fully documented in an autobiography. Became professor of mathematics at Padua, and of medicine at Pavia and Bologna. Prolific author (200 works) of which his greatest work was *Ars Magna* (1545, *Great Art*) which gave the first algebraic solution of cubic equations. In 1570 he was imprisoned for a few months by the Inquisition. He was accused of heresy, particularly for the book *De Astorum Iudiciis* in which he cast the horoscope of Christ, attributing events of His life to astrological influences. As a result he was barred from teaching. Correctly predicted through astrology the exact date of his own death but may have cheated by committing suicide (*see* ABU'L-HASAN IBN YUNUS).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carina

(the Keel; abbrev. Car, gen. Carinae; area 494 sq. deg.) A southern constellation which lies between Vela and Chamaeleon, and culminates at midnight in late January. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, from stars that formed part of the ancient constellation of Argo Navis (the Ship), which had been included by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A conspicuous constellation, the brightest stars in Carina are α Carinae (Canopus or Suhel), which at magnitude -0.6 is the second brightest star in the sky, β Carinae (Miaplacidus), magnitude 1.7, ε Carinae (Avior), magnitude 1.9, and ι Carinae (Aspidiske), magnitude 2.2. There are 14 other stars of magnitude 4.0 or brighter. ι and ε Carinae together with δ and κ Velorum make up an asterism called the False Cross, as it is sometimes confused with the constellation Crux (the (Southern) Cross). Variable stars in Carina include η Carinae, an eruptive, high-luminosity star (observed range -0.8 to 7.9) surrounded by a shell of nebulosity (the Homunculus Nebula) expelled in the nineteenth century, and the Mira-type variables R Carinae (range 3.9–10.5, period about 309 days) and S Carinae (range 4.5–9.9, period about 149 days).

The Milky Way cuts across the north-eastern part of Carina and the constellation contains a number of interesting star clusters and nebulae, including NGC 3532, a large ($60' \times 30'$), spectacular open cluster of about 150 stars between eighth and twelfth magnitude, NGC 2808, a sixth-magnitude globular cluster, and NGC 3372 (the Eta Carinae Nebula), a large (>4 sq. deg.), bright emission nebula centred on η Carinae.

See also: Canopus, Eta Carinae, False Cross.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carlsberg Meridian Telescope

A 7 in (17.8 cm) diameter refractor, formerly known as the Carlsberg Automatic Meridian Circle. It is part of the OBSERVATORIO DEL ROQUE DE LOS MUCHACHOS on La Palma and is dedicated to carrying out high-precision optical astrometry.

It is operated jointly by the Copenhagen University Observatory, the Institute of Astronomy, Cambridge and the Real Instituto y Observatorio de la Armada en San Fernando.

When it was first moved to La Palma in 1984, it was one of the first fully automatic telescopes in the world. Since 1997, the telescope has been operated remotely via the Internet from Britain, Denmark or Spain.

Recent upgrades include installation of a CCD which can observe about 70 000 stars a night down to visual magnitude 17. It is being used to map the northern sky between -3° and $+30^\circ$ in declination.

For further information see
http://www.ast.cam.ac.uk/~dwe/SRF/camc.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carnegie Observatories

The Carnegie Observatories were founded in 1902 by George Ellery Hale. Their first facility was the MOUNT WILSON OBSERVATORY, located in the San Gabriel Mountains above Pasadena, California. Originally a solar observatory, it moved into stellar, galactic and extragalactic research with the construction of the 60 in (1.5 m), and 100 in (2.5 m) telescopes, each of which was the largest in the world at the time of its construction. From the 1940s through the 1970s Carnegie, in partnership with the California Institute of Technology, ran the Mount Wilson and PALOMAR OBSERVATORIES.

Today, the offices of the Observatories remain in Pasadena, but their main observational facility is the LAS CAMPANAS OBSERVATORY located in the southern fringes of Chile's Atacama Desert. There, Carnegie operates the 1 m Swope and 2.5 m du Pont Telescopes. The twin 6.5 m Magellan Telescopes, built in partnership with Harvard, MIT, and the Universities of Michigan and Arizona, will be completed in 2002.

Carnegie Observatories have played a central role in the history of twentieth century astronomy. Major contributions of Carnegie astronomers include Shapley's use of globular clusters to establish the structure of the galaxy, Hubble's discovery of the expansion of the universe, Baade's work on stellar populations, and Sandage's pursuit of the Hubble Constant. Today, the staff of about twenty astronomers works on stellar populations, the structure and evolution of galaxies, and cosmology. The Observatories are one of five departments of the Carnegie Institution of Washington, a private, non-profit organization devoted to scientific research.

For further information see
http://www.ociw.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carrington, Richard Christopher (1826–75)

English amateur astronomer, first person to observe (1859) a solar flare, 'two patches of intensely bright and white light', confirmed by R Hodgson and by the correlated appearance of auroras, seen as far south as Cuba, and magnetic disturbances, observed at the Kew Observatory. Carrington noted the connections, but cautioned that 'one swallow does not make a summer'. Only about 50 flares have been observed in the white light spectral range, as used by Carrington's visual observations with a 4.5 in refractor.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Carter National Observatory

The Carter National Observatory is situated in the Botanic Gardens in Wellington, New Zealand. Opened in 1941, the observatory is equipped with a 41 cm Boller and Chivens, an historic 23 cm Cooke photo-visual refractor and a 36 seat Zeiss planetarium. The staff are involved in research, school and tertiary education programs.

The observatory currently has eight full time members. Research programs are run in collaboration with a consortium of New Zealand and Japanese universities and involve micro-lensing and the search for dark matter.

For further information see
http://www.vuw.ac.nz/~carter/carter.html

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cartwheel Galaxy

A peculiar galaxy in the constellation of Sculptor that derives its name from its wheel-like appearance. The Cartwheel has a ring-shaped 'rim', 150 000 light-years in diameter, that contains billions of recently formed stars and which is dominated by massive clusters of bright blue stars and HII regions. The nucleus, or 'hub' of the galaxy contains a predominantly older population of stars and is surrounded by faint arms, or 'spokes' that stretch out towards the ring. Its striking appearance is the result of a galactic collision. A few hundred million years ago, a smaller galaxy plunged straight through the Cartwheel, which at that time is presumed to have been a normal spiral galaxy. The resulting shock spread out through its gaseous disk like a ripple on a pond, pushing gas and dust before it and compressing it into a ring. The compressed gas clouds collapsed and fragmented, thereby initiating the dramatic burst of star formation that is evident today in the ring. Located at a distance of some 500 million light-years, the Cartwheel is an example of a starburst galaxy (one that is undergoing a major bout of star formation).

See also: galactic nucleus, HII regions, interacting galaxies, spiral arms, spiral galaxy, starburst galaxies, star formation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassegrain Telescope

A type of reflecting telescope based on a design that was produced in 1672 by the Frenchman Guillaume Cassegrain. It utilizes a concave paraboloidal primary mirror and a convex ellipsoidal secondary mirror that is located a short distance inside the focus of the primary. The converging cone of light from the primary is reflected by the secondary, back down the telescope tube and through a hole in the center of the primary to the eyepiece. Because the curved surface of the secondary causes rays of light to converge at a narrower angle than the rays reflected directly from the primary, it increases the effective focal length of the instrument, so enabling an instrument of long effective focal length to be contained within a relatively short tube. Although the focal ratio of a Cassegrain primary mirror is typically in the region of f:3 to f:4, the effect of the secondary mirror enables these instruments to operate at focal ratios in the region of f:10 to f:30.

The standard Cassegrain design produces sharply focused images only in the central parts of its field of view, images away from the center ('off-axis images') being subject to various distortions, or aberrations, such as coma. Good quality images, over fields of view several tens of minutes of arc in diameter, can be obtained with telescopes of the Ritchey–Chrétien design in which the primary is a concave hyperboloid (rather than a paraboloid) and the secondary has a steeper hyperbolic curve than that of a conventional Cassegrain secondary. Another variant on the Cassegrain design is the Dall–Kirkham, which uses a concave ellipsoidal primary and a spherical convex secondary, curves which are easier to make than the paraboloids and hyperboloids of conventional Cassegrains. However, the field of view of a Dall–Kirkham is typically only about one-third of that of a normal Cassegrain.

Most large modern telescopes are based on the Cassegrain design or one of its variants.

See also: field of view, focal length, focal ratio, focus, Gregorian telescope, mirror, reflector, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassegrain, Laurent (*c*. 1629–93)

French teacher and priest, inventor (1672) of the *Cassegrain telescope* configuration, the most widely used reflecting telescope in astronomy. It consists of a concave primary mirror and a convex secondary, in which the compact arrangement of mirrors can be chosen so that aberrations on each mirror cancel each other. Little is known of the life of this shadowy figure, who lived and worked near Chartres and Chaudon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini Dynasty

Cassini, Gian Domenico [Cassini I] (1625–1712); Cassini, Jacques [Cassini II] (1677–1756) son of Gian; Cassini, César-François [Cassini III] (1714–84), second son of Jacques; Cassini, Jean-Dominique [Cassini IV] (1748–1845), son of César-François.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini, César-François [Cassini de Thury, known as Cassini III] (1714–84)

Second son of JACQUES CASSINI, born in Thury, Oise, France. Director of the Paris Observatory (1771), where he concentrated on geodesy and made the first modern map of France which, for navigational purposes, would accurately place the French ports relative to Paris. To everyone's surprise, France was some 20% smaller than previously believed, and the King commented that France had lost more territory to the astronomers than to its enemies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini, Gian Domenico [Giovanni Domenico; Jean Dominique; known as Cassini I] (1625–1712)

Italian-born French astronomer, born in Perinaldo, near Naples. Attracted to astrology in his youth, became Professor at Bologna, during which time he conducted hydrological studies for the Pope to mitigate flooding of the River Po. In 1669 Cassini moved to France and set up the Paris Observatory, remaining director for the rest of his career. He pushed continually for the observatory to acquire the latest technology and make improvements, winning substantial funding. His powerful new telescopes were used to make important findings about the solar system. He saw spots on Mars and measured the planet's rotation period. He also measured the rotation of Jupiter and mapped its spots, bands and flattening at the poles. He timed the revolution of Jupiter's satellites, which allowed the Danish astronomer OLE RØMER to compute the speed of light while he was working in Paris. He made the best map of the Moon until the invention of photography. He observed comets and wrote about planetary and satellite orbits. Although the phenomenon had long been known, possibly from classical times, certainly since the Middle Ages, Cassini was the first to record (1683) scientific observations of the zodiacal light. In 1684 he discovered Iapetus, Rhea, Tethys and Dione, four satellites of Saturn in addition to those previously discovered by HUYGENS. In 1675, he discovered that Saturn's rings are split largely into two parts by a narrow gap—known since as the *Cassini division*. Cassini elaborated on the first major theory of atmospheric refraction, founded on the sine law but opposed the theory of universal gravitation, bequeathing the battle against Newtonianism to his son, Jacques, by whom (and by his grandson and great-grandson) he was succeeded as director of the observatory. The Cassini spacecraft, launched 1987 and scheduled to arrive at Saturn in 2004 to begin exploration of the planet and its moons, is named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini, Jacques [known as Cassini II] (1677–1756)

Astronomer, born in Paris (probably at the Observatory itself), son of JEAN-DOMINIQUE CASSINI, whom he at first assisted, then replaced in official functions as Cassini became blind, and then succeeded as director of the Paris Observatory. Measured proper motion of Arcturus. Cassini II fought continually and unsuccessfully to defend the work of his father and to reconcile observations with the Cartesian theory of vortices. In 1740, understanding the uselessness of his opposition to the new ideas he progressively abandoned his scientific work and handed over the work of the Observatory to his son, César-François.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini, Jean-Dominique [known as Cassini IV] (1748–1845)

Astronomer, born at the Paris Observatory, great-grandson of the first director, grandson of the second, son of the third and himself the fourth, assuming the responsibilities of the directorship during the decline of his father, and the directorship itself in 1784. Persuaded King Louis XVI to restore the observatory, but the French Revolution occurred and Cassini, a monarchist, bitterly opposed the involvement of the new revolutionary government in observatory affairs and lost his position in 1793. Cassini IV spent his old age in his chateau, writing polemics justifying his own position and defending the scientific reputation of the family. The Paris Observatory had been in the charge of the Cassini dynasty for 120 years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassini/Huygens Mission

The Cassini/Huygens mission is a planetary mission designed to explore in detail the Saturnian system. The mission is a joint undertaking between the United States National Aeronautics and Space Administration (NASA) and the EUROPEAN SPACE AGENCY (ESA). It was successfully launched from Cape Canaveral, Florida, on 15 October 1997 (figure 1) for a 7 yr interplanetary journey. The Cassini/Huygens spacecraft comprises the main craft, the Saturn Orbiter, and the Titan Probe, Huygens. The aim of the mission is to carry out a detailed study of the planet SATURN, its rings, satellites and magnetosphere. TITAN, Saturn's largest moon, is a prime target of the mission. A prime goal of the mission is also to study the relationships between the rings and the satellites, the interactions between the magnetospheric plasma and the satellites, the rings and the atmosphere of Titan. Cassini will study not only each part of the Saturnian system but also the system as a whole. The mission carries 26 scientific investigations, which include 18 instrument investigations, and eight interdisciplinary investigations that will rely on the data obtained by more than one instrument. The Saturn Orbiter carries onboard 12 scientific instruments while six instruments have been packed into the Huygens Probe.

Cassini/Huygens is an international undertaking. The planning for the Cassini/Huygens mission was started in 1984 when a joint ESA–NASA study was initiated. It required more than 6 yr of intensive studies and careful planning to allow both agencies to embark in 1990 on what is the most ambitious outer planet mission that has been undertaken so far. All together, 18 nations were involved during the hardware development phase. The ITALIAN SPACE AGENCY (ASI, Agencia Spatiale Italiana) has provided major elements of the Orbiter through a bilateral cooperation programme established with NASA in 1992. The 4 m High Gain Antenna is one of the elements contributed by ASI.

Figure 1. The launch of the Cassini/Huygens spacecraft on 15 October 1997 from Cape Canaveral, Florida.

The journey to Saturn takes 7 yr. Although Cassini/Huygens was launched by the most powerful rocket (Titan IVB/Centaur) that existed at that time, the 5.6 t spacecraft was too heavy to be launched on a direct trajectory to Saturn. The path of its journey includes several planetary encounters that are used for gravity-assist maneuvers to accelerate the spacecraft towards the next planet. The first three planetary encounters took place at Venus on 27 April 1998 and 24 June 1999 and at Earth on 17 August 1999. The last planetary encounter before Saturn will occur in late December 2000 at Jupiter. The Cassini/Huygens spacecraft will arrive at Saturn in July 2004. The nominal mission in the Saturnian system will last 4 yr. The Cassini/Huygens spacecraft will be placed in orbit around Saturn on 1 July 2004, when the Saturn Orbit Insertion (SOI) maneuver will take place. Five months after SOI, the Huygens Probe will be targeted and released from the Orbiter for its encounter with Titan 3 weeks later. In late November 2004, Huygens will enter and descend by parachute through the thick hazy atmosphere of Titan until it reaches the surface. Following the Huygens mission, the Cassini Saturn Orbiter will undertake its own mission of exploration of the Saturnian system in the course of 75 orbits around the planet. Forty-four of the orbits will include a close flyby of Titan at altitudes as low as 850 km above the surface, both for the purpose of making observations and for gravity-assist maneuvers. At the end of the mission, Titan will hold the record as the planetary body most used for gravity assist manoeuvres.

The mission is named in honor of the French–Italian astronomer JEAN-DOMINIQUE CASSINI. In the period 1671–85 he discovered four of Saturn's satellites as well as the large gap in the ring (the Cassini division) which is named after him. The Probe is named in honour of the Dutch astronomer CHRISTIAAN HUYGENS, who discovered Titan in 1655 while he was observing Saturn's rings. Huygens was also the first to explain the true nature of the rings.

The scientific objectives and the payload

The Cassini/Huygens mission will build on the discoveries made by the highly successful VOYAGER missions and will allow a quantum leap in the understanding of the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Saturnian system, the formation and evolution of a planetary system and the origin and evolution of the solar system as a whole. Titan, the largest moon of Saturn, and the second largest moon in the solar system after Jupiter's Ganymede, is a special target of the mission. Titan is a unique type of object in the solar system. It is shrouded by a thick methane-rich, hazy, nitrogen atmosphere. The surface pressure and temperature are 1.5 bar (1.5 times that

Table 1. Orbiter instruments.

Instruments	*Participating countries*	*Measurements*	*Techniques*
• Optical remote-sensing instruments			
Composite Infrared Spectrometer (CIRS)	USA, Austria, France, Germany, Italy, UK	High-resolution infrared spectra, 10–1400 cm^{-1}	Spectroscopy using 3 interferometric spectrometers
Imaging Science Subsystem (ISS)	USA, France, Germany, UK	Photometric images through filters, 0.2–1.1 mm	Imaging with CCD detectors; 1 wide angle camera (61.2 mrad field of view); 1 narrow angle camera (6.1 mrad field of view)
Ultraviolet Imaging Spectrograph (UVIS)	USA, France, Germany	Spectral images, 55–190 nm, occultation photometry, 2 ms; H and D spectroscopy, 0.0004 nm resolution	Imaging spectroscopy, 2 spectrometers Hydrogen–deuterium absorption cell
Visible and Infrared Mapping Spectrometer (VIMS)	USA, France, Germany, Italy	Spectral images, 0.35–1.05 μm (0.073 μm resolution), 0.85–5.1 μm (0.166 μm resolution); occultation photometry	Imaging spectroscopy, 2 spectrometers
• Radio remote-sensing instruments			
RADAR	USA, France, Italy, UK	Ku-band RADAR images (13 777.5 MHz); radiometry, <0.5 K resolution	Synthetic aperture radar; radiometry with a microwave receiver
Radio Science Subsystem (RSS)	USA, Italy	Ka-, S- and X-bands; frequency, phase, timing and amplitude	X- and Ka-band transmissions to Cassini; Ka-, S- and X-band transmissions to the Earth
• Particle remote-sensing and *in situ* measurement instrument			
Magnetospheric Imaging Instrument (MIMI)	USA, France, Germany	(1) Image energetic neutrals and ions <10 keV–8 MeV nucleon; composition (2) 10–265 keV/e ions; charge state; composition; directional flux (3) 20 keV–130 MeV ions; 15 keV to >11 MeV electrons; directional flux	(1) Particle detection and imaging. Ion–neutral camera (time of flight, total energy detector) (2) Charge–energy–mass spectrometer (3) Solid state detectors with magnetic focusing telescope, and aperture controlled ~45Â° field of view
• *In situ* measurement instruments			
Cassini Plasma Spectrometer (CAPS)	USA, Finland, France, Hungary, Norway, UK	Particle energy/charge: (1) 0.7–30 000 eV/e (2) 1–50 000 eV/e (3) 1–50 000 eV/e	Particle detection and spectroscopy: (1) Electron spectrometer (2) Ion mass spectrometer (3) Ion beam spectrometer
Cosmic Dust Analyzer (CDA)	Germany, Czech Republic, France, Norway, UK, USA, ESA/SSD	Directional flux and mass of dust particles in range 10^{-16}–10^{-6} g	Impact-induced plasma cloud currents measurements
Dual Technique Magnetometer (MAG)	UK, Germany, Italy, USA	$\boldsymbol{B}$: dc to 4 Hz up to 256 nT; scalar field dc to 20 Hz up to 44 000 nT	Magnetic field measurement; flux gate magnetometer; vector–scalar magnetometer
Ion and Neutral Mass Spectrometer (INMS)	USA, Germany	Fluxes of positive ions and neutrals in mass range of 2–66 amu	Mass spectrometry
Radio and Plasma Wave Science (RPWS)	USA, Austria, France, Sweden, UK, ESA/SSD	$\boldsymbol{E}$: 10 Hz–2 MHz $\boldsymbol{B}$: 1 Hz–20 kHz Plasma density and temperature	Radio frequency receivers; 3 electric dipole antennas; 3 magnetic search coils; Langmuir probe

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 2. Huygens instruments.

Instruments	*Participating countries*	*Measurements*	*Techniques*
Huygens Atmospheric Structure Instrument (HASI)	Italy, Austria, Germany, Finland, France, Norway, Spain, USA, UK, ESA–SSD	Temperature: 50–300 K Pressure: 0–2000 mbar Electrical conductivity: 10^{-15} Ω m^{-1} to ∞ Acceleration: 1 μg–20 mg Relative permittivity: 1 to ∞ Acoustic: 0–5 kHz, 90 dB at 5 mPa	AC E field: 0–10 kHz, 80 dB at 2 μV m^{-1} Hz$^{-0.5}$ DC E field: 50 dB at 40 mV m^{-1} Meteorology sensors: pressure and temperature; accelerometry measurements. E field receiver; mutual impedance probe; microphone; radar surface echo processing
Gas Chromatograph and Mass Spectrometer (GCMS)	USA, Austria, France	Mass range: 2–146 amu Dynamic range: >10^8 Sensitivity: 10^{-12} mixing ratio Mass resolution: 10^{-6} at 60 amu	Chromotography and mass spectrometry: 3 parallel chromatographic columns; quadrupole mass filter; 5 electron impact sources
Aerosol Collector and Pyrolyzer (ACP)	France, Austria, USA	2 samples: 150–45 km, 30–15 km altitude	3-step pyrolysis: 20 °C, 250 °C, 650 °C
Descent Imager and Spectral Radiometer (DISR)	USA, Germany, France	Upward and downward spectra: 480–960 nm, 0.87–1.7 μm, resolution 2.4–6.3 nm; downward- and side-looking images, 0.66–1 μm; solar aureole photometry, 550 nm, 939 nm; surface spectral reflectance	Spectrophotometry, imaging, photometry, and surface illumination by lamp
Doppler Wind Experiment (DWE)	Germany, Italy, USA	(Allan variance)$^{1/2}$: 10^{-11} (in 1 s), 5–10^{-12} (in 10 s), 10^{-12} (in 100 s), corresponding to wind velocities of 2–200 m s^{-1}, probe spin	Doppler shift of Huygens telemetry signal, signal attenuation
Surface Science Package (SSP)	UK, Italy, USA, ESA–SSD	Acceleration: 0–100g Tilt: ±60° Temperature: 65–110 K Thermal conductivity: 0–400 m W m^{-1} K^{-1} Speed of sound: 150–2000 m s^{-1} Liquid density: 400–700 kg m^{-3} Refractive index: 1.25–1.45	Direct measurements using 'laboratory' methods: impact acceleration; acoustic sounding, liquid relative permittivity, density and index of refraction

on Earth) and 95 K (–178 °C). The tropopause temperature is 70 K at an altitude of 40–45 km. In the upper stratosphere the temperature rises up to 160–200 K. The 1 mbar level is reached in the altitude range 180–200 km. The photochemistry resulting from the photodissociation of methane and nitrogen in the atmosphere makes Titan a very fascinating body to explore. It may harbor complex organic molecules that are thought to be similar to the ones that were making the chemical soup of the young Earth before the appearance of life. The cold environment of Titan has frozen this pre-biological soup in an early state of its development. The exploration of Titan will allow us to investigate the complex chemistry at work in an environment similar to the one that may have prevailed on Earth after the end of the heavy bombardment, 3.8 billion years ago, before the appearance of life.

The scientific objectives of the Cassini/Huygens mission are categorized by the main mission targets.

With regard to the moon Titan:

to determine atmospheric composition, investigate energy sources for atmospheric chemistry, study aerosol properties and cloud physics; measure winds and global temperatures; determine properties of the surface and infer internal structure; investigate the upper atmosphere and ionosphere.

With regard to Saturn:

to determine global temperature and wind fields, cloud properties and atmospheric composition and infer properties of the deep atmosphere; study properties of ionosphere; provide observational constraints on scenarios of Saturn's formation and evolution.

With regard to the ring system:

to study the configuration of the rings and the dynamical processes responsible for their structure; map ring composition and particle size distribution; study

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

interactions between the rings and Saturn's magnetosphere, ionosphere, atmosphere and satellites.

With regard to the icy satellites:
to investigate the composition and distribution of surface materials and define mechanisms of crustal and surface modifications; determine the general characteristics of the satellites, including placing constraints on bulk composition and internal structures; investigate interactions with the magnetosphere.

With regard to Saturn's magnetosphere:
to determine the configuration of the magnetic field; determine properties, including sources and sinks of magnetospheric charged particles; investigate the interaction of the magnetosphere with the solar wind and with Titan's atmosphere and ionosphere.

The Orbiter and the Probe payloads are listed in tables 1 and 2 respectively, in which we also give a brief description of the measurements that will be made and what will be studied. Interdisciplinary investigations that will require data from more than one instrument to be conducted are listed in table 3. An overview of the observations that are planned in the Saturnian system is described below.

The planned observations

Titan

The moon Titan is a prime target of the Cassini/Huygens mission. With a diameter of 5150 km, Titan is the second-largest satellite of the solar system after Jupiter's Ganymede (5262 km). It is larger than the planets Mercury (4878 km) and Pluto (2300 km) and the Earth's moon (3476 km). Titan has a thick, hazy atmosphere which obscures the surface at visible wavelengths. The Huygens Probe will measure the physical properties and determine the composition profile of the atmosphere of Titan. Huygens will also make images of the surface and measure its composition along the descent track. The resolution of the images will improve as the Probe approaches the surface. Near the surface, details of less than 1 m, perhaps even less that 10 cm, should be observable. A radar altimeter will measure the surface topography below the Probe. The wind profile will also be measured. The Cassini Saturn Orbiter and the Huygens Probe will explore Titan with a complementary approach. Each time the Orbiter flies by Titan, it will perform a set of atmosphere and surface remote-sensing observations. The surface of Titan will be mapped by the Orbiter's Titan radar mapper and possibly optical instruments that may be able to see the surface through 'atmospheric windows' in the 0.8–2 μm range. The Orbiter observations will be designed to provide global coverage of Titan's atmosphere properties and of the surface; they will also include re-observations of the atmosphere and surface along the flight path of the Probe.

Table 3. Interdisciplinary investigations.

IDS investigation	*Participating country*	*Brief objectives*
Magnetosphere and plasma	France	Study of plasma circulation and magnetosphere–ionosphere coupling
Rings and dust	USA	Study of rings and dust within the Saturnian system
Magnetosphere and plasma	USA	Study of the plasma environment in Saturn's magnetosphere
Atmospheres	USA	Study of the atmospheres of Titan and Saturn
Satellites	USA	Study of the satellites of Saturn
Aeronomy and solar wind interaction	USA	Study of aeronomy in the Titanian and Saturnian atmospheres
Titan atmosphere–surface interactions	USA	Study of Titan's atmosphere–surface interaction
Titan aeronomy	France	Study of Titan's aeronomy
Titan organic chemistry	France	Study of Titan's organic chemistry

Planet Saturn

Cassini will measure the physical properties of the atmosphere of Saturn and determine its composition all around the globe as a function of altitude. The properties of the ever-changing clouds will be investigated and the global circulation of the atmosphere will be studied. The fast rotation of the planet (10 h 12 min) will allow us to study repeatedly the diurnal variations of atmospheric properties. Because of the alignment of the planet's dipolar axis with its rotation axis, Saturn's magnetic field is different from that of the Earth, Jupiter, Uranus and Neptune. The generation mechanism of Saturn's magnetic field is a prime topic for study (see also SATURN: MAGNETOSPHERE). The detailed study of Saturn's magnetism will allow us to probe the interior of the planet and to understand better how the dynamo effect works inside planet Earth. The interior of the planet will also be probed by gravity measurement techniques based on the precise measurement of the spacecraft's trajectory above the planet. This will allow us to place constraints on the interior of the planet.

Rings

Saturn's famous rings will be imaged under all solar phase angles and from different geometrical perspectives. The configuration of the rings will be observed and the composition and the dynamical processes responsible for the ring structure will be inferred from images and occultations at optical and radio wavelengths. The rings are

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

thought to be composed mainly of water ice particles. The compositional variation of the rings is thought to be due to different types of impurities in the ice. The size distribution and the chemical composition of the ring particles will be obtained. Direct, *in situ* measurements of the dust and meteoroid particles within the rings will be obtained by Cosmic Dust Analyzer. An objective of Cassini is also to search for new moonlets within the rings and to study the interrelation between the rings and satellites, including the imbedded satellites. (See also SATURN: RINGS.)

Icy satellites

The icy satellites offer many new worlds to explore. Saturn has 18 known moons. Cassini will image and map their surfaces over the course of the orbital tour. This will allow us to obtain the composition and distribution of surface material. A major goal of the mission is to study the geological history of the satellites and to study the mechanisms responsible for the modification of their surfaces. Detection of the internal magnetic fields of the satellites is also an important objective of Cassini as Galileo at Jupiter has shown that surprises are to be expected. Radio measurements of the spacecraft's trajectory during each satellite flyby will allow us to infer their bulk compositions and to place constraints on their internal structures. Cassini will search for organic material on the satellite surfaces. One major goal of Cassini is to study the relationship between the icy satellites and the rings. Similarities in composition between the satellites and the rings will be searched for. Cassini will also search for and, if found, characterize the tenuous atmospheres around Saturn's moons. (See also SATURN: SATELLITES.)

Although the full operational capabilities of the Cassini/Huygens spacecraft and of the ground system will only be available 2 yr prior to arrival at Saturn, some cruise science and observations during each planet flyby (Venus, Earth–Moon, Jupiter) were carried out.

A specific science objective has been included in the Cassini/Huygens mission that will take advantage of the inclusion of both X-band and Ka-band uplink and downlink capabilities of the telecommunication subsystem. It will allow the search for gravitational waves to be supported. The gravitational wave experiments will be conducted for periods of at least 40 days duration; they have been scheduled during the three oppositions after the Jupiter flyby in December 2001, December 2002 and January 2004.

The Cassini/Huygens spacecraft

The Cassini/Huygens spacecraft (figure 2) is the largest and most complex outer planetary spacecraft ever built. It is twice the size of the Galileo spacecraft, which is currently exploring the Jovian system; it is 6.8 m tall and 4 m wide. Its launch mass, including Huygens, was 5650 kg. More than half of its launch mass was propellant (3130 kg).

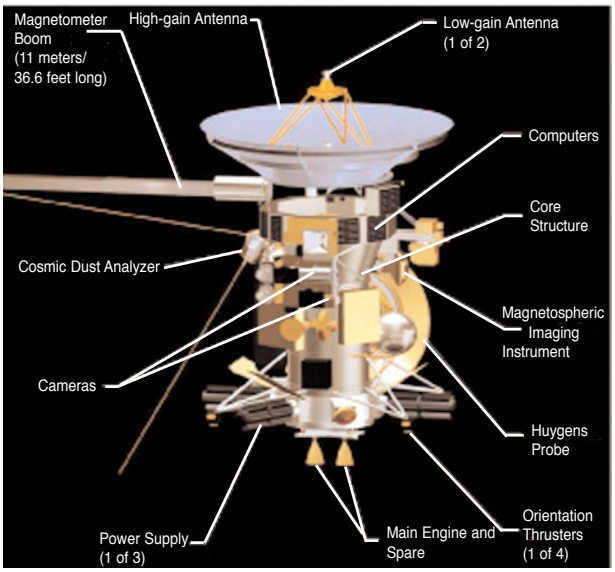


Figure 2. The Cassini/Huygens spacecraft.

The Saturn Orbiter

The Orbiter is a three-axis-stabilized spacecraft. The main body is formed by a stack consisting of the lower equipment module, the propulsion module, the upper equipment module and the High-Gain Antenna (HGA). Attached to this stack are the remote-sensing pallet, which carries the optical payload instruments, the fields-and-particles pallet, which carries most of the plasma particle and the dust instruments, and the Huygens Probe, attached sideways to the propulsion module. Electronic equipment of the radar is attached to the upper equipment module. The long electric antenna (shown in the deployed configuration) and the magnetic antennas of the Radio and Plasma Wave Subsystem (RPWS) are attached to the bottom part of the upper equipment module. The magnetometer boom canister is attached on the upper part of the upper equipment module, below the HGA. The 11 m long magnetometer boom is shown in its deployed configuration in the figure. The electrical power is provided by three Radioisotope Thermoelectric Generators (RTGs) which deliver more than 700 W of electrical power. The spacecraft electronics bus and several instrument electronics subsystems are housed in the upper equipment module which is thermally regulated to provide a benign (~10–20 °C) environment. While in the inner solar system, the spacecraft flies with its HGA pointed at the Sun, using it as a sunshade. At Saturn, using its 20 W X-band transmitter and its 4 m HGA, the spacecraft will be able to communicate with NASA's Deep Space Network at data rates up to 166 kbit s^{-1}. Two Low-Gain Antennas (LGAs) transmit, at a low rate (minimum 40 bit s^{-1}) data and receive telecommands when the HGA cannot be pointed at Earth.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

For propulsion and maneuvers, on its way to Saturn and during the Saturn orbital phase, the Orbiter uses one of the two redundant 445 N engines and 16 smaller 0.5 N thrusters clustered in groups of four in redundant pairs. The gimbal mechanisms of the main engines allow their thrust to be directed through the spacecraft's center of gravity. The main engines use the helium-pressurized hypergolic combination of monomethyl hydrazine ($N_2H_3CH_3$) fuel and nitrogen tetroxide (N_2O_4) oxidizer. The thrusters use hydrazine (N_2H_4). The main engines are covered by a retractable cover when not in use.

The Orbiter Command and Data Subsystem (CDS) receives ground commands via the Radio Frequency Subsystem (RFS). The CDS distributes the commands to the other subsystems, to the Orbiter payload instrument or to the Probe (when attached to the Orbiter). In turn, the CDS collects data from other subsystems, the Orbiter payload or the orbiter-mounted Probe Support Equipment and transfers them to one of the two redundant 1.8 Gbit Solid State Recorders (SSRs). The CDS subsequently retrieves those data from the SSR, formats and delivers them to the RFS for transmission to Earth via either the HGA or one of the two LGAs.

Dynamic control of the orientation of Cassini is provided by the Attitude and Articulation Control Subsystem (AACS). When high spacecraft stability is required, the spacecraft rotation is controlled by Reaction Wheel Assemblies (RWAs). The thrusters are used when fast rotation or acceleration rates are required. The AACS uses two redundant Inertial Reference Units (IRUs), two redundant Stellar Reference Units (SRUs) and a Sun Sensor Assembly (SSA). The AACS is capable of supporting a pointing accuracy of 1 mrad (3.4 arcmin) with a stability of 8 μrad (0.2 arcsec) and rotation rates of 0.02°–1° s^{-1}.

Thermal control of the Orbiter is accomplished by several means, the most visible being the black-and-gold Multi-Layer Insulation (MLI). In addition, the thermal control uses louvers positioned on the 12-bay electronics bus, electric heaters and Radio-isotope Heater Units (RHUs) and heat radiators. The infrared instruments (VIMS and CIRS) are thermally insulated from the spacecraft bus and they provide their own thermal control as an integral part of the instrument.

The Huygens Probe

The Huygens Probe is an entry probe. It has a mass of 330 kg. It comprises two main elements: the entry protection system and the descent module that is enclosed in it. The main elements of the Huygens Probe are shown in figure 3. The probe is attached sideways to the Saturn Orbiter (see figure 2). The Probe separates from the Orbiter with a relative velocity of 0.3 m s^{-1} and a spin rate of 7 rev min^{-1}, which provides stability during the 22 day coast to Titan and during the atmospheric entry. The entry subsystem consists of the front heat shield, which has a diameter of 2.75 m, and the after cover, both covered with silicate-based thermal-protection material for protecting the entry module from the radiative and convective heat fluxes (up to 1 MW m^{-2}) generated during the entry phase in the altitude range 350–220 km, where the Probe decelerates from about 6000 m s^{-1} down to 400 m s^{-1} (Mach 1.5) in less than 2 min. At Mach 1.5, at an altitude of about 170 km, the parachute-deployment sequence is initiated by a pyrotechnic device which pulls out the 2.59 m diameter pilot chute to pull the back cover away from the rest of the assembly. As it goes, the back cover pulls the 8.3 m diameter main parachute from its container. This canopy inflates during the supersonic phase to decelerate and stabilize the Probe through the transonic regime. The front shield is released at about Mach 0.6, 30 s after inflation of the main parachute. In fact the main parachute is sized to provide sufficient deceleration to guarantee a positive separation of the front shield from the descent module.

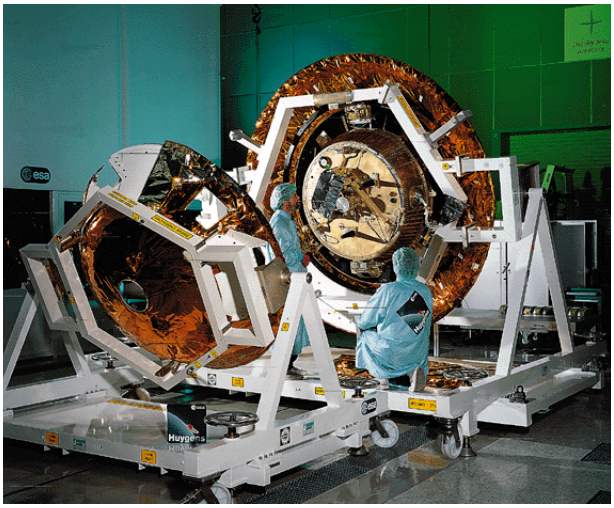


Figure 3. The Huygens Probe during final integration in Europe before it was shipped to NASA's Kennedy Space Center for the integration activities with the Cassini Saturn Orbiter.

Cassini/Huygens mission overview

The Cassini/Huygens spacecraft was launched at 08:43 GMT on 15 October 1997 by a Titan IVB/Centaur rocket from Cape Canaveral Air Force Station in Florida. The 5.6 t spacecraft was too heavy to be injected directly into a Saturn-approach trajectory; it requires several planetary boosts to reach its destination. The interplanetary trajectory (figure 4) includes gravity assists from Venus (26 April 1998 and 24 June 1999), Earth (18 August 1999) and Jupiter (30 December 2000).

Arrival at Saturn of the Cassini/Huygens spacecraft is planned to take place on 1 July 2004. At this point, a major maneuver will be executed while the spacecraft crosses the ring plane. The main engine will

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

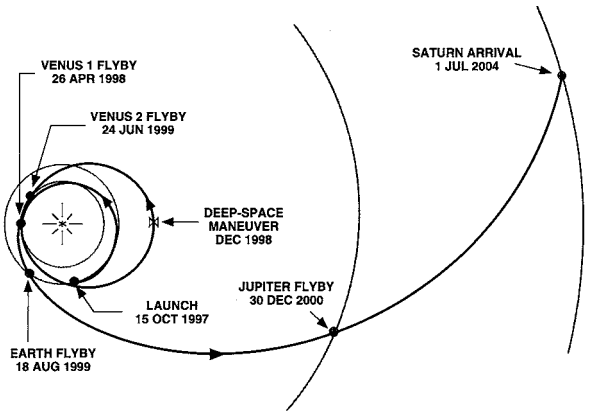


Figure 4. The interplanetary trajectory of the Cassini/Huygens spacecraft.

burn for about 90 min to slow down the spacecraft so that it will be captured in orbit around Saturn. The Probe is to be released from the Saturn Orbiter at the end of the initial orbit around Saturn, nominally 22 days before Titan encounter. Prior to releasing the Probe, the Orbiter is targeted on a collision trajectory with Titan as the Probe has no maneuvering capability by itself. Shortly after Probe release, the Orbiter will perform a deflection maneuver, so as not to enter the atmosphere of Titan. This maneuver will also set up the radio communication geometry between the Probe and the Orbiter for use during the Probe descent. It will will also set up the initial conditions for the subsequent satellite tour which starts after completion of the Probe mission.

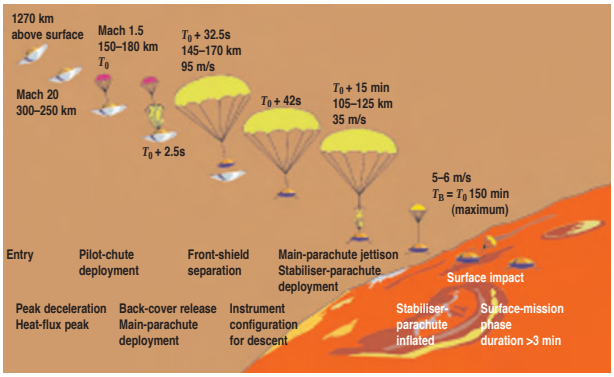


Figure 5. The Huygens entry and descent scenario.

Huygens' entry into the Titan atmosphere is planned for 27 November 2004. The descent under parachute will take a maximum of 2½h. On the way down, Huygens will measure the physical and the chemical properties of the atmosphere and take images of the surface underneath the Probe. During its descent and possibly for a short period on the surface, the Probe will radio the data acquired to the Cassini Saturn Orbiter, which will receive them via its HGA pointed at Titan. The Orbiter will in turn send the data back to Earth after the Huygens mission is over, also using its HGA, then pointed at Earth. The Huygens entry and descent scenario is illustrated in figure 5.

After completion of the Probe mission, the Saturn Orbiter will start its 4 yr tour of the Saturnian system, the so-called satellite tour. In the current mission scenario, it consists of more than 70 Saturn-centered orbits, connected

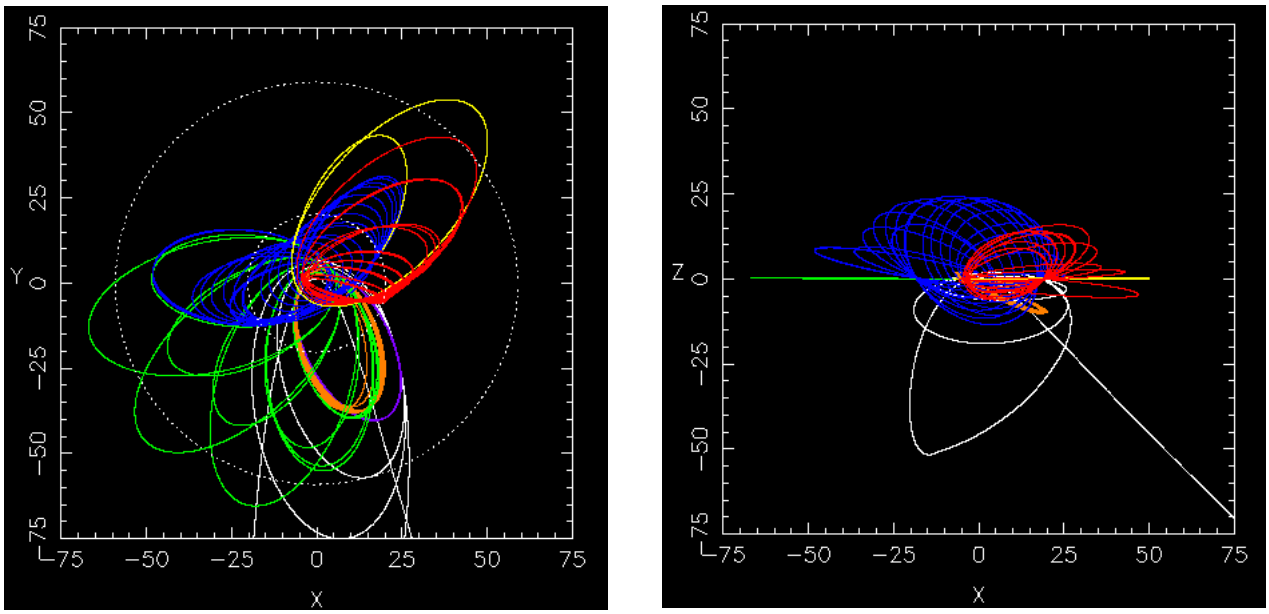


Figure 6. Baseline satellite tour in the Saturnian system; on the left side the projection of the trajectory is shown in the equatorial (X–Y) plane of Saturn, while on the right side the projection perpendicular to the equatorial (X–Z) plane is shown. Dimensions in X, Y, Z are indicated in Saturn radii (1 Saturn radius (R_S) = 60 330 km). Note that the apoapsis of the first orbit lies at ~170R_S, beyond the limit of the figure.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

by Titan gravity-assist flybys or propulsive maneuvers. The size of these orbits, their orientation to the Sun–Saturn line and their inclination to Saturn's equator are dictated by the various scientific requirements, which include Titan ground-track coverage, icy-satellite flybys, Saturn, Titan or ring occultations, magnetosphere coverage, orbit inclinations and ring-plane crossings. The geometry of the tour is illustrated in figure 6.

The flight operations

The flight operations of the Cassini/Huygens mission operations are carried out at the JET PROPULSION LABORATORY (JPL) in Pasadena, California. The uplink and the downlink operations are carried out from the Mission Control Center at JPL. The data are collected via NASA's deep space network and stored on computers at JPL, where the spacecraft's health and status are routinely assessed. The Cassini/Huygens operations are based on a distributed operations concept. In this concept, the instrument teams analyse their instrument data and prepare instrument operation sequences from their home institutes and transmit them to JPL for uplink. During the main mission, from the start of the Saturn encounter until the end of the mission, the Orbiter instrument teams will interact daily with the spacecraft operations team at JPL. The flight operations of the Huygens Probe are carried out from ESA's Operations Center at ESOC in Darmstadt, Germany, where the Huygens Probe Operations Center (HPOC) has been established. All Huygens-related mission activities (periodic in-flight checkouts and Probe mission phase) are carried out from the HPOC. The Huygens telemetry data are routed to the HPOC, where they are processed and distributed to the Huygens investigators.

Jupiter flyby

The Cassini spacecraft has made a successful flyby of Jupiter during a gravity-assisted swing toward Saturn and in addition has been used to try out some of its instrumentation, making important scientific observations of Jupiter and its moons. The Cassini observations were made in conjunction with observations by the Galileo spacecraft, also orbiting Jupiter. The observations by the two spacecraft were complemented by nearly simultaneous images from the Earth-orbiting Hubble Space Telescope and Chandra x-ray Observatory. Earth–based radio and optical telescopes were also used.

The Galileo spacecraft has been in orbit around Jupiter since December 1995 while Cassini observed Jupiter from 1 October, 2000 to 31 March 2001. Cassini made its closest approach on 30 December 2000, at a distance of about 9.7 million kilometers (6 million miles). This gave Cassini a boost from Jupiter's gravity that accelerated the spacecraft by about 2 km s^{-1} (about 4500 miles per hour) and will enable it to reach its ultimate destination, Saturn, in July 2004 for the planned four-year study.

A higher–than–normal drag that was detected on one of Cassini's reaction wheels did not reappear. The reaction wheels are used to rotate the spacecraft in different directions, and the problem led to suspension from 19 December to 28 December of scientific observations that would have required pointing the spacecraft, such as for taking pictures. Only two experiments, the Ion and Neutral Mass Spectrometer and Radio Science, and the Huygens probe for studying Saturn's moon Titan, did not take data during the study of Jupiter.

Observational results include revealing new data on the dynamics of Jupiter's magnetosphere, its radiation belts, aurorae and its interactions with the solar wind. See JUPITER, JUPITER: SATELLITES, JUPITER: RADIO EMISSIONS, JUPITER ATMOSPHERE, IO: PLASMA TORUS.

Web update (31 October 2002)

During July 2002, NASA's Cassini spacecraft took test images of a star that reveals successful results from an extended warming treatment to remove haze that collected on a camera lens during 2001. The quality of the new images is virtually the same as star images taken before the haze appeared. In the most recent treatment, the camera had been warmed to 4 degrees Celsius (39° F) for four weeks ending 9 July. Four previous treatments at that temperature for varying lengths of time had already removed most of the haze. The camera usually operates at –90° C (–130° F).

Bibliography

Hill T W 2002 Magnetic moments at Jupiter *Nature* **415** 965–966

2001 First Results from Cassini: Venus and Earth swingbys (11 articles) *Journal of Geophysical Research* **106 A12**

ESA 1997 *ESA Bulletin* No 92, November (contains five articles describing several aspects of the Cassini/Huygens mission)

NASA 1997 Passage to a Ringed World; the Cassini–Huygens Mission to Saturn and Titan *NASA SP 533*

Russell C (ed) 2000 The Cassini/Huygens Mission *Space Sci. Rev.* in preparation

Wilson A (ed) 1997 Huygens, Science Payload and Mission *ESA SP 1177*

ESA Huygens website http://sci.esa.int/huygens

JPL Cassini website http://www.jpl.nasa.gov/cassini

Jean-Pierre Lebreton and Dennis L Matson

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Cassiopeia

(abbrev. Cas, gen. Cassiopeiae; area 598 sq. deg.) A northern constellation which lies between Cepheus and Andromeda, and culminates at midnight in early October. It is named after Queen Cassiopeia, wife of King Cepheus and mother of Andromeda in Greek mythology, and is usually shown on early celestial charts as a seated figure. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A conspicuous constellation, Cassiopeia is easily recognized by the characteristic letter 'W' (or 'M') formed by its five brightest stars: ε (mag. 3.4), δ (Ruchbah, mag. 2.7), γ (mag. 2.2), α (Shedar, mag. 2.2) and β (Caph, mag. 2.3). There are three other stars of magnitude 4.0 or brighter. γ Cassiopeiae, the central star of the 'W', is a rapidly rotating irregular variable, which spasmodically throws off a shell of material from its equatorial zone and simultaneously fades by up to 1.4 magnitudes. Other interesting variables are R Cassiopeiae, a Mira-type star (range 4.7–13.5, period about 430 days), and RZ Cas, an eclipsing binary (range 6.2–7.7, period 1.2 days). η Cassiopeiae is a fine binary with yellow (G0) and red (M0) components, magnitudes 3.5 and 7.4, separation 12.5″.

The Milky Way passes through Cassiopeia and the constellation contains a number of open star clusters, including M52 (NGC 7654), which contains more than 100 stars between ninth and twelfth magnitude, M103 (NGC 581), which contains about 40 stars between magnitudes 8 and 12, and NGC 225, which has about 20 stars fainter than ninth magnitude. Cassiopeia is also the 'graveyard' of the supernova represented by the powerful radio source Cassiopeia A and of that observed by the Danish astronomer Tycho Brahe (1546–1601) in 1572.

See also: Cassiopeia A, Tycho's star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassiopeia A

The brightest radio source in the constellation of Cassiopeia and the brightest cosmic radio source in the entire sky. It is a supernova remnant, an expanding cloud of gas that was ejected in a catastrophic stellar explosion. From the present size and rate of expansion of the cloud, the explosion that created it must have occurred around the year 1660. Cassiopeia A lies at a distance of about 10 000 light-years. At that distance, a typical supernova should attain a peak brightness greater than that of the brightest naked-eye star and perhaps comparable to the planet Venus, but no bright 'new star' was recorded in Cassiopeia at around that time. For whatever reason, the supernova that produced Cassiopeia A must have been of unusually low optical luminosity.

Although not conspicuous in visible light (only a few faint wisps of nebulosity can be seen) Cassiopeia A radiates across the electromagnetic spectrum from x-ray to radio wavelengths. Its intensity, or flux density, at a wavelength of 20 cm in the radio region, is 2477 janskys, and its angular diameter is about 5 arc minutes.

See also: flux density, radioastronomy, supernovae, supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Castelli, Benedetto (1578–1643)

Mathematician, born in Brescia, Italy, Benedictine monk, professor at Padua. GALILEO's closest scientific collaborator, he defended, and edited Galileo, he helped his sunspot research, inventing the method of projection so as to view safely the Sun's image with a telescope. His book on hydraulics, *Della Misura dell'Acque Correnti, or On the Measurement of Running Waters*, founded modern hydrodynamics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Castor

The star α Geminorum, apparent magnitude 1.58. This is in fact a very complex multiple star system. Observed with a moderately powerful telescope it is seen to be double. The components of this binary system are a pair of white stars, α Gem A (apparent magnitude 1.9, spectral type A1V), and α Gem B (3.0 and A2V). Their current separation is 3.1″ at position angle 76°; the period is 420 years. Each of these component stars is itself a spectroscopic binary, their periods being 2.9 days (α Gem A) and 9.2 days (α Gem B).

The system has another more distant component, α Gem C (the variable star YY Geminorum), 72.5″ distant at position angle 164°. This is an eclipsing binary, its components being a pair of red dwarfs of spectral type M1, their combined magnitude varying between 9.2 and 9.6 in a period of 19.6 h. Although part of the gravitational system, this component is so far from the main pair—about 1000 astronomical units—that its orbital period must be of the order of a million years.

This sextuple star system is situated 52 light-years away, and has a parallax of 0.063″; the combined absolute magnitude of the principal components is 0.6.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cat's Eye Nebula (NGC 6543)

A planetary nebula in the constellation Draco, position RA $17^{h}\ 58.6^{m}$, dec. $+66^{\circ}\ 38.0'$, named for its oval shape and greenish color. It is about $20''$ in diameter, and of ninth magnitude; the central star is tenth magnitude. William Huggins' spectroscopic observation of the Cat's Eye in 1864, revealing a spectrum consisting of three bright lines, provided proof that not all 'nebulae' consist of stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cataclysmic Binaries: Classical and Recurrent Novae

The Latin word *nova*, meaning 'new', was applied historically to the sudden appearance of a bright star, not previously recorded. Such *novae stellae* appear in Eastern records from about 1500 BC and in European records in the past three centuries (many, however, are ancient SUPERNOVAE). The discovery of a different kind of nova, dwarf novae, in the mid-nineteenth century, led to the use of *classical* nova for the original type. Until 1887, when fainter novae began to be discovered by wide-field photography, only six novae had been found in the post-telescopic era. The brightest of these were Nova Vulpecula 1670 (magnitude 2.6 at maximum), Nova Ophiuchi 1848 (mag. 4.2) and Nova Cygni 1876 (mag 3.0).

Post-outburst observations of Nova Oph showed that after a few years it faded to magnitude 13.5, which demonstrated that whatever the cause of the dramatic nova explosion, it had not destroyed the star altogether. The realization that some novae have repeated eruptions (e.g. Nova Corona Borealis in 1866 and 1946) underlined the fact that the process is *cataclysmic*, rather than catastrophic. With current understanding of the cause of eruptions (see below) it is believed that all novae must be recurrent, and that all types of cataclysmic variable must regularly become novae, but the intervals between eruptions are in general much greater than the few centuries of recorded history.

At their brightest, novae typically reach absolute magnitude $M_v \sim -9$, which would make them easily detectable throughout our Galaxy were it not that novae are concentrated towards the Galactic Plane and Galactic Center, which causes strong interstellar extinction of the light from the most distant novae. As a result, although there may be as many as ~80 novae per year erupting in the Galaxy, on average only about three per year are detected. In the Magellanic Clouds and the Andromeda Galaxy (M 31) novae are readily discovered (at maximum being at apparent magnitudes $m_v \sim 10$ and ~15 respectively) and with large ground-based telescopes and CCD detectors novae can be studied even in the Virgo cluster of galaxies, at brightest magnitude $m_v \sim 23$.

A list of novae that reached apparent magnitude 3.0 or brighter is given in table 1. Variable star designations are used in table 1; for example, Nova Herculis 1934 is also known as DQ HERCULIS.

Light curves of novae

The variation of apparent visual brightness during a nova eruption is the most straightforward of measurements to make. Although it is rare to discover a nova much before its maximum brightness, the rise can sometimes be delineated through the chance appearance of the nova on wide-field photographic survey plates. There is a wide range of speeds with which novae pass through their eruption behavior, but all can be approximately described by the schematic curves shown in figure 1.

The initial rise of typically 6–13 magnitudes occurs in one or two days, often with a temporary halt lasting a few hours for fast novae to a few days for slow novae, before a less steep final rise of ~2 mag to maximum. The early decline is smooth for all but the slowest novae, where irregular variations are seen. The *speed class* of a nova is defined by the rate at which the early decline takes place. This is measured as the time t_2 or t_3 taken to fall 2 or 3 magnitudes respectively; the definitions are listed in table 2.

Table 2. Speed classes of novae.

Speed class	t_2 (days)
Very fast	<10
Fast	11–25
Moderately fast	26–80
Slow	81–150
Very slow	151–250

Starting about 3 magnitudes below maximum visual luminosity, through a ~2.5 magnitude drop known as the transition region, novae show a variety of behaviors. Some, particularly the faster novae, continue to decline smoothly. Others pass through a deep minimum, as deep as 7–10 mag, lasting 2 to 3 months, after which the smooth decline continues. Others show quasi-periodic brightness variations with amplitudes 1-1.5 mag and quasi-periods typically in the range 5–25 days.

The final decline to the post-nova is usually smooth, but in a few cases dwarf nova outbursts can begin late in the decline (e.g. Nova Per 1901, which commenced 2–3 mag outbursts in 1950, with intervals of ~2 years).

Absolute magnitudes of novae at maximum brightness, M_v(max), can be found if the distances and interstellar extinctions can be determined. The expansion parallax (see below) for some individual novae, supplemented with maximum brightnesses of novae in the Magellanic Clouds and M31, show that there is a strong correlation between t_2 and M_v(max) or M_{pg}(max) (for blue region photographic magnitudes) known as the 'absolute magnitude versus rate of decline relationship'. This may be expressed by the equation $M_{pg}(\max) = 3.35 \log t_2 - 12.2$. Application of this relationship gives a valuable method of finding the distances to extragalactic systems in which novae occur. Another technique uses the fact that all novae, almost independent of speed class, reach $M_v \sim -5.5$ 15 days after maximum light.

Satellite observations of the ultraviolet brightnesses of novae show that the peaks of the flux distributions move steadily to shorter wavelengths after maximum light. The early decline in visual luminosity is due almost entirely to this redistribution of flux towards ultraviolet wavelengths.

Infrared light curves of novae behave very differently from visual light curves. The infrared flux grows slowly during the initial decline and increases greatly if there is a deep minimum during the transition region. The flux distribution of the IR component is typical of emission

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Bright novae.

Star	Eruption date	Magnitude at maximum	Magnitude at minimum	t_2 (days)	v (km s^{-1})
V603 Aql	1918	−1.1	12.0	4	1700
Q Cyg	1876	3.0	15.6	5	
V476 Cyg	1920	2.0	17.2	7	790
V1500 Cyg	1975	2.2	16.3	2.9	1180
DQ Her	1934	1.3	14.5	67	350
V446 Her	1960	3.0	15.0	5	1380
V553 Her	1963	3.0	15.0	16	580
CP Lac	1936	2.1	16.6	5	2400
GK Per	1901	0.2	13.0	6	1300
RR Pic	1925	1.0	11.9	80	475
CP Pup	1942	0.5	15.0	5	710
V630 Sgr	1936	1.6	19	4	2120
CK Vul	1670	2.6	>21	40	

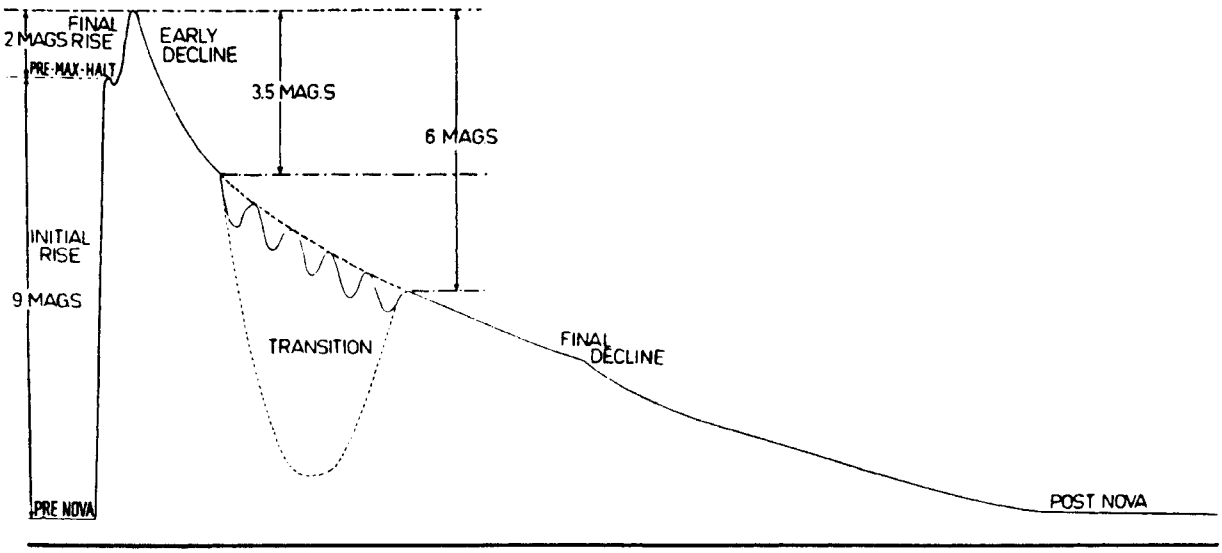


Figure 1.

from dust grains with temperatures ∼1000 K, and is interpreted as absorption of short-wavelength radiation as dust grains grow in expanding and cooling gas ejected by the nova explosion. In extreme cases the dust becomes optically thick in the visible, causing the deep visual minimum.

Adding the visual, ultraviolet and IR fluxes shows that the visual light curve of a nova is very misleading: the total luminosity of a nova in fact remains almost constant at least until well into the transition region. These 'plateau' luminosities of novae during the early phases lie in the range (2–4) $\times$ 10^4 $L_\odot$. These are roughly the Eddington luminosities for stars of about one solar mass. The short-lived peak luminosity can be a few times higher still, especially for fast novae. These observations show that radiation pressure is probably the motivating force for the ejection of gas during a nova explosion.

Spectra of novae

On the final rise to maximum light the spectrum of a nova resembles that of a supergiant; the spectral class is of early B type, moving to later type and reaching about A5 at maximum. The lines are very broad and there is a shortward Doppler shift of the wavelengths; both the shifts and the widths are (1–2) $\times$ 10^3 km s^{-1}, indicating a rapidly expanding optically thick shell. P Cygni line profiles are often seen, again pointing to an expanding shell. After maximum light the changes in spectrum are quite complex, indicative of several expanding envelopes with different velocities. On the initial decline a set of absorption lines resembling an A or F type supergiant, with a blueshift greater than that of the pre-maximum spectrum, develops. This is known as the *principal spectrum*; its velocity displacement is strongly correlated with speed class, with velocities up to 4000 km s^{-1} for the very fast novae and as low as 100 km s^{-1} for the slowest. The absorption

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

lines often show multiple structure, indicative of many subshells or blobs of gas being present in the expanding shell.

Soon after maximum light the strongest lines develop P Cyg profiles and large numbers of emission lines appear, chiefly of H, Ca II, Na I and Fe II in the visible. Later, the principal spectrum (identified by its velocity shift) weakens and forbidden lines of [O I], [O II] and [N II] appear—these are characteristic of gas at low densities.

Another distinct suite of absorption lines—the *diffuse enhanced spectrum*—appears soon after the principal spectrum has become clear, but blueshifted at about twice that of the latter. Yet another system—the *Orion spectrum*—appears one to two magnitudes down from maximum, and is characterized by single components, displaced blueward at least as much as the diffuse enhanced system.

At a late stage in the nova eruption, a *nebular spectrum* develops, at first merely being strengthening of the principal spectrum's forbidden emission lines, but with the ultimate appearance of [O III], [Ne III], giving the nova spectrum an appearance similar to that of a PLANETARY NEBULA. In the highest energy novae successively higher stages of ionization are seen, up to ions typical of coronal conditions: e.g. lines from [Fe X] and [Fe XIV].

The evolution of a nova spectrum is generally understood as being generated by an expanding, inhomogeneous shell, ionized by an intense source of high-energy radiation (temperature 10^5–10^6 K) at its center. The latter must be the central star, heated by the explosion.

The spectra of novae can be used to determine the physical conditions and the chemical composition of the expanding shell. Novae typically show abundances of C, N, O, Ne, Mg and some heavier elements considerably enhanced compared with those of the Sun and other normal stars. The relative abundances differ among the novae; typical enhancements are factors of 10 to 100. The helium-to-hydrogen ratio is also often different from that in the Sun, being higher by up to a factor of 1.5 or lower by as much as a factor of 5. The higher abundance of carbon probably helps the formation of obscuring dust in some of the nova shells.

Many novae show greatly enhanced abundances of neon—sufficient for a class of *neon novae* to have been defined. Most neon novae are fast or very fast, but at least one (Nova Ophiuichi 1988) was a slow nova.

Nova shells

Evidence for ejection of a shell comes not only from the spectroscopically detected expansion, but also from direct imaging. The Hubble Space Telescope was able to resolve the shell of Nova Cygni 1992 467 days after eruption when the angular radius of the shell was only 0″.13. Radio emission from nova shells (in the 2–100 GHz frequency region) enables very long baseline interferometry (VLBI) also to image shells with angular resolutions down to 0″.1. Eventually, at least for relatively nearby novae, the shell may expand sufficiently that it can be seen directly in images made from the ground. Figure 2 shows the present

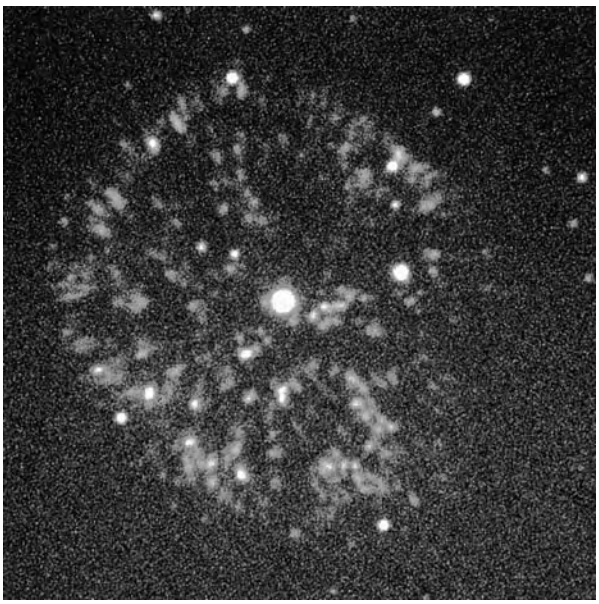


Figure 2.

appearance of Nova Persei 1901; the inhomogeneous structure is confirmed by such images. Many nova shells (e.g. Nova Herculis 1934) have more regular outlines, often elliptical in profile, showing that average ejection velocities vary systematically with direction. This is in accord with some spectroscopic analyses which detect 'equatorial' bands and 'polar' plumes ejected at different rates.

The angular rate of expansion directly measured by VLBI, or from the observed radius of the shell measured on photographs taken a known time (∼ decades) after the nova eruption, can be combined with the spectroscopic expansion velocity to determine the distance of a nova. From simple geometry, the *expansion parallax*, or its reciprocal, the distance d in parsecs, is given by $d = 0.20v/(\mathrm{d}\theta/\mathrm{d}t)$ pc, where v is the expansion velocity in km s^{-1} and $\mathrm{d}\theta/\mathrm{d}t$ is the rate of increase of angular radius, measured in seconds of arc per year. From this relationship we see that a nova at a distance of 1 kpc and expansion velocity 1000 km s^{-1} should show a shell of radius ∼4″ 20 years after eruption. However, with cooling of the central star, expansion away from this central exciting source, and decreasing surface brightness as the shell increases in area, most nova shells in fact cannot be imaged on the sky.

Analysis of the spectra of nova shells can provide estimates of the electron and ion densities in the regions where the emission lines are formed. If the total volume of the shell is known, this can give the total mass of the shell. However, the inhomogeneity of the shell leads to considerable uncertainty in how much of the apparent volume is actually filled with hot gas. Masses obtained spectroscopically lie in the range (1–30) $\times$ $10^{-5}M_\odot$. A possibly more reliable determination is obtained from the radio emission at a time when the shell has become

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

optically thin at radio wavelengths. Then all of the shell material contributes to the radio emission, and this arises from thermal (free–free) emission in the ionized shell. Shell masses determined this way are in the range (5–40) $\times 10^{-5} M_\odot$.

These ejected masses are compatible with the earlier statement that nova eruptions can occur many times in the life of a cataclysmic variable star: if the exploding star has a mass $\sim 1 M_\odot$, it can undergo hundreds or thousands of eruptions before depleting its mass significantly.

Before leaving nova shells, one other effect that may be directly imaged should be mentioned. For eight months after the eruption of Nova Persei 1901 nebulous patches were observed, apparently receding from the nova at ~20 arcmin per year. At the distance of that nova, this angular velocity corresponds to a linear velocity comparable to that of light. It was therefore concluded that the burst of light caused by the nova explosion was illuminating an interstellar dust layer in the vicinity of the nova. Such a phenomenon is known as a *light echo*, but has been seen for only two novae.

The physics of nova eruptions

As with all cataclysmic variables, novae are CLOSE BINARY STARS in which mass is being transferred from a star of normal dimensions (i.e. a main sequence star, or more rarely a giant), known as the secondary, onto a WHITE DWARF primary. At the time it was formed the white dwarf would have had no hydrogen in its interior, which would contain helium or carbon, oxygen or neon, magnesium, according to its mass. It is the continued accumulation of hydrogen-rich gas on the surface of the white dwarf that leads to the explosion that we recognize as a nova eruption.

Apart from an outer layer of about $10^{-4} M_\odot$, the interior of a white dwarf is degenerate. The essential physics of a nova explosion is contained in a comparison of the equation of state for degenerate matter, $P \propto \rho^\gamma$, with that for a perfect gas, $P \propto \rho T$, where P, ρ and T are respectively the pressure, density and temperature of the gas and γ is a constant. Suppose a region of the gas is heated slightly, then the pressure in a perfect gas increases and the gas can expand, thereby absorbing the heat. But the pressure in degenerate gas is independent of temperature, so no mechanical change takes place. If the temperature rises sufficiently, however, the Fermi temperature will be reached, above which the gas ceases to be degenerate—it becomes a perfect gas and its pressure will rapidly increase to match the density and temperature required by the change in equation of state.

Accumulation of hydrogen on the surface of the white dwarf builds a layer, the base of which becomes hotter and denser as it is compressed and is pushed deeper towards the hot interior. Nuclear reactions (at first the proton–proton chain, and later the CNO cycle) begin, which heat the gas. The mass $M_{\rm env}$ of the overlying layer sufficient to cause these nuclear reactions at its base depends quite sensitively on the mass of the white dwarf: it decreases from $\sim 1 \times 10^{-4} M_\odot$ at $0.6 M_\odot$ to $\sim 8 \times 10^{-7} M_\odot$ at $1.4 M_\odot$. This basal layer lies in a region of partial degeneracy, so the heating caused by nuclear reactions increases the temperature without expanding the layer. The increased temperature enhances the rate of energy production and a *thermonuclear runaway* ensues (as in a hydrogen bomb) until the Fermi temperature is reached and then the whole of the outer layer is blown into space. In many cases, a small part of the outer region of the degenerate core may be carried off in the explosion, creating the overabundances of C, O, Mg and Ne that are seen in some novae. Another cause of exceptional C, N or O abundances is that the explosion is so rapid that the CNO cycle does not reach the equilibrium that has produced the abundances of these elements present in normal stars.

If the rate at which mass is transferred from the companion star onto the white dwarf is $\mathrm{d}M/\mathrm{d}t$, then the time between eruptions will be $M_{\rm env}$ divided by $\mathrm{d}M/\mathrm{d}t$. Average mass transfer rates in cataclysmic variables are estimated to be $\sim 5 \times 10^{-9} M_\odot$ yr^{-1}, so the values of $M_{\rm env}$ given above imply intervals of $\sim 2 \times 10^4$ yr for a $0.6 M_\odot$ white dwarf and ~160 yr at $1.4 M_\odot$. We deduce that the observed recurrent novae occur probably at the high-mass end of a spread of white dwarf masses.

Computations of the evolution of a nova shell, using hydrodynamic computer codes, incorporate the initial expansion as the basal layer exceeds its Fermi temperature, plus the effects of the shock wave that occurs in the most energetic explosions, and the effect of radiation pressure. The resulting models match the observed range of nova behaviour quite well (e.g. the constant Eddington luminosity plateau of a nova eruption was predicted before it was observed). The models also show super-Eddington luminosities before the plateau. Early models failed to reproduce the duration of eruption and the decline branch of the eruption light curve, but these are now explained by interaction of the revolution of the secondary star in orbit: this transfers energy and angular momentum to the expanding shell (which forms a common envelope around the central binary) and leads first to a sustained wind phase and then to rapid shut down.

The diffuse enhanced and Orion absorption line systems seen in spectra of erupting novae have been interpreted as evidence for continued ejection of gas after the initial explosive phase. This also is present in the models: the high luminosity of the heated white dwarf produces a RADIATIVELY DRIVEN STELLAR WIND.

As the expanding envelope reduces in density and becomes more transparent it is possible to see further into it, allowing progressively hotter regions to be detected. This is the reason why the peak of the flux distribution moves into the ultraviolet in the early phases of a nova eruption. The freshly revealed surface of the white dwarf can have a temperature $\sim 10^6$ K; a $1.25 M_\odot$ white dwarf will then cool to 50 000 K in about 300 years.

The origin of the quasi-periodic oscillations that occur in some novae during the transition region (figure 1) is not certain; they may be radial pulsations of the envelope that surrounds the central binary star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 3. Recurrent novae.

Star	Dates	Magnitude at maximum	Magnitude at minimum	t_2 (days)
T CrB	1866, 1946	2.0	11.3	3.8
V394 CrA	1949, 1987	7.0	18.8	3.3
RS Oph	1898, 1933, 1958, 1967, 1985	4.3	12.5	4
T Pyx	1890, 1902, 1920, 1944, 1966	6.5	15.4	62
V3890 Sgr	1962, 1990	8.4	17.2	9
U Sco	1863, 1906, 1936, 1979, 1987	8.8	18.5	2.0
V745 Sco	1937, 1989	9.7	21	6.6
LMC-RN	1968, 1990	10.9	20	3

Recurrent novae

Only eight novae are currently known to be recurrent (a few other stars were once thought to belong to this class, but have proved either to be large-amplitude, long-recurrence interval dwarf novae, or are novae which later had dwarf nova outbursts). These are listed in table 3. One of these is in the Large Magellanic Cloud.

There are three subclasses of recurrent novae. The first, comprising T CrB, RS Oph, V3890 Sgr and V745 Sco, all have giant secondaries of M spectral type and hence orbital periods greater than 100 days (an M giant could not fit in a binary of shorter period). From the brevity of eruptions it is possible that some of their eruptions this century have been overlooked. The T CrB stars obey the M_v(max)–t_2 relationship. Their short recurrence times are partly due to the high rate at which giant companions transfer mass to their white dwarf companions, thereby rapidly building up the critical mass M_{env} that leads to explosion.

The second class of recurrent novae is that typified by U Sco and including V394 CrA and LMC-RN. These are among the fastest novae known; they also obey the M_v(max)–t_2 relationship. Again, some of their eruptions may have been missed because of their brevity. The declining parts of their light curves vary from eruption to eruption. All have orbital periods of about 1 day. Their spectra are dominated by helium lines in the post-nova stage.

The final star, T Pyx, is unique. It rises slowly to maximum, taking about 7 days for the initial rise and a further 20 to 30 days to maximum, followed by a slow decline. Its distance is very uncertain, so it is not known whether it obeys the M_v(max)–t_2 relationship. Photographs show a double shell, perhaps the ejecta from the 1944 and 1966 eruptions. It is suspected that in its quiescent state T Pyx may be steadily burning some of its hydrogen through nuclear reactions in its envelope.

Symbiotic stars

Among the SYMBIOTIC STARS are about a dozen that are classified as symbiotic novae. These are related to the T CrB type of recurrent novae: they probably all have long orbital periods (the known ones are from 200 to 1000 days) and some have had more than one eruption. However, the rise to maximum is very slow, in the range of 2 months to 16 years, and the duration of eruption is very long—from years to greater than a century. It is currently thought that the eruptions of symbiotic novae are the result of thermonuclear runaways on white dwarf primaries of lower mass than in the T CrB type.

Novae at minimum light

During the long intervals between eruptions, and after they have fully recovered from their latest eruptions, novae spend at least part of their time as non-outbursting, relatively inconspicuous stellar objects. A good example is DQ Herculis, which had not drawn attention to itself prior to its 1934 eruption, and which remained relatively unstudied in its post-novae state for 20 years before M F Walker discovered it to be an eclipsing binary with an orbital period of 4 h 39 min. This was the first cataclysmic variable star to be recognized as a short-period binary, though UX UMa and AE Aqr were previously known as spectroscopic binaries, but only retrospectively classified as nova-like variables.

Most novae eventually return to their pre-nova brightness (as found on archival photographic plates) and become indistinguishable from nova-like variables. For the brightest novae (table 1) the quiescent magnitude is quite bright. This implies that the currently known nova-likes probably include novae that erupted unnoticed in past centuries—and also that some of the nova-likes are 'pre-novae', in the sense that they will become novae in the future.

As mentioned earlier, some novae pass through a stage as dwarf novae during their post-nova decline. A few novae (Nova Her 1960, Nova Sgr 1919 and recurrent Nova Sgr 1962) showed dwarf nova outbursts prior to eruption; the first two of these started outbursts again in their post-nova phases.

There is evidence that at least some novae fade to very faint magnitudes (implying a large reduction in mass transfer from their secondary stars) centuries after their eruptions. Nova Vul 1670 (table 1) is an example, but more significant is that none of the naked eye novae recorded in Eastern records (and which would be expected to have post-nova apparent magnitudes in the range 10–14) have been recovered in modern surveys for nova-likes. The reason for the low state may be because heating of the secondary star by radiation from the hot post-nova white dwarf causes excessive mass transfer, and the secondary later adjusts to a slightly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

smaller radius, reducing mass transfer for many centuries, which maintains the long-term average rate of transfer determined by orbital evolution.

From the existence of the M_v(max)–t_2 relationship described above, and if nova remnants all had the same absolute brightness M_v(min) at minimum, it would be expected that the *range* $\Delta M_v = M_v$(min)-M_v(max) would be well correlated with rate of decline, t_2. (Note that the range in apparent magnitude $\Delta m_v = m_v$(min)—m_v(max) $= \Delta M_v$, and this is independent of interstellar extinction, since the extinction is applicable to both m_v(min) and m_v(max).) This is indeed the case, but there is a large spread in the relationship, which results from the dominance of the accretion disc at minimum light. At maximum of eruption, a nova would look approximately the same brightness independent of the direction from which it is observed. The minimum brightness, however, depends very strongly on the inclination of the binary orbit: if the system is seen at low inclination, i.e. the disc is seen nearly face-on, it is relatively bright. Seen at high inclination, however, the disc is edge-on and quite faint. The result is an unusually large Δm_v for eclipsing systems like DQ Herculis. The average M_v(min) for novae in quiescence is about +4, but the range is $2 \lesssim M_v(\text{min}) \lesssim 9$, correlated with orbital inclination.

Rare exceptions to this range of minimum magnitude can occur if the white dwarf primary has a strong magnetic field, preventing the formation of an accretion disc. An example is V1500 Cyg, which erupted from a minimum magnitude fainter than 21st (the magnitude listed in table 1 is the post-eruption outburst, probably representing a much higher rate of mass transfer than before eruption).

Bibliography

Bode M F and Evans A (ed) 1989 *Classical Novae* (New York: Wiley)

Gallagher J S and Starrfield S 1978 Theory and observations of classical novae *Ann. Rev. Astron. Astrophys.* **16** 171–214

Gehrz R D 1988 The infrared temporal development of classical novae *Ann. Rev. Astron. Astrophys.* **26** 377–417

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

Brian Warner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cataclysmic Variables: the Magnetic Systems

Cataclysmic variables (CVs) are CLOSE BINARY STARS in which the most massive component is a WHITE DWARF (WD) and the companion is usually a more normal star—commonly a main sequence star but occasionally a giant. The majority of CVs steadily transfer mass from the companion to the WD through an ACCRETION DISK surrounding the WD, but, just as about 5% of isolated WDs have detectable magnetic fields, so do about 25% of the WDs that occur in CVs. If the magnetic field is strong (greater than about 10 MG = 10×10^6 G) it prevents the formation of an accretion disk, and the result is that the gas leaving the companion is magnetically channeled direct to the surface of the WD. The WD itself rotates on its axis at the same rate as the revolution of the binary—it is locked to the secondary by the interaction of their magnetic fields. The prototype of this kind of object is AM HERCULIS, leading to the description of this class as AM Her stars, or (as the infalling gas radiates its energy in the strong magnetic field of the WD and this radiation is therefore cyclotron emission, which is strongly polarized), as polars. The strongest field so far measured in a polar is 230 MG in AR UMa; AM Her itself has one of the lowest fields found in a polar, 14 MG.

If the WD has a weaker magnetic field (less than a few million gauss) its interaction with the (intrinsic or induced) field of the companion is insufficient to force co-rotation, with the result that the WD has a different, usually much shorter, rotation period than the binary period. Furthermore, the field can only prevent formation of a disk quite close to the WD, with the result that an accretion disk does form, but the central part is missing—where the energy density in the field is greater than that of the kinetic inflow of gas and can therefore determine the accretion flow. The prototype of this kind of object is DQ HERCULIS. The group as a whole is usually known as intermediate polars (IPs) (being intermediate between strongly magnetic and non-magnetic), with DQ Her stars preferably kept as a description of the subset of most rapidly rotating WD primaries (e.g. DQ Her itself with 71 s rotation period)—although some authors use DQ Her stars as the general type. The majority of IPs have primaries rotating with periods of tens of minutes. Polarized radiation, modulated at the spin period of the WD, has been detected in only three IPs, presumably those with the strongest fields. Because of the large amount of unpolarized radiation from the (truncated) accretion disk, polarized radiation in IPs is very much diluted; furthermore, the weaker fields result in the cyclotron emission peaking in the infrared, where it is more difficult to detect.

The process of accreting matter onto the WD also transfers angular momentum. However, the result is not straightforward—not only is there a material torque on the WD, due to direct (but non-radial) impact of the gas, there are also torques resulting from the interaction of the magnetic field lines connecting the WD with the accretion disk, part of which is rotating faster than the WD and part is rotating slower. These torques can balance, with the result that the WD rotation period reaches an equilibrium state. However, in general the rotation rates are found to be increasing in some and decreasing in others—at typically only a few seconds per century—as they seek the equilibrium.

The magnetic field strengths in isolated WDs extend down to the current limit of detection—about 50 000 G, so it is to be expected that the same would be true of the WDs in CVs. Such fields are too weak to detect directly in CVs (i.e. by the ZEEMAN EFFECT or by the presence of polarization), but there is indirect evidence of the presence of such weak fields in some CVs through the appearance of dwarf nova oscillations (DNOs) during the outburst of some dwarf novae (also known as U GEM STARS) and intermittently in some nova-like variables (which are rather like dwarf novae stuck permanently in outburst). The DNOs are periodic brightness variations with periods typically in the range 5–40 s and are currently interpreted as magnetic channeling of accreting gas in the very inner region of the accretion disk, just above the surface of the WD. One difference from the IPs is that, whereas IPs have very stable brightness modulations, the DNOs vary rapidly in period—sometimes changing by a factor of 2 within a day. An explanation for this is that the intrinsic fields of the WDs are too small to prevent the accreting gas from slipping around the equator. Whereas in polars and IPs the accreting gas is coupled to the whole body of the WD, in the DNOs we see the effect of an equatorial belt spinning up in response to accreted angular momentum, carrying its (probably self-generated) magnetic field around with it. The existence of long-lived rapidly spinning equatorial belts after dwarf nova outbursts has recently been established through spectra obtained with the HUBBLE SPACE TELESCOPE. Not all dwarf novae or nova-like variables show DNOs, presumably because there are some CVs that have magnetic fields too small to effect any magnetic channeling at all of the accretion flow—in such cases the accretion disk would reach down to the surface of the WD. This may happen also in those dwarf novae where DNOs are seen to cease near the maximum of outburst; at such times the accretion flow is probably crushing the magnetosphere of the WD down to the WD surface.

In all of the classes of magnetic CV, modulated x-ray and EUV emission is seen—at the orbital period in polars, at the WD rotation period in IPs and at the DNO period in outbursts of dwarf novae. Such energetic radiation is emitted from the strongly shock-heated region as the channeled accretion flow impacts the surface of the WD, producing a very high x-ray luminosity (indeed, many of the polars and IPs were first discovered during x-ray surveys of the sky). Some of the radiation leaving such heated spots on the surface of the WD may intercept the surface of the truncated accretion disks in IPs and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

weakly magnetic CVs. In effect, a rotating beam sweeps across the surface of the disk—and also the side of the companion star pointing towards the WD. This radiation is partly processed (degraded) into the optical region of the spectrum, enabling us to detect indirectly the rotating beam. Two optical modulations with different periods can be produced—one through reprocessing from the disk and one from reprocessing on the companion star. In successive sweeps the companion has moved round a little in its orbit, so the signal from the companion has a longer period (this is like the difference between SIDEREAL PERIOD and SYNODIC PERIOD for planets), and this sometimes is the only optical period detectable. In the x-ray region we see the direct beam from the WD. As a result, in some IPs the period (sidereal) of the x-rays and the period (synodic) in the optical differ (their difference in frequency is exactly equal to the orbital frequency—which is how this model for IPs was established in the first instance).

In DQ Her and some outbursting dwarf novae and nova-like variables, observations of the changes in the rapid brightness modulations during eclipse by the companion star show phase changes which again point to a rotating beam emitted from the WD (see DQ HERCULIS).

The accumulation of hydrogen-rich material from the companion onto the surface of a WD leads eventually to a nova explosion (see NOVAE). The magnetic systems are no exception: Nova Cygni 1975 (also known as V1500 Cyg), one of the fastest and brightest novae of the 20th century, was found after it had settled back to its quiescent state to be a polar—or almost so. Its orbital period (determined spectroscopically and photometrically) is 0.139 612 9 d and its WD rotation period (determined from the period of the polarization modulation) is 0.137 154 d. The slight difference was caused by the effects of the nova explosion in 1975, during which the outer layers that were ejected from the WD exchanged some mass and angular momentum with the companion. The rotation period of the WD in V1500 Cyg is observed to be changing at such a rate that synchrony with the orbital period will be regained after about 180 yr. Three other desynchronized polars are known, which must have been undetected novae in recent centuries. There are two IPs among the known novae (Nova Persei 1901 and Nova Puppis 1963).

Bibliography

Hellier C 2001 *Cataclysmic Variable Stars: How and Why They Vary* (Praxis)

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

Wickramasinghe D T and Ferrario L 2000 *Publ. Astron. Soc. Pac.* **112** 873

*Bri*an Warner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Cavendish, Henry (1731–1810)

English chemist and physicist. Cavendish used a sensitive torsion balance (the *Cavendish balance*), which was made to oscillate under the gravitational force of attraction of large masses, to measure the value of the gravitational constant, G, and used his value to determine the mass of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Coordinates

Celestial coordinates are the means of specifying the location of a celestial object within some reference frame. In astronomy, it is necessary and convenient to represent the position of an object, such as a star or a planet, in several different coordinate systems according to the context in which the position is to be used. Coordinates of an object correspond to a particular way of expressing the position of a point with respect to a coordinate frame, such as a set of rectangular axes. In general, each position is represented by a set of three coordinates. In many cases, the distance of the object is not known, so two coordinates are sufficient to represent the direction to the object, although three direction cosines may be used. With the availability of more distance estimates in astronomy, the use of three space coordinates is becoming more common.

Many different coordinate systems are used to specify the positions of celestial objects. Each system depends on the choice of coordinate frame and on the way of specifying coordinates with respect to the frame. The term celestial coordinate frame is used here for a coordinate frame that does not rotate with the Earth (or other planet), so that the coordinates of stars, for example, change only slowly with time (see CELESTIAL COORDINATE SYSTEMS).

In general, an object is moving with respect to the coordinate frame, and the coordinate frame is moving and rotating in 'space'. Therefore, it is necessary to specify the time to which the three spatial coordinates refer and the time for which the coordinate frame is defined. These two times may be the same, but are often different. Also, positions may be of several kinds, according to whether or not allowances have been made for aberration, diurnal rotation, refraction and other factors that affect the direction in which an object is observed. These are referred to as mean, true and apparent places. EPHEMERIDES and catalogs that represent high-precision positions and properties of astronomical objects at given times must always be accompanied by a precise statement that specifies all of these various factors, if the numbers are to be used properly to their full precision.

The vector $\boldsymbol{r}$ from an origin to the object may be represented by rectangular coordinates (x, y, z); that is, by the projections of the distance $\boldsymbol{r}$ on the three axes. It can also be represented by spherical coordinates, in which direction is usually specified by the longitudinal angle (λ) in the x–y-reference plane and the latitudinal angle (β) from the reference plane. Less frequently the polar angle ($90^\circ - \beta$), or complement of the latitudinal angle, is used. The longitudinal and latitudinal angles are given different names and symbols in different coordinate frames.

The geometric transformations of positions between different coordinate systems, and the allowances for physical effects, may be carried out using the techniques of spherical trigonometry or those of vector and matrix algebra. The concept of the celestial sphere may be used for purposes of explanation, with the arcs on the sphere representing angles between directions. The center of the sphere can be located in many different places, but in most cases the sphere will be used to illustrate a particular reference frame with a specific origin. It must be remembered that when an object represented on the sphere is changing radial distances from the center along with transverse motion, these changes must be incorporated into the mathematics of the computation.

The designations used to indicate the principal origins of celestial coordinate frames are as follows.

(1) Topocentric: viewed or measured from the place of observation on the surface of the Earth.
(2) Geocentric: viewed or measured from the center of the Earth.
(3) Selenocentric: viewed or measured from the center of the Moon.
(4) Planetocentric: viewed or measured from the center of a planet (with corresponding designations for individual planets).
(5) Heliocentric: viewed or measured from the center of the Sun.
(6) Barycentric: viewed or measured from the center of the mass of the solar system.

The principal celestial reference planes through the appropriate origins are as follows.

(1) Horizon: the plane that is normal to the local vertical (or apparent direction of gravity) and passes through the observer.
(2) Local meridian: the plane that contains the local vertical and the direction of the axis of rotation of the Earth.
(3) Celestial equator: the plane that is normal to the axis of rotation of the Earth and passes through the Earth's center.
(4) Ecliptic: the mean plane (i.e. ignoring periodic perturbations) of the orbit of the Earth around the Sun.
(5) Planet's meridian: a plane that contains the axis of rotation of the planet and passes through a surface feature.
(6) Planet's equator: the plane that is normal to the axis of rotation of the planet and passes through the planet's center.
(7) Orbital plane: the plane of the orbit of a body around another (e.g. of a planet around the Sun or barycenter).
(8) Invariable plane or Laplacian plane: the plane that is normal to the axis of angular momentum of a system and passes through its center.
(9) Galactic equator: the plane through the central line of the local Galaxy (Milky Way).

The line of intersection of the mean plane of the celestial equator and the ecliptic defines the direction of the equinox (Υ). Using the projection of this direction as the origin, the right ascension (α) is measured in the plane of the equator and celestial (or ecliptic) longitude (λ) is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

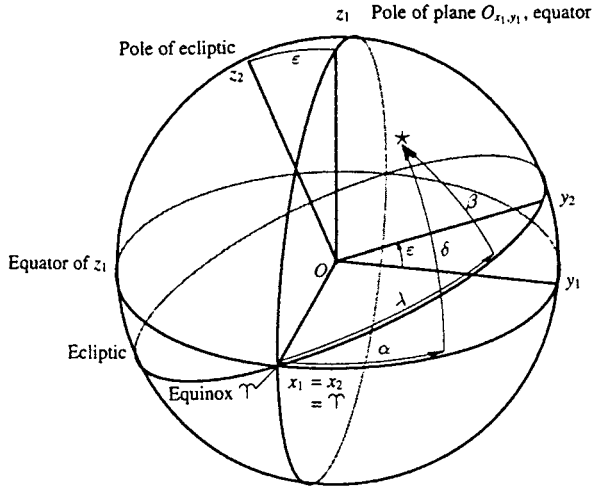


Figure 1. Equatorial and ecliptic reference planes.

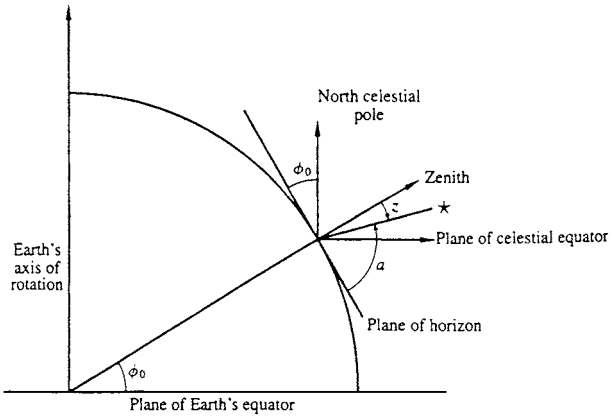


Figure 2. Topocentric coordinates.

measured in the plane of the ecliptic (see figure 1). Right ascension is usually expressed in time measure from 0 h to 24 h, and both right ascension and longitude are measured in the positive (or right-handed) sense. The complement of right ascension with respect to 24 h is known as sidereal hour angle (SHA); in navigational publications it is usually expressed in degrees. Declination is measured from the equatorial plane, positive to the north, from 0° to 90°. Celestial latitude is measured from the ecliptic plane, positive to the north, from 0° to 90°.

The equinox is at the ascending node of the ecliptic on the equator; this is the direction at which the Sun, in its annual apparent path around the Earth, crosses the equator from south to north. It is also referred to as 'the first point of Aries', and it is the vernal (spring) equinox for the northern hemisphere (see figure 1). The inclination of the plane of the ecliptic to the plane of the equator is known as the OBLIQUITY OF THE ECLIPTIC (ε).

The equator and ecliptic are moving because of the effects of perturbing forces on the rotation and motion of the Earth. Hence, the equinox and obliquity change as a function of time, so these celestial coordinate frames must be carefully defined in such a way that they can be related to a standard frame that may be regarded as being fixed in space.

The lines of intersection of the plane of the meridian with the planes of the horizon and equator define the directions from which AZIMUTH (A) and local HOUR ANGLE (h) are measured. Azimuth is measured in the plane of the horizon from the north, increasing in positive value toward the east. Local hour angle is measured in units of time: 1 h for each 15°, positive to the west with respect to the local meridian.

The latitudinal angles with respect to the horizon and equator are known as ALTITUDE (a) and DECLINATION (δ). Altitude is measured positively toward the zenith; in astronomy the zenith distance, angle from local vertical ($z = 90° - a$), is more generally used. Declination (δ) is, measured from the equator, positive toward the north pole of rotation. The ZENITH DISTANCE of the north pole, which is the same as the codeclination of the local vertical (or zenith), is equal to the geographic colatitude of the point of observation.

There are two main topocentric coordinate frames. One is based on the direction of the local vertical, which defines the plane of the horizon; the other is based on the direction of the Earth's axis of rotation, which defines the plane of the celestial equator. The local meridian, which contains the direction of the local vertical and the direction of the axis of rotation, is common to both frames.

The angular coordinates in the topocentric coordinate frame of the horizon and local meridian are known as azimuth and altitude. Azimuth (A) is measured from 0° to 360° from north in the direction of east; that is, in the negative sense with respect to the direction of the zenith. (In some circumstances azimuth may be measured from the south, and so the convention that is used should be stated to avoid ambiguity.) Altitude (a) is measured positively from the horizon toward the zenith. Both the altitude and the azimuth of any celestial object change rapidly as the Earth rotates and therefore for many purposes it is more convenient to use the topocentric frame of the celestial equator and the local meridian. The direction of the north celestial pole is parallel to the Earth's axis of rotation, and the altitude of the pole is equal to the geographic latitude of the point of observation. This is illustrated in figure 2, which shows the Earth as a sphere.

In this topocentric equatorial frame, the angular coordinates are known as local hour angle (LHA, or h) and declination (δ). LHA is measured from 0° to 360°, or from 0 h to 24 h, from south in the direction of west (i.e. in the negative sense with respect to the direction of the north celestial pole), and the LHA of a celestial object increases by about 15°, or 1 h, for every hour of time. Declination is measured from the equator to the celestial pole, positive toward the north. The declination of a celestial object is not changed by the diurnal rotation of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Seidelman P K (ed) 1991 *The Explanatory Supplement to the Astronomical Almanac* (Mill Valley, CA: University Science Books)

P K Seidelman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Equator

The great circle on the celestial sphere obtained by the intersection with the sphere of the plane of the Earth's equator. When the Sun lies in plane, day and night are everywhere of equal length. Known as equinox, this occurs twice each year, when the Sun is at one of the two points at which the celestial equator and the ecliptic intersect.

See also: celestial latitude, celestial longitude, celestial poles, celestial sphere, ecliptic, equinox, obliquity of the ecliptic.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Latitude

The angular distance between the ecliptic and a celestial body measured in a direction perpendicular to the ecliptic, and taking values between 0 and 90°. The position of a celestial body may be expressed in terms of celestial latitude (denoted by β) and celestial longitude.

See also: celestial equator, celestial longitude, celestial poles, celestial sphere, ecliptic, equinox.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Longitude

The angle between the great circle passing through the poles of the ecliptic and the vernal equinox, and the great circle passing through these poles and a celestial object, measured in an anticlockwise (i.e. eastward) direction from the vernal equinox. In other words, it is the angular distance measured along the ecliptic between the vernal equinox and the point on the ecliptic such that the angle between the vernal equinox, this point, and a star is a right angle. Planetary positions are sometimes expressed in terms of celestial longitude (denoted λ) and celestial latitude.

See also: celestial equator, celestial latitude, celestial poles, celestial sphere, ecliptic, equinox.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Mechanics

Celestial mechanics is the study of objects in the universe moving under the influence of a variety of forces. Nowadays the term is generally taken to mean the theory of motion of objects in the solar system, although it can still be applied to the dynamics of stars involved in multiple systems; the gravitational dynamics of stellar clusters and galaxies are usually treated as separate areas of investigation because they make use of different techniques. Celestial mechanics is primarily concerned with the dynamical consequences of the force of gravity between objects but additional forces can operate. These range from the tidal interactions of planets and satellites to the effects of electromagnetic radiation on small particles. Once considered to be a somewhat archaic branch of astronomy, there has been a renewed interest in celestial mechanics in recent years with the realization that its deterministic equations of motion can give rise to chaotic solutions. This has led to numerous applications in the solar system and an improved understanding of long-term dynamical evolution.

Orbital theory

In the 17th century KEPLER derived his three laws of planetary motion empirically based on observations (see KEPLER'S LAWS). Later in the same century NEWTON showed that Kepler's laws were the consequence of a simple inverse square law of force whereby the magnitude of the force of gravitational attraction between two masses, m_1 and m_2, separated by a distance d, is given by

$$F = \mathcal{G}\frac{m_1 m_2}{d^2}$$

where $\mathcal{G} = 6.672 \times 10^{-11}$ m^3 kg^{-1} s^{-2} is the universal gravitational constant. Newton showed that the orbit of one mass with respect to the other would describe a CONIC SECTION (i.e. circle, ellipse, parabola or hyperbola) where values of the position and velocity vectors ($\boldsymbol{r}$ and $\boldsymbol{v}$, respectively) at a particular instant determine the type of path followed. Because the angular momentum of the system is a constant, the relative orbit of the two masses defines a plane—the orbit plane—and so plane polar coordinates fully describe the relative motion. In the case of elliptical orbits the polar equation giving the radial distance, r, as a function of the true longitude, θ, is

$$r = \frac{a(1-e^2)}{1 + e\cos(\theta - \varpi)}$$

where a is the semimajor axis of the ellipse and e is its eccentricity (figure 1). The angle ϖ gives the orientation of the ellipse with respect to a fixed reference line at $\theta = 0$. For the gravitational interaction of two masses the quantities a, e and ϖ are all constant while θ increases non-linearly with time. Elliptical orbits are bounded with r oscillating between between finite limits; consequently they are periodic in nature.

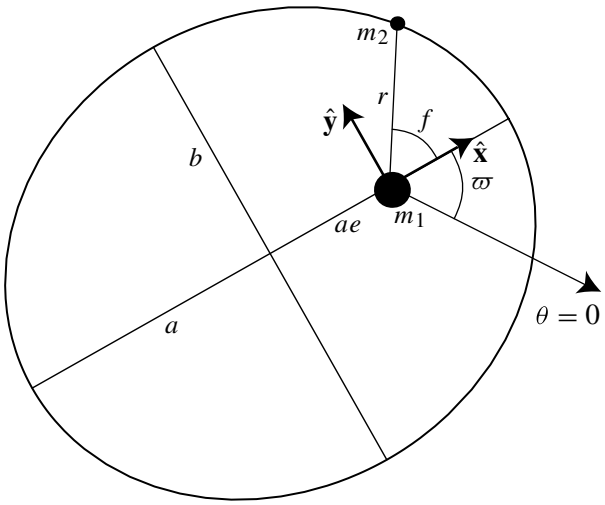


Figure 1. The geometrical properties of planar elliptical motion with semimajor axis a, eccentricity e, semiminor axis $b = a(1-e^2)^{1/2}$ and longitude of pericenter ϖ. The mass m_2 has radial distance r from m_1, true anomaly f, and true longitude $\theta = f + \varpi$. The unit vector in the $\hat{x}$ direction is taken along the major axis of the ellipse directed towards the point of closest approach (the pericenter); the $\hat{y}$ direction is perpendicular to this line in the plane of the orbit. The mass m_1 is at one focus of the ellipse.

The polar equation of the ellipse does not provide a means of calculating the position of an object at a given time. However, this can be done using Kepler's equation which relates the eccentric anomaly, E, to the mean anomaly, M. The components of the position vector with respect to the $\hat{x}$–$\hat{y}$ frame can be written as $r\cos f = a(\cos E - e)$ and $r\sin f = a(1-e^2)^{1/2}\sin E$. Kepler's equation can be written as

$$M = E - e\sin E$$

where E and M are measured in radians. The mean anomaly has the useful property that it increases with time at a fixed rate given by the mean motion, $n = [\mathcal{G}(m_1 + m_2)/a^3]^{1/2}$ with the time needed to complete one orbit (one orbital period) given by $T = 2\pi[a^3/[\mathcal{G}(m_1 + m_2)]]^{1/2}$. Knowing the time at which the mass m_2 is at a particular longitude (usually taken to be the time of pericenter passage, τ) it is possible to calculate the location of the mass m_2 at any subsequent time in the context of the two-body problem. However, because Kepler's equation is transcendental in E it is usually necessary to undertake a numerical solution.

In the case of motion in three dimensions it is necessary to define two additional angles: the orbital inclination, I, and the longitude of ascending node, Ω (figure 2). Furthermore, the argument of pericenter, ω, is used to denote the angle between the ascending node and the pericenter and the longitude of pericenter is now taken to be $\varpi = \omega + \Omega$ where ω and Ω are in different

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

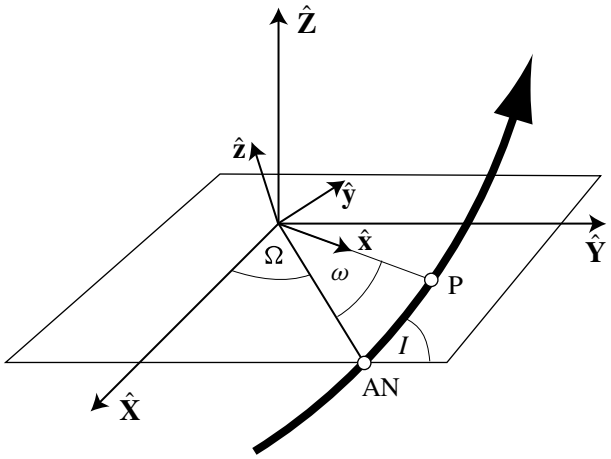


Figure 2. The geometrical properties of three-dimensional elliptical motion along an orbit (thick curve) inclined by an angle I with respect to the $\hat{\boldsymbol{X}}$–$\hat{\boldsymbol{Y}}$ reference plane. The points labeled AN and P denote the ascending node and pericenter, respectively. The ascending node has a longitude Ω with respect to the reference ($\hat{\boldsymbol{X}}$) direction; the pericenter makes an angle ω with respect to the line to the ascending node. The longitude of pericenter is now defined to be $\varpi = \omega + \Omega$.

planes. Note that even though this motion is viewed in a coordinate system inclined relative to the orbital plane, the fundamental properties of the Keplerian orbit remain unchanged.

The motion of two bodies under their mutual Newtonian gravitational attraction (the two-body problem) is one of the few integrable problems in celestial mechanics. It has the property that the orbiting mass returns to the same value of r after f has completed an azimuthal cycle of exactly 2π. The addition of only one more body destroys this symmetry. The problem becomes non-integrable in the sense that no practical mathematical solution is possible although some analytical results can still be derived (see THREE-BODY PROBLEM).

Lunar theory is a branch of celestial mechanics devoted to the study of the Moon's motion. This problem is relatively simple if it is confined to predicting the Moon's path around a spherical Earth. However, the reality is that solar perturbations are also significant and there are more subtle, long-term effects to consider such as modifications to the Earth's spherical shape caused by tides and rotation.

The solar system has a hierarchic structure of several subsystems, each consisting of a central mass with one or more orbiting smaller masses at each level. For example, the planets orbit the Sun but most also have their own small natural satellites. Therefore, when considering the motion of a satellite orbiting close to a planet the perturbations from the Sun can usually be ignored. This does not mean that the orbit is a simple, fixed ellipse; the planet's non-sphericity causes short- and long-term changes in some of the orbital elements. Planetary orbits themselves are not fixed because all planets perturb one another. Therefore, although at any given instant the planet is always following an elliptical path determined by its instantaneous position and velocity (this is referred to as its osculating orbit), the shape and orientation of that ellipse are continually changing owing to perturbations from the other planets. Such changes are small but detectable over time.

In general the smallest objects, such as PLANETARY RING particles and METEOROIDS, are subjected to the widest variety of additional perturbations in the form of non-gravitational forces. Their size means that the effect of incident electromagnetic radiation can affect their orbits. For example, the POYNTING–ROBERTSON EFFECT is caused by the non-isotropic re-emission of radiation and is important for particles whose size is comparable with the wavelength of the radiation. It results in damping of the eccentricity and inclination and a decrease in the semimajor axis regardless of whether the particle is in a heliocentric or planetocentric orbit. This effect is thought to be responsible for the large-scale transport of collisional dust from the asteroid belt through to the inner solar system (see INTERPLANETARY DUST).

In the years following Newton's pioneering work (see CELESTIAL MECHANICS IN THE EIGHTEENTH AND NINETEENTH CENTURIES) considerable effort was devoted to quantifying and then predicting the motions of the objects that make up the solar system. One of the outstanding successes of celestial mechanics was the independent explanations by John Couch ADAMS and LEVERRIER for the observed anomalies in the motion of the outer planets. This led directly to the discovery of the planet Neptune by GALLE in 1847. The improvement in the precision with which planetary orbits could be understood led to an even more momentous discovery in the 20th century. EINSTEIN showed that a small discrepancy of $0.43''$ yr^{-1} in the motion of Mercury's longitude of perihelion was due to the fact that Newton's theory of gravity was incorrect and that his own general theory of relativity could account for the observations (see GENERAL RELATIVITY AND GRAVITATION).

Celestial mechanics gained importance as a practical science with the advent of the space age in the 1960s. The increased number of artificial satellites in Earth orbit meant that the gravitational field of the Earth could be studied in great detail by monitoring satellite orbits and how they changed with time. The need to send spacecraft to other planets led to the use of celestial mechanics in the design of complicated orbital tours (see the GALILEO and CASSINI MISSIONS). The gravity assist technique, whereby a spacecraft changes its orbital energy by means of a close flyby of a planet or satellite, has become the mainstay of the robotic exploration of the outer solar system. Although ground-based observations can be used to estimate the masses of planets and satellites, the careful monitoring of spacecraft trajectories for small deviations has resulted in direct measurements of the gravitational field of many solar system objects.

Gravitational perturbations

Historically perturbation theory in celestial mechanics has been divided into what are called general and special

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

perturbations. General perturbation theory consists of studying the evolution of a perturbed orbit by considering the analytical solution of the differential equations describing the variation of the object's orbital elements or its coordinates. In contrast special perturbation theory consists of analyzing the orbit using a numerical or seminumerical solution of the equations. Although greater insight is gained from an analytical solution, this is frequently impossible in celestial mechanics for all but the simplest of systems. A numerical solution can offer insight into the problem and the availability of fast digital computers allows large numbers of starting conditions to be considered so that some understanding can be achieved.

Consider a central mass m_c and two orbiting masses, m and m', with position vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$ with respect to m_c. The motion of m can be described by the solution to the vector equation

$$\ddot{\boldsymbol{r}} = \boldsymbol{\nabla}(U + \mathcal{R})$$

where

$$U = \mathcal{G}\frac{m_c + m}{r}$$
$$\mathcal{R} = \frac{\mathcal{G}m'}{|\boldsymbol{r}' - \boldsymbol{r}|} - \mathcal{G}m'\frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{r'^3}.$$

Here U represents the potential that gives rise to the two-body Keplerian motion of m around m_c and $\mathcal{R}$ represents the perturbing potential or disturbing function experienced by m due to m'. There are similar expressions for the effect of m on m'. Note that traditionally in celestial mechanics the potential is taken to be a positive quantity. While it is relatively trivial to express $\mathcal{R}$ in terms of Cartesian coordinates, it is more difficult to express it in terms of the orbital elements of m and m'. Nowadays this can easily be accomplished using computer algebra to undertake the necessary series manipulation.

The expansion of $\mathcal{R}$ has the general form

$$\mathcal{R} = \frac{\mathcal{G}m'}{a'}\sum_j S(a, a', e, e', I, I')\cos\varphi$$

where j is an integer and φ is a combination of longitudes (i.e. angles referred to a fixed direction) that can be written as

$$\varphi = j_1\lambda' + j_2\lambda + j_3\varpi' + j_4\varpi + j_5\Omega' + j_6\Omega$$

where the j_i are all integers and $\lambda = M + \varpi = n(t - \tau) + \varpi$ is the mean longitude. Each cosine argument, φ, satisfies the d'Alembert relation whereby $\sum_{i=1}^{6} j_i = 0$. Another important property is that the absolute value of the coefficient of ϖ', ϖ, Ω' and Ω (i.e. $|j_3|$, $|j_4|$, $|j_5|$ and $|j_6|$) is equal to the lowest power of e', e, $s' = \sin\frac{1}{2}I'$ and $s = \sin\frac{1}{2}I$, respectively, that occurs in the accompanying term. All the arguments can be classified according to the absolute value of the sum of the coefficients of the mean longitudes, λ' and λ in the cosine argument. Therefore an expansion that includes Nth-order arguments will involve Nth-order powers of the eccentricities and inclinations.

The disturbing function is used in conjunction with LAGRANGE's equations to give equations for the variations of the orbital elements. To lowest order in e and I these can be written as

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{2}{na}\frac{\partial\langle\mathcal{R}\rangle}{\partial\lambda}$$
$$\frac{\mathrm{d}e}{\mathrm{d}t} = -\frac{1}{na^2e}\frac{\partial\langle\mathcal{R}\rangle}{\partial\varpi}$$
$$\frac{\mathrm{d}\varpi}{\mathrm{d}t} = +\frac{1}{na^2e}\frac{\partial\langle\mathcal{R}\rangle}{\partial e}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{1}{na^2\sin I}\frac{\partial\langle\mathcal{R}\rangle}{\partial\Omega}$$
$$\frac{\mathrm{d}\Omega}{\mathrm{d}t} = +\frac{1}{na^2\sin I}\frac{\partial\langle\mathcal{R}\rangle}{\partial I}$$

where $\langle\mathcal{R}\rangle$ contains the required terms in the disturbing function expansion for an external perturber.

To use Lagrange's equations in some applications in celestial mechanics involves the identification of those terms in the expansion of $\mathcal{R}$ that are likely to be important for the particular problem of interest with the assumption that all the non-important, short-period terms will have zero time-averaged effect. The key orbital elements in this process are the semimajor axes, a and a'.

In the unperturbed problem the mean longitudes, λ' and λ, increase linearly at rates n' and n, respectively, whereas all the other angles in a general argument φ are constant. In the perturbed system λ' and λ are still rapidly varying quantities, but the other angles are slowly varying. Therefore, any arguments which do not involve mean longitudes are slowly varying. These give rise to long-period or secular terms. Under certain assumptions it is possible to derive a secular theory for planetary orbits. The resulting coupled equations of motion can be solved analytically to give approximate orbital elements (except mean longitudes) as a function of time.

Consider a general argument, $\varphi = j_1\lambda' + j_2\lambda$ where a and a' are such that $j_1n' + j_2n \approx 0$ and φ varies with a period longer than either of the two orbital periods. This occurs when there is a simple numerical relationship between the two mean motions or orbital periods. These arguments give rise to resonant terms in the expansion (see RESONANCE). In terms of the semimajor axes the equivalent condition is $a \approx (|j_1|/|j_2|)^{2/3}a'$. The solar system contains more pairs of objects in resonance than would be expected to occur by chance. This is probably due to the effects of tidal torques (see TIDES) causing satellite orbits to expand at differing rates over the age of the solar system.

Resonant terms are localized in semimajor axes. While a particular combination of angles may be slowly varying at one semimajor axis of the perturbed body, the same combination could be varying rapidly at another. In contrast the secular terms can be considered as global. Any argument which is neither secular nor resonant is considered to give rise to a short-period term.

Natural satellites such as the Moon raise tides on their parent planets; the planet itself may also be rotationally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

distorted. In each case it is possible to derive an equivalent disturbing function to represent the potential experienced by the orbiting mass and apply Lagrange's equations. For example, it can be shown that an elliptical orbit of semimajor axis a about an oblate planet of radius R_p would precess at a rate $\dot{\varpi} \approx \frac{3}{2} n J_2 (R_p/a)^2$ where n is the mean motion and J_2 is the second zonal harmonic coefficient of the planet.

One of the major theoretical advances in celestial mechanics in recent years has been the derivation of algebraic mappings to allow perturbed orbits to be followed numerically for extended periods. These mappings (or maps) can be classified into three broad types:

- resonance maps, where the disturbing function at a particular resonant location is derived—the secular part is integrable and the resonant part can be replaced by a sequence of Dirac δ-functions applied as impulses;
- encounter maps, where the orbit is considered as simple Keplerian motion until a conjunction or close approach to the perturbing body takes place—the perturbation is applied as an impulse.
- N-body maps, where the Cartesian equations of motion are re-formulated in a special coordinate system that isolates the two-body Keplerian part from the interaction part—use is made of Dirac δ-functions to treat the interaction part as impulses.

It has now become commonplace to use N-body maps to carry out numerical investigations of long-term stability in the solar system.

Chaotic motion

The fact that motion in the solar system arises from a simple mathematical law led to the belief that discovering the laws that governed the universe was equivalent to being able to predict its past and future state. This opinion was exemplified by LAPLACE's view of the universe as an entirely deterministic system. However, at the end of the 19th century POINCARÉ's study of the three-body problem gave the first indication that deterministic systems could admit unpredictable solutions.

In the solar system chaos has been invoked to explain such diverse phenomena as the KIRKWOOD GAPS in the asteroid belt and the unusual rotation of the Saturnian satellite HYPERION. It is now recognized that chaos has played an important role in shaping the dynamical structure of the solar system.

One of the key problems of celestial mechanics is the stability of the solar system. Modern numerical simulations show that the planetary orbits, especially those of the inner planets, are technically chaotic in the sense that an error as small as 0.01 m in a planet's position today propagates exponentially and it becomes impossible to predict its location in $\sim 10^7$–10^8 yr. Nevertheless, the orbits remain close to their current configurations over timescales comparable with the age of the solar system. The one exception is Mercury where some studies have revealed genuine orbital instability ($e \rightarrow 1$) in $\sim 10^{10}$ yr. The exact mechanism that gives rise to such chaotic behavior is still unknown although secular interactions between planets are thought to be responsible. With the discovery of other planetary systems it is clear that our understanding of these processes can be tested in order to discover which distributions of planetary masses and orbits give rise to configurations that are stable over the lifetime of a star.

Bibliography

Brouwer D and Clemence G M 1961 *Methods of Celestial Mechanics* (New York: Academic)

Murray C D and Dermott S F 1999 *Solar System Dynamics* (Cambridge: Cambridge University Press)

Roy A E 1988 *Orbital Motion* (Bristol: Hilger)

Wisdom J 1987 Urey Prize Lecture: chaotic dynamics in the solar system *Icarus* **72** 241–75

Carl D Murray

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Mechanics in the Eighteenth and Nineteenth Centuries

CELESTIAL MECHANICS (so named by Laplace in 1798) before the twentieth century consisted in the application to the solar system of Isaac NEWTON's law of universal gravitation. By this law, a single body orbiting about a larger body moves in a conic section; the presence of a third body perturbs the motion. Celestial mechanics is concerned with the perturbations thus produced.

Newton's treatment of this problem was fundamentally qualitative. In an approximative formulation, he showed that all the known inequalities of the Moon followed qualitatively from the law of gravitation. In his *Theory of the Moon's Motion* (1702), and in the second edition of the *Principia* (Scholium to Proposition III.25), he quantified his theory by fitting it to observations, and adding several quantitative terms without explaining their derivation.

The THREE-BODY PROBLEM was re-formulated in the 1740s as a set of differential equations. This formulation could be rigorous, containing implicitly all the inequalities implied by universal gravitation. (A solution, it turned out, could be pursued only by successive approximations.) That such a formulation was not achieved earlier is attributable to the lack, among mathematicians within the Leibnizian tradition, of a worked-out calculus of trigonometric functions. It was LEONHARD EULER (1707–83) in 1739 who first gave systematic algorithmic rules for the differentiation and integration of trigonometric functions. By 1744 Euler had derived inequalities of the Moon by solving differential equations, but only his results were published.

In 1746 the Paris Academy of Sciences announced as its prize problem for the year 1748 the formulation of a 'theory of Jupiter and Saturn explicating the inequalities that these planets appear to cause in each other's motions'. The prize-winning essay was Euler's; published in 1749, it provided the first generally accessible account of an analytic resolution of the three-body problem. Meanwhile, from 1746, two members of the Paris Academy's prize commission, ALEXIS-CLAUDE CLAIRAUT (1713–65) and JEAN LE ROND D'ALEMBERT (1717–83), separately set about resolving the problem, and therewith deriving the inequalities of the Moon. Clairaut's lunar theory was published in 1752, d'Alembert's in 1754.

Euler and Clairaut on the equations of motion

Euler wrote equations in polar coordinates for Saturn's motion, with the plane of Jupiter's orbit as the r, ϕ plane

$$\mathrm{d}^2r - r\,\mathrm{d}\phi^2 = -\frac{P\,\mathrm{d}t^2}{2} \tag{E.1}$$

$$2\mathrm{d}r\,\mathrm{d}\phi + r\,\mathrm{d}^2\phi = -\frac{Q\,\mathrm{d}t^2}{2} \tag{E.2}$$

$$\mathrm{d}^2z = -\frac{R\,\mathrm{d}t^2}{2}. \tag{E.3}$$

Here P, Q and R are accelerative forces tending to diminish r, ϕ and z; the factor 1/2 arises from Euler's choice of the unit of acceleration.

Clairaut wrote equations similar to (E.1) and (E.2) for r and ϕ, assuming initially the perturbed body's motion to be confined to a plane

$$r\,\mathrm{d}\phi^2 - \mathrm{d}^2r = \left(\frac{M}{r^2} + P\right)\mathrm{d}t^2 \tag{C.1}$$

$$r\,\mathrm{d}^2\phi + 2\mathrm{d}r\,\mathrm{d}\phi = Q\,\mathrm{d}t^2. \tag{C.2}$$

Here M is the sum of the masses of the central and perturbed body, P is the added radial acceleration from perturbation and Q is the transverse acceleration.

To resolve (E.1) and (E.2), Euler replaced $\mathrm{d}t^2/2$ by $a'^3(\mathrm{d}M')^2$, where a' is Jupiter's mean solar distance and M' its mean motion, and substituted for P and Q the accelerative forces involved when Saturn's motion was referred to an immobile Sun. He then took up successively three different approximative assumptions, in which Jupiter's orbit and Saturn's unperturbed orbit were taken either as circles, or one as a circle and the other as an ellipse. Under each hypothesis, he transformed (E.1) and (E.2) into differential equations for the small departures of r and ϕ from their unperturbed values. (Here trigonometric series were necessary; see the next section.) The equations were resolved by the method of undetermined coefficients.

Clairaut obtained an expression for $\mathrm{d}t$ in terms of r and ϕ by integrating (C.2) twice, and with this eliminated t from (C.1). A double integration of the transformed (C.1) then gave him

$$\frac{f^2}{Mr} = 1 - g\sin\phi - c\cos\phi + \sin\phi\int\Omega\cos\phi\,\mathrm{d}\phi - \cos\phi\int\Omega\sin\phi\,\mathrm{d}\phi \tag{C.3}$$

where f, g, c are constants of integration and

$$\Omega = \frac{(Pr^2/M) + (Qr/M)(\mathrm{d}r/\mathrm{d}\phi) - 2\rho}{1 + 2\rho}$$

ρ being

$$\frac{1}{f^2}\int Qr^3\,\mathrm{d}\phi.$$

To evaluate the integrals in (C.3), it was necessary to replace r in Ω by an approximate function of ϕ; Clairaut used

$$\frac{k}{r} = 1 - e\cos m\phi \tag{C.4}$$

the equation of a precessing ellipse, in which k, e, m are undetermined constants.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The introduction of trigonometric series

The lunar and planetary versions of the three-body problem differ because v, the distance between the perturbing and perturbed bodies, is in the lunar case subject only to small variations, expressible by a few terms of a Taylor series, whereas in the planetary case it varies hugely. In the differential equations v occurs as v^{-3}, which fluctuates yet more wildly. What is needed is an expression for v^{-3} permitting formal integration.

Euler resolved this problem by inventing trigonometric series. To a first approximation, he showed

$$v^{-3} = (a^2 + a'^2)^{-3/2}(A + B\cos\theta + C\cos 2\theta + D\cos 3\theta + \cdots) \tag{E.4}$$

where a, a' are the mean solar distances of Saturn and Jupiter, and θ is the heliocentric angle between the two radius vectors; the coefficients C, D, E, etc, are functions of A and B, while A and B are computable by numerical integration. An improved approximation required taking orbital eccentricities into account, and computing coefficients in the series for v^{-5}, v^{-7}, etc. Euler carried the approximation to the order of the first power of the eccentricities.

In 1767 JOSEPH LOUIS LAGRANGE (1736–1813) introduced an efficient new method for determining the series coefficients on the basis of DeMoivre's theorem. Later theorists carried the approximations to higher powers of the eccentricities.

The method of variation of orbital elements

Euler in treating the perturbed planet's motion in the z-direction replaced (E.3) by two first-order equations, one for the orbital inclination and the other for the longitude of the ascending node. In unperturbed motion these quantities are constant; perturbation makes them variable. Euler's approximate expressions for their variations under perturbation showed the inclination subject only to minor oscillations, while the node retrogressed on the orbital plane of the perturbing planet. This was the first analytic determination of the variation of an ORBITAL ELEMENT.

Euler derived the variations of other orbital parameters in later memoirs, but it was Lagrange who chiefly developed the method. In an essay of 1766 Lagrange showed that the ordinary way of computing perturbations by successive approximations inevitably produces terms proportional to the time, as well as nth-order terms in the $(n+1)$th-order approximation. The method of varying orbital parameters avoids this difficulty. In his further work on planetary perturbations, Lagrange used this method exclusively.

The motion of the Moon's apse

Newton published no derivation of the Moon's apsidal motion. In the *Principia's* first edition he cited results of an attempted derivation; its method, we now know, was flawed. In later editions he deleted all reference to it.

The first analytical derivations by Euler, Clairaut and d'Alembert in the 1740s yielded only about half the observed apsidal motion. Clairaut in November 1747 proposed that the inverse-square law be augmented by an inverse-fourth-power term, whence the remaining apsidal motion could be obtained. Then in May 1749 he retracted the proposal. He had discovered that, if he proceeded to the second-order approximation with respect to the perturbing forces (that is, substituting the terms found in the first approximation, with undetermined coefficients, back into the differential equations, and solving again), he obtained nearly the entire observed apsidal motion. No doubt he had expected the second-order approximation to yield only minor modification of the first-order result, and was surprised that two small terms proportional to the transverse perturbing force could produce so large a difference. Euler, by his own very different route, confirmed Clairaut's result. 'Only since this discovery', he wrote to Clairaut, 'can one regard the inverse-square law of attraction as solidly established...'.

Secular variations, real and supposed, to 1785

Secular variations are usually understood as monotonic changes tending to alter an orbital element unidirectionally. EDMOND HALLEY in the 1690s had detected a secular acceleration in the Moon's mean motion since ancient times; later investigators confirmed it. Halley also assigned a secular acceleration to Jupiter and a secular deceleration to Saturn, to accommodate ancient and modern observations (*Tabulae astronomicae*, 1749). Such secular changes, continued indefinitely, would lead to dissolution of the Earth–Moon and planetary systems.

The first major memoir of PIERRE-SIMON LAPLACE (1749–1827), completed in 1773, demonstrated that, to the order of the squares of the eccentricities, the mean motions of the planets are immune to secular variation from planetary perturbation. To the order mentioned, the coefficient of the term proportional to t^2 in the expression of the longitude was identically zero. The apparent secular changes in Jupiter's and Saturn's rates of mean motion, Laplace conjectured, might be due to cometary perturbation.

In a memoir sent to the Paris Academy in October 1774, Lagrange derived differential equations for those perturbations in the orbital inclinations and nodes of the planets that depend on the orientations and sizes of the orbits, but not on where the planets are in them. To do this, he replaced the tangent of the orbital inclination (θ) and the longitude of the ascending node (ω) by new variables, $u = \theta\sin\omega$, $s = \theta\cos\omega$. The resulting differential equations were linear and of the first order, and had a solution of the form

$$\begin{aligned} s_1 &= A_1\sin(at+\alpha) + B_1\sin(bt+\beta) + C_1\sin(ct+\gamma) + \cdots \\ u_1 &= A_1\cos(at+\alpha) + B_1\cos(bt+\beta) + C_1\cos(ct+\gamma) + \cdots \\ s_2 &= A_2\sin(at+\alpha) + B_2\sin(bt+\beta) + C_2\sin(ct+\gamma) + \cdots \\ &\text{etc.} \end{aligned} \tag{G.1}$$

Here the constants $a, b, c, \ldots$ are the roots of an nth-degree equation, and the remaining constants are determinable

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

from the empirical values of s_1, u_1, s_2, etc, at a given epoch. Lagrange did not solve the entire system (G.1) for all six known planets—an enormous task—but by treating the two large ones, Jupiter and Saturn, as a system apart, obtained four simpler equations. Their solution showed that the nodes and orbital inclinations of these planets oscillated between fixed limits with a period of 51 150 years. Lagrange called these variations secular: they were periodic, to be sure, but their periods were orders of magnitude longer than the periods of ordinary periodic variations, which depend on the positions of the planets in their orbits.

Laplace read Lagrange's memoir of October 1774 on its arrival in Paris, and saw at once (as had Lagrange) that a similar procedure would apply to the eccentricities and aphelia. Thus, introducing the new variables $h = e \sin L$, $k = e \cos L$, where e is the eccentricity and L the longitude of the aphelion, he derived linear, first-order differential equations for $\mathrm{d}h/\mathrm{d}t$ and $\mathrm{d}k/\mathrm{d}t$. The solution was similar in form to (G.1). This analysis was published in 1775.

Laplace now abandoned the problem of planetary perturbations for a decade. The masses of Mercury, Venus and Mars were unknown, and he was uncertain as to the extent of cometary influence on planets.

Lagrange forged ahead. In 1776 he presented a new demonstration of the immunity of the mean motions to secular variation; it contained a classic formulation of the variation of orbital elements, and introduced 'the perturbing function', a potential function from which the perturbative forces in different directions could be derived by partial differentiation. In 1777 he used the perturbing function in the first systematic derivation of all the known integrals of motion for n bodies attracting each other according to the inverse-square law. In 1781 and 1782 he completed the two parts of a major treatise deriving the secular inequalities of all the planets by the method of variation of orbital parameters. In 1783 he applied the same method to the derivation of the zeroth-order periodic inequalities (inequalities independent of the eccentricities and inclinations).

At this point Lagrange realized that his earlier proof of the immunity of the mean motions to secular variation needed qualification: the mean motions could appear to vary secularly owing to the secular variation of other orbital elements. Let the part of the apparent variation in the mean motion that is correlated with a change in mean solar distance, and is thus immune to secular variation, be $\mathrm{d}p$; this part is periodic and dependent on the positions of the planets. The other part, $\mathrm{d}\Sigma$, depends on the variations in other orbital elements, and is 'secular' (independent of planetary positions) when these other variations are. The variation $\mathrm{d}\Sigma$, Lagrange proposed, should be regarded as a variation in the sixth orbital element, the epoch.

$\mathrm{d}\Sigma$ proved to contain secular terms only when the approximation was carried to the second and higher powers of the eccentricities and inclinations. To mitigate the labor of such calculations, Lagrange proposed taking account only of terms that would be greatly augmented by integration. These were terms containing a cosine or sine of an angle π such that $\mathrm{d}\pi = \lambda\,\mathrm{d}p$, where λ is very small; for in the double integration such a term is multiplied by the large factor λ^{-2}. But π is a linear combination of the mean motions for the several planets, and a particular examination of these linear combinations will show which ones lead to small λ. Lagrange found no secular variation in $\mathrm{d}\Sigma$ that would account for Halley's secular accelerations. Laplace would shortly apply Lagrange's calculational proposal not to $\mathrm{d}\Sigma$ but to $\mathrm{d}p$.

Laplace's resolution of major anomalies, 1785–7

On 23 November 1785 Laplace announced the resolution of the anomaly in the mean motions of Jupiter and Saturn. J H Lambert had discovered that since the mid-seventeenth century the rate of Jupiter's acceleration, and the rate of Saturn's deceleration, were decreasing: the inequality could be periodic. Starting from the conservation of *forces vives*, Laplace now found that the secular variations Halley had proposed for these planets were in the ratio that their gravitational interaction would imply. As was well known, two periods of Saturn were nearly equal to five periods of Jupiter; the mean motions n_J and n_S therefore gave a linear combination, $2n_\mathrm{J} - 5n_\mathrm{S}$, that was very small. It followed that the periodic inequality proportional to $\sin[(2n_\mathrm{J} - 5n_\mathrm{S})t]$ could be sizable. On the other hand, its coefficient (which would consist of several terms) would be of the third dimension in the powers and products of the eccentricities and inclinations, and hence could be negligible. Only a complete calculation could decide the question. What Laplace found was the largest perturbation in the solar system, with a coefficient for Saturn of $-49'$, and for Jupiter of $20'$; the period was some 900 years. His new, augmented, theories for the two planets were in agreement with both ancient and modern observations.

Laplace's memoir of November 1785 included an explanation for a resonance Lagrange had discovered in the motions of the first three Galilean satellites of Jupiter, and also an attempted proof that the orbital eccentricities and inclinations of the planets must remain small. The latter proof, as LeVerrier later pointed out, tacitly assumed that all the planetary masses were of the same order of magnitude; it was therefore inconclusive.

By December 1787 Laplace had an explanation for the apparent secular acceleration of the Moon. The average radial component of the Sun's perturbing force is negative, subtracting from the Earth's attraction. The Earth's orbital eccentricity, owing to a secular but cyclical variation, is at present diminishing, causing a slight decrease in the Sun's average perturbing force. Hence the Moon is falling into a lower orbit and increasing its angular speed. Laplace's value for the increment, about 10" per century, agreed with the accepted empirical value. (This analysis was inadequate; see below.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Laplace's *Mécanique Céleste*

The first four volumes of Laplace's *Mécanique Céleste*, which appeared between 1798 and 1805, contained not only an extended exposition of perturbation theory, but its detailed application to the planets and the Moon. For decades these volumes would serve as the textbook *par excellence* for mathematical astronomers. A strict accounting for all the detectable inequalities in the solar system appeared within reach. The masses of Mercury, Venus and Mars could be determined from perturbations. Evidently comets did not perturb the planets detectably.

Laplace achieved his successes by a mixture of methods, choosing one or another as it offered a swifter approach to a particular goal. His procedure for planetary inequalities proportional to the zeroth and first power of the eccentricities and masses was guaranteed to make a full sweep. To obtain the higher-order periodic inequalities, he proceeded by a kind of sharpshooting, which left uncertain whether all terms to a given order of smallness had been accounted for. He added the very long-term periodic inequalities to the mean motions, and used the thus modified mean motions in determining the short-term inequalities. The secular variations were determined by yet another procedure. This diversity of methods prompted the question: could not a single, demonstrably self-consistent procedure be adopted for the determining of all inequalities?

Meanwhile, workers in the Laplacian tradition had a clear enough task. The development of the perturbing function must be carried to higher orders of approximation, and the deduction of higher-order inequalities rendered more systematic. The Baron de Damoiseau and Giovanni Plana pursued these imperatives in their lunar theories. The planetary theories formulated by URBAIN J J LEVERRIER (1811–77) and, after him, by SIMON NEWCOMB (1835–1909), were descended from the same tradition, their distinction lying in their larger observational basis and greater precision.

The discovery of Neptune

The seventh planet, Uranus, was discovered by William Herschel in 1781. In 1821 Alexis Bouvard, in constructing a theory of Uranus, found he could not represent the observations with the expected accuracy of a few arcseconds. After 1821 the discrepancies grew, reaching nearly two arcminutes in the early 1840s. Was an unknown perturbing planet responsible?

In the mid-1840s JOHN COUCH ADAMS (1819–92) in England and LeVerrier in France independently undertook to respond to this question. They took the perturbing planet's orbit to lie in the ecliptic, Uranus's latitudes being minuscule. They accepted the Titius–Bode law, a numerical rule giving approximately the solar distances of the known planets, and implying that the next planet would have a mean solar distance of 38 astronomical units. Values for the orbital eccentricity and perihelion, epoch and mass of the unknown planet had then to be assigned so as to produce the perturbations that would bring Uranus's predicted motions into agreement with observation.

Both researchers succeeded. Adams, on presenting his results to GEORGE BIDDEL AIRY, the Astronomer Royal, in late 1845, met with rebuff: Airy distrusted the Titius–Bode law. LeVerrier sent his prediction to Johan Gottfried Galle of the Berlin Observatory, and on 23 September 1846 Galle turned his telescope to the indicated region of the sky, but failed to find the planetary disk he was seeking. His assistant, Heinrich Louis d'Arrest, suggested consulting the Berlin Academy's new *Star Atlas*. Comparing sky and chart, Galle located an eighth magnitude star not in the chart, 52′ from the position given by LeVerrier. On the 24th it had moved: it was the sought-for planet.

This discovery of Neptune by pure calculation was hailed as a triumph. The planet's orbit could now be calculated from observations. The mean solar distance proved to be 30.1 astronomical units: the Titius–Bode law was wrong. The orbital eccentricity was one-tenth as large as Adams and LeVerrier had supposed, and the perihelion over 100° distant from where they had put it. With the discovery of Neptune's satellite Triton in late 1846, Neptune's mass could be accurately determined; Adams's and LeVerrier's values were too large by factors of 2.65 and 1.88 respectively. Yet for the period from about 1800 to 1850, embracing the conjunction of 1821 between Uranus and Neptune when the perturbations were largest, Adams's and LeVerrier's hypothesized orbits were close to the actual orbit, gave very nearly the correct perturbations in this time-span, and permitted a correct pinpointing of Neptune's position.

The Lagrange–Poisson legacy: Hansen, Hamilton, Jacobi, Delaunay

In 1808 and 1809 Siméon-Denis Poisson (1781–1840) presented to the Paris Academy two memoirs utilizing and developing the Lagrangian method of variation of arbitrary constants. They stimulated Lagrange to further innovation. The Lagrange–Poisson theory was expounded in Lagrange's *Mécanique Analytique*, second edition (vol I, 1811; vol II, 1815).

Poisson introduced the concept of a 'conjugate momentum', p_i, for each coordinate q_i. If $q_i' = \mathrm{d}q_i/\mathrm{d}t$, and T is the kinetic energy, then $p_i = \partial T/\partial q_i'$. He showed that if a, b are two integrals of the system, the 'Poisson bracket'

$$(a, b) = \sum_{i=1}^{n} \left(\frac{\partial a}{\partial p_i}\frac{\partial b}{\partial q_i} - \frac{\partial a}{\partial q_i}\frac{\partial b}{\partial p_i} \right) \qquad \text{(LP.1)}$$

is independent of the time. He further showed that if a_i are the constants of integration that become variable under perturbation, then

$$\frac{\mathrm{d}a_i}{\mathrm{d}t} = \sum_{j=1}^{n} \frac{\partial \Omega}{\partial a_j}(a_j, a_i) \qquad \text{(LP.2)}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where Ω is the perturbing function. Poisson proposed and Lagrange proved that

$$\frac{\mathrm{d}a_i}{\mathrm{d}t} = -\frac{\partial\Omega}{\partial b_i}, \frac{\mathrm{d}b_i}{\mathrm{d}t} = \frac{\partial\Omega}{\partial a_i} \qquad \text{(LP.3)}$$

where the a_i can be the values of the q_i, and the b_i the values of the p_i, at $t = 0$. Equations of the form of (LP.3) would later be termed 'canonical'.

P A Hansen (1795–1874), director of the Seeberg Observatory from 1825 until his death, was a self-taught celestial mechanician. Because Laplace's formulae for perturbed coordinates held only to the first power of the perturbing forces, he soon turned to the method of variation of orbital parameters. This, however, had its own drawbacks: the perturbations of *six* elements had to be determined to find the perturbations of *three* coordinates, and the latter perturbations were often much smaller than the former, so that very small quantities had to be calculated from the differences of large quantities. Hansen sought to obtain a direct computation of the coordinates by rigorous derivation from the Lagrange–Poisson formulae.

The coordinates were functions of the orbital elements and the time; the orbital elements were functions of the time by (LP.2). Using the chain rule for partial derivatives, Hansen developed a two-part procedure. First he obtained the perturbed values of the longitude and radius vector in an instantaneous plane by substituting a 'perturbed time'—the time altered so as to incorporate the perturbative effects—into the standard elliptical formulae. Then he determined the position of the instantaneous plane relative to a plane of reference by means of the variations of two conjugated variables p and q.

Hansen first applied this procedure in a memoir on the mutual perturbations of Jupiter and Saturn; it won the prize of the Berlin Academy in 1831. (G W Hill in his more thorough computation of the Jovian and Saturnian perturbations in the 1880s would follow Hansen's procedure.) Hansen then turned to the lunar problem; his lunar tables, published in 1857, were adopted as the basis for the British and French national ephemerides from 1862 to 1922. The mathematician Arthur Cayley in the 1850s praised Hansen's procedure as the only rigorous treatment then available. Hansen's processes, however, were complicated; few astronomers understood them.

In two papers published in 1834 and 1835 in the *Philosophical Transactions* of the Royal Society of London, the Irish mathematician William Rowan Hamilton laid a new foundation for dynamics. He introduced the 'principal function' S, defined as

$$S = \int_0^t L\,\mathrm{d}t \qquad \text{(H.1)}$$

where $L = T - V$ is the 'Lagrangian', V being the potential energy. S, being evaluated along a solution curve of the differential equations of motion, must have an extremum value, so that $\delta S = 0$. The partial differential equation of the problem, it followed, was

$$\frac{\partial S}{\partial t} + H\left(q_1, \ldots, q_n, \frac{\partial S}{\partial q_1}, \ldots, \frac{\partial S}{\partial q_n}\right) = 0 \qquad \text{(H.2)}$$

where H is the 'Hamiltonian', or total energy, $T + V$. To the equations of motion Hamilton gave the canonical form

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = -\frac{\partial H}{\partial p_i}, \frac{\mathrm{d}p_i}{\mathrm{d}t} = \frac{\partial H}{\partial q_i}. \qquad \text{(H.3)}$$

The German mathematician C G Jacobi (1804–51) generalized Hamilton's theory, showing that it was applicable to non-conservative systems, and that a solution of the dynamical equations of motion was obtainable from a complete solution of (H.2). He also formulated a condition for transforming one set of canonical variables (obeying (H.3)) into another.

Independently of the work of Hamilton and Jacobi, CHARLES DELAUNAY (1816–72) in 1846 proposed determining the lunar perturbations from the initial equations

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \frac{\partial R}{\partial l}, \frac{\mathrm{d}G}{\mathrm{d}t} = \frac{\partial R}{\partial g}, \frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial R}{\partial h}$$

$$\frac{\mathrm{d}l}{\mathrm{d}t} = -\frac{\partial R}{\partial L}, \frac{\mathrm{d}g}{\mathrm{d}t} = -\frac{\partial R}{\partial G}, \frac{\mathrm{d}h}{\mathrm{d}t} = -\frac{\partial R}{\partial H}. \qquad \text{(D.1)}$$

Here (L, l), (G, g), (H, h) are six specially chosen orbital elements, conjugated in pairs in the manner of (LP.3), and R is the perturbing function. ((D.1) was a modification of a similar set proposed by M J Binet in 1841.) Delaunay developed R as a series of cosines of multiples of the relevant angles, to the seventh order of small quantities. His strategy in solving (D.1) was to separate R into two parts, R_1 and $R - R_1$, where R_1 is the largest term in R. He then solved (D.1) with R_1 substituted for R, obtaining new values of L, l, G, g, H, h, which he then substituted into $R - R_1$. Next, $R - R_1$ was to be separated into two parts, and the whole process repeated. At each stage, all the detectable effects due to the part of R under consideration were determined once for all—a major advantage over other methods, in which perturbational terms combined to produce new perturbational terms.

Delaunay's completed lunar theory appeared in two large volumes in 1860 and 1867. Like the earlier theories of Plana and P G D De Pontécoulant (1795–1874), it was a literal theory, giving the coefficients of perturbational terms in algebraic form rather than just numerically. Its chief disadvantage lay in the slow convergence of series.

The secular acceleration of the Moon revisited

In 1853 J C Adams showed that Laplace in his derivation of the Moon's secular acceleration had failed to take proper account of the tangential component of the Sun's perturbing force, and that when this was done, the overall effect deducible from the diminution in the Sun's perturbing force was only 6″ per century. Adams's result was confirmed by Delaunay in 1859. Astronomers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

defending the Laplacian value were unable to demonstrate error in the Adams–Delaunay analysis. In 1865 Delaunay explained how the ocean tides raised by the Moon could slow the Earth's diurnal rotation: the tidal protuberances, being carried eastward of the Moon by the Earth's diurnal rotation, are attracted backward by the Moon, and friction between the tidal water and the solid Earth then tends to slow the Earth's rotation. Result: the Moon appears to accelerate.

The issue was settled in the twentieth century on the basis of ancient eclipses. The retardation in the Earth's diurnal rotation causes an apparent acceleration in the Moon of 20″ per century; this added to the Adams–Delaunay value gives a total acceleration of 26″ per century. The observed value is about 11″; the difference, 15″, is attributable to the reaction of the tides on the Moon, which raises the Moon into a higher orbit where its angular motion is less. Thus the Earth–Moon system evolves.

The lunar theory of G W Hill

The foundations for the century's most innovative lunar theory were laid by the American celestial mechanician GEORGE WILLIAM HILL (1838–1914) in two papers, 'Researches in the Lunar Theory' (1878) and 'On the Part of the Motion of the Lunar Perigee which is a Function of the Mean Motions of the Sun and the Moon' (1877).

Accuracy in earlier lunar theories was chiefly limited by slow convergence in the series giving the coefficients of perturbational terms; the culprit was the parameter m, the ratio of the sidereal month to the sidereal year. Hill's idea was to isolate the role of m and account for it fully at the start, by adopting a new first approximation, not the standard elliptical solution of the two-body problem, but a special solution of a 'restricted' three-body problem, in which the Moon is attracted but does not attract. Thus he set the solar eccentricity and parallax, the lunar eccentricity and inclination, all equal to zero, and used a rotating coordinate frame with origin at the Earth's center, x–y plane in the ecliptic, and x-axis constantly directed to the Sun. The resulting differential equations admitted of the particular periodic solution

$$x = \sum_i a_i \cos(2i+1)\tau$$

$$y = \sum_i a_i \sin(2i+1)\tau \qquad \text{(Hi.1)}$$

where $\tau = \nu(t - t_0)$, t_0 being the time at which the Moon crosses the positive x-axis and ν the frequency of its periodic motion about the Earth; the summations are extended to all integral values of i. By successive approximations, Hill determined the values of the a_i in rapidly converging series in the parameter $m' = m/(1-m)$; the error in these values, when m' was given the value it has for the Earth's Moon, did not exceed two units in the fifteenth decimal. (Hi.1) gives what Hill calls 'the variational orbit', since it embodies the principal part of the lunar inequality known as the variation.

To obtain an eccentric orbit like that of the real Moon, Hill in the second paper mentioned above introduced free oscillations about the radius vector of the variational orbit, and at right angles thereto. The oscillation δp along the radius vector is given by

$$\frac{d^2\delta p}{d\tau^2} + \delta p(q_0 + 2q_1\cos 2\tau + 2q_2\cos 4\tau + \cdots) = 0 \qquad \text{(Hi.2)}$$

an instance of what is now called Hill's equation. To obtain a particular solution in which τ does not appear outside the arguments of the cosines, a constant c must be inserted into the first approximation:

$$\delta p = A\cos(ct\sqrt{q_0} + \varepsilon) \qquad \text{(Hi.3)}$$

where c is the solution of an infinite but converging determinant. Hill obtained a value for c good to nearly the fifteenth decimal; it gives the motion of the lunar perigee as $0.008\,572\,573\,004\,864 \times 360^\circ$. In contrast, Delaunay's series for this constant, though carried to the term in m^9, yields a value accurate only to three significant figures.

With the superiority of Hill's method evident, E W Brown commenced in 1888 the process of developing a complete lunar theory by means of it. The resulting tables were published in 1919, and became the basis of the British and American ephemerides in 1923.

The work of Poincaré

The culminating theoretical achievement of nineteenth-century celestial mechanics was *Les Méthodes Nouvelles de la Mécanique Céleste* by HENRI POINCARÉ (1854–1912), published in three volumes from 1892 to 1899. Using Hamilton–Jacobi theory, Poincaré explored topologically the range of solutions to the three-body problem, starting (as had Hill) from periodic solutions. The topological approach was new. Issues raised by this work, in particular the convergence of series used in celestial mechanics, and the strange behavior of Poincaré's doubly asymptotic solutions, have resurfaced in the second half of the twentieth century in Kolmogorov–Arnold–Moser theory and in chaos theory.

Bibliography

Barrow-Green J 1997 *Poincaré and the Three Body Problem* (Providence, RI: American Mathematical Society)

Gutzwiller M C 1998 Moon–Earth–Sun: the oldest three-body problem *Rev. Mod. Phys.* **70** 589–639

Taton R and Wilson C (eds) 1995 *Planetary Astronomy from the Renaissance to the Rise of Astrophysics, Part B: The Eighteenth and Nineteenth Centuries* (Cambridge: Cambridge University Press) chs 20–22, 28

Curtis Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Poles

The points on the celestial sphere at which the projected axis of the Earth intersects the sphere. The north celestial pole is thus vertically above the terrestrial north pole, and the south celestial pole vertically above the south terrestrial pole. The altitude of the north celestial pole is equal to the latitude of an observer on the Earth's surface, and since there is a star (Polaris, the Pole Star) close to the north celestial pole, it is possible for an observer in the northern hemisphere to obtain a rough estimate of his latitude simply by measuring the altitude of the Pole Star. Due to the rotation of the Earth, stars appear to move in circles centered on the celestial poles. Due to precession the direction of the Earth's axis changes and the position of the celestial poles changes as a result.

See also: celestial equator, celestial latitude, celestial longitude, celestial sphere, ecliptic, equinox.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celestial Sphere

An imaginary sphere of very large radius, centered on the Earth, and to which the stars are considered to be fixed for the purposes of position measurement. Due to the rotation of the Earth on its axis, the celestial sphere appears to rotate around the Earth once a day, and it is convenient to imagine that it is the Earth which is stationary, and the sphere which rotates. The sphere is therefore considered to rotate about an axis joining north and south celestial poles in an east to west direction (i.e. clockwise) at a rate of 15° per hour (of sidereal time). The plane of the Earth's equator, extended outwards to the sphere, marks the celestial equator. At any instant, an observer at a particular point on the Earth's surface can see only one half of the sphere. If he is located at one of the poles, one hemisphere is permanently below the horizon, while at the opposite extreme, if he is located at the equator, then each part of the sphere is visible at some time.

In ancient times, and up to the seventeenth century, it was widely accepted that the stars really were attached to such a sphere, which rotated round the Earth. Although we know that the stars lie at very different distances and that the celestial sphere has no real existence, it is still convenient to retain the concept of positional astronomy.

See also: celestial equator, celestial latitude, celestial longitude, celestial poles, ecliptic, equinox.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cellarius, Andreas ($fl.$ seventeenth century)

Biographical details are unknown for Cellarius, who produced finest celestial atlas of the seventeenth century, the *Atlas Coelestis*, 1660, couched in terms both of the Ptolemaic and Copernican theories of the universe, for example a Ptolemaic view of the Earth as if seen through the starry sphere from an external distant point in space.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Celsius, Anders (1701–44)

Mathematician and astronomer, born in Uppsala, Sweden. As professor of astronomy at Uppsala he participated in MAUPERTUIS' expedition in 1736 to the most northern part of Sweden to measure the length of a degree of latitude along a meridian, close to the pole, and compare the result with measurements near the equator. The expeditions confirmed NEWTON's opinion that the shape of the Earth is an ellipsoid flattened at the poles. Founded the modern observatory in Uppsala in 1741, and made geographical measures, meteorological and astronomical observations. Realized that auroras have magnetic causes through correlation of the dip-angle of a compass needle with auroral activity. Published catalogs of the brightnesses of 300 stars using a photometer that attenuated a star's light until its extinction by a number of glass plates. For his meteorological observations he constructed the eponymous *Celsius thermometer*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Centaurs

A group of objects in the outer solar system whose orbits cross those of the giant planets and are mostly confined between the orbits of Jupiter and Neptune. The first to be discovered was Chiron, in 1977, which was originally given the asteroidal designation (2060) Chiron. In 1989, seven years before perihelion, it was found to have developed a coma and was also given the cometary designation 95P/Chiron. Another object, (5145) Pholus, was found in 1992, and by April 2000 the total had reached 16. It may well be that the Centaurs are objects that have been perturbed inward from the Kuiper Belt, in a sense mirroring the relationship of the near-Earth asteroids to the main asteroid belt. Centaur orbits are unstable, and the fate of some of them may be further perturbation into the inner solar system as 'supercomets'. Their origin in the Kuiper Belt is supported by their reflectance spectra, which show them to be mostly dark red objects, and by their estimated sizes, which have a distribution consistent with that of Kuiper Belt objects. The largest may be (10199) Chariklo, with an estimated diameter of 300 km.

See also: Chiron, Kuiper Belt.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Centaurus

(the Centaur; abbrev. Cen, gen. Centauri; area 1060 sq. deg.) A southern constellation which lies between Vela and Lupus, and surrounds Crux on three sides. It culminates at midnight in early April. Its origin dates back at least to ancient Greece, where it was identified with Chiron in Greek mythology. The brightest stars of Centaurus were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large, conspicuous constellation, Centaurus is easily found. Its brightest star is α Centauri (Rigel Kentaurus, Toliman or Bungula), which at magnitude -0.3 is the third brightest star in the sky. It is actually a triple system consisting of a yellow (G2) primary, magnitude 0.0, an orange (K1) secondary, magnitude 1.4, separation 19.1″, period 81.2 years, and a third (M5) component, magnitude, 11.0, separation 2.2°, known as Proxima Centauri, which, at a distance of 4.2 light-years, is the closest star to the Sun. Other bright stars include β Centauri (Hadar or Agena), a close binary with components of magnitudes 0.6 and 3.9, separation 1.3″, θ Centauri (Menkent), magnitude 2.1, and γ Centauri (Muhlifain), another close binary with components both of magnitude 2.9 (combined magnitude 2.2), separation 1.2″, period 84.5 years. There are 18 other stars of magnitude 4.0 or brighter. Variable stars in Centaurus include the Mira-type stars R Centauri (range 5.3–11.8, period about 546 days) and T Centauri (range 5.5–9.0, period about 90 days).

Among several bright star clusters in Centaurus is ω Centauri (NGC 5139), a magnificent globular cluster, visible to the unaided eye as a fuzzy fourth-magnitude star. Other interesting objects include two open clusters, NGC 3766, which contains more than 60 stars between eighth and thirteenth magnitude, and NGC 5460, which has about 40 stars fainter than eighth magnitude, NGC 3918 (the Blue Planetary), an eighth magnitude planetary nebula described by John Herschel as 'very like Uranus, only half as large again', and NGC 5128 (Centaurus A), a seventh-magnitude elliptical galaxy and strong radio source, bisected by a dark dust lane. Centaurus also contains the x-ray binary Centaurus X-3.

See also: Alpha Centauri.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Center for Astrophysical Research in Antarctica

The Center for Astrophysical Research in Antarctica (CARA) is operated by the US National Science Foundation Science and Technology Center and has its headquarters at the University of Chicago. The astrophysical observatory is located at the South Pole and is used to investigate the origin of structure in the universe.

Instruments deployed there take advantage of the cold, dry and stable conditions which make the Antarctic Plateau the best site on Earth for observations at infrared and submillimeter wavelengths. They allow small telescopes at the South Pole to outperform much larger telescopes at temperate sites.

Observations are made at microwave, submillimeter and infrared wavelengths. Instruments used to study the cosmic microwave background radiation include the Degree Angular Scale Interferometer and the Viper Telescope.

The 1.7 m Antarctic Submillimeter Telescope and Remote Observatory is used to observe molecular clouds, and the Submillimeter Polarimeter for Antarctic Remote Observing maps interstellar magnetic fields.

The 60 cm South Pole Infrared Explorer telescope and the Abu infrared camera demonstrate the site's potential for deep, wide-field, thermal infrared (2–5 μm) imaging of sources in the southern hemisphere.

For further information see
http://astro.uchicago.edu/cara/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Center for High Angular Resolution Astronomy

Georgia State University's Center for High Angular Resolution Astronomy (CHARA) operates the CHARA Array on Mount Wilson, California. This optical/IR interferometric array consists of six 1 m aperture telescopes configured in a Y shape with a maximum baseline of 350 m. Its limiting resolution of 200 microarcsec makes the instrument a powerful tool for studying fundamental properties and surface features of stars.

For further information see http://chara.gsu.edu/CHARA/chara.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Centre National d'Etudes Spatiales

The Centre National d'Etudes Spatiales (CNES) draws up, proposes and conducts France's space policy. Its role is to develop the uses of space, to meet the civilian and military needs of public bodies and of the scientific community, and to foster the development and dissemination of new applications, designed to create wealth and jobs.

CNES conducts French space policy by participating in the programs of the EUROPEAN SPACE AGENCY (ESA). In 1998, France's contribution to ESA represented 58% of the investment subsidy allotted to CNES and nearly 33% of all contributions made to ESA, placing France first among the Agency's contributors. France also plays a crucial role of proposal and training in the evolution of the ESA's objectives and the implementation of its programs. A national program is conducted which supplements the ESA's programs.

CNES has many different partners: space users, for whom it identifies and develops the space techniques and applications best suited to their needs; French industry, to which it gradually transfers its know-how in order to make it more competitive worldwide; laboratories of major scientific organizations with which it works on numerous research programs; defense agencies, alongside which it works to develop a defense program, thus creating synergy between civil and military programs; space agencies of other countries with which it cooperates, chiefly in scientific areas; and developing countries, whose involvement in the use of space technology for teaching, economic growth and resource management is backed by CNES.

The areas in which space policy is implemented are those which involve the major strategic and economic challenges: access to space, with the Ariane program—an ESA program whose launch services are marketed by Arianespace—and the creation of a launch base in French Guiana (Centre Spatial Guyanais, CGS); commercial applications such as Earth observation (Spot, ERS, Meteosat, IASI), and telecommunications (Télécom 2, Stentor, GNSS, Skybridge); scientific programs in conjunction with research institutions, based on European and international cooperation (Topex-Poseidon, Envisat, Jason, Polder, Scarab, Oersted, Cluster, Cassini-Huygens, Soho, Rosetta, Mars Express, Mars Sample Return); activities related to microgravity research and man in space (Perseus with Alice 2, Fertile and Castor, Pégase) and the preparation of experiments designed for the International Space Station; and defense related activities (Helios).

To conduct its operations, CNES has various centers, each featuring specific expertise and assignments: the Paris headquarters; the Launch Vehicle Directorate at Evry, responsible for the Ariane program; the Toulouse center, Centre Spatial de Toulouse, (CST), in charge of preparing and developing space projects for satellites and planetary vehicles, and for running operation facilities and large-scale infrastructures; the Guiana Space Center (CSG) in Kourou, French Guiana, a launch and test center; and balloon launch center (CLBA) at Aire-sur-l'Adour, in the Landes department of southwest France. CNES has 2500 employees spread over these sites.

The CNES budget stands at 13 169 million French Francs, including a state subsidy of 9115 million French Francs and its own resources of 4104 million French Francs. For nearly 20 years, CNES has formed commercial subsidiaries to sell the products and services turned out by space technology. These 20 companies directly employ a workforce of nearly 1000 people.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cepheid Period–Luminosity Relation

The empirical correlation between the luminosity of a Cepheid variable and its period has been known to astronomers, and used by them as a powerful distance determination method, for the better part of the 20th century. Individually, Cepheids have distinctive light curves, rising rapidly in brightness to a maximum, and then falling more gradually to minimum light. This cycle consistently repeats with individual periods ranging from a few days to many hundreds of days, depending on the intrinsic brightness of the star. Collectively, the timescales over which this periodic cycling occurs for Cepheids are tightly coupled to the average luminosities of the stars. The reason for this empirical correlation is extremely well understood, and the applications of the period–luminosity relation are numerous. The implications, specifically for the extragalactic distance scale, are of importance in cosmology.

The period–luminosity relation for Cepheids was first discovered by the Harvard astronomer, Henrietta LEAVITT. She was analyzing the periods and apparent magnitudes of a selection of Cepheid variables all at a common distance in two of the (now known to be) nearby satellite galaxies to the Milky Way, the LARGE and SMALL MAGELLANIC CLOUDS. What Ms Leavitt found was to set in motion one of the most important and long-lasting astronomical activities of the 20th century—the calibration of the size scale of the universe. This was closely followed by the discovery of the expansion of the universe by Carnegie astronomer Edwin HUBBLE (see EXPANDING UNIVERSE), which, when combined, gave rise to a determination of the timescale to the creation of the universe, described by the 'hot big bang' model of modern COSMOLOGY.

Cepheids are supergiant stars, and as such they are among the most luminous objects in those galaxies that are today producing new populations of stars. The immediate precursors to Cepheids, according to stellar evolution theory, are the massive, young O- and B-type stars. These hot (blue) main sequence stars have been recently formed but, because of their high mass and prodigious energy output, they soon evolve away from the hydrogen-core-burning region to cooler surface temperatures and redder colors. In doing so, these enormous stars briefly pass through a zone in which their outer atmospheres are unstable to periodic radial oscillations. This narrow range of temperature triggering the atmospheric instability defines the so-called 'instability strip', where Cepheid pulsation can be observed. High-mass stars pass through the strip at higher luminosity (and slightly cooler temperatures) than lower-mass stars, which pass through the strip somewhat sooner at higher temperatures and at demonstrably lower luminosities. In the color–magnitude diagram the Cepheid instability strip slants upwards and across to the red.

Stated another way, these simple considerations from STELLAR EVOLUTION predict that a mass–luminosity relation exists for Cepheids. Indeed, it is also ensures that there is a radius–luminosity relation for these stars, since the most basic laws of physics demand that the only way for two stars of the same temperature (color) to have different luminosities is for the total surface areas (controlled directly by the radii) to be different. Since temperature implicitly controls surface brightness (through Stefan's law of emissivity, σT^4), and radius explicitly controls the surface area (through geometry, $4\pi R^2$), the combination of these two factors gives the total luminosity.

The interrelation between luminosity L, effective temperature $T_{\rm eff}$ and radius R is easily quantified by the following equation based on the physical arguments given above:

$$L = 4\pi R^2 \sigma T_{\rm eff}^4.$$

Given that the temperature at which these stars are unstable to oscillation covers a relatively small range ($\sim$300 K), luminosity then is primarily a function of the mean radius of the star in the instability strip. Unfortunately, neither the mass nor the radius of any star is an easily observed property, so the relation as written above cannot be used in any practical application for distance determinations (i.e. predicting intrinsic luminosities). If, however, some other (more directly observable) quantity were a function of the mass and/or the radius, one in principle could predict the luminosity using that quantity as a proxy for radius. With empirical hindsight we already know that there is a solution—the one found by Ms Leavitt—the period–luminosity relation. Theory easily explains this empirical relation.

Periods of oscillation at the surface of the Cepheid are intrinsic to the star and are obviously independent of the distance of the observer, but knowledge of an empirical period–luminosity relation immediately suggests that the observed period can be used to predict luminosity. From the physical arguments above, it must then be true that the period of oscillation can stand in for either mass or radius or some combination thereof. Indeed, this is the case.

However, nature provides us with even more; there is also an intrinsic period–color relation for Cepheids. Since both period (P) and color (C) are independent of distance, the P–C relation (once calibrated) is a means of detecting interstellar reddening in a sample of Cepheids in an external galaxy. Individually Cepheids can become redder and fainter owing to intrinsic temperature differences or redder and fainter because of line-of-sight extinction. For a collection of Cepheids, the intrinsic color differences will tend to average out. However, since extinction is a systematic effect, it will tend to move the entire population to redder and fainter levels. Compared with the intrinsic P–C relation the average deviation of the ensemble can be fairly interpreted as being due to interstellar extinction, and the inferred distance corrected for this absorption.

Explicitly correcting for systematic effects such as the dimming by interstellar dust (see INTERSTELLAR MATTER) is best accomplished by observing Cepheids at two or more wavelengths and using the period–color relations to predict statistical corrections for ensemble averages. An

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

alternative approach is to observe the Cepheids at long wavelengths where the effects of interstellar extinction are progressively diminished. For instance, by observing a Cepheid at 2 μm in the near-infrared, as opposed to the 4500 Å (blue) region of the spectrum, the effect of extinction drops by nearly a factor of 10.

For all freely oscillating systems, be they bells ringing, pendulum clocks swinging, planets orbiting, Cepheids pulsating, galaxies rotating, or even the entire universe expanding (or contracting), the primary physical determinant of the timescale (or oscillation period, P) is the mean density of the system (ρ). Density is, by its very definition, solely a function of mass and radius. The mathematical expression of this law of nature is captured in the equation

$$P\rho^{1/2} = Q.$$

At its most basic level, this equation is formally equivalent to the conservation of energy. Kepler's third law of planetary motion, linking the periods of planetary revolution about the Sun with their measured semimajor orbital axes, is another expression of this same formula (see KEPLER'S LAWS). Also, the relation between the expansion rate of the universe as characterized by the HUBBLE CONSTANT and the mean density of the universe compared with its critical density is the same formula in a cosmological setting.

The story behind setting a reliable zero point to the Cepheid P–L relation reads like a microcosm of the parallel quest to set a size scale to the entire universe. Indeed, these two tasks, disparate in scale, are often one and the same problem, recurrently interwined. In many contexts the Cepheid distance scale is the extragalactic distance scale. For instance, the HUBBLE SPACE TELESCOPE (HST) was, in large measure, designed around the expected use of Cepheids to reach galaxies and measure their distances as far out as the Virgo cluster some 10 or 20 times more distant than the most distant galaxies measured by Cepheids from the ground.

Measuring a Cepheid distance to Virgo was once hoped to be the Holy Grail of the distance scale, but this did not turn out to be the case. In the intervening years between the design of HST and its eventual launch and subsequent refurbishment, our view of the universe changed in a fundamental way: large-scale flows are now known to accompany the large-scale structures that fill the nearby universe. In other words, the velocities measured as redshifts for galaxies are not simply composed of a single Hubble expansion component, but they have superimposed flow velocities due to the gravitational perturbations of nearby matter in the form of nearby galaxies, groups and clusters.

The Key Project on the Extragalactic Distance Scale was premised on using Cepheids as distance indicators. Using traditional methods to discover extragalactic Cepheids by imaging in the optical (where the Cepheid amplitudes are large compared with redder wavelengths), the Hubble Space Telescope has been used by several groups including the Key Project to determine distances

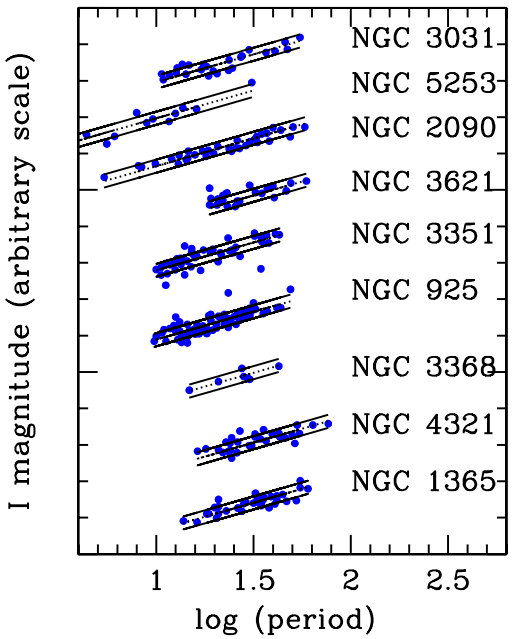


Figure 1. I-band period–luminosity relations for a sample of nearby galaxies (period in days).

to more than two dozen galaxies out to beyond the Virgo and Fornax clusters, nearly an order of magnitude further away than was previously possible using ground-based telescopes. The Cepheid distances to galaxies are then used to calibrate a number of secondary distance indicators for galaxies extending much further into the Hubble flow. With red (I-band) observations (see figure 1) the second color needed for extinction corrections was also obtained and true distances were derived. Follow-up studies of selected objects were also made at 1.6 μm with HST, using a near-infrared camera (NICMOS), to confirm the reddenings and strengthen the distance determinations based on a wavelength that is insensitive to both extinction and metallicity. The chemical composition, or metallicity, poses another plausible complication that needs to be addressed in evaluating the uncertainty associated with Cepheid distances. No significant differences were detected between the optically corrected true distance moduli and those bolstered by the near infrared data.

At the conclusion of the Key Project many secondary distance indicators were calibrated using Cepheid variables. Each of these methods yielded returned values of the expansion rate of the universe, the Hubble constant, falling between 68 and 78 km s^{-1} Mpc^{-1}. However, the distances (and velocities) to galaxies in and around the major clusters were also determined. Corrected for the perturbing effects of the clusters, these field galaxies also show a velocity–distance relation that can be identified with the expansion of the universe. From Cepheid distances alone one then can derive the strength of the relation between velocity and distance, and so locally it is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

found to be 73 km s^{-1} Mpc^{-1}. This value of H_0, whether it is derived exclusively from Cepheids in galaxies nearby or from galaxies tied into the Cepheid distance scale but participating in the flow at 10 000 km s^{-1}, consistently suggest an age of 12 billion years for a low-density universe.

Barry F Madore and Wendy L Freedman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cepheus

(abbrev. Cep, gen. Cephei; area 588 sq. deg.) A northern constellation which lies between Draco and Cassiopeia, and culminates at midnight in late August. It is named after King Cepheus, husband of Queen Cassiopeia and father of Andromeda in Greek mythology. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A rather inconspicuous constellation, the brightest star in Cepheus is α Cephei (Alderamin), magnitude 2.5. There are six other stars brighter than fourth magnitude, including three notable pulsating variable stars: β Cephei (Alfirk; range 3.16–3.27, period 0.19 day), a bluish-white (B2) star, which has a fainter, magnitude 7.8, companion, also bluish-white (A3), separation 13.3″; δ Cephei (the prototype Cepheid variable; range 3.48–4.37, period 5.37 days), a yellow (F5-G2) star, which has a bluish-white (B7) companion, magnitude 6.3, separation 41″; and μ Cephei (the Garnet Star—so called because of its intense red color; range 3.43–5.1, average period about 730 days).

Other interesting double stars in Cepheus include ξ Cephei (Kurhah), which has bluish-white (A3) and pale yellow (F7) components, magnitudes 4.3 and 6.4, separation 7.9″, and Krüger 60 , one of the nearest visual binaries (distance about 13 light-years), which has red (M3 and M4, dwarf) components, magnitudes 9.8 and 11.4, separation 3″, period 44 years, the fainter component of which (Krüger 60B or DO Cephei) is a flare star (max. 10.3). A bright Mira-type variable star is T Cephei (range 7.4–12.9, period about 487 days).

Open clusters in Cepheus include NGC 7160, which contains a couple of dozen stars fainter than seventh magnitude, and NGC 188, which has more than 150 stars between tenth and eighteenth magnitude and is one of the oldest galactic clusters known (estimated age around 9000 million years). There are no bright nebulae or galaxies in the constellation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ceres

The first asteroid to be discovered, by Giuseppe Piazzi on 1 January 1801, and so designated (1) Ceres. The discovery resulted from a search prompted by Bode's law, a numerical relationship between the known major planets which suggested a 'missing planet' between the orbits of Mars and Jupiter. Ceres is by far the largest of the asteroids, with a diameter of 933 km, and accounts for over one-quarter of the mass of the entire main-belt asteroid population. It orbits the Sun at a mean distance of 2.77 AU (414 million km) in a period of 4.60 years; the orbital inclination is 11º, and the eccentricity 0.08. Ceres rotates in 9.08 h, and has a density of 2.7 g cm^{-3}. It is a G-type asteroid, with a reflectance spectrum similar to that of carbonaceous chondrites, and an albedo of 0.11.
See also: BODE'S LAW.

Web update

Recent observations at the European Southern Observatory with the world's first operational virtual telescope 'Astrovirtel' have determined that the newly discovered distant asteroid 2001 KX76 is significantly larger than Ceres. Ceres, the first asteroid (minor planet) to be discovered in the Solar System, has held the record as the largest known object of its kind for two centuries with a diameter about 950 km. Asteroid 2001 KX76 has been found to have a diameter of 1200 km, possibly even 1400 km.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CERN

CERN—the acronym comes from the original French name 'Conseil Européen pour la Recherche Nucléaire—the European Laboratory for Particle Physics, is one of the world's most prestigious research centers. Its business is fundamental physics, i.e. finding out what makes our universe work, where it came from, and where it is going. At CERN, some of the world's biggest and most complex machines are used to study nature's tiniest building blocks. By colliding these minute particles of matter, physicists unravel the basic laws of nature. Amongst other areas of applicability, CERN's research is used in astroparticle physics, with relevance to studies of dark matter, the Big Bang, etc. This research is purely scientific and the results are freely available.

CERN's machines are particle accelerators and detectors. The laboratory's accelerator complex is built around three principal inter-dependent accelerators. The oldest, the Proton Synchrotron (PS), was built in the 1950s and was briefly the world's highest energy accelerator. The Super Proton Synchrotron (SPS), built in the 1970s, was the scene of CERN's first Nobel prize in the 1980s.

Ground-breaking research was performed at the Large Electron Positron collider (LEP) throughout the 1990s. The extraordinary quantity of high-precision results from LEP highlighted the possibility of exciting future discoveries, such as finding the Higgs particle, postulated to confer mass on the particles of matter, and supersymmetry, a profound extension to the relationship between particles of matter and particles carrying the fundamental forces between them.

Fundamental research is CERN's reason for being, but the laboratory also plays a vital role in developing the technologies of tomorrow. From materials science to computing, particle physics demands the ultimate in performance, making CERN an important test-bed for industry. The quality of the technology developed by the laboratory is becoming more and more widely recognized. Today everyone knows the World Wide Web, but not many know that it was invented at CERN, conceived to give particle physicists easy access to their data wherever they happened to be on the planet. The World Wide Web, medical imaging, or advanced techniques for using electronic chips, are just a few of the many recent spin-offs from the fundamental research carried out at CERN.

The next stage in Europe's strategy for particle physics is the Large Hadron Collider (LHC). CERN Council's decision on 16 December 1994 to build the LHC, which has the potential to make some of the most important scientific discoveries of our era, was a milestone in the history of particle physics and of CERN. The LHC is a technologically challenging, superconducting accelerator which will bring protons into head-on collision at higher energies than ever achieved before. It will be installed in the 27 kilometer tunnel that at present houses the LEP collider. Preparations for the LHC are moving ahead at full speed, with several thousand physicists worldwide actively preparing for its arrival, expected in 2006. The tremendous technological challenges facing those who are building the LHC and its detectors are an important driving force for innovation in European industry.

The laboratory sits astride the Franco-Swiss border, west of Geneva, at the foot of the Jura mountains, and was founded in 1954 as one of Europe's first joint ventures. Since then it has become a shining example of international collaboration. From the original twelve signatories of the CERN convention, membership has grown to twenty member states: Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Italy, the Netherlands, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland and the United Kingdom. In 2001 these countries contributed to an annual budget of 1050 million Swiss francs in amounts proportional to their net national income.

For further information see http://www.cern.ch/Public/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cerro Tololo Inter-American Observatory (CTIO)

CTIO is operated by the ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY Inc. (AURA), under a cooperative agreement with the National Science Foundation as part of the National Optical Astronomy Observatories.

The Observatory is about 2200 m above sea level, and 500 km north of Santiago in Chile. Its headquarters are located at La Serena.

On site are six optical telescopes and one radio telescope. The largest instrument is the 4 m Victor M Blanco Reflecting Telescope, whose twin is located at KITT PEAK OBSERVATORY in Arizona. This telescope has been in operation since the early 1970s.

Other instruments at the site include a 1.5 m Ritchey–Chrétien telescope, the YALO 1 m Ritchey–Chrétien telescope, a 92 cm reflector, a 0.6/0.9 m Curtis/Schmidt telescope, the 60 cm Lowell Telescope and 'El Enano' (The Dwarf), the smallest telescope on Tololo.

The 1.2 m radio telescope of the Universidad do Chile is also at the site.

For further information see http://www.ctio.noao.edu/ctio.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cetus

(the Sea Monster or Whale; abbrev. Cet, gen. Ceti; area 1231 sq. deg.) An equatorial constellation which lies between Aquarius and Taurus, and culminates at midnight in October. It is named after the sea monster from which Perseus rescued Andromeda in Greek mythology, though it is sometimes identified as a whale. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

Although the fourth largest constellation, Cetus is rather inconspicuous, its brightest stars being β (Diphda), magnitude 2.0, and α (Menkar), magnitude 2.5. There are seven other stars of magnitude 4.0 or brighter. Interesting stars in Cetus include the prototype long-period variable o Ceti (Mira; range 2.0–10.1, period about 332 days), the flare star UV Ceti (rest magnitude 12.9, max. 6.8) and τ Ceti, magnitude 3.5, one of the two closest Sun-like stars (distance about 11.8 light-years—the other being ε Eridani).

Another interesting object in Cetus is M77 (NGC 1068), a ninth-magnitude spiral galaxy (the brightest Seyfert galaxy).

See also: Mira.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CH Cygni

CH Cygni (CH Cyg) is one of the most enigmatic SYMBIOTIC STARS. Once known as a short-period semiregular variable and standard for the M6 III spectral type, the system began an unusual series of eruptions in 1964. The appearance of a blue continuum and modest H I emission lines led to its reclassification as a symbiotic star despite its lack of the strong higher ionization lines characteristic of most symbiotics. Some of these lines were later discovered with higher-quality spectra acquired in the 1970s and 1980s. Even so, the erratic behavior of the system during quiescence and eruption clearly distinguishes CH Cyg from most symbiotic stars.

CH Cyg has properties in common with *o* Ceti (MIRA) and the symbiotic star R Aqr, two M6–M7 type giants with a white dwarf companion. All are relatively nearby, with distances of 100–300 pc compared with the 1 kpc distance of the typical symbiotic binary. The systems contain pulsating RED GIANTS with periods of ∼100 days in CH Cyg and ∼1 year in Mira and R Aqr; all have orbital periods of decades to centuries. The combination of a late-type giant and a long orbital period in a nearby binary produces interesting behavior not observed in other symbiotic stars, such as Z ANDROMEDAE. Studying this evolution has been very popular; understanding it has proved elusive.

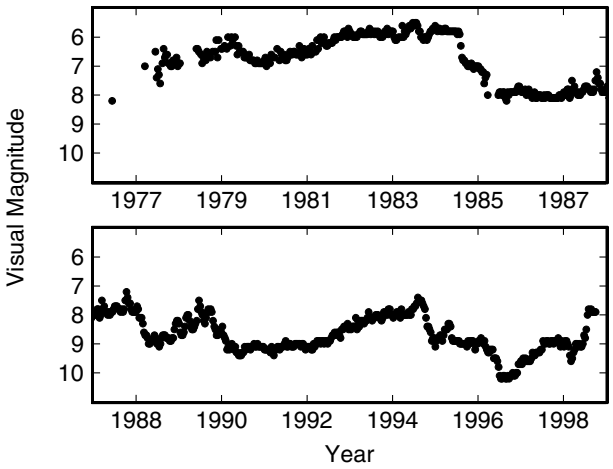


Figure 1. Recent optical light curve of CH Cyg using 10 day means of observations from the AFOEV and VSNET databases. Jet ejection began during the step-like drop in brightness bracketing 1985 in the right corner of the upper panel.

Basic properties

CH Cyg displays a bewildering array of photometric variations. Short-timescale fluctuations are observed at optical and x-ray wavelengths. The optical variations are coherent, with a period of 2500–3000 s and an amplitude that increases from ∼0.02 mag at 7000 Å up to ∼0.3 mag at 3600 Å. X-ray observations indicate a non-variable, soft component with $kT \sim 0.4$ keV and a variable component with $kT \sim 7$ keV. The x-ray variability occurs on timescales as short as 300 s but is not coherent.

Longer timescale variations are also present (figure 1). Semicoherent variations with periods of ∼100 days, ∼770 days, ∼1300 days and ∼32 years appear in the optical and near-infrared light curves. The typical amplitude of these changes is ∼0.5 mag in the optical and ∼0.25 mag in the near-infrared. The amplitudes and phases of these variations change from one decade to the next. Several studies have proposed an additional period of 5700 days produced by eclipses of the emission line source by the red giant star. The optical colors of the system are usually redder during the eclipse.

There is good evidence for RADIAL VELOCITY variations phased with several of the photometric periods. Velocities derived from red giant absorption features indicate periods of ∼760 days and ∼5500 days. The shorter period is similar to the orbital periods of other symbiotic stars. The derived mass function is also comparable with that of other symbiotics if the velocity variations are due to orbital motion. Unlike most symbiotics, however, this period is not always present in the light curve. Irregular light and velocity variations are often observed in single late-type giants, so this period might be a pulsation period or (much less probably) a rotation period. The longer period has more support as the true orbital period of the system. Orbital conjunctions coincide with minima in the emission line intensities. The jet ejection events described below also seem phased with this period. The velocity amplitude leads to reasonable masses for the red giant, 1–3 $M_\odot$, and for the hot component, 0.5–1.0 $M_\odot$.

The spectroscopic variations of CH Cyg are fortunately more tame than the photometric variations. The spectrum always shows evidence for an M-type giant, with a strong red continuum and deep TiO absorption bands. The spectral type of this component varies between M5 and M7. The hot component has three states: (i) an inactive state where the blue continuum and emission lines are absent or barely detectable; (ii) a low state where the blue continuum and H I emission lines are relatively weak; (iii) a high state where the blue continuum is strong and resembles the spectrum of an F-type supergiant (figure 2). Modest H I and Fe II emission lines accompany the F-type continuum in the high state. Some high ionization lines, such as [O III] in the optical and C IV in the ultraviolet, are present during the high and low states. These lines are usually stronger during the low state than the high state, as in other symbiotic stars. However, the emission lines always appear to be weaker relative to the UV or optical continuum than in other symbiotic stars. The origin of this difference is unknown.

The emission line profiles of CH Cyg are complicated and highly variable. Most lines indicate mass motions with velocities of 100–200 km s^{-1}. These lines are usually symmetric about the systemic velocity, but sometimes display extra emission with positive or negative velocity. When the H I lines are strong, they often have broad profiles with deep central absorption reversals. These

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

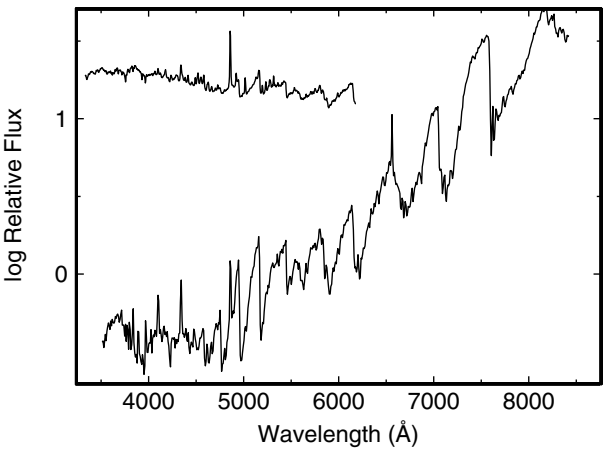


Figure 2. Spectra of CH Cyg in the low state (lower curve) and the high state (upper curve). The low-state spectrum resembles that of a normal M6 giant with weak H I emission lines and a variable blue continuum. In the high state, the blue continuum rises by more than a factor of 10 and resembles that of an F-type supergiant with modest H I emission lines. The right half of the high-state spectrum has been omitted for clarity; it blends smoothly into the low-state spectrum at roughly 7000 Å.

profiles resemble emission lines observed in dwarf NOVAE and other cataclysmic variables. However, the lines often become single peaked, which is not typical of dwarf novae. The variability of the H I lines may be associated with the complex photometric variations, but the available data are not extensive enough to make a clear association with any of the photometric periods.

Surprisingly, some of the most interesting activity of CH Cyg occurs during low states. In 1984, the hot component began to eject material at high velocities shortly after a 2–3 mag drop in the optical brightness. As the F-type continuum and several emission lines faded during this decline, the star developed strong [O III] emission lines indicating material moving at velocities of $\sim$500 km s^{-1}. The intensity of 1–10 cm radio continuum emission increased by a factor of 3 in roughly 1 month. Nearly all of this radio emission was confined to two knots of gas moving away from the central star at a projected velocity of $\sim 1000(d/200\,\mathrm{pc})$ km s^{-1}. This ejection is similar to the bipolar jets observed in ACTIVE GALACTIC NUCLEI and PRE-MAIN-SEQUENCE STARS. Spatially resolved optical spectra clearly associated the [O III] emission with the radio knots. The radio emission declined as the knots moved away from the central source and then rose slightly as the source ejected a second, smaller set of knots moving at the same velocity. This event began $\sim$200 days after the first ejection and was accompanied by a factor of 2 decline in the optical brightness. A third small ejection began $\sim$200 days later but was not associated with a drop in brightness. All of these knots faded over several years, when the system resumed its 'normal' low-state spectrum.

Jet formation associated with a significant drop in the optical brightness began again in 1997–1998. Several optical emission lines simultaneously developed broad features indicating mass motions with velocities of 500–1000 km s^{-1}. The system recovered quickly, rising by 1.5 mag in the ultraviolet during May–July 1998. The 14 yr interval between the two radio outbursts is 1–2 yr shorter than one of the proposed orbital periods.

Interpretation

The origin of CH Cyg's bizarre behavior is unclear. Three emission states—inactive, low and high—are unusual: most symbiotics have low and high states. Multiple strong periodicities in the light curve are also rare. The weak emission line spectrum is atypical, although Mira and a few other symbiotics have similar emission line spectra.

The current picture for the CH Cyg system begins with an interacting binary composed of an M-type giant and a hot WHITE DWARF. The white dwarf accretes material lost by the giant. Energy generated by ACCRETION powers the observed activity. Irregular pulsations of the giant modulate the accretion rate onto the white dwarf, which varies the activity level. It is unlikely that the white dwarf generates much energy by thermonuclear burning of the accreted mass: the luminosity at maximum, $\sim 100 L_\odot$, is small for a white dwarf with a nuclear burning shell.

Adding finer details to this picture has been difficult. Two variants have been proposed. The first envisions a wide binary with an orbital period of $\sim$5500 days. A pulsating red giant is responsible for the $\sim$100 day and the $\sim$700 day periods. A magnetic white dwarf—similar to the white dwarf in Z And—rotates with a period of 2500–3000 s. Accretion onto the magnetic poles of this white dwarf produces the short-timescale variations. Inactive states occur near apastron, when the white dwarf is far away from the red giant and the accretion rate is low. These states are long lived, because the binary spends most of its time at apastron. Active states occur near PERIASTRON, when the accretion rate is high. In this model, cycling between the low and high states results from an imbalance between the magnetic pressure of the rotating white dwarf and the ram pressure of material in the red giant wind. High states occur when the ram pressure exceeds the magnetic pressure. Accretion onto the white dwarf then produces the F-type continuum, although the physical mechanism producing this continuum is unclear. Low states occur when the magnetic pressure exceeds the ram pressure and prevents accretion onto the white dwarf. This 'propeller' mechanism also drives the bipolar jets by a process which is poorly understood at best.

A second concept considers CH Cyg as a triple system. In this model, the red giant and white dwarf inhabit a close binary with a 760 day period. A mundane G dwarf orbits this binary with a 5500 day period. Variants of this picture replace the G dwarf with another M giant, but spectroscopic evidence for two red giants in the system is not compelling. A triple system provides a natural explanation for two persistent modulations of the system, 760 days and 5500 days. Red giant pulsations are still necessary for other long periods; a magnetic white dwarf

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is still responsible for the 2500–3000 s optical pulsation. An advantage of this model is that the close inner binary is similar in structure to a typical symbiotic star. The enhanced mass transfer possible in a closer binary could power the unusual activity of the hot component. A propeller model could still give rise to cycling between high and low states, but the origin of the inactive state is unclear. Other symbiotics stars with ∼700 day orbital periods are never inactive. Short periods of *apparent* inactivity in some symbiotics occur when the red giant occults the hot component. The most popular version of the triple-star model proposes that both red giants eclipse the hot component to produce inactive periods. This geometry is *a priori* unlikely and may not produce long enough inactive periods. Explaining the jet activity in the low state is also a challenge, because almost all other symbiotics eject more material in high states than low states.

Most analyses of CH Cyg have concentrated on photometric data. Detailed models of spectroscopic data will be needed to understand the energetics of the hot component and the dynamics of the NEBULA. Neither proposed model for CH Cyg makes many predictions for future behavior of the system; further work may point to better tests of the models and allow a definitive choice to be made between them.

Bibliography

Hinkle K H, Fekel F C, Johnson D S and Scharlach W W G 1993 The triple symbiotic system CH Cygni *Astron. J.* **105** 1074

Karovska M and Mattei J A 1992 Longterm AAVSO observations of the symbiotic system CH Cygni *J. AAVSO* **21** 23

Kenyon S J 1986 *The Symbiotic Stars* (Cambridge: Cambridge University Press)

Mikołajewski M, Mikołajewska J, Tomov T, Kulesza B and Szczerba R 1990 Symbiotic binaries III. Flickering activity of CH Cygni: magnetic rotator model *Acta Astron.* **40** 129

Munari U, Yudin B F, Kolotilov E A and Tomov T V 1996 UBV-JHKLM photometry of CH Cyg over 1978-1995: dust properties and doubts on the triple star model *Astron. Astrophys.* **311** 484

Skopal A, Bode M F, Lloyd H M and Tamura S 1996 Eclipses in the symbiotic system CH Cygni *Astron. Astrophys.* **308** L9

Scott Kenyon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CH Stars and Barium Stars

The classical barium (or 'Ba II') stars are RED GIANT STARS whose spectra show strong absorption lines of barium, strontium and certain other heavy elements, as well as strong features due to carbon molecules. Together with the related class of CH stars, the Ba II stars were crucial in establishing the existence of neutron-capture reactions in stellar interiors that are responsible for the synthesis of heavy elements. More recently, the Ba II and CH stars have provided strong evidence for mass transfer in binary systems that drastically alters the surface composition of a companion star.

Barium stars

Ba II stars were first recognized by W P Bidelman and P C Keenan in 1951. They are G-K giants whose spectra show an unusually strong absorption line of Ba II at 4554 Å, along with enhanced lines of Sr II and molecular bands of the CH, C_2 and CN molecules. The prototype is the third-magnitude star ζ Capricorni. During the ensuing decade, studies of the spectra at high dispersion revealed that the abundances of numerous heavy elements are enhanced, notably the elements Sr, Y and Zr; rare earths such as La, Ce, Nd and Sm; and of course Ba. It was soon realized that these are exactly the elements expected to be synthesized when 'seed' nuclei in the iron group are exposed to a flux of neutrons and are thus built up to higher atomic weights, on a timescale that is slow compared to the beta-decay timescale. This so-called '*s*-process' leads to high abundances of those chemical elements with low neutron-capture cross sections (these being the 'bottlenecks' in the neutron-capture chain), which are exactly the elements in the atomic-weight peaks around Sr and Ba.

CH stars

The CH stars are a closely related class, actually first recognized even earlier than the Ba II stars by P C Keenan in 1942. They are essentially the Population II counterparts of barium stars, having similar enhancements of *s*-process elements and very strong bands of the CH molecule (see also POPULATION II STARS). However, CH stars have weak lines of the ordinary metals such as iron, since these stars belong to the old, metal-deficient population of our Galaxy.

CH stars are found both in the Galactic halo and in several globular clusters. The cluster ω Centauri contains the richest population of CH stars, about half a dozen.

The luminosity problem

By the early 1970s, the nucleosynthetic origin of the enhanced chemical elements in the Ba II and CH stars was reasonably well understood: they arise in stellar interiors when a flux of neutrons irradiates iron-group seed nuclei. This can happen, for example, during mixing associated with helium thermal pulses which brings together ^{13}C nuclei and alpha particles; each reaction releases a ^{16}O nucleus and a free neutron. Carbon is synthesized during helium burning in the cores of red giants. Subsequently, a mixing process causes the products of these reactions to be exposed at the stellar surface. In the mid-1970s, however, detailed calculations of stellar evolution revealed a severe evolutionary problem: the sort of mixing required (called the 'third dredge-up') only occurs in highly evolved, luminous asymptotic-giant-branch (AGB) stars. Yet the Ba II and CH stars were known to have only the luminosities of ordinary red giants, far too faint for any dredge-up to occur.

The problem was exacerbated by the discovery in 1974 by H E Bond of what he called 'subgiant CH stars,' which are F- and G-type stars of even lower luminosities, extending in some cases down to the main sequence, and showing similar patterns of enhanced heavy elements and carbon. Subsequent work has suggested that 'dwarf barium stars' would be a better name for most of the members of this class, since most of them are not extremely iron-deficient. Even further down the main sequence, at spectral type K, a class of 'dwarf carbon stars' was discovered in the late 1970s. All of these objects are far too faint to have synthesized *s*-process elements and carbon in their own interiors, let alone to be mixing them up to the stellar surface.

Moreover, none of these objects appear to be particularly rare stars; for example, large-scale spectroscopic surveys show that about 1% of red giants are members of the Ba II class, and a similar fraction of F-G dwarfs are barium dwarfs or subgiant CH stars.

The solution

The solution to the luminosity problem came in the early 1980s, when R D McClure and his collaborators established that virtually all Ba II and CH stars are spectroscopic binaries, with fairly long periods and typical orbital separations of the order of 2 AU. This led to the realization that the binary system once contained an AGB star that mixed carbon and *s*-process elements to its surface, and then transferred this 'polluted' material to a companion star. In many cases the companion would have been on or near the main sequence, and it is now seen as a dwarf barium star, dwarf carbon star or subgiant CH star. When the contaminated main-sequence star evolves to become a red giant (or if the mass transfer is from the AGB star directly to a red-giant companion), it will be seen as a classical red-giant Ba II star (or as a CH star if the binary is a metal-poor member of Population II).

The solution confirmed

A direct test of the binary scenario would be to find a white-dwarf companion of a barium star. The above scenario in fact *demands* that all of the companions be white dwarfs, i.e., the remnants of the former AGB star.

In a stunning confirmation of the scenario, E Böhm-Vitense showed in 1980 that ζ Cap has a hot white-dwarf companion, revealed in the ultraviolet by the INTERNATIONAL ULTRAVIOLET EXPLORER satellite. Subsequent UV observations with IUE, and more recently by Böhm-Vitense and her collaborators with the HUBBLE SPACE

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

TELESCOPE, have revealed several more examples of Ba II stars with white-dwarf companions. In many cases, however, the white dwarf has faded to below detectability, and its presence shows up only in the variable radial velocity of the peculiar optical star.

The pollution mechanism

Recent work on the Ba II and related stars has focused on the precise mechanism by which the AGB star contaminates its innocent companion. The two main possibilities are (a) that the AGB star evolves to fill its Roche lobe, and then spills substantial amounts of its envelope material onto the companion; and (b) that the AGB star pollutes the companion only through a stellar wind. In scenario (a), the outcome would usually be a common-envelope interaction, with a spiralling down of the orbit to a much smaller separation, so that the binary emerges as a cataclysmic variable rather than a barium star. However, some authors have argued that stable Roche-lobe overflow is possible, thus avoiding the spiral-down. Scenario (b) is probably more likely at least for most barium stars, both because of their still large orbital separations and because many of them have somewhat eccentric orbits, which would not be the case following Roche-lobe overflow.

Bibliography

Ba II stars were discovered by

Bidelman W P, and Keenan P C 1951 The Ba II stars *Astrophys. J.* **114** 473–6

A comprehensive review article covering work through the early 1980s, including the first demonstration that most or all Ba II stars are spectroscopic binaries, was written by

McClure R D 1984 The barium stars *Pub. Astron. Soc. Pacific* **96** 117–27

CH stars were first recognized by

Keenan P C 1942 The spectra of CH stars *Astrophys. J.* **96** 101–5

Subgiant CH stars were defined by

Bond H E 1974 The subgiant CH stars *Astrophys. J.* **194** 95–107

The first dwarf carbon star was discovered by

Dahn C C *et al* 1997 G 77-61: A dwarf carbon star *Astrophys. J.* **216** 757–66

Details of the formation of Ba II and CH stars have been discussed by

Han Z, Eggleton P P, Podsiadlowski P and Tout C A 1995 The formation of barium and CH stars and related objects *Mon. Not. R. Astron. Soc.* **277** 1443–62

A modern review of Ba II and CH stars can be found in

Jorissen A, Van Eck S, Mayor M and Udry S 1998 Insights into the formation of barium and Tc-poor S stars from an extended sample of orbital elements *Astron. Astrophys.* **332** 877–903

A comprehensive review of all aspects of carbon stars and related objects is

Wallerstein G and Knapp G R 1998 Carbon stars *Ann. Rev. Astron. Astrophys.* **36** 369–433

Howard E Bond and Edward M Sion

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Challis, James (1803–82)

British astronomer, born in Braintree, Essex, England, became Plumian professor and Director of the Observatory at Cambridge. Was persuaded by ADAMS to search for Neptune, but had no up to date chart and had to laboriously observe all the stars in the area twice to find one that moved. Observed Neptune, but only once, in July 1846, and was not re-observing the field of stars with any urgency, so lost priority to J G GALLE and H L D'ARREST in Berlin.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chamaeleon

(the Chameleon; abbrev. Cha, gen. Chamaeleontis; area 132 sq. deg.) A southern constellation which lies between Carina and the south celestial pole, and culminates at midnight in early March. It was first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, the brightest stars in Chamaeleon are α and γ Chamaeleontis, both magnitude 4.1. Interesting objects include δ Chamaeleontis, a wide double star with bluish-white (B2.5) and orange (K0) components, magnitudes 4.5 and 5.5, separation about 6′, the fainter of which (δ^1) has a fainter companion, and NGC 3195, a twelfth-magnitude planetary nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chamberlin, Thomas Chrowder (1843–1928)

Geologist, born in Mattoon, IL, head of the geology department at the University of Chicago; researched on the origin of the Earth and the solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chandra X-ray Observatory

Launched on 23 July 1999 on board the Space Shuttle Columbia from Cape Canaveral, the Chandra X-ray Observatory is the first x-ray astronomy telescope to match the 1/2 arcsecond imaging power and the 0.1% spectral resolving power of optical telescopes. Chandra is named after Subramanian Chandrasekhar, known as Chandra, and author of the Chandrasekhar limit. Chandra has been extremely successful and productive so far. (See http://chandra.harvard.edu for many examples.)

Any experiment that is 10 or 100 times better than its predecessors is virtually certain to make revolutionary discoveries and CHANDRA is demonstrating this truism dramatically. Examples include settling decade-old debates with a single observation: yes cometary x-rays are due to charge exchange; no, most elliptical galaxies do not have much hot gas in their interstellar medium, but they do have their quota of compact binaries; no, cooling flows in clusters of galaxies only cool so far (down to 2×10^7K); and yes, the COSMIC X-RAY BACKGROUND is made up of the light from many quasars—about 1000 in each area of sky the size of the Moon.

Other examples are of completely new discoveries still to be checked out with further studies: a type of quasar that, perversely, is obscured in x-rays, but shines through clearly in optical light; spectral absorption lines that seem to come from a bath of hot gas that our Milky Way galaxy, and the whole of our Local Group of galaxies, is lying in. (There appears to be so much of this 'Warm Hot Intergalactic Medium' or 'WHIM' that it could even be the 'missing mass' that binds the Local Group galaxies together); the relativistically moving jets from quasars shine brightly in Chandra images, as long as clusters of galaxies are across, and brighter when they are further away, thanks to the higher temperature and intensity of the Cosmic Microwave Background, so there are more of these photons to be scattered off the fast moving electrons of the jet, so gaining enough energy to become the x-rays we see. Then there are the nests of super-bright x-ray binaries in starburst galaxies that appear to exceed the EDDINGTON LIMIT of luminosity if they have a normal 10 solar mass black hole in them. Could they imply a new and predicted-not-to-exist population of intermediate mass (100–10(4) solar masses) black holes? The spectra of a neutron star (i.e. one not in a binary or a supernova remnant) shows a black body spectrum with a well determined temperature, together with its distance gives the luminosity and so combining them the area of the NEUTRON STAR. This could be the first example of a new type of star and a new type of matter, a QUARK STAR.

Chandra's resolution leads to 100 times more sensitivity over earlier x-ray missions because background light is imaged away allowing much fainter sources to be seen, and also creates the good spectral resolution. Chandra's resolution stems from its unprecedented x-ray mirror. The x-rays that Chandra images span 0.1–10 keV (12–0.12 nm). X-ray optics only reflect x-rays if they strike glancingly, at about 1°. So x-ray mirrors only bend light slightly, giving x-ray telescopes large f ratios, 8.3 for Chandra. So Chandra's 1.2 m diameter mirror has to have a focal length of 10.070 m. To increase Chandra's area four sets of almost cylindrical mirror shells are nested within one another, giving a total of 800 cm^2. To improve their reflectivity they are coated with iridium, a high-Z, high-density element. Since two reflections are needed to make an image, Wolter type 1 parabolas/hyperbolas are used. With photon wavelengths of only a few nm the mirror surfaces must be smoother than this (<0.04 nm for Chandra), else photons would be scattered to large angles, blurring the image.

To disperse the x-ray spectrum, Chandra has three sets of transmission diffraction gratings, optimized for low (LETG), medium (MEG) and high (HEG) energies. They all have line densities of about 1000 mm^{-1}, and are constructed using semiconductor technology. The gratings behave like objective prisms, making an image of the source for each wavelength. Since a star makes a small image in Chandra, different wavelengths are well separated.

Chandra has two types of detectors: 'ACIS' is made from charge-coupled devices (CCDs), cousins of video cameras CCDs; 'HRC' is made from microchannel plates, used in night vision goggles. HRC can time the arrival of x-ray photons to 16 ms, has the smallest pixels (0.4 arcsec) and has a large (30 arcmin × 30 arcmin) field of view. The HRC has sensitivity to low-energy photons (to E~0.05 keV), and so is used to image the spectra from LETG. ACIS has better efficiency and energy resolution. ACIS uses two types of CCD chips: 'frontside illuminated' (FI) which detect higher energy (E>1 keV) x-rays have low background, while 'backside illuminated' (BI) work well (QE>0.5) down to 0.5 keV. X-ray CCDs determine the energy of each x-ray from the number of electrons it produces. An accidental exposure of the main 16 arcmin × 16 arcmin array of four FI CCDs to energetic (100 keV) protons in the solar wind just after launch degraded their energy resolution. Much of this resolution has been regained by cooling the chips and by special software. The smaller 8 arcmin × 8 arcmin BI chip was unaffected.

Because every photon is recorded individually, x-ray images have four dimensions: position (x,y), energy and time. So, unlike the Hubble Space Telescope, Chandra does not have to be held immobile while exposing an image. Any blur is removed by repositioning each photon in software (using star positions from an 11.2 cm diameter optical telescope) at the Chandra X-ray Center

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

which operates Chandra for NASA from Cambridge, Massachusetts (USA). The ability to slice x-ray images at will is invaluable for analysis, e.g. to extract the spectrum of whatever looks interesting, without having to decide in advance. (The 'Proposers' Observatory Guide' has technical details on Chandra at http://asc.harvard.edu/.)

Chandra is in a high (10 000 km × 140 000 km) orbit. This makes any part of the sky observable without continual blocking by the Earth, increasing available observing time and allowing long uninterrupted observations for studying variable sources. Chandra cannot be reached with the Space Shuttle, so its survival depends on no critical system breaking or running out of supplies. Everything on board is set to last 15 yr or more. No plans exist to build another x-ray telescope with comparable resolution. Fortunately, NASA plans to operate Chandra for at least 10 yr.

Martin Elvis

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Chandrasekhar, Subrahmanyan (1910–95)

Theoretical astrophysicist, born in Lahore, India, Nobel prizewinner (1983) 'for his theoretical studies of the physical processes of importance to the structure and evolution of the stars'. At Cambridge and Copenhagen he developed the theory of white dwarf stars, showing that quantum mechanical degeneracy pressure cannot stabilize a massive star, and that white dwarfs have a maximum mass (the *Chandrasekhar limit*). At the University of Chicago and Yerkes Observatory he investigated and wrote books on stellar structure and evolution, the dynamical properties of star clusters and galaxies, radiative transfer, hydrodynamic and hydromagnetic stability, the stability of ellipsoidal figures of equilibrium, and the mathematical theory of black holes. Like SIR JAMES JEANS, his work was characterized by an intense and protracted investigation of a certain area that appealed to him, which he then in retrospect presented in a definitive coherent book with order, form and structure. He edited the *Astrophysical Journal* for nearly 20 years. NASA's advanced x-ray astronomy facility satellite, AXAF, has been renamed Chandra.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. A 4096 × 4096 pixel CCD with 15 micron pixels packaged for use in a large astronomical spectrograph.

Charge-coupled Device

Charge-coupled devices (CCDs) are silicon-based optical detectors used in most near-UV, visible and near-IR imaging and spectroscopic astronomical instruments. They range in format from television resolution (340 × 512 pixels) to 10 000 × 10 000 pixels. Modern CCDs have intrinsic noise of just a few electrons, dynamic range of 100 000 and quantum efficiency of over 90% throughout most of their useful operating wavelengths. These characteristics make CCDs the detector of choice for most visible astronomical detection requirements.

History and fundamentals

Boyle and Smith invented the CCD at Bell Labs in 1969. While originally intended as a memory device, CCDs have been most widely used as imaging detectors (see e.g. INFRARED ARRAYS, INFRARED IMAGERS, WIDEFIELD CCD IMAGERS). They were first used in astronomy in the late 1970s and rapidly replaced photographic film (see CAMERAS), photodiode arrays and photocathode-based detectors. A large format CCD is shown in figure 1.

The CCD basically consists of a one- or two-dimensional array of metal-oxide-semiconductor (MOS) capacitors in which the gate ('metal') of each capacitor is connected to buses to which external voltages are applied. These gates are actually made of polysilicon rather than metal to allow better light transmission through the frontside structure. Potential well minima are formed in the silicon under the CCD gates with the most positive applied voltage. Electrons are created from incident photons via the photoelectric effect and these photogenerated electrons are collected in the nearest potential well.

CCDs are sensitive to all wavelengths shorter than the cut-off wavelength $\lambda_{\rm cutoff}$ at which an incident photon's energy E is the same as the detection material's bandgap energy. All photons of shorter wavelength can excite an electron from the detector's valance band to its conduction band. For silicon, the most commonly used CCD fabrication material, $E_{\rm bandgap} = 1.12\,{\rm eV}$, so $\lambda_{\rm cutoff} = 1.1\,\mu{\rm m}$.

All gates that are electrically tied together are referred to as a phase. The majority of CCDs used in astronomy are three-phase devices, meaning that three electrical voltages must be applied to shift charge across the CCD. The distance from one phase to the next on the same bus defines the resolution of the detector, and is called a pixel. Typical CCD pixel sizes are 5–30 μm.

The most commonly used CCD in astronomy consists of a two-dimensional array of pixels. The image section of such a device can be viewed as an array of columns and rows of pixels with a single serial shift register at the end of the columns. Charge is shifted (or charge-coupled) one row at a time in all columns at once, shifting the last row into the serial register. The columns are defined by channel stops of highly doped material which confine charge from spreading between adjacent columns. In this manner no gates are required to define pixels on the horizontal direction in the imaging area. The serial register is itself clocked to sequentially shift charge to an output amplifier located at its end. External electronics and a computer are used to convert this analog output sequence of voltages into an image.

Architecture and operation

The operation of a CCD can be divided into three modes: clearing, integration, and readout.

Clearing is required to remove unwanted charge acquired by the device prior to making an integration. One method of clearing a device is to shift the charge from every pixel out of the device without recording the resultant data.

During integration, positive voltages are set on some phases to define potential minima where photoelectrons are collected. The positive voltage is applied to one or two of the three phases. At least one phase of each pixel must be more negative to create a barrier to charge spreading, without which the image would smear. It is possible for this barrier to be created by a fixed implant rather than an applied voltage. During integration, photoelectrons are collected but no shifting occurs.

A CCD must be cooled if the integration time is more than a few seconds to avoid the silicon's self-generated dark signal filling the potential wells. Most astronomical CCDs are cooled to about $-100\,^{\circ}$C in a liquid nitrogen dewar. Many smaller CCDs which are used for telescope guiding applications (where exposure times are just a few seconds) rely on thermoelectric coolers operating at about $-40\,^{\circ}$C to reduce dark signal.

After the integration time has expired, the CCD enters the readout phase. The location of each charge packet created during integration is shifted towards the output when a specific voltage sequence is applied to the buses, as shown in figure 2. Depending on architecture, split parallel clocks can be used to shift half the pixels toward one serial register and the other half toward another serial register, allowing readout in half the time. The serial register can also be split so that charge can be shifted toward amplifiers at each end, also decreasing readout time.

There are several common architectures of scientific CCDs. The entire area of a full frame device is used to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

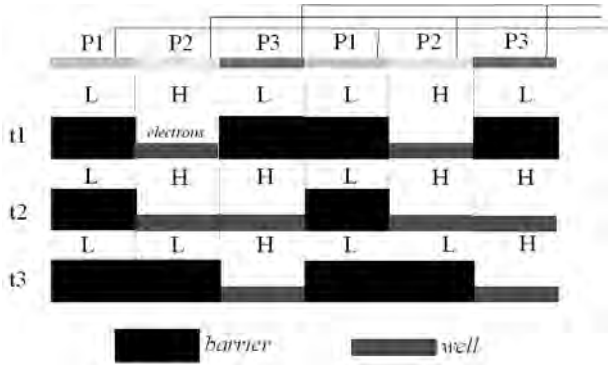


Figure 2. The pixel structure of a typical three-phase CCD showing the applied voltages, potentials, and charge packet locations at three separate times. The times are represented by t1, t2, and t3. Applied voltage is indicated by L (low) or H (high). The three-phases are represented as P1, P2, and P3.

collect images and is the most widely used in astronomy. A full frame CCD requires a shutter to open during integration (or exposure) and close during readout to avoid streaking from the light source while charge is shifted. In a frame transfer CCD, half of the imaging area is covered by an opaque frame-store mask. The unmasked section collects charge during integration. When the exposure is complete, the charge is very rapidly (10^{-6} to 10^{-4} s) shifted to the frame-store region where slower readout can take place while the next integration occurs in the image section. This eliminates the need for a mechanical shutter at the expense of losing half the detector's area for imaging. An interline transfer CCD is similar to a frame transfer device in that it has an opaque bus along each column into which charge can be rapidly shifted after each integration. Interline transfer devices are the most commonly used CCDs for high-light applications (such as TV video) but are seldom used in astronomical applications.

Amplifiers

When electrons are shifted to the end of a serial register they are next detected by the output amplifier. The gate of a transistor (usually a field effect transistor, or FET) is connected to a sense node onto which the charge is shifted from the last serial pixel. The charge appears as a voltage on the gate of the FET due to the capacitance of the node. The conversion factor is typically 1 to 10 μV per electron for modern amplifiers. The output voltage is buffered by the FET to create a measurable voltage across a load resistor located off-chip. This output voltage is easily amplified in the controller electronics and converted to a digital signal by an analog-to-digital converter. The node must be reset before sensing each pixel so that charge does not accumulate from pixel to pixel.

Read noise is a measure of the uncertainty in the output of the CCD and is often the dominant noise source for low-light-level imaging applications. Double correlated sampling (DCS) reduces the uncertainty in the absolute charge level of the node at each reset. Before shifting each pixel, the node is reset with an on-chip Reset FET. The output voltage is sampled and recorded, the pixel to be measured is then shifted onto the node, the node is sampled again, and the difference between the two samples is calculated and recorded as the actual pixel value. By using DCS and low-noise output FETs, read noise of less than two electrons can be achieved.

Quantum efficiency

The quantum efficiency (QE_λ) of a CCD is given by

$$QE_\lambda = (1 - R_\lambda) \exp(-\alpha_\lambda / t)$$

where R_λ is the reflectivity of the incident detector surface, α_λ is the absorption length of photons at the specified wavelength and t is the effective device thickness. Front illuminated CCDs have photons incident on the gate structure which absorbs almost all blue and UV light. These devices are therefore not useful for $\lambda < 400$ nm. In back illuminated (or thinned) CCDs, photons are incident on the backside of the device, opposite the gate structure. These devices must be very thin (<20 μm) for photoelectrons to be collected under the potential minima near the frontside gates. The thinning process is difficult and expensive, but QE is limited only by reflection at the back surface and the ability of silicon to absorb photons. Back illuminated devices can be antireflection coated to obtain peak QE of greater than 95% at specific wavelengths. The QE of CCDs drops significantly toward the near-IR because silicon's absorption length increases dramatically from 100 Å in the UV to 100 μm in the near-IR.

Bibliography

Excellent references on CCDs and optical detector technology are:

Theuwissen A J P 1995 *Solid-State Imaging with Charge-Coupled Devices* (Dordrecht: Kluwer)

Dereniak E L and Crowe D G 1984 *Optical Radiation Detectors* (New York: Wiley)

Many important papers on CCD technology can be found in the various *Proceedings of the SPIE*.

Michael Lesser

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Charlier, Carl Vilhelm Ludwig (1862–1934)

Swedish astronomer, became director of the Lund observatory. Worked in celestial mechanics, instrumentation and statistics, studying the distribution and motions of stars in the solar neighborhood. He showed that the hotter stars and galactic clusters form a flattened system. He proposed a hierarchical model of galaxies in the universe, extending to infinity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chaucer, Geoffrey (*c.* 1343–1400)

Poet, probably born in London, England. Author of the *Canterbury Tales*, which show his familiarity with astrological matters, and *A Treatise on the Astrolabe*, once believed to have been written for a son of Chaucer's, but now thought to be for the son of a friend, Lewis Clifford. The text is the oldest known 'technical manual' in the English language.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chemical Compositions of Stars

In 1835, in a famously inaccurate forecast, the French philosopher Auguste Comte wrote of stars that, 'We understand the possibility of determining their shapes, their distances, their sizes and their movements; whereas we would never know how to study by any means their chemical composition . . .'. At the close of the 20th century the accurate measurement of the abundances of the chemical elements in the visible surface layers of stars of all classes has become a relatively straightforward astrophysical procedure. This is the result of several important developments: (1) a physically realistic theory of the structure of stellar atmospheres; (2) high-powered computer workstations which can efficiently integrate the equation of radiative transfer through a model stellar atmosphere; (3) advanced theoretical codes dealing with physically complex situations, such as departures from local thermodynamic equilibrium (LTE); (4) improvements in the basic atomic data needed to calculate the opacities which produce spectral absorption lines; (5) high-resolution spectrographs equipped with sensitive and linear electronic detectors; (6) access to the full range of wavelengths at which stars emit their light, including ultraviolet wavelengths observed from above the Earth's atmosphere.

The abundances of the chemical elements are a thread that ties together much of what we know about the evolution of stars, galaxies and the universe as a whole. The origin of the universe in the big bang yielded the most abundant of the elements, hydrogen (^{1}H) and helium (^{4}He), as well as small quantities of the isotopes ^{2}H (deuterium), ^{3}He and ^{7}Li. All the remaining elements and isotopes found in nature have been produced by the nuclear reactions that power the stars during the normal course of their evolution, or by rapid NUCLEOSYNTHESIS during the final cataclysmic explosions of certain stars as SUPERNOVAE or, for a few elements, by the spallation of heavier nuclei by COSMIC RAY particles in the interstellar medium. Some of the products of nuclear reactions occurring within a star are injected back into the interstellar medium as the star casts off most of its mass during the brief final stages of its evolution. There they provide the raw material for the formation of later generations of stars, planets and ultimately life itself.

This scenario for the origin of the elements leads to a profound consequence. The relative abundances of the elements are not necessarily constant in time or space. The initial chemical composition of a star is that of the interstellar cloud of gas and dust from which the star formed (see also INTERSTELLAR CHEMISTRY; COSMIC ABUNDANCE OF ELEMENTS). It is a function of the locality of the star's birth in the galaxy and of the age and state of evolution of the galaxy at the time the star formed. The composition of the observable surface layers of a star may be modified over time by processes occurring within the STELLAR INTERIOR. For example, nuclides created by reactions in the star's central core, or in shells surrounding the core, may be dredged up and intermixed with unprocessed material at the star's surface during the course of its evolution. Alternatively, the outer layers of an evolving star may be stripped away by mass loss, leaving its nuclear- processed interior exposed to view. Even in a young unevolved star atoms or ions of individual elements may settle below the observable surface under the pull of gravity or may be driven upward by radiation pressure and achieve high concentrations within the surface layers, if the atmosphere and envelope of the star are highly stable. Finally, a star may accrete processed material from a binary companion undergoing mass loss.

Overall, this is a wonderful complexity. It allows us to use the observed chemical compositions of stars as tracers of the record of evolution of our own and perhaps of other galaxies. It allows us to observe the outcome of processes occurring in the interiors of stars normally hidden from our view. Moreover, at least an approximate knowledge of the chemical composition of a star is needed to deduce the opacity and the mean molecular weight of its material, which are essential input parameters for the computation of theoretical models of its atmospheric and interior structure. Such models are used in turn to perform detailed compositional analyses and to predict other observed stellar properties. The consistency of these theoretical models with the observables they are used to predict—stellar luminosity and color, the detailed distribution of stellar flux with wavelength, spectral line strengths and shapes—gives confidence in or sometimes raises questions about our physical understanding.

How the chemical composition of a star is measured

Only the surface layers of a star are directly observable with an astronomical spectrograph mounted on a telescope. Therefore, the abundances of the chemical elements derived from spectroscopically dispersed starlight reveal only the composition of the star's atmosphere.

Our ability to measure a star's surface composition is enabled by the quantum structure of atoms and ions. Transitions of electrostatically bound electrons between quantum energy states, and the associated absorption or emission of photons, imprint on the stellar spectrum a pattern of discrete lines with wavelengths and relative strengths that are unique for every ionization state of every element in the periodic table. Accurate measurements of the wavelengths and relative strengths of spectral lines allow the identification of which chemical species are present. The measured absolute strengths of the lines—the amount of flux they have removed from or added to the background continuum of light emitted by the star—yield the abundances of those species. The atmospheres of cool stars also contain molecules (see STELLAR ATMOSPHERES: COOL STARS). The quantum structure of their electronic and vibrational energy states produces bands of closely spaced, discrete absorption or emission lines from which abundances can be measured.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

While it is possible to measure abundances from emission lines (lines brighter than the continuum) produced in the hot, tenuous chromosphere or corona of a cool star such as the Sun, the simplest and most accurate determinations are obtained from the absorption lines arising in the denser stellar photosphere (the Fraunhofer spectrum). The reason lies in a set of simplifying assumptions which characterize the physical conditions within the photospheres of many stars rather well, enabling accurate computations of their structures and emitted spectra. These include (1) LTE (the local radiation field is coupled to the local gas kinetic temperature), (2) plane parallel geometry (the photosphere is very thin, compared with the radius of the star), (3) radiative equilibrium (energy is transported primarily by radiation) and (4) hydrostatic equilibrium (the photosphere is essentially motionless). In principle, violations of any of these assumptions can be treated with physically more complex theoretical computations. However, these will not necessarily be of greater reliability, if input physical parameters are inaccurate.

The observed depth of an absorption line centered at wavelength λ, at a position in the line's profile $\lambda + \Delta\lambda$, is a direct measure of the opacity of the stellar atmosphere produced by discrete electronic transitions within atoms or ions of the element in question, as they absorb photons of wavelength $\lambda + \Delta\lambda$. The chemical abundance of the element is derived by comparing the observed strengths and shapes of its spectral lines with a theoretical model of the emergent flux from the star, calculated by solving the equation of RADIATIVE TRANSFER through a model stellar atmosphere in the presence of this line opacity.

A theoretical model stellar atmosphere appropriate to a particular star provides a tabulation of the gas kinetic temperature, pressure and electron density at any depth x below the nominal 'surface' of the star. It is calculated with assumed abundances of hydrogen and helium and at least an approximate estimate of the abundances of heavier elements. From this information the total continuous opacity κ_λ at any wavelength, produced at depth x by the ionization of atoms or ions (bound–free electronic transitions), by molecular dissociation and by electron scattering can be derived. Radiation emerging from the star at all wavelengths is attenuated by these continuous opacity sources and both line and continuous opacity at any wavelength must be included in the calculation of emergent flux. In the absence of spectral lines, that flux is set entirely by the continuous opacity, and thus spectroscopists sometimes refer to the 'line-free continuum' as a reference flux level against which to measure the strength of a spectral line.

The line opacity is expressed in terms of a line absorption coefficient (units $\mathrm{cm}^2\ \mathrm{g}^{-1}$)

$$l_\lambda(x, \Delta\lambda) = \text{const} \times [n_{\mathrm{el},r,s}(x)/N_{\mathrm{el}}](N_{\mathrm{el}}/N_{\mathrm{H}})N_{\mathrm{H}} f \phi_\lambda(x, \Delta\lambda) \times (1 - \mathrm{e}^{-hc/\lambda kT}). \quad (1)$$

The factor $N_{\mathrm{el}}/N_{\mathrm{H}}$, the ratio of the total number of atoms and ions of the element per gram of stellar material to the number of atoms and ions of hydrogen per gram, is the key abundance parameter we seek to measure. It is allowed to vary freely in the calculations until they produce a good match to the observations. All other factors in equation (1) must be evaluated with care lest they corrupt the accuracy of the abundance measurement.

The factor $n_{\mathrm{el},r,s}(x)/N_{\mathrm{e}}$ is the population fraction of the element that lies in the excitation state s of ionization state r from which the line arises. In LTE this fraction is given by the combined Boltzmann and Saha equations. The assumption of LTE is a reasonable approximation of physical conditions in the deeper, denser layers of a stellar atmosphere where many weak and moderately strong spectral lines are formed. However, a more rigorous treatment of departures of energy level populations and ionization equilibria from their LTE values is sometimes required. The number of hydrogen atoms and ions per gram of stellar material, N_{H} in equation (1), follows directly from the abundance mix of hydrogen, helium and metals assumed in the original model atmosphere calculation.

The factor f is the transition probability, the likelihood that the atom will absorb or emit a photon between two particular energy states. It can be determined in the laboratory, for example by measuring the lifetime of the upper level of the transition, selectively populated by laser-induced excitation. This value is then combined with the measured fraction of downward transitions from the upper level that end up in the lower excitation state of the line in question (the branching fraction) to obtain the transition probability. The f value can also be calculated with semi-empirical codes that use the measured energy levels of the atom or from *ab initio* theoretical computations of the structure of the atom, using complex quantum mechanical programs. Systematic errors in f values are notorious as a major cause of errors in abundance measurements. Moreover, the lack of a comprehensive and accurate database of f values can thwart even the most sophisticated non-LTE computations of level populations.

The factor $1 - \mathrm{e}^{hc/\lambda kT}$ in equation (1) corrects for radiatively stimulated emission of photons in the line. Here h is the Planck constant, k is Boltzmann's constant, c is the velocity of light and T is the local gas kinetic temperature.

The function $\phi_\lambda(x, \Delta\lambda)$ in equation (1) describes the microscopic mechanisms which give the spectral line its finite width and shape under the conditions of temperature and density at depth x. These can affect the observed strength of the line (owing to saturation of the radiation available to be absorbed at wavelength $\lambda + \Delta\lambda$). Line broadening arises from a number of processes: (1) Doppler shifts resulting from thermal motion and small-scale turbulence of the gas; (2) the finite energy spread of the upper and lower states of the transition (a result of the Heisenberg uncertainty principle); (3) perturbations of the energy states due to the electrostatic fields of particles in the surrounding gas (Stark and van der Waals broadening); (4) shifting or splitting of the energy levels by magnetic fields (Zeeman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

effect) or by nuclear effects (hyperfine splitting and isotope shifts). Macroscopic mechanisms which also alter the width and shape of the line, but without changing its total strength, include large-scale mass motions (turbulence, convection and mass loss) and stellar rotation. When spectral lines become very strong (highly saturated or black at their centers), they become less sensitive to abundance variations (curve-of-growth effect) and more sensitive in their strength to the function $\phi(x, \Delta\lambda)$. Thus, strong lines present significant challenges to the determination of accurate abundances.

Light emitted at depth x at an angle θ to the surface normal of the stellar atmosphere is attenuated along its path to the surface by a factor $\exp(-\tau_\lambda/\cos\theta)$, where the dimensionless optical depth τ_λ is given by

$$\tau_\lambda = \int_0^x (l_\lambda + \kappa_\lambda)\rho(x)\,\mathrm{d}x \qquad (2)$$

and $\rho(x)$ is the density of stellar material at depth x. The radiation emerges from a unit area of the stellar surface with intensity

$$I_\lambda(x=0,\theta) = \int_0^\infty S_\lambda(\tau_\lambda)\mathrm{e}^{-\tau_\lambda/\cos\theta}\,\mathrm{d}\tau_\lambda/\cos\theta. \qquad (3)$$

The function S_λ is called the 'source function' and is defined as the ratio of energy emitted per unit mass, time, wavelength and solid angle (the emission coefficient) at optical depth τ_λ to the total absorption coefficient $l_\lambda + \kappa_\lambda$. Qualitatively, equation (3) simply states that radiation emitted from greater depths in the atmosphere undergoes greater attenuation before it emerges at the star's surface. In LTE the source function is simply the Planck function describing black body emission at temperature T,

$$S_\lambda = B_\lambda(T) = (2hc^2/\lambda^5)(\mathrm{e}^{hc/\lambda kT} - 1)^{-1}. \qquad (4)$$

The Sun is the only star whose surface is seen as an extended disk which can be spectroscopically observed over the range of angles θ from its center ($\theta = 0^\circ$) to its limb ($\theta = 90^\circ$). Thus, equation (3) is directly applicable in modeling the observed intensity of sunlight at a particular point on the solar disk. For all other stars the relevant quantity is the integrated intensity over the hemisphere of the star facing toward Earth. What is observed at wavelength λ, in the spectrum of a star of radius R at distance d from the Earth, is the flux

$$f_\lambda = 2\pi(R^2/d^2)\int_0^\pi I_\lambda(0,\theta)\,\cos\theta\,\sin\theta\,\mathrm{d}\theta. \qquad (5)$$

The quantity $2\pi R^2/d^2$ may be treated as a scaling factor, which is a byproduct of fitting the observed flux to the integrated emitted intensity computed from a model stellar atmosphere.

Figure 1 gives an example of the process outlined above. Here, we have fitted the observed absorption line spectrum of a chemically peculiar B-type main-sequence star with a theoretically modeled synthetic spectrum in order to derive the abundance of the very heavy element thallium (atomic number 81). In this case the best fit to the spectrum is obtained assuming that thallium is in the form of isotopically pure ^{205}Tl, with an abundance of 4.66. By convention chemical abundances are expressed relative to hydrogen on a logarithmic scale, $\log(N_{\mathrm{el}}/N_{\mathrm{H}}) + 12$. In this star's atmosphere the abundance of thallium exceeds that in the sun by 3.76 in these logarithmic units, i.e. by a factor of $10^{3.76} \approx 5700$.

The Sun as a reference standard for chemical abundances

The abundances of the elements in the Sun and surrounding solar system are more extensively and reliably known than in any other star. Spectra of the SOLAR PHOTOSPHERE are observed with very high spectral resolution and signal-to-noise ratio because of the great brightness of the source, allowing the profiles of a multitude of weak or blended absorption lines to be accurately measured. Moreover, the structure and physical processes of the solar photosphere are well understood not only from theoretical models but also from empirical observations (e.g. the observed variations in the intensity of sunlight from the center to the limb of the solar disk). Photospheric SOLAR ABUNDANCES have been measured for approximately 60 chemical elements, and of these 43 are accurate to 25% or better—a small error bar, considering that the range of abundances spans 12 orders of magnitude.

For the Sun to serve as a useful baseline against which the photospheric abundances of other stars may be compared, it is important to establish (1) that the Sun's surface composition is essentially the same as the composition of the interstellar gas and dust from which it formed (i.e. that it has not been altered by processes endemic to the Sun itself) and (2) that the Sun's photospheric abundances are, at least on average, representative of other stars formed at about the same time in the same region of the Galaxy.

The composition of the solar photosphere is in close agreement with the composition of a rare class of meteorites, type CI carbonaceous CHONDRITES, for those elements for which accurate abundances have been measured in both, as summarized in table 1. The primary exceptions are Li, Be and B which are transported by convection into the solar interior and depleted by nuclear burning at relatively low temperatures. The meteoritic abundances can be measured accurately in the laboratory, providing an important calibration of the accuracy of astrophysical abundance measurements. For the small number of heavy elements for which substantial discrepancies exist between meteoritic and photospheric measurements, it is the photospheric abundances that are highly uncertain.

Radiometric dating of chondritic meteorites indicates that they are as old as the solar system itself (4.55 Gyr). The carbonaceous chondrites contain a relatively high

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

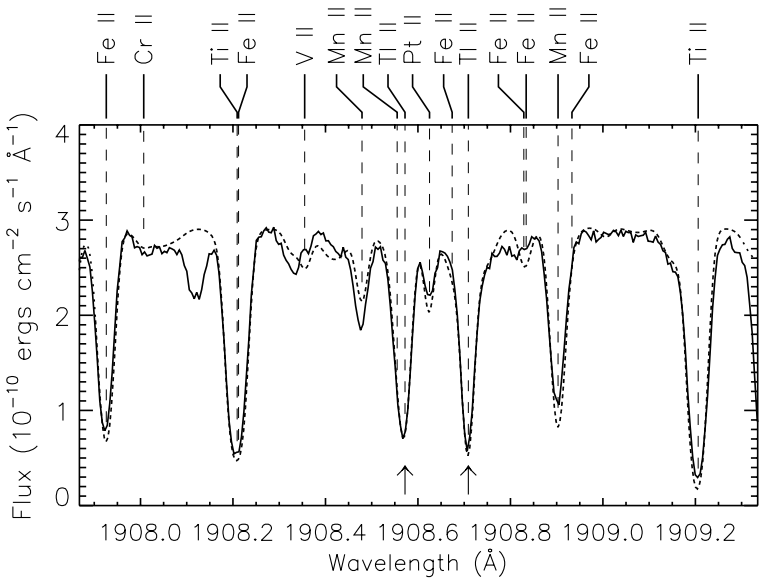


Figure 1. A small portion of the ultraviolet spectrum of the chemically peculiar star χ Lupi, observed with the Goddard High Resolution Spectrograph on board the Hubble Space Telescope (solid curve). Superimposed (dashed curve) is a theoretical spectrum in which the abundance of the thallium isotope ^{205}Tl has been adjusted to fit the two observed absorption lines of singly ionized thallium (Tl II) marked with arrows.

proportion of volatile elements and are believed to have been formed at great distance from the Sun. The close agreement between the composition of CI meteorites and the solar photosphere is strong evidence that the chemical composition of the Sun is (with a few exceptions) the same as that of the pre-solar nebula.

Precise measurements of the chemical compositions of main-sequence stars that bracket the Sun in spectral type (the Sun is type G2 V) indicate that the solar composition is typical for stars similar to the Sun in age, 'metallicity' (as indicated by their iron abundance) and distance of formation from the Galactic center. This is true at least for the dozen or so elements whose abundances have been measured from optical spectra in large statistical surveys.

The abundances of the elements in the solar system are often characterized as 'cosmic abundances' or 'normal abundances'. These terms are overly simplified. However, the solar system does provide a representative sampling of the abundances for a particular time and place in Galactic history, which with few exceptions remain unaltered by processes at work in the Sun itself.

Variations of abundances in time and space

The Galaxy is heavily populated with low-mass stars like the Sun. The theoretical lifetimes of such stars, from the zero-age main sequence to the tip of the red-giant branch, range from about 4.5 Gyr for an F5 dwarf of solar composition (1.3 solar masses) to about 29 Gyr for a K0 dwarf (0.8 solar masses). The Sun's lifetime is approximately 12 Gyr. Consequently, the ages of stars in this large and widely distributed population span the entire history of the Galaxy. Collectively their surface chemical compositions are an encyclopedia of Galactic evolution. In contrast the more massive O-, B- and A-type stars ($\geq$3 solar masses), which are hotter and shine more brightly, have short main-sequence lifetimes $\sim 10^6$–10^8 years. These young stars portray the *present-day* composition and state of evolution of Galactic material.

The composition of an individual star is dictated by the environment in which the star formed. This in turn is set by the history of stellar nucleosynthesis up to that time, by the processes which lead to the ejection of nuclides into the interstellar medium, by the relative amounts of time required for those processes to occur and by the dynamics of the Galaxy, which mixes and homogenizes interstellar material. Although a large variety of nuclear reaction chains occur in stellar interiors, there are only a few that ultimately lead to significant chemical enrichment of the interstellar medium.

There now exist plausible scenarios to explain the astrophysical production of essentially all the chemical elements, thanks to major advances in stellar evolution theory coupled with a large body of modern observational stellar spectroscopy. These are described in the following summary. It is a large step, however, from a plausible scenario to conclusive observational proof, and in some cases considerable work remains to be done.

The most abundant ISOTOPES of elements heavier than boron are thought to be produced primarily in one or more of the following sites: (1) type II supernovae (SNII), (2) type Ia supernovae (SNIa) and (3) highly evolved low- and intermediate-mass stars (1–7 solar masses). Type II supernovae are massive stars (>7–8 solar masses) that end their lives in titanic explosions. Type Ia supernovae are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Measured abundances in the Sun and in meteorites.

Element	Photosphere	Meteorites	Photosphere −meteorites	Element	Photosphere	Meteorites	Photosphere −meteorites
01 H	12.00			44 Ru	1.84 ± 0.07	1.83 ± 0.04	+0.01
02 He	$[10.99 \pm 0.035]$			45 Rh	1.12 ± 0.12	1.10 ± 0.08	+0.02
03 Li	1.16 ± 0.10	3.31 ± 0.04	−2.15	46 Pd	1.69 ± 0.04	1.70 ± 0.04	−0.01
04 Be	1.15 ± 0.10	1.42 ± 0.04	−0.27	47 Ag	(0.94 ± 0.25)	1.24 ± 0.04	(−0.30)
05 B	(2.6 ± 0.3)	2.79 ± 0.05	(−0.19)	48 Cd	1.77 ± 0.11	1.76 ± 0.04	+0.01
06 C	8.55 ± 0.05			49 In	(1.66 ± 0.15)	0.82 ± 0.04	(+0.84)
07 N	7.97 ± 0.07			50 Sn	$2.0 \pm (0.3)$	2.14 ± 0.04	−0.14
08 O	8.87 ± 0.07			51 Sb	$1.0 \pm (0.3)$	1.03 ± 0.04	−0.03
09 F	$[4.56 \pm 0.3]$	4.48 ± 0.06	+0.08	52 Te		2.24 ± 0.04	
10 Ne	$[8.08 \pm 0.06]$			53 I		1.51 ± 0.08	
11 Na	6.33 ± 0.03	6.32 ± 0.02	+0.01	54 Xe		2.23 ± 0.08	
12 Mg	7.58 ± 0.05	7.58 ± 0.01	0.00	55 Cs		1.13 ± 0.02	
13 Al	6.47 ± 0.07	6.49 ± 0.01	−0.02	56 Ba	2.13 ± 0.05	2.22 ± 0.02	−0.09
14 Si	7.55 ± 0.05	7.56 ± 0.01	−0.01	57 La	1.17 ± 0.07	1.22 ± 0.02	−0.05
15 P	$5.45 \pm (0.04)$	5.53 ± 0.04	−0.08	58 Ce	1.58 ± 0.09	1.63 ± 0.02	−0.05
16 S	7.33 ± 0.11	7.20 ± 0.04	+0.13	59 Pr	0.71 ± 0.08	0.80 ± 0.04	−0.09
17 Cl	$[5.5 \pm 0.3]$	5.28 ± 0.06	+0.22	60 Nd	1.50 ± 0.06	1.49 ± 0.02	+0.01
18 Ar	$[6.52 \pm 0.10]$			62 Sm	1.01 ± 0.06	0.98 ± 0.02	+0.03
19 K	5.12 ± 0.13	5.13 ± 0.02	−0.01	63 Eu	0.51 ± 0.08	0.55 ± 0.02	−0.04
20 Ca	6.36 ± 0.02	6.35 ± 0.01	+0.01	64 Gd	1.12 ± 0.04	1.09 ± 0.02	+0.03
21 Sc	3.17 ± 0.10	3.10 ± 0.01	+0.07	65 Tb	(-0.1 ± 0.3)	0.35 ± 0.04	(−0.45)
22 Ti	5.02 ± 0.06	4.94 ± 0.02	+0.08	66 Dy	1.14 ± 0.08	1.17 ± 0.02	−0.03
23 V	4.00 ± 0.02	4.02 ± 0.02	−0.02	67 Ho	(0.26 ± 0.16)	0.51 ± 0.04	(−0.25)
24 Cr	5.67 ± 0.03	5.69 ± 0.01	−0.02	68 Er	0.93 ± 0.06	0.97 ± 0.02	−0.04
25 Mn	5.39 ± 0.03	5.53 ± 0.01	−0.14	69 Tm	(0.00 ± 0.15)	0.15 ± 0.04	(−0.15)
26 Fe	7.50 ± 0.04	7.50 ± 0.01	0.00	70 Yb	$1.08 \pm (0.15)$	0.96 ± 0.02	+0.12
27 Co	4.92 ± 0.04	4.91 ± 0.01	+0.01	71 Lu	(0.76 ± 0.30)	0.13 ± 0.02	(+0.63)
28 Ni	6.25 ± 0.01	6.25 ± 0.01	0.00	72 Hf	$0.88 \pm (0.08)$	0.75 ± 0.02	+0.13
29 Cu	4.21 ± 0.04	4.29 ± 0.04	−0.08	73 Ta		-0.13 ± 0.04	
30 Zn	4.60 ± 0.08	4.67 ± 0.04	−0.07	74 W	(1.11 ± 0.15)	0.69 ± 0.03	(+0.42)
31 Ga	$2.88 \pm (0.10)$	3.13 ± 0.02	−0.25	75 Re		0.28 ± 0.03	
32 Ge	3.41 ± 0.14	3.63 ± 0.04	−0.22	76 Os	1.45 ± 0.10	1.39 ± 0.02	+0.06
33 As		2.37 ± 0.02		77 Ir	$1.35 \pm (0.10)$	1.37 ± 0.01	−0.02
34 Se		3.38 ± 0.02		78 Pt	1.8 ± 0.3	1.69 ± 0.04	+0.11
35 Br		2.63 ± 0.04		79 Au	(1.01 ± 0.15)	0.87 ± 0.02	(+0.14)
36 Kr		3.23 ± 0.07		80 Hg		1.17 ± 0.08	
37 Rb	$2.60 \pm (0.15)$	2.41 ± 0.02	+0.19	81 Tl	(0.9 ± 0.2)	0.83 ± 0.04	(+0.07)
38 Sr	2.97 ± 0.07	2.92 ± 0.02	+0.05	82 Pb	1.95 ± 0.08	2.06 ± 0.04	−0.11
39 Y	2.24 ± 0.03	2.23 ± 0.02	+0.01	83 Bi		0.71 ± 0.06	
40 Zr	2.60 ± 0.02	2.61 ± 0.02	−0.01	90 Th		0.09 ± 0.02	
41 Nb	1.42 ± 0.06	1.40 ± 0.02	+0.02	92 U	(< -0.47)	-0.50 ± 0.04	
42 Mo	1.92 ± 0.05	1.97 ± 0.02	−0.05				

Abundances are in the form $\log(N_{\rm el}/N_{\rm H}) + 12.00$, where N_i is the abundance by number of atoms per gram of material. Values in square brackets are best estimates from non-photospheric data (e.g. solar wind or corona). Values in parentheses are uncertain. Taken from Grevesse N, Noels A and Sauval A J 1996 Standard abundances *Cosmic Abundances (ASP Conf. Ser. 99)* pp 117–26, reproduced courtesy of the Astronomical Society of the Pacific Conference Series.

carbon–oxygen (CO) white dwarfs in close binary systems that explode in nuclear deflagration of carbon following mass transfer from or merger with the binary companion. The envelopes of low- and intermediate-mass stars are chemically enriched by an orderly sequence of interior nuclear processing during their lifetimes. These envelopes are shed as stellar winds and planetary nebulae as the stars evolve off the asymptotic giant branch (AGB).

The more massive AGB stars are believed to yield most of the carbon and nitrogen in the Galaxy. Oxygen and the 'α elements', whose most abundant isotopes are integral multiples of the helium nucleus (Mg, Si, S, Ar, Ca and Ti), are produced predominately by SNII. The 'odd-Z' light elements, Na and Al, evidently originate in the more massive and metal-rich SNII. Elements of the 'iron peak' (V, Cr, Mn, Fe, Co, Ni) come from both types of supernovae, but SNIa give a substantially higher yield of these elements.

Elements heavier than the iron peak arise from the successive capture of neutrons. In NUCLEAR PROCESSING during helium-shell thermal pulses in the interiors of low-mass (1–3 solar mass) AGB stars, neutron captures

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

occur steadily but infrequently, allowing time for beta decay of the neutron to occur before subsequent neutron captures. This is the 's- (slow) process'. S-process elements, whose most abundant isotopes are created by this process, include Sr, Y, Zr, Nb, Ba, La, Ce and Nd.

Explosive events such as supernovae are characterized by high neutron fluxes. 'R- (rapid) process' elements are created in such environments when the capture of neutrons occurs so rapidly that beta decay cannot take place prior to subsequent neutron captures. The sudden termination of neutron captures at the completion of the explosive event yields a pattern of abundances in the ejected material rich in beta-unstable high-neutron-number isotopes, which subsequently decay to more stable nuclides. As measured in the solar system, approximately 2/3 of the 49 stable elements heavier than Ge consist purely or predominantly of r-process isotopes. However, these are very-low-abundance, trace constituents of solar material. The pattern of *relative* abundances among the solar r-process elements appears nearly identical to that observed in very old, metal-poor stars in the halo of the Galaxy. This suggests that all r-process nuclides originate from a single site, most probably from SNII.

Figure 2 provides an example of how stellar chemical abundances are used to probe Galactic evolution. By convention, an abundance ratio given in square brackets is the logarithmic abundance value relative to the abundance in the solar system, $[A/B] = \log (N_A/N_B)_{star} - \log (N_A/N_B)_{sun}$. Between metallicity [Fe/H] $= -1$ and -4, corresponding to old stars in the Galactic halo, the measured abundance of α elements relative to iron exceeds the solar value on average by about +0.4 dex. There is a large star-to-star scatter which is partly real and partly measurement error. However, for the generally younger disk-population stars with metallicity [Fe/H] $= -1$ to 0 or higher, [α/Fe] drops steadily toward the solar value as metallicity increases. This pattern is interpreted as resulting from the time lag between the rapid evolution (<0.02 Gyr) of massive stars to become SNII and the much slower evolution (0.1–10 Gyr) of the progenitors of CO white dwarfs to become SNIa. The SNII produce high yields of the α elements and small amounts of iron. A burst of massive star formation early in the Galaxy's life would result soon thereafter in the ejection by SNIIs of material relatively rich in α elements into the interstellar medium, providing the raw material for the formation of low-metallicity, enhanced-α stars. In roughly a few hundred million years the first SNIa would have evolved from their 7–8 solar mass progenitors, yielding protostellar material significantly enriched in iron-peak elements, but with little additional production of the α elements. At this point the [α/Fe] ratio would begin to decrease, and this trend would continue as progressively lower-mass SNIa progenitors reached the endpoint of their evolution. This scenario is supported by stellar abundance measurements, such as those plotted in figure 2. The r-process elements follow a trend with [Fe/H] somewhat similar to the α elements. It is this evidence that suggests that SNII may also be responsible for their production.

In the innermost 7 kpc of the Galactic disk the oldest, most metal-deficient stars ([Fe/H] $= -1$ to -0.5) have larger enhancements of [α/H] than do their counterparts at larger distances from the Galactic center. Also, the enhancement of α-element abundances persists to higher values of metallicity in the stars closer to the Galactic center. This suggests that STAR FORMATION, particularly of massive stars that become SNII, probably began earlier and has proceeded at a greater rate in the inner part of the Galactic disk.

There is considerable evidence that the chemical composition of the Galactic disk has not evolved homogeneously. For example, it is very difficult to discern a clear relationship between [Fe/H] and stellar age in the solar neighborhood, although such a relationship would be expected from simple models of Galactic evolution. The youngest stars have [Fe/H] ~ 0 and the oldest have [Fe/H] ~ -1. However, there is a large and real scatter in metallicity at all ages in between. The natural variation in iron abundances among stars of the same age is estimated to be about $\pm$25–50% (1σ). The causes of this dispersion are not fully understood. It is clear that metal enrichment of the interstellar medium from which stars form is not only a function of time but of location. Important factors may include (1) local variations in initial mass function and star formation rates, (2) episodic infall of low-metallicity material (which can dilute the interstellar gas) onto localized regions of the disk and (3) the rate at which the Galactic disk becomes well mixed with gas recycled through stellar nucleosynthesis.

The stars of the Galactic halo are metal deficient, averaging about -1.6 in [Fe/H]. The current record for the most metal deficient star observed is held by the halo giant CD $-38\,245$ with [Fe/H] $= -4.01$. Intensive searches for zero-metallicity stars in the Galaxy have been unsuccessful. This suggests that the earliest epoch of nucleosynthesis may have occurred prior to the formation of the Galaxy itself, in protogalactic substructures.

The halo stars contain a significantly higher proportion of r-process elements than the Sun. For very old stars with low metallicity, [Fe/H] < -2.5, the dispersion from star to star in the abundances of r-process elements is much larger than for the presumably younger stars of higher metallicity. It is likely that the early chemical evolution of the halo was dominated by a relatively small number of SNIIs, with a wide range of yields of heavy elements. The composition of the halo interstellar material was very inhomogeneous early in the life of the Galaxy but became less so as time progressed and the ejecta of discrete SNII events became more thoroughly mixed.

Surface abundances as a window into the stellar interior

Stars spend most of their lives on the hydrogen-burning main sequence. In this phase the chemical composition of their surfaces, with few exceptions, offers no hint of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

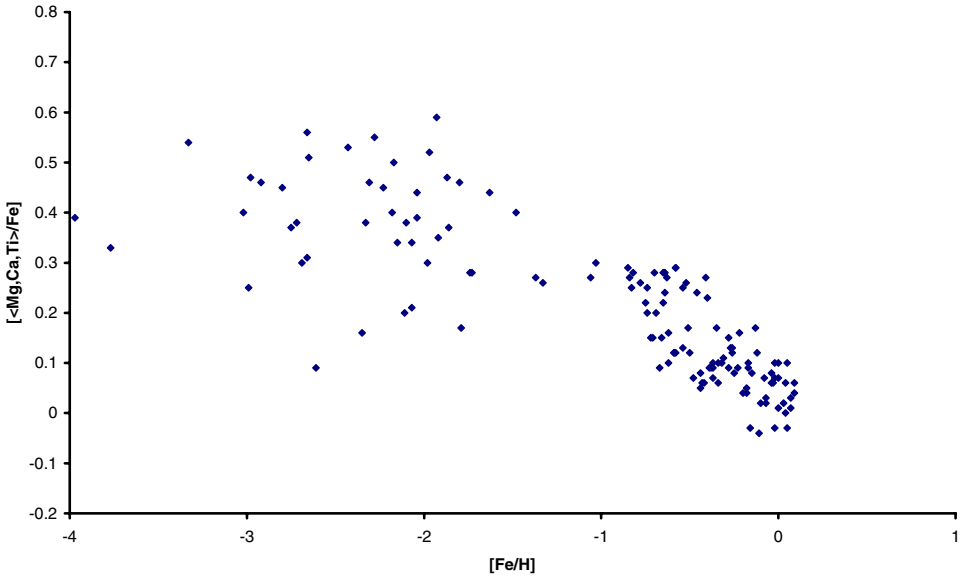


Figure 2. Each point represents, for an individual star, the average measured abundance of three 'α elements' (magnesium, calcium and titanium) relative to iron, plotted as a function of the measured abundance of iron itself. All abundances are in logarithmic units, relative to the Sun—the coordinates (0,0) correspond to solar abundances. Stars with [Fe/H] < -1 are found in the Galactic halo and those with [Fe/H] $\ll -1$ are extremely old. Stars with [Fe/H] > -1 are found in the disk of the Galaxy. The inflection near [Fe/H] $= -1$ is thought to represent the onset of major iron production in type Ia supernovae. (Data are taken from Nissen P E *et al* 1994 *Astron. Astrophys.* **285** 440–50, Edvardsson B *et al* 1993 *Astron. Astrophys.* **275** 101–52, Norris J E, Peterson R C and Beers T C 1993 *Astrophys. J.* **415** 797–810, Magain P 1989 *Astron. Astrophys.* **209** 211–25 and Gratton R G and Sneden C 1988 *Astron. Astrophys.* **204** 193–218.)

the nuclear transmutations taking place deep within their interiors. However, when a star exhausts the supply of hydrogen in its central core it undergoes relatively rapid and dynamic readjustments of its structure associated with the various phases of post-main-sequence evolution. The products of nuclear burning may then be dredged to the surface by convection (possibly assisted by rotationally induced circulation) or exposed to view by copious mass loss or both. Thus, the observed surface compositions of evolved stars provide important empirical underpinnings for the theory of STELLAR EVOLUTION.

Modifications of surface composition in this manner can take several forms: (1) changes in the relative abundances of carbon, nitrogen and oxygen and the carbon isotopes which mark the burning of H into He by the CNO cycle; (2) enhancements of the absolute abundance of carbon produced as a product of the 'triple-alpha' reaction which converts He to C; (3) enhancements of s-process elements created by the capture of neutrons released, for example, during the conversion by the 'alpha process' of ^{14}N into ^{25}Mg and of ^{13}C into ^{16}O; (4) depletion of hydrogen and enrichment of carbon, nitrogen, oxygen and s-process elements as mass loss strips away the hydrogen-rich stellar envelope. In a binary system mass transfer from a highly evolved star to a lower-mass companion can result in similar abundance modifications of the latter's atmosphere, which then displays abundance anomalies apparently at odds with its less-evolved state.

The terminology of spectral classification of evolved stars with 'anomalous' surface abundances (e.g. stars of types C, R, S, N, Ba, CH, HdC, RCrB, OBN, OBC, WN, WC, WO) is rooted in the history of observational stellar spectroscopy going back more than a century. It conveys no information about the progression of stellar evolution, as it is understood today, which has as its logical byproduct the development of these abundance 'anomalies'. This modern framework is well illustrated by the projected future course of evolution of the Sun. In broad outline the evolutionary scenario for the sun is similar to that of all stars up to about 7–8 solar masses. The major distinction among stars in this mass range relates to the manner in which they begin to burn helium into carbon in their central cores late in life—explosively in the case of stars less than about 2.5 solar masses, quiescently for the more massive stars.

After exhaustion of the Sun's central hydrogen fuel supply, its helium core will contract, hydrogen shell burning will commence, envelope expansion will occur and the Sun will evolve away from the main sequence and up the red-giant branch (RGB). In the initial stages of evolution up the RGB the Sun's convective envelope will extend deeply into the interior and, in this 'first dredge-up phase', will bring to the surface the products of CN nuclear processing which occurred while the Sun was on the main sequence. The solar photosphere will become enhanced in nitrogen and reduced in carbon, with a decrease of the ^{12}C/^{13}C ratio. Such changes are observed on the surfaces of metal-rich field stars like the Sun as they begin their ascent up the RGB in good agreement with the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

predictions of standard stellar evolution theory. Farther up the RGB, however, the Sun may exhibit additional abundance modifications of the same type, indicating dredge-up of CN-processed material from deep layers near the hydrogen-burning shell. Bright metal-poor RGB stars in GLOBULAR CLUSTERS can also show the products of ON-processing and proton-capture reactions at the stellar surface. Enhanced nitrogen, depleted oxygen, very low $^{12}C/^{13}C$ values (3–4), and abnormal abundances of sodium, aluminum and magnesium are all observed. For both metal-rich field RGB stars and bright metal-poor globular cluster RGB stars this extra mixing with deep interior regions is not predicted by basic theoretical models of the stellar interior. Standard models for these more evolved RGB stars contain an extended radiative region between the hydrogen-burning shell and the convective envelope that prevents the two from coming into contact. The resolution of this quandary is believed to reside in the influence of stellar rotation, which creates circulation currents capable of transporting CN- and ON-processed material upward to the base of the convection zone (see also STELLAR ROTATION ON THE MAIN SEQUENCE). This is a good example of the insight surface abundances can provide into otherwise undetectable processes hidden from view within the stellar interior. In general the observed surface abundances of RGB stars appear to depend in a complex (and often puzzling) manner on whether the stars are metal rich or metal poor, and whether they are found in globular clusters or in the field. 'Environmental pollution' from the ejecta of highly evolved stars in the near vicinity, particularly in the crowded stellar neighborhoods within globular clusters, is another process that is sometimes evoked to explain such complexities.

At the tip of the RGB, the Sun will achieve a high temperature in its degenerate central core, inducing 'helium-flash' ignition of the triple-alpha reaction that converts helium to carbon. The helium-burning core will join the outward-moving hydrogen-burning shell as the sources of the Sun's luminosity on the 'zero-age' horizontal branch. Subsequently, as the fuel supply of the helium core is depleted, the Sun will advance into a phase characterized by nuclear burning alternately within a helium shell and a hydrogen shell. During this 'two-shell' period its envelope will once again expand as it evolves up the AGB. Periodic thermal flashes associated with ignition of the helium shell will produce a plentiful supply of carbon and also of s-process elements, which will be dredged upward through the convective envelope to the surface. The Sun may first become an S star, with atmospheric enhancements primarily of s-process elements. At this stage its spectrum may even reveal technetium (^{99}Tc), a radioactive s-process element with such a short half-life (2×10^5 yr) that its presence essentially proves that nuclear-processed material has been dredged up from the deep interior very recently. In due course, farther up the AGB, the Sun will become a classical carbon or N-type star, whose spectrum is characterized by strong bands of CH, CN, C_2 and C_3. Ultimately, much of the Sun's hydrogen-rich envelope will be ejected as a PLANETARY NEBULA, leaving behind a WHITE DWARF core composed of carbon and oxygen, probably overlaid with thin surface strata of hydrogen on top of helium, segregated by gravitational settling. If the Sun had a less massive binary companion, its atmosphere could be polluted by the Sun's ejected envelope and acquire the anomalous abundances associated, for example, with barium (Ba) stars.

For stars more massive than about 7–8 solar masses, evolution induces surface abundance changes primarily as a result of large-scale mass loss in strong stellar winds. The WOLF–RAYET (WR) STARS are good examples. They are classified as a spectroscopic sequence: WN stars with enhanced nitrogen and helium, WC stars enhanced in carbon and WO stars enriched in oxygen. This ordering reflects progressively deeper exposure of the stellar core by mass loss, revealing the products first of hydrogen burning, then of helium burning and finally of carbon burning.

Atmospheric abundance anomalies in main-sequence stars

For a century stellar spectroscopists have been intrigued and bemused by the main-sequence stars classified as Ap ('p' denotes 'peculiar'), Bp, Am ('m' stands for 'metallic line'), Si, SrEuCr, HgMn or more generally CP (for 'chemically peculiar'). The atmospheres of these stars display astounding departures from solar abundances in many elements, including enhancements of heavy elements such as Hg by factors ranging up to 10^5–10^6. Large anomalies in the isotopic mixture of some elements are also occasionally observed (see, e.g., figure 1). There are no known nuclear processes capable of explaining these anomalies. The most important clue to their interpretation is the location of these stars in the HERTZSPRUNG–RUSSELL DIAGRAM. Collectively, they constitute 10–20% of all main-sequence stars with effective temperatures between approximately 9000 K and 16 000 K. Stars in this temperature range are too hot to possess extensive convective envelopes (the thin hydrogen convection zone dies out completely around 10 000 K) and too cool to have significant stellar winds. Many of them possess strong dipolar or quadrupolar magnetic fields. Both magnetic and non-magnetic CP stars tend to be intrinsically slow rotators. This combination of factors implies that the CP stars must possess hydrodynamically very stable atmospheres and envelopes. In such a quiescent environment atoms and ions are susceptible to gravitational settling and radiatively driven diffusion. The competition between gravity and radiation results in some ions drifting downward out of view below the photosphere, while others are being driven upward into the photosphere by radiation pressure transmitted through strong spectral lines or continua at wavelengths at which the star strongly emits flux.

This scenario is qualitatively supported by the observations. For example, the largest abundance enhancements often occur in elements that were originally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

very low in abundance in the material from which the star formed (i.e. virtually all elements heavier than the iron peak), while the initially abundant elements (e.g. He, C, N) tend to be somewhat depleted. The abundant elements saturate the radiation field in their strongest spectral lines so that they absorb relatively little radiation pressure and (in the absence of any other material circulation or mixing) tend to sink. The converse is true of initially trace elements. Some ions (e.g. B II, Zn II) can absorb so much radiation pressure that they are blown entirely away from the star, presaging the onset of the stronger stellar winds at higher effective temperatures. Magnetic field lines also serve to guide the migration of ions under gravity and radiation pressure, producing surface abundance inhomogeneities or 'star spots' on the magnetic CP stars.

Radiatively driven diffusion and gravitational settling of atoms and ions have significance beyond the bizarre abundance patterns of CP stars. Theoretical diffusion models suggest that helium, second only to hydrogen in abundance, has been reduced in the Sun's photosphere by about 10% owing to gravitational settling. These processes may be at work deeper in the interiors of stars in regions of radiative equilibrium, modifying the opacity and mean molecular weight of stellar material.

The only other example of main-sequence stars which exhibit abundance anomalies in their atmospheres (excepting binaries that have accreted material from an evolved companion) is the very massive stars classified as OBN. There is evidence of CNO processing of material seen at the surface of these stars, i.e. enhanced nitrogen and reduced carbon abundances. Since these stars possess radiative envelopes, the most likely mechanism for transport of CNO-processed material from the deep interior is meridional circulation, which is a natural consequence of their normal, rapid rotation.

Bibliography

Cox P A 1989 *The Elements: Their Origin, Abundance and Distribution* (Oxford: Oxford University Press)

Holt S S and Sonneborn G (ed) 1996 *Cosmic Abundances (ASP Conf. Ser. 99)* (San Francisco, CA: Astronomical Society of the Pacific)

McWilliam A 1997 Abundance ratios and galactic chemical evolution *Ann. Rev. Astron. Astrophys.* **35** 503–56

Michaud G and Tutukov A (ed) 1990 *Evolution of Stars: the Photospheric Abundance Connection (IAU Symp. No.145)* (Dordrecht: Kluwer)

Mihalas D 1978 *Stellar Atmospheres* 2nd edn (San Francisco, CA: Freeman)

Smith K C 1996 Chemically peculiar hot stars *Astrophys. Space Sci.* **237** 77–105

David Leckrone

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chicxulub Basin

A terrestrial impact feature about 170 km in diameter centered near the coastal town of Puerto Chicxulub on the northern coast of Mexico's Yucatán Peninsula. The structure is well preserved but not apparent on the surface, being buried beneath about a kilometer of sediments. Its presence was suspected when arc-like formations showed up in plots of surface-gravity variations. Supporting evidence came in the form of microtektites (microscopic glass beads formed from impact-melted surface material) found in Haiti. Confirmation was provided by offshore drilling carried out in a search for new oilfields. The age of the Chicxulub basin is 65 million years, and its size is consistent with the impact of a 10 km diameter asteroid. Such an impact would have had catastrophic consequences for life and the environment. It seems likely that this was the event responsible for the mass extinctions that occurred at the end of the Cretaceous period 65 million years ago, leaving its mark as a sedimentary layer enriched in iridium at what is known to stratigraphists as the K/T boundary.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chinese Astronomical Society

The Chinese Astronomical Society (CAS), founded in 1922, now has eleven professional commissions and six working committees. The total membership of the CAS is 1651, including 272 IAU members. The secretariat is located at the Purple Mountain Observatory.

For further information see http://www.bao.ac.cn/cas/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chinese Astronomy

Astronomy has been a subject of interest to Chinese people since ancient times. As early as the sixteenth century BC, a supernova was recorded on an animal bone used in divination. In ancient China, the main mission for astronomers was to determine the farming seasons and to predict important events for the future according to the phenomena that appeared in the sky. In the minds of rulers in ancient times astronomy had been given special status among the disciplines of natural sciences because of the belief that it helps predict the future, and fate. As a result, astronomers had been appointed as royal officials since the twenty-fourth centre BC. A royal observatory was established in the West Zhou Dynasty (eleventh century BC–771 BC) and continued its existence into the beginning of the twentieth century.

Because in ancient times astronomy was serving the needs of rulers, it was treated as ASTROLOGY to some extent. for divination purposes, observation and predictions of solar and lunar eclipses were given special attention. The conjunctions of bright planets as well as appearances of comets, meteors and new stars were considered to have divinatory value. Therefore the movements of the Sun, Moon and planets, and rare astronomical phenomena were carefully and systematically observed. Thus a huge body of astronomical phenomena were recorded. The unique status of astronomy in ancient China guaranteed a separate astronomical chapter in the historical chronicle and more than 10 000 individual records of solar and lunar eclipses, sunspots, auroras, meteors and meteorites, comets, novae and supernovae, the occultation of planets by the moon and other phenomena were saved. An account of comets in the eleventh century BC and a record of sunspots in 28 BC are considered as the earliest recorded observations of these phenomena in astronomical history. China's records of historical supernovae are the most complete in the world. Among them, the recording of the 1054 supernova has been used to determine the age of the Crab pulsar. The records of eclipses have been applied to calculate the secular acceleration of the Earth's rotation (see also CHINESE ASTRONOMY IN HISTORY).

Astronomy enjoyed great development as a very important tool for making calendars and keeping time. Since the twenty-fourth century BC, a lunisolar calendar system, combining the lunar and solar calendars, was established and this system continues to exist even today in China. The length of a year is determined by the period of apparent motion of the Sun. The length of a month is determined by the synodic period of the Moon. The two calendar systems are coordinated by seven leap months in 19 years.

Since the seventh century BC gnomons were made to determine the time of the summer solstice and winter solstice, so as to determine the length of a year. The length of a year was taken to be 365.25 days in the calendar in use in the fifth century BC. In 1277, Guo Shoujing, a well known astronomer in the Yuan Dynasty, used a horizontal rod to replace the tablet in the gnomon and added a pinhole-imaging device to increase the sharpness of the shadow. The precision acquired by using the improved device resulted in an error of only about several arcseconds of the latitude of the Sun. The precise calculation of length of a year was recorded in a calendar book in 1199. According to this calendar, a year has exactly 365.2425 days. This calculation was confirmed by Guo Shoujing's observation in 1281. This is very close to today's value of the tropical year.

To keep time, a clepsydra (a water clock) was commonly used since the eighth century BC. A water-powered globe was built to show the revolution of celestial sphere by ZHANG HENG, the famous astronomer in the Han dynasty, in the third century. The water globe was improved by Yi Xing and Liang Lingzhan in the year 723, in which the escape wheel was invented. Yi Xing also deduced from his observations that the difference of altitude of the polar star was constant in a fixed distance along the terrestrial meridian. This knowledge was used by him to measure the size of the territory of the country.

From the second century BC to the thirteenth century AD, observatory instruments and observations improved dramatically. To determine the positions of stars and planets, an ARMILLARY SPHERE (an equatorial mounted instrument) was built in the first century BC. This instrument consisted of a polar axis, a sighting tube and, equatorial and horizontal circles with a scale of 365.25 degrees. In the fourth century BC, coordinates of 121 stars were listed in Shi Shen's star catalog. A more detailed catalog that included 283 constellations and 1464 stars had been compiled in the third century AD by Chen Zuo with China's own special constellations and naming system. However, even more remarkable was a star atlas that was stored in Dunhuang grottos, in which 1350 stars and 283 constellations were recorded early in the eighth century. Besides, as a result of detailed observations some laws were discovered. One of them was the discovery of precession by Yu Xi in 330. Careful observations in the fourth century also led to the discovery that comets' tails always point away from the Sun and that comets always reflect the Sun's rays.

Between the eighth century BC and the third century AD, different hypotheses regarding the definition and concept of the universe were actively debated and recorded. Among them, one hypothesis suggested that the Earth and stars float in an unlimited space. Theoretical thought—for instance, discussions on the structure of the universe—was considered to be useless to life and neglected in the Tang dynasty (618–906). Furthermore, people outside the royal observatory were forbidden to study astronomy from the Tang dynasty onwards. Unfortunately, therefore, the imperial management of astronomy restrained the free development of studies of astronomy in ancient China. Precise measurements of the position and movements of planets were not emphasized enough, except their synodic periods, and this absence of precision and the dominance of astrology in ancient China

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

resulted in much slower development of the theory of planetary motion than in Europe. Despite an abundance of many great astronomical discoveries, studies of those observations were not systematic and did not lead to many breakthroughs in astronomical theories. From the fourteenth to fifteenth centuries, the development of astronomy in China slowed down.

At the beginning of the seventeenth century, Mathoeus Ricci and other missionaries came to China from Europe and worked in cooperation with Chinese astronomers. They introduced modern western astronomy to China including the Copernican theory, telescopes and spherical astronomy.

Modern astronomy in China

In 1911, with the ending of China's last dynasty, the Qing Dynasty, the royal observatory was closed. The Chinese Republic later established the Central Observatory in Beijing. In 1958, BEIJING ASTRONOMICAL OBSERVATORY was rebuilt. Zo-Se Observatory was built near Shanghai in 1900, where a 40 cm refractor, the first large telescope used in astronomical study in China, was installed. Zi-Ka-Wei observatory, built in 1872, began to broadcast the time signal in 1914. These two observatories were combined to become the present SHANGHAI ASTRONOMICAL OBSERVATORY. In 1935, the PURPLE MOUNTAIN OBSERVATORY (PMO) was established as an institute of the Academy. Telescopes were installed one after another. In 1975, Yunnan Astronomical Observatory was separated from PMO to become an independent observatory. Shaanxi Observatory was built as the time service center of China in 1966. The majority of astronomers and instruments work in the five observatories managed by the Chinese Academy of Sciences.

The Beijing Astronomical Observatory (BAO) is engaged in astrophysical research including the abundance of the elements, short period ceiphids, supernova surveys and AGNs, starburst galaxies, optical identification of x-ray sources, large-scale structure of galaxies, narrow-band surveys, radio astronomy and solar physics. The optical, radio and solar observing sites are separately located in Xinglong, Miyun and Huairou along the Great Wall. The BAO manages the Beijing Regional Warning Center of the International Space Environment Service. The solar magnetic field of active regions and solar radio eruptions are monitored. Predictions of events of solar activity are provided. A 2.16 m telescope, solar magnetic field telescope, 1.26 m infrared telescope, and meter-wave radio synthesis telescope are installed in BAO. World Data Center-D for Astronomy is also located at the BAO.

The Shanghai Astronomical Observatory (SAO) focuses more on studies of astro-geodynamics and astrometry. VLBI measurements, satellite and lunar laser ranging are used to determine Earth rotation parameters. The variability of the baseline is being interferometrically measured in cooperation with Japanese, German and American observatories. They are also beginning to develop some astrophysical studies in SAO. Open cluster variables in spherical clusters and N-body simulations are the main areas of interest in astrophysical research. Two 25 m radio dishes have been installed in Shanghai and Urumqi stations and are working as VLBI units of EVN. A 1.56 m reflector is operated in SAO. The Eastern Asia–Pacific Space Geodynamics Network was established in SAO in 1994, in which more than 20 VLBI, GPS and SLR stations of about 10 countries participate.

Scientists at the Purple Mountain Observatory (PMO) conduct research on the solar system, molecular astronomy and satellite dynamics. Minor planets, planetary physics, artificial satellite dynamics and astronomical almanacs are traditional research fields. Researches and techniques in millimeter wave astronomy have made tremendous progress in recent years. A 13.6 m mm-wave telescope has been installed in the Qinghai plateau.

The fine structure of the Sun, rotation of stars, close binaries, stellar structure and BL Lac objects are the main fields of research in Yunnan Astronomical Observatory. Scientists here are conducting research into artificial satellite dynamics, astrometry and solar radio eruption. A 1 m telescope and several small telescopes are in use in Kunming, the capital of Yunnan province.

Institutions providing astronomical education report to the Ministry of Education. These institutions include: the Department of Astronomy of Nanjing University, the Department of Astronomy of Beijing Normal University and the Department of Astrophysics of Peking University. Though no observation facilities are available in these universities, scientists and faculties from these universities can use telescopes and observations from different observatories based on the quality of the applications.

The High Energy Astrophysics Center of the Institute for High Energy Physics, the Center for Astrophysics of China Science and Technology University, astrophysics groups in Nanjing Normal University, Jiaotong University and Huazhong Normal University are active in researches on high-energy astrophysics and plasma astrophysics.

In Teipei, the Institute for Astrophysics of the Academia Sinica was built in 1990. This institute adds two dishes to the Sub-Millimeter Array (SMA) and supports and shares 10% of BIMA to cooperate with astronomers in the US. The Institute for Astronomy was established in the Central University at Zhongli, a town south of Teipei; its 70 cm telescope was installed in Mount Yu. A solar oscillation net with five detectors installed in Big Bear, the Canary Islands, Tashkent, Beijing and Teipei has operated in Tsinghua University since 1991. In Hong Kong, astrophysics research groups are actively working in Hong Kong University, Hong Kong Chinese University and Hong Kong Science and Technology University.

The Beijing Planetarium was established as a center for popular astronomy in 1957. The Astronomical Museum in Taipei was built in 1996, in which the exhibition area covered 50 000 m^2. Hong Kong Planetarium was built in 1980. More than 50 small planetaria are scattered across the country.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The CHINESE ASTRONOMICAL SOCIETY (CAS) is a non-government organization for astronomers and scientists. The CAS is an affiliate of the INTERNATIONAL ASTRONOMICAL UNION (IAU). The Astronomical Association, located in Teipei, is an affiliate of the IAU. The Hong Kong Astrophysics Society was established in 1997.

Most of the telescopes used in the observatories of the mainland are made by the Center for Research and Manufacture of Astronomical Instruments in Nanjing in cooperation with the observatories. New initiatives for future astronomical instruments in China have been scheduled. The Large Multi-Object Spectroscopy Telescope (LAMOST) was approved as a national project in the 1990s. LAMOST is an alt-azimuthally mounted reflecting Schmidt telescope. The optical system consists of a 6×6.7 m spherical mirror fixed on the foundation and facing north. An aspheric reflecting mirror lies at the center of the sphere 40 m away. Spherical aberration will be eliminated in real time with the help of active optics. The effective aperture is 4 m. The focal length is 20 m. The field of view is about 5° giving a linear diameter 1.75 m. Due to the large FOV, about 4000 fibers can be attached to the focal surface so the spectra of 4000 objects can be simultaneously observed. The LAMOST project will be completed in 2004.

Financial support for astronomy in China comes mainly from CAS, the Ministry of Education, the Ministry of Science and Technology, the National Natural Sciences Foundation (NNSF) and the National Planning Committee. The CAS and the Ministry of Education provide operational expenses for their affiliate observatories and departments, which amount to US$10 million per year approximately. NNSF supports individuals and groups as well as moderate and small programs of about US$1 million per year. The National Planning Committee supports large projects involving the construction of large scientific instruments. The budget of LAMOST is US$30 million. The Ministry of Sciences and Technology funds the larger research projects, which have budgets of about US$1 million a year.

Qibin Li

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chinese Astronomy in History

This article outlines knowledge of the sky and the celestial motions that originated in Chinese culture from earliest times to 1912, when the use of traditional techniques ended along with the Chinese empire.

The Chinese experience of astronomy for the past two thousand years repays study because, unlike Europe, the level of knowledge and practice remained high without interruption for more than two thousand years and because it was based on styles of work quite different from those prevalent in the West.

Setting

All of the great historical civilizations have exchanged ideas and artifacts since the New Stone Age. China was so isolated from Europe, the Middle East and India that outside scientific influence remained controllable until the mid-19th century. It remained a centrally ruled empire larger than all of Europe for more than 2000 yr. Even in brief periods of division, the continuity of its high culture, unlike that of Europe for much of the Middle Ages, was never interrupted.

Despite China's enormous extent and local diversity, its literate elite, because of their classical educations, shared not only a language but values and patterns of thought. Among these values was overwhelming prestige for the civil service. One generation of talented astronomers after another sought the only employer ready to provide large, long-term investment in scientific research and development.

The government supported large-scale astronomical projects, from the 2nd century BC on, because of the ideology that justified its power. It claimed that the imperial house was chosen by Heaven (that is, the divine order of nature) to shape society in its image. The monarch's personal virtue and charisma, not his managerial skills, kept him in touch with the cosmic order and enabled him to keep the state in harmony with it.

This doctrine included the notion of the Mandate of Heaven. Heaven chose an exceptionally virtuous family to rule. When eventually its descendants no longer maintained that virtue, Heaven transferred the mandate to a worthier family. The conquest that ended a dynasty and began a new one was thus not a crime—if it succeeded.

There was an early warning system. Certain anomalous phenomena in the sky (or, less often, on Earth) signaled that the ruler must reform or be dethroned. Auspicious omens acknowledged a sage reign—or, in different circumstances, encouraged a sage rebel.

Prompt notification was essential. As a result, the imperial court included an amply staffed Bureau of Astronomy. Its responsibilities for ASTROLOGY and mathematical astronomy were intimately related. Its officials scrutinized the sky for omens, recorded them, determined their meaning from the Bureau's archives and reported them so that the emperor and his officials could determine how to respond. Accounts of this process suggest that astrology was important not as a source of objective predictions, right or wrong, but as a ritual form for the open discussion of policy.

Equally important was moving as many phenomena as possible from the realm of the ominous to that of the predictable, an effort that improved computation. For instance, by the late 3rd century AD new moons on the last day of the month had ceased to be omens. Because it was impossible to consistently forecast solar ECLIPSES that would be visible from a given observatory, unpredicted eclipses remained the chief concern of rulers.

The emperor symbolically expressed his control over time as well as space by issuing the annual CALENDAR (actually an ephemeris or almanac that included eclipses and planetary phenomena). He expected the Bureau to incorporate in it the most accurate predictions possible, particularly those of eclipses.

As a result of this ongoing state sponsorship, the published historical record offers unbroken, dated records of most important celestial phenomena, and detailed accounts of the evolving methods used to make EPHEMERIDES, for more than two millennia.

Beginnings

Records written on bone in the 15th century BC include observations of phenomena in the sky. Because these documents come from royal archives, eclipse sightings are prominent and figure as portents. They are dated using a cycle of 60 days, uninterrupted until modern times.

In the 4th and 3rd centuries the warring states used a variety of discrepant calendars. All were based on the tropical year and the lunation. They were generated by counting from an epoch chosen as the beginning of a year, month, 60 day cycle and day. The astronomical year began at the winter solstice, the month at the new Moon—not an observable event—and the day at midnight. Thus the epoch of the Triple Concordance system of 7 BC was exactly set at midnight of sexagenary day 1, 143 127 yr before the winter solstice of 24 December 105 BC.

Since the civil year needed an integral number of days per month and months per year, months of 29 or 30 days roughly alternated, so that a civil year of 12 months would be 354 or 355 days long. Chinese adopted intercalation, at first using what is known in the West as the METONIC CYCLE. They added 7 extra months every 19 yr. That schema spaced intercalary years of 383 or 384 days equally to yield a long-term average of $365\frac{1}{4}$ days per year. Over the centuries, more elaborate schemes of intercalation, and then computation of true lunations, yielded a mean tropical year that by the 13th century AD was 365.2425 days.

Observations

The first clearly instrumental observations measured the shadow cast by an upright gnomon higher than a human. They cannot be dated certainly before the 4th century BC. The longest shadow of the day determined the moment of noon and the north–south direction. Once astronomers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

noticed that the noon shadow varied in length from day to day, it became possible to define the beginning of the year as the day of the longest noon shadow, the winter solstice. At night, sighting along a north–south line on stars passing the gnomon yielded their meridian transits, which were important before the invention of graduated instruments.

The development of graduated armillary instruments with a sighting tube is a vexed topic. The earliest known document that requires the use of a simple dioptra is the *Gan Shi xing jing* (Star canons of Gan De and Shi Shen), the data in which have been dated by error analysis at ca 70 BC. Geng Shouchang used a simple equatorial armillary in 52 BC; an ecliptic ring was added by AD 84 and horizon and meridian rings by 125.

Chinese degree measurement was not derived from the Babylonian sexagesimal system. It was based on the *du*, originally a measure of time intervals between meridian transits, but after about AD 180 redefined as the distance of one day's mean solar travel. There were initially $365\frac{1}{4}$ *du* in a circle, and the number changed with the value for year length. The precision of the first star catalogue was 1/8 *du*; this increased to 0.05 *du* by AD 1279.

The coordinates of observation, as in modern astronomy but unlike those of Europe before the Renaissance, were equatorial, measured from the winter solstice point rather than from the Sun's position at the spring equinox. There is little to choose between the two systems, since the Sun, Moon and planets orbit close to the ecliptic, and the stars rotate parallel to the equator. It was necessary in either case to convert positions along one to positions along the other. Early Chinese accomplished this with simple numerical methods; by the reform of 1279, gradual moves in the direction of spherical trigonometry greatly increased accuracy.

The time of observations was generally measured by water clocks, of which several examples survive from the last two centuries BC. The precision of records was at first ca 1/2 h. Despite great elaboration over the centuries, typical precision by 1300 was perhaps 0.4 h, and significantly better for solar eclipse data.

Coordinates of celestial objects were, as in Europe, measured from markpoints spaced closely enough that observers without instruments could estimate with fair accuracy the distance of objects from them in *du*. Unlike the 12 houses of the zodiac, the Chinese *xiu* (lunar lodges) were a set of 28 unequally spaced stars in prominent constellations. This was similar to the coordinates used in India and called *nakshatra* and those of the Islamic world called *al-manāzil* (see INDIAN ASTRONOMY, ISLAMIC ASTRONOMY). The lunar lodges were in principle as satisfactory as the zodiac, but occasional unintended changes in their determinative stars affected accuracy when new predictive methods were tested against ancient observational records.

Endless polemics have failed to settle where and when the lodges originated. Their 28 names first appear together on a lacquer box buried in 433 BC. Only in sources of the 4th or 3rd century do they record coordinates of observations.

Calendar reforms

The first unification of China, in 221 BC, led to a single calendar, but after a century the system was regularly a day off in predicting new moons. After a reform of the calendar in 104 BC, in 7 BC astronomers expanded the calendar-making techniques into an astronomical system (*li*) that included all predictable phenomena, including eclipses and planetary motions. This expansion was motivated as much by two emperors' desire to renew their political order along lines determined by cosmic religion as by a drive toward technical innovation.

The Triple Concordance system of 7 BC was not revised piecemeal but replaced by a complete new computational system in AD 85. Because integral replacement became the norm, and imperial prestige depended on accurate prediction, 101 systems were proposed before 1912, of which 41 saw official use.

Like Ptolemy's *Almagest*, this Triple Concordance system set out step-by-step procedures for generating the complete ephemeris mentioned above. A clerk with only basic arithmetical skills could carry out the entire process, consulting a built-in table that obviated a complex computation.

Chinese computational methods differed from those of the Greeks in their numerical or algorithmic, rather than geometric, approach. Predicting solar eclipses, which amounts to finding the intersection of a shadow cone with a spherical Sun, is difficult without geometry. That is why it remained a central problem.

Chinese began by counting off mean solar and lunar cycles to find when the two bodies intersected. The results were so poor that astronomers apparently counted off 135 months from an observed eclipse to predict another one, a more successful technique. However, the point of mathematical astronomy was predicting without having to look at the sky. Yang Wei in AD 237 did away with reliance on prior observations by introducing the concept of eclipse limit (in modern terms, the distance of the conjunction from the intersection of the orbits). Techniques for computing apparent rather than mean lunar (AD 223) and solar (ca. 604) motions, and their combination in predicting lunations (619), were extremely important. Elaborate linear interpolation algorithms (fifth order by 1064) and prototrigonometric methods (beginning at about the same time) increasingly approximated the results of geometric analysis.

Development

The bureaucratic character of astronomy did not change markedly over the centuries. Because appointments tended to be sinecures, often hereditary, innovation often came from outsiders—amateur astronomers who had access to information about previous systems published in the dynastic histories. The originator of a new system, after it was thoroughly tested, was often appointed to the Bureau. Some security-minded dynasties forbade the private study of astronomy, cutting off this source of new methods. However, other administrations tested

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

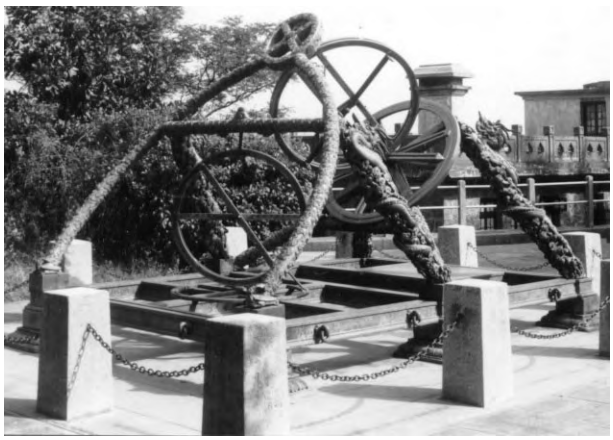


Figure 1. Abridged armillary of ca 1270, now at Purple Mountain Observatory, Nanjing.

officials to raise the level of practice or included technical questions on the general civil service examination to motivate widespread study of astronomy.

Since the government valued astronomical knowledge, and remained the only source of massive support for research, its large projects took the science in several new directions. These included the following

- Observatories were built with a large array of instruments. The huge observational ARMILLARY SPHERE of 1442, reproducing a design of 1074, and the 'abridged armillary' (equatorial torquetrum, *jianyi*), the design of which goes back at least to ca 1270, survive (figure 1).
- Elaborate demonstrational instruments, such as a series of water-driven, escapement-controlled astronomical clocks were constructed. Their high point was the enormous clock that Su Song built ca 1090 (figure 2).
- Large-scale surveys of longitude and latitude were carried out to increase accuracy of data recording and to investigate such matters as the differing visibility of eclipses. That of 1279 covered 6000 miles north–south and 2000 miles east–west.
- Fixed stars were repeatedly catalogued, especially to fix the determinative stars of the lunar lodges. Ten such surveys became the basis of official astronomy between 964 and 1106 (figure 3).

Despite the conservatism inherent in the official focus and stereotyped form of astronomy, the government facilitated improvements in every aspect of the ephemeris. A few important ones not already noted are (in order of the date at which official use began, or otherwise that of writing):

- Precession (literally 'annual difference', *sui cha*, 330), defined as the time difference between the tropical and sidereal year. This constant corresponded to a shift of the winter solstice point 1 *du* per 50 yr. Chinese did not see it as a rotation of the equatorial about the ecliptic pole.
- Exact times of new moons were determined from observations of lunar eclipses (384).
- Sustained investigation of apparent planetary motions was carried out (roughly 729 on). Originally the emphasis on eclipse prediction led to neglect of this problem.
- An approximation to spherical trigonometry applied to the celestial sphere (1279) was combined with a cluster of other changes, such as substituting decimal parts for mixed fractions and discarding the traditional dependence on counting off cycles from an early epoch.

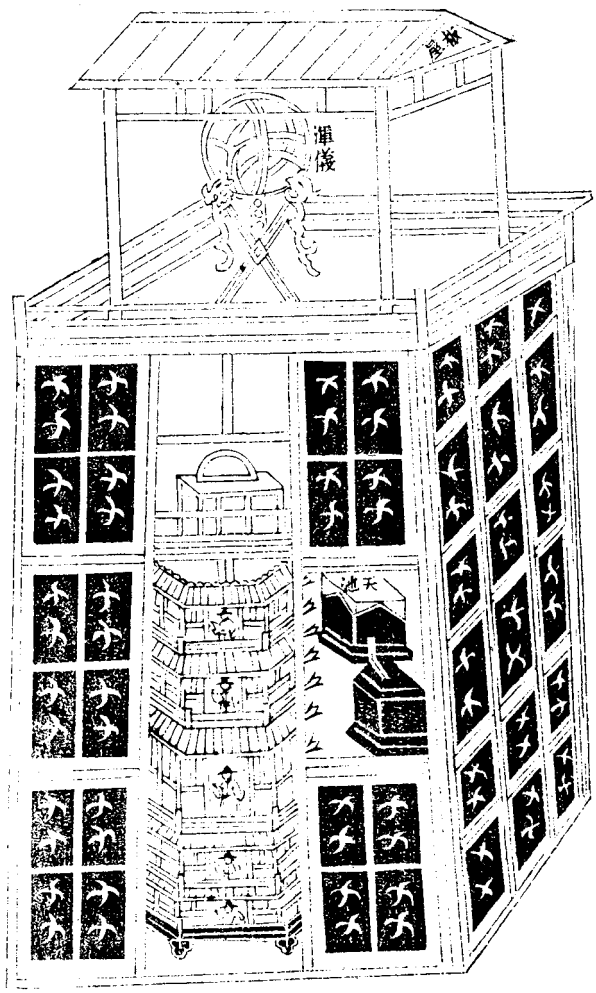


Figure 2. The great astronomical clock of 1089, from the contemporary *Xin yixiang fa yao*.

After the revolution of 1911, the traditional approach was discarded, not because of its particular shortcomings but because of a blanket move toward modernization. After over a century of imperialism and incessant war,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

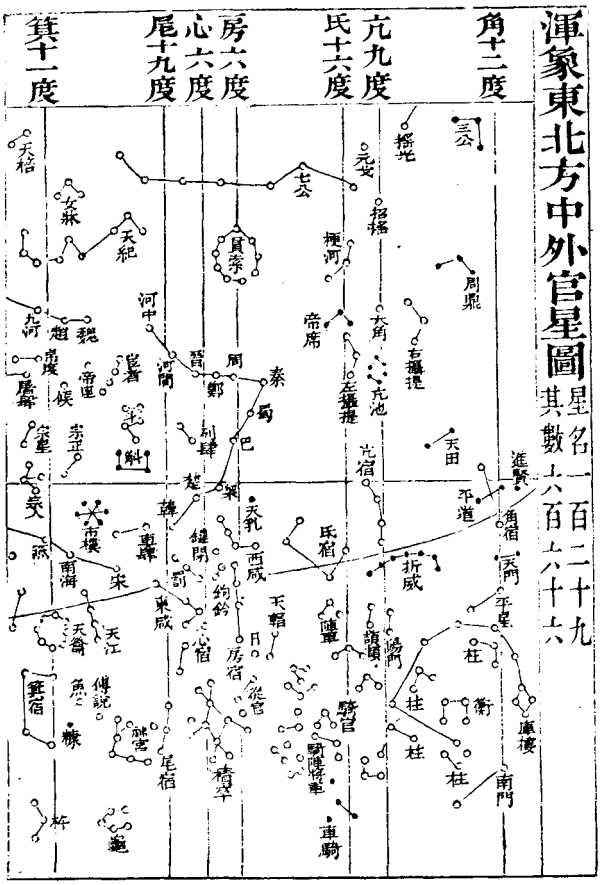


Figure 3. Part of a star map of the late 11th century, from *Xin yixiang fa yao*.

China was too poor to invest massively in modern astronomy at something approaching the world level until after 1975 (see CHINESE ASTRONOMY).

International influence

Since China was the predominant culture in East Asia, the institutions and methods of its official astronomy provided a foundation for counterparts (always with local character) in Korea, Japan and Vietnam.

The desire for accurate prediction of solar eclipses motivated Chinese to hire experts in the geometric methods of the Greek tradition. They employed Indians from 665 on, Muslims during the Mongol occupation (1279–1367) and Jesuit missionaries in the 17th and 18th centuries. The latter, in fact, by offering the Manchu invaders their skills at casting cannon and computing the mandatory imperial ephemerides, took control of the Astronomical Bureau. Even when the Jesuits introduced European techniques, they remolded them into the traditional form of the *li*. Although they mentioned Copernicus and other important European figures in their Chinese writings, the constraints of the Counter-Reformation prevented them from describing revolutionary changes in cosmology.

Applications in modern astronomy

Although narrow-minded positivists ignore the history of astrology, in China the main legacy of traditional to modern science has come not from computational techniques but from detailed astrological records of observations, unequalled elsewhere. The fullest ancient supernova data that have been correlated with pulsar phenomena are generally Chinese. Information on the time, magnitude, and duration of large numbers of solar and lunar eclipses has cast light on the secular acceleration of the Moon and Earth. Dates of every return of Halley's comet for over 2000 yr have clarified its motion. Records, some multiple, on 147 distinct appearances of sunspots and of about twice that many auroras have proved valuable for investigating the relations between the two phenomena and testing hypotheses about sunspot minima.

Bibliography

Chen Z 1980–1984 *Zhongguo tianwenxue shi (History of Chinese Astronomy)* 3 vols (Shanghai: Shanghai Renmin) (the most comprehensive history)

Chuang W and Wang L 1988 *Zhongguo gudai tian xiang jilu zong ji (Ancient Chinese Records of Celestial Phenomena: General)* (Nanjing: Jiangsu Kexue Jishu) (detailed compilation of data from a broad range of primary sources)

Needham J and Wang L 1954 *Science and Civilisation in China* (Cambridge: Cambridge University Press) (vol 3 [1959] is the only comprehensive history in a Western language, but use with care)

Sivin N 1995 *Science in Ancient China* (Aldershot: Variorum) (includes five studies of Chinese astronomy, and an annotated, selected bibliography of Western-language sources)

Stephenson F R and Clark D H 1978 *Applications of Early Astronomical Records* (Bristol: Hilger) (this and Stephenson's other publications draw on Chinese materials)

Nathan Sivin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chiron

Chiron is among the present-day population of minor planets called CENTAURS, whose orbits cross those of the giant outer planets Saturn, Uranus and Neptune. Chiron was the first Centaur discovered, and perhaps is the most unusual due to its cometary activity and other characteristics. Centaurs are probably former residents of the KUIPER BELT, and so Chiron may be considered our first and most accessible emissary from the far reaches of the outer solar system. Studies of Chiron and other Centaurs and Kuiper Belt objects allow us to probe the formation and evolutionary history of the solar system.

Chiron orbits the Sun between Saturn and Uranus on a 50-year orbit, ranging in heliocentric distance from 8.5 AU to just under 19 AU (an AU, or astronomical unit, is the distance between the Earth and the Sun, and is roughly equal to 1.5×10^{13} cm). The specific orbital properties of Chiron are given in table 1. Chiron passed perihelion on 14 February 1996, and was the focus of an international observation campaign at that time.

Table 1. Orbital properties of Chiron.

Period	50.2 years
Eccentricity	0.38
Inclination	6.94°
Perihelion distance	8.443 AU
Aphelion distance	18.765 AU
Semimajor axis	13.633 AU
Argument of perihelion	339.1°
Longitude of ascending node	209.4°

Chiron's diameter is about 180 km; the range of recent measurements taken at various wavelengths is about 150 to 370 km, though occultation and IR observations support the value near 180 km. Chiron has an absolute magnitude of $H = 6.5$ mag; over the range of its orbit, its apparent magnitude varies from 16 to 20 mag. The visible lightcurve is comparable to that in the ultraviolet, with a period of 5.9 h and a peak-to-peak variation of about 0.1 mag.

Measurements at visible wavelengths have shown that Chiron typically has a comparatively neutral, or gray, color (meaning that it does not preferentially reflect or absorb particular wavelengths of light that illuminate it). This gray color is independent of rotational phase, though there have been observations showing some color changes possibly connected to its cometary activity. For example, ultraviolet observations show a similar neutral color during quiescent periods, though during an outburst, Chiron exhibited a possible blue color slope and the ALBEDO (reflectivity) increased.

Some key physical properties of Chiron are given in table 2. The significant digits given in these values are roughly indicative of their uncertainties.

The discovery of Chiron follows an historical progression of discovering new classes of minor planets in regions of the solar system previously thought to be empty. The first asteroid, CERES, was discovered by Piazzi in 1801,

Table 2. Physical properties of Chiron.

Diameter	~180 km
Synodic rotation period	5.91783 h
Lightcurve amplitude	7–10%
Absolute magnitude	6.5
$B - V$ color	0.70
Asteroid spectral type	C
Visual geometric albedo	0.05
Ultraviolet geometric albedo	0.08
Magnitude slope parameter	0.15

and was soon followed by discoveries of PALLAS, JUNO and VESTA over the next six years. As telescopes, detectors and techniques improved, the observational limits of the asteroidal bodies progressed with time from the main belt, moving both outward (through the Cybele, Hilda and TROJAN ASTEROID groups) and inward (through Amor, Apollo and Aten groups).

In 1977 the horizon of the small bodies in our solar system was extended to orbits as distant as Uranus, with the discovery by Charles Kowal of object 1977 UB, later designated as asteroid number 2060, and named Chiron. Chiron implied a new class of small bodies called Centaurs—outer planet crossers. A total of 17 Centaurs had been discovered as of the end of 1999. Fifteen years after the discovery of Chiron, our view of the solar system was expanded further by the first discovery of a Kuiper Belt object, 1992 QB_1, by David Jewitt and Jane Luu. This was the first of hundreds of such objects that now have been found to have orbits beyond Neptune. These observational discoveries convincingly established that the vast outer planet and trans-Neptunian regions are, indeed, teeming with small bodies.

Chiron's connection to the rest of the solar system

Chiron is believed to be an escapee from the Kuiper Belt, a reservoir of COMETS and larger PLANETESIMALS postulated by Edgeworth and Kuiper in the 1940s and 1950s. One primary line of evidence for this origin of Chiron is dynamical. Studies by Shio Oikawa and Edgar Everhart, and more recently by Luke Dones, Harold Levison and Martin Duncan, have shown that Chiron's orbit is unstable due to perturbations by the giant planets on timescales of $<10^6$ years. This, and Chiron's comet-like activity (discussed in the next section), provide strong, yet circumstantial, evidence that Chiron is a recent addition to the planetary region.

Around the same time as the dynamical instability of Chiron's orbit was discovered, orbital evolution simulations first began to provide convincing support for the idea that orbits for most of the short-period comets could not be derived from a distant, spherical OORT CLOUD. As discussed by Martin Duncan, Thomas Quinn and Scott Tremaine in papers in 1988 and 1990, a postulated disk source in the outer solar system (the as-of-yet undiscovered Kuiper Belt) must be the source of such objects. Dynamical simulations have shown that objects in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the Kuiper Belt between 30 and 50 AU exhibit dynamically chaotic orbits due to gravitational perturbations by the giant planets. As such perturbations occur, the perihelion distances of an object in this region can rapidly decrease so that the object's orbit eventually crosses Neptune's orbit. Once this occurs, the object will almost certainly experience a close encounter with Neptune, which then can throw the object into either a much larger or much smaller orbit. Those objects that are thrown into smaller orbits can then be influenced by strong perturbations by the other giant planets. In this way, objects in the Kuiper Belt can successively evolve into orbits like those of the short-period comets and of the Centaurs, like Chiron.

Additional evidence that Chiron came from the Kuiper Belt is physical. Chiron's diameter of approximately 180 km, its color and the indications of volatiles on its surface are consistent with the size estimates, observed colors and expected surface chemistry for some types of Kuiper Belt objects.

Given the dynamical instabilities of Centaur orbits, the plausible method for moving Kuiper Belt objects into the solar system, and the physical similarities between Centaurs and Kuiper Belt objects, it is therefore likely that Chiron and the other Centaurs are unique transition objects, connecting the outer solar system's Kuiper Belt and inner solar system objects such as Jupiter family comets.

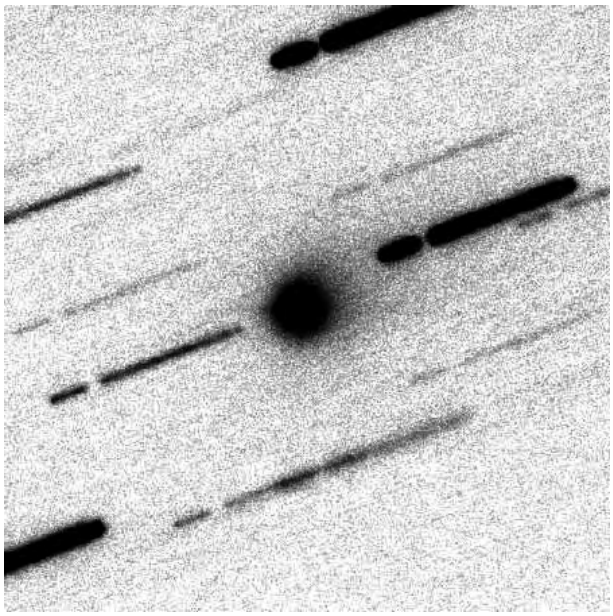


Figure 1. An *R*-band filter CCD image of Chiron obtained by Karen Meech and Graham Knopp with the University of Hawaii's 2.2 m telescope on 20 February 1993. Chiron was 9.31 AU from the Sun and 8.35 AU from the Earth, and had an apparent visual magnitude of 16.1 mag. North is towards the top and east is to the left of this 1.8×1.8 arcmin^2 image. Evidence of a coma can be seen as the haze surrounding Chiron; the dust extends about 15 arcsec (9×10^4 km) to the northwest. To correct for Chiron's motion against the background stars, this image is a composite made from 33 1 min exposures that subsequently were shifted to align with Chiron, and then added together to create an image with an effective exposure time of 33 min. Because of this 'shift-and-add' procedure, the stars appear as trails.

Observations: comet or minor planet?

As alluded to above, one of Chiron's unique features among the currently known Centaurs is that it has displayed a transient COMA, which has shown levels of activity that can be as high at aphelion as at perihelion. Indicative of its unusual nature, Chiron has two accepted designations, '(2060) Chiron' and '95P/Chiron', depending on whether one considers it as a minor planet or as a comet. However, Chiron is about 50 times larger than a typical comet.

Unusual activity was first detected by Dave Tholen, William Hartmann and Dale Cruikshank in 1988 when Chiron's brightness doubled while it was still at a distance of 12 AU from the Sun. Soon thereafter, a coma with a size of roughly 10^5 km around Chiron was observed by Karen Meech and Michael Belton. An image of Chiron and its coma is shown in figure 1. One explanation for this coma is that it comprises submicron dust particles that are escaping (entrained in volatile outgassing) from a low atmosphere. Marc Buie, James Elliot, Schelte Bus and more than a dozen collaborators observed Chiron during a stellar occultation in 1993, and their data show evidence of jet-like features, a common property in comets and inferred from models and other observations of Chiron. These jets imply that the coma is formed from a few active regions rather than a more pervasive sublimation over the surface.

The first ultraviolet spectrum of Chiron was obtained in 1995 by Noah Brosch with the long-wavelength spectrometer of the INTERNATIONAL ULTRAVIOLET EXPLORER satellite in low-resolution (6 Å) mode. One year later, the Faint Object Spectrograph on the HUBBLE SPACE TELESCOPE made ultraviolet observations at 2 Å resolution of Chiron during one of its outbursts. This 'outburst' consisted of a large (about 60%) increase in Chiron's flux, but curiously there was no evidence of an extended coma as one might expect in such a case. The source of this significant increase is unknown, and the analysis by Joel Parker, Alan Stern, Michel Festou and Michael A'Hearn suggests it may be tantalizing evidence of a highly unusual phase function, recoating of the surface from an unobserved outburst or a cold, persistent, near-surface 'fog' (small-scale-height atmosphere) that experienced a peculiar, unresolved outburst. Also, there is evidence that Chiron's color changed from gray to slightly blue during this outburst. The coincidence of this apparent color change with the flux increase may be the signature of small Rayleigh-scattering grains in Chiron's coma or on its surface associated with the higher albedo.

Chiron has exhibited activity at all points of its orbit, and there is a wide range in both the amplitude (from a few per cent to more than a factor of two) and the duration

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(from weeks to months) of these outbursts. In fact, Chiron was more active at its aphelion in 1970 than it was in following decades when it was closer to the Sun. Such an analysis (done by Schelte Bus, Edward Bowell, Alan Stern and Michael A'Hearn) of Chiron's activity was made possible by examining historical photographic plates from observatory archives.

The source of Chiron's activity is still an open question. Since Chiron is active at such large distances—e.g. beyond Saturn—its activity cannot be driven by sublimation of water-ice, which is not volatile at this distance (water-ice was first detected on Chiron by Michael Foster, Simon Green, Neil McBride and John Davies in 1998, and has been seen on other Centaurs by Robert Brown, Dale Cruikshank, and their collaborators). However, model calculations performed by Diana Prialnik, Noah Brosch and Drora Ianovici show that Chiron's observed activity pattern could be due to CO that is initially trapped in amorphous water-ice but then released when the ice is crystallized. Chiron's activity must be driven by ices that are volatile at these distances, and possible candidate species include CO, which is a common source of activity in comets, N_2, which is common on Pluto and Neptune's moon Triton, or CH_4. The variability of Chiron's outbursts, the apparent lack of a correlation between activity and distance from the Sun, and the observations of jet-like features imply that Chiron's surface may be unevenly covered with inert material.

Knowledge of the species driving Chiron's activity is the key to understanding its connection to other objects and regions of the solar system. In 1995, Maria Womack and Alan Stern made millimeter observations of Chiron when it was at a heliocentric distance of 8.5 AU. They detected carbon monoxide with a column density of $N(\mathrm{CO}) = 9 \times 10^{12}$ cm^{-2}, implying a production rate of $Q(\mathrm{CO}) = 2 \times 10^{28}$ s^{-1} if CO comes directly from the nucleus. This is comparable to the CO production rate of comet Hale–Bopp at heliocentric distances of 3.8 AU. The Hubble Space Telescope ultraviolet spectra provide the best upper limits to date for a number of other species. The data showed no evidence of coma emissions in CS λ2576, OH λ3085 or any other coma species, and they determined column abundance upper limits of $N(\mathrm{CS}) < 4.3 \times 10^{13}$ cm^{-2} and $N(\mathrm{OH}) < 1.5 \times 10^{14}$ cm^{-2}, corresponding to parent molecule production rates of $Q(\mathrm{H_2O}) < 4 \times 10^{30}$ s^{-1} and $Q(\mathrm{CS_2}) < 4 \times 10^{28}$ s^{-1}, respectively.

Chiron's future

It is still beyond the capability of modern instruments to make regular and detailed observations of the member objects of the Kuiper belt, so Chiron, due to its relative proximity, provides astronomers with the best opportunity to examine the properties of an object that was recently in the Kuiper Belt. Observations must continue to be made of its sporadic outbursts to discern the underlying mechanisms of its activity and the source of its coma. A significant step would be to make unambiguous detections and abundance measurements of the species comprising the coma gases. By the time Chiron returns for its next perihelion pass on 3 August 2046, our astronomical instrumentation and our understanding of Centaurs will hopefully have advanced to the point where we will be able to make detailed observations and analyses of its physical properties with respect to the Kuiper Belt. Perhaps we will even take advantage of Chiron's semi-centennial passage to launch a space mission to rendezvous with this unique emissary from the outer limits of our solar system.

Bibliography

Altenhoff W J and Stumpff P 1995 Size estimate of 'asteroid' 2060 Chiron from 250 GHz measurements *Astron. Astrophys.* **293** L41–L42

Brosch N 1995 The first UV spectrum of 2060 Chiron *Mon. Not. R. Astron. Soc.* **276** 1185–90

Bus S J *et al* 1996 Stellar occultation by 2060 Chiron *Icarus* **123** 478–90

Bus S J, Bowell E, Stern S A and A'Hearn M F 1991 Chiron: evidence for historic cometary activity *The Int. Conf. on Asteroids, Comets, Meteors* (Houston: Lunar and Planetary Institute) p 34

Campins H, Telesco C M, Osip D J, Rieke G H, Rieke M J and Schulz B 1994 The color temperature of (2060) Chiron: a warm and small nucleus *Astron. J.* **108** 2318–22

Dones L, Levison H F and Duncan M 1996 On the dynamical lifetimes of planet-crossing objects *Completing the Inventory of the Solar System (ASP Conf. Ser. 107)* (San Francisco: Astronomical Society of the Pacific) ed T W Rettig and J M Hahn pp 233–244

Duncan M, Quinn T and Tremaine S 1988 The origin of short-period comets *Astrophys. J. Lett.* **328** L69–L73

Edgeworth K E 1949 The origin and evolution of the solar system *Mon. Not. R. Astron. Soc.* **109** 600–9

Elliot J L *et al* 1995 Jet-like features near the nucleus of Chiron *Nature* **373** 46–8

Fanale F P and Salvail J R 1997 The cometary activity of Chiron: a stratigraphic model *Icarus* **125** 397–405

Foster M J, Green S F, McBride N and Davies J K 1999 Detection of water ice on 2060 Chiron *Icarus* **141** 408–10

Fulle M 1994 Spin axis orientation of 2060 Chiron from dust coma modeling *Astron. Astrophys.* **282** 980–8

Hartmann W K 1990 The changing face of Chiron *Astronomy* **18** 44–8

Hartmann W K, Tholen D J, Meech K J and Cruikshank D P 1990 2060 Chiron—colorimetry and cometary behavior *Icarus* **83** 1–15

Jewitt D and Luu J 1992 Submillimeter continuum observations of 2060 Chiron *Astron. J.* **104** 398–404

Kowal C T and Gehrels T 1977 Slow-moving object Kowal *IAU Circular* **3129** 1

Kuiper G P 1951 On the origin of the solar system *Astrophysics: A Topical Symposium* ed J A Hynek (New York: McGraw Hill) pp 357–424

Lebofsky L A, Tholen D J, Rieke G H and Lebofsky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M J 1984 2060 Chiron—visual and thermal infrared observations *Icarus* **60** 532–7
Luu J X 1993 Cometary activity in distant comets—Chiron *Publ. Astron. Soc. Pacific* **105** 946–50
Meech K J and Belton M J S 1990 The atmosphere of 2060 Chiron *Astron. J.* **100** 1323–38
Meech K J, Buie M W, Samarasinha N H, Mueller B E A and Belton M J S 1997 observations of structures in the inner coma of Chiron with the HST planetary camera *Astron. J.* **113** 844–62
Oikawa S and Everhart E 1979 Past and future orbit of 1977 UB, object Chiron *Astron. J.* **84** 134–9
Parker J Wm, Stern S A, Festou M C and A'Hearn M F 1997 Ultraviolet observations of Chiron with the HST/FOS: examining a Centaur's gray matter *Astron. J.* **113** 1899–1908
Prialnik D, Brosch N and Ianovici D 1995 Modelling the activity of 2060 Chiron, *Mon. Not. R. Astron. Soc.* **276** 1148–54
Stern S A and Campins H 1996 Chiron and the Centaurs: escapees from the Kuiper belt *Nature* **382** 507–10
Sykes M V and Walker R G 1991 Constraints on the diameter and albedo of 2060 Chiron *Science* **251** 777–80
Tholen D J, Hartmann W K and Cruikshank D P 1988 (2060) Chiron *IAU Circ.* **4554** 2
Womack M and Stern S A 1997 Observations of carbon monoxide in (2060) Chiron *28th Ann. Lunar and Planetary Sci. Conf.* (Houston: Lunar and Planetary Institute) p 1575

Joel Wm Parker

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chladni, Ernst Florens Friedrich (1756–1827)

German physicist (acoustics). He collected meteorites. He claimed on the basis of a comprehensive catalog of accounts of finds and falls that he compiled that they fell from the sky. This view was comprehensively rubbished by the scientific establishment, especially the French Academy of Sciences, until convinced otherwise by BIOT's investigation of the fall of meteorites at L'Aigle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chondrites

The chondrites are the most common type of METEORITE falling to Earth from space. They acquire their name because most of them contain 'chondrules' (from the Greek, 'chondros', a grain) which are silicate beads sometimes found in large numbers. The chondrules have always fascinated scientists. The inventor of the petrographic microscope, H C Sorby, described them as droplets of fiery rain. They are essentially glassy beads made by a violent but brief heating event that caused dust grains to form millimeter-sized melt droplets. However, the cause of the heating remains a mystery. The most important feature of the chondrites is that, with the exception of a few highly volatile elements, they have the same composition as the Sun. They are also extremely ancient rocks, having formation ages comparable with the age of the solar system (4600 million years). Finally, they have unique textures suggestive of accumulation of diverse components with little or no subsequent alteration. In contrast, the ACHONDRITES are meteorites that have distinctly non-solar composition and are igneous rocks that formed by various types of volcanism.

The largest chondrite seen to fall from space is the 4000 kg Jilin, while the largest whose fall was not observed but for which the meteorite found on the surface of the Earth is the Tsarev chondrite that was found in Russia in 1968. The most famous chondrite is probably the ALLENDE METEORITE, which fell in Mexico in 1969. About 2 t were recovered strewn across farmland in Chihuahua province. The Allende chondrite is famous for the large number of white silicate aggregates with extraordinary chemical and isotopic properties. The 691 g Semarkona chondrite, which fell in India in 1940, is scientifically well known because it is the member of the largest chondrite class that most closely resembles original solar system matter.

Meteorites usually produce spectacular effects as they fall through the atmosphere, fireballs, many sonic booms and even smells are produced, and the surfaces of the surviving stones are blackened where a thin skin of material melted during atmospheric passage. The meteorites usually break into many pieces that are strewn over an ellipse on the Earth with the heaviest pieces travelling furthest. The Pultusk meteorite produced 100 000 fragments. About 850 chondrites have been seen to fall and placed in the world's museums, and nearly 20 000 fragments have been found on the Earth, mainly in North America, north Africa, central and western Australia and, especially, Antarctica.

This article describes the classification and basic properties of chondrites, their physical history as determined from isotopic studies, radiometric dating and petrology, and some current ideas for their origin.

Classification and basic properties of chondrites

The chondrite classes are indicated in table 1, along with information on their major properties. Also included are two classes of 'primitive achondrites'. These have traditionally been regarded as achondrites because of their igneous textures, but their composition is solar and it is now assumed that they are chondrites that melted without changing composition.

Although chondrites have the same proportions of most elements as the Sun, two volatile elements that are not present in solar proportions are sulfur and oxygen. The abundance of these elements, oxygen in particular, varies significantly among the chondrites. This has important implications for the minerals present. In the EL and EH chondrites (collectively termed 'enstatite chondrites'), iron is in the metal and sulfide form, but little or none is in the silicate form. The dominant silicate mineral is enstatite, $MgSiO_3$, after which the class is named. The H, L and LL chondrites (collectively termed 'ordinary chondrites') contain more oxygen than the enstatite chondrites. Thus while some of the iron exists as metal and sulfide, much is in the silicate form as the stony (silicate) minerals olivine and pyroxene. (Olivine is a solid state solution of Mg_2SiO_4 and Fe_2SiO_4, the Fa term in table 1 referring to the percentage of Fe_2SiO_4 present, while pyroxene is a solid solution of $MgSiO_3$ and $FeSiO_3$, the Fs term in table 1 referring to the percentage of $FeSiO_3$ present.) The CI, CM, CV and CO chondrites (collectively termed 'carbonaceous chondrites') contain more oxygen than the other chondrites and in some the iron no longer exists in the metal form but is entirely replaced with magnetite and similar minerals. The CARBONACEOUS CHONDRITES also contain higher amounts of volatile constituents than the other chondrites, the CI chondrites being 20% water and therefore best considered as rather hard balls of mud.

These properties can conveniently be displayed by a plot of iron in the metal form against iron in the sulfide and silicate form (figure 1(*a*)). On such a plot, first used by Nobel laureate Harold Urey and his colleague Harmon Craig, the chondrites will lie on a diagonal with oxygen-poor meteorites on the top left and oxygen-rich meteorites on the bottom. The diagonal will be the 'solar total iron' line if the ratio of iron to silicon is equal to the solar value. However, the chondrites move away from the solar diagonal, indicating that, in addition to the differences in the amount of oxygen they contain, some chondrites contain different amounts of total iron.

Other properties seem to vary with these chondrite classes, such as the abundance of highly refractory (non-volatile) elements that show small but significant decreases from solar as we go from carbonaceous to ordinary to enstatite chondrites. Similarly, the relative abundance of the three isotopes of oxygen in the different chondrite classes is highly characteristic. Carbonaceous chondrites (CI chondrites excepted) contain a greater proportion of the major oxygen isotope (^{16}O) than rocks from the Earth and Moon, whereas enstatite chondrites and CI chondrites have similar proportions to the Earth and Moon and ordinary chondrites have a lower proportion of ^{16}O. The biggest effects in oxygen isotopes are displayed by the CV chondrites and appear

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Classes of chondrites and some of their properties.

	Statistics		Texture				Composition				
Class	Falls (number)	Finds (number)[a]	Chondrule (vol%)	Chondrite (μg)	Metal (vol%)	Metal (μg)	Fe/Si (a/r)	Mg/Si (a/r)	Fa,Fs (mol%)	$\delta^{17}O$[b] (‰)	$\delta^{18}O$[b] (‰)
Carbonaceous chondrite											
CI	5	0 (2)	0	–	0	–	0.86	1.05	nd,nd	9.23	16.84
CM	13	4 (156)	5–15	270	≤0.1	–	0.80	1.05	15,2	3.0	10.9
CO	5	37 (81)	40	150	3.6	nd	0.77	1.05	14.3,4.0	−3.0	−2.6
CV	6	11 (81)	40	1000	2.3	nd	0.78	1.03	8.4,4.0	−15.4	−11.3
CK	1	7 (73)	15	700	0	–	0.78	1.09	30.9,26.9	−1.8	−5.8
CR	3	12 (76)	55	700	10.7	nd	0.82	1.05	4.7,3.4	2.29	6.25
Ordinary chondrites											
H	311	974 (6881)	70	380	8.4	190	0.81	0.96	19.3,16.8	2.9	4.1
L	425	855 (6332)	70	620	4.1	160	0.57	0.93	25.2,20.9	3.5	4.6
LL	73	114 (960)	70	750	2.0	140	0.52	0.94	31.3,25.2	3.9	4.9
R	1	4 (10)	12–40	380	<0.1	–	0.60	0.94	38,0–32	5.52	5.07
Enstatite chondrites											
EH	8	4 (63)	20	280	10.1	41	0.95	0.68	0.4,0.3	3.0	5.6
EL	7	8 (30)	20	480	10.2	78	0.62	0.81	0.4,0.3	2.7	5.3
Primitive achondrites											
Bra[c]	0	4 (6)	–	–	nd	nd	0.65	1.16	33,18	1.61	3.48
Aco[d]	1	2 (12)	–	–	2	nd	0.70	1.06	11.9,12.6	3.73	1.00

[a] Non-Antarctic finds, with all finds in parentheses.

[b] The δ symbol refers to the difference in parts per thousand of the $^{17}O/^{16}O$ and $^{18}O/^{16}O$ ratios between the sample and a standard sample of ocean water.

[c] Brachinites.

[d] Acopulcoites.

(a/r) = atomic ratio

to be related to the white silicate aggregates that are characteristic of this class.

The main questions posed by the chondrite classes is that, given that chondrites are ancient and that they are essentially solar in composition, how did these variations in iron chemistry, amount of iron and the other properties come about? What do they tell us about conditions in the early solar system?

One clue to the nature of the process that gives rise to these properties is the nature and abundance of the chondrules and the relative sizes of the chondrules and metal grains (figure 1(*b*)). The abundance of chondrules falls from about 75 vol% in the ordinary chondrites to 35 vol% in enstatite and CV and CO chondrites to less than 10 vol% in CM and essentially zero in CI chondrites. Chondrules are highly diverse, some containing iron-rich silicates and being relatively rich in volatile elements and some containing iron-poor silicates are being depleted in volatile elements. Thus it is at least arguable that some of the oxygen and refractory element trends in figure 1(*a*) reflect the abundance and proportion of the different types of chondrules. Metal grain size also varies considerably from class to class. Enstatite chondrites have the smallest and ordinary chondrites the largest metal grains, while carbonaceous chondrites generally have little or no metal.

History of chondrites

If one looks at a series from the same class under the microscope there is often a wide range of textures, with the chondrules being distinct in some chondrites and blurred in others. The chondrules are essentially grains of olivine and pyroxene surrounded by a material that is glassy in some chondrules, translucent in others and crystalline in yet others. Similarly, the matrix between the chondrules and metal grains is fine grained in some chondrites and coarse grained in others. In addition to these physical trends, the composition of mineral grains and phases becomes more uniform as textures become blurred. These effects are all attributed to heating of the rock after its assembly. The effect is known as metamorphism, and it is described by a numbering system from 3 to 6 with unmetamorphosed type 3 and heavily metamorphosed meteorites being type 6. From detailed studies of the chemistry of phases in chemical equilibrium we can estimate that types 3, 4, 5 and 6 correspond approximately to temperatures of 400–600 °C, 600–700 °C, 700–750 °C and 750–950 °C, respectively. Types 1 and 2 refer to chondrites metamorphosed in the presence of considerable amounts of water (∼50 vol%), the type 1 at 100–150 °C and type 2 at <20 °C. From compositional profiles in the metal of ordinary chondrites, caused by incomplete equilibration, we can estimate that cooling rates following metamorphism were typically 10–100 K Myr^{-1}.

Another process that most if not all the chondrites have suffered is reflected in the naked-eye appearance of the chondrites. They often appear to be made up of fragments of previously existing rocks. The Fayetteville meteorite (figure 2(*a*)) is an excellent example

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

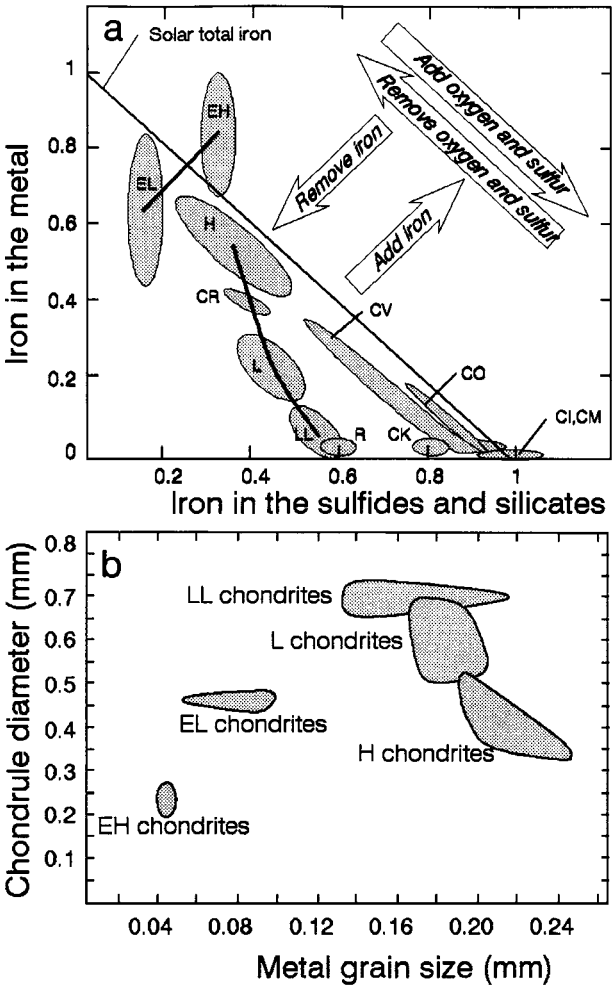
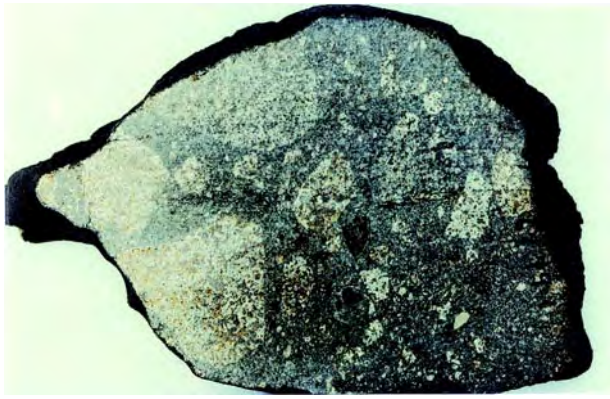
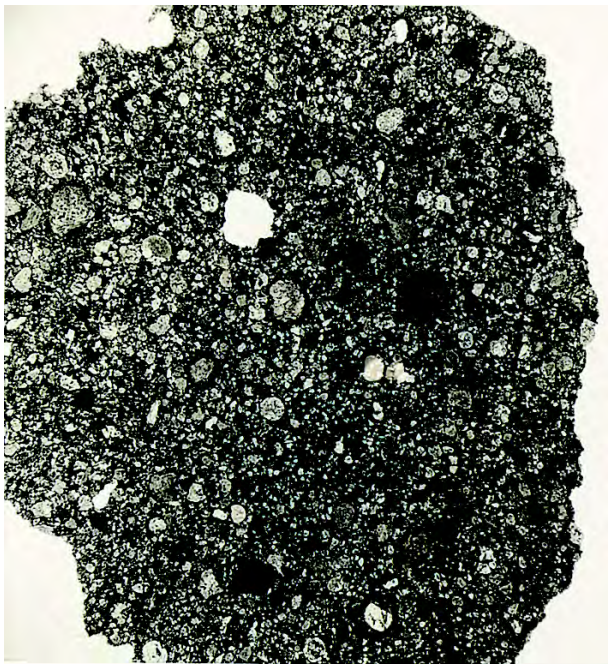


Figure 1. The basic properties of chondrites are summarized in these two plots. (*a*) Iron exists in three forms, as metal, sulfide and silicates. Plotting iron in the metal form (as ratio of iron atoms to silicon atoms) against iron the sulfides and metal produces several different fields, each representing a different chondrite class. How the classes relate in terms of adding or removing iron, sulfur and oxygen are indicated by the large arrows. (*b*) Physically, the chondrite classes are characterized by the size of their chondrules and metal grains. The carbonaceous chondrite classes are not shown because they contain little or no metal, but they do contain very large amounts of fine-grained silicate matrix.



(*a*)



(*b*)

Figure 2. (*a*) A cut face of the chondrite that fell at Fayetteville, Arkansas, on 26 December 1934. The black fusion crust produced during atmospheric passage can be seen around the edge. The interior consists of light fragments of normal H chondrite embedded in a darker 'matrix' of crushed H chondrite material mixture that is physically similar to the 'soil' on the Moon. Fayetteville is a chondrite that is described as a gas-rich regolith breccia fron the surface of an airless body. The cut face is 9 cm in its longest dimension. (Johnson Space Center photograph S84-32404.) (*b*) A slice from the Indarch EH3 chondrite cut so thin that light can pass through and reveal its internal texture. The sulfide and metal grains are opaque and appear black, while the many chondrules appear as light gray circular objects typically 0.5 mm in diameter. The longest dimension of the thin section is 1.4 cm. (Smithsonian Institution photograph M-1249.)

of this texture. Technically, the process is referred to as 'brecciation' and the rocks are referred to as 'BRECCIAS'. Fragments of perfectly normal chondrite are sometimes surrounded by a darker 'soil' made of crushed chondrites and being rich in inert gases, carbon and crystals that have been damaged by radiation. These meteorites are referred to as 'gas-rich regolith breccias'. They were the surface debris of some airless body exposed to solar wind and energetic solar radiations. The Apollo astronauts found similar rocks on the Moon.

In addition to metamorphsim and brecciation, many meteorites show evidence for sudden excursions to high temperatures followed by very rapid cooling. Crystals are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cracked and deformed, and metal and sulfide minerals are melted and forced down cracks, and sometimes minerals acquire unusual textures. These effects can be reproduced in the laboratory by exposing the chondrites to shock waves produced by plastic explosives or large guns. The effects are classified as a series of levels, S1, S2, S3, S4, S5 and S6, that refer to shock pressures of <5 GPa, 5–10 GPa, 10–20 GPa, 20–30 GPa, 30–50 GPa and >50 GPa, respectively (1 Gpa, or gigapascal, is equal to about 10 000 atm). The source of the shock wave experienced by the chondrites is almost certainly impact between two ASTEROIDS in space.

Some of the best information we have on the history of chondrites is provided by various dating methods (figure 3). As soon as the meteorite is removed from its original parent body, and acquires meter-sized dimensions, COSMIC RAYS bombard the stone and create isotopes not otherwise present. From these isotopes it is possible to calculate the duration of exposure to cosmic rays. It is frequently found that large numbers of chondrites of a given class were exposed at the same time, and this is strong evidence that all or most of the chondrites were released from the same parent object at that time. The best example is the H chondrite class, most of whose members became exposed to cosmic rays 8 million years ago. Several other chondrite classes display similar peaks in their exposure age distributions.

A different dating method leads to the same conclusion for the other major chondrite class, the L chondrites. The 'Ar–Ar age' is based on the radioactive decay of ^{40}K to ^{40}Ar; the potassium is measured in the laboratory from the abundance of another experimentally produced isotope, ^{38}Ar. Most of the L chondrites have Ar–Ar ages of about 500 million years, and many of these are intensely shock heated. It thus seems that 500 million years ago all these separate L chondrites experienced a common heating event that caused a loss of the ^{40}Ar gas and a resetting of the Ar–Ar clock. They must also have been part of the same parent object at that time. The exposure and Ar–Ar ages are evidence that members of the large chondrite classes came from one or two objects in space.

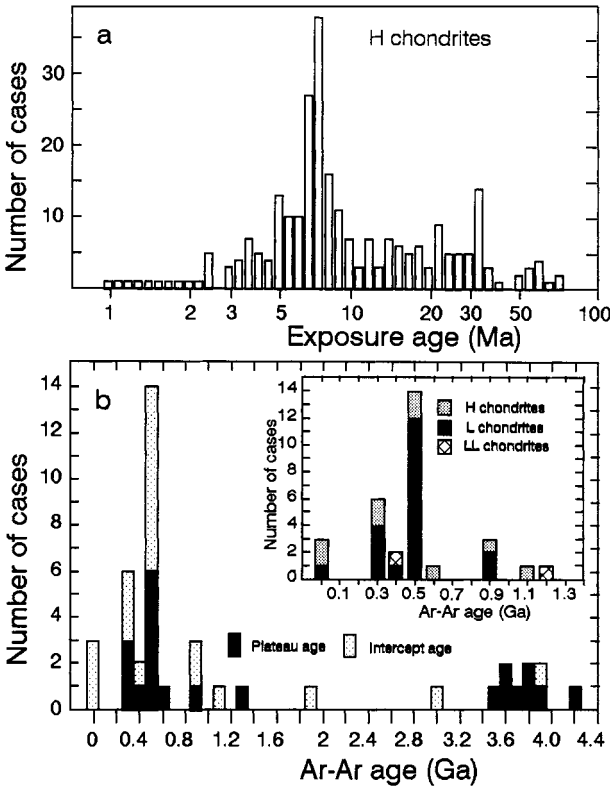


Figure 3. Two kinds of ages provide important insights into the history of the chondrite meteorites. (*a*) The duration of exposure to cosmic rays, calculated from the abundance of isotopes produced by cosmic ray bombardment of the meteorite. The peak in the histogram at 8 million years indicates that many or most of the individual meteorites falling to Earth were exposed to cosmic rays at the same moment and must have been part of the same single object prior to that break-up event. The data shown are for the H chondrites but other classes often show similar tendencies towards preferred cosmic ray exposure ages. (*b*) The Ar–Ar age essentially measures the time during which the ^{40}Ar accumulated in the meteorite as a result of radioactive decay of ^{40}K. The peak in the histogram at 500 million years shown mainly by the L chondrites indicates that at this time a large number of L chondrites had the Ar–Ar 'clock' reset to zero by loss of previously accumulated ^{40}Ar. Thus, most members of this class also experienced a common heating event at one time, indicating that they too were part of the same object at the time of that event. 'Plateau' and'intercept' ages refer to different ways to handle the data.

Origin of chondrites

So what were the objects from which the chondrites are coming? A handful of meteorites, mostly chondrites, were observed sufficiently well during fall that their orbits could be determined. The orbits resemble those of the NEAR-EARTH ASTEROIDS, coming within about 0.9 AU of the Sun and going out to the main ASTEROID BELT at about 2.5 AU (1 AU, the 'astronomical unit', is the distance of the Earth from the Sun, 150 million kilometers). The relationship between comets and asteroids is unclear, and as many as half the near-Earth asteroids could be extinct comet nuclei, but many could have come from the main asteroid belt.

The spectrum of the sunlight reflected from the surface of asteroids provides clues to their surface composition. Over 2000 asteroids have now been sorted into classes in this way, and many appear to resemble meteorites (table 2). However, asteroids with spectra resembling those of the ordinary chondrites, the chondrite class that dominates the Earth's meteorite flux, are very rare among the asteroids. Perfect matches are found for less that 1% of asteroids, and even if one assumes that some form of space weathering is affecting the asteroid surfaces, then less than 11% of the asteroids could be the source of the Earth's major chondrite classes. There are essentially two factors that determine which meteorites reach the Earth besides the abundance of their parent asteroids in the asteroid belt. These are the orbital

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mechanism for transferring objects from the asteroid belt to Earth and the Earth's atmosphere. Passage through the Earth's atmosphere destroys most of the fragile CI and CM meteorites but leaves unscathed the tough ordinary chondrites. Nearly one-half of the classified asteroids in the main belt have spectra resembling those of CI or CM chondrites while only a few per cent of meteorites reaching the surface of the Earth are members of these classes.

Table 2. The asteroid classes, their relative abundance and meteorite matches.

Class	Abundance (%)	Meteorite match[a]
D	6.2	IDP
P	4.3	IDP
C	22.8	CI,CM
T	1.3	Sulfur-rich irons
B+G+F	9.1	CI,CM
Q	0.41	H,L,LL
V	0.41	HED
R	0.10	?
S (IV)	11.0	Lod, Win, OC, irons
S (others)	25.9	Pal, Ure, Bra, Lod, Mes, Win, Ste
A	0.71	Pal,
M	4.3	Irons, EC
E	1.3	Aub

[a] Meteorite classes most closely resembling the asteroid class. The chondrite classes are H, L, LL (ordinary chondrites, OC), CI, CM, CV, CR, CO, EH and EL (enstatite chondrites, EC); Win, winonites; EC, enstatite (EH and EL) chondrites; Bra, brachinites. Other classes: Ste, Steinbach; Mes, mesosiderites; HED, howardites–eucrites–diogenites; Aub, aubrites; Ure, ureilites; Pal, pallasites; Lod, lodranites; IDP, interplanetary dust particles.

We think that most of the ordinary chondrite classes are coming from single parent bodies, and spectral data show these objects to be rare in the asteroid belt. Recently it has been suggested that the H chondrites are coming from asteroid 6 Hebe, the '6' indicating that it was the sixth asteroid discovered. This 200 km object not only has a surface spectrum similar to that of H chondrites, mixed with some related IRON METEORITES, but also is located in space where both Saturn and Jupiter exert a strong periodic gravitational influence that could send meteorites from Hebe to Earth.

If we assume, as seems likely, that the various chondrite classes are coming from different asteroids, then how did the different classes originate? How did their parent asteroids have different surfaces? This is equivalent to asking how the chondrules in the meteorites formed and how the metal and silicate proportions came to differ from class to class. Most scientists argue that chondrite formation and metal–silicate separation occurred in the NEBULA, that lightning or other pulses of energy shot through the nebula dust to produce chondrules and that the metal and silicate grains swirling around in the nebula became separated, perhaps by aerodynamic sorting by the gas or perhaps by some other process. However, the present writer has suggested that both chondrule formation and metal–silicate separation occurred on the surfaces of asteroids. I have suggested that chondrules formed by impact and that metal and silicate grains became separated as gases from the interior of the asteroid, mainly water, passed through the dry surface layers, size- and gravity-sorting chondrules and metal, and adjusting their relative proportions, as they escaped to space.

So the origin of meteorites is essentially the same as the origin of asteroids. The nebula from which the Sun formed fragmented from the larger interstellar cloud and underwent gravitational collapse to produce the proto-Sun and a number of rings of material that quickly coalesced into planets. At the location of the asteroid belt, 2–3 AU from the Sun, the gravitational forces from the major planets (Jupiter and Saturn) were so strong that coalescence was prevented and a large number of tiny PLANETESIMALS formed. Some of these grew by ACCRETION but most fragmented by impact to form the asteroid belt much as we observe it today.

Many of the asteroids contained enough of the radioactive isotope of aluminum, ^{26}Al, that large amounts of heat were generated and they melted to produce the many igneous classes of meteorite. The daughter product of ^{26}Al decay, ^{26}Mg, has been found in aluminum-rich phases of certain meteorites. However, some chondrites were heated without melting, and many were not heated at all. Chondrites are very clearly primitive early solar system materials. Just to underscore this point grains of diamond, graphite, silicon carbide and aluminum oxide have been found in chondrites that have isotopic properties indicating that they were not produced in our own solar system but were ejected from the outer atmospheres of a variety of stars. They are presolar, INTERSTELLAR GRAINS. Thus, while chondrites are widely regarded as heralds bringing us information on the early solar system, in a very real sense they are a connection between our laboratories and processes occurring in interstellar space and distant stars

Bibliography

Kerridge J F and Matthews M S (ed) 1988 *Meteorites and the Earth Solar System* (Tuscon, AZ: University of Arizona Press)

Hewins R H, Jones R H and Scott E R D (ed) 1996 *Chondrules and the Photoplanetary Disk* (Cambridge: Cambridge University Press)

Huang Shaoxing, Akridge D G and Sears D W G 1996 Metal-silicate fractionation in the surface dust layers of accreting planetesimals: implications for the formation of ordinary chondrites and the nature of asteroid surfaces *J. Geophys. Res.* **101** 375–85

Symes S J K, Sears D W G, Akridge D G, Huang Shaoxing and Benoit P H 1998 The crystalline lunar spherules: their formation and implications for the origin of meteoritic chondrules *Meteoritics Planetary Sci.* **33** 13–29

Derek Sears

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Christiansen, Wilbur Norman (1913–)

Australian radioastronomer, built a radioastronomy interferometer array, consisting of 32 antennas in an east–west line. It was built to investigate the Sun, and expanded with a north–south array (in what was called, after Christiansen, the *Chris-cross*) to produce true images. This radio-interferometer was used to make the first purely Earth rotation synthesis maps. The Fourier transforms, however, had to be done by hand, so a single contour map took half a year of hand computations to produce. Also carried out a preliminary 21 cm survey of interstellar hydrogen and detected two of the Milky Way's spiral arms.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Christie, Sir William Henry Mahoney (1845–1922)

British astronomer, worked at the Royal Observatory, Greenwich, as Chief Assistant (assistant director) before succeeding Airy as Astronomer Royal (director) in 1881. Turned the Royal Observatory towards astrophysics (as opposed to positional astronomy). He installed large collecting area telescopes (e.g. the 28 in refractor, still in operation, used originally for double star observations). He brought the status of the Prime Meridian of the world's international time zone system back for Greenwich from the Washington Conference of 1884. He expeditiously executed the work of measuring the positions of all the stars in the large Greenwich Zone (northern region) of the Astrographic Catalog. Founded and first edited *The Observatory* magazine, which still preserves a human face for British astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromatic Aberration

An optical defect whereby light rays of different wavelengths are focused at different points along the optical axis of a lens. Because the refractive index of glass is a function of wavelength, shorter-wavelength light is refracted (deflected) to a greater extent than longer-wavelength light. Consequently, shorter-wavelength light (for example, blue light) is brought to a focus closer to the lens than longer-wavelength light (for example, red). Since white light consists of a mixture of wavelengths, the image of a star, formed by a simple lens, consists, in effect, of a series of different-colored images spread out along the optical axis. The image formed at the blue focus, for example, consists of a blue dot surrounded by a halo of false color consisting of blurred images of varying sizes in the light of other wavelengths. Similarly, the image formed at the red focus consists of a red dot, surrounded by a blurred halo of colors. Between these two foci, there is a plane at which the smallest blurred image occurs, this image being called the circle of least confusion. Since this coincides approximately with the focus for light in the yellow-green part of the spectrum, to which the human eye is most sensitive, the best visual image is normally obtained when the eyepiece is adjusted to put this wavelength region into sharpest focus.

The spread of foci along the optical axis is called longitudinal chromatic aberration, and the spread in the plane containing the circle of least confusion is called transverse chromatic aberration. The degree to which the refractive index of an optical material, such as glass, varies with wavelength, is called the dispersion.

The effects of chromatic aberration can be considerably reduced by means of compound objective lenses that consist of two or more appropriately shaped components, each made of glass with different dispersions. An achromatic lens (or achromatic doublet), which has two components, can bring two wavelengths to the same focus and reduce the spread in focal lengths for the other wavelengths by about a factor of ten. A three-element apochromat reduces the spread even further.

Because the laws of reflection are independent of wavelength, chromatic aberration does not affect the mirrors that are used in reflecting telescopes.

See also: focal length, objective lens, refractor, spherical aberration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere

The chromosphere (Greek $\chi\rho\omega\mu\alpha$, color) is a gaseous shell atmosphere surrounding the Sun, and more generally stars, with a typical thickness of 2000 km, lying above the SOLAR PHOTOSPHERE and containing mainly hydrogen and helium gas.

Historical introduction

The total eclipse of 1860 may be regarded as marking the first serious study of the chromosphere. During eclipses the moon appears surrounded by a colored circle with some protruding structures, called PROMINENCES, which extend through the external corona and finer structures like 'poils' in French that would be the spicules. During the eclipse of 1970 C A Young used a spectrograph and a tangential slit. When the moment of totality arrived he was rewarded by the sudden appearance of numerous emission lines called a flash spectrum. This phenomenon lasted only a few seconds since the Moon rapidly covered the lowest layers of the chromosphere. In 1901 J Evershed published the spectrum of the chromosphere and discussed the similarities to and differences from the photospheric spectrum (table 1). During the latter part of the 19th century a new kind of instrument, *the spectroheliograph*, was developed at Mt Wilson by G E Hale and at Meudon by H Deslandres (see SPECTROHELIOGRAPH/SPECTROHELIOSCOPE). Observing in the H and K lines of ionized calcium (Ca II), they were successful and discovered the regular pattern of the chromosphere called the chromospheric network (see CHROMOSPHERE: NETWORK). The interpretation of these lines in emission above the limb was puzzling and nobody had in mind at that time that the chromospheric temperature could be higher than the photospheric one. We had to wait until 1942 for the discovery of the hot CORONA surrounding the Sun. A new instrumental development helped to give a better understanding into the chromosphere. In 1933 B Lyot published a report on birefringent filters with a narrow bandwidth which allowed the spicules at the limb to be seen. These filters became quite widely distributed and used in the 1950s and the fine structure of the chromosphere has been systematically studied, using the Hα line, as bushes of fine bright and dark structures tied in the network. These structures are called mottles. Outside of an eclipse the chromosphere can be observed with CORONAGRAPHS. This kind of instrument was first developed by B Lyot. The principle is based on the introduction of an artificial moon occulting the photospheric disk.

The principal characteristics of the chromosphere are its thickness of 2000 km and its basic temperature and pressure. Beyond the solar limb the chromospheric spectrum is characterized by emission lines. On the disk these lines appear as dark lines in the solar spectrum. They correspond to absorbed radiation at discrete wavelengths by elements of the chromosphere. The principal lines in the visible spectrum have been identified and are called the Fraunhofer lines (table 1). The intensity of the lines reflects the physical conditions of the plasma which absorbs or emits the radiation and is a good diagnostic of the plasma temperature. Above the photosphere, the temperature continues firstly to decrease to a minimum of 4200 K and then increases again.

Table 1. The strongest lines in the chromospheric spectrum (3200–9100 Å) classified by their intensity.

Wavelength (Å)	Element	Name of the line
6562.82	H I	Hα
4861.33	H I	Hβ
3933.66	Ca II	K
3968.47	Ca II	H
5975.65	He I	D_3
4340.47	H I	Hγ
8542.09	Ca II	
4101.74	H I	Hδ
8662.14	Ca II	
3999.65	H I	H_8
3970.07	H I	Hϵ

Morphology and dynamics of the quiet chromosphere

The aspect of the chromosphere depends strongly on the line used for the observations, and more precisely on the wavelength in which the image is obtained, e.g. Ca II H and K, Hα, He II 304 Å. The solar disk observed in Fraunhofer lines has a diameter slighter larger than the photospheric disk and is covered by a pattern very different from the granular pattern of the photosphere. In the H and K lines of Ca II the images of the chromosphere show a bright network surrounding more or less large cells which have a diameter of the order of 30 000 km (figure 1). This cell pattern overlies the photospheric supergranulation which is a manifestation of large-scale solar convection, an important mechanism for magnetic flux transport.

The H and K lines have a special profile with extremely broad absorption wings and a central emission. The central emission of the K line is called K3, the two adjacent maxima are called K2 (K2v violet wing and K2r red wing) and far in the wings it is called K1. Similar definitions are valid for the H line. The spectroheliograms in K1 show the photospheric pattern with granulation and sunspots. In K3 there is a mesh-like brightness pattern over the entire solar surface, which represents the network and the supergranules or the intranetwork, also called the internetwork. The network is more prominent in two symmetric zones north and south of the equator, located at tropical latitudes. In K2 very small bright points are visible over all the solar surface.

Network

The chromospheric network is directly related to the magnetic field. It appears brighter than the mean chromosphere owing to pressure and temperature variations and possibly heating in the magnetized atmosphere. The strength of the central part of the K line as measured in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

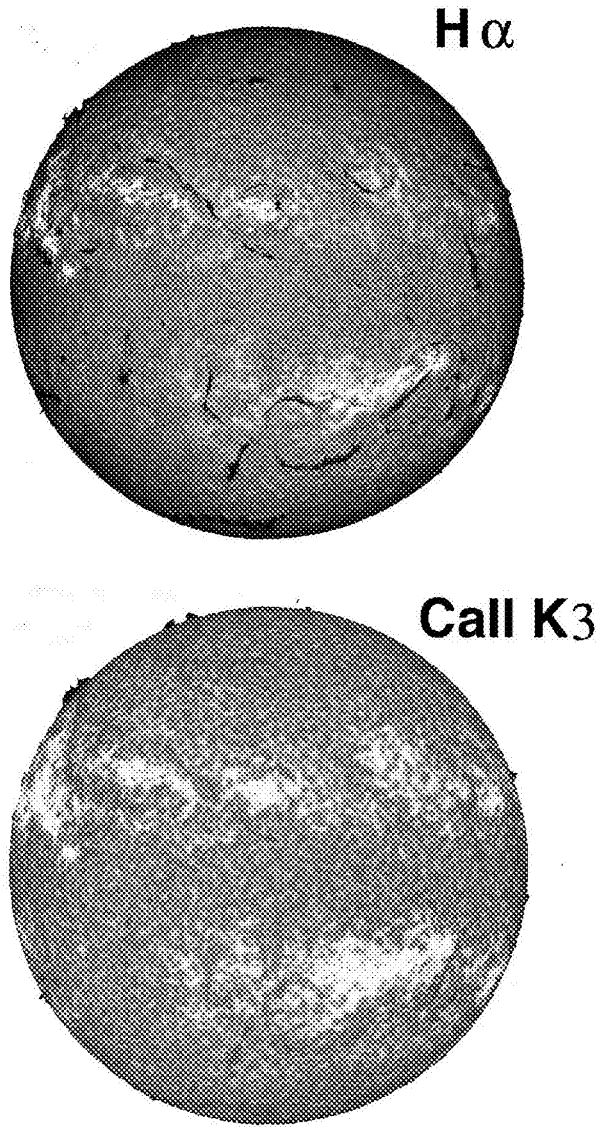


Figure 1. The chromosphere observed in the lines Hα of hydrogen and Ca K3 of ionized calcium in spectroheliograms obtained at Meudon (Courtesy of the Observatoire de Paris).

a 1 Å band is used as an index (K index) to define the local magnetic flux density in the underlying photosphere and is used for stars to define the degree of magnetism of the star. In the quiet Sun the K index increases linearly with the magnetic flux density.

However, this law is not true in magnetized regions (exceeding 400 G), where the power law decreases with an exponent lower than unity. This form is called the canopy effect (see CHROMOSPHERE: MAGNETIC CANOPY). The most conspicuous examples of canopy phenomena are associated with sunspots. The field forms a canopy over a non-magnetized region between 600 and 1000 km in the chromosphere (figure 2). The network is the site of the anchorage of coronal loops extending to coronal altitudes.

The Hα spectroheliogram (figure 1) has some similarities to the K3 spectroheliogram with two bright zones parallel to the equator. The network contrast is nevertheless weaker than in K3.

In the Lyman line (Lα) the network is much more enhanced than in Hα and is filamentary in appearance.

Bright grains

In K2 of the Ca II line, and better in UV lines, the network and the internetwork are grainier. The grains have a high contrast at wavelengths representative of the temperature minimum, e.g. in the continuum at $\lambda \sim 1600$ Å. The space imager of the TRACE satellite shows such grain patterns at high resolution. The continuum UV bright grains are observed everywhere as bright points. They have a characteristic size of 1000 km and a brightness temperature excess of 30–360 K.

In the network the bright points are generally associated with magnetic elements which collide. Most of the time they have opposite polarities and the bright points are due to heating of the plasma after reconnection. Magnetic flux is transformed to heat.

In the intranetwork, in high-resolution spectroheliograms obtained in the K2 or H2 wings of Ca II lines, bright grains (Ca II bright grain) are again well observed. For many years it was believed that they were due to the presence of small magnetic tubes like network bright points. The relationship of these bright grains to small magnetic elements has been subject to a long debate because the coalignment was not conclusive. It has been now proved that the bright grains are due to the behavior of chromospheric oscillations which produce shock waves (figure 2).

Fine structures

The Hα spectroheliograms (figure 1) show the chromospheric network but more distinctively the existence of very tiny small elongated structures. Gathered around sunspots are structures called fibrils (CHROMOSPHERE: FIBRILS). Thicker and longer structures in active regions or between active regions are called filaments. Tiny structures visible in emission above the limb are spicules. The small bushes of fine structures anchored in the network are mottles. These fine structures are controlled by the magnetic field which confines the plasma. The magnetic field plays an important role in the chromosphere (see below the definition of the plasma β).

Mottles

The mottles are irregular threads of large brightness contrast which link together to form quasi-regular pattern around and above supergranules in the photosphere. The mottles appear differently at different wavelengths in the Hα line. At the center of the line bright mottles are low lying (700–3000 km above the photosphere), while the dark mottles form a coarse fragmentation, but the admixture of bright and dark features makes the large-scale pattern less distinct than in the Ca II H and K lines in the quiet Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

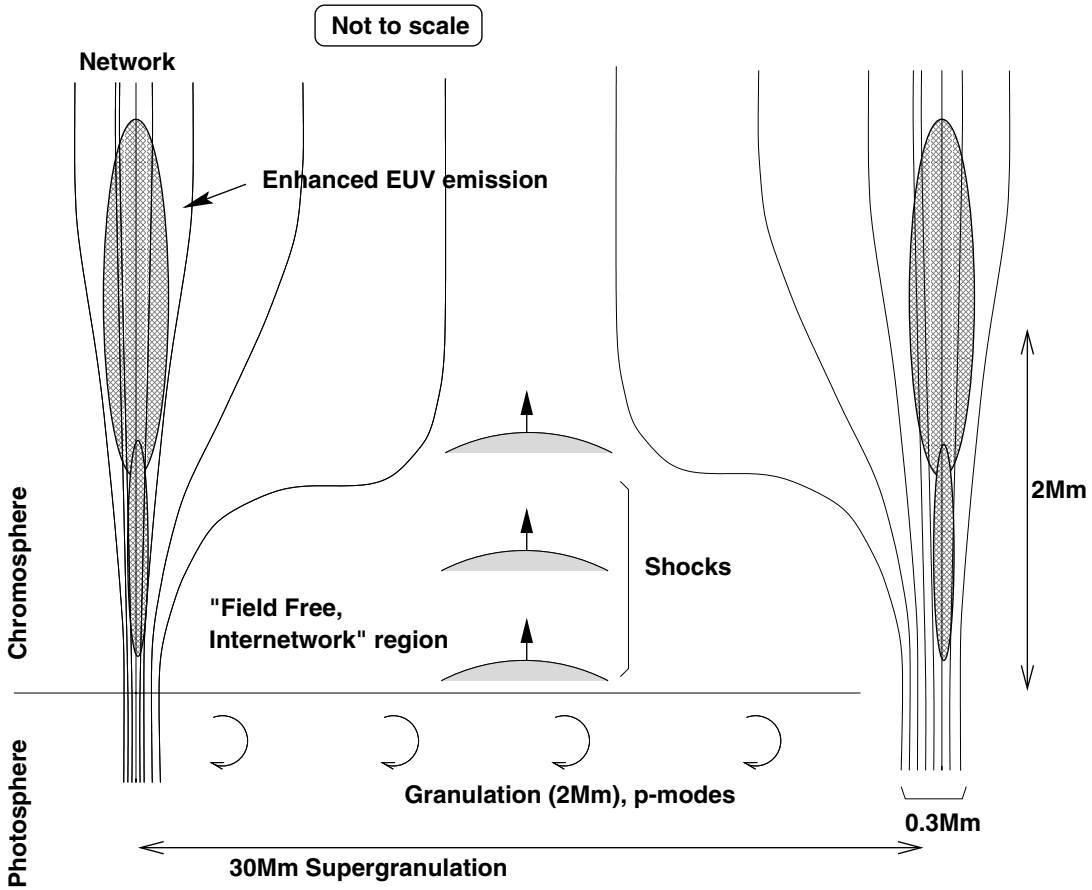


Figure 2. Diagram of the chromosphere (courtesy of P Judge).

Sometimes the fine structures take specific arrangements such as rosettes which have alternately bright and dark fine mottles radiating outward from a common bright center (figure 3). The bright core corresponds to a magnetic element of the network. The visibility of bright Hα mottles fades quickly as the observing wavelength is shifted from the center of the line. At 0.6 Å towards the blue wings of the K line only dark elongated mottles are found. Their intermittent clusters outline the supergranule cells. Seen near the limb at this wavelength position on either blue or red wings, the dark mottles retain their narrow shapes and often form in clusters or bushes.

In radio wavelengths (at 6 cm) the fine-scale network is resolved. All bright 6 cm features overlie bright Ca II K mottles.

The typical lifetime of mottles is 12–20 min. They are the sites of upward and downward motions with shorter lifetimes. A quasi-periodicity of 5–10 min is observed. Downward motions are well observed at the base of the mottles. They are evidently responsible for some bright grains owing to the compression of the plasma or an increase of the magnetic field strength. These downflows explained the common redshift observed in the network. The mottles can be regarded as cylinders of thickness $D = 500$–1000 km and of temperature $T = 8000$–$15\,000$ K. The pressure is about 0.5 dyn cm^{-2} and velocities are between 5 and 10 km s^{-1}. The difference between dark and bright mottles may be explained as due to an increase in gas pressure.

Spicules

At the limb, jet-like structures called spicules are observed extending to the hot corona (figure 4). Spicules have a temperature of 10 000 K and density of 10^{-13} g cm^{-3}. They rise to a maximum height of about 10 000 km with a velocity of about 20 km s^{-1}. Between 5000 and 30 000 spicules are present on the disk at any one time. The large number of spicules combined with their measured velocities implies an upward mass flux two orders of magnitude larger than that of the solar wind. This mass flux is large enough to fill the corona completely in 3 h. How does the mass flow come back to the chromosphere? It is tempting to try to find an answer in the downflow measured in the network and strongly visible at the basis of the mottles. Are mottles of the same nature as spicules? This has also been a subject of long debate and is still a matter of controversy. Nevertheless, spicules have been

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

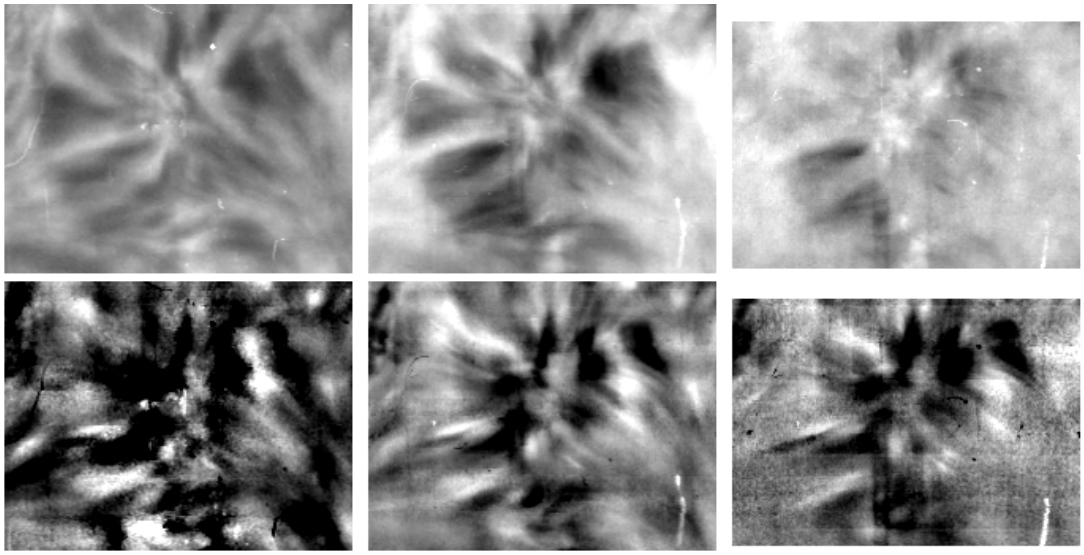


Figure 3. Observations of chromospheric mottles arranged as a rosette observed in the Hα line with the Multi-channel Subtractive Double Pass (MSDP) SPECTROGRAPH (Pic du Midi, Turret Dome). From left to right, the maps are obtained from near the line center to the far wings (H$\alpha \pm 0.25$ Å, ± 0.5 Å, ± 0.75 Å). The upper panels display intensity maps of a zone of dimensions 38 500 km $\times$30 000 km. The lower panels display Doppler-shift maps: dark regions represent downflows, bright regions upflows.

identified in filtergrams obtained for the disk in the far blue wing of the Hα line; which implies the existence of strong upward motions in these structures.

Physical conditions of the quiet chromosphere

The exploration of physical phenomena in the solar atmosphere by means of numerical simulations of atmospheric structure and spectral line formations is an essential complement to high-resolution observational study of the Sun. The principal goal of modeling efforts is to develop diagnostics of physical conditions and processes in the solar plasma. However, it is difficult to achieve this goal because of the variety of velocity fields, the thermal inhomogeneities, the small-scale geometry and the limit of computing resources.

It is universally agreed that the chromosphere is variable and inhomogeneous. Moderate spatial resolution observations in a wide variety of wavelength regions are combined into histograms which indicate the relative frequency of occurrence of different brightness levels. Semi-empirical model atmospheres have then been constructed to match the wavelength behavior of each brightness level. Fully three-dimensional model computational treatment of fine structures is too demanding to be practical, but simpler models in one dimension seem to give useful results. Such one-dimensional atmospheric models are derived for the quiet Sun, the sunspots and other structures. Using these models interpretation of the line profiles is possible and leads to knowledge of the physical parameters of the plasma where the line is formed.

Semiempirical models

To derive quantitative information about the physical parameters from the observed radiation, we have to understand the different processes concerning absorption of the incident radiation and emission by the atoms of the atmosphere. This parameter is called the source function of the plasma. We need to solve the equations of radiative transfer in an atmosphere where the assumption of local thermodynamic equilibrium (LTE) is not always satisfied. As a first step in the modeling the plasma is assumed to be static and the magnetic field is neglected. The non-LTE assumption involves detailed calculation of the statistical equilibrium of each atom (mainly hydrogen, the most important species) with as many atomic levels as necessary. For the basic model of the chromosphere a plane atmosphere is assumed as a simple geometry. A mean 1D semi-infinite atmospheric model is described by the variation in height of the temperature: the temperature minimum is 4300 K at around 500 km; it then rises to a plateau of 6000–10 000 K between 1000 and 2100 km. In the same height range (500–2100 km) the electron density decreases from 2.5×10^{11} to 2.5×10^{10} cm^{-3}, the gas pressure from 1.3×10^{3} to 1.5×10^{-1} dyn cm^{-2} and the density drops from 5×10^{-9} to 5×10^{-14} g cm^{-2}.

To fit the various types of radiation observed from different regions of the chromosphere different 1D models have been computed, e.g. models for intranetwork, network, SOLAR CHROMOSPHERIC PLAGE regions and flares. The brightness observed in Hα flares is explained by a rapid increase of the gas temperature in the chromosphere to nearly coronal temperatures. The brightness of plages is also due to higher temperatures, but for these structures

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

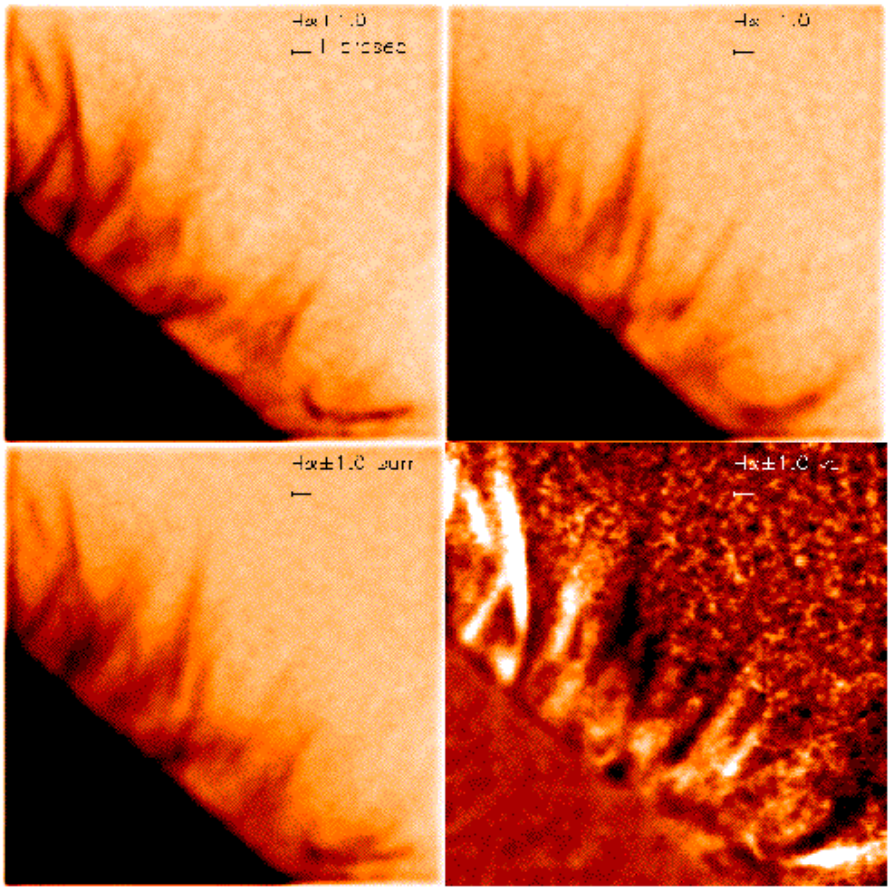


Figure 4. Spicules observed in the Hα line of hydrogen at Sacramento Peak (upper panels, from left to right in the red and blue wings, lower panels the sum and difference of intensities of the wings). The difference represents the velocity. (Courtesy of Skoutchmy.)

the magnetic field plays an important role and cannot be neglected.

Bifurcated models

Another view of the chromosphere is obtained from observations in the CO bands (band of carbon monoxide). These bands are optically thick at the base of the temperature inversion. The core of the strongest CO lines is not centrally reversed, contrary to predictions of the semiempirical models. Instead, the CO features exhibit a cool core which indicates the presence of a low temperature ($T < 4000$ K) above the temperature minimum height. Bifurcated models have been constructed to explain the formation of CO. It is admitted that a part of the chromosphere could be cold enough (see CHROMOSPHERE: THERMAL BIFURCATION).

Such a cool model ($T \sim 4000$ K until 1500 km height) has also been newly proposed in the simulation of intranetwork bright grains (chromospheric oscillations). The semiempirical models may not necessarily reflect an average chromosphere. This is an open question.

Formation of the lines

Semiempirical models nevertheless allow one to explain the central reversal of the H and K Ca II lines. The shape of the line profiles depends strongly on the profile of the temperature through the chromosphere and their regime of formation (by collisions or by radiation). If the line is controlled by collisions, as for example the K line of Ca II, the ratio between absorption and emission (the source function) increases in the chromosphere and the core of the line is self-reversed. If the line is controlled by radiation, as for example the Hα line, the source function is decreasing everywhere and the core of the line is in absorption without reversal.

UV lines are in emission because their source functions are higher than the chromospheric continuum intensity. For the Lyman lines the wings are formed in the chromosphere with the core in the transition zone at still higher temperatures.

At radio wavelengths (millimeter and centimeter), however, the radiation is nearly that of a black body at chromospheric temperatures.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetized atmosphere and fine structures

While the photosphere is dominated by convection owing to its high gas pressure, the chromosphere is well structured by the magnetic field. The plasma is frozen in the magnetic field lines and so outlines the magnetic field structures. This property is useful to visualize the structures. Magnetic flux tubes, as they emerge because of magnetic buoyancy are pushed and concentrated into flux elements of a few hundred km in diameter, as a result of convective motions seen as granules and supergranules and of the so-called convective collapse. This concentration of flux leads to the formation of the chromospheric network, e.g. pores, faculae and sunspots. Photospheric motions and magnetic buoyancy may be the origin of the continued modifications of the coronal magnetic field and the dynamical events observed, i.e. on a small scale reconnection and on the large scale expansion. In the chromosphere the ratio between the gas pressure and the magnetic pressure β ($\beta = 8\pi p/B^2$) is already less than unity in most places, which implies that the magnetic pressure dominates. Flux tubes play an important role in the upward transport of energy and mass into the corona. The chromospheric velocity amplitudes are larger than the photospheric ones and can reach a mean value of 1–1.5 km s^{-1}. This confirms the validity of the continuity equation ($\rho VA =$ constant) in flux tubes with ρ the density, A the tube cross-section and V the velocity of the plasma.

Fine structures overlie the mean chromosphere. A discontinuity in the gradient of physical parameters of the atmosphere does exist and semi-infinite models are no longer valid; cloud-model techniques are more appropriate. In the classical case the structure is supposed to be a relatively thin cloud over a uniform atmosphere, and both of them contribute to the observed profiles. Since the first proposal of such a technique in 1964 by J Beckers, many investigations using such models have been carried out for mottles, waves and filaments.

Using the cloud-model method to analyse spectroscopic data, velocities as large as those of the spicules have been detected in mottles. Spicules should be mottles observed under certain physical conditions such as high speeds. The dynamics in the fine structures is various and large. Many models are proposed to explain the flows in such structures: they are based either on pressure-gradient pulses at the base of the structures or on siphon flows, or are a manifestation of wave motions in flux tubes.

Waves and heating

Oscillations in the chromosphere have been detected quite recently, with significant differences in both power and phase spectra between the intranetwork and network. Oscillations of around 3 min exist mainly in the intranetwork.

Three minute oscillations

The pattern of the intranetwork oscillations is well correlated with the pattern of p-mode oscillations in the photosphere. However, the slow network oscillations are not. They correspond to interferences with shock waves from the 5 min photospheric oscillations and 180 s period pervasive standing oscillations. The latter oscillations have an 80 Mm horizontal wavelength. They are not associated with magnetic flux tubes. The identification of the interiors of network cells with the basal chromosphere has revived interest in the origin of grains which are observed in Ca II H2v and K2v and intermittently present in C I. The 3 min oscillation could be associated with spatially unresolved 'clusters' of grains. Recent simulations of the generation of Ca II K2v bright grains by shocks of waves have been performed quite successfully by M Carlsson. The trigger mechanism of these waves is not yet fully understood.

Low-frequency oscillations

The network power is dominated by low-frequency oscillations with periods larger than 5 min (8–20 min). In bright network regions, at the boundaries of cells, random motions on scales of 2–10 arcsec have been detected. More recently, slow oscillations of 6–7 min have been identified by using the Lyman lines observed with the spectrometer SUMER aboard the space mission SOHO. These waves seem to propagate upwards and they are not spatially coherent. One possibility is that these oscillations are related to the pseudo-periodic behavior of mottles, or spicules, since they can produce such wave-trains. The bright features strong in the Hα far wing or in Ca II would correspond to the base of mottles and would be therefore of a magnetic nature.

Heating

Many oscillations are observed in the chromosphere. Are these waves heating the chromosphere? When acoustic waves travel upwards through the solar atmosphere, the density gradient results in an increase of the wave amplitude (V), so that the product ρV^2 remains roughly constant. Taking into account the observed velocities at low levels, it is possible to infer the height at which the amplitude reaches the sound velocity and thereby leads to a shock-wave formation. Shock waves can dissipate mechanical energy into heat by non-linear processes. Extensive calculations have been performed to explore the efficiency of such processes. It has been shown that heating by shock waves can be efficient in the low chromosphere, but at higher levels the acoustic waves should become evanescent and the available energy becomes too small to compensate for the radiative losses which are deduced from the observed spectrum. The recent discovery of bright grains has been explained in terms of the simulation of waves traveling in the chromosphere.

The active chromosphere

The active chromosphere involves all the manifestations of the solar activity observed in chromospheric lines (table 1). The solar activity is directly related to the magnetic activity and is manifested by the existence of active centers and newly emerging magnetic flux.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

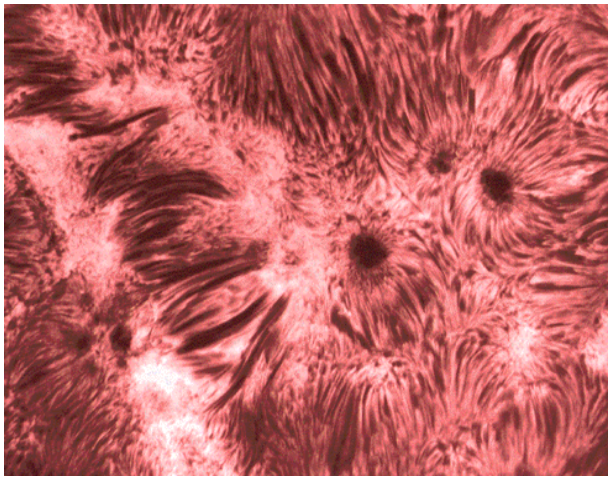


Figure 5. Chromospheric fibrils observed around sunspots and arch filament system in a active center (Pic du Midi, Turret Dome).

Active center

An image obtained in the Hα line at Pic du Midi (figure 5) shows the fibril pattern surrounding a sunspot. These structures are tied into the penumbra of the spot. They can be radially oriented or turning either clockwise or counterclockwise. It was recently discovered that the fibril pattern follows a clockwise orientation in the southern hemisphere and counterclockwise in the northern hemisphere. This helicity law seems to be important in the global view of the Sun's activity.

Some dark structures emerge with large velocities from the sunspot. They were discovered by Ellerman and are called Ellerman bombs (see SOLAR CHROMOSPHERE: ELLERMAN BOMBS). They are characterized by very broad emission profiles in the Hα line.

In an active center bright regions are observed, the so-called plages, or faculae. In old groups the sunspots disappear first and only faculae remain, which subsequently disperse and vanish.

Relation to the solar cycle

The normalized distribution of areas of active regions, using Ca II K plages or magnetic flux, varies with the phase of the solar cycle. The proportion attributed to plages and strong network in the total magnetic flux change varies by different amounts across the cycle, e.g. by a factor of 8 for plages, a factor of 40 for strong network regions and a factor of 10 for an activity complex.

The total enhanced network and plage signal is larger during active periods. The active network accounts for a large part of the total solar irradiance variation. This apparent solar cycle variation of enhanced intensity contrast is, nevertheless, only 2% for the network and 9% for plages.

Newly emerging magnetic flux

New magnetic flux emerges continuously and the full magnetic pattern of the Sun is renewed within 40 h. The emergence of flux has different consequences in the upper atmosphere and different signatures according the place where it appears. When new flux emerges at the center of supergranules it moves toward the edges and interact with the network magnetic flux. In the chromosphere the signature of an emerging dipole is an arch filament system. These arch-shaped tubes rise, reach an altitude of around 5000 km, bringing up denser material than in the surrounding atmosphere. At both ends of the arches the material falls owing to gravity with velocities of about 20 km s^{-1} which cannot be, nevertheless, considered as free-fall motions. In Hα the dense material appears as absorption features but in UV (SOHO) or in soft x-ray (YOHKOH the Japanese satellite) it appears in emission and brighter than the surrounding material.

Otherwise, the dynamic phenomena discussed above are related to solar activity. Twisting and shearing of flux tubes may set up currents along the field lines. On large scales differential rotation will induce an extended current system which increases continuously, leading to an outer expansion of magnetic field lines, which is visible with Yohkoh. The effective electrical conductivity of the solar corona is so high that the field structures resulting from flux emergence cannot immediately relax to a current-free state. The presence of currents can lead to the formation of different structures such as prominences, with any disruptions leading to eruptions such as SOLAR CORONAL MASS EJECTIONS or FLARES.

Bibliography

Bray R J and Loughhead R E 1974 *The Solar Chromosphere* (London: Chapman & Hall)

Lites B.W 1984 *Chromospheric Diagnostics and Modelling* (NSO Observatory Publishing)

Rutten R J and Schrijver R J 1994 *Solar Surface Magnetism* (Dordrecht: Kluwer)

Schmieder B 1997 Flows through the magnetized structured solar atmosphere *Lecture Notes in Physics* ed G Simnett, C Alissandrakis and L Vlahos (Berlin: Springer) p 139

Tsiropoula G, Alissandrakis C E and Schmieder B 1993 The fine structure of a chromospheric rosette *Astron. Astrophys.* **271** 574

Vernazza J E, Avrett E H and Loeser R 1981 Structure of the solar chromosphere *Astrophys. J.* **45** 635–725

Brigitte Schmieder

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Emerging Flux Regions

A description of the emergence of large concentrations of magnetic fields (active regions) at the solar surface is presented. The dark structures known as SUNSPOTS are a consequence of this process.

The emergence of an active region on the solar surface is one of the most prominent processes that can be followed by looking at the Sun on a daily basis. Regions that before emergence had the normal bubble convective appearance (GRANULATION) will in a few days develop dark sunspots and SUNSPOT PORES and bright FACULAE. At visible wavelengths (very often in the core and wings of the red Hα line) we can observe active region emergence in the deepest observable layers, the SOLAR PHOTOSPHERE and CHROMOSPHERE. Nowadays, we can also follow this process in the UV and x-ray spectral range from space, where we track active region emergence in the outermost layer of the Sun, the CORONA. It is in the chromosphere and in the corona where all the energetic processes relating to flux emergence can be followed more easily. The interaction of the newly emerging flux and preexisting field lines usually triggers a large number of FLARES (the sudden, transient release of energy probably related to magnetic reconnection). The dynamics observed in all these wavelength ranges clearly shows that active region emergence should be understood as the result of buoyancy rise of a magnetic structure from deeper layers in the Sun.

The magnetic field lines that form an active region are thought to be generated by some kind of DYNAMO mechanism (like almost all cosmic magnetic phenomena). The exact location of this dynamo within the Sun is not clearly established, but helioseismic studies suggest that the generation and storage of the dynamo fields take place close to the base of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE). On the solar surface, the emergence of localized, vast amounts of magnetic fields suggest that, at the bottom of the convection zone, the field is organized into isolated magnetic flux tubes or ropes. These magnetic flux tubes are thought to be formed as toroidal loops in the interior of the Sun (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES). Their stability there remains controversial but it is clear that they develop a strong tendency to rise (at least, in part) towards the surface. This buoyancy comes from the extra pressure exerted by the magnetic field that reduces the gas pressure inside the loop (pressure equilibrium is a very efficient mechanism which can be assumed for these tubes). The lower gas pressure can be achieved only through lower temperatures and/or gas densities inside them. Thus, we are led to consider magnetic tubes, which, having lower densities, will develop some uncompensated buoyancy and rise (similar to bubbles in water, but with other physical agents playing an important role). The irruption of these magnetic bubbles into the solar atmosphere represents the emergence of an active region.

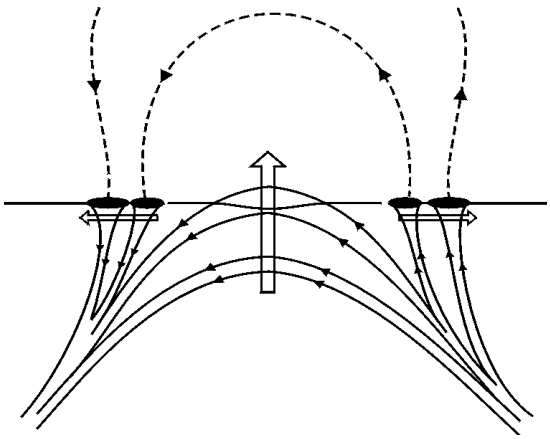


Figure 1. Illustration of the structure of an emerged Ω loop on the solar surface. The arrows indicate the proper motion of the field lines. The pores at the footpoints (black ellipses) separate from the central emerging region, where new horizontal flux is emerging. The dashed curves represent previously emerged field lines and already fill the coronal volume (from Zwaan 1992).

An active region usually has a clear dividing line between the two polarities (by polarities we mean here magnetic fields with their vectors pointing out of the Sun and fields pointing into the Sun). After emergence, the two polarities are seen at the surface separated in longitude. Magnetic field lines start at one polarity (figure 1, right), arch above the solar surface filling the solar corona (as seen from spacecraft such as Yohkoh and SOHO) and end at the other polarity (figure 1, left). In the deep (unobservable) layers, the field lines of both polarities dive down and connect with the system of toroidal flux tubes. This configuration of the field lines after emergence is called the Ω-shaped loop configuration by solar astronomers. An idealized sketch is shown in figure 1.

Observational description of emerging flux regions

The emergence of active regions in the solar atmosphere is a relatively easy phenomenon to spot in any of the spectral bands used to monitor the Sun. This is particularly true if we compare it with the more elusive reverse process of active region disappearance. Already in the late 1960s Bruzek (1967) presented a classical description of the emergence process as seen in the core of Hα (i.e. in the chromosphere). He coined the term arch filament system (AFS) to describe the system of parallel loops that connect the footpoints of an emerging flux region (EFR). These arching loops have since then been taken as the signature of the birth of an active region (see figure 2; a description of this EFR can be found in Strous (1994)). As seen in the chromosphere, the central portion of the AFS shows ascending velocities of up to 10 km s^{-1}, reflecting the upward buoyancy-driven motion of the magnetic system. Closer to the footpoints of the EFR, the loops show

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

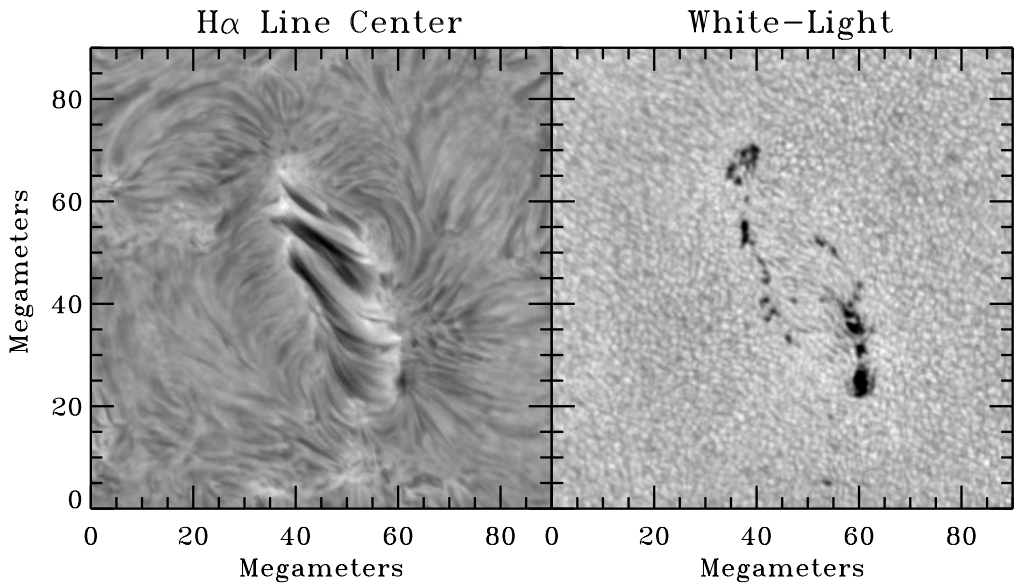


Figure 2. An EFR seen in the Hα line center (left) and white light (right) The footpoints of the EFR correspond to the two almost vertical alignments of pores. In between can be seen elongated intergranular lanes in the white-light image and the AFS in the Hα picture (the light and dark structures arching from one footpoint to the other). The images were taken at the Swedish Vacuum Solar Telescope on La Palma (Canary Islands, Spain). Image courtesy of L Strous/Swedish Vacuum Solar Telescope, Royal Swedish Academy of Sciences.

receding velocities of several tens of km s^{-1} produced by matter falling back along the magnetic field lines. During the process of emergence, the footpoints separate at a rate of a few km s^{-1} while new flux appears in between. The loops composing the AFS are in constant change (lifetimes of individual filaments $\approx$30 min) while they interact with new field lines rising from below. After a few hours, the flux system at the footpoints is organized into several pores, dark structures seen in white light where the normal granular appearance of the photosphere is absent. The magnetic field in these structures is able to impede the normal heat transport by convection (i.e. the granulation) and the reduced temperature of the gas makes them appear dark in contrast to the surrounding atmosphere. In the course of the days following the first sighting of the AFS, several bursts of magnetic flux emergence can occur and, quite often, it is in these later bursts where the bulk of the magnetic flux makes its way through the photosphere. The pores formed from all the different emergence episodes can eventually coalesce to give rise to a sunspot. This process is driven by the subsurface topology of the flux loop more than by any flow that may prevail there. This usually occurs in the leading (with respect to solar rotation) footpoint of the EFR and a stable round sunspot is observed. In the following footpoint, if a sunspot forms at all, it usually corresponds to a less stable (shorter-lived) structure with an irregular shape. This difference between the leading and following footpoints is very well documented but its origins are far from being fully understood. The AFS is seen while new flux emerges from below. This usually takes no more than 4–5 days.

Flux emergence as described above is consistent with the idea of an Ω-shaped magnetic loop that is rising through the solar convection zone (Zwaan 1992). During this rise, the loop passes through a highly stratified external atmosphere and interacts with convective motions. All this probably leads to the break up of the single initial loop into several loops (as depicted in figure 1). At the surface, each of them produces an emergence episode. The fact that different episodes can later give rise to pores that merge into a single sunspot indicates that deeper down the flux system has retained its identity.

Observations of EFR were first made in the chromosphere owing to the accessibility of the process as seen in Hα. Nowadays, the EFRs have also been followed in the photospheric layers, where we find dark alignments between the footpoints with elongated granules showing how the magnetic field perturbs the normal granulation seen elsewhere (see the white-light image in figure 2). Spectropolarimetric observations in photospheric lines have revealed downflows of up to 2 km s^{-1} close to the footpoints. These downflows are photospheric counterparts of the much stronger chromospheric downdrafts (due to the lower densities) mentioned above. In the photosphere, upflowing velocities of the central parts of the EFR have been detected that reach values of 1 km s^{-1}. It is clear that the field lines speed up as they rise through these layers to reach even higher upward velocities in the chromosphere.

Recent photospheric observations made with the Advanced Stokes Polarimeter (ASP; Lites *et al* 1998) have clarified several issues that had remained unanswered

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

before the use of high-sensitivity full Stokes polarimetry. This instrument allows reliable estimates of the mean vector magnetic field within the resolution element and the area fraction occupied by the field. These observations have clearly detected the horizontal (transverse) field lines as they pass through the photosphere, representing the signature of the top part of the emerging loops. The intrinsic field strength measured at the locations of the horizontal field lines is in the range 200–600 G, i.e. at or below equipartition with the kinetic energy of the granular motions. This result was expected from the fact that these fields are able to disturb the granulation pattern (Zwaan 1992). Fields in the range of 1200 G, or higher, are seen only in the vertical footpoints, in the form of either faculae or pores. It has been argued that the stronger fields found in the footpoints are achieved after the convective collapse process driven by the ubiquitous downflow found there. While this may well be the case, ASP observations have shown that the downflow is seen over regions that are already in the kilogauss regime; thus the downflow itself cannot be taken as the signature of convective collapse. The only way to ascertain whether this process is responsible for the high (twice equipartition) fields found in the vertical footpoints is by identifying a region of vertical weak field strength where a downflow develops and, after a short time interval (minutes), develops kilogauss strengths. This has not yet been observed and it represents an important challenge of modern spectropolarimeters in the present solar cycle.

Complex patterns of emergence

The Ω-shaped loop configuration of an active region described so far represents a simple bipolar system that describes most of the emergence cases observed on the Sun. This configuration fits in a natural way with Hale's polarity law. This law states that in a given solar hemisphere, during one solar cycle, the polarity of most of the active regions is oriented such that the positive polarity (field lines pointing out of the Sun) precedes the negative polarity (field lines pointing into the Sun). The reverse order would be found in the other hemisphere. The bipolar character of the Ω-shaped structures in each hemisphere will give rise to Hale's law. However, a small percentage of the emergence cases (around 1%) cannot be described by this bipolar configuration. Some active regions develop what is called a δ configuration, in which a single spot (the δ spot) contains within its penumbrae two umbral cores that have opposite polarities. The order of the two polarities is generally opposite to that predicted by Hale's law (in the above example the negative-polarity umbral core will precede the positive one). This δ-spot configuration, although rare, is very important because it represents the most flare-productive regions observed at the surface. δ groups are responsible for almost all great flares. The existence of complex neutral lines (where the polarity changes sign) seems to favor the occurrence of very energetic processes.

We have recently started to understand (Linton *et al* 1999) that this complex configuration may result from the evolution of toroidal loops as they pass through the solar convective zone. Some configurations of the magnetic field in these loops are liable to develop a twist of more than 90° in the top part of the Ω loop, thus inverting the polarity order. This kink instability develops during the rise of the loop. It will thus take place only if the timescale of the instability is shorter than the rise time. The inverted portion of the loop can turn into a compact sunspot displaying the two polarities at the surface. In some cases, the evolution can be complex enough (through field-line reconnection) to generate closed magnetic systems that ascend through the solar surface (Lites *et al* 1994).

Bibliography

Bruzek A 1967 On arch-filament systems in spotgroups *Sol. Phys.* **2** 451–61

Linton M G, Fisher G H, Dahburg R B and Fan Y 1999 Relationship of the multimode kink instability to δ-spot formation *Astrophys. J.* **522** 1190–205

Lites B, Low B C, Martínez Pillet V, Seagraves P, Skumanich A, Zoe F, Shine D and Tsuneta S 1994 The possible ascent of a closed magnetic system through the photosphere. *Astrophys. J.* **446** 877–94

Lites B, Skumanich A and Martínez Pillet V 1998 Vector magnetic fields of emerging solar flux I. properties at the site of emergence *Astron. Astrophys.* **333** 1053–68

Strous L 1994 Dynamics in solar active regions: patterns in magnetic-flux emergence *PhD Thesis* Utrecht University

Zwaan C 1992 The evolution of sunspots *Sunspots: Theory and Observations (NATO ASI Series vol 375)* ed J Thomas and N O Weiss, pp 75–100

V Martínez Pillet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Fibrils

Fibrils are dark, elongated, curvilinear patterns in the CHROMOSPHERE as seen through a filter that isolates the Hα spectral absorption line of hydrogen. Well-developed fibrils occur in and around SOLAR ACTIVE REGIONS (figure 1) and SOLAR FILAMENT CHANNELS.

Like much of the traditional nomenclature used to describe the solar atmosphere, the term fibril is derived from Latin ('little fiber') and originates from the rich but limited view afforded by Hα images. Modern research on fibrils aims to relate their appearance to physical properties of structures in the atmosphere. An important component of this goal is to understand the relationship between fibrils and the fine structure of the quiet chromosphere. As seen in Hα, the quiet chromosphere is dominated by bright and dark mottles, structures which, although often elongated, are characteristically shorter in length, shorter-lived and show less large-scale organization than fibrils.

The prevailing hypothesis is that the magnetic field is the most important determinant of chromospheric fine structure. The difference in Hα appearance between quiet and active regions should therefore reflect differences in the organization of the field. There is considerable indirect evidence to support this hypothesis and none that contradicts it; but there are still few direct measurements of the magnetic field in the chromosphere.

Appearance

Traditionally fibrils are identified with dark features; but, as is apparent from figure 1, brighter and darker Hα features are closely associated. Even when the orientation of fibrils in a particular area is clear, it is often difficult to follow an individual dark feature for more than a few thousand kilometers. Large fibrils extend as far as 30 000 km; these are sometimes called chromospheric threads. The apparent width of fibrils ranges from 2000–3000 km down to the limit of resolution of ground-based images, about 300 km. Fibrils probably have structure or substructure on even finer scales, as the best images hint. The apparent aspect ratio (length to width) of fibrils ranges from about 3–4 (a less elongated feature would probably not be called a fibril) to as high as 30–40.

When an active region is born as an emerging flux region, the short fibrils that appear are called arch filaments (see CHROMOSPHERE: EMERGING FLUX REGIONS). The dark fibrils are intermixed with features that are distinctly brighter than the background chromosphere. Since emerging flux regions are always magnetically bipolar, and the opposite ends of an arch filament point to the opposite polarities, arch filaments support in a simple geometry the general hypothesis that fibrils follow magnetic lines of force. As the active region develops, the single arch filament system is replaced by a growing welter of fibrils that reflects the increasing size and magnetic complexity of the region. Sizable SUNSPOTS are surrounded by fibrils that begin in the white-light penumbra but often extend, in a superpenumbra, well beyond the penumbral boundary (see SUNSPOT PENUMBRA: STRUCTURE AND ACTIVITY). Superpenumbral fibrils in a complex active region are often strongly curved and non-radial. This is to be expected if fibrils follow the magnetic field. When the photospheric magnetic field is complex, even a potential field above the photosphere may be curved and non-radial around flux concentrations; when the photospheric field is highly non-potential, as it is often measured to be, strong curvature and 'S' shapes are expected. Thus, simple fibril patterns leading from one sunspot to another are the exception rather than the rule. Moreover, there are typically multiple episodes of flux emergence as the region grows. The fibril pattern fragments as magnetic field lines from newly emerged flux reconnect with existing flux.

A filament or filament channel may form within or on the boundary of the active region. Fibrils are indistinct within the channel; to either side, fibrils are aligned with the long axis of the channel. Measurements of the horizontal magnetic field within filaments indicate that the field runs predominantly along the axis of the filament, although the filament separates areas of opposite magnetic polarity at the photospheric level.

Long fibrils may form around the boundary of an active region, particularly in the directions of adjacent active regions, suggesting again that the fibrils follow magnetic connections.

A decaying active region is often dominated by a single large and regular sunspot. Close to the spot, its fibrils extend nearly radially outward. Distinct superpenumbral fibrils may extend twice as far as the penumbral boundary and occasionally exhibit a vortical pattern due to interaction with the solar rotation.

A large fibril often appears to obscure brighter structures, suggesting that the fibril is higher in the solar atmosphere. In this respect fibrils are analogous to dark mottles in quiet regions. It is difficult to give a range of heights for fibrils. Height is difficult to measure in individual cases, varies along a fibril, and varies between superpenumbral fibrils, short fibrils in magnetically complex areas, and long fibrils that could as well be called active-region loops. The appearance of fibril systems seen in perspective near the limb suggests that fibrils are elevated by 3000 km or more above the background chromosphere. Arch filament systems are seen to extend to heights of 10 000–15 000 km.

Dynamics

An arch filament system lasts for 1–3 days. An individual arch filament has a shorter lifetime, about 10–30 min. The top of an arch filament system rises with an apparent velocity $v \lesssim 20$ km s^{-1}, consistent with its identification with emerging magnetic flux. Material appears to flow downward along the arches with $v \approx 30$–50 km s^{-1}.

The general fibril pattern of a well-developed region evolves gradually, over a period of hours, and can remain recognizable for days. In this respect fibrils are more stable than dark mottles in quiet regions. The natural

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

interpretation is that the overall magnetic structure of an active region is determined by large and relatively long-lived concentrations of magnetic flux, whereas dark mottles surround small flux concentrations, usually associated with the CHROMOSPHERE NETWORK, that change on timescales of minutes to hours. However, individual fibrils have a shorter lifetime than their geometrical pattern, ranging from about 1 min for the smallest fibrils to 20 min for the largest.

The superpenumbral fibril pattern of an isolated sunspot changes even more slowly than the fibril pattern of a mature active region. The lifetime of an individual superpenumbral fibril is about 3–40 min, somewhat longer than the lifetime of a fibril in a younger active region but still much shorter than the lifetime of the pattern. The most characteristic dynamic feature of superpenumbral fibrils is a material flow along their length, inward and downward, of 5–20 km s^{-1}.

This inverse Evershed effect in the chromosphere contrasts with a slower outward flow in the penumbra at the level of the photosphere (see SUNSPOTS: EVERSHED EFFECT). Chromospheric outflow also occurs in a minority of superpenumbral fibrils.

It is difficult, using ground-based observations disturbed by atmospheric turbulence, to decide whether there is a one-to-one correspondence between the flow field and Hα brightness. Some high-quality observations suggest that flow channels usually coincide with fibrils, but not precisely or always. Like fibrils, individual flows often change beyond recognition in a matter of minutes, whereas a flow pattern can persist for hours. Active-region chromospheric flows not associated with sunspots have been studied less thoroughly than penumbral and superpenumbral flows.

Interpretation

Why are fibrils dark? How high are they in the solar atmosphere? How hot and how dense are they? What is the direction and strength of the magnetic field within them? How do these physical quantities compare with their values outside a fibril at the same height?

Fibrils have been recognized and studied for many years, yet the answers to these questions are surprisingly tentative for two main reasons: because fibrils are defined by their appearance in Hα, and it is easier to acquire high-quality images in this absorption line than it is to interpret them; and because it is difficult to measure magnetic fields in the chromosphere.

The quantitative interpretation of Hα images can be framed in terms of a contrast profile, $C(\lambda) = (I - I_0)/I_0$, where $I(\lambda)$ is the Hα intensity profile in the fibril and $I_0(\lambda)$ is a reference profile representing the inter-fibril chromosphere. Although the contrast profile is not a necessary construct in the task of reproducing $I(\lambda)$, it is conceptually useful for two reasons. First, it captures a defining characteristic of a fibril: that it be dark *relative* to its surroundings in the core of Hα. Second, if fibrils are indeed higher in the atmosphere than inter-fibril regions, it may be possible to approximate the fibril as a cloud-like feature that is illuminated from below by a background radiation field that is close to that of the average chromosphere. Most analyses of Hα contrast profiles employ (or critique) some variant of the cloud model described below.

Several observational considerations apply to contrast profiles. First, the wavelength resolution should be a small fraction of the width of the Hα line, which is about 1.5 Å (full width at half depth). Spectrographs and the best narrowband filters routinely achieve 100–125 mÅ (full width at half maximum transmission), but many Hα images are acquired with filters that have a bandpass of 0.25–0.5 Å. A theoretical contrast profile must be convolved with the filter bandpass before comparison with such images. Second, filtergrams should be acquired at more than one wavelength; otherwise it will not be possible to separate changes in the shape of the line profile from wavelength shifts due to the Doppler effect. The images at different wavelengths should not be separated in time by more than a minute or so, lest the fibril undergo significant evolution during the sequence. Third, the comparison between theoretical and observed contrast profiles should, ideally, take the effects of atmospheric seeing and instrumental stray light into account—but this is difficult. Finally, the concept of a background profile is only a first approximation. The inter-fibril atmosphere is not necessarily like the 'quiet' photosphere, nor constant from place to place.

The basic cloud model (attributed to J Beckers) follows the transfer of spectral line radiation through a discrete feature:

$$I = I_0\,\mathrm{e}^{-\tau} + \int_0^{\tau} S\,\mathrm{e}^{-t}\,\mathrm{d}t$$

where $I_0(\lambda)$ and $I(\lambda)$ are the intensities entering and leaving the feature, $S(\lambda)$ is the line source function, and $\tau(\lambda)$ is optical depth. Detailed nonlocal thermodynamic equilibrium (NLTE) analysis of the formation of the solar Hα line has established that, within 2–3 Doppler widths of line center: (a) the absorption coefficient has a Gaussian (Doppler) profile; and (b) the source function is nearly independent of wavelength, reflecting the dominance of non-coherent scattering. The first result, together with an assumption that the Doppler width of the absorption coefficient is constant along the line of sight, means that the optical depth will also have a Gaussian form

$$\tau(\lambda) = \tau_0 \exp[-(\lambda - \Delta\lambda_v)^2/\Delta\lambda_D^2]$$

where λ is measured from the center of the reference profile I_0, $\Delta\lambda_D$ is the Doppler width, and $\Delta\lambda_v$ is the line-of-sight velocity of the feature with respect to the reference. The second result, together with an assumption that the source function is constant along the line of sight through the cloud, allows the transfer equation to be integrated to yield

$$C(\lambda) = \left(\frac{S}{I_0(\lambda)} - 1\right)(1 - \mathrm{e}^{-\tau(\lambda)}). \qquad (1)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

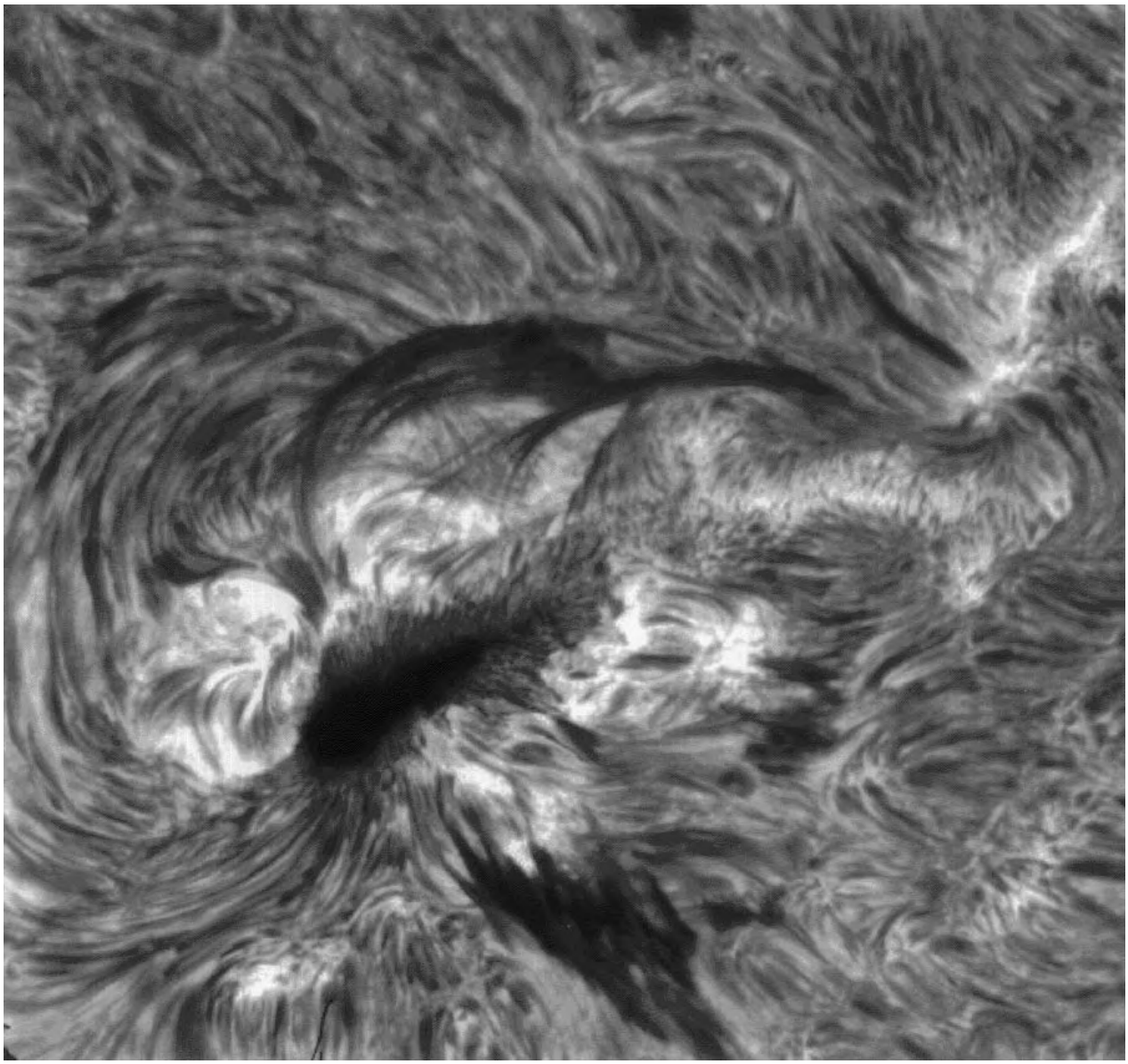


Figure 1. Chromospheric fibrils in a solar active region. Seen through a filter that isolates the core of the Hα absorption line, fibrils appear as thin, elongated dark features up to 30 000 km long, intermixed with brighter features (the area shown is about 200 000 km on a side). The largest and darkest features, above and below the sunspot, are called active prominences or filaments rather than fibrils. Image from Big Bear Solar Observatory.

The model is applied by adjusting the parameters S, τ_0, $\Delta\lambda_D$ and $\Delta\lambda_v$ through nonlinear least-squares fitting to achieve best agreement between equation (1) and the observed contrast profile.

A number of questions naturally arise concerning the applicability of the cloud model. Can it produce contrast profiles that agree with observations? Is it physically consistent? Is it realistic? What is learned from the fitted parameters?

Figure 2 illustrates the variety of contrast profiles that the cloud model yields for various reasonable combinations of the parameters. The contrast depends on wavelength in the line and can be negative (dark) or positive (bright). The profile is asymmetric if the cloud has a non-zero line-of-sight velocity.

The cloud model has been more extensively applied to structures in the quiet Sun (bright and dark mottles) than to fibrils and loops in active regions. The majority of observed contrast profiles of mottles can be adequately fitted by the model. Figure 3 gives an example of the model fitted to observations of an active-region loop.

The form of equation (1) does place two simple

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

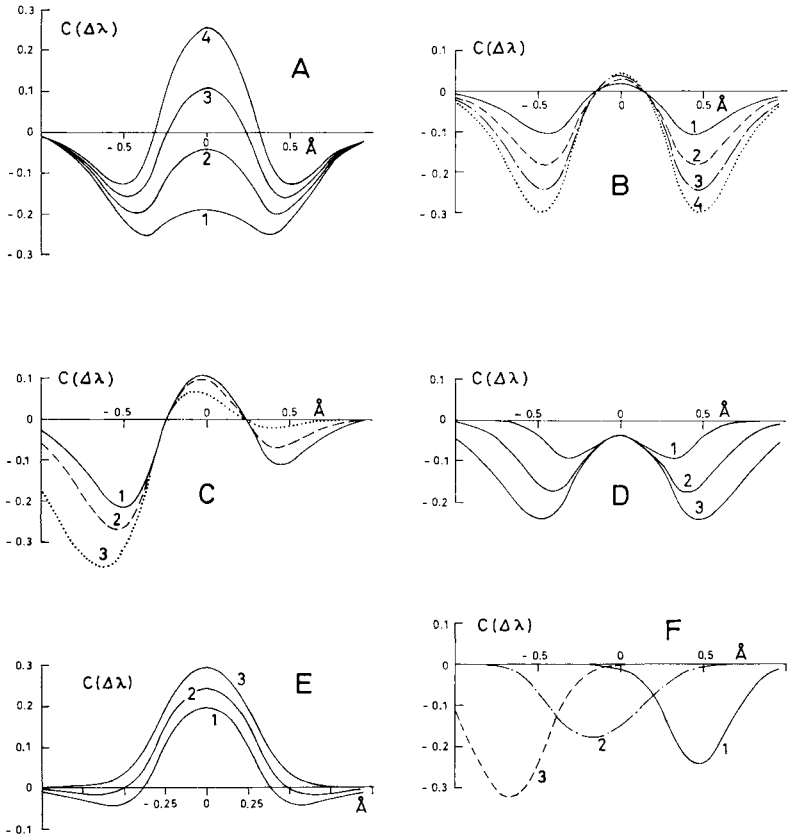


Figure 2. Theoretical contrast profiles from the cloud model. Each panel shows the effect of varying one parameter while holding the others constant. A: varying S/I_0. B: varying τ_0. C: varying $\Delta\lambda_v$. D: varying $\Delta\lambda_D$. From Grossman-Doerth U and von Uexküll M 1971 *Solar Phys.* **20** 31–46.

restrictions on the theoretical contrast profile. First, if $S < I_0(0)$, then $S < I_0(\lambda)$ for all λ, because the reference profile has its minimum intensity at $\lambda = 0$. Thus, if a feature is dark at unshifted line center, then it must be dark everywhere in the line. For fibrils, this seems to be true within the observational uncertainties. Second, since

$$\frac{C(+\lambda)}{C(-\lambda)} = \frac{1 - e^{-\tau(+\lambda)}}{1 - e^{-\tau(-\lambda)}}$$

and the right-hand side is always positive, the contrast at equal wavelength displacements from unshifted line center must have the same sign. This condition is automatically satisfied for features that are dark at all wavelengths; it is sometimes observed to be violated for other types of features.

In short, a cloud model can be reasonably fitted to most dark features in Hα. However, this does not ensure that the model is physically meaningful. A basic objection is that, for any particular geometry of the absorbing structure with respect to the underlying atmosphere, the source function is determined by the optical depth, Doppler absorption width and bulk velocity of the gas in the structure. Thus, it is not physically consistent to fit the parameters of the cloud model independently. When the cloud model is compared with treatments in which the source function is derived from NLTE theory, it is found that the results agree fairly well for 'high' clouds—those that subtend an angle much less than one radian as seen from the base of the background chromosphere. For 'low' clouds, the cloud model is unreliable: it may not fit at all, and even when it does, the values of the fitted parameters need not correspond to their physical values. A low cloud violates a basic assumption of the model, that the clouds have no back-reaction on the underlying chromosphere.

Other assumptions of the cloud model (constancy of the source function, Doppler width and bulk velocity along each line of sight, and global invariance of the reference line profile) are not as fundamental and can be relaxed in the form of multi-cloud or differential cloud models.

Whatever model is fitted to the observations, the goal is to estimate the values of physically interesting parameters. From Hα observations, the gas density and pressure can be estimated fairly well, but the kinetic temperature is poorly constrained.

The profile-integrated absorption coefficient in the Hα

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

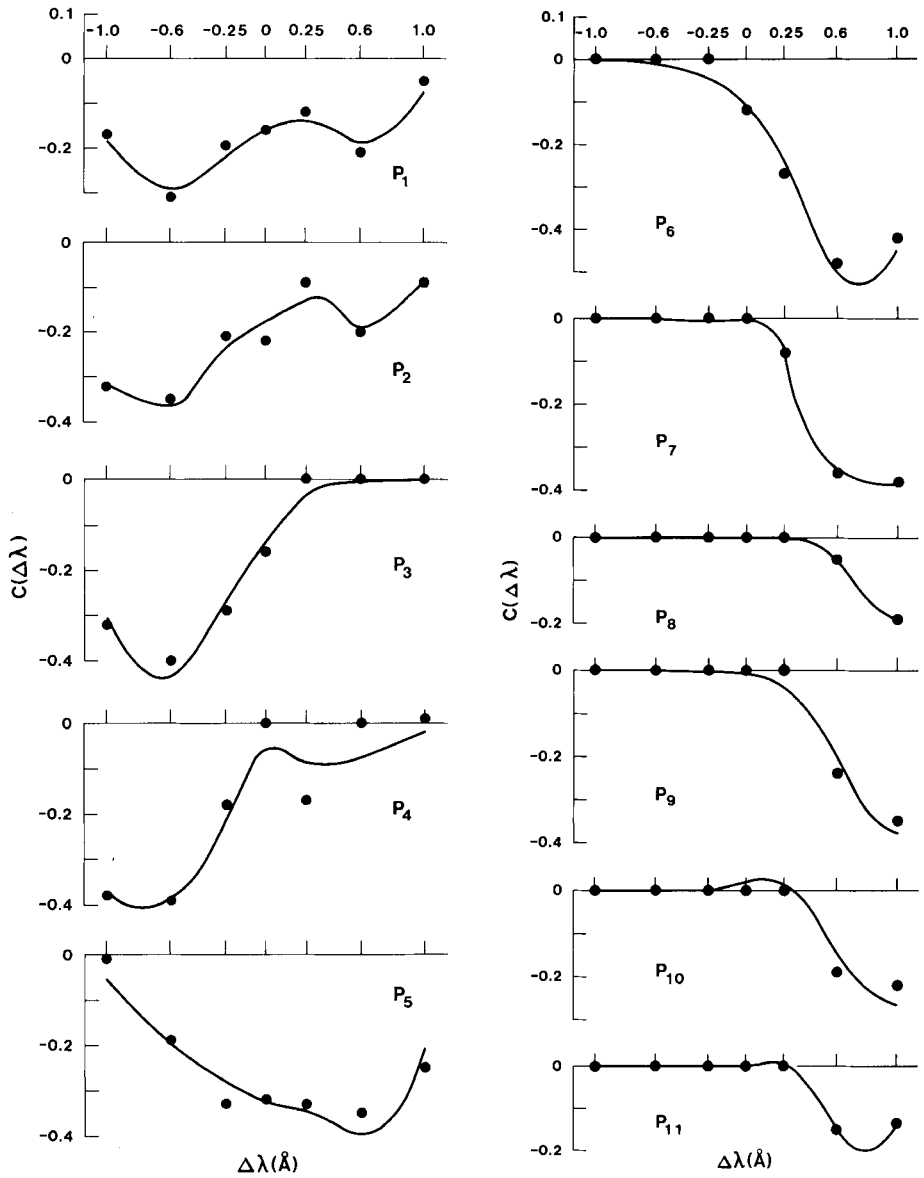


Figure 3. Observed contrast profiles (points) and fitted cloud models (curves) for an active-region loop. (From Bray R and Loughhead R 1983 *Solar Phys.* **85** 131–140.)

line (neglecting stimulated emission) is given by

$$\kappa_\lambda = \frac{\pi e^2}{mc}\frac{\lambda^2}{c} f_{\rm abs} N_2$$

where e and m are the charge and mass of the electron, $f_{\rm abs}$ is the absorption oscillator strength, and N_2 is the density of absorbers (here, the second level of hydrogen). If this absorption is distributed over a Doppler profile, the absorption coefficient at the center of the profile is

$$\kappa_0 = \frac{\pi^{1/2} e^2}{mc^2}\frac{\lambda^2}{\Delta\lambda_D} f_{\rm abs} N_2.$$

Within the cloud approximation, $\tau_0 = \kappa_0 L$, where L is the thickness of the structure along the line of sight, and, inserting numerical values

$$N_2 = 7.3 \times 10^7 \tau_0 \Delta\lambda_D / L \ {\rm cm}^{-3}. \qquad (2)$$

A cloud model gives τ_0 and $\Delta\lambda_D$. The thickness L is estimated from the observed transverse dimension of the fibril (atmospheric seeing and substructure along the line of sight can both contribute to an estimate higher than the true value). Equation (2) than leads to a typical range $N_2 = (1\text{–}5) \times 10^4$ cm^{-3} for a variety of dark features including fibrils. NLTE models of gas slabs irradiated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by photospheric and chromospheric radiation consistently yield an approximate relationship between N_2 and N_e, the electron density: $N_e = 3.2 \times 10^8 N_2^{1/2}$. Applying this relationship to the range just given for N_2 yields a range $N_e = (3\text{–}7) \times 10^{10}$ cm^{-3}. The total number density of hydrogen (neutral plus ionized), N_H, can be derived if the ionization fraction $x \equiv N_e/N_H$ is known. NLTE models yield values of x as small as 0.4 for thin ($L \approx$ 500 km), cool ($T \approx 7000$ K) slabs to essentially complete ionization for temperatures much above 10^4 K, leading to a derived range $N_H = (0.3\text{–}2) \times 10^{11}$ cm^{-3}. As the ambient density at heights greater than a few thousand kilometers is $N_H \approx 10^9$ cm^{-3}, fibrils are typically much denser than their surroundings. The derived gas pressure, $p_g = N_e kT(1 + 1.085/x)$, has a smaller range because of the opposing effects of temperature and ionization, $p_g = 0.1\text{–}0.3$ dyn cm^{-2}. All these ranges should be treated cautiously, as encompassing the majority, but certainly not all, structures that could be called fibrils.

It is not surprising that the temperature within fibrils is poorly constrained by Hα observations because the source function is dominated by photoionization due to the external radiation field rather than by internal collisions. NLTE slab models set an upper limit of about 2×10^4 K; at higher temperatures, the contrast is positive at line center for all reasonable values of the other parameters. The typical range $\Delta\lambda_D = 0.4\text{–}0.6$ Å yields a consistent but also weak constraint from the relationship $\Delta\lambda_D = (\lambda/c)(\xi^2 + 2kT/m)^{1/2}$: $T = (7\text{–}40) \times 10^3$ K for a microturbulent velocity in the range $\xi = 10\text{–}20$ km s^{-1}. As the ambient temperature at heights greater than a few thousand kilometers is $T \approx 10^6$ K, fibrils are typically much cooler than their surroundings.

The strength of the magnetic field within fibrils is not known with any confidence from direct measurement. Limited measurements of the direction of the field using chromospheric lines are consistent with the hypothesis that fibrils follow magnetic lines of force. This is also consistent with the fact that a field of 10 G corresponds to a magnetic pressure $p_{mag} = B^2/8\pi = 4$ dyn cm^{-2}, an order of magnitude greater than the gas pressures estimated above. The field is expected to exceed 10 G throughout the high chromosphere and low corona of an active region and may often exceed 100 G, so $\beta \equiv p_{gas}/p_{mag} \ll 1$ should apply within a fibril. The dynamic pressure $\rho v^2/2$ of gas flowing at 30 km s^{-1} is comparable to p_{mag} for $B = 10$ G.

The line-of-sight velocity of the material at some point along a fibril is sometimes estimated either from the Doppler shift of the line (the wavelength displacement of the bisector of a chord at a chosen depth in the line profile) or by noting the wavelength at which that material appears darkest (has the highest absolute value of contrast). Velocities estimated by the naïve contrast method are almost uncorrelated with velocities inferred from the cloud model and should not be used. Velocities estimated from the Doppler shift typically agree in sign with cloud velocities but are several times smaller in absolute value. Velocities derived from the cloud model should be more accurate than either of the simpler methods. Velocities so derived for arch filament systems and large fibrils are often supersonic at some points along the structure for any allowable value of the gas temperature.

Future research

Although fibrils are defined by their appearance in Hα, the analysis of Hα spectra reveals physical substance behind the definition: fibrils are dynamic structures that follow magnetic field lines but are denser and cooler than their surroundings. They are important elements of the active chromosphere. Future research on fibrils should therefore transcend the narrow objective of observing and explaining their appearance in Hα and bring to bear a full range of diagnostics. Particularly helpful would be: bi-dimensional (imaging) spectroscopy in temperature-sensitive visible lines, as well as Hα; UV/EUV imaging spectra from spacecraft; and magnetic measurements in chromospheric lines such as Hβ and He I 1083 nm. As many of these measurements as possible should be simultaneous. Such a range of observations would help to pin down not only the physical conditions within fibrils but the dynamic relationship of the fibril to the structure of the magnetic field and the coronal plasma in an active region.

Bibliography

Alissandrakis C E, Tsiropoula G and Mein P 1990 Physical parameters of solar Hα absorption features derived with the cloud model *Astron. Astrophys.* **230** 200–12

Bray R J and Loughhead R E 1974 *The Solar Chromosphere* (London: Chapman and Hall)

Cram L E 1986 Interpretation of Hα contrast profiles of active region loops *Astrophys. J.* **300** 830–5

Foukal P 1971 Morphological relationships in the chromospheric Hα fine structure *Solar Phys.* **19** 59–71

Tsiropoula G and Schmieder B 1997 Determination of physical parameters in dark mottles *Astron. Astrophys.* **324** 1183–9

Douglas Rabin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Heating Mechanisms

Mechanical heating

Chromospheres (and coronae) exist because of mechanical heating. Mechanical heating comprises all processes which convert nonradiative hydrodynamic or magnetic energy into thermal energy, that is, into microscopic random thermal motion. The physics of chromospheres (and coronae) is different from the physics of STELLAR ATMOSPHERES and interior layers, where the energy input is exclusively due to radiation, convection and thermal conduction. As observations show that the chromospheric emission depends strongly on the magnetic field, one subdivides the chromospheric heating processes into pure hydrodynamic mechanisms and magnetic mechanisms. Both types of mechanisms can be subdivided further into fast and slow processes. For the magnetic mechanisms the fast or wave mechanisms are called ac (alternate current) mechanisms and the slow mechanisms dc (direct current) mechanisms.

A mechanical heating mechanism consists of three processes, the *generation* of the energy carrier, the *transport* of this mechanical energy and its *dissipation*. Ultimately the mechanical energy comes from the nuclear processes in the stellar core. The energy generated there is transported in the form of radiation and convection to the stellar surface, where in the surface convection zone the mechanical energy is generated. The mechanical energy generation is due to the gas motions of the convection zone, which are largest in the regions of smallest density near the top boundary of that zone. Consequently the mechanical energy is generated in a narrow surface layer.

As current observations cannot tell with certainty whether a solar surface region is devoid of small-scale magnetic fields or not and because heating in current channels with diameters of meters cannot be resolved at the present time, it is difficult to identify a specific heating mechanism from the list of proposed processes. For an extensive list of proposed heating mechanisms see Narain and Ulmschneider (1996) and also the article on CORONAL HEATING MECHANISMS. Usually several mechanisms act at the same time. Surprisingly only by stellar observations (where the chromospheres are reduced to point sources) was one able to definitively identify acoustic waves as the important basic mechanism for stellar chromospheres. The reason for this is that for stars the effective temperature $T_{\rm eff}$ can be changed by 1 order of magnitude, the gravity g by 4, the metallicity by 3 and the rotation rate by 2 orders of magnitude. This greatly enlarges the range of theoretical predictions and thus permits crucial observations.

In the following subsections the necessity of mechanical heating, the zoo of mechanisms, the mechanical energy generation calculations, the observed mechanical heating rates and the possible total chromospheric heating picture are discussed in greater detail.

Necessity of mechanical heating

Consider a gas element in the chromosphere where flows introduced by the solar wind can be neglected. The total heating rate, that is, the amount of net heat per s, $\Phi_{\rm T}$ (erg cm^{-3} s^{-1}), flowing into the element across its boundaries, is given by

$$\Phi_{\rm T} = \Phi_{\rm R} + \Phi_{\rm J} + \Phi_{\rm C} + \Phi_{\rm V} + \Phi_{\rm M} = 0$$

as in dynamical equilibrium the chromosphere does not show time dependence. The terms on the right-hand side stand for radiative, Joule, thermal conductive, viscous and mechanical heating. A typical empirical chromospheric cooling rate from the standard solar model is $-\Phi_{\rm R} = 10^{-1}$ erg cm^{-3} s^{-1}. Consider a typical acoustic or magnetohydrodynamic disturbance in the solar chromosphere with characteristic values, size $L =$ 200 km, temperature $\Delta T = 1000$ K, velocity $\Delta v =$ 3 km s^{-1} and magnetic field perturbation $\Delta B = 10$ G. Using appropriate values for the thermal conductivity $\kappa_{\rm th}$, viscosity $\eta_{\rm vis}$ and electrical conductivity $\lambda_{\rm el}$ one finds (in erg cm^{-3} s^{-1})

$$\Phi_{\rm C} = \frac{\rm d}{{\rm d}z}\kappa_{\rm th}\frac{{\rm d}T}{{\rm d}z} \approx \frac{\kappa_{\rm th}\,\Delta T}{L^2} \approx 3\times 10^{-7}$$

$$\Phi_{\rm V} = \eta_{\rm vis}\left(\frac{{\rm d}v}{{\rm d}z}\right)^2 \approx \frac{\eta_{\rm vis}\,\Delta v^2}{L^2} \approx 1\times 10^{-7}$$

$$\Phi_{\rm J} = \frac{c_{\rm L}^2}{16\pi^2\lambda_{\rm el}}(\nabla\times B)^2 \approx \frac{c_{\rm L}^2\,\Delta B^2}{16\pi^2\lambda_{\rm el}L^2} \approx 7\times 10^{-5}$$

where $c_{\rm L}$ is the light velocity. It is seen that in the chromosphere $\Phi_{\rm C}$, $\Phi_{\rm V}$ and $\Phi_{\rm J}$ can be neglected relative to $\Phi_{\rm R}$ and that the main energy balance is between radiative cooling and mechanical heating. If mechanical heating were switched off then one would obtain radiative equilibrium, $\Phi_{\rm R} = 4\pi\kappa(J - S) = 0$. That is, in a radiative relaxation time scale of about a fraction of an hour radiative equilibrium would be reached. Here κ is the opacity, J the mean intensity and S the source function.

However, the heating rates $\Phi_{\rm C}$, $\Phi_{\rm V}$ and $\Phi_{\rm J}$ can become much bigger if the length scale L becomes very small. For acoustic waves as well as slow-mode MAGNETOHYDRODYNAMIC (MHD) and longitudinal MHD tube waves, this is accomplished by shock formation; for magnetic cases, by the formation of current sheets. Heating terms of the small L type are collected in the mechanical heating term $\Phi_{\rm M}$. Mechanical heating thus occurs when small scales form naturally.

The zoo of heating mechanisms

The *hydrodynamic mechanisms* consist of pulsational waves with periods larger than the acoustic cut-off period, $P > P_{\rm A} = 4\pi c_{\rm S}/\gamma g$ (where $c_{\rm S}$ is the sound speed and $\gamma = 5/3$ the ratio of specific heats), and acoustic waves where $P < P_{\rm A}$. These hydrodynamic mechanisms dissipate via shocks. The *magnetic mechanisms* are the acoustic-like slow-mode MHD waves and longitudinal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

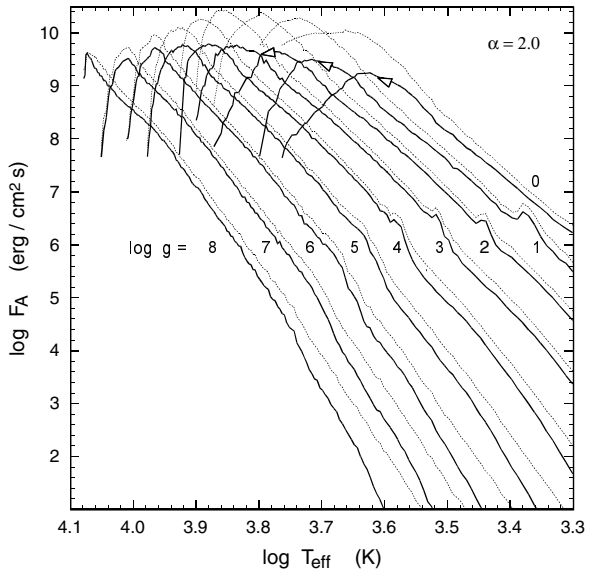


Figure 1. Acoustic wave energy fluxes (drawn).

tube waves as well as the transverse and torsional Alfvén waves. The latter two wave types are difficult to dissipate in the chromosphere and numerous dissipation mechanisms, mode coupling, resonance heating, turbulent heating, have been studied. In addition magnetoacoustic surface waves are present which dissipate via mode coupling, resonant absorption and phase mixing. The dc mechanisms associated with magnetic field reconnection very likely operate at heights above the chromosphere. The so-called turbulent and explosive events seen in the C IV transition layer line are thought to be caused by such reconnection processes. For more details see Narain and Ulmschneider (1996).

Mechanical energy generation

Mechanical energy is generated by the turbulent gas motions in the convection zone. Numerical convection studies and observations indicate that the turbulent flows can be described by a Kolmogorov-type energy spectrum. Using such a spectrum where the rms turbulent velocity is provided from convection zone models (which depend on $T_{\rm eff}$, g and the mixing-length parameter α, where $\alpha = 2$ agrees well with observations), acoustic wave energy fluxes for the Sun and a large number of stars have been computed using the Lighthill–Stein theory (Ulmschneider *et al* 1996). Lighthill's theory of quadrupole sound generation is in excellent agreement with observations in terrestrial applications. Acoustic fluxes thus obtained are shown in figure 1. The Lighthill–Stein theory also provides acoustic frequency spectra which extend from the cut-off period $P_{\rm A}$ to about a factor 100 smaller with a maximum roughly at period $P_{\rm M} = P_{\rm A}/5$. For the Sun one finds typical values of the total acoustic wave energy flux of 1.7×10^8 erg cm^{-2} s^{-1} and period $P_{\rm M} = 60$ s, using a mixing-length parameter $\alpha = 2.0$.

The fluxes of longitudinal and transverse waves in magnetic tubes have been computed for the Sun by squeezing and shaking a tube model by external time-dependent turbulent flows at a given excitation height. Figure 2 shows the instantaneous and averaged generated longitudinal (top) and transverse (bottom) tube wave fluxes as function of time. The figure shows the very stochastic nature of the generation process. The wave energy generation occurs in spurts which is in good agreement with observations. The spurts are caused by sudden enhancements in the flow velocity. In figure 2 the same time-dependent turbulent velocity fluctuations were applied. For the Sun a typical average longitudinal wave flux of 1.5×10^8 erg cm^{-2} s^{-1} and for the transverse flux a value about 30 times higher were found with a flat maximum extending from $P_{\rm M} = 90$ s to larger periods. In view of the same acoustic wave flux inside and outside the tube one should realize that the gas density inside the tube is roughly by a factor of 6 smaller than the external density. Relative to this low density, the wave flux inside the tube is large, leading to rapid shock formation, but also to severe NLTE effects (departure from the thermodynamic equilibrium) which strongly modify the radiative losses.

Empirical mechanical heating rates

Standard empirical solar chromosphere models indicate that active regions (where magnetic fields dominate) need mechanical fluxes of $F_{\rm M} = 1.3 \times 10^8$ erg cm^{-2} s^{-1}, while quiet regions 1.4×10^7 erg cm^{-2} s^{-1}. These values have been multiplied by a factor of 2.3 to account for the line losses by Fe II.

Direct measurements of velocity fluctuations have been used to estimate the acoustic flux in the solar chromosphere. Because velocity fluctuations are observed as Doppler shift fluctuations of the cores of spectral lines, one must be careful to correct for seeing and for the effects of the line contribution function. Deubner (1988) finds $F_{\rm M} = 2 \times 10^7$, 1.2×10^6, 4.5×10^5 erg cm^{-2} s^{-1} at the heights 300, 800, 1500 km, respectively.

The emission cores of the Ca II H and K and of the Mg II h and k lines are reliable indicators of chromospheres. Plotting for many stars the empirical core emission fluxes in these lines versus $T_{\rm eff}$ one finds that the fluxes show a well-defined lower limit which extends e.g. for Ca II from $F_{\rm Ca} = 1 \times 10^7$ erg cm^{-2} s^{-1} for early F stars to 2×10^4 erg cm^{-2} s^{-1} for late M stars. This limit is called *basal flux line*. It has been shown recently that this basal emission can be reproduced by acoustic heating (Buchholz *et al* 1998). However, one finds that stars with the same $T_{\rm eff}$ and gravity g show different chromospheric emission and that this emission is larger when the stars rotate more rapidly. As rapidly rotating stars with a surface convection zone generate more magnetic fields via the dynamo mechanism it is clear that the excess heating is caused by magnetic heating mechanisms. A rapidly rotating star has more than an order of magnitude more chromospheric emission than the basal emission.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

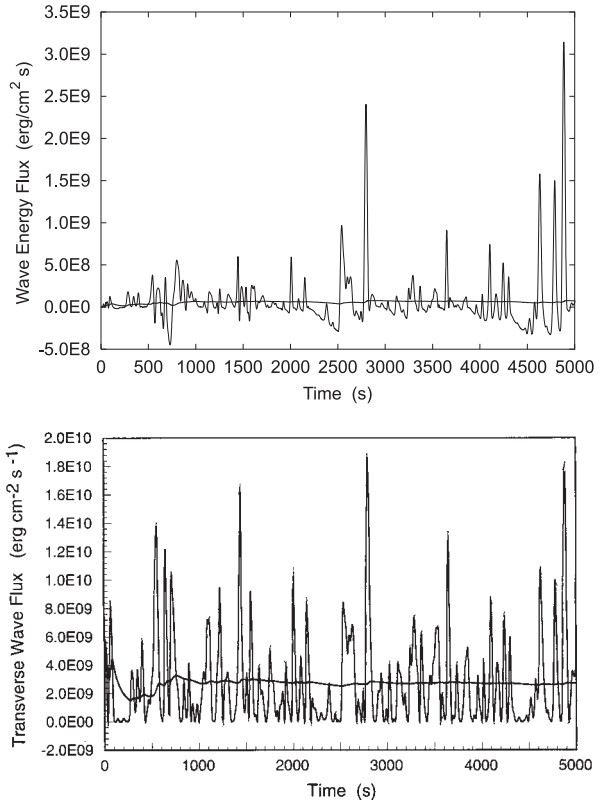


Figure 2. Instantaneous and average longitudinal (top) and transverse (bottom) MHD wave energy flux.

The total chromospheric heating picture

The successful reproduction of the basal flux line by acoustic wave heating models clarified that the fundamental heating process in chromospheres are acoustic waves. This mechanism operates independently of rotation. The faster a star rotates the more it is covered by magnetic fields which at photospheric and subphotospheric heights appear in the form of magnetic flux tubes. The fluctuating gas flows in the turbulent convection zone outside the tubes excite longitudinal, transverse and torsional MHD tube waves. By shock formation the longitudinal waves in the lower and middle chromosphere heat the flux tubes which explains the excess emission in the magnetic regions. With more tubes on the star, more tube wave energy is generated. The much more efficiently produced transverse waves have two problems: they suffer severely from leakage of the wave energy away from the tube and they are difficult to dissipate. Very likely they are responsible for the heating of the high chromosphere and the transition layer. Here stellar observations of the C IV line (Wood *et al* 1996) indicate that, in addition to wave heating, dc heating by microflares generated by reconnection events augments the Alfvén wave heating. These reconnection events are the result of slow horizontal convective motions (called foot-point motions) which lead to twisting and braiding of the magnetic flux tubes. Thus the convection zone is the source not only of ac heating but also of dc heating. The probably less efficiently generated torsional waves very likely traverse the chromosphere without dissipation and become important in the corona and in the stellar wind. An efficient three-dimensional simulation of the magnetohydrodynamics in stellar convection zones should in the future be able to completely explain the chromosphere phenomenon which, in absence of any input from interstellar space must, after all, be completely dependent on the interior of the star.

Bibliography

Buchholz B, Ulmschneider P and Cuntz M 1998 Basal heating in main-sequence stars and giants: results from monochromatic acoustic wave models *Astrophys. J.* **494** 700–14

Deubner F-L 1988 Observations of solar oscillations *Pulsation and Mass Loss in Stars* ed R Stalio and L A Willson (Dordrecht: Kluwer) pp 163–79

Narain U and Ulmschneider P 1996 Chromospheric and coronal heating mechanisms II *Space Sci. Rev.* **75** 453–509

Ulmschneider P, Theurer J and Musielak Z E 1996 Acoustic wave energy fluxes for late-type stars *Astron. Astrophys.* **315** 212–21

Wood B E, Linsky J L and Ayres T R 1996 Evaluating possible heating mechanisms using the transition region line profiles of late-type stars *Astrophys. J.* **478** 745–65

P Ulmschneider

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Magnetic Canopy

The magnetic canopy is a layer of magnetic field which is directed parallel to the solar surface and located in the low CHROMOSPHERE, overlying a field-free region of the SOLAR PHOTOSPHERE. It has a field strength of the order of 0.01 T and covers a large area fraction of the solar surface. The magnetic canopy can be compared with the canopy of the rain forest: the tree-trunks correspond to the MAGNETIC FLUX TUBES that rise in the vertical direction from the photospheric level, while the spreading branchy foliage compares with the horizontally spreading magnetic field. As in the case of the rain-forest canopy, the magnetic canopy has its own life and is not readily accessible to observation.

The magnetic canopy can be observationally inferred from magnetograms taken in chromospheric spectral lines or, theoretically, from a magnetohydrostatic extrapolation of the easily observable photospheric magnetic field.

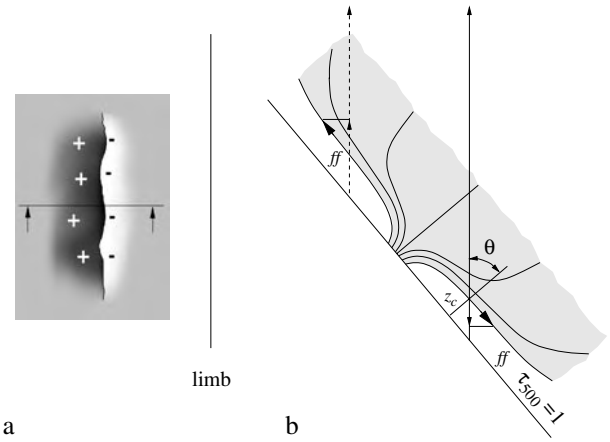


Figure 1. (*a*) Sketch of a chromospheric magnetogram near the solar limb and (*b*) the corresponding magnetic configuration. The base of the magnetic canopy is located at a height z_c above the solar surface.

Observational evidences of the magnetic canopy

At the beginning of the 1980s R G Giovanelli and H P Jones extensively studied the magnetic canopy, based on earlier suggestions and observational data by W C Livingston. They used magnetograms of the network field near the solar limb (see CHROMOSPHERE: NETWORK), taken in chromospheric spectral lines within the infrared and the visible wavelength range, e.g. the 854.2 nm triplet line of Ca II and the Mg I b_2 line at 517.3 nm. These magnetograms typically show a *polarity inversion* across a line that coincides with the limbward edge of a unipolar magnetic network field. (The magnetic network is seen in photospheric magnetograms, outlining the supergranule boundaries.) Towards disk center from the polarity-inversion line a diffuse field (dark area in figure 1(*a*)) extends well over a region which shows no magnetic field at the photospheric level and which has the same (normal) polarity as the magnetic network. On the limbward side this patch of normal polarity shows a fringe of opposite polarity. The polarity inversion and fringe pattern is a very common feature of chromospheric magnetograms of the network field near the limb. It also occurs in active regions and sunspots.

Figure 1(*b*) shows the magnetic structure that most naturally explains such a magnetogram. It shows the front elevation of the section indicated in figure 1(*a*). The normal to the solar surface (the height level at which the continuum optical depth at $\lambda = 500$ nm is unity ($\tau_{500} = 1$)) is inclined by an angle θ with respect to the line of sight, corresponding to an observation near the solar limb. Although the magnetic network field is fragmented and consists of many small-scale magnetic flux tubes, in figure 1(*b*), for clarity, it is represented by a single vertically oriented flux tube. At the network boundary the field is free to spread in the horizontal direction, thereby forming the magnetic canopy which, overlies the field-free region *ff*. The magnetic polarity inversion of the chromospheric magnetogram can now be readily understood as due to the canopy field in the following way. Lines of sight passing through the network field receive a magnetic field component toward the observer and so do lines of sight passing through the canopy field to the left (disk center) side of the network (dashed line of sight), explaining the extended diffuse network field. The canopy field to the right (limbward) side of the network, however, gives rise to a line of sight component of opposite direction (solid line of sight), which causes the fringe of opposite polarity. Note that the corresponding magnetogram of the magnetic network at the photospheric level would be of exclusively positive polarity. Towards the limb, the network fields apparently move closer together with increasing line of sight aspect angle θ, giving rise to an ever narrower *fringe pattern* of alternating polarity in the magnetogram.

By using a standard solar model atmosphere, the radiation transfer equation for polarized light is integrated along a line of sight that passes through the canopy field like the solid one indicated in figure 1(*b*). Assuming the atmosphere to be field free up to a height z_c and the canopy to consist of a homogeneous layer of magnetic field of strength B_c, the canopy height z_c can be derived from a comparison of the thus computed with the observed magnetogram signal. Since only the magnetic field close to the *canopy base* contributes substantially to the magnetogram signal (because of the rapid decrease of density and field strength with height), the assumption of a homogeneous, purely horizontally directed canopy field is a justifiable approximation. Such calculations have resulted in canopy heights of 600–1000 km above the solar surface (corresponding to the lower chromosphere) in quiet-Sun regions and as low as 200 km in ACTIVE REGIONS. Contrary to this, simple flux-tube models of the magnetic network show the canopy field to form in the *upper* chromosphere and the CORONA. Only models that allow for substantial differences between the atmosphere within the flux tube (*internal atmosphere*) and outside it (*external*

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

atmosphere) are capable of producing a low-lying canopy field. SUNSPOTS are also observed to produce a low-lying canopy field, which extends over a large distance beyond the penumbra (superpenumbra; see SUNSPOT PENUMBRA) with a canopy height of one hundred to several hundred kilometers. In this case, unlike to the network canopy, magnetic tension forces play a dominant role in keeping the canopy at low altitude.

The flux-tube model of the magnetic canopy

Given the magnetic field strength of the network field at the solar surface ($\tau_{500} = 1$), B_0, and the surface fraction occupied by this field, the magnetic filling factor f_0, the simplest model that can be constructed consists of an ensemble of static, rotationally symmetric, vertical flux tubes that are evenly distributed at the $\tau_{500} = 1$ level. It is usually assumed that at this level the field strength of the vertical component is constant across the flux tube, $B_z(r) = B_0 = \text{const.}$, and that the internal and external atmospheres are both plane parallel (no variation in lateral but only in the vertical direction). Given additionally the gas pressure p_g as function of height for the internal and external atmospheres, the magnetic field configuration can be computed for all heights, including the canopy and chromospheric region, by solving the magnetohydrostatic equation (see MAGNETOHYDRODYNAMICS).

A representative example of such a flux tube is shown in figure 2, indicating magnetic field-lines separating equal amounts of magnetic flux Φ. The flux-tube surface is distinguished by a discontinuity (in the case of an ideal plasma) in physical variables in such a way that the magnetic pressure, $p_{\rm mag} = B^2/2\mu$, exerted by the flux-tube surface field equals the difference in gas pressure between external and internal atmosphere, i.e.

$$B^2/2\mu|_{\rm surface} = p_{\rm g\,ext} - p_{\rm g\,int} \qquad (1)$$

where $\mu = 4\pi \times 10^{-7}$ V s A^{-1} m^{-1} is the magnetic permeability taken as that of free space. This equation is an immediate consequence of Ampère's law and is still valid in the presence of magnetic curvature forces. In the extreme case of a totally evacuated flux tube, $p_{\rm g\,int} = 0$, and we obtain $B_{\rm max}(z) = [2\mu\, p_{\rm g\,ext}(z)]^{1/2}$, which amounts to about 0.17 T at the solar surface level and to about 0.02 T at a height of 500 km. Because of the exponential decrease of the external gas pressure $p_{\rm g\,ext}$ with height, the magnetic flux tube expands, while keeping the total magnetic flux Φ constant.

It can be seen from figure 2 that the magnetic field in the outer parts of the tube becomes horizontal at some critical height, thereby forming the canopy field. This horizontally spreading field merges with the equal-polarity field of neighboring flux tubes, in which case it bends back and outward into the vertical direction again, or it connects with opposite-polarity neighboring tubes to form a loop-like structure. The first case is shown in figure 2, where the canopy field merges with the neighboring field of the same polarity at a height of

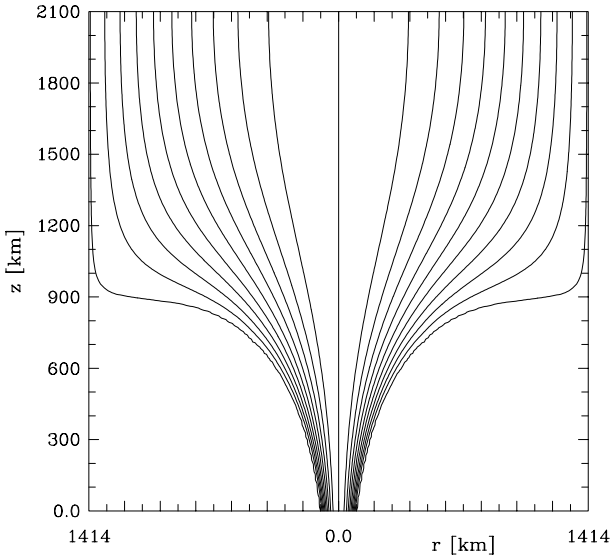


Figure 2. Magnetic flux tube. The field lines spread into a horizontally extending canopy field at a height of approximately 800 km and merge with the field of neighboring flux tubes of equal polarity at a radius of 1400 km.

900 km. The horizontal distance over which the canopy field stretches depends on the local filling factor, which varies between ≈0.2 (PLAGE regions) to 5×10^{-4} (quiet Sun). Thus, a canopy field may cover a large area of field-free photosphere of the quiet Sun. In figure 2 it stretches over only a moderate horizontal span (from 700 to 1400 km in a height region from 800 to 1100 km) owing to a filling factor of 5×10^{-3} that is ten times larger than that of a quiet-Sun region.

The height at which the canopy field sets in, the canopy height z_c, critically depends on the combination of external and internal atmosphere. A canopy field with a base height in accordance with observations ensues only when the flux-tube atmosphere shows a chromospheric temperature rise while the temperature in the external atmosphere continues to fall beyond the height of the temperature minimum of the flux-tube atmosphere. Under these circumstances, the gas pressure within the flux tube rapidly approaches the surrounding gas pressure at a critical point above the height of the temperature minimum. According to equation (1), this forces the flux-tube surface to spread in the horizontal direction because the finite surface field cannot reach beyond the critical height where $p_{\rm g\,ext} - p_{\rm g\,int} = 0$, irrespective of field strength or filling factor.

Is it possible that the external atmosphere remains cool up to the canopy base? An analysis of the solar spectral lines of the carbon monoxide molecule as well as radiative equilibrium models that include the effect of the CO molecule show that over a large fraction of the solar surface the temperature must strongly drop *below* the classical temperature minimum, so that the atmosphere *bifurcates* into a hot and a cool component near

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the height of the classical temperature minimum, which is at about 500 km above the solar surface (see CHROMOSPHERE: THERMAL BIFURCATION). It is very likely that the hot and the cool component represent the internal and external flux-tube atmospheres, respectively. Thus, the occurrence of the canopy field seems to be tightly connected with the thermal bifurcation of the solar atmosphere.

Consequences of the magnetic canopy

As a consequence of the thermal bifurcation, the atmosphere immediately below the canopy base is relatively cool ($\leq$4000 K) while the canopy itself may be already at chromospheric temperature, depending on the balance of heating and cooling. Hence, the canopy base represents a *thermal boundary layer*, which has important consequences for the formation of spectral lines in the lower chromosphere, such as the central emission core of Ca II H and K or Mg II h and k. The formation of these emission cores is suppressed by the cool atmosphere below the canopy base. On the other hand, since the flux-tube atmosphere shows a chromospheric temperature rise, the emission cores fully develop along the flux-tube axis and in its vicinity, the size of which depends on the expansion rate of the flux tube. This is in accordance with the observed enhanced chromospheric emission in network regions and the absence of it (with the exception of sporadic bright grains) in intranetwork regions (see SOLAR PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS). The model thus shows emission lines in the low chromosphere to be *spatially intermittent* on scales of about 1 arcsec and to be strongly correlated with the intense magnetic field at the photospheric level, both predictions being in good agreement with observation. The resonance lines of Mg II form slightly higher in the atmosphere than those of Ca II, probably mostly above the canopy base. Correspondingly, these lines show emission cores also in the intranetwork region.

A canopy field in conjunction with a spatially separate mass motion beneath the canopy (e.g. convective motion) may strongly influence the shape of Zeeman sensitive spectral lines when observed in polarized light (SOLAR MAGNETIC FIELD: MEASUREMENT and POLARIMETRY). The Stokes V profile of a spectral line (the difference between right-hand and left-hand circularly polarized light as a function of wavelength), which is anti-symmetric under normal conditions, may assume an extreme asymmetric shape depending on the velocity and the canopy height.

The magnetic canopy probably bears consequences for the formation and strength of acoustic shocks that result from upward traveling waves generated at photospheric and convection zone levels. These waves preferentially shock at around a height of 1 Mm, thus near the canopy base, giving rise to the Ca II H_{2V} and K_{2V} bright grains (see SOLAR CHROMOSPHERIC OSCILLATIONS). The precise emission properties of these grains are also expected to be modified by the presence of the magnetic canopy.

As another consequence of the magnetic canopy, the frequencies of solar acoustic p-mode oscillations of high modal degree are shifted to larger values. For a given mode, this FREQUENCY shift is proportional to the square of the canopy field strength, which is approximately proportional to the photospheric filling factor, f_0. This is a possible explanation for the variation of the frequencies of p-mode oscillations over the SOLAR CYCLE and their correlation with magnetic activity.

Bibliography

Evans J E and Roberts B 1992 Interpretation of solar-cycle variability in high-degree p-mode frequencies *Nature* **355** 230–2

Giovanelli R G 1980 An exploratory two-dimensional study of the coarse structure of network magnetic fields *Sol. Phys.* **68** 49–69

Jones H P and Giovanelli R G 1982 Magnetic canopies in unipolar regions *Sol. Phys.* **87** 37–42

Solanki S K, Livingston W and Ayres T 1994 New light on the heart of darkness of the solar chromosphere *Science* **263** 64–6

Steiner O 1994 Theoretical models of magnetic flux tubes: structure and dynamics *Infrared Solar Physics (IAU Symposium 154)* ed D M Rabin (Dordrecht: Kluwer) pp 407–21

Oskar Steiner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Network

The chromospheric network on the Sun was discovered by HALE in 1892 with his newly invented spectroheliograph. It is a patchy pattern (Hale called it a 'reticuled structure') covering the solar disk outside active regions that appears when the Sun is imaged in spectral diagnostics formed in the CHROMOSPHERE (figure 1).

The network pattern has a honeycomb appearance at characteristic cell sizes of 10–40 arcsec and consists of small magnetic patches which outline, although incompletely, the borders of the supergranulation cells in the underlying photosphere (SOLAR PHOTOSPHERE: SUPERGRANULATION). Each magnetic patch measures a few arcsec and consists probably of a cluster of flux tubes that are either already physically merged at chromospheric heights (1000–5000 km above the photosphere) or appear merged due to radiation scattering (SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: OBSERVATIONS).

The chromospheric network is to high precision co-spatial with the magnetic network observed with magnetographs in the photosphere. The reason for the co-spatiality is the kilogauss field strength of the photospheric flux tubes. Such high field strength goes together with near-evacuation, therefore large buoyancy, and therefore radial (upright) orientation (SOLAR PHOTOSPHERE).

The network patches appear bright in chromospheric lines (the Ca II H and K resonance lines at $\lambda = 396$ nm and 393 nm, the inner wings of the hydrogen Balmer series Hα line at 656.3 nm, and many emission lines in the mid-ultraviolet). The observed brightness increases with the apparent magnetic flux density in the underlying photosphere, effectively measuring the spatial flux tube density. The intensity enhancement is presumably due to magnetism-related heating, but the precise mechanism has not been identified (CHROMOSPHERE: HEATING MECHANISMS).

The network pattern is dynamic on timescales of hours. Its evolution is governed by the continuous appearance of new magnetic field within supergranulation cells, the subsequent field migration to the cell boundaries imposed by the supergranular flows, further migration along the supergranular boundaries, and eventual field disappearance through dispersion, reconnection, subduction or rise into the upper atmosphere.

Composition

The magnetic flux tubes that make up the magnetic network are seen individually in the highest-resolution pictures of the photosphere. In particular, they are most sharply imaged in the Fraunhofer G-band around $\lambda = 430.5$ nm. This band is made up of molecular CH lines that brighten considerably, in not-understood fashion, at locations where strong-field flux tubes in intergranular lanes are dynamically interacting with the surrounding convective flows. The flux tubes show up on G-band filtergrams as tiny bright points with 0.1–0.2 arcsec diameter—just at the diffraction limit of the best telescopes (SOLAR PHOTOSPHERE: FILIGREE; SOLAR TELESCOPES AND INSTRUMENTS: GROUND).

The patches in the overlying chromosphere appear more diffuse, with grainy structure at 1–2 arcsec scale. In ground-based observations they are best seen when imaging the solar chromosphere in the cores of the Ca II H and K lines. These are the strongest spectral lines of the optically accessible part of the solar spectrum because calcium is an abundant element and most calcium particles reside in the singly-ionized ground state throughout the photosphere and low chromosphere. In addition, the network patches are well visible in the ultraviolet lines from the lower ionization stages of abundant elements such as CI and CII in the mid-ultraviolet. They remain visible in ultraviolet lines from higher ionization stages up to formation temperature $\approx 10^5$ K, and are also evident in Lyman α at $\lambda = 121.5$ nm (the strongest line in the solar spectrum).

More complex morphology is seen in Hα than in Ca II K because Hα has additional contributions from much higher layers, due to its high excitation energy. Only at high temperature and relatively high density does it become sufficiently opaque. When observed at Hα line center the network has dark extensions called mottles or, when they are longer than a few arcsec, fibrils (CHROMOSPHERE: FIBRILS). These extend into spicules when seen at the limb (CHROMOSPHERE: SPICULES). Such filamentary structures undoubtedly portray magnetic field topology and they may represent the lower parts of coronal loops (CORONAL LOOPS), but their behavior is very dynamic and clearly more complex than static loop models would predict. Static models of flux tube expansion with height (SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: THEORY) predict that at much lower levels the magnetic field should already funnel out to nearly homogeneous space filling (CHROMOSPHERE: MAGNETIC CANOPY). The observed mottle, fibril and spicule flows map the more vertical field lines since the gas moves freely only along the field. The more inclined flows do not reach as high.

The spatial coincidence with the magnetic network in the photosphere can be used to determine the magnetic polarity of the chromospheric network. It tends to be mixed-polarity over large areas in the quiet Sun and to become unipolar near active regions or to remain unipolar in the remains of previous active regions. Large unipolar areas also occur near the solar poles.

Topology and pattern evolution

The cellular appearance of the chromospheric network is incomplete, in contrast to the supergranulation cell boundaries observed on photospheric Dopplergrams. On low-resolution Ca II K spectroheliograms the cells are more easily outlined than on high-resolution Ca II K filtergrams, illustrating that the supergranulation flows do not produce field filling of the cell boundaries everywhere. In addition, there are mobile, small patches of field concentration within the supergranulation cells that haven't yet made

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

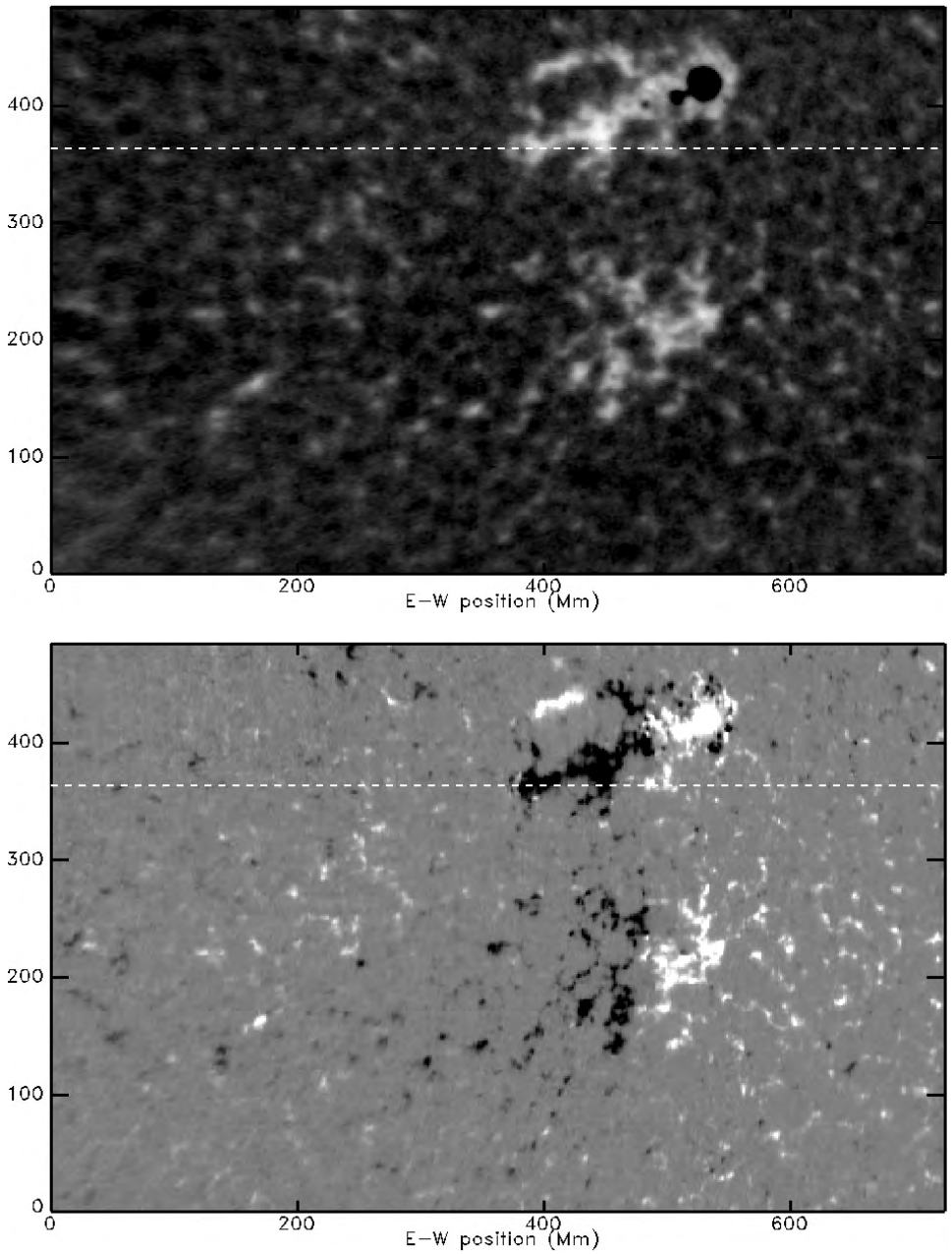


Figure 1. Chromospheric network. Upper panel: solar image in the Ca II K line taken on 21 December 1994 from Antarctica. Dashed line: solar equator. Upper bright patch: small active region with sunspot (blanked out). Lower bright patch: active network. Remainder of the field: quiet network. Lower panel: corresponding photospheric magnetogram taken at the same time from Kitt Peak. Black and white: apparent longitudinal magnetic flux density in both polarities.

it to a cell boundary. It is likely that many of these mobile patches have recently emerged from below the surface.

The cell sizes of the chromospheric network are larger (32 Mm diameter) when determined with classical autocorrelation methods than when topological methods are used that identify cells from gradient diagnostics (13–18 Mm diameter). There seems to be no direct relation between cell size and magnetic content, although in non-quiet plage areas the field-free cells seem to be much smaller. The spatial cell characteristics are dominated by the spatial density of the upflow centers, as is also the case for the smaller-scale granulation (SOLAR PHOTOSPHERE: GRANULATION). Indeed, both granulation and supergranulation appear to form as bubbles starting from a 'seed' location, i.e. a center of outflow, and both are surface filling.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetic patches originally emerge in bipolar ephemeral regions inside supergranular cells, and are subsequently transported to the supergranulation cell boundaries, where they stay confined. The motion of the magnetic concentrations can be described by a random walk in response to the continually evolving convective motions. On the granulation scale the granular buffeting makes magnetic concentrations move in a random walk with typical step lengths equal to a granular radius ($R \simeq 0.5$ Mm) and time steps equal to a granular lifetime ($T \simeq 10$ min). The flow patterns across supergranulation cells survive longer than the individual small-scale granular flows so that the patches are eventually collected in the supergranular boundaries by the supergranular outflows. They subsequently move along these boundaries and so outline the supergranular boundaries. This can be observed both for the chromospheric network on Ca II K filtergrams and high-resolution ultraviolet images, and on photospheric magnetograms.

Along the supergranular boundaries, the magnetic field appears concentrated in patches with an average magnetic flux content of 2.5×10^{18} Mx (Mx, Maxwell $= 10^{-8}$ Weber; 1 G $=$ 1 Mx cm^{-2}) varying from well below 10^{18} Mx to more than 15×10^{18} Mx per patch and with an exponential drop-off in frequency of occurrence. The observed Ca II K excess brightness $I_{\rm K}$ above the non-magnetic background ('basal flux') increases with the apparent magnetic flux density ϕ in the underlying photosphere according to the power law $I_{\rm K} \propto |\phi|^{0.65}$ (but with considerable scatter around this relation). Note that the apparent flux density is often confused with intrinsic solar magnetic field strength; it is primarily a measure of the spatial density of unresolved flux tubes with intrinsic field strength $|\boldsymbol{B}| \approx 1400$ G over the angular resolution element of the telescope.

On the supergranulation scale the cell pattern evolution causes another random walk with typical step lengths equal to a supergranular cell radius ($R \simeq 6 - 9$ Mm) and typical time steps equal to the characteristic supergranular lifetime ($T \simeq 24$ h).

While diffusing along supergranular boundaries, the flux patches may interact with one another. When field concentrations of opposite polarity meet, they may cancel against each other, i.e. vanish from sight. This may be due to actual field reconnection, but may also indicate that magnetic fields move upwards or downwards through the observed layer, in the form of small vertical loops with their apex pointed downward or upward respectively. When patches of the same polarity meet, flux merging may be observed. The reverse process, fragmentation, is also seen.

The overall exponential flux distribution function can be described by a 'magnetochemistry' model that describes the emergence of new concentrations, fragmentation into pieces, and disapperance through cancellation or merging. Taking the frequencies of these processes into account, the maximum lifetime of an individual field concentration is estimated to be no more than about 40 h. Thus, the fine structure of the quiet network cannot be recognized after a longer time span.

Dynamical properties

The precise co-spatiality between the chromospheric network and the photospheric magnetic field and the power-law relation between network brightness and apparent flux density (primarily measuring spatial flux tube density) strongly suggest that a magnetically dominated heating mechanism operates in or along the kilogauss flux tubes that make up the network (CHROMOSPHERE: HEATING MECHANISMS). However, many attempts have failed to find wave modes that would betray kinetic energy transfer from the photosphere where the flux tubes are buffeted by the granular flows up to the chromospheric heights where wave dissipation might contribute the heating that is required to explain the observed brightness. So far, all network oscillation studies have only identified rather slow periodicities, 5 min or longer. This result is in stark contrast with the roughly 4 to 2 min acoustic waves that are ubiquitously present in the internetwork regions. It is not even clear whether the observed slow modulations describe wave modes (such as breaking internal gravity waves) or just random motions enforced by the foot-point swaying. Thus, the actual mechanism through which the network patches are bright, while undoubtedly due to dynamical flux tube physics, is yet to be identified.

Bibliography

Lites B W, Rutten R J and Kalkofen W 1993 Dynamics of the solar chromosphere I. Long-period network oscillations *Astrophys. J.* **414** 345–56

Schrijver C J 1995 Basal heating in the atmospheres of cool stars. Observational evidence and theoretical support *Astron. Astrophys. Rev.* **6** 181–223

Schrijver C J, Title A M, van Ballegooijen A A, Hagenaar H J and Shine R A 1997 Sustaining the quiet photospheric network: the balance of flux emergence, fragmentation, merging, and cancellation *Astrophys. J.* **487** 424–36

Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

Robert J Rutten and Hermance J Hagenaar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Spicules

As early as 1877 Father Secchi described the jet like, elongated structures in the solar chromospheres which were named spicules by Walter Orr Roberts in 1945. They are observed in the optical chromospheric emission lines of hydrogen (especially the 656 nm Hα line), ionized calcium (especially the 393 nm and 397 nm H and K lines) and helium lines (especially the 588 nm and 1083 nm neutral helium lines). Figure 1 shows a series of images of the CHROMOSPHERE taken at different wavelengths in the Hα line showing the spicular structures. Except for the lowest heights above the solar limb, up to 1500 km, the spicules are clearly visible. Typical spicule properties are: width 900 km, average height 8000 km, lifetime 15 min apparent velocities 25 km s^{-1}.

From images like the one in figure 1 one derives the spatial properties of the chromosphere and its structure. Most or all of the chromospheric emission in these lines appears to come from spicules, so that it is fair to say that the solar chromosphere, at least at heights exceeding 1500 km above the photosphere, consists of spicular material. The decrease of average chromospheric brightness with height is the result of the decrease of the individual spicule brightness and of the number of spicules with height. When observed in the Hα line, the line most used, we observe an average height of spicules of 8000 km, somewhat less near the solar equator, more at the poles. It should be noted, however, that their upper boundary is not well defined, so that the observed height depends on the observing conditions. Individual spicules can extend significantly higher with heights above 10 000 km being observed. In the literature one finds references to so-called 'super-spicules' especially in the discussion of solar limb observations made in far-ultraviolet lines like the ionized helium line at 30.4 nm. Super-spicules are significantly larger than spicules; they are structures which in their properties lie between spicules and transient solar prominences (SURGES). Not included in the present context, they will be briefly discussed at the end of this section. The width of spicules is observed to average 900 km, which is probably an upper limit since it is uncorrected for the imaging effects of the Earth's atmosphere (SEEING).

Dynamics

The dynamics of spicules is most interesting. In their initial appearance they appear 'geyser-like' in the sense that there is an upward surge of material with velocities reaching 25 km s^{-1} lasting for some 5 min, but with a path which is frequently slanted with respect to the local vertical direction. After reaching a typical maximum height of 9000 km in 5 min, the ejection stops. It is followed by a fading of the spicule brightness or a return of the emitting material to the photosphere. On balance many more spicules appear to go up than come down. The total life cycle for those that descend is about 15 min. Since the spicular material is largely ionized, motions tend to follow the direction of the magnetic field lines. The spicular orientations and motions are therefore a good proxy for the magnetic field structure pattern in this part of the solar atmosphere. Direct measurements of the magnetic field strengths by means of the Zeeman, or possibly Hanle effect are not available yet (see SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). Although there are significant temporal and local spatial variations in the orientations, there are clear changes in the average orientations of spicule directions, and hence magnetic fields, with solar latitude. Near the solar equator the orientations, although on the average vertical, vary greatly from spicule to spicule. Near the poles the orientations are more uniform, mostly close to vertical, but sometimes inclined systematically. In polar regions the inclination appears to follow the orientations of coronal plumes. Because of the systematic group behavior of spicules there, the patterns are sometimes referred to as 'porcupine' or 'wheat field' configurations.

Astronomers are always concerned that the apparent motions seen in motion pictures like those of the Hα chromosphere might not correspond to real motions. Apparent motions can result from other effects, like temporal changes in atomic excitation conditions. The only way of checking the reality of the motions is by means of observations of the Doppler shift of the spectral line. This is a tricky thing to do, however, since the apparent, 'proper motions' occur in the plane of the sky whereas the Doppler shift correspond to line-of-sight motions. The comparison therefore is statistical, and requires knowledge of the varying tilts of spicules. Observations to date, taking these considerations into account, have indeed shown the apparent motions to be actual material motions. Doppler shifts appear to vary across spicules resulting in a tilt of the spectrum emission lines. These variations are often interpreted in terms of a rotation of spicules.

Physical properties

In addition to telling us about velocities, spectral analysis of the spicular emission provides important information about the physical conditions in spicules. Spectral observations are available in a number of different lines: the hydrogen Balmer lines (Hα through Hε), neutral helium lines (D3 line at 588 nm, 1083 nm line), and the ionized calcium lines at 393, 397, 850 and 854 nm). The observed emission of these lines indicate temperatures in the vicinity of 15 000 K, probably lower (<10 000 K) at heights below 2000 km, and electron densities decreasing from 1.5×10^{11} cm^{-3} at $h = 2000$ km to 3×10^{10} cm^{-3} at $h =$ 10 000 km. At the lower heights spicules are embedded in the what is normally referred to as the chromosphere–corona transition region. At the greater heights they are embedded in the million degree CORONA itself, but there must be a local transition region between each spicule and this corona. Nothing is known about these local transition regions. Future high-resolution observations from space of higher temperature emission lines in the ultraviolet (for example the CIV lines at 155 nm and the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

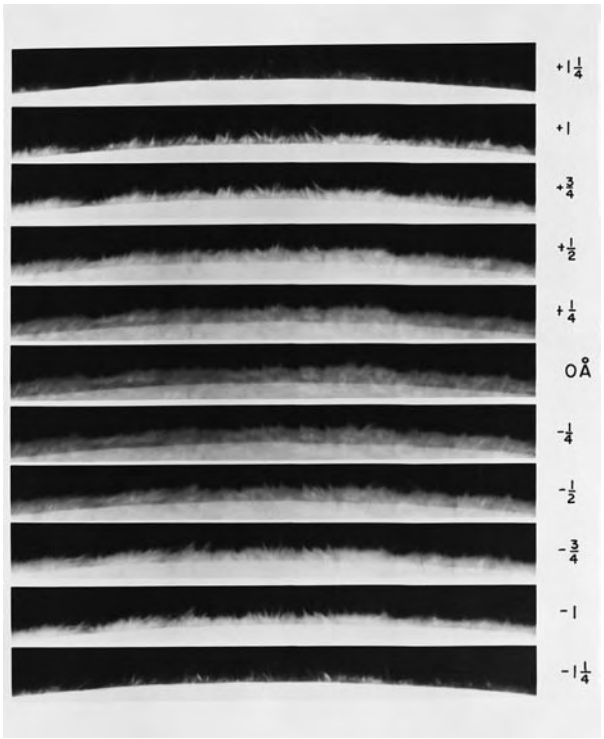


Figure 1. Images of the chromospheric spicule patterns as seen through a 0.025 nm wide Lyot filter at different wavelengths in the chromospheric Hα emission line. (Courtesy R B Dunn, National Solar Observatory at Sacramento Peak.)

helium Lyα line at 30.4 nm) can be expected to clarify that. The electron pressure in spicules at a height of 8000 km equals 0.09 dyn cm^{-2} as compared to 0.02 dyn cm^{-2} in the surrounding corona. The difference is large enough that it may be real, indicating possibly a significant overpressure in the chromospheric spicules. In addition one derives from the width of the spectral lines in spicules a non-thermal broadening of about 20 km s^{-1}, part of which may result in a turbulent pressure. An isolated magnetic field of only 1 G results in a dyn cm^{-2} magnetic pressure of 0.04 dyn cm^{-2}, a variation of 1 G in an average field of, for example, 12 G in a variation of 1 dyn cm^{-2}. Except for channeling the spicular motion, for which the field has to be significantly larger than 1 G, the magnetic field variations between spicule and the surrounding corona have to be less than 1 G, with the magnetic field strength variation, if any, being less in spicules.

This sensitivity in the pressure balance to small changes in the magnetic field strengths also points out large effects that the magnetic field can have on the dynamics of spicules. Quite a large number of theoretical models exist for the origin of spicules, all involving magnetic field effects. Because of the similarity in their dimension to that of the solar convection cells which are observed at the solar surface as GRANULATION, and because of the similarity in birth rates of spicules and granules within the higher magnetic field concentrations seen at the solar supergranule boundaries (see later in this section) it has been suggested that spicules are the result of the convection–magnetic field interactions. What is clearly needed are improved observations of spicule spectra and images, including higher spatial resolution observations of their velocities and, if possible, magnetic fields. Also, observations of spicules as observed in projection against the solar disk are invaluable. That is the main topic of the following discussion.

Appearance against solar disk

When observing structures like PROMINENCES or spicules in the higher solar atmosphere against the solar disk one might see those structures either as dark features in absorption or as bright structures in emission. Such structures are dark when the amount of background radiation from the solar disk that is absorbed is larger than the amount emitted by the structure itself. The ratio of the emission to absorption coefficient in the structure is referred to as the 'source function' of the structure, or in short S. A structure is bright when S exceeds the background intensity I, dark when it is less. The contrast depends on the optical thickness τ, so that the intensity difference ΔI from the background equals: $\Delta I = (I - S)(1 - e^{-\tau})$. This simple model of the visibility of structures overlying the bright solar background is frequently referred to as the 'cloud model'. The same model actually holds for terrestrial clouds and for any other similar configuration anywhere. For the conditions existing in the solar atmosphere the source function S can be taken as wavelength and direction independent. The background intensity is a function of wavelength (Hα is a strong Fraunhofer absorption line) and of position on the solar disk (solar limb darkening). It is therefore quite likely that the sign of the contrast of spicules when seen in projection against the solar disk changes with position in the Fraunhofer line and with disk position. With the physical conditions in spicules described earlier, one might expect spicules to be dark when seen in the line wings for most of the solar disk. Near the core of the H line and at the extreme limb one expects bright contrast structures. Right at the solar limb, where the disk brightness falls off rapidly and actually vanishes outside the limb, the expected behavior of spicule contrast is very complex. It is therefore very difficult to trace the spicules from outside the limb back onto the solar disk in order to unambiguously directly identify disk structures with spicules. It is, however, generally accepted that the dark, narrow, elongated 'mottles' seen in monochromatic images in the wings of the Hα line (see figure 2) are to be identified with chromospheric spicules in projection on the solar disk. The $(1 - e^{-\tau})$ term in the cloud model is also wavelength dependent because the optical thickness τ is. It does not affect the sign of the contrast, only its magnitude.

Given this identification of spicules as seen against the solar disk, one can determine a number of other important

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 2. Monochromatic image of the Sun in the wing of the hydrogen Balmer α line of a region near the solar limb. The lower half of the image shows a small active region. Closer to the solar limb in the upper half of the image one recognizes the patterns of dark elongated mottles which are commonly identified with the chromospheric spicules as observed outside the solar limb. Note that they are arranged in 'hedgerow' or 'porcupine' like patterns which form the so-called chromospheric network. This network, foreshortened when seen near the solar limb, surrounds the supergranulation convection cells described elsewhere. (Courtesy R B Dunn, National Solar Observatory at Sacramento Peak.)

parameters. It is clear from images like those shown in figure 2 that spicules are not uniformly distributed across the solar disk. They are located in the so-called chromospheric network which borders the photospheric granulation cells (see CHROMOSPHERE: NETWORK). Although it is estimated that there are about 10 times fewer spicules than granules on the Sun, the number density of granules and spicules, as well as their birth rate, within these network areas are comparable. This suggests a physical relation between the granulation structure and spicule occurrence in these regions, with the well known local solar magnetic field enhancement in these network elements of about 100 G almost certainly playing an important role. However, attempts at establishing observationally such a connection have so far failed. The suspicion is that limited spatial resolution and possibly spicules originating well above the photosphere may be the cause of these failed attempts. But it may be, of course, that there is no such connection either with the granules themselves or with the intergranular lanes.

Role of spicules

What role do spicules play in relation to other solar aspects? From the physical parameters it is easy to estimate the kinetic energy flux and mass flux in the combined spicules. In the upward moving spicules the kinetic energy flux is negligible compared with other energy transfers to and from the corona. It is a factor of 100 less than, for example, the conductive energy losses downward from the solar corona. The mass flux is a totally different matter. In the upward moving phase the spicule mass flux exceeds the mass loss of the solar corona through the SOLAR WIND by two orders of magnitude. In that sense spicules may be of broader interest rather than being a phenomenon of only parochial interest.

No generally accepted magnetohydrodynamical model exists for spicules. A number of models have been proposed using a variety of physical mechanisms including different kinds of wave phenomena, instabilities, magnetic reconnection, etc based on velocity and magnetic fields associated for example with granules and supergranules. They all claim to explain many of the observed spicule properties. Improved observations of spicules will be needed to distinguish between these models. New tools like solar adaptive optics and high-resolution UV observations which are likely to be available in the near future are likely to clarify the situation.

There are larger structures visible especially well in ultraviolet lines like the He^+ Lyα line at 30.4 nm which are similar in behavior to spicules. They are referred to as 'super-spicules' and are mostly seen at high solar latitudes. Attempts to relate these structures to Hα images of the chromosphere and prominences have failed. Occasionally exceptionally long spicules in Hα are also seen in the solar polar regions. They have been called 'macro-spicules'. Little is known about the behavior and origin of either of these larger structures.

J M Beckers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chromosphere: Thermal Bifurcation

Thermal bifurcation refers to the existence of 'cool clouds' of gas ($T < 3500$ K) at high altitudes in the solar atmosphere, within what was thought to be a uniformly hot ($T \sim 7000$ K) CHROMOSPHERE. It also refers to a cooling instability, promoted by carbon monoxide molecules, which operates in places where the chromospheric mechanical heating is weak. The cool clouds were discovered through the anomalous darkening of strong 4.7 μm CO absorption lines near the extreme limb; first recognized in the early 1970s, and more recently confirmed in the surprising off-limb emissions of the infrared CO bands. Because carbon monoxide is both the principal diagnostic of the cool gas and a possible contributor to its formation, the low-temperature protrusions into the chromosphere have been deemed the 'COmosphere'. The underlying cause of the cool gas and how much of the volume of the low chromosphere it might occupy remain controversial.

Background

In the early 1970s, R W Noyes and D N B Hall reported curiously depressed brightness temperatures of carbon monoxide absorptions in infrared (4.7 μm) spectra taken close to the edge of the solar disk. The 'cool cores' of the CO lines—as low as 3700 K—were surprising because they were as cool as the coldest temperatures thought to occur in SUNSPOTS, arguably the chilliest features of the solar atmosphere. Noyes and Hall also noted striking intensity fluctuations of the infrared CO lines, presumably in step with high-altitude compressions and rarefactions caused by the global p-mode oscillations.

At the time, a controversy had developed whether the minimum temperature at the base of the chromospheric temperature inversion was as hot as the 4400–4500 K indicated by Ca II 393 nm and Mg II 279 nm or instead the 4200 K suggested by continuum intensities in both the far-ultraviolet and submillimeter regions. During the $T_{\rm min}$ debate, the curious discovery by Noyes and Hall was dismissed as an unimportant oddity. The authors themselves had attributed the dark cores of the strong CO lines to an effect connected with the narrow, cool subduction lanes of the photospheric granulation pattern. In many minds, the anomalous CO temperatures thus were associated with a small-scale phenomenon of the SOLAR PHOTOSPHERE, of questionable relevance to the global atmosphere, and certainly not to the overlying chromosphere.

In the late-1970s, a powerful new optical–infrared instrument—the 1 m Fourier transform spectrometer (FTS)—was commissioned at Kitt Peak. The $T_{\rm min}$ debate had rekindled interest in alternative ways to trace the thermal stratification of the outer photosphere. T R Ayres (a proponent of the 'hot' $T_{\rm min}$) and L Testerman revisited the anomalous extreme-limb behavior of the CO lines, confident that the high-precision FTS measurements would bring the CO $\Delta v = 1$ bands into agreement with other $T_{\rm min}$ proxies. However, not only did Ayres and Testerman confirm the earlier result of Noyes and Hall, but also they concluded that the CO anomaly could be explained most naturally by adding a new layer to the solar atmosphere; a pervasive cold component overlapping the lower reaches of the traditional hot chromosphere.

In parallel, Ayres offered an explanation for the out-of-character cool high-altitude gas. His 'thermal bifurcation' mechanism was an elaboration of the key role of CO surface radiative cooling—something like a reverse greenhouse effect—originally described independently by T Tsuji and H R Johnson in the context of the photospheres of red giant stars. The idea is as follows. The vibration–rotation transitions of CO very efficiently convert gas collisional energy into infrared photons. Those photons that escape the region carry away thermal energy and thus cool the gas. When there is little or no mechanical heating in the outer layers of the photosphere, the strong surface cooling in the 4.7 μm CO bands drives the gas to a low-temperature state, balancing the mild heating in the H^- continuum (optically thin absorption of mainly visible photons from the background photospheric radiation field). If a small amount of mechanical heating is introduced on top of the slight radiative component, however, the gas will warm and the CO molecules will begin to dissociate. The loss of CO reduces the cooling, leading to additional warming and further dissociation. The ensuing thermal runaway ('heating catastrophe') halts only above about 6000 K, where partial ionization of hydrogen liberates enough electrons to collisionally energize high-temperature cooling species such as Ca^+ and Mg^+. If the mechanical heating then is removed, the gas first cools down to the H^- boundary temperature ($\sim$4900 K), at which point CO begins to form, which cools the gas further; more CO forms, and so forth, leading to a 'cooling catastrophe' that returns the region to the low-temperature branch of the 'bifurcated' energy balance.

In this view, the outer atmosphere of the Sun can exist in two rather different thermal states—one cold, the other hot—depending on the amount of mechanical energy deposited locally. The thermal bifurcation mechanism naturally operates only at high altitudes; once the CO $\Delta v = 1$ bands become optically thick, the infrared photons are trapped in the source region and the radiative cooling is strongly suppressed.

In the thermal bifurcation picture, the classical hot chromosphere is confined to small-scale overheated structures bathed in a pervasive cool high-altitude layer close to radiative equilibrium. Figure 1 is a cartoon contrasting the historical and new views of the solar outer atmosphere.

Photosphere–chromosphere interface

The classical picture, illustrated in the left-hand panel, is a plane-parallel stratification; a hierarchy of distinct layers, controlled by different energy budgets, but through which the temperature runs smoothly. The layered atmosphere represents the idealized response of a hydrostatically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

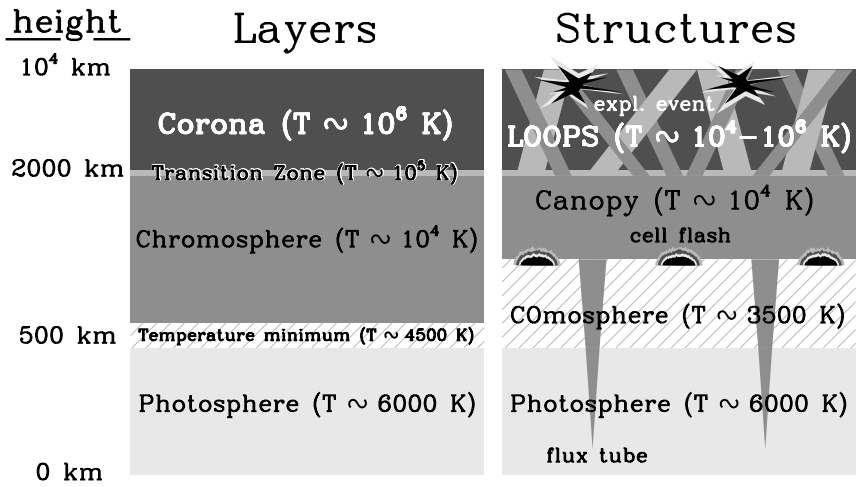


Figure 1. Classical and modern views of solar atmosphere. Height '0' is at $\tau_{500\ \mathrm{nm}} = 1$.

stratified gas in which the mechanical heating is uniform horizontally and varies slowly with height. The key layers and their interrelationships are described elsewhere in this volume.

The new paradigm—structures—is illustrated on the right-hand side of figure 1. It is not particularly new at all, but has been around in one form or another ever since the first high spatial resolution pictures of the chromosphere and corona were taken decades ago.

A jumble of magnetic fields—organized into a hierarchy of loop-like features—constitutes the CORONA. Its underside terminates in a canopy, occupying what previously would have been called the upper chromosphere (see CHROMOSPHERE: MAGNETIC CANOPY). The canopy itself is further rooted mainly in the supergranulation network by slender funnel-shaped flux tubes. These penetrate down through the photosphere into the convection zone, ultimately connecting to the deep source region of the magnetic flux (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES). Interactions between the tangled fields in and above the canopy give rise to localized explosive events traced by high-velocity jets in far-ultraviolet lines (see TRANSITION REGION: EXPLOSIVE EVENTS).

Much of the 'chromospheric' emission forms in the canopy, threaded by warm gas with large horizontal density contrasts. Additional contributions come from the downward extensions of the flux tubes, which evidently are hotter than the surrounding 'quiet' medium into at least the middle photosphere. A third component arises from cell flashes at the base of the canopy. These transient events are seen commonly in the interiors of the supergranulation cells. They are thought to be caused by outward-running acoustic disturbances that steepen and form shocks at high altitudes.

The canopy is ubiquitous and its density striations can be traced in Hα filtergrams. The cell flash and low-altitude flux tube components are compact, giving rise to transient and persistent bright points, respectively, in Ca K 393 nm spectroheliograms.

Between the upper photosphere ($h \sim 500$ km) and the base of the canopy ($h \sim 1000$ km) lies the COmosphere, the cool zone mentioned earlier, replacing what historically was designated the low chromosphere.

The essential aspects of the thermal bifurcation phenomenon were verified in the late 1980s in a series of *ab initio* model solar atmospheres constructed by L Anderson and R G Athay. The authors found that very low temperatures ($T < 3000$ K) are possible at high altitudes (dominated there by CO cooling); the CO-cooled zone can persist in the face of low-level mechanical heating but it disappears abruptly once the heating exceeds a critical level. At the same time, 2D radiation–hydrodynamic simulations of convection, with high-altitude CO cooling, by M Steffen and D Muchmore also demonstrated that the strong surface emission in the $\Delta v = 1$ bands can maintain a cold ($T \lesssim 3000$ K), horizontally uniform layer above the classical $T_{\min}$ in the face of the fluctuating granular radiation field and overshooting.

Beyond the Sun, the thermal bifurcation mechanism was recognized as one facet of the more general phenomenon of autocatalytic 'molecular cooling catastrophes'. These operate in a diversity of cosmic environments; for example driving the fragmentation and collapse of molecular clouds, or forging silicate grains in the bloated outer envelopes of red supergiants.

Current research

In 1992, inspired by the growing interest in infrared solar physics—not only the CO controversy but also the major benefits for magnetic field studies—the National Solar Observatory (NSO) installed a large infrared grating in the 13.8 m main spectrograph of the 1.5 m McMath–Pierce telescope. The mosaic grating, one of the largest in the world, was salvaged from the mothballed horizontal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

infrared spectrograph where Noyes and Hall had carried out their pioneering observations of the CO 4.7 μm bands two decades earlier.

Off-limb emissions

In April 1993, on a day of particularly good seeing during engineering tests on the newly refurbished spectrograph, W Livingston of the NSO staff discovered the remarkable off-limb emissions of the CO lines.

The limb corresponds to the point on the projected disk of the Sun where the line of sight becomes transparent in the continuum; thus the background intensity falls away. The CO lines can be in emission beyond the continuum 'edge' if they still are optically thick along the raypath: the thermally radiating—albeit cool—CO lines shine against essentially dark sky. (On the disk, the glare of the hot continuum radiation causes the cool CO cores to appear in 'absorption'.) The observation showed directly that molecular gas must exist well above the altitude where conventional models place the chromospheric temperature inversion; simulations suggest that cool gas is present up to perhaps 1000 km.

Figure 2 depicts the off-limb emissions in the 4.67 μm region. The observations were obtained during the morning of 9 May 1996, under clear skies with moderately good seeing. They were acquired with a long-slit cryogenic camera system that superseded the single-channel scanner used by Livingston. The spectrograph slit was placed across the north limb, perpendicular to it. The figure represents the coaddition of the five sharpest spectrograms from a run of 1000 short exposures. The coadded frame was divided by an average on-disk spectral trace to normalize the intensities and suppress absorptions from the Earth's atmosphere.

The CO lines are the mounds of emission, extending beyond the bright continuum edge, in figure 2 (top); each 'bump' corresponds to an individual CO transition, all observed simultaneously, at the same point on the limb. Figure 2 (bottom) depicts the translimb extensions, measured bin by bin across the spectrum; '0' refers to the apparent continuum edge. Although the maximum displacements in the cores of the strongest CO lines ($\sim 0.7''$) are only about half the effective instrumental resolution (because of seeing, diffraction and scattering), they can be measured accurately relative to the continuum. (In the same way that tiny parallaxes—hundredths of an arcsecond—can be measured on plates with spatial resolution no better than 1 arcsec.)

The off-limb CO emissions are built up along an enormous pathlength through the low chromosphere. They tell us directly that there is some cool gas present somewhere along the sightline, but not where or how much. The key question is how widespread the cool gas truly is.

Surface maps of temperature and velocity

By sequentially stepping the McMath–Pierce telescope, the new long-slit camera system allows one to build up strip-maps of key CO line parameters: central absorption intensity (related to surface temperature), Doppler shift (a measure of line-of-sight velocity) and line broadening (tracing local turbulence). The spatial scanning can be done repetitively to make 'movies'.

First use of the new system for spectral imaging of CO was in October 1993 by H Uitenbroek, R Noyes and D Rabin. Additional CO thermal and velocity mapping was conducted by T Ayres and D Rabin the following spring. Both studies concluded that the evident patchwork quilt of spatial variations in the CO brightness is associated mostly with the oscillation interference pattern, particularly at high altitudes (i.e. in strong CO lines). Occasional persistent CO 'bright points' were seen, coincident with Ca K network fragments (regions of strong magnetic field arrayed along the stagnation points of the large-scale horizontal flow pattern of the supergranulation). There was no evidence in either study for a significant population of very dark points. That absence constrains alternative explanations for the anomalous limb-darkening of the CO bands.

How pervasive is the cool gas?

To reproduce the extreme-limb darkening and off-limb emissions of the CO line cores requires either (1) cool material with a high filling factor (say, $\sim$80%) at high altitudes or (2) a smaller covering fraction (say, $\sim$20%) of much colder structures. In the first case, the intensity behavior is controlled everywhere by the pervasive cool component, and an image looking straight down on the solar surface should appear relatively bland and featureless, apart from the relatively small fluctuations due to the p-mode pattern. In the second case, the isolated cold, but transversely opaque, 'pillars' dominate only near the limb where they effectively 'shadow' any hot structures (the latter naturally are transparent in CO). At disk center, however, the cold regions (with $\Delta T \sim 1000$ K) should show up as small-scale 'dark spots' with the $\sim$20% filling factor. So far—within the sustained few-arcsecond resolution encountered under good conditions at Kitt Peak—nothing remotely like the expected sprinkling of dark points has been seen. The opposite scenario—pervasive COmosphere—thus appears to be more promising.

Dynamical COmosphere?

An alternative view of the COmosphere has been proposed by M Carlsson and R F Stein. One-dimensional radiation–hydrodynamics simulations suggest that the 500–1000 km altitude range might be in a continually evolving thermal state owing to the passage of steepening acoustic wave-trains from below. At those low densities, the CO chemical formation timescales are long. The molecular population cannot respond to the instantaneous temperature perturbations, but instead attains an equilibrium based on the time-average thermal profile (rather cool in the Carlsson and Stein model). Such a steady-state background of CO, existing in a fluctuating thermal environment, might be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

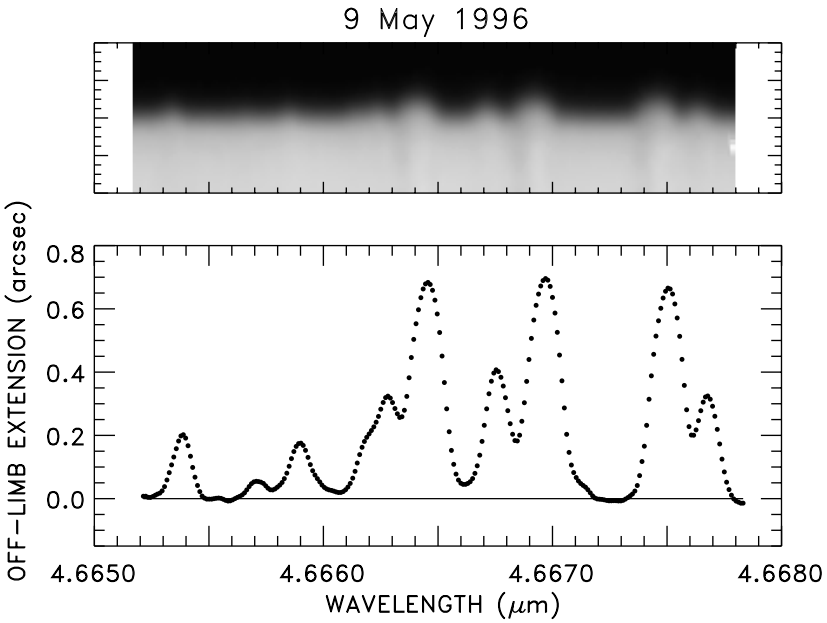


Figure 2. Off-limb emissions of CO lines.

capable of explaining the off-limb emissions and extreme-limb darkening of the infrared $\Delta v = 1$ lines. At the same time one would record significant time-averaged ultraviolet signatures from the same spatial volume (mostly from the high-temperature peaks, owing to the exponential sensitivity of the thermal emission (planck function) at short wavelengths).

A second dynamical explanation has been offered by H Uitenbroek and R Noyes. They suggest that the COmosphere is a byproduct of the expansive cooling of convective granules overshooting into the stable layers of the high photosphere.

The observational consequences of these proposals should be pursued further. In addition, the CO chemistry (and cooling) should be incorporated in the dynamical simulations.

Implications

In the final analysis, the COmosphere very likely is a relatively passive zone of the outer atmosphere, really just an outward extension of the photosphere. Its importance comes in how one attributes the radiative activity of different levels of the atmosphere in a global assessment of the energy budget. The gas density is large in the 500–900 km range compared with the higher-lying canopy zone. Because the radiative cooling depends on $\sim n^2$, the intermediate altitude range plays a strategic role in the energy budget. One can imagine that a calculation of total radiative losses (which must balance the input heating, and therefore are a proxy for it) will yield very different answers if there is a temperature *rise* in those layers (classical chromosphere) versus a temperature *drop* (COmosphere). If the thermal model is biased, gross errors in the inferred mechanical energy input are possible. (The source of the chromospheric heating is a longstanding mystery in solar–stellar physics.)

Finally, it should be mentioned that thermal bifurcation effects are relatively weak in G-type stars like the Sun and probably absent among the even warmer F types. That is because at a critical $T_{\rm eff}$, a few hundred kelvins hotter than solar, the H^- boundary temperature becomes too warm for a seed population of CO molecules to form and trigger a cooling cascade. Conversely, the mechanism becomes increasingly important toward the later spectral classes, especially prominent among the cool, low-gravity red giants.

Summary

The stratification of the solar atmosphere once was thought to be well understood, based on a wealth of ultraviolet and optical diagnostics. Over the past two decades, however, studies of infrared molecular spectra have unearthed troubling anomalies. Strong lines of the 4.7 μm fundamental bands of carbon monoxide display very cool brightness temperatures at the extreme limb and curious off-limb emissions protruding hundreds of kilometers into the supposedly hot chromosphere. A straightforward—but controversial—proposal is that the low chromosphere is not hot at all, but instead is permeated by CO-cooled 'clouds', a COmosphere, if you will.

Bibliography

Anderson L S and Athay R G 1989 Model solar chromosphere with prescribed heating *Astrophys. J.* **346** 1010–18

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ayres T R 1981 Thermal bifurcation in the solar outer atmosphere *Astrophys. J.* **244** 1064–71

Ayres T R and Rabin D 1996 Observations of solar carbon monoxide with an imaging infrared spectrograph I thermal bifurcation revisited *Astrophys. J.* **460** 1042–63

Carlsson M and Stein R F 1995 Does a nonmagnetic solar chromosphere exist? *Astrophys. J.* **440** L29–L32

T R Ayres

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cidenas (fl. fourth century BC)

Babylonian astronomer, flourished at Sippra, discovered the precession of the equinoxes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Circinus

(the Compasses; abbrev. Cir, gen. Cirini; area 93 sq. deg.) A southern constellation which lies between Centaurus and Triangulum Australe, and culminates at midnight in early May. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2.

A small, inconspicuous constellation, the brightest stars in Circinus are α Cirini, a binary with white (F1) and orange (K5) components, magnitudes 3.2 and 8.6, separation 15.7″, and β Cirini, magnitude 4.1. Other interesting objects include NGC 5315, a tenth-magnitude planetary nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Circumpolar Stars

Stars that never set as seen from a particular location. The requirement for this to happen is that the star's polar distance is less than the observer's latitude—thus from a location in latitude 52° North, stars of NPD less than 52° (i.e. with declinations of between 0° and +38°) are circumpolar and will be seen to circle around the north celestial pole. In practice it is difficult to observe objects very close to the horizon, so a more realistic requirement would be for the star's polar distance to be at least 5° less than the latitude. Observed from the Earth's geographical poles all stars are circumpolar, whereas from the Equator none are circumpolar. Circumpolar stars transit the meridian twice a day, at upper and lower culmination alternately.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clairaut, Alexis Claude (1713–65)

French mathematician, infant prodigy, wrote *Théorie de la Figure de la Terre*, a treatise dealing with the shape of rotating solid bodies, calculated the effect that Saturn and Jupiter would have on the 1759 appearance of HALLEY's Comet and predicted its return to within a month. He recalculated NEWTON's theory of the motion of the Moon, wrote *Théorie de la Lune*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clark, Alvan (1804–87), Clark, George Bassett (1827–91), and Clark, Alvan Graham (1832–97)

Alvan was an astronomer, telescope maker, born in Ashfield, MA. Founded, with his sons George (born in Lowell, MA) and Alvan Graham (born in Fall River, MA), Alvan Clark & Sons, makers of optical lenses for telescopes, first significant American contribution to astronomical instrument-making. On testing a lens on a newly manufactured telescope in 1862 Alvan Graham Clark discovered the white dwarf companion of Sirius, Sirius-B.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clarke, Arthur C (1917–)

Science fiction writer and space visionary, born in Minehead, Somerset, England. Worked during the second World War on radar ground controlled approach equipment, became president of the British Interplanetary Society. In 1945 he wrote a paper 'Extra-terrestrial relays' which described the principle of satellites in geostationary orbits. The geostationary orbit at 42 000 kilometers has been named *the Clarke orbit* by the International Astronomical Union. Clarke inspired satellite applications for weather forecasting. He developed the concept of the rotating space station, pictured by Stanley Kubrick in the movie *2001: A Space Odyssey*, scripted by Clarke.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Classification of Stellar Spectra

How does a scientist approach the problem of trying to understand countless billions of objects? One of the first steps is to organize the data and set up a classification scheme which can provide the best insights into the nature of the objects. Perception and insight are the main purposes of classification. In astronomy, where there are 'billions and billions' of stars, classification is an ongoing process, with new discoveries either being incorporated into the classification scheme or remaining isolated as peculiarities worthy of further study. With new telescopes and new detectors opening fresh windows to the universe, the need for classification is likely to persist as an essential ingredient of astronomical progress.

If we were able to travel quickly among the stars, we would notice their size, color and surface brightness first. Being earthbound, we can see only the Sun in detail, but there are many different criteria which we can use to classify stars and other objects: position, brightness, polarization, color, spectrum, surface brightness, movement and variability. Of these, the spectrum yields the most information about the nature of the star.

Astronomy is not an experimental science, stories of alien abductions notwithstanding. Unlike most scientists, astronomers are limited to observations from a distance. Stars and galaxies cannot be heated over a Bunsen burner; chemicals cannot be added to a star in order to learn from the reaction. Cosmic experiments do occur, but the timing is serendipitous, so the observer must be patient and watch, with no choice over the parameters of the event. To maximize the chances of catching a cosmic experiment, large surveys of stars and galaxies have been and will be conducted.

Because of this need to organize large sets of observational data, a sophisticated classification philosophy was developed by W W Morgan at the Yerkes Observatory of the University of Chicago, and was given the title 'The MK process'. Interestingly, the astronomical development anticipated the thinking of the Bavarian philosopher, Ludwig Wittgenstein, on the subject of the use of specimens in classification (Wittgenstein 1953).

The most important function of a classification system is to provide a general reference frame into which most objects to be studied can be fitted. Statistics based on the classification scheme can provide powerful insights into the nature of the objects under consideration. Trends can be observed among the normal objects and rare or pathological objects can be identified easily. Isolating peculiarities is also important in astronomy, because the detailed study of peculiar stars often provides new insight into the nature and behavior of normal stars. Classification is the root of our understanding of stars and STELLAR EVOLUTION.

In addition to the development of a general reference frame and the discrimination of peculiar spectra, classification techniques can be used very successfully for the determination of distances to stars. Although the method is referred to as 'spectroscopic parallax', the name is misleading because the technique does not use actual trigonometric PARALLAXES directly. It is based on the assumption that *stars with similar spectra have similar fundamental parameters*. This assumption allows astronomers to use nearby stars with good trigonometric parallaxes to provide a calibration which can then be applied to distant stars with similar spectra. Until recently, moderately hot or large stars, which are among the apparently brightest stars because of their intrinsic luminosity, but which are actually much more rare per unit volume than stars like the Sun, had no counterparts near enough in distance for the technique of trigonometric parallax to be effective, so indirect methods were developed. The indirect methods include the use of DOUBLE STARS and STAR CLUSTERS for bootstrapping from common stars with known distances to the uncommon hot, HIGH-LUMINOSITY STARS by assuming the same distances for the members of a cluster. Using data from the astrometric satellite HIPPARCOS, it is now possible to extend the distance calibrations by a factor of 10, which allows direct calibrations of red and yellow giant stars as well as many of the stars hotter than the Sun; however, the supergiants and the hottest stars are still beyond the reach of high-quality trigonometric parallaxes. For the luminosity calibration of these rare but luminous stars, binary and cluster-fitting techniques are still used.

The process used for classification is comparison of the unknown spectrum with the known spectra of well-studied standard stars (specimens), using pattern-recognition techniques involving all the features in a stellar spectrum. When good-quality spectra are used, and the unknown star is carefully compared with a good set of standard stars, it is possible to obtain very reliable and informative results. Poor results are usually due to an inadequate or inappropriate grid of standard-star spectra.

The history of classification of stellar spectra

Father ANGELO SECCHI (1818–78), a Jesuit astronomer working in the 1860s at the Vatican Observatory, was the first to provide a general classification scheme for stellar spectra. He looked at thousands of stars with a visual spectroscope and small telescopes (15–25 cm aperture), but his observations were not photographic, so the blue–violet region of the spectrum was not prominent, mainly because the human eye is not very sensitive to deep blue wavelengths. The coolest stars dominate his classification scheme, forming two of his four classes, partly because they are among the apparently brightest stars in the visual region of the spectrum, and partly because of the striking appearance of their molecular bands. In modern terms, his four classes are

I. stars hotter than the Sun,
II. stars like the Sun,
III. stars cooler than the Sun with titanium oxide absorption bands in the yellow and red regions of the spectrum, and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

IV. stars cooler than the Sun with carbon molecular absorption bands in the yellow and red regions of the spectrum.

The class IV stars are easily distinguished from class III by their redder color, by the positions of the bands and by the appearance of the carbon bands, which fade in the opposite direction from those of titanium oxide.

By the end of the 19th century, Edward C Pickering and others realized the importance of photographic spectra and objective-prism telescopes for classifying very large numbers of stars using a uniform scheme with consistent criteria. The 'HD' project at Harvard College Observatory was sponsored by Henry Draper and carried out primarily by ANNIE JUMP CANNON, who by the mid-1920s had classified more than a quarter of a million stars on a homogeneous, one-dimensional system. The HD system, which now corresponds more or less to the temperature scale of modern astronomy, originally was based chiefly on the strength of the hydrogen lines and types were assigned in alphabetical order: A, B, C, ... Later, when it was discovered that the hydrogen lines reach a maximum strength (A-type stars) at temperatures lower than those of B-type stars and O-type stars, and that some of the letters are redundant, the sequence of letters was changed to the order used today for the temperature reference frame: OBAFGKM(SRN). The SRN types are in parentheses because their temperatures are parallel to the GKM stars, the spectra differing in chemical composition rather than temperature.

There are various mnemonics devised for (and by) students to help remember the sequence, but the most common is 'Oh, Be A Fine Girl (Guy), Kiss Me Sweetly Right Now'. Dr Nancy Houk, at the University of Michigan, is reclassifying all of the HD catalogue stars on the two-dimensional MK system. The results of her work in the first four of six volumes are shown in figure 1.

Also at Harvard, Antonia Maury isolated a class of peculiar stars which she classified with a lowercase 'c' placed in front of the temperature class. Her important insight formed the basis of the second dimension, which corresponds to surface gravity and is related to the luminosities of stars.

Many other systems have been proposed, but most of them are intended for the classification of stars in particular regions of the HERTZSPRUNG–RUSSELL DIAGRAM and are used to advantage only in that part of the HR diagram. The BCD system of Barbier, Chalonge and Divan, which measures the Balmer jump directly, is a good example; it is powerful for its purpose, but is not good as a general reference frame. Others for a variety of reasons have not achieved wide acceptance. The Mount Wilson system developed by Adams in the 1930s is another example; in that case the calibration was used for the classification, so the parameters changed whenever a new model was introduced. Adams' system became very cumbersome and inefficient. For further historical examples, see the article by Kurtz in *The MK Process* (Garrison 1984) and the excellent history of spectroscopy by Hearnshaw 1986: *The Analysis of Starlight*.

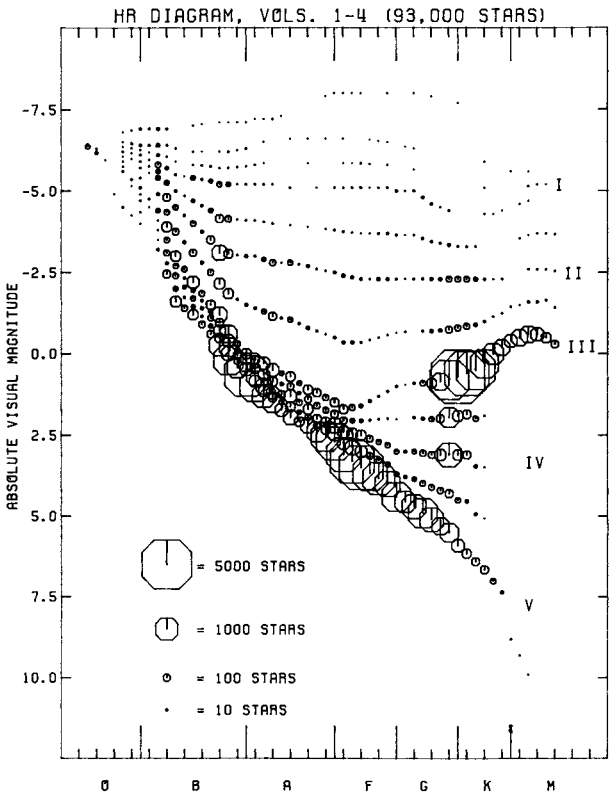


Figure 1. Hertzsprung–Russell (HR) diagram for all the stars contained in the first four of six volumes of the Michigan Spectral Catalog, a massive undertaking by Nancy Houk. It is an illustration of how MK spectral types can be used as a general reference frame to study the characteristics of stellar populations.

At the University of Chicago's Yerkes Observatory in 1943, W W Morgan, P C Keenan and E Kellman proposed a system of classification of stellar spectra known as the 'MKK system'. The 'MK system' (revised by Morgan and Keenan in 1953 from the MKK) is virtually the only one in wide use today as a general reference frame. Its success is due in large part to the philosophy of classification developed by W W Morgan—the MK process, of which the MK system is a subset. The general reference frame for the MK system is illustrated in figures 2 and 3.

The MK system

Classification systems are created for a reason. In using a particular system, it is important to be aware of why it was created, how it was constructed, what its useful limits are, how it has evolved and what credibility it has achieved. Collectively these correspond to the mandate of the system. The mandate of the MK system is to describe the blue–violet region of the spectrum of a star in terms of a set of standard stars, without reference to photometry, theory or other information.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

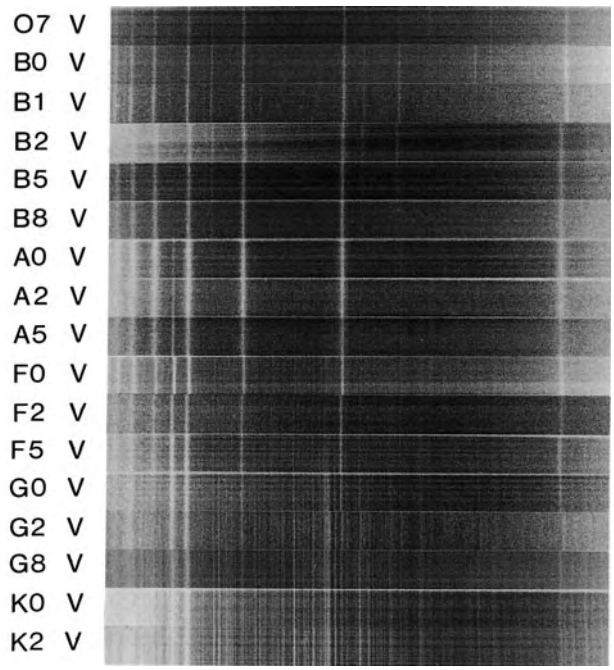


Figure 2. Montage of photographic spectra of standard stars illustrating the temperature classes for stars along the main sequence of the HR diagram. The sense of the sequence is that the O-type stars are hot and the K-type stars are cool. Notice that there are considerable differences in pattern along the sequence. Temperature is the primary variable in stellar spectra.

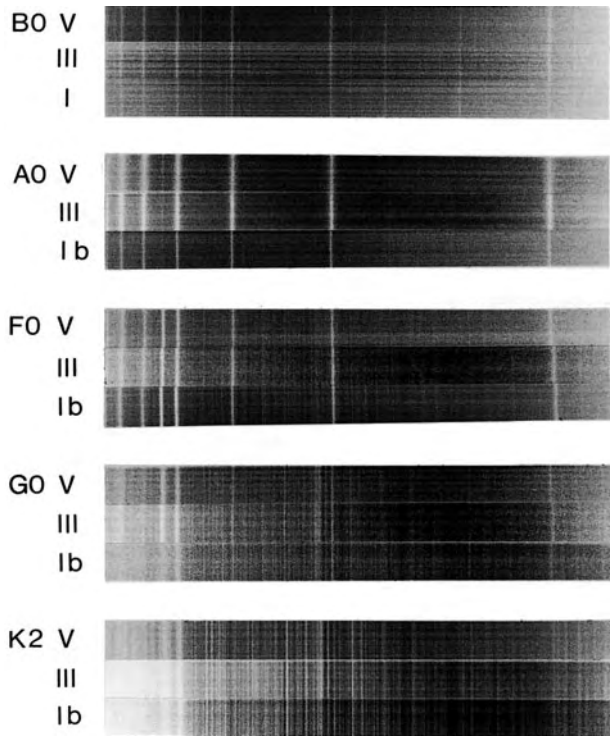


Figure 3. Montage of photographic spectra of standard stars illustrating the luminosity class sequences (V, III, I) for key spectral types (B0, A0, F0, G0, K0). The sense of the luminosity class sequence is that class I is brightest and class V is faintest for a given spectral type. Notice that the differences among the luminosity classes are somewhat more subtle than those for the temperature types, but are still quite easy to distinguish.

The value of the separation of the MK system from photometry and theory initially is that, at the interface between any two independent systems, there is information which is lost if one system is allowed to influence another. It is much more productive to compare two autonomous systems, assuming that both are carefully done, than to declare one or the other superior. For example, there is not a one-to-one relationship between the photometry and the spectrum, although, in general, the systems are correlated and can be used for a calibration. Another example is theory, which involves many simplifying assumptions and is continuously evolving (as it should). In this case, the derived parameters change with new advances in the models, so it is best not to mix the theory and the classification *a priori*.

The use of line-strength ratios helps with the comparison of spectra and enables a high level of discrimination of spectral types and luminosity classes. The MK system uses the following nomenclature.

(a) The Harvard sequence (OBAFGKMSRN) is retained for the major divisions of the temperature axis (figure 2), in order to preserve some continuity with the great work of the Henry Draper Catalog. O-type stars are the hottest and most massive, whereas M stars are the coolest. We know now that the carbon (R and N) stars are approximately the same temperature as the G, K and M stars except that the ratio of carbon to oxygen is reversed and that the S stars have similar temperatures to the M stars except that they show ZrO absorption bands.

(b) Each letter is followed by subdivisions 0–9, except for the O-type stars for which O3 is the hottest type.

(c) A luminosity-related class is added, representing indirectly the absolute brightness of the star, as indicated by features sensitive to surface gravity and microturbulence. Roman numerals are used to represent the luminosity class, with the sense that supergiants are I, giants are III and main sequence stars (called dwarfs, although they are similar to the Sun) are V (figure 3). Sometimes the subdivisions a, ab and b are used to refine the luminosity classification in the sense that ab is the middle of the box and a is among the brightest stars in the box.

(d) There is no single third dimension representing abundance because, in peculiar stars, elements can be individually affected. Thus the peculiarity is noted by the element's symbol (Fe, Ba, C, CH, CN, etc), followed by a number 1–5, which is positive for overabundance and negative for underabundance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(e) When there are enough specimens in a peculiar class, the definition of 'normal' may be expanded to include them. This happened with supergiants, rapid rotators, emission-line stars and others. Thus, there are several additional symbols used in the classifications. A few examples will illustrate the principles. Rapid rotators are designated by the letters n or nn, depending on the rotation rate. Stars with hydrogen lines in emission are labeled with the letter e, which signals a detailed description of the star's spectrum in the notes; since emission lines are usually variable on time scales of years, the details are not given in the general spectral type but rather in the notes, accompanied by the date. Hot, O-type stars with helium, carbon, nitrogen or oxygen in emission are marked with the letter f. Very hot stars losing mass at a prodigious rate are classified as WOLF–RAYET (WR) STARS.

(f) If a star does not fit into the system at all, it is given the designation of the standard specimen which matches it most closely and the letter p is attached, indicating a peculiar (and probably especially interesting) star. The peculiarity is then described in a note.

(g) Typical examples of classifications are as follows: G2 V (the Sun's classification), indicating a star of moderate temperature and mass which is in the stable hydrogen-fusing stage of its life; M6 IIIe, indicating a cool giant star with hydrogen emission which is probably fusing helium into carbon for energy, and B0 Ib (close to the classification of the star which later became supernova 1987A), representing a very hot, large, very massive star which is evolving rapidly (astronomically speaking); K0 IIIb Fe2, CH1, CN−2 signifying a moderately cool giant star, slightly less luminous than most giants, with strong metals (Fe2), somewhat strong G-band (CH1) as well as weak CN; A5p (Cr, Eu) describing a moderately hot star with strong lines of chromium and europium.

The MK process

The success of the MK system is due mainly to 'the MK process', an Aristotelian, inductive approach to classification in general, which is based on particular specimens (standard stars in the case of stellar spectra) and is independent of calibrations from theory or photometry. It is not based on theoretical models or averages as in a Platonic approach, but it is autonomous and is based on existing objects. The classification of a stellar spectrum is a description in terms of the set of standard stars, so the actual classification remains the same when the calibration changes owing to advances in photometric techniques, the theory of stellar atmospheres or other methods of determining temperature, surface gravity, atmospheric structure or chemical composition. Each classification box is represented by one or more specimens rather than by its physical parameters. For example, the standard star α PsA = HD216956 represents the box A3 V in the MK system. The spectrum of an unclassified star is compared carefully with that of the standard star and determined to be *like* or *not like* α PsA. If it is *like* α PsA, it is given the classification A3 V; if it is *not like* α PsA, it is compared with another standard star and the process continues until a match is found. This is not as tedious as it seems, because a quick perusal of the spectrum of the unknown star will tell the experienced observer (or a computer program) which group of standard stars is most likely to match. A current calibration of the box, based on the calibration of the standard stars, is then applied to determine the physical parameters for the unknown star. With the large number of stars which have been classified, it would be a nightmare to have to redo the work every time new physical data become available. The value of an autonomous classification methodology is that, when the calibration changes from year to year or person to person, the classification does not have to be redetermined.

The MK process methodology outlines a way to extend the MK system beyond its wavelength definitions and standards. New autonomous systems, independent of the MK system and of each other, can be created to cover new categories of objects in and beyond the Milky Way Galaxy.

New wavelength regions: extending the classification

The general reference frame for MK types is the blue–violet region of the spectrum, which was dictated originally by a combination of the old prism spectrographs which cut off the ultraviolet (UV) and the relative insensitivity of yellow, red and infrared (IR) emulsions in the 1930s. Even today, with satellites, high-UV-throughput instruments and CCDs which are very sensitive in the red and IR, the blue–violet remains the best choice for a general reference frame, because it contains many lines and bands with a wide range of excitation. For particular classification problems in particular parts of the HR diagram, however, it is often a great advantage to use other wavelengths.

The energy distribution for the hottest stars peaks in the UV where there are strong lines suitable for classification of O- and B-type stars. Classification criteria and atlases have been published by NASA astronomers Janet Rountree and Nolan Walborn. The UV spectra are influenced strongly by stellar mass loss (winds) and the assumption is made that mass loss is correlated with the luminosity of the star, which is not always strictly true; there are exceptions. Also, cool stars do not have much light in the UV, and their UV spectra are dominated by chromospheric lines. Thus, the UV can be used advantageously for particular projects involving hot stars but is not good as a general reference frame.

The energy distribution for the coolest stars peaks in the IR and falls off rapidly toward the blue–violet. The IR is the best region for molecular bands, which are the most useful criteria for cool stars. This is an advantage for classifying the temperatures of cool stars, but there are very few atomic lines in the IR and lines from the terrestrial atmosphere interfere strongly. Blue spectra are needed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

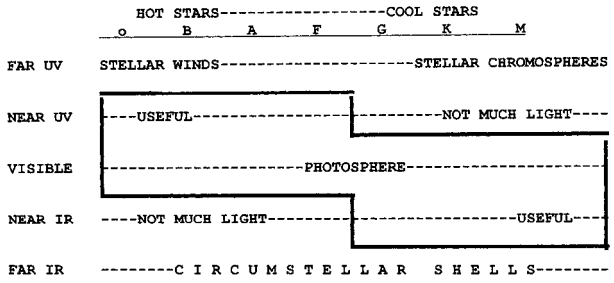


Figure 4. Table summarizing the choices of wavelength regions, as well as some of the general advantages and disadvantages for the various types of stars.

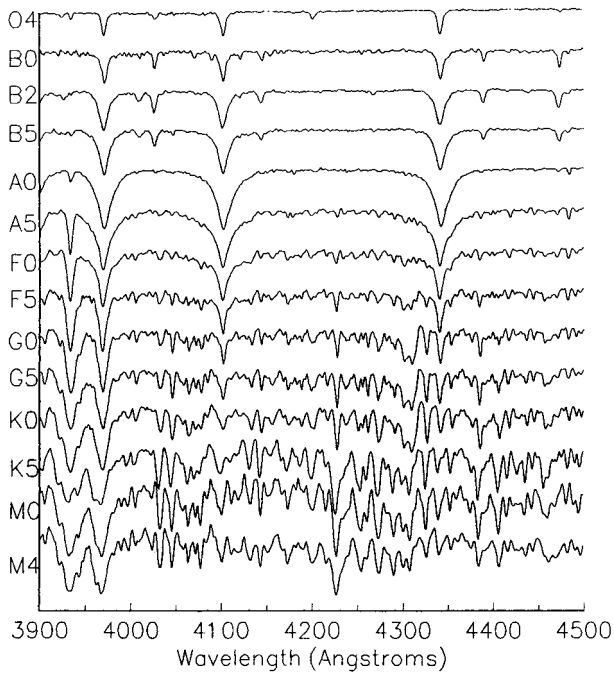


Figure 5. Similar to figure 2, but digital data have been used. The digital spectra cover a comparable wavelength region and have signal-to-noise ratios equivalent to those of the photographic spectra. The sequence of standard stars is an illustration of the spectral variations due to temperatures, for stars on the main sequence.

to properly determine the metallicities because there is much more variety of atomic lines of various excitations. Hot stars have less energy in the IR and the spectra are influenced strongly by circumstellar matter. Thus, the IR can be used advantageously for particular problems involving cool stars but is not good as a general reference frame. See figure 4 for a summary.

The digital dialect

Most pre-1980 classification atlases were photographic, but most of the work since then has been digital. CCDs are now used universally. There are some differences in use of the two media; however, the problems unique to the digital revolution have been worked out. Whereas photographic spectra are logarithmic (density) and two dimensional, digital spectra are linear and are usually represented as one-dimensional tracings (figure 5). In pattern recognition, the eye prefers to look at surfaces rather than line tracings. Digital detectors are much more efficient than photographic emulsions, although observers often underestimate the overall photographic signal-to-noise ratio (S/N) in determining equivalent digital S/N. The best widened photographic spectra used for classification have S/N of approximately 200–300, so the digital equivalent of the best photographic work should be similar. Many observers work with much less, and the accuracy of their results suffers. However, when working with extremely faint objects, it is necessary to compromise. Good digital classifications are possible with $S/N = 50$ and mediocre ones with $S/N = 35$, but below that the results are likely to be unreliable.

Another difference in approach is in the process of comparison with standard stars. Because of the non-linearity of the photographic process, the exposures (i.e. S/N) for the standards had to be similar. However, with linear digital data, it is advisable to take the bright standards at very high S/N (several hundred) even if it is only possible to take the unknown at very low S/N.

With digital data, care must be taken to prevent loss of information or introduction of false information during the reduction process, especially during rectification. Also, when comparing tracings, it is important to use ratios of close lines and to look at the line strength rather than the sharp bottoms of the lines, because, for strong lines, the bottom of the line is at lower S/N and the wings of the lines are far more reliable and important. Thus, digital data, while much more efficient, versatile and easy to work with than photography, must be used with care to give superior results.

Automating the classification process

The tradition in MK spectral classification involves comparing visually the entire blue–violet spectrum of a star to find a match with one of the standard stars. Humans are very good at pattern-recognition, which is difficult for computers. However, powerful computer techniques have been developed for automation by simulating the pattern-recognition process. (For a review, see the article by Kurtz in Garrison (1984).) The most successful methodologies, which allow computers to do as well as or better than humans on normal stars, are the minimum metric distance (MMD) and neural net (NN) approaches. Each has advantages and each has been tested on large data sets. Automation techniques can be applied to great advantage when dealing with data sets of millions of stars. In such cases, when new types of peculiarities are found, they are flagged for human consideration. This means not only that very large numbers of stars can be classified but that human experts still will be needed to interpret the results.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Natural groups

The MK system was set up to be used with spectra of high enough resolution to permit detection of the most important peculiarities, yet low enough to reach stars as faint as possible. This compromise is satisfied in the range of resolution 0.1–0.3 nm. There are situations, however, when it would be useful to have some information for stars as faint as possible, which would mean extending the MK process to lower resolutions. W W Morgan developed the concept of natural groups, which uses the strongest, most outstanding features of the spectrum to identify uniquely particular types of stars, even at extremely low resolutions. Examples are the A0 V stars, whose hydrogen lines are stronger than in any other normal type of star and can be detected at resolutions as low as 2–3 nm. Thus if the hydrogen lines are still visible when a spectrum's resolution is degraded to this amount, the star must be approximately A0 V. At very low (e.g. 3 nm) resolution the G8–K2 III stars are uniquely outstanding, because their temperatures and surface gravities are optimum for peak visibility of the broad CN bands. Similarly the coolest stars can be classified uniquely and easily at very low resolution by use of their TiO and VO bands in the red region of the spectrum and their CO bands in the IR domain. The powerful techniques of natural groups can thus be used as an initial filter to define a set of objects for further study but are not good for definitive types for individual stars because the three regions mentioned happen to be also the locations of many peculiar and interesting stars whose peculiarities are not visible at such low resolutions.

Integrated spectra: binaries, clusters and galaxies

The concept of natural groups also can be used to disentangle the complexities of integrated spectra of binary stars, globular star clusters and galaxies. If, for example, the CN bands are strong in the integrated spectrum of a galaxy nucleus, there must be a substantial contribution to the light by G8–K2 giant stars. MgH is a molecular band that is prominent only in cool dwarfs; if it is not visible in the integrated spectrum, the cool dwarfs must not be present in enough numbers to counteract the light from the giants, which, while much rarer, are also much more luminous. W W Morgan and his associates carried the idea to the limit, using extremely low resolution (1000 nm mm^{-1}), which produced spectra that resembled tadpoles. The Cyg OB6 Association was discovered using the tadpoles to reveal highly reddened, very faint, massive, hot stars, which were distinct from cool stars with similar colors. These natural-group techniques have not been fully exploited in the interpretation of the stellar content of clusters and galaxies.

Objective prisms, multiobject spectrographs and surveys

Taking individual spectra for large surveys is very time consuming. One way to improve efficiency is to use a wide-field SCHMIDT TELESCOPE fitted with a thin prism across the aperture, thus producing spectra of all the stars in a large area. In the early 1900s, Annie Jump Cannon used objective-prism spectra for assigning spectral types on the Harvard system to more than a quarter of a million stars over the entire sky. Dr Nancy Houk's reclassification of all the Henry Draper stars on the two-dimensional MK system is also based on objective-prism spectra.

Fiber optics are being used to obtain many spectra of faint objects simultaneously. Modern multiobject spectrographs can be used to take low-dispersion spectra of hundreds of stars or galaxies at a time, yielding thousands of spectra per night. Automated classification techniques could be used to great advantage by providing preliminary, on-line spectral types for the stars and signaling the observer when unusual spectra are detected (e.g. a supernova in a distant galaxy), but this idea has not been explored effectively yet.

Calibration: effective temperature, logarithm of gravity and abundances

The MK process is inductive. The classifications are based on a set of standard stars only and are independent of the results of theory, photometry, all other external influences. After the classifications have been performed, it is useful to calibrate the system, to tie it into the other methods as well as the fundamental parameters of $T_{\rm eff}$, log g and [Fe/H].

An autonomous approach such as the MK process is useful because there is information in the interface which would be lost if the methods were not *a priori* independent. With the new data from satellites such as HIPPARCOS, it is then possible to calibrate the luminosities for nearby stars and thus to obtain distances for all stars with similar spectra. The theory of STELLAR ATMOSPHERES can be used to calibrate spectral types, luminosity classes and abundance indicators in terms of the fundamental parameters of $T_{\rm eff}$, log g and abundances of key elements.

The principle of complementarity

When comparing the results of various approaches to the interpretation of stellar spectra in terms of the fundamental parameters such as $T_{\rm eff}$, log g and [Fe/H], it is most productive to think of the methods as complementary, with important information in the interface. For example, while there is general agreement between the carefully measured colors of a star and its spectrum, there is not a one-to-one correlation between photometry and the spectrum, mainly because they refer to different parts of the stellar atmosphere. The photometry refers primarily to the bottom of the photosphere, whereas the spectrum arises from several layers of the photosphere. Different results from the two techniques might indicate a peculiar structure in the atmosphere. If both techniques have been used carefully, there is usually a good reason for the difference and many types of peculiar stars have been discovered through the disagreement between their colors and their spectra.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Summary
An inductive process of classifying stellar spectra has been outlined. With careful attention to comparison with standard specimens, a high level of discrimination can be achieved. The MK process is a powerful tool for astrophysical insight. W W Morgan, creator of the MK system of spectral classification, has written 'It could be said, with some truth, that the spectral forms have helped to classify themselves, by showing where they are most comfortable'.

Bibliography

Corbally C J, Gray R O and Garrison R F 1994 *The MK Process at 50 Years: A Powerful Tool for Astrophysical Insight (Astronomical Society of the Pacific Conf. Ser. vol 60)* (San Francisco, CA: Astronomical Society of the Pacific)

Garrison R F 1984 *The MK Process and Stellar Classification* (Toronto: David Dunlap Observatory)

Garrison R F 1988 *Publ. Astron. Soc. Pac.* **100** 1036

Hearnshaw J 1986 *The Analysis of Starlight* (Cambridge: Cambridge University Press)

Keenan P C and McNeil R C 1976 *An Atlas of Spectra of the Cooler Stars: Types G, K, M, S, and C* (Columbus, OH: Ohio State University Press)

Morgan W W and Keenan P C 1973 Spectral classification *Ann. Rev. Astron. Astrophys.* **11** 29

Morgan W W, Abt H A and Tapscott J 1978 *Revised MK Spectral Atlas for Stars Earlier than the Sun* (Chicago, IL: Yerkes Observatory)

Morgan W W, Keenan P C and Kellman E 1943 *An Atlas of Stellar Spectra* (Chicago, IL: Yerkes Observatory, University of Chicago Press)

Wittgenstein L 1953 *Philosophical Investigations* (New York: Macmillan)

R F Garrison

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clavius, Christoph (1538–1612)

Mathematician, born in Bamberg, Germany. Became a Jesuit in Rome, and wrote a number of textbooks, including a version of Euclid's *Elements*, a commentary on the *Sphere* of Sacrobosco, and books on algebra, the astrolabe, and practical arithmetic and geometry. He participated in the commission for the reform of the calendar that led, in 1582, to the institution of the Gregorian calendar. In his astronomical books, he opposed the Copernican system on both physical and scriptural grounds, but at the end of his life he was faced with telescopic observations by GALILEO, which he was able to repeat, of the phases of Venus and the satellites of Jupiter. He certified what Galileo had seen as real and called on astronomers to consider how the celestial orbs may be arranged in order to save these phenomena but could not bring himself to abandon the Ptolemaic view.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clementine

Joint project between the US Strategic Defense Initiative Organization and NASA, launched in January 1994. Designed to test military sensors and spacecraft components in the space environment. Also intended to make scientific observations of the Moon and near-Earth asteroid 1620 Geographos. Obtained global ultraviolet and infrared imaging, laser ranging altimetry coverage from 60° S to 60° N, and charged particle measurements. Obtained the first data to indicate the presence of water ice inside permanently shaded craters at the lunar poles. Failed before it could reach Geographos.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Climate

Earth's climate may be defined as the global physical condition, averaged over some period of time (typically decades or longer), of the EARTH'S ATMOSPHERE, OCEAN and ice sheets. It is the presence of a relatively dense atmosphere—third among the solid bodies of the solar system—that makes Earth habitable. Without the blanketing of infrared energy radiated from Earth's surface and lower atmosphere, conditions would be well below freezing almost everywhere around the globe. While the atmosphere is the most immediate and palpable part of the Earth's climate system, at least to land-based life, the ocean plays a crucial role. The ocean and atmosphere exchange heat and water vapor (as well as carbon dioxide and other gases). The much greater mass of the ocean compared with the atmosphere provides a damping effect on the climate system, but paradoxically might also drive it from one particular state to another (for example from glacial to interglacial climate). This dual role of the ocean is a consequence of the presence of ocean currents, which in turn depend on the distribution of heat and salt (salinity) with location and depth. Absent our deep oceans, Earth's climatic regime would be drastically different from the present state and past states recorded in the geologic record. Ice sheets also play a role in strongly reflecting the Sun's energy and in storing fresh water that when melted changes the ocean's salinity and global sea levels.

The primary energy source for the climate is the Sun. Rotation of the Earth, seasonal variation of the polar orientation and long-term orbit changes all serve to modulate the amount of solar energy received from place to place and on various timescales. Despite rather dramatic variations in climate through history, a remarkable fact derived from the geologic record is the essential stability of liquid water over large parts of the Earth's surface at least back through 4 billion yr ago. The mass and energy exchanges between the various components of the climate system, and indeed their interactions with the crust of the Earth, must have maintained this essential stability in the face of large progressive changes in the energy output of the Sun implied by stellar evolution theory. The habitability of the Earth is therefore far more complex than simply its favorable position some 150 million km from the Sun.

Worldwide monitoring of the climate has been possible only in the last few decades owing to improvements in transportation, instrumentation and information-transfer technologies. Direct measurements, i.e. *in situ*, represent the first kind of climate data to be collected, and include pressure, temperature, water content (humidity), wind speed, salt content (at oceanic sites), cloud and precipitation conditions. Surface stations grew in number from about 60 in 1875 (which were primarily in the United States and Western Europe) to about 7000 stations reporting daily today. Upper-atmosphere data are more difficult to collect, but by the late 1980s about 800 stations (mostly near airports) launched balloon-borne packages ('rawinsondes' or 'radiosondes') twice daily to measure meteorological parameters at altitudes up to 30 km.

Satellite monitoring began with the launch of TIROS 1, the first satellite designed specifically for weather observations, in 1960. In the ensuing four decades monitoring of cloud and precipitation patterns by orbiting satellites has become an established part of weather and climate studies and has expanded to include remote observations of atmospheric chemistry, atmospheric energy balance, sea state and many other parameters. Much of this information is now publicly available, essentially instantaneously, via the world wide web (interested readers can most easily obtain access through the web site of their particular nation's weather service).

Examination of the climates of the three other large solid bodies in our solar system with substantial atmospheres—Venus, Mars and Titan—reveals very different climate regimes. Venus exhibits a massive CO_2 atmosphere that stands as a stunning example of the effect of greenhouse warming. Mars' thin atmosphere of CO_2 is driven by a combination of processes familiar to terrestrial climatologists but with significant differences in the seasonal cycling of CO_2 into and out of the Martian polar regions, and the occurrence of occasional global dust storms. In addition to the deep atmospheres of the outer planets, the thick atmosphere of Saturn's moon Titan supports a modest greenhouse and a temperature profile with a shape very similar to the Earth's, although hundreds of degrees Celsius colder. By studying the climates of these other solid worlds it is possible to gain some insight into the conditions required in other planetary systems for habitable planets to exist (see also MARS: ATMOSPHERE, VENUS: ATMOSPHERE, TITAN).

Humanity is changing the chemistry and energy balance of the Earth's atmosphere by introducing large amounts of carbon dioxide, chlorofluorocarbons (CFCs) and other products of our industrial–agricultural civilization into the air and water. The CFCs reduce the thickness of the Earth's high-altitude ozone layer, which protects exposed organisms from lethal ultraviolet energy from the Sun. Carbon dioxide and other gases enhance the infrared blanketing effect in the atmosphere, hence warming the planet as well as changing the behavior of the climate system on a timescale of decades. Informed decisions regarding controls on emission of greenhouse and ozone-destroying gases require a deep understanding of how the climate system works, an understanding that is being gleaned from new generations of surface, airborne and spaceborne instrumentation (see also GREENHOUSE EFFECT, OZONE HOLE).

Components of the Earth's climate system

Atmosphere

The atmosphere is what we most associated with climate, and it plays host to the short-term and spatially localized variability termed weather. Earth's atmosphere is mostly molecular nitrogen (N_2), constituting 78% of the air

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by number. Approximately 21% of the atmosphere is molecular oxygen (O_2), a product of plant and microbial photosynthesis that reacts with common minerals and would disappear in less than 10 million years in the absence of these processes. Water vapor, because it condenses to form clouds, and is connected through condensation and evaporation to the Earth's oceans and land, is highly variable from place to place and with altitude (see WATER CYCLE). On a global average its mixing ratio near the base of the atmosphere is 1%. Carbon dioxide (CO_2) is the primary supply of carbon to the plant and bacterial photosynthetic cycle and currently constitutes 330 parts per million (0.000 33) of the atmosphere. The total mass of Earth's atmosphere is estimated to be 5×10^{18} kg, determined through the atmospheric density–altitude structure.

The atmosphere can actually be thought of as concentric layers of gas each of which is distinguished by particular mechanisms of energy transport and hence temperature profile. The *troposphere* is characterized by falling temperature with altitude (except near the surface, where locally temperature can rise with altitude in a thin *inversion* layer). In the troposphere, convection and radiation carry heat upward from the surface which is warmed by visible solar radiation. It is here that nearly all clouds form and weather occurs, land–sea–air exchange of gases takes place and the biosphere exists. The troposphere ends at the *tropopause*, about 8 km above in the surface in the polar regions and 13 km in the tropics (with substantial seasonal variations as well). Here the temperature profile reverses direction, and temperature begins to increase with altitude in the *stratosphere*. The stratosphere is warmed by absorption of solar ultraviolet and infrared radiation, largely by ozone (O_3) as part of its photochemical cycle of creation and destruction. As ozone declines, cooling by emission of radiation to space by CO_2 leads to a halt in the rising temperature at the *stratopause* roughly 50 km above the surface; above the temperature falls with height in the *mesosphere*. Beyond 100 km altitude photolysis (breaking of a molecule by absorption of photons) of O_2 and ionization increase the temperature steeply with altitude in the *thermosphere*, where conduction replaces radiation as the key energy transport process. In the ensuing discussion of climate only the stratosphere and troposphere are considered, within which lies over 90% of the total atmospheric mass.

The Earth's rotation and differential solar heating as a function of latitude and time of year drive atmospheric circulation patterns. Solar heating is greatest at the equator and least at the poles; hence the atmosphere must transport heat from equator to pole. It does so through the so-called Hadley circulation, in which warm air rises and is transported to higher latitudes, cools and sinks. The Hadley circulation does not consist of a single cell in the case of the Earth; instead three cells in each hemisphere move heat poleward. As seen in figure 1 rising air at the cell boundaries diverges, leading to belts of low-pressure air, whereas sinking air converges at cell boundaries yielding zones of higher pressure. Since the Earth rotates this pattern of high and low belts is accompanied by the prevailing *zonal* wind pattern also shown in figure 1. The *westerlies* increase in speed with altitude, reaching a maximum just below the stratosphere as the narrow *jet streams*, which are responsible for the movement of mid-latitude storm systems. The *trade winds*, blowing from the east, are a primary driver of the oceanic Gulf Stream through frictional stresses on the sea surface.

Superimposed on the primary circulation pattern described above is a series of secondary high- and low-pressure systems associated with perturbations in the basic flow. The Earth's rotation creates an effective *Coriolis* force that acts at right-angles to the direction of the wind: to the right in the northern hemisphere and left in the southern. At latitudes higher than 10°, and altitudes above the frictional boundary layer of the atmosphere, the Coriolis force is sufficiently strong that it nearly fully balances the pressure gradient force between high- and low-pressure centers. This *geostrophic* balance causes areas of high pressure to spin in a clockwise direction and low-pressure areas to spin in a counterclockwise direction, in the northern hemisphere—and in the opposite sense in the southern. Low-pressure areas are thus *cyclonic* (in the northern hemisphere) regions where rising air forms clouds and precipitation systems if enough moisture is available. Anticyclonic high-pressure areas are characterized by subsidence, evaporation and hence drying; they are generally fair-weather features (again the direction of circulation is reversed in the southern hemisphere). Some high- and low-pressure systems are persistent features, either seasonally or annually, but may fluctuate in strength and position from year to year. Others are spatially and temporally variable, such as mid-latitude storm systems and hurricanes.

Mid-latitude and tropical cyclonic systems exist because of perturbations in the flow that grow with time. *Baroclinic instabilities* are triggered by the existence of a meridional (north–south) temperature gradient in a geostrophic atmosphere and hence are responsible for storm systems at mid-latitudes (where the Coriolis effect is strong). *Barotropic instabilities*, on the other hand, operate at equatorial latitudes, and are a class of instability in which only kinetic (not thermal) energy is transferred from the background rotation of the atmosphere to the growing perturbation. Barotropic instabilities are thought to produce tropical depressions, which in some cases become tropical storms or hurricanes, in the equatorial belt of low pressure called the *intertropical convergence zone* (ITCZ).

In contrast to the geostrophic balance of the terrestrial and Martian atmospheres, on Venus and Saturn's moon Titan stratospheric winds are more rapid than the planetary rotation. The resulting strong centrifugal force dominates over the Coriolis force, leading to so-called *cyclostrophic* balance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

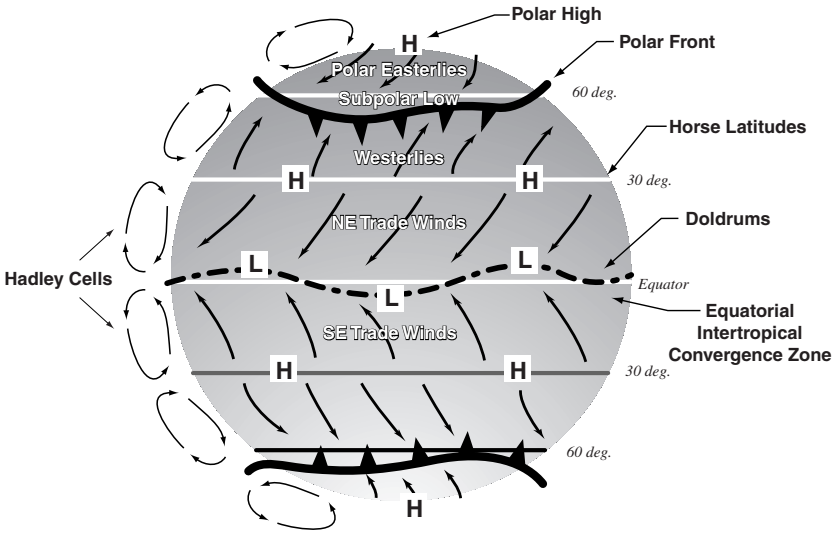


Figure 1. Persistent features of the Earth's atmospheric circulation. Shown is the pattern of permanent high- and low-pressure belts and the Hadley cells that move heat from equator to pole in each hemisphere. The doldrums are the equatorial regime of very calm air; horse latitudes refer to the narrow belts of high pressure over the oceans at about 30–35° latitude. Labeled also is the pattern of atmospheric winds corresponding to the pressure belts; the mid-latitude and higher winds in the southern hemisphere mirror those in the north and are not indicated.

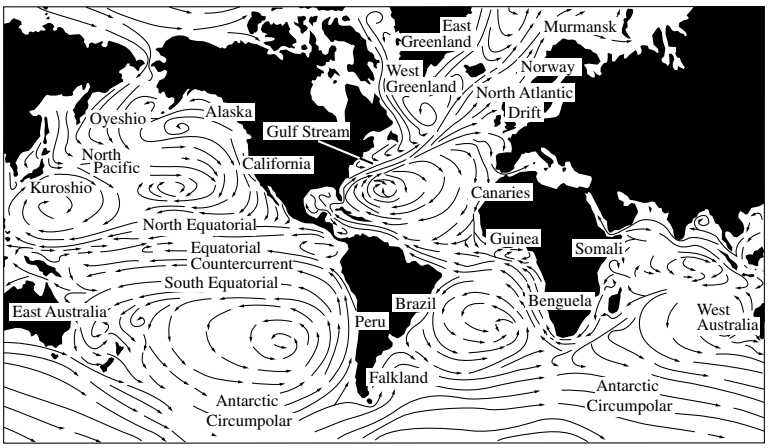


Figure 2. Ocean circulation patterns. Primary surface currents during wintertime are shown, after Peixoto and Oort (1992).

Hydrosphere

The Earth is unique among the solar system's planets in having a massive layer of liquid water, our ocean. In addition to oceanic water, fresh water systems are distributed over the Earth's surface in the form of rivers and lakes. There is also a substantial amount of water circulating through the upper crustal layer of the Earth's interior. This water mixes with the deep ocean on a wide range of timescales, from very short to geological (i.e. millions of years). The liquid water system as a whole is referred to as the *hydrosphere*, and has a mass estimated at 1.7×10^{21} kg, the vast majority of which is in the ocean. Thus the amount of water in contact with the atmosphere on timescales much shorter than geological amounts to 200 times the mass of the Earth's atmosphere, and strongly modifies the response of the climate to changes in external conditions, as well as creating forcing functions of its own. The oceans effectively store large amounts of thermal energy (originally sunlight) that is later released as sensible or latent heat. The temperature of the ocean has a direct effect on the amount of water vapor, the principal greenhouse gas, released into the atmosphere; the oceans also store large amounts of CO_2 (the other primary greenhouse gas) in a temperature-dependent fashion. The average depth of the ocean is 4 km; some of the deepest portions mix with the upper layers or the atmosphere only occasionally, while the uppermost part of the ocean exchanges gases and heat with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the atmosphere very rapidly and continuously. The ocean moves heat from equator to pole in amounts comparable with what is transported by the atmosphere.

In contrast to the atmosphere there is a large-scale day-to-day uniformity of temperature near the surface of the ocean and a slow seasonal shift. Like a large indoor swimming pool the ocean damps out temperature variations in the atmosphere and controls the lower atmospheric humidity (water vapor content) over large areas of the Earth. The ocean exhibits circulation patterns that bring warm currents of water to high latitudes and cold water to equatorial regions (figure 2). These reduce the north–south gradient of temperature in the atmosphere. The Atlantic Gulf Stream is the best known example of the former. The southward Pacific current off the California coast is a familiar example of the latter, making summer ocean water very cold as far south as Los Angeles. Surface currents are driven by wind stresses, and convergence or divergence of flows leads to vertical motions (downwelling or upwelling) below the surface. This is not, however, the whole story. The oceans contain about 3 parts per thousand dissolved salts, which both lower the freezing temperature of the water and change its density. The former effect allows annual mean temperatures at high latitudes to be nearly 2 °C below freezing. The density effect is such as to make waters of increasing salinity denser.

The deep circulation patterns in the ocean are poorly known, because of a dearth of observations. Thermal and saline (*thermohaline*) density effects, compressibility of the ocean water (actually quite important at large depths) and mass conservation associated with return flow of water from surface currents drive overall ocean circulation. Because of compressibility and thermal effects, very cold, deep, ocean water may not exchange with the near surface more frequently than hundreds of years. The positions of the continents play a key role in the circulation, and in particular may affect the efficiency of north–south heat transport. The role of wind in driving ocean currents and the effect of currents in redistributing heat available to the atmosphere results in a complex, coupled ocean–atmosphere climate system.

The premier example (and possibly the most important in determining global climate state; see below) is the Gulf Stream to North Pacific conveyor belt. The Gulf Stream current is powered by the stress of trade winds exerted on the ocean surface in the tropics, and of the westerlies at mid-latitudes. The current moves warm waters northeastward, along the United States east coast. Evaporation increases the current's salinity; exchange of heat with the atmosphere along the route leads to a net cooling of the water and warming of the atmosphere. The resulting progressive, northward densification reaches a critical stage near Iceland, where the Gulf water becomes denser than the surroundings and sinks. The overturn is aided by winds which blow colder surface waters aside, allowing deeper parts of the current to rise, release heat, and cool rapidly. The now-bottom water flows southward, around Africa and then northward to the colder North Pacific. There it rises to the surface and the divergence pushes ambient waters aside, some moving southward toward the start of the Gulf Stream. The most important consequence of this circulation is to transport significant amount of equatorial heat to the North Atlantic region, strongly moderating the climate of Europe and hence preventing the buildup of massive ice sheets.

Cryosphere

The cryosphere comprises the Earth's permanent and seasonal ice sheets and snow, as well as permafrost, which together account for 80% of the planet's fresh water supply. The major components are the ice sheets of Antarctica, Greenland, permafrost in North America and Asia, sea ice, mountain glacier systems and seasonal snowfields. Although only about 2% of the Earth's water is frozen, the cryosphere has a profound effect on present-day climate and on the glacial–interglacial oscillations of the past 2 million years. The very high reflectivity of water ice (which is sensitive to particle size and hence formation conditions and age, as well as to impurities) rejects by reflection solar radiation in the visible and near-UV parts of the spectrum. In the thermal infrared, water ice is quite dark, and hence acts as a very efficient nighttime radiator of terrestrial thermal radiation. The large latent heat of melting of water ice also acts as a strong heat sink, particularly during colder epochs in which more massive glaciation exists than at present.

The cryosphere amplifies the effect on climate of the northern–southern hemisphere dichotomy in distribution of continents and oceans. Because the northern hemisphere has far more land area than does the southern, development of massive continental glaciation to mid-latitudes, a feature of glacial epochs, is predominantly a northern hemisphere effect. This in turn may tie the timing of onset of glaciations to configurations of the Earth's orbital and spin parameters that produce minima in annually averaged northern hemisphere insolation. On much shorter (seasonal) timescales, however, the presence of a large polar continent in the south and polar ocean in the north affects the relative responses of the poles to the decreased wintertime insolation.

Greenhouse warming

The suitability of Earth for life is in large measure a result of the stability of liquid water at the surface; the annually averaged global mean ocean surface temperature is 288 K. Yet simply equating the annually averaged amount of solar radiation absorbed by Earth's surface to the outgoing bolometric thermal radiation produces an effective temperature of 255 K, well below the freezing point of water (even with salts). The high surface temperature of the Earth, and hence the stability of liquid water, is a consequence of the higher mean opacity of the atmosphere in infrared (where thermal radiation is released toward space) than in optical wavelengths (where the peak of the solar spectrum occurs). The term

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

greenhouse effect is indelibly affixed to this basic property of the terrestrial atmosphere, despite that fact that in a greenhouse thermally driven convection plays essentially no role at all, while it is important in regulating the lower temperature of the Earth's atmosphere.

The wavelength distribution of photons streaming from the Sun, approximated as a blackbody at 6000 K, peaks at about 0.6 μm wavelength (see BLACK-BODY RADIATION), where atmospheric gases have almost no absorption (with the exception of the so-called ozone *Chappuis* bands). Thus in a cloudless atmosphere sunlight would be virtually free to reach the surface (although Rayleigh scattering by molecules of the air preferentially diffuses shorter wavelengths out of the solar beam, leading to the familiar blue sky). Clouds reflect about 20% of the incoming sunlight, while the surface reflects less than 10% on average. About 2/3 of the sunlight incident on Earth is absorbed by the ground or lower atmosphere and converted to photons with a spectral distribution that, if approximated by a blackbody, corresponds to a temperature of 255 K (absent the greenhouse effect; 288 K with the greenhouse warming). The peak of the spectral distribution of these thermalized photons is in the mid-infrared part of the spectrum, at about 15–20 μm wavelength, and the atmosphere is very opaque there. Absorption bands of water vapor and CO_2 dominate, with holes in the absorption spectrum filled in by CH_4, O_2, O_3, nitrous oxides and even man-made CFC's.

At ground level the atmosphere is highly opaque beyond 5 μm wavelength, but the absorptions decrease with altitude as atmosphere column abundance (from a given level to space) and pressure decline. By 10 km the atmosphere is largely transparent again. Thermal photons streaming upward from the ground have mean free paths much shorter than the inbound solar radiation whose energy flux they must balance. Absorption and reradiation of the thermal photons redirect a portion of the flux back downward; the downward-directed flux decreases with altitude as the atmospheric transparency increases. As a result of this inefficiency in outward thermal transport relative to inbound solar visible radiation, the surface temperature must rise above the effective temperature to ensure a balance between ingoing and outgoing energy flux. Altitudes above the surface also must experience elevated temperatures, but progressively less so as one moves outward. Thus the temperature increases between the altitude at which infrared photons are free to stream outward, the *mean radiating level*, and the surface.

The steepness of the temperature increase is modified from the greenhouse value by a number of other energy transport processes. Most important is convection. When the temperature gradient steepens to exceed the idealized, adiabatic lapse rate, the air becomes unstable and bulk vertical motions (convection) transport the heat upward at the adiabatic lapse rate. In the absence of water the relevant adiabatic lapse rate is the so-called *dry* value of 10 K km^{-1}. Water condensation during cloud formation releases latent heat which allows for *moist* (sometimes, *wet*) convective transport of heat at temperature gradients well below the dry adiabat. To a first approximation the observed mean temperature gradient in the Earth's atmosphere is the moist adiabatic lapse rate. Further modification of the temperature profile occurs as winds advectively transport heat from one location to another, and storm systems move both moisture and heat in horizontal and vertical directions. Figure 3 summarizes the various mechanisms for upward and downward energy transport.

The essential point about the greenhouse effect should not be obscured by the complexity of multiple transport processes. Inbound energy flux must be balanced by outgoing, and if the atmosphere is less transparent to outgoing (thermal) radiation than to incoming (solar) radiation, the surface temperature is elevated until the balance is achieved. However, the response of the Earth's climate to the increase in greenhouse gases is difficult to predict because of the large number of degrees of freedom. Modifications to cloud formation, precipitation, surface ice coverage, ocean circulation, temperature and CO_2 solubility all represent responses to increased greenhouse gases that can provide positive or negative feedbacks. The effect of an incremental increase in greenhouse gases can be predicted (incrementally higher surface temperature), but discrete changes (particularly large ones) can simply push the Earth's climate into a different state.

An extreme example of the greenhouse effect is available for study on Venus, where the surface pressure is 92 bar and the atmosphere is 96% carbon dioxide. Because a layer of reflective sulfuric acid and water clouds covers the globe, the surface of Venus receives only a few per cent of the sunlight incident on the top of the atmosphere—less than the Earth receives at its surface despite the latter being 40% more distant from the Sun. Yet the surface temperature of Venus is 730 K, a value that reflects the extreme infrared blanketing afforded by the high CO_2 surface pressure, as well as contributions from minor species such as water vapor (roughly 100 parts per million of the atmosphere). As with the Earth outward energy transport is largely by convection, but on Venus it is entirely dry convection.

Mars' thin atmosphere has a mean surface pressure of only 7.5 mbar and is almost devoid of water. Therefore even though it is 95% carbon dioxide (which yields a CO_2 partial pressure larger than on Earth), absorption bands of carbon dioxide are very narrow in frequency, lacking the broadening afforded by a high pressure of nitrogen; in addition water absorption bands are almost nonexistent. Thus there is only a very small greenhouse effect of 5 K. Solar heating of the surface drives convection and hence produces a troposphere extending to 40 km. However, the lapse rate is much less than dry adiabatic because of a suspension of dust particles in the atmosphere, which absorb sunlight at all altitudes through the troposphere. In some years global dust storms load the atmosphere with so much dust that the ambient atmospheric temperature profile is strongly perturbed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

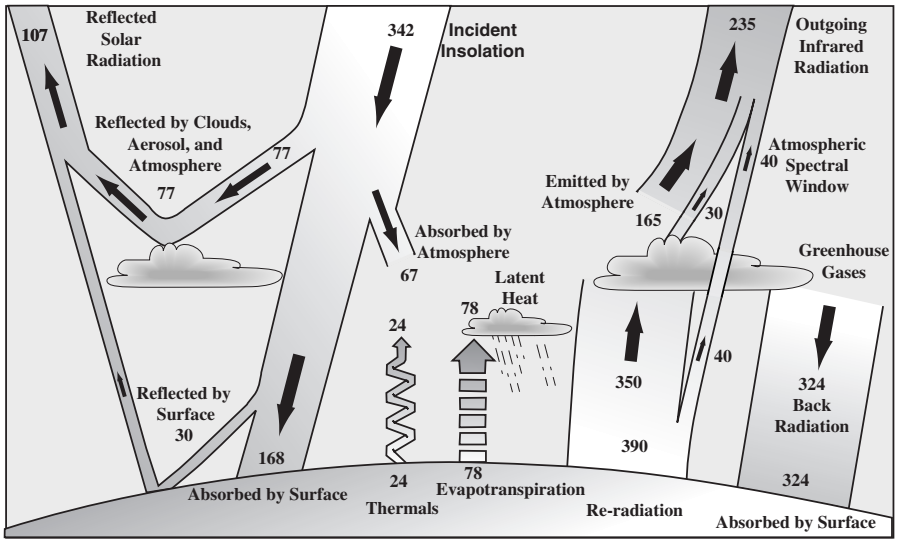


Figure 3. Contributions to sources and sinks of energy in the Earth's atmosphere. Left half of the figure is fate of sunlight (mostly optical), right half the history of reradiated thermal energy, and non-radiative processes are sketched in the center. Energy fluxes in watts per square meter. Redrawn from Trenberth *et al* (1996).

General circulation models: ocean and atmosphere

The fundamental theoretical tool both for weather forecasting and for climate change studies is the numerical *general circulation model*, or GCM. The Earth is divided into a grid with cell sizes as small as practical for the available computational resources. In each grid the following equations are applied: (a) horizontal equations of motion; (b) hydrostatic law; (c) continuity equations; (d) energy balance (first law of thermodynamics); (e) equation of state for moist and dry air; (f) balance equation for water vapor. These form a closed system with boundary conditions specified by meteorological measurements or (for paleoclimate models) geological data and assumptions. With high-speed computers atmospheric motions can be coupled to those of the oceans, leading to ocean–atmosphere GCMs. For the ocean component similar equations to (a)–(f) apply, except that (e) is replaced by the equation of state for ocean water and (f) is replaced by a balance equation for salinity. Over vegetated areas vapor and energy transfer from and to plants by evapotranspiration must also be included. Equations are integrated to track the time evolution, in three dimensions, of atmospheric circulation, intensity and position of high- and low-pressure centers, temperature, heat and mass transfer between atmosphere and ocean, and other parameters.

Because of the vast difference in timescales between atmosphere and ocean some simplification of the ocean model (for example a shallow ocean representing the uppermost well-mixed layer), or averaging of the atmospheric processes, must be made. Increasingly full-up simulations coupling the two are being accomplished as computational power increases. However, significant limitations remain. In the atmospheric case, processes occurring on scales below the grid size can have significant effect. For example, orographic lifting of air masses by localized or topographically complex mountains, small-scale variations in land use patterns or intricate geographical relationships between dry land and bays or seas often cannot be handled well by GCMs. The formation of precipitation is highly sensitive to subscale processes, and hence GCMs are less successful at tracking this process in detail. As we shall see below, variations in precipitation patterns over land and sea may have served as the trigger for profound flips in climate state, and it possible that GCM simulations, particularly if time averaged to track decadal or longer changes, may miss these.

Sufficient observations of the Martian atmosphere have been made to enable application of GCMs of considerable sophistication. The total absence of a global ocean on Mars means that there is no large-scale heat sink to dampen climate variations on seasonal or longer timescales and no source of water vapor to modify atmospheric lapse rates and serve as a source of massive cloud formations. Instead the Martian climate is driven mostly by seasonal cycles of CO_2 and dust. The polar regions become so cold during the winter (150 K) that the primary atmospheric constituent, CO_2, condenses out to form an extended cap in the winter hemisphere. The annual cycling of CO_2 from one winter hemisphere to the other modulates the bulk atmospheric pressure by 30%. This phenomenon, with no analogue on the Earth or Venus, has been measured directly by the VIKING and MARS PATHFINDER landers. Among other effects, it drives substantial variation in the nature of the equatorial to mid-latitude Hadley circulation, which comprises one Hadley

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cell at the solstices and two at the equinoxes. At higher latitudes in winter and spring, traveling disturbances exist that are akin to terrestrial storm systems; these are absent in summer. Such systems are detected both by lander meteorological stations and by very tenuous traces of water clouds seen in global imagery.

Climate events: ENSO and the Asiatic monsoon

The El Niño–Southern Oscillation is a climatic phenomenon whose existence and worldwide connections have been recognized only in the past few decades. It provides an excellent example of atmosphere–ocean interactions and the ability of regional ocean-driven perturbations to cause worldwide changes in weather patterns. On intervals of roughly 2–10 yr an unusually strong ocean current pushes warm, low-salinity surface waters against the west coast of South America, where it then moves southward. This phenomenon is referred to as *El Niño*, a co-option of a traditional term used by Peruvian fisherman to describe the annual warming of offshore waters right after Christmas.

Accompanying the warming of eastern Pacific waters is an associated rise of air temperatures. Strong El Niños may persist for more than a year, and perturbations to Pacific ocean currents are generated over many thousands of miles to the west of South America. The atmospheric disturbance accompanying the oceanic changes is even more widespread. During El Niño years, drought conditions occur in Australia, Indonesia, southeastern Africa and northeastern South America, while unusually heavy rains may occur in Ecuador, northern Peru, eastern equatorial Africa and areas south of India. The alteration in weather patterns had traditionally been considered a separate phenomenon involving shifts in positions of seasonal high- and low-pressure patterns, known as the *Southern Oscillation*. It was not until the 1960s that the causal link between it and El Niño was established.

The mechanics of the oceanic perturbation involve waves generated in the Pacific which push warm water toward the east, raising sea level there slightly and depleting the amount of warm water in the western Pacific. The onset and dissipation of this event involve a complex, poorly understood series of interactions between the changed sea surface temperature distribution and the atmospheric temperature pattern. What is not understood at all is the origin of the ENSO phenomenon in its entirety. Proposals range from the idea that the ocean circulation pattern itself drives the oscillation, all the way to the unlikely notion that undersea eruptions contribute heat to the deep ocean and stimulate the effect. Also poorly understood is whether ocean–atmosphere oscillations other than in the equatorial Pacific might exist. Finally, the speculation has been made that ENSO is symptomatic of an increasingly warm climate on the edge of instability, and that cooler times during the Holocene did not experience such oscillations.

The substantial summer rains in the Indian subcontinent and surrounding parts of Asia are a part of the seasonal shift of wind pattern called the Asian monsoon. The great mass of the Tibetan plateau, the large land area of the Asian continent with its concomitant extreme seasonal temperature contrasts and the location of the ITCZ all play a role in defining the monsoon climate regime. Most dramatic is the seasonal shift in winds in the spring, during which mid-level winds change from westerly to easterly, and extremely hot and dry conditions in late spring give way to moist onshore flow as summer commences. A complex continental airflow is initiated leading to large amounts of rainfall in places such as India. The amount of rain delivered both locally and over the whole region varies dramatically from one year to the next and appears to be correlated with the strength of El Niño. Since India and some other Asian nations depend on the monsoon for significant parts of their agricultural water supply, the predictability of monsoon variability and its connection to worldwide phenomena such as ENSO represent climate issues of great human importance.

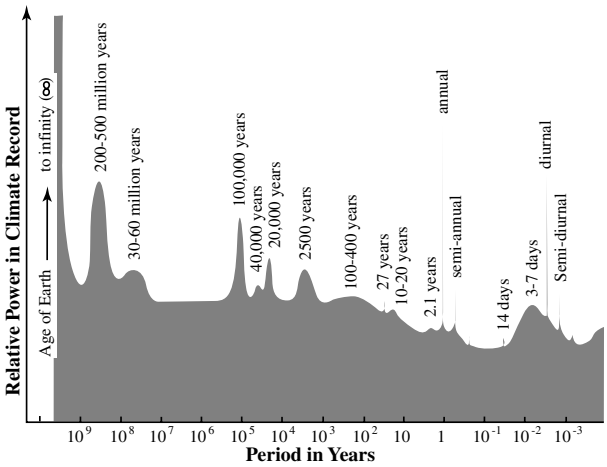


Figure 4. Temporal climate scales. Variance as a function of timescale in climate based on direct observations and proxy indicators from the climate record. After Peixoto and Oort (1992).

Climate change in the past

Information on climate variations comes, on the longest timescales (billions to hundreds of millions of years), from oxygen isotopes in cherts, and geologic evidence including geographical distribution of fossil species and signatures of glaciation. On intermediate timescales (millions to tens of millions of years), isotopic abundances in seafloor sediments (carbon and hydrogen) provide a more detailed climate record. Finally, on shorter timescales a variety of climate proxies are available, from isotopic data in seafloor sediments throughout the Pleistocene, to ice core data (isotopic temperature proxies, carbon dioxide content, etc) and on to tree-ring cores, streamflow data and historical records for the current Holocene stage.

Figure 4 is a power spectrum of atmospheric temperature variations on timescales ranging from geologic to diurnal. The origin of the various peaks, from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the longest to shortest timescales, is as follows. Isotopic data suggest an overall decline in surface temperature from the time of the earliest sediments, possibly due to declining levels of CO_2 in the Earth's atmosphere. The abundance of this gas is thought to be regulated by complicated interactions between weathering of silicates (leading to trapping of CO_2 as carbonates on the seafloor) and volcanic extrusion of CO_2 associated largely with plate tectonic recycling of oceanic crust and sediments. Geomorphological and isotopic data hint at one and maybe two periods of deep, nearly global ice ages prior to 500 million years ago, as well as (with fossil data) climate variation on hundred million year timescales. Such variations probably reflect the changing positions of the continents, which affect oceanic circulation and hence transport of heat, along with associated changes in seafloor spreading rates, volcanism and mountain building activity, all of which affect the atmospheric CO_2 abundance. The weak variation seen on 10 million year timespans might be volcanically or tectonically driven, but could reflect as well stochastic effects such as oscillations in solar luminosity, and even passage of the solar system in and out of galactic molecular clouds.

The strong signals of climate variation on 10 000–100 000 yr timescales are almost certainly a result of variations in the Earth's axial tilt, orbital eccentricity and phasing of various orbital parameters. These *Milankovich cycles* are a natural consequence of the gravitational effect of the Moon and other planets on Earth, and lead to complex cyclical modulations of the seasonal and annually averaged absorption of solar radiation by the Earth. The variance is strong because the modulations led to oscillations of the Earth, over the last 2 million years, between states of significant northern hemisphere glaciation with duration of order 100 000 yr and warm interglacial periods typically 1/10 the glacial duration. The weak signature of climate fluctuations on the thousand-year scale is related to the mid-Holocene *climate optimum* and subsequent cooling. On historical timescales, the peculiar Little Ice Age which began several centuries ago and ended in the 19th century can be seen as a strong variance, as well as ENSO-type oscillations on decadal and semi-annual scales. Earth's orbital motion, axial tilt and spin (in their current relative orientations) provide strong fluctuations on annual, seasonal and diurnal timescales.

Two particular well-studied examples of climate conditions in the past have relevance to our understanding of human-induced climate change. Isotopic data and geological evidence in the form of fossil distributions and geomorphology argue that the Earth during the time of the Cretaceous (65–100 million years ago) was free of permanent ice sheets. To explain the data the equator-to-pole temperature difference must have been 1/2–1/4 what it is at present, with an equator not much warmer than at present but poles 20–40 °C warmer. Simply doubling or quadrupling the atmospheric CO_2 level does not explain Cretaceous climate, because it raises equatorial temperatures too much. Instead, enhanced oceanic transport of heat from equator to pole is invoked in GCMs that fit the geologic and isotopic data. The origin of such enhanced oceanic heat transport is not well understood; possibly the difference in continental configuration during the Cretaceous could have played a role.

The Younger Dryas episode some 11 000 yr ago, during the transition from cold glacial to warmer Holocene conditions, is especially noteworthy. It was a cold snap in continental areas around the North Atlantic that lasted no more than a century, during which glaciers reversed their retreat. The origin of the Younger Dryas episode paradoxically lay in the melting of the glaciers themselves. Glacial meltwater accumulated in parts of eastern Canada and flowed down the Mississippi with volume rivaling the present Amazon river. Glaciers blocking this massive flow of freshwater from entering the North Atlantic retreated sufficiently by about 11 000 yr ago that the meltwater began flowing eastward. The supposition is that this influx of freshwater to the North Atlantic diluted the saline Gulf Stream sufficiently to prevent its sinking near Iceland, and hence reduced or shut off the associated release of heat. The Gulf circulation stopped, replaced by a stable upper layer of very cold water that chilled Northern Europe and Canada and reinitiated the buildup of glaciers through expanded annual snowfalls and reduced snowmelt. As the freshwater flow subsided the Gulf current returned and warming began again. Very recent ice core data reveal many other rapid episodes of climate fluctuation in glacial and interglacial epochs of the Pleistocene, some on the scale of decades or less.

The above interpretation of the Younger Dryas episode has led to speculation that a similar reversal triggered in the North Atlantic could be in store if human activities accelerate global warming (see below). GCMs predict enhanced rainfall at high latitudes as CO_2 levels increase in the coming century. Sufficient rainfall could decrease the salinity of North Atlantic waters to such an extent that the subsidence of water off Iceland is interrupted as in the Younger Dryas. The resulting shift in climate is hard to predict using GCMs, but conceivably could be in the direction of increased glaciation with positive feedback afforded by the high reflectivity of water ice. The specifics of this putative response to global warming is less important than the general lesson of the Younger Dryas that perturbations associated with climate change in one direction can cause the climate system to flip into a mode that evolves in a quite different direction.

Human induced global warming

Beginning in the mid-19th century CO_2 levels in the atmosphere began to increase as a result of human activities, including burning of fossil fuels and deforestation. The present CO_2 abundance is 30% higher than the base level before the increase. In addition, human activities are enhancing the abundances of CH_4 and CFCs in the atmosphere which add to the infrared absorption along with CO_2. Increasing the greenhouse gas content of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the atmosphere forces the mean radiating level to move upward; in response the mean surface temperature must increase to balance incoming solar flux. However, the complexity of the various effects associated with such a change, including cloud and precipitation shifts as well as the role of the ocean in storing and releasing heat and carbon dioxide require use of a GCM to predict in detail the resulting climate changes.

Direct temperature records over the past hundred years as well as proxy temperature indicators available throughout the Holocene suggest that the last decade was exceedingly warm, perhaps the warmest of the past 1000 yr. Much political debate surrounds the interpretation of the data (natural fluctuation or human-induced global warming?) and the utility of GCMs in predicting local or regional impacts. The climate may respond linearly to increasing greenhouse gas content, at least for a while. However, the Younger Dryas and other evidence of rapid climate shifts in the geologic record serve as a warning that human activities, in perturbing the composition of the atmosphere, will force us to deal with substantial and probably undesirable rapid shifts in climate.

The future of climate studies

The future of climate studies will depend, as over the last half-century, on cooperative international programs of monitoring climate and evaluating climate data. Special programs, often time limited and sometimes motivated by political as well as scientific imperatives, benefit the field of climate studies through additional funding and attracting talented students to the field. The International Polar Years in 1882–3 and 1932–3 were early examples focused on data collection in the arctic. The International Geophysical Year from 1957 to 1958 demonstrated the feasibility of much broader-scale, if not global, coordination of data collection on a range of Earth studies including climate. More recently the World Meteorological Organization and the United Nations Environmental Program established the International Panel on Climate Change, which has produced several major published assessments of the state and near-term future of the Earth's climate.

Spacefaring nations coordinate satellite observations and, in the case of the United States and Europe, are in the process of implementing major space-based systems to monitor Earth's climate with a level of accuracy and breadth of coverage far exceeding previous efforts. These satellite systems represent current investments of billions of dollars and euros and will grow in the future. Impressive increases in computational capability will benefit study of climate in two ways, allowing more elaborate and higher-fidelity climate models and enabling collection and processing of very large data sets on Earth's climate. New computational approaches to processing the information and the availability of the world wide web make climate data more understandable and accessible to researchers and policymakers alike. The gradual accumulation of spacecraft data on Mars, Venus and Titan has permitted GCMs to be applied to these other three solid planetary bodies with substantial atmospheres. Such efforts test our understanding of the physics of climate in regimes very different from that of Earth itself. As our experience with the Earth's climate grows in detail alongside the exploration of other planetary climates our understanding of the processes involved will deepen dramatically.

Bibliography

Broecker W S 1997 Thermohaline circulation, the Achilles heel of our climate system: will man-made CO_2 upset the current balance? *Science* **278** 1588–92

Houghton J T 1986 *The Physics of Atmospheres* (Cambridge: Cambridge University Press)

Mitchell J F B 1989 The 'greenhouse' effect and climate change *Rev. Geophys.* **27** 115–39

Peixoto J P and Oort A H 1992 *Physics of Climate* (New York: AIP Press)

Taylor K 1999 Rapid climate change *Am. Sci.* **87** 320–7

Trenberth K E, Houghton J T and Meira Fliho L G 1996 The climate system: an overview *Climate Change 1995: the Science of Climate Change* ed J T Houghton, L G Meira Filho, B A Callender, N Harris, A Kattenberg and K Maskell (Cambridge: Cambridge University Press) pp 51–65

Jonathan I Lunine

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Close Binary Stars

Often a star is found to have a companion star which is very close to it. Often also the two stars would fit comfortably inside our solar system, whose radius is about a ten-thousandth of the distance to the next nearest star. Sometimes the two stars are so close together as to be virtually touching. In a few cases they are genuinely touching, and even overlapping, so that we may have a single figure-of-eight-shaped entity (a 'contact' binary) consisting of two stellar cores within a single thin envelope.

Most close binaries (CBs) are discovered by one of two techniques: (i) they may show a periodic Doppler shift due to motion in a Keplerian orbit (SPECTROSCOPIC BINARY STARS, or SBs) or (ii) they may show periodic eclipses as one star passes in front of the other (ECLIPSING BINARIES, or EBs). SBs and EBs have periods ranging from ∼30 yr down to a few days, and even a few minutes. Specifically, we set a period of 30 yr as an upper limit for 'close'. This implies separations, by Kepler's third law, of less than the orbit of Saturn, approximately.

There are other ways of recognizing some classes of CB. For instance, eruptions of the NOVA and dwarf nova category are indicative of binarity, and so (often, but not always) is the production of x-rays. However, the binarity of such objects is only known because a very high proportion of them are found on close inspection to be either SBs or EBs (or both). The statistical distribution of periods in CBs is not yet well determined. Although several thousand SBs and EBs are known, their detection is subject to many selection effects. For example, short periods are usually easier to recognize than long periods. The most systematic studies tend to show that longer periods, say 3–30 yr, are more common than short, say 3–30 days, but only by a modest factor, perhaps a factor of 5. Among stars similar to the Sun, probably about 20% are CBs as defined above, while among massive stars of spectral type O the proportion must be much higher because about 25% have periods of up to only about 30 days. Longer-period spectroscopic orbits among O-type stars are hard to recognize, because such hot and massive stars tend to have an intrinsic spectral 'jitter' which can conceal the orbital motion.

If two solar-type stars are in orbit about each other (strictly, about their center of gravity) with a period of about 10 yr, they would be about as far apart as Jupiter from the Sun. We might not think of this as close, because their separation would be almost a thousand times as large as their radii. However, as stars evolve they become larger, ultimately by a factor of several hundred in radius and perhaps even a thousand (see STELLAR EVOLUTION). Clearly when one star is almost as large as the orbit in which it finds itself, something rather drastic must happen. There is much observational evidence for this, some direct and some indirect. Directly, we observe a number of binaries in which a stream of gas is flowing from one star to the other. Indirectly, we see a number of CBs containing compact stellar remnants: WHITE DWARFS (WDs), NEUTRON STARS (NSs) and even BLACK HOLES (BHs). These remnants may appear fairly far apart in their present binaries but must be descendants of stars which were once much larger. They may be in binaries that are too close at present ever to have contained normal stars, so that some interaction must have taken place in the past.

CBs are of interest for at least three very different reasons. Firstly, careful measurement of the orbit, especially in systems which are both SB and EB, can lead to determination of the masses and radii of stars. This is in fact the only reliable way of determining such fundamental parameters. Secondly, they show a wide range of physical processes that are not to be seen in single stars, for instance gas streams that may be torn off one component and accreted by the other. Thirdly, they allow us the opportunity to test conclusions about the evolution of stars. This article will concentrate mainly on the third aspect.

A star of given initial mass begins the major part of its life at a modest radius, comparable with the Sun's, when nuclear fusion (the burning of hydrogen to helium) becomes established in its central core. It is a 'main sequence' (MS) star, and remains so, expanding slowly by a factor of about 2.5 in radius until it has exhausted the hydrogen fuel in its center. Then it evolves more rapidly to a large radius, high luminosity and low surface temperature: it becomes a RED GIANT (RG). The central energy-producing core by contrast becomes more compact. The lifetime as an RG is typically ∼10–30% of the lifetime in the MS band. Finally either (a) it reaches such a high luminosity that it is able to blow off most of its outer layers through a strong stellar wind or (b) in more massive stars the core becomes so compact that gravity causes it to implode, and the sudden release of gravitational energy in the core blows the envelope off in a supernova explosion. The remnant in either case is a compact inert object, a WD in case (a) and an NS or a BH in case (b). The lifetime of this remnant is extremely long, but at the same time the object is very faint and therefore inconspicuous.

It appears to be a rather fundamental property of the evolution of stars that, as they age, their cores become smaller and their envelopes become larger. The present Sun has about 50% of its mass in the inner 50% of its radius. However, the Sun will become an RG in about 5 Gyr, and at a late stage in this process the inner 50% of mass will have shrunk into a core (effectively a WD) of one-twentieth of its present radius while the remaining 50% will have expanded into an envelope about 300 times its present radius. At the same time its luminosity will have increased to several thousand times its present value. More massive stars go through rather similar stages, but more rapidly the greater the mass.

Evidently, such expansion must somehow be modified if the Sun had a binary star companion in, say, an orbit like the Earth's. This orbit has a radius about 200 times that of the Sun. In fact when the Sun has reached a radius of only about 80 times its present value, the presence of this

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

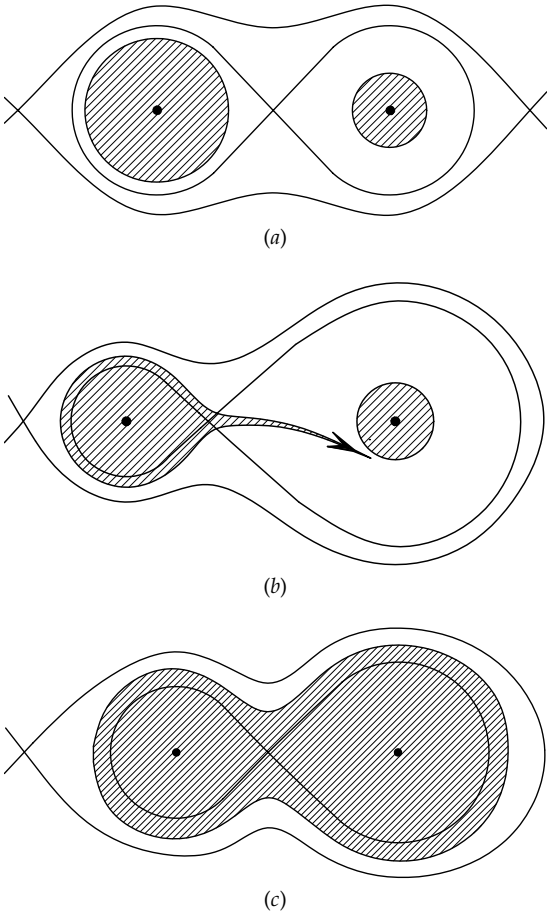


Figure 1. Critical surfaces for two stars in a CB. (*a*) The two stars both underfill their critical surfaces, forming a stable detached (D) pair. (*b*) One star fills its critical surface, to form an unstable semidetached (S) pair. (*c*) Both stars overfill their critical surface by the same amount, to form a stable contact (C) pair.

hypothetical companion in the Earth's orbit would cause the Sun's envelope to become unstable. The critical size is determined by its 'ROCHE LOBE' (figure 1). For any pair of masses, and any orbital separation, there is a critical figure-of-eight-shaped surface which has one component at the centre of each lobe. If the star's volume grows to exceed the volume of its Roche lobe, the star becomes unstable. Its material should pour out of its lobe through the crossover point, and into the other lobe.

Several stars are observed to be doing just this. For instance, in Algol (β Per), the first EB to be discovered (by Goodricke, in 1783), the radius of one component is closely equal to its lobe radius, and the presence of a gas stream in the system has long been recognized by the presence of emission lines. In an apparent paradox, the star which fills its lobe is less massive than the companion, contradicting the theoretical expectation that the more massive star should be the one to become unstable first. The answer to this is that the star which was originally more massive did indeed become unstable first, but then lost so much of its outer layers because of its instability that it is now the less massive by a large margin. Most, but probably not all, of the mass lost was accreted by the companion, increasing the margin even more.

Four aspects of stellar evolution are particularly important in shaping our expectations about binary stars.

(i) Stars tend to increase in size with age, at least as long as the star still has nuclear fuel to burn. They preserve their initial mass for most of their life, but at a late stage of high radius and high luminosity they eject a considerable fraction of their mass into space, leaving a compact, fuel-exhausted remnant.

(ii) The rate of evolution increases strongly with the degree of evolution; an extreme RG is evolving a hundred times or more faster than on the MS.

(iii) The rate of evolution is very sensitive to the initial mass of the star: if one component of a binary is say 20% more massive than the other, it can be expected to run through its entire life and become a compact remnant, while the other remains little changed on the MS. Stars whose initial masses are less than the Sun cannot evolve significantly at all, in the lifetime (believed to be ∼10–15 Gyr) of the universe.

(iv) The nature of the remnant (WD, NS or BH) probably depends mainly on the initial mass of the star, although with a great deal less certainty than the three previous propositions. WDs appear to come from stars no more than 6 or 7 times as massive as the Sun, BHs arguably from stars more than 30–50 times as massive as the Sun and NSs from intermediate masses. There may be some substantial overlap, however, perhaps depending on some apparently minor properties of a star such as its rotation rate, or the strength of its magnetic field.

The preliminary expectation is that, just as the evolution of a single star is mainly dependent on a single quantity, its initial mass, so the evolution of a binary will mainly depend on three quantities, the two initial masses and the orbital period (or equivalently, separation) (see CLOSE BINARY STARS: FORMATION AND EVOLUTION). The last of these will determine how far in its evolution the star of larger initial mass can proceed before it becomes unstable by filling its Roche lobe and transferring mass to its companion. Beyond this point, the evolution of both stars can become markedly different from what it would have been if they were not in the binary.

The three quantities referred to above, two masses and a period, determine (by Newton's law of gravity) not only the separation but also the angular momentum. It is a cardinal principle of physics that in an isolated system angular momentum is conserved. The transfer of mass between the stars, once one fills its Roche lobe, will alter the period and separation of the two stars, but should not (one might suppose) alter its angular momentum. This principle allows us to predict what the future evolution would be after the onset of mass transfer (usually referred to as Roche-lobe overflow). On this basis, we expect roughly the following:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(a) the loser continues to lose mass until it is stripped down to its compact (WD) core, the orbit becoming substantially wider in the process;
(b) the gainer, which is probably still an MS star unless its initial mass was rather closely equal to the gainer's, will start to evolve faster, and will in turn become large until it fills its own Roche lobe.

In practice, however, the situation appears more complicated. The likely explanation is that stars, especially in binaries, lose mass, not just to each other but to (in effect) infinity, so that the system is not isolated. The mass lost to infinity can carry angular momentum. If this mass is relatively rich in angular momentum, compared with the original binary, the loss can make the binary smaller. If it is relatively poor, it can make the binary wider, even (somewhat surprisingly) so wide that it becomes unbound, the two components themselves going to infinity in different directions.

Two rather extreme examples may illustrate this. The star 0957 − 66 (see table 1) is a binary of two WDs, both with masses about one-third of solar. The orbital period is less than 1.5 h. Even the total mass is too low for there to have been any evolution, and so the system must have lost at least a third of its original mass; more realistically, at least 60%. Also, the angular momentum is too low, by an even larger factor, for the system to have once contained two MS stars capable of significant evolution. The system must have lost at least 85–90% of its original angular momentum.

At the other extreme, we noted above that a high proportion of massive (O-type) MS stars are in CBs. O-type stars can be expected to evolve into NSs, and yet NSs are observed to be overwhelmingly single; only about 5% of radio pulsars are found in binaries. This suggests that the supernova explosion that accompanied the collapse of the core to NS size removed a good deal of mass, easily more than 50% of the system's original mass, while removing disproportionately little of the angular momentum. Such a process can disrupt the binary entirely, giving one and in the longer term two isolated NSs. A variant of this is that the supernova explosion has a degree of asymmetry to it, so that the remnant of the explosion receives a 'kick' which can be strong enough to eject it from the system.

In understanding the evolution of CBs we must therefore pay attention to possible processes causing mass loss or angular momentum loss, in addition to any Roche-lobe overflow that is to be expected. A few processes are listed here.

(a) Gravitational radiation is bound to cause angular momentum loss, though normally on a very slow timescale. However, the WD pair mentioned above should spiral into each other in less than 1 Gyr.

(b) Magnetic activity can cause mass loss. This activity is believed to be the driving force in the solar wind. Such wind is much stronger in evolved RGs than in MS stars, especially the relatively rapidly rotating RGs that are often found in CBs. The magnetic field, linking the star to the wind, can cause 'magnetic braking', a possibly efficient angular momentum loss process. A 0.1% loss of mass might be accompanied by a 10% loss of angular momentum.

(c) Even single stars lose much mass, perhaps 50–80% of the initial mass, because of much stronger 'superwinds' in their highly luminous latest stages. Some of this may be accreted by the companion even if neither star ever fills its Roche lobe. This mass loss will probably not cause much angular momentum loss, and the final binary might be about as close as the initial binary, or even slightly wider.

(d) Roche-lobe overflow can cause a catastrophic runaway if either the mass ratio (loser/gainer) is large or the loser has a deeply convective envelope such as is expected in an RG. The mass shed by the loser, instead of accumulating on the gainer, may accumulate in a 'common envelope' around the system and ultimately be expelled, carrying not only a large amount of mass but also a disproportionately large amount of angular momentum, so that the final binary may be a great deal smaller than the initial one. This process may even lead to coalescence, the remaining object being a rapidly rotating single star.

(e) As indicated above, a supernova explosion may remove much mass, and either increase or decrease the angular momentum in the remaining mass. A well-aimed 'kick' may even cause coalescence, the NS remnant hitting the MS companion and sinking to its centre.

Many observed binaries appear to have been modified by one or more of these processes, possibly in addition to Roche-lobe overflow.

Table 1 lists a number of binaries and their astrophysical parameters such as period, masses, eccentricity, separation and radii (where known). They demonstrate a wide range of evolutionary states, for each component. As well as the states MS, RG, WD, NS, BH we include two more, pre-main-sequence (PM) and hot remnant, which we describe below. A state such as 'RG + MS; S' means that there is an RG and an MS star in a 'semidetached' (S) state—figure 1(*b*)—where the larger star fills its Roche lobe. D refers to a 'detached' state—figure 1(*a*)—and C to a 'contact' state—figure 1(*c*).

TY CrA is an apparently very young binary, where the less massive star is still shrinking towards the MS; hydrogen fusion has not yet achieved equilibrium. When it does, the system will somewhat resemble GG Lup, where the less massive star is also the smaller. α Vir (Spica), GG Lup, RT And, α Aur (Capella) and V2291 Oph are all normal systems, some very close and some wider, in which no Roche-lobe overflow has yet occurred. In each case the more massive star will fill its Roche lobe in due course. In α Aur we see that both stars are giants: in comformity with point (iii) above they have nearly equal masses. V2291 Oph has quite a long period ($\sim$1 yr). If it were $\gtrsim$20 times longer still, the RG might well evolve without Roche-lobe overflow, but with a final superwind episode—process (c) above—that would turn it into a WD without much effect on either the period or the other star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Some close binaries, and their probable evolutionary states.

Name	State	P	e	a	M_1	M_2	R_1	R_2
TY CrA	MS + PM; D	2.89		14.5	3.16	1.64	1.80	2.08
α Vir	MS + MS; D	4.01	0.13	28	11	7	8.1	4
GG Lup	MS + MS; D	1.85	0.15	12	4.12	2.51	2.38	1.73
RT And	MS + MS; D	0.63		3.7	1.24	0.91	1.26	0.90
α Aur	RG + RG; D	104		161	2.61	2.49	11.4	8.8
V2291 Oph	RG + MS; D	385	0.31	424	3.86	2.95	32.9	2.73
α CMa	WD + MS; D	50 yr	0.59	4300	0.95	2.2	0.008	1.7
EG52	WD + WD; D	20 yr	0.18					
ζ Cap	WD + RG; D	2380	0.28					
CI Cyg	RG + MS; D	855	0.15	500:	1.7:	0.6:		
BF Cyg	HR + RG; D	756.8	0.3:	400:	0.5:	1:	0.5:	75:
UU Sge (Abell 63)	HR + MS; D	0.465		2.5	0.63	0.29	0.33	0.29
DQ Her	WD + MS; S	0.194		1.4	0.60	0.40		0.49
λ Tau	MS + MS; S	3.95		22	1.9	7.2	5.3	6.4
β Per	RG + MS; S	2.87		14	0.8	3.7	3.2	3.1
V1379 Aql	HR + RG; D	20.7	0.09	44	0.30	2.27	0.05	9.0
0957 − 66	WD + WD; D	0.061		0.6	0.32	0.37		
AM CVn	WD + WD; S	0.012:						
Z Her	RG + MS; D	3.99		15	1.31	1.61	2.7	1.85
RT Lac	RG + RG; S	5.07		16	0.63	1.57	4.6	4.3
ϵ CrA	MS + MS; C	0.591		3.5	1.5	0.15	2.2	0.7
ϕ Per	HR + MS; D	127		230	1.14	9.3		
CV Ser	HR + MS; D	29.7		125	6	23		
QV Nor	NS + MS; D	3.73	0.08	28	1.3	20		17
Cyg X-1	BH + MS; D	5.60						
1913 + 16	NS + NS; D	0.323	0.62	2.8	1.442	1.386		
V404 Cyg	BH + RG; S	6.47			>6			
1855 + 09	NS + WD; D	12.33	0.000 02					

Period (P) in days, unless specified; eccentricity (e) is zero unless stated; semimajor axis (a) in solar radii; masses (M) and radii (R) in solar units. A colon indicates substantial uncertainty. Evolutionary 'state' is explained in the text.

This would leave it like α CMa (Sirius), where the WD has only half its companion's mass but was presumably once the more massive component. Further evolution of the second star in V2291 Oph would be similar (still supposing that the period were longer) and leave a wide pair of WDs like EG52.

However, ζ Cap shows that some significant interaction can take place with orbital periods of a few years. Here one star, presumably the more massive originally, is a WD, but the companion RG has a spectrum unusually rich in barium (Ba)—and also some other elements. These elements are expected to be formed by nuclear fusion only at a very late stage in evolution, shortly before the superwind stage. They must have been produced by the first star, ejected in its superwind, and accreted by the second star where it now shows up in the spectrum.

BF Cyg is possibly in a state that ζ Cap will evolve to, when its giant component is even larger. This 'symbiotic' system shows a combination of a cool, very large star and a small hot object, possibly a hot remnant of an RG that has lost its envelope but is not yet a WD. The hot remnant seems to be interacting with a wind from the current RG component, that perhaps half-fills its Roche lobe. When it does fill its Roche lobe, the interaction may be quite dramatic—case (d) above—leading to a common-envelope phase and leaving a WD + WD pair, probably intermediate in period between 0957 − 66 and EG52.

CI Cyg is another symbiotic system, but here the hot object is arguably an MS star, heated by accretion from the RG wind, rather than a WD. It is therefore less rather than more evolved than ζ Cap (and probably of substantially less mass). A common-envelope episode looks rather likely, because the future loser, the RG, is perhaps three times as massive as the gainer. The mass expelled in the common envelope may briefly shine as a planetary nebula in the UV light from the exposed RG core (a hot remnant), and if the period is severely decreased by the common-envelope phase, as expected, we may get an object like UU Sge, which is surrounded by the planetary nebula Abell 63. A number of other planetary nebulae have CB central stars, which are most easily understood as remnants of a common-envelope phase (see BINARY STARS IN PLANETARY NEBULAE).

In UU Sge the hot remnant might be expected to cool off quite soon to a WD, and then the system might not change further because the other star has too little mass to evolve—point (v). However, angular momentum loss by process (b) might well operate, since cool rapidly rotating MS dwarfs are well known to be magnetically active. This would shrink the binary until it might become

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

like DQ Her, where the cool dwarf fills its Roche lobe. Such binaries are 'cataclysmic' variables, which show episodic outbursts. These are due to a combination of instability in the accreting gas flow (dwarf novae) and thermonuclear instability in the gas recently accreted on the surface of the WD component (classical novae) (see CATACLYSMIC BINARIES: CLASSICAL AND RECURRENT NOVAE). DQ Her had a classical nova outburst in 1934.

Systems of short period initially, such as GG Lup, α Vir and RT And, should reach Roche-lobe overflow much sooner than say CI Cyg, perhaps even before making the transition from MS to RG. λ Tau and β Per (Algol) are two such binaries with ongoing Roche-lobe overflow. The first may be descended from a relatively massive system, intermediate between α Vir and GG Lup, the second from something intermediate between RT And and GG Lup. The loser in β Per is now an RG, but was probably MS when Roche-lobe overflow began. Unlike in the wider systems such as V2291 Oph or CI Cyg, there is no room for the RG, in the future, to evolve to large radius and a massive WD core. Instead the RG in β Per is likely to lose all its envelope when the core is still a low-mass object. The product will be a system like V1379 Aql, where the hot remnant (not yet a WD) is only 10% of its companion's mass, and is presumably the remnant of a former RG. The companion in V1379 Aql has itself evolved from the MS to become an RG, and soon will fill its own Roche lobe. The severe mass ratio will probably make this a common-envelope event—point (d)—and the next stage will be two low-mass WDs, in a very short-period orbit, presumably like $0957-66$. Gravitational radiation—point (a)—will cause these WDs to spiral in towards each other, on quite a short timescale (about 10 million years), to give a third episode of Roche-lobe overflow. The system may resemble AM CVn, which appears (but with considerable uncertainty) to be an ultrashort-period binary of two WDs with the larger (and less massive) component overflowing its Roche lobe.

Some binaries, such as Z Her, are wide enough for one component to become at least a small RG before Roche-lobe overflow. It is odd here that the more evolved component is slightly less massive than its companion. This may be because of point (b) above; magnetic activity may cause much faster than normal mass loss by stellar wind, shortly before Roche-lobe overflow begins. However, the companion is itself quite evolved, and may also be an RG when Roche-lobe overflow begins. We may therefore get a double-RG semidetached binary, like RT Lac. The later evolution may also take the system through stages like V1379 Aql, $0957-66$ and AM CVn.

In a system initially like RT And, where the two components are almost touching even though they are hardly evolved, we can expect that very shortly after Roche-lobe overflow the components actually come into contact (figure 1(c)). Both nuclear evolution and magnetic braking—point (b)—may work to make the system even smaller during this contact phase, but can do so only if the mass flow is on average reversed, so that the less massive star is gradually absorbed into the more massive. This may be the origin of a system like ϵ CrA, which must be at quite an advanced stage since one component is 10 times as massive as the other. An oddity in this and other contact binaries is that the two components are at roughly the same temperature and luminosity, whereas if they were separate the less massive component would be at least 1000 times fainter. This seems to indicate that when two stars are in contact the hotter one heats up the cooler one by sending energy through the narrow 'neck' joining them (figure 1(c)). Presumably mass flows slowly in the opposite direction, until the less massive star is entirely swallowed up, and what remains is a single star giving little indication of its former binary history.

If one component is massive enough, it may be destined to end up as an NS or a BH, rather than a WD. α Vir may well produce two NSs in due course, λ Tau probably only one NS and one WD. ϕ Per perhaps represents a late stage in the future evolution of α Vir, when Roche-lobe overflow has stripped the initially more massive star down to a hot remnant that is probably not very far from a supernova explosion. The longer period in ϕ Per is not inconsistent with a much shorter initial period, since conservation of angular momentum should tend to make the period increase as Roche-lobe overflow progresses. CV Ser might be the comparably late stage of a substantially more massive binary still, and arguably the more evolved but less massive component could be heading for a BH rather than NS. Although we argued earlier that a supernova explosion is likely to disrupt a binary, blowing the two components apart into separate single stars, it evidently does not always do this, since binaries such as QV Nor and Cyg X-1 have an NS and a BH companion respectively. These ultracompact companions are identified by the fact that they are highly luminous in x-rays (see X-RAY BINARY STARS). The BH is inferred in Cyg X-1 because, although its mass is not well known, it must on any reasonable model be several times higher than the maximum possible mass of an NS.

In both QV Nor and Cyg X-1 there is a normal star as well as a compact remnant. This normal component must itself grow with age, and lead to a second episode of Roche-lobe overflow. However, in these cases the mass ratio is very extreme, which by point (d) suggests a common-envelope phase with the likelihood that the orbit shrinks by a large factor. This may be the origin of a double-NS binary like 1913 + 16. No BH + NS binary has yet been recognized, but we can probably expect a small number of them in the Galaxy.

A rather large initial mass ratio seems required to produce a system like V404 Cyg, where one component is a BH while the other is a quite modest RG. Perhaps this started with a period as long as several years, with one star of 40 solar masses and the other of only 2 or 3. We know of no such systems directly, but they would be very hard to recognize. The more massive star may have briefly become an extreme RG, filled its Roche lobe and then suffered massive envelope ejection and simultaneous

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

orbital shrinkage by the common-envelope mechanism. The remnant of the massive star then collapsed in a supernova explosion to produce a BH, but, perhaps rather flukily, not disrupting the new short-period binary. Then the low-mass component evolved and now suffers Roche-lobe overflow. The RG should lose all its envelope, and end up as a WD: the result would be somewhat like the binary 1855 + 09, although with a BH rather than NS. To obtain 1855 + 09 itself, we might need moderately less massive original stars, but the history might be similar.

The science of understanding the evolutionary processes by which CB stars have reached the configurations in which they are observed is still in its infancy. By no means all of the small selection in table 1 fit comfortably with present understanding, and several much more difficult systems are known. To add to uncertainty, several stars turn out on close enough inspection to be triple, and the third star may be capable of interacting with the first two to give an extra layer of complexity.

Peter Eggleton

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Close Binary Stars: Formation and Evolution

A binary system consists of two gravitationally bound stars that orbit their common center of gravity (COG). Observations indicate that the majority of stars are members of binary systems or even of multiple systems consisting of three, four or more stars. The multiple systems tend to be hierarchical, that is triples usually consist of a close pair with the third companion orbiting at much larger distance, etc. Single stars such as the Sun are less common. It is therefore impossible to develop a comprehensive theory of stellar evolution without including effects of binary evolution.

We speak of CLOSE BINARY STARS when the two stars are so close that they influence each other's evolution profoundly. Most commonly this is effectuated by the tidal forces (see below) by which the stars are deformed and, when one of the stars expands beyond a certain critical surrounding surface (ROCHE LOBE), cause mass exchange between the two stars. Since stars swell up considerably during their lifetime, often by several orders of magnitude, even binaries with orbital periods of decades can be close binaries. G P Kuiper (1941) was the first to recognize the importance of Roche lobes for binary stars. Close binaries can be classified into three broad classes: *detached* (D), *semidetached* (SD) and *contact* (C), depending on whether none (D), only one (SD) or both stars fill their critical Roche surface for mass exchange.

The prototype semidetached close binary system is ALGOL (β Persei); its periodic (2.87 days) brightness variations were first discovered in 1667 but it was only in 1783 that John Goodricke explained these variations in terms of an eclipse of one star by the other in a close binary system. After the introduction of stellar spectroscopy at the end of the 19th century this idea was shown to be correct. However, with the advent of stellar evolution calculations on electronic computers around 1950 the Algol system appeared paradoxical: how can the less massive ($0.8M_\odot$) subgiant component be further evolved than its more massive ($3.7M_\odot$) main-sequence companion? It was known that a more massive star always evolves faster owing to its higher central temperature and correspondingly faster rate of thermonuclear fusion. Furthermore, it was discovered that the subgiant components of Algol-type systems do not obey the usual mass–luminosity relation but that the luminosities and radii were significantly larger. In 1955 J A Crawford came up with the solution of the 'Algol paradox': initially the subgiant was the more massive component until it expanded beyond its Roche lobe and transferred a large fraction of its mass to the less evolved main-sequence companion. The currently observed peculiar subgiants in Algol-type systems are thus remnant stars that have lost most of their mass. Mass transfer between two stars is a basic feature of the evolution of close binary systems whereby very diverse systems can be formed. Obviously, the currently observed binary configuration depends on the binary's evolutionary phase and on its initial conditions (mainly the initial masses, orbital period and orbital eccentricity), i.e. on the primordial binary system. Obviously, these initial conditions are determined by the binary formation process which is, however, still shrouded in mysteries. Observations of PRE-MAIN-SEQUENCE BINARIES indicate that binaries are formed as a byproduct of star formation, but the details remain obscure.

In the past few decades, when satellite observations of cosmic UV and x-ray sources became available, much attention has been given to binaries which harbor compact stars. Large amounts of high energy are liberated by internal dissipation when the transferred mass falls into the deep potential well of a compact companion. Examples of such systems are the CATACLYSMIC BINARIES (white dwarf accretor) and X-RAY BINARIES (neutron star or black hole accretor).

Tidal deformation of binary stars

When the orbital separation D between two binary components is large compared with the dimension of the largest star, the dynamics of the two bound stars can be adequately described by a system of two point masses. However, when the size R of the largest star is not much smaller than the orbital separation D one has to take into account the internal structure of the star(s). The difference between the companion's attractive gravitational force at the close and the far sides of the other star is governed by Newton's inverse square law. The ratio between the companion's gravitational force at the two extreme points of both stellar hemispheres of the larger star is thus $[(D-R)/(D+R)]^{-2}$. This ratio can differ significantly from unity if R/D is not much smaller than unity. Clearly the mutual gravitational attraction not only keeps the two stars together in their orbit but also squeezes the stars by pulling them apart along the line connecting the two stellar centers. This deformation is a manifestation of the 'tidal force', i.e. the difference of the gravitational force over the star. If the orbit is eccentric and/or the stellar spins are not synchronous with the orbital revolution the tidal deformation of the close binary components is obviously time dependent. The decrease of orbital energy through inevitable dissipation of these periodic tidal deformations generates tidal torques between the stars that tend to circularize (and shrink) the orbit and tend to synchronize the stellar spins with the orbital revolution. The time scale for this tidal relaxation process declines rapidly with increasing R/D, i.e. with increasing tidal squeezing of the binary components. For sufficiently close binaries the tidal deformation of the stars can become so severe that surface matter in one of the stars (or both) becomes unbound and starts flowing off to the companion, a phenomenon known as *Roche lobe overflow* (RLO).

Principles of mass transfer in close binaries

Let us adopt the so-called 'Roche approximation', i.e. we assume that the tidal torques have acted sufficiently

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

long for the binary orbit to have become circular (i.e. both stars describe circles about their common COG). We also assume that, as a result of tidal torques, the stars spin about their individual COG with the same angular speed as the orbit. Because the internal mass distribution of stars exhibits a strong concentration to the stellar center (generally increasing with age) we may, as a first approximation for the gravitational field in the rarefied outer layers, assume that all the mass is concentrated at the stellar center. Let us now consider a test particle (i.e. a point mass with negligible mass compared with the stellar masses M_1 and M_2) drifting outside the two stars. Introducing a frame rotating with the same angular speed $\Omega = [G(M_1 + M_2)/D^3]^{1/2}$ as the orbital motion about the COG, we can express the acceleration experienced by this particle as:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = -\frac{GM_1}{r_1^3}\vec{r}_1 - \frac{GM_2}{r_2^3}\vec{r}_2 - \vec{\Omega} \times (\vec{\Omega} \times \vec{r}) - 2\Omega \times \vec{v}$$

where G is Newton's gravitational constant, $\vec{r}_1$ and $\vec{r}_2$ are the vectors from the two stellar masses to the test particle and $\vec{r}$ is the vector from the COG to the particle. The last two terms on the right describe the centrifugal and the Coriolis acceleration in our rotating frame. It can be shown that the acceleration of a co-rotating test particle (with $\vec{v} = \vec{0}$) can be expressed as the gradient of the effective gravitational potential (corrected for the centrifugal effect):

$$\Psi = -\frac{GM_1}{r_1} - \frac{GM_2}{r_2} - \Omega^2\frac{d^2}{2}$$

where d is the particle's distance to the orbital axis through the COG.

Surfaces of equal Ψ are known as Roche equipotential surfaces; see figure 1. The cross sections with the orbital plane contain five special points, called *Lagrangian points* L_1, $L_2, \ldots, L_5$, where the acceleration of the comoving particle vanishes, i.e. where the gravitational acceleration due to the two stars is precisely cancelled by the centrifugal acceleration away from the orbital axis. The equipotential surface through the inner Lagrangian point L_1 consists of two pear-shaped Roche lobes and forms the biggest closed surface around each of the two stars, merging at L_1. It can be shown that the Roche equipotential surfaces inside the stars coincide with surfaces of equal pressure and density: the Roche equipotentials (approximately) give the shape of the two stars in the binary. What happens if one of the two stars expands and fills its Roche lobe? We have seen that a test particle at rest in L_1 experiences no net acceleration. However, inside the lobe-filling star acts another force due to the gradient of the thermal pressure in the stellar gas. If the star is detached this pressure force balances the inward pull of gravity of the star (corrected for the centrifugal force due to the stellar spin). But since the net gravitational force (corrected for rotation) is zero in L_1 the pressure force there is unbalanced and blows a jet of gas through the (saddle point) region around L_1 into the the companion's Roche lobe. There it will be rapidly accelerated to supersonic speeds by the companion's gravitational attraction. The transferred mass (or at least a fraction of it) will ultimately fall onto the companion. As the mass flow through L_1 carries orbital angular momentum it cannot fall straight onto the companion and—if the latter is compact enough for the stream not to hit it directly—it will form an accretion disk around the companion in which part of the infall energy is dissipated and radiated away. The mass exchange profoundly alters the further evolution of both stars.

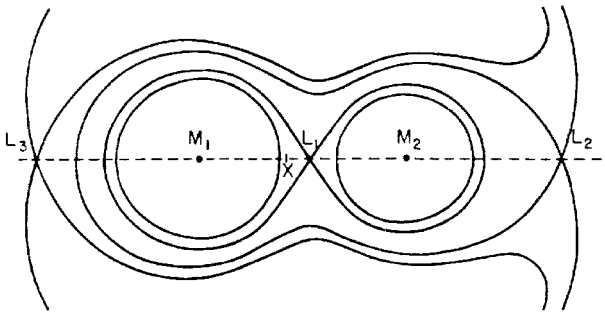


Figure 1. The cross section with the orbital plane of some special Roche equipotentials. The surface through the inner Lagrangian point L_1 forms the two Roche lobes around the stars. The cross denotes the binary's center of mass.

Orbital evolution

The process of mass exchange is a complicated hydrodynamic phenomenon the details of which are still poorly understood. The broad picture is as follows. Let us assume for simplicity that all the matter that flows into the companion's Roche lobe is ultimately accreted by the companion and that the system's orbital angular momentum

$$J_{\mathrm{orb}} = \left(\frac{GM_1^2M_2^2D}{M_1 + M_2}\right)^{1/2}$$

is *conserved* during the mass exchange. Part of the orbital angular momentum could be absorbed (through tidal torques) by the stellar spins but, except for extreme systems, this can usually be neglected. Because the amount of mass and angular momentum stored in the transferred, but not yet accreted, matter can be neglected relative to that of the two stars, the new orbital separation after exchange of a given amount of mass is given by the condition that J_{orb} is constant. If we differentiate the above expression with respect to time (indicated by a dot) then the condition $\dot{J}_{\mathrm{orb}} = 0$ yields

$$\frac{\dot{D}}{D} = -2\left(1 - \frac{M_1}{M_2}\right)\frac{\dot{M}_1}{M_1}$$

which shows that mass transfer from the more massive star (M_1) to the less massive star ($\dot{M}_1 < 0$) yields $\dot{D} < 0$ and thus gives rise to a *shrinking* orbit. On

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the other hand, when $M_1 < M_2$ mass transfer to M_2 *widens* the orbit. Observations of many binary systems confirm this basic picture, although many of the details are still poorly known. For example, observations of circumstellar matter around SD close binaries indicate that some of the transferred matter may escape accretion by the companion. However, except in extreme situations, the assumption of no mass loss from the system seems generally to yield at least a qualitatively correct picture.

Response of mass-losing star

Since the more massive star (primary) in the binary has the largest radius and evolves fastest it will be the first to expand to its Roche lobe in initially detached close binary systems. We have just seen that the orbital separation D decays in response to mass transfer to a less massive component. The size of the Roche lobe around the primary shrinks with D and it depends on the primary's response to the loss of surface matter on what time scale the mass transfer is continued. This response depends on the stratification in the primary's envelope: a deep *convective envelope* is characterized by a steep temperature gradient. When mass is lost at its surface, rapid adiabatic expansion of the newly exposed surface layers—to restore global hydrostatic equilibrium—then brings up relatively hot material, making the star initially bigger than its thermal equilibrium configuration. On the other hand, when the envelope is predominantly *radiative*, for which the temperature gradient is less steep, adiabatic expansion causes the star to be smaller than its thermal equilibrium size. Obviously, RLO from a deep convective envelope is intrinsically unstable because the larger the mass loss, the stronger the rapid adiabatic expansion and the larger the primary's extension beyond the Roche lobe. Figure 2 shows the schematic evolution of a binary undergoing Roche lobe overflow (RLO). We can distinguish three time scales for the mass transfer.

Dynamical time scale mass exchange occurs when the primary's quick adiabatic response to the mass loss brings the primary's surface beyond the shrunken Roche surface. The primary cannot be kept inside its Roche lobe, so that its surface layers will be depleted by continuous expansion through L_1. This situation occurs when the primary is a (super)giant and has a deep convective envelope. A similar unstable situation occurs when the (more massive) contact star is degenerate (a white dwarf) because such stars always expand when losing mass.

Thermal time scale mass exchange occurs when after the quick adiabatic expansion that restores global hydrostatic equilibrium the primary remains inside its shrunken Roche lobe and no runaway mass exchange takes place. The mass transfer is then driven by the primary's (slower) thermal relaxation: by restoring thermal equilibrium in its perturbed outer layers the star will re-expand beyond its Roche lobe. This type of Roche lobe overflow occurs in primary's with (predominantly) radiative outer layers and is self-regulating because the thermal relaxation and consequent expansion compete with the shrinking due to rapid mass loss.

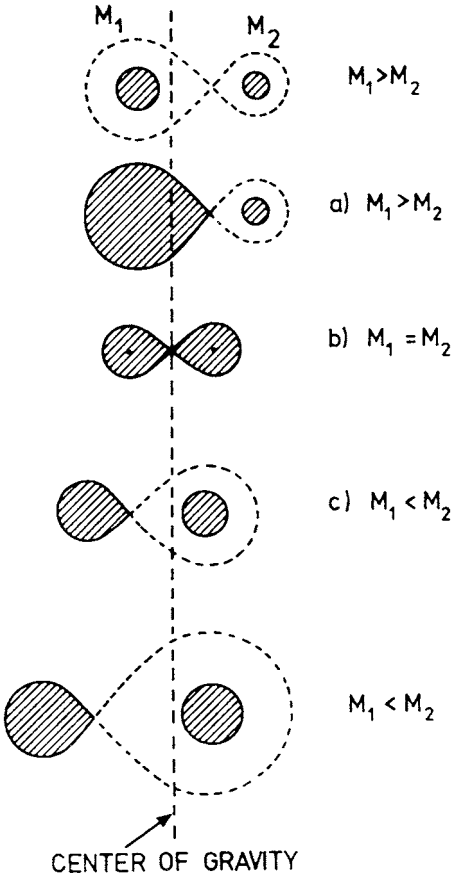


Figure 2. Subsequent stages of evolution of a close binary system in which the primary at the onset of RLO (in frame (*a*)) has a predominantly radiative envelope. In the upper frame the star expands on a nuclear time scale towards its Roche lobe. The phases (*a*) and (*b*) correspond to rapid thermal time scale mass transfer which terminates at (*c*). From then on the binary expands and mass transfer occurs on the intrinsic evolutionary expansion time of the now less massive contact star.

Nuclear time scale mass exchange occurs when the contact star is less massive than the accretor and the binary separation and generally the Roche lobe widen. Under these circumstances mass exchange can only be sustained by the primary's evolutionary expansion driven by thermonuclear fusion reactions in its interior. This situation occurs in the later phases of mass transfer after the mass ratio of the binary has been reversed and yields the slowest mode of mass exchange.

Response of accreting companion

Let us now consider what happens to the accreting secondary during incipient RLO. Being the less massive star in the system the accreting detached companion generally has a longer thermal time scale than the contact star. Therefore its thermal equilibrium becomes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

severely perturbed by the rapid accretion of the transferred material, especially because the accreted material carries significant kinetic energy after falling down from L_1. This kinetic energy is mostly dissipated at or near the accretor's surface, heating it considerably. The (generally) small fraction of the orbital angular momentum carried along by the accretion stream can bring the accretor in rapid rotation. It is likely that the accretor will swell up to or beyond its Roche lobe, especially during a dynamical time scale mass transfer phase, but possibly also during a thermal time scale exchange. In that case the binary soon evolves into a contact system with a common envelope around the two Roche lobes. The subsequent evolution of such contact systems is poorly understood.

Cases A, B and C of mass exchange

There exist three main evolutionary phases during which a star expands and may first fill its Roche lobe, namely the phase of core hydrogen burning (main sequence phase) and the phases of rapid core contraction towards helium ignition and towards carbon ignition. Kippenhahn and Weigert (in 1967) introduced the corresponding classification into respectively cases A, B and C of mass exchange.

Case A. The primary fills up its Roche lobe while still on the main sequence. Once Roche lobe overflow begins the primary's thermal equilibrium becomes increasingly perturbed and a rapid phase of thermal time scale mass exchange follows by which the mass ratio q is more than reversed. After that the evolution slows down to a nuclear time scale. The Roche lobe filling star has become an overluminous core hydrogen burning subgiant. Many eclipsing variables of this type are known. Numerical simulations show that case A mass exchange almost always leads to a contact binary.

Case B. The primary is a (sub)giant burning hydrogen in a shell when it reaches its Roche lobe. The initial rapid phase of mass transfer can either be on a dynamical time scale (when the primary has developed a deep convective envelope) or on a thermal time scale when the convection in the envelope is less developed. The subsequent slower phase of mass transfer to the now more massive companion proceeds either on the thermal time scale of the contracting helium core or on the nuclear time scale of the hydrogen shell source in the less massive primaries that have developed a compact electron degenerate helium core.

For primaries originally more massive than $\simeq 3M_\odot$ the mass exchange terminates when the primary ignites helium in the core whereby the remnant of the primary (almost stripped of its hydrogen envelope) shrinks inside its Roche lobe. Case B in massive binaries thus produces almost pure helium stars (WOLF–RAYET STARS (WR stars)) in orbit with a main sequence companion. The more massive helium stars (roughly $M_{\rm He} > 2.5M_\odot$) subsequently explode as a type II supernova: the remnant core collapses to a neutron star, or even a black hole.

Primaries originally less massive than $3M_\odot$ terminate the rapid mass transfer phase when the stellar expansion slows down with the onset of electron degeneracy in the highly compressed helium core. The subsequent slow mass exchange produces binaries with mass ratios as small as 1/5 or 1/10. The main sequence secondary is now much brighter than its less massive (overluminous) subgiant contact companion. If the orbital inclination is favorable such a binary system shows eclipses and will be classified as an Algol-type binary. The mass transfer terminates when the contact star's hydrogen envelope is depleted by both shell burning and mass loss. The remnant envelope shrinks within its Roche lobe and the primary evolves into a low-mass helium white dwarf.

Case C. This corresponds to mass transfer from evolved stars with very extended envelopes and has been little studied partly in view of the difficulties in treating the extended cool outer layers of the contact star.

Figure 3 shows the orbital period values that correspond to the three classes of mass exchange for a binary consisting of a $7M_\odot$ plus $3M_\odot$ (or $5M_\odot$) star. To compare the size of the Roche lobe with the usually spherical stellar models one determines the Roche radius: that is the radius of a sphere with the same volume as the Roche lobe.

Non-conservative evolution of binaries

We have already noted that part of the mass stream through L_1 in semidetached binaries may leave the system, but so far we have no means to quantify this reliably. However, there exist other ways by which mass may be lost from a binary. Observations show that many stars lose mass in the form of stellar winds, independent of a possible binary companion. The companion may in fact capture part of this ejected wind material, so that mass exchange can occur even in detached binaries. However, the transfer rates are generally much smaller than for RLO. In (super)giant stars and very massive early-type stars the stellar wind losses take away a significant fraction of the star's mass. A fast isotropic wind which decouples from the star at its surface leaves the angular momentum of the remaining mass invariant and thus widens the binary orbit as the gravitating mass in the binary decreases.

Stars with surface magnetic fields strong enough to couple to the ejected ionized wind material and enforce corotation out to several stellar radii from the surface are spun down by the magnetic torque. In SD binary systems effective tidal interaction tends to keep the stellar rotation synchronous with the orbit, so that the spin angular momentum carried off by the magnetically coupled wind must be compensated by the system's orbital angular momentum. An effectively coupled wind can thus enforce a decrease of the orbital separation. If these wind losses can occur simultaneously with RLO the mass transfer rate is enhanced by the induced shrinking of the Roche lobe. It is commonly believed that such '*magnetic braking*' occurs in systems with G- and K-type stars, e.g. in cataclysmic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

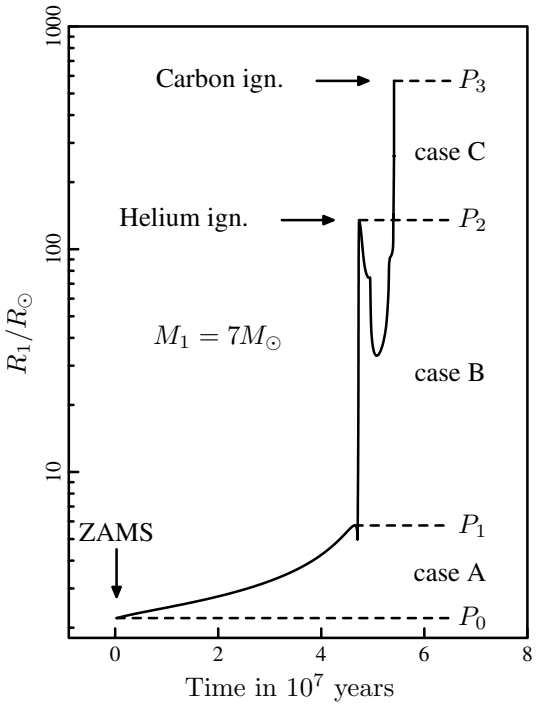


Figure 3. The radius expansion of a $7M_\odot$ star defines the critical orbital periods for the three cases A, B and C of RLO. The size of the Roche lobe is determined by the mass ratio M_1/M_2 and the orbital period (separation). A binary with an orbital period as short as P_0 would be semidetached from the very beginning of its main sequence phase. For a $3M_\odot$ ($5M_\odot$) companion the critical periods are (in days): $P_0 = 0.52$ (0.47), $P_1 = 2.2$ (2.0), $P_2 = 247$ (225) and $P_3 = 2137$ (1951). For an initial period of say 50 days the binary would evolve as case B: the primary would fill its Roche lobe after exhaustion of hydrogen in the core.

variables (CVs). The observationally inferred spin-down of G stars with age gives us information about this magnetic braking although it is not at all certain that these values may be applied to rapidly rotating binary stars.

Another non-conservative effects follows from Einstein's general theory of relativity which predicts that a compact binary system emits *gravitational radiation* carrying off angular momentum. The resulting rate of orbital angular momentum loss is

$$\frac{\dot{J}}{J} = -\frac{32}{5}\frac{G^3}{c^5}\frac{M_1 M_2}{D^4}(M_1 + M_2)\,\mathrm{s}^{-1}$$

where c is the speed of light. In close binaries with orbital periods less than about half a day the induced orbital decay becomes significant.

Common envelope binaries

Little is known about the transfer of mass and energy between two stars in contact binaries and their evolution cannot be simulated reliably. The best known class of contact binaries are the W Ursae Majoris (W UMa) systems which consist of stars of spectral type G and K, with a typical mass ratio of 2. These stars have convective outer layers and thus a convective common envelope which is thought to engulf the Roche lobes of both stars and to lie inside the equipotential surface through L_2. Contact binaries with massive early-type stars (which have radiative envelopes) are also observed. Furthermore there is observational evidence for contact binaries with cool supergiant components with deep convective envelopes.

Spiral-in evolution

It was suggested by B Paczynski (in 1976) and is now commonly believed that, under certain circumstances, the evolution of common envelope binaries can be very violent in contrast to the observed, apparently stable, systems mentioned above. A critical factor seems the binary's mass ratio $q = M_1/M_2$. If the contact star in a close binary is much more massive and bigger than its companion it will absorb a non-negligible fraction of the orbital angular momentum when tidal forces try to keep its spin synchronous with the speeding up of its companion's narrowing orbital revolution. Even before the SD phase the natural expansion of the primary slows down its spin whereby tides become active to resynchronize the star. Because of the primary's absorption of orbital angular momentum the companion must spiral slightly inward after which it revolves even faster in its orbit. In order to speed up the spin of the big star towards synchronism ever more angular momentum needs to be tapped from the orbit and for a sufficiently big primary the binary undergoes a catastrophic tidal instability (G H Darwin, 1908) whereby the companion dives into the primary's envelope. The companion then moves with supersonic speeds through the primary's envelope whereby shock waves violently dissipate orbital energy, leading to even faster orbital decay. The details and precise outcome of this complicated phenomenon are unknown. In many cases the primary's envelope may be expelled by the dissipated orbital energy, so that virtually only its core remains in a tight orbit with the perhaps ablated companion. The observed planetary nebulae with double cores (about 5–10% of the population) may be the result of such evolution. Another observational link exists, namely with the CVs, compact binaries with a white dwarf component which is surrounded by a luminous accretion disk. These systems show weak periodic outbursts (dwarf novae) or strong eruptions (novae).

Formation and evolution of CVs

CVs are thought to have evolved through a spiral-in phase because the WHITE DWARF component must be the remnant core of a former big giant star which can never have been accommodated in the currently observed tight CV binary systems with typical orbital periods less than 10 h. Presumably the unevolved main-sequence companion of the white dwarf, originally in a wide orbit, spiralled into the much bigger and heavier giant (white dwarf progenitor) owing to a tidal instability and evaporated the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

latter's extended envelope, so that only the degenerate core (white dwarf) remained. For orbital separations (just after the spiral-in phase) larger than a critical value the binary evolves slowly to a wider system when the unevolved star expands to its Roche lobe and a slow nuclear time scale mass transfer sets in. In the narrower systems magnetic braking of the main-sequence star is thought to dominate, whereby the binary separation decays even when the main-sequence star fills up its Roche lobe and transfers matter to the more massive white dwarf. The enhanced mass transfer generates a hot, luminous accretion disk around the white dwarf and the binary has become a CV.

The magnetic braking time scale being much shorter than the contact star's nuclear time scale, the nuclear evolution of this star is effectively frozen and it evolves down the main sequence, remaining roughly in thermal equilibrium. However, when the orbital period has been decreased to about 3 h the continuing mass loss has mildly perturbed the contact star's thermal equilibrium. Observations indicate that no CV binaries exist with a period between roughly 3 and 2 h: the so-called *period gap*. The period gap is believed to be caused by the sudden termination of magnetic braking with consequent interruption of mass transfer as the star shrinks inside its Roche lobe as it restores thermal equilibrium. It is thought that magnetic braking terminates when the contact star has become fully convective—which occurs roughly at this stage of evolution in numerical simulations. By then the only angular momentum losses from the binary are due to gravitational radiation.

Only when the orbit has decayed and the Roche lobe has tightened again around the low-mass star can the RLO continue. When continuing RLO has reduced the contact star's mass to about $0.1M_{\odot}$ its thermal time scale becomes longer than the time scale for mass loss, i.e. longer than the time scale for gravitational radiation losses. Further mass loss perturbs its thermal equilibrium significantly and the star evolves away from the main sequence. At this time the nuclear fusion reactions in its interior begin to fade, whereby the electron gas becomes degenerate. After a while the contact star becomes so degenerate that its mass–radius relation reverses and it begins to expand in response to any mass loss. From then on the mass transfer rate is dictated by the condition that the Roche lobe, and thus the stellar separation, must increase. The binary system thus passes a *minimum orbital period* which, for hydrogen rich envelopes, turns out to be about 80 min. The observed absence of CV systems below $P_{\rm orb} \simeq 80$ min confirms that the mass transfer in these compact CVs is indeed driven by gravitational radiation.

Formation and evolution of high-mass x-ray binaries (HMXBs)

Massive x-ray binaries consist of a massive, early-type (OB) star in orbit with a NEUTRON STAR which captures matter from the massive star's stellar wind through gravitational focusing. The existence of a neutron star implies that the binary survived a violent SUPERNOVA (SN) explosion by which the neutron star was formed after the collapse of the primary's burnt out stellar core. It is believed that before the SN explosion the binary evolved through a case B mass transfer by which the mass ratio reversed, so that the helium star remnant of the primary had become the less massive component when it exploded as a SN (van den Heuvel and Heise, 1972). Had it been the more massive star that exploded the binary would have been dispersed, unless the explosion was asymmetric and happened to generate the right net kick in the direction opposite to the orbital motion so that the binary could survive.

In the *Bondi–Hoyle* approximation the neutron star's effective wind interception radius is $R_A \simeq GM_2/V^2$, where V is the velocity of the (supersonic) wind material near the neutron star (in its rest frame). For typical wind speeds R_A is larger than the neutron star's geometric radius. The wind velocities are high, of order the escape speed from the primary's stellar surface, so that the companion can accrete only a small fraction of the emitted wind. The wind matter heats up and emits x-rays when it collides with the neutron star, producing a bright x-ray source (e.g. Vela X-1); this corresponds to stage (*e*) in figure 4. In some cases (Cen X-3) incipient RLO may be responsible for the x-ray emission, although full-blown thermal time scale mass transfer from the massive star presumably leads to disappearance of the x-ray source by self-absorption and to a subsequent spiral-in of the neutron star and evaporation of the massive star's envelope. The spiral-in may produce a new, but now much more compact, x-ray binary system like Cyg X-3 with an orbital period of only 4.8 h, corresponding to stage (*g*) in figure 4. Infrared spectroscopy has shown that the companion of the neutron star is a nitrogen-type WR star. Such stars are helium stars with masses in excess of about $5M_{\odot}$, so that the helium star in Cyg X-3 is probably the bare core remaining after spiral-in of the neutron star into the much more massive original secondary. When the second helium star has evolved to the SN phase the binary may be disrupted, leaving two single neutron stars in which the youngest, because of its rapid spin and strong magnetic field, acts as a radio pulsar. If circumstances are favorable (supernova kick) and the binary remains bound, we are left with a compact binary pulsar system, of which several are observed and in which the orbital decay by gravitational radiation can be accurately measured because of the accurate clock provided by a rapidly spinning neutron star (radio pulsar). Observations by J H Taylor (in 1986) beautifully confirmed the prediction from general relativity.

Formation and evolution of low-mass x-ray binaries (LMXBs)

Our Galaxy contains about 100 low-mass x-ray binaries in which the neutron star (or BLACK HOLE) receives mass from a low-mass contact star. Figure 5 shows a typical HMXB and LMXB system. Only for a minority of these systems has the orbital period been measured; most orbital periods turn out to be less than 20 h. However, from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

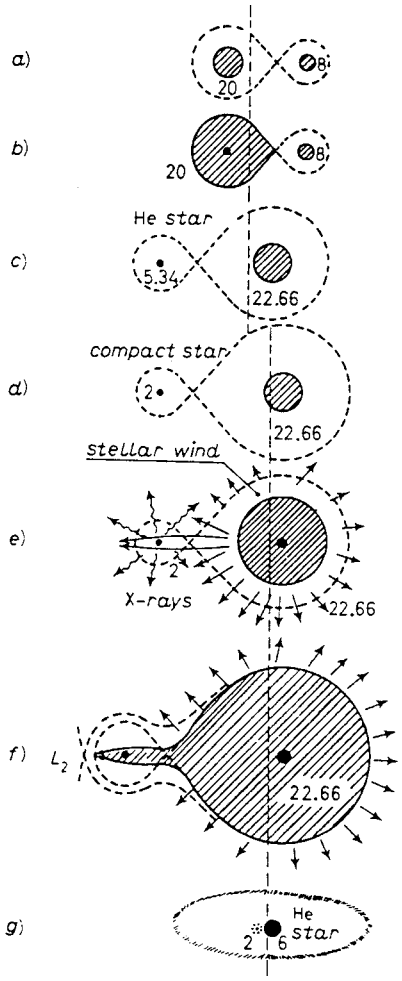


Figure 4. Stages in the evolution of an HMXB. It is assumed that the SN explosion of the primary leaves a $2M_\odot$ neutron star: (*a*) $t = 0$, $P = 4.7$ days; (*b*) $t = 6.17 \times 10^6$ yr, $P = 4.7$ days, onset of RLO; (*c*) $t = 6.20 \times 10^6$ yr, $P = 10.86$ days, end of RLO and beginning of WR phase; (*d*) $t = 6.76 \times 10^6$ yr, $P = 12.63$ days, He star (WR star) has exploded as an SN; (*e*) $t = 10.41 \times 10^6$ yr, $P = 12.63$ days, the normal star becomes a supergiant and its strong stellar wind powers a bright x-ray source; (*f*) $t = 10.45 \times 10^6$ yr, $P = 12.63$ days, onset of second phase of RLO, x-ray source disappears owing to absorption, followed by a spiral-in of NS; (*g*) $t \simeq 10.47 \times 10^6$ yr, $P \simeq 4$ h, onset of second WR stage. Finally, at $t \simeq 11 \times 10^6$ yr, the second He star explodes as an SN; survival or disruption of binary pulsar depends on remnant mass and the possible kick from the SN explosion. (After van den Heuvel.)

the existence of wide binaries consisting of a radio pulsar with a white dwarf companion one expects the existence of LMXBs in which the neutron star is powered by RLO from the expanding (sub)giant progenitor of the white dwarf, a phase during which the binary widens on the nuclear time scale of the (sub)giant. Few of these systems are observed but Cyg X-2 probably belongs to this class. Like in CVs, the observationally inferred mass transfer rates in *compact* LMXBs seem higher than can be driven by

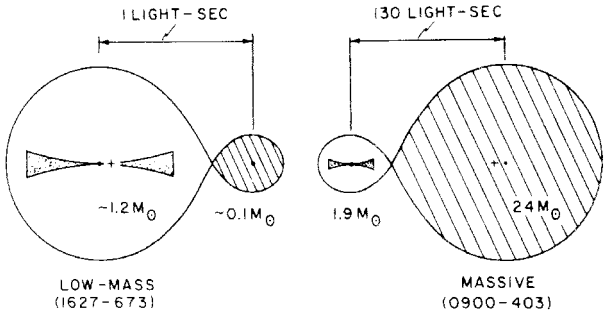


Figure 5. Comparison of the very different dimensions of a low-mass x-ray binary and a high-mass x-ray binary. The optical stars are depicted to fill their Roche lobes (hatched). The radial extent of the accretion disks (stippled) around the neutron stars is sketched nearly to scale. The dots and the crosses (+) mark the centers of mass of the individual stars and the system, respectively. (After Bradt and McClintock.)

the slow nuclear expansion of the low-mass contact star. Apparently the mass transfer is enhanced by the angular momentum loss driven contraction of the binary, like in CVs. It appears there is a *bifurcation* into compact and wide LMXBs: short-period systems, in which magnetic braking and gravitational radiation driven angular momentum losses dominate, evolve to ever smaller size, while systems with periods beyond a critical initial period (of order 1 day), beyond which these angular momentum losses become ineffective, evolve on the nuclear expansion time of the (sub)giant contact star and become very wide systems. A major difference between compact LMXBs and CVs is that accretion onto the neutron star in LMXBs liberates far more energy which may be partly intercepted by the (thus heated) contact star. The accreting matter can spin the neutron star up to millisecond spin periods. It has been suggested that the accreted rotational energy may ultimately *evaporate* the low-mass contact star in compact LMXBs. That would explain the apparent absence of the (orbital) period gap between about 3 and 2 h in LMXBs which is seen in the population of CVs. The idea is that after termination of magnetic braking (when the contact star becomes fully convective) the mass transfer is interrupted, so that the neutron star is suddenly unveiled and becomes a rapidly spinning *radio pulsar* whose high-energy emission may evaporate the companion before it refills its Roche lobe. This evaporation would also explain the observed existence of rapidly spinning single-millisecond radio pulsars.

From the fact that our Galaxy contains a similar number of HMXBs (only 50% less than the number of LMXBs) whereas the lifetime of a HMXB must be orders of magnitude shorter than that of an LMXB, we can deduce that the formation of an LMXB is a much rarer event than the birth of an HMXB. Therefore it is no surprise that the formation of LMXBs is far less straightforward than that of HMXBs. One of the formation scenarios that have been suggested is *accretion-induced collapse* (AIC)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a massive white dwarf, whereby the white dwarf is driven over the *Chandrasekhar limiting mass* and collapses to become a neutron star. This process requires precise fine tuning with mass transfer rates in a narrow range to a rare sort of massive white dwarf (consisting of O–Ne–Mg). Such white dwarfs can only be formed in binaries with a restricted range of primary masses and orbital periods. Mass transfer to more common intermediate-mass carbon–oxygen white dwarfs leads in most cases to explosive carbon burning by which the degenerate star is completely destroyed in a type Ia SN explosion.

Some LMXBs have black hole companions (see BLACK HOLE CANDIDATES IN X-RAY BINARIES) and these cannot be produced by AIC, so that formation channels different from AIC must be active in producing LMXBs. In the special high star density environment of globular clusters LMXBs may be formed when a low-mass main-sequence star is captured by a passing compact star, but this cannot explain the occurrence of LMXBs in the rest of the Galaxy. Spiral-in evolution after an SN explosion of a massive helium star might explain the existence of neutron star or black hole components in compact LMXBs, but only under rather extreme conditions can the post-spiral-in systems evolve into LMXBs.

Bibliography

Reviews on close binary evolution are

Paczynski B 1971 *Ann. Rev. Astron. Astrophys.* **9** 183

Shu F and Lubow S H 1981 *Ann. Rev. Astron. Astrophys.* **19** 277

Thomas H-C 1977 *Ann. Rev. Astron. Astrophys.* **15** 127

Reviews of the evolution of x-ray binary systems and of binary pulsar systems are

Phinney E S and Kulkarni S R 1994 *Ann. Rev. Astron. Astrophys.* **32** 591

van den Heuvel E P J 1992 *Interacting Binaries (Lecture Notes Saas Fee Advanced Course 22)* ed H Nussbaumer and A Orr (Berlin: Springer) (in the same course is a review on cataclysmic variables by M Livio)

Verbunt F 1993 *Ann. Rev. Astron. Astrophys.* **31** 93

A popular account of x-ray binaries can be found in

van den Heuvel E P J and van Paradijs J 1993 *Sci. Am.* (November)

Gerrit J Savonije

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clouds in Planetary Atmospheres

What are clouds? The answer to that question is both obvious and subtle. In the terrestrial atmosphere clouds are familiar as vast collections of small water drops or ice crystals suspended in the air. In the atmospheres of Venus, Mars, Jupiter, Saturn, Saturn's moon Titan, Uranus, Neptune, and possibly Pluto, they are composed of several other substances including sulfuric acid, ammonia, hydrogen sulfide, ammonium hydrosulfide, methane, and complex organic molecules. The study of clouds touches on many facets of atmospheric science. The chemistry of clouds is tied to the chemistry of the surrounding atmosphere. Clouds can be both a source and a sink for condensable molecules as they moisten or dry the air. Local cooling or heating and local differences in moisture content due to cloud formation change the buoyancy of air parcels which drives local dynamics (frontal systems, thunderstorms and hurricanes are a few familiar examples). Clouds influence the global climate by reflecting or absorbing solar radiation while trapping thermal radiation. The fate of life on Earth a few centuries from now may hinge on their uncertain role in mitigating the greenhouse effects of increased emission of carbon dioxide gas from biomass and fossil fuel burning. This article examines what we know about clouds in the atmospheres of the planets.

Cloudy planetary atmospheres

Saturn's largest satellite Titan and all the planets except Mercury have cloudy atmospheres (see PLANETARY ATMOSPHERES). These can be categorized into three main groups. Water is the principal cloud constituent on the relatively thin atmospheres of Earth and Mars. Venus and Titan have thick cloud and haze layers which are influenced strongly by chemical reactions initiated by sunlight in the high atmosphere. The giant planets Jupiter, Saturn, Uranus and Neptune have reducing atmospheres dominated by hydrogen which combines with trace amounts of nitrogen, sulfur and carbon to form clouds of ammonia (NH_3), ammonium hydrosulfide (NH_4SH), hydrogen sulfide (H_2S) and (for Uranus and Neptune) methane (CH_4). Methane clouds were inferred to be present in Pluto's atmosphere although the interpretation of the data is ambiguous.

The Earth

We know most about the nature of clouds in the EARTH'S ATMOSPHERE and so it is instructive to begin a survey of planetary clouds with a review of terrestrial clouds. Water is by far the main constituent of terrestrial clouds but some other constituents (mainly sulfur and nitrogen) are often present. Water can exist as vapor, liquid drops, or frozen ice crystals. The amounts of water in these various forms depends on the complex history of each air parcel as it moves over land or water. Clouds take on characteristic forms which are controlled by the dynamical processes responsible for their formation. By looking at photos of clouds taken from space a sense of the dynamical regime can be revealed.

Visual categorization of cloud types is a science that goes back many hundreds of years, and much of the terminology is Latin. Meteorologists in the present day group clouds into three main groups depending on the altitude (low, middle and high), with sub-categories within each group. In the lowest 2 km altitude are the cumulus, cumulonimbus, stratus, stratocumulus, and nimbostratus. Within the middle layer (2–8 km) are altostratus and altocumulus. At the highest level in the troposphere are cirrus, cirrostratus and cirrocumulus. At still higher altitudes in the stratosphere and at very low concentrations (except immediately after a volcanic eruption) are small sulfuric acid drops at many latitudes and ice particles which form the polar noctilucent clouds (thin clouds seen in the polar dawn and dusk).

Cumulus clouds are relatively small (a few km in diameter), detached, puffy clouds whose tops and sides appear brilliant white in direct sunlight. In this form they would be impossible to see from space unless observed with a nearby camera with high spatial resolution. They can be seen as single clouds, in fields of many, and may develop into large towers of several grouped together. As moisture becomes more available from ground evaporation these clouds can grow and merge to form cumulonimbus which are heavy and dense, with a towering structure. These can be sites of precipitation and lightning. As seen from above they often show a flat top with a tail extending downwind. The morphology of the tail appears similar to a blacksmith's anvil, and the word 'anvil' is often applied to these features. The clouds are classified as low clouds because the base is within the lowest level, but in its entirety the cloud extends through all three levels, with the anvil providing a source of water downstream in the highest level.

Stratus clouds form a fairly flat, uniform layer, dark gray blanket when seen from below. They often produce drizzle or snow. Stratocumulus appear much like stratus except that they show detail (patchiness, ragged edges) and are less dense in places, permitting more sunlight to penetrate and showing a lighter shade of gray. Seen from above, they appear clumpy and sometimes reveal periodic roll structures (cloud streets) which can extend for many kilometers. Motion within these clouds can occasionally produce cumulus-type structures. Large (~1000 km) banks of stratus and stratocumulus clouds can be seen in satellite photographs. Nimbostratus can be extremely deep with a top nearly at the top of the troposphere and usually produce precipitation.

Altostratus clouds are distinguished from stratus mainly by the location of their base which is higher in the troposphere. A colored corona around the Sun can sometimes be seen. This is a consequence of the way ice crystals scatter light. Altocumulus also have their base higher in the atmosphere compared to cumulus. These clouds are usually thin and may take on a multitude of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

forms from thin and flat to detached clumps to castle-shaped morphology.

Clouds in the high troposphere are composed almost entirely of ice particles. By analogy with the lower clouds, the clouds which are confined to the highest layers (cirrus, cirrostratus, and cirrocumulus) can range from broad flat sheets to clumpy, distinct elements. There are many sub-categories of cirrus which characterize the many ways these clouds can appear (smooth, stranded, with hooks, in cells, grainy or undular, resembling skeletal structures, etc). All of this morphological richness is rooted in the equally rich dynamic wind field and a complex interplay between the supply of vapor or crystals, and cloud microphysical processes (nucleation, condensation, growth, evaporation, aggregation, precipitation). Much of the cirrus owes its origin to water supplied by high altitude entrainment of air from upstream cumulonimbus anvils and nimbostratus cloud tops. Atmospheric halos and coronae are often seen within thin cirrus clouds, a result of sunlight scattered by large water ice crystals.

How do we understand the complexity of the different cloud types and what do they tell us about the processes? This is an area of active research but already a great deal is known. The existence of a cloud points to a process where an air parcel containing water vapor (say, from evaporation over the ocean or over moist land surface) was transported to a location where the temperature is cooler—cool enough that the vapor pressure of water in the air is higher than the condensation vapor pressure at the air temperature. In the lowest 12 km or so altitude (the troposphere) the temperature decreases with altitude, so a parcel of moist air which buoyantly rises from the ground will eventually reach a level where the temperature is cool enough to support condensation and cloud formation. Low-lying clouds have so much moisture that they need not rise very far for the temperature to reach the condensation temperature. Another way to cool the air and promote cloud formation is to mix warm moist air with cool dryer air as a polar cold front moves through. A third way is to lift air as it flows over a mountain. A cloud forms above or nearby a mountain when the air flowing over cools as it reaches a higher altitude. Clouds formed in this way are called orographic clouds because they remain fixed next to a topographic feature as the air flows past. The same process operates in the Martian atmosphere near the tops of the high volcanos and in Neptune's atmosphere above a large vortex (the Great Dark Spot).

Cumulus cloud formation is initiated by natural convection as sunlight warms a moist ground. As a warm, moist buoyant parcel of gas rises it reaches a condensation level where the temperature drops below the condensation temperature and the cumulus cloud forms at its base. An aid to convection in the terrestrial atmosphere is the density difference due to both the temperature difference (a warm parcel is less dense) and a compositional difference (a moist parcel is less dense than dry gas for the Earth's atmosphere). In the giant planet atmospheres a moist parcel is more dense than dry gas which inhibits rather than promotes moist convection. A third factor which enhances moist convection is the additional heat released as latent heat of condensation when the vapor condenses in the cloud. In a quiescent situation where cumulus or stratus clouds form and dissipate at an equal rate there is an equilibrium between the supply of moist air from the ground and evaporation and precipitation in the cloud. If there is a great deal of moisture available from the ground and if the initial convective motion is quite vigorous an instability can grow, with the buoyancy carried by the moist convective parcels amplifying to the point that very large amounts of moisture are injected at very high altitudes (cumulonimbus towers and anvils). Stratus clouds result from a more sluggish dynamical regime where the air is more stable and where there is a stable and abundant supply of moisture to maintain the cloud against precipitation and evaporation.

Space-based views of clouds on all the planets are dominated by the large-scale features. At large scales three types of cloud features are commonly seen in the Earth's atmosphere. Each of them has an easily recognizable morphology and motion which are rooted to the dynamical regime forcing cloud formation. Mesoscale convective systems consist of fields of cumulonimbus clouds whose anvils merge into a single cloud shield with a horizontal extent much larger than the individual anvils. Mesoscale convective systems may be long (~100 km) and narrow (a few km) with extensive precipitation occurring over a large scale. *Cumulus convection* is driven by small-scale (100 km or less) convective cells. Small-scale cells can combine to form the much larger mesoscale convective systems seen in satellite images.

Of special importance for the dynamics of the terrestrial atmosphere is the organized deep convection near the equator. Time-averaged maps of cloudiness reveal a band near the equator that moves in latitude with the seasons. This is the Inter Tropical Convergence Zone (ITCZ). Its existence is prompted by converging air mass at low altitude driven by large-scale Hadley circulation in both hemispheres. Abundant water vapor near the surface is available for a massive amount of latent heat transport. Moist convection within the ITCZ is vigorous and is responsible for a large amount of heat and moisture transferred to the upper atmosphere within the ITCZ.

Atlantic hurricanes and Pacific typhoons are tropical cyclones which constitute another commonly observed large-scale cloud structure. They form over oceans and derive their energy from moist warm tropical ocean air. Cyclones rotate in the counterclockwise direction in the northern hemisphere with lifetimes of the order of 10 d. Hurricanes and typhoons have return flow in a narrow column at the core which is free of clouds. There remains some controversy about what special conditions are required to transform commonly occurring tropical cyclones into intense systems that occur relatively rarely. A common thread to the various models involves a very effective coupling between cumulus-scale moist

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. The Solid State Imaging camera on the Galileo Orbiter captured this image of clouds over the southern Pacific ocean and Antarctica on 12 December, 1990 when the spacecraft was 1.6 million km from Earth. Extensive cloud formation can be seen, with prominent curved cloudy frontal systems associated with extratropical cyclones. The dark ocean provides the contrast which makes the clouds easy to see in visible-wavelength images. Clouds and surface ice are difficult to distinguish due to lack of contrast. In broad-band visible wavelengths there is no way to get a sense of cloud altitude, and the cloud type categories developed from ground-based experience are not apparent.

convection with the large-scale wind field via an instability called the conditional instability of the second kind, or CISK. According to most models the ocean surface temperature must be at least 26 °C, which agrees with observation and leads to the conclusion that these intense storms will be more frequent and/or more intense should the sea surface warm in the future from global warming.

A third commonly observed large-scale cloud structure is the extratropical cyclone which produces extensive precipitation in the middle latitudes. Several of them can be seen in figure 1. They are familiar as weather fronts ahead of cold low-pressure regions . The clouds that form at these fronts do so by *sloping convection* (as opposed to cumulus convection described above). Sloping convection occurs when one body of air overrides another, forcing moist air upward. Clouds of all types are associated with these large-scale features and also with the two others mentioned above.

With that brief introduction to terrestrial clouds we are now in a position to explore clouds on other planets. The best place to begin is with the atmosphere of Mars which contains water ice clouds and displays many of the dynamical features familiar to terrestrial meteorology (see MARS: ATMOSPHERE).

Mars

Our knowledge of the Martian atmosphere is based on images and spectra taken over many years from ground-based observatories as well as the Hubble Space Telescope, several Soviet and US flyby or orbiting satellites, and landers, most notably Mariners 4, 6, 7 and 9, Mars 2, 3, 4 and 5, Viking 1 and 2, Phobos 1 and 2 and Mars Global Surveyor, with more on the way or planned. As of the beginning of 1999 three craft (VIKING Landers 1 and 2 and MARS PATHFINDER) have landed on the surface and sent back *in situ* information on atmospheric composition, temperature, pressure and cloud and dust opacity. Many more landers are planned for the next decades.

The Martian atmosphere is made up almost entirely of carbon dioxide (CO_2) with trace amounts of water, argon and several other constituents. Mean surface pressure is 8 mb (compared to a pressure near 1 b at sea level on Earth). Surface temperature varies between about 140 K and 250 K depending on latitude and season. Carbon dioxide condenses to form an extensive frost layer on the surface of Mars at high latitude when the surface temperature drops slightly below 150 K. Occasionally the temperature in the polar atmosphere several km above the surface drops low enough to form a CO_2 ice haze, but almost all the bright white clouds seen on Mars are composed of water ice.

Water ice also condenses at high latitude during winter, and at lower latitudes as well during the cold night. The polar ice caps and the Martian soil serve as reservoirs of water which absorb or release water on diurnal and seasonal cycles as well as on much longer-term cycles tied to slow changes in the Martian orbit and tilt. Outflow channels on the surface are now dry but point to a much wetter Mars in the distant past. The present amount of water in the atmosphere amounts to only about 10 precipitable micrometers on average, too little to be important as a source of buoyancy via latent heat release. Thus none of the terrestrial cloud types (cumulonimbus or stratonimbus on the small scale, hurricanes and mesoscale convective systems) which feed on latent heat release and an abundant supply of water are seen on Mars.

The most fruitful approach to understanding water clouds on Mars is to note their similarity to cirrus clouds in the high, cold terrestrial atmosphere, with the addition of the effects of topography operative in the Mars atmosphere. Lee wave clouds are often seen on Mars downstream of high volcanic mountains and near craters at lower elevations. It is common to see lee wave clouds with many linear undulations extending for more than 100 km downstream. Clouds form most often in the early morning and late afternoon when the temperatures are cool. They are abundant in northern summer in the Tharsis region, a high plateau home to several of the highest Martian volcanos. Figure 2 shows this, along with large-scale features that appear to trace planetary-scale Rossby waves at high latitude much as do extratropical cyclone

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

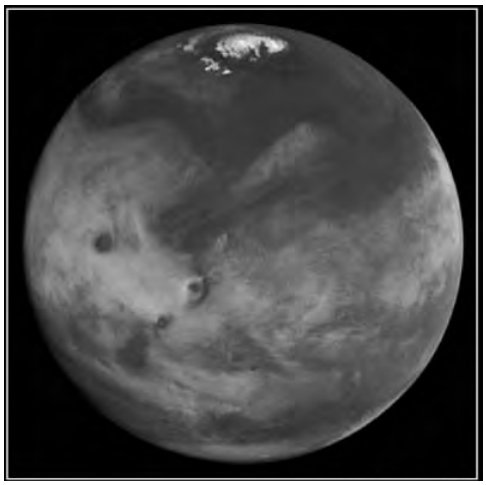


Figure 2. The surface of Mars is dark at blue wavelengths and produces a good background for showing bright water ice clouds and north polar frost deposits seen in this image taken by the Hubble Space Telescope Wide Field and Planetary Camera 2 on 20 May, 1997. Features as small as 45 km can be seen. Ice clouds cover most of the Martian mid-latitudes and are especially prominent in the Tharsis Montes region to the left of the center of the planet. Four of the highest volcanic peaks can be seen poking though the clouds. The curved cloud contours in the upper left are suggestive of planetary-scale waves similar to those which drive large extratropical cyclones in the terrestrial atmosphere. (Photo courtesy of P. James, T. Clancy and S. Lee.)

cloud bands on the Earth. At night the surface can radiate to space efficiently and cool dramatically, leading to the formation of low-lying ground fog which dissipates soon after sunrise.

Although not classified as cloud, dust in the Martian atmosphere plays an important role in atmospheric dynamics and cloud formation. The Martian atmosphere usually contains enough dust to be a significant absorber of solar energy, heating the atmosphere during the day and promoting radiative cooling at night. Strong thermal contrasts at the edge of the polar frost deposit, coupled with the low density of the Martian atmosphere, can lead to high surface winds which pick up dust, sometimes progressing to a global dust storm that can blanket the planet, obscuring even the highest volcanos. After several months most of the dust sediments out and the atmosphere gradually returns to a cooler state more conducive to the formation of water ice clouds.

Venus

Venus and the Earth have similar size and mass but their atmospheres and cloud structures differ markedly (see VENUS: ATMOSPHERE). The Venus atmosphere is composed primarily of CO_2. It is massive (90 bars at the surface) and hot (740 K) near the surface. Because of its great fluid mass, slow rotation and greenhouse effect there is little horizontal thermal contrast (at most a few K from equator to pole) and consequently the dynamical regime is quite different than on the Earth or Mars.

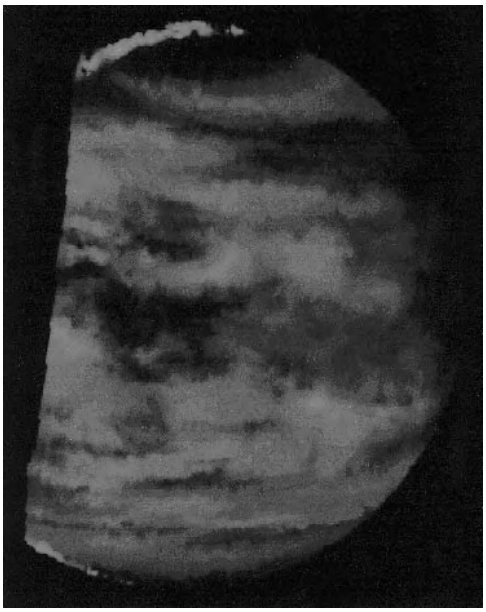


Figure 3. This image of the Venus nightside taken by the Galileo Near Infrared Mapping Spectrometer instrument at a wavelength of 2.3 microns in the near-infrared is bright where infrared light emitted by the hot surface is less attenuated by overlying clouds which block the radiation in the darker regions. The contrasts are produced mostly by variations in the number of large particles in the deepest of the three main cloud layers. These variations are due to atmospheric motions which are most vigorous in the lower cloud region.

Venus as seen through a telescope appears bright and almost featureless due to a ubiquitous thick layer of cloud. Subtle contrasts can be seen at violet and near-ultraviolet wavelengths. The material responsible for the contrast has not been identified with confidence but suggested candidates include elemental sulfur and the compound $FeCl_3$. Sulfur dioxide gas (SO_2) has been detected at the cloud tops, and the visible clouds are composed of liquid sulfuric acid (H_2SO_4) drops. At near-infrared wavelengths there are several spectral intervals (windows) where the gas opacity is relatively low, permitting infrared light radiated by the hot surface to penetrate the clouds. There is considerable cloud opacity even at those wavelengths however, so the contrasts seen in the near-infrared tell us mostly about the patchiness of the lower cloud layers which are more turbulent than the upper layer we see at visible wavelengths. These contrasts can be seen in figure 3.

Many spacecraft from the United States and the former Soviet Union have visited Venus. Several probes have descended into the atmosphere and measured pressure, temperature, gas composition, the internal radiation field and cloud particle number densities and sizes. The clouds extend over a very large vertical range. There are three main cloud layers at altitudes between about 45 and 70 km. Thin haze layers extend above and below those boundaries. Several experiments revealed the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

main cloud region to be composed of three layers. The top layer is concentrated between 59–66 km altitude. The middle layer lies between 49–57 km altitude and the lower layer is most dense in a 2 km region just below the 49 km boundary, at least as measured by the particle counter data on the Pioneer Venus large probe. Ground-based and Galileo near-infrared images (from the Near-Infrared Mapping Spectrometer—NIMS) later showed this layer to be highly variable in space and time (see figure 3).

The composition of the cloud particles is dominated by sulfuric acid, but some evidence suggests that an additional constituent may be responsible for the largest particles in the lower cloud layer. In the top cloud and haze sunlight breaks up SO_2 molecules which combine with water to form sulfuric acid with a concentration near 85% at the top, increasing at deeper levels. In the middle and lower clouds there is insufficient ultraviolet light to drive photochemistry and thermochemical processes become important in establishing chemical equilibrium. At the deepest levels the atmosphere and surface interact and the surface acts as a source and sink of sulfur and other constituents. The particle spectrometer on the Pioneer Venus large probe identified three sizes of particles. Mode 2 particles have radius close to 1 micron and dominate the top and middle clouds. Mode 1 particles are smaller (0.4 microns mean radius) and account for a few tenths of the optical depth in the upper cloud while mode 3 particles are larger (several microns or more) and provide much of the mass in the lower cloud.

The dynamical regime under which these clouds form is mostly stable in the upper two layers (altitudes above 49 km). These clouds are analogous to stratiform clouds in the terrestrial atmosphere. The subdued contrasts seen in the highest cloud move with the wind speed (about 100 m s^{-1}) at the visible cloud tops (near 70 km altitude) while the contrasts seen in the near-infrared window regions (e.g. figure 3) move at a slower rate (50 m s^{-1}, consistent with wind speeds measured by tracking entry probes). Estimates of the number of large particles in the lower cloud deck made by various entry probes varied by large factors. This observation supports the idea that the contrasts seen at near-infrared wavelengths are indicative of highly variable concentrations of large (several microns radius or larger) particles in the lower cloud deck. The lower cloud is a region of dynamical instability in the Venus atmosphere. The mechanism responsible for driving the Venus circulation remains unclear, although it seems likely that interactions between waves and the mean flow, initiated by absorption of sunlight in the cloud region, play a central role.

Titan

Like Venus, TITAN's atmosphere is a massive one with a dense haze layer and slow rotation and may be driven by the same mechanism. But the chemistry of the Titanian and Venusian atmospheres could hardly be more different. Titan's atmosphere is composed mainly of nitrogen with a few percent of methane and other hydrocarbons and nitriles (molecules with H, C, and N). Titan's atmosphere is cold (about 94 K near the surface and 71 K at the temperature minimum near 40 km altitude). At those temperatures methane may condense near the temperature minimum but some infrared measurements from the Voyager IRIS experiment suggest that a methane cloud may not be present. If that is the case methane would be supersaturated near the temperature minimum. Supersaturation does not occur in the terrestrial atmosphere (with respect to water) but can be achieved in the laboratory under controlled conditions. It will be difficult to resolve this issue until the Huygen's Probe descends into Titan's atmosphere in 2004.

Photolysis of methane in the high atmosphere leads to formation of ethane and other hydrocarbon and nitrile molecules. Processing of these molecules by long-term exposure to sunlight and charged-particle bombardment from space eventually produces a polymer haze which covers the satellite and blocks the surface from view at visible wavelengths. This haze is very dark at ultraviolet wavelengths and becomes bright, with little absorption at red wavelengths. It is composed of aggregates of small particles whose mean radius is close to 0.066 microns.

Chemical models predict a number of haze and cloud layers deeper in the atmosphere, composed of more volatile hydrocarbons. Methane is the most volatile and may condense below the temperature minimum (altitude range 10–40 km). A little higher in the atmosphere ethane may form a condensation cloud. Less volatile hydrocarbons and nitriles are expected to form condensate layers at higher altitudes, including C_2H_2, C_3H_4, C_3H_8, C_4H_2, HCN, C_2N_2, HC_3N, C_4N_2 and many heavier hydrocarbons. Carbon dioxide and water ice particles are also possible. Of these, only HC_3N and C_4N_2 particles have been detected spectroscopically. Detection of many of them is made difficult by Titan's obscuring photochemical haze at higher altitude. The spectral signature of C_4N_2 (dicyanoacetylene) ice was seen by the Voyager infrared spectrometer only in the polar haze of Titan. Analysis of these observations reveals that (1) ice particle mean radius is near 5 μm, and (2) the ratio of condensate to vapor mole fraction as inferred from the data is about 100 times more than expected from steady-state equilibrium of the constituents, accounting for creation of the vapor at high altitudes from photochemical production, and depletion of the vapor and condensate at low altitudes due to condensation and sedimentation of particles. The large disequilibrium ratio is probably a result of seasonal variation of photochemical destruction of vapor at high altitude and high latitude working in conjunction with a significant thermal time lag deeper in the atmosphere due to long thermal time constants.

There is almost no contrast on Titan at visible wavelengths apart from a hemispheric north/south reflectivity contrast which reverses over a several-year period as Titan responds to seasonal changes in insolation. There is a corresponding hemispheric difference in particle size in the upper haze which in turn is related to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

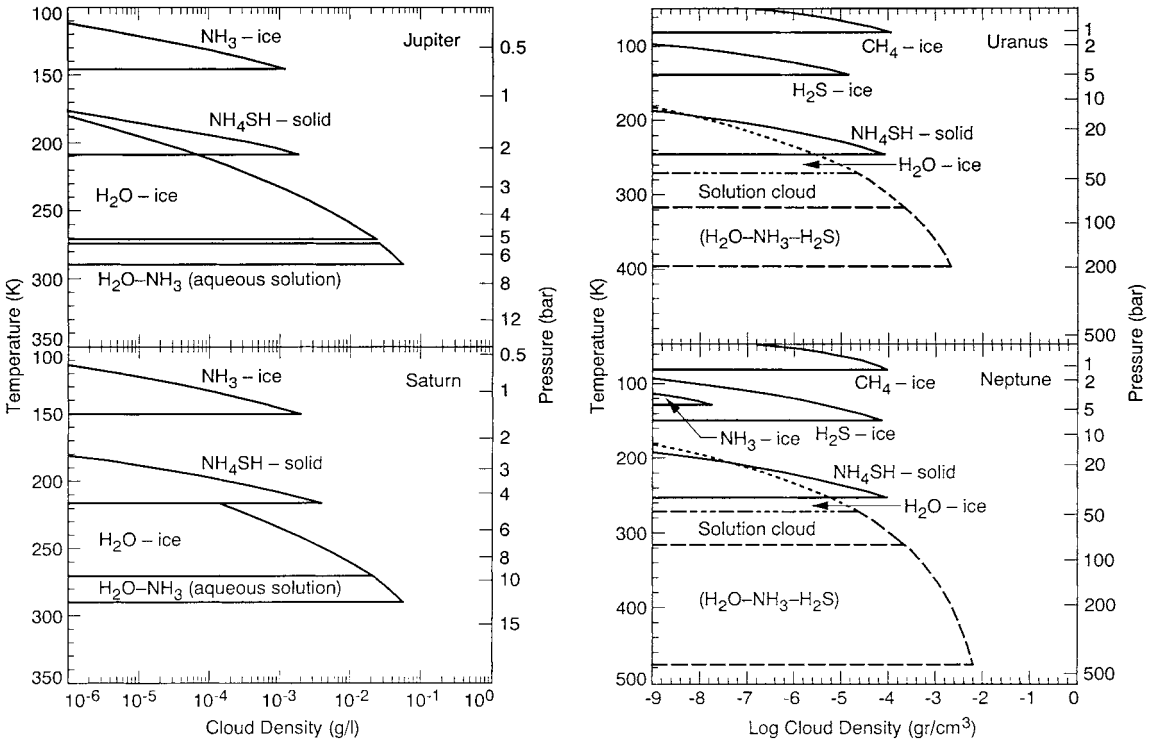


Figure 4. Thermochemical equilibrium models for the giant planet atmospheres indicate the vertical regions and compositions of ices expected to form clouds. Within these vertical regions actual cloud locations and densities are determined largely by atmospheric dynamics and microphysical processes. (Figures for Jupiter and Saturn were constructed from models by S K Atreya and M Wong, based on S K Atreya and P N Romani (1985) *Planetary Meteorology* ed G E Hunt (Cambridge: Cambridge University Press) pp 17–68. Those for Uranus and Neptune were first published by I de Pater *et al* (1991) *Icarus* **91** 220–233, © Academic Press.)

a hemispheric-scale rising and sinking motion which responds to seasonal change. Atmospheric aerosols contribute significantly to the heat balance in the stratosphere through their strong absorption of sunlight and therefore are important to the energy balance and circulation.

The giant planets

All of the giant planets Jupiter, Saturn, Uranus and Neptune have deep atmospheres composed mainly of hydrogen and helium with trace amounts of cloud-forming molecules water, ammonia (NH_3), ammonium hydrosulfide (NH_4SH), hydrogen sulfide (H_2S) and methane (CH_4). All of these except methane can condense in the atmospheres of Jupiter and Saturn. The atmospheres of Uranus and Neptune reach sufficiently cold temperatures near the temperature minimum at 0.1 b pressure to condense methane. Condensate clouds form in layers with the most volatile forming at higher altitudes where the temperatures are cooler.

Our ideas about composition are based on a combination of observations from the ground and spacecraft together with models of chemical reactions at the temperature and composition of the atmosphere. So far all of our measurements, except those from the Galileo probe which entered Jupiter's atmosphere (see GALILEO MISSION TO JUPITER), are made remotely with instruments from the Earth or on spacecraft (see JUPITER: ATMOSPHERE). The Galileo probe sampled a part of Jupiter's atmosphere that is thought to be depleted in condensable molecules, much like a desert region on the Earth, and so the probe measurements do not tell us about the global properties of clouds and cloud-forming constituents.

As a starting point in understanding the vertical locations and compositions of clouds we rely on thermochemical equilibrium models which make predictions from measured or inferred chemical abundance in the deep atmosphere, the pressure and temperature as measured by a variety of instruments, and knowledge of the chemical thermodynamics of the available molecules. This exercise can tell us at what altitudes to expect clouds and what their composition should be but it is not adequate for predicting the number densities of cloud particles at a given location because other processes such as atmospheric dynamics and cloud particle microphysics determine those quantities.

The cloud forming molecules mentioned above are observe to be or are predicted to be a factor of two or more abundant than they would be for a mix of solar atmosphere material at the same temperature and pressure. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

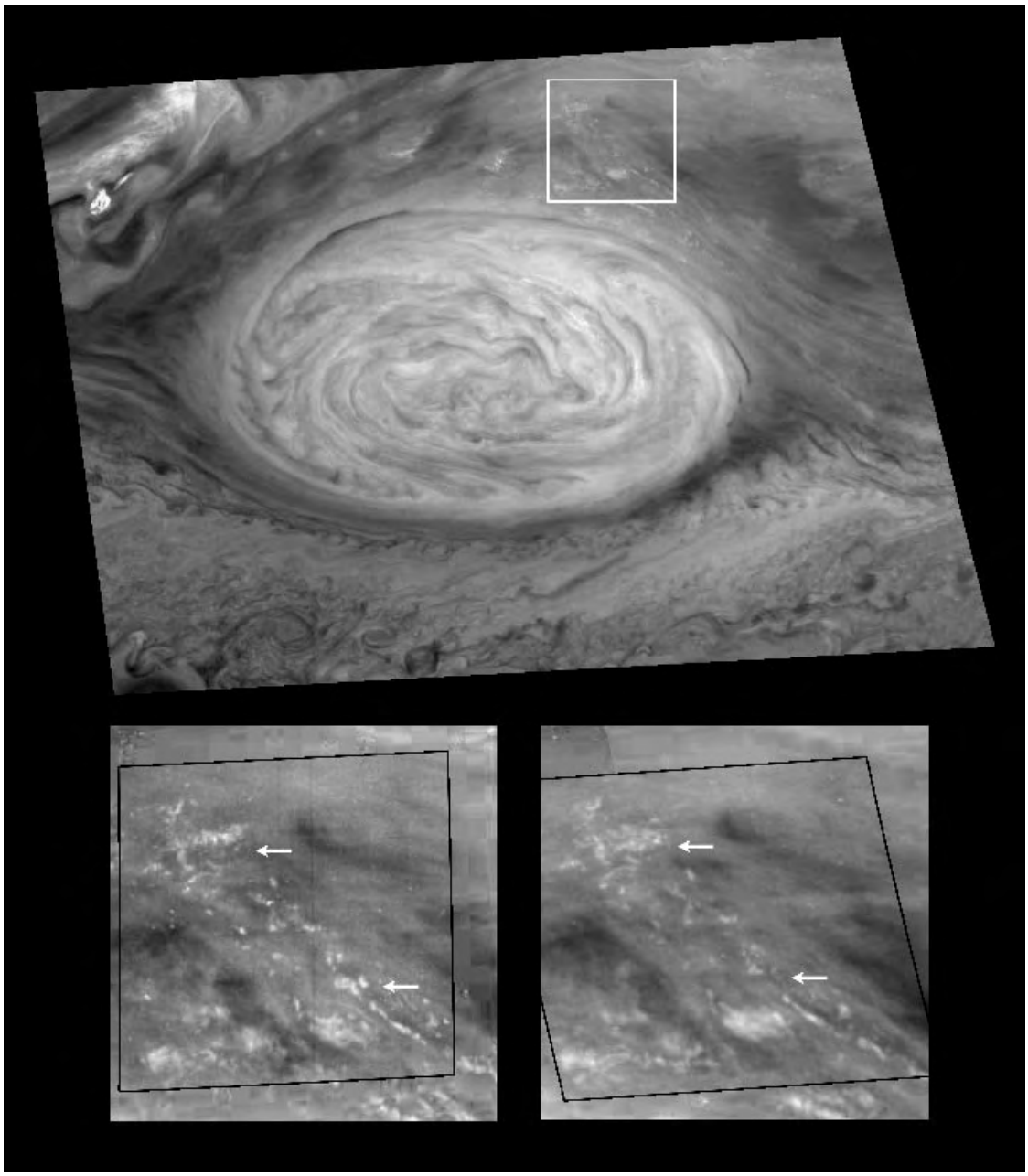


Figure 5. Jupiter's Great Red Spot and environs as seen by the Galileo Solid State Imager in 1996. The Great Red Spot centered in the top panel rotates counterclockwise. The embedded images shown in the two bottom panels reveal small puffy clouds about 50 km in diameter, possibly analogous to large terrestrial cumulus clouds. The time difference between the two bottom panels is 70 min.

enhancement of 'solar' abundance increases from Jupiter to Saturn to Uranus and Neptune, reflecting the fact that the larger planets were able to retain more of the light elements hydrogen and helium during their formation. An increase in the amount of available condensable molecule will result in cloud formation at a deeper level where the temperatures are warmer.

Predicted cloud locations for the four giant planets are shown in figure 4. JUPITER and SATURN form one pair with similar cloud types, while URANUS and NEPTUNE, which are colder, form a second. A water–ammonia solution cloud is predicted to form at pressure levels starting near 8 b in Jupiter's atmosphere and 15 b in Saturn's atmosphere. At higher altitudes (lower pressure and temperature) are clouds of ammonium hydrosulfide and ammonia ices. The topmost cloud in both cases is ammonia ice. A spectral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

signature for ammonia ice was mysteriously not seen until recently in near-infrared spectra from the European Infrared Space Observatory (ISO) and the Galileo NIMS instrument. Spectral signatures of ices in the other clouds have not been observed. One of the Galileo Probe experiments detected particles near the 1.6 b level in Jupiter's atmosphere whose composition could be either ammonium hydrosulfide or water. It did not see a water cloud at the pressure levels where thermochemical equilibrium models predict, probably because water in the region it sampled is depleted due to a downwelling of dry air from the cold upper troposphere. It is not known how extensive a deep water cloud may be on any of the giant planets.

Uranus and Neptune are cold enough to condense methane which forms the top cloud in those atmospheres (see URANUS AND NEPTUNE: ATMOSPHERES, IONOSPHERES AND MAGNETOSPHERES). The next-deeper cloud, at a few bars pressure, is probably H_2S ice but may contain some ammonia as well. At deeper levels (10 to a few hundred bar pressure) are NH_4SH and water clouds.

The visibility of clouds varies markedly among the giant planets. Jupiter's atmosphere is filled with white-yellow clouds and darker yellow-brown clouds with a few regions that are more red (the GREAT RED SPOT) or gray (the hot-spot regions where upwelling infrared radiation at 5 microns wavelength can be seen). Jupiter's Great Red Spot is a large (about 3 times the size of the Earth) oval rotating in the anticyclonic sense (opposite to terrestrial storm systems). It has been observed for more than a century and probably owes its stability to the shear of the east–west zonal jets in which it is embedded. It is shown in figure 5.

All of the ices mentioned above are white at visible wavelengths. If these were the only cloud constituents Jupiter would appear as a featureless white ball. The colors and contrasts are due to a coloring material whose origin and composition is uncertain. Candidate materials include compounds of hydrogen bonded to sulfur, nitrogen and carbon (hydrocarbons) created by photolysis in the upper troposphere or by bombardment from high-energy charged particles precipitating from the magnetosphere. The latter process is almost certainly responsible for the formation of polar stratospheric hazes in the stratospheres of Jupiter and Saturn.

Cloud contrasts are subdued for Saturn and almost absent for Uranus. These planets as well as Neptune have deep cloud structures which reduce contrast and block upwelling thermal radiation. Contrasts are dominated by the jet structures, although there is not a one-to-one correspondence between cloud reflectivity and wind speed and direction. Saturn shows seasonal variations in cloud altitude. These result from solar heating of the upper atmosphere which influences the static stability and cloud altitude. These effects may well be occurring as well on Uranus and Neptune but seasonal changes for those planets occur on timescales long compared to our ability to observe them. Individual ovals and spots do form in the Saturn atmosphere. They can be seen most easily in contrast-enhanced images (figure 6).

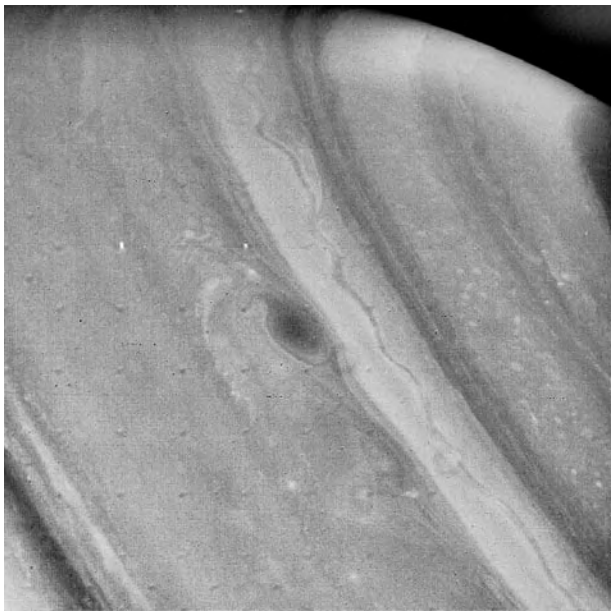


Figure 6. A dark oval and many smaller features can be seen embedded in the bright-dark cloud bands which line up with the jet streams in Saturn's atmosphere. A dark, kinked curve called the ribbon feature is centered in the bright band at 47° north latitude, the location of a high speed (150 m s^{-1}) jet. Cloud contrasts serve as markers of the wind and of the jet instabilities. This image was obtained by the Voyager 2 Imaging Subsystem in 1981.

The VOYAGER spacecraft viewed Uranus at a time when only one hemisphere was illuminated by the Sun. Only a few low-contrast cloud features were seen against an almost uniform thick cloud cover. More recently the near-infrared camera on the Hubble Space Telescope photographed a number of clouds in the opposite hemisphere (figure 7).

A number of cloud types were seen during the close Voyager flyby of Neptune in 1989. Several large ovals were seen. The largest of them was named the Great Dark Spot (figure 8). Like Jupiter, some coloring constituent of uncertain origin and composition is responsible for cloud contrast and color. These ovals move around and oscillate in longitude. The aspect ratio and tilt of the long and short axes of the ovals also oscillate. These oscillations can be understood with models that take into account the vorticity of the spot, its depth in the atmosphere, and the wind shear in which it is embedded.

Other types of clouds were seen in the Neptune atmosphere which are reminiscent of cirrus clouds and lee-wave clouds seen high in the terrestrial and Martian atmospheres. They are striated, they evolve rapidly with time (too rapidly to use as tracers of wind) and some of them form clumps that are tied to and overlay the larger ovals. The White Companion to the Great Dark Spot is the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

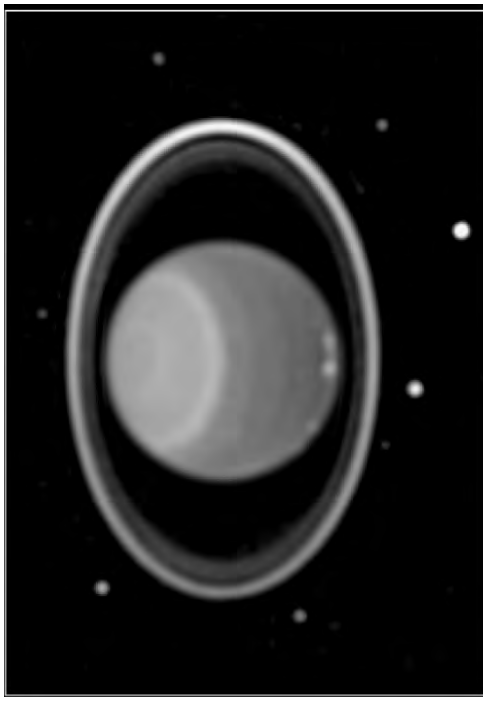


Figure 7. Near-infrared images obtained by the NICMOS camera on the Hubble Space Telescope provide the best views of clouds in the Uranus atmosphere. The entire planet is covered by clouds but the south polar region is covered by a relatively high, thick cloud layer and appears bright in this view. Small clouds, newly discovered in this image taken in 1997, appear in northern low latitudes, just above the terminator. The rings and 8 of the uranian satellites are also visible.

prime example of this. It was seen even in ground-based images in one of the methane absorption bands sensitive to high clouds. It can be seen in figure 8. These clouds are probably composed of methane ice, with the underlying thick cloud bank being H_2S ice.

Cloud microphysics

Cloud microphysics is the study of processes that govern cloud particle formation and evolution on the microscopic scale. It addresses such questions as 'What conditions are required for clouds to form? How rapidly do particles grow? What controls the particle size distribution?' The study of cloud microphysical processes goes beyond the microscopic. A broader goal is to understand the underlying processes which govern cloud structure and dynamics over large scales. A considerable amount of work has been invested in answering these questions for terrestrial clouds where *in situ* measurements can be made and coupled with laboratory and theoretical investigations. There has been some attempt to understand the microphysics of clouds on the other planets, although the observational constraints are much weaker.

Condensate particles can form either by direct condensation from supersaturated vapor or by condensing on a pre-existing foreign particle (dust or sulfate aerosol, for example). These are called homogeneous and heterogeneous nucleation, respectively. Homogeneous nucleation begins when molecules in the vapor phase combine to form a small cluster which can serve as a seed nucleus. The cluster molecules may evaporate, or more molecules from the vapor may condense, depending on which outcome is favored by thermodynamics.

The Gibbs free energy is the thermodynamic quantity which controls particle growth or evaporation. Gibbs free energy depends on two terms, particle surface area times surface tension plus particle volume times KT times the natural log of the vapor pressure divided by the saturation vapor pressure, or

$$\Delta G = 4\pi a^2\sigma - \tfrac{4}{3}\pi a^3 n_l K T \ln(e/e_l).$$

In the above equation ΔG is the change in Gibbs free energy associated with the formation of a particle of radius a, σ is the surface tension of the liquid, n_l is the number density of molecules in the liquid, K is the Boltzmann constant, e is the vapor pressure and e_l is the saturation vapor pressure at temperature T.

A plot of ΔG as a function of radius a reveals that the Gibbs free energy for particle formation is always positive for vapor pressure less than or equal to the saturation vapor pressure. This is thermodynamically unfavorable and any particle which may exist initially will evaporate. When the vapor pressure is higher than the saturation vapor pressure the Gibbs free energy becomes negative at some particle radius. There is a critical radius, a_c, where condensation onto the particle surface reduces the Gibbs free energy and the particle will grow by condensation.

Under normal circumstances in the terrestrial atmosphere it is extremely difficult to grow a particle up to the critical radius because the ratio of vapor pressure to saturation vapor pressure is never very high. A much more common initiation of the growth process depends on the existence of cloud condensation nucei (CCN) which may be composed of small dust or sulfate particles on Earth. A wide variety of CCN particle compositions may be present, but only a small fraction of them are conducive to growth. Particles whose surface structure is similar to that of water or ice (such as silver iodide crystals used for cloud seeding) make good CCN. Another class of particles which make even better CCN are hygroscopic sulfates and nitrates. An abundance of small CCN leads to an abundance of small droplets. Conversely, in an environment with few CCN, few water droplets will form, but the ones that do will take advantage of the available vapor to grow to relatively large size. Images of the ocean surface under some conditions reveal ship tracks as trails of small cloud particles whose formation was initiated by an abundance of CCN in the exhaust plumes of the ships.

Condensation alone is insufficient to grow particles in clouds to the point of precipitation over the coarse of a few hours. Another mechanism, growth by collision and merging of drops (coalescence) or sticking of ice

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

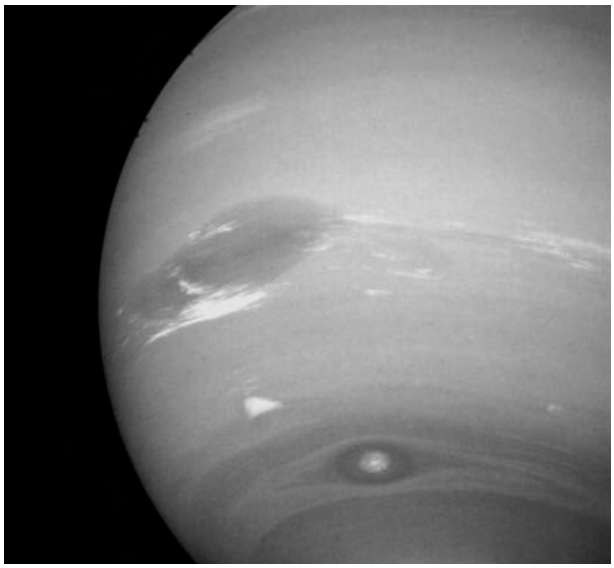


Figure 8. Neptune's Great Dark Spot (GDS) is the large dark oval to the left of center in this 1989 image by the Voyager Imaging Subsystem camera. The Bright Companion cloud forms at higher altitudes at the southern edge of the GDS. Several other bright clouds downstream (to the right) of the GDS and another large oval are also visible.

particles (coagulation), produces much more rapid growth when the particles become large enough (about 20 μm radius) that sedimentation velocities become important. The largest particles sediment faster than the smaller ones, and collide with other particles as they descend through the cloud. This behavior leads to a process which can rapidly grow large (up to about 2 mm) particles. Even larger particles can form in updrafts or parcels which cycle between updrafts and downdrafts.

These ideas introduce the interplay between cloud microphysics and vertical and horizontal winds. Another coupling is between radiation and clouds. At the top of the cloud layer the cloud cools by emitting thermal radiation to space. This cooling in turn feeds a circulation pattern with sinking cool air replaced by rising warmer air from below and within the cloud. Thus the three major physical processes (cloud microphysics, atmospheric dynamics, and radiation) are all coupled, and this coupling must be taken into account in cloud formation/evolution models.

Cirrus clouds form near the top of the troposphere where the temperatures are cold (some 40 °C lower than the freezing temperature of water). Clouds are supplied with vapor from the tops of large convective systems carried horizontally by the wind shear. Both liquid and ice can exist, with the liquid supercooled by more than 30 °C. Ice crystals can grow to relatively large sizes (on the order of 100 μm) as they descend. Ice crystals forming at a variety of temperatures produce a wide variety of crystal habits, but the dominant forms are hexagonal. Typical crystal angles are 120° and 90°. These lead to halo features, the most common at a scattering angle of 22°. Flat plates tend to orient in one direction as they fall due to their aerodynamical profile. Particle orientation leads to certain types of halos.

Microphysical models of Venus, Mars, Titan, and the giant planets

Venus

Our knowledge of the Venus cloud structure is considerable, thanks to extensive measurements from multiple instruments on several probes which descended though the atmosphere, as well as remote spacecraft and ground-based instruments. Models of the cloud microphysics must take into account the coupled chemical composition of the particles and gas (both change with altitude), photochemical processes at the top of the atmosphere, the temperature profile of the atmosphere, and vertical transport due to eddy motions and particle sedimentation.

Microphysical models attempt to reproduce the observed particle chemistry (sulfuric acid/water solution whose concentration varies with altitude as determined by temperature) and the size distribution and number density of the particles which also vary with altitude. Recalling that there are three principal cloud layers in the Venus atmosphere, the topmost is governed by photochemical processes which act on sulfur dioxide and water to produce a sulfuric acid cloud. The top cloud is close to steady state with little temporal variation. Particle growth matches loss by sedimentation. Particle growth is governed by the available sulfate produced by photochemistry. Sulfur dioxide is supplied by vertical transport from below the clouds. A bi-modal size distribution is observed in the top cloud with one mode having mode radius 0.3 μm, and the other with mode radius 1 μm.

The lower and middle clouds form by condensation from existing vapor rather than by photochemistry which drives particle formation in the top cloud. Two vapor constituents (H_2SO_4 and H_2O) are present in varying amounts. Small particles are present below the base of the lower cloud and these can serve as condensation nuclei. Model simulations show that the two most important parameters of the model are the vertical transport, as characterized by an eddy diffusion profile which is a function of altitude (it reaches a peak at 54 km altitude in the model), and the nucleation mechanism. The nucleation mechanism is thought to be a combination of heterogeneous nucleation on insoluble cores and an activation mechanism on soluble (hygroscopic) cores. A numerical model accounts for the balance in vapor pressures of H_2SO_4 and H_2O as a function of temperature, the supply of vapor via eddy diffusion, loss rates due to sedimentation, and continuity of particle mass as particles grow. The model reproduces the observed tri-modal particle size distribution. The smallest particles (less than 0.6 μm) are the insoluble cores. The largest particles (typically 3–4 μm radius) are those which nucleate near the top of the cloud. As they fall they grow by collision and condensation, becoming larger at lower altitudes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars

Water molecules condense on dust nuclei in the Martian atmosphere to form water ice clouds. The water abundance is not high and so ice condensation requires both cold temperatures and availability of dust condensation nuclei. At very cold temperatures carbon dioxide can condense on the water ice particles.

Images from space and from the Viking and Mars Pathfinder landers have shown that ice cloud formation on Mars is highly variable. When Mars is farthest from the Sun the temperatures are colder and ice clouds are more extensive and more frequent. But even during the few weeks of the Mars Pathfinder mission cloud formation near the landing site was highly variable.

There are several major differences between ice cloud formation on Mars and the Earth. The water abundance in the Martian atmosphere is much less than that for Earth. As a consequence there are so few particles the coagulation process can be ignored, leaving nucleation, condensation and sedimentation as the factors which determine cloud formation and particle size distribution. The smallest particles are the same size as the condensation nuclei. The ability of Martian dust particles to serve as nucleation sites is highly uncertain because not much is known about the mineralogy, which has a strong effect on nucleation. Clay minerals might be present, and they typically are efficient nucleators. Other mineral types are inefficient. One way to account for the uncertainty in modeling this process is to make the contact parameter a parameter of the microphysical model. The energy associated with ice particle formation is proportional to the factor $(2+m)(1-m)^2$ where m is the contact parameter. Values of m close to 1 (e.g. for silver iodide, for example) reduce the energy of formation leading to rapid nucleation. Typical terrestrial soil particles have m near 0.4.

Microphysical models which employ the temperature profile as determined by the Mars Pathfinder instruments find ice cloud formation in two layers. The highest is in the altitude range 30–50 km. A second layer forms at the temperature minimum near 10 km altitude. Competition between growth by nucleation and condensation and loss by sedimentation defines the size distribution. Particles in the upper layer have mean radius near 1.5 μm, whereas those in the lower layer have mean radius near 2.5 μm.

Titan

As is the case for the Earth, Venus, and the giant planets, clouds or haze in the stratosphere form as a result of photochemical processes, whereas those in the troposphere form from condensation. At the deepest level there should be methane clouds. Other condensate clouds, possibly with a variety of compositions, may form at intermediate levels.

Methane cloud formation near the temperature minimum (below 40 km altitude) is highly uncertain. The Voyager infrared spectra have been interpreted by some to imply that methane is highly supersaturated near the temperature minimum. This cannot occur in the presence of a substantial cloud, and so would imply that condensation occurs only rarely, forming large ice (or liquid in the lowest few km) particles which fall rapidly to the surface (rain without clouds). That scenario was proposed prior to the interpretation of high global supersaturation of methane. Titan and the Earth may differ significantly in terms of the number of condensation nuclei available to initiate particle formation. On the Earth there are numerous CCN mixed up from the surface. On Titan the only CCN available may come from slow sedimentation of the overlying photochemical haze and from other condensates. The number density from that source is thought to be quite low compared to that for the Earth. Low CCN density would lead to fewer particles, but those that form can grow by condensation to large size (radii greater than 50 μm) because their formation does little to deplete the available vapor. Both microphysical models and recent ground-based near-infrared spectra imply patchy and variable clouds of large methane ice or liquid particles which form at an altitude near 15 km.

Titan's stratospheric photochemical haze is composed of complex hydrocarbons and nitriles. The starting point for this process is the photolysis of methane and nitrogen high in the atmosphere (higher than 300 km altitude). Although the process is slow, over geologic time enough methane should have been processed to generate 100 m thick surface deposits of liquid ethane and solid organic material, provided a surface reservoir of methane has been available to replenish that which is destroyed over billions of years. The haze particles are undoubtedly complex organics and nitriles, although not enough is know to be more specific regarding the composition.

Microphysical models for Titan's stratospheric haze lump the chemical processes leading to formation of condensable organics into one adjustable parameter, the mass production rate of aerosol-forming molecules at the top of the haze layer. Nucleation is assumed to be initiated by formation of small clusters of molecules which grow through collisions. Models include particle charging by UV and cosmic ray events. Charge is a free parameter (number of electrons per unit particle radius) and serves to inhibit growth by coagulation. Traditional microphysical models assumed the particles remain spherical as they grow. This would be expected if the particles are liquid. More recent models incorporate laboratory results and theoretical predictions which favor solid particles. As solid particles collide they stick together to form aggregates. As the aggregates collide with each other they form larger aggregates which have a fractal nature. Cluster–cluster aggregates have fractal dimension close to 2 (versus spheres which have fractal dimension 3). Near the top of the haze layer ($\sim$400 km) the growth process is via molecules which stick to individual particles. The individual particles grow to some size limit determined by the pressure. Both theory and observation indicate the individual particles, or monomers have radii near 0.066 μm. Mutual collisions among the monomers give rise to clusters and to cluster–cluster aggregates as the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

particles settle. The cluster aggregate idea can explain the combination of strong linear polarization in light scattered near 90° scattering angle and the strong scattering of light in the forward direction.

Giant planets

Microphysical studies of the condensate clouds in the tropospheres of the giant planets rely on the familiar principles which describe growth by condensation, coagulation, and coalescence and loss by evaporation and sedimentation. However, these models are extremely crude because the number and composition of potential cloud condensation nuclei are unknown. Furthermore cloud formation depends strongly on atmospheric dynamics which are not taken into account. There are no *in situ* measurements except for the Galileo probe which found few particles in its descent path through a volatile-depleted patch of Jovian atmosphere. Some microphysical models predict that the methane clouds on Uranus and Neptune are 10–100 times more massive than any clouds on Earth. However, the observations show very little methane cloud mass. Observational evidence for the water cloud is lacking, although higher clouds interfere. The ammonia cloud near the top of the troposphere on Jupiter and Saturn is predicted from microphysical models to be a weakly precipitating cloud similar to terrestrial cirrus, and observations are consistent with that idea.

Coupled photochemical/cloud microphysical models are able to account for the optical depth and size distribution of hydrocarbon haze particles in the stratospheres of Uranus and Neptune, provided the mass production rate and particle charge are adjustable parameters. The models account for particle sizes (all submicron) and vertical profiles as derived from Voyager data. Eddy diffusion transports methane to the upper atmosphere where photochemical processes lead to formation of diacetylene, acetylene, and ethane ice particles which condense near 0.1 mb, 2.5 mb, and 14 mb, respectively. As the particles fall to pressures greater than 600 mb they evaporate as the temperature rises. Organic polymers can also form, probably via solid-state photochemistry within the original ice particle during the settling lifetime. This process can account for the weak absorption seen in reflected sunlight.

Bibliography

Grinspoon D H, Pollack J B, Sitton B R, Carlson R W, Kamp L W, Baines K H, Encrenaz Th. and Taylor F W 1993 Probing Venus's cloud structure with Galileo NIMS *Planet. Space Sci.* **41** 515–42

Houze R A 1993 *Cloud Dynamics* (New York: Academic)

Kahn R 1984 The spatial and seasonal distribution of Martian clouds and some meteorological implications *J. Geophys. Res.* **89** 6671–88

West R A, Strobel D F and Tomasko M G 1986 Clouds, aerosols and photochemistry in the Jovian atmosphere *Icarus* **65** 161–217

West R A, Baines K H and Pollack J B 1991 Atmospheric clouds and aerosols in the Uranian atmosphere *Uranus* ed J T Bergstralh, E D Miner and M S Mathews (Tucson, AZ: The University of Arizona Press) pp 296–324

Robert West

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clown Face Nebula (NGC 2392)

A planetary nebula, also known as the Eskimo Nebula, in the constellation Gemini, position RA $07^{h}\ 29.2^{m}$, dec. $+20^{\circ}\ 55'$. It is bluish, $13''$ in diameter, and of ninth magnitude, with a tenth-magnitude central star. The blue-green nebula's hazy outer regions are thought to resemble an Eskimo's hood or clown's ruff.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cluster

Two European Space Agency missions to send four identical spacecraft to explore the interaction between Earth's magnetosphere and the solar wind. SOHO and Cluster made up the first ESA 'Cornerstone'. Cluster I was lost in the failed maiden launch of the Ariane 5 rocket in 1996. Cluster II, which replicates the original mission, will be launched from Baikonur in summer 2000. Four spacecraft will provide the first three-dimensional measurements of large- and small-scale phenomena in the near-Earth environment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cluster Membership

Our Milky Way Galaxy contains numerous stellar systems, the members of which are thought to have a common origin and are gravitationally bound. The galactic GLOBULAR and GALACTIC OPEN CLUSTERS are the most important representatives of such systems. Less conspicuous are the low-density and gravitationally unbound stellar associations. There is a tight link between the open clusters, associations and the galactic disk which is readily visible on the starry sky. However, the galactic disk is very rich in stars, such that the open clusters and associations, projected against this background, start to blend with it. This blending, or in other words the cluster contamination by galactic disk stars, becomes severe at fainter magnitudes. Handling the cluster contamination is pretty much part of the problem of determination of cluster membership. The galactic globular clusters, on the other hand, are so rich in stars that a casual background star rarely is adopted as a cluster member. This may not be the case if the globular cluster is poorly populated and/or is located near the center of the Galaxy where the background star density reaches a maximum.

Why do we need to identify the cluster members as confidently as possible? Observation of STAR CLUSTERS is important for comparison of the theories of STELLAR EVOLUTION and STELLAR FORMATION with observations. In fact, the stellar evolution theory for stars younger than a few Gyr almost entirely rests on observations of open star clusters and the Sun. If the observations of a star cluster are contaminated by background stars, it inevitably will affect the cluster's HERTZSPRUNG–RUSSELL (HR) DIAGRAM and the derived mean properties of the cluster may be in error. In turn, stellar formation theory requires knowledge of the stellar initial mass function (IMF), which in principle can be deduced from the observations of star clusters. Again, the contamination together with completeness of cluster membership are the crucial issues in an attempt to derive the IMF.

A prerequisite to a thorough study of any specific star cluster and association is a membership determination, which is valuable in its own right. Membership determination presents a paradoxical situation. On the one hand the concept is easy to grasp but on the other hand, the practical implementation has kept generations of astronomers busy and surely will in the future. This paradox is sometimes overlooked even by workers in the field and is not adequately acknowledged by others using cluster-study results.

The following discussion has a passing connection to the group of statistical methods known as clustering techniques, which are often used in the analysis of galaxy distributions. In the case of star clusters, the underlying assumption is that their existence usually is well established, whereas individual membership probabilities may not be known.

Membership probability

Since the star clusters are projected onto the Galactic star field, we always deal with a mix of two populations: the cluster (c) stars and field (f) stars. There is no readily available criterion which will give us a clear discrimination between these two populations. Thus, we can only estimate a star's probability, p, of being a member of the cluster, which can be defined as follows:

$$p = \frac{F_c}{F_c + F_f} \tag{1}$$

where F_c and F_f are some distribution functions (often called the probability density functions, pdf) for the cluster and field respectively. These functions, in turn, are dependent on some physical quantities and their errors characterizing the cluster and field stars.

In principle, any membership criterion based upon physical or other phenomena is useful, as long as it allows one to separate the cluster and field population to some degree. A true cluster star will probably pass all 'filters' set by different approaches if the assumptions are correct and unbiased, whereas a field star is expected to fail on at least one of them.

Usually, star clusters have been discovered as local, projected star-density enhancements on the sky. As a first cut, one can draw an imaginary circle on the sky around such enhancement and declare that every star within the circle is a cluster member. Unfortunately, at fainter magnitudes this approach leads to poor results. As an example, let us consider the case of a relatively nearby open cluster, Messier 39. Armed with *a priori* knowledge of true cluster members, it is possible to estimate the mean probability for each star to be a cluster member at different apparent magnitudes within a radius equal to 15 arcmin around the cluster center. In table 1 a cluster membership probability of 75% means that 3/4 of those stars are cluster members, whereas 1/4 are field stars. Since this is a probability, we cannot tell which specific stars are the cluster members. Obviously, the lower the probability, the lower the chances of finding a cluster star in the sample.

As seen from table 1, down to V ~11 mag the number of cluster stars is high enough to let us ignore the presence of field stars. However, at fainter magnitudes the cluster members virtually vanish into a sea of field stars. This is a very common picture in open cluster studies.

Fortunately, there are a few features which are different for stars in clusters and those in the galactic field. For example, due to their common origin and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Membership probabilities.

	Radius (arcmin)	
	15	55
Visual magnitude, *V*	Probability estimate (%)	Proper motion probability (%)
8	75	96
9	87	83
10	83	86
11	20	71
12	7	66
13	4	63
14	2	30

gravitational field the cluster stars are located within a limited volume rarely exceeding 20 pc in diameter. In the HR or the color–magnitude diagrams cluster stars are mainly located within a narrow band along the main sequence, with the exception of very young open clusters in which the less massive stars are scattered well above the so-called zero-age main sequence. On the other hand, the distribution of galactic field stars in the color–magnitude diagram merely forms a wide and diffuse cloud. This difference in appearance of color–magnitude diagrams is one frequently used method to delineate the cluster stars from the field. Although it never deals with calculation of membership probabilities, in essence it boosts membership probabilities *en masse*. However, moving down to fainter magnitudes, there always is a critical region of the color–magnitude diagram where field stars pollute the diagram beyond recognition. Ironically, at the moment this region occupied by the faintest members of star clusters, such as white and brown dwarfs is at the center of attention of many astronomers for very good reasons. These objects can provide an independent absolute age scale and may reveal some secrets of planet formation. Apparently, there is a need to derive real cluster membership probabilities, not just estimates.

Any star's motion can be divided into two measurable quantities—tangential velocity on the sky or the so-called proper motion, and radial velocity along the line of sight. These two velocities are heavily used in calculating the membership probabilities and certainly warrant a closer look.

For a typical open cluster, the scatter of field-star motions is approximately up to a hundred times higher than the dispersion of internal motion velocities around the cluster center of mass. Essentially, this disparity is the key in determination of cluster membership using star motions. Of course, each cluster has its own systemic motion within the Milky Way. Membership determination is much easier if this motion is significantly different from the mean motion of field stars.

Proper motion as a tool for membership

Although the existence of PROPER MOTION has been known since its discovery by Edmond Halley in 1718, it is not easy to measure a star's proper motion.

First, proper motions are very small. A typical unit for measuring the proper motion is a milliarcsecond per year (mas yr^{-1}) and very few stars have motions exceeding 1000 mas yr^{-1}. In the cluster world, the Hyades cluster members possess the largest proper motion, around 100 mas yr^{-1}. Second, the standard ground-based technique for determination of proper motion requires a concerted long-term observational effort. Traditionally, the positions of stars are measured on two or more sets of photographic plates taken at different epochs spanning at least several decades. Then, all measured coordinates are reduced into a common coordinate system at a fixed epoch, allowing a linear time-dependent coordinate term for each star (a.k.a. proper motion) in addition to the conventional global polynomial terms. The reliability of a proper motion is characterized by the ratio of the proper motion to its standard error. The rule of thumb is, the longer the time span, the more precise the proper motion. It is relatively easy to obtain proper motions at the 1 mas yr^{-1} accuracy level, although sub-mas yr^{-1} accuracies are still a challenge. Moreover, for most of the open clusters there is virtually no early plate material available, and when it is, the plates frequently do not contain faint enough stars.

The importance of proper motion in cluster membership determination was understood long ago, but a thorough exploitation of proper motion started only with the introduction of photography in astronomical observations during the last decades of the nineteenth century. Since then, about 100 out of 1200 known open clusters have at least one determination of proper motion using the photographic method, as opposed to about 500 clusters having had photometric observations accomplished.

A special place in cluster membership studies is occupied by the recent very successful HIPPARCOS mission and planned future space astrometry missions, such as SIM, FAME and GAIA. Hipparcos has made a significant impact on the nearby open cluster and association membership and census. Due to expected positional accuracies at the unprecedented micro-arc-second level, the future space missions, FAME and GAIA, will be in a position to provide excellent membership information for virtually all registered open clusters. These future missions are expected to revolutionize studies of the Milky Way, including the star clusters.

As said before, membership determination using proper motion utilizes the fact of quite different dispersions in the velocities of cluster and field stars. In practise, however, each determination of proper motion has some error. In other words, the true proper motion is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

always somewhat blurred due to measuring and other errors. If these errors are comparable with the dispersion of field stars (usually around 5–10 mas yr^{-1}) and systemic motion of a cluster does not differ significantly from the mean motion of field stars, then cluster member segregation is impossible.

Consider again the open cluster Messier 39, which has a rather large sample of stars having proper motions with accuracies slightly above 1 mas yr^{-1}. The working tool to analyse proper motion is the so-called vector-point diagram (VPD). In this diagram (figure 1) with axes usually oriented along right ascension and declination, each dot is the end of a proper motion vector. Two populations of dots can be noticed: a large concentration around (0,0) and a smaller one around (–6, –16) mas yr<figref id="EAA2854fig1">. The latter essentially represents the cluster stars. Since errors are the major source of cluster-star dispersion in VPDs, the distribution of cluster stars is very close to Gaussian. The dispersion of field stars is mainly caused by true motion of the galactic stars which, generally speaking, do not obey any simple distribution; however, superposition of many smaller distributions of a poorly known nature produces a distribution which appears Gaussian. This fact is used by astronomers to calculate the membership probabilities just from proper motions in the following manner. The cluster and field proper motion distribution functions, F_c and F_f, are expressed explicitly as two separate Gaussians with their own parameters. The sum of these distributions, $F_c + F_f$, essentially represents the distribution of vector points in the VPD. A number of mathematical techniques are devised to find the parameters of these distributions. Once they are known, it is trivial to calculate individual membership probabilities using equation (1). The second set of mean probabilities in table 1 shows how dramatically the mean membership probability increases at fainter magnitudes with respect to the would-be probabilities of a smaller central star sample. Despite the area of the sky under the survey being more than ten time larger, the proper-motion-based probabilities still are very reliable, and what is more important, such probability is provided for each star in the sample. These probabilities are totally independent of any astrophysical theory predictions, for instance how the cluster color–magnitude diagram should look. This provision is critical in the case of 'weird' stars, such as blue stragglers, variables, binaries and other exotic objects. Figure 2 shows the color–magnitude diagram for all stars in the region of Messier 39 (figure 2(*a*)). The cluster's upper main sequence is barely visible; however, with the help of proper motion virtually all field stars are gone (figure 2(*b*)) and the main sequence is beautifully delineated.

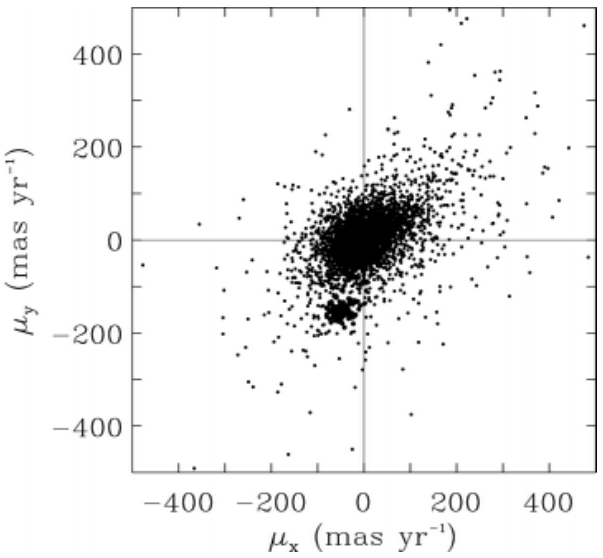


Figure 1. Vector-point diagram of proper motions in the region of open cluster Messier 39. Each dot is the end of a star's proper-motion vector originating from the zero point. The cluster stars can be recognized as a small clump around (–6, –16) mas yr^{-1}.

On the other hand, proper motion is not an absolute criterion in deciding the membership of a single star of particular interest. Even a high membership probability, by the very nature of probabilities, leaves a chance for a field star. Thus, if possible, another criterion, such as radial velocity, should be used.

Radial velocities

RADIAL VELOCITY determinations are made by measuring the Doppler shift in the spectra of stars, i.e. the displacement of spectral lines, $\Delta\lambda$, with respect to their wavelength at rest, λ_0, as determined from laboratory measurements. In the classical framework this displacement is proportional to the velocity along the line of sight

$$v_r = c\frac{\Delta\lambda}{\lambda_0} \qquad (2)$$

where c is speed of light. The radial velocities of open cluster stars, as measured from the solar system, rarely exceed 10–20 km s^{-1}, hence a required accuracy of radial velocities to ±1 km s^{-1} or better necessitates that the wavelength displacement must be determined to a few parts in a million, which is a challenge even to modern spectroscopy. The most frequently used technique for radial velocity measurements is the cross-correlation of an observed spectrum with a high-quality template (optical or synthetic) of a reference spectrum. This way it can be done directly, while observing as in the CORAVEL spectrometer, or numerically, using the CCD spectrograms taken for one star at a time as in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

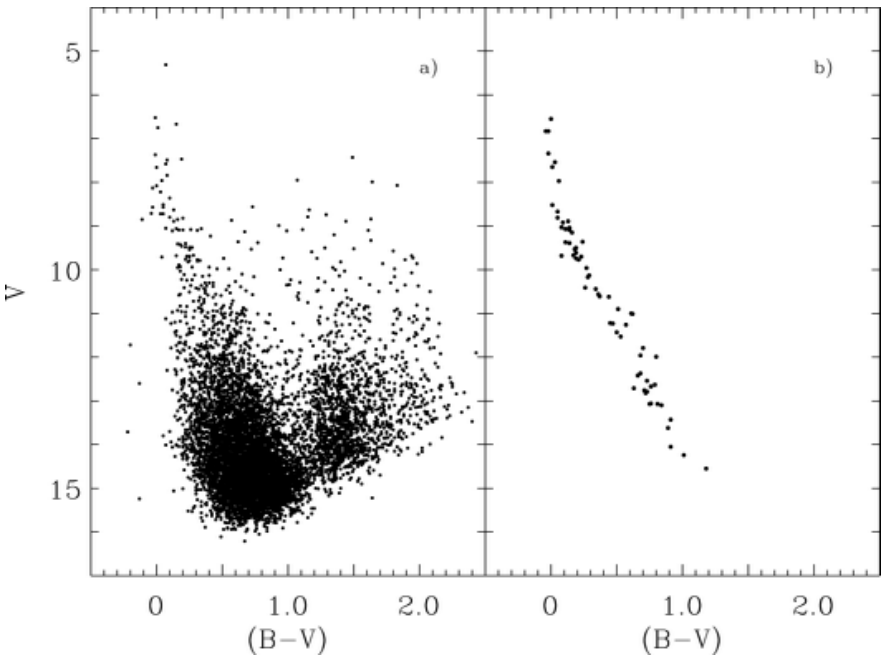


Figure 2. Color–magnitude diagram in the region of open cluster Messier 39: (*a*) all stars; (*b*) only probable cluster members.

French/Swiss ELODIE and CORALIE designs and digital speedometers developed at the Harvard-Smithsonian Center for Astrophysics, USA, or for many stars simultaneously (cf the Hydra multi-object spectrograph on the WIYN telescope at Kitt Peak, USA). The most accurate radial velocities have been achieved by placing an absorption cell (HF or iodine) in front of the spectrograph, but this technique requires a high level of light flux and is mainly used in extra-solar planet hunting around bright stars.

The spectra of late-type (F-G-K) stars contain a large number of small-width metal absorption lines which makes the application of cross-correlation techniques relatively easy. This explains why the vast majority of cluster stars having precise radial velocities are of late spectral type. Due to the scarcity of spectral features coupled with their strong dependence upon the physical conditions in STELLAR ATMOSPHERES, and often large rotational velocities of hot stars, it is much more difficult to measure radial velocities of early-type O, B and A stars which usually are the brightest in open clusters. Recent advances in modelling the stellar atmospheres and tuning these models to a star's actual physical parameters promise to close this gap soon.

Figure 3 shows the distribution of radial velocities in the open cluster NGC 188. In some sense it resembles the vector-point diagram, though just in one coordinate. In principle, it is possible to model the cluster and field star distributions, but for two stumbling blocks. First, the radial velocity measurements are telescope-time prohibitive and, hence, the star samples are always optimized to include a maximum number of potential cluster members. In other words, such selection is greatly biased towards cluster members. This causes a severe under-representation of field stars and, consequently, undermines the statistical grounds for calculating the membership probabilities. Secondly, about half or even more of cluster members can be binaries (some are even triple systems), like those marked in grey in figure 3. The orbital motion will induce a time-dependent offset relative to the mean motion of the binary system. To detect such an offset motion, the observations must span a considerable part of the orbital period. If the offset goes undetected, then a star could be erroneously categorized as a field star. While this may appear annoying for the determination of cluster membership, most researchers thrive on binaries, since they provide such fundamental astronomical data as masses, radii and luminosities, and have a profound impact on star formation history and dynamical evolution of stellar systems (see also BINARY STARS IN GLOBULAR CLUSTERS).

For these reasons, the determination of cluster membership from radial velocities is so far limited to a simple division of stars into cluster members and non-members, depending whether or not their radial velocities fall within the 3σ window around the cluster's mean radial velocity. Nevertheless, even such a rudimentary membership in combination with a membership from proper motions weeds out field stars quite effectively. To date, about 200 open clusters have some sort of radial velocity measurements, mainly for the giant-branch stars. Only a dozen of them, mainly the nearby clusters, have measurements for the main-sequence stars, but this is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

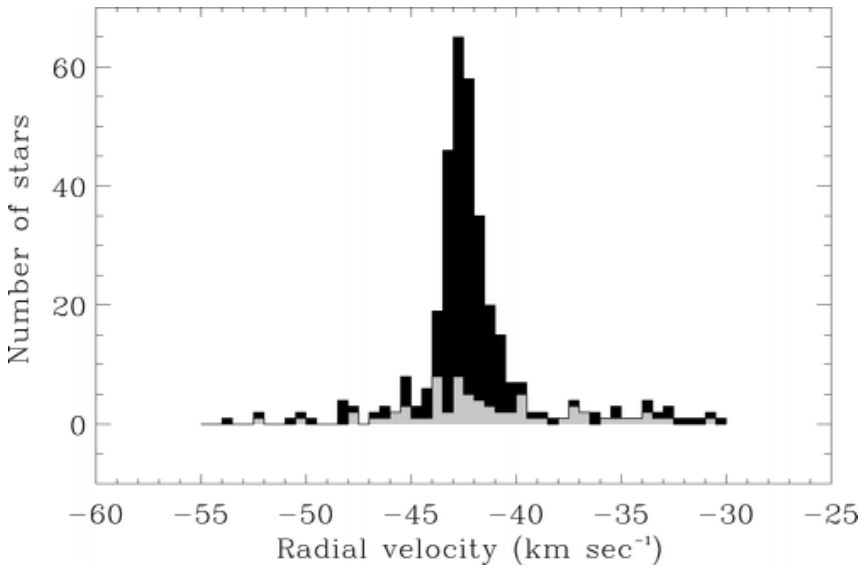


Figure 3. A histogram of radial-velocity distribution in the open cluster NGC 188. The grey area indicates the number of stars showing variable radial velocity. The central peak is the location of mainly the cluster stars. (Courtesy of R D Mathieu and C Dolan, University of Wisconsin-Madison, USA.)

rapidly improving with new advances in the multi-object spectrograph design and the availability of large telescopes.

Membership in associations

The stellar association membership presents a real challenge to anyone attempting to determine them. Most common are the so-called OB ASSOCIATIONS. These associations are gravitationally unbound star systems, recognizable by some excess density of O and B spectral type stars against the sky background and a small internal velocity dispersion. The latter does not exceed a few km s^{-1} and, as a result, the association can be detected by common kinematics of its members, though in a less consistent manner than for the star clusters. Since the associations cover tens to hundreds of square degrees on the sky at extremely low density (a few stars per square degree), before the Hipparcos mission the membership information was limited to the bright stars and even it was not reliable. The Hipparcos measurements are ideally suited for membership determination in nearby OB associations out to the distance of ≈650 pc.

The large spatial extent of associations prompts researchers to use the so-called moving cluster technique. All association stars are moving so that a receding association approaches a location on the sky known as the covergence point. In the case of associations this is only partially true because the star motions are slightly diverging. Apparently, the velocity dispersion perpendicular to the direction of the association's motion is just a combination of an intrinsic velocity dispersion convolved with the observational errors. The distribution of this velocity dispersion is used to calculate membership probabilities for associations. With the advent of Hipparcos parallaxes and proper motions it is possible to derive an elliptical membership probability cylinder in 3D velocity space, which constitutes the so-called 'spaghetti' method introduced by Dutch astronomers. Now a much better census of the stellar content in stellar associations is known to within 1 kpc of the Sun.

Globular clusters

The main thrust in globular cluster astrometry is towards the determination of absolute proper motions which allow one to calculate the orbits. The knowledge of orbits is crucial in understanding the content and properties of presumably a rather heterogeneous system of globular clusters, some possibly captured during galactic merger events (see GLOBULAR CLUSTER SYSTEMS IN INTERACTING GALAXIES).

High-precision relative proper motions are required to identify potential field stars projecting onto the cluster. Thus, the Yerkes refractor plate material has produced clean samples of cluster stars in a dozen globular clusters. Since the globular clusters are rather distant objects, only upper parts of the cluster color–magnitude diagram are visible on old photographic plates. It is expected that the Hubble Space Telescope (HST) archival CCD frames will be useful in producing deep membership studies in globular clusters. The positional accuracy on the HST CCD frames is sufficiently high to reach a sub-mas yr^{-1} accuracy level for proper motions in a few years, a feat impossible from ground-based observations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Embedded clusters

This is a class of very young open clusters surrounded by thick clouds of molecular gas and dust, presumably leftovers from star formation. Usually these clouds block out most of the star light. However, in the near-infrared wavelengths the clouds become fairly transparent. With the advent of panoramic near-infrared wavelength detectors it is possible to sample the nearby ($d < 1$ kpc) embedded clusters down to and well below the hydrogen burning limit. Such clusters are ideal places to study the stellar initial mass function (IMF), since stellar evolution and dynamical effects have minimally changed the content and mass of embedded clusters. On the other hand, even at infrared wavelengths the extinction in the direction of cluster members can be non-uniform, which coupled with the emission due to the presence of circumstellar disks and other remnants of star formation make the mass determination rather difficult.

For the handful of well-studied embedded clusters, cluster membership is still not available for a couple of reasons. First of all, the density of stars in these clusters is much higher than in the surrounding field, thus allowing one to apply a simple statistical field-star subtraction. Secondly, the time span between the first and second epoch imaging-array frames is currently too short to derive reliable proper motions. In addition, the whole field of INFRARED ASTRONOMY is rapidly evolving and the future instrumentation will most likely provide new means to find the cluster membership.

Alternative membership method

The faint end of cluster sequences in the HR diagram is invariably contaminated by field stars due to a couple of aggravating circumstances. First, the cluster luminosity function (in essence, the number of stars per unit magnitude) at around the solar luminosity and fainter is getting flat, mainly due to the effect of dynamical 'evaporation' of cluster stars. On the other hand, the field-star luminosity function is steadily growing. In this situation, the observed ratio of cluster to field stars is plummeting and so are the membership probabilities. Second, at faint magnitudes membership is plainly lacking because of the shallow old-plate limiting magnitude and unacceptably low light-flux levels for radial velocity measurements.

It turns out that the majority of young stars of spectral type A and later have considerable coronal activity, showing up in x-rays. The ROSAT x-ray satellite, operating from 1990–8, was the prime instrument to map out the 'soft' (0.1–2.4 KeV) x-ray sky. It was quickly realized that almost all nearby clusters of Pleiades age and younger should be visible on the ROSAT x-ray 'sky', and these positive identifications can be used as a valuable membership criterion. Using the ROSAT data, nearly a dozen open clusters have their membership lists enlarged substantially at the faint end. Although, it is not entirely clear if all fainter cluster stars have such x-ray activity, it is nevertheless a powerful alternative membership determination technique to complement classical kinematical methods. Perhaps its potential somewhat diminishes in the areas of nearby OB associations where the fraction of x-ray bright field stars is higher.

Future

Despite the huge effort put into the determination of star cluster and association membership, there is still plenty of work ahead. Not a single cluster is known, for which we have discerned all cluster members. To make matters more complicated, the Hipparcos observations have revealed a number of cluster stars beyond the tidal radius for the nearest open clusters. The tidal radius is the distance beyond which the cluster loses its stars due to Galactic gravitational tidal forces. This means that the all-sky proper motion and parallax surveys, such as the planned GAIA mission, should be used to derive a decisive and complete cluster or association membership. The parallax will provide a hitherto missing positional coordinate—the distance. The cluster membership can be rigorously derived only with a full bundle of three positional coordinates and three velocity components in hand, which, compared with the potential energy of a cluster, ultimately defines whether or not the star is likely to be bound to a cluster. The instantaneous positional accuracy of the planned astrometric missions is so high that it would take only a couple of months to achieve the best ground-based proper motion accuracy, as opposed to the several decades needed before.

Bibliography

Baranne A, Mayor M and Poncet J L 1979 CORAVEL—a new tool for radial velocity measurements *Vistas Astron.* **23** 279–316

Cabrera-Caño J and Alfaro E J 1990 A non-parametric approach to the membership problem in open clusters *Astron. Astrophys.* **235** 94–102

Randich S, Schmitt J H M M, Prosser C F and Stauffer J R 1996 The x-ray properties of the young open cluster around α Persei *Astron. Astrophys.* **305** 785–805

Vasilevskis S 1962 The proper motions of open clusters *Astron. J.* **67** 699–706

de Zeeuw P T, Hoogerwerf R, de Bruijne J H J, Brown A G A and Blaauw A 1999 A Hipparcos census of the nearby OB associations *Astron. J.* **117** 354–99

Imants Platais

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cluster of Galaxies

A collection of galaxies that is held together by gravity. Clusters may contain from a few to a few thousand member galaxies. Small clusters, with up to a few dozen members, are referred to as 'groups', the Milky Way Galaxy, for example, being a member of the Local Group, which contains at least 25 members. Most galaxies are members of groups or binary pairs. Larger clusters contain hundreds or thousands of members and, typically, have diameters of a few megaparsecs (about 10 million light-years). Rich (densely populated) clusters are divided into regular clusters and irregular clusters.

Regular and great regular clusters are fairly symmetrical spherical systems with a strong concentration of galaxies towards their centers. They frequently contain several thousand members, the great majority of which are elliptical or lenticular (S0) galaxies, and usually contain one or more giant elliptical (cD) galaxies at their centers. It is believed that numerous collisions and mergers have taken place in clusters of this kind, and that giant galaxies have grown by absorbing smaller ones, a process that is sometimes called 'galactic cannibalism'. The shortage or absence of gas-rich spirals within such clusters is believed to be the result of gas being stripped out of them, or converted rapidly into stars, during collisions and close encounters. Such clusters often contain substantial quantities of tenuous high-temperature gas (with temperatures of the order of 10^8 K), which radiates at x-ray wavelengths. Irregular clusters are more diffuse and less structured. They contain a mixture of all types of galaxy, including substantial numbers of spiral galaxies. Within clusters of this kind, close encounters and collisions are relatively rare.

The nearest large cluster is the Virgo cluster, an irregular aggregation of more than a thousand galaxies that lies at a distance of about 50 million light-years. The nearest rich regular cluster is the Coma cluster, which lies at a distance of about 300 million light-years, and is estimated to contain up to ten thousand galaxies.

Studies of the internal dynamics of clusters (the speeds at which the member galaxies move around within the cluster) imply that most clusters can hold themselves together by gravitational forces only if they contain at least ten (and in some cases up to a hundred) times as much dark matter as luminous matter.

See also: Coma Cluster, dark matter, elliptical galaxies, Hubble classification, interacting galaxies, Local Group, spiral galaxy, superclusters and the Local Supercluster, Virgo Cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Coalsack

A large and prominent dark nebula in the constellation Crux, encroaching into neighboring Centaurus and Musca, centered approximately on position RA $12^{h}\ 50^{m}$, dec. -63°. It is easily visible against the bright background of the Milky Way, and measures nearly 7° by 5°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Collimation

The process whereby the components of an optical system, such as a telescope, are positioned and aligned to give optimum performance. For a refractor, each lens should be centered on, and perpendicular to, the optical axis of the telescope. Cassegrain reflectors, and related types, are collimated by ensuring that the axis of the primary mirror passes through the center of the secondary mirror and that the centers of the eyepiece and the secondary mirror lie on this same axis. Where light from the primary is reflected through a right-angle (as in the Newtonian reflector), the system is collimated when the axis of the primary mirror and the axis of the eyepiece intersect at the center of the diagonal flat mirror and are inclined to its surface by the same angle.

A collimated beam of light is one that consists of a bundle of parallel rays. A collimator is a device (lens or mirror) that produces a collimated beam. This may be achieved, in principle, by arranging for light to pass through a small aperture or pinhole placed at the focus of a lens; after passing through the lens, the emerging beam of light should consist of parallel rays.

See also: Cassegrain telescope, eyepiece, focal length, mirror, Newtonian telescope, reflector, refractor, spectroscope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Collisionless Processes in Astrophysical Plasmas

Plasmas, also called the fourth state of matter to distinguish them from solids, fluids and electrically neutral gases, are ionized gases. This generally implies that their thermal energy exceeds the first ionization potential such that a considerable fraction of their atoms (or molecules) is ionized, although even at lower thermal energies substantial ionization in a gas may also be produced by other processes, for instance by external radiation. It also implies that plasmas consist of free charged particles, in general positively charged ions and negatively charged electrons. Exceptions from this general rule are 'gravitational plasmas' which are dilute inert gases exposed to a gravitational field or behaving under the action of their own gravitation and showing similarities to ordinary plasmas. Other exceptions are plasmas built of electrons and positrons, positively and negatively charged ions, dusty plasmas where the dust grains may carry any amount and any sign of charges, electroweak plasmas where the interaction is mediated by massive W and Z bosons instead of photons and, finally, chromodynamic plasmas which consist of asymptotically free quarks and gluons.

All such 'plasmas' may be found in cosmic space at different locations and at different cosmological times depending on the available interaction or thermal energies. For instance, shortly after the big bang at about 10^{-10} s the plasma was in the electroweak state (see COSMOLOGY: STANDARD MODEL). With transition below 100 GeV temperature in energy units at this time the W, Z bosons became photons. Similarly, before about 10^{-4} s the cosmological temperature was high enough to allow for the chromodynamic plasma state sometimes called the quark–gluon soup. When the temperature dropped below nuclear binding energies the nuclei froze out. At temperatures well below 0.5 MeV at several seconds the plasma became ordinary plasma. This is the state we are dealing with here. Ordinary plasmas have been abundant in the universe since then until the present. However, until about 300 000 yr after the big bang all the ordinary matter in the universe was in the plasma state. After decoupling of matter from light at about this time the plasma state is still found in many, though only in localized regions, and is no longer of cosmological origin. Therefore, today, the claim sometimes made that almost all matter in the universe would be in the plasma state must be taken with care. Much of it is, in particular most of the hot matter, but by far not all of it. This is even more true as we have little idea yet what the presumably largest amount of gravitating matter in the universe, the mysterious dark matter, is really made of (see DARK MATTER: ITS NATURE).

Quasineutrality and screening

Since ionization produces equal charges of opposite signs, real plasmas are quasineutral. This means that, looked at from the outside, the plasma carries zero total (electric) charge. Embedding the plasma in an external electric field or separating one of the charge components out turns the plasma into non-neutral plasma with entirely different behavior. Looked at from the inside, the plasma is by no means electrically neutral. This is due to the high mobility of the charge carriers of which it is built. Each particle, electron or ion, is in fast thermal motion. It represents a microscopic charge density element, a microscopic current element and because of its motion a microscopic radiating antenna element. It is thus the source of an electromagnetic field. Because of this the plasma is subject not only to moving charges but also to highly fluctuating electromagnetic fields on all time and spatial scales. These electromagnetic fields act on each of the charges in the plasma giving the plasma its characteristic properties that distinguish it from the other states of matter.

Quasineutrality is an important property. It implies that the plasma volume, on average, carries equal numbers of opposite charges. However, because of their different masses, ions and electrons are usually in grossly different states of motion, with the ions nearly immobile and the electrons very fast moving. This implies that on average many electrons pass close to an ion thereby screening the electric field of the ion from the rest of the plasma. From Poisson's equation for the charge density it then follows that the electric field of an ion charge $+e$, with the elementary charge e it carries, is not felt outside a distance $\lambda_D = v_e/\omega_{pe} = (\epsilon_0 k_B T_e/n_0 e^2)^{1/2}$ from the ion. The electric potential $\phi(r)$ of the ion as function of radial distance r from the ion in the surrounding electron cloud assumes the dependence $\phi(r) = (e/4\pi\epsilon_0 r)\exp(-r/\lambda_D)$. The $1/r$ dependence is the usual Coulomb law, while the exponential describes the screening. This screening distance λ_D is the so-called Debye length. It can be written as the ratio of the electron thermal speed $v_e = (2k_B T_e/m_e)^{1/2}$ and the electron plasma frequency $\omega_{pe} = (n_0 e^2/\epsilon_0 m_e)^{1/2}$ and thus depends on the electron temperature T_e and average plasma density n_0 (k_B, m_e, ϵ_0 are Boltzmann's constant, electron mass, and free space dielectric constant, respectively).

Classification

Plasmas can be classified according to their number densities. The denser a plasma the higher is the probability that the particles collide with each other. Plasma processes under the condition of frequent collisions are called collisional. When the collisions are rare or practically absent they are called collision free. Sometimes this terminology is applied to the plasma itself, but it should be kept in mind that what is meant hereby is that the processes taking place are either slower or faster than the typical time between binary collisions in the plasma. In this spirit the degree of collisionality is defined relative to a given time scale. Hence it is not a strict definition. Because of the available long time scales, many astrophysical plasmas can be considered to be collisional. However, for a large class of sufficiently dilute plasmas even in astrophysics the time

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scales are short enough for these plasmas to fall under the description of collisionless plasmas. Collisionless plasma processes take place in stellar winds and the interstellar medium, in the plasmas of molecular clouds, stellar atmospheres, the various magnetospheres of magnetized stars, jet plasmas and in some regions in accretion disks. Moreover, most of the processes leading to shock waves in astrophysics which are visible in either the optical or the x-ray emission belong to the class of collisionless processes or in short to collisionless plasmas.

It is also believed that the environment of black holes contains largely plasmas that are collision free for the processes taking place there. The strict definition of a collisionless process in plasma is the following. Given the time scale τ of the process under consideration and the typical frequency $\nu_c = 1/\tau_c$ of binary collisions between the particles constituting the plasma, i.e. the inverse of the time τ_c between two such collisions, a plasma is called collisionless when $\tau\nu_c > 1$. In order to determine the degree of collisionality of a plasma one thus needs to know both the Spitzer–Coulomb collision frequency ν_c and the typical time scale of the process.

Relaxation time scales

Impose an inhomogeneity or, otherwise, any deviation from thermal equilibrium onto a plasma. Because of the highly dynamical state of plasmas such deviations will readily relax, trying to get rid of the excess energy and tending to return to equilibrium conditions. Equilibria are determined by detailed force balance. In a plasma, in particular in a collisionless plasma, such a force balance is not trivially achieved. According to statistical mechanics, in thermodynamic equilibrium the distribution function should be of the canonical form of a Boltzmann energy distribution. In velocity space this distribution takes the form of a Maxwellian bell-shaped distribution function. Any deviation from this distribution will tend to be smeared out within a certain relaxation time until the distribution returns to a Boltzmannian.

The characteristic relaxation times of the various possible relaxation processes can be calculated from a Fokker–Planck equation that describes the evolution of the non-equilibrium distribution function towards thermal equilibrium. The most important relaxation time is that of approaching the Maxwellian equilibrium distribution function. There is a typical relation between the times for electrons and ions to approach such an equilibrium that can be written as

$$\tau_{ee} : \tau_{ii} : \tau_{ei} \rightarrow \left(\frac{m_e}{m_i}\right) : \left(\frac{m_e}{m_i}\right)^{1/2} : 1. \qquad (1)$$

This shows that the electrons on their own, because of their high mobility, relax after a time τ_{ee} which is faster by the root of the electron-to-ion mass ratio than the time the ions need to relax to equilibrium. Similarly, the ions relax faster by the same ratio than the total plasma, composed of ions and electrons, needs to reach one common temperature for ions and electrons. Because of this, electrons can often be described by isotropic Maxwellian distributions, while electrons and ions may have grossly different temperatures. This exchange of energy among the different components of the plasma depends on the number of collisions the particles experience. Because of the relatively long range of the Coulomb interaction in highly dilute and hot plasmas such as the astrophysical plasmas and because of the generally long astronomical time scales astrophysical plasmas on large spatial and temporal scales may sometimes not be really collisionless.

The collision frequency is defined as $\nu_c = \sigma_c n_0 \langle v \rangle$, where σ_c is the collisional cross section and $\langle v \rangle$ is the average velocity of the particles. The problem is to find an expression for σ_c. In a dilute plasma where all particles interact only via Coulomb forces, the regime applicable to astrophysical plasmas, the collisional cross section is inversely proportional to the fourth power of the thermal velocity of the particles. Hence, the collision frequency becomes $\nu_c \propto n_0/T^{3/2}$. Thus the number of collisions per second decreases rapidly with increasing temperature T (usually the electron temperature which in many cases is the higher one of the two, electron T_e and ion T_i temperatures) and with decreasing density. It is the hot dilute plasmas that tend to be collisionless. Hence, both the collisional relaxation time $\tau_c = \nu_c^{-1}$ and the mean free path of the particles, $\lambda_{\rm mfp} = \langle v \rangle/\nu_c = (n_0\sigma_c)^{-1}$ increase as $T^{3/2}$ with temperature.

Similarly to the above collision time one can define a relaxation time for the energy exchange $\tau_W = W|{\rm d}t/{\rm d}W|$ where W is particle energy, a relaxation time τ_M for the return of a small deviation of the distribution function to the Maxwellian and a relaxation time τ_{eq} for electrons and ions to achieve temperature equilibrium. Typical values for the collision frequency and the relaxation time scales for various astrophysical plasmas are given in table 1. These are based on the above binary collision frequencies.

Table 1. Collision frequencies and relaxation times.

Symbol	Expression	Conditions	Description
τ_{ce}	$0.01T_e^{3/2}/n$	$v_e \gg v_{e\rm th}$	e–e collisions[a]
τ_{ci}	$0.6T_i^{3/2}/n$	$v_i \gg v_{i\rm th}$	i–e collisions[a]
τ_{We}	$0.03T_e^{3/2}/n$		Energy relaxation
τ_{Wi}	$1.38T_i^{3/2}/n$		Energy relaxation
τ_{Me}	$0.02T_e^{3/2}/n$		Isotropization

[a] Density in cm^{-3}, temperature in K.

In the presence of an electric field $\boldsymbol{E}$ the particles experience acceleration or deceleration. For very fast particles, as mentioned above, the friction term can no longer prevent the particles from running away in the electric field. Such electric fields exist in restricted spatial regions. The fraction of electrons that escape is given by

$$\Delta n/n \approx 2.3 \times 10^{13} E_D T_e^{1/2}/\exp{(8E_D/E)^{1/2}} \qquad (2)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where $E_D \approx 4 \times 10^{-10} Z n_e T_e^{-1}$ is the Dreicer electric field, another fundamental parameter of the plasma. For small electric fields no particles will run away, but the larger the electric field becomes the more particles will escape collisional confinement and leave the position to form beams or extended tails on the distribution function.

Description of plasmas

Plasmas without collisions must be treated within full kinetic theory. This implies that one has to solve the combined Vlasov–Maxwell system of equations for both electrons and ions. This system of equations is extraordinarily involved. The Vlasov equation is the collision-free one-particle Boltzmann equation which, given the Hamiltonian $\mathcal{H}(\boldsymbol{p}, \boldsymbol{x}, t)$ and defining the one-particle probability distribution function $f(\boldsymbol{p}, \boldsymbol{x}, t)$, is

$$\partial_t f + \{\mathcal{H}, f\} = 0 \tag{3}$$

where the braces are Poisson brackets and $\boldsymbol{p} = m\gamma_r \boldsymbol{v}$ is the particle momentum, γ_r the relativistic γ factor, $\boldsymbol{v}$ particle velocity and $\boldsymbol{x}$ position. The explicit version of this equation neglecting gravity becomes

$$\partial_t f + \boldsymbol{v} \cdot \boldsymbol{\nabla} f + \frac{q}{m}(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \cdot \boldsymbol{\nabla}_v f = 0 \tag{4}$$

where $\boldsymbol{\nabla}_v \equiv \partial/\partial \boldsymbol{v}$ is the gradient in velocity space. The electric and magnetic fields appearing in equation (4) are related to the distribution function f via the expressions

$$\rho_q = \sum_{i=e,i} q_i \int \mathrm{d}^3 \boldsymbol{v}\, f_i \tag{5}$$

$$\boldsymbol{j} = \sum_{i=e,i} q_i \int \mathrm{d}^3 \boldsymbol{v}\, \boldsymbol{v} f_i \tag{6}$$

for the electric space charge ρ_q and current $\boldsymbol{j}$ densities. These enter the electrodynamic (Maxwell) equations

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j} + c^{-2}\, \partial_t \boldsymbol{E} \qquad \nabla \cdot \boldsymbol{B} = 0$$

$$\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho_q/\epsilon_0. \tag{7}$$

Solutions of this extremely complicated set of equations can be found only in the most simple cases. However, whenever one of the physical time scales is much longer than the relaxation times, the collisions are frequent enough to allow for a fluid description. In such a description the fluid variables (density n, bulk velocity $\boldsymbol{u}$, pressure tensor $\mathbf{P}$, etc) are defined as the ascending moments

$$\mathcal{M}_i^{(i)} = \int \mathrm{d}^3 \boldsymbol{v}\, \boldsymbol{v} \ldots_{(i \text{ times})} \boldsymbol{v} f_i. \tag{8}$$

Under such conditions the plasma can be treated within the MAGNETOHYDRODYNAMIC (MHD) approach which is strictly valid only in the strongly collisional state. Collisionless MHD is, strictly speaking, a contradiction in itself when applied to plasmas even though it has found extensive application and led to many useful results.

A fluid description may also result for collision-free plasmas when some ordering process is at work that forces the particles to behave fluid like. Gyration around a strong magnetic field may have such an effect. This then leads, under certain conditions, to the so-called Chew–Goldberger–Low anisotropic MHD. Also, collisionless plasmas very frequently generate internal correlations, turbulent interactions and other effects that force the plasma particles to exert collective behavior. In such cases the ordinary binary collisions are replaced by so-called 'anomalous' collisions that may also allow for a fluid description. Anomalous collisions are a consequence of plasma wave turbulence.

Generalized Ohm's law

Because of the enormous mass differences between electrons and ions $m_e/m_i = 1/1836$ it is sometimes convenient to consider only the dynamics of the ions. This applies when considering processes that change on a much longer time scale $\tau = \omega^{-1}$ than the electron plasma oscillation time ω_{pe}^{-1} or $\omega \ll \omega_{pe}$. In this case the electron dynamics partially decouples from that of the ions. The momentum equation for electrons can then be used as the definition equation for the electric field. Rearranging it in order to express $\boldsymbol{E}$ through the electron quantities one has

$$\boldsymbol{E} + \boldsymbol{u}_e \times \boldsymbol{B} = -\frac{m_e \nu_c}{e}(\boldsymbol{u}_e - \boldsymbol{u}_i) - \frac{1}{en}\boldsymbol{\nabla} \cdot \mathbf{P}_e - \frac{m_e}{e}\frac{\mathrm{d}\boldsymbol{u}_e}{\mathrm{d}t} \tag{9}$$

which is an extended form of Ohm's law. Here $\boldsymbol{u}$ is the bulk velocity and $\mathbf{P}$ the pressure tensor. Introducing the current density $\boldsymbol{j} = e(\boldsymbol{u}_i - \boldsymbol{u}_e)$, equation (9) assumes the ordinary form

$$\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B} = \eta \boldsymbol{j} - \frac{1}{en}\boldsymbol{j} \times \boldsymbol{B} - \frac{1}{en}\boldsymbol{\nabla} \cdot \mathbf{P}_e - \frac{1}{\epsilon_0 \omega_{pe}}\frac{\mathrm{d}\boldsymbol{j}}{\mathrm{d}t} + \frac{m_e}{e}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}. \tag{10}$$

Here $\boldsymbol{u}$ is to good approximation the center-of-mass velocity, $\eta = \nu_c/\epsilon_0 \omega_{pe}^2$ the resistivity and the second term on the right-hand side is the Hall term. In a collisionless plasma the first term disappears, and Ohm's law is entirely governed by the Hall term, the electron pressure gradient and the electron inertia in the two last terms, while the left-hand side of equation (10) describes the frozen-in property (see below) widely used in MHD. The terms on the right of Ohm's law modify this condition to a large extent in a real collisionless plasma. The Hall term introduces effects of non-coplanarity, helicity and anisotropy, while the pressure gradients and in particular the inertial terms may violate the frozen-in state.

Properties of collisionless plasmas

Neglecting collisions in the plasma implies that the plasma is ideally conducting. This should not be confused with the plasma being superconducting, however. On the contrary, ideal plasmas may contain magnetic fields, and these fields cannot be expelled from the plasma because they are frozen in into the plasma. In order to pull the magnetic field out, the frozen-in condition has to be broken at least locally by generation of anomalous resistivities that destroy the non-collisionality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Frozen-in concept

The frozen-in property follows from the universal validity of Faraday's law $\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$ even for the plasma, and from the (simplified) Ohm's law $\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B} = \eta \boldsymbol{j}$. The latter gives the proportionality between the electric field $\boldsymbol{E}$ and the current density $\boldsymbol{j}$. The coefficient is the resistivity $\eta = m_e \nu_c / n_0 e^2$ which itself increases with collision frequency. The left-hand side of Ohm's law is a simple consequence of the Lorentz transformation in moving media. Taken together these two equations lead to the fundamental induction equation for a conducting and moving medium such as a plasma:

$$\frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times \boldsymbol{u} \times \boldsymbol{B} + \frac{\eta}{\mu_0} \nabla^2 \boldsymbol{B}. \qquad (11)$$

This equation shows that for $\eta \to \infty$ the magnetic field in the plasma moving with velocity $\boldsymbol{v}$ is zero which implies that in the moving frame the magnetic field does not change in time and hence, as argued above, is frozen in into the plasma.

Gyration

Physically this implies that in the moving frame the plasma particles themselves are tied to the magnetic field lines, gyrating around them in circles of radius $r_B = v_\perp / \Omega$ and gyration period $\tau_B = 2\pi/\Omega$. The gyration frequency is given by $\Omega = eB/m$ and is proportional to the magnetic field. This gyration is simply imposed on the charges by the Lorentz force $\boldsymbol{F}_\mathrm{L} = \pm e \boldsymbol{v} \times \boldsymbol{B}$. Locally it is perpendicular to the magnetic field $\boldsymbol{B}$ and particle velocity $\boldsymbol{v}$. It does not perform work on the particles. It therefore conserves particle energy and merely turns the component of the particle velocity perpendicular to the magnetic field into a circular orbit around $\boldsymbol{B}$.

Breaking the frozen-in condition

In order to take the plasma particles away from the magnetic field in collisionless plasma one thus needs anomalous processes. They generate anomalous transport coefficients. Presumably, in all collisionless astrophysical plasmas such processes are at work. Otherwise, the frozen-in property can also be broken by extraordinarily steep gradients in density, temperature, composition or magnetic field in the plasma. Their length scales must be of the order of the gyroradii of the particles involved. Compared with all astrophysical scales, these are enormously small (see table 2). Gradients on these scales are thus not only unobservable but probably also unrealistic. The main mechanism which leads to anomalous transport in a collisionless plasma is the turbulence.

Turbulence is generated via the excitation of plasma waves. Wave excitation proceeds via instabilities. There is an enormous number of possible instabilities in collisionless plasmas depending on the nature of the free-energy sources available, i.e. weak inhomogeneities, beams, jets, shear in field and velocity, electric currents, external forces exerted on the plasma and so on. Each of the instabilities excites a limited range of frequencies ω and wavevectors $\boldsymbol{k}$ which are mutually related by the dispersion relation $\omega = \omega(\boldsymbol{k})$.

Plasma waves

Because of its finite and relatively high temperature, the plasma is a thermally active medium. This means that the velocities of the particles constituting the plasma cover a wide range in velocity space, each of the particles being capable of absorbing or spending energy in interaction with an electric field fluctuation. In general, a particle that is moving at lower speed than the phase velocity of such fluctuations will be pushed by the wave and will absorb fluctuation energy. This leads to wave damping. On the other hand, a particle moving faster than the wave will push the wave and increase the fluctuation energy. Because of this nature it is clear that only a restricted number of wave modes can propagate in plasma. These are the modes that are very weakly absorbed by the plasma particles. The dispersion relations for the allowed waves can be found from an eigenvalue solution of the system of basic equations linearized around the equilibrium. Clearly, there will always be weak collisionless (or Landau) damping, for some of the particles will always be slow with respect to the wave and will absorb wave energy. However, in many cases this damping is so weak that it can be neglected.

When speaking of such wave modes or fluctuations it is not meant that the waves are all electromagnetic in nature as in ordinary electrodynamics. The property of the plasma of being an active medium implies that entirely new wave modes can evolve in the plasma body or at its surface. These modes are unknown in free space, and most of them cannot leave the plasma but are confined to the plasma. The large variety of such possible modes reaches from sound waves and purely electrostatic fluctuations on the electron and ion scales to all kinds of electromagnetic fluctuations, purely magnetic fluctuations of long wavelengths, and mixed magnetoacoustic waves.

The almost absolute invisibility of the majority of plasma waves looked at from a remote position as is common in astrophysics presents an enormous difficulty for astrophysics in inferring anything about the very nature of interactions in a collisionless astrophysical plasma. Any information must necessarily rely on secondary effects leading to the observation of radiation that can be detected on Earth. Since no *in situ* observations on any astrophysical object will ever be possible the true processes will forever be hidden to us. It is therefore of vital interest to infer information about the effects collisionless plasma processes may have on the radiation emitted from such objects. Unfortunately, very little is known about such effects at the current time.

Radiation

The most desirable effect is the direct emission of radiation in a certain wave band from that one may make

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Parameters of astrophysical plasmas.

Region or object	Density n (cm^{-3})	Temperature T (K)	Magnetic field B (G)	Plasma frequency f_{pe} (Hz)	Cyclotron frequency f_{ce} (Hz)
Inter galactic	$\leq 10^{-5}$	10^5–10^6	$\leq 10^{-6}$	<30	<2
Inter stellar	10^{-3}–10	10^2	$\geq 10^{-6}$	$3 \times (10^2$–$10^4)$	≥3
Pulsars					
surface	10^{12}	–	10^8–10^{13}	10^{12}	10^{16}–10^{18}
magnetosphere[a]	$<7 \times 10^{10}(r/R)^{-3}$	$\sim 10^{10}$	$\sim 10^{10}(r/R)^{-3}$	$3 \times 10^{18}(r/R)^{-3/2}$	$3 \times 10^{18}(r/R)^{-3}$
White dwarfs (atmosphere)	10^{10}	10^7	10^6	10^{20}	3×10^{12}
H I	10^{-3}	10^2	10^{-7}	3×10^2	3×10^{-1}
H II	10^2–10^3	10^3–10^4	10^{-6}	(1-3) $\times 10^5$	3×10^0
Planetary nebulae	10^3–10^5	10^3–10^4	10^{-4}–10^{-3}	$3 \times (10^5$–$10^6)$	$3 \times (10^2$–$10^3)$
Magnetosphere (stars)	10^{12}	10^4–10^6	10^3–10^5	10^{10}	$3 \times (10^5$–$10^7)$
Sun					
Chromosphere	10^{12}	2×10^4	10^3	10^{10}	(2-3) $\times 10^5$
Corona	10^8–10^{12}	10^6–10^7	10^{-3}–10^1	10^8–10^{10}	$3 \times (10^3$–$10^7)$
Solar wind at 1 AU	10^0–10^4	10^2–10^3	10^{-6}–10^{-4}	10^4–10^6	3–300
Earth					
Magnetosphere	10^{-2}–10^4	10^4–10^8	10^{-5}–10^{-1}	10^0–10^5	$3 \times (10^1$–$10^5)$
Ionosphere	10^3–10^6	10^2–4×10^3	10^{-1}	3×10^5–10^7	3×10^5

[a] R is the star radius.

inferences about the emission mechanism and hence the state of the collisionless plasma at the source. Here we are not interested in the ordinary *synchrotron* or *bremsstrahlung* (free–free) emissions. The former is emitted by energetic charged particles that gyrate in a magnetic field configuration. It allows us to map the remote magnetic fields in direction and magnitude. The latter, being emitted from very hot weakly *collisional* plasmas, permits the determination of the hot temperature and density at the location of the emitting source.

The important processes taking place in collisionless plasmas are without exception based on the assumption of instability, generation of plasma waves and interaction of these waves among themselves or with the particle constituents. There are only very few cases when these interactions directly lead to emission of radiation. Such cases are restricted mainly to two processes: the maser amplification of radiation in strongly magnetized very dilute plasmas containing a weakly relativistic trapped particle component, the so-called *electron cyclotron maser* mechanism, and radiation from fast electron beams passing through a dense, hot but otherwise relatively weakly magnetized plasma. Cyclotron maser emission is restricted to underdense plasmas satisfying the condition $\omega_{pe}/\omega_{ce} < 1$, which is satisfied only in strong magnetic fields. Since large-scale magnetic fields are usually weak in astrophysics, this condition restricts it to the vicinity of strongly magnetized objects. Examples of such processes are found in interplanetary space in the non-thermal radio emissions from the Sun and the planets. In astrophysics similar processes are expected to take place in the very strongly magnetized pulsar magnetospheres, the coronae of magnetized stars and possibly also in radio or x-ray jets ejected from massive objects.

Collisionless plasma processes play a key role also in a number of other effects such as the acceleration of small groups of charged particles into energetic beams and possibly plasma jets, the formation of collisionless shock waves, the magnetic reconnection process and the generation of anomalous transport. In order to describe these phenomena a brief discussion of the plasma instability and turbulence mechanisms follows.

Plasma instability and turbulence

In particular cases when the particle distribution function deviates from the Maxwell–Boltzmann distribution, the plasma undergoes instability. Because the plasma is thermally active, it contains a bath of spontaneously excited thermal noise of all wavelengths. At given temperature T the level of fluctuation never falls below this thermal level with the latter given by the Stefan–Boltzmann law. Decreasing the fluctuation level thus requires cooling of the plasma by means of adiabatic expansion or radiative losses in cases when the plasma is optically thin to the radiation or from its surface as in the case of thermal bremsstrahlung radiation.

Linear wave growth

The instability picks a particular range of wavelengths and frequencies out of the thermal noise and lets it grow in time and space. For one single mode of wavevector $\boldsymbol{k}_0$ out of this range, the energy density of the thermal noise level is $W_0 \equiv \frac{1}{2}\epsilon_0|\boldsymbol{E}_0|^2 + \frac{1}{2}|\boldsymbol{B}_0|^2/\mu_0$. Since the linear wave mode $[\boldsymbol{E}(t,\boldsymbol{x}), \boldsymbol{B}(t,\boldsymbol{x})] \propto \exp(-\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k}_0 \cdot \boldsymbol{x})$, the energy density in the mode grows with time as

$$W(\boldsymbol{k}_0, t) = W_0(\boldsymbol{k}_0)\exp[2\gamma(\boldsymbol{k}_0)t] \tag{12}$$

where, for instability, the growth rate $\gamma(\boldsymbol{k}_0)$ is the positive imaginary part of the wave frequency. It is the main goal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

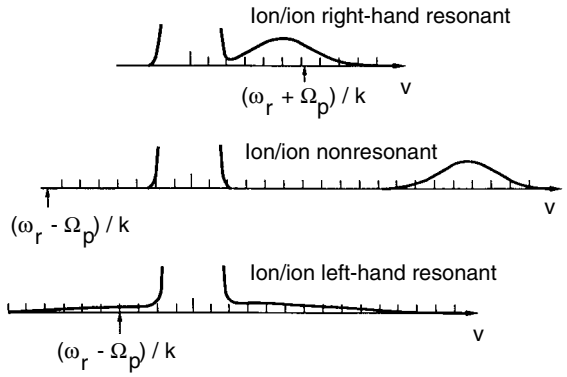


Figure 1. Reduced ion distribution functions corresponding to three electromagnetic ion–ion beam instabilities and associated cyclotron resonant velocities (after Gary, 1993).

of *linear plasma theory* to calculate $\gamma(\boldsymbol{k}_0)$. The growth rate itself is a function of the wave frequency, wavenumber and the properties of the free-energy source. The exponential growth of the wave given by the growth rate has a number of important effects. First, it readily depletes the energy of the free-energy source and transforms it into electromagnetic energy unless the source is continuously re-established. Second, the wave energy for exponential growth readily reaches a level that is comparable with the local kinetic energy level $n_0 k_B T$ of the plasma itself. Once this happens, linearity is violated, unless the wave energy is transported away by some process such as wave propagation or convection on a time scale faster than the wave grows. The plasma then enters the non-linear state.

An example of an electromagnetic instability occurring at collisionless shocks is the ion–ion beam instability. Whenever two ion components have a drift velocity v_d greater than some threshold, the corresponding free energy can lead to instability growth. Let us assume, for simplicity, that the two ion components, in the case of a shock the upstream plasma representing the core and the backstreaming ions representing the beam, are drifting Maxwellians. There are three electromagnetic instabilities that may arise from such a configuration: the right-hand resonant, the non-resonant and left-hand resonant ion–ion instabilities. Figure 1 shows schematically the ion distribution function (core and beam) corresponding to the three instabilities and the associated cyclotron resonant speeds. At low densities and low drift velocities the ion–ion right-hand resonant instability is the growing mode of lowest threshold. The waves propagate in the direction of the beam. In the limit of zero drift velocity this mode is the right-hand circularly polarized positive helicity mode on the magnetosonic whistler branch. In the case of large beam velocities and high beam densities a non-resonant instability is possible. When the beam has a large temperature, the left-hand instability is possible where the waves are resonant with those beam ions which have velocity (in the core frame) opposite to the beam velocity.

Nonlinear effects

In the non-linear state large-amplitude waves start decaying into sidebands, thereby broadening the spectrum. They may interact with other waves transforming energy into remote parts of $\boldsymbol{k}$-space. They may exert radiation pressure on the plasma, causing modulation of their propagation properties. This introduces ripples into the plasma. The waves can be trapped, reflected and become modulated in wavelength. They also may interact with the particle component. Low-energy particles in the wave frame can become trapped in the potential wells of the wave and transported away, thereby becoming accelerated by the wave until they are injected from the wave into the plasma as beamlets. Particles can be reflected from the wave and become accelerated to moderate and high energies in a kind of Fermi process. Some of these processes modify the properties of the collisionless plasma in such a way that anomalous collisions are caused.

The plasma containing a high level of the kind of turbulence described above presents a large number of hurdles for a test particle that tries to pass across the plasma. The particles will be scattered at waves, become trapped, becoming untrapped after a certain time, energized or decelerated. All these processes modify the free-particle orbit and lead to scattering of the particle in energy, path and pitch angle. In a more macroscopic description this process can be described as anomalous transport. All active collisionless plasmas thus exhibit anomalous transport properties. In each single case it is, however, extremely difficult to determine the kind of anomalous transport that is realized in the plasma under the particular conditions. In fact, the whole problem of anomalous transport remains to be solved. It is therefore safe to say that we do not know what the transport properties of collisionless astrophysical plasmas really are. Only in a few simple limits, in particular the quasilinear one, have such collisionless 'anomalous' transport coefficients so far been derived.

It should be added that on the long temporal and large spatial scales one encounters in astrophysical applications any plasma process will always have long reached the quasi-stationary and highly saturated non-linear state of fully developed turbulence. Since this is a natural fact one must be very cautious in applying linear plasma theory in astrophysics. Since a fully non-linear theory of plasma turbulence does not exist, this considerably obscures the astrophysical application of plasma physics.

Anomalous transport

The transport coefficients (resistivity, heat conductivity etc) entering the fluid description of plasmas have to be calculated within the fully developed turbulence theory. Within a first and lowest level approximation this is usually done for selected cases in the framework of quasilinear theory. The philosophy of this theory is straightforward.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fokker–Planck approach

Consider an instability that excites linearly growing waves of frequency $\omega(k)$ in a certain waveband of width $\Delta\omega$, Δk. The growth rate is $\gamma(\omega, k)$. Once these waves reach substantial amplitude their presence cannot be neglected any more energetically in comparison with other energy sources and in particular with the free energy driving the instability. In such a case the waves start reacting on the plasma and the free energy. The simplest effect is a quenching of the instability by the waves leading to saturation at a level that is determined by equation (12) with vanishing growth rate. The growth rate $\gamma[f(\boldsymbol{v}, t)]$ itself depends functionally on the modified distribution function $f(\boldsymbol{v}, t)$ which for its part evolves according to the Fokker–Planck equation. Writing this as

$$\partial_t f = \partial_v \cdot \mathbf{D}(W, \gamma, \boldsymbol{v}) \cdot \partial_v f \tag{13}$$

and knowing the explicit form of the velocity space diffusion coefficient $\mathbf{D}(W, \gamma, \boldsymbol{v})$, the expressions for W, γ, $\mathbf{D}$ together with the evolution equation (12) for $W(t)$ and the Fokker–Planck equation (13) form a closed system from which $f(t \to \infty)$ and $W(t \to \infty)$ can be determined. Once the final wave energy density $W(\infty)$ is known, it determines the anomalous collision frequency. This can be seen from the fact that the classical Coulomb collision frequency

$$\nu_c \propto \frac{\omega_{pe}}{n\lambda_D^3} = \frac{\omega_{pe} W_0}{n k_B T} \tag{14}$$

can be expressed in terms of the wave energy W_0 in the thermal fluctuations. Replacing W_0 with $W(\infty)$ gives the estimate (with α a numerical coefficient)

$$\nu_{\rm an} = \alpha \omega_{pe} \frac{W(\infty)}{n k_B T} \tag{15}$$

for the anomalous collisions.

Non-magnetized case

In non-magnetized plasmas (or in one-dimensional plasmas with the direction of current parallel to the magnetic field $\boldsymbol{B}$) the dominant current instability is ion acoustic wave turbulence. In this case one finds the Sagdeev formula

$$\nu_{\rm an,ia} = 0.01 \omega_{pi} \frac{v_d}{c_{ia}} \frac{T_e}{T_i}. \tag{16}$$

Here $\omega_{pi} = (m_e/m_i)^{1/2}\omega_{pe}$ is the ion plasma frequency, $c_{ia} = (k_B T_e/m_i)^{1/2}$ is the ion acoustic wave velocity and $v_d = |\boldsymbol{j}/en|$ is the current drift velocity. This formula demonstrates that the anomalous collision frequency can be large in dense plasmas of high plasma frequency carrying strong currents and containing hot electrons.

Magnetized case

In magnetized plasmas a variety of other wave modes may contribute to the generation of anomalous collisions and transport. Within the quasilinear approach the most important are those instabilities which are caused by macroscopic gradients in density, magnetic field and temperature leading to so-called drift instabilities. Their typical oscillation frequency is mediated by a combined motion of ions and electrons, the lower-hybrid resonance $\omega_{lh} \approx (\omega_{ce}\omega_{ci})^{1/2} = \omega_{ce}(m_e/m_i)^{1/2}$, and the anomalous collision frequency becomes

$$\nu_{{\rm an},lh} \approx \left(\frac{\pi}{2}\right)^{1/2} \left(\frac{r_{ci}}{4L_\nabla}\right)^2 \omega_{lh}. \tag{17}$$

Here, L_∇ is the relevant gradient scale, and r_{ci} is the thermal ion gyroradius. These collision frequencies can become very large, of the order of ω_{lh} itself. They can cause substantial plasma diffusion close to Bohm diffusion $D_B = k_B T_e/16eB$. Microscopic plasma turbulence is thus capable of substantially changing the transport properties of a collisionless plasma. Quasilinear theory may in many cases be superseded by other strongly non-linear effects, such as particle trapping in large-amplitude wave fields, modulation of the plasma by the 'radiation pressure' of the large-amplitude waves, coherent interactions and non-linear wave–particle interactions that are not included in the above theory. Moreover, in relativistic plasma radiation cooling of the electrons may modify the conditions and induce high-ion-temperature instabilities. These lead to low-frequency electromagnetic turbulence that results in plasma vortices and mixing effects.

Reconnection

MAGNETIC RECONNECTION (figure 2) is among the most important applications of anomalous transport. Reconnection describes the merging of two oppositely directed magnetic fields across the current sheet that separates them in highly conducting plasma. This problem is of marked importance for astrophysical plasma physics. Reconnection has two effects: it changes the magnetic field topology and it leads to bulk plasma acceleration.

In vacuum reconnection does not provide any difficulties as the magnetic flux can rearrange in arbitrary ways. However, in ideally conducting collisionless plasmas merging of two oppositely directed magnetic field configurations approaching each other is inhibited by the frozen-in condition. When the plasma is collisional, merging is made possible in some narrow region, the so-called diffusion region, by the diffusive (last right-hand) term in equation (11). The initially anti-parallel fields diffuse across the diffusion region and rearrange into an X-type configuration. The two formerly unconnected field lines become suddenly connected. What is much more important, however, is that plasmas to both sides of the current layer that before merging could not mix after reconnection find themselves on the same field line and will mix along the field even outside the diffusion

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

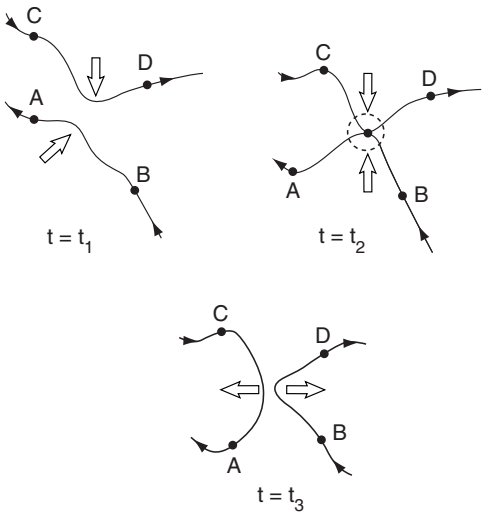


Figure 2. The schematics of the sequential change in magnetic topology and plasma acceleration during reconnection. The dashed circle surrounds the diffusion region. Plasmas initially (time $t = t_1$) on two separate field lines (A and C, B and D) are found on the same field line after reconnection took place ($t = t_3$). Field line stresses are released and converted into bulk plasma acceleration as indicated by the open arrows.

region. In addition, the strong bending of the newly connected field lines in the current sheet causes magnetic stresses that tend to relax and accelerate the bulk plasma up to flow velocities of the order of the Alfvén velocity, $v_a = B/(\mu_0 n m_i)^{1/2}$. In the so-called Petschek model of reconnection two standing slow-mode waves extend from the diffusion region and define the wedge-shaped high-speed flow region. Most of the acceleration actually occurs in these slow-mode shocks. The Petschek model is based on a fluid (MHD) picture. In a collisionless plasma it is not clear whether or not such slow-mode shocks actually form during the reconnection process. In the small diffusion region the Joule dissipation causes local heating of the plasma, and the electric field generated in the reconnection process further accelerates groups of particles up to high energies.

In a collisionless plasma a deviation from non-ideality can support reconnection in the diffusion region. Such a non-ideality may be provided by anomalous resistivities based on anomalous collision frequencies. This requires strong currents and/or plasma gradients until the current instabilities exceed the threshold. However, collisionless reconnection can also proceed without anomalous resistivity. Within the region close to the X-line the ions can be considered to be unmagnetized. The frozen-in condition for the electrons can be broken by the electron inertia and/or by a non-gyrotropy of the electron particle distribution function in the X-line environment.

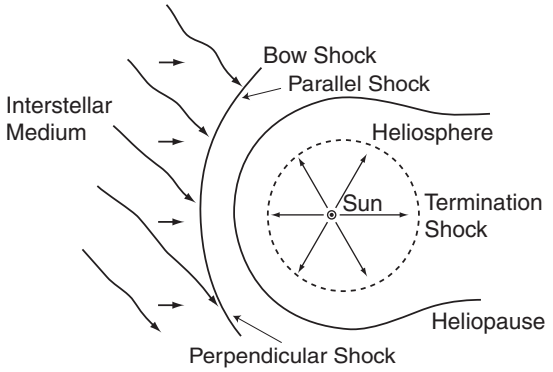


Figure 3. The heliospheric bow shock resulting from interaction of the heliospheric system with the interstellar medium: an example of a collisionless shock. In the upper part of the figure the shock is quasiparallel (magnetic field lines normal to shock); in the lower part it is quasiperpendicular (field lines tangential to shock). Inside the heliosphere a termination shock appears: the termination shock of the solar wind.

Collisionless shocks

Another important field of application of collisionless plasma theory is the dissipation in collisionless shocks. An example of a collisionless shock system is shown in figure 3. It consists of the termination shock of the solar wind resulting from the interaction between the highly supersonic solar wind stream and the interstellar medium. In addition a bow shock is produced outside but in front of the heliopause by its motion through the interstellar cloud.

Sub- and supercritical shocks

In this case the anomalous transport processes are of lesser importance. In a collisional plasma the width of the shock transition is of the order of the collisional mean free path. This is too large by many orders of magnitude in collisionless plasma where the shock transition is typically a few ion inertial lengths, c/ω_{pi}. The physics of collisionless shocks depends on essentially three parameters, the upstream magnetosonic Mach number $M = u/(v_A^2 + c_{ia}^2)^{1/2}$, the angle θ_{Bn} between the upstream magnetic field and the shock normal, and the chemical composition of the plasma. According the the Mach number one distinguishes subcritical ($M < M_c$) and supercritical ($M > M_c$) shocks. The critical Mach number $M_c = u_d/c_{ia,d}$ is defined as the ratio of the downstream velocity to the ordinary downstream (index d) sound speed. With respect to θ_{Bn} one distinguishes quasiperpendicular ($\theta_{Bn} > 45^\circ$) and quasiparallel ($\theta_{Bn} < 45^\circ$) shocks. For perpendicular shocks the critical Mach number is $M_c = 2.76$. It decreases with increasing angle θ_{Bn}. The chemical composition may lead to weakening of the shock profile.

In subcritical shocks the anomalous collisional dissipation in the shock front caused by micro-instabilities generated there is roughly sufficient for the required increase in entropy, heating and generating the turbulent disorder behind the shock that leads to dissipation of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

excess upstream energy stored in the flow. In supercritical shocks the flow energy supply to the shock from upstream is too fast. The instabilities in the shock front do not have sufficient time to evolve to a level sufficiently high for effective dissipation. In this case the shock front acts like a mirror and starts reflecting ever more ions and electrons back upstream into the onstreaming flow. In this case it becomes vital to distinguish between quasiperpendicular and quasiparallel shocks. This is important because in collisionless plasma the particles are tied to the magnetic field and can only move along the magnetic field. This implies that in quasiperpendicular shocks the reflected particles stay at the shock front or are convected back into the downstream region. The upstream width of the region containing the reflected particles is of the order of the average gyroradius of the reflected particles.

Quasi-perpendicular shocks

The specularly reflected and subsequently gyrating particles cause a foot-like signature in any *in situ* recording of $\boldsymbol{B}$. During this gyration the specularly reflected ions are accelerated in the $\boldsymbol{v} \times \boldsymbol{B}$ electric field to about twice the upstream bulk speed. Since these ions remain to be tied to the magnetic field, they become convected with the magnetic field into the downstream region. The region immediately behind the shock ramp contains the transmitted distribution plus the accelerated and transmitted specularly reflected ions.

The transmitted specularly reflected ions gyrate in bunches around the magnetic field lines. Such distributions are unstable with respect to the generation of plasma waves. They excite waves that in the downstream region lead to dissipation and plasma heating. In high Mach number shocks, as well as in shocks of supernova remnants, the interaction of specularly reflected ions with background electrons in addition excites ion acoustic waves which may give rise to rapid anomalous heating of the electrons in the shock ramp.

Quasi-parallel shocks

In the case of a quasiparallel shock the situation is dramatically different. Here the reflected particles can move far upstream along the magnetic field into the onflowing stream. Since at the same time the field is convected downstream toward the shock, this is true for particles having velocities along $\boldsymbol{B}$ higher than the speed of the flow.

Such fast particles form a counterstreaming beam in the flow and excite beam–beam instabilities in the otherwise undisturbed super-magnetosonic supercritical flow. These instabilities evolve into large-amplitude turbulence which is continuously convected towards the shock. The region filled with this turbulence is called foreshock. Since the foreshock is turbulent, energy is dissipated here, and the flow is slightly decelerated. Thus the foreshock effectively marks the grossly widened shock transition.

When the low-frequency upstream waves are convected with the flow into the shock they interact with upstream energetic particles and steepen into large-amplitude pulsations. The shock itself can be viewed as a patchwork of such large-amplitude pulsations which pile up and determine the downstream plasma state.

In the case of more parallel shocks the interaction of the upstream plasma with the hot downstream plasma in the region of overlap at the shock ramp also causes an ion–ion beam instability. The waves generated by this so-called interface instability are confined to the region downstream of the shock transition where they contribute to dissipation and heating. The wavelength of the interface waves increases with increasing shock Mach number. In the low to medium Mach number regime the interface waves have small wavelengths and are damped within a short distance downstream. At higher Mach number shocks the interface waves have large wavelengths and determine the downstream turbulence. In parallel shocks the downstream state is thus determined by the upstream turbulence. Quasiparallel shocks are thus highly non-stationary and oscillatory with the main dissipation provided by the shock interaction with the upstream waves in place of the microturbulence in the shock ramp itself. This shows that in collisionless plasmas the mechanism for shock formation is very different from that of collisional fluids.

Particle acceleration

An important field in collisionless plasma astrophysics is the mechanism of acceleration of particles to high energies. Collisionless shocks are among the main agents for accelerating ions as well as electrons to high energies whenever sufficient time is available. Particles reflected from the shock and from scattering centers behind the shock in the turbulent compressed region when coming back across the shock into the turbulent upstream region have a good chance to experience multiple scattering and acceleration by first-order Fermi acceleration. Second-order or stochastic Fermi acceleration in the broadband turbulence downstream of collisionless shocks will contribute to acceleration. In addition, ions may be trapped at perpendicular shocks. The trapping forces are provided by the electrostatic potential of the shock and the Lorentz force excerted on the particle by the magnetic and motional electric fields in the upstream region. With each reflection at the shock the particles gyrate parallel to the motional electric field, picking up energy and surfing along the shock surface. All these mechanisms are still under investigation, but there is evidence that shocks play a most important role in the acceleration of cosmic rays and other particles to very high energies.

In addition to acceleration in shocks one expects that particles are also accelerated in the reconnection region. Despite the shocks, collisionless plasmas are capable of generating large potential differences along the magnetic field. These are known as double layers or electrostatic shocks. In most cases the potential will not consist of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a continuous drop but will compose itself of myriads of little microscale or mesoscale potential drops confined to cavitons, solitary wave structures or phase space holes generated in non-linear wave–particle interaction. The plasma divides itself into several components which either support these structures or destroy them by transit time damping leading to acceleration.

Concluding remarks

There is no *in situ* information available from astrophysical plasmas. So one is forced to refer to indirect methods and analogies with accessible plasmas. Such plasmas are found only in near-Earth space. Actually, most of the ideas about and models of the behavior of astrophysical plasmas have been borrowed from space physics and have been rescaled to astrophysical scales. Measurements and models developed for the description of space plasmas often serve as paradigms for astrophysical plasmas. These models encompass stellar wind flows, pulsar magnetospheres, magnetospheres of magnetized stars and other planets including extra-solar planets, magnetospheres near black holes, the interaction of stellar winds with the surrounding INTERSTELLAR MATTER, expansion of SUPERNOVA REMNANTS, the whole domain of reconnection, collisionless shocks, plasma jets, and some non-thermal radiation mechanisms.

However, the large spatial and long temporal scales in astrophysics and astrophysical observations do not allow for the resolution of the collisionless state of the plasmas. For instance, in the solar wind the collisional mean free path is of the order of a few AU. Looked at from the outside, the heliosphere (figure 3), the region which is affected by the solar wind, will thus be considered collision dominated over time scales longer than a typical propagation time from the Sun to Jupiter. On any of the smaller and shorter scales this is wrong, because collisionless processes govern the solar wind here. Similar arguments apply to stellar winds, molecular clouds, pulsar magnetospheres and the hot gas in clusters of galaxies. One should thus be aware of the mere fact of a lack of small-scale observations. Collisionless processes generate anomalous transport coefficients. This helps in deriving a more macroscopic description. However, small-scale genuinely collisionless processes are thereby hidden. This implies that it will be difficult, if not impossible, to infer anything about the real structure, for instance, of collisionless astrophysical shock waves. Similarly, the processes of non-thermal radiation and acceleration, and the kinetic behavior of the gas in clusters of galaxies, are only accessible by exploiting the analogy with accessible collisionless plasmas in the close environment of the Earth and the heliosphere. From a theoretical point of view, collisionless plasma processes require the solution of the collision-free Boltzmann equation (Vlasov equation) combined with the full set of Maxwell's equations. This is a formidable task; it is often easier to perform particle-in-cell simulations of such processes where macro-ions and electrons are followed in time in the self-consistently calculated magnetic and electric fields. If only low-frequency processes are studied, the electrons are often approximated as a fluid and only the ions are treated as particles (hybrid simulation method).

Bibliography

The literature about collisionless astrophysical plasmas is sparse. Sporadic information can be found in the following proceedings:

Brinkmann W, Fabian A C and Giovanelli F (eds) 1990 *Physical Processes in Hot Cosmic Plasmas (NATO ASI Series C, vol 305)* (Dordrecht: Kluwer)

Epstein R I and Feldman W C (eds) 1986 *Magnetospheric Phenomena in Astrophysics (AIP Conf. Proc. 144)* (New York: AIP)

Fälthammar C-G *et al* (ed) 1988 *Plasma and the Universe* (Dordrecht: Kluwer)

Pallavicini R (ed) 1988 *Hot Thin Plasmas in Astrophysics (NATO ASI Series C vol 249)* (Dordrecht: Kluwer)

Sagdeev R Z (ed) 1993 *Nonlinear Space Plasma Physics* (New York: AIP)

Watanabe T, Kosugi T and Sterling A C (eds) 1998 *Observational Plasma Astrophysics: Five Years of YOKOH and Beyond* (Dordrecht: Kluwer)

Zank G P and Gaisser T K (eds) 1992 *Particle Acceleration in Cosmic Plasmas (AIP Conf. Proc. 264)* (New York: AIP)

More specialized monographs on plasma astrophysics deal with its radiational aspects, such as the books

Benz A O 1995 *Plasma Astrophysics* (Dordrecht: Kluwer)

Melrose D B 1980 *Plasma Astrophysics* vols 1 and 2 (New York: Gordon and Breach)

An excellent monography on pulsar magnetospheres is

Michel F C 1991 *Theory of Neutron Star Magnetospheres* (Chicago, IL: University of Chicago Press)

although the collisionless side of the coin is not treated explicitly there. Instead the particle and fluid pictures dominate. More about collisionless pulsar physics is contained in the controversial volume

Beskin V S, Gurevich A V and Istomin Ya N 1993 *Physics of the Pulsar Magnetosphere* (Cambridge: Cambridge University Press)

Reviews of collisionless space plasma physics are given by

Baumjohann W and Treumann R A 1996 *Basic Space Plasma Physics* (London: Imperial College Press)

Treumann R A and Baumjohann W 1997 *Advanced Space Plasma Physics* (London: Imperial College Press)

Linear collisionless plasma instabilities are discussed in

Gary S P 1993 *Theory of Space Plasma Microinstabilities* (Cambridge: Cambridge University Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

An excellent presentation of the basic kinetic collisionless plasma theory is found in

Montgomery D C and Tidman D A 1964 *Plasma Kinetic Theory* (New York: McGraw-Hill)

Nonlinear theory is described in

Davidson R C 1972 *Methods in Nonlinear Plasma Theory* (New York: Academic)

Rudolf A Treumann and Manfred Scholer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Colombo, Giuseppe ('Bepi') (1920–84)

Colombo was a mathematician and engineer, born in Italy, professor at the University of Padua. He explained the peculiar rotation of Mercury. The planet turns three times during two revolutions around the Sun, in a previously unsuspected resonance (see MERCURY). Using this feature, Colombo discovered the orbital maneuver that enabled Mariner 10 (see MARINER MISSIONS) to pass close to Mercury three times in 1974–75. He proposed this manoeuver to NASA, and tripled the value of the mission. Colombo is also famous for his development of the concept of space tethers. While space-based tethers have been studied theoretically since early in the 20th century, Colombo invented the idea of using a long tether to support a satellite from an orbiting platform in a gravitationally stabilized system. He designed the tethered satellite system (TSS) to investigate plasma physics and the generation of electricity in the upper atmosphere. Two technology tests of a tether system, TSS-1 and TSS-1R, were carried out in July 1992 and February 1996 by Italy using the NASA Shuttle. In the more recent test, the tether deployed to 20 km but broke without useful result. The ESA mission to Mercury scheduled for 2009 is named Bepi Colombo in affectionate recognition of his contributions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Columba

(the Dove; abbrev. Col, gen. Columbae; area 270 sq. deg.) A southern constellation which lies between Lepus and Pictor, and culminates at midnight in mid-December. It represents Noah's dove and was introduced by the Dutch theologian and geographer Petrus Plancius, who formed it from stars to the south of Canis Major and included it on a celestial chart in 1592.

A small, inconspicuous constellation, the brightest stars in Columba are α Columbae (Phact), magnitude 2.6, and β Columbae (Wazn), magnitude 3.1. There are three other stars of magnitude 4.0 or brighter. Interesting objects include μ Columbae, a famous 'runaway star', which is one of three stars (the other two being 53 Arietis and AE Aurigae) that are diverging at high velocity ($\sim$120 km s^{-1}) from a point in Orion from which they are thought to have been ejected as a result of a supernova explosion a few million years ago, and NGC 1851, a seventh-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Coma

The distortion of off-axis images produced by lenses or mirrors. Rays of light that enter a lens, or reflect from a mirror, at an angle to its optical axis (the line passing perpendicularly through the center of the lens or mirror), form pear-shaped images. The magnitude of the distortion increases with increasing distance from the optical axis. This particular optical aberration derives its name from the 'comet-like' appearance of such images.

A spherical mirror does not suffer from coma but does suffer from spherical aberration. If the latter is corrected, by deepening the curve of the mirror's surface to a paraboloid, coma is introduced. The effects of coma diminish in proportion to the square of the focal ratio and, for visual purposes, are of little consequence except close to the edge of the field of view in instruments with focal ratios of f:8 or more. Unless coma is corrected, its effects are very obvious in instruments with substantially shorter focal ratios and, in such cases, acceptable-quality images are restricted to the central part of the field of view. A parabolic mirror may be corrected for coma by placing a suitably shaped lens, or lenses (or mirrors), in front of its reflecting surface. Coma may be reduced in a compound lens by an appropriate selection of curvatures for its various surfaces.

See also: field of view, focal ratio, spherical aberration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Coma Berenices

(Berenice's Hair; abbrev. Com, gen. Comae Berenices; area 386 sq. deg.) A northern constellation which lies between Canes Venatici and Virgo, and culminates at midnight in early April. It is named after the lock of hair of Queen Berenice II of Egypt (*c*. 269–221 BC), which, according to legend, she gave to the gods for the safe return of her husband, Ptolemy III Euergetes, from battle. It was introduced by the Dutch cartographer Gerardus Mercator, who included it on a celestial globe in 1551.

A small, rather inconspicuous constellation, the brightest star in Coma Berenices is of magnitude 4.2. α Comae Berenices (Diadem) is an interesting close binary with pale yellow (F5) components, both of magnitude 5.2, separation ∼0–0.9″, period 25.8 years, the orbit lying almost edge-on, so the motion is practically along a straight line. Other interesting objects include Melotte 111, a wide (5°) open cluster of about 40 stars between fifth and tenth magnitude, M53 (NGC 5024), an eighth-magnitude globular cluster, M64 (NGC 4826, the Black-Eye Galaxy), a ninth-magnitude spiral galaxy, and NGC 4565, a tenth-magnitude edge-on spiral galaxy.

There are many galaxies in the southern part of Coma Berenices that belong to the great Virgo Cluster, and in the north-east of the constellation lies the distant Coma Cluster. The north galactic pole is also located in the constellation.

See also: Black-Eye Galaxy, Coma Cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Coma Cluster

Coma, with Virgo, is one of the best-studied GALAXY CLUSTERS. The cluster is located at right ascension $12^h59^m48.7^s$, declination $+27^\circ58'50''$ (J2000) and lies almost at the north Galactic pole, $(l, b) = (58^\circ, +88^\circ)$. In the major catalog of galaxy clusters by Abell, its designation is Abell 1656 (see ABELL CLUSTERS). The mean redshift of the cluster is approximately 6900 km s^{-1}, which puts it at a distance of $69h^{-1}$ Mpc (where $H_0 = 100h$ km s^{-1} Mpc^{-1}). Coma is one of the richest nearby clusters, having 650 confirmed member galaxies and, including dwarfs, probably as many as 2000 galaxies in total. It has two central, dominant galaxies and therefore is classified as a binary cluster.

Coma has long been taken as the archetypal rich cluster. It appears regular and roughly spherical, with a strong central concentration. The spherical symmetry and general regularity meant that it was considered to be a good (if rare) example of a cluster that had achieved dynamical equilibrium and was therefore amenable to straightforward theoretical analysis. Work over the last two decades, however, has gradually revealed that Coma is far from relaxed, with many substructures. The complexity of the dynamical system reveals clues to the process of cluster formation.

The total mass of Coma has been estimated from observations of its gravitational effects on the galaxies and hot x-ray gas in the cluster. Assuming that the cluster is in dynamical equilibrium, the total mass is found to be much greater than the observed mass in the galaxies and x-ray gas. This result, first obtained by ZWICKY in 1933, remains one of the strongest pieces of evidence that the universe is predominantly composed of some form of unseen DARK MATTER.

History

William Herschel was the first to note the concentration of nebulae in the constellation COMA BERENICES in 1785 (see figure 1). The first catalog of 108 nebulae in Coma was given by Wolf in 1902 and expanded by Curtis to over 300 objects in 1918. The recession velocities for a few members of the cluster were measured by Hubble and Humason in 1931.

The place of the Coma cluster as a touchstone of extragalactic astronomy and cosmology was established when Zwicky derived an estimate for the mass of the cluster in 1933 and found it to be at least $2 \times 10^{14}h^{-1}M_\odot$, which was much greater than could be accounted for by the visible matter. This was the first evidence suggesting that some form of dark matter existed in the universe.

Cluster galaxies

Richness

Coma has 276 member galaxies brighter than $B = 18$ within $0.75h^{-1}$ Mpc (0.67°) of the cluster center and approximately 900 members brighter than $B = 20$ out to $1.5h^{-1}$ Mpc (1.33°). The distance modulus for Coma is $34.2-5\log h$, so that $B = 20$ corresponds to $M_B = -14.2$; about half the members of the Local Group of galaxies are dwarf galaxies with absolute magnitudes fainter than this, so the total number of galaxies in Coma is about 2000.

Luminosity function

The number of galaxies as a function of luminosity has been determined in Coma at optical, infrared, ultraviolet and radio wavelengths. The optical and near-infrared luminosity function (LF) is not very well fitted by the standard Schechter functional form and appears to have three main regimes: the number of galaxies per unity luminosity increases rapidly between the brightest cluster member at $B = 12.6$ and $B \approx 16$ ($L \approx 3 \times 10^9h^{-1}L_\odot$), and then levels off to $B \approx 18$ ($L \approx 0.5 \times 10^9h^{-1}L_\odot$), before increasing as a power law, $N(L) \propto L^\alpha$, for the dwarf galaxies at fainter magnitudes. The LF varies with position in the cluster. In the central regions the LF in fact has a peak at around $B \approx 16.5$ and a dip at $B \approx 17.5$ before rising steeply. This may be due to mass (luminosity) segregation as the cluster approaches equipartition of energy in the center, or due to increased merging in the cluster core or due to different evolutionary histories for the central galaxies. The steepness of the power law for dwarf galaxies also depends on position in the cluster, going from $\alpha \approx -1.3$ in the core to $\alpha \approx -1.8$ on the periphery.

Morphologies

As in most clusters, the galaxies in the core of Coma are mostly ellipticals and lenticulars, while the outskirts have a higher proportion of spirals. For the brightest 200 galaxies in the cluster, the mix is approximately E:S0:Sp = 30%:55%:15%. The variation in the morphological mix is reflected in the colors of the galaxies, with red galaxies dominating in the center and blue galaxies on the periphery. There is an excess of early-type galaxies in the region between the NGC4839 subcluster and the main cluster that show evidence of recent star formation, which may have been triggered by interaction with the tidal field of the cluster or the intracluster medium (ICM) as these objects fell into Coma. More generally, the E and S0 galaxies in the cluster are strongly concentrated along a northeast–southwest axis while the spirals show a more diffuse and isotropic distribution.

Dominant galaxies

There are three dominant (D or cD) galaxies in the Coma cluster. Two of these (NGC4874 and NGC4889) lie in the cluster core and cause it to be classified as a binary cluster (Bautz–Morgan class II). NGC4874 ($B = 12.8$) is a cD galaxy with an extended halo and an absolute luminosity of $5.8 \times 10^{10}h^{-1}L_\odot$ that is located close to the peak of the distributions of galaxies and x-ray gas. NGC4889, 0.12° east of NGC4874, is even brighter ($B = 12.6$, $6.9 \times 10^{10}h^{-1}L_\odot$), although it lacks an extended halo. Both NGC4874 and NGC4889 are estimated to have masses (within $80h^{-1}$ kpc) of $1.4 \times 10^{13}h^{-1}M_\odot$. NGC4874 is a strong radio source but NGC4889 is not. The third dominant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

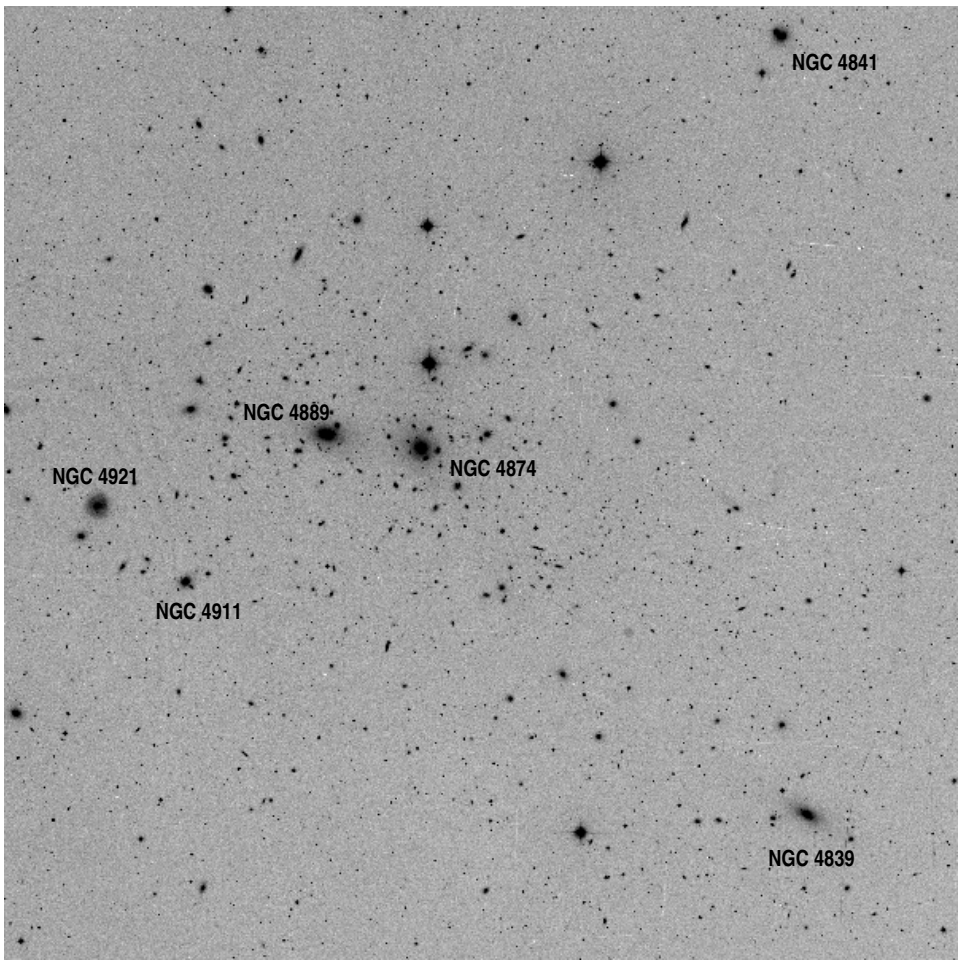


Figure 1. The visual appearance of the central region of the Coma cluster. This image, from the Palomar Observatory Sky Survey, is centered at $12^{\rm h}59^{\rm m}15^{\rm s} + 27^\circ55'00''$ (J2000) and covers $1.2^\circ \times 1.2^\circ$; north is at the top, east is at the left. It shows the core of the cluster around NGC4874 and NGC4889 and the southwestern group around NGC4839. (Credit: National Geographic Society, California Institute of Technology, Space Telescope Science Institute.)

galaxy is NGC4839 ($B = 13.5$), which lies well outside the cluster core, 0.67° away to the southwest of NGC4874. It is another cD galaxy, with luminosity $3.0 \times 10^{10} h^{-1} L_\odot$. It is also a head–tail radio source, with the tail pointing away from the cluster center. Each of the three dominant galaxies is the leading member of a subcluster within Coma (see below).

Dynamics

REDSHIFTS have been measured for 650 cluster members. The line-of-sight velocity distribution within the cluster can be resolved into two Maxwellian (Gaussian) distributions. One of these is the main body of the cluster, with a mean velocity of 6900 km s^{-1} and a one-dimensional velocity dispersion of 1100 km s^{-1}. The second is the group of galaxies around NGC4839, with a mean velocity of 7300 km s^{-1} and a dispersion of 300 km s^{-1}. The E and S0 galaxies which dominate the main cluster have a Maxwellian velocity distribution, indicating that they are approximately in dynamical equilibrium. However, the spiral galaxies in the cluster have a broader and flatter velocity distribution, which suggests that they are freely falling onto the cluster.

Intracluster medium

Coma has long been known to be a strong source of x-ray emission. The total x-ray luminosity of the cluster is approximately $2.5 \times 10^{44} h^{-2}$ erg s^{-1}. The x-ray emission comes from the ICM, a hot (8 keV $\approx 9 \times 10^7$ K) plasma with a mass at least as large as the mass in galaxies. Figure 2 shows the deepest x-ray image of the whole Coma cluster, taken with the ROSAT satellite. The cluster appears fairly regular apart from the obvious second peak in the x-ray emission centered on NGC4839, which has a luminosity of $10^{43} h^{-2}$ erg s^{-1}, typical of a bright group of galaxies. The temperature of this subcluster is higher than expected given its luminosity, suggesting that it has been shock-heated by its interaction with the cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

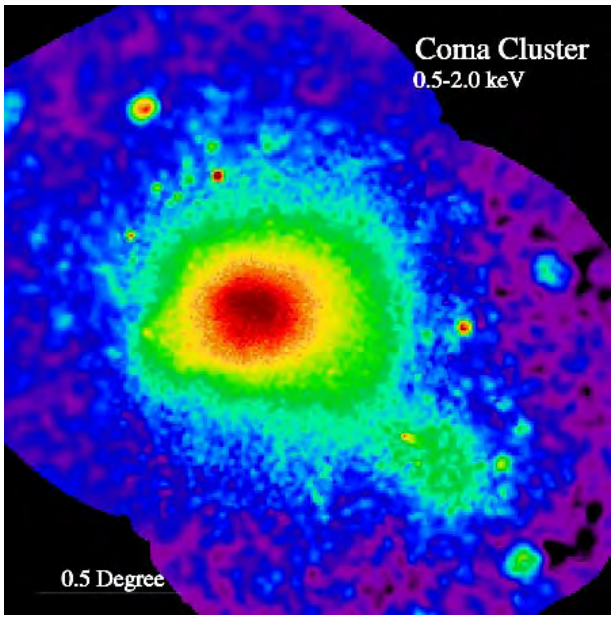


Figure 2. An x-ray image of the Coma cluster obtained with the ROSAT satellite, showing both the main cluster and the NGC4839 group to the south-west. (Credit: S L Snowden, High Energy Astrophysics Science Archive Research Center, NASA.)

Detailed analysis of the x-ray gas distribution shows that there is a double core in the cluster, with peaks associated with NGC4874 and NGC4889, surrounded by a more symmetrical diffuse component. The luminosity and temperature of the gas in these peaks are more consistent with a rich group or poor cluster than with individual galaxies. Unlike many other clusters, Coma does not possess a central cooling flow. This presumably is due to the disruption of the central regions associated with the interaction between the NGC4874 and NGC4889 subclusters. As well as the double core, there is also a distinct 'tail' of x-ray emission running outwards from the cluster center, passing through NGC4911 and ending near NGC4921. This feature has a luminosity of $10^{42}h^{-2}$ erg s^{-1} and an estimated gas mass of a few times $10^{11}h^{-1}M_{\odot}$. It may be associated either with another subcluster around NGC4911 or with the passage of the NGC4874 group through the cluster.

The temperature of the cluster generally decreases out from the center but also shows hot and cool regions. Initial maps of the x-ray gas temperature show evidence for variations associated with the substructures in the cluster. Mapping the temperature of the cluster with future x-ray satellites such as AXAF will greatly improve our understanding of the conditions in the ICM and the dynamics of the cluster. Images of the cluster in the extreme ultraviolet obtained with the EUVE satellite indicate that there is also a substantial amount of sub-10^6 K gas in Coma.

Coma is one of only 10 clusters possessing a diffuse radio halo (the 'Coma C' source); none of these clusters has a COOLING FLOW. This halo is probably produced by the *in situ* acceleration of relativistic electrons by a large-scale magnetic field. The conditions required to create a persistent cluster-wide radio halo have probably been produced by the on-going merger between the NGC4874 and NGC4889 subclusters, although the mechanism by which the kinetic energy of the merger is converted into radio emission remains unclear.

Structure

Detailed analysis of the distribution and dynamics of the galaxies and x-ray gas shows that there are at least four subclusters in Coma. The NGC4874 and NGC4889 subclusters are in the final stages of a merger at the core of the cluster, while the NGC4839 subcluster is either entering Coma for the first time from the direction of the nearby cluster Abell 1367 or is returning from that direction having already made its first passage through the cluster core. The fact that NGC4839 has a radio 'tail' pointing away from the main cluster indicates that the subcluster is inward bound and moving through a fairly dense ICM. Another subcluster has been identified around NGC4911, and there is some evidence for other, lower-level, structures.

There is a significant relation between structures on scales ranging from $10h^{-1}$ kpc up to $100h^{-1}$ Mpc. The major axes of the dominant galaxies and the distribution of E and S0 cluster galaxies and the x-ray gas are all approximately aligned with the filamentary structure (the 'GREAT WALL') running through Coma southwest to Abell 1367 at $cz = 6500$ km s^{-1} and northeast to Abell 2197/2199 at $cz = 9100$ km s^{-1}. The inference is that the cluster has been built up predominantly of material, both subclusters and individual galaxies, which have fallen in along this filament.

Mass and matter content

Ignoring the substructure and treating Coma as a spherical system with both the galaxies and x-ray gas in hydrostatic equilibrium, the total mass of the cluster is estimated to be $(3.1 \pm 0.5) \times 10^{14}h^{-1}M_{\odot}$ inside $1h^{-1}$ Mpc and $(6.5 \pm 2.5) \times 10^{14}h^{-1}M_{\odot}$ inside $3h^{-1}$ Mpc. Even within the limited context of these assumptions this is a wide range. For determinations of cluster mass using galaxy dynamics, this is partly due to the unknown anisotropy in the galaxy orbits (whether, at a given radius, the galaxies are predominantly on isotropic, radial or circular orbits); for determinations based on x-ray observations, the uncertainty lies in the unknown temperature distribution. As well, there is the unknown distribution of the dark matter which dominates the total mass of the cluster.

All current observations are consistent with a model in which 'light traces mass' (i.e. the dark matter has the same distribution as the visible matter), the galaxies are on isotropic orbits and the cluster has a uniform temperature. However, they are also consistent with models which depart significantly from this simplest case. The best-fitting models suggest that the ICM and galaxies have somewhat more extended distributions than the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

dark matter. The core assumption of all these models, namely that Coma is in hydrostatic equilibrium, is clearly violated at some level. The extent to which the observed substructure invalidates the conclusions of the equilibrium models is not yet clear, although early indications are that the mass estimates are reasonably accurate.

Zwicky's 1933 estimate of the V-band mass-to-light ratio (M/L) in the Coma cluster was about $100hM_\odot/L_\odot$; today's mass estimates correspond to values in the range 300–$400hM_\odot/L_\odot$. This exceeds M/L for individual galaxies by a factor of 10 or more, implying that there is of order 10 times as much mass in dark matter as there is in galaxies. However, the cluster M/L falls well short of the M/L of a critical-density ($\Omega = 1$) universe, for which $M/L \approx 1600hM_\odot/L_\odot$. The ratio of the luminous baryonic mass (i.e. the mass in galaxies and x-ray gas) to the total cluster mass is about 0.15 inside $1h^{-1}$ Mpc and 0.2–0.4 inside $3h^{-1}$ Mpc. This is considerably higher than the baryon fraction $\Omega_B h^2 = 0.012 \pm 0.002$ that is predicted by big bang nucleosynthesis for an $\Omega = 1$ universe (even for $h = 0.5$). Taken at face value, the mass-to-light ratio and baryon fraction of Coma both suggest a low-Ω universe.

Bibliography

Comprehensive and up-to-date reviews on all aspects of the Coma cluster, with references to the literature, can be found in

Mazure A, Casoli F, Durret F and Gerbal D (ed) 1998 *Untangling Coma Berenices: a New Vision of an Old Cluster* (Singapore: World Scientific)

The discovery of the Coma cluster is described in

Herschel W 1785 *Philos. Trans. R. Soc. London* **75** 213

The classic paper measuring the mass of the Coma cluster is

Zwicky F 1933 *Helv.Phys.Acta* **6** 110

Matthew Colless

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Coma, Cometary

The roughly spherical, temporary atmosphere of gas and dust expelled from a comet's nucleus as a result of solar heating when the comet is sufficiently close to the Sun. The nucleus and coma are together known as the comet's *head*. The coma, which is typically drawn into a teardrop shape by the solar wind, contains neutral and ionized gas molecules which shine by fluorescence and reflected sunlight. The coma typically develops at 3–4 AU from the Sun and can grow to 10^5–10^6 km, reaching its maximum extent just after the comet's perihelion passage.

See also: cometary atmospheres.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Combustion

The process whereby a substance is combined with oxygen with the production of heat. Burning is a familiar example of this process. The energy required to propel chemical rockets is provided by the combustion of fuel with an oxidant at very high temperatures. A common oxidant is liquid oxygen (often denoted by LOX). Others include hydrogen peroxide and nitrogen tetroxide.

See also: kinetic theory of gases.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Arend–Roland (C/1956 R1)

A long-period comet discovered jointly by Sylvain Arend and Georges Roland in November 1956. It reached perihelion (0.32 AU) on 8 April 1957, and passed closest to Earth (0.57 AU) on 21 April. The comet attained a peak brightness of magnitude 0, and developed a dust tail that was 30º long at its greatest extent, but it was most notable for its antitail, which at best was half the length of the main tail and appeared as a pronounced spike extending from the head in the opposite direction to the tail. Anti-tails are an effect of perspective: when the Earth is close to the comet's orbital plane and the comet is heading in the general direction of the Earth, the curvature of the dust tail makes the tail visible on either side of the head. The inclination was 120º and the eccentricity given as 1, making the aphelion distance and period indeterminately large.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Bennett (C/1969 Y1)

A long-period comet discovered by Jack Bennett in December 1969. It passed closest to Earth on 26 March 1970, having reached perihelion (0.54 AU) on 20 March, and was then at its brightest, at magnitude 0. The comet was best placed for observation in April, when its tail reached around 11°, some estimates putting it at 25°. Jets of material emitted by the nucleus spiraled out as the nucleus rotated. Bennett was one of the first comets for which a vast hydrogen cloud—surrounding the head at 13 million km, around 10 times the Sun's diameter—was detected, by ultraviolet observations made from Earth orbit. The eccentricity was 0.995, inclination 90°, period around 1700 years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Biela (3D/Biela)

A periodic comet that disintegrated in the nineteenth century and gave rise to the briefly active Bielid meteor shower; the 'D' in the prefix indicates 'defunct'. The comet was discovered in 1772 by Jacques Montaigne, and recovered by Jean-Louis Pons in 1805, but on neither occasion was it observed for long enough for an orbit to be computed. This was accomplished by Wilhelm von Biela when he recovered the comet in February 1826. He was able to follow it for 72 days, establishing its identity with the comets of 1772 and 1805, and it thus became the third known periodic comet. At its 1845/6 return the comet was observed to have split into two nuclei, one brighter than the other, separated by around 300 000 km. Both nuclei were seen to return separately in 1852, differing more in brightness and now over 2 million km apart. That was the comet's final apparition: there was no sign of either nucleus on what should have been the next favorable return in 1865/6, but in 1872 a meteor storm occurred, suggesting that the comet had completely disintegrated. There were further notable displays of what became known as the Bielid meteor shower at intervals corresponding to the comet's period of just over 6.6 years. The comet's disintegration into a meteor stream provided strong confirmation of the relationship between comets and meteors, first proposed by Giovanni Schiaparelli in 1866. At their final appearance, the twin nuclei (designated 3D-A and 3D-B) had inclination 13°, eccentricity 0.76 and period 6.62 years.

See also: Bielids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet de Chéseaux (C/1743 X1)

A long-period comet discovered by Dirk Klinkenberg on 9 December 1743, and independently by Philippe de Chéseaux four days later. It was also known as Klinkenberg's Comet, but it was de Chéseaux who computed its orbit and his name is now more commonly associated with it. This was one of the brightest comets of modern times, reaching an estimated magnitude -7 in February 1744. Its multiple tails—more than 10, according to some reports—formed a great fan that was visible in the pre-dawn sky before the comet itself rose. The comet reached perihelion (0.22 AU) on 1 March. Its inclination was 47°, and its eccentricity indistinguishable from 1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Donati (C/1858 L1)

A long-period comet discovered by Giovanni Donati in June 1858. It reached its greatest magnitude of −1 shortly before perihelion (0.58 AU) on 20 September, and was closest to Earth (0.5 AU) on 9 October. Its appearance was impressive, with a prominent curved dust tail stretching an estimated 60°, and two thin gas tails. Concentric, sharply defined shells of material ahead of the nucleus suggested that material was being ejected mainly from one vent when the rotation of the nucleus carried the vent round to the sunward side; the rotation period was estimated from the separation of the shells to be 4.6 h. The eccentricity was 0.996, the inclination 117°, and the period roughly 2000 years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Encke (2P/Encke)

The short-period comet with the shortest known period at around 3.3 years. It was discovered independently by Pierre Méchain in 1786, by Caroline Herschel in 1795, and by Jean-Louis Pons in 1805 and again in 1818. In 1821 Johann Encke calculated the orbit of Pons's 1818 comet and predicted its return to perihelion in 1821 to an accuracy of one day. Two years later he established this comet's identity with those observed in 1786, 1795 and 1805, and Encke's Comet, as it was known, thus became the second known periodic comet. Its early discovery and short period give it the record for the most observed returns to perihelion (0.33 AU), the 60th occurring in 2000. The comet seems to have been in a stable orbit for several thousand years, which poses the problem of why the historical record contains no sightings of it before 1786. The comet is now quite faint (sixth magnitude) and shows only a short tail, presumably having shed most of its dust and gas. However, at fourth magnitude it was a naked-eye object at several returns in the early nineteenth century. The orbit shows signs of slow decay, the present period being 3.28 years; its aphelion distance is 4.06 AU, the eccentricity is 0.85 and the inclination 12°. Debris from the comet forms a meteor stream which the Earth intersects twice a year, producing the Taurid shower in November and the daytime Beta Taurid shower in June.

See also: Taurids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Family

A group of comets with broadly similar orbital elements. The distribution of aphelia of members of a comet group peaks at a value corresponding to the distance of a particular major planet from the Sun. Perturbations by that planet have been responsible for changing the diverse original orbits of the comets into those they now occupy. It was once thought that all the major planets possessed comet families. Only Jupiter's comet family is now recognized as real, those of Saturn, Uranus and Neptune being explained in terms of orbital resonances with Jupiter. Jupiter's comet family itself contains over two-thirds of all the comets officially designated as being of short period. An example of a comet recently captured by Jupiter into its family is 16P/Brooks 2, which in 1886 passed just two Jupiter radii from the planet, an encounter that changed its period from about 30 years to 7 years. The orbits of Jupiter's comet family are continually decaying, and members are eventually ejected from the solar system by a gravitational 'slingshot', or tidally disrupted. Fresh captures replenish the stock.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Giacobini–Zinner (21P/Giacobini–Zinner)

A short-period comet discovered independently in 1900 by Michel Giacobini and in 1913 by Ernst Zinner. With a period of 6.61 years it is, like the majority of short-period comets, a member of Jupiter's comet family. It reached its greatest magnitude of 7 when it made its closest approach to Earth in 1946, passing at a distance of 0.26 AU. A brief outburst raised its magnitude to 6 for a few days, and similar short outbursts have been observed at other returns. These outbursts may be linked to storm-level activity in the comet's associated meteor shower, known as the Giacobinids, as occurred in 1946. The perihelion distance is 1.03 AU, aphelion distance 5.99 AU, eccentricity 0.71 and inclination 32°. In 1985 the comet became the first to be visited by a space probe when the International Cometary Explorer passed within 7800 km of its nucleus. A shock front was detected where the ionized species in the coma encountered the solar wind.

See also: comet family, Giacobinids, International Cometary Explorer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Hale–Bopp (C/1995 O1)

Comet C/1995 O1 (Hale–Bopp) was discovered simultaneously by the two American amateur astronomers Alan Hale and Thomas Bopp on 23 July 1995. At that time, it was at 7.1 AU from the Sun and had already a total visual magnitude of 10, which is more than 100 times brighter than COMET HALLEY at the same distance. It passed perihelion on 1 April 1997, at 0.91 AU from the Sun. The orbit of comet Hale–Bopp is inclined by 89º over the ecliptic. Its orbital period was about 4200 yr before entering the inner solar system and changed to 2400 yr after gravitational perturbation by Jupiter. It is thus a long-period comet, presumably coming from the OORT CLOUD.

As soon as comet Hale–Bopp was discovered and its orbit determined, it was clear that it would be a very bright object at perihelion, one of the brightest ever recorded comets. Indeed, it reached a total visual magnitude of −1 around perihelion. With favorable viewing conditions, it was a naked-eye object in the northern hemisphere for more than 2 months. At that moment, its gas production reached 10^{31} water molecules s^{-1} (about 300 t s^{-1}, 10 times the rate of comet Halley). It may thus be counted as one of the 'Great comets' of our epoch.

The campaign of observations of comet Hale–Bopp

The early discovery of comet Hale–Bopp gave ample time to organize world-wide observations with major instruments. Comets Hyakutake and Hale–Bopp were the first comets of exceptional brightness (surpassing the apparitions of comet West in 1976 and P/Halley in 1986) to be observed with the powerful means of astronomy at the end of the 20th century: the versatile imaging possibilities of CCDs, the new spectroscopic ground-based facilities in the infrared, millimetric and submillimetric domains, and space observatories such as the Infrared Space Observatory (ISO), the Hubble Space Telescope (HST) and x-ray satellites (see COMET HYAKUTAKE for a discussion of x-ray observations of comets). It is thus not a surprise that a wealth of new results followed.

Coma dust patterns, nucleus rotation and size

A complex pattern of dust jets (up to seven) was seen in the inner coma (see COMA, COMETARY). These jets showed either spirals or a radial pattern, depending on the viewing geometry of the moment. Close to perihelion, a pattern of expanding shells was conspicuous even for observers with modest-size instruments. This phenomenon traces the unequal activity of the nucleus surface: 'active regions' are periodically exposed to solar heating as the comet nucleus is rotating. The observation of the time evolution of the jet (or shell) pattern allowed a period of 11.4 h for the rotation of the nucleus to be determined precisely and its axis to be assessed.

A nucleus diameter of 40–80 km was determined from observations with the Hubble Space Telescope, the Infrared Space Observatory and ground-based radiometry at millimetric wavelengths. This is the largest COMETARY NUCLEUS among those for which the size could be assessed (comet P/Halley had an elongated nucleus of 'only' 8 km × 8 km × 16 km).

The sodium tail

One of the highlights of the observations of comet Hale–Bopp was the detection of a third kind of COMETARY TAIL (in addition to the dust and ionic tails): a thin, straight tail revealed by the yellow D lines of atomic sodium at 589 nm, which shows up at an angle from the other straight ionic tail and distinct from the broad, curved dust tail. Emission of the sodium lines had already been observed in a number of comets, with hints for a sodium tail suggested in a couple of them, but this was the first time that the presence of the sodium tail was so conspicuous. The sodium D lines are emitted through fluorescence: photons from the Sun are absorbed and re-emitted isotropically; a transfer of momentum results, accelerating the sodium atoms away from the Sun at velocities which could exceed 80 km s^{-1}. The origin of the cometary sodium is ill understood. Decomposition of a gas-phase parent molecule seems to be ruled out. Sodium is rather released from cometary dust particles.

Evolution with heliocentric distance

The early detection of the comet allowed observers to study the evolution of this exceptional comet over a very large span of heliocentric distances and to see the 'turning on' of the sublimation of various molecules from the nucleus ices as the comet approached the Sun. Just after discovery, at 7 AU, carbon monoxide (at radio wavelengths) and the CN radical (in the visible spectrum) were already identified. Water (traced by the OH radical observed in the near UV or at radio wavelengths) showed up at only 4.8 AU. Cometary activity, which is driven by water ice sublimation at short heliocentric distances, is governed by the sublimation of more volatile species, especially carbon monoxide, at long distances. As the comet approached closer to the Sun, the apparitions of methanol (CH_3OH), hydrogen cyanide (HCN), hydrogen sulfide (H_2S), formaldehyde (H_2CO) and methyl cyanide (CH_3CN) were successively observed (see figure 5 of COMETARY ATMOSPHERES).

The same pattern was observed in reverse order as the various molecules 'turned off' when the comet receded from the Sun. In 2001, at about 14 AU from the Sun, comet Hale-Bopp is still observed to be active. Images obtained at the MPG/ESO 2.2 m telescope at La

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Silla still show the prominent, curved jet-like structure in the coma that was observed earlier, with a total size of the coma of around 2 million km. Carbon monoxide is also still strongly outgassed, as seen in radio spectroscopic observations at the SEST 15 m radio telescope, also at La Silla.

New molecules

An extensive spectroscopic investigation of comet Hale–Bopp at all wavelengths, using state-of-the art instruments, resulted in the detection of many new molecular species. Comet Hale–Bopp is thus now the comet for which the chemical composition is the best known (see COMETARY ATMOSPHERES). The list of cometary molecules is not closed, however, since many unidentified spectral features were observed.

The composition of volatiles in comet Hale–Bopp is found to be strikingly similar to the chemical abundances of interstellar ices (as recently observed with ISO) and to those observed in interstellar hot molecular cores and bipolar outflows around protostars.

Radio observations

In addition to the already-known cometary parent molecules (H_2O, CO, CO_2, CH_3OH, H_2CO, HCN, H_2S) and to the ones just discovered in comet Hyakutake (NH_3, HNCO, CH_3CN, HNC, OCS), several molecules were identified by radio spectroscopy (see figure 4 of COMETARY ATMOPHERES): the relatively complex organic molecules formic acid (HCOOH), acetaldehyde (CH_3CHO), methyl formate ($HCOOCH_3$) and formamide (NH_2CHO), the nitrile HC_3N, thioformaldehyde (H_2CS), all with abundances of the order of 10^{-4} relative to water. The sulfur oxides SO and SO_2 were observed with abundances of about 0.3%, making them an important sulfur repository. The NS radical was discovered, the origin of which (from the nucleus ices or from an unknown parent molecule) is still a puzzle.

Several ions were also observed for the first time by radio spectroscopy: CO^+, HCO^+, H_3O^+.

Deuterated water (HDO) was observed from one of its submillimetric transitions, with a D/H ratio of 3×10^{-4}, as in comets 1P/Halley and Hyakutake. The deuterated variety DCN of hydrogen cyanide was also observed with D/H of 2.5×10^{-3}. As discussed in COMET HYAKUTAKE, these values are intermediate between the cosmic D/H ratio and the ratios observed in interstellar molecules and provide a critical test for formation theories of the solar system.

Infrared observations

Ground-based spectroscopic observations in the infrared with the Cryogenic Echelle Spectrometer (CSHELL) on the NASA Infrared Telescope Facility (IRTF) in the 2.8–4.8 μm region allowed the observation of CO, CH_4, C_2H_2, C_2H_6, OCS, NH_3, of several radicals through their fundamental bands of vibration and of water through 'hot' bands which are not affected by telluric absorption.

The apparition of comet Hale–Bopp also coincided with the operation of the Infrared Space Observatory which could, for the first time, secure a cometary spectrum in the infrared range from 2.5 to 190 μm. The water molecule, which is so difficult to observe from the ground because of telluric absorption, could be observed in many details through its fundamental bands of vibration at 2.7 and 6.5 μm wavelengths and its rotational lines in the far infrared (180 μm). Carbon dioxide (CO_2) was observed through its vibrational band at 4.25 μm: this was the second comet in which this molecule (which cannot be investigated from the ground because of strong telluric absorption) was directly observed, after P/Halley where it was observed *in situ* with the infrared spectrometer aboard the VEGA probe. Its abundance was about 20% relative to water (at a distance of 2.9 AU from the Sun when the ISO observation was made), which makes it a major cometary volatile.

Distributed sources of molecules

Previous to observations of comet Hale–Bopp, it was already known that some 'parent' molecules were not coming from the sublimation of nucleus ices, but from a still ill-known 'distributed source' within the coma. This was demonstrated for CO and H_2CO by, e.g. *in situ* observations of P/Halley by the Giotto space probe.

Infrared long-slit spectroscopic observations of comet Hale–Bopp showed that CO was produced half from a distributed source and half from the nucleus when the comet was at less than 2 AU from the Sun. A distributed source for OCS was also found. The Institut de Radioastronomie Millimétrique (IRAM) interferometer on Plateau de Bure (France) and the Berkeley Illinois Maryland Array (BIMA) and the Owens Valley Radio Observatory (OVRO), both in California, could map at radio millimetric wavelengths the radio emission of some molecules, using a technique successfully applied before for comet Hyakutake. This allowed the space distributions of some species such as CO, H_2CO, SO, HNC, HCO^+ to be investigated and clues to their origin to be obtained.

The exact nature of these 'distributed sources' of molecules is still an enigma. It seems likely that they could be the organic grains observed *in situ* by the space probes which explored comet Halley (the so-called 'CHON particles'), but the desorption mechanism of molecules from these grains is still to be elucidated.

The HNC molecule (an unstable isomer of hydrogen cyanide HCN) was found to have an abundance ratio relative to HCN varying from 0.03 to more than 0.20 as the comet approached the Sun. This is not expected if HNC were well mixed with HCN in cometary ices. A possibility is that HCN is converted to HNC through charge-exchange chemical reactions within the coma, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

reaction rates becoming more important when the density is higher, i.e. when the comet is more productive, closer to the Sun. Interferometric observations at IRAM indeed suggested that HNC has a distributed source.

Characterization of cometary dust

It was already known that cometary dust contained silicates from ground-based observations of their characteristic band around 10 μm and that part of them could be crystalline in some comets. The record of the full infrared spectrum of comet Hale–Bopp by ISO immediately revealed a series of additional spectral features at 16, 19.5, 23.5, 27.5 and 33.5 μm which were identified to be magnesium-rich olivine (forsterite Mg_2SiO_4), a special kind of silicate (see figure 6 of COMETARY ATMOSPHERES). A more quantitative analysis of the ground-based and ISO spectra of comet Hale–Bopp showed that its dust was essentially composed of forsterite, pyroxenes (another class of silicates such as $MgSiO_3$), amorphous silicates and spectrally featureless material which could be the organic-coated grains previously observed by the space probes which explored comet Halley.

The spectrum of comet Hale–Bopp in this spectral domain closely resembles that of the dust disks around some young or even evolved stars, such as that of HD 100546, an intermediate star between β Pictoris or Vega-like objects and Herbig Ae–Be stars, which were also observed by ISO. Thus, the dust contained in such disks appears to be similar to cometary dust.

In contrast, interstellar dust only shows silicates which are in the amorphous phase. Conversion of silicates from the amorphous to the crystalline phase, to our knowledge, can only occur through an annealing process at relatively high temperature (>1000 K). We still have to understand how comets such as Hale–Bopp could have incorporated apparently unprocessed interstellar volatiles on the one hand and evolved, crystalline silicates on the other hand.

Bibliography

A'Hearn M F, Boehnhardt H, Kidger M and West R M (eds) 1999 *Proc. 1st Int. Conf. on Comet Hale–Bopp, Earth Moon Planets* **77–79**

West R M 1995–2000 Comet Hale–Bopp web pages http://www.eso.org/outreach/info-events/hale-bopp/index.html

Jacques Crovisier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Halley (1P/Halley)

Comet Halley is probably the most famous comet on record. The reason is that it is the only bright comet, easily visible with the unaided eye, that comes back so often, namely with a periodicity that is never very different from 76 years. Since such a periodicity is roughly comparable to the duration of a human life, grandparents talk about it to their grandchildren and oral tradition has established its fame.

The 1682 passage of comet Halley

Until the 1682 passage of comet Halley, none of its previous passes had been identified as coming from the same object. At that time, no comet was known to be periodic, and the shape of the cometary trajectories was unknown. In particular, Johannes Kepler assumed them to be straight lines. However, a couple of years before the passage of comet Halley, Sir Isaac Newton established for the first time that a parabola fitted the observed trajectory of the Great Comet of 1680 much better than a straight line. A parabola was suggested as one of the possible solutions given by his new theory of universal gravitation.

In 1703, Isaac Newton's friend, EDMOND HALLEY, undertook to compute parabolic orbits for those 24 historical comets for which enough accurate data were available. Among the 24 parabolas, three of them were not only of the same size, but they were superimposed upon each other in space. Their dates of passage were separated by 75 and 76 years. He concluded that the three parabolas were actually the extremity of an extremely elongated ellipse. The three anonymous comets were therefore the three passages of the same single comet coming back periodically. Halley, rightly proud of his results, wrote (in Latin):

> *If, in agreement with our predictions, this comet comes back again in 1758, fair posterity will recognize that its return has been first predicted by an Englishman.*

The 1759 return

The comet was found back at Christmas 1758 by a German amateur; it passed at perihelion in March 1759. Since the return predicted by Halley had been verified, the so far anonymous comet was called for the first time Halley's comet or, by modern astronomers, comet Halley. Its 1758 passage played a pivotal role in continental Europe in convincing people that Newton's law of universal gravitation was correct. The reason is that Alexis Clairaut, using the existence of planetary perturbations to correct the comet's orbit for the first time, predicted the 1759 perihelion passage with an accuracy of one month.

The 1835 return

By comet Halley's 1835 return, improvements in optical instruments revealed for the first time the existence of physical phenomena taking place in the comet's head. In particular, Friedrich Bessel's drawings show jets, rays and fans that seem to be ejected sunwards before being repelled away from the Sun. Bessel interpreted his drawings by his now classical 'fountain model', already suggestive of vaporizing gases dragging dust away, sunwards off the nucleus. This looked like a confirmation of Laplace's ideas, written in 1803, about the existence of a solid frozen nucleus.

The 1910 return

Photography and spectroscopy brought a very large amount of data for the first time. In 1931, Bobrovnikoff published a major monograph of the 1910 passage, in which he used 709 photographs and several scores of spectrograms. The radicals and ions that had been identified in the heads and tails of former comets originated from only three elements: carbon, nitrogen and oxygen. Comet Halley's 1910 passage added hydrogen to this list (the Fortrat band of CH near 4300 Å). Bobrovnikoff assumed that the four identified elements (H, C, N, O) came from three 'parent' molecules vaporizing from the nucleus, namely carbon dioxide CO_2, ammonia NH_3 and water H_2O.

Fragmentation and ionization of these three molecules were able to explain all the observed spectra: CN, C_2, C_3, CH, OH, NH, NH_2, and the ions CH^+, OH^+, N_2^+, CO^+ and CO_2^+ (including later identifications). These data are the baseline upon which Fred Whipple would build his icy conglomerate model of the COMETARY NUCLEUS in 1950. The need for more parent molecules only became apparent later, because the absence of molecular collisions in most cometary coma was not observed previously.

The 1986 return: atomic abundances

With the help of space missions, comet Halley's 1986 passage brought a wealth of new data, coordinated by the International Halley Watch and preserved on compact disks in the Comet Halley Archives. The mean ratios in numbers of atoms of the light elements found in the volatile fraction of comet Halley were:

$$\mathrm{H/O} = 1.9 \pm 0.4$$

$$\mathrm{C/O} = 0.20 \pm 0.05$$

$$\mathrm{N/O} = 0.10 \pm 0.05$$

$$\mathrm{S/O} = 0.01 \pm 0.005.$$

The dust-to-gas mass ratio was found to be $M = 0.8 \pm 0.2$ for comet Halley. Its elemental abundances were found to be quasi-solar for the 17 elements measured, except for

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hydrogen, which was depleted by a factor of about 500 in the comet with respect to the Sun. There was also a smaller discrepancy for Si, twice as abundant, and Fe, half as abundant in the comet as in the Sun, so that the ratio Fe/Si is four times smaller in comet Halley than in the Sun.

This anomalous ratio recalls the same type of anomaly in carbonaceous chondrites, where the Fe/Si ratio varies from 25% to 75% solar. COMETS and CHONDRITES are believed to have a similar origin, although a slightly different thermal history (chondrites were heated more than comets). The composition of the terrestrial planets also suggests that a variable iron deficiency exists in the inner solar system with respect to the Sun.

The 1986 return: molecular data

All chemical species identified in previous comets, plus a few more, were identified in comet Halley during its 1986 passage. They are listed in table 1.

The *mean* chemical composition of dust in comet Halley has also been established in percentages by mass. Dust contains 33% organics and 67% inorganics. The organic fraction is half unsaturated hydrocarbons and half more complex molecules also containing oxygen, nitrogen and sulfur. The inorganic fraction is three quarters silicates, the rest being FeS, graphite and sulfur; both fractions also contain bound water, probably hydration water.

The volatile fraction of comet Halley contained (in numbers of atoms) about 78.5% H_2O, 4.5% HCO–OH, 4.2% H_2CO, 3.5% CO_2, 1.0% CO; 2.5% N_2, 1.0% HCN, 0.8% NH_3, 0.8 N_2H_4, 0.4% $C_4H_4N_2$; 1.5% C_2H_2, 0.5% CH_4, 0.2% C_3H_2; 0.1% H_2S, 0.05% CS_2 and 0.05% S_2.

The composition of individual grains of dust from comet Halley has revealed a large dispersion of elemental abundances from grain to grain, although it seems possible to classify them into four distinct families:

Group A (carbon-rich) mean 54% C, 12% O, 10% metals (37 grains).
Group B (oxygen-rich) mean 54% O, 10% C, 15% metals (18 grains).
Group C (Mg–Si rich) mean 81% metals (but 5% Fe), 2% C, 2% O (10 grains).
Group D (iron-rich) mean 33% Fe, but 9% Mg, 5% Si, 8% C, 4% O (11 grains).

These submicroscopic grains seem to represent interstellar dust grains from four different sources, coming initially from different stellar environments, and which have never been thermally processed before sedimenting together to the mid-plane of the protosolar nebula, before accreting into comets.

Halo of vaporizing grains

The existence of an extended halo of vaporizing icy particles has been detected around the nucleus of comet Halley, in particular by its production of CO, probably coming from the photodissociation by the solar light of the molecules of formic acid HCO–OH and formaldehyde H_2CO, which have both been detected in the organic grains dubbed CHON by astronomers. Other icy grain halos had been detected before in other comets.

Isotopic ratios

Before comet Halley's 1986 passage, the only isotopic ratio measured in comets was the $^{12}C/^{13}C$ ratio measured from spectra of the C_2 radical. The spectra were difficult to measure because of the blend of the C_2 band with NH_2. Approximate ratios from 70 to 140 were assumed to be consistent with the terrestrial ratio of 89, rather than with the interstellar ratio of 43±4. However, the analysis of the ^{13}CN spectral lines in comet Halley gave an isotopic ratio of 65±9 for carbon in the CN radical. This isotopic ratio depends on the molecule considered because it comes from different fractionation effects. The interstellar ratio also varies from 40 to 90 in different molecules. The fact that some dust grains in comet Halley had different isotopic ratios for carbon can be interpreted as being due to the different origins of these grains in interstellar space, confirming the interpretation of their large dispersion in elemental abundances mentioned previously.

The D/H isotopic ratio has also been measured in comet Halley; first measurements had large error bars, but they have been considerably improved recently. Deuterium now yields 320 parts per million (ppm) in Halley's water, that is about twice as large as the concentration of deuterium in our oceans (about 156 ppm). This large deuterium enrichment has been confirmed in two recent bright comets (Hyakutake and

Table 1. Chemical species identified in comet Halley.

(a)	By spectroscopy: *Organic molecules*: C, C_2, C_3, CH; CN, HCN, CH_3CN; HCO, H_2CO, CO_2 *Inorganic molecules*: NH, NH_2, NH_3, NH_4; H, O, OH, H_2O; S, S_2 *Ions*: C^+, CH^+, CO^+, CO_2^+; N_2^+; OH^+; H_2S^+; Ca^+
(b)	By mass spectrometer during spacecraft flyby: H_3O^+, $(H_2CO)n^+$, NH_4^+, NH_3^+, CH_4^+, CH_3^+, S_2^+, CS^+, CS_2^+, $C_3H_3^+$ HCO–OH and H_2CO in dust grains

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hale–Bopp). Compared with the deuterium in the atmosphere of Jupiter, which represents a fair estimate of its content in the primeval nebular gas, the deuterium enrichment in the water molecules of comet Halley, as well as that in the two other bright comets, is by a factor of 16, whereas the enrichment in the water of our oceans is by a factor of eight.

Comet Halley's orbit

The present orbit of comet Halley is an extremely elongated orbit of eccentricity $e = 0.967$. With its 76-year period, its perihelion distance of 0.587 astronomical units (AU) sets its aphelion at almost 35 AU. Its inclination to the ecliptic is 17.8º but, because it orbits in a retrograde direction, its inclination is usually said to be 162.2º.

More than 30 of its previous passages have been computed backwards with an outstanding accuracy, which has allowed 30 passages to be retrieved unambiguously in historical documents. Its period has varied irregularly from 79.6 years (from 451 to 530 AD) to 74.4 years (from 1835 to 1910). The changes in its period come from the different positions of the giant planets, the attractions of which change its orbit slightly. In contrast, its orientation in space hardly changes.

The fate of comet Halley

All observed comets are short-lived because they come repeatedly into the inner solar system, where they decay in the solar heat; periodic comets usually disappear after two or three hundred passages. Comet Halley is no exception. Its frozen nucleus vaporizes gases that drag dust away; this produces those transient phenomena for which comets are so well known. Their heads and tails, constantly lost to space, are the best evidence of their decay. Such a decay often ends in a splitting of the nucleus. Eventually, all comets fade into invisibility. Minor fragments may remain forever along their orbits, or they may decay into finer and finer dust, if they do not eventually hit one of the planets.

Since we still see comets billions of years after the formation of the solar system, there must be a permanent source of new comets.

Sources of pristine comets

Indeed, *two* reservoirs of pristine comets have been found, where a large number of comets have remained since the origin of the solar system. One of them is the KUIPER BELT, a ring of millions of comets on quasicircular orbits beyond the orbit of Neptune; the other is the OORT CLOUD, a gigantic sphere centered on the Sun, the diameter of which is one thousand times the size of the planetary system; it contains billions of comets. These two reservoirs keep the comets in their pristine state in the deep cold of space.

The Kuiper Belt is somewhat perturbed by resonances with the orbit of Neptune; this produces the SHORT-PERIOD COMETS (periods between 3 and 30 years) whereas Galactic tides and sometimes the action of a passing star nearby perturb the Oort Cloud and produce the LONG-PERIOD COMETS (periods more than 200 years).

Comet Halley is a special case: with only a few other comets, its period lies in the gap between short- and long-period comets; for this reason, it should probably be called an 'intermediate-period' comet. Consequently, its origin is not completely certain.

Origin of comet Halley

The very fact that comet Halley is on a retrograde orbit makes it more probable that it was captured from the Oort Cloud. The reason is that all Kuiper Belt comets turn in the prograde direction, as do all the planets. A close encounter with a planet, which would have switched its direction drastically, is a highly improbable event. If comet Halley came from the Oort Cloud some millennia ago, then it shared a common origin with all the other comets of this cloud. The Oort Cloud was formed during the accretion of the giant planets. The growth of these very massive objects ejected a large fraction of the icy PLANETESIMALS that formed in their zones, into the Oort cloud. Uranus and Neptune were the most effective; Jupiter and Saturn became too massive and developed a gravitational field that ejected most of the icy bodies too far away into space to be kept by the Sun's gravitation inside the Oort Cloud.

So the ultimate origin of comet Halley lies probably among the icy planetesimals that accreted in the zones of Uranus and Neptune. The numerous isotopic anomalies mentioned before in the dust grains of comet Halley confirm that the interstellar grains that accreted into such an icy planetesimal had never been heated before; if it had been formed in the zone of Jupiter, the larger temperature of the PROTOPLANETARY DISK (near 220 K) would have induced neutral isotopic-exchange reactions with the nebular hydrogen, which would have also diminished the deuterium enrichment of its water, from 16 down to 6. Since this was not the case, this seems to be the best proof of the ultimate origin of comet Halley. Its deuterium enrichment corresponds to unmodified interstellar conditions. In the outer solar system, it was too cold for neutral isotopic-exchange reactions to work.

The nucleus of comet Halley

Although very small, the cometary nucleus is the only permanent feature of a comet. Head and tail are transient phenomena that appear only near the Sun. They are remnants of the decay of the nucleus into dust and gas. It was known by indirect means that cometary nuclei were objects of a few to a few tens of kilometers, but the nucleus of comet Halley is the first nucleus that has been photographed from nearby (figure 1). This was done in 1986 during the flyby of the GIOTTO spacecraft (European Space Agency).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

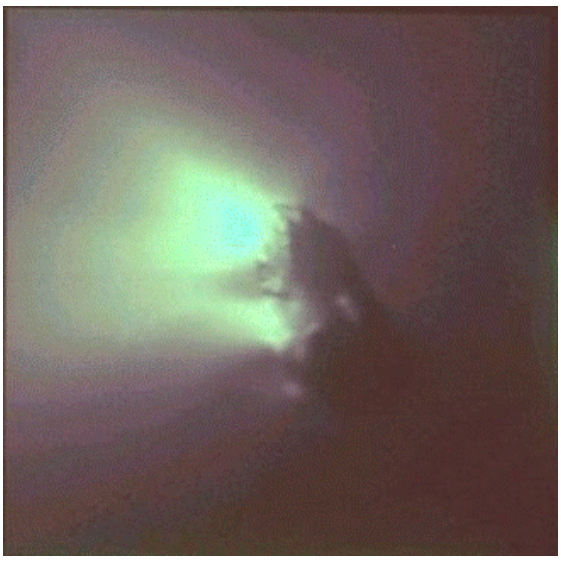


Figure 1. Comet Halley's nucleus, seen from a distance of 4000 kilometers by the Giotto spacecraft during its close passage on 13–14 March 1986. The nucleus is black and irregular; its vents eject gas and dust to the left.

The pictures revealed that it was quite irregular in shape; 90% of its surface area is covered by a black crust that is assumed to be made of silicate grains glued with black organic material looking like soot and clearly containing a large amount of carbon. The loss of water ice and more volatile ices in the crust does not extend very deep; the crust is pierced with many openings that act as erupting craters or fountains from which the inside snows vaporize. What we see now is clearly the result of many passages close to the Sun, where the surface of the nucleus was heated enough to lose its volatiles and weld its grains together. However, the heat conduction of the crust allows the nearby layers of volatiles to vaporize during each passage to perihelion, and their pressure has to become large enough to keep open many 'fountains', which feed the coma and the tails, until the comet has cooled down enough by receding from the solar heat. The nucleus of comet Halley is an elongated body measuring about 8 by 15 km, which seems to have resulted from two roughly spherical icy planetesimals that stuck together.

Comet Halley's 1986 passage has played an important role in clarifying many features and properties of comets, including their origin. It now seems rather well established that cometary nuclei have been accreted from icy interstellar grains that had sedimented first into the mid-plane of the protosolar nebula. The still existing comets are more or less pristine remnants of the stuff which was not used to make the planets.

Bibliography

More information about the remarkable co-ordination between amateur and professional astronomers under the name of the 'International Halley Watch' (IHW), at the occasion of the 1986 passage of Comet Halley can be found in the publication detailed below. This organization was sponsored by the International Astronomical Union and supported by NASA.

Delsemme A H 1991 The 1986 passage of comet Halley *The Comet Halley Archive Summary Volume* ed Z Sekanina (Washington: NASA-JPL) pp 317–30

An excellent book on the results known soon after Comet Halley's 1986 passage can be found in:

Mason J (ed.) 1990 *Comet Halley, Investigations, Results, Interpretations* (New York: Ellis Horwood)

A more recent reference relevant to the space exploration of Comet Halley is:

Eberhart P 1998 Composition of Comets: The *In Situ* View. *Proc. IAU Colloq. 168* ed M A Hearn Astron Soc. Pacific Conference Series (in press)

Armand H Delsemme

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Hyakutake (C/1996 B2)

Comet C/1996 B2 (Hyakutake) was discovered on 30 January 1996 by the Japanese amateur astronomer Yuji Hyakutake. (It should not to be mistaken for another 'comet Hyakutake' discovered by the same person 5 weeks before, C/1995 Y1, which was much less spectacular.) It was soon found out that comet Hyakutake was to make a close approach to Earth on 25 March 1996, at only 0.102 AU (15 300 000 km). It passed perihelion on 1 May 1996 at 0.230 AU from the Sun. This body is a long-period comet (about 9000 yr) with an orbital plane inclined by 125º over the ecliptic; it is presumably coming from the OORT CLOUD.

Close approaches of comets to Earth

Several COMETS are known to have passed peculiarly close to the Earth. Among those with well-known orbits, the closest approach was made by comet D/1770 L1 (Lexell) in 1770: only 0.015 AU (i.e. 2 400 000 km, seven times the Moon-to-Earth distance). In the 20th century, 7P/Pons–Winnecke passed at 0.039 AU in 1927, 73P/Schwassmann–Wachmann 3 at 0.062 AU in 1930 and C/1983 H1 (IRAS–Araki–Alcock) at 0.031 AU and C/1983 J1 (Sugano–Saigusa–Fujikawa) at 0.063 AU in 1983. The close approach of comet Hyakutake at 0.102 AU is thus rare but not exceptional. This comet, however, was a relatively highly productive one, which expelled about 2×10^{29} water molecules s^{-1} (about 5 t s^{-1}) when it was at about 1 AU from the Sun: this is no less than 1/5 of the production of COMET HALLEY at the same distance. This made the comet a very bright object at the end of March 1996, reaching a zero total visual magnitude when at its brightest. It was a conspicuous naked-eye object for several days with a tail extending up to 90º.

The campaign of observation of comet Hyakutake

A campaign of observation, mustering most large astronomical instruments, had to be organized at short notice. It collected many new results on cometary science, just before the equally successful observations of comet C/1995 O1 (Hale–Bopp).

The nucleus, size and rotation

A radar echo could be obtained from the comet using the 70 m Goldstone antenna in California. Since the radar signal is proportional to Δ^{-4} (Δ being the distance of the observer to the target), only comets at a close distance could be observed, and, until now, only a handful of them could be detected by radar techniques: 2P/Encke, 26P/Grigg–Skjellerup, C/1983 H1 (IRAS–Araki–Alcock) and 1P/Halley. From these observations, the COMETARY NUCLEUS of comet Hyakutake was found to be 2–3 km in diameter. Compared with the high level of activity of the comet, this small size indicates that most of the surface is covered with exposed, sublimating ice. Despite the small distance to the observers, the nucleus was too small to be resolved by optical means.

The rotation period of the comet nucleus was determined to be about 6.3 h from near-nucleus imaging and from the comet's light curve.

Several secondary nuclei were observed from images taken just after the close approach to Earth of the comet. Each fragment was surrounded by its own coma. They receded in the anti-solar direction with a velocity of about 12 m s^{-1} and progressively disappeared. Obviously, they were not due to a major break-up of the nucleus, such as those which occurred to comets C/1975 V1 (West) or D/Shoemaker–Levy 9, but rather to fragments of the external crust of the nucleus torn off by cometary activity. It is probable that such a phenomenon is common among active comets but can only be observed at the favor of a close approach.

New molecules

Before the observations of comet Hyakutake, only a handful of parent molecules had been firmly identified in comets: H_2O, CO, CO_2, CH_3OH, H_2CO, HCN, H_2S. Spectroscopic observations, mainly in the radio and infrared domains, allowed the identification of many more species (see also COMETARY ATMOSPHERES.

Ammonia (NH_3), suspected to be the progenitor of the NH and NH_2 radicals, was observed through its centimetric lines with an abundance with respect to water of about 0.7%. This makes this molecule the main repository of cometary nitrogen. Other nitrogen compounds were observed as minor species by their millimetric radio lines: isocyanic acid (HNCO), methyl cyanide (CH_3CN) and hydrogen isocyanide (HNC, an isomer of hydrogen cyanide HCN unstable in the conditions of terrestrial laboratories). A sulfuretted molecule, carbonyl sulfide (OCS), was also detected by radio spectroscopy.

Deuterated water (HDO) was observed by its radio line at 465 GHz. The corresponding [D]/[H] ratio is 3×10^{-4}, a value close to that observed by the mass spectrometers of Giotto during their flyby of comet Halley. This value shows an enrichment of about a factor of 10 over the 'cosmic' abundance of deuterium in the primitive solar nebula ([D]/[H] = 2.5×10^{-5}), where deuterium was mainly contained in the HD molecule. It is not, however, as large as that observed in some interstellar molecules ([D]/[H] = several 10^{-3}). Deuterium enrichment can naturally result from cold-temperature chemical reactions, such as those which occur in the interstellar medium. The enrichment observed in comets suggests that deuterium was not re-equilibrated in cometary water and that cometary material retained, at least partly, the signature of interstellar matter.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The apparition of comet Hyakutake also gave the opportunity to make the first successful interferometric observations of a comet at millimetric wavelengths, soon assessed by further observations of COMET HALE–BOPP.

Infrared observations used a new generation of high-resolution, sensitive spectrometers such as the Cryogenic Echelle Spectrometer (CSHELL) on the NASA Infrared Telescope Facility (IRTF). CO and HCN were observed for the first time in the infrared, and several hydrocarbons were revealed. Methane (CH_4), suspected for a long time to be the parent molecule responsible for the presence of the CH radical, but never yet firmly identified, was observed with an abundance of about 1% relative to water. Ethane (C_2H_6) was also observed with almost the same abundance, as well as acetylene (C_2H_2), which is an unsaturated hydrocarbon (see figure 3 of COMETARY ATMOSPHERES). The presence of both saturated and unsaturated hydrocarbons shows that cometary molecules are not the result of equilibrium chemistry. Carbonyl sulfide (OCS) was also detected in the infrared.

Ultraviolet observations with the Hubble Space Telescope revealed the S_2 molecule, seen for the second time in a comet, the first one being C/1983 H1 (IRAS–Araki–Alcock) at its very close approach to Earth. Obviously, this molecule which has a very short lifetime can only be detected by high spatial resolution observations of close comets. How this molecule could be incorporated in cometary matter, or created by chemical processing of cometary ices or within the inner coma, is not yet understood.

X-ray observations

A totally unexpected result was the detection of x-ray emission from comet Hyakutake. This results from the use of three x-ray satellite observatories, initially dedicated to the observation of high-energy emissions from the Sun or galactic and extragalactic sources: the Röntgen X-ray Satellite (ROSAT) which made the initial detection of the comet, the Rossi X-ray Timing Explorer (XTE) and the Extreme Ultraviolet Explorer satellite (EUVE; see figure 2(a) of X-RAY ASTRONOMY). This discovery was soon confirmed in comet Hale–Bopp, and the emission of several other comets could also be retrieved from the archive data of x-ray satellites. This phenomenon is therefore not rare and could be observed even in some faint comets. It opens quite a new prospect for cometary observations and its interpretation is a new challenge for cometary physics.

High-energy radiation such as x-rays cannot be emitted by a comet on its own, which is an inert, cold body. Rather, this emission comes from the interaction of cometary matter with high-energy solar radiation or solar particles. The exact nature of this interaction has been elucidated recently by a reexamination of the EUVE data on comet Hyakutake and from independent observations of comet C/1999 S4 (LINEAR) by the Chandra X-ray Observatory, which show emission lines of highly excited ions. Charge transfer from heavy ions in the solar wind to neutral atoms in the cometary atmosphere excites cometary atoms to high electronic levels, and their de-excitation results in x-ray emission. This rules out other proposed mechanisms such as scattering of solar x-ray emission by very small (attogram) cometary grains.

Bibliography

West R M 1966–2000 Comet Hyakutake web pages http://www.eso.org/outreach/info-event/hyakutake/

Jacques Crovisier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Ikeya-Seki (C/1965 S1)

Among the small bodies of the solar system, comet Ikeya-Seki—the Great Comet of 1965—is an outstanding object for more than one reason. It belongs to an exceptional class of comets, the Kreutz sungrazing group, an ensemble so called because its members have extremely small closest approach distances, thus brushing the Sun at perihelion (see COMETS: KREUTZ SUNGRAZING). Ikeya-Seki is one of the brightest members of the family, certainly the most brilliant and best observed in our epoch. In close proximity to the Sun, part of the refractory component of the cometary material (the 'dust') is vaporized owing to the very intense heat so that heavy, metallic elements, which are not seen at larger distances, are set free and can be detected. Numerous emissions due to such elements were indeed photographed and identified in detail for the first time in the optical spectra of Comet Ikeya-Seki. Similarly, it was on this comet that the first infrared multiband observations were made and used to derive some further information on the cometary dust. Ikeya-Seki will also be remembered as one of the most spectacular comets of this century, with one near the beginning (the Great Daylight Comet of January 1910) and another close to the end (Hale–Bopp, the Great Comet of 1997); these two latter comets were not Kreutz sungrazers, but they were quite intrinsically bright, especially COMET HALE-BOPP.

Discovery: designations

Comet Ikeya-Seki was discovered on 18 September 1965, independently by two Japanese amateur astronomers, Kaoru Ikeya and Tsutomu Seki. Thus it bears the names of two very strenuous and successful 'comet hunters' who systematically and patiently scrutinize the sky on every clear night in search of new fuzzy moving objects. To the name is added a year designation. Thus, this comet is known in the literature as Ikeya-Seki (1965f) or (1965 VIII). According to a new system adopted recently it is also designated as C/1965 S1 (Ikeya-Seki).

Our famous comet, it should be noted, must not be confused with another Ikeya-Seki discovered about two years later, 1967n=1968 I=C/1967 Y1.

The orbit: a sign of parentage

Only a few days after the discovery, it was realized that the comet would pass very near to the Sun and when, a little later, enough observations had been secured for a reliable trajectory to be determined, the orbital elements left no doubt: in fact, a new member of the sungrazer family was coming in. It was actually the eighth member to be rather well known at that time. All these objects approach the inner solar system from almost exactly the same direction in space and they follow very nearly identical paths, the principal difference being in the period of revolution (500–1100 years, for those that are known), hence in the timing of their appearance.

The idea that some comets form a group or family means that a physical or genetic relationship is thought to exist among these individual objects. This concept as applied to a number of comets with very similar orbits and extraordinarily small perihelion distances dates back to the end of the nineteenth century. KIRKWOOD, and after him Kreutz, who first studied in great detail the orbits of sungrazing comets, contended that these comets were formed by the disintegration of an ancient, giant parent comet. The sungrazer family has been discussed further in the twentieth century, particularly by Marsden and by Öpik, after the passage of Comet Ikeya-Seki. More recently, Bailey and his collaborators have proposed that long-term secular gravitational perturbations in the solar system produce gradual changes in the orbital elements of some comets, leading to sungrazing trajectories; they suggest that such transformations into a sungrazing phase are not infrequent in the dynamical evolution of long-period comets. Let us mention briefly here a special parental link between two members of the family: the orbital elements of comet Ikeya-Seki bear so close a resemblance to those of the Great September Comet of 1882 (1882 II or C/1882 R1), for example their perihelion distances differ by only 0.000 04 AU (1 AU, one astronomical unit, is the mean distance from the Earth to the Sun), less than 1% of the solar radius, that these comets were probably born together when their direct parent split at perihelion (this may be the Great Comet of 1106, as indicated by Marsden). Thus, 1882 II and 1965 VIII are twin sisters—not identical twins, however, since the former was appreciably bigger and brighter than the latter.

The orbit of Ikeya-Seki itself is shown schematically in figure 1. It is an extremely elongated ellipse which takes the comet right through the solar corona at perihelion: it is then at a distance from the Sun's surface of only two-thirds of the solar radius, that is a little less than 500 000 km, whereas at the other end of the orbit (at aphelion) it is about 30 billion kilometers away. A sungrazing comet typically pays us but a lightning visit and provides a striking example of Kepler's second law, according to which the line between the Sun and the orbiting body (the radius vector) sweeps over equal areas in equal intervals of time (see KEPLER'S LAWS). Thus, Ikeya-Seki swung through one-half turn around the Sun in less than four hours (its velocity at perihelion was 480 km s^{-1}), whereas it takes some 900 years for it to make a complete revolution!

From the brightness of the comet at discovery and shortly thereafter, it was foreseen, and widely announced, that Ikeya-Seki would be an impressive sight when close to the Sun. Spectacular it was indeed (figure 2) reaching

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

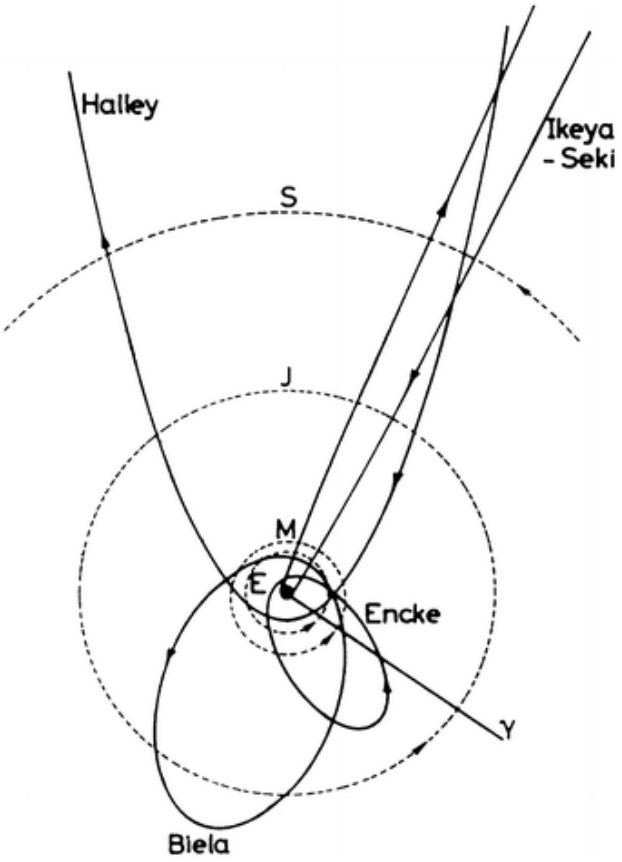


Figure 1. Orbits of Comet Ikeya-Seki (1965 VIII or C/1965 S1) and some other famous comets represented in a simplified manner, in projection onto the plane of the ecliptic (orbital plane of the Earth); also shown for comparison, as dotted lines, are the orbits of a few planets (Earth, Mars, Jupiter and Saturn). The orbit of Ikeya-Seki is a highly eccentric ('plunging') ellipse, close to a parabola. At perihelion the comet truly skims over—'grazes'—the Sun, flying at a distance from the center of the star of only 1.67 times its radius. This exceptionally small perihelion distance, less than one-hundredth of an astronomical unit (AU), could not be shown on the scale of the figure. On the other hand, the aphelion, almost 200 AU away, is out of scale, well outside the planetary system. On this scale again, the orbits of the other members of the sungrazer family to which Comet Ikeya-Seki belongs could not be distinguished from this comet's trajectory. The orbit in space is tilted 38º to the ecliptic; the orbital element 'inclination' taken as 142º indicates that the orbit is retrograde, like that of Comet Halley, which means that the motion is clockwise as seen from the north celestial pole, as opposed to the prograde motion of the planets and of the other two comets. Because the major part of their orbit lies south of the ecliptic, the sungrazing comets are not easy objects for observers in northern latitudes.

an astronomical magnitude of about −10 to −11 at its brightest (i.e. comparable in brightness to the Moon, somewhere between quarter-phase and full moon, or several hundred times brighter than Venus at greatest brillancy) and exhibiting a splendid tail (sometimes 30º

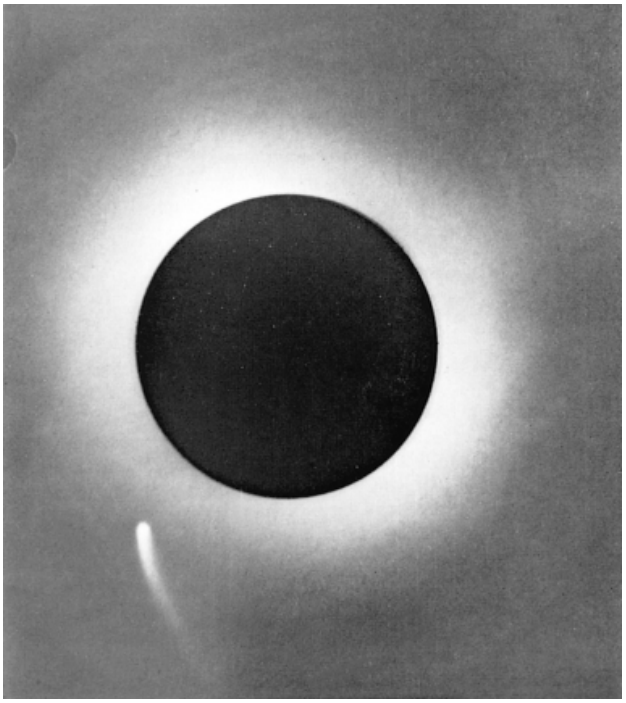


Figure 2. Photograph of Comet Ikeya-Seki obtained two hours before its perihelion passage on 21 October 1965. North is up, east to the left. The Sun is occulted by a black disk surrounded here by scattered solar light (not due to the corona). The swift, accelerated (northeasterly) motion of the comet causes the tail particles to lag far behind the radius vector sun-comet. (Photograph obtained at the Norikura Corona Station of the Tokyo Astronomical Observatory. Reproduced from Hirayama T and Moriyama F 1966 Observations of comet Ikeya-Seki (1965f) *Publ. Astron. Soc. Japan* **17** 433–6.)

long) for a few weeks around perihelion (figure 3). Unfortunately, however, the shape and orientation of the orbit of a sungrazer render its observation rather difficult. So, owing to the small angular distance of Comet Ikeya-Seki to the Sun and its position in the sky, always south of the celestial equator, the magnificent show it gave was reserved essentially for amateur and professional astronomers in privileged sites with very clear skies, at not too large geographical latitudes. It was well observed in Japan, Australia and New Zealand, South Africa, southern Europe, South America, Mexico and in the southwestern part of the United States. Like many other sungrazers, Ikeya-Seki faded away quite rapidly; it was observed for less than four months after perihelion.

Physical characteristics: size, structure

Fragmentation is a common occurrence amongst sungrazers. For example, comet 1882 II was observed to have broken up into more than five pieces near perihelion; similarly, several reports indicated that at least three subnuclei (one much brighter than the others) appeared right after Ikeya-Seki's closest approach to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3. This picture was taken on 29 October 1965, when Comet Ikeya-Seki was 0.45 AU (67 million km) from the Sun and 1.05 AU (157 million km) from the Earth. The tail extends over a little more than 30º, corresponding to a total length of almost 90 million km (true length corrected for projection effect). Notice the presence of structure in the tail. The wavy or helicoidal pattern in the distant part of the tail may be due to the electromagnetic interaction of very small electrically charged cometary grains with the solar wind, as suggested by Horanyi and Mendis. (Photograph by E A Harlan, at the Lick Observatory, Mount Hamilton, California. Reproduced from Heck A 1985 Astronomes amateurs et observations cométaires *Ciel et Terre, Bull. Soc. Roy. Belge d'Astron.* **101** 115–20.)

Sun. Sungrazing objects are formed, we believe, by a multistep process which starts with the breaking apart of the progenitor, many centuries or millennia previously, followed by the disruption of the successive descendants when they return to the Sun's neighborhood. This most probably produces a very large number of small sungrazers, a notion supported by the current discovery of more and more such 'pygmy' sungrazers (in J E Bortle's phrase) with the coronagraphs aboard the Solar and Heliospheric Observatory (SOHO).

The question which comes to mind when a sungrazing comet is seen before perihelion is indeed will it survive? In short, the answer is the bigger fellows will (even if fractured), the pygmies will not. Size and internal strength are the determining factors for avoiding breakup or complete dissipation. As they pass so close to the solar furnace, these comets are subjected to very intense thermal stresses and a more or less important fraction of their mass will be vaporized. Besides, they experience another physical influence from the Sun, a tidal effect. Since the central gravitational attraction is inversely proportional to the square of the distance, the action of the Sun is stronger on the parts of a passing body closest to it than on the most distant parts. This difference may be significant if the solid body is larger than a certain critical limit and is in addition not sufficiently strongly structured, it will then not be able to resist the differential gravitational force and it will split or break up: this is called tidal disruption.

On the basis of dynamical, evolutionary and statistical arguments and using observational data pertaining to the best known sungrazers, Ikeya-Seki in particular, several astronomers have been able to draw inferences regarding the origin and the orbital stability of this class of comets, as well as on some of their physical properties. Consider, for instance, the differential velocities acquired by the fragments of a sungrazer after perihelion passage. If, as Öpik did, we ascribe these differences to differences in the radial deceleration or 'rocket effect' produced on the subnuclei of different sizes as they eject gases in a general sunward direction, we can estimate the dimensions of these various fragments. Applying this method to Ikeya-Seki and using the observed successive separations of the two main companions, one determines that the brighter of these had a diameter between 5 and 10 km, the fainter one being about twice as small. These dimensions are compatible with the lower limit obtained by Sekanina for the diameter of the parent nucleus (3–5 km) on the basis of a different assumption, namely that the relative motions of the subnuclei are attributable to the separation velocity imparted at breakup (this is identified with the equatorial rotational velocity of the parent comet, rotation also being a possible contributing cause to disruption). With an assumed density of 1 g cm^{-3}, the mass of Comet Ikeya-Seki may have been of the order of 10^{14} kg (some five times less than the mass of COMET HALLEY). The mass of the twin sister 1882 II was probably 100 times greater, that of the 'gigantic' ancestor of the sungrazers some two orders of magnitude higher still, i.e. a mass of say 10^{18} kg. This huge number must be put in an astronomical perspective: it represents but a mere one-ten-millionth or so of the Earth's mass!

Since cohesive strength and size are closely related in tidal breakup, knowledge of the latter (at a given distance from the central body) can be used to derive some information regarding the former. Even though the inferred values are approximate, they are rather instructive. By means of a simplified model (non-rotating, homogeneous, spherical nucleus), applied to the data available for Ikeya-Seki and some other members of its family, Öpik showed that the forces of cohesion which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hold these bodies together are much weaker than those of any known relevant materials (like terrestrial rocks or the parents of meteorites), except meteoric 'dust balls' (loosely bound assemblages of dust grains observed as meteors when they enter our atmosphere), which are characterized by crushing strengths of the order of 10^4 dyn cm^{-2}, about four (respectively five) orders of magnitude lower than the tensile (respectively compressive) strengths of rocks. These findings are generally consistent with WHIPPLE's comet model which describes the nucleus as a loose, porous aggregate of icy and mineral matter. Low cohesive strength is also in accord with the fact that fragmentation of comets is often observed to occur even at large distance from a massive body like the Sun or Jupiter, hence in the absence of any tidal action. Finally, it is worth recalling that although rather fragile, cometary nuclei are nevertheless resistant enough to survive a grazing approach to the Sun, the larger ones in any case being observed in the form of several persistent components after their splitting experience.

Clues on the nature and composition of cometary dust

In spite of the great difficulties involved in observing an object located only a few degrees away from the Sun in the sky, considerable efforts were devoted to obtaining data on Comet Ikeya-Seki around perihelion, particularly spectroscopic data. The strong incentive arose from the fact that this sungrazer was offering a very rare—so far unique—opportunity to get direct evidence on the composition of cometary dust by remote observations. Our very limited knowledge of the relative abundances of the heavy elements in comets otherwise comes essentially from the study of meteoroids of cometary origin or from *in situ* measurements (possible only in Halley's comet until now). Thus, quite a number of spectra of Ikeya-Seki were secured on 20 and 21 October 1965, at several observatories (Ratcliffe, South Africa; Sacramento Peak, New Mexico; Lick, California; Kitt Peak, Arizona; Haute-Provence, France).

The evaporation (or rather sublimation) of the non-volatile substances which constitute the cometary dust requires relatively high temperatures, above approximately 1000 K, reached only when a comet is sufficiently near to the Sun. Huebner has studied the sublimation of refractory components by the solar radiation and determined the vaporization rate and lifetime of dust particles as a function of their size, of the latent heat of their constituent materials and of the heliocentric distance (r). It is clear that the metallic elements which are chemically bound in the 'rocky' material, in the form of silicates or oxides can only be liberated when r is less than 0.1–0.2 AU.

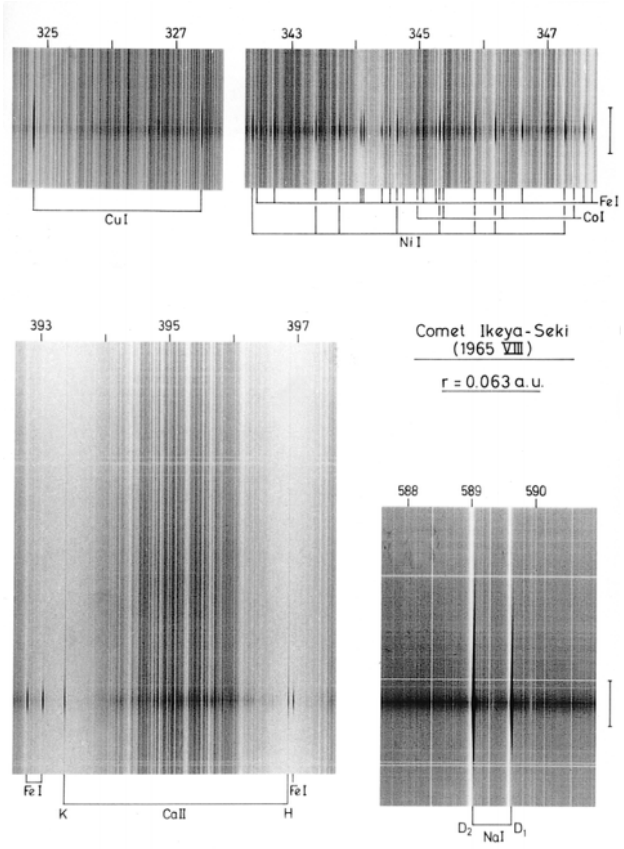


Figure 4. Sample portions of spectra obtained when Comet Ikeya-Seki was about 14 solar radii from the Sun's center after perihelion, in broad daylight. The background is thus the daytime sky spectrum, where numerous Fraunhofer (absorption) lines are seen. On top of this appear (a) a narrow strip of solar radiation scattered by dust particles surrounding the comet's nucleus, the 'continuum', and (b) a number of very short cometary emissions (in black on these negative reproductions) due to neutral copper, iron, cobalt, nickel. The sodium and ionized calcium (H and K) emissions are much longer. The spatial scale is indicated by vertical bars on the right, representing 10 000 km on the comet. The original spectral dispersion was 0.85 mm $Å^{-1}$. A total of approximately 500 emission lines were recorded in the range covered by the series of spectra, 3300–6000 Å; nearly 70% of these emissions belong to Fe I. It is noteworthy that the molecular emissions (not shown here) of CN and, although much weaker, of CH are still present so close to the Sun. (Spectra taken by Livingston and collaborators, Kitt Peak National Observatory, Arizona. Reproduced from Arpigny *et al* 1999 *Atlas of Cometary Spectra* Kluwer, Dordrecht.)

A few sections from these extraordinary spectra of Comet Ikeya-Seki are illustrated in figures 4 and figures 5. Numerous emission lines were observed due mainly to neutral atoms of the iron group, V, Cr, Mn, Fe, Co, Ni and Cu, but also to K, Ca and Ca^+, in addition to Na which first appears near 1 AU from the Sun. This confirmed in particular the presence of a few lines of iron and nickel which had been reported by Copeland and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lohse after their observations of comet 1882 II with a visual spectroscope. This identification had been questioned in the meantime on the basis of an ill-founded criticism ignoring the peculiar physical conditions prevailing in a comet's atmosphere, a low-density medium very far from the thermodynamic equilibrium generally found in laboratory experiments. The observed atomic emissions are produced, like the molecular emissions seen further away from the Sun, by resonance—fluorescence (absorption of solar radiations by the atoms or molecules, followed by the emission of the various possible transitions issued from the upper energy states excited in this way). Analysing a number of spectra of Ikeya-Seki, Arpigny used the measured line intensities to estimate the relative abundances of the various atoms. A notable result was that some elements, especially Al, Si, Ca and Ti, were found to be appreciably underabundant when referred to Fe and compared with the solar system abundances. Mg was a marginal case. Since these elements are the constituent atoms of high-temperature condensates (e.g. corundum, perovskite, melilite, spinal), a possible interpretation of the deficiencies is that the temperature in the coma of Ikeya-Seki was not high enough when the analysed spectra were taken (around 14 solar radii or 0.065 AU from the Sun) for such resisting materials to be decomposed to any significant degree. No actual measurements of the dust temperature, T_d, were performed at this heliocentric distance, nor any closer to perihelion, but indirect evidence suggests that T_d was probably in the range 1000–1500 K near 0.065 AU. It may be noted that Mg-rich olivines and pyroxenes, recently advocated to be important silicatic components of cometary dust grains, also contain some of the underabundant elements. More detailed quantitative values could not be derived because the available data were unfortunately not sufficient, but it seemed nevertheless safe to conclude that the elemental composition of the dust in the comet was similar to that of the Sun and of chondritic meteorites.

On the other hand, the spatial extension of the Na D emission out to large distances from the center of the comet (see figure 5) despite the very short lifetime of sodium against ionization (of the order of 50 s at $r = 0.04$ AU or 8.6 solar radii) indicates, as pointed out by Spinrad and Miner, and by Huebner, that the Na atoms must be embedded in a compound (or compounds) with a high latent heat of vaporization and thus be produced in a distributed source, not directly from the nucleus. From a study of the sodium velocity field it was also inferred that the precursor dust grains must have been fairly large, more than 20 μm in radius. At the same time, it appears that the cometary sodium seen near 0.5–1 AU necessarily originates from particles of a different nature, containing less refractory Na-rich substances.

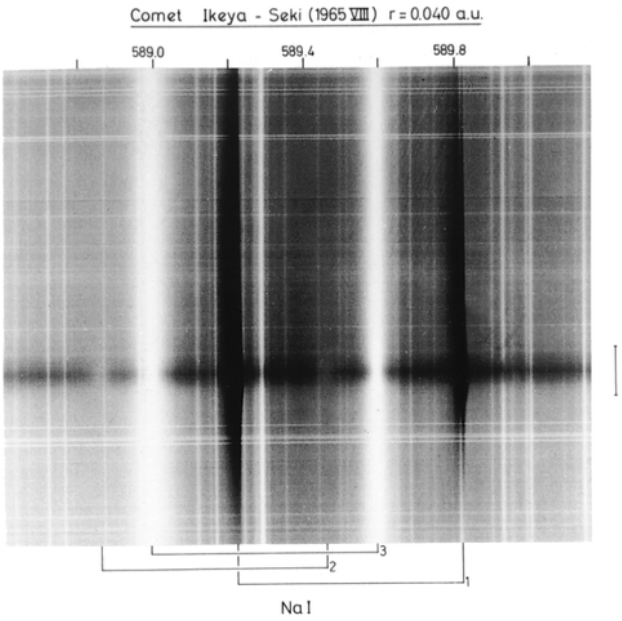


Figure 5. The sodium D lines in comet Ikeya-Seki at $r = 8.5$ solar radii before perihelion (about 2º angular distance), at very high dispersion, 5 mm $Å^{-1}$. This spectrum provides a nice illustration of the Doppler effect. Three doublets can be distinguished: (1) the cometary emissions, here displaced longward by the comet's geocentric velocity ($\dot{\Delta} = +118$ km s^{-1}), (2) the D lines in absorption in the dust scattered solar radiation (the 'continuum'), shifted toward shorter wavelengths according to the sum of the heliocentric and geocentric radial velocities ($\dot{r} + \dot{\Delta} = -188 + 118 = -70$ km s^{-1}), and (3) the same in the background sky spectrum (which gives the reference wavelength scale). (Spectra secured by Spinrad and collaborators, Kitt Peak National Observatory, Arizona. Reproduced from Arpigny *et al* 1999 the *Atlas of Cometary Spectra* Kluwer, Dordrecht.)

Besides these spectroscopic observations, the first infrared observations of a comet were conducted on Ikeya-Seki by Becklin and Westphal who detected the thermal emission by the dust grains and followed its evolution as a function of r, from 0.2 to 0.5 AU both before and after perihelion. From the energy distribution at 1.6, 2.0, 3.4 and 10 μm, these authors found that the temperature of the grains was higher than the temperature that would be attained by a gray or black body in equilibrium with the solar radiation at the heliocentric distance of the comet; they also deduced some information on the wavelength dependence of the emissivity of the grain material, which was found to be much lower in the infrared than in the visible. These findings were confirmed in many comets thereafter. It has now been established that the particles responsible for the thermal emission are very small (submicron- to micron-sized) grains made of silicates mixed with absorbing (probably carbonaceous) materials. These tiny particles absorb efficiently the strong visible solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

radiation, but cannot radiate easily in the infrared; they are therefore overheated. Polarization measurements were made in the tail of the comet, revealing an abrupt variation in the degree of polarization, a reversal from +20% to –40%, which was interpreted as due to a segregation according to grain size along the tail, the grains being presumably composed of slightly absorbing silicates. Attempts were made to observe Comet Ikeya-Seki in the ultraviolet from rockets, as well as in radio wavelengths, but these failed for various reasons, unfortunately.

Conclusion

We started off by stating that Comet Ikeya-Seki was an exceptional object, which is true in some important respects, but at the same time we have reasons to believe that it bears definite similarity to other non-sungrazing comets. Besides the presence of very small, hot dust grains just mentioned and found in many comets, we also note, for instance, that beyond $r = 0.4$ AU, its spectrum showed the usual molecular emission bands due to CN, C_3, CH, C_2 etc. Hence some of its physical and chemical properties may be considered as typical of comets in general. To know these common features in much more detail, especially with regard to the heavy element abundances and the chemical composition of the dust, we await the appearance of another bright member of the Kreutz sungrazing group. We shall then be even better prepared, with present-day highly performing instruments, to study a sungrazer, in particular by multiwavelength, well calibrated measurements, observing the visitor in the ultraviolet, optical, infrared and radio ranges, when close to the Sun and at distance. It will be very important, for the near-perihelion phase, to plan rather tight time sequences and to be ready for the investigation of the spatial distributions of the various emissions: all these observations will have to be carried out in the space of a few hours!

Bibliography

Arpigny C 1979 Relative abundances of the heavy elements in comet Ikeya-Seki (1965 VIII) *Les Éléments et leurs Isotopes dans l'Univers (22nd Liège International Astrophys. Coll.)* (Liège: Institut d'Astrophysique, Université de Liège) pp 191–7

Bailey M E, Chambers J E and Hahn G 1992 Origin of sungrazers: a frequent cometary end-state *Astron. Astrophys.* **257** 315–22

Becklin E E and Westphal J A 1966 Infrared observations of comet 1965f *Astrophys. J* **145** 445–53

Huebner W F 1970 Dust from cometary nuclei *Astron. Astrophys.* **5** 286–97

Sky Telescope 1965, 1966 November, December, January (comprehensive description of the observations of comet Ikeya-Seki 1965 VIII with many illustrations)

Spinrad H and Miner E D 1968 Sodium velocity fields in comet 1965f *Astrophys. J* **153** 355–66

Whipple F L 1985 *The Mystery of Comets* (Washington, DC: Smithsonian Institution Press)

Claude Arpigny

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet IRAS–Araki–Alcock (C/1983 H1)

A long-period comet discovered independently in 1983 by the IRAS infrared satellite, Genichi Araki and George Alcock. At its closest to Earth, on 11 May, it was just 0.03 AU (4.6 million km) away—the closest approach of any comet since Lexell's Comet of 1770. At this time it was a large diffuse object of magnitude 2, moving rapidly across the sky. There was no discernible tail, just a diffuse coma 2° or 3° across, and more extended on the sunward side of the nucleus. Radar observations showed the nucleus to have a diameter of 9.3 km. Perihelion (0.99 AU) was on 21 May. The period is approximately 1000 years; the eccentricity is 0.99, and the inclination 73°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Kohoutek (C/1973 E1)

A long-period comet discovered by Luboš Kohoutek in 1973 March. At that time it was near the orbit of Jupiter, and very bright for a comet at that distance. This led to predictions that it would be exceptionally bright at perihelion (some estimates putting it as high as magnitude −12, rivaling the full Moon), which attracted great media interest. However, the comet failed to brighten significantly as it approached the Sun. Following its perihelion passage (0.14 AU) on 28 December, it appeared in the evening sky in January 1974 at magnitude 4 with a tail 25° long. In this respect Comet Kohoutek failed to live up to its promise. But because of the great expectations it was targeted more than any previous comet had been, and was the subject of an international effort that would be surpassed only by the 1986 International Halley Watch. Comet science was significantly advanced as a result. It is possible that Kohoutek was making its first visit from the Oort Cloud to the inner solar system. This would account for its brightness at the distance of Jupiter, as it would then have been a pristine object outgassing the most volatile of its constituents. The inclination is 14°, and the eccentricity is quoted as 1, meaning that its period and aphelion distance are almost indeterminately large.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet LINEAR (C/1999 S4)

Comet C/1999 S4A (LINEAR) was found by the automatic minor planet search programme LINEAR based in New Mexico. In July 2000, the comet disintegrated completely as it made its closest approach to the Sun, at 71 million miles. At least 16 fragments, resembling 'mini-comets' with tails, were then imaged by the HUBBLE SPACE TELESCOPE.

Results from observations of the debris include evidence that comet impacts may have played a significant role in the formation of LIFE ON EARTH by providing most of the water in the Earth's oceans, as well as organic material. LINEAR is the first comet observed to have a composition that would allow it to carry the same type of water as found in oceans on Earth.

Pam Spence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Swift–Tuttle (109P/Swift–Tuttle)

A short-period comet discovered in 1862 independently by several observers, the first of which were Lewis Swift and Horace Tuttle. At its brightest it reached magnitude 2 and showed a tail over 25° long. Subsequent calculations indicated that the period was around 120 years, but searches in the early 1980s failed to find it. The comet was not recovered until September 1992, as a 'new' comet reported by Tsuruhiko Kiuchi. At this apparition it reached fifth magnitude and developed a 7° tail; perihelion (0.96 AU) was on 12 December. The nucleus showed one particularly active jet, observations of which indicated a rotation period of 2.9 days. This activity is thought to be responsible for non-gravitational forces which cause the period to vary from a value predicted from gravitational perturbations. The comet was subsequently identified with Kegler's Comet of 1737, and with two comets recorded in Chinese annals from 68 BC and AD 188. On its present orbit the comet has an estimated period of 135 years; its eccentricity is 0.96 and its inclination 113°. Debris from the comet forms the Perseid meteor stream.

See also: Perseids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comet Tempel–Tuttle (55P/Tempel–Tuttle)

A short-period comet discovered independently by Ernst Tempel in December 1865 and by Horace Tuttle in January 1866. It was determined to have a period of around 33 years. Studies by Giovanni Schiaparelli showed a very close match between the orbital elements of the comet and those of the Leonid meteor stream, which had produced a spectacular meteor storm in November 1866; this was the first successful demonstration of a link between comets and meteors. The comet was not seen on its next two expected returns in 1899 and 1932 (and there were no Leonid meteor storms in those years). It was recovered in 1965, but only as a distant sixteenth-magnitude object (and another spectacular Leonid storm followed in 1966). At the 1998 apparition it reached fifth magnitude; as on previous occasions, it showed no tail. In 1366 it is calculated to have reached magnitude 3 and to have passed just 0.023 AU (3.4 million km) from the Earth—the third-closest approach of any recorded comet. The perihelion distance is 0.98 AU; the eccentricity is 0.90, and the inclination 163°.

See also: Leonids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmic Rays: Propagation in the Heliosphere

Cosmic rays consist predominantly of atomic nuclei stripped of their electrons, with a small admixture, of order 1%, of electrons and positrons. Their energies range from a few MeV to more than 10^{20} eV, but only cosmic rays with energies below a few 10s of GeV are significantly affected by heliospheric processes. Because all cosmic rays carry an electrical charge, propagation of cosmic rays in the heliosphere is controlled by their interaction with the interplanetary magnetic field. If there were no magnetic field, cosmic rays would pass through the solar system without hindrance.

The processes that affect propagation of cosmic rays through the field are (1) gyration about the lines of force of the magnetic field, (2) scattering from irregularities in the field, (3) drift motions imposed by the large scale gradients and curvature of the field, and (4) adiabatic deceleration, which arises as a result of their coupling through the field to the expansion of the radially outflowing SOLAR WIND. The gyrational motion allows the particles to propagate easily along a field line, which corresponds to the so-called guiding center of the particle's motion, but prevents significant flow perpendicular to the field. The scattering slows the motion along the fields but can also produce transport perpendicular to the field by changing the guiding center of the particle from one field line to an adjacent one. The drifts are determined by the global geometry of the interplanetary field and provide for organized, as opposed to diffusive, cross-field flows. The adiabatic deceleration changes the energy of particles in the solar wind and is the principal source for the flux of particles observed at energies below a few hundred MeV/nucleon in the inner heliosphere. Since particles with these low energies in the interstellar medium are excluded from the inner heliosphere by the solar wind and interplanetary magnetic field, the low-energy cosmic rays observed near Earth entered the heliosphere at higher energies and were decelerated to their observed energies. A cosmic ray may lose several hundred MeV to adiabatic deceleration during propagation from the boundary of the heliosphere in to the orbit of Earth, with a wide distribution of energy losses depending on the actual path taken by the individual particles.

A comprehensive description of the present state of theory and observation of the propagation of cosmic rays in the heliosphere can be found in the book '*Cosmic Rays in the Heliosphere*' (Fisk *et al* 1998). This brief article can only summarize the most important points.

The interplanetary magnetic field, which controls the propagation, has its origin at the Sun and is carried outward by the electrically conducting plasma of the solar wind (see SOLAR WIND: MAGNETIC FIELD). Since the source of the heliospheric magnetic field is the solar field, the overall magnetic structure of the heliosphere is determined by the magnetic structure of the Sun. This structure is not static but undergoes dramatic changes in the course of an 11 yr solar activity cycle. These changes strongly affect propagation of the cosmic rays.

At solar minimum, the overall solar field is predominantly that of a dipole with a small tilt, of order 10° or so, from the rotation axis of the Sun. As discussed in the article on the solar wind magnetic field, field lines from the open polar coronal holes expand near the Sun and are drawn out by the solar wind to fill each hemisphere of the heliosphere. Thus the interplanetary magnetic fields in the north and south hemispheres of the heliosphere have opposite magnetic polarity. The opposing polarities are separated by a thin near-equatorial current sheet. As a result of rotation of the Sun under the radially expanding solar wind, the individual field lines trace out Archimedean spirals in interplanetary space. The solar wind has typical velocities of $\sim$400 km s^{-1} near the equator and 800 km s^{-1} at high latitudes. The rate of solar rotation Ω about 13.8° day^{-1} near the equator, and the solar wind velocity V_{sw} determine the angle ϕ of the spiral to the radial direction through the relation

$$\phi = \tan^{-1}(r \cos\theta \Omega / V_{sw}) \qquad (1)$$

where r is the radius and θ the latitude. Near the ecliptic this angle is about 45° at the orbit of Earth, and it increases to $\sim$80° near the orbit of Jupiter. Since the heliosphere most likely has a radius of order 100 AU, the field is essentially tangential through most of the heliosphere. If cosmic rays were to follow the field their path lengths from the solar wind termination shock at about 100 AU to Earth at 1 AU could be as great as $\sim$5000 AU. In actuality, the cosmic rays experience scattering from field irregularities during their propagation, and the effects of scattering and the continuing outward convection of the field by the solar wind result in a significant reduction in the cosmic ray intensity in the inner heliosphere compared with that in nearby interstellar space. This reduction in intensity is called solar modulation, and it is effective at all times, even at solar minimum.

In the course of a SOLAR CYCLE, the effective tilt of the solar dipole field increases as solar activity increases, and large local fields associated with active regions begin to make significant contributions to the overall field, in effect increasing the importance of higher-order multipoles in the field. At solar maximum, the dipole component is no longer the dominant component, and the field at the Sun and, by extension, in interplanetary space, becomes very complicated. In addition, the increased solar activity results in increased generation of transient disturbances, such as SOLAR CORONAL MASS EJECTIONS, in the solar wind, which further distort the magnetic field. In response to this increase in field complexity, the cosmic ray intensity throughout the heliosphere is reduced markedly from its solar minimum levels.

As activity diminishes following solar maximum, the field begins to simplify and, as solar minimum approaches a new dipole component, with a sign opposite that of the previous solar minimum, becomes dominant. Thus the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

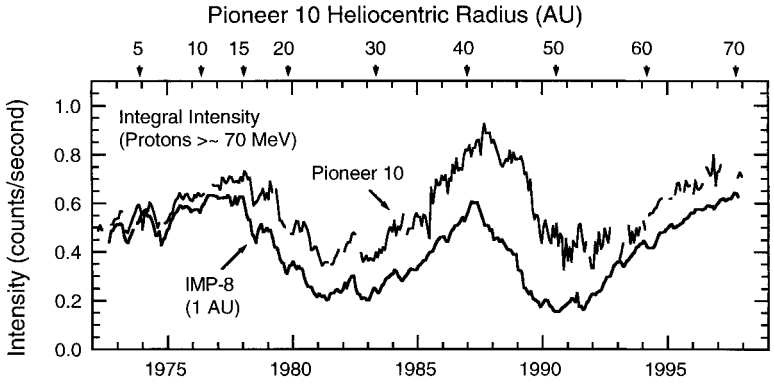


Figure 1. 27 day average intensities of cosmic rays, primarily protons with energy >70 MeV, as measured by University of Chicago instruments on IMP-7 and 8 in Earth orbit and on Pioneer 10 to radii beyond 60 AU from the Sun. IMP intensities have been normalized to match the Pioneer intensities observed at launch in 1972 to correct for the different acceptance factors of the instruments. Effects of the 11 yr solar activity cycle on the cosmic ray intensity are evident, as is the outward propagation of changes in modulation.

solar field, and the heliospheric field with it, reverse sign every 11 yr. For cosmic rays the result of this evolution is the observed 11 yr variation in the cosmic ray intensity, known as the solar cycle modulation (see COSMIC RAYS). As an example, figure 1 shows the effects of this modulation as measured at Earth and at PIONEER 10 to distances beyond 60 AU from the Sun during two cycles of solar activity. The outward propagation of the changes in modulation, representing convection of the changed magnetic structure by the solar wind, is clearly visible, especially for the brief solar minimum of 1987 when the modulation at Earth began to increase even as it was still decreasing at Pioneer 10. Since the average solar wind velocity near the ecliptic is about 400 km s^{-1}, or 0.23 AU day^{-1}, changes observed at 1 AU may not be observed until many months later in the outer heliosphere.

Also apparent in figure 1 is the gradually increasing flux measured at Pioneer 10 compared with that at IMP-8. This increase results from the radial gradient of the cosmic ray intensity, reflecting the fact that the cosmic rays in the outer heliosphere undergo less modulation than those in the inner heliosphere. The magnitude of the gradient is a few per cent per AU, depending on particle energy and species, and is largest in the inner heliosphere. The largest gradients observed, as large as 15–20% AU^{-1} in the inner heliosphere, are for the low-energy anomalous components (see ANOMALOUS COSMIC RAYS) which are accelerated at the solar wind termination shock. Study of the solar cycle modulation provides some of our best information concerning the propagation of cosmic rays in the heliosphere.

Other information that is of critical importance for determining the nature of transport of charged particles in the heliosphere comes from the study of solar energetic particle events, the so-called solar cosmic rays. This aspect has been thoroughly reviewed by Kunow *et al* (1990). During the onset of a solar particle event, the first-arriving particles typically arrive in a very anisotropic flow along the magnetic field direction. The time profile of the arriving particles provides information concerning the scattering properties of the interplanetary magnetic field. Figure 2 shows a time–intensity profile for one of the largest solar energetic particle events ever recorded, which was observed by ground-based detectors of atmospheric radiation induced by cosmic rays in February 1956. This event has features typical of many solar particle events. The rapid onset implies that the particles injected in a burst at the Sun arrive with little scattering between the Sun and the Earth, suggesting a mean free path between scatterings of the order of 1 AU. The slower exponential decay implies the existence of significant scattering beyond the orbit of Earth, which acts to confine the particles in the heliosphere and delay their escape into the interstellar medium. The observations reproduced here in fact provided the first evidence for the existence of the heliosphere. At high energies, such as shown in figure 2, the time scale of a typical event is hours. At lower energies of a few MeV, which can be observed only from space, the time scale for an event from onset until return to normal quiet time intensities may be many days, even weeks for large events.

Scattering of cosmic rays by the interplanetary magnetic field is a result of the interaction between the charged particles spiraling along the field lines and irregularities in the field. If the irregularities are represented by a spectrum of waves through a Fourier transform, the strongest scattering is produced by waves whose wavelength is resonant with the spiraling motion of the particles as they move along the field. Thus the scattering effect depends on the energy and the pitch angle (angle of the velocity vector with respect to the magnetic field) of the particles and on the mean field strength. The effect of the scattering is to modify the pitch angle and/or to displace the guiding center of the particle's gyromotion normal to the field, by, on average, about one gyroradius.

An extensive literature exists concerning attempts to derive the particle scattering properties of the field from observations of the spectrum of irregularities in the field. (For entry points, see Giacalone 1998, Burger and Hattingh

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

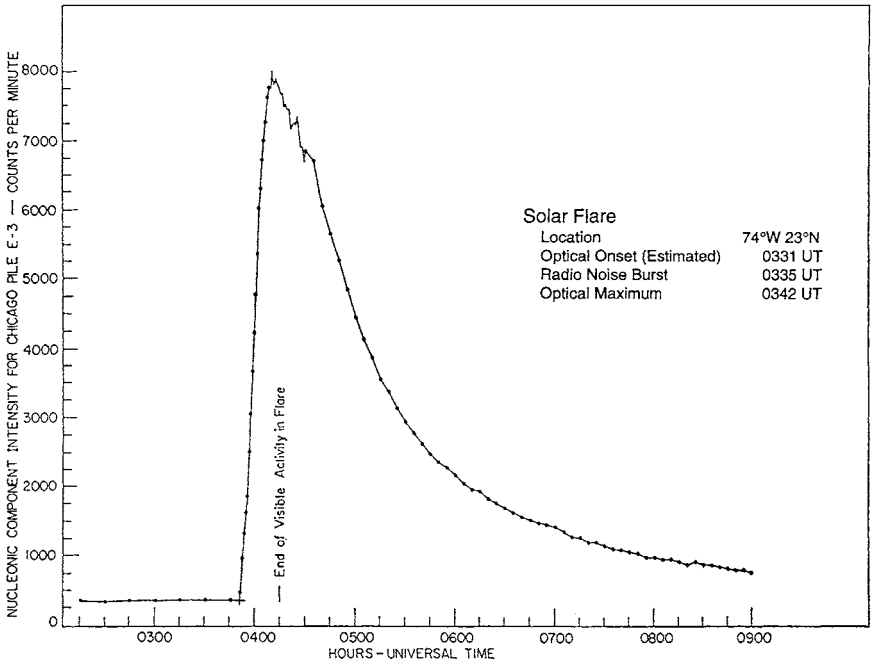


Figure 2. Intensity of ground-level neutrons produced by protons with energies above about 1 GeV interacting in the atmosphere following the 23 February 1956 solar flare, one of the largest ever recorded. The prompt rise implies a long scattering mean free path between the Sun and the Earth, while the extended decay provided the first evidence for an extended scattering region beyond the Earth that we now know as the heliosphere. (Adapted from Meyer *et al* 1956.)

1998). Unfortunately, the effort has still not been fully successful.

In modeling the effects of scattering on modulation of cosmic rays, and of the propagation of solar energetic particles, it is usual to describe the scattering by introducing a diffusion coefficient, κ which may be broken into components $\kappa_\parallel$ parallel to the field and $\kappa_\perp$ perpendicular to the field. In particular, computational models for the modulation of cosmic rays typically use the radial diffusion coefficient

$$\kappa_{rr} = \kappa_\parallel \cos^2\phi + \kappa_\perp \sin^2\phi \qquad (2)$$

as a parameter, where ϕ is the spiral angle of the interplanetary field with respect to the radius vector from the Sun. Throughout most of the heliosphere, $\cos^2\phi$ is approximately zero so that the values of the diffusion coefficient reflect primarily diffusion across the mean field. Typical forms for the diffusion coefficient are $\kappa = A\beta P^\alpha (B_0/B)$ where β is the particle velocity as a fraction of the speed of light, P is the particle magnetic rigidity, or momentum per unit charge, A is a constant and B is the magnetic field strength referred to a reference field B_0. The value of the exponent α is in the range 0–1, and the values of κ_r found to fit observations are typically a few times 10^{22} cm^2 s^{-1} at the orbit of Earth for protons with energies of about 1 GeV. The values increase with radius as a result of the decline in B, which, when the field becomes nearly tangential, decreases as $1/r$.

The deduced values of κ_{rr} pose a particular challenge for theories of the propagation since it has been difficult to achieve rapid enough transport across the mean field direction to match observations. For cross-field propagation due only to scattering, a rough upper limit on the perpendicular diffusion coefficient comes from assuming translation by one gyroradius in the period of one gyration about the field. For ~1 GeV particles near the orbit of Earth, this limit is about 10^{20} cm^2 s^{-1}, or less than 1% of the size of the diffusion coefficient deduced from observations.

Direct observations also clearly require much more efficient transport. As an example of observations which require rapid latitudinal transport, figure 3 shows the persistence to heliospheric latitudes of ~70°–80° of variations in the cosmic ray intensity with the 26 day period of the solar rotation that were induced by solar wind stream interactions near the equator. As shown in the figure, fast solar wind overtaking slow wind produces a region of compressed field, which yields a reduced diffusion coefficient that locally increases the strength of the modulation of cosmic rays. In addition, shocks which accelerate particles to energies of a few MeV may form on the leading and trailing edges of the interaction region. Since the streams are rooted at the Sun, the whole structure rotates with the Sun, giving rise to the name corotating interaction regions (CIRs) (see SOLAR WIND: COROTATING INTERACTION REGIONS). Near solar minimum, the CIRs are confined primarily to the heliospheric equatorial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

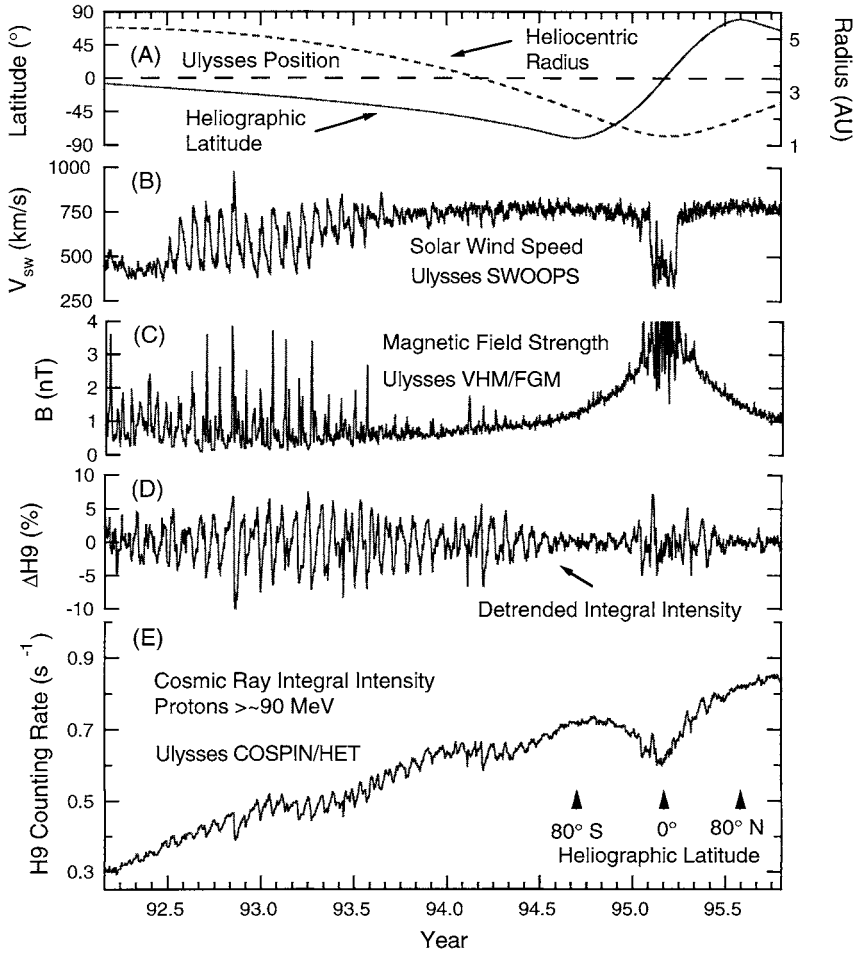


Figure 3. (A) Ulysses position in heliocentric radius and heliographic latitude (B) 6 h average solar wind speeds and (C) daily average magnetic field strengths measured by the Ulysses SWOOPS (Bame *et al* 1992) and VHM/FGM (Balogh *et al* 1992) experiments, respectively; (D) daily average detrended integral intensity of cosmic rays derived by subtracting 27 day running average intensities from (E) the daily intensities measured by the COSPIN/HET experiment (Simpson *et al* 1992). The modest increase of the cosmic ray intensity towards high latitudes is clearly shown by the dip near the equator as Ulysses traveled rapidly from the south to the north polar regions in 1994–1995. The overall increase in intensity from 1992 through to 1995 is the result of decreasing modulation in the approach to solar minimum. (Adapted from Zhang 1997.)

zone where the high-latitude fast wind and the equatorial slow wind can interact. Observation of the cosmic ray intensity variations caused by CIRs at high latitude requires that the cross-field transport take place in a time short compared with the 26 day rotation period of the Sun. Otherwise, the variations would be wiped out. This requires cross-field diffusion coefficients of order 10–20% of the size of diffusion coefficients parallel to the field, or more than a factor of 10 greater than can be explained by scattering theory. At lower energies of a few MeV, recurrent increases attributed to particles accelerated at the CIR shocks have also been seen at high latitudes, raising even greater difficulties for transport theory. For a full discussion of the observations see Simnett *et al* (1998).

Transport across the mean field can be greatly enhanced by systematic motions of the field lines with respect to the mean field direction. Two mechanisms have been identified that may contribute significantly to propagation of energetic particles across the field. The first, first recognized by Jokipii and Parker (1969), is that the solar magnetic field is dragged around on the surface of the Sun by the large-scale convective motions that form the supergranulation network. Since the interplanetary field is rooted in the solar field, the transverse motions imposed by this convective turbulence are carried out into the interplanetary medium. As a result, the interplanetary field lines themselves undergo a random walk in both longitude and latitude. This braiding of the interplanetary field lines then provides an efficient means of enhancing the effect of scattering in transporting particles across the mean field, since a single scattering may transfer a particle's guiding center to a field line that diverges rapidly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

from the original field line.

A more ordered motion of the field lines has been suggested by Fisk (1996), who considered the implications for the field lines of the observations that (a) the magnetic axis of the Sun is inclined to the rotation axis and, in the corona, seems to rotate rigidly with the equatorial rotation rate of the Sun and (b) the photosphere, in which individual field lines are rooted, exhibits differential rotation, rotating significantly more slowly at polar latitudes than at the equator. The result is significant latitudinal mixing of field lines, providing direct magnetic connections between equatorial and polar regions of the heliosphere over a radial range of order 15 AU. Longitudinal mixing would also be expected, but the model was proposed, and has been most closely examined, as a means of explaining the rapid latitudinal transport of particles that seems to be required by Ulysses observations (see SOLAR WIND: ULYSSES).

Despite many years of work, however, a convincing understanding of cross-field diffusive transport remains to be achieved. In particular, it is very difficult to derive experimental tests to isolate the effects of the field line random walk or the mixing driven by the differential rotation. All that is certain is that particles, while guided by the field lines, show surprising ability to diffuse in the direction perpendicular to the mean field as well.

A more systematic motion across the field is produced by the gradient and curvature drifts in the heliospheric magnetic field. The spiral nature of the average interplanetary field provides curvature on a global scale, and the progressive weakening of the field as the solar wind carries it outwards through the heliosphere provides a global gradient in the field. By the basic physics of charged particle motion in a non-uniform field, therefore, a globally organized drift of cosmic rays through the heliosphere is generated. The sense of the drift is such that when the north pole of the Sun is a positive magnetic pole, positively charged cosmic rays (nucleons) drift from the polar regions of the heliosphere toward the equator and then rapidly outward along the equatorial current sheet. Electrons, on the other hand, drift inwards along the current sheet and towards the poles above and below the current sheet. 11 yr later, when the magnetic poles of the Sun have the reverse polarity, the drift patterns are also reversed. The drift velocities can be substantial. For particles with magnetic rigidity P, charge q and velocity β, the drift velocity is given by

$$\boldsymbol{V}_{\mathrm{d}} = (\beta P c/3q)\boldsymbol{\nabla} \times (\boldsymbol{B}/B^2) \tag{3}$$

where $\boldsymbol{B}$ is the vector interplanetary field. For 0.1 GeV protons near the orbit of Earth in an ideal Parker spiral field, this velocity is approximately 300 km s^{-1}, comparable with the solar wind velocity. At higher energies, the ideal drift velocity can substantially exceed the solar wind velocity through much of the heliosphere. However, these velocities are directed and not diffusive. Thus, since for the period shown in figure 3 the drift velocities of protons were directed from the poles towards the equator, it would be difficult to invoke drifts to explain the observation of intensity variations at high latitudes impressed by CIRs at low latitudes.

The effect of drifts is substantially diminished in the real field of the heliosphere by the irregularities in the field imposed both at the Sun and by the evolution of structures in the solar wind. CIRs, coronal mass ejections, shear flows in the solar wind, waves and turbulence in the solar wind flow, and many other non-uniformities distort the fields and interrupt the drift flow. At low latitudes, the variable tilt of the heliomagnetic equator translates into waviness of the heliospheric equatorial current sheet as the Sun rotates and the field is carried outwards. This waviness, which varies during the solar cycle, increases the length of the drift path along the sheet and also the probability of loss of particles from the current sheet region. At high latitudes, where from equation (1) one might expect nearly radial fields as $\cos\theta$ approaches 0, the effect of small transverse fluctuations imposed by convective motions at the Sun destroys the field coherence necessary for the full effect of drifts. As first noted by Jokipii and Kota (1989), the fluctuations produce transverse components in the field that vary with radius as $1/r$, whereas the radial components vary as $1/r^2$. Thus beyond a few AU from the Sun transverse components dominate, and the field is essentially tangential. As a result, it is as difficult for cosmic rays to penetrate the polar regions of the heliosphere as it is in the equatorial regions. This has been confirmed by observations from the Ulysses mission, which show both the transverse fluctuations in the field and a change of less than a factor of 2 in cosmic ray intensity from the equator to the pole near solar minimum.

Nevertheless, the effect of drifts, although reduced, has been demonstrated by the observation of variations in the modulation dependent on the sign of the particle charge, to which all major modulation processes except drifts are insensitive, and by confirmation of many predictions of drift models that are difficult to understand in the absence of drifts, such as the alternating broad and narrow shapes of the cosmic ray intensity profiles at successive solar minima (cf Figure 1), reflecting the alteration of the drift paths through the heliosphere imposed by the reversal of the solar dipole field.

The present state of the understanding of the propagation of charged particles in the heliosphere is mixed. The important processes have been characterized in gross terms and models have been developed that, using a number of adjustable or observationally defined parameters, produce reasonable and consistent descriptions of the observed phenomena. However, detailed physical understanding of the characteristics of the propagation remains as work in progress.

Bibliography

Balogh A *et al* 1992 *Astron. Astrophys.* **92** 221–36
Bame S J *et al* 1992 *Astron. Astrophys.* **92** 237–66
Burger R A and Hattingh M 1998 *Astophys. J.* **505** 244–51

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fisk L F 1996 *J. Geophys. Res.* **101** 15 547–54
Fisk L A, Jokipii J R, Simnett G M, von Steiger R and Wenzel K-P (ed) 1998 *Cosmic Rays in the Heliosphere* (Dordrecht: Kluwer) (Fisk L A, Jokipii J R, Simnett G M, von Steiger R and Wenzel K-P (ed) 1998 *Space Sci. Rev.* **83**)
Giacalone J 1998 *Space Sci. Rev.* **83** 351–63
Jokipii J R and Kota J 1989 *Geophys. Res. Lett.* **16** 1–4
Jokipii J R and Parker E N 1969 *Astrophys. J.* **155** 777–98
Kunow H *et al* 1990 *Physics of the Inner Heliosphere, 2: Particles, Waves and Turbulence* ed R Schemm and E Marsch (Berlin: Springer) pp 243–342
Meyer P, Parker E N and Simpson J A 1956 *Phys. Rev.* **104** 768–83
Simnett G M *et al* 1998 *Space Sci. Rev.* **83** 215–58
Simpson J A *et al* 1992 *Astron. Astrophys.* **92** 365–400
Zhang M 1997 *Astrophys. J.* **488** 841–53

R Bruce McKibben

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmological Model

An idealized picture of the overall structure of the universe and its evolution that may be tested by comparing its predictions with observational data. For simplicity, most cosmological models ignore individual galaxies and clusters, and treat the matter and radiation content of the universe as if it were smeared out into an idealized smooth distribution that is sometimes called 'the substratum'. Practically all cosmological models incorporate the cosmological principle (that the universe is homogeneous and isotropic) and assume the universality of physical laws (that the laws of nature are everywhere the same). If the universe is homogeneous (the same everywhere), then all fundamental observers (observers who are at rest relative to the substratum in their vicinity) see the same sequence of events in the history of the universe, this sequence of events defining a universal cosmic time.

Most cosmological models are based on the general theory of relativity, in which gravititation is treated as a phenomenon arising from the curvature of space (or, strictly, four-dimensional spacetime) that is induced by the presence of massive bodies. Flat space has zero curvature (in flat space, the shortest distance between two points is a straight line and the sum of the angles in a triangle is 180°). Positive curvature corresponds to a universe of finite volume that is curved in an analogous way to the surface of a sphere (in such a space, 'parallel' lines eventually meet, and the angles in a triangle add up to more than 180°). Negative curvature is analogous to the shape of a saddle (in negatively curved space 'parallel' lines eventually diverge, and the sum of the angles in a triangle is less than 180°).

Different model universes may be generated by selecting different values for fundamental quantities such as mean density and curvature, and by choosing zero or non-zero values for an additional term (the cosmological constant) that appears in Einstein's equations. The standard cosmological models are developed from equations derived by Aleksandr Friedmann (1888–1925) in 1922 and are known, therefore, as Friedmann models. If the cosmological term is assumed to be zero, there are three Friedmann models that equate to the Big Bang universe: the open model (in which space is negatively curved and infinite), the closed model (in which space is positively curved and finite) and the 'flat' model (zero curvature, infinite space). The open universe expands forever, the closed universe expands to a finite size then collapses, and the flat universe is just, but only just, capable of expanding forever. The flat model is also known as the Einstein–de Sitter universe. If a positive value for the cosmological constant is included, models can be constructed in which the expansion rate of the universe accelerates.

See also: Big Bang theory, cosmological principle, critical density, general theory of relativity, inflationary universe, oscillating universe, steady-state theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmological Principle

A fundamental postulate of cosmology that states that, on the large scale, the universe is homogeneous (one large region of space is the same as any other large region of space) and isotropic (looks the same in every direction). This principle implies that our Galaxy is not located at any special or privileged position in the universe and that any observer on any galaxy will see the same general features of the universe as we do. The cosmological principle is sometimes called the Copernican principle after the Polish cleric who, in 1543, dethroned the Earth from its previously assumed central position in the cosmos.

An extension of the cosmological principle, known as the perfect cosmological principle, postulates that, on the large scale, the universe looks the same everywhere, in all directions, *at all times*. If valid, the perfect cosmological principle would require that the large-scale appearance of the universe does not change with time and would preclude the Big Bang model of the universe. It was proposed in 1948 by Hermann Bondi and Thomas Gold as the basis of their formulation of the steady-state theory. Although the cosmological principle is widely assumed to be valid (although the degree to which it is true depends on the scale over which homogeneity and isotropy is averaged), the perfect cosmological principle is untenable in the light of modern observational evidence.

See also: cosmological model, cosmology, steady-state theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmology

The study of the structure, origin and evolution of the universe as a whole.

Observationally, luminous matter in the universe is distributed in a clumpy fashion, being aggregated together into individual galaxies, groups of galaxies, clusters and superclusters. On the very large scale, however, the universe appears to be homogeneous (i.e. is the same everywhere) and isotropic (i.e. looks the same in every direction). This implies that all observers will see the same large-scale view of the universe regardless of where they are located. No individual galaxy or observer occupies a privileged position, and the universe has no definable center or edge.

During the 1920s, Edwin P Hubble (1889–1953) showed that all galaxies beyond our immediate local group have redshifts in their spectra and that the magnitude of the redshift is proportional to the distance of the galaxy. If the redshift is interpreted as a Doppler effect, Hubble's observations implied that all galaxies are receding with speeds proportional to their distances, a relationship which is now known as the Hubble law. Subsequent observations have confirmed that the Hubble law holds good to very large distances. These observations imply that each galaxy (or, strictly, each cluster of galaxies), is receding from every other one and that the whole universe is expanding.

The observed recession of the galaxies is consistent with the Big Bang theory, which implies that the universe originated a finite time ago (probably between 10 and 15 billion years ago) by expanding from a hot dense initial state. The hot Big Bang theory provides a good explanation of many facets of the observed universe, including the recession of the galaxies, the relative abundance of the lightest chemical elements, and the existence and properties of the cosmic microwave background radiation (a dilute background of microwave radiation that permeates the universe). A more recent adjunct to the theory, the inflationary hypothesis, contends that at a very early stage in its history, the universe experienced a brief, but very dramatic, period of accelerating expansion. This hypothesis, if valid, provides an explanation for several features of the universe, including its homogeneity and isotropy, which cannot directly be explained by the standard Big Bang model. However, how, when, and in what order galaxies, clusters and superclusters were formed remains an unresolved question.

Another major issue in modern cosmology is the question of whether or not the universe will continue to expand forever. In the absence of any other force, gravity is expected to be slowing the rate of expansion. If the mean density of the universe exceeds a particular value, called the critical density, the expansion will cease at some time in the future and the universe will then begin to contract, slowly at first and then ever more rapidly until all of its matter and radiation collides in a 'big crunch'. A universe of this kind is called 'closed'. If the mean density is less than critical, the universe will expand forever, and is called 'open'. The dividing line between these two alternatives is a universe where the mean density is exactly equal to the critical density and in which the speeds of recession of the galaxies slow down ever closer to zero but do not become zero until the infinite future. This case is called the 'flat' universe (or the Einstein–de Sitter universe) because, in such a universe, space, on the large scale, has zero net curvature.

Whether the universe is open, flat or closed may, in principle, be determined by measuring the rate at which the expansion is decelerating or by determining the mean density of matter and radiation in the universe. Although many theoretical cosmologists favor 'flat' or 'closed' models, and the observational data are far from conclusive, the balance of recent evidence appears to favor the open model.

See also: Big Bang theory, cosmological model, cosmological principle, deceleration parameter, expanding universe, Hubble constant, inflationary universe, redshift, steady-state theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmology Standard Model

COSMOLOGY in the modern sense of quantitative study of the large-scale properties of the universe is a surprisingly recent phenomenon. The first galaxy RADIAL VELOCITY (a blueshift, as it turned out) was only measured in 1912, by Slipher. It was not until 1924 that Hubble was able to prove that the 'nebulae' were indeed large systems of stars at vast distances, by which time it was clear that almost all galaxies had spectral lines displaced to longer wavelengths. Subsequent observations increasingly verified Hubble's (1929) linear relation between distance d and the recessional velocity inferred if redshift was interpreted as a Doppler shift:

$$v = Hd.$$

The theoretical groundwork for describing the universe via GENERAL RELATIVITY was already in place by the mid-1920s, so that it was not long before the basic observational fact of an expanding universe could be given a relatively standard interpretation. The main observational and theoretical uncertainties in this interpretation concern the matter and energy content of the universe. Different possibilities for this content generate very different COSMOLOGICAL MODELS. The purpose of this article is to outline the key concepts and practical formulae of importance in understanding these models, and to show how to apply them to astronomical observations.

Isotropic spacetime

Modern observational cosmology has demonstrated that the real universe is highly symmetric in its large-scale properties, but it would in any case make sense to start by considering the simplest possible mass distribution: one whose properties are homogeneous (uniform density) and isotropic (the same in all directions). The next step is to solve the gravitational field equations to find the corresponding metric. Many of the features of the metric can be deduced from symmetry alone—and indeed will apply even if Einstein's equations are replaced by something more complicated. These general arguments were put forward independently by H P Robertson and A G Walker in 1936.

Consider a set of 'fundamental observers', in different locations, all of whom are at rest with respect to the matter in their vicinity. We can envisage them as each sitting on a different galaxy, and so receding from each other with the general expansion (although real galaxies have in addition random velocities of order 100 km s^{-1} and so are not strictly fundamental observers). A global time coordinate t is supplied by the time measured with the clocks of these observers—i.e. t is the proper time measured by an observer at rest with respect to the local matter distribution. The coordinate is useful globally rather than locally because the clocks can be synchronized by the exchange of light signals between observers, who agree to set their clocks to a standard time when for example the universal homogeneous density reaches some given value. Using this time coordinate plus isotropy, we already have enough information to conclude that the metric must take the following form:

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)[f^2(r)\,\mathrm{d}r^2 + g^2(r)\,\mathrm{d}\psi^2].$$

Here, we have used the equivalence principle to say that the proper time interval $\mathrm{d}\tau$ between two distant events would look locally like special relativity to a fundamental observer on the spot: for them, $c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$. Since we use the same time coordinate as they do, our only difficulty is in the spatial part of the metric: relating their $\mathrm{d}x$ etc to spatial coordinates centered on us.

Distances have been decomposed into the product of a time-dependent scale factor $R(t)$ and a time-independent comoving coordinate r. It is clear that this metric nicely incorporates the idea of a uniformly expanding model with no center. For small separations, where space is Euclidean, we have a simple scaling of vector separations: $\boldsymbol{x}(t) : R(t)\,\boldsymbol{x}(t_0)$. The same law applies irrespective of the origin we choose: $\boldsymbol{x}_1(t) - \boldsymbol{x}_2(t) : R(t)\,[\boldsymbol{x}_1(t_0) - \boldsymbol{x}_2(t_0)]$, and so every observer deduces $\boldsymbol{v} = H\,\boldsymbol{r}$

Because of spherical symmetry, the spatial part of the metric can be decomposed into a radial and a transverse part (in spherical polars, the angle on the sky between two events is $\mathrm{d}\psi^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$). The functions f and g are arbitrary; however, we can choose our radial coordinate such that either $f = 1$ or $g = r^2$, to make things look as much like Euclidean space as possible. Furthermore, the remaining function is determined by symmetry arguments.

Consider first the simple case of the metric on the surface of a sphere. A balloon being inflated is a common popular analogy for the expanding universe, and it will serve as a two-dimensional example of a space of constant curvature. If we call the polar angle in spherical polars r instead of the more usual θ, then the element of length, $\mathrm{d}\sigma$, on the surface of a sphere of radius R is

$$\mathrm{d}\sigma^2 = R^2\left(\mathrm{d}r^2 + \sin^2 r\,\mathrm{d}\phi^2\right).$$

It is possible to convert this to the metric for a 2-space of constant negative curvature by the device of considering an imaginary radius of curvature, $R \to {}_{\mathrm{i}}R$. If we simultaneously let $r \to \mathrm{i}r$, we obtain

$$\mathrm{d}\sigma^2 = R^2\left(\mathrm{d}r^2 + \sinh^2 r\,\mathrm{d}\phi^2\right).$$

These two forms can be combined by defining a new radial coordinate that makes the transverse part of the metric look Euclidean:

$$\mathrm{d}\sigma^2 = R^2\left(\frac{\mathrm{d}r^2}{1 - kr^2} + r^2\,\mathrm{d}\phi^2\right)$$

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

where $k = +1$ for positive curvature and $k = -1$ for negative curvature.

This is in fact the general form of the spatial part of the Robertson–Walker metric. To prove this in 3D, consider a 3-sphere embedded in four-dimensional Euclidean space, which is defined via the coordinate relation $x^2 + y^2 + z^2 + w^2 = R^2$. Now define the equivalent of spherical polars and write $w = R\cos\alpha$, $z = R\sin\alpha\cos\beta$, $y = R\sin\alpha\sin\beta\cos\gamma$, $x = R\sin\alpha\sin\beta\sin\gamma$, where α, β and γ are three arbitrary angles. Differentiating with respect to the angles gives a four-dimensional vector (dx,dy,dz,dw), and it is a straightforward exercise to show that the squared length of this vector

$$|(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z, \mathrm{d}w)|^2 = R^2[\mathrm{d}\alpha^2 + \sin^2\alpha(\mathrm{d}\beta^2 + \sin^2\beta\ \mathrm{d}\gamma^2)]$$

which is the Robertson–Walker metric for the case of positive spatial curvature.

This $k = +1$ metric describes a closed universe, in which a traveler who sets off along a trajectory of fixed β and γ will eventually return to their starting point (when $\alpha = 2\pi$). In this respect, the positively curved 3D universe is identical to the case of the surface of a sphere: it is finite, but unbounded. By contrast, the $k = -1$ metric describes an open universe of infinite extent; as before, changing to negative spatial curvature replaces sin α with sinh α, and α can be made as large as we please without returning to the starting point. The $k = 0$ model describes a flat universe, which is also infinite in extent. This can be thought of as a limit of either of the $k = \pm 1$ cases, where the curvature scale R tends to infinity.

The Robertson–Walker metric may be written in a number of different ways. The most compact forms are those where the comoving coordinates are dimensionless. In terms of the function

$$S_k(r) = \begin{cases} \sin r & (k=1) \\ \sinh r & (k=-1) \\ r & (k=0) \end{cases}$$

the metric can be written as

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}r^2 + S_k^2(r)\,\mathrm{d}\psi^2].$$

The most common alternative is to use a different definition of comoving distance, $S_k(r) \to r$, so that the metric becomes

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\,\mathrm{d}\psi^2\right).$$

There should of course be two different symbols for the different comoving radii, but each is often called r in the literature. Finally, a common alternative form of the scale factor is where its present value is set to t)\equiv

$$a(t) \equiv R(t)/R_0.$$

The redshift

How does this discussion relate to HUBBLE'S LAW: $v = Hr$? Comoving coordinates are time independent, so the proper separation of two fundamental observers is just $R(t)\,\mathrm{d}r$, and differentiation gives Hubble's law, $v = H(R\,\mathrm{d}r)$, with

$$H = \frac{\dot{R}}{R}.$$

At small separations, the recessional velocity gives the Doppler shift (see DOPPLER EFFECT):

$$\frac{\nu_{\mathrm{emit}}}{\nu_{\mathrm{obs}}} \equiv 1 + z \simeq 1 + \frac{v}{c}.$$

This defines the REDSHIFT z in terms of the shift of spectral lines. What is the equivalent of this relation at larger distances? Since photons travel on null geodesics of zero proper time, we see directly from the metric that

$$r = \int \frac{c\,\mathrm{d}t}{R(t)}.$$

The comoving distance is constant, whereas the domain of integration in time extends from t_{emit} to t_{obs}; these are the times of emission and reception of a photon. Photons that are emitted at later times will be received at later times, but these changes in t_{emit} and t_{obs} cannot alter the integral, since r is a comoving quantity. This requires the condition $\mathrm{d}t_{\mathrm{emit}} / \mathrm{d}t_{obs} = R(t_{emit}) / R(t_{obs})$, which means that events on distant galaxies time dilate according to how much the universe has expanded since the photons we see now were emitted. Clearly (think of events separated by one period), this dilation also applies to frequency, and we therefore obtain

$$\frac{\nu_{\mathrm{emit}}}{\nu_{\mathrm{obs}}} \equiv 1 + z = \frac{R(t_{\mathrm{obs}})}{R(t_{\mathrm{emit}})}.$$

In terms of the normalized scale factor $a(t)$, $a(t) = (1 + z)^{-1}$. Photon wavelengths therefore stretch with the universe, as is intuitively reasonable.

The meaning of the redshift

For small redshifts, the interpretation of the redshift as a Doppler shift ($z = v/c$) is quite clear. What is not so clear is what to do when the redshift becomes large. A common but incorrect approach is to use the special-relativistic Doppler formula and write

$$1 + z = \left(\frac{1 + v/c}{1 - v/c}\right)^{1/2}.$$

This is wrong in general, but it is all too common to read of the latest high-redshift QUASAR as 'receding at 95% of the speed of light'. The reason the redshift cannot be interpreted in this way is because a non-zero mass density must cause gravitational redshifts.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

However, although the redshift cannot be thought of as a global Doppler shift, it is correct to think of the effect as an accumulation of the infinitesimal Doppler shifts caused by photons passing between fundamental observers separated by a small distance:

$$\frac{\delta z}{1+z} = H(z)\,\delta\ell(z)$$

(where $\delta\ldots$ is a radial increment of proper distance). This expression may be verified by substitution of the standard expressions for $H(z)$ and $\mathrm{d}\ldots/\mathrm{d}z$. The nice thing about this way of looking at the result is that it emphasizes that it is momentum that is redshifted; particle de Broglie wavelengths thus scale with the expansion, a result that is independent of whether their rest mass is non-zero.

An inability to see that the expansion is locally just kinematical also lies at the root of perhaps the worst misconception about the BIG BANG THEORY. Many semipopular accounts of cosmology contain statements to the effect that 'space itself is swelling up' in causing the galaxies to separate. In fact, objects separate now only because they have done so in the past; a pair of massless objects set up at rest with respect to each other in a uniform model will show no tendency to separate (in fact, the gravitational force of the mass lying between them will cause an inward relative acceleration). In the common elementary demonstration of the expansion by means of inflating a balloon, galaxies should be represented by glued-on coins, not ink drawings (which will spuriously expand with the universe).

Dynamics of the expansion

The equation of motion for the scale factor can be obtained in a quasi-Newtonian fashion. Consider a sphere about some arbitrary point, and let the radius be $R(trr$, where r is arbitrary. The motion of a point at the edge of the sphere will, in Newtonian gravity, be influenced only by the interior mass. We can therefore apparently write down immediately a differential equation (Friedmann's equation) that expresses conservation of energy: $(\dot{R}r)^2/2 – GM/Rr = constant$. In fact, this equation really requires general relativity: the gravitation from mass shells at large distances is not Newtonian, because space is curved, and so we cannot employ the usual argument about their effect being zero. Nevertheless, the result that the gravitational field inside a uniform shell is zero does hold in general relativity, and is known as Birkhoff's theorem. General relativity becomes even more vital in giving the constant of integration in Friedmann's equation:

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2.$$

Note that this equation covers all contributions to ρ, i.e. those from matter, radiation and vacuum; it is independent of the equation of state.

It is sometimes convenient to work with the time derivative of the Friedmann equation, for the same reason that acceleration arguments in dynamics are sometimes more transparent than energy ones. Differentiating with respect to time requires a knowledge of $\dot{\rho}$, but this can be eliminated by means of conservation of energy: $\mathrm{d}(\rho c^2 R^3) = -p\,\mathrm{d}(R^3)$. We then obtain

$$\ddot{R} = -4\pi G R(\rho c^2 + 3p)/3.$$

Both this equation and the Friedmann equation in fact arise as independent equations from different components of Einstein's equations for the Robertson–Walker metric.

The Friedmann equation is so named because FRIEDMANN was the first to appreciate, in 1922, that Einstein's equations admitted cosmological solutions containing matter only (although it was Lemaître who in 1927 both obtained the solution and appreciated that it led to a linear distance–redshift relation). The term Friedmann model is therefore often used to indicate a matter-only cosmology, even though his equation includes contributions from all equations of state. A common shorthand for relativistic cosmological models, which are described by the Robertson–Walker metric and which obey the Friedmann equation, is to speak of FRW models.

Density parameters etc

According to the Friedmann equation, the 'flat' universe with $k = 0$ arises for a particular critical density. We are therefore led to define a density parameter as the ratio of density to critical density:

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}.$$

Since ρ and H change with time, this defines an epoch-dependent density parameter. The current value of the parameter should strictly be denoted by Ω_0. Because this is such a common symbol, it is normal to keep the formulae uncluttered by normally dropping the subscript; the density parameter at other epochs will be denoted by $\Omega(z)$. If we now also define a dimensionless (current) Hubble parameter as

$$h \equiv \frac{H_0}{100\ \mathrm{km\ s^{-1}\ Mpc^{-1}}}$$

then the current density of the universe may be expressed as

$$\begin{aligned}\rho_0 &= 1.88 \times 10^{-26}\Omega h^2\ \mathrm{kg\ m^{-3}} \\ &= 2.78 \times 10^{11}\Omega h^2 M_\odot\ \mathrm{Mpc^{-3}}.\end{aligned}$$

A powerful approximate model for the energy content of the universe is to divide it into pressureless matter ($\rho : R^{-3}$), radiation ($\rho : R^{-4}$) and vacuum energy (ρ independent of time—i.e. there is a non-zero cosmological constant). The first two relations just say that the number

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

density of particles is diluted by the expansion, with photons also having their energy reduced by the redshift; the third relation applies for Einstein's cosmological constant. In terms of observables, this means that the density is written as (where the normalized scale factor is $a = R/R_0$).

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_\mathrm{v} + \Omega_\mathrm{m}a^{-3} + \Omega_\mathrm{r}a^{-4})$$

In terms of the deceleration parameter,

$$q \equiv -\frac{\ddot{R}R}{\dot{R}^2}$$

the $\ddot{R}$ form of the Friedmann equation says that

$$q = \Omega_\mathrm{m}/2 + \Omega_\mathrm{r} - \Omega_\mathrm{v}.$$

Lastly, it is often necessary to know the present value of the scale factor, which may be read directly from the Friedmann equation:

$$R_0 = \frac{c}{H_0}[(\Omega - 1)/k]^{-1/2}.$$

The HUBBLE CONSTANT thus sets the curvature length, which becomes infinitely large as Ω approaches unity from either direction. Only in the limit of zero density does this length become equal to the other common measure of the size of the universe—the Hubble length, c/H_0.

Solutions to the Friedmann equation

The Friedmann equation may be solved most simply in 'parametric' form, by recasting it in terms of the conformal time $\mathrm{d}\eta = c\,\mathrm{d}t/R$ (denoting derivatives with respect to η by primes):

$$R'^2 = \frac{8\pi G}{3c^2}\rho R^4 - kR^2.$$

Because $H_0^2R_0^2 = kc^2/(\Omega–1)$, the Friedmann equation becomes

$$a'^2 = \frac{k}{\Omega - 1}[\Omega_\mathrm{r} + \Omega_\mathrm{m}a - (\Omega - 1)a^2 + \Omega_\mathrm{v}a^4]$$

which is straightforward to integrate provided that $\Omega_\mathrm{v} = 0$. Solving the Friedmann equation for $R(t)$ in this way is important for determining global quantities such as the present age of the universe, and explicit solutions for particular cases are considered below. However, from the point of view of observations, and in particular the distance–redshift relation, it is not necessary to proceed by the direct route of determining $R(t)$.

To the observer, the evolution of the scale factor is most directly characterized by the change with redshift of the Hubble parameter and the density parameter; the evolution of $H(z)$ and $\Omega(z)$ is given immediately by the Friedmann equation in the form $H^2 = 8\pi G\rho/3–kc^2/R^2$. Inserting the model dependence of ρ on a gives

$$H^2(a) = H_0^2[\Omega_\mathrm{v} + \Omega_\mathrm{m}a^{-3} + \Omega_\mathrm{r}a^{-4} - (\Omega - 1)a^{-2}].$$

This is a crucial equation, which can be used to obtain the relation between redshift and comoving distance. The radial equation of motion for a photon is $R\,\mathrm{d}r = c\,\mathrm{d}t = c\,\mathrm{d}R/\dot{R} = c\,\mathrm{d}R/RH$. With $R = R_0/(1+z)$, this gives

$$R_0\,\mathrm{d}r = \frac{c}{H(z)}\,\mathrm{d}z = \frac{c}{H_0}[\Omega_\mathrm{v} + \Omega_\mathrm{m}a^{-3} + \Omega_\mathrm{r}a^{-4} - (\Omega - 1)a^{-2}]^{-1/2}\,\mathrm{d}z.$$

This relation is arguably the single most important equation in cosmology, since it shows how to relate comoving distance to the observables of redshift, Hubble constant and density parameters. The comoving distance determines the apparent brightness of distant objects, and the comoving volume element determines the numbers of objects that are observed. These aspects of observational cosmology are discussed in more detail below.

Lastly, using the expression for $H(z)$ with $\Omega(a)–1 = kc^2/H^2R^2$ gives the redshift dependence of the total density parameter:

$$\Omega(z) - 1 = \frac{\Omega - 1}{1 - \Omega + \Omega_\mathrm{v}a^2 + \Omega_\mathrm{m}a^{-1} + \Omega_\mathrm{r}a^{-2}}.$$

This last equation is very important. It tells us that, at high redshift, all model universes apart from those with only vacuum energy will tend to look like the $\Omega = 1$ model. This is not surprising given the form of the Friedmann equation: provided that $\rho R^2 \to \infty$ as $R \to 0$, the $–kc^2$ curvature term will become negligible at early times. If $\Omega \neq 1$, then in the distant past $\Omega(z)$ must have differed from unity by a tiny amount: the density and rate of expansion needed to have been finely balanced for the universe to expand to the present. This tuning of the initial conditions is called the flatness problem and is one of the motivations for the applications of quantum theory to the early universe.

Matter-dominated universe

From the observed temperature of the microwave background (2.73 K) and the assumption of three species of neutrino at a slightly lower temperature, we deduce that the total relativistic density parameter is $\Omega_\mathrm{r}h^2 \approx 4.2–10^{–5}$, so at present it should be a good approximation to ignore radiation. However, the different redshift dependences of matter and radiation densities mean that this assumption fails at early times: $\rho_\mathrm{m}/\rho_r : (1 + z)^{–1}$. One of the critical epochs in cosmology is therefore the point at which these contributions were equal: the redshift of matter–radiation equality

$$1 + z_\mathrm{eq} \simeq 23\,900\Omega h^2.$$

At redshifts higher than this, the universal dynamics was dominated by the relativistic-particle content. By an

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

interesting coincidence, this epoch is close to another important event in cosmological history: recombination. Once the temperature falls below $\approx 10^4$ K, ionized material can form neutral hydrogen. Observational astronomy is only possible from this point on, since Thomson scattering from electrons in ionized material prevents photon propagation. In practice, this limits the maximum redshift of observational interest to about 1100; unless Ω is very low or vacuum energy is important, a matter-dominated model is therefore a good approximation to reality.

By conserving matter, we can introduce a characteristic mass M_*, and from this a characteristic radius R_*:

$$\frac{4\pi G}{3c^2}\rho R^3 = \frac{c}{H_0}\frac{\Omega[k(\Omega-1)]^{3/2}}{2} \equiv \frac{GM_*}{c^2} \equiv R_*$$

where we have used the expression for R_0 in the first step. When only matter is present, the conformal-time version of the Friedmann equation is simple to integrate for $R(\eta)$, and integration of $\mathrm{d}t = \mathrm{d}\eta/R$ gives $t(\eta)$:

$$R = kR_*[1 - C_k(\eta)]$$

$$ct = kR_*[\eta - S_k(\eta)].$$

The evolution of $R(t)$ in this solution is plotted in figure 1. A particular point to note is that the behavior at early times is always the same: potential and kinetic energies greatly exceed total energy and we always have the $k = 0$ form $R{:}t^{2/3}$.

Radiation-dominated universe

At high enough redshifts, the $R{:}t^{2/3}$ law will fail, because radiation pressure will become important. At these redshifts, it is an excellent approximation to ignore the effects of spatial curvature, so that the Friedmann equation for a matter–radiation mix is

$$\dot{a}^2 = H_0^2(\Omega_\mathrm{m}a^{-1} + \Omega_\mathrm{r}a^{-2}).$$

This may be integrated to give the time as a function of scale factor:

$$H_0 t = \frac{2}{3\Omega_\mathrm{m}^2}[(\Omega_\mathrm{r} + \Omega_\mathrm{m}a)^{1/2}(\Omega_\mathrm{m}a - 2\Omega_\mathrm{r}) + 2\Omega_\mathrm{r}^{3/2}]$$

which goes to $\frac{2}{3}a^{3/2}$ for a matter-only model and to $a^2/2$ for radiation only. At early times, the scale factor thus grows as $R{:}t^{1/2}$.

One further way of presenting the model's dependence on time is via the density. Following the above, it is easy to show that

$$t = \begin{cases} \left(\dfrac{1}{6\pi G\rho}\right)^{1/2} & \text{(matter domination)} \\ \left(\dfrac{3}{32\pi G\rho}\right)^{1/2} & \text{(radiation domination).} \end{cases}$$

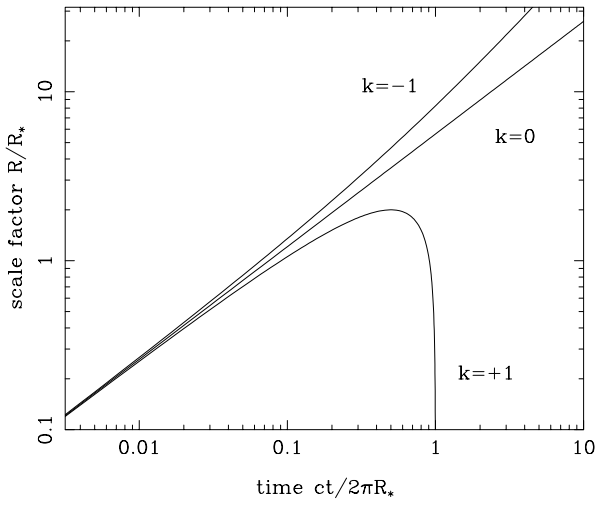


Figure 1. The time dependence of the scale factor for open, closed and critical matter-dominated cosmological models. The logarithmic scale is designed to bring out the early-time behavior, although it obscures the fact that the closed model is a symmetric cycloid on a linear plot of R against t. Reproduced from *Cosmological Physics* (Cambridge University Press).

Models with vacuum energy

The solution of the Friedmann equation becomes more complicated if we allow a significant contribution from vacuum energy—i.e. a non-zero cosmological constant. The Friedmann equation itself is independent of the equation of state, and just says $H^2R^2 = kc^2/(\Omega–1)$, whatever the form of the contributions to Ω. In terms of the cosmological constant itself, we have

$$\Omega_\mathrm{vac} = \frac{8\pi G\rho_\mathrm{vac}}{3H^2} = \frac{\Lambda c^2}{3H^2}.$$

The reason that the cosmological constant was first introduced by Einstein was not simply because there was no general reason to expect empty space to be of zero density, but because it allows a non-expanding cosmology to be constructed. This is perhaps not so obvious from some forms of the Friedmann equation, since now $H = 0$ and $\Omega = \infty$; if we cast the equation in its original form without defining these parameters, then zero expansion implies

$$\rho = \frac{3kc^2}{8\pi GR^2}.$$

Since Λ can have either sign, this appears not to constrain k. However, we also want to have zero acceleration for this model, and so need the time derivative of the Friedmann equation: $\ddot{R} = -4\pi GR(\rho + 3p)/3$. A further condition for a static model is therefore that

$$\rho = -3p.$$

Since $\rho = -p$ for vacuum energy, and this is the only source of pressure if we ignore radiation, this tells us that $\rho = 3\rho_\mathrm{vac}$ and hence that the mass density is twice the vacuum density. The total density is hence positive and $k = 1$; we have a closed model.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Notice that what this says is that a positive vacuum energy acts in a repulsive way, balancing the attraction of normal matter. This shows that the static model cannot be stable: if we perturb the scale factor by a small positive amount, the vacuum repulsion is unchanged whereas the 'normal' gravitational attraction is reduced, so that the model will tend to expand further (or contract, if the initial perturbation was negative).

de Sitter space

The endpoint of an outwards perturbation of Einstein's static model was first studied by DE SITTER. This universe is completely dominated by vacuum energy and is clearly the limit of the unstable expansion, since the density of matter redshifts to zero while the vacuum energy remains constant. Consider again the Friedmann equation in its general form $\dot{R}^2 - 8\pi G\rho R^2/3 = -kc^2$: since the density is constant and R will increase without limit, the two terms on the lhs must eventually become almost exactly equal and the curvature term on the rhs will be negligible. Thus, even if $k \neq 0$, the universe will have a density that differs only infinitesimally from the critical, so that we can solve the equation by setting $k = 0$, in which case

$$R \propto \exp(Ht) \qquad H = \left(\frac{8\pi G\rho_{\rm vac}}{3}\right)^{1/2} = \left(\frac{\Lambda c^2}{3}\right)^{1/2}.$$

An interesting interpretation of this behavior was promoted in the early days of cosmology by Eddington: the cosmological constant is what caused the expansion. In models without Λ, the expansion is merely an initial condition: anyone who asks why the universe expands at a given epoch is given the unsatisfactory reply that it does so because it was expanding at some earlier time. It would be more satisfying to have some mechanism that set the expansion into motion, and this is what is provided by vacuum repulsion. This tendency of models with positive Λ to end up undergoing an exponential phase of expansion (and moreover one with $\Omega = 1$) is exactly what is used in inflationary cosmology to generate the initial conditions for the big bang.

The steady-state model

The behavior of de Sitter space is in some ways reminiscent of the steady-state universe, which was popular in the 1960s. This STEADY-STATE THEORY drew its motivation from the philosophical problems of big-bang models—which begin in a singularity at $t = 0$, and for which earlier times have no meaning. Instead, Hoyle, Bondi and Gold suggested the perfect cosmological principle in which the universe is homogeneous not only in space but also in time: apart from local fluctuations, the universe appears the same to all observers at all times. This tells us that the Hubble constant really is constant, and so the model necessarily has exponential expansion, $R : \exp(Ht)$, exactly as for de Sitter space. Indeed, de Sitter space is a steady-state universe: it contains a constant vacuum energy density and has an infinite age, lacking any big-bang singularity. However, de Sitter space is a rather uninteresting model because it contains no matter. Introducing matter into a steady-state universe violates energy conservation, since matter does not have the $p = -\rho c^2$ equation of state that allows the density to remain constant. This is the most radical aspect of steady-state models: they require continuous creation of matter. The energy to accomplish this has to come from somewhere, and Einstein's equations are modified by adding some 'creation' or '*C*-field' term to the energy–momentum tensor:

$$T'^{\mu\nu} = T^{\mu\nu} + C^{\mu\nu} \qquad T'^{\mu\nu}_{;\nu} = 0.$$

The effect of this extra term must be to cancel the matter density and pressure, leaving just the overall effective form of the vacuum tensor, which is required to produce de Sitter space and the exponential expansion. This *ad hoc* field and the lack of any physical motivation for it beyond the cosmological problem it was designed to solve was always the most unsatisfactory feature of the steady-state model, and may account for the strong reactions generated by the theory.

Bouncing and loitering models

Returning to the general case of models with a mixture of energy in the vacuum and normal components, we have to distinguish three cases. For models that start from a big bang (in which case radiation dominates completely at the earliest times), the universe will either recollapse or expand forever. The latter outcome becomes more likely for low densities of matter and radiation, but high vacuum density. It is, however, also possible to have models in which there is no big bang: the universe was collapsing in the distant past, but was slowed by the repulsion of a positive Λ term and underwent a 'bounce' to reach its present state of expansion. Working out the conditions for these different events is a matter of integrating the Friedmann equation. For the addition of Λ, this can only in general be done numerically. However, we can find the conditions for the different behaviors described above analytically, at least if we simplify things by ignoring radiation. The equation in the form of the time-dependent Hubble parameter looks like

$$\frac{H^2}{H_0^2} = \Omega_{\rm v}(1 - a^{-2}) + \Omega_{\rm m}(a^{-3} - a^{-2}) + a^{-2}$$

and we are interested in the conditions under which the lhs vanishes, defining a turning point in the expansion. Setting the rhs to zero yields a cubic equation, and it is possible to give the conditions under which this has a solution, which are as follows.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

(1) First, negative Λ always implies recollapse, which is intuitively reasonable (either the mass causes recollapse before Λ dominates, or the density is low enough that Λ comes to dominate, which cannot lead to infinite expansion unless Λ is positive).
(2) If Λ is positive and $\Omega_m<1$, the model always expands to infinity.
(3) If $\Omega_m>1$, recollapse is only avoided if Ω_v exceeds a critical value

$$\Omega_v > 4\Omega_m\{\cos[\tfrac{1}{3}\cos^{-1}(\Omega_m^{-1}-1)+\tfrac{4}{3}\pi]\}^3.$$

(4) If Λ is large enough, the stationary point of the expansion is at $a<1$ and we have a bounce cosmology. This critical value is

$$\Omega_v > 4\Omega_m\{f[\tfrac{1}{3}f^{-1}(\Omega_m^{-1}-1)]\}^3$$

where the function f is cosh if $\Omega_m<0.5$, otherwise cos. If the universe lies exactly on the critical line, the bounce is at infinitely early times and we have a solution that is the result of a perturbation of the Einstein static model.

In fact, bounce models can be ruled out quite strongly. The same cubic equations that define the critical conditions for a bounce also give an inequality for the maximum redshift possible (that of the bounce):

$$1+z_B \leqslant 2f[\tfrac{1}{3}f^{-1}(\Omega_m^{-1}-1)].$$

A reasonable lower limit for Ω_m of 0.1 then rules out a bounce once objects are seen at $z>2$.

The main results of this section are summed up in figure 2. Since the radiation density is very small today, the main task of relativistic cosmology is to work out where on the Ω_{matter}–Ω_{vacuum} plane the real universe lies. The existence of high-redshift objects rules out the bounce models, so that the idea of a hot big bang cannot be evaded.

At this point, we have reproduced one of the great conclusions of relativistic cosmology: the universe is of finite age, and had its origin in a mathematical singularity at which the scale factor went to zero, leading to a divergent spacetime curvature. Since zero scale factor also implies infinite density (and temperature), the inferred picture of the early universe is one of unimaginable violence. The term big bang was coined by FRED HOYLE to describe this beginning, although it was intended somewhat critically. The problem with the singularity is that it marks the breakdown of the laws of physics; we cannot extrapolate the solution for $R(t)$ to $t<0$, and so the origin of the expansion becomes an unexplained boundary condition. It was only after about 1980 that a consistent set of ideas became available for ways of avoiding this barrier, in the form of inflationary cosmology.

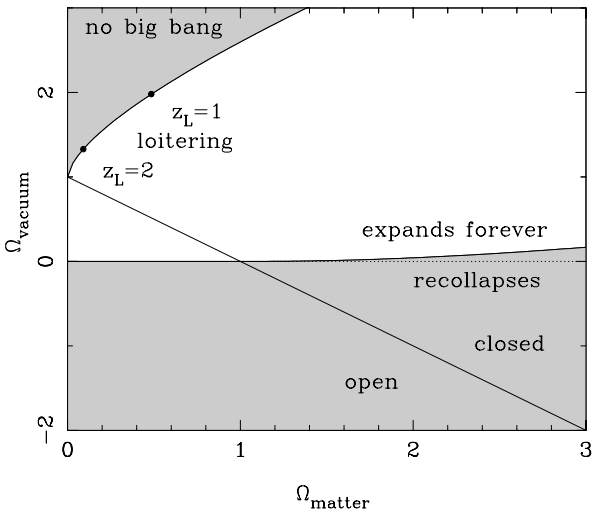


Figure 2. This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega>1$ are always spatially closed (open for $\Omega<1$), although closed models can still expand to infinity if $\Omega_v\neq0$. If the cosmological constant is negative, recollapse always occurs; recollapse is also possible with a positive Ω_v if $\Omega_m\gg\Omega_v$. If $\Omega_v>1$ and Ω_m is small, there is the possibility of a 'loitering' solution with some maximum redshift and infinite age (top left); for even larger values of vacuum energy, there is no big-bang singularity. Reproduced from *Cosmological Physics* (Cambridge University Press).

Flat universe

The most important model in cosmological research is that with $k = 0 \Rightarrow \Omega_{total} = 1$; when dominated by matter, this is often termed the Einstein–de Sitter model. Paradoxically, this importance arises because it is an unstable state: as we have seen earlier, the universe will evolve away from $\Omega = 1$, given a slight perturbation. For the universe to have expanded by so many e-foldings (factors of e expansion) and yet still have Ωd1 implies that it was very close to being spatially flat at early times. Many workers have therefore conjectured that it would be contrived if this flatness was other than perfect.

An alternative $k = 0$ model of greater observational interest has a significant cosmological constant, so that $\Omega_m + \Omega v = 1$ (radiation being neglected for simplicity). This may seem contrived, but once k=0 has been established, it cannot change: individual contributions to Ω must adjust to keep in balance. The advantage of this model is that it is the only way of retaining the theoretical attractiveness of $k = 0$ while changing the age of the universe from the relation $H_0t_0 = 2/3$, which characterizes the Einstein–de Sitter model. Since much observational evidence indicates that $H_0t_0\simeq1$, this model has received a good deal of interest in recent years. For this case, the Friedmann equation is

$$\dot{a}^2 = H_0^2[\Omega_m a^{-1} + (1-\Omega_m)a^2]$$

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

and the $t(a)$ relation is

$$H_0 t\,(a) = \int_0^a \frac{x\,\mathrm{d}x}{[\Omega_\mathrm{m} x + (1-\Omega_\mathrm{m})x^4]^{1/2}}$$

which integrates to

$$H_0 t\,(a) = \frac{2}{3}\frac{S_k^{-1}[(a^3|\Omega_\mathrm{m}-1|/\Omega_\mathrm{m})^{1/2}]}{(|\Omega_\mathrm{m}-1|)^{1/2}}.$$

Here, k in S_k is used to mean sin if $\Omega_\mathrm{m}>1$, otherwise sinh; these are still $k=0$ models. This $t(a)$ relation is compared with models without vacuum energy in figure 3. Since there is nothing special about the current era, we can clearly also rewrite this expression as

$$H(a)t\,(a) = \frac{2}{3}\frac{S_k^{-1}\{[|\Omega_\mathrm{m}(a)-1|/\Omega_\mathrm{m}(a)]^{1/2}\}}{[|\Omega_\mathrm{m}(a)-1|]^{1/2}} \simeq \frac{2}{3}\Omega_\mathrm{m}(a)^{-0.3}$$

where we include a simple approximation that is accurate to a few % over the region of interest ($\Omega_\mathrm{m}\gtrsim 0.1$). In the general case of significant Λ but $k\simeq 0$, this expression still gives a very good approximation to the exact result, provided that Ω_m is replaced by $0.7\Omega_\mathrm{m}-0.3\Omega_v+0.3$ (Carroll *et al* 1992).

Horizons

For photons, the radial equation of motion is just $c\,\mathrm{d}t=R\,\mathrm{d}r$. How far can a photon get in a given time? The answer is clearly

$$\Delta r = \int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{R(t)} = \Delta\eta$$

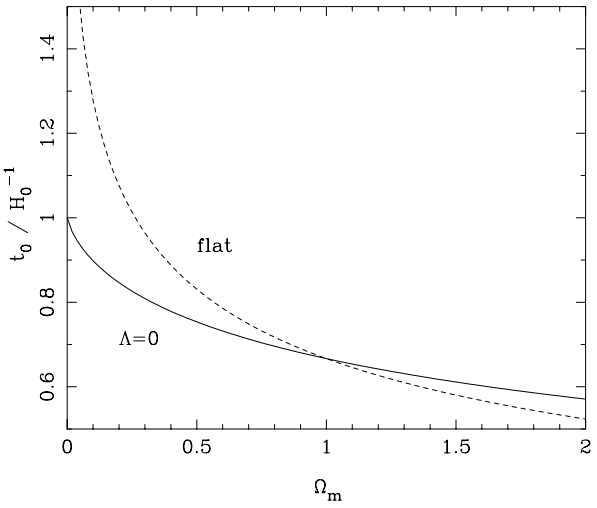


Figure 3 Age of the universe versus matter density parameter. The solid curve shows open models; the dotted curve shows flat models with $\Omega_\mathrm{m}+\Omega_v=1$. The accelerating effects of vacuum energy allow $H_0t_0\gtrsim 1$ for $\Omega_\mathrm{m}\gtrsim 0.25$ in the flat case, which is much older than the open model of the same density. Reproduced from *Cosmological Physics* (Cambridge University Press).

i.e. just the interval of conformal time. What happens as $t_0\to 0$ in this expression? We can replace $\mathrm{d}t$ by $\mathrm{d}R/\dot{R}$, which the Friedmann equation says is $:\mathrm{d}R/(\rho R^2)^{1/2}$ at early times. Thus, this integral converges if $\rho R^2\to\infty$ as $t_0\to 0$, otherwise it diverges. Provided that the equation of state is such that ρ changes faster than R^{-2}, light signals can only propagate a finite distance between the big bang and the present; there is then said to be a particle horizon. Such a horizon therefore exists in conventional big-bang models, which are dominated by radiation at early times.

A particle horizon is not at all the same thing as an event horizon: for the latter, we ask whether Δr diverges as $t\to\infty$. If it does, then seeing a given event is just a question of waiting long enough. Clearly, an event horizon requires $R(t)$ to increase more quickly than t, so that distant parts of the universe recede faster than light. This does not occur unless the universe is dominated by vacuum energy at late times. Despite this distinction, cosmologists usually say 'the horizon' when they mean 'the particle horizon'.

Observations in cosmology

The main task of observational cosmology is to use observed quantities such as fluxes and angular sizes in order to deduce intrinsic quantities such as luminosities and physical sizes for a distant object. This conversion requires a knowledge of distance, which is not directly measurable; all that we know about a distant object is its redshift. Observers therefore place heavy reliance on formulae for expressing distance in terms of redshift. For high-redshift objects such as quasars, this has led to a history of controversy over whether a component of the redshift could be of non-cosmological origin:

$$1+z_\mathrm{obs} = (1+z_\mathrm{cos})(1+z_\mathrm{extra}).$$

For now, we assume that the cosmological contribution to the redshift can always be identified.

We now assemble some essential formulae for interpreting cosmological observations. Our observables are redshift, z, and angular difference between two points on the sky, $\mathrm{d}\psi$. We write the metric in the form

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}r^2 + S_k^2(r)\,\mathrm{d}\psi^2]$$

so that the comoving volume element is

$$\mathrm{d}V = 4\pi[R_0 S_k(r)]^2\,R_0\,\mathrm{d}r.$$

The proper transverse size of an object seen by us is its comoving size $\mathrm{d}\psi$; $S_k(r)$ times the scale factor at the time of emission:

$$\mathrm{d}\ell = \mathrm{d}\psi\;R_0 S_k(r)/(1+z).$$

Probably the most important relation for observational cosmology is that between monochromatic flux density and luminosity. Start by assuming isotropic emission, so

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

that the photons emitted by the source pass with a uniform flux density through any sphere surrounding the source. We can now make a shift of origin, and consider the Robertson–Walker metric as being centered on the source; however, because of homogeneity, the comoving distance between the source and the observer is the same as we would calculate when we place the origin at our location. The photons from the source are therefore passing through a sphere, on which we sit, of proper surface area $4\pi[R_0 S_k(r)]^2$. However, redshift still affects the flux density in four further ways: photon energies and arrival rates are redshifted, reducing the flux density by a factor $(1+z)^2$; opposing this, the bandwidth $\mathrm{d}\nu$ is reduced by a factor $1+z$, so the energy flux per unit bandwidth goes down by one power of $1+z$; finally, the observed photons at frequency ν_0 were emitted at frequency $\nu_0(1+z)$, so the flux density is the luminosity at this frequency, divided by the total area, divided by $1+z$:

$$S_\nu(\nu_0) = \frac{L_\nu[(1+z)\nu_0]}{4\pi R_0^2 S_k^2(r)(1+z)}.$$

A word about units: L_ν in this equation would be measured in units of W Hz^{-1}. Recognizing that emission is often not isotropic, it is common to consider instead the luminosity emitted into unit solid angle—in which case there would be no factor of 4π, and the units of L_ν would be W Hz^{-1} sr^{-1}.

The flux density received by a given observer can be expressed by definition as the product of the specific intensity I_ν (the flux density received from unit solid angle of the sky) and the solid angle subtended by the source: $S_\nu = I_\nu\ \mathrm{d}\Omega$. Combining the angular size and flux–density relations thus gives the relativistic version of surface-brightness conservation. This is independent of cosmology:

$$I_\nu(\nu_0) = \frac{B_\nu[(1+z)\nu_0]}{(1+z)^3}$$

where B_ν is surface brightness (luminosity emitted into unit solid angle per unit area of source). We can integrate over ν_0 to obtain the corresponding total or bolometric formulae, which are needed for example for spectral-line emission: The form of these relations lead to

$$S_{\mathrm{tot}} = \frac{L_{\mathrm{tot}}}{4\pi R_0^2 S_k^2(r)(1+z)^2}$$

$$I_{\mathrm{tot}} = \frac{B_{\mathrm{tot}}}{(1+z)^4}.$$

the following definitions for particular kinds of distances: angular-diameter distance is

$$D_{\mathrm{A}} = (1+z)^{-1} R_0 S_k(r)$$

luminosity distance is

$$D_{\mathrm{L}} = (1+z)\ R_0 S_k(r).$$

Angular-diameter distance versus redshift is illustrated in figure 4.

The last element needed for the analysis of observations is a relation between redshift and age for the object being studied. This brings in our earlier relation between time and comoving radius (consider a null geodesic traversed by a photon that arrives at the present):

$$c\ \mathrm{d}t = R_0\ \mathrm{d}r/(1+z).$$

So far, all this is completely general; to complete the toolkit, we need the crucial input of relativistic dynamics, which is to give the distance–redshift relation.

Distance–redshift relation

The general relation between comoving distance and redshift was given earlier as

$$R_0\ \mathrm{d}r = \frac{c}{H(z)}\ \mathrm{d}z = \frac{c}{H_0}[\Omega_\nu + \Omega_{\mathrm{m}} a^{-3} + \Omega_{\mathrm{r}} a^{-4} - (\Omega - 1)a^{-2}]^{-1/2}\ \mathrm{d}z.$$

For a matter-dominated Friedmann model, this means that the distance of an object from which we receive photons today is

$$R_0 r = \frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{(1+z')(1+\Omega z')^{1/2}}.$$

Integrals of this form often arise when manipulating Friedmann models; they can usually be tackled by the substitution $u^2 = k(\Omega{-}1)/\Omega(1{+}z)$. This substitution produces Mattig's formula, which is one of the single most useful equations in cosmology as far as observers are concerned:

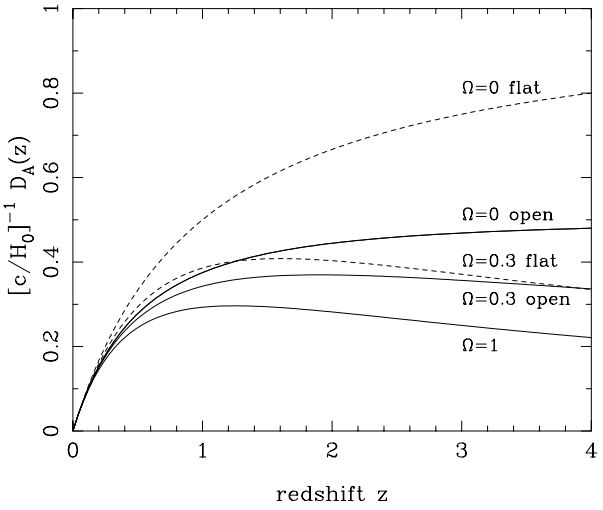


Figure 4. A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid curves show models with zero vacuum energy; dashed curves show flat models with $\Omega_{\mathrm{m}} + \Omega_v = 1$. In both cases, results for $\Omega_{\mathrm{m}} = 1,0.3,0$ are shown; higher density results in lower distance at high z, owing to gravitational focusing of light rays. Reproduced from *Cosmological Physics* (Cambridge University Press).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

$$R_0 S_k(r) = \frac{2c}{H_0} \frac{\Omega z + (\Omega - 2)[(1 + \Omega z)^{1/2} - 1]}{\Omega^2 (1 + z)}.$$

There is no such compact expression if one wishes to allow for vacuum energy as well. The comoving distance has to be obtained by numerical integration of the fundamental $\mathrm{d}r/\mathrm{d}z$, even in the $k = 0$ case. However, for all forms of contribution to the energy content of the universe, the second-order distance–redshift relation is identical, and depends only on the deceleration parameter:

$$R_0 S_k(r) \simeq \frac{c}{H_0} \left(z - \frac{1 + q_0}{2} z^2 \right).$$

The sizes and flux densities of objects at moderate redshift therefore determine the geometry of the universe only once an equation of state is assumed, so that q_0 and Ω_0 can be related. At larger redshifts, this degeneracy is broken, and accurate measurements of the distance–redshift relation can in principle determine the parameters Ω_m, Ω_v etc independently.

Recent observations

NASA's Far Ultraviolet Spectroscopic Explorer (FUSE) satellite has given astronomers a glimpse of the ghostly cobweb of helium gas left over from the Big Bang, which underlies the universe's structure. The helium is not found in galaxies or stars but spread thinly through space. The observations help confirm theoretical models of how matter in the expanding universe condensed into a web-like structure pervading all the space between galaxies. The helium traces the architecture of the universe back to very early times. This structure arose from small gravitational instabilities seeded in the chaos just after the Big Bang.

Outstanding issues

As outlined above, the basic isotropic models of relativistic cosmology depend on four main parameters: the present rate of expansion and the present contributions to the total density of non-relativistic matter, ultrarelativistic matter and vacuum. The appearance of distant objects depends on these numbers, so in principle it is possible to determine these parameters, and also to investigate the weakest class of non-standard cosmologies—in which there may exist additional contributions to the density, with more exotic equations of state. All such possibilities can be investigated empirically within the framework of FRW models.

A deeper issue is to ask whether the basic assumption of isotropy and homogeneity is valid, and if so, why this should be. Studies of large-scale structure and anisotropies in the microwave background suggest that deviations from the Robertson–Walker metric are limited to fractional perturbations at about the 10^{-5} level, so the basic metric seems a good zero-order model. However, the existence of a particle horizon that is small at early times means that it is a major surprise to find the universe to be nearly homogeneous over regions that have only recently come into causal contact. This is one of a number of peculiarities in the standard model that cry out for explanation: like the basic fact of expansion and the near-perfect flatness, these are puzzles of the initial conditions that require explanation by a more complete theory, such as INFLATION.

Lastly, there are a number of peculiar features which relate to our status as observers. Unless $\Omega = 1$, we live near a special time—at which the contributions to the Friedmann equation from spatial curvature or vacuum energy are comparable with that of non-relativistic matter. There exists a body of ideas under the heading of the 'ANTHROPIC PRINCIPLE' which attempt to quantify the selection effects imposed by the need for intelligent observers. In some cases, these arguments are relatively uncontroversial: we should not be surprised that the universe is now roughly as old as a typical star, since stars are needed to make the heavy nuclei needed for interesting chemistry. Whether such reasoning explains all the features of the observed universe is likely to remain controversial. However, at a practical level, the standard isotropic cosmological models provide a context within which this difficult debate can at least be conducted with confidence.

Web update (31 July 2002)

A team of 27 astronomers led by Professor George Efstathiou of the University of Cambridge has published strong evidence for the existence of dark energy using the clustering pattern of 250 000 galaxies in a large volume of the universe surveyed with the Anglo–Australian Telescope at Siding Spring in New South Wales, Australia. By comparing the structure in the universe now, some 15 billion years after the Big Bang, with structure observed in the cosmic microwave background radiation, which preserved information about what the universe was like when it was only 300 000 years old, the Anglo–Australian team could apply a simple geometrical test to elucidate the composition of the universe. Their results show that the universe is full of vacuum energy, completely consistent with the earlier supernovae results.

Christopher Kochanek of the Harvard–Smithsonian Centre for Astrophysics in Cambridge, Massachusetts and Neal Dalal of the University of California, San Diego have used radio telescopes and gravitational lensing to search for cold dark matter. They have studied seven galaxies, each magnified by four nearer ones. Because each lensing galaxy is in a slightly different position, the researchers got four different images of each of the seven distant galaxies. The four images should have been

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

identical. But each is actually slightly different. The difference was enough to have been caused by the kind of clumps of dark matter around lensing galaxies that mathematical models predict.

Web Update references:

Efstathiou G *et al* February 2002 *Monthly Notices of the Royal Astronomical Society* **330**, No. 2

Dalal N and Kochanek C S (in press) 2002 Direct detection of CDM substructure. *Astrophysical Journal*

Bibliography

Kriss G A *et al* 2001 *Science* **293** 1112–1116

Carroll S, Press W H and Turner E L 1992 *Ann. Rev. Astron. Astrophys.* **30** 499

Peacock J A 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

Peebles P J E 1993 *Principles of Physical Cosmology* (Princeton, NJ: Princeton University Press)

Weinberg S 1972 *Gravitation and Cosmology* (New York: ₙiley)

John Peacock

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Cosmology: A Brief History

Cosmology is the branch of astronomy that deals with studies of the large-scale structure of the universe. Observationally it requires data on the most remote objects while theoretically it demands the largest possible extrapolations of the basic laws of physics. Despite these severe constraints, cosmology has of late emerged as a very important branch of science where predictions can be made and tested.

When did modern cosmology begin? Indeed, one should go back to Isaac NEWTON, and his correspondence with Richard Bentley from 10 December 1692, to 17 January 1693 (see Whiteside 1976). It is interesting to read Newton's attempts to construct the model of a homogeneous and isotropic but static universe and his realization that it is unstable. Later attempts within the Newtonian framework, before relativity came on the horizon, were by C Neumann and H Seeliger during 1895–1896. In 1934, W H McCrea and E A Milne demonstrated how Newtonian ideas of gravitation and dynamics can be suitably adapted to give standard models of relativity.

The advent of the GENERAL THEORY OF RELATIVITY in 1915 offered a possible resolution of the conflicts which were beginning to surface between the Newtonian laws of dynamics and gravitation and SPECIAL RELATIVITY. General relativity linked the phenomenon of gravitation to the geometry of space and time. Only 2 years after he proposed the theory, ALBERT EINSTEIN made a bold attempt to apply it to construct a model of the entire universe.

Like Newton, Einstein also found that a static model was not permitted by his 1915 equations of relativity and introduced the so-called COSMOLOGICAL CONSTANT, λ, which implied (in the Newtonian approximation) a repulsive force that varied directly with distance. The static model that emerged required the universe to be closed. Einstein felt that the emergence of such a model was a demonstration of a unique and consistent relationship between spacetime geometry and the matter contents of the universe.

However, a paper by W DE SITTER in the same year demonstrated that the model was not unique. de Sitter found a model universe which was empty but expanding. Although it was considered esoteric at the time, this model has played a key role in cosmology on a number of later occasions.

In the second decade of this century, there was no systematic study of galaxies, although, by 1914, observers of diffuse nebulae, like V M Slipher, had reported nebular shifts, mostly redshifts, that indicated a radial recessional motion of these nebulae. However, despite these findings, the general belief in a static universe was quite strong and de Sitter's solution was treated more as a curiosity. Indeed, in 1922–1924, A Friedmann and later in 1927 (independently) Abbé Lemaître obtained models of the expanding universe for which the cosmological constant was not required, but these were also ignored by Einstein and others.

Meanwhile, understanding of the universe on the observational front was also growing. In 1924 E P HUBBLE had established, through the use of Cepheid variable stars, that the Andromeda Nebula is so far away that it has to be extragalactic. Indeed in the next few years the existence of extragalactic nebulae as galaxies in their own right began to be established.

However, it was the announcement of the velocity–distance relation of these nebulae by Hubble in 1929 that turned the tide in favor of these models. For, after a careful analysis of data on nebular redshifts, Hubble arrived at what is today known as 'HUBBLE'S LAW', namely that the radial velocity of a typical galaxy away from us is proportional to its distance from us. More exactly, the data show that the redshift of a galaxy increases with its faintness. If the redshift is interpreted as Doppler shift and faintness as due to distance, then Hubble's law follows. Although there might be other interpretations of the data, all cosmological models had to take cognizance of this basic fact about the universe. Indeed it was later realized that in his 1927 paper Lemaître had predicted a linear velocity–distance relation of this kind.

Thus, soon after Hubble's law became accepted, Einstein saw that a static model was unrealistic and abandoned the cosmological constant as the 'greatest blunder' in his life. There were others, however, who thought otherwise and, even today, this constant continues to feature in cosmological literature. The reader interested in knowing who did what and when in those early days may wish to see the historical account by North (1965).

The big bang models

The assumption of homogeneity and isotropy allows the cosmologist to define a 'cosmic time'. The spatial sections at a given cosmic time are supposed to be homogeneous and isotropic. H P Robertson in 1935 and A G Walker in 1936 independently worked out the most general line element describing such a spacetime. Taking any observer as the local origin of spherical polar coordinates (r, θ, ϕ) and t for the cosmic time, the Robertson–Walker line element is given by

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - S^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right]. \quad (1)$$

The function $S(t)$ is the scale factor mentioned earlier: its increase with time signifies the expansion of the universe. The constant k in the above is a parameter specifying whether the space t = constant is of positive ($k = +1$), negative ($k = -1$) or zero ($k = 0$) curvature.

The simplest FRIEDMANN model is the Einstein–de Sitter model jointly advocated by Einstein and de Sitter (1932) which has $k = 0$. For pressure-free matter (often called dust) this model has $S \propto t^{2/3}$.

That the geometrical features of the model are linked to its physical matter contents is demonstrated by the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

different behavior of these models for different matter density ρ. Thus we define the following quantities:

$$H(t) = \frac{\dot{S}}{S} \qquad \rho_c = \frac{3H^2}{8\pi G} \tag{2}$$

as the HUBBLE CONSTANT and the critical density at epoch t. We will denote their values at the present epoch t_0 by suffix zero. The density parameter is defined by

$$\Omega = \frac{\rho}{\rho_c}. \tag{3}$$

Then for the Friedmann solutions we have the following result: the universe is closed for $\Omega_0 > 1$ and open otherwise ($k = 0, -1$). Actually the case $k = 0$ is the marginally open case with $\Omega_0 = 1$; if the density exceeds ρ_c the universe is of the closed type. This is why the density ρ_c is called the closure or critical density.

In all models the scale factor was zero at some epoch in the past, commonly called the big bang epoch. At this epoch the curvature of spacetime was infinite and so was the density of matter and radiation in the universe. What about the future behavior of the universe? There the answer depends on the geometry of space. In the open models the universe expands for ever, with the scale factor going to infinity. In the closed models the scale factor attains a maximum value before decreasing back to zero.

In the above argument it is assumed that the matter in the universe is in the form of dust. This is a reasonable approximation at present when pressures are small and matter density far exceeds the density of radiation. However, one can show that

$$\rho_{\text{matter}} \propto S^{-3} \qquad \rho_{\text{radiation}} \propto S^{-4} \tag{4}$$

so that, at a sufficiently early epoch when S was small enough, the radiation term dominated over the matter term. This of course does not alter the earlier conclusion about the existence of the big bang epoch; in fact we now conclude that the universe was infinitely hot at that epoch.

In the Robertson–Walker spacetimes, the redshift is simply related to the scale factor. Calculation shows that a source with redshift z is being observed at an epoch when the scale factor of the universe was $(1+z)^{-1}$ times its present value. Observations at the present epoch indicate that matter density is at least $\sim 10^3$ times the radiation density. Thus we can estimate that the universe was radiation dominated at epochs prior to that of redshift $\sim 10^3$.

With the realization that the basic Friedmann models give an adequate description of the expanding universe, there have been many developments in cosmology in the last five decades that are based on these models. These developments may broadly be divided into investigations of (a) large-scale structure, through observations of discrete sources, (b) early history of the universe, through observations of relics, (c) evolution of the universe from particles to galaxies, (d) basic physical laws operating in the extreme conditions a few moments after the big bang, and (e) alternative cosmologies. We will briefly outline a few historical results.

Observations of discrete sources

A relativistic cosmological model uses curved spacetime and as such there are effects of non-Euclidean geometries that may, in principle, be observable. This was the expectation which prompted optical and radio astronomers of the 1950s and 1960s to push their observing capabilities to the limit. By observing the distributions of discrete source populations (galaxies, quasars, radio sources, x-ray sources, etc) the cosmologist hoped to find which of the various theoretical models came closest to reality.

The observational tests included (i) the measurement of Hubble's constant, (ii) the extension of Hubble's law to galaxies of large redshifts, (iii) the counts of galaxies and radio sources out to larger and larger distances, (iv) the angular diameter–redshift relation and (v) the relationship of surface brightness of a galaxy to its redshift. For details of these cosmological tests see recent textbooks and review articles, e.g. Sandage (1988) and Narlikar (1993).

The trend of such studies has shifted, however, from determining the geometry of the universe to determining how the discrete sources evolve. These studies are expected to tell us about the evolution of the physical environment of the universe, but so far no clear picture has emerged amidst a series of parameter-fitting exercises.

A key measurement that continues to be controversial is that of Hubble's constant. Hubble originally obtained the value of 530 km s^{-1} Mpc^{-1}, but in retrospect we find that there were several systematic errors in his measurements. For a long time, as the value of the constant steadily came down, there was continuing controversy about its true value, which was believed to lie between 50 and 100 km s^{-1} Mpc^{-1}.

Even today there are several calibration problems. However, it is only now that a clear appreciation of various practical issues is emerging, and different approaches are beginning to converge to a value of H_0 somewhere between 55 and 70 km s^{-1} Mpc^{-1}.

Relics of the early universe

The BIG BANG THEORY hinges on the fact that, at a time $t = 0$, the universe came into existence in a singular event. Thus no physical description of the original event is possible, although physical theories can examine the subsequent behavior of the universe. One of the early attempts to go close to the big bang epoch was made in the late 1940s by GEORGE GAMOW, who appreciated the fact that the early universe was radiation dominated, that is, its contents were made up of photons and other particles which were mostly relativistic in their energies. Thus, one could approximate the equation of state by pressure $p = \frac{1}{3}\rho$, both p and ρ being dependent on temperature as its fourth power, as for radiation in thermal equilibrium. Gamow and his collaborators Ralph Alpher and Robert Herman worked out the physics of the universe when it was around 1–200 s old. (A paper in 1948 by Alpher, H Bethe and Gamow on this topic led to it being called the 'α–β–γ' theory!)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamow had hoped to demonstrate that, in the high temperatures prevailing in this era, particles such as neutrons and protons would be synthesized into heavier nuclei, thereby determining the chemical composition of the universe. In the end, this work was partly successful in that light nuclei such as deuterium, helium, etc could be made in the primordial soup, but not the heavier ones such as carbon, oxygen and metals. Later it became clear from the important work of Burbidge, Burbidge, Fowler and Hoyle in 1957 (referred to as the B^2FH theory) that these nuclei are made in stars. Nevertheless, the abundances of light nuclei worked out according to the modern version of Gamow's pioneering attempt show a broad agreement with the observed ones.

A further check on the early hot universe scenario was the discovery in 1965 of the COSMIC MICROWAVE BACKGROUND by Penzias and Wilson. Gamow, Alpher and Herman had predicted such a background as the relic of the early era, although the discoverers had been unaware of these results. In their 1948 paper Alpher and Herman had predicted a relic background with an estimated temperature of 5 K. Present big bang calculations, however, cannot estimate the temperature of the background: it has to be taken as a parameter prescribed by observations. It should be mentioned that in the early 1960s R H Dicke and his colleagues had independently arrived at the Gamow–Alpher–Herman prediction of relic radiation and were setting up a detector for the radiation when they were anticipated by Penzias and Wilson.

The most spectacular development of recent years has been the success of the COBE satellite in measuring the spectrum (in 1990) and small-scale anisotropy (in 1994) of the microwave background. The background shows a black body temperature of 2.7 K and is highly homogeneous, with temperature fluctuations $\Delta T/T \sim 6 \times 10^{-6}$.

The abundances of light nuclei and the microwave background, its spectrum and anisotropy have provided strong prima facie support for the big bang scenario. There were constraints and challenges too.

Evolution of structure in the universe

A major challenge in cosmology has been to demonstrate how, in the standard big bang model, first nucleons and leptons evolved out of more primordial particles and from them eventually the large-scale structures in the universe formed; all this in a manner consistent with the radiation fluctuations found by COBE.

Of particular interest in this work is the role of the inflationary phase first discussed independently by A Guth, K Sato and D Kazanas during 1980–1981. The basic idea is the following. The big bang universe was infinitely hot at $t = 0$, but its temperature dropped with time according to $t^{-1/2}$. In this process, the matter in it underwent a phase transition and its effect was, for a very brief period, to inflate the universe at an exponential rate, much like the old de Sitter universe. The changes in the spacetime vacuum generate a force that simulates the cosmological constant first introduced by Einstein. This is the force that 'drives' the universe so fast with an exponential growth at a time scale of around 10^{-36} s.

Most theories of structure formation rely on initial fluctuations as they evolve through INFLATION and their subsequent growth. The latter takes place through gravitational interaction and clustering. Here cognizance must be taken of the interaction of the growing lumps of inhomogeneities not only with the visible matter but also with DARK MATTER. In particular, the results are sensitive to the type of dark matter, 'cold' or 'hot' or a mixture of both.

Dark matter is the name given to matter that is not normally seen through any waveband of the electromagnetic radiation. It was FRITZ ZWICKY who in 1933, first pointed out the possible existence of the 'missing mass' in clusters of galaxies. However, it took nearly four decades for the astronomical community to catch up with him! In the 1970s, the studies of motions of clouds of neutral hydrogen showed that they were moving with near constant rotational speeds around a typical spiral galaxy, even if they were located at progressively larger distances beyond the visible mass of the galaxy. These flat rotation curves indicated that the mass $M(R)$ of a galaxy up to a distance R from its centre increases approximately in proportion to R even if R vastly exceeds the visible boundary of the galaxy.

Likewise, Zwicky's expectations about hidden mass in clusters were also borne out with the findings that the galaxies in a typical cluster were moving with such high speeds that if one uses the virial theorem for a relaxed cluster

$$2T + \Phi = \text{constant} \tag{5}$$

where T is the kinetic energy and Φ the gravitational potential energy, then one needs a lot of hidden mass to make up for the latter. What is this dark matter made of and how much of it exists in the universe? This important question has been hotly debated but remains unanswered.

The present cosmological model building has to take various constraints into account and, within the big bang framework, the need to resurrect the cosmological constant is being strongly felt. It is, however, correct to say that the model building exercise is yet to settle down to a well accepted set of parameters including the value of this constant.

Alternative cosmologies

From time to time there have been alternatives proposed to the big bang cosmology, although the majority of cosmologists have always believed in the validity of the latter. The STEADY STATE THEORY proposed in 1948 by H Bondi, T Gold and F Hoyle livened up the cosmological scenario by offering a clearly testable alternative. This cosmology had the spacetime geometry described by the model proposed by de Sitter in 1917, although the physical rationale was different. The discovery of the microwave background in 1965 robbed the theory of much of its

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

credibility. Other major initiatives in the field were the Brans–Dicke cosmology proposed by C Brans and R H Dicke in 1961, as a theory with its origins in Mach's principle, and the cosmology proposed by P A M Dirac in 1973, based on attempts to explain the very large dimensionless numbers that appear in cosmology and microphysics.

Lately, in 1993, the steady state theory has been revived in the modified form called the quasi-steady state cosmology (QSSC), by F Hoyle, G Burbidge and J V Narlikar.

Conclusions

As the observational details about the universe become more and more focused, the big bang cosmology becomes more and more constrained. For example, one long standing discrepancy has not yet been resolved: the ages of stars in some very old globular clusters are in the range 12–15 billion years, which is larger than the timespan of the standard model! This is another reason for reviving the λ term, for its inclusion can increase the age of the universe.

In the last analysis, what cosmological theory survives would depend on how the observational challenges are met. Unlike the situation at the start of this century, when there were hardly any cosmological parameters to constrain the theory, we now suffer from the embarrassment of riches. Let the fittest theory survive.

Bibliography

Narlikar J V 1993, *Introduction to Cosmology* (Cambridge: Cambridge University Press)

North J D 1965 *The Measure of the Universe* (Oxford: Oxford University Press)

Sandage A 1988 *Ann. Rev. Astron. Astrophys.* **26** 561

Whiteside T (ed) 1976 *Mathematical Papers of Isaac Newton* vol 7 (Cambridge: Cambridge University Press) pp 233–8

J V Narlikar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmos

Series of more than 2300 Russian/Soviet military, research and scientific satellites. Most operated for a few weeks or months and few details of their payloads or results were released.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cotes, Roger (1682–1716)

Mathematician, born in Burbage, Leicestershire, England. At the age of 26 he became the first Plumian Professor of Astronomy. He edited the second edition of NEWTON's *Principia*, bringing out subtleties and clarifying issues in a correspondence with Newton which, while friendly at the start, cools towards the end, presumably under the burden of over-numerous, impatiently received editorial suggestions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cowling, Thomas George (1906–90)

British mathematician, professor at Leeds. Worked on stellar structure (energy transfer by convection and radiation), developed the theory of stellar and planetary magnetic fields.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crab Nebula (M1, NGC 1952)

A supernova remnant in the constellation Taurus, position RA $05^{h}\,34.5^{m}$, dec. $+22^{\circ}\,00'$. The supernova of AD 1054 that produced the nebula reached a magnitude of -6, visible in daytime. The nebula was discovered in 1731 by John Bevis and independently in 1758 by Charles Messier, prompting him to start compiling his list of objects that might be confused with comets. It was named by Lord Rosse for its superficial resemblance to a crab. The Crab is $6'$ by $4'$ in extent and of eighth magnitude. Its outer regions consist of twisting filaments of hydrogen expelled by the supernova, appearing red on photographs and traveling outward at over 1000 km s^{-1}. The inner region glows with the pale yellow light of synchrotron radiation triggered by electrons emitted by the Crab Pulsar at the center, the core of the star that exploded as a supernova. This inner region makes the Crab Nebula the best-known example of a plerion—a supernova remnant with a 'filled' center. The Crab emits strongly in radio waves (and has the radio designation Taurus A) and x-rays (Taurus X-1).

See also: Crab Pulsar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crab Pulsar

The pulsar PSR 0531 + 21 (previously known as NP 0532), situated at the heart of the Crab nebula in Taurus. One of the first pulsars to be discovered, in 1968, and the first to be positively identified with a visible object. Known since 1948 as the powerful radio source Taurus A, the Crab nebula had been discovered in 1964 to be a powerful source of x-rays as well—the first to be optically identified beyond the solar system.

The discovery of the first pulsar (PSR 1919 + 21) by Jocelyn Bell (b. 1943) and Antony Hewish (b. 1924) at Cambridge in November 1967 was all the more remarkable because of its rapid pulse rate, with a period of only 1.337 s, and the one part in a million consistency with which this period was maintained. As further pulsars were discovered (more than 20 within a year), some were found to have periods of even less than a second. Such rapidly rotating sources implied that the objects were of very small size, and hence were of unprecedentedly high density. This led to considerable effort by physicists to devise a theoretical model which would provide a satisfactory explanation. The concept of super-dense neutron stars had been proposed as early as 1934 by Walter Baade (1893–1960) and Fritz Zwicky (1898–1974); they had also suggested that they might be associated with supernova explosions. In 1939 Zwicky suggested that the vast energy released in a supernova explosion might arise from the creation of a neutron star, and with remarkable foresight suggested that the Crab Nebula might contain such an object.

Foremost among the theorists considering the newly discovered pulsars was Thomas Gold (b. 1920) at Cornell University, New York, who published a paper in 1968 suggesting that a spinning neutron star was indeed the only possible mechanism that could explain such a rapidly pulsating object. The theory was virtually confirmed by the discovery of the Crab pulsar a few months later, with its dramatically short period of only 33.3 ms. The south-westernmost of a pair of stars visible near the center of the Crab nebula had been identified in 1942 by Zwicky and Rudolph Minkowski (1895–1976) as the most likely supernova remnant. Observations made in January 1969 at the Steward Observatory in Tucson, Arizona, confirmed that this sixteenth-magnitude star showed the same rapid fluctuations in brightness as the pulsar—the first optical identification of a radio pulsar. Only a few such optical/radio identifications are known, owing to their very low luminosity at optical wavelengths.

Subsequent observations have shown the Crab pulsar's rotation to be decelerating by 36.4 ns per day. This loss in rotational energy arises from its conversion into synchrotron radiation, which is emitted into the surrounding nebula along paths constrained by a strong dipolar magnetic field and gives rise to the lighthouse-like 'flashes' as the star rotates. Minkowski's suggestion that the neutron star is the power source of the nebula has thus been triumphantly vindicated. Measurements of the expansion of the nebula have confirmed its identification with a supernova explosion observed in AD 1054, indicating that this is one of the youngest pulsars yet discovered—hence its very rapid rotation.

As the most easily observable pulsar, and being active over much of the electromagnetic spectrum, the Crab pulsar has remained one of the most important targets of astrophysical research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crabtree, William (1610–44)

Instrument-maker and astronomer, born in Broughton, Lancashire, England, studied sunspots, made precise observations of the planets, including a transit of Venus. Crabtree collaborated with HORROCKS who predicted the transit. Horrocks described his friend Crabtree as 'a man who has few superiors in mathematical learning' and described how awestruck Crabtree had been on perceiving the little disc of Venus moving against the face of the Sun: 'rapt in contemplation he stood, motionless, scarce trusting his senses through excess of joy'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cramér–Rao Limit (or minimum-variance bound)

When it is difficult to calculate the variance (statistical error) of a parameter, the more readily calculated Cramér–Rao limit in statistics can be used as an estimate. The Cramér–Rao limit is in fact a lower bound to the variance but it is often about the same size. The reality of this assumption has to be investigated case by case, e.g. by Monte Carlo simulations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crater Constellation

(the Cup; abbrev. Crt, gen. Crateris; area 282 sq. deg.) A southern constellation which lies to the south-west of Virgo, and culminates at midnight in mid-March. It represents the cup of the god Apollo in Greek mythology (see Corvus). Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A small, rather inconspicuous constellation, the brightest stars in Crater are δ Crateris, magnitude 3.6, and γ Crateris and α Crateris (Alkes), both magnitude 4.1, which form an isosceles triangle pointing away from Virgo. γ is a binary, having a magnitude 9.5 companion, separation 5.2″. There are no bright star clusters, nebulae or galaxies in Crater.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Craters

Craters are bowl-shaped cavities which have been formed by an explosion. On planetary bodies, the few examples of past or present volcanism (Mars, Venus, Io) are not of the explosive type. Therefore, nearly all craters on PLANETARY SURFACES result from the impact of a METEOROID at very high speeds (tens of km s^{-1}). PLANETARY EXPLORATION has demonstrated that impact craters are ubiquitous features of the geology of solid solar system bodies. For most objects, internally driven evolution (tectonics, volcanism) has stopped after a few million to a few hundred million years. Impact cratering at all scales has therefore been a major factor for the evolution of planetary surfaces.

First, a family portrait of craters throughout the SOLAR SYSTEM will be drawn, with their striking similarities and intriguing differences. We will present the early controversies on lunar craters, the only such planetary features which can be seen with the naked eye, and the evidence from which lunar craters as well as nearly all other craters in the solar system are now recognized as impact features instead of volcanic craters.

In a further section, the origin of impacting bodies will be discussed. The evolution of the impact rate with time is closely related to the formation process of the solar system (see SOLAR SYSTEM: FORMATION). Our understanding of collision processes in the solar system has recently much improved with the advent of chaotic dynamical models.

We will then discuss the physics of impact cratering. The formation of craters is ruled by relatively simple scaling laws over an extremely wide range of impacting masses. The evolution of large craters is strongly influenced by gravity. At the higher end of the size range, the relaxation of the planetary crust modifies the crater shape. Understanding these crater formation and evolution processes is essential when attempting to derive the ages of solar system surfaces from the cratering record.

Finally, the consequences of crater formation at all scales on the evolution of a planetary surface will be examined. They range from the formation of major impact basins several thousand kilometers in diameter to thorough modifications of the mineralogy and texture of surface soils (or regoliths). Perhaps the most dramatic cratering event of all may have been the giant impact of a Mars-sized body on the Earth which is now considered to be at the origin of the formation of the Moon.

Craters in the solar system

The most direct evidence that high-velocity objects impact planetary bodies is the observation of a shooting star, which shines brightly due to its very high temperature upon atmospheric entry. The typical mass of these particles is a few milligrams. In the nineteenth century, METEORITES with masses of a few hundred grams to several tonnes were identified as extraterrestrial objects. The other line of evidence can be traced back to the observation by GALILEO of round features at all scales on the lunar surface, the lunar craters and basins (see MOON). Up to the middle of the twentieth century the origin of these features was controversial, as they were attributed either to volcanic activity or to impact processes. The matter was finally laid to rest with the Apollo program. In the mean time, the best preserved terrestrial impact crater had been discovered in northern Arizona (of course, it was known for centuries by the Navajos): METEOR CRATER, with its diameter of 2 km and its depth of more than 300 m, is a very young feature formed 40 000 years ago (figure 1). It shows a striking similarity with bomb craters, with its bowl shape and slightly raised rim. Indeed, impact and explosion cratering are closely related, and a lot has been learned about the physics of cratering from the near-surface nuclear tests of the 1950s and early 1960s.

Figure 1. Meteor Crater, near Flagstaff in northern Arizona, is the youngest and best preserved large impact crater on the Earth, with a size of 2 km. It was formed by the impact of an iron meteor a few tens of meters in diameter. The similarity with craters generated by explosions is striking, with a bowl shape and raised rim.

The extraordinary adventure of planetary exploration during the last 40 years has demonstrated the ubiquitous character of impact cratering in the solar system. Indeed, the first impression given by the first Mars fly-bys was that it was, somewhat disappointingly, just a larger version of the Moon. These observations happened to cover the older, southern hemisphere of the planet which is indeed dominated by impact craters. When the Mariner 9 orbiter finally obtained a global coverage, it revealed the tremendous volcanoes and canyons which characterize the Tharsis region. The large cavities at the top of these volcanoes have been formed by collapse, not by an explosion, and are therefore calderas, not craters. The PIONEER and VOYAGER MISSIONS to the systems of outer planets have observed enormous impact features, such as the Valhalla basin on GANYMEDE (figure 2). At the other end of the solar system, MERCURY was revealed as a near twin to our Moon, its surface being pocked with craters of all sizes. Only 50% of the surface of Mercury has been imaged, so that surprises cannot be excluded once the other hemisphere is observed. More recently, close-range observations of ASTEROIDS showed craters with sizes up to 30% of the diameter of the body (figure 3).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

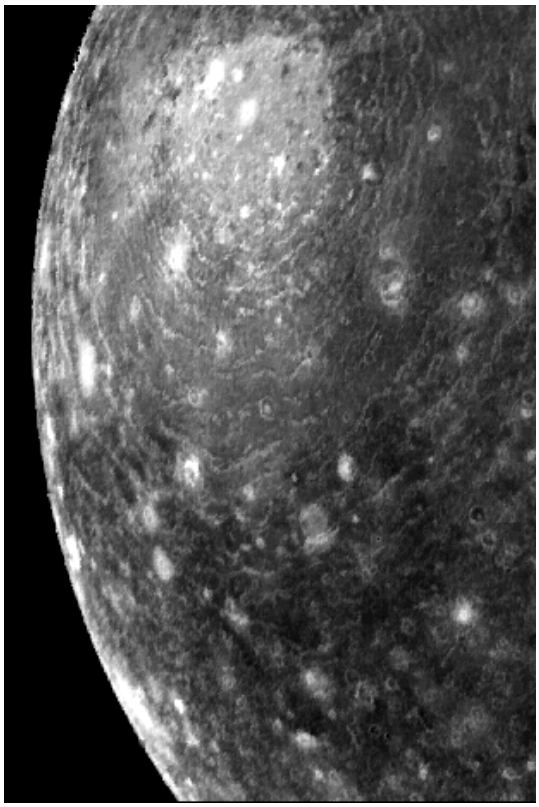


Figure 2. The Valhalla basin is an impact structure nearly 2000 km in diameter on Callisto, the outer large satellite of Jupiter. The small vertical extent and concentric rings reveal the importance of relaxation processes in the icy lithosphere. This structure is pockmarked by later impacts which saturate the surface.

Figure 3. This image of Ida, a 40 km sized asteroid in the main belt, was obtained by the Galileo spacecraft on its way to Jupiter. The global shape of the asteroid is indented by major impact craters, while smaller craters are ubiquitous on the surface.

The wealth of observations from planetary missions has also shown that several planetary bodies exhibit much smaller densities of impact craters. This is the case for Venus, Io, Europa and icy satellites such as Enceladus (satellite of Saturn) or Miranda (satellite of Uranus). The example of the Earth suggests that this situation arises when resurfacing processes have erased all but the latest impact craters, i.e. when the surface is younger than a few hundred million years. The cratering record is therefore clearly linked with the age of the surface, which leads to a discussion of the origin of impacts in the solar system.

The history and characteristics of impacts in the solar system

Much of what we know about the time evolution of cratering comes from the study of lunar samples gathered by the Apollo and Luna programs in the early 1970s. The potassium–argon radioisotopic chronometer is reset by shock events, as the daughter product, argon, is a noble gas which is expelled under high pressures. These analyses provided the first absolute ages of major impact events on the Moon, such as that which formed the IMBRIUM BASIN, 3.8 billion years ago. The age of lunar mare, which provided clean plates for registering impacts from 4 to 3 billion years ago could be determined from the rock formation ages of lunar basalts (uranium–lead and rubidium–strontium chronometers). Since 1970, the advent of chaotic dynamics has much improved our understanding of the evolution of protoplanetary swarms during the formation of the solar system. These theories also provided a satisfactory model of the transfer of asteroids or asteroidal fragments to the inner solar system by planetary perturbations. From these various sources of information, there is now a wide consensus on the following scenario:

(a) The condensation of the solar nebula resulted in the formation of a very large number of small bodies, the planetesimals in the inner solar system and the cometesimals in the outer solar system.
(b) The planetary embryos first grew by direct collisions, which resulted in a few million years of extremely heavy bombardment.
(c) Once the direct feeding zones were cleaned up, gravitational perturbations induced further collisions or near encounters between the surviving objects. The remaining primordial population of small bodies was then progressively cleaned up by collisions with planets or ejection in the OORT CLOUD. This phase, the accretionary tail, lasted more than 500 million years according to the ages of lunar basins. It left three major reservoirs: the Oort cloud, corresponding to cometesimals ejected at large distances by the giant planets, the main belt of asteroids, where the accretion of a planet was prevented by the proximity of the proto Jupiter, and the KUIPER BELT of cometesimals, where the density was too low to form a planet.
(d) During the remaining 4 billion years, collisions in the main belt, planetary and stellar perturbations have resulted in a steady state flux of secondary objects, represented nowadays by near-Earth asteroids (from the main belt), short-period comets (from the Kuiper belt) and long-period comets (from the Oort cloud). The relative contribution of comets and asteroids to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the impact rate is highly uncertain. It is expected to depend on the distance to the Sun, with comets representing up to one-third of the impacts on the Earth and Moon. On the Moon, this constant impact rate corresponds to the formation of a kilometer sized crater every 30 000 years. On the Earth, it is doubled as impactors are focused by the strong gravitational field, but this also makes the craters smaller. The 2 km wide Meteor Crater, formed 40 000 years ago, is therefore quite typical taking into account the large surface fraction covered by oceans.

The impact velocities depend strongly on the distance to the Sun and on the source region. As a simple rule, the relative velocity of an impactor with a planetary body is a major fraction of its orbital velocity, long-period comets having the largest impact velocity. For the Earth–Moon system, with its orbital velocity of 29.8 km s^{-1}, the typical impact velocity is 20 km s^{-1}. For Mercury (50 km s^{-1} orbital velocity) it is likely to be more than 30 km s^{-1}. In the main belt, a major source region, the relative velocities are less than 10 km s^{-1} for asteroidal fragments and more than 10 km s^{-1} for comets, which have a much larger eccentricity. The planetocentric velocity of satellites (1 km s^{-1} for the Moon) only becomes a factor for the inner satellites of Jupiter (Io: 17 km s^{-1}).

The mass distribution of impacting bodies is not modified by gravitational perturbations. It is therefore expected to be very similar for objects from the same source region throughout the solar system. As a simple approximation, the flux of objects above a given mass is inversely proportional to the mass, with a steeper dependence at small masses and a flatter dependence at large masses. The dependence of cometary and asteroidal contributions with heliocentric distance is still uncertain.

The physics of impact cratering

The impact cratering process is closely related to the formation of craters by explosions. This relationship can be inferred from the comparison between impact velocities (5 to 30 km s^{-1}) and the highest detonation speed of chemical explosives (7 km s^{-1}). Therefore, an impactor delivers an energy per kilogram which is in the range of those dissipated by high-yield explosives. All non-grazing impacts are extremely inelastic, forming a bubble of gas with an initial temperature of several 10 000 K. At these very high velocities, it is the expansion of this bubble of gas which results in the formation of the crater, once again quite similarly to a chemical or nuclear explosion.

As the shock wave expands, it incorporates an increasing mass of material, which slows down its velocity. The initial expansion velocity is related to the impact velocity, hence is higher than the speed of sound. During this hypersonic phase, the shock wave is at the outer edge of the expanding cavity. Once the expansion has slowed below the speed of sound, the compression wave expands faster than the edge of the cavity. These P-waves eventually propagate far from the impacts. Impact events are therefore a source of low-magnitude seismic activity.

With a given impact velocity, the distance at which the specific energy per kilogram drops below a given value is proportional to the radius of the impactor. As an example, at 10 radii from the impact, the cavity corresponds to 500 times the impacting volume, hence a mean distributed energy 1/500 the impact energy, irrespective of the scale. Assuming a simple equipartition between kinetic and thermal energies, the expanding and decelerating shock wave defines four major zones:

(a) Very close to the impact, the energy per kilogram is higher than the vaporization energy of the material. The mass of gas formed is similar to the mass of the impactor at velocities of 20 km s^{-1}.
(b) Farther out, the energy per kilogram is higher than the melt energy. The mass of liquid is typically 10 to 20 times the mass of the impactor at 20 km s^{-1}.
(c) At a distance from the impact which is proportional to the size of the impactor and depends on the specific cohesive energy of the material, the shock wave becomes unable to fragment the material. This distance defines the radius R_S of the transient cavity, the so-called 'strength crater', within which the material is set loose from the substrate.
(d) Outside the transient cavity, the propagating shock wave may generate non-destructive fractures and shock pressure effects.

At any stage of the expansion, the velocity of the ejecta is expected to be similar to that of the shock wave. The slowest ejecta, coming out from the edge of the transient cavity, are therefore expected to leave at a velocity V_S which is proportional to the square root of the cohesive energy of the material. This simple scaling law for ejecta velocities has been experimentally tested for materials as different as basalt and pumice sand. The parabolic trajectory of ejecta is then determined by the surface gravity of the impacted body. For the slowest ejecta, the travelled distance is V_S^2/g, which defines the gravity radius R_g. If the strength radius R_S is smaller than R_g, most ejecta end up far from the crater. Its final morphology is then that of the transient cavity. For such 'strength craters', the observed depth to diameter ratio is about 0.4, close to the 0.5 value which would correspond to our simplistic model of a spherically expanding shock wave (the difference results from the enhanced disruption of material near the free surface).

The final shape of larger craters is dominated by gravity. When R_S is larger than R_g, ejecta are pulled back into the raised rims typical of large lunar craters and partially fill the impact cavity. Most of the energy is now expended as gravitational potential energy, so that the size of gravity craters does not depend on the cohesive strength of the material. The transition occurs at very small scales for non-cohesive material in a strong gravity field. As an example, R_g is only 1 m for the lunar regolith. For a rocky substrate, or for low-gravity bodies such as asteroids, this transition is on a kilometer scale. As material falls back into the transient cavity, it is not surprising that gravity craters are much shallower than strength craters, with depth to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

diameter ratios as low as one to five. Furthermore, lateral transport of the loose rim material tends to flow back into the crater, so that gravity craters gradually fade out with a lifetime proportional to the square of their size.

For the largest craters (more than 20 km in diameter on the Moon), the lithosphere cannot support the initial crater topography against gravitational forces. Partial collapses lead to the formation of complex craters, characterized by faults, slumped walls and terraces. Before these long-term relaxation processes, a rebound in the lithosphere forms a central peak. This is similar to the impact of a rain drop in a puddle, after which a small droplet is shot back. At scales of more than a few hundred kilometers, impact craters are named basins, around which the propagating shock wave generates concentric rings. The readjustments of the crust explain the relatively small vertical extent of very large impact structures (up to 10 km for lunar basins, even less for basins on icy satellites). The northern regions of MARS represent a very specific case, as there are large amounts of subsurface ice at increasing depths towards the south. The fluidized ejecta from craters in the right size range generate 'splash craters' characterized by a sharp outer boundary of the lobate ejecta blanket.

At the upper end of the size range, major cratering events can occur at a scale similar to the size of the impacted body itself. The propagation of the shock wave can no longer be considered in the infinite planar surface approximation, but must be analysed in three dimensions, taking into account possible reflections at the surface. Shock wave focusing may be responsible for the irregular terrains observed in regions antipodal to major impacts on several solar system bodies, such as Mercury. Experimental simulations have been done at a scale of less than 10 cm which showed that the maximum crater size beyond which complete fragmentation occurs is about 40% of the size of the impacted body. The cratering record in the solar system is a remarkable demonstration of the validity of scaling laws for impact cratering: subcritical craters have indeed been observed for a wide range of solar system bodies, in particular Phobos (25 km in size), the largest of the two satellites of Mars (see PHOBOS AND DEIMOS), with its large crater Stickney (9 km in diameter), but also asteroids such as Ida and Mathilde and icy satellites such as Mimas. In the case of Phobos, linear grooves originating from Stickney and propagating over the whole surface suggest that global fracturation has occurred, stopping just short of fragmentation. Models show that in the main belt of asteroids, all objects smaller than about 200 km have a lifetime shorter than the age of the solar system. Fragmentation does not necessarily means destruction: if the body is large and/or weak, its self-gravitational energy is much larger than its cohesive energy. In such a case, the fragments will reassemble in 'rubble piles' with an equilibrium ellipsoidal shape if the body is rotating. These objects are expected to represent a large fraction of asteroids of intermediate sizes. The most dramatic cratering event in the inner solar system was probably the giant impact of a Mars-sized body with the Earth which is now widely considered to be responsible for the formation of the Moon.

The simple proportional scaling between the size of the impacting body and the size of the crater in the strength regime is very well substantiated by experiments. The dependence on the impacting velocity is much more complex. The analogy with explosion craters led to the hypothesis that the volume of the crater is simply proportional to the energy of the impact, hence the square of the impact velocity. This assumes correctly that the formation of the gas bubble can be considered as a point source when compared with the size of the crater. This 'energy scaling' assumption seemed consistent with the extreme inelastic character of the crater formation process. Experimental results in weakly cohesive material suggested a dependence on momentum, not energy, hence a volume which is proportional to the impact velocity. Recent results all lie in the range between these two models. The model dependent variation with velocity is a major problem when attempting to adapt the well calibrated lunar production curve to determine absolute crater production rates on other solar system bodies: as we have seen, impact velocities can be substantially higher (Mercury) or lower (asteroids, icy satellites) than on the Moon.

Once formed on the surface of a solar system body, craters are not eternal. Geological processes and erosion can erase the cratering record. On the Earth, most of the surface is renewed on time scales of millions of years. On Mars, there is a striking difference in terms of cratering record between the old southern hemisphere and the northern hemisphere, which remained geologically active for more than 3 billion years. Even on inactive bodies devoid of atmosphere, such as the Moon or Mercury, two major external processes are at work: impact obliteration and impact erosion. Obliteration occurs when the formation of a new crater wipes out all or part of a pre-existing crater. The lifetime against obliteration increases for large craters, which survive all but the biggest local impacts. This process was dominant during the accretionary tail, when the flux of impactors was very large. Impact erosion results from the progressive modification of the shape of the crater by small impacts, which are very numerous (typically 1000 times more frequent for a size which is 10 times smaller than the original impactor). Erosion processes are strongly dependent on gravity: in low-gravity environments, such as small satellites or asteroids, ejecta from small impacts will be removed from the vicinity of the crater, leading to a simple 'sand-blasting' erosion process. In a significant gravity field, such ejecta travel preferentially downslope, at a rate which is proportional to the slope angle and inversely proportional to the gravity. Highs (the rim) and lows (the bottom of the crater) tend to relax towards the horizontal quite similarly to the dissipation of a heat pulse, hence the lifetime is proportional to the square of the crater diameter. These old craters with subdued topography are visible close to the terminator, where they saturate the surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

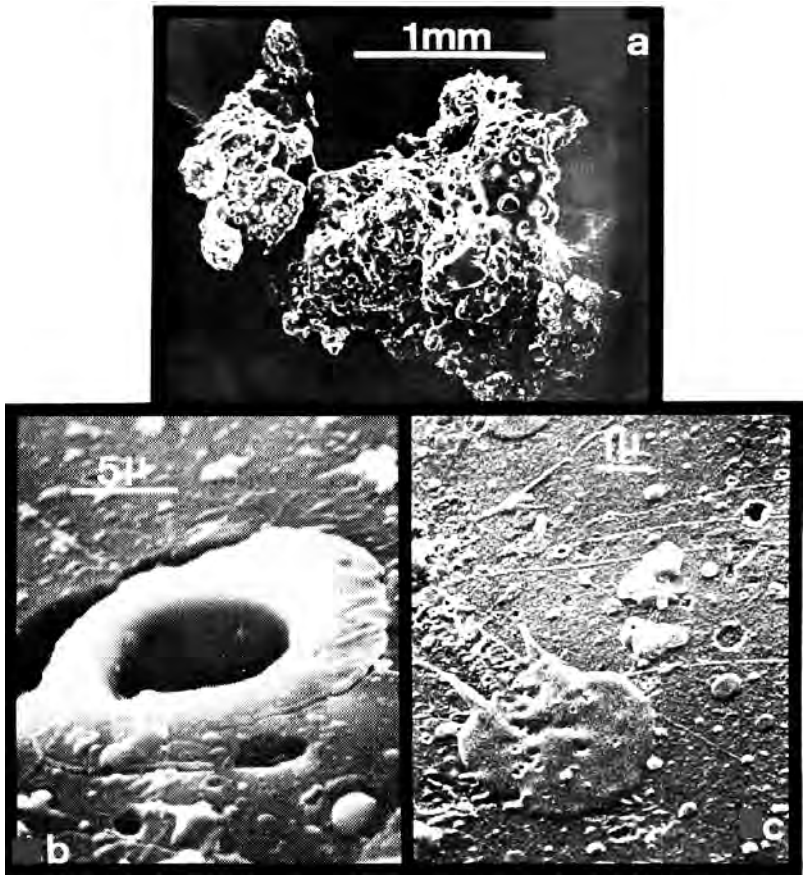


Figure 4. Three examples of small-scale impact processes in the lunar regolith: a millimeter-sized glassy agglutinate (*a*), a microcrater a few micrometers in diameter (*b*) and a glassy splash on a lunar dust grain (*c*).

Craters as a chronometer of the surface of solar system bodies

As small craters have shorter lifetimes, the crater size frequency distribution of a surface presents two domains: the saturation region, at small sizes, where craters are in a steady state, and the production region, characterized by a steeper dependence on size, corresponding to the large craters which have in general survived since their formation. Only the latter is of interest when attempting to use the cratering record for determining the age of a surface. The size distribution in the production region is roughly similar for surfaces throughout the solar system, from Mercury to the satellites of Jupiter (see JUPITER: SATELLITES). This is considered to be a consequence of the scattering of impactors across the solar system by gravitational perturbations, which do not modify the mass distribution of a population.

There is no controversy if one sticks to relative ages: on a given body, a region with few large craters is younger than another region with a higher density of such features. If the age is younger than 4 billion years, in the steady-state domain of secondary impacting bodies, it is expected to be proportional to the density of large craters. Beyond that, the steep increase in flux means that age differences of a few 100 million years may result in very different crater densities. As the overall rate is different for each region of the solar system, the determination of absolute ages from the cratering record is to a large extent model dependent. As an example, the ages derived for old Mars terrains can differ by more than a billion years, and a similar controversy is building up for EUROPA, a satellite of Jupiter. Its highly spherical surface could be resurfaced over short time scales according to the proponents of a global ocean a few 10 km below the surface.

The Moon represents up to now a unique situation: the samples obtained by the APOLLO and LUNA missions provide absolute ages for nine regions, two in highlands and seven in mare, which span more than 1.2 billion years. Thus, lunar samples provide a calibration for cratering ages to within a few 10 million years. The controversies on the age of surfaces for other planets will be resolved as soon as even a few samples are available. A sample return from Mars can be expected within the next 10 years. Even a single absolute age determination will strongly constrain the ages of all surface units on the planet by calibrating the crater production curve. Alternately, one could consider

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 5. Copernicus is one of the most spectacular ray craters on the lunar surface and has a diametre of 110 km. As it lies near the center of the nearside, it was one of the first craters to be studied extensively from the ground. Its rays extend over more than 1000 km. A smaller ray crater, Kepler, lies to the west of Copernicus.

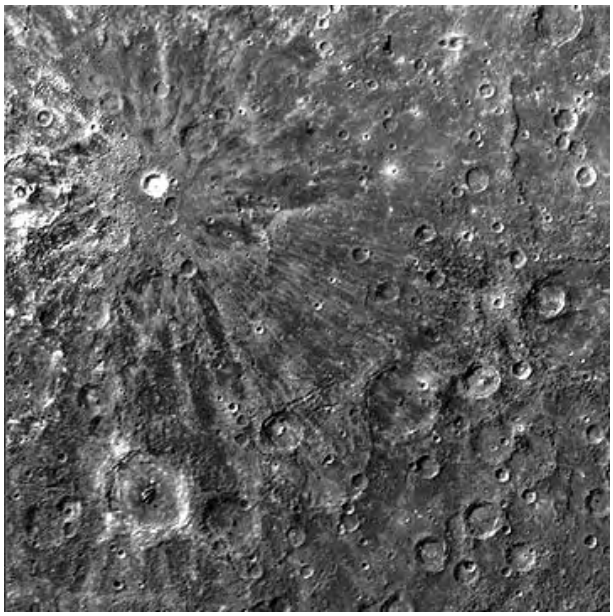


Figure 6. Copley is a 30 km sized ray crater on Mercury. The relatively weak contrast between the rays and the underlying material when compared to Copernicus may be attributed to the efficient maturity processes on the closest planet to the Sun, where impact velocities and particle fluxes are highest.

in situ determinations of radioisotopic ages, but this still represents a formidable challenge

The large-scale effects of impact cratering

Impact cratering is a major process for the evolution of the surfaces of solar system bodies. Its role can be discussed as a function of the age of the surface. Few bodies have retained a high level of internal activity until now. Apart from our Earth, only VENUS and Mars have had significant magmatic episodes during the last two billion years. Tidal heating still drives extensive internal processses for the two innermost Galilean satellites, IO and Europa. Similar processes have occurred on DIONE and ENCELADUS, two satellites of Saturn, as well as MIRANDA, a satellite of Uranus. On Ganymede, phase changes of the ice play a major role, and the surface of cometary nuclei is ablated by the sublimation of volatiles. Erosion can be a major factor for solid planetary bodies with an atmosphere (Earth, Mars, Venus and TITAN). For surfaces younger than a few hundred million years, cratering is a rare, catastrophic event which plays a minor role for the local topography.

At the other end of the scale, the surfaces of asteroids, small icy satellites, CALLISTO and the highlands of Mercury and the Moon are more than 4 billion years old. They have witnessed the high-flux period of the accretionary tail. Their trademark is a saturation with craters of all sizes. These very high fluxes have fractured the crust over several tens of kilometers. This fractured layer is called a megaregolith. For medium size asteroids, the megaregolith can extend to the center, resulting in a rubble pile structure.

The fractures induced by the largest basins extend to the base of the planetary crust. A few hundred million years later, the heat from the decay of long-lived radionuclides triggers magmatic episodes in the mantle. The lava rises to the surface in the fractured regions and floods preferentially the large basins. This process is at the origin of mare on the Moon and Mercury. Recent advances in planetary astronomy (Hubble Space Telescope, adaptive optics) have made possible the identification of such basaltic regions on VESTA. The same cycle has therefore occurred on differentiated asteroids, but on a much shorter time scale, given the much smaller size of these bodies. The heat source is still under debate. On the Earth, plate tectonics has completely erased the traces of early major impacts. Large-scale crustal weakening due to impacts may, however, have played a role in the very early stages of plate tectonics. Mars represents an interesting intermediate case: the southern hemisphere is lunar like, with the two large basins Hellas and Argyre, while the northern hemisphere (with a extension to the south over the Tharsis region) is dominated by tectonic and volcanic processes (see MARS: SURFACE).

During the last 4 billion years, the impact rate was much lower, and the scale of cratering effects correspondingly smaller. Here and there, a few major

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

impacts still formed large craters (Copernicus and Tycho are the largest craters of this type on the Moon). The steady shower of small impacts resulted in the formation of a debris layers a few meters to a few tens of meters in thickness, the REGOLITH. Regoliths are ubiquitous features of the surfaces of atmosphereless bodies throughout the solar system. Remote sensing techniques which probe the uppermost layers do not give direct information on the bedrock, but on the overlying layer of dust and rocky fragments. Fortunately, in such a situation lateral transport can only be triggered by impacts, hence it has a very inefficient random walk behavior. The lateral mixing scale is a few hundred meters on large bodies such as Mercury, the Moon or large icy satellites. It is inversely proportional to the gravity, which explains the relative homogeneity of the surfaces of small asteroids observed by Galileo (Gaspra and Ida) and Near (Mathilde).

The small-scale effects of impact cratering

The remarkable self-similarity of impact cratering extends down to microscopic sizes. A tiny impact on a single rock or grain on a planetary surface creates a microcrater (figure 4(*b*)) which follows the same scaling rules as macroscopic strength craters in the bedrock. Above a diameter in the range of 40% of the size of the rock or grain, fragmentation occurs. As a result, the size of regolith particles slowly decreases from the coarse fresh EJECTA to a very fine dust (typically 50 μm in size). This process should be quite similar on rocky and icy bodies, as ice has a mechanical behavior similar to a rock at the very low temperatures of the outer solar system.

On rocky bodies, two impact related processes tend to increase particle size. Coating by liquid splashes (figure 4(*c*)) or vapor deposits plays a relatively minor role. The liquid fraction from centimeter-scale impacts can weld together grains, forming glassy agglutinates (figure 4(*a*)), or solidify in flight, forming glass spherules. The proportion of glassy particles increases as a function of the time of exposure to space of a sample, which is defined by the local history of meter-sized craters. The study of lunar samples shows that on the Moon, after a few ten million years within the uppermost centimeter, a steady state is reached in the evolution of the size distribution: agglutination exactly compensates fragmentation while the proportion of glassy particles continues to increase, reaching more than 50% in the most exposed lunar soils.

This evolution of a soil sample from the coarse, crystalline fresh EJECTA to a fine-grained, glass-rich soil is called maturity. It also involves exposure to solar wind particles and cosmic rays. Its most dramatic effect is the lowering of the albedo by more than 30%, as glassy particles are darker than the original crystalline grains. The fresh ejecta form bright rays extending radially from the younger large craters (figures 5 and 6). With the increase of maturity, these rays progressively fade into the background. Maturity is a balance between large-scale cratering, which controls the rate at which fresh material is brought onto the surface, and sub-centimeter-sized impacts which control the formation rate of glassy particles. It is difficult to extrapolate the well known lunar situation to other rocky bodies. However, on Mercury the impact velocities are typically twice as large, generating more liquid material, and the regolith should be more mature than on the Moon. The situation on icy bodies is more difficult to assess. One may expect liquid water droplets formed during impacts to play a similar role in welding together ice grains into 'icy agglutinates' and solidifying in flight, forming ice spherules, but the time evolution of icy regoliths is still highly spectulative.

Yves Langevin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crescent

The phase of a body in the solar system when some but less than half of its sunlit side is visible. The only objects to show crescent phases to observers on the Earth are those that can pass between the Earth and the Sun: the inferior planets, Mercury and Venus, and the Moon. However, objects outside the Earth's orbit have been imaged in the crescent phase by spacecraft.

See also: phases of the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crimean Astrophysical Observatory

The Crimean Astrophysical Observatory (CrAO) is one of the largest scientific centers in the Ukraine and former Soviet Union. CrAO was formed in June 1945 by the Academy of Sciences of the USSR from the Simeiz Department of the Pulkovo Observatory. Since 1991 CrAO has been controlled and financed by the Ministry of Science and Technology of Ukraine.

The main part of the Observatory is located in Nauchny, about 12 km SE of Bakhchisaraj. The main telescopes are: the 2.6 m Shajn telescope, two 1.25 m telescopes; a ground-based gamma telescope (GT–48) with the mirrors covering an area of 54 m^2, the 1.2 m Solar Tower telescope; as well as a few other smaller instruments.

The Department of Radioastronomy, with its 22 m radio telescope (RT–22) for mm and cm radio waves, is located on the seashore, near Simeiz. RT–22 participates in international programs in the global VLBI network for geodynamical and astrophysical investigations.

The Simeiz Station, founded in 1873, participates in the global network of Satellite Laser Ranging Stations for studies of Earth dynamics, and is located on Mount Koshka at an altitude of 346 m.

CrAO possesses equipment for astrophysical observations over a wide spectral range from gamma rays to radio waves of stars and galaxies, as well as the Sun and solar system.

The main subjects of investigation are: solar activity, helioseismology, structure and chemical composition of stellar atmospheres, stellar and solar magnetic fields, variable stars, extragalactic astronomy, planets, asteroids and geodynamics. Currently, next-generation ground-based and space telescopes as well as astrophysical equipment are being designed and manufactured at CrAO.

More than ten devices have been designed and manufactured by CrAO for space observation. In 1983–89 space observations were carried out with the station 'Astron' launched into a high orbit and equipped with a 0.8 m ultraviolet Ritchey–Cretien telescope, designed and manufactured at the CrAO.

There are 350 employees at the CrAO including about 100 scientists, 16 senior and 40 PhDs.

For further information see
http://www.crao.crimea.ua.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Critical Density

The mean density of a universe which is just capable of expanding forever. In such a universe (an Einstein–de Sitter universe) the rate of expansion is slowed by gravity and approaches ever closer to zero. If the mean density exceeds the critical value, the expansion will cease at a finite time in the future and the universe thereafter will begin to collapse. If the mean density is less than the critical value, the universe will continue to expand at a rate which decreases towards a constant value.

The magnitude of the critical density is given by $3H_0^2/8\pi G$, where H_0 is the present value of the Hubble constant and G the gravitational constant. Depending on the precise value of the Hubble constant, the value of the critical density is in the region of (5–20) $\times 10^{-27}$ kg m^{-3}, with a probable value in the range (6–10) $\times 10^{-27}$ kg m^{-3}. The latter figure is equivalent to an average of about 3–6 hydrogen atoms per cubic meter of space. Many theoreticians believe that the universe should have a density equal to, or indistinguishably close to, the critical value, but the average density deduced from the amount of visible (luminous) matter is only of the order of 1% of the critical density. Observations of the rotation rates of galaxies and the dynamics of galaxy clusters indicate that there is at least ten times as much dark matter as luminous matter in the universe, but there are growing doubts as to whether there is enough dark matter to give a mean density as high as, or higher than, the critical value.

The ratio of the actual mean density to the critical density (the density parameter) is denoted by the symbol Ω (omega). $\Omega > 1$ if the actual density exceeds the critical value, $\Omega = 1$ if the actual density is equal to the critical density and $\Omega < 1$ if the actual density is less than the critical value.

See also: cosmological constant, cosmological model, cosmology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crust

The thin outer layer of a differentiated planetary body. Crusts consist of rock and/or ice—rock predominating in the inner solar system, ice in the outer—and have a different composition from the underlying mantle. *Primary crust* is the surface layer formed when the planet first differentiates. It preserves the cratering record from the early bombardment in the solar system's history. Primary crust is found for example in the highlands of Mercury and the Moon, and many of the satellites in the outer solar system. *Secondary crust* is formed when the mantle melts and volcanism leads to the first phase of resurfacing; the floodplains of the lunar maria are secondary crust, as is the whole surface of Venus. *Tertiary crust* is produced on geologically active worlds by the recycling of primary and secondary crust material. On the Earth and Io, continuing volcanism produces tertiary crust; on the Earth plate tectonics is continually recycling the crust, which is also modified by processes such as sedimentation and erosion.

Neutron stars are believed to have a crust consisting predominantly of iron.

See also: differentiation, neutron stars, plate tectonics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Crux

(the (Southern) Cross; abbrev. Cru, gen. Crucis; area 68 sq. deg.) A southern constellation which lies to the north of Musca and is surrounded on three sides by Centaurus. It culminates at midnight in late March. The brightest stars of Crux were included by the ancient Greeks in Centaurus and were cataloged as such by Ptolemy (*c*. AD 100–175) in the *Almagest*, but European navigators separated them out in the sixteenth century and they were included as a constellation in their own right on a celestial globe by the Dutch theologian and geographer Petrus Plancius in 1598.

Crux is the smallest but also one of the most conspicuous of the 88 constellations and (together with Orion and the asterism of the Plough in Ursa Major) one of the best known. Its brightest star, α Crucis (Acrux) is a triple system consisting of two bluish-white (B0.5 and B1) components, magnitudes 1.3 and 1.6, separation 4.1″, and a third, unseen component which revolves around the former in 75.8 days. Other bright stars include β Crucis (Becrux or Mimosa), magnitude 1.3, γ Crucis (Gacrux), magnitude 1.6, and δ Crucis, magnitude 2.8. A line from γ through α (along the longer axis of the cross) points roughly towards the south celestial pole, while a line from δ through β (along the shorter axis) points roughly towards β Centauri. There is one other star brighter than fourth magnitude, ε Crucis, magnitude 3.6, which lies between and slightly offset from the line joining γ and α, and rather spoils the symmetry of the cross—which, if anything, more closely resembles a kite.

Interesting objects in Crux include μ Crucis, a wide binary star system with bluish-white (B2 and B5) components, magnitudes 4.0 and 5.1, separation, 35″, the Jewel Box (NGC 4755), a magnificent open star cluster centered on κ Crucis, and the Coalsack, an area of dark nebulosity silhouetted against the Milky Way which runs through the constellation.

See also: Coalsack.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cubewano

Any of the main stream of Kuiper Belt objects, at 41–47 AU mean distance from the Sun, inclination of orbit ranging up to 30°, eccentricity up to 0.1. Named after the designation, 1992 QB1, of the first Kuiper Belt object, which was discovered by Dave Jewitt (University of Hawaii).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Culmination

The passage of a celestial body across an observer's meridian (also known as 'transit'). Upper culmination (upper transit) is the passage of the body through the point at which its altitude above the horizon is greatest (and zenith distance is least). For example, the Sun reaches upper culmination at noon. Lower culmination (lower transit) is the crossing further from the zenith. If the body is circumpolar, the point of lower culmination will be above the horizon; otherwise, lower culmination occurs when the body is at its maximum angular distance below the observer's horizon.

See also: altitude, horizon, diurnal motion, meridian, transit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Culpeper, Nicholas (1616–54)

Astrologer and herbalist, born in London. His tradition continues to the present day, with lotions and remedies on sale in his name.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cunitz [Cunitia], Maria (1610–64)

Astronomer, translated the works of KEPLER, making his work accessible. Her abilities in astronomy were so outstanding that she was called Urania Propitia, i.e. 'she who is closest to the muse of astronomy'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Curtis, Heber Doust (1872–1942)

Astronomer, born in Muskegon, MI, became director of the Allegheny Observatory of the University of Pittsburgh and professor of astronomy and director of the Astronomical Observatories of the University of Michigan. Surveyed nebulae with the Lick Observatory Crossley reflecting telescope, identified that some long thin-looking nebulae were identical with the spiral nebulae, but seen edge on with a band of obscuring matter like that seen centrally in the Milky Way. Interpreted the Zone of Avoidance, in which few or no spiral nebulae were seen near the Milky Way, as being due to the obscuration by the Milky Way's obscuring matter, and concluded that spiral nebulae were Milky Way galaxies outside our Galaxy. Engaged in the 'great debate' with HARLOW SHAPLEY in Washington 1920, arguing correctly against Shapley's view that spiral nebulae were minor objects within an immensely bigger Milky Way structure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cusp Cap

A brightening at one or other of the tips—*cusps*—of the crescent phase of Venus, as seen from Earth. Cusp caps were first reported by the German amateur astronomer Baron Franz Paula von Gruithuisen in 1813, and have been recorded by telescopic observers ever since. They were named by analogy with the Earth's polar caps; early observers fancied they were seeing glimpses of a possibly Earth-like surface through breaks in the Venusian cloud cover. Cusp caps take the form of bright patches at the horns of the crescent. On the side of a cusp cap away from the cusp itself there is often a darker band known as a *cusp collar*. Both caps and collars show short- and long-term variations in brightness and size.

It was once thought that the cusp caps were contrast effects, but images from the Mariner 10 and Pioneer Venus spacecraft showed that the brightenings, along with other 'deformities' in the planet's limb and terminator as observed from Earth, are real. They are related to the planet's atmospheric circulation. On Venus the predominant pattern is a single Hadley cell in each hemisphere which circulates from the equator almost to the poles. The cusp collars are thought to represent the northern- and southernmost limits of the circulation pattern, and the variations in the caps and collars are thought to reflect short- and long-term variations in the Hadley-cell circulation.

See also: Venus: atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cygnus

(the Swan; abbrev. Cyg, gen. Cygni; area 804 sq. deg.) A northern constellation which lies between Cepheus and Vulpecula, and culminates at midnight in late July. Its origin is uncertain, though it was known to the ancient Greeks, who identified it with one of the forms assumed by Zeus during his amorous pursuits, or with other mythological swans. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large, conspicuous constellation, Cygnus is easily recognized by the cruciform configuration of its five brightest stars, which gives it its alternative name of the Northern Cross. Its brightest star, α Cygni (Deneb or Arided), magnitude 1.3, marks the tail of the swan and is one of the stars that, together with α Lyrae (Vega) and α Aquilae (Altair), makes up the asterism of the Summer Triangle. Other bright stars include γ Cygni (Sadr), magnitude 2.2, ε Cygni (Gienah), magnitude 2.5, δ Cygni, a close binary with bluish-white (B9.5) and white (F1) components, magnitudes 2.9 and 6.6, separation 2.5″, period about 800 years, and β Cygni (Albireo), in the head of the swan, a particularly fine visual binary with orange and bluish-white (K3 and B8) components, magnitudes, 3.1 and 5.1, separation 34″, the former of which is itself a very close binary. There are ten other stars of magnitude 4.0 or brighter.

Other interesting double stars include o^1 Cygni, a triple system consisting of orange (K2) and bluish-white (A5) components, magnitudes 3.8 and 4.8, separation 338″, and a third, bluish-white (B5) component, magnitude 7.0, separation 107″ from the latter, and 61 Cygni, a binary with orange (K5 and K7) components, magnitudes 5.2 and 6.1, separation 30″, period about 650 years, the primary of which was the first star to have its trigonometrical parallax measured, in 1838. Interesting variable stars in Cygnus include the Mira-type variable χ Cygni (range 3.3–14.2, period about 409 days), the eruptive variable P Cygni (range 3–6, last outburst 1655) and the brightest U Germinorum star, SS Cygni (range 7.7–12.4, mean period 49.5 days), which is also an x-ray binary.

Cygnus lies in a particularly rich part of the northern Milky Way, which is at its brightest between γ and β, and is divided to the east by the Great Rift (or Cygnus Rift), a band of dark, absorption nebulae that continues through Aquila and Serpens Cauda to Ophiuchus, where it broadens out. Among the many open star clusters and nebulae in Cygnus are M39 (NGC 7092), a fifth-magnitude open cluster which contains a couple of dozen stars between seventh and tenth magnitude, the Veil Nebula (NGC 6960, 6974, 6979 and 6992–5), which is the optical component of the Cygnus Loop, the remnant of a supernova that exploded about 30 000 years ago, and NGC 7000 (the North America Nebula), a large (2°) area of mixed nebulosity.

Also in Cygnus are Cygnus X-1, an x-ray binary whose smaller component may be a black hole, and Cygnus A, a powerful extragalactic radio source thought to be two galaxies in collision.

See also: Albireo, Cygnus A, Cygnus X-1, Deneb, Great Rift, North America Nebula, 61 Cygni; SS Cygni.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cygnus A

The brightest radio source in the constellation of Cygnus, and the second-brightest cosmic radio source in the sky. This radio source, which had been detected in 1944 by Grote Reber (1911–), was shown in 1951 to coincide in position with a remote galaxy of unusual appearance (the galaxy appeared double, but is now known to be a single galaxy crossed by a lane of dust). Cygnus A was the first radio source to be detected beyond the confines of our own galaxy and the first example of what is now called a radio galaxy.

Cygnus A, which lies at a distance of about 750 million light-years, has a radio output a million times more powerful than that of a conventional galaxy like the Milky Way. At radio frequencies, it has a classic double-lobed structure, the main body of radio emission emanating from two elongated clouds that extend to a distance of some 200 000 light-years on either side of the center of the galaxy. The central radio source is linked to the outer lobes by two long, narrow filaments that are composed of energetic radio-emitting electrons that have been ejected from the core of the galaxy. The galaxy itself, which is exceptionally massive (about 10^{14} solar masses), is also an x-ray source, the x-ray emission coming from a distribution of hot gas with a temperature in the region of 10^8 K.

The intensity, or flux density, of Cygnus A at a wavelength of 20 cm in the radio region is about 1495 janskys.

See also: active galaxy, flux density, radio galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cygnus Loop

A large supernova remnant in the constellation Cygnus, measuring nearly 3° across. Some arcs of the Loop, known collectively as the Veil Nebula (or sometimes the Cirrus Nebula), are visible at optical wavelengths; radio, infrared and x-ray images reveal the complete Loop. The brightest part of the Veil is NGC 6992–5, to the east (centered on position RA $20^{h}\ 56.4^{m}$, dec. +31° 43′); the western section is NGC 6960 (RA $20^{h}\ 45.7^{m}$, dec. +30° 43′). Other sections are NGC 6974 (RA $20^{h}\ 50.8^{m}$, dec. +31° 52′) and NGC 6979 (RA $20^{h}\ 51.0^{m}$, dec. +32° 09′).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cygnus X-1

Cygnus X-1 is one of the strongest x-ray sources. It is the first celestial object for which we had reasonably convincing evidence that it is a BLACK HOLE. Its x-ray properties include an ultra-soft spectrum, compared to massive x-ray binaries containing a neutron star, rapid ($\sim$ 1 s) flickering, and high/low flux states with different spectral characteristics. In 1971, a RADIO SOURCE appeared at the time of a transition between the two flux states. This source is coincident with the O9.7Iab supergiant HDE226868. This star was found to be a single-line spectroscopic binary and ellipsoidal variable with an orbital period of 5.6 days.

Subsequent x-ray observations showed that the x-ray flux varies with the orbital period and another period found in the light curve, thereby confirming that Cygnus X-1 is a massive X-RAY BINARY. The radial velocities of H and He II emission lines in the optical spectra show that mass is flowing from HDE226868 toward Cygnus X-1. The emission line profiles suggest that mass is being transferred by a focused stellar wind.

Since Cygnus X-1 is not an eclipsing binary, we cannot determine the mass of the x-ray source with precision. However, we can place strong limits on critical system parameters by using a light curve synthesis program to model the ellipsoidal light variations and absorption line widths of HDE22686. The models are further constrained by the effective temperature of the star, the absorption line radial velocity curve, and the distance estimated from the increase in various effects of the interstellar medium with distance determined by measuring stars in the field around HDE226868. Fortunately, results from the modelling are not sensitive to assumptions that must be made during the process.

This analysis yields a strong lower limit on the mass of Cygnus X-1, $M_{\rm x} > 7M_{\odot}$. This is more than twice the upper limit for the mass of a neutron star, $M_{\rm ns} < 3.2M_{\odot}$, according to the theory of GENERAL RELATIVITY. Cygnus X-1 was the first object for which there was a convincing argument that it is a black hole. It was nearly a decade after the first recognition that Cygnus X-1 must be a black hole before another example was discovered.

The x-ray source

The strong x-ray source Cygnus X-1 was first discovered by a rocket-borne x-ray detector in 1964. It is one of the five strongest x-ray sources in the energy range 2–10 keV. It has two distinct x-ray 'states', originally labeled *high* and *low*. The x-ray spectrum is harder, that is, has a greater proportion of high-energy photons, in the low state, not only because of a decline in the low-energy x-ray flux, but also because the x-ray flux for energies greater than 10 keV increases at the same time. Similarly, the high-energy flux declines when the low-energy flux rises. As a result, it is now common practice to label the 'states' *soft* and *hard* respectively. Long-term monitoring of Cygnus X-1 has shown that the different states can last for weeks to years. There are no characteristic time scales for the two states. Moreover, at least two other 'states' have been observed at least once. These states must reflect changes in the accretion rate or structure of the ACCRETION DISK, but as yet no one has offered a model that explains all of the observations.

The x-ray flux variations are not confined to the state changes. In fact, the x-ray flux varies on time scales from $\sim$1 ms to at least the time scales associated with the state changes. These include flux variations that are modulated with the orbital period and a variable, longer period that may be a precession period of some structure in the system. These are superimposed on a background of shot noise and occasional flares.

The 5.6 day x-ray flux variation is easiest to understand. The minima last for 27% of the orbital period and are centered precisely around the time of superior conjunction of the black hole with the O-star. The minima are deeper in the lower energy bands. The phasing and spectral shape of the minima indicate that the variation is due to photoelectric absorption by the focused stellar wind from the O-star. These variations are observed only during the hard state. The anticorrelation of the Hα emission line equivalent width with the soft x-ray flux during the 1996 soft state lends support to this idea.

Monitoring of the 3–12 keV flux of Cygnus X-1 from 1969 to 1980 showed that the average flux varies with a period $P = 294 \pm 4$ days. The x-ray flux is at the maximum level for approximately 40% of the 294 day period. The decline to minimum flux, which is about 75% of maximum flux, takes half as long as the recovery back to maximum flux. From August 1996 to September 1998, during the hard state following the 1996 soft state, both the 2–10 keV and 20–100 keV flux varied with a period of 142 ± 7 days. The amplitude in the 2–10 keV band was nearly double that observed earlier and the shape of the flux curve was almost a mirror image of the the one for the 294 day period. It remains to be seen whether the curious relationship between the periods and light curves is anything more than a remarkable coincidence.

The short-time-scale variations of the x-ray flux can be described by shot noise, a type of noise resulting from the superposition of a large number of disturbances (shots) that occur at random. The shots have a characteristic time scale of about 20 ms, but their characteristics vary in a complex fashion with energy and x-ray state. The shots all rise more slowly to peak intensity than they decline. The peaks of the shots in the higher energies are narrower and more asymmetric than those in the low energy bands. During the hard and transition states, the average hardness of the shots is lower than the steady emission. During the soft state, the average hardness of the shots is greater than the steady emission. The hard x-ray shots lag behind the soft x-ray shots by an amount that is a function of energy. This delay is smallest for the hard state, and slightly larger than the soft state, and much larger for the transition state. A separate study of the Fourier time lag, the time lag between the same phase in two different

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

time series, reached similar conclusions using x-ray hard state data. This implies that nearly all of the hard and soft flux variations occur simultaneously. At the same time, the time series in different energy channels show nearly complete coherence down to ≈0.02 Hz, and after a slight loss of coherence near 10^{-2} Hz, recovered back to nearly unity at 0.005 Hz. This is difficult to understand if the hard and soft flux variations are produced nearly simultaneously, since the hard photons are scattered many more times than the soft photons.

The hard state spectrum of Cygnus X-1 can be represented by the sum of a soft thermal component and a hard, non-thermal power-law component with a high-energy cut-off near 150 keV. The power-law spectrum is modified by reprocessing features, in particular a weak iron fluoresence line at 6.4 keV and possibly a Compton reflection hump near 30 keV. During a soft state, the soft thermal component increases by a factor of several while the hard non-thermal power-law component declines by 10–20%. All four spectral components are usually explained in the context of accretion disk corona models in which the thermal radiation originates in an accretion disk and is Comptonized in a hot plasma, the 'accretion disk corona'. These models can successfully reproduce the spectral variations between the soft and hard states if the accretion disk extends inwards to the last stable orbit around the black hole and the central corona fills a small volume during the soft state, and the inner edge of the disk is much further from the black hole and the central corona much larger in the hard state. However, there is no generally accepted model for the geometry or dynamics of the accretion disk/accretion disk corona system. At this time, it is not clear that any of the models have succeeded in reproducing anything more than the x-ray spectrum. Attempts to explain the short-time-scale variability have had limited success, at best, and the precession has been treated qualitatively as a separate problem.

The optical light curves

In order to determine the masses of stars in a BINARY SYSTEM we need to determine its orbital inclination i, where i is the angle between the orbit plane and the plane of the sky. In principle, when there are no eclipses, and a binary system cannot be resolved, optical polarimetry is the best method to determine i. However, HDE226868 is so faint that the photometric errors render the polarimetric result, $25^\circ < i < 65^\circ$, essentially useless for determining i.

Because of the nature of the mass transfer in the Cygnus X-1 system, we expect that HDE226868 comes very close to filling its ROCHE LOBE. This means that the star will be strongly distorted into a roughly ellipsoidal shape with the long axis of the ellipsoid aligned with the line between the star and the black hole and the end of the ellipsoid directed toward the black hole drawn out into a point. Regions of the star that are farther from the center will have lower surface brightnesses than those that are close to the center. This is called gravity darkening or gravity brightening.

If there are no other light sources and we view the tidally distorted star from the direction of the orbit plane, i.e. $i = 90^\circ$, the light curve will have a shallower, rounded minimum at orbital phase $\phi = 0.0$, superior conjuction of the black hole, a sharper, deeper minimum at $\phi = 0.5$, inferior conjunction of the black hole, and broad equal maximum between the minimum. The luminous flux is smaller at the conjunctions because the areas of the star projected on the sky are smaller, the surface brightnesses of those regions are lower because of the effects of gravity darkening. The depths of the minima depend on temperature. In this case, the depths will increase with decreasing wavelength because of the high effective temperature of HDE226868.

In the more general case, the amplitude of the light variation depends on the mass ratio $q = M_x/M_{opt}$, where M_x is the mass of the black hole, and M_{opt} is the mass of HDE226868, the fill-out factor ρ, the ratio of the volume radius of the star (i.e. the radius of a sphere with volume equivalent to that of the star) to the radius of a sphere with a volume equivalent to the volume of the Roche lobe, and orbital inclination angle i. Larger q and ρ will increase the tidal distortion of the star and thereby increase the amplitude of the light curve. As i is decreased from 90°, the amplitude of the light curve is reduced and the differences in amplitude and shape between the two minima gradually disappear. In addition, at low inclinations, the gravity darkening effects become so small that the amplitude of the light curves are independent of wavelength in the optical region at least.

The optical light curves of HDE226868 observed during the hard state of the x-ray source look approximately like those expected for a low-inclination system. The amplitudes of the U, B and V light curves are approximately 0.025 mag. The light curve is phased as expected with respect to the radial velocity curve. If HDE226868 fills its Roche lobe, the semi-amplitude of the light curve is almost independent of the mass ratio, and it can be used to place a lower limit on the orbital inclination angle $i \geq 27^\circ$.

There are two problems that make it difficult to work with the light curves. First, the rms scatter about the light curve is 0.022 mag, almost half the amplitude of the intrinsic variability. This is comparable to the variability seen in photometry of single, luminous OB stars and is probably due to the same cause—either density inhomogeneities in their winds or multiperiodic non-radial pulsation. Unfortunately, this extra intrinsic variability means that we have to average observations over several years get reliable ellipsoidal light curves. Second, an additional component of scatter in the ellipsoidal light curves is due to periodic changes in the shape of the light curve.

After the 5.6 day ellipsoidal light curves have been subtracted from the luminous fluxes in the U and B bands obtained in the during the x-ray hard state between 1972 and 1983, they have light curves with $P = 294$ and the same shape and phasing as the x-ray flux in data taken between 1972 and 1983. The variation is largest,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

0.0061 ± 0.0012 mag, in U, smaller in B, and statistically undetectable in the V band. Comparison of the 5.6 day light curves obtained during the x-ray bright and faint phases of the 294 day period show that when the x-ray source is brighter, the optical light curves are brighter at all phases except around the time of inferior conjunction of the x-ray source. The phasing and relative amplitude of the optical light curves suggests that the optical variations are due to a hot source between the x-ray source and HDE226868. No models have been proposed that explain why the optical and x-ray fluxes vary in phase. The orbital modulation of the optical light curves disappeared during the 1996 soft x-ray state, while the 142 day long-term modulation continued. The orbital modulation returned to normal when the x-ray source returned to the hard state.

Clearly these results raise serious questions about the standard assumptions that are made in various schemes to derive the mass of the black hole in Cygnus X-1. A careful inspection of the many mean light curves is not encouraging. None of them is a good fit to a standard ellipsoidal light curve model.

HDE226868: spectroscopic observations

The optical spectrum of HDE226868 has been classified O9.7Iab(p-var), where the peculiarity relates to weak, variable emission in the He II λ468.6 nm line. Weak variable emission is also present in the Hα, Hβ and Hγ lines. The absorption line spectrum is very similar to those of comparable supergiants, but the star's atmosphere may be overabundant in He and C. RADIAL VELOCITIES measured from the absorption lines show that HDE226868 is a single-line spectroscopic binary with a orbital period $P = 5.5998$ days and radial velocity semiamplitude $K_{\rm opt} = 74.7$ km s^{-1}. This yields a mass function

$$f(M) = \frac{M_{\rm x} \sin^3 i}{(M_{\rm opt} + M_{\rm x})^2} = 0.25 \pm 0.01.$$

Measures of the radial velocities of the emission lines yield a velocity curve where the maximum velocity is shifted about $0.75P$ from maximum radial velocity of the absorption lines. This indicates that the emission lines are formed in gas that is lying on or near the line joining the two stars and is flowing from the visible star to the x-ray source. This shows that the visible star must fill, or nearly fill, its Roche lobe. Detailed analyses of the emission profiles show that they are blends of two components, a P Cygni profile[1] that follows the absorption line radial velocity curve and a broad, roughly symmetric, emission feature that is shifted $0.86P$ with respect to the absorption line velocity curve.

These analyses indicate that the emission lines form in a region less than $0.1R_{\rm opt}$ above the surface of HDE226868, where $R_{\rm opt}$ is the star's radius. The average flow velocity in this region is less than 100 km s^{-1}. The width of the emission features, FWHM < 400 km s^{-1} is consistent with the velocity gradient expected in a focused stellar wind. The flow velocity and velocity gradient in a gas stream are unlikely to be as large so close to the star.

The emission from the outflow is not occulted. If the star fills 90% of its Roche lobe and we make a reasonable allowance for error in the position where the emission forms, this implies that $i < 55^\circ$. The limit is reduced significantly if the star fills a larger fraction of its Roche lobe.

The upper limit on the peak intensity of Hα emission from the accretion disk is 3% of the continuum flux from HDE226868. This is not surprising because the accretion disk is expected to be much fainter than HDE226868.

The mass of the compact object

The light and velocity curves of HDE22686 do not provide the information required to solve for the critical parameters of the binary system, especially the mass of the x-ray source. It is possible, however, to obtain significant limits on the important parameters by combining information derived from these curves with other data. If the spin and orbital angular momentum vectors are aligned, the ratio of the projected rotation velocity $V \sin i$ to the absorption-line velocity curve semi-amplitude $K_{\rm opt}$ is independent of the orbital inclination angle

$$\frac{V \sin i}{K_{\rm opt}} = \rho\Omega \left(\frac{1}{q} + 1\right) \Phi(\Omega, q)$$

where Ω is the ratio of the star's spin angular velocity to its orbital angular velocity, q is the mass ratio, and $\Phi(\Omega, q)$ is the Roche lobe volume radius divided by the system separation. This relation can be used to derive q for a grid of ρ and Ω.

Theoretical calculations show that there will be catastrophic Roche lobe overflow that would smother the x-ray source if $\rho = 1$. Moreover, theoretical calculations of emission line profiles for focused wind models show that for $\rho < 0.9$ the wind density is not high enough along the line joining the stars to produce the observed emission. The circularity of the orbit suggests that tidal forces will have had enough time to bring the star into near synchronous orbit, so it is reasonable to confine our grid to values of Ω near unity.

The amplitude of the light curve depends on q, ρ, Ω and i. The shape of the star is modeled with a potential that includes the effects of radiation pressure from both the star and x-ray source and asynchronous rotation, to first order. The modeling process is begun by guessing values of q and i, and these are used to obtain the component masses, the radius and surface gravity of the visible star. The V magnitude and extinction can be used to relate the visible star's radius to its distance d in kpc, $R_{\rm opt} = (9.7 \pm 0.9)d$. The distance can then be used to compute the luminosities of the visible star and x-ray source, which are required to estimate the ratio of

[1] A P Cygni profile consists of a blue shifted-absorption trough and a red-shifted emission line that joined at approximately the wavelength of the line in the reference frame of the star. They form in spherically symmetric outflows from a star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the radiative to gravitational acceleration that are required when including the effects of radiation in calculating the potentials. The fluxes and specific intensity profiles are calculated using a model atmosphere code. The specific intensity profiles are calculated for $T_{\rm eff} = 27\,500$ K and surface gravity that is appropriate for the parameters selected.

These data are input into a light curve synthesis code, where they are integrated over the surface of the star while taking into account the effects of gravity darkening and limb darkening. If the FWHM of the flux line profile and the amplitude of the light curve do not match the observations, q and i are adjusted, and the process is repeated until a match is obtained. Since the amplitude of the light curve is most sensitive to i, and the FWHM depends entirely on q, the selection of adjustment is relatively easy and the process converges within a few iterations. The 1972 to 1983 V light curve was used for this purpose because it was the one that seemed least affected by the perturbations noted earlier.

Fortunately, the results of this exercise are insensitive to the assumptions made during the calculations. These include the value of the macroturbulent velocity used in calculating the theoretical line profiles, the inclusion or omission of the effects of radiation pressure, and coalignment of the angular momentum vectors. For near synchronous rotation and a generous allowance for the uncertainties, $M_{\rm opt} = 33 \pm 9\ M_\odot$ and $M_{\rm x} = 16 \pm 5\ M_\odot$. These values are consistent with distance estimates based on the UV and optical colors of HDE226868 and studies of the interstellar absorption as a function of distance for stars in the surrounding field. The absolute lower limit on the distance based on these data is $d > 1.7$ kpc. This corresponds to $M_{\rm opt} = 20\ M_\odot$ and $M_{\rm x} = 7\ M_\odot$. An alternate analysis based on unified atmosphere-wind models fitted to the spectra of HDE226868 obtained somewhat lower values, $M_{\rm opt} = 18\ M_\odot$ and $M_{\rm x} = 10 M_\odot$. This result also depends implicitly on the light curves to obtain the inclination, though less so than the first method, and no error estimates were given. If they had been, it is likely that the errors for the two methods would have overlapped.

Bibliography

The history of the observations and modeling of HDE226868 can be traced through the following papers:

Bolton C T 1975 Orbital elements and an analysis of models for HDE226868 = Cygnus X-1 *Astrophys. J.* **200** 269–77

Gies D R and Bolton C T 1986 The optical spectrum of HDE226868 = Cygnus X-1: II spectrophotometry and mass estimates *Astrophys. J.* **304** 371–88

Sowers J W, Gies D R, Bagnuolo W G Jr, Shafter A W, Wiemker R and Wiggs M S 1998 Tomographic analysis of Hα profiles in HDE226868/Cygnus X-1 *Astrophys. J.* **506** 424–30

A good summary of the x-ray properties of Cygnus X-1 can be found in

Liang E P and Nolan P L 1984 Cygnus X-1 Revisited *Space Sci. Rev.* **38** 353–84

For a good review comparing Cygnus X-1 with other stellar systems that are believed to contain black holes see

Cowley A P 1992 Evidence for black holes in stellar binary systems *Ann. Rev. Astron. Astrophys.* **30** 287–310

C T Bolton

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cysat, Johann Baptist [known as Cysatus] (1586–1657)

Swiss Jesuit astronomer who used a telescope to discover the Orion Nebula in 1618.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

d'Alembert, Jean-Baptiste Le Rond (1717–83)

French mathematician, born in Paris. A foundling, he was named after the place where he was left, the church of St Jean Baptiste de Rond. He published *d'Alembert's principle* (the principle that the action of a force produces an equal and opposite reaction in the mass that it is accelerating) in *Traité de Dynamique* (1743), which was a new interpretation of Newton's dynamics. He used his new formulation of Newton's laws to study physical astronomy and planetary motions. As scientific editor, he was one of the group of French intellectuals, the so-called Encyclopedists, including Diderot, Voltaire, Montesquieu and Rousseau, who composed the monumental 28 volume encyclopedia *Dictionaire Raisonné des Sciences, des Arts et des Métiers*, expounding the philosophy of the French Enlightenment, including mathematics and astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

d'Arrest, Heinrich Louis (1822–75)

Born Berlin, Germany, worked at the Berlin Observatory and Copenhagen Observatory. At Berlin he participated with JOHANN GALLE in the discovery of Neptune. Discovered comets and asteroids and observed nebulae, including the variable nebula of JOHN HIND.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Danish Space Research Institute

The Danish Space Research Institute (DSRI), was established in 1968. The present scientific staff is 18 physicists. DSRI is involved in providing scientific instruments for the Russian Spectrum-X-Gamma mission, the ESA Gamma-Ray Observatory Integral, the ESA Cosmic Background Radiation mission Planck Surveyor, and magnetometers for several missions including the Argentinian SAC-C satellite.

For further information see http://www.dsri.dk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dante Alighieri (1265–1321)

Italian poet, born in Florence, Italy, author of *The Divine Comedy*, the story of the poet's journey through Hell and Purgatory (guided by Virgil) and through Heaven (guided by Dante's wife Beatrice, to whom the poem is a memorial). The poem brings together the whole medieval picture, integrating a religious and moral perspective with a geographical and astronomical one, depicting a changeless universe ordered by God. This is how the medieval world imagined the Ptolemaic system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dante, Egnatio [Egnazio, Ignazio] (1536–86)

Astronomer, instrument-maker and cartographer. Born in Perugio, Italy, of a father who also apparently made astronomical instruments, he became a Dominican monk, and published works on Sacrobosco's *Sphere*, and other astronomical and mathematical topics. He made maps under the patronage of Cosimo I de' Medici, and a large terrestrial globe, which still exists. He was later professor of mathematics at Pisa and Bologna. In 1574 he determined by observation that the equinox was 11 days earlier than the calendar then in use. Dante continued to refine the observations to quantify this accumulated error, and, being appointed cosmographer and mathematician to Gregory XIII he was commissioned to map the papal states and to lead work to reform the calendar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dark Matter in Galaxies

Dark matter in spiral galaxies

SPIRAL GALAXIES are flat rotating systems. The stars and gas in the disk are moving in nearly circular orbits, with the gravitational field of the galaxy providing the inward acceleration required for the circular motion. The rotation of these galaxies is usually not like a solid body: the angular velocity of the rotation typically decreases with radius. To a fair approximation, assuming Newtonian gravity, the rotational velocity $V(r)$ at radius r is related to the total mass $M(r)$ within radius r by the equation $V^2(r) = GM(r)/r$, where G is the gravitational constant.

The radial variation of the rotational velocity (the rotation curve) is most readily measured from the gas in the disks. The emission lines of ionized gas in the inner regions are measured with optical spectrographs. With radio synthesis telescopes, rotation curves can be measured from the neutral hydrogen (H I) which emits a narrow spectral line at 1420 MHz (21 cm wavelength). The interest in measuring rotation curves of spiral galaxies is that they give a direct measure of the radial distribution of the total gravitating mass.

Until the early 1970s, most of the rotation data for spirals came from optical observations which did not extend beyond the luminous inner regions. At that time, the optical rotation curves seemed consistent with the distribution of luminous matter. With the construction of RADIO TELESCOPES like the Westerbork Radio Synthesis Telescope in The Netherlands, it became possible to measure the distribution and rotation of the H I in spiral galaxies. It was soon discovered that the H I in many spirals extended far beyond the starlight, and that the H I rotation curves in such galaxies often showed nearly constant rotational velocity out to the radial limits of the data.

This was unexpected, because a flat rotation curve means that the total mass of the spiral within some radius r increases linearly with r, while the total luminosity approaches a finite asymptotic limit as r increases. It soon became clear that a large amount of invisible gravitating mass (more than 90% of the total mass in some examples) is needed to explain these flat rotation curves.

The problem is illustrated in figure 1. The rotation curve comes from H I observations of a well studied spiral, NGC 3198. The curve labelled *disk* is the rotation curve that would be expected if the surface mass density in this galaxy were proportional to the light distribution shown in the upper panel. In this analysis, the constant of proportionality (the ratio of mass to light) was made as large as possible, with the criterion that the total expected rotation curve should not exceed the observed rotation curve. The gas in the galaxy also contributes to the expected rotation curve, as shown by the curve labelled *gas* (see also GAS IN GALAXIES). The contributions to the expected rotation curve from stars and gas must be added in quadrature to derive the total expected rotation curve. From the curves in this figure, there is no way that the stars and gas together can produce the flat observed rotation curve. An additional massive and extended distribution of dark matter is needed. The rotation contribution for a simple dark halo model is shown in the figure: the model is chosen so that the dark halo plus disk plus gas together give the rotation curve that passes through most of the observed points in the figure. At the radial limit of the data, the dark halo is providing most of the total gravitational field.

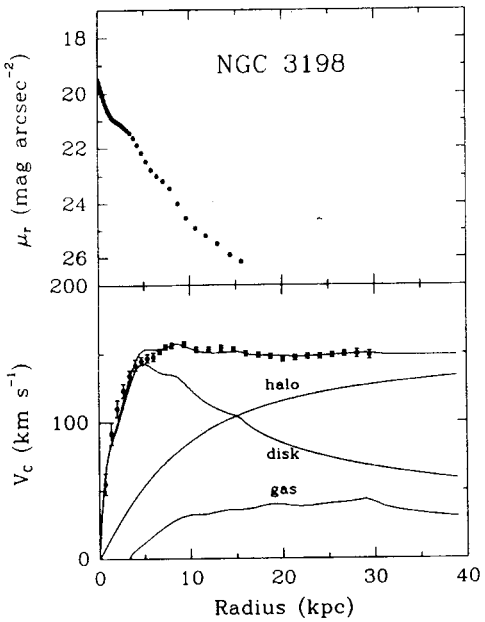


Figure 1. The upper panel shows the R-band radial surface brightness distribution of the spiral galaxy NGC 3198. The lower panel shows its H I rotation curve (points). The curve labelled *disk* shows the expected rotation curve if the surface density distribution followed the surface brightness distribution in the upper panel. The curve labelled *gas* is the contribution to the rotation curve from the observed gas. Together, the gas and the disk cannot reproduce the observed flat rotation curve at large radius. An extra gravitating component, the dark halo, is needed. The curve labelled *halo* is the rotation curve of the adopted dark halo model: the three labelled rotation curves, when added in quadrature, produce the total rotation curve that passes through the observed points. (From Begeman K 1987 *PhD Thesis* University of Groningen.)

The halo model in figure 1 is a minimum halo, in the sense that the contribution of the disk to the rotation curve was made as large as possible. If a lower mass-to-light (M/L) ratio had been adopted for the disk, then a more centrally concentrated halo would be needed to make up the larger discrepancy between the observed and expected rotation curves. It is difficult to measure the M/L ratio for disks independently of the rotation curve itself, and there is still a lot of controversy about the correctness of the maximum disk approach as shown in figure 1.

The situation shown in figure 1 is typical of almost

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

all DISK GALAXIES that have extended H I distributions, including the small DWARF IRREGULAR GALAXIES that are also in rotational equilibrium. In almost every example, a massive dark halo that dominates the enclosed mass $M(r)$ at large r is needed to explain the observed rotation curves. The nature of this dark matter remains unknown (see DARK MATTER: ITS NATURE), and is one of the great problems of modern astrophysics.

For decompositions as shown in the figure, the dark halo is often modelled for simplicity as a spherical system with a radial volume density distribution of the form $\rho(r) = \rho(0)/(1 + r^2/a^2)$, where a is a scale length and $\rho(0)$ the central density. This particular distribution is chosen because it generates a flat rotation curve at large radius.

The spherical shape is simply an assumption. It turns out to be very difficult to measure the shapes of dark halos. The thin neutral hydrogen layer in edge-on spirals is observed to flare beyond the edge of the stellar disk, where the vertical restoring force of the disk is reduced. The extent of this flaring can be used to estimate the flattening of the dark halo (a flat halo would lead to less flaring). Another idea is to study the POLAR RING GALAXIES, in which material is rotating in two orthogonal planes so the gravitational field in both planes can be measured. Both of these approaches have uncertainties, but they hint that dark halos may be significantly flattened.

In figure 1, the gently rising rotation curves for the halo and for the gas are roughly similar in shape out to the radial limit of the observations. This is again typical for disk galaxies with extended H I distributions, and hints that the gas and the dark matter distributions are in some way closely related. Some authors have argued that the dark matter may be in the form of cold gas, but this remains contentious.

Dark matter in our Galaxy

Although our location in the disk of our Galaxy makes it difficult to measure its rotation curve out to large radius, other less secure mass estimators become possible in this situation. For example, the velocities of halo stars out to distances of 100 kpc from the Sun can be used to estimate the total galactic mass. Some halo stars pass through the solar neighborhood at velocities in excess of 600 km s^{-1}. If these stars are bound to the Galaxy, then they provide another mass estimator. The ANDROMEDA GALAXY, at a distance of 770 kpc, is approaching the Milky Way at about 120 km s^{-1}, which provides yet another mass estimator by assuming that the combined mass of the two galaxies is sufficient to turn around their initial expansion within the lifetime of the universe. All of these estimators together indicate that the enclosed mass of the Milky Way increases linearly with radius out to at least 150 kpc and that its total mass is about $15 \times 10^{11} M_\odot$. The luminous mass is unlikely to exceed $1.2 \times 10^{11} M_\odot$, so the mass of our Galaxy is probably more than 90% dark.

If the dark matter is in the form of compact objects of stellar or substellar mass (e.g. jupiters, white or brown dwarfs, neutron stars), then dark halo objects passing between the observer and a distant star cause the distant star to brighten as they pass, through GRAVITATIONAL LENSING. The duration of the brightening depends on the mass and velocity of the lens and the geometry of the event. For stars in the MAGELLANIC CLOUDS, the typical duration of the brightening would be a few months for dark halo objects of solar mass. Several groups have attempted to detect dark objects in this way. The MACHO experiment monitored the brightnesses of several million stars in the LMC over a period of about 7 years, and has detected about 15 lensing events. If these lensing objects lie in the galactic halo, then they provide about half of the total dark matter and their individual masses are about $0.5 M_\odot$. This mass is typical of white dwarfs and suggests that these dark objects could be old WHITE DWARFS, perhaps remnants of a very early and very massive burst of star formation. There is no evidence for a significant population of luminous stars associated with the dark halo, so a rather special stellar mass function would be needed to produce large numbers of old white dwarfs without producing even larger numbers of old luminous stars. It remains possible that the lenses are not in the galactic halo. For example, if the LMC is sufficiently extended along the line of sight, the lenses could be objects in the LMC lensing other objects in the LMC.

Dark matter in elliptical galaxies

The evidence for dark matter in disk galaxies is secure because their dynamics is so simple. ELLIPTICAL GALAXIES are not supported by rotation, so rotation curves cannot be used to study their dark matter content. Other dynamical mass indicators are needed. Some of the largest ellipticals are embedded in an envelope of hot x-ray emitting gas. If the radial temperature distribution of this hot gas is known, its hydrostatic equilibrium can be used to measure the total mass of the parent elliptical. The motions of stars, globular clusters and planetary nebulae in the outer regions of elliptical galaxies can also be used to estimate their masses, although this requires some assumptions about the orbits of these tracer objects. The weak gravitational lensing of background galaxies provides an interesting and independent statistical way to measure the dark matter content of elliptical and disk galaxies.

The evidence indicates that the largest elliptical galaxies contain substantial amounts of dark matter. We can use the mass-to-light ratio as a useful measure of the dark content of a galaxy. An old stellar population without dark matter has a mass-to-light (M/L) ratio of about 5 in solar units, depending on its metallicity. Disk and elliptical galaxies can have M/L ratios as high as $\sim$80. Some of the smallest elliptical galaxies, the dwarf spheroidal companions of the Milky Way, show very high dark matter content, with M/L ratios $\sim$100. For these systems, the masses are estimated from velocities measured for many individual stars. Again, some assumptions about the stellar orbits are needed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Formation of dark halos

In the current picture of GALAXY FORMATION, clumps of dark matter mixed with baryons come together to form a galaxy. The nature of the dark matter is unknown, but it is believed to be non-dissipative. The dark halos then gradually build up into weakly flattened spheroidal structures while, in disk galaxies, the rotating baryons dissipate into flat disks in near-circular motion within their dark halos. There are now enough data on dark halos to determine how their properties change with the brightness of the visible galaxy. It turns out that the faintest galaxies have the densest dark halos, about 1000 times denser than the dark halos of the brightest galaxies. This indicates that the halos of lowest mass emerged first from the expanding universe, when the density of the universe was high, as expected from theoretical arguments.

For more information on dark matter in the universe, see the article on GALAXY CLUSTERS.

Bibliography

For a useful recent overview of work on dark matter in galaxies see

Persic M and Salucci P (ed) 1997 *Dark and Visible Matter in Galaxies (ASP Conf. Ser. 117)* (San Francisco: Astron. Soc. Pacific)

Ken Freeman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dark Matter: Its Nature

What is the stuff that makes up most of the mass in the universe? The answer to this simple-sounding question is far from obvious, and actually presents one of the greatest unsolved mysteries of astrophysics, cosmology, and elementary particle physics. With conventional astronomical methods one can only see the 'luminous matter', notably in the form of stars, which reveals its presence by the emission of light. On the other hand, one can determine the gravitating mass of various systems such as spiral galaxies or galaxy clusters from their dynamical properties, and one finds a huge discrepancy relative to luminous matter. Assuming the usual law of gravity, one is led to conclude that there are large amounts of 'dark matter', a term first introduced by FRITZ ZWICKY in his seminal paper of 1933 where he studied the dynamics of GALAXY CLUSTERS.

The physical properties of dark matter can be constrained by several powerful astrophysical and cosmological arguments which disfavor 'baryonic matter' as a main constituent. This term refers to hydrogen, helium, and the heavier elements which, besides electrons, consist of protons and neutrons, falling into the 'baryon' category of ELEMENTARY PARTICLES. But some new form of matter appears to hold galaxies and galaxy clusters gravitationally together! The most popular explanation is that of 'particle dark matter' which goes back to Cowsik and McClelland who speculated in 1973 that neutrinos could play this role. However, while recent experiments indicate that neutrinos do have mass—which has been an open question for decades—it looks impossible to attribute all of the dark matter to these weakly interacting particles.

One is thus led to postulate hitherto undetected elementary particles for the cosmological dark matter. On the other hand, there are already independent particle-theory motivations for certain new particles which could well play this role. The dark-matter problem thus provides one of several links between particle physics and cosmology; it is a key ingredient of 'astroparticle physics', or alternatively 'PARTICLE ASTROPHYSICS'. The laws of the microcosm of elementary particles and the macrocosm of the largest structures in the universe, inner space and outer space, are closely intertwined!

Perhaps the most remarkable development of the 1990s is that the physics of dark matter has turned into a truly experimental science. If any of the popular speculations about the nature of dark matter are correct, this mysterious stuff may well turn up in one of the current or near-future direct search experiments.

In the following, the astrophysical motivation for the reality of dark matter and the most important astrophysical constraints on its nature will be discussed. Well-motivated candidates, current search strategies and preliminary results will then be reviewed.

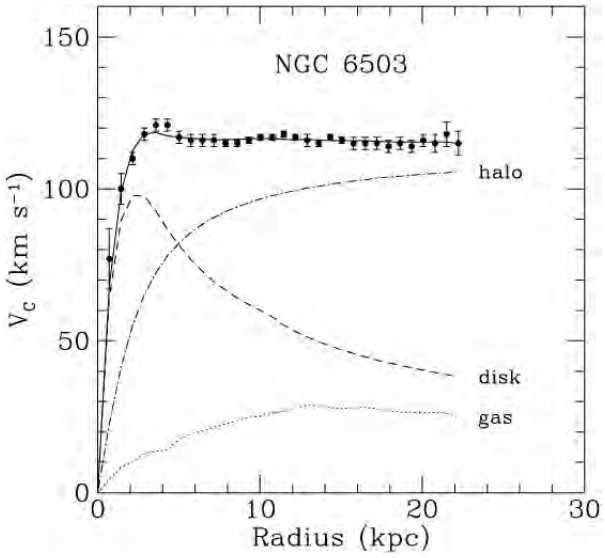


Figure 1. Rotation curve of the spiral galaxy NGC 6503 as established from radio observations of hydrogen gas in the disk (K Begeman *et al Mon. Not. R. Astron. Soc.* **249** 439 (1991)). The dashed curve shows the rotation curve expected from the disk material alone, the chain curve from the dark-matter halo alone.

Dynamical evidence

Rotation curves of spiral galaxies

Why are astronomers so sure that there are large amounts of dark matter lurking everywhere in the universe? The flat rotation curves of SPIRAL GALAXIES provide perhaps the most impressive evidence. These systems consist of a central bulge and a thin rotating disk. It is natural to measure its orbital velocity as a function of galactocentric radius by virtue of the Doppler shifts of spectral lines. Galaxy disks tend to contain neutral hydrogen which can be observed by its 21 cm line emission, allowing one to measure the rotation curves to much larger radii than with optical tracers.

The example of figure 1 illustrates the general behavior of the rotation curves. The orbital velocity rises from the center outward until it reaches a value of the order of 100 km s^{-1}, where it stays constant out to the largest measured radii, a systematic trend already diagnosed by Freeman in 1970. This behavior is entirely unexpected because the surface luminosity of the disk falls off exponentially with radius, implying that the mass M of luminous matter, mostly stars, is concentrated around the galactic center. Thus one expects a Keplerian decline of the orbital speed, $v_{\rm rot} = (G_N M/r)^{1/2}$ (Newton's constant G_N, radius r), in analogy to the planetary motions in the solar system—see the dashed line in figure 1.

The difference between the expected and measured rotation curve is ascribed to the gravitational effect of

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

dark matter. A number of arguments suggest that this material is not part of the galactic disk itself. First, in our galaxy the vertical distribution of stars together with their velocity dispersion reveals that there is no significant amount of disk dark matter. Second, a thin self-gravitating disk is dynamically unstable. Third, the hydrogen is vertically more extended than would be expected if all of the gravitating matter were in the disk, especially at large radii ('hydrogen flaring').

An overall picture of spiral galaxies emerges where the bulge and disk are dynamically subdominant components immersed in a huge dark-matter halo. It is not crucial that this halo be strictly spherical; it may well be somewhat oblate or even triaxial.

For the direct detection of dark matter, our own Milky Way is the most interesting galaxy. Its rotation curve conforms to the standard picture with an approximate plateau value for the rotation velocity of 220 km s^{-1}. The dark-matter density in the solar neighbourhood implied by models of the halo is 300MeVcm^{-3} within about a factor of two, i.e. roughly the mass equivalent of a hydrogen atom per 3 cm^3. (Note that the atomic mass unit corresponds to 931.5 MeV.)

Cosmic density contribution

The contribution ρ of a given matter component to the overall density of the universe is usually expressed in terms of the 'omega parameter' $\Omega=\rho/\rho_{\rm crit}$. Here, $\rho_{\rm crit} = 3H^2{}_0/(8\pi G_{\rm N}) = h^2\ 1.88\tilde{\rm A}$—$10^{-29}$ g cm^{-3} is the critical density with H_0 the present-day cosmic expansion parameter ('Hubble constant'). It is usually written as $H_0 = h$ 100 km s^{-1} Mpc^{-1}; different measurements span the approximate range h = 0.5–0.8. In the framework of the usual Friedmann–Lemaître–Robertson–Walker cosmology the spatial geometry of the universe is Euclidean for Ω = 1 ('flat universe'), the spatial curvature is negative for $\Omega<1$ ('open universe'), and positive for $\Omega>1$ ('closed universe').

The cosmic density contribution of galaxies is related to the measured luminosity density of the universe, (1.7 ± 0.6)–$10^8\ h\ L_\odot$ Mpc^{-3} in the visual spectral band (solar luminosity $L_\odot$). Most of the light is produced by stars; thus multiplying with the mass-to-light ratio for a typical stellar population of a few $M_\odot/L_\odot$ gives us the luminous matter density. One finds

$$\Omega_{\rm lum}h = 0.002 - 0.006$$

to be a consistent range. Therefore, luminous matter alone is far from the critical density.

Typical mass-to-light ratios of galaxy halos are at least around 30 h as far as the measured rotation curves reach, providing a cosmic density of $\Omega_{\rm gal}\gtrsim$0.03–0.05. The rotation curves tend to stay flat out to the largest measured radii and thus the true size of dark-matter halos is not known. Estimating their extent from galactic satellite dynamics yields $\Omega_{\rm gal}h$ = 0.2–0.5.

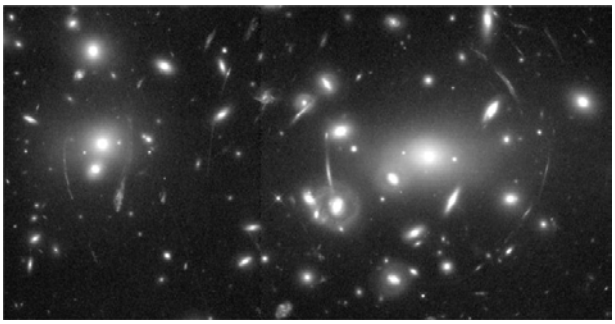


Figure 2. Hubble Space Telescope image of the galaxy cluster Abell 2218, showing a number of arcs and arclets around the two centers of the cluster. They are distorted images of background galaxies, caused by the cluster's gravitational lens effect. (NASA HST Archive.)

Clusters of galaxies

Galaxy clusters are the largest gravitationally bound systems in the universe. Zwicky first noted in 1933 that the velocities of the member galaxies tend to be so large that huge amounts of dark matter are needed to bind them gravitationally. Taking a cluster mass-to-light ratio of around 300 $M_\odot/L_\odot$ as representative for the universe leads to a cosmic matter density of $\Omega_{\rm M}$ = 0.1–0.4.

Recently it has become possible to measure cluster masses from the image distortion of background galaxies caused by the cluster's gravitational light deflection ('weak gravitational lensing'); an example is shown in figure 2. One uses the statistical distribution of these 'arclets' to reconstruct the shear field of gravitational image distortions and from there one can derive cluster mass distributions. (see GRAVITATIONAL LENSISNG BY CLUSTERS OF GALAXIES). They confirm the large mass-to-light ratios inferred by Zwicky's method.

When X-RAY TELESCOPES became available in the mid-1960s, galaxy clusters were found to be the most powerful x-ray sources in the sky. The emission is extended over the entire cluster and thus reveals the presence of large amounts of hot gas ($T = 10^7$–10^8 K) where x-rays are produced by electron bremsstrahlung. The mass in this 'x-ray gas' is typically 10–20% of the total, i.e. clusters contain more mass in hot gas than in stars. Detailed studies reveal a cluster baryon fraction of typically $f_{\rm B}h^{3/2}$ = 0.03–0.08.

Large-scale flows

On large scales the motion of galaxies is dominated by the overall cosmic expansion. Still, they exhibit 'peculiar velocities' relative to the overall cosmic flow; our group of galaxies moves at 627±22 km s^{-1} relative to the reference frame defined by the cosmic microwave background radiation. These motions are attributed to the action of gravity over the age of the universe, caused by the inhomogeneities of the matter density. The observed large-scale velocity fields, together with the observed

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

galaxy distributions, are then translated into a measure for the matter density required to explain the large-scale flows. One finds

$$\Omega_M > 0.3$$

and even larger values by related methods which are more model-dependent.

Flat universe

The critical value $\Omega_{tot} = 1$ for the cosmic mass–energy density, corresponding to an overall Euclidean (flat) spatial geometry, is favored to avoid fine-tuning the cosmic initial conditions. In an expanding universe, Ω_{tot} quickly evolves away from 1 towards either 0 or ∞ so that the near-flatness of the present-day universe suggests $\Omega_{tot} = 1$ as an exact identity. Moreover, 'inflationary models' of the early universe generically produce a flat geometry even though one can construct specialized models which circumvent this outcome. The critical density exceeds the luminous mass by about two orders of magnitude so that the universe, if it is flat, contains more than 99% dark matter.

Astrophysical constraints

Big-bang nucleosynthesis

The first question about the nature of dark matter is whether it could not consist of ordinary material in some non-luminous form, perhaps stellar remnants such as neutron stars or molecular hydrogen clouds which are difficult to observe. However, the overall baryon abundance is severely constrained by big-bang NUCLEOSYNTHESIS.

When the universe was about three minutes old, the initial protons and neutrons formed helium at a mass fraction of about 22–25%, together with some traces of deuterium (D or ^{2}H), ^{3}He and ^{7}Li. Within the standard big-bang picture, these primordial light-element abundances depend only on the cosmic baryon density (see BARYOGENESIS).

The abundance most sensitive to $\Omega_B h^2$ is that of deuterium. Its measurement in intergalactic hydrogen clouds has recently become possible by observing quasar absorption lines. While this novel approach holds much promise toward a precision determination of the primordial deuterium abundance, one currently finds both high and low values of $D/H \div 2\text{–}10^{-4}$ and $2\text{–}10^{-5}$, respectively, which are mutually inconsistent unless the baryon distribution is vastly inhomogeneous on large scales. They span a range roughly corresponding to

$$\Omega_B h^2 = 0.005 - 0.024$$

which is consistent with the helium and lithium observations. The more favored low deuterium value corresponds to a high baryon content.

This range for Ω_B is depicted in figure 3 as a function of the Hubble expansion parameter together with the luminous mass density of (1) and the lower dynamical mass limit of (2). The currently favored range for H_0 is 50–80 km s^{-1} Mpc^{-1}, implying a gap between the cosmic baryon density and both the luminous and the dynamical matter density. Apparently there is a significant fraction of 'dark baryons' which never made their way into galaxies and stars, and more importantly, lots of unidentified non-baryonic dark matter. The apparent gap between the baryons and the dynamical mass density corresponds well to the baryon fraction implied by the x-ray gas in galaxy clusters discussed earlier.

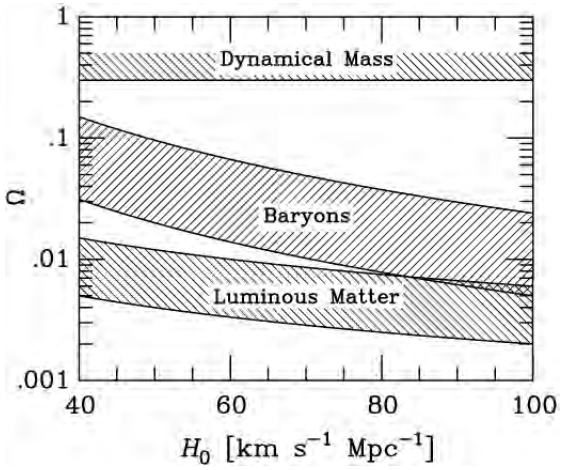


Figure 3. Cosmic matter components as a function of the assumed present-day Hubble expansion parameter.

Structure formation

The standard theory of cosmic structure formation holds that at some early time the universe was almost perfectly homogeneous, apart from tiny density modulations, which were enhanced by the action of gravity as time went on, leading to the formation of galaxies, clusters, and large-scale coherent structures. It is often argued that the primordial density variations came from quantum fluctuations in the very early universe which were boosted to macroscopic scales during a phase of exponential expansion ('INFLATIONARY UNIVERSE').

The amplitude of the density fluctuations at the epoch when the COSMIC MICROWAVE BACKGROUND radiation decoupled from the ambient plasma has now been inferred from several experiments, notably the COBE satellite, which measured the temperature fluctuations across the sky. An amplitude of the fluctuation spectrum consistent with these measurements is too small to allow the observed structures to form if the medium consists only of baryons and radiation. Weakly interacting particles fare better because they are not held up by photon pressure ('dark-matter boost'). This is a generic argument against baryonic dark matter even though there may be possibilities to circumvent it.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The matter which makes up the cosmic 'fluid' can diffuse, wiping out part of the initial density fluctuations. This effect is particularly important for weakly interacting particles which can diffuse far until their momentum has been sufficiently redshifted by the cosmic expansion; they erase the primordial fluctuation spectrum up to a scale which is larger for less massive particles. One speaks of 'hot dark matter' if the fluctuations are wiped out below scales which later correspond to galaxies, while 'cold dark matter' has this effect only on subgalactic scales. The dividing line ('warm dark matter') corresponds to a particle mass in the keV range.

The consensus is now almost universal that some variant of a cold dark matter scenario is probably how our universe works, where structure forms by the gravitational instability mechanism from a nearly scale-invariant spectrum of primordial density fluctuations. It strongly disfavors both baryons and low-mass particles such as massive neutrinos as the main ingredients of the cosmic matter cocktail.

Cosmic microwave background

The cosmic microwave background radiation holds a wealth of cosmological information. Its very presence and its uncannily precise black-body nature are the most striking proofs of the hot big-bang cosmogony. The COBE satellite and more recent ground-based experiments have measured tiny angular temperature variations with typically 10 μK amplitudes which already provide tight constraints on theories of structure formation and thus on the nature of dark matter.

The most important information is contained in the power spectrum of the temperature sky map. While there is still a lot of scatter in the data, they already seem to confirm the appearance of features known as 'acoustic peaks' or 'Doppler peaks'. Their angular scale and amplitude provide invaluable information on the properties of the universe at radiation decoupling.

There are two approved satellite missions, NASA's Microwave Anisotropy Probe (MAP), to be launched in 2001, and ESA's PLANCK, to be launched around 2007, which will take full-sky temperature maps at fine angular resolutions. It is thought that these experiments will ultimately be able to determine the most important cosmological parameters on the 1% level, notably the baryon fraction and total dark matter content. There remain degeneracies between different combinations of parameters, however, which will need to be broken by other methods.

Candidates and searches

Neutrinos

The presence of large amounts of dark matter in the universe is almost uncontroversial, and the case against baryons as a main component is quite compelling, leaving us with particle dark matter (PDM) as one possible solution. The candidate particles would have to be relics from the early universe where they would have been produced thermally in the hot and dense plasma, or by some non-equilibrium process.

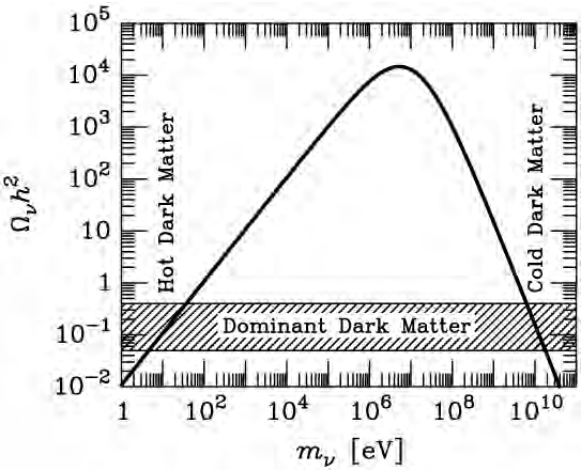


Figure 4. Cosmic mass density as a function of the assumed neutrino mass ('Lee–Weinberg curve'). The hatched band indicates the required range for Ωh^2 of the dominant particle dark matter component.

The minimum required relic density is $\Omega_{\rm PDM} \gtrsim 0.2$ if we use (2) and allow for a significant baryon fraction. The observed age of the universe together with the measured expansion rate yields $\Omega h^2 \gtrsim 0.4$ so that

$$\Omega_{\rm PDM} h^2 = 0.05 - 0.4$$

where $h \gtrsim 0.5$ has been used.

Of the known particles, neutrinos are the only candidates. Their relic density as a function of the assumed mass is shown in figure 4. There are crudely as many background neutrinos in the universe as there are microwave photons. Multiplying this number by m_ν explains the increasing branch of figure 4 and gives $\Omega_\nu h^2 = \sum m_\nu / 93$ eV, whence the required neutrino mass is 4–40 eV. There is a second solution for a neutrino mass of a few GeV on the decreasing branch; the number of relic neutrinos is reduced by annihilations for $m_\nu \gtrsim 1$ MeV, where they are still in thermal equilibrium when the cosmic temperature falls below their mass.

The laboratory limit of about 20 MeV for ν_τ, and much less for ν_μ and ν_e, precludes the large-mass solution among the known neutrino families. The low-mass solution, however, is problematic from the perspective of structure formation because they represent 'hot dark matter'. In addition, there is a problem with neutrinos filling the dark-matter halos of galaxies as they cannot be arbitrarily densely packed in phase space. This 'Tremaine–Gunn argument' requires $m_\nu \gtrsim 20$ eV for typical spiral galaxies, and a few 100 eV for dwarf galaxies so that neutrinos could not be the dark matter on all scales.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Finally, the positive evidence for neutrino masses from recent oscillation experiments points to sub-eV mass differences between the three flavours so that their masses would have to be much larger than their differences ('degenerate masses') if they were to be dark matter. The largest cosmologically allowed mass of about 40 eV then distributes itself among the three families so that each of them must be lighter than about 13 eV, exacerbating the Tremaine–Gunn problem.

Weakly interacting massive particles (WIMPs)

Neutrinos on the decreasing branch in figure 4 would play the role of cold dark matter and are thus favored by cosmology, but in this mass range they do not exist. One thus postulates novel weakly interacting massive particles (WIMPs; see WIMPS AND MACHOS) to fill the gap. On the other hand, from a theoretical perspective one postulates a 'SUPERSYMMETRY' which predicts a new partner to every known particle. The lightest supersymmetric particle in the form of a 'neutralino' could well play the WIMP role. Independently of the dark-matter problem, the search for supersymmetry is one of the main goals of the most ambitious particle accelerator ever, the Large Hadron Collider (LHC) which is currently under construction in Geneva at the CERN Laboratory.

If the dark matter consists of some form of WIMPs, then our Milky Way galaxy should be filled with a 'gas' of these almost collisionless particles which would permeate everything, including our laboratories. The experimental search for galactic WIMPs has turned into an entire branch of non-accelerator particle physics, with numerous groups mounting more and more sensitive detectors. Usually one tries to measure the tiny energy depositions caused by a rare collision of a galactic WIMP with a nucleus in a crystal. In semiconductor crystals, especially germanium, one searches for the resulting electronic signal, in sodium iodide crystals for the scintillation light, or one may cool the crystal to very low temperatures and search for the minuscule heating caused by WIMP collisions; an example for such a cryogenic setup is shown in figure 5.

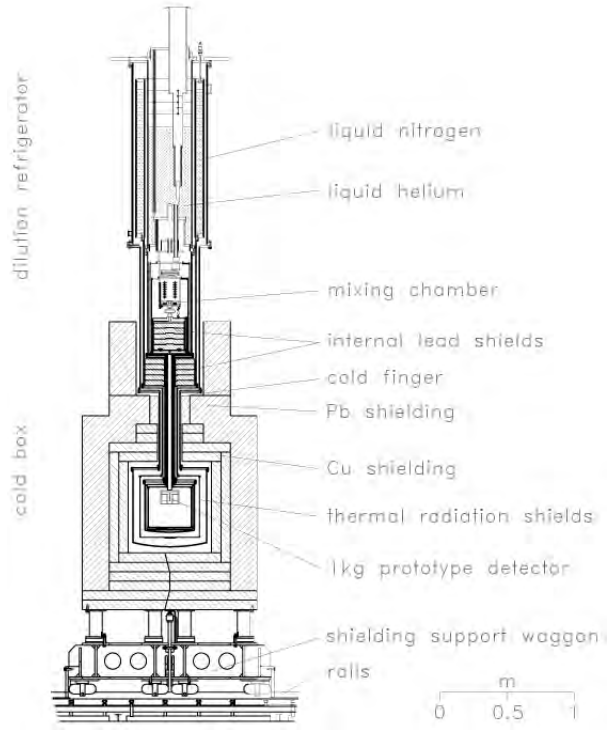


Figure 5. Schematic view of the experimental setup of CRESST, located in the Gran Sasso underground laboratory near Rome (Italy), as an example for a cryogenic dark-matter experiment.

The main experimental problem is the extremely small expected signal rate. In detail, it depends on the assumed WIMP properties and target material, but a typical number is below 1 event kg^{-1} d^{-1}. To reduce natural radioactive contaminations one must use extremely pure substances, and the cosmic-ray background requires locations deeply underground, for example in deep mines.

There exist other 'indirect' search methods. For example, WIMPs traversing the Sun or Earth will occasionally collide with a nucleus there, lose enough energy to get trapped, and ultimately annihilate with other trapped WIMPs. The resulting high-energy neutrinos from the Sun and the center of the Earth may well show up in existing and future neutrino telescopes.

While supersymmetric models are quite flexible in their predictions, the current round of experiments has reached the sensitivity where one can begin hoping to find these elusive dark-matter candidates.

Axions

A cold dark matter candidate *sui generis* is provided by axions, low-mass bosons which are postulated in the framework of quantum chromodynamics, the theory of the strong interaction among quarks. Other than neutrinos and WIMPs, these particles would arise from a non-thermal process in the early universe, producing essentially a Bose condensate, i.e. highly occupied, quasi-classical oscillations of the axion field. The axion mass would lie in the 10^{-5} eV regime, corresponding to GHz oscillation frequencies of the axion field.

Axions would have tiny electromagnetic interactions, implying that galactic dark-matter axions would drive a microwave cavity, provided it is placed in a strong magnetic field. Two such experiments are now in operation, one in Livermore, California (US Axion Search), the other in Kyoto, Japan (CARRACK). Their sensitivity is enough to detect the feeble output of a microwave cavity, induced by galactic dark-matter axions. Again, within the next few years one has a realistic chance to detect dark matter if the underlying hypothesis of the existence of axions is correct.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Dim stars (MACHOs)

As long as particle dark matter has not been found, one may continue to speculate that the arguments against baryons are wrong and that galactic halos consist of ordinary matter in some non-luminous form. Obvious candidates are dim stars which have been termed massive astrophysical compact halo objects (MACHOs) as a deliberate pun on the WIMP hypothesis.

The physical nature of MACHOs could be stars which are too small to shine brightly (brown dwarfs or M dwarfs) or burnt-out stellar remnants (WHITE DWARFS, NEUTRON STARS, BLACK HOLES). Of course, stellar remnants seem implausible, among other reasons because they would have to arise from a large population of normal stars, of which there is no trace in the halo.

A practical search method involves the gravitational light deflection caused by a galactic MACHO which happens to pass near the line of sight to a background star. While the effect is too weak to produce several resolved images of the target star, its apparent brightness will temporarily increase because the MACHO's gravitational field effectively focuses more light into the telescope ('gravitational microlensing'). Such events are rare so that one must monitor millions of stars simultaneously to obtain a reasonable event rate. The Large and Small Magellanic Clouds as target regions have enough bright stars at a suitable distance and have been used since the early 1990s by the MACHO and EROS Collaborations in an attempt to find non-luminous stars making up the Milky Way halo.

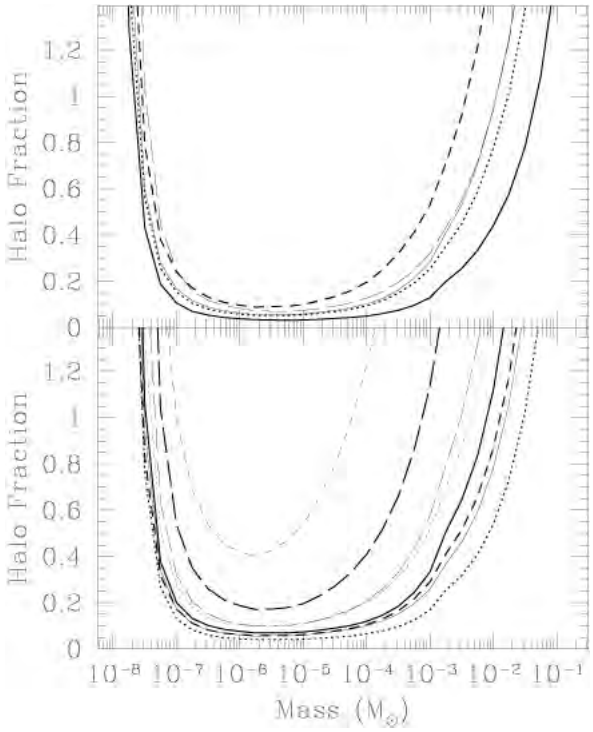


Figure 6. Upper limit to the Milky Way halo fraction that can be attributed to MACHOs as a function of their assumed mass from a combined analysis of the MACHO and EROS data. The different line styles correspond to different halo models; the upper panel shows those used by the EROS Collaboration, while the lower panel the ones of the MACHO Collaboration. In each case the solid curve represents a 'standard halo'. (C Alcock *et al Astrophys. J.* **499** L9 (1998).)

The duration of the temporary brightness excursion of a lensed star depends on the relative distances between observer, lens and target star, and on the lens mass. The non-observation of short-time events has allowed these experiments to exclude MACHOs over a wide mass range as a dominant halo component (figure 6), but does not yet exclude brown dwarfs as a possibility.

On the other hand, about a dozen positive events have turned up which, if attributed to MACHOs, would point to a mass around 0.5 $M_\odot$. However, the early excitement about the apparent discovery of some or all of the galactic dark matter has given way to a more sceptical assessment—the apparent mass range of the observed events simply does not seem to make sense. Perhaps the least troubling interpretation is that one is not seeing MACHOs but normal stars as lenses, which is possible if there is an unrecognized population of stars between us and the Large Magellanic Cloud, or even by stars within the Magellanic Clouds themselves if their distribution is different from what had been thought. Thus, while the observed MICROLENSING events are no doubt real, the question of where and what the lenses are remains for now wide open. (See also DARK MATTER IN GALAXIES.)

Primordial black holes

While stellar remnants seem implausible, black holes which formed in the early universe cannot be excluded as dark-matter candidates. From a structure-formation perspective, they represent cold dark matter; from a search perspective, they could be MACHOs. The main objection is the lack of a plausible production mechanism in the early universe. Surely, as long as particle dark matter remains undiscovered, and as long as the observed microlensing events remain unexplained, primordial black holes should not be brushed aside.

Cosmological constant (vacuum energy)

It is possible, and indeed implied by particle physics theory, that gravitating energy is associated with the ground state of quantum fields, i.e. with the vacuum. In the equations of general relativity it would play the role of a 'COSMOLOGICAL CONSTANT', an optional term which is permitted by the symmetries of the equations, but which is not motivated by the Newtonian limit. Vacuum energy would form a perfectly homogeneous background and thus cannot explain galactic rotation curves or dark matter in galaxy clusters. On the other hand, it could well explain the gap between dynamical measures of dark

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

matter and the critical density (figure 3). We could have a Euclidean universe even if 'real matter' contributes only a sub-critical mass density.

The most counterintuitive property of vacuum energy is that it is not diluted by the cosmic expansion but rather becomes more and more dominant relative to normal matter as time goes on. It cannot be measured locally; a modification of the cosmological redshift–distance relationship is the main observable consequence.

Type Ia SUPERNOVAE as standard candles at cosmological distances have recently provided some exciting evidence for a cosmological constant. While being far from definitive, it suggests that the universe could be Euclidean with something like 70% of the critical density in vacuum energy and 30% in matter. This scenario accommodates effortlessly the absolute and relative amounts of baryonic and non-baryonic matter implied by big-bang nucleosynthesis, the x-ray gas in galaxy clusters, dynamical indicators for dark matter, and a flat geometry of the universe.

Modified gravity

The hypothesis of particle dark matter requires non-trivial and perhaps bewildering extensions of the standard model of particle physics. As long as the nature of dark matter has not been positively identified it may seem no more radical to modify general relativity such that there is no need for dark matter. It has sometimes been argued that the hypothesis of dark matter is just a parameterization of our ignorance of the physical laws which apply on large astrophysical scales where no independent tests of the validity of general relativity exist.

In one phenomenological approach known as modified Newtonian dynamics (MOND), gravitational accelerations a below a certain limit a_0 are given by $a^2/a_0 = G_N M/r^2$ (Newton's constant G_N). With $a_0 \approx 10^{-10}$ m s^{-2} this approach is surprisingly successful at explaining a broad range of dark-matter phenomena related to dwarf galaxies, spiral galaxies and galaxy clusters. Unfortunately, MOND lacks a relativistic formulation so that it cannot be applied in a truly cosmological sense.

Before modifications of general relativity can be taken seriously they must pass relativistic tests. An important case are galaxy clusters where large amounts of dark matter are indicated by non-relativistic methods *à la* Zwicky (virial theorem) as well as by relativistic indicators (gravitational lensing). As virial and lensing masses seem to agree well in several cases, scalar–tensor extensions of general relativity are in big trouble, if not ruled out entirely.

No serious attempt has been made to discuss truly cosmological phenomena such as structure formation and microwave background temperature fluctuations in the framework of alternative theories of gravity. At present no covariant theory of gravity is known that can explain the dark-matter problems on all scales.

Conclusion

Over the past decade, the idea has become almost commonplace that most of the stuff in the universe consists of non-baryonic matter, perhaps in the form of neutralinos or axions which are motivated in particle physics for independent reasons. Yet this remains a radical conjecture which has often been likened to the Copernican revolution when Earth and with it Man was moved from the center of creation to some unspectacular average position. Probably the next big step in this second Copernican revolution will be precise measurements of the angular temperature fluctuations of the cosmic microwave background, which are expected to confirm or refute the apparent discrepancy between the baryon content of the universe and its dynamical mass density. Even then, however, the second revolution will not be complete without a direct and positive identification of the dark-matter particles or objects. Until that happens, perhaps by one of the ongoing experimental searches, one should keep an open mind. The true solution of the dark-matter problem may not have been thought up yet.

Web update (31 July 2002)

A team of 27 astronomers led by Professor George Efstathiou of the University of Cambridge has published strong evidence for the existence of dark energy using the clustering pattern of 250 000 galaxies in a large volume of the universe surveyed with the Anglo–Australian Telescope at Siding Spring in New South Wales, Australia. By comparing the structure in the universe now, some 15 billion years after the Big Bang, with structure observed in the cosmic microwave background radiation, which preserved information about what the universe was like when it was only 300 000 years old, the Anglo–Australian team could apply a simple geometrical test to elucidate the composition of the universe. Their results show that the universe is full of vacuum energy, completely consistent with the earlier supernovae results.

Christopher Kochanek of the Harvard–Smithsonian Centre for Astrophysics in Cambridge, Massachusetts and Neal Dalal of the University of California, San Diego have used radio telescopes and gravitational lensing to search for cold dark matter. They have studied seven galaxies, each magnified by four nearer ones. Because each lensing galaxy is in a slightly different position, the researchers got four different images of each of the seven distant galaxies. The four images should have been identical. But each is actually slightly different. The difference was enough to have been caused by the kind of clumps of dark matter around lensing galaxies that mathematical models predict.

Web Update references

Efstathiou G *et al* February 2002 Monthly Notices of the Royal Astronomical Society **330, No. 2**

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Dalal N and Kochanek C S 2002 Direct detection of CDM substructure *Astrophysical Journal* (in press)

Bibliography

Börner G 1993 *The Early Universe* 3rd edn (New York: Springer)

Carr B 1994 Baryonic dark matter *Annu. Rev. Astron. Astrophys.* **32** 531–90

Kolb E W and Turner M S 1990 *The Early Universe* (Redwood City, CA: Addison-Wesley)

Tremaine S 1992 The dynamical evidence for dark matter *Phys. Today* February, pp 28–36

Trimble V 1987 Existence and nature of dark matter in the universe *Annu. Rev. Astron. Astrophys.* **25** 425–72

Tyson A 1992 Mapping dark matter with gravitational lenses *Phys. Today* June, pp 24–32

Georg G Raffelt

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Dark Nebula

An interstellar cloud of gas and dust which absorbs light from nearby sources, also known as an *absorption nebula*; the light is re-emitted as infrared radiation or scattered, making the nebula appear dark. At higher galactic latitudes, away from the Milky Way, dark nebulae appear merely as particularly star-poor regions of the sky, but near or against the Milky Way they show up strongly in contrast with the bright star fields around them. Prominent dark nebulae seen in silhouette against the Milky Way are the Cygnus Rift in the northern hemisphere and the Coalsack in the southern. A small but strikingly distinctive dark nebula is the famous Horsehead nebula in Orion. The smallest are known as globules. Dark nebulae consist predominantly of molecular hydrogen and are believed to be sites of star formation. The true nature of dark nebulae—that they are not simply voids—was first recognized by E E Barnard, who photographed and cataloged them; the results were published in 1927 in *A Photographic Atlas of Selected Regions in the Milky Way*. A more comprehensive catolog of dark nebulae by Beverly Lynds appeared in 1962.

See also: Coalsack, Cygnus, globule, Horsehead nebula, interstellar molecular clouds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Darwin, George Howard (1845–1912)

Mathematician and astronomer, born in Downe, Kent, England, a son of Charles Darwin, became Plumian professor of astronomy and experimental philosophy at Cambridge University. He applied mathematical theory to the dynamics of the Sun–Earth–Moon system. He added physical effects into gravitational theory, including the frictional effects of tidal action and the shapes which the Earth and Moon take up as they rotate. From his work, he proposed a theory of the origin of the Moon, namely that it was pulled from a molten Earth early in its history by tidal action of the Sun, the first time that the origin of a planet had been explained from its current dynamical state.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Davenport, William (1782–1823)

Irish astronomer, born in Dublin, became the second director of the Armagh Observatory. During eight fruitless years of idleness and decline he contributed practically nothing to astronomy until he committed suicide with a shotgun in the study of the observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

David Dunlap Observatory

The David Dunlap Observatory (DDO) is located 25 km north of the center of Toronto, Canada, in the town of Richmond Hill at 79d 25′ 20″W and +43d 51″ 46″N. Operated as a research and educational facility of the Department of Astronomy at the University of Toronto, DDO is the site of a 1.88 m reflecting telescope, the largest in Canada. Two smaller telescopes plus computer, library and support facilities are available for the use of astronomers. The observatory was made possible with a generous donation by the Dunlap family in June 1928 through the efforts of the first director Clarence A Chant. First light for the 1.88 m telescope occurred on the night of 31 May 1935.

The telescope has remained in productive use to the present day contributing to research in several areas including: determining the radial velocities of stars in the solar neighborhood; pioneering work on Cygnus X-1, the first detected black hole candidate; cataloging of globular clusters; analyses of several classes of variable stars, especially Cepheid, RR Lyr-type and close-binary stars; and studies of the galactic interstellar medium. The 1.88 m telescope has been upgraded with computer controls and modern instrumentation and is dedicated to spectroscopic studies of mainly galactic objects. A variety of public programs, courses and school visits are also offered at DDO.

For further information see
http://ddo.astro.utoronto.ca/ddohome.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dawes, William Rutter (1799–1868)

English clergyman and keen-eyed amateur astronomer. Observed double stars, discovered Saturn's Crêpe Ring independently of WILLIAM BOND. Gave a useful empirical formula for the resolving power of a telescope known as *Dawes limit*: for a telescope of aperture d cm a double star of separation $11/d$ arcseconds or more can be resolved.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dawes' Limit

An empirical measure of the resolving power of a telescope that was devised by William Rutter Dawes (1799–1868), a keen-eyed observer of double stars. According to the Dawes' criterion, the resolving power of a telescope is given by $R = 0.115/D$, where R denotes resolving power expressed in seconds of arc and D denotes the aperture of the telescope (in meters). For example, the Dawes' limit for a telescope of 0.1 m (100 mm) aperture would be $0.115/0.1 = 1.15$ arcsec. The Dawes' limit gives resolving powers that are some 20% better that the theoretical values that are given by the Rayleigh limit. This reflects the fact that skilled observers, under ideal conditions, may be able to resolve double stars that are marginally closer together than the theoretical limit, a limit that is defined in a convenient, but to some extent arbitrary, fashion.

See also: aperture, Rayleigh limit, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Day

Traditionally, the time taken for the Earth to rotate once around its axis, but conventionally taken to be 86 400 SI seconds unless otherwise specified. There are several different definitions of the day. The sidereal day is the time interval between two successive upper transits of the vernal equinox; it is, in effect, the time taken for the Earth to rotate on its axis through an angle of 360° with respect to the background stars. The apparent solar day is the time interval between two successive noons (upper transits of the Sun across an observer's meridian). Because (a) the Earth moves round the Sun in an elliptical orbit at a variable rate, and (b) the apparent annual path of the Sun relative to the stars (the ecliptic) is tilted at an angle to the celestial equator, the right ascension of the Sun changes at a non-uniform rate and the time interval between successive noons (and hence the duration of the apparent solar day) is not precisely constant. The mean solar day is the time interval between two successive upper transits of a hypothetical object called the mean sun, which moves along the celestial equator, relative to the background stars, at a uniform rate (equal to the average angular rate at which the real Sun appears to move along the ecliptic). On the assumption that the Earth rotates on its axis at a uniform rate (which is not strictly accurate), mean solar days are of equal duration.

Because the Earth revolves round the Sun at an average angular rate of just under 1° per day, the Earth has to rotate through an angle of nearly 361° between successive upper transits of the Sun across an observer's meridian. Consequently, the mean solar day is just under 4 min longer than the sidereal day and the sidereal day is equal to 23 h 56 min 04 s of mean solar time.

See also: apparent solar time, diurnal motion, International Atomic Time, mean solar time, sidereal time, Universal Time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

DB Pulsating Stars

Pulsating WHITE DWARF stars with nearly pure helium atmospheres exist in a narrow temperature range near 25 000 K. At this writing, there are eight of these stars known. Collectively, they go by two names: the V777 Her stars, named after the VARIABLE STAR designation of the first known star in the class, and the more informative name, the DBV stars. The name DBV follows the standard convention for white dwarf stars and tells us that they are of spectral type DB, indicating a nearly pure helium atmosphere, and the V suffix indicates they are variables of that spectroscopic class; this is how we will refer to them.

The STELLAR PULSATIONS in the DBV stars are global in nature; their study provides a window into the deep interiors of these dense stars, and thereby a chance to learn about the physics of matter under extreme conditions of density and temperature inaccessible in terrestrial laboratories.

Stars that begin their lives with total masses below about $10M_{\odot}$, or well over 90% of all stars, end their lives as white dwarf stars. Of these, roughly 30% will become DB white dwarf stars. Since the DBV stars are otherwise normal, anything we learn about their internal structure applies to other DB stars as well. Studies of the DBV stars thereby serve to constrain our understanding of one of the two most significant endpoints for STELLAR EVOLUTION. This is of increasing importance as we are learning to use calculations of the ages of the cool white dwarf stars as chronometers to measure the ages of the various STELLAR POPULATIONS in the galaxy.

The DBV stars are particularly significant to the more than 300 yr old science of pulsating variable stars in that they are the first class of pulsating variable stars whose existence was predicted theoretically prior to their discovery. Recently, the pulsating subdwarf B stars joined the DBV stars in the distinction of being predicted in advance of their discovery. Together they indicate the rapid advances in our understanding of pulsating variable stars.

What we know about the DB pulsators

The DBV stars share a set of common observed properties with the other classes of pulsating white dwarf and pre-white-dwarf stars, the DOV (PG 1159-035, GW Vir) stars, the PNNV stars, and the DAV (ZZ Ceti) stars. The observed pulsations occur on timescales of 100–1000 s and are the result of modulations in the total intensity of the starlight. The amplitudes range from the limits of detectability at or near 1 part in 10 000 up to 30%. Typical amplitudes of the peak-to-peak variations in the light curves are of order 10% and appear to be modulated on timescales of days to years. All objects are multiperiodic with complex light curves. The top panel in figure 1 is a portion of a typical light curve of the DBV prototype, GD 358 (V777 Her).

The similarities in observed properties of all of the pulsating white dwarf and pre-white-dwarf stars are reflected in similarities of the theoretical interpretations. The largest-amplitude frequencies present in the light curves are the normal modes of oscillation of the star. The frequencies of the normal modes are consistent only with the class of non-radial pulsation modes known as g-modes. These modes have gravity as their restoring force and are characterized by fluid motions of the star primarily along gravitational equipotential surfaces. Modes with relatively small radial displacements are energetically more favorable for high-gravity objects such as white dwarf stars. The high gravity combined with relatively slow rotation rates, and the absence of strong magnetic fields, makes the stars almost completely spherical. Just as in all spherical potential problems in physics, such as the familiar quantum mechanical description of the atom, spherical harmonics provide an appropriate representation of the normal modes of oscillation in the white dwarf stars.

In all of the non-radial pulsating stars, similarly to energy levels in the atom, the frequency of each observed mode is associated with three quantized numbers, k, l and

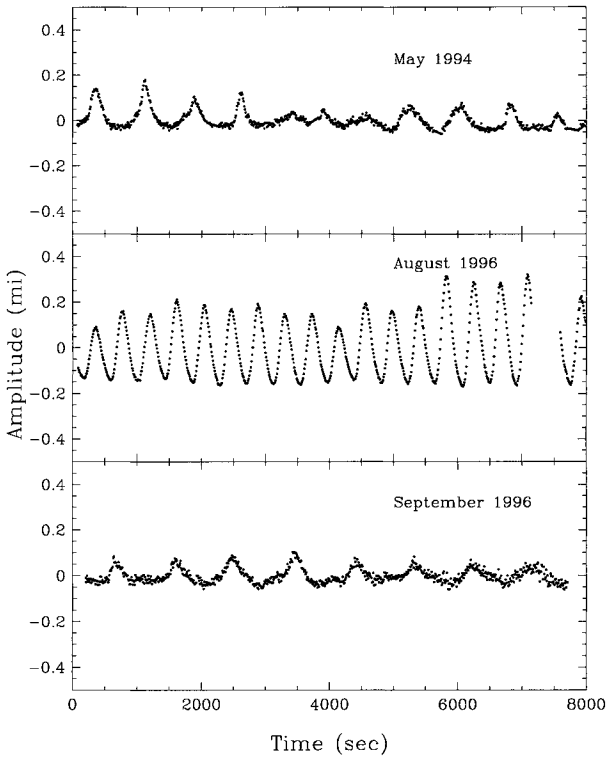


Figure 1. Portions of light curves of the DBV prototype GD 358 (V777 Her). The top panel is a typical light curve from the 17 yr it has been observed. The middle panel is a portion of the light curve during a surprising event when one of the lower-amplitude pulsations completely dominated the light curve with a very large amplitude and a nearly perfectly sinusoidal pulse shape. The lower panel illustrates the return to more normal appearing light curve after this event.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

m:

$$\sigma = \sigma_{k,l,m}$$

The first of the quantized numbers, k, corresponds approximately to the number of radial nodes in the eigenfunction, the second to the degree (sometimes referred to as the order) of the spherical harmonic and the third is the azimuthal quantum number. For a completely spherical system, the frequencies of the modes are degenerate with respect to m. In a real star this degeneracy is likely to be broken by non-zero rotation or possibly a magnetic field.

We can visualize the degree as the number of nodal circles passing around the star. At the photosphere this divides the star into alternating bright and dark regions which give rise to the observed light intensity variations. Because we are observing the whole disk, the alternating regions are averaged over the disk. This causes a geometric cancellation of the bright and dark regions making it difficult, or impossible, to observe intensity variations for $l \geq 5$ with current detectors. We note that spectroscopic techniques of Doppler mapping used in other bright non-radial pulsators are not possible here because the white dwarfs are both fainter and broad lined. This is even more acute for the DBV stars because the helium lines do not have the sharp NLTE core we observe in the hydrogen lines in the DAV stars. See also NON-RADIAL STELLAR PULSATIONS.

Theoretical calculations of the origin of the observed modes suggest that they are self-excited, or amplified from small perturbations always present in the star, and that the physical process responsible for the self-excitation is associated with the partial ionization of the helium in the surface layers. The nature of this excitation and its relation to the convection associated with the partial ionization are difficult problems currently being explored. Our understanding of this process is in its infancy.

Asteroseismology of the DB white dwarf pulsators

The theoretical calculation of the eigenfrequencies for models of the DBV stars, as for the other white dwarf and pre-white-dwarf pulsators, can be carried out to considerable accuracy in the adiabatic approximation. This neglects the change in energy of a mode and is for the most part independent of the detailed structure of the partial ionization zone and any dynamical processes taking place there. Because time is the quantity we can measure most accurately in science, frequencies are also the most reliable of the quantities we measure observationally for the pulsating variable stars.

The accuracy in both the measurement and the calculation of frequencies lends itself to their intercomparison as a way to learn about the internal structure and evolution of the stars. This process is called asteroseismology.

Using the techniques of asteroseismology we have made significant progress in measuring the total mass, surface layer structure, rotation rate and law, magnetic field strength and the absolute luminosity of the DBV stars. This last gives us an important asteroseismological distance measure.

Our expectation, based on the linear theory, is that for each degree, l, we should find a series of multiplets, corresponding to different radial overtones, each with $2l+1$ components. The separation between the different modes of the multiplets is related to the rotation rate of the star and is a weak function of l. This should give us an independent measure of the rotation rate of the star for each l observed. The spacing between the periods of the multiplets of different radial overtone is set, as you might expect from the nature of the restoring force for g-modes, by the buoyant frequency averaged over the star, weighted by the eigenfunctions. This spacing is roughly uniform in period, and the average value provides a sensitive determination of the mass of the star.

Deviations in the spacing of consecutive modes from the average value are caused by the way that the eigenfunctions sample the star. This is strongly influenced by the compositional layering of the star that has resulted from its prior nuclear burning history augmented by chemical diffusion in the strong gravitational field. This stratification produces resonant mode trapping, which leaves its signature in periodic variations of the spacings of the modes from the global average. This pattern in the frequency distribution is called a trapping cycle and its determination makes possible an accurate measure of the surface layer chemical stratification.

Asteroseismological analysis depends on accurate measurements of all of the frequencies present in the star. The more frequencies present the greater the potential scientific return. Fortunately most of the DBV stars are quite rich in observed frequencies; however, this richness exacts a price. In order to disentangle the many closely spaced frequencies it is necessary to observe with as few gaps in the data as possible for a period long enough to resolve the multiplets.

To accomplish this task for pulsating white dwarf stars and related objects, the WHOLE EARTH TELESCOPE (WET) network was established in the late 1980s and is currently in operation. This network uses many optical facilities around the globe to provide essentially continuous coverage of a selected target. Telescope time allocation constraints and the phase of the Moon typically set the total length of these observing runs—new Moon is required because the white dwarf stars are typically quite faint. The concept is to defeat the encroachment of daylight at any one sight: a kind of an anti-British-Empire where we would say 'The Sun never rises on the Whole Earth Telescope'.

To date, the most extensively observed DBV star is the prototype star, GD 358 (V777 Her). Nearly 150 h of continuous WET observations made possible accurate determinations of the total stellar mass, surface He-layer mass, luminosity and the distance to GD 358. The best fitting values were $(0.61 \pm 0.3) M_{\odot}$ for the total stellar mass, -5.70 $(+0.18 - 0.30)$ for the logarithm of the fractional He-layer mass, $\log(L/L_{\odot}) = -1.3$ $(+0.09 - 0.12)$ for the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

luminosity and a distance of 42 ± 3 pc. For GD 358 we also have an observed parallax value of 36 ± 4 pc. For now, we have these impressive results for only one DBV, but results based on observations of more DBV stars with the WET should follow in the next several years.

Mysteries for the future

With all the progress we have made in understanding the frequencies and their distributions in the DBV stars, we have made very little progress in understanding the pulse shapes in the light curves. All but one currently known has a pulse shape with sharp maxima and broad flat minima (see the top and bottom panels of figure 1 for an illustration of the shapes of the typical pulses). These pulses are very asymmetric about the mean light level as well as being non-sinusoidal. Many linear combination frequencies (for the most part first sums and differences with the sums having the highest amplitudes) are often present with much higher amplitude than the actual harmonics of the modes. The cause, if we can find it, holds a clue to the physics of the pulsations and in particular their excitation.

We have also observed changes of an order of magnitude or more in the relative amplitudes of particular modes. Perhaps the most dramatic instance of this came during simultaneous Hubble Space Telescope and ground-base observations of GD 358, when a single mode rose to amplitude of 30% and assumed a completely sinusoidal pulse shape. A sample from this light curve is in the middle panel of figure 1. The timescale on which this occurred was of order a day, and the timescale to return to its more normal multiperiodic light curve was of the same order. A portion of the light curve after the more normal behavior resumed is shown in the bottom panel of figure 1. What is the cause of these dramatic changes? Are they periodic? We simply do not yet know, but the answer should prove to be exciting.

One of the most vexing problems in the DBV stars is the role of convection in determining the pulsation properties. The convective turnover times near the top of the convection zone are estimated to be of order 1 s, very short compared with the pulsation timescales. Yet most theoretical investigations have assumed that the convective flux is frozen in during the pulsations. This difficult problem is currently being studied.

One very interesting prospect is to try to identify frequencies that are very stable in both period and amplitude to try to measure evolutionary period changes by monitoring the rate of change of the phases of the pulsation. If this can be done, as seems possible so far in the DAV stars (which have been studied for nearly twice as long), we can measure the evolutionary timescales in these stars and calibrate the white dwarf cooling calculations necessary for using the white dwarf stars as galactic chronometers.

Perhaps the greatest mystery of the DB white dwarf stars is their origin. How is a DB white dwarf progenitor different from one that will produce a DA white dwarf? This is especially puzzling given that the mean masses of the DA and DB stars seem to be about the same. Further, along the white dwarf cooling track, the He-rich objects disappear around 40 000 K and do not reappear until just above the temperature of the DBV stars near 30 000 K. Understanding of these evolutionary questions will be dramatically improved if we have accurate measures of the surface layer mass for a significant number of DB white dwarfs, as well as the He-layer mass in the DA white dwarf stars.

The tantalizing results for GD 358 illustrate the potential of asteroseismology for measuring the surface He-layer mass, but they also raise a puzzling question: why is the mass so thin? Standard evolutionary theory suggests the logarithm of the fractional layer mass should be closer to -2 or -3. One possible explanation is that the observed value is not a measure of the actual total He mass but the diffusion boundary as the He percolates up from below. It is also possible that this boundary is not the base of the He layer, but rather the boundary between the relatively lighter ^{3}He and the heavier ^{4}He. If this is the case, then the total layer mass may be approximately the value expected from evolutionary theory. The intriguing possibility of this isotopic separation may have other astrophysical consequences as well. Perhaps it is possible that GD 358 is not typical of the other DB stars for some reason. Only extensive WET-style observations of more DBV stars will tell.

Bibliography

Some flavor of the excitement of the prediction and discovery of the DBV stars can be found in

Winget D E, Robinson E L, Nather R E and Fontaine G 1982 Photometric observations of GD 358—DB white dwarfs do pulsate *Astrophys. J.* **262** L11

The DBV stars have been recently reviewed by

Bradley P A 1995 The DBV stars: progress and problems *Balt. Astron.* **4** 311–20

A recent review of the asteroseismology of white dwarf stars can be found in

Winget D E 1998 Asteroseismology of white dwarf stars *J. Phys.: Condens. Matter* **10** 11 247–61

An excellent discussion of the WET as an instrument can be found in

Nather R E, Winget D E, Clemens J C, Hansen C J and Hine B P 1989 The Whole Earth Telescope—a new astronomical instrument *Astrophys. J.* **361** 309

The WET observations of the prototype of the class are presented in

Winget D E *et al* 1994 Whole Earth Telescope observations of the DBV white dwarf GD 358 *Astrophys. J.* **430** 839

D E Winget

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

de Chéseaux, Philippe Loys (1718–51)

Swiss astronomer, observed several clusters and 'nebulous stars', and compiled a catalog of their positions, giving it to Reaumur, who presented it to the Académie des Sciences in 1746, but it was not otherwise published. Eight of the nebulae were original discoveries. De Chéseaux was the first to formulate OLBERS' paradox, applied to the Milky Way. If the stars extended indefinitely in all directions every sight-line would end on a star's surface and the Milky Way should be as bright as the surface of the Sun. De Chéseaux suggested that since this was not so there was some sort of attenuation in space, dimming the more distant stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

de Sitter, Willem (1872–1934)

Cosmologist, born in Sneek, Netherlands, and started a rather traditional career in astronomy, observing with DAVID GILL at the Cape of Good Hope, determining the masses and orbits of the satellites of Jupiter, determining the variation of the rotation of the Earth, and studying the distribution and motions of stars. He became director of the Leiden Observatory, and turned to a range of problems in the theory of relativity. One fundamental tenet of special relativity is the constancy of the velocity of light, and in particular that c. does not depend on the velocity of its source. By analysing the velocity curves of binary star systems, de Sitter proved this. De Sitter created a mathematical theory of general relativity based on four-dimensional space-time and looked at the theory's astronomical consequences. His work inspired ARTHUR STANLEY EDDINGTON'S expedition to measure the gravitational deflection of light rays passing near the Sun during the eclipse of 1919. De Sitter created a general relativistic theory of the Universe, and found solutions that describe an empty Universe, contrary to the principle formulated by ERNST MACH. The solutions that he found described an expanding universe, a theoretical result spectacularly confirmed by EDWIN HUBBLE's observations of the recession of galaxies. De Sitter worked with EINSTEIN on the *Einstein–de Sitter* model of the universe that is the simple basis for current more complex Big Bang models.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deceleration Parameter

A quantity, denoted by the symbol q, that describes the rate at which the expansion of the universe is slowing down. The value of the deceleration parameter at the present epoch in the history of the universe is denoted by q_0.

Different values, or ranges of values, of q_0 correspond to different cosmological models. For example, if q_0 is greater than $\frac{1}{2}$ the universe is closed and will eventually collapse, whereas if q_0 is less than $\frac{1}{2}$, the universe is open, and will expand forever at a finite rate. The 'flat', or Einstein–de Sitter, universe corresponds to $q_0 = \frac{1}{2}$. The case $q_0 = 0$ corresponds to a universe that expands at a constant rate. A negative value of q_0 corresponds to a universe in which the expansion is accelerating.

In principle, it should be possible to determine the value of q_0 observationally. For example, for a population of uniformly luminous sources (for example, a set of identical supernovae within remote galaxies) the relationship between apparent brightness and redshift is dependent on the value of the deceleration parameter. Although measurements of this kind are notoriously difficult to make and to interpret, recent observations tend to favor q_0 values of less than $\frac{1}{2}$ and hence favor open universe models.

See also: cosmological constant, cosmological model, Hubble diagram, redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Declination

The angular distance of a celestial body north or south of the celestial equator. In other words, the angle between the celestial equator and a star, measured in a direction perpendicular to the celestial equator. Declination may take values between 0 and 90°, and is taken to be positive for an object north of the celestial equator, and negative for an object south of the celestial equator. Declination is often abbreviated to dec, or denoted by δ. The position of a star is normally expressed in terms of right ascension and declination.

See also: celestial coordinates, celestial equator, great circle, right ascension.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dee, John (1527–1608)

Navigator, mathematician, traveler, polymath, mystic, charlatan, astrologer, model for Shakespeare's Prospero and King Lear, and court intriguer. Born in London, he became a navigation instructor, applying Euclidean geometry to navigation and building the instruments to do so. He advised expeditions seeking the Northwest passage to the Pacific via Canada. He cast horoscopes for Elizabeth I, receiving the title Royal Astrologer; he claimed to have put a hex on the Spanish Armada and caused the bad weather that wrecked the fleet. He published trigonometric theorems for determining parallax of the new star of 1572. He designed a large-radius astronomicus for THOMAS DIGGES to observe it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deep Impact

NASA Discovery mission scheduled for launch in 2004. Intended to send a 500 kg copper projectile into Comet Tempel 1 in July 2005. A camera and infrared spectrometer on the spacecraft will study the resulting icy debris and pristine interior material.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deep Space (Space Technology)

Series of NASA technology demonstration missions under NASA's New Millennium programme. DEEP SPACE 1 (launched October 1998) carries 12 advanced technologies, including autonomous navigation and ion propulsion. It may be directed to encounter near-Earth asteroid 1992 KD and two comets. Deep Space 2 (launched January 1999) comprises two small surface penetrators, part of the failed Mars Polar Lander mission. Deep Space 3/Space Technology 3 will comprise two spacecraft for testing space interferometry. Launch is scheduled for 2003. Deep Space 4 was cancelled in June 1999.

Breaking news update (30 April 2002)

NASA's Deep Space 1 mission, which successfully tested 12 advanced space technologies and captured excellent images of a comet, came to an end on December 18, 2001. Engineers sent a final command turning off the ion engine, which has used up 90% of its xenon fuel. The spacecraft will remain in orbit around the Sun, operating on its own. Its radio receiver will be left turned on, in case future generations want to contact the spacecraft.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deep Space 1

Deep Space 1 (DS1) was launched in October 1998 as part of NASA's New Millennium Program and completed its primary mission testing ion propulsion and 11 other advanced, high-risk technologies in September 1999. The DS1 spacecraft flew past asteroid (9969) Braille at a distance of 26 km in July 1999. The asteroid is elongated and irregularly shaped, with a longest side estimated at 2.2 km and its shortest side 1 km. The spacecraft's infrared spectrometer confirmed that the small asteroid is similar to Vesta. It may be ejecta from an impact on Vesta or a sibling with common origin.

NASA extended the mission, taking advantage of the ion propulsion and other systems to undertake an encounter with Comet 19P/Borrelly. The comet was discovered in 1904 by Alphonse Borelly and has a 7 year period. The spacecraft's closest approach to Comet Borrelly in September 2001 was at a distance of only 2200 km (1400 mi). The encounter provided high (100 m) resolution pictures of the comet, sending back black-and-white photographs of the inner core of the 10 km long (more than 6 mi long) nucleus. The pictures showed smooth, rolling plains with brighter regions which seem to be the source of dust jets seen in the comet's coma. Rugged land was found at both ends of the comet's nucleus with many high ridges. This rough terrain contains very dark patches that appear to be elevated compared with surrounding areas.

DS1 also measured the types of gases and infrared waves around the comet and how the gases interacted with the solar wind. A very strong interaction occurs between the solar wind and the comet's coma. Near DS1's closest approach to the nucleus, the solar wind picked up charged water molecules from the coma, slowing the wind sharply. The data show that the flow of ions around the comet's rocky, icy nucleus is not centered on the comet's nucleus. Ions in the turbulent flow are heated to about 106 K (2×10^6 ºF).

Breaking news update (30 April 2002)

NASA's Deep Space 1 mission, which successfully tested 12 advanced space technologies and captured excellent images of a comet, came to an end on December 18, 2001. Engineers sent a final command turning off the ion engine, which has used up 90% of its xenon fuel. The spacecraft will remain in orbit around the Sun, operating on its own. Its radio receiver will be left turned on, in case future generations want to contact the spacecraft.

Bibliography

Images of Comet Borelly are available on the web at http://nmp.jpl.nasa.gov/ds1/images.html

Pam Spence

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Degenerate Matter

Highly compressed matter in which the normal atomic structure has broken down and which, because of quantum-mechanical effects, exerts a pressure that is independent of temperature.

At the very high temperatures and pressures that exist inside stars, matter is almost completely ionized and forms a gas of nuclei and electrons (a plasma). When a star has consumed all its nuclear fuel, it shrinks under the action of its own gravity to form, in most cases, a compact white dwarf. As the star shrinks, and its constituent electrons are forced closer together, the volume of space available to each electron rapidly decreases. The Pauli exclusion principle implies that no more than two electrons (electrons with opposite spin) can have the same position and momentum within the limits imposed by the Heisenberg uncertainty principle. As the electrons are pushed closer together, and their separations in position decrease, the differences in their momenta increase, so generating a pressure that, unlike the pressure in a conventional (ideal) gas, depends only on the density of the gas, not on its temperature. Matter in this state is said to be electron-degenerate, and the pressure exerted by the electrons is called electron-degeneracy pressure. Electron-degenerate matter has densities of around 10^6 kg m^{-3}. At very much higher densities, a similar phenomenon, called baryon degeneracy, occurs for protons or neutrons.

Provided its mass is not too great, electron degeneracy pressure prevents a white dwarf from contracting further under its own gravity. If the mass of a collapsed star exceeds about 1.4 solar masses (the Chandrasekhar limit), gravity will overwhelm electron degeneracy and further collapse will ensue. During the ongoing collapse, electrons combine with protons to form neutrons, so producing a neutron star. Because neutrons, too, are subject to the Pauli exclusion principle, at high enough densities (around 4×10^{-17} kg m^{-3}) they form a neutron-degenerate gas that prevents further collapse of the star. The maximum inert mass that can be supported by baryon degeneracy pressure is given by the Oppenheimer–Volkov limit, a limit that is not precisely known but which is thought to be in the region of two to three solar masses. This gives an upper limit for the mass of a neutron star. If a collapsing star exceeds this mass it will form a black hole.

See also: baryons, black hole, electron, uncertainty principle, ion, neutron stars, Pauli exclusion principle, quantum mechanics, quantum theory, white dwarfs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delambre, Jean Baptiste Joseph (1749–1822)

Astronomer, born in Amiens, France. Produced tables of the location of planets and their satellites from observations made at his private observatory. Succeeded Méchain on the project to determine the size of the Earth, measuring the section from Dunkirk to Rodez. The detailed account of this project forms a three-volume book, and became the base of the metric system, defining the meter. Delambre himself made calculations of the orbit of Uranus, and in his later years wrote a monumental *History of Astronomy*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delaunay, Charles Eugene (1816–72)

Celestial mechanician, born in Lusigny-sur-Barse, France, died at sea (near Cherbourg) after just two years as director of the Paris Observatory. Studied lunar theory and worked out formulae that gave the longitude, latitude and parallax of the Moon as infinite series. The work was accurate but impractical as the series converged slowly.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delisle, Joseph-Nicolas (*fl.* eighteenth century)

French astronomer, member of the Académie, hired CHARLES MESSIER as a draftsman, and as a recorder of astronomical observations. Visited ISAAC NEWTON in London in 1724. Calculated the return path for HALLEY's Comet in 1758–9 and published a map which guided the recovery search, showing the predicted path of return calculated at 10 day intervals. Messier, as his observing assistant, was set to observe the appropriate sectors, and succeeded as the first professional astronomer to see Halley's Comet on its return on 21 January 1759. He had, however, been preceded on Christmas night, 1758, by a German farmer named Palitzsch living near Dresden. Delisle retired and Messier succeeded him as Astronomer of the Navy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delphinus

(the Dolphin; abbrev. Del, gen. Delphini; area 189 sq. deg.) A northern constellation which lies between Pegasus and Aquila, and culminates at midnight in late July. It represents either the messenger that the god Poseidon in Greek mythology sent to fetch the sea nymph Amphitrite to be his bride, or the dolphin that was said to have rescued Arion, a semilegendary poet and musician of Lesbos, who had thrown himself into the sea when threatened during a voyage to Corinth. The brightest stars of Delphinus were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small but conspicuous constellation, the brightest stars in Delphinus are β Delphini (Rotanev), a close binary, integrated magnitude 3.6, with components of magnitudes 4.0 and 4.9, separation 0.2″, α Delphini (Sualocin), magnitude 3.8, and ε Delphini, magnitude 4.0. Interesting objects include γ Delphini, another binary, with orange (K1) and bluish-white (A2) components, magnitudes 4.3 and 5.2, separation 9.8″, and NGC 7006, an eleventh-magnitude globular cluster which, at a distance of about 115 light-years, is one of the most remote.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delta Aquarids

A meteor shower that takes place in late July and early August. There are two radiants in the constellation Aquarius, one at dec. -17° coming to maximum on 29 July, and the other, slightly weaker, at dec. -10° and coming to maximum around 6 August. The parent body of this shower is unknown.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Delta Cephei

A variable star discovered in 1784 by John Goodricke (1764–86), a young deaf-mute astronomer from York who two years previously had explained the variability of Algol. The first discovered of an important class of variable stars which are now known as Cepheid variables, δ Cephei and other Cepheids of relatively long period now form the subclass 'classical Cepheids'. Cepheids are characterized by the regularity of their periods and the form of their light-curves, and provide a means of measuring distances by the application of the period–luminosity relation. This is especially valuable in the case of those Cepheids found in other galaxies.

The apparent magnitude of α Cephei varies between 3.48 and 4.37 in a period of 5.36634 days; the light curve showing a rise from minimum to maximum in about 1.8 days, followed by a much slower fade to minimum in 3.6 days. The spectral type is G2Ib. It is in fact a multiple star system, having two close companion stars of sixth and thirteenth magnitude. With a parallax of 0.003", the system is 980 light-years distant and is very luminous (absolute magnitude -3.32).

Observations of classical Cepheids begun in the 1890s revealed that their surface temperatures and radial velocities also vary, exactly in phase with their brightness; this indicates that they are actually expanding and contracting in diameter. This pulsation gives rise to the observed variations in temperature and luminosity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Democritus of Abdera (*c*. 470–*c*. 380 BC)

Philosopher, born in Abdera, Thrace, Greece. Wrote many mathematical works, of which none survive. Suggested that the Milky Way was made up of crowded individual stars. Best known as an atomist philosopher, he claimed that the universe was a mechanical system obeying fixed laws, explaining its origin through atoms moving randomly and colliding to form larger bodies and worlds. In many ways Democritus was the first modern scientist.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deneb

The star α Cygni, 'the Tail of the Swan'. One of the three bright stars which form the asterism of the Summer Triangle, with Altair and Vega. It is a white supergiant of spectral type A2Ia, apparent magnitude 1.25. It is a very distant star (approx. 3230 light-years, parallax 0.001″), and is the type-star of a class of pulsating variables, varying rather erratically over a range of ± 0.04 magnitudes. It is very luminous, its absolute magnitude being about -8.7.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Denning, William Frederick (1848–1931)

Possibly a journalist and certainly an amateur astronomer. Born in Redpost, Somerset, England. The spectacular meteor storm of 1866, and a fireball in 1869, focused his interest on meteor astronomy. In 1877 he demonstrated a steady night by night movement in the Perseid meteor radiant, which proved that meteors came from showers of dust distributed along the path of a comet. Had the distinction to be cited in H G Wells' *The War of the Worlds* (published 1898): 'Then came the night of the first falling star. It was seen early in the morning rushing over Winchester eastward, a line of flame, high in the atmosphere. Hundreds must have seen it, and described it as leaving a greenish streak behind it that glowed for some seconds. Denning, our greatest authority on meteorites stated that the height of its first appearance was about ninety or one hundred miles.' Published the *General Catalogue of the Radiant Points of Meteoric Showers and of Fireballs and Shooting Stars Observed at More Than One Station*, discovered several comets and Nova Cygni 1920.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Density

The amount of mass contained within a unit volume of material. The mean density of an astronomical body, such as a star or a planet, is simply equal to its mass divided by its volume. In practice, however, the density of such a body increases towards the center. Thus, for example, the density of the surface rocks on the Earth is about half the mean density, while the central density is about 2.5 times the mean value. Astronomical bodies exhibit a wide range of densities. Examples are given below, taking the density of water (1000 kg per cubic meter) as unity.

Object	Mean density (water = 1)	Mean density (kg m^{-3})
Neutron star (typical)	4×10^{14}	4×10^{17}
White dwarf (typical)	10^6	10^9
Earth	5.5	5.5×10^3
Jupiter	1.3	1.3×10^3
Sun	1.4	1.4×10^3
Red giant (typical)	10^{-4}	10^{-1}
Red supergiant (typical)	10^{-7}	10^{-4}
Interstellar gas cloud	10^{-22}	10^{-19}

See also: atmospheric pressure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Descartes, René (1596–1650)

Mathematician and philosopher, born in La Haye (now Descartes), Touraine, France, settled in Holland. His work, *La Géométrie*, formulated geometry in terms of algebra, from which comes the concept of *Cartesian coordinates*. Studied Aristotelian philosophy and was attracted to mathematics, and the purely logical analysis of practically everything. Wrote *Discours de la Méthode pour bien Conduire sa Raison et Chercher la Vérité dans les Sciences*, expressing optics, the weather and geometry in terms of mathematics. Wrote *Principia Philosophiae* and attempted to put the whole universe on a mathematical basis. Since he did not believe in action at a distance, he assumes that the universe is filled with matter that constitutes a system of vortices which carry the Sun, the stars, the planets and comets in their paths, which for unexplained reasons are ellipses. Despite the problems with the vortex theory it was championed in France for nearly 100 even after NEWTON showed it was impossible as a dynamical system, and it remains a boast in France about one's logical prowess to call oneself a Cartesian.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Deslandres, Henri Alexandre (1853–1948)

French astronomer, became director of the Paris and Meudon Observatories. Spectroscopist (molecular spectra). Measured radial velocities and rotations of planets and stars. Independently of George Hale, he invented the spectroheliograph and studied the solar chromosphere and solar activity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Detection of Faint Objects

In order to detect a faint object against the sky background, the object must have a signal level in contiguous pixels that is above a set threshold. Thus, the detection limit is set by a limiting isophotal threshold and a minimum object area. The area is typically that of the SEEING disk, i.e. the area of a stellar image. The limiting isophote is typically 2 times the background fluctuations of single pixels. These parameters can be optimized by attempting to detect negative images, which will be entirely due to noise, and then adjusting the minimum area and limiting isophote until the number of noise images is reduced to an acceptable number.

Sky background

The determination of the sky background relies on the assumption that most of pixels in the image are measurements of the sky level, which varies only slowly across the image. The usual procedure is to subdivide the image into tiles, determine the value of the sky background in each tile and then smoothly interpolate the background value at each pixel in the image. The limiting isophote is referenced to this smooth background.

The size chosen for the tiles is a compromise between using small tiles to follow the sky variations and using large tiles to ensure that most of the pixels in any tile are sky, not image, pixels. The sky intensity assigned to each tile should be the intensity value which occurs most frequently in that tile.

Taking the average value of all the pixels in the tile would bias the background too high because of the high values of image pixels that would be included in the average. A much better estimate of the sky background than the average value is the median value obtained by sorting the pixels by intensity values and taking the intensity value from the pixel half way between the first and last after sorting. Typically the median value is used, although more computationally expensive methods for estimating the modal value, the intensity value which occurs most frequently, are sometimes used. These involve smoothing a histogram of pixel values and taking the intensity value corresponding to the peak of the smoothed histogram. The smoothing should be done by convolving the histogram with a Gaussian that has a width similar to that of the histogram. Using the peak of the unsmoothed histogram is too unstable. The histogram must resolve the distribution of intensities around the histogram peak.

Signal-to-noise ratio

The accuracy of the flux measurement of an individual detection is given quantitatively by the signal-to-noise ratio, R. The fractional error is $1/R$. Let S be the detected PHOTON rate in counts per second from the object. The signal, s, is the number of photons detected in the integration time, t; thus $s = St$. The noise, n, is the fluctuation in the number of detected photons, estimated from Poisson statistics as $n = s^{1/2}$. Thus $R = s/n = (St)^{1/2}$ for the case where the only noise source is the signal itself.

Modern astronomical images are obtained with CCD CAMERAS which have readout noise, N, equivalent to a noise signal N^2, associated with each pixel. Additional noise sources are the noise from the sky background in counts s^{-1}/pixel, B, and the dark current in counts s^{-1}/pixel, D. If a is the number of pixels covered by the image, the total noise in the image is that from the sum of the source, the background, the dark current and readout noise,

$$n = (St + Bta + Dta + N^2a)^{1/2} \tag{1}$$

and the signal-to-noise ratio is given by

$$R = \frac{St}{n} = \left[\frac{St}{1 + a(B + D)/S + N^2a/St}\right]^{1/2}. \tag{2}$$

Thus, the integration time, t, required to obtain a given signal-to-noise ratio, R, is

$$t = \frac{R^2}{2S}\left\{1 + a\frac{B + D}{S} + \left[\left(1 + a\frac{B + D}{S}\right)^2 + \frac{4N^2a^2}{R^2}\right]^{1/2}\right\}. \tag{3}$$

For example, consider the integration time required to detect a 22.0 magnitude galaxy in the V band, at a signal-to-noise ratio of 10, with a 48 in telescope using a thermoelectrically cooled back-illuminated CCD detector when the Moon is 7 days old. Assume that the pixel size is 1.1 arcsec × 1.1 arcsec, the galaxy image will be spread over 4 pixels and the cooler operates at −35 °C.

Table 1. Sky brightness.

Lunar age (days)	U	B	V	R	I
	(mag arcsec^{-2})				
0	22.0	22.7	21.8	20.9	19.9
3	21.5	22.4	21.7	20.8	19.9
7	19.9	21.6	21.4	20.6	19.7
10	18.5	20.7	20.7	20.3	19.5
14	17.0	19.5	20.0	19.9	19.2

Table 1 gives approximate sky brightness in various wavelength bands as a function of the age of the Moon. The sky background in the V band from a 7 day old Moon has a surface brightness of about 21.4 magnitudes arcsec^{-2}. The area of a galaxy covering 4 pixels is $a = 4 \times 1.1 \times 1.1 = 4.84$ arcsec^2. Table 2 gives the response of the CCD in terms of detected photons per second from a 20.0 magnitude star observed with a 1 m (39.4 in) telescope and the readout noise per pixel. From table 2, we have $N = 7$ counts/pixel. For a 48 in telescope, the number of counts per second in table 2 must be increased by a factor of $(48/39.4)^2 = 1.48$. The signal from a $V = 22.0$ galaxy is given by $S = 1.48 \times 23 \times 10^{-0.4(22.0-20.0)} = 5.39$ counts s^{-1}, and the sky background is given by $aB = 4.84 \times 1.48 \times 23 \times 10^{-0.4(21.4-20.0)} = 45.8$ counts s^{-1}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. CCD response.

Read noise (counts/pixel)	U	B	V	R	I
		(counts s^{-1})			
7	3.5	20	23	26	12

Table 3 gives the number of dark counts per pixel per second as a function of temperature for a thinned CCD operated in MPP mode, typically devices with 46.0 pA cm^{-2} at 293 K. For a cooler operated at −35 °C, and a galaxy covering 4 pixels, $aD = 4 \times 2.14 = 8.56$ counts s^{-1}.

Table 3. Dark noise.

Temperature (°C)	Dark (counts s^{-1}/pixel)
−30	4.17
−35	2.14
−40	1.18
−45	0.58
−50	0.27
−55	0.13

Substituting the values into equation (3) gives $t = 209$ s to obtain a signal-to-noise ratio of $R = 10$ for a $V = 22.0$ magnitude galaxy.

Bibliography

A computer program for detecting and cataloging faint galaxies is described by

Bertin E and Arnouts S 1996 *Astron. Astrophys.* **117** 393

The software may be downloaded via anonymous ftp from the URL ftp://ftp.iap.fr/pub/from_users/bertin/sextractor/

Bruce A Peterson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Detectors: Photography

Photography has had a long and productive partnership with astronomy that is now drawing to a close. Although practical photography appeared in 1837, and solar photography was well established in the 1860s, it was to be almost 50 yr before the technology had advanced enough to be seriously applied to imaging the night sky. After the first demonstrations of its value as a detector of the unseen in 1882–3, by 1890 it had largely displaced the human eye as the main astronomical detector, certainly for objects beyond the solar system. This advance made possible the new science of astrophysics. By 1900 photography had revolutionized the way astronomy was done and within 30 yr had completely changed the perception of our place in the universe. It also profoundly influenced the design of astronomical telescopes and instruments.

It was not until the late 1870s that the long exposures necessary for deep astronomical photography became possible. Before that, photographic emulsions were either too insensitive to be useful or (in the case of wet collodion emulsions) dried out and became insensitive during the necessarily long exposure. The introduction of the 'dry gelatine' process changed that and, for the first time, photographic materials could be prepared well in advance of exposure. Processing could be similarly delayed, and the invention of the dry plate spawned an industry that removed the mystique and practical difficulty from the photographic process, making it widely available. It was in this context that it was found that long-exposure photography could reveal stars that were invisible to the human eye when applied to the same telescope. By 1890, photography was revealing an unseen universe of faint stars, nebulae and galaxies that allowed the burgeoning science of astrophysics to flourish.

Photography as an astronomical detector

In astronomy, photography combines the quite different roles of detector and recording medium. As a detector it remains useful because of its almost unlimited sensitive area, high resolution and signal to noise, excellent uniformity, ready availability and low cost. It is also a very efficient, stable and compact recording medium, with the advantage that the images can be examined easily with a simple light table. However, the detector is non-linear recorder, nor are the analogue records are easily searched, rendered quantitative or duplicated, and it is not easy to make the detector itself into a quantitative tool. Although now largely displaced by electronic image detection, photography still has an important role in modern astronomy, especially in survey and patrol work and in the detection of faint light from extended objects.

The light-sensitive layer of all common photographic materials is a thin coating of gelatine on a suitable substrate, usually glass or transparent film in the case of negatives, paper or plastic for prints. Dispersed throughout the gelatine are minute crystals of silver halides, usually silver bromide or bromo-iodide for the emulsion types used in astronomy. These may be treated with dyestuffs during manufacture to extend the native blue–UV sensitivity of the silver salts to longer wavelengths or to enhance sensitivity to specific parts of the spectrum, such as the 656 nm hydrogen alpha line. Emulsions with useful (although low) sensitivity to wavelengths as long as 1.2 μm have been made.

Astronomical photography is a very challenging situation for the photochemist. Because light levels are usually very low, the emulsion must be highly sensitive. However, if good sensitivity is achieved, the ever-present glow of the night sky masks the signals from faint celestial sources. The primary aim in the early days of astronomical photography was to have a detector as sensitive to faint light as possible, in other words, with a high quantum efficiency.

This property was of interest to the commercial manufacturer as well as the astronomer, and astronomers were fortunate that C E Kenneth Mees, founding director of the Eastman Kodak research laboratories, had an interest in astronomy. This led to the introduction of the extremely sensitive Kodak 'Spectroscopic Plates' in the 1930s. These were materials designed to have high sensitivity to faint light, where exposure times were necessarily long. They were used for recording astronomical spectra as well as the direct imaging of faint light. Although it was eventually realized that these tasks were often quite different image recording problems, these fast, coarse-grained, low-resolution products were widely used in the astronomical community until about 1990.

Although low resolution and high granularity were not too important with long-focus telescopes with large plate scales, it was not the ideal combination for recording faint, finely structured spectra in the absence of non-image background light, nor for recording any kind of image where the object of interest was fainter than natural night sky airglow. The introduction of fast, Schmidt-type telescopes in the 1930s, with their finely detailed images and small image scales, served to emphasize the poor imaging properties of the early materials. Eventually (in the late 1960s) the fast, grainy materials were displaced by new products designed for faint object detection, not high speed. It was gradually realized that the important factor in image detection was detective quantum efficiency (DQE), rather than simple quantum efficiency:

$$\mathrm{DQE} = \frac{(S/N)^2_{\mathrm{out}}}{(S/N)^2_{\mathrm{in}}}.$$

DQE is usually expressed as a percentage, where

$$(S/N)_{\mathrm{out}} = \frac{\log_{10} \mathrm{e}\, \gamma}{\sigma_{\mathrm{D}}}$$

where γ is the contrast, or slope of the characteristic (H&D) curve of the photographic material at the sky background density, and σ_{D} is the rms diffuse density noise. Considered in this way, the older, highly sensitive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

emulsion types (e.g. Kodak Type 103a) had a DQE of about 0.1% at best, while the Kodak IIIa types that replaced them have DQEs of 3–5%. A more recent development, Eastman Kodak's Tech Pan, has an even higher DQE.

The result of these considerations was for the manufacturers to make the emulsions more contrasty (to discriminate faint signals in noise) and finer grained (to improve the signal statistics). However, although finer grains give greatly improved imaging properties, they always yielded less sensitive emulsions. This nexus was broken in the early 1970s, when it was found that baking emulsions just before exposure gave much improved long-exposure performance, mainly by eliminating low-intensity reciprocity failure (LIRF). LIRF was reflection of the electronic inefficiencies in the light detection process which are important at the low photon arrival rates found in astronomy. The processes that counter these inefficiencies are generally known as 'hypersensitization' (hypering). They do little to improve the 'snapshot' or short-exposure sensitivity of the products concerned, but can improve the sensitivity for exposures of many minutes by a factor of 30 in some cases.

Further advances, especially the use of hydrogen as a reduction sensitizer in 1974, and the exposure at the telescope of the hypersensitized material in an inert atmosphere, have enabled extremely fine-grain materials to be used on all telescopes with exposure times comparable with the earlier generation of coarse-grained emulsions. The fine-grained emulsions and use of hydrogen originated in the Eastman Kodak research laboratories, but many astronomical practitioners became skilled at the complex hypering processes, often gas treatments that made the use of the new materials a practical possibility. They also developed their own variations of these processes, skills that have been of particularly important with Schmidt-type instruments specializing in survey work.

Practical considerations

The high-contrast nature of modern photographic emulsions has made them much more difficult to handle, a problem compounded by hypering. In general, high contrast implies that the exposure latitude is very limited. Moreover, it was soon found that the level of exposure was much more critical with the contrasty emulsions than with their relatively low-contrast forebears. The ideal situation was to adjust the exposure so that the density produced by the night sky AIRGLOW was at a developed density of about 1.0 above the chemical fog of the plate. This ensured that the minimum density on the plate was close to the maximum contrast of the emulsion, thus maximizing output signal to noise. However, this implied adjusting the effective exposure within a range of about 10%, a much tighter tolerance than is usual in photography. This in turn implied a knowledge of the (variable) night sky brightness and of the effective sensitivity of the material, which had been greatly increased by hypering. The influence of these variables is reduced by regular measurement of the night sky brightness and an extensive program of testing the effects of hypersensitizing.

A further difficulty with high-contrast emulsions was their extreme sensitivity to processing variations and to mishandling, especially after hypering. The overall effect of the newer emulsions was to force rapid improvements in photographic technique that are reflected in the quality of the recently completed sky surveys and which also positively benefited the last few years of photographic observing on large reflectors.

Photography as a recording medium

It is with Schmidt telescopes that the large sensitive area and high spatial resolution of modern photographic emulsions are best utilized, and it is no accident that Schmidt telescopes are still the main users of photography today. These instruments have fast focal ratios, typically between $F/1$ and $F/3$ and cover wide angles of sky, often many degrees across. Their continued use of photography is not conservatism; rather it is a recognition that the high information content of the wide-field images captured by these telescopes is most efficiently recorded with modern, hypersensitized photographic emulsions. In addition, the world's observatories have substantial photographic archives from larger telescopes and these are a valuable resource.

The UK Schmidt Telescope at Siding Spring in Australia and its northern counterpart, the Oschin Schmidt at Mt Palomar in California, have focal planes measuring 356 mm square, covering $6.6^\circ \times 6.6^\circ$ of sky (see MOUNT STROMLO AND SIDING SPRING OBSERVATORIES, PALOMAR OBSERVATORY). Equipped with hypersensitized Tech Pan film they have a detector with effectively 5 μm pixels operating at a DQE of much better than 5%. Although this is an order of magnitude less sensitive than a CHARGE-COUPLED DEVICE (CCD), CCDs are still limited to much smaller fields and pixels that are effectively 2 or 3 times larger, so the total observing efficiency can still be higher with photography. This has not prevented the very effective use of CCDs in Schmidt telescopes.

Although the basic solid-state mechanisms of photodetection in CCDs and photographic plates are similar, there the similarity ends. After exposure, the latent images on photographic plates and film are developed by immersing the material in a sea of electrons, a chemical reducing agent (developer) that amplifies the photo-electric effect of the photons by a factor of a billion or so. After the removal of the unexposed silver halide ('fixation') the final result is matrix of metallic silver particles suspended in a thin, clear gelatine layer. This is the photographic negative, and in many applications this is the final photographic stage. However, this negative can be copied to produce large numbers of identical-looking duplicates, or portions of it can be photographically enhanced in various ways to extract otherwise hidden information. The large fields are also useful for making color images of the sky.

Much more often nowadays, the original negatives are digitized with specialized scanning machines to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

convert the 10^5–10^6 analogue images buried in the silver grains into digital form. This enables many plates of the same field to be digitally 'stacked', which effectively improves $(S/N)_{\rm out}$ by a factor close to $N^{1/2}$ where N is the number of plates. However, it should be noted that even the most modern current measuring machines such as SuperCOSMOS in Edinburgh, which scans in 10 μm steps, may not be able to extract all the useful information from an exposure which has an inherent resolution around 5 μm. Multiple exposures can also be combined photographically to great effect, and photographic imaging methods such as unsharp masking are also still useful with original plates, where the maximum photographic density in well-exposed areas may be well above the density that can be probed by digitizing machines.

Nonetheless, from these databases, accurate positions, parallaxes and and proper motions (transverse angular velocities) for stars can be derived, as well as information about VARIABLE STARS. With extended objects, the position angles and counts of GALAXIES (for example) can be measured and the detection of faint, extremely extended galaxies that are normally undetectable by visual inspection is simplified. More sophisticated manipulations on sets of plates, especially measurements of the color of objects, lead to large catalogues of high-redshift QUASARS and other unusual objects and recently the detection of a large galaxy in the process of merging with the Milky Way.

It seems likely that the high sensitivity and excellent storage capacity of photographic plates will continue to be an economical and convenient way to gather and store data for some time to come, although those data will be increasingly analyzed and disseminated in digital form.

Bibliography

Eccles M J, Sim M E and Tritton K P 1983 *Low Light Level Detectors in Astronomy* (Cambridge: Cambridge University Press)

Malin D F 1988 The age of the specialist in astronomical photography *Astrophotography* ed S Marx (Berlin: Springer)

Meaburn J 1976 *Detection and Spectrometry of Faint Light* (Dordrecht: Reidel)

Smith A G and Hoag A A 1979 Advances in photography at low light levels *Ann. Rev. Astron. Astrophys.* **17** 43

David Malin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Detectors: Superconducting Tunnel Junction Devices

The absorption of a photon of a wavelength λ (nm) in a superconductor is followed by a series of fast processes which involve the breaking of Cooper pairs by energetic phonons created by the hot electrons produced as the atom relaxes after the initial photoabsorption. The result of this cascade essentially is that the photons' energy is converted into a population of free charge carriers known as quasiparticles in excess of any thermal population. For typical transition metals this conversion process ranges from nanoseconds (niobium) to microseconds (hafnium). At sufficiently low temperatures (typically about an order of magnitude lower than the superconductor's critical temperature T_c) the number density of thermal carriers is very small while the average number of excess carriers N_0 created as a result of the photoabsorption process can be written as

$$N_0(\lambda) \sim 7 \times 10^5/\lambda\ (\mathrm{nm})\Delta(T/T_c).$$

Here the wavelength is expressed in nm and the temperature-dependent energy gap $\Delta(T/T_c)$ is in meV. Thus in a superconductor such as tantalum with $T \ll T_c$ (4.5 K) the initial mean number of free charge carriers created $N_0(\lambda)$ is $\sim 10^3$ eV^{-1}.

The variance on $N_0(\lambda)$ depends on the variance in the partition of the energy of the photons between productive phonons (phonons with an energy $\Omega > 2\Delta$ which can break Cooper pairs) and phonons which are essentially lost from the system ($\Omega < 2\Delta$). The population of $\Omega < 2\Delta$ phonons evolves with time as the average energy of the increasing quasiparticle population relaxes, through quasiparticle phonon emission, towards the bandgap. The variance $\langle N_0 \rangle$ depends on the superconductor's bandgap Δ and its Fano factor F such that

$$\langle N_0 \rangle \sim 7 \times 10^{-4}\mathrm{F}/[\lambda\ (\mathrm{nm})\Delta(T/T_c)].$$

Expressing this variance in terms of the wavelength resolution we have

$$\mathrm{d}\lambda_\mathrm{F}\ (\mathrm{nm}) \sim 2.8 \times 10^{-3}\ \lambda^{3/2}[F\Delta(T/T_c)]^{1/2}.$$

It has been shown that $F \sim 0.2$ for elemental superconductors such as niobium and tin (Kurakado 1982, Rando *et al* 1992). This therefore represents the fundamental Fano limited resolution of any superconductor. Thus a superconductor such as tantalum with $T \ll T_c$ irradiated with photons of wavelengths covering the x-ray to the near infrared $\lambda \sim 1$, 10, 100 and 1000 nm then $\mathrm{d}\lambda_\mathrm{F} \sim 0.001$, 0.033, 1.07 and 34 nm respectively (Peacock *et al* 1997, 1998).

The quasiparticles produced through photoabsorption can be detected by applying a d.c. potential across two such films separated by a thin insulating barrier, forming a superconducting tunnel junction (STJ). This potential bias favors the transfer of quasiparticles from one film to the other through quantum mechanical tunneling across the barrier. The detector signal is therefore represented by the current developed by this tunnel process. After initial tunneling, a quasiparticle can tunnel back, therefore contributing many times to the overall signal (Grey 1978). On average each quasiparticle will contribute $\langle n \rangle$ times to the signal through an average of $\langle n \rangle$ tunnels before it is lost from the system through traps etc (Poelaert *et al* 1999). Hence the mean number of effective charge carriers $N = nN_0$. The multiple tunnel process leading to n the average number of tunnels per quasiparticle is of course subject also to statistical fluctuation (Goldie *et al* 1994). The fluctuations due to the Fano process and that arising from the tunnel process can be added in quadrature such that the overall limiting resolution for a perfectly symmetrical superconducting tunnel junction can be written as

$$\mathrm{d}\lambda_\mathrm{T}\ (\mathrm{nm}) \sim 2.8 \times 10^{-3}\lambda^{3/2}\Delta(T/T_c)^{1/2}(F + 1 + 1/n)^{1/2}$$

for the case when $n \geq 2$. Figure 1 illustrates this tunnel junction limited resolution for a number of elemental superconductors based on the parameters summarized in table 1. Note this expression for the tunnel limited resolution $\mathrm{d}\lambda_\mathrm{T}$ can be further generalized to any superconductor compound or proximized bilayer through the use of the approximate BCS relation in the weak coupling limit of $2\Delta = 3.5kT_c$, where k is Boltzmann's constant. Deviations from this relation are small even for strongly coupled superconductors such as niobium and are also summarized in table 1. Thus in terms of the critical temperature we can write

$$\mathrm{d}\lambda_\mathrm{T}\ (\mathrm{nm}) \sim 1.1 \times 10^{-3}\lambda^{3/2}T_c^{1/2}(F + 1 + 1/n)^{1/2} \quad (n \geq 2).$$

Typically n is of order 10–100 and depends on the size and nature of the STJ.

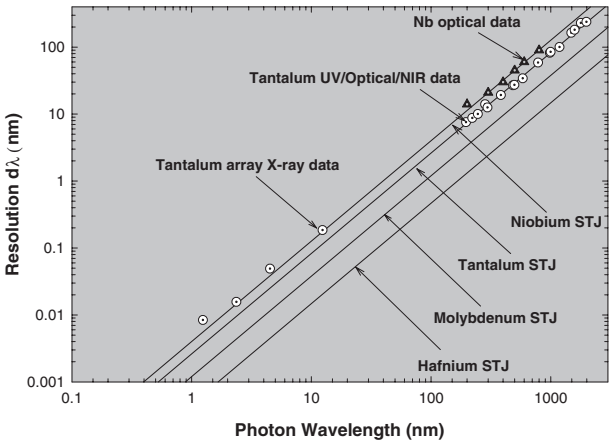


Figure 1. The tunnel limited resolution of a number of elemental superconductors as a function of wavelength from the x-ray to the near infrared. The experimental data derived from niobium and tantalum STJs are also shown.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Key properties of elemental superconductors used in the fabrication of STJs.

Material	T_c (K)	Ω_D (meV)†	Δ (meV)	τ_{qp} (ns)‡	τ_Ω (ns)‡	H_c(G)§	$2\Delta/kT_c$
Niobium	9.20	23.7	1.550	0.149 (0.386)	0.004 (0.009)	1980	3.9
Vanadium	5.30	32.7	0.800	5.3 (4.8)	0.013 (0.012)	1420	3.5
Tantalum	4.48	20.7	0.664	1.8 (2.4)	0.023 (0.016)	830	3.4
Aluminium	1.14	36.9	0.172	100 (453)	0.242 (0.246)	105	3.5
Molybdenum	0.92	38.8	0.139	2077 (2963)	0.405 (0.579)	95	3.5
Hafnium	0.13	21.7	0.021	∼(8000)	(20)	13	3.9

† The Debye energy.
‡ The quasiparticle (τ_{qp}) and phonon (τ_Ω) characteristic times taken from Kraft *et al* (1998). Data in brackets are based on recent data on the phonon related characteristics of these materials.
§ The critical magnetic field.

Astrophysical rationale

Having established the limiting wavelength resolution $d\lambda_T$ of an STJ it is worth considering the applications of such a resolution within such fields as astrophysics.

X-ray wavelengths

Any PLASMA at a temperature above $\sim 10^6$ K radiates the bulk of its energy as x-rays from collisional excitation line emission and thermal bremsstrahlung continuum processes. At low temperatures ($<10^7$ K) the bulk of this photon emission is in the form of emission lines radiated in the soft x-ray band (SXB) with $\lambda > 0.5$ nm ($\sim$2 keV), while at higher temperatures continuum emission processes dominate. This only arises as a result of the fact that at higher temperatures the majority of ions with a low atomic number are completely stripped of their electrons while the remaining ions are hydrogenic or helium-like species of sulphur and iron.

High-resolution x-ray spectroscopy (see X-RAY ASTRONOMY) provides the ability to determine the electron and ion temperatures the electron density and the relative abundance of the elements, as well as establishing the degree of thermal and ionization equilibrium. While the measurement of the intensity of the hydrogenic and helium-like lines from the same element is an important ion temperature indicator, it is the ability to resolve the satellite lines, e.g. the resonance, forbidden and intercombination lines, from helium-like species which can determine the key characteristics of the x-ray emitting plasma in a model independent manner. Table 2 summarizes these key transition wavelengths for some of the most abundant elements expected to be present in an astrophysical plasma together with the tunnel limited resolution of a tantalum based STJ at these wavelengths. Clearly such an STJ, provided it can achieve a measured resolution close to $d\lambda_T$, should be capable of resolving these key transitions. To illustrate this point figure 2 shows the response of a tantalum STJ to the large complex of lines (the Fe-L complex) around 1 nm, expected to be radiated from an optically thin plasma having a temperature $\sim 10^7$ K. In this example SOLAR ABUNDANCES and ionization equilibrium were assumed, and for clarity the continuum emission has been suppressed. The majority of lines are easily resolvable with such a tantalum STJ enabling the temperature to be uniquely determined

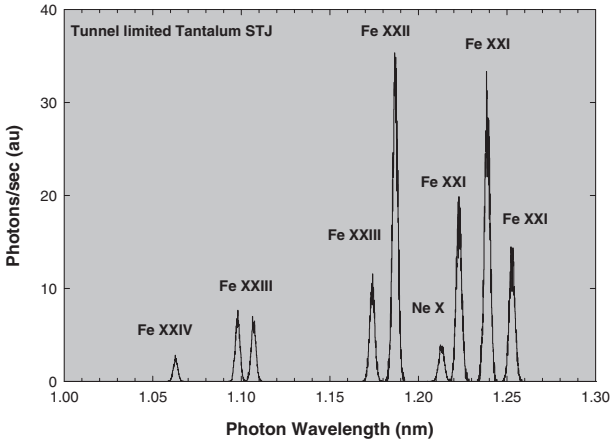


Figure 2. The simulated response of a tantalum based STJ to the Fe-L complex of lines around 1 nm from a hot optically thin plasma at a temperature of 10^7 K. Practically all the lines are resolvable by such a detector.

through the measurement of the relative intensity of the lines from the same ion. Also, through the relative intensity of lines from different elements such as Fe and Ne, the relative abundances can be established. Note that the intensity ratio of resonance lines from different ions of the same element together with line centroids allows one to deduce either the degree of ionization equilibrium or possibly distance to the object, through the determination of the REDSHIFT z. Needless to say, a high spectral resolution is required for such observations. This resolution can be achieved using a tunnel limited tantalum STJ but is completely impossible with conventional solid state devices.

To illustrate the high degree of sensitivity of various emission line strengths to plasma temperature figure 3(a, b) shows the simulated tantalum STJ response to emission line spectra from a hot solar abundant optically thin plasma in equilibrium over the wavelength region covered by the hydrogenic and helium-like oxygen lines $\sim$2 nm. Again for clarity the continuum has been suppressed. The ion temperature T was taken to be $\log T \sim 6.4$ and 6.8 respectively. This temperature can be established directly from the line intensities

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. The principal line transitions of the most abundant helium-like lines.

Element	$1s^2$–$1s2p^1P$ (R)	$1s^2$–$1s2p^3P$ (I)	R–I (nm)	$1s^2$–$1s2p^2P$ (F)	I–F (nm)	$d\lambda_T$
Iron	0.1850	0.1859	0.0009	0.1867	0.0008	0.0002
Sulphur	0.5039	0.5067	0.0028	0.5099	0.0032	0.0009
Silicon	0.6650	0.6690	0.0040	0.6737	0.0047	0.0014
Oxygen	2.1602	2.1804	0.0202	2.2100	0.0296	0.0083

R = Resonance line, I = intercombination line and F = forbidden line transition. All line wavelengths are in nm. $d\lambda_T$ (nm) the tantalum tunnel limited STJ resolution determined at the resonance line wavelength.

without recourse to modeling the underlying continuum bremstrahlung spectrum which anyway provides only a measure of the free electron temperature.

UV/optical wavelengths

In optical and UV spectroscopy high resolution normally implies a resolving power $\lambda/d\lambda > 10^4$. From figure 1 it is clear that none of the classical superconductors forming the basis of current STJs under development (those based on Nb, Ta, Al, Mo or Hf) could achieve such a resolving power. In fact a superconducting critical temperature $T_c \ll 100\ \mu$K is implied to achieve a resolving power of 10^4 leading to the development of STJs based on such elemental superconductors as rhodium. Of course things are not quite this simple, with the temporal characteristics associated with the production of the free excess charge carriers being a function of the critical temperature (see table 1) while the phonons with $\Omega > 2\Delta$ have wavelengths significantly larger than the thickness of the film. Thus such low-temperature superconductors may well be significantly slower in their overall in response.

Given that the resolution of a typical STJ based on tantalum is not appropriate for high- or even medium-resolution spectroscopy what are the alternative key attributes which such a device can bring to the fields of OPTICAL/ULTRAVIOLET ASTRONOMY? Two features are important: (a) The timing characteristics ($\leq$10 μs) coupled to the broadband spectral capability may make this the ideal spectrophotometer: objects such as pulsars and flare stars may be ideal objects with which to observe with narrow field small arrays. (b) The efficiency at UV wavelengths which, if coupled to a large format array (a panoramic detector), may allow for the development of an efficient broad band imaging spectrometer with which to determine the low-resolution spectra of very faint objects allowing for very deep field surveys. Such surveys could allow the determination in a single exposure of the broadband spectra and possibly therefore the redshift z (and therefore age) of all objects in the field through the measurement of the Lyman edge and the Lyman emission lines—the Lyman forest. Note that the observed wavelength $\lambda_o = \lambda_R(z+1)$, where λ_R is the rest wavelength. Thus the classical Lyman edge would appear at $\sim$400 nm at $z \sim 3$. This is close to the optimum performance for a tantalum based STJ where it has an efficiency of $\sim$70% and a resolution of $\sim$20 nm. It is, however, clear that STJ devices based on lower-temperature superconductors such as hafnium would allow the clear evaluation of redshift.

Current performance of STJs

The key factors described in this section involve the basic performance of tantalum based STJs building on the earlier work in niobium. The validation of the basic equations discussed earlier relating to both N_0 and $d\lambda_T$ with both tantalum and niobium devices give some confidence in the ultimate successful development of lower-temperature elemental superconducting tunnel junctions such as those based on hafnium (Kraft *et al* 1998).

X-ray wavelengths

The efficiency of any detector is an important parameter when considering practical applications. Figure 4 illustrates the efficiency of a tantalum based STJ as a function of photon wavelength for the case when one film about 100 nm thick is used as the primary detection element. For comparison the efficiency of a hafnium film is also shown. Little difference exists between such films given their similar atomic numbers although their tunnel limited resolution should be very different. At a wavelength of 2 nm the efficiency is $\sim$75%; however, the situation rapidly degrades at shorter wavelengths with an efficiency of only 5% at 0.2 nm. While the thickness of these films can undoubtedly be increased to 200–300 nm, beyond this various loss mechanisms may become important such that the spectral resolution could be expected to degrade. Of course in practice the efficiency at the longer wavelengths will be lower than that indicated in figure 4 due to the fact that unlike in the optical/UV where back illumination is the mode of operation, at x-ray wavelengths the photon enters the detector through the front (front illumination). This means that some fraction of the x-rays at long wavelengths are absorbed in the top film oxide layer and also the top film, if the bottom film is used as the primary detection film, as well as the top contact. This latter point is important when considering arrays in which a significant amount of top contact wiring together with a SiOx insulation layer is required.

While the resolution in the medium x-ray band (MXB $\sim$ 0.1–0.5 nm) has not yet reached the tunnel limits indicated by figure 1, the situation in the soft x-ray band (SXB $\sim$ 0.5–10 nm) is close. Figure 5 illustrates the measured spectra from a tantalum STJ forming part

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

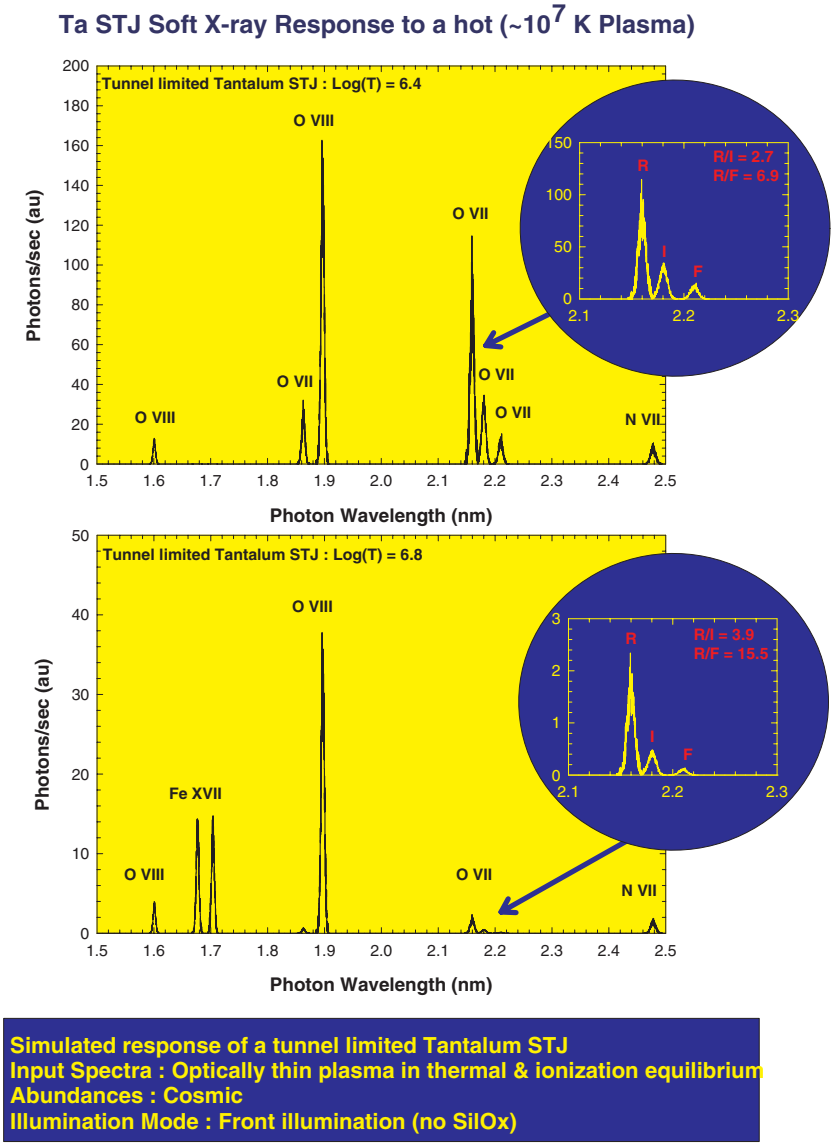


Figure 3. The simulated response of a tantalum based STJ to the helium and hydrogenic lines of oxygen at a temperature T of (*a*) $\log T = 6.4$ and (*b*) $\log T = 6.8$, from an hot optically thin plasma. Note the difference in scale between the two spectra as well as the radically different line ratios when the temperature has changed by only a factor of 2.5.

of a 6 × 6 element array illuminated by monochromatic radiation of various wavelengths in the SXB (Rando *et al* 1998). Each device was 25 × 25 μm and consisted of two films each 100 nm thick. Only those photons absorbed in the base film, which are separated from top film and substrate events by their distinct signal risetime are shown here. Typical resolutions $d\lambda \sim 0.015$ nm (3.5 eV) at $\lambda \sim 2.4$ nm (∼500 eV) were measured and are indicated in figure 1. While these data have not yet achieved the tunnel limited resolution ($d\lambda_T \sim 0.01$ at $\lambda \sim 2.4$ nm) the cause is determined to be a spatial variation in the detector gain which provides an additional variance dependent on the square of the photons energy and which contributes to the overall variance (Verhoeve *et al* 1998).

UV/optical wavelengths

At optical and UV wavelengths, where the photon energy is very small, spatial effects on the resolution are unimportant. Here it is rather that the signal is low such that the signal to noise ratio is the dominant factor governing the measured resolution. At these wavelengths the photons enter the detector through the substrate, which can either be sapphire or magnesium fluoride depending on the short-wavelength cut-off required (Peacock *et al* 1998). The theoretical efficiency of a tantalum device deposited on a sapphire substrate with this mode

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

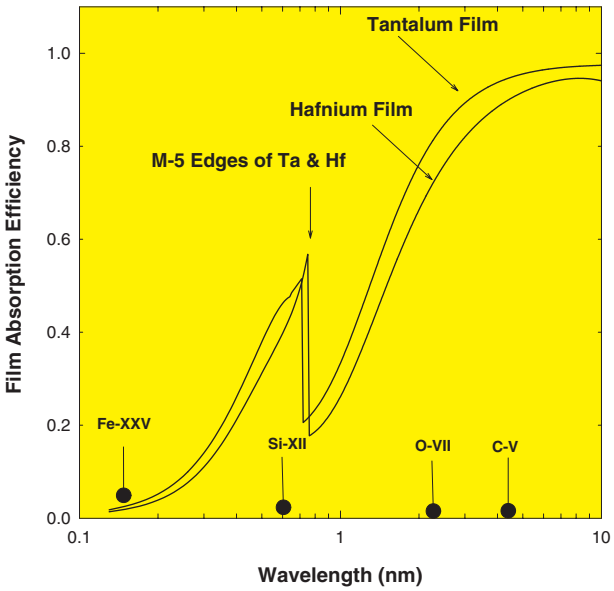


Figure 4. The efficiency of a tantalum and hafnium 100 nm film to the absorption of x-rays as a function of wavelength. Some key emission lines are indicated.

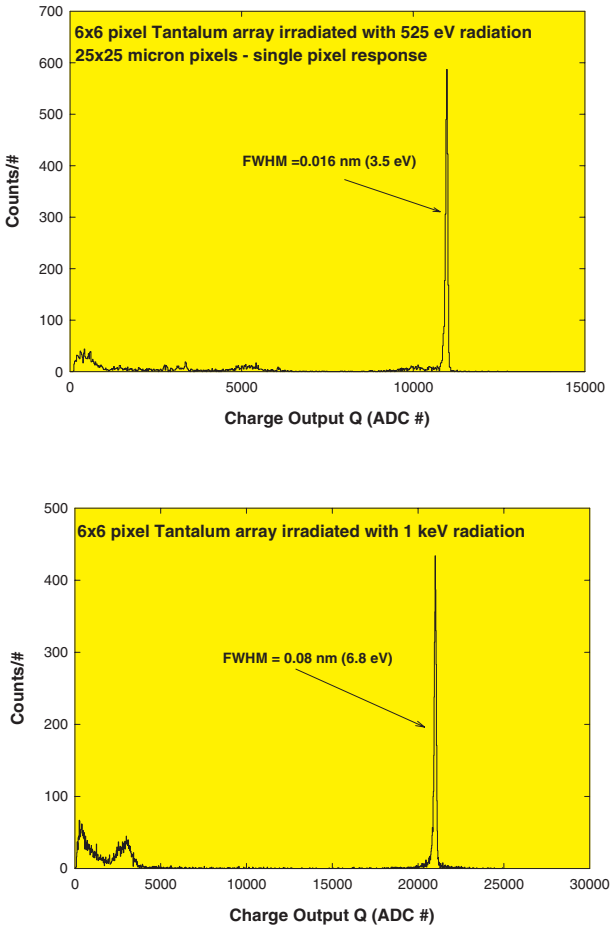


Figure 5. The spectra from a single pixel of a 6 × 6 tantalum array illuminated by monochromatic soft x-rays.

of illumination is very high. All photons are absorbed in the high-quality epitaxial tantalum base film. Efficiencies of ∼70% from 200–600 nm are expected limited at the short wavelength by the cut-off of the sapphire substrate. Such efficiencies have been experimentally confirmed.

To illustrate the broad band response of this type of photon counting spectroscopic detector, figure 6 shows the charge spectrum from a single tantalum based device when illuminated with optical light via a grating monochromator. This grating response covers four orders from 296 nm to 1183 nm, i.e. from the UV to the NIR. Not only are the various orders well resolved but the charge output as a function of wavelength can be precisely determined leading to a wavelength linearity which is very high. These types of measurements allow the determination of the wavelength resolution across a broad waveband and are shown for both tantalum and niobium based devices in figure 1.

Conclusion

The STJ based currently on tantalum or niobium has now been developed to a stage where practical small format arrays (3 × 3 and 6 × 6 pixel) have been produced which provide similar performance to optimized single devices. Such arrays are already being developed into instruments for ground based optical astronomy (Rando *et al*). The performance of these arrays at UV and SXB wavelengths is such that practical instruments can now be considered for space based applications. The key specific points which have been experimentally demonstrated can be summarized as:

(a) Tunnel limited resolutions have been achieved at optical and UV wavelengths.
(b) High efficiency has been shown at UV wavelengths.
(c) Resolutions are within a factor of two of the tunnel resolution in the SXB.
(d) High-speed photon counting has been realized (10 kHz).

Areas where development of the basic tantalum device are still required can be summarized as:

(a) Reduction of the spatial contribution to the resolution allowing for the demonstration of tunnel limited resolution in the MXB.
(b) Improvement in the efficiency of the device in the MXB
(c) Reduction in the various absorbing materials (contacts, SiOx etc) which reduce the efficiency in the SXB.
(d) Production of very large format arrays.
(e) Development of large format application specific readout electronics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

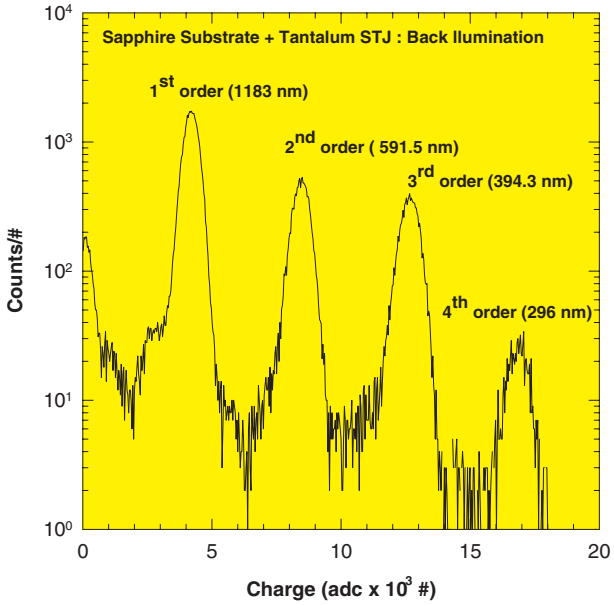


Figure 6. The charge spectrum from irradiation of a tantalum STJ by photons of wavelength 1183 nm. The various orders from the grating monochromator are easily discernible and provide a excellent technique with which to establish the linearity of the device.

Bibliography

Goldie D *et al* 1994 *Appl. Phys. Lett.* **64** 3169
Grey K 1978 *Appl. Phys. Lett.* **32** 392
Kaplan S *et al* 1976 *Phys. Rev.* B **14** 4854
Kraft S *et al* 1998 *Proc. SPIE* **3445** 226
Kurakado M 1982 *Nucl. Instrum. Methods* **196** 275
Peacock A *et al* 1997 *Astron. Astrophys. (suppl)* **123** 581
Peacock A *et al* 1998 *Astron. Astrophys. (suppl)* **127** 497
Poelaert A, Kozorezov A, Peacock A and Wigmore K 1999 *Phys. Rev. Lett.* **82** 1257
Rando N *et al* 1992 *Nucl. Instrum. Methods* A **313** 173–85
Rando N *et al* 1998 *Proc. SPIE* **3445** 74
Verhoeve P *et al* 1998 *Appl. Phys. Lett.* **72** 3359

A Peacock

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Diamond Ring

An effect observed at the onset and end of the totality phase in a total eclipse of the Sun. Just as the extreme edge of the Sun's disk is about to disappear behind (or emerge from behind) the Moon's disk, a bright arc of sunlight is seen, giving the impression of a ring. In conjunction with Baily's Beads-type bright points, the ring has highlights like a diamond ring.

See also: eclipse.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dichotomy

The phase of a body in the solar system when exactly half of its sunlit side is visible. The term is used in particular for the half-phases of the inferior planets, Mercury and Venus; in correct usage it is not applied to the Moon, for which the terms first quarter and last quarter are preferred. Other bodies in the solar system can be imaged at dichotomy only by spacecraft.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dicke, Robert Henry (1916–)

Physicist, born in St Louis, MO, professor at Princeton (1946–84). He did not believe EINSTEIN's general theory of relativity and conducted numerous experiments to challenge its famous tests. He attempted to determine whether the Sun was oblate (squashed at the poles) and could thus cause alterations to the orbit of Mercury which had been interpreted in Einstein's favor. He re-measured the gravitational deflection of starlight by the Sun. He measured the position of the Moon by laser beams reflected back to Earth. He determined the age of the oldest stars. All this was to look for effects that would provide evidence of his own theory. This was all in vain, and his challenge was unsuccessful. He was, however, a believer in the theory that the universe began with a Big Bang, and had identified the possibility that there would be a microwave remnant of that event. He was building a radiotelescope to look for it when ROBERT WILSON and ARNO PENZIAS found it while engaged on other studies. Some claimed that Dicke should have shared in the Nobel prize with them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Differential Rotation

The variation with latitude of the angular rotation rate of the Sun. From studies of the motion of sunspots across the solar disk, carried out between 1853 and 1861, English astronomer Richard Carrington (1826–75) showed that the angular rotation rate of the solar surface decreases, and the rotation period increases, with increasing latitude north or south of the solar equator. Although sunspots are seldom observed at latitudes much higher than about 30°, measurements of rotation rates at a wide range of latitudes can be obtained by measuring Doppler shifts in the wavelengths of spectral lines at different points on the Sun. Recent observational data indicate that the sidereal rotation period of the photosphere increases from 24.8 days at the solar equator, to 26.1 days at latitude 30°, about 31 days at latitude 60° and approximately 35 days at the poles.

Helioseismology (a technique for studying the structure and motions of the solar interior) indicates that differential rotation extends down to the base of the convective zone, which is located at a depth equal to 29% of the Sun's radius, but that below this level the solar interior has a near uniform rotation period of just under 27 days.

See also: Doppler effect, helioseismology, sidereal period, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Differentiation

The process by which a planetary body (a solid planet or major satellite) acquires a layered structure with dense materials at its center, in a core, and less dense materials in an overlying mantle and crust. A body that has formed by accretion has a homogeneous composition, with materials of different densities distributed uniformly throughout its volume. In order for its constituent materials to begin separating out, the interior of a body must begin to melt. The heat required for melting comes from a number of sources. They include gravitational compression (temperature increases with pressure, heating the center), the kinetic energy of further impacts (which is converted into heat energy, heating the surface), radiogenic heating (energy is released when radioactive elements decay, heating the whole body) and tidal forces (exerted by a planet on a satellite, deforming it and thus heating the interior). Once a body is at least partially molten, dense materials can start to gravitate toward the center, where they form a core, while less dense materials begin to float toward the surface, where they form a mantle and crust. Differentiation is thus driven by gravity, which also plays a part in the melting that makes it possible, so the process can take place only in bodies that have grown sufficiently large. The temperature at which melting can commence depends on the melting points of a body's constituent materials.

Differentiation of the terrestrial planets typically began with the formation of a *core*. Once molten at the center, dense material (in the Earth's case, predominantly iron and nickel) would quite quickly segregate toward the center. The gravitational potential energy released by core formation would have been converted into enough heat to melt all of the overlying layers, forming a rocky *mantle*. After this initial phase, with the outer layers solidifying as heat was radiated into space, the other factors played a greater part. Radiogenic heating in the mantle produced partial melting, forming pockets of lighter magmas which rose to the surface, erupting as volcanoes, a process which continues on the Earth and, quite possibly, Venus. Impacts, particularly in the period known as late heavy bombardment which ended about 3.9 billion years ago, provided the energy to melt the entire surface, creating global magma oceans in which lighter material formed by chemical segregation floated to the surface (in much the same way as slag forms at the top of a blast furnace), to solidify as a *crust*.

Larger satellites in the outer solar system are also believed to have undergone differentiation. For Jupiter's Galilean satellites, tidal heating may well have been a contributory factor at some stage in their internal evolution. Io, Europa and Ganymede are thought to be fully differentiated, with predominantly iron cores surrounded by silicate mantles (molten in Io's case) and ice crusts, Ganymede having in addition a substantial icy outer mantle. The most distant of the four, Callisto, is thought to have only a partially differentiated rock/ice interior. Mid-sized icy satellites of Saturn and Uranus may well have experienced differentiation, mainly through radiogenic and tidal heating. This would have been possible through the formation of what is known as a eutectic melt of ammonia and water, which is molten at a temperature as low as 176 K, and behaves under the prevailing conditions rather as magma on the Earth. The strange surface of Miranda is thought to have resulted partly from an incomplete differentiation that affected some parts of the surface only.

In the asteroid belt, Vesta is the only body for which there is good evidence for differentation having occurred, and which has survived intact the catastrophic collisions that completely disrupted other differentiated bodies to produce, for example, the M and R class asteroids, and ion meteorites.

See also: accretion, core, crust, mantle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Diffraction

The bending and spreading out of waves (for example light waves or water waves) that occurs when they pass by the edge of an opaque object or through a narrow slit or aperture. Light waves from a distant point-like source advance as a series of wavefronts ('crests' and 'troughs') that lie at right angles to the direction in which the light is propagating (like ocean waves rolling up onto a beach). Diffraction occurs because each point on an advancing wavefront acts as a tiny source of wavelets that spread out spherically. When the wavefronts from the source are interrupted, by encountering an opaque obstruction or by passing through a slit or aperture, secondary wavefronts spread out from the edges of the obstruction or aperture. This causes some of the light to pass into what ought (if light rays traveled purely in straight lines) to be the geometric shadow of the object, and causes the edges of shadows cast by opaque objects to be fuzzy rather than perfectly sharp.

When a wavefront is imaged after passing through a slit-shaped aperture, the resulting image consists of a narrow central band of light together with, on either side, a number of successively fainter bands. This pattern, which is called a diffraction pattern, arises from the mutual interference of waves originating at different parts of the aperture. The paths along which these waves have traveled in order to reach the same point differ in length, so the waves arriving at a particular point are out of phase to a greater or lesser extent. Where the path difference is precisely half a wavelength (or an integral number of wavelengths plus half a wavelength) they will be 180 degrees out of phase (a crest from one coinciding with a trough from the other) and will cancel out. Where the path difference is a whole number of wavelengths, they will be in phase (a crest coinciding with a crest) and will add together to produce a crest of larger amplitude. The central brightness maximum occurs where a large number of waves reinforce each other. The first minimum arises where the path difference from opposite sides of the aperture is half a wavelength, so the waves from one side cancel the waves from the other. The next maximum occurs where the path difference is a full wavelength, and so on. When light passes through a circular aperture, the resulting diffraction pattern consists of a central spot (which is called the 'Airy disk'), into which most of the light is concentrated, surrounded by a series of concentric light and dark rings (or 'fringes').

The magnitude of any diffraction effects depends on the wavelength of the light and the size of the aperture (the longer the wavelength, or the smaller the aperture, the larger the diffraction pattern). Diffraction affects the quality and sharpness of images produced by optical systems. In particular, the theoretical resolving power of a telescope is determined by the size of the Airy disk that is produced when a point-like source, such as a star, is imaged.

See also: Airy disk, aperture, diffraction grating, resolving power, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Diffraction Grating

A plate on which a large number of parallel grooves or slits have been cut and which spreads light into its constituent wavelengths, so producing a spectrum. When a collimated beam of light falls on the grating, light is diffracted by each of the grooves or slits and, in certain particular directions, waves of different wavelengths leaving the different parts of the grating interfere with each other to produce a spectrum. A diffraction grating that transmits light is called a transmission grating and a grating that works by reflecting light from the surfaces of the grooves is called a reflection grating. Most astronomical spectrographs employ reflection gratings.

The resolving power of a grating (its ability to reveal fine detail in a spectrum) depends upon the total number of grooves that it contains. Gratings used for optical astronomical purposes typically have between 100 and 1000 grooves per millimeter and contain between 1000 and 50 000 grooves in all.

A simple grating spreads the incoming light into a large number of spectra, the fraction of the light going into any one spectrum being relatively low. In order to overcome this problem, the reflecting surfaces of the grooves are orientated at a particular angle (the grating is said to be 'blazed') so as to concentrate most of the light (up to about 90%) into one of these spectra.

See also: diffraction, interference, interferometer, spectroscope/spectrograph, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Diffuse Interstellar Bands

In 1922 Berkeley/Lick Observatory graduate student Mary Lea Heger included two unidentified features in a published list of 'stationary' absorption lines seen in the spectra of distant binary stars. Spectral lines due to stars in a binary system shift back and forth in wavelength owing to the Doppler effect, while the mystery lines did not. Astronomers soon realized that 'stationary' lines are formed in interstellar space and thus do not share the motion of the stars in an orbiting system — but the two features found by Heger remained unidentified. More than a decade later Paul Merrill, who followed Heger as a graduate student at Berkeley/Lick, investigated the unknown features and found about 10 more. Merrill dubbed them diffuse interstellar bands (DIBs) owing to their great widths (as compared with interstellar spectral lines due to atoms and ions).

Now, 80 years after Heger's first notice of the DIBs, more than 200 of the features have been found (see figure 1) — and not even one has been identified. The behavior of the DIBs is consistent with an interstellar origin, in that they grow in strength with distance and total gas abundance, but their wavelengths and widths do not match the spectral features of any known atom, ion or molecule.

Historical developments

Almost immediately following Merrill's initial work on the DIBs in the mid-1930s, several prominent astronomers suggested that molecules in space might be responsible, as this was about the same time when the first simple (two-atom) interstellar molecules were discovered. However, the unknown bands were very broad, not consisting of separate, sharp lines as would be expected from small molecules. A more difficult challenge for the molecular hypothesis was that the density of gas in interstellar space is very low, seemingly too low to allow the formation of complex molecules.

Astronomers turned their attention to another hypothesis: that the DIBs are formed by solid particles in space known as interstellar dust grains. Solid particles can absorb at specific wavelengths to create bands similar to the DIBs, if the particles contain impurity atoms or attached atoms or molecules on their surfaces. For about three decades starting in the 1950s the dust grain hypothesis was favored by most researchers, but again no specific substances or particles could be identified with any of the DIBs. Furthermore, careful examination of specific DIBs failed to reveal some of the properties expected of solid-state absorbers, such as bandwidth and profile variations with varying grain size, or polarization within the bands.

Thus the dust hypothesis was already in trouble when, starting in the 1970s, several factors began to point back in the direction of a molecular origin for the DIBs. The first significant development was the discovery of many new and increasingly complex molecules in dark interstellar clouds, where molecules emit at radio wavelengths, releasing energy gained in collisions with each other. It became apparent that molecules can form and survive in interstellar space, at least in the very densest clouds if not in the more diffuse regions where the DIBs arise. Then a series of papers published in the late 1970s showed how molecules, even relatively small ones, could form broad, featureless spectral bands similar to the DIBs. However, the strongest impetus for the molecular hypothesis came in 1984 with the discovery that large organic molecules are abundant in space, even in diffuse regions. In that year the French scientists Jean Puget and Alain Léger showed that certain interstellar features seen at infrared wavelengths can be attributed to large hydrocarbon molecules, specifically a class known as polycyclic aromatic hydrocarbons (PAHs). Almost immediately several astronomers suggested that PAHs, or more likely their ions (having lost one electron), could produce the DIBs. This hypothesis remains viable, as do other classes of molecules such as carbon chains, which are abundant in dense clouds.

Since the mid-1980s progress has been rapid, for at least two reasons: (1) the availability of very accurate astronomical measurements of the DIBs, taking advantage of new technology in telescopes and light-detecting

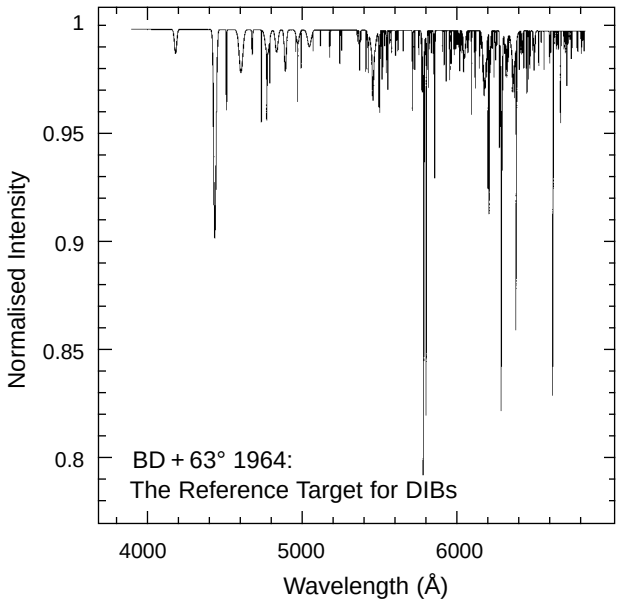


Figure 1. The diffuse interstellar bands in the spectrum of a distant star. All of the dips in the graph shown here are DIBs in the spectrum of the star BD + 63 1964. The particles that create these spectral bands have not been identified, but are thought to be complex organic molecules in interstellar space (O'Tuairisg *et al* (2000), *A&A Suppl.* **142**, 225–238).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

devices such as charge-coupled devices; (2) the involvement in DIB research of scientists, such as chemists, from fields other than astronomy.

Properties of the DIBs

Today more than 200 DIBs have been found, and additional weak ones appear almost every time a new level of sensitivity is reached in the observations. All of the DIBs correlate roughly with interstellar parameters such as total gas column density or the extinction of starlight due to interstellar dust — but none correlates perfectly with these quantities or with each other. Subgroups of DIBs have been found to correlate internally better than the bands as a whole, leading to the identification of 'families' that may have a common origin. In any event it has become clear that multiple carriers must be involved, given the sometimes widely disparate behavior of different DIBs.

In general the DIBs are strongest in diffuse interstellar regions rather than the dark, dense clouds where radio-emitting molecules are observed. In fact the DIBs become weaker with increasing cloud density, suggesting that they may be formed by ionized species that become neutral and alter their spectra in denser environments.

The widths of the DIBs range from less than 0.1 nm to nearly 3 nm. Most lie between roughly 400 and 800 nm, with a greater concentration of features toward the longer wavelengths. Features appear in the near-infrared portion of the spectrum as well, although observations there become difficult owing to atmospheric interference. It is not known whether DIBs occur in the ultraviolet because of observational constraints making their detection difficult.

The wavelengths and profiles of the DIBs are, within observational limits, constant. Minor profile variations may be seen from one line of sight to another, but these are usually attributable to Doppler shifts in separate interstellar clouds that contribute to the absorption. A few of the narrower DIBs display substructure in their profiles which is usually interpreted as due to rotational–vibrational spectra of molecules, although other interpretations have been proposed, such as isotopic shifts. The broadest DIBs, which tend to lie toward shorter wavelengths than most of the narrow ones, appear to lack any detectable fine structure. The widths and profiles of these bands have been attributed to natural broadening (damping) in molecules with very short-lived upper states.

In a very few cases some of the DIBs have been seen in emission rather than absorption. The best-known region where such emission has been observed is a carbon-rich reflection nebula called the Red Rectangle. In this nebula wavelength shifts in the DIB emission bands vary with distance from the exciting star in a manner that appears consistent with variations in rotational excitation of molecules.

An order-of-magnitude estimate of the quantities of molecules needed to produce all the observed DIBs places constraints on the elements that may be responsible. Only universally common species such as hydrogen, carbon, nitrogen, oxygen and a few others are viable; other elements are too rare. Thus hydrocarbons or carbon chains, as well as other types of organic molecules, satisfy the abundance constraint and are therefore plausible on these grounds as well as the spectroscopic properties mentioned above.

Current Status and Future Prospects

Ultimately the goal is to create and measure, in the laboratory, the particles that produce the DIBs. Only when species are found whose spectra precisely match the observed interstellar bands (while not predicting features that should be seen in space but are not) will we be able to claim identification. Achieving this goal is a daunting task, given the enormous variety of viable species, each with its own unique spectrum, and given also the practical difficulties involved with laboratory studies of large molecules, especially molecular ions which are favored by many researchers as the most probable candidates for the DIB carriers.

Despite the obstacles, several research groups around the world are pursuing the DIBs, in the laboratory and at the telescope. There have already been some near-misses, i.e. species displaying enough of the required properties to raise hopes, but so far all have fallen by the wayside. However, given the unprecedented current level of interest and intensity of effort in attacking the problem, there is reason for optimism that at least some of the DIBs may be identified soon. When that happens, astronomers will have a new tool for probing the physics and chemistry of interstellar space, probably gaining as well new information on the abundances and types of organic species that were present in space long before life formed on Earth.

Bibliography

Herbig G B 1995 The diffuse interstellar bands *Anna. Rev. Astron. Astrophys.* **33** 19–74

Salama F, Bakes E L O, Allamandola L J and Tielens A G G M 1996 Assessment of the polycyclic aromatic hydrocarbon–diffuse interstellar band proposal *Astrophys. J.* **458** 621–636

Salama F, Galazutdinov G A, Krelowski J, Allamandola L J and Musaev F A 1999 Polycyclic aromatic hydrocarbons and the diffuse interstellar bands: a survey *Astrophys. J.* **526** 265–273

Snow T P 2001 The unidentified diffuse interstellar bands as evidence for large organic molecules in the interstellar medium *Spectrochim. Acta A* **57** 615–626

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Tielens A G G M and Snow T P 1995 *The Diffuse Interstellar Bands* (Dordrecht: Kluwer)

Theodore P. Snow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Digges, Leonard (*c* 1520–*c* 1559) and Digges, Thomas (1545/6–95)

Both were English astronomers, opticians and military engineers. Thomas was born in Wotton, Kent, England, and incorporated his father's work on optics and ballistics into his own publications. He was tutored by JOHN DEE. In 1573 Thomas Digges published *Alae seu Scalae Mathematicae*, a work on the position of the supernova of 1572, showing it had no parallax, i.e. was at a great distance, beyond the Moon, thus showing that changes took place in the stellar sphere. The observations demonstrate that Digges was the best observer of his time. Digges went on to translate COPERNICUS' works and added his own ideas of an infinite universe with the stars at varying distances in space to his *Perfit Description of the Caelestial Orbes* in 1576.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dione

A mid-sized icy satellite of Saturn, discovered by Giovanni Cassini in 1684. Its diameter is 1120 km, and it orbits at a distance of 377 000 km. Its density of 1490 kg m^{-3}, the highest of Saturn's mid-sized satellites, indicates a composition of rock and ice. There are two basic terrain types on Dione: cratered terrain and cratered plains; the cratering on the plains indicates that the resurfacing process that created them took place long ago. The largest crater is the 231 km diameter Amata. The plains are crossed by troughs and fissures, the longest being the 394 km Palatine Chasma, which is up to 8 km wide. There are also features that look like the Moon's sinuous rilles (which are known to be collapsed lava tubes), and strange wispy patterns on the surface. A small satellite, Helene, is co-orbital with Dione.

See also: co-orbital satellites, Saturn: satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Direct Motion

Angular motion in the prevailing direction, also known as *prograde motion* (as opposed to retrograde motion, in the opposite sense). Motion of a body in the solar system, either axial rotation or orbital revolution, is direct if it is in the same direction as that of the Sun's rotation: counter-clockwise as viewed from the Sun's north pole. The majority of solar system objects show direct motion, which is a relic of the motion of the rotating disk of material from which the Sun and the planets formed. Bodies in direct orbits have orbital inclinations of less than 90°; bodies with direct rotation have axial inclinations of less than 90°. All the major planets, most planetary satellites, all but a handful of asteroids and the majority of comets have direct orbits. Of the major planets, Mercury, Earth, Mars, Jupiter, Saturn and Neptune rotate in the direct sense. Most satellites rotate directly; rotations for asteroids and comets are known only in a small number of cases. The term is also used for the regular movement of solar system bodies from west to east on the celestial sphere.

Compare: retrograde motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Disconnection Event

A discontinuity in the gas tail of a comet in which the tail appears to break off from the coma, and a new tail begins to grow. A disconnection event occurs when the comet crosses a sector boundary. Entrained within the solar wind is a magnetic field. As the Sun rotates, the field lines are wound into a spiral pattern. The magnetic field has regions of north and south polarity, and the sector boundary is the two-dimensional surface that separates these regions. It too is wound into a spiral, so a comet can undergo several disconnection events in a single apparition if it makes several crosses of sector boundary lines. For example, Halley's Comet underwent 19 disconnection events at its return in 1986.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Disk Galaxies

Stellar disks are the main structural features of disk galaxies, which we divide into the spiral, S0, and irregular morphological classes. The disks are highly flattened, with approximate circular symmetry. In spiral and irregular galaxies, the disk contains gas as well as stars, and fine structure is common, including the spiral arms which define that class. Our nearest disk galaxy is our own MILKY WAY, a spiral galaxy; the Sun lies in the plane of the disk, about 8 kpc from its center.

The stars and gas of a galaxy disk follow near-circular orbits around the center, in the plane of the disk. The disk does not rotate rigidly like a turntable, but differentially: stars in the outer regions take longer to complete an orbit than those close to the center. The orbital motion of the stars and gas supports the disk against collapse under gravity. Organized rotation dominates all other motions; random velocities account for very little of the kinetic energy, so that disks are dynamically 'cold', or at least 'cool'. In the Milky Way, disk stars near the Sun move at about 200 km s^{-1}, taking about 250 Myr to complete an orbit, but their random motions are typically only 30 km s^{-1}. Because the disks are cool, they tend to be unstable to forming internal substructures such as bars and spiral arms.

Giant disk galaxies, with luminosities more than about $610^9 L_{\dot{o}}$ (our Milky Way is roughly three times brighter), are composite systems. The round inner bulge is much denser than the disk, and 'hotter'; the stars have large random motions. Within the bulge is a yet denser star cluster, the galactic nucleus; some nuclei contain massive black holes at their centers. Giant galaxies follow a morphological sequence, named for its originator, Edwin Hubble; see GALAXIES: CLASSIFICATION. The S0 galaxies have large central bulges or spheroids, and a smooth stellar disk; spiral arms, gas and star formation are normally absent. Along the sequence from Sa through Sb to Sc, the central bulge becomes smaller, while the prominence of the spiral arms increases, along with the fraction of gas and newly formed stars.

Dwarf galaxies are smaller, less luminous, and less massive than the giants. They are also more diffuse, with reduced densities of stars and gas, and they lack the dense round central bulges. At the boundary between dwarfs and giants are the Sd galaxies, with very thin stellar disks, and only rudimentary spiral structure. The trend towards less organized optical structure continues through the Sm Magellanic irregular class to the dwarf irregulars (dIrr), which are the least massive and least luminous, with blue luminosities below $10^8\ L_{\dot{o}}$. All of these classes of dwarf disk galaxies are rich in gas, and have relatively large contributions from young stars.

Structures of galaxy disks

Stellar disks

The Milky Way is the only galaxy where the three-dimensional structure is well known; elsewhere, we see only two-dimensional projections of the galaxies. By observing many galaxies, randomly distributed over all possible viewing angles, we deduce the average three-dimensional structures of the various classes. In disk galaxies that we view nearly face-on, perpendicular to the plane containing the disk, the projected brightness of starlight declines smoothly with radius. At a given wavelength λ, the run of intensity $I_\lambda(r)$ with radius r from the disk center, is roughly exponential: $I_\lambda(r) \approx I(0)\, e^{-r/h_r}$. The *radial scale length* h_r is the e-folding length for the starlight integrated vertically through the disk. In most spiral galaxies, h_r is 1–5 kpc; dwarfs can have h_r<1 kpc while in some peculiar giant galaxies h_r>10 kpc.

In most disk galaxies, near the center where $r<h_r$, the luminosity rises above the inward extrapolation of the exponential law describing the main disk. We can define the disk central brightness $I_\lambda(0)$ by extending the disk model to $r=0$; the total luminosity of the disk is simply $L_\lambda = 2\pi h_r^2 I_\lambda(0)$. Measured values of the central disk brightness are commonly around $I_\lambda^*(0) \approx 150\ L_{\dot{o}}\ pc^{-2}$ in blue light; this is the *Freeman peak*. It is now recognized that the Freeman value reflects an approximate upper bound to disk surface brightness. While few galaxy disks have brighter centers, disks exist with lower surface brightness, ranging down to a few percent of the Freeman luminosity density. These are the *low-surface-brightness* galaxies. One caution: although it is usual to extrapolate the exponential law in this way, it is not clear if the stellar disks really extend into the central galaxy.

We see our own galaxy, the Milky Way, as a luminous path across the sky on a dark night; the Sun is near the mid-plane of our disk, so we observe the stellar disk in projection from within. The disks of other galaxies also appear as narrow bands of light when we see them edge-on. If an edge-on disk contains only stars, with no interstellar matter to block light, we can use brightness maps to derive the distribution of stellar density 'above' the mid-plane in the z-direction.

Stellar disks have complicated vertical structures. In our own Milky Way, the youngest stars are concentrated near the disk mid-plane, with exponential *scale heights* h_z<0.1–0.2 kpc (see GALACTIC THIN DISC). Older stars form a 'thick disk' (see GALACTIC THICK DISC) with a scale height $h_z \approx 1$ kpc. The vertical scale height of each component of a stellar disk depends on its velocity dispersion perpendicular to the disk; higher speeds yield thicker disks, since the stars can travel further from the midplane. Young stars are born from clouds of gas that are dynamically 'cold'; they have little vertical kinetic energy, and are confined close to the mid-plane. As they orbit, stars feel the gravitational forces from large lumps of gas and

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

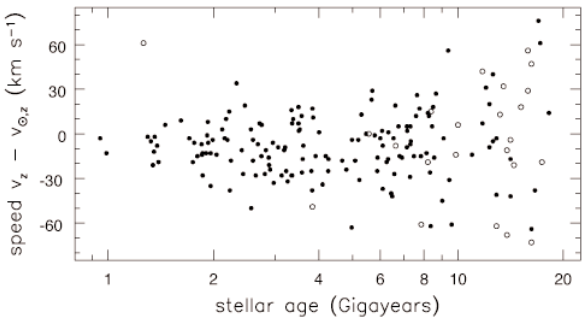


Figure 1. For nearby main-sequence F stars, velocity v_z–$v_{z'\odot}$ perpendicular to the Galactic plane, measured relative to the Sun. Open circles show stars with less than 1/4 of the Sun's iron abundance. Older stars tend to be metal-poor and faster-moving. Data from Edvardsson *et al* 1993 *Astron. Astrophys.* **275** 101.

stars that are present in the spiral arms. The repeated tugs 'heat' them, increasing their random motions both vertically and in the plane of the disk: see figure 1.

In external galaxies, the vertical light distributions is often nearly exponential at large z, but flattens near the mid-plane. The average vertical scale heights derived from fits to the luminosity profiles of giant disks are smaller than the radial scale lengths, typically by factors of ≈5–10. Some 'superthin' Sd galaxies have even larger ratios, while irregular galaxies are typically fluffy: see figure 2. When we observe a galaxy disk edge-on, there is the additional complication that we integrate the light over a range of radii. The projected light distribution along the mid-plane $z = 0$ is not proportional to the surface brightness $I_\lambda(r)$ that we would measure if the galaxy was seen face-on. For example, the radial brightness profile of an exponential disk that is edge-on will be fit by modified Bessel functions, which have a shallower drop-off with radius than the exponential distribution.

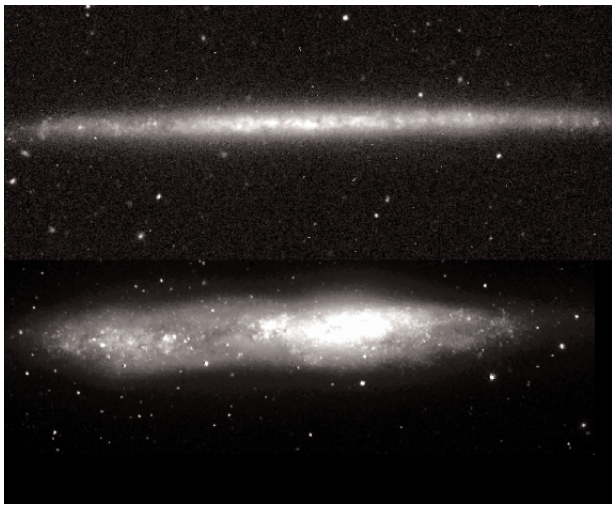


Figure 2. *B*-band images of two very different disk galaxies. Top: 'superthin' Sd UGC 7321, viewed ≲2° from edge-on. This is a small galaxy, with $L_B \approx 10^9\ L_\odot$ (L Matthews, WIYN telescope). Bottom: barred Magellanic or irregular NGC 55, about 10° from edge-on. Its linear size is about half that of UGC 7321, while it is twice as luminous. Note the fluffy disk, and off-center concentration of light (A Ferguson).

Figure 3. *V*-band image of Sa galaxy NGC 4594, the 'Sombrero'. This is a luminous galaxy ($L_V \approx 8$–$10^{10}\ L_\odot$) with a very large bulge, and numerous globular star clusters around it (A Cole, WIYN telescope).

Gas disks

Except in the S0 galaxies, the disk contains gas as well as stars. The interstellar medium of a disk galaxy like the Milky Way is richly varied. It includes phases ranging from molecular cloud complexes with densities of ~10^2 particles cm^{-3}, to diffuse highly ionized gas with densities of ~10^{-2} cm^{-3} and temperatures >10^6 K. The gas is also threaded with energetically significant magnetic fields, and bathed in the cosmic rays trapped by the field. Interstellar matter is perhaps most readily seen as dark clouds or lanes, where dust grains associated with the gas block our view of the stars behind. In the southern and northern Milky Way, dust is apparent as individual dark clouds, such as the 'coal sacks', along the plane of the bright disk. In optical pictures of edge-on disk galaxies such as figure 3, dust in the disk causes the characteristic dark band across the mid-plane. See also INTERSTELLAR MATTER.

Interstellar dust grains (see INTERSTELLAR GRAINS) absorb and scatter starlight at wavelengths from the near infrared to soft x-rays. Interstellar dust grains are small—most have diameters between 0.001 μm and 1 μm—and dust opacity rises rapidly with decreasing wavelength. We get our best views of the distribution of stars in

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

spiral galaxies in the infrared spectral region at λ~2 μm, the longest wavelengths where direct stellar radiation is still important. In the thermal infrared region, at wavelengths 5–300 μm, we see radiation from the dust grains themselves, warmed by the starlight that they have absorbed; most of the power is in the far infrared (FIR), at λ>50 μm. Images of galaxies at FIR wavelengths show emission where there are stellar (or other) heating sources, and the dust absorbs their light efficiently. The most spectacular star-forming disk galaxies, the STARBURST GALAXIES such as M82, contain so much dusty gas that it blocks starlight even at infrared wavelengths. There, the distributions of young stars can only be mapped from the FIR emission, and the radio emission of ionized gas near massive stars, or of the young remnants of supernova explosions.

A typical giant spiral galaxy has about 5–10×10^9 $M_\odot$ of cool interstellar gas. About half of this is clumped into dense regions, close to the midplane of the disk, where hydrogen is in molecular form (H_2). Young stars are born in these cool molecular clouds. Since H_2 is a symmetric molecule, it produces no strong emission, and its ultraviolet absorption bands become difficult to observe in dense regions of the interstellar medium. The locations and amount of H_2 are usually derived from observations of tracer molecules which do produce radio frequency line emission, such as CO or HCN. Molecular gas is usually concentrated in the inner galaxy, where it forms a very thin disk, with scale height h_z<0.1–0.3 kpc.

The remainder of the interstellar gas is more diffusely distributed; most of it is neutral (H I) or ionized (H II) hydrogen. The spatial distribution of the H I in disk galaxies is the most easily measured, through the hyperfine 21 cm emission line. This component is usually quite thin with vertical scale height h_z<1 kpc, and a radial distribution that extends further from the center than the stellar disk. The outer regions of most galaxy disks appear to consist mainly of H I gas. Warm ionized gas, within 1–2 kpc of the midplane, is traced from the recombination emission lines of hydrogen. The disk is surrounded by a much hotter diffuse halo of gas at 10^5–10^6 K; we detect it in x–ray emission, or by ultraviolet absorption spectroscopy of resonance lines produced by heavier elements within the hot gas.

The first unambiguous evidence for a giant halo of hot gas around a nearby spiral galaxy much like our own Milky Way was found by astronomers using NASA's CHANDRA X-RAY OBSERVATORY. Chandra found a diffuse halo of x–ray gas, radiating at a temperature of almost 3 million degrees, around galaxy NGC 4631, approximately 25 million light years from Earth. The Chandra image reveals a halo of hot gas that extends for approximately 25 000 light years above the disk of the galaxy. One important feature of the x–ray emission from NGC 4631 is that it closely resembles the overall size and shape seen in the radio emission from the galaxy. This indicates that there may be a close connection between the outflows of hot gas, seen in x–rays, and the galaxy's magnetic field, revealed by radio emission. A Hubble image of NGC 4631 shows filamentary, loop–like structures enclosing enhanced x–ray–emitting gas and emanating from regions of recent star formation in the galaxy's disk. Together, these data clearly show that the hot gas is heated by clusters of massive stars and is now expanding into the halo of the galaxy.

The interplay between the gaseous and stellar components of disks is extremely complicated. A gaseous disk experiences dissipation, and can radiate away the energy associated with random motions, so that it tends to collapse in the vertical direction. However, energy and momentum are fed into the interstellar medium, by fast stellar winds, ionizing radiation, and supernova explosions from massive stars. These increase the internal energy of the gas, causing it to expand vertically; in extreme cases it can even escape as a galactic wind. Since stars are made from dense cool gas, star formation can be

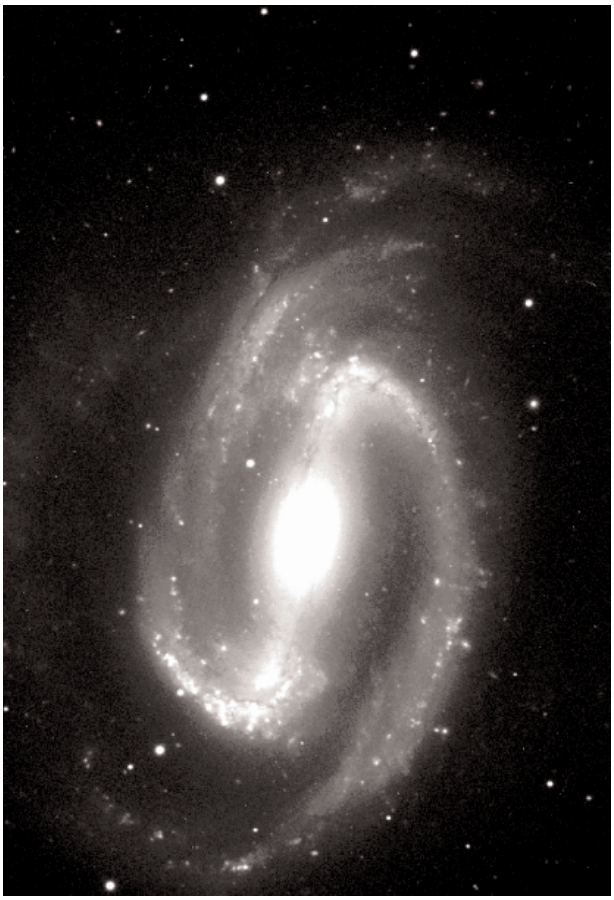


Figure 4. The barred galaxy NGC 1300, classified SBb or SBbc, with 'grand design' spiral arms. The arms trail; their tips point opposite to the galaxy's sense of rotation. Note the dust lanes on the leading edge of the bar, and on the concave sides of the spiral arms (WIYN telescope).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

a self-regulating process. If many young stars are made, they can blow the gas away, or heat it so much as to prevent further starbirth. See also GAS IN GALAXIES.

Spiral arms and bars

The photogenic arms of spiral galaxies are the most striking luminous structures in any galaxy atlas. In a minority of giant galaxies, such as M81 or NGC 1300 (see figure 4), we see a 'grand design' spiral: a pair of well-defined arms winds over azimuth angles of more than 90° and extends in radius over much of the optically visible stellar disk. More frequently, the spiral arms are broken up into shorter segments, and more than two arms are present. In 'flocculent' disks, the spiral consists of many small arm fragments; there is no organized spiral structure on large scales. Gas seems to be essential to forming spiral arms; the S0 galaxies lack both gas and spiral pattern. Strong spiral arms form most readily in the rapidly rotating massive disks of luminous galaxies. Lower luminosity (and less massive) galaxies have messy spiral patterns, or none. The Sd systems often have ill-defined arms, while spiral arms are totally absent in the low mass dIrr galaxies. The decline in spiral arm intensity and frequency with decreasing galaxy mass suggests that gravity is important in shaping the spiral structure.

To the surprise of astronomers, the galaxy NGC 4622 appears to be rotating in the opposite direction to what they expected. Astronomers are puzzled by the clockwise rotation because of the direction the outer spiral arms are pointing. Most spiral galaxies have arms of gas and stars that trail behind as they turn. But this galaxy has two leading outer arms that point toward the direction of the galaxy's clockwise rotation. To add to the conundrum, NGC 4622 also has a trailing inner arm that is wrapped around the galaxy in the opposite direction it is rotating. Astronomers suspect that NGC 4622 interacted with another galaxy. Its two outer arms are lopsided, meaning that something disturbed it. The new Hubble image suggests that NGC 4622 consumed a small companion galaxy. The galaxy's core provides new evidence for a merger between NGC 4622 and a smaller galaxy. This information could be the key to understanding the unusual leading arms.

Spiral arms can be characterized by their pitch angle, the angle at which the arm cuts through a circle around the galaxy center. Arms in Sa galaxies are tightly wound, with small pitch angles, while those in later-type spirals are more open, with pitch angles of 10–30°. Spiral arms generally trail, with the tips of the arms pointing opposite to the direction of galactic rotation. In figure 4, we see dark dust lanes on the concave sides of the spiral arms; these show where the disk gas is compressed as it flows into the arms. Thus the spiral stands out when observed in H I, or tracers of molecular gas such as CO emission. The bright young stars which make the spiral arms so prominent have been formed from this compressed gas.

Massive stars have short lives; those hot enough to ionize the gas around them last less than 5 Myr, and lifetimes are generally <30 Myr. This is only about 10–20% of a galactic rotation period. So the hot stars that we see in the grand-design spiral arms must have been born in or near them; if they were made elsewhere in the disk, they would die before they could be concentrated into the spiral arms. A classic problem is whether spiral arms stimulate star formation, or simply act to organize it. In the former case, we would expect star formation to be more vigorous in galaxies with grand-design spiral patterns than in flocculent galaxies. But observational comparisons show no significant difference in the rates; the arms seem to act more as star formation organizers.

The young stars in spiral arms radiate 10–100 times as much power per unit mass as older stellar populations, in which the most massive stars are no longer present. Even small fractions of young stars can have major observable consequences. These effects are most pronounced in blue and ultraviolet light, where these hot stars produce most of their luminosity. Young massive stars also photoionize the gas from which they were born, producing brilliant ionized nebulae, known as H II regions. The H II regions are excellent tracers of spiral structure, since they are easily observed in their bright atomic emission lines, from hydrogen recombination and collisionally excited lines of common elements such as nitrogen, oxygen, and sulfur.

About half of all disk galaxies show a central linear bar, containing up to a third of the total light. The ratio of the long to the short axis of the bar can be as extreme as 1:5, as in figure 4. Gas is not required for a bar: the disks of S0 galaxies are as likely to be barred as are the gas-rich spiral galaxies. In Hubble's classification, 'B' is added to indicate a bar; NGC 1300 is of type SBb. Bars are the offspring of stellar disks; they appear to be vertically thin, like the disk, and not round like the bulge. The main bars of giant galaxies typically extend to radii of a few kiloparsecs.

Bulges and nuclei

In less luminous Sc and Sd galaxies, and some dwarfs, the approximately exponential disk extends to the center of the galaxy; for these pure disks, the central surface brightness is simply $I_\lambda(0)$. Low-luminosity galaxies are often asymmetric, with no well-defined center; e.g. the brightest region may not lie at the kinematic center of the system. The Large Magellanic Cloud is a prime example of an off-center disk galaxy; here, the center of rotation is outside the bright stellar bar. The irregular galaxy NGC 55, in figure 2, also shows an off-center concentration of light.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

By contrast, most giant disk galaxies are reasonably symmetric. Extra light is also present in the inner few kiloparsecs, above that expected from the main disk. This is the central *bulge*: a spheroidal stellar system normally consisting of stars with ages ≧3 Gyr (see GALACTIC BULGE). Bulges typically contain little gas, except near their centers. Spheroidal bulges are rotating, but the large random stellar motions allow them to be thicker in the *z*-direction, perpendicular to the disk midplane, than the surrounding stellar disk. The 'Sombrero', shown in figure 3, is an outstanding example of a disk embedded in an unusually large oblate stellar bulge. Smaller bulges, like those of the Milky Way or our neighbor Sb galaxy M31, are much more common. By contrast, some spirals have inner disks. Like bulges, these small disks have much higher surface brightness than the main stellar disk; but they are relatively flat, roughly axisymmetric, and dynamically 'cooler' than spheroidal bulges. Inner disks often support star formation and have their own spiral structure.

Images of the centers of disk galaxies taken at high angular resolution reveal a variety of structures. In some galaxies, small nuclear bars are present, often nested within a larger main bar; these have sizes of <1 kpc. The available evidence suggests the nuclear bars are independent entities, that do not rotate at the same rate as their main bars. At or near the very center is the *nucleus*, a tiny (r<10 pc) object containing the highest density of stars in the galaxy, up to $10^6\ L_{\odot}\ \mathrm{pc}^{-3}$ (see GALACTIC NUCLEUS). Sometimes a massive black hole, with M_{BH}>$10^6\ M_{\odot}$, is also present as well. When gas or stars are swallowed by the black hole, huge amounts of gravitational energy can be released, producing a luminous active galactic nucleus (AGN) that can compete with the power output of all of the stars in the galaxy. However, most nuclei are not AGNs, and derive much of their luminosities from the stars they contain. As a result they are relatively faint, with optical luminosities of 10^6–$10^8\ L_{\odot}$.

Dynamics

The stars of a galaxy move under gravity; other forces are rarely important. In the disk, stars are widely separated in space, so we have the additional simplification that close collisions between them are rare, and can be safely ignored. Thus their motions can be studied as a response to the galaxy's mean gravitational field, which must be computed self-consistently from the mass distribution; this is a challenge, since the system has ~10^{11} particles (see also STELLAR DYNAMICS). By contrast, gas responds both to gravity and to hydromagnetic forces. Since dense gas clouds do collide, they can dissipate the energy of random motions. Because of this, cool disk gas rapidly settles onto closed periodic orbits about the galaxy center; in an axisymmetric disk, these are the circular orbits.

Rotation and mass distribution

We can use Doppler-shifted emission lines from the gas to find the mass of a disk galaxy. If we measure the speed $V(r)$ of gas in a circular orbit at radius r, we can use the radial-force equation $V^2(r)/r = GM(<r)/r^2$ (which is exact in a spherical system) to make a rough estimate of the mass $M(<r)$ within that radius. In giant disk galaxies, $V(r)$ climbs steeply over the inner kiloparsec or so, to a maximum that is typically at V_{max}~150–300 km s^{-1}, and then stays roughly constant. In dwarfs, the rise is more gradual and peak speeds are lower.

We can compare the observed rotation curves of disk galaxies with what we would expect if their mass had been concentrated entirely in the stars and gas. Even when the assumed ratio of mass to light in the disk is adjusted so that gas and luminous stars account for as much of the galaxy's rotation as possible, we would expect the rotation speed to fall beyond 2–3 times the scale length hr of the exponential stellar disk. Instead, we usually see that $V(r)$ remains approximately constant to the edge of the gas disk, implying that $M(<r)$ continues to rise; the outer parts of the galaxy contain much mass but emit little light. A spherical halo of *dark matter* is generally invoked to provide the required extra mass. Dark matter usually accounts for between 50% and 90% of a disk galaxy's total measured mass, with dwarf galaxies generally having a higher proportion than giants. See also DARK MATTER IN GALAXIES.

More luminous galaxies rotate faster on average, which tells us that they are more massive. The peak rotation speed V_{max} increases with the galaxy's luminosity L, roughly as $L : V_{\mathrm{max}}{}^{\alpha}$, with α~4: this is the *Tully–Fisher relation*. If galaxies contained no dark matter, we might be able to understand the TULLY–FISHER RELATION. But since the speed V_{max} is set largely by the unseen material, while the luminosity comes mainly from the stellar disk, the link between them is puzzling. Nevertheless, the Tully–Fisher relation is very useful in finding distances to galaxies and galaxy groups, since we can estimate the true luminosity of a galaxy from its measured rotation speed.

Disk stability

Galaxy disks live on the edge of gravitational instability; galactic bars, and the spiral arms, are the consequences. The mutual gravitational attraction of disk stars and gas clouds tends to pull them together. On the other hand, since the period of an orbit around the galaxy center increases with radius, the disk's rotation tries to shear any feature into a trailing spiral arm segment. In a 'cool' disk, where the random motion of the stars is small enough that it does not take them outside the spiral arms, their gravity can reinforce a spiral pattern by attracting other stars into it, and so help the spiral to grow. But if the disk contains little mass, or its stars have large random motions, then gravity will be

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

insufficient to amplify a spiral, and no strong arms will develop.

Locally, an axisymmetric disk is unstable, and a spiral will grow, if the Toomre 'Q' parameter $Q = \kappa\sigma_R/3.36G\Sigma \lesssim 1$. This criterion specifies that the dispersion σ_R of stellar velocities in the radial direction must be small enough by comparison with the surface density Σ of mass in the disk; κ is the epicyclic frequency with which a star oscillates radially about the nearest circular orbit, and G is the gravitational constant. Computer simulations of a 'cool' axisymmetric disk of 'stars' attracting each other by gravity, and initially following nearly circular orbits, show that a spiral pattern generally grows if $Q \lesssim 1$. As it does so, the stars develop larger random motions, and Q rises; so we never expect to see a stellar disk with $Q<1$. Near the Sun, the density in the disk $\Sigma \sim 50$ $M_\odot$ pc^{-2}, and $\kappa \approx 35$ km s^{-1} kpc^{-1}, while for stars about as old as the Sun, the velocity dispersion $\sigma_R \approx 30$ km s^{-1}. Hence $Q \sim 1.4$, which is safely greater than unity.

The simulated disks generally grow a two-armed spiral, and often a straight central bar in addition. The 'stars' stream through the spiral, slowing down and lingering in its gravitational potential well, and then pass on through it; the spiral is a *density wave*, a stellar traffic jam. We believe that the same is true in real galaxies. The pattern of the arms turns in the same sense as the disk's rotation, at some angular rate Ω_p, which is set by the angular frequency of rotation $\Omega(r)$ and the epicyclic frequency $\kappa(r)$ in the region where the spiral is strong.

Why do the arms of a spiral galaxy trail, pointing against the direction of galactic rotation? When a trailing spiral is present, gravitational forces cause the inner parts of the disk to exert a torque on the outer disk, which transfers angular momentum outward and allows material at small radii to move inward. The spiral torques decrease the energy of the disk's rotational motion, transferring it to increase the random speeds of the stars. As the disk 'heats up', the spiral eventually disappears. The disk of an S0 galaxy, where no gas is present, would probably behave in the same way: any spiral pattern would be short lived. But in a gas-rich disk, stars freshly born from the cool gas have very small random speeds. Continued addition of these new stars is probably important in prolonging the life of a spiral pattern, or in re-creating it periodically.

The gravitational pull of the spiral arms affects gas even more strongly than the stellar disk, because the random speeds of gas clouds are only 5–10 km s^{-1}, much lower than for the stars. Except immediately around the corotation radius, where the angular speed $\Omega(r)$ of a circular orbit is close to Ω_p, the linear speed $r[\Omega(r)-\Omega_p]$ with which gas moves into the spiral arm is supersonic. Shocks develop, compressing the gas enormously as it flows into the arm; we see dark dust lanes there. After about 10 Myr, young stars are born in the compressed gas and begin to shine; the Hα emission from gas ionized by these stars is not concentrated on top of the dust lane, but 'downstream' of it. Radiation from the hottest of these stars also splits some of the H_2 molecules apart into atomic hydrogen, H I.

Like spirals, the figure of a galactic bar is not static, but rotates with some pattern speed Ω_p. Unlike spiral arms, a bar truly captures its stars. Within the bar, stars and gas no longer follow near-circular paths, but are trapped near elongated orbits that close on themselves as seen by an observer moving with the bar's rotation. Disk gas cannot flow in toward the galaxy center unless it can get rid of its angular momentum; the strongly asymmetric gravitational forces of a bar help it to do just that. The closed oval orbits converge on each other near the ends of the bar. As gas on the orbits approaches those regions, shocks form. The gas is compressed, and we see dark dust lanes, like those along the leading edge of the bar in figure 4. In the shock, the gas loses part of its energy of forward motion as heat, so it drops down onto more tightly bound orbits closer to the center of the galaxy. Bars may be important in feeding gas in towards the central galaxy.

History

Near the Sun, the oldest disk stars are about 10 Gyr old. By examining the number of stars of each mass that are present today, we can deduce that the Milky Way's disk has been making new stars at a roughly constant rate over the last few gigayears. We live in an Sc (or Sbc) spiral galaxy; the disks of Sd galaxies, and the irregulars, are bluer than ours, implying a larger proportion of newly formed stars. Either their rates of star formation have increased recently, or they simply began their star-forming histories later. The redder colors of Sa and S0 galaxies tell us that star formation in their disks is decreasing; by now, most or all of their gas supply is used up.

Near the Sun, figure 1 shows that younger stars contain a larger proportion of elements heavier than helium. The disk has formed successive generations of stars, each producing metals to contribute to the already enriched interstellar gas. In all but the lowest-luminosity disk galaxies, we see a color gradient; the centers are redder than the outer parts. This is because the inner parts contain both a larger proportion of old stars, and stars that are richer in heavy elements. The bulges and inner disks have already made many generations of stars, while the outer disks have lagged behind, producing their stars more slowly, and remaining gas-rich to the present day. See also GALAXY EVOLUTION.

Galaxy formation: inside-out or outside-in?

How might a galaxy like our own Milky Way come into being, out of the hydrogen and helium gas resulting from the Big Bang? Some would argue that giant disk galaxies

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

have grown outward from their central parts (see GALAXY FORMATION). The dense inner bulge, with its low specific angular momentum, formed early on, from a denser-than-average lump of gas that collapsed under its own gravity; it rapidly used up all its gas in forming stars. The disk material, which has more angular momentum per unit mass, would have fallen in later, from greater distances. It must have remained gaseous until almost all its random motion had been dissipated, allowing the material to move inwards and settle onto circular orbits in a thin disk. Only then could the oldest disk stars have been born. Perhaps the disks are still being assembled today, as clouds of cool gas fall in, renewing the supply of raw material to make new stars. The dark matter, unable to dissipate energy, could not fall inwards; hence the dark halo is less concentrated to the galaxy center than are the luminous stars and gas.

Other astronomers point out that the stars of the bulge do not seem to predate the oldest disk stars near the Sun. They offer an alternative idea: the disk formed first, and the bulge grew within it. The disk gas would have flowed slowly inwards, perhaps coaxed towards the center by spiral arms or a galactic bar. When the inner part of the disk had reached a sufficient density, it would have become unstable and formed a bar, which then puffed up vertically into a bulge. This model has received some support from computer simulations. Also, in some bulges the gas is seen to follow oval orbits similar to those in galactic bars, as expected if these bulges are in fact thickened bars. The 'inside-out' enthusiasts retort that the bulge of a galaxy like the Sombrero (figure 3) is much more luminous than the disk, and so is unlikely to be an outgrowth from it. There may be no single route to making a disk galaxy.

Bibliography

Binney J and Merrifield M 1998 *Galactic Astronomy* 3rd edn (Princeton, NJ: Princeton University Press)

Sparke L S and Gallagher J S 2000 *Galaxies in the Universe: an Introduction* (Cambridge: Cambridge University Press)

Jay Gallagher, Linda Sparke and Lynn Matthews

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Dispersion

In optics, the spreading of a mixture of wavelengths of light (colors) into its consituent wavelengths. The refractive index of a medium (e.g. glass), which determines the extent to which a ray of light is deflected when it enters that medium at any angle other than perpendicular to its surface, is the ratio of the speed of light in a vacuum to the speed of light in a medium. Because different wavelengths of light travel at different speeds in a medium (they all travel at the same speed in a vacuum), the refractive index of a medium is wavelength dependent. Dispersion is a measure of the extent to which refractive index varies with wavelength.

When white light (a mixture of wavelengths) passes through a glass prism, the different wavelengths are refracted by differing amounts and are spread out into a band of wavelengths (a 'rainbow' band of colors), which is called a spectrum. The dispersion of a spectrograph (a device for producing a spectrum) is a measure of the extent to which the various wavelengths are spread out at the focal plane of the instrument. For a particular operating wavelength, λ, dispersion $= \delta\lambda/\delta x$, where $\delta\lambda$ is the change in wavelength over a distance δx along the length of the spectrum. It is usually expressed in nanometers (nm) of wavelength per millimeter (mm) of length (i.e. nm mm^{-1}). The dispersions of astronomical spectrographs usually lie in the range from about 200 nm mm^{-1} to about 0.01 nm mm^{-1}, the lower dispersions (more than a few tens of nanometers per millimeter) generally being useful only for survey work.

The wavelength dependence of refractive index (i.e. dispersion) is also responsible for preventing a simple lens from focusing all wavelengths of light at the same point, the spread of focal points for different wavelengths being called chromatic aberration. Dispersion also arises in the phenomenon of diffraction (different wavelengths are diffracted by different amounts) and when light of different wavelengths is scattered from small particles (the shorter the wavelength, the greater the degree of scattering).

See also: chromatic aberration, diffraction, diffraction grating, refractor, spectroscope/spectrograph, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Distances (from Antiquity to 1900)

Ancient views on the scale of the planetary system

Before the dissemination of Greek culture in the 'Hellenistic' period that followed the conquests of Alexander the Great in the late 4th century BC, notions of astronomical distances were purely speculative. Although ancient cultures developed accurate surveying methods, the mathematical techniques required for astronomical distance measurement came only with formal Euclidian geometry.

For the Greeks, who believed the Earth to be spherical and at rest at the center of the cosmos, the baseline for distances within the planetary system was the Earth's radius. ERATOSTHENES OF CYRENE (*ca* 276–195 BC) devised a method and made the first measurement of the Earth's dimensions (see figure 1). He knew that at the summer solstice the Sun cast its rays directly down a well in the Egyptian town of Syene, while at Alexandria, further north and on roughly the same meridian, the sunlight falling on a gnomon cast a shadow of 1/50 of a circle. Assuming that Syene was 5000 stades south of Alexandria, and knowing that distance to be 1/50 of the Earth's circumference, Eratosthenes calculated the circumference to be 250 000 stades. Depending on the length of a stade—there is uncertainty about which value he employed—he may have been within 8% or even better of the modern value.

ARISTARCHUS OF SAMOS (*ca* 310–230 BC) provided two elegant geometric demonstrations concerning the scale of the cosmos; both would remain central to astronomical thought until the 17th century. First, Aristarchus recognized that the relationship between the Sun's and Moon's positions at quadrature provides a means to calculate their relative distances from Earth (see figure 2). For then the angle Sun–Moon–Earth is 90°, and so if the observer measures the angle Moon–Earth–Sun, then the shape of the triangle will be known, and this will give the ratio of the lengths of the Earth–Moon and Earth–Sun distances.

Aristarchus adopted 87° for the second angle and concluded that the Sun was some 19 times more distant than the Moon. There are two practical problems with this method: it is extremely difficult to know the moment of quadrature by observing when the terminator is precisely at the half-way point on the Moon's face and it is no easy task to observe the Moon–Earth–Sun angle at that time. In fact, Aristarchus's value of 87° may have been to illustrate his method rather than the result of actual measurement.

Aristarchus's second demonstration used eclipse geometry to determine the absolute sizes of the Moon and Sun. During a total solar ECLIPSE, Sun and Moon subtend the same angle, so the ratio of their apparent diameters is the same as the ratio of their distances, which he believed to be 19:1. To obtain absolute distances, he analyzed the situation during a total lunar eclipse, assuming the width of the shadow cone at the Moon's distance to be twice the lunar diameter. From his analysis, he concluded,

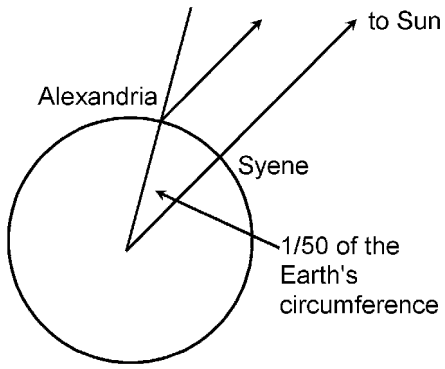


Figure 1.

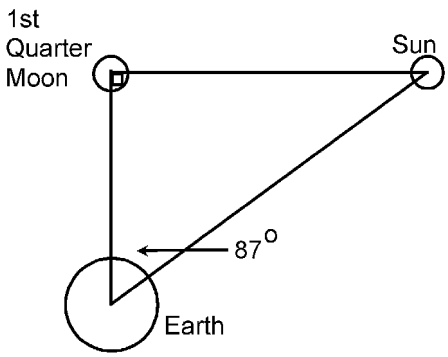


Figure 2.

in mean terms, the Moon's diameter to be about one-third that of the Earth, and the Sun's some $6\frac{3}{4}$ times the Earth's. Curiously, Aristarchus does not calculate the absolute distances in his treatise, but, given a lunar apparent diameter of $\frac{1}{2}^\circ$, and since this was two-thirds the radius of the Earth, its distance was some 80 Earth radii; the Sun's distance (19 times greater than the Moon's) was therefore about 1520 Earth radii.

Hellenistic astronomers recognized the importance of the Sun's PARALLAX, the angle subtended by the Earth's radius as seen from the center of the Sun. Ideally, one could determine this by viewing the Sun from two distant locations at the same time, or by observing from the same location several hours apart. Since the Earth's radius was known, the absolute distance to the Sun could be calculated if this angle were determined. However, measuring parallax by direct observation of the Sun is impossible, as the Sun's atmosphere does not provide a distinct 'edge', and the angle is so small that naked-eye instruments would be useless. HIPPARCHUS of Rhodes (*fl* 141–127 BC), the greatest influence on Ptolemy, developed alternative methods. Using solar eclipse observations taken from two locations, he calculated the lunar parallax, giving him minimum and maximum distances to the Moon in Earth radii. By a somewhat different employment of Aristarchus's eclipse diagram, and *assuming* a solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

parallax of 7 arcmin as the smallest parallactic angle that might be visible, he found a slightly smaller absolute lunar mean distance, just over 67 Earth radii.

PTOLEMY (*fl* AD 127–141) knew of Hipparchus's methods, but attacked the problem in a different way. Studying lunar eclipse data, he was able to obtain the lunar parallax and a (maximum) distance of a little more than 64 Earth radii. By using the eclipse diagram, the solar parallax could then be calculated, which allowed the Sun's absolute distance to be determined. This came to 1210 Earth radii (giving a parallax of nearly 3 arcmin), very close to the 19 times the lunar distance that Aristarchus had calculated. This has suggested to some modern historians that Ptolemy chose his values carefully to obtain the required result. In any event, he could not have observed solar parallax directly, so while his lunar distance is reasonable, his solar distance is an order of magnitude too small.

The order and sizes of the planets

Greek astronomers were largely agreed on the order in which the planets lay in their heights above the Earth. The Moon, being so large and having the shortest SIDEREAL PERIOD, was the closest object. The Sun, although it subtended approximately the same visual angle, took a year to complete its circuit of the sky, so was further away. Mars, Jupiter and Saturn had increasingly longer sidereal periods, so it seemed reasonable to locate them beyond the Sun and in that order, while the stars lay on the outermost sphere. However, since Mercury, Venus and the Sun all have a sidereal period of 1 yr, there was no compelling reason to order them in any particular way. In Ptolemy's *Almagest* the motion of each planet was represented by its own geometrical model, and so in this work Ptolemy did not need to commit himself to any particular order, but, in his later *Hypotheses of the Planets*, he argued for the order Moon–Mercury–Venus–Sun. He also provided a physical picture of the universe similar to the concentric material spheres of Aristotle. Believing the cosmos to be a plenum (a space filled with matter as opposed to a void), he held that the outer surface of one planetary sphere coincided with the inner surface of the next sphere, so that the cosmos was a nesting set of such spheres. Thus he took the lunar apogee to be equal to the perigee of Mercury, and this distance was the radius of both the outer boundary of the lunar sphere and the inner boundary of the sphere of Mercury. Mercury's apogee, similarly, was equal to the perigee of Venus, and so on.

The ratio of the perigee to the apogee of any particular planet could be calculated from its geometrical model. Ptolemy's nesting spheres therefore allowed him to derive absolute distances for every planet if he could calibrate his scheme by knowing the actual value of any one of the parameters. As he had already calculated that the Moon's apogee distance was just over 64 Earth radii, the remaining distances followed. The Sun's mean distance was 1210 Earth radii, or a little less than 8 million kilometers using the modern value for the Earth's radius. The outer boundary of Saturn's sphere was 19 865 radii.

Ptolemy similarly provided the angular and actual diameters of the Sun, Moon and planets, although the angular diameters of the planets were in fact subjective and based essentially on their perceived brightnesses. Saturn, for example, was said to have an angular diameter 1/18 that of the Sun when the planet was at its mean distance. As this mean distance was known, it followed that Saturn's actual diameter was about 4.3 times that of the Earth.

Since the 'fixed' stars were immediately beyond the outer boundary of Saturn's sphere, they were some 20 000 Earth radii above the Earth. Ptolemy estimated the apparent diameters of the first-magnitude stars to be 1/20 that of the Sun, and from this he calculated their volumes to be over 94 times that of the Earth.

Islamic astronomers did little more than tinker with these values (see also ISLAMIC ASTRONOMY). Al-Farghani for example in the 9th century estimated the apparent diameters of stars of lesser magnitudes, and concluded that a sixth-magnitude star had a volume 18 times that of the Earth.

Latin translations of Arabic versions of Ptolemy and his Muslim commentators became widely known in western Europe during the 12th and 13th centuries. While few European astronomers of the time could deal with the complexities of the *Almagest*, most educated people were aware of the nesting sphere arrangement and its distance scale.

Distances in the Copernican system

The heliocentric model of the planetary system, published in 1543 by NICOLAUS COPERNICUS (1473–1543) in his *De revolutionibus orbium coelestium*, established by elementary geometry not only a unique ordering of the planets but their (broadly correct) relative distances, in multiples of the Earth–Sun distance, later termed the astronomical unit. In Ptolemaic astronomy, the sphere of the Moon served to calibrate the entire system, but in Copernican astronomy lunar motion was now referred to the Earth, and was no longer part of the planetary order. However, Copernicus was satisfied with the ancient value of the solar parallax, 3 arcmin, and as a result the radius of orbit of the outermost planet, Saturn, was for him little more than half the size it had been for Ptolemy.

There was no observational compulsion to prefer Copernicus over Ptolemy. On the contrary: if Copernicus was correct, then one should be able to detect annual or stellar parallax, the apparent movements of the stars that one would expect to see if they were being observed by astronomers located on an Earth that was moving in annual orbit, and this no one succeeded in doing. There were two possible inferences: either Copernicus was mistaken or the stars were far more distant than had previously been thought, so distant that existing instrumentation was insufficiently sensitive to detect this parallax. TYCHO BRAHE (1546–1601) was to carry pre-telescopic instrumentation to near its limits of accuracy,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and he estimated that his failure to detect annual parallax meant that it would have to be less than 1 arcmin and so—if Copernicus was right—the stars would be at least 700 times further than Saturn; this being so, even a third-magnitude star would have a volume 200 times that of the Sun. To have a universe at whose center was a solar system consisting of relatively tiny bodies, surrounded by an enormous and totally empty space before one came to stars of colossal size at its periphery, made no sense. Tycho's own geoheliocentric system, by contrast, retained Copernicus's relative geometry for the planets and the ancient value for solar parallax. It located the stars at a mere 14 000 Earth radii, and he estimated their volumes at between 68 Earth volumes for a bright star and 1/3 for a faint one.

In 1610 GALILEO (1564–1642) turned the telescope into an astronomical tool, although it was to be some decades before measuring instruments were supplied with telescopic sights. KEPLER's third law made it possible for the relative distances of planets from the Sun to be determined, in astronomical units. If therefore a planetary distance could be measured directly by triangulation, the astronomical unit could be calculated. It was Kepler himself who suggested this approach, using Mars, Mercury and Venus as targets. Tycho had attempted to measure the parallax of Mars, but Kepler's analysis of these data suggested a parallax no greater than 4 arcmin, which would reduce solar parallax to 2 arcmin. Taking into account corrections for refraction, Kepler later reduced this figure by half (see KEPLER'S LAWS).

The possibility that planets might be seen to TRANSIT across the face of the Sun had been recognized since Ptolemy's time, but was only now taken seriously. Kepler, whose own interest was primarily in planetary apparent diameters (as he believed planetary volumes were proportional to their distances from the Sun), correctly calculated that both Mercury and Venus would transit in 1631, on dates separated by 1 month. In the event Venus was not observed as it had already set for European locations, but three astronomers managed to obtain observations of Mercury. One successful observer, PIERRE GASSENDI (1592–1655), was surprised to find Mercury's image so small, much smaller than Kepler and others had expected. This fuelled the debate over planetary sizes and their relationship to distance, but contributed nothing to the measurement of parallax.

JEREMIAH HORROCKS (1618–1641), although a believer in Kepler's elliptical orbits, realized that planetary distances and sizes based on harmonies were incompatible with Gassendi's observation. In recomputing Venus's orbit, Horrocks discovered that it would transit the Sun in 1639. His observations showed a disk of about 1 arcmin. By invoking a dubious proportionality argument, Horrocks claimed that, as seen from the Sun, all planetary disks (including the Earth's) would appear to have an apparent diameter of 28 arcsec. This, he said, implied a solar parallax of 14 or 15 arcsec, which increased the Sun's distance enormously.

Horrocks's writings did not appear until many years after his death, and so had little impact, but several of his contemporaries were also of the opinion that the traditional value for the solar parallax was too large. G-B RICCIOLI (1598–1671), whose *Almagestum novum* of 1651 was the last major geocentric work, provided better observational data on planetary diameters and refraction, and a smaller solar parallax. However, Riccioli and some of his correspondents were still trying to build a distance scale observing lunar dichotomy. Better telescopic observations were required, and these came only with the invention of the micrometer. Although discovered accidently in England around 1640, it was not generally known until the 1660s. CHRISTIAAN HUYGENS (1629–1695) employed a crude micrometer to improve on the observations of his predecessors, and the instrument was further developed by Adrien Auzout (1622–1691), Pierre Petit (*ca* 1596–1677) and others. JOHN FLAMSTEED (1646–1719), the first Astronomer Royal, used the screw micrometer to good effect.

Now armed with telescopes superior in their ability to measure small angles, astronomers could attack planetary parallaxes anew. G-D CASSINI (1625–1712), when professor in Bologna, became interested in the question of how one finds the Sun's position while correcting for both parallax and refraction. If the solar parallax were quite small, then refraction at various altitudes above the horizon should be fixed throughout the year, but if parallax were larger, then one required variable refraction to account for the observations. This had been a serious issue for Tycho who, through assuming a solar parallax of 3 arcmin, required several different refraction tables, although he thought refraction negligible at altitudes above 45°. When Cassini arrived in Paris in 1669, members of the Académie Royale des Sciences debated this issue keenly. Observations in the tropics, when compared with those in Europe, might be decisive. As Mars was to be in a favorable opposition in 1672, Cassini urged an expedition to the colony at Cayenne, close to the equator. Jean Richer (died 1696) was dispatched, while Cassini made observations in Paris.

The results from Cayenne, while showing that refraction correction does indeed require only a single table, could not provide a parallax value that was unequivocal. Cassini found several values for Mars' parallax around 24 arcsec. Flamsteed, in England, was able to make a single observation with a micrometer later that year; although giving several versions of his results, by 1673 he claimed a horizontal parallax for Mars of about 25 arcsec, which implied a solar parallax of 10 arcsec. Instrumental errors were such that one could not rely on these measurements and the computed value for solar parallax, but it was now evident that solar parallax was truly small, less than 15 arcsec.

Transits of Venus and solar parallax

During the next half-century, English, French and Italian astronomers measured Mars' parallax at oppositions, obtaining solar parallax values ranging from 9 to 15 arcsec.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

EDMOND HALLEY (1656–1742) dismissed observations of Mars as unreliable and argued for making careful measurements of a transit of Venus. He had himself observed a transit of Mercury from St Helena in 1677, but, on comparing his results with the single European observation, obtained a solar parallax of 45 arcsec, clearly too large. After careful consideration of Venus's orbit, Halley was able to calculate the dates of future transits of both Mercury and Venus, and these he announced in the *Philosophical Transactions* in 1691. The next pair of Venus transits would occur in June of both 1761 and 1769; another pair would follow in December of both 1874 and 1882. A second paper by Halley, published in 1716, laid out the method for observation, with suggestions of locations where the transit might be observed in 1761. Observers should time the transit from ingress on the Sun's limb to egress. One's geographic location determined where Venus's path would lie on the Sun's face and how long the transit would take.

Halley's younger French contemporary, J-N DELISLE (1688–1768), championed the transit method and proposed an alternative technique for observers who were not in a position to time the entire transit: if one could measure accurately the time of either ingress or egress, one could compare results with another observer so long as the longitudes of both locations were known accurately. For the same event, the time difference between two stations was related to the length of Venus's path across the Sun.

Delisle also pressed his contemporaries to measure the transit of Mercury which occurred in 1753. The results were disappointing, but the practice was valuable. In the run-up to the 1761 transit of Venus, the French worked to improve orbital calculations, with Delisle publishing a map and times a year before the event. Both the French and the British despatched expeditions overseas but most observations were taken in Europe, in a number of countries. With such a wide variety of observers and their differing experience, the results were equivocal, with estimates of the solar parallax ranging from 8 to nearly 11 arcsec.

With improved planning and support, the 1769 transit was even better observed, with some 77 stations involved world-wide. Again, results varied, but not so widely as in 1761. Part of the problem was lack of precise longitudes for distant stations. Atmospheric seeing, the quality of telescopes and observers' experience were also important factors. Furthermore, it was very difficult to determine the precise moment of ingress or egress, because observers noted a luminous ring around Venus and because of the 'black-drop effect', where a small, dark link seemed to join Venus's outline with the Sun's limb momentarily before or after contact. Nevertheless, Delisle's value of 8.6 arcsec was widely accepted, although during the following century, with the emergence of better statistical methods and advances in celestial mechanics, the value was revised upwards slightly.

A second pair of transits followed in 1874 and 1882 and were the focus of major international efforts with better instrumentation, which now included photography. However, observational teams arrived at a range of values between 8.76 and 8.85 arcsec, a discrepancy equivalent to about 2 million kilometers. Armed with data from four transits and revised longitudes for observing stations, SIMON NEWCOMB (1835–1909) calculated several values for solar parallax, ranging between 8.79 and 8.85 arcsec. After further work on the speed of light with A A MICHELSON (1853–1931), Newcomb settled on 8.80 arcsec. In the late 1870s, DAVID GILL (1843–1914) made another attempt at Mars' parallax at opposition, obtaining a value for solar parallax of 8.78 arcsec, close to Newcomb's early estimate.

Even before the 1874 transit of Venus, J G GALLE (1812–1910) suggested that minor planets might be better targets, as they showed no visible disk. Measurements of three in the late 1880s allowed Gill to obtain a value of 8.80 arcsec, which within limits of error almost matched the revised value provided by Newcomb. In 1898, it was calculated that a newly discovered minor planet, EROS, would have a perihelion distance of only 0.27 astronomical units from Earth at its opposition in 1900. However, even then, visual and photographic observations yielded two different values for the solar parallax, 8.806 ± 0.004 and 8.807 ± 0.003 arcsec, so placing the Sun at more than 140 million kilometers from Earth at its mean distance. These values would stand for another 40 yr, until observations of Eros's opposition in 1930 were reduced and published.

Stellar distances

When Galileo became converted to Copernicanism, he realized that the detection of annual parallax was a priority. Assuming the stars lay at differing distances and that the brightest were the nearest, he suggested that examination of DOUBLE STARS might avoid some of the practical difficulties involved in observations spread over many months and being made with instruments liable to warp with the passage of the seasons. If the fainter star of a double was much more distant than its brighter partner, then the fainter star could be regarded as a quasi-fixed reference point provided by a helpful Nature, and the astronomer might observe the motion of the brighter star relative to the fainter. However, as so often he contented himself with offering the suggestion for others to implement.

After Kepler taught astronomers to think dynamically and to view the planetary system as composed of small bodies in orbit about a massive Sun, the detection of annual parallax ceased to be a test of Copernicus's hypothesis and became instead the preferred method of determining the distances of the stars. The problems were practical, and not least among these was refraction, whose effects were poorly understood. ROBERT HOOKE (1635–1703) sought to avoid this difficulty by observations of the star GAMMA DRACONIS, which passed directly overhead his lodgings in London. To ensure that his instrumentation was not warped by the passage of the seasons, Hooke devised a zenith telescope which he incorporated into the fabric of his house. However, he made only four observations of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the star, and other astronomers found his claim to have detected annual parallax unconvincing.

The same technique was applied with much greater care in the 1720s by an amateur, SAMUEL MOLYNEUX (1689–1728), and his friend JAMES BRADLEY (1693–1762), a future Astronomer Royal. Molyneux commissioned a zenith telescope from the maker George Graham; with it, he and Bradley found that there was indeed an annual movement in gamma Draconis, but that it was 3 months out of phase with the expected annual parallax. Bradley therefore commissioned another zenith telescope, with a wider field of view, and this enabled him to establish the pattern of this annual movement in a number of stars. Eventually he realized that the cause lay in the aberration of light, the influence of the Earth's velocity on the direction from which starlight reaches the observer. It was proof of the motion of the Earth, but not the expected one. Bradley estimated from his failure to detect annual parallax that the distances of the stars must be at least 400 000 astronomical units.

Meanwhile other investigators were attempting to estimate stellar distances on the assumption that the stars were physically similar, so that the apparent brightness of a star was inversely proportional to the square of its distance from the observer. The problem was, how in practice to make the comparison between the apparent brightness of the Sun and, say, SIRIUS. Huygens attempted to do this by making a hole in a screen, of a size such that the fraction of the solar surface visible through the hole was equal in brightness to Sirius. From this he estimated that Sirius lay at 27 664 astronomical units.

A better method was proposed by JAMES GREGORY (1638–1675) in 1668. Assuming a knowledge of the dimensions of the solar system, Gregory replaced Sirius by a planet of equal brightness, and then calculated the ratio of the light that came directly from the Sun to the light that came from the Sun via the planet. Isaac NEWTON (1642–1727) repeated the calculation with better values for the distances within the solar system, and concluded that Sirius lay at about one million astronomical units. However, his estimate was known only to intimates until its posthumous publication, which occurred only a few months before Bradley's announcement in 1729 of the aberration of light and its implications. These two results—Newton's estimate of an actual distance for Sirius of one million astronomical units on the assumption of physical equality among the stars and Bradley's estimate of a minimum stellar distance of 400 000 astronomical units—established the scale of the stellar universe, and alerted astronomers to the exquisite delicacy involved in the measurement of annual parallax.

There was an understandable loss of nerve, and only at the beginning of the 19th century did attempts to measure annual parallax resume. Telescopes had improved in quality, and it was becoming realized that large proper motion might be a better clue to nearness than apparent brightness. In particular, the discovery that 61 CYGNI had a proper motion of no less than 5 arcsec per annum suggested that this star might be among the nearest.

The early decades of the 19th century saw a rash of mistaken claims to the detection of annual parallax, and sceptical astronomers would need convincing evidence before believing any future claim. Such evidence was provided by FRIEDRICH WILHELM BESSEL (1784–1864) of Königsberg, where there was a fine 16 cm heliometer by Joseph Fraunhofer. In 1837 and 1838 he observed 61 Cygni an astonishing 16 times or more every clear night, and at the end of 1838 he announced the star had a parallax of about 1/3 of a second of arc. What carried conviction was the match between theory and the pattern of Bessel's observations.

Bessel had been encouraged by the preliminary results of Friedrich Georg WILHELM STRUVE (1793–1864) of Dorpat Observatory in Russia, who had been observing VEGA with a Fraunhofer refractor of 24 cm, the largest in the world. Vega was bright and had a large proper motion. Struve observed the star from 1835 and derived a parallax of 1/8 of a second of arc, although he later doubled this value. A third determination was announced in January 1839 by THOMAS HENDERSON (1798–1844) following his return from the Cape of Good Hope, where the stars he had observed included ALPHA CENTAURI. This star was bright, was a double with wide angular separation and had a large proper motion, all recognized criteria for nearness. On reducing his observations Henderson derived a parallax of over 1 second of arc.

For the remainder of the 19th century, measurements of annual parallax proved difficult and the results were often contradictory, even after the invention of photography. Furthermore, measurements of distances beyond the nearest stars were hardly to be expected, given that as late as 1900 the very existence of other galaxies was thought unlikely. However, William Herschel (1738–1822) had pioneered methods that would bear fruit in the 20th century, even if he had limited impact on his contemporaries and immediate successors (see HERSCHEL FAMILY). In 1785 he had published a diagram of a cross-section of the Galaxy, using the distance of Sirius as a yardstick. He had assumed that within the Galaxy the stars were distributed uniformly, and that his telescope could penetrate to the borders in all directions, so that the number of stars visible in any field of view was an index of the distance to the border of the Galaxy in that direction. Both assumptions he later abandoned, but his diagram had illustrated the potential power of hypotheses of uniformity when combined with stellar statistics. It was after reasoning along similar lines that he told the poet Thomas Campbell in 1813, 'I have observed stars of which the light, it can be proved, must take two million years to reach the earth' (Lubbock C A 1933 *The Herschel Chronicle* (Cambridge: Cambridge University Press) p 336

Bibliography

Hoskin M 1982 *Stellar Astronomy: Historical Studies* (Chalfont St Giles: Science History Publications)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Newcomb S 1895 *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy* (Washington, DC: US Government Printing Office)
Van Helden A 1985 *Measuring the Universe: Cosmic Dimensions from Aristarchus to Halley* (Chicago, IL: University of Chicago Press)
Woolf H 1959 *The Transits of Venus* (Princeton, NJ: Princeton University Press)

Richard A Jarrell and Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Distribution of Galaxies, Clusters, and Superclusters

The distribution of matter in space is the key to understanding the past, present, and future of the universe. If the average mass density on the largest scales exceeds a critical level, then the expansion of the universe will eventually cease and reverse, ultimately ending in a 'Big Crunch'. Conversely, if the mean mass density in the universe is less than this critical value, the expansion of space will continue without end. The distribution of matter also reveals much about the cosmic history of the universe: matter was very smoothly distributed shortly after the Big Bang but is much less so now. When observations of how galaxies and CLUSTERS OF GALAXIES are spread throughout space are combined with observations of their spectroscopic and morphological properties, astrophysicists can place important constraints on how these systems formed, evolved, and, ultimately, elucidate the nature of the growth of structure in the universe. In particular, the relationship between the distribution of luminous matter (e.g., hot gas, stars, galaxies) and non-luminous matter (e.g., non-baryonic particles) is determined by the thermal properties of the non-baryonic 'dark' matter, the history of star formation, gravity, and the cosmological model. Hence, accurate predictions for the spatial distributions of galaxies and clusters as functions of time are fundamental requirements for any viable theory of how structure in the universe formed and evolves. Studying the distribution of galaxies and clusters over a range of distances and epochs has, consequently, been an area of much active research over the past 20 years. While there is still much that is unknown about the underlying mass distribution in the universe, substantial progress has been made owing to breakthroughs in astronomical instrumentation and detector technology, numerical simulation software and high-speed computer processors, and particle physics. What we do know is that the distributions of galaxies and clusters show a remarkable degree of inhomogeneity over distances spanning at least 150 megaparsecs (Mpc) and that the clustering properties of galaxies appear to depend on both the local environment the galaxy resides in and on the intrinsic properties of the galaxy. The precise nature of the DARK MATTER and the role of gas hydrodynamics (e.g., STAR FORMATION) in the evolution of structure, however, presently remain unsolved problems in astrophysics.

The cosmological principle, dark matter, and the clustering hierarchy

A guiding principle of cosmology has been that if one looks out to large enough distances, then the universe should be homogeneous and isotropic. This assumption is often referred to as the 'COSMOLOGICAL PRINCIPLE'. By isotropic, we mean that the global properties of the universe (and, therefore, the galaxy distribution) are not dependent the viewing direction. By homogeneous, we mean that observers in different parts of the universe will record the same basic picture of the universe at a given cosmic epoch. The gist of the cosmological principle is that there are no preferred directions or locations in the universe. Present observations suggest that the cosmological principle is valid but only on distance scales in excess of 150 Mpc or so. On smaller scales, we do see spatial variations in galaxy properties (and possibly even in certain cosmological parameters) which are presumably tied to substantial fluctuations in the underlying distribution of matter in space.

Early maps of the galaxy distribution on large scales, generated prior to the 1980s, were two dimensional—reproducing the distribution of galaxies on the plane of the sky. The third dimension, distance from our Galaxy, was only available for a small fraction of those galaxies. Nonetheless, these maps showed the distribution of galaxies to be highly non-random: when one galaxy is found, the likelihood of finding a nearby companion galaxy is significantly higher than Poisson statistics would predict.

Clustering in the universe is detected on many different scales. Our Milky Way Galaxy is a member of a group (named the LOCAL GROUP) of 24 galaxies spread over a region about 1 Mpc in radius. The Local Group contains a total mass in the range 3×10^{12} to 10^{13} solar masses. Groups like this are relatively common: on average, there is approximately one group for each 1800 cubic Mpc volume. Clusters of galaxies are approximately comparable in size, spanning 1–5 Mpc, but are considerably more massive and rarer than galaxy groups (figure 1). Clusters contain hundreds of galaxies in their central regions. The total cluster mass is usually at least 10^{14} solar masses and the richest systems can be 50 times more massive than that. The mass density within a cluster is typically at least 100 times the universal closure density[1]. Clusters of galaxies are the largest systems known to have reached dynamic equilibrium, a state in which the cluster's gravitational potential energy is twice its internal kinetic energy. Gravitationally bound structures may exist on larger scales still. Although quite rare, clusters of clusters (known as SUPERCLUSTERS) have been detected. These systems can contain in excess of 10^{16} solar masses, including up to 10 individual galaxy clusters spread across regions 10–30 Mpc in extent. Some superclusters are suspected of being bound systems but it is highly

[1] The universal closure density is 1.88×10^{-29} grams per cubic centimeter. This is the minimum mean mass density required to eventually halt the present expansion of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

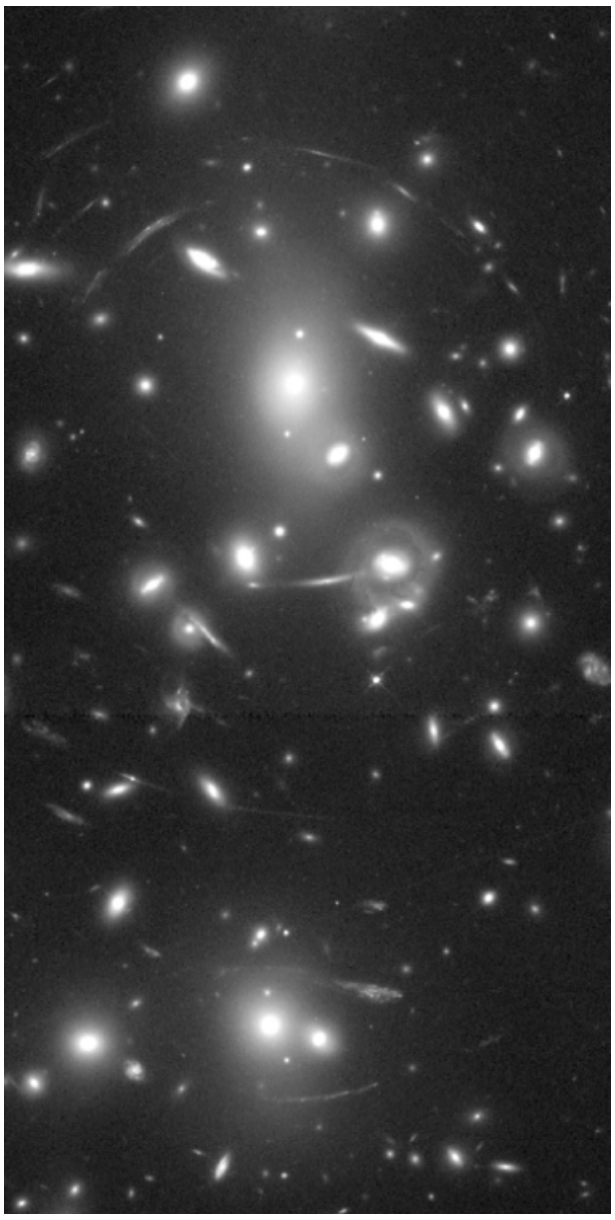


Figure 1. The galaxy cluster Abell 2218 as seen by the Hubble Space Telescope. The field of view spans a distance of about 415 kpc. The curved, faint galaxies are actually distant galaxies lying well beyond the cluster whose appearances have been distorted by the effect of the cluster's mass on the curvature of space. Such distortion, known as gravitational lensing, is one way the mass of a cluster can be estimated.

unlikely they have had sufficient time to reach dynamical equilibrium.

Structure on scales in excess of 100 Mpc was not discovered until wide-area, three-dimensional galaxy surveys were first undertaken in the early 1980s. Three-dimensional surveys are essential for discovering the true nature of the galaxy, cluster, and supercluster distributions because projection effects hide substantial

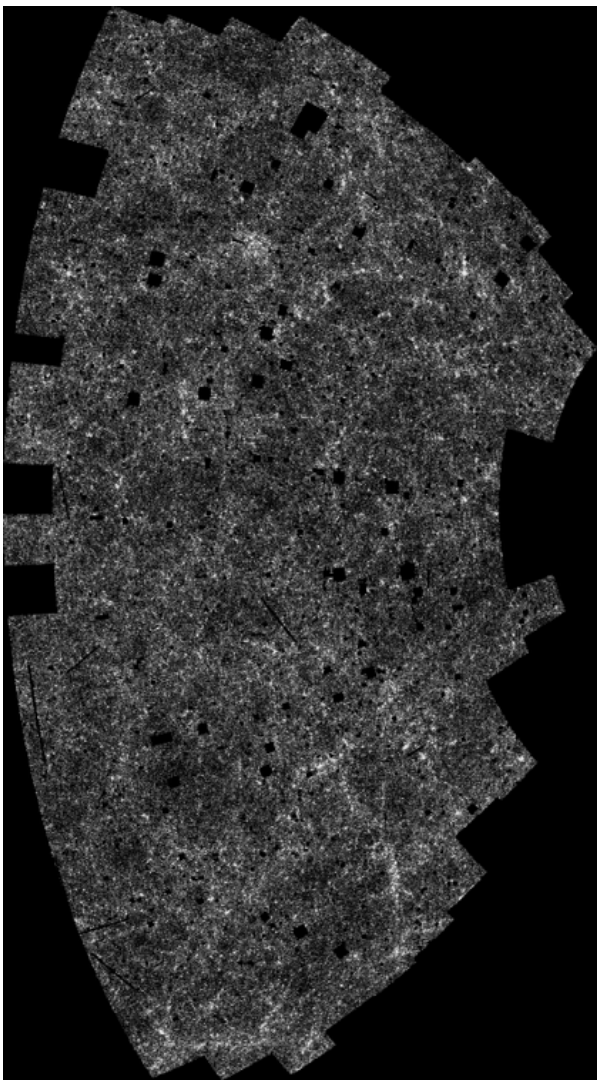


Figure 2. A map derived from digitized photographic sky survey plates showing the angular distribution of over 2 million galaxies in the southern galactic hemisphere. Blank regions in the map were skipped due to the presence of very bright stars.

structures in two-dimensional maps like that in figure 2. Indeed, an amazing bubble-like network of voids and sheets of galaxies and clusters, extending for hundreds of Mpc, is revealed in these spatial surveys.

The clustering patterns that we see in the universe, from galactic scales and up, are believed to be governed by the gravitational potential of ubiquitous dark matter. The galaxies, which comprise less than 5% of all the matter in the universe, are simply tracers of the dark matter distribution. This remarkable realization is based on observations which infer the total mass in galaxies and in clusters of galaxies from the distribution of orbital velocities of the material inside these systems. The dynamical mass estimates derived in this way are typically 5–10 times larger than the masses estimated by summing up the luminous components (e.g., stars in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

case of galaxies; galaxies and hot gas in the case of clusters). The implication is that most of the matter in the universe has not yet been directly detected by conventional astronomical instrumentation.

The nature of this dark matter is still not well understood. The hypothesis that dark matter consists solely of BARYONS (e.g., neutral gas, dilute plasma, massive planets, white dwarfs, neutron stars, and black holes) appears to be in conflict with observations of hot gas in clusters of galaxies (the gas comprises only about 10–20% of the total cluster mass), the tendency for dying stars to eject a substantial fraction of their mass back into interstellar space (and, hence, locking baryonic matter in stellar remnants is a very inefficient process), and limits on the number of massive planets, brown dwarfs, and very low mass stars from both gravitational micro-lensing surveys and deep Hubble Space Telescope images (the space density of such objects is too low to account for a significant fraction of the inferred total mass density in the universe).

The spatial distribution of galaxies and clusters can help constrain the nature of the dark matter because dramatically different configurations are predicted depending, for example, on whether or not the dark matter particles are highly relativistic (dubbed hot and cold dark matter, respectively). Candidates for non-baryonic dark matter include neutrinos, axions, and the weakly interacting massive particles predicted by supersymmetry models (photinos, higgsinos, zinos, and solar cosmions). Programs to detect the more exotic of these cosmological non-baryonic particles are underway.

Voids, sheets, filaments, and spikes

The frothy spatial distribution of about 10 000 galaxies in two nearly opposing directions is shown in figure 3. The data were obtained at observatories located in Arizona and Chile. In this figure, known as a cone diagram, each galaxy's position is shown as a point. Our own galaxy is located at the point where the wedges meet. The distance represented by the length of each wedge is about 120 Mpc. Several striking features are evident in this galaxy map (none of which were predicted by theory prior to the observations): large regions of very low galaxy density, somewhat misleadingly referred to as voids, extend for tens of Mpc. The voids are surrounded by a sheet-like network of galaxies often extending for several hundred Mpc. Indeed, one such coherent feature can be seen in figure 3 crossing from the mid-region of the northern cone diagram into and across the southern cone diagram. This feature has been dubbed the 'GREAT WALL'.

Clusters of galaxies can also be seen in the map as highly elongated clumps oriented towards the center of the diagram. This elongation is an artifact of the large, orbital velocities of galaxies about the cluster centers which can often approach 1000 km s^{-1}. These orbital velocities distort the spatial positions determined from the redshift measurements and result in the elongated appearance of clusters in such diagrams. If we could observe the cluster galaxies using a map based on distances not dependent on the REDSHIFT, such elongation would vanish.

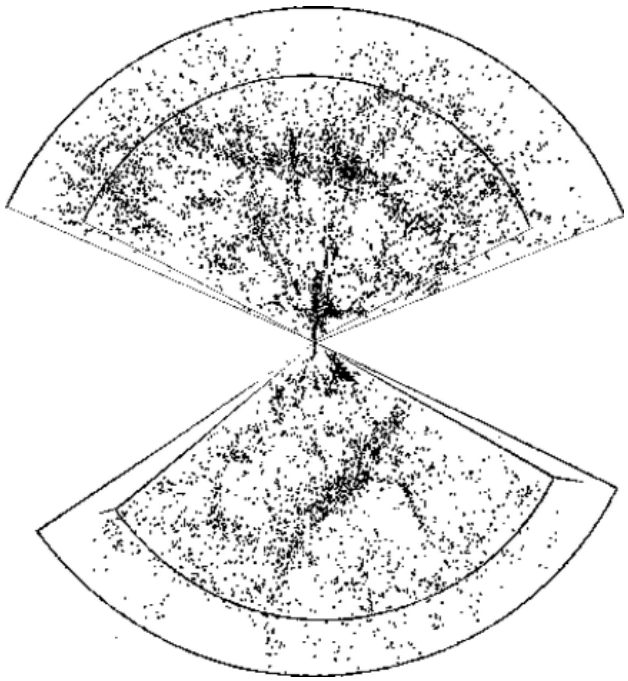


Figure 3. The spatial distribution of approximately 10 000 galaxies extending out to a distance of 120 Mpc. Galaxies from a small portion of the northern galactic hemisphere are shown in the upper cone diagram and those from a comparable region in the southern galactic hemisphere are shown in the lower cone diagram. These data are from a large redshift survey coordinated by the Center for Astrophysics at Harvard University.

The structures in this map are clearly comparable with the depth of the survey and, thus, wider and deeper surveys were undertaken in the 1990s to assess just how large the largest structures in the universe are (see GALAXY REDSHIFT SURVEYS). One technological development which made very large redshift surveys feasible was the multi-object spectrograph (MOS). The 10 000 redshifts above were measured one at a time on telescopes which had narrow fields of view ((1/7)th the area subtended by the full Moon). Multi-object spectrographs today allow up to 600 redshifts to be measured simultaneously from galaxies spread over an area of a few square degrees (roughly 15 times the area subtended by the full Moon).

One of the first large MOS surveys to be completed, known as the Las Campanas Redshift Survey (LCRS), measured about 24 000 redshifts in two directions 132° apart. The distribution of galaxies from this survey is shown in figure 4. The depth of the LCRS

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

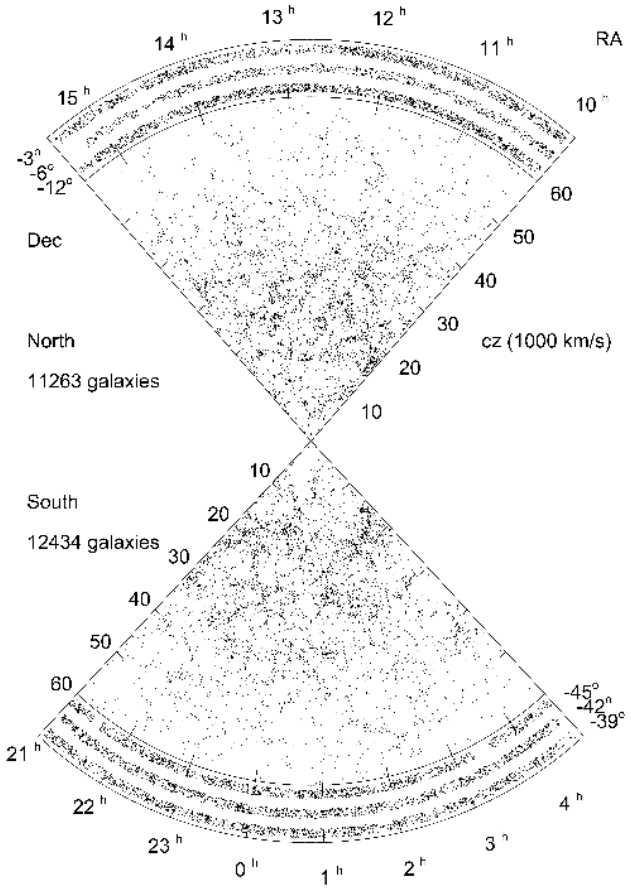


Figure 4. The spatial distribution of approximately 24 000 galaxies extending out to a distance of 600 Mpc. These are the results from the Las Campanas Redshift Survey conducted by an international team of astronomers using a Chilean telescope equipped with a multi-object spectrograph. The 3 bands going across the top and bottom show the projected distribution of the galaxies on the sky. Although the depth of this survey is substantial, it only subtends 0.3% of the entire sky.

was nearly 5 times greater than that in figure 3. Yet the structures found were comparable in size suggesting that perhaps there are no substantial inhomogeneities much larger than 120 Mpc. However, the LCRS only includes distances for a fraction of all the galaxies in this part of the sky due to time and telescope constraints. As a consequence, the structure in the LCRS map is not as clearly delineated as that in a survey which completely samples the galaxy distribution, such as the one in figure 3. Any conclusions regarding the existence of structures on scales larger than 120 Mpc or so must therefore be tempered by the fact that the survey has only done a partial sampling of the galaxy and cluster distribution.

The results of the most extensive redshift survey of clusters of galaxies is shown in figure 5. The radius of the survey is 240 Mpc and includes about 50% of the ~1000 clusters actually residing in this volume.

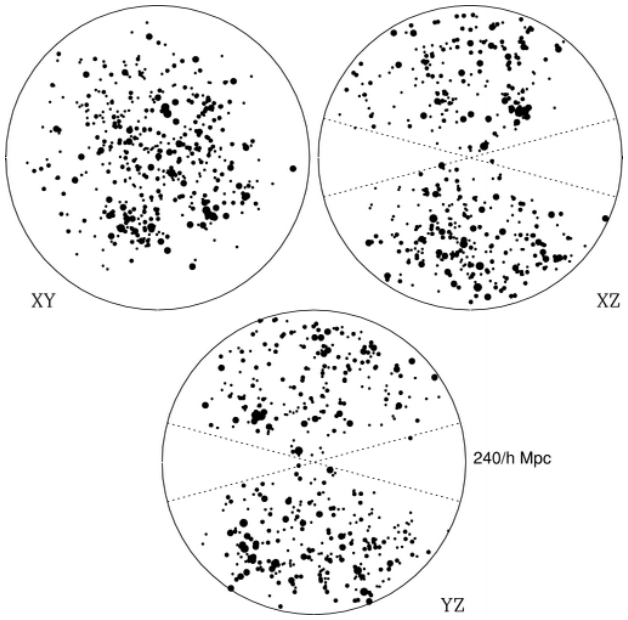


Figure 5. Three views of the spatial distribution of the richest ~480 clusters within a radius of 240 Mpc, centered on our Galaxy. Each dot represents a cluster and the dot size is proportional to the number of bright galaxies in the central region of the cluster. The empty wedges in the XZ and YZ projections are due to the blockage by the disk of our Galaxy.

Quantifying clustering

The constraints that the above surveys place on theories of structure formation and dark matter models come from a variety of statistical quantities which measure the degree to which the galaxy and cluster distributions differ from a purely random configuration. One of the most widely used of these statistics is the two-point correlation function, $\xi(r)$. The two-point correlation function measures the probability of finding one galaxy separated by a distance r from another galaxy. The observations above reveal that $\xi(r)$ is well represented by a power law of the form

$$\xi(r) = \left(\frac{r_0}{r}\right)^{\gamma} \quad (1)$$

where r_0 is known as the correlation length. The correlation length is the distance scale at which there are twice as many pairs of objects than expected if the distribution of objects were completely random. For galaxies, the best-fit parameters for $\xi(r)$ are $r_0 = 5.5\pm1$ Mpc and $\gamma = 1.8\pm0.05$. If the distribution of galaxies were random, one would expect $\xi(r) = 0$ at all scales. The constraints on $\xi(r)$ mirror the visual impression given by the redshift surveys—the galaxy distribution is strongly clustered, not random. Remarkably, the correlation function of clusters of galaxies has a slope nearly identical to that for galaxies ($\gamma = 1.9\pm0.3$) but the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

correlation length is significantly larger, $r_0 = 20\pm8$ Mpc. The agreement between the shape of $\xi(r)$ for galaxies and clusters suggests that these two rather different types of stellar systems trace the same underlying mass distribution. An additional constraint from the redshift surveys of galaxies and clusters is that the power-law representation of $\xi(r)$ appears to be valid on scales up to at least 40 Mpc and perhaps as high as 60 Mpc.

The correlation function parameters above are mean values for large samples of galaxies and clusters. The values are not universal, however. There is a highly significant dependence of the correlation length on the luminosity of the galaxies being studied and on their morphology. For example, it has recently been established that the correlation function of the most optically luminous galaxies ($L_{opt} \geq 5 \times 10^{43}$ erg s^{-1}) have correlation lengths of ~16±2 Mpc, approaching that of clusters of galaxies, whereas the intrinsically faintest galaxies have correlation lengths as low as 4 Mpc. It has also been established that the likelihood of finding a neighbor galaxy within 1 Mpc from an elliptical is about twice as high as that for finding a companion within the same volume from a spiral galaxy. Thus, although all the distribution of all galaxies is determined by the underlying matter distribution, elliptical galaxies and/or luminous galaxies appear to be more efficient tracers of regions in space with the highest densities. The cluster correlation function parameters also appear to depend weakly on the cluster's x-ray luminosity: the most luminous x-ray clusters ($L_X \geq 1.4 \times 10^{44}$ erg s^{-1}) have $r_0 \approx$ 35±15 Mpc, compared to $r_0 \approx 15\pm4$ for the least rich (and least x-ray luminous) systems.

The power spectrum, $P(k)$, is the Fourier transform of $\xi(r)$ and provides a complementary but important constraint on the clustering properties of galaxies and clusters: broad features in the correlation function are narrow in Fourier space and vice versa (see CORRELATION FUNCTION AND POWER SPECTRA IN COSMOLOGY). An analysis of the distribution of galaxies in the LCRS revealed there is a significant 'spike' in the power spectrum on scales near 100 Mpc. Deeper redshift surveys, known as 'pencil beam' surveys because they cover very small solid angles but probe great distances, also have detected excess clustering signatures on scales of up to 130 Mpc. Indeed, the results from the first pencil beam surveys showed there appeared to be quasi-periodic spikes in the radial distribution of galaxies, occurring every 120–130 Mpc. These features are surprising since it is not understood why there should be such a preferred size scale. One interpretation is that structures on large scales, like the network of walls and voids seen in the wide area redshift surveys, have characteristic sizes. But what determines these characteristic scales? Some of this excess power could be the lingering effects of the properties of the plasma that existed in the very early universe. But at present we do not really understand the origins of these remarkable features in the galaxy distribution. Nonetheless, the global shape of the galaxy and cluster power spectra are similar, suggesting again that clusters and galaxies are tracing similar perturbations in the matter distribution. The amplitude of $P(k)$ for the richest clusters is, on average, a factor of ~2–3 higher than that for poorer clusters, consistent with differences seen in their respective $\xi(r)$. In turn, the rich cluster $P(k)$ amplitude is about 12–16 times higher than that derived for galaxies in the LCRS.

'Biasing'

The above results demonstrate that not all extragalactic objects trace the underlying mass distribution in the same way. Indeed, this conclusion represents a substantially new paradigm in cosmology. Prior to the early 1980s it had usually been assumed that a one-to-one relation existed between fluctuations in the distribution of galaxies and fluctuations in the distribution of mass. We now realize that not only is the relationship not a direct one but it probably is not even a linear one and may depend on time and size scale, as well. Galaxies and clusters are thus deemed 'biased' tracers of the matter distribution and much research is now in progress to understand the physical mechanisms which can account for this more complex relation between the galaxy, cluster, and matter distributions.

The dependence of the clustering properties, such as the correlation length, on the mass of the system have been explained in terms of the statistics of 'rare peaks' in a distribution. Fluctuations in the density of mass in the universe occur on many scales. Superposed on the largest fluctuations are smaller ones. If galaxies and clusters of galaxies form at sites which achieve some minimum density enhancement, then there will be a preference for the most massive and most luminous objects to form at those sites which already have elevated mass density levels by virtue of sitting atop the peak of a larger scale fluctuation. The highest density peaks in the mass distribution, however, occur less frequently—one has to search a larger volume to find a large density peak than a small peak. It has been proposed that a universal relation can be made between the mean separation between objects, d_i, and their correlation length, r_0, of the form

$$r_0 = (0.3 \pm 0.1) d_i \tag{2}$$

implying that objects with larger separations (which tend to be more massive systems) will exhibit stronger clustering[2]. Does the trend between clustering strength and mass extend to superclusters? The answer is unclear because there are too few superclusters known at present to perform robust computations. Surveys initiated in

[2] The mean interobject separation is defined as $d_I = (n_i)^{-1/3}$ where n_i is the mean space density of the objects.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1998 may yield enough additional superclusters to address this question.

Although early versions of structure formation theory based upon a universe dominated by cold dark matter (CDM) did predict a trend towards higher correlation lengths as the galaxy and cluster mass increased, these same models did not come close to predicting the actual observations in which the richest clusters are 10–20 times more strongly clustered than are galaxies. The resolution of this discrepancy was to postulate that galaxies and clusters are biased tracers of the mass. These biased CDM models are more successful at reproducing the actual galaxy distribution, but there remain other significant observations with which the models do not agree very well. This is probably because either the dark matter properties assumed are not accurate and/or the specific biasing model is not accurate.

Statistics like $\xi(r)$ and $P(k)$ are not sufficient to unambiguously describe the galaxy and cluster distributions. Distributions that appear dramatically different to the eye can often have similar two-point correlation functions. This is because the eye is sensitive to high-order moments of the distribution (such as 'wall'-like bands of galaxies surrounding large, underdense volumes). Statistics which measure the high-order moments of the distributions thus provide superior discrimination between cosmological models as well as some of the best constraints on the degree of biasing and, thus, on the reliability of galaxies and clusters as tracers of the mass. Current constraints on the higher moments (orders 3–6) of the galaxy and cluster distributions suggest that while galaxies and clusters are sampling the same underlying matter distribution the biasing between them is non-linear.

Cosmic flows

The above observations demonstrate both the power and limitations of mapping structure with redshift surveys. While such surveys succeed in revealing extensive and intricate structures on many scales, the physical interpretation of the structure is complicated by the fact that the tracers of the structure are biased. However, there is a complementary approach to mapping structure which takes advantage of the fact that all the matter in the universe is subject to its own mutual gravitational force. Inhomogeneities in the mass distribution can induce motions of galaxies and clusters (dubbed 'peculiar' velocity by astronomers) which are unrelated to the expansion of space. On scales larger than about 10 Mpc, a three-dimensional map of these motions can be used to infer the masses and probable locations of the responsible structures because the induced velocities are linearly proportional to the gradients in the gravitational field[3]. The beauty of such measurements is that they provide a more direct method for mapping the mass distribution and many of the concerns about biased tracers are moot. Peculiar velocity observations are, however, difficult to perform accurately.

The observational study of large-scale cosmic velocity fields began in 1976, with the discovery that the temperature of the COSMIC MICROWAVE BACKGROUND (CMB) is a few millidegrees warmer in one direction and equally cooler in the opposite direction. This signature was quickly interpreted as due to motion of our Galaxy relative to the CMB, motion that is assumed to be due to an imbalance in the gravitational tug of the various structures which surround our Galaxy. (If matter were uniformly distributed around us, the gravitational forces would cancel, on average, and no motion would be induced. However, as seen in the redshift surveys matter is very far from uniformly distributed.) Subsequent measurements of this anisotropy from the Cosmic Background Explorer (COBE) satellite have yielded a highly accurate measurements of this motion: 620 km s^{-1} towards the direction l = 270.8°, b = +29.1°, when measured with respect to the center of the Local Group. Since the CMB is the most distant reference frame possible (it originates at a redshift of ~1000, when the universe was only a few hundred thousand years old), it is quite reasonable to presume that the structures causing this motion are much closer.

In order to measure the peculiar velocity of an extragalactic object, one needs to know both its redshift and its distance. In this case, the distance must be determined *independently* from the redshift. This is because the observed redshift contains contributions from both the expansion of the universe and from any gravitationally induced velocity. The contribution solely from the expansion (referred to as the recession velocity) can be determined directly from the redshift-independent distance via HUBBLE'S LAW. The peculiar velocity is then the difference between the observed redshift and the recession velocity. There are presently at least seven independent methods which can be used to determine distances to galaxies and clusters, albeit not all methods work on all galaxy systems nor do they all yield the same measurement accuracy. However, the common principle in all of these methods is that there exists some observable quantity (other than redshift) that can be used to estimate the distance. In many of these methods, the distance is inferred by comparing an estimate of the object's intrinsic luminosity with its apparent (observed) luminosity.

[3] On scales less than 5 Mpc, the relationship between induced velocity and mass becomes highly nonlinear making reconstruction of the underlying structure from peculiar velocities extremely difficult.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Using such techniques, redshift-independent distances (and, hence, peculiar velocities) can be established for large samples of galaxies and clusters in different parts of the sky. From such datasets, the motions of galaxies can be charted and the presence of fluctuations in the matter distribution then inferred. The two largest galaxies in the Local Group, our Milky Way and the Andromeda galaxy, are separated from one another by about 700 kpc. Two massless particles sitting that far apart in our expanding universe would be traveling away from one another at more than 50 km s^{-1}. In fact, these two galaxies are moving towards one another at about 70 km s^{-1}. This observation has led to estimates of the combined mass of these two spiral galaxies. The entire Local Group itself is moving due, in part, to the gravitational pull of the VIRGO CLUSTER which lies about 16 Mpc away. But the Virgo cluster mass is sufficient to explain only about 40% of the CMB velocity. Additionally, the CMB velocity vector is misaligned with the position of the Virgo cluster by 47°. Additional structures at larger distances must, therefore, play a role.

One of the first peculiar velocity surveys to detect these structures was conducted in the mid-1980s using distances to over 300 elliptical galaxies. A startling result was found: a vast number of galaxies within a sphere nearly 100 Mpc in diameter appear to share a common motion towards a massive aggregate of matter. Such a motion is referred to as a 'bulk flow' and since so many galaxies participate in it, the aggregate was dubbed the 'Great Attractor'. The bulk flow vector was 600±104 km s^{-1} towards $l = 312°$, $b = 6°$ (angular error is ~±10º), close to the CMB velocity of 620 km s^{-1} but about 45º from its apex. Subsequent surveys indicate that the 'Great Attractor' is a moderately rich supercluster lying between 40–50 Mpc from our Galaxy. The Great Attractor itself may be participating in a bulk flow suggesting that the origin of the flow is due, in part, to structure on scales larger than 50 Mpc.

Do such bulk flows extend to scales of 100 Mpc or larger? If so, then the cosmological principle is not valid on scales even as large as 100 Mpc. With this key question as a driver, several groups set out to explore the velocity field on scales 4 times larger than the original 'Great Attractor' survey. Initial results from some of these surveys suggest there is evidence for bulk flows of 600–800 (±150–300) km s^{-1} which extend out to scales of 100 Mpc or more. The different surveys have not all yielded consistent directions for these flows, however. There is some evidence to suggest that these very large-scale flows point to within ±30º of the Shapley supercluster, a system with a mass of a few × 10^{16} solar masses located 150 Mpc from the Local Group and towards $l = 312°$, $b = 30°$. The results are controversial and this is presently a research area of much active observational work. One challenge of measuring bulk flows on such large-scales is that the signal is small and often must be extracted from noisy data. If very large-scale bulk flows turn out to be a reality, then present models for structure formation in the universe will require substantial revision (if not outright rejection).

The Hubble bubble?

Gravitational attraction towards massive structures is not the only means available for generating a large bulk flow, however. If the expansion of the universe were dependent on one's location, then flows could be a direct consequence. Another way to ask this question is: is the Hubble constant (H_0) really constant? The answer is no. A more accurate description of this term would be the Hubble parameter. The Hubble parameter does vary with cosmic age—the temporal dependence is a consequence of the deceleration of the expansion caused by the matter in the universe. The amplitude of the time dependence, however, only becomes significant when surveys span very large distances (and, hence, a large dynamic range in 'look-back' time). Variations in the value of H_0 as a function of location, however, are also possible if the fluctuations in the mass distribution are sufficiently large. The expansion of space within a large underdense region will proceed at a faster rate than that in a large overdense region. The size of the voids and walls seen in the galaxy distribution are, in principle, large enough to yield measurable variations in the Hubble parameter on scales of 100–200 Mpc. As figure 3 shows, the region of space surrounding our Galaxy is underdense relative to, for example, the 'Great Wall' structure which lies about 70–100 Mpc away. One possible explanation, then, for a large peculiar velocity on the scale of 100–200 Mpc is that our Galaxy resides in a region of space where the local expansion rate is faster than that on much larger scales. To date, the observational constraints on the spatial variation of the Hubble parameter are less than 5% on scales of 150 Mpc. Although small, this can still amount to a few hundred km s^{-1} and thus such effects may indeed be at work.

Future observations and outstanding problems

One key challenge posed by all these observations is to understand how the universe evolved from its highly homogeneous state shortly after creation to the present epoch where large inhomogeneities on scales of 100 Mpc, or more, are relatively common. The models which appear most consistent with the majority of the observations are those which call for a CDM-dominated open universe and/or an open universe with CDM plus some hot dark matter component such as neutrinos. The observations may also be consistent with a non-zero cosmological constant (Λ). If the large-amplitude, large-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scale bulk flows persist, however, some revisions to these models will be required.

While much progress has been made in mapping the structures in the universe, we still have a ways to go. At present, the fraction of the local universe for which we have reliable maps of the galaxy and cluster distributions is a mere 1%, akin to trying to infer the geography of the entire surface of the Earth from a map of the eastern half of the United States. Some fundamental astrophysical problems in the study of the distribution of galaxies, clusters, and superclusters still remain unsolved. Are there galaxian structures on scales beyond 300 Mpc? Are there bulk flows which indeed extend beyond 100 Mpc? Just how complex is the relationship between the galaxy distribution and the matter distribution? What structure existed when the universe was half or even one tenth its present age? What is the precise nature of the dark matter in the universe?

Given the limited amount of telescope time available to conduct surveys of the galaxy and cluster distributions, astronomers have had to balance the desire to cover large areas of the sky (and hence sample many nearby regions of space) and the desire to probe the galaxy distribution at great distances (to study the evolution of structure). Fortunately, as multi-object spectrographs improve and more large-aperture telescopes are constructed, including a few observatories solely dedicated to spectroscopic surveys, the trade off between area and depth is becoming less of an issue. Facilities exist now which will enable astronomers to conduct surveys that both cover moderately large areas and extend to quite large distances (and, hence, probe farther back into time). Three-dimensional surveys of a million or more nearby galaxies, and of thousands of galaxy clusters, are underway. The study of large-scale structure at high redshift is also now becoming accessible. It is likely the next decade will see many breakthroughs in this particular area thanks to an array of large optical and near-infrared optimized telescopes being commissioned over the next several years.

The COBE satellite provided constraints on the distribution of fluctuations in the matter distribution in the very young universe but only on extremely large scales (greater than 600 Mpc). CMB observatories in space to be launched early in the 21st century will yield exquisite constraints on the early matter distribution on scales comparable with those probed by existing redshift surveys and, thus, provide a direct point of comparison for evolutionary studies. Like the peculiar velocity surveys, the CMB studies are attractive because the fluctuations in the microwave background are direct, unbiased tracers of the underlying matter distribution.

But our ability to observe the universe must be matched with an equal ability to interpret and understand the observations. Thanks to revolutions in computer processing power and data storage capacity, experiments to simulate large segments of the cosmos are now underway (see UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION). These simulations track the trajectories of a billion particles as they evolve over cosmic time to form structure. Such new tools will provide theoretical predictions at a level of detail and accuracy hitherto unavailable. And since the differences between some cosmological models are quite subtle, such a capability is indispensable.

Bibliography

Clark S *Towards the Edge of the Universe: A Review of Modern Cosmology* (Chichester: John Wiley & Sons and Praxis)

Dressler A 1994 *Voyage to the Great Attractor: Exploring Intergalactic Space* (New York: Knopf)

Fairall A *Large Scale Structures in the Universe* (Chichester: John Wiley & Sons and Praxis)

Ferris T *Coming of Age in the Milky Way* (New York: Doubleday)

Turok N (ed) *Critical Dialogues in Cosmology* (Singapore: World Scientific)

Marc Postman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Diurnal Motion

The apparent daily motion of a celestial body across the sky from east to west caused by the axial rotation of the Earth. Diurnal motion causes stars to appear to trace out circles centered on the celestial poles. For an observer at the north or south pole of the Earth, the diurnal motion of stars is along circles parallel to the horizon, while, viewed from the equator, stars trace out paths that intersect the horizon at right angles. In general, the arcs along which stars appear to move intersect the horizon at an angle equal to 90° minus the observer's latitude.

See also: axis, celestial sphere, circumpolar stars, day, horizon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Divided Instruments

Although the division of the zodiac into 360° probably derives from Egypt or Assyria around 2000 BC, there is no surviving evidence of Mesopotamian cultures embodying this division into a mathematical instrument. Almost certainly, however, it was from Babylonia that the Greek geometers learned of the 360° circle, and by c. 80 BC they had incorporated it into that remarkably elaborate device generally called the 'Antikythera mechanism'. This mechanism, found in 1900 by sponge divers in the remains of a ship which sank around 180 BC off the Greek island of Antikythera, is a sophisticated geared instrument that could be used to compute the motions of the planets. It is ironic, therefore, that the earliest known graduated astronomical instrument should already be one of great complexity, pointing to both a highly developed astronomical geometry and a complementarily mature state of instrumental craftsmanship in classical Greece.

Ptolemy's influence

In his encyclopaedic *Mathematical Composition*, or *Almagest*, of c. AD 150, Claudius PTOLEMY describes the uses of a variety of astronomical instruments capable of measuring the positions of the heavenly bodies from fixed coordinates by means of the 360° scale. These include the astrolabon and quadrant, as well as adaptations of the linear properties of the circle whereby angles could be measured using straight rods arranged in triangular configurations by means of sines and tangents.

Ptolemy's astrolabon is really a set of armillary rings, in brass or bronze, aligned with the main planes of the sky, such as the zodiac, celestial equator and meridian. Graduated into 360°, these rings could be used to track the motions of astronomical bodies across the sky, so that tables of observations could be compiled, from which theories could be generated. The two-dimensional 'astrolabe' was constructed by an ingenious flattening of these rings onto a series of plates to produce a stereographic projection of the sky in the form of a planisphere. It was the Arab astronomers of the 7th century AD who developed the astrolabe in the form in which it became familiar in the Latin West in the 12th century, turning it into a dual-function instrument that could be used both to observe the altitudes of bodies and to perform practical calculations in spherical trigonometry.

The division of angles into ever-decreasing units of minutes and seconds, where the accuracy of each digit of division could be relied on geometrically and scientifically, became a concern to those cultures that placed a high premium on astronomical research. In the post-classical world, this meant the Arabs, who developed two classes of graduated astronomical instruments. The first of these were large instruments, often built in masonry, in which accuracy was achieved by the greatness of the radius employed. The 1424 observatory of ULUGH BEG of Samarkand with its 130 ft radius meridian sextant set in a drum-shaped building housed the most spectacular and successful instrument of this type, which was capable, in theory, of measuring the solar declination and stellar, lunar and planetary altitudes to within a few arcseconds.

The second class of graduated instruments to come out of the medieval Islamic world were what might be called reckoning instruments. These included the astrolabe, hand quadrant and forms of cross-staves. They were not sufficiently accurate to re-define any astronomical constants, but were intended, with their scales divided to 1° or perhaps 30 arcmin, to measure solar or other data for the daily business of time-finding, navigation, surveying, religious liturgy or astrology. This class of instruments generated the world's first trade in divided scientific artefacts of which a large number of individual examples, especially astrolabes, survive in museum collections around the world.

The astrolabe

The astrolabe came into Europe via Muslim contacts (see ASTROLABE). The astrolabe was not an instrument of original research but a reckoning instrument, whereby one could establish local time by Sun and stars, as well as compute a wide range of calendrical features. Each instrument carried two 360° scales, on the back and front of the main brass plate. The rear scale was used to make celestial observations by means of a sighting alidade, while the front, in conjunction with the stereographically projected sky of 'climate' plates, could be used to calculate the aforementioned horological and calendrical data.

Studies in a range of surviving European astrolabes, made between c 1380 and 1650, indicate that several geometrical configurations, employing scribing compasses and copy-templates, were used in various workshops to generate the 360° scales, and by 1500 there seem to have been established centers, such as Florence and Augsburg, for the manufacture of astrolabes, dials and similar mathematical instruments.

Tycho Brahe

It was the Danish nobleman TYCHO BRAHE, between 1572 and 1598, who became the first European to devise relatively large-radius graduated instruments intended for original research. These included a collection of ARMILLARY SPHERES (set in zodiac and equatorial coordinates), quadrants, sextants, rulers of parallax (or triqueta), globes and clocks with which he performed the first fundamental revision of the positions of the heavenly bodies since Ptolemy 1300 yr before.

Unlike Ulugh Beg and other Muslim astronomers with whose observations he was familiar, Tycho abandoned the idea of building very large radius instruments after a 14 cubit (19 ft) radius quadrant which he had built with Paul Hainzel of Augsburg showed signs of distorting under its own weight. Tycho went on to become the first person to analyze graduated instruments as engineering structures, realizing that stability of the entire instrument was essential if the scale graduations were to remain true. Instead of making instruments of tens of feet radius,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

as a way of making the graduations bigger, he adopted and greatly improved the transversal or diagonal scale: a device going back to Richard Cantzlar and LEVI BEN GERSON. In the diagonal scale, the degree space is seen not simply as a line, but as a two-dimensional area extending across part of the radius. A line, drawn diagonally across the area, and divided into 30 or 60 equal spaces, could then allow a degree to be read to a small fraction from an overall radius of only 6 ft or so. All of Tycho's mature instruments carried diagonals, as did those of most of his successors—Pierre GASSENDI, Giovanni CASSINI, Johannes HEVELIUS and John FLAMSTEED—over the next 130 yr. Tycho Brahe was proud to inherit the triquetum, or set of so-called 'Ptolemy's Rulers', of NICOLAUS COPERNICUS presented to him by Johannes Hannow, Canon of Ermeland. The Rulers consisted of three straight rods and read an angle by using the geometrical relationship of the tangent to the radius and circumference of a circle. The instrument's advantage lay in the relative ease whereby one could divide a straight line into equal parts as opposed to a 90° arc, although, as Tycho found after exhaustive testing, the instrument was not suitable for critically accurate observations.

The ultimate quality of Tycho's positional observations—averaging 1 arcmin, as opposed to 10–15 arcmin for his predecessors—derived from his constant analysis, re-designing and increasing specialization of function of his observatory instruments. Tycho's concern with error detection and correction when it came to both the geometrical and the engineering construction of his instruments constituted a landmark in instrument technology and convinced scientists that their observations, and the plausible theories that they could build on them, were only as good as the graduated scales at their disposal.

Renaissance Europe saw an explosion of ingenuity in the devising of angle-measuring instruments, aimed at the specialist needs not just of astronomers, but also of navigators, surveyors, gunners and mathematical teachers. The principles implicit in Ptolemy's Rulers, whereby straight rods were adapted to read angles, were made use of to produce the mariner's cross-staff, the baculus and their variants. Indeed, even some astronomers of distinction did important angle-measuring work with straight instruments, the most outstanding of whom were Thomas HARRIOT, Pierre Gassendi and Jeremiah HORROCKS.

Intellectual pressures

The intellectual pressures generated by the Copernican revolution, initially exemplified in the path beaten by Tycho Brahe, encouraged European astronomers to search for and quantify a stellar parallax by which the Earth's motion could be physically demonstrated. Also, after 1670, and especially after NEWTON'S *Principia* in 1687, critical measurements were seen as lying at the heart of the quantification of the solar system and the physical testing of gravitation theory. By 1680, therefore, the improvement of graduated instruments had become an ongoing concern, especially in England, where craftsmen such as Thomas TOMPION, George GRAHAM and Jesse RAMSDEN not only made considerable financial profits from pioneering innovation but also won Fellowships of the Royal Society and international academic recognition as well.

The quadrant, which was the basic angle-measuring instrument of the period 1575–1790, had its divisions struck off on its limb using the canons of Euclidean geometry. The radius of any circle will naturally strike off six 60° arcs into the circumference of the same circle, or go $1\frac{1}{2}$ times into a quadrant. A bisected radius—physically exemplified by the beam compass—then strikes off 30°, so that 60° + 30° produce a perfect quadrantal arc. Further bisections will divide the quadrant down to 15° spaces, after which a variety of geometrically 'impure' trisections and quinquesections will take one down to single degrees, and then to minutes.

The micrometer, telescopic sight and clock

Following Tycho Brahe's lead in the construction of more compact instruments with geometrically complex scales, 17th- and 18th-century instrument makers invented a number of systems whereby diagonals of various forms, combined with vernier scales and micrometers, enabled an instrument of 6 or 7 ft radius to perform much more accurately than a simple masonry quadrant of 100 ft. Indeed, the period 1640–1670 saw an extraordinary increase in the sophistication and accuracy of graduated instruments used in astronomical observatories. The large private observatory of Johannes Hevelius, the instruments of which were described and illustrated in minute detail in his *Machina coelestis* (1673), both built on the Tychonic tradition and at the same time influenced Giovanni Cassini in Paris and John Flamsteed in Greenwich. Also, around 1640, the Yorkshireman William GASCOIGNE invented the micrometer and developed the first workable telescopic sight. Between them these two inventions enabled an over 40-fold increase in the practical sighting of astronomical instruments beyond what was possible with the naked eye. By exploiting the ancient properties of the tangent to the circle, and now made vastly more delicate as a result of precision screw-threads and optical reading, the micrometer made it possible to measure down to a single arcsecond. Neither the telescopic sight nor the micrometer became common, however, until brought to the attention of the wider world of learning by Robert HOOKE in 1667, and they had been only partly used by Hevelius. Also, when the oscillating pendulum was successfully harnessed to the going train of a clock by Christiaan HUYGENS in 1658, and was improved by Hooke's and Joseph Clement's development of the 'seconds pendulum' and anchor escapement around 1669, the new precision clock suddenly became a major angle-measuring instrument in its own right, whereby, in conjunction with a meridian quadrant, it could be used to divide up the sky into hours, minutes and seconds of right ascension.

The ROYAL OBSERVATORY, GREENWICH, under John Flamsteed, the first Astronomer Royal, became the first place

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in Europe where these new optical, micrometric and horological techniques of celestial angle measuring were used systematically and with spectacular success, between 1675 and 1719, although valuable work was also done by Cassini and Jean PICARD in Paris, and by Ole Christensen RØMER in Copenhagen.

Cross-checking

During the 18th century, graduated instrument design and use were dominated not only by an ever-increasing pressure towards greater accuracy but also by the evolution of cross-checking and error-detecting techniques whereby both craftsman and astronomer could quantify the errors of their instruments.

Fundamental in this line of development was the 8 ft radius quadrant which George Graham FRS built for Edmond HALLEY at Greenwich in 1725. In addition to its 90° scale, Graham equipped this instrument with a second scale of 96 equal parts, the whole of which could be drawn by 'pure' geometrical bisection: i.e. radius = 64 parts then bisected to 32:16:8:4:2:1, by beam compasses.

These two scales, drawn 1 in apart from the same center, acted as cross-checks on each other, when read against their corresponding micrometers, to reach a new threshold of accuracy. Then, in 1750, John Bird completed a new quadrant for James BRADLEY, Halley's Greenwich successor. Bird's 8 ft quadrant carried a similar specification to Graham's, although its 90° scale incorporated no trisections, for all of its degrees and minutes were struck off from a perfectly straight 'scale of equal parts' by means of beam compasses. Bird's quadrant's 1 arcsec accuracy was also achieved by an all-brass construction, which avoided bimetallic tensions, and by the fitting of superior micrometers. It was this quadrant that enabled Bradley to perform the most accurate angular measurements made to date, while a comparable 6 ft Bird quadrant made it possible for Tobias MAYER of Göttingen to compile the first set of lunar tables that were sufficiently accurate for the longitude at sea to be found.

The circle

By the 1770s, however, inventors such as Jesse Ramsden and John and Edward TROUGHTON, perhaps inspired by the published researches of Michel Ferdinand d'Albert d'Ailly, duc de Chaulnes, had come to realize that full-circular instruments were the only way forward when it came to achieving ever more critical accuracies, for an astronomical circle was capable of being checked and re-checked by means of reticule microscopes set at 180°, 90°, 60° and 30° across its radius in a way that a quadrant was not. Ramsden's 36 in theodolite for the early Ordnance Survey, 1787, and his 5 ft circle for Guiseppe PIAZZI in Palermo, Sicily, in 1789, were the first circular instruments to demonstrate the spectacular accuracy breakthrough permitted by the full-circle design. Subsequent circles made by Edward Troughton for John POND, Stephen Groombridge and others up to 1806 all demonstrated that the 1 arcsec ceiling of the quadrant could be surpassed to 0.1 arcsec with the circle.

All of these circles had to be meticulously hand-graduated, although in 1766 Jesse Ramsden had begun experiments whereby a 36 in diameter, finely denticulated wheel, acted on by a precision worm-wheel, could 'mass produce' nautical instruments, such as sextants, that had become so necessary in finding the longitude by lunars. By the 1840s, however, the vastly increased machine-tool skills of the industrial revolution enabled William Simms to engine-divide the 6 ft transit circle which the Astronomer Royal George Biddell AIRY had designed, and commissioned into the Royal Observatory, in 1850. When fully adjusted, the Airy Transit Circle, by Ransome & Simms, could measure meridian angles to 0.01 arcsec and was to remain in constant daily use until 1954.

Between 1680 and 1850, it was British craftsmen who made most of the major innovations in graduated instruments, although by 1850 the firms of George Friedrich von Reichenbach and Traugott Lebrecht Ertel were producing superlative instruments for an increasingly international market. It was the advent of photography that fundamentally altered the traditional uses of graduated instruments. Also, especially after the establishment of the *Carte du Ciel*, 1887, measurements taken from a photographic plate, rather than through a telescope eyepiece, showed the way forward when it came to measuring star positions.

The European Space Agency's 'HIPPARCOS' stellar photographic project now represents the state of the art in contemporary angle measuring, while even the Royal Greenwich Observatory's last great circle, the Carlsberg Meridian Circle on La Palma in the Canary Islands, measures and records the meridian passage of stars under computer control.

Bibliography

Bennett J A 1987 *The Divided Circle: a History of Instruments for Astronomy, Navigation and Surveying* (Oxford: Phaidon-Christie)

Chapman A 1995 *Dividing the Circle: a History of Critical Angular Measurement in Astronomy, 1500–1850* 2nd edn (Chichester: Praxis-Wiley)

Allan Chapman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Divini, Eustachio (1610–85)

Telescope-maker, born in San Severino (Ancona), Italy, he made telescopes, some as long as 16 m; he made astronomical observations with some of them, including a map of the Moon, and a work on Saturn.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dixon, Jeremiah (1733–79)

Surveyor and astronomer, born in Bishop Auckland, Durham, England. In 1760 the Royal Society chose CHARLES MASON for an expedition to observe the transit of Venus, and Mason chose Dixon as his assistant, successfully observing the transit at the Cape of Good Hope on 6 June 1761. In 1763 Mason and Dixon traveled to Pennsylvania and Maryland, to define the disputed boundary between the two states; this established the *Mason–Dixon* line, which became synonymous with the boundary between both the political and geographic north and south of America. In 1769 Dixon observed the second transit of Venus of the eighteenth century.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dobson Unit

The Dobson unit is a measure of atmospheric ozone content. The total amount of ozone in a column of air is expressed as the thickness of the layer of ozone formed on compressing the column to standard temperature and pressure conditions (1 atmosphere and 0 °C). A layer thickness of 0.01 mm corresponds to 1 Dobson unit (DU). 300 DU is a typical global average.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dobsonian Telescope

A simple, cheap and portable form of Newtonian reflecting telescope mounted on an altazimuth mounting. Attached to a simple fork mounting that rotates in altitude (elevation) and azimuth (parallel to the horizon), with Teflon strips as the bearing surfaces, a Dobsonian telescope, provided it is well balanced, can be moved quickly and smoothly by hand to point towards any part of the sky. Because of its simple mode of construction, aperture for aperture, the Dobsonian is significantly cheaper to construct than other designs. Consequently this design enables amateur astronomers to acquire instruments of relatively large aperture and light-gathering power at relatively low cost.

See also: altazimuth mounting, light-gathering power, Newtonian telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dollond, John (1706–61)

Instrument-maker, born in London, Dollond was the first manufacturer to produce achromatic lenses commercially, starting in 1758. The achromat consisted of two lenses in different types of glass. The idea had been put forward by a London barrister, Chester Moor Hall (1703–71), but Dollond carried out the experimental work to realize the achromatic lens. One lens was made of relatively common crown glass, but the other glass, flint, was produced in small quantities. English glassmakers supplied English manufacturers with the best quality and sold rejects elsewhere. Dollond had first choice, and with the protection of his patent exploited his position. But British astronomers had access to excellent instruments and produced excellent science! When the patent expired, the price of achromatic telescopes halved. John Dollond's work was continued by his sons, Peter and John, and by the latter's nephew George Huggins (who, such was the Dollond family's prestige, took the name George Dollond).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dome

A circular, very shallow-sided hill on the surface of a planetary body. Lunar domes are found in the maria, and are typically 10–15 km across, with gradients of no more than a few degrees. They are volcanic in origin, and some have a small pit at their center; they may mark the tops of the volcanic vents through which the lava that formed the maria emerged.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dominion Astrophysical Observatory

The Dominion Astrophysical Observatory began operating the 1.8 m telescope in 1918, added a 1.2 m telescope in 1962, and then joined the National Research Council of Canada in 1970. The site near Victoria, BC now contains the headquarters of the HERZBERG INSTITUTE OF ASTROPHYSICS, the Institute administration, groups supporting optical astronomy, submillimeter astronomy, instrumentation, and the Canadian Astronomy Data Center.

For further information see
http://www.hia.nrc.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dominion Radio Astrophysical Observatory

The Dominion Radio Astrophysical Observatory began operating in 1959, and joined the NATIONAL RESEARCH COUNCIL in 1970. It became part of the Herzberg Institute of Astrophysics in 1975. The site near Penticton, BC has a 26 m radio telescope, a seven-antenna synthesis telescope on a 600 m baseline and two telescopes dedicated to monitoring the solar radio flux at 10.7 cm. This part of the Institute also supports the development of correlators and the design of future radio telescopes.

For further information see http://www.hia.nrc.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Donati, Giambattista [Giovanni Battista] (1826–73)

Astronomer, born in Pisa, Italy. Director of the observatory at Florence, he discovered the brilliant comet (*Donati's Comet*) of 1858. He was the first to observe a comet's spectrum (Tempel's Comet of 1864). Inspired the foundation of Acretri Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Doppelmayr, Johann Gabriel (1671–1750)

Born in Nuremberg, Germany. Widely traveled, he wrote on astronomy, spherical trigonometry, sundials and mathematical instruments, as well as biography and history. He made no discoveries himself, but successfully disseminated the work of others.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Doppler Effect

The change in the observed frequency and wavelength of a source of radiation resulting from its motion towards, or away from, the observer. If a source is approaching, the frequency of the radiation is increased (more wavecrests per second reach the observer than would be the case if the source were stationary relative to that observer) and its observed wavelength decreased. If the source is receding, the frequency is decreased and the wavelength of the radiation is increased.

The principle was suggested in 1842 by the German mathematician, Christian Doppler (1803–53). In 1848, the French physicist Hippolyte Fizeau (1819–96) showed that the wavelengths of spectral lines would be affected in the same way. Consequently, lines in the spectrum of a receding source would be displaced towards the longwave (red) end of the spectrum and those in the spectrum of an approaching source would be displaced towards the shortwave (blue) end. The term 'redshift' is applied to the former case and 'blue shift' to the latter.

The observed wavelength, λ', differs from the rest-wavelength (the wavelength measured in a frame of reference in which the source is stationary), λ, by an amount $\Delta\lambda$, which depends upon the radial velocity (speed directly towards or away from the observer) of the source. For velocities, v, which are small compared with the speed of light, c, the wavelength shift is given by:

$$\frac{\Delta\lambda}{\lambda} = \frac{\lambda' - \lambda}{\lambda} = \frac{v}{c}.$$

If the velocity of the source relative to the observer is an appreciable fraction of the speed of light, the relativistic expression:

$$\frac{\Delta\lambda}{\lambda} = \left(\frac{c+v}{c-v}\right)^{1/2} - 1$$

must be used.

The radial velocity of a light source may be determined by comparing the observed wavelengths of its spectral lines with the wavelengths that those lines would have if the source were stationary. The Doppler effect, therefore, provides the means by which the radial and rotational velocities of stars and galaxies may be determined.

See also: redshift, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Doppler, Johann Christian Andreas (1803–53)

Born in Salzburg, Austria, Doppler studied and taught mathematics in Vienna. On the verge, because of economic hardship, of emigrating to America, he was offered posts in Prague. Despite huge teaching loads, he was able to carry out some research of his own (in the face of complaints of neglect by his students). In 1842 read a paper to the Royal Bohemian Society 'On the colored light of the double stars and certain other stars of the heavens', articulating the *Doppler principle* which relates the frequency (or wavelength) of a source to its velocity relative to an observer. Doppler showed that there would be a change of frequency of both light and sound waves if the source were moving; the change of frequency (or wavelength) is called the *Doppler shift*. He tried to illustrate his theory with an application to the colors of double stars, but the effect is actually too small to produce a major shift of color. He foresaw that his effect 'will in the not too distant future offer astronomers a welcome means to determine the movements and distances of distant stars'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dorado

(the Goldfish, or Swordfish—sometimes shown as Xiphias; abbrev. Dor, gen. Doradus; area 179 sq. deg.) A southern constellation which lies between Pictor and Reticulum, and culminates at midnight in early December. It was first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, the brightest stars in Dorado are α Doradus, magnitude 3.3, and β Doradus, one of the brightest Cepheid variables (range 3.5–4.1, period 9.84 days). There are no other stars brighter than fourth magnitude. The constellation contains most of the Large Magellanic Cloud (LMC, which extends into the neighboring constellation of Mensa), in which interesting objects include S Doradus, an extremely luminous eruptive variable (range 8.6–11.5), and NGC 2070 (the Tatantula Nebula), a large (40′) emission nebula centered on the star 30 Doradus. The part of the LMC in Dorado was also the location of Supernova 1987A.

See also: Large Magellanic Cloud, Supernova 1987A, Tarantula Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Double Star

The description 'double star' is given to any pair of stars that appear close together in the sky. Double stars fall into two distinct classes: those pairs that are actually close together in space and are therefore physically linked, and those that may be very far apart in space but happen to lie in almost the same direction.

A pair of stars that really is close together has a common gravitational field and is termed a 'binary system'. The term 'physical double' is also sometimes used. Although there are often two principal components, many systems have additional members and so constitute a 'multiple star' system. Stars that appear to be double only because they happen to lie in almost the same line of sight are termed 'optical doubles'. There is no way of distinguishing between the two classes from a single observation, but while the relative position of the components of an optical double remains fixed, or changes by the straight-line relative proper motion vector, members of a binary system are in orbit around a common center of gravity and their accelerating orbital movements can be revealed by an extended series of observations. The orbital periods are usually very long, measured in decades or even centuries, so that these observations must be made over a correspondingly long period of time.

Statistical analysis shows that a large majority of the closer double stars are physical binary systems rather than optical pairs, but many double stars are difficult to identify with certainty as close binary or multiple systems, unless they are included in a major program of observation which will require the use of a powerful telescope for extended periods every few years throughout much of its working life. Such lengthy programs can rarely be allowed to take precedence over work that offers more immediate results, unless the suspected double stars in question are of particular interest, for example as members of an important moving cluster of stars.

See also: binary system, multiple star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Double Stars: Visual and Imaging Observations by Amateurs

Measurements of the position angle and separation with time of visual double stars allow the apparent orbit of the two stars to be calculated and hence the total mass of the two stars derived (figures 1 and 2). It is still one of the best ways of calculating stellar masses but the long period of many visual binary systems (typically a few hundred years) means that the number of reliable orbits is still very small (currently about 50).

Amateur observers have a long and distinguished record in the annals of double star observing. S W Burnham started out with a small refractor and later made it his life's work, holding down a day job as a clerk in a Chicago court and at weekends observing at Yerkes Observatory. He later became a professional. In England, T H E C Espin discovered several thousand faint and wide pairs with his reflectors and significant contributions were also made by W R Dawes (both, incidentally, clergymen). The greatest amateur was the Frenchman Paul Baize, by day a pediatrician, who made 24 000 measurements and computed nearly 500 orbits and is tenth on the all-time list of observers.

The filar micrometer (figure 3), invented in the 1640s and still in use today, is one of the chosen instruments for those who want to push their telescopes to the limit of resolution. The human eye is still the best detector for pairs which are close or unequal, and a good 35 cm telescope will allow pairs as close as 0.25 arcsec to be measured. Another effective instrument is the double-image (or Lyot) micrometer which is perhaps better than the filar for small separations but cannot match the range of separations that the filar can handle.

While many bright pairs have orbits which require little further adjustment there are a large number of fainter pairs which require attention. In addition the

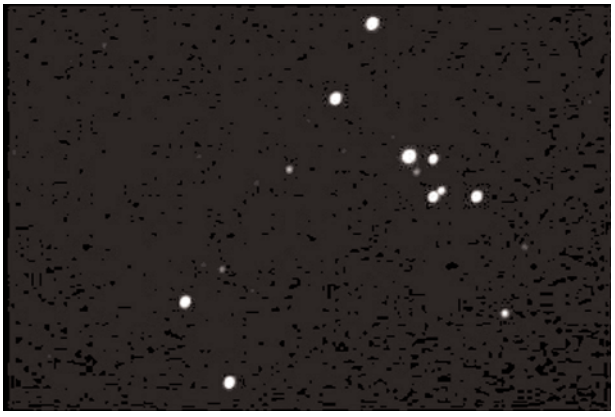


Figure 1. The multiple star HJ 684 located in the cluster NGC 1662 in Orion, imaged by Martin Nicholson. The bright close pair center right is CD—magnitudes 10.3 and 12.5, PA is 308.72°, and separation 10.26 arcsec. These components were previously measured in 1910 and 1987 only. The field size is 10 arcmin × 6 arcmin. North is to the top and east to the left. This is one of six 2 s exposures. Credit: Martin Nicholson, Daventry Observatory.

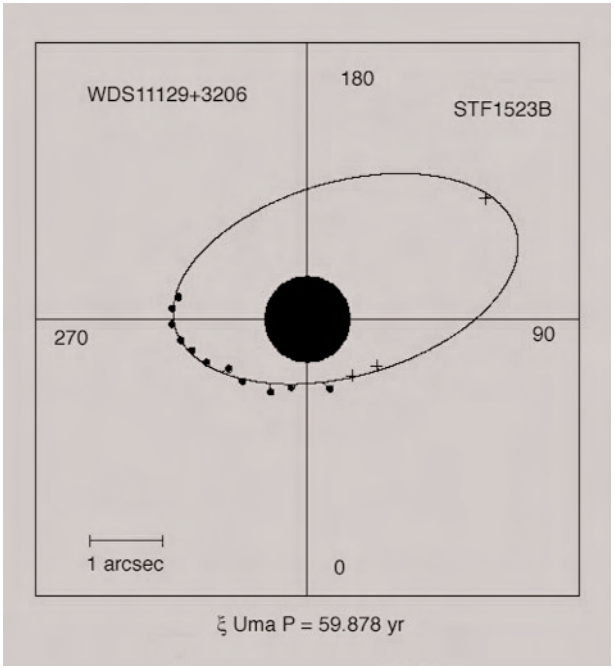


Figure 2. The pair ξ UMa—observed by R W Argyle between 1970 and 2002. The full curve shows the current orbit for this pair. The full dots represent means of five or more nights, the crosses one to five nights. The solitary single cross is an observation made in 1970 and the motion is clockwise. The radius of the central circle is the angular resolution of the telescope—in this case 0".56.

Figure 3. The RETEL filar micrometer used by the writer. It uses parallel wires (one movable and one fixed) to measure separation and a third wire at right angles to measure position angle. The whole field is illuminated by means of a variable-intensity red light-emitting diode. Credit: Richard Sword, Institute of Astronomy, Cambridge.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Hipparcos satellite has discovered thousands of new pairs most of which have not yet been followed from the ground and all of which need to be.

While the filar micrometer has been the mainstay of this work for centuries, the advent of charge-coupled device (CCD) cameras means that there is an opportunity for amateurs equipped with small to medium apertures to make substantial contributions (see CCDs, FILTERS AND CCD PHOTOMETRY FOR AMATEURS). In particular there are many thousands of faint pairs on the various Schmidt sky surveys which are completely unrecorded and measured. Using a 12 in Schmidt–Cassegrain and an SBIG CCD for instance, Martin Nicholson, working at Daventry near Northampton, has found many new pairs using the Guide Star Catalogue. Pairs as close as 5 arcsec and as faint as magnitude 15 can be imaged routinely with his 12 in Meade LX200 and SBIG ST7-E CCD. Reduction of the CCD frames is made using the IRAF suite of image-processing routines.

Another advantage of the CCD camera is that, with the appropriate filters, it can be used to determine the colors of each component of resolved double stars. Many pairs in the Washington Double Star catalogue have not been observed photometrically and much remains to be done here.

In the field of double star observing, there is much that modern amateurs can do. They can continue to monitor the longer-period pairs and lay down measures for future orbital determination. The United States Naval Observatory webpage points to all the data needed. The WDS catalogue is as inclusive as its maintainers can make it (it currently contains over 99 000 entries) and the Sixth Orbital catalogue lists all the orbits deemed worthy of printing. The quality of each orbit has been determined—grade 1 is definitive and grade 5 is indeterminate. Any pairs with orbits of grade 2 or lower could usefully be observed.

Amateurs with a large enough aperture (30 cm or more) could examine and measure the new double stars in the Hipparcos or Tycho catalogues. All are listed in the WDS catalogue which is available both on-line and as a CD-ROM which can be requested free of charge from USNO.

In addition, the power of the modern Schmidt surveys can be utilized to measure and catalogue many faint pairs which have been ignored up until now. Finally, a CCD could be used to great effect on a photometric program of double stars. Many pairs have the magnitudes assigned to them by their discoverers—usually eye estimates. Little is known about the colors of these stars or how many are variable. With a double star there is a built-in comparison for photometry.

Those intending to make a serious contribution must consider how to publish their measurements. These can be submitted to professional journals such as *Astronomy and Astrophysics* while specialist amateur groups such as the Webb Society, the Iberoamerican League for Astronomy (LIADA) and the Société Astronomique de France also offer this service.

Bibliography

Iberoamerica League of Astronomy (LIADA), contact Francisco Rica, e-mail FRICA0@terra.es

Lord C, website http://www.brayebrookobservatory.com (a good site containing a comprehensive description of the Lyot micrometer)

Société Astronomique de France—Double Star Section, contact Dr Edgar Soulié, e-mail SOULIE@drecam.cea.fr (long-established French group)

United States Naval Observatory website http://ad.usno.navy.mil (the best source for data pertaining to visual double stars)

Webb Society website http://www.webbsociety.freeserve.co.uk (encourages observation and publishes measures on a regular basis)

Robert Argyle

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Douglass, Andrew Ellicott (1867–1962)

Astronomer, born in Windsor, VT, investigated the relationship between sunspots and climate by examining the growth rings of ancient Arizona pine and sequoia and coined the term 'dendrochronology' for the method by which archaeologists and historians could use tree-rings to date wooden remains.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

DQ Herculis

The old NOVA DQ Herculis (Nova Herculis 1934) is the prototype of its CATACLYSMIC BINARY sub-class, distinguished by a rapidly rotating weak-field magnetic WHITE DWARF accreting primary. The magnetic moments in the DQ Her white dwarfs are weak enough (10^{32}–10^{34} G cm^3) that the Keplerian-disk ACCRETION flow is broken up only a few stellar radii above the white dwarf surface. The infall is still energetic enough that accretion columns (or curtains) emit significant soft x-ray/UV radiation. These rotating radiation sources illuminate and photoionize the disk surface as a sweeping beam. The subsequent electron–ion recombination yields a sweeping wave of Balmer emission from the disk surface, and this 'reprocessed' radiation can in certain geometries be observed as a periodic photometric signal.

Table 1. DQ Her system parameters.

Orbital period	P (hr)	4.56
Mass ratio M_2/M_1	q	0.66
Primary mass	$M_1/M_\odot$	0.60 ± 0.07
Secondary mass	$M_2/M_\odot$	0.40 ± 0.05
Orbital separation	$a/R_\odot$	1.41 ± 0.05

DQ Her itself is an ECLIPSING BINARY system with $P_{\rm orb} = 4^{\rm h}39^{\rm m}$. Other observationally determined system parameters are given in table 1. In the late 1950s the system held for many years the record for the shortest photometric period in an astronomical source—the 71-s white dwarf spin period—which has an observed amplitude of roughly 1%. What remains remarkable today is the system's high inclination of $i \approx 89.5^\circ$. The white dwarf primary is eclipsed by the edge-on ACCRETION DISK, and only a thin sliver of the back face of the the accretion disk contributes to the pulsed radiation received from DQ Her. Variations in the phase of the 71-s signal during eclipse ingress and egress—first detected in 1972—led to the rotating magnetic accretor model for DQ Her and related systems. Significant variations in phase observed *outside* eclipse result from the self-eclipse of a non-axisymmetric disk, and provide a means to probe the time-averaged structure of an accretion disk at high state, i.e. a disk that has a high mass-transfer rate and high viscosity.

Eclipse-related phase variations

Figure 1 shows schematically but approximately to scale a $q = 0.67$ system viewed from an inclination angle of $i = 80^\circ$, and helps convey the origin of the eclipse-related phase shifts. In the figure, the mean opening angle and radius of the reprocessing region are indicated (see also figure 2), and the black wedge represents the soft x-ray beam sweeping in a prograde direction. If we assume for now that the disk has a constant radius and opening angle for all azimuthal angles, then out of eclipse the times of pulse maxima occur when the beam is centered on the back face of the disk. As the disk reprocessing region is covered during eclipse ingress, the pulses appear to arrive progressively earlier as the latter portion of the pulses are eclipsed by the secondary, arriving one-quarter cycle early just before the total eclipse of the reprocessing region. During egress, the situation is reversed: the first pulse arrives one-quarter cycle late relative to the out-of-eclipse ephemeris, and pulses arrive progressively earlier as the reprocessing region is uncovered.

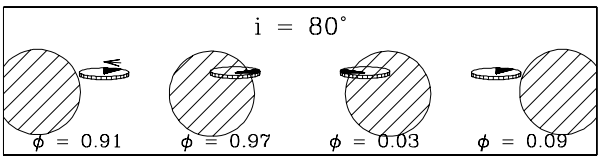


Figure 1. A line drawing (approximately to scale) of DQ Her passing through eclipse, as viewed from an inclination angle of $i = 80^\circ$. The EUV/x-ray beam is represented as a wedge on the disk surface, with the direction of rotation indicated with the arrow. The accreting white dwarf is located at the apex of the wedge. During ingress (egress) the reprocessed pulses will arrive earlier (later) than when out of eclipse. Note that the non-zero disk opening angle results in a front/back asymmetry in the visibility of the disk surface. For inclinations $i \geq 86^\circ$ only a thin strip at the top of the back rim is visible—the white dwarf is eclipsed by the disk.

If as in figure 1 the opening angle of the reprocessing region was constant, then the time center of the pulse would always correspond to the beam passing the back center of the disk, and the only out-of-eclipse phase variations would be those arising from the light travel time delay across the orbit. However, DQ Her exhibits significantly larger variations in the phase of the 71-s signal, and these must result from variations in the rim height of the reprocessing region.

The time-averaged disk structure

Numerical hydrodynamical simulations (smoothed particle hydrodynamics, SPH) of a three-dimensional disk at equilibrium yield an estimate of the radial extent of the reprocessing region and any dependence of the region's rim height as a function of azimuthal angle. Figure 2 shows a top view (x–y projection) of 20 000 SPH simulation particles, center of mass, the stellar surfaces, the orbit with $P_{\rm orb} = 71$ s, the radial extent of the reprocessing region, and the disk opening angle versus azimuthal angle results. Also shown are the viewing angles at the beginning (end) of eclipse ingress (egress). These are consistent with the size of the reprocessing region determined by finding the disk inflection point beyond which surface elements are eclipsed from the primary by intervening disk material. The reprocessing region is found to be nearly circular, simplifying the analysis to modeling a cylinder of non-uniform height. The heavy solid line in the figure shows the opening angle of the reprocessing region (height of the cylinder) as a function of azimuthal angle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

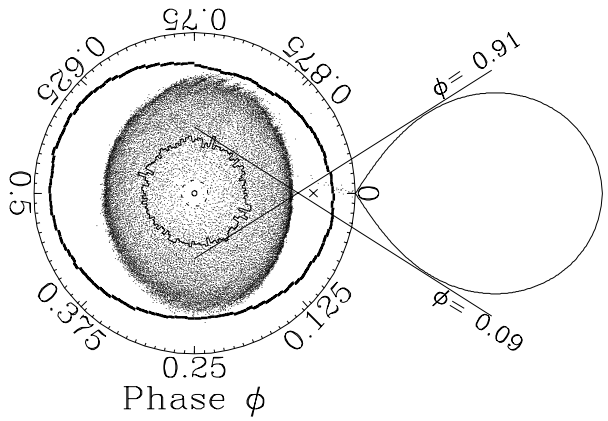


Figure 2. The stellar surfaces and 20 000 SPH (x, y) positions are shown, in addition to the system center of mass '×', the radius with orbital period of 71 s (dotted line), the radius of the reprocessing region, and the opening angle of the reprocessing region (heavy line). The angular scale shows units of phase centered on the primary, and viewing angles of $\phi = 0.91$ and 0.09 are shown.

There are several features of note in this figure. First, most of the mass—evident from the surface density of points—flows in the non-circular outer streamlines in the disk. This is likely to be a general result for a disk in a permanent high state. Second, and also likely to be a general result, the reprocessing region is nearly circular with a radius roughly half the total radius of the disk. The opening angle of the reprocessing region is largest in the quadrants where the disk radius is smallest, and vice versa. This result is a consequence of the requirement that the angular mass flux must be constant for all azimuthal angles in an equilibrium disk. The double-sinusoid shape of the rim of the reprocessing region will affect the observed phase of the pulse produced by the sweeping beam of ionizing radiation.

The O–C diagram

The O–C phase diagram of the observed 71-s signal averaged over several orbits is obtained by fitting a 3-cycle sine curve in a window that shifts by one cycle between fits. Modeling the reprocessing region as a cylinder of variable height with geometry fixed in the co-rotating frame and determined by time-averaging the hydrodynamical results (figure 2), it is possible to calculate the expected O–C phase diagram by integrating over the visible surface of the back cylinder wall.

The observational and numerical results are shown in figure 3—the agreement is remarkable. The best match occurs for the startlingly small range $89.5^\circ \leq i \leq 89.7^\circ$. Note that if the mean opening angle $\langle\theta\rangle$ given by the numerical simulations is unphysically large, then the estimates of the system inclination would need to be even closer to 90°

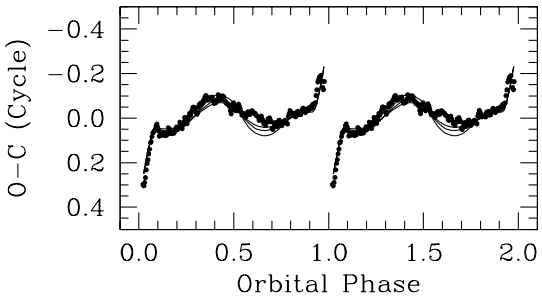


Figure 3. The O–C diagram calculated from the SPH disk structure as a function of orbital phase and for inclination angles ($i = 89.5$, 89.6, and 89.7°. The match is excellent except near orbital phase $\phi = 0.6$.

Spectroscopy

Hubble Space Telescope time-series spectral observations of the Vis/UV continuum and emission lines show that while the UV continuum source is completely eclipsed and the He II emission region nearly so, the N V and Si IV eclipses are deep but not complete, and C IV lines show only a 75% drop. Stellar winds have been suggested as the source of the emission lines, but the behavior is also simply consistent with the geometry in figure 2 if the outer disk is a source of optically-thin emission. The figure shows that while the reprocessing region is completely eclipsed, the outer disk is not.

The general model for DQ Herculis is now reasonably well in hand, but additional observations and simulations are required to understand completely this unique system.

Bibliography

Wood M A *et al* 2000 *Baltic Astronomy* in press
Patterson J 1994 *Publ. Astron. Soc. Pacific* **106** 209
Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press) ch 9

Matt A Wood

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Draco

(the Dragon; abbrev. Dra, gen. Draconis; area 1083 sq. deg.) A northern constellation which lies between Ursa Major and Cepheus, and encloses Ursa Minor on three sides. The center of the constellation culminates at midnight in May, though part of it is on the meridian from mid-February to late July. Draco represents the dragon Ladon in Greek mythology, which guarded the golden apple tree that was given by Mother Earth to the goddess Hera when she married Zeus, and which was killed by Hercules, who is usually shown in early celestial charts with a foot upon Draco's head. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large but inconspicuous constellation, the brightest stars in Draco are γ Draconis (Eltanin), magnitude 2.2, η Draconis, magnitude 2.7, and β Draconis (Rastaban or Alwaid), magnitude 2.8. There are ten other stars of magnitude 4.0 or brighter, including α Draconis (Thuban or Adib), magnitude 3.7, which was the northern pole star about 4800 years ago. Other interesting stars include the binaries ν Draconis, which has bluish-white/white (A4 and A6) components, both of magnitude 4.9, separation $62''$, one of which has an unseen companion, and μ Draconis, which has pale yellow (F7) components, both of magnitude 5.8, separation $2.2''$, and a third component of magnitude 13. Another interesting object is NGC 6543, a ninth-magnitude planetary nebula, in which, in 1864, William Huggins first observed 'nebular lines' typical of a large mass of hot luminous gas.

The meteor shower known as the Giacobinids (Draconids) appears to radiate from the constellation.

See also: Gamma Draconis, Giacobinids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Drake, Frank Donald (1930–)

Born in Chicago, in 1958 he joined the newly founded National Radio Astronomy Observatory (NRAO) in Green Bank, West Virginia, and attempted the first search for extraterrestrial broadcasts, a two week observation of the stars Tau Ceti and Epsilon Eridani. Developed the *Drake equation* as a way to estimate how many intelligent, communicating civilizations there are in our galaxy. The Drake equation multiplies the number of stars in the Milky Way by the fraction which have orbiting planets, then by the number of planets per star that are capable of sustaining life, then by the fraction of these planets where life actually evolves, then by the fraction of these where the life is intelligent and communicates, and then by the fraction of the planet's life during which the communicating civilizations live. This represents an estimate of the number of communicating civilizations in the Galaxy. The real value of the Drake equation is not in the answer itself, but the questions that are prompted when attempting to come up with an answer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Draper, Henry (1837–82)

American pioneer of astronomical photography, born in Prince Edward County, VA. Made the first photograph of a stellar spectrum (Vega). His assistant WILLIAMINA FLEMING analysed the objective prism photographs of stellar spectra and formed the basis for the classification of stars, with the data eventually published by ANNIE CANNON as the *Henry Draper Catalog of Stellar Spectra*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dreyer, John Louis Emil (1852–1926)

Danish astronomer, born in Copenhagen, who, after working at Birr Castle in Ireland, became director of Armagh Observatory, studying the many nebulae he had previously observed at Birr. At a time when it was not clear whether spiral nebulae were inside our own Galaxy (the Milky Way) or were indeed, as was suspected, island universes outside, he established through careful measurements that they had no proper motion and were likely to be distant. Compiled the *New General Catalogue of Nebulae and Clusters of Stars* (known as NGC) which remains the principal catalog of nebulae and galaxies; it was based on observations by the Herschels (see HERSCHEL FAMILY) and by LORD ROSSE, Dreyer and associates at Birr and Armagh. When government funds for observing were withdrawn from Armagh, as a result of the disestablishment of the Church of Ireland, Dreyer turned to astronomical history, publishing TYCHO BRAHE'S observations and *The History of the Planetary System from Thales to Kepler*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dubhe

The star α Ursae Majoris, apparent magnitude 1.81, one of the leading pair of stars in the Plough known as 'the Pointers'. The name comes from an abbreviation of the Arabic for 'the Back of the Great Bear'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dudley Observatory

The Dudley Observatory, in Schenectady, New York, is a private foundation supporting research and education in astronomy, astrophysics and the history of astronomy. Chartered in 1852, it is the oldest organization in the US, outside academia and government, dedicated to the support of astronomical research. For more than a century it was a world leader in astrometry, with such achievements as publication in 1937 of the *General Catalog of 33,342 Stars for the Epoch 1950*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dumbbell Nebula (M27, NGC 6853)

A planetary nebula in the constellation Vulpecula, position RA $19^{h}\,59.6^{m}$, dec. $+22^{\circ}\,43'$, and the first to be discovered, by Charles Messier in 1764. It measures $8'$ by $4'$ and is of eighth magnitude. The bipolar expulsion of material from the thirteenth-magnitude central star gives it its twin-lobed appearance; it was named by Thomas Webb, though it resembles a bow-tie more than a dumbbell.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dunér, Nils Christoffel (1839–1914)

Swedish astronomer, confirmed Galileo's measurement of the rotation of the Sun spectroscopically with the Doppler shift, and showed the solar rotation to be differential (faster at the equator).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dunlop, James (1793–1848)

Of Scottish origin, Dunlop was Superintendent at the Parramatta Observatory, New South Wales and compiled the *Catalogue of 7835 Stars from Observations made at the Observatory at Parramatta*, for his catalog of 621 nebulae, for his list of double stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dunsink Observatory

Designed by Henry Ussher, the first astronomy professor of Trinity College Dublin, Dunsink Observatory (1783) incorporated modern designs such as ventilation in the observation room and a free-standing telescope support column. Among the past directors were the mathematician W R Hamilton (1827–65) and stellar dynamicist H C Plummer (1912–21). The main remaining instrument is a Grubb 12 ft (3.66 m) refractor (1868). Part of the Dublin Institute for Advanced Studies since 1947, the observatory currently concentrates on high-energy studies of extragalactic objects.

For further information see http://www.dinsink.dias.ie/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dust Tail, Cometary

The part of a comet consisting of dust grains, carried away from the nucleus by the outflow of gas from the coma during a close approach to the Sun, which are propelled away from the comet's head by radiation pressure. Not all comets develop a dust tail (or a gas tail), but both types of tail always point away from the Sun. The combined effects of radiation pressure and the Sun's gravity acting on the dust particles make dust tails curved, in contrast to gas tails, which are straight. Dust tails shine only by reflected sunlight, so their color is yellow. Their length is typically 10^6–10^7 km, and like gas tails they begin to form at around 2 AU from the Sun. Most of the particles in a dust tail are less than 1 μm across. Larger particles remain in the comet's orbit and form meteor streams.

See: cometary atmosphere, cometary tails, meteor stream, Poynting–Robertson effect.

Compare: gas tail, cometary.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dusty Circumstellar Disks

What are circumstellar (CS) disks and why should we care about them? Disks appear at many scales in astronomy—from SPIRAL GALAXIES to PLANETARY RINGS. In the present article our concern is with rings or disks that contain particulate material and which orbit stars at various stages in their life cycles. For main sequence stars such as the Sun, the primary interest in such disks is that they could be related to the formation and/or existence of PLANETS and smaller objects like COMETS and ASTEROIDS. After stars have evolved beyond the main sequence orbiting disks are rare and often portend unusual, unexpected, and still unexplained phenomena.

The discovery of dusty CS disks around stars was perhaps the most notable achievement of the INFRARED ASTRONOMY SATELLITE (IRAS) which flew in 1983. IRAS, a joint venture of the United States, the Netherlands and the United Kingdom, found that some bright nearby stars, including VEGA and FOMALHAUT, emit much more radiation at wavelengths between 25 and 100 μm than can be produced solely by the stellar photospheres. The spectral energy distribution of this excess infrared (IR) radiation is generally consistent with models wherein the emission arises from dust particles in circumstellar disks whose characteristic radii are a few tens or hundreds of astronomical units (AU). These grains absorb the optical and ultraviolet light from the central star and then radiate this energy away in the IR. Such stars have been dubbed 'Vega-like'.

At most stars, the particles emit most vigorously at wavelengths between 25 and 200 μm which are difficult or impossible to study from the ground. So, observational progress has been slow in the years since IRAS flew. Indeed, for about 15 years, excepting one star (β PICTORIS), no one was able to image the distribution of dust particles around the Vega-like stars.

Recently, there have been significant observational breakthroughs. From the ground, with very sensitive detectors on large telescopes, images have, at last, been obtained of disks that orbit a few of the Vega-like stars (including Vega itself). These observations have been of thermal emission from the grains at IR wavelengths between 10 and 25 μm and, also, at much longer wavelengths in the radio band near 1 millimeter. In addition, the NICMOS camera on the HUBBLE SPACE TELESCOPE (HST) has imaged 1–2 μm radiation scattered by grains orbiting in disks and rings around a few stars.

Although most of the detected material in most of the disks consists of small particles, at a few stars modest amounts of gas have also been seen. If one is interested in the formation of comets, asteroids and planets, the dust particles that comprise the disks are relevant to all. On the other hand, the principal relevance of the gas is to the formation of giant gassy planets such as Jupiter and Saturn.

The present article is structured along the lines of STELLAR EVOLUTION. We pick up the story about 10 million years after the formation of a star at the last stages of the pre-main sequence evolution. This is the time when massive disks that surround many T TAURI STARS have, by and large, dissipated and only relatively little gas and dust remain. The formation of larger bodies—'PLANETESIMALS' and planets—is either well along or it will not happen at all. We next consider disks around main sequence stars. These are either remnants of the star formation process or else the debris from destruction of larger objects like comets or asteroids. Evidence for the existence of dusty debris from ancient comets is seen in our own planetary system in the vicinity of the orbit of the Earth.

Finally, we consider stars in post-main sequence stages of evolution—the RED GIANT STARS and WHITE DWARFS. Surprisingly, a small percentage of such stars appears to be surrounded by orbiting dusty disks. The origins of most of these disks are mysteries whose eventual understanding will, no doubt, reveal some interesting and important astronomical phenomena.

We conclude with a brief discussion of some properties of circumstellar dust clouds, in particular, how one may estimate their masses from measured fluxes at submillimeter wavelengths.

The evolution of stars and their circumstellar disks

Pre-main sequence evolution

When intermediate-mass stars form they are often surrounded by massive disks of dust and gas that astronomers can detect via strong IR dust emission and radio emission from carbon monoxide molecules (see YOUNG STELLAR OBJECTS). After a few million years the small particles in the disk that are responsible for the IR emission begin to disappear, either because they clump together or because they are lost from the system. The timescale for loss of small particles can be determined by measuring the amount of material as a function of the age of the young stellar objects.

We want to know how long it takes for the dust particles to disappear because this constrains the time it takes to form planetesimals, i.e. bodies that look like present-day comets and asteroids. We want to know how long it takes the gas to dissipate because this will give an upper limit to the time available to form gas giant planets.

Young stars were initially discovered to lie in or near dusty, interstellar molecular clouds. These stars usually have ages less than a few million years, often the ages are less than one million years. These are the T Tauri stars of about a solar mass or somewhat less and the Herbig Ae

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars that are a few times more massive than the Sun and which will evolve into A-type main sequence stars like Vega and Fomalhaut. It has not been easy to find somewhat older T Tauri stars, say in the 10–20 million year range, because such stars are hard to identify with infrared and optical techniques. Recently, all-sky surveys of x-ray emission, especially with the German ROSAT satellite, have enabled detection of many stars that appear to be old T Tauri stars. Such stars can be discovered because they emit x-rays strongly compared to older, main-sequence stars.

Not only has the ROSAT satellite enabled detection of some oldish T Tauri stars in regions surrounding various dusty molecular clouds, but it has enabled the clear recognition of the nearest region of recent star formation—the TW Hydrae Association—named in honor of its first known member, the classical T Tauri star, TW Hya. Combination of ROSAT data, measurements of lithium abundances from optical spectra, and distance determinations with the Hipparcos satellite, indicate that the TW Hya Association is about 10 million years old and only about 50 pc from Earth. This is more than two times closer to Earth than any other known region of significant star formation. It is remarkable that this closest T Tauri Association has gone unrecognized for so long. The primary reason is that the massive molecular cloud from which the Association stars formed has dissipated; the only clouds that remain in the vicinity are those that orbit the stars, i.e. CS disks.

Because of their proximity to Earth, disks around TW Hya Association stars can be studied with special clarity. One of the most interesting Association members is an A-type star, HR 4796. Initially, IRAS discovered that about 0.5% of the total energy from HR 4796 is emitted at infrared wavelengths—HR 4796 is the dustiest A-type star in the Bright Star Catalog. (The BSC lists various properties of nearly 10 000 of the brightest stars in the sky, most of which are bright enough to be seen with the naked eye.) By comparison, other relatively dusty A-type stars like Vega and Fomalhaut emit a 100 times smaller percentage of their total energy at IR wavelengths.

Recently, the 10–20 μm thermal emission from dust heated by HR 4796 was imaged at telescopes in Hawaii and Chile. At about the same time, an image of a narrow ring of dust particles that scatter light at a wavelength of 1.1 μm was obtained with the NICMOS camera on the HST (see figure 1, top panel). The stable containment of the ring particles and the abrupt inner and outer truncation of the ring around this relatively young star implies dynamical constraints imposed by one or more related, but as yet unseen, bodies. How such bodies, presumably planets, could have formed so quickly at such large distances from the star presents an interesting challange to those who model planet formation.

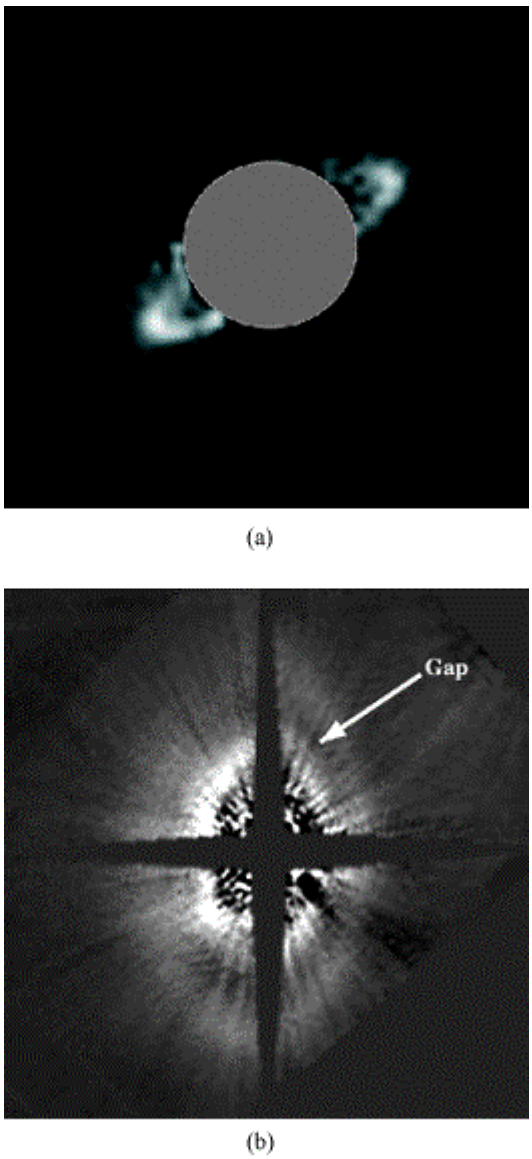


Figure 1. Images obtained at a wavelength of 1.1 μm with the NICMOS infrared camera on the Hubble Space telescope in 1998. The upper panel is HR 4796, a star of age about 10 million years located 67 pc from Earth. The region near the star is unusable and is masked out. The diameter of the surrounding inclined dust ring which reflects 1.1 μm starlight toward the Earth is about 150 AU. The lower panel is HD 141569, a star about 100 pc from Earth whose age is probably less than 10 million years. The diameter of the inclined dusty disk is about 700 AU and there is an indication of a gap between the brightly reflecting dust close to the star and the fainter dust further out. Gaseous carbon monoxide molecules detected in orbit around HD 141569 with a radio telescope seem to be confined primarily to the region interior to the gap. See text for discussion of the structure of the dust at these two stars. (Images courtesy of G Schneider, A Weinberger and P Lowrance.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Another A-type star imaged by NICMOS is HD 141569, seen in the lower panel of figure 1. This star is not a member of the TW Hya Association but it was chosen for NICMOS imaging because it has a lot of surrounding dust—about 1% of its energy is emitted at infrared wavelengths—and also surrounding molecular gas. The mass contained in hydrogen molecules orbiting HD 141569 is about 10 Earth masses or more. (No gas has been seen near HR 4796.)

A potentially very exciting aspect of the NICMOS/HST image of HD 141569 is the apparent gap between inner and outer dusty regions. Such a gap would be anticipated if a substantial planet were forming there. HD 141569 is orbited by a companion star, possibly even two companion stars, a few disk diameters away that could also influence the dynamics of the dusty disk.

The brightest molecular emission seen at a star of age about 10 million years is from TW Hya itself (see figure 2). Future observations of these molecules (and others yet to be detected) with sensitive interferometers that operate at millimeter wavelengths promise to help us to better understand the chemical conditions in a potential protoplanetary disk. The interferometers could reveal changes in chemistry as a function of distance from TW Hya. Such chemical data may be helpful to planetary scientists who try to understand the chemistry of our own forming planetary system 10 million years after the formation of the Sun. Chemistry in disks that surround other T Tauri stars can also be studied with future interferometers, but TW Hya should remain unique because it is the closest classical T Tauri star, so finer details can be resolved in its CS disk and, also, because it is the oldest known T Tauri star that still retains substantial quantities of surrounding molecular gas.

The main sequence

Closer to Earth than the TW Hya Association is the best studied of all the Vega-like stars—β Pictoris, a naked eye star located in the southern hemisphere. The age of this A-type star is uncertain; based on its position in a HERTZSPRUNG–RUSSELL diagram, it may be as young as about 10 million years and just arriving on the main sequence, but it could perhaps be as old as 100 million years. This range of plausible ages is narrowed significantly if β Pic formed at about the same time as two low mass (M-type) stars GL 799 and GL 803; because the three stars appear to be moving through space together, they probably did form together. The estimated age of the two M stars is 20±10 million years. Additional evidence in support of this young age is the fact that β Pic is the second dustiest A-type star in the Bright Star Catalog and it seems clear that the amount of dust around stars declines with age (see below). Proper interpretation of the various phenomena seen at β Pic, in the context of the formation of planetary systems, depend on its age.

The disk around β Pic was imaged at a wavelength of 9000 Å with a camera with coronagraphic optics within a year after IRAS discovered the large amount of excess infrared emission. The image revealed that the Earth lies very nearly in the plane of the β Pic disk which extends many hundreds of AU from the star. The initial coronagraphic studies and many subsequent ones in the optical and infrared from the ground and with the HST have revealed many interesting aspects of the β Pic environment. For example, there is a diminution of dust density within a few tens of AU from the star compared to the dust abundance further out along the disk. This might be due to planet formation, but it could also be due to dust destruction or ejection.

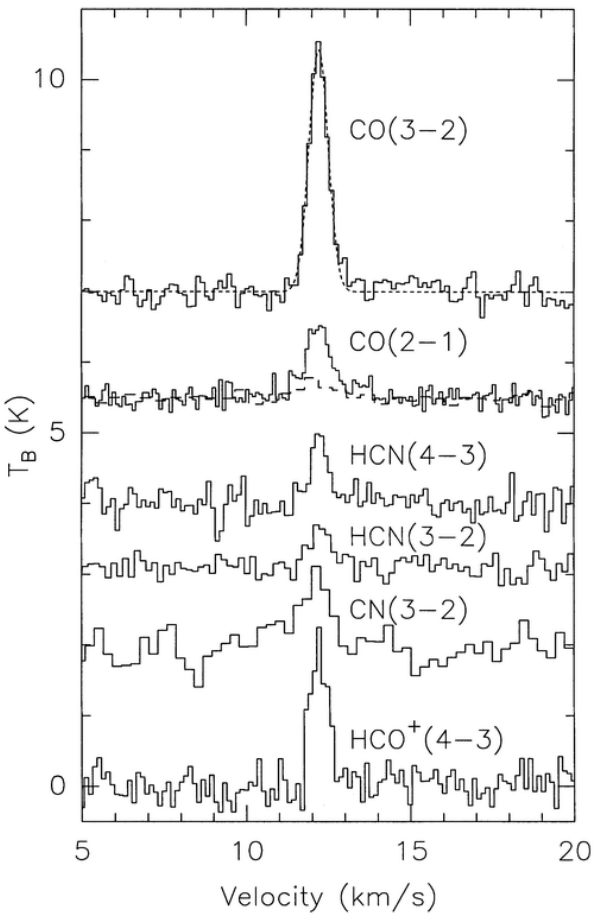


Figure 2. Molecular spectra observed toward TW Hya with the James Clerk Maxwell Telescope at the Mauna Kea Observatory. The abscissa is heliocentric radial velocity and the ordinate is the relative main beam brightness temperature (T_B). The two spectra labeled CO(2-1) are of the ^{12}CO (solid curve) and ^{13}CO (dashed curve) isotopomers (the ^{13}CO spectrum has been expanded along the ordinate by a factor of two). The short dashed curve overlaying the CO(3-2) profile represents the best fit Gaussian function.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Optical spectra of ionized calcium lines show time-variable components. These have been interpreted as due to comet-like objects falling onto β Pic, sometimes as many as a few per week. Modest masses of carbon monoxide (CO) molecules and neutral carbon atoms have been detected in the ultraviolet spectrum of β Pic with HST. It is suggested that these gaseous species result from evaporation of comet-like bodies out to ~100 AU from β Pic. In spite of deep searches, radio astronomers have not yet been able to detect CO near β Pic.

Precise optical images of the disk were interpreted initially as showing a warp, but recent coronagraphic images obtained with the HST suggest that the apparent warp may rather be the superposition of two disk-like components inclined with respect to each other by 4º. Such a structure might be produced by the gravitational influence of a giant planet in orbit around β Pic.

Finally, images obtained at a wavelength of 850 μm show the unexpected presence of an emission peak offset from β Pic by about 34 seconds of arc to the southwest (see figure 3(b)). It is not known if this emission is associated with the star or is, rather, a totally unrelated background galaxy located near the edge of the observable universe! If the emission is associated with β Pic, then it is about 650 AU from the star. It must be due to warm dust grains, since companion stars and planets would radiate too faintly at 850 μm to be detected from Earth. The dust cannot be associated with a companion star, since such a star would easily be seen in the IR. Perhaps there is a brown dwarf or giant planet companion embedded in the dust (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS)? But, if so, then unless β Pic is quite young, it is hard to understand how so much dust could be retained for so long near the companion. Perhaps there is an asymmetrical ring that surrounds β Pic well outside of the inner, previously known disk. But, if so, then how did the outer ring get to be so asymmetric with so much dust toward the south-west and so relatively little toward the north-east?

Observations in 2001 by the European Southern Observatory's New Technology Telescope reveal dusty disks surrounding brown dwarfs in the Orion Nebula, suggesting that they are more star-like than planet-like and have the potential to form accompanying planetary systems. This finding is based on detailed observations with SOFI, a specialized infrared-sensitive instrument at the European Southern Observatory's (ESO) 3.5 m New Technology Telescope (NTT) at the La Silla Observatory.

Moving along to older stars, 850 μm images of three are displayed in figure 3. Vega and Fomalhaut are nearby A-type stars, whereas ε Eri (K-type) is even nearer but with mass only slightly less than that of our Sun. These stars were chosen for imaging because they are among the strongest of the excess IR emitters detected by IRAS. The images seen in figure 3 are the first ever of the dust distribution around these three stars. The easiest to interpret is Fomalhaut which is consistent with a ring of material located about 80 AU from the star. As in the case of β Pic, we view Fomalhaut more or less in the plane of its dusty disk. The dust at ε Eri also is distributed in a ring-like manner, but here we view the disk face on; that is, the dust orbits the star at distance of about 60 AU and nearly in the plane of the sky.

An obvious interpretation of the dust rings at ε Eri and at Fomalhaut is that we are seeing analogs of the so-called KUIPER BELT of comets that orbits the Sun just outside of the orbits of Pluto and Neptune. In the case of ε Eri and Fomalhaut, we are not seeing comets directly, but rather millimeter-size dust particles that are produced by the (partial) destruction of comets, probably many more than exist in the Kuiper belt. Given that the ages of ε Eri and Fomalhaut are estimated to be of the order of a few 100 million years, we may speculate that we are seeing structures that are now quite similar to what our own Kuiper belt must have been like billions of years ago.

The regions of relatively low 850 μm emission inside of the dust rings at ε Eri and Fomalhaut might be caused by the formation of planets. In the case of Fomalhaut, another conceivable explanation for the central cavity could be sublimation of icy dust particles during the lifetime of the disk. But such an explanation will not explain the region of relatively low dust abundance near ε Eri because the star is insufficiently luminous to vaporize icy grains.

The image of ε Eri has a surprising feature—a peak of emission in the ring to the east and slightly south of the star. Assuming that this is a real non-uniformity in the ring structure, then it is not easily explained. Perhaps a planet or a 'cosmic collision' can produce structure in the distribution of dust particles.

Finally, Vega presents yet another mystery. As may be seen in figure 3(c), the peak 850 μm emission seems to lie off to one side of the star. As at β Pic, it is conceivable, albeit unlikely, that the emission peak is due to a distant background galaxy. But if the peak in 850 μm emission is due to dust in orbit around Vega, then, as at ε Eri and β Pic, we do not understand the cause of the emission. In particular, since Vega is probably at least 300 million years old it seems most unlikely that dust could be retained in orbit around a substellar companion for so long.

Vega and Fomalhaut with probable ages of hundreds of millions of years are less dusty than HR 4796 and β Pic whose ages are ten to tens of millions of years. More generally, far IR observations of main sequence stars in nearby young clusters and in the field indicate that as stars age, on average, less and less dust is retained in their vicinity. And for stars older than about 20 million years there is little evidence for orbiting gas. Most

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

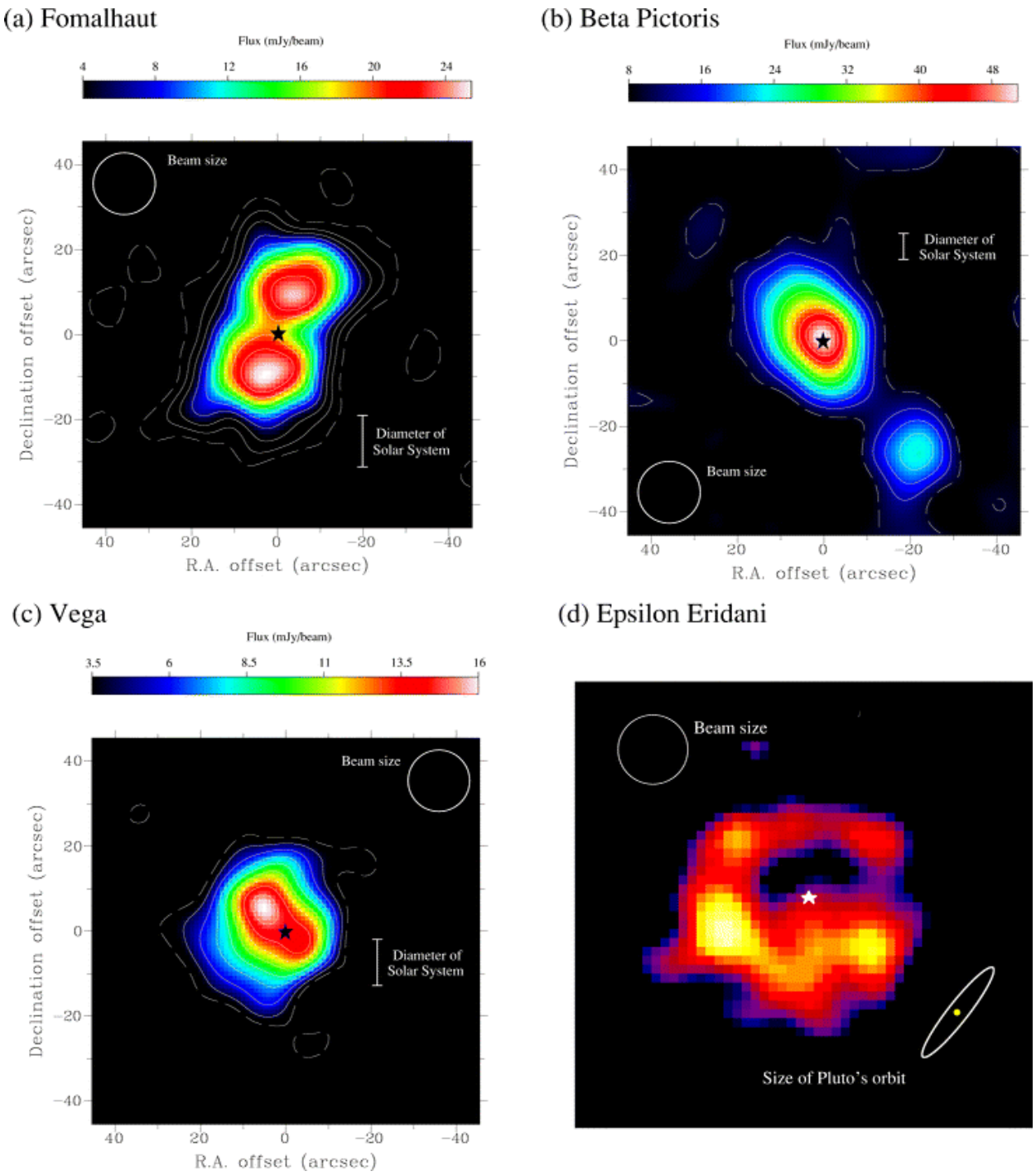


Figure 3. Dust emission around Fomalhaut [(a), upper left], β Pictoris [(b), upper right], Vega [(c), lower left] and ε Eri [(d), lower right] at a wavelength of 850 μm. The images were obtained with the SCUBA bolometer array on the James Clerk Maxwell Telescope. North is up and east is to the left. In each map the star is located at the position of the 'star' symbol. In each panel the diameter of the JCMT beam at half power is shown by the circle. (Images courtesy of W Holland.)

likely then, the preponderance of the gas and dust seen in CS disks around all but the very youngest main sequence stars is the product of destruction of comets, much as the dust component of the ZODIACAL LIGHT seen in the inner parts of our solar system is a product of cometary debris.

First ascent giant stars

The particulate matter in CS disks around main sequence stars may be envisioned as a natural consequence of the star formation process—leftover material that was not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

incorporated into large objects or, alternatively, particles generated by (partial) destruction of the large objects. In either case, by the time an intermediate mass star evolves off the main sequence one expects very little dust to remain. So, the discovery by IRAS of quite large amounts of dust near a small but not insignificant percentage of ordinary first ascent G and K giant stars came as quite a surprise.

Three models have been proposed to explain the presence of substantial dust near these giant stars: (1) mass ejection; (2) destruction of comets and particulate matter in a Kuiper-belt-like region as the central star increases its luminosity as it leaves the main sequence; and (3) stellar heating of dust in random interstellar clouds that, by chance, lie near some G and K giants. None of these models can be readily reconciled with all existing data. So, at this moment, we do not understand the origin and nature of the dust near G and K giant stars. In addition, gas has never been detected at any of these dusty systems.

Asymptotic and post-asymptotic branch giant stars

Space telescopes have detected strong mid- and far-IR emission from luminous, red giant stars with either carbon-rich or oxygen-rich photospheres. In most cases, the IR emission is produced by dust grains entrained in outflowing winds, i.e., not in a CS disk. But a few of these infrared-bright evolved stars have been found to possess peculiar characteristics of various sorts. For example, a handful of stars with carbon-rich photospheres heat nearby, associated silicate dust grains which formed in an oxygen-rich environment. When measured at such stars, the width of CO rotational lines are typically only one or two km s^{-1} wide and much narrower than lines characteristic of outflowing winds. That is, the narrow linewidths indicate velocities characteristic of orbital motion of gas many hundreds of AU from the stars, rather than gas escaping from the stellar surface.

A plausible, but not yet proven, model for most of these stars with unusual infrared and/or CO spectra is the following. The primary, giant star, is a member of a binary system. When the primary ejects large amounts of mass while it is on the asymptotic giant branch, the gravitational field of the secondary deflects some of this outflowing material into a circumbinary disk, diameter ~1000 AU, which is gravitationally bound to the two stars. The companion star necessary in this scenario has already been detected near some, but not all, of the unusual giant stars. At least a few of the disks show evidence for large scale clumpiness, the nature of which is not now understood.

White dwarfs

A dusty CS disk in orbit around a white dwarf could be detected as excess IR emission. At this moment, in spite of searches at hundreds of white dwarfs, we know of only one, G29-38, that is surrounded by dust grains—these emit strongly at wavelengths between 2 and 20 μm. The origin of the dust is a mystery. Perhaps G29-38 is embedded in a moderately dense interstellar cloud and dust from the cloud has been gravitationally drawn into a disk orbiting the star. Or perhaps there has been some kind of cosmic catastrophe near G29-38, say a collision of two moderately massive objects that has partially shattered them. Or perhaps G29-38 is pulling material off of a nearby but unseen companion. Or perhaps one of any number of other exotic hypotheses.

Properties of the circumstellar dust

Astronomers have expended substantial effort to understand the radiation charactertistics of small dust particles. Early work was on grains in interstellar clouds—such particles typically have submicron dimensions. Observed CS disks have higher densities than interstellar clouds and larger particles are present, either from the sticking together of small particles or from destruction of larger bodies. Perhaps the most interesting astrophysical aspects of the CS dust at a given star are its total mass and spatial distribution. Images such as figure 3 give a large scale picture but, since the dust is heated by the central star, the dust visibility depends on its temperature as well as its spatial location. For example, the sharp cutoff in 850 μm emission outside of the bright 'blobs' seen to the north and south of Fomalhaut is due in part to diminishing dust temperature as well as dust density.

At a given distance from a star the temperature of a dust particle depends on its size relative to the wavelength of the peak of its thermal infrared emission. If the particle is small with respect to the wavelength of its peak IR emission, then such a particle will radiate inefficiently and its temperature will rise above the temperature a large particle would have at the same location. Near this peak of emission, for large and small particles alike, the amount of emission depends sensitively on the dust temperature because of the exponential in the denominator of the Planck function.

So, the appearance of the dust emission at a given wavelength results from a convolution of dust mass and temperature. To derive dust mass it is best to use submillimeter data (as in figure 3) because:

(1) one is on the Rayleigh–Jeans portion of the Planck curve so that intensity is only linearly proportional to dust temperature;
(2) optical depths are small so one is sensitive to all of the mass; and
(3) one can see all of the dust, hot and cold and big and little, at least up to a point.

Particles that are much larger than the wavelength under investigation have relatively low surface-to-volume ratio and the opacity per gram of material

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

diminishes. Therefore, for measurements at, say, one millimeter wavelength, 'dust grains' larger than about 1 cm in size contribute little to the emission compared to smaller particles, even if most of the mass is in the larger ones.

Hence, submillimeter fluxes can provide only lower bounds to mass in orbit around stars—such observations will not reveal the presence of large objects like comets, asteroids and planets, at least not directly. The existence of such bodies can be deduced indirectly if it can be shown that the detected dust results from breakdown of larger objects. This is likely the case at stars as old as Vega and Fomalhaut. For younger stars like β Pic or the pre-main sequence stars discussed above, we still do not know if the detected dust is left over from the star formation process or is the debris from breakdown of larger objects or some combination of the two.

As an illustrative example, consider the submillimeter flux (F_ν) measured at Earth that is emitted by an optically thin cloud or disk a distance R away.

$$F_\nu = \pi B_\nu 4\pi a^2 Q_\nu n_{gr} V / 4\pi R^2. \quad (1)$$

Here πB_ν is the surface emittance of a single grain per unit area, $4\pi a^2$ is the emitting area of a grain of radius a, n_{gr} is the number of grains per unit volume, and V is the total volume of the cloud. Q_ν is the efficiency with which the grains absorb and emit waves of frequency ν. The total mass (M) of dust in the disk may be written

$$M = \rho N 4\pi a^3 / 3 \quad (2)$$

where N is the total number of grains in the cloud, and ρ is the density of a typical grain. Combining equations (1) and (2) yields

$$M = F_\nu R^2 / B_\nu k_\nu. \quad (3)$$

At submillimeter wavelengths, where $h\nu \ll kT$, the Planck function, B, can be written simply as $2kT/\lambda^2$. Thus, the calculated grain mass is not sensitively dependent on the grain temperature (which is often not well determined).

In equation (3), k_ν is the opacity, that is, the absorption cross section per gram of material at frequency ν. For typical grain materials, the absorption cross section can be much smaller than the total extinction cross section—that is, the grains often have large albedos at wavelengths near a millimeter. The opacity may be written as

$$k_\nu = (\pi a^2 Q_\nu)/(4\pi a^3 \rho / 3) = 3Q_\nu / 4a\rho. \quad (4)$$

This is the absorption per unit mass of dust, whose units are usually given in the literature as $cm^2\ g^{-1}$. Sometimes the opacity quoted in the literature is the absorption per unit mass of dust and gas. To obtain this from submillimeter dust emission (which is the measured quantity) requires an assumption about the ratio of the mass contained in gas to that contained in the dust particles; this ratio is often taken to be 100, the value typical of the interstellar medium. However, in CS disks the gas-to-dust ratio is often much smaller than 100 and it is usually uncertain or unknown, so it is better to quote opacities in terms of dust mass only.

For quite some time, k_ν at wavelengths near one millimeter was uncertain by at least an order of magnitude, resulting in corresponding uncertainty in dust masses. Recent laboratory and theoretical efforts have improved the situation so that now k_ν is probably known to a factor of 3 or so. In any event, one cannot obtain total masses contained in CS disks because objects of dimensions much larger than a few cm are effectively invisible even if they carry most of the mass of the disk (because large objects have such small opacities per gram).

That said, submillimeter fluxes from stars like β Pictoris and HR 4796 that are just arriving on the main sequence indicate total dust masses about 20 times the mass of our Moon. The most common particle sizes lie in the range between 10 μm and a few mm. For older stars like Vega and Fomalhaut total dust masses are 10 times less.

The future

Until the past few years study of circumstellar disks from the ground, with but few exceptions, has not been very rewarding. Analysis of IRAS data obtained over 15 years ago has dominated the field. Recently, sensitive new array detectors at submillimeter wavelengths and in the mid-infrared (10–20 μm) have finally enabled a suite of important ground-based discoveries. At about the same time, the NICMOS near-IR coronagraphic camera on the Hubble Space Telescope obtained stunning images of CS disks around a few young stars.

The coming decade promises great advances in ground-based, airborne, and space-based studies of CS disks. On the ground, millimeter interferometer arrays at high, dry sites in Hawaii and Chile will produce unprecedently detailed pictures of the structure of CS disks. The SOFIA airborne observatory will measure the spatial extent and spectrum of dust emission throughout the far-IR where the CS disks around most Vega-like stars emit most intensely. And from space, the SPACE INFRARED TELESCOPE FACILITY (SIRTF) will enable detection of many new Vega-like systems that emitted too faintly for detection by IRAS. Many such systems may be discovered around stars of solar mass or less that were mostly out of the reach of IRAS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Artymowicz P 1997 Beta Pictoris: An early solar system? *Ann. Rev. Earth Planet. Sci.* **25** 175–219

Hildebrand R 1983 The determination of cloud masses and dust characteristics from submillimeter thermal emission *Q. J. R. Astron. Soc.* **24** 267–82

Lagrange A-M, Backman D and Artymowicz P 2000 Planetary material around main sequence stars *Protostars & Planets IV* ed V Mannings, A P Boss and S S Russell (Tucson: University of Arizona Press)

Pollack J B *et al* 1994 Composition and radiative properties of grains in molecular clouds and accretion disks *Astrophys. J.* **421** 615–39

Zuckerman B and Becklin E E 1993 Submillimeter studies of main-sequence stars *Astrophys. J.* **414** 793–802

Ben Zuckerman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dutch Astronomy

The Netherlands has a strong international position in astronomical research, particularly in the fields of radioastronomy, the interstellar medium, structure and dynamics of the Milky Way and other galaxies, stellar and solar physics, and, more recently, high-energy astrophysics. The prominent position is strongly linked to the quality of the research education, to which traditionally much care is devoted. Many graduate students gain postdoctoral positions at premier institutions abroad, or obtain good research positions outside astronomy. As early as the 1930s many astronomers educated in the Netherlands were called to leading positions abroad. For example, between 1970 and 1993 the position of Director General of the European Southern Observatory (ESO) was filled by Dutchmen, Blaauw (1970–5), Woltjer (1976–87) and van der Laan (1988–92) respectively.

Organizational structure

External factors at both international and national levels have important influences on astronomical research in the Netherlands. Intergovernmental organizations like the EUROPEAN SPACE AGENCY (ESA) and the EUROPEAN SOUTHERN OBSERVATORY (ESO) determine, in broad outline, the opportunities for research in the Netherlands, especially in the long-term.

At a national level, the following three organizations determine the research strategy.

The Netherlands Research School for Astronomy (NOVA)

Traditionally, most of the research in the Netherlands is carried out at the astronomical institutes at the universities of Amsterdam, Groningen, Leiden and Utrecht. These institutes are the Astronomical Institute Anton Pannekoek (Amsterdam), the KAPTEYN ASTRONOMICAL INSTITUTE (Groningen), the Astronomical Institute Utrecht and LEIDEN OBSERVATORY. The institutes in Leiden and Utrecht have long historic traditions, having been founded around the middle of the seventeenth century. In 1992, the university institutes federated in the Netherlands Research School for Astronomy (Nederlandse Onderzoekschool Voor Astronomie, NOVA). NOVA's mission is to carry out frontline astronomical research in the Netherlands, and to train young astronomers at the highest international level. Since 1992, all graduate astronomy education in the Netherlands has been concentrated in NOVA. In 1998 the Dutch government initiated a research program, the so-called 'In Depth Strategy', to identify and stimulate national focus points of excellent scientific research. To meet the selection criteria, such focus points were compared with the best foreign institutes in their field of research. In addition they had to have the potential to develop even further and become centers of excellence of world-class reputation. The training of outstanding young scientists was also one of the main objectives of the program. Out of the 34 applications covering all academic disciplines, six national research schools were selected by the Netherlands Organization for Scientific Research (NWO). NOVA was highest ranked among the selected research schools and received a significant grant for the period 1999–2005 to set up an innovative research program (see later for more information). Discoveries with large international telescopes are often made by groups, which have built a novel auxiliary instrument, such as a special purpose spectrograph of a new detector system. In this respect, NOVA initiated Dutch involvement in several instrumentation programs for the ESO Very Large Telescope (VLT) and the VLT Interferometer. For more information the reader should visit the NOVA web site http://www.strw.LeidenUniv.nl/nova.

The Netherlands Foundation for Research in Astronomy (NFRA)

The NFRA was founded in 1949 by Professor JAN HENDRIK OORT to detect radio emission from the Sun and the Milky Way. It was set up as an inter-university foundation under the umbrella of the Netherlands Organization for the Advancement of Pure Research (ZWO, now called NWO). It developed and operated the radio telescopes at Kootwijk in the early 1950s, the 25 m DWINGELOO OBSERVATORY Telescope, which came into operation in 1956, and the WESTERBORK SYNTHESIS RADIO TELESCOPE (WSRT), which was opened in 1970. The Kootwijk and Dwingeloo telescopes produced the first radio maps of the Milky Way in the radiation of the 21 cm line of atomic hydrogen. In the 1970s, the WSRT (at that time consisting of 12 25 m antennas on a baseline of 1600 m, though now 14 antennas on a 3 km baseline) was the largest radio telescope in the world and produced many radio maps of external galaxies to study their kinematical and dynamical behavior.

In the 1980s, NFRA significantly broadened its activities by joining the United Kingdom in developing and operating the ISAAC NEWTON GROUP (ING) OF TELESCOPES on the Canary Island of La Palma, which became operational in 1985. At the same time, NFRA also joined the JAMES CLERK MAXWELL TELESCOPE (JCMT), a 15 m radio telescope in Hawaii dedicated to the observation of electromagnetic radiation in the (sub)millimeter wavelength range. The JCMT collaboration also involved the United Kingdom and Canada. The Netherlands has a 20% share in the ING, as well as in the JCMT. The present arrangements are that the Netherlands will continue to participate in the ING and JCMT until 2009. For the optical instrumentation program, on a national level, the NFRA collaborated with the Kapteyn Sterrenwacht Werkgroep (KSW) in Roden. The KSW was a joint effort of the universities of Groningen and Leiden to focus their technical resources to achieve their contributions to develop and construct instrumentation for the ING. In 1996, the KSW merged with NFRA and moved to Dwingeloo to start an optical/infrared instrumentation group. Their first major project is a collaboration with a French group to build VISIR, the VLT Imager and Spectrometer for the mid-infrared. This instrument will be delivered at the ESO VLT in 2001.

Early in the 1990s, NFRA refocused its activities on the WSRT to upgrade this facility with state-of-the-art multifrequency front-end receivers to improve the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sensitivity and the flexibility. The upgrade also included a new backend correlator to significantly increase the spectral bandwidth, a pulsar backend (PUMA, largely developed at Utrecht University), new observing modes for VLBI, and filters and software tools to reduce the interference by human communication (TV, radio and mobile phones). The upgraded WSRT was completed by the start of the 21st century. It will allow astronomers to make extensive studies of the formation and evolution of galaxies in the early universe by observing the redshifted 21 cm line of atomic hydrogen in the range of redshifts up to $z = 4$ (or equivalently, back in time when the universe was only 10% of its present age).

NFRA has also been an active partner in the European VLBI Network (EVN). The EVN is a consortium of radioastronomy institutes in 12 European countries and China to jointly carry out radio observations at very high angular resolution (up to milliarcseconds) using very long baseline interferometry (VLBI) techniques. The EVN initiated and supervised the establishment of the JOINT INSTITUTE FOR VLBI IN EUROPE (JIVE). JIVE was set up in 1993 to provide an advanced data processing facility for VLBI in Europe and to support EVN operations. NFRA is the host institute for JIVE. The astronomical operations of JIVE began in the fall of 1999.

In 1995 NFRA started an initiative to seek for world-wide support to develop a next generation radio telescope with a detecting area of the order of 1 km^2 to study the evolution of the early universe and the epoch of the start of galaxy formation. This program is called the Square Kilometre Array (SKA). In order to make such a project financially feasible, one has to remove all moving parts from the telescope. Therefore, NFRA has chosen the technology of phased arrays to achieve flat antennas. All signal processing electronics will be integrated at chip level. This technique will enable the use of software methods for adaptive suppression of external interference. It will also allow for the use of multiple (up to 100 or more) beams on the sky covering the entire visible sky. Current phased array antennas, mainly used in military applications, have narrow spectral bandwidth, while astronomers demand very wide-band systems to study the evolution of galaxies from the early universe to the present epoch. Other countries, like Australia, Canada, China, India and the US, also participate in the technology development for SKA. It is likely that it will take a decade or more to develop the SKA concept before construction will become feasible.

For more information on the activities at NFRA the reader might visit the NFRA website at http://www.nfra.nl.

Space Research Organization of the Netherlands (SRON)

The scientific space research activities in the Netherlands are concentrated in SRON, a foundation within the framework of NWO. Its mission is to initiate, develop, build and use internationally outstanding instruments for scientific research in and from space. The major SRON responsibilities are to develop a national program for scientific space research, to maintain the work program of the SRON institute, which incorporates laboratories in Utrecht and Groningen, and to provide advice with regard to Dutch participation in scientific projects within the framework of ESA.

Many instruments for astrophysical research built by SRON (and its predecessor) in the past, found their place in national satellites, ESA satellites and NASA satellites. Initially, instruments were lifted above the atmosphere by means of sounding rockets and balloons. In 1968 the first Dutch experiment went into space on board an American satellite to measure electrons in cosmic radiation. In the 1970s, a Utrecht experiment did some pioneering work on measuring star spectra in the ultraviolet spectral range. The first completely Dutch satellite was ANS (ASTRONOMISCHE NEDERLANDSE SATELLIET), equipped for photometry in the ultraviolet spectral range and for the detection of x-ray sources. The launch of the European satellite COS-B, designed for a first survey of the Galaxy in gamma radiation, took place in August 1975. Three years later NASA launched the Einstein satellite with a Utrecht x-ray reflection grating on board. It was followed by the Solar Maximum Mission in 1980, with the Utrecht HXIS instrument on board. The second (partial) Dutch satellite, IRAS, and the European EXOSAT (European X-ray Observatory Satellite) were launched in 1983, both containing important scientific contributions from the Netherlands. The data from all these missions have formed a source of pioneering astronomical research for many years.

On 17 November 1995 ESA's Infrared Space Observatory (ISO) was launched. One of the four instruments on board the satellite was built under SRON responsibility: the Short Wavelength Spectrometer (SWS). During a very successful life of 28 months more than 9000 observations were done on stars, planets, galaxies, dust and nebulae. All SWS observations now reside in the Dutch ISO Data Analysis Centre.

SRON provided major contributions to the following astrophysical instruments, which are still operational in space as of early 2000:

- COMPTEL (1991), a gamma-ray detector-spectrometer on board the NASA observatory CGRO;
- the Wide Field Camera for the Italian–Dutch x-ray satellite Beppo-SAX (April 1996);
- the Low Energy Transmission Grating Spectrometer aboard Chandra (formerly called AXAF). Chandra was successfully launched in July 1999;
- the x-ray spectrometer for the ESA satellite XMM, which was successfully launched in December 1999.

Dutch space research specializes in high-energy astrophysics (x-ray research) and low-energy astrophysics (infrared and submillimeter research). For the first decade of the 21st century, SRON has focused its astronomical program on harvesting the scientific data obtained with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

above-mentioned instruments and on the construction of HIFI, an important instrumental contribution to the submillimeter detection system for the ESA 'cornerstone' satellite FIRST/PLANCK. More information is available at the website http://www.sron.nl.

Focus on three research areas in astronomy and astrophysics

World-wide, one can identify a dozen or so main areas of astronomical and astrophysical research which will flourish in the next decade because of major breakthroughs, driven by the next generation of telescopes which cover the entire electromagnetic spectrum, by the jump in resolution provided by interferometry in the infrared and submillimeter regimes, and by the powerful new computational capabilities for data analysis and theoretical modeling. The Netherlands is very strong in three of these areas. Each area addresses deep questions and together they converge on NOVA's main science theme for the coming decade: the life-cycle of stars and galaxies. These three areas are as follows.

Formation and evolution of galaxies from the distant past to the present epoch

Galaxies contain billions of stars, as well as interstellar gas and dust, and are embedded in dark halos of unknown constitution. Astronomers are able to look back in time by observing galaxies at ever-greater distances. Because light travels at finite velocity, distant objects are seen at a time when the universe was younger. The expansion of the universe causes light to be redshifted, so that the most distant galaxies are those with the highest redshift. How did galaxies form? What processes have occurred between high redshift and the present? Do evolved galaxies contain relics, which are clues to their formation? What are the influences of the environment, of nuclear activity, and of the original large-scale distribution of dark matter? What is the role of massive black holes in galactic nuclei?

In the last few years rapid new developments in instrumentation have provided unique opportunities to find answers to these questions. The Hubble Space Telescope and the ESO Very Large Telescope (VLT) have reached their full potential. Optical telescopes with a diameter of eight or more meters have started to produce exciting data. Dutch astronomers have access to the equivalent of 30% of an 8 m telescope through their membership of ESO. Also, the upgraded WSRT and the JIVE VLBI data processor have led to new research opportunities. As a result, the distant and therefore early universe has now come within observing range. Fast progress is now being made in the study of proto-galaxies and in the large-scale structure of the universe at only one-tenth of its present age.

Birth and death of stars: the life cycle of gas and dust

New stars continue to be born deep inside molecular clouds in galaxies. The birth process leads to a circumstellar disk of gas and dust from which planets and comets may subsequently form. What are the physical processes that lead to these new solar systems, and how do they evolve? How is the chemical composition of the gas and dust involving the major biogenic elements modified during the collapse from the cold, tenuous interstellar medium to the dense protoplanetary material? During the late stages of their life, stars eject nucleosynthetically enriched material into the interstellar medium. How does this material drive the chemical evolution of a galaxy and of the newly formed stars therein? What drives the mass loss, and how does it influence stellar evolution?

The availability of radio telescopes that are able to detect electromagnetic radiation at millimeter and submillimeter wavelengths (like the JCMT) and the success of ESA's Infrared Space Observatory (ISO) have recently led to the possibility of penetrating far into the dusty areas around young and old stars. This is leading to new insights into the physical and chemical processes that occur at the birth and death of stars and into the formation of planets. In this extremely interdisciplinary area new laboratory simulations in the Raymond and Beverly Sackler Laboratory for Astrophysics at Leiden University as well as computer models further contribute to our understanding of the life cycle of gas and dust.

Final stages of stellar evolution: physics of neutron stars and black holes

At the end of its life, a massive star explodes and ejects its outer layers. The stellar core collapses to form a neutron star or a black hole. These are the densest objects that exist, and the ones with the strongest gravitational fields. What are the properties of matter at the extreme density in the interior of a neutron star? What are the observational signatures of black holes? Can we observationally verify the extraordinary predictions of general relativity for the properties of curved space-time near these objects? How do particles and radiation behave near these compact objects? What happens when two compact objects orbiting each other eventually merge? Is this the origin of the most powerful explosions we know, the enigmatic gamma-ray bursts?

Rapid progress is being made in this field as a result of the availability of new satellite measurements with, among others, Dutch instruments, for example in GRO and Beppo-Sax. The significance of x-ray research in extragalactic objects is growing fast. A breakthrough made in mid 1997, to which Dutch researchers made an important contribution, is the insight that the well-known 'gamma-ray bursts' originate in very distant galaxies; it is now certain that the phenomenon involves extremely high-energy acceleration processes, presumably related to the formation of compact objects (neutron stars and black holes) or the fusing of two such objects into a double star. Using the new x-ray observatories XMM (ESA) and Chandra (NASA), to which SRON has contributed important instrumentation, the study of high-energy processes will become possible in the surroundings of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

compact objects with unparalleled sensitivity, and spatial and spectral resolution.

There are many connections between the above areas of research. Stars form in galaxies from interstellar material, and at the end of their lives return chemically enriched material to the interstellar medium from which new generations of stars and planets form. The most massive and luminous stars evolve fastest, and leave neutron stars and black holes. This life cycle causes evolution in the stellar population of a galaxy as a whole. Recent technological advances have made it possible to observe this evolution all the way back to epochs when the universe was less than 10% of its present age. Unravelling the history of galaxy formation in the universe, and so understanding the life cycle of stars and galaxies, requires full knowledge of astrophysics and state-of-the-art observations, which are now within reach.

Observing facilities available for Dutch astronomers

Astronomers in the Netherlands have access to the telescopes of the European Southern Observatory (ESO) in Chile, to the UK/NL Isaac Newton Group on La Palma, to the James Clerk Maxwell Telescope (JCMT) on Mauna Kea, to the upgraded Westerbork Synthesis Radio Telescope (WSRT), and to the Joint Institute for VLBI in Europe (JIVE). Together these facilities cover the optical and submillimeter/centimeter wavelength windows. During 2000, ESO's Very Large Telescope, the world's largest optical/infrared telescope, will be completed. ESO plans to combine the light of the four main 8 m telescopes with that of several smaller telescopes to form a near-infrared interferometer (VLTI), which will provide unprecedented spatial resolution. Furthermore, a joint European–US ATACAMA LARGE MILLIMETER ARRAY (ALMA) is planned in Chile, coordinated by ESO for the European side. Initial operations are expected around 2007. The Netherlands will make several major contributions to the VLT and the VLTI. They are: VISIR (see earlier); OmegaCam, a wide field ($1^\circ \times 1^\circ$) optical camera for the 2.5 m VLT Survey Telescope; MIDI, the first instrument on the VLTI to do imaging spectroscopy in the mid-infrared; SINFONI, a near-infrared integral field spectrograph with a large field of view making use of adaptive optics technology. All these projects are collaborations with other institutes elsewhere in Europe, mainly in Germany and France. Visit the NOVA website for further information. The Netherlands is also participating in two technology development studies for ALMA; the NOVA/SRON study for the receiver that will cover the atmospheric window in the 600–720 GHz range, and the NFRA participation in an international study on an advanced correlator. Furthermore, in order to develop knowledge of optical/infrared interferometry and its applications to future astronomy programs, NOVA and ESO initiated a VLTI Expertise Centre (NEVEC) at the Leiden Observatory in 1999.

Dutch astronomers also have access to observing time on a variety of space observatories. The NASA/ESA Hubble Space Telescope, two major new high-energy satellites launched in 1999 (Chandra, XMM), and the Italian–Dutch x-ray satellite BeppoSAX. ESA's submillimeter Cornerstone Observatory FIRST/PLANCK is scheduled for launch in 2007. Together this represents an investment by the world-wide space agencies on the order of US$ 15 billion. Furthermore, the databases obtained with ESA's astrometric satellite Hipparcos and infrared satellite ISO will, at least for the next decade, contribute to exiting new results.

Wilfried Boland

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dwarf Elliptical Galaxies

DWARF SPHEROIDAL GALAXIES were first identified by Shapley, who had noticed two very diffuse collections of stars on Harvard patrol plates. Although these systems had about as many stars as a GLOBULAR CLUSTER, they were of much lower density, and hence much larger radius, and thus were considered distinct galaxies. These two, named Fornax and Sculptor after the constellations in which they appear, are in fact satellites to the MILKY WAY and are part of a family currently numbering nine that orbit the Galaxy. Studies by Baade and others showed that the stellar populations are old, based on the presence of RR LYRAE STARS and the apparent lack of star formation regions, prompting a question of link between these galaxies and the ELLIPTICAL GALAXIES. The galactic dwarf spheroidals are themselves part of a larger class of galaxy called dwarf ellipticals (dE), which extend in luminosity from that of the faintest dwarf spheroidals ($M_V = -9$) to around $M_V = -17$, a good example being NGC 205. These galaxies are referred to as 'spheroidals' by some authors. The basic isolating characteristics of the class are low surface brightness and smooth light distribution. The radial light distributions tend to be exponential, which they have in common with IRREGULAR and SPIRAL GALAXIES and which distinguishes them from the ellipticals, which have more of a power-law light profile. Also, the surface brightness (either central or average value) decreases with decreasing luminosity, whereas for ellipticals surface brightness increases with decreasing luminosity. There is a range of luminosities for which both dwarf ellipticals and ellipticals exist—the difference in the types is then best demonstrated by the remarkable differences in structure between that of NGC 147, a dwarf elliptical, and M32, an elliptical, both of which are companions to M31 and which have similar luminosities (figure 1).

dEs are very common. In the LOCAL GROUP, there are 19 known (nine around M31, one unattached to a large galaxy, and the nine of the Milky Way already mentioned). The nearby group of M81 has 12 known. The largest collections of dEs, however, are found in the rich clusters. Virgo, for instance, has had over 400 dEs cataloged to the relatively bright luminosity of $M_V = -13$, although the dwarfs as a population contribute less than 10% of the total light of the cluster. There is some indication that the relative number of dEs to giant galaxies in a group goes up as the number of giants increases. The luminosity function of dEs was shown by Sandage, Binggeli and Tammann to have an exponential form, thus entailing the exponential part of the Schechter function which represents the luminosity function of all galaxies. The slope of this function in CLUSTERS OF GALAXIES shows significant variations among clusters studied, with the interesting result that clusters with large cD galaxies have proportionately fewer dE galaxies (slope $\alpha \sim -1$) than do those clusters without such a dominant galaxy ($-1.4 \geq \alpha \geq -2.0$). The conclusion here may be that cD halos are formed by the disruption of dwarfs, although the mechanism for doing so is unidentified, and it is as likely that the problem needing explaining is actually the preponderance of dEs in non-cD clusters. The three dimensional shapes of dEs appear not to be disks, despite their exponential light profiles; rather, the flattening distributions indicate spheroidal or ellipsoidal shapes because of the lack of very flat systems. The lack of nearly circular galaxies argues against the galaxies being oblate, and although prolate systems are permitted by the data, it is most likely that, as a group, these galaxies are triaxial.

Because of the intrinsic faintness of these galaxies, there is not a surfeit of kinematic information about them, although what is known has proved to be crucial to the study of DARK MATTER IN GALAXIES. Studies of the velocities of individual stars in the galactic dwarf spheroidals have shown that the least luminous galaxies (UMi and Draco) have velocity dispersions around 10 km s^{-1} similar to the dispersions found in the brightest galactic dwarf spheroidal, Fornax, which is nearly 100 times more luminous. The high dispersions result most simply in total M/L values of order 100 for the faint dwarfs, and since the stellar M/L values are order unity, there is thus strong evidence for unseen matter. Modeling attempts to explain the large velocity dispersions by interactions and disruptions with the Milky Way have largely been unsuccessful, because these interactions, while disturbing the structure of dwarfs galaxies, do not significantly alter the central velocity dispersions. More detailed structural models of the galaxies, in which dark matter halos are included along with the observed luminous matter distribution, have confirmed the large M/L ratios and also shown that the central densities of dark matter can be very high, near 1 $M_\odot$ pc^{-3}, much larger than the density of visible stars.

Spectral observations in integrated light of more distant galaxies have further shown that the dEs are slow rotators and are in general not rotationally flattened, thus placing them on common ground with the large ellipticals, and making them different from the low-luminosity ellipticals, which are rotationally flattened. The galactic dwarf spheroidal UMi, which is quite flat in projection ($e = 0.6$), has also been shown to have little rotation, the data coming from individual constituent stars in this case. The velocity dispersions obtained in integrated light of luminous dwarf ellipticals tend to indicate normal M/L ratios; i.e. the central areas are not dark matter dominated.

A salient feature in nearly half of the luminous dEs is the presence of a distinct nucleus. The most widely known nucleus is surely that of NGC 205, although this nucleus is in fact quite different from those of most other galaxies in that it is a young star cluster, part of the star formation event first noticed by Baade. The nuclei in other dEs appear to be the same age as the surrounding galaxy. For a small nucleus, it can appear simply as a star cluster, occasionally miscentered in the host galaxy; in the cases where the nucleus is large, it tends to blend in with the structure of the surrounding galaxy rather

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

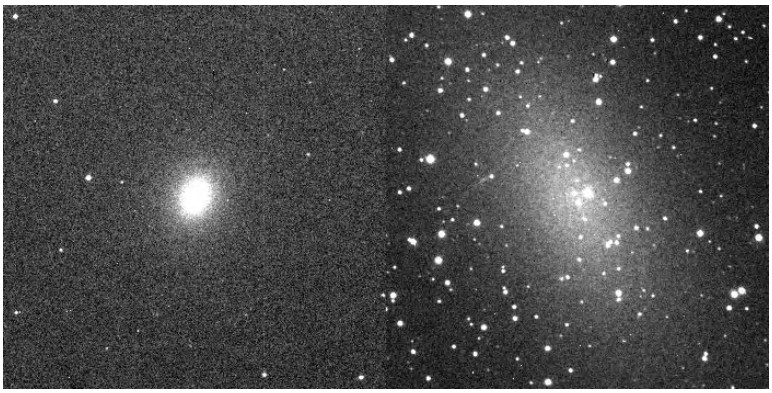


Figure 1. A comparison of an elliptical and a dwarf elliptical galaxy with similar luminosities but very different structures, both companions to M32. Left: M31. Right: NGC 147, a dwarf elliptical, shown at the same scale.

than appear as an imposed star cluster. Their luminosities range upwards to $M_V = -12$, which is somewhat brighter than the brightest globular clusters known anywhere. The presence of a nucleus is weakly correlated with a number of other properties. They occur more frequently in more luminous galaxies. They occur more often in rounder galaxies. In clusters, the nucleated galaxies tend to be more concentrated as a group than do non-nucleated galaxies. Also, the nucleated galaxies tend to be redder than non-nucleated galaxies, although there is little color difference between the nucleus and the surrounding stars. None of these tendencies yet points to an origin of the nuclei, however, other than to support the obvious idea that the nuclei are secondary star formation events, where gaseous material fell to the center of a newly formed galaxy and built up a region of high stellar density. HST images of a nearby nucleus (which contains nearly 5% of the total galaxy luminosity) show a central density as high as that found in galactic globular clusters (figure 2).

Like all types of galaxies, dEs contain globular clusters. The presence of them is directly related to the host galaxy luminosity, and the scaling factor appears to be more like the ellipticals with a relatively high number of clusters per unit luminosity (4 per unit $M_V = -15$), rather than like the spirals which are cluster poor. The faintest known galaxy with globular clusters is still Fornax, which has 5. Colors and metal abundance measures of dE globulars indicate that they are more metal poor than the surrounding halo in the mean; a similar situation is found in ellipticals.

Ideas that the stellar populations are similar to those of globular clusters (i.e. uniformly old) have had to be modified, first because of the indirect indicators of young age such as the presence of anomalous Cepheid variable stars, the second parameter problem (a red horizontal branch in a metal-poor system) and upper asymptotic giant branch stars, and later more directly by color–magnitude diagrams (particularly of Carina), which show multiple main sequence turnoffs, indicating several discrete epochs of STAR FORMATION. Many of the Local Group

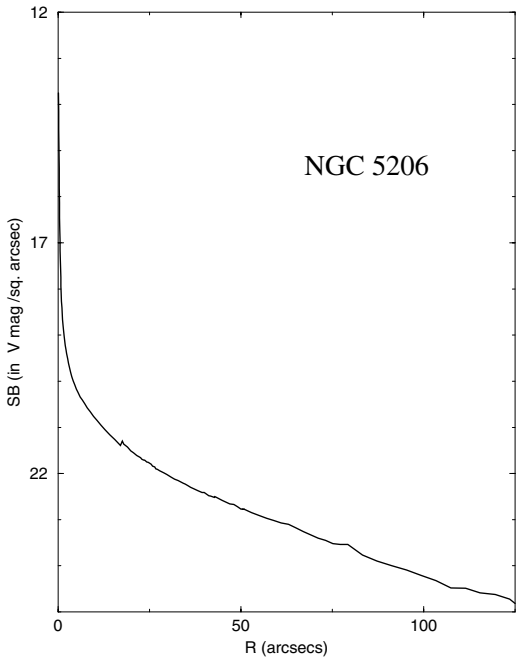


Figure 2. Radial light profile of NGC 5206

dwarf spheroidals show such evidence for star formation more recent than 10 Gyr. Direct evidence for current star formation among Local Group dEs is found only in NGC 185 and 205, however. More distant galaxies also show some signs of recent star formation, in that the integrated colors show large variations among galaxies of similar luminosity; some colors are bluer than can be explained by low metallicities. It is apparent that star formation in dEs is a more complex thing than in globular clusters. The acquisition of more star formation material can explain some cases, but others require self-regulation of star formation, via stellar winds perhaps. Galaxies such as the Phoenix dwarf and LGS3 are thought to be transition objects, in which star formation is about to end,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

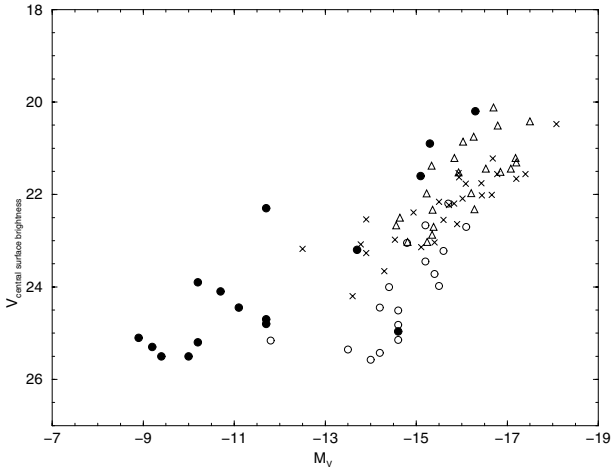


Figure 3. Relation between luminosity and surface brightness for dwarf ellipticals. Open symbols are cluster galaxies; full symbols are Local Group and M81 galaxies.

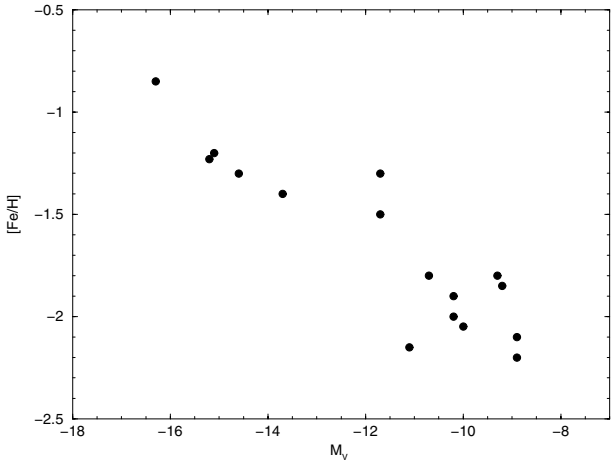


Figure 4. Relation between luminosity and metallicity for dwarf ellipticals.

after which time the galaxy would take on the appearance of a dE. Studies of all the dwarfs in the Local Group (including the irregulars) show that dwarf elliptical star formation histories merge smoothly into those of DWARF IRREGULAR GALAXIES. A useful summary might be that stars began forming in dwarf ellipticals 15 Gyr ago, but that star formation may have ceased only a few Gyr ago.

dEs as a class have a number of remarkable relations between observed properties. As mentioned above, the central or mean surface brightnesses are directly related to the total luminosity, over a range of about 8 magnitudes (figure 3). The central values used for this relation are extrapolated values, so that any nuclei present are ignored. This relation is not tight, however, and some have even doubted that it exists, noting that surface brightness selection effects in galaxy searches would tend to produce such a relation. The central surface brightnesses of these galaxies are of course well below that of the night sky. However, more recent deep searches, while finding the very interesting large (and hence luminous) low-surface-brightness galaxies, have not found them in the abundance that would fill the surface brightness–luminosity plane uniformly. Thus, it seems likely that while large deviations from a relation do exist, there is a preferred or dominant relation. (One has to be careful to distinguish large, low-surface-brightness dwarf elliptical galaxies, with M_V fainter than -18, from the large, low-surface brightness disk galaxies, such as Malin I.) A second relation is that of mean stellar metallicity (Z) and total luminosity (L) (figure 4). The metal abundances used in this relation are those found from color–magnitude relations, and thus the galaxies measured so far are all at least within the reach of HST. Over a range of 8 magnitudes, the metallicity changes by 1.4 dex (a factor of 25, $Z \propto L^{0.4}$), and the relation is fairly tight. A question then arises: with two relations relating three parameters, L–Z, and L–SB, one tight and one not so tight, is the relation between SB and Z tight or not? That is, is the true parameter that dictates mean metallicity during galaxy formation the galaxy luminosity (total mass), or rather the surface brightness (surface density)? Observations of one large low-surface-brightness galaxy in the M81 group provide the answer here. This galaxy lies well off of the main SB–L relation, and also lies well off of the main SB–Z relation. However, it does lie on the L–Z relation formed by Local Group galaxies. Thus it is luminosity that dictates the metallicity. It is interesting that in the faintest dwarfs, while the luminous matter is a small fraction of the total mass, it is the luminous matter that dictates the metallicity.

The existence of the metallicity–luminosity relation and the surface brightness–luminosity relation as well as the complex star formation histories have provided an impetus for searches for theories of how these galaxies formed. Formation ideas are varied, from the detritus models wherein the small companion dE galaxies are the leftover material of large galaxy formation, to formation in tidal tails of interacting galaxies, to that of Babul and Rees which is a more *ab initio* scheme. Babul and Rees have shown that the collapse of weak density perturbations can be delayed because of the photoionizing effects of the intergalactic UV radiation field at redshifts greater than 1. Thus star formation in dwarf galaxies may have begun relatively late and may have resulted in that class of galaxies being excessively bright at intermediate redshifts, perhaps leading to the excess number of blue galaxies apparent at those redshifts. Larson showed that the more massive dwarf precursors were better able to retain gas (which otherwise would have been lost via supernova explosions) and hence had longer-lived star formation epochs, resulting in higher central densities and higher metallicities. Since gas loss in self-gravitating systems cannot explain both the observed form of the decline in metallicity and surface brightness with luminosity in dEs, Dekel and Silk developed a model which included the presence of dark matter, wherein mass loss from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

supernova winds becomes important at a threshold value of mass and allows both stated conditions to be met. Being derived from hierarchical scenarios, the dwarfs should also predominate in areas of low density (i.e. voids), something for which there is much counterevidence for.

There are obvious connections between dEs and star-forming dwarfs (dIrrs)—low surface brightness, low metallicity, low luminosity and even similar flattening distributions. However, in detail there are problems with the idea that dEs are simply the extension of dIrrs—those galaxies that ran out of star forming gas first. As a class, dwarf ellipticals are smaller than dIrrs—there are many more large dIrrs than dEs. Thus it is hard to obtain all of the dEs by stopping the star formation and allowing the dIrrs to fade—the length scales would be wrong. (It would be possible to obtain some fraction of the dEs this way of course.) It has been difficult to compare the metallicities of dEs ([Fe/H] in general, from giant branch locations) with those of dIrrs ([O/H] in general, from ionized gas measures) to check the idea that at the same luminosity, corrected for fading for dIrrs, the two types have similar metallicities. Preliminary work does show that it is marginally possible for the observed L–Z relation of dEs to have come from that of dIrrs, but the amount of luminosity fading required is excessive. It is certainly true that dEs and dIrrs come from related progenitors, although the precise parameters separating those families is not yet known.

Bibliography

Babul A and Rees M 1992 *Mon. Not. R. Astron. Soc.* **255** 346
Dekel A and Silk J 1986 *Astrophys. J.* **303** 39
Ferguson H and Binggeli B 1994 *Astron. Astrophys. Rev.* **6** 67
Larson R 1974 *Mon. Not. R. Astron. Soc.* **169** 229
Mateo M 1998 *Ann. Rev. Astron. Astrophys.* **36** 435

Nelson Caldwell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dwarf Galaxy

A small galaxy with a very low mass and luminosity. Dwarf galaxies may be elliptical or irregular in shape but, in either case, are usually sparsely populated with stars, have low surface brightnesses, low overall luminosities, and consequently are hard to detect. It is usually possible to look straight through a dwarf galaxy. Dwarf ellipticals are populated by old stars and contain little, if any, gas, whereas dwarf irregulars contain younger stars and substantial quantities of gas. Dwarf elliptical galaxies, which typically contain from a few hundred thousand to a few million stars, have diameters of a few thousand light-years and luminosities ranging from about 10^5 to about 10^7 solar luminosities, are probably the most abundant species of galaxy in the universe. They are denoted by 'dE' in the Hubble classification scheme. Dwarf irregulars, which appear to be less abundant than dwarf ellipticals, contain from about 10^6 to 10^8 (or, exceptionally, up to about 10^9) stars. The smallest galaxy known is GR-8, a dwarf irregular with a diameter of 1000 light-years.

See also: elliptical galaxies, galaxy, Hubble classification, irregular galaxy, Local Group, Sagittarius Dwarf Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dwarf Irregular Galaxies

Dwarf irregular galaxies at optical wavelengths are small, faint and appear to be unstructured and irregular in shape. They are typically gas-rich, metal-poor systems, with varying levels of STAR FORMATION occurring in a haphazard manner across the galaxy. Their structure does not normally include any distinctive features, such as spiral arms or nuclei, and their kinematics seems to be dominated by random motions or relatively slow solid-body rotation. There is not an extremely precise definition of what is a DWARF GALAXY and what is not, but the most typical demarcation between a dwarf galaxy and a larger ('normal') galaxy is that dwarfs are typically fainter than an absolute B-band magnitude $M_B = -16$. This is not a strict rule, as the Small Magellanic Cloud (with $M_B = -17$) is often considered to belong to the class of dwarf irregular galaxies. The lumpy irregularities in the optical morphology of dwarf irregular galaxies are due to randomly distributed star formation regions, which may be individually quite bright, but the overall surface brightness is typically extremely low ($\Sigma_0 \sim 23$–25 mag arcsec^{-2}). Because this type of galaxy is so numerous and they can easily vary in brightness quite dramatically owing to increased star formation rates, there is strong evidence to suggest that they are cosmologically important population of galaxies.

There are several classes of dwarf galaxy, not only dwarf irregular galaxies, and the difference between dwarf irregular galaxies and other classes of dwarf galaxy, such as 'blue compact dwarf' galaxies, 'dwarf spheroidal' galaxies and 'dwarf elliptical' galaxies, is often a question of semantics and has not always been precisely defined. These types of dwarf galaxy on average fall into the same low surface brightness class, but differ in their current rates of star formation. Blue compact dwarf galaxies have one or more compact, high star formation rate regions (e.g. NGC 1705, NGC 1569, VII Zw403). DWARF SPHEROIDAL GALAXIES, on the other hand, currently have no ongoing star formation, and no detectable gas, but they can show evidence of having had a complex star formation history, including one, or more, distinct episodes of star formation (e.g. Carina, Leo I, Sculptor). There are also a small number of what appear to be transition objects, which have extremely low rates of current star formation, but with evidence of more active star formation in the recent past (e.g. Pegasus, Phoenix, Antlia). The existence of these transition objects leads to the hypothesis that dwarf spheroidal, irregular and blue compact galaxies are all the same type of object, observed at different evolutionary phases. DWARF ELLIPTICAL GALAXIES (e.g. M32, NGC 147, NGC 185) usually have evidence only for a fairly old stellar population, and it is not entirely clear that they can fall into the same global scheme as the other types of dwarf galaxy suggested above.

Dwarf irregular galaxies are the most common type of galaxy in the region surrounding our own Galaxy, the so called LOCAL GROUP. In a recent census, 22 out of 41 galaxies within the Local Group can be classified as dwarf irregular type galaxies. Similar results, showing the dominance of dwarf irregular populations, can also be seen for other nearby groups of galaxies (or often called CLUSTERS OF GALAXIES), such as Virgo and Fornax. There are also indications from redshift surveys that the universe as a whole is dominated by small irregular type galaxies. It is of historical interest to note that the first object to be '... definitely assigned to the region outside the Galactic system ...' (i.e. to be extragalactic) by Edwin Hubble in 1925 was NGC 6822, one of the nearest examples of a dwarf irregular galaxy. Pioneering work by Walter BAADE, starting in 1944, taking the first steps in understanding galaxy evolution, was based on studies of the resolved stellar populations in dwarf galaxies. Baade noted that the dwarf elliptical galaxies he looked at contained only red stars, namely old stars (which he called population II), and dwarf irregular galaxies contain red and blue stars, which indicate a younger population (which he called population I). He further noted that while an exclusively old population has been observed, a young population was never seen without an old underlying, fainter, population of red stars, which is called 'Baade's Sheet'.

Because of their small size and low surface brightness dwarf irregular galaxies can be very hard to detect, even in the Local Group, and to the present day we are still finding new dwarf irregular galaxies (e.g. Antlia in 1997, by a group from Cambridge, England). The first major survey of these faint and often unregarded smudges on photographic plates was undertaken by the Swedish astronomer Holmberg, in 1950 in the M81 and M101 galaxy groups. A further extensive specific search for field dwarf galaxies was undertaken by van den Bergh in 1959 in Canada (the DDO catalogue). Of the present-day dwarf irregular members of the Local Group 50% have been found since 1971, and we cannot rule out that we will find many more extremely low surface brightness objects in the Local Group.

The distribution of neutral hydrogen (H I) gas in dwarf irregular galaxies is very clumpy and irregularly distributed and is frequently much more extended than the optical emission. There have been very few cases in which molecular gas has been detected. Molecular gas, typically, means carbon monoxide, CO, which is relatively straightforward to detect and is considered to be a tracer for the presence of molecular hydrogen, H_2, which is the most important constituent in the star-formation process. Given the importance of molecular gas to the star-formation process it must exist in these galaxies, as we see young stars. It is possible that the extremely low metallicity of dwarf irregular galaxies makes the molecular gas more difficult to detect, because the relative fraction of CO to H_2 molecules is much smaller than in our Galaxy.

Dwarf irregular galaxies are also so-called 'metal-poor' galaxies. This means that the abundance of 'metals', namely elements with atomic numbers greater than 2, relative to the abundance of hydrogen is much lower

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

than is observed in our Sun (the solar abundance). The metallicity of dwarf irregular galaxies ranges between 2.3% of the solar value (e.g. Leo A) and around 10% of the solar value (e.g. IC 1613). The typical elements observed to determine the metallicity of nearby galaxies are carbon, oxygen and nitrogen, which are detectable in optical and UV emission lines in currently star-forming regions (so-called, H II regions), and iron has been measured in spectra of individual stars of various ages. The extremely low abundances of metals relative to hydrogen, typically in the range 2–15% of solar neighborhood values, can be taken as a sign either of youth or of a sporadic star formation history, or a combination of both.

The star formation history, or luminosity and color evolution, of dwarf irregular galaxies shows evidence of not having been a smooth function of time. That is to say, they have been subject to sporadic increases in the average rate of star formation, so-called 'bursts' of star formation. These bursts may make a dwarf irregular galaxy look much more like a blue compact dwarf galaxy. Evidence for these bursts of star formation comes both from the direct observations of the resolved stars (the stellar population) in nearby dwarf irregular galaxies, a census of which gives information about the star formation history that has created the currently seen age and metallicity distribution, and also from studies of the luminosity and colors of much more distant, late-type galaxies seen at intermediate redshift (so-called faint blue galaxies).

It is possible to derive very accurate star formation histories from resolved stellar populations going back to the oldest star formation episodes in a galaxy. Low-mass stars can live a very long time, and thus leave a longstanding fossil record which can be directly interpreted as the star formation history of a galaxy through the entire history of the universe. Current evidence shows that every dwarf galaxy in the Local Group has had a different star formation history from every other. Dwarf irregular galaxies typically appear to contain old populations, as seen from the red giant branch stars, but this has only been unequivocally proven from the detection of RR LYRAE STARS in one or two cases (e.g. IC1613). Dwarf irregular galaxies thus not only have irregular morphology but also have had irregular star formation histories, dominated by short 'bursts' of star formation, rather than a more or less constant rate of star formation with time (as for spirals) or an epoch of formation a long time ago (as for ellipticals). This is consistent with what we know about blue compacts and dwarf spheroidals as well.

Nearby dwarf irregulars are currently forming stars at an extremely low rate, and although this low rate may be more or less a constant underlying rate, through time every so often this rate does increase. When this happens it is possible that a dwarf irregular galaxy would then be classified as blue compact dwarf galaxy, and if this rate ever goes to zero they would be dwarf spheroidals.

The surveys to study the properties of the very distant, so-called 'faint blue galaxies' seen in all deep images of the sky have come to the conclusion that the majority of these objects are at intermediate redshifts and are late-type STARBURST GALAXIES. An analysis of the number of faint blue galaxies in deep surveys, combined with their luminosity and color variations, gives an indication of how short and how intense these bursts of star formation must be for us to detect the number of faint blue galaxies that we do. This type of analysis suggests that they are quite small and undergo extremely short extremely intense bursts of star formation. Detailed imaging of these objects from the Hubble Space Telescope shows a large majority to be irregular in shape. Thus what is known about these very distant, very faint but very numerous faint blue galaxies is perfectly consistent with what we can assume about the present-day characteristics of the dwarf irregular galaxies we see and can study in great detail in the nearby universe and Local Group. Dwarf irregular galaxies are thus a very important link between the nearby universe, which we can study relatively easily, and the most distant objects in the universe, GALAXIES AT HIGH-REDSHIFT, which we can never hope to be able to study in the same detail as Local Group galaxies.

Bibliography

Baade W 1968 *Evolution of Stars and Galaxies* (Cambridge, MA: Harvard University Press)

Da Costa G S 1998 Dwarf galaxies *Stellar Astrophysics for the Local Group* (Cambridge: Cambridge University Press) pp 351–406

Hodge P 1986 *Galaxies* (Cambridge, MA: Harvard University Press)

Mateo M 1998 Dwarf galaxies in the Local Group *Ann. Rev. Astron. Astrophys.* **36** 435

Meylan G and Prugniel P 1994 *Dwarf Galaxies* (Garching: European Southern Observatory)

Eline Tolstoy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dwarf Spheroidal Galaxies

Our galaxy, the MILKY WAY, is surrounded by a swarm of dwarf galaxies each composed of 100 000 to billions of stars. Many of these galaxies are easy to see; for example, the MAGELLANIC CLOUDS are visible to the naked eye for observers in the Southern Hemisphere. Many of the dwarfs are far more difficult to spot; some of these were discovered telescopically in the nineteenth century. One class of nearby dwarf galaxy eluded discovery for even longer. In the late 1930s, Harlow Shapley and Walter Baade reported the discovery of a new type of 'star system' in the constellation Sculptor (figure 1). Barely visible on their original plates, this new object was the first example of a dwarf spheroidal (dSph) galaxy ever found. It is now known to be one of nine such galaxies orbiting the Milky Way.

By convention, the nearby dSph galaxies are usually named after the constellation in which they are located (see table 1). The nine dSph galaxies closest to the Milky Way are all very likely gravitationally bound to our Galaxy. The remaining dSph galaxies listed in the table reside within the LOCAL GROUP of galaxies, and all but one are satellites swarming around the nearby the ANDROMEDA GALAXY. In recent years dSph galaxies have been discovered in large numbers within other nearby GROUPS OF GALAXIES and GALAXY CLUSTERS. When these numbers of known dSph galaxies are extrapolated to the universe as a whole, it becomes apparent that dSph galaxies are by far the single most common type of galaxy by number.

If dSph galaxies are so common, how did they elude discovery for so long? Even in the darkest, most remote sites on Earth, the sky glows faintly with optical radiation. This light comes mostly from atomic and molecular emission from the upper parts of our own atmosphere, from sunlight reflected off dust scattered throughout the solar system and within our atmosphere, and from individually invisible, but very numerous faint galaxies and stars. Most galaxies have surface brightnesses comparable to that of the night sky; that is, the total light they emit over the total area of the galaxy as seen in the sky is comparable to the amount of light the sky itself emits. These high surface brightness galaxies—SPIRAL and ELLIPTICAL GALAXIES—are the ones most commonly pictured in popular astronomy books. They represent what most people think about when they imagine a 'typical' galaxy. In contrast, dSph galaxies have far lower surface brightnesses, glowing faintly with a surface brightness as low as only 1% that of the night sky. These galaxies are the closest examples of true LOW SURFACE BRIGHTNESS systems.

Finding such ghostly objects would be similar to spotting a 40 W lightbulb in front of a large searchlight or discerning a 1 m rise in a 100 m tall plateau from far away. Sculptor was found in the 1930s only because it is one of the brightest nearby dSph galaxies, it is located in a particularly dark part of the sky, and many photographic plates were available by that time to confirm its discovery. Many of the most recently discovered dSph galaxies were not found by eye, but from sensitive electronic measurements of photographic images of large regions of the sky. Because they are so difficult to spot, it seems certain that many more nearby dSph galaxies remain to be discovered.

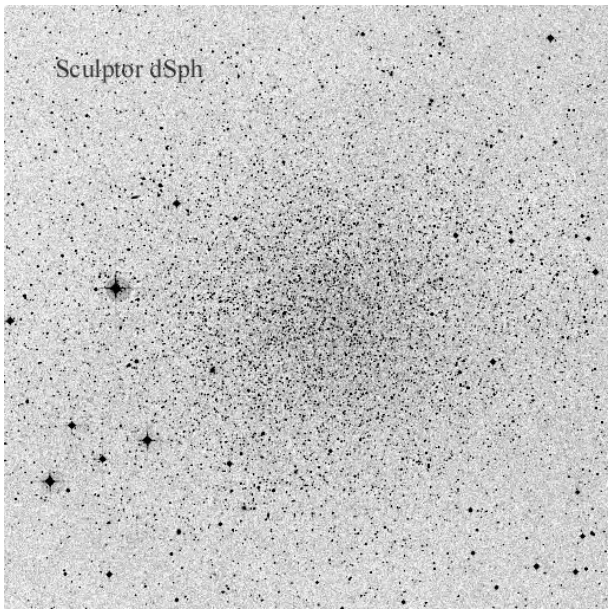


Figure 1. An negative image of the Sculptor dSph galaxy, a nearby satellite of our home Galaxy, the Milky Way. Sculptor is located about 220 000 light-years away and the galaxy is clearly resolved into individual faint stars as seen in the picture. The bright stars in the image are most likely located in our Galaxy somewhere between us and Sculptor. Some of the fainter objects are background galaxies far behind Sculptor. For comparison, the full Moon would easily fit within the field of view shown in the figure. This picture is used by permission from the digitized Southern Sky Survey from Space Telescope Science Institute; the original plate was obtained at the Anglo-Australian Observatory with the UK Schmidt Telescope.

Global properties of dwarf spheroidal galaxies

As their name implies, dSph galaxies are among the smallest, least luminous galaxies known. Ursa Minor, for example, shines with a total luminosity of about 300 000 times the luminosity of the Sun. By comparison, the Magellanic Clouds—dwarf companions of our Milky Way—emit as much light as 2 billion Suns. Even some of the Milky Way's GLOBULAR CLUSTERS emit more light than the faintest dSph galaxies. Most dSph systems show a centrally concentrated structure with a core region of nearly constant, but extremely low, surface brightness. The outer parts of dSph galaxies slowly fade into the night sky; in many cases, no well-defined outer boundary can be measured. The central cores range from 1000–3000 LIGHT-YEARS in diameter, while the ill-defined outer boundaries typically extend 3–10 times further out; the dimensions of the Local Group dSph galaxies (see table 1) are representative. The ultra-low core surface brightnesses

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Dwarf spheroidal galaxies (after Mateo).

Galaxy	Year of discovery	RA 2000	Declination 2000	Parent*	Distance (million light-years)	Luminosity (million Suns)	Mass (million Suns)	Core diameter (light-years)	Outer diameter (light-years)
NGC 205	1864	00 40.4	+41 41	And	2.6	370	740	850	3500
NGC 185	1864	00 39.0	+48 20	And	2.0	130	130	500	8000
NGC 147	1864	00 33.2	+48 31	And	2.3	130	110	550	10 000
Sagittarius	1994	18 55.1	−30 29	MW	0.08	18	100–1000	1800	30 000+
Fornax	1938	02 40.0	−34 27	MW	0.45	16	68	1500	7700
And VII	1998	23 26.5	+50 42	And	2.5	5.7	—	520	—
Leo I	1955	10 08.5	+12 19	MW	0.81	4.8	22	700	2700
And I	1972	00 45.7	+38 00	And	2.6	4.7	—	1200	10 200
And II	1972	01 16.5	+33 26	And	1.7	2.4	60	675	7100
Sculptor	1938	01 00.2	−33 43	MW	0.26	2.2	6.4	360	4750
And VI	1998	23 51.7	+24 36	And	2.7	1.4	—	380	—
And III	1972	00 35.3	+36 31	And	2.5	1.1	—	390	1900
And V	1998	01 10.3	+47 38	And	2.6	1.0	—	500	—
Tucana	1985	22 41.8	−64 25	none	2.9	0.6	—	425	2250
Leo II	1950	11 13.5	+22 09	MW	0.66	0.6	9.7	525	1575
Sextans	1990	10 13.1	−01 37	MW	0.28	0.5	19	1100	10 500
Carina	1977	06 41.6	−50 58	MW	0.33	0.4	13	680	2230
Ursa Minor	1955	15 09.2	+67 13	MW	0.21	0.3	23	650	2100
Draco	1955	17 20.3	+57 55	MW	0.27	0.3	22	590	1850

Notes:* Name of Galaxy about which the listed dwarf orbits. And = the Andromeda galaxy; MW = the Milky Way; none = indicates that the galaxy appears to be unbound to any identified parent galaxy.

and small sizes are the features that principally distinguish dSph systems from all other classes of galaxies.

Many dSph galaxies also exhibit well-defined central nuclei that are far more compact than the galactic cores. In many cases, the optical appearance of the nuclei resemble that of large globular clusters. Nuclei in dSph galaxies are most common in the more luminous systems, but, even among the brightest dSph systems, not all possess nuclei. Only one dSph galaxy in the Local Group unambiguously contains a nucleus: NGC 205, a companion of the Andromeda Galaxy. These galaxies range in shape on the sky from nearly circular to highly elliptical. In terms of the HUBBLE CLASSIFICATION system, they range from E0 to E7. It is not unusual for the core of a given galaxy to exhibit a different elliptical shape and orientation than the outer region of the same system (see figure 3 of NGC 205 in the entry the Local Group).

Most of the radiation from dSph galaxies is emitted in the form of starlight in the optical portion of the ELECTROMAGNETIC SPECTRUM. The lack of strong emission lines, infrared, or radio radiation suggests that these galaxies are generally devoid of an interstellar medium composed of gas or dust floating freely between the stars. Recent observations have revealed a possible, though complicated, relation between some nearby dSph galaxies and clouds of neutral hydrogen gas visible only at radio wavelengths. These galaxies also emit very little light that can be attributed to extremely young stars or star-forming regions. This suggests that dSph galaxies are currently dormant systems that formed most of their stars more than about 1 billion years ago.

Nearby dwarf spheroidal galaxies

Because most distant dSph galaxies appear as featureless smudges of light, we must turn to the closest dSph satellite galaxies within the Local Group to try to understand the true nature of these little, but extremely common, systems. At the time of the initial discovery of Sculptor by Shapley and Baade, it was clear that dSph galaxies were composed of stars similar to those found in globular clusters. Moreover, dSph galaxies and globular clusters were the only sorts of objects known to inhabit the outermost regions of the halo of our Galaxy. These features and the simple structure of both types of star systems seemed to indicate that dSph galaxies were simply puffed up versions of otherwise normal globular clusters.

This belief persisted for many years, but in time enough evidence accumulated to begin to suggest that something far more complicated was going on within dSph galaxies. The Ursa Minor galaxy, for example, appeared to be so extended that it could not possibly be gravitationally bound. The implication was that we were witnessing the disruption of this dwarf galaxy as it passes close by the Milky Way. The problem with this was that independent calculations of the tidal forces suggested that Ursa Minor was sufficiently far from the Milky Way that it should not yet be falling apart. As astronomers began to carry out detailed studies of the individual stars in the closest dSph systems, they confirmed that the stars in dSph galaxies are deficient in heavy elements compared with the Sun. Oddly, the global properties of the stars in color–magnitude diagrams of dSph galaxies did not confirm these low element abundances. This seemingly contradictory behavior is known as the second

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

parameter problem and it is present in nearly every dSph system. These dwarfs also contain luminous CARBON STARS and unusual VARIABLE STARS that are not found in globular clusters.

In the early 1980s the first studies of dSph galaxies employing modern electronic detectors and large-aperture telescopes revealed two unexpected features that helped explain these puzzles. The first was that the stars with individual dSph galaxies appear to be moving far too fast relative to one another to remain bound within such small galaxies. That is, if we estimate the mass of each galaxy based on the number of visible stars, we find that every one of these galaxies appears to be disintegrating before our very eyes, not just Ursa Minor! Because this seemed highly improbable, the assumption that we see all the mass in dSph galaxies was questioned. One alternative model is that the galaxies are much more massive than expected from their visible matter because they contain DARK MATTER. Dynamical models that include dark matter do adequately explain the internal motions of the stars in all dSph systems. In the most extreme cases, only 1% of the mass of the galaxy is visible. The rest of the mass is composed of dark matter whose nature, unfortunately, remains a complete mystery. If this is true, dSph galaxies have the largest proportion of dark mass of any known type of galaxy and can therefore claim to be the 'darkest' galaxies known. Curiously, dark matter models also suggest that no dSph galaxy has a mass less than about 10 million times the mass of the Sun.

The second revolutionary observation proved that, unlike globular clusters, dSph galaxies contain significant numbers of 'intermediate-age stars'. Thus, these galaxies appear to have formed stars over a large fraction of the lifetime of the universe while globular clusters generally formed nearly all of their stars very soon after the universe first formed. Modern studies of nearly all the nearby dSph galaxies reveal evidence of a complicated star formation history, each galaxy being different from all the others much like the ages of children may differ greatly from family to family even within a small town or neighborhood. How and why individual dSph galaxies have been able to support extended and episodic star formation histories remains unclear.

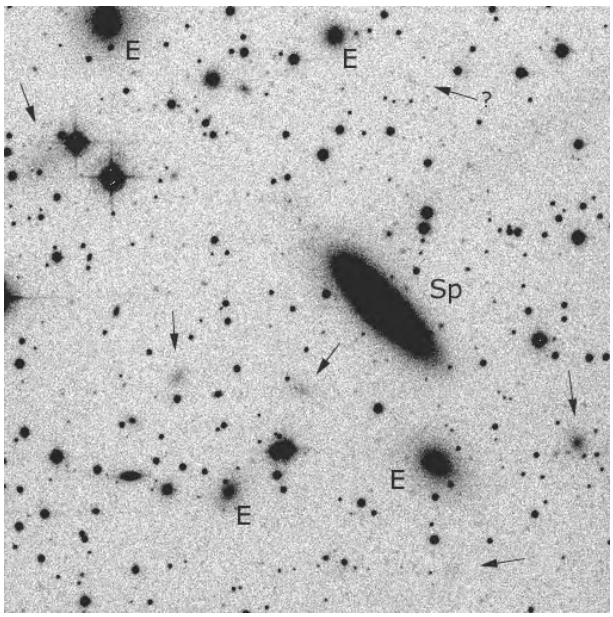


Figure 2. A close-up of a small region of the nearby Centaurus galaxy cluster. Five dSph galaxies within the cluster are denoted with arrows, and a sixth uncertain case is identified with an arrow and a question mark. The galaxies denoted with an 'E' are normal dwarf elliptical galaxies in the cluster and the elongated galaxy denoted 'Sp' is a spiral galaxy of the cluster viewed at a nearly edge-on orientation. Note how extremely diffuse and faint the dSph galaxies are relative to the other galaxies and to the brightness of the night sky itself. These features uniquely distinguish these dwarfs from normal ellipticals. Clusters such as Centaurus appear to be typical: dSph galaxies are the most common single type of galaxy known in the present-day Universe. This negative image was obtained at the Las Campanas Observatory of the Carnegie Institution of Washington and was provided courtesy of Kristin Chiboucas.

Dwarf spheroidal galaxies throughout the universe

Every galactic group or cluster that has been adequately studied has revealed large numbers of dSph galaxies. Most of these galaxies appear to be quite similar to the ones found in the Local Group, but some interesting variations are also found. For example, the nearby M81 group contains a few dSph galaxies that are considerably more extended and have even lower central surface brightnesses than the known local systems. Nearby galaxy clusters such as the VIRGO, FORNAX, CENTAURUS and COMA CLUSTERS contain hundreds to thousands of individual dSph galaxies (figure 2). Dwarfs with nuclei are particularly common in these environments, especially among the brighter galaxies. In most cases, the nuclei appear to contain stars that are distinctly different—possibly older or with different elemental abundances—from the majority of stars in the parent dSph galaxies.

Dwarf spheroidal galaxies appear to concentrate more towards the centers of galactic groups and clusters than any other type of galaxy, including spiral galaxies, large elliptical galaxies or the relatively dust-free S0 DISK GALAXIES. This is also observed in the Local Group where the vast majority of all dwarf companions nearest the Milky Way and the Andromeda Galaxy are dSph systems. Astronomers speculate that proximity of dwarfs to larger galaxies or the centers of large clusters helps remove gas from the smaller systems as a result of tidal forces or perhaps from interactions with isolated gas clouds also orbiting the larger systems. Either process can truncate star formation in the affected systems, allowing them to fade to very low surface brightnesses and ultimately resulting in the formation of dSph galaxies. Alternatively, dwarf galaxies near larger companions or near the dense centers of clusters may accelerate their star formation due

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to interactions with their larger neighbors. Such galaxies would run out of gas and appear today to contain mostly middle-aged stars.

Two observations do suggest that dSph galaxies may interact tidally with their massive parents. First, none of these galaxies are found extremely close to a large galaxy, implying that tidal forces destroy dSph systems that venture too near to their parents. In the case of the Milky Way, the limiting distance appears to be about 150 000 light-years. Second, we actually see one example of the disruption occurring today. The Sagittarius dSph is unique among all known dSph galaxies in its enormous size and low density, both indicative of tidal disruption. This case really does appear to be one of a dSph galaxy in its final death throes. Models of its interaction with the Milky Way suggest that within about one billion years, Sagittarius will have dissolved away as a distinct galaxy, its stars scattered through the halo of our Galaxy.

Beyond about 100 million light years, dSph galaxies appear too small and too faint to be identified easily. However, their presence may still be inferred indirectly. Deep counts of galaxies have revealed a population of so-called faint blue galaxies located at such great distances that we are seeing the systems as they appeared many billions of years ago. Some of these could conceivably be dSph galaxies caught during one of their many strong episodes of star formation. Such events make the galaxies brighter and much bluer and consistent with the properties of at least some of the faint blue galaxies observed. Faint IRREGULAR GALAXIES appear to have global properties similar to dSph galaxies, but with the addition of young stars and a significant interstellar medium. Perhaps these represent modern-day examples of galaxies that have undergone evolution from a young, blue phase and are changing into typical dSph galaxies. Because dSph galaxies are so common and because most popular models of GALAXY FORMATION in the early universe suggest that small galaxies formed first, it may even be possible that dSph galaxies represent the basic unit of galaxy formation. If so, the dSph systems we see today may represent fossils of objects similar to the very first galaxies ever formed.

Bibliography

van den Bergh S 2000 The galaxies of the Local Group (Cambridge: Cambridge University Press) (A galaxy by galaxy description of the Local Group; many historical comments included. All of the local dSph galaxies are included.)

Hodge P 2000 *An Atlas of Local Group Galaxies* (Dordrecht: Kluwer) (An atlas of photos of all Local Group galaxies including all the nearby dSph systems.)

Impey C and Bothun G 1997 Low surface brightness galaxies *Ann. Rev. Astron. Astrophys.* **35** 267–307 (A technical review of low surface brightness galaxies that offers some insight into dSph in distant environments.)

Mateo M 1998 Dwarf galaxies of the Local Group *Ann. Rev. Astron. Astrophys.* **36** 435–506 (A recent technical review of the Local Group including the dSph galaxies.)

Mario Mateo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dwingeloo Observatory

The Dwingeloo 25 m Telescope in Dwingeloo, The Netherlands, was inaugurated in 1956 and was for a short period of time the largest telescope in the world. For many years the Dwingeloo Telescope was the bread-and-butter instrument of the Dutch radioastronomy community. Its main operating frequency band around 1400 MHz allowed studies of the distribution of neutral hydrogen in our Galaxy and other nearby galaxies. The all-sky survey of the 21 cm line radiation of neutral hydrogen and the discovery of an unknown galaxy behind the plane of the Galaxy show that the DWT was a viable instrument until the end of its active life. The Dwingeloo Telescope was closed for observations in 1998 after 42 years of service.

The Dwingeloo Telescope was operated by the Netherlands Foundation for Research in Astronomy (NFRA), which is funded by the Netherlands Organization for Scientific Research (NWO).

For further information see http://www.nfra.nl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dyer Observatory

The Arthur J Dyer Observatory is the principal astronomical facility of Vanderbilt University. It is located about 16 km south of Nashville near Brentwood, Tennessee at an elevation of 345 m above sea level.

Built in 1953, the observatory houses a Seyfert 60 cm Cassegrain/reflector–corrector telescope and the DeWitt 30 cm Cassegrain, which is equipped for photoelectric photometry and long-focus photography. The telescopes are sometimes opened to the public.

For further information see http://www.dyer.vanderbilt.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamical Reference Frame

The dynamical reference frame is a realization of a dynamical reference system which is based on the motions of the solar system (see SPACE AND TIME REFERENCES: CONCEPTS). It was the standard reference frame until it became possible to observe the positions of very distant radio sources with great accuracy.

The principle of the definition of a dynamical system of coordinates is to set the condition that the motion of celestial bodies can be represented by a solution of a system of differential equations of motion written in a fixed triad. This is verified by the fact that there is no need of any additional acceleration to account for the motion. To do away with a linear acceleration, one sets the center of the coordinates on the barycenter of the solar system, assuming that one takes into account the Coriolis acceleration due to the motion around the center of the Galaxy. Stopping the rotation of the axes is the major problem in building a dynamical frame, because there are no materialized invariant directions and one is reduced to describe as precisely as possible the motions of fiducial points accessible to observations.

Actually, there is a fixed reference plane defined by the solar system. It is the invariable, or Laplacian, plane. This is the plane normal to the axis of angular momentum and passing through the barycenter. This plane can be defined theoretically, but it is not observable. In addition, it is somewhat dependent upon the completeness of the bodies of the system included in its determination, and hence is unfit for practical use.

So, in practice the choice for the definition of the dynamical reference frame is the mean equatorial plane, the ecliptic plane, which is the mean orbital plane of the Earth without planetary perturbations, and one of the intersections of these planes, the equinox, which is the fiducial point.

The Earth, being our habitat and base for observations, becomes a natural basis for most reference frames. The equatorial plane is the most obvious and necessary reference plane, since the motions of all observatories are with respect to that plane. It can be determined from observational data either by determining the celestial pole or from diurnal motions of celestial sources. However, it is not fixed and moves due to gravitational forces by the Sun, Moon and planets on the bulge of the Earth and due to the dynamics of the many components of the interior, surface and exterior of the Earth. These motions are called precession, nutation and polar motion (see EARTH'S ROTATION).

Another natural plane of reference is the orbital plane of the Earth. This is subject to motions due to the perturbations of the planets and is changing continuously. The mean orbital plane of the Earth is called the ECLIPTIC. As long as a general theory by Newcomb was used to determine the ephemeris of the Earth, the ecliptic was well defined by the mean elements. With a numerical integration for the ephemeris of the Earth, the time period required to define the mean orbit is uncertain. The plane is also dependent on the ephemerides used to determine it.

The equinox is defined as the intersection of the eclipitic and the mean equator of the Earth. In determining this intersection from numerically integrated EPHEMERIDES, a distinction must be made. The equator can be taken as the instantaneous plane or as a uniformly moving plane. The historical determination was made based on the uniformly moving plane and that is the definition that has continued to be used. The difference between the two possible definitions of the equator is approximately a tenth of an arc second. For calendars, seasons, and holiday definitions the equinox is specified as the time when the Sun passes through either of the intersection points, i.e. when the apparent longitude of the Sun is zero or 180°.

Since both the equator and the ecliptic move due to solar, lunar and planetary gravitational forces on the shape of the Earth, and perturbations by the planets on the Earth's motion, the equinox moves with time. Therefore, the dynamical reference frame had to be specified for a specific epoch and observations and predictions were transformed to and from standard epochs to the chosen times of observations. The dynamical reference frame served as the fundamental reference frame of astronomy until 1998, when it was replaced by the International Celestial Reference System (ICRS), based on extragalactic radio sources, and thus fixed in space (see EXTRAGALACTIC REFERENCE FRAMES).

The new fiducial point does not have to revolve as the equinox did. The definition was arbitrary, but due to the inter-dependence of the definitions of time and coordinate systems, some optimum choices were possible.

The realization of the dynamical reference frame has been based on fundamental catalogs determined from observations of bright stars. These catalogs contained positions and proper motions based on some value of precession for the motion of the equinox. The fundamental catalogs were of two series, the US and German series (Eichhorn 1974). The Fundamental Catalogues series (FK3, FK4, and FK5) were used until 1998. The FK3 and FK4 were based on Newcomb's value of precession and a dynamical theory of motion of the solar system. FK5, introduced in 1984, was based on a revised value of precession and Jet Propulsion Laboratory Development Ephemeris DE 200. The introduction of the FK5 was also the occasion for the introduction of a new theory of nutation, new astronomical constants, and new definitions of timescales.

It was recognized that the FK5 defined a catalog equinox which was the origin of the right ascensions within the catalog. While this was attempted to be equivalent to the dynamical equinox defined by the solar system ephemerides, there was inevitably a difference between these two origins, called the catalog and dynamical equinoxes respectively, and this difference could be a function of time. The new ICRS defines a reference frame which is supposed to have its origin equivalent to the dynamical equinox of J2000.0, but again

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

there will inevitably be a difference between the equinox determined from the ephemerides and the ICRS origin. This difference will be dependent upon the ephemerides which will now be determined from observations reduced directly on the ICRS,

The motion of the equatorial plane is due to the torque of the Sun, Moon and planets on the dynamical figure of the Earth, it has been divided into long and short period terms. The smooth long-period motion of the mean pole of the equator about the pole of the ecliptic with a period of about 26 000 yr is lunisolar precession. The short-period motion, determined from the NUTATION theory, is made up of many terms with periods less than 18.6 yr, with a maximum amplitude of about 9″.0 (see EARTH'S ROTATION: THEORY).

The motion of the ecliptic is due to the gravitational action of the planets on the Earth's orbit and makes a contribution to precession known as planetary PRECESSION. If the equator were fixed, this motion would cause a precession of the equinox of about 12″ per century and a decrease in the obliquity of the ecliptic of about 47″ per century. The combination of lunisolar and planetary precession is called general precession. The causes of precession and nutation are the same, only the periods are different. There is some consideration being given now to eliminating the distinction.

At this time the officially adopted constant of general precession in longitude per Julian century at epoch J2000.0, $p = 5029.''0906$, is known to be incorrect by about $-0.''2957$ per century. The IAU 1980 Theory of Nutation is known to need improvement. The nutation theory for a rigid Earth has been improved and extended to smaller terms. However, the transformation from the rigid Earth model to the real non-rigid Earth model has not been determined to be consistent with observational data. Therefore, for accuracies of ± 1 mas the 1980 theory of nutation is satisfactory. For better accuracies, it is recommended that numerical values determined and distributed weekly by the International Earth Rotation Service (IERS) be used for both precession and nutation.

Bibliography

Eichhorn H 1974 *Astronomy of Star Positions* (New York: Frederick Unger)

Seidelmann P K (ed) 1991 *The Explanatory Supplement to the Astronomical Almanac* (Mill Valley, CA: University Science Books)

P Kenneth Seidelmann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamics

A galaxy is made up of stars in the same way that a crystal is made up of atoms, and the galaxy's large-scale properties reflect the forces that act between its component stars in the same way that the structure of a crystal reflects the nature of interatomic forces. Galaxies and crystals have strongly contrasting large-scale properties because a galaxy is held together by a force that has very different properties from a typical interatomic force.

For most purposes we may neglect the physical size of stars and consider them to be point particles that interact only through gravity. Gravity is a uniquely long-range force, so the net force on an individual star depends directly on the locations of every other star in the galaxy. By contrast, the net force on an atom in a crystal depends only on the location of up to a few dozen near neighbors.

Given that there are in excess of 10^{11} stars in a galaxy such as our own Milky Way, progress in STELLAR DYNAMICS would be slow if we had to know the locations of every star before we could calculate the motion of a single star. Fortunately, the problem can be simplified by arguing that the net force on a star does not depend sensitively on the precise locations of the billions of stars that are rather distant from it, so an excellent estimate of the contribution these stars make to the net force can be obtained by mentally smearing the mass of these stars into a smooth continuum of density. Mathematically, if in some region there are ν stars per unit volume, then we calculate the gravitational attraction from a smooth distribution of matter of density $\rho = \langle m \rangle \nu$ units of mass per unit volume, where $\langle m \rangle$ is the average mass of a star. The required gravitational attraction is easy to calculate if we first determine the gravitational potential that is generated by the density distribution ρ.

Replacing individual stars by a smoothed-out distribution of mass yields an indifferent estimate of the part of the force on a star that arises from the star's near neighbors. In particular, it completely neglects the fact that the star sometimes experiences short intense periods of acceleration when a neighboring star passes unusually close. This neglect of 'collisions' can be made good at a subsequent stage of the analysis, when collisional effects are added in to the 'collisionless dynamics' that are determined by approximating the system's mass distribution by a smooth one.

The longer a stellar system evolves, the more important collisional phenomena become. As a general rule, the time required for collisional effects to become important in a system of N stars is $\sim N/8 \ln N$ times the mean time it takes a star to cross the system. This time is substantially larger than the age of the universe in the case of a typical galaxy but significantly shorter than its age in the case of a typical STAR CLUSTER.

In this article we review the basic principles that govern the dynamics of galaxies and star clusters. More information on most topics will be found in the monograph [2], which will not be explicitly cited again.

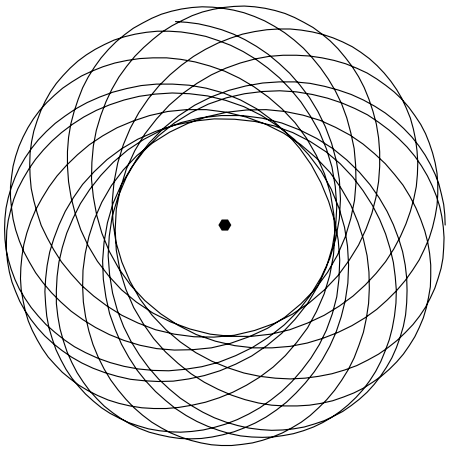


Figure 1. A typical orbit in a spherical potential.

Collisionless dynamics

A stellar system is usefully thought of as a collection of orbits in the smoothed-out potential of the system, each of which is populated by a certain number of stars. Consequently, a primary task of stellar dynamics is to understand what orbits can arise in a given gravitational potential.

The easiest case to consider is a spherical potential. Then every star moves in the plane perpendicular to the star's ANGULAR-MOMENTUM vector. Within this plane the star describes a rosette (figure 1). The orbit is best quantified by its angular momentum, $\boldsymbol{L}$, and its 'radial action' $J_{\rm r}$, which quantifies the vigor of the star's radial excursions: a circular orbit has zero radial action, $J_{\rm r} = 0$. The direction of the angular-momentum vector $\boldsymbol{L}$ defines the orbital plane, while the magnitude of the vector, L, controls the size and eccentricity of the orbit – if $J_{\rm r}/L$ is small, the orbit is nearly circular, and the orbit is physically small if L is small and large if L is large.

Few galaxies have spherical potentials, but many have potentials that are nearly axisymmetric. How do the orbits in a spherical potential change if we squash the potential along some axis, so that it remains axisymmetric about that axis? In a squashed potential, orbits are not confined to planes, but we can obtain a reasonably accurate picture of them by considering them to be confined to precessing planes. That is, the orbit is approximately confined to a plane that is inclined at some angle to the potential's symmetry axis, and slowly rotates about this axis. On the plane the orbit remains a rosette. The shape of this rosette is controlled by $J_{\rm r}$ and a number called I_3 in the same way that the shape of a rosette in a spherical potential was controlled by $J_{\rm r}$ and L. The inclination of the precessing orbital plane to the potential's symmetry axis is determined by the ratio I_3/L_z of I_3 to the star's angular-momentum about the potential's symmetry axis.

The potentials of many galaxies deviate significantly from axisymmetry, so let us ask what happens if we squash an axisymmetric potential along an axis perpendicular to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

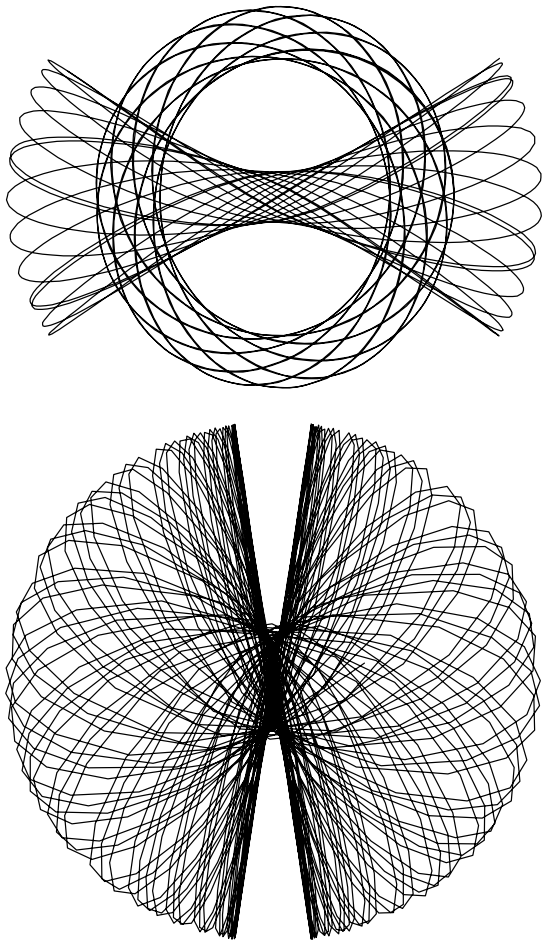


Figure 2. Three orbits in a triaxial potential. The contours of constant gravitational potential are elongated horizontally with axis ratio 0.9.

its symmetry axis. Figure 2 shows that the consequences of deforming a potential from axisymmetric to triaxial symmetry can be dramatic. While some orbits deform only slightly, others strongly amplify the deviation of the potential from axisymmetry.

When the potential of a galaxy is triaxial, it will usually tumble around one of its principal axes. The orbits shown in figure 2 are in a potential whose axes are fixed rather than tumbling. Figure 3 shows one important consequence of letting the potential tumble. The rate at which the potential tumbles defines a characteristic radius, the corotation radius, $r_{\rm CR}$, at which a star that moves around with the potential moves on a circular orbit. At radii much smaller than $r_{\rm CR}$, the rotation of the potential's figure is unimportant. At radii larger than about $r_{\rm CR}/10$ the orbits are profoundly modified by the figure rotation, and at radii close to $r_{\rm CR}$ a large fraction of the orbits are chaotic. Chaotic orbits such as that shown in figure 3 look rather ragged and have a propensity to change their character suddenly and irreversibly. It is likely that the steady accumulation of such changes drives the evolution of some galaxies, for example LENTICULAR GALAXIES.

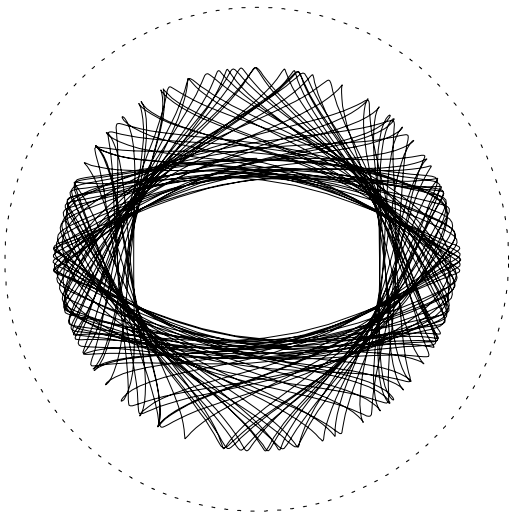


Figure 3. A chaotic orbit in a rotating barred potential. The dashed curve shows the potential's corotation radius.

Entropy and equilibrium

In stellar dynamics we seek statistical information in the sense that we would like to know the probability of the occurrence of any given configuration—in which given numbers of stars are on each orbit. In ordinary statistical physics the probability P that a given quantum state occurs can usually be obtained from the Gibbs distribution, $P \propto \mathrm{e}^{-\beta E}$, where $\beta = 1/kT$ is the inverse temperature and E is the energy of the state. Can this distribution be carried over into stellar dynamics?

The answer to this question is 'no'. In fact, the concept of thermodynamic equilibrium ultimately fails for any self-gravitating system, including every stellar system. The reason for this failure is very interesting and can be understood by considering a self-gravitating system of mass M and characteristic radius R. The energy of the system is then of order $U = -GM^2/2R$, and, by the 'virial theorem', the mean-square random velocity of the system's constituent particles (atoms or stars), $\langle v^2 \rangle$, is of order $\langle v^2 \rangle = GM/R$. As in ordinary kinetic theory, we identify the system's temperature as

$$T = \frac{\langle m \rangle \langle v^2 \rangle}{3k} = \frac{GM\langle m \rangle}{3kR} = -\frac{2U}{3Nk}$$

where k is Boltzmann's constant and N is the number of particles (atoms or stars) in the system. Now imagine putting the system into thermal contact with a slightly cooler reservoir of heat. Heat will flow out of the system into the reservoir. By conservation of energy, the system's internal energy U will become more negative and, by the equation above, T will increase. This behavior is quite the reverse of what we observe in every-day life: when heat

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

flows out of a coffee cup, it cools down rather than heats up.

We can show that this anomalous behavior of self-gravitating systems drives them away from thermal equilibrium by considering such a system to be made up of two parts: the core and the envelope. If the system is very centrally concentrated, the core is orders of magnitude denser than the envelope and the core's equilibrium is very little affected by the presence of the envelope, so the relation between internal energy and temperature that we derived above can be applied to the core alone. The envelope is confined by a mixture of the 'external' gravitational field of the core and its own self-gravity, so its temperature and internal energy will be connected by a relation that lies somewhere between that derived above for a perfectly self-gravitating system and that appropriate for a classical gas, in which U is positive and $T = \frac{2}{3}U/Nk$. In general we may write $T = \alpha U/Nk$, where $-\frac{2}{3} \leq \alpha \leq \frac{2}{3}$ depending on the degree to which the envelope is self-gravitating.

If initially the core and the envelope are in thermal equilibrium, they must have a common temperature T. Consider a thermal fluctuation that transfers a small quantity of heat, δU, from core to envelope. This transfer causes the temperature of the core to rise by $\delta T_{\rm core} = 2\,\delta U/3Nk$ (because the energy of the core changes by $-\delta U$), while the temperature of the envelope will rise by $\delta T_{\rm env} = \alpha\,\delta U/Nk$. Since the envelope must be to some extent self-gravitating, $\alpha < \frac{2}{3}$ and $\delta T_{\rm core} > \delta T_{\rm env}$. Consequently, after the fluctuation, the core is hotter than the envelope, and the temperature gradient between core and envelope starts to drive a steady flow of heat from core to envelope. This flow increases the temperature difference between core and envelope, and a 'gravithermal catastrophe' ensues in which $T_{\rm core}$ rises inexorably.

The sequence of events that we have just described prevents any classical self-gravitating system, whether it be a star or a galaxy, from achieving thermal equilibrium. None the less, the concept of thermal equilibrium is invaluable in understanding the structure of a star, because in a star there is usually a well-defined temperature at any given radius since individual atoms or photons do not stray far between collisions. In a stellar system, by contrast, stars move at constant energy through wide ranges in radius, and the distribution of stellar velocities at a given radius is not characteristic of any particular temperature.

Although the analogy between the structure of a star and a stellar system is imperfect, the theory of STELLAR EVOLUTION does provide valuable insights into the dynamics of stellar systems. Stars start out in life relatively homogeneous. Gradually the gravithermal catastrophe causes the core to contract and heat up, and the envelope to expand and cool, with the observational consequence that the star becomes a RED GIANT. All stellar systems are at some point on this trek towards every greater central condensation. It is a journey without end because one can show that, no matter what its current configuration, the entropy of a stellar system can be increased by a transfer of energy from the dense interior to the more rarefied exterior.

Dynamics of star clusters

What varies from system to system is the rate at which such transfers are proceeding and the physical processes responsible for them. The systems that evolve in closest analogy with a star are star clusters. In these systems, which contain from a few hundred to perhaps a hundred thousand stars, collisional processes are important in two contexts. First, they effect a transfer of heat from the contracting core to the expanding envelope. Second, they enable energy to be released in the core in much the same way that in a stellar core nuclear reactions release energy. The mechanism of this energy release is as follows. When a single star approaches a BINARY STAR closely, it exchanges energy with the binary and will often form a triple star in which all three stars are bound. Subsequently, two of the three stars will move towards one another, releasing energy which is used to eject the third star at speed. Hence, the net result of the formation of the triple star is the replacement of a relatively slow-moving free star by a much faster one. Over time, this faster star will tend to lose energy to stars that it encounters as it moves through the cluster. Binary stars, being more massive than most single stars, tend to sink to the core. Moreover, the rate at which any given binary star will suffer close encounters with single stars is highest in the dense environment of the core. So most of the energy that is liberated by binary–star encounters heats the core rather than the envelope. It follows that a star cluster resembles a star in that its core is heated and its envelope transmits a flux of thermal energy outwards.

A star cluster's core responds to heating by expanding, and thus reducing the rate at which binary–star encounters heat it. Consequently, oscillations in the density of the core can arise, in which expansion and decreased heating is followed by contraction and increased heating, and then the cycle repeats [6].

Dynamics of elliptical galaxies

ELLIPTICAL GALAXIES differ from GLOBULAR CLUSTERS principally in scale: whereas a globular cluster contains of order 10^5 stars, a massive elliptical galaxy contains in excess of 10^{12} stars. In addition to being much more massive than globular clusters, elliptical galaxies are physically larger too. In fact, a massive elliptical galaxy may have several thousand globular clusters bound to it, with ample space for them to move through the galaxy just like stars. The vast number of stars in an elliptical galaxy, and the huge volume within which they are distributed combine to render collisional effects entirely unimportant. So whereas collisional effects are constantly moving a globular cluster down a well-defined evolutionary path, an elliptical galaxy spends most of its time stuck in a particular configuration. What physical processes got it into the configuration in which it is now frozen?

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The answer to this question is not entirely clear, but undoubtedly two important processes are mergers and cannibalism.

In the absence of collisional effects, an individual star can change the energy of its orbit only when the overall galactic gravitational potential is changing. This fact implies that the evolution of a collisionless galaxy is concentrated into brief episodes of violent change: revolution rather than evolution. During one of these episodes of change, rapid change in the overall gravitational potential causes individual stars to gain and lose energy fairly rapidly. This vigorous exchange of energy between stars in turn drives the rapid evolution of the overall potential.

An episode of rapid change in the gravitational potential can be initiated by a close encounter with another galaxy. If the intruding galaxy comes close enough and moves slowly enough, the two galaxies become a binary galaxy within which the gravitational potential is constantly changing as the galaxies swing around each other while moving alternately closer together and further apart.

As the gravitational potential changes, some stars gain energy and others lose it, but overall more energy is picked up than lost by stars. The net gain of energy by stars is mirrored by a net loss of energy from the relative orbit of the two galaxies. Hence the galaxies spiral together and eventually merge (see GALAXIES: INTERACTIONS AND MERGERS).

The time required for the galaxies to merge completely depends on their relative initial masses. If the two galaxies are initially of comparable size, merging is complete by the time the galaxies have orbited around each other two or three times. If one galaxy is much less massive than the other, more orbits are required for merging—in order of magnitude, the number of orbits required is equal to the mass of the bigger galaxy divided by the mass of the smaller. When this ratio is large, the larger galaxy seems to eat the smaller, and one speaks of 'cannibalism'.

The configuration of the galaxy that emerges from a merging event depends significantly on the mass ratio of the merger. If the mass ratio is close to unity, the merger product tends to be triaxial in shape and at most slowly rotating. Larger mass ratios are more likely to produce axisymmetry and significantly rotating merger products [7].

Nuclear black holes

It is likely that all substantial galaxies contain massive central BLACK HOLES—we know for certain that our own Milky Way has a black hole of mass $\sim 2 \times 10^6 M_\odot$, and that our nearest giant neighbor, the ANDROMEDA GALAXY M31, has an even more massive nuclear black hole. The presence of nuclear black holes in galaxies that merge has important consequences for the dynamics of the merger remnant. First, and most conspicuously, the irregular and time-varying gravitational potential of the merging galaxies feeds a significant fraction of any gas in the system onto the black holes, which are then liable to become ultraluminous as they accrete some of this abundant gas supply. In the final stage of the merger, when the nuclei of the two galaxies have spiralled very close together, the black holes can have a significant impact on the distribution of stars in their vicinity. Indeed, the merger leads to the formation of a binary black hole that initially moves through a centrally concentrated sea of background stars. One by one the stars gain energy from the binary black hole and move out to larger radii. Over time the density of the stellar sea ceases to be centrally concentrated. In consequence of losing energy to background stars, the orbit of the binary black hole shrinks and shortens in period until it begins to be relativistic in the sense that its period is not extremely long compared with the time light requires to cross it. When this stage is reached, the radiation of GRAVITATIONAL WAVES by the orbit becomes the dominant mechanism of energy loss. As the orbit shortens further in period, the intensity of the radiated gravitational waves and the rate of orbital decay rapidly accelerate. The merger of the holes must then ensue, surely one of the most awesome events the universe is ever called upon to witness.

Over a long time-scale, the final, merged, black hole can profoundly modify the enveloping galaxy [5, 8]. The galactic merger will frequently have left the latter triaxial. Its triaxiality is sustained by 'box' orbits like that shown in the top panel of figure 2. A star on such an orbit is liable to pass very close to the galactic center, where the black hole resides. If on some passage it comes very close to the black hole, it will be deflected through a large angle and emerge on a very different orbit, like that shown in the bottom panel of figure 2. Clearly, when a sufficient number of stars have been transferred in this way from elongated orbits to fat orbits, the galaxy will become less elongated overall. Moreover, this is liable to be a run-away process: transfers of stars from elongated to fat orbits make the overall galactic potential less elongated, and this reduction in the elongation of the potential reduces the elongation of the orbits of other stars, that have passed nowhere near the black hole.

The details of these processes—the ejection of stars from galactic centers by binary black holes and the probably catastrophic erasure of triaxiality following scattering of high-energy stars by a merged black hole—have not been fully worked out [9], but it is likely they account for the observed fact that lower-luminosity elliptical galaxies have cuspy and probably axisymmetric centers, while more luminous systems have more nearly homogeneous and probably triaxial centers [1].

Dynamics of disk galaxies

Many, perhaps essentially all, stars form in centrifugally supported disks of gas. Consequently, the dynamics of disks of gas and stars is of vital importance for galactic dynamics (see also DISK GALAXIES).

Imagine a disk of stars in which every star is on a circular orbit. This is an equilibrium configuration, but it turns out to be a violently unstable one that will quickly break up into rings of enhanced density separated by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

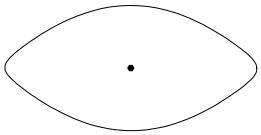


Figure 4. In a suitably rotating frame of reference the rosette orbit of figure 1 closes. The filled point marks the centre of the potential.

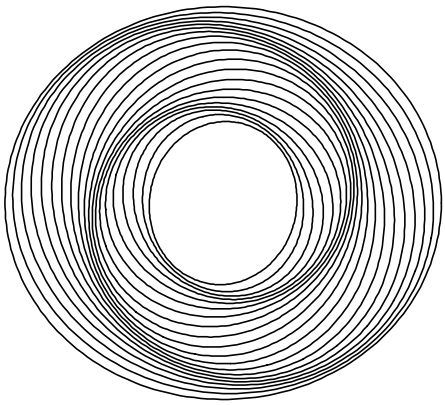


Figure 5. When elliptical orbits are arranged so that the orientation of each major axis changes smoothly with radius, spirals of enhanced density are generated.

density depressions. To avoid rapid fragmentation, the stars of a disk must have random velocities in addition to rotating around the galactic center, and the higher the surface density of the disk, the larger the random velocities of stars need to be if fragmentation is to be avoided.

Since the stars of a stable disk have random velocities, they have to be on non-circular orbits, so they must be on rosette-like orbits such as the one shown in figure 1. If we view such an orbit from a frame of reference that rotates at the right frequency, ω_p, the orbit looks closed as in figure 4. If we populate such an orbit with many stars uniformly distributed in phase around the orbit, we have an elliptical hoop of stars, whose long axis rotates steadily at the angular velocity ω_p—we say that the hoop 'precesses' with frequency ω_p. Notice that the precession frequency ω_p is entirely distinct from the frequency Ω at which individual stars circulate around the hoop—Ω is generally much greater than ω_p. Figure 5 shows what happens if we superimpose many hoops of stars, each with a slightly different radius and initial orientation of its long axis. Spiral arms of enhanced density appear.

Figure 5 describes the essence of how the arms of SPIRAL GALAXIES are constructed. What makes the dynamics of these systems extremely complex is the fact that, in an axisymmetric potential, the precession frequency ω_p varies with the radius of the hoop. In fact, the precession frequency of the innermost orbits in the figure will generally be higher than that of the outermost orbits shown, with the consequence that the spiral pattern will gradually wind up into an ever tighter spiral.

In reality, the spiral arms will contribute a significant non-axisymmetric component to the gravitational potential, and we need to ask how this component affects the precession frequencies of individual hoops. The non-axisymmetric component of the potential applies torques to each ring, and these torques can either increase or decrease the hoop's precession frequency. The idea behind the normal mode theory of spiral structure is to find a configuration of the hoops—essentially a rule, $\phi_0(a)$, for the initial orientation of a hoop's major axis as a function of semimajor axis length, a—such that the torques speed up the slowest hoops and slow down the fastest ones, so that all hoops rotate at the same rate, and the spiral pattern that they form rotates like a rigid body. In these circumstances, ω_p is called the 'pattern speed' of the spiral arms.

Determining the function $\phi_0(a)$ is a mathematically challenging problem and the theory of spiral structure has not often been approached by this direct route. The usual approach is to start from the insight that a self-gravitating disk of stars is a medium that is capable of supporting waves, in the same way that the surface of a pond is. If we disturb the stars at some radius, r_0, we will change the way they contribute to the overall surface density near r_0, and thus we will change the gravitational potential in which other, previously undisturbed stars move. Hence, the motion of these other stars will be disturbed, and they will in turn disturb yet other stars. In fact, our disturbance will spread from radius r_0 like a ripple over a pond.

When waves move in a confined region, such as a teacup or a piano wire, resonant frequencies arise. Waves with these frequencies can travel from a source, and, after possibly multiple reflections from the boundaries, return to the point of origin with the same phase regardless of the number of times that they have been reflected. Because the waves always return with the same phase, they interfere constructively and achieve a significant net amplitude. The resonant frequencies of a simple system such as a piano wire can be exactly determined by finding the frequencies at which waves that are reflected numbers of times interfere constructively. The resonant frequencies of a teacup cannot be found exactly in this way, but good approximations to them can be found thus.

The 'density-wave' theory of spiral structure interprets spiral structure as generated by resonant spiral waves and yields approximate values for the possible pattern speeds and shapes of spiral arms. A significant complication of the theory arises because a typical galactic disk does not have well-defined boundaries like those of a teacup or a piano wire. Boundaries of a sort emerge once one chooses a pattern speed ω_p for the desired spiral pattern. These boundaries are defined by radii at which a star moving on a nearly circular orbit in the undisturbed axisymmetric potential of the galaxy can resonate with a disturbance of frequency ω_p. The inner boundary is the radius of the 'inner Lindblad resonance' (ILR). At this radius the star rotates faster than ω_p, so it regularly overtakes the spiral wave pattern, and the frequency at which it experiences the wave coincides with the frequency, ω_r, of its

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

natural oscillations in radius. The outer boundary is the edge of an annulus in which spiral disturbances cannot propagate. This annulus is centered on the 'corotation resonance', which is the radius at which a star could move on a circular orbit at frequency $\omega_{\rm p}$. Spiral structure arises when leading spiral waves propagate out towards corotation and are reflected and amplified into trailing waves at the edge of the corotation annulus. The amplified trailing waves then propagate in towards the ILR, where they are refracted into leading waves that move back out to corotation, completing the cycle.

A piano wire is not fundamentally changed by the passage along it of waves. A stellar disk, by contrast, is significantly modified by spiral waves. Such waves increase the random velocities of stars that are in resonance with them. This increase in random velocities is a reflection of the tendency of trailing spiral arms to transport angular momentum out from the ILR and thus to enable stars at the ILR to move inwards. In fact, spiral arms achieve for stellar disks something very similar to what mergers achieve for elliptical galaxies: an increase in the central concentration and entropy of the system. Since the ability of a disk to sustain spiral arms decreases as the random velocities of stars increase, spiral structure will eventually die away in a purely stellar disk. Spiral galaxies such as our own have sustained strong spiral structure for a Hubble time because they have star-formation rates that are fairly constant in time. Young stars have small random velocities and so enhance the disk's ability to support spiral arms.

If one releases from an axisymmetric equilibrium a disk in which stars have very small random velocities, the disk very quickly becomes strongly non-axisymmetric. The non-axisymmetry is not primarily due to spiral structure, however, but to a tumbling bar. The dynamics of a tumbling bar is analogous to that of the spiral arms shown in figure 5: large numbers of orbits are elongated with major axes that rotate at a common frequency, $\omega_{\rm p}$. A bar differs from a spiral disturbance in that (i) individual orbits tend to be much more strongly elongated in a bar, and (ii) in a bar the major axes are all aligned, irrespective of the radius of the orbit. A large proportion of spiral galaxies, including the Milky Way, have bars at their centers. In a strong bar, a significant fraction of all stars have a chance to pass arbitrarily close to the galactic nucleus, where a massive black hole is expected to lurk. In close analogy to what happens in a triaxial elliptical galaxy, deflection of stars by the gravitational potential of the black hole can lead to the dissolution of the bar [10].

Dynamics of the Milky Way

We live far out in the disk of a BARRED SPIRAL GALAXY, and from this vantage point can study many of the physical processes described above [1]. Figure 6 is a plot of the random velocities of groups of solar-neighborhood stars versus the age τ of the stellar group: the open circles show the group's rms velocity $\sigma_{\rm r}$ in the direction to the Galactic center, and the full points

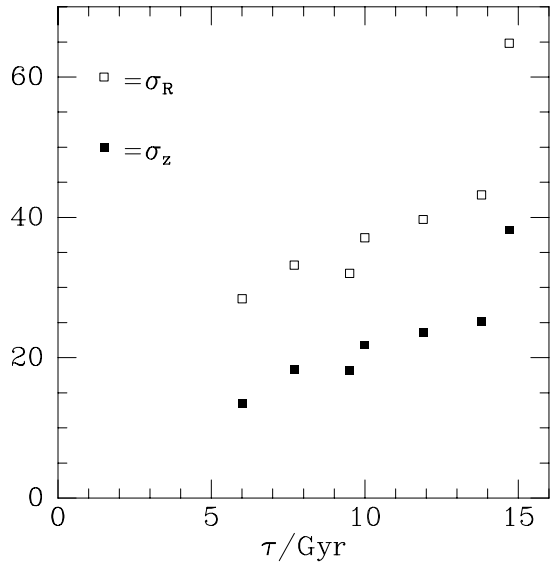


Figure 6. Velocity dispersion versus age for stars in the solar neighborhood. Data from [11].

show the rms velocity σ_z perpendicular to the Galactic plane. The unambiguous increase in rms velocity with age demonstrates that, although collisional processes are unimportant for the Milky Way, the solar neighborhood is evolving dynamically. We believe that spiral waves drive the increase in σ_r. These waves cannot be directly responsible for the increase in σ_z, however. Clouds of molecular gas with masses in the region of $10^5 M_\odot$ are thought to be responsible for this increase: the gravitational fields of these objects deflect stars that have picked up significant random motion within the plane into the direction perpendicular to the plane, so that they henceforth make significant excursions above and below the plane.

Figure 6 shows that the random velocities of stars are highly anisotropic in the sense that the random velocity perpendicular to the plane is not much larger than a half of that in the radial direction. The velocity dispersion in the tangential direction, σ_ϕ, lies between these two extremes, and there is an interesting connection between the ratio σ_ϕ/σ_r and the way the speed of a circular orbit varies with galactocentric radius.

Figure 7 shows the difference between the Sun's velocity in the direction of Galactic rotation and the mean velocity in this direction of various stellar groups as a function of the random velocity, S^2, of each stellar group. We see that the Sun is moving faster around the Galactic centre than any stellar group, and the amount by which a stellar group lags the rotation of the Sun tends to increase with the group's random velocity. The physical origin of this observation is that the hotter a stellar disk is, the more slowly it needs to spin to be in centrifugal equilibrium in a given gravitational potential; a hot stellar disk has significant pressure support. Each stellar group forms an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

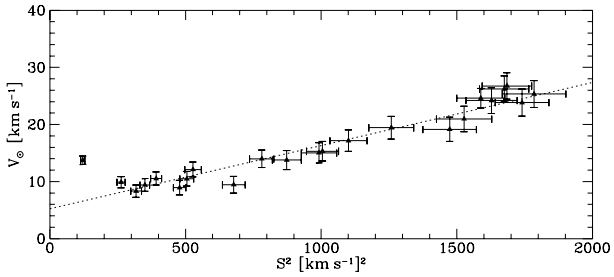


Figure 7. The velocity of the Sun in the direction of Galactic rotation relative to the mean velocity of different stellar groups. S^2 is the mean-square random velocity of each group. Data from [3].

independent disk, which spins at its own rate. The Sun happens to belong to one of the coolest, fastest-spinning disks, and within that disk happens just now to be one of the fastest-circulating stars (see also SUN AS A STAR).

When we look at stars that lie 1 or 2 kpc from the Sun, we see that the mean velocity of the stars at a given place varies from point to point around the Galaxy. This effect mainly reflects the fact that the Galactic disk is a swirling accretion disk, a sort of vortex, in which stars stream around the Galactic center at a speed that varies with radius. In fact, historically observations of this effect were a primary source of data from which the Galactic potential's run of circular speed with radius was inferred. Unfortunately, the data are not easy to interpret, in part because, as we have just seen, the speed with which a population of stars circulates depends on the population's velocity dispersion, and in part because, as we shall now see, the pattern of stellar flow within the Galaxy is in reality more complex than simple circular motion.

If we identify a group of stars, we can measure its 'velocity-dispersion tensor', which is the 3×3 matrix whose elements σ_{ij}^2 are the averages $\langle v_i v_j \rangle$ of products of velocity components. The diagonal elements of this matrix are just the squares of the velocity dispersions that we have just been discussing: $\sigma_{rr}^2 = (\sigma_r)^2$. If the Galaxy were axisymmetric, the off-diagonal elements would vanish. In fact the component $\sigma_{r\phi}^2$ is significantly non-zero for every group of solar-neighborhood stars. Responsibility for this observational result probably lies with two agents. For groups of stars that have small random velocities and therefore stay near the Sun, the non-zero values of $\sigma_{r\phi}^2$ are probably caused by spiral structure. In the case of a group of stars with large ($\sim$50 km s^{-1}) velocities relative to the Sun, the non-zero value of $\sigma_{r\phi}^2$ is probably caused by the bar at the Galactic center. The length of this bar is somewhat uncertain, but probably lies near 3 kpc, rather less than half the Sun's distance from the Galactic center, R_0. Stars with larger velocities relative to the Sun travel on their orbits to sufficiently small Galactocentric radii that they are profoundly influenced by the non-axisymmetric potential of the bar.

Figure 8 is a plot against radius r of the mass $M(r)$ contained within a sphere of radius r around the Galactic center. This plot was made by studying the kinematics of gas clouds and stars that lie near the Galactic centre and then applying simple dynamical theory. For radii greater than about a parsec, $M(r)$ decreases with r, but, at smaller r, M levels out at a value of order $2 \times 10^6 M_\odot$. The natural interpretation of this result is that a black hole of this mass sits at the Galactic center and that stars make a comparable contribution to M only at radii larger than about a parsec.

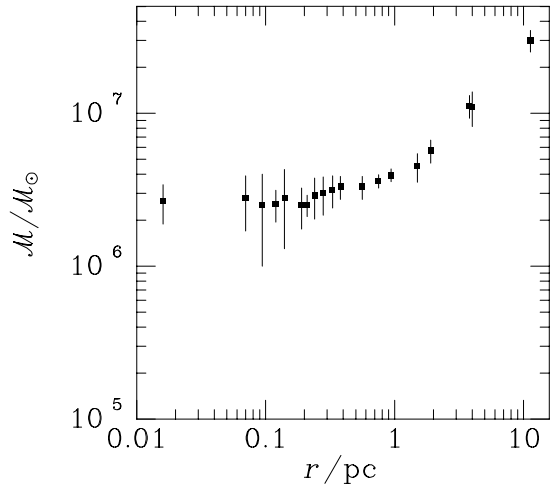


Figure 8. The mass $M(r)$ contained in a sphere of radius r around the Galactic center. Data from [4].

Bibliography

[1] Binney J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)
[2] Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)
[3] Dehnen W and Binney J J 1998 *Mon. Not. R. Astron. Soc.* **298** 387
[4] Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219
[5] Gerhard O E and Binney J J 1985 *Mon. Not. R. Astron. Soc.* **216** 467
[6] Goodman J 1987 *Astrophys. J.* **313** 576
[7] Weil M and Hernquist L 1995 *Astrophys. J.* **431** L79
[8] Norman C A, May A and van Albada T S 1985 *Astrophys. J* **296** 20
[9] Rees M J and Hut P 1992 *Mon. Not. R. Astron. Soc.* **259** 27
[10] Sellwood J A and Moore E M 1999 *Astrophys. J.* **510** 125
[11] Strömgren B 1987 *The Galaxy (NATO ASI)* ed G Gilmore and R Carswell (Dordrecht: Reidel) p 229

James Binney

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamics Explorer (Explorer 62 and 63)

Two NASA satellites launched together in August 1981 to study Earth's ionosphere, auroras and magnetosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamo Theory

Dynamo theory is that branch of MAGNETOHYDRODYNAMICS which is concerned with the self-excitation of magnetic fields in any large rotating mass of conducting fluid in motion (usually turbulent). Self-exciting dynamo action is believed to account for the very existence of magnetic fields in astrophysical systems, whether at the planetary, stellar, or galactic scale (see GEODYNAMO, DYNAMOS: SOLAR AND STELLAR).

The process of self-excitation occurs when fluid motion across a magnetic field generates the current distribution which is itself the source of the 'given' magnetic field. This principle is most simply illustrated by the example of the 'homopolar' disk dynamo sketched in figure 1: a conducting disk rotates about its axis with angular velocity Ω, and a current path between its rim and its axle is provided by the wire twisted as shown in a loop around the axle. Suppose that a current $I(t)$ flows in the loop, the current circuit being closed through the axle of the disk and the disk itself. This current generates a magnetic flux Φ across the disk, and, provided that the conductivity of the disk is not too high, this flux is given by $\Phi = MI$ where M is the mutual inductance between the loop and the rim of the disk. Rotation of the disk induces an electromotive force $E = \Omega\Phi/2\pi$ which drives the current I; the equation for $I(t)$ is then

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI = \mathcal{E} = \frac{M\Omega I}{2\pi} \tag{1}$$

where L and R are the self-inductance and resistance of the complete current circuit. The device is evidently unstable to the growth of I (and so of Φ) from an infinitesimal level if

$$\Omega > 2\pi R/M. \tag{2}$$

In this circumstance, the current grows exponentially, as does the retarding torque associated with the Lorentz force distribution in the disk. Ultimately, the disk angular velocity slows down to the critical level $\Omega_0 = 2\pi R/M$ at which the driving torque G just balances this retarding torque, and the current can remain steady. In this equilibrium, the power supplied $\Omega_0 G$ is equal to the rate of Joule dissipation RI^2 (assuming that frictional torques are negligible).

This simple system exhibits the dual property of any self-exciting dynamo: the current in the system generates the magnetic field (without the assistance of any 'external' source), and motion of a conductor across the field generates the current that maintains it. Note the following properties of the system which re-emerge in the more general fluid context, and which appear to be more or less universal concomitants of dynamo action: (1) the system exhibits 'differential rotation' in that the disk rotates, while the wire is at rest (the gradient of angular velocity is here concentrated at the two sliding contacts); (2) in order for the system to act as a dynamo, the wire must be twisted in the same sense as the sense of angular

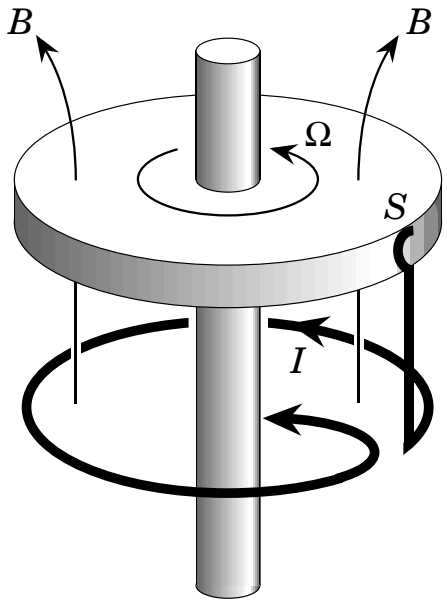


Figure 1. The self-exciting homopolar disk dynamo. Note that the wire is twisted in the same sense as that of the rotation Ω.

velocity of the disk, i.e. there is a definite 'handedness' in the system, a property which, in the fluid context, is generally encountered as a non-zero 'helicity' of the fluid flow (see MAGNETIC HELICITY).

There is a further intriguing property of this simple system, however, which indicates the dangers of too simplistic an argument when dealing with a medium of very high conductivity. If the conductivity of the disk were infinite, then we know that the flux of magnetic field across its rim must remain constant by Alfvén's theorem (magnetohydrodynamics) a result clearly incompatible with exponential growth of flux. What happens in this situation is that an additional current is induced round the rim of the disk which is just such as to maintain the constancy of the trapped flux. This perhaps indicates that, when we move to the much more difficult fluid context, great care will be needed in analysing the high-conductivity limit, which is of particular relevance in stellar and galactic contexts.

The stretch–twist–fold dynamo

Consider an initially circular magnetic flux tube of small cross-section embedded in a highly conducting incompressible fluid. We may easily imagine a doubling of the intensity of the magnetic field by a movement analogous to the doubling of an elastic band: first stretch the tube to double the radius, then twist and fold it in the manner indicated in figure 2. This process evidently doubles the net magnetic flux around the (now composite) tube; if the process is repeated over and over again, then a successive doubling, analogous to exponential growth, occurs. This is what is conventionally known as the stretch–twist–fold dynamo. It relies on the idea that, in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

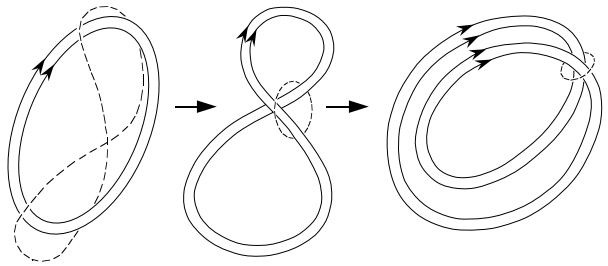


Figure 2. The stretch–twist–fold action applied to a magnetic flux tube; the reciprocal action on a linked material circuit is indicated by the dashed curve.

a perfectly conducting fluid, 'magnetic lines are frozen in the fluid', so that, in a highly conducting fluid, a similar behavior is to be expected. Although the idea is physically plausible, it is difficult to establish the existence of this type of dynamo by rigorous mathematical argument. One difficulty is that the stretch–twist–fold action necessarily introduces a twist within the tube itself (as can be easily seen by applying the twist–fold action to a ribbon in the form of a circle).

If we trace backwards in time the effect of the above distortion on a Lagrangian material circuit round the doubled flux tube (the dashed curve in figure 2), then it is easy to see that this curve must have started as a twisted curve embracing the circular tube twice. Hence the stretch–twist–fold action has the reverse effect on this Lagrangian circuit. An alternative magnetic flux tube, starting in the position of the dashed curve, would thus be diminished by the same distortion. Things are evidently not as simple as they might appear at first sight.

Formal definition of dynamo action

We start from the magnetic induction equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{\nabla} \wedge (\boldsymbol{u} \wedge \boldsymbol{B}) + \eta \nabla^2 \boldsymbol{B} \tag{3}$$

where $\boldsymbol{B}$ is the magnetic field, $\boldsymbol{u}$ the velocity field and η the magnetic diffusivity of the fluid. $\boldsymbol{B}$ also satisfies the supplementary condition $\mathrm{div}\boldsymbol{B} = 0$. If the velocity field $\boldsymbol{u}$ is steady (i.e. $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$), then we may look for solutions of equation (3) in the form of 'eigenmodes',

$$\boldsymbol{B}(\boldsymbol{x}, t) = \mathrm{Re}\, \hat{\boldsymbol{B}}(\boldsymbol{x}) e^{pt}. \tag{4}$$

The amplitude function $\hat{\boldsymbol{B}}(\boldsymbol{x})$ then satisfies the equation

$$p\hat{\boldsymbol{B}} = \boldsymbol{\nabla} \wedge (\boldsymbol{u} \wedge \hat{\boldsymbol{B}}) + \eta\, \nabla^2 \hat{\boldsymbol{B}} \tag{5}$$

which, when coupled with appropriate boundary conditions, defines an eigenvalue problem for the exponent p, the field $\hat{\boldsymbol{B}}$ being then the corresponding eigenfunction. p may in principle be real or complex: $p = p_r + \mathrm{i}\, p_i$. If $p_r > 0$ then the corresponding solution of equation (3) exhibits dynamo action: there is evidently a systematic increase in intensity of the magnetic field with growth rate p_r. If $p_i = 0$, the dynamo action is a simple exponential growth, while if p_i is non-zero, then the field at any point oscillates with exponentially growing amplitude.

Solution of the above eigenvalue problem evidently depends on the parameter η which appears in equation (5). Let us suppose that the velocity field is characterized by a length scale l and a velocity scale u_0, so that the magnetic Reynolds number $R_m = u_0 l/\eta$ may be defined. In astrophysical contexts, we are usually concerned with the situation in which R_m is very large, and particular attention must therefore focus on the behavior of p and $\hat{\boldsymbol{B}}$ as R_m tends to infinity. This consideration has led to a formal distinction between 'slow' and 'fast' dynamos. Roughly speaking, the dynamo is fast if p_r is proportional to u_0/l (i.e. independent of η) in the limit $R_m \to \infty$. If, on the other hand, $p_r l/u_0 \to 0$ as $R_m \to \infty$, then the dynamo is slow.

The stretch–twist–fold dynamo considered above is perhaps the best candidate for characterization as a fast dynamo; in a single period of the stretch–twist–fold cycle, it exhibits a doubling of flux, but at the cost of introducing a certain degree of 'fine structure' in the field. Actually, this appears to be a property of any fast dynamo. Equation (5) can be easily analyzed in the formal limit $\eta = 0$. It has been shown that a fast dynamo can exist only if the corresponding field $\hat{\boldsymbol{B}}$ exhibits fine structure on scales of order $R_m^{-1/2}$. In the limit $\eta = 0$, such fields become non-differentiable.

Attempts have been made to circumvent this difficulty by widening the class of solutions of equation (3) to which the term 'fast dynamo' may be ascribed. The reader is referred to the book by Childress and Gilbert (1995) for an account of investigations, based largely on Lagrangian and associated mapping techniques, in this area.

We may note that the homopolar disk dynamo is, for the reasons indicated in the introduction, definitely a slow dynamo, i.e. its growth rate tends to zero as the magnetic Reynolds number, based on disk angular velocity and conductivity, tends to infinity.

Cowling's theorem

In planetary and stellar contexts, it was natural that early investigations should focus on axisymmetric systems. In 1934, COWLING established that axisymmetric dynamo action is impossible, a result which, for many years, falsely led people to assume that all dynamo action was impossible (see Cowling 1957). It is important, nevertheless, to understand the basic nature of Cowling's argument, if only to see how the associated difficulties may be circumvented. Consider an axisymmetric magnetic field generated by currents contained entirely within a sphere (figure 3). The 'toroidal' currents within the sphere (i.e. those in the ϕ direction in a system of cylindrical polar coordinates r, ϕ, z) act as the source of the 'poloidal' part of the magnetic field $\boldsymbol{B}_P$ (i.e. that part with components in the r and z direction). This poloidal field, being associated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

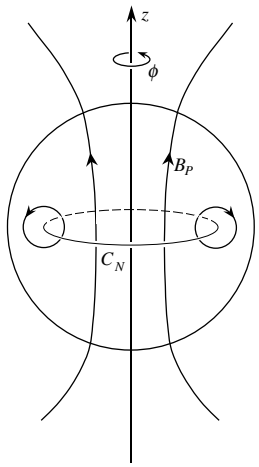


Figure 3. Cowling's theorem: induction cannot maintain current in the neighborhood of the circle C_N if the flow and field are axisymmetric.

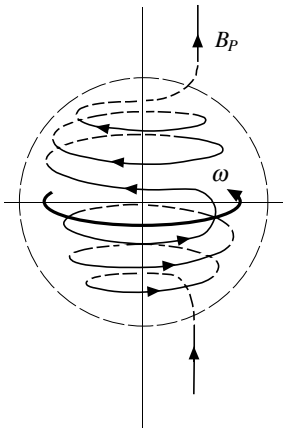


Figure 4. Generation of toroidal field by differential rotation with a variation of angular velocity along each field line.

with a compact source, has at most a dipole structure at a great distance from the sphere. The Earth's magnetic field of course provides an excellent example of this phenomenon. The lines of force of this dipole field must have at least one 'O-type' neutral point, and this neutral point must lie inside the sphere of conducting fluid. There is an associated circle C_N of neutral points around the axis of symmetry. The toroidal current distribution can be driven only by fluid motion across the poloidal magnetic field and this effect is absent on C_N where $\boldsymbol{B}_P = 0$. Consequently any toroidal current on C_N must necessarily decay through normal Ohmic diffusion associated with finite conductivity ($\eta > 0$). Thus, steady maintenance of an axisymmetric magnetic field in a finite volume of conducting fluid is impossible.

This argument of course needs to be formalized, but this can be done, and indeed many extensions of Cowling's theorem have been proved. Essentially, it is well established that, if a steady magnetic field is to be maintained in a finite region containing fluid of finite conductivity, then the field, and the motions that support it, must be non-axisymmetric. This theorem accounts for much of the mathematical difficulty of the subject and for the long period that elapsed following the proof of the theorem before explicit examples of dynamo action in a homogeneous fluid were discovered.

The generation of toroidal field by differential rotation

One mechanism, which is essentially axisymmetric in character, is nevertheless of great importance in the dynamo context. Let us suppose for the moment that a poloidal magnetic field is maintained by some (as yet unspecified) mechanism. Consider the action of a purely toroidal velocity field on this magnetic field. If the angular velocity is constant along every field line, then these field lines are simply rotated around the axis of symmetry with the corresponding angular velocity (the law of 'isorotation'). If, on the other hand, there is any gradient of angular velocity along a field line, then that field line will tend to be drawn out in the toroidal direction, an effect that will be limited only by diffusion effects (figure 4). If the magnetic Reynolds number is large, then the field line may be 'cranked' round the axis many times before this diffusion effect intervenes. Detailed analysis based on equation (3) indicates that the toroidal field that is generated increases until it is $O(R_m)$ times the initial poloidal field. At this stage, in general, Ohmic diffusion establishes a steady state.

This is evidently a powerful mechanism whereby toroidal fields may be generated from a pre-existing poloidal field. The great difficulty is to explain how it is that the poloidal field can be generated in the first place. For this purpose, it is absolutely necessary to consider non-axisymmetric effects. The greatest impetus to this subject has undoubtedly come from the assumption at the outset that the fluid motion is fully turbulent, an assumption that is of course utterly reasonable in stellar and galactic contexts. Even in the case of planetary interiors, although the fluid motions are extremely weak (of the order of fractions of a millimeter per second) it seems highly likely that these motions, while not necessarily fully turbulent, will nevertheless have a significant non-axisymmetric ingredient.

Mean field electrodynamics

The idea behind mean field electrodynamics is that we are generally concerned in dynamo theory with the development of a 'large-scale' magnetic field, i.e. with a field that is coherent over the global scale of the planet or star or galaxy under consideration. The turbulence, or other random motion, which plays a part in the dynamo process in general has a very much smaller characteristic length scale. There is therefore a significant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scale separation which can be exploited in the construction of a theory for the large-scale field.

Let us first suppose that, in some region of fluid which contains many constituent eddies of the turbulence, the large-scale ingredient of the field may be treated as nearly uniform. The velocity $\boldsymbol{u}$ distorts the mean field $\boldsymbol{B}_0$, generating a fluctuating component $\boldsymbol{b}$. If we average equation (3) over the scale l of the turbulent fluctuations, then we obtain an equation for the mean field $\boldsymbol{B}_0$, namely

$$\frac{\partial \boldsymbol{B}_0}{\partial t} = \boldsymbol{\nabla} \wedge \boldsymbol{\mathcal{E}} + \eta \nabla^2 \boldsymbol{B}_0. \tag{6}$$

Here, a term appears resulting from the interaction between two fluctuating fields, namely

$$\boldsymbol{\mathcal{E}} = \langle \boldsymbol{u} \wedge \boldsymbol{b} \rangle. \tag{7}$$

This is the 'mean electromotive force' associated with the fluid motion. We have assumed that the average of $\boldsymbol{u}$ itself and similarly the average of $\boldsymbol{b}$ are zero. Nevertheless, any correlation between $\boldsymbol{u}$ and $\boldsymbol{b}$ will result in a non-zero value for $\boldsymbol{\mathcal{E}}$.

In order to proceed further, it is necessary to obtain an equation for the fluctuating field $\boldsymbol{b}$. This is obtained by simply subtracting equation (6) from equation (3) to give

$$\frac{\partial \boldsymbol{b}}{\partial t} = \boldsymbol{\nabla} \wedge (\boldsymbol{u} \wedge \boldsymbol{B}_0) + \boldsymbol{\nabla} \wedge (\boldsymbol{u} \wedge \boldsymbol{b} - \langle \boldsymbol{u} \wedge \boldsymbol{b} \rangle) + \eta \nabla^2 \boldsymbol{b}. \tag{8}$$

This equation evidently establishes a linear relationship between $\boldsymbol{b}$ and $\boldsymbol{B}_0$, and hence, via equation (7), between $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{B}_0$. Since $\boldsymbol{B}_0$ is nearly uniform, this linear relationship may be expressed in the form

$$\mathcal{E}_i = \alpha_{ij} B_{0j} + \beta_{ijk} \frac{\partial B_{0j}}{\partial x_k} + \cdots \tag{9}$$

where the coefficients α_{ij}, β_{ijk}, and so on, do not depend on the field $\boldsymbol{B}_0$ but are determined (in principle) solely by the statistical properties of the turbulence $\boldsymbol{u}$ and the magnetic diffusivity η.

The implications of equation (9) are most easily understood if it is assumed that the background turbulence is isotropic (i.e. there is no 'preferred direction'), so that the tensor coefficients α_{ij}, β_{ijk}, etc. are also isotropic. In this case, we may write

$$\alpha_{ij} = \alpha \delta_{ij} \qquad \beta_{ijk} = \beta \epsilon_{ijk} \tag{10}$$

and the corresponding expression for $\boldsymbol{\mathcal{E}}$ is then given by

$$\mathcal{E} = \alpha \boldsymbol{B}_0 - \beta \, \boldsymbol{\nabla} \wedge \boldsymbol{B}_0 + \ldots. \tag{11}$$

The corresponding form of equation (6) then becomes

$$\frac{\partial \boldsymbol{B}_0}{\partial t} = \alpha \nabla \wedge \boldsymbol{B}_0 + (\eta + \beta) \nabla^2 \boldsymbol{B}_0. \tag{12}$$

There is here a very important effect to notice, an effect first described in a seminal paper by Steenbeck, Krause and Rädler in 1966 (see Krause and Rädler 1980). Since $\boldsymbol{\mathcal{E}}$ is a polar vector, whereas the magnetic field $\boldsymbol{B}_0$ is an axial vector, the coefficient α in equation (11) must be a pseudo-scalar, i.e. a quantity which changes sign under change from a right-handed to a left-handed frame of reference. By contrast, the coefficient β is a pure scalar, independent of frame of reference. It is evident from equation (12) that β simply serves to augment the molecular diffusivity of the fluid, i.e. we can describe β as a turbulent diffusivity. By contrast, the coefficient α is associated with a term of novel structure in the induction equation; moreover, for a field of very large length scale, it is this term that dominates the right-hand side. The effect, appropriately described as the 'α effect', is crucial to an understanding of dynamo theory.

Dynamo instability associated with the α effect

To understand this effect in the simplest possible way, it is sufficient to consider a field structure $\boldsymbol{B}_0(x)$ satisfying the Beltrami condition

$$\boldsymbol{\nabla} \wedge \boldsymbol{B}_0 = K \boldsymbol{B}_0 \tag{13}$$

where K is a constant (positive or negative). The ABC field

$$\boldsymbol{B}_0 = \begin{pmatrix} B \cos Ky + C \sin Kz \\ C \cos Kz + A \sin Kx \\ A \cos Kx + B \sin Ky \end{pmatrix} \tag{14}$$

provides an example of such a field, as may be easily verified, but there are many others (see FORCE-FREE MAGNETIC FIELDS). Since, for such a field, $\nabla^2 \boldsymbol{B}_0 = -K^2 \boldsymbol{B}_0$, equation (11) simplifies to the form

$$\frac{\partial \boldsymbol{B}_0}{\partial t} = \alpha K \boldsymbol{B}_0 - (\eta + \beta) K^2 \boldsymbol{B}_0. \tag{15}$$

Hence $\boldsymbol{B}_0$ is proportional to e^{pt}, where

$$p = \alpha K - (\eta + \beta) K^2 \tag{16}$$

and dynamo instability occurs if $p > 0$, a condition that is always satisfied if $\alpha K > 0$ and $|K|$ is sufficiently small, i.e. provided that the scale $L = 2\pi/|K|$ is sufficiently large. This condition is certainly consistent with the initial assumption that the scale of the mean field be large compared with the scale of the turbulence. Thus, in the presence of a non-zero α effect, the medium is always unstable to the growth of magnetic field perturbations on a sufficiently large length scale.

This effect is one of the most remarkable discoveries of the last 50 years. It indicates how order, represented by the large-scale magnetic field, may emerge from chaos, represented by the small-scale turbulence in the medium.

Evaluation of α

It remains, however, to develop a convincing theory for the determination of the all-important parameter α in terms of the statistical properties of the background turbulence. We have already noted that α is a pseudo-scalar; it can therefore be non-zero only if the statistical properties of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

turbulence have a crucial property of 'lack of reflectional symmetry'. The simplest measure of a lack of reflectional symmetry in a turbulent flow is its mean helicity

$$\mathcal{H} = \langle \boldsymbol{u} \cdot \operatorname{curl} \boldsymbol{u} \rangle \tag{17}$$

the mean of the scalar product of the velocity field and the vorticity field curl $\boldsymbol{u}$. This is analogous to magnetic helicity. The helicity is related to the topological structure of the vorticity field, which is conserved in the classical situation in which vortex lines are frozen in the fluid. It may be expected that, if this mean helicity is non-zero, then α will in general be non-zero also.

It is indeed possible to calculate α explicitly in any circumstances in which the fluctuating ingredient of the magnetic field $\boldsymbol{b}$ is weak compared with the mean field. In such circumstances, it is found that

$$\alpha = -\frac{1}{3}\eta \int\int \frac{k^2 H(k,\omega)}{\omega^2 + \eta^2 k^4}\, \mathrm{d}k\, \mathrm{d}\omega \tag{18}$$

where $H(k,\omega)$ is the helicity spectrum of the turbulence, satisfying the defining relationship

$$\langle \boldsymbol{u} \cdot \operatorname{curl} \boldsymbol{u} \rangle = \int\int H(k,\omega)\, \mathrm{d}k\, \mathrm{d}\omega. \tag{19}$$

Thus, α is linearly related to the helicity spectrum of the turbulence. In these integral expressions, k is a wavenumber magnitude and ω a frequency. Note that, in the limit as $\eta \to 0$, the expression (18) for α depends critically on the behavior of the helicity spectrum for small values of frequency ω. Unfortunately, it is in this double limit ($\eta \to 0$, $\omega \to 0$) that the theory yielding the expression (18) is at its least convincing.

Physical mechanism of the dynamo process

The dynamo process associated with the α effect may be represented schematically as in figure 5. Background turbulence with helicity provides an α effect, the dynamo parameter being related to the helicity through equation (18) (or some higher-order version of this equation). This α effect generates an electromotive force (EMF) $\boldsymbol{\mathcal{E}}$ parallel to $\boldsymbol{B}_0$ (through the equation $\boldsymbol{\mathcal{E}} = \alpha \boldsymbol{B}_0 + \cdots$) and this EMF drives a mean current $\boldsymbol{J}_0$ parallel to $\boldsymbol{B}_0$. The resulting mean field has non-zero magnetic helicity, and so the tubes of force of $\boldsymbol{B}_0$ are generally linked. The prototype linkage is shown in figure 5; the current driven round each tube provides (via Ampère's theorem) the mean magnetic field in the other tube.

Three vital conditions ensure the success of this process:

- the turbulence must have non-zero helicity;
- the fluid must have finite non-zero magnetic diffusivity;
- the fluid domain must be large enough to provide space for the growth of the large-scale magnetic field.

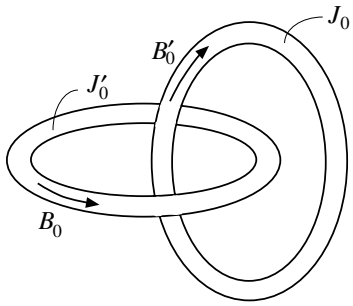


Figure 5. The dynamo cycle associated with the α effect. The current J_0 is associated with the field B_0, and the current J_0' with B_0'

The α–ω dynamo mechanism

As we have seen above, the α effect in isolation can easily give rise to dynamo instability; we have shown this to be true in a medium of infinite extent; however, a similar mechanism certainly occurs in any sufficiently large spherical domain of conducting fluid in turbulent motion.

If the fluid in such a domain is rotating with mean angular velocity Ω, then turbulent convection within the domain in general leads to a state of 'differential rotation' in which the inner regions rotate more rapidly than the outer regions. As we have seen, differential rotation provides a powerful mechanism whereby toroidal field can be generated from poloidal field. The α effect provides the vital complementary mechanism whereby poloidal field can be regenerated from toroidal field. If the turbulence is isotropic, then the same α effect contributes to the regeneration of toroidal field from poloidal field, but the differential rotation effect is generally much more powerful.

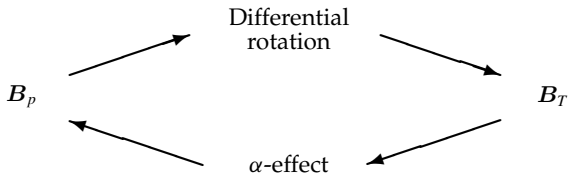


Figure 6. Field regeneration through the α–ω dynamo cycle.

These considerations lead to consideration of dynamos of 'α–ω type' as indicated schematically in figure 6. Here, the dynamo cycle consists simply of two stages: generation of toroidal from poloidal field by differential rotation and regeneration of poloidal from toroidal field by the α effect. Many numerical investigations of the resulting mean field dynamo equations have demonstrated the efficiency of this type of self-exciting dynamo.

Back-reaction of Lorentz forces

The above outline description of the process of dynamo action is purely 'kinematic', in that it assumes that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

velocity field, or in the case of turbulent flow the statistics of the velocity field, is prescribed independently of the magnetic field. In circumstances in which dynamo action occurs, the amplitude of the magnetic field grows exponentially in time, and so therefore does the Lorentz force, which is quadratically related to the magnetic field. This Lorentz force reacts back upon the fluid motion so that the velocity field itself depends on the magnetic field that it generates. The dynamo process then enters a nonlinear regime, whose complexities can generally be unravelled only through numerical simulation of the full magnetohydrodynamic equations.

It is also necessary to specify the nature of the forces that sustain the motion, and the associated dynamo, against dissipative effects which now include viscosity as well as Joule dissipation. This is equivalent to specifying the source of energy of the fluid motion; this source may be of thermal origin (as in the solar context); in the context of the terrestrial dynamo driven by fluid motion in the liquid core the source of energy is more likely to be associated with segregation of light material during the process of solidification of the inner core, a process that provides buoyancy forces of 'compositional' rather than thermal origin. Either way, the need to include consideration of the buoyancy field adds significantly to both the physical and the mathematical complexity of the problem.

The study of fully non-linear dynamos, taking account of the diverse possible sources of energy, is an area of very vigorous current research, and great advances are to be expected over the next 10–20 yr.

Bibliography

Childress S and Gilbert A D 1995 *Stretch, Twist, Fold: the Fast Dynamo* (Berlin: Springer)

Cowling T G 1957 *Magnetohydrodynamics* (New York: Interscience)

Krause F and Rädler K-H 1980 *Mean Field Magnetohydrodynamics and Dynamo Theory* (Oxford: Pergamon)

Moffatt H K 1978 *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge: Cambridge University Press)

Parker E N 1979 *Cosmical Magnetic Fields* (Oxford: Clarendon)

Zeldovich Ya B, Ruzmaikin A A and Sokoloff D D 1983 Magnetic Fields in Astrophysics (London: Gordon and Breach)

Keith Moffatt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamos: Solar and Stellar

The direct observation of magnetic fields on the surface of the Sun (see SOLAR MAGNETIC FIELD), as well as on the surface of other stars (see MAGNETIC FIELDS IN STARS), directly raises the question of the origins of these magnetic fields. The simplest hypothesis is that these magnetic fields are simply the remnants of magnetic fields carried in by the accreting gas during the formation phase of such stars, but a number of observational facts make this explanation untenable for stars such as the Sun: the observed fields are globally relatively weak, but locally very strong, much stronger than would be expected of decaying remnant fields dating back to the formation of the Sun, and the global magnetic fields are observed to vary systematically in a roughly periodic fashion (in the Sun's case, this periodicity is roughly 22 yr (see SOLAR ACTIVITY INDICES)). These observations instead argue for the presence of a process which is capable of re-generating the magnetic field of a star such as the Sun on a time scale of roughly 11 yr; this is the process referred to in the following as a (stellar magnetic) dynamo (see DYNAMO THEORY).

At the outset, it is important to distinguish between the large-scale (i.e. stellar-scale) and small-scale behavior of stellar magnetic fields, which can be drastically different. In the Sun's case—the only case for which we have detailed, spatially resolved observations—it is the large-scale fields that show the cyclic, activity-related temporal variations; these large-scale fields are weak (1–10 G), in the sense that their energy density is substantially smaller than that associated with either the thermal energy content of the Sun's surface layers or the kinetic energy density of these same layers. In contrast, the small-scale solar fields, i.e. the fields on spatial scales comparable with the convective surface motions, can be strong (up to 3000 G), in the sense that their energy density can be of the order of, or even larger than, the kinetic energy density of the surface layers and is of the order of the plasma internal energy density of these layers. These locally strong fields show a much more complex temporal behavior, much of which is not directly tied to the cyclic behavior of the large-scale fields (see SOLAR PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS). This temporal distinction between the behavior of large-scale and small-scale magnetic fields is a crucial argument supporting the current view that the origins of the large- and small-scale fields may be unrelated.

Mean field solar dynamos

As discussed elsewhere (see DYNAMO THEORY), the challenge for solar and stellar dynamo theory was originally to construct a physically plausible mechanism for re-generating both the poloidal and the toroidal components of the Sun's relatively weak large-scale magnetic field. (It is important to note that the importance of small-scale solar magnetic fields was not appreciated at the time, primarily because direct measurements of these fields were not possible.)

Since the 1940s, it has been recognized that differential rotation in the interior or surface layers of the Sun could generate toroidal magnetic flux from a pre-existing poloidal magnetic component; the missing link was a similar physical process re-generating the poloidal component from the toroidal one. This link was found in the mid-1950s by E N Parker. He showed that motions of conducting PLASMA associated with MAGNETIC BUOYANCY and convection, which carry (toroidal) magnetic flux radially outward, will be subjected to Coriolis forces if the star rotates; this rotation-dependent force twists the rising and falling plasma bubbles precisely the right way so that the trapped toroidal fields produce a poloidal component. With the addition of a phenomenological model for an enhanced eddy diffusivity, Parker was able to construct a solar magnetic dynamo model which reproduced all of the classical features of large-scale solar magnetic activity: the 22 yr periodic variation of the toroidal and poloidal fields, the butterfly diagram of magnetic field emergence, showing flux emergence closer and closer to the solar equator as one half of a magnetic cycle proceeds, and the Hale polarity law of local flux emergence, which states that the leading polarity of emerging bipolar fields always lies closer to the solar equator.

In Parker's original model, the differential rotation needed to re-generate the large-scale toroidal field was assumed to be the result of radial differential rotation, i.e. Parker assumed that the Sun rotates more quickly in its interior than in its surface layers; at that time, this differential rotation was unobserved, and therefore unconstrained. Later workers such as H Babcock produced variants of Parker's model which took advantage of the observed (surface) DIFFERENTIAL ROTATION in latitude and were able to produce dynamo models which were similar in successfully describing the large-scale features of the Sun's magnetic field. Additional work was carried out in studying the evolution of the emerged magnetic fields. Thus, R Leighton showed that the asymmetry of magnetic flux emergence associated with Hale's polarity law, when combined with a simple random walk model of flux diffusion on the solar surface (consistent with Parker's eddy diffusion model), naturally leads to a migration of surface magnetic fields to the solar poles in which the following polarity dominates at the poles. In this picture, the observed small-scale magnetic fields were viewed as detritus of the decaying large-scale fields and as the seeds for the next generation of large-scale poloidal fields.

By the mid-1970s, magnetic dynamo models of the type first described by E N Parker were well understood from both the phenomenological and the formal, mathematical perspectives: the Potsdam group of dynamo theorists, as well as the work of K Moffatt and collaborators in the UK, and Parker and collaborators in the US, had firmly established the foundations of mean field theory in MAGNETOHYDRODYNAMICS. However, two advances in the 1970s and early 1980s led to a questioning of mean field theory. First, high spatial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

resolution observations of surface solar magnetic fields by J Stenflo, J Harvey and others began to raise serious questions about the applicability of mean field theory to the Sun—observations indicated that while the large-scale fields were indeed weak, this was not true of the small-scale fields. In mean field theory, it is assumed that the small-scale fields are weak, in the sense that they are not strong enough to react back on the fluid motions; in contrast, direct measurements showed that these small-scale fields were, in contradiction, strong enough to affect fluid motions. These back reactions were capable, in principle, of destroying the underpinnings of mean field theory; for example, it is by no means clear that the large eddy diffusion required by mean field theory can exist if the small-scale fields significantly affect small-scale fluid motions. In addition, computing advances in the early 1980s had reached the stage that it became possible to solve the fully nonlinear magnetohydrodynamic equations for a convective spherical shell, albeit for very limited spatial resolution, and such simulations by P Gilman and collaborators showed that the large-scale behavior of the numerical solutions did not correspond very closely to either mean field theory or observations of the Sun's magnetic cycle.

Recent work

Much of the recent work on solar and stellar dynamos has grown out of the concerns raised about mean field dynamo models, as well as new observational and computational tools. For example, the realization that small-scale magnetic fields may play an important role in the solar dynamo led to the exploration of new classes of dynamo models for the large-scale magnetic fields by M Schüssler and collaborators. These models take as their starting point the notion that the magnetic field in the solar interior is in a highly filamentary, flux tube, form (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES). The exploration of the consequences of such a dynamo model, including studies of the dynamics of rising magnetic flux tubes in the convection zone (see SOLAR INTERIOR: CONVECTION ZONE), is a very active area of research today.

The advent of solar seismology (HELIOSEISMOLOGY) (see HELIOSEISMIC OBSERVATIONS) allows us to focus on another aspect of solar dynamo theory, namely the interior rotation rate of the Sun, and the precise location of the dynamo. Concerns about the precise location of the solar dynamo were raised as early as the late 1970s, when it was suggested by N O Weiss, E Spiegel, R Rosner and others that the Sun's toroidal fields might be plausibly produced in the subadiabatic undershoot layer lying beneath the solar convection zone, where these fields could be stably stored. This idea has received much recent attention because helioseismological observations have indicated that there is a layer of substantial differential rotation (the 'tachycline') lying at the base of the solar convection zone (see SOLAR INTERIOR: ROTATION), i.e. precisely in the region where the earlier models had placed the toroidal field production region. This area of research is also currently very active.

In the early 1970s, S Vainshtein and Y Zel'dovich demonstrated the existence of a new class of dynamos—the so-called 'fast dynamos'—whose properties did not depend on the molecular diffusivity in the limit of very large magnetic Reynolds numbers. This type of dynamo has the property that rotation is not essential to its existence, and it is thought that the maximal spatial scale of the magnetic fields produced by such a dynamo is comparable with the maximal spatial scale of the driving turbulent motions (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION). At the time, the connection between such dynamos and specific astrophysical magnetic fields was still unclear. Recent advances in computing power now allow direct numerical simulations of the fully nonlinear magnetohydrodynamic equations that have far greater spatial dynamic range than what could be done in the 1980s; as a result, it is now feasible to carry out detailed numerical experiments of fast dynamos. Using flows which are variously driven by specified external random 'stirring' or by imposed thermal fluxes (thus driving thermal convective flows), F Cattaneo, D Hughes and collaborators have been able to show that magnetic fields which are locally in equipartition are readily produced by such flows. The volume filling factor of the strong fields depends on the precise mixing characteristics of the flows themselves; and the morphology of the surface flows and surface magnetic flux in the convectively driven calculations uncannily resembles recent observations of solar magnetic fields and surface convective motions obtained from the MDI instrument on board the SOHO spacecraft. These calculations—which show that the magnetic field in chaotic flows has a highly fractal spatial structure—also call into question the existence of magnetic 'flux tubes' as discrete, relatively long-lived entities. This area of research is also extremely active today.

Stellar dynamo problems

Quite aside from the precise mechanism by which stellar magnetic fields are produced, we also do not have a good understanding of a number of additional aspects of stellar magnetic activity.

Magnetic activity at the end of the main sequence

Surveys of the nearby stars have shown that stellar activity as measured by the x-ray to bolometric luminosity ratio does not change significantly as one proceeds down the main sequence to low-mass stars (see SOLAR–STELLAR CONNECTION and STELLAR ACTIVITY). However, surveys of very low mass stars (e.g. stars that may be, or are, classified as brown dwarfs) suggest that some of them may well be inactive (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS). Although the lack of evidence is not evidence for absence, it is known that some sort of transition from the above-described convection-driven activity should occur at sufficiently low mass: for example, we know that the observed activity of the giant gaseous planets in our solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

system is not solar like at all, but is instead the result of the interaction of planetary magnetospheres with the solar wind (see PLANETARY MAGNETOSPHERES); for gaseous planets, magnetic field 'stirring' at the planetary surfaces by surface flows is a negligible source for their activity, while such stirring by photospheric motions is precisely how we believe stellar chromospheres and coronae are energized. Thus, observations of both the Sun and stars, as well as planets in our solar system, allow us to conclude that rotating, self-gravitating objects (stars and giant gaseous planets) produce magnetic fields in their interior via a dynamo process, that these objects are all 'active' and that the underlying basic physics that drives the activity in these two cases is may be drastically different in its origins. However, exactly how this transition between solar-like behavior and planetary behavior occurs is not yet understood.

A related issue revolves around the nature of convection, and the nature of the interior dynamo, as one proceeds to lower- and lower-mass objects. Classical stellar structure theory predicts that stars become fully convective at spectral type ∼dM5 (see HERTZSPRUNG–RUSSELL DIAGRAM and STELLAR INTERIORS), and, based on recent ideas that the base of the convection zone should play a particularly important role in magnetic dynamos, one would expect to see a transition in activity behavior at ∼dM5—yet no such transition has been observed. How can we understand this? It may be possible that, in the presence of even relatively modest interior magnetic fields, low-mass stars never become fully convective, i.e. that magnetic fields stabilize the deep interior in the mean, but a definitive answer remains elusive at the time of writing.

Magnetic A stars

Observations of late spectral type B–A main sequence stars have established the existence of a subclass of such stars that show the presence of extremely strong surface magnetic fields, with reports of extreme examples of fields with strength in excess of 30 kG. Temporal variations of the polarized Zeeman signal from these stars show only variations related to stellar rotation; secular variations that might be interpreted as cycle-related changes have not been observed from these stars. The implication is either that these magnetic fields are primordial in origin or that the dynamo process responsible for these fields has a substantially longer characteristic time scale than stellar dynamos associated with the surface convection zone of late-type main sequence stars.

Spot stars

Observations of main sequence stars of spectral type K and M have shown that the more active stars of this type can display substantial bolometric luminosity variations that have been attributed to the presence of large 'star spots' on their surface (see RED DWARF STARS). While solar sunspots typically occupy areas of order 10^{-6}–10^{-5} of the visible solar disk, these starspots can occupy 30% or more of the visible surface of the active star. While there has been a revolution in the study of such stellar surface features with the introduction of Doppler imaging, little is understood theoretically about the formation of such spots. For example, the type of local fast dynamo process discussed earlier, which gives good account of the small-scale magnetic flux production on the solar surface, cannot be trivially extended to explain these enormous magnetic features.

Bibliography

Rosner R 2000 Magnetic fields of stars: using stars as tools for understanding the origins of cosmic magnetic fields *Phil. Trans. R. Soc.* **358** 689–709

Robert Rosner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dyson, Sir Frank Watson (1868–1939)

Astronomer, born in Measham, Leicestershire, England, Astronomer Royal for Scotland and afterwards for England. Directed (with SIR ARTHUR STANLEY EDDINGTON) the 1919 eclipse expedition which confirmed the deflection of starlight by the Sun's gravity, as predicted by Einstein's general theory of relativity. Directed measurements of terrestrial magnetism, latitude and time at the Royal Observatory and initiated the radio broadcast of time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eötvös, Baron Lóránd [Roland] von (1848–1919)

Hungarian physicist, born in Pest (now part of Budapest), became professor of experimental physics there. Worked on wide range of physical problems including gravitation, and invented the *Eötvös balance*, a torsion balance. With it, he tested (in what became known as the *Eötvös experiment*) the equivalence principle that gravitational mass and inertial mass are equivalent; he found that they were identical to less than five parts in a billion. The equivalence principle was originally due to GALILEO and given prominence by EINSTEIN in the theory of general relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eagle Nebula (IC 4703)

An emission nebula in the constellation Serpens, position RA $18^{h}\ 18.8^{m}$, dec. $-13^{\circ}\ 47'$. It measures about $30'$ across and surrounds the open cluster M16 (NGC 6611); it is often mistakenly given the cluster's designation. Its shape is vaguely reminiscent of an eagle in flight, though some dark nebulosity near its center bears a closer resemblance to an eagle. The nebula is the site of star formation, especially in huge twisting columns of dark nebulosity nicknamed 'elephant trunks' or 'the pillars of creation'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth: Space-based Observations

In the wake of World War II, the US Naval Research Laboratory began experimenting with German-designed V-2 rockets as well as smaller, lighter launch vehicles called 'sounding rockets'. In order to monitor which way the rockets were pointing during their flights, scientists began putting aircraft gun cameras on them. During one of these flights, on 7 March 1947, the first space-based picture of Earth was taken at an altitude of 100 miles over New Mexico. In 1950, upon developing the film after another of these flights, scientist Otto Berg discovered a series of pictures of a huge tropical storm over Brownsville, Texas. He pieced the pictures together into a mosaic of the region enveloped by the storm, thus demonstrating the potential for space-based cameras to help us monitor our changing world (see also ROCKETS IN ASTRONOMY).

Interest quickly grew among the international community of Earth scientists in using sounding rockets to study the Earth's upper atmosphere. In 1952, the International Council of Scientific Unions proposed declaring 1957 the 'International Geophysical Year' (IGY) to satisfy a growing desire among scientists to expand the study of our home planet to include the whole Earth system and its surroundings. In 1954, the IGY committee challenged participating countries to launch a satellite as part of the IGY activities. On 29 July 1955, President Dwight Eisenhower announced that the United States would launch about six Earth satellites as a contribution to the IGY. On the following day, Premier Nikita Kruschev said the Soviet Union would also launch satellites during the IGY, and the race was on between the superpowers.

In October 1957, the Soviet Union successfully launched Sputnik I—a 22-inch spherical satellite weighing 175 pounds—into a low-Earth orbit. In January 1958, the United States successfully launched Explorer I, followed by the launch of Vanguard I two months after that. But the Russians' success and the United States' initial failures spread a sense of crisis in the latter country. The Americans declared space exploration a national priority, and in 1958 President Eisenhower asked his Science Advisory Committee to write a policy on space. Later that year Congress signed the act creating the National Aeronautics and Space Administration (NASA). Since then, NASA has played an international leadership role in developing and deploying satellite technologies for Earth observation.

A space-based perspective

The Soviet–American race to demonstrate technological superiority yielded a fortunate byproduct for Earth scientists—a newfound perspective from which to study our world. In 1959 the Russians, again followed closely by the Americans, launched the first humans into space. Looking back, the cosmonauts and astronauts saw the Earth as a tiny oasis of life—unique, as far as we know—adrift in a seemingly lifeless sea. Against the cold, black backdrop of space, these space pioneers observed how beautiful, how warm and inviting, and yet how small and fragile our world seemed. Looking closer, they saw changes occurring, some rapidly and some slowly, all over the planet. EARTH'S ATMOSPHERE seemed to almost boil with activity as clouds formed and dissipated daily. Vortices of violent storms raged across the surface. Changes on land were evident as volcanic eruptions and large wildfires sent plumes of smoke high into the atmosphere. At night, the extensive spread of humanity could clearly be seen as clusters of city lights outlining the Earth's habitable land masses. Outer space, scientists found, is an ideal perspective from which to observe and measure changes on our home planet as they happen.

In April 1960, NASA launched the Television and Infrared Observation Satellite (TIROS) into a mid-latitude orbit around the Earth. Followed later that same year by TIROS II, and by TIROS III a year later, these were the first satellites designed to study Earth's rapidly changing weather patterns. They collected and beamed back to Earth thousands of images of cloud cover as well as images of hurricanes, tropical storms and weather fronts that might have otherwise gone undetected for days using conventional methods.

Over the decades, the art and science of space-based observations of the Earth expanded and evolved. Scientists developed satellite 'remote sensors' that are sensitive to regions of the electromagnetic spectrum other than visible light, allowing us to observe things we would not ordinarily see. For instance, designing TIROS II and subsequent satellites to measure infrared light (basically heat) enabled scientists to produce 'false color' images of clouds at night. ('False color' images are made by assigning red, green and blue values to differing wavelengths of radiant energy.) In 1964, the TIROS design was succeeded by the more advanced Nimbus series of weather satellites, and NASA began to explore other orbit strategies for its spacecraft. The Nimbus satellites flew in near-polar, Sun-synchronous orbits, allowing scientists to piece their data together into mosaic images of the entire globe. Other satellites, such as the 1966 Applications Technology Satellite (ATS-1), flew in 'geosynchronous' orbits, where they had fixed views over roughly the same spots on Earth throughout their lifetimes. The ATS-1 provided images over the same areas progressively every 30 minutes, providing meteorologists with more updates of changes occurring in near-real time. In May 1974, NASA launched the first in the series of geostationary weather satellites, called SMS-1. In October 1975, NASA launched the National Oceanic and Atmospheric Association's (NOAA) first Geostationary Operational Environmental Satellites

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(GOES), each with a fixed and continuous view of the changing weather patterns beneath them. The GOES 8 and 10 satellites are still in operation, providing many of the daily weather images over North America that meteorologists display in their television evening news shows today.

Earth is a dynamic planet

Change is perhaps the only constant in our planet's history. Since the Earth's beginning about 4.5 billion years ago, natural climate and environmental conditions on our planet have been in constant flux. Solar variability, volcanic eruptions, meteor impacts, the emergence of life, the formation of an atmosphere rich in oxygen and greenhouse gases, changing ocean circulation patterns, wildfires—over the millennia these and other geological forces shaped an intricately intertwined global climate system that we are only just beginning to understand.

Fast forward to the late 1800s and witness the arrival of a new and profound force for change—humankind's Industrial Revolution. The Industrial Revolution not only ushered in a new era of prosperity for many, it also gave us the tools to dramatically reshape our environment. But the impact we are having on our environment appears to have set the wheels of climate change in motion at a rate not seen on Earth for millions of years. Over the last 100 years, the human population has tripled. In that time, we have consumed fossil fuels in increasing amounts to feed our rapidly industrializing economies, raising atmospheric carbon dioxide levels by 25 per cent. We have burned away forests to make room for our cities and roads, releasing carbon dioxide as well as eliminating trees that would otherwise be re-absorbing some of the gas from the air during photosynthesis. Over the last three decades, humans have burned more than 750 000 square kilometers of forestland every year. We introduced chlorofluorocarbons into the air, which destroy ozone in the stratosphere. Without its stratospheric ozone shield, the surface of our world would be exposed to the Sun's harmful ultraviolet rays, and life as we know it could not exist. Due, at least in part, to the rising levels of carbon dioxide, the average global temperature has risen by 0.5 ºC over the last century, and some computer models predict it could rise another 1 ºC over the next century. Already we see the polar ice caps melting at an alarming rate, while an average of two to three glaciers disappear worldwide every week. Geologists speculate that the US Glacier National Park will contain no glaciers by the year 2050. (See also GREENHOUSE EFFECT.)

What precisely are the causes of these changes? Which changes can be attributed to natural variability, and which are human-induced? How will these changes affect weather and climate patterns in the twenty-first century? How will climate and environmental change impact the quality of life for our grandchildren, and theirs? To help answer these questions and more, NASA developed a new series of satellite remote sensors, beginning with the Earth Resources Technology Satellite (ERTS) launched in 1972 (and later renamed Landsat 1).

From observing to measuring changes on Earth

Altogether, NASA launched a total of seven Landsats, the most recent, Landsat 7, was launched in April 1999. It and Landsat 5 are the only two of that series currently in operation. Since the first Landsat mission, the series' primary objective has been to measure changes on the land surface at high resolution. Whereas the first Landsat could collect images at about 65 m per pixel (at nadir), Landsat 7 can see the surface at resolutions as fine as 15 m per pixel. Such detailed imagery allowed scientists to find geologic features such as meteor impact craters previously undiscovered. Additionally, because healthy green plants reflect and absorb infrared light differently from dry brown plants, scientists found they could use Landsat data to monitor the health of croplands and forestlands over large areas. Landsat also provides an excellent data set for gauging the rates of deforestation, drying inland water bodies, expanding desert regions and sprawling urban centers. Researchers around the world routinely use Landsat data for these and myriad other applications.

In 1973, two researchers at the University of California, Irvine, published their theory that man-made chlorofluorocarbons (CFCs) could damage the Earth's stratospheric ozone shield. Two years later, Congress asked NASA to develop a 'comprehensive program of research, technology, and monitoring of phenomena of the upper atmosphere'. In particular, Congress' intent was to ascertain the 'health' of the ozone layer. In 1978, NASA launched the last of the Nimbus spacecraft series, Nimbus 7, carrying two new sensors called the Solar Backscatter Ultraviolet (SBUV) instrument and the Total Ozone Mapping Spectrometer (TOMS). Sensitive to radiant energy in the ultraviolet region of the electromagnetic spectrum, the Nimbus 7 sensors took advantage of the fact that molecules and aerosol particles reflect certain wavelengths of ultraviolet rays while ozone absorbs others at different levels in the atmosphere. By analyzing the amount of ultraviolet energy reflected back up to the spacecraft, researchers could produce profiles of how thick or thin the ozone was at different altitudes and locations.

But, ironically, the 'OZONE HOLE' over Antarctica was not discovered until May 1985, when a British researcher using a ground-based Dobson ozone spectrophotometer, found that stratospheric ozone over Halley Bay was about 40 per cent less than what had been measured during the previous winter. Surprised, NASA researchers hastily reviewed their TOMS data and

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

found that it too had detected a dramatic loss of ozone over all of Antarctica. Why had they not they discovered the phenomenon earlier? Unfortunately, the TOMS data analysis software had been programmed to flag and set aside data points that deviated greatly from expected measurements, and so the initial measurements that should have set off alarms, perhaps years earlier, were simply overlooked. In short, the TOMS team failed to detect the ozone hole because it was much more severe than scientists expected.

In the years following the discovery of the ozone hole, NASA satellites recorded depleting ozone levels over Antarctica that grew worse with each passing year. In response, in 1987, 43 nations signed the 'Montreal Protocol' in which they agreed to reduce the use of CFCs by 50 per cent by 2000. In 1991, NASA's Upper Atmosphere Research Satellite (UARS) mapped levels of chlorine monoxide in the stratosphere and quickly demonstrated that there is a direct link between the presence of chlorine, the formation of chlorine monoxide during winter in the Southern Hemisphere, and the destruction of ozone. Was a global disaster averted? Will levels of stratospheric ozone stabilize and return to normal in the coming years? We do not know yet, only time and more data will tell.

Towards predicting climate change

As we make the transition to the twenty-first century, the Earth's environment will be the focus of many agricultural, industrial, societal and political decisions. We recognize that we need more and better data to help us understand how current environmental change trends may impact climate years or even decades in the future. To collect these data, in 1991, NASA established the Earth Observing System (EOS) program in response to a US presidential initiative to provide in-depth scientific understanding about the functioning of the Earth as a system. The EOS charter is to collect, at a minimum, a new 15-year data set on which to base a long-term and comprehensive examination of our planet's climate system. EOS is a multinational endeavor comprising three parts: (1) a series of advanced satellite remote sensors; (2) a robust new computer network (called EOSDIS) for processing, storing and distributing EOS data; and (3) about 850 scientists working in many nations and in many Earth science disciplines who will use these new data in their research.

Already, NASA has launched four EOS spacecraft in the 1990s, including Landsat 7, QuikScat, ACRIMSAT and the EOS flagship, Terra, that was launched in December 1999. There are 15 more EOS satellites scheduled to launch through to 2003. Rather than focusing on individual disciplines, the data from these EOS satellites will be integrated into a more wholistic study of the Earth that includes seven overarching themes:

- radiation, clouds, water vapor, precipitation and atmospheric circulation;
- ocean circulation, productivity and exchange with the atmosphere;
- tropospheric chemistry and greenhouse gases;
- land ecosystems and hydrology;
- snow, ice and glacier extent;
- ozone and stratospheric chemistry;
- volcanoes and climate effects of aerosol.

With EOS, for the first time ever for a major Earth observation program, the goals include freely sharing the resulting data with both scientists and civilian organizations alike. This treatment of data contrasts sharply with previous satellite missions for which public access to data was quite costly. In fact, some EOS data are being directly broadcast to anyone anywhere who has a compatible receiving station and the capacity to process and store such a huge flow of information. As its name suggests, EOS is a large-scale, long-term collaborative mission involving scientists in government agencies, academia and industry from many nations.

Conclusion

Today, Earth scientists' goals are to not only observe weather patterns around our world, but to determine the causes and effects of climate and environmental change. With increasingly sophisticated satellite remote sensors, we can measure a wide range of geophysical parameters (such as surface temperature, distribution of clouds and aerosol particles, the abundance of trace gases in the atmosphere, or the distribution and types of life on land and in the ocean) with unprecedented accuracy and resolution. Moreover, we can now measure how changing certain aspects of the climate system (such as cloud cover) can have a 'ripple effect' through other aspects of the climate system (such as surface temperature, precipitation, the radiation budget). Scientists are feeding these new satellite data, collected throughout the Earth's climate system, into sophisticated new computer models that, ultimately, will enable them to predict climate changes months, years or even decades before they occur. If we are to become better stewards of our home planet—if we are to leave abundant natural resources to our grandchildren and to their grand-children—then we must continue monitoring our planet with satellite sensors ever improving our understanding of how the Earth system works.

David D Herring

Michael D King

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth's Atmosphere

Composition and physical conditions of the neutral atmosphere

Our planet is surrounded by a thin envelope of gas, its atmosphere. This atmosphere, containing large amounts of molecular oxygen and therefore ozone, which protects us from the harmful solar UV radiation, and greenhouse gases like water vapor or carbon dioxide, maintaining mild climatic conditions, is unique in the solar system. It plays a crucial role in preserving at the surface of the Earth an environment suitable for the preservation and development of life, and must be considered as a key component of the Earth system. The average pressure at the ground level is 1013 mbar, with a column atmospheric mass of 1 kg cm^{-2} and a density of 2.7×10^{19} molecules/cm^3. The average temperature at the surface of the Earth is 288 K, or 15 °C. The atmosphere is mainly composed of molecular nitrogen (78.1%), molecular oxygen (20.9%) and less than 1% of argon, with an average molecular mass of 29 atomic mass units (table 1). A few precipitable centimeters of water are also present in the atmosphere, either as vapor or in the form of water droplets, or ice particles, in the clouds. The volume mixing ratio of water vapor is highly variable, but is everywhere less than 1%. Carbon dioxide, despite its small relative abundance (0.04%), plays a crucial role in the radiative budget of the planet. Minor constituents of Earth's atmosphere are listed in table 2.

The chemical composition of the atmosphere was discovered by the French chemist Antoine-Laurent de Lavoisier in the second part of the eighteenth century from the study of oxidation reactions. Lavoisier, at the same time as the English scientist Joseph Priestley, came to the conclusion that a particular constituent of air (oxygen) has the property of combining with metals to form oxides from which it can be released by heating. The existence of a second constituent (nitrogen), which does not enter the formation of oxides, was also required by experimental results. At this time, instruments for measuring the most basic physical properties of the atmosphere (temperature, pressure, rain, wind, humidity) had already been invented and the first systematic measurements of temperature were just beginning to be performed, for example at the Royal Meteorological Institute of the Netherlands. The vertical exploration of the atmosphere, first illustrated by Pascal's experiment on the Puy de Dome in 1647, was continuing with the help of kites and balloons: the first meteorological sounding, realized in 1783, reached an altitude of 3400 m. At the end of the nineteenth century, two French meteorologists, Hermitte and Besançon, performed systematic launches of balloons carrying pressure and temperature sensors. The stratosphere was discovered in 1899 by Léon Teisserec de Bort. During the twentieth century observations rapidly developed through the use of stratospheric balloons and airborne measurements, whereas in the most recent period, remote sensing from satellites has become the most powerful means of monitoring the global structure of Earth's atmosphere.

The vertical structure of the atmosphere is controlled by the law of hydrostatic equilibrium: density adjusts itself in such a way that the (downward) gravitational force is balanced by the (upward) force induced by the vertical pressure gradient. Assuming that air is a perfect gas, it can be shown that the atmospheric pressure decreases exponentially with altitude, with values of 100 mbar and 10 mbar at 15 km and 30 km altitude respectively. Considering its mass, Earth's atmosphere is equivalent to a homogeneous layer of 8 km thickness. Up to 100 km altitude, the atmosphere is mainly composed of neutral species. Constituents are mixed by turbulent viscosity, resulting from wind and wave activity, and the vertical distribution of each constituent does not depend on its own molecular mass. To a first approximation, the species which are not affected by condensation or chemistry, like molecular nitrogen, molecular oxygen and argon, have the same vertical distribution and are therefore uniformly mixed. The region below 90 km altitude, where constituents are uniformly mixed, is called the 'homosphere'.

Because the temperature decreases with altitude, as described later, and because the saturation pressure of water vapor is an increasing function of the temperature, the water vapor mixing ratio decreases with altitude and is only a few ppmv (part per million in volume) above 15 km. Another important example of a non-uniformly mixed constituent is ozone (O_3), whose abundance is controlled by chemistry. Ozone results from the dissociation of molecular oxygen by solar ultraviolet photons, which yields oxygen atoms able to form ozone by reaction with molecular oxygen. This reaction occurs between 15 km and 50 km in the stratosphere. Although the ozone mixing ratio does not exceed 0.1 ppmv, the ozone layer is quite an efficient shield against harmful solar ultraviolet radiation, because photons in the wavelength range 200–300 nm easily break down the ozone molecule. The quantity of ozone present in the stratosphere is equivalent to a 3 mm thick layer of pure ozone in STP conditions (standard pressure and temperature: 1013 mbar, 273 K). The presence of a well-marked stratosphere is an oustanding characteristic of Earth. The atmospheres of our two 'sister' planets, MARS and VENUS, which are not supplied with oxygen by photosynthesis, have no stratosphere.

Historically, ozone was measured for the first time at the beginning of the 1950s, by using the method of solar OCCULTATION. The earliest ozone measurements were obtained by the Naval Research Laboratory (NRL) using captured V2 rockets, which with their inertial stabilization offered an excellent platform. The occultation method consists of observing the Sun, or a star, setting or rising behind the limb of Earth. In this way, the absorption spectrum of the upper atmosphere can be observed as a function of altitude, yielding the vertical distribution of absorbing species. Accurate ozone measurements were obtained up to 70 km. The same method allowed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Principal constituents of the atmosphere. (From Banks and Kockarts (1973).)

Constituent	Molecular mass (amu)	Molecular mass ($g \times 10^{24}$)	Composition by volume (%)	Molecular concentration at STP (cm^{-3})	Atmospheric thickness (atm cm)	Atmospheric column content (cm^{-2})
N_2	28.02	46.51	78.08	2.10×10^{19}	6.24×10^5	1.68×10^{25}
O_2	32.01	53.12	20.95	5.63×10^{18}	1.68×10^5	4.50×10^{24}
Ar	39.96	66.33	0.93	2.51×10^{17}	7.47×10^3	2.01×10^{23}
CO_2	44.02	73.06	0.03	8.87×10^{15}	2.64×10^2	7.09×10^{21}
Air	28.97	48.08	100.00	2.69×10^{19}	8.00×10^5	2.15×10^{25}

Table 2. Minor molecular constituents of the atmosphere. (From Banks and Kockarts (1973).)

Molecule	Ratio by volume	Remarks
CO_2	3×10^{-4}	Small variation
H_2O	10^{-2}–10^{-5}	Variable, particularly in troposphere
CH_4	1.5×10^{-6}	Fairly constant, mixed in troposphere
N_2O	2.5×10^{-7}	Approximately constant, mixed in troposphere
H_2	5×10^{-7}	Probably mixed in troposphere
O_3	10^{-8}–10^{-7}	Maximum in stratosphere
CO	5×10^{-8}–2×10^{-7}	Variable
NO_2	5×10^{-10}–2×10^{-8}	Variable, effect of NO

measurement of O_2 absorption as high as 140 km, whereas it was firmly believed that this gas was totally absent above an altitude of 90–100 km.

Above 90 km altitude, in the 'heterosphere', the different species separate from each other due to the effect of gravity. Viscous forces are no longer able to mix the constituents and each one distributes according to its own molecular mass. In this altitude range, the atmosphere is flooded with ultraviolet photons and energetic particles, and the atmospheric molecules are dissociated and ionized. Atoms, which are lighter than molecules, extend up to higher altitudes: atomic oxygen is the major constituent above 200 km altitude, whereas hydrogen, together with helium, dominates above 700 km. The average molecular mass of atmospheric molecules therefore decreases with altitude in the heterosphere. Ultimately, through their thermal kinetic energy, hydrogen atoms are able to escape from the terrestrial atmosphere in the upper regions where collisions are not frequent.

The general structure of the vertical temperature profile is given in figure 1. It is based on the nomenclature first introduced by Sydney Chapman in the 1930s. In the lower part of the atmosphere, below 10–15 km, termed the 'troposphere', the temperature decreases with altitude with an average rate of 6.5 °C km^{-1}, nearly equal to the humid adiabatic temperature lapse rate. The heat source is the surface, illuminated by the visible solar radiation. Heat is transferred to the troposphere through three different processes: evaporation of water and release of latent heat through the formation of clouds; infrared emission and absorption by greenhouse gases; and sensible heat flux. Vertical transport of water vapor and heat is ensured by convective motions. The transparency of Earth's atmosphere in the visible range, which allows most solar

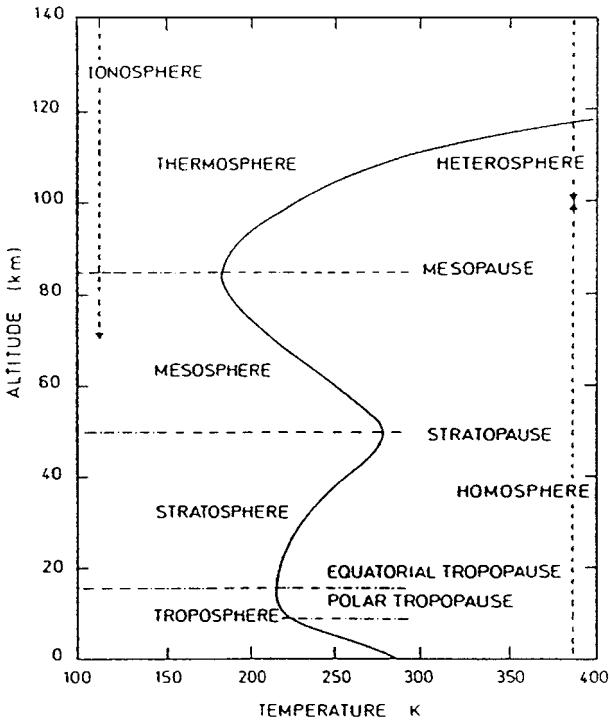


Figure 1. Thermal structure of Earth's atmosphere. (From Brasseur and Solomon (1994).)

photons to reach the ground, is a major determinant of the Earth's climatic system. On Venus, for example, the very dense lower atmosphere of carbon dioxide, with a pressure at the ground level of ≈ 90 bar, prevents the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

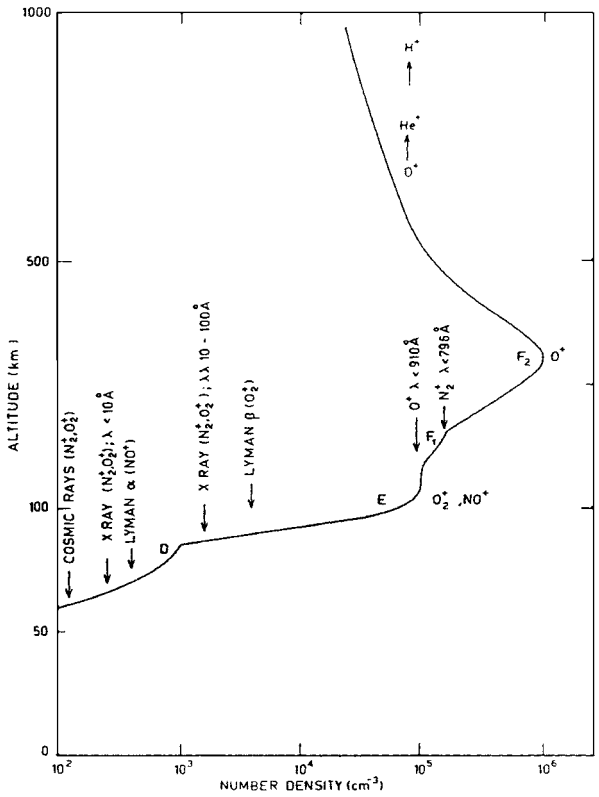


Figure 2. Average daytime electron distribution with principal ions in the various layers of the atmosphere. (From Banks and Kockarts (1973).)

solar visible radiation from reaching the ground, yielding a completely different behavior of the climatic system (see VENUS: ATMOSPHERE). Above 15 km, in the 'stratosphere', temperature increases from 220 K up to 270 K at 50 km altitude. The stratosphere is directly heated by the solar ultraviolet radiation, absorbed by ozone, whereas cooling occurs through infrared emission of carbon dixoxide, and to a lesser extent water vapor. Above 50 km altitude, in the 'mesosphere', the ozone concentration decreases and the temperature decreases regularly up to 90 km altitude, where it is less than 200 K. Above 90 km, in the 'thermosphere', the enhanced ultraviolet radiation results in an increase of temperature with altitude, with a high diurnal and seasonal variability due to the low thermal inertia of the tenuous atmosphere. Heat absorbed in the thermosphere is transported downward by thermal conduction and radiated in the infrared at the top of the mesosphere.

The ionized environment of Earth

The ionized atmosphere, or 'ionosphere', is composed of several well marked layers. This layer structure is due to the variety of ionizing fluxes (UV and x-rays, energetic particles) and ionized species. Although a few ions are present in the 60–80 km range, ionization is substantial only above 80–90 km. The main ionospheric layers are denoted by D, E and F. They extend from the base of the D layer, at 60 km altitude, up to 400 km, above the peak of the F_2 region, as shown in figure 2. The most deeply penetrating energy fluxes are cosmic rays, x- and solar Ly-α photons, which are responsible for IONIZATION in the D layer and are very sensitive to solar activity. Soft x-rays (λ = 1–10 nm) and long-wavelength UV radiation are the main ionization sources in the E layer. Major ions are atomic and molecular oxygen, together with nitrogen monoxide. Whereas O_2^+ and N_2^+ are created by ionization of the corresponding molecules, which are the dominant species, according to the following reaction scheme

$$O_2 + h\nu \rightarrow O_2^+ + e^-$$

NO^+ is formed through the charge-exchange reaction

$$O^+ + N_2 \rightarrow NO^+ + N.$$

The vertical distribution of ions is complex. O_2^+ and NO^+ dominate in D, E and F_1 regions, whereas O^+ is the major ion in the F_2 region. Above 1000 km, light ions like H^+ and He^+ are dominant. Electron density peaks at 300 km, with a maximum value of 10^6 cm^{-3}. This value presents a strong diurnal variability, due to the lack of solar photons during the night. At night, the electrons completely disappear in the D layer and their abundance considerably decreases in E and F_1 layers. The behavior of the F_2 layer is more complex, because transport by vertical diffusion and winds plays an important role in the topside ionosphere.

Historically, the role of the conductive ionized envelope of our planet in reflecting radio waves was discovered at the very beginning of the twentieth century. The first transcontinental radio communication was established by Marconi. At the end of the 1920s, Appleton in England and Tuve in the United States developed the first instrument dedicated to the teledetection of the ionosphere, the 'ionosonde', which dominated research on high atmospheric physics until the beginning of the space era. Modern ground-based observations of the physical parameters of the ionosphere (density, temperature, wind) use the technique of the 'incoherent scattering radar', which consists of illuminating the ionosphere by a radio wave beam and analysing the spectral characteristics of the scattered signal.

At increasing altitudes, the role of the Earth's intrinsic magnetic field becomes important. In the 'magnetosphere', the motion of particles is mainly controlled by electromagnetic forces (see MAGNETOSPHERE OF EARTH). Because the magnetic field deflects the charged particles of the SOLAR WIND, which blows from the dayside to the nightside with a velocity of about 400 km s^{-1}, it carves in the flow of the solar wind an elongated cavity. This cavity presents to a first approximation an axial symmetry around the Sun–Earth axis, contrary to the spherically symmetric low layers of the ionosphere. Its shape is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

approximately parabolic in the solar direction, where it extends up to about 12 Earth radii. In the downwind direction, the magnetosphere is almost cylindrical and forms a magnetic tail with a length of several AUs (1 AU = 1 Astronomical Unit, that is the Sun–Earth distance, or 1.5×10^{11} m), which can be compared to cometary tails. The magnetosphere is composed of an extremely tenuous medium, with a typical density of one particle per cubic centimeter, where the collisions between particles play a negligible role. The inner boundary of the magnetosphere, the plasmapause, is located at about 4 Earth radii from Earth center in the equatorial plane. The region inside the plasmapause, called the 'plasmasphere', is filled with a relatively cold ($T < 10\,000$ K) and dense ($n \approx 100$ particles/cm^3) plasma originating in the ionosphere (see MAGNETOSPHERE OF EARTH: PLASMASPHERE).

A few of the incident solar wind particles (about 1%) penetrate the magnetosphere. Some of them precipitate at high latitude along the lines of the magnetic field lines and reach the low atmospheric levels: an electron of typical energy 10 keV can penetrate down to 100 km altitude before being substantially dragged. Precipitating energetic particles (mainly electrons) produce, through ionization of species like atomic oxygen, followed by recombination and fluorescence, the well-known and visually stunning phenomenon of auroras. The luminosity pattern of an AURORA, which reflects the shape of Earth's magnetic field lines, anchored in the deep Earth and disturbed at high altitude by the flow of precipitating particles, is extremely various, from quiescent, diffuse big auroras to the complex, beautiful patterns of auroral 'bands' and 'arcs'.

In a more detailed way, the primary electron is the source of many secondary electrons at altitude, through ionization reactions of the type

$$A + e^- \rightarrow A^+ + e^- + e^-.$$

Ionized species recombine, either by 'radiative'

$$A^+ + e^- \rightarrow A^* \rightarrow A + h\nu$$

or by 'dissociative' recombination

$$XY^+ + e^- \rightarrow X^* + Y^* \rightarrow X + Y + h\nu$$

by emitting photons. The most frequently observed emissions are the green and red lines of atomic oxygen at 558 nm and 630–636 nm respectively. A violet coloring, due to the N_2^+ line at 391 nm, may sometimes be observed. The red line of oxygen is produced by radiative de-excitation of the first excited state O(^{1}D) of oxygen. Because the radiative lifetime of O(^{1}D) is long, about 2 mn, fluorescence in this line cannot occur in the lower ionosphere ($z <$ 200 km), where the excited state is deactivated by collisions with air molecules in a time smaller than the radiative emission time. For this reason, red auroras occur only at high ionospheric levels.

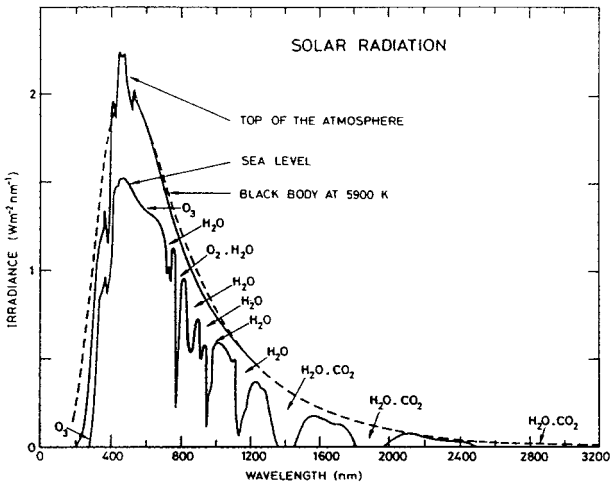


Figure 3. Spectrum of solar radiation outside Earth's atmosphere and at sea level, including comparison with black body radiation at a temperature of 5800 K. (From Banks and Kockarts (1973).)

Atmospheric absorption, scattering and scintillation

The UV and visible radiation of the Sun is partially or totally absorbed by Earth's atmosphere. At wavelengths of less than about 300 nm, stratospheric ozone molecules are easily dissociated by solar photons. At longer wavelengths, in the visible range where the solar radiation spectrum peaks, there is much less absorption, although some absorption occurs around 600 nm due to ozone. At wavelengths longer than 700 nm, absorption occurs mainly by excitation of vibrational (and rotational) levels of H_2O and CO_2 molecules, as well as O_2 in the forbidden band at 760 nm. As can be seen in figure 3, an extinction continuum is also present. It is due to the scattering of photons by air molecules, mainly N_2 and O_2, and cloud particles. The average reflectivity (or 'albedo') of Earth's atmosphere and surface is 0.3. Only 70% of incoming solar photons are absorbed by Earth's system, the remainder being scattered to space, principally by clouds which cover on average 50% of the planet.

In the field of astronomical ground-based observation, the phenomenon of ATMOSPHERIC REFRACTION, in the presence of turbulent motions, is also of great significance. The troposphere and stratosphere of the Earth are highly turbulent because of winds and wave activity: the kinetic energy of large-scale atmospheric motions (typically kilometric) is progressively transferred to small scales (typically millimetric), where dissipation occurs by viscous friction. This process is accompanied by local fluctuations of the density, therefore of the refractive index. The planar wave front received from a star is distorted as the light passes through the atmosphere. Both intensity and propagation direction are randomly modified due to the variations in the cross-sectional area and orientation of bundles of light rays. The fluctuations in the intensity are responsible for the well known phenomenon of SCINTILLATION which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

results from traveling inhomogeneous air masses at altitude. The associated variations of the propagation direction yield a loss of angular resolution, which depends on the turbulent scales and their vertical distribution, as well as of wind patterns. It is theoretically shown and experimentally verified that atmospheric turbulence results in an angular dispersion, or 'SEEING', in the range from 0.2 to 1 arcsec, depending on the altitude of the observation site and local weather conditions. The seeing is therefore larger than the diffraction-limited resolution of a 1 m class telescope. The only way to overcome this intrinsic limitation is to use adaptative optics, the principle of which is to monitor the phase and adjust the mirror's shape in real time.

Global radiative budget and dynamics

Despite substantial atmospheric extinction, as previously described, Earth's atmosphere is largely transparent to solar photons in the visible part of the spectrum. As a result, the visible radiation of the Sun is mainly absorbed by Earth's surface. It causes heating of soil, water, vegetation and the adjacent air, resulting in the existence of a troposphere. The surface, which is heated, radiates at long wavelength, in the infrared portion of the spectrum, mainly from 7 μm up to 100 μm. Although the energy of the emitted infrared photons is not large enough to dissociate or ionize molecules, contrary to the solar ultraviolet radiation, it is able to increase the internal vibrational and rotational energy of absorbing molecules, principally H_2O, CO_2 and O_3. That excess energy is transferred to the main atmospheric molecules as kinetic energy by molecular collisions. Conversely, cooling of the atmosphere occurs through emission of radiation, due to molecular state transitions between vibrational and rotational levels. Some of this radiation is directed towards Earth's surface, resulting in a substantial GREENHOUSE EFFECT. In the absence of any atmosphere, the average surface temperature would be 255 K, well below the freezing point of water, and no life could exist. At ground level nearly as much radiation is received from the atmosphere, in the infrared, as from the Sun, in the visible. The very high temperature at the surface of Venus (470 °C) is mainly due to the strong greenhouse effect of the massive carbon dioxide atmosphere; the black body temperature of Venus at the top of sulfuric acid clouds where thermal photons are released to space is about −40 °C, lower than the black body temperature of Earth (−18 °C) because of the high reflectivity of Venus' clouds.

The dynamics of the low atmosphere, thermally and dynamically coupled with the oceans, is controlled by the redistribution of heat from low-latitude regions, where it is preferentially deposited by solar visible radiation, to the whole planet. Indeed, the infrared cooling is more uniformly distributed at the surface of the planet than solar heating, with an excess downward flux of radiative energy in the intertropical region and the reverse situation at high latitudes. The Earth can be compared to a Carnot machine, with both a warm source (low-latitude regions) and a cold source (high-latitude regions), which converts thermal energy into work. The mechanical energy produced is used for generating large-scale organized motions able to transport heat towards the poles. The efficiency of the machine is weak, much lower than the maximum value allowed by the laws of thermodynamics ($\approx$ 10%). For a flux received from the Sun of 340 W m^{-2} in global, diurnal and latitudinal average, the mechanical power required to maintain the large-scale circulation is only 2.4 W m^{-2}, that is an efficiency of only 0.7%. Despite this low efficiency, the terrestrial 'wind machine', which feeds the global hydrological cycle, is of prime importance for maintaining life at the surface of Earth.

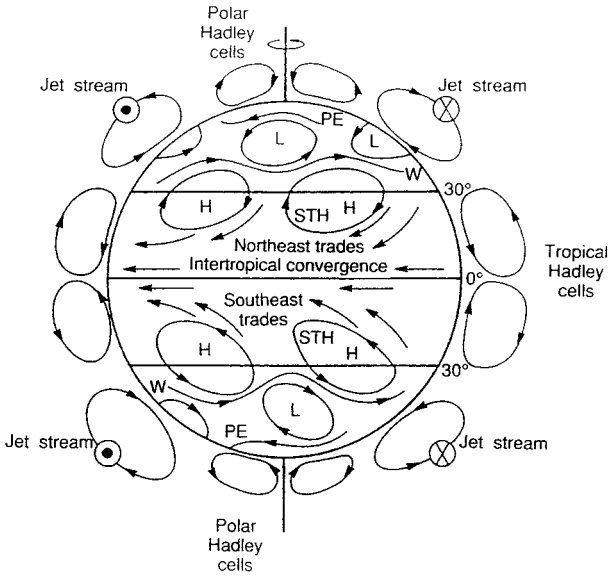


Figure 4. Principal features of the atmospheric circulation. (From Graedel and Crutzen (1993).)

Because the heating of the Earth by the deposited solar light is greatest at the equator, the hot equatorial air expands and moves upward. Consequently, air at low altitudes and from higher latitudes moves toward the equator to take its place. This equatorward flow is balanced by a return flow at higher altitudes, as summarized in figure 4. This poleward moving air progressively cools by infrared radiation to space. At about 30° latitude, the cooled denser air moves downward, closing what is called the 'Hadley circulation' on both sides of the equator. The equatorial upwelling branch of the Hadley cell occurs in the 'intertropical convergence zone' (or ITCZ). This is characterized by heavy precipitation due to strong upward convective motions, yielding massive water condensation and subsequent rain, and relatively low surface pressures. Condensation plays an important role in controlling and enhancing the tropical Hadley circulation, through the release of large amounts of energy (the 'latent heat of condensation'). On the other hand, the regions of subsiding air are much dryer and surface pressure relatively high. Due to the Coriolis force,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which is associated with the rotation of the Earth and results in a deflection to the right of the direction of motion in the northern hemisphere and to the left in the southern hemisphere, equatorward motions of air in the tropical latitudes result in 'trade winds', blowing from the northeast in the northern hemisphere and from the southeast in the southern hemisphere. The flows at mid and high latitudes are much less regularly ordered than in the tropics because of the strong influence of travelling high and low pressure systems. Because of the Coriolis force, the winds tend to rotate clockwise around high-pressure systems (called 'anticyclonic circulation') in the northern hemisphere and counterclockwise in the southern hemisphere. Circulation around low-pressure systems ('cyclonic circulation') occurs in the opposite direction. In the mid-latitude regions, poleward heat transfer is ensured through sporadic air mass exchanges across 'warm fronts' and 'cold fronts'. This mechanism is due to the baroclinic instability, or literally 'pressure sloping instability', related to the horizontal temperature gradient, generating poleward motions of warm air and equatorward motions of cold air. Contrary to what happens in the tropical regions, the atmospheric flow at mid-latitudes is globally westerly (that is from the west). At higher altitudes, at the polar regions, heat transport is accomplished through the formation of cells rather similar to the tropical Hadley cells. It must be noted that about one-third of the poleward heat transfer takes place in the oceans.

The Martian climate, although colder, may be compared with the terrestrial climate. Global circulation on Mars also consists of trade winds in the intertropical zone and cyclonic/anticyclonic systems at mid-latitudes. This is due to similarities in the way in which the atmosphere is heated from below, as well as in the inclination of the rotation axis (23.5° for Earth, 25.2° for Mars) or the duration of the day, which differs by only 40 min (see MARS: ATMOSPHERE). Nevertheless, the absence of water oceans on Mars results in striking differences, like the development of a single Hadley cell with a wide rising branch over the summer hemisphere and a wide subsiding branch over the winter hemisphere, quickly reversing around the equinoxes.

The dynamics of the atmosphere is intimately coupled with the dynamics of the ocean, able to transport heat over large distances. Surface waters are dragged off along with winds and, because of heat exchanges with the atmosphere, modify the atmospheric thermal and wind regimes. One example of such a surface oceanic current is the Gulf Stream. Westerlies at mid-latitudes push surface water to the east, whereas easterly trade winds in the ITCZ act in the opposite way. This results in a clockwise rotation of water bodies, with a northward warm current along the eastern coast of America. During winter, cold and dry continental air coming from the west is heated and loaded with moisture when passing over the Gulf Stream, which results in strong perturbations and snow falls. The dynamical interaction between the atmosphere and the ocean at low latitudes may have dramatic consequences, mainly through the combined actions of evaporation and atmospheric convection, yielding the formation of vigorous cumulonimbus convective cells, or 'hot towers' (figure 5), pumping air and moisture upward and associated with massive rainfalls, lightning (which is electrical discharges between the positively charged upper layers of clouds and negatively charge lower layers, or Earth's surface) and winds. Such systems, associated with strong perturbations of the pressure field, sometime become hurricanes and tornados, rapidly rotating under the effect of the Coriolis force.

Figure 5. *Challenger* Space Shuttle orbiter photo showing a system of cumulonimbus clouds.

One of the most important, planetary scale, phenomena resulting from the interaction between the ocean and the atmosphere is the El Niño southern oscillation (ENSO). In normal conditions, trade winds push equatorial pacific waters from the Peruvian coasts to Indonesia. Surface waters are progressively heated during their slow westward migration through the Pacific ocean and reach a temperature of 29 °C near Indonesia, which results in a strong longitudinal thermal contrast. The ENSO phenomenon, which is still incompletely understood, is provoked by a weakening of trade winds, therefore thermal contrast, with a subsequent warm eastward countercurrent. This current, episodically appearing near Peruvian coasts right after Christmas (hence, the reference to 'the child'), is well known by the fishermen of this area, and is associated with massive rains. In the 1920s, Sir Gilbert Walker, director of the Indian Meteorological Office, found evidence of an anticorrelation between the atmospheric pressure in Tahiti, in the central Pacific, and the pressure in Darwin, in Australia. In 1969, J Bjerkness established the link between the warm oceanic countercurrent and the associated meteorological perturbations. Cumulus systems normally present over Indonesia are displaced several thousands of kilometers to the east, which tends toward amplifying, and sometimes reversing, trade winds. This positive retroaction ef-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

fect yields dramatic consequences, with severe drought in Indonesia and heavy rains in Ecuador and northern Peru.

Biochemical cycles of main atmospheric species

From a chemical point of view, the atmosphere of the Earth is strongly coupled with Earth's solid and liquid surfaces through a variety of biogeochemical cycles. Earth's atmosphere is generally supposed, although this is still a controversial question, to have been outgassed from the interior, in a short episode of intense differentiation in the first 100 million years of Earth's life, 4.5 billion years ago. The primitive Earth atmosphere contained mainly carbon dioxide and water vapor, with possibly some traces of methane and other reduced gases. It is thought that a substantial amount of water may have been brought by comets and meteorites over a rather long period of time after accretion (a few hundred million years) and it is not excluded that Earth's atmosphere is partly of cometary and meteoritic origin. Oxygen at present contained in our atmosphere is the result of photosynthetic activity in the biosphere. It progressively accumulated in the atmosphere during the Precambrian era and reached its present level 300 million years ago, when the first land animals appeared. The atmospheric oxygen content is thought to be regulated by the action of marine photosynthetic activity, which removes a small fraction of carbon from the atmosphere through the formation of organic reduced carbon, a very small part of which precipitates downward through the ocean and is finally incorporated into sea-floor sediments. Oxygen atoms left behind by reduced carbon, which is trapped in sedimentary rocks for typical times of a few hundred million years (due to sea floor expansion), are responsible for the presence of a permanent oxygen atmosphere on Earth. Ultimately, reduced carbon is rejected by volcanoes or resurfaced metamorphic rocks above subduction zones and oxidized by the atmosphere, closing the cycle.

Similarly, the 'oceanic biological pump' plays a crucial role in regulating the atmospheric content of carbon dioxide, through the so-called 'carbonate–silicate' cycle. Carbon dioxide is released by volcanism and metamorphic processes from the solid Earth to the atmosphere and ocean at a rate which is estimated to be large enough to double the amount of carbon in the ocean and atmosphere in a few hundred thousand years. It must be noted that the terrestrial ocean contains much more carbon dioxide than the atmosphere (40 000 Gt versus 750 Gt; $1\ \mathrm{Gt} = 10^{12}$ kg). The carbon contained in carbonate rocks, if entirely released to the atmosphere, would represent $\approx$ 50–100 bar of CO_2, which is similar to the present content of the Venusian atmosphere. Due to the rapid exchange of CO_2 between the atmosphere and the ocean, the residence time of a CO_2 molecule in the atmosphere is less than 10 years (less than 1000 years in the ocean). For this reason, the ocean and the atmosphere must be treated as a single reservoir when studying their interaction with the solid planet, which acts over much longer time scales. Because carbon dioxide is released to the atmosphere, it must be returned to the solid Earth at an equal rate. The mechanism of this return is well known. It is the weathering of silicate minerals by rain water, containing dissolved carbon dioxide, followed by the deposition of carbonate minerals in sedimentary rocks. Cations produced by silicate weathering (Ca^{2+}, Mg^{2+}) are carried by rivers to the oceans, where they react with carbonate ions (CO_3^{2-}) to precipitate as carbonate minerals. In this way, oxidized carbon is trapped in sea-floor sediments in the form of limestone. Because the rate of supply of cations by weathering strongly depends on humidity and temperature conditions, and because CO_2 is a greenhouse gas, it is thought that this cycle has a controling influence on the budget of carbon in the combined atmosphere–ocean system.

In spite of the assumed stabilizing effect of the carbonate–silicate cycle over geological times, climatic conditions on Earth must be regarded as having evolved through a number of different stages over a long period of time. An example of global climatic instability is provided by the glacial–interglacial cycles which occurred during the Quaternary. The chemical composition and temperature of the ancient atmosphere may be retrieved from ice cores drilled on the Antarctic glacier. It is possible to trace back the Earth's climatic history during the last few hundred thousand years. Periodicities of 100 000, 41 000 and 23 000 years are evidenced and, as primarily inferred by Milankovitch, they are the consequences of variations in the motion of Earth through space. Orbital variations may be described by three parameters: the degree of ellipticity, the angle of tilt between the axis of rotation of Earth relative to the Earth–Sun orbit and the orientation of the axis of rotation with respect to the orbital path, which varies because of the precession phenomenon. All these three parameters undergo predictable changes due to the gravitational attractions exerted by the Sun, the Moon and the planets. These long-term variations modify the spatial and seasonal sequences of solar energy deposition at the surface of our planet, resulting in variations of Earth's climate. For example, at the end of the last glacial period, about 10 000 years ago, the inclination of the Earth's rotational axis relative to the ecliptic plane was about 24.5° (at present 23.5°) and the closest approach of the orbiting Earth to the Sun occurred in July (at present January). This past condition resulted in about 8% more solar radiation reaching the northern hemisphere in the summer and about 8% less in winter time. Although the astronomical effects are weak by themselves, they are relayed by strongly nonlinear feedback effects which result in large-amplitude climatic variations. Besides, the climate is also thought to exhibit large variations by itself, without any changes to astronomical effects.

Effects of anthropogenic activities on climatic evolution and atmospheric composition

It is thought that anthropogenic activities, responsible for a rapid increase in gas and particle emissions in the atmosphere, are potentially able to destabilize the global

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

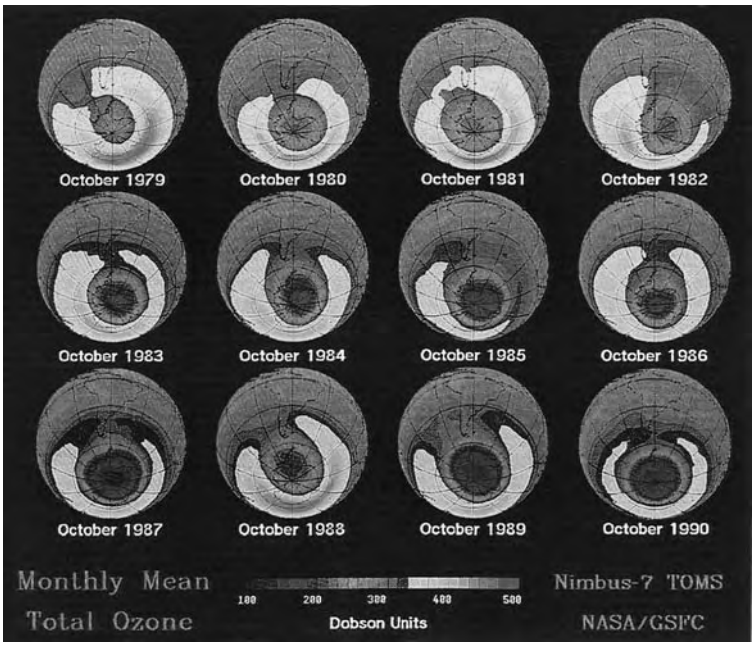


Figure 6. Total column ozone concentrations (in Dobson units; 100 Dobson = 1 mm thick layer of pure ozone in STP conditions) over Antarctica as determined by TOMS (Total Ozone Monitoring Satellite) during October from 1979 to 1990. (From Graedel and Crutzen (1993).)

climate of Earth. The carbon dioxide content of the atmosphere has increased from 275 ppmv at the end of the eighteenth century up to 350 ppmv at the present time. This increase is due to a variety of worldwide human activities. Until about 1950, the release of CO_2 was mainly due to the oxidation of organic matter in agricultural soils. Currently, the largest net source of CO_2 is fossil fuel combustion. Tropical deforestation also contributes to the atmospheric increase, although to a lesser degree. Because carbon dioxide is a greenhouse gas, this increase is expected to enhance the infrared opacity of the atmosphere, resulting in an increase of Earth's surface temperature. In addition to this direct effect, related secondary effects could play an important role, like the expected increase in the water evaporation rate, yielding both an additional warming, because water vapor is a greenhouse gas, and a counteracting cooling due to the increase of cloud cover and therefore of the average planetary albedo. It is emphasized that H_2O is the most important greenhouse gas, blocking the infrared radiation from Earth's surface more efficiently than CO_2 does. The temperature record for the past century shows that about 0.5 °C of warming has occurred in the last 130 years and that the last two decades were the warmest during this period. Although some natural variability cannot be excluded, it is thought that this general warming is the consequence of the regular atmospheric CO_2 increase.

Another consequence of anthropogenic gaseous emissions is the decrease of the stratospheric ozone content, due to the destructive action of chlorine contained in CFCs (chlorofluorocarbons). CFCs (mainly $CFCl_3$ and CF_2Cl_2) are not broken down in the troposphere. Over a period of a few years, these very stable gases move from their release point at the surface up in the stratosphere, where the solar ultraviolet radiation is intense enough to break them and release active chlorine species (Cl, ClO), able to destroy ozone. Satellite observations, performed since the beginning of the 1980s, show that stratospheric ozone regularly decreases at a global scale, with a rate of a few per cent per decade. This decrease is attributed to the action of CFCs. It has been decided at the Montreal Conference of 1990 to stop the emission of CFCs, but the long characteristic time of this destructive process prevents any significant improvement being obtained before the middle of the next century. The most dramatic illustration of stratospheric ozone destruction is the Antarctic 'OZONE HOLE', a rapid and accelerating decrease in the ozone over the Antarctic each September and October. The decrease in total ozone content may reach a factor of three. As soon as it was discovered, in 1985, attention was focused on the possibility of an enhanced chlorine destruction cycle, which was rapidly confirmed by observations and experimental work.

The strong decrease of stratospheric ozone over the Antarctic during the austral spring, which is illustrated in figure 6, is due to the very low temperatures occurring in the polar stratosphere of the southern hemisphere. During the polar winter, in the absence of solar radiation (from June to September), the temperature in the stratosphere can be as low as −85 °C, because polar air is enclosed in the 'polar vortex', a very large whirlwind which develops during winter over the polar regions. These

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

very low temperatures are responsible for the formation of 'polar stratospheric clouds' (or PSCs). These clouds are composed of pure water or mixed water/nitric acid ice. The surfaces of the PSCs serve as reaction sites for chemical processes able to transform unreactive chlorine species to reactive ones. These catalytic processes take place during the four months of the polar night and produce, from inactive HCl and $ClONO_2$ species, large quantities of HOCl and Cl_2 (see later). The combined effects of denitrification, through freezing and sedimentation of HNO_3 crystals, and chlorine radicals released at sunrise (Cl, ClO) from the photodissociation of HOCl and Cl_2, yield a massive destruction of ozone in the polar vortex. Incidently, other anthropogenic constituents, like bromine, play a similar, although less pronounced, role in the destruction of stratospheric ozone.

At stratospheric levels, the photodissociation of molecular oxygen by ultraviolet radiation at wavelengths of less than 240 nm produces atomic oxygen, which further combines with molecular oxygen to produce ozone

$$O_2 + h\nu \rightarrow O + O$$
$$O + O_2 + M \rightarrow O_3 + M.$$

Ozone is rapidly photodissociated back to atomic oxygen by photons in the wavelength range 200–300 nm

$$O_3 + h\nu \rightarrow O_2 + O.$$

'Odd oxygen' species (O, O_3) may recombine together according to the reactions

$$O_3 + O \rightarrow 2O_2$$
$$O + O + M \rightarrow O_2 + M.$$

The previous cycle, first identified by Sydney Chapman in the 1930s, must be completed by a set of other catalytic cycles evidenced in the last 20 years. A few minor constituents play a crucial role in destroying atmospheric ozone through catalytic cycles of the following type

$$X + O_3 \rightarrow XO + O_2$$
$$O_3 + h\nu \rightarrow O + O_2$$
$$O + XO \rightarrow X + O_2$$
$$\text{net: } 2\,O_3 + h\nu \rightarrow 3\,O_2.$$

Along this chain, the radicals X and XO catalyse the conversion of O_3 into O_2. The produced oxygen atom regenerates the catalyser X. The Chapman cycle is obtained for X = O. The hydroxyl radical, produced by the photodissociation of water vapor molecules, also plays a significant role (X = OH). Paul Crutzen found in 1970 that nitrogen oxides NO and NO_2 (X = NO), partly produced by agricultural activities and fossil fuel combustion, are quite efficient catalysers for ozone destruction. Mario Molina and Sherwood Rowland, who obtained together with P Crutzen the Nobel Prize for chemistry in 1995, have exhibited the crucial role of chlorine contained in CFCs on the chemistry of stratospheric ozone. Cl and ClO released by CFCs in the stratosphere are found to be still more efficient than NO in destroying ozone, through the previously mentioned catalytic cycle (X = Cl). Numerical models of stratospheric chemistry have been constructed in order to examine the consequences of all cycles previously described on the long-term evolution of ozone. The decrease of stratospheric ozone measured during the last 15 years is globally well reproduced by these models.

Bibliography

Banks P M and Kockarts G 1973 *Aeronomy* (New York: Academic)

Brasseur G and Solomon S 1994 *Aeronomy of the Middle Atmosphere* (Dordecht: Reidel))

Graedel T E and Crutzen P J 1993 *Atmospheric Change, an Earth System Perspective* (New York: Freeman)

Holton J R 1979 *An Introduction to Dynamic Meteorology* (New York: Academic)

Lewis J S and Prinn R G 1984 *Planets and their Atmospheres* (New York: Academic)

Eric Chassefière

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth's Interior

The interior of the Earth can be most simply described as a series of concentrically nested spheres, the main ones being, from the outside in, the CRUST, MANTLE, and CORE. This gross structure, recognized by seismologists early in the twentieth century, appears to be analogous to that of most silicate bodies—planets, satellites and probably the asteroids or their parent bodies—except the very smallest. The process of DIFFERENTIATION, by which these zoned structures were developed, probably resulted from extensive melting very early in the history of the solar system. The discovery that solid bodies are grossly differentiated was one of the most important geological findings of twentieth century space exploration, beginning with the Apollo expeditions to the Moon.

The Earth's interior will be described from the outside in, which is also the order of decreasing knowledge. The crust is now rather well-understood, or at least well described; the mantle can be described within certain constraints; and the core is understood least of all. Knowledge of the interior comes from a wide range of sources. The exposed interior is roughly the province of *geology*; study of the hidden interior, obviously almost the entire volume of the Earth, is the province of *geophysics*. Geologists study what can be seen, geophysicists what cannot, although there is complete overlap between the two fields.

Crust

There are two main divisions to the Earth's crust (figure 1), oceanic and continental. Continental crust can be summarized, simplistically, as old, thick and granitic. Oceanic crust can similarly be summarized as young, thin and basaltic. Furthermore, the oceanic crust is demonstrably mobile; as shown on the map, it is forming continuously by basaltic volcanism along mid-ocean ridges, and moving away from these ridges by the process of sea-floor spreading.

The outline just given of the Earth's crust is a subset, so to speak, of an even broader scheme, the now well-known theory of plate tectonics, which arose within about two years in the mid-1960s by the sudden coalescence of several lines of evidence. The Earth's crust, or more correctly its *lithosphere* (whose definition will be given shortly), is now believed to consist of a number of *plates*. Plates are segments of the lithosphere, relatively inactive and rigid, bounded in the simplest examples (e.g. the Nazca Plate) by some combination of oceanic ridges (or spreading centers), transform faults and subduction zones. The salient aspect of plate tectonic theory is that these plates are considered to be mobile, moving horizontally more or less in the manner proposed by Alfred Wegener in his classic continental drift theory. Plate tectonics alone fills many books, but its main features can be understood with the aid of figure 1. To begin our descent into the interior of the Earth, let us examine the vertical structure of the continental crust (figure 2).

Continental crust is an infinitely complex collection of igneous, sedimentary and metamorphic rocks. However, the cross section in figure 2 is offered as representing the most important features of the oldest and most fundamental part of this crust, the granite-greenstone terrains of Precambrian (older than about 600 million years) areas such as the Canadian Shield. Relatively few people have more than glimpsed such terrain, and even most geologists have only a general acquaintance with it. The reason is basically that continental crust as a whole is largely blanketed with relatively young rocks, such as the sandstones, shales and limestones exposed in the walls of the Grand Canyon. Furthermore, even Precambrian shields are criss-crossed with Proterozoic (2.5 billion- to 600 million-year-old) mountain belts or roots thereof. However, underlying most such mountain belts is rock of the sort shown here. Granite is the familiar pink igneous rock found on building fronts, cemetery headstones and monuments. Greenstone, less familiar, is essentially a greenish assemblage of metamorphosed volcanic rocks, originally basalts and related types.

The continental crust is thus underlain in most areas by rocks of the granite-greenstone type. From top to bottom, one will encounter successively belts of metamorphosed volcanics (greenstones), intruded by granitic rocks (many of which are not granites in the strict petrographic sense). At greater depths, these rocks are succeeded by what appear to be the metamorphic equivalents of granites, greenstones and sedimentary rocks, collectively termed granulites. These are dense, thoroughly recrystallized and highly deformed rocks.

A recent discovery, from seismic reflection profiling, is that unlike upper crustal rocks, and obviously unlike the deformed rocks of mountain belts, the lower crust granulites are largely flat-lying or within 20 to 30 degrees of the horizontal. These granulite layers were probably formed by a combination of deformation and recrystallization, acting on what were originally layers of sedimentary or volcanic rock. Readers in the United Kingdom can see (weather permitting) outstanding exposures of granulite along the northwest Scottish coast, around the village of Scourie.

The base of the continental crust is the Mohorovičić discontinuity, discovered by the seismologist of that name early in the century. The Moho, as it is commonly called, typically occurs at depths of 40 kilometers, but in areas such as the Andes or the Tibetan Plateau the crust is much thicker and the Moho may be 70 km or more deep. It is generally believed to be a compositional boundary, marked by a sudden increase in density as indicated by the increasing velocity of seismic waves. However, there is much to learn about this widely known boundary.

Oceanic crust differs from continental crust in almost every way. Its gross composition is basaltic, the basalt occurring, from the surface down, as lava flows, dikes (nearly vertical, sheet-like intrusive bodies), and near the base of the crust, the metamorphosed equivalents of these basaltic rocks. Newly formed oceanic crust can be seen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

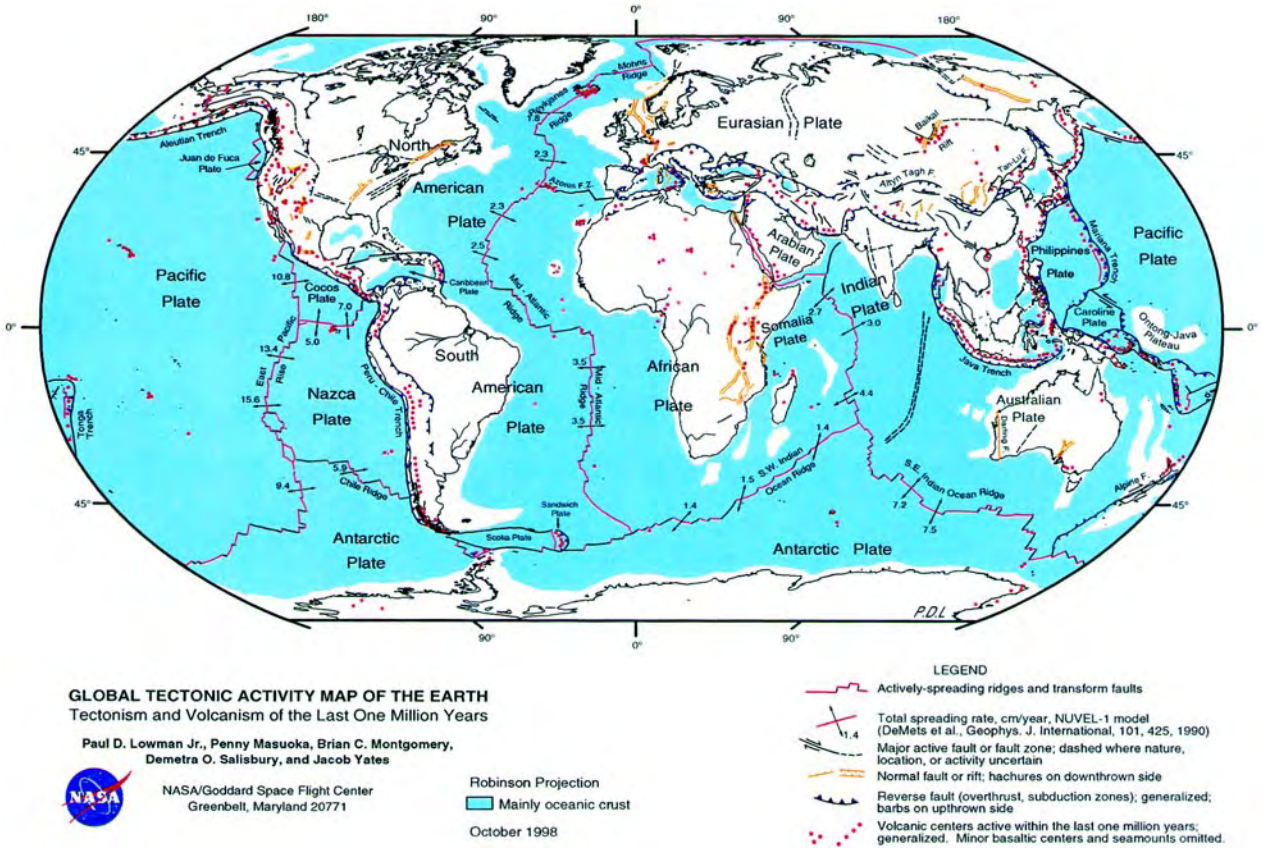


Figure 1. Tectonic and volcanic activity map of the Earth, covering the last one million years. Modified from Lowman (1982). **This figure is reproduced as Color Plate 24.**

in places like Iceland, located on the Mid-Atlantic Ridge. The crust under ocean basins is typically much thinner than under continents, the Moho generally lying at depths of 5 to 10 kilometers.

The origin of the oceanic crust is now, we believe, fairly well understood, an understanding that even skeptics will concede to be a major triumph for plate tectonic theory. In the context of this article, the dominant characteristic of oceanic crust is its youth and mobility. Because oceanic crust is eventually recycled through the mantle in subduction zones (figure 1), very little of it is more than about 200 million years old, in contrast to the 4 billion year age of the oldest exposed continental rocks, the Acasta Gneiss in Canada.

Mantle

Let us begin our descent into the mantle by returning to the deferred definition of *lithosphere*. The term crust is both a compositional and mechanical one, referring to the boundary between the silica-rich rocks of the crust and the silica-poor, iron- and magnesium-rich, rocks of the mantle. The term lithosphere in contrast is essentially a mechanical one, referring to an outer layer of the Earth that is relatively strong, elastic and rigid. This lithosphere includes, and is much thicker than, the crust.

Under ocean basins it varies from nearly zero at mid-ocean ridges to, very roughly, 50 to 100 km near subduction zones (where the oldest crust descends). However, under the old and stable parts of continents, such as the Canadian Shield, the lithosphere is much thicker, at least 200 km and in some estimates considerably more. Although the essence of the definition is mechanical, the subcontinental lithosphere in particular may have formed by extraction of basaltic magmas through geologic time, thus having a compositional connotation. In plate tectonic theory, horizontal crustal motions such as sea floor spreading and continental drift are considered the expression of motions (geometrically, rotations on a sphere) of the lithosphere, the thinner included crust being carried along with this lithosphere. (The term *plate*, to remind the reader, refers to the lithosphere, not just the crust.)

Below the lithosphere lies the *asthenosphere*. It is here that our ignorance of the Earth's interior becomes acute, for the thickness and extent of the asthenosphere, and in fact its very definition, are controversial. Physically, the asthenosphere is defined as a weak zone, or shell, characterized by low rigidity and strength, expressed as a decrease in the velocity of seismic compressional ('P') waves. In classic geophysics, the asthenosphere was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

considered the zone of flowage responsible for the well-known post-glacial rebound of Scandinavia and northern Canada. In plate tectonic theory, it is thought to be the low-resistance layer on which plates move. Several lines of evidence indicate that the asthenosphere is largely a thermal feature, related to either partial melting or at least decreased shear strength expressing increasing temperature with depth. However, some geophysicists, partly because of the relationship of asthenospheric depth to oceanic crust age, define it on the basis of geochemical characteristics that will not be covered here.

Our descent into the mantle will be guided by P J Wyllie's comprehensive diagram (figure 3), showing the main inferred characteristics of the Earth's interior all the way to the center some 6300 km down.

The composition of the mantle has been very slowly uncovered by a century of work in several fields. Since it is exposed almost nowhere—and there only in limited, highly metamorphosed form—mantle composition has been an unusually difficult problem. Some volcanoes bring up samples from depths of perhaps 200 kilometers, diamonds being the celebrated example, but these are our only direct source of knowledge. Seismic properties have traditionally been the main basis for estimates of mantle composition, although they provide essentially only constraints on composition. Basaltic and related magmas are clearly of mantle derivation, and some 75 years of theoretical and experimental petrology pioneered by N L Bowen in the 1920s have, combined with other lines of evidence, led to fairly detailed profiles of the mantle, shown in figure 3.

The upper mantle is believed to consist of iron- and magnesium-rich rocks similar to those found at the surface, with names unfamiliar even to many geologists, such as lherzolite. However, as shown in figure 3, pressure increases rapidly with depth; a thousand kilometers down, the pressure is roughly half a million times atmospheric pressure at the surface. With increasing pressure, minerals invert to a series of denser and denser polymorphs (i.e. same chemical composition, but different crystal structure), such as perovskite (an iron-magnesium silicate). They may dissociate to form new mineral assemblages, that is, new rocks of varieties never found at the Earth's surface.

These phase changes are the result of the increase in pressure and temperature. The melting point of most solids is raised by increasing pressure, in a complex way, and it is believed that the mantle is close to its melting curve below the asthenosphere. There are several seismically identified discontinuities in the mantle, of which the one 650 km down is the most prominent. This discontinuity, actually a zone some tens of kilometers thick, is thought by most petrologists to represent the transition from lherzolite to perovskite. Regardless of its actual nature, this discontinuity is geophysically important as the boundary between the upper and lower mantle.

One of the most controversial topics in geophysics is the nature of motion in the Earth's mantle. It has been speculated for most of the century that despite its solid nature, the mantle is in slow continuous motion, specifically by thermally driven convection currents. The confirmation of crustal mobility in the form of sea-floor spreading, and other lines of evidence, has greatly strengthened the concept of mantle convection. However, it is by no means clear just what pattern or patterns such currents follow. It was originally proposed, in the 1930s, that the entire mantle was convecting, a view still held by many eminent geophysicists. However, the 650 km discontinuity is clearly a density barrier to whole-mantle convection. The pattern may consist of upper and lower mantle convection cells.

Another complication is the possible existence of more localized mantle motions, plumes of hotter material rising slowly through the mantle, perhaps analogous to the popular 'lava lamps'. There is good seismic evidence for at least some mantle plumes, for example under the volcanically active Yellowstone National Park. However, it is not known where such plumes originate; some workers consider them to begin at the core–mantle boundary, and to rise some 2900 km almost to the Earth's surface.

One of the longest-known and best-confirmed phenomena of plate tectonics is subduction, the downward movement of slabs of oceanic crust and lithosphere complementary to sea-floor spreading (where crust is created, as shown on figure 1). These slabs must sink to several hundred kilometers depth, inverting to denser phases as they go, and contributing to the heterogeneity of the mantle. Subduction thus has the interesting implication that surficial geologic processes can influence the nature and behavior of the mantle at enormous depths.

Core

It has been known for centuries that the interior of the Earth must be much denser than the crust, since the specific gravity of the entire globe is about 5.5, in contrast to that of, for example, granite, which is about 2.8. Striking confirmation that the Earth could not be granite all the way through was provided by the discovery that RADIOACTIVITY was a major source of heat in the Earth. But the radioactivity of granite in the crust alone was enough to account for observed heat flow, which would be enough to melt the interior if it were all granite.

Seismologists Gutenberg and Weichert identified a major discontinuity about 2900 km down early in the twentieth century, which was rapidly recognized as the boundary of a discrete core. Its physical properties corresponded roughly to that of iron, and the existence of iron METEORITES demonstrated that iron could be segregated in the center of a planet. The core does not transmit shear waves, leading to the conclusion that it was liquid. However, further seismologic studies by I Lehmann in the 1930s showed that the inner part of the core, below about

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

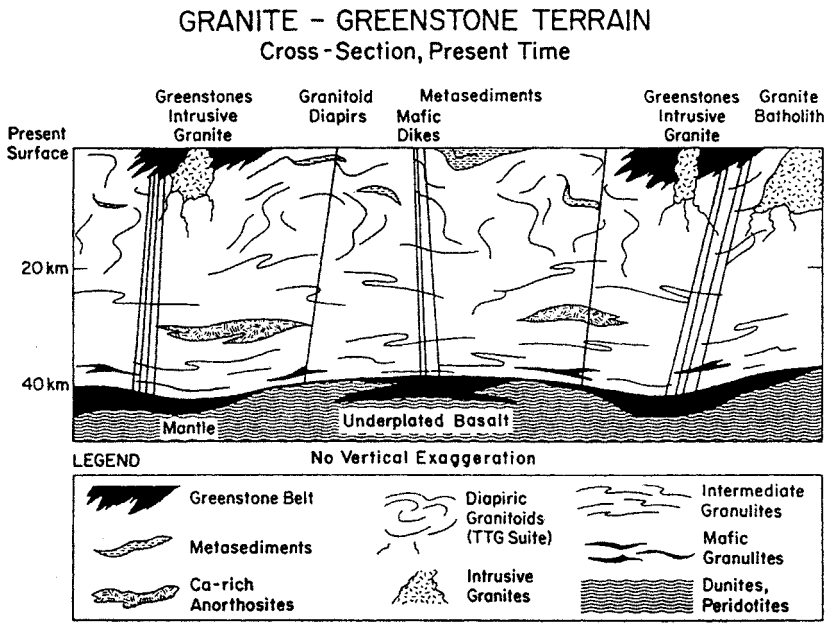


Figure 2. Structure and composition of continental crust as represented by granite-greenstone terrains (e.g. Superior Province, Canadian Shield). From Lowman (1989).

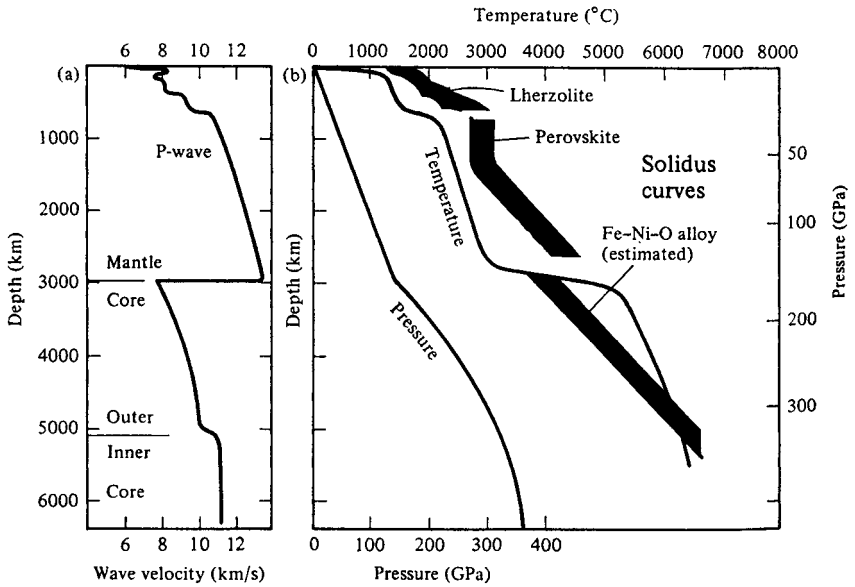


Figure 3. (Left) Seismic wave velocities (P-waves) in the interior of the Earth as a function of depth. (Right) Inferred phases in the interior of the Earth as function of depth, pressure, and temperature. From Wyllie (1992).

5000 km depth, was solid, an iron (presumably) sphere roughly the size of the Moon.

This general physical picture of the Earth's core is agreed upon by most geophysicists. However, continuing progress in seismology and geomagnetism are refining it rapidly. The main geomagnetic field is now plausibly explained as the result of convection currents in the liquid outer core, which, cutting the lines of force of the Earth's field, constitute a self-regenerating dynamo mechanism. This is not perpetual motion; interested readers should consult the listed references.

There is now evidence that the core–mantle boundary has considerable relief, on the order of a kilometer, perhaps related to the rise of mantle plumes if such do in fact exist. These irregularities may cause slight but real variations, over periods of years, in the EARTH'S ROTATION. It may be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

helpful to visualize the Earth not as a solid rotating sphere, but as two such spheres, the inner one being the Moon-sized solid core separated from the outer one (the solid mantle) by the liquid segment of the core. Complications in the motion of such a decoupled system can easily be imagined.

Estimates of core compositions have been considerably refined; nickel is almost certainly a major alloying element, with traces of other siderophile elements such as platinum that are depleted in the Earth's crust and upper mantle. There are also probably substantial amounts of a lighter alloying element in the core, probably oxygen or silicon.

Summary

Our knowledge of the interior of the Earth has increased enormously through the twentieth century. Furthermore, we now know something of the interiors of other planets, providing the basis for comparative studies (see PLANETARY INTERIORS). However, the interior below a few tens of kilometers will be inaccessible to direct study for centuries to come, perhaps forever, and study of the inside of the Earth will be a fruitful field for geophysical research for many years to come.

Bibliography

Anderson D L 1992 The Earth's interior *Understanding the Earth* ed G C Brown *et al* (Cambridge: Cambridge University Press) ch 3

Anderson D L 1995 Lithosphere, asthenosphere, and perisphere *Rev. Geophys.* **33** 125–49

Brown G C, Hawkesworth C J and Wilson R C L (eds) 1992 *Understanding the Earth* (Cambridge: Cambridge University Press) p 551

Dziewonski A M, Forte A M, Su W and Woodward R L 1993 Seismic tomography and geodynamics *Relating Geophysical Structures and Processes: the Jeffreys Volume (Geophysical Monograph 76, IUGG Volume 16)* ed K Aki and R Dmowska (Washington, DC: American Geophysical Union) pp 67–105

Hide R, Clayton R W, Hager B H, Spieth M A and Voorhies C V 1993 Topographic core–mantle coupling and fluctuations in the Earth's rotation *Relating Geophysical Structures and Processes: the Jeffreys Volume (Geophysical Monograph 76, IUGG Volume 16)* ed K Aki and R Dmowska (Washington, DC: American Geophysical Union) pp 107–20

Jeanloz R 1993 Chemical reactions at the Earth's core–mantle boundary: summary of evidence and geomagnetic implications *Relating Geophysical Structures and Processes: the Jeffreys Volume (Geophysical Monograph 76, IUGG Volume 16)* ed K Aki and R Dmowska (Washington, DC: American Geophysical Union) pp 121–7

Lowman P D Jr 1982 A more realistic view of global tectonism *J. Geol. Ed.* **30** 97–107

Lowman P D Jr 1984 Formation of the earliest continental crust: inferences from the Scourian Complex of northwest Scotland and geophysical models of the lower continental crust *Precambrian Res.* **24** 199–215

Lowman P D Jr 1989 Comparative planetology and the origin of continental crust *Precambrian Res.* **44** 171–95

Rudnick R L and Fountain D M 1995 Nature and composition of the continental crust: a lower crustal perspective *Rev. Geophys.* **33** 267–309

Wyllie P J 1992 Experimental petrology: Earth materials science *Understanding the Earth* ed G C Brown *et al* (Cambridge: Cambridge University Press) ch 4

Paul D Lowman Jr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth's Rotation

In the absence of internal or external perturbation, the Earth would rotate in space around its center of mass at a uniform angular velocity and around a fixed rotation axis lying at 23°26′ with respect to the ecliptic pole axis and directed approximately towards the northern star. The mean motion is a rotation in space every 23 h 56 min 4.10 s but, in reality, the Earth being composed of various non-rigid parts and not being isolated in space, the rotation is perturbed by a wide variety of geophysical and astronomical processes. These perturbations can be investigated in terms of causes and effects and are subject to theoretical investigations (see also EARTH'S ROTATION: THEORY).

Irregularities of the Earth's rotation—historical perspective

It is worthwhile to approach the Earth's rotation from a historical perspective. The chronology of the discovery of the Earth's rotational irregularities also reflects the progress of astrometric and geodetic techniques and time measurement.

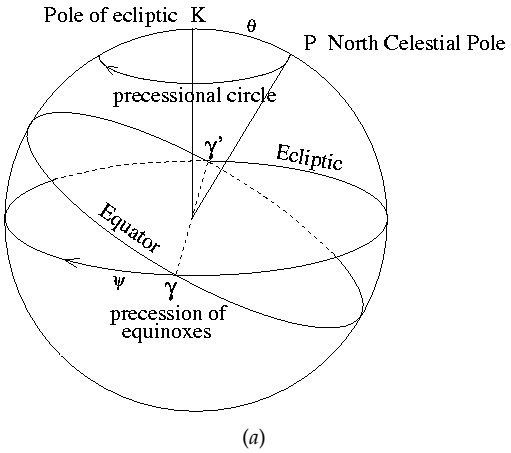
Precession

The first irregularity discovered in the Earth's rotation, and the largest one, does not concern the angular speed itself but the direction of the rotation axis with respect to the stars. In the Western world, as early as the 2nd century BC, the Greek astronomer HIPPARCHUS discovered that the vernal equinox, that is the direction of the Sun when it crosses over the equatorial plane towards the northern hemisphere, precesses with respect to the background stars. In other words, the equatorial plane turns clockwise at a fixed inclination of 23°26′ (obliquity) with respect to the ecliptic (figure 1(*a*)). The rate is of about 50″ yr^{-1} and thus the vernal equinox moves through all the constellations of the zodiac over the 25 800 yr precession period. At present the vernal equinox is in the constellation Pisces and is slowly approaching Aquarius, as shown in figure 1(*b*).

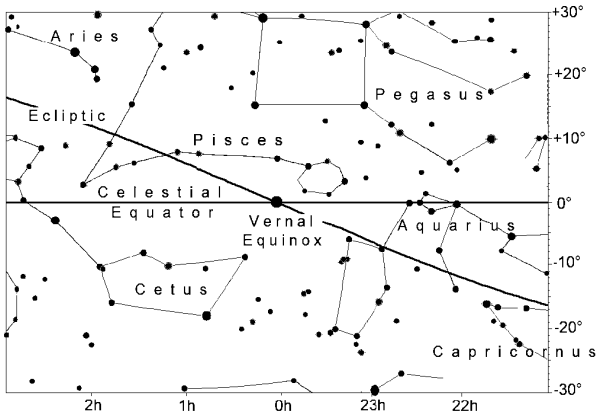
The diurnal rotation axis is perpendicular to the equatorial plane. Because of the PRECESSION, this axis describes clockwise a cone around the pole axes of the ecliptic, completing a turn in 25 800 yr, as shown in figure 1(*a*). At present the Earth's rotation axis points towards the Pole Star or North Star, but in 13 000 yr the precession of the rotation axis will mean that the bright star Vega in the constellation Lyra will be approximately at the North Celestial Pole, while in another 25 800 yr Polaris will once again be the Pole Star.

Nutation

Other irregularities in the Earth's rotation axis motion with respect to the stars are periodic oscillations of the rotation axis, called NUTATIONS, superimposed on the precession and predicted in the *Principia Mathematica* of Newton. These periodic oscillations are of much smaller amplitudes, no larger than 1′. Therefore they were not discovered until the advent of astrometric techniques at the beginning of the 18th century (see ASTROMETRY: HISTORY).



(*a*)



(*b*)

Figure 1. (*a*) Precession of the Earth's rotation axis (P) around the ecliptic pole (K). (*b*) Precession of the vernal equinox through the constellations.

In 1748, the English astronomer Bradley first observed the most important term of the nutation, associated with the regression of the lunar orbit's line of nodes, which has a period of 18.6 yr and an amplitude of the order of 10″. Other terms, up to several thousand in most accurate models, are at least 10 times smaller.

Polar motion

EULER predicted in 1765 that the rotation axis would also show a motion with respect to an Earth-fixed reference frame if the axis of rotation is not coincident with the principal axis of inertia—even if the Earth is isolated and a rigid body. The American astronomer Chandler took that important step with the discovery, in 1891, of that motion from observations of the geographical latitude of astronomical observatories. However, the main component is a 433 day periodic term, differing from that predicted by Euler (305 days) based on the Earth's ellipticity, because of the non-rigidity of the Earth. It has

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

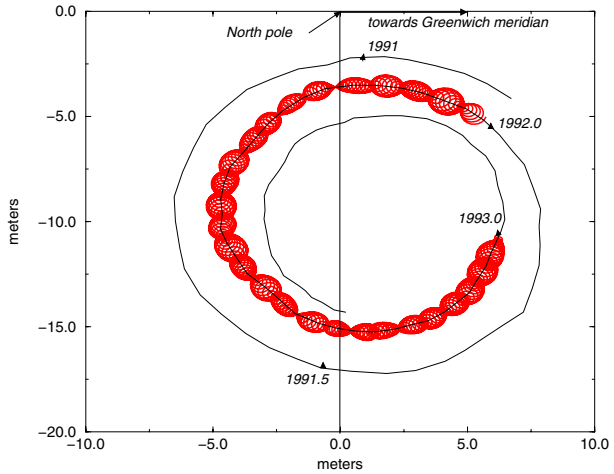


Figure 2. Motion of the Earth's rotation axis on the crust. Notice the long periodic polar motion composed of the annual term and the Chandler wobble, and the diurnal clockwise oscillations due to precession–nutation.

a variable amplitude up to 0.3″ (corresponding to 10 m on the surface of the Earth). As well as the Chandler wobble, an important annual term up to 0.1″ exists and is interpreted as the effect of seasonal mass displacements. Moreover, the precession–nutation gives rise to a small retrograde diurnal polar motion of the rotation axis in the Earth, as shown in figure 2 (see also POLAR MOTION AND LENGTH OF DAY).

Variations in length of day

Earth's diurnal rotation was humanity's time keeper until the discovery by de Sitter (1927) and Spencer Jones (1939) of the secular deceleration of the Earth's rotation speed and afterwards of its seasonal variations by Stoyko (1937) thanks to the improvement of precision clocks. Actually the rotation rate of the Earth's crust exhibits complicated fluctuations in speed to several parts in 10^{-8} (corresponding to a variation of a few milliseconds (ms) in the length of day, or LOD). These fluctuations are mostly seasonal (1 yr, half a year cycle) and also exhibit oscillations of all sorts of longer and shorter periods, such as the so-called 'decade fluctuations'. Moreover, analyses of solar eclipse observations recorded by Babylonians, Greeks, Arabs and Chinese indicate that the rotation rate decreases secularly, which causes an increase of the LOD of about 2 ms century^{-1}.

All these variations are the consequences of the angular momentum balance. This involves changes in the Earth's rotation due to the torques acting on the solid Earth as well as those due to changes in the mass distribution within the solid Earth, which modify its inertia tensor. This will be described in more detail in the section 'Dynamical description'.

Effect of the Earth's rotation on science and everyday life

As will be seen later, measuring the Earth's rotation involves nowadays large-scale, expensive observing techniques. Before we describe the state of the art, it is necessary to stress the importance of knowledge of the Earth's rotation by underlining its consequences in fundamental science, technology and everyday life.

The Earth's rotation and time

The Earth's rotation rules our whole life through the alternation of the days and nights. Therefore it constitutes the most natural unit for reckoning time. Inhabitants of the Earth are much more concerned about the diurnal cycle of the Sun than those of the star. From an Earth-bounded point of view, the Sun is moving slightly more slowly than a star, because the Sun is orbiting counterclockwise annually around the Earth whereas the diurnal sidereal rotation is clockwise. Hence the solar day is about 4 min longer than the sidereal day. Because of the variations of the Sun's velocity (or equivalently of the orbital velocity of the Earth in a heliocentric reference frame) and because of the OBLIQUITY OF THE ECLIPTIC, the solar day is subject to small variations of up to 30 s during the year. A mean solar day has been defined and conventionally taken as 86 400 s or 24 h. The diurnal cycle of the Sun constitutes the fundamental unit of the majority of calendars. Differences between CALENDARS concern the choice of greater cycle for reckoning the days (the month and the year), that is for establishing the dates. The Gregorian calendar, which constitutes nowadays the international reference, has been designed in order to keep the coincidence between the seasons and the occurrences of the equinoxes and solstices, of which the directions in space vary owing to the precession. As the tropical year is not a multiple of the number of days, the year of the Gregorian calendar contains an integer number of days but is subject to variations from one year to another (365 or 366 days), so that the mean year remains equal to the tropical year.

The succession of the mean solar days enables us to define a time scale, the so-called UNIVERSAL TIME. Thanks to the advent of the atomic clock, the Earth's rotation rate has been shown to be not uniform and consequently Universal Time has been given up as a time scale reference. The international time unit, the second of atomic time, has been chosen to be equal to the average of the second of the mean solar day for the epoch 1900. The INTERNATIONAL ATOMIC TIME scale, the Temps atomique international (TAI), is derived naturally with respect to the Universal Time (UT1) scale. In order to maintain the coherence between atomic time and universal time, an intermediary atomic time scale is maintained for civil use. This is the *Coordinated Universal Time* (UTC), which differs from the TAI by an integer number of seconds so that it never shifts from the UT1 by more than 0.9 s.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Earth's rotation is crucial for physical theories

The Earth's rotation provides a validation of Newtonian dynamics. Newton asserts a distinction between inertial reference frames and non-inertial frames, undergoing acceleration with respect to the inertial frame, and in which the so-called non-inertial effects, such as centrifugal and Coriolis effects, appear. Assuming that a star-fixed reference frame is inertial, the rotating crust is a non-inertial frame. In turn the mass elements of the Earth should be subject to centrifugal effects, implying an equatorial bulge and a lower gravity at the equator than at the poles. This was shown in the 18th century by geodetic measurements and time pendulum recording at different latitudes and constituted a very important validation of Newtonian concepts. The second proof was obtained with the famous FOUCAULT'S PENDULUM designed by him in 1851, which exhibits an apparent turning of its oscillation plane under the Coriolis effect due to the Earth's rotation.

Geophysical effects of the Earth's rotation

The Earth's rotation strongly influences the atmosphere. First of all, the succession of the day and night induces daily fluctuations of temperature, and in turn changes of the pressure, and wind fields. Secondly the air motions are ruled partly by the Coriolis effect. This allows understanding of the wind patterns, especially the direction of the predominant winds with latitude.

The Earth's rotational irregularities have a role in geophysical phenomena. As an example, the decreasing rate of the Earth's rotation could induce a decrease of the Earth's flattening because of the decreasing centrifugal effect.

Geometric description—The Earth orientation parameters

The above considerations focus on the Earth's angular velocity vector. However, the astrometric or geodetic observations do not depend directly on this quantity but on the orientation of the ground-based stations and the celestial reference frame involved in the observations. Numerical methods, such as least-squares analyses, allow us to determine the orientation from the rough observations. In order to harmonize the determinations carried out around the world, the international scientific community has adopted a unique set of Earth orientation parameters (EOPs). The basic idea is to separate the modeled orientation changes from the unknown ones. The model provides the effect of mean diurnal rotation and of the precession–nutation. The difference between the real orientation and the model is described by five orientation parameters.

Firstly, two *celestial pole offsets* compensate for the defect in the precession–nutation model of the polar geographic axis; they are constrained by convention to contain oscillations having periods greater than a few days. The precession–nutation model and the celestial pole offsets define the spatial direction of an axis whose intersection with the celestial sphere is called the Celestial Ephemeris Pole. The word 'ephemeris' refers to the fact that the position of the pole in space is specified by the ephemeris provided by the conventional model for nutation.

The third parameter, the universal time variations $UT1 - UTC$, integrates the irregularities of the Earth's rotation speed. The Universal Time UT1 is the time scale built on the Earth's rotation angle, assuming that the rotation angle is proportional to UT1. As seen previously, the Coordinated Universal Time UTC is an atomic time scale, evolving uniformly with respect to the international atomic time (TAI).

Finally the two polar motion coordinates, x and y, provide the position of the Celestial Ephemeris Pole in the terrestrial frame, that is the shift between the Celestial Ephemeris Pole and the North Geographic Pole, but expressed in the terrestrial frame. It can be shown theoretically that the Celestial Ephemeris Pole represents as a first approximation the diurnal average of the Earth's rotation axis. Therefore the polar motion of the Celestial Ephemeris Pole, restricted to its long periodic variations, represents the motion of the rotation axis on the crust.

Actually three angles are sufficient to describe the defects in the *a priori* orientation of a terrestrial frame in a celestial frame. However, these angles include variations from several days or years (precession–nutation, long periodic polar motion) to 1 day and less. To estimate these parameters would require very intensive observations, which at present is not routinely achieved but will be soon as technology advances.

Monitoring the Earth's orientation

Evolution of the techniques

Astrometric and geodetic techniques provide the orientation as a function of time of the terrestrial frame with respect to the celestial frame from which the EOPs, including UT1, polar motion and precession and nutation quantities, can be derived. International services, successively since 1900 the *International Service of Latitude* (ILS), the *International Service of Polar Motion* (IPMS) and *the Bureau International de l'Heure* (BIH), have been in charge of collecting and combining the measurements in order to provide the best estimates of polar motion on a regular basis. The parameter UT1 was provided on a regular basis by the BIH from 1967. All measurements were based on optical observations done by Danjon astrolabe, or by transit, zenithal and circumzenithal instruments; they were provided for a 5 day interval, but the atmosphere limited their precision to 0.01″. Since 1972, several space geodetic techniques have been used for this purpose by BIH, beginning with satellite Doppler tracking and lunar laser ranging (LLR) and then, in 1978, satellite laser ranging (SLR) and very long baseline interferometry (VLBI) on extragalactic radiosources. Other satellites techniques, such as the Global Positioning System (GPS) and DORIS, have been considered in recent years. All these techniques provide time or frequency measurements. This is the primary information used by analysis centers for determining the EOPs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

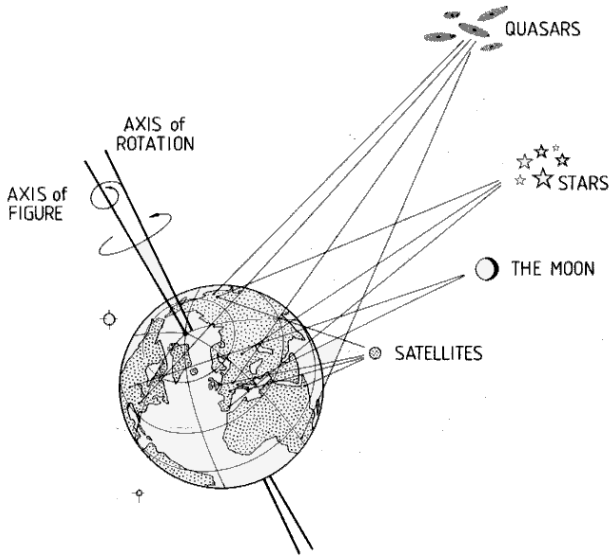


Figure 3. The techniques for observation of the Earth's rotation: the beacons are stars for astrometry, quasars for VLBI, satellites for SLR and GPS, Moon for LLR.

Today the astrometric methods have given way to the space geodetic techniques in routine Earth rotation measurements (see figure 3).

Since 1988, the international service in charge of the EOPs and the corresponding reference frames is the *International Earth Rotation Service* (IERS) whose *Central Bureau* is at Paris Observatory. The present accuracy is of the order of 0.2 mas (1 cm) for daily EOP determinations and can reach 0.5 mas for interval of the order of hours thanks to intensive VLBI or GPS sessions. This has led to a dramatic improvement in the determination of the EOPs and consequently of the knowledge of the details of the Earth's rotation and therefore of global geophysical parameters and geophysical processes influencing the Earth's rotation.

Modern geodetic techniques

In VLBI the basic quantity is the delay in wave front arrival at two radio telescopes from distant quasars. In the ideal case, this delay would be only a function of the orthogonal projection of the baseline between the two telescopes on the direction of the radiosource. The real case is more complicated because this delay contains non-geometrical effects, mainly due to clock synchronization and bending of the wave front in the atmosphere. Anyway, information on the Earth's orientation is contained in VLBI delays. Its modeling depends on many parameters which have to be estimated simultaneously. Until 1998 VLBI observation sessions have been carried out during 24 h at a few days' interval. They generally involved a set of baselines, observing consecutively several radiosources constituting the International Celestial Reference Frame. A distinction is established between arc parameters which are constant during the session (such as the coordinates of the radiosources, the EOP) and the local parameters (such as the clock parameters, the atmospheric delay) which differ from one session to another. Wave front arrivals are recorded on magnetic tapes independently at each radio telescope. All tapes are shipped to a central facility where they are cross-correlated to determine the series of differences in arrival time (the delay) and the rate of change of the delay (the delay rate) on the VLBI baselines (see figure 4).

Laser ranging techniques are currently used on satellites (satellite laser ranging or SLR) and the Moon (lunar laser ranging or LLR). This provides information on exactly how far the satellite is from the station at a particular time, which, in turn, can be used to determine the Earth's orientation in space. A very short laser pulse is transmitted from a telescope in a ground station, and it is retroreflected by a corner cube reflector on a satellite or the Moon back to the station. The round trip time is measured, which is equivalent to measuring the distance. It is now possible to accurately measure this up to 50 ps or better, which is equivalent to an accuracy of 1 cm or less.

GPS uses a constellation of satellites that broadcast signals to stations on the Earth. From the time delay of the broadcast signal and a precise determination of the satellites' orbits, the position of the station can be determined. A network of approximately 30 permanent stations is operated in order to help to establish a terrestrial reference frame. This network can be used to detect changes in the rotation rate of the Earth (the excess LOD) and in the orientation of its axis of rotation (the polar motion). High-frequency variations in the length of day are determined daily and are precise to $\pm 20\ \mu$s day^{-1}.

Comparison of the techniques

Table 1 provides a comparison of the performances techniques for monitoring the Earth's rotation. It turns out that the VLBI technique is the most efficient because the celestial reference frames realized by VLBI analyses are the most stable, and in turn the long periodic oscillations in the Earth's orientation, especially the precession–nutation and the several days' fluctuations in the LOD, are well determined. The international celestial reference frame itself consists of a set of radiosource coordinates provided by VLBI analyses. Satellite techniques do not provide such a stable reference frame. Because of many non-modeled effects in satellite orbits, the corresponding reference frame is only suitable for monitoring short-period terms of the Earth's orientation changes. Polar motion, which appears in a celestial reference frame as a quasi-diurnal counterclockwise oscillation of the geographic axis, is very well determined by satellite techniques, at the same accuracy level as VLBI.

The satellite techniques must be connected to an absolute UT1 value provided by an independent technique such as VLBI. This is necessary because a uniform rotation of the satellite orbit nodes cannot be distinguished from the Earth's rotation. However, because VLBI does not operate continuously, GPS can give better real-time LOD estimates to fill in the gaps.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

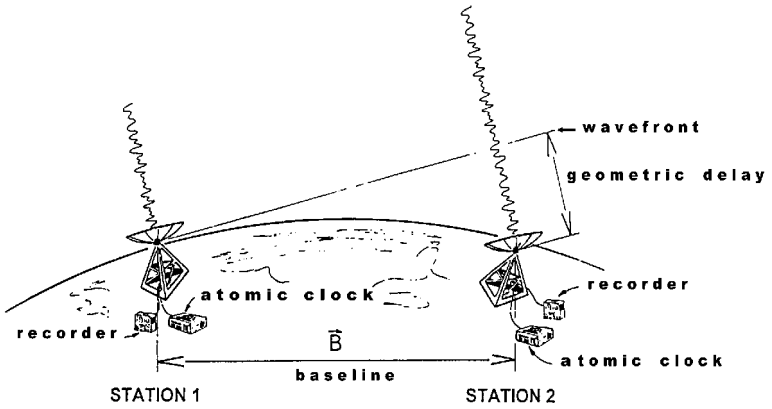


Figure 4. Observation of a quasar by VLBI.

Table 1. Potentialities of modern geodetic techniques according to the nature of the Earth's orientation parameters.

	VLBI	SLR	LLR	GPS
Precession–nutation	Up to 20 yr terms: 0.2 mas	Not currently applicable	Not currently applicable	Up to 13.6 day terms; VLBI necessary
Universal time	0.01 ms	0.02 ms	0.2 ms	0.02 ms; VLBI necessary
Polar motion	0.2 mas	0.2 mas	Not applicable	0.2 mas

Accuracy level of the determination is also reported. Unit for angles is milliarcsecond (mas), unit for UT1 is millisecond (ms).

Dynamical description

The dynamical causes of the Earth's orientation changes can be classified in two kinds: astronomical ones and geophysical ones. The astronomical causes, at our present state of knowledge, are reduced to the torque produced by the tidal gravitational attraction of celestial bodies on the Earth. The geophysical causes are associated with mass displacements in the Earth's layers (solid core, fluid core, mantle, oceans and atmosphere) or with interactions between the Earth's layers of an electromagnetic and a mechanical nature (gravitation, pressure and viscosity).

Let $\vec{H}$ be the Earth's angular momentum and $\vec{L}$ the external torque acting upon the Earth. In the geocentric quasi-inertial frame I (its axes, centered at the geocenter, do not rotate in space), the angular momentum balance is written as

$$\left(\frac{\mathrm{d}\vec{H}}{\mathrm{d}\vec{t}}\right)_I = \vec{L}. \qquad (1)$$

It is more convenient to consider a body-fixed frame **K**. Let $\vec{\omega}$ be the Earth's angular velocity vector; in **K** equation (1) provides the dynamic Euler equation:

$$\left(\frac{\mathrm{d}\vec{H}}{\mathrm{d}\vec{t}}\right)_K + \vec{\omega} \times \vec{H} = \vec{L}. \qquad (2)$$

The Earth's angular velocity vector consists of the constant diurnal rotation of the figure axis, that is the vector $(0, 0, \Omega)$ in the body-fixed system, and small perturbations $(\Omega m_1, \Omega m_2, \Omega m_3)$, m_1 and m_2 being for polar motion and m_3 for the relative variations in the Earth's rotation.

Solving equation (2) for a rigid body without any external torque results in a rotational eigenmode, the Eulerian wobble.

If one considers an elastic or inelastic body, we have to take into account the motion of each body element with respect to the system **K**. This causes variable increments of the moment of inertia c_{ij} and an additional angular momentum vector with components (h_1, h_2, h_3) in **K**. Expressing equation (2) as a function of the Earth's rotation vector, the Earth's inertia matrix, the additional angular momentum due to the non-rigidity of the Earth and the external torque and neglecting the cross-terms of the increments m_i, c_{ij}, h_i provides the so-called Liouville equations. The changes in the Earth's rotation and polar motion thus appear as the response of a differential system to a three-dimensional excitation vector. The excitation mechanisms result from a change in the Earth's angular momentum due to the application of an external torque, from a change in the Earth's inertia tensor and from internal redistribution of its angular momentum. Polar motion appears to be mainly excited by mass redistribution and LOD (i.e. m_3) by mass motions. Because of its linearity, such a system enables us to uncouple the sources of perturbations such as the effects of mass redistribution within the different parts of the Earth and the astronomical torque.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The classical way to solve the equations of the Earth's rotation involves two steps. Firstly the equations are solved for a rigid Earth model, only accounting for the astronomical forcing. Then one estimates to what extent this 'rigid solution' is modified by introducing the non-rigid structure of the Earth and geophysical effects such as air mass displacements. These two steps are described in more detail in the following.

Astronomical forcing

The Moon and to a lesser degree the Sun exert a gravitational tidal torque on Earth's equatorial bulge. The Moon contributes more than twice as much as the Sun owing to its proximity, which compensates for its smaller mass. This torque, lying in the equatorial plane, exhibits in an inertial system spectral components ranging from half a day to several thousand years, according to the periodic variations in the orbital elements, causing in turn variations in the direction of the geographic polar axis and rotation axis. These variations are classically described by two Euler angles, a *precession angle* ψ in the ecliptic plane and an inclination *angle* θ or obliquity, giving the offset between the Earth's axis (geographic or rotation axis) and the axis of the ecliptic poles (see figure 1(*a*)).

In the case of a rigid Earth model the problem of the precession–nutation is reduced to a purely astronomical problem. Solving the Liouville equations, we obtain the rigid Earth's angular velocity vector $(\omega_i^{\text{rigid}})_{i=1,2,3}$ as a function of the external torque and of the dynamical flattening of the Earth. By integration of kinematical relations linking the component of the angular velocity vector to the time derivative of the EOPs, the Euler angles of the geographic axes with respect to the axes of the quasi-inertial frame can be derived. They provide the variations in ecliptic longitude ψ and obliquity θ.

The quasi-stationary torque component, that is the term turning clockwise with sidereal period for a terrestrial observer, induces secular variations of the Euler angles. The most important concerns the angle ψ and amounts to $50.29''\ \text{yr}^{-1}$. This is the so-called *astronomical precession*. In turn the rotation axis describes, as a first approximation, a cone around the axis of the ecliptic poles in 25 800 yr. The periodic variations of the tidal torque, ruled by the motion of the Sun and the Moon in a geocentric reference frame, induce periodic variations in the rotation axis direction, the so-called *astronomical nutation*. At equal torque amplitudes the effect is more important for greater periods. Indeed the longer is the period the longer is the duration during which the periodic variation can take place. Mathematically the Euler angles results from the torque integration and hence from the division of any spectral component by its frequency. This explains why the biggest nutation terms are not necessarily induced by the largest torque components. As an example, the largest nutation term, in 18.6 yr, is a quasi-circular term of amplitude about $10''$, which is more than 200 times the amplitude of the term in 13.6 days, induced by the strongest tidal component. Moreover, because of the planetary perturbations on the Earth's orbit (i.e. the ecliptic), the obliquity is decreasing at a rate of $0.46''\ \text{yr}^{-1}$ (i.e. the axis of rotation is less and less tilted).

In the case of a rigid Earth model, the angular velocity is constant except for very small variations of diurnal and semidiurnal periods and amplitudes up to 2×10^{-6} s in UT1, due to the effect of the non-zero component of the luni-solar tidal torque in the equatorial plane.

Geophysical effects

The case of a real Earth is more complicated than considered in the previous section.

First, there are deformations under the influence of tidal gravitational forcing and centrifugal forcing which induce moment of inertia changes. The inertia changes of tidal origin are responsible for a modification of the amplitudes of nutations as well as for the zonal variations of the Earth's rotation at the same periods as the nutations and amplitudes up to 1 ms in UT1, both effects being scaled by the elasticity factor (Love number k) of the mantle. The elastic deformations of centrifugal origin are responsible for the major part of the modification of the Chandler period as compared with the Euler period.

Second, the fluid core and the mantle do not rotate rigidly; the fluid core naturally has a clockwise nearly diurnal wobble with respect to the crust, because its main inertia axis does not coincide with the mantle inertia axis and this causes by reaction a similar effect on the mantle. The period of this rotational free motion, also referred to as the free core nutation, is strongly dependent on the flattening of the core–mantle boundary and has been shown to influence the near-resonance tidal amplitudes. In a celestial frame, the free core nutation would appear as a clockwise oscillation. Theory predicts resulting perturbations of clockwise nutations of a rigid Earth, as confirmed by VLBI observations. Solving for the rotation of a non-rigid Earth model thus requires us to write Liouville equations for the mantle and the core separately and to solve both these equations simultaneously. Finally the angular velocity vector of the mantle $(\omega_i)_{i=1,2,3}$, induced by the external forcing, is expressed as a function of the rigid solution $(\omega_i^{\text{rigid}})_{i=1,2,3}$. Present models of precession–nutation, taking into account the rigid solution and non-rigid effects such as the presence of a fluid core inside the Earth with a flattening that deviates from hydrostatic equilibrium, agree with observations at the submilliarcsecond level. The value of the dynamical flattening of the Earth, which is a fundamental parameter of the models, is derived from the observed precession rate.

In addition, we have to consider the displacement of masses inside the Earth's superficial layer. These geophysical processes influence the Earth's rotation as much as the Moon and the Sun, but their energy is mostly released in the long periodic range (several days) with respect to the crust, in contrast to the gravitational influence which is daily retrograde. Most of them are independent of the Earth's rotational perturbation;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

they are triggered by thermal, tidal and electromagnetic influences. However, the variation of the centrifugal forces associated with the wobble of the rotation axis induces a change of the water height in the oceans, called the 'pole tide'. This phenomenon plays a fundamental role in accounting for the period of the Chandler wobble.

The changes of air and oceanic mass distribution involve variations of the Earth's rotational speed as well as a long periodic displacement of the rotation axis with respect to the crust. Annual components up to 2 ms in the LOD and up to 200 mas in pole motion are induced by the atmospheric circulation. The observed subseasonal components (10–100 days) can be explained by the combined effect of the atmosphere and the oceans. The most common approach for computing the effects of the fluid layers is based on the angular momentum time series of these layers (quantities c_{ij} and h_i) and the Liouville equations. Meteorological centers provide us with atmospheric angular momentum (AAM) series from pressure and winds assimilation models. The oceanic data are much less common, so that the oceanic angular momentum results mostly from modeling. The correlation between LOD and AAM is so well established that forecasts of atmospheric angular momentum provided by the major centers are now being investigated for prediction of the Earth's rotation.

Some geophysical forcing of larger periods (more than several years) can be expected from the observations. This might involve the superficial layers (for instance the El Niño cycle in the atmosphere and in the ocean) and the coupling between the core and the mantle. Such mechanisms can explain the decade-long fluctuations of the LOD. The major contribution to secular change in the LOD is attributed to tidal friction which transfers Earth's spin angular momentum to the lunar orbit, the complementary contribution originating from the effect of the post-glacial rebound on the Earth's principal moment of inertia.

The observed secular motion of the pole with respect to the crust, called the 'secular pole trend', of about 0.3″ century^{-1} in the direction about 70°W is mostly attributed to mass redistributions accompanying and following the termination of the Pleistocene glaciation 20 000 yr ago. Other geophysical causes are also under consideration such as a global rise in sea level or trends in groundwater storage and recent glacial activity.

It can be considered that variations in the Earth's rotation (precession, nutation, polar motion and LOD) are now quite well explained by the various geophysical effects described above (mass redistributions, influence of the fluid core, forcing of the atmosphere and the oceans, etc), except for a very puzzling component, the Chandler motion of the pole. As seen above, this component has the period of a free rotational mode of the Earth, about 433 days. The theory shows that, in the absence of any external torque, the rotation axis turns around the figure axis in the terrestrial frame with this period. As the Earth is not a perfect elastic body, this motion should be dissipated in thermal energy. However, no damping of the Chandler motion has been observed. Consequently it is excited by some mechanism, which is not yet elucidated. Neither the combined effect of the atmosphere and the oceans nor earthquakes seem to give the appropriate forcing.

Recent progress

Diurnal and subdiurnal determination of the polar motion and universal time

Vast progress in astrometric and geodetic observing techniques has taken place in the last few decades, as seen below. The progress is not only restricted to the increasing accuracy of the EOPs. Since the 1990s the intensification of geodetic observations by VLBI and GPS has enabled the determination of subdaily variations of the polar motion and the Earth's rotation speed. Since 1999 VLBI observations have been done continuously and more intensively by different antenna networks working alternatively. New phenomena have been put forward, which are diurnal and semidiurnal changes in the polar motion and in the Earth's rotation speed. These variations are mostly explained by the oceanic tides. Other components of non-tidal origin might be detected, but no firm statement can be made yet.

Free core nutation

Furthermore, the modeling of Earth's orientation changes has been improved. Until the 1980s the conventional precession–nutation model was that of a rigid Earth. However, the non-rigidity effects are up to 100 mas, which are 1000 times the accuracy of the VLBI observational results. Since 1980 much interest has thus been focused on the refinement of the nutation model by taking into account the deformation of the mantle as well as the interaction processes between the fluid core and the mantle. The parameters of such a model are of primary geophysical interest, such as the flattening of the fluid core. Models constrained by VLBI observation allow one to deduce these parameters. From the observed period of the free core nutation (430 days), it has been found that the fluid core has a flattening greater than that associated with a hydrostatic equilibrium, corresponding to a 500 m increase of its equatorial radius. Observation of geophysical processes, such as earthquakes, does not enable such a fundamental result to be obtained.

Future

New discoveries of Earth's rotational irregularities can be expected from the increasing performance of modern geodetic techniques. Nevertheless, these techniques rely on very costly processings. The ring laser gyro based on the Sagnac effect and frequently used in navigation systems constitutes an interesting alternative, because it does not need any astrometric observations and allows one to do astronomy *in situ*. With this technique, still in development for monitoring the Earth's rotation, oscillations at a time scale lower than 1 h could be detected.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Lambek K 1980 *The Earth's Variable Rotation* (Cambridge: Cambridge University Press)

Melchior P 1978 *The Tides of the Planet Earth* (Oxford: Pergamon)

Morritz H and Mueller I I 1988 *Earth Rotation* (New York: Ungar)

Munk W H and MacDonald G J F 1960 *The Rotation of the Earth: a Geophysical Discussion* (Cambridge: Cambridge University Press)

Nicole Capitaine and Christian Bizouard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth's Rotation: Theory

The Earth's rotation theory is devoted to explaining the variations of the Earth's rotation. Additionally to the uniform rotation around an axis close to the geographic North pole, the Earth undergoes variations in rotation speed (including changes in the length of the day); furthermore, the orientation of the rotation axis changes with time, inside the Earth as well as with respect to inertial space. In a terrestrial frame, the motion of the rotation axis around the mean geographic pole is called polar motion; in the celestial frame, the motion of the rotation axis in space is called PRECESSION/NUTATION.

Variation of the Earth's rotation rate, variation of the length of day

The variation of the length of day (LOD), as measured by geodetic observations over recent decades, is at the level of several milliseconds for periods shorter than a few years (see the article on the EARTH'S ROTATION). These short-term variations are mainly induced by atmospheric and oceanic effects on the Earth. They can be computed from models constrained by meteorological data by considering (1) that the atmospheric pressure (or indirectly the ocean pressure) pushes on the Earth's topography (e.g. on the mountains) thereby producing a torque which changes the Earth's rotation, (2) that the density variations in the atmosphere give an additional torque (gravitational torque), and (3) that the winds also induce a friction torque on the Earth's surface. Similar torques are also introduced by the ocean. Altogether these phenomena induce acceleration and deceleration of the Earth's rotation at short time scales and, in turn, associated variations in the LOD. The Earth's variation of the LOD can be computed and compared with the ANGULAR MOMENTUM of the EARTH'S ATMOSPHERE, as done in figure 1.

Besides these variations, there are decade variations of the LOD (see above and the remaining differences in figure 1 between the measured LOD variations and those computed from the atmosphere) which are usually attributed to an exchange of angular momentum with the core because of the correlation between the surface magnetic field variations and the variations of the LOD at these time scales. This angular momentum transfer is due to torques at the CORE–MANTLE boundary (CMB): (1) a topographic torque, due to the pressure of the fluxes acting on the CMB topography (bumps and valleys which can be as high as a few kilometers); (2) a viscous torque, related to the viscosity of the liquid core; (3) an electromagnetic torque, related to the Lorentz force induced at the CMB by the electromagnetic field; (4) a gravitational torque, related to the gravitational interaction between the core and the mantle of the Earth. The topographic torque and the electromagnetic torque are usually believed to be the most important ones at long time scales. Also, the gravitational torque has attracted a lot of attention recently. Because there are no direct observations of the core flows to lead to well-constrained models such as exist for the ocean and atmosphere, only indirect calculations of the core angular momentum (and thereby of the torque acting on the CMB) can be obtained. Their variations can be calculated, if the motions in the core are characterized by rotation of fluid cylinders, from the flows at the top of the core derived from variations in the poloidal magnetic field observed at the surface. Further, these flows depend on poorly constrained additional assumptions to resolve formal non-uniqueness in the solution. Consequently, it is very speculative to give the associated changes of the Earth's angular momentum or the torques acting on the inner core–outer core boundary (ICB) or CMB, without any other constraint. Nevertheless, the Earth's rotation variations being well correlated with the magnetic field at decade time scale, data of variations of the LOD can be used to constrain the core angular momentum and the torques at decade time scale.

There is also a very long-term trend in the variation of the LOD as observed from paleontological data (coral growth in time) and from eclipse data (lag between the date deduced from a uniform rotation and the date given in past records); this very long-term trend is related to the deceleration of the Earth's rotation induced by tidal friction. Due to friction, the tidal bulge created by the classical tidal torque has a phase angle of about 5 degrees with respect to the Earth–Moon position. Because of this, the Moon exerts an additional torque which tends to break the Earth's rotation. This induces a linear trend in the solar time corresponding to an LOD increase of about two milliseconds over one century (this would almost not be seen on figures showing time-dependent functions over 14 years as presented here). This tidal friction is the reason why in the Devonian period, for instance, there were only 22 hours in a day.

Besides the tidal torque which is the predominant mechanism in reducing the Earth's spin, there is another long-term component acting to decrease the LOD (by about half a millisecond per century) and related to the decrease of the Earth's oblateness associated with deglaciation after the ice ages.

Polar motion

The polar motion in recent times is within a 20 m size square at the Earth's surface around the North pole (see POLAR MOTION AND LENGTH OF DAY). There are two main components: (1) one which has an annual period and is believed to be due to the atmosphere, but might well be additionally caused by other seasonal phenomena such as groundwater changes, and (2) another which has a period of about 435 days and is related to a normal mode of the Earth called the Chandler Wobble (CW). The CW is related to the fact that the Earth has an equatorial bulge inducing an excess of mass at the equator with respect to the pole (the Earth is oblate and has an equatorial radius larger than the polar radius by 21 km). The excitation of the CW is not well established but is generally believed to be mainly due to the atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

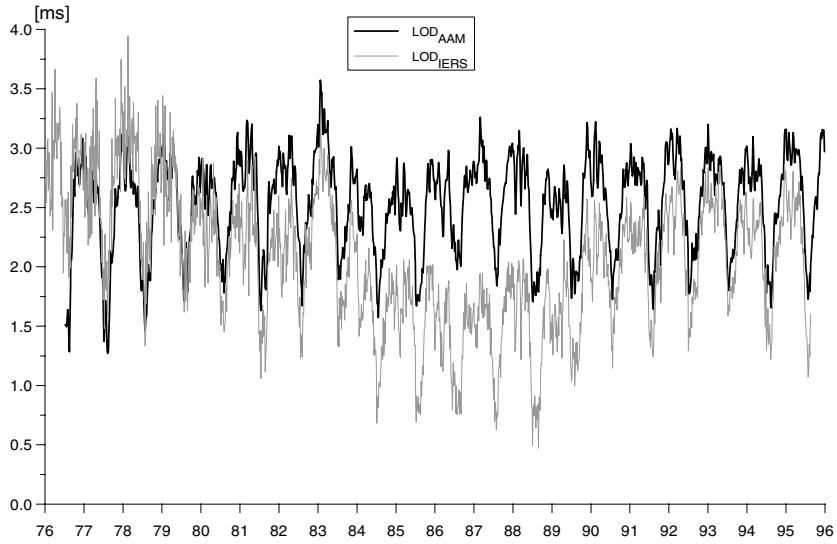


Figure 1. Comparison of the atmospheric angular momentum and the variations of the LOD for the last 14 years.

Additionally to these two periods there is a long-term trend called polar wander which is explained by internal-mass redistributions over long periods such as that associated with post-glacial rebound and mantle convection.

Precession/nutation

The luni-solar attraction on the Earth's equatorial bulge induces a torque which tends to rock the equator toward the ecliptic. The Earth responds to this as a gyroscope; its rotation axis undergoes a large motion around the perpendicular to the ecliptic of about 50 arcseconds per year. The opening of the associated cone is the angle between the equator and the ecliptic (23° 26′). The relative positions of the Moon, the Earth and the Sun being not fixed in space, in addition to this precessional motion, there exist periodic motions of which the periods are directly related to the periods involved in the relative positions of Moon, Earth and Sun: the nutations. The largest nutation amplitude corresponds to the period of the node of the Moon's orbit around the Earth (18.6 years) on the ecliptic. Figure 2 shows schematically the precession and nutations.

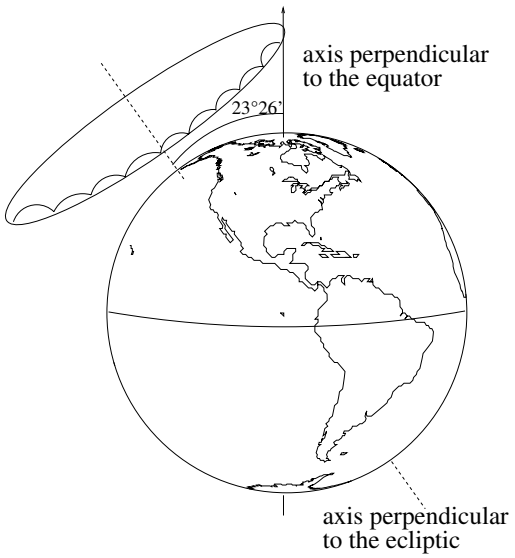


Figure 2. Schematic representation of precession and nutations.

There are actually very precise models to compute the Earth's orientation in space (for space navigation for instance). These models are based, on one hand, on computation of the Earth's nutation for a rigid body and, on the other hand, on the transfer function for a non-rigid Earth.

The rigid Earth nutation series are computed from CELESTIAL MECHANICS starting from EPHEMERIDES which give the relative positions of the Earth, Moon, and Sun (and to a minor extent the relative positions of the planets) as a function of the time variation of the fundamental argument (e.g. the longitude of the Moon). Present-day theories have an accuracy at a few microarcsecond (μas) level (1 μas $= 10^{-3}$ mas, and 1 mas is the angle with which one would see a ping-pong ball on the top of the Empire State building in New York from Europe). This accuracy is better than the nutation observation accuracy. These theories do not only include the nutations related to the direct gravitational interaction between the Moon and the Earth, and between the Earth and the Sun, but also all the second-order effects such as the perturbation of the Sun on the Earth–Moon interaction, the perturbation of the Moon on the Earth–Sun interaction, the direct and indirect effects of the planets, the effect of the presence of the equatorial bulge of the Earth (J_2-tilt effect) and the effects of the nutations on each other (coupling effect).

The accuracy of non-rigid Earth nutation transfer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

functions is currently not as good as for the rigid Earth nutation series. The models used generally consider that there is (1) an inelastic, solid, deformable inner core, (2) a liquid outer core, and (3) an inelastic, deformable mantle. They take into account the non-hydrostatic equilibrium state of the Earth which is very close to the hydrostatic ellipsoid, but not exactly, because of the presence of mass heterogeneities inside the Earth (as seen from seismic mantle tomography or from the non-ellipsoidal contributions to the geoid). The global Earth dynamical flattening of the models corresponds to the observed value (derived from the observed precession), and the CMB topography accounts for the non-hydrostatic bumps related to the deformations induced by mantle convection. The nutations computed from the convolutions between the rigid Earth nutation series and the transfer function are influenced by resonances with free modes of the Earth, especially the free core nutation (FCN) which has a period of about 432 days in the celestial reference frame. The computed normal mode periods are very close to the observed values, particularly due to the consideration of non-hydrostatic equilibrium. Some very recent models account additionally for electromagnetic coupling at the CMB. These models are precise at the sub-milliarcsecond ($<$ mas) level.

Additionally to these computations of nutations in response to the luni-solar (and planetary) gravitational attractions, the atmosphere has a non-negligible influence at the VLBI observation precision level. The atmosphere exerts a pressure torque, a gravitational torque and a friction torque on the Earth at diurnal time scale in a frame tied to the Earth. The friction torque is believed to have a negligible contribution to nutation. The atmospheric one-solar-day effect on the Earth induces a perturbation on the prograde annual nutation at a level of a tenth of a mas. Similarly, the diurnal ocean tides induce perturbation in the principal nutations. The indirect effect of the ocean due to the atmospheric forcing must also be taken into account. The atmospheric and oceanic torques cause an angular momentum transfer, and consequently, the effects on nutations can also be computed from the angular momentum balance equation, instead of calculating the torques. The so-obtained perturbation of the nutation is at the level of the tenth of a mas for the atmosphere effects, and of a mas for the ocean tide effects.

Veronique Dehant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earthshine

A faint illumination of the Moon's night side by sunlight reflected from the Earth. It is visible when the Moon shows a crescent phase, as a pale grayish ghostly glow over the remainder of the disk. Earthshine is traditionally known as 'the old moon in the new moon's arms'; seldom-used alternative names are *earthlight* and the *ashen light*. Earthshine is difficult to see less than a day after new moon because the Moon is then very low in a twilit sky. It is most noticeable about two days before or after new moon. At three or four days from new moon the illuminated portion of the disk is large enough to begin to swamp the pale earthshine. A lesser factor affecting the strength of visible earthshine is the phase of the Earth as would be seen from the Moon. At new moon the Earth is full, and reflects the most sunlight toward the Moon. Three or four days from full, the Earth would be gibbous and not nearly as bright. The intensity of earthshine also varies from one lunation to another, mainly as a result of changing levels of cloud cover over the hemisphere of the Earth presented to the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eckert, Wallace John (1902–71)

Computer scientist and astronomer. Born in Pittsburgh, PA, Eckert was a pioneer of the use of IBM punched card equipment for astronomical calculations. As director of the US Nautical Almanac Office he introduced computer methods to calculate and print tables instead of relying on human 'computers'. When, later, he became director of the Watson Scientific Computing Laboratory at Columbia University he directed the construction of computers, tailor-made to calculate the positions of the planets and Moon. It is hard to imagine the world of astronomy today without the computers and methods that he pioneered.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eclipse

An eclipse of the Sun is an OCCULTATION at which the Moon comes between the Earth and the Sun, blocking some or all of the sunlight from reaching an observer. Eclipses result from the fortuitous circumstance that the angular sizes of the Sun and the Moon as seen from Earth are the same to within about 10%, although in absolute diameter they differ by a factor of 400.

Types of solar eclipses

Solar eclipses can be total, annular or partial. At a total eclipse, the cone of lunar shadow is intercepted by the Earth's surface. The intersection traces out a path that instantaneously is close to an ellipse but that over a period of hours traces out a path thousands of km long but only up to about 300 km wide. Observers in this umbra, from which the solar photosphere is entirely blocked, see a total eclipse, while observers in a much larger area ahead or behind or to the sides of the umbra are in the penumbra and see a partial eclipse. Only during the total phase of a total solar eclipse does the CORONA become visible, and so it is mainly these total eclipses that are scientifically valuable. Totality can last from fleeting instants up to about 7 min (figure 1).

Because of the ellipticities of the orbits of the Moon around the Earth and the Earth around the Sun, mainly the former, the conical shadow of the Moon sometimes does not reach the Earth's surface. At those times, observers on Earth see an annulus of photosphere surrounding the Moon, and the eclipse is said to be an annular eclipse. Since the photosphere is approximately a million times brighter than the full Moon, an annular eclipse with 10% or even only 1% of the photosphere showing allows so much sunlight to hit the Earth's atmosphere that the sky remains blue and the corona is not visible. Annularity can last as long as 12 min.

Total and annular eclipses are known as central eclipses. Occasionally, less than 10% of the time, the umbra reaches the earth for only part of a central eclipse, and an annular–total solar eclipse occurs. Sometimes the conical umbra of the Moon's shadow is directed beyond the Earth's poles, and no total or annular eclipse occurs on Earth. These eclipses are partial only.

Although the mean semidiameters of the Sun and the Moon are 960" and 931" respectively, the ellipticities of 0.055 and 0.017 for the Moon around the Earth and the Earth around the Sun, respectively, provide enough range in apparent sizes to account for annular and total eclipses. The Moon can appear as much as 10% smaller than the Sun or 8% larger.

The umbra, if it were to hit the Earth vertically, could not exceed about 270 km in diameter, but the projection onto the globe can almost double this figure.

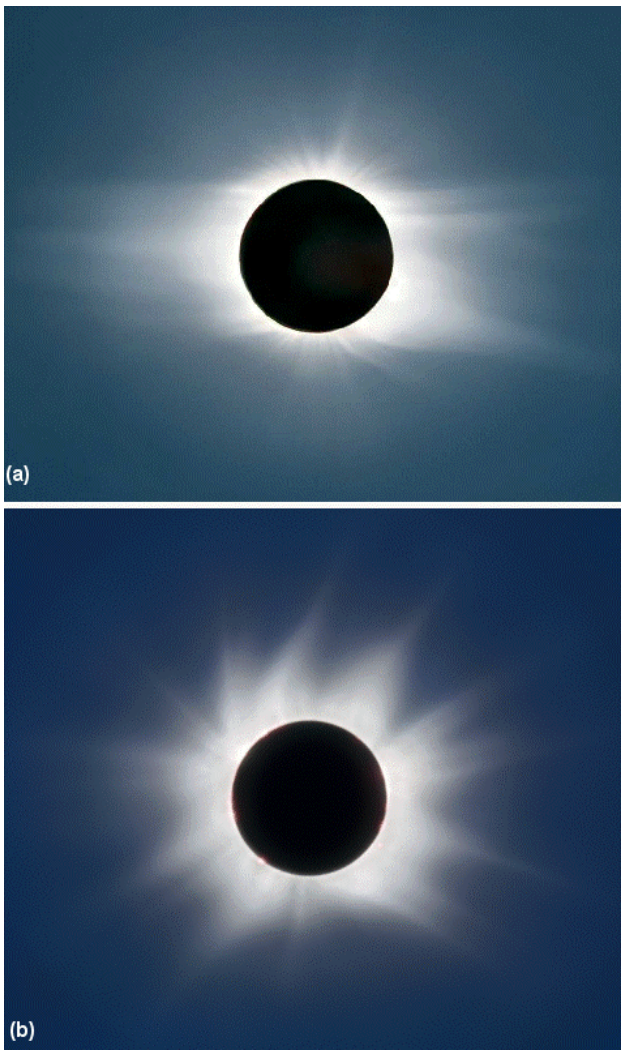


Figure 1. (*a*) The total solar eclipse of 26 February 1998, with coronal streamers at equatorial latitudes and polar tufts nearer the poles. The image was processed from several individual photographs in order to show the corona over a wide dynamic range of intensities. © 1999 Jay M. Pasachoff and Wendy Carlos. All rights reserved. SOHO–Williams College collaboration with the support of the Committee for Research and Exploration of the National Geographic Society.

(*b*) The corona of the eclipse on August 11. Processing by Wendy Carlos of eclipse photographs by Jay Pasachoff. © 2000 by Wendy Carlos and Jay M. Pasachoff.

Frequencies of solar eclipses

The 20th century contained 224 solar eclipses, including 147 central eclipses and 77 partial eclipses. There are between two and five solar eclipses in any given year, although only about every 18 months is there a total eclipse visible somewhere on Earth. An observer waiting at one location on Earth would see a total solar eclipse on average only once every 375 years, a rate that depends on latitude. However, observers willing to travel can observe

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

eclipses much more often. Even for a single country, statistics over a 3000 yr interval (Williams 1996) show that total eclipses are visible in England on average every 44 yr. Over the same interval, eclipses are visible at London every 330 yr and at Edinburgh every 186 yr. The paths of the total eclipses of 21 June 2001 and 4 December 2002 cross at the west coast of southern Africa, giving a region there two total eclipses within 18 months.

The umbral shadow moves across the Earth at about 3400 km h^{-1}. Earth's rotation allows surface points to diminish this relative velocity. The maximum rotation speed is about 1700 km h^{-1} for a point at the equator, so the relative velocity of the umbra diminishes for a time to about 1700 km h^{-1}, a speed at which the supersonic Concorde once kept up with the 1973 eclipse for 74 min. Normal jets, however, extend totality only slightly. Orbiting satellites would travel through the umbra with an even greater mismatch of speeds, so natural eclipses are not observed from space.

Predicting eclipses

Eclipses are predicted with a method descended from that worked out by F Bessel in 1824. Eight Besselian elements are given in a frame of reference of the Moon's shadow, as explained by Espenak (1987) and in the *Explanatory Supplement to the American Ephemeris* (1992). The Besselian elements of an eclipse are related to a fundamental plane that is perpendicular to the Moon's shadow and that passes through the center of the Earth. The coordinates x, y and z are then oriented east, north and parallel to the axis of the lunar shadow, respectively. The Besselian elements are then x and y for the shadow, the radii of the penumbra and umbra on the fundamental plane, L_1 and L_2 respectively, the direction of the shadow's axis on the celestial sphere, given by declination and hour angle d and μ respectively, and the angles with which the penumbra and umbra, respectively, intersect the axis, f_1 and f_2. These Besselian elements are tabulated or can be calculated by a variety of computer programs.

F Espenak and J Anderson, for each total eclipse, produce a NASA Reference Publication with tables and maps, taking over a function that was formerly covered by US Naval Observatory Circulars. See http://umbra.nascom.nasa.gov/eclipse/. Programs are now available for even the amateur market that accurately predict eclipse paths. The CD-ROM *RedShift* is an example. The *Scientific Astronomer* package for *Mathematica* includes eclipse-prediction capabilities.

Details of the calculations use a constant k for the fractional mean radius of the Moon–Earth, and this constant has been empirically adjusted to account for the observed average size of the Moon and Sun. Whether an eclipse is total depends, after all, on the complete coverage of the Sun by the lunar disk, while a phenomenon known as Baily's beads result from the last bits of solar photosphere shining through the deepest valleys on the lunar limb. So depending on assumptions, different calculations result in slightly different numbers for the frequency of total eclipses. The IAU value (1982) is $k = 0.272\ 507\ 6$, but a smaller value of $k = 0.272\ 281$ is better used for computing totality. Calculations taking lunar librations into account must be made for individual eclipses.

A solar eclipse is but an especially well-aligned new moon. The planes of the orbits of the Moon around the Earth and the Earth around the Moon are inclined by 5º. They intersect at only two points, known as the ascending node and the descending node. Only when both the Sun and the Moon are close to a node simultaneously can an eclipse occur. Because the Moon cannot move out of a node in a half-month interval, there is always a lunar eclipse 2 weeks before or after a solar eclipse, although most of these eclipses are not total. These eclipse seasons are separated by 173.3 days, giving an eclipse year of 346.6 days. In years that have three eclipse seasons, a total of five solar eclipses can occur, a circumstance that will not occur until 2206.

The saros

A coincidence in the cycles of the Sun and the Moon results in a repeat of eclipses with approximately the same characteristics every 18 yr $11\frac{1}{3}$ days, a period named the 'SAROS' by Edmond Halley. The period may have been discovered by the Babylonians.

The coincidence is several, including the following: (a) 223 synodic months = 6585.32 days, where a synodic month is the period of the lunar phases; 242 nodical months = 6585.36 days, where a nodical month (also called a draconic month, after the mythical Chinese dragon who supposedly devoured the Sun at eclipses) is the interval at which the Moon passes successive nodes in its orbit; 239 anomalistic months = 6585.54 days, where an anomalistic month is the period over which the Earth–Moon distance varies; 19 eclipse years = 6585.78 days, where an eclipse year is twice the interval between successive eclipse seasons, the 38 day period during which the Sun and Moon are sufficiently near the nodes that eclipses can occur.

The coincidence of 223 synodic months = 242 nodical months means that an eclipse occurs $6585\frac{1}{3}$ days = 18 yr $11\frac{1}{3}$ days after a previous eclipse. (Depending on how leap years fall, the interval can be 18 yr $10\frac{1}{3}$ days or 18 yr $12\frac{1}{3}$ days.) The 1/3 day means that the path of the successive eclipse is 1/3 of the way around the globe with respect to the earlier one. The further coincidence of the anomalistic month means that total eclipses are followed by total eclipses and, in addition, that long ones are followed by long ones after the saros interval, and so on. For example, the saros series that includes the longest eclipses (with maximum durations of totality in

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

parentheses) includes 30 June 1973 (7 min 4 s), 11 July 1991 (6 min 53 s), 22 July 2009 (6 min 39 s), 2 August 2027 (6 min 22 sec) and so on. Throughout this saros, the Moon's diameter is 8% larger than the Sun's. Successive eclipses drift in latitude, and the duration of the saros series now active range from 1244 yr to 1514 yr.

Historical eclipses

Records of eclipses can be found on cuneiform tablets from ancient Babylonia and on records from China, giving thousands of years of data. The earliest surviving Babylonian record is the eclipse of 731 BC. Shang dynasty Chinese sources range from ca 1350 BC, although reliable records date only to about 700 BC. Stephenson (1997) has analysed a range of such ancient eclipses to determine the rotation rate of the Earth, since whether an eclipse is total or not at a given point depends on this rate. Ancient documents can often be analysed to see whether the eclipse was total, since the difference between total eclipses and other solar eclipses is so dramatic. Stephenson continues his study through medieval European records and Arab chronicles, to provide a graph of the length of the day over time based on over 400 eclipse observations. He finds a difference in ΔT, the difference between mean solar time and ephemeris time, of approximately 20 000 s over a range of over 2700 yr. Much of the change in the rotation rate of the Earth derives from lunar and solar tides, but he also finds non-tidal contributions, perhaps resulting from post-glacial uplift affecting the terrestrial oblateness. A day was approximately 42 ms shorter in 500 BC than it is now, although the graph shows a periodic term in addition to a trend.

Stephenson discusses in detail each of the ancient reports and assesses its reliability. For example, many of the supposed eclipse reports allow of alternative interpretations. Perhaps the most famous ancient eclipse is the one of 28 May 585 BC, which Herodotus (died ca 425 BC) related (as reported by Pliny in the 1st century AD) was predicted by Thales of Miletus and stopped a war between the Medes and the Lydians. Stephenson's conclusion, however, is that Herodotus has exaggerated and that the eclipse was not actually predictable at that time. Stephenson states that 'the first person to give the true explanation of eclipses was Anaxagoras (500–428 BC),' but quotes Neugebauer as concluding that eclipses were not predictable for hundreds of years after that.

Eclipse phenomena

Eclipse calculations specify the times of the contacts. First contact is the first meeting of the disks of the Moon and Sun; it is noticeable to an observer looking through a filter with unaided eye within a few seconds. Second contact is the complete coverage of the solar photosphere by the Moon; it marks the beginning of totality. Third contact is the first appearance of the solar photosphere after totality; it marks the end of totality. Fourth contact marks the departure of the lunar disk from the photosphere. It is little observed, since observers often lose interest after the excitement of totality.

The interval between first and fourth contact is commonly 3 h and can be as long as 4 h. For most of the time between first and second contacts, the eclipse would not be noticeable to anyone who did not know that it was occurring. However, for the last 15 min or so before second contact, the sky color noticeably changes and shadows grow strangely sharper, since they are now being cast by a crescent instead of a sphere. During the last couple of minutes before totality, shadow bands, low-contrast ripples of light and dark, are often seen to move rapidly across the landscape. They are very difficult to photograph and are better observed photoelectrically. Since they are caused by phenomena in the Earth's upper atmosphere, they are of little interest to solar physicists, but atmospheric physicists might similarly be more interested in them than in the solar phenomena that are soon to become visible.

As the crescent narrows, within the last minute before second contact it begins to break up into individual beads, known as Baily's beads. They are named after the English astronomer Francis Baily, who noted them at the eclipse of 1836, though they had earlier been commented on. Depending on the sky quality, the solar corona may or may not be visible during this period. The last Baily's bead glows so brightly compared with the other solar phenomena visible—the chromosphere and/or the innermost corona—that it looks like the diamond on a ring and is known as the diamond-ring effect. It often lasts only a few seconds but can last nearly a minute, depending on the depth of the deepest valley on the edge of the Moon that is aligned properly to transmit photospheric light. The glitter of a diamond is often matched in images by internal scattering in camera lenses. (See figure 2).

As the diamond diminishes, a pinkish narrow rim on the Sun becomes visible. It got the name 'chromosphere' from its colorful appearance at eclipses. Depending on how much larger in angular size the Moon is than the Sun, you may be able to see the full chromosphere at one time or you might see only the leading edge at second contact and then the trailing edge at third contact. The pink color comes largely from Hα, and has an admixture of other chromospheric emission lines. Spectra taken over this interval show the photospheric absorption lines fading and chromospheric emission lines flashing into view (that is, appearing abruptly), the flash spectrum.

As the diamond ring and chromosphere disappear, you have totality (figure 3). The shape of the corona becomes visible, with CORONAL STREAMERS obvious. The shape of the corona varies over the solar-activity cycle. At solar minimum, most of the streamers visible

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

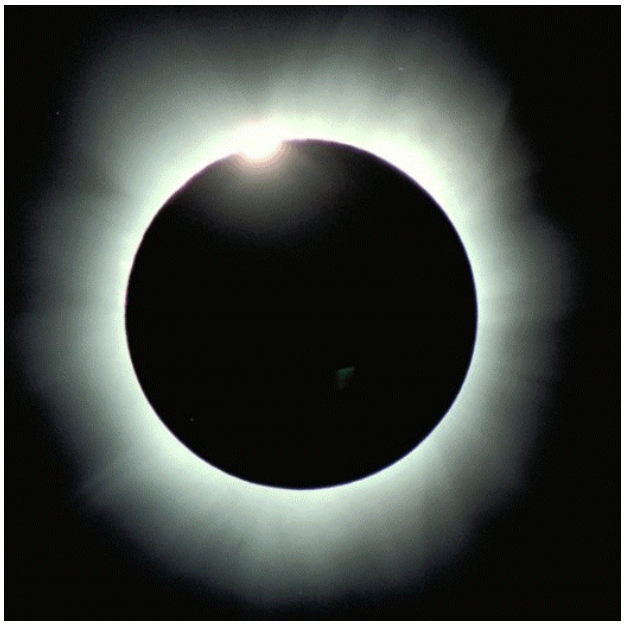


Figure 2. The diamond ring effect beginning totality at the 2001 total solar eclipse imaged from Zambia by the Williams College Eclipse Expedition.

Figure 3. The corona in the sky surrounding the dark disk of the Moon during the total solar eclipse in India of 1995. (Jay M Pasachoff; Williams College, with the support of the Committee for Research and Exploration of the National Geographic Society and of the US National Science Foundation.)

are equatorial, and polar tufts show near the poles. Isophotes of the corona appear elongated. At solar maximum, streamers are visible at more latitudes and isophotes of the corona appear nearly round.

SOLAR PROMINENCES have become visible during the diamond-ring effect and chromosphere period, and if sufficiently large may remain visible throughout totality.

All too soon, within a period no longer than about 7 min 30 sec and often much less, a brightening region of the corona becomes visible as the Moon progresses across the Sun, the pinkish chromosphere becomes visible again and a second diamond ring appears to mark the end of totality. The rest of the eclipse phenomena, including Baily's beads and final crescents and later partial phases, proceed in reverse order from the ingress.

Eclipse safety

The brightness of the solar corona during totality is approximately that of the full moon, and the corona during totality is equally safe to watch. However, staring at the crescent or other partial phases of the eclipse can be hazardous to eyesight, so warnings should be given to the populace about how to watch an eclipse safely (see SUN: SAFE OBSERVATION). All too often, the distinction between the hazards of the partial phases and the safe glory of totality is lost or not understood by reporters or by the public. One often runs across a popular belief that some special rays are emitted by the Sun during an eclipse that make it unsafe to watch or even to be outdoors. This belief is totally untrue, of course; there is less of everything during an eclipse, including less solar radiation.

Next, we discuss viewing methods for the partial phases of an eclipse; it is important to note that the filters described are only for the partial or annular phases and must be removed to view totality.

On a normal day, one is not tempted to stare at the Sun, and the eye-blink reflex normally prevents you from doing so. However, during the partial phase of an eclipse, people's interest has been called to the Sun, and during the crescent phases the total intensity of sunlight may be inadequate to activate the eye-blink reflex even though the specific intensity of regions of the solar photosphere that is still visible remains high. Thus people should observe the Sun only through special filters. These filters cut out all but about 1/100 000 of the photospheric light, reducing the photosphere to a safe level for observation.

The key factor in safe solar filters, aside from their absorbing or reflecting sufficient levels of solar intensity, is that they do so evenly across the spectrum—that is, they are of neutral density (ND). With each level of density given logarithmically, filters that pass only 1/100 000 the light are ND5. Solar filters made of aluminized Mylar, a coated plastic, are very popular and are available inexpensively. As long as these Mylar filters are undamaged, without creases or pinholes, they are safe to look through. Unfortunately, they are now often distributed in the form of eyeglasses, which may tempt people to stare at the Sun through them, rather than as simple rectangles to hold up and look through. Although the solar intensity seen through them has been reduced to a safe level, it is still a good extra safety precaution never to stare at the Sun.

For telescopic use, filters made of chromium deposited on glass pass the requisite amount of photospheric light to photograph and view the partial phases while giving images of excellent definition. A

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

very slight increase in transmission at the red end of the spectrum often gives a pleasing tint to the solar image. Welders' glass of #14 or #15 is also safe for visual or photographic viewing, although the color is usually less pleasing. Photographic 'ND filters', such as Wratten filters, are not safe for use at eclipses, since they are transparent in the infrared rather than being ND across the spectrum. Similarly, crossed polarizers are also unsafe.

A very safe way to view the partial phases is with a pinhole camera. Such a device is no more than a small hole, perhaps 2 mm–5 mm across, in a piece of cardboard or aluminum foil (perhaps taped in place over a larger hole in cardboard). This small hole is held 0.5 m–1 m or so above a piece of paper or cardboard, and one looks down at this second surface with the Sun over one's back. The shape of the partial phase is visible on this second surface and is completely safe to look at. The interstices among the leaves of a tree often act as their own natural pinholes, and the ground or other surfaces under trees often show multiple pinhole images of the partial eclipse and its progression.

An analogous method using an optical device is 'eyepiece projection', where an eyepiece is cranked beyond its normal focal point and used to project an image of the partially eclipsed Sun on a paper, cardboard or other screen. One looks only down at the screen, never up through the telescope.

Though the eye is not damaged by occasional unaided glances at the Sun, which often occur accidentally on normal days, it can be immediately and permanently damaged by the focused image of the Sun through a telescope or binocular. So if a telescope or binocular is used in a direct-view mode, a solar filter must be taped or otherwise secured very carefully to the front of the optical device so that it cannot blow off or fall off. Filters that go on the eyepiece end are always liable to crack or be otherwise damaged under the concentrated solar radiation, so are much less safe than filters above the objective lens. Viewing through a properly filtered telescope or using eyepiece projection has the advantage over pinhole cameras of showing sunspots, which do not normally show in pinhole images.

Many people have heard of viewing the partial phases through smoked glass, achieved by holding glass in a flame. This method is undesirable because of the unevenness of the coating and the ease in its being damaged. A later traditional method is to use two or three layers of fogged and developed black-and-white film. As long as the film contains silver, which not all of the new black-and-white films do, and it is completely exposed to light and then is developed to full density, it should be safe for use, since the silver provides ND. Color film and slides are transparent in the infrared and must not be used as filters. Fogged and developed x-ray film works well, but one must be careful not to be misinterpreted when giving instructions, because I have heard of someone using regular x-rays and looking through the light parts that show the bones. Similarly, some compact disks (CDs) work as solar filters, if they are coated sufficiently densely, but novices should be warned not to look through the hole.

All too often, the need for instructions on how to observe safely is misinterpreted by school officials or by the press to mean that people should be warned not to look at the eclipse directly. Since so much of the glory of watching the eclipse comes from the rapid changes in ambient light and from the extreme contrasts of the eclipse phenomena themselves, much of that effect is lost in watching an eclipse on television. Stories of students being prevented from watching eclipses by school personnel abound and are most unfortunate since students can be inspired by observing such natural phenomena. Further, when teachers exaggerate the hazards when telling students about eclipses, students can soon learn the true situation and regret having missed seeing the phenomenon. Then they will lose trust in the teachers and may not take their advice on other matters, including health-related ones like taking drugs or smoking.

The occasion of an eclipse is an excellent one for public information about science. It is important to give accurate information early on to the general public and to students. It is further important to distinguish between the partial (or annular) and total phases of an eclipse. Since the corona is approximately 1/1 000 000 the brightness of the photosphere, even if 99% of the photosphere is hidden, the remaining 1% is still 10 000 times brighter than the corona. So the general public should be informed that a so-called 99% eclipse is very far from totality, and that to see the exciting eclipse phenomena they must journey to the zone of totality.

Eclipse science

One of the main advantages of solar eclipses is that the solar corona can be observed with the latest instruments and with cognizance of the latest theoretical ideas. Thus the types of discoveries that have been and can be made at eclipses changes over time.

Since only a few minutes at most of totality are available at eclipses, it is usually desirable to extend or continue observations at a forthcoming eclipse. Further, new ideas about the solar atmosphere may suggest new observations to be made. In addition, eclipses sample the solar-activity cycle, with different prominences and coronal features visible at each.

Eclipse expeditions began in the 18th century and reached maturity during the 19th. The first coronal photograph was taken in Koenigsberg in 1851. Prominences were discovered to be intrinsic to the Sun rather than artifacts during the 1860 expedition to Spain. The diamond ring effect was first photographed at the 1859 expedition, and photographs at the 1870 and 1871

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

expeditions were used to prove that the corona was intrinsic to the Sun by comparing photographs from widely separated locations.

Perhaps the most famous eclipse discovery was made at the 1868 eclipse in India. A spectrograph revealed the brightest emission lines of the CHROMOSPHERE for the first time. At first a bright yellow line was thought to be the D line of sodium, but it was soon realized that it was displaced from the D_1–D_2 sodium doublet, and was called D_3. D_3 was so bright that it was realized both from the eclipse results and independently elsewhere that it could be viewed outside of eclipse. The element that caused the D_3 line in the spectrum was called 'helium', from the Sun god Helios, since it was known to exist only in the Sun. Helium was not identified on Earth until 1895. The flash spectrum was discovered at the eclipse of 1870, when it was thought to reveal a 'reversing layer'.

Starting with the total eclipse of 1869, additional emission lines were discovered that came from the corona, and were said to come from 'coronium'. The periodic table had not been completed at that time, so it was not obvious that 'coronium' was any less an element than 'helium'. In slitless spectra, in which the brightness of the chromospheric and/or corona itself is used to define the part of the image that is to be dispersed, the chromospheric emission lines appear as arcs while the coronal lines appear as circles. The brightest coronal lines in the visible part of the spectrum, all that could then be studied, are the 'green line' at 530.3 nm and the 'red line' at 637.4 nm. A dozen fainter coronal lines have also been identified in the visible. The source of these spectral lines was not identified until 1939 and following years, when they were shown to be from 13-times-ionized iron ([Fe XIV], where the neutral state is Fe I and the square brackets indicate that the lines are forbidden by the quantum-mechanical selection rules) and from nine-times-ionized iron ([Fe X]), respectively. This result was the proof, backing other lines of evidence that had not been generally accepted, that the corona is very hot, containing gas at millions of kelvins.

The observations by Eddington and colleagues at the 1919 eclipse expeditions that stars very near the Sun were deflected slightly, as predicted by Albert Einstein in his general theory of relativity, led to the acceptance of Einstein's theory. The Sun was in the Hyades during the eclipse, giving an especially rich star background. Eclipse observations for the purpose of testing relativity have been repeated several times since, and the old data from 1919 and 1922 have been reanalysed, but there are better methods of now of verifying the same principles and this Einstein experiment is no longer a major part of eclipse research.

Rocket missions are sometimes flown during eclipses. The Lyman-alpha corona was discovered during a rocket flight that took slitless spectra during the 1970 eclipse. The Lyman-alpha image was not expected because of the high temperature of the corona, but even at millions of kelvins sufficient hydrogen remains un-ionized to scatter enough photospheric light to be observed. This Lyman-alpha corona is now observed from telescopes on the space shuttle and on satellites. Other eclipse rockets are flown at the same time as eclipses but out of totality, for comparison with eclipse results. High-resolution x-ray images are obtained for this purpose, for example.

The corona observed at eclipse comes from several different sources. The K-corona (from the German word *Kontinuierlich*) is caused by photospheric light scattering off coronal electrons. In the scattering process, Doppler effects broaden the photospheric absorption spectrum to oblivion, or nearly so, and polarize the light. The F-corona (from Fraunhofer) is the light from the photosphere that scatters off dust in interplanetary space. Although dust is vaporized very close to the Sun, by the orbit of Mercury there is sufficient dust to make an important contribution to the corona as viewed from Earth. This scattering preserves the Fraunhofer absorption spectrum and does not polarize the light, so spectra and polarization measurements are commonly used to 'separate' the F- and K-coronas. The emission lines of highly ionized species, like the coronal green and red lines, are measurable only very close to the solar disk, just above the lunar disk at eclipses. The relative contributions of these different parts of the corona are shown in figure 4. The dust itself can be measured in the infrared, and is sometimes called the T-corona (from thermal). The F-corona is the inner part of the zodiacal light.

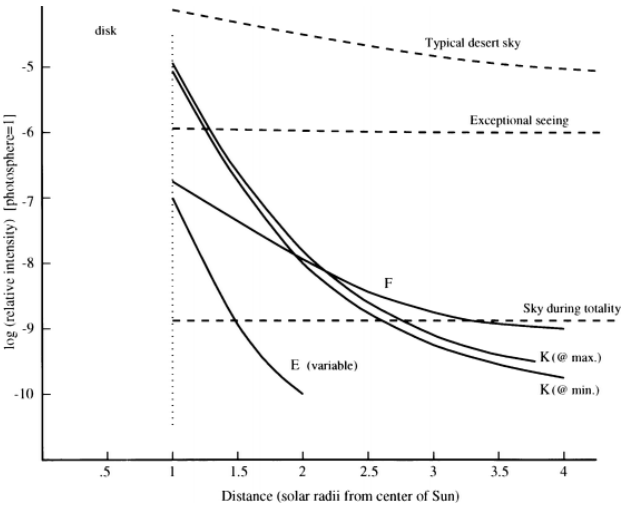


Figure 4. The intensities of different parts of the corona. K is the continuous (spectrum) corona, F is the Fraunhofer (absorption line) corona and E is the set of emission lines. (From Golub L and Pasachoff J M 1997 *The Solar Corona* (Cambridge: Cambridge University Press); after van der Hulst H C 1953 *The Sun* ed G Kuiper (Chicago, IL: University of Chicago Press).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

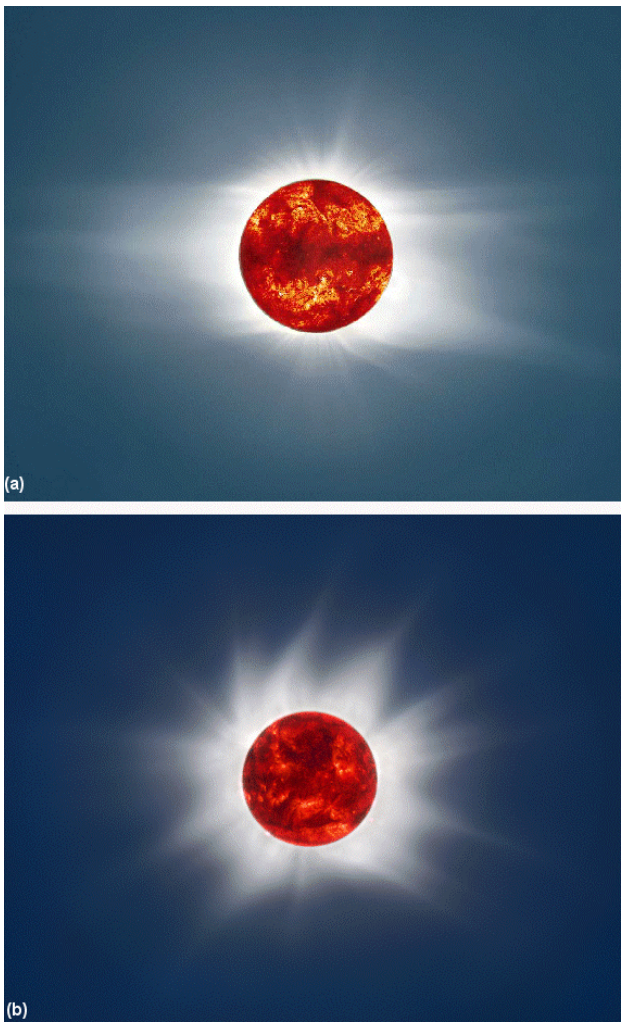


Figure 5. (*a*) The total solar eclipse of 26 February 1998, with an image of the solar disk based on observations from the Extreme-ultraviolet Imaging Telescope (EIT) on the Solar and Heliospheric Observatory (SOHO). The EIT central image shows the coronal temperature, and was computed as the ratio of two of their individual images that show the corona in different temperature ranges. © 1999 Jay M. Pasachoff and Wendy Carlos. All rights reserved. Central image from the EIT/SOHO team, NASA Goddard Space Flight Center. SOHO–Williams College collaboration with the support of the Committee for Research and Exploration of the National Geographic Society.

(*b*) Composite image from August 11 total solar eclipse with SOHO EIT image of the solar disk superimposed on coronal features photographed by the Williams College Eclipse Expedition. SOHO EIT images courtesy of the EIT team at NASA Goddard Space Flight Center, combined by Wendy Carlos.

Since the corona diminishes in intensity so rapidly with distance from the solar limb, a factor of about 1000 in the innermost solar radius above the limb, ordinary film or CCDs cannot capture the intensity of the whole corona on a single image. Accordingly, G Newkirk worked out a method in the 1960s, following earlier work by Laffineur and others, of making an eclipse camera with an ND filter that is radially graded in density so that it is densest near the limb and progressively less dense radially outward. Thus the coronal image that passes the filter is flatter in intensity than the original image, and can be captured on a single piece of film. Ten such images were taken, from 1966 through 1998. Radial filters have also been used by other observers. At present, computer techniques are allowing similar effects to be created using multiple images taken with varying exposure times, as was shown in figures 1 and 4. Such images show the shapes of coronal streamers especially well and thus reveal the SOLAR MAGNETIC FIELD.

Although the innermost corona has been observable from certain mountaintop observatories since the coronagraph was invented by Lyot in the 1930s, those observations were limited to regions closer to the limb than were visible at eclipses. The K-coronameter now at Mauna Loa Observatory and some other coronagraphs can now be used to observe the corona farther out, but are still limited in coronal observability. The brightness of the daytime sky, even at high altitude from such coronal sites as Mauna Loa or Haleakala, remains too high to carry out the full range of coronal studies.

With solar satellites aloft viewing the corona every day, we have entered a new era in coronal eclipse research. There are still niches for which the Sun can be better studied from the ground than at an eclipse, and scientific work at eclipses must be pointed to those studies. For example, the spatial region within the limit of the occulting disk of the innermost coronagraph of SOHO can be observed at an eclipse. Cadences can be achieved at eclipses that are quicker than those available from SOHO, Yohkoh, or TRACE, which has the potential of solving the problem of CORONAL HEATING MECHANISMS. Further, regions of overlap can be analysed to the mutual benefit of eclipse and space researchers. For example, space instruments can be calibrated by comparison with eclipse results, or the forbidden lines visible from the ground can be compared with the permitted lines of the same ionic species observed from space. Eclipse coronal observations can extend the LASCO coronagraph observations into the solar limb, within LASCO's 1.8 solar radius occulting disk (and even the 1.1 solar radius disk of a defunct component). The comparison of ground-based eclipse observations with space results can extend the value of either type of observation (figure 5).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

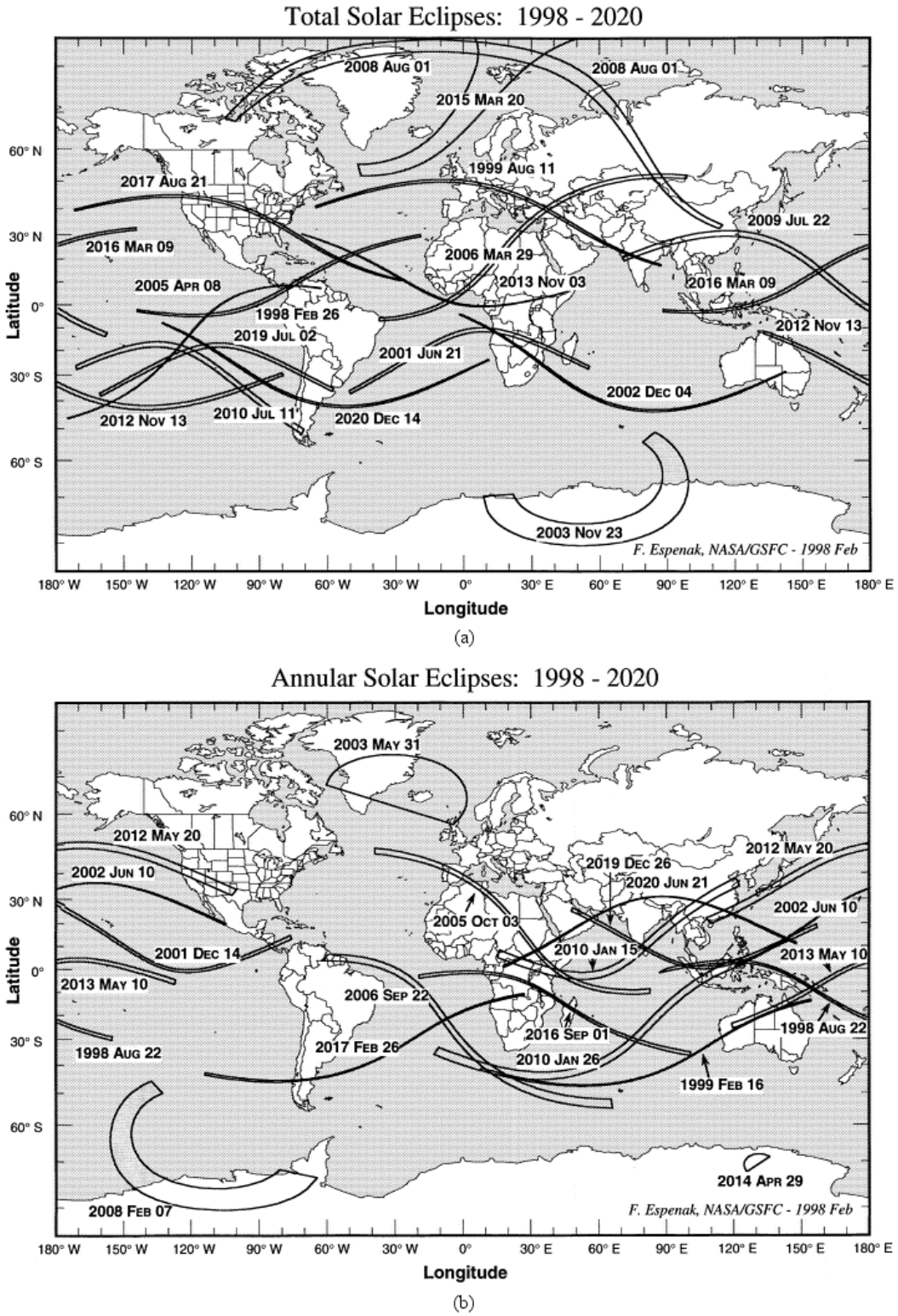


Figure 6. (*a*) Total solar eclipses, 2000–2010. (Courtesy of Fred Espenak, NASA/GSFC). (*b*) Annular solar eclipses, 2000–2010. (Courtesy of Fred Espenak, NASA/GSFC).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Total solar eclipses, 2000–2010.

Date	Maximum duration	Path of totality
21 June 2001	4 m 57 s	Southern Africa (Angola, Zambia, Zimbabwe, Mozambique, Madagascar)
4 December 2002	2 m 4 s	Southern Africa (Angola, Zambia, Botswana, Zimbabwe, South Africa, Mozambique), Pacific Ocean, ending at Western Australia
23 November 2003	1 m 57 s	Antarctica
8 April 2005	0 m 42 s	Total only at maximum over the Pacific Ocean
29 March 2006	4 m 7 s	West (Ghana, Togo, Benin, Nigeria, Niger, Chad, Libya, Egypt), Turkey, Russia, top of Caspian Sea, Kazakhstan
1 August 2008	4 m 27 s	Northern Canada, Greenland, Russia (Siberia), western Mongolia, China
22 July 2009	6 m 39 s	India, Bangladesh, northernmost Myanmar, China
11 July 2010	5 m 20 s	Pacific Ocean, ending over southern Chile and Argentina

Future eclipses

Total eclipses occur about every 18 months (figure 6(*a*)), table 1 and annular eclipses occur with about the same frequency (figure 6(*b*)). The partial phases are visible for thousands of kilometers to the sides of the zones of totality or annularity.

Bibliography

Espenak F 1987 *Fifty Year Canon of Solar Eclipses: 1986–2035 (NASA Reference Publication 1178)* (Greenbelt, MD: NASA)

Golub L and Pasachoff J 1997 *The Solar Corona* (Cambridge: Cambridge University Press)

Golub L and Pasachoff J 2001 *Nearest Star: The Exciting Science of Our Sun* (Cambridge, MA: Harvard University Press)

Mucke H and Meeus J 1992 *Canon of Solar Eclipses –2003 to +2526* (Wien: Astronomische Büro)

Mouradian Z and Stavinschi M (ed) 1997 *Observational and Theoretical Problems Pertaining to Solar Eclipses* (Dordrecht: Kluwer)

Stephenson F R 1997 *Historical Eclipses and Earth's Rotation* (Cambridge: Cambridge University Press)

Williams S 1996 *UK Solar Eclipses Since the Year 1* (Leighton Buzzard: Clock Tower)

Zirker J 1995 *Total Eclipses of the Sun* 2nd edn (Princeton, NJ: Princeton University Press)

World Wide Web

Working Group on Eclipses of the International Astronomical Union: http://www.williams.edu/astronomy/IAU_eclipses

Eclipse predictions: http://umbra.nascom.nasa.gov/eclipse/

Golub and Pasachoff, *The Solar Corona*: http://www.williams.edu/astronomy/corona

Golub and Pasachoff, *The Nearest Star*: http://www.williams.edu/astronomy/neareststar

Jay M Pasachoff

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eclipsing Binary Stars

We have an eclipsing binary when the stars of an orbiting pair periodically pass in front of one another, thereby dimming light from the system. Bright eclipsing stars such as Algol and ϵ Aur excite the interest of even casual observers, and many are within reach of binoculars. Both stars are eclipsed in most examples, although usually one of the alternating eclipses is much more prominent than the other, and eclipses of only one star are possible. The existence of eclipsing binaries is expected when one realizes that many stars occur in binaries[1], although whether they are common or rare is a matter for further thought and observation. Historically, the discovery of eclipsing binaries was one of the first indications of binary nature. (See VARIABLE STARS for early history and BINARY STARS: OVERVIEW.) Although initial mental images may be of simple spherical stars, a host of interactive phenomena includes tides, winds, impacting streams, magnetic starspots and transfer of energy, while results provide checks, insights, and discoveries for the fields of stellar structure and evolution. Classes are named after prototypes, such as the abundant RS CVn, Algol and W UMa binaries. Typically a class originates with its recognition in terms of common observational traits, followed by several false starts toward understanding. The class may reach mature status, as have the mentioned classes, when there is widespread agreement on its coherence and on the structure and evolution of its members—of course with surprises at all stages. The central thread that has led from classification to structure and evolution is the idea of limiting lobes and the natural morphological types built around them. Thus we have detached (see BINARY STARS: DETACHED), semidetached, overcontact and double-contact binaries—somewhat pictorial names with formal meanings that are developed below. The eclipsing binary concept has grown to include eclipses of and by circumstellar disks, eclipses of fluorescent circumstellar gas by one or both stars and eclipses of the stars by attenuating clouds of circumstellar material (see DUSTY CIRCUMSTELLAR DISKS). The best-known eclipsing binary results are accurate fundamental data on masses, luminosities and radii of stars (MLR data), while classical problems concern the origin, structure, evolution and general nature of binaries, with impact on our understanding of star clusters and galaxies.

Information from eclipsing binaries

Eclipsing binaries have primarily been discovered by photographic surveys, with a few turned up by accident in photoelectric work. They can be found at great distances and many are known in other galaxies, although limiting magnitudes and spatial resolution set practical limits. Larger light loss (eclipse depth) leads to greater likelihood of discovery and usually to greater astrophysical usefulness. The orbit is edge on to the observer for an inclination, i, of 90° and high $\sin i$ obviously favors deep eclipses, as does a large ratio of star size to orbit size. *Total eclipses* (entire star covered) and *annular eclipses* (only a ring of surface not eclipsed) are favored by unequal star sizes and, of course, by high $\sin i$. A *complete eclipse* is one that is either total or annular, while a *partial eclipse* is neither total nor annular. The deepest eclipses occur when a star with most of the system luminosity (in a given photometric bandpass) is totally eclipsed, as in RW Tau. Annular eclipses, partial eclipses and eclipses of stars with low relative luminosity are typically much less deep and may even be undetectable. Total eclipses detectable only in x-rays can occur when a normal star has a hot neutron star companion. For circular orbits, the star with higher average surface brightness over the eclipsed area, which ordinarily means the star of higher surface temperature, necessarily has the deeper eclipse. The situation is more complicated for an eccentric orbit binary where one eclipse may be near periastron and rather deep, with the other near apastron and quite shallow or absent. Decrease of $\sin i$ reduces eclipse durations and depths, but only slightly for inclinations near 90°. Larger reductions occur when $\sin i$ becomes too low for complete eclipses.

Brightness is measured relative to a judiciously chosen constant comparison star, nearby on the sky, while Doppler shifts are relative to the spectrum of a stationary light source. Velocity curves for both stars are ordinarily required for full results on absolute dimensions. The roles of light and velocity curves can be envisioned via a motorized scale model with glowing stars. Place the model in the sky (surmounting practical difficulties) at a distance of a few kilometers and observe its light curve, which could be indistinguishable from that of a real binary. By applying knowledge of eclipses and other photometric effects we might find relative dimensions, but nothing in the light curve tells us the model's tiny absolute dimensions. Radial velocity curves deal with real lengths and can testify to the model's absolute size but will not provide a picture. We therefore need both light and velocity curves to derive both image information, including relative dimensions, and the absolute scale.

Among phenomena that make light curves informative are tidal and rotational distortions, mutual heating, gravity brightening, limb darkening, magnetic star spots and accretion hot spots. Brightness variations due to tides and heating are collectively known as proximity effects. They tend to be prominent in eclipsing light curves because the same characteristic that makes eclipses likely also enhances tides and heating. Statistically, that characteristic is *large star size relative to orbital separation*, although tidal distortion is more directly dependent on star size relative to limiting lobe size. Prior to about 1970 proximity effects were mainly regarded as hindrances to eclipse analysis, while spots were not widely seen as real. With an evolution in thinking, all these effects are now recognized as valuable indicators of physical and geometrical binary star properties.

[1] Estimates range from somewhat over half to 90+% of stellar systems being binary or multiple.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ellipsoidal variation is named for the approximately ellipsoidal figure of a tidally distorted star. It is periodic on the orbit period, being concave-downward between conjunctions and concave-upward around the conjunctions (figure 1). That is, tidally distorted stars appear brightest from the side and dimmest end on. In the rather extreme case of overcontact, the overall binary system is shaped somewhat like a dumbbell as in figure 1. The term *ellipticity effect* also is used, particularly when a distinction is made between variation due to shape (*geometric ellipticity*) and to the consequent surface brightness distribution (*photometric ellipticity*). Photometric ellipticity arises because of gravity brightening, where local surface flux depends on effective gravity. Theory predicts that bolometric flux is proportional to effective gravity in the case of radiative energy transfer and roughly to the 0.3 power of gravity for convective transfer. The first prediction is from the theory of RADIATIVE TRANSFER, while the second comes from numerical models of convective stellar envelopes. The geometric and photometric brightness variations have roughly equal amplitudes (radiative case), with similar waveforms, and add to a total variation almost twice that from geometry alone (figure 1). In the convective case, total variation is only moderately greater than the pure geometric effect. Ellipsoidal variation is largest for overcontact binaries with strong tides and radiative envelopes such as TU Mus, where it can reach amplitudes of order 30% of mean system light. Amplitudes are quite noticeably smaller, although still important, for the very abundant W UMa type overcontact binaries with deep convective envelopes. Many binaries have sin i too low for eclipses but still show ellipsoidal variation, as does the bright star Spica, which thereby varies about 3%. Ellipsoidal variation also can 'hide' from observation by affecting only the dim member, as in ALGOL where the lower mass star has a very large tide but only 2% of the luminosity of its nearly spherical companion. The gravity effect not only brightens a system as seen from the side, but also reduces the inward-facing brightness of each star relative to its outward-facing side, so that it appears dimmer at superior than at inferior conjunction. Ellipsoidal amplitudes tend to be small for eccentric orbits because the stars are only briefly near periastron, where tides are strongest. Ellipsoidal waveforms become more complicated with increased orbital eccentricity because tides are then variable and at best quasi-static. A likely example of aperiodic dynamical tides is GP Vel, the site of x-ray source Vela X1.

Radiative heating of each star by the other produces the *reflection effect*. Thermal re-emission from a reflection cap on the inward side causes a light curve elevation around superior conjunction as part of a periodic variation that can be described by a few Fourier terms. Note that this difference at the conjunctions is opposite in sign from the corresponding one for ellipticity. The typical overall result of the two reflection effects seen together is a hill around secondary eclipse (figure 2) and intereclipse variation that is concave-upward. (The upward concavity is very subtle and usually is overpowered by any small ellipsoidal variation.) The enhancement near secondary eclipse arises from the separate reflection effects being a half-cycle out of phase and the cooler star having the larger effect. One might expect local energy conservation in incident versus re-emitted radiation, but theory and observation indicate that irradiated stars with deep convective envelopes conserve energy only globally and re-emit about half of incident energy locally. Thus the *bolometric albedo* parameter, (locally re-emitted energy)/(incident energy), is unity for radiative envelopes but only about 0.5 for convective envelopes, which have reduced reflection amplitudes.

Limb darkening is a wavelength-dependent aspect phenomenon caused by the temperature gradient in a stellar atmosphere and made familiar by observations of the Sun. It alters eclipse shapes and depths in ways that are mainly subtle but can be obvious in annular eclipses. Light curves of RS CVn type, BY Dra type and some other categories show variable 'maculation' waves caused by heavily spotted areas (figure 3). In some binaries these magnetic starspots drift in longitude and their maculation waves accordingly migrate steadily in phase, perhaps superposed with ordinary ellipsoidal and reflection variation.

While each light curve effect has something to tell, eclipses deserve their top billing, as even minimal eclipses greatly delimit possible configurations. That fact becomes apparent when one tries to understand a binary with several kinds of variation but no eclipses. Multibandpass observations are particularly useful when the star temperatures differ greatly, as short and long wavelengths respectively emphasize contributions from the hotter and cooler stars. Parameters derivable from eclipses are listed in table 1, with comments. Among these, P, e, ω, M_2/M_1 and a reference time are also derivable from radial velocities in suitable circumstances, while i appears in the velocity parameters $a_{1,2}$ sin i. Thus an i from light curves allows extraction of a_1 or a_2 from single-lined velocities, or both from double-lined velocities, and thus contributes to establishing the absolute scale and absolute masses.

There is much active research on unusual and bizarre binaries where information retrieval can be difficult but well worth the trouble, as it can point to new directions in stellar evolution. Quasi-periodic and irregular variations call attention to interesting characteristics and evolutionary histories. For example, RS CVn binaries were first noticed because of light curve waves due to magnetic starspots, but the evolutionary significance of RS CVns now runs far beyond that phenomenon. However, quite regular timewise behavior sometimes stimulates a flood of research when its implications are appreciated. The stars of a W UMa binary exchange thermal energy through a connecting neck, yet are remarkably stable and long lived. The issue of why they are stable has generated a large literature over many decades, so observationally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

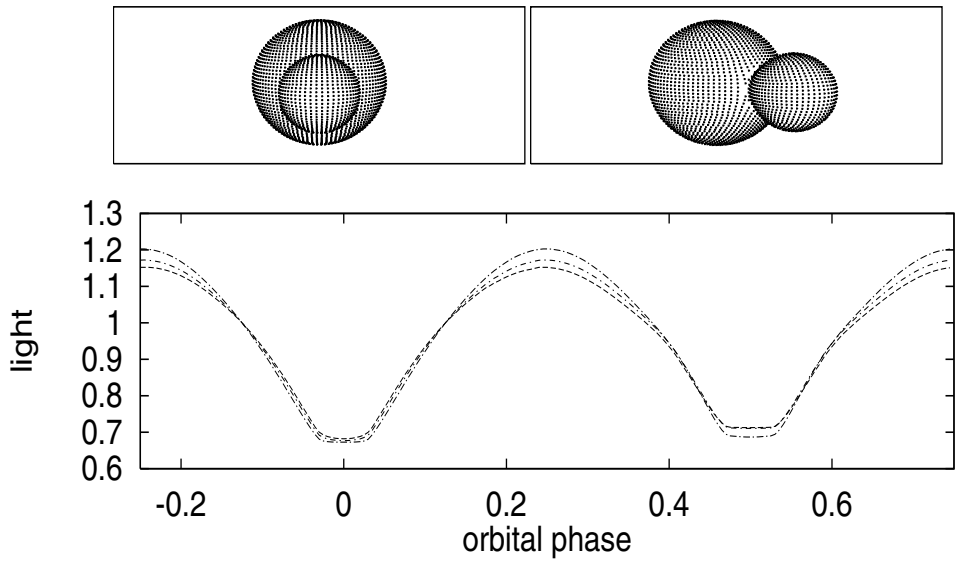


Figure 1. The effect of gravity brightening on the ellipsoidal variation of a tidally distorted binary is shown by the three curves in the lower panel. The largest amplitude is for the full effect, as expected for radiative energy transfer. The smallest amplitude shows the purely geometrical part of ellipsoidal variation. The in-between amplitude is for the reduced gravity brightening expected for convective energy transfer. The top panels are computer simulations of the binary at two places in its orbit.

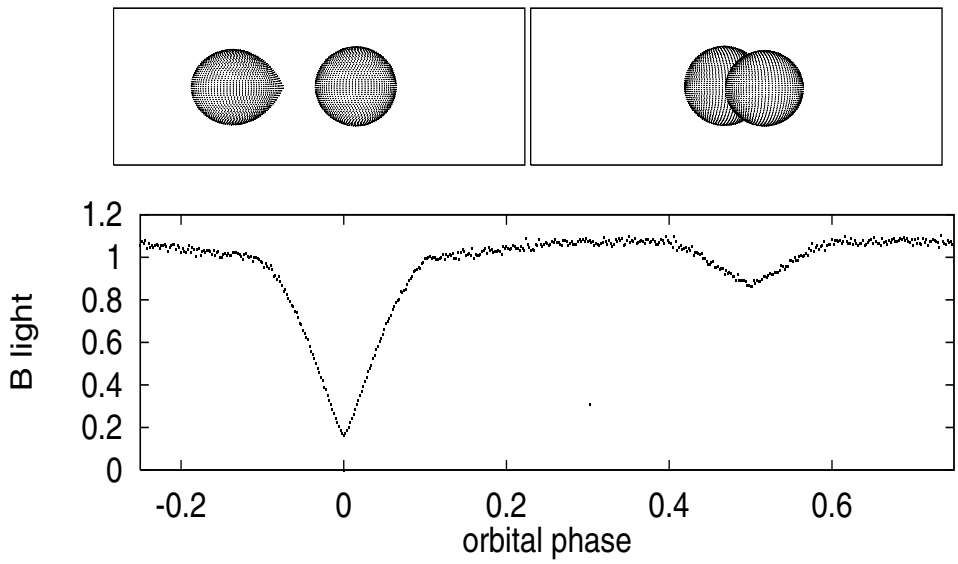


Figure 2. The reflection effect is seen in this simulated blue light curve of Z Vulpeculae (lower panel). Notice the hill around secondary eclipse as the heated inward-facing side of the lower-temperature star turns toward the observer. The low-temperature star is the one with the large tide, as shown in the computer pictures of the top panels. It is nearly the same size as its hot companion, but far less massive and therefore much more distorted.

simple binaries can be theoretically intricate. Nature makes many contenders for the title of strangest binary, but β Lyr and ϵ Aur have continued as leading candidates for most of a century. Subsets of their observations look deceptively simple, but straightforward interpretations lead quickly to contradictions. Binaries that break the rules make us re-examine our assumptions. Light curves also differ in intricacy, apart from whether the behavior is understood. Certain red giants with small hot companions (SYMBIOTIC STARS, especially symbiotic novae) show eclipses of fluorescent stellar winds and chromospheres that exhibit great detail, especially in spectroscopy and multiwavelength photometry. Although the fundamental phenomenon seems to be understood, much is being

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

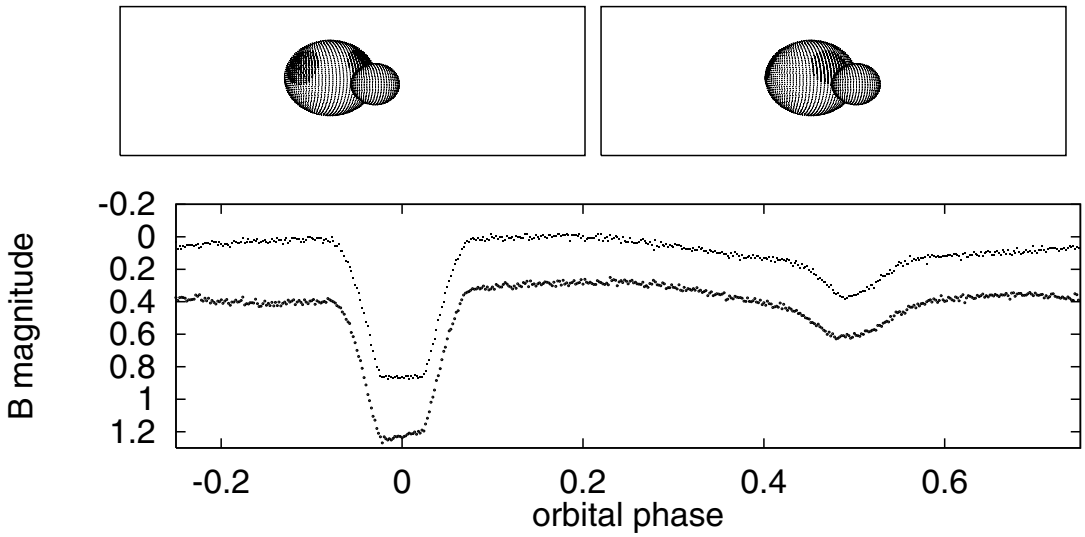


Figure 3. Here the RS CVn type binary AR Lacertae shows its spots. Of course, spot details are unknown, so the simulation shows two broad areas that represent heavily spotted regions, according to the approximate situations in 1977 (top left) and 1981 (top right). Changes from 1977 (upper curve) to 1981 can easily be seen in the machine-generated light curves of the lower panel.

Table 1. Parameters determined from eclipsing binary light curves.

Parameter	Comment
Orbital period, P	$\mathrm{d}P/\mathrm{d}t$ may also be a parameter
Reference time, t_0	Usually the time of a conjunction
Orbital inclination, i	$i = 0^\circ$ in plane of sky
Surface potentials, ψ_1, ψ_2	Potentials specify surface dimensions (*e.g.* R_1, R_2)
Orbital eccentricity, e	e =(distance between foci)/(length of major axis)
Argument of periastron, ω	Angle between line of nodes and orbit's major axis
Gravity brightening exponents, β_1, β_2	Symbols and definitions differ among authors
Bolometric albedos, A_1, A_2	Discussed in text
Mean surface effective temperatures, T_1, T_2	Temperatures that would exist without reflection effect
Mass ratio, M_2/M_1	Determinable only in favorable circumstances
Bandpass luminosities, L_1, L_2	Conventionally normalized to their sum
Limb darkening coefficients, x_1, x_2, y_1, y_2	Various approximate relations commonly used
Spot latitude, longitude, size, temperature	Spots are circular in most models
Third light, l_3	Represents any third source—usually a star

learned by matching good models to observations.

Star figures and system morphology

Tidal dissipation has dragged many eclipsing binaries into SYNCHRONOUS ROTATION (for brevity, co-rotation), which for circular orbits means that axial rotation and orbital revolution have the same period. Their tidal distortions are static in a co-rotating frame, with stretching along the line of centers. Rotation produces polar flattening, so the combined tidal and rotational figure is egg like with the narrow end toward the companion and the polar dimension smaller than the side dimension. The shape can often be approximated by a triaxial ellipsoid (a *geometrical model*), although not in the more distorted cases.

Today's *physical models* are based on level surfaces of constant potential energy per unit mass that can be larger, smaller or accurately match certain critical surfaces that have important roles in light curve analysis as well as in binary star evolution. The configuration for co-rotation, where the entire system rotates with one angular velocity, is most easily viewed in the co-rotating frame. A null effective gravity point must exist on the line of centers between the stars where the opposing star gravities plus centrifugal force add to zero. It marks a saddle point in potential, with the downhill direction away from the null gravity point toward both stars. The *inner critical surface* passes through the null point and is the largest equipotential a star can fill without spilling gas onto its companion. It consists of a *limiting lobe* for each star, and the term *contact* in binary star morphology refers to contact of a star with its lobe rather than of one star with another. Ordinarily a star cannot be larger than its limiting lobe because it would then rapidly lose material, although stable overcontact configurations exist

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with both stars exceeding their limiting lobes, neither star being a willing receptacle for material from the other. If undercontact, accurate contact and overcontact are respectively symbolized **U**, **A** and **O**, then the binary combinations **UO** and **AO** are effectively ruled out except for extremely brief intervals. Otherwise pressurized layers would face onto a vacuum. The morphological categories that are ordinarily attainable are called *detached* (**UU**), *semidetached* (**UA**) and *overcontact* (**OO**). No known evolutionary process leads to the **AA** case in co-rotating binaries and any evolutionary size change would convert **AA** to **UU**, **UA** or **OO**. However, the **AA** case (*double contact*) probably occurs in non-synchronous binaries when evolutionary expansion causes one star to fill its lobe while fast rotation causes the other lobe to shrink onto its star. RZ Sct and U Cep may be examples. There also is an outer critical surface that marks the largest an overcontact binary can be without having catastrophic flow from a 'back' null point of effective gravity. That point is behind the less massive star on the line of centers. All cases are illustrated in figure 4.

Analysis

A modest number of simple rules connect light curves with system characteristics. Assume the time scale to be in orbital phase (*i.e.* folded on the period) and the brightness scale to be in light (not stellar magnitude). Then for circular orbits the two eclipses have equal durations, while for eccentric orbits the eclipse that occurs nearer to apastron is the longer one. In the circular case for spherical stars, the sine of half the phase duration of either eclipse, $\sin \Theta_e$, is a lower limit on the sum of relative radii, with direct application if $i = 90^\circ$. Allowing for inclination, the relation is $(R_1 + R_2)/a = (\cos^2 i + \sin^2 i \sin^2 \Theta_e)^{1/2}$. An obvious point is that the ratio of light lost to light remaining at the bottom of a total eclipse is essentially the luminosity ratio of the smaller to the larger star, in a given observational bandpass. This rule becomes virtually exact in the absence of tides and mutual irradiation and allows exceptionally accurate determination of luminosity *ratios*. The ratio of radii is determined most strongly from a reasonably deep annular eclipse, combined with a total eclipse. Then the ratio of light loss in the annular eclipse to the light remaining in totality is nearly the square of the ratio of smaller to larger star radii, with small discrepancies due to limb darkening and other brightness distribution effects. For circular orbits, the eclipses are equally spaced, by half a cycle, while for eccentric orbits they are equally spaced only if we 'look along the major axis', with $\omega = 90^\circ$ or 270°. The argument of periastron ω tells how the orbit is situated in angle within its own plane. If the orbit rotates (apsidal motion) then the eclipses have excursions about the 0.0 and 0.5 cycle points that are 180° out of phase (one goes 'left' while the other goes 'right'). The position angle of the nodal line, Ω, cannot be found from a light or radial velocity curve because neither is affected by a rotation of the binary about the line of sight. However, Ω can be found from polarimetric variation for a few binaries.

In a physical light curve model, the gravitational potential energies of the two stars, ordinarily considered as point masses, and centrifugal potential energy are added to form a total potential energy per unit mass, ψ, with

$$\psi = \psi(\mathcal{R}, \theta, \phi, M_1, M_2, F_r, D). \qquad (1)$$

Here θ, ϕ are latitude and longitude and M_1, M_2 are the masses. One usually works with a dimensionless version whose length unit is the orbital semimajor axis, a, and in which mass ratio replaces the absolute masses. The rotation parameter F_r and instantaneous separation D generalize ψ to non-synchronism and eccentric orbits at the cost of certain, usually inconsequential, approximations. To specify a star surface, the radial coordinate, $\mathcal{R}$, is required to follow a surface of constant ψ. Equation (1), or in practice the dimensionless form, is numerically inverted with fixed $\psi, M_2/M_1, F_r, D$ to compute a relative radial coordinate, $r = \mathcal{R}/a$, as a function of θ and ϕ (i.e. to compute a grid of surface points). Thus equation (1) defines a family of surfaces in the parameter ψ, one of which corresponds to the actual surface of a given star, with ψ_1, ψ_2 replacing the mean star radii of geometrical models as star size parameters. The magnitude and direction of the potential gradient, $\nabla\psi$, respectively determine local effective gravity and surface orientation. Many schemes have been published for integrating the light from these surfaces while accounting for eclipses and the various surface brightness effects. Such computations became practical with the machines of ca 1970.

Morphology can be built into light curve analyses. For example, it might be reasonably certain that a binary is semidetached, although a discussion of reasons would exceed available space. If so, the contact star's surface potential is not a free parameter but is the lobe potential and can be calculated from other parameters, most notably the mass ratio. The functional relation thus defined can be used to eliminate an entire dimension of incorrect solutions. The lobe filling condition can be exploited not only for general solution improvement but sometimes to determine particular parameters that otherwise would not be derivable. The prime example is the *photometric mass ratio*. Ordinarily a mass ratio estimate requires some means to find the relative location of the center of mass, which a light curve does not provide. However, light curves of semidetached and overcontact binaries do tell the location of the null point of effective gravity, which is determined by the mass ratio. Lobe size is fixed by the condition that the lobe contains the null point. Thus, in a semidetached binary, star size determines lobe size, which in turn determines mass ratio. For overcontact, the dumbbell's overall shape (ratio of bulge sizes) is conditioned by the pinch of its waist around the null point, so essentially R_2/R_1 determines mass ratio. A common misconception is that a photometric mass ratio mainly derives from ellipsoidal variation. However, the actual information sources are eclipses, particularly total-annular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

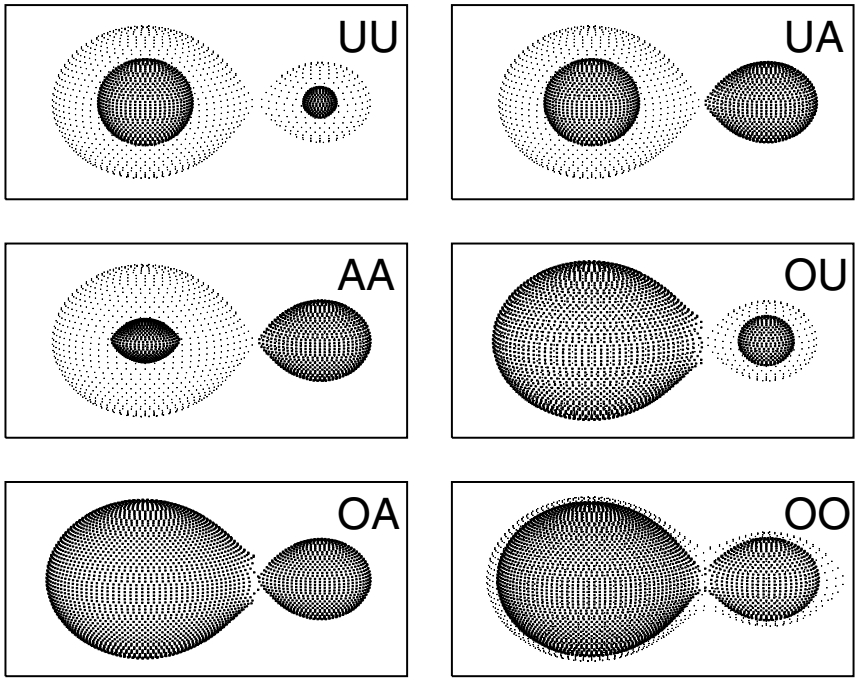


Figure 4. All combinations of undercontact (**U**), accurate contact (**A**) and overcontact (**O**) are illustrated. Stars are shown with the larger dots and limiting lobes as ghostly surfaces (smaller dots). The well-known detached (**UU**), semidetached (**UA**) and overcontact (**OO**) cases are at top left, top right and lower right, respectively. The overcontact binary is surrounded by an outer contact surface that limits its maximum size. The newer category of double contact (**AA**) is at middle left, where the actual limiting lobe coincides with the star, and the ordinary (co-rotation) lobe also is shown for reference. Double contact binaries have one component limited in size by fast rotation. The **OU** (middle right) and **OA** (lower left) configurations are ordinarily unphysical, as discussed in the text.

eclipses, that determine accurate relative star dimensions. A photometric mass ratio allows one to find all the mass and absolute dimension information that normally requires velocity observations of both components, even when velocities of only one component exist.

Recent and future developments

A new data source has appeared in spinoff from the several gravitational microlens surveys that automatically produce thousands of light curves. Although precision is not as good as for data gathered in the usual way and radial velocities are lacking, observers can sift through the survey light curves to find interesting binaries for more thorough observation. Extra-Galactic eclipsing binaries thus identified and subsequently observed with large telescopes can improve the HUBBLE CONSTANT. Stellar interferometry promises a new era of inroads into the domain dominated by eclipsing binaries. In principle, interferometric images could replace light curve analyses, although the baselines and general capabilities needed for typical distant examples seem beyond forseeable expectations. However, major instrumental advances are on the way. In the nearer term, advanced interferometry will forge connections between the fields of eclipsing and visual binaries.

Although one naturally thinks of eclipses in the observational context of light curves, there are direct links of eclipses to other observations. One link is spectral disentangling, a computationally intensive use of eclipses to isolate entangled binary spectra. Several distinct procedures exist and all achieve good separation of apparently hopelessly blended spectra, sometimes even when there is no phase without serious blending. The essential idea is that each eclipse modulates only one of a blended pair of spectra and does so in a known or determinable way. Disentangling schemes average over random noise very effectively because they not only fit entire spectral windows (as opposed to individual lines) but fit all spectra simultaneously. Another kind of simultaneous solution fits combinations of *types* of observations, such as light, velocity and polarimetry curves. The main advantage is reduced parameter count, as some parameters (*e.g.* eccentricity, mass ratio, inclination) are common to two or more data types. Still another development is rotation measurement from light curves rather than from spectral line broadening, which can sometimes be done because rotation affects star figures and surface brightness distributions and therefore affects eclipse circumstances.

Omissions due to space limitation are covered by the bibliography and include background areas such as history, contributors, the role of computers, 'badly behaved' binaries, famous classical problems,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

interconnections with spectroscopic binaries and full descriptions of eclipsing binary categories. Also covered there are further ways to observe eclipses, such as by POLARIMETRY. Disks and other circumstellar structures as well as extended atmospheres and spots of various kinds are also mentioned in the bibliography. However, proper treatments of these items and of exotic new discoveries incorporating white dwarfs, neutron stars and black holes are to be found only in the general literature.

Bibliography

Kallrath J and Milone E F 1999 *Eclipsing Binary Stars, Modeling and Analysis* (Berlin: Springer)

Kopal Z 1959 *Close Binary Systems* (New York: Wiley)

Kuiper G P and Johnson J R 1956 Dimensions of contact surfaces in close binaries *Astrophys. J.* **123** 90

Smith R C 1984 The theory of contact binaries *Q. J. R. Astron. Soc.* **25** 405

Wilson R E 1994 Binary star light curve models *Publ. Astron. Soc. Pac.* **106** 921

R E Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ecliptic

A great circle on the celestial sphere which represents the apparent annual path of the Sun in its motion relative to the background stars. It is so called because eclipses can occur when the Moon crosses it. Due to the motion of the Earth around the Sun, the Sun appears to move across the celestial sphere, completing one full circuit in a year. The ecliptic, in fact, represents the intersection of the orbital plane of the Earth with the celestial sphere and, because the equator of the Earth is inclined by an angle of approximately $23\frac{1}{2}^{\circ}$ to the orbital plane, it follows that the ecliptic is inclined to the celestial equator by the same amount (the obliquity of the ecliptic). The ecliptic intersects the celestial equator at two points, the vernal equinox and the autumnal equinox.

See also: celestial equator, celestial poles, celestial sphere, equinox, obliquity of the ecliptic.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ecosphere

The shell-shaped region around a star within which the surface temperature of a suitable planetary body is conducive to the existence of life. Both the diameter and the thickness of the ecosphere depend on the type of star, in particular on its surface temperature. In the solar system the ecosphere extends from just outside the orbit of Venus to just within the orbit of Mars, making Earth the only suitable planet for life, by this definition. (Note that the Moon is in the Sun's ecosphere, but does not possess the conditions to make it suitable for life.) In general, hotter stars have a larger and wider ecosphere; cooler stars a smaller and narrower one. A dwarf star, though, would have a much smaller and narrower ecosphere than a giant star with the same surface temperature.

The traditional concept of an ecosphere as defined above was framed to apply to surface life (on land or in surface oceans) and to lifeforms similar to those familiar on the Earth. In the 1990s it became clear that 'life' in its broadest sense may cover a wider spectrum and may exist in more unusual environments. Micro-organisms known as extremophiles have been found which, as the name implies, thrive in extreme environments, in particular at temperatures much higher or lower than other organisms can tolerate. In addition, there are deep-ocean and subsurface organisms which do not rely on photosynthesis, and can withstand temperatures of 1100 °C and 1200 °C, respectively, at depths of 5 km. It has been speculated that life may exist in the subsurface ocean of Jupiter's satellite Europa, for example. If life does turn out to be possible in this and other 'extreme' environments, the standard definition of ecosphere will need to be rewritten.

See also: life on other worlds

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eddington, Sir Arthur Stanley (1882–1944)

Astrophysicist, born in Kendal, Westmorland, England, became Plumian professor of astronomy and director of the Cambridge Observatory. Eddington's work on the theory of relativity was described by EINSTEIN as 'the finest presentation of the subject in any language' and from Greenwich, he led one of the two 1919 solar eclipse expeditions which confirmed the predicted deflection of starlight by gravity. His lifetime's work concerned the internal structure of stars. He discovered the mass–luminosity relationship, calculated the abundance of hydrogen and explained the pulsation of Cepheid variable stars and the very high densities of white dwarfs. He was one of the first to identify nuclear reactions as the source of power in stars. He studied the pressure of the transport of energy through a star by radiation, and showed that above a certain luminosity, the *Eddington limit*, a star cannot through the force of gravity maintain its equilibrium and will disrupt. In his later years he pointed out some surprising numerical coincidences between atomic and cosmological quantities, but his cosmological theory about the reasons why has not been accepted at all.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Edgeworth, Kenneth Essex (1880–1972)

Amateur astronomer, born in Streete, County Westmeathe, Ireland. His 1943 paper 'The evolution of our planetary system' is the first reference to a reservoir of comets beyond the planets, a theory which foreshadowed the concept of the KUIPER belt, sometimes referred to in Ireland and Britain as the *Edgeworth–Kuiper belt*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Education

In surveys of first-year undergraduate science students at British universities, astronomy and space stand out most prominently as the areas of science that attracted the students into science. There is something grand and inspiring about the Universe that motivates young people, and indeed the not-so-young, to study astronomy to a greater or lesser degree.

Science is now taught in virtually all schools. In general astronomy is regarded as a minor part, but it is regarded as important to understand the basic astronomical facts such as that the Earth is spherical, that its rotation causes day and night, that it orbits the Sun and its tilt during the orbit produces the seasons, that we live in a galaxy of stars and that there are other galaxies in the Universe.

Motivated school students can take astronomy as a separate examination subject in many countries; in Britain, for example, there is an annual paper in astronomy in the portfolio of tests (the GCSE) that are taken at about age 15. This encourages the candidate to undertake actual observational programs as well as learning from books. Significantly, however, only the most elementary mathematics is included in the British GCSE—and astronomy is above a mathematical subject.

Local astronomical societies are invaluable in education, because many of them have observatories of their own (see AMATEUR ASTRONOMICAL SOCIETIES) and nearly all of them organise lecture programs with qualified speakers. The programmes may be informal or may be structured as an adult education course.

Of course, public observatories are of immense importance, and in some countries are both common and highly successful—notably in the United States and Japan. In Britain they are rare, perhaps because there is the problem that during a 'public night', when the telescopes are made available, the skies are too often cloudy! A series of robot telescopes now exists for education, placed in good observing sites. In California the most advanced network is based at Mt Wilson Observatory and is called Telescopes in Education (http://tie.jpl.nasa.gov/). The UK is building on Hawaii the 2-metre telescope Faulkes Telescope for public access (http://www.faulkes.com/).

One very important development concerns the spread of planetaria. A PLANETARIUM provides an 'artificial sky'; the viewer sits inside a large dome, and the celestial bodies are shown by means of a special projector. The effect can be amazingly realistic, and in a large planetarium it is hard to remember that you are not looking at the real sky. There are hundreds of planetaria world wide, see the International Planetarium Society at http://www.ips-planetarium.org/

One great advantage of a planetarium is that its sky can be made to show phenomena which the real sky cannot, except under special circumstances, for instance, a total eclipse of the Sun. The sky can be shown as it will be at any hour, and from any part of the world; time travel is easy—one can, for instance, see the sky as it was at the time of the Crucifixion. The movements of the planets can be speeded up, and even the proper motions of the stars can be shown, whereas the real stars are so far away that the constellation patterns remain sensibly unchanged for periods of many lifetimes.

Planetaria have traditionally been based on opto-mechanical projection devices. Recent advances in computer technology are making possible digital projectors, controlled in almost infinite variety by software. The possible shows are almost unrestricted in subject matter, and can range from the purely fun (simulated trips into a black hole for example) to the seriously technical (e.g., teaching navigation).

Many universities cater for astronomy in one way or another, so that a degree can be taken either in 'pure' astronomy or in a modular form (physics and astronomy being a favorite) (see PROFESSION). About 50 UK universities teach astronomy modules in specialist science degrees (i.e. half of them). In the USA, where degree level studies have a higher diversification than in the UK, astronomy is often a popular choice for the science component of a liberal arts degree and for study as a special subject in science degrees.

National astronomical societies (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS) have education subsections, such as the American Astronomical Society's Working Group on Astronomical Education (http://www.aas.org/~wgae/wgae_index.html) and the Royal Astronomical Society's Education Committee. The Astronomical Society of the Pacific is targeted explicitly towards the dissemination of information to the public including through educational programmes (http://www.aspsky.org/education.html).

The International Astronomical Union, through its Commission 46 on the Teaching of Astronomy (http://physics.open.ac.uk/IAU46/) supports a travelling telescope and summer schools for students of astronomy in the developing world, where there is increasing interest. In South Africa, the new African government supports astronomy education as a means to develop the aspirations of the South African people into its modern era.

The US Association for Astronomy Education (http://solar.physics.montana.edu/aae/), the UK Association for Astronomy Education (supported by the Royal Astronomical Society) and the European Association for Astronomy Education (based at the European Southern Observatory) offer events and teacher training in astronomy (http://www.star.ucl.ac.uk/~aae/aaehomep.htm), with the avowed intent to modernise astronomy education in schools.

Paul Murdin and Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Effective Temperature Scale and Bolometric Corrections

The conversion from an observational quantity, such as the color index or the spectral type, to the effective temperature ($T_{\rm eff}$) of a star is known as the effective TEMPERATURE SCALE. Bolometric corrections are required in the calculation of the luminosity of a star if the flux from the star has not been observed over the entire ELECTROMAGNETIC SPECTRUM.

The $T_{\rm eff}$ of a star is defined to be the temperature of a BLACKBODY RADIATOR with the same radius and the same total energy output (i.e. luminosity) as the star. Practically speaking, the $T_{\rm eff}$ corresponds quite closely to the temperature of the visible surface of a star. Knowledge of any two of the three variables, $T_{\rm eff}$, radius and luminosity, allows calculation of the third variable. Thus to calculate directly the $T_{\rm eff}$ of a star, both the luminosity and the radius of a star must be known. These two quantities can be found for only a limited number of stars. The luminosity of a star requires knowledge of the bolometric flux of a star (which includes radiation across the entire electromagnetic spectrum) and thus requires observations from space, and is well known for only a few hundred stars. Fluxes are thus for the most part measured in only a limited region of the spectrum—usually the optical. To derive bolometric fluxes from these measurements, a correction, termed a bolometric correction, is required. The radius of a star requires knowledge of the angular diameter of a star and the distance to the star. Again, this quantity is known for only a limited number of stars. Thus, for most stars, the $T_{\rm eff}$ cannot be determined in a fundamental way. Astronomers must therefore rely on other methods, including measuring a proxy for the $T_{\rm eff}$ (for instance the color index or the spectral type), and then finding a way to convert from this proxy to the $T_{\rm eff}$. This conversion is called the effective temperature scale.

The determination of the effective temperature

The most straightforward way to derive the effective temperature scale, i.e. the conversion between an observational quantity such as the B–V color index, or the spectral type of a star, and the $T_{\rm eff}$ is to determine the $T_{\rm eff}$ for as many stars as possible over a wide range of temperatures, and then relate these effective temperatures to the observed quantity. How does one determine the $T_{\rm eff}$ of a star?

Fundamental effective temperatures

The fundamental way to derive the $T_{\rm eff}$ of a star is to use the Stefan–Boltzmann law in the following form:

$$L = 4\pi\sigma R^2 T_{\rm eff}^4 \qquad (1)$$

where L is the luminosity of the star, σ is the Stefan–Boltzmann constant and R is the radius of the star. Unfortunately, neither the luminosity nor the radius can be measured directly. What can be measured from the Earth are the monochromatic flux, F_ν, which is the energy from the star incident on a unit area on the surface of the Earth per unit time per unit frequency interval, and the angular diameter (θ_D) of the star. These quantities can be related to the $T_{\rm eff}$ in the following way:

If $\mathcal{F}_\nu$ represents the flux emitted from the stellar surface, and F_ν represents, as above, the flux from the star measured at the Earth, then it necessarily follows, from the conservation of energy (ignoring interstellar extinction and absorption by the atmosphere of the Earth), that

$$4\pi r^2 F_\nu = 4\pi R^2 \mathcal{F}_\nu$$

where r is the distance to the star. Integrating the flux at the Earth over the entire electromagnetic spectrum yields

$$\int_0^\infty F_\nu \, {\rm d}\nu = F_{\rm bol}$$

where $F_{\rm bol}$ is the total or bolometric flux from the star, and

$$L = 4\pi R^2 \int_0^\infty \mathcal{F}_\nu \, {\rm d}\nu = 4\pi\sigma R^2 T_{\rm eff}^4.$$

Consequently

$$F_{\rm bol} = (R/r)^2 \sigma T_{\rm eff}^4.$$

Now, the angular diameter of a star, measured from the Earth is $\theta_D = 2(R/r)$. Thus, finally

$$F_{\rm bol} = \tfrac{1}{4}\theta_D^2 \sigma T_{\rm eff}^4.$$

Angular diameters can be measured for nearby stars using a number of techniques, including optical phase-coherent or amplitude interferometry, intensity interferometry, speckle interferometry (see INTERFEROMETRY: GROUND) and lunar OCCULTATIONS. It is possible to determine actual radii of stars from the analysis of ECLIPSING BINARY STARS. To determine the bolometric flux of a star, its flux must be observed over a wide range of frequencies in the electromagnetic spectrum. For most stars (except for the very hottest stars), the majority of the flux from the star is concentrated in the wavelength range ~100 nm (ultraviolet) to ~12 μm (infrared). This requires observations from spacecraft, as the ultraviolet and much of the infrared are strongly absorbed by the Earth's atmosphere. Spacecraft which have been useful in this regard include the International Ultraviolet Explorer (IUE), the Sky Survey Telescope on the ESRO TD1 satelite, the Infra-Red Astronomical Satellite (IRAS), the Infrared Space Observatory (ISO) and the Hubble Space Telescope (HST).

Bolometric corrections

If the bolometric flux of a star has not been measured directly, it is still possible to estimate it from the stellar flux over a much more limited wavelength range using bolometric corrections. Bolometric corrections are usually defined in terms of magnitudes. Thus, if M_V represents the absolute magnitude of a star measured in the Johnson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

photometric V-band, and $M_{\rm bol}$ the absolute bolometric magnitude calculated from the entire spectrum, then the bolometric correction ($B.C.$) is defined as

$$B.C. \equiv M_{\rm bol} - M_V. \quad (2)$$

Then the luminosity of the star can be calculated from the equation

$$\log\left(\frac{L_*}{L_\odot}\right) = 0.4[M_V(\odot) - M_V(*) + B.C.(\odot) - B.C.(*)] \quad (3)$$

where the $\odot$ quantities refer to the Sun. If the radius of the star is known, then the $T_{\rm eff}$ follows from equation (3) and equation (1).

Tables of bolometric corrections may be calculated using stars for which complete flux distributions are available, or from theoretical stellar flux distributions. Equation (2) can be rewritten

$$B.C. = 2.5 \log \frac{\int_0^\infty F_\nu S_\nu \, {\rm d}\nu}{\int_0^\infty F_\nu \, {\rm d}\nu} + C_1$$

where S_ν is the photometer response for the wavelength band of interest (in equation (2), the Johnson V-band). C_1 is an arbitrary constant which is usually selected according to convention. The most commonly used convention selects C_1 so that the $B.C.(\odot) = -0.07$. Another convention selects C_1 so that the minimum $B.C.$ is zero, and yet another sets C_1 so that the $B.C.(\odot) = -0.19$. Which convention is chosen (in light of equation (3), where the *difference* in the bolometric corrections of the Sun and the star is the relevant quantity) is not important, just so long as one is consistent. Bolometric corrections are usually tabulated against B–V or the spectral type. When tabulated against B–V the $B.C.$ does not show a strong dependence on the luminosity type of the star. Table 1 gives a recent tabulation of the bolometric correction versus B–V for dwarf and giant stars. Supergiants deviate slightly from the values given in this table. Table 2 tabulates the bolometric correction versus spectral type. In this table separate scales for the different luminosity classes are given. In both tables the convention that $B.C.(\odot) = -0.07$ is adopted. Notice that for both cool and hot stars, the magnitude of the bolometric correction is quite large. This is because, for these stars, most of the flux is outside the optical region. For such stars, the bolometric correction is not very reliable. Instead of equation (2) for the definition of the bolometric correction some sources use the equation $B.C. \equiv M_V - M_{\rm bol}$, in which case the bolometric corrections are the negatives of those given in tables 1 and 2.

Other methods of determining the effective temperature

Many stars are too far away for the measurement of their angular diameters, and/or many have not been adequately observed in the ultraviolet or in the infrared. For these stars other techniques are available to determine their effective temperatures, but not in a fundamental way.

Table 1. $T_{\rm eff}$ and bolometric corrections as functions of B–V.

B–V	$B.C.$	$T_{\rm eff}$ (K)
−0.35	−4.72	56 700
−0.30	−3.23	33 600
−0.25	−2.18	22 700
−0.20	−1.53	17 000
−0.15	−1.01	13 700
−0.10	−0.61	11 700
−0.05	−0.33	10 400
0.00	−0.16	9530
0.05	−0.05	8920
0.10	0.00	8450
0.15	0.02	8080
0.20	0.03	7770
0.25	0.04	7480
0.30	0.03	7220
0.35	0.03	6960
0.40	0.02	6730
0.45	0.01	6500
0.50	−0.01	6280
0.55	−0.04	6080
0.60	−0.06	5890
0.65	−0.09	5720
0.70	−0.12	5560
0.75	−0.16	5410
0.80	−0.20	5280
0.90	−0.29	5050
1.00	−0.38	4840
1.10	−0.49	4660
1.20	−0.61	4480
1.30	−0.75	4320
1.40	−0.92	4150
1.50	−1.15	3980
1.60	−1.54	3770
1.70	−2.46	3460
1.80	−5.54	2940

For all but the hottest stars the observed flux distribution (i.e. F_ν or F_λ versus wavelength) of a star may be compared with theoretical flux distributions from model STELLAR ATMOSPHERES. The best determinations of $T_{\rm eff}$ using this method are for stars which have been observed over a wide wavelength range. Figure 1 shows an example of such a fit for the star β Leo, which has been observed in the ultraviolet (using the IUE satellite), and the optical and near-infrared using ground-based telescopes. Even if a star has not been observed outside the optical range, ground-based observations can be used to obtain an adequate fit with theoretical flux distributions. Because theoretical models are used, this method of determining the $T_{\rm eff}$ cannot be regarded as fundamental; the derived effective temperatures are only as good as the models.

For very hot stars (for example the early O-type stars), the $T_{\rm eff}$ is so great that the maximum of the continuum flux of these stars is in the extreme UV, and thus even space-based observations fall on the long-wavelength or Rayleigh–Jeans tail of the flux distributions of these stars, the slope of which is very insensitive to the $T_{\rm eff}$. For such stars, it is necessary to adopt an entirely different technique to estimate the effective temperature. This technique

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

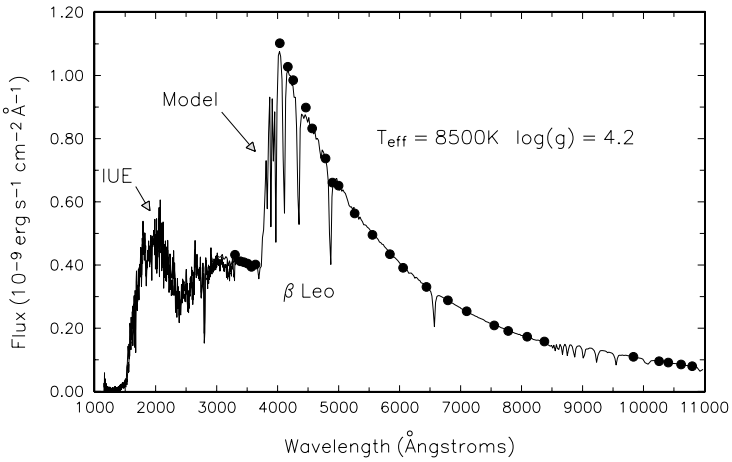


Figure 1. A model fit to the observed fluxes of the star β Leo. The full circles indicate ground-based spectrophotometry in the optical and near-infrared. The data marked IUE (extending from 1100 Å to 3200 Å) represent an ultraviolet spectrum of the star obtained with the International Ultraviolet Explorer (IUE) satellite, while the solid line extending from 1100 Å to 11000 Å (and partially obscured by the IUE data) represents the fitted model calculated with the model stellar atmospheres program ATLAS9. The effective temperature of the star is determined to be 8500 K.

involves fitting the spectral line profiles of lines of singly ionized helium (He II), neutral helium (He I) and hydrogen with theoretical line profiles. Such comparisons require the calculation of non-LTE stellar atmosphere models which include the effects of radiation scattered back from the stellar wind onto the photosphere, the so-called wind blanketing effect. Including the effects of wind blanketing can change the effective temperature by as much as 10%.

Lacking an observed flux distribution, certain spectral features can be used to determine the effective temperature. For instance, in the B-type stars the size of the Balmer jump, a discontinuity near 3650 Å in the stellar flux due to the photoionization of hydrogen, is temperature sensitive. In the F- and G-type stars, the profiles of the Balmer lines of hydrogen are quite sensitive to the $T_{\rm eff}$. In the late-type stars (G- and K-type stars), ratios of the line depth of lines due to neutral metals can be quite sensitive to the $T_{\rm eff}$ if one of the lines has a low excitation potential while the other has a relatively high excitation potential. In such a case, the line with the low excitation potential grows rapidly in strength with decreasing $T_{\rm eff}$, while the high excitation potential grows less rapidly or even declines in strength. Such techniques can yield a temperature precision on the order of 20 or 30 K, although the absolute temperature is more uncertain.

The effective temperature scale

Once the effective temperatures of a sufficient number of stars have been determined (preferably in a fundamental way), the relationship between an effective temperature proxy, such as the $B{-}V$ index and the effective temperature, can be tabulated. This is known as the effective temperature scale. One recent tabulation is given in table 1. Notice the unsuitability of the $B{-}V$ color index for this purpose for the very hot stars; a small change in $B{-}V$ yields a large change in the $T_{\rm eff}$. Some improvement can be obtained by using the $U{-}B$ index, but both are subject to the same fundamental limitation mentioned above: for the hottest stars, observations in the optical and even the near-UV lie too far out along the temperature-insensitive Rayleigh–Jeans tail of the flux distribution. As an alternative, tabulations of the $T_{\rm eff}$ and the bolometric corrections are often made against the spectral type. Table 2 contains such a tabulation derived from a number of recent sources. For the cool, luminous stars, effective temperatures are not yet well determined. In the B-, A-, and early F-type stars, the $T_{\rm eff}$ and bolometric corrections are not strongly dependent on the luminosity class.

Applications of the effective temperature scale and future developments

Arguably the most important diagram in stellar astronomy is the HERTZSPRUNG–RUSSELL (HR) DIAGRAM. In its theoretical form, this diagram plots stars in the $T_{\rm eff}$, luminosity plane. In its most common observational form, also known as the color–magnitude diagram, stars are plotted in the color index (commonly $B{-}V$), absolute visual magnitude (M_V) plane. The HR diagram is the arena on which much of modern-day stellar astrophysics takes place. In particular, stellar evolutionary tracks and isochrones calculated from theoretical models are plotted on the HR diagram. To compare these theoretical calculations with observations of real stars and clusters of stars it is necessary to convert from the theoretical quantities of $T_{\rm eff}$ and luminosity to $B{-}V$ and M_V or other similar observed quantities. Indeed, one of the greatest barriers to the advancement of stellar evolutionary theory is the inaccuracy of our knowledge of the effective temperature scale.

Determination of the $T_{\rm eff}$ is often the first step to the derivation of other fundamental parameters of a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. $T_{\rm eff}$ and bolometric corrections as functions of spectral type.

	Luminosity Class V		Luminosity Class III		Luminosity Class I	
Spectral type	$T_{\rm eff}$ (K)	$B.C.$ (mag)	$T_{\rm eff}$ (K)	$B.C$ (mag)	$T_{\rm eff}$ (K)	$B.C.$ (mag)
O3...	46 400	−4.27	46 500	−4.27	44 500	−4.14
O5...	44 300	−4.13	42 300	−3.98	40 300	−3.82
O7...	40 100	−3.81	38 100	−3.64	36 100	−3.46
O9...	35 900	−3.44	34 000	−3.27	32 000	−3.08
B0...	29 800	−2.87	28 000	−2.70	27 000	−2.60
B2...	21 380	−2.04	19 000	−1.78	17 000	−1.53
B5...	15 000	−1.24	15 000	−1.24	13 500	−0.98
A0...	10 000	−0.25	10 000	−0.25	10 000	−0.25
A5...	8200	0.02	8200	0.02	8200	0.02
F0...	7250	0.04	7250	0.04	7250	0.04
F5...	6600	0.01	6500	0.01	6400	0.00
G0...	5850	−0.07	5650	−0.10	5450	−0.15
G5...	5570	−0.12	4980	−0.31	4750	−0.44
K0...	5270	−0.21	4820	−0.39	4510	−0.59
K5...	4560	−0.56	3920	−1.25	3710	−1.68
M0...	4050	−1.05	3730	−1.63		
M2...	3750	−1.60	3610	−1.95		
M4...	3120	−4.15				
M6...	2590					

star. For instance, knowledge of the $T_{\rm eff}$, the luminosity (from M_V and bolometric corrections via equation (3)) and the surface gravity, which can be determined from spectral analysis, leads to the radius and mass of the star. Knowledge of the mass can then be used, in conjunction with stellar evolutionary theory, to determine the age of the star.

The effective temperatures and bolometric corrections for the hottest and the coolest stars are still uncertain. The newly launched FAR ULTRAVIOLET SPECTROSCOPIC EXPLORER (FUSE) satellite (which has a spectral coverage of 905–1195 Å) helps to sample more of the energy distributions of the hottest stars, and can thus help in the determination of the effective temperatures. This, plus continual advances in stellar atmosphere models for these stars, should lead to a significant improvement in our knowledge of the fundamental parameters of these stars in the near future. The determination of fundamental effective temperatures for the coolest stars, especially the HIGH-LUMINOSITY STARS, will likely continue to be problematical, as the very definition of the radius and the $T_{\rm eff}$ in the very extended atmospheres of these stars is difficult.

Bibliography

Berriman G, Reid N and Leggett S K 1992 Effective temperatures of M Dwarfs *Astrophys. J.* **392** L31–3

Fitzpatrick E L and Garmany C D 1990 The H–R diagram of the Large Magellanic Cloud and implications for stellar evolution *Astrophys. J.* **363** 119–30

Flower P J 1996 Transformations from theoretical Hertzsprung–Russell diagrams to color–magnitude diagrams: effective temperatures, $B–V$ colors and bolometric corrections *Astrophys. J.* **469** 355–65

Gray D F 1992 *The Observation and Analysis of Stellar Photospheres* 2nd edn (Cambridge: Cambridge University Press)

Kurucz R L 1992 *ATLAS9 Stellar Atmosphere Programs and 2 km/s Grid* CD-ROM 13 (Cambridge, MA: Smithsonian Astrophysical Observatory)

Richard Gray

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Effelsberg Radio Telescope

The 100 m diameter radio telescope has been operated by the MAX-PLANCK-INSTITUT FÜR RADIOASTRONOMIE since 1972 near Effelsberg, Germany. The telescope is fully steerable and its pioneering 'homologous' design minimizes elevation-dependent gravitational deformations, enabling routine observations to wavelengths as short as 6 mm. Scientific programs include high-sensitivity continuum surveys, extensive polarization studies of our own and external galaxies, pulsar research, and atomic and molecular line spectroscopy. The telescope frequently participates in VLBI observations.

For further information see
http://www.mpifr.bonn-mpg.de/index_e.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Egg Nebula (CRL 2688)

A planetary nebula in the constellation Cygnus, position RA 21^{h} 02.3^{m}, dec. $+36^{\circ}$ $42'$. It is young, possibly having left its red giant phase only a few hundred years ago. The central star, which is obscured by dust, is emitting powerful jets and has shed several shells of material.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Einstein Observatory (HEAO-2)

The second in the series of HIGH ENERGY ASTROPHYSICAL OBSERVATORIES was launched by an Atlas-Centaur rocket on 13 November 1978. Soon after its insertion into a 470 km circular orbit inclined at 23.5º to the equator, HEAO-2 was named the Einstein Observatory, in celebration of the centenary of Albert Einstein's birth.

This groundbreaking NASA X-ray mission revolutionized X-RAY ASTRONOMY. As the first satellite to be equipped with focusing X-ray mirrors, Einstein not only observed sources 10 000 times fainter than its predecessor, UHURU, but it was the first observatory to image X-ray sources.

The four nested paraboloid–hyperboloid mirrors of the grazing-incidence telescope (the largest had a diameter of 58 cm) were able to focus incoming X-rays with energies in the range 0.25–4 keV (see GRAZING INCIDENCE OPTICS). The photons were then analyzed by four interchangeable instruments that could be rotated into the focal plane of the telescope. The field of view varied with the X-ray energy, but was typically about one square degree.

Einstein carried five main instruments:

(1) The High Resolution Imager (HRI) provided an imaging capability with a spatial resolution of approximately 2 arcsec, but had a field of view less than half a degree in diameter.
(2) The Imaging Proportional Counter (IPC) had a greater sensitivity than the HRI and covered the entire field of view of the telescope, but it had a poorer spatial resolution (approximately 1 arcmin) and a very limited spectral capability.
(3) The Solid State Spectrometer (SSS) had a moderate spectral resolution and a 6 arcmin diameter circular field of view.
(4) The Focal Plane Crystal Spectrometer (FPCS) provided high spectral resolution in selected energy bands.
(5) The Monitoring Proportional Counter (MPC) measured the 1–20 keV X-ray flux of the source(s) being observed simultaneously by the telescope.

The leader of the project and the principal investigator for all five instruments was X-ray pioneer RICCARDO GIACCONI.

The first object studied by Einstein was the BLACK HOLE candidate Cygnus X-1. More than 5000 X-ray sources were subsequently catalogued during the observatory's 30-month operational life—including more than 100 hot, young stars in the ORION NEBULA alone. Einstein's ability to detect faint sources meant that it was able to study objects such as SUPERNOVA REMNANTS, X-RAY BINARIES, GALAXY CLUSTERS, ACTIVE GALAXIES and QUASARS. Its high spatial resolution meant that many of these could be linked to optical counterparts for the first time.

Einstein's observations influenced stellar models by showing that the X-ray flux from many stars was thousands of times greater than predicted (see SOLAR–STELLAR CONNECTION). Complex shells of X-ray emitting gas were mapped around several supernova remnants, while 'on–off' emissions from the spinning X-ray PULSAR at the heart of the CRAB NEBULA were also studied. Other supernova remnants were discovered in the LARGE MAGELLANIC CLOUD.

The observatory also confirmed the existence of X-ray sources in both the spiral arms and the nucleus of the ANDROMEDA GALAXY, those near the center being much more luminous. A powerful X-ray source was detected at the center of the radio galaxyCENTAURUS A, while additional X-ray emissions were aligned with its optical jet. Other studies revealed clouds of extremely hot gas surrounding some members of galaxy clusters and filling the space between them (see COOLING FLOWS). Formerly unseen emissions from distant quasars were also revealed by extended observations, indicating that the diffuse X-ray background may be caused by numerous faint sources.

The mission ended in April 1981, when the satellite ran out of attitude control gas. It reentered the atmosphere on 25 March 1982. (see X-RAY TELESCOPES, NEUTRON STARS and NGC 4151)

Peter Bond

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Einstein, Albert (1879–1955)

Physicist born in Ulm, Württemberg, Germany, described the photoelectric effect (for which he received the Nobel prize in 1921) and created the theory of special relativity in 1905 in his spare time, while an employee of the Swiss patent office. The theory of relativity was based on two hypotheses, that the laws of physics had to have the same form in any frame of reference and that the speed of light remained constant in all frames of reference. He deduced that mass and energy were equivalent. Sought to extend the special theory of relativity to phenomena involving acceleration using the principle of equivalence, in which gravitational acceleration was held to be indistinguishable from acceleration caused by mechanical forces. Gravitational mass was therefore 'equivalent' to inertial mass. While a professor in Prague he predicted the gravitational deflection of light, for example how light from a distant star, passing near the Sun, would appear to be bent slightly, shifting the star's apparent position away from the Sun. This was proposed as the first experimental evidence in favour of Einstein's theory and verified by EDDINGTON's and DYSON's observations at the solar eclipse of 1919. Developed the general theory of relativity while at Zurich and Berlin. The success of the 1919 test of relativity began Einstein's establishment as a scientific icon, with the London *Times* headlining its account 'Revolution in science—new theory of the universe—Newtonian ideas overthrown'. By contrast his work in Berlin was disrupted by demonstrations with an anti-Jewish taint, and eventually Einstein took up a post at Princeton as the Nazis took power. At Princeton he began his work to unify the laws of physics, an ambitious, possibly over-ambitious, task which physicists continue. Became active in the peace movement and took on the role of a wise and independent scientific counsellor to the world. His name is taken for many institutions etc, perhaps the most important astronomically being the Einstein x-ray astronomy satellite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

EISCAT

The EISCAT Scientific Association is an international research organization operating three geophysical research incoherent scatter radar systems, an ionospheric heater and a dynasonde in northern Scandinavia. It is funded and operated by the research councils of Norway, Sweden, Finland, Japan, France, the United Kingdom and Germany.

Two of the VHF and UHF radar transmitters are located close to the city of Tromsø in Norway, while the third, which became operational in 1996, is located near Longyearbyen on the Arctic island of Spitzbergen. The radars measure plasma densities, temperatures and velocities in the polar and auroral ionosphere.

For further information see
http://www.eiscat.no/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ejecta

Material thrown out by an impact on the surface of a planetary body, particularly by a crater-forming event. Ejecta consists mostly of material excavated from the body that has been struck. It can form rays, extending radially from the crater; an *ejecta blanket*, a continuous deposit of ejecta surrounding the crater; or secondary craters, caused by larger fragments of ejected debris. Material that has been on a surface tends to darken with age through exposure to radiation, so ejecta, which consists of fresh subsurface material, is frequently bright by contrast.

See also: craters, ray, secondary crater.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Electromagnetic Radiation

Radiation, consisting of an electric and a magnetic disturbance, which travels in a vacuum at a characteristic speed known as the velocity of light (approximately 300 000 km s^{-1}). In other words, a light wave consists of periodically varying changes in electric and magnetic fields.

Usually electromagnetic radiation is thought of as a wave motion, the distance between successive crests being the wavelength. Thus, for example, visible light has a wavelength of a few hundred nanometers (a nanometer $= 10^{-9}$ meters), compared with a radio wave with a wavelength of the order of meters. The number of wavecrests passing a fixed point in one second is the frequency. This equals the velocity of light divided by wavelength. Thus, short wavelength corresponds to high frequency.

The full range of electromagnetic radiation is divided, in order of increasing frequency and decreasing wavelength into: radio waves, microwaves, infrared rays, visible light, ultraviolet light, x-rays and gamma rays.

Electromagnetic radiation may also be described in terms of quanta or photons ('packets' or 'particles' of energy). The energy associated with electromagnetic radiation is inversely proportional to wavelength and directly proportional to frequency. Thus, the shorter the wavelength, the more energetic the photon.

The human eye responds to different wavelengths of light by recognizing different colors (red light has a longer wavelength than blue light).

Most of our information about the universe has been obtained by measurements of electromagnetic radiation of one kind or another reaching us from space.

See also: electromagnetic spectrum, spectrum, synchrotron radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Electromagnetic Spectrum

The complete range of electromagnetic radiation, from the shortest to the longest wavelength. By convention, the electromagnetic spectrum is divided into a number of different wavebands. From the shortest to the longest wavelengths, the broad principal divisions are: gamma rays (wavelengths shorter than 0.01 nm), x-rays (0.01–10 nm), ultraviolet (10–390 nm), visible (390–700 nm), infrared (700 nm–1 mm), radio (1 mm upwards). Although human eyes are sensitive only to wavelengths in the range from about 390 nm to about 700 nm, the 'optical' region of the spectrum is generally taken to encompass the wider waveband from 310–1000 nm (310 nm–1 μm). Further subdivisions in common astronomical usage are: hard x-ray (0.01–0.1 nm), soft x-ray (0.1 nm–10 nm), extreme ultraviolet (EUV, or XUV: 10 nm–91 nm), near-infrared (1–4 μm), mid-infrared (4–40 μm), far-infrared (40 μm–350 μm; i.e. 40 μm–0.35 mm), sub-millimeter (0.35–1 mm), millimeter-wave (1–10 mm), microwave (1 mm–0.3 m).

Most incoming radiation is absorbed by the atmosphere or reflected back into space, but radiation in the wavelength ranges 310 nm–1100 nm, and from about 2 cm to about 30 m, can penetrate to ground level, these bands being known, respectively, as the optical and radio 'windows'. A few narrow bands of radiation in the near- and mid-infrared regions of the spectrum, and radiation in the sub-millimeter and millimeter-wave regions, can be studied from good, high-altitude, observing sites.

See also: electromagnetic radiation, gamma-rays, infrared astronomy, optical astronomy, radioastronomy, ultraviolet astronomy, wavelength, x-rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Electron

A stable elementary particle with a mass of 9.1091×10^{-31} kg (equivalent, in energy terms, to 0.511 MeV) and a negative charge of 1.602×10^{-19} coulomb. The electron has a spin of 1/2 (in units of the Planck constant, h, divided by 2π) and is a member of the lepton family. Electrons can exist either as free particles or as the negatively charged components of atoms. A complete atom contains equal numbers of negatively charged electrons and positively charged protons and is, therefore, electrically neutral. A flow of electrons (for example, along a wire, or in some other conducting medium) constitutes an electric current. For historical reasons, electrons are sometimes referred to as beta particles, and a stream of electrons as beta radiation.

In the simplest model of the atom (the Bohr model), electrons are considered to travel around the atomic nucleus (which consists of protons and neutrons) in certain specific orbits, each of which corresponds to a particular energy level of the atom. If an electron absorbs a quantity of energy equal to the difference in energy between two levels, it will jump up to a higher ('excited') level. When an electron drops from a higher level to a lower one, energy is emitted in the form of a photon. In complex atoms, the constituent electrons are located in concentric shells, the maximum number of electrons in each shell being dictated by quantum mechanics and the Pauli exclusion principle. The maximum permitted number of electrons in the innermost shell is 2 (in this case, the two electrons must have opposite spins), in the second shell, 8, in the third, 18, and so on. When a shell contains the maximum number of electrons, it is said to be filled. The innermost shells usually fill up first, any remaining electrons being available to take part in chemical interactions with other atoms (these electrons are called valence electrons). Atoms with filled shells, such as helium (2 electrons) and neon ($2 + 8 = 10$ electrons), are particularly stable, more energy being needed to raise one of their electrons to a higher level, or to remove it altogether, than is the case with atoms that have partially filled outer shells.

The magnitude of the charge on the electron (or 'electronic charge'), which is denoted by the symbol 'e', is used as the unit for describing the charges on subatomic particles and atomic nuclei.

See also: atom, electronvolt, elementary particles, ionization, leptons, Pauli exclusion principle, photon, proton, quantum mechanics, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Electronvolt

A unit of energy that is used to describe the energies of subatomic particles and photons, or the energy levels of atoms. One electronvolt (symbol eV), which is the kinetic energy gained by an electron when it is accelerated through a potential difference of one volt, is equivalent to 1.602×10^{-19} J. Commonly used multiples are keV (10^3 eV), MeV (10^6 eV), and GeV (10^9 eV).

See also: atom, electron, energy, photon, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Electroweak Force

The united force that encompasses the electromagnetic force and the weak nuclear force. The unification of these two forces is described by a theory that was devised during the 1960s by Sheldon Glashow, Steven Weinberg and Abdus Salam according to which, at high enough energies, the electromagnetic force and the weak nuclear interaction behave in exactly the same way.

The electromagnetic force is conveyed between real particles by virtual photons (photons which come into existence in accordance with the Heisenberg uncertainty principle). According to Glashow, Weinberg, and Salam, the weak nuclear interaction, which gives rise to radioactive decay, is conveyed (or 'mediated') by very massive virtual particles, the W^+, W^-, and Z^0 bosons. The range of a force is determined by the masses of its force-carrying particles (the higher the mass, the shorter the range). Because the electromagnetic force is conveyed by photons, which carry energy but have zero rest-mass, it is a long-range force (a virtual photon of zero energy can travel an infinite distance). Because the W and Z particles have masses in the range 88 to 100 times that of the proton, their range is very small (less than 10^{-17} m). Electroweak theory predicts that at energies in excess of about 100 GeV, W and Z bosons are produced as frequently as photons and the two forces (electromagnetic and weak) behave as a unified force. The discovery of W and Z particles, in high-energy experiments carried out in 1983 using the large proton synchrotron at CERN (the European particle physics laboratory near Geneva), provided strong confirmation of electroweak theory.

See also: fundamental forces, gauge bosons, grand unified theory, uncertainty principle, photon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Element, Chemical

A substance composed of atoms, each of which has the same atomic number (i.e. contains the same number of protons in its nucleus), and which cannot be broken down by chemical processes into two or more other substances. The identity of a chemical element is determined by its atomic number, its chemical properties being determined by the arrangement of the electrons in the various shells that surround the nucleus (in a complete, neutral, atom, the number of orbiting electrons is equal to the number of protons in its nucleus). Although all atomic nuclei of the same element contain the same number of protons, they may contain differing numbers of neutrons and may, therefore, have different mass numbers (the mass number being the total number of protons and neutrons in the nucleus). Atoms of the same element with different mass numbers are called isotopes.

Each chemical element is denoted by a chemical symbol. For example, hydrogen is denoted by H, helium by He, carbon by C, iron by Fe, and so on. The basic symbol is amplified by the addition of the mass number (A) as a superscript and the atomic number (Z) as a subscript. Thus, an element, X, would be denoted, symbolically, by ${}^{A}_{Z}\mathrm{X}$. For example, hydrogen (with one proton and no neutrons in its nucleus) is represented by ${}^{1}_{1}\mathrm{H}$, helium (with two protons and two neutrons, so that $A = 4$ and $Z = 2$), by ${}^{4}_{2}\mathrm{He}$, iron (26 protons, 30 neutrons), by ${}^{30}_{26}\mathrm{Fe}$, and so on. Of the 92 'naturally occurring' elements, the heaviest of which is uranium, 90 have been found on Earth. The remaining two (technetium and prometheum) have been synthesized experimentally and have also been identified in the spectra of certain stars. In addition, there are at least 15 transuranic ('beyond uranium') elements, elements with atomic numbers greater than 92, which have been synthesized by bombarding various isotopes with (usually) neutrons or alpha particles. A few of the transuranic elements, all of which are radioactive, have been found in nature in microscopic trace quantities.

See also: atom, cosmic abundance of elements, electron, isotope, neutron, proton, radioactivity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elementary Particles

Particles that at one time were thought to be the most basic, indivisible, building blocks of atoms and all forms of matter. This category of particle includes the stable leptons (i.e. electrons and neutrinos), the stable baryons (i.e. protons and neutrons) and photons. While electrons and neutrinos are still considered to be truly fundamental particles, which have no internal structure and cannot be broken down into smaller components, protons and neutrons are known to be composed of particles called quarks and are not, therefore, fundamental.

See also: baryons, electron, fundamental particles, hadrons, leptons, neutrinos, neutron, photon, proton, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ellipse

The conic section obtained by cutting a right circular cone by a plane which makes an angle relative to the base of the cone less than the angle made by the side. It is an oval curve having within it two points or foci (singular: focus) such that the sum of the distances from each focus to any point on the curve is a constant. The greatest diameter of an ellipse is the major axis (half of this is the semi-major axis), and the least diameter the minor axis. The ellipse is symmetrical about both its axes. The curve when rotated about either axis forms the surface called the ellipsoid of revolution, or a spheroid.

The foci are located on the major axis, one on either side of the center of the ellipse; the greater the separation of the foci, the more flattened the ellipse. A measure of the separation of the foci is given by the eccentricity, e, which can take values between 0 (the circle) and 1 (the case of the parabola). For an ellipse of semi-major axis a and eccentricity e, the distance from the center of the ellipse to either focus is ae.

The path of a heavenly body moving around another in a closed orbit in accordance with Newton's gravitational law is an ellipse. In the solar system one focus of such a path about the Sun is the Sun itself.

See also: conic sections, hyperbola, orbital elements, parabola.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elliptical Galaxies

Elliptical galaxies are smooth, quiescent star-piles, devoid of any of the spectacular structures found in SPIRAL GALAXIES. They have no disk, no spiral structure and only small amounts of the gas and dust. As a result there are no obvious symptoms of continuing star formation: no H II regions or young star clusters. They are the simplest galactic systems comprising just a single component, relatively bright in the center but fading rapidly with increasing radius. Elliptical galaxies are found mostly in the denser regions of the universe, from rich clusters to small groups; truely isolated ellipticals being relatively rare. Figure 1 shows the central region of the Coma cluster which is dominated by a population of ellipticals. The most luminous elliptical galaxies are amongst the brightest galaxies we know of and this, combined with their apparent simplicity, has made them targets of detailed study, not only so that we can understand their own evolution but also so that they can be used as standard candles to determine distances.

The simple structure of elliptical galaxies is reflected in their place in HUBBLE'S CLASSIFICATION. They are characterized by a single number, the ellipticity $\varepsilon = 10(1 - b/a)$, where b and a are the projected angular extent of the short and long axis of the galaxy on the sky. E5 galaxies have a long axis twice that of the short axis and are almost the flattest ellipticals. Some galaxies have been classified E6 and E7 but these have almost always been found to harbor stellar disks. Because the classification depends on the *projected* axial ratio it is not a physical classification. By making the simple assumption that ellipticals have the two long axes equal with the third axis shorter (i.e. they are oblate, like a tangerine) Hubble was able to invert the observed distribution of projected shapes to estimate the true distribution. He estimated that on average ellipticals have a short axis about 65% of the length of the long axis with a dispersion of about 15%. This picture remained current until the late 1970s but has been dramatically revised since then. In recent years attempts to produce a physical classification for ellipticals, and new ways of estimating their intrinsic shapes, have been developed (see later).

In this article I will limit myself to a discussion of classical elliptical galaxies. These range from the brightest galaxies, which are about 40 times as bright as the Milky Way, to dwarf galaxies that are 100 times fainter than the Milky Way and that are typically found as companions to more massive galaxies (e.g. M32, the companion to the Andromeda nebula, M31). I will not include DWARF SPHEROIDAL GALAXIES which are sometimes, erroneously, refered to as dwarf ellipticals. Luminous elliptical galaxies have roughly the same colors as K-giant stars giving them a yellow-orange hue; less luminous galaxies have slightly bluer colours. In appearance and spectral characteristics the bulges of spiral galaxies resemble low-luminosity elliptical galaxies and historically the bulge component of M31 was taken as typical of ellipticals. There are many parallels between ellipticals and bulges and I will compare and contrast them throughout this article.

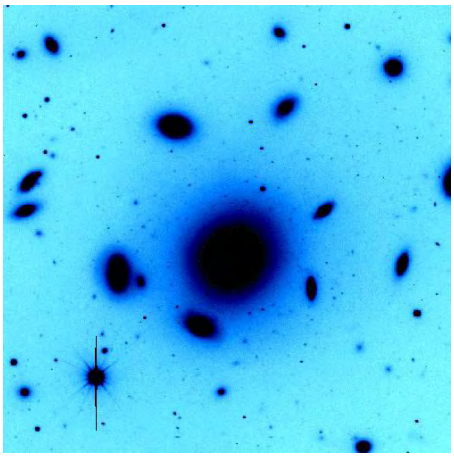


Figure 1. The center of the Coma cluster of galaxies, dominated by the bright elliptical galaxy NGC 4874, which is surrounded by a swarm of fainter, featureless elliptical galaxies. (Image courtesy of the Isaac Newton Group of Telescopes, La Palma and the SMAC (Streaming Motions of Abell Clusters) team.)

In the next section I will draw together the ways in which we have determined the distribution of light and mass as a function of radius within elliptical galaxies, including a discussion of attempts to refine our views of the intrinsic shapes of elliptical galaxies and attempts to detect their massive dark halos. I follow that with a discussion of the special properties of elliptical galaxy cores, where we often find separate components, even rotating in the opposite direction to the bulk of the galaxy, will be discussed. I also discuss the extensive searches for black holes in the centers of ellipticals. Then I describe how we learn about the evolution of the stars that make up ellipticals and how the large-scale properties of elliptical galaxies scale with each other: luminosity, mass, surface density, color, heavy element abundance and rotation. Finally I bring together all of the facets of elliptical galaxies and present the current thinking about how elliptical galaxies formed and evolved.

Structure, size, luminosity and mass

Elliptical galaxies are so called because the contours of constant intensity (called isophotes) are concentric ellipses. This is a very good description of most elliptical galaxies but there are small but significant variations from this simple form. High-quality images show (a) systematic changes in ellipticity with increasing radius, the most common form being a steady increase in the flattening and (b) that the position angle of the major axis of the isophotes changes with increasing radius. In most cases these changes are small (one or two ellipticity classes or a few degrees) but in a small minority they can be dramatic.

The sizes of ellipticals are measured by fitting an analytic form to the fall in brightness (averaged in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

concentric circular annuli) with increasing radius: the luminosity profile. The most commonly used form was proposed by Gerard de Vaucouleurs in 1948 on the basis of images taken on photographic plates. He proposed

$$I(R) = I_e \exp\{-7.67[(R/R_e)^{1/4} - 1]\}$$
$$= I_e 10^{\{-3.33[(R/R_e)^{1/4}-1]\}}$$

where R_e, the *effective radius*, is the radius that contains half the total light of the galaxy and I_e is the surface brightness (the amount of light from a square arc second of the galaxy) at R_e. This law can be integrated to give a finite total light of $I_{\rm tot} = 7.22\pi {R_e}^2 I_e$. de Vaucouleurs' law, as it is known, fits remarkably well to many of the most accurately determined luminosity profiles of elliptical galaxies and provides a convenient measure of the size and total luminosity of galaxies. Effective radii range from ∼20–30 kpc for giant galaxies to hundreds of pc for the smallest. Total blue luminosities range from 100 billion solar luminosities ($M_B \sim -24$) to 25 million solar luminosities ($M_B \sim -15$). A generalization of this law, proposed by Sersic, which replaces the $\frac{1}{4}$ power by another free parameter $\frac{1}{n}$, fits a wider range of galaxy luminosities and by allowing $n = 1$ even fits the luminosity profile of spiral galaxies well. Other laws have been proposed to describe the presence of a constant surface brightness core in some elliptical galaxies and to account for the extended low surface brightness haloes found in others. Details of these are given in Binney and Merrifield (1998) which is an excellent reference for further reading.

As well as ellipticity changes and isophote twists, elliptical galaxy isophotes exhibit small deviations from pure ellipses that have become the basis for a more physical classification of elliptical galaxies. Accurate determinations of the boxy or disky nature of isophotes have been made consistently even though they arise from contributions to the light of at most a few per cent. Remarkably many properties of elliptical galaxies vary depending on the type and strength of these isophote distortions and recent work links the disky ellipticals in a sequence with lenticular galaxies leaving boxy ellipticals as a separate class. The kinematics, stellar populations, radio power and x-ray emission from elliptical galaxies all show a dependence on isophote shape and these global differences have led to sugestions that these two flavors of elliptical galaxy have different formation histories. Further clues to the formation history of elliptical galaxies are found in the faint, incomplete, shells that have been identified around many galaxies (see figure 2). These weak features may occur in as many as 30% of elliptical galaxies and are thought to be the remnant of a small disk galaxy that has fallen into an elliptical and been destroyed.

Determining the luminosity profile is the first step to calculating the mass of elliptical galaxies. The simplest assumption is that the light traces the mass and so, for a given assumed intrinsic shape, the observed surface brightness (which is the integral along the line-of-sight of the luminosity density) is related to the projected surface mass density by a constant factor, the mass-to-light ratio (M/L) of the stars.

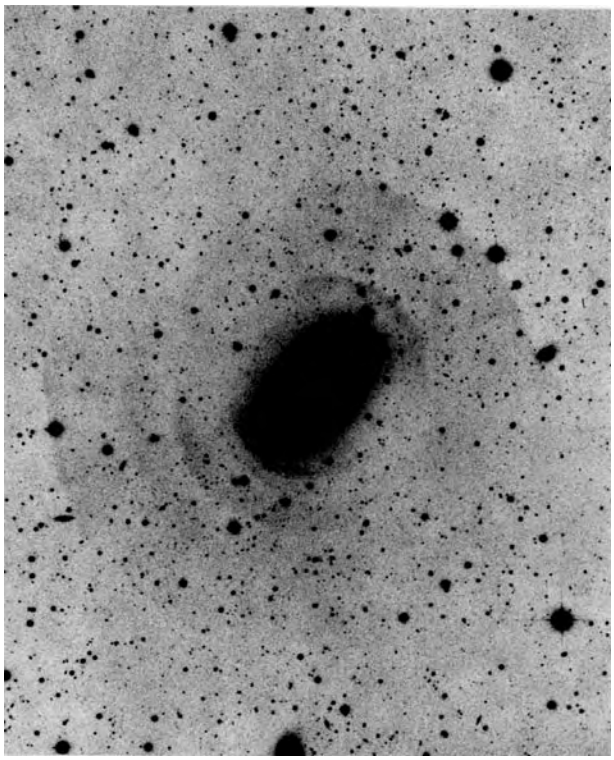


Figure 2. The spectacular shell system surrounding the elliptical galaxy NGC3923. Careful photographic processing, known as unsharp masking, first revealed this extensive set of shells which continues into the central region of the galaxy. (Copyright Anglo-Australian Observatory, photograph by David Malin.)

Kinematics and intrinsic shapes

Measurements of the rotation and random motions of the stars in ellipticals are made from the shifts and width of absorption lines in the spectra of the galaxies along the major axis. The random motions are characterized by the value of the dispersion of the Gaussian that best fits the width of the absorption lines: the velocity dispersion, σ. Early measurements revealed that more luminous galaxies have broader lines such that the total luminosity $L \propto \sigma^4$. This is known as the Faber–Jackson relation after the first people to draw attention to this in 1975. This discovery ignited an interest in using measurements of galaxy velocity dispersions to estimate the intrinsic luminosity of ellipticals and thus determine their distances. This possibility stimulated researchers to make surveys of galaxy dispersions and these produced a much better understanding of the relationships between their structure, dynamics and stellar populations that will be discussed under 'Global properties'.

Until the late 1970s it was assumed that the flattened figures of elliptical galaxies arose because the stars in the equatorial plane are supported by their rotation about the galaxy minor axis. However, as soon as the rotation velocities of the most luminous elliptical galaxies could be measured it was discovered that they are too small to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

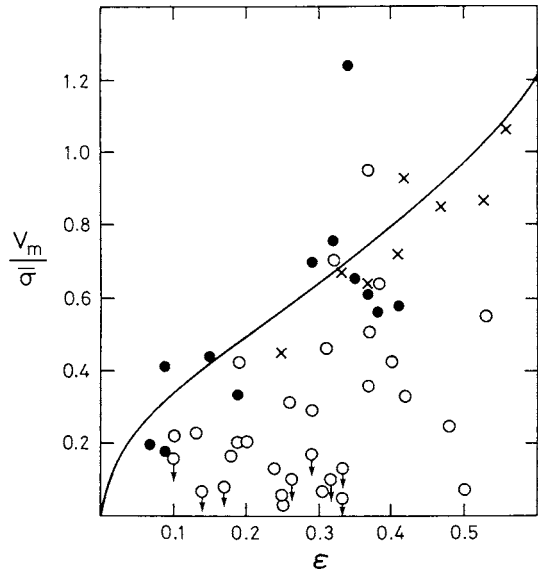


Figure 3. The maximum rotation velocity (V_m) normalized to the average velocity dispersion ($\bar{\sigma}$) plotted against apparent ellipticity (ε) for a sample of luminous ellipticals (open circles), faint ellipticals (full circles) and the bulges of spiral galaxies (crosses). The curve represents the expected relation between rotation and flattening for oblate galaxies with isotropic velocity dispersions that are supported by rotation. The bulges and faint ellipticals appear to conform to this model but the luminous ellipticals have significantly lower rotation velocities. Note that when galaxies are viewed edge-on the ellipticity is greatest and they are expected to have the highest rotation velocities; as they are viewed closer to pole-on they are seen to be rounder and V_m is also reduced. Thus if a galaxy is viewed at a range of inclinations to the line-of-sight it is expected to remain roughly on the predicted line. (Figure from Davies R L *et al* 1983 *Astrophys. J.* **266** 41.)

account for the observed flattened figures. Furthermore, when kinematic maps of ellipticals were made some dramatic examples of substantial rotation on the minor axis were found. This discovery led to a reassessment of the nature of elliptical galaxies. Figure 3 shows the rotation velocity, normalized to the central velocity dispersion, plotted against ellipticity for luminous elliptical galaxies, fainter ellipticals and the bulges of edge-on spiral galaxies. Also plotted is the expected relation for galaxies that are oblate and rotationally supported. Luminous elliptical galaxies are deficient in rotation by a large factor. Rotation is dynamically unimportant for luminous ellipticals; they are supported by the *random* motions of the stars. These random motions are not isotropic and this anisotropy produces the flattened figures observed. This discovery also challenged the assumption that ellipticals have oblate shapes and opened up the possibility that they could be prolate (two short axes equal, like a sausage) or even triaxial where all three axes have different lengths (like most potatoes).

We now believe that most luminous elliptical galaxies are triaxial and that the observed isophote twisting arises from the projection onto the sky of the changing intrinsic axial ratio with radius. This hypothesis also accounts for the rotation sometimes observed on the projected minor axis. This view has been established statistically from both the distribution of observed shapes of thousands of galaxies and from observations and models of the kinematics of many tens of galaxies. The average intrinsic shape for luminous elliptical galaxies is, however, poorly constrained. Most estimates are close to perfectly triaxial (where the intermediate length axis is the average of the long and short), e.g. a:b:c::1:0.85:0.7 with considerable scatter. These statistical studies suggest that intrinsically spherical galaxies are rare but, for example, cannot rule out a distribution of intrinsic shapes that is a mixture of oblate and prolate figures. The discovery that elliptical galaxies are not flattened by their rotation has important implications for our ideas about how they form. Models where isolated massive clouds of gas collapse dissipationally and form stars rapidly always produce galaxies which rotate faster than is observed. These results strengthened the view that elliptical galaxy formation proceeds through a sequence of mergers (of objects on misaligned orbits), to produce a galaxy with low total low angular momentum.

Figure 3 also shows that the less luminous elliptical galaxies and the bulges of spiral galaxies fall on or near the line describing isotropic oblate galaxies, showing that these objects have substantial kinematic similarities and are both qualitatively different from their more luminous brethren. Whereas luminous elliptical galaxies have complex kinematic structures, figure 3 shows that low-luminosity elliptical galaxies and the bulges of spiral galaxies have isotropic velocity dispersions and oblate figures flattened by rotation. There are many other characteristics of elliptical galaxies that vary with luminosity and that I will discuss in the section on 'Global properties'.

Anisotropy and mass-to-light ratio

Precise determination of the distribution of mass within elliptical galaxies requires detailed measurements of their luminosity profile and kinematics together with self-consistent dynamical models. Both analytic models based on the Jeans equations and models that construct galaxies by selecting orbits from a library (Schwarztschild's method) have been successfully used (STELLAR DYNAMICS). For luminous ellipticals infering the distribution of mass within a galaxy on the basis of these data and models is complicated by the degeneracy between changes in M/L, and changes in the orbital anisotropy of the stars. Using only measurements of the rotation and dispersion we cannot distinguish between an increase in M/L and an increase in one component of the velocity dispersion at the expense of another. Techniques based on measuring the *shape* of absorption line profiles have been developed to break this degeneracy. The shape of absorption lines is determined by the luminosity distribution and the number of stars at each velocity along the line of sight. In the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

center of a galaxy an increase in dispersion may indicate an increase in M/L *or* an increase in the fraction of radial orbits. The latter will, however, produce a line profile that is less sharply peaked than a Gaussian. Thus by measuring deviations from a Gaussian profile as well as its width we can determine whether stars are predominantly on tangential or radial orbits.

This ambiguity can be avoided altogether for low-luminosity ellipticals or spiral bulges as we believe them to have isotropic dispersions so determining the mass distributions in these systems is more straightforward. Similarly if galaxies possess disks of excited or ionized gas we can trace the mass distribution in the same way as for spiral galaxies.

A review of the topics covered in this section can be found in de Zeeuw and Franx (1992).

Dark haloes: mass at large radius

Spiral galaxy rotation curves have been used to show that they are embedded in dark haloes that contain three to ten times as much mass as is visible in stars, gas and dust. Can dark haloes be detected in elliptical galaxies? The velocity dispersion in many galaxies remains roughly constant (outside the central region) to one or more effective radii. As galaxies get rapidly fainter away from the center does this mean we are detecting a dark halo or is the fraction of stars on circular orbits increasing with radius? In the outer parts of a galaxy, if there are more stars on circular orbits the absorption lines will be less peaked than a Gaussian and have truncated wings. In a few elliptical galaxies where the measurements extend to two to three effective radii the line of sight velocity profile has been used to eliminate the possibility of tangential anisotropy and indicate the presence of a dark halo. Similar conclusions are drawn from the velocities of GLOBULAR CLUSTERS and PLANETARY NEBULAE at large radii in elliptical galaxies.

An important probe of the mass distribution at large radius is the hot x-ray emitting gas that has been detected out to five to ten effective radii in a large number of elliptical galaxies. This gas is enriched in heavy elements and is thought to originate from the stellar component of the galaxies through mass loss during the late stages of stellar evolution. By assuming that this gas is isothermal and in hydrostatic equilibrium it is possible to estimate the mass out to very large radii from the distribution of x-ray emission. These measurements have confirmed that elliptical galaxies reside in massive haloes qualitatively similar to those of spiral galaxies. With the advent of more capable x-ray satellites that can determine the temperature of the x-ray gas the simple assumptions required to estimate masses have been confirmed for a small number of galaxies.

Amongst the most luminous elliptical galaxies are those that possess an extended (compared with a de Vaucouleurs law) low surface brightness halo, known as 'D-type' galaxies, often found in the centers of rich clusters when they are classified morphologically as 'cD galaxies'. As well as high luminosity and extended halos these galaxies often have more than one nucleus, that is several smaller galaxies are superimposed on the image of the galaxy center. cD galaxies appear to be the product of many interactions or mergers of cluster galaxies with the central dominant galaxy. Several attempts have been made to measure the stellar kinematics in these out to large radii to establish the M/L at large radii and determine whether the stars in the halo are moving in the cluster potential. In a few such galaxies we observe the velocity dispersion rising with increasing radius; however, they do not reach the cluster dispersion, nor do they reach the velocity dispersion expected on the basis of the velocities of the multiple nuclei, so the interpretation of these motions remains controversial.

The centers of ellipticals

The center of a galaxy is a unique environment. The potential well is deepest at the center and any material that dissipates energy in collisions, such as dust or gas, will fall to the center relatively rapidly. For a giant galaxy this will take at most a few hundred million years. Galaxies of all types exhibit special phenomena at their centers. The high-resolution Hubble Space Telescope surveys of the centers of ellipticals often reveal patches or disks of dust and central stellar disks. The centers sometimes host powerful radio sources (see ACTIVE GALACTIC NUCLEI) or even star formation. We believe that radio galaxies are powered by massive black holes, and as a result the centers of elliptical galaxies have been identified as likely sites of black holes.

Black hole searches

Studies of quasars and galaxies at high redshift reveal that active nuclei were much more common at high redshift ($z \sim 2$), sufficiently common that we infer that most luminous present-day galaxies underwent an active phase and therefore should contain a massive central black hole. Searches for central black holes started in the late 1970s with observations of the luminous radio galaxy M87 (pictured in M87: THE MASSIVE GALAXY). From the rapidly increasing central velocity dispersion it was concluded that a black hole of 3 billion solar masses was present in the nucleus. The interpretation of these data was controversial because of the uncertainty introduced by the degeneracy between orbital anisotropy and mass distribution. These important observations stimulated a great deal of investigation of anisotropy and stability, new models were produced and more refined observations were made that increased our confidence in the case for a massive black hole in M87. In the 1990s measurements by the Hubble Space Telescope (HST) of the velocity of the excited gas close to the nucleus revealed a gradient in velocity of ± 500 km s^{-1} across the central 40 pc of the galaxy implying a mass of $2.4 \times 10^9 M_\odot$ contained within a radius of 20 pc. This provided strong confirmation of the black hole hypothesis independent of the uncertainties concerning the mix of stellar orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Hubble Space Telescope (HST) has carried out extensive imaging surveys of galactic nuclei finding many galaxies that have stellar cusps (luminosity profiles that continue to increase in surface brightness to the limit of resolution of the telescope). Other galaxies have been found with resolved cores where the luminosity profile reaches a constant brightness at a finite radius. These measurements are important for evaluating the value of any central mass but a nuclear cusp alone cannot be used to conclude that a black hole is present. When combined with kinematic measurements, made mostly from the ground, we have been able to address the more general question of the frequency of black holes in elliptical galactic nuclei. More than 20 galaxies have been studied in detail. A general trend has emerged: more massive galaxies host more massive black holes. The black hole has roughly 0.2–0.6% of the total galactic mass. This relationship is very uncertain and exhibits a factor of three scatter in the black hole mass, but if true it has wide-ranging implications (see ACTIVE GALAXIES: OBSERVATIONS). However, not all galaxies are *required* by the observations to host a black hole.

A good review of the evidence for supermassive black holes in galaxies of all kinds can be found in Kormendy and Richstone (1995).

Kinematically decoupled cores

As measurements of rotation velocities and velocity dispersions were made with greater resolution and precision an increasingly large number of galaxies were found to have central kinematics that were decoupled from those of the bulk of the galaxy. The most dramatic examples have the central region (typically 0.1–0.3R_e) rotating in the opposite direction from the rest of the galaxy, some have cores rotating about an axis less misaligned, in yet others the cores have very much higher rotation and lower velocity dispersion than the main galaxy. The latter suggests that there is a central disk embedded in the elliptical. These components often occur in the less luminous elliptical galaxies together with with disky isophotes and stellar cusps. In addition to these extended components the HST surveys of galactic cores have revealed many small (100 pc) disk components in ellipticals. Decoupled kinematics are not limited to ellipticals, one early-type DISK GALAXY has been found where two identical co-extensive disks rotate with the same velocity but in opposite directions!

Although at first such decoupled systems were considered rare, taking account of projection effects and the difficulty of detecting small or weak components we now believe that between one- and two-thirds of all elliptical galaxies have extensive central components with decoupled kinematics (i.e. that are a kiloparsec or so across and have been discovered from the ground). The most likely origin for the decoupled central components in ellipticals would appear to be through the accretion of material with different specific angular momentum from that of the main body of the galaxy. The existence of decoupled cores has been widely cited as evidence that elliptical galaxies have experienced significant accretions in the past; however, the colors, metallicity and age of the stellar populations appear to be indistinguishable from those of the main galaxy itself. We will return to the possible origin of these components in the final section.

Stellar populations

In the Milky Way we have used the concept of STELLAR POPULATIONS as an important tool in unraveling our Galaxy's history. We would like to extend these ideas to external galaxies. The colors and absorption line strengths measured in the spectra of the integrated light of elliptical galaxies are rather similar to those of the bulge of the Milky Way. Such considerations led us to believe that elliptical galaxies are old, single stellar populations devoid of star formation, which are now largely quiescent. The oldest stars in the Milky Way are found in the globular cluster population. Elliptical galaxies possess globular cluster populations in proportion to their luminosity (see GLOBULAR CLUSTERS) and the colors of these populations suggest that a substantial fraction of them are as old as the globular clusters in the Milky Way, i.e. the oldest stars in elliptical galaxies are as old as the oldest stars in our galaxy.

Even the nearest luminous elliptical galaxies cannot be resolved into individual stars from the ground, and only a handful can be resolved with HST so we are forced to try to unravel the integrated light from all the stars. By analysing the integrated spectrum we aim to discover what kind of stars constitute an elliptical galaxy. We approach this by using a library of stellar spectra or model stellar atmospheres covering a range of luminosity and effective temperature. We combine these using our knowledge of the relative numbers of stars in a population of a particular age and metallicity based on evolutionary models (see STELLAR EVOLUTION, STELLAR ATMOSPHERES). In this way we attempt to synthesize the observed integrated spectrum from known stellar components. In the optical ($\lambda = 0.5\ \mu$m) we find that about half the light arises from main sequence stars and half from giant stars but that ellipticals have an excess of light in the near infrared ($\lambda = 1$–$2\ \mu$m) and in the ultraviolet ($\lambda = 0.2\ \mu$m). The former can be attributed to stars with a higher metal content than the Sun, whereas the latter are likely to arise from extreme horizontal branch stars (see also HOT STELLAR POPULATIONS IN ELLIPTICAL GALAXIES). Both types are missing from the local stellar libraries.

We would like to infer the star formation history of ellipticals from their integrated spectra, but assigning a unique age and metallicity to an elliptical galaxy using this technique has proved remarkably troublesome. The result of increasing the age of the stars in the models is very similar to that of increasing their metallicity; both make the spectra redder and increase the strength of absorption lines. In recent years techniques using the strength of the hydrogen Balmer lines as age indicators have enabled separate age and metallicities to be determined and the initial applications of these methods suggest that elliptical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

galaxies in rich clusters form a constant age sequence (say roughly 10 Gyr old) with metallicity varying from three times solar in the most luminous galaxies to one-third solar in those 100 times fainter. Samples of galaxies drawn from lower-density environments exhibit a much wider range in age, some having luminosity weighted ages as young as 2 Gyr! The symptoms of a young stellar population (<5 Gyr old) are easier to detect beacuse younger stars are much more luminous (per unit mass) than older stars so even a small mass fraction can significantly modify the integrated spectrum of an elliptical. These results suggest an environmental dependence on the star formation history of ellipticals such that galaxies in dense regions formed completely at early epochs whereas those in less dense environments have continued to experience star formation (perhaps involving only a small fraction of the total mass) until recent times. With the advent of HST and large ground-based telescopes it has become possible to observe galaxies at high redshift in sufficient detail to evaluate how they have changed over a considerable fraction of the age of the universe, and therefore measure galaxy evolution directly. I will discuss these powerful observations in the final section.

Global properties

Almost all the physical properties of elliptical galaxies vary smoothly with increasing luminosity. I have already discussed how the degree of rotational support decreases with increasing luminosity, a fundamental difference in the orbital make-up of faint and luminous ellipticals. This trend was only discovered through detailed study of the kinematics of tens of elliptical galaxies. While this work is becoming easier with the development of area spectrographs such as SAURON, many more galaxies have been characterized by single central spectra. These measurements have resulted in the discovery of a number of global relationships between properties that are easily measured for hundreds of galaxies. Remarkably there are tight correlations between parameters which are measured locally at the center of the galaxy, such as velocity dispersion, and global parameters such as size and brightness. I summarize these global relations here. They form much of the basis for our understanding of how elliptical galaxies formed and evolved.

The color–magnitude relation

The most luminous elliptical galaxies are slightly redder than the faintest ellipticals. This is a small effect but it has very small scatter (a few hundredths of a magnitude) and is well established in nearby clusters of galaxies. There has been a debate over the origin of this effect, is it due to a higher average metal content in the luminous galaxies or are they on average older? The color–magnitude relation has been observed as a tight correlation in clusters at redshifts in excess of 0.5 and this makes the interpretation in terms of age changes unlikely. This result is established in clusters, the simplest assumption to make is that the color–magnitude relation is universal, but this is not yet proven. We believe that the color–magnitude effect arises because the deeper potential wells of luminous galaxies can retain the metals produced in supernova explosions more effectively.

The Mg–σ relation

This correlation involves the strength of the magnesium absorption features at a wavelength of 0.52 μm and the stellar velocity dispersion σ. Galaxies with high-velocity dispersions have stronger magnesium absorption lines. (This is the best developed correlation but other absorption line strengths have also been used.) This relation also exhibits very little scatter, <0.02 magnitudes, and because it is a distance-independent correlation galaxies from a wider range of environments can be included. Once again the correlation could arise from changes in age or metallicity, but measurements in rich clusters out to $z = 0.4$ are consistent with the local relation changed only through the stellar evolution that occurs in that time interval (about 4 Gyr). This suggests that the correlation arises because high-σ galaxies are metal rich. Tests of this relation in different galactic environments have failed to find any systematic differences.

The Faber–Jackson relation and the Fundamental Plane

Earlier I introduced the Faber–Jackson relation: the increase in central velocity dispersion with luminosity, $L \propto \sigma^4$. In the 1980s surveys of elliptical galaxies were made to measure the non-Hubble motions of galaxies. Using the high-quality data collected it became apparent that an extra parameter, the galaxy surface brightness, increased the precision of the L, σ^4 relation as a distance indicator. Thus a new relationship, which has become known as 'the Fundamental Plane' superceded the L, σ^4 relation. The Fundamental Plane relation is:

$$\log R_e = 1.25 \log \sigma + 0.33 \langle SB_e \rangle + \text{constant}.$$

where SB_e is the effective surface brightness I_e expressed in magnitudes and $\langle SB_e \rangle$ is the average surface brightness within R_e so that $\langle SB_e \rangle = 3.61 SB_e$.

In the size, velocity dispersion, surface brightness space elliptical galaxies are not distributed uniformly, they fall on the above plane with relatively small intrinsic scatter of about 15% in R_e. The Fundamental Plane (FP) is close to what would be expected on the basis of the virial theorem if all galaxies have the same M/L ratio. The difference with the observed relation can be used to restrict the possible range in M/L of ellipticals to a factor of three. Similarly, the width of the plane can be used to restrict the scatter in M/L at any point in the plane to be less than 12%.

The scatter in the FP and the Mg–σ relations are sensitive to the range in *both* the age and metallicity of the stellar population. However, the scatter in the FP is more sensitive to the range of ages, whereas the scatter in the Mg–σ relation is more sensitive to the range of metallicities. Using these two together the range in the age and metallicity of ellipticals has been restricted to 30%,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

i.e. luminous ellipticals vary from say 10–13 billion years old and have metallicities in the range 3 ± 1 times solar abundance.

Structural properties
Many structural properties vary smoothly with luminosity or central velocity dispersion. I have discussed most of them already but bring them together here. Statistically, large, luminous, high-velocity dispersion elliptical galaxies have higher x-ray luminosity, host more powerful radio sources, have triaxial figures and anisotropic velocity distributions and have boxy isophotes and flat central luminosity profiles. Alternatively the small, low-luminosity ellipticals and spiral bulges have lower x-ray and radio luminosities, have oblate figures and isotropic velocity distributions, and have disky isophotes and cuspy central luminosity profiles. These differences have been used to argue that luminous-boxy and faint-disky galaxies formed and evolved differently and should be part of separate morphological sequences. This hypothesis is attractive but has yet to address how two physically distinct classes obey the same color–magnitude, FP and Mg–σ relations if indeed careful study proves that they do.

The morphology–density relation
Elliptical galaxies are gregarious, they are most often found in regions of high galaxy density. The mix of galactic morphological types is a strong function of environment, in rich clusters approximately 40% of the galaxies are ellipticals and another 50% are lenticulars with the remaining 10% being spiral galaxies. In less dense environments the fraction of ellipticals falls dramatically while the spiral fraction increases. In the lowest density environments less than 10% of galaxies are classified as elliptical with spiral and lenticular galaxies being found in roughly equal proportions. The same analysis has been carried out for clusters of galaxies at $z \sim 0.4$ imaged by the HST. The ellipticals still dominate the galaxy population in rich clusters at $z \sim 0.4$ but the fraction of lenticular galaxies has dropped to 20%, with spirals accounting for 30% of the rich cluster population. It is clear that in rich clusters the elliptical galaxy population is already in place by $z \sim 0.4$ but that there are very many fewer early type galaxies with disks than we find in rich clusters today.

How and when did elliptical galaxies form?

We believe that the distribution of galaxies traces out the large-scale structure of dark matter and that the dark matter haloes of galaxies assemble hierarchically with the most massive haloes forming most recently. Only through understanding the behavior of the baryons within these haloes can we hope to understand how the luminous galaxies we observe evolve.

The high central densities and relaxed structure of elliptical galaxies are taken as strong evidence that they underwent substantial gaseous dissipation during formation. Furthermore the small scatter observed in their global relations suggests that they are old (with a formation redshift greater than $z = 2$–3) and coeval. These considerations have led to models in which elliptical galaxies form through the dissipative collapse of a massive gas cloud resulting in a huge starburst which turns almost all the gas in a halo into stars simultaneously, say 10 billion years ago. The most recent measurements at sub-millimeter wavelengths have detected galactic masses of cool dust at $z = 2$–3 that are forming stars at a very high rate. These sub-millimeter sources have been identified as the progenitors of the most luminous cluster ellipticals we see today. However, models show that the monolithic collapse of a massive gas cloud produces an unrealistically rapid rotation in the end product and furthermore the collapse of massive objects at high redshift appears to contradict the hierarchical assembly of galaxies.

These drawbacks have led to alternative suggestions that ellipical galaxies formed through a sequence of galactic mergers. There is much empirical evidence for this: low rotation velocities, decoupled cores and shells are all a natural consequence of accretions and mergers. In the alternative model disk galaxies merge together to generate elliptical galaxies during the formation of groups and clusters of galaxies. Theoretical investigations of galaxy mergers using n-body techniques have successfully demonstrated that disk galaxies undergoing mergers will produce elliptical galaxies with de Vaucouleurs' law luminosity profiles and low rotation velocities. In cosmological simulations the bulk of the stars that comprise present day ellipticals form in merger events that occur typically at $z = 1$–2. The galaxies continue to grow in mass and luminosity through mergers until they fall into massive clusters where the typical velocities of galaxies get so large that mergers are inhibited. This scenario accounts for why ellipticals are predominantly found in high-density regions and predicts that on average ellipticals in rich clusters will be older (in the sense of a luminosity weighted age) than those in less dense regions.

These two formation mechanisms predict different age distributions of elliptical galaxies, particularly as a function of galactic environment. We would like to infer the star formation history of ellipticals from their integrated spectra, in particular to distinguish between the hypotheses that the stars formed in a single event at high redshift, $z > 2 - 3$ or that star formation was much more extended with perhaps some star formation occurring at $z \ll 1$. Such tests are under way. It has been suggested that the two flavors of elliptical galaxy, luminous-boxy and faint-disky, may have different evolutionary histories with luminous-boxy galaxies formed predominantly through a dissipationless merging sequence that produce triaxial anisotropic galaxies. The luminosity weighted ages of the two flavors of ellipticals have yet to be determined.

The powerful telescopes at our disposal now, in particular the HST, have enabled us to study the evolution of elliptical galaxies directly. We have known for more than two decades that the fraction of blue galaxies in rich clusters at high redshift ($z = 0.5$) is much higher than

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

we find in local clusters. HST has revealed this blue excess as disk galaxies and has established that at $z = 0.5$ elliptical galaxies have already assembled. It seems very likely that these massive cluster ellipticals did form rapidly and early at $z > 3$ and they ceased star formation before $z = 1$. Perhaps ellipticals in lower-density environments will be found to have experienced a more extended and recent series of mergers, giving them younger mean ages. Such a formation scenario takes ingredients from both the monolithic collapse and merger hypotheses for the origin of ellipticals but already faces apparent contradictions with the empirical measurements.

Bibliography

An excellent general text on the Milky Way and other galaxies is

Binney J J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)

In particular chapters 4, 5 and 11 cover material on the morphology, stellar population and kinematics of elliptical galaxies.

The dynamics of elliptical and bulges is reviewed comprehensively in

de Zeeuw P T and Franx M 1992 *Ann. Rev. Astron. Astrophys.* **29** 239

An excellent overview of the searches for black holes in galactic nuclei can be found in

Kormendy J and Richstone D 1995 *Ann. Rev. Astron. Astrophys.* **33** 581

A more advanced treatment of galaxy dynamics which includes many of the topics introduced here can be found in

Binney J J and Tremaine S 1998 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)

Roger L Davies

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elongation

In general use, the angular separation between the Sun and a planet (or other body orbiting the Sun) or the Moon, i.e. the angle Sun–Earth–object. A more accurate definition, which takes account of the fact that the orbits of bodies orbiting the Sun are inclined to the plane of the ecliptic (the Earth's orbital plane), is that the elongation of a body is the difference between its celestial longitude and the Sun's. Elongation is measured in degrees east or west of the Sun. If a body's elongation is 0° it is at conjunction, if 90° it is at quadrature, and if 180° it is at opposition. The term is also used for the angular separation between a planet and one of its satellites, i.e. the angle planet–Earth–satellite, measured east or west of the planet.

The elongations of the inferior planets, Mercury and Venus, are restricted between certain limits. These maximum angular separations from the Sun are termed *greatest elongation east*, when the planet sets at the latest time after the Sun at a particular apparition, and *greatest elongation west*, when it rises at the earliest time before the Sun at a particular apparition. In Mercury's case, greatest elongation varies between $17^\circ\ 50'$ when the planet is at perihelion (closest physical distance to the Sun) and $27^\circ\ 50'$ when it is at aphelion (farthest from the Sun). For Venus, which has a much less eccentric orbit, greatest elongation varies between 45° and 47°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elsasser, Walter M (1904–91)

German-American physicist, suggested that the Earth's rotation sets up eddy currents in the liquid core which generate the Earth's observed magnetic field.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Emission Spectrum

A spectrum that consists of a number of bright emission lines, each line corresponding to the emission of light at a certain particular wavelength. An atom or ion will emit a photon of a particular energy (and hence of a particular wavelength) when one of its constituent electrons spontaneously drops down from one of its higher ('excited') permitted energy levels to a lower energy level (i.e. when the electron makes a 'downward transition'). The various permitted transitions correspond to a series of lines of different wavelengths, the wavelength of each line being inversely proportional to the magnitude of the energy gap between the relevant energy levels. In addition to undergoing electronic transitions (like atoms and ions) molecules may also emit radiation at characteristic wavelengths through downward transitions in their vibrational and rotational states.

In order to emit radiation through downward transitions, atoms, ions and molecules must first be excited to higher energy levels. This may occur through radiative excitation (absorbing incoming photons) or through energetic collisions between atoms, ions or molecules. Emission will also occur when an ion captures an electron (this process is usually called 'recombination'). Usually, an electron will be captured into an excited level of the atom or ion and will then drop down to lower levels in a series of steps, each of which corresponds to the emission of light of one particular wavelength. When an electron is captured, the difference between the energy that it possessed before being captured, and its energy that is has when bound to the atom or ion, is radiated away. Depending on the range of energies of the free electrons that are available to be captured, the series of lines in an emission spectrum may end in an emission continuum (emission over a continuous range of wavelengths shorter than that of the shortest-wavelength individual line).

An emission spectrum is characteristic of a high-temperature, low-density gas. Typical sources of emission spectra include fluorescent strip lights, the chromospheres and coronae of stars, luminous nebulae, active galactic nuclei and quasars.

See also: absorption spectrum, atom, chromosphere, continuous spectrum, corona, electromagnetic radiation, element (chemical), energy, ion, ionization, molecule, photon, quantum theory, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Empedocles of Akragas (*c*. 492–*c*. 432 BC)

Lived in Acragas (now Agrigento, Sicily), as the author of a physical-cosmological poem entitled *On Nature*, which developed the ideas of the Miletian philosophers like XENOPHANES who believed that the different forms of matter are manifestations of basic elements in different mixtures. Xenophanes suggested that everything was composed of two elements, water and earth, and Empedocles extended this to four: earth, water, air and fire. He believed that the Sun was a reflection of fire, and is credited as the originator of the concept that light has a finite velocity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Enceladus

A mid-sized icy satellite of Saturn, discovered in 1789 by William Herschel. Its diameter is 500 km and it orbits Saturn at a distance of 238 000 km. The part of Enceladus imaged in the Voyager mission has two distinct types of terrain: areas with a medium density of cratering, and almost crater-free plains with extensive fault features—straight and curved grooves and ridges, such as Daryabar Fossa and Samarkand Sulci. The satellite possesses an albedo of 1.0, the highest known, which suggests that its surface is very fresh. So, not only has Enceladus been geologically active in the past, it may still be so today. Its bulk density is 1120 kg m^{-3}, indicating a largely water-ice composition. The activity is made possible by the formation of what is known as a eutectic melt of ammonia and water which is molten at a temperature as low as 176 K. The energy for the activity could be supplied by tidal heating, produced by the satellite's slightly elliptical orbit and a 2 : 1 orbital resonance with Dione. Enceladus orbits at the center of Saturn's tenuous E ring; meteoric debris from the satellite is believed to be the source of the ring's small particles.

See also: Saturn: satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Encke, Johann Franz (1791–1865)

Astronomer, born in Hamburg, became professor and director of the observatory at Berlin. Discovered a comet in 1805, which he later found to be identical with a comet observed in 1818. *Encke's Comet* has a period of 3.3 years (the shortest of any comet), and Encke predicted its return for 1822 (seen only in the southern hemisphere) and in 1825 (which Encke observed from the Seeberg Observatory near Gotha). Presided at the Berlin Observatory over work on calculation of the orbits of asteroids, and the discovery of Neptune by GALLE with the help of star charts that Encke had edited. In 1838, discovered the gap between the A- and F-ring around the planet Saturn (now known as the *Encke division*).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Energy

A physical quantity that, traditionally, is defined to be the ability to do work, where 'work' is the application of a force over a distance. The energy that a body possesses by virtue of its motion is called kinetic energy, the kinetic energy (KE) of a body of mass m traveling at velocity v being given by: $\mathrm{KE} = 1/2mv^2$. In order to change the velocity of a body, a force must be applied, the resultant change in the kinetic energy of the body being equal to the work done by the force. The energy that a body possesses by virtue of its location is called potential energy. When a body of mass m is raised to a height h the potential energy (PE) that it gains is given by: $\mathrm{PE} = mgh$, where g is the acceleration due to gravity (this particular expression assumes a constant gravitational field; in practice, the gravitational attraction exerted on the body decreases with increasing height). Potential and kinetic energy are interchangeable. If the body is released from height h, it will accelerate towards the ground, so converting its potential energy into kinetic energy.

Energy exists in a wide variety of forms. For example, heat is essentially the kinetic energy contained in the motions of atoms, molecules, or ions. Electromagnetic radiation is a form of energy, the energy of a photon (a 'particle' of electromagnetic radiation) being inversely proportional to the wavelength of the radiation. According to the special theory of relativity, energy (E) and mass (m) are equivalent and interchangeable, the relationship between the two being given by $E = mc^2$, where c denotes the speed of light. Because c is a large number, a small mass is equivalent to a very large quantity of energy. The conversion of mass to energy through nuclear reactions powers the stars. The law of conservation of mass-energy states that the total amount of mass-energy in an isolated system is constant (one form can change into the other, but the total sum of mass and energy remains the same).

The SI unit of energy is the joule (J). A unit that is commonly used in particle physics is the electronvolt (eV).

See also: electromagnetic radiation, electronvolt, kinetic theory of gases, mass, photon, special relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Environmental Threat to Astronomy

Astronomy is fortunate in being one of those sciences currently making important advances in understanding, well supported by superb, advanced, technology. Revolutionary new data are being acquired which are radically impacting upon our cherished concepts from the solar system to the depths of the universe. Sadly, the state of astronomy at the dawn of the twenty-first century is not one of unbounded optimism, for the very technology which has produced the current, unparalleled, scientific advances is also capable of inflicting catastrophic damage on observational astronomy—the lifeblood of astronomical science.

Observational astronomy is concerned with detecting extremely weak electromagnetic radiation emitted by cosmic objects. The limiting strength that can be observed is determined by the strength of the background noise against which the radiation is detected. Natural backgrounds are weak (except in the case of the infrared) and therefore small enhancements of natural background levels will significantly raise the level of detection of the faintest objects. Large backgrounds will be disastrous for all serious astronomical research. The activities of mankind produce large increases in astronomical backgrounds in many different ways.

Electromagnetic pollution

Electromagnetic pollution at radio wavelengths

The average mobile telephone, if at the distance of the Moon, has sufficient power output to be among the brightest radio astronomical objects. RADIO TELESCOPES suffer a cascade of radio signals generated on the ground and in space. These signals come from as simple a thing as flicking an on/off switch, through electrical machinery in its myriad forms to telecommunications transmissions and signals from position-fixing satellites. Many of these signals are weak, but any form of communications transmission is powerful (multimegawatts). Because of the nature of radio telescopes, they can pick up signals originating from directions other than that at which the telescope is pointed because of side lobe responses. Such side lobe responses can swamp the observed cosmic signal. Radio observatories can be partially protected from transmissions generated terrestrially by retreat to remote valleys well protected by mountains, but there is no protection at all from aircraft transmissions or transmissions from orbiting satellites. The exploitation of space for telecommunications purposes is growing explosively with the launch of satellite clusters such as Iridium. Despite the careful regulation of radio-frequency space by the International Telecommunications Union, residual noise at significant levels enters the bands reserved for RADIO ASTRONOMY and other passive services (not-transmitting) from neighbouring bands used commercially, through insufficiently attenuated signal tails or signal harmonic frequencies or inadvertent (and sometimes deliberate) tuning through or into the reserved radio astronomy bands. Despite ingenious procedures developed by radio astronomers, such noise frequently degrades, and often swamps, the cosmic signals being observed

Figure 1. Light pollution at Stonehenge. Light pollution looking north-east at Stonehenge (UK), 2 October 1996. Stonehenge is in a rural area! (Copyright D McNally.)

Electromagnetic pollution at optical wavelengths

Optical wavelength electromagnetic pollution is particularly simple to demonstrate. It is clearly visible (see figure 1). The growth of outdoor lighting installations over the last 50 years has produced a high level of background illumination in cities, towns and urban areas which spills a considerably way into rural areas. The light spill from one urban area may overlap the light spill from neighboring urban areas. In the UK there are few areas left where the background is the naturally occurring night sky glow—the essential limiting factor for astronomical observation. The pollution results from the fact that most luminaires now in use have significant upward light spill. It is possible for modern luminaires to have no upward light spill and have sufficient sideways cut-off to put the light emitted where it is needed. The requirement to cut energy consumption to reduce release of CO_2 into the atmosphere is forcing the pace of outdoor lighting design. It is now possible to light motorways and roads, outdoor playing fields, items of architectural note etc in such a way that the luminaires emit no lights upwards and the illumination is confined laterally to the area required to be illuminated—the light spill from these high-grade schemes is confined to the light reflected from the surface of the ground and such light scattered by atmospheric particulates (see later). Unhappily such high-quality designs are the exception rather than the norm.

Because outdoor lighting is so necessary to twenty-first-century lifestyles, outdoor lighting is globally distributed. This means that not only are observatories established more than 50 years ago engulfed in overly lit urban areas, but even the remoter prime observatory sites are seeing growth in local outdoor lighting. As telescope diameters increase and detectors become more sensitive, so more observation is directed towards

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

objects who faintness approaches the natural background. Astronomers have argued that outdoor lighting should not increases the background by more than 10% of the natural background. Even that excess is not acceptable to those studying contributions to the natural sky glow. Most urban backgrounds exceed the natural background by at least a factor 100 and in the centres of major cities by factors of in excess of a million. This has a serious effect on optical observational astronomy.

Astronomers are also concerned about the spectral character of the lamps used in outdoor lighting. The low-pressure sodium lamp emits virtually all its output over a very narrow wavelength range—it is almost monochromatic. Clearly this lamp is preferred by the astronomical community. It is a highly energy efficient lamp. It is not universally liked as it gives poor rendering of color. However, addition of small amounts of white light at locations where color rendering is important can be very effective. Nevertheless there is pressure to use multispectral lamps (e.g. high-pressure sodium, halide) and the recent trend towards such lamps will progressively prevent astronomical use of some spectral ranges—whereas now only a very short spectral range occupied by low-pressure sodium lamps is unusable.

While OPTICAL ASTRONOMY, unlike radio astronomy, has trends working to its benefit, there is no escaping the fact that the background at optical wavelengths is increasing as are projects near prime observatory sites which demand high-intensity outdoor lighting (e.g. a new prison project is planned on Hawaii near the prime observatory sites on Mauna Kea which houses, *inter alia*, the world's two largest telescopes).

Exploitation of space

Exploitation of space has brought additional degradation of astronomical observing conditions over and above the effect on radio astronomy mentioned above. The use of space has meant the accumulation of space debris. There are now over 8200 pieces of trackable debris (size over 10 cm in low Earth orbit) in space and substantially greater amounts of smaller debris which can cause significant damage to operational spacecraft. Debris is now posing a threat to all operational satellites including orbiting astronomical observatories like the HUBBLE SPACE TELESCOPE.

The number of operational and spent satellites and large pieces of debris is so great that many astronomical deep sky images now carry trails caused by sunlight reflected by this subset of space debris. Fifty per cent of deep sky images carry at least one trail and 25% carry multiple trails. Such trailing also affects the very much smaller sized photometric fields of view. Since most of the trailing occurs within the 2 h after sunset, or before sunrise, the astronomical community lives with the image degradation or curtailment of the efficiency of operation that trailing causes.

Of more concern are proposals to place solar reflectors in space to produce the appearance of objects up to the brightness of 100 times that of the full Moon. The Moon is sufficiently bright to suspend astronomical observation of the faintest celestial objects for two weeks out of every four. Currently tests of the Znamya2 and Znamya2.5 solar reflectors have been attempted, but the tests could not be fully implemented. Were this technology to be proved, the astronomical nightmare is, that despite being only operational in twilight, such reflectors might be used for commercial or other advertising. Thus far, proposals for advertising reflectors have been turned down by the space agencies. But one is prompted to ask, should advertising from space become acceptable, would not advertisers then want systems in space which could exploit dark time (indeed there has been a suggestion for advertising holograms to be generated in space in dark time)? Were that to become the case, optical astronomy would die.

There is one proposal for the use of space reflectors which is of great concern—namely to use a system of such reflectors to collect enough solar energy to generate up to 50% of the world's demand for electricity by 2050. Solar radiation would be converted in space to microwaves for transmission to Earth where the microwave energy would be converted to electricity. The efficiency of the process is said to be less than 10%. What happens to the remaining 90% lost in the process? It is admitted that such a proposal would have serious consequences for ground-based optical astronomy and would be disastrous for ground-based radio astronomy. Such extension of the exploitation of space is of very great concern to the astronomical community.

Others areas of concern

Atmospheric extinction. The atmosphere contains naturally occurring particulates and aerosols augmented by those generated by the activities of mankind through dumping of industrial effluent into the atmosphere, emissions from vehicles, creation of dust etc. As pointed out in earlier, atmospheric particulates increase the scattering of outdoor lighting. Atmospheric extinction also decreases the energy received from cosmic sources as well as reddening that light. Increase in extinction is largely an urban phenomenon and relatively local except in the case of aircraft contrails. Generation of extinction in the northern hemisphere seems not to be spreading in the southern hemisphere, as yet. This is a problem requiring vigilance and study in case a serious problem is developing—particularly in regard to the consequences of aircraft contrails.

Accumulation of heavy molecules in the stratosphere. Infrared observers have found that the accumulation of heavy molecules in the stratosphere is increasing the extinction in the limited atmospheric windows available to ground-based infrared observatories. While some of these molecules are now banned (to prevent further damage to the ozone layer) it is not yet clear if replacement molecules will lead to any improvement for INFRARED ASTRONOMY. This is again an area for sustained vigilance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Global warming and atmospheric heat dumping. It is agreed that the Earth's atmosphere is in a warming phase (see GREENHOUSE EFFECT). A consequence of rising temperatures is enhanced cloudiness—a trend reinforced by contrail formation in the absence of rapid dispersal (see earlier). Global warming may lead to changes in weather patterns, one consequence of which could be climatic degradation at prime observatory sites. As a corollary, in urban areas there is atmospheric dumping of heat from industrial processes down to heat loss from buildings. One clear consequence of such local and largely urban, heating is enhanced atmospheric turbulence and consequent degrading of astronomical SEEING.

Ground vibration. Industrial operations and other activities of mankind produce vibration. Vibration is a local phenomenon, but low-amplitude vibrational effects can be felt at distances of up to 10 km from source. This can be a serious problem for astrometric observatories in particular. Observatories free from vibration in the past may become susceptible because of new industrial development in the local neighborhood.

Other concerns. The above listing of areas of concern is not exhaustive. New concerns will appear and old concerns develop unforeseen and more serious consequences. There are still concerns regarding space activities that could result in terrestrial pollution of the nearby solar system bodies—which may affect the integrity of future exploration of those bodies. There is concern about the consequences of HAPS (High Altitude Platform Systems) on both radio and optical astronomy. There is concern in optical astronomy about the consequences of using optical/IR laser beams (e.g. ESA's SPOT4) for satellite communications. There will no doubt be many other concerns downstream. Astronomy is in consequence a highly beleagured science.

Protection measures

The science of astronomy has taken steps to protect itself. Despite many well equipped observatories on urban sites, the trend has always been for observatory sites to retreat ever further from centers of population (see OBSERVATORIES AND THEIR LOCATION). However, there is now no scope for further retreat except possibly through greater use of Antarctica. Space is not the user friendly, if costly, place it once seemed. Astronomers have therefore entered into local agreements to give special protection to prime observatory sites. So far these agreements have worked well, but it must be remembered that those living close to such sites may find these agreements overly restrictive in the future. Radio astronomers have been in detailed negotiation, nationally and internationally (through IUCAF, the InterUnion Committee on Allocation of Frequencies), with the International Telecommunications Union on allocation and defence of the radio astronomy bandwidth allocation. This has been extremely successful but is becoming a highly demanding business given the hunger commercially for bandwidth. These activities deserve the strongest community backing.

Astronomers have taken their concerns to the public. Light pollution is a topic which the public can readily appreciate and indeed emphathize with. Many people resent that they can no longer see the spectacle of a dark night sky. But it is hard to convey the dire situation faced by radio astronomy, though opposition to space reflectors seems to command global support.

In July 1999, the astronomical community had an opportunity to present their concerns to the UN at a symposium at UNISPACE III. The UN have taken on board for consideration astronomical concerns above the impact of exploitation of space and the problem that telecommunications pose for radio astronomy. Such consideration must of its nature be long term. One consequence will be that the astronomical community will have to commit significant resources to quantitative assessment of its observing environment. This will be essential in obtaining any future international agreement to protect astronomical science.

The astronomical community will also have to give greater attention to public information. Considerable effort is now expended and has demonstrably paid off. We still need to reach a wider audience since there is wide public interest in, and sympathy for, astronomical observation but it is coupled with widespread ignorance of how the activities of mankind impact upon astronomy. The outlook may well be bleak but it is not without considerable hope.

Bibliography

A wide ranging readable account of the basic situation can be found in

McNally D (ed) 1994 *The Vanishing Universe* (Cambridge: Cambridge University Press)

McNally D 1996 Adverse environmental impacts on astronomy, *Q. J. R. Astron. Soc.* **37** 129

Three IAU conferences have been on this topic:

Crawford D L (ed) 1991 *Light Pollution, Radio Interference and Space debris* (IAU Colloquium 112)

Isobe S and Hirayama T (ed) 1998 *Preserving the Astronomical Windows* (*IAU Joint Discussion Kyoto General Assembly*) (Publ. Astron. Soc. Pacific)

Crawford D L, Cohen J and Sullivan W T (ed) 2000 *Preserving the Astronomical Sky* (IAU Symposium 196) (Publ. Astron. Soc. Pacific) at press

Other resources:

Committee on Radio Frequencies 1996 *Handbook in Radio Astronomy*

Flury W (ed) *ESA Conference Reports on Space Debris Assessment*

Kovalevsky J (ed) 1992 *The Protection of Astronomical and Geographical Sites* (Gif-sw-Yvette: Editions Frontieres)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Research Council 1995 *Orbital Debris: a Technical Assessment* (Washington, DC: National Acadamy Press)

Simpson J A (ed) 1994 *Preservation of New Earth Space for Future Generations* (Cambridge: Cambridge University Press)

D McNally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ephemerides

In everyday language, 'ephemeris' indicates what occurs daily; the ephemeris of the day is the list of the remarkable events of this day. By extension, the astronomical ephemerides are tabular listings of the positions of a celestial body, allowing one to know its position for a given time. Today, ephemerides may have a new form in order to improve the accuracy of the positions deduced from them. Ephemerides provide representations of the motion of celestial bodies.

The ephemerides of positions are thus the representation of a motion. The most common ephemerides are those in the form of tables of numbers. But it is not the only possible form and, nowadays, it is no longer the best form because there are much more powerful ways to describe a motion. The quality of an ephemeris for the representation of a motion depends on two factors: the use of a small number of data (to avoid too large tables) and the possibility of good precision (to give the weakest error compared to the 'true' position).

Before discussing the various types of representation used for the ephemerides, let us see first of all how the motions of celestial bodies belonging to the solar system are described. One will see hereafter that the ephemeris itself depends on the nature of the motion which is to be described.

Planetary theories and ephemerides

History

Ephemerides have been published since the fourteenth century. The first ephemerides intended for use in planetary dynamics appeared in the French *Connaissance des Temps*, first volume of which was published in Paris in 1679. This first volume provided a tool for reducing the observations of solar system bodies. It contained rise and set times for the Sun, Moon, Mercury, Venus, Mars, Jupiter and Saturn, their transits to the Paris meridian, their longitudes, their latitudes and their declinations. Later, Lalande compiled the tables of the Sun and the planets which were used as a basis for the ephemerides published in the *Connaissance des Temps* until 1808. In 1767, an ephemeris of the same kind designed for the improvement of Astronomy, Geography and Navigation was published in London: the *Nautical Almanac and Astronomical Ephemeris*.

During the years 1770–1825, Laplace studied the consequences of the application of Newton's law of gravitation on the motion of the bodies of the solar system and solved several problems concerning planetary theories. The representation of these motions in the ephemerides evolved continuously, but until the end of the eighteenth century, the *Connaissance des Temps* presented the ephemerides month by month, the Sun day by day and the planets six days by six days. The time interval evolved (decreased) in order to improve the accuracy of the positions, day by day or more, depending on the motion of the body. From 1838, the ephemerides were published body by body. They were daily for the Sun, three days by three days for Mercury, six days by six days for Venus and Mars, eight days by eight days for Jupiter, ten days by ten days for Saturn and fifteen days by fifteen days for Uranus. For the Moon, a smaller interval was necessary (cf table 1). In 1863, the ephemerides were provided daily for Mercury, Venus, Mars, Jupiter and Saturn and four days by four days for Uranus and Neptune.

Table 1.

(185) Année 1808.

JOURS.	ASCENSION DR. ☾		DÉCLINAISON DE LA LUNE.			
	A MIDI.	A MINUIT.	A MIDI.	A 6 HEUR.	A 12 HEUR.	A 18 HEUR.
	D. M.	D. M.	D. M.	D. M.	D. M.	D. M.
1	48. 59	55. 9	17. 17. B	17. 47. B	18. 14. B	18. 37. B
2	61. 22	67. 40	18. 58.	19. 15.	19. 29.	19. 40.
3	74. 1	80. 25	19. 48.	19. 52.	19. 53.	19. 50.
4	86. 50	93. 15	19. 45.	19. 35.	19. 22.	19. 6.
5	99. 41	106. 6	18. 47.	18. 24.	17. 57.	17. 28.
6	112. 28	118. 50	16. 55.	16. 19.	15. 41.	14. 59.
7	125. 8	131. 25	14. 15.	13. 28.	12. 38.	11. 46.
8	139. 41	143. 56	10. 52.	9. 56.	8. 57.	7. 57.
9	150. 11	156. 26	6. 55.	5. 51.	4. 46.	3. 40. B
10	162. 43	169. 5	2. 32. B	1. 24. B	0. 15. B	0. 54. A
11	175. 31	182. 4	2. 4. A	3. 14. A	4. 24. A	5. 33.
12	188. 44	195. 33	6. 42.	7. 49.	8. 56.	10. 1.
13	202. 33	209. 43	11. 4.	12. 5.	13. 4.	13. 59.
14	217. 5	224. 38	14. 52.	15. 42.	16. 28.	17. 10.
15	232. 20	240. 11	17. 48.	18. 21.	18. 49.	19. 12.
16	248. 7	256. 7	19. 30.	19. 43.	19. 51.	19. 53.
17	264. 5	272. 0	19. 50.	19. 42.	19. 28.	19. 10.
18	279. 48	287. 26	18. 47.	18. 19.	17. 47.	17. 11.
19	294. 53	302. 6	16. 31.	15. 47.	15. 1.	14. 11.
20	309. 7	315. 55	13. 18.	12. 23.	11. 27.	10. 28.
21	322. 31	328. 55	9. 28.	8. 27.	7. 25.	6. 22.
22	335. 9	341. 15	5. 18.	4. 13. A	3. 9. A	2. 5. A
23	347. 14	353. 7	1. 0. A	0. 3. B	1. 7. B	2. 10. B
24	358. 56	4. 43	3. 12. B	4. 13.	5. 14.	6. 13.
25	10. 28	16. 14	7. 11.	8. 8.	9. 4.	9. 58.
26	22. 1	27. 50	10. 50.	11. 40.	12. 29.	13. 16.
27	33. 43	39. 39	14. 0.	14. 43.	15. 23.	16. 1.
28	45. 40	51. 46	16. 36.	17. 9.	17. 39.	18. 6.
29	57. 57	64. 12	18. 31.	18. 52.	19. 11.	19. 26.
30	70. 32	76. 56	19. 38.	19. 46.	19. 51.	19. 53.
31	83. 23	90. 51	19. 51.	19. 46.	19. 37.	19. 25.

N

The development of celestial mechanics

The nineteenth century was very rich in progress in celestial mechanics and the motion of the solar system's bodies became well modeled (see CELESTIAL MECHANICS IN THE EIGHTEENTH AND NINETEENTH CENTURIES): it is even thanks to the defects noted on the ephemerides of Uranus that Neptune was discovered through calculations by Le Verrier and Adams. These defects did not result from a bad evaluation of the perturbations from Jupiter and Saturn: it was necessary to admit the existence of an additional body beyond Uranus. After the discovery of Neptune by Galle in Berlin, observations were realized and the elements of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2.

LUNE 1978. 81

ASCENSION DROITE, DÉCLINAISON, PARALLAXE, DEMI-DIAMÈTRE.

TEMPS DES ÉPHÉMÉR.	ASCENSION DROITE	VAR. pour 1m	DÉCLINAISON	VAR. pour 1m	PARALLAXE	DEMI-DIAMÈTRE
Janv. h	h m s	s	° ′ ″	″	′ ″	′ ″
0 0	10 41 1,537	1,99046	+ 5 33 33,47	− 9,4788	55 37,29	15 9,33
3	10 46 59,975	1,99226	5 4 58,69	9,5733	55 41,75	15 10,55
6	10 52 58,772	1,99445	4 36 7,38	9,6624	55 46,31	15 11,79
9	10 58 57,999	1,99705	4 7 0,50	9,7461	55 50,99	15 13,07
12	11 4 57,729	2,00004	3 37 39,06	9,8243	55 55,77	15 14,37
15	11 10 58,037	2,00346	3 8 4,03	9,8969	56 0,65	15 15,70
18	11 16 58,999	2,00731	2 38 16,46	9,9638	56 5,64	15 17,06
21	11 23 0,692	2,01160	2 8 17,37	10,0249	56 10,73	15 18,45
1 0	11 29 3,196	2,01632	1 38 7,82	10,0800	56 15,93	15 19,86
3	11 35 6,593	2,02150	1 7 48,89	10,1291	56 21,23	15 21,31
6	11 41 10,963	2,02715	0 37 21,68	10,1720	56 26,63	15 22,78
9	11 47 16,390	2,03325	+ 0 6 47,30	10,2087	56 32,12	15 24,27
12	11 53 22,959	2,03983	− 0 23 53,08	10,2388	56 37,71	15 25,80
15	11 59 30,754	2,04688	0 54 38,32	10,2624	56 43,40	15 27,35
18	12 5 39,862	2,05441	1 25 27,20	10,2793	56 49,18	15 28,92
21	12 11 50,368	2,06243	1 56 18,50	10,2893	56 55,04	15 30,52
2 0	12 18 2,360	2,07093	2 27 10,97	10,2922	57 0,99	15 32,14
3	12 24 15,925	2,07990	2 58 3,31	10,2878	57 7,03	15 33,78
6	12 30 31,150	2,08938	3 28 54,21	10,2762	57 13,14	15 35,45
9	12 36 48,123	2,09932	3 59 42,32	10,2568	57 19,33	15 37,13
12	12 43 6,930	2,10975	4 30 26,24	10,2297	57 25,58	15 38,84
15	12 49 27,659	2,12066	5 1 4,56	10,1945	57 31,91	15 40,56
18	12 55 50,393	2,13204	5 31 35,81	10,1511	57 38,30	15 42,30
21	13 2 15,219	2,14389	6 1 58,50	10,0993	57 44,74	15 44,06

Table 3.

LUNE 2000 DT = 4 jours

ASCENSION DROITE ET DÉCLINAISON APPARENTES, DISTANCE À LA TERRE
Équinoxe et équateur vrais de la date.

	du 0 janvier 0h au 4 janvier 0h				du 4 janvier 0h au 8 janvier 0h			
	asc. droite	déclinaison	distance		asc. droite	déclinaison	distance	
	13.675691 41	*−4.883403 0*	*397 117.928*		*16.810860 50*	*−18.326871 2*	*406 313.750*	
0	15.230730 28	−12.144665 6	402 707.316	0	18.486695 79	−19.955278 2	405 017.941	0
1	1.565330 38	−6.763099 8	4 580.555	1	1.686828 82	−0.724416 6	−1 947.259	1
2	0.012841 92	0.539832 4	−995.952	2	0.008533 47	0.919814 6	−622.419	2
3	0.002266 53	0.041387 4	17.794	3	−0.002630 90	0.012138 2	26.796	3
4	−0.000296 47	−0.000282 0	4.477	4	−0.000133 99	−0.003849 0	−2.405	4
5	−0.000012 49	−0.000021 8	−0.439	5	0.000038 99	−0.000141 5	−0.147	5
6	0.000000 20	−0.000002 1	−0.002	6	0.000001 84	0.000023 3	0.024	6
7	0.000000 12	0.000000 0	0.001	7	−0.000000 33	0.000001 6	0.001	7
8	0.000000 02	0.000000 1	0.000	8	−0.000000 03	−0.000000 2	0.000	8

	du 8 janvier 0h au 12 janvier 0h				du 12 janvier 0h au 16 janvier 0h			
	asc. droite	déclinaison	distance		asc. droite	déclinaison	distance	
	20.179333 68	*−19.751707 8*	*402 472.530*		*−0.518049 87*	*−7.513 7252*	*389 872.633*	
0	21.841653 19	−14.334297 8	396 691.308	0	1.139704 31	1.685 9847	380 458.908	0
1	1.650634 52	6.161313 4	−5 298.789	1	1.678602 71	9.299 8571	−9 665.124	1
2	−0.011391 30	0.703505 5	−519.483	2	0.025362 77	0.029 4605	−182.489	2
3	0.000677 73	−0.042473 8	−1.577	3	0.004572 62	−0.073 8086	76.686	3
4	0.000382 10	−0.001909 2	0.744	4	0.000026 64	−0.003 3364	7.860	4
5	−0.000004 22	0.000153 3	0.419	5	−0.000034 36	−0.000 2094	0.035	5
6	−0.000002 09	−0.000015 2	0.013	6	−0.000002 91	0.000 0019	−0.052	6
7	0.000000 19	−0.000001 6	−0.001	7	−0.000000 27	0.000 0041	−0.003	7
8	0.000000 00	0.000000 2	0.000	8	0.000000 00	0.000 0003	0.000	8

the motion of Neptune were then known enough so that precise ephemerides could be published.

After such an achievement of celestial mechanics, Le Verrier undertook to completely rebuild the theories of the motion of the main planets. He derived the theories and the tables of these bodies. Gaillot, a collaborator of Le Verrier, then improved the theories of the four large planets, Jupiter, Saturn, Uranus and Neptune. At the same time, the theoretical work of Hansen gave rise to the Russian astronomer Kowalski's theory of Neptune and, especially, to Newcomb's global theory of planets.

The tables of Le Verrier and Gaillot were used in the ephemerides of the *Connaissance des Temps* until 1984 but the tables of Kowalski and those of Newcomb were the sources of the ephemerides of Neptune published in the *Connaissance des Temps* from 1863 to 1883. For Uranus, the tables of Newcomb were used from 1877 to 1883. *The Nautical Almanac* also used Le Verrier's tables; the American Ephemeris used Hansen's work and Newcomb's tables for the Sun and the planets, the tables of Hansen and Brown for the Moon, until 1983. In 1960, the British and the American ephemerides published the same tables, the name of which became *The Astronomical Almanac* in 1981.

At the time of Le Verrier and Gaillot, the ephemerides published in the *Connaissance des Temps* acquired a form which was used continuously until 1979 (cf table 2). Note that in 1877 the rectangular coordinates of the Sun were introduced and the ephemerides of Uranus and Neptune were provided daily in the *Connaissance des Temps*.

Ephemerides nowadays

Significant transformations have been made to the ephemerides in the last few years. In 1980, the tabular presentation of the ephemerides of the *Connaissance des Temps* was replaced by a representation in Chebyshev polynomials (see later); much more compact and well adapted to the development of microcomputers (cf table 3). Tables 1, 2 and 3 show the progress in the representation of the ephemerides of the Moon.

Since 1984, the *Connaissance des Temps* has used new analytical theories of the motions of the Sun, the Moon and the planets, elaborated at the Bureau des Longitudes and adjusted on the numerical integration of the Jet Propulsion Laboratory (JPL) DE200. In 1984, the *Astronomical Almanac* changed the sources of its ephemerides and used the numerical integration DE200/LE200 of the JPL. JPL is now using an improved ephemeris DE405/LE405. Theories and numerical integration will be described further.

Theories of the motions

Since the beginnings of astronomy, modeling the motion of the bodies of the solar system has been a challenge. First, these motions were empirical extrapolations of the observations carried out; the first tables thus came from a purely kinematic analysis of the observed motions. The precision of these first tables was obviously poor and could be increased only with the improvement of the precision of the observations themselves.

Then came calculations based on gravitational theories in which the parameters were deduced from observations. Since Newton, the dynamic laws were known and it was then important to include these in the equation and to take into account all the gravitational effects which have some influence on the bodies. Lagrange's theoretical search relative to the planetary problem resulted in modeling the long-term evolution of the orbits by a linear differential system which coupled the eccentricities and the inclinations. This result is fundamental.

Ephemerides may be made from theoretical equations by solving the equations into trigonometric series, which are time consuming to calculate without computers. Another way is the direct calculation through numerical integration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elements of celestial mechanics

General remarks on planetary theories

The motion of planets around the Sun is a particular case of the *N*-BODY PROBLEM which has no exact solution for *N* larger than 2. All the bodies attract each other in accordance with the law of gravitation but it is considered that the planets have a small mass relative to the one of the central body (the Sun). Approximate solutions of the problem have to be made thanks to the theory of perturbations, where the coordinates are functions of the time t, of the masses of the involved bodies and of the constants of integration. One obtains these solutions by deriving analytical theories or by carrying out numerical integrations.

Analytical theories

The coordinates are obtained in the form of combinations of analytical algebraic and trigonometrical functions of the time t and the parameters of the problem, masses and constants of integration. The calculation of a position using these theories is long but simple since it is sufficient to substitute the time parameter in the series.

Numerical integrations

Numerical integrations give the numerical values of the coordinates and velocities for some values t_0, $t_0 + h$, $t_0 + 2h$, etc, t_0 being the initial time and h being the step of integration. The methods of numerical integration are well adapted to calculations by computer and they were especially used at the Jet Propulsion Laboratory in the USA.

The time argument of the ephemerides

It is advisable to now say some words on the 'time' argument of the ephemerides. Indeed, to know a position at a given moment, which moment to seek in the ephemerides? Until 1834, one used the real solar time of Paris in the *Connaissance des Temps*. Then, because of the existence of more reliable clocks, one used the mean time of Paris. In 1916, following an International Convention, one used the mean time of Greenwich. These time scales were all related to the rotation of the Earth known as uniform at that time. When irregularities were put into evidence in this rotation, astronomers introduced a new uniform time scale for use in the ephemerides, based on the rotation of the Earth around the Sun. Ephemeris Time compiled from Newcomb's theory of the Sun, or, as it was in the *Connaissance des Temps*, Temps Uniforme compiled from Le Verrier's theory of the Sun, were used. From 1984, Terrestrial Time, a uniform time scale derived from atomic clocks which were much more stable than celestial motions (see TIME SCALES: ATOMIC), was introduced.

Reference systems

Knowledge of the positions and the motions of celestial bodies requires a precise notion of the reference systems of space and temporal coordinates where one will locate the positions described by the ephemerides. We saw previously that the time argument of the ephemerides could not be taken in any time scale. In the same way, we will see that the spatial coordinates to be used must be attached to a reference system through a reference frame. One can distinguish the celestial and the terrestrial reference systems.

The celestial reference systems are either dynamical (based on the dynamical study of the motions of solar system's bodies) or kinematical (based on the statistical study of the motions of far bodies such as stars, galaxies and quasars). The dynamical reference celestial systems, first built in Newtonian mechanics, are built today in relativistic mechanics which ties the time argument to the spatial variables (see REFERENCE FRAMES AND TIME SCALES IN GENERAL RELATIVITY). For the needs of the current ephemerides, the definitions given previously for the time argument are sufficient. To link the positions of the celestial bodies to an observer located on the Earth, one will need also to define a terrestrial reference system.

The terrestrial reference system requires the knowledge of a model of Earth, its form and its deformations depending on time. The determination of a topocentric position requires such a system to know the position of the observer relative to positions of celestial bodies that are commonly determined in a geocentric celestial reference system.

The celestial reference frames

A reference system is a theoretical concept and the construction of ephemerides requires the knowledge of a reference frame which is the concrete materialization of a reference system. A celestial reference frame will thus be associated with dynamical theories, constants and star catalogs.

A reference frame is defined by a plane of reference and a given direction. For the ephemerides currently published, the plane of reference is the equator of the Earth (equatorial reference frame) or the ecliptic plane of the orbit of the Earth around the Sun (ecliptic reference frame). The direction of reference is that of the equinox, ascending node of the ecliptic on the equator, intersection of the two planes. These planes and this direction are related to the Earth and we will define various reference frames depending on the theories and the constants used to build them. Then, for each reference frame thus built, we will fix the planes and the direction of reference to obtain a perfectly defined reference frame.

Reference frames related to stars

These reference frames are defined by the positions and the PROPER MOTIONS of stars. The star catalogs provide a value of the proper motions which makes it possible to obtain, by a simple calculation, the position of stars at a given date. However, the Astronomisches Rechen Institut of Heidelberg, Germany, has published each year since 1960, following an international agreement, the apparent positions of the 1535 stars of the FK4 and FK5 catalogs in the *Apparent Places of Fundamental Stars*. New reference frames using stars of the HIPPARCOS catalogue and the very

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

far radiosources are under development to produce on international system of reference.

Dynamical reference frames related to the solar system's bodies
The dynamical models of the bodies of the solar system adjusted on radar or laser observations are independent of the stars of the catalogs and define a plane and a direction of reference. Associated with a theory of PRECESSION–NUTATION, they define a reference frame. Let us quote the reference frame related to numerical integrations DE405/LE405 of the Jet Propulsion Laboratory and the one related to the analytical theories ELP2000/VSOP82 of the Moon and the planets of the Bureau des Longitudes.

Coordinates systems

Once given a reference frame, it is necessary to choose the coordinates to be used in order to provide the positions in the ephemerides. For any reference frame, a theory for the precession–nutation is associated and one defines:

- positions in the mean reference frame (the fundamental plane is a mean plane, i.e. affected only by precession). An epoch must be given to fix the reference frame; this epoch is conventional (for example 1950 for reference frame FK4 and 2000 for the FK5).
- positions in the true reference frame (the fundamental plane is affected by precession and nutation). The epoch of the reference frame is that to which one gives the positions: the reference frame is not fixed, but this reference frame is the only one directly accessible to the observer.

In the reference frames thus defined, the coordinates may be:

- geometrical—they are the positions really occupied by the body at the date given, without taking into account the speed of the light;
- apparent—they are positions as seen by an observer being at the center of the reference frame. They are referred to a true reference frame of the date. The speed of the light and the annual aberration are taken into account.
- astrometrical or astrographic—these coordinates are directly comparable with the star catalogs. They may be deduced from the apparent coordinates by the transformation of precession–nutation to come back to a fixed reference frame (the one of the date of the star catalogs) and by the suppression of the annual aberration. The light time is still taken into account.

The ephemerides of position
Depending on the positions to be represented in the ephemerides, we have to choose a fundamental plane and some coordinates. For the ephemerides for the terrestrial observer, the fundamental plane will be the terrestrial equator and the coordinates to be provided will be the right ascension (angle counted towards the east starting from the equinox, along the terrestrial equator) and the declination (angle counted towards the north above the equatorial plane). The radius vector, the distance from the center of the reference frame to the considered body, may also be provided. This distance does not depend on the reference frame, except by taking into account the light time.

The ephemerides of the Moon, the Sun and main planets give apparent coordinates mainly for the guidance of telescopes. For Pluto, the asteroids and the comets, they provide the astrometrical coordinates directly comparable with the star catalogs. In the case of natural satellites, differential coordinates relative to the planet in a tangential plane with the celestial sphere, directly comparable with the photographic exposures or images from CCD cameras, are provided.

Let us quote the topocentric coordinates, centered on an observer situated on the Earth: these coordinates should be calculated for each observer, so, in order to transform the apparent coordinates into the topocentric ones, a small correction has to be added.

Positions may also be given in non-spherical coordinates such as rectangular coordinates: the x-axis being directed towards the equinox of the reference frame and the z-axis towards the celestial North Pole of this reference frame. Such ephemerides are mainly provided for space navigation.

Ephemerides concerning stars
The proper motions of stars are not sufficiently important to justify the publication of an ephemeris daily. Only the ephemeris of the polar star is provided daily in several ALMANACS, since this allows the setting up of telescopes.

Ephemerides for physical observations
In order to observe from the Earth the surfaces of solar system's bodies, it is necessary to define the parameters which will characterize the rotation of the body around its axis, to give some numerical values and to define a coordinate system in which the points on the surfaces may be located (planetographic or planetocentric reference frames). One then calculates the points of the body which are visible from the Earth at a given date. It is also useful to know quantities related to the illumination of the apparent disc. The definitions given hereafter were ratified by the International Astronomical Union.

Definition of the parameters of the rotation
The North Pole is, by definition, that of the two poles of the axis of rotation which is in the north side of the invariant plane of the solar system (near to the ecliptic). The parameters of the rotation of the solar system's bodies are, for each body:

- the axis of rotation which is given by the right ascension and the declination in the system J2000 of the direction of its North Pole, depending on time;
- the position of the meridian line origin which is given by the angle which separates the ascending node of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

equator of the planet on its orbit from this meridian line origin, angle counted in the equator of the planet and depending on time. A constant at the origin of time will be added if the meridian line origin is linked to a perceptible detail characteristic on the surface of the body.

Definition of the coordinates system
Two different coordinates systems are necessary to locate a point on the surface of a body:

- the planetocentric coordinate system where the planetocentric latitude of a point is the angle made by the vector, joining the center of the body and the point, with the equatorial plane; and where the planetocentric longitude is the dihedral angle between the meridian of the point and the meridian origin, counted positively towards the east on the body, independently of the rotation. This system is used for the study of the motion of a satellite over the body.
- the planetographic coordinate system used to chart the surface of the body. The planetographic latitude of a point of the surface is the angle made by the normal to the surface at this point with the equatorial plane. If the body is assimilated with an ellipsoid of equatorial radius a and of polar radius b ($b < a$), the planetocentric and planetographic latitudes φ and φ' of a point of the ellipsoid are connected by

$$\tan \varphi' = \tan \varphi / (1 - f)^2$$

where f is the flatness of the ellipsoid defined by $f = (a - b)/a$.

The planetographic longitude is defined in a way similar to the planetocentric longitude. It is also counted starting from the meridian line origin but in the direction opposed to rotation. It results from this that, for a distant observer, the planetographic longitude of the center of the apparent disc always increases with time, whatever the direction of rotation of the body. In order to respect a historical tradition and to be able to continue to use old charts, this rule admits exceptions for example for the Earth, the Sun and the Moon whose rotations are direct.

Accuracy of the ephemerides

The accuracy of the ephemerides is the result of three cumulative precisions chosen for the compilation of the ephemerides.

- The internal precision of the dynamical theories used to model the motions depends on the importance of the terms of the series neglected and may be evaluated by comparison to a numerical integration. Note that the numerical integrations have their own internal accuracy depending on the algorithms and the computers used. The internal precision is generally the best.
- The external precision of the theory depends on the accuracy of the observations used for the adjustment of the constants of integration (the external precision is the one limiting the total precision of the ephemerides).
- The precision of the representation is chosen when adjusting the number of coefficients and the length of the intervals of validity (this precision is chosen in order to avoid degrading the resulting accuracy).

Representation of the ephemerides

Ephemerides in tabular form
The oldest form of the ephemerides is the publication in the form of tables of positions, the interval of which depends on the variation of the values of the positions. The faster this variation, the smaller the interval should be between two values. A simple reading of the table and a visual interpolation are sufficient to find the observed object in the sky. However, in order to get a higher precision, a nonlinear interpolation is necessary. This will need more complex calculations using the published data. For a higher precision, a solution is to decrease the length of the time interval between two values and therefore to increase the volume of data to be published. The *Astronomical Ephemerides* and the *Annuaire du Bureau des Longitudes*, nowadays, and the *Connaissance des Temps* until 1979, contain tabulated ephemerides. Let us quote the motion of the Moon: in order to ensure an internal precision of a few units of 0.01 arcsecond, the ephemeris is tabulated with a 3 h interval. For each coordinate, this representation with a step of 3 h requires 2920 values annually (table 3). Instead, with the tabulated ephemerides, 7440 values were necessary to reach the same accuracy, but with more computer calculations (table 2). Let us quote also the case of the Galilean satellites of Jupiter which have an extremely fast motion around Jupiter. A tabulated ephemeris would have required a step every 20 min which would have led to a volume of data too large to be published in a printed ephemeris. The solution was of two kinds. The first way was to provide formulae, complex and long, to be implemented which would provide a precision of one tenth of a second of degree. The second way was to offer the reader a graphical chart, the reading of which led to a precision of one second of degree. This last method is quite weak but is sufficient to identify the bodies at the time of an observation.

The common widespread use of pocket calculators and microcomputers implies that the simplicity of the formulation is no longer fundamental in the compiling of the ephemerides. It is more important to preserve a good precision of representation over the period of time covered by the ephemeris. Thus, instead of evaluating a function $f(t)$ as a table of computed values at short intervals of time, it is better to establish valid functions of approximation of $f(t)$ on large intervals of time, with the concern to minimize the errors and the total volume of the data.

Approximation of a function
The goal of the ephemerides is thus to represent a function $f(t)$, derived either by the result of a numerical integration

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or by an analytical function of the time, which describes the motion of a body or the temporal evolution of a phenomenon.

In the majority of the cases, the Chebyshev series are now used for the realization of the ephemerides. Why? In fact, any function can be approximated on a given interval of time by a polynomial, the degree and the number of terms of which increase as the interval of time increases. Such an approximation is not optimized and the volume of data to be published is significant. On the other hand its use is very simple since it is enough to substitute time in the polynomials thus built. Other representations are much better: let us try to explain why.

Chebyshev approximation

The functions to be approximated are always defined for all the values of the interval. The principle of the Lagrangian approximation is to adjust a polynomial of degree n on $n + 1$ tabulated values. It would obviously be more desirable to take into account all the values of the function. For this reason, the Chebyshev approximation is chosen: the polynomials of Chebyshev have the interesting characteristic of having the error regularly distributed on the considered interval and being stable during evaluation. Moreover, among the various types of polynomial approximations, the approximation of Chebyshev is the one with the lower degree and the one on which the error does not exceed a given value. Another interesting fact is that the last retained terms offer an estimate of accuracy of the polynomial representation.

The approximation function of the ephemerides will thus represent the true function by developments in Chebyshev polynomials defined on successive intervals of times. This method of representation makes it possible to compress the volume of data and to calculate the positions of the bodies quickly. For these reasons, it is very well adapted to the needs of astronomers (amateurs, professionals, theoreticians, astrometrists, etc).

The Chebyshev ephemerides appear as a succession of coefficients $a_0, a_1, \ldots a_n$ for a coordinate given on an interval $[t_1, t_2]$. The use of these ephemerides is as follows. The calculation of the value of the coordinate at the time t of the interval $[t_1, t_2]$ is done by carrying out the change of variable: $x = -1 + \{2(t - t_1)\}/\{t_2 - t_1\}$. Therefore x belongs to the interval $[-1, +1]$. The Chebyshev polynomials of the first kind $T_n(x)$ are given by the relation: $T_n(x) = \cos(n \arccos x)$. One can also calculate them using the relation of recurrence: $T_n(x). = 2x.T_{n-1}(x) - T_{n-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$. The position is then calculated through the following sum: $a_0.T_0(x) + a_1.T_1(x) + \cdots + a_n.T_n(x)$ will provide the result.

Approximation by mixed functions

As we have seen, the polynomial approximation is not the only one. The introduction of mixed functions, makes it possible to take into account the quasi-periodical character of the functions appearing in the ephemerides. Indeed, a function of a quasi-periodical nature may be approximated on a given interval by a development in polynomials of relatively low degree n, but the fact that the function is quasi-periodical on the interval, which is the case in the motion of the bodies of the solar system, is not taken into account.

Let us see, through an example, how to benefit from the quasi-periodical nature of a function. Let the even function be $f(x) = \cos x + \cos 2x$. It is much better represented with the base of functions $1, \cos x, \cos 2x, \cos 3x \ldots$ than with the polynomials $1, x^2, x^4, x^6, \ldots$ or with any other basis of even polynomials.

However, an ephemeris is not so simple but it is possible, on a given interval of time, to make the approximation using several periodic functions, the period of which is suggested by the physical nature of the problem (for example the period of revolution of a celestial body), and also using some secular terms (or some polynomials of time) because of the not rigorously periodic nature of the described motion.

The representation using mixed functions is particularly well adapted to the motions of natural satellites of planets, and more generally for the motions of short period bodies, the motion of which is never far from a periodical motion. In the *Connaissance des Temps*, the differential coordinates of the natural planetary satellites are described by coefficients of mixed functions on intervals covering several revolutions of the satellites and are easy to calculate using the following formula:

$$f(t) = a_0 + a_1 t + b_0 \sin(\nu t + \varphi_0) + b_1 t \sin(\nu t + \varphi_1) + b_2 t^2 \sin(\nu t + \varphi_2) + c_0 \sin(2\nu t + \psi_0)$$

where $2\pi/\nu$ is a period close to the period of revolution of the body; a_j, b_j, and c_j, are the amplitudes; φ_j and ψ_j are the phases ($j = 0, 1, 2$). This formula rises from the choice of the selected basis. The mixed terms have the form: $at^k \sin(p\nu t + \psi)$ (k and p: integers; a: amplitude; ψ: phase). The maximum values of k and p are adjusted according to the required precision. If the interval of representation is widened, it will be necessary to enrich the basis by mixed terms in power of time and multiple of the basic frequency. However, if the interval decreases, it is necessary to modify the basis in order to avoid the loss of precision (this is not the case with a polynomial representation, the precision of which increases when the interval decreases).

The status of current research on ephemerides

The first way to improve ephemerides (i.e. to improve the accuracy of the positions provided by the ephemerides) is to improve the theories used for the motions of the celestial bodies. At the present time, an effort has been made to take into account all the gravitational effects occurring in the solar system and to include the relativistic effects. Improved observations are needed and, for the inner planets (Mercury to Mars), the radar observational method has led to the abandonment of optical observations. For the giant planets, optical observations are still the only way to observe the positions of the planets together with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the small number of positions sent by the space probes. The observation of planetary satellites is now preferable to the use of direct observations of the planets themselves. At last, the determination of the DYNAMICAL REFERENCE FRAME (a reference frame deduced from the motion of the Earth around the Sun) and the linkage to the independent stellar reference frame (no longer based on stars but on very far sources; see also EXTRAGALACTIC REFERENCE FRAMES) is the last step for the improvement of the ephemerides, solved by spatial observations from artificial satellites such as Hipparcos in the past and Fame or Gaia in the future.

The second way to improve ephemerides is to use better representations of the motions. In fact, the problem is mainly in the reduction of the amount of data. Chebyshev approximation is still in use for that purpose.

Various supports used for the publication of the ephemerides

Printed ephemerides

Although microcomputers are omnipresent for astronomical calculations, the printed ephemerides are still abundantly fused by professional and amateur astronomers. Why? First of all, the observer often needs a confirmation or a simple approximate position. Opening a book and directly reading a position is easy. One can also find:

- hours of rising and setting of the Sun and the Moon (and also of principal planets);
- positions of the bodies of the solar system and bright stars for a terrestrial observer;
- ephemerides for the physical observations of these bodies.

These ephemerides are in general of weak precision and a tabulated form is sufficient.

Let us quote also some specialized ephemerides such as the *Nautical Almanac* or the *Ephémérides Nautiques* published for navigational use. However, not all the printed ephemerides provide tabulated positions. Graphical representations may be provided (essentially for the Galilean satellites of Jupiter) as well as coefficients for polynomials (in the case of high precision ephemerides) needing the use of a pocket calculator or a microcomputer. This is the case for the *Connaissance des Temps* since 1979 which before then was used for the manual programming of telescopes or more simply for the tests and the checks of more complex software.

The following is a list of the main printed ephemerides published nowadays:

- by the Institut de Mécanique Céleste/Bureau des Longitudes, France: *Connaissance des Temps*, *Annuaire du Bureau Des longitudes*, *Ephémérides Nautiques*;
- by H M Nautical Almanac Office of the former Royal Greenwich Observatory now at the Rutherford Appleton Laboratory, United Kingdom and by the Nautical Almanac Office of the USNO, USA: *Astronomical Almanac*, *Nautical Almanac*:
- by the Institute of Applied Astronomy, St Petersburg, Russia: *Ephemerides of Minor Planets*;
- by the Astronomisches Rechen Institut, Heidelberg, Germany: *Apparent Places of Fundamental Stars*.

Electronic ephemerides

The paper medium is traditional for the ephemerides which have existed for centuries. Nowadays, the use of the printed ephemerides is limited to specific help. The search for a body in the sky or the guidance of a telescope following a fast object of the solar system is now made in an automatic way by using electronic ephemerides.

The first way is to have the printed ephemerides put on a medium (diskettes, CD-ROM) associated with a program of reading and of calculation of a position at a given date. It is obvious that the calculations of positions using the basic dynamical theories is not possible in real time because of the complexity of calculations and of the volume of data to be handled. The reading of tabulated values followed by their interpolation (not easy) or better, the reading of coefficients followed by precise calculations of the associated positions is today commonly used. For the calculations of positions not taking place in real time, it is also possible to carry out interactive calculations online using a specialized server. Web servers on the Internet allow these calculations now and ephemerides of high precision are thus available for all. The high precision ephemerides used by professional astronomers and by the space agencies are provided, of course, in an electronic form. The following is a list of addresses of the most interesting professional servers reachable on the Internet.

Server of the Jet Propulsion Laboratory, Pasadena, CA, USA: http://ssd.jpl.nasa.gov/

Server of Institut de Mécanique Céleste/Bureau des Longitudes, Paris, France: http://www.bdl.fr/

Server of the US Naval Observatory, Washington, DC, USA: http://aa.usno.navy.mil/AA/

Server of the HM Nautical Almanac Office, United Kingdom: http://www.ast.cam.ac.uk/

Server of the Minor Planet Center, Cambridge, MA, USA: (ephemerides of asteroids and comets exclusively) http://cfa-www.harvard.edu/iau/services/CS.html/

Bibliography

Introduction aux Éphémérides Astronomiques (Bureau des Longitudes) Editions de Physique

Explanatory Supplement to the Astronomical Ephemeris (US Naval Observatory/Royal Greenwich Observatory)

Jean-Eudes Arlot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Epicurus of Samos (341–270 BC)

Greek philosopher, founded Epicurean philosophy in Athens, which divided the universe into 'atoms' and the 'void.' Epicureans regarded happiness as the chief human good, attainable through freedom from anxiety and fear, and their materialism was despised in subsequent years as anti-Christian.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Epoch

A particular instant of time used for reference purposes. In the context of the measurement of stellar positions, the epoch of observation is the date on which the observational data were obtained. A catalog for the epoch 2000.0, for example, would list positions, valid for that date. To obtain positions at some other epoch, the effects of proper motion must be calculated. Other effects such as nutation and aberration have also to be taken into account in order to obtain a precise position for an object on a particular date.

Note that the epoch of observations of star positions does not have to be the same as the equinox to which the coordinate system is referred. The positions of a high proper motion star at epochs 1972.3, 1984.5, 1999.1 could be plotted in the coordinate frame of the equinox 1950.0, for example.

In the context of planetary motion, the term 'epoch' may be used to describe the time of perihelion passage (the time at which the planet makes its closest approach to the Sun), this being one of the elements of the orbit.

See also: equinox, precession, proper motion, nutation, aberration, orbital elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Epsilon Aurigae

This star, sometimes known as 'Almaaz' or 'Al Anz', is an eclipsing binary of the same type as Algol, but of exceptionally long period. It is a pale yellow supergiant of spectral type F0Ia, apparent magnitude 3.03v. It is intrinsically brilliant (absolute magnitude −6.0), its luminosity having been estimated as about 200 000 times that of the Sun, but it is 2040 light-years away, with a parallax of 0.002″. The eclipses last for about a year, and occur at intervals of 27.1 years; the next eclipse is due in 2010.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Epsilon Lyrae

Though not very bright, this is a favorite test object for powerful binoculars and small telescopes. Situated 1.6° north-east of Vega, it is known as 'the Double Double'. The star can just be seen to be double with the unaided eye, under good conditions by observers with acute eyesight. With binoculars it is easily separated, the component stars being 3.5′ apart at a position angle of 173°. It is in fact a quadruple star system, each component being itself a binary. With a good 3 inch (7.5 cm) refractor both can be resolved. The system is 160 light-years distant (parallax 0.020″).

The first component system, ε^1 Lyr, which has a combined apparent magnitude of about 4.7, consists of two white stars separated by 2.5″ at position angle 353°, with an orbital period of about 1165 years. The apparent magnitude of ε^1 Lyr A is 5.06, its spectral type A3V and absolute magnitude 1.6; ε^1 Lyr B has an apparent magnitude of 6.19, spectral type F1V and absolute magnitude 2.7.

The second system, ε^2 Lyr, has a combined magnitude of about 4.6 and is very similar to the first pair. It again comprises two white stars, separated by 2.3° at position angle 87°, orbital period 585 years. The apparent magnitude of ε^2 Lyrae C is 5.30, its spectral type A5V and absolute magnitude 1.8; ε^2 Lyr D has an apparent magnitude of 5.51, spectral type F0Vn and absolute magnitude 2.0.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equation of Time

An equation giving the difference between apparent solar time (or sundial time) and mean time. Thus:

$$\text{equation of time } (E) = \text{apparent solar time (AT)} - \text{mean time (MT)}.$$

A positive value E means that apparent time is ahead of mean time. A negative value of E means that apparent time is lagging behind mean time.

The difference arises for two reasons. First, since the Earth moves round the Sun in an elliptical orbit, when the Earth is near perihelion it moves faster than when it is near aphelion; consequently, the apparent motion of the Sun relative to the background stars is more rapid near perihelion than near aphelion. This variation in the Sun's apparent motion affects the length of the apparent solar day. Second, although the real Sun's apparent motion is along the ecliptic, the variation in the length of the apparent solar day depends upon the daily change in the right ascension of the Sun. This change in RA is equal to the projection of the Sun's ecliptic motion onto the celestial equator, and is maximum near solstices and minimum near equinoxes.

As a result of the combination of these two effects, mean time and apparent solar time can differ by up to 16.3 min. Apparent time is ahead of mean time between about 16 April and about 15 June and, again, between about 2 September and about 26 December, the difference reaching maxima of about 3.75 and 16.3 min on 15 May and 4 November, respectively. Apparent time lags behind mean time during the rest of the year, the difference reaching maxima of 14.3 min and 6.3 min on 12 February and 27 July, respectively.

See also: epoch, mean solar time, sidereal time, sundial.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equatorial Mounting

A telescope mounting in which one axis of rotation (the polar axis) is aligned parallel to the Earth's axis, and the other (the declination axis) is aligned at right angles to this axis, in the plane of the celestial equator. Rotation about the polar axis allows the telescope to be pointed towards celestial bodies with different right ascensions or hour angles, while rotation about the declination axis enables the telescope to be swivelled perpendicular to the celestial equator in the direction of increasing or decreasing declination. The advantage of the equatorial mounting is that the apparent diurnal motion of a celestial body can be followed by driving the telescope in the opposite direction to the Earth's rotation (i.e. driving from east to west) at a rate of one revolution per sidereal day around the polar axis only. The disadvantages of the equatorial mounting are that it is more complex and expensive to construct than the altazimuth mounting, and gravitational loadings change in a complex way as the telescope moves, so making it difficult to compensate for structural flexure. Nevertheless, until comparatively recently, the overwhelming majority of large telescopes were set up on equatorial mountings.

There are many varieties of equatorial mounting. Popular varieties include the German, cross-axis and fork mountings. In the German mounting, the declination axis sits across the top end of a short polar axis, with the telescope at one end and a counterbalancing weight at the other. In the cross-axis mount, a long polar axis is supported at top and bottom by two pillars, or piers, and the telescope is attached to a declination axis that crosses the polar axis roughly mid-way between its top and bottom ends. Once again, the weight of the telescope is counterbalanced by a weight placed at the opposite end of the declination axis. In the fork mounting, the telescope swings in declination about two pivots attached at opposite sides of the upper end of an open U-shaped fork, the bottom of which is attached to a short polar axis.

See also: altazimuth mounting, celestial coordinates, celestial equator, declination, diurnal motion, hour angle, right ascension, sidereal time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equator-S

Small German satellite launched in December 1998 to investigate Earth's equatorial magnetosphere. Operations began in February 1999, but ceased after three months due to a system malfunction. The first satellite to receive positional information from GPS satellites while beyond geostationary orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equinox

An instant at which the Sun crosses the celestial equator; the Sun is then vertically overhead at the equator, and day and night have equal duration at every point on the Earth's surface. The apparent annual path of the Sun on the celestial sphere is inclined to the celestial equator and intersects it at two points. The terms vernal equinox and autumnal equinox are applied to these points.

Star positions are measured in celestial coordinates based on a particular vernal equinox, say the equinox of the year 2000. Thus, the word 'equinox' applied in the context of the position of a star defines the coordinate system being used.

See also: celestial equator, celestial sphere, ecliptic, First Point of Aries, First Point of Libra, precession.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equuleus

(the Little Horse; abbrev. Equ, gen. Equulei; area 72 sq. deg.) A northern constellation which lies between Delphinus and Pegasus, and culminates at midnight in early August. Its origin is uncertain, though its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

Equuleus is the second smallest of the 88 constellations (and the smallest in the northern sky) with only one star, α Equulei (Kitalpha), magnitude 3.9, brighter than fourth magnitude. Interesting objects include ε (or 1) Equulei, a quadruple star system consisting of two pale yellow (F5 and F7) components, magnitudes 6.1 and 6.4, separation 0.9″, period 101 years, the former of which has an unseen companion which revolves around it in a period of 2.03 days, and a fourth, yellow (G0) component, magnitude 7.4, separation 10.3″.

There are no bright star clusters, nebulae or galaxies in Equuleus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eratosthenes of Cyrene (*c*. 276–*c*. 196 BC)

Astronomer, philosopher, geographer and geometrist (developed theorems in prime number theory), born in Cyrene (now Shahhat, Libya) became librarian in Alexandria. Eratosthenes made a surprisingly accurate measurement of the size of the Earth. Having heard that at the summer solstice the Sun illuminated the bottom of a well at Syene (now Aswan, south of Alexandria), and was therefore overhead, he observed the angle of the Sun off the zenith on the same day at Alexandria. He measured the distance between the two cities, allegedly by counting the rotations of the wheel of a carriage driven between them, and determined the circumference of the Earth as 250 000 stadia (47 000 km, compared to the true value of 40 000 km). He also measured the tilt of the Earth's axis with great accuracy obtaining the value of 23 degrees 51 minutes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Erfle, Heinrich Valentin (1884–1923)

Optician, born in Duerkheim, Germany, head of the telescope department for CARL ZEISS, Jena; the *Erfle wide field eyepiece* is named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eridanus

(the River; abbrev. Eri, gen. Eridani; area 1138 sq. deg.) A southern constellation which lies between Cetus and Orion and extends from the celestial equator to within 32° of the south celestial pole. The center of the constellation culminates at midnight in mid-November, though its brightest star is on the meridian a month earlier. Eridanus has been associated with several different terrestrial rivers, notably the Nile, which also lies south–north, and the Po, into which, in Greek mythology, Phaëthon, son of the Sun-god Helios, was said to have fallen after Zeus struck him with a thunderbolt during his abortive ride of his father's chariot. The brightest stars of Eridanus south to about dec. −40° were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*; the more southern ones (including its brightest star) were added in Renaissance times and are shown, for example, in Bayer's *Uranometria* of 1603.

A large but rather inconspicuous constellation, the brightest stars in Eridanus are α Eridani (Achernar), which at magnitude 0.5 is the ninth brightest star in the sky, β Eridani (Cursa), magnitude 2.8, and γ Eridani (Zaurak), magnitude 3.0. There are 14 other stars of magnitude 4.0 or brighter. Interesting stars include ε Eridani, which at magnitude 3.7 is one of the nearest naked-eye stars (distance 10.5 light-years) and a similar star to the Sun, θ Eridani, a fine binary with bluish-white (A4 and A1) components, magnitudes 3.3 and 4.4, separation 8.3″, o^2 Eridani (Keid), a triple-star system with an orange (K1) primary, magnitude 4.4, a white dwarf secondary, magnitude 9.5, separation 93″, and a third, red dwarf (M4.5) component, magnitude 11.2, which orbits the latter, separation 9.0″, period about 248 years, and has a mass of only about 0.2 that of the Sun, and 32 Eridani, another binary with yellow and bluish-white (G8 and A2) components, magnitudes 4.8 and 6.1, separation 6.5″. Other interesting objects include NGC 1535, a tenth-magnitude planetary nebula, and NGC 1291, a ninth-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eros

The first Amor asteroid to be discovered, independently by Gustav Witt and Auguste Charlois in 1898, designated (433) Eros. The asteroid orbits the Sun at a mean distance of 1.46 AU, or 218 million km (perihelion 1.13 AU, aphelion 1.78 AU), in a period of 1.76 years; inclination 11°–, eccentricity 0.22. Its rotation period is 5.3 h. Eros is an S-type asteroid, with a reflection spectrum similar to that of ordinary chondrite and stony-iron meteorites. It was the prime target of the NEAR Shoemaker probe, which entered orbit around Eros on 14 February 2000. Images returned showed an irregular, elongated object measuring 33 × 13 km, with a 7 km diameter crater. The slight deviation to the probe's trajectory revealed that Eros has a bulk density of about 2.5 g cm^{-3}. At 3:02:10 EST on 12 February 2001, the NEAR Shoemaker spacecraft landed on the surface of Eros, the first spacecraft ever to land on an asteroid. It traveled its last mile by cruising down to the surface of Eros at a gentle 4 mph (1.9 m s^{-1}), finally coming to rest after its 2 million mile journey. The last image snapped by the spacecraft was a mere 394 ft (120 m) from the asteroid's surface and covered an area 20 ft (6 m) square.

See also NEAR SHOEMAKER

Bibliography

Veverka J *et al* 2001 *Nature* **413** 390–393

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Escape Velocity

The minimum velocity required in order that an object can leave the surface of, or depart from any other point within the gravitational field of, a massive body and recede to a large distance. In principle, if atmospheric resistance is ignored, a projectile fired vertically from the surface of a massive body at its escape velocity will continue to move away with a speed which decreases with increasing distance but does not decline to zero until it has receded to an infinite distance. An object moving through the gravitational field of a massive body will, in general, follow a parabolic trajectory; for this reason, escape velocity is sometimes called parabolic velocity. The escape velocity at the surface of the Earth is about 11.2 km s^{-1} (6.96 miles). The velocity of escape from the less massive Moon is about 2.4 km s^{-1} at its surface. A planet cannot long retain an atmosphere if the planet's escape velocity is low enough to be near the average velocity of the gas molecules making up that atmosphere.

See also: weight, zero gravity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eta Aquarids

A meteor shower that takes place in late April and May. The radiant lies in the constellation Aquarius. The Eta Aquarids occur when the Earth intersects the descending node of the meteor stream from Halley's Comet; the Orionids in October are produced by the Earth's passage through the ascending node. Because Halley's orbit is retrograde, Eta Aquarid meteoroids impact the Earth at a high relative velocity and produce fast meteors. There are records of this shower dating back to 74 BC.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eta Carinae

One of the most enigmatic stars known. It is an S Doradus-type eruptive variable, situated at the heart of the bright diffuse nebula NGC 3373. It is obscured by the Homunculus Nebula, a bright cloud of gas and dust ejected by the star itself during a major outburst in 1843, when it reached an apparent magnitude of -0.8, becoming temporarily brighter than all other stars except Sirius. Its apparent magnitude is at present 6.2 but varies erratically over decades and it has been recorded as faint as 7.9. It is an extremely powerful source of infrared radiation, and is believed to be a highly luminous blue star much of whose visible radiation is absorbed by the nebula and re-emitted in the infrared. Because of the obscuration its spectral type, parallax and absolute magnitude cannot be determined with any certainty.

It has been variously suggested that η Carinae is a very young star in its pre-main-sequence stage, or an extremely old star in the closing stages of its evolution, which is now believed more likely. The last phase of its existence could well result in one of the most spectacular supernova explosions ever witnessed, but there is no means of reliably predicting when this will occur.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eta Carinae Nebula

A bright, diffuse nebula, NGC 3373, surrounding the star η Carinae. About 2° across and divided by a dark, V-shaped, obscuring dust-lane, it contains a number of interesting stars and star clusters, and also two major included nebulae—the bright Homunculus Nebula and the dark, absorbing Keyhole Nebula. Its distance has been estimated as about 8000 light-years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ether (or Aether)

A transparent, weightless medium which was imagined by nineteenth-century physicists to fill all space, and to provide the medium through which electromagnetic waves could propagate. Conventional physical ideas of the time did not admit of waves that could travel through 'nothingness'. The concept of the ether came to be associated with the idea of absolute space, the ether providing a background to the universe against which all motion could be measured.

Experiments were devised, notably the Michelson–Morley experiment, to try to measure the motion of the Earth through the ether, and these all failed to detect any such motion, however small. The null result of these experiments was explained by the theory of relativity, first published in 1905. The ether, being undetectable, became an unnecessary hypothesis, and has been abandoned (along with the Newtonian concept of absolute space).

See also: Michelson–Morley experiment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eudoxus of Cnidus (*c*. 400–*c*. 347 BC)

Greek astronomer, mathematician, born in Cnidus (on the Resadiye peninsula, now Turkey). First astronomer to produce a detailed description of the stars and constellations. His work is now lost but known through the extensive borrowing by ARATUS in his *Phenomena*. Proposed a geocentric system for the solar system in which the Sun, Moon and planets moved on spheres centered on the Earth. The model failed to account for variations in the observed diameter of the Moon or changes in the brightness of planets, which were correctly interpreted to indicate that their distances were changing. He noticed some of the discrepancies with observations and tried to adjust the model by postulating that each sphere had its poles set at an angle to the next sphere. His model contained no mechanical explanation and was only a mathematical description, but the spheres came later to be regarded as having a physical reality. Some of his mathematics became a foundation for parts of Euclid's *Elements*, indeed may have been due to him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Euler, Leonhard (1707–83)

Swiss mathematician, born in Basle. Developed theories of the planets, the Moon, and the tides. His formulated the concept of *Eulerian angles* in which to express dynamics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Euro-Asian Astronomical Society

The Euro-Asian Astronomical Society (EAAS) was created in April 1990, and is active mainly in New Independent States (NIS), the former USSR territory. Its members include professional NIS astronomers, astronomical associations and institutions worldwide.

Its purpose is to maintain and develop NIS astronomy and to reinforce scientific contacts between astronomers of NIS and the rest of the world.

The society organizes astronomical meetings, astronomical education and related topics. It supports library programs and provides free of charge, or very cheap, subscriptions to Russian-language astronomical/space research journals and magazines for astronomical institutions of NIS and some Central European countries (except Russia).

Its publications include a professional English Journal, Astronomical and Astrophysical Transactions, a popular Russian publication translated into English, *Universe and Ourselves*, and an electronic newsletter, in Russian but translated into English.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Europa

Europa is the smallest of the four large 'Galilean' satellites of JUPITER. Table 1 gives its basic parameters. Little known before 1979, when the VOYAGER spacecraft returned the first close-up images, it has recently been a prime target of the Jupiter-orbiting GALILEO MISSION, which has provided major advances in our understanding of Europa. Its icy surface and rocky interior, and strange surface topography of low-relief ridges and chaotic terrain, make Europa one of the most unusual objects in the solar system. Most intriguing is the possibility that this unusual surface overlies a liquid ocean, which might conceivably support life.

Table 1. Properties of Europa.

Radius	1561 km
Density	3005 kg m^{-3}
Mean distance from Jupiter	671 000 km
Orbital eccentricity	0.0101
Surface composition	H_2O ice, hydrated minerals (perhaps salts or H_2SO_4); SO_2; O_2; CO_2; H_2O_2
Interior composition	Silicate rock; Fe/FeS; H_2O

Origin

Europa was presumably formed at the same time as its three siblings, IO, GANYMEDE, and CALISTO, from the same cloud of material, the proto-Jovian nebula, that formed Jupiter itself. Temperatures in the proto-Jovian nebula would have been highest near Jupiter, with ice condensing only in the outer parts. Europa formed near the inner margin of the ice zone, and so acquired a little ice in addition to rock and iron. The solid material in the nebula accumulated into ever-larger pieces which were drawn together by mutual gravitational attraction, until four large bodies, the nascent Galilean satellites, became large enough to sweep up the remaining debris. After formation, the satellites were driven away from Jupiter by their tidal interaction with the planet. Eventually the inner three became locked by mutual gravitational perturbations into the present orbital resonance, in which Io orbits Jupiter twice for every orbit of Europa, and Europa orbits twice for every orbit of Ganymede, a circumstance that has profound consequences.

Interior

Europa's mean density, 3000 kg m^{-3}, is much greater than the density of water, 1000 kg m^{-3}, but is significantly less than the density of around 3500 kg m^{-3} that would be expected if the interior were made entirely of silicate rocks. Its interior is strongly differentiated, with a probable iron or iron sulfide core, a thick mantle of silicate rock, and an outer layer of water or ice. The DIFFERENTIATION of Europa indicates that at some time its interior has been hot and mobile enough for the heavier constituents to settle towards the center. Heating during its formation, radioactive decay, or tidal heating may have been responsible. Unlike its neighbor Ganymede, Europa does not generate a significant internal magnetic field, though as the origin of Ganymede's field is a mystery we are not yet sure what this means for Europa. However, Europa perturbs Jupiter's magnetic field in a way that implies an electrically conducting layer, such as a liquid ocean, not far below Europa's surface. The same appears to be true on Callisto, a geologically dead moon that otherwise looks very different from Europa. The orbital resonance with Io and Ganymede makes Europa's orbit significantly non-circular, so although Europa rotates synchronously, keeping the same face to Jupiter, the tidal bulge raised by Jupiter changes size and orientation slightly during the orbit as Europa's distance from Jupiter varies. This continual distortion dissipates energy inside Europa, in a milder version of the tidal heating that produces Io's volcanos. Dissipation is greatest in the parts of the interior that deform the most during the orbit, but the amount of deformation of the ice or rock depends on its viscosity,

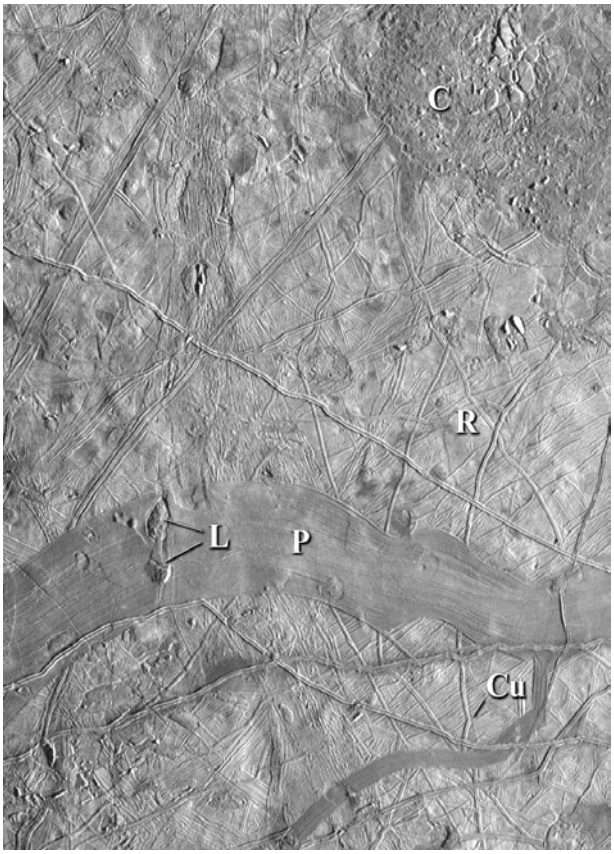


Figure 1. Galileo image of a 215–280 km region of Europa's surface, showing the wide variety of surface features. Examples are shown of ridged plains (R); chaos (C); lenticulae (L); a pull-apart band (P); and one 'cusp' of a cuspate ridge (Cu). Note the almost complete absence of impact craters.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

which depends on its temperature, which in turn depends on the amount of tidal heating, so there are complex feedback mechanisms that are difficult to model. However, it seems likely that most of the dissipation occurs in the lower, warmer, part of the solid ice layer, and that the heating rate is roughly 0.05 W m^{-2} or less.

Surface geology

Europa's surface can be divided into several very distinct types of terrain (figure 1), each of which is described separately below.

Ridges

Much of Europa's surface is made up of innumerable overlapping ridges, tens to hundreds of meters high and up to thousands of km long. The more recent ridges tend to stand higher than the older ones, and are often flanked by diffuse regions of lower ALBEDO. Most frequently, the ridges are double, with a central trough, but some have multiple components. Some are remarkably straight, while others have irregular, lobed, margins, and others have strikingly regular 'cuspate' forms, like the path of a bouncing ball seen from the side. These are most likely formed from propagating fractures that spread in a direction that changes periodically during each orbit of Europa because of changing stresses from Jupiter's tides. The most extensive ridges are global in extent, and with their dark margins form the dark 'fractures' that are Europa's most distinctive feature in global views. The double ridges seem to start life as simple fractures, which then somehow gradually develop ridges on either side. If the fracture opens further, additional ridges may appear within the central trough, producing complex sets of ridges. The ridges may develop by upwarp of the material on either side of the fractures by shallow intrusions along the fractures, or by accumulation of debris on either side of the fracture. The debris may accumulate either by 'volcanic' eruption along the fracture, or squeezing of material out of the fracture by its periodic opening and closing, perhaps by daily tides.

The ridges form a global pattern that provides evidence that Europa's outer shell is detached from its interior, and due to Jupiter's tides rotates very slightly faster than the interior, completing a full rotation relative to Jupiter over thousands or millions of years. The shell will be distorted and stressed as it rotates over the tidal bulge that always faces Jupiter, and the expected stresses from this process very nicely explain the global alignment of the ridges, and the apparent changes in their orientation with time. Rhythmic changes in the stress direction during each orbit of Jupiter may even explain the remarkable 'cuspate' ridges, if each 'bounce' of the ridge forms during a single Europan day.

Pull-apart features

In many areas, the ridges are broken by smoother, darker, bands where the surface has clearly been pulled apart: features on either side of the band match up perfectly, as in figure 1. The smooth material in the band is new material that has filled the gap as the surface on either side separated, much like sea-floor spreading on Earth. The global distribution and orientation of the pull-apart features suggests that they open in response to stress from the daily tides. Somewhere on Europa, there must have been compression of the surface to make room for the observed spreading, and it is odd that no unambiguous compressional features have yet been identified.

Chaos and lenticulae

Large areas of Europa's surface have been fractured and broken up to form 'chaos'. Large chaos regions may be more than 100 km across, while the smaller but similar lenticulae are typically 10 km in diameter. A possible evolutionary sequence for lenticulae can be inferred. In places the original surface is simply domed up; elsewhere the dome is fractured, frequently to the point where the original surface texture is unrecognizable. Chaos regions are similar but larger. The fractured blocks of original surface have often been moved horizontally, rotated, or even tilted: much of the nearby original surface must have been destroyed to make room for this motion. Chaos and lenticulae are generally darker than the original surface: the dark mottled terrain seen in global views is mostly chaos. Both types of terrain appear to have formed by disruption of the surface from below, probably due to enhanced local heating. However, whether the heating involved large-scale melting, with the broken crustal fragments floating like icebergs, or simply the rising of warm, mobile, solid ice towards the surface, is not yet clear. Feedback between ice viscosity and tidal heating, discussed above, may produce local thermal runaway and explain the patchy distribution of the chaos regions.

Impact features

Europa's surface is more lightly cratered than most PLANETARY SURFACES, an indication of its relative youth. The largest craters seen so far, Tyre and Callanish, roughly 80 km in diameter, are composed of many concentric circular fractures surrounding a rough central region. The concentric fractures suggest that the impacts penetrated the rigid outer layer of Europa into a fluid layer below: the fractures would have formed as material rapidly flowed towards the center of the hole created by the impact. The Galileo spacecraft has shown craters larger than 25 km in diameter are extremely flat compared to similar sized craters on other bodies, and are also surrounded by multiple concentric rings. It is likely that these craters have penetrated the ice into the probable

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

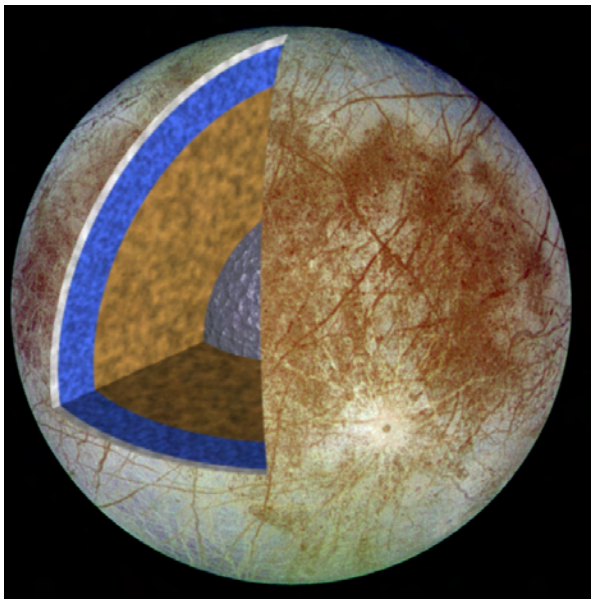


Figure 2. Full-disk view of Europa from Galileo. The bright impact crater Pwyll is conspicuous. A cutaway shows a schematic of the interior structure: the cold, rigid outer icy crust (white), the possible liquid water ocean or warm ice layer (blue), the silicate layer (brown), and the central iron core (silver). This figure is reproduced as Color Plate 28.

ocean below. Smaller impacts, such as the bright ray crater Pwyll (figure 2), 25 km in diameter, look superficially like lunar craters, with a sharp rim and central peak, and thus probably did not penetrate to a fluid layer. Taken together, these constraints imply a thickness of 10–20 km, though the models are incomplete, and uncertainties remain for the rigid outer layer, They also do not prove that the underlying fluid layer is liquid water: very warm, mobile, ice is also possible. Pwyll-sized craters are extremely flat compared to lunar craters, showing that even the 'rigid' outer layer is soft enough below the surface to flow and fill in the original hole.

Age of the surface

The age of the geological features on Europa's surface is difficult to determine precisely. No detectable changes occurred in the 17 years between the Voyager and Galileo images: if we assume that changes occur smoothly and uniformly over Europa's surface, this implies an average surface age of at least 30 million years. The surface age can also be estimated from the abundance of impact CRATERS, combined with estimates of the rate at which impacts are expected to occur. Cratering rates can be determined, though not very accurately, from the number of observed comets and asteroids near Jupiter. Our best estimates suggest that Europa's surface age is somewhere between 10 and 300 million years. If this age estimate is correct, it is very likely that geological activity on Europa continues today: the chance that activity would have persisted for 95–99% of Europa's history and then stopped just before humans appeared on the scene is small. This raises another mystery, however. Chaotic terrain on Europa is more recent than ridged plains in almost all areas of Europa so far surveyed. If Europa's surface is young, why would geological activity have changed so dramatically so recently? One possibility is that activity on Europa is cyclic, driven by complex interactions between its interior and its orbit, and that at some future date the style of activity will change back from chaotic terrain formation to ridged plains formation.

Interaction with Jupiter's magnetosphere

Europa is immersed in Jupiter's enormous magnetosphere, which rotates with Jupiter once every ten hours (see JUPITER: MAGNETOSPHERE), much faster than Europa's 3.5-day orbital period. The magnetosphere carries with it a dense plasma of energetic ions, dominated by sulfur and oxygen ions that have escaped from Io. Lacking a strong intrinsic magnetic field that could deflect the plasma, Europa's surface, particularly the trailing hemisphere (the side that faces backwards as Europa orbits Jupiter), is thus continually bombarded by energetic ions. The bombardment has two main effects: implantation and erosion.

The strongest direct evidence for implantation is an absorption feature in Europa's ultraviolet spectrum, due either to sulfur dioxide or sulfur–oxygen bonds in some other form. The feature is strongest on the trailing hemisphere, where bombardment is strongest. There is also a pervasive darkening of Europa's trailing hemisphere at longer ultraviolet wavelengths. It is likely that some of the bombarding sulfur ions become embedded in the water ice of Europa's surface, and combine with oxygen from the surrounding water to produce the UV absorption feature. It is also possible that the UV feature is due to sulfur intrinsic to Europa, concentrated on the trailing side by sputtering away of water ice, as explained in the next paragraph.

Each ion that collides with Europa's water ice also ejects many water molecules, or fragments of water molecules, from the surface—a process known as sputtering. Most of the material re-impacts Europa elsewhere on its surface, but a significant fraction is moving fast enough that it escapes into space. It is estimated that, on average, 5–20 cm of ice is lost from Europa every million years, and 20–80 cm is transported large distances across the surface. The fact that geological features on Europa as small as tens of meters in size have not been worn down to nothing by this process provides additional evidence that the surface is relatively young, less than 100 million years old. Sputtering of the surface ice also supports a tenuous oxygen atmosphere around Europa, which was detected by the Hubble Space Telescope. There is also a faint but detectable halo of sodium

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

around Europa, perhaps also derived ultimately from Io.

Recent images from the Hubble Space Telescope show a faint far–ultraviolet auroral 'footprint' on Jupiter's upper atmosphere at the point where a Jovian magnetic field line connnects Europa to Jupiter. This shows that Europa is electrically conducting, probably through its tenuous ionosphere, and drives electrical currents between itself and Jupiter as it moves through the Jovian magnetic field. Io and Ganymede show similar, but stronger, auroral footprints.

Ocean? Life?

The most intriguing question about Europa is whether there is an OCEAN of liquid water beneath the ice that might even support life. Many observations, including the pull-apart features, evidence for nonsynchronous rotation of the outer shell, the fractured and rotated blocks in the chaos regions, and the unusual appearance of Europa's impact craters, point to the existence of a soft, mobile layer a few kilometers beneath a rigid outer crust. The big question is then whether this layer is warm ice not far below its melting point, or is actually liquid. The evidence for nonsynchronous rotation, and electromagnetic evidence for a liquid layer beneath the surface, are perhaps the strongest evidence for an ocean rather than warm ice. Hybrid models, in which a 'lithosphere' of cold, rigid ice overlies warmer, convecting ice, which in turn overlies a liquid ocean, are also plausible.

If an ocean has existed over most of Europa's history, conditions there may be suitable for the development and sustenance of life. Organic (carbon-based) molecules that provide the raw material for life are common in the solar system, and while sunlight probably does not penetrate the outer ice shell, tidal heat may have supplied sufficient energy for the chemical reactions necessary for life. A commonly drawn analogy is with the living communities that exist independently of sunlight, sustained only by chemical and thermal energy from the EARTH'S INTERIOR, around the hydrothermal vents or 'black smokers' on the floors of the Earth's oceans. Currently, however, we have no direct evidence for life on Europa, and further exploration will be required to determine whether oceans or life really exist there (see also LIFE ON OTHER WORLDS).

Future exploration

Plans are being made for a Europa Orbiter mission that could be launched as early as 2003 and might enter Europa orbit around 2008, with the primary goal of establishing the existence (or otherwise) of a subsurface ocean. In addition to the usual cameras and spectrographs, the spacecraft would probably carry a radar to probe beneath the surface, with the possibility of detecting any ocean directly. A laser altimeter and gravity measurements would directly detect the distortion of Europa's surface as its distance from Jupiter changes—a thin ice shell over an ocean would distort more than a thick layer of solid ice. After the orbiter mission, landers that could sample the surface directly, and eventually probes that could drill through the ice to the ocean (if it exists) are being considered. The technical challenges of these missions, particularly the large quantity of fuel required to rendezvous with Europa, and the harsh radiation environment, are extreme, and overcoming them will be challenging. Europa will keep many of its secrets for a while longer.

Bibliography

Clarke J T *et al* February 2002 Ultraviolet emissions from the magnetic footprints of Io, Ganymede and Europa on Jupiter *Nature* **415** 997–1000

Turtle E P and B A Ivanov 2002 Numerical simulations of impact crater excavation and collapse on Europa: Implications for ice thickness. *Lunar Planet. Sci.* **XXXIII** abstract 1431

World Wide Web

Galileo mission site:
http://www.jpl.nasa.gov/galileo/images/europa/eurimages.html

John Spencer

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Europa Orbiter

Proposed mission to investigate Jupiter's moon Europa. Part of NASA's Outer Planets/Solar Probe Project. Possible launch in November 2003. May carry a radar sounder to measure the thickness of the surface ice and detect any underlying liquid ocean. Other instruments would reveal details of the satellite's surface and interior processes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

European Astronomical Society

The European Astronomical Society was founded in 1990 to 'contribute to and promote the advancement of astronomy, in its broadest sense, in Europe by all suitable means'. It counts over 1500 members and has 18 affiliated societies. It publishes a newsletter and organizes every year a Joint European and National Astronomical Meeting (JENAM).

For further information see http://www.iap.fr/eas/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

European Southern Observatory (ESO)

ESO is an intergovernmental European organization with its headquarters, scientific and technical divisions in Garching near Munich. It was founded in 1962 to establish and operate an astronomical observatory in the southern hemisphere and to promote and organize cooperation in European astronomical research. Its member states are Belgium, Denmark, France, Germany, Italy, The Netherlands, Sweden and Switzerland.

The original observatory at La Silla is located in one of the most arid areas in the world, the Chilean Atacama Desert, 600 km north of Santiago. Since the land was acquired in 1964, 14 optical telescopes have been built on La Silla mountain, 2400 m above sea level.

The most recent addition (in 1989) was the 3.5 m New Technology Telescope (NTT), one of the world's most advanced optical telescopes. The NTT operates with the SUSI camera and the multimode EMMI (visual) and SOFI (infrared) instruments.

The largest instrument at La Silla is a 3.6 m telescope equipped with the ADONIS adaptive optics unit and a high-resolution spectrograph. An infrared instrument and a high-accuracy radial velocity spectrograph are also planned.

Other facilities include the MPG/ESO 2.2 m wide-field instrument, the 1.52 m telescope with its new FEROS spectrograph, and the Danish 1.54 m telescope. The newly installed Swiss 1.2 m Leonard Euler telescope with the CORALIE echelle spectrograph is successfully 'hunting' for planets around other stars.

The 15 m Swedish–ESO submillimeter telescope (SEST) became operational in 1987. A new bolometric array will pave the way for the ATACAMA LARGE MILLIMETER ARRAY (ALMA).

ESO is also completing the Very Large Telescope (VLT) on the 2635 m high Mount Paranal, 130 km south of Antofagasta in Chile. The VLT will comprise four 8.2 m reflecting unit telescopes and several moving 1.8 m auxiliary telescopes. Their light beams will eventually be combined in the VLT interferometer (VLTI). When completed, the VLT will be the largest optical telescope in the world. On October 29, 2001, ANTU and MELIPAL, two of the four VLT 8.2 m Unit Telescopes were successfully linked. Light from the southern star Achernar (Alpha Eridani) was captured by the two telescopes and sent to a common focus in the observatory's Interferometric Laboratory. Following careful adjustments of the optical paths, interferometric fringes were soon recorded, proving that the beams from the two telescopes had been successfully combined in phase. From an analysis of the observed pattern (the fringe contrast), the angular diameter of Achernar was determined to be 1.9 milli-arcsec. At the star's distance (145 l.y.), this corresponds to a size of 13 million km. The observation is equivalent to measuring the size of a 4 m long car on the surface of the Moon.

For further information see http://www.eso.org/

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

European Space Agency (ESA)

An international organization whose task is 'to provide for and to promote, for exclusively peaceful purposes, cooperation among European states in space research and technology and their space applications'.

ESA has 14 member states: Austria, Belgium, Denmark, Finland, France, Germany, Italy, Ireland, The Netherlands, Norway, Spain, Sweden, Switzerland and the United Kingdom. Portugal became the fifteenth member in 2000. Canada takes part in some projects under a cooperation agreement.

The agency was born in 1975 from the merging of two existing organizations: ESRO (the European Space Research Organization) which developed satellites, and ELDO (the European Launcher Development Organization) which was responsible for development of a European launch vehicle.

ESA's activities cover space science, Earth observation, telecommunications, space segment technologies including orbital stations and platforms, ground infrastructures and space transportation systems, as well as basic research in microgravity.

ESA's headquarters are in Paris. Other ESA establishments include: the European Space Research and Technology Centre (ESTEC) in The Netherlands; the European Space Operations Centre (ESOC) in Germany; the European Space Research Institute (ESRIN) in Italy; the European Astronaut Centre (EAC) in Germany and a launch center at Kourou in French Guiana.

The science program, in which all member states participate, is one of the agency's mandatory activities. The program originated in the days of ESRO, which launched seven scientific satellites. ESA has played a leading role in an additional 13 scientific missions (including the Hubble Space Telescope) since the agency began operating on 31 May 1975.

The ESA science program is based on a long-term plan, known as Horizons 2000, which was adopted in 1985. The most expensive and ambitious missions, which are called Cornerstones, include the Solar and Heliospheric Observatory (SOHO) and Cluster II (combined as the Solar–Terrestrial Science Program), the X-ray Multi-Mirror (XMM) mission, the Rosetta comet mission and the Far Infrared Space Telescope (FIRST).

Four missions are under study for the next three Cornerstones: the BepiColombo Mercury orbiter, the LISA gravitational wave observatory, the GAIA astrometric interferometer and the IRSI Infrared Space Interferometer mission.

ESA has also launched a number of small and medium class missions, including Hipparcos, Ulysses, the Infrared Space Observatory (ISO) and the Huygens probe to Titan. In addition, the International Gamma-ray Laboratory (INTEGRAL) is scheduled for launch in 2001.

Medium missions have now been replaced by smaller, cheaper missions which offer faster development times and more frequent launch opportunities. These are divided into Flexible missions, which are purely scientific, and SMART (Small Missions for Advanced Research in Technology) missions, in which demonstration of new technologies takes precedence over science. The first example of a Flexible mission will be Mars Express, scheduled for launch in 2003. The main purpose of the SMART-1 mission to the Moon, due to launch in 2002, will be to demonstrate ion propulsion.

Also under study are MiniSTEP, a 2003 mission in collaboration with NASA, to test the equivalence principle, and an ESA contribution to the Next Generation Space Telescope.

For further information see
http://www.esa.int/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Evening Star

A name given to Venus when it is visible in the west after sunset. Ancient astronomers believed that morning and evening apparitions of Venus were of two different planets; the evening planet was given the name Hesperus. As the evening star, Venus is moving from superior conjunction (when it lies behind the Sun) to inferior conjunction (when it is between the Earth and the Sun), and is visible for longest when it reaches the position known as greatest elongation east, when its angular separation from the Sun is greatest. The name 'evening star' is sometimes given to evening apparitions of Mercury.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Evershed, John (1864–1956)

English astronomer, discovered the *Evershed effect*, the radial motion of the gases in sunspots (*see also* GEORGIO ABETTI).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Excitation and Ionization

Excitation: general properties

The excitation of a region reflects the physical conditions taking place in the astrophysical environment. The state of excitation is generally determined from the observation of spectral lines arising from different excited levels of an atom, ion or molecule. The analysis of the emitted spectrum allows the determination of the column densities[1] of the emitting levels, provided that the RADIATIVE TRANSFER of the emitted spectrum has been taken into account and that the fundamental atomic or molecular physics data are known. In some cases, absorption studies also allow the derivation of excitation conditions. This is currently performed in visible and UV absorption studies of a variety of atomic and molecular species towards diffuse and translucent clouds lying in front of a bright background star.

An *excitation temperature* can be defined from the ratio of the column densities of two excited states expressed by Boltzmann's law where the kinetic temperature is replaced by the excitation temperature $T_{\rm exc}$:

$$N(J)/N(J-1) = (g_J/g_{J-1})\exp(-\Delta E_{J\to J-1}/kT_{\rm exc})$$

where $N(J)$ is the column density of a level labelled by the J quantum number, g_J is the corresponding statistical weight, $\Delta E_{J\to J-1}$ the difference of the level energies and k is Boltzmann's constant. The ratio of the column densities is assessed to be the same as the ratio of the abundances. The physical significance of the temperature may be questioned but the value is useful as a clue to the excitation process.

The comprehensive time variation of the atomic, ionic or molecular level populations is expressed as the balance of excitation and deexcitation processes and can be put to zero at steady state if the timescales involved are shorter than the evolution time of the astrophysical region. The general set of equations is expressed as:

$$n_u\left(\sum_{k\neq u}(n\gamma_{u\to k}+R_{u\to k})+\sum_{k<u}A_{u\to k}\right) = \left(\sum_{u\neq k}n_k(n\gamma_{k\to u}+R_{k\to u})+\sum_{k>u}A_{k\to u}n_k\right).$$

$\gamma_{u\to k}$ stands for the collision rate coefficient in $\mathrm{cm^3\ s^{-1}}$, which induces the u to k transition due to a perturber of density n. For sake of simplicity, only one type of perturber has been written in the expression but several partners may be involved. In a hot ionized gas, electrons and protons are the most efficient perturbers, in atomic interstellar clouds, atomic hydrogen and helium are the exciting agents whereas in comets water molecules are the most abundant species. The excitation and deexcitation collision rate coefficients satisfy the so-called microscopic detailed balance equations:

$$g_u\gamma_{u\to k}\exp(-E_u/kT) = g_k\gamma_{k\to u}\exp(-E_k/kT).$$

The collision deexcitation rate coefficients depend generally smoothly on the temperature, and are often expressed as a power law of the temperature. $A_{u\to k}$ represents the spontaneous emission probabilities in $\mathrm{s^{-1}}$. The emitted lines can be of different kinds. Permitted lines are electric dipole transitions. Astrophysical environments are often favorable to forbidden line emission[2], because they are far from equilibrium. A famous example is the green nebular oxygen lines connecting the ground ^{3}P level to ^{1}D level which is an electric quadrupole transition or the green coronal Fe^{13+} line connecting the two fine structure levels of the ground ^{2}P state which is a magnetic dipole transition. $R_{k\to u}$ represents photon stimulated processes. Radiative excitation is described by $R_{k\to u}$, the absorption probability. Stimulated emission may also occur under specific conditions, for example when a population inversion is completed: maser emission is indeed observed in dense clouds associated with some embedded energy sources and is usually characterized both by variability and high polarization. Stimulated processes are proportional to the incident radiation field intensity, which depends on the astrophysical environment, and to the Einstein $B_{k\to u}$ coefficient, which is a pure atomic or molecular property.

If the radiation field involved in the radiative induced processes depends on the line itself, then on the level populations, the equations become nonlinear and the system has to be solved iteratively. This is the case in the envelopes of SUPERNOVAE as in SN 1987A or in molecular emission of abundant species such as CO in dense interstellar clouds.

At thermal equilibrium, the collisions dominate and the excitation temperature is equal to the real (kinetic) temperature. A convenient concept is the critical density which is the lower limit above which thermodynamical equilibrium is satisfied. It is defined as the ratio between the radiative deexcitation probability and the collisional deexcitation rate coefficient connecting two levels. The populations of the different levels are only dependent on the equilibrium between collisional excitation and deexcitation. However, the astrophysical environments are most often far from thermal equilibrium and the excitation follows then from detailed microscopic processes.

[1] The column density is defined by the integration of the density of the species along the line of sight. If the medium is homogeneous, it is simply given by the product of the density by the length of emitting region. Its unit is $\mathrm{cm^{-2}}$.

[2] A forbidden transition is characterized by a low emission probability since it arises from at least second-order perturbation terms in the matter–radiation interaction. In other words, forbidden lines result from *non*-electric dipole transitions, i.e. magnetic dipole or electric quadrupole.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sources of excitation

Collisional excitation

In GASEOUS NEBULAE, many common ions have low-lying energy levels with excitation potentials of order kT_e, where T_e is the electron temperature. Then the electrons in the extended tail of the Maxwell distribution corresponding to about 10 000 K have sufficient energy (≈2–4 eV) to excite the first spectral terms of O^+, O^{2+}, N^+, S^+. Once these levels are excited, they radiate spontaneously back to the ground state via forbidden transitions. The emissivities of the corresponding lines are simply $j_\lambda = N_u A_{u\to\ell} hc/\lambda$, where A is the Einstein coefficient of the transition taking place at the wavelength λ. Studying intensity ratios of lines corresponding either to close or to very different excitation energies, one will probe the density or temperature conditions. It is important to recognize that these transitions not only are diagnostics of physical conditions, but represent also the path through which the kinetic energy of the incident particles is converted into photon energy, depicting in that way a cooling process. In the cooler interstellar mostly atomic environments, the fine structure line of C^+, whose transition energy expressed in K is 92 K, is a major coolant of these regions at about 100 K. With a wavelength of 158 μm, in the far-infrared, this transition has been extensively studied by the Infrared Space Observatory (ISO) satellite launched in November 1996 by the European Space Agency (see INFRARED SPACE MISSIONS). Also detected by ISO are the electric quadrupole lines of molecular hydrogen observed in dense warm shocked regions at about 500–1000 K. In dense cold molecular clouds, rotational transitions of CO and its isotopes as well as other molecular species become valuable probes of the medium and the main coolants.

Radiative excitation

It has been recognized earlier that other excitation mechanisms are involved. This is the case of the fluorescence, so-called Bowen, lines in the visible observed in the spectra of high excitation PLANETARY NEBULAE since 1935. The important factor is the fluorescence efficiency defined as the ratio between the photon rate of fluorescent emission and the creation rate of pumping photons. In the equilibrium state, the rate of creation of pumping photons is equal to the total destruction rate including all loss processes. The incident pumping radiation is provided by the available He^+ Ly α photons whose wavelength at 30.3780 nm is almost coincident with a resonance line of O^{2+} at 30.3799 nm. The emission of vibrationally excited molecular hydrogen in REFLECTION NEBULAE is another manifestation of radiative excitation. The photons produced by nearby bright stars excite electronic transitions of molecular hydrogen which subsequently fluoresces and cascades inside the ground electronic state via vibrational infrared transitions. Finally, let us mention the excitation conditions derived from absorption studies of the CN molecule in diffuse and translucent clouds. It turns out that the rotational temperature of about 2.8 K reflects the excitation by the cosmic background radiation field with which the molecule is in quasi-equilibrium. This interpretation required more that twenty years.

Ionization: general properties

The IONIZATION of a medium is determined by the abundance of electrons. The number density of ions is then obtained by the requirement that the total charge is zero. The concomitant presence of electrons and ions could suggest that permanent electric fields be present. However, due to the inverse square law force applying between electric charges, the distance over which an appreciable electrostatic field can be created by a charged particle embedded in a plasma is determined by the Debye length L:

$$\frac{1}{L^2} = \frac{e^2}{\varepsilon_0 kT} \sum n_{i0} Z_i^2$$

usually very small compared to the cosmic scales involved. For fully ionized hydrogen $L = 69(T/n)^{1/2}$ in SI units. Electrostatic forces may however come in to play when large magnetic fields are also present over large distances.

The ionization of a medium is of crucial importance for its large-scale properties. For selected ranges of the ionization rate, several phases may coexist in pressure equilibrium. Each stable phase is determined by the onset of a new cooling mechanism or the decline of a heating mechanism, all of which are highly dependent on the relevant microscopic atomic and molecular processes.

Ionized species present in astrophysical regions involve a large variety of atomic positive ions, some with a very large nuclear charge due to extreme conditions of density and temperature. Molecular ions are also present in cold regions such as interstellar clouds and cometary environments. Negative ions are rarely considered due to their low stability against surrounding photons or collisions. The role of H^- is however undeniable and its presence in the atmospheres of cool stars is conspicuous from the continuum opacity arising in the photodetachment process. Large molecules such as polycyclic aromatic hydrocarbons (PAH) or hydrogenated amorphous carbons (HAC) which are invoked to explain the unidentified infrared emission lines at 3.3, 6.2 and 12 μm, may be somewhat negatively charged due to their large attachment probabilities. Interstellar dust particles may also be partially ionized; their charge is positive in the outer edges of interstellar clouds and may be negative in the central parts of molecular clouds.

In thermodynamic equilibrium, the ionic abundances are dependent on the temperature and the electronic density via the Saha equation. If we consider an atomic species A in two arbitrary states of ionization r and $r + 1$, then the number densities of the corresponding populations are related by the following expression: $n_{r+1} n_e / n_r = f_{r+1} f_e / f_r$. f_r, f_{r+1} and f_e are the partition functions of the ions of charge r and $r + 1$ and of the electrons, respectively. If one keeps only the ground-state contribution, labelled by a superscript 0, in the partition

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

functions of the ions, the Saha equation has the more familiar expression:

$$\frac{n(A^{r+1})n_e}{n(A^r)} = 2\frac{g^0_{r+1}}{g^0_r}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} \exp(-\Phi_r/kT).$$

g^0_r, g^0_{r+1} are the statistical weights of the ions in their ground level, and Φ_r is the ionization energy of the A^{r+} ion. This approximation is often used to calculate the sources of opacity of bound–bound and bound–free transitions of the available elements in the stellar interiors for their various stages of ionization.

Sources of ionization

Photoionization

In gaseous nebulae, ions result from the balance between photoionization of atoms followed by radiative recombination. Photoionization is induced by energetic photons emitted by central bright stars surrounded by mostly hydrogenic gas. The ejected electrons then recombine with the protons to give back atomic hydrogen. However, excited electronic levels are efficiently populated in the recombination process, which cascade subsequently by emitting radiative transitions in the radio, millimeter, visible and ultraviolet wavelength range. The beautiful blue or red colored shapes displayed in nebular observations are indeed coming from the Balmer α and Balmer β lines arising from recombination. Other minor elements are present in these regions showing up in emission spectra over the full range of the electromagnetic spectrum. In the radio range, emission occurs between Rydberg levels with large values of the principal quantum number ($n \approx 100$). The analysis of the recombination spectrum allows the determination of the electronic density and the temperature which is a few thousand K, much below the temperature corresponding to thermal ionization. In addition to discrete spectral lines, the interaction of the electrons with the ions gives rise to a weak continuous radiation which, in the radio and infrared regime, is dominated by free–free transitions due to bremsstrahlung. The corresponding photons delineate the distribution of the nebular gas and allow the derivation of the density of electrons via the so-called emission measure[3].

Photoionization is also the ionization agent in SUPERNOVAE REMNANTS where a compact galactic x-ray source illuminates the gas. Photoionization can then involve inner shell electrons and is often followed by the emission of one or more Auger electrons. These ionization processes are heating mechanisms of the gas whereas the subsequent emission of the gas is a cooling process. Other processes may contribute to the heating as, for example, direct Compton scattering on the electrons. The ionization and temperature of the gas may reach a stationary state depending on the gas pressure and the flux and spectrum of radiation.

[3] The emission measure is the integration along the line of sight of the square of the electron density. Its unit is usually cm^{-6} pc.

The ionization equilibrium can be modified by additional processes when they compete efficiently with one of the basic mechanisms mentioned above. Charge transfer with atomic hydrogen may drive the equilibrium towards ionization stages which photoionization models, including radiative and dielectronic recombination processes, cannot account for. This occurs in planetary nebulae for several low and intermediate excitation lines in C^{2+}, N^+, O^+, etc. This recognition has been the origin of a large number of theoretical studies on charge transfer collisions involving multiply charged ions.

Collisional ionization

In the absence of ionizing radiation, the local ionization state is mainly controlled by collisions of ions with thermal electrons as, for example, in the solar CORONA. Electron impact ionization is also taking place in the hot interstellar gas component which has been found thanks to O^{5+} absorption. The dominant processes are electron impact excitation and ionization followed by radiative recombination, dielectronic recombination and bremsstrahlung. Dielectronic recombination occurs at specific resonant energies corresponding to weakly bound autoionizing states leading to a much larger rate coefficient so that dielectronic recombination dominates radiative recombination in high temperature plasmas by factors of about a hundred. If ionization comes only from thermal electrons, the master equations describing the variation with time of the fraction f_r of element X in each ionization stage involve the electron density in front of each term. So, at steady state where $\mathrm{d}/\mathrm{d}t = 0$, the electron density is removed so that the ionization balance of the ion fractions is only dependent on temperature. We should point out that this situation cannot be described by a state of thermodynamic equilibrium since it does not result from detailed balance of time-reverse processes (e.g. photoionization versus radiative recombination, electron impact versus three-body recombination). In dense molecular clouds, the ionization results from the interaction of the gas with COSMIC RAYS, which are energetic particles such as protons and alpha particles probably produced during the explosion of supernovae. The ionization fraction of the medium is presumably small and cannot be determined directly since all ions present are not observable. Then, the ionization is deduced from millimeter observations of molecular ions (HCO^+) and interpreted from chemical models. The OH radical is another tracer of cosmic rays in cold interstellar clouds since the first step of its chemistry involves the presence of protons or H_3^+ ions. The inferred cosmic ionization rate is of the order of 10^{-17} s^{-1}.

Evelyn Roueff

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Exit Pupil

The image of the objective formed by the eyepiece of a telescope. All of the rays collected by the objective pass through the exit pupil, which is where the pencil of rays emerging from the eyepiece has its smallest diameter and the illumination of the image is greatest. The exit pupil, therefore, is the best place at which to place the pupil of the eye when observing with a telescope.

The diameter of the exit pupil (D_e) is equal to the aperture of the telescope (D) divided by the magnification (M). For example, a telescope with an aperture of 200 mm and a magnifying power of 50 will produce an exit pupil with diameter $D_e = D/M = 200/50 = 4$ mm. The diameter of the exit pupil must be less than or equal to the diameter of the pupil of the eye if all the light collected by the telescope is to enter the observer's eye. To meet this condition, the magnification must be greater than or equal to D/D_p, where D_p is the diameter of the pupil of the eye. Since the diameter of the dark-adapted pupil is about 7 mm, the minimum magnification that will ensure that exit pupil is smaller than the observer's pupil is approximately $D/7$, where D is the aperture expressed in mm ($D/0.007$ if D is in meters).

See also: aperture, eye relief, magnification, objective lens.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Exobiology

Exobiology (=bioastronomy, astrobiology) is the study of life in the universe. More precisely, it is the study of the origins, evolution and distribution of life in the universe and of the structures and processes associated with life itself (see also LIFE: ORIGINS AND POSSIBLE DISTRIBUTION IN THE UNIVERSE).

Historical background

From myths to the emergence of a new scientific field

Is there life elsewhere? This is certainly one of the oldest question asked by humankind. Mentions of it have been found, for instance, in Epicurus' 'Letter to Herodotus', 300 BC. Since antiquity, all over the world, many authors have thought about, discussed and written their feelings about this fundamental problem. In the 13th century, the Chinese philosopher Teng Mu wrote 'how unreasonable it would be to assume that apart from the Earth and the sky that we see, there are no other skies, nor other Earths'. Some have perished because of this idea. GIORDANO BRUNO (1548–1600) was burned by the Roman Catholic Church, because, as he wrote in his 'Del' infinito universo e mondi', he thought that 'There are countless suns and countless earths orbiting around their sun, exactly as our seven [sic] planets orbit around our Sun. Living beings inhabit those worlds.' GALILEO GALILEI (1564–1642) was promoting analogous ideas but, fortunately, he just escaped a similar destiny, by abjuring his 'mistakes and heresies' in front of the Court of the Supreme and Universal Inquisition. Later on, the question of life elsewhere was considered by many writers, philosophers and even scientists such as Kepler, Fontenelle, Kant, Goethe and Huygens. FLAMMARION, at the end of the 19th century was probably one of the first to ask 'What to do to bring evidence of these other worlds?'. The long story of the detection of so-called canals on the Martian surface brought a false answer to the question of life on Mars until the beginning of the 20th century. Then the development of more and more powerful tools for exploring other worlds, by remote sensing, and more recently by *in situ* techniques, opened a large avenue in this area and marked the true emergence of this new scientific domain of exobiology (see also LIFE ON OTHER WORLDS).

A new field is born

When NASA started the Apollo program, with the project of sending men to the Moon and bringing lunar samples to the Earth, the question of the possible presence of (indigenous) life on our natural satellite arose logically. The problem was in fact double : if there were life forms on the Moon, it was essential to avoid any biological contamination of our satellite by terrestrial organisms, and—even more important—of the Earth by hypothetical lunar organisms present in the lunar samples carried back to our planet. Thus, in the 1960s, planetologists and microbiologists alike started to work together to tackle this question. This was the time when the recipient of the Nobel Prize Award (1958) in Medicine and Physiology,

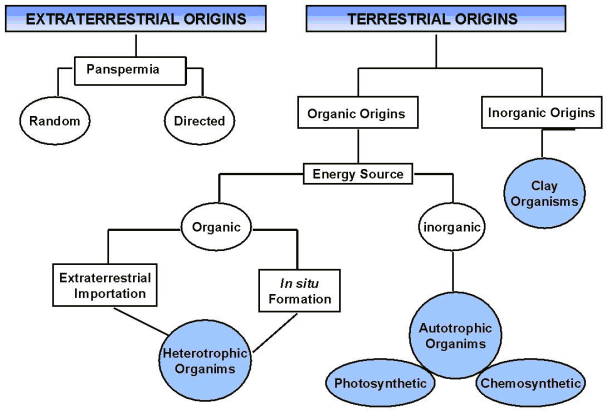


Figure 1. Diagram of theories of the origin of Life on Earth (from W L Davis and C P McKay).

Joshua Lederberg, invented the word 'Exobiology'. It then became a fully recognized scientific field, spanning many classical domains, including astrophysics, astronomy, planetology, physics, chemistry, biochemistry and biology, as well as human sciences and even mathematics.

The various approaches

Indeed, the approaches abound. The more logical and direct way is to search for life on extraterrestrial planetary objects. This can be done by remote sensing measurements or even, with the availability of space technologies, by *in situ* exploration. It can also be carried out by studies of extraterrestrial samples transferred to the Earth through natural processes (meteorites, micrometeorites, cosmic dust particles and cometary materials) or collected in space by robots or manned missions. Remote sensing techniques can even be applied to search for biological activity on extrasolar planets, when techniques become powerful enough to detect Earth-like objects and the presence of molecular oxygen and ozone in large amounts in their atmosphere (assuming that this is an unambiguous signature of life). The other approach is to assume that what happened on Earth is a common event in the universe and that there are numerous places in our Galaxy where life arose and evolved toward intelligent and technologically developed living systems. Then, by listening to the universe in the right direction and at the right frequency (presumably radiowaves) and at the right time, one should be able to detect signals from extraterrestrial civilizations. This is the SETI (search for extraterrestrial intelligence) approach mainly based on radio observations. There is a less direct approach but fully logical: the study of the terrestrial example, and, first of all, of the origin of life on Earth. By putting together the pieces of the puzzle of the origins of the first living system on our planet, one should be able to put some constraints in the conditions necessary for the emergence of life, in general, on a planetary body. Furthermore, the variety,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

distribution and limits of life on Earth would yield the—so far—unique example of distribution and evolution of life in the universe.

The terrestrial example

What is life ?

The only example we have—the terrestrial one—allows us to characterize life through a set of properties the simultaneous presence of which is specific to living systems. The most important are the properties of autoreproduction, of evolution and of mutation with a very high level of information included in the structures of the living systems, and with the property of autoregulation against the constraints induced by its environment. The terrestrial example shows that, starting from a simple unicellular living system, our last universal common ancestor (LUCA), a primitive prokaryotic cell, biological evolution has been able to induce a fantastic biological diversity. In spite of the many differences between the incredibly wide variety of species which live on Earth today, life on Earth shows a large unity. In fact these differences are seen mainly at the macroscopic level. As soon as we look at living systems with a microscope, we can observe that all are made of cells. From one in the unicellular, to billions in the pluricellular, the cell can be considered as the structural unit of life. Eukaryotes have cells with a clearly differentiated nucleus and prokaryotes are cells without a nucleus, but all cells have roughly the same structure and include the same kind of macromolecules, isolated from the outside by a membrane made of lipids. This selective frontier allows, through complex chemical mechanisms, the selection and concentration of all molecules necessary to the building of the cell, to its development and to its replication. The cell is the microscopic chemical plant for life , the place where all the life specific, chemical processes, are in progress, involving molecules also specific to life.

The building blocks

The unity of life is even clearer at molecular level. All living systems on Earth use the same type of molecules and, first of all, macromolecules: nucleic acids and proteins. In all (terrestrial) living systems cellular multiplication is programmed from the chromosomes, made of nucleic acids. These macromolecules include the genetic information which makes one living system different from another. The message is stored in deoxyribonucleic acid (DNA). It is transcribed and transferred to the ribosome (the cellular plant) where it is translated and converted, with the help of the genetic code, allowing the synthesis of proteins, the other macromolecules fundamental to life. Proteins play a crucial role in almost all biological processes. They can act as a very specific and powerful catalyst, with the protein enzymes, as a transport and storage molecule (e.g. hemoglobin transporting oxygen), or as a mechanical support, such as collagen, etc.

Nucleic acids are made of a very large number of units, called nucleotides, constituted of a five-carbon sugar, ribose in RNA or deoxyribose in DNA, linked to a phosphate molecule and to a one- or two-ring organic molecule called a nucleotide base (such as adenine). The smallest DNA molecule uses several thousand nucleotides and the largest one several billions. However, all use only four different nucleotides, only differing by their nucleotide base. Similarly, all RNAs on Earth are made of only four different nucleotides, using the same kinds of bases as DNA. From this alphabet of only four letters, nature can make a very large number of words, by using many letters and by changing their sequence from one to another nucleic acid.

The primary structure of proteins is based on the same principle, although their building blocks—called amino acids—are simpler than nucleotides. Among the very large number of different amino acids, proteins use only 20. This alphabet of 20 letters provides, again, a fantastically rich vocabulary, based on long words using a large variety of sequences. Furthermore, these molecules (with the exception of the simplest amino acid, glycine) possess an asymmetric carbon and consequently can exist under two different configurations, called enantiomers, and labelled 'L' and 'D', symmetric but not superimposable (this property is called 'chirality', and the molecules are said to be 'chiral'). Only one is used by life: the 'L' configuration for amino acids and the 'D' for sugars, and consequently for nucleotides.

To understand the origin of life it is essential to understand the origin of proteins and nucleic acids but also the origin of chirality in the living systems.

The origins of life

The study of the origins of life on Earth has probably been one of the most successful approaches in exobiology, even before the emergence of the field. Historically, the problem of the origin of life had for centuries almost only one answer: spontaneous generation. Although debated and discarded at least for macroscopic life, in the middle of the 19th century many still though that life could emerge spontaneously and suddenly from inert matter. With his—now historical—experiment on spontaneous generation Pasteur demonstrated in the 1860s that the observation of the spontaneous creation of microorganisms from non-living matter is only a false interpretation of observational data and can be explained by the biological contamination of the samples via the microorganisms present in the ambient air. Then, for more than 50 yr, the theory of the origin of life was mainly based on the concept of panspermia. It assumes that living germs are present everywhere in the universe and can travel in the cosmos, simply pushed by the pressure of radiation released by stars. They would have reached the solar system and seeded the Earth. The panspermia theory is not fully given up today; however, the generally accepted theory is based now on a thoroughly different concept of 'chemical evolution'. Already emerging from several writings of the 19th century, in particular in Darwin's letters, the concept of chemical evolution was first clearly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

introduced and described by the Soviet biochemist Oparin, in 1924. The main idea is to assume that life arose on the Earth after a long evolution of organic (containing chemically reduced carbon, i.e. carbon in a non-oxidized form) but inert materials preceding Darwin's biological evolution and allowing, by natural chemical processes of a now so-called 'pre-biotic chemistry', the formation of structures of increasing complexity, up to the first replicating system. For Oparin, our ancestor would have been a 'heterotrophic' (living system unable to synthesize itself the organic molecules it needs) microorganism, fed by the many organic compounds present on the primitive Earth, produced in the atmosphere and in the primordial oceans by abiotic (without the involvement of life) processes. This vision has now markedly changed and it seems more likely that the first living system was an 'autotroph' (living system able to synthesize itself the organic molecules it needs), feeding from the inorganic molecules, in particular carbon dioxide, present in its environment. In addition several scenarios are invoked to explain the origin of the simple organics which are involved in this prebiotic chemistry. However, the general concept of a chemical evolution from simple to complex organic compounds, and to the first bio-macromolecules capable of replication, remains the same.

Table 1. Some key molecules of prebiotic chemistry.

Name	Formula	Prebiotic role
Hydrogen cyanide	HCN	Source of amino acids, of purine and pyrimidine bases
Cyanoacetylene	HC_3N	Source of pyrimidine bases
Cyanogen	NC-CN	Condensing agent
Liquid water	H_2O	Fundamental solvent
Formaldehyde	HCHO	Source of sugars
Hydrogene sulfide	H_2S	Source of S-containing compounds
Phosphate	PO_4^{3-}	Source of P-containing compounds

The first experimental evidence of this theory was provided in 1953 by the publication of the now well-known experiment by Stanley Miller. By simulating in a reactor the evolution of a model of the primordial atmosphere of the Earth under energy deposition, in the presence of liquid water, Miller was able to obtain amino acids. This also opened the new field of prebiotic chemistry. Since 1953, hundreds of similar experiments have been carried out, using various conditions (energy sources, composition of the gas mixture, etc). They show that the production of organic molecules, including amino acids and nucleotide bases, can be obtained, but only if the starting atmosphere is chemically reducing. They also show the importance of several simple organic molecules, such as formaldehyde, HCHO, hydrogen cyanide, HCN, cyanoacetylene, HC_3N and other nitriles, because of their chemistry in liquid water. Now, although we have no direct information on the conditions on the primitive Earth, the current models are in favor of an oxidizing atmosphere. This drastically reduces the importance of atmospheric processes in the formation of organics on the primitive Earth. However, other sources can be envisaged. The deep-sea hydrothermal vents, which include many chemical ingredients able to induce prebiotic syntheses, are now considered to have been a possible place on the primitive Earth where chemical evolution to life started. Another possibility is to assume that a large part of the starting carbonaceous ingredients necessary to prebiotic chemistry was imported from space. Indeed, the Earth is continuously bombarded by extraterrestrial materials (meteorites, micrometeorites, interplanetary dusts and cometary matter) a noticeable fraction of which includes organics. It is likely that these bombardments were even more important on the primitive Earth and thus may have been a source of organics.

Whatever the scenario is, it is well established today that life was largely present on the Earth 3.5 billion years ago, from the age of the oldest stromatolites (small geological structures resulting from the interaction of microorganisms with their environments). They also show that this was already a complex life, capable of photosynthesis. It is likely that the first living systems were simpler. The recent discovery of the catalytic properties of RNA and of an 'RNA World' suggests that these biomacromolecules may have been the first replicating system. However, many still doubt, because of the experimental difficulties in synthesizing prebiotically these molecules, and even their building blocks (the nucleotides). On the contrary, amino acids and their polymers are much easier to obtain. The question is still open.

If there is life elsewhere, there is no reason to assume that it must have the same appearance as life we know on Earth, because of the variety and diversity of terrestrial life at the morphological level. However, at the molecular level, the terrestrial example may be a more general one. The importance of liquid water and of the chemistry of carbon in the prebiotic terrestrial processes is probably not restricted to the Earth: water is an abundant molecule and carbon an abundant element in the universe. The discovery of an active organic chemistry in the interstellar clouds clearly supports the idea that organic chemistry is universal. It is thus reasonable to assume that extraterrestrial life is also based on carbon chemistry and liquid water, even if the hypothetical extraterrestrial biochemistry may not necessarily be identical to the one we know.

Exobiology in the solar system

Many objects of the solar system are of exobiological interest: not only Earth, but also comets, Mars, Europa, and even the giant planets and some of their satellites, such as Titan and Triton, because of the very complex organic chemistry which is involved in their environments.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Characteristics of exobiologically important planets and satellites of the solar system.

Planet or satellite	Density ($g\ cm^{-3}$)	Distance to the sun (AU[a])	Mean surface temperature (°C)	Surface pressure (bar)	Atmosphere	
					Main composition	Organic compounds
Earth	4	1	22	1	N_2, O_2	Many
Mars	3.7	1.5	−50 to −70	0.007	CO_2, N_2	None
Jupiter		5	No surface tropopause: −255		H_2, He, CH_4 CH_4	Several hydrocarbons
Europa	3	5	−150 to −180		Negligible	No but possibility of subsurface liquid water
Saturn		10	No surface tropopause: −190		H_2, He, CH_4	Several hydrocarbons
Titan	1.9	10	−180	1.5	N_2, CH_4	Several hydrocarbons and nitriles
Neptune			No surface tropopause: −220		H_2, He, CH_4	Several hydrocarbons, HCN
Triton			−235	14×10^{-6}	N_2	Methane

[a] 1 UA $\approx 150 \times 10^6$ km.

Meteorites, micrometeorites and comets

Until 1969, the only extraterrestrial samples available in our laboratory were METEORITES, in particular carbonaceous CHONDRITES. The detailed study of meteorites has not revealed the presence of extraterrestrial microorganisms in those objects. The recent story of the Martian meteorite ALH 84001 does not undermine this statement. However, it is now fully established from careful chemical analysis of carbonaceous chondrites that this class of meteorites is rich in organic compounds. The several per cents of carbonaceous matter includes molecules of biological interest, such as amino acids and even nucleotide bases. Furthermore, the recent re-analyses of some of these meteorites shows the presence of an excess of L enantiomers in the amino acid content of extraterrestrial origin. This should have a strong input on our understanding of the origin of chirality in the terrestrial living systems. COMETS may have also played a major role in the origin of life on Earth, by importing water and organics on primitive Earth. In any case, they are important objects for exobiology because of the cometary organic chemistry. Although the molecular composition of the cometary nucleus is still not known today, comets are assumed to be very rich in organics, made of C,H,O and N atoms. The future cometary missions, in particular the ESA Rosetta mission with *in situ* analysis of the cometary nucleus, should provide in the first decade of 2000 essential information on the cometary nucleus.

Mars and Europa

Thus far, two planetary objects seem to be the most promising for a direct search for extraterrestrial life in the solar system. The first is the planet MARS and the second is EUROPA, one of the Galilean satellites of Jupiter.

In spite of its dimensions (about half of the Earth) and the absence—now—of a dense atmosphere and of liquid water, Mars is still considered as the best target. VIKING has not been able to detect the presence of microorganisms (autotrophs or heterotophs, see the definitions above) in the Martian soil. However, it has shown that the red planet was probably covered by liquid water and had a dense atmosphere at the time life had already emerged on Earth. Since, at that time, the Earth and Mars were very similar, it is logical to assume that what occurred on the Earth also occurred on Mars. Then both planets have evolved differently. Because of a tectonic activity too light to recycle carbon dioxide (essential for keeping the surface warm by the greenhouse effect), Mars lost most of its atmosphere and became colder and colder. It also lost its liquid surface water, which may now be trapped as a solid 'permafrost' in the Martian soil. Also, under such conditions, the atmosphere became strongly oxidizing; the oxidants, such as hydrogen peroxide, diffused in the soil, and all Martian life gradually disappeared from this hostile

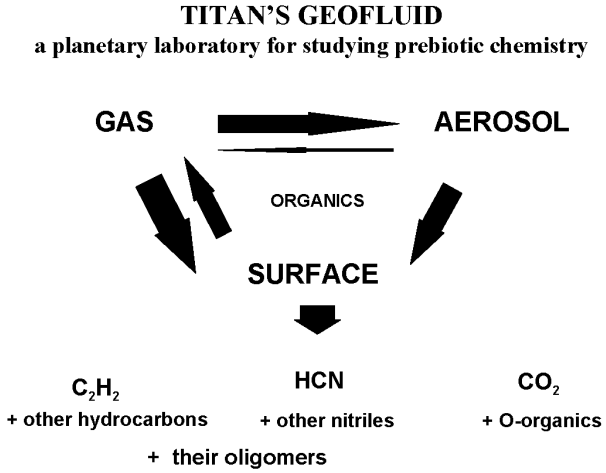


Figure 2. The several couplings in Titan's geofluid.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

environment, at least on and near its surface. Although the presence of life as extant on Mars is not fully excluded, future missions to Mars will have to embark scientific instrumentation to search for traces of extinct life in the Martian soil, and deep enough (at least a few meters deep) to reach depths where the oxidation processes are held to be negligible. Such missions are in preparation now.

The interest of exobiologists in Europa is more recent. VOYAGER data of the outer solar system have provided enough information to build models of the internal structure of many of the giant planets and their satellites. Among the three main models of Europa, one assumes that the icy surface of the Jovian satellite corresponds to a water ice crust several kilometers deep, covering an ocean of liquid water several tens of kilometers down. If such an ocean has hydrothermal activity, the presence of living organisms, prokaryotic microorganisms and even macroscopic eukaryotes, cannot be ruled out. However, the detection of such living systems will not be an easy task.

Titan

Another planetary object was examined by the Voyager mission and became, subsequently of prime interest for exobiology: TITAN. Voyager demonstrated that the largest satellite of Saturn, second by size in the solar system, has a dense atmosphere, mainly made of molecular nitrogen (more than 90%), with a noticeable concentration of methane (several per cent). Although Titan is much colder than the Earth, its atmospheric thermal profile is very similar to the terrestrial one, with a troposphere (95–74 K), a tropopause (74 K) and a stratosphere (74–170 K), owing to the presence of greenhouse gases (mainly methane and hydrogen) and antigreenhouse components (clouds and aerosols). Such a main chemical atmospheric composition is very favorable to organic syntheses. From laboratory experiments and theoretical modeling, one can expect many organics in Titan's atmosphere, mainly hydrocarbons and nitriles. Indeed, several organic compounds have already been detected: they all belong to the two classes of organics. Furthermore, the most important nitriles of terrestrial prebiotic chemistry—hydrogen cyanide, cyanogen and cyanoacetylene—have been detected in Titan's atmosphere. This environment is also very rich in aerosols and, probably, in clouds. Thus, it is also the only known satellite the surface of which is not visible from outside. Although the chemical nature of these aerosols has not been ascertained directly so far, it seems likely that they are mainly organic, with a nucleus of macromolecular nature. Furthermore, models of the surface and its interaction with the atmosphere strongly suggest that it is partly covered by seas and lakes of liquid methane and ethane (a necessary source and sink, respectively, of these two compounds on Titan).

With the exception of liquid water, absent on this too cold a planetary object, because of its environment, very rich in organics, in the gas and aerosol phases, and the possible presence of liquid bodies on its surface, Titan shows many similarities with primitive Earth. Studying Titan now should provide much key information on the possible couplings in the organic-related processes between the different parts of a planetary geofluid. Furthermore, study of Titan's chemistry, owing to the absence of liquid water, could also provide crucial data on the role played by liquid water in exobiology.

The Cassini–Huygens mission launched in October 1997 will reach the Saturn system in 2004 (see CASSINI MISSION). Then the Cassini NASA spacecraft will become an artificial satellite of Saturn and will release the Huygens ESA probe which will enter Titan's atmosphere. During 2.5 h of descent, the six scientific instruments aboard the probe will perform a detailed physical and chemical analysis of Titan's atmosphere (gas and aerosols) and surface. During the nominal duration of the mission (until 2008), the 12 scientific instruments of the Cassini orbiter will also provide essential information of exobiological importance on Titan's environment.

Exobiology outside the solar system

The interstellar medium

Organic chemistry is specific neither to the Earth nor to the solar system. It seems to be widely present in the whole universe. One hundred and ten different molecules have already been detected in the interstellar medium: a large majority of those (87) are organic (and only seven include silicon atoms). A complex organic chemistry probably occurs in interstellar clouds, in the gas and in the solid phase of the interstellar dust. Such chemistry may have also been involved in the molecular cloud from which the presolar nebula, and then the solar system, arose. Studying this chemistry and its evolution during the accretion phase of the presolar and the solar system is of great exobiological importance.

Extra-solar planets

Since the first—now confirmed—detection of a planet around the star 51 Peg by the Swiss astronomers Mayor and Queloz in 1996, more than a 20 EXOPLANETS have been discovered around close-by stars. Because of our current technologies, such discoveries are limited to Jupiter-size planets. However, they already show a wide variety of planetary objects in the universe, since the detected giant planets are very different from the solar system's ones, being much closer to their star and much warmer. There may be Titan-like satellites orbiting around these giant exoplanets having dense atmospheres, together, contrary to Titan, with liquid water. These planetary systems may also have Earth-like planets, which should be detectable with our future technologies. We should even be able to detect, by remote sensing, signatures of life on these hypothetical planets. Although this awaits confirmation, the presence of molecular oxygen in large abundance, and hence of ozone, in a planetary atmosphere may turn out to be a fair indication. Detection of intense infrared bands of ozone on an exoplanet should strongly suggest the presence of life on it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 3. The main 'SETI' experiments carried out up to now (from J Tarter).

Year of the experiment	Observer(s)	Site	Target	Duration (h)
1960	Drake ('Ozma')	NRAO	2 stars	150
1964–1965	Kardashev and Sholomitskii	Crimea	2 quasars	80
1970–1972	Slysh, Pashchenko, Rudnitskii and Lekht	Nancay	5 OH masers	2
1979–1982	'Serendip' Team	California	Random	400
1981–1988	Biraud and Tarter	Nancay	343 stars	∼600
1990–	Colomb 'Meta II'	Argentina	Whole sky visible from Argentina	Permanent
1992–1993	'HRMS' NASA Team	Arecibo	25 stars	200
1992–	Bowyer, Werthimer and Donnelly ('Serendip III')	Arecibo	30% of the sky visible from Arecibo	50
1995	Phoenix Project	Australia	311 stars	2600

SETI

Another approach assumes that life has evolved, Earth like, toward intelligence and technologically advanced civilizations: we thus have to search for signals from such hypothetical extraterrestrial intelligence (SETI). As proposed by Cocconi and Morrison as early as in 1959, the radiowave domain seems to be the most appropriate for interstellar communications. Furthermore, there may be 'magic' wavelengths in the universe, corresponding to emission lines of very abundant compounds, such as the specific radio line of hydrogen at 21 cm. The first SETI experiment was carried out by Frank Drake in 1960 (see SETI INSTITUTE). With his project 'Ozma', he listened during several weeks to two nearby stars (Epsilon Eridani and Tau Ceti) at the magic wavelength of 21 cm. Since then, several tens of similar experiments have been performed, including recently, with a dedicated detector of several millions of channels. No clearly identified signal of extraterrestrial intelligence has been detected (yet). However, only a very small fraction of our universe has been explored thus far. Although we have not yet received any clear message from an extraterrestrial civilization, we have sent already several outward. The Pioneer X and XI spacecraft, launched in the early 1970s, which now travel outside the solar system, carry a metallic frame engraved with information on the solar system, the Earth and terrestrial life. Similarly, Voyager 1 and 2, launched in 1977, carry a videodisk about our planet. A message has also been sent from the Arecibo radiotelescope toward Messier 31. Of course, all these actions are only symbolic, but they clearly show that, today, the idea that there may be life elsewhere, including intelligent life, is largely accepted by the scientific community as well as the political community.

Current studies and prospects

Exobiology is present in every field and almost all exobiological approaches are effectively followed up to this day. A wide scientific community is working in the field of life origins, as demonstrated by the size (about 500 members) of the International Society for the Study of the Origin of Life (ISSOL). In this field, the likely future privileged studies will probably be the search for replicating molecules simpler than RNA, prebiotic syntheses in the submarine hydrothermal vents (through theoretical modeling as well as laboratory experiments) and studies on prebiotic processes from micrometeorites. The 2000 millennium will see a large increase of the exobiological explorations of the solar system. Many missions to Mars are already planned in the 2000–2010 period, with dedicated instrumentation to search for past and present life on the red planet. This period should also lead to important discoveries about comets, with the Rosetta missions, and about Titan, with Cassini–Huygens. We can even expect a mission to Europa, with a search for life in its hypothetical oceans. In the 2010–2020 period, if not before, one could expect, with a technology similar to the one planned in the IRSI (ex-Darwin) mission, the detection of Earth-like extrasolar planets and of extraterrestrial biological activity through the discovery of ozone. Whatever these discoveries, it is clear that the new millennium should drastically encourage multidisciplinarity. Exobiology, as one of the largest multidisciplinary fields, should thus, placed in the foreground, be the field of the future.

Bibliography

Brack A (ed) 1998 *The Molecular Origins of Life: Assembling Pieces of the Puzzle* (Cambridge: Cambridge University Press)

Bylinsky G 1981 *Life in Darwin's Universe* (New York: Doubleday)

Carle G, Schwartz D and Huntington J (eds) 1992 *Exobiology in Solar System Exploration (NASA SP 512)*

Chela-Flores J and Raulin F (eds) 1998 *Exobiology: Matter, Energy, and Information in the Origin and Evolution of Life* (Dordrecht: Kluwer)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cosmovici C B, Bowyer S and Werthimer D (eds) 1997 *Astronomical and Biochemical Origins and Search for Life in the Universe (IAU Colloquium No 161)* (Bologna: Editrice Compositori)

Doyle L R (ed) 1996 *Circumstellar Habitable Zones* (Travis House)

Hornek G 1995 Exobiology, the study of the origin, evolution and distribution of life within the context of cosmic evolution: a review *Planet. Space Sci.* **43** (1/2) 189–217

Horowitz N H 1986 *To Utopia and Back, the Search for Life in the Solar System* (New York: Freeman)

Jakosby B 1998 *The Search for Life on Other Planets* (Cambridge: Cambridge University Press)

François Raulin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Exosat (European X-ray Observatory Satellite)

European Space Agency x-ray observatory launched in 1983. Designed to locate and identify cosmic x-ray sources. A highly elliptical 356 km × 191 581 km Earth orbit allowed continuous observations for up to 72 hours. Payload consisted of three instruments to study the x-ray spectrum between 0.04 and 50 keV. Suffered a number of equipment failures but made important studies of binary stars and galaxies, including discovery of several new transient x-ray sources. Operations ended April 1986.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Expanding Universe

A universe that expands with time. Although the possibility had been raised earlier through theoretical work carried out by Willem de Sitter (1872–1934), Aleksandr Friedmann (1888–1925), and the Abbé Georges Lemaître (1894–1966), that our universe is expanding was first demonstrated observationally in 1929 by Edwin P Hubble (1889–1953), through his measurements of the redshifts in the spectra of galaxies.

In our expanding universe, each galaxy (or cluster of galaxies) is receding from every other one, the speed of recession of one galaxy from another depending on the separation between them. If motions induced by local concentrations of mass are ignored, the galaxies and clusters behave as if they were at rest in an expanding space. A useful analogy is to consider the whole of three-dimensional space to be represented by the (two-dimensional) surface of a balloon and to represent galaxies and cluster by spots stuck to this surface. As the balloon expands, each 'galaxy' moves away from every other one but there is no unique center to this expansion (each dot is equivalent to every other one, and the *surface* of a balloon does not have a center).

See also: Big Bang theory, cosmological model, cosmology, Hubble law, redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Explorer

Series of small NASA experimental, scientific and research satellites. Explorer 1, launched in 1959, discovered Earth's van Allen radiation belts. They included the Interplanetary Monitoring Platforms (10 satellites which investigated the Earth–Moon magnetic environment over a complete solar cycle), the Radio Astronomy Explorers (Explorers 38 and 49) and the Small Astronomy Satellites (Explorers 42, 48 and 53). In March 1999, Explorers 50 (IMP 8) and 67–74 (EUVE, SAMPEX, RXTE, FAST, ACE, SNOE, TRACE and SWAS) were still operating. Of the 72 successful Explorer missions, five operated for 10 years or more. They include the ISEE 3/ICE spacecraft (14 years), the IUE spacecraft (19 years), and IMP 8 (still operational after 25 years). (See also INTERNATIONAL SUN–EARTH EXPLORER, DYNAMICS EXPLORER, UHURU.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extended Atmospheres

During a total ECLIPSE of the Sun, when the Moon blocks out the intensely bright disk of the Sun, a faint white 'halo' can be seen surrounding the Sun (see figure 1). This halo, officially called the Sun's 'CORONA' (i.e. 'crown'), exists mainly because there are electrons in the outer solar atmosphere: even though the disk of the Sun is hidden from direct view, the electrons scatter a fraction of the light from the disk towards our eyes. The fraction which is scattered towards us is small, typically no more than about 1 part per million of all the photons which emerge from the Sun. (Farther out from the Sun, there is also some scattering from dust particles.) The fraction of 1 in a million has to be compared with the sky brightness near the Sun due to the molecules and dust in our own atmosphere: at the best mountain observing sites on Earth, the intensity of pure blue sky may be low enough that the innermost corona is just visible to the naked eye. More typically, the brightness of clear sky near sea level is a thousand times brighter than the corona. This explains why the corona is too faint to be seen by the eye outside of eclipse. But when the corona *can* be seen, it may be quite irregular in shape, extending in certain directions much more than in others. The minimum extent of the corona is over the North and South poles of the Sun, especially when the Sun is close to the minimum level of SUNSPOT activity. The greatest extent of the corona occurs in features called 'streamers', which can be seen to distances of as much as a solar radius or more beyond what is usually thought of as the 'edge' of the Sun.

Also visible for a short period of time during certain phases of a total eclipse is a rose-colored aureole of light close to the Sun (see figure 2). This colored aureole, known as the 'CHROMOSPHERE' (i.e. 'color sphere'), also extends out beyond the usual 'edge' of the Sun, although not as much as the corona.

Together, the chromosphere and the corona constitute what may be called the 'extended atmosphere' of the Sun. It is now known that such extended atmospheres also exist around other stars.

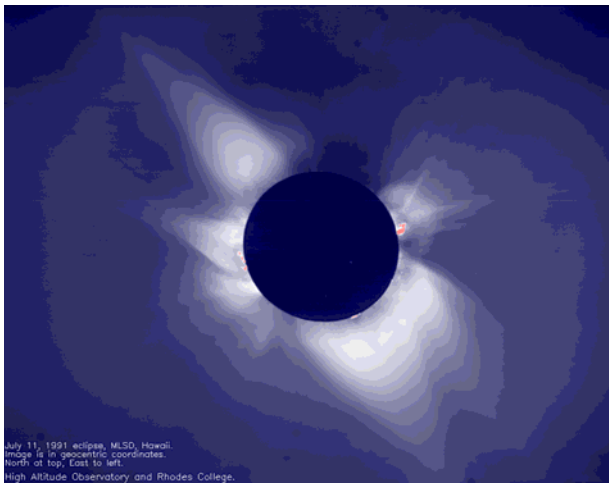


Figure 1. The solar corona: an irregular halo of light which can be seen by the naked eye during an eclipse of the Sun. The image was obtained by observers from the High Altitude Observatory and Rhodes College on 11 July 1991.

Figure 2. The solar chromosphere: a narrow band of reddish light which can be seen briefly by the naked eye close to the edge of the eclipsed Sun.

The non-extended case

In order to appreciate the significance of *extended* atmospheres in the Sun and other stars, it is instructive first to discuss atmospheres which can be referred to as 'non-extended'. This discussion will serve to highlight the conditions which must be satisfied in order that a star may develop an atmosphere which can be truly be referred to as 'extended'.

Two objects which are prominent features in the sky appear to have sharp edges: the Moon and the Sun. This seems like a trivial observation, but it tells us something informative. As far as the Moon is concerned, it is hardly surprising that we see a sharp edge at all times: after all, the Moon is made of solid material, solid enough that twelve men have walked on its surface. When we look at the edge of the Moon, we are viewing a region where the solid surface stops, and the vacuum of space begins. Since this is a discontinuous transition, it is no wonder that the edge appears sharp to us. But the Sun does not have a solid surface: it is composed of gas, which tends to diffuse out in all directions. Shouldn't this cause the 'edge' of the Sun to appear fuzzy? And yet the Sun appears to have a sharp edge: when you watch the Sun setting, there comes a moment when the last sliver of the solar disk is just visible, and then, in an instant, it disappears from view.

Why does the edge of the Sun seem sharp? The reason has to do with the fact that the light we see when we look at the Sun comes from gas which is not entirely free to diffuse wherever it wants to go: the force of *gravity* holds on to the gas. If gravity had its way, it would drag all of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the gas down into the center of attraction at the center of the Sun. But gravity does not have its way: the inward force of gravity is balanced by the pressure which the gas exerts. Gas pressure arises from the fact that the individual atoms are not bound tightly together, but are free to move in all directions. Because of this freedom of movement of its constituent particles, a gas has a natural tendency to want to diffuse away and fill all available space.

Under some conditions, the opposing tendencies of gravity and pressure can come into balance with each other, and the gas can remain stationary. But it is sometimes impossible to find a balance between gravity and pressure: in such a case, the gas cannot be stationary. Both of these cases apply in different parts of the solar atmosphere, and they help us understand why the Sun has a non-extended atmosphere in one region and an extended atmosphere in another region.

Hydrostatic equilibrium

When the attractive force of gravity is in exact balance with the tendency of gas to diffuse, the gas is said to be in 'hydrostatic equilibrium' (HSE). The equation which describes HSE can be derived by requiring that when an extra parcel of gas is added to the atmosphere, the extra weight of that parcel must be balanced by extra pressure. The equation which describes this balance is

$$\frac{\mathrm{d}p}{\mathrm{d}r} = -g\rho \tag{1}$$

where p is the gas pressure, r is the radial distance from the center of the Sun, g is the acceleration due to gravity and ρ is the gas density.

A particularly simple solution of the HSE equation can be obtained if we consider the special case of a perfect gas at a uniform temperature. In such a gas, the pressure exerted by the gas is related to the density of the gas according to Charles' law:

$$p = \frac{R_g \rho T}{\mu}. \tag{2}$$

Here, T is the temperature of the gas (on an absolute scale), μ is the mean molecular weight of the gas (on a scale where a hydrogen atom has a weight of unity), and R_g is referred to as 'the gas constant'. The value of R_g is defined as the ratio of Boltzmann's constant to the mass of a hydrogen atom: R_g ($\approx$8.3 J mol^{-1} K^{-1}) is a measure of the energy in the random (thermal) motions of 1 g of hydrogen atoms at a temperature of 1 K above absolute zero.

Inserting equation (2) into equation (1), and assuming for the moment that g and μ are also constant, we can solve for density ρ as a function of radius r:

$$\rho = \rho_0 \exp[-(r - r_0)/H]. \tag{3}$$

Here, ρ_0 is the density at the level in the atmosphere where the radial distance equals r_0. The quantity H, which has the dimensions of length, is related to the parameters of the atmosphere by the expression

$$H = \frac{R_g T}{g\mu}. \tag{4}$$

To see the significance of the length H, note that when we measure the density of the gas at a height H above $r = r_0$, equation (3) tells us that the local density is smaller than the density at $r = r_0$ by a factor e. Every time we move up by a *linear* distance H, the density falls off by a *multiplicative* factor e. Thus, if we move up to a height of $10H$ above r_0, the local density will be smaller than at $r = r_0$ by a factor of e^{10}, i.e. by a factor of about 22 000. This means that, for every 22 000 atoms at $r = r_0$ there is only about one atom at $r = r_0 + 10H$. Another way of looking at this is to consider that, of 22 000 atoms at $r = r_0$, all of which have a tendency to diffuse outwards into space, only one has enough energy (i.e. fast enough speed) to climb up against gravity to heights of $10H$ above r_0. The number of atoms which are able to climb up to even greater heights (say, $20H$) is even smaller (by a further factor of 22 000). Clearly, the number density of atoms falls off very rapidly with increasing height in this example of HSE.

The quantity H, which measures how rapidly the density falls off with height, is referred to as the 'scale height' of the atmosphere. Physically speaking, a significant role for the scale height has to do with the amount of gas in the atmosphere: if all the gas in an atmosphere were forced to have the same density, then the thickness of the atmosphere would be equal to H. That is, if I imagine that I am standing at a region in the atmosphere where the local number density of the gas is n_0, then the total number of particles lying in a column of area A above my head equals n_0 times H times A. This result will be useful below when we estimate the number density of particles in the corona.

Now that we have an expression for H, we can understand why the edge of the Sun is sharp: it has to do with the numerical value of H in the layers of the solar atmosphere that we can see with our eyes.

The scale height in the solar photosphere

The layers of the Sun which we can see are called the 'SOLAR PHOTOSPHERE' (literally, the 'light-sphere'): these layers contain the atoms of gas which scattered the photons of light for the last time before the light started on its journey through the (relative) vacuum of interplanetary space towards our eyes. Gas in the photosphere has a temperature T of about 5000 K. Moreover, the gas consists of about 90% neutral hydrogen atoms and about 10% helium: this leads to $\mu \approx 1.3$. The acceleration due to gravity in the photosphere g_{phot} can be estimated from Newton's formula $g_{\mathrm{phot}} = GM_\odot/R_\odot^2$, once we know the mass of the Sun $M_\odot \approx 2 \times 10^{30}$ kg and the radius of the photosphere $R_\odot \approx 7 \times 10^8$ m. ($G \approx 6.67 \times 10^{-11}$ N m^2 kg^{-2} is Newton's constant of gravitation.) Inserting the above values, we find $g_{\mathrm{phot}} \approx 270$ m s^{-2}.

Now we are in a position to evaluate the scale height H_{phot} in the photosphere: $H_{\mathrm{phot}} \approx 120$ km.

The 'edge' of the photosphere: how sharp is it?

Suppose I point a telescope at the edge of the Sun and line up a cross-hair exactly with the photosphere. The density

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the gas there has a certain value, ρ_0, and this is enough to scatter the photons from the Sun for the last time and make the photons come towards me. I measure the intensity of the light by counting how many photons come into my detector every second from this first pointing, P_1. Now suppose I offset the pointing of the telescope to a second position: I set the cross-hairs on gas which lies 1200 km above the photosphere. The density of the gas at this new position, which is $10H_{\rm phot}$ above the first position, is 22 000 times less than at the first position. With such a low density, there will be roughly 22 000 less scatterings occurring, and as a result, the number of photons I will count from this second pointing P_2 will be less than P_1 by 22 000. Since P_2 is so much less than P_1, the gas at pointing 2 appears dark in comparison to the photosphere. Thus, in moving from P_1 to P_2, I have moved from bright sunlight into what is essentially darkness: by all definitions, I have moved past the 'edge' (or limb) of the Sun.

Now in the context of how sharp the edge of the Sun appears to us, the question is: how far, in angular terms, did I have to offset the telescope to go from pointing 1 to pointing 2? The answer depends on the distance to the Sun. This is known to be about 150 million km. At such a distance, in order to shift a telescope cross-hair a distance of 1200 km, I must move the telescope through an angle of between 1 and 2 arcseconds. The human eye cannot resolve such a small angle (as small as a golf ball seen from a distance of a few miles). As a result, when I look at the edge of the Sun, I see the Sun going from bright (photosphere) to dark (gas at a height of $10H_{\rm phot}$) within an angular distance that is so small that I cannot resolve it. The edge of the Sun therefore *appears* essentially discontinuous to me even though, in principle, the gas density actually falls off smoothly and continuously according to the predictions of HSE.

The chromosphere

During an eclipse, when the chromosphere appears above the edge of the Moon, the light comes mainly from spectral lines emitted by hydrogen (see HYDROGEN SPECTRUM). The strongest of these in the visible spectrum is referred to as Hα. The wavelength of Hα is 6563 Å, which is in the red region of the spectrum. It is the predominance of this line that gives the chromosphere its reddish color. Careful measurements of the way in which the intensity of Hα and other spectral lines varies with height enable observers to determine a 'scale height' for each line: this is the height over which the intensity of the line falls off by a factor of e. Different lines yield somewhat different results, but average values are many hundreds of kilometers, or even 1000 km.

Although a model is required in order to interpret strictly the intensities of spectral lines in terms of local gas densities, the data indicate that scale heights in the chromosphere are certainly larger than $H_{\rm phot}$. Hydrostatic equilibrium, where $H = R_g T/g\mu$, is a good approximation for the chromosphere, and as a result, the chromosphere requires temperatures which are hotter than in the photosphere. The excess temperature in the chromosphere compared to the photosphere is required to be a few thousand degrees.

The corona: what are the temperatures?

The image of the corona in figure 1 indicates that the corona extends in certain directions to an appreciable fraction of a solar radius before fading out. Measurements of the way in which the brightness fades out can be used to assign also a 'scale height' $H_{\rm cor}$ to the density of the coronal material. Typical values of $H_{\rm cor}$ are one or two tenths of a solar radius, i.e. $H_{\rm cor} \approx 10^8$ m. This is almost 1000 times larger than the scale height in the photosphere.

This factor of almost 1000 indicates that the Sun truly does have an outer atmosphere which is 'extended' compared to what we see in the photosphere.

Does hydrostatic equilibrium apply to the corona? The answer depends on what part of the corona one is discussing. First, let us discuss regions close to the Sun, within a few tenths of a solar radius of the photosphere. (We will return in the next section to regions farther out.) Close to the Sun, HSE is probably a good approximation. Therefore, in order to reproduce a coronal scale height $H_{\rm cor}$ which is almost 1000 times larger than in the photosphere, we need to have a coronal temperature which is almost 1000 times as large as in the photosphere. That is, the coronal temperatures must be a few million degrees.

Can this be true? Can gas in the extended solar atmosphere really be almost 1000 times hotter than the gas in the photosphere? The answer is: yes! The surest proof that such hot gas is present is provided by instruments which detect x-rays from the Sun. An example is shown in figure 3. This image was constructed by the Soft X-ray Telescope (SXT) on board the YOHKOH spacecraft. SXT responds mainly to photons which are emitted by gas at temperatures of 2–3 million K. The SXT image shows that the corona does not emit uniformly: there are bright spots (where photon intensity is large) and dark spots (the so-called 'coronal holes') where photon intensity is so low that the detector records 'nothing'. In the dark 'coronal holes' (e.g. near the bottom of the image in figure 3), we can see the circular outline of what our eyes see when we look at the Sun: the photosphere. Our eyes are sensitive to photons with wavelengths of a few thousand Å, and in such wavelengths, the photosphere of the Sun (with its temperature of about 5000 K) is very intense. That is why the Sun (or more strictly, its photosphere) appears so bright to our eyes. But the detectors in SXT are looking for photons with wavelengths of a few Å: the photosphere emits very few of these, and as result, SXT 'sees' the photosphere as a dark object.

The brightest features in the SXT image are compact regions where a trained eye can see so-called 'CORONAL LOOPS': these appear as partial arcs of circles that start at one position on the surface of the Sun, loop upward to a certain height, and then return to the solar surface. When images of the Sun in x-rays are compared with magnetic field data, it is found that the x-ray loops are associated with regions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

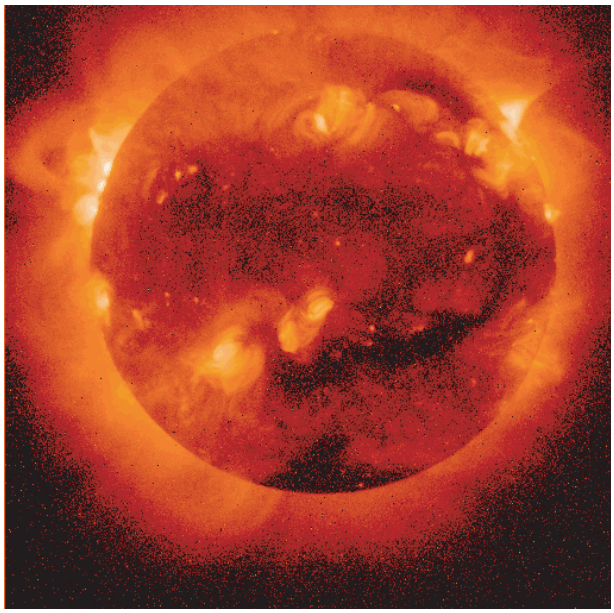


Figure 3. The Sun as seen in x-rays: the image was obtained by the Soft X-ray Telescope on board the YOHKOH satellite in April 1999.

of strong magnetic field. Each loop rises out of a region where the magnetic field has one polarity (say northward), and returns to the Sun in a region of the opposite polarity.

The discovery that magnetic fields serve to organize the structure of the solar corona was a key step in astronomers' attempts to understand how the gas in the corona can be heated to temperatures of almost 1000 times the photospheric temperature. In the 'CORONAL HOLES' , magnetic field lines are open into interplanetary space: ionized gas is free to stream out along these open field lines. The essential difference between coronal regions which are dark in x-rays and those which are bright in x-rays has to do with the magnetic field topology: the fields are open or closed (respectively) in dark or bright regions of the corona.

The corona does not have a unique temperature. In active regions, some material may reach temperatures $T_{\max}$ as high as several million degrees. In the quiet sun, $T_{\max}$ may be only 1–2 million degrees. In the flaring Sun, $T_{\max}$ may be several tens of millions of degrees. But in all cases, there are also measurable quantities of gas with temperatures below $T_{\max}$: in fact, by observing the Sun in spectral lines which are formed at a variety of temperatures, one can determine how much gas is present at various temperatures. The distribution of material at various temperatures is called the differential emission measure (DEM). Typically, the shape of the DEM is a power law with positive slope between temperatures of the order of 10^5 K and $T_{\max}$. Below 10^5 K, the DEM has a negative slope. Detailed study of the shape of the DEM may eventually allow one to infer something about the physical nature of the mechanism which is responsible for heating the corona.

The corona: what are the densities?

Now that we know the scale height $H_{\rm cor}$ in the low corona (where HSE applies), we can use the observed brightness to estimate roughly the density of the material at the base of the corona, $n_{\rm cor}$ electrons m^{-3}. To do this, let us imagine what happens to photons which emerge from 1 m^2 of the solar photosphere during an eclipse of the Sun. These photons stream away from the Sun, mostly along the radial direction, and by the time they have passed through the corona, the total number of electrons $N_{\rm e}$ which they have passed by in their 1 m^2 column is $n_{\rm cor} \times H_{\rm cor} \approx 10^8 n_{\rm cor}$. There is a finite probability $p_{\rm es}$ that, when a photon passes by an electron, the photon will be scattered: the quantity which determines $p_{\rm es}$ is the so-called Thomson cross-section, $\sigma_{\rm e}$ which is about 10^{-28} m^2. (The value of $\sigma_{\rm e}$ a few times $\pi r_{\rm e}^2$ where $r_{\rm e}$ is the classical radius of the electron: $r_{\rm e} = {\rm e}^2/mc^2$. Here, e and m are the charge and mass of the electron, and c is the speed of light.) The probability that a photon will scatter after passing through $N_{\rm e}$ electrons cm^{-2} is $p_{\rm es} \approx N_{\rm e} \times \sigma_{\rm e}$. Inserting the above values, we find that

$$p_{\rm es} \approx 10^{-20} n_{\rm cor}. \qquad (5)$$

Now the observed intensity of the corona is about 10^{-6} times the photospheric intensity. This indicates that $p_{\rm es}$ has an empirical value of order 10^{-6}. Referring to equation (5) above, we see that $n_{\rm cor}$, the electron number density in the low corona, must be of the order of 10^{14} m^{-3}. Since the major constituent of the Sun is hydrogen, the most abundant element in the corona is also hydrogen. In the photosphere, hydrogen exists in a form which is almost entirely electrically neutral, with one electron bound to each proton. In the corona, the temperature is high enough to ionize the hydrogen essentially completely: for each free electron, therefore, there is also one free proton.

Breakdown of hydrostatic equilibrium: 'super-extended' atmosphere

We mentioned above that in the low-lying parts of the solar corona, hydrostatic equilibrium (HSE) is probably satisfied. But once we admit the existence of gas at temperatures of a million degrees of more, the assumptions of HSE eventually break down. To see why this is so, we return to equation (1). So far, we have treated the gravitational acceleration g as a constant: this is a good approximation as long as we are considering gas within a few scale heights of the photosphere. (The value of g at a radial distance of $r = R_\odot + 10H_{\rm phot}$ differs from that at $r = R_\odot$ by only a few tenths of 1%.) But when we consider things on a larger scale, g can no longer be regarded as constant: at a radial distance r from the Sun, we need to use the local value of g, i.e. $GM_\odot/r^2$. This changes the solution of equation (1) in a fundamental way. Inserting the radially varying g into equation (1), and again assuming for simplicity that the gas has a unique

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

temperature, we find that the solution now takes on a strikingly different form:

$$\rho = \rho_0 \, \mathrm{e}^{\{[A(r_0/r-1)]\}}. \tag{6}$$

Here, $A = GM_\odot \mu / R_g T r_0$ is a measure of how effectively the thermal pool fills up the Sun's gravitational well. Equation (6) describes how the density in the solar corona would behave if the corona were in hydrostatic equilibrium. A remarkable feature of the density distribution in equation (6) is that as we move farther and farther from the Sun, i.e. as r tends to infinity, the density of the coronal material does *not* fall to zero. Instead, the density approaches a constant value: $\rho(\infty) \rightarrow \rho_0 \, \mathrm{e}^{-A}$.

Inserting $T = 10^6$ K, and use the solar parameters mentioned above (except that $\mu \approx 0.5$ in fully ionized hydrogen), we find $A \approx 12$. Now, as we have seen, the gas at the base of the solar corona has a number density of some 10^{14} protons m^{-3}. At infinity, a corona in hydrostatic equilibrium would have a number density which is less than this by e^{-12}, or 6×10^{-6}. Thus, the number density of the solar corona at infinity would be about 6×10^8 protons m^{-3}. Such a density is much larger than the density of gas in the interstellar medium (ISM): the latter contains typically 10^6 protons m^{-3}.

Because the ISM has a density (and pressure) which is many times smaller than the corona in hydrostatic equilibrium, it is impossible for the ISM to contain the corona: the latter has a pressure which is too high to be confined.

What happens when HSE breaks down? To answer that, we note that the equation of HSE is itself only a special case of a more general equation which describes the law of conservation of momentum. For a gas of density ρ moving radially outward at speed v, the momentum per unit volume is ρv. Conservation of momentum for this gas is described by the equation:

$$\rho \frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\mathrm{d}p}{\mathrm{d}r} - \rho g + F \tag{7}$$

where F is a general symbol for any other forces which may be present in the corona apart from pressure and gravity. (Magnetic fields can give rise to extra forces which are important for the corona.) In the case where $F = 0$, and HSE is satisfied (see equation (1) above), the right-hand side of equation (7) is exactly zero, and equation (7) can be satisfied by the solution in which the gas does not move, $v = 0$: the absence of motion justifies the term 'static' in 'hydrostatic equilibrium'. But if pressure forces do not balance gravity, then the right-hand side of equation (7) is non-zero. As a result of the unbalanced forces, the gas begins to accelerate in response to Newton's famous second law of motion. Instead of hydro-'static' conditions, we now have to deal with unbalanced forces (i.e. dynamics). The onset of acceleration in the radial direction means that the corona must expand. This expansion of the coronal gas gives rise to a flow which is named the 'SOLAR WIND'.

It is important to note that when we talk of the solar wind, we are *not* talking about evaporation, as if a small fraction of the coronal material were 'boiling off': there is nothing evaporative about the process described by equation (7). The solar wind involves a truly hydrodynamic expansion *of the entire corona* (see also SOLAR WIND: CORONAL ORIGINS).

In the simplest case of steady flow, the velocity of outflow v does not depend on the time, but varies with radial distance. In this case, equation (7) can be written in the form

$$v \frac{\mathrm{d}v}{\mathrm{d}r} = -\frac{1}{\rho}\frac{\mathrm{d}p}{\mathrm{d}r} - \frac{GM_\odot}{r^2} + F. \tag{8}$$

The details of how rapidly the wind accelerates, i.e. the shape of v as a function of radial distance r, depend on the temperature (which controls the pressure) and especially on the nature of any extra forces F. But regardless of the details of F, we expect that at great distances from the Sun, all terms on the right-hand side of equation (8) will become small. This means that $\mathrm{d}v/\mathrm{d}r \rightarrow 0$ at great distances from the Sun. In other words, v tends to a constant value (the terminal speed) far from the Sun. The value of the terminal speed depends on where on the Sun the wind originates: empirically, wind from the polar regions of the Sun is found to have the highest terminal speed, some 700–800 km s^{-1}. The slowest observed wind speeds (about 300 km s^{-1}) are found to originate in streamers which are more common near the Sun's equatorial regions. The differences between winds from different locations on the Sun arise at least in part from differences in the forcing term F in equation (8). Over the polar coronal holes, waves on the magnetic field may help to provide momentum and energy to the wind.

The solar wind extends far beyond the visible edge of the Sun, far beyond even what the eye sees during an eclipse: in fact the wind expands past the orbit of the Earth, and past the orbits of all of the known planets. No-one yet knows for sure how far out the wind expands. The most remote spacecraft which have been able to sample the wind (Voyagers 1 and 2), are at the time of writing (April 1999) about 50% farther from the Sun than Pluto's mean distance, and they still see the solar wind expanding. Some theories make predictions about where the solar wind will be stopped by running into the ISM: but the Voyagers have not yet reached such a point (see VOYAGER MISSION).

The fact that the solar corona extends throughout the solar system gives a new twist to the definition of 'extended atmosphere': it can come as a shock when one first realizes that the Earth on which we live is itself embedded in the Sun's (super)-extended atmosphere.

Mass flux carried by the solar wind

How much mass does the solar wind carry away from the Sun? Spacecraft measurements of the solar wind as it passes Earth's orbit indicate that the mean solar wind density is some 5–10 $\times$ 10^6 protons m^{-3} while the mean velocity is 300–400 km s^{-1}. These measurements indicate that the Sun loses mass at the rate of a few million tons per

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

second. This is comparable to the rate at which mass is lost in the core of the Sun due to thermonuclear reactions. If the Sun has lost mass at a constant rate during its evolution (lasting some 10^{17} s so far), the wind has not caused any significant alteration to the mass of the Sun during its life: the change in mass is no more than 0.1%.

Other stars

The existence of chromospheres on stars of spectral classes G, K and M has been known for decades: the chromospheres reveal their existence in the form of emission lines of hydrogen and other elements in the spectrum.

Since about 1980, when the first x-ray telescopes on spacecraft with enough sensitivity to detect stars became operational, it has been discovered that stars of almost all types emit x-rays. The strongest sources of x-rays among single stars occur in dwarfs of spectral class M: some M dwarfs emit almost 1% of their entire luminosity in the form of x-rays. Compared to the Sun (where the x-ray output amounts to only one part in 10^5 or 10^6 of the entire luminosity), M dwarfs are truly efficient generators of coronae. Moreover, the coronae in M dwarfs contain components which are much hotter than in the Sun: the x-rays from certain M dwarfs require source temperatures of several tens of millions of degrees. In the Sun, temperatures as high as this are achieved only temporarily in short-lived explosions (flares). On the basis of data from the ground, it has been known for decades that M dwarfs are sites of large flares and strong magnetic fields. It seems likely that the strong heating of their coronas is connected to magnetic activity.

Cool giant stars exhibit evidence of chromospheric material, but not of coronal material. In such stars, there are, however, comparatively dense but comparatively slow winds, with mass loss rates which exceed the solar wind value by several orders of magnitude. These winds can be considered as extended atmospheres of such stars: although the photosphere of Betelgeuse (a red supergiant) has dimensions of a few hundred solar radii (i.e. a few AU), emission from the wind in certain lines can be traced out to distances of the order of 10 000 AU.

What heats an extended atmosphere?

As we have seen, the presence of an extended atmosphere depends on excess temperatures. What causes this excess heating? Attempts to answer that question have attracted a lot of attention in the course of the past few decades. The quest for an answer comes down to identifying a source of mechanical energy which can do work on the gas.

As far as chromospheric heating is concerned, a consensus seems to be emerging: acoustic waves from turbulent convection beneath the photosphere can heat at least the lower regions of the chromosphere. As these waves propagate upwards in the atmosphere, reaching gas of progressively lower density, the pressure fluctuations of the acoustic waves grow in intensity until shock waves are formed. The shocks cause localized heating of the gas, enough to reproduce many of the observed features of chromospheric emission. Dissipation of the waves is severe enough that in the Sun, almost no waves survive even as far as the upper chromosphere. In other stars, generation of acoustic waves by their convection may in some cases (especially M dwarfs) have properties which are more favorable to the survival of acoustic waves in the upper chromosphere: in such stars, the acoustic waves with longest periods may even help to provide some heat to the low corona.

In the chromosphere, the deposition of mechanical energy is balanced by radiation in spectral lines of abundant elements in certain stages of ionization. As long as these lines remain available for radiation, deposition of energy does not necessarily result in severe local heating: the deposited energy may simply be radiated effectively by slight increases in temperature. As a result of this competition between gain of energy and loss by radiation, the chromosphere on average may contain one or more regions where the temperature remains almost constant with increasing height. These temperature 'plateaus' may occur between about 6000 K and 20 000 K. The process of hydrogen ionization also serves to absorb mechanical energy without having the local temperature increase significantly: as a result, the average chromosphere in the Sun displays the remarkable property that the mean proton density remains essentially constant (within a factor of ± 2) over a range of altitudes where the mean hydrogen atom density falls off by a factor of 10^4. But overall, there is on average an inexorable rise in temperature with increasing height, and each of the elements becomes progressively more ionized. Eventually, one reaches a height where the temperature is such (between 10^4 K and 10^5 K) that *none* of the abundant elements is in stages of ionization where strong spectral lines are available. In such conditions, addition of extra mechanical energy inevitably causes the temperature to climb rapidly to millions of degrees. This rapid transition to coronal temperatures occurs over a narrow range of altitudes.

As regards coronal heating, magnetic effects must dominate in the source of mechanical energy. One widely discussed candidate is waves which propagate along the magnetic field: these waves can dissipate their energy in a variety of ways. For example, in active regions, a magnetic loop may act as a sort of resonant cavity which the wave can cause to 'ring' in such a way that the wave energy is deposited in the loop. And in coronal holes, waves on the magnetic field may dissipate and deposit both momentum and energy into the coronal gas. In active regions, miniature magnetic explosions ('nano-flares') may be occurring often enough that they can keep some closed magnetic loops hot.

Data from the SOHO spacecraft may help to pin down the mechanism which cause coronal heating. For example, there is already evidence that in coronal holes the coronal heating mechanism heats protons more efficiently than it heats electrons. And heavy ions (such as oxygen) are heated even more than protons.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The existence of the solar corona dates back to antiquity. But the first realization that coronal temperatures were indeed of order 1 million K came from

Grotrian W 1939 *Naturw.* **27** 214
Edlen B 1941 *Arkiv Mat. Astron. Fys.* **28B** (1) 1

Reports on the scale heights of chromospheric emission lines, as well as a detailed study of the chromosphere using a non-LTE approach to radiative transfer in optically thick lines can be found in the book

Thomas R N and Athay R G 1961 *Physics of the Solar Chromosphere* (New York: Interscience)

The above book also discusses reasons to expect the presence of plateaus in the chromospheric temperature.

For a discussion of the 'thermostatic' effects of hydrogen ionization in the chromosphere, see

Athay R G 1981 *The Sun as a Star* ed S Jordan, NASA SP-450 p 85

The physics of acoustic heating of the time-averaged chromosphere has been studied extensively by P Ulmschneider and colleagues at Hiedelberg. For an up-to-date summary of this work, see

Ulmschneider P 1999 *Solar and Stellar Activity: Similarities and Differences, Astron. Soc. Pacific Conf. Ser.* vol 158 eds C J Butler and J G Doyle (San Francisco: Astronomical Society of the Pacific) p 260

For an alternate point of view concerning acoustic heating of the quiet chromosphere, see

Carlsson M and Stein R F 1995 *Astrophys. J.* **440** L29

For a critical review of various mechanisms which have been proposed for coronal heating, see

Zirker J B 1993 *Solar Phys.* **148** 43

For a summary of the heating of the active corona by resonant absorption of magnetic waves, see

Davila J M 1991 *Mechanism of Chromospheric and Coronal Heating* eds P Ulmschneider *et al* (Berlin: Springer) p 464

The concept of 'nanoflare' heating of closed magnetic loops in the solar corona was proposed in

Parker E N 1987 *Solar Phys* **111** 297

The original work on the outflow of the solar wind can be found in

Parker E N 1963 *Interplanetary Dynamical Processes* (New York: Interscience)

For discussions of how specific dissipation mechanisms of waves on the magnetic field can be used not only to heat the coronal gas but also to drive the wind, see, e.g.

Tu C Y and Marsch E 1997 *Solar Phys.* **171** 363
Cuseri I *et al* 1999 *Astrophys. J.* **514** 989

Dermott Mullan

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extragalactic Astronomy *c.*1900–1950

In 1931 the Dutch astronomer WILLEM DE SITTER (1872–1934) declared that 'Never in all the history of science has there been a period when new theories and hypotheses arose, flourished, and were abandoned in so quick succession as in the last fifteen or twenty years'. de Sitter was looking back on a period of turmoil and upheaval in the physical sciences, in which perhaps most importantly for astronomers the long-standing debate on the nature of NEBULAE had been settled and in which the entirely unexpected expansion of the universe had been proposed and widely accepted. These two epoch-making developments were due mainly to astronomers who in the first three decades of the 20th century made observations with giant optical telescopes of a class of nebulae, the spiral nebulae, which, it would become generally accepted after the mid-1920s, were galaxies. In fact, in these three decades extragalactic astronomy, as astronomers later would understand it, was established. This new field was in many ways a distinctly American activity—in considerable part because it was in the United States that the most powerful telescopes and the best-equipped astrophysical observatories were to be found, together with a sizable body of well-trained astrophysicists interested in problems of the nebulae.

Early history

The 18th and 19th centuries had seen wide ranges of opinion on the nature of the nebulae, the two most extreme being that all nebulae are star systems external to our Galaxy, so-called 'island universes', or that all the nebulae are indeed truly nebulous. However, by the end of the 19th century it had become widely accepted that none of the visible nebulae was an external galaxy. External galaxies might exist, but if so they lay beyond the reach of even the most powerful telescopes.

Three main pieces of evidence were reckoned by most astronomers to tell heavily against the island universe theory: (1) the distribution of the nebulae (it was generally agreed that the nebulae very largely avoid the plane of the Milky Way but, if the nebulae were truly independent galaxies, why should their position be correlated with our own Galaxy's orientation?), (2) the well-established gaseous nature of some nebulae and (3) the 'nova' of 1885 (now accepted to have been a supernova) that had flared in the heart of the Andromeda nebula. (The nova had reached a luminosity of about 1/10th of the entire nebula. How, astronomers asked, could a single star rival the brightness of a vast assembly of stars? Their answer was that it could not. The nebula was not a vast stellar assembly.)

At the start of the 20th century, most astronomers also judged the vast majority of nebulae to be 'spiral nebulae', that is, to exhibit a spiral shape. Lord ROSSE and the Parsonstown observers in the middle of the 19th century had been the first to discern the spiral form of a number of nebulae, but in the late 1890s JAMES E KEELER (1857–1900), while photographing nebulae and star clusters with the 36 in Crossley reflector at Lick Observatory in California, recorded vastly more spirals on his plates than he had expected; in fact, he reckoned, 120 000 of them were within reach of his telescope. In 1899 the German astrophysicist Julius Scheiner (1858–1913)—working at Potsdam, the leading German center for astrophysics—detected absorption lines in the spectrum of the largest, and so presumably one of the nearest, spirals, the Andromeda nebula. As it was well-known that the spectra of star clusters exhibited absorption lines, Scheiner contended that the Andromeda nebula itself was a star system that lay beyond the Galaxy. At first Scheiner's arguments won few adherents since nearly all astronomers were convinced that spiral nebulae developed into either single stars or sparse clusters. However, around 1910 a few investigators started to challenge this position. Led by E A Fath at Lick and Max Wolf at Heidelberg, they extended Scheiner's researches on the spectra of spirals and found that while some spirals displayed bright lines—well known by this time to be characteristic of gas clouds—most of those examined had spectra dominated by dark absorption lines. Perhaps, after all, some astronomers decided, the spirals are clusters of stars, maybe even distant galaxies.

During the 1890s Keeler had measured the shifts of the spectral lines of a number of planetary nebulae and, interpreting the shifts as Doppler shifts, he had determined their radial, or line-of-sight, velocities. They had radial velocities, on the whole, a little higher than, but essentially similar to, those of the stars. The measurement of the radial velocity of a spiral proved a much more daunting problem because of the low surface brightness of the spirals, but in 1913, V M SLIPHER (1875–1969) of Lowell Observatory in Arizona argued that the Andromeda nebula was rushing towards the Sun at 300 km s^{-1}, the highest speed then recorded for an astronomical body. Slipher's find was so puzzling that a few suspected its correctness. Doubts were, however, swiftly dispelled when others confirmed Slipher's result and Slipher himself obtained similar values for further spirals.

The observations of the spectra of spirals had persuaded some astronomers that the spirals were, or might well be, external galaxies, and the measured radial velocities won still more to this view. The velocities were so much higher than those of the confirmed members of the Galaxy that some astronomers decided that the spirals could not be a part of the same dynamical system. One result was that, by the mid-1910s, the island universe theory, a theory that a decade or so earlier had been almost completely out of favor, had become a leading topic of debate.

The shift of opinion about external galaxies had also been influenced by two other factors: (1) the wide-spread suspicion that the Galaxy itself has a spiral structure and (2) the 'grandeur' of the island universe theory. In 1911, for example, the brilliant English astrophysicist A S EDDINGTON (1882–1944) had declared that the island universe theory 'opens up to our imagination a truly magnificent vista

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of system beyond system—of universe beyond universe'. The theory's appeal to the scientific 'taste' of astronomers alone had, nevertheless, not been enough: the shift in opinion about external galaxies had also required the hard evidence of the spectrum and radial velocity observations of the spirals.

Novae in spirals

In 1917 astronomers found more evidence that bore on the debate about nebulae when NOVAE were detected serendipitously in a number of spirals. The first were found by H D CURTIS (1872–1942), an astronomer at the Lick Observatory, and a similar discovery was made shortly after by G W RITCHEY at Mount Wilson Observatory, also in California. Astronomers now eagerly searched old plates of spirals for further novae. By late 1917 sufficient numbers of novae had been located for Curtis to dismiss the 1885 nova in the Andromeda nebula as abnormally bright, and thus not a reliable indicator of the nebula's distance. Curtis now compared the brightness of the typical novae in spirals with the average brightness of novae within the Galaxy, and in so doing he estimated these spirals to be a hundred times more distant than the galactic novae, thereby, in the opinion of Curtis and other astronomers, providing yet more evidence for the island universe theory. Indeed, by the end of 1917 nearly all of the leading students of spiral nebulae judged it likely that they were external galaxies.

A key issue here is that the Galaxy was also the measuring rod against which supposed galaxies were measured, and contemporary estimates of the dimensions of the Galaxy played a part in determining the status of the island universe theory. This point became particularly important when in late 1917 HARLOW SHAPLEY conceived and fashioned a bold new model of the Galaxy in which its diameter is roughly 300 000 ly—roughly 10 times larger than the commonly accepted estimates—and in which the Sun is placed some tens of thousands of light-years from its center. Certainly Shapley came to be loathe to accept that the spirals were comparable with his gigantic stellar system and he suggested that they were truly nebulae, but he wavered on the status of the spirals and between 1917 and 1920 he changed his mind three times on their relation to the Galaxy. Shapley's writings on the spirals did nevertheless bring him an invitation to participate in what has become one of the best-known, but often misunderstood, events in 20th century astronomy: the so-called 'Great Debate'.

The 'Great Debate' and internal motions

In April 1920 Shapley and Curtis both delivered talks on 'The scale of the universe' to the National Academy of Sciences, one of America's most prestigious scientific bodies.

Shapley chose to speak at a very popular level (most likely his chief concern at the Debate was to appeal to the members of the audience concerned with selecting the next director of the Harvard College Observatory, a job Shapley coveted), and it was only after the Debate, with the published papers of Shapley and Curtis, that the protagonists actually came to grips with each other's arguments. In his published paper Curtis maintained that the spirals were external galaxies, arguing chiefly on the basis of the by now familiar observations of their spectra, radial velocities and novae. Curtis also attacked Shapley's estimate of the size of the Galaxy. Curtis and many others had advocated a comparable-galaxy theory (that is, a theory in which the spirals were not just external galaxies, but in addition were comparable in size with the Galaxy). He admitted that if the spirals were as large as Shapley insisted the Galactic system to be, then the spirals would be at such enormous distances that the novae within them would have to be impossibly bright, thereby undermining the case for the comparable–galaxy theory.

During the late 1910s and early 1920s another class of observations came to play a leading role in the island universe debate: the measurements of the internal motions of the spirals. ADRIAAN VAN MAANEN (1884–1946), a Mount Wilson astronomer, began his investigations of this topic in 1915 when he scrutinized photographic plates, taken some years apart, of the giant spiral M101 (that is, number 101 in Charles Messier's catalog of nebulae). His goal was to detect and measure the size of any changes in appearance in the intervening years, and he decided that M101 did indeed exhibit a rotation, or, more likely in his view, motions along its spiral arms. At first the internal motions did not obviously conflict with the island universe theory. The English applied mathematician and theoretical astronomer J H JEANS (1877–1946) in fact exploited van Maanen's measures to champion the island universe theory and to buttress his own theories on the origin and development of spiral nebulae. Jeans' theories were very influential, especially among the Mount Wilson astronomers, and they were brought before a wide audience in his *Problems of Cosmogony and Stellar Dynamics* of 1919.

In his 'Great Debate' paper Shapley wrote little on external galaxies and largely confined himself to arguing that their fate was inextricably bound to his theory of the Galaxy. However, Shapley did cite the measurements of the internal motions of spirals made by van Maanen, claiming that these could not fit into a scheme in which the spirals were galaxies 300 000 ly in diameter—the size he claimed for our own Galaxy—as the sizes of the internal motions then implied velocities for the outlying regions greater than the velocity of light. This was, he judged, a clear reduction to absurdity of the island universe theory.

In the early 1920s van Maanen measured several more spirals and found that his results agreed very well with his earlier ones for M101. Some astronomers harbored doubts about the internal motions nevertheless. The motions were almost the only evidence against the island universe theory of spirals; and if the spirals were not galaxies, what were they? In particular, how could one interpret the spectrum, radial velocity and nova observations? These considerations persuaded some astronomers that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

van Maanen's scrutinies had not revealed real motions. Thus while the island universe theory lost some support, many judged it to be, on the whole, still much the best available explanation of the properties of spiral nebulae.

Hubble and Cepheids

The sudden, and unexpected, resolution of the debate on the existence of external galaxies was effected in late 1923 and 1924 by a series of discoveries made by the Mount Wilson astronomer EDWIN P HUBBLE (1889–1953).

Hubble had completed his doctoral dissertation on 'Photographic investigations of faint nebulae' in 1917 at a time when the island universe theory was very much in fashion and he seems to have been inclined towards it from this date on. After a brief period in the US army, he joined the Mount Wilson staff in 1919. In 1923, he decided to obtain plates of the Andromeda nebula in the hope of recording novae flaring in the nebula. For someone anxious to establish an accurate distance to the nebula this program had much to recommend it. In particular, by observing more novae he would be able to determine with greater precision the mean apparent brightness of novae in the nebula and thus estimate their distance—and so the nebula's—with more certainty than had been the case. He was also excellently placed to pursue such a program because he had access to the world's most powerful telescope, the Mount Wilson 100 in, and the rich resources of the world's leading astrophysical observatory.

He indeed found novae in the nebula, but in early 1924, to his surprise, he also decided that a star he had earlier identified as a nova was in fact a Cepheid variable. Hubble constructed a light curve for the Cepheid variable and rapidly secured its period. Then by assuming (as Shapley had done earlier in his studies of the Galaxy) that Cepheids of the same period have the same intrinsic luminosity wherever they are found, Hubble had been able to exploit the Cepheid period–luminosity relationship to calculate the nebula's distance. His result of 900 000 ly placed it far beyond the boundaries of even Shapley's model of the Galaxy. Astronomers also quickly agreed that Hubble had resolved the outer parts of the nebula into swarms of stars. Moreover, he soon found more Cepheids in the Andromeda nebula and some other spirals that supported his early findings. The Cepheids were accepted by the opponents of the island universe theory to be accurate distance indicators to the spirals, and the Cepheids' unambiguous message was that the spirals were indeed external galaxies, a message confirmed by almost all of the other evidence. The only discrepancy came with the internal motions, but soon nearly everyone rejected them as hopeless exaggerations of any real motions, and the case for their accuracy had collapsed under the weight of opposing evidence. Within a year or two of Hubble chancing upon the first Cepheid, the debate on the existence of island universes had effectively ended.

Once Hubble had settled the question of galaxies, astronomers began to turn increasingly to other problems associated with the spirals. One was how to classify

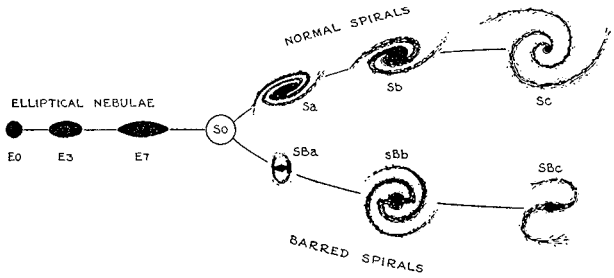


Figure 1. Hubble's 1936 sequence of nebular types.

them. In the 1910s Curtis had divided the nebulae into three main groups: diffuse nebulae, planetary nebulae and spirals. While Curtis's decision to create one grouping of 'spirals' had drawn attention to them as possible external galaxies, by the mid-1920s this scheme seemed to some astronomers to be far too restrictive and alternative classifications were advanced. The most influential was Hubble's, first proposed in 1922 and later elaborated by him (figure 1). Although Hubble claimed that the scheme was purely descriptive and independent of any theory, he was an enthusiastic advocate of Jeans's theories on the origin and development of extragalactic nebulae and this revealed itself clearly in his scheme. By 1926 Jeans had rejected van Maanen's measures as worthless, but in Hubble he had found another champion of his own theories. Hubble agreed with Jeans that there was one basic type of extragalactic nebula and that the various stages of development of this one type could explain the appearance of nearly all the observed extragalactic nebulae. Jeans placed the elliptical nebulae, which he believed to be enormous galaxy-sized clouds of dust and gas, at the beginning of the evolutionary sequence of these nebulae. He argued that the end of the elliptical stage is marked by the build up of spiral arms from the material of the nuclear regions. As this process continues, the nebula increases in size, until the arms are wide open and the nucleus inconspicuous. As a spiral ages its arms start to break up into condensations. These are proto-stars or proto-star clusters, and, as the spiral develops, the region of star formation moves inexorably inwards.

The Jeans theory of nebular evolution was rejected by, among others, Shapley and the Swedish student of nebulae, Knut Lundmark (1889–1958), who asserted that even elliptical nebulae were star systems, and who both emphasized the importance of possible gravitational interactions with other galaxies for the evolution of an individual galaxy. For this reason Hubble always used the term 'extragalactic nebulae' while Shapley employed 'galaxies', and these terms became shibboleths dividing East and West coast astronomers in the United States (indeed it was not until after Hubble's death in 1953 that 'galaxies' was universally adopted). Shapley and Lundmark both proposed hypotheses in which the Galaxy is composed of a number of smaller, interacting galaxies. They had done so in part because, despite the uncertainty

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

about the size of the Galaxy, by 1930 it seemed likely to the majority of astronomers that it was far larger than any other known star system.

The expansion of the universe

Shortly after he had turned seriously to the classification of nebulae, Hubble began to tackle another outstanding and perplexing problem in extragalactic research: whether or not there is a relationship between the distance of a galaxy and the size of its REDSHIFT.

V M Slipher, as we saw earlier, was the first to measure the radial velocity of a spiral nebula, and by 1914 he had been able to announce the radial velocities of 15 spirals. To secure these exceptionally faint and elusive spectra he required lengthy exposure times, sometimes as long as 80 h. Slipher's progress was therefore slow, and others were usually content to leave these difficult and tedious labors to him. However, as Slipher doggedly measured more velocities, a curious pattern began to emerge: nearly all of them were shifts to the red end of the spectrum; that is, if the spectral shifts were interpreted as Doppler shifts, all but a few were velocities of recession. In addition, the radial velocities were generally very much larger than those of the stars or gaseous nebulae, and their very sizes prompted some to query the Doppler, or velocity, interpretation of the spectral shifts.

Meanwhile, another development suggested that the radial velocities might be correlated with such distance-related variables as the diameters of the spirals. This possibility arose largely through a hypothesis advanced by the Dutch astronomer Willem de Sitter. In 1917 de Sitter had discovered an alternative solution to that found by ALBERT EINSTEIN (1879–1955) to the field equations of the general theory of relativity. de Sitter's solution, moreover, predicted a relationship between redshift and distance for distant stars and galaxies. As astronomers usually understood it, this relationship was not just the result of a recession of far-off bodies. Rather, the intrinsic properties of space and time in de Sitter's solution cause clocks to run more slowly the further they are from the observer. Hence, the atomic vibrations within a distant galaxy appear to slow down, the frequency of light decreases, the wavelength increases and a redshift is observed. de Sitter's solution applies strictly to an empty universe, but de Sitter himself and a number of his contemporaries thought that, if the density of matter was sufficiently low, it might be applicable to the real universe.

Even before the significance of de Sitter's model had been realized, there were speculations about a redshift–distance relation. However, as general relativity became more widely discussed and the notion became established of developing theoretical models of the entire universe that could be compared with actual observations, so de Sitter's model of the universe gained in plausibility, and in the 1920s it became generally agreed that, if de Sitter's model approximated the actual properties of the universe, then a relationship between redshift and distance was to be expected. Although there was much debate over the exact form of the relationship, it became widely accepted that, the greater the distance of a body, the larger should be its redshift.

Early attempts by C A Wirtz, Lundmark and others to establish a relation failed to gain support principally because the distances to the spirals were not known accurately, and these pioneers had been compelled to employ, for want of anything better, the very crude and, to other astronomers, unconvincing distance indicators of constant absolute magnitudes and constant absolute diameters.

Hubble, however, with the aid of the Mount Wilson 100 in reflector, was far better equipped than his competitors to investigate the redshifts and distances of extragalactic nebulae. He was also able to call on MILTON HUMASON (1891–1972), an outstanding observer, to make the demanding and time-consuming measurements of the redshifts.

To determine the distances of galaxies, Hubble exploited a carefully constructed distance 'ladder' that used a series of interlocking distance indicators, the most important rung of which was the Cepheid period–luminosity relationship. By applying this ladder Hubble obtained the distances to 24 of the 46 extragalactic nebulae whose redshifts had been determined, almost all by Slipher. The other 22 nebulae were not used in his primary proof because he reckoned they were not in groups and average apparent magnitudes could not be taken. When Hubble plotted the redshifts of the 24 nebulae against their distances he judged that a linear redshift–distance relation was the simplest way of representing his data, indicating a systematic recession of the nebulae represented by $V = kr$ (in which the k term is today known as Hubble's constant H, V is the velocity and r the distance). Furthermore, when distances were calculated (by means of apparent magnitudes) to the remaining 22 nebulae whose redshifts had been measured, they too, in Hubble's judgement, fitted the linear relation well.

By 1929 such a relation was not unexpected; indeed, in 1930 Hubble himself told de Sitter that 'the possibility of a velocity–distance relation among nebulae has been in the air for years—you, I believe, were the first to mention it'. The attempts of Wirtz, Lundmark and others to determine observationally the nature of the relation had failed to win general acceptance, but they had undoubtedly helped to prepare the way for Hubble's researches. Hubble had worked within a well-defined problem area, using a well-defined method, and had expressed his argument with care, stressing the solid observational basis of the relation, particularly the reliability of his distance measurements. His achievement was to persuade his colleagues that there was indeed a linear relation, at least to a first approximation.

The linear relation was swiftly incorporated into theoretical models of the universe. By early 1930 mathematical cosmologists had become acutely aware of the inadequacies of the static solutions of the field equations of general relativity as representations of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

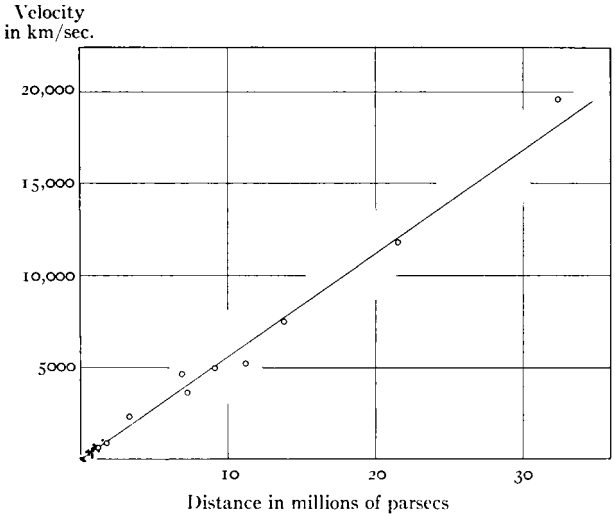


Figure 2. Hubble and Humason's 1931 apparent velocity–distance diagram. The dark circles in the bottom left hand of the plot represent the data from Hubble's 1929 paper on 'A relation between distance and radial velocity among extra-galactic nebulae'.

universe, but at this point a paper by the Belgian cosmologist GEORGES LEMAÎTRE (1894–1966), ignored when first published in 1927, was 'rediscovered' and was now welcomed enthusiastically. The combination of Lemaitre's theorizing and Hubble's handling of the observations convinced many that the static solutions had failed because the universe is in fact expanding, and the redshift–distance relation a natural consequence of this expansion. Expanding models, widely neglected during the 1920s, were now eagerly explored and after 1930 accorded a central role in extragalactic astronomy and cosmology. Indeed, by this time a research enterprise had come into being that centered on the nature and evolution, as well as the relationships between different sorts, of galaxies and their place in an expanding universe.

Hubble, however, preferred not to identify the redshifts definitely as Doppler shifts. Also, he regarded his 1929 paper as a preliminary report and not the publication of a completed piece of research. He particularly wanted to test the linear relation against more distant nebulae, and in 1931 he and Humason announced their latest findings (figure 2). Now he calculated that the most remote extragalactic nebula lay at an estimated distance of 32×10^6 pc, a large increase when compared with the furthest at 2×10^6 pc in the 1929 plot. There were, moreover, many more points on the 1931 diagram because 40 more redshifts were available, and Hubble judged the linear relation to be much clearer than in 1929. This rapid extension had been made possible by the combination of an extremely efficient spectrographic objective on the 100 in Mount Wilson telescope—at this time still the most powerful in the world—and Humason's skill as an observer. In the 1931 paper Hubble also sought to distinguish carefully between observation and interpretation. He emphasized that Humason had measured redshifts, and that they were not necessarily Doppler shifts. Although, as we have noted, many astronomers and mathematicians believed that the redshift–distance relation signified an expansion of the universe, others were incredulous at the size of the 'velocities' Humason was finding: if the redshift of one nebula was a genuine radial velocity, then it was receding at a staggering 19 700 km s^{-1}. Furthermore, if one simply extrapolated the present expansion back in time with no allowance for a possible acceleration in the expansion, the redshift–distance relation implied that there was a time about 2×10^9 yr ago when all the material in the universe had been packed together. For the numerous astronomers who accepted Jeans's estimate of 10^{13} yr as the age of the Galaxy, this time was far too recent, especially as the time scale for stellar evolution was also expected to be of this order. How, they asked, could the age of the Galaxy and its stars be much greater than the age of the universe? In consequence, the time scale problem spawned several (but by general agreement unconvincing) attempts to invent an alternative mechanism to the Doppler effect that would explain the redshifts.

While the debate on the nature of the redshifts has still not been finally settled, by the early 1950s the time scale problem had dissolved: astronomers had concluded that revisions of the cosmic distance scale meant that the present expansion rate did not conflict with estimates of the ages of the stars and galaxies. There were two main reasons for this change of opinion. First, astronomers now concluded that Hubble had mistaken H II regions—clouds of ionized gas—for the brightest stars in extragalactic nebulae. Since these clouds are intrinsically more luminous than stars, it meant that the corresponding distances had been underestimated. Secondly, the period–luminosity relationship for the Cepheids had been dramatically revised by WALTER BAADE (1893–1960).

Baade, a German-born astronomer, had joined the Mt Wilson staff in 1931. However, when the US entered World War II he was counted as an enemy alien. He did, nevertheless, receive special permission to use the Mt Wilson telescopes, and, with the aid of red-sensitive photographic plates and the war-time black-out of Los Angeles (the city was already a significant source of light pollution for the observatory), Baade managed to resolve into stars the central regions of the Andromeda nebula. Baade went on to argue that there are two distinct types of stars in spiral galaxies, population I (young blue stars found in dusty regions) and population II (old red stars found in dust-free regions). He later decided that there are two different period–luminosity relationships and the determining factor deciding which of the two a Cepheid belonged to was its population type. A major consequence of Baade's findings was that the distances established by means of Cepheids of population I had to be increased by a factor of 2, one consequence of which was that astronomers decided to double the galactic distance scale.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clusters

Until 1926 the studies of the arrangement of the nebulae had been largely restricted to their distribution on the sky, but in that year Hubble made the study three dimensional when he examined the frequency distribution of diameters and apparent magnitudes (both being approximate indicators of distance) of some 400 extragalactic nebulae. He found that the numbers of nebulae to different limits (roughly, distance limits) were proportional to the volumes of space represented by the limits. Thus he argued there is an approximately uniform density of nebulae in all directions (with the exception of the Galaxy and the Virgo cluster of galaxies), given the assumption that the relative number of giant, normal and dwarf galaxies was the same in each of his samples. During the 1930s Hubble made further extensive studies of the space density of extragalactic nebulae as a function of the distance from our Galaxy. His method was to count extragalactic nebulae to various limiting magnitudes over widely spread areas of sky. Such counts confirmed his belief that, although nebulae are found singly, in groups and even occasionally in great clusters, when very large regions of space are compared the tendency to cluster averages out and one region is very much like any other.

Shapley, however, argued that Hubble was mistaken. In 1932, he published the Shapley–Ames catalog of galaxies, which, he reckoned, was complete to magnitude 12.9. To him, the catalog revealed a general lumpiness of distribution, with galaxies tending to form clusters and lumps, the most notable of which was the Virgo cluster. Shapley later extended the limits of his surveys to the 18th magnitude; again he reckoned clustering was a prominent feature. Indeed, by the early 1950s, clustering had become widely accepted as a major characteristic of the large-scale distribution of galaxies, and Hubble's hypothesis of a general field occasionally disrupted by groups or clusters of galaxies had fallen from favor. In 1952, G de Vaucoleurs, using the Shapley–Ames catalog and elaborating on a number of earlier speculations, also argued for the existence of clusters of clusters, or superclusters, to explain the distribution of bright galaxies on the sky.

Concluding remarks

At the start of the 20th century astronomers had believed that our Galaxy constituted the entire visible universe, but within 30 yr it had become accepted that space contains a multitude of visible galaxies. These galaxies, moreover, were found to display a redshift–distance relation, a relation that was generally taken to indicate that the universe is expanding. These triumphs convinced astronomers that future progress in cosmology would probably be dependent on the construction of ever more powerful optical telescopes. The giant 200 in reflector at Mount Palomar, completed in 1948, was in part a product of this belief. The redshift–distance relation and the models of an expanding universe had indeed produced an exchange between theory and data in studies of the large-scale properties of the universe that had been undreamt of in the early years of the century. Yet it was to be the unexpected extension of observations beyond the narrow range of optical wavelengths, to radio waves with wavelengths greater than 10 m and γ waves with wavelengths shorter than 10^{-12} m, that was to do most to fashion the next stage in the development of extragalactic astronomy.

Bibliography

Berendzen R, Hart R and Seeley D 1976 *Man Discovers the Galaxies* (New York: Science History Publications)

Hetherington N (ed) 1993 *Encyclopedia of Cosmology* (New York: Garland)

Hoskin M 1982 *Stellar Astronomy: Historical Studies* (Chalfont St Giles: Science History Publications)

Hubble E P 1936 *The Realm of the Nebulae* (New Haven, CT: Yale University Press)

Smith R W 1982 *The Expanding Universe: Astronomy's 'Great Debate' 1900–1931* (Cambridge: Cambridge University Press)

Smith R W 1990 Edwin P Hubble and the transformation of cosmology *Phys. Today* **43** 52–60

Robert W Smith

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extragalactic Background Light

Extragalactic background light (EBL) is the integrated flux from all extragalactic sources, including those which are not individually detected. In keeping with the COSMOLOGICAL PRINCIPLE—that the universe should appear homogeneous and isotropic to a typical observer—the EBL is expected to have a uniform mean level on large angular scales over the sky.

The cosmological significance of the EBL was first recognized as early as the 1700s, when astronomers imagined the universe to be a static volume of space, uniformly filled with stars. In this cosmological picture, the number of stars at a distance, r, increases as r^2, canceling the $1/r^2$ fall-off in the apparent brightness of a single star and making the total flux from objects populating every shell of space a constant value. This view led HEINRICH WILHELM OLBERS and others to conclude that the nighttime sky should be as bright as the daytime sky. The dilemma presented by the fact that the sky is (relatively) dark at night became known as OLBERS' PARADOX. This apparent conflict is now easily explained by the expansion of the universe, the finite speed of light and, most importantly, the finite lifetimes of stars. However, Olbers' paradox stands as an excellent illustration of the power of background measurements to test our model of the universe.

Ironically, while the night sky is far from dark, extragalactic sources contribute less than 1% of the sky brightness from the ultraviolet (UV) to the infrared (IR). Terrestrial, zodiacal and Galactic sources produce foreground emission which dominates the sky flux, making the diffuse EBL very difficult to measure, but also causing strong bias against the individual detection of faint or low surface brightness galaxies and intergalactic stars. Measurements of the diffuse EBL are immune to surface brightness selection effects. The EBL therefore provides a complete census of the energy density in the universe, making it a very powerful constraint on models of STAR FORMATION and the baryonic content of the universe.

Sources of the EBL

The majority of the EBL at UV to IR wavelengths is produced by stars at all REDSHIFTS since the first stars formed. In the far-UV (∼1000–1600 Å), Lyman continuum absorption by hydrogen in the intergalactic medium limits the objects contributing to the EBL to those at redshifts below $z \sim 1600$ Å/912 Å $- 1 \sim 0.8$. In addition, the UV light from galaxies is produced by the hot, massive stars, which are also responsible for most of the nuclear processing of hydrogen into helium and heavier elements (generally called 'metals'). The far-UV EBL thus traces the global star formation and metal production rate in the last half of a Hubble time.

The sources of the EBL become more complicated at optical (3000 Å–1 μm) and IR (1–1000 μm) wavelengths because the light we detect in those bands originates from objects over a wider range of redshifts and, consequently, emitted wavelengths. The optical EBL includes the direct emission from STELLAR POPULATIONS out to redshifts of $z \sim 10$. The energy density from stellar populations peaks at 1–1.5 μm at most stellar ages, so that the optical EBL will include both older populations at low redshift and young populations at high redshift.

Finally, much of the stellar emission in the universe will be contribute to the far-IR background, both because the light from the earliest generations of stars (most distant stars) will be redshifted to longer wavelengths and because UV–optical photons will be absorbed by dust grains in the emitting galaxy and re-radiated as thermal emission at infrared wavelengths ($\lambda \gtrsim 10$ μm). Dust attenuation of short-wavelength light is very efficient and may prevent more than 30% of the star formation in the universe from being directly detected at wavelengths shorter than 1 μm.

Non-stellar emission, such as the gravitational potential energy from accreting black holes and emission from decaying particles, will also contribute to the EBL. It is also possible that a significant fraction of the baryons in the universe might reside in a primordial INTERGALACTIC MEDIUM of ionized gas which might be identified by its contribution to the EBL. While hot gas at a temperature of $T \sim 10^6$ K would emit x-rays, and cool gas ($T \sim 10^4$ K) would produce dramatic absorption features in quasar spectra, gas at a temperature of $T \sim 10^5$ K would be detectable only by its contribution to the EBL through redshifted He II emission at 304 Å and Lyα at 1216 Å.

Measuring the EBL

Figure 1 shows the surface brightnesses of the foreground sources as a function of wavelength along lines of sight where these sources are faintest—at Galactic latitudes greater than 65° and ecliptic latitudes greater than 30°. For comparison, EBL detections and upper limits are shown in figure 2, along with models of the expected EBL and lower limits derived by summing the flux from resolved extragalactic sources (i.e. detected galaxies) and models of the EBL. At all wavelengths, the cumulative flux from foreground sources is roughly 100 times that of the EBL.

Ultraviolet (1000–2500 Å)

Because the atmosphere is opaque to photons with wavelengths shorter than about 3100 Å, UV observations must be made from rockets or from space. Early efforts to detect the far-UV background were inspired by the fact that the only bright foreground source documented at the time—ZODIACAL LIGHT—drops off sharply below 4000 Å and becomes insignificant below 2000 Å. However, the measurement was more difficult than expected and results varied by 3 orders of magnitude. With hindsight, it is clear that poor spatial resolution made subtraction of Galactic stars extremely difficult in these experiments and that the instrumental backgrounds were poorly understood. In addition, line emission from atoms and simple molecules (H I, O, and NO) at altitudes above 300 km and solar Lyα photons contributed significant flux to the rocket-based measurements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

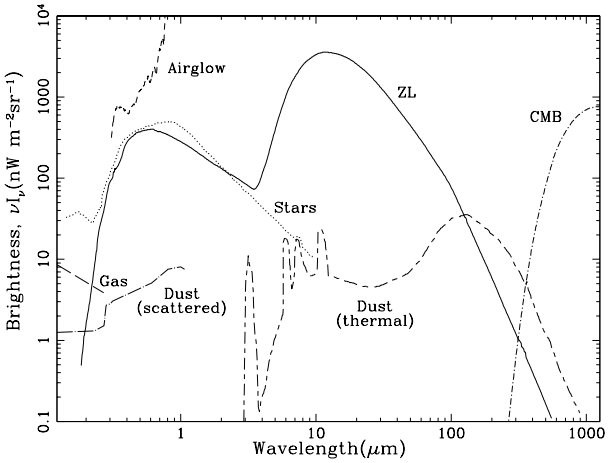


Figure 1. The flux from foreground sources is shown for lines of sight at Galactic latitudes greater than 65° and ecliptic latitudes greater than 30°, where foregrounds are faintest. The curve marked by short dashes at 0.3–1 μm shows the airglow contribution to ground-based observations: below 0.3 μm, the atmosphere is opaque; above 1 μm, the atmosphere is too bright for background observations. The solid curve shows zodiacal light, with the scattered component peaking around 0.6 μm and the thermal component peaking around 10 μm. The short-long dashed curve shows thermal emission from interstellar (Galactic) dust, with strong emission features around 3, 6, 8 and 10 μm. Scattered light at UV to optical wavelengths from the same dust is marked by a dot–dashed curve. The long dashed curve shows the Galactic UV contribution composed of atomic and molecular line emission and continuum emission from interstellar gas (see text). The dotted curve shows the flux from discrete Galactic stars fainter than $V = 6.0$ mag. The dot–short dashed curve shows the cosmic microwave background radiation (2.73 K).

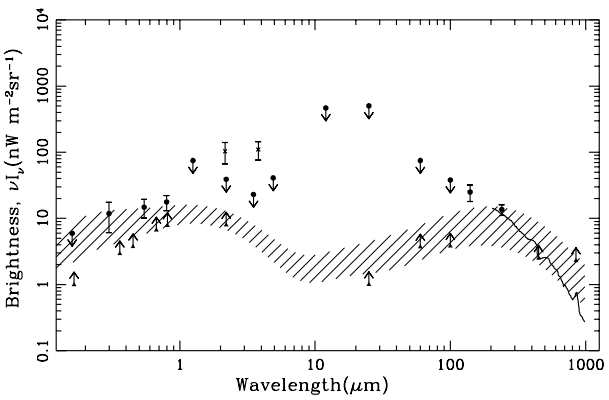


Figure 2. Measurements of the surface brightness of the EBL from the UV to the IR are indicated by filled dots and by the dark, solid curve. Upper limits are indicated with downward arrows. Upward arrows indicate lower limits to the EBL from the integrated flux of detected sources (galaxy counts). The hatch-marked region shows the range predicted by various models based on UV–optical source detections when dust obscuration is included (see Dwek *et al* (1998) for details).

Rapid progress came around 1980, with the advent of UV detectors with better resolution and greater sensitivity. Results in the last few decades have consistently shown that the majority of the UV background originates within the Galaxy, with only a small component being isotropic and therefore possibly extragalactic. The majority of the Galactic contribution is diffuse, dust-scattered starlight. The scattered flux can be estimated based on the intensity of the thermal emission from the same dust, which identifies the column density of dust, and simple scattering models incorporating empirically and theoretically determined dust properties. The previously poor agreement between modeling efforts now seems well explained by evidence that interstellar dust is strongly forward scattering. The scattered UV flux therefore depends on the scattering angle between the line of sight and the incident starlight from the center and plane of the Galaxy. Line emission from [O III] 5007 Å, C IV 1550 Å and O III] 1663 Å has also been detected in correlation with the column density of neutral hydrogen, as have H_2 fluorescence (1550–1650 Å), two-photon, bound–free and free–free continuum emission, and Lyα (1216 Å) and Hα line emission, all from warm ($T \sim 10^5$ K), high-latitude gas. The intensity of these components is proportional to the cosecant of the Galactic latitude, which approximates the path length through smoothly distributed gas.

Measurements and limits in the last 10 yr have converged towards EBL fluxes in the range 2–5 nW m^{-2} sr^{-1}, with a minimum estimate of 1 nW m^{-2} sr^{-1} from the cumulative flux in detected sources. Emission from galaxies is apparently responsible for the majority of this flux, while quasars and active galactic nuclei are estimated to contribute less than 0.2 nW m^{-2} sr^{-1} and flux from any intergalactic medium is limited to less than 0.2 nW m^{-2} sr^{-1} between 912 and 1216 Å.

Optical (3000 Å–1 μm)

The Earth's own atmosphere was the greatest impediment to early efforts at measuring the optical EBL. Atmospheric emission (AIRGLOW) is not only at least 100 times brighter than the EBL but also can vary by several per cent on timescales of minutes, as can atmospheric extinction. These effects make it nearly impossible to correct for atmospheric emission and extinction to the required accuracy. The only demonstratedly successful method for dealing with the atmosphere is to avoid it by conducting measurements of the EBL from space. The dominant foreground source contributing from beyond Earth's atmosphere is sunlight scattering off of the large ($\geq$10 μm), rough, INTERPLANETARY DUST grains which concentrate in the ecliptic plane. This scattered zodiacal light is also at least 100 times brighter than the flux of the EBL, even at ecliptic latitudes of $|b| > 35°$. As the solar system is harder to escape than the Earth's atmosphere, it is fortunate that the scattering involved is well described by Mie scattering theory and is only weakly wavelength dependent, with scattering becoming more efficient by ~5% per 1000 Å toward longer wavelengths. The mean zodiacal light flux

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can therefore be measured using the strength of solar absorption features (FRAUNHOFER LINES) which are preserved in its spectrum.

Optical CCD detectors typically have subarcsecond resolution, which makes resolution and subtraction of stars in our own Galaxy relatively easy. However, as in the UV, interstellar dust scatters incident starlight, producing diffuse Galactic light at optical wavelengths. Thermal emission from dust and the correlation of dust with neutral hydrogen can both be used to identify the lines of sight with the lowest dust column density, and simple scattering models can be used to estimate and subtract the low-level contribution which remains. The scattering properties of the dust are inferred from observations of the diffuse Galactic light at near-UV wavelengths where it dominates the sky and also from observations of a few neutral hydrogen (H I) clouds and planetary nebulae, where dust column densities are high.

Ground-based efforts to measure the optical EBL started in the 1970s. Airglow was identified either by geometrical modeling or by beam switching between a clear line of sight and a 'dark' dust cloud at high Galactic latitude in order to isolate the foreground from background sources. In 1980, the *Pioneer 10 Explorer* obtained data not only from above the atmosphere but from beyond the zodiacal dust cloud at a distance of 3 AU from the Sun. While it avoided both airglow and zodiacal light, the *Explorer*'s poor resolution made the subtraction of even discrete Galactic sources (i.e. stars) very difficult. As in the UV, rapid progress has been made with the advent of high-resolution detectors above the atmosphere aboard the *Hubble Space Telescope*. Detections of the EBL at 3000, 5500 and 8000 Å have recently been reported based on coordinated HST imaging (see HUBBLE DEEP FIELDS) and ground-based spectroscopy.

Infrared (1 μm–1000 μm)

Any warm body emits thermal radiation, including telescopes, the Earth's atmosphere and the Earth itself. Observations of the IR EBL must therefore be made from space with cryogenically cooled detectors. From space, thermal emission comes from dust trapped in the solar system and in the Galaxy. Subtraction of these foregrounds depends on spectral modeling using multiwavelength observations and geometrical modeling based on all-sky coverage. The most ambitious and successful IR measurement to date was undertaken with the *Cosmic Background Explorer* (*COBE*), launched in 1989. Prior to *COBE*, progress was made using short-lived, rocket-borne telescopes and satellites intended for observations of discrete sources, such as the *Infrared Astronomical Satellite* (*IRAS*).

Zodiacal light from scattering dies off at a wavelength of a few microns, but thermal re-emission of absorbed sunlight from the same dust peaks around 10 μm. This thermal emission is recognizable from the motion of the Earth in its orbit through the zodiacal dust. The annual variation of the sky brightness in all directions can be modeled empirically and subtracted. Once the thermal zodiacal component is removed, some individual Galactic sources can be resolved and subtracted. Flux from unresolved, discrete Galactic sources can be determined at each wavelength based on detailed statistical models of galactic sources and their spatial distribution. The diffuse emission from interstellar dust is brightest at around 100 μm. This component can be estimated and subtracted at wavelengths longer than 100 μm by scaling the zodiacal light-subtracted, 100 μm observations using a temperature model of the dust.

The *COBE* observations produced 3σ detections at 140 and 240 μm and 2σ upper limits at 100 μm and shorter.

Testing the big picture

The difference between the cumulative flux in optically detected sources and the detected optical EBL indicates that roughly 50% of the flux from individual sources is beyond the spectroscopic detection limit of current redshift surveys (see figure 2 and also GALAXY REDSHIFT SURVEYS). Recent results show that the difference may be partially explained by low surface brightness objects which are not included in standard catalogs and missed flux in low surface brightness regions of the objects which are catalogued. At longer wavelengths, however, the individual sources which produce the IR EBL have not been completely identified and may be powered by stellar or non-stellar sources. More generally, upper limits, lower limits and detections of the EBL all seem to indicate that the energy density in extragalactic sources is roughly constant over 4 orders of magnitude in wavelength (0.1–1000 μm). This spectral shape alone is a very interesting constraint on our models of the universe. The mean EBL at any wavelength is a complicated function of emission rates from stars, dust extinction, evolution of galaxies and the initial mass function and evolution of the stellar populations themselves. Meaningful interpretation of the EBL therefore requires modeling these variables as a function of redshift. Such models—which characterize our picture of the universe—are constructed from observations of resolved sources as a function of redshift.

One method for modeling the EBL is based on measurements of the rest-frame UV emission of individual galaxies, which indicates the star formation rate as a function of redshift and subsequently the EBL as a function of wavelength. Uncertainties in the dust attenuation of UV and optical light in the emitting galaxies limit the accuracy of these models. Other uncertainties include the stellar initial mass function and spectral energy distributions, which together identify the number of stars forming with a given mass and the spectral output from those stars. Models of the EBL based on this method need to include significant ($\gtrsim$30%) dust obscuration and re-radiation in order to provide reasonable agreement with observations of the infrared EBL. This necessity indicates that star formation regions at all redshifts are largely dust enshrouded and optically dark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The conversion rate of neutral hydrogen gas into stars as a function of redshift can be inferred from the statistics of absorption features in quasar spectra. This provides another method of modeling the EBL. As with the models based on UV rest-frame emission, the dominant uncertainty in these models is the fraction of optical light attenuated by dust and the characteristic temperature of the dust. As metals are produced through NUCLEOSYNTHESIS in the same hot, young stars whose output dominates the EBL, the mean EBL also can be predicted by the total metal mass density in the local universe. CLUSTERS OF GALAXIES are a convenient place to measure metal density, because any metals which may be ejected from galaxies will be trapped in the cluster potential. The rate of metal production can also be traced by absorption features caused by intervening systems in quasar spectra as a function of redshift. The total metal enrichment in the universe as identified by either method agrees with the total EBL energy density detected to date to within a factor of 2–5.

After more than 25 yr of work in the field, measurements or well-quantified upper limits have been found for the EBL over the full wavelength range from the UV to the IR. To within the measurement and modeling uncertainties, the results at all wavelengths produce a generally self-consistent picture of the total star formation rate in the universe, the total metals produced in stars and the total energy output from nucleosynthesis. Additionally, at wavelengths shorter than 1 μm, the diffuse background can be accounted for by sources which are well understood and largely present in surveys of individually detected galaxies. Efforts to better quantify the EBL UV and mid-IR wavelengths are underway in the form of new satellite missions and programs to identify the sources of the IR background.

Bibliography

Bowyer S and Leinert Ch (ed) 1989 *Galactic and Extragalactic Background Radiation (IAU Symposium 139)* (Dordrecht: Kluwer)

Calzetti D, Livio M and Madau P (ed) 1995 *Extragalactic Background Radiations* (Cambridge: Cambridge University Press)

Dwek E *et al* 1998 The COBE diffuse infrared background experiment search for the cosmic infrared background: IV. Cosmological implications *Astrophys. J.* **508** 106

Harrison E 1987 *Darkness at Night* (Cambridge, MA: Harvard University Press)

Leinert Ch *et al* 1998 The 1997 reference of diffuse night sky brightness *Astron. Astrophys. Suppl. Ser.* **127** 1

Rebecca A Bernstein

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extragalactic Reference Frames

The conceptual basis of reference frames defined by extragalactic objects is straightforward: that the universe as a whole does not rotate so very distant objects cannot have an overall rotational motion. Experimentally, the global rotation of the universe is less than 10^{-12} arcsec yr^{-1} as inferred from the 3 K microwave background radiation. At the distance of 10^8 parsec, even if an object were moving transversely at the speed of light, its angular velocity would be less than 0.6×10^{-3} arcsec yr^{-1}, while an object moving at a physically more reasonable speed comparable to the Sun would show a motion of 10^{-6} arcsec yr^{-1}, entirely undetectable by current technology. Since neither systematic universal motion nor random motion at such great distance is measurable, it is reasonable to construct a static celestial reference frame on the basis that such objects are fixed in the sky.

The idea of using extragalactic objects to define the celestial reference frame is quite old, having been discussed by Herschel and Laplace some 200 years ago, even before there was proof that such objects existed. The discovery of external galaxies and their distances in the early twentieth century raised the possibility of their use as fiducial objects for the celestial reference frame. However, their fuzzy, faint optical images made it quite difficult, and ASTROMETRY continued to use bright stars despite the complications of significant random and systematic motions. Fortunately, observations in the radio frequencies in the 1960s revealed a class of objects that has proven to be suitable for astrometry of the highest order, albeit with some limitations. These are QUASARS, compact radio sources which are pointlike in the optical band but with large REDSHIFTS implying cosmological distances.

Very long baseline interferometry

The technique for measuring the positions of quasars and similar objects is VLBI (very long baseline interferometry). In this application, VLBI is a geometric technique that measures directions in space with great accuracy. Schematically, using a pair of distant antennas, the difference in the time of arrival at the two stations of signals emitted by a radio source is measured with a precision of a few picoseconds. For a typical baseline of 5000 km, this corresponds to a fraction of a milliarcsecond. This time difference (delay) and its first derivative (delay rate) are reconstructed by offline correlation of the signals (actually radio noise usually buried in the thermal noise of the electronics) recorded at the two ends of the baseline. A key piece of equipment at the stations is the clock used to time events. VLBI makes use of hydrogen masers with stability of one part in 10^{14} for both time and frequency, and the raw data are recorded digitally at rates up to 1 Gbit s^{-1}. In current practice a network of up to 20 stations observes 50–80 radio sources around the sky multiple times during a 24 h session. Since a delay and a delay rate are generated from each pair of stations for each time a source is observed, there can be upwards of 20 000 delay/delay rate data points from one observing session. VLBI observations suitable for high-precision astrometry began in the late 1970s with the development of dual-frequency receivers and the Mark III VLBI data acquisition system, and an active worldwide observing program continues.

The analysis of VLBI data for astrometric and geodetic purposes requires the estimation of source positions, station positions, nutation corrections and parameters characterizing the behavior of the clock and atmosphere at each station as well as careful modeling of all geophysical and astronomical effects. Each separate session provides the relative positions of the observed sources with some uncertainty. On different days different sets of sources are observed. It is the overlap of some common sources from one day to another that connects all the positions together. From all the overlaps and directly measured positions come the complete set of relative positions. Repeated measurements, either directly in one day or indirectly from different days, improve the precision of the inferred positions. Having varied combinations of baselines and sets of observed sources strengthens the overall geometry and weakens the correlations between the source coordinates and the numerous phenomena modeled in the analysis. In particular, a mixture of north–south and east–west baselines is needed to determine both components of source position. Since VLBI measurements give relative positions, the orientation of the ensemble of positions, i.e. the pole and right ascension origin, must be determined from external information. This is a necessary step in assigning the numerical position values that define the reference frame. In practice the intermediate results are manipulated as the usual right ascension and declination values, but the fundamental VLBI measurements are of the arclengths between sources, the precision of which is contained in the full covariance matrix of the set of source positions. The earliest VLBI positions in 1971 were no better than a few arcseconds, but by the mid 1980s a few milliarcseconds had been achieved.

Unlike stellar astrometry, which must rely on a conventional theory of PRECESSION and NUTATION to transform true date coordinates to mean coordinates of the standard epoch, VLBI astrometry is sensitive to the errors in the standard precession/nutation model, which amount to several milliarcseconds. To achieve the full accuracy inherent in the stability of extragalactic positions, it is necessary to allow for corrections to the precession/nutation theory. This can be accomplished in two ways, either by adjusting the precession constant and various nutation coefficients or by estimating the offset between the actual pole and the modeled pole for each day of observation.

Foundation of the extragalactic reference frame

The process of laying the conceptual foundation for the extragalactic reference frame, identifying its fiducial objects and developing the fundamental catalog was begun at the Twentieth IAU General Assembly in Baltimore in 1988 by a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

resolution stating 'The International Astronomical Union should adopt a celestial reference frame based upon a consistent set of coordinates for a sufficient number of extragalactic objects when the required observational data have been successfully obtained and appropriately analyzed. ... This reference frame is likely to be based, initially at least, exclusively upon radio astrometry.' The Working Group on Reference Systems headed by J Hughes was established with subgroups for time (B Guinot), coordinate frames and origins (J Kovalevsky), astronomical constants (T Fukushima), and the theory of nutation (D McCarthy). The fundamental concepts (see also SPACE AND TIME REFERENCES: CONCEPTS) of what is now the ICRF (International Celestial Reference Frame) were developed by the subgroups and extensively discussed at the IAU Colloquium 127 on Reference Systems held in Virginia Beach in 1990. The general theory of relativity for space and time, the use of extragalactic radio objects, and the continuity between the existing stellar and the anticipated extragalactic celestial reference frames were summarized in nine recommendations adopted as a resolution by the Twenty-first IAU General Assembly in Buenos Aires in 1991. A working group under C de Vegt was formed to select the radio sources to be used with emphasis on the need to connect the radio and optical frames. The Twenty-second IAU General Assembly in The Hague in 1994 adopted the list of some 600 sources and established a new working group under L Morrison to define the positions. This work was done by a subgroup of VLBI and celestial reference system specialists. The resulting ICRF catalog was adopted by the Twenty-third IAU General Assembly in Kyoto in 1997 and came into effect on 1 January 1998, replacing the FK5. A Working Group for the International Celestial Reference System (ICRS) under F Mignard was formed to carry on the ICRF work, to extend the frame to other objects and frequencies, and to elucidate the consequences of the ICRF.

Definition of the ICRF extragalactic reference frame

According to the new adopted rules, the fundamental directions of the celestial reference system will remain fixed in space; they will no longer be dependent on modeling the motion of solar system objects. The fiducial objects will be monitored. The adopted positions may be revised when improved information is available, but the coordinate axes will be maintained by implementing the statistical condition that the coordinates of selected defining sources show no global rotation from the old set of coordinates to the new realization.

The data used to determine the ICRF radio source positions included 1.6 million pairs of dual-frequency delays and rates from 1979 to the middle of 1995. The data were contributed from several geodetic and astrometric observing programs, the vast majority of the observations coming from the former but most of the sources coming from the latter. The design of the analysis was intended to provide (1995) state of the art source positions, in particular to free the results from contamination from precession/nutation errors, anomalous source behavior, geodetic noise and systematic atmospheric effects. Simultaneous dual-frequency observations provided the calibration of ionospheric dispersion. Celestial pole offsets in obliquity and longitude were estimated for each day. Positions of sources with excessive scatter or apparent motion (as determined from test solutions) were estimated separately for each day as were station positions. An elevation limit of 6° was adopted to balance the need for maximum observing geometry against the difficulty in correctly modeling the atmosphere at low elevations. Atmospheric gradients were estimated to remove a significant systematic effect. Both delays and rates were used to maximize the information available and to decrease correlations.

Because of the large number and high quality of observations, the formal errors for the positions were exceedingly small and not a meaningful measure of uncertainty. It was necessary to consider several other effects to assign realistic errors. One was the statistical validity of the formal errors. Another was the cumulative influence of all modeling errors and editing decisions. Yet another was the magnitude of specific, identifiable errors that could have distorted the results. A consideration of all known error sources indicated that a realistic error estimate could be made by inflating the formal errors by a factor of 1.5 followed by a root-sum-square increase of 0.25 milliarcsecond. For the most frequently observed sources 0.25 milliarcsecond was the dominant error.

The sources from the ICRF solution were divided after the fact into three categories: 212 defining sources that met a series of quality criteria; 294 candidate sources that failed one or more of the quality criteria, e.g. too few observations or observing interval too short, and 102 'other' sources (whose positions were estimated each day). While the distribution over the sky of the ICRF as a whole is quite uniform, the defining sources are sparser in the southern hemisphere because of a dearth of VLBI stations and limited observing programs. The 'other' sources include a number with relatively bright optical counterparts that could be used to connect the radio and stellar frames.

The axes of the frame were aligned to the International Earth Rotation Service (IERS) realization of the ICRS with an error of $\sim$0.020 milliarcsecond. The Hipparcos catalog, which is the optical realization of the celestial reference frame (see HIPPARCOS AND TYCHO CATALOGUES), was aligned to the ICRF, largely through VLBI observations of radio stars. The ICRF also provides the framework for various space geodetic techniques to measure the terrestrial positions and velocities that constitute the terrestrial reference frame and to monitor variations in the rotation of the Earth.

Future improvements

The two chief phenomena leading to random and systematic errors in the ICRF are the atmosphere and source structure. While the propagation delay at zenith from the dry atmosphere can be accurately calibrated from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

local meteorological data, the dry delay in the line of sight to the observed source is subject to modeling error. The delay from water vapor cannot be calibrated accurately, and both delays are subject to random temporal and spatial fluctuations that must be estimated from the VLBI or other microwave data.

Quasars are not point radiators. Source structure can cause their position to appear slightly different depending on baseline geometry, and changes in source structure can cause apparent movement. Detailed source maps show that intrinsic variations in structure can be quite extreme, ranging from relatively compact naked-core objects, to compact double sources, to complex core-jet objects. The situation is complicated by the fact that compact extragalactic RADIO SOURCES are known to have variable intensity and to have frequency- and time-dependent intrinsic structure. Consequently, unknown and unmodeled source structure effects may be introduced into the reference frame. The sources can be characterized after mapping by the degree to which the delay and delay rate are affected by the structure over the full range of possible baseline geometries, and the worst sources can be rejected for most astrometric purposes.

As of the beginning of 1999, about 800 000 more observations including about 50 new sources had been acquired by geodetic and astrometric programs since the ICRF was constructed in 1995. These were integrated with the data and analysis used for the ICRF to update the original catalog. The original ICRF analysis was not copied exactly to incorporate new data, but the new positions are consistent with the ICRF at its stated level of accuracy, about 0.25 milliarcsecond in overall source positions and about 0.020 milliarcsecond in orientation. Positions were determined for new sources, and updated positions and errors were found for candidate sources with additional observations. The positions and uncertainties of the defining sources were not changed since they are the actual realization of the ICRF, and the work to include more recent data does not lead directly to the replacement of the existing ICRF. The defining sources will be used in a no-net-rotation condition relative to the ICRF to ensure that positions from new analyses are aligned with the ICRF.

There are several features that distinguish the ICRF from conventional stellar catalogs that formerly defined the celestial reference frame. First, the quasars are optically faint so connection to the stellar frame is complicated. Second, while the positional history of the radio sources is known, the results of future observations cannot be predicted. Consequently, new sources must be observed to replenish and expand the list of candidates. Third, as new observations accumulate, it should be possible to move candidate sources up or down the scale of usefulness in subsequent ICRF realizations. The problem of position changes related to source structure variation may be solved in the future if the application of source structure information permits the identification and use of truly kinematically static points in the sky. Since the original VLBI delay and delay rates are accessible, they can always be reused in improved analysis. When models and estimation techniques are significantly improved and sufficient data are accumulated, a new ICRF can be constructed *de novo*.

Bibliography

Ma C, Arias E F, Eubanks T M, Fey A L, Gontier A-M, Jacobs C S, Sovers O J, Archinal B A and Charlot P 1998 The International Celestial Reference Frame realized by very long baseline interferometry *Astron. J.* **116** 516–46

Sovers O J, Fanselow J L and Jacobs C S 1998 Astrometry and geodesy with radio interferometry: experiments, models, results *Rev. Mod. Phys.* **70** 1393–454

Chopo Ma

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extrasolar Planetary Companions and Brown Dwarfs

The first convincing detection of both extrasolar planets (EXOPLANETS) and brown dwarfs came during the 1990s[1]. The first planetary system found around another star is actually a very special case: the host is a 6.2 ms pulsar, a billion year old NEUTRON STAR which should formerly have erupted as a supernova. From precise timing of the radio pulses, A Wolszcan and D A Frail determined the orbits of three inferred terrestrial-mass companions in 1992. The leading model for the formation of pulsar planets involves post-supernova capture of a small amount of material from a preexisting companion into an ACCRETION DISK. The apparent existence of planets in such an exotic environment may tend to indicate that planet formation can take place under a wide variety of conditions, but we do not consider this case further here.

The inferred existence of the first Jovian-mass extrasolar planet in a normal stellar environment was found by detection of periodic radial velocity variations in the optical spectrum of the Sun-like G star 51 Peg; the discovery of 51 Peg B was announced in 1995 by M Mayor and D Queloz of the University of Geneva. Over the next few years several groups utilized this the Doppler technique to report dozens of discoveries.

In the same few-month period of 1995, the first bona fide brown dwarf—instantly recognizable as a substellar object—was reported by a group of Johns Hopkins University and California Institute of Technology scientists (see bibliography). In contrast to the first case, the latter was a direct detection of the brown dwarf as a resolved, distant companion to the nearby M dwarf Gliese 229. Since this object, called GL 229B, features methane (CH_4) rather than carbon monoxide (CO) in its infrared spectrum, its surface temperature was inferred to be near 1000 K—much too cool to be a star. Methane is also the dominant form of carbon in the atmosphere of Jupiter. In figure 1 we display a recently published infrared spectrum of GL 229B with the methane and water bands labelled. For comparison are shown spectra at similar wavelengths of the late M dwarf vB10 and the L dwarf GD 165B (see the final section). These considerably warmer objects show H_2O and CO, but not CH_4.

In retrospect, as will be discussed, GL 229B may not have been the first brown dwarf to be discovered. Several different techniques have yielded candidate substellar-mass objects over the last 10 yr. In the next section we summarize the techniques that have been tried or are planned for the future. Then we summarize briefly some findings and their implications as of early 1999. In figure 2, kindly made available by Dr Alan Boss, the distribution of most known extrasolar planets and brown dwarfs as of early 1999 is shown as a plot of mass (in M_J) versus orbital semimajor axis (in AU). Other features of this plot will be discussed subsequently.

Direct detection techniques

The direct detection of GL 229B was made possible by its photospheric radiation. In a later section the discoveries of more luminous brown dwarf candidates in both young star clusters and the field will be discussed. It is obviously a much more difficult task to detect directly a cold planet. In the visible the reflected light from a Jupiter analog is typically $\sim 10^{-9}$ that of the primary star. In the thermal infrared (5 μm and greater), this contrast may be as small as 10^{-4} for an old planet. Direct imaging of a planetary or cold brown dwarf companion requires that the noise in the point spread function of the primary's light profile be reduced accordingly. As of this writing, only a few brown dwarf companions (and no planets) have been detected directly in this way. For the future, ADAPTIVE OPTICS systems with a coronographic capability are being developed for the new generation of 8–10 m telescopes.

For example, in 2001, The ADONIS (Adaptive Optics Near Infrared System) instrument mounted at the European Southern Observatory (ESO) 3.6 m telescope at the La Silla Observatory has obtained an image of a dust disk around the star iota Horologii, already known to possess a planet at least twice as heavy as Jupiter in an Earth-like orbit. The dust disk around iota Horologii is quite extended; it is detected to a distance of about 65 AU (10 000 million km) from the star. From the elongation, it appears that the inclination of the disk is about 42º.

Also promising for direct detection of planets is ground-based interferometric imaging in the infrared (see also INTERFEROMETRY: GROUND). Direct imaging of Earth-like planets will require development of a space-borne interferometer with a baseline of ~100 m and perhaps 4 m apertures.

Astrometric detection

The astrometric signature of an orbiting planet is a periodic wobble proportional to both the planet's mass and its orbital radius (see ASTROMETRY: TELESCOPES AND

[1] For this discussion, a brown dwarf is defined as a configuration incapable of sustained hydrogen fusion reactions, with a mass between about $0.01M_\odot$ and the stellar mass limit of $(0.075–0.08)M_\odot$. These values correspond very nearly to 10 and 75–80 Jupiter masses (M_J), respectively. A giant planet is defined as having a mass less than about $10M_J$ and thus is incapable of even limited nuclear burning of deuterium. In addition, theoretical calculations tend to favor a lower mass limit for formation of isolated entities (brown dwarfs) of the order $10M_J$. An upper limit to the mass of a Jovian planet can be inferred from the limited amount of gas in a circumstellar disk and its finite lifetime; this has been suggested to be not more than several Jovian masses.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

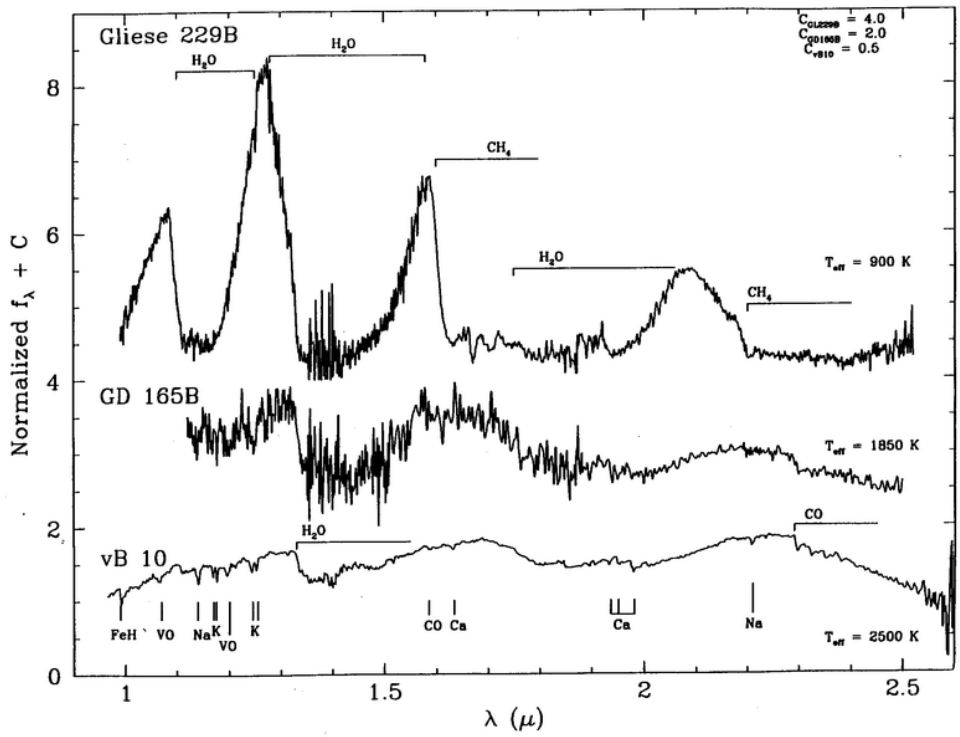


Figure 1. The near-IR spectrum of GL 229B compared with spectra of vB10 (M8 V) and GD 165B (L2 V, see the last section of this article). The great strengths of the CH_4 and H_2O bands in GL 229B give it a blue J–K color. From Oppenheimer B R, Kulkarni S R, Matthews K and van Kerkwijk M H 1998 *Astrophys. J.* **502** 932.

TECHNIQUES). It is therefore nicely complementary to the Doppler radial velocity technique which is biased in favor of nearby companions; this technique is also insensitive to the viewed orbital inclination. A Jupiter at 4 AU from a solar-type star 10 pc away would produce an astrometric amplitude of only 0.5 milliarcsec (mas). The sensitivity of an astrometric measurement to 1 mas is shown as the long–short dashed diagonal line in figure 2, from top left to bottom right. Objects falling to the right of the line are detectable by this method, if the orbital periods are not prohibitively long. Jupiter by itself would be a marginal detection in a solar system clone. In the future as astrometric precision is refined from milliarcseconds down to microarcseconds (from space), this technique can become very powerful. The reality of the astrometric companion alleged for Lalande 21185 (figure 2) is now questionable. To date no definitive detection of a planetary companion has been accomplished by astrometry.

Photometric detection

Several ground and space programs and proposals are underway which will attempt to detect planets and brown dwarfs undergoing transits of a parent star. The Jovian planet will dim a G star primary by the order of 1% (the ratio of surface areas), a terrestrial planet by only 0.01%. Another photometric signature to be exploited is gravitational MICROLENSING. With suitably high time-resolution sampling, the microlensing surveys can reveal the photometric signatures of foreground substellar objects in the light of background stars such as those observed in the crowded fields of the GALACTIC BULGE.

Detection by periodic radial velocity variations

As discussed in the introduction, it is the Doppler technique which has yielded bountiful discoveries of extrasolar planets in several years of work. The general expression for the semiamplitude, K, of the stellar radial velocity induced by an orbiting companion is

$$K = \frac{2\pi G}{P^{1/3}} \frac{m_{\rm p} \sin i}{(M_* + m_{\rm p})^{2/3}} \frac{1}{(1-e^2)^{1/2}}.$$

Newton's refinement of Kepler's third law then yields $m_{\rm p} \sin i$ and the orbital radius a (see KEPLER'S LAWS)[2].

[2] Originally, it was also proposed that some mechanism other than orbital motion was responsible for the apparent periodic Doppler signature. However, subsequent research has shown that orbital motion is the only plausible interpretation, despite the unexpectedly short orbital periods implied for some of the planetary companions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

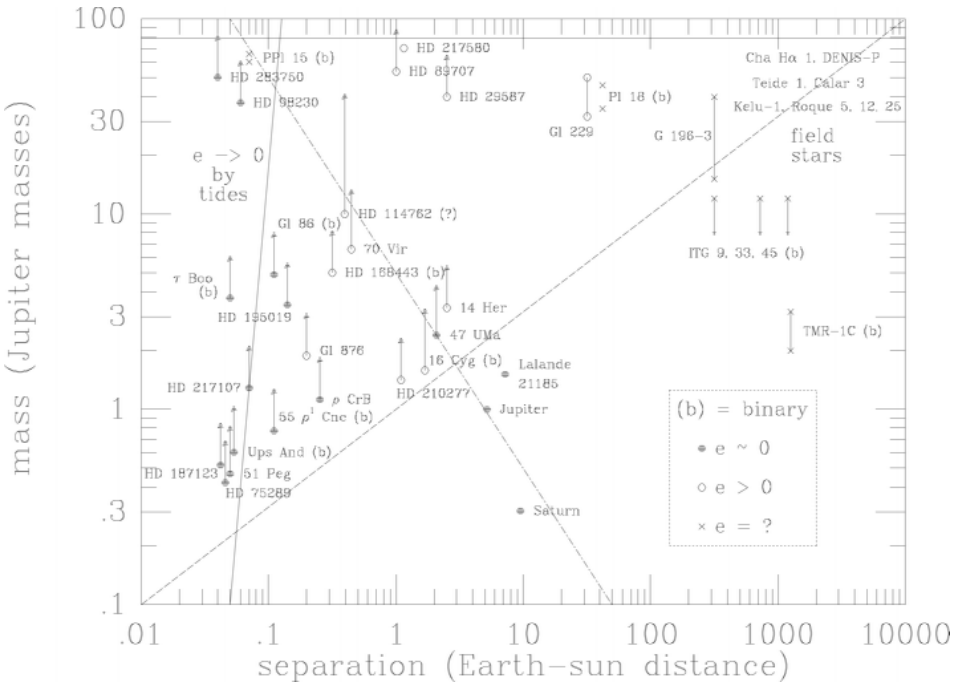


Figure 2. Extrasolar planets and brown dwarfs known as of early 1999, masses (M_J) versus separation (in AU). Arrows indicate that most mass estimates determined from the Doppler technique are lower limits. Filled, open circles and ×s indicate circular, eccentric and unknown eccentricity orbits. Jupiter and Saturn are also plotted. Some field brown dwarfs are plotted arbitrarily in the upper right with rough mass estimates. To the left of the solid line orbits should be circularized owing to tides (51 Peg B objects). The dashed line gives sensitivity for Doppler amplitude of 30 m s^{-1}. The nearly perpendicular long–short dash line gives complementary sensitivity for an astrometric wobble of 1 mas.

The high accuracies to which K is now measured are due to the development of precise spectrographs designed for this purpose (see SPECTROGRAPHS: HIGH-RESOLUTION SPECTROGRAPHS). The Sun orbits the barycenter of the solar system with a typical velocity of only 13 m s^{-1}. A few groups have achieved measurement errors as low as 3 m s^{-1}, accurate enough to detect the effects of Jupiter and Saturn at least marginally if they were to observe a solar system clone. Their various techniques involve superposing the wavelength calibration scale directly on the stellar spectrum, to remove even small effects of spectrograph flexure when the telescope moves. The accuracy of the Doppler technique in ferreting out periodic radial velocity variations in the spectra of solar-type stars is limited directly by the 'noise' due to motions in the stellar atmospheres. The dispersion of velocity measurements increases linearly with the stellar rotational velocity, which is known to correlate with general stellar activity. An active star may even display a Doppler-like velocity signature at the stellar rotation period.

The dashed line approximately perpendicular to the astrometric line in figure 2 maps the sensitivity of a radial velocity amplitude of 30 m s^{-1}; this shows the complementary nature of these techniques. Since virtually all extrasolar planet discoveries were made using the Doppler technique, the known cases fall above and to the left of this line. The arrows indicate that the mass determinations are always lower limits, since the orbital inclination is generally unknown. Thus many possible brown dwarfs in the upper left may in reality be low-mass stars, the result of selection bias. In fact the evidence currently is that there are few brown dwarf companions found close to stars in the solar neighborhood. Some objects in the upper right include single brown dwarf candidates to be discussed later.

Properties of extrasolar planets

The preceding discussion indicates that the current list of known objects is biased by the selection constraints of the Doppler and astrometric methods. In particular the sensitivities to Jovian planets at the separations observed for those in our solar system are marginal, and the time baselines are still too short for even one orbital period to be fully sampled. That Jovian planets may be found inside of 1 AU separation is now obvious, and these of course are found most easily. Two distinct groups deserve special discussion.

One group including the first object 51 Peg B lie in circular orbits of less than 0.1 AU separation. Objects falling to the left of the solid curve near 0.1 AU in figure 2 lie so close to the parent star that the orbits must be circularized, as is observed to be the case.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Theoretical work on Jovian planet formation argues strongly that they could not accrete and grow to Jovian mass in the hot inner region of the circumstellar nebula. It is believed that they must have migrated inwards before the gaseous phase of the disk ended, since the dissipation of the orbital angular momentum depends on the presence of viscous gas. It is possible that some Jovian planets accrete onto the host star and may even enhance the surface abundance of some elements.

A second peculiar group are the Jovian planets and possible brown dwarf companions in eccentric orbits. These lie to the right of the tidal circularization line, as they should. A planet with the most elongated orbit detected so far (September 2001) is HD 80606, which moves between 5 and 127 million km from the central star. Since the standard model of accretion in a gaseous disk to form a massive planet should circularize its orbit, it is likely that a dynamical interaction of two of these bodies occurred. It is also possible for a planet to be ejected from the system.

The upsilon Andromedae solar system includes both types: there is a close $>0.7M_J$ planet of period 4.6 days in a circular orbit, a $>2M_J$ planet with a 241 day period in a mildly eccentric orbit and a $>4.6M_J$ planet in an eccentric orbit with a 3.5 yr period. In summary, the distribution of objects plotted in figure 2 is certainly biased heavily, but the discoveries nonetheless demonstrate how versatile the formation of solar systems can be.

Brown dwarfs in young clusters and star-forming regions

Young, nearby clusters have offered the opportunity to detect brown dwarfs when they are relatively luminous and at a predicted luminosity and age. Primarily on the basis of ground-based surveys at wavelengths of 0.8–2 μm, dozens of brown dwarf candidates have been found in photometric surveys of several young CLUSTERS such as the Pleiades and star-forming regions in Orion and rho Ophiuchus. The systemic membership of the candidates may be confirmed by astrometric and radial velocity measurements, as well as the fact that the candidates fall on a sequence in a color–magnitude diagram predicted by models of gravitationally contracting, protostellar objects. In 2001, observations by the European Southern Observatory's New Technology Telescope revealed dusty disks surrounding brown dwarfs in the Orion Nebula, suggesting that they are more star-like than planet-like and have the potential to form accompanying planetary systems. This finding is based on detailed observations with SOFI, a specialized infrared-sensitive instrument at the European Southern Observatory's (ESO) 3.5 m New Technology Telescope (NTT) at the La Silla Observatory.

The detection of lithium is another important constraint on the existence of objects below the stellar mass limit. In particular, lithium is believed to undergo destructive fusion reactions at an interior temperature ($\sim(2–3)\times10^6$ K) slightly below the limit for proton–proton fusion. Fortunately, calculations of the interior structures of very-low-mass stars and brown dwarfs predict that energy is transported primarily by convective motions from the centers to the very surfaces. This means that, if lithium is hot enough to be depleted at the center where the temperature is highest, it will eventually be depleted throughout the configuration. Therefore, if the resonance lines of Li are observed in stellar atmospheres (at the strength expected for the abundance at which the gaseous body should have formed), hydrogen is also not being fused in the deep interior—i.e. the configuration is a brown dwarf. Numerous such detections of the Li lines have been reported, and an Li luminosity boundary consistent with theory has been found in the Pleiades. This 'lithium test' thus enhances the likelihood that Pleiades substellar candidates Teide 1, Calar 3 and the Roque objects (figure 2, upper right) are part of the substellar extension of the Pleiades mass function as low as $30M_J$.

Brown dwarfs in the local Galactic disk

Follow-up observations of 'field' point sources detected in the near-infrared sky surveys DENIS (Deep Near Infrared Survey) and 2MASS (Two Micron All Sky Survey) have shown that brown dwarfs exist in the field as single objects. 2MASS in particular has discovered ~100 brown dwarf candidates (as of this writing). While warmer than GL 229B, these infrared-selected objects are cooler than the M dwarf stars, with $T_{\rm eff} < 2000$ K.

In 2001, two examples have been discovered of small, cool stars that resemble brown dwarfs but are actually the remnants of ordinary stars that have been whittled down to cool Jupiter-sized bodies over billions of years by spilling material over to a white dwarf companion star. Using the UK Infrared Telescope (UKIRT) in Hawaii, the first direct evidence of such stars was obtained by taking infrared spectra with the CGS4 instrument of two variable binary star systems: LL Andromedae and EF Eridani. The observers took advantage of periods when the flow of material between the two stars in these binary systems temporarily stops. At these quiescent times, UKIRT can distinguish the radiation coming from the cool donor star. In the case of LL Andromedae, the signature of methane was detected at a wavelength of 2.2 microns. This shows that the donor star's temperature is around 1300 K, similar to a 'T-type' methane brown dwarf. In EF Eridani, the donor star is a little warmer, at around 1650 K, similar to an'L-type' brown dwarf. According to theory, the estimated mass of these cool stars is near four-hundredths the mass of the Sun, or 40 times the mass of the planet Jupiter, with estimated distances of between 100 and 130 light years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

An assessment of a complete subsample of 2MASS L dwarfs suggests that substellar objects are comparable with or outnumber the stars in the solar neighborhood, though contributing only modestly to the mass of the local disk (see GALACTIC THIN DISK). It appears that only young brown dwarfs—or those relatively few likely to have masses just below the stellar mass limit—are observed in the L dwarf temperature range and are readily detectable by 2MASS and DENIS. This implies that the great majority have cooled to much lower temperatures. Indeed, field objects showing methane in the spectra and with temperatures closer to that of GL 229B (~1000 K) have now also been found, by the Sloan Digital Sky Survey (SDSS) and by 2MASS. It may well turn out that the majority of the Sun's substellar neighbors have cooled to temperatures of several hundred kelvin, cooler even than GL 229B.

Breaking news update (31 January 2002)

Observations by NASA's Hubble Space Telescope of the gas-giant planet orbiting the yellow, Sun-like star HD209458, 150 l.y. from Earth have revealed the presence of sodium in light filtered through the planet's atmosphere when it passed in front of its star as seen from Earth. The planet was discovered in 1999 by its subtle gravitational pull on the star. The planet is 70% the mass of Jupiter. Its orbit is tilted nearly edge-on to Earth, which allows repeated transit observations. The planet is only about 6 million km (4 million mi) from the star. At this distance the planet's atmosphere is heated to around 1000 K.

At the time of this update 66 extrasolar planetary systems are confirmed, consisting of 74 planets and 7 multiple planets systems. See http://www.obspm.fr/encycl/catalog.html. Among recent discoveries are three Jupiter-sized planets travelling in circular orbits similar to that of Earth and Mars in our own solar system.

Bibliography

A thorough review of the discovery of extrasolar planets and the implications as of about 1997 is given by

Marcy G W and Butler R P 1998 *Annu. Rev. Astron. Astrophys.* **36** 57

parts of this article borrow from the format and content of that review. An excellent theoretical review of the formation of protostellar disks and the diversity of possible planets is

Lissauer J J 1995 *Icarus* **114** 217

A review of prospects for space-based interferometric imaging of terrestrial planets is given by

Woolf N and Angel J R P 1998 *Annu. Rev. Astron. Astrophys.* **36** 507

The discovery of the first extrasolar Jovian planet was reported in

Mayor M and Queloz D 1995 *Nature* **378** 355

The discovery of GL 229B is published in

Nakajima T, Oppenheimer B R, Kulkarni S R, Golimbowski D A, Matthews K and Durrance S T 1995 *Nature* **378** 463

and in

Oppenheimer B R, Kulkarni S R, Matthews K and van Kerkwijk M H 1995 *Science* **270** 1478

The first isolated or 'field' methane dwarf, SDSS1624 + 0029, is published in

Strauss *et al* 1999 *Astrophys. J.* **522** L61

James W Liebert

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extrasolar Planets

The discovery of the first planets around a PULSAR (1992) and around a main sequence star different from our Sun (1995) has opened, after centuries of speculation, a new era in astronomy. The motivations for the search for 'other worlds' are presented. A comparative review is given of different techniques for their detection. Special attention is paid to the planetary parameters which can be derived in each method. The first discoveries and their astrophysical consequences are presented. A review of new questions and future projects beyond these first discoveries is given.

Introduction

The search for 'other worlds' is one of the oldest well-formulated scientific questions. It was raised for the first time in 'modern' terms by the Greeks (Democritus, Epicurus). Since then, it has inspired many thinkers (Giordano Bruno, Fontenelle, Goethe, Kant, …), but it was pure speculation. There had been a first attempt in the 17th century by HUYGENS (1698), who tried to see planets around other stars. He quickly realized that this observation was far beyond the capabilities of his modest telescope. After that, for centuries, seeing such planets could only be a dream. A new hope appeared in the early 20th century after the astrometric and spectroscopic detection (explained further in this article) of 'unseen' companions to stars (i.e. very faint stars). However, the expected planetary masses were to small to be reached by these methods. Astronomers had to become patient again. A slim hope was raised by the Dutch astronomer P VAN DE KAMP in the 1950s: he thought that he had detected a planet around BARNARD'S STAR (one of our closest stellar neighbors), but it was later proven to be an artefact. The first important step was made by Frail and Wolszczan when they detected the first planets around a pulsar. It was a great encouragement for planet hunters, although the star itself (a neutron star) was too odd to resemble our Sun. The good news was heard in October 1995, when the Swiss astronomers Mayor and Queloz claimed the detection, thanks to the RADIAL VELOCITY method using a telescope at the Haute Provence Observatory (south of France), of a planet around the star 51 Peg, similar to the Sun. Understandably enough, there was some skepticism after decades of mistakes, but only for a few weeks. This great discovery makes us live in a very special epoch, where progress on the 23-century-old question 'Are we alone in the universe?' has made a major step and the question can be investigated scientifically.

This first discovery was followed by many others; since then, more and more planets have been discovered (by radial velocity measurements of the parent star) each year and there are at present (September 2001) 66 of them (see also EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS). Table 1 gives the current list (updated frequently on the internet (Schneider 1999)). Furthermore, for one of the stars (upsilon Andromedae), three planets have been detected. For another star, one could measure the radius of the planet (30% larger than Jupiter); this observation is a definite proof that the stellar wobble is indeed due to a giant, gaseous, planet.
In 2001, a two-planet system was detected (around the star HD 82943) in which one orbital period is nearly exactly twice as long as the other. HD 74156 is another two-planet system with a Jupiter-like planet and a more massive planet farther out. A giant planet moving in an orbit around a Sun-like central star (HD 168443) already known to possess a planet has been detected, as well as a planet with the most elongated orbit detected so far (HD 80606), moving between 5 and 127 million kilometers from the central star.

In one of the systems, 55 Cnc, a dust disk has been found around the star, first as an IR excess by ISO (Dominik *et al* 1998), and subsequently as an image at the IRTF (Trilling and Brown 1998). The inclination of the disk as deduced from the image shape is ~25°.

After this turning point it is possible to go further, and astronomers are now willing to address several questions.

1. How frequent are other planetary systems?
2. Are they like ours (number, masses, radii, albedos of the planets; radii, eccentricities, inclinations of their orbits)?
3. What types of environments (atmospheres, magnetospheres, rings etc) do they have and what are their properties?
4. What are the other physical properties of planets (rotation period, atmospheric circulation, presence of oceans, …)?
5. How do they form and evolve?
6. How do these features vary with the type of the central star (mass, spectral type, chemical composition, age, binarity etc)?

In addition to these purely astrophysical questions, there is a related query, shared by every human: 'is there life elsewhere in the universe?' This question raises the problem of the different forms of life we can expect outside the solar system and the problem of the definition of life. The convential view is that life can develop on Earth-like planets in the 'habitable zone' (where the temperature is about 300 K) around stars. Beyond the detection of these planets, it is even possible to search for spectroscopic signatures of life on these 'habitats'.

The detection methods for extrasolar planets

To investigate these planetary systems, astronomers use several methods of detection and observation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Extrasolar planets.

Star	M_{min} (Jupiter masses)	Period (days)	Semimajor axis (AU)
HD 83443	0.35	2.9861	0.038
HD46375	0.25	3.024	0.041
HD 187123	0.54	3.097	0.042
Tau Boo	4.14	3.313	0.047
BD-10 3166	0.48	3.487	0.046
HD 75289	0.46	3.508	0.048
HD 209458	0.63	3.524	0.046
51 Peg	0.46	4.231	0.052
Ups And b	0.68	4.617	0.059
Ups And c	2.05	241.3	0.828
Ups And d	4.29	1308.5	2.56
HD 168746	0.24	6.409	0.066
HD 217107	1.29	7.130	0.072
HD 130322	1.15	10.72	0.092
HD 108147	0.34	10.881	0.098
55 Cnc	0.93	14.66	0.118
GJ 86	4.23	15.80	0.117
HD 195019	3.55	18.20	0.136
HD 192263	0.81	24.35	0.152
Rho CrB	0.99	39.81	0.224
HD 168443	8.13	58.10	0.303
GJ 876	2.07	60.90	0.207
HD 16141	0.22	75.80	0.351
HD 114762	10.96	84.03	0.351
70 Vir	7.42	116.7	0.482
HD 52265	1.13	118.96	0.49
HD 1237	3.45	133.8	0.505
HD 37124	1.13	154.8	0.547
HD 89744	7.17	256.0	0.883
HD 134987	1.58	260.0	0.810
HD 169830	2.96	230.4	0.823
HD 12661	2.83	264.5	0.825
Iota Hor	2.98	320.0	0.970
HD 177830	1.24	391.0	1.10
HD 210277	1.29	436.6	1.12
HD 82943	2.24	442.6	1.16
HD 222582	5.18	576.0	1.35
16 CygB	1.68	796.7	1.69
HD 10697	6.08	1074.0	2.12
47 UMa	2.60	1084.0	2.09
14 Her	5.44	1700.0	2.84
Confirmed pulsar planets			
PSR 1257+12	0.015 $M_\oplus$	25.34	0.19
	3.4 $M_\oplus$	66.54	0.36
	2.8 $M_\oplus$	98.22	0.47
	~100 $M_\oplus$	~170	~40
PSR B1620-26	$1.2 \leq M \leq 6.7$	61.8–389	10–64

Extrasolar planets can be detected by either direct or indirect methods. Each detection method is characterized by some observables which are related to the instrinsic physical parameters of the planet. These parameters are their mass M_P, radius R_P, temperature T_P, distance a from the parent star, orbital period P, brightness L_P and distance D from the solar system. The potential success of a given method depends naturally on its instrumental limitations.

Direct imaging

The most natural gesture to detect a planet is trying to 'see' it. However, planets have generally no intrinsic emission, at least in the optical wavelength range. One

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can only detect their illumination by the parent star. However, one then faces a severe problem: the light reflected by the planet is extremely faint and, at the same time, the planet is extremely close to its star, so that it is very difficult to separate them. Indeed, for a planet orbiting at a distance a from a star with a brightness L_*, the stellar light is dispersed over the whole area $4\pi a^2$ of the sphere of radius a on which the planet is located. The planet receives only a small fraction of this light, proportional to the surface πR_P^2 of the planetary disk seen by the observer. The planet thus acquires by reflection a brightness L_P given by

$$L_P = L_* \frac{\text{planetary disk area}}{\text{area of the sphere at the planet's distance}} = \frac{L_*}{4}\frac{R_P^2}{a^2}.$$

Because the planetary radius is much smaller than its distance to the star, the brightness ratio L_P/L_* is very small: it is 2.5×10^{-9} for a Jupiter.

On the other hand, the planet is seen by the observer at an angular distance $\alpha = a/D$ from the star, i.e. very close to it (fractions of arcseconds). The planet, in addition to being very faint, is then embedded in the light halo of the star owing to the telescope's diffraction. The addition of the planet's faintness and its closeness to the star makes its detection very difficult. To circumvent this difficulty, several tricks have been invented: 'nulling imaging', 'dark speckle imaging', 'phase mask coronography', 'hyper-telescope', …. All of them are intended to extinguish the star as much as possible (without extinguishing the planet). They use the wavelike properties of light by making the star interfere destructively with itself, without affecting the nearby planet. For observations from the ground, there is an additional source of degradation of images: the atmospheric turbulence. One can, in that case, increase the sharpness of images by 'adaptive optics', a technique which counterbalances the deformation of images by the atmosphere by deforming slightly, in real time, the telescope mirror (see INSTRUMENTATION: ADAPTIVE OPTICS).

The ADONIS (Adaptive Optics Near Infrared System) instrument mounted at the European Southern Observatory (ESO) 3.6 m telescope at the La Silla Observatory has obtained an image of a dust disk around the star iota Horologii, already known to possess a planet at least twice as heavy as Jupiter in an Earth-like orbit. The dust disc around iota Horologii is quite extended; it is detected to a distance of about 65 AU (10 000 million km) from the star. From the elongation, it appears that the inclination of the disk is about 42º.

Despite the difficulties presented by this direct detection, astronomers put much effort into it because it is the only way to access several properties of the planet. As an example, the careful study of the time evolution of the planet's reflection, in different colors, can reveal its diurnal rotation period, its seasonal variation, the relative amount of water at the surface (oceans) etc.

Dynamical perturbation of the star by the planet

Instead of seeing the planet directly, one can search for its effects on the parent star. The first of these effects is the gravitational perturbation of the star's position by the planet. When the planet makes a revolution around its parent star, in fact both objects are in orbit around their common center of mass. Thus the star makes a small circular orbit with a radius $a_* = aM_P/M_*$ and a period naturally equal to the planetary orbital period P (see figure 1 by G Marry).

For the observer, this small orbit results in perturbations of three of the star's observables: its radial velocity V_R (the velocity in the direction of the star's line of sight), its position α on the sky with respect to a fixed star and the time of arrival of signals regularly emitted by the star. These three observables are perturbed with a period equal to the planet's orbital period and with an amplitude proportional to the planet's mass. The measurement of this amplitude therefore gives a determination of the planet's mass. In fact, for radial velocity measurement and time-of-arrival measurements a slight complication arises: if the inclination of the planetary orbit is not known, the method gives only a minimum value of the mass. The best example of regularly spaced signals is the periodic emission of pulsars. The first planets (in fact three and perhaps four) were discovered by this method by Wolzsczan and Frail around the pulsar PSR 1257+12. The amplitude of the radial velocity variation is 13 m s^{-1} for a Jupiter at 5 AU from its parent star. Figure 2 shows the periodic variation of the radial velocity of the star 51 Peg as a function of time ($\phi = 1$ corresponds to $t = 4.23$ days) measured by Mayor and Queloz. It was the first planet ever detected around a Sun-like star. After this first discovery, an interesting new step was the detection, by the same method, of three planets around the star upsilon Andromedae.

The precision in radial velocity measurements is at present 3 m s^{-1}. Thus it becomes possible to detect Saturns by radial velocity measurements. However, there is no hope of detecting an Earth since the radial velocity perturbation produced by such a planet is only 10 cm s^{-1}.

Planetary transits

The planet can produce a drop in the star's light during transits of the star's disk by the planet. The detection of a TRANSIT in the star's lightcurve requires two conditions to be satisfied.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

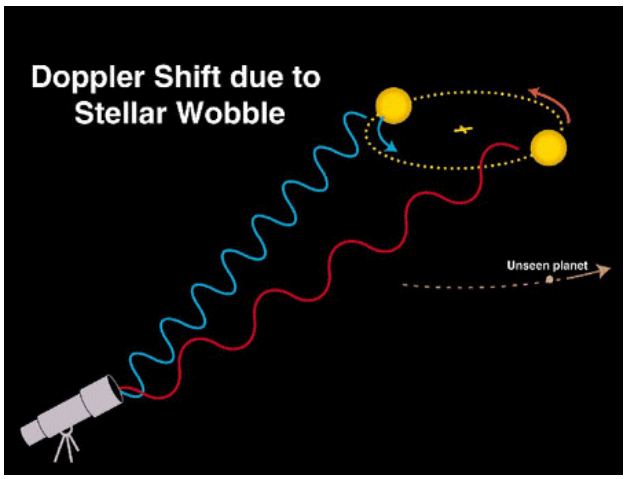


Figure 1. Doppler shift due to stellar wobble.

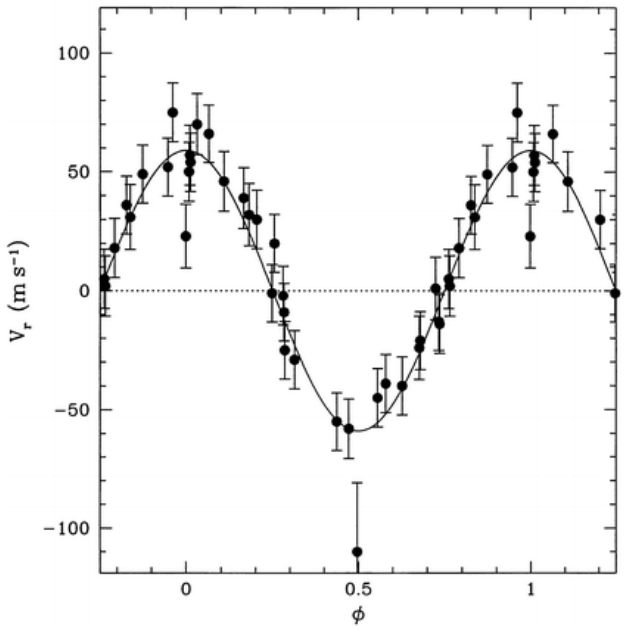


Figure 2. The variation of the radial velocity of 51 Peg (from Mayor M and Queloz D 1995 *Nature* **178** 355).

1. The orbital plane of the planet must be correctly oriented: for random orientations, the geometric probability p is $p = R_*/a$. For a Jupiter (an Earth) around a $1R_\odot$ star, this probability is 10^{-3} (respectively 0.5%).
2. The precision of the star's photometry must be better than the depth $\Delta F/F = (R_P/R_*)^2$ of the transit. For a $1R_{\rm Jup}$ ($1R_\oplus$) planet the drop is 1% (respectively 10^{-4}). In ground-based observations, the photometric precision is at best 0.1%. In space, it can reach a few 10^{-5}. Thus, while Jupiter-size planets can be detected in this way from the ground, one can detect Earth-sized planets from space. In fact, it is at present the only method capable of detecting and investigating further Earth-like extrasolar planets.

Late in 1999, the first planetary transit was detected from the ground (around the star HD 209458), leading to the first determination of the radius of an extrasolar planet; this transit was confirmed by an inspection of the photometric archives of the Hipparcos satellite.

Gravitational lensing

The planet can produce a gravitational amplification A_G of the light of background stars, increasing with the planet's mass and its distance to the observer. This amplification can reach factors up to 100 when the planet lies at several kiloparsecs, i.e. as far as the Galactic center. The GRAVITATIONAL LENSING method is thus suitable only for very distant planets, difficult to observe afterwards by any other method. Furthermore, a lensing event is seen only once and it is not possible to investigate a planet at 4 kpc any further by any other method. This makes the lensing method less attractive. Nevertheless, astronomers have recently shown by this method that less than 25% of stars have Jupiter-mass companions at distances larger than 3 AU.

First lessons from present findings

As often happens in science, what has been found was largely unexpected. The first discoveries of extrasolar planets have managed to produce great surprises and a few conclusions can already be drawn. There are three main lessons from these discoveries, all equally important.

First, the global result is that about 5% of main sequence stars have a giant planet at a distance less than 2 AU. Before 1995, astronomers had no idea whether planetary systems were very rare or not. There could, for instance, have been less than one planetary system for 1000 stars. By itself this knowledge is very precious for planetary formation theory. The theoreticians now have to explain this proportion. Also, it immediately raises a new, important question. This proportion being valid for giant planets, does it hold for telluric (i.e. small solid) planets? The latter could indeed *a priori* be very different as their formation and evolution mechanisms are significantly different. This question is of course important for future searches for life on Earth-like planets.

A second conceptually important consequence comes from the discovery of the first transitory planet HD 209458. Since a transitory body has been observed and its radius has been measured (giving a value 30% larger than Jupiter), it constitutes the first proof that the stellar wobble is due to a planet and not to other causes (such as stellar oscillation, small companion black hole,…). It also rules out an *a priori* possibility: that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

planet is a heavy (0.6 Jupiter mass) planet entirely made of solid rock instead of gas. In that case, its radius would have been ~0.3, instead of 1.3, Jupiter radii. Thus, from transits we can learn much about the planet constitution.

The third lesson is that these discoveries have given unexpected results.

1. While astronomers were expecting to find planets around main sequence stars, the first planets were discovered around stars where they were not expected, namely around pulsars. It is not clear how these planet-mass objects can be formed: have they survived the supernova explosion or have they condensed in the supernova remnant? If they were planets from the pulsar progenitor, they should have very eccentric orbits, while for PSR 1257+12 the eccentricity is almost 0. Indeed, because of the conservation of their angular momentum $J = [M_* a_P(1 - e^2)]^{1/2}$ and total energy $E_{\rm pot} + E_{kin}$ after the mass loss $M_{*1} \rightarrow M_{*2}$ during the supernova explosion, they acquire an eccentricity given by $1 - e^2 = M_{*1}(2M_{*2} - M_{*1})/M_{*2}^2$. Furthermore, the orbit remains bound only if $M_{*2} \geq M_{*1}/2$.
2. The first giant planets were found 100 times closer to their parent star than expected. Indeed, for main sequence stars with masses from $0.1M_\odot$ to $1M_\odot$, the planets were expected to form at a distance from 4 to 5 AU. This is the distance at which the equilibrium temperature is below 160 K, to allow for the formation of icy cores of giant planets. It was thus an incomprehensible surprise to find planets as close as 0.05 AU to their parent star. In fact the solution to this problem seems to lie in the notion of 'orbital migration' introduced 10 yr before the planets were discovered. When a planetary core has started to grow up to a mass M_P in a protoplanetary disk, it creates, by gravitational attraction of the surrounding gas and dust, a gap with a radius $R_H \approx a_P(M_P/M_*)^{1/3}$; this is the radius of the 'Hill sphere' inside which the gravitational attraction of the planetary core is larger than the attraction of the parent star. There are then tidal forces between the core and the inner and outer edges of the gap, whose imbalance depends on the difference in density between the edges. The imbalance can be such that these tidal forces pull the orbit toward the star.
3. Some of the planets have orbits with unexpectedly large eccentricities. This makes it unlikely that they were formed in a protoplanetary dust disk, since the internal dissipation circularizes the orbits of the dust grains. They were thus more likely to have formed in a way similar to very-low-mass stars, i.e. by gravitational condensation of the gas, with no solid core. For 16 Cyg B, the eccentricity comes perhaps from the perturbation by the companion star 16 Cyg A. Eccentricity of orbits can also be the result of migration: during their migration (at different rates) one planet can expel another from the system; from angular moment conservation, it then results that the resulting orbits must be very eccentric.

Future perspectives

These first discoveries have encouraged astronomers to accelerate their searches and many projects are at present ongoing or under study.

Future ground-based and space-based facilities

Radial velocity searches. ESO will make radial velocity searches in a dedicated program (to start in 2001) with the 3.6 m telescope at La Silla with a precision of 1–2 m s^{-1}. There is also a Swiss program, monitoring 800 stars on a dedicated 1.5 m telescope, that started in 1998, at La Silla. At the 10 m Keck telescope, Marcy *et al* have started to monitor 400 stars with a precision of 3 m s^{-1}. Several other groups conduct such radial velocity searches, or plan to do so, with more moderate precisions.

Astrometric search. The FAME astrometric mission, to be launched by NASA in 2004, is intended to measure during a few years the positions of millions of stars with a precision of 50 μarcsec (μas). It should detect up to a few hundred Jupiters. The Space Interferometric Mission (SIM) is a NASA project, devoted mainly to astrometry and, secondarily, to imaging. The nominal goal for astrometric precision is 4 μas. As far as extrasolar planets are concerned, this would give SIM the ability to detect Saturn-mass (Earth-mass) planets at up to 15 pc (respectively 5 pc). The GAIA project, if selected by ESA in the future, would measure the positions of 10^9 stars in the Galaxy with a precision of 1 μas (10 μas) for stars with $V = 5$ (respectively $V = 15$). For solar-type stars, it could detect 'big Earths' (with 10 Earth masses) around ~100 stars at up to 10 pc and Jupiter-mass planets at up to 200 pc around about 5×10^5 stars. In the latter case, if the proportion of ~5% of stars having planetary systems holds, one can estimate that GAIA would detect ~25 000 extrasolar Jupiters.

Imaging On the ground, large telescopes such as the Keck telescope, the VLT and the Large Binocular telescope will try to search for planets by imaging in the future, by different interferometric techniques. However, their imaging performances are yet not well known.

In space, the projects are more clearly identified (although the situation could change with new plans).

The NASA NGST Project (Next Generation Space Telescope), to be launched after 2007 if approved, is a large (~8 m) telescope. Its orbit is not definitely chosen; perhaps around the Earth–Sun Lagrange point L2. If a coronographic mode were to be implemented, it could detect Jupiters around nearby stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Other space projects, based on interferometric nulling, were initially motivated by exobiologic purposes. They did lead to the DARWIN project, presented in 1993 as an ESA cornerstone candidate, quickly followed by the NASA Terrestrial Planets Finder (TPF). DARWIN is, in its initial configuration, a free-flyer mission made of five 1.5 m telescopes separated by 20–100 m, while TPF is initially based on a rigid linear structure. The main difficulty encountered by these projects is that, in order to detect planets at 10 μm, it is necessary to avoid the IR thermal emission of the solar system's zodiacal dust; the only solution found to circumvent this difficulty is to send the mission to 5 AU from the Sun. A complete unknown is also the level of zodiacal IR thermal emission of other planetary systems. Extrasolar IR zodiacal light similar to the solar system's zodiacal light gives a background 300 times larger than the planet's emission. The example of 55 Cnc points out the danger that such backgrounds can even be 10 times larger than the solar system's zodiacal light. Another project, in the visible, has recently been suggested: the Epicurus project proposed by A Labeyrie to ESA. In the visible and near-IR range, the zodiacal light problem is avoided. A much more ambitious project is under study to make multipixel images of an extrasolar Earth-sized planet: the Planet Imager project, a multikilometer space interferometer which could be built around 2030 (do not forget that from now to 2030 there is the same time span as that from 1970 to today).

Transits. Programmes from space are under way, with different statuses.

- The Hubble Space Telescope carried out photometric monitoring in July 1999, continuously over 9 days, of 20 000 stars in the globular cluster 47 Tuc. It should have detected a few tens of Jupiters in close orbits. Preliminary results show that zero planets are detected; this may be due to the submetallicity of the cluster's stars.
- COROT is an ESA/CNES mission (to be launched in 2004) dedicated half to asteroseismology and half to planetary transits. It is in fact the only currently approved space mission largely dedicated to the search for extrasolar planets. With its 25 cm aperture, it should detect, during its two and a half years of operation, tens of planets with radii down to ~$2R_\oplus$ at orbital distances up to ~0.5 AU from their parent star. For M stars, some of these planets should fall in the 'habitable zone' of the star (defined as the distance at which the planet's temperature allows for liquid water).
- The KEPLER project is entirely devoted to the search for planetary transits. It is much more ambitious than COROT but is not yet approved. With its 1 m aperture and its 100 deg^2 field of veiw, it should be able to detect, after 5 yr of operation, a few hundred Earth-like planets in the habitable zones of stars.
- There is also a similar ESA project under study: Eddington.

Extrasolar planetary environments and comets. Given the extreme difficulty of detecting extrasolar planets, it seems premature to detect their surroundings such as PLANETARY RINGS etc and extrasolar COMETS. Surprisingly, most of these surroundings are detectable, some of them even more easily than planets themselves.

Rings. Rings of extrasolar planets have such a low mass (less than an Earth mass) that they have no observable dynamical effect on the star's motion. However, as shown by the example of Saturn's ring, they can be significantly more extended than giant planets (their surface being 4 times the projected surface of giant planets). They are thus obviously easier to observe by methods sensitive to the surface of objects to be detected, namely direct imaging and transits.

Satellites. Giant satellites of planets are, at least in the solar system, at most as large as telluric planets. Since to detect them by imaging it is necessary to separate them from the parent planet, the required separation power needs kilometric baselines in space.

The transit method is more promising and can be exploited in two different ways.

- Detection of the transit made by the satellite itself is marginally possible for giant (Titan-like) satellites in front of a low-radius (M or K) star monitored with a photometric precision of 10^{-4} or better (as is the case for the COROT mission and the Kepler project).
- Even if the satellite transit itself is not detected because of lack of photometric precision, it disturbs the timing of the transit of the giant planet by an amount $\Delta T_S \approx 15$ min (very easy to detect) for an Earth-mass satellite in orbit around a Saturn-like planet.

Magnetospheres. At least one planet in the solar system has a radio decametric emission stronger than that of the (quiet) Sun. This is probably due to a strong quasi-dipolar magnetic field of internal origin, producing an extended PLANETARY MAGNETOSPHERE in which particle acceleration takes place, inducing strong non-thermal radio emission.

These emissions would be detectable at 15 pc with the best receptors (Jupiter itself would be detectable at 0.5 pc). Planetary magnetic fields are usually attributed to turbulent rotation–convection motion of conducting material inside the core. Scaling laws predict a magnetic field of a few gauss for the already discovered extrasolar planets. Further in the future, a radio telescope on the dark side of the Moon would be free of anthropogenic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

parasites and of ionospheric absorption. It would allow for radio detection below 10 MHz.

Comets. In the star β Pictoris, intense spectroscopic monitoring shows rapidly (on time-scales of days) varying absorption lines. One can interpret them in terms of absorption by infalling evaporating comets at a rate of about one infall per day. If this interpretation is correct, this method could succeed for this particular star because it is surrounded by an exceptional reservoir of small bodies, as shown by the presence of a dust disk. For random stars, it is easier to detect comets by the transit of their dust tail, in a way similar to the detection of planets. For an amount of comets equivalent to that of our solar system, the detection rate by this method is about one comet per month.

Search for life on extrasolar planets

Habitable zones. What could be called 'life' is the subject of an immense debate (Schneider 1998; see also LIFE ON OTHER WORLDS). Let us restrict ourselves here to the standard conservative choice of carbon-based organic chemistry on a solid planet with liquid water. The planet must be sufficiently massive to retain its atmosphere, but not so massive as to retain its hydrogen, which would be lethal for standard life. Its mass must then be between $0.1M_\oplus$ and $10M_\oplus$. With pressure conditions given by this mass range, liquid water can be present if the planet temperature is about 300 K.

A planet orbiting at a distance a around a star with a radius R_* and a temperature T_* acquires by heating an equilibrium temperature T_P given by

$$T_P = \frac{(1-A)^{1/4}}{\sqrt{2}} \left(\frac{R_*}{a}\right)^{1/2} T_*$$

where A is the mean 'albedo' (reflectance) of the planet surface. From equation (2), a planet having a temperature $\simeq 300 \pm 20$ K must be located at a distance a from the star given by (for $A = 0.55$) $a = 0.7R_*(T_*/300)^2$. Depending on the type of the central star, this distance runs from $\simeq 0.1$ AU for cool ($T_* = 3000$ K) dwarf stars to $\simeq 2$ AU for hot ($T_* = 6500$ K) larger stars. For hotter stars the habitable zone is larger than 2 AU, but they evolve too rapidly to have stable temperature conditions over the billions of years required for the development of life (unless one invokes less conservative scenarios).

Remote detection of signatures of life. Suppose that an Earth-like planet is discovered in the habitable zone of its parent star. Is it possible to detect the presence of life on it? The answer is that it is possible to detect spectroscopic signatures of what is perhaps due to complex organic chemistry. One can think of two types of signatures.

1. *Search for oxygen.* The first type of approach rests on the observation that on Earth all the molecular oxygen and ozone in the atmosphere are of biogenic origin. To then separate the planet from its parent star, an interferometer in space with a baseline of at least 20 m is required (the DARWIN project).

However, this method suffers from at least one artefact due to the photodissociation of water under some conditions (such as the continuous supply of water by a large quantity of infalling comets).

In addition, oxygen delivery is a secondary consequence of Earth-type photosynthesis and we do not know to which extent it is accidental or universal. For instance, there are many photosynthetic bacterial species on Earth which do not deliver oxygen as a byproduct of their photosynthetic activity (producing sulfur instead).

2. *Search for chlorophyll.* A more general signature rests on the remark that, whatever the details of an extrasolar biochemistry might be, it must make use of a molecular converter which absorbs the stellar light in the intense part of its spectrum to transform it into chemical energy. On Earth, this converter is chlorophyll, but it could be other molecules on other planets. On Earth, the chlorophyll absorbs up to 80% of light in the band 400–700 nm. It would be seen as an absorption band of a few per cent in the planetary spectrum at wavelengths where the star is the brightest. To minimize the risk of confusion with artefacts (for instance absorption bands of silicates), the amount of absorption can be correlated with the planet's seasons.

Conclusion

For the first time we have a clear answer to a 2300 yr old question: 'are there somewhere [i.e. around other stars] other worlds like ours?' (EPICURUS 300 BC). The first discoveries have been full of surprises and we should be prepared for the unexpected.

In this new era of astronomy, one can already design a plan for the long-term future with well-identified steps in a clear global strategy for the entire coming century, with even a tentative agenda:

(a) first images of extrasolar planets (around 2005);
(b) detection of Earth-like planets (2005);
(c) search for spectral signatures of life (2010–15);
(d) multipixel imaging of extrasolar planets (>2025);
(e) multipixel spectroscopy of extrasolar planets (>2030).

In the second half of the century, when the most interesting planets (possibly with signs of biological activity) will have been identified, the pressure will become strong to design a mission to go there (in one way or another). Let us not forget that future astronomers born today will most likely experience these times

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

around 2100. A prospective view beyond that would be pure speculation.

Web update (15 August 2001)

A second planet has been found orbiting the star 47 Ursae Majoris located in the asterism the Plough or Big Dipper in the constellation of Ursa Major. The new planet is at least three-fourths the mass of Jupiter and orbits the star at a distance that, in our Solar System, would place it beyond Mars but within the orbit of Jupiter.

In addition, new observations from the Submillimeter Wave Astronomy Satellite (SWAS) provide the first evidence that extra-solar planetary systems contain water, a molecule that is an essential ingredient for known forms of life. From its vantage point in orbit above the absorbing effects of water in Earth's atmosphere, SWAS is capable of detecting the distinctive radiation emitted by water vapor in space.

An aging giant star designated by astronomers as IRC+10216, also known as CW Leonis, located 500 light-years (almost 3000 trillion miles) from Earth in the direction of the constellation Leo, is apparently vaporizing a surrounding swarm of comets, releasing a huge cloud of water vapor. The observations of water vapor around IRC+10216 suggest that other stars may be surrounded by planetary systems similar to our own. Over the past decade, more than 50 stars have been shown to have large planets in orbit around them, but little is known about the composition of those planets.

Breaking news update (31 January 2002)

Observations by NASA's Hubble Space Telescope of the gas-giant planet orbiting the yellow, Sun-like star HD209458, 150 l.y. from Earth have revealed the presence of sodium in light filtered through the planet's atmosphere when it passed in front of its star as seen from Earth. The planet was discovered in 1999 by its subtle gravitational pull on the star. The planet is 70% the mass of Jupiter. Its orbit is tilted nearly edge-on to Earth, which allows repeated transit observations. The planet is only about 6 million km (4 million mi) from the star. At this distance the planet's atmosphere is heated to around 1000 K.

At the time of this update 66 extrasolar planetary systems are confirmed, consisting of 74 planets and 7 multiple planets systems. See http://www.obspm.fr/encycl/catalog.html. Among recent discoveries are three Jupiter-sized planets travelling in circular orbits similar to that of Earth and Mars in our own solar system.

Bibliography

Angel R and Woolf N 1998a Searches for Earth-like planets around stars *Annu. Rev. Astron. Astrophys.* **36**

Angel R and Woolf N 1998b Searching for life in other solar systems *Sci. Am.* **9**(1) 22

Boss A 1998 *Looking for Earths* (New York: Wiley)

Clark S 1998 *Extrasolar Planets. The Search for New Worlds* (Wiley-Praxis series in Astronomy and Astrophysics) (New York: Wiley)

Dominik C, Laureijs R, Jourdain de Mouizan M and Habing H 1998 A Vega-like disk associated with the planetary system of ρ^1Cnc *Astron. Astrophys.* **329** L53

Epicurus 1957 Letter to Herodotus *The Stoic and Epicurian Philosophers* ed W J Oates pp 3–15:5

Huygens C 1698 *The Celestial Worlds Discover'd* facsimile edition 1968 (Cass)

Marcy G, Cochran W and Mayor M 2000 Extrasolar planets around main sequence stars *Protostars and Planets IV* ed Mannings, Boss and Russell (University of Arizona Press) in press

Mayor M and Queloz D 1995 A Jupiter-mass companion to a solar-type star *Nature* **378** 355

Perryman M 2000 Extra-solar planets *Rep. Prog. Phys.* **63** 1209

Schneider J 1998 The search for life outside the solar system *Exobiology: Matter, Energy and Information in the Origin and Evolution of Life in the Universe* ed Chela-Flores and Raulin (Dordrecht: Kluwer) p 319

Schneider J 2000 *The Extrasolar Planets Encyclopaedia* http://www.obspm.fr/planets

Wolszczan A and Frail D 1992 A planetary system around the millisecond pulsar PSR1257+12 *Nature* **255** 145

Jean Schneider

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Extreme Ultraviolet Explorer (EUVE)

NASA spacecraft launched on July 7, 1992 to carry out a full-sky survey in the extreme ultraviolet range of the spectrum. It carried three EUV telescopes, each sensitive to a different wavelength band. A fourth telescope performed a high sensitivity search of a limited sample of the sky in a single EUV band.

EUVE re-entered the Earth's atmosphere at approximately 23:59 UTC on January 31 2001. According to calculations made by the United States Space Command Space Control Center, EUVE re-entered the atmosphere over central Egypt. The object was not designed to survive re-entry intact and was expected to break apart and mostly burn up in the atmosphere. U.S. Space Command cannot confirm if any pieces survived re-entry. Science operations ended in December 2000. During its eight years in orbit, EUVE was the first astrophysics mission to explore the extreme ultraviolet and helped to bridge the gap in our understanding of this previously unknown spectrum. EUVE observed more than 1,000 nearby sources, including more than three dozen objects outside the Galaxy.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Eye Relief

The distance between eyepiece and the exit pupil. In the case of a compound eyepiece consisting of two or more lenses, the eye relief is the distance from the final lens to the exit pupil. For comfortable visual observing, the eye relief should be in the region of six to ten millimeters.

See also: exit pupil, eyepiece, field of view, magnification.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eyepiece

A magnifying lens used to view and enlarge the image produced at the focus of a telescope. For any given telescope, the magnification achieved is inversely proportional to the focal length of the eyepiece. A telescope used for visual work will normally have a range of eyepieces of, for example, 40 mm, 25 mm, 12 mm and 6 mm focal lengths. Used with a telescope of focal length 1200 mm, these eyepieces would produce magnifications of $\times 30$, $\times 48$, $\times 100$ and $\times 200$ respectively.

The eyepiece is normally mounted in a short tube (the drawtube) that can be slid in and out to enable the sharpest focus to be achieved. For visual use, the eyepiece is normally placed beyond the focal plane of the telescope at a distance such that the rays emerging from the eyepiece from each part of the image enter the eye along parallel paths. The eye is then relaxed (as if looking at an infinitely distant object) and less likely to become tired. For an eyepiece consisting of a single lens, this condition is achieved when the distance between the focal plane of the telescope and the eyepiece is equal to the focal length of the eyepiece.

The main parameters of an eyepiece are focal length, exit pupil, eye relief, and apparent angular field of view. Focal length determines the magnification achieved. The exit pupil is a circle through which all rays emerging from the eyepiece pass, and eye relief is the distance between the rear of the eyepiece to the exit pupil. For comfortable viewing this should normally be in the region of 6 to 10 mm. The apparent angular field of view, or acceptance angle, is the angular diameter of the circle of vision that can be seen when the eye is placed at the exit pupil. Apparent fields of view for eyepieces range from about 30° to as much as 85°, depending on the design. The observed field of view (the angular diameter of the region of sky that is visible through the telescope) is equal to field of view of the eyepiece divided by the magnification. For example, if an eyepiece with an acceptance angle of 50° produces a magnification of $\times 100$, the observed field of view will be 0.5° (approximately equal to the apparent diameter of the Moon).

There are many different types of eyepiece. Although the simplest is a single biconvex lens, an eyepiece of that kind suffers markedly from various aberrations, in particular chromatic and spherical aberration. In practice, eyepieces consist of at least two lenses, a field lens, which accepts rays of light from the telescope objective or mirror, and an eye lens, from which light emerges into the eye. Huygenian and Ramsden eyepieces consist of two plano-convex lenses (in the former case, the field lens is placed inside the focal plane of the telescope and the image is formed, inconveniently, between the two lenses). Although cheap to produce, they suffer from chromatic aberration and have modest fields of view (35°–45°). The Kellner is similar to the Ramsden, but uses as its eye lens an achromatic doublet, which greatly reduces chromatic aberration. Better correction for aberrations, good fields of view (about 50°) and good eye relief are provided by more complex eyepieces such as the orthoscopic (in which the field lens is a triplet and the eye lens plano-convex) and the Plössl (which consists of two identical achromatic doublets). For certain purposes (for example, observing star clusters or nebulae of large angular size), wide-field eyepieces such as the Erfle or Nagler are advantageous. Eyepieces of this kind, which consist of many elements (typically six or more), offer angular fields of view of up to 85° but are very expensive to produce.

See also: chromatic aberration, exit pupil, eye relief, field of view, focal length, focus, magnification, spherical aberration, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Faber–Jackson Relationship

A correlation between the luminosity of an elliptical galaxy and the velocity dispersion (the spread in velocities) of the stars in its central regions that was established in 1976 by S M Faber and R E Jackson. The light that emerges from the central regions of an elliptical galaxy is the combined light of large numbers of stars, some of which (measured relative to the center of the galaxy) are approaching the observer and others receding. Consequently, the Doppler effect broadens the lines in the galaxy's spectrum by an amount that depends on the spread of velocities among its constituent stars (lines in the spectra of approaching stars are blue-shifted to shorter wavelengths, lines in the spectra of receding stars are red-shifted to longer wavelengths, and the resulting lines in the spectrum of the galaxy are spread out over a range of wavelengths). The velocities of its constituent stars, and hence the velocity dispersion, depend on the mass of the galaxy (the higher the mass, the greater the velocities) and, as a general rule, the higher its mass, the more stars the galaxy contains. Consequently, it is logical that there should be a link between velocity dispersion and luminosity.

See also: Doppler effect, elliptical galaxies, Tully–Fisher relationship.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fabri [Fabry], Honoré (1607–88)

Born in Virieu-le-Grand, Dauphiné, France, a Jesuit, he held posts in a variety of cities in France and in Rome, where he was briefly imprisoned for unorthodoxy. He was an astronomer, physicist and mathematician and studied Saturn's rings, disputing about their nature with HUYGENS for five years. Discovered the Andromeda nebula, developed a theory of tides which was based on the action of the Moon and pioneered calculus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fabricius, David (1564–1617) and Fabricius, Johannes (1587–1616)

Lutheran pastor and astronomer in Osteel, East Frisia (northwest Germany), discoverer (1596) of the first known variable star, *mira stella* ('wonderful star'), now simply Mira (Omicron Ceti). Fabricius observed the star at its brightest and thought it was a nova, after which Holwarda noticed that a star in Cetus cataloged by PTOLEMY and TYCHO was missing but then it reappeared. Eventually the long periodic cycle of Mira was recognized. In 1611, his son Johannes returned home from the Netherlands with a telescope. Johannes discovered spots on the disc of the rising Sun, which the father and son investigated, first by direct observation, then switching to a camera obscura because of the Sun's brightness. They discovered how the spots rotated behind the Sun and back into view. Johannes wrote an un-influential tract on Sunspots, *De Maculis in Sole Observatis, et Apparente earum cum Sole Conversione Narratio* ('Narration on Spots Observed on the Sun and their Apparent Rotation with the Sun'), which was overshadowed by SCHEINER's similar discoveries a year later. Johannes died young, and David was killed with a shovel a year later by an angry peasant, whom he had accused of stealing a goose.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fabry, Marie Paul Auguste Charles (1867–1945)

Physicist, born in Marseilles, France, became professor at Marseilles and Paris, co-inventor with Alfred Perot of the *Fabry–Perot interferometer* (two partially reflective plates, almost parallel, in an etalon), discoverer of the ozone layer in the upper atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Faculae

Patches of enhanced brightness in the solar photosphere that are often seen in white light images of the Sun. They can normally be observed only when located relatively close to the edge of the visible disk (the limb) where the background brightness of the photosphere is less intense than at the center of the disk. They form inside active regions where localized magnetic fields heat their material to temperatures several hundred kelvin higher than that of their surroundings. Although often associated with sunspot groups, they may also be seen before spots emerge and after they disappear.

See also: active region, solar photosphere, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fallows, Fearon (1789–1831)

English astronomer, born in Cockermouth, Cumberland. Disenchanted with university life at Cambridge, he won the job of the first director of the Royal Observatory at the Cape, in South Africa, to provide navigational and time information for the ships of the Royal Navy that called into the harbor. He established the observatory at the base of Table Mountain, in what is now Observatory, a suburb of Cape Town, and began observing after a series of misfortunes and tribulations with the Admiralty, builders, staff, servants, his wife and relatives, accidents and snakes. His health suffered. He died aged 42 of the effects of scarlet fever, and is buried in the observatory grounds. A sad story.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

False Cross

The asterism formed by the four stars δ and κ Velorum, and ε and ι Carinae, all of the second magnitude, which make up a cross of about $10^\circ \times 6^\circ$. It is so named because it is sometimes mistaken for the Southern Cross (Crux) by observers unfamiliar with the southern sky. There is a superficial resemblance, but Crux is more compact (about $7^\circ \times 5^\circ$) and comprises rather brighter stars. The two crosses can in fact be quite easily distinguished, as the False Cross appears to recline, whereas the long axis of the Southern Cross runs almost exactly north–south, and the latter has the additional indicator of two bright stars (α and β Centauri) to the east of and in line with its short axis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Far-Infrared and Submillimetre Telescope (FIRST)

European Space Agency 'Cornerstone' mission to be launched in 2007. Carries a 3.5 metre diameter mirror, the largest ever used for an infrared space telescope. Its three instruments—two cameras and a high-resolution spectrometer—will observe far-infrared and submillimetre emissions from gas and dust heated by young stars, and should detect galaxies forming in the very early Universe. The instruments will be cooled close to absolute zero (−273 °C) using liquid helium cryostat technology developed for the ISO mission.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Far-Ultraviolet Spectroscopic Explorer (FUSE)

The first large-scale space mission to be fully planned and operated by a US academic establishment—Johns Hopkins University in Maryland. Launched June 1999. Began as a NASA-managed project in the mid-1980s, but later restructured to reduce cost and development time. Designed for high resolution spectroscopic studies at far-ultraviolet wavelengths. It will examine hydrogen and deuterium gas clouds between stars in the Milky Way and nearby galaxies, giving new information about galaxy evolution and star formation. (See also EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

FAST (Fast Auroral Snapshot Explorer/Explorer 70)

NASA Small Explorer mission to observe and measure rapidly changing electric and magnetic fields and particle flows in the acceleration region above Earth's aurora. Launched in August 1996.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fermi, Enrico (1901–54)

Italian physicist, created the first controlled chain reaction, founded Argonne National Laboratory. His work on the properties of electrons (spin-half particles like electrons are called fermions after him, and the study of their properties is called *Fermi–Dirac statistics*) enabled the pressure source in white dwarf stars to be identified, and white dwarf star properties to be calculated by CHANDRASEKHAR. Articulated the *Fermi paradox*, suggesting that the time-scale for a space-faring race to colonize a galaxy was small compared to the life of a galaxy, and so in any given galaxy all suitable planets should have been occupied; but this seems not to have been the case for the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

FG Sagittae (FG Sge)

FG Sge ($\alpha = 20^\circ11'56''$, $\delta = +20^\circ20'04''$, epoch = 2000) is a very peculiar giant star undergoing extremely rapid evolutionary changes and nucleosynthetic processing. It has evolved at a remarkable pace over the last century changing among other properties its appearance from a hot (O-type) star to a luminous, cool (K-type) giant today. FG Sge is classified as a 'born-again giant' currently experiencing its second stage as an asymptotic giant branch (AGB) star. The rapid and rare metamorphosis is likely the result of re-initiated nuclear burning when the star was approaching the domain of stellar embers as a WHITE DWARF. The newly generated nuclear energy has ballooned the star back to giant dimensions and simultaneously caused the cooling of the stellar surface. As one of just a few such known stars, FG Sge provides astronomers with an unusual opportunity to study stellar evolution and nucleosynthesis in real time. Furthermore, it resembles the R CORONAE BOREALIS (R CrB) STARS and thus may shed light on their unclear immediate ancestry.

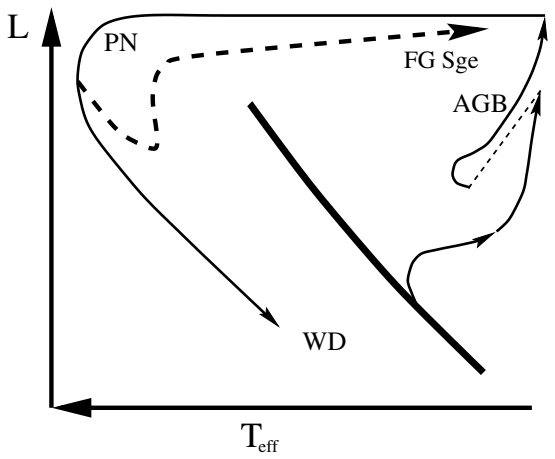


Figure 1. A schematic Hertzsprung–Russell diagram showing the likely evolution of FG Sge in terms of surface temperature and luminosity. Rather than fading to become a white dwarf (WD) after having departed from the asymptotic giant branch (AGB) and passing through a planetary nebula (PN) phase, the star has experienced a final He-shell flash causing it to briefly return to the AGB regime for the second time.

Born-again giants

Normally stellar evolution proceeds on time-scales of millions or billions of years. Notable exceptions are SUPERNOVAE, for example, though such cataclysmic events only represent the ultimate end as normal stars. To witness stellar evolution without a simultaneous complete disruption of the star on a time-scale more appropriate to the lifetimes of humans is consequently exceptionally rare. 'Born-again giants', such as FG Sge, offer a unique opportunity to study stellar evolution and nucleosynthesis during the very act, which gives important information as to how the late stages of evolution for stars similar to the Sun may proceed.

Towards the end of the AGB phase, stars of intermediate mass (a stellar mass on the main sequence of $M_* \approx 1$–$9M_\odot$) successively experience relatively long periods of quiescent hydrogen burning and brief periodic outbursts of helium burning. The latter phenomenon is called thermal pulses or 'He-shell flashes', as the He burning is thermally unstable. The thermal pulses during the AGB are responsible for substantial mixing of the interior by the development of convective shells (dredge-up), which in turn leads to further nuclear processing. Such dredge-ups are believed to eventually turn the AGB star into a CARBON STAR. The thermal pulses also extinguish the outer H-burning shell by pushing it outwards to lower temperatures. Once the He burning has died down, H burning is once again ignited and the story repeats itself after about 10^4–10^5 years.

Finally, however, due to the ever-decreasing envelope mass resulting from the intense mass loss and nuclear processing, the AGB stage is terminated. The dispersal of the cooler envelope allows the hotter core to be glimpsed, causing the star to evolve rapidly towards hotter surface temperatures in the HERTZSPRUNG–RUSSELL (HR) DIAGRAM. The still present nuclear energy production soon fizzles out, leaving the star to cool as it descends the white dwarf cooling track. Eventually it fades to oblivion when no nuclear energy is able to replace the thermal energy lost as an ever-more faint stellar ember. A schematic representation of a typical post-AGB evolution is shown in figure 1.

Though the above scenario is the most likely chain of events, descending the white dwarf cooling track may not be a dead-end. Theoretical estimates suggest that in about 10% of the stars going through a PLANETARY NEBULA stage, the pre-white dwarf will experience one more He-shell flash, causing an additional loop in the HR diagram. Due to the heating of the interior during the gravitational contraction of the core, the temperature may rise sufficiently to again trigger He burning. The liberated nuclear energy is sufficient to expand the star back to its former glory as a cool, luminous giant. If the outer H-burning shell is still active, the final He-shell flash will only cause the star to swell up but not bring down the outer envelope gas to the high temperatures of the interior due to the entropy barrier posed by the H burning. However, if H burning has ceased, no such obstacle prevents an almost complete ingestion of the still H-rich envelope, which in turn triggers H burning as well as associated nucleosynthesis and mixing with the interior. In such a case one can expect large modifications of the surface chemical composition, in particular regarding the elements H, Li, C, N and the slow neutron capture (s-process) elements.

The lifetime of such a rejuvenated giant is, however, short since quiescent He burning is never established. Within 100–1000 years—depending on the remaining envelope mass—the born-again giant will trace back its

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

former path when the stellar envelope will again be dispersed to finally become a white dwarf, this time a point of no return. The short lifetimes of born-again giants ensure that this phenomenon will be an observational rarity, in spite of the relatively many stars which are expected to undergo such a phase. Since the final thermal pulse may occur during a planetary nebula phase, this scenario can naturally explain the presence of nebulae surrounding some cool, luminous stars in the vicinity of the AGB, which otherwise would be difficult to understand.

The evolution of FG Sge

The rapid evolution of FG Sge during the twentieth century can be understood as the result of a final He-shell flash in a post-AGB star, as illustrated in figure 1. The best evidence that FG Sge is indeed a born-again giant is its surrounding nebula Henize 1-5, which directly implies that FG Sge very recently must have been a hot planetary nebula nuclei (with a surface temperature $T_{\rm eff} \geq 50\,000$ K). It is not possible to conclude exactly when the thermal pulse occurred but it most likely happened towards the end of the nineteenth century. Since then the star has been evolving towards cooler surface temperatures as a result of the overall expansion. Due to the cooling, the star has steadily become visually brighter; the visual magnitude decreased from $\approx 14^{\rm m}$ in 1880 to $\approx 9^{\rm m}$ in 1990. As yet it is unclear whether it has also increased in bolometric luminosity since the lower temperatures have shifted most of the emergent flux to optical wavelengths. According to theoretical calculations, the evolution towards the AGB regime proceeds at essentially constant luminosity after an initial rapid fading and subsequent brightening (not observed in the case of FG Sge). The evolutionary time-scale of about 100 years to reach the AGB again is consistent with the theoretical estimates for a current stellar mass of $M_* \approx 0.6 M_\odot$ (implying a mass on the main sequence of about $3 M_\odot$).

Historical spectra revealed that FG Sge was a hot star ($\approx$O3, with $T_{\rm eff} \approx 43\,000$ K) in 1880, from which it has continually evolved towards later spectral types (see CLASSIFICATION OF STELLAR SPECTRA). Today it is roughly a K2 giant, but exact classification is complicated by its extraordinary rich and peculiar stellar spectrum. To attribute a surface temperature of the star from say, 1980, until today is therefore rather uncertain, though various photometric and spectroscopic estimates suggest $T_{\rm eff} \approx$ 5000–6000 K. Accompanying the changes in appearance, the pulsational period has increased from around 5 days in the 1930s to about 100 days during the early 1990s. This can be qualitatively understood as a consequence of the expansion of the star, probably from not much bigger than the Earth to a giant perhaps 100 times the size of the Sun. From a pulsational analysis, a stellar mass of $M_* \approx 0.8 M_\odot$ can be deduced but unfortunately such estimates for FG Sge are still hampered by uncertainties.

It is not clear whether the final He-shell flash occurred before or after H burning had been terminated. Substantial modifications of the surface elemental abundances can only be expected in the latter case. There are indications that the amount of the *s*-process elements (e.g. Y, Zr, Ce and La) in the surface layers increased significantly by a factor of about ten within just a few years around 1970. Furthermore, during the 1980s FG Sge turned into a carbon star, as evidenced by the appearance of strong bands of molecular carbon (C_2) in the spectrum. Only in one other star, the similar born-again giant SAKURAI'S OBJECT, has direct evidence for current nucleosynthesis and mixing been witnessed. These observed hints of mixing suggest that H burning had already ceased before the final thermal pulse, but the details are far from properly understood.

Such surface abundance alterations are only predicted to occur at lower surface temperatures when the outer convection zone extends into layers in which nuclear processing has modified the chemical composition. It is also difficult to explain the surface enrichment of *s*-process elements prior to carbon, as theoretically one would expect the two to occur simultaneously. Furthermore, hydrogen is clearly still present in large amounts in the surface layers, while it is expected that during the ingestion of the envelope, a large portion of the present hydrogen should be burnt through CNO processing. There are indications that FG Sge may be modestly H deficient, suggesting an incomplete ingestion of the H-rich envelope, but this remains to be convincingly verified. In the alternative scenario in which H burning was still active when the thermal pulse occurred, any modifications of the surface abundances are not possible to explain without invoking other, more speculative, processes.

In August 1992 FG Sge suddenly experienced a drastic decline. Within a few weeks the star faded in the visual by more than four magnitudes. Closely thereafter an infrared excess, presumably related to circumstellar dust, was discovered which had not been present a few years earlier. The two events are therefore probably related: dust had started to condense around the star, which obscured the visual photosphere. The behavior has continued since 1992 to the present day with slow recoveries toward normal light interrupted by further rapid visual declines, as seen in figure 2. The star has, however, not yet fully returned to its initial brightness before the fadings were first observed. The variability suggests that the dust formation only occurs in localized patches rather than simultaneously around the whole star, and that fadings are only observed if the condensation happens to take place along the line of sight. This behavior closely resembles the visual variability of the R CrB stars, into which FG Sge may currently be evolving. The details behind the dust formation are poorly understood, both in FG Sge and the R CrB stars, but it is presumably related to stellar mass loss, as evidenced by the appearance of CO absorption from circumstellar gas in connection with the first fading of FG Sge. The apparent congruity between the timing of the onset of the declines and the pulsational phase of the star is suggestive of the existence of a causal connection, and further that the dust forms close to the stellar surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

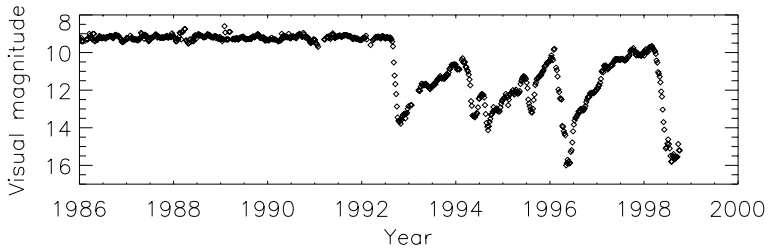


Figure 2. Visual magnitude estimates of FG Sge from 1986 to the present showing the drastic R CrB-type visual variability which started in August 1992. The data have been kindly provided by the AAVSO international database based on observations of variable star observers worldwide. The measurements have been averaged over five-day periods for the sake of clarity.

Related objects

Though FG Sge is clearly a truly unusual object it is not unique. There are at least two other stars which have been observed to undergo a born-again giant phase, and there are others which are suspected to recently have experienced a similar evolutionary stage. Both V605 Aql (Nova Aql 1919) and the recently discovered Sakurai's object have experienced a final He-shell flash during the twentieth century, and in fact evolved on an even shorter time-scale than FG Sge.

V605 Aql was initially classified as a NOVA when it brightened by more than four magnitudes between 1919 and 1921, before it soon faded to obscurity again. The rapid unfolding of events unfortunately has prevented the outburst phase from being studied carefully with modern instrumentation. Today the star is surrounded by the planetary nebula Abell 58, which contains evidence that V605 Aql indeed was a born-again giant recently. The outer parts are H-rich but the central knots of Abell 58 are H-deficient, which suggests that V605 Aql actually became H-deficient in connection with the final He-shell flash. The significantly faster evolutionary time-scale of just a few years compared with FG Sge, indicates that the stellar mass is higher, perhaps as high as $1.0 M_{\odot}$.

Another related object is Sakurai's object, which promises to hold important clues to the understanding of the born-again giant phenomenon. It was discovered in 1996 but had probably experienced its final He-shell flash a couple of years earlier, unnoticed. It has already cooled down to similar temperatures as FG Sge. More dramatically, it shows clear evidence of current mixing and nucleosynthesis which must have occurred subsequent to the thermal pulse. Since discovery, the surface abundance of hydrogen has decreased by a factor of ten while both the amount of lithium and *s*-process elements for example have increased by similar factors. These alterations are likely the result of H-burning, *s*-processing and Li-production through the Cameron–Fowler Be-transport mechanism. Furthermore, Sakurai's object has recently started to show visual variability similar to the R CrB stars and the appearance of an infrared excess, like FG Sge. Its even faster rise-time in brightness suggests it has a higher mass, likely similar to V605 Aql. Sakurai's object is also surrounded by a faint planetary nebula.

Furthermore, FG Sge seems to resemble the R CrB stars in terms of the visual variability with occasional drastic fadings. FG Sge has, however, not yet been shown to be H-deficient which is the other defining property of this group, though there are indications it may be. FG Sge may thus be evolving into an R CrB star. A study of its elemental abundance distribution is therefore expected to reveal whether it is in fact linked to these stars. Such an analysis is of great importance but is unfortunately encumbered by the complexity of the stellar spectrum. Further support for the hypothesis comes from Sakurai's object which not only shows typical R CrB variability but also a very similar chemical composition, suggesting a common evolutionary background.

Bibliography

FG Sge has attracted a great number of studies due to its special significance in terms of stellar evolution. The historical evolution of FG Sge is discussed in the following recent articles:

Blöcker T and Schönberner D 1997 Stellar evolution of low and intermediate-mass stars: III an application of evolutionary post-AGB models: the variable central star FG Sagittae *Astron. Astrophys.* **324** 991–7

Gonzalez G, Lambert D L, Wallerstein G, Kameswara Rao N, Smith V V and McCarthy J K 1998 FG Sge—a newborn R CrB star? *Astrophys. J. Suppl.* **114** 133–49

Kipper T 1996 The story of FG Sagittae *Hydrogen-deficient Stars* ed C S Jeffery and U Heber (San Fransisco: Astronomical Society of the Pacific) pp 329–40

van Genderen A M and Gautschy A 1995 Deductions from the reconstructed evolutionary and pulsational history of FG Sagittae *Astron. Astrophys.* **294** 453–68

Van Genderen and Gautschy references to the earlier literature on the subject. One such reference, however, deserves to be mentioned here due to its importance: the changing chemical composition of FG Sge—the first time this has been observed in a star. This was first described in:

Langer G E, Kraft R P and Anderson K S 1974 FG Sagittae: the *s*-process episode *Astrophys. J.* **189** 509–21

Martin Asplund

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Field of View

The angular diameter of the area of sky that is visible through a telescope. When a telescope is used visually, this equates to the area of sky that can be seen when the eye is placed at the exit pupil of the eyepiece. The apparent angular field of an eyepiece is its acceptance angle: the apparent angular diameter of the circle of vision that can be seen when the eye is placed at the exit pupil. Typically this is about 40°, but, for wide-field eyepieces, can exceed 70°. The field of view of the telescope (θ) is related to the apparent field of the eyepiece (θ') as follows: $\theta = \theta'/M$, where M is the magnification.

For most telescopes, the field of view is less than one degree, a few tenths of a degree being typical. Schmidt telescopes, however, have fields of view of up to ten degrees.

See also: exit pupil, eyepiece, magnification.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Filament

A long, dark absorption feature seen against the solar disk in monochromatic light (light of one particular wavelength). Filaments are clouds of gas, with temperatures in the range 5000 K to 10 000 K, and lengths ranging from tens of thousands to hundreds of thousands of kilometers, which are suspended above the chromosphere by magnetic fields. They are normally located along the boundary between regions of opposite magnetic polarity on the solar surface (the 'neutral line').

Filaments appear dark because they absorb light from the background solar disk. Although they also emit light (they radiate an emission line spectrum) they are much fainter than the disk. For example, when viewed in the hydrogen alpha line (of wavelength 656 nm), a typical filament has no more than 10% of the brightness of the disk and appears dark by contrast with its surroundings. However, when solar rotation carries a filament across the edge of the solar disk, it then appears as a bright feature (a prominence) against a dark background.

Most filaments have lifetimes of about a month, but larger ones survive for several months. Occasionally a filament will erupt outwards and disappear within a few hours, such an event being called a 'sudden disappearance'. If the rising filament crosses the solar limb (the edge of the visible disk) it becomes visible as an eruptive prominence.

See also: chromosphere, emission spectrum, prominence.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Filar Micrometer

A device that is used for the accurate visual measurement of the angular separation and relative orientation of two neighboring astronomical objects, such as the two component stars of a visual binary. It consists of a fixed 'horizontal' wire and two 'vertical' wires (i.e. orientated at right angles to the 'horizontal' wire), one of which is fixed and the other movable, the movable wire being adjusted in position by turning a finely threaded screw. The position of the movable wire is read off from a finely graduated drum that is rotated by the screw. The wires are positioned at the focal plane of the telescope and viewed by an eyepiece in which the wires and the astronomical objects are both sharply in focus. The entire micrometer assembly can be rotated around the optical axis of the telescope, its orientation being indicated on a finely graduated circular scale.

To make a measurement of the separation between two stars, the micrometer is rotated until one of the fixed wires (the 'horizontal' one) lies along the line between the two stars. The telescope is then adjusted until one of the stars lies on the intersection of the two fixed wires. The reading on the scale is noted. The movable wire is then shifted along until it intersects the second star, and a second reading taken. The difference in readings gives the linear distance through which the wire has moved. If the focal length of the telescope is known, the angular separation that corresponds to the measured linear separation can then be calculated. The orientation (or position angle) of the line joining the two stars is read from the circular scale.

See also: eyepiece, focus, position angle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Filterbanks

Filterbanks are instruments which divide the signal from a radio telescope into a number of frequency channels and measure the power in each of the channels. The output of the filterbank is a stream of digital data sent to a computer for analysis. Filterbanks can take several forms including analog filterbanks, digital filterbanks, fast Fourier transform (FFT) machines and software filterbanks. These different types of backend use very different means to accomplish essentially the same signal processing goals. Correlators can also be used as filterbanks (see the articles on CORRELATORS and PULSAR BACKENDS for more information).

Filterbanks are most commonly used to observe radio pulsars. A brief summary of the observational properties of pulsars will serve as an example of the use of filterbanks and also of how they work. For more information, see the article on PULSARS.

Pulsars and dispersion

Pulsars—highly magnetized rotating neutron stars—are observed as very regular pulses of broadband amplitude-modulated radio noise. These flashes of radio waves can be thought of as beams from a cosmic lighthouse sweeping through the Galaxy; when the beam intersects the Earth we see the pulsar's flash, and then it disappears until the beam comes around again. A pulsar's broadband non-thermal emission is characterized by a power-law radio spectrum

$$S \propto \nu^{-\alpha} \tag{1}$$

where S is the flux density (power per unit area per unit frequency), ν is the frequency and α ($\sim$2) is the spectral index.

The pulses of radio waves are emitted nearly simultaneously at all frequencies, but the interstellar medium (ISM) causes the different frequencies to propagate through space at different speeds (see INTERSTELLAR MATTER). Because of free electrons in the interstellar medium the signal at a lower frequency travels slower, and is therefore detected by telescopes later than that at a higher frequency. This dispersion is similar to the dispersion of white light into a rainbow of colors when it passes through a prism, and is the result of a medium having a frequency-dependent index of refraction. The propagation speed of the pulse (group velocity) as a function of frequency is given by

$$v_{\mathrm{g}} = c\left(1 - \frac{n_{\mathrm{e}}e^2}{\pi m_{\mathrm{e}}\nu^2}\right)^{1/2} \tag{2}$$

where c is the speed of light in a vacuum, n_{e} is the density of free electrons (assumed small), and e and m_{e} are the charge and mass of an electron respectively.

Dispersion greatly complicates observations of radio pulsars. At two given frequencies, the noise pulses from a pulsar will arrive at different times, separated by an amount proportional to the dispersion measure (DM), which is the integrated electron column density along the line of sight to the pulsar (DM $\equiv \int_0^d n(s)\,\mathrm{d}s$, expressed in units of pc cm^{-3}). Over a narrow bandwidth $\Delta\nu$, the time delay in seconds due to dispersion is given by

$$\Delta t_{\mathrm{s}} \simeq 8300\mathrm{DM}\Delta\nu_{\mathrm{MHz}}\nu_{\mathrm{MHz}}^{-3} \tag{3}$$

where $\Delta\nu_{\mathrm{MHz}}$ and ν_{MHz} are the bandwidth and observing frequency measured in megahertz. Figure 1 illustrates dispersion; the pulse arrival time varies by about a quarter of the pulse period over the observed bandwidth.

While dispersion complicates pulsar observations, it is not without benefit. By measuring the DM from the pulse time delay between different frequencies, we can estimate the distance to a pulsar based on our knowledge of the density of free electrons in the Galaxy in the direction of the pulsar. The density of electrons is higher in the Galactic disk, so generally speaking a given DM implies a shorter distance to a pulsar in the disk than to one at high Galactic latitude. Also, we observe a higher density of electrons toward the Galactic center than toward the outer Galaxy.

Filterbanks

The dispersive effects of the ISM discussed in the previous section demonstrate the need for pulsar filterbanks. The pulse profile (intensity as a function of time or pulse phase) observed in a bandwidth $\Delta\nu$ will be smeared in time by an amount Δt according to equation (3) (this smearing can be thought of as a convolution of the intrinsic pulse profile with a boxcar or top-hat function of width Δt). In order to resolve the details of the pulsar's light curve, we would like to have $\Delta t \ll P$, where P is the pulse period; this restricts us to small bandwidths. However, the signal-to-noise ratio increases as the square root of the bandwidth, and pulsars are weak sources, so we desire large bandwidths. The solution to this problem is to observe the pulsar in a number, n_{chan}, of narrow channels, each with bandwidth $\Delta\nu$. This gives the desired small Δt corresponding to a single channel bandwidth $\Delta\nu$, in addition to the increased signal-to-noise ratio corresponding to a total bandwidth $B = n_{\mathrm{chan}}\Delta\nu$. To correct for the dispersive delay between channels, an appropriate time offset is applied to each channel's output before the various frequency channels are combined. This signal processing solution is known as incoherent dedispersion; information about individual frequency components inside a single channel is destroyed by this method.

The number of channels in a filterbank is a primary factor in its complexity and cost, so n_{chan} is an important design parameter. The total bandwidth B is set by the telescope receiver, and it is generally desirable for the backend instrument to use all available bandwidth to maximize the signal-to-noise ratio. The maximum DM and Δt are set by science goals, with Δt being a small fraction of the minimum pulsar period a filterbank is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

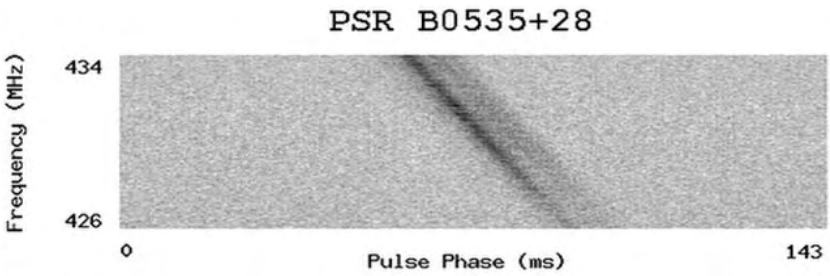


Figure 1. Dispersed pulse profile. The dispersed integrated pulse profile of PSR B0535 + 28, observed over an 8 MHz bandpass in 128 separate frequency channels with the Penn State Pulsar Machine at the Arecibo Observatory. Darker grays represent higher flux levels. Note that the pulse arrives significantly later at lower frequencies. This pulsar has $P = 143$ ms and DM $= 39.7$ pc cm^{-3}.

designed to study. From equation (3) we can see that the required number of channels for a filterbank operating at a central frequency ν is given by

$$n_{\rm chan} \simeq 8300 \frac{B_{\rm MHz} \rm DM}{\Delta t_{\rm s} \nu_{\rm MHz}^3}. \quad (4)$$

A filterbank can be thought of as a set of $n_{\rm chan}$ radio tuners working in parallel. Each tuner gets an identical input from the telescope (the antenna), and is tuned to a different frequency channel. If we imagine that these tuners are attached to power meters rather than speakers, we can watch the meters to continually monitor the power in each frequency channel. Now, if these power meters are connected to a computer which records the measurements, we have the basic workings of an analog filterbank. Note that this is different from the operation of most spectrum analysers, which sweep through the spectrum and measure the power in only one channel at a time; the analogy here would be to a single radio tuner which slowly sweeps across the bandwidth.

An alternative signal processing approach is coherent dedispersion. This technique completely removes the linear transfer function imposed by the dispersive ISM, and eliminates all dispersive smearing from the pulse profile. Coherent dedispersion can be thought of as a filterbank with an arbitrarily large number of arbitrarily narrow channels.

Analog filterbanks

Analog filterbanks are probably the easiest class to understand. We will use a highly simplified four-channel filterbank, depicted schematically in figure 2, to illustrate the operation of this type of instrument. The filterbank's input signal comes from the telescope feed; note that this signal is often mixed down (shifted to a lower frequency) from the frequency observed on the sky to simplify signal transmission and processing. An illustrative input spectrum (power per unit bandwidth as a function of frequency) is depicted in column (i) of the figure. This telescope feed signal is split four ways and passed directly to the input of each channel. From this point forward the signal is divided into four parallel channels, each processing a different part of the input spectrum. The input spectrum is filtered to select a smaller bandwidth for processing. Common filter types include LC and crystal filters, depending on the required passband characteristics. Columns (ii) and (iii) of the figure show the passband (transmission as a function of frequency) of each filter as well as the spectrum of the filtered signal, which is the multiplicative product of the input spectrum and the filter passband. At this point, we have simply divided the input spectrum up into several smaller sections. Next, in column (iv), the power in each channel is measured with a square-law detector. This detection is usually accomplished by passing the signal through a diode which is biased so that the output current is proportional to the square of its input voltage. This means that the current in the circuit at this point is proportional to the squared electric field of the signal received by the telescope, which is in turn proportional to the intensity of the observed source. This post-detection current is then integrated over an appropriate time interval with a simple RC circuit, analog-to-digital (A/D) converted to produce digital data, and finally sent to a computer for analysis. These data are proportional to the intensity of the source in the channel's passband integrated over the sampling time interval.

In order to receive all of the signal from a pulsar, both polarizations (either linear or circular) must be processed by the filterbank. This requires that all the steps described above prior to square-law detection be performed separately on each polarization. If polarization information is not required, the post-detection intensity signals can be combined to produce one data stream per frequency channel.

An example of an analog filterbank is the Penn State Pulsar Machine (PSPM), which is used at the 305 m radio telescope at Arecibo, Puerto Rico. This filterbank divides its 8 MHz bandwidth into 128 frequency channels and can measure the power in each channel as often as every 12 μs. This instrument is used for searching for new pulsars as well as timing known pulsars.

The number of channels in an analog filterbank can be effectively increased through the technique of Time Multiplexed Frequency Expansion (TMFE), as depicted in figure 3. Say we want to double the number of channels

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

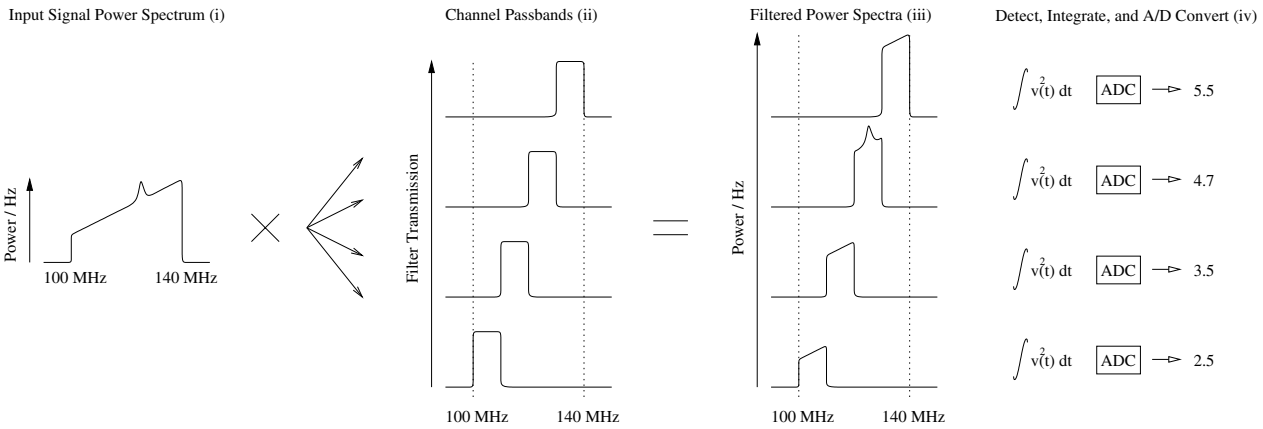


Figure 2. Analog filterbank schematic. The input spectrum is filtered, detected, integrated and A/D converted. See text for a detailed explanation.

in the filterbank in figure 2. Using TMFE, the input signal is first filtered into two channels, then each channel is digitized at baseband. The digitized data from both channels are stored in two sections of a single buffer; when this buffer is full the data are stored in a second buffer. At this point, a digital-to-analog (D/A) converter plays back the data from the low-frequency section of the first buffer, but at twice the original sample rate. This expands the bandwidth to 40 MHz—twice the original bandwidth of the channel, or the entire bandwidth of both channels. After mixing this signal up from baseband to the proper frequency for the filterbank input, the signal is fed into the filterbank. During this time, the four output channels of the filterbank contain data from only the lower half of the input bandwidth; thus we have now divided half of the bandwidth into four channels. These data are recorded for later analysis. Next, we follow the same playback procedure for the high-frequency section of the first buffer, and the filterbank divides this half of the bandwidth into four channels as well. So the output of the filterbank alternates, in this example, between four channels from an input passband of 100–120 MHz and four channels from an input passband of 120–140 MHz. Our effective channel bandwidths are now 5 MHz, as opposed to the 10 MHz of the original filterbank without TMFE. Since the TMFE system plays back its data at twice the sample rate, both sections of the first buffer can be replayed in the time it takes to fill the second buffer, so there is no loss of observing efficiency. The data in the second buffer are now fed into the filterbank while the first buffer is filled with new data, and the process repeats. It is easy to see that we can increase the effective number of filterbank channels by a factor of N by using N TMFE filters, A/D converters and buffer subdivisions, and by running the D/A converters at N times the sample rate. A TMFE system with $N = 4$ is in operation at the Parkes 64 m telescope in Australia, in conjunction with a pre-existing analog filterbank.

Digital filterbanks

The steps taken inside a digital filterbank are similar to those in the analog filterbank, but they are accomplished through different electronic means. In a digital filterbank, the input voltage signal from the telescope is sampled and digitized. This digitized signal is then subdivided into channels by digital filters, for instance finite impulse response (FIR) or infinite impulse response (IIR) filters. As an example, in the FIR case, the filter is programmed with coefficients of the discrete Fourier transform of the desired filter passband. The output of the filter for a given input data stream is then the convolution of a series of contiguous input data with the filter coefficients. This appears very different from the analog filtration process, but the Convolution Theorem guarantees that the result will be the same. As the filtered signals are already digital numbers, 'square-law detection' is accomplished by simply squaring the data, and integration by simply summing the desired number of samples.

The Navy–Berkeley Pulsar Processor (NBPP) is a digital filterbank used at the Nançay Observatory in France. This instrument has 96 channels; its FIR filters can be programmed to cover a total bandwidth of up to 168 MHz. It is designed to search for pulsars at observing frequencies above 1 GHz.

Digital filters can also be used to accomplish coherent dedispersion, as in the Green Bank–Berkeley Pulsar Processor (GBPP), at the National Radio Astronomy Observatory in Green Bank, West Virginia. After the digital bandpass filtering described above, this instrument uses a series of high-performance FIR devices to counteract the dispersion in each channel. These custom-designed PULFIR chips convolve the data stream with up to 1024 coefficients, which correspond to the dispersive transfer function for a given DM.

Fast Fourier transform filterbanks

Another way of producing what is effectively a filterbank involves no filters at all, but rather uses integrated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

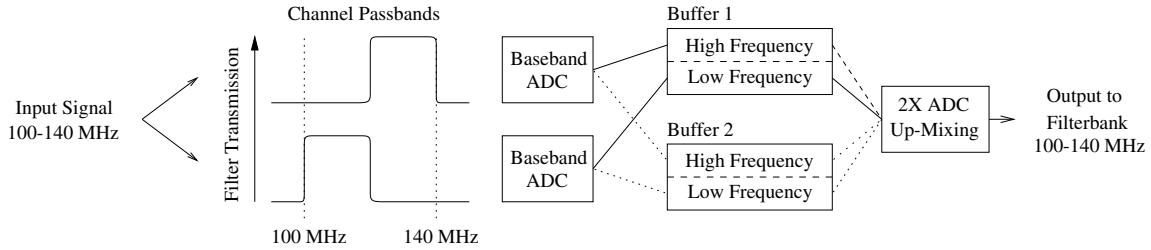


Figure 3. TMFE schematic. The signal is filtered, digitized, buffered and D/A converted sequentially at, in this example, twice the sample frequency. See text for a detailed explanation.

circuits optimized for performing fast Fourier transforms on digitized signals. As with the digital filterbank, the telescope signal is first digitized. However, rather than sending the digitized signal through a series of filters, one per channel, it is fed into an FFT chip. This dedicated computer calculates the discrete Fourier transform of a segment of the data stream, which is the discrete amplitude spectrum of the signal. To obtain the power spectrum of the signal, we need only to square the amplitude spectrum, multiplying by its complex conjugate. As this discrete power spectrum is nothing more than the power contained in individual frequency channels, we have effectively created a filterbank. These data can now be integrated by summing the output over a desired number of cycles, as with the digital filterbank.

Since we can only calculate the FFT of a short segment of data at one time, the Fourier transform we compute will be the convolution of that intrinsic to our input signal and that of the window function which is applied to it. In the simplest case, this window function is unity for the samples that make up the short input data segment and zero for all earlier and later times. This simple window function has as its Fourier transform a sinc function. When convolved with the transform of the signal, the result is that a significant amount of the amplitude, and hence power, of each component of our intrinsic spectrum is lost in the sidelobes of this sinc function. We can minimize this loss of power into spectral sidelobes by Hanning smoothing the input data. This involves multiplying the input data set by a window function which tapers to zero at the ends. The Fourier transform of this window function has a broader main peak, but much smaller sidelobes, than in the previous case. This decreases our spectral resolution slightly, but the power scattered into the sidelobes is greatly reduced; the result is a much cleaner power spectrum. With Hanning smoothing, the data near the ends of our input segment are weighted much less heavily than those near the center of the segment. To compensate for this loss, the input segments are commonly overlapped, meaning that only perhaps half of the input segment is replaced for each new FFT.

An example of an FFT filterbank is the Arecibo Observatory Fourier Transform Machine (AOFTM), which synthesizes 1024 channels covering a 10 MHz bandwidth and has 102 μs time resolution.

Software filterbanks

One final way of creating a filterbank is to simulate one using computer software. In this system, the signal from the telescope is Nyquist sampled (with a sample frequency of at least $2B$), digitized and recorded for later analysis. This method is the most computationally intensive and generates huge amounts of data, but the hardware is relatively simple and all of the information in the telescope signal is preserved. The data are later processed by a general purpose computer, which can be programmed to perform coherent dedispersion, or to simulate a digital filterbank with FIR or IIR filters, or an FFT machine. Software filterbanks will become more and more practical as computing power increases; at present they are limited to relatively small bandwidths compared with the other types of filterbanks. The computing power required to process each megahertz of bandwidth in real time is about 50 million floating point operations per second (MFlops) for a simulated incoherent filterbank and 200 MFlops for coherent dedispersion.

An example of this kind of instrument is the Caltech Baseband Recorder (CBR), used at Arecibo Observatory. This instrument covers a 10 MHz bandwidth, taking data at a rate of 20 MB s^{-1}. The data are analysed off-line using supercomputers.

Bibliography

A summary of signal processing issues in the context of pulsar observations is given by:

Hankins T H and Rickett B J 1975 Pulsar signal processing *Methods in Computational Physics. Volume 14—Radio Astronomy* ed B Alder, S Fernbach and M Rotenberg (New York: Academic) pp 55–129

A more detailed description of digital filterbanks can be found in:

Backer D C, Dexter M R, Zepka A, Ng D, Werthimer D J, Ray P S and Foster R S 1997 A programmable 36 MHz digital filter bank for radio science *Publ. Astron. Soc. Pacific* **109** 61

Bryan A Jacoby and Stuart B Anderson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Filters

Filters are used in astronomy for two main purposes, to either lower the overall intensity of the light or to restrict the wavelength range of the light that is being measured. When the eye was the only astronomical detector, filters were scarcely used, but astronomy has always adopted the new technologies of the day, be it photography, photomultiplier tubes, image tubes, CCDs or IR arrays. As the new detectors invariably have a wider wavelength response and/or improved sensitivity than previous detectors, filters became very important in order to match previously used bandpasses and to sample the flux of astronomical sources at different wavelengths thus providing some astrophysical information on the source's temperature and other properties. Some filters work by selective absorption and transmittance: these are the colored glass filters. Others work by selective reflection and transmittance: these are usually referred to as interference filters.

An interference filter comprises a multi-layer of thin films that selectively transmits or reflects light of particular wavelengths. They fall into two broad types, bandpass filters (usually narrow bands) and edge filters, that transmit above or below a certain wavelength; dichroic beamsplitters are edge filters. The Fabry–Perot tunable filter is a unique interference filter that can be used to spectacular effect on emission-line objects (see also TUNABLE IMAGING FILTERS). Finally, there are filters used for measuring polarized light and studying astronomical magnetic phenomena.

Colored glass filters

The first filters used in astronomy were colored glass (Schott or Corning) or colored gelatin (Kodak). These were used with photographic emulsions to define B, V, R and I bands. Colored glasses had been made for thousand of years for expensive glass vessels and ornaments. Nowadays, colored glasses are produced in one of two ways, either by ionic coloration or by absorption and scattering from a suspension of colloidal particles that are produced in the glass and controlled in size by heat treatment after an essentially colorless glass is made. Figure 1 shows some sample transmittance curves of these glass filters.

The ionic glasses are made by dissolving particular salts (such as cobalt or nickel oxide) in glass. The ionic coloration of the UG (violet) and BG (blue) series of glasses produces a spectral transmittance curve resembling a bell-curve of half-width between 100 and 200 nm but most of these glasses also transmit red light beyond 700 nm as well, thus the violet glasses appear purple to the eye.

The second type of colored filter glass are a series of sharp edge (or short-wave cut-off) filters, the WG, GG (sulphur and cadmium sulfide), OG (cadmium selenide) and RG (gold) series which absorb light blueward of a quite sharply defined wavelength. The filters are made with their short-wave cut-offs ranging from 400 to 800 nm in steps of 20 nm. To the eye, the WG glasses are essentially colorless, the GG series go from colorless through light yellow to dark yellow, the OG are orange-red and the RG series are rose to ruby in color.

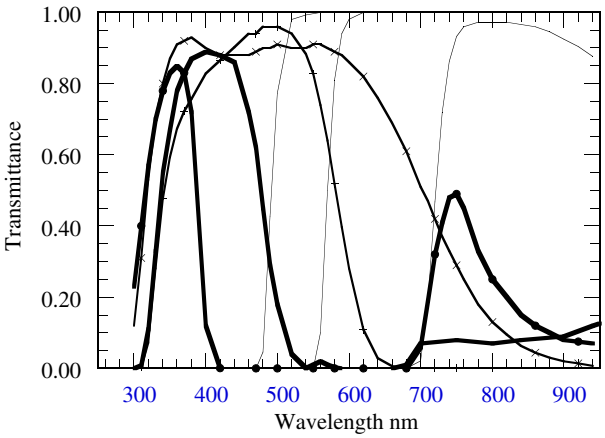


Figure 1. The transmittances of a sample of colored glass filters. The thick curves are for the ionic colored glasses UG1 (dots) and BG12. The medium thick curves with symbols are S8612 and KG3 used as red cut-off glasses. The thin curves are three of the family of short-wave cut-off glasses GG495, OG570 and RG9.

Schott also make two special (nearly) colorless glasses (BG39/40 and the KG series) that can be used as long-wave cut-off filters; however, these glasses have long-wave cut-offs that fall more slowly (over about 250 nm) than the short-wave cut-offs rise (over about 50 nm) thus producing an asymmetrical spectral transmittance curve when used together to define a bandpass. Where a steeper cut-off is desirable, it is necessary to use an edge interference coating (see below) in combination with a Schott short-wave cut-off filter. An important use for the BG39/40 and S8612 glasses is also as a 'red-leak' blocker, absorbing light beyond 700 nm where most of the Schott blue and violet glasses transmit. Polished copper sulfate crystals or a copper sulfate solution have also been used in astronomy as a UV transmitting blue filter to block red leaks, but the improved UV transmission of S8612 glass, although not as high as that of copper sulfate, has enabled it to replace copper sulfate for most imaging applications.

Combination filters and bandpasses

The ionic colored glasses define broad bell-shaped bandpasses in the violet, blue or green regions; however, the choice is restricted. Better definition has traditionally been obtained by combining one of those filters with one of the long-pass glass filters, or by using a long-pass glass filter and the red sensitivity cut-off of the detector, for instance the particular sensitization of a photographic emulsion (O, G, D, E, J, F, N) or the fall-off in the sensitivity of a photocathode (S5, S11, S20, GaAs). But with the use of wide wavelength response CCD detectors and the restriction of photographic emulsions to the broad band Tek Pan film, it is now necessary to have filter defined

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

bandpasses. This is also true in the IR and the UV, where in the past, the atmospheric windows were often partly used to define the edges of the bandpass resulting in the bandpass differing between seasons and with observing altitude. The standard U, B, V and R bandpasses can be defined using only glass filters but better R and I bandpasses can be made using interference edge filters for the red cut-offs. The bandpass filters for the SLOAN DIGITAL SKY SURVEY have been defined in this fashion, the short-wave cut-off coming from the Schott color filter and the long-wave cut-off from an interference edge filter. It is necessary to use BG39 with other BG glasses and S8612 with UG glasses to eliminate the red leak of those glasses.

Combination glass filters have been used successfully with CCD systems to match the standard UBVRI system (see MAGNITUDE SCALES AND PHOTOMETRIC SYSTEMS). The thicknesses of the glasses can be adjusted to tune the passband for better results with detectors having different wavelength responses.

Practical consideration

From a practical viewpoint, many of the BG and all the UG glasses are difficult to use. They are hard to polish, tarnish in air easily and the high concentration of ions in the available glasses often means that they are brittle and the required blue or violet color is produced with only a thin piece (1 mm or less) of the glass. This is unfortunate, as it means that it is difficult to produce good optical quality glass sandwiches in a large enough size to cover the focal plane of the large SCHMIDT TELESCOPES where it is desirable to have such a blue and ultraviolet filter for use with the new Tek Pan film. However, up to 25 cm square filters suitable for CCD arrays can readily be made. Large glass pieces are usually unavailable from stock and if ordered have to await a new melt. Some glasses, e.g. UG2 and BG37, are no longer made.

Neutral density filters

Although astronomical objects are generally very faint, it is sometimes necessary to attenuate light. The NG series of Schott filters provides a large range of attenuation and are especially useful when it is better to absorb the light rather than reflect it away. However, the attenuation is not especially constant with wavelength and the glasses absorb shortward of 400 nm. For astronomical objects therefore it is usually preferable to use filters made from fused silica that have been coated with a metallic alloy. This provides excellent attenuation that, although not identical for all wavelengths, is usually close enough to be easily calibrated. These coated fused silica filters can be used from the UV to 2 μm. For IR wavelengths germanium substrates are used. These filters are also made with a wedge or radial gradient of attenuation that is useful in non-quantitative situations.

Interference filters

Thin film coating technology has matured, and excellent interference filters are available from many manufacturers using computer designed coatings and computer controlled coating plants.

Bandpass filters

Bandpass interference filters are made by coating a glass or fused silica substrate with alternating layers of $\lambda/4$ reflective and $\lambda/2$ transmitting dielectric materials to make a single or multiple Fabry–Perot interferometer. The spacings between the layers is designed to enhance the transmittance at the wavelength of interest. By varying the reflectivities, the number of layers and the number of Fabry–Perot cavities the shape and width of the bandpass can be altered.

The more cavities, the squarer the bandpass is. Square bandpasses are desirable for emission line photometry so that the response is uniform across the field (changing angle) and for a range of RADIAL VELOCITIES. But for some astronomical purposes, such as broad band continuum photometry, square bandpasses are not desirable as they produce too great a response difference as a particular spectral feature moves in and out of the band due to radial velocity differences or chemical and physical differences between objects. Similarly, square bandpasses can also result in big differences when a filter is replaced by another whose bandpass is not precisely the same. For that reason, combination glass filters, or glass and interference edge filters, have been preferred for continuum photometry. However, rapid advances in coating technology, including rugate filter technology, now enable arbitrary shaped bandpasses with high throughput to be made. It is likely that interference filters in the future will increasingly replace most colored glass filters.

Interference filters are produced as a glass sandwich, usually between two colored blocking glasses. Because of the hygroscopic nature of the coatings the edges are sealed with epoxy and encased in an aluminum cylinder. The filters are normally between 5 and 10 mm thick depending on the diameter and the manufacturer.

Shifts in bandpass

The bandpass of an interference filter shifts blueward as the incident angle deviates from the normal. This can be useful as it enables a blueward tuning of the filter to be done. Typically a 5 degree tilt shifts the central wavelength of a 656 nm filter by about 0.7 nm while a 10 degree tilt shifts it by 2.7 nm. As most imaging is done with a converging beam the bandpass will broaden and shift blueward depending on the f/ratio of the telescope and position in the field. It is necessary to take this into account when designing narrow band imaging filters.

Temperature will also shift the bandpass as the distance between the layers changes. Typical optical interference filters at room temperatures show shifts of about 0.003% of the peak wavelength per °C. This is about 10 times less than a typical glass edge filter, such as GG395 or OG590.

Edge filters

Unlike bandpass filters, edge filters do not contain cavities. Generally they consist of quarter-wave and modified quarter-wave layers that produce an 'edge' in wavelength

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

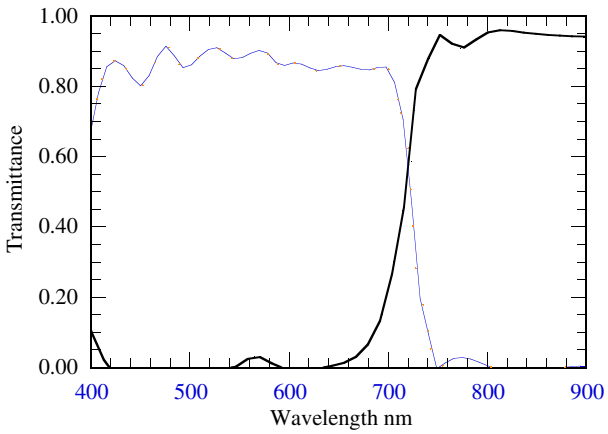


Figure 2. The transmittances of typical multilayer (interference) edge filters. The thin curve is for the short-pass filter used for defining the red edge of a passband or as a heat reflecting dichroic filter. The thicker line is for a long-pass filter that could be used as a dichroic beamsplitter in a spectrograph, reflecting blue light and transmitting red light.

space as the filter changes from reflection to transmission. Figure 2 shows examples of typical short-pass and long-pass filters. Short-pass edge filters are often used with long-pass glass filters to define astronomical photometric bandpasses. *Dichroic mirrors* are edge filters used at an angle of 45°. In optical spectroscopy they are normally used to reflect light blueward of some limit into one spectrograph and transmit the remainder into another. They are often also used to reflect IR light to an IR instrument and transmit the optical light to a guider.

Fabry–Perot tunable filters

A Fabry–Perot interferometer has been introduced at two observatories that has revolutionized the way in which high-efficiency intermediate to narrowband imaging can be carried out. Advances in coating techniques, precise spacing control of the interferometer plates and an innovative charge shuffling technique on the CCD detector have made this possible. More details can be found at (http://www.aao.gov.au/local/www/jbh/ttf/). The filter can be tuned between 370 and 1000 nm and the resolution chosen between 100 and 1000. At low resolutions (below 300), normal broad band BVRI filters can be used to block sidebands. At high resolution (1000) a set of intermediate interference filters are needed for sideband suppression. Unprecedented observations have been made with this new instrument.

Rugate filters

A rugate filter is an interference coating in which the refractive index of the coating varies continuously through its thickness. This provides greatly enhanced flexibility in filter design and extremely high levels of rejection. Multiple passbands can be produced, for instance to transmit two emission lines of interest, or to reject certain wavelengths, such as the strongest night sky emission lines or intense laser lines for safety goggles.

Antireflection coatings

Typical glass surfaces reflect about 4% of the normally incident visible and near-IR light. Silicon and germanium reflect over 60%. To avoid light loss and problems associated with reflected stray light from bright stars and the sky, antireflection coatings are usually applied to all optics including the filters. Multilayer coatings can be designed to produce extremely low reflectivities over a wide wavelength range, but for filters, a single-layer coating designed for the central wavelength of the filter is usually adequate. The most effective single layer is a $\lambda/4$ thickness of material whose refractive index is the square root of that of the glass.

Bibliography

Information on glass filters can be obtained from Schott Filter Glasses (Schott: Jena Glaswerk Schott & Gen., Mainz); on optical interference filters from the Oriel Corporation (http://www.oriel.com), Melles Griot (http://www.mellesgriot.com) and Barr Associates (http://barrassociates.com) also for rugate filters; for infrared filters (http://barr-associates-uk.com).

Michael Bessell

Table 1. Camera focal lengths and fields.

Name	Focal length	F no.	Aperture	Scale	Field
Hasselblad	80mm	2.8	60mm	60 arcsec	30 arcdeg
Nikon	400	4.5	90	12	6.8
40 inch	8128	8	1016	0.6	0.3

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finé, Oronce (1494–1555)

Cartographer, mathematician and military scientist, born in Briancon, France, professor of mathematics at the Collège Royal in Paris, wrote on astronomical instruments and astronomy, suggested that eclipses of the Moon could be used to determine longitude. A map of the world that he produced in 1531 uses the name 'Terra Australia' for the first time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finder

A small telescope of low magnification and relatively wide field of view that is attached to the main telescope and can be adjusted to be exactly parallel to the optical axis of the main instrument. Because its field of view is wider than that of the main telescope, it is much easier for the observer to locate the object of interest in the finder (provided it is bright enough to be seen or near something bright enough to be seen) than in the narrow field of the main instrument. Once the object has been identified, the telescope controls are adjusted to bring it to the center of the field of view of the finder (which may be marked by crosswires). It should then be readily visible within the field of view of the main telescope.

Modern telescopes that are computer controlled can position themselves on a celestial object to arc-second accuracy and have little need for a finder telescope.

See also: field of view, guide telescope, magnification, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fireball

An exceptionally bright meteor, traditionally defined as one that exceeds the planets in brightness. Since Venus at greatest brilliancy reaches magnitude -4.7, a fireball is now generally taken to be a meteor with a magnitude of -5 or greater. About one meteor in every thousand recorded is a fireball. Some meteor showers, for example the Geminids, are particularly productive of fireballs. Among sporadic meteors, fireballs are most likely to occur in the spring. Fireballs are produced by larger than average meteoroids and are usually of asteroidal rather than cometary origin, and therefore denser. A sufficiently large object will not only produce a fireball but will be only partially consumed by its passage through the Earth's atmosphere, and will fall to the surface as a meteorite. A fireball that generates a sonic boom is known as a bolide.

See also: bolide.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

First Point of Aries

An alternative term for vernal equinox. At one time (some 2000 years ago) the vernal equinox lay in the constellation of Aries but, due to precession, this is no longer the case and the vernal equinox has moved into Pisces.

See also: celestial equator, celestial sphere, ecliptic, equinox, First Point of Libra, precession, solstices.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

First Point of Libra

An alternative term for autumnal equinox. Due to precession, this point no longer lies in the constellation of Libra but, instead, is located in Virgo.

See also: celestial equator, celestial sphere, ecliptic, equinox, First Point of Aries, precession, solstices.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fisher, Sir Ronald Aylmer (1890–1962)

Statistician, born in London, England. After studying astronomy using AIRY's manual on the *Theory of Errors* he became interested in statistics, and laid the foundation of randomization in experimental design, the analysis of variance and the use of data in estimating the properties of the parent population from which it was drawn. Invented the maximum likelihood method for estimating from random data.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fission

The breaking apart of a body into smaller fragments. In the context of nuclear physics the term, nuclear fission, refers to the splitting of a heavy atomic nucleus into two or more lighter nuclei with the release of energy. The mass of the nucleus prior to fission is greater than the combined masses of the fragments, the difference in mass, Δm, being released as a quantity of energy, ΔE ($\Delta E = \Delta m c^2$, where c denotes the speed of light).

For example, the capture of a neutron by a uranium-235 nucleus causes that nucleus to break into two fragments (a nucleus of molybdenum-98 and a nucleus of xenon-136), together with two neutrons and a substantial quantity of energy (212 MeV). Because the reaction liberates additional neutrons, a chain reaction is possible in a mass of uranium. If the reaction is allowed to build up, unchecked, a nuclear explosion will take place (this is the basis of the atomic bomb), while if the rate of the reaction is controlled, by regulating the supply of neutrons, a steady output of energy can be achieved (this is the basis of the nuclear reactor).

See also: atom, electronvolt, energy, fusion, mass, neutron.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fitzgerald [FitzGerald], George Francis (1851–1901)

Irish physicist, born in Kill-o'-the Grange, Monkstown, Co. Dublin, became professor at Trinity College Dublin, identified the reason for the result of the Michelson–Morley experiment, as the shrinkage of a body due to motion at speeds close to that of light. LORENTZ independently came to the same conclusion, and the effect is called the *FitzGerald–Lorentz contraction of relativity*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Five College Radio Astronomy Observatory

The Five College Radio Astronomy Observatory (FCRAO) is a research organization within the University of Massachusetts which aims to conduct astronomical research, to develop state of the art microwave instrumentation, and to train both graduate and undergraduate students with support from the National Science Foundation and the Commonwealth of Massachusetts.

FCRAO operates the 14 m telescope located in New Salem, Massachusetts for observations within the 3 mm wavelength band. Recent instrumentation highlights include the development of large format focal plane array receivers. FCRAO students, scientists and visiting observers have exploited the capabilities of the system to conduct unique, wide field imaging studies of the molecular interstellar medium in the Milky Way, detailed studies of star forming regions and the molecular gas content in external galaxies.

FCRAO is a partner with the Instituto Nacional de Astrofisica, Optica y Electronica (INAOE) in Mexico to construct the Large Millimeter Telescope, a 50 m antenna on Cerro La Negra. This facility will be capable of efficient operations between 85 and 345 GHz to explore the early epoch of galaxy formation, the structure of nearby galaxies, and the physical and chemical processes within the molecular interstellar medium.

For further information see
http://donald.phast.umass.edu/~fcrao/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fizeau, Armand (1819–96)

French physicist, first measured the speed of light by reflecting it over an 8 km distance through a rapidly rotating toothed disk known as the *Fizeau wheel*, which was used to measure the time of flight over that distance. Suggested using a twin-aperture interferometer to measure stellar diameters, calculating that interference fringes would vanish at a separation related to the size of the star's disc. Idea followed successfully by MICHELSON.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flagstaff Field Center

The US Geological Survey Flagstaff Field Center was established in 1963 to provide geologic information about the Moon and to help train astronauts. After Apollo, research expanded to support robotic exploration of the planets and the study of climate and human impacts on the natural environment. Future research (terrestrial and extraterrestrial) will reflect the increasingly global nature of society's needs, quests and problems.

For further information see http://wwwflag.wr.usgs.gov/usgsflag/space/space.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flame Nebula (NGC 2024)

A bright nebula in the constellation Orion, position RA $05^{\mathrm{h}}\,41.9^{\mathrm{m}}$, dec. $-01^{\circ}\,51'$. About $\frac{1}{2}^{\circ}$ across, its illuminating source is hidden behind a patch of dark nebulosity that crosses its center.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flammarion, Nicolas Camille (1842–1925)

French astronomer and geophysicist, born in Montigny-le-Roy, observed double stars, and planets from his observatory at Juvisy, south of Paris. Very well-known for his popularization of astronomy, including many lavishly illustrated books (for example *L'Astronomie Populaire*) and a journal *L'Astronomie*, still published. It was Flammarion who, as shown by Arthur Beer and Bruno Weber, drew, sometime before 1888, a much reproduced 'German medieval woodcut' showing the Ptolemaic uinverse with a pilgrim looking through a celestial sphere to the mechanism beyond.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flamsteed, John (1646–1719)

Born in Denby, Derbyshire, England, the first Astronomer Royal, a term then synonymous with director of the Royal Observatory, Greenwich, whose creation he had proposed directly to King Charles II. The observatory was completed in 1675, and the King appointed Flamsteed at a salary from which he had to provide his own instruments. Flamsteed was a skilled observer, making the most accurate measurements of the positions of the stars and Moon then known, for the purpose of navigation (particularly for determining longitude). He was a perfectionist and when SIR ISAAC NEWTON, on the Board of Visitors (governors) of the observatory, pressed Flamsteed to release preliminary data on the Moon, to test the law of gravitation, they clashed. HALLEY intervened on behalf of Newton and edited the observations for immediate publication, over Flamsteed's objections. Flamsteed bought and burned 300 of the edition of 400, and the book is very rare. Flamsteed left his definitive observations to be published by his widow as *Historia Coelestis Britannica*. He introduced a system of letters and numerals for stars, in sequence by right ascension within a constellation, and known as *Flamsteed numbers*, e.g. 61 Cygni.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flare

A localized explosive release of energy, usually from a site located within a complex solar active region, in which up to 10^{25} J can be liberated within a few thousand seconds. A typical flare reaches peak brightness within the first 5 min of the event and declines more slowly over the next 20 min. Some flares can last for up to three hours.

Flares emit radiation over practically the entire electromagnetic spectrum from hard (short-wavelength) x-rays, or even gamma rays, to radio waves, the bulk of the energy being emitted at x- and extreme ultraviolet wavelengths. They eject streams of high-energy subatomic particles, including electrons (which in some cases are accelerated to speeds in excess of half the velocity of light), protons and small numbers of heavier nuclei, and expel bulk clouds of plasma through the corona into interplanetary space. As the high-speed particle streams and plasma clouds plow through the corona, they stimulate the emission of microwave and radio radiation across a wide range of frequencies. These dramatic events also send shock waves across the surface of the Sun and into its interior and trigger various kinds of prominence phenomena.

Flares are believed to be caused by the sudden release of magnetic energy that had previously been stored in the twisted magnetic fields of complex sunspot groups or active regions by means of a process called magnetic reconnection. This occurs when oppositely directed field lines come into contact, sever, and join up to form new field structures—for example a closed loop with its 'feet' connected to the solar surface and an open U-shaped structure above. In the process, part of the energy contained in the magnetic field is converted into other forms of energy, such as thermal (heat) energy and the kinetic energy of particle streams. Localized heating at the site of the reconnection produces much of the soft (longer wavelength) x-ray and extreme ultraviolet radiation, while electrons accelerated down the loops produce bursts of hard x-rays and hydrogen-α (optical) radiation where they plow into the chromosphere. Hard x-rays are also produced above the site of the flare by shock waves that heat the coronal plasma to temperatures of around 2×10^8 K. Like an elastic string, a magnetic field line has tension in it. Consequently, when a magnetic field structure severs and reconnects, it catapults bulk plasma outwards.

Flares have a variety of effects on the Earth. High-speed particles from solar flares take about 30 min to travel to the Earth. Low-energy particles and disturbances produced in the solar wind by the flare take from a few hours to a couple of days to arrive. Charged particles plunging into the upper atmosphere give rise to auroral displays. Flare-generated x-ray and ultraviolet radiation disturbs the Earth's ionosphere and can therefore disrupt terrestrial radio communications. Clouds of ejected plasma, and disturbances in the solar wind and interplanetary magnetic field, squeeze and stretch the Earth's magnetosphere, giving rise to magnetic storms and surges in power transmission lines.

See also: active region, aurora, chromosphere, corona, geomagnetic storm, magnetosphere of Earth, prominence, solar wind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flash Spectrum

The emission-line spectrum of the solar chromosphere. The name derives from the fact that when the Sun is observed with the aid of a spectroscope, during a total solar eclipse, the faint chromospheric emission lines suddenly flash into view at the instant when the much more brilliant light of the photosphere is blotted out completely by the disk of the Moon. American astronomer Charles Young first observed the flash spectrum in 1870.

The visible spectrum of the chromosphere consists of a very weak continuum on which are superimposed thousands of emission lines. When the solar disk is viewed directly, these faint emission lines cannot be seen against the intensely luminous photosphere. Instead what is seen is the continuous spectrum of the photosphere together with the superimposed pattern of dark absorption lines that results from absorption of light as it moves out through the photosphere and chromosphere. When, as at a total eclipse, chromospheric gas can be seen at the limb (the edge of the solar disk) without the background photospheric light, the emission lines show up clearly. Because different lines become prominent at different temperatures, observations of the chromospheric spectrum made at different heights above the limb yield information about how the temperature of the solar atmosphere varies with altitude. During a total solar eclipse, the appearance of the flash spectrum changes rapidly as the Moon's disk advances across the thin chromospheric layer at the solar limb, and therefore allows observers to receive light from different altitudes above the solar surface.

See also: absorption spectrum, chromosphere, continuous spectrum, emission spectrum, photosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fleming, Williamina Paton Stevens (1857–1911)

Astronomer, born in Dundee, Scotland, emigrated to Boston with her parents and worked as a maid in the home of EDWARD PICKERING, the director of the Harvard Observatory. Unhappy with the work performed by his existing employees and declaring that his maid could do a better job, Pickering hired her for clerical work and mathematical calculations. Fleming devised a system of classifying stars according to their spectra, obtained by Pickering from images of the stars obtained by photographing the sky through a prism placed over a wide-field camera. This resulted in a catalog of the 'spectral types' of 10 000 stars, published in 1890 as the first *Draper Catalog of Stellar Spectra* and the basis of the several volumes (1918–24) of the definitive *Henry Draper Catalog* drawn up by ANNIE CANNON. Discovered 222 variable stars and 10 novae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fluid Dynamics

Fluid dynamics is a subfield of continuum mechanics which deals with the theory of motion of a fluid (hydrodynamics) or gas (gas dynamics).

Fluid dynamics is studied theoretically, experimentally and by means of numerical simulations. The phenomena of fluid motion are governed by well known continuum laws of physics: the conservation of mass, momentum and energy. Whether these continuum laws accurately describe a given physical system depends both on the mechanical and thermodynamic state of the system (i.e. its density, temperature, etc) and on the kind and strength of the interaction of its constituents (e.g. the scattering rate of gas particles). The matter of astrophysical objects (e.g. of a star or galaxy) can often be approximated as a gas or fluid, i.e. the equations of fluid dynamics are adequate to describe the astrophysical phenomena. Hereafter, for simplification, the word fluid will be used as a synonym for both fluid and gas.

The conservation laws describing the motion of a fluid can be formulated as set of nonlinear partial differential equations. Analytic solutions of this set of equations only exist for simplified highly symmetric flow problems, and the mathematical theory is often difficult. For this reason, observations of fluid motion in nature, in the laboratory and with computers are also essential for understanding the motion and behavior of fluids.

In astrophysics the numerical tool is often the only one available to the researcher. Moreover, as most astrophysical conditions are inaccessible in the laboratory, and as astrophysical fluid motion may occur on time scales long compared with the lifespan of humans or deep inside astrophysical objects, numerical simulation is the only means to study such fluid motion. The study of astrophysical fluid flows is further complicated by the effects of self-gravity, which must be considered in many astrophysical flow problems, by the enormous range of length scales and time scales to be covered in the simulations, and by a variety of other physical effects which must be taken into account frequently. The latter include radiation transport (of photons and neutrinos), heat conduction, radiative cooling, ionization and recombination of atoms, magnetic fields, energy generation by thermonuclear reactions, flow velocities near the speed of light and in strong gravitational fields, and the properties of dense matter.

Applicability of fluid dynamics

Two requirements must be fulfilled in order for the continuum approximation of fluid dynamics to hold.

Firstly, one has to assume that the microscopic behavior of single particles can be neglected. This requires that $\lambda \ll L$, where λ is the collisional mean free path of the particles and L is a characteristic macroscopic linear dimension of the system, or a scale over which the distribution function varies significantly. If the above equation holds, the concept of a fluid element can be introduced, the linear size of the fluid element being small compared with L but large compared with λ. This implies that the number of particles in the fluid element is large, and hence mean physical quantities, like, for example, the fluid density ρ or the fluid velocity $\boldsymbol{v}$ can be defined for the fluid element. The velocity of the individual particles is then given by $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$, i.e. it possesses a random component $\boldsymbol{w}$ above the mean. Because $\lambda \ll L$, however, the particles only perform a random walk about the mean motion $\boldsymbol{v}$, and consequently the fluid element remains well defined during the evolution except for some small loss of particles near its surface, which, if necessary, can be well described by an additional diffusion term in the hydrodynamic equations.

Secondly, the forces between particles must be saturating or 'short range', because otherwise collective effects must be taken into account. Formally this can be expressed by the requirement that the energy per particle E/N must approach some constant value when the number of particles tends to infinity. For a fluid element with volume V one can then define an energy density $\varepsilon \equiv E/V = n(E/N)$ where $n \equiv N/V$ is the number density, and a pressure $p = n(\partial\varepsilon/\partial n) - \varepsilon$ exerted on the 'walls' confining the fluid element. Examples of non-saturating forces are gravity and electromagnetic forces which both scale as $(\sim r^{-2})$. Consequently, gravity must be included in the hydrodynamic equations as a macroscopic external force. Although electromagnetic forces are long range too, they saturate in electrically neutral systems due to screening.

Equations of fluid dynamics

The equations of fluid dynamics can be derived by a phenomenological approach or within the framework of the KINETIC THEORY. In the phenomenological approach one postulates certain relations between stress and rate of strain and between heat flux and temperature gradient and develops the fluid dynamic equations from the universal laws of conservation of mass, momentum and energy. The required constants of proportionality between stress and rate of strain and between heat flux and temperature gradient are called transport coefficients and must be determined experimentally.

The equation that results from requiring the conservation of mass in a fluid flow is called the *continuity equation*. The conservation of momentum law or Newton's second law applied to a fluid yields a vector equation known as the *Navier–Stokes equation*. The First Law of Thermodynamics or the conservation of energy law imply a fluid dynamic equation which is called the *energy equation*.

In the kinetic theory approach the fluid dynamic equations are derived from the *Liouville equation* by defining the transport coefficients in term of certain integral relations which involve the dynamics of colliding particles. The drawback to this approach is that the interparticle forces must be specified in order to evaluate the collision integrals. Both the phenomenological approach and the kinetic theory approach yield the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

same fluid dynamic equations provided equivalent assumptions are made during their derivations.

Mathematically speaking the five fluid dynamic equations form a set of coupled nonlinear partial differential equations which are of first order in time and of second order in space. They describe the evolution of mass density ρ, momentum density $\rho\boldsymbol{v}$ and (total) energy density e in time and space. In addition to the evolution equations, it is necessary to supply a relationship between fluid properties in order to close the system of equations. The equation of state is an example of such a relationship. It relates the thermodynamic variables pressure p, density ρ and temperature T or internal energy density ε.

In many astrophysical applications of fluid dynamics, additional equations and/or additional source terms have to be considered, which for example describe composition changes and energy generation due to nuclear reactions, the diffusive transport of matter or radiation (diffusion equation), magnetic fields and their coupling to matter (magnetohydrodynamics), the evolution of radiation fields and their interaction with matter (radiation hydrodynamics), and self-gravity. Moreover, in certain applications the conservation equations have to be formulated consistently with SPECIAL or GENERAL RELATIVITY.

Form of the equations

Here the equations of fluid dynamics are given without derivation. The various quantities appearing in the equations below are defined as follows: $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$ is the particle velocity, $\boldsymbol{v}$ the average fluid velocity, $\boldsymbol{w}$ the statistical fluctuation around the mean flow, ρ is the fluid density, $\rho\boldsymbol{v}$ is the fluid momentum, and $\rho\varepsilon$ (erg cm^{-3}) is the kinetic energy density about the mean motion (i.e. the internal energy), respectively.

The *continuity equation*

$$\frac{\partial\rho}{\partial t} + \mathrm{div}(\rho\boldsymbol{v}) = 0$$

describes the conservation of mass, the *Navier–Stokes equation* of motion

$$\frac{\partial}{\partial t}(\rho\boldsymbol{v}) + \mathrm{div}(\rho\boldsymbol{v}\otimes\boldsymbol{v}) + \mathrm{div}\underline{\underline{\Pi}} = -\rho\mathrm{grad}\Phi \qquad (1)$$

describes the transport of (relative) momentum in the flow. The total energy equation (assuming that the particles have no internal degrees of freedom)

$$\frac{\partial}{\partial t}\left(\frac{\rho}{2}|\boldsymbol{v}|^2 + \rho\varepsilon\right) + \mathrm{div}\left[\left(\frac{\rho}{2}|\boldsymbol{v}|^2 + \rho\varepsilon\right)\boldsymbol{v}\right] + \mathrm{div}\vec{h} + \mathrm{div}(\underline{\underline{\Pi}}\boldsymbol{v}) = -\rho\boldsymbol{v}\mathrm{grad}\Phi$$

finally describes the conservation of energy. $\underline{\underline{\Pi}} \equiv \rho\langle\boldsymbol{w}\otimes\boldsymbol{w}\rangle$ is the pressure tensor, from which one usually separates out the trace of the symmetric dyadic $\boldsymbol{w}\otimes\boldsymbol{w}$, i.e. one writes

$$\underline{\underline{\Pi}} = P\underline{\underline{I}} - \underline{\underline{\pi}}$$

where

$$P \equiv \tfrac{1}{3}\rho\langle|\boldsymbol{w}|^2\rangle$$

is the isotropic fluid (or gas) pressure

$$\underline{\underline{\pi}} \equiv \rho\langle\tfrac{1}{3}|\boldsymbol{w}|^2\underline{\underline{I}} - \boldsymbol{w}\otimes\boldsymbol{w}\rangle$$

is the viscous stress tensor and $\underline{\underline{I}}$ is the unit tensor.

$$\varepsilon \equiv \tfrac{1}{2}\langle|\boldsymbol{w}|^2\rangle$$

is the specific internal energy (erg g^{-1}), and

$$\vec{h} \equiv \rho\langle\boldsymbol{w}\tfrac{1}{2}|\boldsymbol{w}|^2\rangle$$

is the conduction heat flux. The operator $\langle\rangle$ denotes an average over the one-particle Boltzmann distribution function.

In the absence of viscosity the Navier-Stokes equation reduces to the *Euler equation*

$$\frac{\partial}{\partial t}(\rho\boldsymbol{v}) + \mathrm{div}(\rho\boldsymbol{v}\otimes\boldsymbol{v}) + \mathrm{grad}\,P = -\rho\mathrm{grad}\Phi.$$

For the special case of an *adiabatic flow* the total energy equation reduces to the simpler form

$$\frac{\partial e}{\partial t} + \mathrm{div}[(e+P)\boldsymbol{v}] = -\rho\boldsymbol{v}\mathrm{grad}\Phi$$

where $e = \frac{1}{2}\rho|\boldsymbol{v}|^2 + \rho\varepsilon$ is the total energy density (erg cm^{-3}).

Flow regimes

Fluid flows can be uniquely classified into different flow regimes depending on the value of certain characteristic dimensionless flow parameters. Different flows having the same dimensionless flow parameters are called similar.

Depending on the degree of compressibility of a fluid under forces acting on it, it can be treated as either being *incompressible* or *compressible*. In the former case the fluid dynamic equations can be simplified considerably. If the state of the fluid is independent of time at every point in space occupied by the fluid, the flow is called *stationary*. If the opposite is true, one has an *unsteady flow*. For stationary flows *Bernoulli's equation* holds, which says that the sum of the specific kinetic energy ($|\boldsymbol{v}|^2/2$) and of the enthalpy of the fluid is constant along a streamline. A *streamline* is defined as a curve whose tangent vector at each point of the curve has the same direction as the velocity field of the flow at that point.

If the fluid moves with a velocity smaller or larger than the sound speed, the flow is called *subsonic* and *supersonic* respectively. The Mach number $Ma \equiv |v|/c_s$, which is the ratio of the absolute value of the flow velocity $\boldsymbol{v}$ and the sound speed c_s, is the parameter characterizing this aspect of a flow. *Transonic flows* have $Ma \approx 1$. Supersonic flows accessible in the laboratory have Mach numbers $Ma \lesssim 10$, while astrophysical flows can be extremely supersonic ($Ma > 100$).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

When a flow accelerates (decelerates) from subsonic (supersonic) to supersonic (subsonic) speeds *shock waves* occur in the flow. A shock wave is a front across which there is a nearly discontinuous, finite jump in pressure, with corresponding jumps in temperature, density, and other fluid properties. Shocks are nonlinear waves whose behavior is much more complex than that of (linear) sound waves, which are periodic adiabatic pressure perturbations of the flow. Because of their nonlinear character linear superposition does not apply in shock wave interactions. Flows entering a shock at an oblique angle are deflected away from the normal direction. The deflection is due to a decrease of the normal component of the fluid velocity entering the wave. Hence kinetic energy is dissipated into heat in a shock wave, increasing the entropy of the fluid. Conservation of mass, momentum and energy must also hold across shock waves. This fact can be used to derive the *Rankine–Hugoniot relations* relating the fluid states on both sides of the shock wave.

One further discriminates between *viscous* and *inviscous* or *ideal* flows depending on the amount of viscosity of the fluid. The viscosity of a fluid varies, in general, with the thermodynamic state of the fluid. Except for superfluids the viscosity of a fluid is never exactly zero, but it may be sufficiently small to consider the fluid for all practical purposes as being inviscous or ideal. Quantitatively the viscous character of a fluid is determined by comparing the size of the viscous drag acting on the fluid with its inertia. The relevant flow parameter is the *Reynolds number* $Re \equiv |\boldsymbol{v}|l/\nu$, where l is a characteristic linear dimension of the flow, and ν is the kinematic viscosity of the fluid. Viscous flows have $Re < 1$. Most astrophysical flows have $Re \gg 1$ (Reynolds numbers larger than 10^{10} are not uncommon), because of the large size of the astrophysical objects.

A common feature of terrestrial and in particular of astrophysical flows is their tendency to undergo a spontaneous transition from *laminar* (i.e. smooth, regular type of flow) to *turbulent flow*. This occurs if the Reynolds number of the flow exceeds a critical value $Re_{\rm crit}$ beyond which effects due to inertia dominate over friction by such an amount that the flow becomes unstable to turbulence. From laboratory experiments one finds critical Reynolds numbers for fully developed turbulence in the range $1000 \lesssim Re_{\rm crit} \lesssim 2000$.

According to J O Hinze 'Turbulent fluid motion is an irregular condition of flow in which the various fluid quantities show a random variation with time and space coordinates, so that statistically distinct average values can be discerned'. Turbulent motion is not irregular in either space or time alone, but both. Turbulence can be generated by friction at fixed walls ('wall turbulence') or by flow of layers of fluids with different velocities past or over one another ('free turbulence'). In real viscous fluids, viscosity effects will result in the conversion of kinetic energy of flow into heat, i.e. turbulent flow is dissipative in nature. Hence, if there is no continuous input energy into the flow, the turbulent motion will decay. If the turbulence has quantitatively the same structure in all parts of the flow field, the turbulence is said to be *homogeneous*. If the statistical features have no preference for any direction, the turbulence is called *isotropic*. At a given point in the turbulent domain a distinct pattern is repeated more or less regularly in time, and at a given instant a distinct pattern is repeated more or less regularly in space. Turbulent motions in which distinct patterns can be discerned are described quantitatively by introducing the notion of *scale of turbulence*: a certain scale in time and a certain scale in space. Turbulent motion, in addition to its scale, is characterized by the violence or intensity of the turbulence fluctuations, which is defined by the root-mean-square value $w' \equiv \sqrt{w^2}$ of the random component $\boldsymbol{w}$ of the momentary value of the velocity $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$ about the mean flow velocity $\boldsymbol{v}$. Often the intensity of the turbulence is defined by the *relative intensity* $w'/|\boldsymbol{v}|$.

Flow instabilities

Fluid flows are subject to a variety of flow instabilities, the most common ones being the buoyancy driven *Rayleigh–Taylor instability* and the *Kelvin–Helmholtz instability* occurring in shear flows.

Rayleigh–Taylor instabilities occur, for example, when two fluids of different densities are superposed one over the other or accelerated towards each other (see figure 1). According to a linear stability analysis the arrangement of two fluids is stable if the lighter fluid is on top of the heavier fluid. Otherwise the arrangement of fluids is unstable.

Rayleigh–Taylor instabilities are ubiquitous in astrophysical flows. A particular example is Rayleigh–Taylor instabilities induced by the propagation of a shock wave through the envelope of a massive star which explodes as a (core collapse) SUPERNOVA (see figure 2). Since the energy given to the shock wave in a supernova explosion is much larger than the binding energy of the mass outside the collapsed core, gravity is dynamically unimportant for the propagation of the shock wave through the stellar envelope. In this case the Rayleigh–Taylor instability is driven by pressure and density gradients of opposite signs, the 'effective acceleration' being the (negative) pressure gradient.

Kelvin–Helmholtz instabilities occur whenever the velocity field of the flow has a gradient perpendicular to the (local) flow direction, i.e. in the presence of shear flow. The Kelvin–Helmholtz instability gives rise to a roll-up of the shear layer. It is responsible for the formation of mushroom shaped flow structures seen, for example, in large atmospheric explosions, and in connection with Rayleigh–Taylor instabilities (see figures 1 and 2). In the latter case the rising bubbles and sinking spikes resulting from the Rayleigh–Taylor instability cause a shear flow with respect to the surrounding matter.

Fluid dynamics and nuclear burning

Much of our understanding of astronomy and astrophysics is inseparably connected to our understanding of fluid

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

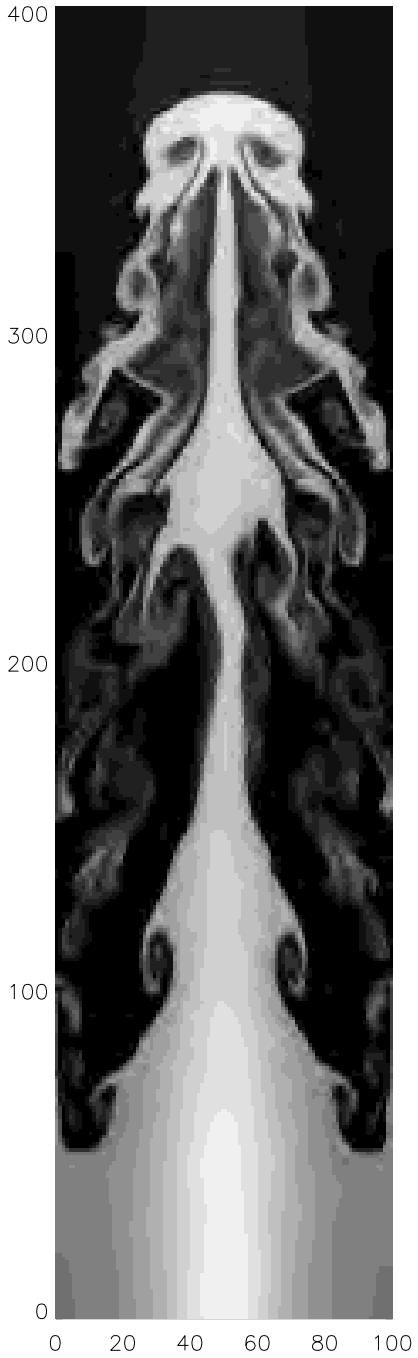


Figure 1. Rayleigh–Taylor instability in a dense fluid (bright) supported against gravity (directed upward!) by a lighter fluid (dark). The horizontal interface between the two fluids (initially located at 200) has been perturbed by a small sinusoidal perturbation. The typical structure of a narrow, high-density finger penetrating into the low-density region and a broader, low-density bubble at the edge of the grid penetrating into the high-density region is clearly visible. The mushroom cap structure on the top of the high-density region, which is created by the Kelvin–Helmholtz instability, is also apparent.

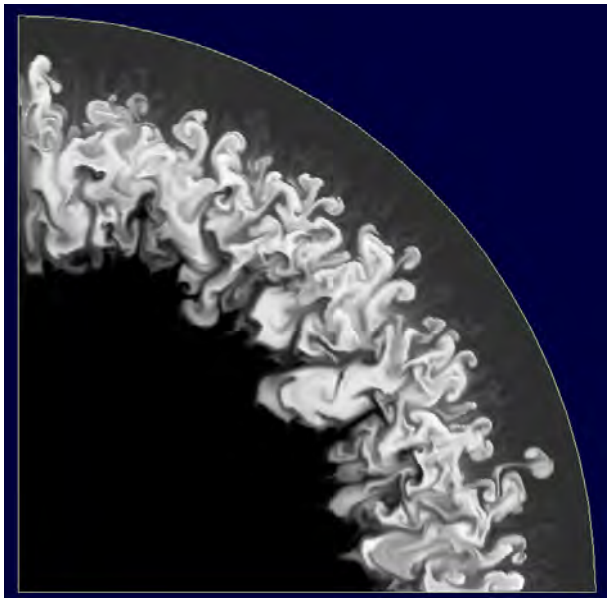


Figure 2. Simulation of Rayleigh–Taylor instabilities induced by the propagation of a shock wave through the envelope of a massive star which explodes as a (core collapse) supernova. The initially spherical helium shell (bright grey) of the progenitor star has been shredded by the instability. The snapshot is taken about 4 h after the onset of the explosion.

dynamics and nuclear burning. Thermonuclear processes and gravitational binding are the two major sources of energy for astronomical flows. The release of energy in an astrophysical PLASMA gives rise to changing pressure gradients and so to fluid motion. Such fluid motion, in turn, will transport ashes and fuel to and from the flame zone. This gives rise to a subtle and difficult problem. The degree of difficulty of the problem is closely related to the nature of the coupling between the nuclear burning and the fluid flow. Special cases, in which this coupling is simpler, can be treated more reliably. First, nuclear reactions release energy in the form of radiation and kinetic energy of particles; at stellar densities this energy acts to heat the medium in which the reactions occur. The heating can modify the pressure, and thereby modify the flow. Second, nuclear reactions change the composition of the material, which in turn may modify its equation of state and its opacity, which in turn may modify the pressure and the heating, and again the flow. Conversely, the flow may remove ashes and bring fuel to the flame, countering the natural tendency of the flame to exhaust its fuel.

Many different types of burning occur in astrophysical problems. Burning can range from quiet hydrostatic burning, such as occurs in the center of most stars, to the explosive burning which can occur, for example, in models of thermonuclear supernova explosions. Burning can occur in laminar or turbulent flows, the fluid motion being produced either by external causes and/or by the burning itself. Hence, the interaction between burning and flow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

must be considered.

Hydrostatic burning is the least violent form of burning and occurs in stars that are hydrostatic. In this case, the reaction rate is slow enough that energy can be transported from the burning region by radiation or conduction as rapidly as it is generated. A similar case is a subsonically convective flame region, in which heat and ashes are removed by slow convection, and new fuel brought in.

Detonations are the most violent form of burning one encounters. Astrophysical detonations generally occur only under degenerate conditions. In its simplest form, detonation wave theory is easily understood. As a shock propagates into unburned fuel, it compresses and heats the material beyond the ignition point. Immediately behind the shock is the reaction zone, in which the fuel burns. In simple detonation theory the width of the reaction zone is neglected, and the detonation front is treated as a sharp discontinuity. In this case jump conditions can be derived for the change in the fluid variables across the front in much the same way as is done for a simple shock. Note, however, that in order to obtain the post-detonation state one must first choose a detonation velocity, because unlike the case for simple shocks, the front velocity is not determined from the jump conditions. The simple detonation theory assumes that the reaction rate is infinite and that the reaction zone has zero width. A slightly more complex treatment is used in the Zeldovich–von Neumann–Doering model. This theory assumes that the shock, which is taken to be infinitely thin, is followed by a reaction zone of finite width. The final state obtained after the fuel is completely burned is exactly the same as for the simpler theory described above.

Deflagrations represent a much less violent form of burning than detonations, but in many ways are more complex. They result when the burning is unable to produce sufficient overpressure to create a shock which is strong enough to ignite the fuel. The motion of the front is usually very subsonic. Burning is initiated by the diffusive transfer of heat from the hot ashes behind the front into the cold fuel. Although, for the case of a thin front, deflagrations must obey the same jump conditions as detonations, the propagation velocity now depends on the rate of heat transfer. Compared to detonations another major difference is that the pressure and density decrease behind the deflagration front, and in the reference frame in which the front is stationary, the velocity increases. In the simplest theory deflagrations are described as flow discontinuities, which propagate subsonically. If the flow is laminar, a physically more accurate description is that of a conductive or diffusive laminar flame, whose width and speed is determined by the rate of energy generation and the transport of energy and composition. In most situations, however, even this description is too simple, because deflagrations or subsonic flames are subject to a variety of instabilities, which can significantly influence their shape (wrinkling of the flame) and propagation speed. Moreover, because the propagation speed of the flame is subsonic, flame and flow do interact. In particular, if burning occurs in a convective or turbulent environment, depending on the relevant length scales and time scales, physically quite different combustion regimes must be considered.

Relativistic fluid dynamics

Some astrophysical phenomena require a relativistic formulation of fluid dynamics. A special relativistic formulation must be used if the flow velocity reaches a significant fraction of the speed of light, or if the internal energy of the fluid becomes comparable to its rest mass. Such conditions occur in jets from quasars, active galactic nuclei and some galactic compact binaries, and in gamma-ray burst sources. General relativistic flows involving strong gravitational fields and relativistic speeds are encountered in accretion flows onto compact objects like neutron stars or black holes, in the formation or merging process of neutron stars and black holes, and are most likely at the heart of the processes leading to the formation of relativistic jets and gamma-ray bursts.

An important difference between Newtonian and relativistic fluid dynamics is the presence of a maximum velocity, i.e. the speed of light in a vacuum in the latter case. A flow can be relativistic in two ways. In a kinematically relativistic jet the beam gas moves with relativistic velocity, while in a thermodynamically relativistic jet the specific internal energy of the beam gas is large compared with its rest mass energy.

Special relativistic astrophysical flows are encountered in extragalactic jets (see ASTROPHYSICAL JETS). These are highly collimated supersonic outflows from ACTIVE GALACTIC NUCLEI. Extragalactic jets are thought to be formed as a consequence of mass accretion onto a central rotating super massive BLACK HOLE in an active galactic nucleus being fed by interstellar gas and gas from tidally disrupted stars. Most models of the (initial) collimation and acceleration of astrophysical jets involve magnetohydrodynamic processes at or near the inner edge of a rotating magnetized accretion disk with corona. The combined action of thermal and magnetic pressure gradients, of centrifugal force and of gravity drives an outflow, which does collimate itself. General relativistic effects seem to be crucial for a successful launch of the jet. A Newtonian treatment of extragalactic jets is justified only at distances larger than several kiloparsec from the central black hole. At smaller scales, and in particular at parsec scales, jets propagate at relativistic speeds. Hence, special relativistic hydrodynamic simulations must be used in order to describe their propagation correctly (see figures 3 and 4).

Equations for ideal special relativistic flows

The equations of special relativistic fluid dynamics for an ideal fluid are most conveniently expressed as a set of conservation equations

$$\frac{\partial D}{\partial t} + \mathrm{div}(D\boldsymbol{v}) = 0$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

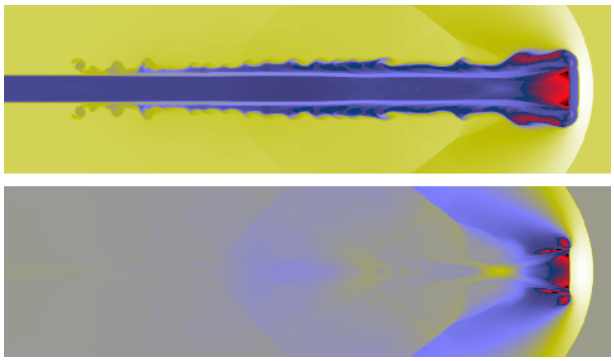


Figure 3. Color-coded contour plots of the logarithm of the proper rest-mass density (top) and pressure (bottom) of a hot, relativistic jet ($v_b = 0.99c$, $Ma = 1.72$). The maximum values are coded in white. Decreasingly smaller values are coded in green, bright blue, dark blue, red and the minimum values are coded in black. The jet has a 'naked' beam and almost completely lacks internal structure. The beam is surrounded by lobes (near the jet head) instead of a cocoon and is in pressure equilibrium with the shocked ambient medium. **This figure is reproduced as Color Plate 20.**

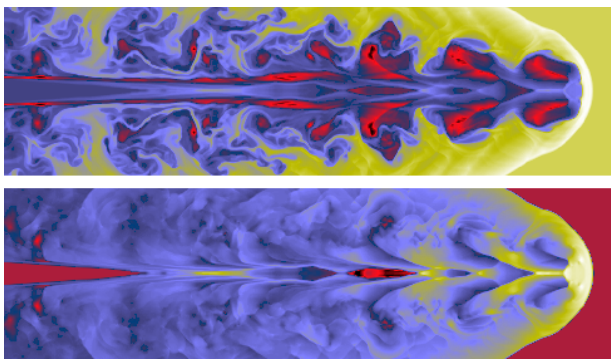


Figure 4. Same as figure 3, but for a cold, relativistic jet ($v_b = 0.99c$, $Ma = 6.0$). The jet has a beam with internal oblique shocks and an extended turbulent cocoon. **This figure is reproduced as Color Plate 21.**

$$\frac{\partial S^i}{\partial t} + \mathrm{div}(S^i \boldsymbol{v}) + (\mathrm{grad}\, p)^i = 0, \qquad i = 1, 2, 3$$

$$\frac{\partial \tau}{\partial t} + \mathrm{div}[(\tau + p)\boldsymbol{v}] = 0$$

for the conserved rest-mass density D, momentum density $\boldsymbol{S} = (S^1, S^2, S^3)$ and energy density τ, respectively. These variables, which are defined in a fixed frame, are related to quantities in the local rest frame of the fluid through

$$D = \rho W$$
$$\boldsymbol{S} = \rho h c^{-2} W^2 \boldsymbol{v}$$
$$\tau = \rho h W^2 - p - Dc^2$$

where ρ, p, $\boldsymbol{v}$, W, h and c are the proper rest-mass density, the pressure, the flow velocity, the flow Lorentz factor $W = (1 - v^2/c^2)^{-1/2}$, the specific enthalpy $h = c^2 + \varepsilon + p/\rho$, and the speed of light in a vacuum.

Numerical integration of the fluid dynamic equations

In astrophysical applications one usually has to solve the hyperbolic set of the inviscous and compressible Euler equations instead of the parabolic set of the viscous and compressible Navier–Stokes equations, because (except in shock waves) viscosity and heat conduction due to physical processes are often negligibly small. However, because of truncation errors and because of finite spatial resolution, all numerical methods proposed for the solution of the Euler equations possess some finite intrinsic numerical viscosity, whose form and size depends on the numerical method, on the grid resolution, and to some extent of the flow problem itself. The effect of the numerical viscosity on the solution can only be quantified by resolution studies, i.e. by performing a series of simulations of the same flow problem using different grid resolution. Thus, when simulating astrophysical flow one is in the strange situation that one tries to solve the inviscous Euler equations, but instead solves some viscous variant, which is also different from the Navier–Stokes equations.

The evolution equations are a set of coupled nonlinear first-order (in time) partial differential equations. One way of solving this set of equations is to discretize them in time and space. By this discretization the partial differential equations are transformed into a set of coupled nonlinear algebraic equations, which can be solved on a computer with appropriate numerical techniques. The algebraic equations are not uniquely determined, because a given differential equation can be discretized in different ways. Obviously, one tries to choose those algebraic systems which minimize the truncation errors, and which possess some additional properties, like being stable and consistent. This is at the heart of the 'art of computing'.

Guided by the conservation properties underlying the hydrodynamic equations applied mathematicians and physicists have been able to develop accurate and stable high-resolution finite volume schemes In finite volume schemes the (finite) computational domain is discretized into a finite number of zones or cells (typically 10^2 to 10^3 per spatial dimension). The hydrodynamic scalar fields (e.g. density) and the components of the vector fields (e.g. momentum) are approximated by a set of discrete values, which are the (approximate) cell averages of the respective variable. Time is discretized too, the evolution being approximated by a series of finite time steps (typically 10^3 to 10^5 per simulation).

Besides finite volume schemes, another completely different method is widely used in astrophysics for integrating the hydrodynamic equations. This method is smoothed particle hydrodynamics, or SPH for short. In SPH the fluid or gas is described by a finite set of 'particles' (typically 10^3 to 10^5), which move with the flow. SPH is a free-Lagrange method, in which spatial gradients are evaluated without the use of a grid.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The literature on fluid dynamics is extensive. Thus, the following references can only provide entry points into the field.

The derivation of the fluid dynamic equations using the phenomenological approach can be found, for example, in the book by

Shu F H 1991 *The Physics of Astrophysics* (Mill Valley, CA: University Science Books)

The kinetic theory approach to the equations of fluid dynamics is described in detail by

Hirschfelder J O, Curtiss C F and Bird R B 1964 *Molecular Theory of Gases and Liquids* (New York: Wiley)

A general introduction to fluid dynamics can be found in

Landau L D and Lifshitz E M 1982 *Course of Theoretical Physics, Vol.VI, Fluid Mechanics* (Oxford: Pergamon)

For a modern description of numerical techniques for the solution of the equations of fluid dynamics one can consult, e.g.

Laney C B 1998 *Computational Gasdynamics* (Cambridge: Cambridge University Press)

or with special emphasis on astrophysical fluid flow including also various astrophysical applications

LeVeque R J, Mihalas D, Dorfi E A and Müller E 1998 *Computational Methods for Astrophysical Fluid Flow* ed O Steiner and A Gautschy (Berlin: Springer)

Ewald Müller

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flux Density

A measure of the radiation arriving from a source at a particular frequency. The *flux* of radiant energy is the quantity of energy per second (watts) passing through unit surface area (1 meter squared) perpendicular to the direction of the source and expressed as W m^{-2}. The flux density is the flux of radiant energy within a unit interval of frequency (a frequency band with a width of 1 hertz) and is expressed as W m^{-2} Hz^{-1}.

The flux densities for most cosmic radio sources are extremely low. A unit of flux density used by radioastronomers is the jansky (Jy), where 1 Jy = 10^{-26} W m^{-2} Hz^{-1}.

See also: radioastronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Focal Length

The distance between the center of a lens, or the reflective surface of a mirror, and its focal point, or focus (the point at which rays of light from a point source, located effectively at an infinite distance, converge to form an image).

See also: focal ratio, mirror, objective lens.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Focal Ratio

The ratio of the focal length of a lens or mirror to its aperture. For a lens or mirror of focal length, F, and aperture, D, the focal ratio, f, is given by $f = F/D$. For example, a lens of 10 m focal length and 2 m aperture would have a focal ratio of $10/2 = 5$; this would be written as 'f:5'.

The focal ratio, or f-number, of a camera lens may be altered by means of an adjustable diaphragm that changes its aperture. For a lens or mirror of given focal length, increasing the aperture (and thereby reducing the focal ratio) increases the illumination of the image. Because, in general photography, this decreases the exposure time that is needed in order to record a particular image, the term 'focal ratio' is synonymous with the 'speed' of an optical system (for example, an f:2 system is faster than an f:16 system).

See also: aperture, focal length.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Focus

(1) In optics, the point at which parallel light rays from a distant point source are brought together by a lens or mirror to form an image. Light rays traveling parallel to the optical axis (the line that passes perpendicularly through the center of a lens or mirror) of an ideal convex lens or concave mirror converge to intersect at the focus, or focal point, of the lens or mirror. Rays arriving at an angle to the optical axis, from point sources or extended objects (sources of finite angular size), form images at the focal plane, a plane that passes through the focal point perpendicular to the optical axis. In practice, the focal plane may be slightly curved.

In a conventional refractor, light is brought to a focus by the objective lens, the distance between the objective and the focus being the focal length. In a reflecting telescope, the focal point of light rays reflected from the primary mirror (the mirror that collects incoming light) is called the prime focus. Instrumentation, or even an observer, may be placed directly at the prime focus of a large reflector. Most reflecting telescopes use one or more additional mirrors to reflect the converging cone of light, and the focal point, to a different location. In the Newtonian reflector, a small flat mirror, set at an angle of 45° to the optical axis, reflects the converging cone of light to a focal point (the Newtonian focus) at the side of the telescope tube. In the Cassegrain system, a convex secondary mirror, located a short distance before the prime focus, changes the angle of convergence of the rays and reflects light back through a central hole in the primary to the Cassegrain focus. Because of the magnifying effect produced by their curved secondaries, Cassegrain and related systems have effective focal lengths that are several times longer than the focal length of their primary mirrors.

In the Nasmyth system, which can be used with telescopes on altazimuth mountings, light reflected from a convex secondary passes back down the telescope tube and is then reflected from a flat tertiary (third), through a hollow altitude axis (the axis round which the telescope turns in elevation), to a focus (the Nasmyth focus) that remains in the horizontal plane no matter in what direction the telescope itself is pointing. In the coudé system, which may be used with equatorially mounted telescopes, light reflected from a convex secondary is then reflected from a plane tertiary mirror to a plane quadric (fourth) mirror which, in turn, reflects light down the polar axis of the mounting to a fixed focal position—the coudé focus.

See also: altazimuth mounting, Cassegrain telescope, equatorial mounting, mirror, Newtonian telescope, objective lens, reflector, refractor.

(2) In an ellipse, a focus (plural: foci) is one of two points on the major axis such that the sum of the distances from each of the two foci to any point on the ellipse is a constant.

See also: conic sections, ellipse.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fomalhaut

The star α Piscis Austrinus, 'the Mouth of the Fish', apparent magnitude 1.17, spectral type A3V. Situated at a distance of 25.1 light-years (parallax 0.130″), the star has an absolute magnitude of 1.7. Although a white star, it often appears reddish to northern hemisphere observers, due to a combination of its low declination (-29.6°) and atmospheric pollution. It is notable for observers in the British Isles as the southernmost first-magnitude star visible to them, just above the southern horizon in September and October. From southern England it reaches a maximum meridian altitude of about 10° at the end of September.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fontenelle, Bernard le Bouyer de [Bernhard le Bovier de] (1657–1757)

Born in Rouen, France, a Cartesian philosopher, wrote on the history of mathematics and the philosophy of mathematics and science; wrote *Entretiens sur la Pluralité des Mondes* (1686) in which he argued for the existence of other worlds—planets orbiting other stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Force-Free Magnetic Fields

In MAGNETOHYDRODYNAMICS (MHD), which describes the fluid behavior of PLASMAS, the magnetic field $\boldsymbol{B}$ exerts a force, the Lorentz force, given by $\boldsymbol{F} = (1/c)\boldsymbol{J} \times \boldsymbol{B}$, in terms of the field and its electric current density $\boldsymbol{J} = (c/4\pi)\nabla \times \boldsymbol{B}$, where c is the speed of light. In many circumstances, the electrical conductivity of the plasma is so large as to be taken infinite, in which case electric currents change only by the inductive effect of the flowing plasma, with negligible Ohmic decay. In the absence of plasma motion, electric currents, once set up, persist with no decay. A particular form of such an equilibrium state is one in which the current density is parallel to the magnetic field, as described by the equation

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \alpha \boldsymbol{B}. \tag{1}$$

The Lorentz force vanishes in the plasma and the magnetic field is said to be force free. In equation (1), the proportionality factor α is allowed in general to vary in space, subject to the condition

$$\boldsymbol{B} \cdot \boldsymbol{\nabla}\alpha = 0 \tag{2}$$

requiring α to be constant on a line of force, in order that Maxwell's equation

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{3}$$

is satisfied.

Force-free fields are expected in the outer solar atmosphere where the plasma is electrically highly conducting but is so tenuous that its pressure and other body forces may be neglected compared with the typical strength of the Lorentz force that the magnetic field and its electric current are capable of producing. The Lorentz force $\boldsymbol{F}$ may be written as $\boldsymbol{F} = -\nabla(B^2/8\pi) + (1/4\pi)(\boldsymbol{B} \cdot \nabla)\boldsymbol{B}$, in terms of a pressure force and a tension force, which mutually cancel when the magnetic field is force free. The magnetic pressure $B^2/8\pi$ is a measure of the magnitude of the Lorentz force the field is capable of exerting. Force-free magnetic fields are therefore expected when the ratio of the gas pressure to the magnetic pressure, denoted by β in the plasma physics literature, is much smaller than unity.

Equations (1) and (2) pose a nonlinear problem for the magnetic field except when the function α is a constant. In this exceptional case, equation (2) is trivially satisfied, and solving equation (1), for some constant value of $\alpha = \alpha_0$, may be transformed to solving the vector Helmholtz equation

$$\boldsymbol{\nabla}^2\boldsymbol{B} + \alpha_0^2\boldsymbol{B} = \boldsymbol{0} \tag{4}$$

subject to suitable boundary conditions formulated for a physical problem. The astrophysical and plasma physics literature contains many works on the solutions to this linear problem. The general nonlinear problem is not readily treated, with few known solutions. In the following, we describe briefly some of the general physical properties of force-free magnetic fields.

The nature of the Lorentz force is such that the magnetic pressure force may be balanced by the magnetic tension force only in some bounded region of space. The magnetic pressure has an outward exertion which cannot be globally confined by the magnetic tension force alone. This is a fundamental property expressed by Chandrasekhar's virial relation:

$$\int_V \frac{B^2}{8\pi}\,\mathrm{d}V = \frac{1}{4\pi}\int_{\partial V}\left[\frac{1}{2}B^2\boldsymbol{r}\cdot\,\mathrm{d}\boldsymbol{S} - (\boldsymbol{B}\cdot\boldsymbol{r})(\boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S})\right] \tag{5}$$

valid for any force-free field $\boldsymbol{B}$ in some volume V with boundary surface ∂V. If the force-free field is self-confining, we could take the volume V to be all space, in which case the surface integral in equation (5) would vanish since the field intensity would fall with distance as fast as a dipole magnetic field. A contradiction then arises from equation (5) unless $\boldsymbol{B} = \boldsymbol{0}$. This classical result implies the impossibility of creating a magnetically self-confining plasma without rigid walls in the laboratory. In the astrophysical context, this property shows that force-free magnetic fields need to be anchored to regions where the magnetic field exerts nonzero Lorentz force on the embedding plasma.

In the solar atmosphere, only the magnetic field in the thin layer called the photosphere can be measured with any useful spatial resolution. Equation (5) offers a means of estimating the energy of the magnetic field in the atmosphere above, taken in the force-free approximation, in terms of the measured vector field at the photosphere. In this popular model, the volume V is often taken to be the space above the photosphere treated as an infinite plane. A basic motivation of this modeling work is to demonstrate the amount of free energy that can be stored in the parallel electric currents of a force-free field. The stored energy is believed to be the origin of the enormous amount ($\sim 10^{31-32}$ erg) liberated in a large SOLAR FLARE. The application of force-free fields to the solar atmosphere has served to illustrate observed magnetic structures and to illuminate basic physical points. Application with the use of measured photospheric vector fields as inputs is an involved and difficult numerical simulation effort, as a result of nonlinearity, difficulties associated with the proper posing of the boundary conditions and the ill-posed nature of the mathematical problem.

In a perfect electrical plasma conductor, the magnetic flux is frozen into the plasma so that the field topology cannot change in time. If a magnetic field, not necessarily in force equilibrium, is taken to be given and is then deformed by displacing the embedding perfectly conducting plasma to an extremum in its total magnetic energy, the result would be the force-free field as described by equations (1) and (2), where the scalar function α takes a distribution in space that renders the solution $\boldsymbol{B}$ topologically identical to the given initial field.

The variational problem stated above is difficult to treat but has very interesting physical implications.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The solution to the variational problem may not be unique, in which case the nonunique solutions may be distinguished between the force-free fields which are linearly stable and those which are linearly unstable (see MAGNETOHYDRODYNAMIC INSTABILITIES). Linear magnetohydrodynamic stability of force-free magnetic fields has been treated extensively in the literature, with the general result that the rigid anchoring of the force-free field lines to the base of an atmosphere provides some degree of stability so long the lines of force are not too lengthy within the atmosphere. Most of the force-free fields studied in linear stability analyses are one- or two-dimensional fields. Extension to non-trivial variation in three-dimensional space, except for the case of constant α fields, is limited although some simple examples have been found, including some with realistic geometry suggestive of observed solar magnetic structures. These examples were found to be linearly stable.

More interesting is the possibility that an extremum in magnetic energy corresponds to a force-free field which necessarily contains magnetic tangential discontinuities in order to satisfy the requirement that the force-free field has a prescribed field topology. Mathematically, this is related to the fact that, in variational problems, the set of smooth minimizing functions to approach an extremum is not necessarily compact. If the extremum magnetic field must have a tangential discontinuity and is a minimum in energy, a magnetic field having the same magnetic topology may physically relax into such a state with the formation of the inevitable tangential discontinuity. Physical plasmas have large but finite electrical conductivities so that the formation of the tangential discontinuity in such a plasma would proceed until the thickness of the discontinuity is small enough for the electrical resistivity, however weak, to dissipate the sheet. This process is a consequence of high electrical conductivity, which, remarkably, creates the physical circumstance for inevitable dissipation of electric currents. The idea was proposed by E N PARKER as the origin of plasma heating under astrophysical conditions.

The turbulent MHD relaxation of plasmas at large magnetic Reynolds' numbers proceeds by the formation of magnetic tangential discontinuities which dissipate by MAGNETIC RECONNECTION (see also MAGNETOHYDRODYNAMICS: MAGNETIC RECONNECTION AND TURBULENCE). With resistive effects important only at these reconnection sites, the change in magnetic field topology proceeds in such a way that, as a measure of the complexity in the field topology, the MAGNETIC HELICITY is conserved in the global volume of the plasma, as opposed to the more stringent requirement of no change in field topology. This global conservation of magnetic helicity is a constraint that sets the extremum in the total magnetic energy to be one of a force-free field having a constant $\alpha = \alpha_0$, described by equation (4). The value of α_0 is then fixed such that the linear force-free field has the conserved amount of total magnetic helicity. This provides a physical basis for singling out the linear force-free fields from the nonlinear ones. Application of this theory due to B Taylor to the laboratory plasmas has been very successful. Extending the application to solar and astrophysical magnetic fields is not straightforward because of the problem of gauge dependence of magnetic helicity in the case of magnetic fields which thread across the boundary of its physical domain.

Bibliography

Canfield R and Pevtsov A (ed) 1999 *Magnetic Helicity in Space and Laboratory Plasmas* (Washington, DC: American Geophysical Union)

Low B C 1988 *Astrophys. J.* **330** 992

Parker E N 1994 *Spontaneous Current Sheets in Magnetic Fields* (Oxford: Oxford University Press)

Priest E R 1982 *Solar Magnetohydrodynamics* (Dordrecht: Reidel)

Boon Chye Low

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fornax

(the Furnace; abbrev. For, gen. Fornacis; area 398 sq. deg.) A southern constellation which lies between Sculptor and Eridanus, and culminates at midnight in early November. It was introduced as Fornax Chemica (the Chemical Furnace) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2. The shortened form was adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest star in Fornax is α Fornacis, a binary with pale yellow and yellow (F6 and G7) components, magnitudes 3.8 and 7.0, separation 4.5″, period 314 years (?). There are no other stars brighter than fourth magnitude. Interesting objects include the Fornax Dwarf Galaxy (ESO 365-G4, also called the Fornax System), a dwarf spheroidal galaxy in the Local Group, about 540 000 light-years distant, NGC 1365, a ninth-magnitude barred-spiral galaxy, and NGC 1316, another ninth-magnitude spiral which is the optical counterpart of the strong radio source Fornax A. The last two are members of the conspicuous Fornax Cluster of galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Foucault, Jean Bernard Léon (1819–68)

Physicist, born in Paris. Became professor at the Paris Observatory and member of the Longitude Institute. Collaborated with FIZEAU in the measurement of the velocity of light using the Fizeau wheel. After they quarrelled, independently of Fizeau he measured the speed of light in water, and found it to be slower than in air, as predicted by the wave theory of light, and against the corpuscular theory. First scientist (with Fizeau) to photograph the Sun (on a daguerreotype). Developed the *Foucault pendulum*, at first hanging a 2 m long pendulum in his basement, and a 10 m one in the Paris Observatory, so watching the rotation of the pendulum's oscillating path and confirming the rotation of the Earth. After this he carried out a demonstration with a very long pendulum (60 m) in the dome of the Pantheon, causing a public sensation. Showered with honours, and appointed to the staff of the Paris Observatory, he set about improving its telescopes, in the course of which he invented the knife-edge test for the figure of a mirror, the process of silvering glass mirrors (replacing the speculum metal mirror), and pneumatic mirror support systems.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Foucault's Pendulum

A freely swinging pendulum mounted in such a way as to reduce its physical contact with the Earth to an absolute minimum. An ideal pendulum of this type would, if set swinging in a particular plane at one of the Earth's poles, continue to swing in that same plane relative to the 'fixed stars', and the Earth would rotate underneath it. To the observer on the Earth's surface, the pendulum would be seen to rotate its plane of swing through 360° each day. The period of rotation of the plane of the pendulum increases with decreasing latitude on the Earth's surface. No rotation is observed for a pendulum at the equator. Observation of the behavior of such a pendulum is direct evidence of the fact that the Earth rotates on its axis. This type of pendulum is named after the French physicist, Léon Foucault (1819–68).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fowler, Alfred (1868–1940)

Professor at the Imperial College of Science and Technology, in South Kensington, London, a master of spectroscopy, especially for solar phsyics. He identified titanium oxide in cool stars, magnesium hydride in sunspots, carbon monoxide in comets and ozone in the Earth's upper atmosphere. When the Bohr theory of the atom appeared, he made contributions to atomic physics, identifying the spectrum of ionized helium and determining the value of the Rydberg constant and the proton–electron mass ratio. Organized, and became the first Secretary of, the International Astronomical Union.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fowler, William Alfred ['Willy'] (1911–95)

American nuclear physicist, born in Pittsburgh, PA, worked at CalTech. Nobel prizewinner (1983). While at the Kellogg Laboratory studied the nuclear reactions of protons with the isotopes of carbon and nitrogen, the very reactions in the CN cycle identified at that time by HANS BETHE as occurring in the stars. Established the science of nuclear astrophysics. Found that there was a gap in the sequence of stable nuclei at mass 8, as at mass 5. These mass gaps spelled doom for GEORGE GAMOW'S brilliant idea (the alpha-beta-gamma theory) that all nuclei heavier than helium (mass 4) could be built by neutron addition, one mass unit at a time, in the Big Bang. Measured the rates of nearly all nuclear reactions of astrophysical interest, and co-authored, with Margaret Burbidge, Geoffrey Burbidge and FRED HOYLE, the famous 1957 paper, 'Synthesis of the elements in stars', known as B^2FH. They showed how the cosmic abundances of essentially all the nuclides from carbon to uranium could be explained as the result of nuclear reactions in stars, using hydrogen and helium produced in the Big Bang. After 1964 Fowler worked on theoretical problems involving supernovae, gravitational collapse and nucleo-cosmo-chronology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Franklin-Adams, John (1843–1912)

English businessman and amateur astronomer; compiled the *Franklin-Adams Photographic Atlas of Star Positions* (1913). The *Franklin-Adams camera* was a 25 cm aperture, 10 degree field telescope, later re-erected in Johannesburg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fraunhofer Lines

Dark lines in the spectrum of the Sun, first studied by Joseph von Fraunhofer (1787–1826) in 1814. He labeled the most prominent of these by the letters A to K, ranging from the red end of the spectrum to the violet. Most are absorption lines due to the absorption of light by atoms in the outer regions of the Sun; for example, a prominent pair of close dark lines in the yellow part of the solar spectrum is the 'D' lines of the element sodium. A few Fraunhofer lines orignate in molecules in the Earth's atmosphere.

See also: solar spectroscopy, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fraunhofer, Joseph von (1787–1826)

Instrument-maker, born in Straubing, Germany, began life as a wood turner then as a glazier, educated himself in optics. Joined the instrument workshop of Georg Reichenbach and Joseph Utzschneider making and developing lenses, and taking up optical glass making for the microscopes, opera glasses, loupes and the small and large astronomical telescopes and heliometers which the company built. Fraunhofer calculated, designed and tested every instrument, wrote the manuals and watched the disassembling and packaging process for the bigger telescopes, including the big refractor for the Russian observatory in Dorpat, used by STRUVE to measure 3000 double stars. This was the first telescope mounted equatorially and driven about the polar axis with a clock. Around 1813 Fraunhofer began to research the chromatic properties of different sorts of glass and rediscovered the dark lines in the Sun's spectrum, as first seen by WILLIAM WOLLASTON and now known as *Fraunhofer lines*. Fraunhofer identified and measured the wavelengths of 574 lines, cataloging them with letters (A to Z for the strong ones, as for example the 'sodium D-lines', lower case letters for the weaker, e.g. the magnesium b-line), which are still used today. His motivation for the study seems to have been to provide wavelength standards to make refractive index measurements although his work is remembered as pure science. He followed up this work with studies of the spectra of the light of electric sparks, different light sources and stars. Fraunhofer sold his firm to Merz.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Freja

Swedish–German satellite, launched in October 1992. Carried instruments to study Earth's aurora and associated wave and particle phenomena.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Frequency

The rate at which specific events occur. In the context of electromagnetic radiation, the term is taken to mean the number of wavecrests per second passing a particular point. For light of wavelength λ, since the speed of light $= c$, the frequency, $f = c/\lambda$.

The SI unit of frequency measurement is the hertz (Hz), measuring the number of waves passing a fixed point per second. Thus, 1 wave per second would correspond to a frequency of 1 Hz, 100 000 waves per second to 100 000 Hz, etc. Common multiples are kilohertz (kHz) = 1000 Hz, megahertz (MHz) = 1000 000 Hz, and gigahertz (GHz) = 1000 000 000 Hz. An older unit of frequency is cycles per second (c/s); 1 Hz = 1 c/s.

See also: wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Freundlich, Erwin Finlay (1885–1964)

Born in Biebrich, Germany, collaborator with EINSTEIN, making measurements (successfully) of Mercury's orbit and the gravitational redshift (unsuccessfully) to confirm the general theory of relativity. Fled Nazi Germany to become the Napier Professor of Astronomy at the University of St Andrews in Scotland.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Friedman, Herbert (1916–)

Astrophysicist, born in New York City, worked at the US Naval Research Laboratory in Washington, used rockets for astronomical observations, including solar x-ray astronomy using Geiger counters and detectors developed from them. Showed in 1964 that a celestial x-ray source coincided with the Crab Nebula supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Friedmann [Fridman], Aleksandr [Alexander] Aleksandrovich (1888–1925)

Mathematician, born in St Petersburg, Russia, where he studied and first worked. After exciting war-time and revolutionary experiences he returned to Petrograd (as it had been re-named, before becoming Leningrad and now St Petersburg again) where he taught mathematics and mechanics. Friedmann took up an interest in EINSTEIN's general theory of relativity, and showed in a classic paper 'On the curvature of Space' that the radius of curvature of the universe can be either increasing or a periodic function of time. Einstein and de Sitter had considered a stationary type of universe, and Einstein at first claimed that Friedmann's solution does not satisfy the field equations, but conceded after further study that Friedmann was right. Friedmann's solution foreshadowed the concept of the expanding universe. Just as COPERNICUS made the Earth orbit round the Sun, so Friedmann made the universe expand.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Frost, Robert (1874–1963)

American lyrical poet, born in San Francisco, amateur astronomer. His poetry contains numerous astronomical images, including the image of the farmer who, unknowingly, picked up a meteorite and built it into a prosaic stone wall. He had a telescope in his farmhouse in Derry, New Hampshire and each of his children was given the task of following one star (cf. his poem, *Choose Something Like a Star*).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fuertes Observatory

Located on the North Campus of Cornell University, Ithaca, New York, it contains a 30 cm refractor. Earlier in the twentieth century, the observatory was used to tell the time by noting when certain stars passed overhead. Today it is used largely as a teaching facility for introductory astronomy classes. It is also used by the Cornell Astronomy Club.

For further information see http://astrosun.tn.cornell.edu/teach/fuertes.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fundamental Catalogs

Catalogs that list stars whose positions and proper motions have been very accurately determined over many years by positional measurements relative to telescopes, rather than to other stars. A number of early catalogs are still of value because the determination of positions and proper motions is an on-going exercise in which old observations are combined with the latest measurements; a notable example is the *General Catalog of 33 342 Stars* (*abbrev.* GC) compiled by the American astronomer Benjamin Boss (1880–1970) and published in five volumes in 1936–7.

The most important fundamental catalogs are the FK series, begun in Berlin in 1879. The third of these, the *Dritter Fundamentalkatalog des Berliner Astronomischen Jahrbuchs* (*abbrev.* FK3), published in 1934 contained the positions of 1535 stars with a limiting magnitude of about 7.5, for the epoch 1950.0. It was adopted as the standard list of fundamental stars by the International Astronomical Union (IAU) in 1935, and a series of annual volumes of *Apparent Places of Fundamental Stars* was commenced in which the predicted positions of the fundamental stars are tabulated at 10 day intervals, using the FK data.

Succeeding catalogs, the FK4 and FK5, were published in 1963 and 1988 respectively by the Astronomisches Rechen-Institut, Heidelberg, on behalf of the IAU. The catalog positions are calculated for a standard epoch, so that they provide a fundamental reference frame to which observations of the positions of all other bodies may be referred. Precise positional observations of the fundamental stars made at many observatories around the world, mainly with transit circles (and, in more recent years, photographic zenith tubes and prismatic astrolabes), are combined to update and improve the catalog positions and proper motions for this list of fundamental stars.

See also: fundamental stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fundamental Forces

The forces that govern the various interactions between particles. The four fundamental forces, or interactions, are: gravitation, the electromagnetic force, the strong nuclear force, and the weak nuclear force.

The gravitational force acts between all particles and is always attractive; it pulls matter together. Gravitation is the force that binds together the solar system, the galaxies, and clusters of galaxies. The electromagnetic force controls the behavior of electrically charged particles (particles of like charge repel, whereas opposite charges attract). The electromagnetic force is the binding force of atoms; the electromagnetic interaction between its positively charged nucleus and its surrounding cloud of negatively charged electrons holds an atom together. The electromagnetic force also controls the absorption and emission of electromagnetic radiation. The strong nuclear force binds together the nucleons (protons and neutrons) in an atomic nucleus and prevents the like-charged protons from flying apart. Particles that are acted on by the strong nuclear force (including protons and neutrons) are known collectively as hadrons. The strong nuclear interaction between individual hadrons is believed to be a remnant of a more powerful color force that acts between quarks inside hadrons. The weak nuclear force causes the radioactive decay of certain atomic nuclei. In particular, it governs the process called beta decay whereby a neutron breaks up into a proton, an electron and an antineutrino.

The strong and weak nuclear interactions are short-range forces that are effective only within the nuclei of atoms, the range of the strong force being about 10^{-15} m and that of the weak force about 10^{-17} m. The electromagnetic and gravitational interactions are long-range forces, their strengths being inversely proportional to the square of distance (if the distance is doubled, the force is reduced to a quarter of its previous value). Thus, in principle, the gravitational attraction between two masses, or the electromagnetic force between two charged particles, does not decline to zero until the separation between them becomes infinite. Although in terms of interactions between individual particles, gravitation is far weaker than the electromagnetic force, because matter on the large scale tends to be electrically neutral, gravitation controls the overall dynamics of planets, stars, and galaxies.

In the present-day universe, at everyday energy levels, the four forces are distinct from one another and have different strengths (in order of decreasing strength: strong nuclear, electromagnetic, weak nuclear, gravitation). At very high energies (particle energies in excess of about 10^{11} eV) the weak and electromagnetic forces merge into a single 'electroweak' force. According to Grand Unified Theories (GUT), the strong and electroweak forces will lose their separate identities and behave as a single unified force at particle energies in excess of about 10^{24} eV (about 10^{12} times higher than can be achieved in present-day particle accelerators). Although a complete theory (a 'theory of everything') that embraces all four forces has not yet been achieved, it is widely believed that at still higher energies, gravitation will unite with the others into a single 'superforce'. If this hypothesis is valid, the universe was dominated by the superforce in the first microscopic instant of the Big Bang. Thereafter, as the universe expanded and cooled (and particle energies decreased), the unified force broke up into the four separate forces.

According to modern quantum theories, the various fundamental forces are conveyed between real particles by means of 'virtual' particles (particles that are permitted temporary existence by the Heisenberg uncertainty principle). The force-carrying particles, or 'gauge bosons', for each of the forces are as follows: electromagnetic force, photons; weak nuclear interaction, intermediate vector bosons; strong nuclear interaction between hadrons, pions (strong interaction, or 'color force', between quarks, gluons). The existence of the hypothesized force-carrying particle for gravitation (the graviton) remains unproven.

According to Kaluza–Klein theory, which is named after physicists Theodore Kaluza and Oscar Klein, the universe began with more than four dimensions (current versions of this kind of theory require at least ten dimensions), of which only four expanded to form space (which has three dimensions—length, breadth, height) and time, the others remaining so tightly 'curled up' on themselves that they are not directly perceptible. Theories of this kind suggest that the fundamental forces are manifestations of these hidden dimensions.

See also: electroweak force, gauge bosons, grand unified theory, hadrons, uncertainty principle, mesons, quantum theory, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fundamental Particles

Subatomic particles that cannot be divided into smaller components. The fundamental particles include leptons (electrons, muons, tau particles, and neutrinos), quarks and gauge bosons particles which, so far as is known at present, do not have an internal structure composed of other, more basic, particles.

See also: elementary particles, gauge bosons, leptons, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fundamental Stars

A definitive list of 1535 stars, adopted by the International Astronomical Union in 1935, whose positions and proper motions have been very accurately determined over many years and are regularly updated by positional observations of very high precision. The data are regularly published in fundamental catalogs and provide a reference-frame for position- and motion-determining observations of all other bodies.

See also: fundamental catalogs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fusion

The process in which two lighter atomic nuclei are combined to form a heavier atomic nucleus. Very high temperatures are normally required in order for atomic nuclei to collide with sufficient energy to overcome their mutual electrostatic repulsions (each atomic nucleus has a positive charge, the magnitude of which depends on the number of protons it contains). Fusion that occurs under high-temperature conditions is called thermonuclear fusion.

Fusion reactions involving light elements release large amounts of energy. The mass of the resulting nucleus is less than the combined masses of the two original nuclei, the difference in mass, Δm (which is known as the mass defect) being released as energy (E) in accordance with Einstein's relationship, $E = \Delta mc^2$, where c denotes the speed of light. The fusion of elements up to iron (atomic mass 56) results in the release of energy. The fusion of elements heavier than iron requires an input of energy. Thermonuclear fusion provides the energy that powers stars. In main sequence stars, such as the Sun, fusion reactions (the proton–proton reaction or the carbon–nitrogen–oxygen cycle) convert hydrogen to helium. In red giants, helium is converted to carbon by a process called the triple-alpha reaction and, in highly evolved high-mass stars, fusion reactions synthesize a succession of elements up to iron.

See also: atom, element (chemical), energy, fission, mass, nucleosynthesis, special relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gagarin, Yuri Alekseyevich (1934–1968)

Russian cosmonaut, born in Gzhatsk, Russia (renamed Gagarin after its favourite son). Trained in the Soviet air force, in 1961 he became the first man to travel in space, completing an orbit around the Earth in the Vostok spacecraft.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Astronomy *c.* 1900–*c.* 1950

By the early years of the 20th century it was widely believed that the great majority of the objects in our stellar system were close to the Galaxy's center, that the Sun was near its center, that the stars moved randomly in space and that the Galaxy included all the objects in the visible universe. There was, however, no general agreement on the Galactic system's size and detailed shape. Also, while some astronomers were prepared to accept in principle that external galaxies existed, they placed these other galaxies beyond the reach of even the most powerful telescopes.

Half a century later a very different conception of the Galaxy had been fashioned. In the early 1950s the orthodox view was that the Galaxy had a diameter of 30 000 pc and was a disk-like system with a central bulge. It also possessed a spiral structure with the Sun placed about 8500 pc from the center. All astronomers now accepted that the Galaxy rotated differentially and that it was only one of a multitude of galaxies.

Galactic models in the early years of the century

There were two principal models of the Galaxy in the first decade of the 20th century: the spiral and the ellipsoidal. The hypothesis that the stars in the Galaxy are arranged in a spiral pattern, which had first been proposed, in the middle of the nineteenth century, owed most to the Dutch astronomer Cornelis Easton (1864–1929). In 1900 he had described 'A new theory of the Milky Way' in which he had noted that the latest observations of nebulae had demonstrated the spiral to be a much commoner form than had previously been supposed. Might the spiral structure, Easton asked, be the plan on which the Galaxy was designed?

A principal aim of some astronomers in the 19th century had been to discover the actual distribution of the stars in space and hence the size and structure of the Galaxy. To succeed they judged they would have to disentangle the two causes of a star's apparent brightness: its distance and its luminosity. For want of anything better, these pioneers had sometimes been driven to employ the assumptions of constant absolute magnitudes and constant spatial densities for the stars, but both were by 1900 long accepted as grossly inadequate. At the end of the 19th century and in the early years of the 20th, however, astronomers forged novel methodological and conceptual tools in their attempts to discern the nature of the Galaxy. These tools gave rise to 'statistical astronomy' which many hoped would eventually enable them to discover the true arrangement of the stars. As the astronomer–historian A Pannekoek expressed it, 'here [is] a new kind of astronomy, which may be called "statistical astronomy". It appears where we have to deal not with stars individually but, with hundreds or thousands or even millions of them. Then the nature of the problems has changed with the object; we do not ask which stars, but how many stars have certain characteristics (color, spectrum, duplicity) or certain values of the parameters (temperature, density, luminosity, magnitude). Counting supplies the measuring. The positions (in the sky or in space) do not matter, but the densities of distribution (over the sky or over space). Statistical laws of distribution are the objects and the working instruments of the astronomer who is dealing with thousands and millions of the heavenly host'.

In the early years of the 20th century the three most notable exponents of, and innovators in, statistical astronomy were C V L CHARLIER (1862–1934) at Lund, Hugo von Seeliger (1849–1924) at Munich and J C KAPTEYN (1851–1922) at Groningen. Seeliger's approach consisted basically of counting the total number of stars between successive magnitude limits. These counts and the mathematical analyses of the counts enabled him to estimate (1) the flattening of the stellar system toward the plane of the Milky Way, (2) the relative extent of the Galaxy and (3) the general distribution of stars as a function of magnitude and galactic latitude. Further, for his 'Betrachtungen über die räumliche Vertheilung der Fixsterne' of 1898, Seeliger constructed an ellipsoidal model of the Galaxy which had a diameter of about 3000 pc along the galactic plane and stretched for about 1300 pc at right-angles to the plane. However, to secure detailed information on the actual spatial density an astronomer needed information on the variations in the intrinsic brightness of stars and, in particular, the luminosity function—that is, the number of stars per unit volume of space in successive intervals of absolute magnitude.

At the start of the 20th century so few trigonometric parallaxes of stars were known that an attempt to determine the luminosity law directly was out of the question. Kapteyn's response to the paucity of parallax data was to develop various mathematical techniques to exploit the more easily obtained observations of magnitude and proper motion; of these the most valuable involved an empirical expression for the average parallax of a group of stars as a function of their magnitudes and proper motions. With the aid of these techniques, Kapteyn approximated the luminosity law and then calibrated the dependence on distance of the space density of the stars.

An early product of Kapteyn's scrutiny of the proper motions of the stars was his discovery, announced in 1904, that the stars tended to move in two distinct directions. Kapteyn supposed this phenomenon to be the result of two intermingled streams of stars moving relative to one another. While an explanation of Kapteyn's find became one of the outstanding problems of galactic astronomy, the immediate and surprising message of the star streams was that the stars do not move randomly, as had been assumed, but that there is an order amidst the apparent jumble of stellar proper motions.

Kapteyn was also anxious to secure within a reasonable time representative samples of the immense amount of data that he believed necessary to a proper study of the Galaxy. He therefore proposed in 1906 that astronomers should obtain as much information as they

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

could about 206 selected stellar areas uniformly spread over the sky. This plan was soon taken up by a number of observatories.

Despite their hopes of achieving a greater understanding of the structure of the Galaxy, statistical astronomers labored under a major handicap: the lack of firm knowledge about interstellar absorption. The possibility of the absorption of star light made it difficult to decide whether the observed thinning out of the stars with increasing distance was genuine or whether it was the result of absorption. Statistical astronomers were particularly uneasy because the observed change in density seemed to place the Sun in a nearly central, and apparently privileged, position in the Galaxy.

In the late 19th century E E BARNARD (1857–1923), a renowned American observational astronomer, had noticed dark regions in his photographs of the Milky Way. Such regions had been known since the time of William Herschel and astronomers had pondered whether they were areas genuinely devoid of stars or whether dark obscuring clouds were blocking the light of the stars that lay beyond. Until the 1910s many astronomers accepted that the dark regions were indeed holes or rifts, but Barnard's later photographs of the Milky Way were largely instrumental in convincing his colleagues that they were really clouds of obscuring material situated among the stars.

By about 1916, however, astronomical opinion had moved toward the view that while localized obscuring clouds did exist, the general absorption of star light in space is insignificant. This shift was largely due to the examination by HARLOW SHAPLEY, then a young astronomer at Mount Wilson Observatory in California, of the color of the stars in globular clusters. Shapley had examined in detail the globular cluster M13 in Hercules. In 1915, he had estimated this cluster to be about 30 000 pc away, a much larger distance than the contemporary estimates of the diameter of the Galaxy. If absorbing material had been present it would, Shapley asserted, have manifested itself by reddening the light from the stars of the Hercules cluster: the color indices of the stars would then have been more positive than for a sample of similar stars in the solar neighborhood. Yet this was not the case, and so Shapley claimed that the extinction and reddening of star light could effectively be ignored in researches of the Galaxy. These studies freed Kapteyn of any reservations he had over assigning a nearly central position to the Sun in the Galaxy.

Shapley's model of the Galaxy

In his celebrated *Stellar Movements and the Structure of the Universe* of 1914, A S EDDINGTON (1882–1944) summarized the picture that modern researches had painted of the stellar system: 'It is believed that the great mass of the stars. . .are arranged in the form of a lens- or bun-shaped system. That is to say, the system is considerably flattened towards one plane. . .. In this aggregation the Sun occupies a fairly central position. . .. The median plane of the lens is the same as the plane marked out in the sky by the Milky Way, so that when we look in any direction along the galactic plane (as the plane of the Milky Way is called), we are looking towards the perimeter of the lens where the boundary in most remote. . . . The thickness of the system, though enormous compared with ordinary units, is not immeasurably great. No definite distance can be specified, because it is unlikely that there is a sharp boundary; there is only a gradual thinning out of stars. The facts would perhaps be best expressed by saying that the surfaces of equal density resemble oblate spheroids. To give a general idea of the scale of the system, it may be stated that in directions towards the galactic poles the density continues practically uniform up to a distance of about 100 parsecs; after that the falling off becomes noticeable so that at 300 parsecs it is only a fraction (perhaps a fifth) of the density near the Sun. The extension in the galactic plane is at least three times greater. These figures are subject to large uncertainties'.

Just a few years later, Shapley claimed that the Galaxy has a diameter of about 100 000 pc, a staggering increase on the commonly accepted size. Shapley, moreover, placed the Sun in an eccentric position thousands of parsecs from the Galactic center.

All of the main components involved in Shapley's invention of the big Galaxy were present by late 1917: his concern for the peculiar distribution of the globular clusters (they seemed to be almost all located in one part of the sky), a conviction that the Galaxy was much larger than his contemporaries conceded and a knowledge of the distances of the globular clusters. By early 1918, Shapley had decided that the globular clusters outlined the stellar system, and from their distances he had calculated that the Galaxy is about 100 000 pc in diameter and that the Sun is 20 000 pc from its center. Further, the globular clusters are symmetrically distributed throughout space on either side of the Galactic plane, and the center of the system of globular clusters defines the center of the Galactic system itself. The local group, which contains all of the stars ordinarily studied, is a few times larger than a globular cluster.

Not all of the elements of this radical model were new, however. Shapley's idea of the eccentric positioning of the Sun within the Galaxy was not the novelty he liked to suppose. The idea has an ancestry going back to the 18th century and only 9 yr earlier, in 1909, Karl Bohlin, a noted Swedish astronomer, had even tied it to the proposal that the globular clusters are symmetrically placed about the center of the Galactic system. Bohlin's hypothesis provoked little discussion, perhaps because of the strange version of the island universe theory that he had linked to his Galactic theory. In fact, in 1915 Shapley had explicitly rejected Bohlin's scheme because the distances Shapley had derived to the globular clusters, together with his estimates of the dimensions of the Galaxy, at that time clearly ruled out a model in which the clusters surrounded the center of the Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shapley's model was nevertheless unexpected. Astronomers had no inclination to demolish the contemporary notions of the structure and size of the Galaxy. Shapley was also arriving at his surprising conclusions on the basis of a new methodology. Instead of advancing meticulously outwards from the region of the solar neighborhood, as the statistical astronomers did, Shapley went to the other extreme of exploiting the remote globular clusters to advance into the Galaxy. Also, the two approaches did not seem to link up, and to many astronomers it seemed inherently unreasonable that the refined and familiar tools of statistical astronomy would give so bad a value for the dimensions of the Galaxy and such a poor answer for the location of the Sun.

In the early 1920s, Kapteyn wrote two papers that capped his life's work in Galactic astronomy and offered a vision of the Galaxy again very different from Shapley's. The first was composed in 1920 in collaboration with P J van Rhijn. The authors argued that the Sun is close to the center of the Galaxy, and that the limits of the Galactic system, which they represented as a flattened ellipsoid, were reached at a distance of about 1250 pc at right-angles to the Galactic plane and at 9000 pc along the plane. Two years later Kapteyn wrote what he called a 'First attempt at a theory of the arrangement and motion of the sidereal system'. He reckoned that the limits of the Galactic system, which he took to be where the star density sank to 1/100th of the density in the solar neighborhood, were found at roughly 1700 pc at right-angles to the plane and 8500 pc along the Galactic plane. Kapteyn thereby proposed dimensions for the Galaxy that, although small compared with Shapley's, were far larger than the estimates of the mid-1910s, and so the 'Kapteyn universe', as this model became known, was itself a notable departure from the previously prevailing orthodoxy.

Many astronomers reckoned there were weaknesses and flaws in Shapley's model, especially in his scheme of distances. These he obtained with the aid of three interlocking methods. Some of the larger, and apparently closer, clusters contained stars that Shapley identified as Cepheid variables. Yet Shapley's calibration of the period–luminosity relationship relied on a statistical analysis of the proper motions of only 11 stars. Furthermore, despite the advance in accuracy achieved in L Boss's catalogue of proper motions, the proper motions of these eleven Cepheids were still very uncertain.

To reach a cluster that contained no visible Cepheids, Shapley examined its 30 most luminous stars. He then discarded the five brightest in case they were field stars and not physically associated with the cluster. For the remaining 25 he calculated their mean luminosity. Shapley now derived the absolute mean magnitude for these 25 stars by following the same procedure for the clusters of known distance (that is, those containing visible Cepheids). He could now find the distances not only to those clusters in which he could detect Cepheids but to any in which he could distinguish the 30 brightest stars. For the very distant globular clusters that would not yield to either of these approaches, he employed their apparent angular diameters as distance indicators. To calibrate this final link in his chain of distance determinations he drew an 'apparent diameter versus distance' curve for the globular clusters calibrated by clusters that had submitted to the previous techniques.

It was to the initial Cepheid calibration that astronomers objected most strongly. If Shapley had erred at this point, then the claimed distances for the globular clusters, on which rested his theory of the Galaxy, would be undermined. To many astronomers it seemed imprudent to base a revolutionary conception of the Galaxy on a statistical analysis of a mere 11 stars. One of these astronomers was H D CURTIS of Lick Observatory in California.

In 1920 Curtis and Shapley discussed 'The scale of the Universe' at a meeting of the National Academy of Sciences in Washington. This encounter—the so-called 'Great Debate'—led in the following year to the publication of papers in a *Bulletin of the National Research Council* by Shapley and Curtis under the same title as their Washington addresses. The papers, however, were pitched at a much higher level than the Washington talks (Shapley, in fact, was most anxious in the 'debate' to impress those members of the audience concerned with selecting the next director of the Harvard College Observatory, a position he coveted), and a lengthy section of Curtis's 1921 paper offered a critique of Shapley's use of the Cepheids as distance indicators. In Curtis's opinion the proper motion data on the Cepheids were meager, and he was far from convinced that a period–luminosity relationship existed for the Cepheids.

Shapley appreciated that his calibration of the period–luminosity relationship was open to criticism and so, to counter Curtis's attack on the Cepheids, he employed in his own paper the B stars as the principal distance indicators to the globular clusters. But Curtis, like many other astronomers, rejected Shapley's claim that the B stars were reliable 'standard candles'.

It would, however, be a mistake to believe that around 1920 astronomers were faced with a stark choice between the Galactic models of Kapteyn and Shapley. By no means every astronomer was prepared to side with either Shapley or Kapteyn. Certainly the spiral model had numerous adherents. One result, then, was that despite Shapley's vigorous attempts to defend his model of the Galaxy, in the early 1920s it probably had more opponents than advocates.

Support for Shapley's models: Cepheids and Galactic rotation

In 1923, R E Wilson, an established American astronomer, re-calibrated the period–luminosity relationship for the Cepheids. Although Wilson concluded that Shapley's calibration probably needed altering by an amount that would decrease the distances derived from the Cepheids by a factor of around 1.3, his answer was sufficiently close to Shapley's that Shapley's model now gained in credibility.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In the next year, EDWIN HUBBLE, using the 100 in telescope at Mount Wilson, detected Cepheid variables in a number of spiral nebulae and soon demonstrated to the satisfaction of nearly all astronomers the extragalactic nature of the spirals (see EXTRAGALACTIC ASTRONOMY *c.* 1900–1950). In so doing, Hubble further underlined the reliability of the Cepheids as distance indicators and thereby made Shapley's model of the Galaxy still more plausible.

In 1927, more evidence came from an unexpected source for the two central features of Shapley's model of the Galaxy, the role of the globular clusters in outlining the Galaxy and the eccentric position of the Sun. This source was the rotation of the Galaxy. That the Galactic system rotated had long been suspected; indeed, its flattened form seemed to be the natural consequence of rotation. The spectroscopic measures in the 1910s and 1920s of the rotation of spiral nebulae had further strengthened the belief among astronomers that the Galaxy rotates.

In 1926, J H OORT (1900–92), a young Dutch astronomer based at Leiden Observatory, contended that our local system of stars is rotating around the center of the Galaxy. By so doing, he hoped to explain the well-known, but enigmatic, asymmetry of the motions of the high-velocity stars and globular clusters. The local system of stars, Oort argued, is moving more quickly than the high-velocity stars which are really a subsystem of stars rotating relatively slowly. In consequence, as the local system moves 'through' the high-velocity stars, these appear to drift asymmetrically with respect to the local system.

The following year Oort announced that, as a result of his attempt to verify directly Lindblad's theory of Galactic dynamics by an analysis of the motions of the globular clusters and high-velocity stars, he had secured firm evidence of a differential rotation of the Galaxy. In 1924 BERTIL LINDBLAD (1895–1965), through his attempts to interpret the phenomenon of star streaming, had considered a possible rotation of the Galaxy. By 1925 he had decided that the motions of the constituents of the Galaxy were explicable on the hypothesis that the Galaxy is divided into a series of subsystems, each of which has rotational symmetry about a common rotational axis. Each subsystem possesses a different speed of rotation, and hence a different degree of flattening, and he stressed that, while the high-velocity stars do not belong to the same dynamical system as those of low velocity, they must be related to the rest of the Galaxy since their motions are symmetrical with respect to the Galactic plane. Lindblad, as Oort was soon to do, calculated the dynamical center of the Galaxy to be very close in Galactic longitude to the center of the globular cluster system as defined by Shapley. However, there was at first a basic difference between Oort's theory and that of Lindblad. As Oort recalled in 1971: 'Lindblad had tentatively supposed that the principal part of the greater galactic system is formed by an ellipsoid of constant density, and that as a consequence its rotation is approximately the same as that of a solid body. The systematic effects I found in 1927 indicated that the actual motions were quite different, and that the angular velocity increased strongly toward the center of the galaxy, indicating a large concentration of mass in the central area'. The detection of Galactic rotation had been very much 'in the air' and Oort had presented what most astronomers saw as its observational verification. Support for Oort's arguments, moreover, soon came from other investigators.

The 'discovery' of Galactic rotation, astronomers quickly agreed, gave support to Shapley's claims that the center of the Galaxy is defined by the system of globular clusters and that the Sun is eccentrically placed. There was, nonetheless, one point where the conclusions of Oort and Lindblad conflicted with Shapley's model. Oort had reckoned that the distance to the dynamical center of the Galaxy was about 6000 pc, roughly one-third of Shapley's estimate. Lindblad also thought that Shapley's distance to the Galactic center needed revision, and he suggested that it be reduced by a factor of 0.6. These estimates were, nevertheless, soon to be modified and brought into close agreement.

Trumpler's investigations of open clusters

The impression that the effects of a general interstellar absorption within the Galaxy were insignificant was rudely disturbed in 1930 by the publication of the results of R J TRUMPLER'S study of the open Galactic clusters. Trumpler (1886–1956), an astronomer at the Lick Observatory, had assumed that the clusters of similar form have the same dimensions; by comparing the observed angular diameter of a cluster with the assumed average linear diameter of the subclass to which the cluster belonged, he derived a value for the cluster's distance. Trumpler had also examined the magnitude and spectral types of the stars within the clusters. Then, by constructing a Hertzsprung–Russell diagram for each cluster and comparing the observed diagram with a standard diagram, Trumpler secured another value for the cluster's distance. He thereby calculated that the two distance indicators to the clusters gave systematically different answers: the more distant the cluster, the more the two distance values differed.

The reason for this deviation, astronomers soon agreed, was that interstellar absorption diminished the brightness of the cluster stars but did not affect appreciably the measured diameters of the clusters. The effect of absorption, therefore, was to make the distance to a cluster obtained from its Hertzsprung–Russell diagram appear greater than it actually was. Although a number of earlier investigations can, with the benefit of hindsight, be seen to have pointed towards the existence of a general interstellar absorption, it was Trumpler's analysis of the Galactic clusters—probably because it was more extensive and complete than previous researches—that convinced astronomers that absorbing matter is present throughout the Galaxy, not just in isolated clumps.

The difference between Oort's distance to the dynamical center of the Galaxy and Shapley's estimate of the distance to the center of the system of globular clusters could now be simply explained: in his calculation of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

size of the Galactic system Shapley had not allowed for the dimming effect of a general interstellar absorption, and as a result he had overestimated the distance to the Galaxy's center.

The vindication of the spiral hypothesis

By the mid-1930s, then, it was widely believed that the Sun is about 10 000 pc from the galactic center, but there was still no consensus on the overall size or structure of the Galaxy. J S PLASKETT, director of the Dominion Astrophysical Observatory in Canada, argued in 1935 that all astronomers would probably agree that the Galaxy's diameter is no greater than 40 000 pc, although some would limit it to 30 000 pc. In 1937 B J BOK, of Harvard Observatory and one of the leading students of the Galaxy, warned astronomers that the only well-established fact about the Galaxy's structure was the eccentric location of the Sun. Bok did, however, propose a detailed working model of the Galaxy in which the Galactic center lay in the direction of Sagittarius and in which a spiral arm passes from Carina through the Sun towards Cygnus. Shapley's model of the late 1910s, Bok asserted, had proved valuable for two decades, but now it needed replacing.

The suggestion that the Galaxy has a spiral structure was, as Bok acknowledged, hardly novel. However, in the late 1920s it had gained in plausibility after Hubble had demonstrated the spiral nebulae to be external galaxies. It was not, however, until the mid-1950s that astronomers judged there was conclusive evidence of the Galaxy's spiral structure. The new, and decisive, evidence did not come from the traditional statistical astronomy. Rather, it followed a shift in the way astronomers probed the Galaxy's structure with the emphasis now on the scrutiny of particular sorts of Galactic objects, as well as comparisons with external galaxies. Because of the Sun's eccentric position within the plane of the Milky Way, astronomers accepted they were surrounded by a confusing complexity of star and dust clouds, but by investigating the galaxies—star systems which seemed to contain the same kinds of objects as the Galaxy—astronomers judged they were better able to gain insights into the design of the Galaxy itself.

In his studies in the 1940s with the 100 in telescope at Mount Wilson, WALTER BAADE (1893–1960) had noticed that the bright blue stars lay along the spiral arms of the Andromeda galaxy. Shortly afterwards, in 1951, a group of astronomers led by W W Morgan investigated the distances of the bright blue stars in our Galaxy and found that they too seemed to trace out spiral arms. In 1952 Bok listed nine 'tracers' for discerning the spiral arms. One of these tracers exploited a new tool for the examination of the remote regions of the Galaxy: the 21 cm radiation.

For much of the Second World War the Netherlands had been under German occupation, but copies of the *Astrophysical Journal* still reached the Observatory at Leiden. Two papers by GROTE REBER, one of the pioneers of radio astronomy, appeared in the *Journal* in 1940 and 1944, and the second paper included a crude radio map of the Galaxy. Oort, the Director at Leiden, was prompted to suggest that if spectral lines could be observed in the radio region it would enable the kinematics and dynamics of the whole Galactic system to be analyzed since the radio region is unaffected by obscuring dust.

In 1944 a celebrated seminar was held at Leiden to discuss Reber's papers. At this seminar H C VAN DE HULST, following Oort's proposals, predicted that atomic hydrogen in interstellar space might emit an observable radio line at a wavelength of 21 cm. It was, however, to be 6 yr before the line was detected and strong evidence secured for what had been a theoretical possibility.

Although some astronomers were slow to appreciate the opportunities offered by the 21 cm radiation, its advocates claimed that it would help them investigate some fundamental problems concerning the Galaxy: the Galaxy's spiral structure, the distribution of mass and velocity in the Galaxy, the distribution of neutral hydrogen in the plan of the Galaxy and the presence of neutral hydrogen associated with the large complexes of obscuring dust and gas clouds. For its proponents, the great advantage of the 21 cm radiation was that it could penetrate clouds of obscuring matter without being diminished in intensity. Indeed, the elucidation of the Galaxy's fine structure would lie mainly with the infant science of radio astronomy.

Stellar populations

In 1921 Shapley and Curtis had spoken for most astronomers when they agreed that stars 'in clusters and in distant parts of the Milky Way are not peculiar—that is, uniformity of conditions and of stellar phenomena naturally prevails throughout the galactic system'. The view that each region of the Galaxy contained the same sorts of stars as all other regions was finally to be undermined in the 1940s by Baade.

Until 1943 all attempts to resolve the nucleus of the Andromeda galaxy into stars had, by general consent, failed. However, in 1943 Walter Baade equipped the 100 in Mount Wilson telescope with photographic plates especially sensitive in the red region of the spectrum. He thereby succeeded in resolving the nucleus into a swarm of faint red stars. Baade further concluded that the brightest stars in the nucleus were about 100 times fainter than the blue stars in the spiral arms of the Andromeda galaxy. These red stars, moreover, were similar in type and luminosity to the brightest stars found in globular clusters.

Astronomers had accepted since the 1910s that the outer regions of a number of spiral nebulae were much bluer than their nuclei. After his resolution of the nucleus of the Andromeda galaxy, Baade introduced the concept of 'stellar populations' to explain this apparent anomaly and to help interpret the structure of galaxies. Different populations of objects, he argued, have different spatial distributions. What he termed 'population I' objects were associated with the spiral arms of galaxies and included blue supergiant stars and galactic clusters. Baade linked

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

'population II' with elliptical galaxies, globular clusters and the nucleus of spiral galaxies.

The concept of populations also led to a reinterpretation of the motions of the Galaxy's stars. Even before the term 'stellar population' had been coined, astronomers had noted a correlation between the astrophysical characteristics of stars and their kinematical behavior. When plotted on a Hertzsprung–Russell diagram the high-velocity stars produced an array similar to that of a globular cluster, but the equivalent diagram for the low-velocity stars resembled that for a galactic cluster. Now astronomers argued that the different diagrams could be explained in terms of populations: the low-velocity stars are population I objects while the high-velocity stars belong to population II.

In 1952 Baade announced another surprising discovery. There was, he asserted, an error in the calibration of the period–luminosity relationship. The estimated distances to the external galaxies, which were based on this relationship, had therefore to be revised. Indeed, the distances had to be approximately doubled (see also EXTRAGALACTIC ASTRONOMY *c.* 1900–1950). The re-calibration of the period–luminosity relationship and the resulting doubling of the distances to the galaxies also meant that the estimates of the diameters of the galaxies had to be doubled: at last the lingering unease that the Galaxy might be truly exceptional, a continent amidst the islands of external galaxies, was dispelled.

Bibliography

Holmberg G 1999 *Reaching for the Stars: Studies in the History of Swedish Stellar and Nebular Astronomy, 1860–1940* (Lund: Lund University)

Hoskin M 1982 *Stellar Astronomy: Historical Studies* (Chalfont St Giles: Science History Publications)

Paul E R 1993 *The Milky Way Galaxy and Statistical Cosmology 1890–1924* (New York: Cambridge University Press)

Smith R W 1982 *The Expanding Universe: Astronomy's 'Great Debate' 1900–1931* (Cambridge: Cambridge University Press)

Robert W Smith

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Bulge

Bulges are the central visible spheroidal systems in DISK GALAXIES. The prominent bulges of early-type galaxies are exemplified by that of the SOMBRERO GALAXY, NGC 4594, shown in figure 1. The bulge of our own galaxy, which is of later type, is dominated by the disk, as is seen in figure 2.

A bulge is essentially a STELLAR POPULATIONS concept, the key properties being the age, kinematics and chemical composition of the stars of the bulge. The kinematics of the stars yield a certain structure, which is the nominal defining property of a galactic bulge. The general properties of bulges are well reviewed by Wyse *et al* (1997). The history of the study of the galactic bulge is recounted by Frogel (1988). Here we deal in turn with the age, kinematics, structure and chemistry of the bulge of the MILKY WAY.

Age

Our preconceptions about the stellar population of the galactic bulge are that it is old, like the halo, and that, like elliptical galaxies, it has high density and must have formed early in the galaxy's life, and perhaps in terms of cosmic time as well. But it is important to look at the *data* on the age of the bulge population.

Holtzman *et al* (1993, 1998) used the wide field planetary cameras of the Hubble Space Telescope to image main sequence stars in Baade's window (a relatively unobscured line of sight into the bulge at impact parameter 500 pc). They found a population of mixed age characteristics with stars younger than the Sun relatively deficient compared with the solar neighbourhood. Ortolani *et al* (1995) studied the same field and concluded that the age and other characteristics were similar to that of metal-rich globular clusters. Frogel *et al* (1999) have mapped the inner bulge in the infrared, and conclude that there is a young component of the stellar population observed near the galactic centre, but that it declines in density much more quickly than the overall bulge population, and is undetectable beyond 1° from the galactic centre.

While the main sequence provides a democratic census of stellar ages, VARIABLE STARS offer good tracers of stellar populations of certain ages. Oort and Plaut (1975) identified the RR LYRAE STARS in a number of similar windows. These are clearly old stars as a rule, the youngest simple stellar population with RR Lyrae stars being NGC 121 in the Small Magellanic Cloud, whose age is 11.5 Gyr according to Stryker *et al* (1985). Menzies and Whitelock (1985) found that the long-period variables in the bulge are of relatively short period ($\sim$200 days) by the standards of the galactic disk with bolometric luminosity approximately -4 mag. A similar luminosity was obtained for OH/IR stars (the terminal phase of asymptotic giant branch evolution) by Wood *et al* (1998). This also suggests a population from the age interval represented by NGC 121 to 47 Tucanae, a metal-rich Galactic globular cluster.

We conclude that the stellar population of the bulge of the galaxy is indeed predominantly old. We may be witnessing bulge formation in real (or at least 'lookback') time in the very high star formation rates ($10M_\odot$/year) at redshifts greater than 3 in Lyman dropout galaxies (Steidel *et al* 1996).

Kinematics

Our preconception of bulges from the stellar populations paradigm of Baade is that they are scaled down elliptical galaxies. The data, however, are more specific. Davies *et al* (1983) found that at every luminosity both ellipticals and bulges follow the isotropic rotator model. This model, described succinctly by Binney and Tremaine (1987), assumes an axisymmetric system that rotates about its symmetry axis. The model is flattened by rotation to an ellipticity ε, which is uniquely (in a model in which the orbits of the stars are fully isotropic) a function of the ratio of the rotation velocity and the central velocity dispersion.

Structurally, the bulge was traditionally difficult to get a quantitative fix on, although de Vaucouleurs and Pence (1978) were able to determine an effective radius, $r_e = 2.5$ kpc. The effective radius is the radius which contains half the light. Morrison *et al* (1990) found that this *over*predicted the halo density in the solar neighbourhood. Terndrup (1988) found that standard spheroid models (the Bahcall and Soneira (1980) model, in particular) *under*predicted the number of giant stars in the bulge. The K and M giants in the bulge fit a de Vaucouleurs law with $r_e = 0.25$ kpc, a full order of magnitude less than de Vaucouleurs and Pence found. Similar compactness was found for OH/IR stars by Habing (1986). It really seems that the bulge is a separate component from POPULATION II near the Sun. This would not be surprising if the bulge and halo were formed in different ways, for example the bulge by dissipative collapse at an early epoch and the halo by accretion subsequently (see HALO, GALACTIC).

The infrared has proved the answer to determining structural parameters of the bulge. Binney *et al* (1997) find axis ratios of (1:0.6:0.4). Zhao *et al* (1994) show that the bulge is really structurally a bar. This is not consistent with the isotropic rotator model, and one is left wondering what fraction of bulges are bars (as the Milky Way's apparently is), and what fraction resemble ELLIPTICAL GALAXIES.

Chemistry

Our preconception from spectroscopy is that the bulge is a super metal-rich stellar population (Whitford 1978), i.e. the metallicity is higher than the Sun's. The data, however, show a broad distribution (Minniti *et al* 1995), from -1 to 0.5 in [Fe/H] in Baade's window (McWilliam and Rich 1994), i.e. from a tenth to three times solar abundance. Rich (1990) showed that this distribution fits the simple model of chemical enrichment (Pagel and Patchett 1975), where the bulge behaves like a closed system, in which gas steadily, without our specifying the timescale, and eventually completely, turns into stars, never being replenished at any time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

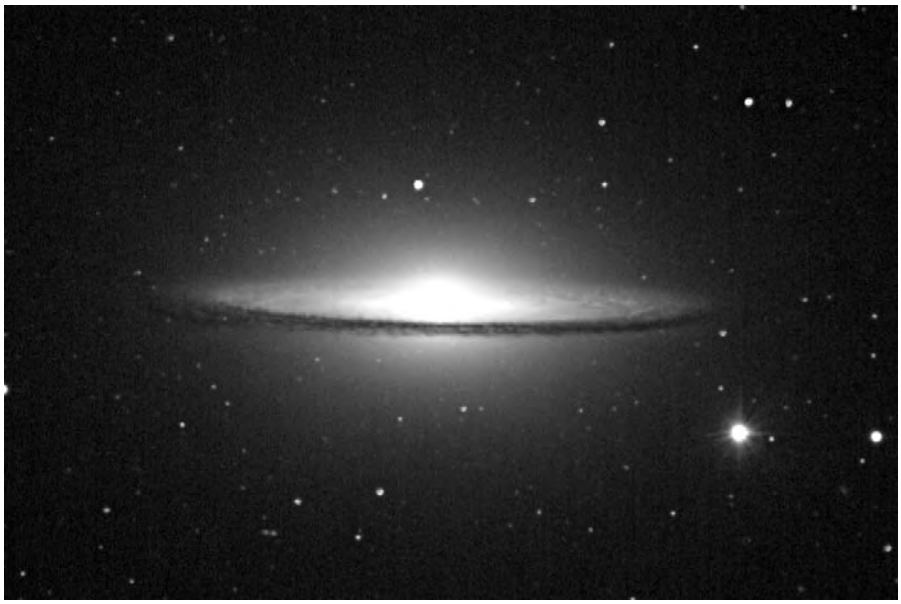


Figure 1. The Sombrero galaxy, NGC 4594. This is an Sa galaxy with a prominent bulge. (Picture from Siding Spring Observatory by courtesy of Michael Bessell.)

Figure 2. The bulge of our own galaxy is much less prominent relative to the disk. The Milky Way's generally attributed Hubble type (Sbc) reflects this. (Picture from Siding Spring Observatory by courtesy of Michael Bessell.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wood *et al* (1998) found that young stars in the bulge are between two and four times SOLAR ABUNDANCE. Jablonka *et al* (1996) studied a sample of bulges, finding a direct relationship between metallicity and luminosity and noting an enhancement of the element ratio Mg/Fe relative to the solar ratio. In principle, this means rapid initial enrichment in the bulge, relative to a more leisurely enrichment history in the disk's interstellar medium, dominated by iron producing type Ia supernovae (Wheeler *et al* 1989, Gilmore *et al* 1989).

Star clusters

A separate population of bulge GLOBULAR CLUSTERS has been identified by Barbuy *et al* (1998). There is a prevalence of core collapsed clusters among them. Core collapsed clusters have power-law radial density distributions, as opposed to the classical (King model) flattened cores. Rich *et al* (1999) note the tendency of these clusters to blue horizontal branches and depletion of red giants, a phenomenon which may result from the high frequency of stellar collisions in the dense environments of the cluster centres.

Formation

A very plausible model of the formation of the galactic bulge has the majority of the stars form through the dissipative and violent collapse of protogalactic gas clouds in the first billion years of cosmic history. There is clear evidence, however, that the bulge has added to its stock of stars since that time. A number of processes, such as mergers of dwarf galaxies and disk instabilities, could be responsible for that extended population.

Bibliography

Barbuy B, Bica E and Ortolani S 1998 *Astron. Astrophys.* **333** 117
Bahcall J and Soneira R 1980 *Astrophys. J.* (suppl) **44** 73
Binney J, Gerhard O and Spergel D 1997 *Mon. Not. R. Astron. Soc.* **288** 365
Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)
Davies R, Efstathiou G, Fall S, Illingworth G and Schechter P 1983 *Astrophys. J.* **266** 41
de Vaucouleurs G and Pence W 1978 *Astrophys. J.* **83** 1163
Frogel J 1988 *Ann. Rev. Astron. Astrophys.* **26** 51
Frogel J, Tiede G and Kuchinski L 1999 *Astrophys. J.* **117** 2296
Gilmore G, Wyse R and Kuijken K 1989 *Ann. Rev. Astron. Astrophys.* **27** 555
Habing H *et al* 1985 *Astron. Astrophys.* **152** L1
Holtzman J *et al* 1993 *Astrophys. J.* **106** 1826
Holtzman J *et al* 1998 *Astrophys. J.* **115** 1946
Jablonka P, Martin P and Arimoto N 1996 *Astrophys. J.* **112** 1415
McWilliam A and Rich R 1994 *Astrophys. J.* (suppl) **91** 749
Menzies J and Whitelock P 1985 *Mon. Not. R. Astron. Soc.* **212** 783
Minniti D *et al* 1995 *Mon. Not. R. Astron. Soc.* **277** 1293
Morrison H, Freeman K and Flynn C 1990 *Astrophys. J.* **100** 1191
Oort J and Plaut L 1975 *Astron. Astrophys.* **41** 71
Ortolani S *et al* 1995 *Nature* **377** 701
Pagel B and Patchett B 1975 *Mon. Not. R. Astron. Soc.* **172** 13
Rich R 1990 *Astrophys. J.* **362** 604
Rich R *et al* 1998 *Astrophys. J.* **116** 1295
Steidel C *et al* 1996 *Astron. Astrophys. Lett.* **462** L17
Stryker L *et al* 1985 *Astrophys. J.* **298** 544
Terndrup D 1988 *Astrophys. J.* **96** 884
Wheeler C, Sneden C and Truran J 1989 *Ann. Rev. Astron. Astrophys.* **27** 279
Whitford A 1978 *Astrophys. J.* **226** 777
Wood P, Habing H and McGregor P 1998 *Astron. Astrophys.* **336** 925
Wyse R, Gilmore G and Franx M 1996 *Ann. Rev. Astron. Astrophys.* **35** 637
Zhao H, Spergel D and Rich R 1994 *Astrophys. J.* **108** 2154

Jeremy Mould

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Latitude

The angular distance between the galactic equator and celestial body, measured perpendicular to the galactic equator, and taking values between 0 and 90°. It is denoted by b^{II} to signify that the value of galactic latitude relates to the galactic equator defined in 1959 by the International Astronomical Union by reference to the centerline of the Milky Way defined by radio observations of neutral hydrogen. The previous system is denoted b^{I}.

See also: galactic longitude, galactic plane, galactic year.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Longitude

The angular distance between the direction of the galactic center and the point on the galactic equator perpendicularly below (or above) a celestial body; it is measured in an anticlockwise direction, and takes values between 0 and 360°. Galactic longitude is denoted by l^{II} to signify that it is referred to the position of zero longitude agreed in 1959 by the International Astronomical Union (and coinciding with the radio source Sagittarius A which is presumed to be the galactic center). The old system (denoted l^{I}) was referred to the point of intersection of the galactic equator and the celestial equator, and was offset from the galactic center by some 30°.

See also: galactic latitude, galactic plane, galactic year.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Metal-Poor Halo

Most of the gas, stars and clusters in our Milky Way Galaxy are distributed in its rotating, metal-rich, gas-rich and flattened disk and in the more slowly rotating, metal-rich and gas-poor bulge. The Galaxy's halo is roughly spheroidal in shape, and extends, with decreasing density, out to distances comparable with those of the Magellanic Clouds and the dwarf spheroidal galaxies that have been collected around the Galaxy. Aside from its roughly spheroidal distribution, the most salient general properties of the halo are its low metallicity relative to the bulk of the Galaxy's stars, its lack of a gaseous counterpart, unlike the Galactic disk, and its great age. The kinematics of the stellar halo is closely coupled to the spheroidal distribution. Solar neighborhood disk stars move at a speed of about 220 km s^{-1} toward a point in the plane and 90° from the Galactic center. Stars belonging to the spheroidal halo do not share such ordered motion, and thus appear to have 'high velocities' relative to the Sun. Their orbital energies are often comparable with those of the disk stars but they are directed differently, often on orbits that have a smaller component of rotation or angular momentum. Following the original description by Baade in 1944, the disk stars are often called POPULATION I while the metal-poor halo stars belong to POPULATION II.

GLOBULAR CLUSTERS are the most readily identifiable component of the Galactic halo. Roughly 150 are known and most, but not all, occupy a spheroidal distribution, have radial velocities that are much larger than expected for disk-like rotation, and are metal poor. Figure 1 shows the distributions of the metal abundances, [Fe/H] ($\equiv \log[n(\mathrm{Fe})/n(\mathrm{H})]_\star - \log[n(\mathrm{Fe})/n(\mathrm{H})]_\odot$). Figure 2 shows the distribution of clusters' distances from the Galactic plane as a function of [Fe/H]. Two 'populations' of globular clusters exist: one belonging to the metal-weak halo, and a more metal-rich one that is a disk, possibly a 'thick disk' population (see GALACTIC THICK DISK). Field stars show very similar behaviors in metallicity and kinematics.

Globular cluster masses range from a few hundred $M_\odot$ (AM-4) to several million $M_\odot$ (ω Centauri = NGC 5139). Despite their large masses, globular clusters contribute only a small fraction (≈2%) to the total mass of the metal-poor stellar halo. Field stars dominate the mass. The total mass of the metal-poor stellar halo may be estimated using the density of such field stars in the solar neighborhood (≈0.2–0.5% of the disk stars) and the cumulative distribution of the mass in globular clusters as a function of Galactocentric distance. In field star and cluster stellar mass, the metal-poor halo amounts to only about 10^9 $M_\odot$, about 1% of the Galaxy's total stellar mass.

Despite its minor contribution to the Galaxy's stellar mass budget, the halo has been studied extensively to answer a wide variety of questions. In 1918, Shapley exploited the positions and approximate distances to globular clusters to identify the direction of and distance to the Galactic center, R_{GC}. Globular clusters and the halo population 'standard candle' RR LYRAE VARIABLES are still used to measure R_{GC}. The recognizability of globular clusters and RR Lyraes and their large range in R_{GC} provide good 'test particles' to map out the Galaxy's gravitational potential and discern the distribution of both the stars and the DARK MATTER. Most work on the halo population, however, arises from our desire to understand the Galaxy's formation process and its early history, and, in fact, to discern the chemical abundances produced by the BIG BANG since it appears that the first stars to be formed in our Galaxy lie in the Galactic halo. What were the abundances of those earliest stars? How did they change and how rapidly? How did the Galaxy form and evolve?

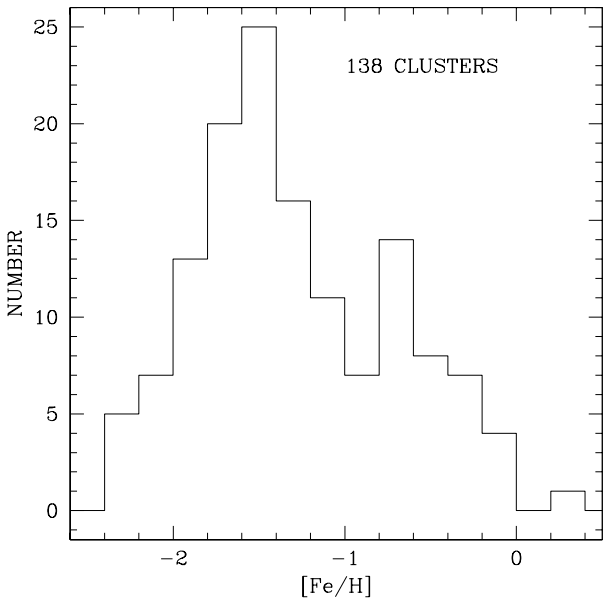


Figure 1. The distribution of [Fe/H] values for globular clusters.

Chemical abundances

Figure 1 employs [Fe/H] as the measure of 'metallicity'. For brighter field stars and the brightest stars in globular clusters, this may be determined by employing high-resolution, high-signal-to-noise spectroscopy to measure line strengths. In conjunction with atomic line parameters and an appropriate model stellar atmosphere, one may derive element-to-hydrogen abundance ratios such as [Fe/H]. The process is often iterative, involving tests to make certain that the derived abundances do not depend on the lower energy levels of the atomic transitions (to check the model's effective temperature), on the line strengths (to assess the effects of turbulent line broadening) and on gravity-sensitive and gravity-insensitive lines. The technique is often referred to as a 'fine analysis' or a 'curve of growth' analysis.

Because of long exposure times, such analyses of field and cluster stars are not always employed to estimate [Fe/H] values. Instead, metallicity 'indicators' are calibrated using results available from such analyses

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

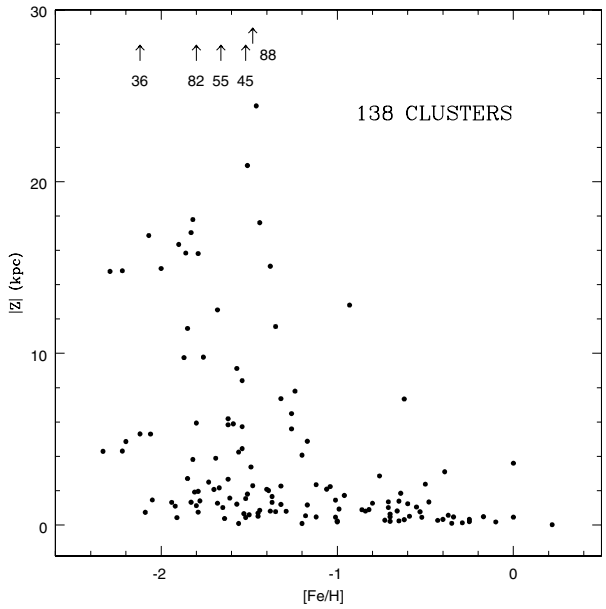


Figure 2. The distance from the plane of globular clusters as a function of [Fe/H].

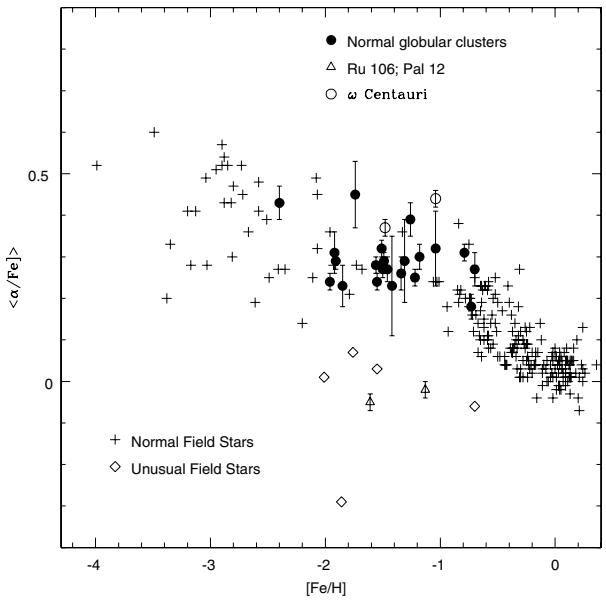


Figure 3. The behavior of Mg, Si and Ca abundances relative to iron as a function of [Fe/H] for field and cluster stars.

and then exploited to study larger and/or fainter samples of stars or clusters. Lower-dispersion spectra are easier and faster to obtain than high-dispersion spectra, and one technique that exploits this method is called ΔS since it employs a measured apparent difference in spectral type. In RR Lyrae variables, one measures the spectral type using the hydrogen lines and compares that with that defined from the 3933 Å line of Ca II. The difference is an indicator of the calcium abundance. A similar method, called W', exploits the three strong lines of Ca II at 8498, 8542 and 8662 Å and is used to estimate [Fe/H] (assuming a constant or at least monotonic relation between [Ca/H] and [Fe/H]) for red giant stars in globular clusters. One may also employ photometric methods. As [Fe/H] increases, both the predicted and the observed colors of stars along the red giant branch become redder, and thus the red giant branch color is a good metallicity indicator. Quantitatively, one measures the dereddened $B - V$ or $V - I$ color indices at the magnitude level of a cluster's horizontal branch. The resultant values, $(B-V)_{0,g}$ and $(V-I)_{0,g}$, are calibrated using [Fe/H] values available for a well-studied subsample of clusters.

High-resolution spectroscopy of metal-poor halo stars is valuable for measuring the abundances of elements other than iron to study the chemical evolution in the early Galaxy and, in fact, those produced by the Big Bang. Five results are especially noteworthy.

(1) Figure 1 shows two roughly Gaussian distributions of [Fe/H]. Simple models may explain such behavior. In brief, a gaseous system with an initial metallicity, $[Fe/H]_{init}$, produces stars, some of which explode as supernovae, and the metallicity of the gas rises. Low-mass, long-lived stars then preserve the record of this enrichment process. The fraction of stars with higher metallicities rises until roughly 50% of the gas has been converted into stars or expelled from the system. At that point, the metallicity continues to rise, but the fraction of stars with higher metallicities diminishes owing to the decreasing gas supply, eventually reaching $[Fe/H]_{final}$ when all the gas is gone. The low mean metallicity of the halo is thought to arise because most of the gas was expelled from the halo before significant supernova-induced enrichment could occur. This gas may have formed much of the Galaxy's bulge. Figure 1 also suggests that two separate chemical evolution histories have been involved in the production of the metal-poor halo and the moderately metal-poor thick disk.

(2) Globular clusters are fairly massive systems, yet only one, ω Centauri, shows clear signs of metallicity variations that might be due to self-enrichment during the formation of the cluster. Since the colors of red giant branch stars depend on [Fe/H], the failure to detect *spreads* in colors of clusters' red giant branch stars implies a lack of significant self-enrichment. Given the small total mass in any one globular cluster, this may not be surprising. Some primordial variations in selected elements have been claimed, however.

(3) The abundances of the 'α' elements, oxygen, magnesium, silicon, calcium and titanium appear to be enhanced relative to iron by a factor of between 2 and 3 in metal-poor stars and clusters. Figure 3 shows the behavior of elements other than oxygen in field and in clusters, from which it is clear that ['α'/Fe] $\approx$ +0.4 for [Fe/H] <-1.0. The cause is thought to be the dominance of type II SUPERNOVAE, which are the explosive deaths of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

massive, and short-lived, stars. Models predict that such stars will produce all the elements, from helium to carbon and oxygen through the iron peak and beyond into the neutron capture domain. The decline in ['α'/Fe] is thought to arise from the later appearance of type IA supernovae, thought to be due to the deflagration–detonation of white dwarfs that exceed the Chrandrasekher mass limit owing to mass transfer. Their ejecta should be very rich in iron-peak elements and relatively deficient in lighter elements. Thus figure 3 may be the signature of the initial type II events, followed later ($\approx 10^9$ yr) by a mixture of type Ia and type II supernovae. Oxygen is the most abundant heavy element, but, unfortunately, disputes continue about the trend of [O/Fe] versus [Fe/H]. Analyses of OH lines in the ultraviolet suggest that [O/Fe] increases linearly as [Fe/H] declines, but analyses of O I in main sequence stars and of [O I] in red giants show the same behavior as in figure 3.

(4) Neutron capture elements, including those manufactured in the S-process and the R-process, can reveal the relative importance of these processes and, hence, the relative contributions to nucleosynthesis in the young Galaxy by the stars responsible for these processes. The s-process may arise largely in the late evolutionary stages of intermediate-mass stars, while the r-process is thought to occur within type II supernovae. As in the case of the 'α' elements, changing ratios of s-process and r-process abundances indicate changing relative contributions by asymptotic giant branch stars and supernovae. If the models are correct, r-process production should dominate the first several hundred million years of chemical evolution. Further, some of the r-process elements, such as thorium, are radioactive and, in principle, may have their abundances measured in stars, enabling their ages to be determined.

(5) The most metal-poor stars formed out of material enriched only slightly from stellar evolution and supernovae and a more easily detectable contribution from the big bang. Lithium, beryllium and boron are especially interesting since their production ratios are determined by the baryon density, its smoothness and the effects of cosmic rays. In almost all metal-poor stars that do not have deep convection zones that could carry lithium to deeper layers where it can be destroyed by proton capture, the lithium abundances appear to be constant, about a factor of 10 below the current interstellar medium value, and indicative of a low baryon density universe. There are, however, some puzzling exceptions, including a few metal-poor stars with little or no detectable lithium and a few others with enhanced lithium abundances.

Distances and motions

Radial velocities of stars in the field and in clusters may be measured directly using stellar spectroscopy and the Doppler effect. Tangential velocities may be determined only if a star's PROPER MOTION and distance are both measured since $v_{\rm tan} = 4.74\ \mu d$, where μ is the proper motion in arcsec per year and d is the distance in parsecs. If all these are known, one may determine velocities in the Galactic frame. In the solar neighborhood, these are called U, V and W, and are directed toward ($\ell = 180^\circ$; $b = 0^\circ$), ($\ell = 90^\circ$; $b = 0^\circ$) and ($b = 90^\circ$), respectively. With a model Galactic gravitational potential, Galactic orbits may be estimated, including apogalacticon, $R_{\rm apo}$, perigalacticon, $R_{\rm peri}$, and maximum distance from the Galactic plane, $Z_{\rm max}$. Stars and clusters at different $R_{\rm GC}$ may be handled similarly, but Π, Θ and Z replace U, V and W, retaining the same (local) vector orientations. Three-dimensional motions have been measured for over 30 globular clusters and over 1000 metal-poor field stars.

The distance scale for metal-poor stars remains a controversial but crucial topic. Distances to stars help to determine their velocities as well as their absolute magnitudes. The former are necessary in studies of relations between kinematics versus chemistry that provide insight into the Galaxy's dynamical history. The latter are vital because it is the luminosity of the main sequence turn-off that is used to estimate the relative and absolute ages of globular clusters.

Trigonometric PARALLAXES have been measured for a significant number of metal-poor main sequence stars, and such data may be used to calibrate relations between absolute magnitude, color and metallicity. These may in turn be used to estimate distances to those globular clusters with high-precision photometry of their faint main sequence stars. Another 'standard candle' is the RR Lyrae variable. It appears that $M_V({\rm RR}) = 0.20[{\rm Fe/H}] + 0.8$, although the zero point is uncertain by about 0.2 mag. Measurement of horizontal branch magnitudes, corrected for reddening, thus enables measurement of cluster distances. In fact, using the relation for M_V(RR) and measuring the apparent V magnitude of the main sequence turn-off yields the M_V value for the turn-off, *independent* of reddening.

Ages

Two lines of evidence suggest that the most metal-poor stars and clusters with [Fe/H] < -2.0 have the same age to the limit of our abilities to measure them. First, spectroscopic evidence indicates that the r-process dominates over the s-process in such systems, suggesting that intermediate-mass stars had not yet polluted the interstellar medium, and implying that the age differences probably do not exceed 10^9 yr. Figure 3 also suggests a similar timescale in that supernovae of type II dominated the production of iron. Second, evidence favoring common ages comes from color–magnitude diagrams. Recent ground-based and space-based observations have extended the comparisons to large distances, including NGC 2419 with $R_{\rm GC} = 90$ kpc. Figure 4 shows that this cluster is indistinguishable in age from M92, which lies only 9 kpc from the Galactic center, and is similar in age to several other low-metallicity ([Fe/H] < -2.0) clusters with $R_{\rm GC}$ values ranging from 6 to 9 kpc. Again, the maximum age spreads are roughly a billion years or less. Analyses of field star ages using Strömgren photometry also show no sign of age differences among the most metal-poor stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

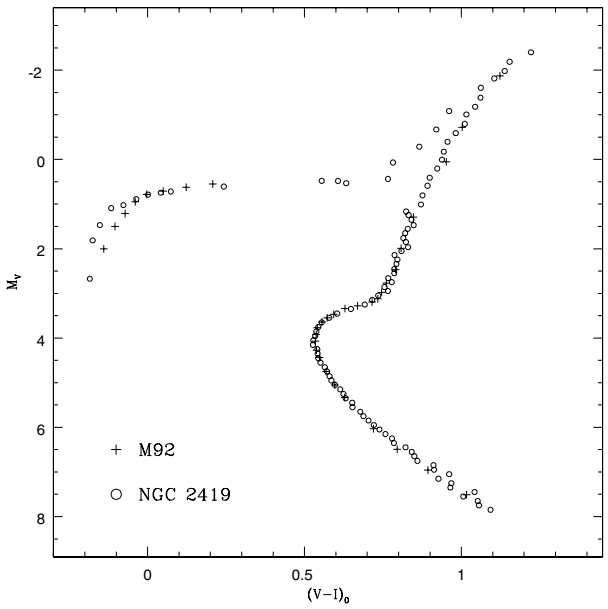


Figure 4. Superposed color–magnitude diagrams of the middle halo globular cluster M92 and the distant cluster NGC 2419, taken from Harris *et al* (1997).

The next important chronological questions are whether the metal-poor and metal-rich field stars and clusters differ in age and whether this difference is a function of their location in the Galaxy. The age differences imply the speed at which the metal-poor halo was chemically enriched, and differences in initial densities of the young Galaxy's central, middle and outer halo may have produced differences in the rate at which stars and clusters formed and the rate at which the metallicity enrichment proceeded. At large R_{GC}(>80 kpc), it appears that intermediate-metallicity clusters, with [Fe/H] ≈ -1.5, are of order 2×10^9 yr younger than comparable metallicity clusters with R_{GC} values of 6–12 kpc. The implication is that the metallicity enrichment proceeded more slowly in the outer halo. Thus the differences in gas density between the outer and middle halo may have been an important factor in the Galaxy's evolutionary rates. We do not yet know whether these trends of equal ages at the lowest metallicities and smaller age spreads continue into the innermost halo ($R_{GC} < 3$ kpc).

The age spread among the intermediate-metallicity and relatively metal-rich ([Fe/H] ≈ -1) clusters in the middle halo remains a matter of debate. Field stars appear to show age spread at such intermediate metallicities. There are clear examples of significant age differences among clusters, the pair NGC 288 and NGC 362 being one of the best examples. Another pair of unusually young clusters is Ruprecht 106 and Palomar 12, which appear to be several billion years younger than average. In fact, the ['α'/Fe] ratios for these clusters are close to solar and, hence, very low for their [Fe/H] values, as shown in figure 3. Their behavior may be explained by very young ages (but then the question is why are they so metal poor?) or by an origin perhaps quite different from that of the bulk of the metal-poor globular clusters. In any case, there seems little doubt that *some* globular clusters are younger than average, but how many and by how much? To attempt to detect smaller age differences than possible from the main sequence regime, shown in figure 4, attention has turned to the HORIZONTAL BRANCH STARS, identified in the figure as those stars with $M_V \approx +0.5$ mag. Theory predicts that, for intermediate-metallicity clusters, younger clusters might be distinguished from older clusters by the distribution of stars along the horizontal branch: redder indicates younger ages. NGC 362, with a red horizontal branch, is indeed younger than NGC 288, which has a very blue horizontal branch. The problem is actually quite complex, and this 'second parameter', the color of the horizontal branch compared with those of other clusters with similar [Fe/H] values (the 'first parameter'), may also be affected by stellar interactions in dense environments. Indeed, some clusters, such as NGC 2808 and NGC 6229, show bimodal horizontal branches yet no signs of two epochs of star formation. What can be said is that *some* clusters with intermediate metallicities in the middle halo are younger than average.

The absolute ages of the Galaxy's oldest globular clusters are of great interest but two key problems remain. First, if star formation began first in the Galaxy's central regions, the oldest clusters, those with low [Fe/H] values but lying close to the Galactic center, have not yet been studied. Of greater concern, however, is the dichotomy in results for M_V(RR) that has arisen from the Hipparcos satellite's trigonometric parallaxes. Absolute magnitudes for field metal–metal main sequence stars have been used to estimate distances and main sequence turn-off luminosities for comparably metal-poor globular clusters. Ages range between 12 and 14 billion yr. However, the M_V(RR) values derived are brighter than those of field RR Lyrae variables obtained with Hipparcos. On the other hand, the detection of thorium (and other r-process elements) in the spectra of some metal-poor field stars, most notably CS22892-052, has enabled estimates of stellar ages from radioactive decay. Values near 14 billion yr appear favored.

Formation and evolution of the halo

Early work on the chemical abundances, velocities and ages of what we now call the metal-poor halo field stars and clusters culminated in a classic paper by Eggen *et al* (1962). They derived U, V and W velocities for 221 field dwarf stars selected from two catalogs: one of nearby stars and one of stars with large proper motions (hence a good source of high-velocity stars). They estimated metallicities from the 'ultraviolet excess', $\delta(U-B)_{0.6}$, which in essence is an indicator of the degree of heavy element line blanketing in the near ultraviolet. The velocities were combined with a model Galactic gravitational potential to estimate the stars' orbital eccentricities, and the most important result

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

was that the orbital eccentricity correlated well with the metallicity. The orbital eccentricities of the most metal-poor stars were found to be uniformly high, indicating that the stars were formed on plunging orbits. Assuming metallicity is a proxy for time, the conclusion was that the Galaxy began to form stars while it was still in a state of collapse. The large spherical halo could then be explained as the earliest stage of the Galaxy, following which it rapidly collapsed and settled into a disk configuration. The process was rapid because the plunging orbits implied a nearly free-fall timescale, of order 2×10^8 yr, and it was a coherent process since the study did not reveal stars with low metallicities (the oldest stars) with disk-like orbits or stars with high metallicities (the more recently formed stars) with high-velocity halo-like orbits.

The first major challenge for the rapid–coherent collapse model was the study of Galactic globular clusters by Searle and Zinn (1978). One key point of the collapse model is that globular clusters should all have the same age, at least to within our abilities to measure them. Searle and Zinn noted that clusters with large R_{GC} tend to have redder than average horizontal branch colors. If the 'second parameter' is age (which appears to be true for the most distant clusters), then the process was not as rapid as originally believed. A slower, coherent contraction would still be a reasonable interpretation of the data, except that in such a case enough time would have passed to permit the remaining gas to be enriched by supernovae. As the gas contracted to form the disk, one would then expect the innermost stars and clusters to have higher metallicities since star formation should have continued longer. A radial gradient in metallicity would have been established, but Searle and Zinn failed to find any evidence of such a gradient for R_{GC} values larger than about 8 kpc. Based on their cluster data, they proposed that the halo was formed out of 'protogalactic fragments', each with an individual metallicity enrichment history. The assemblage of these fragments would then have resulted in a dynamically hot halo whose field stars and globular clusters have a range in ages yet no metallicity gradient owing to the random, 'incoherent' nature of the process. Additional support for the 'incoherent' model came from studies of field stars. Kinematically biased studies of stars selected only from proper motion catalogs (the 'Carney–Latham' sample) and stars selected without kinematical bias by Norris revealed a lack of correlation between metallicity and kinematics, as shown in figures 5 and 6.

An obvious argument favoring an accretion origin for some, perhaps most, of the halo was the discovery of the SAGITTARIUS DWARF GALAXY by Ibata *et al* (1994). This is a large dwarf galaxy on close approach to our Galaxy. In an astronomically short time, it will shed many stars, including its four globular clusters (M54, Terzan 7, Terzan 8 and Arp 2), into the Galaxy's halo. Several other signs point to the importance of accretion as a source of some of the halo's stars and clusters.

First, if the evolution of the halo into the disk was a coherent process, the halo should possess modest net prograde rotation. Figures 5 and 6 show that if the local Galactic rotational velocity, Θ_0, is 220 km s^{-1}, then the net halo rotation is very small. Further, Majewski (and others) have found that stars 5 kpc or more from the plane may possess net retrograde rotation. This suggests that at least some fraction of the halo's stars did not participate in the

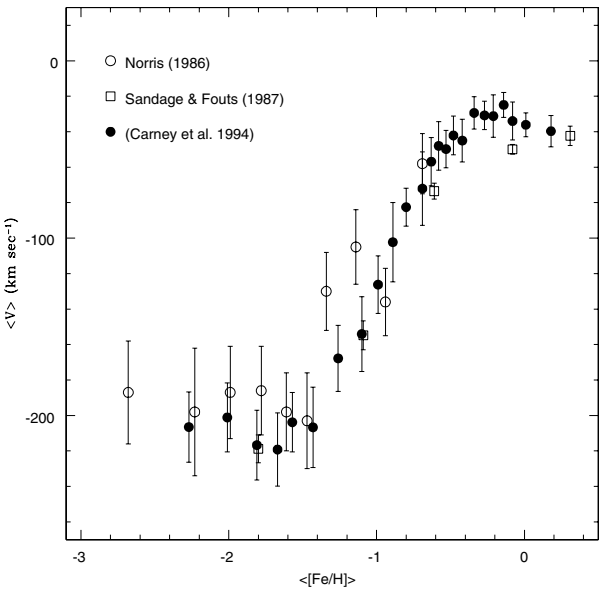


Figure 5. Average velocities V as a function of [Fe/H] using equal bin sizes from kinematically biased (Sandage and Fouts 1987; Carney *et al* 1994) and unbiased (Norris 1986) samples of field stars.

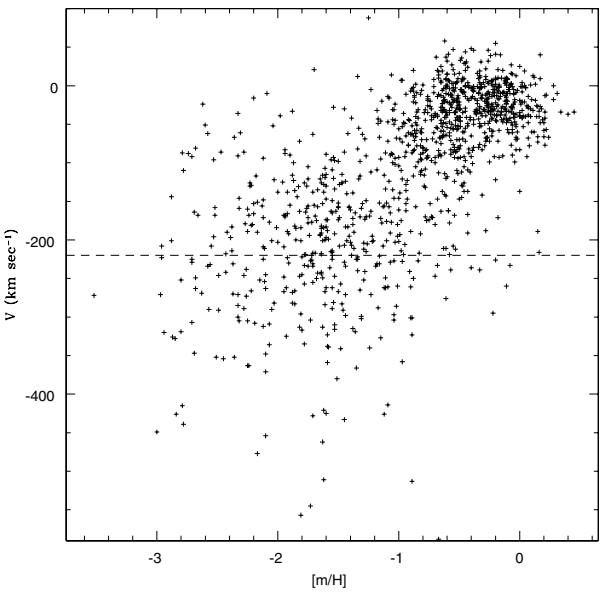


Figure 6. The velocity V versus [m/H] from Carney *et al* (1994). The dashed line divides prograde and retrograde Galactic orbital motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

coherent Galaxy formation process. Other studies have suggested that globular clusters with horizontal branch colors implying younger-than-average ages also show signs for retrograde rotation. The evidence is weak given the one-dimensional velocity information and the small sample sizes, however.

Another possible signature of independent origin and evolution is chemistry. If two gas clouds produce stars at similar rates and with similar mass functions, the supernova enrichment history should be the same, and in consequence identical trends of [α/Fe] versus [Fe/H] should result. However, if one cloud forms stars more slowly, the appearance of the effects from the type Ia supernovae should first appear when the mean metallicity is still 'stuck' at a lower value, perhaps much lower than seen in the Galaxy at [Fe/H] ≈ -1. Ruprecht 106 and Palomar 12, identified in figure 3, could be explained by delayed star formation in a small 'fragment' or dwarf galaxy that was accreted by the Galaxy. A small number of field stars have also been found with unusually low [α/Fe] ratios, the most extreme being BD+80 245. An interesting additional feature of these stars is that they are generally very high velocity, with orbits that carry them out to $R_{GC} = 20$–40 kpc. Ruprecht 106 and Palomar 12 also lie at fairly large distances, $R_{GC} = 17$ and 14 kpc, respectively. On the other hand, not all stars with such large apogalacticon values or clusters with such large distances have unusual abundance ratios. Studies of such systems appear to more frequently find the enhanced [α/Fe] values typical of middle halo stars and clusters.

Another avenue to explore for signs of accretion is the idea that disintegrating satellite galaxies would leave behind a stream of stars (and gas, if they possess any). The MAGELLANIC STREAM is an example of a spoor that is apparently induced by the Galaxy's tidal effects on the MAGELLANIC CLOUDS. (Note this also indicates that the Clouds are also doomed to merge with the Galaxy.) The Magellanic stream traces almost an entire great circle on the sky, and the stream also passes near the positions of the Draco and Ursa Minor dwarf spheroidal galaxies. Another such great circle includes the Fornax, Leo I, Leo II and Sculptor dwarf galaxies, as well as a number of Galactic globular clusters. Another search technique recognizes that each distant dwarf galaxy or globular cluster in orbit about the Galaxy has an orbital pole and that, in the absence of three-dimensional motions, the pole occupies a point along a great circle in the sky. Intersections of such great circles suggest common orbital poles and have been used to identify possible streams. However, until accurate proper motions become available for more globular clusters and the distant dwarf spheroidal galaxies, the reality of these common origins for some such systems must remain speculative. Finally, common origins for larger numbers of objects but with all three components of velocity known may show up in the form of velocity 'substructure'. A dynamically hot population such as the halo should have stars well dispersed in velocity space. Observed 'clumpings' of stars at high velocities indicate either that the halo is not well mixed kinematically or that these clumps arose because of tidal destruction of small stellar ensembles by the Galaxy's tidal forces.

Even if the evidence cited above supports accretion of proto-Galactic 'fragments' or dwarf galaxies, this does not necessarily rule out that many of the metal-poor, high-velocity stars were part of a coherent formation process, as outlined by Eggen *et al*. While some subsets of globular clusters may show net retrograde rotation, others show prograde rotation. While some globular clusters may be younger than average, many appear to be coeval with other clusters of similar metallicity. Some globular clusters and field stars, especially with large R_{GC} or R_{apo} values, have unusual [α/Fe] values, but most do not, even clusters like NGC 7006 ($R_{GC} \approx 35$ kpc) and numerous extreme-velocity field stars. Furthermore, metal-poor stars close to the Galactic plane show not only prograde rotation but signs of a radial metallicity gradient, as do subsets of globular clusters. Both processes, coherent evolution plus accretion, have almost certainly contributed to the formation of our Galaxy's metal-poor halo, but a much larger and more complete census will be required to determine the relative numbers of natives and immigrants.

Bibliography

The original division of the local stellar neighborhood into 'high velocity' and 'low velocity' was made by

Oort 1926 *Groningen Obs. Publ. 40*

This was followed much later by the recognition that some stars are extremely deficient in heavy element abundances:

Aller 1960 *Astron. J.* **65** 399
Aller and Greenstein 1960 *Astrophys. J. Suppl.* **5** 139
Baschek 1959 *Z. Astrophys.* **48** 95
Chamberlain and Aller 1951 *Astrophys. J.* **114** 52

The ΔS calibration calibration may be found in

Clementini *et al* 1995 *Astron. J.* **110** 2319
Fernley and Barnes 1996 *Astron. Astrophys.* **312** 197

W' is discussed by

Da Costa and Armandroff 1995 *Astron. J.* **109** 2553

An excellent discussion of chemical evolution models is provided in

Pagel *Nucleosynthesis and Chemical Evolution of Galaxies*

The halo–bulge connection is summarized by

Wyse and Gilmore 1992 *Astron. J.* **104** 144

[α/Fe] is discussed at length by

Wheeler *et al* 1989 *Ann. Rev. Astron. Astrophys.* **27** 279

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

while light element abundances reflective of the big bang are discussed by

Boesgaard and Steigman 1985 *Ann. Rev. Astron. Astrophys.* **23** 319

and the Type II *versus* Type Ia supernovae effects are illustrated by

Wyse and Gilmore 1988 *Astron. J.* **95** 1404

Globular cluster space velocities may be found in

Dinescu *et al* 1997 *Astron. J.* **114** 1014
Dinescu *et al* 1999 *Astron. J.* **117** 277
Odenkirchen *et al* 1997 *New Astron.* **2** 477

Good examples of the relative contributions of s-process versus r-process elements as a function of metallicity are presented by

Gratton and Sneden 1994 *Astron. Astrophys.* **287** 927
Krishnaswamy-Gilroy *et al* 1988 *Astrophys. J.* **327** 298

Ages of metal-poor field stars were derived by

Marquez and Schuster 1994 *Astron. Astrophys. Suppl.* **108** 341

The relative ages of the most metal-poor clusters are summarized well by

Harris *et al* 1997 *Astron. J.* **114** 1030
Stetson *et al* 1996 *Publ. Astron. Soc. Pac.* **108** 560

The distant intermediate metallicity clusters are discussed by

Stetson *et al* 1999 *Astron. J.* **117** 247

The age difference between NGC 288 and NGC 362 is summarized by

Sarajedini *et al* 1997 *Publ. Astron. Soc. Pac.* **109** 1321

The 'second parameter' as a possible age indicator is presented by

Lee *et al* 1994 *Astrophys. J.* **423** 248

Radioactive dating of field stars is described by

Cowan *et al* 1997 *Astrophys. J.* **480** 246

Some of the key papers dealing with the origin of the halo are

Eggen, Lynden-Bell and Sandage 1962 *Astrophys. J.* **136** 762
Searle and Zinn 1978 *Astrophys. J.* **225** 357

They lay out the basic arguments in favor of a rapid, coherent collapse of the halo prior to the formation of the disk (Eggen *et al*) and an alternative involving independent evolution of 'proto-Galactic fragments' during or following the formation of the Galaxy. Thorough summaries of globular cluster kinematics, including early versions of figures 1 and 2, are in

Da Costa and Armandroff 1995 *Astron. J.* **109** 2533
Zinn 1985 *Astrophys. J.* **293** 424

The latest results from the 'Carney–Latham' sample are discussed

Carney *et al* 1996 *Astron. J.* **112** 668

while the work by Norris may be found in

Norris 1986 *Astrophys. J. Suppl.* **61** 667

The possibility that stars far from the plane are on retrograde orbits was first raised by

Majewski 1992 *Astrophys. J. Suppl.* **78** 87

The discovery paper of the Sagittarius dwarf galaxy is

Ibata, Gilmore and Irwin 1994 *Nature* **370** 194

Good examples of searching for accretion patterns by lining up objects on great circles on the sky are summarized by

Fusi Pecci *et al* 1995 *Astron. J.* **110** 1664

and the use of orbital poles is given by

Lynden-Bell and Lynden-Bell 1995 *Mon. Not. R. Astron. Soc.* **275** 429

Velocity 'substructure' is discussed well by

Majewski *et al* 1994 *Astrophys. J. Lett.* **427** L37

Bruce W Carney

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Nucleus

The central core of a galaxy. In elliptical galaxies and in the central bulges of spiral galaxies, the nucleus corresponds to the location at which the optical brightness reaches a maximum, and to the region within which stars are most densely concentrated. Irregular galaxies do not contain well-defined nuclei. Active galaxies, which comprise a few percent of the total number of galaxies, contain intensely bright, compact nuclei, called active galactic nuclei (or AGNs), that radiate strongly over a wide range of wavelengths (from x-ray to radio) and in many cases display substantial variability.

The nucleus of our own Galaxy, the Milky Way, contains a compact source of radio emission, called Sagittarius A* (pronounced Sagittarius A-star), is less than 15 astronomical units in diameter and appears to coincide in position with a relatively weak source of x-radiation. Ionized gas clouds within a few light-years of the center are moving at speeds of up to several hundred kilometers per second under the gravitational influence of a central concentration of some 5 million solar masses, only about half of which appears to be contained in stars. These observations, taken together, are consistent with the hypothesis that the galactic nucleus contains a black hole of some 2.5 million solar masses surrounded by an accretion disk. The Milky Way Galaxy, therefore, appears to contain a miniature, and low-powered, version of an active galactic nucleus. Because the nuclei of some other 'normal' galaxies display similar characteristics, it has been suggested that many, perhaps all, galaxies contain massive black holes at their centers.

The Andromeda galaxy has a double nucleus, the brighter component of which may be the core of another galaxy that, following a galactic collision, has been absorbed by the larger system.

See also: active galactic nucleus, elliptical galaxies, galaxy, interacting galaxies, Milky Way Galaxy, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Open Clusters

Galactic open clusters are beautiful objects to view through telescopes, especially when blue and red supergiant stars are present in the cluster field. They are stellar systems containing from a few hundred to several thousand stars and are readily distinguished on sky photographs by a density of stars much larger than that in the surrounding field. About 1200 such objects have been catalogued so far and about half of them have been investigated at least once. Typical clusters are, by order of increasing age, the Orion nebula cluster, the Pleiades, Hyades and Messier 67. The brighter objects have been listed in the Messier catalogue, while most are recorded in the New General Catalogue (NGC) of Dreyer (1888). Open clusters are different from GLOBULAR CLUSTERS in many aspects. They are found mainly within or close to the Galactic plane and have ages ranging from a few million to ten billion years, while globular clusters populate the Galaxy halo and are as old as the universe. Consequently, open clusters may be used to describe the history and evolution of our Galaxy after the formation of the disk.

The strong interest for open clusters results from their unique and fundamental properties: the stars in a cluster are at the same distance from the Sun, (the depth effect is negligible for open clusters at distances larger than about 300 pc), they have the same chemical composition (the cosmic scatter on the chemical composition is much smaller than the precision of the determination of element abundances) and the same age (the age spread at cluster formation amounts only to a few million years and appears to be important only for very young open clusters with ages less than 10 million years). The stars, however, differ by their mass: the most massive stars in the youngest open clusters have masses of about 80 solar masses, and the lower mass limit is below 0.08 solar masses, reaching the domain of brown dwarfs.

Consequently, open clusters offer many opportunities of investigating the dependence of astrophysical phenomena on the stellar masses (by working in one cluster both the chemical composition and age are fixed) or on ages (by selecting stars of a given mass in several clusters of a given chemical composition). For example, the x-ray observations obtained by the ROSAT satellite for clusters of different ages have permitted the study of the decrease of the x-ray emissivity of solar-type stars with the increasing ages of the parent clusters.

Open clusters seem to be really ideal tools, but there are a number of problems to solve before one can fully benefit from their diagnostic power. Like other (distant) objects, open clusters suffer from the effects of absorption of light by interstellar matter, which has to be first estimated. The CLUSTER MEMBERSHIP of the stars is the second important problem to solve. On photographs or CCD frames, cluster members and field stars are all projected on the same plane and there is no direct way to separate them. Additional observations are needed to separate the cluster members from the field stars.

Fortunately, this can be achieved, at least for clusters not too far away, by using another property of open clusters: all member stars share the same spatial motion; if this were not the case, the group would rapidly dissolve. Therefore, by measuring their PROPER MOTIONS relative to the background stars or galaxies on photographic plates taken at a large epoch difference, which may now reach 100 yr, it is possible to compute the membership probability of each star and select the candidate members. The RADIAL VELOCITY provides also a powerful criterion to select the cluster members, but was a rather difficult and slow method to implement until recently, because the spectra used to be recorded on glass plates. New more efficient detectors and new specialized instruments have led to the increased use of this cluster-membership selection method.

Basic cluster parameters are the amount of interstellar absorption (reddening), the distance, chemical composition and age. The last of these is not a quantity directly obtained from the observations but through the comparisons with theoretical models. When enough clusters have been observed and their parameters determined, then astrophysical and Galactic applications can be undertaken. We are facing a case of extreme information concentration: on CCD frames thousands of stars have to be measured to obtain a few parameters describing the open clusters. Also, for example, when enough ages have been obtained, one parameter, the slope of the age distribution, will summarize the data resulting from large observing efforts.

The importance of open clusters

Different astrophysical problems are investigated with open clusters of different ages.

Very young open clusters, with ages less than a few million years, are the direct results of star formation processes and are still associated with the cloud from which they are born, which gives rise to spectacular bright nebulae, such as the Orion nebula. Some clusters are still embedded in clouds which are completely dark at visible wavelengths. They are only revealed by near-infrared imaging. An extensive description of very young open clusters is crucial to constrain the various scenarios of molecular-cloud fragmentation and STAR FORMATION. Ideally, one would wish to address the following topics:

- the number of stars of various mass formed,
- the percentage of binaries among stars of different mass,
- the distribution of their orbital parameters,
- the initial distribution of stellar rotation,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- the sequence of star formation and the importance of age spread,
- the radial structure and the importance of initial mass segregation.

Observations indicate that stars frequently form in clustered environments, either in rich clusters of many hundreds to many thousands of stars or in smaller groups and aggregates containing of order ten to a few tens of stars. It is only recently that properties of extremely young clusters are beginning to be well characterized.

Young open clusters, covering the range from 10's to 1000 million years, are especially interesting because of the variety of their stellar content. They contain several kinds of VARIABLE STARS and of 'abnormal' stars: emission-line stars and chemically peculiar stars. When the distance and age of the parent clusters are known, the intrinsic properties of these stars are determined more accurately than by any other means.

Intermediate-age and old open clusters, older than 1000 million years, have been mostly useful to understand the main trends and details of STELLAR EVOLUTION, because the evolutionary effects are more evident with a spectacular red giant branch. The best-observed old cluster, Messier 67 (NGC 2682), is a prototype for this kind of morphology and contains numerous RED GIANT STARS. Because cluster stars have the same age, each cluster defines an isochronous sequence in the photometric equivalents of the Hertzsprung–Russell (HR) diagram. Such sequences can be computed from grids of evolutionary models and compared with the observed data.

Properties of open clusters

The color–magnitude diagram

The main tool used to study open clusters is the HERTZSPRUNG–RUSSELL DIAGRAM in which the star's luminosity is plotted as a function of the temperature. In such a diagram each point represents the observations obtained for one star. The original form of the HR diagram, based on the absolute magnitude and the spectral types, has been replaced by a photometric equivalent form which displays the apparent magnitude as a function of a color index.

A color index represents the ratio of the fluxes measured through two filters: for example, the widely used *B*–*V* index results from the ratio of the fluxes emitted in the blue (*B*) and visible (*V*) parts of the stellar energy distribution. Several photometric color indices are closely related to the effective temperature.

The apparent *V* magnitude differs from the absolute magnitude by a constant which is called the distance modulus of the cluster. This quantity, denoted by $m - M$, is related to the distance r, expressed in parsecs (pc), by the simple relation $m - M = 5 \log r - 5 + A_V$, where A_V represents the correction for the interstellar absorption. The visual absorption A_V is itself proportional to the color excess: $A_V = RE\ (B - V)$, where $E\ (B - V) = (B - V) - (B - V)_0$ is the difference between the star's intrinsic color $(B - V)_0$ (not affected by interstellar absorption) and that observed. R is the ratio of total to selective absorption and must be determined observationally. Often a value of R of 3.2 (mag/mag) is assumed. These simple relations show that the determination of the amount of interstellar absorption is very important for the determination of the distance.

Examples of typical color–magnitude diagrams are shown in figures 1, 2 and 3, based on data extracted from the WEBDA database described in the final section. In very young open clusters, such as the Orion trapezium cluster (figure 1), only the most massive stars have completed their formation and have reached the ZAMS, while the less massive stars are still in the phase of contraction toward the ZAMS. Membership of these PRE-MAIN-SEQUENCE STARS is difficult to determine from photometric criteria only and spectroscopic observations are needed to search, for example, for the presence of emission in the Balmer Hα hydrogen line or of lithium absorption lines, which are both signs of youth. When possible, proper motions and radial velocities are of course also used to confirm the membership.

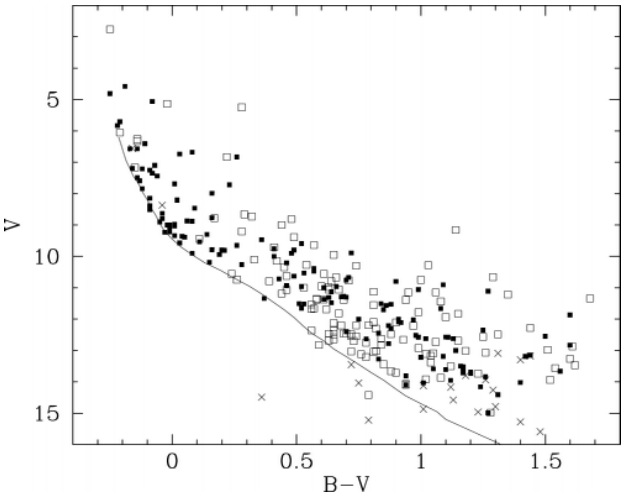


Figure 1. Color–magnitude diagram for the Orion nebula cluster. The star brighter than $V = 10$ have completed their formation, while the stars fainter than this limit are still contracting toward the zero age main sequence (ZAMS) (continuous line). Filled squares are for the stars with membership probability larger than 50%, open squares for stars with probability less than 50%, and crosses for stars without membership information. The nebulosity in the cluster region absorbs the light of some stars and alters their colors, making them redder. This is called differential reddening. Additional information is needed to separate the stars really belonging to the cluster from the field stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

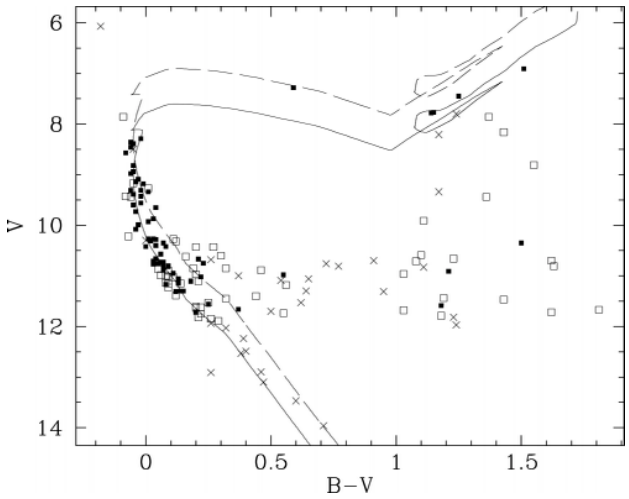


Figure 2. Color–magnitude diagram for NGC 2287 (M41). Symbols are related to the cluster membership of the stars (filled squares, more than 50% chance of being a member; open squares, less than 50%; crosses, no information). The continuous curve is a theoretical isochrone for 240 million years. The dashed curve is the same isochrone shifted by 0.75 mag to mimic the binary boundary. Both curves delineate the main sequence band and reproduce quite well the observed morphology.

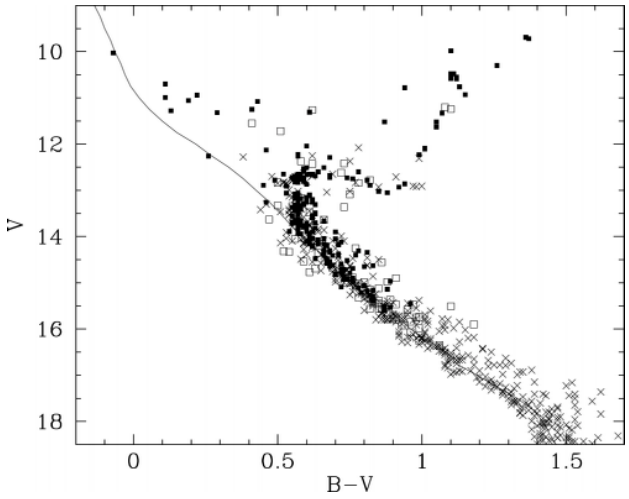


Figure 3. Color–magnitude diagram for NGC 2682 (M67). In this characteristic diagram, the main sequence is well populated. The upper main sequence is bending toward the right (redder color) under the effect of stellar evolution. After a small gap, resulting from a rapid phase of evolution at the end of the hydrogen burning phase, the structure of the stars changes and the stars move toward the right of the diagram. They become red giants and start ascending the red giant branch. The helium burning phase lasts for a longer time and produces a small clump of stars with similar magnitudes and colors, located at V ~10.5 and $B - V$ ~1.05. Blue stragglers are located to the left of the main sequence bending ($B - V < 0.50$), in a position where stars are not expected to remain according to the cluster age. The thickness of the main sequence is due to the presence of binary stars. Symbols are the same as those in figure 2.

If all stars were single, the sequence observed in the color–magnitude diagrams would be very thin. However, even with the high-precision photometric data obtained with the Hubble Space Telescope, the main sequence presents a finite thickness. The major part of the scatter is produced by BINARY STARS. Obviously, unresolved double stars are brighter than single stars of similar temperature. Because the magnitude is on a logarithmic scale, the addition of the fluxes of two identical stars makes the binary system brighter by 0.75 mag. The main sequence band, defined as the zone between the ZAMS and the upper binary ridge, is then populated by binary stars which have mass ratios less than 1 (the mass ratio $q = M_B M_A^{-1}$ is by definition always smaller than 1). The smaller the mass ratio, the closest the stars to the ZAMS. When the ratio of the luminosities (L_A/L_B) reaches the value of 100, the presence of the secondary is no longer distinguishable.

Stellar rotation and peculiarities also introduce scatter on the star position within the main sequence band by affecting the magnitude and/or the colors.

However, the main characteristic visible in color–magnitude diagrams is the effect of the stellar evolution. At any age, the more massive stars are more evolved than low-mass stars, because the speed of evolution depends on the initial mass. This explains why the upper part of the main sequence is evolved away from the ZAMS and presents a characteristic shape for each cluster and age. The comparison of the observations with theoretical isochrones permits at the same time the determination of the ages of open clusters and the testing of evolutionary models. For 10 yr, models with larger cores, including the effects of convective-core overshooting, have been preferred to the so-called canonical ones, because they better reproduce the observed diagrams.

Luminosity and mass functions

Young open clusters are the direct products of the processes of star formation and therefore provide fundamental clues to the understanding of cloud fragmentation. Ideally, if we were able to count the number of stars of different masses, we should be able to describe the end-product of the formation processes, i.e. the initial mass function. What is observed in fact is the luminosity function, defined as the number of stars in magnitude bins, and masses are deduced from a mass–luminosity relation. As a rule, most clusters conform to the Salpeter law, although the universal character of this law has not yet been fully demonstrated. It is not a trivial task to decide whether the differences observed between the luminosity or mass functions in different open clusters are real or produced by biases and incompleteness. The complete description of the stellar population, although of fundamental importance, is quite difficult to derive because we face the problems of membership determination, stellar multiplicity detection

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and the completeness of the samples, because the radial extension and structure of the clusters are still poorly known. In addition, because of the possibilities for stars to escape from their parent clusters, the observed present-day mass function can be related to the cluster initial mass function only through a thorough knowledge of the dynamical evolution, which is difficult to achieve.

For these reasons, for none of the clusters do we have complete information, although meaningful progress has been made in several areas.

Radial and spatial structure

The structure of open clusters has been for a long time mainly a topic for theoreticians and numerical simulation. All we knew was derived from models, because relevant observations were not available. As more proper motions and photometric studies became available, it was realized that open clusters were not limited to the bright central area which is prominent on photographs, but that they extend to much larger distances. The problem is, however, to identify the stars in the surroundings. The statistical approach consisting in counting stars at larger and larger distances from the center until one reaches the level of the sky background obviously fails when the density of cluster stars is lower than the fluctuation of the number of stars in the sky background. Therefore, there is no other solution but to identify the stars one by one in the cluster corona with the help of proper motions and radial velocities, which ensures that the stars are moving together with the same spatial motion, and of photometric information, which proves that the stars are at the same distance.

The radial distributions of stars have confirmed the theoretical predictions: in most clusters the massive stars are found close to the cluster center, while less massive stars are more uniformly distributed over the cluster area. This mass segregation is caused by star interactions or collisions which redistribute the kinetic energy and make the massive stars fall toward the cluster center. The same phenomenon is also true for the binaries, for which the sum of the masses makes them heavier than single stars. However, recent observations show that the time evolution of mass segregation is not straightforward. As a cluster emerges from its parent cloud, it undergoes dynamical changes and the most massive stars are found at the cluster center. We are facing a dilemma: either the massive stars are formed at the cluster center and remain there or they are brought to the center by some forces, for example dynamical friction in the cloud. From this initial state, a fine mass segregation will develop by dynamical evolution of the cluster until the mass difference between the maximum and minimum masses becomes small and the star distribution becomes more uniform, when stars have more or less the same masses.

Another prediction of the models is that open clusters lose stars by an evaporation process. During encounters one star may acquire enough energy and speed and become able to leave the cluster, especially after the collision of a single star and a binary system, during which star exchange may happen, with the concomitant ejection of one component. This phenomenon has the consequence that open-cluster lifetimes are limited: if they are not massive enough to retain their stars, the cluster will dissolve within a few 100 millions years. As recent observations of star forming regions have shown that stars predominantly form in clusters, the field stars are explained as resulting from the disruption of many clusters. The verification of this hypothesis requires numerous comparisons of the properties of field and cluster stars. Tidal interactions of clusters with molecular clouds in the Galactic disk also contribute efficiently to the disruption of the clusters.

Rotation distribution

Since the beginning of the 20th century and until recently, high-resolution spectra could be recorded on glass plates only for relatively bright stars. Therefore information on stellar rotation in open clusters was limited to the brighter stars on the upper main sequence. Furthermore, the quantity determined was not exactly the rotation speed, but the product of the equatorial velocity and the inclination angle i of the rotation axis, i.e. the quantity $V \sin i$. Therefore it was not possible to easily know whether a star with a small $V \sin i$ measurement was rotating slowly or whether it was seen more or less pole on. The velocity is maximum if the star is observed equator on, i.e. if the angle is close to 90º ($\sin i = 1$). New instruments and larger telescopes have allowed the $V \sin i$ observations to be extended to solar-type stars and the distribution of projected rotation along the main sequence to be derived.

For O- to middle F-type stars the maximum rotation shows a moderate decrease, but, because of the presence of the factor $\sin i$, most values between 0º and the maximum possible rotation are observed. However, at about 7000 K, the onset of convection in the stellar atmosphere rapidly brakes the rotation. The initial phase of braking is rather rapid, reducing the projected rotation to values of $V \sin i$ less than 10 km s^{-1}. Subsequently the effect is not as marked and the rotation is more gradually reduced during the star main sequence lifetime.

The observations of main sequence solar-type stars in open clusters of various ages, between 1 million and 1 billion years, have permitted us to understand how rotational braking develops. There seems to be no uniform law: each star is braked depending on its initial velocity and mass. The initial spread is thought to be linked to the time at which decoupling of the accretion disk surrounding the pre-main-sequence stars occurs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binary frequency

Stars are usually formed in binary or multiple systems, although the true frequency, which may vary between 25% and 60%, is not exactly known, because we are not yet able to detect all possible binaries along the main sequence, down to the faintest stars. Frontier technology (SPECKLE IMAGING OF BINARY STARS, INTERFEROMETRY and ADAPTIVE OPTICS imaging) now allows the detection of binary systems with separations between 0″.05 and 1″.0. Further progress in spectroscopic observations will be possible with new multifiber spectrographs which will be able to obtain the spectra of some 100 stars simultaneously. Systematic observations of representative clusters have already been undertaken, which will produce, in the next few years, a great development of our knowledge of the duplicity of stars in various open-cluster environments, ages and richness.

The radial-velocity spectrometers, developed in the mid 1970s, have permitted the measurement of the radial velocities of solar-type stars in nearby clusters and the determination of the orbital parameters of most SPECTROSCOPIC BINARIES discovered. Main sequence binaries with periods smaller than 8 days have circular orbits, while all others have elliptical orbits, the mean eccentricity being around 0.3. This cut-off period results from the effects of tidal interaction between the two stars in binary systems, which are stronger during the pre-main-sequence phase because the stellar radii are larger and the atmospheric convection zone is more developed. The transition limit between circular and elliptical orbits is around 150 days for binaries with a red giant primary and 1500 days for supergiant stars.

Comparison of the binary frequencies and separation distributions in various environments, namely from loose star forming regions, such as the Taurus–Auriga complex, to the dense cluster cores, as in Orion where the mass density reaches 10 000 solar masses pc^{-3} in the 0.2 pc central region, have shown that these characteristics are not the same. Do these differences originate from the formation processes themselves or do some dynamical interactions modify the original state? The answer is not known yet.

Variable stars

Various kinds of variable stars have been found in open clusters, as seems normal because intrinsic variability depends on the position in the HR diagram, as is the case for the δ Cephei variables. However, their presence in open clusters permits us to establish or study the relations between their periods, luminosity and chemical composition. Comprehensive surveys have been performed to detect other types of variable stars, called β Cephei variables, among the B-type stars in the upper main sequence of young open clusters. Surveys have also been undertaken with bidimensional detectors (CCDs) to search for δ Scuti stars, another kind of pulsating variable which presents often a combination of several periods of pulsation simultaneously.

Apart from the intrinsic pulsating variables, there are additional kinds of variables in open clusters: the flare stars in very young clusters, ECLIPSING BINARY STARS and stars in which the variability is induced by rotational modulation. Continuous observations are required to perform an extensive detection of variable stars. However, at the time of photoelectric photometry, stars were observed one by one and therefore the number of observations obtained for each star was not sufficient to have a complete detection of variablity. The advent of CCD cameras now allows us to obtain large numbers of frames over various time intervals, which enables the determination of variability periods up to several tens of days.

One kind of variable star discovered in open clusters in the beginning of the 1980s is related to the variability induced by the presence of spots on the star's surface and the rotation of the stars. Because the spots are less luminous than the star surface, the presence of spots makes the stars a little less luminous, by a few 0.01 magnitudes. The rotation produces a cyclic variation which is often quite stable over time intervals of months or years, depending on the lifetime of the spots. Such observations are at present very active because the derived periods are the direct rotation periods. If the equatorial diameters of the stars can be determined, the equatorial speed can be computed.

Recent surveys in old open clusters have permitted the discovery of relatively large numbers of contact binaries (binary stars so close that they reach contact) all along the main sequence and even among the stars in position of blue stragglers.

Blue stragglers

BLUE STRAGGLERS form a population of stars which has been originally discovered in old open clusters. They remain in the color–magnitude diagrams beyond the observed turn-off, i.e. they are brighter and bluer than the normal end of the main sequence. According to the masses deduced from their luminosities, these stars should not be present at all in such a position, because they should have evolved away from the main sequence. Figure 3 (M67) illustrates the positions of these stars in a well-studied old open cluster. A variety of plausible explanations exist, from delayed star formation through binary mass transfer, or coalescence, to internal mixing. Probably, several mechanisms are needed to explain the variety of properties observed.

Space observations

The observation of open clusters with instruments on board satellites requires great sensitivity and high spatial resolution to resolve the sources and cross-identify them

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with cluster stars. Not all major satellites were able to be efficient in this respect.

HIPPARCOS

The HIPPARCOS satellite was not designed to do imaging and could observe only four stars per square degree. Therefore the main results were obtained on nearby open clusters which cover large areas of the sky. Hipparcos provides fundamental estimates of the cluster distances because the observations are independent of any *a priori* astrophysical assumption on the star structure or chemical composition. This is very important because the cosmic-distance scale, through the determination of the zero point of the CEPHEID PERIOD–LUMINOSITY RELATION, is mostly based on the distance of open clusters.

The best observed cluster was the Hyades and Hipparcos' trigonometric parallaxes and absolute proper motions permitted the distance of the cluster center to be determined with an unprecedented accuracy: 46.34±0.27 pc, i.e. 151 ly. The precision of the individual parallaxes also allowed a three-dimensional view of the cluster to be built. Because of its proximity, the Hyades is the only cluster where the three-dimensional distribution of stars can be observed.

Good distances have also been determined for nine further open clusters closer to the Sun than 300 pc, including the Pleiades, Praesepe and Coma Berenices clusters. The mean cluster parallaxes have not been obtained from the individual parallaxes, but a global solution, based on all measurements of all member stars in the cluster field, has been computed from the great-circle intermediate data. Results are in agreement with what was expected, except for the Pleiades, for which the distance obtained from the Hipparcos satellite is 10% shorter than that determined from the ground by classical techniques. So far (March 2000), the reason for this discrepancy has not been identified: some authors incriminate the satellite itself (but why in the Pleiades only?), others question the properties of Pleiades stars, especially the adopted chemical composition. As many as 2158 transits for 54 stars in the Pleiades were used to compute the final solution.

IUE

The International Ultraviolet Explorer (IUE) satellite worked in a star by star mode and many observations were targeted at massive O-type stars in young open clusters, because they mostly emit ultraviolet radiation. The main goals were the study of the ultraviolet spectra of young stars of known ages but also the properties of the interstellar matter and especially the clouds from which the stars have been formed. The ultraviolet observations from the IUE satellite contributed to the determination of the absorption law in several clusters still associated with reflection nebulae, such as NGC 2244, NGC 2264 and NGC 6530.

ROSAT

Results from the Einstein x-ray satellite had shown that x-ray emission is a common property of solar-type stars and that most main sequence dwarf stars in the Hyades do emit x-rays. This result prompted the observation of several open clusters with the ROSAT x-ray observatory, covering an age range from a few million to more than a billion years. The resulting trend is very clear: the intensity of the x-ray emission is lower in older clusters. This can be understood because the x-ray emission depends on the star's activity, which is itself related to the rotation rate and age. Studies of stellar rotation have shown that the rotation decreases with increasing age. The angular-momentum loss results from the interaction of the stellar winds and magnetic field.

In clusters for which observations have been obtained over a long time, variability in x-rays has been observed. Flaring activity explains the large flux increase, but the data are not complete enough to provide information on the variations of x-ray activity during the whole stellar cycles.

One interesting by-product of the ROSAT x-ray imaging is the discovery of faint member stars in nearby clusters, through their x-ray fluxes, as for example in the young southern open clusters IC 2391 and IC 2602. These two clusters are spread over large areas and photometric and astrometric surveys did not really succeed in identifying faint members among the several ten of thousands of stars measured on Schmidt plates (over such areas, CCDs are completely useless). However, because of the youth of these clusters, faint members were easily detected thanks to their x-ray emission. Ground-based follow-up observations confirmed the membership of most candidate stars.

The Hubble Space Telescope

The HUBBLE SPACE TELESCOPE (HST) produced beautiful images of very young open clusters still associated with nebulosity and of dense clusters which cannot be completely resolved with ground-based instruments. Two objects were more intensely observed: NGC 3603 and the Orion nebula cluster.

NGC 3603 is a dense, rich and very young open cluster which is known as the densest concentration of early-type massive stars in our Galaxy, at a distance of about 7 kpc. It is in the core of the most massive visible giant H II region and is considered as a starburst and compared with the R136 cluster in the Tarantula nebula in the Large Magellanic Cloud. JHK photometry and spectral classification of the 14 brighter stars have revealed the presence of three Wolf–Rayet (WN) stars and a dozen very massive O-type stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Orion nebula cluster, also known as the Trapezium cluster, is one of the most studied star-forming regions. The color–magnitude diagram based on the HST near-infrared photometry shows that the star formation appears to be remarkably coeval: 80% of the stars have an inferred age less than 1 million years (Myr) and 15% have an age between 1 and 4 Myr. The age spread is therefore only a few million years. Several new binary stars were found with separation below 1 arcsec, a domain which is not easily accessible from ground-based imaging. However, the most impressive result is the direct discovery of accretion disks around pre-main-sequence stars which showed up very clearly on HST images as dark DUSTY CIRCUMSTELLAR DISKS. The exceptional resolution of the HST images has also permitted the discovery of a phenomenon which has been called proplyds: when the low-mass stars are close enough to the cluster center and hence subjected to the strong winds emitted by the massive stars of the Trapezium, the material on the side of the disk facing the massive stars is heated by the hot radiation.

Prospects

The number of faint stars in open clusters has, so far, been the main difficulty in the systematic collection of spectroscopic information for the low-mass stars on the lower main sequence. The advent of CCD cameras solved this problem and photometry is now currently done to very faint magnitudes, although the measurement precision is degraded for the fainter objects. The development of large (8–10 m) telescopes (Keck, VLT, Gemini, Subaru) and of multi-object spectrographs opens completely new areas for the observations of open clusters. For example, better resolution at the center of dense or remote clusters will be possible. Infrared capabilities of bi-dimensional detectors on large telescopes will produce a wealth of new observations on embedded clusters and star-forming regions that should provide a much better understanding of the process of star formation. With multi-object spectrographs, astronomers will be able to obtain spectra of about 100 stars at the same time, greatly reducing the need for observing time, with an increased precision. These new facilities will greatly change our way of thinking and of working. We can therefore expect to have answers to the many basic questions, and among them the initial mass function, star multiplicity and chemical composition, simply because of the immense possibilities offered by the new telescopes and instruments, as well as future satellites.

One potentially interesting domain, not yet really explored, is the study of stellar cycles, analogous to the solar cycle, but for many stars along the main sequence of open clusters differing in ages and chemical compositions. The required precision lies within the capability of modern instruments. Such a program would be a long-term project best adapted for institutes which have their own telescope and are not subject to regular telescope-time request. The ouput of such long-term systematic observations will undoubtedly provide many astrophysically fundamental results.

Database

Published data for stars in open clusters have been collected in a database developed at the Institute of Astronomy of the University of Lausanne by the author. It contains, in spring 2000, information for more than 350 000 stars in about 600 open clusters covering the domains of astrometry, photometry, spectroscopy and bibliography. The database is publicly accessible on the world wide web, at the URL http://obswww.unige.ch/webda/. A powerful feature is the possibility of getting basic data for stars from scanned maps or plotted charts. It contains also a large archive of original data files.

Bibliography

Friel E 1995 The old open clusters of the Milky Way *Ann. Rev. Astron. Astrophys.* **33** 381–414

Gilmore G and Howell D (ed) 1998 *The Stellar Initial Mass Function* (*Astronomical Society of the Pacific Conference Series, vol 142*)

Janes K A (ed) 1991 *The Formation and Evolution of Star Clusters* (*Astronomical Society of the Pacific Conference Series, vol 13*)

Micela G, Pallavicini R and Sciortino S (eds) 1997 Cool stars in clusters and associations: magnetic activity and age indicators *Memorie della Società astronomica Italiane* **68** (4) (Firenze)

Payne-Gaposchkin C 1979 *Stars and Clusters* (Cambridge, MA: Harvard University Press)

Jean-Claude Mermilliod

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

Galactic Plane

The plane of the disc of our Galaxy. It follows a line running more or less through the center of the Milky Way, and is inclined to the celestial equator by an angle of about 62°. In the current system of galactic coordinates defined by the IAU in 1959, the galactic plane is defined by $b^{\mathrm{II}} = 0^{\circ}$ and is the central line of the distribution, as determined by radioastronomers, of neutral hydrogen in the Milky Way.

See also: galactic latitude, galactic longitude, galactic year.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Thick Disk

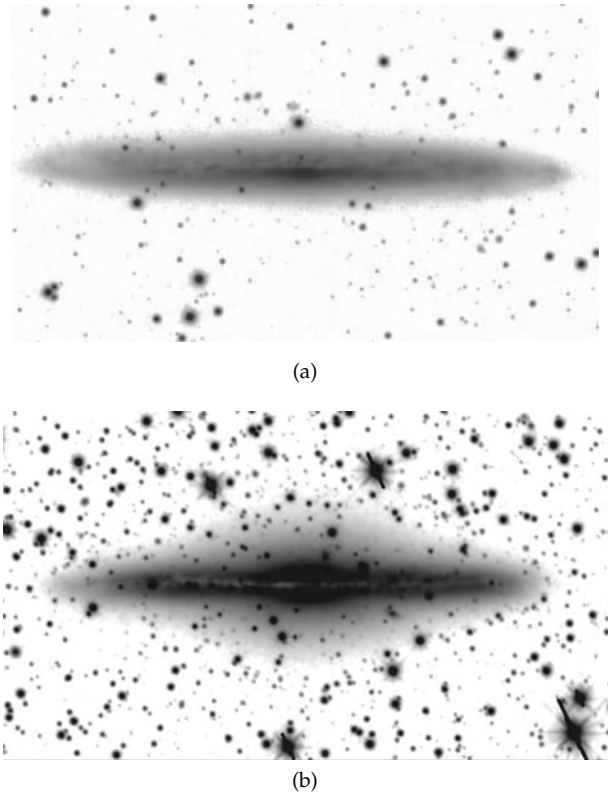
The MILKY WAY (the Galaxy) possesses a number of different stellar components—often referred to as populations, following the pioneering work of WALTER BAADE in the 1940s. Each of these components is described by a fairly well-defined set of observed distributions of kinematics, chemical abundance and age and has a distinctive spatial distribution. A comparison of these populations with the predictions of galactic formation theories and paradigms leads to insight into the manner in which the Galaxy formed and has evolved. Following Baade it was soon appreciated that there existed three critically important populations—the halo, the disk and the bulge, which have masses (1–3) $\times$ $10^9 M_\odot$, $6 \times 10^{10} M_\odot$ and $10^{10} M_\odot$, respectively[1]. (As the names suggest, the halo has a spheroidal distribution, the disk is a highly flattened rotating system, and the bulge comprises a well-defined central concentration. We note also that the disk and halo are often referred to as POPULATION I and II, respectively.) It also became evident that intermediate components exist, and the thick disk is one of these. It is intermediate in its properties between those of the halo and the disk (now renamed the thin disk; see GALACTIC THIN DISK). Insofar as the halo is old (age ~12–15 Gyr) and believed to pre-date the thin disk (age $\leq$10 Gyr), interest in the thick disk is driven by the working hypothesis that an understanding of it will lead to a clearer picture of the early evolution of the Galaxy, giving insight perhaps into the transition phase from halo to disk, or alternatively into phenomena which thickened an already existing thin disk.

The present discussion will proceed as follows. In the next section we shall trace the historical recognition of the Galaxy's thick disk, after which a description will be given of our current understanding of its properties. A discussion is then given of the main competing ideas which have been put forward to explain the various observations, followed by a brief confrontation of the observations with these concepts. It will be clear that at the time of writing (1999) our understanding is incomplete. We conclude with a discussion of prospects for resolving the problem.

Historical development

As noted above, the need for a component intermediate between the basic disk and halo populations soon became evident. Of particular interest in the present context was the designation, at an important Vatican Conference in 1958, of intermediate population II, although its implications for Galactic formation and evolution remained unclear. This component would later become recognized as the thick disk.

An important further clue came from studies of external DISK GALAXIES where it became clear that edge-on systems showed a variety of disk thicknesses. The upper panel of figure 1 shows the brightness distribution for the edge-on disk galaxy NGC 5907, from the recent work of Morrison and her colleagues, where the density distribution perpendicular to the disk can be well represented by an exponential, decreasing as $\exp(-z/z_0)$, where z is the vertical distance from the plane and z_0 is called the scale height, and in this object has the value 400 pc. Observations of other edge-on systems, however, revealed examples of a more complicated disk structure. Perhaps the most famous of these is NGC 891, which is shown in the lower panel of figure 1. This could be well explained in terms of *two* disk components having very different exponential scale heights—and astronomers were led to think, fairly naturally, in terms of 'thin' and 'thick' components. In NGC 891 the thinner component has a vertical exponential scale height of 400–650 pc and is not unlike the situation one finds in the Galaxy. The second, thicker, component has an exponential scale height of 1500–2500 pc. (For reference in figure 1 the disk of NGC



(a)

(b)

Figure 1. (Upper panel) The distribution of red light in the edge-on disk galaxy NGC 5907, which can be well represented in the vertical direction by a single exponential component. (Lower panel) The red light distribution in NGC 891. Here two exponential components ('thick' and 'thin') are required to explain the observations. The case has been made in the astronomical literature that the Milky Way seen edge on would look rather like NGC 891. (This diagram originates from the work of H L Morrison and her colleagues, and was kindly supplied by them.)

[1] The reader should distinguish the luminous halo, discussed here, from the 'dark' unseen halo, which is inferred from dynamical arguments, and contains ~50–90% of the Galaxy's mass.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

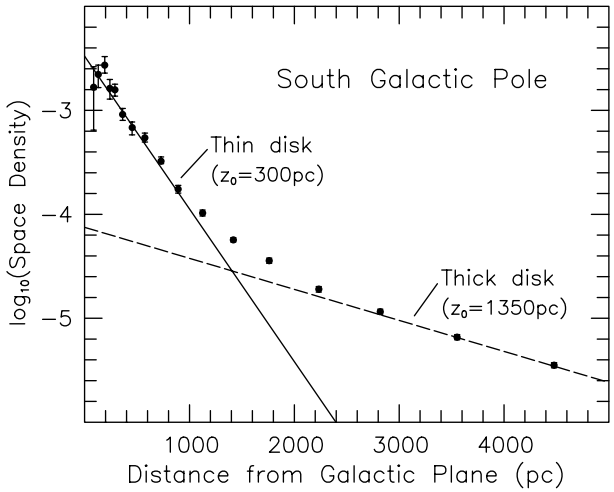


Figure 2. The logarithm of the space density of stars towards the south Galactic pole, as a function of distance from the Galactic plane, from the work of Gilmore and Reid. An exponential falloff of a single component would be represented by a straight line in this diagram and would clearly not be an adequate representation of the data. The authors fitted a two-component model, which they identify as the thin disk (scale height 300 pc) and thick disk (scale height 1350 pc).

891 extends some 35 kpc from end to end.) It is this second component, first given the name 'thick disk' in external systems, which is under discussion here for the Milky Way.

In the Galactic context, the term 'thick disk' was introduced by Gilmore and Reid in 1983 to explain their observation that star counts towards the south galactic pole were not well fitted by a single exponential, but rather could be nicely represented by two such components—one with scale height 300 pc and the other with scale height 1350 pc. The ratio of the density of the two components at the GALACTIC PLANE was 0.02. A representation of their seminal diagram is shown in figure 2. It is important to appreciate that the normalization of thick/thin $= 0.02$ at the Galactic plane is an order of magnitude larger than that of the halo population, which in the solar neighborhood is halo/thin ~ 0.001. That is, for every 1000 stars chosen at random in the solar neighborhood, one expects only one to be a member of the halo population, and 20 to belong to the thick disk. While the thick disk may be a minority component, in the solar neighborhood it far outweighs the halo.

For those who like to follow the impact of new instrumentation on the subject, it is interesting to note that this new development followed the implementation of large (wide-field) SCHMIDT TELESCOPES and powerful measuring machines necessary for the analysis of the large photographic plates on each of which huge numbers (hundreds of thousands) of stellar images were recorded. In a similar vein, historians of science will note that the 'thick disk' of Gilmore and Reid was also present in the reports of star counts towards the north Galactic pole by Yoshii in 1982.

Figure 2 deserves closer scrutiny. It is clear that two exponential components, 'thick' and 'thin' (shown by the straight lines), provide a sufficient explanation of the observations. However, are they a necessary requirement? It is trivial to demonstrate that a multicomponent or continuous distribution, based on existing observational data, can also be presented which very nicely fits the observations in figure 2. The reader should be aware that this is more than a semantic point. If, on the one hand, one accepts that the data are best represented by two components, 'thick' and 'thin', one will be led naturally to seek an explanation in terms of discrete phenomena. Indeed, it is probably fair to suggest that most workers have come to the view that the Galactic thick disk has resulted from the merger of a small satellite with the Milky Way. If, on the other hand, one accepts that the observed behavior in figure 2 represents a continuous, extended disk configuration, one will seek alternative explanations, involving perhaps dissipative collapse scenarios, multiple mergers or clumpy galactic collapse formation. What needs to be done, and which has not been systematically done to this point, is to compare the extensive and growing observational material with the predictions of both types of models—discrete and continuous.

Properties of the thick disk

Initial estimates

Work by Gilmore and his coworkers in the early 1980s led to the suggestion that the thick disk had abundance and kinematic characteristics quite intermediate between those of the halo and thin disk populations. At that time the mean chemical abundances, [Fe/H][2], of the halo and thin disk were -1.6 and -0.2, respectively: that of the thick disk was -0.6. The stellar velocity dispersions in the direction perpendicular to the Galactic plane were $\sim$90 and $\sim$15 km s^{-1} for halo and thin disk: for the thick disk the value was 60 km s^{-1}. Also, for the asymmetric drift—the velocity by which material lags behind the solar neighborhood in its rotation about the Galaxy—the halo and thin disk had values $\sim$220 and $\sim$15 km s^{-1}, respectively: that of the thick disk was 100 km s^{-1}. This laid the ground for the suggestion that the thick disk was a discrete component, with origins quite different from those of the other populations. Later investigations would support the abundance result but lead to more moderate kinematics, closer to that of the thin disk, with vertical velocity dispersion 40 km s^{-1}, and asymmetric drift $\sim$40 km s^{-1}.

Recent determination of thick disk parameterization

Current estimates of the properties of the thick disk and the other Galactic disk and halo populations are presented in table 1. The situation has become much more complicated in the past two decades not only for the disk but also for the halo, leading to greater difficulty in understanding the manner in which the Galaxy formed and has evolved.

[2] $[\mathrm{Fe/H}] = \log_{10}[N(\mathrm{Fe})/N(\mathrm{H})] - \log_{10}[N(\mathrm{Fe})/N(\mathrm{H})]_{\odot}$, where N(Fe) and N(H) are the number densities of iron and hydrogen, respectively.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Some population characteristics of disk and halo components in the solar neighborhood.

Component	Scale height (pc)	⟨[Fe/H]⟩	σ_U, σ_V, σ_W [a] (km s^{-1})	V_{lag} [a] (km s^{-1})	Age (Gyr)	ρ/ρ_{tot} [b]
Old thin disk	300	−0.3	30, 20, 15	15	≤10	0.95–0.98
Thick disk	800–1500	−0.6	65, 55, 40	40	12–15	0.02–0.05
Metal-weak thick disk	1400:	−1.2:	Unknown	40	(12–15):	(0.0005–0.002):
Flattened halo (also called old, low or collapsed halo)	1600–2000	−1.6: [c]	130:, 100:, 90: [c]	160	12:–15	0.0008:
Spherical halo (also called younger, high or accreted halo)	Spherical	−1.6: [c]	130:, 100:, 90: [c]	270	12:	0.0002:

[a] σ_U, σ_V and σ_W are velocity dispersions in the directions away from the Galactic center, toward Galactic rotation and toward the north Galactic pole, respectively. V_{lag} ($=V_{solar\ nbd.} - V$) measures the asymmetric drift, the velocity by which the component lags the solar neighborhood in its systemic rotation.

[b] Ratio of the density of the component to total density in the solar neighborhood. We assume $\rho_{halo}/\rho_{disk} = 0.001$.

[c] Decomposition of the two halo components has not yet been achieved. The tabulated values are those determined for their admixture in the solar neighborhood. The values of the individual components are thus uncertain.

Substructure in the halo

To understand the disk one must also appreciate the situation with the halo. Although it comprises only a minuscule fraction of the mass of the Galaxy, it contains the oldest material and yields crucial clues to the earliest phases of the system. Current indications suggest that the halo is a multiple population, with components which originated during both an initial dissipative collapse (as described in 1962 in the seminal work of Eggen, Lynden-Bell and Sandage) and the later accretion of smaller fragments which had evolved independently of the central concentration, and later came into dynamical equilibrium with it (as advocated in the equally important work of Searle and Zinn in 1978). The recent discovery of the SAGITTARIUS DWARF SPHEROIDAL GALAXY, currently in catastrophic collision with and being destroyed by the Milky Way, attests indisputably to the operation at some level of the merger scenario. In table 1, the halo is represented by two components, described by different workers by various descriptors. This is no doubt a gross oversimplification, and current efforts are being directed to ascertaining how much discreteness remains evident in the accreted component.

Substructure in the thick disk?

The situation with the thick disk has also become more complicated. In table 1 one sees what one might call the 'canonical' thick disk first appreciated by Gilmore and Reid, with ⟨[Fe/H]⟩ ∼ −0.6, and kinematics much closer to that of the old thin disk than originally claimed. Observational evidence now also exists to support the case for the existence of material with disklike kinematics and [Fe/H] <–1.0—called by some the 'metal-weak thick disk'. The meager data available for this component are also presented in table 1. The case for the population remains to some extent controversial, but evidence continues to grow that some ∼10–30% of material with −1.6 < [Fe/H] < −1.0 has disklike kinematics. There are even GLOBULAR CLUSTERS in this abundance range (old, populous, star clusters previously believed surely to be members of the halo) with recent well-determined space velocities which are characteristic of the disk rather than the halo.

The question which begs to be answered, if the metal-weak thick disk is real, is whether it is yet another discrete component of the disk created by merger or part of a more continuous structure, formed perhaps in a dissipative manner. Another tantalizing question provoked by table 1 is the significance of the similarity in many respects of the metal-weak thick disk and the flattened halo component.

Gradients in the thick disk?

A critical consideration for an understanding of the thick disk is the existence or otherwise of abundance and kinematic gradients within it. Since most of the available data exist for material not too far from the solar neighborhood, and the thin disk far outweighs the thick disk close to the Galactic plane, one must always be concerned that leakage of the thin disk, due to observational errors, into putative thick disk samples will simulate the effect of vertical gradients. That said, most workers advocate that no vertical abundance gradient exists, or if it does it is small and currently at the limit of detection. Future insight into this question may come from analysis of external edge-on systems. Indeed, van der Kruit and Searle, in an early analysis of NGC 891 (see figure 1, lower panel), report a gradient in the photometric properties of the thick disk component of that system which is most easily interpreted as one in abundance.

Concerning kinematics, and in particular the asymmetric drift, there are two schools of thought. The first contends that the drift is a function of height above the Galactic plane (and is given by $V_{lag} \sim 18z + 15$ km s^{-1}, where z is the vertical distance from the plane in kpc), while the second finds no such dependence and reports that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

drift is constant with height, with $V_{\rm lag} = 40$ km s^{-1}, at least within ~3000 pc of the plane.

Little information exists on the question of the dependence of kinematics on abundance in the thick disk population(s) or on gradients as a function of distance from the Galactic center. These are questions for the future.

To a large extent, resolution of the nature of the thick disk and its implications for Galactic disk cosmogony will turn on the issue of these interdependences of position, abundance and kinematics.

Age of the thick disk

The thick disk is old. Indeed, it contains the relatively metal-rich, so-called disk globular clusters, such as the brilliant 47 Tucanae, the age of which lies in the vicinity of 12 Gyr. Studies of thick disk field stars suggest that they too have ages similar to that of 47 Tuc. It is important to appreciate, however, that these estimates are accurate only to ~2–3 Gyr. With this in mind then, one should operate with care in seeking to answer the important cosmogonical question of whether the thick disk has the *same* age as the Galactic halo, for which estimates lie in the range 12–15 Gyr. Some workers advocate that the thick disk is younger than the halo by ~2–4 Gyr. In the opinion of this reviewer, at least, the issue remains unresolved at the ~2–3 Gyr level.

Models of thick disk formation

Numerous mechanisms have been suggested for the thickening of galactic disks. Foremost among others, these include internal collisional heating of the thin disk, merger with external systems and dissipative collapse. The different mechanisms make different predictions concerning the abundance, kinematic and age profiles of the thick disk and, in principle, comparison between observation and theory should enable the relevant process(es) to be identified. For example, the dissipative models predict abundance and kinematic gradients, while the minor merger ones appear not to do so. One might also not be surprised to find discontinuities between those properties of the thick and thin disks if the merger scenario is relevant. A slow dissipative formation mechanism might lead to a range in ages in the thick disk, intermediate between that of the halo and the old thin disk, while the age profile in merger scenarios will be driven by the conditions which existed at the time of the merger. Collisional heating of the thin disk will produce a thick component with characteristics which are closely related to those of the thin disk, modulo the manner in which the heating occurred.

One should always bear in mind that not one, but rather a combination of mechanisms may have contributed to the disk thickening. We briefly discuss the three mechanisms.

Internal disk heating

The disk of a galaxy will be heated by the scattering of its constituents by giant molecular clouds and transient density waves associated with spiral arms, and this is believed to be the origin of an observed dependence of velocity dispersion on age in thin disk material and to be responsible for vertical velocity dispersions less than 20–30 km s^{-1}. Only if one is prepared to accept the existence of massive objects such as black holes or clusters of Jupiters, of mass $>10^6 M_\odot$, is one able to produce velocity dispersions similar to those found in the thick disk.

Mergers

Computer modeling of the mergers of minor satellites with disk galaxies show very nicely how disk thickening might occur, given favorable relative orbits (prograde approaches strongly preferred) and satellite density profiles (dense satellites preferred). Studies, such as those of Quinn, Hernquist and Fullagar, show how effectively this mechanism can operate. Some of their results are shown in figure 3, where the satellite has 0.1 times the mass of the disk. One sees that over a relatively short time of a few Gyr the satellite is shredded and becomes part of the disk which itself has been thickened. The model makes important predictions among which is the finding that there will be essentially no change in the asymmetric drift of the thickened disk from its initial value. No vertical abundance gradient is produced in this type of model, although an existing one would be preserved. A caveat of the mechanism is that its success hinges on the density profile of the incoming satellite, and it has been demonstrated by Huang and Carlberg that mergers with low-density satellites, of the kind one observes in the local neighborhood now, do not produce disk thickening, but rather lead to disk tilting as they are assimilated into the larger system.

Since the thick disk is old, indeed as old as the halo of the Galaxy to within the observational uncertainty of 2–3 Gyr, the model requires that the most recent merger of this kind occurred $\gtrsim 10$ Gyr ago.

It is important to recall that the Searle–Zinn formation scenario for the halo of the Galaxy invokes the need for the merger of fragments in the forming system. Given that the age of the thick disk is commensurate with that of the halo, one should not be surprised to find that assimilation of fragments (not necessarily as large as those discussed in the above simulations) played a role in determining the thickness of the disk.

Dissipative collapse formation

In 1976, Larson modeled the dissipative collapse of a rotating system and produced galaxies with thick disks having an abundance gradient. Note that his work actually pre-dates the identification of this component in the Milky Way. A more recent variation of the model, which gives insight into the manner in which the thickening may have occurred, comes from the work of Noguchi, who simulated the collapse of clumpy star-forming galaxies in an effort to explain the peculiar morphology of high-redshift systems. The evolution of a typical galaxy is shown in figure 4. Beginning with a rotating ensemble of gaseous and dark matter particles, a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

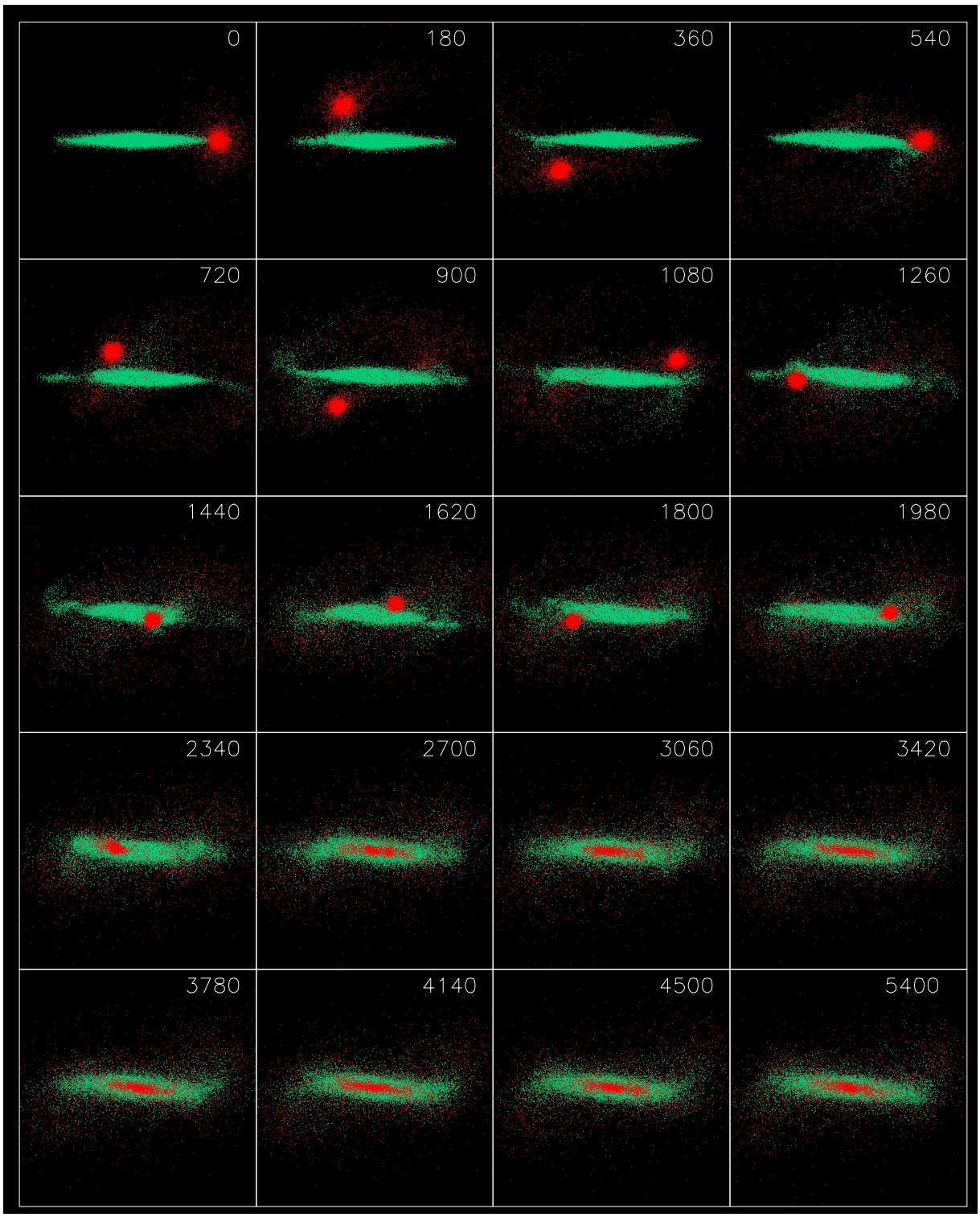


Figure 3. The simulation of a satellite–disk merger from the work of Quinn, Hernquist and Fullagar. The satellite has 10% the mass of the disk, and approaches on a prograde orbit inclined by 30° with respect to the disk plane. The number in the top right of each panel is elapsed time in units of 10^6 yr. (This diagram was kindly provided by P J Quinn, L Hernquist and D Fullagar.) **This figure is reproduced as Color Plate 25.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

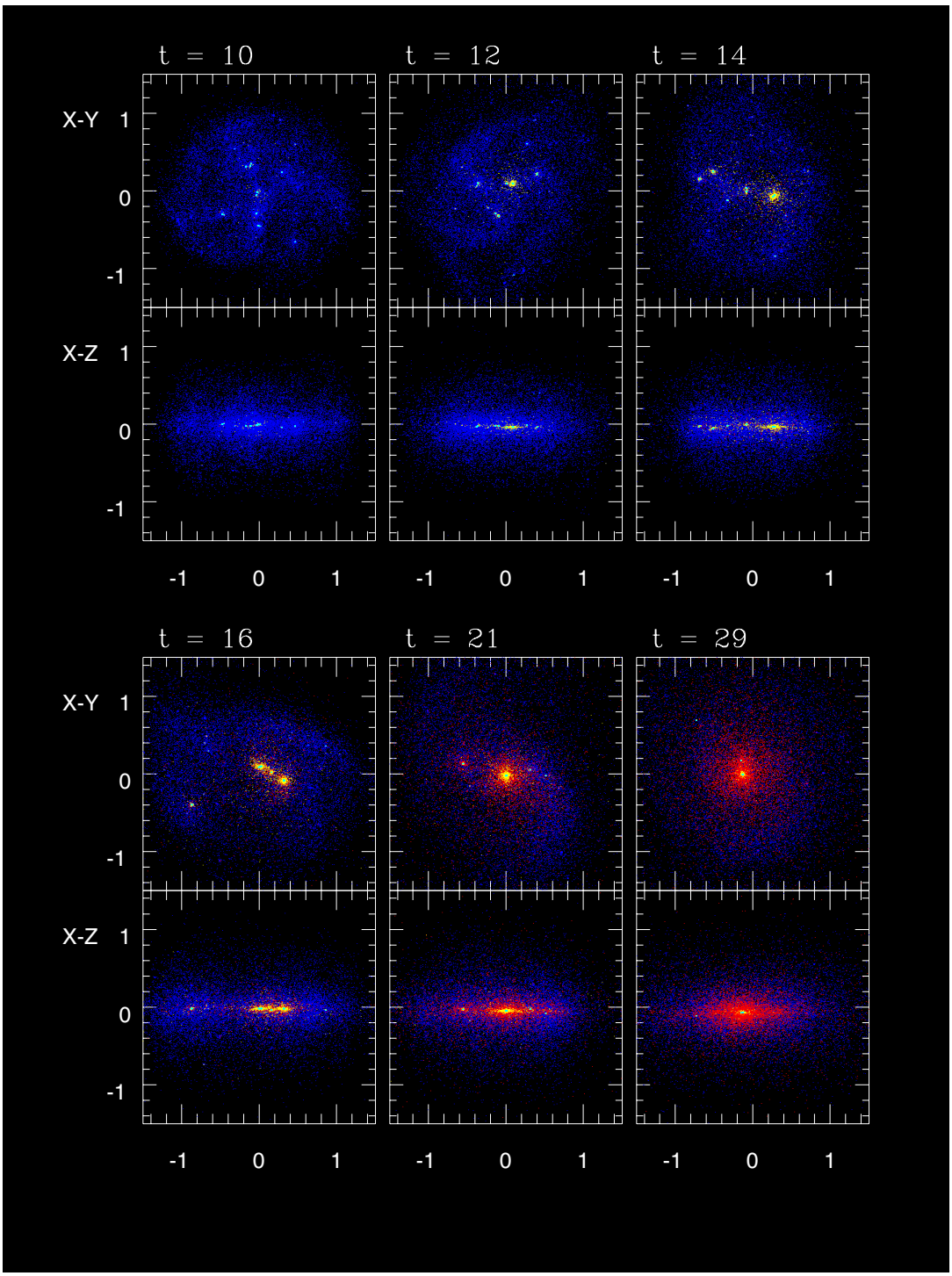


Figure 4. The simulation of the collapse of a rotating proto-galaxy by Noguchi. In all panels the abscissa is the distance along the x-axis of the simulation and represents the distance from galactic center. Panels with y-axis labeled X–Y show the simulation projected onto the disk, while those labeled X–Z view the simulation edge on. The panels are labeled with time, where the unit is 7×10^7 yr. The stellar particles are color coded: blue is used for objects younger than 10^8 yr, yellow between this and 5×10^8 yr and red for older ones. (This diagram was kindly provided by M Noguchi.) **This figure is reproduced as Color Plate 26.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

disk system forms on a timescale of 1 Gyr. Its evolution is critically determined by the formation of distinctive clumps of mass $10^9 M_\odot$. Dynamical friction causes these to spiral towards the galactic center, creating a bulge, while during the process stars and gas scatter off the clumps, thickening the disk. The scale height of the old stellar component in the model was 1.9 kpc, a little larger than that observed in the Milky Way, but not unlike the value observed for the thick disk of other galaxies (see also GALAXY EVOLUTION).

It will not have escaped the reader that we may have just entered the region of semantics. Perhaps the clumps in this formalism and the small merger fragments discussed at the end of the previous section are not unrelated. Indeed, the relative importance of the accretion and dissipative collapse mechanisms may turn on the spectrum of structure sizes in the vicinity of a forming galaxy.

Comparing the observations with model scenarios

What then may one conclude concerning the nature of the thick disk and what it has to tell us about the manner in which the disk formed? It is probably not unreasonable to conclude that most workers are inclined to the view that it resulted from a minor merger some ∼10–12 Gyr ago. One can certainly reproduce most of the observations in this manner, modulo, for example, the final word on the existence or otherwise (and their interpretation) of gradients, which are not predicted in this model. One might also be concerned by the necessity of mergers of the disk with 'hard' as opposed to 'soft' satellites noted in the previous section.

Another challenge for this scenario is to explain the nature of the so-called 'metal-weak thick disk'. Is it also discrete, and also produced by a merger? The critic might also well ask: is discreteness of the thick disk population(s) necessarily required by the observations, or are they really part of a continuous disk configuration? One will find in the bibliography below the challenge of Nemec and Nemec to all who would make claims on this matter: 'In most models of the Galaxy the number of components is based on preconceived ideas or on a subjective assessment of the observational data ... More objective ... procedures for estimating the number of components and goodness-of-fits tests ... for comparing competing models are available.'

Problems also abound for those who would advocate the alternative world view of a dissipative origin of the thick disk. While models of dissipative Galactic collapse exist in the astronomical literature, too few details are available with which one may confront the observations, and no detailed comparison is possible at this time.

Future prospects

The most limiting factor in efforts to understand the population structure of the Galaxy has been our necessary reliance on solar neighborhood samples within a few hundred parsecs of the Sun, where the important halo and thick disk stars represent only very small fractions, ∼0.001 and ∼0.02–0.05, respectively, of field stars. Ideally, one would like to observe samples well away from the Galactic plane, at distances of a few kpc, where the contaminating influence of the thin disk will be minimized. Fortunately, with the current advent of large multifiber spectrographs on 4 m class (and larger) telescopes which will permit the simultaneous observation of several hundred objects at these distances, this is now becoming possible. In principle, one may now obtain samples of several thousand stars (near main sequence turnoff dwarfs) in various critical directions (towards, for example, the Galactic poles, center and anticenter, and in directions of Galactic rotation) with a view to separating out the populations discussed above and to more strongly constraining the reality or otherwise of gradients, dependences of kinematics on abundance and substructure within components.

Because of their different characteristics, the halo and disk will give different signatures in different directions in the Galaxy, and one will be able to disentangle thick disk and halo several kpc from the Galactic plane and to look for the distinctive features of dissipation and merger events. One will be able to determine whether there is any evidence for vertical and radial gradients in abundance and kinematics, together with the density and (systemic) rotational properties of the various subcomponents.

When these data are available one will be in a sound position to confront the results of the various models of the manner in which the thick disk formed.

Bibliography

Croswell K 1995 *The Alchemy of the Heavens: Searching for Meaning in the Milky Way* (New York: Anchor)

Gilmore G, Wyse R F G and Kuijken K 1989 Kinematics, chemistry, and structure of the Galaxy *Ann. Rev. Astron. Astrophys.* **27** 555–627

Majewski S R 1993 Galactic structure surveys and the evolution of the Milky Way *Ann. Rev. Astron. Astrophys.* **31** 575–638

Nemec J M and Nemec A F L 1991 Mixture models for studying stellar populations. I—Univariate mixture models, parameter estimation, and the number of discrete population components *Publ. Astron. Soc. Pac.* **103** 95–121

Norris J E 1996 The relationship between the halo and disk populations *Formation of the Galactic Halo ... Inside and Out* ed H Morrison and A Sarajedini (San Francisco, CA: ASP) pp 14–23

John E Norris

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Thin Disk

Of the components of our MILKY WAY GALAXY, the thin disk is the most prominent part to our eyes. It manifests itself as the band of faint light that we see encircling the whole sky. Except for the bulge in the direction of the center of our Galaxy, the stars that make up the Milky Way as we see it are part of the thin disk, just as our Sun is part of that same population of stars. A remarkable fact is that when we look at the Milky Way from the southern hemisphere in winter—when the center of the Galaxy is almost at the zenith at midnight—the appearance is strikingly similar to somewhat later Hubble class, edge-on galaxies such as NGC 891 (of Hubble type Sb).

The young disk

As we will see below, the thin disk is made up of stars of a large range of ages. The youngest have only just formed out of the interstellar medium; the gas and dust between the stars out of which stars are being formed is generally considered to be part of the young disk. The gas can be mapped easily using the 21 cm line of neutral hydrogen (H I) and this has been done extensively for our Galaxy since RADIOASTRONOMY started to be practised in a systematic manner in the 1950s. It took until the 1970s, with the advent of aperture synthesis radio telescopes, until it could be done with sufficient angular resolution in external galaxies. The neutral hydrogen extends usually much further than the stellar disk, often exhibiting a deviation from the plane of the inner galaxy. This can be seen in edge-on galaxies as a curving upward on one side and a symmetric curving downward at the opposite. In moderately inclined galaxies there are kinematic signatures of these so-called warps.

The measurements of the neutral gas have to be supplemented by observations of the distribution of the molecular gas, most abundant in molecular hydrogen H_2. This, however, cannot be observed directly and its distribution is inferred from radio measurements of spectral lines of carbon monoxide, CO. This is transferred into distributions of H_2 by a still rather dubious 'X factor' which descibes the ratio of H_2 versus CO. This X factor is predicted on theoretical grounds and the dubious assumption is made that it is the same from galaxy to galaxy and with galactocentric radius in spite of known variations in heavy-element abundance and radiation field (the latter being important for photodissociation (see PHOTODISSOCIATION REGIONS)). The molecular gas is found to be concentrated to a large extent in giant molecular clouds (GMCs) that contain masses up to $10^6 M_\odot$ (see INTERSTELLAR MOLECULAR CLOUDS). In these GMCs star formation appears to be occurring.

The interstellar gas then is distributed in a very thin layer—much thinner than the older disk stars—with a thickness of order 100 pc. The molecular gas is more concentrated toward the galaxy's center. The velocity distribution at any position in the disk is largely isotropic with (one-dimensional) velocity dispersions of order 7 km s^{-1}.

STAR FORMATION occurs in groups. This is because for the conditions in the interstellar medium a gas cloud can only be unstable with respect to its own gravity if its mass exceeds about 10^5 or so $M_\odot$. This so-called Jeans' mass can be derived crudely as follows. Consider a spherical region of radius R and density ρ, or mass $M = \frac{4}{3}\pi R^3 \rho$. The gravitational (potential) energy can be calculated to be

$$\Omega = -\frac{3GM^2}{5R}.$$

The kinetic energy due to the random motions with dispersion $\langle V^2 \rangle^{1/2}$ is roughly

$$T = \tfrac{1}{2} M \langle V^2 \rangle.$$

The virial theorem then tells us that this region will not contract under its own gravity as long as $\Omega + 2T$ is positive, or in other words as long as there is sufficient kinetic energy to overcome the gravitational force. This then translates into a maximum radius for stability that is called the Jeans' length,

$$R_{\rm Jeans} = \left(\frac{5\langle V^2 \rangle}{4\pi G \rho} \right)^{1/2}$$

and the mass contained within this radius is the Jeans' mass.

In the interstellar medium near the Sun the gas density is of order 1 H atom cm^3 and the velocity dispersion is about 7 km^{-1}. Then $R_{\rm Jeans} \sim 400$ pc. This is more than the thickness of the gas layer. A circular area of the disk with this radius contains about $10^6 M_\odot$ in the form of gas.

Clusters of newly formed stars are called associations. There are two types, depending on the kind of stars seen in them. The most prominent are the OB ASSOCIATIONS; their brightest stars are heavy O and B stars and they are usually accompanied by H II REGIONS, where the interstellar gas is ionized by the UV radiation from these hot stars. The second type is called T associations; these are made up of T TAURI STARS, which are stars of solar mass or somewhat less contracting on their way to the main sequence. The existence of these two types of associations is one basis for a proposed concept of bimodal star formation, in which massive and less massive stars are formed independently.

The distribution of OB associations can be studied in much detail. In agreement with the fact that associations are formed out of the interstellar medium recently, they are also found in our Galaxy in a thin layer, comparable in thickness with that of the gas. The local system of OB stars deviates from that of the plane of the Galaxy; it is confined to a narrow, tilted layer; this is called GOULD'S BELT, since it is reflected in the distribution of OB stars on the sky and this is also generally along a great circle inclined to that of the Milky Way. The velocity distribution of OB stars also deviates from being isotropic, although the velocity dispersion is comparable with that of the gas. The long

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

axis of the velocity distribution does not point toward the Galactic center (has a so-called vertex deviation); this is probably the result of local disturbances in the smooth density distribution in the disk due to the spiral structure of the Galaxy.

Open clusters

Stars are being born in groups or clusters, each representing of the order of $10^5 M_\odot$. Most clusters dissolve in the course of time as a result of the gravitational interactions with the surroundings. Irregularities, such as those associated with spiral structure, are believed to play a major role in this, but also condensations in the interstellar medium, such as the GMCs, are important. It is not known which stars belong to the same cluster in which the Sun was one of the stars being born. Finding out about this is almost certainly impossible. Yet, some clusters or at least their more solidly bound central parts survive, in some cases as long as the age of the disk.

So, we have a whole range of clusters of varying ages. Their ages can be derived from their HERTZSPRUNG–RUSSELL DIAGRAMS. The youngest are the associations discussed above, the oldest—such as the cluster NGC 188—have ages up to about 10^{10} yr. Local examples are the Hyades (consisting of relatively bright stars over a fairly large area of sky and having an age of almost 10^9 yr) and the Pleiades (the brightest stars being visible to the naked eye and having an age of about 10^8 yr).

Star clusters are entities that have been able to retain their structure and to form a relatively securely bound group of stars. Clusters that have dissolved by gravitational tidal effects still may survive as recognizable clusterings in phase space. Even if the prominent structure of clusters in coordinate space has disappeared, their signatures in velocity space may survive on a longer timescale. Structure in the distribution of space velocities of common stars in the neighborhood of the Sun does show clear records of the dynamical history of the disk in the form of the so-called moving groups: collections of stars that are isolated in phase space and appear to share other properties such as age and/or abundance of heavier chemical elements.

The old disk

The old disk contains most of the mass in stars in the disks of our and other galaxies. In spite of the prominence of spiral structure in pictures of DISK GALAXIES, the underlying smooth component contains most of the light (and mass). The fact that spiral structure is so evident in atlases of galaxies derives from the fact that photographic print emphasize the gradients (photography essentially being a logarithmic process). Digital surface photometry shows that even the bright H II regions in the SPIRAL ARMS and the dustlanes on the inside of the arms are only relatively minor disturbances on the underlying brightness distribution. In a vertical column through the disk of our Galaxy in the solar neighborhood, the OB stars contribute only a third or so to the integrated surface brightness. For what follows we can take in external galaxies the surface brightness of the disk to be mainly representing the distribution of older disk stars, even at relatively blue optical wavelengths. The light from the older disk stars is of course dominated by their brightests members, i.e. the giant branch stars among them.

From studies of the surface brightness distributions in spiral galaxies—starting with those by de Vaucouleurs in the 1940s and 1950s—it was found that the radial fall-off from the center onwards is exponential. That is to say, the surface brightness can be fitted to a function of the form

$$L(R) = L_0 \exp\left(-\frac{R}{h}\right)$$

where h is called the scale length. The surface brightness $L(R)$ is expressed here in $L_\odot$ pc^{-2}, but in practice is measured in magnitudes per arcsec^2 and indicated by μ. For the majority of the bright spiral galaxies, Freeman in 1970 found that the central surface brightness was remarkably constant at 21.6 B magnitudes arcsec^{-2}. In the meantime, many galaxies with fainter central surface brightnesses have been found, but these are mostly small dwarf galaxies or a minority of so-called low-surface-brightness galaxies (constituting a small fraction of the mass in stars in the universe), but the result stands for the common disk galaxies such as our own.

The cause of this exponential nature of the distribution of stars appears to be the distribution of angular momentum in the gas from which galaxies form. The angular momentum probably results from tidal interactions between protogalaxies in the early universe, when they were much closer together. Irregularities in their mass distribution made them susceptible to gravitational tidal fields from their neighbors, resulting in a spinning-up of these gas clouds. It is commonly accepted that this, when the protogalactic gas clouds detached themselves from their surroundings in the expanding universe, resulted in entities closely resembling a uniform-density, uniformly rotating sphere. That is to say, the more or less spherical density distribution had a roughly uniform density and rotated everywhere with the same angular velocity. If this distribution of angular momentum is conserved during the collapse of the gas (after some of it was taken away as the stars that form population II in the halo) into a disk, then a roughly exponential distribution as observed will follow. This assumes that the gravitational field is mostly determined by that of a dark halo with a density distribution falling off as R^{-2} as implied by the rotation curves of spiral galaxies. The exponential nature of the light (and presumably mass distributions in galaxy disks) seems to reflect the effect of how galaxies acquired their angular momentum in the early universe.

The older disk stars form a system that is clearly thicker than that of the young population (see also GALACTIC THICK DISK). From the common stars in the solar neighborhood this follows both from counts that can be used to derive their spatial distribution as well as from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

their kinematics. The velocity dispersion of stars such as our Sun locally is of the order of 20 km s^{-1} in the direction perpendicular to the Galactic plane. This is much higher than that of the young population and consequently the thickness of the thin disk of older stars must be much larger than that of the gas and very young stars. Since this system of old disk stars is almost certainly effectively self-gravitating (that is to say, these stars themselves provide essentially all of the gravitational force in the direction perpendicular to the plane; the z-direction) it can be expected that the spatial density distribution should approximate that of an isothermal sheet. This is a stratified layer in which the velocity dispersion is constant with distance from the plane.

It is well known that such an isothermal, stratified distribution, where stars move under the collective gravitational force of themselves, can be written as

$$\rho(z) = \rho(0)\,\mathrm{sech}^2\left(\frac{z}{z_0}\right)$$

where the vertical scale parameter z_0 is related to the velocity dispersion $\langle V_z^2\rangle^{1/2}$ by

$$z_0 = \frac{\langle V_z^2\rangle}{\pi G\sigma}$$

with σ the surface density (integrated over a vertical column):

$$\sigma = 2\rho(0)z_0.$$

Comparison with the distribution of stars in the solar neighborhood and their kinematics shows that this is a remarkably good approximation to the actual situation. The inferred vertical density distribution at larger distances from the plane approximates an exponential:

$$\mathrm{sech}^2\left(\frac{z}{z_0}\right) \approx 4\exp\left(-2\frac{z}{z_0}\right).$$

This has indeed been observed in the surface brightness distribution in external disk galaxies when seen edge on. What is more remarkable is that the scale parameter z_0 is to an excellent approximation independent of distance from the center. Using the exponential radial distribution mentioned above, it follows that the space distribution of old disk stars in a spiral galaxy can be described very well by

$$\rho(R,z) = \rho(0,0)\exp\left(-\frac{R}{h}\right)\mathrm{sech}^2\left(\frac{z}{z_0}\right).$$

Dynamical heating

When we look at more detail into the distribution of velocities of common stars in the solar neighborhood we find that their velocity dispersions increase with age. Apparently, stars are being born with the velocity distribution of the interstellar medium from which they form and subsequently gain velocity. Although the interstellar medium appears—as a result of collisions between gas clouds—to be essentially isotropic, the process by which stars on average gain in space motion quickly makes their velocity distribution anisotropic. The gain in velocity is at first rather rapid, but then seems to level off when the age is a few billion years. In the radial direction the gain in velocity is largest, leveling off at about 40 km s^{-1}, while in the vertical direction it reaches only about 20 km s^{-1}. This difference is possible because the motion in the vertical direction is independent of the motion in the Galactic plane. The increase of space motion of disk stars with age is usually referred to as dynamical heating or as secular evolution.

There are two processes that may be responsible for this dynamical heating. The first, proposed by Spitzer and Schwarzschild in the 1950s, is by scattering of stellar orbits by encounters with large mass concentrations in the interstellar medium in the disks; these were later identified with the GMCs mentioned above. This will statistically speed up the motions of the stars, but will quickly become less efficient and the space motions of the stars will level off as observed. This is so because, when the motions increase, stars will spend a larger fraction of the time out of the thin layer where the GMCs reside and the scattering will quickly become less effective.

The second mechanism was proposed in the 1960s by Barbanis and Woltjer and is the scattering of the stars by encounters with spiral arms. These also constitute variations in the gravitational field. This mechanism is less effective in the vertical direction, because the period of vertical motion of stars is very different from that from encountering the spiral structure. In the radial direction there is much more of a resonance between these encounters and radial oscillations and the effect is more efficient. Theoretical studies indicate that the first mechanism would predict the ratio of the vertical to radial velocity dispersions to be larger than observed, so it is likely that (at least in our Galaxy in the neighborhood of the Sun) the second mechanism has been more effective. Still, at various positions in different galaxies the actual dynamical heating is probably accomplished by both processes in varying relative importance.

Random motion of the old disk stars in galaxies is vital for matters of stability. If there were no relative motion of the stars, the disks of old stars would not be able to be as smooth as they are; any statistical density fluctuation would quickly result in gravitational collapse and the disks would break up in many dense concentrations. This is a dilemma that goes back to the days of Isaac Newton: if every star attracts every other one then why would they not have collapsed into a dense concentration? Newton's incorrect answer was that stars are distributed evenly in space, such that every star is pulled equally in all directions by the other stars. We now know that the correct answer is that stability derives from the relative motions of the stars, providing in essence a kind of pressure that prevents these gravitational instabilities from occurring.

Now return to galactic disks. We have seen that the random motions of the stars can stabilize any distribution

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

up to a certain radius. For flat disks we can rewrite this Jeans' criterion as follows. In an infinitesimally thin disk we have a surface density σ and a velocity dispersion $\langle V^2\rangle^{1/2}$. Then the kinetic energy in a region with radius R is

$$T = \tfrac{1}{2}\pi R^2 \sigma \langle V^2\rangle$$

and the potential energy is roughly

$$\Omega = \frac{2\pi^3}{3} G\sigma^2 R^3.$$

For this case of a very thin disk then the Jeans' length is

$$R_{\rm Jeans} = \frac{3\langle V^2\rangle}{2\pi^2 G\sigma}.$$

For the old stellar disk in our Galaxy in the neighborhood of the Sun, the surface density is about $80 M_\odot$ pc^{-2} and the velocity dispersion roughly 40 km s^{-1}. Then this radius comes out as about 5 kpc. So we find that random stellar motions can stabilize the old stellar disk only up to moderate scales.

So something else has to be responsible for stabilizing the disk over larger length scales. This stability can be provided by differential rotation, which means that the angular rotation velocity varies with galactocentric distance. Parts further out have a lower angular velocity and as a result of this, seen from any point in the disk, these parts lag behind in the motion around the center, while the parts further in go ahead. This appears for not too large distances as a rotation around this reference position, and the angular velocity at which this occurs is described by the so-called Oort constant B. If the rotation velocity around the center of the galaxy is $V_{\rm rot}$ then

$$B = -\frac{1}{2}\left(\frac{V_{\rm rot}}{R} + \frac{{\rm d}V_{\rm rot}}{{\rm d}R}\right).$$

This provides a centrifugal force that can counteract the gravity and in this manner can provide stability. Consider again a circular area in the disk with radius R and surface density σ. For a particle at the edge this centrifigul force is B^2R. On the other hand, the gravitational force from the matter within R is approximately that due to the mass M inside R, or $GM/R^2 = \pi G\sigma$. Now, contrary to the Jeans' stability mechanism there is a *minimum* radius for stability, because as R increases the centrifugal force due to differential rotation increases. This minimum radius occurs where this centrifugal force equals the attraction due to gravity and is

$$R_{\rm min} = \frac{\pi G\sigma}{B^2}.$$

In our Galaxy in the solar neighborhood the Oort constant $B \sim -13$ km s^{-1} kpc^{-1}. This then gives $R_{\rm min} \sim 5$ kpc.

As Alar Toomre first described in 1964, a galactic disk is thus stable at all length scales, if this minimum radius is equal to or smaller than the Jeans' length. From the numbers given, we see that the disk of our Galaxy is indeed stable, but only just. So, disks can be stabilized at small scales by random motions and at large scales by differential rotation. Putting the two together then gives for a disk with given surface density and differential rotation a minimum velocity dispersion required for stability:

$$\langle V^2\rangle^{1/2}_{\rm min} = \left(\frac{2\pi}{3}\right)^{1/2} \frac{G\sigma}{|B|}.$$

Stars that have a small velocity with respect to the the general rotation will, in a coordinate system that rotates along with this general rotation, descibe a small elliptical orbit (an epicycle) in which it goes around in a period $2\pi/\kappa$ and

$$\kappa^2 = -4B\frac{V_{\rm rot}}{R}.$$

κ is the so-called epicyclic frequency. In actual galaxies the rotation velocity $V_{\rm rot}$ is constant with radius over most of the disk and then $\kappa = -2\sqrt{2}B$.

Then we can rewrite the equation above as

$$\langle V^2\rangle^{1/2}_{\rm min} = 4.09\frac{G\sigma}{\kappa}.$$

In a more detailed analysis and for arbitrary rotation curves, Toomre derived this minimum velocity dispersion actually as

$$\langle V^2\rangle^{1/2}_{\rm min} = 3.36\frac{G\sigma}{\kappa}.$$

Disks of spiral galaxies, including our own, do not appear to be violently unstable. However, there is spiral structure and star formation, so some instability must occur. As it turns out, as best as we can measure the disks are close to the border between stability and instability. This presumably is the cause of the fact that star formation is regulated as a moderately slow process, whereby the conversion of the interstellar gas into stars proceeds at a reasonable rate, as observed from the distribution of stellar ages in the solar neigborhood. As we can derive from the Toomre minimum velocity dispersion criterion above, the minimum dispersion decreases with galactocentric radius, the epicycle frequency decreasing as R^{-1}, but the surface density σ exponentially. This is indeed roughly as observed. Furthermore, it implies also a decrease in the velocity dispersion in the direction perpendicular to the plane in such a way that it roughly produces the constant thickness of the thin disk (the radial constancy of the scale height parameter z_0).

Bibliography

The material presented above (and many more related matters) is discussed in more detail in

Gilmore G, King I R and van der Kruit P C 1990 *The Milky Way as a Galaxy* (University Science Books)

Piet van der Kruit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galactic Year

A term that is sometimes used to describe the orbital period of the Sun around the galactic center, approximately 225 million years.

See also: galaxy, Milky Way Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxies at High Redshift

Galaxies at high REDSHIFT are very distant galaxies and, since light propagates through space at a finite speed of approximately 300 000 km s^{-1}, they appear to an observer on the Earth as they were in a very remote past, when the light departed them, carrying information on their properties at that time. Observations of objects with very high redshifts play a central role in cosmology because they provide insight into the epochs and the mechanisms of GALAXY FORMATION, if one can reach redshifts that are high enough to correspond to the cosmic epochs when galaxies were forming their first populations of stars and began to shine light throughout space.

One fundamental prediction of the theory of the big bang, which has found empirical confirmation in the discovery of the cosmic background radiation, is that early in its evolution the universe consisted only of matter and radiation coupled in thermodynamical equilibrium and homogeneously distributed in space. No galaxies or stars or any other structure could exist in such physical conditions, except for minuscule primordial fluctuations of density, superimposed on the otherwise extraordinarily smooth distribution of matter and energy by quantum physical processes during the first instant of existence of the universe. Today, the universe is highly inhomogeneous, with the matter organized in a hierarchy of structures such as stars, galaxies, clusters and superclusters of galaxies. Understanding the mechanisms that led to the transition from the homogeneous early universe to the structured one observed at the present epoch is of central importance to cosmology and fundamental physics. Yet, they remain poorly understood and still defeat empirical investigation.

Our current ideas attribute the formation of the cosmic structures to the action of gravity, which amplified the primordial density fluctuations until they collapsed under their own self-gravity and became bounded structures. According to this paradigm, the epoch of galaxy formation and the properties of the nascent galaxies were decided by the properties and the relative abundances of the matter and the energy that permeated the universe shortly after the big bang. Galaxies are the building blocks of the distribution of matter in space, and an important step to decipher the history of the cosmic evolution is to identify the first galaxies that formed in the universe and study their properties at the time of their formation. The search for 'primeval galaxies' has traditionally played a central role in observational cosmology for more than three decades and, even after the spectacular accomplishments of the last 5 yr or so, is still driving much of extragalactic research.

The redshift and the expanding universe

One of the most important scientific discoveries of this century is the expansion of the universe. This makes the galaxies and the other larger structures that populate the cosmic space recede away from each other with a velocity that is progressively higher for objects that are separated in space by larger distances. If the recession velocity between two objects is small compared with the speed of light, its value is directly proportional to the distance between them, namely

$$v_r = H_0 d$$

where the constant of proportionality H_0 is called the 'HUBBLE CONSTANT'. For larger recession velocities this relation is replaced by a more general one calculated from the theory of general relativity. In each cases, the value of H_0 provides the recession velocity of a pair of galaxies separated by unitary distance, and hence sets the rate of the expansion. Since if the universe has been expanding at a higher rate it has taken a shorter time to go from the size at the big bang (close to a mathematical point) to the current size, a higher value of the Hubble constant also means a younger universe. Recent measurements of the Hubble constant place its value in the range between 60 and 70 km s^{-1} Mpc^{-1} (1 Mpc $= 3.1 \times 10^{22}$ m). This corresponds to an age of the universe between 10 and 15 Gyr (1 Gyr $= 10^9$ yr).

A result of the cosmic expansion and of the finite speed of light is that the light emitted from any given galaxy is observed by a remote observer (i.e. one located on a different galaxy, for example the Milky Way) as having a longer wavelength than it had at the emission. This is because the cosmic expansion has caused the space between the two galaxies to increase during the lapse of time between emission and detection, and while the light was traveling from one galaxy to the other. This causes a 'stretching' of the wavelength of the light, namely a shift towards longer (therefore redder) wavelengths. The fractional change in the wavelength with respect to the one at emission as measured by the observer is called the 'redshift'. It follows that, the higher the redshift, the longer the stretch in wavelength, and hence the more distant the two galaxies.

The redshift provides a measure of the distance to a galaxy as well as of the lapse of time between the instant of emission and that of observation. Thus, an observer can see the galaxy and study its properties as it was at the epoch of the emission. Galaxies located near ours have very small redshifts and are seen as they were a very short time ago compared with the age of the universe. In practice, local galaxies are contemporaries of our own, and hence the present cosmic time corresponds to redshift zero. The value of the redshift of objects seen at the time of the big bang is infinity. Figure 1 shows the relationship between redshift, distance and the corresponding age of the universe (i.e. the lapse of time from the big bang to that redshift) for two different assumptions for the cosmological parameters Ω and Ω_Λ, namely the total density of matter and the energy density due to the cosmological constant, respectively. In both cases the Hubble constant has been assumed to be 65 km s^{-1} Mpc^{-1}.

In practice, redshifts are measured from the spectra of the galaxies, namely diagrams that plot the intensity of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

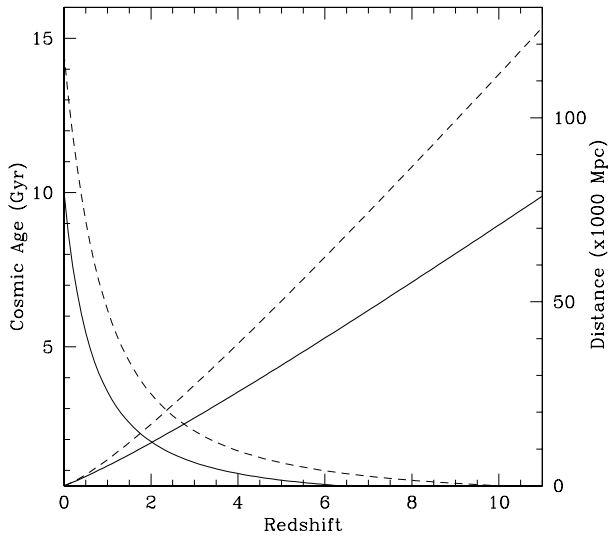


Figure 1. The relationship between redshift, cosmic age and distance for two different world models, both spatially flat (i.e. with Euclidean geometry). The Hubble constant is assumed to be $H_0 = 65$ km s^{-1} Mpc^{-1} in both cases. In one model only matter is present in space, namely $\Omega = 1.0$ (continuous curves); in the other model the space is filled with both matter and the cosmological constant, with $\Omega = 0.3$ and $\Omega_\Lambda = 0.7$ (dashed curves). The thin curves represent the cosmic age at the various redshifts. The cosmic age at the present epoch (redshift zero), is 10 and 14.5 billion years in the two models, respectively. At the time of the big bang (infinite redshift) the cosmic age was zero. Within the accuracy of the plot, redshift $z = 10$ can already be identified as infinite. The thick curves represent the distance of a source as a function of its redshift, expressed as $\times 1000$ Mpc. As a comparison, the Andromeda galaxy is at about 0.8 Mpc from the Milky Way and the Virgo cluster at about 18.5 Mpc. The most distant galaxies observed so far are located at a redshift $z \sim 5.5$.

a galaxy's light (also referred to as flux) as a function of its wavelength, by comparing the observed wavelengths of atomic and molecular features (either emission or absorption lines) with those of the same features observed in laboratories on Earth (i.e. at redshift zero). The top panel of figure 2 shows the model ultraviolet (UV) spectrum of a star-forming galaxy at redshift $z = 3$. Absorption lines can be observed as deep 'dents' in the spectrum. Another pronounced feature visible in the figure is the sudden dimming of the light at wavelengths shorter than the 'Lyman limit', as we shall discuss later.

The search for galaxies at high redshift

Because of their extreme distances, galaxies at high redshift necessarily appear to an observer as faint objects, even if their absolute luminosity is large. However, only a very small fraction of all the faint galaxies observed in a deep image of the sky are located at high redshifts. This is because galaxies have absolute luminosities that cover a wide range, extending over several orders of magnitude, and fainter galaxies are also much more abundant in space than brighter ones. As a result, deep images of the sky are crowded by a myriad of relatively close but intrinsically faint galaxies, while intrinsically luminous galaxies at high redshift are comparatively very rare. This 'contamination' by interlopers is in fact so severe that, without some criterion to cull them from the faint, nearby ones, it would be totally impractical to search for high-redshift galaxies by randomly measuring the redshifts of samples of faint galaxies until the very distant ones are found. Because of this elusiveness, high-redshift galaxies came to be regarded as the 'holy grail' of cosmological research.

Theoretical expectations of the mechanism of formation of the so-called spheroids, namely ELLIPTICAL GALAXIES and bulges of SPIRAL GALAXIES, provided a criterion to identify high-redshift galaxies. The stars that constitute the spheroids are among the oldest observed in the local universe, and currently there is no appreciable ongoing activity of STAR FORMATION in these systems. These facts indicate that the spheroids formed their stars during a very early cosmic epoch in a relatively short period of time. Since this epoch marked the first significant episodes of star formation during the cosmic evolution (note that about $\sim 1/2$ of all the stars of the present-day universe are found within spheroids), it has been traditionally identified with the epoch of galaxy formation, and the nascent spheroids themselves have been identified with the primeval galaxies.

The theoretical expectations predicted that the entire stellar content of a spheroid formed during the gravitational collapse of the proto-cloud of gas from which the structure originated. Simple physical arguments show that the duration of such a collapse is of the order of the time of free fall which, for a galaxy with the mass of the Milky Way, is about one hundred million years. If a whole galaxy's worth of stars is to be formed in such a relatively short time, the rate of formation of the stars must have reached very high values, a condition known as a starburst. For example, a galaxy like the Milky Way has a stellar mass of about $10^{11} M_\odot$[1], and if this assembled during a burst that lasted $\sim 10^8$ yr, the star formation rate must have been $\sim 10^3 M_\odot$ yr^{-1}. As a comparison, the star formation rate of the Milky way today is about $1 M_\odot$ yr^{-1}. A galaxy undergoing its first starburst is expected to emit intense UV radiation including, in particular, a very strong emission line at the wavelength $\lambda_\alpha = 1216$ Å called Lyman-α. This feature is produced by the recombination of hydrogen atoms, a gas which is very abundant in young, star-forming galaxies, photoionized by the intense UV light of young, massive stars.

Estimates of the age of the stellar populations of the spheroids place the epoch of the bursts in the redshift range $2 < z < 7$. The UV spectrum of such galaxies would then be observed redshifted to optical and near-infrared wavelengths and be detectable by ground-based telescopes and electronic detectors. For example, the

[1] The mass of the Sun is represented by the symbol $M_\odot$, and it has been measured to be $M_\odot = 1.98 \times 10^{30}$ kg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyman-α of a nascent spheroid at redshift $z = 3.5$ today is observed at $\lambda_{\rm obs} = (1 + z)\lambda_\alpha \sim 5500$ Å, which is in the middle of the visible band. If at redshift $z \sim 7$ or larger the Lyman-α line is redshifted into the infrared portion of the spectrum.

Large observational campaigns to identify primeval galaxies by means of their Lyman-α emission line started in the early 1980s, when solid-state electronic imaging detectors replaced the traditional photographic plates. The electronic detectors, also known as CHARGED-COUPLED DEVICES, overcame the characteristic limited sensitivity and difficulty of calibration of the plates, which had so far limited the study of very faint objects. In the early 1990s, electronic detectors sensitive to infrared radiation also became available, which made it possible to extend the searches for Lyman-α emitters to higher redshifts than those probed by optical detectors. These observations were designed to be sensitive to sources characterized by the presence of a strong emission line. Some consisted of narrow-band imaging, namely of images taken through narrow filters that can isolate the line emission, if its redshift is such as to place it at the wavelength of the filter. The presence of the emission line is detected from galaxy's relatively brighter apparent luminosity in the narrow-band filter compared with that through a conventional filter. Other searches consisted of spectroscopy of regions of the sky and were designed to directly detect the emission line from the recorded spectra.

Interestingly, except for a handful of objects, Lyman-α emitters at high redshifts were not found, while numerous detections of relatively bright objects were expected if a whole population of galaxies (i.e. the spheroids) formed through the mechanism of gravitational collapse. The lack of detections provided support for competing models for the formation of the spheroids. These predict that spheroids were assembled through hierarchical merging of smaller subgalactic structures. These small systems themselves have formed through a gravitational collapse before they merged into larger galaxies. However, because they would be significantly smaller than the present-day bright spheroids, they are predicted to have been correspondingly fainter than massive spheroids at the time of their formation, in this way eluding detection.

Unfortunately, the lack of detections of bright Lyman-α emitters at high redshifts did not conclusively constrain the mechanism of formation of the spheroids. On the one hand it did not prove the merging scenario. On the other hand it did not disprove the monolithic collapse scenario, since other explanations can be put forward to explain the lack of Lyman-α emitters. The most important of these is the presence of dust, which can form in nascent galaxies in a relatively short time, of the order of the free-fall time, as a result of the high rate of supernova events that take place in these systems. Even relatively small amounts of dust can have large effects on the Lyman-α luminosity, because this emission line is resonant. This means that, contrary to non-resonant lines, whose photons can travel long distances within the interstellar gas before interacting with other atoms, the Lyman-α photons are constantly being emitted, absorbed and re-emitted again by virtually all the hydrogen atoms that they encounter in their path. This greatly increases the geometrical path of the photons within the gas cloud before they can leave it. This, in turn, greatly increases the chances for the photons of being absorbed and destroyed by dust grains present in the cloud, resulting in a selective dimming of the luminosity of the galaxies at the Lyman-α wavelength.

Very recently, thanks to the extraordinary sensitivity of telescopes of large aperture, such as the 10 m Keck, Lyman-α emitting galaxies at high redshift have been eventually identified (Hu *et al* 1998). Interestingly, when sufficient numbers of these galaxies were observed that their abundances in space as a function of the luminosity (the so-called 'LUMINOSITY FUNCTION OF GALAXIES') could be characterized with sufficient precision, it became clear that those luminous enough to be considered plausible proto-spheroids were too rare. By then, however, a massive population of star-forming galaxies at high redshift (with or without Lyman-α emission) had already been identified by means of another technique. While the properties of these galaxies provided an explanation for the paucity of Lyman-α emitters, they also provided plausible candidates for the present-day spheroids.

The Lyman-break technique

A completely different technique to search for star-forming galaxies at high redshift was proposed in the early 1990s. This technique exploits another major feature of the UV spectra of galaxies with ongoing star formation, namely the hydrogen ionization edge or Lyman limit.

Star-forming galaxies are very luminous at UV wavelengths and have a characteristic 'blue' spectrum, i.e. rich in radiation of short wavelengths, which is 'flat', namely the intensity of the light does not depend on the wavelength. In a diagram that plots the light intensity as a function of wavelength such a spectrum looks approximately horizontal, as the top panel of figure 2 shows. However, ionizing radiation, namely light with wavelength shorter than 912 Å, although copiously produced inside the galaxies, cannot escape them (and thus be observable), because it is entirely absorbed by the hydrogen gas, which is very abundant within and around star-forming galaxies. When a hydrogen atom is hit by a photon with wavelength shorter than 912 Å, it becomes ionized, that is its electron is stripped from the proton at the expense of the energy of the photon itself, which is destroyed.

As a result of this absorption of ionizing photons, the spectrum as recorded by an observer external to the galaxies has a very pronounced 'dimming' of more than an order of magnitude at wavelengths shorter than the Lyman limit. This is called the Lyman break (see figure 2, top panel). Additional absorption is also introduced by the intergalactic gas that the light from the galaxy travels through in its journey towards the observer. This increases the amplitude of the break even further (and also

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

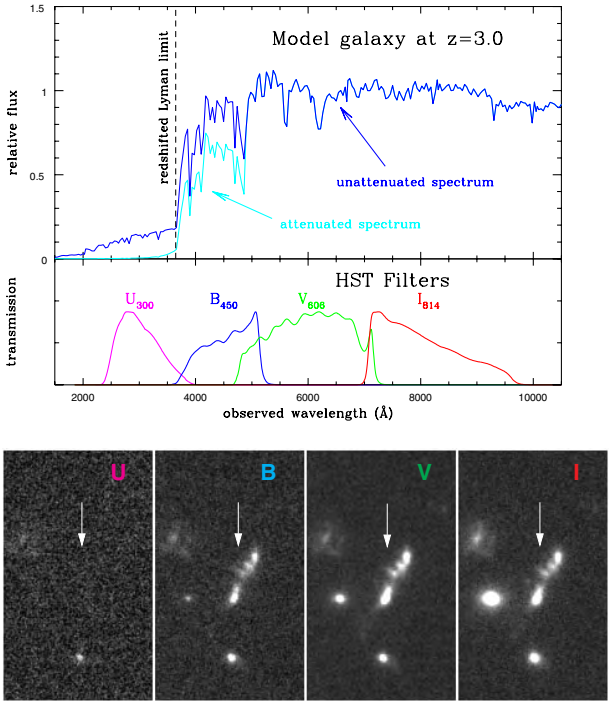


Figure 2. The UV spectrum of a star-forming galaxy observed at redshift $z = 3$, and the idea behind the Lyman-break technique. The top panel shows the spectrum redshifted at $z = 3$, namely with the intrinsic wavelengths increased by a factor $1 + z = 4$. Notice the sharp decrease of the intensity of the light at wavelengths shorter than the Lyman limit. The 'attenuated' spectrum illustrates the absorption of UV light by intervening neutral hydrogen. The middle panel shows the transmission curves of the four filters used to observe the Hubble Deep Field (HDF) with HST. When observing through a given filter, only radiation with wavelengths within the filter curve can reach the detectors. The bottom panels reproduce the images of a real Lyman-break galaxy at $z = 2.8$ (the multiple object with a sausage-like morphology indicated by the arrow) through each filter. Notice how the galaxy 'drops out' of the U filter, since this samples the spectrum to the left of the Lyman limit. Three other galaxies can be observed which do not drop out from the U filter (they are actually detected in the filter), but rather progressively dim, indicating that they are at relatively modest redshifts.

slightly attenuates the spectrum shortly before the break–see figure 2).

The Lyman break abruptly interrupts the 'flatness' of the spectra of star-forming galaxies, and provides the telltale clue to identify them. For example, at redshifts around $z \sim 3$ the Lyman break is shifted from 912 Å in the rest frame to ∼3700 Å in the observer's frame, well into the optical band, observable from the ground. When observed through a set of filters that straddle the Lyman break, star-forming galaxies at those redshifts appear relatively bright in those filters that probe the spectrum longward of the break (the I and V filters of figure 2), they dim a little just before the break (the B filter), and are extremely faint (or not visible at all) in those filters that probe shortward of the break (the U filter). Thus, by accurately measuring the light of galaxies in a carefully selected set of filters, it is possible to cull high-redshift galaxy candidates from the much more abundant galaxies of similar apparent luminosity placed at modest distances. This is illustrated in the middle and bottom panels of figure 2, which show the transmittance curves of filters used for this technique and one of the high-redshift galaxy candidates found in this way, respectively. These candidates are then followed up with spectroscopic measurements to confirm their redshifts. In practice, the selection of candidates is efficiently done with telescopes of middle size (e.g. 4–5 m), while the spectroscopic measurements to confirm the redshifts require telescopes of the 8 m class or larger. The candidate shown in figure 2 has been confirmed to be at redshift $z = 2.8$.

The wavelength of the filter bandpass is what determines the redshift range of the candidates. The filter suite shown in the figure is very sensitive in the redshift range $2 \lesssim z \lesssim 3.5$. However, by excluding the U band and using the B band in lieu to probe shortward of the Lyman limit, together with the V and I bands to probe longward of it, one can target higher redshift intervals, which in this case $3.5 \lesssim z \lesssim 4.5$. Other redder filters can be used for even higher redshift intervals.

The 'Lyman-break technique' turned out to be very sensitive and efficient. Although the exact numbers depend on the targeted redshift range, the fraction of the candidates that are confirmed at high redshifts by the spectroscopic measures is very high. For example, galaxies in the range $2 \lesssim z \lesssim 3.5$ are identified with essentially 100% efficiency. In the range $3.5 \lesssim z \lesssim 4.5$ the efficiency is still ∼50%. Two major surveys have been made using the Lyman-break technique. The largest one has been carried out by Steidel and collaborators (Steidel *et al* 1996, 1999) using ground-based telescopes for both the selection of the candidates and the spectroscopic confirmation. To date, this survey includes about 2000 candidates at $z \sim 3$, of which ∼1000 have spectroscopically measured redshifts, and another ∼300 candidates at $z \sim 4$ with ∼60 spectroscopic redshifts.

Another important survey has been made from space using the data collected by the Hubble Space Telescope during the observations of the HDF (HUBBLE DEEP FIELD) survey (Madau *et al* 1996). Although significantly smaller in size than the ground-based survey owing to the limited coverage of sky area, the HDF has allowed the identification of significantly fainter Lyman-break galaxies than possible from the ground, allowing researchers to study how the properties of these systems change with their luminosity. Because the archive is freely accessible to the world community, several groups have used the HDF data to identify high-redshift galaxies, and there have been reports of possible detection of galaxies with redshifts as high as $z \sim 5.5$ in the HDF.

The nature of high-redshift galaxies

The samples of high-redshift galaxies made available by the Lyman-break technique have eventually allowed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

researchers to carry out empirical studies of their properties, opening the distant universe to the entire investigation. These studies include both statistical analysis of the large samples themselves as well as follow-up observations, including high-angular resolution imaging with the Hubble Space Telescope and imaging and spectroscopy at near-infrared wavelengths (to study the rest-frame optical ones) from large ground-based facilities and again from HST. This work has allowed us to test our fundamental ideas on galaxy and structure formation on an empirical basis.

One important thing must be kept in mind when interpreting the galaxies' properties. The UV luminosity of star-forming galaxies (i.e. the intensity of the UV spectrum) is powered by the emission of young, massive stars, those approximately 20 times more massive than our Sun, while less massive stars are responsible for the optical and infrared luminosity. Massive stars are produced during star formation together with less massive ones, but, in contrast to the latter, which can live up to several billion years (depending on their mass), they only live about a million year or so, and then they explode as SUPERNOVAE. If the star formation continues in a galaxy at a steady rate, new massive stars constantly replace the dead ones, and the UV luminosity of the galaxy remains constant. The higher the star formation rate, namely the higher the number of massive stars formed in a given time, the larger the UV luminosity of the galaxy. At the same time low-mass stars keep piling up, increasing the optical luminosity of the galaxies. If star formation ceases, massive stars die off within a few 10^6 yr, and the UV luminosity fades away. The galaxy, however, remains visible at optical and infrared wavelengths because of the long-lived less-massive stars. The longer the duration of the star formation phase, the larger the amount of small-mass stars formed and the higher the optical and infrared luminosity. Thus, the UV luminosity of star-forming galaxies is a direct measure of their star formation rate, while the optical luminosity is linked to the amount of small stars that have been formed.

The assembly of galaxy structures

High-resolution images at optical wavelengths obtained with HST have revealed a variety of morphologies and sizes (Giavalisco *et al* 1996). Since at these high redshifts optical images probe the UV light, they provide information on the regions with active star formation. The images show that some galaxies are compact systems, with relatively smooth and regular morphology that bears a pronounced resemblance to the spheroids observed in the present-day universe. One example of such galaxies is shown in figure 3. Other galaxies are also regular, but they have more diffuse light profiles (i.e. the variation of the brightness from the center to the outer regions) and look more similar to present-day spirals, although familiar features of these systems such as the spiral arms have not been identified at high redshift. Still other galaxies are irregular and fragmented, often showing multiple compact nuclei embedded in diffuse nebulosity, and overall morphology that cannot be classified in terms of the traditional galaxy types. One example of this category is the galaxy at $z = 2.81$ shown in the bottom panel of figure 2.

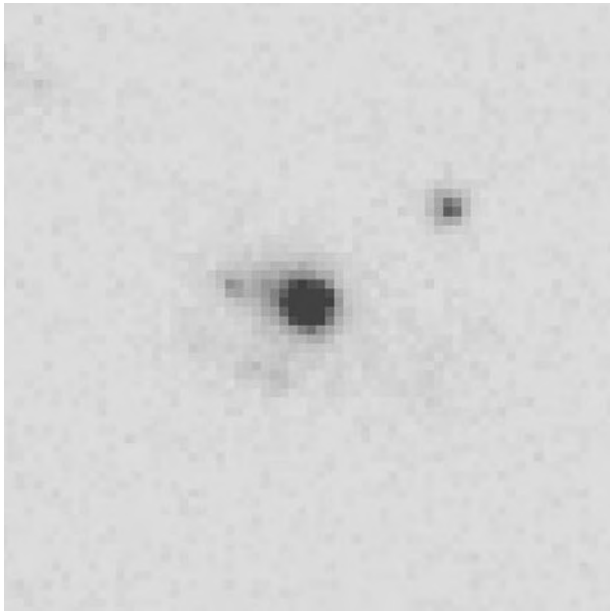


Figure 3. Picture of a Lyman-break galaxy at redshift $z = 2.96$ taken with the Hubble Space Telescope and the WFPC2 camera (the faint source to the upper left of the galaxy is unrelated). The image has been taken through the *V* filter F606W. The angular resolution is 0.05 arcsec, corresponding to a spatial scale of $\sim$0.5 kpc, and the size of the square image is 3.2 seconds of arc, or 27 kpc. The galaxy shows a regular morphology, and it is characterized by the same light profile commonly found in elliptical galaxies and bulges (the 'spheroids') of the present-day universe. The regular morphology of this galaxy contrasts with the fragmented structure found in the Lyman-break galaxy at $z = 2.81$ shown at the bottom of figure 2 (the third image from left of this galaxy has also been taken with the V filter F606W).

This diversity points to a variety of formation mechanisms. The compact, spheroidal galaxies can be explained if they formed as a result of a violent dynamical process, either a gravitational collapse of a proto-cloud or the merging of a number of discrete subunits, the more diffuse ones through the constant accretion of gas onto a rotationally supported disk, where it is converted into stars. In either case, the interesting fact is that the light profile of these galaxies strongly suggests that they are dynamically evolved and stable systems. At the observed rates of star formation and if left undisturbed in the course of evolution, they will have evolved into what are today elliptical and spiral galaxies of medium mass and luminosity, respectively. The fact that intense star formation seems to be occurring in these systems after the main dynamical event that gave origin to their structure (and most likely triggered the star formation activity) took place suggests that continued hierarchical merging after

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

redshift $z \sim 3$ is not necessary for the formation of some galaxies. One prediction in this case is that the mass of the regular Lyman-break galaxies is similar to that of present-day elliptical and spiral galaxies.

On the other hand, the irregular and fragmented morphology of other Lyman-break galaxies suggests that intense star formation occurs during interactions and merging events. In this case the expectation is that the mass of the forming galaxies progressively increases during the evolution (i.e. at smaller redshifts) and hence Lyman-break galaxies at $z \sim 3$ should also include small, submassive systems.

Direct measurements of the mass of distant galaxies are very difficult at the present, even with the most powerful of the current telescopes, whereas they will be routine with large-aperture space telescopes such as the proposed NASA 8 m Next Generation Space Telescope. Great progress in understanding the mechanisms of galaxy formation is expected with such a facility.

The cosmic evolution of star formation

One of the most important results that came out of the discovery of galaxies at high redshift is the possibility to reconstruct the evolution of the activity of star formation in the universe over a very large stretch of cosmic time. The temporal interval currently probed extends from the present epoch to when the universe was about 10% of its current age.

The amount of stars being produced in the universe at any given epoch can be estimated from the abundance and UV luminosity of star-forming galaxies at that epoch. The abundance is measured by their volume density, namely the number of galaxies in a given volume of space, for example 1 Mpc^3 (2.9×10^{67} m^3), while the luminosity is directly obtained by the photometry of the sources detected in the images. The counts of Lyman-break galaxies down to the faintest available UV luminosity, therefore, can be used to derive a measure of the UV luminosity density. Recalling that the UV luminosity is proportional to the star formation rate, this can then be expressed as star formation density, namely the amount of stellar mass formed in a given volume of the universe at a given redshift. A practical unit of measure for the cosmic star formation density is expressed as solar mass per year per cubic megaparsec, or $M_\odot\ \mathrm{yr}^{-1}\ \mathrm{Mpc}^{-1}$.

The diagram in figure 4 (top) shows the cosmic star formation density plotted (on a logarithmic scale) as a function of redshift. The different symbols indicate different surveys that have measured the cosmic star formation density in a variety of redshifts intervals, marked by the horizontal bars. Also shown are the uncertainties on the measurements, marked by the vertical bars.

Taken at face value, the diagram suggests that cosmic star formation activity started sometime just before redshift $z \sim 5$, gradually increased towards smaller redshifts, reaching a peak at redshifts $z \sim 1$ (about 1/3 of the cosmic age after the big bang) and then rapidly decreased towards the present epoch (i.e. $z = 0$) to the same levels as at $z \sim 5$. If confirmed, this would be a very interesting result, because it would empirically support the idea that during its early evolution the universe experienced a time when it contained no stars (dark age).

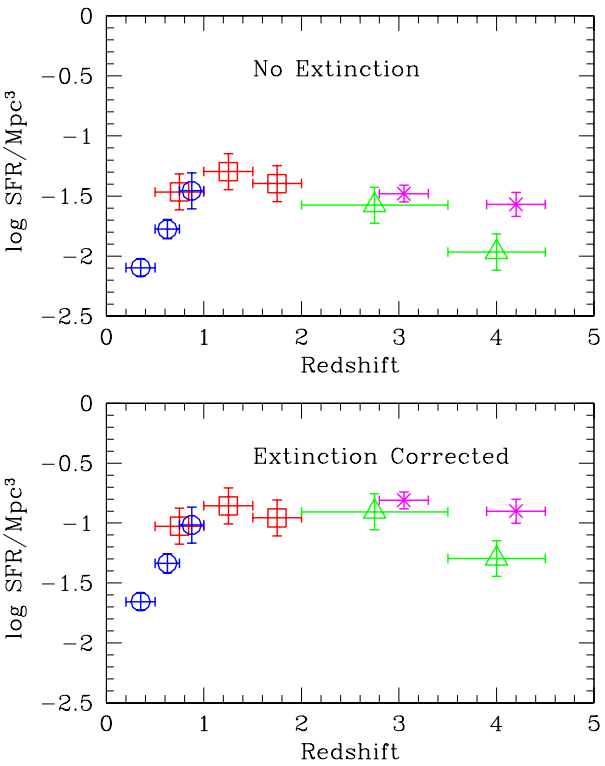


Figure 4. The history of cosmic star formation. Both plots show observations of the star formation density, namely the stellar mass produced every year in a cube of space 1 Mpc on a side, as a function of redshift. The y-axis is plotted on a logarithmic scale, and the mass is measured in units of the solar mass, i.e. 2×10^{33} g. The vertical bars for each data point represent the uncertainties on the measurements. The horizontal bars represent the redshift intervals covered by each data point. The top panel shows the data as observed; those in the bottom panel have been corrected for dust absorption, assuming that the properties of the dust do not change with redshift and are the same as observed in the local universe. Notice that the uncorrected data seem to suggest that the cosmic star formation began sometime prior to redshift $z \sim 5$, progressively increased until it reached a peak at $z \sim 1$ and then sharply decreased towards the present epoch, i.e. $z = 0$. The corrected data, however, do not support this interpretation, suggesting instead that the epoch of the onset of star formation in the universe has not yet been observed by the current data, and must be searched for at redshifts higher than $z \sim 5$.

However, these measurements must be taken with great caution, since they can be subject to systematic errors that can bias the results. As should be clear, the measures of abundance are of a statistical nature, since they are based on galaxy counts, and therefore they rely on the assumption that the observed samples at high redshift

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are fair representations of the population of star forming galaxies at their epochs. This would not be the case, however, if the samples were too small, because in this case they could be subject to a statistical fluctuations (cosmic variance) and be either overpopulated or underpopulated in galaxies, depending whether an overdensity or underdensity region, respectively, has been (by statistical chance) targeted.

Another more serious and insidious source of systematic error is that measures of UV luminosity can be affected by unknown amounts of dust obscuration. The interstellar medium of nascent galaxies becomes polluted by dust in short time scales, as this is produced in the supernova explosions that end the short lives of massive stars. Dust affects the observed luminosity of these sources, because it absorbs UV and optical radiation and converts it into infrared radiation. Moreover, the absorption is more severe for radiation with shorter wavelengths than with longer ones. As a result, the observed UV spectra are fainter and redder than they would be if dust were not present, and an observer would conclude from them that the galaxies are less luminous and hence forming stars at a lower rates than they actually are.

Qualitatively, the presence of dust in the Lyman-break galaxies at $z \sim 3$ and $\sim$4 has been revealed by a number of indicators. These include the observed UV spectra, which are systematically somewhat redder (richer in radiation with longer wavelengths) than those of known dust-free galaxies with similar properties. Another indicator is the intensity of the rest-frame optical emission lines (observed at near-IR wavelengths, because of the redshift) which are comparatively too strong for the observed UV luminosity, consistently indicating the presence of dust reddening. Quantitatively, however, it is extremely difficult to estimate the amount of obscuration suffered by the galaxies and, therefore, attempt a correction to recover the intrinsic luminosity and star formation rates. This is because the correction depends on the properties of the dust (the so-called 'extinction curve') and on the shape of the 'unreddened' spectra (i.e. those that would be observed without dust obscuration). Since these are not precisely known, the correction can only be computed after making some assumptions, and thus it will depend on them.

The bottom panel of figure 4 shows the evolution of the cosmic star formation activity after corrections for dust obscuration have been made, assuming that the properties of the dust and the intrinsic spectra of the galaxies at high redshift are similar to those of starburst galaxies in the local universe. If these assumptions are realistic, the correction at $z \sim 3$ is about a factor of 5, implying that the amount of star formation at those cosmic epochs was actually 5 times higher than what one would naively derive using the observed data without any dust correction. Changing the assumptions to cover all the known cases of dust properties and spectra of star-forming galaxies would change the correction, making it as small as a factor of 2 or as large as a factor of 15.

Irrespective of the exact value of the correction, however, the important fact is that there is no obvious evidence for a decline of the cosmic star formation activity towards high redshifts (i.e. past $z \sim 1$) when some amount of dust correction is attempted. This means that, going toward the highest redshifts probed by the current galaxy surveys, namely $z \sim 4.5$ or $\sim$10% of the cosmic age, really there is no evidence that the cosmic star formation activity is decreasing from the level reached at $z \sim 1$. In other words, there is no evidence from the current data that, going from large redshifts toward small ones, we are seeing the end of the dark era and the beginning of the epoch of star formation, namely the transition from the epoch when the universe did not contain stars in appreciable quantities to when it started to efficiently form them. It seems now very likely that to detect and study this transition we will have to identify galaxies at even higher redshifts than we currently are capable of doing.

Are there more high-redshift galaxies?

One interesting question to ask is whether the Lyman-break technique, i.e. the selection of distant galaxies from their UV emission, returns all the galaxies that are physically present at the targeted redshifts. Star-forming galaxies with a large amount of dust or old galaxies, namely with previously formed stellar populations and no star formation activity, have very little UV luminosity, if any at all, and cannot be found with the Lyman-break technique. Discovering such galaxies at very high redshifts (say, $z > 3$), if they exist, would have enormous consequences. On the one hand, if UV-dark star-forming galaxies are present in large numbers, this means that we have severely underestimated the amount of stars formed early in the universe. This implies that today there are many more stars and heavy elements[2] than we currently observe in the universe. On the other hand, if old galaxies are already present at very high redshifts, this would imply either that we have underestimated the age of the universe or that we do not understand very well the physical conditions of the early universe or time scale of STELLAR EVOLUTION.

Recent imaging observations at sub-millimetric wavelengths (Lilly *et al* 1999) have unveiled a population of sources with properties consistent with star-forming galaxies at high redshift whose UV radiation is being either partially or completely absorbed by dust and re-emitted at far-infrared wavelengths (observed as submillimetric wavelengths because of the redshift). Interestingly, some of these sources have been identified as Lyman-break galaxies at $z \sim 3$ (Barger *et al* 1999) and, conversely, Lyman-break galaxies at $z \sim 3$ have been observed as submillimetric sources (Chapman *et al* 2000). Many others have eluded the efforts to identify them and assign them a redshift, and at present the redshift distribution

[2] Elements heavier that helium are produced in large quantities only in the interior of stars and released into space when these explode as supernovae. Before the beginning of star formation the universe essentially consisted only of hydrogen and helium.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the submillimeter population is unknown. Of course, similarly unknown is the extent to which the Lyman-break galaxies and the submillimeter galaxies overlap, i.e. are the same objects. Assuming that the latter are placed at $z > 2$ and adding together all their luminosities yields a total energy output at infrared wavelengths that is a factor of several larger than that of the energy output at UV wavelengths by the Lyman-break galaxies. We have seen, however, that the UV luminosity of these sources is likely to be underestimated by a factor of several owing to dust obscuration, and, when a correction is included, the two contributions become comparable. This is consistent with the possibility that the two galaxy populations are the same one. In such a case the submillimetric emission would be powered by the UV light absorbed by dust and re-emitted at infrared wavelengths[3].

Unfortunately, with the current instrumentation it is not possible to identify the nature of the submillimetric sources with great confidence, and establish whether they have or do not have UV emission, which from high redshifts would be observed at optical wavelengths. This is largely because the angular resolution of the images is very coarse (about 15 seconds of arc) and many faint optical sources can be found within a circular region of the sky with such a large diameter, making the identification with optical galaxies very difficult. Future generations of submillimeter telescopes, such as the ALMA project that will be commissioned in Chile in 2007 as a joint collaborative project of the US, Europe and Japan, will have the sensitivity and resolving power to accurately image the sources and allow the identification with optical ones (or demonstrate that optical counterparts are very rare). It will be very interesting to see whether current UV survey have accounted for all the star formation activity in the young universe, whether most of it has actually been hidden by dust, or else if the universe was already populated by 'old' objects when it was only ∼10% or less of its current age.

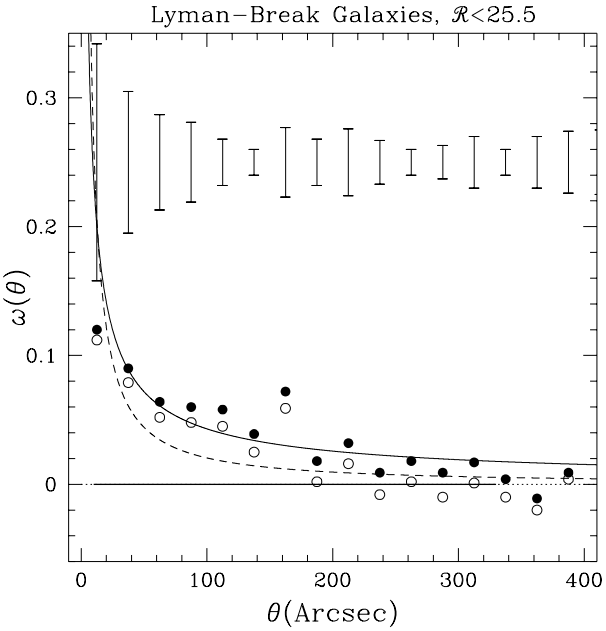


Figure 5. The angular correlation function of Lyman-break galaxies with apparent magnitude brighter than $\mathcal{R} = 25.5$. This quantity measures the probability of finding a pair of galaxies in the sky separated by an angle θ (measured in seconds of arc) as a function of the angle itself. The diagram shows that close galaxy pairs are much more likely to be found than loose ones, indicating the strong tendencies of Lyman-break galaxies to cluster in space. For clarity, the error bars have been plotted on the top of the figure. Filled and empty circles indicate two different estimators, used to test for the possible existence of systematic errors in the measurements. The continuous and dashed curves are the fits to the filled and empty circles (data points), respectively. The clustering and abundances of Lyman-break galaxies have been found to agree with the predictions of the theory of gravitational instability for dark matter halos, lending strong empirical support to the idea that galaxies have formed within halos that provided enough gravity to make gas condense and convert into stars.

Galaxy formation and dark matter

Another fundamental avenue of research opened by the identification of forming galaxies at high redshift is the possibility of testing the idea that gravity has been the force responsible for galaxy and structure formation.

If gravity has assembled the cosmic structures, then one prediction is that galaxies have formed in those regions of space where enough mass had condensed and produced the gravitational pull to confine the gas in a relatively small volume of space, promoting and facilitating its conversion into stars.

What are these mass condensations that seeded galaxy and star formation? There is compelling evidence that the majority of the mass present in the universe (about 90% of it) is not in the form of visible matter but is dark, either because it is cold or because it interacts very weakly with the electromagnetic radiation. Its presence, however, is detected through the gravitational effects, such as motion, that it induces in other, visible objects. The nature of the dark matter is unknown. It has been proposed that it might consist of ordinary matter such as objects similar to planets or unborn stars (brown dwarfs), or of more exotic types of matter, such as subatomic particles (e.g. massive neutrinos), other species of particles predicted by quantum theories (e.g. axions, photinos, gravitinos) or even of such systems as primordial mini BLACK HOLES. It is beyond the scope of this article to discuss dark matter in detail (but see DARK MATTER: ITS NATURE), and in the following we shall refer to it as a 'cosmological fluid' that, together with the regular visible matter, fills the space in the universe, and consider only its gravitational effects.

If gravity has assembled structures out of the cosmic fluid, then one specific prediction of the theory is that the spatial distribution of these structures (we shall refer to these structures as 'the halos') is not homogeneous,

[3] The equality of the two terms is expected in this case from the principle of conservation of energy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

but highly clustered. In other words, the halos are not expected to be distributed in space at random, but preferentially located next to each other, defining local concentrations that are also preferentially found next to other concentrations, and so on. Thus, if the visible part of galaxies (i.e. the stars) forms when gas condenses inside the dark matter halos, then the spatial distribution of forming galaxies must be the same as that expected for the halos themselves.

One of the most remarkable properties of the Lyman-break galaxies is that their spatial distribution is, within the uncertainties of the measurements, the same as that expected for the dark matter halos (Giavalisco *et al* 1998; Adelberger *et al* 1998). Figure 5 illustrates the strong spatial clustering of the galaxies. It shows the probability of finding a pair of galaxies separated by an angle θ in the sky as a function of the angle itself (measured in seconds of arc). This quantity is known as the angular correlation function and it is traditionally represented by the symbol $\omega(\theta)$. The data show that pairs of Lyman-break galaxies separated by small angles in the sky are much more likely to be found than pairs with large separations. This means that the galaxies have a strong tendency to cluster in space, namely to be physically closer to each other than in a homogeneous (random) distribution. If the galaxies were randomly distributed in space, the function $\omega(\theta)$ would have been flat, namely equal to a constant numerical value for any angular separation, indicating that pairs could have been found equally likely at all angular separations.

Because the distances to the galaxies are known, it is possible to transform the angular separations in the sky into physical separations in space and derive the spatial scales over which Lyman-break galaxies are clustered. These scales turn out to be of the order of 3–5 Mpc, the exact values depending on the Hubble constant H_0 and the cosmological parameters Ω and Ω_Λ, which are not yet precisely known. These observed scales have been found to be the same as those predicted by the theory of gravitational instability. The theory also predicts the abundances of the galaxies (i.e. the number of galaxies in a given volume of space), and, remarkably, the observed numbers of galaxies also agree very well with the predicted ones.

The predictions of the theory of gravitational instability concern structures that form out of the 'cosmic fluid' through the action of gravity. The simultaneous agreement of the observed clustering strength and spatial abundances of the Lyman-break galaxies with the analogous quantities predicted for the dark matter halos shows that the visible component of galaxies has the same properties predicted for the halos. This is consistent with the idea that visible galaxies and dark halos are physically associated and in fact are the same structures. This evidence provides strong empirical support for the notion that star formation takes place within the gravitational potential provided by the halos (recall that the dark matter accounts for ∼90% of the mass), which would act as 'condensation seeds', making it possible for the gas to condense and transform into stars. The observed clustering properties and abundances of Lyman-break galaxies represent a remarkable success for the theory and show that main ideas behind the paradigm of galaxy formation are generally robust.

Bibliography

Adelberger K L, Steidel C C, Giavalisco M, Dickinson M, Pettini M and Kellogg M 1998 *Astrophys. J.* **505** 18

Barger A J, Cowie L L, Smail I, Ivison R J, Blain A W and Kneib J-P 1999 *Astron. J.* **117** 2656

Chapman S *et al* 2000 *Mon. Not. R. Astron. Soc.* at press

Giavalisco M, Steidel C C, Adelberger K L, Dickinson M E, Pettini M and Kellogg M 1998 *Astrophys. J.* **503** 543

Giavalisco M, Steidel C C and Macchetto D 1996 *Astrophys. J.* **470** 189

Hu E, Cowie L L and McMahon R G 1998 *Astrophys. J.* **502** 99

Lilly S J *et al* 1999 *Astrophys. J.* **518** 641

Madau P, Ferguson H C, Dickinson M E, Giavaisco M, Steidel C C and Fruchter A 1996 *Mon. Not. R. Astron. Soc.* **283** 1388

Steidel C C, Adelberger K L, Giavalisco M, Dickinson M E, Giavalisco M and Pettini M 1999 *Astrophys. J.* **519** 1

Steidel C C, Giavalisco M, Pettini M, Dickinson M and Adelberger K 1996 *Astrophys. J.* **462** 17

Mauro Giavalisco

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxies: Classification

Galaxies are majestically beautiful objects, and anyone who has seen a galaxy through a telescope, or even our own Milky Way Galaxy with the unaided eye, cannot help but wonder at the nature of these remote star systems and what they have to say about our universe and ourselves. Much of what is known about galaxies began with a simple classification of their appearance as seen on direct photographs taken with large observatory telescopes. Galaxies present a wide variety of forms, or morphologies, and can be naturally divided into categories in much the same way as living organisms can be divided into genera and species. However, galaxy morphology presents problems for classification that would not be encountered in biological taxonomy and that make it difficult to classify them with the precision that is normally associated with living species. Still, visual galaxy classification continues to be useful at a time when galaxies have never been better understood. Classification provides order to the daunting variety of forms even if we do not yet know how the different forms came about. Classification also provides a framework for further studies and suggests a logical approach to studying galaxies.

Factors influencing galaxy classification and morphology

The main problem with classifying galaxies is that no galaxy can be brought into a laboratory so that it can be viewed from any particular direction or distance. Most galaxies have preferred planes of symmetry, and ideally for classification we would like to view them all from directly above their preferred planes of symmetry. However, these planes are oriented randomly to the line of sight, making the appearance of many galaxies very dependent on viewing geometry. Also, galaxies are spread over a wide range of distances, such that the farther they are away from us, the harder it is to see details of their morphology.

The appearance of many galaxies also depends on the wavelength of light in which they are viewed. Photography and imaging of galaxies is usually done through filters which transmit a relatively small portion of their spectral energy distribution, which refers to the distribution of their light with wavelength. With some exceptions, spectral energy distributions of galaxies are dominated by starlight, with a small fraction of their light coming from glowing interstellar gas. Starlight tends to be mostly continuous thermal radiation (meaning all wavelengths are emitted) whose color is determined by the surface temperature of the star. Hot stars (surface temperatures of 20 000 K or more) emit most of their light as ultraviolet radiation and appear bluish in color, while cool stars (surface temperatures of 4000 K or less) emit most of their light as infrared radiation and appear reddish in color. When imaged in short-wavelength light, such as blue or ultraviolet, the appearance of galaxies tends to emphasize hot, blue and relatively young massive stars, but when viewed in long-wavelength light, such as red or infrared, the appearance of galaxies tends to emphasize cooler and generally older less massive stars. Young stars tend to be less smoothly distributed and much less frequently present than old stars, so many galaxies appear more patchy and uneven in blue filters than in red filters.

The presence of interstellar material known as dust also has a wavelength-dependent effect on the appearance of a galaxy. Dust consists of fine particles of heavy atoms and light elements that are thought to be produced by processes connected with the evolution of stars. In highly flattened galaxies, dust collects within a thin plane at the galaxy's mid-section; in other, less flattened galaxies dust may be distributed more randomly. Since interstellar dust scatters short-wavelength light more effectively than long-wavelength light (by virtue of the small size of the particles compared with the wavelength of visible light), the effects of dust on galaxy morphology and classification are more serious on blue light images than on red light images. Dust also impacts the appearance of galaxies at far-infrared wavelengths, where thermal emission from dust can actually dominate over the contribution of starlight. At longer (radio) wavelengths, starlight from galaxies can be very weak, and the appearance of a galaxy can be determined mainly by SYNCHROTRON RADIATION (nonthermal electromagnetic radiation produced by relativisitic electrons spiraling in strong magnetic fields) or by emission from cold neutral atomic hydrogen at 21 cm wavelength.

The expanding universe can affect the appearance of galaxies. The cosmological REDSHIFT both shifts and stretches the spectral energy distribution of galaxies which would impact their appearance in any set of fixed standard filters. For example, the blue light appearance of a galaxy may be seen only in a red filter because the redshift has shifted that part of the spectrum into the red. Thus, galaxies at high redshift can be difficult to classify and compare with those observed at low redshift.

The appearance of a galaxy is not expected to be perfectly fixed with time. Galaxies tend to rotate, new stars may be born and older stars may die, and interstellar gas may be consumed or replenished at rates dependent on both internal and external factors. Galaxies tend to be large compared with their typical separations, so that interactions may impact their structure over a long period of time. Because of the finite speed of light, very distant galaxies are seen as they were when the universe was much younger than it is now, and their appearance may have been influenced by the different conditions that existed at those times.

Finally, the morphology of a galaxy is also seriously influenced by its total mass and its environment. Massive galaxies tend to be much more structured and well ordered, more luminous and of higher average surface brightness than low-mass galaxies, such that catalogs over-represent high-mass galaxies and under-represent low-mass galaxies. Because of this selection effect, the galaxy classification systems in use today apply mainly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to massive galaxies. There do not necessarily exist low-mass counterparts of all of the known types of high-mass galaxies. For this reason, statistics of galaxy types must take into account the luminosity dependence of morphology in order to get an accurate picture of the entire population of galaxies. There is also a well-established correlation between a given galaxy's morphology and the density of surrounding galaxies (see GALAXY MORPHOLOGY–DENSITY RELATION). Certain types of galaxies are prevalent in high-density environments such as rich galaxy clusters, while others are prevalent in lower-density environments. This correlation is one of the primary observations that any theory of galaxy morphology and evolution would have to explain.

Classification of nearby galaxies at optical wavelengths

For historical reasons, the first classifications for galaxies were made at optical wavelengths and were entirely restricted to relatively nearby galaxies. From 1781 to 1847, Sir William Herschel and his son, John (see HERSCHEL FAMILY), searched the sky for 'white nebulae', objects which were later proved to be external galaxies. Through large telescopes, the Herschels were able to detect different degrees of central concentration, apparent flattening and mottling in these objects. More complex structure could be seen in a few of the brighter cases, but it was William Parsons, Third Earl of ROSSE, who, in 1845, added the attribute 'spiral' to some members of the Herschels' 'white nebulae'. This is when galaxy morphology began to get very interesting. Lord Rosse used a 72 in speculum metal reflector, the largest telescope in the world in his day.

The main classes of galaxies were identified not from visual observations but from photography with large telescopes using plates with emulsions sensitive to the blue region of the spectrum. Photography became important in astronomy towards the end of the 19th century, and by 1920 the main classes of bright galaxies were identified. After establishing the extragalactic nature of the 'white nebulae' in 1924, EDWIN HUBBLE developed a classification system for galaxies that, with some refinements, is still in use today.

The Hubble classification system

A schematic of the HUBBLE CLASSIFICATION system as it is used today is shown in figure 1, and examples of each type are shown in figures 2, 3 and 4. Initially, Hubble distinguished three main classes of galaxies: elliptical galaxies, which were smooth and largely featureless systems having a round or elliptical shape; spiral galaxies, which were highly flattened disk-shaped systems having spiral 'arms' in the disk; irregular galaxies, which showed a chaotic appearance. As with living species, subclasses can be distinguished within these main classes. For ELLIPTICAL GALAXIES, symbolized by the letter E, the only other criterion that Hubble was able to use for classification was the ellipticity of the isophotes. Hubble appended to the letter E a numerical index of ellipticity: $n = 10(1 - b/a)$, where b/a is the ratio of the minor axis dimension to the major axis dimension. E0 galaxies are round while E6 galaxies are highly elongated.

The uppermost panel of figure 2 shows two ellipticals, NGC 221 and Maffei 1. In many elliptical galaxies, the light declines smoothly with increasing distance r from the center as $r^{1/4}$. Some 40% of ellipticals are known to have dust which breaks the characteristic smoothness of the light distribution. The dust may be scattered in patches or collected in planes of symmetry. The intrinsic shapes of ellipticals are nontrivial to deduce, but it is generally believed that many are triaxial, meaning that the semi-axis radii of isophotes (or contours of equal brightness) along three perpendicular symmetry planes are unequal. The projected index n in the classification En does not connect directly to a preferred intrinsic shape, and indeed n may have little intrinsic significance to an actual galaxy. This has led some authors recently to propose replacing the index n with a subclassification based on the deviations of the isophotes of ellipticals from a perfect elliptical shape. The range of shapes, from boxy to pointy or disky, can be tied to more meaningful physics than the index n.

Among the nearest galaxies, the most common type is DWARF ELLIPTICAL or DWARF SPHEROIDAL GALAXIES. These galaxies tend to have much lower masses and surface brightnesses than the giant ellipticals. The nearest normal major elliptical to our Galaxy is Maffei 1, in the far northern sky. Although less than 10 million light-years distant, Maffei 1 is heavily obscured by foreground interstellar dust in our own Galaxy and is difficult to study at optical wavelengths.

The symbol for spiral galaxies in the Hubble classification system was S, but these systems offered four criteria for subclassification. SPIRAL GALAXIES tend to be composite systems consisting of a smooth, relatively spherical bulge of old stars superposed within a highly flattened disk consisting of spiral arms and often young massive stars lining the arms. Spirals differ according to the presence or absence of a bar-shaped feature, the relative prominence of the bulge, the degree to which the arms are open or tightly wrapped and the degree to which the arms are lined by discrete, resolved objects. The last three aspects seem to correlate in that galaxies having large bulges also have smooth, tightly wrapped arms while those having little or no bulge have patchy, open spiral arms. These characteristics led to the Sa–Sb–Sc and SBa–SBb–SBc 'tuning fork' classification of spirals shown in figure 1, where the B stands for bar. In BARRED SPIRAL GALAXIES, the arms break from the ends of the bar, while in nonbarred spirals the arms break directly from the bulge region. Galaxies of types Sa and SBa are frequently referred to as 'early-type' spirals in astronomical literature while those of types Sc and SBc are frequently referred to as 'late-type' spirals, nomenclature which draws from terms used in stellar astronomy to connote an evolutionary sequence. Studies by H Shapley and G de Vaucouleurs led to the extension of the spiral sequence to types considered to be 'later' than Sc. At types Sd and SBd, the bulge is very

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

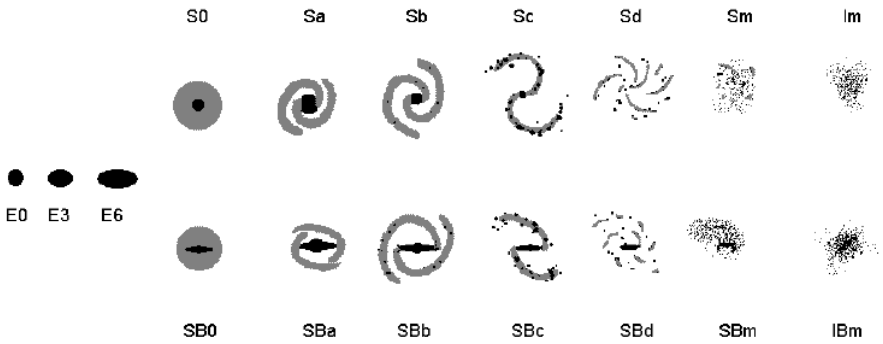


Figure 1. The Hubble sequence of galaxy types modified with additional types.

small or absent and the structure becomes increasingly open and highly resolved. The types Sm and SBm in figure 1 represent Magellanic-Cloud-type spirals. For a long time, the MAGELLANIC CLOUDS, two nearby galaxies seen in the far southern sky that are physical companions of our Galaxy, were regarded as IRREGULAR GALAXIES, but in the 1950s G de Vaucouleurs demonstrated that both have spiral structure with a characteristic asymmetry and no bulge component.

The majority of the galaxies that are listed in galaxy catalogs tend to be spirals of relatively high luminosity. In the Sa galaxy shown in the upper left panel of figure 3, the arms are smooth, tightly wrapped around the center, and the bulge is large compared with the disk. In the SBa galaxy in the upper right panel of figure 3, the bright bulge is crossed by a luminous bar. The bars of galaxies in general are composed of rather old stars compared with spiral arms.

The middle panels of figure 3 show typical examples of Sb and SBb galaxies, defined to be intermediate in bulge-to-disk strength, pitch of arms and resolution of arms compared with Sa,SBa and Sc,SBc galaxies. The Sb example, NGC 5985, shows a multi-armed spiral pattern. The SBb example, NGC 5850, shows a well-resolved spiral pattern of intermediate pitch angle.

The two lowest panels of figure 3 show typical examples of Sc and SBc galaxies. These tend to have the most well-developed and open spiral arms with a high degree of knottiness or resolution of the arms. The bulges tend to be very small compared with the extended disks. The objects that dot the arms tend not to be individual stars but unresolved clusters or loose associations of massive young stars. Such stars have spectral classes of O and B, and hence the clusters are usually referred to as OB ASSOCIATIONS. In the Sc example NGC 5457, these associations are several hundred light years in diameter and include hundreds of such stars.

The galaxies shown in the two upper panels of figure 4 are examples of Sd (NGC 45) and SBd (NGC 2500) types. In NGC 45, the spiral arms are weak and the bulge is very small; the bulge can only contribute a few per cent or less to the total luminosity of the system in this case. This is also true for NGC 2500, which has a clear bar. Note how both NGC 45 and NGC 2500 have many scattered 'knots', most of which are likely to be OB associations. The lower left panel of figure 4 shows a classic Magellanic barred spiral, NGC 4618, of type SBm. Although NGC 4618 has two clear arms, one arm is much brighter than the other and the center of the bar does not coincide with the center of symmetry of the outer disk light. These are characteristics of the SBm class.

Irregulars were originally characterized as 'nondescript' and lacking rotational symmetry. Although some are generally chaotic looking at relatively high surface brightness levels, at low light levels a faint, more symmetric disk can often be seen. It is now known that many irregulars do rotate and have bar-like structures in their disks. The properties of irregulars are so varied, and the range in their total luminosities is so great, that they have been the subject of much recent research. Irregular galaxies resembling the Magellanic Clouds but lacking any clear spiral structure are now referred to as Im or IBm types. The middle right panel of figure 4 shows a Magellanic irregular, NGC 2366, type IBm. This galaxy has numerous bright OB associations but no clear spiral structure. Its highly elongated shape suggests that it might be highly inclined.

In 1936, Hubble revised his classification system to include a fourth major galaxy class: S0 galaxies, which were armless DISK GALAXIES representing the transition from ellipticals to fully developed spirals. Some authors believe that the addition of this class destroyed the simple beauty of Hubble's original system. S0 galaxies are still an enigma because their relationship to spirals and ellipticals is still very unclear. Some authors believe that S0s represent spirals that were stripped of their residual gas by interactions in a cluster environment. Thus, rather than being transition types they form a separate sequence parallel to the spiral sequences. Other S0s may indeed be true transition forms between spirals and ellipticals. Note that elliptical and S0 galaxies are collectively referred to as 'early-type' galaxies, while irregulars are referred to as 'late-type' galaxies. These bracket the 'early-' and 'late-type' spirals. Some S0s have clear bars and are referred to as the SB0 type in figure 1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

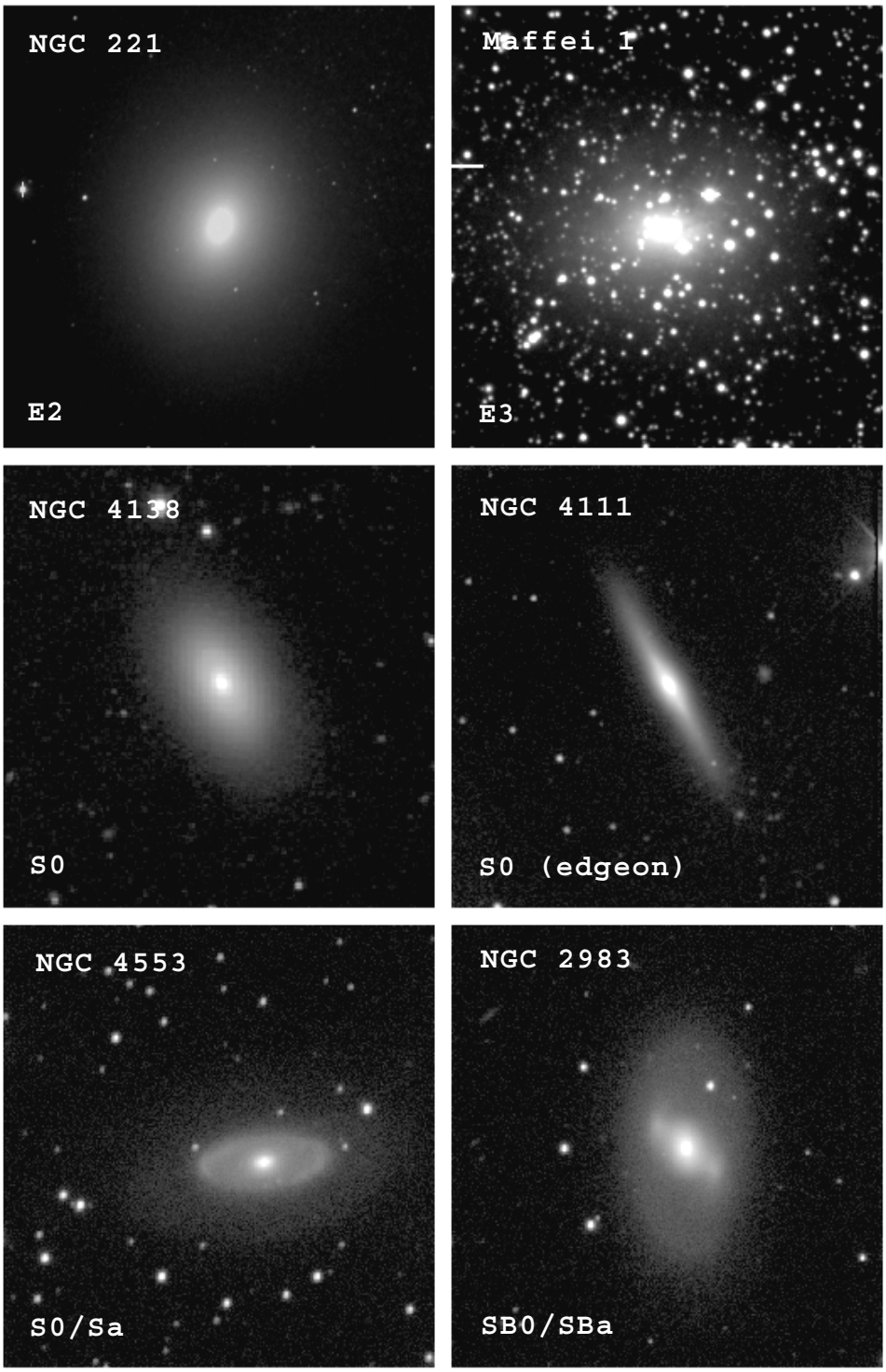


Figure 2. Examples of elliptical, early S0 and late S0 galaxies.

The morphology of S0 galaxies is best revealed by edge-on examples. When edge on, an S0 galaxy clearly shows its highly flattened, disk shape (see NGC 4111 in the middle right panel of figure 2). However, while edge-on spirals reveal dark lanes of dust in their midplanes, the best S0 galaxies show no such midplane of dust, implying

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

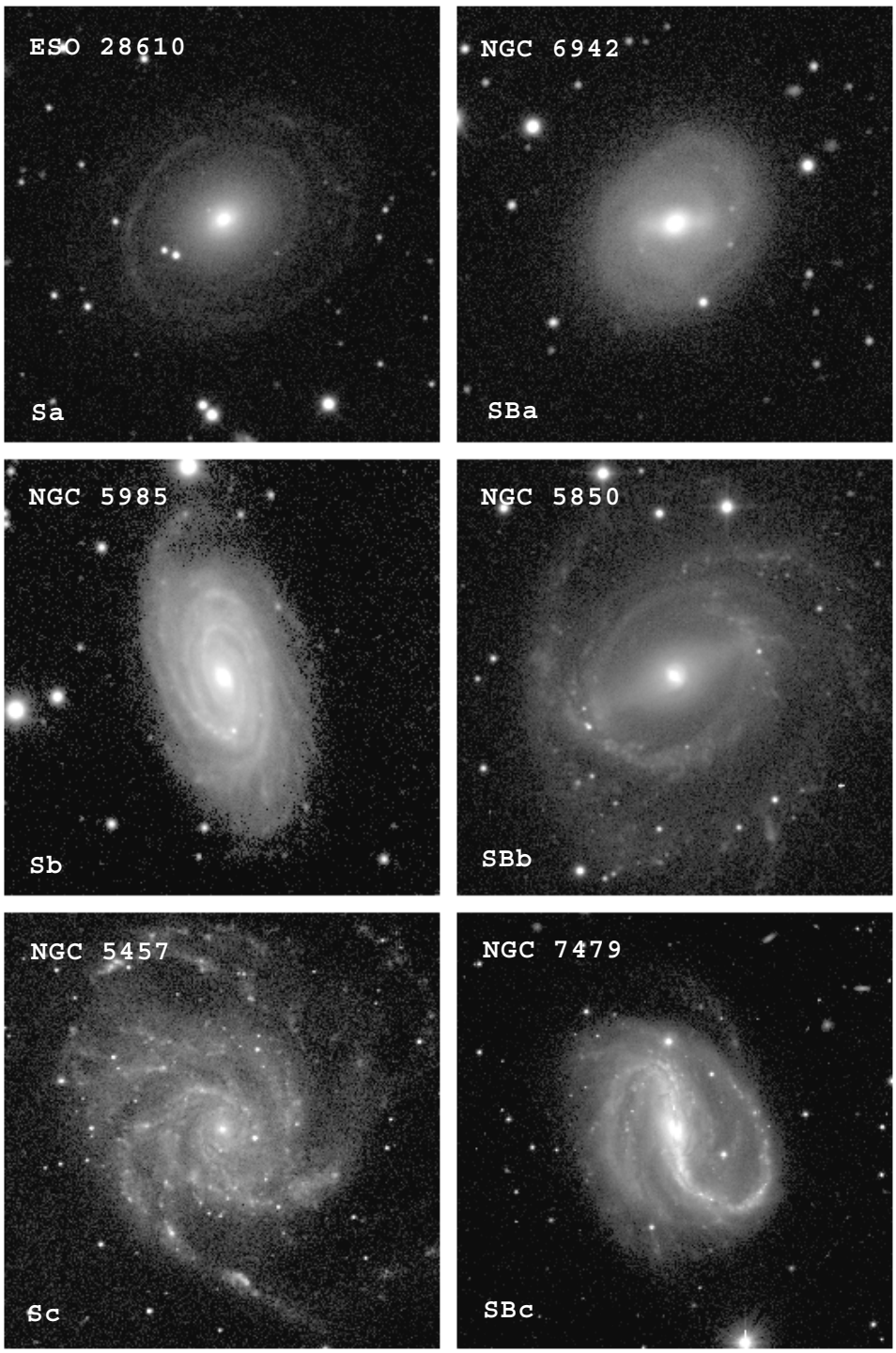


Figure 3. Examples of early- and late-type spiral galaxies.

the absence of much interstellar gas. This accounts for their smooth appearance and relatively red colors compared with spirals. Some 'late' edge-on S0s do show a midplane of dust.

At minimum, a true S0 galaxy should have no spiral arms, a clear bulge, and a clear disk or outer envelope of light. NGC 4138, shown in the middle left panel of figure 2, is a good example seen in an inclined but not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

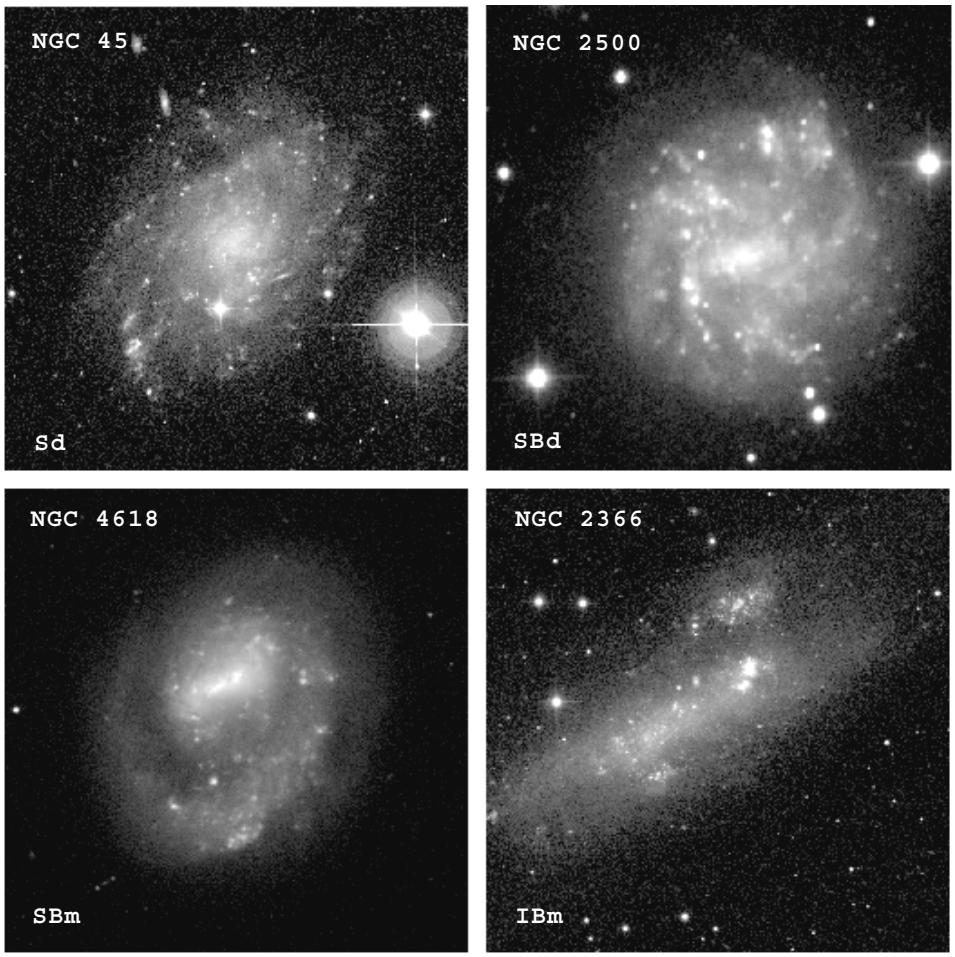


Figure 4. Examples of very-late-type and Magellanic spiral and irregular galaxies.

an edge-on view. This galaxy includes a weak ring of light just outside the bulge, which has led some authors to consider the galaxy to be a late S0. Seen face-on, the earliest S0s can be misclassified as ellipticals on inadequate image material. Late S0s have clearer differentiation of structure. Two examples are shown in the bottom panels of figure 2. NGC 4553 is a late S0 with a clear ring around the center. NGC 2983 is an SB0 with a well-differentiated bar and extended disk component. Both galaxies have also been classified as early-type spirals, although any spiral structure is very weak in both cases.

The relative luminosities of the bulges of S0 galaxies to their disks is a question connected to how we should interpret them. If they tend to have large bulge-to-disk luminosity ratios, then their placement in the Hubble sequence between ellipticals and spirals is likely to be correct. However, if some S0s have a relatively small bulge-to-disk ratio, then they could be 'stripped' spirals and should be placed on a sequence parallel to spirals.

The arrangement of ellipticals, S0s, spirals and irregulars along a sequence as shown in figure 1 is known as the Hubble sequence. The Hubble sequence is considered among the most important aspects of galaxies because certain properties of galaxies vary smoothly along the sequence. Interpreting what the sequence actually means is difficult because the mean luminosity of the galaxies is not the same at each type. The greatest spread in luminosity is found near the ends of the sequence, among the ellipticals and the late-type spirals and irregulars. An extremely important observation is how elliptical and S0 galaxies dominate the populations of rich and dense CLUSTERS OF GALAXIES, while spirals dominate looser clusters and the general field. This is the 'morphology–density' relation alluded to in the previous sections.

Other approaches to galaxy classification

Hubble's classification system is easy to use because it focuses on gross characteristics of galaxies. All of the original Hubble classes are rather broad in the galaxies they cover, but greater attention to detail is possible and other authors have proposed somewhat different points of view for classification, building on Hubble's idea.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

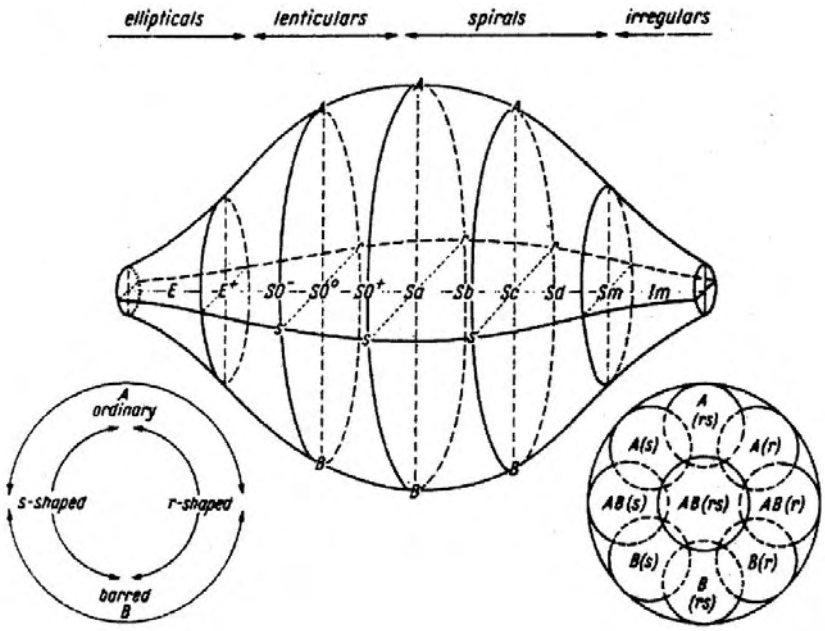


Figure 5. Illustration of the de Vaucouleurs revised Hubble classification system. In this system, S0 galaxies are referred to as 'lenticulars'. The lower left and right circles illustrate the arrangement of families and varieties within a cross-section through the system.

For example, rather than a two-pronged 'tuning fork' classification, G de Vaucouleurs proposed a classification volume whose main dimension is the basic Hubble type (i.e. E, S0, Sa, Sb, Sc, etc) and whose secondary dimensions are the family (presence or absence of a bar) and variety (presence or absence of a feature known as an inner ring; see figure 5). de Vaucouleurs' classifications provide a better description of what a galaxy looks like without being too complicated. Still, there are more than 100 'cells' in the de Vaucouleurs system compared with about 10 in Hubble's original system. The de Vaucouleurs system is recognized mainly for the addition of spiral types later than Sc, called Sd and Sm, and for the notation SA, SAB and SB to denote continuity of the bar characteristic. SA galaxies are true nonbarred (ordinary) galaxies, SB galaxies are true barred galaxies, while SAB galaxies are intermediate in apparent bar strength. Galaxies having a ring-shaped pattern in the inner regions are denoted as being of the (r) variety while those having no ring are said to be pure spirals, or the (s) variety. Galaxies having weak or partial, broken rings are assigned a variety of (rs), meaning a pseudoring is present. Examples of all nine possible combinations of families and varieties are illustrated in figure 6. The galaxies shown all have types of Sa or Sb. Inner rings are fairly common among these types, but are very much rarer among Sc types and later and nonexistent among ellipticals and early S0s.

In 1960, S van den Bergh demonstrated that galaxy morphology is influenced by the total luminosity or energy output of a galaxy. He assigned not only modified Hubble types to spiral and irregular galaxies, but also luminosity classes symbolized in a manner similar to stellar luminosity classes. The most luminous spirals or irregulars are luminosity class I while the least luminous are luminous class V. Categories II, III and IV have intermediate luminosities between these extremes. Most galaxies of luminosity classes IV and V are irregulars or very late-type spirals (Sd or Sm types). Of the spiral galaxies illustrated in this article, NGC 210, NGC 1300, NGC 1433, NGC 5457, NGC 5985 and NGC 7479 are luminosity class I or I–II, NGC 45 and NGC 2500 are luminosity class III, while NGC 2366 is luminosity class IV–V, as judged by A Sandage and G Tammann. Note how much more contrasted the spiral structure of the high luminosity class galaxies is compared with the intermediate or lower luminosity classes.

Spiral galaxies may also be classified strictly according to the appearance of their arms. Some spiral galaxies show a 'grand-design' spiral structure, consisting of two well-defined symmetric and long spiral arms. The first attempts to understand the nature of spiral structure in galaxies focused on grand-design spirals because they showed the best-defined spiral patterns. In other galaxies, however, the spiral arms are only chaotic or fragmentary and do not appear to be global in nature; such spiral patterns are often called 'flocculent'. The Hubble classification system and its revisions do not distinguish between these extremes in spiral arm morphology. D and B Elmegreen brought greater attention to these differences and proposed recognizing them with arm classes (ACs) ranging from 1 for the purely flocculent types to 12 for the best grand-design types. Intermediate-looking spirals

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

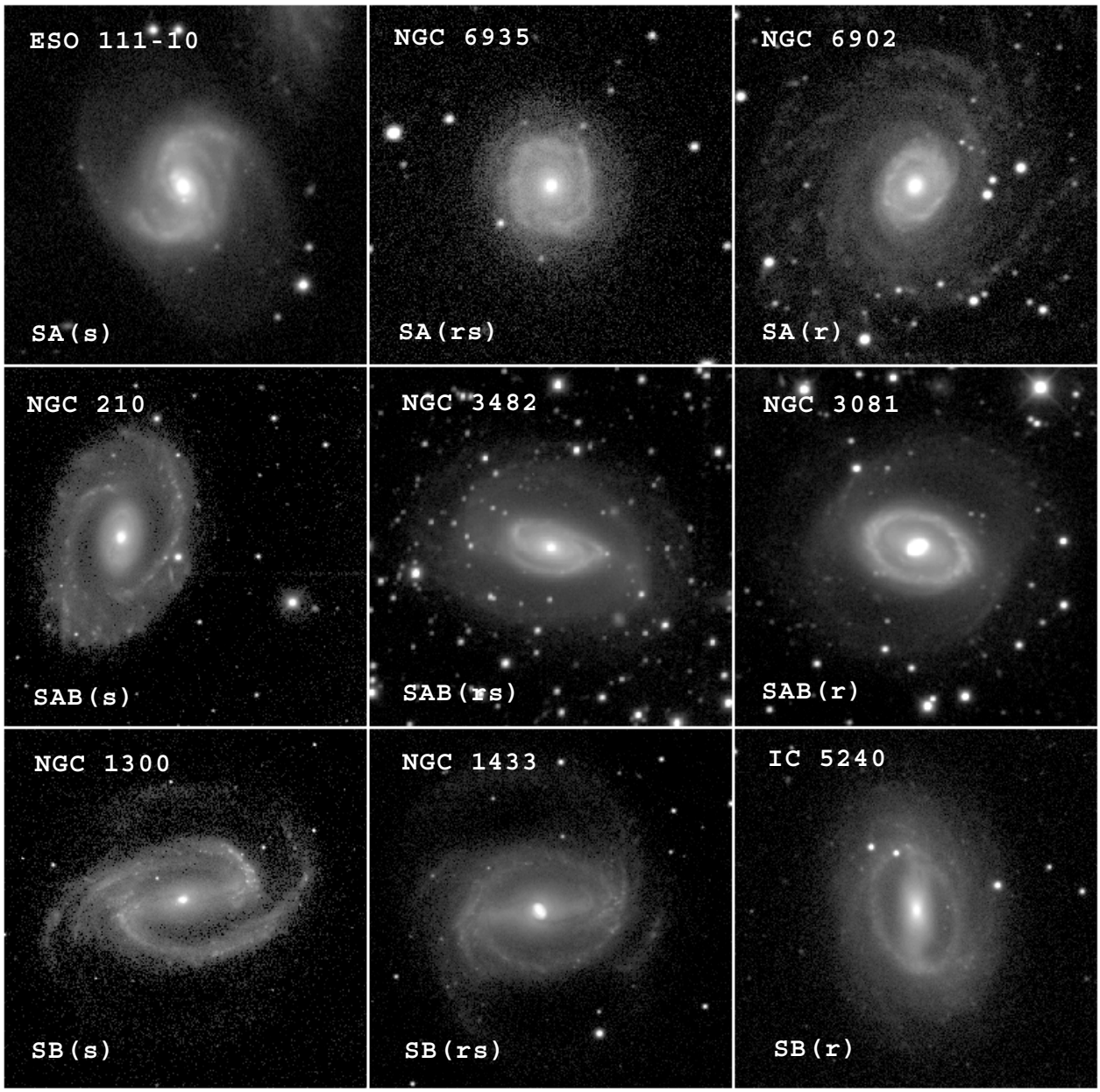


Figure 6. Examples of galaxies having different families and varieties in the de Vaucouleurs revised Hubble classification system.

have intermediate arm classes. In this article, NGC 1300 is an example of an AC 12 spiral while NGC 45 and NGC 2500 are examples of AC 1 systems. The distinction is worth recognizing because the origin of spiral patterns may differ among the different arm classes. Recognizing them allows one to do statistics to identify environmental factors that may be relevant.

W W Morgan proposed another classification system for galaxies in 1958 that tied galaxy morphology to the then-current ideas of STELLAR POPULATIONS in galaxies. The degree of central concentration of a galaxy was used to define a spectral classification system (population group) based on form alone. Galaxies with strong central concentration were known to be dominated by spectral class K giants (population group k) in their central areas, while those with little central concentration were dominated by spectral class A stars (population group a). The elliptical, S0 and early-type spirals illustrated in this article are all population group k systems, while the late-type spirals (Sd, Sm) and irregulars (Im) tend to be population group a. Morgan also identified an extremely important rare type of galaxy known as a cD galaxy. These

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

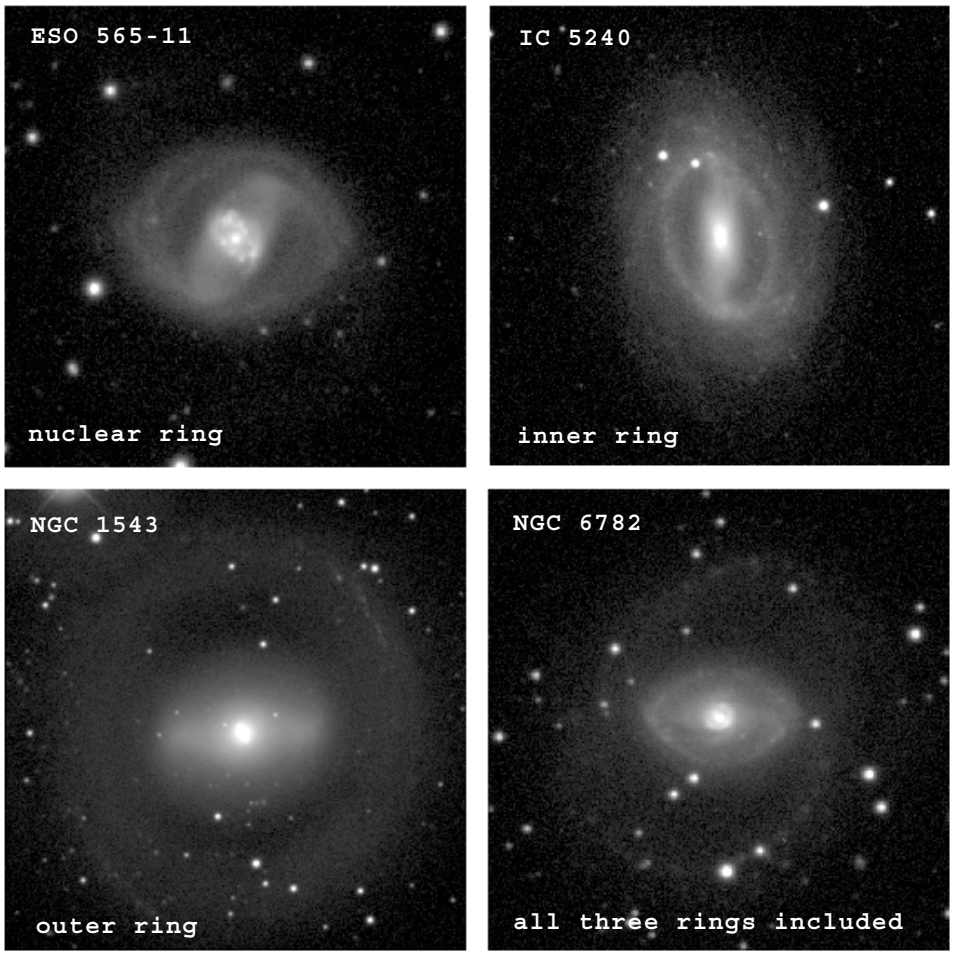


Figure 7. Examples of ringed galaxies.

are extremely large elliptical-like systems often found at the centers of rich galaxy clusters. They are characterized by a shallow light profile, low central concentration, and an extended envelope of light. The origin of such systems has been a topic of much research, and mergers of other galaxies is one possible interpretation.

Other classes or aspects of galaxy morphology

Ringed galaxies

A ringed galaxy is a normal spiral or S0 galaxy having a ring or ring-like pattern as part of the light distribution. In the jargon of normal galaxy morphology, there are three different types of rings: nuclear rings, which are small rings of active star formation often found in the centers of early-type barred galaxies, inner rings, which usually envelop the bars of barred spirals, and outer rings, which are large and faint structures having a diameter about twice that of the bar in barred galaxies. Of the three different types of rings, only inner rings are included as a major part of galaxy classification in the Hubble system, as noted in the previous section. The four panels of figure 7 show examples of all three ring types, one of which (IC 5240) is also included in figure 6. The classification symbols (nr), (r) and (R) are used for nuclear, inner and outer rings, respectively, and all three ring types may coexist in the same galaxy (see NGC 6782 in the lower right panel of figure 7). Inner rings and pseudorings can be identified in more than 50% of normal giant galaxies, and all three ring types are most frequent among early-type barred spirals.

Of the many morphological features of galaxies we have so far discussed, rings are among the best understood. They are features that seem to be generated by the way a bar can redistribute residual gas in a galactic disk, collecting the gas into ringed-shaped patterns that eventually condense into stars. The patterns are tied to locations known as orbital resonances, places where gas clouds move in step with the rotation rate of a bar pattern. Thus, rings and bars are features which are intimately connected.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

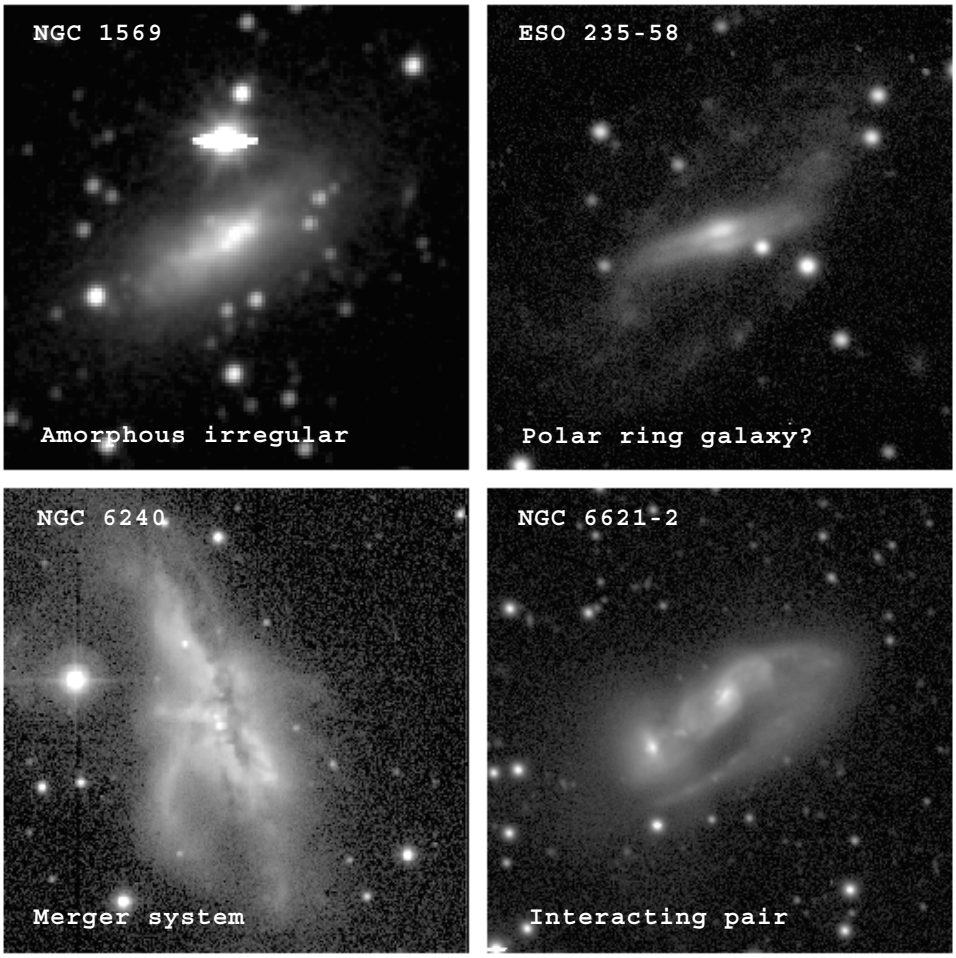


Figure 8. Examples of interacting galaxies and an amorphous irregular galaxy.

Amorphous irregular galaxies

Although classical irregular galaxies tend to be late-type systems with a considerable degree of patchiness and irregularity, another class of irregular galaxies has such amorphous characteristics that some astronomers considered them to belong near the transition between S0 and spiral galaxies. An example, NGC 1569, is shown in the upper left panel of figure 8. These objects tend to have much smoother distributions of luminosity than classical Magellanic irregulars. The distinction has been recognized since 1950, when E Holmberg suggested the term 'Irr I' be applied to normal late-type irregulars and 'Irr II' to these amorphous irregulars. de Vaucouleurs applied the term 'I0' to some of these objects, while Sandage prefers the general term 'amorphous' galaxies. The hallmark of the class is that the spectral character of the unresolved light is consistent with the presence of many young stars, unlike most other galaxies in the transition region between spirals and S0s. Interactions may play a role in the creation of these objects.

Interacting galaxies

Some of the most complex morphologies seen in the galactic population form directly as a result of gravitational interactions with neighboring galaxies (see GALAXIES: INTERACTIONS AND MERGERS). Interactions can be mild and may do no serious damage to the symmetry of a galaxy; however, such interactions may help to generate the formation of bars or even grand design spiral structure. Stronger interactions can lead to asymmetry and peculiar structure, and possibly even features known as 'tidal tails'. The strongest interactions can be so disruptive that the galaxies involved lose their individual identities and merge into a single, different type of object. It is currently thought that many giant elliptical galaxies formed as a result of the mergers of two spiral galaxies rather than naturally. Figure 8 shows a few cases of interacting systems. NGC 6240 in the lower left panel has been shown by numerous authors to be a remnant of a merger between two once separate galaxies; the system has not settled completely into a stable configuration and shows much peculiar structure. In the lower right panel, NGC 6621-2

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is a pair of clearly interacting galaxies. Interactions of this nature are not rare and many galaxies may go through a phase of interaction throughout their evolution. Gravity has drawn out material into the form of an apparent bridge between the two systems.

Certain types of interactions can lead to very special and very rare galactic morphologies. For example, if a small galaxy plunges directly through the center of a larger disk-shaped galaxy, a collisional ring galaxy can result. These unusual systems can look like smoke rings in lacking a central nucleus. The ring in this case represents a shock wave moving outward, much like a water wave that results when a pebble is dropped in a pond. Another type of morphology can result when a small companion galaxy, usually a gas-rich irregular, passes close over the pole of a gas-poor S0 disk galaxy. In this circumstance, the interaction can disrupt the companion and capture it into a polar orbit. Eventually, the material that made up the companion spreads along the whole orbit, producing a POLAR RING GALAXY. The upper right panel of Figure 8 shows a likely example of a polar ring galaxy in formation. Known as ESO 235-58, this object actually shows a midplane dust lane dividing the central disk. Although the central section looks like a bar, bars do not normally have dividing dust lanes crossing them from end to end. The central section of ESO 235-58 must be an edge-on disk galaxy, and the other material is orbiting out of the plane of that disk.

Galaxy classification and morphology in the next decade

The era of the orbiting HUBBLE SPACE TELESCOPE (HST) has added a new dimension to galaxy morphology and classification that could not be tapped from previous ground-based observations. HST can provide images of galaxies at large cosmological redshifts of the same quality and resolution used to define the main morphological classes on ground-based images. GALAXIES AT HIGH REDSHIFT are being seen as they were when the universe was much younger than it is now. As noted at the beginning of this article, we do not expect that galaxy morphology is fixed and unchanging, but that change must have occurred. Thus, by studying high-quality images of galaxies at high redshift, we can see what typical changes must have occurred. Current results suggest that interactions between galaxies were more frequent in the past and that some galaxies had not yet settled into well-organized shapes. Thus, morphology contains clues as to how galaxies evolve, and galaxy classification will continue to be refined and adjusted to account for new findings.

The sheer numbers of galaxies available for classification will also have an impact on how astronomers will be classifying galaxies. A large survey can provide high-quality images of more than a million galaxies, and it is impractical to classify so many galaxies visually by one or more experts. Instead, the coming decade will see more and more the use of computers to objectively classify large numbers of galaxies for cosmological and other research.

Galaxies can also be classified in wavelength bands other than the blue-light band used to develop the Hubble classification system. The tendency for blue filters to emphasize hot, young, massive and relatively rare stars means that the classification system tends to overestimate the significance of these objects. In the next decade, galaxy classification will be revised in order to group galaxies according to their appearance in infrared light, which tends to emphasize the stars that dominate the mass distribution. This will be especially important for intermediate to late-type spirals.

Bibliography

Hubble's original paper on galaxy classification was summarized in

Hubble E 1926 Extra-galactic nebulae *Astrophys. J.* **64** 321–69

His revision to include S0 galaxies is described in

Hubble E 1936 *The Realm of the Nebulae* (New Haven, CT: Yale University Press) (reprinted 1958 New York: Dover)

The de Vaucouleurs revised Hubble classification system, including Magellanic spirals and irregulars, is outlined in

de Vaucouleurs G 1959 Classification and morphology of external galaxies *Handb. Phys.* **53** 275

Large atlases illustrating Hubble's final revised classification system are found in

Sandage A 1961 *The Hubble Atlas of Galaxies* Carnegie Institution of Washington Publ. no 618

Sandage A and Bedke J 1994 *The Carnegie Atlas of Galaxies* Carnegie Institution of Washington Publ. no 638

van den Bergh luminosity classes of galaxies were first described and later further outlined in

van den Bergh S 1960 A preliminary luminosity classification of late-type galaxies *Astrophys. J.* **131** 215–23

van den Bergh S 1960 A preliminary luminosity classification for galaxies of type Sb *Astrophys. J.* **131** 558–73

Sandage A and Tammann G 1987 *A Revised Shapley–Ames Catalog of Bright Galaxies* Carnegie Institution of Washington Publ. no 635

The Morgan system of spectral form classifications of galaxies is given in

Morgan W W 1958 A preliminary classification of the forms of galaxies according to their stellar population *Publ. Astron. Soc. Pac.* **70** 364–91

Arm classifications for galaxies are described in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Elmegreen D and Elmegreen B 1987 Arm classifications for spiral galaxies *Astrophys. J.* **314** 3–9

Ringed galaxies are reviewed by

Buta R and Combes F 1996 Galactic rings *Fund. Cosm. Phys.* **17** 95–281

Collisional ring galaxies are reviewed by

Appleton P N and Struck-Marcell C 1996 Collisional ring galaxies *Fund. Cosm. Phys.* **16** 111–220

Polar ring galaxies are reviewed by

Whitmore B C *et al* 1990 New observations and a photographic atlas of polar-ring galaxies *Astron. J.* **100** 1489–522

An up-to-date review of the whole field of galaxy morphology and classification is provided by

van den Bergh S 1998 *Galaxy Morphology and Classification* (Cambridge: Cambridge University Press)

Image credits

NGC 221, NGC 4618, NGC 5985, and NGC 7479 were obtained by S Odewahn and R Gal with a green filter and the Palomar 1.5 m telescope. NGC 45, NGC 2366, NGC 2500, and NGC 5457 are blue light images obtained by D Elmegreen and coworkers with telescopes at the Kitt Peak National Observatory. NGC 4111 and NGC 4138 are near-infrared images from the Ohio State University bright galaxy survey, supported by a grant from the National Science Foundation. For these early-type galaxies, there is little difference between the appearance in the near-infrared and the appearance in a standard blue light filter. W C Keel provided the images of NGC 6240 and NGC 6621-2. The image of NGC 6240 was obtained in blue light at the European Southern Observatory while that of NGC 6621-2 was obtained in red light at the Kitt Peak National Observatory. The remaining images were obtained by the author and coworkers with telescopes at the Kitt Peak National Observatory and the Cerro Tololo Inter-American Observatory, and are in blue light only. All images are based on electronic detectors and are in logarithmic units.

Ronald Buta

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxies: Interactions and Mergers

For much of the 20th century—once the vast distances to galaxies were known—galaxies were thought to be 'island universes', forming and evolving in isolation with no contact between one another. In this picture, the processes which shape the galaxies we see today are uniquely determined by the initial conditions under which galaxies form and processes completely internal to the galaxies themselves.

As galaxy catalogs began to grow, however, more and more examples of paired galaxies were found, as well as many PECULIAR GALAXIES with long, luminous 'plumes' and 'tails' emanating from their bodies. These galaxies were found to have anomalously blue colors, arguing that their star-forming properties were quite different from normal galaxies. Examples were found of galaxy pairs so strongly perturbed that they were suggested as possible examples of actual merging encounters. Meanwhile, astronomers began to use computer simulation to study the effects of nearby companions on galaxies and found that many of the properties of these peculiar galaxies could be explained through gravitational interactions and mergers of galaxies.

From these studies, an alternative picture of GALAXY EVOLUTION began to grow. Rather than evolving in isolation, galaxies are found in clusters and groups and can interact quite strongly with their nearby companions (see GALAXY CLUSTERS). These interactions can have a profound impact on the properties of galaxies, resulting in intense bursts of STAR FORMATION, the onset of QUASAR-like activity in galactic nuclei and perhaps even the complete transformation of SPIRAL GALAXIES into ELLIPTICAL GALAXIES. Studies of galaxies in the early universe show a significant fraction of interacting and merging systems, and theories of cosmological structure formation indicate that most galaxies have had some form of strong interaction during their lifetime. Rather than being rare events, galaxy interactions may be the dominant process shaping the evolution of the galaxy population in general.

Dynamics of Interacting Galaxies

The evolution of interacting galaxies is governed largely by gravitational effects. That these interactions can have a profound effect on the participating galaxies is clear from the observational record. Morphologically, interacting galaxies are found to sport long bridges and tails, stellar bars and/or enhanced spiral structure and often severely distorted main bodies (see figure 1). At first, many astronomers believed the long streamers of stars and gas seen emanating from some interacting galaxies to be shaped by magnetic fields or nuclear jets—the thinness and linearity of these features made a gravitational origin seem unlikely. However, computer models of interacting galaxies in the early 1970s by Alar and Juri Toomre showed convincingly that these streamers were the simple consequence of gravitational tides acting on rotating disk galaxies, and these streamers were dubbed 'tidal tails'.

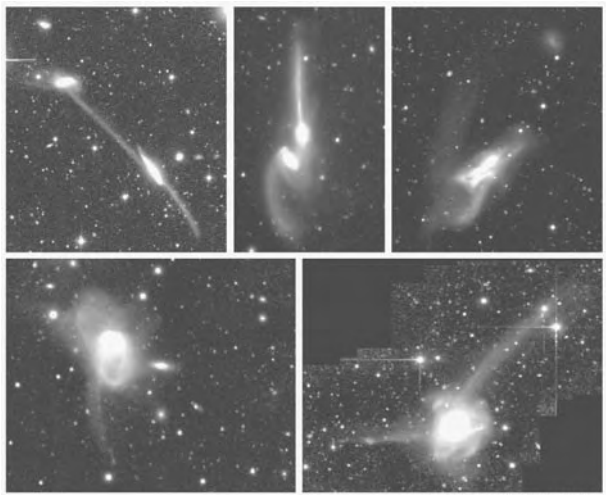


Figure 1. A montage of interacting and merging systems. Clockwise from upper left: Arp 295; NGC 4676 (The Mice); NGC 520; NGC 7252; NGC 3921. Image from Hibbard and van Gorkom 1996 *Astron. J.* **111** 655.

To understand the development of tidal tails, recall that tidal forces—the differential gravitational forces from a nearby mass—act to stretch an object radially. In this manner, our own Moon raises bulges on the surface of the Earth which give rise to our oceanic TIDES. Similarly, when two galaxies experience a close encounter, their tidal fields stretch one another radially. This stretching, combined with the galaxies' rotation, causes the stars and gas in the outskirts of each galaxy to 'shear off' from their parent galaxies. Material on the far side of each disk—away from the companion galaxy—is ejected into long, thin tidal tails, while material on the near side is drawn towards the companion. Depending on the encounter geometry, the nearside material may actually form a physical 'bridge' between the galaxies, along which material may flow from one galaxy to another. Because of the coupled effects of gravitational tides and galactic rotation, the development of tidal features depends strongly on resonances between the rotational and orbital motions of the galaxies. Prograde encounters, in which the galaxies' sense of rotation and orbital motion are matched, are most effective at tail-building. The lack of spin–orbit resonances in retrograde encounters, on the other hand, acts to suppress the formation of tidal tails.

The early computer models of Toomre and Toomre demonstrated these basics of building tidal tails and opened the door to the use of computer simulation to study the evolution of interacting galaxies. Because of the long timescales involved in galaxy interactions—hundreds of millions or billions of years—observations of interacting systems show only individual 'snapshots' of a complex evolutionary process. Piecing together these varying snapshots into a coherent sequence is a task made difficult by the unknown initial conditions of the different interactions. However, computer simulations

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can be used to study a wide variety of well-defined galaxy interactions, and by studying the evolution of these models the *observational* snapshots can be more easily placed into a evolutionary context.

Computer simulations employ a Lagrangian technique known as N-body modeling, wherein galaxies are represented by N discrete particles whose initial positions and velocities sample the phase space distribution of a normal disk galaxy. Current simulations employ 10^6–10^7 particles, so that each particle represents 10^4–10^5 stars (simulating galaxies star by star is still well beyond current computational abilities). The N-body model is then advanced forward in discrete time steps by calculating the net gravitational acceleration acting on each particle from the other particles and then advancing each particle forward in time given its position, velocity and acceleration (for the interstellar gas, hydrodynamic forces must also be considered; these are discussed in a later section).

One example of such a computer model of colliding galaxies is shown in figure 2. In this model, two equal-mass spiral galaxies are placed on an initially parabolic orbit with a closest approach of $2.5h$, where h is the exponential scale length of the disk stars. One disk is perfectly prograde while the second is inclined by 71° to the orbital plane. As the galaxies first collide, the tidal forces act to distort each disk, launching the tidal tails. Moving at parabolic velocity, the galaxies quickly pass one another and move apart, reaching maximum separation at $T \sim 40$ (unit time is roughly 10 million years). After this point, the galaxies reverse their motions and fall back on one another, merging together to form a ellipsoidal merger remnant around $T \sim 60$.

After first passage, in addition to forming the extended tidal tails, the galaxies themselves develop a strong internal response to the interaction. The passage of a companion seeds an $m = 2$ gravitational perturbation in the disk of each galaxy; these perturbations can then be amplified into strong spiral arms or even dramatic central bars by the self-gravity of the disk. The detailed response of the galaxies depends crucially on their internal structure: if the stellar disk dominates the mass distribution in the inner portion of the galaxy, the disk self-gravity can easily amplify the perturbation into a strong bar. On the other hand, if the galaxy is bulge or dark matter dominated in its interior, the disk is more stable against growing $m = 2$ bar modes, and spiral structure is the more likely outcome.

What causes the rapid merging of the two galaxies? Certainly there is a transfer of orbital energy to internal motions of the stars in each galaxy—it is this energy transfer that provides the energy necessary to launch the tidal tails. The amount of energy carried away by the tidal tails amounts to only a few per cent of the total orbital energy, however; if this was the sole energy sink, interacting galaxies would slowly spiral together over many orbital periods, rather than exhibit the rapid merging seen in figure 2. Indeed, early investigations into the collisions of galaxies argued that subsequent mergers would be quite rare, unless the collision happened to be extremely penetrating.

In fact the luminous portions of galaxies represent only a small fraction of the total galactic mass—galaxies are embedded in massive, extended 'dark matter halos'. These dark matter halos extend to many tens or even hundreds of kiloparsecs, such that encounters where the luminous galaxies seem to pass by one another may in fact be penetrating encounters for the dark halos. When this happens, dynamical friction ensures rapid orbital decay. As the galaxies pass through each other's dark halos, they set up a trailing wake in the halo mass distributions, creating a gravitational drag on the galaxies' relative motion. As a result, energy and angular momentum are transferred from the binary orbit to the internal motions of the dark halo—the orbit decays and the halos are spun up. It is this ability for dark matter halos to absorb orbital energy and angular momentum that makes galaxy mergers possible; without halos, mergers would be rare indeed (see DARK MATTER IN GALAXIES).

Starbursts and active nuclei in interacting galaxies

With the launch of the infrared astronomy satellite (IRAS) in 1983, a new population of galaxies was identified which have extremely high, quasar-like luminosities ($\geq 10^{12}\ L_\odot$) in the infrared (8–1000 μm). This emission is believed to come from hot ($T \sim$ 50–60 K) dust reradiating energy from a central starburst or active nucleus. The exact details of the energy generation—starburst versus AGN—remain controversial, and many ultraluminous infrared galaxies (ULIRGs) show evidence for both starburst *and* AGN activity. The difficulty of course lies in the fact that the nuclei of ULIRGs are, almost by definition, extremely dusty environments, and seeing into the central regions is a very difficult task. Evidence suggests that in some cases the nuclear dust is optically thick even at x-ray wavelengths; finding unequivocal evidence of an AGN in such systems will be very hard indeed.

While the details of the energy source in ULIRGs remain elusive, the causal connection between ULIRG activity and galaxy interactions is clear. Nearly all ULIRGs show morphological evidence for strong interactions, through double nuclei, severely distorted isophotes or strong tidal features. Yet the converse is *not* true; that is, not all strongly interacting galaxies show ultraluminous activity. In fact, the most dramatic nearby mergers, such as the Antennae (NGC 4038/39) or NGC 7252, have experienced violent collisions, yet have infrared luminosities an order of magnitude less than the extreme ULIRGs. Clearly the detailed triggering mechanism for ultraluminous activity must depend on a variety of factors.

How do mergers trigger ultraluminous activity? Regardless of whether the central engine is a starburst or an AGN, the first prerequisite is a large supply of fuel in the form of nuclear gas. If galaxy collisions can drive gas inwards from the galaxies' disks into their nuclei, this nuclear inflow can feed the central engine. Once the gas reaches the inner kiloparsec, it can either fragment

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

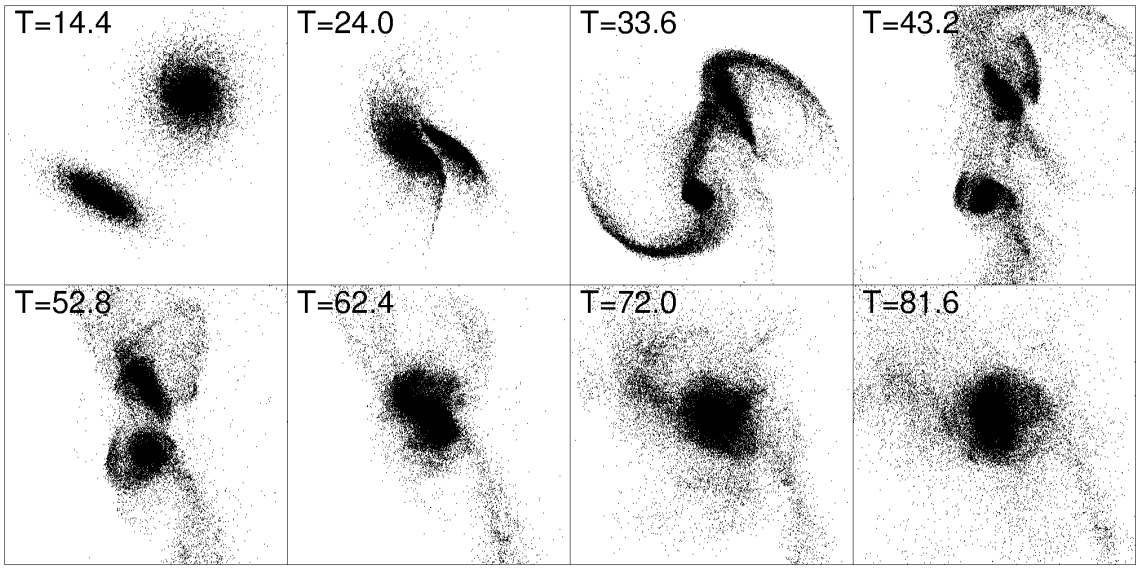


Figure 2. An N-body model of two equal-mass merging disk galaxies. Time is shown in the upper left of each panel, where unit time is ∼10 million years. After first collision at $T \sim 24$, the galaxies pass by one another before turning around on their orbits and merging to form an elliptical-like object in the last panel. From Mihos and Hernquist 1996 *Astrophys. J.* **464** 662.

and form stars (at a rate of $\sim 100 M_\odot$ yr^{-1} to provide the observed luminosity) or continue to flow inwards and fuel an AGN.

The first step, then, is to drive gas inwards from the disk to the nucleus. Computer simulations have demonstrated the efficacy of interactions and mergers at driving these nuclear gas flows. These simulations must also describe the hydrodynamic evolution of the interstellar gas, and do so using a Langrangian technique known as 'smoothed particle hydrodynamics', in which the gas is represented by discrete fluid elements (particles) which carry the local thermodynamic and hydrodynamic properties of the fluid. These properties are updated according to hydrodynamic conservation laws, and artificial viscosity is included to model shocks. This technique is ideal for simulating gas flows in interacting galaxies, which lack the symmetry and geometry needed for efficient grid-based hydrodynamical algorithms.

As an example, figure 3 shows the evolution of the interstellar gas in the merger model of figure 2. Shortly after the galaxies first collide, and the strong bars and spiral features form from the self-gravitating response of the disks, gas shocks and crowds along the leading edge of these features. Because of the offset between the stellar and gaseous density peaks, the gas feels a strong gravitational torque from the stars, losing angular momentum and flowing inwards towards the nucleus. How far inwards it can flow depends on the detailed response of the disk. If the disk develops a strong bar, the gas can flow into the central kiloparsec on a dynamical timescale; however, if the disk is stable against bar formation, the gas tends to 'hang up' in the inner few kiloparsecs. In this case, once the galaxies ultimately merge a second phase of inflow occurs, as gas is driven further inwards to the nucleus owing to the strong hydrodynamic and gravitation torques at work.

Because the inflow is tied to the (largely internal) gravitational response of the galaxies to the interaction, the triggering of inflow and nuclear activity depends on a variety of factors, such as the internal structure of the galaxies involved and the orbital geometry of the collision. Prograde collisions drive inflow and activity more rapidly than retrograde collisions, owing to the spin–orbital coupling of the encounter. If the encounter involves DISK GALAXIES with massive central bulges, the inflow should be delayed until the galaxies merge, owing to the disk stability provided by the central bulges. Conversely, without bulges, disk galaxies should be more prone to early bar formation, inflow and central activity. Similarly, galaxies in which the disk contributes little to the dynamical mass (e.g. low surface brightness disk galaxies) will also be more resistive to inflow, as the self-gravity of the disk is much weaker. Of course, in any case the galaxies themselves must have sufficient disk gas to drive inwards to begin with; collisions of gas-poor spirals will be less effective at fueling ultraluminous levels of activity.

Once triggered, the ultraluminous infrared phase probably does not last long—less than $\sim 10^8$ yr. A variety of arguments support this claim. Stellar population synthesis models successfully explain the optical and infrared spectra of ULIRGs with a burst of star formation $\sim 10^7$–10^8 years old. Similar numbers come from gas depletion arguments: if the luminous activity arises from star formation, the gas depletion time is given by $\tau_{\rm gas} \sim M_{\rm gas}/\dot{M} \sim 10^7$–$10^8$ yr. Of course, if the luminosity comes from accretion onto an AGN, the same amount of gas can sustain ultraluminous activity for a much longer period

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

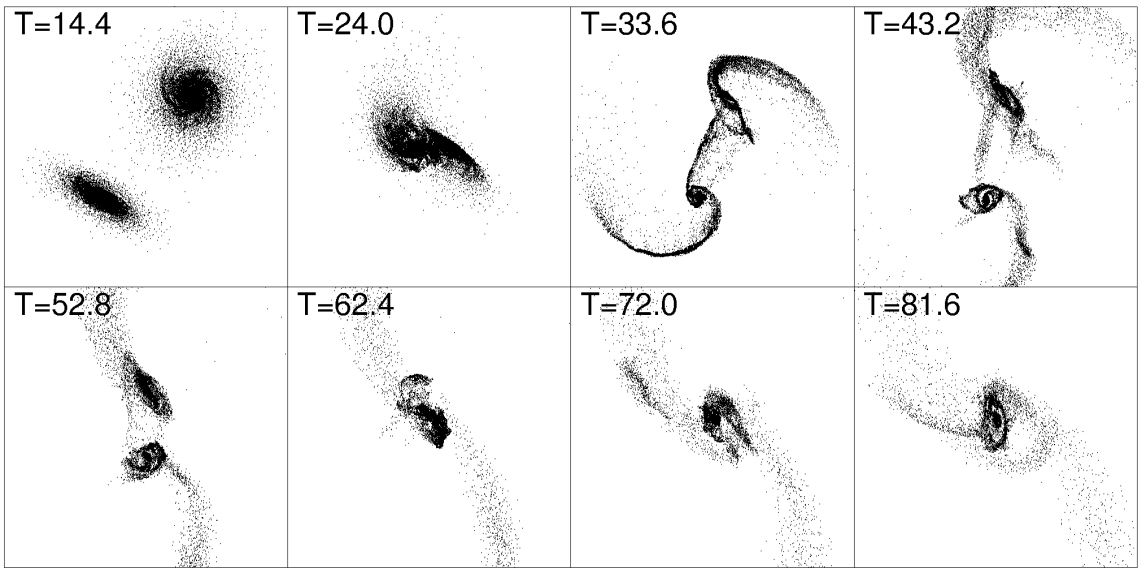


Figure 3. The evolution of the gas in the *N*-body model shown in figure 2. Time is shown in the upper left of each panel, where unit time is ∼10 million years. After the first collision, gas shocks along spiral arms and flows inwards into the inner few kiloparsecs ($T \sim 48$) and then finally into the nucleus itself when the galaxies ultimately merge. From Mihos and Hernquist 1996 *Astrophys. J.* **464** 662.

of time. However, the fact that nearly all ULIRG systems show strong dynamical evidence that they are in the late stages of a merger—evidence which fades rapidly once the merger is complete—argues that the ultraluminous phase cannot last much longer than a half-mass dynamical timescale, $\sim 10^8$ years. In the latter case of AGN accretion, the end of the ultraluminous IR phase may be marked by either an end to AGN fueling or else the destruction of the 'dust shroud' by the intense UV radiation field of the AGN. In this second scenario, rather than running out of gas, the ULIRG may simply evolve into a *bona fide* optical quasar. The detailed evolution of ULIRGs is a subject of intense current study.

Merger remnants and elliptical galaxies

Once a merger is complete, and the ULIRG phase (if any) is over, what is left behind? In 1977, Alar Toomre suggested that the remnants of disk mergers could account for the population of elliptical galaxies in the universe. Because of violent relaxation, the merging process would effectively 'scramble' the stellar disks, giving the remnant the surface brightness profiles and large velocity dispersions characteristic of elliptical galaxies. Toomre argued that if the merging timescale is $\sim 5 \times 10^8$ yr, extrapolating the number of nearby on-going mergers (∼10) over the total age of the universe (assuming the merger rate scales like the $t^{5/3}$ expected from a flat distribution of binding energy) yields a total of ∼750 merger remnants, similar to the number of elliptical galaxies found in the nearby field. Furthermore, Toomre argued, if the remnants of these spiral mergers did not constitute the present-day elliptical galaxy population, where are they now?

This 'merger hypothesis' for the formation of elliptical galaxies idea also had observational support from studies of the peculiar galaxy NGC 7252 (see figure 4) by François Schweizer. While this galaxy possesses two gas-rich tidal tails (indicating a merger of two late-type spirals), it also has the surface brightness profile expected for an elliptical galaxy. Schweizer argued that we were catching the elliptical formation process in the act and that NGC 7252 would evolve into a normal elliptical galaxy given time.

While the merger hypothesis could explain many of the qualitative properties of ellipticals, several detailed objections were raised. The central phase space density of spirals is lower than that of ellipticals; since violent relaxation preserves phase space density, merger remnants should have phase space densities too low to compare well with elliptical galaxies. Also, the specific frequency of GLOBULAR CLUSTERS (the number of globular clusters per unit luminosity) is much larger for ellipticals than spirals. Finally, because of the high relative velocity of galaxies in clusters, mergers should be less common in these environments, yet that is where the elliptical fraction is highest. How then could mergers of disk galaxies produce the present-day elliptical galaxy population?

Recent developments have relieved some of these concerns. While it is true that stellar phase space density is preserved during mergers, this constraint does not apply to the interstellar gas, which can dissipate energy and flow inwards. Subsequent star formation can raise the central phase space density in the merger and in principle produce an elliptical-like nucleus. Interaction-induced star formation may also account for the differing globular cluster specific frequencies of spirals and ellipticals.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

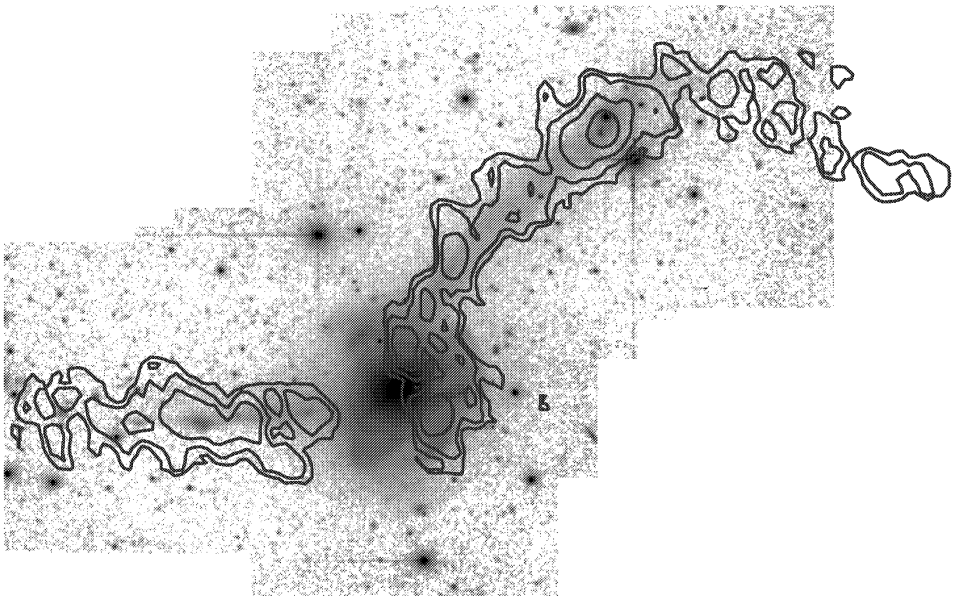


Figure 4. The merger remnant NGC 7252. The image shows the optical light from NGC 7252, while the contours show the distribution of neutral hydrogen gas. Note the gas-rich tails, the loops of starlight surrounding the main body and the smooth, elliptical-like appearance of the inner regions. From Hibbard *et al* 1994 *Astron. J.* **107** 67.

Recent Hubble Space Telescope observations indicate that galaxy collisions can result in the formation of a large number of young, compact star clusters. If these clusters can survive the dynamical environment of tidal shocks and mass loss, perhaps they are the progenitors of the old globular clusters observed in elliptical galaxies. Finally, the problem of the high relative velocity of galaxies in clusters can be surmounted by realizing that structure forms hierarchically. Galaxy clusters are the last objects to form; galaxies first assemble into smaller groups which fall together to form a larger galaxy cluster. The galaxy mergers which produce the current cluster ellipticals may have occurred long ago in small groups (with low velocity dispersions), before the cluster formed as a single virialized structure.

The merger hypothesis thus remains a viable mechanism for the formation of elliptical galaxies. We can now ask whether the merging process leaves behind any signatures which can be used to constrain the fraction of ellipticals which formed in this manner. The nuclear starbursts should result in the formation of a central population of stars which is distinct in a number of ways from the older stars in the original galaxies. This population may manifest itself through features ('seams') in the surface brightness profiles of ellipticals, through strong age and metallicity gradients, and through kinematically distinct nuclear kinematics. On larger scales, unless the starburst completely processes the interstellar gas into stars, a small but significant amount of cold gas may remain in the remnant. In the outer portions of the remnant, the relaxation timescale is long and low surface brightness tidal feature—'shells' and 'loops'—can survive for many billions of years, indicating a strong collision in the galaxy's past. Observationally, a significant fraction of elliptical galaxies show kinematically distinct cores, residual gas and dust, and large-scale tidal features. Through studies such as these, it may be possible to constrain the fraction of ellipticals formed through disk mergers.

Of course, in a universe where structure forms hierarchically, all objects form from the coalescence of smaller objects. From this point of view, it may be wiser not to ask 'did mergers form ellipticals?'. but rather 'what merged to form ellipticals?'. Observational studies have shown that cluster ellipticals must have formed very early, perhaps even before massive disk galaxies had formed. If so, the merging objects were probably very different from the types of galaxies we see involved in nearby mergers. Because the collapse time for structure is faster in high-density environments, cluster and field ellipticals have a different formation timescale, and (under the merger hypothesis) should have different progenitors. Cluster ellipticals may form from the rapid assembly of many small progenitors, while field ellipticals may form through the more classical merger hypothesis picture of two merging spiral galaxies. These different histories should manifest themselves through differences in the structural, chemical and stellar population mix between cluster and field samples. The observational data, however, do not show as strong an effect as that predicted by models of hierarchical structure formation. The question of whether most ellipticals were formed through major merger events remains open.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Minor mergers and satellite accretion

So far we have focused on so-called 'major mergers'—mergers of two large, roughly equal-mass disk galaxies. While these represent some of the most spectacular collisions we observe, they are far from the most common. Because of the shape of the galaxy luminosity function, which rises at fainter luminosities, encounters more commonly involve a galaxy interacting with a small satellite companion. However, it turns out that even these 'minor mergers' can have a dramatic impact on the evolution of galaxies.

That minor mergers occur frequently can be seen from even a casual inspection of the environment of our own MILKY WAY. The Milky Way is surrounded by 14 known satellite galaxies, the largest and most massive being the LARGE and SMALL MAGELLANIC CLOUDS. At a galactocentric distance of 50 kpc, the Magellanic Clouds are believed to orbit our Galaxy once every few billion years. The effect of this interaction is clear: tidal forces from the Milky Way have torn a long stream of gas—the MAGELLANIC STREAM—from these companions, and it has been argued that some of the younger Milky Way globular clusters may have been tidally stripped from the Magellanic Clouds. Recently, the discovery of the SAGITTARIUS DWARF GALAXY plunging through the galactic disk on the far side of the Galaxy has given us an up-close view of the tidal destruction of a merging dwarf galaxy. A similar system of companions surrounds the nearby Andromeda Galaxy, including the bright dwarf elliptical galaxies NGC 205 and M32. Indeed, if models of structure formation in the universe are to be believed, all bright galaxies should have a large number of satellite companions.

Unlike major mergers, where dynamical friction is so efficient that the galaxies merge after only a few perigalactic passages, the extreme mass ratio (>10:1) of minor mergers ensures that the orbital decay of the satellite's orbit is slow. For example, it is estimated that the orbit of the Magellanic Clouds will decay in another ∼10 Gyr, giving the Clouds a total 'survival time' of ∼10 orbital periods. Rather than the sudden violence of a major merger, these minor mergers are more properly a relatively slow accretion event, and the dynamical evolution of these systems is markedly different. In particular, the tidal field of the host galaxy will act to gradually strip material from the orbiting companion, such that the merger may not play to completion—the companion may be totally destroyed before reaching the inner portion of the host galaxy. Since the tidal radius of a orbiting companion is roughly proportional to $(\rho_s/\rho_G)^{1/3}$ where ρ_s and ρ_G are the average mass density of the satellite and host galaxy, respectively, it is the low-density, diffuse satellite companions which are preferentially destroyed as they orbit their host. Low-density DWARF IRREGULAR and DWARF SPHEROIDAL GALAXIES will certainly be tidally disrupted at relatively large distances from their hosts, while compact companions (such as Andromeda's M32) may survive the accretion process well into the host.

As the satellite is stripped on its orbit, this stripped material can be incorporated into the luminous halo of the main galaxy. Over several orbits, this stripped material will gradually spread out along the satellite's orbital path, making a kinematically distinct 'tidal stream' in the halo. In 1978, Leonard Searle and Robert Zinn proposed that, rather than forming in a single monolithic event, the halo of the Milky Way formed through continual accretion of material much like the infall of satellite galaxies. To reproduce the observed halo luminosity of the Milky Way ($L_B \sim 10^9 L_\odot$), the Milky Way must have accreted on the order of a few hundred small satellites during its lifetime to build the halo in this manner. Whether or not such an evolutionary picture is consistent with the Milky Way's dynamically cold disk (see below) or with the long dynamical friction timescales involved remains unclear, but a variety of observations argue that some fraction of the Milky Way's halo—and presumably those of other spirals as well—formed through satellite mergers. Observations of halo stars show the presence of substructure, moving groups and possible streams of globular clusters and dwarf galaxies in the halo. The presence of young A stars in the otherwise-old halo also argues for recent accretion events, while tidally stripped material from the Sagittarius dwarf extending out more than 30° from the core demonstrates explicitly how satellite accretion can feed the Galactic halo.

If the satellite is sufficiently dense and massive, its accretion can do more than simply build up the stellar halo of a galaxy. Such satellites more readily survive into the inner galaxy; once there, they can strongly influence the dynamical evolution of the disk. An N-body simulation of such an accretion event is shown in figure 5, where a prograde satellite with mass 10% that of the disk merges with a larger disk galaxy. As this massive satellite plunges through the disk, it scatters disk stars off their orbits, heating and thickening the stellar disk. As the satellite galaxy falls to the plane of the disk through dynamical friction, conservation of total angular momentum dictates that the disk warp and tilt in response. The resulting warped, thickened disk is qualitatively similar to those seen in some nearby galaxies (such as NGC 3628) and inferred in our own Milky Way.

As a massive satellite spirals into the inner disk, it can also induce strong spiral arms and bars through a resonant response in the disk. Interstellar gas can be compressed along these features and flow inwards towards the nucleus owing to gravitational torquing from the bar and arms, similar to, but at a more modest level than, the major-merger-induced nuclear inflows. Thus these minor mergers can also drive increased disk and nuclear star formation rates or activate a central engine. When the merger is complete, a significant fraction of the disk gas may have been swept from the disk and into the nucleus. The disk itself has been kinematically heated, and spiral structure destroyed, by the passage of the companion through the inner disk. The induced bar may buckle in three-dimensional space, scattering stars above and below the disk plane, in a pseudo-bulge. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

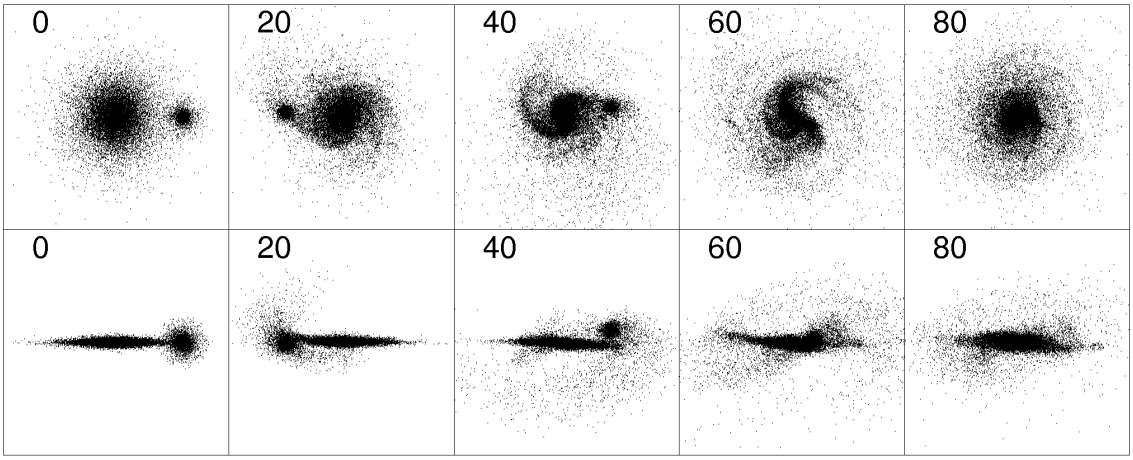


Figure 5. An N-body model of a large disk galaxy accreting a small satellite companion. The companion is initially on a circular orbit around the primary, inclined 30° from the disk plane. Top row shows a face-on view of the accretion; bottom row shows an edge-on view. As the satellite falls into the inner regions of the host, it excites strong spiral arms and a central bar, and thickens and warps the disk. Note also the stripping of material from the companion as it is accreted. From Hernquist and Mihos 1995 *Astrophys. J.* **448** 41.

morphological, kinematic and star-forming properties of the disk have all been radically altered by this interaction, suggesting that even minor mergers can drive significant evolution in disk galaxies.

Bibliography

The subject of galaxy interactions covers such a wide range of topics that it is impossible to give a comprehensive literature review. Instead, I have attempted to outline some of the original material in the field, while also giving more recent reviews on the broader subjects.

Recent reviews on the subject of galaxy interactions include

Barnes J and Hernquist L 1992 *Ann. Rev. Astron. Astrophys.* **30** 705

Kennicutt R C, Schweizer F and Barnes J E 1998 *Galaxies: Interactions and Induced Star Formation* (Berlin: Springer)

Some of the first detailed N-body calculations of interacting galaxies were done by

Toomre A and Toomre J 1972 *Astrophys. J.* **178** 623

while more modern calculations can be found in

Barnes J and Hernquist L 1996 *Astrophys. J.* **471** 115

Mihos J C and Hernquist L 1996 *Astrophys. J.* **464** 662

The detailed properties of ULIRG galaxies, and the connection between mergers, ULIRGs, and quasars was first described in

Sanders D B *et al* 1988 *Astrophys. J.* **325** 74

and have been recently reviewed in

Sanders D B and Mirabel I F 1996 *Ann. Rev. Astron. Astrophys.* **34** 749

The theoretical motivation behind the merger hypothesis for elliptical galaxy formation can be found in

Toomre A 1977 *The Evolution of Galaxies and Stellar Populations* ed B M Tinsley and R B Larson (New Haven, CT: Yale Observatory Press) p 401

while some of the first observational groundwork is in

Schweizer F 1982 *Astrophys. J.* **252** 455

The accretion model for the galactic halo was originally proposed by

Searle L and Zinn R 1978 *Astrophys. J.* **225** 357

while a recent review of observational and theoretical work on the structure and formation of the halo is given in

Majewski S R 1996 *Ann. Rev. Astron. Astrophys.* **34** 749

Models of satellite infall can be found in

Hernquist L and Mihos J C 1995 *Astrophys. J.* **448** 41

Johnston K V, Spergel D N and Hernquist L 1995 *Astrophys. J.* **451** 598

Walker I R, Mihos J C and Hernquist L 1996 *Astrophys. J.* **460** 121

C Mihos

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy

A large assemblage of stars, gas and dust that is held together by the mutual gravitational interaction between its constituents. Galaxies contain between a few million and about ten trillion (10^{13}) stars together with differing proportions of interstellar material (gas and dust).

Most galaxies can be categorized as *normal* galaxies, but a few percent (those that exhibit unusual characteristics such as having a compact, variable nucleus, radiating predominantly non-stellar radiation, and having a disturbed appearance) belong to the category of *active* galaxies. Normal galaxies may be classified in a number of ways, the most common system being the Hubble classification scheme, which divides galaxies into the following principal types: elliptical, spiral, barred spiral and irregular. Elliptical galaxies are spherical or ellipsoidal in shape, are dominated by old red stars and contain very little interstellar gas. Spiral galaxies consist of a central nucleus of relatively closely packed stars surrounded by a flattened, disk-shaped distribution of stars, gas and dust, within which luminous gas clouds and the brightest, youngest, stars are gathered into spiral-shaped 'arms' that appear to spread out from the nuclear region. In barred spirals, the nucleus is elongated, or bar-shaped, and the spiral arms emanate from either end of the nuclear bar. Irregular galaxies display no obvious pattern or structure. Spirals and barred spirals contain a larger proportion (typically about 10%) of gas and dust than ellipticals. The gas content of irregulars varies considerably, but is typically about 20%.

The masses of galaxies range from a few million solar masses, in the case of dwarf ellipticals, to a few trillion (10^{13}) solar masses, in the case of the most massive ellipticals. Their luminosities range from about 200 000 to a trillion (10^{12}) times that of the Sun and their diameters mostly range from a few thousand light-years to a few hundred thousand light-years. Although most of the ellipticals are relatively small, 'giant' ellipticals are found in the cores of clusters of galaxies, a subset of these, the 'supergiant' (cD) galaxies, having diameters of up to five million light-years.

Measurements of the rotation curves of spiral galaxies (graphs of the velocities at which stars and gas clouds revolve at different distances from the centers of galaxies) indicate that stars and gas clouds in the outer regions of galaxies are moving as fast as, or faster than, those that are located closer to their centers. These results imply that up to 90% of the mass of a galaxy consists of dark matter (matter that emits no detectable radiation but exerts a gravitational influence), much of which lies in its outer regions.

The galaxy to which the Sun belongs is known as the Galaxy (with a capital 'G'), or the Milky Way Galaxy. It is a spiral or barred spiral system that contains more than 100 billion stars and which has an overall diameter (measured across its disk) of about 100 000 light-years.

See also: active galaxy, barred spiral galaxy, cluster of galaxies, dark matter, dwarf galaxy, elliptical galaxies, galactic nucleus, Hubble classification, irregular galaxy, lenticular galaxy, Milky Way Galaxy, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Clusters

Galaxy clusters are the largest stable structures in the universe. Galaxy clusters have been key astrophysical objects in the development of our current understanding of a host of issues.

- Dark matter was first detected in galaxy clusters.
- Galaxy clusters contain gas at a temperature of about 10^8 K and are very luminous x-ray sources.
- The theory of gravitational instability, which is the current paradigm for the formation of all structure, was first developed to understand galaxy clusters.
- Clusters contain a unique mixture of galaxy types, including the biggest and many of the oldest.

Galaxy clusters were first tentatively recognized in the sky surveys of NEBULAE in the 19th century. The two most prominent in the northern sky are the VIRGO CLUSTER, which is the approximate center of the local supercluster at a distance of about 20 Mpc from the Milky Way, and the COMA CLUSTER (figure 1), which is a spectacularly 'rich' cluster at REDSHIFT of about 6900 km s^{-1}. The Virgo cluster lies near the north Galactic pole and provides us with a convenient local sample of a relatively low-richness cluster. Coma is also located in the north Galactic cap and stands out as an unusual structure in the whole northern extragalactic sky.

When galaxy clusters were first noticed in the 1800s, the size and structure of our own Galaxy was not known, let alone the distinction between various types of nebulae in our Galaxy and the extragalactic nebulae, which turned out to be other galaxies. Hubble's distance scale of the late 1920s was an essential underpinning to the physical investigation of galaxy clusters. The first step was momentous. ZWICKY measured the velocities of individual galaxies in the Coma cluster. Then, making the somewhat bold assumption that the cluster was in a stable equilibrium, he derived its total mass. The ratio of the mass to the total luminosity of cluster galaxies is an astonishing value of about $200M_\odot$ per $L_\odot$ (adjusted to current-day values of the HUBBLE CONSTANT). The Sun itself has a mass-to-light ratio of unity, and mixes of stars in an average galaxy have M/L values ranging from about 1 to about 10, depending on their Hubble type. Zwicky's conclusion was that most of the mass was 'missing' from the census of visible light. This conclusion was controversial until the 1970s, the primary issue being the dynamical state of clusters which was unclear partly because there was no clear theory of how such massive objects could come together. In the late 1960s Peebles (1970) did one of the first n-body simulations to show how a cluster could form from small density fluctuations in an expanding universe. Then, with the first imaging x-ray telescopes in space it was conclusively established that large clusters of galaxies contained gas at a temperature of about 10^8 K. Such hot gas required that there be a large, deep potential well to keep it contained. X-ray observations of clusters have become one of the major tools for first finding galaxy clusters and for investigating the structure of the DARK MATTER distribution.

In the 1980s GRAVITATIONAL LENSING was discovered in galaxy clusters, more or less as predicted by Zwicky. This elegant technique is both simple and powerful. The data consist of nothing other than a picture of a cluster at a moderate redshift, say roughly 1/2, where the lensing effect has a reasonable 'lever arm'; the picture will have a high density of background galaxies without too many foreground sources (which suffer no lensing effect). The gravitational field of the cluster causes the background galaxies to be slightly tangentially elongated (on the average), which directly measures the mean surface density of the cluster. Very deep images with exceptionally high angular resolution (such as from the Hubble Space Telescope) reveal more detail which can be used to investigate smaller-scale structure in the dark matter distribution.

A famous theoretical paper of Sunyaev and Zeldovich (SZ) predicted that as COSMIC MICROWAVE BACKGROUND (CMB) radiation passes through a cluster its radiation would be affected in two ways. First, the very hot electrons in the cluster would Compton scatter the 3 K photons of the CMB, leading to a slight increase in the energy of each photon. If the cluster is moving with respect to the CMB, then there is an additional Doppler effect of light being red- or blueshifted by reflection from a moving mirror. By comparing the CMB radiation near to the cluster with that seen coming through the cluster the SUNYAEV–ZELDOVICH EFFECT can be used to map the distribution of hot gas in clusters. This technique is best done at fairly short wavelengths where the confusing effects of radio sources are minimized. Radio telescopes that operate efficiently for these measurements at very high frequencies are now being constructed.

The galaxy content of clusters is unique. The GALAXY MORPHOLOGY–DENSITY RELATION (Dressler) states that the low-density field is dominated by spiral-type galaxies but as one moves to regions of ever higher density the fraction of spirals declines and elliptical and S0 galaxies dominate in regions of very high galaxy density, such as the centers of galaxy clusters. Spirals are actively forming stars whereas E and S0 galaxies have very low star formation rates. Understanding this relation is a key consideration in theories of GALAXY FORMATION. Although E and S0 galaxies are fairly rare outside of clusters, they do provide about 15% or so of the field galaxy mix. The cD galaxy type, a supergiant elliptical with an extended envelope, is unique to clusters. Relative to field galaxies, cluster galaxies are older and contain a larger fraction of their stars in a spheroidal component as opposed to a flat disk. One possible inference is that galaxy merging was more important in clusters than in the field.

Finding galaxy clusters

Clusters are rare objects, containing about 1% of all galaxies. Finding them efficiently with minimal biases remains a research problem. The problem is that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. The color image of the central region of the Coma cluster was taken at KPNO with 0.9 m telescope and T2KA CCD in the B, R and I bands as part of Omar López-Cruz's PhD thesis. The field of view covers 23 arcmin × 23 arcmin. Preprocessing was done with IRAF by López-Cruz and the color composition was done by Ian Shelton. **This figure is reproduced as Color Plate 22.**

method of finding the clusters can significantly influence the properties of the cluster sample. A trivial sampling bias is that practical samples are only complete to some limiting flux as measured in some waveband which usually, but not always, leads to a declining number of clusters with increasing redshift.

The Abell catalog

The Palomar Sky Survey of the 1950s provided the first nearly uniform optical coverage of the northern celestial sphere. Abell's 1958 paper describing his use of this huge database remains a classic of experimental design which in many ways has served as a prototype for other surveys for galaxy clusters. Abell's opportunity at the time was exceptional; however, the fact that his catalog has remained the fundamental low-redshift sample even today is testament to the care which he approached this problem at a time when the nature of galaxy clusters was far from clear. Earlier work of Zwicky at Caltech and Neyman, Scott and Shane at Lick had made it clear that separating galaxies into cluster and field populations was an ambiguous process. On the other hand, prominent clusters such as Virgo and Coma were readily identifiable physical objects, of which there must be other examples scattered throughout the universe. Abell devised a set of approximately redshift-independent criteria for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

identifying galaxies. The technique makes the reasonable and verifiable assumption that the tenth-brightest member has a statistically constant luminosity. From the apparent brightness an approximate distance can then be deduced, allowing clusters at different redshifts to be considered on the same basis. Abell made the inspired choice of 1.5 Mpc (for $H_0 = 100$) for the characteristic projected radius inside which he counted galaxies to measure the richness. For the range of cluster memberships he considers, from 50 to about 500, this is comparable with the 'virialization' radii of the clusters. The result is Abell's exceptionally durable list of about 3000 clusters, culled from approximately 10 000 square degrees of sky. From this catalog Abell draws the important conclusion that clusters are not randomly distributed (see ABELL CLUSTERS).

X-ray cluster catalogs

X-ray astronomy transformed the study of clusters. Since the Abell clusters have internal line-of-sight velocity dispersions ranging from about 500 to 1000 km s^{-1} any gas within them is, in equilibrium, heated to a temperature of between 10^7 and 10^8 K. Gas at these temperatures radiates primarily in the x-ray part of the spectrum, with characteristic energies of a few kiloelectronvolts (keV). The first x-ray satellites found that clusters were very bright, implying x-ray luminosities in the range from 10^{43} to 10^{45} erg s^{-1}. This x-ray luminosity is about 2–4 orders of magnitude less than the luminosity of all the visible galaxies. However, the galaxies are powered by nuclear energy whereas the x-ray gas ultimately derives most of its energy from its gravitational energy release as it falls into the cluster.

Normal galaxies in the field produce little x-rays. Therefore at low redshifts clusters of galaxies are readily found, since the field is effectively 'dark'. However, the x-ray sky is fairly 'bright', probably as a result of large numbers of x-ray-emitting galaxies and ACTIVE GALACTIC NUCLEI (AGN) at high redshift. However, these distant sources effectively act to provide a uniform background, upon which the clusters are superposed. On the whole, the result of x-ray cluster finding at low redshift is to confirm the reality of most of Abell's clusters. The x-ray data are not sensitive to the projection of unrelated field galaxies and so usually give a better indication of the overall size of the cluster.

X-RAY TELESCOPES have not had very large apertures. Consequently it is difficult to do deep surveys over large sky areas. The Einstein Medium Sensitivity Survey (EMMS) and the Rosat Deep Cluster Survey (RDCS) are both 'serendipitous' surveys, which take advantage of deep x-ray exposures of unrelated objects, such as M stars. After imposing strict limits on any association, one sorts through the 'other regions' of the x-ray image of such an object. EMSS examined about 800 square degrees and RDCS about 100 square degrees, but to an x-ray flux limit about an order of magnitude fainter. The EMSS clusters extend from about redshift 0.05 to 0.55. The RDCS clusters extend nearly up to redshift 1. However, both of these samples contain only about a dozen sources at redshifts beyond about half of their highest redshift.

The x-ray surveys helped to clarify the issue as to where galaxy clusters 'end'. They have a characteristic radius of about 1 Mpc, inside of which the gas is hot and the galaxies are on approximately random orbits, as opposed to infalling nearly radial orbits outside the cluster. This distinction between cluster and field leads to the result that about 1% of all galaxies are in rich clusters.

New x-ray satellites will probe to much higher redshifts. In particular there are plans for the XMM satellite to use x-rays to detect clusters to redshift 2, where most other techniques become unreliable and inefficient.

New techniques

The hot gas in galaxy clusters is also seen in projection against the 2.7 K CMB radiation. Zeldovich and Sunyaev (1969) pointed out that as the CMB photons traverse the cluster they will, on average, be heated slightly. Therefore the Planck spectrum of the CBR is boosted upward slightly in energy. The result in the Rayleigh–Jeans tail at low frequency is that the cluster appears to be 'cool' with respect to the surrounding sky. The temperature decrement is of order a few hundred microkelvin. The unique power of the technique is that the SZ effect has no redshift dependence and can in principle find clusters to very high redshift. In practice a serious limitation is the sensitivity and the sky coverage of existing mm-wave radio telescopes. Mapping even a single square degree is a considerable challenge today, but will become nearly routine with future facilities.

Weak gravitation lensing directly 'images' the dark matter in the universe (Kaiser and Squires 1993) based on the gravitationally induced distortions of distant galaxies viewed along a line of sight that passes close to a mass concentration. This technique obviously depends on having observations that reach to high enough redshifts that large numbers of galaxies behind the clusters of interest are reached. There is the practical complication that galaxies are themselves non-circular so one must observe sufficient numbers so that the shear induced by a cluster, say of order 10% image distortion, can be confidently detected. As large-area CCD cameras become available this technique will probably become very important.

The galaxy population of most clusters contains far more relatively red, 'early-type' galaxies than are present in the field. An imaging survey with color information (Yee *et al* 1999) therefore can seek out concentrations of these objects on the sky. Because the red sequence changes its color with redshift it also provides a redshift indicator. Large-area optical and infrared imaging surveys will make this technique feasible.

There should be a dramatic change in the availability of carefully defined samples of galaxy clusters to redshifts well beyond 1 in about the year 2000. Up to the year 2000 Abell's sample has stood alone as the only really large one. X-ray selected samples have found objects slightly beyond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

redshift 1, but the numbers beyond even redshift of 0.5 are only of order one dozen. These new samples will open up a huge new range of astrophysical investigations.

Galaxies in clusters

The properties and evolutionary scenarios of cluster galaxies are a large and significant subject in their own right. A fundamental observation is the morphology–density relation (Dressler 1980) where it is noted that spheroid-dominated 'early-type' galaxies are much more frequent in clusters than the disk-dominated 'late-type' galaxies which predominate in the field. Quantitatively there is a smooth trend of fraction of early type as a function of the local environmental density. This relation is an important clue that the clustering environment plays a significant role in determining the properties of galaxies. This relation implies that most clusters have a radial gradient in their galaxy populations, with the central region largely composed of early-type galaxies and the outer regions beginning to resemble the field in the mix of galaxy types and colors. The red ELLIPTICAL GALAXIES in the centers of clusters are generally agreed to be the oldest normal galaxies in universe, as indicated by their exceptionally red colors in most photometric bands, which reflects the continuum light of the dominant stars of the red giant branch, and the strength of various lines which are fairly well calibrated against a similar mix of stars that have individual age dates within our own Galaxy. Of course the age of the stars does not necessarily tell when the galaxy was assembled into a single unit. Clusters contain many examples of the S0 type of galaxy, which was originally suggested by Hubble as a hypothetical type needed to fill out his tuning fork diagram. The S0s resemble the larger bulge DISK GALAXIES, Sa and Sb, but have disks that are somewhat thicker and have essentially no spiral features.

A second important observation is that clusters contain an increasing fraction of star-forming, late-type galaxies with increasing redshift (Butcher and Oemler 1984). This is possibly related to the increasing numbers of blue galaxies in the field with increasing redshift, which, through infall, could mix in with the cluster galaxies. In both cases the detailed mechanisms leading to these observational effects need to be better understood, which has provided part of the motivation for building new telescopes.

Clusters, by definition, have a brightest cluster galaxy (BCG). Besides being extremely luminous, the BCG is generally found at the center of the cluster, which is probably a consequence of dynamical friction on the large mass of the BCG. Generally the BCG is an elliptical-type galaxy, quite frequently of the unusual subtype 'cD' (cluster dominant), which is an elliptical with an extended envelope about the normal surface brightness profiles of elliptical galaxies, de Vaucouleurs' $r^{1/4}$ relation. These envelopes of extended stellar light are sufficiently faint that its detailed properties have not been studied in any detail. Two intriguing possibilities for the origin of the envelope are that it formed out of a 'cooling flow' in the x-ray gas in an extended galaxy formation period and that it is composed of stars that were originally part of other cluster galaxies but subsequently tidally stripped from them by the strong gravitational field in the central region of the cluster.

Dark matter in clusters

The Coma cluster has been a Rosetta stone for understanding the contents of the universe, in particular as an indicator of the total mass and gas density in the universe. Zwicky (1933) obtained spectra of a number of galaxies in the cluster from which he deduced that the galaxies were probably moving at random in a virialized potential with a line-of-sight velocity dispersion, σ_1, of about 1000 km s^{-1}. He estimated the virial radius to be somewhat more than $1/h$ Mpc, leading to a virial mass estimate, $M_V = 3G^{-1}\sigma_1^2 r_v$, of about $10^{15} M_\odot$. The sum of the luminosities of the cluster galaxies was about $5 \times 10^{12} L_\odot$. These numbers indicate that the mean amount of mass per unit luminosity is about $200h^{-1} M_\odot/L_\odot$, whereas both observation and modeling indicate that the stars themselves have a mean mass-to-light ratio of about $10 M_\odot/L_\odot$. The discrepancy is huge, highly significant and confirmed many times over in independent observations. One is led inexorably to the conclusion that the bulk of the mass of clusters is of some unknown 'dark' form.

On the basis that clusters were large, massive systems whose physics is entirely dominated by gravity, Oort (1958) proposed that the mass-to-light ratios of galaxy clusters were representative of the universe as a whole. Arithmetic then indicates that the mean mass density of the universe is the luminosity density times the mass-to-light ratio. The Canadian Network for Observational Cosmology (CNOC) survey established that the virial mass was a reliable indicator of the total mass in a cluster and simultaneously measured the luminosity density of the field, finding that $\Omega \simeq 0.19 \pm 0.06$ (Carlberg *et al* 1996, 1997). The two interesting implications of this result are that the total mass of clusters is well beyond what could be formed by the baryons created in the big bang (Copi *et al* 1995) and furthermore that it is well below $\Omega = 1$, the value that a flat, matter-dominated universe would have and the simplest result of an inflationary cosmology.

Dark matter has only been detected by its gravitational influence in astronomical observations. If general relativity were seriously in error then dark matter need not exist. However, we hastily add that general relativity has been both experimentally and theoretically scrutinized at length with no flaws having yet emerged. The physical makeup of dark matter is completely unknown at this time, although there are a wide range of particle physics candidates for its possible constitution.

Hot gas in clusters

The discovery that clusters emitted prodigious amounts of fairly hard x-rays conclusively demonstrated that clusters were bound, substantially virialized structures. The x-ray emission is produced via thermal bremsstrahlung in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

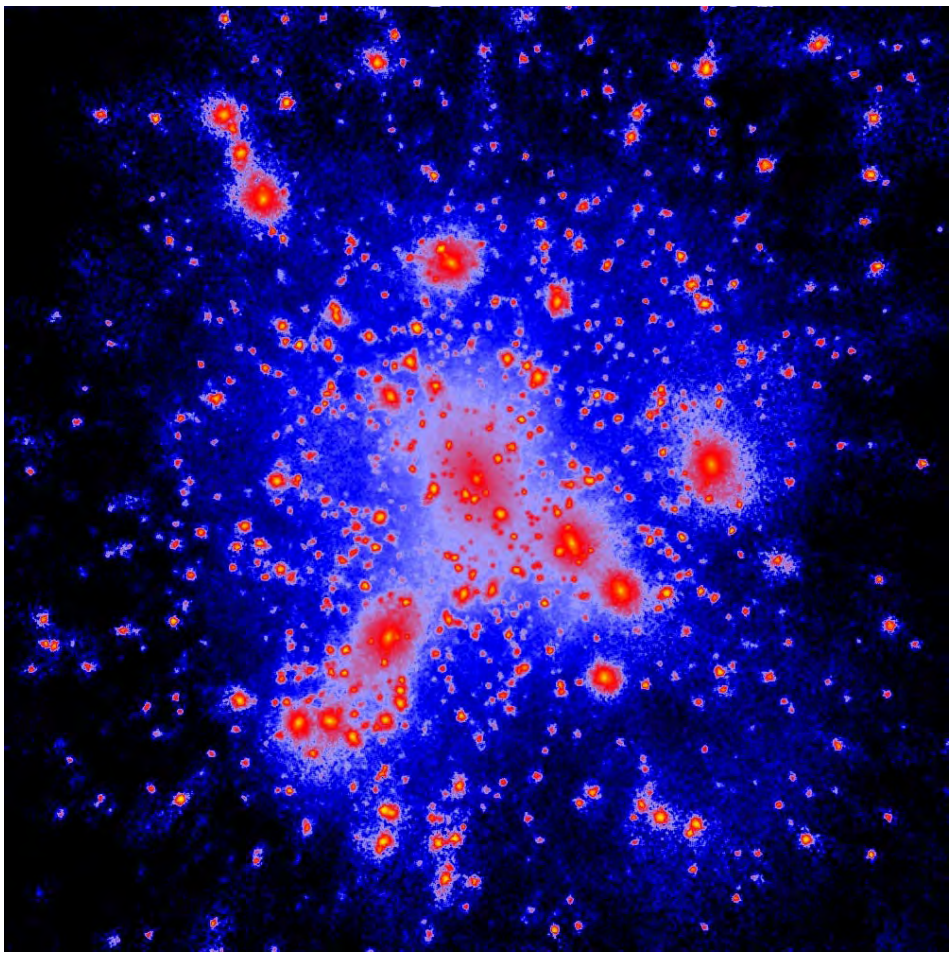


Figure 2. An n-body simulation of a cluster in a CDM model universe, courtesy of Ben Moore (University of Durham), Tom Quinn, Joachim Stadel, George Lake (University of Washington). The picture is color coded with the local density of dark matter. As the dark matter halos fall into the cluster their outer envelopes are tidally stripped off to join a cluster background but the very dense cores remain intact and orbit within the cluster. **This figure is reproduced as Color Plate 23.**

hot gas. The temperatures of around 10 keV imply that the electrons are slightly relativistic, which becomes an issue for some precision measurements of the state of the gas, particularly in the SZ effect. A public access computer code allows anyone to do a detailed computation of the emitted spectrum from an optically thin astrophysical plasma (ftp://heasarc.gsfc.nasa.gov/software/plasma_codes/raymond). One substantial result is that the x-ray temperature is identical (within the current errors) to the thermalized mean motion of the galaxies, indicating that both are in the same equilibrium with respect to the dark matter. X-ray emission drops as the density squared, meaning that the x-rays rapidly become hard to measure towards the edge of the cluster.

One of the most intriguing and controversial inferences from the x-ray data comes from the straightforward calculation of the cooling time of the cluster gas. In quite a few clusters the gas within the central few hundred kiloparsecs will cool in less than a Hubble time, in the absence of a heat source. The implied mass deposition onto the center of the cluster ranges up to as much as $1000 M_\odot$ per year. Where does this mass go? It seems very unlikely to be completely dark, and, in spite of evidence that there is some residual star formation in the BCG, it is far below the massive amount in the cooling flow. The increasing sensitivity and angular resolution of x-ray satellites should help resolve this issue.

As the CMB radiation with a Planck spectrum at 2.7 K passes through a cluster Compton scattering boosts the mean energy of the photons slightly (Zeldovich and Sunyaev 1969; Sunyaev and Zeldovich 1980). Even at the center of a rich cluster the decrement is slightly short of 1 mK. Of great value is that the thermal effect, which measures the electron pressure along the line of sight, $\int p_e \, dl$, has a frequency dependence that is different from the kinematic SZ effect, which depends on the peculiar velocity of the cluster with respect to the CMB. The result is that precision measurements of the flux difference from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the CMB as a function of frequency can provide a precise estimate of the cluster's peculiar bulk flow velocity, which is of great interest for constraining the amount of large-scale clustering in the universe.

Clusters also contain magnetic fields which become dynamically important near the centers of clusters (Kim *et al* 1990). The origin of this field is not known. If it is intrinsic to the cluster, as opposed to, say, being expelled from galaxies, it is quite puzzling.

The clustering of clusters

Abell noted that within his catalog of clusters the clusters had a tendency to lie near one another. Hauser and Peebles (1973) re-examined this and concluded that the clustering could be described with the two-point correlation function with a correlation length probably larger than that of galaxies, later refined by Bahcall and others. The strong clustering of clusters posed a puzzle, which ultimately helped inspire Kaiser's (1984) theory of biasing. The idea is quite simple, just that clusters are much rarer than galaxies themselves. Within a Gaussian random field the rare objects, such as galaxy clusters, will be more correlated than more typical objects. The clustering data are in quite good accord with the theoretical expectation. The clustering data are then consistent with a hierarchical theory for the formation of clusters from an expanding dark matter distribution. Much more work remains to be done on larger samples to check the theory in much more detail.

Formation and evolution of clusters

Clusters are such massive aggregations of matter that in the absence of a theory for their formation it was argued that they were simply chance superpositions of unbound galaxies. The idea of gravitational instability led to a plausible solution. At some early stage in the universe very small amplitude density fluctuations are imposed upon the dark matter. With time these grow as a power law, eventually reaching an overdensity of about 5 times the critical value, after which they begin to collapse. Both the amplitude and the timescales for this process work out quite reasonably.

In the early 1980s a particular fluctuation spectrum, the cold dark matter (CDM) spectrum, became a widely accepted basic description of many of the large-scale structural properties of the universe. The CDM spectrum flows from a consideration of how a scale-free spectrum created at some early time should be altered as the universe expands (see, for instance, Kolb and Turner 1990). Clusters, because they fairly accurately reflect the dark matter properties of the universe, are often used to normalize the amplitude of the spectrum on scales where the spectrum is strongly non-linear.

Beyond fairly simple analytic approaches such as the Press–Schechter theory (1974) for the numbers of clusters the most powerful approach to detailed predictions of the outcome of a CDM theory and clusters is to do a detailed n-body simulation (figure 2). Because clusters are relatively rare, one per $10^6 h^{-3}$ Mpc^3, one must simulate a large segment of the universe to find a single cluster (Colberg *et al* 1999) or do a specialized simulation of a small segment of the universe (Moore *et al* 1998). The agreement with the data is generally strikingly good, leading to an overall impression that the basic picture is in place and can serve as a powerful tool to test the cosmological model.

Although the basic physics of the hot gas in clusters is in place its detailed simulation and how the galaxies form remains a very active field of research. The combination of new x-ray satellites, precision SZ measurements, increased computer power and increasing sophistication of simulation will lead to rapid progress over the next decade or so.

Bibliography

Abell G O 1958 *Astrophys. J. Suppl.* **3** 211
Butcher H and Oemler A 1984 *Astrophys. J.* **285** 423
Carlberg R G, Yee H K C, Ellingson E, Abraham R, Gravel P, Morris S L and Pritchet C J 1996 *Astrophys. J.* **462** 32
Carlberg R G, Yee H K C and Ellingson E 1997 *Astrophys. J.* **478** 462
Colberg J M, White S D M, Jenkins A and Pearce F R 1999 *Mon. Not. R. Astron. Soc.* **308** 593
Copi C J, Schramm D N and Turner M S 1995 *Science* **267** 192
Dressler A 1980 *Astrophys. J.* **236** 351
Hauser M G and Peebles P J E 1973 *Astrophys. J.* **185** 757
Kaiser N 1984 *Astrophys. J.* **284** L9
Kaiser N and Squires G 1993 *Astrophys. J.* **404** 441
Kim K-T, Kronberg P P, Dewdney P E and Landecker T L 1990 *Astrophys. J.* **355** 29
Kolb E W and Turner M S 1990 *The Early Universe (Frontiers in Physics)* (Reading, MA: Addison-Wesley)
Moore B, Governato F, Quinn T, Stadel J and Lake G 1998 *Astrophys. J.* **499** L5
Oort J H 1958 *La Structure et l'Évolution de l'Univers Onzième Conseil de Physique* ed R Stoops (Brussels: Solvay Institute) p 163
Peebles P J E 1970 *Astrophys. J.* **75** 13
Press W H and Schechter P 1974 *Astrophys. J.* **187** 425
Suniaev R A and Zeldovich I B 1980 *Mon. Not. R. Astron. Soc.* **190** 413
Yee H K C, Gladders M D and Lopez-Cruz O 1999 *Photometric Redshifts (PASP Conf. Ser.)* ed Storri-Lombardi and Weymann (astroph/9908001)
Zeldovich Y B and Sunyaev R A 1969 *Astrophys. Space Sci.* **4** 301
Zwicky F 1933 *Helv. Phys. Acta* **6** 110

Raymond Carlberg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Evolution

The galaxies that we observe in the universe at the present time exhibit a remarkable variety of properties, such as morphology, colors, luminosity and dynamics. There are galaxies which are one hundred times brighter than our own Milky Way and others that have one thousandth its luminosity. Some galaxies have very regular shapes, other are amorphous. There are galaxies with spheroidal shape that have little amount of rotation, and whose stars have large 'thermal motions' (like the molecules of a hot fluid), and galaxies that resembles rotating disks, where self-gravity and centrifugal force balance each other. There are galaxies with high amounts of gas (mostly hydrogen and helium) and dust, and a vigorous activity of STAR FORMATION. Other galaxies consist mostly of old, evolved stars, indicating that the activity of star formation in these systems ended in a remote past.

Why do galaxies show such a variety of properties? How could they form and diversify in this way, and when did they do so? Is a given property an indication of a specific evolutionary phase in the life of a galaxy, or have galaxies with different properties formed and evolved in different ways? Which physical mechanism is each property the manifestation of?

Such questions are among the most important in the study of cosmology, and yet they still elude a consistent and complete explanation. There was a time during the cosmic evolution that followed the BIG BANG when there were no galaxies, and the universe was filled with matter and radiation uniformly distributed throughout the space. Only very tenuous perturbations of the density of matter and radiation altered this otherwise almost perfect homogeneity. Then the universe underwent a transition from this phase to the highly structured one that characterizes the current epoch. GALAXY FORMATION and evolution are the results of a complex sequence of events that occurred during this transition. Their importance rests in the fact that the chronology and dynamics of these events are intimately linked to the global cosmological properties of the universe (i.e. the age of the universe, if it will expand for ever or if it will eventually collapse again in a Big Crunch, or if the expansion is an accelerated one or not) and to those of the matter that fills the cosmic space (e.g. the composition of the dark matter and its abundance relative to the ordinary, visible matter). The goal of galaxy evolution is to reconstruct back in time the physical mechanisms that led to the present-day galaxies and to explain them in terms of the properties of the matter and of the cosmology.

The morphology, structure and luminosity of galaxies

The galaxies in the present-day universe have been extensively studied for more than seventy years and, as a result, we have now a fairly accurate knowledge of their physical properties, such as the distribution of luminosity, abundances, morphology, age, chemical composition, and internal kinematics. This knowledge forms the background against which observations of the properties of galaxies at earlier cosmic epochs are compared to test and quantify galaxy evolution.

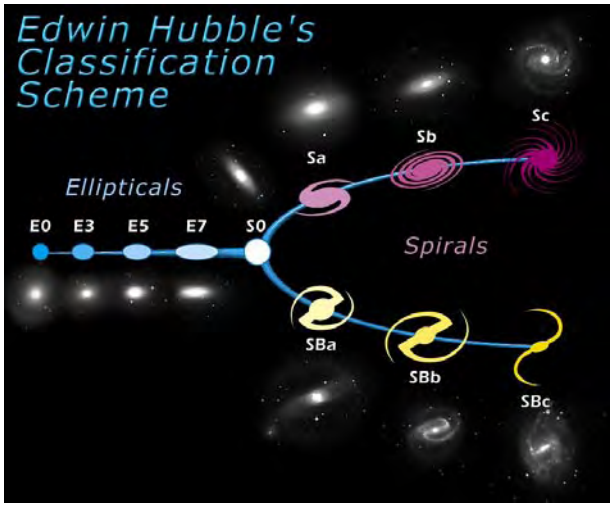


Figure 1. The Hubble morphological classification. Elliptical galaxies are ordered from E0 to E7, according to their ellipticity. The S0 type essentially represents elliptical galaxies that also have a disk. Spiral galaxies are divided into normal and barred types, and are ordered from the 'a' to the 'd' sub-types (including mixed type such as 'bc' and 'cd' not shown in the figure), according to the increasing importance of the disk component over the bulge. Irregular galaxies are not shown. (Photo credit: STScI OPO.)

Galaxy morphology

The morphology of the luminous component of galaxies, namely stars, gas and dust, is probably the most obvious manifestation of the diversity of galaxy properties. During the first half of the 19th century, EDWIN P HUBBLE classified galaxy morphologies into the four principal types of ellipticals, lenticulars, spirals and irregulars. Figure 1 shows the classical representation of the Hubble classification, also referred to as the *Hubble sequence*.

Essentially, the HUBBLE CLASSIFICATION scheme is based on the relative proportions of the two major structural parts of the luminous component of galaxies, namely the bulge and the disk. The bulge is generally made of relatively old and evolved stars, which are characterized by red colors, and is shaped like an ellipsoid with various degrees of flattening and oblateness (ellipsoids with no flattening and oblateness are spheres). In general the bulge contains no appreciable amount of dust and gas, and therefore has no activity of star formation. The disk is composed by a mixture of old and young stars and by gas and dust, and it is frequently site of star formation activity. Its colors are generally bluer than those of the bulge. The most visually striking feature of the disk is the frequent presence of the 'SPIRAL ARMS', density waves that generate in the disk and which contain active regions of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

star formation. Other articles in this encyclopedia discuss galaxy morphology in greater detail.

Bulge and disk do not only differ for the composition of their stellar populations and interstellar medium, but also for their dynamics. Galaxies are large concentrations of mass, and they would collapse under the effect of their own self-gravity if some opposing force were not there to keep them in dynamical equilibrium. Disks are kept in equilibrium by rotation, which provides the centrifugal force to oppose gravity. Bulges have modest or no rotation, and are sustained against gravity by the 'velocity dispersion' of their stars, which is similar to the motions of molecules in a hot gas. These motions provide the pressure that contrasts gravity and keeps the structure in dynamical equilibrium.

ELLIPTICAL GALAXIES consist only of the bulge component and have no disk. LENTICULAR GALAXIES are essentially ellipticals with a very thin, often almost invisible disk, which gives them the shape of a lentil. SPIRAL GALAXIES have both a bulge and a disk with spiral arms. If a bar is present, this contains the bulge at its center and has the spiral arms departing from its extremities. IRREGULAR GALAXIES tend to resemble spiral ones, except that the bulge, the disk and the spiral arms are not regularly defined. There are also amorphous galaxies, which have morphology that cannot be easily classified into any of the main types.

Initially, it was believed that the Hubble sequence was the manifestation of an evolutionary process such that galaxies originated as ellipticals, which were therefore called the 'early types', and evolved to become spirals and irregulars, which were called the 'late types'. Although it has been shown that such ideas were incorrect, nonetheless the nomenclature has remained, and today it is common, for example, to refer to ellipticals and lenticulars as 'early-type galaxies'.

The luminosity function

One very powerful tool that describes in a quantitative way two of the most important properties of galaxies is a statistic called the *luminosity function*, which provides information on the relative frequency in space of galaxies with given luminosity. The reason to couple together luminosity and abundance is an important one. As will become clear later, the luminosity of galaxies is a measure of either their stellar content or of their current activity of star formation or of both (depending on the evolutionary phase of the galaxy). The stellar content, in turn, is generally correlated to the total mass of the galaxy, to its star formation history as well as to the history of merging with other galaxies. Finally, the current activity of star formation provides a measure of the rate at which the stellar content of galaxy is increasing.

The abundance of galaxies in space (or volume density) reflects another set of important physical properties. Galaxies form within virialized structures, namely systems that have collapsed under the effect of their self-gravity and have decoupled their dynamical evolution from that of the cosmic mass distribution (for example, they have stopped following the cosmic expansion). The theory of gravitational instability predicts that the number of virialized systems that have formed at any given time depends on their mass, with more massive systems being less abundant than less massive ones. It also predicts that, for a given mass, the number of systems depends on the density parameters Ω and cosmological constant Ω_Λ and on the spectrum of the primordial density perturbations that emerged from the Big Bang[1]. However, only galaxies that have formed enough stars or that have enough star formation (and whose light is not obscured by dust) can be observed. Thus, by coupling together the information on the luminosity and abundances of galaxies, the luminosity function provides synthetic information on the formation and evolution of both the structural and visible components of galaxies as well as on the cosmological parameters. A more detailed discussion on the luminosity function can be found in the article LUMINOSITY FUNCTION OF GALAXIES.

The luminosity function of local galaxies has been directly measured with good accuracy. It is found that fainter galaxies are much more common than brighter ones. For example, although galaxies with luminosity similar to that of the Milky Way or the Andromeda galaxy[2] are relatively common, fainter galaxies such as the Magellanic Clouds are much more frequent. Bright galaxies such as the big ellipticals often found at the centers of clusters of galaxies are much rarer. Irregular galaxies are typically fainter and more abundant than galaxies with regular morphology.

Star formation and merging: the engines of galaxy evolution

Clearly, a morphological classification is useful only if it reflects some important physical property that can be directly linked to the origin and evolution of the galaxies. What are the physical properties of the Hubble sequence? We have seen that the age of the stellar populations and the dynamical state play an important role, but how do they relate to the history of star formation and how can we use them to reconstruct its evolution? How can we exploit the information contained in the luminosity function?

Star formation

Let us illustrate the role of star formation with an example. Let us consider the case of a spiral galaxy which contains 10^{11} stars similar to the Sun. This would be approximately the size of a galaxy like the Milky Way. Let us imagine that

[1] The cosmological parameters Ω and Ω_Λ measure the total mass content of the universe and the strength of the cosmological constant, a form of repulsive force that may cause the universe to keep expanding at an accelerated pace. Their combined value decides if the universe will expand for ever or it will collapse into a Big Crunch. Together with the Hubble constant, they determine the age of the universe

[2] Such galaxies are commonly referred to as 'L^* galaxies'. The symbol L^* is used to indicated the value of the blue luminosity of galaxies similar to the Milky Way. Its value is $L^* = 10^{12} L_\odot$, where $L_\odot$ denotes the luminosity of the Sun or 4×10^{33} erg s^{-1}

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

this galaxy is experiencing an episode of star formation, and that it is generating new stars at a rate of, say, 10 solar masses per year or $M_\odot\ \mathrm{yr}^{-1}$ (this is the total stellar mass, not the number, of new stars). The STELLAR MASSES of freshly produced stars are not all equal, but vary approximately between 100 solar masses to about one tenth of solar mass (i.e. stars whose mass is between one hundred times and one tenth the mass of our Sun. The symbol that indicates the mass of the Sun is $M_\odot$). The relative proportions of stars of a given mass produced during an episode of star formation are, to first approximation, the same everywhere and are described by the so-called initial mass function (IMF). These proportions are such that for every star of mass M being produced, there are about 2.5 times as many stars with mass $M/2$ being also produced, implying that low-mass stars are formed much more frequently than high-mass stars. Another way to state this is that when a given total amount of stellar mass is being produced in a galaxy, this is distributed prevalently in low-mass stars.

While low-mass stars carry the bulk of the new stellar mass produced during an episode of star formation, they contribute very little to the luminosity of the event. This is because high-mass stars are much more luminous than low-mass ones. A star of 100 $M_\odot$ is about 1000 000 times brighter than the Sun. Even if there is only one star of 100 $M_\odot$ every 500 stars of 1 $M_\odot$, it alone outshines all of them, whose contribution to the total luminosity is negligible. After the first million years of continuously forming stars at a rate of 10 $M_\odot$ per year, the galaxy of our example has put together a total stellar mass equal to $10 \times 1000\ 000 = 10^7 M_\odot$, about one ten-thousandth of its initial stellar mass. Of these, about 20 000 are massive stars. Because each of these massive stars is 1000 000 times brighter than stars like the Sun, their total luminosity will approximately equal that of 2×10^{10} of these. Thus, during the star formation activity the galaxy increases its luminosity by $\sim 20\%$. Its colors also become bluer, because the radiation emitted by massive stars is very rich of light at short wavelengths, which include the blue and ultraviolet ones. Had the galaxy initially been fainter, containing say 10^{10} Sun-like stars, the same episode of star formation would have trebled its initial luminosity.

Although fainter, however, low-mass stars live much longer than high-mass ones. A star of 100 $M_\odot$ lives about one million years, after which it explodes as supernova, while a star like the Sun keeps burning its hydrogen nuclear fuel for about 10 billion years, one hundred thousand times longer. Thus, about one million years or so after the end of the star formation, when all the massive stars have died out, the luminosity of the galaxy eventually becomes again dominated by low-mass stars, both the recently created new ones and the ones that existed before the episode of star formation. If star formation continued at the same rate for one million years, the galaxy of our example, which started with 10^{11} Sun-like stars, would create another 10^7 of them, or only 0.1% of the original amount. Once the massive stars have died out, the luminosity of the galaxy in its post-starburst phase is again powered by these stars and it returns to the original value. Of course, star formation of longer duration would result in a larger increase of final luminosity and stellar mass. The luminosity of the galaxy during the phase of star formation, however, depends only on the rate at which stars are being produced because this is what determines the amount of massive stars that can be observed at a given time. Also, after star formation ends, the colors of the galaxy also return to be dominated by the light of the low-mass stars, which is redder than that of the massive stars whose colors dominated the spectral energy distribution of the galaxy during the star-forming phase.

Luminosity, stellar mass and color are not the only changes induced in a galaxy by star formation. Another fundamental product is the creation of the 'metals', namely of elements heavier than helium. The chemical composition of the universe as it emerged from the Big Bang practically consisted only of hydrogen and helium, in the proportion of about 80% and 20% by mass, respectively. Essentially all the heavier elements that we observe, both on Earth and elsewhere in the universe, have been synthesized in the interiors of stars by reactions of nuclear fusions at epochs subsequent to the Big Bang. The metals synthesized in the interiors of stars are released to interstellar space during SUPERNOVA explosions. Once the chemically enriched material has been dispersed, it becomes available to fuel further star formation and, with successive generations of star formation episodes, the general metal content of the galaxy (and of the universe as well) increases with time.

Merging

Another important mechanism that drives galaxy evolution is merging, which can transform the morphology of galaxies and promote a substantial amount of star formation over a short period of time. It also induces changes in the volume density and stellar content, and therefore affects the evolution of the luminosity function. Merging of galaxies is observed at all cosmological epochs where galaxies have been detected, and it is also predicted in all theories of galaxy evolution. See GALAXIES: INTERACTIONS AND MERGERS.

When two or more galaxies merge, the final structure that forms may not bear any resemblance to the original components. The outcome of the process depends on the relative masses of the mergers and on their types. If two galaxies with approximately the same mass merge, the perturbation to the gravitational equilibrium of the two systems is very large and takes place on a relatively short time scale (of the order of the free-fall time scale, which is $\sim 10^7$–10^8 years for a galaxy similar to the Milky Way). The gravitational potential of the forming system varies very rapidly throughout the whole structure during the event, a process known as *violent relaxation*, and when this happens the final equilibrium configuration of the system is a spheroid, similar to a bulge or an elliptical galaxy, supported against its self-gravity by the velocity dispersion. If the two merging systems were two spiral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

galaxies, their disks would be destroyed during the event, and the gas contained in them transformed into stars on a time scale of the order of a few free-fall times, giving origin to a very powerful burst of star formation. In this case, depending on the gas content, the stellar mass of the final galaxy can increase to significantly larger values than the sum of the two mergers, and the luminosity during the event can reach prodigious values.

If the merging systems have largely different masses, and consist of one massive galaxy and one or more sub-massive ones, the event is not as traumatic as above. Although the small galaxy is destroyed and loses its identity, it only causes a relatively minor perturbation to the structure of the massive galaxy, which in general retains its properties and morphology. If there is available gas, either in the large galaxy or in the accreting one, the perturbation can induce episodes of star formation that may significantly increase the luminosity of the galaxy, even if they do not in general appreciably change its stellar mass.

Reconstructing the evolution from present to past

The morphological and dynamical differences of present-day galaxies have been interpreted as evidence of different evolutionary history (Kennicutt 1998). The STELLAR POPULATIONS of elliptical and lenticular galaxies as well as the bulges of spirals are very old, and hence the activity of star formation must have ended rather early during cosmic evolution. Direct estimates of the age of the stars place the epoch of their formation to when the universe was about $\sim 20\%$ of its current age. For example, if the universe is 14 billion years old, most ellipticals, lenticulars and bulges already assembled the bulk of the stars by the time the universe was about 3 billion years old. This means, for example, that to form a total stellar mass of $\sim 10^{11} M_\odot$ in a few billion years, star formation rates must have been very high, of the order of $10^{11}/10^9 \sim 10^2\ M_\odot\ \mathrm{yr}^{-1}$. Thus, the 'spheroids' must have have formed their stars during intense but relatively short bursts of star formation.

In the disk of spiral galaxies star formation is observed to proceed at a relatively low pace but in a continuous fashion, and the age and composition of the stars suggest that this was the case in the past as well, although the rate of star formation was probably higher in the past. Luminosity and colors (the so-called spectral energy distribution) of DISK GALAXIES, therefore, largely reflect those of the forming stellar population. At the onset of the first episode of star formation, the spectral energy distribution closely resembled that of young massive stars, since only these objects populated the galaxy. At more advanced stages, when a significant older stellar population had built up, the spectral energy distribution became redder, reflecting the presence of the low-mass stars. For such a mechanism to take place on a time scale comparable to the age of the universe it requires a reservoir of gas to be continuously available to feed the activity of star formation. This gas is thought to accrete on the disk from the outer regions of the galaxy.

The morphological and dynamical properties of galaxies are decided when the body of the structure forms by gravitational collapse of a proto-cloud or of discrete sub-structures as in the case of merging. If the collapsing structure has enough angular momentum (in other words, enough rotation), it eventually settles on a final equilibrium configuration which has the shape of a disk, where the centrifugal force balances gravity. On the other hand, if the structure does not have enough angular momentum, there is no centrifugal force to oppose gravity, and the collapse continues until the velocity dispersion of the sub-components of the structure can halt it. The equilibrium configuration in this case is that of a spheroid, whose morphology is significantly more compact than that of a disk.

This difference in the collapse dynamics have consequences in the star formation mechanisms as well. The violent collapse that gives origin to a spheroid deeply changes the gravitational potential of the forming structure in a short time scale (of the order of the free-fall time). This large perturbation can drive a violent burst of star formation throughout the structure that converts all the available gas into stars on a time scale of several free-fall times (i.e. 10^8–10^9 yr). When the burst ends because of gas exhaustion, no more significant star formation is possible if the galaxy does not accrete additional gas by interaction and merging with other galaxies. The stars, which are essentially coeval, then begin a phase of passive evolution.

Observing the evolution from past to present

Directly comparing the properties of earlier galaxies with the present-day ones is another powerful avenue of investigation which has been made possible by the technological advances in telescopes and instrumentation achieved over the last 30 years. Thanks to the travel time imposed by the finite speed of light, approximately 1.6 billion miles per year, galaxies at earlier cosmic epochs can be studied simply by looking at distant galaxies. This is because a distant galaxy appears to an observer as it was in the past, at a look-back time equal to the travel time of the light from the galaxy to the observer. For very distant galaxies the look-back time is a large fraction of the entire age of the universe. Table 1 one shows the relationship between the luminosity distance[3], redshift, cosmic age, look-back time, and fractional age of the universe. These quantities have been calculated assuming a universe with a Hubble constant equal to $H_0 = 65\ \mathrm{km}^{-1}\ \mathrm{Mpc}^{-1}$ and with density parameter and cosmological constant equal to $\Omega = 0.3$ and $\Omega_\Lambda = 0.7$, respectively. It is common practice to use the redshift, rather then the distance, to provide a measure of the look-back time (for more information about the redshift see the article on GALAXIES AT HIGH REDSHIFT in this volume). Galaxies with redshift smaller than $z \sim 0.1$,

[3] This is the distance that enters in the inverse-square law, which states that the observed flux ϕ of the source is inversely proportional to the square of the distance d from the source to the observer, or $\phi = L/4\pi\ d^2$, where L is the luminosity of the source.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or distance shorter than $\sim$ 500 Mpc, are considered to be local systems observed in the present-day universe. Note that the epoch when the universe was half its present age (which is 14.5 billion years with the adopted cosmological parameters) corresponds to redshift $z \sim 0.7$, or to a distance of $\sim$ 4600 Mpc, and that the bulk of the cosmic evolution, say $\sim$ 80% of the cosmic age, took place from redshift $z \sim 2$ to the present time (redshift $z = 0$). Note also the little increment of look-back time at very high redshifts. Finally, we recall that the cosmic microwave background radiation originated at $z \sim 1300$.

Table 1. Relationship between luminosity distance D in Mpc, redshift z, cosmic age (in billion years), look-back time (LBT) (in billion years) and fractional age as % of the present age. $H_0 = 65$ km s^{-1} Mpc^{-1}; $\Omega = 0.3$; $\Omega_\Lambda = 0.7$.

D	z	Age	LBT	%
0	0.0	14.51	0.00	100.00
496	0.1	13.11	1.40	90.35
1672	0.3	10.83	3.68	74.64
3051	0.5	9.08	5.43	62.58
4587	0.7	7.72	6.79	53.20
7116	1.0	6.20	8.32	42.73
11 749	1.5	4.53	9.99	31.22
18 735	2.0	3.48	11.04	23.98
21 965	2.5	2.77	11.74	19.09
27 378	3.0	2.26	12.24	15.58
32 935	3.5	1.91	12.60	13.16
38 610	4.0	1.63	12.88	11.23
50 240	5.0	1.24	13.27	8.55
74 339	7.0	0.81	13.71	5.58
111 830	10.0	0.50	14.01	3.45
657 150	50.0	0.05	14.46	0.34
1370 345	100.0	0.02	14.50	0.14
19 222 400	1300.0	0.0004	14.51	0.003

Faint blue galaxies

Currently, the observations have studied relatively well galaxies at redshifts up to $z \sim 1$. From $z \sim 1$ up to $z \sim 2.5$ there is a gap of systematic observations due to the difficulty, with current technology, of measuring redshifts of faint sources in that range. At redshift higher than $z \sim 2.5$, and up to $z \sim 5$, star-forming galaxies have been observed and studied in large numbers (but not other types of galaxies, if they existed at those epochs) thanks to a special observing strategy called the 'Lyman-break technique' (see the article GALAXIES AT HIGH REDSHIFT in this volume), fine-tuned to be sensitive in that redsfhit range. More recently, there have been a few reports of identification of star forming galaxies with redshifts as high as $z \sim 6.2$, and also of candidates at redshifts $z \sim 10$. Great efforts are currently being put into the design and construction of very large telescopes, such as ground-based ones with mirrors of 30 meters or more, or the 8-meter space-based Next Generation Space Telescope (NGST), capable of detecting galaxies at $z \sim 20$, if they exist there.

The first surveys of very faint galaxies were carried out in the late 1970s and early 1980s, following the advent of the first CCD detectors, and consisted mostly of deep imaging surveys (Tyson 1977). Although they did not provide direct information on the distance to the galaxies (and thus on the epoch at which they were observed), these observations found evidence that galaxies' properties have changed over time. If the distribution of *intrinsic* luminosities and the abundances of galaxies did not evolve, then from the knowledge of the luminosity function at the present time one can accurately predict the numbers of distant (i.e. earlier) galaxies that have a specified distribution of *apparent luminosities*. This is due to the fact that the apparent luminosity of a galaxy is related to its intrinsic luminosity by the inverse-square law. Thus, under the assumption that the distribution and abundances of absolute luminosties (i.e. the luminosity function) is the same at any epoch, one can compute the distribution of apparent luminosities of galaxies expected in a deep survey simply by summing together the contribution of the luminosity function from each epoch. Very interestingly, the observations revealed that the number of faint galaxies is much higher than one would expect from such a calculation, with the excess increasing at fainter galaxies and being more pronounced at bluer wavelengths than at redder ones. Figure 2 shows this effect by plotting the number counts of galaxies per square degree observed in the sky at various wavelengths versus their apparent luminosity, expressed in magnitude[4] (bottom) or physical units (top). Plotted in the graph are the observed counts in the K, I, B and U bands, respectively, corresponding to wavelengths of 22 000 (near-infrared), 8000 (red end of visible radiation), 4000 (blue radiation) and 3600 Å (violet end of visible radiation). The different symbols represent data from various observers, and provide an indication on the uncertainties of the observations. The curves are the theoretical predictions based on the local luminosity function. As can be seen, the excess in the counts over the predictions is very strong in the U and B bands, where it becomes noticeable already at around magnitude 22 and it sharply increases at fainter levels. The excess becomes progressively smaller towards redder wavelengths, and in the K band it is noticeable only at the faintest levels.

While deep multi-band imaging showed that the number density and/or the luminosity of blue galaxies is rapidly evolving going towards earlier epochs, without an indication of the distance to these faint galaxies, however, it is not possible to understand from these data alone when such evolution took place and whether the evolution of the blue galaxies was synchronous or not with other types.

The advent of more sensitive CCD detectors and advanced instumentation during the late 1980s and early 1990s made it eventually possible to carry out large faint spectroscopic surveys and measure the redshifts of

[4] The magnitude m of a galaxy with apparent luminosity f is defined as $m = -2.5 \log(f) + c$, where c is an arbitrary number that sets the zero point. Because of the negative sign, the scale of magnitude is such that brighter sources have smaller magnitudes and vice versa.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

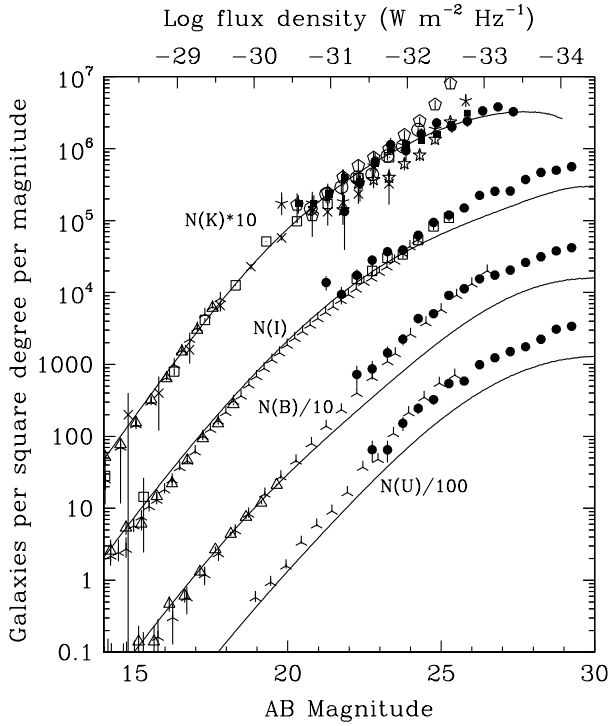


Figure 2. The excess of blue galaxies in the faint counts. The vertical axis reports the surface density of galaxies in the sky in units of galaxies per square degree per magnitude interval. The horizontal axis on the bottom reports the magnitude (faint is to the right, bright is to the left), the one on the top reports the luminosity in logarithmic scale expressed in physical units of Watts per square meter per Hertz. Counts in the K, I, B and U bands are plotted. Different symbols denote different observations. The curves are the theoretical predictions of the counts assuming that the galaxies of the present-day universe have not evolved. The faint excess is very strong in the U and B band and becomes progressively weaker in the I and K bands. (From Ferguson 1998.)

the galaxies, and hence their distance. This, in turn, made it possible to measure the intrinsic luminosity of the galaxies, and hence to derive their luminosity function at various cosmic epochs, mapping its evolution in time (Lilly *et al* 1995). These works showed that most of the faint blue galaxies observed in the imaging surveys are intrinsically relatively small systems located at moderate to intermediate redshifts, i.e. $z \sim 0.3$ to $z \sim 1$, which are undergoing a robust activity of star formation. Their closest counterparts in the local universe are the irregular galaxies, similar to the Magellanic Clouds. Interestingly, the luminosity function of blue galaxies was found to evolve very rapidly from the present epoch up to the earliest epoch studied ($z \sim 1.3$), implying that the luminosity (and hence the star formation rate) and abundance of these galaxies were progressively higher and higher in the past.

The observations also showed that the luminosity function of galaxies that have the same intrinsic spectral energy distribution as the ellipticals and spirals of the local universe, hence which are redder than the faint blue galaxies, is characterized by a much more modest evolution over the same cosmic epochs. This is shown in figure 3, which plots the measures of the luminosity function of red and blue galaxies in a number of reshift intervals, as specified in the labels (note that there are not red galaxies in the range $z = 1.00$ to $z = 1.30$). The vertical axes represent the number density in units of galaxies per cubic Mpc, the horizontal ones the luminosity expressed as absolute magnitudes (this is the apparent magnitude a source would appear to have if placed at a distance of 10 pc from the observer). The circles show the data, the continuous line is the fit to the data, the dashed line is the fit to the data of the redshift bin $z = 0.2$ to $z = 0.5$ (used to show the relative evolution), and the dotted line is the fit to the luminosity function of local galaxies. Thus, galaxies with colors similar to those of the Milky Way (or redder) were already formed and in place by redshift $z \sim 1$ (namely for the last two-thirds of cosmic time) and have been evolving very modestly since then. Blue galaxies, on the other hand, have undergone a spectacular evolution over the same epoch, progressively diminishing their abundances and dimming their luminosity (i.e. reducing the star formation rate) towards the present time.

Solving one piece of the puzzle

With the repair of the *Hubble Space Telescope* (*HST*) in 1994 it eventually became possible to image distant galaxies with enough spatial resolution to classify their morphology and compare it to that of the local galaxies (Abraham *et al* 1996). The unprecedented angular resolution of the *HST* imaging has shown that galaxies up to redshifts $z \sim 1$ have morphologies that can still be classified according to the traditional types of ellipticals, lenticulars, spirals and irregulars (Glazebrook *et al* 1995). Consistently with what was found in the redshift surveys, the *HST* imaging surveys showed that elliptical and lenticular galaxies underwent essentially only passive evolution[5] during the last half of the universe lifetime (Schade *et al* 1999). Over the same period of time, the disks of spiral galaxies showed some moderate evolution induced by an overall increased of the activity of star formation, although their abundance was found to be comparable to that in the present-day universe (Lilly *et al* 1999). In sharp contrast to the modest evolution of ellipticals and spirals, however, the *HST* images showed that irregulars galaxies underwent a spectacular amount of evolution, both in luminosity and abundance, over the same period of time (Ellis 1997). Thus, *HST* has clarified the nature of the faint blue galaxies by showing that they consist of irregular galaxies, and has provided an independent confirmation that these systems were much more abundant in the past than they are today, and that they were also forming stars at a higher pace. These galaxies, in effect, have been the drivers of the

[5] The expression 'passive evolution' denotes the evolution in luminosity of a galaxy induced by the natural evolution of its stars

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

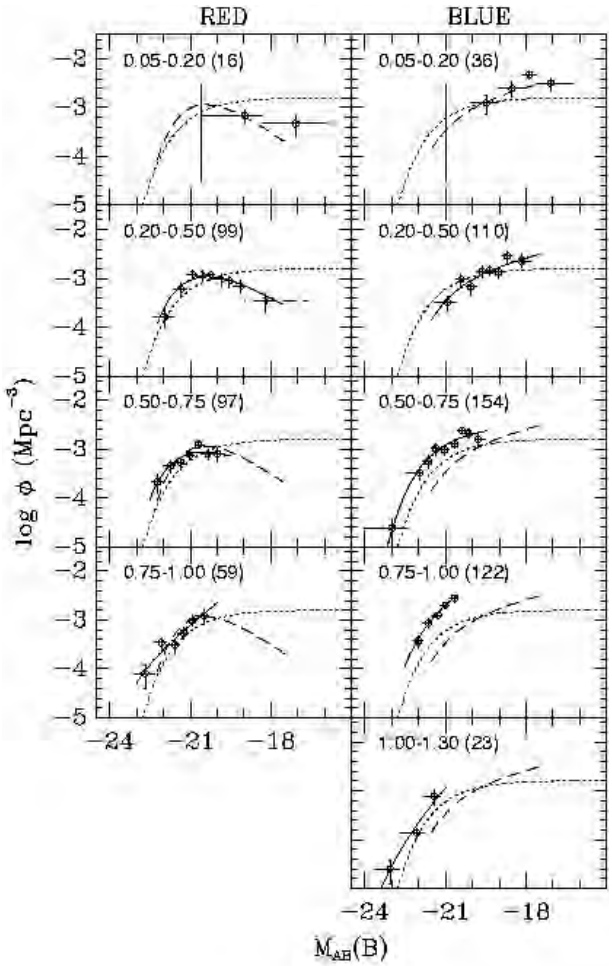


Figure 3. The evolution of the luminosity function of red and blue galaxies. Each panel contains measure of the luminosity function in the specified redshift range (data points) and a mathematical fit to the data point (continuous curve). Also shown is the luminosity function of local galaxies (dotted line) and the fit to the data of the redshift bin $z = 0.20$–$z = 0.50$ (dashed line), useful to see the differential evolution. The vertical axes represent the volume density of galaxies in units of galaxies per cubic Mpc. The horizontal represents the intrinsic luminosity, expressd in absolute magnitudes in the *B* band. The absolute magnitude is the magnitude a source would appear to have if placed at a distance of 10 pc from the observer. (From Lilly *et al* 1995.)

evolution of the overall luminosity function (i.e. that of all galaxies, regardless of their morphology or colors) and of the star formation history of the universe during the last 8 billion years of cosmic age. Figure 4 shows four examples of faint blue galaxies imaged with *HST*. Note the highly irregular and fragmented morphologies, which consist of compact regions that host active star formation embedded in more diffuse structures composed by a mixture of new and older stars and of gas and dust.

Figure 4. Four images of faint blue galaxies taken with the *Hubble Space Telescope* (Griffith *et al* 1998). Images like these helped understanding the nature of these population of galaxies, which underwent a spectacular evolution in luminosity and volume density over the last 8 billion years of the cosmic time. These galaxies are very similar to the Irregular types found in the present-day universe, however they were much more abundant in the past. Where the descedants of these rapidly evolving galaxies are at the present epoch is still precisely not known. (Photo credit: STScI OPO, R Griffiths (Carnegie Mellon University and NASA).)

Looking at the big picture

The study of galaxies over the last two-thirds of the cosmic age has shown clear patterns in the evolution of galaxies. What are today small galaxies seem to have undergone most of their evolution over this time. Although we do not yet know with certainty what and where the present-day counterparts of the faint blue galaxies are at the present epoch, there is mounting evidence that these systems have gradually merged with larger galaxies and have disappeared. As we have discussed earlier, the accretion of a small system onto a large one does not alter the properties of the latter as when a major merging event between two similar systems takes place. Another possibility that has been proposed is that at least a fraction of them have faded beyond detectability of the present intrumentation, once the burst of star formation has ended.

Over the same cosmic time, large galaxies such as ellipticals and spirals evolved much more modestly, being essentially already in place by redshift $z \sim 1$. This implies that they completed most of their formation at earlier epochs, and that to understand their evolution one must search at even earlier epochs.

The search for the progenitors of the so-called 'bright galaxies' has traditionally occupied a central role in observational cosmology (see the article on GALAXIES AT HIGH

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

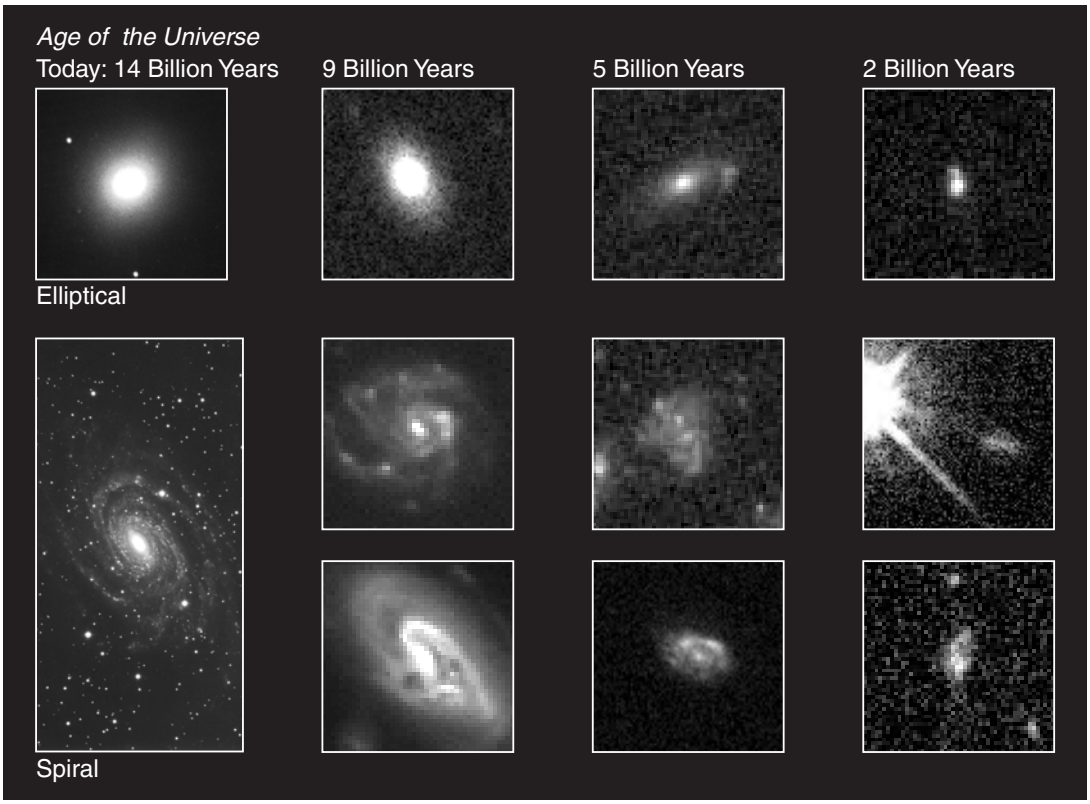


Figure 5. Comparison between galaxies observed at various cosmic epochs (expressed as time after the Big Bang). The traditional morphological types can be recognized already at ~ 5 billion years. At earlier epochs the morphology seems to differ from that of later epochs, although galaxies do exhibit various degrees of compactness and irregularity in their morphologies, ranging from galaxies with light profiles similar to that of ellipticals and bulges to those with disk-like morphology, in resemblance to the Hubble types. (Photo credit: STScI OPO, M Dickinson, M Giavalisco and NASA.)

REDSHIFT in this volume). One of the most important results of the extragalactic research of the last five years has been the identification of a massive population of star-forming galaxies at redshifts larger than $z \sim 2.5$, observed at a time when the universe was younger than 20% of its current age (Steidel *et al* 1996; Giavalisco *et al* 1996). These galaxies, also known as Lyman-break galaxies, are observed during an intense activity of star formation which has contributed a significant fraction of all the stars that we count in the present-day universe, between 20% and 50% according to current measures. Thus, the stars formed in these galaxies can account for a large fraction (if not all) of the stellar mass currently contained in spheroids since, having formed so early in the cosmic evolution, they must necessarily be very old today. These arguments strongly support the idea that the Lyman-break galaxies are the direct progenitors of the present-day spheroids, observed at a time when they were forming the bulk of their stars.

Unfortunately, because of the gap in the observations between redshift $z \sim 1$ and $z \sim 2.5$, this association currently cannot be tested by directly following the evolution of the galaxies at progressively lower redshifts. Equally unknown is whether the Lyman-break galaxies were the first to appear in the universe or if other galaxies preceded them in evolution. These will be major activities of research when instruments of improved sensitivity for ground-based telescopes and the Next Generation Space Telescope will come on line.

Figure 5 shows a comparison between elliptical and spiral galaxies of the present-day universe (observed with ground-based telescopes) with their counterparts observed at progressively higher look-back times (observed with *HST*). The galaxies in the right panels are Lyman-break galaxies at redshift $z \sim 3$. Note that for these objects, a classification into morphological types is not really possible. However, in resemblance to the Hubble sequence, these galaxies do show various degrees of compactness and irregularity in their morphologies, ranging from systems with the light profile similar to that of ellipticals and bulges to systems with disk-like morphology. In the three panels on the right of figure 5 we have represented the cases of a galaxy with the same type of light profile as a present-day spheroid (top), of one with the same light profile as a disk (middle), and of one with an irregular profile (bottom).

One possible evolutionary path is that the Lyman-break galaxies have formed the bulk of their stars in intense

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

bursts of star formation at high redshifts and then evolved mostly passiveley. In this scenario, the final outcome will be an elliptical galaxy when at the end of the collapse the progenitor has exhausted all the gas reservoir. Or it will be a spiral galaxy when the proto-bulge has still sufficient gas around it (or it is able to strip gas from another galaxy it happens to interact with), to give origin to a disk where star formation can occurr with low rate and over a prolonged period of time.

Another possibility is that the Lyman-break galaxies have evolved into the present-day galaxies through successive generations of hierarchical merging. In such a scenario, elliptical galaxies and bulges are easily explained, since we have seen that the outcome of a major merging episode is a spheroid. In order to form lenticular and spiral galaxies, it is necessary that, following the merging episode that originated the bulge, the nascent galaxy has not only a sufficient gas reservoir but also the possibility to evolve undisturbed, with no further merging events, so that the disk can develop (disks are destroyed in a merging event). According to this scheme, elliptical galaxies would be spheroids that have not developed a disk, either because of the lack of gas or because they formed recently and simply did not have the time to do so. They would be, therefore, structurally 'young' systems, even if their stellar populations are old. Spirals, on the other hand, must have had the time to develop the disk, and thus they are 'structurally' old even if they contain a large fraction of young stars in the disk. Interestingly, while the data show that the occurrence of merging seems to be higher in distant galaxies, its relative importance in the evolution of galaxies of different types has not yet been quantified. Thus, it is not known, from an empirical point of view, for which galaxies merging has been instrumental for their formation, and for which it only provided a 'perturbation' in an otherwise structurally passive evolution.

The picture that seems to emerge is that galaxy evolution proceeded along different paths, according to the different types of the galaxies, with massive galaxies forming first and smaller systems later. In other words, the mass seems to be the key parameter that decides the evolutionary destiny of galaxies. This is consistent with the paradigm that gravity has been the primary mechanism for the formation of structure at all scales in the universe, including galaxies. The paradigm states that the cosmic structures form as a result of the amplification and gravitational collapse of density perturbations in the mass distribution. Since perturbations grow as a result of their self-gravity, a prediction of the paradigm is that more massive perturbations grow faster than smaller ones, and hence they can virialize and begin the process of star formation sooner.

The importance of this paradigm goes beyond explaining the formation of galaxies, since it forms the conceptual ground for our understanding of many of the observations in cosmology and, ultimately, of the evolution of the universe as a whole. Yet, despite its importance, this paradigm has never be subject to a rigorous empirical test. Reconstructing the history of the evolution of galaxies from their earliest appearance in space to the present time will provide one the first and most important empirical constraints to the whole theoretical framework of cosmology.

Bibliography

Abraham R *et al* 1996 *Astrophys. J.* **107** 1
Kennicutt R 1988 *Ann. Rev. Astron. Astrophys.* **36** 189
Ellis R S 1997 *Ann. Rev. Astron. Astrophys.* **35** 389
Ferguson H C 1998 The Hubble Deep Field *Proc. STScI May Symp.* ed M Livio, S M Fall and P Madau p 181
Giavalisco M, Steidel C C and Macchetto F D 1996 *Astrophys. J.* **470** 189
Glazebrook K, Ellis R, Santiago B and Griffiths R 1995 *Mon. Not. R. Astron. Soc.* **275** 19
Lilly S J *et al* 1999 *Astrophys. J.* **500** 75
Lilly S J *et al* 1995 *Astrophys. J.* **455** 108
Schade D *et al* 1999 *Astrophys. J.* **525** 31
Steidel C C, Giavalisco M, Pettini M, Dickinson M and Adelberger K 1996 *Astrophys. J.* **462** 17
Tyson J A 1988 *Astrophys. J.* **96** 1

Mauro Giavalisco

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Formation

A galaxy is the environment in which stars are born and die, and distant galaxies are the luminous beacons that enable us to probe the distant universe. Our galaxy, the Milky Way, is one of billions of such systems in the observable universe. How galaxies formed represents a central theme in modern cosmology.

At first glance, the universe appears to contain two different types of galaxies: galaxies with disk-like morphologies and galaxies with spheroidal morphologies. This basic distinction breaks down, however, once individual disk and spheroidal galaxies are examined in detail, since most disk galaxies contain small spheroidal components at their centers and most spheroidal galaxies contain small disks at their centers.

Disks in and of themselves contain a wide variety of features. Most DISK GALAXIES exhibit spiral arms with a large range of winding angles and contrast. Approximately half of all disk galaxies also contain a highly elongated bar structure near their center, the bars possessing a variety of axial ratios. Some disks also show moderate deviations from planarity toward their edges (warps) while other disks show significant lanes of dust across their observed profiles. Similarly, ELLIPTICAL GALAXIES, while possessing relatively uniform profiles compared to disk galaxies, show a significant variety of substructure. At least half of ellipticals have detectable shell structure and others distinct cores. Finally, all galaxies, irrespective of type, show great variations in the amounts and spatial distribution of gas, dust, stars, and metal abundances as well as their basic surface brightnesses, luminosities, colours, and masses.

Despite their considerable diversity, galaxies show a remarkable degree of uniformity as well. The profiles of disk and ellipticals are remarkably homologous, the global structural parameters of disk and spheroidal galaxies define a tight two-dimensional plane, and the colours and apparent star formation histories of both spiral and elliptical galaxies show a striking correlation with luminosity. Of great significance is that most galaxies are very slowly evolving structures, both chemically and dynamically. Their properties were acquired long ago, at or soon after the epoch of galaxy formation.

Galaxies began as clouds of primordial gas, hydrogen and helium. Even before galaxies condensed into distinct clouds, infinitesimal density fluctuations were present in the EXPANDING UNIVERSE. These originated at very early epochs in an inflationary phase transition from a universe that initially was relatively uniform. Fluctuations grew in strength under the inexorable influence of self-gravity. Eventually, clouds developed that fragmented into stars. Much of the detailed physics in this schematic of GALAXY EVOLUTION is now understood.

This review begins with a discussion of the cosmological world model in which galaxies form, discusses the processes by which the initial perturbations are established, presents the theory for the growth of these perturbations into collapsing and eventually virialized objects, illustrates the importance of gas cooling in the formation of galaxies, outlines the processes by which galaxies acquire angular momentum, and concludes by summarizing the basic observations and theory of disk and elliptical galaxy formation.

World model

We begin by providing some background on the standard world model and the primordial fluctuations out of which galaxies are believed to have grown.

The apparent homogeneity and isotropy of the observable universe, both in terms of its large-scale structure and the cosmic infrared microwave background radiation—the almost constant 2.73 K blackbody radiation background in which the universe is immersed—motivate the assumption that the universe is both homogeneous and isotropic. By homogeneous, we mean that every point in space statistically resembles every other point in space. By isotropic, we mean there is no point in space where any direction differs statistically from any other direction in space.

Assuming universal homogeneity and isotropy, Einstein's theory of GENERAL RELATIVITY can be used to show that the evolution of the scale of the universe follows the two independent equations:

$$\frac{\dot{a}^2 + k}{a^2} = \frac{8\pi G}{3}\rho \tag{1}$$

$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2} = -8\pi G p \tag{2}$$

collectively known as Friedmann's equations, where a is a measure of the size of the universe, p is the pressure, k is the curvature, ρ is the density, and G is Newton's constant. Combined with the equation of state, these equations completely determine $a(t)$, $\rho(t)$, and $p(t)$.

These equations can be recast into the form:

$$\frac{\ddot{a}}{a} = H_0^2\left[\Omega_{\Lambda,0} - \Omega_0(1+z)^3/2\right] \tag{3}$$

$$\begin{aligned}\frac{\dot{a}}{a} &= H_0 E(z) \\ &= H_0\sqrt{\Omega_0(1+z)^3 + \Omega_{R,0}(1+z)^2 + \Omega_{\Lambda,0}}\end{aligned} \tag{4}$$

where $\Omega_0 = 8\pi G\rho_0$, $\Omega_{R,0} = 1/(H_0 a_0 R)^2$, $\Omega_{\Lambda,0} = \Lambda/3H_0^2$, ρ_0 is the matter density of the universe at the present epoch, Λ is the vacuum energy density or cosmological constant, and R is a constant with units of length. Hubble's constant, denoted by H_0, characterizes the rate at which the universe is expanding at the present epoch. It is the constant of proportionality relating an object's distance D to its rate of recession v:

$$v = H_0 D. \tag{5}$$

Note that $\Omega_{R,0} = 1 - \Omega_0 - \Omega_{\Lambda,0}$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

As the universe expands, the gravitational attraction of the mass inside it slows this expansion. This deceleration may or may not be enough to slow this expansion sufficiently so that the universe recollapses. The case where there is sufficient matter to cause such a recollapse corresponds to a universe where the universal geometry is closed, i.e., $k = +1$. The case where there is not sufficient matter to force such a recollapse corresponds to two separate geometries: one in which the universal geometry is flat ($k = 0$) and one in which the universal geometry is open ($k = -1$). A universe with a flat geometry is known as Einstein–de Sitter.

The time evolution of the universal scale length a is amenable to the following simple analytic solution in the case of an Einstein–de Sitter universe where ordinary matter dominates the energy density ($p = 0; \rho \propto (1+z)^{-3}$):

$$a = a_0(t/t_0)^{2/3} \tag{6}$$

where t_0 and a_0 are the current age and size of the universe, respectively.

For an open universe, the solution is given in terms of the following parametric equations:

$$a = \frac{\Omega_0}{(1-\Omega_0)^{3/2}} \frac{c}{2H_0} (\cosh\Theta - 1) \tag{7}$$

$$t = \frac{\Omega_0}{(1-\Omega_0)^{3/2}} \frac{1}{2H_0} (\sinh\Theta - \Theta) \tag{8}$$

while for a closed universe, the parametric equations are

$$a = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{c}{2H_0} (1 - \cos\Theta) \tag{9}$$

$$t = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{1}{2H_0} (\Theta - \sin\Theta). \tag{10}$$

The turn-around time t_m for this universe occurs when $\Theta = \pi$, so from equation (10), we find

$$t_\mathrm{m} = \frac{\Omega_0}{(\Omega_0-1)^{3/2}} \frac{\pi}{2H_0} \tag{11}$$

for the turn-around time.

Requiring that the curvature be identical everywhere in space-time, the most general way of expressing the concept of distance is the Friedmann–Robertson–Walker metric:

$$\mathrm{d}l^2 = a^2R^2\left[\mathrm{d}\chi^2 + f^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\right] \tag{12}$$

where

$$f(\chi) = \begin{cases} \sin\chi, & k = +1 \\ \chi, & k = 0 \\ \sinh\chi, & k = -1. \end{cases} \tag{13}$$

The two-dimensional analogue to the Friedmann–Robertson–Walker metric is

$$\mathrm{d}l^2 = a^2\left[\mathrm{d}\chi^2 + f^2(\chi)\mathrm{d}\theta^2\right] \tag{14}$$

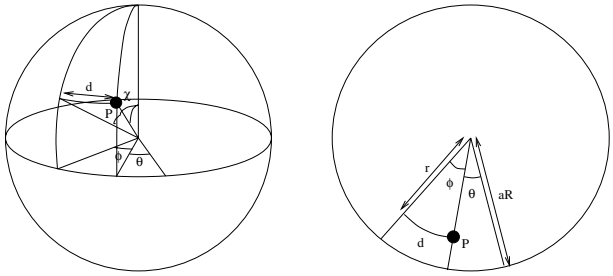


Figure 1. The topology of space in a closed ($k = +1$) two-dimensional universe. The coordinates (χ, θ) define the position of a point P in this universe. The angular width ϕ of an object of size d is illustrated.

which is more amenable to our everyday intuition, especially in the case of a closed $k = +1$ universe where

$$\mathrm{d}l^2 = a^2\left[\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\theta^2\right]. \tag{15}$$

This is simply the expression for the distance on the surface of a sphere where θ is the azimuthal coordinate and where χ is the tangential coordinate (figure 1).

The wavelengths of photons expand, or REDSHIFT, along with the universe. The redshift of an object is a measure of how much the universe and, therefore, the wavelength of that object's photons have expanded until the present. A redshift of zero indicates the present. Quantitatively, we express the relationship between the size of the universe a and its redshift z as

$$a \propto \frac{1}{1+z}. \tag{16}$$

Using solutions to Friedmann equations, one may derive both the age of the universe and the effective distances to objects which existed at earlier epochs. By integrating up the infinitesimal times

$$\begin{aligned} \mathrm{d}t = \frac{1}{\dot{a}}\,da &= \frac{1}{aH_0E(z)} \frac{a\,\mathrm{d}z}{1+z} \\ &= \frac{\mathrm{d}z}{H_0E(z)(1+z)} \end{aligned} \tag{17}$$

one can compute the age of the universe:

$$t = \frac{1}{H_0} \int_{z=0}^{\infty} \frac{\mathrm{d}z}{E(z)(1+z)}. \tag{18}$$

Similarly, one may readily derive expressions for the distance though there is one subtlety. Two different measures of distance are standardly discussed in cosmology: the angular-size distance and the luminosity distance. Both distances are defined so that the standard expressions involving these quantities apply. The former, the angular-size distance, commonly denoted D_A, is defined in analogy with the expression

$$\theta = \frac{d}{D_\mathrm{A}} \tag{19}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where d is an object's intrinsic size and θ is the angular size of the object on the sky in radians. Similarly, the latter, the luminosity distance, commonly denoted $D_{\rm L}$, is defined in analogy with the expression

$$f = \frac{L}{4\pi D_{\rm L}^2} \tag{20}$$

where f is the observed flux and L is the intrinsic luminosity.

We now provide a heuristic derivation of the above equations. Imagine that the light from some object with size d, redshift $z_{\rm obs}$, and tangential coordinate χ converges to $\chi = 0$ and $z = 0$ on paths where $\mathrm{d}\theta = 0$. Along this path, the expression for the metric reduces to $\mathrm{d}l^2 = a^2R^2\mathrm{d}\chi^2$. The integrated coordinate distance χ is then

$$\begin{aligned}\chi &= \int_{z=0}^{z_{\rm obs}} \frac{\mathrm{d}l}{aR} = \int_{z=0}^{z_{\rm obs}} \frac{c\,\mathrm{d}t}{aR} \\ &= \frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}} \frac{c\,\mathrm{d}z}{E(z)}.\end{aligned} \tag{21}$$

The object is observed at tangential coordinate χ at some previous time, and the distance of this object (at χ and $z_{\rm obs}$) from the z-axis (see figure 1) is

$$r(z) = a(z)R\sin\chi = \frac{a_0 R}{1+z}\sin\chi. \tag{22}$$

Clearly, the angle ϕ emanating from the top of the sphere intersecting our object of size d at some distance r from the z-axis is d/r, so

$$\phi = \frac{d}{r}. \tag{23}$$

Identifying $D_{\rm A}$ in equation (19) with $r(z)$ in equation (23) yields

$$\begin{aligned}D_{\rm A} = r(z) &= \frac{a_0 R}{1+z}\sin\chi \\ &= \frac{a_0 R}{1+z}\sin\left[\frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}}\frac{c\,\mathrm{d}z}{E(z)}\right].\end{aligned} \tag{24}$$

A heuristic derivation of the luminosity distance is similarly possible. Imagine the light emitted from some object at redshift $z_{\rm obs}$ and $\chi = 0$ propagates to some χ reaching it at $z = 0$. This light would have spread out over a ring of radius $2\pi r(0)$, where $r(0)$ is again the distance from the points at χ and the z-axis. Hence, $f = L/2\pi r(0)$. The obvious three-dimensional analogue is $f = L/4\pi r(0)^2$. Multiplying the flux by a factor of $1/(1+z)$ to account for time dilation and $1/(1+z)$ to account for the energy loss due to photon redshifting, one obtains

$$f = \frac{L}{4\pi r(0)^2(1+z)^2}. \tag{25}$$

Identifying $r(0)(1+z)$ in equation (25) with $D_{\rm L}$ in equation (20) yields

$$\begin{aligned}D_{\rm L} &= r(0)(1+z) \\ &= a_0 R(1+z)\sin\left[\frac{1}{H_0 a_0 R}\int_{z=0}^{z_{\rm obs}}\frac{c\,\mathrm{d}z}{E(z)}\right].\end{aligned} \tag{26}$$

As in our derivation of the angular and luminosity distances, it is easy to see that the apparent surface brightness of objects decreases as $(1+z)^{-4}$, a factor of $(1+z)^{-1}$ due to time dilation, a factor of $(1+z)^{-1}$ due to a redshifting of the photons, and a factor of $(1+z)^{-2}$ to account for its larger angular size on the sky.

Origin of fluctuations

INFLATION is a popular scenario for producing the small matter overdensities, or seeds, on which mass accretes and galaxies later form. The popularity of inflation derives from the relatively natural explanation it provides for establishing the initial conditions from which our universe seems to have evolved. Not only does it explain the homogeneity of the universe on large scales, but it also solves the problem of the relative absence of objects like magnetic monopoles or other types of massive topological defects. It provides a natural explanation for the apparent flatness of the universe although the observational evidence for this still might be considered preliminary.

Inflation is initiated at the temperature scale corresponding to the breaking of the symmetry of grand unification ($T \sim 10^{16}$ GeV, at $t \sim 10^{-35}$ s) as the universe expands and cools through a brief period of supercooling. The associated release of scalar field energy results in an exponential increase in the scale factor, and corresponding reduction in temperature, and is followed by a period of reheating when the associated field kinetic energy thermalizes. This early epoch of exponential expansion is driven by the evolution of a scalar inflation field ϕ and potential V, the field evolving as

$$\ddot{\phi} + 3\frac{\dot{a}}{a} + \frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} = 0. \tag{27}$$

The pressure and density are

$$p_\phi = \dot{\phi}^2 + V \tag{28}$$

$$\rho_\phi = \dot{\phi}^2 - V \tag{29}$$

where the inflation potential is such that $V >> \dot{\phi}^2$, so $\rho = -p = V$ as in the case of a non-zero cosmological constant.

Zero point fluctuations of the scalar field are imprinted on the causal horizon scale, which increases exponentially. The causal horizon is greatly enlarged, to encompass all of the universe observed today, and the reheating process imprints a scale-invariant distribution of energy density fluctuations. A perturbative analysis of this scalar field's evolution predicts a power spectrum with slope close to unity: $P(k) \propto k$, where k is the wavenumber defined via the Fourier transformation of the underlying Gaussian-distributed density field $\rho(x,t)$: $\delta_k = \int \rho(x,t)\mathrm{e}^{\mathrm{i}k\cdot x}\mathrm{d}^3x$. This power spectrum is called the Harrison–Zeldovich scale-free power spectrum. The amplitude of the fluctuations is not predicted by the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

inflationary theory. However the prediction of scale-invariance is generic to many, although not all, inflationary models. Scale-independent curvature fluctuations are equivalent to density fluctuations on the horizon scale of fixed amplitude. On superhorizon scales, an invariant $\delta\phi = G\,\delta M/L\,c^2 = \delta\rho/\rho)(L/ct)^2 = (\delta\rho/\rho)\,L^2_{\rm comoving}\,(a/ct)^2$ can be found which does not change with time.

The energy density of photons and other massless particles scales as $(1+z)^4$ while the energy density of nonrelativistic massive particles scales as $(1+z)^3$. Consequently, there is an epoch of matter-radiation equality before which the energy density of the universe is dominated by radiation (massless particles), and after which the energy density of the universe is dominated by matter (massive particles). The redshift at this epoch is

$$1 + z_{\rm eq} = \frac{\rho_{m_0}}{\rho_{r_0}} = 3 \times 10^4\,\Omega\,h^2 \tag{30}$$

where the matter density is $\rho_{\rm m} = \rho_{m_0}(1+z)^3$ and relativistic density is $\rho_{\rm r} = \rho_{r_0}(1+z)^4$. The horizon scale at this epoch is $L_{\rm eq} \equiv 2\,ct_{\rm eq} = 12\,h^{-1}$ Mpc, in comoving coordinates.

During the radiation-dominated era, perturbations in the primordial power spectrum $(\delta\rho/\rho)$ grow in proportion to the square of the universal scale-factor a^2. The growth of perturbations stalls once they get inside the horizon (Hubble radius) since the time scale for the growth of perturbations $(\propto (\rho_{\rm mat}G)^{-1/2})$ is large compared to the expansion rate of the universe $(\propto (\rho_{\rm tot}G)^{-1/2})$. However once the universe becomes matter-dominated, perturbations resume growth, but now in proportion to a, the universal scale-factor, instead of the square of this scale.

The stalling of growth on small scales within the Hubble radius imprints itself on the power spectrum which survives the transition to a matter-dominated universe. Subhorizon density fluctuations have a distribution with scale $\delta\rho/\rho \propto L^{-2}_{\rm comoving} \propto M^{-2/3}$, on scales $\lesssim L_{\rm eq}$, while on larger scales growth suppression implies that $\delta\rho/\rho \approx$ constant. Mathematically, we include this effect by multiplying the primordial power spectrum $P_{\rm p}$ for linearized fluctuations in a stochastic density field by the transfer function

$$P_{\rm f}(k) = T(k)\,P_{\rm p}(k) \tag{31}$$

to obtain the final processed power spectrum $P_{\rm f}$. An approximate analytical fit of a transfer function calculated numerically is given by

$$T(k) = \frac{\log(1+2.34q)}{(2.34q)} \times \left[1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4\right]^{-1/4} \tag{32}$$

where $q = (k/{\rm Mpc}/(\Omega h^2))$ (Peacock 1997, Bardeen *et al* 1986) and $h = H_0/(100\,{\rm km\,s^{-1}\,Mpc^{-1}})$. Here we note that the above power spectrum results in a bottom-up sequence of evolution, smaller scale fluctuations reaching large amplitude before the large-scale fluctuations.

Dissipationless non-radiative matter, known otherwise as DARK MATTER, is now a standard part of the paradigm for galaxy formation. It is important in reconciling the power spectrum observed in the COSMIC MICROWAVE BACKGROUND with the local power spectrum observed in galaxies. Nonbaryonic dark matter makes this possible because the growth of fluctuations is able to commence immediately after the era of matter-radiation equality. The growth of fluctuations in the baryonic component is, by contrast, suppressed because of its coupling to radiation background. Eventually, as the universe expands, it becomes cool enough so that radiation decouples from the baryons, an event known as last scattering. At this point, the mass scale above which fluctuations can grow suddenly drops from $\sim 10^{16} M_\odot$ to $\sim 10^6 M_\odot$, and so the fluctuations in the baryonic component are free to grow with the rest of the universe. Since the dark matter component is the densest portion of the universe, it controls the growth of fluctuations, and soon after decoupling, the baryonic component is boosted in amplitude by $\sim (1+z_{\rm eq})/(1+z_{LS})$, z_{LS} being the redshift at last scattering.

Remarkably, the acoustic modes set up in the baryonic component prior to last scattering are observable via their imprint on the cosmic microwave background radiation which provides a snapshot of fluctuations at last scattering. This may be seen as follows. The dispersion relation for fluctuation growth at rate $e^{i\omega t}$ is

$$\omega^2 = k^2 V_{\rm S}^2 - 4\pi G\rho \tag{33}$$

($V_{\rm S}$ being the sound speed) for time scales short compared to the expansion time, or equivalently for wavelengths $2\pi/h \ll ct$. Such wavelengths oscillate as sound waves with amplitude proportional to $\exp(ik\,V_{\rm S}\,t)$. Inflation specifies primordial curvature fluctuations which are time-invariant on superhorizon scales. The corresponding amplitude corresponds to the mode $\cos(k\,V_{\rm S}\,t)$ which is finite as $k \to 0$. At $t_{\rm LS}$ the maximum amplitude in $\delta T/T$ is attained by fluctuations which have entered the horizon and satisfy $kV_{\rm S}t = \pi n$, with $n = 1, 2, \ldots$. Clearly the wave that crests on the horizon at $t_{\rm LS}$ and undergoes oscillations systematically experiences both Compton drag and diffusive damping. The net result is the amplitude is diminished and only the first three or so peaks are detectable. Doppler effects contribute 90 degrees out of phase to the gravitational potential induced $\delta T/T$ peaks, and the net result is a series of peaks corresponding to the maximum in $\delta T/T = |\frac{1}{3}\delta\phi - \underline{r}\cdot\underline{v} + \frac{1}{3}\delta\rho/\rho|$. The predicted enhancement of the first peak of wavelength $2\,V_{\rm S}\,t_{\rm LS}$ is as large as a factor of three relative to the $k \to 0$ or Sachs–Wolfe potential fluctuation limit that corresponds to the inflationary imprint of primordial, nearly scale-invariant curvature fluctuations.

The observational situation is as follows. The COBE DMR experiment measured potential fluctuations on angular scales $\geq 10^\circ$. The last scattering horizon is about 1°. Several experiments report evidence for the first acoustic peak on this scale, and definitive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

measurements with large sky coverage that provide adequate foreground discrimination were performed in 2000 by the Boomerang and Maxima balloon experiment collaborations (de Bernardis *et al* 2000, Hanany *et al* 2000). The situation is complicated by the contribution from the time-varying potential perturbations $\delta T/T \sim \int \dot{\phi}\, \mathrm{d}t$ which arise in models with $\Omega < 1$ at curvature-dominated epochs, i.e. $z \lesssim \Omega^{-1} - 1$ and are expected at angular scales $\gtrsim 1^\circ$. A final complication is the possible tensor mode contributions; this gravity wave background contributes to the energy density in the radiation-dominated era and only affects the low order multipoles corresponding to angular scales that are larger than the horizon angular scale at t_{eq}.

Linear evolution of density fluctuations

Gravitation magnifies the initial perturbations in the matter distribution, both dark and baryonic. While these perturbations initially grow linearly, eventually these perturbations stop expanding, break away from Hubble flow, and then collapse and virialize. Galaxies form from the gas which cools onto the center of these collapsed masses, called dark halos.

An analysis of the growth of the initial matter power spectrum begins with the basic fluid equations in terms of the density ρ, the velocity $\boldsymbol{u}$, the gravitational potential Φ, the pressure p, and the time t:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0, \quad \rho \frac{D\boldsymbol{u}}{Dt} = -\nabla p - \rho \nabla \Phi. \tag{34}$$

For convenience, we change variables to comoving coordinates $\boldsymbol{x} = \boldsymbol{r}/a$, to peculiar velocities $\boldsymbol{v} = a\mathrm{d}\boldsymbol{u}/\mathrm{d}t$, to dimensionless overdensities $\delta = \rho/\bar{\rho} - 1$, and to conformal times $\tau = t/a$. Making these substitutions, the fluid equations become

$$\boldsymbol{v} = \dot{\boldsymbol{x}} \tag{35}$$

$$\dot{\delta} + \nabla \cdot [(1+\delta)\boldsymbol{v}] = 0 \tag{36}$$

$$\dot{\boldsymbol{v}} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} + \frac{\dot{a}}{a}\boldsymbol{v} = -\frac{\nabla p}{\rho} - \nabla \Phi \tag{37}$$

while Poisson's equation reads

$$\nabla^2 \Phi = 4\pi G \bar{\rho} a^2 \delta. \tag{38}$$

The behavior of the modes can be obtained by taking the divergence of Euler's equation, eliminating $\nabla \cdot \boldsymbol{v} = 0$ by the continuity equation, taking Φ, δ, and $\boldsymbol{v}$ to be small, and then linearizing the equations. One obtains

$$\ddot{\delta} + \frac{\dot{a}}{a}\dot{\delta} = \frac{\nabla^2 p}{\rho} + 4\pi G \bar{\rho} a^2 \delta. \tag{39}$$

If we consider a pressure-free universe, this equation involves no spatial derivatives, so its solution can then be written as

$$\delta(\boldsymbol{x}, \tau) = A(\boldsymbol{x}) f_1(\tau) + B(\boldsymbol{x}) f_2(\tau). \tag{40}$$

Using the fact that $\Omega = 8\pi G\rho/3H^2 = 8\pi G\rho a^2/3\dot{a}^2$, we can rewrite equation (39) as

$$\ddot{\delta} + \frac{\dot{a}}{a}\dot{\delta} - \frac{3}{2}\Omega \left(\frac{\dot{a}}{a}\right)^2 \delta = 0. \tag{41}$$

For an Einstein–de Sitter universe ($\Omega = 1$), $a \propto \tau^2 \propto t^{2/3}$ and equation (41) becomes

$$\ddot{\delta} + 2\frac{\dot{\delta}}{\tau} - 6\frac{\delta}{\tau^2} = 0. \tag{42}$$

For solutions, one obtains a growing mode $\delta \propto D(\tau) \propto \tau^2 \propto t^{2/3}$ and a decaying mode $\delta \propto D(\tau) \propto \tau^{-1} \propto t^{-1/3}$. $D(\tau)$ is known as the linear growth factor. Given the small size of the perturbations at the era of matter–radiation equality, one ignores the decaying mode solution and simply considers the growing mode. For the more general case of a universe without a cosmological constant (i.e., $\Omega_\Lambda = 0$), it can be shown (Peebles 1980) that

$$D(\tau) = 1 + \frac{3}{x} + \frac{3(1+x)^{1/2}}{x^{3/2}} \ln\left[(1+x)^{1/2} - x^{1/2}\right] \tag{43}$$

where $x = \Omega^{-1} - 1 \propto a$.

Before finishing this discussion, we examine the evolution in position and velocity of a test particle because these quantities will prove useful later when we look at how galaxies acquire ANGULAR MOMENTUM. Since $\nabla^2 \Phi_0 = 4\pi G \bar{\rho} a^2 \delta$ is a constant in the linear regime for an Einstein–de Sitter universe, the linearized version of Euler's equation can be immediately integrated to give

$$\boldsymbol{v} = -\left(a^{-1} \int D\mathrm{d}\tau\right) \nabla \Phi_0. \tag{44}$$

Integrating the peculiar velocity gives us:

$$\boldsymbol{x} = \boldsymbol{x}_0 - \left(\int \frac{\mathrm{d}\tau}{a} \int D\mathrm{d}\tau\right) \nabla \Phi_0. \tag{45}$$

Since the growth factor $D(\tau)$ is a solution to the fluctuation growth equation $\frac{\mathrm{d}}{\mathrm{d}\tau}(a\dot{\delta}) = 4\pi G \bar{\rho} a^3 \delta$ (see equation (39)), it follows that

$$\boldsymbol{x} = \boldsymbol{x}_0 - \frac{D(\tau)}{4\pi G \bar{\rho} a^3} \nabla \Phi_0 \tag{46}$$

and

$$\boldsymbol{v} = -\frac{\dot{D}}{4\pi G \bar{\rho} a^2} \nabla \Phi_0. \tag{47}$$

This essentially Lagrangian view to the growth of perturbations is due to Zeldovich (1970).

Mean square fluctuations

It is useful to consider the evolution of the mean square mass fluctuations filtered on various spatial scales R because of the information it provides on the collapse of structure on these spatial scales. Due to the scale-free growth of initial perturbations, the power spectrum $P(k, \tau)$

$$P(k, \tau) = \left| \int \mathrm{d}^3 x \delta(\boldsymbol{x}, \tau) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \right|^2 \tag{48}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

grows as $D(\tau)^2$ just as $\delta(\boldsymbol{x}, \tau)$ does.

Assuming that a section of the power spectrum can be written as

$$P(k, \tau) \propto D(\tau)^2 k^n \tag{49}$$

where n is the power spectrum index, we find that the mean square fluctuations filtered on a particular mass scale are

$$\begin{aligned}\sigma^2(M) &= \left\langle (dM/M)^2 \right\rangle \\ &= \int \mathrm{d}^3 k W(kR) P(k, \tau) \\ &\propto D(\tau)^2 R^{-3-n} \\ &\propto D(\tau)^2 M^{-(n+3)/3}\end{aligned} \tag{50}$$

where $W(kR)$ is a window function that filters out spatial scales of R or smaller. The top hat filter of radius R

$$W(k) = \frac{3}{kR^3}(\sin kR - kR\cos kR). \tag{51}$$

is one commonly used window function. Note that $\sigma(M, 0) = (M/M_{nl})^{-(n+3)/6}$ where $M_{nl} = 1 \times 10^{15}\sigma_8 \Omega h^{-1} M_\odot$ corresponds to mass fluctuations of amplitude unity on $8h^{-1}$ Mpc radius spheres. The factor σ_8 is $\sigma(M, 0)$ at $8h^{-1}$ Mpc for the mass density: galaxy count fluctuations have unit amplitude on this scale.

Halo scaling relations

The growth of fluctuations becomes nonlinear and start collapsing into virialized objects when

$$\left\langle (\delta M/M)^2 \right\rangle = 1. \tag{52}$$

From equation (50), this implies that

$$M \propto D(t)^{6/(3+n)}. \tag{53}$$

The density ρ, size r, and temperature T of collapsed objects then scale as

$$\rho \propto (1+z)^3 \tag{54}$$

$$r \propto (M/\rho)^{1/3} \propto (1+z) D(t)^{6/(3+n)} \tag{55}$$

$$\begin{aligned}T \propto V_\mathrm{c}^2 &\propto \frac{GM}{r} \propto M^{2/3}\rho^{1/3} \\ &\propto (1+z)^1 D(t)^{4/(3+n)}.\end{aligned} \tag{56}$$

These scaling relationships will be useful in later discussions, particularly when we examine the cooling of gas in virialized structures.

Spherical collapse model

To understand the nonlinear growth of structure, it is convenient to consider the idealized spherical 'top hat' collapse model. Here, one supposes there to be spherically symmetric regions of radius R and of uniform overdensity $\bar{\delta}$ in an otherwise uniform universe at some initial time t_i. At these early times, the universe will be approximately Einstein–de Sitter, so we can express this overdensity as the sum of the growing and decaying modes:

$$\delta = \delta_+ \left(\frac{t}{t_\mathrm{i}}\right)^{2/3} + \delta_- \left(\frac{t}{t_\mathrm{i}}\right)^{-1}. \tag{57}$$

We take the matter in this region to be expanding at the approximately the same rate as the universe, and therefore we require that the peculiar velocity be zero:

$$\tfrac{2}{3}\delta_+(t_\mathrm{i}) - \delta_-(t_\mathrm{i}) = 0. \tag{58}$$

Therefore, $\delta_+ = \frac{3}{5}\delta$. According to Birkhoff's theorem, in a spherically symmetric situation, matter external to the sphere will not influence its evolution, so it follows that

$$\frac{\mathrm{d}^2 R}{\mathrm{d}t^2} = \frac{-GM}{R^2} = \frac{-4\pi G}{3}\bar{\rho}(1+\bar{\delta})R \tag{59}$$

which is identical in form to the equation for the evolution of the cosmological scale factor a,

$$\frac{\mathrm{d}^2 a}{\mathrm{d}t^2} = \frac{-GM}{a^2} = \frac{-4\pi G}{3}\bar{\rho}a. \tag{60}$$

Therefore, the size of the region R evolves like the cosmic scale factor a but with an initial density parameter $\Omega_\mathrm{p}(t_\mathrm{i})$ given by

$$\Omega_\mathrm{p}(t_\mathrm{i}) = \frac{\rho(t_\mathrm{i})(1+\bar{\delta})}{\rho_\mathrm{c}(t_\mathrm{i})} \tag{61}$$

where $\rho_\mathrm{c}(t_\mathrm{i})$ is the critical density ($3H_\mathrm{i}^2/8\pi G$ at time t_i). By analogy with the solutions for the universe, the region will collapse if $\Omega_\mathrm{p} > 1$. By analogy with equation (4), the expansion of the region evolves according to

$$\left(\frac{\dot{a}}{a}\right)^2 = H_\mathrm{i}^2 \left[\Omega_\mathrm{p}(t_\mathrm{i})\frac{a_\mathrm{i}^3}{a^3} + (1 - \Omega_\mathrm{p}(t_\mathrm{i}))\frac{a_\mathrm{i}^2}{a^2}\right]. \tag{62}$$

Eventually, the region stops expanding, turns around, and collapses. At the turn-around time t_m, $\dot{a} = 0$, implying that $a/a_\mathrm{i} = \Omega_\mathrm{p}(t_\mathrm{i})/(1 - \Omega_\mathrm{p}(t_\mathrm{i}))$. The density at turn-around t_m is then

$$\rho_\mathrm{p}(t_\mathrm{m}) = \rho_\mathrm{c}(t_\mathrm{i})\Omega_\mathrm{p}(t_\mathrm{i})\left[\frac{\Omega_\mathrm{p}(t_\mathrm{i}) - 1}{\Omega_\mathrm{p}(t_\mathrm{i})}\right]^3. \tag{63}$$

By analogy with the solutions to the cosmological equations for closed universes (see equations (9)–(11)), we can determine t_m as

$$\begin{aligned}t_\mathrm{m} &= \frac{\pi}{2H_\mathrm{i}}\frac{\Omega_\mathrm{p}(t_\mathrm{i})}{[\Omega_\mathrm{p}(t_\mathrm{i}) - 1]^{3/2}} = \frac{\pi}{2H_\mathrm{i}}\left[\frac{\rho_\mathrm{c}(t_\mathrm{i})}{\rho_\mathrm{p}(t_\mathrm{m})}\right]^{1/2} \\ &= \left[\frac{3\pi}{32 G \rho_\mathrm{p}(t_\mathrm{m})}\right]^{1/2}.\end{aligned} \tag{64}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

For an Einstein–de Sitter universe, the density $\rho(t_m)$ of the universe at turn-around is simply

$$\rho(t_m) = \frac{1}{6\pi G t_m^2} \tag{65}$$

so that

$$\chi = \frac{\rho_p(t_m)}{\rho(t_m)} = \left(\frac{3\pi}{4}\right)^2 \approx 5.6. \tag{66}$$

Formally, after turn-around, the region will collapse to a point at $t = 2t_m$. Of course, before that happens, shell crossings will occur. From the virial theorem, dissipationless matter collapses to a radius that is one-half the turn-around radius. Since $a \propto t^{2/3}$, the universe will expand by $2^{2/3}$. Hence, the density of the region relative to the background universe at collapse t_c is

$$\frac{\rho_p(t_c)}{\rho(t_c)} = (2^{2/3})^3 8\chi \approx 180. \tag{67}$$

An extrapolation of the linear growth estimate at time t_c yields

$$\delta_+(t_c) = \frac{3}{5}\delta_i \left(\frac{2t_m}{t_i}\right)^{2/3} = \frac{3}{5}\left(\frac{3\pi}{2}\right)^{2/3} \approx 1.68. \tag{68}$$

The assumption of a uniform spherical overdensity in an otherwise uniform universe is quite unrealistic; in fact, collapse typically proceeds toward the creation of a large number of two-dimensional pancakes. Nevertheless, numerical simulations show the basic scalings derived here to be roughly correct and useful for making simple analytic estimates.

Form of collapsed structures

For many years, the profiles of collapsed halos were taken to be that of a isothermal sphere:

$$\rho(r) = \frac{V_c^2}{4\pi G r^2} \tag{69}$$

where r is the radius and V_c is the circular velocity. Over the last few years, however, the detailed N-body simulations of Navarro *et al* (1997) have shown that the collapsed matter profile is better fitted by the double power-law profile:

$$\rho(r) = \rho_{crit} \frac{\delta_0}{(r/r_s)(1 + r/r_s)^2} \tag{70}$$

where r_s is the core radius where the slope of the power-law changes from -1 near the center to -3 at large radii and r_{200} is the virial radius of the halo. More recent high resolution simulations find a somewhat steeper innermost slope to the density profile, possibly as steep as -1.5 (Jing and Suto 2000, Moore *et al* 1999).

Press–Schechter model

A convenient and somewhat approximate description of the mass function of halos at particular redshifts is given by the Press–Schechter formalism derived heuristically by Press and Schechter (1974) using linear growth theory and the spherical 'top hat' model. The mass function of non-linear objects is computed with the aid of linear theory on the assumption that the probability distribution of density fluctuations at a given mass and redshift is Gaussian, centered on the mean $\sigma(M, z) \equiv \left\langle (\delta\rho/\rho)^2 \right\rangle^{1/2}$.

Since the set of initial perturbations is Gaussian-distributed and remains so under linear growth, we can write the distribution of mass fluctuations δ_M as

$$P(\delta_M)\mathrm{d}\delta_M = \frac{1}{\sqrt{2\pi}\sigma(M)} \exp\left(\frac{-\delta_M^2}{\sigma(M)^2}\right) \mathrm{d}\delta_M \tag{71}$$

where $\sigma(M)$ is equal to the mean square mass fluctuations $\langle (\delta M/M)^2 \rangle$ defined earlier.

The fraction of points where the mean density inside a radius R exceeds δ_c is

$$P_{>\delta_c}(M) = \int_{\delta_c}^{\infty} P(\delta_M)\mathrm{d}\delta_M. \tag{72}$$

Press and Schechter (1974), using the spherical collapse model as a heuristic guide, took all regions where the mean density exceeded the critical value needed for linear collapse, i.e., $\delta_c \approx 1.68$, to be part of collapsed structures at least as massive as $M = 4\pi\bar{\rho}a^3R^3/3$.

Then, to distinguish the fraction of structures which have just collapsed to a mass M and those that are part of a bigger structure at least as massive as $M + \mathrm{d}M$, we subtract $P_{>\delta_c}(M + \mathrm{d}M)$ from $P_{>\delta_c}(M)$. Notice that by doing this, we completely ignore the possibility that the mass scale just collapsing is contained within a larger mass that is just collapsing, a complication known as the 'cloud-in-cloud' problem. Another problem is that only half of the points are associated with an overdensity and therefore become part of any collapsed structure. Press and Schechter (1974) elected to solve this problem somewhat arbitrarily by multiplying the number of structures at a given mass scale by a factor of 2 with the vague understanding that this represents flow from underdense to overdense regions.

The number of halos $n(M)$ with masses between M and $M + \mathrm{d}M$ is

$$n(M)M\mathrm{d}M = 2\rho_M[P_{>\delta_c}(M) - P_{>\delta_c}(M + \mathrm{d}M)]\mathrm{d}M, \tag{73}$$

ρ_m being the average density at the redshift in question and the mass term M on the left-hand-side accounting for the fact that more massive halos are associated with more points of reference. Rewriting this, we get

$$\begin{aligned} n(M)\mathrm{d}M &= \frac{2\rho_M}{M}\left|\frac{\mathrm{d}P_{>\delta_c}(M)}{\mathrm{d}M}\right|\mathrm{d}M \\ &= \frac{2\rho_M}{M}\left|\frac{\mathrm{d}P_{>\delta_c}(M)}{\mathrm{d}\sigma(M)}\right|\left|\frac{\mathrm{d}\sigma(M)}{\mathrm{d}M}\right|\mathrm{d}M. \end{aligned} \tag{74}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finally, since the mean square fluctuations $\sigma^2(M)$ evolve as $D(\tau)^2$ according to linear growth (see equation (50)), we can rewrite the expression as

$$n(M)\mathrm{d}M = \frac{2\rho_\mathrm{M}}{M^2}\frac{\delta_\mathrm{M}}{\sigma(M)}\frac{1}{\sqrt{2\pi}\sigma(M)} \times \exp\left(\frac{-\delta_\mathrm{c}^2}{2\sigma(M)^2}\right)\left|\frac{\mathrm{d}\sigma(M)}{\mathrm{d}M}\right|\mathrm{d}\ln M. \quad (75)$$

This is the Press–Schechter mass function at redshift z. Despite the extremely heuristic nature of this derivation and the many problems already discussed, it agrees remarkably well with numerical simulations, and thus has proven very useful in characterizing the growth of structure.

Cooling processes

Galaxies are much less massive than the mass scales ($\sim 10^{14} M_\odot$) going nonlinear in the universe today, so clearly galaxies must be more than simply virialized structures. The key seems to be the process of cooling and the time scale for the settling of baryons into the centers of their dark matter halos. There are several cases to consider. Clearly, if the cooling time of a gas is larger than the Hubble time, the gas cannot have evolved much over the history of the universe. If the cooling time is smaller than the Hubble time but larger than the dynamical time, the gas will suffer slow quasistatic collapse into the center of the virialized halo. On the other hand, if the cooling time is smaller than the dynamical time, the ambient gas will undergo runaway cooling and collapse to the center of the virialized halo. It is this case, where the cooling time is much shorter than the dynamical time scales for ACCRETION or merging (Binney 1977, Rees and Ostriker 1977, Silk 1977) that is relevant for the formation of galaxies.

There are four important processes by which gas in halos cools: Compton cooling, free–free emission (bremsstrahlung), recombination, and collision-induced de-excitation.

We begin with a consideration of the Compton cooling process. When low-energy photons pass through a gas of non-relativistic electrons, they scatter off the electrons with the Thompson cross-section σ_T:

$$\sigma_\mathrm{T} = \frac{8\pi}{3}\left(\frac{e^2}{m_\mathrm{e}c^2}\right)^2 \quad (76)$$

where m_e is the mass of the electron and e is the charge of the electron. Some photons are scattered up in energy and some are scattered down, but the net effect is to slow the electrons relative to the frame of the cosmic microwave background radiation. The mean shift in photon energy per collision is

$$h\overline{\Delta\nu} = \frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}h\nu \quad (77)$$

where k is Boltzmann's constant, h is Planck's constant, ν is the frequency of a photon, and T_e is the temperature of the electron gas. In a sea of photons with temperature T_γ, the mean rate of energy loss per electron is

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}\sigma_\mathrm{T}aT_\gamma^4 \quad (78)$$

where a is the Stefan–Boltzmann constant. The cooling time t_cool is

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT_\mathrm{e}}{n_\mathrm{e}\frac{4kT_\mathrm{e}}{m_\mathrm{e}c^2}\sigma_\mathrm{T}aT_\gamma^4} = \frac{3m_\mathrm{e}c^2}{8\sigma_\mathrm{T}aT_\gamma^4} \sim 2.1\times10^{12}(1+z)^{-4}\ \mathrm{yr}. \quad (79)$$

High temperature (10^6–10^7 K) primordial gases are almost entirely ionized. Under these circumstances, the dominant cooling mechanism is due to the acceleration of electrons off the bare H^+ and He^{2+} nuclei. This results in a cooling rate per unit rate per unit volume:

$$\frac{\mathrm{d}E}{\mathrm{d}t} \propto n_\mathrm{e}n_\mathrm{H}T^{1/2}. \quad (80)$$

The cooling time t_cool here is approximately equal to

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT}{\frac{\mathrm{d}E}{\mathrm{d}t}} = 6.6\times10^9\frac{T_6^{1/2}}{n_{-3}}\ \mathrm{yr} \quad (81)$$

where $T_6 = T/10^6\,\mathrm{K}$ and $n_{-3} = n/10^{-3}\mathrm{cm}^{-3}$.

On the other hand, low temperature (10^4–10^5 K) primordial gases are only partially ionized. Here cooling is dominated by two processes: one where electrons recombine with ions resulting in the release of a photon (recombination) and one where electrons collide with partially ionized atoms, thereby exciting them to a state which they escape by the release of a photon. The total cooling rate can be expressed as

$$\frac{\mathrm{d}E}{\mathrm{d}t} \propto n_\mathrm{e}n_\mathrm{H}f(T). \quad (82)$$

The latter process is the dominant one, and for primordial abundances, the function $f(T)$ can be approximated as $2.5(T/10^6\,\mathrm{K})^{-1/2}\,\mathrm{erg\,cm^3\,s^{-1}}$. The cooling time t_cool is then

$$t_\mathrm{cool} = \frac{\frac{3}{2}nkT}{\frac{\mathrm{d}E}{\mathrm{d}t}} = 3.0\times10^9\frac{T_6^{3/2}}{n_{-3}}\ \mathrm{yr}. \quad (83)$$

We compare these cooling time scales with the dynamical time scales $t_\mathrm{dyn} \sim \sqrt{1/G\rho}$. We consider a uniform spherical cloud with mass M in virial equilibrium with f fraction of its mass in dissipative baryonic matter and the rest in dark, dissipationless matter, so the gas mass M_g is equal to fM. For this mass configuration,

$$t_\mathrm{dyn} \sim \sqrt{\frac{1}{G\rho}} \sim \sqrt{\frac{1}{Gn/f}} \sim 6.5\times10^9 f^{1/2}n_{-3}^{-1/2}\ \mathrm{yr}. \quad (84)$$

Now, we compare this dynamical time scale with the cooling times derived for each of the cooling processes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

discussed above. At early times, $z > 10$, the Compton cooling process dominates, and since

$$n_{-3} = 2.3 \times 10^{-2} f(1+\delta)(\Omega_0 h^2)(1+z)^3 \tag{85}$$

the dynamical time scale goes as

$$t_{\rm dyn} \sim 3.0 \times 10^9 (\Omega_0 h^2)^{-1/2}(1+z)^{-3/2}\ {\rm yr} \tag{86}$$

and the relative time scale τ goes as

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} \approx 6 \times 10^2 (1+z)^{-5/2}. \tag{87}$$

Thus, Compton cooling will be important at $z > 7$ where $\tau < 1$. Notice that the relative time scale does not depend upon mass, temperature, or density. Therefore, if galaxies formed at early times, they would have no preferred scales.

The relevant temperatures for lower mass halos ($\lesssim 10^{12} M_\odot$) are less than 10^6 K. In this case, line cooling dominates and the relative time scale τ goes as $t_{\rm cool}/t_{\rm dyn} \propto (T^{3/2}/\rho^{1/2}) \propto ((M^{2/3}\rho^{1/3})^{3/2}/\rho^{1/2}) \propto M$. Therefore, the $\tau = 1$ line runs parallel to lines of constant mass. To determine the mass limit more precisely, we look at τ:

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} = 0.4 T_6^{3/2} f^{-1/2} n_{-3}^{-1/2}. \tag{88}$$

We can relate this to the mass of a spherical cloud model in virial equilibrium by using the following relation from the virial theorem:

$$\frac{3kT}{2\mu} = \frac{0.3GM}{R}, \tag{89}$$

where μ, the mean molecular weight, is roughly equal to half the proton mass $m_{\rm p}$ since the medium is ionized. From this, it follows that $T^3 \propto \rho M^2 \propto n f^{-1} M^2$ and then that

$$M_{\rm g} = 1.2 \times 10^{13} T_6^{3/2} f^{3/2} n_{-3}^{-1/2} M_\odot. \tag{90}$$

Hence,

$$\tau = \frac{M_{\rm g}}{1.2 \times 10^{13} f^2 M_\odot} = \frac{M}{1.2 \times 10^{13} f M_\odot}. \tag{91}$$

This sets the mass limit below which gas can effectively cool to form structures. For $f \sim 1$, the mass limit is much larger than the typical limiting galaxy mass ($\sim 10^{12} M_\odot$), but for smaller values consistent with constraints set by big-bang NUCLEOSYNTHESIS ($f \sim 0.05 - 0.1$), the mass limit is comparable to these limits.

On the other hand, for higher mass halos ($\gtrsim 10^{12} M_\odot$), the relevant temperatures are greater than $10^6 K$. Here the dominant cooling mechanism is bremsstrahlung, and the relative time scale ratios τ go as $t_{\rm cool}/t_{\rm dyn} \propto (T^{1/2}/\rho^{1/2}) \propto ((M^{2/3}\rho^{1/3})^{1/2}/\rho^{1/2}) \propto R$, so the $\tau = 1$ line runs parallel to lines of constant radius. To determine the limiting radius more precisely, we look at τ

$$\tau = \frac{t_{\rm cool}}{t_{\rm dyn}} = T_6^{1/2} f^{-1/2} n_{-3}^{-1/2}. \tag{92}$$

Using the spherical cloud model again, we solve for R in terms of the other variables,

$$R = \sqrt{\frac{3fkT}{0.8\pi\mu^2 G n_{-3}}} = 610 f T_6^{1/2} n_{-3}^{-1/2}\ {\rm kpc} \tag{93}$$

and so

$$\tau = \frac{R}{610 f^{3/2}\ {\rm kpc}}. \tag{94}$$

Hence, for $f \sim 0.1$, massive gas clouds of radii greater than 20 kpc can efficiently cool. Since this length is smaller than the typical cluster size, cooling is not very efficient in clusters, and therefore the gas simply suffers slow quasistatic collapse.

The galaxy cluster mass function

In the previous section, we discussed two important different regimes for virialized masses, one in which the cooling time was longer than the dynamical time and one in which it was shorter than it. In the former regime, one obtains GALAXY CLUSTERS where most of the gas remains hot and in the latter regime one obtains galaxies where much of the halo gas has apparently cooled. In either case, one can use Press–Schechter theory to calculate the mass function, and with simple assumptions about the conversion of gas into stars or other luminous objects, one can convert this into a LUMINOSITY FUNCTION OF GALAXIES.

Perhaps the most direct comparison with observations is via the mass function of galaxy clusters. The shape, expected to be exponential plus a power law tail, fits the prediction remarkably well, to the extent that cluster masses are well determined. Three techniques are used to estimate cluster masses: galaxy velocity dispersion and distribution, hot gas temperature and distribution, and gravitational lensing maps. The first two methods assume virial equilibrium. All three methods give consistent results, to within a factor of 2 in mass. One can compare the characteristic cluster mass, determined by the fitting function

$$\frac{{\rm d}N}{{\rm d}M} \propto M^{-\alpha} \exp\left(-(M/M_{\rm nl})^\beta\right), \tag{95}$$

with the predicted value of $M_{\rm nl}$ taken from field galaxy counts and a bias factor that has to be empirically deduced. Indeed, $M_{\rm nl}$ corresponds to a typical observed cluster mass. The normalization of the cluster mass function depends both on the mean density and $\sigma(M, Z)$, with an exponential sensitivity to $\sigma(M, Z)$. Only five percent of galaxies are in clusters, which can therefore account for perhaps one percent of the critical density. Clusters are therefore rare objects, typically 3σ fluctuations. The number density of clusters is controlled by both the mean density and σ_8, in the combination $\sigma_8\,\Omega^{0.6} \approx 0.7 \pm 0.2$. The scale $8h^{-1}$ Mpc, corresponding to unit amplitude of the optical counts and the mass M_8 of a typical cluster, is used for normalization, and $\sigma(M, 0) = \sigma_8 (M_8/M)^{(n+3)/6}$ where $M_8 = 4\pi(8h^{-1}{\rm Mpc})^3 \Omega\bar{\rho}$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The galaxy luminosity function

The Press–Schechter formulation can also be used to derive the galaxy luminosity function, which is described by the Schechter function

$$\frac{\mathrm{d}N}{\mathrm{d}L} \propto L^{-\alpha} \exp(-L/L_*) \tag{96}$$

where $L_* \approx 10^{10} L_\odot$ and $\alpha \approx 1$ to 1.5 depending on the galaxy selection criterion. There are two ingredients that must be incorporated in connecting mass to luminosity in order to obtain a satisfactory comparison of theory and observation. A characteristic luminosity L_* must be derived, and the formation of low mass objects into luminous objects must be inefficient, since the mass function tail has slope M^{-2} whereas the luminosity function is flatter. The characteristic luminosity follows from the fact that cooling is efficient only for masses up to $\sim 10^{12} M_\odot$. Then, assuming a typical baryon fraction, an age of $\sim 10^{11}$ yr, and a standard mass function for stars, one derives a characteristic luminosity of $\sim 10^{10} L_\odot$.

Angular momentum

As structures grow and collapse in the early universe, they exert tidal torques on each other, and this provides each collapsing mass with some angular momentum. This angular momentum, in turn, is important in determining the final properties of the disk and elliptical galaxies which form inside these collapsed structures.

The angular momentum of a collapsing halo can be expressed as

$$\boldsymbol{J} = \int_{\mathrm{V}} \mathrm{d}^3 x \bar{\rho} a^3 (a\boldsymbol{x} - a\bar{\boldsymbol{x}}) \times \boldsymbol{v} \tag{97}$$

where $\bar{\boldsymbol{x}}$ is the center of mass for the volume. Using equation (47), we express $\boldsymbol{v}$ as $-a\dot{b}\nabla\Phi_0$ where $b(\tau) = D/4\pi G\bar{\rho}a^3$. For convenience we expand $\nabla\Phi_0$ in a Taylor series around the point $\boldsymbol{x}$:

$$\begin{aligned} \nabla\Phi_0|_{\boldsymbol{x}} &= \nabla\Phi_0|_{\bar{\boldsymbol{x}}} + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \frac{\partial^2 \Phi_0}{\partial \boldsymbol{x}^2}|_{\bar{\boldsymbol{x}}} \\ &= \nabla\Phi_0|_{\bar{\boldsymbol{x}}} + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \underline{\underline{T}} \end{aligned} \tag{98}$$

where $T_{jl} = \partial^2\Phi_0/\partial x_{\mathrm{j}} \partial x_{\mathrm{l}}$. Rewriting this, we get

$$J_{\mathrm{i}}(\tau) = -a\dot{b}\epsilon_{ijk} T_{jl} \int_V (x_{\mathrm{l}} - \bar{x}_{\mathrm{l}})(x_{\mathrm{k}} - \bar{x}_{\mathrm{k}}) \bar{\rho} a^3 \mathrm{d}^3 x \tag{99}$$

or

$$J_{\mathrm{i}}(\tau) = -a\dot{b}\epsilon_{ijk} T_{jl} I_{lk} \tag{100}$$

where I_{lk} is the inertial tensor.

We now estimate how J_{i} scales. Since I_{lk} scales as a^2 until collapse while T_{jl} continues to scale as $\nabla\Phi/a^2 \sim (D/a)/a^2 \sim 1/a^2$, each structure effectively acquires angular momentum from its neighbors until collapse. Since the collapse of a structure occurs when $\delta \sim 1$, $b \sim D(\tau)/4\pi G\bar{\rho}a^3 \sim 1/4\pi G\bar{\rho}a^2 \sim 1/\nabla^2\Phi$ from Poisson's equation and the relation $D(\tau) \propto a$, so T_{jl} scales as $\nabla^2\Phi_0 \sim 1/b$. $\underline{\underline{I}}$ scales as $MR_0^2 \sim MR^2/a^2$. Hence,

$$\begin{aligned} J_{\mathrm{i}}(\tau) &\sim -a\dot{b} T_{jl} I_{lk} \sim a\dot{b} \frac{1}{b} \frac{MR^2}{a^2} \\ &\sim \frac{\dot{b}}{\dot{a}b} \frac{\dot{a}}{a^2} M(M/\rho)^{2/3} \\ &\sim \Omega^{0.6} H(\Omega H^2)^{-2/3} M^{5/3} \\ &\sim \Omega^{-0.07} t^{1/3} M^{5/3}. \end{aligned} \tag{101}$$

It is standard to construct a dimensionless quantity which characterizes the angular momentum that each collapsed mass has acquired via tidal torques. This quantity is called the dimensionless angular momentum λ, and it can be expressed as

$$\lambda = \frac{|\boldsymbol{J}||E|^{1/2}}{GM^{5/2}}. \tag{102}$$

Noting that $|E| \sim M^2/R \sim M^2(\rho/M)^{1/3} \sim M^{5/3}(\Omega H)^{1/3} \sim \Omega^{1/3} M^{5/3} t^{-2/3}$, we see that $\lambda \sim |\boldsymbol{J}||E|^{1/2} M^{-5/2} \sim \Omega^{-0.07} t^{1/3} M^{5/3} \Omega^{1/6} t^{-1/3} M^{5/6} M^{-5/2} \sim \Omega^{0.1}$. Therefore, the distribution of dimensionless angular momenta λ is essentially independent of a halo's mass, collapse time, or even the basic world model. N-body simulations (Warren *et al* 1992, Cole and Lacey 1996, Catelan and Theuns 1996) and analytical treatments (Steinmetz and Bartelmann 1995) find a distribution which is well fitted by the expression

$$p(\lambda) = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\left[-\frac{\ln(\lambda/\bar{\lambda})^2}{2\sigma_\lambda^2} \right] \frac{\mathrm{d}\lambda}{\lambda} \tag{103}$$

where $\bar{\lambda} = 0.05$ and $\sigma_\lambda = 0.5$.

Disk formation

Disk galaxies make up the dominant component of the local galaxy census. Disk galaxies are known to have exponential profiles ($I(r) \propto \exp(-r/r_{\mathrm{d}})$), to have significant fractions of dust and stars, to still be undergoing some STAR FORMATION, and to be rotationally supported. They are extremely flattened objects and can appear very elongated if viewed edge-on. They also frequently have long bars and spiral structures. It is because of this latter feature that these galaxies are often called SPIRAL GALAXIES.

Most of the global disk properties, e.g., mass, luminosity, size, and metallicity, define a two-dimensional manifold with little scatter about that manifold. It is more well-known in terms of its two-dimensional projections, in particular, the well-known TULLY–FISHER RELATION between luminosity and circular velocity. There are two main views on this tight relationship: one in which these processes as consequences of self-regulating mechanisms for star formation in disks (e.g. Silk 1997) and one in which this is simply the consequence of the cosmological equivalence of mass and circular velocity (e.g. Mo *et al* 1998).

In the past, disk galaxies were thought to have surface brightnesses tightly distributed around 21.65 b_{J} mag/arcsec^2 (Freeman 1970). Shortly after this claim

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

was made, arguments were made that there was a strong selection bias against low surface brightness galaxies and in reality the spread in surface brightness extended to much lower values (Disney 1976). Recently, there have been a large number of efforts to quantify the bivariate luminosity–surface brightness distribution (de Jong 1996, McGaugh 1996, Dalcanton *et al* 1997, Sprayberry *et al* 1997, de Jong and Lacey 1999). While the results are somewhat different in terms of their details, they suggest that the surface brightness distribution of galaxies peaks around 22 b_J mag/arcsec^2 with a spread of $\sim 1 - 1.5$ mag/arcsec^2. The luminosity function of spiral galaxies is also nicely described by the Schechter function.

The typical values of the dimensionless angular momentum of collapsed halos (~ 0.05) are considerably smaller than that of the largely flattened centrifugally-supported disk galaxies we observe in our universe today ($\sim 0.4 - 0.5$), so considerable dissipation must occur to produce these disks. Without the presence of dissipationless dark matter, the collapse would proceed in such a way that the total angular momentum $\boldsymbol{J}$ and the total mass m would be conserved, but the energy would scale as $1/R$ where R is the collapse factor, so that $\lambda \propto JE^{1/2}M^{-5/2} \propto 1/\sqrt{R}$. The disk would then need to collapse by a factor of $(0.5/0.05)^2 \sim 100$ to obtain its observed dimensionless angular momenta, and this would take longer than the age of the universe for a 10-kpc disk! However, if the gas cloud collapses inside a dark matter halo, for which it represents only a fraction f of the mass, then the angular momentum $\boldsymbol{J}$ and mass $\boldsymbol{M}$ would scale by a factor f and the energy $\boldsymbol{E}$ would scale by a factor f^2, so that $\lambda \propto JE^{1/2}M^{-5/2} \propto 1/(f^{1/2}R^{1/2})$. For a typical estimate of the baryon fraction, $f \sim 0.1$, the gas cloud would then only need to collapse by a factor of 10, easily accommodated in current theories.

Despite the simplicity of this picture, a significant portion of the available gas cools to form GALAXIES AT HIGH REDSHIFT. Detailed simulations which follow the evolution and merging of these galaxies into larger and larger systems produce disks whose sizes are much smaller than those observed (Steinmetz and Navarro 1999) because of substantial angular momentum transfer from the baryons to their dissipationless halos.

Feedback

A nonnegligible fraction of stars end their lives as SUPERNOVAE, injecting much of this energy into the ambient gaseous medium. This energy serves to heat the gas, either expelling it from the star-forming environment or making it too hot to be conducive to star formation. Hence, the formation of stars serves to suppress further star formation and hereby regulates itself. This process is quite logically called feedback. The presence of feedback, particularly in disk galaxies, explains why the conversion of gas into stars frequently requires ten to hundreds of dynamical time scales ($\sim 10^{10}$ yr) instead of just several dynamical time scales ($\sim 10^8$ yr).

Feedback also provides the preferred explanation for the flattening of the luminosity function relative to the mass function at low masses (see the section above on galaxy luminosity function). DWARF GALAXY potential wells are shallow, and interstellar gas is readily energized above the escape velocity and therefore blown out in a galactic wind. Evidence for galactic winds is commonly found for STARBURST GALAXIES, often of relatively low mass.

Elliptical galaxy formation

Ellipticals make up the other principal component of the local galaxy census. Ellipticals possess elliptical isophotes with projected ellipticities $\epsilon = a/b$ (a being the major axis and b the minor) ranging from 0 to 0.7, the former being denoted an E0 and the latter an E7. Low redshift ellipticals possess an abundance of low-mass stars and are therefore very red. The lack of short-lived blue stars is generally taken as an indication that these galaxies are very old and have not formed stars for at least 5–10 Gyr. Like spirals, the luminosity function for ellipticals can also be described by a Schechter function, but with a much shallower faint-end slope (Bromley *et al* 1998, Folkes *et al* 1999). Unlike spirals, ellipticals are predominantly found in dense regions, i.e., galaxy clusters (Dressler 1980).

Ellipticals are known to have approximately de Vaucouleurs surface brightness profiles:

$$I(r) \propto \exp(-7.67(r/r_e)^{1/4}) \tag{104}$$

where r is the radius and r_e is the half-light radius. To higher order, the surface brightness profiles of ellipticals show an important dichotomy. Some ellipticals, known as disky ellipticals, appear to have power-law profiles all the way into their center, and other ellipticals, known as boxy ellipticals, exhibit a sharp break from this power-law at some radius near the center.

Like spirals, the global structural properties of ellipticals are known to populate a two-dimensional manifold, commonly known as the fundamental plane. These are known according to various names: the Faber–Jackson (Faber and Jackson 1976) relationship ($L \propto \sigma^4$), the Kormendy luminosity–radius (Kormendy 1977) relationship, and the D_n-σ (Dressler *et al* 1987) relationship. It has largely been agreed that the fundamental plane is essentially a consequence of the virial theorem and a relatively homologous formation scenario where the mass-to-light ratio varies as a small power of the mass ($M/L \propto M^{1/6}$).

There are two prevailing scenarios for the formation of elliptical galaxies: one in which ellipticals formed as the result of mergers from spiral galaxies and one in which ellipticals formed at high redshift from monolithic collapse. We begin by presenting the monolithic collapse scenario.

Monolithic collapse

One possible mechanism for the formation of elliptical galaxies is the early formation of stars from the gas collapsing onto the center of a dark halo. Early collapse and fragmentation into stars prior to the collapse of the halo can constitute the core of the elliptical while

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars formed from the secondary gas infall can constitute the shallower wings. An examination of the velocity-dispersion rotational-rate relationship demonstrates that ellipticals are essentially pressure-supported and that rotational flattening is not important in imparting ellipticity to these galaxies. In fact, detailed comparisons show that the dimensionless angular momenta of slow-rotating ellipticals seems to be no larger 0.05. In order to obtain the typical mass M ($\sim 10^{11} M_{\odot}$), radius R (~ 10 kpc), and angular momentum without recourse to dissipation, it would be necessary for the halo to collapse at redshifts beyond 10. On the other hand, with dissipation, one could easily obtain galaxies of the desired mass and radius, but the dimensional angular momentum would be too large (unless the initial angular momentum for the halo just happened to be particularly small), and the galaxy would resemble a disk.

Merger-based origin

Another mechanism for the formation of ellipticals is through the mergers of spiral galaxies. This mechanism provides a natural way of resolving the angular momentum problem, the crucial point being that since the spin angular momenta are randomly oriented with respect to each other, the resultant spin angular momentum for the formed elliptical can be considerably smaller than the spin angular momentum of the colliding disks. There are a number of other attractive features to this scenario. First, there are numerous examples of disk galaxies merging to form objects with de Vaucouleurs profiles in the local universe (Schweizer 1982, 1986), and it is quite conceivable that mergers were more frequent in the past. Secondly, nearly half of elliptical galaxies (Malin and Carter 1983, Schweizer and Ford 1984) possess features, such as shells or other sharp features, indicative of mergers or an otherwise violent formation. Thirdly, detailed N-body simulations of collisions between disk galaxies embedded in dark halos produce galaxies with de Vaucouleur profiles similar to those found in nearby ellipticals. Fourth, the GLOBULAR CLUSTER populations around ellipticals have bimodal metallicity distributions, indicative of a multi-stage formation scenario (Ashman and Zepf 1992, Zepf and Ashman 1993). All these features point toward the conclusion that at least *some* ellipticals formed by merging.

Before discussing the relative merits of the two formation scenarios for ellipticals, it is interesting to look at several of the difficulties which only arise in the merging scenario because of the close relationship between ellipticals and their progenitors (spirals). First, the energy per particle and phase space density are higher at the centers of ellipticals than any observed spiral, and therefore the merging process must be accompanied by a great deal of gas dissipation and cooling both to form a much deeper central potential and to obtain the high phase space density observed there if we presume this scenario is correct. In fact, nuclear starbursts are frequently observed to accompany such mergers (Schweizer 1990). Second, the number of globular clusters ($\sim 10^4$–$10^5 M_{\odot}$ compact star clusters) per unit luminosity for ellipticals is typically 4–10 times larger than that for spirals (van den Bergh 1990), so disk-disk mergers must result in the formation of a large number of globular clusters if we presume this scenario is correct. Finally, while ellipticals might be expected to show relative alpha-to-iron abundances typical of spirals, ellipticals contain significantly larger abundances of alpha elements than iron elements, so a substantial fraction of the stars present in ellipticals must have formed in the merger events between two spiral galaxies.

A comparative evaluation

The principal observational differences between the monolithic collapse and merger scenarios for elliptical formation concern their predictions for the formation history of ellipticals. Monolithic scenarios tend to form elliptical galaxies at very high redshifts ($z > 3$) while the elliptical population builds up more gradually in hierarchical scenarios.

Consequently, the merger scenarios, with their more diverse and contemporary formation histories, show more scatter in both the colour-magnitude relationship and the fundamental plane than monolithic collapse scenarios. Observationally speaking, ellipticals show a high degree of uniformity both in their small color-magnitude scatter, i.e. $\sigma(U - V) = 0.15$ (Bower *et al* 1992) and their tightness around the fundamental plane (Renzini and Ciotti 1993). This observed tightness about the fundamental plane extends to $z \sim 1$ (Aragon-Salamanca *et al* 1993, Stanford *et al* 1998). The observed tightness supports a monolithic collapse scenario where ellipticals form early and somewhat coevally. Of course, in hierarchical scenarios, most galaxies assemble quite early ($z \sim 2$) in the rich clusters, where the most compelling examples of tight fundamental planes are observed, so apparent difficulties with this scenario are not as strong as they first might seem (Kauffmann and Charlot 1998a).

Due to the different formation times for ellipticals, these scenarios also yield remarkably different predictions for the evolution in the number density of early-type galaxies as a function of redshift. While there has been an increasing number of studies reporting a devolution in the number and luminosity of ellipticals at high redshift relative to that found in the local universe (Kauffmann and Charlot 1998b, Kauffmann *et al* 1996, Zepf 1997, Barger *et al* 1999, Menanteau *et al* 1999) as would be expected in a hierarchical scenario where their formation is more gradual, these results remain somewhat controversial (Broadhurst and Bouwens 1999, de Propris *et al* 2000).

Another important difference between these scenarios is the star formation rates they predict at high redshift. In the hierarchical scenario, galaxies start out small and slowly build up to the massive entities we observe in the universe today. Clearly, we do not expect large star formation rates here at early times except possibly when two galaxies merge. On the other hand, in the monolithic scenario, ellipticals need to undergo huge star formation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rates ($\sim 100\,M_\odot$/yr) to form the typical $10^{11}\,M_\odot$ stars observed in nearby giant elliptical galaxies at high redshift since there is only a period of $\sim 10^9$ yr available. In fact, very few galaxies with these huge star formation rates ($\sim 100\,M_\odot$/yr) have been found in either emission-line searches or Lyman-dropout searches at moderate redshifts ($1 < z < 5$), pointing to either a high redshift of formation or dust-enshrouding. Recently, however, SCUBA results and subsequent follow-up work have revealed a population of ultraluminous infrared galaxies at moderate redshifts with high enough star formation rates ($\sim 100\,M_\odot$/yr) to match those needed in a monolithic collapse scenario. Nevertheless, the exact nature of this population, its number density, and its relevance remain unclear.

Bulges

Many spiral galaxies feature a bulge, or a spheroid, at their centers. Spheroids resemble elliptical galaxies in many important respects including their overall appearance and placement in the fundamental plane. This suggests that bulges are nothing but elliptical galaxies upon which gas has later accreted. Note, however, that somewhat contrary to ellipticals, in particular ellipticals with boxy isophotes, is the presence of considerable rotational flattening in many bulges (Davies *et al* 1983, Davies 1987). This is in agreement with what one might expect from dissipational collapse and, in particular, from the formation of bulges via disk instabilities (van den Bosch 1998).

Summary

While there are many things we do not understand about galaxy formation, many pieces of the picture now seem to be clear. Galaxies seem to form in a homogeneous, isotropic universe that is expanding according to Friedmann's equations. Inflation, though not unique, appears to be a relatively successful way of producing the scale-free spectrum of density fluctuations out of which galaxies have formed. Growth of the fluctuations can be followed initially with linear growth theory and later using a spherical collapse model. Press–Schechter theory provides a relatively successful way of putting these ingredients together to predict the mass spectrum of collapsed objects. The relative magnitudes of the cooling and dynamical time scales are important for determining the mass range of galaxies, galaxies forming when the cooling time is smaller than the dynamical time. Disk galaxies form from the cooling of gas onto the centers of collapsed halos, the gas settling into a disk supported by its angular momentum. Elliptical galaxies, on the other hand, seem to form by disk–disk mergers or by gas cooling within a halo of low intrinsic angular momentum (monolithic collapse).

Many important questions remain. For example, what is the relative importance of different mechanisms for the formation of both ellipticals and bulges? How do the sizes, luminosities, star formation rates, number densities, and metallicities of various galaxy types evolve over the history of the universe? What mechanisms are responsible for the tight correlation between the global properties of ellipticals and spirals? While theoretical simulations are becoming increasingly sophisticated, the inherent nonlinearity of galaxy formation processes make the role of new observations tantamount. To give the reader a taste of the improvements we will see in the next ten years in probing galaxy formation in the most remote regions of the universe, in figure 2 we have included some simulations for a hierarchical merging model using two current generation instruments (WFPC2 and NICMOS) and two future generation instruments (ACS and NGST). The obvious increase in depth will clearly bring our already moderately mature understanding of galaxy formation further into focus.

Bibliography

Aragon-Salamanca A, Ellis R S, Couch W J and Carter D 1993 *Mon. Not. R. Astron. Soc.* **262** 764
Ashman K M and Zepf S E 1992 *Astrophys. J.* **384** 50
Bardeen J M, Bond J R, Kaiser N and Szalay A S 1986 *Astrophys. J.* **304** 15
Barger A J, Cowie L L, Trentham N, Fulton E, Hu E, Songaila A and Hall D 1999 *Astron. J.* **117** 102
Binney J 1977 *Astrophys. J.* **215** 483
Bower R G, Lucey J R and Ellis R S 1992 *Mon. Not. R. Astron. Soc.* **254** 601+
Broadhurst T and Bouwens R 2000 *Astrophys. J.* **530** 53
Bromley B C, Press W H, Lin H and Kirshner R P 1998 *Astrophys. J.* **505** 25
Catelan P and Theuns T 1996 *Mon. Not. R. Astron. Soc.* **282** 455
Cole S and Lacey C 1996 *Mon. Not. R. Astron. Soc.* **281** 716+
Dalcanton J, Spergel D N, Gunn J E, Schmidt M and Schneider D P 1997 *Astron. J.* **114** 2178+
Davies R 1987 *Structure and Dynamics of Elliptical Galaxies* IAU Symposium vol 127, ed T de Zeeuw (Dordrecht: Reidel)
Davies R L, Efstathiou G, Fall S M, Illingworth F and Schechter P L 1983 *Astrophys. J.* **266** 41
de Bernadis P *et al* 2000 *Nature* **404** 955
de Jong R and Lacey C 1999 *Astrophys. Sp. Sci.* **269** 569
de Jong R S 1996 *Astron. Astrophys.* **313** 45
de Propris R, Sanford A, Eisenhardt P, Dickenson M and Elston R 1999 *Astron. J.* **118** 719
Disney M J 1976 *Nature* **263** 573
Dressler A 1980 *Astrophys. J.* **236** 351
Dressler A, Lynden-Bell D, Burstein D, Davies R L, Faber S M, Terlevich R and Wegner G 1987 *Astrophys. J.* **313** 42
Faber S M and Jackson R E 1976 *Astrophys. J.* **204** 668
Folkes S, Ronen S, Price I, Lahav O, Colless M, Maddox S, Deeley K, Glazebrook K, Bland-Hawthorn J, Cannon R, Cole S, Collins C, Couch W, Driver S, Dalton G, Efstathiou G, Ellis R, Frenk C, Kaiser N, Lewis I, Lumsden S, Peacock J, Peterson B, Sutherland W and Taylor K 1999 *preprint* astro-ph/9903456
Freeman K C 1970 *Astrophys. J.* **160** 811+
Hanany S *et al* 2000 *preprint* astro-ph/0005123
Jing Y P and Suto Y 2000 *Astrophys. J* **529** L69

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

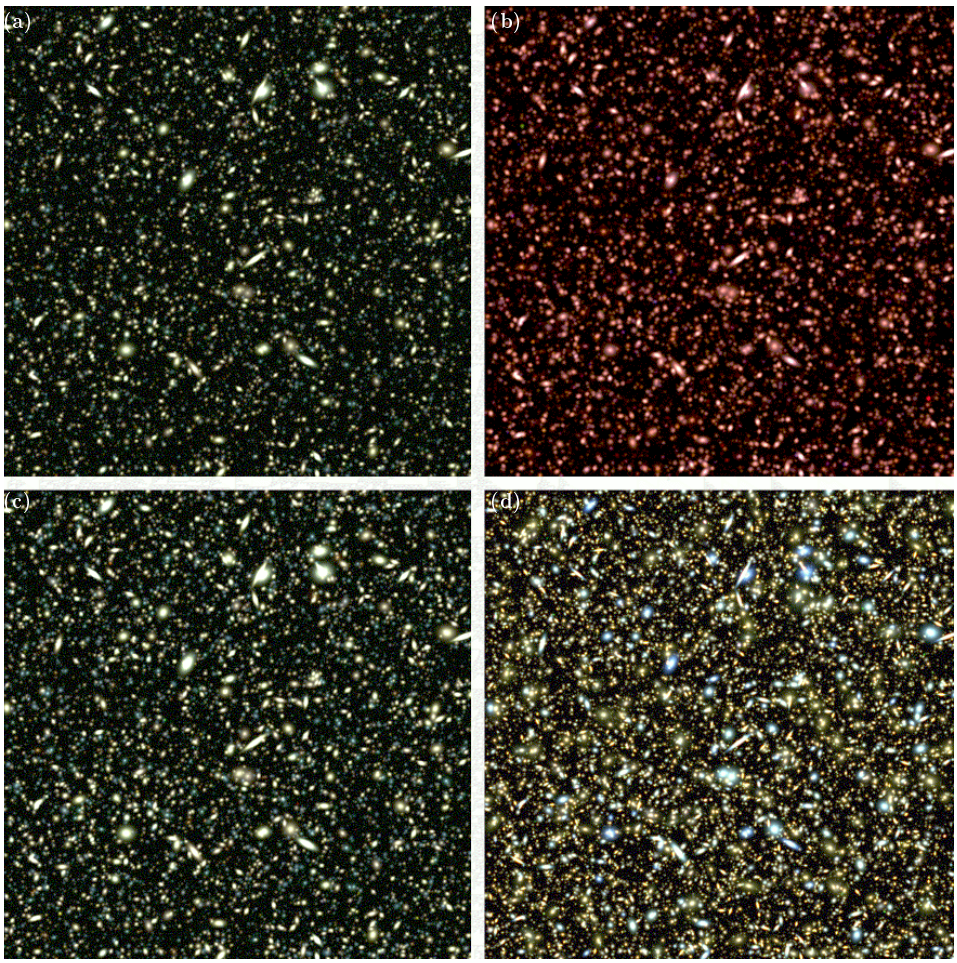


Figure 2. A simulation of HDF-depth ($\sim$ 30 orbits) exposure over a $7' \times 7'$ field of view for WFPC2 (a), NICMOS (b), ACS (c) and NGST (d). WFPC2 and NICMOS are instruments currently on the Hubble Space Telescope. ACS is an instrument that will be placed on the Hubble Space Telescope in 2001, and NGST is a completely new space telescope, which will be launched in $\sim$ 2007. A hierarchical merging model was used as the model inputs for the simulation (see Bouwens and Silk 1999 for details). 5 μm, 3 μm and 1 μm wavelengths are assumed for the RGB channels in the false-colour NGST image while for the NICMOS image, the K (2.2 μm), H (1.6 μm) and J (1.2 μm) bands are assumed. **This figure is reproduced as Color Plate 18.**

Kauffmann G and Charlot S 1998a *Mon. Not. R. Astron. Soc.* **294** 705+
Kauffmann G and Charlot S 1998b *Mon. Not. R. Astron. Soc.* **297** L23
Kauffmann G, Charlot S and White S D M 1996 *Mon. Not. R. Astron. Soc.* **283** L117
Kormendy J 1977 *Astrophys. J.* **218** 333
Malin D F and Carter D 1983 *Astrophys. J.* **274** 534
McGaugh S S 1996 *Mon. Not. R. Astron. Soc.* **280** 337
Menanteau F, Ellis R, Abraham R, Barger A and Cowie L 1999 *Mon. Not. R. Astron. Soc.* **309** 208
Mo H J, Mao S and White S D M 1998 *Mon. Not. R. Astron. Soc.* **295** 319
Moore B, Quinn T, Governato F, Stadel J and Lake G 1999 *Mon. Not. R. Astron. Soc.* **310** 1147
Navarro J F, Frenk C S and White S D M 1997 *Astrophys. J.* **490** 493+
Peacock J A 1997 *Mon. Not. R. Astron. Soc.* **284** 885
Peebles P 1980 *The Large-Scale Structure of the Universe* (Princeton, NJ: Princeton University Press)
Press W H and Schechter P 1974 *Astrophys. J.* **187** 425
Rees M J and Ostriker J P 1977 *Mon. Not. R. Astron. Soc.* **179** 541
Renzini A and Ciotti L 1993 *Astrophys. J.* **416** L49
Schweizer F 1982 *Astrophys. J.* **252** 455
Schweizer F 1986 *Science* **231** 227
Schweizer F 1990 *Dynamics and Interactions of Galaxies* ed R Wielen (Berlin: Springer)
Schweizer F and Ford W K 1984 *Bull. Astron. Astrophys. Soc.* **16** 889+
Silk J 1977 *Astrophys. J.* **211** 638
Silk J 1997 *Astrophys. J.* **481** 703+
Sprayberry D, Impey C D, Irwin M J and Bothun G D 1997 *Astrophys. J.* **482** 104+
Stanford S A, Eisenhardt P R and Dickinson M 1998 *Astrophys. J.* **492** 461+

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Steinmetz M and Bartelmann M 1995 *Mon. Not. R. Astron. Soc.* **272** 570
Steinmetz M and Navarro J F 1999 *Astrophys. J.* **513** 555
van den Bergh S 1990 *Dynamics and Interactions of Galaxies* ed R Wielen (Berlin: Springer)
van den Bosch F C 1998 *Astrophys. J.* **507** 601
Warren M S, Quinn P J, Salmon J K and Zurek W H 1992 *Astrophys. J.* **399** 405
Zel'dovich Y B 1970 *Astron. Astrophys.* **5** 84+
Zepf S E 1997 *Nature* **390** 377+
Zepf S E and Ashman K M 1993 *Mon. Not. R. Astron. Soc.* **264** 611+

Joseph Silk and Rychard Bouwens

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Morphology–Density Relation

The galaxy morphology–density relation describes the strong correlation that exists between the mix of the basic morphological types of galaxy—elliptical (E), spiral (Sp) and lenticular (S0)—in a given volume of space and the local density of galaxies within that volume.

Morphological types of galaxy

As was first recognized by Edwin HUBBLE, galaxies fall into distinct morphological groups fundamental to which are two basic structural features: a more or less flat *disk* of stars and sometimes gas and dust, and a *spheroidal* component. ELLIPTICAL (E) GALAXIES are pure spheroidal systems and were named after the smooth, ellipsoidal shaped light distributions that they project onto the sky. SPIRAL (Sp) GALAXIES comprise both a disk and a spheroid and are distinguished by the conspicuous spiral arm structure in their disks. The LENTICULAR (S0) GALAXIES are, in comparison, somewhat of a transitional object, containing both a spheroid and a disk, but the latter exhibiting no spiral structure.

As well as recognizing these basic morphological types and establishing a classification scheme for them—what is known as Hubble's 'tuning fork' sequence—Hubble was also amongst the first to recognize that the morphological mix of galaxies, that is the relative numbers of E, S0 and Sp types, changed considerably in different environments. The most conspicuous evidence for this came from contrasting the central regions of nearby rich clusters of galaxies—the densest galaxy environments known—with the lower-density regions outside clusters, commonly referred to as the 'field'. The former were seen to be dominated by E and S0 types, with Sp galaxies being almost absent, whereas the field was rich in spirals and E and S0 types were in the minority. A variation also in the morphological mix within clusters has since been well established, with the Sp fraction found to rise and the E and S0 fractions to decrease with increasing radius from the center of rich, regular clusters.

Quantitative studies

The first serious attempt to quantify these well perceived trends was made by Oemler in 1974. He morphologically classified the galaxy populations within 10 nearby rich clusters and derived the global fractions of E, S0 and Sp galaxies in each. These fractions appeared to roughly segregate the clusters into three different classes which Oemler designated 'cD', 'spiral poor' and 'spiral rich'. The cD clusters were those dominated by E and S0 galaxies (making up ∼35% and ∼45% of the cluster population respectively) and corresponded to those with a central cD galaxy and a smooth, centrally concentrated galaxy distribution. In contrast, Sp galaxies constituted about 60% of the population in spiral-rich clusters and these were systems which happened to be highly irregular in shape with little if any degree of central concentration. Spiral-poor clusters possessed intermediate properties both in terms of their overall structure and morphological fractions (E = 15%, S0 = 55%, Sp = 30%). Oemler suggested that these three different types of cluster represented successive stages in the cluster dynamical evolution process which in itself drove a change in the morphological mix.

While this approach of measuring the global morphological fractions in clusters is useful in revealing the gross differences that exist between the different classes of cluster and also between clusters and the field, it falls far short of identifying the fundamental physical variable to which morphological mix is related. For one, radial gradients in the morphological mix are seen within clusters, suggesting that cluster-centric radius is important. On the other hand, this quantity is meaningless in irregular clusters where there is no obvious center and the distribution of galaxies is chaotic and lumpy. Motivated by these problems, Dressler used a sample of 55 rich clusters to further explore the variation of morphological mix and its dependencies. Noticing that the high-density galaxy clumps scattered within the irregular clusters in his sample had a similar morphological mix to the dense cores of the regular clusters, he investigated whether *local galaxy density* was a quantity more germane to the change of morphological mix with environment. This indeed proved to be the case, with strikingly well-behaved and monotonic relations found between morphological fraction and local projected galaxy density over five orders of magnitude in the latter. These are shown in figure 1 and, as Dressler emphasized, they provide very strong evidence that it is the environment *local* to a galaxy that is most fundamental in determining its morphology. It is the behavior encapsulated in this diagram that is commonly referred to as 'the galaxy morphology–local density relation'.

To construct figure 1, Dressler measured some 6000 galaxies, mostly within 2.1 deg^2 fields centered on each of his 55 clusters (all of which were at $z < 0.06$), but also within 15 similarly sized 'field' regions. His imaging material consisted of photographic plates, the majority of which were taken on the 100 inch Du Pont telescope at LAS CAMPANAS OBSERVATORY in Chile. These were used to visually assign each galaxy a morphological type (E, S0, Sp or Irr (irregular)) and to measure its position, the magnitude of its spheroid component and its ellipticity. To measure local density, the 10 nearest (projected) neighbors to each galaxy were identified and the area within which they were contained computed. This allowed the local surface density of galaxies (in galaxies Mpc^{-2}) to be calculated, after having first made a correction for field galaxy contamination. To account for the different redshifts of the clusters, this calculation of the local density was always made down to the same absolute magnitude ($M_V = -20.4$; $H_0 = 50$ km s^{-1} Mpc^{-1}) in each cluster. Marginal distributions were then formed for each morphological type (E, S0, Sp + Irr), giving the number of galaxies as a function of local galaxy density. In combination, these

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

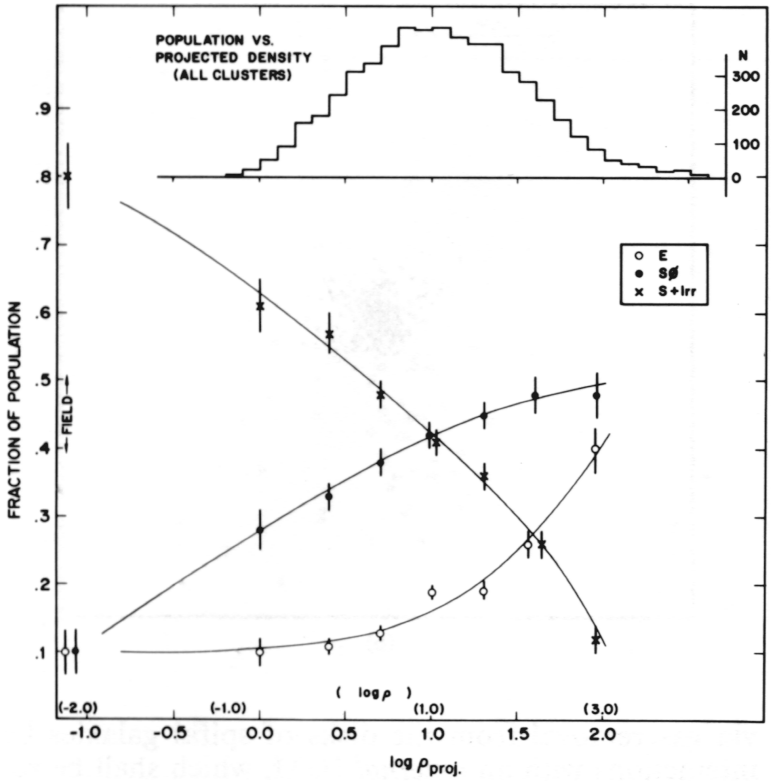


Figure 1. Dressler's galaxy morphology–local density relation. Plotted are the fractions of E, S0 and Sp + Irr galaxies as a function of the logarithm of the local projected galaxy density (see text for details). The upper histogram shows the numbers of galaxies contributing to each of the data points in the lower plot, with the exception of the left-most points which represent the field. The scale shown on the upper side of the horizontal axis indicates the logarithm of the true spatial density of galaxies (in galaxies Mpc^{-3}).

then provided the *fractions* of E, S0 or Sp + Irr types at a given local density value.

The fact that the morphology of a galaxy is closely related to the density of galaxies in its vicinity would also suggest that the local environment is of key importance in the formation and/or evolution of a galaxy (see also GALAXY EVOLUTION, GALAXY FOMATION). However, whether its influence is felt mainly at the time of formation, or as the galaxy subsequently evolves, or during both phases, remains an open question. The paucity of spirals (and the abundance of S0 galaxies) in the cores of regular clusters has long been cited as evidence for the environment playing a large role in their evolution, their encounter with either the hot intracluster gas, other cluster galaxies, or the strong tidal field of the cluster leading to a termination of their star formation and a morphological conversion into S0 galaxies. Dressler, on the other hand, argued that the local environment had its greatest influence at the formation stage, since his galaxy morphology–local density relation showed the S0 fraction to vary slowly with density to the extent that S0s still made up a significant proportion of the galaxy population in regions where the density was too low for gas removal and disruption mechanisms to be effective. Furthermore, the galaxy morphology–local density relation appears to be no different in irregular clusters than in regular clusters, despite the former being insufficiently evolved, dynamically, for these same processes to be in operation.

The relation at high redshift

The study of the galaxy morphology–local density relation at high redshift, and hence at a significantly earlier epoch, provides a direct means of discriminating between these 'formation' and 'evolutionary' scenarios (see GALAXIES AT HIGH REDSHIFT). This became observationally possible for the first time in the early 1990s with the launch of the HUBBLE SPACE TELESCOPE (HST) which, once repaired, provided the one-tenth of an arcsecond spatial resolution necessary to discern the morphologies of galaxies at cosmological distances. Images taken with HST's Wide Field and Planetry Camera of rich clusters at redshifts of $\sim$0.5, have allowed the galaxy morphology–local density relation to be viewed as it was some 6–8 billion years ago. This has yielded a number of major results. Firstly, the fraction of E galaxies seen in these distant clusters is, on average, no different to that seen in present-day clusters. This would imply that ellipticals are old systems which pre-date the clusters and the local environment is more important to their formation, with greater numbers being formed in regions of higher density. In contrast, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

proportion of S0 galaxies in the distant clusters is two to three times smaller than that observed nearby, with a proportional increase in the Sp fraction. This confirms that certainly the cluster environment, at least, had a major influence on the evolution of spirals, converting them into S0s. It also verifies, morphologically, a well known effect seen in distant clusters both photometrically and spectroscopically, namely the presence of many more blue, star-forming galaxies in comparison to clusters at low redshift (named after its discoverers as the 'Butcher–Oemler' effect). As far as the galaxy morphology–local density relation is concerned, gradients in the morphological mix with changing local density are only found in distant clusters which have a regular and centrally concentrated structure, whereas they are completely absent in irregular clusters. Thus the relationship appears to evolve in a hierarchical way, establishing itself first in the richer and denser galaxy systems and then slowly working its way down into the poorer and less dense clusters and groups.

Remaining issues

The final word on the *local* galaxy morphology–local density relation has not rested with Dressler's original work. As part of the HST-based studies of the relation at higher redshift, Dressler's original data set was reanalysed in order to incorporate better, morphologically dependent field galaxy corrections. This had a small but noticeable effect, reducing slightly the spiral fraction at the lowest projected densities. There has also been debate over whether the local projected galaxy density is the sole determinant of morphological mix. A key point in this regard is that if a relation based on projected density was universal, then that could not be the case for the relationship between morphological mix and three-dimensional space density, implying a dependence on some further property that varied from cluster to cluster. It is of note in this context that in Dressler's original study, some marginal evidence was found for a shift in the population levels in clusters with the strongest x-ray emission and thus highest central concentration.

A strong case has also been made for cluster-centric radius alone being the determinant of morphological mix in clusters on the basis that it gives much tighter correlations than local galaxy density or indeed any other cluster property. However, it is not quite this simple, since while the correlations with radius might be stronger, they show a very different behavior in regular and irregular clusters, with the gradients in population mix being much stronger in the former than in the latter. Furthermore, it remains hard to reconcile with the reality of sub-clustering within clusters which has concentrations of galaxies of very similar morphological mix located over a broad range of cluster-centric radius.

These issues notwithstanding, the galaxy morphology–local density relation still remains one that is recognized as being fundamental to galaxies, their morphology and the relationship of this property to environment. The debate surrounding the relation serves as an important reminder that nature is not so simple that it can be described perfectly by a single relationship, but most likely there are secondary factors that come into play and that there are exceptions to the average behavior that this important relation represents.

Bibliography

Butcher H and Oemler A Jr 1978 *Astrophys. J.* **279** 18
Couch W J, Barger A J, Smail I, Ellis R S and Sharples R M 1998 *Astrophys. J.* **497** 188
Couch W J, Ellis R S, Sharples R M and Smail I 1994 *Astrophys. J.* **430** 121
Dressler A 1980 *Astrophys. J.* **236** 351
Dressler A 1984 *Ann. Rev. Astron. Astrophys.* **22** 185
Dressler A, Oemler A Jr, Butcher H R and Gunn J E 1994 *Astrophys. J.* **430** 107
Dressler A, Oemler A Jr, Couch W J, Smail I, Ellis R S, Barger A, Butcher H, Poggianti B M and Sharples R M 1997 *Astrophys. J.* **490** 577
Hubble E and Humason M L 1931 *Astrophys. J.* **74** 43
Oemler A Jr 1974 *Astrophys. J.* **194** 1
Salvador-Sole E, Sanroma M and Jordana J J R 1989 *Astrophys. J.* **337** 636
Whitmore B C, Gilmore D M and Jones C 1993 *Astrophys. J.* **407** 489
Zwicky F 1938 *Publ. Astron. Soc. Pacific* **50** 218

Warrick Couch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Redshift Surveys

The aim of galaxy redshift surveys is to map the three-dimensional distribution of many thousands of galaxies, in order to understand the properties of this distribution and what it implies about the contents and evolution of the universe. It has been known since observations by Hubble in the 1930s that galaxies are not distributed at random, but occur in structures of various sizes. For example, our own LOCAL GROUP contains two large galaxies (our Galaxy and Andromeda) and some 20 medium to very small galaxies. The Local Group is located in the outskirts of the Local Supercluster (see SUPERCLUSTERS AND THE LOCAL SUPERCLUSTER), which is a roughly disk-shaped structure containing around 1000 known galaxies and centered on the well-known VIRGO CLUSTER.

In the 1970s, Peebles and coworkers carried out pioneering statistical measurements of galaxy clustering by analyzing the two-dimensional distribution of galaxies on the sky, notably from the Lick catalogue; it is possible to infer statistical information about the three-dimensional distribution from this, although individual structures cannot be mapped in detail as many structures are superposed in projection on the sky. To obtain the third dimension (i.e. distance) for each galaxy, the galaxy REDSHIFT is used as an approximate distance indicator. The recession of galaxies due to the expansion of the universe causes the observed atomic features in a galaxy spectrum to be Doppler shifted to longer wavelengths relative to their laboratory wavelengths. By Hubble's law the recession velocity v of a galaxy is linearly proportional to its distance r, therefore $v \approx H_0 r$ where H_0 is the HUBBLE CONSTANT; the law is not exact as there is a small extra term due to the 'peculiar velocity' of the galaxy due to the gravitational attraction of the galaxy's neighbors.

Therefore, if we make a three-dimensional map of galaxies using their angular positions on the sky and redshift as the radial direction, this will be a fairly accurate map of the 'true' positions; the scale of our map will depend on the Hubble constant, and the galaxies will be slightly displaced radially from their 'true' positions by the unknown peculiar velocities, but the overall picture will be quite realistic.

Redshift–distance relation

Quantitatively, the redshift z is defined by the fractional wavelength shift:

$$z \equiv \frac{\lambda_{\rm obs} - \lambda_{\rm rest}}{\lambda_{\rm rest}} = \frac{\lambda_{\rm obs}}{\lambda_{\rm rest}} - 1$$

where $\lambda_{\rm obs}, \lambda_{\rm rest}$ are observed and rest wavelengths. Typical wide-area galaxy surveys sample redshifts $z \lesssim 0.1$, so the non-relativistic Doppler formula $z \approx v/c$ is adequate, giving $r \approx cz/H_0$. For low redshifts, the redshift is often quoted for convenience as cz in units of km s^{-1} rather than the dimensionless z, e.g. for the Coma cluster $cz \approx 7000$ km s^{-1} or $z \approx 0.0233$. The distance is then $7000\,{\rm km\,s^{-1}}/H_0$, which is generally written in the literature as $70h^{-1}$ Mpc where $h \equiv H_0/(100\ {\rm km\ s^{-1}\ Mpc^{-1}})$; i.e. the uncertainty in H_0 is absorbed into the scale factor $h \approx 0.7$.

Strictly speaking, the redshift is not a 'true' Doppler effect but really a general-relativistic effect arising from the overall expansion of the universe, whereby photon wavelengths are stretched proportionally to the scale factor of the universe:

$$1 + z = \lambda_{\rm obs}/\lambda_{\rm em} = a_{\rm obs}/a_{\rm em}$$

where $a_{\rm em}, a_{\rm obs}$ are the scale factors at the times of photon emission and observation, and thus $a_{\rm obs}/a_{\rm em}$ is the ratio by which the universe has expanded over this interval. This leads to Hubble's law at low z, while the distance–redshift relation becomes non-linear at $z \gtrsim 0.2$ and depends slightly on cosmological parameters, but the essential point of a monotonic redshift–distance relation still applies.

Measurement of redshifts

Galaxy redshifts are usually measured by standard optical spectroscopy (e.g. figure 1) since it is a good compromise between speed and accuracy. The galaxy's spectrum having been taken, the redshift is obtained either by 'cross-correlation' with a template spectrum to obtain the relative shift in log(wavelength) or by directly identifying and fitting the positions of individual spectral features, e.g. emission lines. The typical precision is 50–100 km s^{-1}, which is quite adequate since galaxy peculiar velocities are larger than this. Redshifts may also be measured with radio telescopes using the 21 cm hydrogen line, which gives very accurate redshifts but is somewhat more time-consuming and is generally only possible at lower redshifts. 'Photometric' redshift estimates can also be obtained from multiband photometry: these are useful for evolutionary studies, but their precision is not good enough for detailed mapping of structure.

Density fluctuations

The main reason for the great interest in mapping the distribution of galaxies is that on large scales it provides a clear 'window' back to conditions in the early universe. If the universe had started out perfectly smooth (apart from the random locations of atoms), it was long ago apparent that galaxies and clusters would not have formed by now; so there must have been some mechanism in the early universe causing the density to vary slightly from place to place. Gravity makes overdense regions pull in more matter from their surroundings, while matter is pulled out of underdense regions and they become more underdense, so the amplitude of these density fluctuations grows with time. On small scales, it is difficult to extract information about the initial fluctuations, because old stars in our Galaxy have made some 50 orbits in its lifetime, so any initial substructures have largely been smeared out over time. Likewise, a galaxy in a typical rich cluster has made around five orbits around the center of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

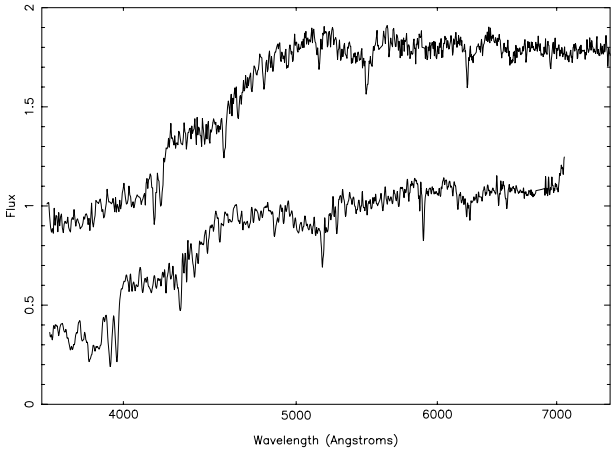


Figure 1. A typical (early-type) galaxy optical spectrum (upper line), with a spectrum of a similar galaxy at zero redshift (lower line). The redshift between spectral features is easily seen as the horizontal wavelength shift, in this case $z = 0.058$.

cluster. However, over the scales of superclusters (tens of megaPARSECS), galaxies have deviated from the pure Hubble expansion by only a fraction of a supercluster's size, so the character of the initial fluctuations is preserved; therefore, mapping the large-scale distrubution of galaxies gives us a clear view of density fluctuations in the early universe. Around 1980, the theory of INFLATION led to some fairly specific predictions about the nature of the initial density fluctuations; the precise way these grow with time depends on the amount and type of dark matter in the universe, so it was realized that the observed pattern of galaxy clustering today provides information about these unknowns.

Recent redshift surveys

Also in about 1980, advances in detector technology meant that a single galaxy redshift could be measured with a moderate-aperture optical telescope in tens of minutes instead of the few hours required with photographic plates, so it became feasible to map the three-dimensional galaxy distribution with surveys of large numbers of galaxy redshifts, notably the first Center for Astrophysics (CfA) survey of Davis and Huchra in 1982. For quantitative measurements, it is necessary for a survey to have well-defined selection criteria such as surveying all galaxies (or a random subsample) to a given apparent magnitude limit over a chosen region of sky. The observed density of galaxies in the survey then falls with redshift, since nearby all but the faintest galaxies are included, while at large distances only rare intrinsically bright galaxies are included; but this 'selection function' may be corrected for in the analysis.

Several important developments occurred from further surveys during the late 1980s, including the discovery that large VOIDS up to 50 Mpc in diameter with a galaxy density below 10% of the mean were quite common; and the discovery of substantial 'bulk flows' of galaxies towards a large supercluster in the constellations of Hydra–Centaurus, known as the Great Attractor. These surveys showed that superclustering is a ubiquitous phenomenon, and that the old picture of isolated clusters in a random 'sea' of field galaxies is not correct. Rather, the galaxy distribution has a sponge-like topology, with both high- and low-density regions forming an interconnected network, with high-density flattened 'walls' separating the voids; clusters are commonly located at the intersections of the walls.

The discovery of anisotropy in the COSMIC MICROWAVE BACKGROUND (CMB) by COBE in 1992 provided a major boost to the field, since the size of the COBE anisotropy was close to the predictions from theoretical models of structure formation 'matched' to the existing measurements of present-day galaxy clustering.

At the present time, the largest existing galaxy redshift surveys are as follows: the CfA-2 survey of Huchra and Geller, containing about 15 000 galaxies selected from photographic plates over most of the Northern hemisphere to a blue magnitude of about 15.5, giving a median depth around $90h^{-1}$ Mpc. This survey discovered a structure known as the 'Great Wall' spanning approximately $100h^{-1}$ Mpc across the northern sky, and containing the Coma cluster; an extension to the southern sky known as SSRS also showed wall-like structures.

The largest current 'all-sky' survey is the IRAS PSCz survey, which contains about 15 000 galaxies detected by the Infra-Red Astronomical Satellite (IRAS); this is a successor to earlier IRAS-based surveys such as the 'QDOT' and '1.2 Jansky' surveys. The PSCz survey consists of a sample of galaxies selected by far infrared brightness (above 0.6 J at 60 μm); these long wavelengths are almost immune to 'extinction' by dust in the plane of our Milky Way Galaxy, which limits the coverage of optical surveys. Also, the IRAS satellite scanned the whole sky, so the problem of matching ground-based northern and southern hemisphere catalogues does not arise. The downside is that IRAS preferentially selects star-forming late-type spiral galaxies; ellipticals contain little dust and so are intrinsically much fainter in the far infrared. The median depth of the PSCz survey is similar to CfA-2, but the average galaxy density is roughly $3\times$ lower so it provides a wider but sparser map of the galaxy distribution. The PSCz survey covers 84% of the whole sky, and so is the main source at present for comparison of the 'local' galaxy density with bulk flow measurements and our Local Group motion with respect to the CMB; the predicted direction of motion calculated from PSCz agrees within 15° of the measured Local Group motion.

The largest survey completed to date is the Las Campanas Redshift Survey: this contains 25 000 galaxies to a greater depth (a median redshift $z \approx 0.08$), covering three long thin 'slices' (about $90^\circ \times 1.5^\circ$) in each of the north and south Galactic hemispheres. This survey made use of the new technology of fiber-fed spectrograph (see SPECTROGRAPHS: FIBER-FED SPECTROGRAPHS), whereby spectra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of many galaxies (in this case 128) can be obtained simultaneously by placing optical fibers at the known locations of galaxies in the focal plane of the telescope. The other ends of the fibers are then arranged in a line along the entrance slit of the spectrograph, and a two-dimensional detector is used (e.g. a CHARGE-COUPLED DEVICE (CCD)) to record all the spectra simultaneously. A small number of fibers are placed on 'blank' regions of sky so that the night-sky spectrum can be subtracted from the measured (galaxy+sky) spectra.

Statistics of galaxy clustering

Among the simplest statistics describing galaxy clustering is the two-point correlation function $\xi(r)$; given two small volumes of space $\mathrm{d}V_1, \mathrm{d}V_2$ separated by distance r, the probability of finding galaxies in both is defined to be $\mathrm{d}P = [1 + \xi(r)]\bar{n}^2\,\mathrm{d}V_1\,\mathrm{d}V_2$, where $\bar{n}$ is the mean galaxy density. Thus $\xi(r)$ may be thought of as the excess probability (relative to random) of finding a second galaxy at distance r from a randomly-chosen first galaxy. Observationally, it is found that $\xi(r)$ is well approximated by a power law

$$\xi(r) \approx \left(\frac{r}{r_0}\right)^{-1.8}$$

where $r_0 \approx 5h^{-1}$ Mpc is the correlation length. The fractional variance $\sigma^2(r)$ of galaxy counts (above shot-noise) in a randomly placed volume V is given by

$$\sigma^2(r) = \frac{1}{V^2}\int\!\!\int_V \xi(r_{12})\,\mathrm{d}V_1\,\mathrm{d}V_2$$

this is found to be approximately unity for spheres of radius $8h^{-1}$ Mpc. The correlation function and the power spectrum $P(k)$ form a Fourier-transform pair, i.e.

$$\xi(r) = \frac{1}{(2\pi)^3}\int P(\boldsymbol{k})\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r})\,\mathrm{d}^3k$$

assuming statistical isotropy leads to

$$\xi(r) = \frac{1}{2\pi^2}\int_0^\infty k^2 P(k)\frac{\sin kr}{kr}\,\mathrm{d}k.$$

The importance of $P(k)$ is that if the initial density fluctuations formed a Gaussian random field (i.e. phases of different Fourier modes of the fluctuations were uncorrelated) as suggested by inflationary cosmologies, then $P(k)$, or equivalently $\xi(r)$, suffices to give a complete statistical description.

Many other statistics have been investigated (e.g. three-point and higher correlations, topology, probability distributions of counts in cells etc); the evidence at present is consistent with initially Gaussian fluctuations. While the present-day fluctuations are definitely non-Gaussian on small scales, the amount of this is approximately consistent with non-linear gravitational evolution from Gaussian initial conditions, as estimated from e.g. large N-body simulations.

In reality there are a number of complications to the above, notably biasing (i.e. the uncertain relationship between the distributions of galaxies and the underlying mass) and redshift-space distortions, whereby the galaxy peculiar velocities distort the mapping from real space to redshift space: clustering is smeared out by random motions on small scales, but on large scales it is enhanced because superclusters are expanding more slowly than the Hubble flow, so appear 'squashed' along the line-of-sight direction. The latter effect is marginally detected in current surveys, and can provide a measurement of the parameter $\beta \equiv \Omega^{0.6}/b$ where b is the bias parameter.

Surveys in progress

While the current surveys give us a good general picture of the galaxy distribution, they are still limited in statistical precision because they contain only a modest number (typically a few tens) of individual superclusters and voids. To obtain more precise statistical measurements, especially over the scales ∼50 Mpc which are currently being probed by degree-scale CMB experiments, larger surveys are very desirable. At the present time, there are two much larger surveys in progress: the 2 degree field (2dF) Galaxy Redshift Survey at the Anglo-Australian Telescope, and the Sloan Digital Sky Survey using a special-purpose 2.5 m telescope in New Mexico. The 2dF instrument consists of several components: a set of large corrector lenses giving a 2° diameter field, two sets of 400 optical fibers, and a robotic positioner. The positioner places the fibers with a positional accuracy of 30 μm using magnetic buttons on a metal plate, to correspond to the known positions of target galaxies, which are drawn from the APM Galaxy Survey based on scans of Schmidt plates. The positioner takes about 1 h to position a set of 400 fibers, so to avoid dead-time there are two sets of fibers, and the positioner is at work placing the 'next' set of fibers while the telescope is observing the 'current' set. Observations have been in progress since late 1997, and as of late 1999 redshifts have been measured for 77 000 galaxies and 7000 quasars (see figure 2); when complete at the end of 2001, the survey will contain redshifts of 250 000 galaxies and 20 000 quasars.

The Sloan Digital Sky Survey (SDSS), undertaken by a consortium of some 100 mainly US-based astronomers, aims to measure redshifts for 1 million galaxies and 100 000 quasars using a dedicated 2.5 m telescope in New Mexico. A novel feature of this project is that it will be making an imaging survey using CCDs in five optical passbands from ultraviolet (350 nm) to near-infrared 850 nm, using a very large camera containing 30 CCD detectors; this will be the first CCD survey to cover a substantial fraction (∼25%) of the entire sky. The target galaxies and quasars will be selected from this imaging survey and spectra taken using a fiber-optic spectrograph with 640 fibers over a 2.5° field of view. As of late 1999, the SDSS telescope and instruments are built and now starting a commissioning phase. It is planned to complete the survey in approximately 5 yr of observing.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

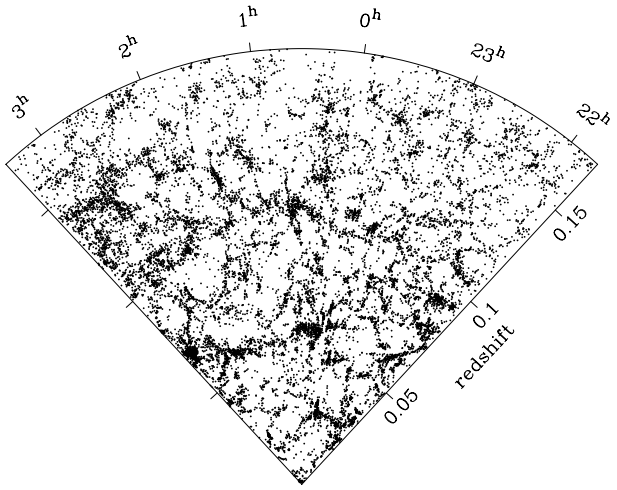


Figure 2. The distribution of 16 500 galaxies in a thin slice of the universe, comprising a mostly complete subregion of the 2dF Galaxy Redshift Survey. Our galaxy is located at the apex: the angular coordinate is right ascension, radial coordinate is redshift. This slice is 4° thick in the 'out-of-paper' (declination) direction. Numerous superclusters, sheets and voids are clearly visible.

For the relatively local universe, a new full-sky survey is planned based on the ongoing 2 Micron All Sky Survey (2MASS) infrared imaging survey. This will use a number of small telescopes in both hemispheres to obtain redshifts for the brightest 100 000 galaxies selected in the infrared K-band (2.2 μm). This K-band selection should combine the high sampling density of the optical surveys with the immunity to Galactic extinction of the IRAS surveys and will give a very detailed map of the local universe within 100 Mpc.

In addition, there are two surveys now planned which aim to study galaxy clustering when the universe was approximately half its present age: using multislit spectrographs on the new generation of 8 m telescopes, the VIRMOS survey on the ESO Very Large Telescope and the DEEP survey on the Keck telescope both aim to survey around 10 000 galaxies at a redshift of $z \sim 0.5$–1, to study the growth of structure over time.

Bibliography

Several reviews giving both a theoretical and observational perspective are contained in

Dekel A and Ostriker J (eds) 1999 *Formation of Structure in the Universe* (Cambridge: Cambridge University Press)

Another longer review is by

Strauss M and Willick J 1995 *Phys. Rep.* **261** 271

Detailed treatments of the theory of density fluctuations are given by

Peebles P J E 1980 *The Large-Scale Structure of the Universe* (Princeton, NJ: Princeton University Press)

and by

Peacock J A 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

Will Sutherland

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galaxy Structure

A very large body of information has been collected on the structure of the Galaxy. Here we can only concentrate on a small selection of the most recent and most important observational facts and theoretical findings. The MILKY WAY GALAXY belongs to the Local Group, a smaller group of three large and over 30 small galaxies, and is the second largest (after the Andromeda Galaxy M31) but perhaps the most massive member of this group. The Milky Way is most likely a BARRED SPIRAL GALAXY of Hubble type SBc. We are situated within the outer regions of its spiral disk, only about 14 ly above the equatorial symmetry plane but about 8 kpc or 26 400 ly from the Galactic Center (1 pc ~ 3.29 ly ~ $3 \times 9.5 \times 10^{12}$ km). The Milky Way forms a luminous band spanning all around the sky. This symmetry plane is also called the 'Galactic Equator'. In addition to the luminous stars there is also a dark band due to the obscuration of stellar light via gas and dust within the GALACTIC PLANE. The center of the Galaxy lies in the direction of the constellation Sagittarius, close to the border of both neighbor constellations Scorpius and Ophiuchus. The coordinates of the center (Rogers *et al* 1994) are $\alpha(2000) = 17^h\ 45^m\ 40.045^s \pm 0.010^s$ and $\delta(2000) = -29^o\ 00'\ 27.9'' \pm 0.2''$.

Main structural components

Since the Sun is located within the Milky Way disk it is exceedingly difficult to reliably determine its large-scale structure and dynamics, i.e. characteristics of the main components and the shape of the Galactic rotation curve.

The reason for that difficulty is mainly distance ambiguities and the location of the Sun within the Galactic obscuring dust layer. In summary the main structural components are the disk including the barred bulge with a central cusp, and most likely four spiral arms, a halo as well as a more extended corona including the Magellanic clouds and at least seven nearby dwarf galaxies and a large number of globular star clusters (see figure 1). The total mass of this central section of the Galaxy is (with separations from the center of $\leq$5 kpc) about 2×10^{11} solar masses ($M_\odot = 1.985 \times 10^{33}$ g). Here the molecular gas and dust as well as the youngest brightest stars of POPULATION I are located in a flat disk structure. The older population I stars form a less flattened distribution. This part of the Galaxy is surrounded by a probably ellipsoidally shaped halo of POPULATION II stars out to a radius of about 20 kpc. Population II stars are also found in the barred bulge of the Galaxy. The mass of this system is comparable with the mass in the inner section of the Milky Way. Within the halo the mass of the Galaxy amounts to about $4 \times 10^{11} M_\odot$ (Fich and Tremain 1991). The Galactic Halo may have been formed largely by accretion and tidal disruption of satellite systems. This is supported by numerous observational facts: the discovery of a retrograde rotating stellar group, patchiness in the kinematics of halo stars (Majewski *et al* 1996), and the discovery of an elongated dwarf galaxy in Sagittarius (Ibata *et al* 1994). In addition most of the Milky Way satellites appear to be located near two great circles in the sky (Lynden-Bell 1976). The corona extends out to approximately 100 kpc and may account for as much as $(3–6) \times 10^{12} M_\odot$ (Fich and Tremaine 1991). It exceeds by far the mass contained in the spiral disk and bulge region.

Main tracers of structure and dynamics

The entire solar system is orbiting the Galactic Center at a distance of ~8 kpc, on a nearly circular orbit. It is moving at about 250 km s^{-1} toward $\alpha = 21^h\ 12^m$, $\delta = +48^o$ 19'. Figure 2 shows a rotation curve derived from observations and consistent with model calculations. At this velocity it takes about 220 million years to complete one orbit. The Sun has orbited the Galactic Center about 20–21 times since its formation about 4.6 billion years ago.

There is still a controversy on the shape of the rotation curve of the Galaxy in the deep halo. The rotation velocity may ultimately decay as in accord with the Keplerian law at a distance of 50 kpc. However, measurements indicate that the rotation curve of the Milky Way essentially stays flat. It actually may even still increase at the distance of the MAGELLANIC CLOUDS and the MAGELLANIC STREAM. Murai (1997) has clarified, on the basis of recent observational data on the spatial location of the LMC, SMC and the Magellanic Stream and their radial velocity distribution, that the halo of the Galaxy has a huge dark matter component that results in a flat rotation curve with a terminal velocity of the order of 250 km s^{-1}. However, for the spiral arms found outside the corotation of the bar, the dark halo may very well have some effect. In model calculations that include a halo component (e.g. Englmaier and Gerhard 1999; see below) resulting in a flat rotation curve, the halo contribution to the radial force at the solar circle is 23%.

The main tracers of the Milky Way structure are the stars and the line emission of the molecular (mostly traced via the line emission of the ^{12}CO $J = 1–0$ rotational transition at 2.6 mm) and atomic (traced via the 21 cm hyperfine-structure transition of atomic hydrogen—H I) gas in the disk. Through data from the Hipparcos astrometry satellite great progress has been made in understanding the stellar distribution and kinematics (e.g. Creze *et al* 1998; Bienaymé 1999). The stellar structure of the Milky Way can be investigated much better in the infrared light than in the optical, since the extinction is lower than in the visible light. The COBE satellite has provided an infrared image of the Milky Way's central region. The near-infrared (NIR) is a particularly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

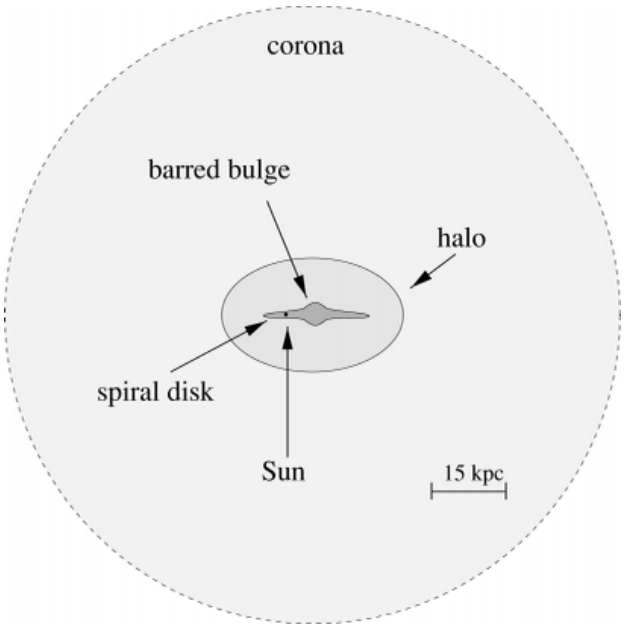


Figure 1. Main structural components of the Milky Way system. The Milky Way Galaxy in the center of the field is seen edge on from outside the corona that also contains the Magellanic Clouds, nearby dwarf galaxies and a large number of star clusters.

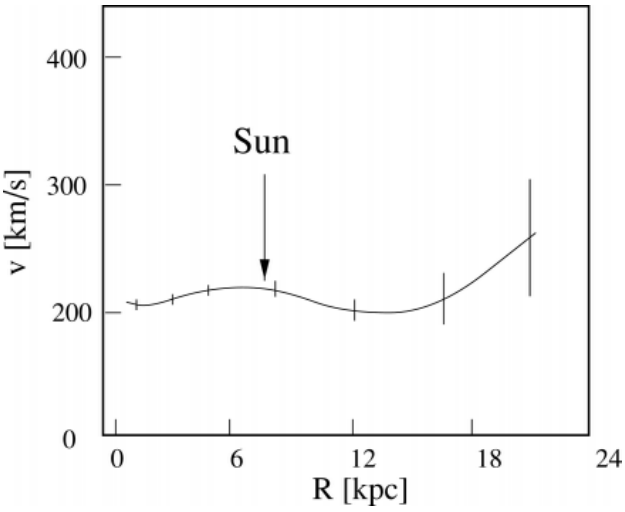


Figure 2. Rotation curve for the Galaxy (after Fich and Tremaine 1991). The vertical bars indicate the approximate uncertainty.

advantageous wavelength window for observing the distribution of old, evolved stars in the Galactic disk and bulge. These stars are important because they are tracers of the overall stellar mass distribution. At shorter wavelengths extinction is a serious problem, and at longer wavelengths the flux starts to be dominated by dust emission. As determined from NIR 2.4 μm mapping of the stellar continuum emission (Kent *et al* 1991) the radial surface brightness profile of the disk has an exponential scale length of about 3.0 kpc. In the solar neighborhood, the scale height is about 250 pc and with increasing radius it decreases to about 160 pc at a distance of 5 kpc from the center. This is in agreement with star counts perpendicular to the Galactic plane.

The bar

In the last few years, independent evidence points to the existence of a bar in the inner Galaxy. This evidence results from NIR photometry (Blitz and Spergel 1991; Weiland *et al* 1994; Dwek *et al* 1995), from IRAS source counts (Nakada *et al* 1991; Whitelock and Catchpole 1992; Nikolaev and Weinberg 1997; Stanek *et al* 1997), as well as from the large microlensing optical depth towards the bulge (Paczynski *et al* 1994; Zhao *et al* 1996) and possibly also from stellar kinematics (Zhao *et al* 1994). This evidence is summarized in recent reviews, e.g. Gerhard (1996) and Kuijken (1996).

The best current models for the distribution of old stars in the inner Galaxy are based on the NIR data from the DIRBE experiment on COBE. The gravitational potential of the stars can then be connected with the available gas and stellar kinematic observations in order to study the large-scale structure and dynamics of the Milky Way Galaxy.

Since even at a wavelength of 2 μm the extinction towards the Galactic nuclear bulge is important, the DIRBE data has to be corrected for the effects of extinction. Spergel *et al* (1996) use a three-dimensional model of the dust distribution to clean the data. These data have then been used by Binney *et al* (1997) to fit a non-parametric model of the NIR emission under the assumption of eight-fold (triaxial) symmetry with respect to three orthogonal planes. The best fits to the photometry are obtained when the Sun is assumed to lie ~14±4 pc above the plane. For a fixed geometry the recovered emissivity is essentially unique (Binney and Gerhard 1996; Bissantz *et al* 1997). However, matching physical models are possible over the full range of likely bar orientation angles ϕ. This angle is defined as the angle in the Galactic plane between the major axis of the bar at $l > 0^\circ$ and the Sun–center line. It covers a range from 15º to 35º. Using the most favored angle of $\phi = 20^\circ$, the deprojected luminosity distribution shows an elongated bulge with axis ratios 10:6:4 and semimajor axis of ~2 kpc, surrounded by an elliptical disk extending up to ~3.5 kpc on the major axis and ~2 kpc on the minor axis. Figure 3 shows a comparison between the 3.4 μm flux density distribution as measured by the COBE satellite (full contours) and a model (dashed contours; Binney *et al* 1997) that assumes a bar angle to the line of sight of $\phi = 20^\circ$. The asymmetry of the highest contours is a clear indication for the presence of a bar as seen from the position of the Sun within the Galactic plane. The enhanced flux at positive Galactic longitudes indicates the location of the tip of the bar that is closest to us.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

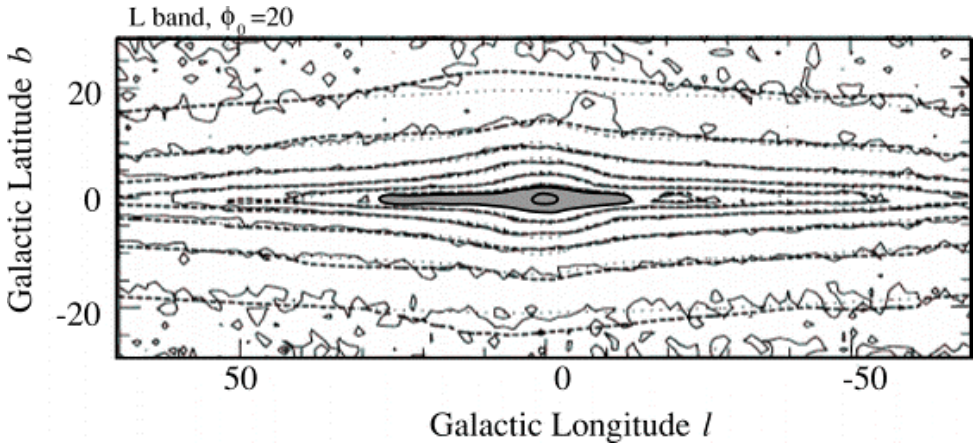


Figure 3. A comparison between the dust-corrected 3.4 μm flux density distribution as obtained by the COBE satellite (full contours, see also Arendt *et al* 1994) and a model (dashed contours; Binney *et al* 1997). In order to demonstrate that the Galaxy is not axisymmetric we filled in the top two contour lines.

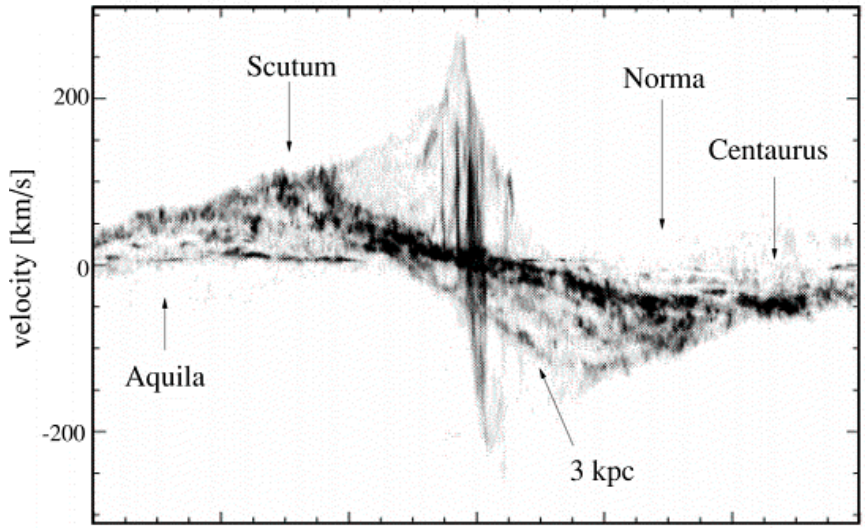


Figure 4. Galactic longitude velocity diagram of the ^{12}CO line emission as a function of Galactic longitude and velocity integrated over latitudes between $b = \pm 2^{\circ}$ (Dame *et al* 2000). The directions of some arm tangents have been indicated. Comparison with figure 5 shows where the spiral arms are located within the Galactic disk.

In all models there is a local density minimum ~2.2 kpc down the minor axis. The subsequent maximum ~3 kpc down the minor axis may be associated with the Lagrange point L_4. The identification of this point as well as the length of the bar are consistent with a bar pattern speed of $\Omega_b \sim 60$–70 km s^{-1} kpc^{-1} (Binney *et al* 1997). Model calculations assuming an eight-fold or four-fold symmetry for the disk spiral structure indicate that the deprojected intensity distributions recovered from the cleaned COBE data could reflect spiral structure within the Milky Way with density contrasts greater than ≥3 in the NIR.

The spiral structure

The SPIRAL ARMS can be traced by the distribution of atomic and molecular gas as well as the distribution of young bright stars and the emission line nebulae. The corresponding directions and distances with respect to the Sun can then be used to map out the local spiral structure. Figure 4 shows the distribution of the ^{12}CO line emission as a function of Galactic longitude and velocity integrated over latitudes between $b = \pm 2^{\circ}$.

The spiral arms manifest themselves via an increase in projected density of brightness of the corresponding tracers especially when the line of sight is tangential to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the spiral arm. This technique was first applied to H I data (e.g. Kerr and Westerhout 1964) and has thereby led to a first picture of the distribution of atomic gas and a first indication of a spiral structure in the Galaxy. More recent modern H I surveys of the Galactic disk (e.g. Stark *et al* 1992; Burton 1985) primarily aim at high sensitivity, low contamination due to side lobes of the antenna beam and a large velocity coverage in order to trace the high-velocity H I gas located towards the center of the Galaxy and in individual clouds in and above the plane. Spiral arms are named after the constellations in which they are most prominent. The results indicate that the Sun lies near the inner edge of the so-called ORION ARM (see figure 5). The Perseus arm is located about 2 kpc beyond the Sun and the Aquila and Carina arm sections are approximately located at a distance from the Sun of 2 kpc towards the Galactic Center (see figure 5). From the study of H II regions, molecular clouds and the Galactic magnetic field it appears that the Milky Way may have four main spiral arms (Georgelin and Georgelin 1976; Caswell and Heynes 1987; Sanders *et al* 1985; Grabelsky *et al* 1988; Vallée 1995).

Dame *et al* (1987, 1999) have combined large-scale surveys of the ^{12}CO(1–0) integrated line intensity of the entire Galactic plane and specific nearby clouds to produce a panorama of the entire Milky Way in molecular gas at an angular resolution of ½º. Their compilation exhibits a sky coverage and resolution which are comparable with those of early 21 cm surveys. A comparison with other population I tracers demonstrates that the currently available CO surveys probably provide an almost complete inventory of nearby molecular clouds. The ^{12}CO(1–0) line, however, traces in most cases the $\tau = 1.0$ surfaces of optically thick molecular gas. A much better census of the amount of molecular gas can be obtained via the optically thinner ^{13}CO(1–0) line. Stark (1987) presents a Galactic survey of 47 000 positions observed in the ^{13}CO(1–0) line emission. The data show a clear presence of voids in the molecular gas disk of the Milky Way. These regions contain many times fewer giant molecular clouds (GMCs) than do adjacent regions of similar size. This is interpreted as a signature of a spiral structure. The difference between the arm and interarm regions in the inner Galaxy is only clearly visible in the distribution of GMCs. Small molecular gas clouds are present throughout the disk. It is possible that GMCs are formed as interstellar gas enters a spiral arm and that they break up into small molecular or atomic clouds when the gas leaves the arm. The molecular gas disk is flatter than the atomic gas disk. The Gaussian scale height (i.e. thickness perpendicular to the plane) at the solar circle within 1 kpc is ~74 pc (Dame *et al* 1987). The H I atomic gas scale height is ~100 pc at 3 and ~220 pc at 9 kpc distance from the center (Malhotra 1995).

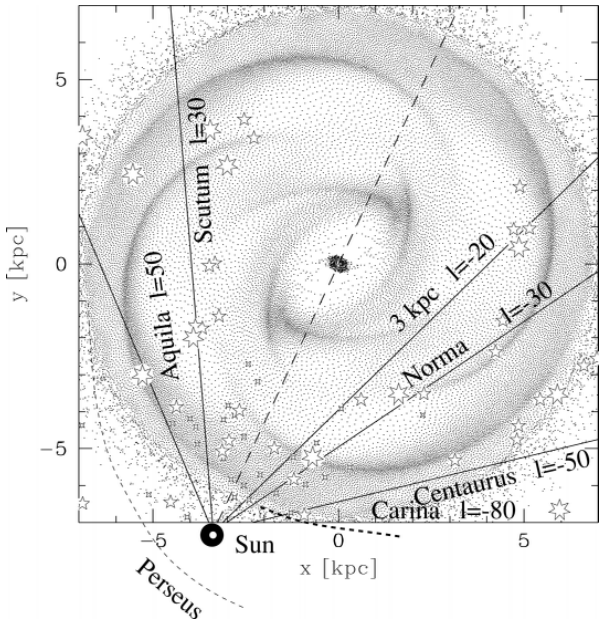


Figure 5. The gas flow in the Galactic disk under the influence of the barred bulge with a bar angle $\phi = 20º$, corotation radius at ~3.1 kpc as well as a dark halo component (Englmaier and Gerhard 1999). The location of the Sun is shown at $x = -2.7$ kpc and $y = -7.5$ kpc just outside the frame. The directions of some arm tangents and the corresponding approximate galactic longitudes have been indicated. The Orion arm is possibly a connection between the Aquila and Perseus arm in the vicinity of the Sun. The image is a result of hydrodynamical two-dimensional modeling using 20 000 particles in a smoothed particles hydrodynamics (SPH) code described in Englmaier and Gerhard (1997). The barred bulge has a size and orientation (4 × 2.4 kpc) which is almost identical to the central region void of particles.

The currently best model of Englmaier and Gerhard (1997) provides a coherent interpretation of most of the observed gas dynamical features. The four-armed spiral structure outside corotation reproduces quantitatively the five main spiral arm tangents at $l \leq 60º$ that are observed in several tracers. In this model the observed 3 kpc arm is part of a spiral arm structure that starts at the ends of the bar and extends into the corotation region. Orbits along the bar are called x1 orbits. They allow transitions to the inner x2 orbits perpendicular to the bar (Contopoulos and Papayannopoulos 1980). This model exhibits an inner gas disk with a cusped orbit shock transition to an x2 orbit disk of radius $R \sim 150$ pc (see inner portion of figure 5). The corotation radius of the bar is fairly well constrained at $R \sim 3.5 \pm 0.5$ kpc. The Galactic terminal velocity curve from H I and CO observations out to $l \pm 45º$ (~5 kpc) is approximately described by a maximal disk model with constant mass-to-light ratio for the NIR bulge and disk.

The center of the Galaxy

At a distance of only 8 kpc the Galactic Center is the closest nucleus of a galaxy, 100–1000 times closer than

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the nearest extragalactic systems. It is thus a unique laboratory in which physical processes that are also relevant for nuclei of other galaxies can be studied with the highest angular resolution possible. The gas and dust on the line of sight to the center, however, make it impossible to observe it at optical or UV wavelengths. It is only accessible in the radio, infrared and x-ray domain. The optical extinction amounts to 30 magnitudes (A_V = 30^m) or an attenuation of visible light by a factor of 10^{-12}. This corresponds in the NIR *K* band at a wavelength of 2.2 μm to an extinction of only $A_K = 3.3^m$.

The central stellar cluster

Catchpole *et al* (1990) showed from NIR star counts that the density distribution of stars near the Galactic Center can be modeled as a power law r^{-p} with $p \sim 2.2 \pm 0.2$ for $K = 6^m$–8^m stars. This is similar to the value of $p \sim 2.0 \pm 0.2$ derived from the distribution of OH/IR stars near the center (Lindqvist *et al* 1992). Using radial velocities of the OH–IR stars and the assumption of isotropy, Lindqvist *et al* determined the mass distribution inside ~100 pc. The corresponding mass density profile has $p \sim$ 1.5 between ~20 and ~100 pc and steepens inside ~20 pc. The overall slope is in approximate agreement with the value of $p \sim 1.8$ originally found by Becklin and Neugebauer (1968). Because of the limited resolution the central cusp cannot be recovered from the DIRBE NIR data and has to be re-introduced to model the dynamics. The total mass, however, is small compared with the mass of the barred bulge. Inside the central parsec of the Galaxy the stellar number density counts show that all the ~ 600 stars in the central stellar cluster brighter than 16^m at 2.2 μm can be described very well by the number density distribution expected for an isothermal stellar cluster with a core radius of approximately 0.2–0.3 pc (Eckart *et al* 1993, Eckart and Genzel 1996).

Phenomena within the central few parsecs

The central few parsecs of the Milky Way is a region full of interesting phenomena (figure 6). A circumnuclear disk or ring consisting of a system of orbiting molecular filaments at a distance of 1.5–5 pc is connected to a dense central stellar cluster via several ridges, filaments and streamers of gas and dust that form the so-called 'mini spiral' close to the central compact (see above) stellar cluster that connects on large-scale NIR images visually to the stellar bulge and disk population at approximately the radius of the circumnuclear ring.

The circumnuclear disk is probably fed by gas infall from dense molecular clouds at distances larger than 10 pc from the center. The mini spiral represents the brightest part of the thermal radio source Sgr A (West). The central 1.5 pc is comparatively devoid of interstellar matter and is referred to as the central cavity. Right at the center of the stellar cluster there is located a compact radio source Sagittarius A* (Sgr A*) with a radio

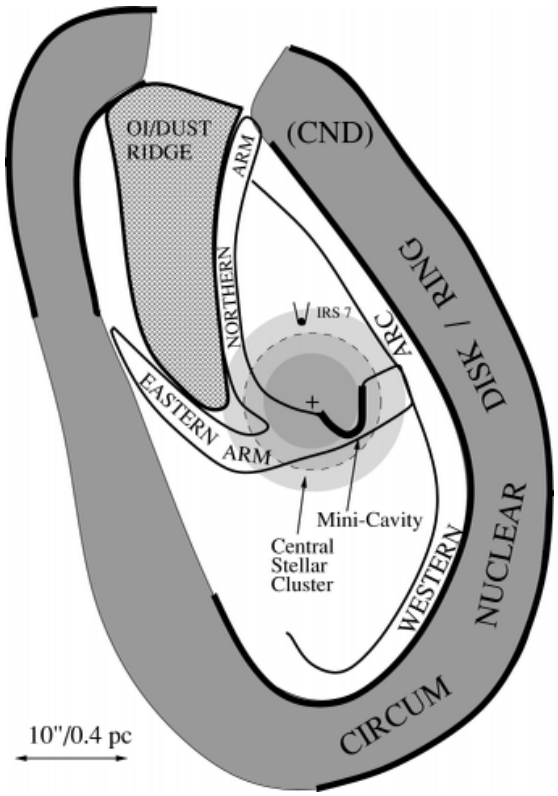


Figure 6. Schematic picture of the different structures in the Galactic Center. The 'mini spiral' arms—the eastern and northern arm as well as the 'bar'—connect the central stellar cluster region with the circumnuclear gas ring (after Genzel *et al* 1994). The brightest star at 2 μm wavelength, IRS 7 (infrared source), has been labeled. The position of the compact radio source Sgr A* is also indicated as a cross.

luminosity of ~2×10^{34} erg s^{-1}, corresponding to a flux density of 1.1 Jy at λ = 2 cm, and a size of less than 2.4×10^{13} cm (Rogers *et al* 1994; Krichbaum *et al* 1994). This source is the most likely candidate for a possible massive black hole at the center of our Galaxy (e.g. Rees 1984, Genzel and Townes 1987; Lo 1989). Along the arms of the mini spiral gas and dust are orbiting about the center where they interact with the intense nuclear wind originating from the central cluster of very massive and luminous stars. Recent subarcsecond angular resolution line imaging (Eckart *et al* 1995; Tamblyn *et al* 1996) and integral field spectroscopy (Krabbe *et al* 1995; Genzel *et al* 1996) have shown that several of the brightest members of the central cluster—the so-called IRS16 complex—are He I emission line stars. These objects show in their spectra prominent broad He I and H I recombination lines in emission. From non-LTE stellar atmosphere modeling of the observed emission characteristics of several of the He I stars Najarro *et al* (1994, 1997) have inferred that these objects are moderately hot (17 000–30 000 K), very luminous ((1–30) × $10^5 L_\odot$) massive stars whose helium-rich surface

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

layers are expanding as powerful stellar winds with velocities of 200–800 km s^{-1} and mass loss rates of (1–70) × $10^{-5} M_{\odot}$ yr^{-1}. These stars have formed within the last few million years and provide the dominant fraction of the total luminosity of the central parsec. Imaging spectroscopy with the integral field imaging spectrometer shows that within the central 0.3–0.4 pc of the stellar cluster bright late-type stars (supergiants and the brightest AGB stars) are absent but that the core is surrounded by a ring of red supergiants–AGB stars showing strong CO band head absorptions in their NIR spectra (Genzel *et al* 1996; see also earlier work by Sellgren *et al* 1990; Haller *et al* 1996).

The central black hole

The gas and stellar velocities increase toward the position of Sgr A* indicating the presence of a large compact mass. The evidence for a dark central mass of (1–3) × $10^{6} M_{\odot}$ in the Galactic Center has been steadily growing over the last two decades from observations of RADIAL VELOCITIES of gas and stars (Lacy *et al* 1980; Serabyn and Lacy 1985; Genzel *et al* 1985; Sellgren *et al* 1990; Krabbe *et al* 1995; Haller *et al* 1996; Genzel *et al* 1996). A reliable estimate of the enclosed mass and its compactness, however, can only be made if the full velocity field (radial and transverse components) is known. Furthermore, stars as tracers of the velocity field are much more reliable than gas, since they are not influenced by magnetic pressure or frictional forces. Using stars as tracers of the gravitational potential therefore allows us to measure the mass content and mass concentration within the central parsec of our Galaxy. Recent review articles on the phenomena and physical properties of the Galactic Center are Genzel *et al* (1994), Blitz *et al* (1993), Mezger *et al* (1996) and Morris and Serabyn (1996).

The stellar PROPER MOTION (Eckart and Genzel 1996; Ghez *et al* 1998) and radial velocity dispersions (Genzel *et al* 1996) are in very good agreement, indicating that the stellar velocity field is on average close to isotropic. Comparing individual images from different observing epochs one finds within the central arcsecond (0.04 pc) several fast-moving stars with velocities in excess of 1000 km s^{-1} in the immediate vicinity (0.01 pc) of Sgr A*. From the stellar radial and proper motion data, one can infer that a dark mass of (2.61±0.35) × $10^{6} M_{\odot}$ must reside within about a light week of the compact radio source. Its density must be 2.2 × $10^{12} M_{\odot}$ pc^{-3} or greater. There is no stable configuration of normal stars, stellar remnants (neutron stars or stellar black holes) or substellar entities that can account for the combination of both the high mass and the high mass density. The combination of these data—especially the proper motion measurements of the stars in the central arcsecond—now provides compelling evidence for a massive BLACK HOLE at the core of the Milky Way.

In 2001 the Chandra X-ray telescope observed a flare of X-ray energy produced from where the lip of a supermassive black hole at the center of the Milky Way Galaxy should be. The flare dimmed and brightened over a period of ten minutes, the time taken for light to travel about 150 million kilometers, meaning that it originated from an object no larger in diameter than the distance of the Earth to the Sun. This implies that the mass at the center of our Galaxy (about 2.6 million times the mass of our Sun) is crammed into that space. According to the laws of general relativity, only a black hole can be that dense.

Bibliography

Arendt *et al* 1994 *Astrophys. J.* **425** L85

Bally J, Stark A A, Wilson R W and Henkel C 1987 *Astrophys. J. Suppl.* **65** 13

Becklin E E and Neugebauer G 1968 *Astrophys. J.* **151** 145

Bienaymé O 1999 *Astron. Astrophys.* **341** 86

Binney J J and Gerhard O E 1996 *Mon. Not. R. Astron. Soc.* **279** 1005

Binney J J, Gerhard O E and Spergel D 1997 *Mon. Not. R. Astron. Soc.* **288** 365

Binney J J, Gerhard O E, Stark A A, Bally J and Uchida K I 1991 *Mon. Not. R. Astron. Soc.* **252** 210

Bissantz N, Englmaier P, Binney J J and Gerhard O E 1997 *Mon. Not. R. Astron. Soc.* **289** 651

Blitz L, Binney J, Lo K Y, Bally J and Ho P T P 1993 *Nature* **361** 417

Blitz L and Spergel D N 1991 *Astrophys. J.* **379** 631

Burton W B 1985 *Astron. Astrophys. Suppl.* **62** 365

Burton W B and Liszt H S 1978 *Astrophys. J.* **225** 815

Caswell J L and Haynes R F 1987 *Astron. Astrophys.* **171** 261

Catchpole R M, Whitelock P A and Glass I S 1990 *Mon. Not. R. Astron. Soc.* **247** 479

Cohen R J and Few R W 1976 *Mon. Not. R. Astron. Soc.* **176** 495

Contopoulos G and Papayannopoulos T 1980 *Astron. Astrophys.* **92** 33

Creze M, Chereul E, Bienaymé O and Pichon C 1998 *Astron. Astrophys.* **329** 920

Dame T M *et al* 1987 *Astrophys. J.* **322** 706

Dame T M *et al* 2000 in preparation

Dwek E *et al* 1995 *Astrophys. J.* **445** 716

Eckart A and Genzel R 1996 *Nature* **383** 415

Eckart A, Genzel R, Hofmann R, Sams B J and Tacconi-Garman L E 1993 *Astrophys. J.* **407** L77

Eckart A, Genzel R, Hofmann R, Sams B J and Tacconi-Garman L E 1995 *Astrophys. J.* **445** L26

Englmaier P and Gerhard O E 1997 *Mon. Not. R. Astron. Soc.* **287** 57

Englmaier P and Gerhard O E 1999 *Mon. Not. R. Astron. Soc.* **304** 512

Fich M and Trenaine S 1991 *Annu. Rev. Astron. Astrophys.* **29** 409

Genzel R, Crawford M K, Townes C H and Watson D M 1985 *Astrophys. J.* **297** 766

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Genzel R, Hollenbach D J and Townes C H 1994 *Rep. Prog. Phys.* **57** 417
Genzel R, Thatte N, Krabbe A, Kroker H and Tacconi-Garman L E 1996 *Astrophys. J.* **472** 153
Genzel R and Townes C H 1987 *Annu. Rev. Astron. Astrophys.* **25** 377
Georgelin Y M and Georgelin Y P 1976 *Astron. Astrophys.* **49** 57
Gerhard O E 1996 *Proc. IAU Symp. 169* ed L Blitz and P Teuben
Gerhard O E and Vietri M 1986 *Mon. Not. R. Astron. Soc.* **223** 377
Ghez A M, Klein B L, Morris M and Becklin E E 1998 *Astrophys. J.* **509** 678
Grabelsky D A, Cohen R S, Bronfman L and Thaddeus P 1988 *Astrophys. J.* **331** 181
Haller J W, Rieke M J, Rieke G H, Tamblyn P, Close L and Melia F 1996 *Astrophys. J.* **456** 194
Ibata *et al* 1994 *Nature* **370** 194
Kent S M, Dame T M and Fazio G 1991 *Astrophys. J.* **378** 131
Kerr F J and Westerhout G 1964 Distribution of hydrogen in the Galaxy *Galactic Structure* vol 5, ed A Blaauw and M Schmidt (Chicago, IL: University of Chicago Press)
Krabbe A *et al* 1995 *Astrophys. J.* **447** L95
Krichbaum T P, Schalinski C J, Witzel A, Standke K J, Graham D A and Zensus J A 1994 *The Nuclei of Normal Galaxies* ed R Genzel and A I Harris (Dordrecht: Kluwer) p 411
Kuijken K 1996 *Barred Galaxies: Proc. IAU Colloq. 157* ed R Buta, D A Crocker and B G Elmegreen (San Francisco, CA: Astronomical Society of the Pacific) p 504
Lacy J H, Townes C H, Geballe T R and Hollenbach D J 1980 *Astrophys. J.* **241** 132
Lindqvist M, Habing H J and Winnberg A 1992 *Astron. Astrophys.* 259 118
Liszt H S and Burton W B 1980 *Astrophys. J.* **236** 779
Lo K Y 1989 *Proc. 136 IAU Symp.* ed M Morris, p 527
Lynden-Bell 1976 *Mon. Not. R. Astron. Soc.* **174** 695
Malhotra S 1995 *Astrophys. J.* **448** 138
Majewski S R, Munn J A and Hawley S L 1996 *Astrophys. J.* **459** L73
Mezger P G, Duschl W J and Zylka R 1996 *Astron. Astrophys. Rev.* **7** 289
Morris M and Serabyn E 1996 *Annu. Rev. Astron. Astrophys.* **34** 645
Murai T 1997 *Proc. IAU Symp.* **186** 14
Mulder W A and Liem B T 1986 *Astron. Astrophys.* **157** 148
Najarro F, Hillier D J, Kudritzki R P, Krabbe A, Genzel R, Lutz D, Drapatz S and Geballe T R 1994 *Astron. Astrophys.* **285** 573
Najarro F, Krabbe A, Genzel R, Lutz D, Kudritzki R P and Hillier D J 1997 *Astron. Astrophys.* **325** 700
Nakada Y, Onaka T, Yamamura I, Deguchi S, Hashimoto O, Izumiura H and Sekiguchi K 1991 *Nature* **353** 140
Nikolaev S and Weinberg M D 1997 *Astrophys. J.* **487** 885
Paczynski B, Stanek K Z, Udalski A, Szymanski M, Kaluzny J, Kubiak M, Mateo M and Krzeminski W 1994 *Astrophys. J.* **435** L113
Peters W L 1975 *Astrophys. J.* **195** 617
Rees M 1984 Annu. Rev. *Astron. Astrophys.* **22** 471
Rogers A E E *et al* 1994 *Astrophys. J.* **434** L59
Sanders D B, Scoville N Z and Solomon P M 1985 *Astrophys. J.* **289** 373
Sellgren K, McGinn M T, Becklin E E and Hall D N B 1990 *Astrophys. J.* **359** 112
Serabyn E and Lacy J H 1985 *Astrophys. J.* **293** 445
Spergel D N, Malhotra S and Blitz L 1996 *Spiral Galaxies in the Near-IR* ed D Minniti and H-W Rix (Berlin: Springer) p 128
Stanek K Z, Udalski A, Szymanski M, Kaluzny J, Kubiak M, Mateo M and Krzeminski W 1997 *Astrophys. J.* **477** 163
Stark A A 1987 *Proc. IAU Symp.* **115** 495
Stark A A, Gammie C F, Wilson R W, Bally J, Linke R A, Heiles C and Hurwitz M 1992 *Astrophys. J. Suppl.* **79** 77
Tamblyn P, Rieke G H, Hanson M M, Close L M, McCarthy D W and Rieke M J 1996 *Astrophys. J.* **456** 206
Vallée J P 1995 *Astrophys. J.* **454** 119
Wada K, Taniguchi Y, Habe A and Hasegawa T 1994 *Astrophys. J.* **437** L123
Weiland J L *et al* 1994 *Astrophys. J.* **425** 81
Whitelock P A and Catchpole R W 1992 *L A U Symp.* **149** 503W
Zhao H-S, Rich R M and Spergel D N 1996 *Mon. Not. R. Astron. Soc.* **282** 175
Zhao H-S, Spergel D N and Rich R M 1994 *Astron. J.* **108** 2154

Andreas Eckart

Peter Englmaier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gale, Walter Frederick (1865–1945)

Australian banker and financier, amateur astronomer, born in Paddington, NSW, discovered a number of comets and double stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galilean Telescope

A refracting telescope that uses a long focal length converging lens and a short-focus diverging lens to produce a magnified image of a distant object. Named after the Italian scientist Galileo Galilei (1564–1642), who designed and built a telescope of this type in 1609, the Galilean refractor produces an upright image but suffers from having a very small field of view. This particular optical system is still used in opera glasses because the design is compact.

See also: eyepiece, objective lens, refractor, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galileo Galilei

Galileo Galilei (1564–1642) is one of the most significant figures in the long history of astronomy. Within a short two years (1609–11), the telescopes he fashioned revealed details of the Sun, Moon and planets that announced the end of ancient astronomy and the direction of the new. The success of the telescope heralded a new era for natural philosophy generally, when it would no longer be constrained by the limitations of the human senses.

Life and works

Born near Pisa, Galileo enrolled in medicine at the University of Pisa in 1581. His interests soon turned to mathematics. When he left the university without a degree in 1585, he began an intense study of mathematics and natural philosophy to prepare himself for a university position. Appointed to the chair of mathematics in Pisa in 1589, he composed a treatise (left unpublished), *De motu*, in which he questioned, on largely conceptual grounds, some of the most distinctive features of the Aristotelian account of motion. In 1592, he was appointed to the chair of mathematics at the University of Padua. There he lectured on, among other things, the application of spherical geometry to planetary astronomy, and busied himself with the design of instruments for practical use in surveying and gunnery, as well as for facilitating rapid numerical calculation.

In a letter of thanks to KEPLER (1597) on receipt of the latter's newly published *Mysterium Cosmographicum*, he associates himself with the Copernican views defended in this work, claiming that these views had already enabled him to explain 'many physical effects' (probably alluding to the tides) but that he had hesitated to publish for fear of public ridicule. Historians are still divided as to whether he really *was* committed to the COPERNICAN SYSTEM at so early a date. In 1604, a new star was observed in Sagittarius, challenging the Aristotelian doctrine of the unchangeability of the heavenly bodies. Already critical of Aristotelian views on motion, Galileo offered three public lectures on the NOVA, his first venture in astronomy. He seems to have concentrated on explaining how the star's lack of parallax implies its great distance from Earth and hence the changeability of the non-terrestrial region. His energies at this time were, however, mainly devoted to mechanics.

In 1609, a chance event diverted him from his researches in mechanics for what would become a 25 year interval. When he heard about a spyglass invented in Holland that made distant objects seem near, he immediately grasped its potentialities for the furtherance of his own career. How much about its mode of construction he learnt from others and how much he worked out for himself is not quite clear. What *is* clear, however, is that his tested skill in instrument-making gave him an advantage over his competitors and a lead in telescope-making that he maintained over the crucial two-year period (1609–11) within which all the major astronomical discoveries within the reach of a telescope of this type were made by him (see GALILEAN TELESCOPE). By August 1609 he could present the Doge of Venice with a roughly 9x telescope. Impressed by the military and commercial potential of the '*perspicillum*', a grateful Venetian Senate promised to double his salary.

However, Galileo saw even greater potential for himself elsewhere. By January 1610 he had constructed a 20x instrument and had discovered Earth-like irregularities on the Moon, four satellites orbiting Jupiter, and stars beyond numbering that were invisible to the naked eye. Later in 1610 came moving sunspots, the phases of Venus and the oddly changing shape of Saturn. Galileo rushed his first discoveries into print, dedicating the *Sidereus Nuncius* (*Sidereal Messenger*) to his former pupil, now Cosimo II, Grand Duke of Tuscany, and dubbing Jupiter's moons the 'Medicean' planets. These shrewd gestures gained him the position he had long coveted as Mathematician and Philosopher to the Grand Duke. His little book caused an instant sensation and made him Europe's most famous astronomer.

Other works quickly followed. His *Discourse on Bodies in Water* (1612) listed his newest astronomical discoveries, notably the periods of Jupiter's satellites, his most considerable contribution to calculational astronomy. His *Letters on Sunspots* (1613) mocked the views of the Jesuit astronomer, Christopher Scheiner, who represented sunspots as clouds circling the (still unchanging) Sun. But trouble was brewing in another quarter. The new discoveries might refute the Aristotelian claim that the Earth is the immobile center of rotation of the planets. But what of the passages in Scripture that describe the Sun as in motion or the Earth as immovable? Aristotelians could and did call theological witness in their support. Galileo immediately grasped the challenge that this could offer to the Copernican system, of whose truth he was now convinced. In a letter to Benedetto Castelli (1613) and a much longer treatise, the *Letter to the Grand Duchess Christina* (1615), formally addressed to his patron's mother, he laid out in persuasive fashion the hermeneutic principles that ought to guide the interpretation of Scripture when a literal reading opposed a finding of 'sense or necessary demonstration'. But it was to no avail. In February 1616, the Roman Congregation of the Index banned Copernicus' *De Revolutionibus* 'until it be corrected', declaring the Copernican theses regarding the Earth's motion and the Sun's rest to be 'contrary to Scripture'. Cardinal Robert Bellarmine, the leading theologian of the Congregation of the Holy Office, was instructed to inform Galileo officially that the Copernican doctrine was not to be held or defended; if he proved resistant, he was to be given a personal injunction not even to discuss the forbidden view. Whether this injunction was delivered, and if so under what circumstances, has been a matter of debate since the Holy Office file on Galileo was made public more than a century ago.

What motivated the Cardinal-members of the Holy Office was not the defence of an ancient and theologically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

comfortable Earth-centered cosmology. In their eyes, what was at stake was the authority of Scripture. This was, after all, the time when tensions between Reformers and the theologians of the Counter-Reformation were at their most intense. And the issue that most clearly divided the two sides was the proper interpretation of the Bible. Here was Galileo proposing his own non-literal interpretation of literal-sounding Scriptural passages, and unable to offer a demonstration for the conflicting Copernican claims, the only condition under which the Roman theologians would have been prepared to concede the propriety of a non-literal approach to the disputed passages.

Their error, as we would see it today, was not primarily scientific. It was in the first instance theological, in that their principles of Scriptural interpretation were at fault; they would have done well to heed Galileo's reminder of an older and more sophisticated set of principles going back to St Augustine. Their error was also philosophical in that they, and indeed Galileo himself, assumed that a claim about the natural world had to be demonstrated in order to count as properly scientific. Both errors are understandable in the theological and philosophical contexts of the day. But they were to have far-reaching consequences. What would be needed was a return to the more liberal canons of Scriptural interpretation favored by Augustine and a better understanding of the limits of theory-proof in natural science. Both would be slow in coming.

Forced to abandon his plans for a Copernican treatise, Galileo turned in 1618 to the problem of the nature of comets. Three bright comets appeared in that year, setting off a flurry of speculation. TYCHO BRAHE had carefully tracked the comet of 1577 and had concluded that it lay far beyond the Moon, that it was orbiting the Sun not the Earth, and that its motion was either non-uniform and circular or else perhaps oval. Galileo published the *Discourse on the Comets* (1619) under the name of his friend, Mario Guiducci, claiming that Tycho's parallax arguments would work only if the comet were a physical body, not if (as Galileo himself maintained) they were no more than reflections of sunlight from vapors of earthly origin lying beyond the moon. He criticizes Tycho for suggesting non-circular orbits and non-uniform motions, but has to admit that his own theory does not square with the mass of observational reports. Another target of the *Discourse* is Orazio Grassi, a Jesuit natural philosopher, who responded vigorously to the attack and to whom Galileo replied in devastating style with *The Assayer* (1623). The book was quickly recognized as a literary masterpiece. It is also significant for the exposition it gives of Galileo's own natural philosophy: the Book of Nature is written in the language of mathematics, the qualitative properties of material objects are merely subjective products of human sensation, and matter is composed of minute corpuscles in motion. His treatment of comets, however, the topic that had occasioned the book in the first place, is sprinkled with inconsistencies, his least successful venture in astronomy.

In 1623, Cardinal Maffeo Barberini, an avowed admirer of Galileo's work, was elected Pope, as Urban VIII. Galileo, encouraged, asked permission to resume work on his Copernican treatise. Urban allowed him to proceed, with the proviso that the treatment be hypothetical. The two men appear to have understood this restriction in very different ways. Galileo evidently took it to mean that he could make the best case possible for the Copernican hypothesis. Urban wanted to ensure that an explanation of observed effects in terms of causes not themselves observed would not be presented as demonstrative; to do so would, in his view, challenge Divine omnipotence. (This type of objection is traceable to the nominalist critics of Aristotle in the fourteenth century.) God could, after all, have brought about the same effects by means of other hidden causes. Urban may also have had in mind the status traditionally assigned to 'mathematical' astronomy, its goal merely to 'save the phenomena' (describe the observed planetary positions correctly), whereas for true causes one has to turn to 'physical' astronomy.

Galileo's Copernican treatise gradually took shape, delayed by his recurrent ill-health as well as by the complexity of the rhetorical task he faced. Urban's directive meant that he could not claim to demonstrate the Earth's motion, yet he could hardly forget that, according to the theologian-critics of Copernicanism in 1616, the literal interpretation of the debated passages in Scripture had to take priority if the 'scientific' claim fell short of demonstration. There were lengthy negotiations with the censors in Rome and Florence, but eventually the *Dialogue on Two Chief World Systems* made its appearance in 1632. Almost immediately an outcry began among Roman theologians who had not forgotten the prohibition implicit in the 1616 decree: the *Dialogue* clearly *did* defend the proscribed view. But Urban's reaction was more personal; in his eyes Galileo had betrayed the trust placed in him. A commission of inquiry referred the case to the Holy Office; Galileo was summoned to Rome and put on trial in early 1632.

Galileo could claim that his book had, after all, received the censors' approval. But against this, the investigative judge could argue that the censors had not been told about the personal injunction given him in 1616 not even to discuss Copernicanism. Against this in turn, Galileo could introduce in evidence a letter from Bellarmine that seemed to call in question whether in fact a personal injunction of that severity *had* in fact been given him. The depositions continued; Galileo argued, although implausibly, that his book did not really urge the superiority of the Copernican view. Finally, a plea-bargain was struck; Galileo agreed to abjure publicly the prohibited claims of the Earth's motion and the Sun's rest and was sentenced to prison, the sentence promptly commuted to house-arrest.

Back finally at home, he returned to the long-interrupted manuscript on mechanics. In 1638, the *Two New Sciences* was published in Holland. Here he laid out the two laws of fall that were to become the foundation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of modern mechanics, following a method that appealed to experience yet also sought the certainty of geometrical demonstration. It came as final affirmation for one who in his later years had known more than his share of tragedy.

Telescopic discoveries

Galileo's contributions to astronomy were of two very different kinds. First were the stunning observational discoveries of 1609–11. These did not involve the calculational feats of his great predecessors, Ptolemy, Copernicus and Kepler. Galileo was fortunate to be the first to be enabled to see in the heavens what no one before him had seen; what made him a great astronomer was his ability to grasp the significance for cosmology of what he saw. Second was his furtherance of the case for the Copernican system. He did more than any other single person to bring about the displacement of the Earth from its privileged place at the cosmic center with the vast shift in world view that followed.

His lunar observations, vividly chronicled in the *Sidereus Nuncius*, made the greatest impact, perhaps, on the general public. Patches of light and shadow could now be plausibly identified as Earth-like features. Prominent among them were what seemed to be lunar mountains, whose peaks could be seen as spots of light at lunar dawn while the lowlands below them remained in shadow. And the mountains' own shadows shortened as the lunar noon approached, just as one would expect. Further, the faint illumination of the dark part of the Moon's face could best be understood as 'earthlight', sunlight reflected from Earth, at its brightest when the Moon lies between Earth and Sun so that the 'full Earth' is shining on it; Earth and Moon thus reflect sunlight in the same way.

The discovery that four satellites circled Jupiter showed that, contrary to ARISTOTLE's doctrine of natural motion, there is more than one center of rotation in the planetary system. It also eliminated an objection to the Copernican claim that the Earth carried its Moon along as it circled the Sun. Galileo took a great deal of trouble with the difficult problem of computing the periods and the times of occultation by Jupiter of each of the four satellites (see JUPITER: SATELLITES). He hoped that it would allow the compilation of tables of occultation times at a reference base like Florence, constituting a celestial 'clock' sufficiently accurate for the calculation of longitude for ships at sea, a crucial need for navigators on the long sea voyages then becoming common. The practical difficulties involved in making the requisite observations while at sea defeated his plans, but this method of measuring longitude came to be relied on by mapmakers later in the century.

The phases of VENUS from crescent to full and back again became Galileo's most potent argument against the older cosmology. If Venus and the Sun both orbit the Earth, with Venus below the Sun, Venus should always show a crescent shape to Earth, if illuminated by the Sun. If, however, Venus and Earth both orbit the Sun, Venus should run through the same sequence of phases as does our own moon. And this Galileo observed it to do, enough to undermine the Aristotelian–Ptolemaic world system irreparably. The appearances of Venus were, however, compatible with the TYCHONIC SYSTEM which had the Sun orbit the unmoving Earth carrying the entire planetary system with it, a major reason why astronomers who were unwilling, for physical or theological reasons, to embrace the Copernican system, preferred the Tychonic alternative.

Galileo saw the SUNSPOTS as something like terrestrial clouds, close to or on the Sun's surface and constantly changing in shape. Their rotation testified to a solar rotation taking about a month. Here again, the warrant for his hypothesis, suggested by familiar terrestrial analogy, was the observable consequences that could be drawn from it. In the *Dialogue*, he made use of the observed shapes of the sunspot paths to argue that these shapes would follow naturally if the annual motion of the Earth be granted, whereas they would demand an implausible combination of four motions on the Sun's part if the Earth's motion be denied.

These discoveries brought to the fore an issue for which the methodology of Aristotelian science was ill-prepared. How are the natures of distant and unfamiliar objects to be discovered? The intuitive grasp of necessary relations between essence and property on which Aristotle's account depended required a direct familiarity with the natures themselves. But with sunspots, comets and the like, one had to resort, instead, to the imagination and test the resultant construct by the observable consequences drawn from it. Galileo's treatment of the hypothetical analogy between the lunar and terrestrial surfaces offers a particularly elegant example of this sort of retroductive reasoning from effect to cause that relies for its warrant on explanatory success, rather than on deductive derivation from intuitively warranted principle. Galileo never explicitly acknowledged this shift, one with far-reaching consequences for later science.

Arguing for the Copernican system

The *Dialogue on Two Chief World Systems* is a masterpiece of intricately interlocking argument in support of the Copernican system. Galileo had to deal first with the Aristotelian arguments against the Earth's motion. This would require a new mechanics; indeed, some have claimed that such a goal may have prompted his groundbreaking early work on motion. He had to challenge Aristotle's way of distinguishing motions into natural and violent; the EARTH'S ROTATION could be neither. Against the objection that the motion of the Earth, if real, should be perceivable, he had to claim the imperceptibility of shared uniform motion (allowing circular motion to count as uniform). Against the strongest Aristotelian objection that on a rotating Earth objects would tend to fly off, he had to formulate a version of the principle of inertia. But surely if the Earth were to rotate fast enough, experience with spinning millstones showed that objects would fly off? Without an understanding of gravity, Galileo's response to this objection was ingenious but finally unconvincing.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Responding to objection was, of course, not going to be enough. Positive arguments in this domain, Galileo notes, are of two sorts, one based on astronomical observations, the other on terrestrial effects. And the first kind, based on shared celestial motions, can only afford 'greater probability', he admits. If we observe a diurnal motion shared by *all* the celestial bodies, is it not simpler to suppose that the motion is really that of the Earth? The argument here, Galileo emphasizes, is not just from economy. It is based on the ability to *explain* the motion. Whereas *he* can give an account of the Earth's inertial rotation, his opponents have nothing to offer in regard to the complex causes required to keep so large and so complex a system revolving around a relatively small Earth. The carrier spheres are no longer believable (he does not mention Tycho's role in banishing them), so how could the myriad stars (now multiplied enormously in number by telescopic testimony) retain their positional relationship with one another over the ages, as they are observed to do? What sorts of forces could bring this about? Galileo enumerates seven arguments of this general kind in favor of the Earth's rotation. Though he does not say so, these arguments would also tell against the third 'chief world system', that of Tycho, which he studiously ignores throughout the *Dialogue*.

The shared-motion argument for the *annual* motion of the Earth distinguishes between the planets and the fixed stars. It is a harder case to make since Galileo has no explanation to offer for this motion. Further, it leads to awkward questions about the absence of stellar parallax. Galileo conjectures in response that the stars must be at an enormous distance from us. Oddly, he makes no use of the 'coincidence' arguments proposed by Copernicus and Kepler: various features of the planetary motions (e.g. the fact that planets are brightest when in opposition) are 'natural' in the Copernican system, unexplained coincidence in its rival.

He realized that to achieve the sort of proof that his critics, particularly his theological critics, demanded, he would have to produce a causal inference from some terrestrial phenomenon to the Earth's motions as its only possible explanation, the classical Aristotelian requirement for demonstration. As early as the 1590s, he had speculated that the ocean tides might furnish the needed proof. He obviously meant this to be the culminating argument of the *Dialogue*; indeed, he had to be ordered not to work a reference to it into the book title. The double motion of the Earth, allied with the peculiarities of individual sea coasts, could, he labored to show, explain the general features of the TIDES. Though some modern scholars are sympathetic to his tidal argument, his readers were (and still are) on the whole unconvinced. The tie between the tides and the lunar phases had been noted since antiquity and did not easily fit into Galileo's scheme. (He rejected outright the possibility that lunar attraction might be responsible.) More seriously, in order to meet Aristotelian objections he had already asserted, on several grounds, the imperceptibility in principle of terrestrial effects of the Earth's motions. How could he now turn around and rely on such terrestrial phenomena as the tides to provide *demonstration* of these motions?

The cumulative effect of the *Dialogue* arguments was, thus, mixed. The arguments *against* the Aristotelian–Ptolemaic system were decisive; it had no serious support from astronomers from this time forward. The positive arguments *for* the Copernican alternative were much less decisive. The tidal argument did not convince. The shared-motion arguments had to bear the burden, and this they did, provided one settles for a high degree of probability. And this is just what the scientists of Galileo's century were finding it very difficult to do.

Bibliography

Drake S 1978 Galileo At Work: His Scientific Biography (Chicago: University of Chicago Press)

McMullin E 1967 Galileo, man of science *Galileo: Man of Science* ed E McMullin (New York: Basic Books) pp 3–51

McMullin E 1978 The conception of science in Galileo's work *New Perspectives on Galileo* ed R Butts and J Pitt (Dordrecht: Reidel) pp 209–57

Shea W R 1972 *Galileo's Intellectual Revolution* (London: Macmillan)

Shea W R 1990 Galileo Galilei: an astronomer at work *Nature, Experiment, and the Sciences* ed T H Levere and W R Shea pp 51–76

Swerdlow N M 1998 Galileo's discoveries with the telescope and their evidence for the Copernican theory *The Cambridge Companion to Galileo* ed P Machamer (Cambridge: Cambridge University Press) pp 244–70

Van Helden A 1984 Galileo and the telescope *Novità Celesti E Crisi Del Sapere* ed P Galluzzi (Firenze: Barbèra) pp 149–58

Van Helden A 1989 Galileo, telescopic astronomy, and the Copernican system *Planetary Astronomy from the Renaissance to the Rise of Astrophysics* ed R Taton and C Wilson (Cambridge: Cambridge University Press) pp 81–105

Wallace W A 1992 *Galileo's Logic of Discovery and Proof* (Dordrecht: Kluwer)

Ernan McMullin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galileo Mission to Jupiter

The Galileo mission to JUPITER provided several historic firsts in planetary and solar system exploration: the first *in situ* sampling of an outer planet atmosphere, the first extended study of an outer planet, its satellites and magnetosphere from an orbiter, the first images of an asteroid, together with the first confirmation that asteroids can have gravitationally bound moonlets and, by an amazing set of circumstances, the only direct view of the impact of comet Shoemaker–Levy 9 with Jupiter.

The spacecraft consisted of two main components: an atmospheric entry probe and a planetary orbiter. The Galileo probe was designed and built by Hughes Space and Communications Group under contract to NASA Ames Research Center. The Jet Propulsion Laboratory designed and built the Galileo orbiter and had overall responsibility for mission design and operations. The German government was a partner in the mission through its provision of the orbiter propulsion subsystem and two science experiments. Scientists from six nations participated in the mission.

The principal scientific objectives of the mission were (1) to study the composition, structure and dynamics of the jovian atmosphere, to a depth having a pressure of at least 10 bar, (2) to study the composition, geology, surface morphology and interior structure of the planet-sized Galilean moons Io, Europa, Ganymede and Callisto and (3) to determine the structure of, and characterize the plasma physics in, the vast jovian magnetosphere. In spite of some serious spacecraft hardware problems which occurred on the way to Jupiter, the Galileo Mission met all these objectives, plus it produced some first-order science prior to arrival at Jupiter. The instrument suites aboard the probe and orbiter are given in table 1. All instruments functioned well.

The Galileo spacecraft arrived at Jupiter on 7 December 1995, after having been launched on 18 October 1989 by the space shuttle Atlantis from Kennedy Space Center. The original launch date for the mission was to have been January 1982, but various delays, many not associated with the spacecraft, led to a target launch date of May 1986. However, the Challenger shuttle accident occurred in January 1986, delaying Galileo's launch for yet another 3 yr. The launch delays meant that the trajectory to Jupiter had to be completely redesigned because of Shuttle payload, safety and launch limitations. The ingenious trajectory redesign resulted in a trajectory that included a close flyby of Venus (February 1990) and two close flybys of Earth (December 1990 and December 1992) for gravity assists to gain sufficient energy to reach Jupiter. By serendipity this trajectory allowed the first-ever flyby reconnaissance of two asteroids and the direct viewing of the impacts with Jupiter of the kilometer-sized fragments of comet Shoemaker–Levy 9.

Table 1. Galileo scientific payload.

Investigation	Purpose
Probe	
Atmospheric Structure Instrument	Jovian atmosphere thermal structure
Energetic Particle Detector	Inner radiation belt parameters
Lightning/Radio Emission Detector	Jovian atmosphere optical–radio lightning properties
Helium Abundance Detector	Jovian atmosphere relative helium abundance
Nephelometer	Jovian atmosphere cloud parameters, location
Net Flux Radiometer	Solar–planetary radiative flux in jovian atmosphere
Neutral Mass Spectrometer	Jovian atmosphere composition, 2–150 amu
Doppler Wind Experiment	Zonal wind profile in jovian atmosphere
Radio Scintillation Experiment	Turbulence, signal absorption in jovian atmosphere
Ground Based Doppler Wind[a]	Zonal wind profile in jovian atmosphere
Orbiter	
Solid-State Imager	Image Galilean satellites, Jupiter circulation
Near-Infrared Mapping Spectrometer	Jupiter and satellite composition, temperature
Ultraviolet Spectrometer	Gases and aerosols in jovian atmosphere
Extreme Ultraviolet Spectrometer	Ion, airglow emissions
Photopolarimeter/Radiometer	Characterize atmospheric aerosols, energy balance
Magnetometer	Magnetic field strength and direction
Energetic Particle Detector	High-energy ions and electrons
Plasma Detector	Characteristics of low-energy particles
Heavy Ion Counter	Fluxes and composition of heavy ions
Plasma Wave Detector	Electric and magnetic waves, wave–particle interaction
Dust Detector	Properties of dust near Jupiter
Radio Science	Celestial mechanics, gravity fields, atmospheric structure

[a] Not an originally selected Galileo investigation, but endorsed by the Galileo Project Steering Group.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The probe was separated from the orbiter on 13 July 1995. At this point both the orbiter and the probe were about 50 million miles from Jupiter and 5 months away from arrival. The probe plunged into the atmosphere of Jupiter on 7 December 1995, 14:04 PST, at a latitude of 6.5ºN, surviving the most difficult atmospheric entry ever attempted in terms of relative atmospheric entry speed, exceeding 47 km s^{-1}. The probe returned data for just over an hour, until it reached a pressure level in the atmosphere of 22 bar. On the same date, after having received and recorded the probe data, the orbiter was the first ever to be placed in orbit about an outer planet and returned data about the jovian system during a nominal mission lasting until December 1997. The orbiter is still functioning at the time of writing this article. The primary purpose of the extended mission beyond the nominal 2 yr is to reconnoiter the jovian moon Europa. Europa is of high interest because of the possible existence of a subsurface liquid water ocean, which conceivably could harbor biology.

En route to Jupiter, the spacecraft suffered a major anomaly on 11 April 1991, when the high gain antenna (HGA) failed to deploy properly. Until this point, the HGA had been folded against a central mast and was protected by Sun shields so that heat loads on the HGA would be within tolerances. Despite repeated attempts to solve the problem, the HGA was not functional for the rest of the mission. Mission communications were carried out using the low-gain broad-beam antenna, with a subsequent reduction in communication rate by a factor of 1000. It is to the credit of all the mission teams that, in spite of this severe problem, Galileo met almost all its fundamental science objectives. The major impact to data return came in the total number of images returned, reduced from the expected approximately 100 000 during the nominal mission to less than 2000, with another approximately 2000 images to be returned during the 2 yr Galileo Europa portion of the mission.

Science highlights en route to Jupiter

Before reaching Jupiter the spacecraft took the first close-up pictures of two asteroids. On 29 October 1991 the spacecraft flew by asteroid GASPRA at a closest approach of about 1600 km. The images returned revealed a cratered, complex and irregular body approximately 19 km by 12 km by 11 km, apparently covered by a thin regolith. On 28 August 1993 Galileo encountered the second asteroid, IDA, with a closest-approach distance of about 2400 km, and discovered that Ida had a small moon (about 1.5 km in diameter). Dactyl, as the moonlet was named, was the first confirmed moon of an asteroid and was estimated to be about 100 km from Ida's centroid. Ida itself was considerably larger than Gaspra, being approximately 56 km by 24 km by 21 km, and had a rotation period of 4.6 h about an offset axis.

Twenty-one kilometer-sized fragments of comet Shoemaker–Levy 9 collided with Jupiter in July of 1994, at a time when the Galileo spacecraft was about 1.5 AU from Jupiter (see SHOEMAKER–LEVY 9–JUPITER COLLISION). The set of collisions of the fragments of SL-9 with Jupiter was one of the most intensely observed events by astronomers in history, but, because of the geometry of its trajectory, Galileo was the only platform that had a direct view of the impact sites because they occurred on the side of Jupiter facing away from Earth. Galileo returned key data concerning the time history and light curves of several of the fragment impacts, information which otherwise would have been unobtainable.

As Galileo passed close by VENUS and the Earth–Moon system on its way to Jupiter, unique sets of observations were accomplished. Among the measurements at Venus were high spatial and temporal resolution imaging of atmospheric motions at cloud levels and compositional measurements from the near-infrared mapping spectrometer. High spatial resolution images of both the Earth and the far side of the Moon were obtained, and a wonderful set of images of the entire Earth–Moon system produced a movie in which the Earth is seen rotating as the Moon orbits around it.

Science highlights at Jupiter

Jupiter

One of the most important scientific objectives of Galileo was to determine the composition of the jovian atmosphere. Composition information is important in order to develop an understanding of solar system formation and evolution. Prior to Galileo, it was of course known that Jupiter was made mostly of hydrogen and helium, but the precise value of the relative abundance of helium was still uncertain to about 25%. Furthermore, although the jovian abundance of carbon was fairly well established, the jovian abundances of the next most abundant non-noble volatile elements in the solar system, oxygen, nitrogen and sulfur, were poorly constrained, if at all. Noble gases other than helium had not even been detected (see JUPITER: ATMOSPHERE).

The Galileo probe made direct measurements of atmospheric composition between levels in the atmosphere corresponding to pressures from 0.4 bar to 22 bar. Indirect measurements of the ammonia abundance profile with depth were made using the probe telemetry signal strength as received by the orbiter. The He/H_2 mass ratio was measured to be 0.24, somewhat higher than the value near 0.18 inferred from the earlier Voyager flybys of Jupiter and closer to the value thought to be appropriate for the proto-Sun, 0.275. Carbon (in the form CH_4), nitrogen (in the form NH_3) and sulfur (in the form H_2S) were all measured to have relative abundances to hydrogen about 3 times the respective solar values,

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

although there is still some uncertainty as to the ammonia determination. Prior to Galileo, remote observations indicated an ammonia abundance less than twice solar. The Galileo ammonia measurements of 3–4 times solar imply that planetesimals which brought heavier elements to Jupiter originated from colder regions of the early solar system than previously thought. Another surprise was the relative abundance of oxygen (in the form H_2O), which was found to be depleted with respect to the solar value, as was the noble gas neon. It is at present thought that the depletion in H_2O does not represent the planet as a whole, but rather local or regional variation in the H_2O abundance. To date, the Galileo orbiter has not detected solar amounts of water in Jupiter's atmosphere either, but all data have not yet been analysed.

Other surprising aspects of the Galileo probe data, which may be related to the region in which the probe descended, were the dependencies on depth of the mixing ratios of the condensibles NH_3, H_2S and H_2O. The probe entered an anomalously clear region of the jovian atmosphere in terms of cloud layers, known as a 5 μm hot spot because emitted radiation from the deeper atmosphere is detected remotely in these relatively clear regions. Three cloud layers were expected, but only one tenuous cloud layer was clearly detected, corresponding most likely to the expected second cloud deck of ammonium hydrosulfide, NH_4SH. The probe probably started measurements below the topmost cloud, expected to be NH_3 ice crystals, and did not encounter an expected water cloud. All the condensible abundances varied dramatically with depth, but the variations were not correlated with condensation levels or with each other. The typical profiles consisted of highly depleted mixing ratios at depths near 0.4 bar, increasing to mixing ratios a few times solar values at depths near and above 10 bar. The one exception was water, which reached a value about 30% solar at the deepest levels probed. Various possible explanations for these observations have appeared, but at the present time there is not a consensus as to the proper interpretations and scientific implications.

The Galileo probe established that the dominant zonal (east–west) winds, which relate to the banded appearance of the planet and which have been observed ever since cloud motions could be followed with telescopes on Jupiter, extend to at least the depth penetrated by the probe. Zonal wind speeds increased from about 70 m s^{-1} at cloud levels near 0.4 bar to 175 m s^{-1} at a pressure level near 5 bar and then remained approximately constant to a depth corresponding to 22 bar. The fact that the winds extend to 22 bar probably implies that the winds are driven from deep in the jovian atmosphere, conceivably very deep.

The thermal structure of the atmosphere below the 1 bar pressure level was measured to be close to that corresponding to a dry adiabat, as had been expected. However, in the upper atmosphere at heights above 1 μbar, temperature rapidly increased with height, reaching a value (depending precisely on choice of temperature at the initial altitude of measurement) of 900 K near 1000 km above the 1 bar level. Solar heating is too small to be the energy source for the upper atmosphere heating. Most probably the energy source is a combination of dissipation of atmospheric gravity waves propagating from the deeper atmosphere with heating of the atmosphere by energetic particle precipitation from the magnetosphere.

Galilean satellites

The Galilean satellites are of high scientific interest for several reasons. In conjunction with Jupiter, these large moons form a system that in certain ways is analogous to the solar system and hence a system that can reveal information regarding the physics of planet formation and evolution (see JUPITER: SATELLITES). Furthermore, these bodies are highly interesting in their own right. IO is the most volcanically active body in the solar system. EUROPA is of major interest because of the possibility, discussed below, of a subsurface water ocean which could harbor biology. GANYMEDE is the largest moon in the solar system, being the size of the planet Mercury. The outermost satellite, CALLISTO, is the same general size as Ganyemede, but appears rather different from the other three satellites in terms of geophysical history.

A much more comprehensive knowledge of the Galilean satellites of Jupiter is now emerging from the Galileo orbiter data. The Galileo orbiter passed close enough to each of these satellites so that tracking of the spacecraft allowed measurement of the quadrupole moment of the gravitational field, from which it is possible to develop models of the satellite's interior. At the same time, magnetic field measurements were made during the satellite flybys which could be used to assess whether the magnetic field perturbation measured was due to an induced field or to an intrinsic field of the satellite itself. Combining these sets of data, the following picture seems to emerge. The inner three satellites, Io, Europa and Ganymede, are strongly differentiated. The best model fit is a three-layer model consisting of a metallic core, rocky envelope and (except for Io) water ice outer shell. The magnetic signature of Ganymede implies an internally generated magnetic field, consistent with the existence of a metallic core. Io's magnetic field has not yet been measured adequately to distinguish what type of magnetic field, if any, it has. The preferred explanation for Europa's magnetic signature is an induced magnetic field driven by time-varying processes in Jupiter's magnetosphere near Europa's orbit. This implies induced currents in Europa flowing in a near-surface conducting layer, such as a salty subsurface ocean. However, an intrinsic field cannot yet be ruled out. Callisto, the outermost Galilean satellite, is partially

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

differentiated with an ice–rock mixture, such that the rock fraction is a slowly increasing function of depth. Callisto's magnetic signature implies an induced field qualitatively similar to that of Europa, requiring a near-surface conducting layer. Such a near-surface conducting layer on Callisto is difficult to understand. Neither an icy crust nor a conducting ionosphere can explain the magnetic observations, and known sources of heating are not expected to be large enough to cause a near-surface liquid or slushy conducting layer.

Spectra of the satellite surfaces imply that, besides water ice, the surfaces of Ganymede and Callisto could contain CH, CO_2, CN, SO_2 and perhaps an SH-radical-containing material. Sulfur-bearing minerals may also be possible constituents. Certain impact scars on Europa seem to contain saline hydrates mixed with the dominant constituent, water ice. Evaporite, hydrated minerals produce good spectral matches to various other features on Europa. The surface of Io of course does not contain water ice because of the active heat release by volcanism, but, besides silicate materials, SO_2 ice grains of size smaller than 500 μm are almost everywhere.

A current question of extreme scientific interest is whether Europa has a liquid water ocean which might harbor biology beneath the frozen surface. Theoretical models of Europa's interior show that the same tidal heating mechanism that produces the volcanoes on Io may act to a lesser degree within Europa to maintain temperatures above the melting point of water ice at depths of a few tens of kilometers. At the present time, probably the most indicative evidence that a subsurface slushy or liquid layer exists consists of Galileo orbiter images, such as that of the Conamara region presented in figure 1. This image shows a chaotic surface terrain, in which it appears that a thin ice crust has been disrupted and moved about in chunks, which resembles disruption of polar ice packs on the Earth during spring thaws. The size and geometry of the 'chunks' imply that a layer of liquid water or soft ice not too far below the surface could have driven the relative motions and rotations apparent in the image. The age of the surface has yet to be unambiguously determined. Depending on which model is used, crater counts imply either a young age of a few million years or less or, instead, a rather old surface of a few billion years. Currently, therefore, evidence exists that at least at some time in Europa's history there existed a 'slushy' near-surface layer. The challenge now is to determine whether that is the present situation.

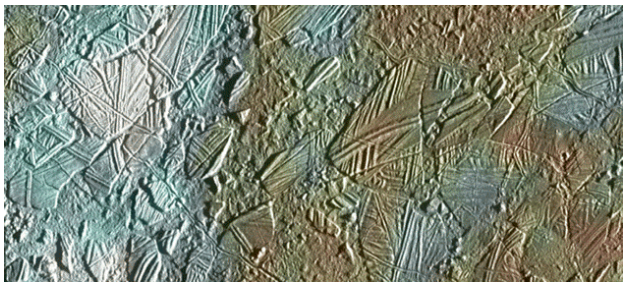


Figure 1. Image of an approximately 70 km × 30 km area of Europa in the Conamara region. The colours in the image have been enhanced for visibility. White and blue colours delineate regions which have been covered by fine dust particles ejected when the crater Pwyll was formed 1000 km to the south (north is to the top of the image, the Sun illuminates the surface from the right). Mineral contaminants are probably responsible for the reddish-brown colours. Crustal plates up to 13 km in size appear to have been created and then moved into new positions, indicating water or soft ice not too far beneath the surface at the time of crustal movement.

Magnetosphere

The magnetosphere of Jupiter is a vast system of plasma and magnetic fields that interacts in one way or another with every object in the jovian system (see JUPITER: MAGNETOSPHERE). The magnetic fields within the magnetosphere originate in Jupiter and the Galilean satellites and from plasma currents within the magnetosphere. The Galileo orbiter has provided key information about magnetospheric geometry which, when combined with previous spacecraft data, can be used to give an overall view of magetopause and bow shock shape and location. On the Sun side of Jupiter, the magnetic equatorial bow shock occurs at about $80R_J$ (jovian radius), and the equatorial magnetopause is near 70 R_J, each value being somewhat time variable depending on solar wind conditions. Galileo measured magnetic perturbations near each of the Galilean satellites, as described above, indicating either internally generated magnetic fields or fields induced by magnetospheric plasma interactions with the satellites. New information obtained within Io's plasma torus revealed a much stronger electromagnetic interaction with Io than had been expected (see IO: PLASMA TORUS). Europa has a tenuous neutral oxygen atmosphere that is both generated and removed by plasma sputtering processes that depend strongly on electrodynamic conditions near Europa. Comparison of modelling results with Galileo and other data implies an oxygen column density of about 5×10^{18} m^{-2}.

Summary

Galileo has been a pioneering mission in many ways. In spite of the early launch delays, the high gain antenna malfunction, and having to wait for many years to get to Jupiter, Galileo has provided a picture of the jovian system that is unprecedented. As with all great planetary missions, the results from Galileo form the basis for, and point the way to, further exploration of the jovian system.

In 2001, the Galileo mission was given the go-ahead for a mission extension that includes five more fly-bys of the Jovian moons before a final plunge into the crushing pressure of the giant planet's atmosphere. Galileo's mission has previously been extended twice. On 25 May

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

2001, Galileo passed about 123 km (76 mi) above Callisto. in August 2001, the Galileo spacecraft flew past the north pole of Jupiter's moon Io, passing through a swarm of sulfurous particles hurled into space by a previously unknown volcano. Scientists had not expected the spacecraft to encounter volcanic 'ash'. A different volcano named Tvashtar had been lofting a plume seven months earlier, but Galileo saw no sign of that plume during its Io flyby in early August. Instead material was erupting from a previously unknown volcano just 600 km (370 mi) from Tvashtar. On October 1, 2001, at 0123 UT, the Galileo spacecraft approached the closest yet to the surface of Io, to within 181 km (112 mi) of ground level near Io's south pole.

In 2002, having completed its imaging mission, Galileo will continue studies of Jupiter's massive magnetic field with seven instruments. In November 2002, it will swing closer to Jupiter than ever before, dipping within about 500 km (300 mi) of the moon Amalthea. Scientists will use Galileo measurements to determine the mass and density of Amalthea. They will also study dust particles as Galileo flies through Jupiter's gossamer rings and seek new details of the magnetic forces and the densities of charged particles close to the planet. Galileo's final orbit will take an elongated loop away from Jupiter. Then, in August 2003, the spacecraft will head back for a direct impact and burn up as it ploughs into Jupiter's 60000 km-thick atmosphere.

Breaking news update (30 April 2002)

The Galileo spacecraft resumed gathering scientific information at about 4:00 Universal Time on January 18, 2002, after commands radioed from Earth took the Jupiter orbiter out of the passive standby mode it entered on the previous day. On Thursday January 17 at about 13:41 Universal Time, while making a close approach to Jupiter's innermost Galilean moon, Io, the spacecraft detected a computer reset, which caused it to enter a 'safe' mode. In this mode, onboard fault protection software instructs the spacecraft cameras and science instruments to stop taking data and places them in a safe state awaiting further instructions from the ground. The reset was apparently caused by exposure to the intense radiation environment at Io's distance from Jupiter. Since the spacecraft began orbiting Jupiter in 1995, it has endured a cumulative radiation exposure about three-and-a-half times what it was originally designed to withstand. After resuming operations, Galileo passed within about 102 km (63 mi) of Io. Planned observations for the remainder of the spacecraft's current swing near Jupiter include a series of images of the planet's atmosphere, a farewell color study of its icy moon Europa and navigational imaging of the small moon Amalthea.

A recent image from NASA's Galileo spacecraft adds evidence to a theory that Callisto may hold an underground ocean. The image shows a part of Callisto's surface directly opposite from the Valhalla basin where Callisto was punched by a major collision. The opposition point shows no effect from the impact. Points opposite major impact features on some similar-size worlds, such as Mercury and Earth's Moon, show lumpy terrain attributed to seismic shocks from the distant impacts. The new image is consistent with a 1990s' model proposing that a liquid layer could be acting as a shock absorber inside Callisto.

Bibliography

Detailed descriptions of the Galileo spacecraft, its trajectory to Jupiter, and the instruments on the probe and orbiter can be found in articles appearing in

1992 *Space Sci. Rev.* **60** (1–4)

Various observations conducted during the transit to Jupiter were reported in the following publications. Regarding the flyby of Venus, see articles in

1991 *Science* **253** 1457–612

Regarding Earth–Moon, see

Belton *et al* 1994 *Science* **264** 1112–15

Two special issues reporting the Galileo observations of asteroids Gaspra and Ida were, respectively

1994 *Icarus* **107** 1–97
1996 *Icarus* **120** 1–245

Regarding Galileo observations of the impact of Shoemaker–Levy 9 with Jupiter, see

Hord *et al* 1995 *Geophys. Res. Lett.* **22** 1565–8
Weismann *et al* 1995 *Abstr. Lunar Planet. Sci. C.* **26** 1483

Science results and parameters from the probe entry and descent into Jupiter's atmosphere were first published in

1996 *Science* **272** 837–60

Orbiter results at Jupiter were first published in

1996 *Science* **274** 377–403

Subsequently many papers have appeared discussing Galileo observations. The reader is referred here to three special journal issues which should provide good starting points to pursue a particular question. Two special issues

1998 *J. Geophys. Res.-Planet* **103** (E10)
1998 *J. Geophys. Res.-Planet* **103** (E9)

provide comprehensive analyses of probe and magnetospheric observations, respectively. Remote sensing results from the Galileo orbiter appear in

1998 *Icarus* **135** (1)

Richard Young

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Galle, Johann Gottfried (1812–1910)

Astronomer, born in Pabsthaus, Germany. In 1846, at the Berlin Observatory, he discovered the planet Neptune, whose position had been calculated by LEVERRIER. Leverrier had written to Galle asking him to search for the 'new planet' at a predicted location. Galle, together with his assistant HEINRICH D'ARREST, began a search on the same night that they received the letter. At D'Arrest's suggestion, Galle used the latest star chart that had only just been produced by Encke. In 30 minutes they had located a star not on the map, confirming that it was the new planet on the following night by its motion relative to the other stars. Galle wrote to Leverrier, saying, 'Monsieur, the planet of which you indicated the position really exists'. Leverrier replied, 'I thank you for the alacrity with which you applied my instructions. We are thereby, thanks to you, definitely in possession of a new world'. Galle also detected Saturn's C-ring and observed minor planets, attempting to determine the solar distance through the observation of the close approach of one (see DAVID GILL, HAROLD SPENCER JONES).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamma Draconis

Sometimes known by its ancient name Eltanin (or Etamin), the 'Head of the Dragon', from its description in the revision of Ptolemy's *Syntaxis* published by Ulugh Beigh in the fifteenth century. The star has an apparent magnitude of 2.24 and is of spectral type K5III. At a distance of 148 light-years, it has a parallax of 0.022″ and an absolute magnitude of −1.0.

This second-magnitude star was, for nearly 300 years, of considerable importance to positional astronomers, especially those observing from the Royal Observatory at Greenwich. Due to its declination (51° 29.3′) being approximately equal to the latitude of Greenwich (51° 28.6′), it transits there very close to the zenith: 4′ north of the zenith in the time of the first Astronomer Royal, John Flamsteed (1646–1719), and only 2′ north by the time of the seventh Astronomer Royal, G B Airy (1801–92), who described it as 'the birth star of modern astronomy'.

Transiting zenithally is very advantageous, for not only does it simplify the telescope mounting required, and hence improves stability, but the positional errors caused by the effects of atmospheric refraction are zero at the zenith. These advantages had long been realized. In 1669 Robert Hooke (1635–1703) set up a vertical telescope at Gresham College in the City of London, in an attempt to measure the annual parallax of γ Draconis and hence its distance. It was not then realized how distant the stars are, the parallax of 0.022″ would in fact have been impossible to measure with the instruments of Hooke's time. In 1677 Flamsteed had a 90-foot well dug at Greenwich, to act as the 'tube' of a vertical refractor to observe the star, but observations made with it were fraught with difficulty and proved to be of little use.

In 1725 James Bradley (1693–1762), Savilian Professor of Astronomy at Oxford, wished to observe γ Draconis for the same purpose. He joined the wealthy English amateur astronomer Samuel Molyneux (1689–1728) at Kew, where they used Molyneux's 24.5 foot zenith sector by George Graham (1675–1752). Like Hooke, they obtained varying results for the zenith distance of the star at transit, but these did not seem to be explained by an annual variation due to parallax. Intrigued by these discrepancies, Bradley commissioned an excellent new 12.5 foot zenith sector from Graham, and set it up at his home in Wanstead. After making many more observations of the star Bradley arrived at the explanation of the variations in its observed position—and announced the discovery of the aberration of light. Further work led to his discovery of nutation, announced in 1748. Following his appointment as the third Astronomer Royal in 1742, Bradley persuaded the Admiralty to purchase his zenith sector so that these observations could be continued at Greenwich.

As the principal 'zenith star' γ Draconis continued to be observed at Greenwich for a further two centuries. In 1851 Airy introduced a reflex zenith tube of his own design to continue the observation of this and other much fainter zenithal stars, and visual observations of greatly increased accuracy were made with this instrument over the next 60 years. From 1911 to 1950 it was superseded by the floating zenith tube previously used at Cambridge by Bryan Cookson (1874–1909).

The success of Airy's design for the reflex zenith tube led to the development of the first photographic zenith tube by F E Ross of the US Naval Observatory, Washington, in 1911. A much improved and fully automated version was designed and installed at the Royal Greenwich Observatory at Herstmonceux in 1955, the first of a network of similar instruments set up at major observatories around the world. These proved to be the most accurate instruments ever constructed for the astronomical determination of time and latitude—a remarkable tribute to Airy's design, a century earlier, of an instrument specifically to observe γ Draconis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamma Rays

The most energetic form of electromagnetic radiation. Gamma rays have wavelengths shorter than 0.01 nanometers (10^{-11} m) and individual photon energies greater than about 10^5 electron volts (eV). Despite their high energies and penetrating power, gamma rays from cosmic sources are absorbed well above ground level by the atmosphere. Even the most energetic gamma rays so far detected, with wavelengths of around 10^{-9} nm and energies in excess of 10^{12} eV, are absorbed by the time they have penetrated down to about 10 km above ground level.

The term 'gamma rays' may also be written as 'γ-rays', where γ is the Greek letter 'gamma'.

See also: electromagnetic radiation, electromagnetic spectrum, electronvolt, gamma-ray astronomy, Geminga.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamma-Ray Astronomy

Gamma-ray photons are the most energetic ones of the whole electromagnetic spectrum. Like all photons with wavelengths shorter than 10 nm, they are absorbed by the Earth's atmosphere. Detection of γ-ray emission from space requires then the use of stratospheric balloons, rockets and satellites, which became feasible only with the advent of the space era.

The borderline between the spectral regions of x-rays and γ-rays is sometimes placed at an energy of $E \sim 30$ keV (wavelength $\lambda \sim 4.13 \times 10^{-9}$ cm or a frequency of $\nu \sim 7.2 \times 10^{18}$ Hz). At lower energies, radiation from astrophysical sources is dominated by thermal emission (i.e. resulting from radiation in equilibrium with matter, in optically thick media), while at higher energies non-thermal emission prevails. Alternatively, the x–γ-ray borderline may be placed at 511 keV (the energy equivalent of the electron rest mass); above this 'limit' electrons behave relativistically. In any case, the use of large photon collectors becomes impossible above a few tens of keV: photons have wavelengths smaller than typical interatomic distances in solids and cannot be reflected by any kind of mirror (contrary to what happens at longer wavelengths).

The γ-ray domain extends over 12 decades in energy, as much as all other spectral regions put together. However, above a few tens of GeV, the number of photons received from even the brightest sources is so small that their observation during the limited lifetime of a satellite is no longer possible (see VERY-HIGH-ENERGY GAMMA-RAY SOURCES). Those very high energy (VHE) photons are detected through their interaction with the Earth's atmosphere, producing cascades of secondary particles.

Astronomy in the γ-ray domain faces several serious handicaps.

- The impossibility of using large collectors to focus the γ-ray photons implies that the detector itself should be as large as possible.
- A detector placed outside the Earth's atmosphere is constantly bombarded by energetic cosmic ray particles, coming from the outer space or trapped in the Van Allen radiation belts. The interactions of these particles with the detector induces very high backgrounds, which are difficult to attenuate with passive or active shielding. As a result, the sensitivity of γ-ray detectors increases only as the square root of their collecting area.
- Astrophysical sources emit generally very small numbers of γ-ray photons. Statistically significant detections then require very long exposures, which are not always compatible with the generally short lifetime (a few years) of an observatory in space.

These difficulties explain why γ-ray astronomy is the last window of the electromagnetic spectrum to be opened to our observation of the universe. However, it has undergone a spectacular development in the 1990s and has grown to a mature astrophysical discipline. It offers information that is impossible to obtain in other wavelengths, allowing us to probe the close surroundings of a black hole or the interiors of a supernova, to identify nuclear isotopes and to obtain clues to the origin and acceleration of cosmic rays, either in our Galaxy or in the very active central regions of remote galaxies.

A short history of γ-ray astronomy

The opening of the γ-ray window was preceded by a relatively long period of theoretical studies concerning the γ-emissivity of potential astrophysical sources. Most of these ideas were put forward by scientists working in fields other than astronomy (particle and nuclear physics and cosmic ray physics). Following the detection of atomic hydrogen in the Galaxy (through its characteristic 21 cm line; see RADIOASTRONOMY) Hayakawa realized in 1952 that the interaction of COSMIC RAYS with the hydrogen atoms of the interstellar medium should give rise to γ-ray emission. Six years later, Morisson suggested two known radiosources (the CRAB NEBULA and CYGNUS A) as possible γ-ray emitters. Thus, theoreticians realized quite early that relativistic particles could generate photon emission in both extremes of the electromagnetic spectrum.

Progress in observational γ-ray astronomy has been too slow to test those ideas, because of the smallness of early detectors and underestimated background fluxes. The first γ-ray emission of extraterrestrial origin was detected during the solar flare of March 1958. However, it was only 10 yr later that γ-ray emitters outside our solar system were detected: the US satellite OSO-3 (launched in 1961) recorded a few hundred photons of energy >50 MeV from the inner Galaxy, while a balloon-borne experiment detected photons in the 100–300 keV range from the direction of the Crab nebula. A few months later, radioastronomers discovered the CRAB PULSAR; subsequent analysis of the γ-emission from the Crab revealed a periodicity similar to that in the radio band. During a solar flare in August 1972, a spectrometer aboard the US satellite OSO-7 recorded the first γ-ray spectrum of the Sun, revealing several lines of nuclear origin, as well as the electron–positron annihilation line at 511 keV. Also, the following year a US team reported the detection of γ-ray bursts of extraterrestrial origin by the military satellites of the Vela series (launched to monitor the Earth's atmosphere for nuclear explosions).

After these early discoveries, the γ-ray sky was explored in some detail in the 1970s by the US SAS-2 and the European COS-B satellites, which provided the first high-energy images of the Milky Way, detailed observations of the first known γ-ray pulsars and the first detection of extragalactic γ-ray sources. Extrasolar γ-ray spectroscopy started in the same period, with the discovery of the electron–positron annihilation line from the galactic centre direction by a balloon-borne spectrometer. This discovery was confirmed in the early 1980s by a γ spectrometer aboard the US HEAO-3 satellite, which discovered also the first extrasolar γ-ray

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

line emission of nuclear origin, namely the 1.8 MeV line of the radioactive nucleus ^{26}Al.

The pace of discoveries slowed down in the 1980s, as a result mainly of the lack of dedicated γ-ray satellites. The explosion of the supernova SN1987A in the nearby galaxy of the Large Magellanic Cloud on February 23 1987 offered the γ-ray community an unprecedented opportunity: the SMM (Solar Maximum Mission) satellite, as well as several balloon borne instruments, detected γ-ray lines from the decay of radioactive ^{56}Co (predicted long ago to be the source of power for the late optical light curves of supernovae). Finally, the early 1990s constitute the golden era of γ-ray astronomy, thanks to the discoveries of the Soviet–French SIGMA mission (see GRANAT) and the US COMPTON GAMMA RAY OBSERVATORY (CGRO).

Emission processes of γ-ray photons

Nature creates γ-ray photons in a variety of ways. If the temperature of the environment is high enough ($>10^8$ K), γ-photons are produced as thermal emission of the astrophysical PLASMA and they generally have a blackbody spectrum. The interiors of stars in their advanced evolutionary stages reach such high temperatures, but they cannot be observed, because the star's dense material makes it opaque to photons of all wavelengths.

The close neighborhood of a strongly gravitating body, such as a black hole accreting matter from a companion star, may also reach high temperatures, as the gravitational energy of accreted matter turns into heat. The detection of γ-ray photons produced in this way is a unique tool for studying the interactions of compact objects with their environment as well as their energetics.

However, astrophysical γ-rays are mainly produced by various non-thermal processes, involving in general the interactions of high-energy particles. For instance, high-energy electrons radiate γ-rays by interacting electromagnetically with nuclei, photons or intense magnetic fields; the corresponding processes are known as bremsstrahlung, inverse Compton emission and synchroton emission, respectively, and they can obviously take place also in a high-temperature plasma. In addition nuclear interactions of protons with energies higher than $\sim$1.2 GeV produce secondary unstable particles, pions and mesons; γ-ray photons are produced by the decay of some of these particles, the most prolific producers being the π^0 pions. The resulting γ-ray spectrum has a characteristic maximum around 70 MeV (corresponding to half the rest-mass energy of π^0).

All these thermal and non-thermal processes give rise to the emission of γ-rays with a continuous spectrum of energy. There are also processes producing γ-rays with discrete energies, involving in general the de-excitation of atomic nuclei. Just as the more familiar optical lines are caused by the electron transitions in atoms, nuclear lines result from the much more energetic transitions (typically, in the MeV range) between energy levels of the atomic nuclei. This makes γ-ray spectroscopy an invaluable diagnostic tool for nuclear astrophysics, since it allows for an unambiguous identification of isotopic species (radiation originating from electrons enables one in general, to identify chemical elements, but not isotopes).

The excited nuclear states are populated either through the decay of an unstable nucleus or through high-energy nuclear interactions. In the former case, we obtain information about the synthesis of radioactive nuclei in various sites (mostly explosive ones such as supernovae) and the physics of these sites. In the latter case, we learn about the nature of the energetic particles (e.g. their composition, spectrum, acceleration mechanism etc). γ-ray lines from the de-excitation of nuclei are routinely observed in the Sun's photosphere during solar eruptions; in some cases, the nuclei are produced through neutron captures, in a medium sufficiently dense to favor these reactions, but also sufficiently thin to allow γ-rays to escape.

Finally, the annihilation of an electron and a positron also gives rise to emission in the γ-ray domain. Positrons may be produced in astrophysical sites through the decay of π^+ pions or through the β^+ decay of radioactive nuclei. Depending on the conditions they may annihilate directly with electrons or combine first with them to form positronium. In the former case (at high temperatures, $T>10^6$ K) two γ-ray photons of energy 0.511 MeV are produced. In the latter case, the resulting emission depends on the state of positronium: in 25% of the cases it is formed in the *para* (singlet, $^1S_0$) state and two photons of 0.511 MeV are again emitted; in 75% of the cases it is formed in the *ortho* (triplet, $^3S_1$) state and a continuous spectrum of three photons, each with an energy of $\leq$0.511 MeV, is then produced.

The γ-ray sky

The poor resolving power of current γ-ray instruments is one of the main problems in γ-ray astronomy (see GAMMA-RAY TELESCOPES). Observations in this energy range usually define 'error boxes', regions of the sky where the source most probably lies. Instruments launched in the 1990s provide error boxes of $\sim$1 arc min at low energies (the SIGMA telescope) and $\sim$5—10 arc min at higher energies (the EGRET instrument aboard CGRO), to be compared with, for instance, $\sim$0.1 arc sec in the visible. Identifying a source in these conditions is much more difficult than at other wavelengths. In the case of a variable source, the time-scale of variability sets an upper limit to the dimensions of the source (which cannot be larger than the distance light travels during this time-scale).

The sources of the γ-ray sky are extremely varied: some of them appear as 'points' (i.e. with angular dimensions smaller than the resolution of γ-detectors) and others as extended sources, both in our Galaxy and in the remote universe.

Stars on the main sequence (i.e. burning hydrogen in their cores) do not, in principle, radiate γ-rays and only the Sun has been observed to do so sporadically. During their advanced evolutionary stages, some stars may eject radioactive isotopes through their powerful

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stellar winds and become γ-ray emitters (red giants, Wolf–Rayet stars). However, radioactivity is mostly produced during stellar explosions (supernovae and, to a lesser extent, novae). On the other hand, stellar remnants may produce copious amounts of γ-rays in several cases: rotating neutron stars may convert a fraction of their rotational energy into high-energy photons (γ-ray pulsars), while accreting black holes convert the energy of accreted matter into low-energy γ-rays in their vicinity and present a steady or variable emission. Some of these objects produce jets of high-energy particles which interact with their environment to produce γ-rays. Finally, the emission mechanism of the mysterious γ-ray bursts is not understood yet, although it—most probably—involves compact objects (neutron stars and black holes).

The Sun

Since the first detection of its γ-ray emission, in 1958, the Sun has been regularly monitored by US, Soviet and Japanese satellites (OSO, HEAO, Prognoz, SMM, Hinotori, Yohkoh, CGRO). γ-rays are emitted continuously, but particularly during a brief period of intense activity, associated with the 11 yr cycle of solar flares (see SOLAR FLARES: GAMMA RAYS). During this period the energy stored in unstable magnetic configurations of the solar surface layers is suddenly liberated; particles are accelerated to high energies and their interactions produce γ-rays. The study of the γ-ray emission provides information on the acceleration sites, their physical conditions and their composition.

The low-energy part of the continuum γ-ray spectrum of the Sun results from bremsstrahlung emission of accelerated electrons interacting with the solar atmosphere. At higher energies (100 MeV) part of the emission is due to the decay of π^0s produced by the interactions of high-energy accelerated protons.

The de-excitation of atomic nuclei (excited by high-energy collisions) gives rise to prompt γ-ray lines, which may be narrow (when ambient nuclei are excited by energetic protons and α particles) or broad (when fast nuclei are excited by ambient protons and α particles). The most intense narrow lines are those at 4.438 MeV and 6.129 MeV, resulting from the de-excitation of the abundant ^{12}C and ^{16}O nuclei, respectively; the lines of several other nuclei (^{15}N, ^{20}Ne, ^{24}Mg, ^{28}Si and ^{56}Fe) have also been detected (figure 1). The relative intensities of these lines provide information on the isotopic composition of the Sun's surface.

The most intense solar γ-ray line, at an energy of 2.223 MeV, results from the de-excitation of deuterium nuclei (D), produced by neutron captures on protons; since free neutrons are unstable (they decay to protons within 10 min) the ones involved in the formation of deuterium are ejected by accelerated heavy nuclei. The ejected neutrons have to slow down to the thermal velocities of the ambient medium before being captured by the protons and thus they give rise to delayed γ-ray line emission. The delay time (about 100 s) provides information on the ambient density ($\sim 10^{17}$ particles cm^{-3}) and the location of the emission site in the solar photosphere (at a depth of $\sim$300 km).

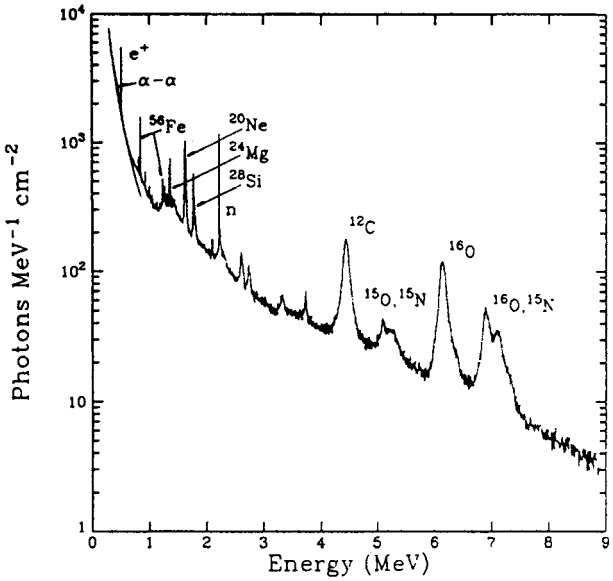


Figure 1. γ-ray spectrum of a solar flare modeled by Murphy *et al* (1991).

The second most intense line observed in the solar γ-ray spectrum results from the annihilation of positrons with electrons. Positrons are produced by β^+ decay of short-lived radioactive nuclei and of π^+ pions (both produced in high-energy interactions). Before annihilating, positrons have to slow down and their γ-ray emission is delayed (as in the case of neutrons).

Stellar explosions and remnants

Radioactive nuclei are produced by thermonuclear reactions in high-density environments, which are opaque to γ-rays. The photons of the radioactive decays interact with the ambient electrons and lose energy via Compton scattering; the medium is heated and the energy is ultimately radiated at longer wavelengths. Only in the case of a site violently expanding after an explosion (SUPERNOVAE, NOVAE) or suffering mass loss through a stellar wind (WOLF–RAYET STARS or stars in the asymptotic giant branch stage) can the γ-ray photons escape, since the opacity decreases with decreasing density.

The characteristic energy of escaping γ-ray lines enables one to identify isotopic species, while their intensity provides information on the amounts that have been synthesized and on the physical conditions of the corresponding stellar zones. The shape of the lines may give information on the structure (velocity and density profiles) of the stellar ejecta. These properties make γ-ray spectroscopy a unique tool for nuclear and stellar astrophysics.

These theoretical ideas, put forward mostly by D D Clayton in the 1960s, were spectacularly confirmed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in 1987: the 0.847 MeV and 1.238 MeV lines of the decay of ^{56}Co were observed in SUPERNOVA 1987A, a massive star ($20M_{\odot}$) that exploded in the nearby galaxy of the Large Magellanic Cloud. Their detection confirmed that ^{56}Fe, the most strongly bound stable nucleus in nature, is produced in the form of unstable ^{56}Ni (through the decay chain ^{56}Ni $\longrightarrow$ ^{56}Co $\longrightarrow$ ^{56}Fe). It also confirmed that the observed late optical luminosity of supernovae (decreasing by a factor of $\sim$2 every 75 days) is due to the energy released by the radioactive decay of ^{56}Co (which has a half-life of 77 days). Finally, the appearance of the γ-ray lines about 6 months earlier than expected from theoretical models suggested that some of the ^{56}Ni nuclei, produced in the dense inner zones of the star, had been mixed into the outer layers.

The extraordinary opportunity offered by SN1987A was due to its proximity; indeed, it was the first supernova visible to the naked eye in the past four centuries, since supernovae are mostly detected in distant galaxies. Moreover, theory suggests that exploding massive stars (characterized as supernovae of type II or SNII) produce moderate amounts of ^{56}Ni ($\sim 0.1M_{\odot}$) and remain opaque to γ-rays for a long time ($\sim$1 yr, i.e. until the radioactivity of ^{56}Co has dropped to low levels). Supernovae of another class, SNIa, are much brighter γ-ray line emitters. They are white dwarfs in binary systems, undergoing a thermonuclear explosion, which is triggered by the accretion of mass from their companion star; they produce $\sim$10 times more ^{56}Ni and become transparent to γ-rays much earlier than SNII, because of their smaller envelope mass. Unfortunately, observations show that SNIa are less frequent than SNII and their occurrence in a closeby galaxy is unexpected, on statistical grounds. The first SNIa to be (marginally) detected in the light of the ^{56}Co lines was SN1991T in the Virgo cluster of galaxies, at an estimated distance of 55 million light- years (by CGRO). Future γ-ray line observations of extragalactic SNIa will provide important information on the physics of those poorly understood objects.

Long-lived radioactive nuclei can emit characteristic γ-ray lines hundreds or thousands of years after the supernova explosion, while they are evolving in an environment transparent to γ-rays. The 1.156 line of ^{44}Ti (half-life 60 yr) has been detected by CGRO in the remnant of Cas-A, presumably a massive star that exploded about 300 yr ago, at a distance of 7000 light-years from the Earth. The ^{44}Ti emission may help to reveal remnants of recent supernova explosions that are unobserved in other wavelengths, because of obscuration by surrounding material (i.e. dense molecular clouds or layers of dust).

Finally, for tens of thousands of years after the explosion the SUPERNOVA REMNANT may radiate high-energy γ-rays, produced by accelerated electrons and protons interacting with the ambient medium. The Crab remnant, observed since the late 1960s, as well as five other sources detected by CGRO, belongs to this class of γ-ray emitters.

γ-ray pulsars

PULSARS are rapidly rotating neutron stars, first detected in the late 1960s through their periodic radioemission. More than 550 radio pulsars are known to day, with periods ranging from a few ms to several seconds. Before the launch of CGRO, only the Crab and VELA PULSARS were known to emit periodically in the γ-ray domain. Since then five more objects have been added to the list and all of them (except Geminga) emit also in the radiofrequencies (figure 2).

The fact that only a small fraction of the radio pulsars emits in γ-rays suggests that a special environment is required for γ-rays to be produced. The required energy is presumably provided by the rotational energy of the neutron star, which slows down in the process; however, the emission mechanism as well as the role of the magnetic field and of the pulsar environment are not well understood. A hint may lie in the fact that the younger pulsars are more active in the low-energy ($\sim$1 MeV) domain than the older ones, which have in general weaker magnetic fields and emit essentially in the $>$100 MeV range.

GEMINGA is the only known pulsar with no detectable radio emission. Many years after its detection in γ-rays (by SAS-2), it represented an enigma, since it had not a counterpart in any other wavelength. Its period ($\sim$0.237 s) was first measured in x-rays by the German satellite ROSAT and confirmed in γ-rays by CGRO. Its distance, evaluated in the early 1990s to $\sim$100 light-years, makes it the closest known neutron star.

Accreting compact objects

NEUTRON STARS and BLACK HOLES are the end products of the evolution of massive stars and they may be found in binary systems (as happens with about half of the stars we observe in the Galaxy). Matter from the companion star may then be accreted onto the compact object, forming around it an ACCRETION DISK. As matter spirals in, heat is generated at the expense of gravitational energy and the inner disk may reach high temperatures. According to theory, disks around neutron stars may never reach temperatures higher than $T \sim 10^8$ K (corresponding to thermal emission in the x-rays), whereas disks around black holes may become much hotter ($T \sim 10^9$ K) and radiate in the low-energy γ-ray domain, up to a few hundreds of keV. γ-ray observations would allow then neutron stars to be distinguished from stellar black holes.

Observations with SIGMA and CGRO allowed a few black hole candidates in our Galaxy to be identified, on the basis of their γ-ray signature. Some of the sources show steady emission, while others present a variable behaviour (presumably as a result of variation in the accretion rate or of instabilities developed in their accretion disk). A typical example of this class of transient sources (or γ-ray novae) is the source GRS 1124-68, alias Nova Muscae. Detected in January 1991, it has been observed in radio, x-rays and γ-rays, its luminosity in the last of these energy ranges exceeding that of the Crab (which is the strongest γ-ray source in the sky).

An extremely important manifestation of variable γ-ray activity is the excess emission detected in the 300–600 keV range from the source 1E1740.7-2942, located

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

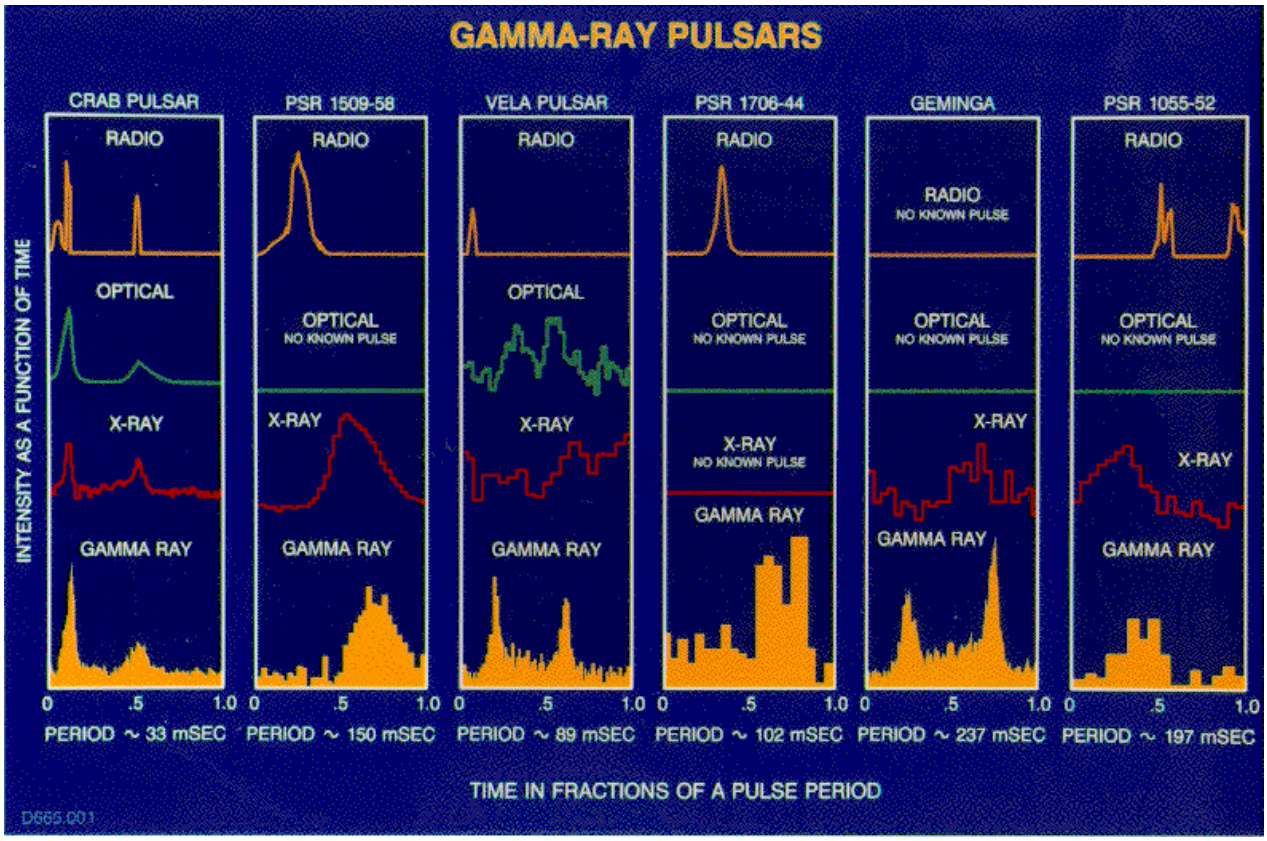


Figure 2. The folded light curves (intensity as a function of time) of various γ-ray pulsars at different energies (from radiowaves to γ-rays). A straight line means no known pulses. Explaining the variety of pulse shapes and energy dependences is a challenge to theorists.

in the galactic center region; observed for 2 days in October 1990 by SIGMA, this 'paroxysmic' activity has been interpreted in terms of electron–positron annihilation in the vicinity of a stellar mass black hole. Follow up radio-observations of the source revealed radio-emitting jets, presumably due to synchrotron emission of accelerated electrons. This behaviour (compact 'erupting' γ-ray source + radio jets moving at relativistic velocities) has also been observed in a few other sites in our Galaxy. These objects form the family of 'micro-quasars' (by analogy with the much more powerful and distant QUASARS) and represent ideal laboratories for the study of gravitational and relativistic astrophysics.

Diffuse γ-ray emission in the galaxy

Some of the 'products' of stellar activity may pervade the interstellar medium and give rise to diffuse γ-ray emission. This is the case of cosmic rays (energetic particles, accelerated by supernova explosions), long-lived radionuclei (decaying away from their production sites) and positrons.

Diffuse γ-ray emission from the Galaxy was first detected in 1968 by OSO-7 and galactic maps were provided by COS-B and CGRO. The emission correlates, in general, with features of the galaxy known from observations in other wavelengths: the molecular 'ring' at a galactocentric distance of 10 000–15 000 light-years, the spiral arms or a few molecular clouds, i.e. sites of rich gaseous content and active star formation. The analysis of the local γ-ray spectrum shows that at low energies (1–30 MeV) it results mainly from bremsstrahlung radiation of cosmic ray electrons; at higher energies, the emission of π^0s dominates. The LARGE MAGELLANIC CLOUD is the only other normal galaxy where γ-ray emission of this type has been detected.

During its million-year lifetime, radioactive ^{26}Al from more than 10 000 Wolf–Rayet stars and supernova explosions has been accumulated in the Galaxy's disk. Its characteristic 1.8 MeV line has been detected by HEAO-3 in 1984 and mapped by CGRO as diffuse emission along the galactic plane (figure 3). The HEAO-3 detection revealed that stars continue to enrich the Galaxy with their nucleosynthesis products. The galactic ^{26}Al map reveals, better than any other tracer, where the sites of current nucleosynthesis lie in the Milky Way.

Balloon-borne spectrometers detected in the 1970s the electron–positron annihilation line (at 0.511 MeV) from the Galactic center direction. For some time it was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

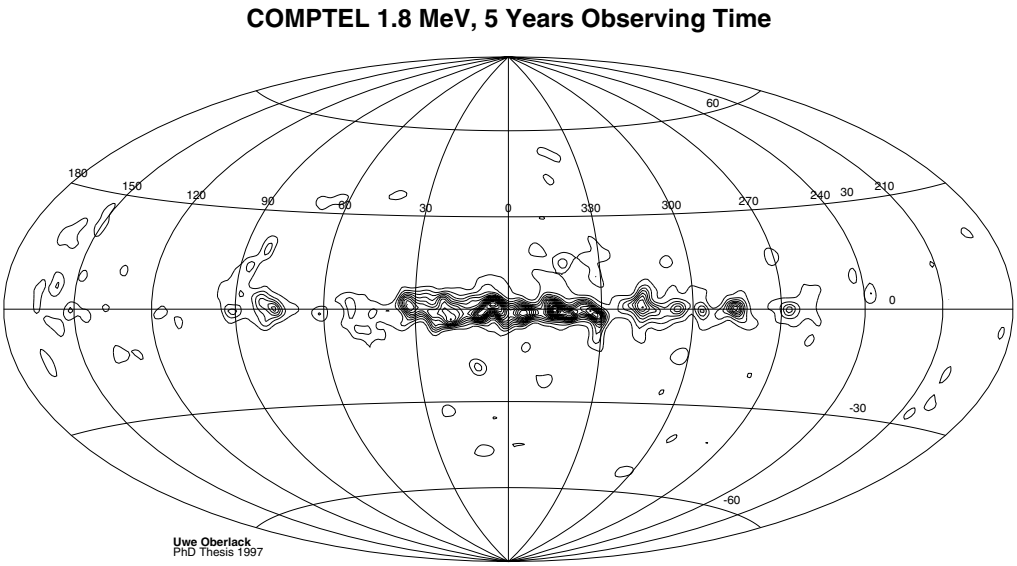


Figure 3. Map of the Galaxy in 1.8 MeV (^{26}Al line). Data from COMPTEL on CGRO.

thought that this emission was variable, but this variability was presumably due to the different properties of the instruments used. CGRO established the existence of diffuse (and steady) 0.511 MeV emission in the Milky Way, with ∼80% coming from the galactic bulge and the remaining ∼20% from the disk. Positrons produced by the β^+ decay of radionuclides, such as ^{26}Al, ^{56}Co and ^{44}Sc, are (most probably) at the origin of this galactic emission.

γ-ray bursts

Cosmic γ-ray bursts (GRBs), discovered in the late 1960s by the Vela satellites, are flashes of high-energy radiation which appear at random times from random directions, briefly dominating the γ-ray sky (from a few milliseconds to a few seconds) and then fading away. For more than a quarter of a century their nature remained a mystery: their non-thermal γ-ray emission presented no regularities at all, they did not seem to radiate in other wavelengths and it was impossible to evaluate their distances or to associate any known object with them.

The situation improved considerably in the 1990s. The BATSE instrument aboard CGRO detected about 1 burst per day for several years and established that their distribution in the sky is isotropic (figure 4). Such a distribution is naturally obtained from sources located at cosmological distances (several 10^8–10^9 light-years); in that case, their observed frequency implies about one burst per galaxy per million years, i.e. they represent events much rarer than supernovae. The Italian–Dutch x-ray satellite BeppoSAX allowed for an accurate location of GRBs, and in several cases relatively long-lived emission ('afterglow') at x-ray, optical and radio wavelengths was detected. In at least one case the afterglow optical spectra showed redshifted absorption lines due to the presence of remote intervening gas clouds, supporting the idea that the GRB is at a cosmological distance.

To be detectable at large distances, GRBs must radiate huge amounts of energy, about 10^{53} erg on average. Collisions involving neutron stars and/or black holes may produce such energies, mainly in the form of neutrinos, but the exact mechanism of the energy conversion to γ-rays is not elucidated yet. 'Fireball' models, involving material expanding at relativistic energies, seem the most promising at present.

Active galactic nuclei and blazars

The first detection of an extragalactic γ-ray source, the BRIGHT QUASAR 3C 273, was reported in 1978 by a team working with the COS-B satellite. Several other objects were subsequently detected in the low-energy γ-ray band (<1 MeV) and identified with the highly variable, moderately distant (redshift $z < 0.05$) ACTIVE GALACTIC NUCLEI (AGNs). In the 1990s, the EGRET instrument aboard CGRO detected high-energy γ-ray emission (>100 MeV) from several dozens of powerful and distant ($z > 2$) sources, a class of AGNs now known as BLAZARS; these objects radiate in their 'active' phase more energy in high-energy γ-rays (10^{45} erg s^{-1} on average) than in all other wavelengths.

The short time-scale of variability of AGNs (a few days to weeks for blazars), combined with their high γ-luminosities, suggests that accretion onto a massive black hole ($(10^6$–$10^9)M_\odot$) is the most probable power source. This idea could explain, in a unified framework, both the low-energy γ-emission of AGNs and the high-energy γ-emission of blazars. In the former case, the γ-rays are emitted from the hot inner regions of the accretion disk surrounding the black hole (as is the case with accreting stellar mass black holes in our Galaxy); in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

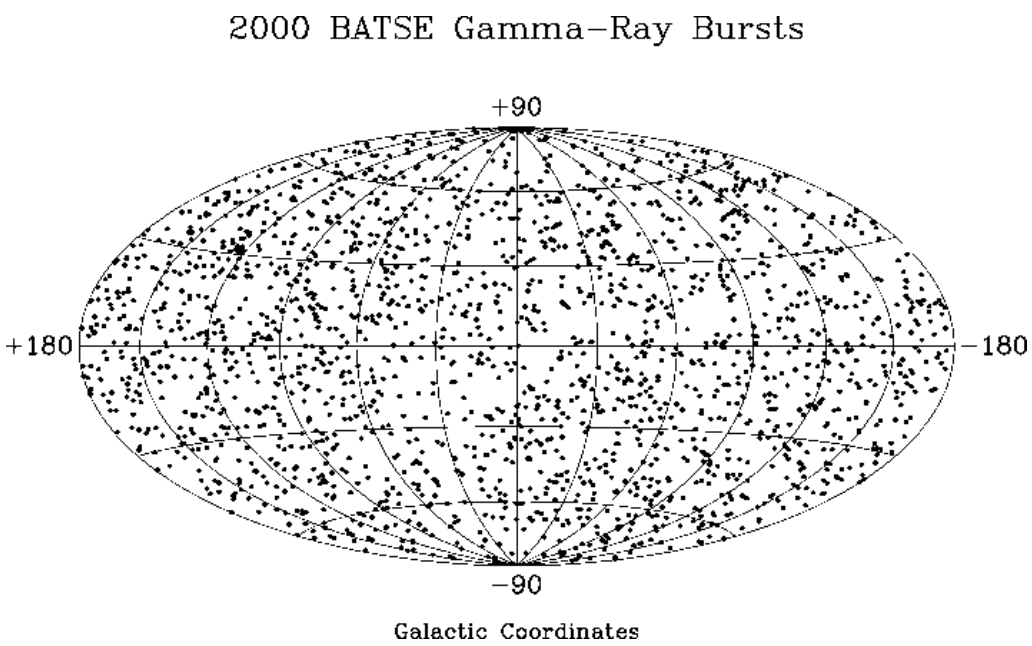


Figure 4. Isotropic distribution of GRBs in the sky.

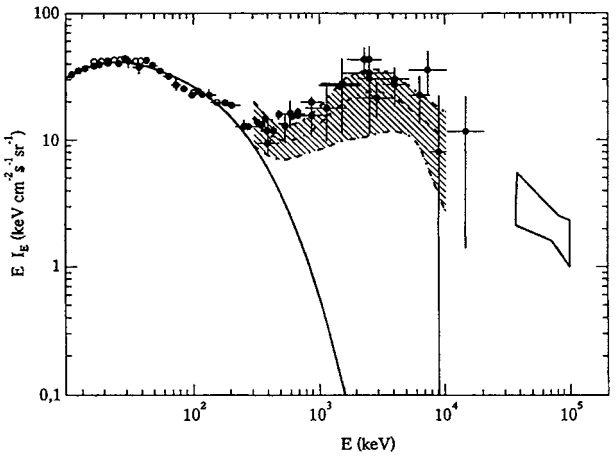


Figure 5. Diffuse extragalactic γ-ray background. Data compiled by Paul and Laurent (1998).

the latter case γ-rays are emitted by relativistic particles moving in diametrically opposite highly collimated jets (the intensity of the emission being enhanced along the direction of motion). This unified scheme seems able to explain not only the γ-ray properties of AGNs and blazars, but also their observed features in other wavelengths.

Diffuse extragalactic background

Since the early days of γ-ray astronomy, observations established the presence of a uniform, diffuse, isotropic γ-ray emission in the sky, most probably of extragalactic origin (figure 5). The most natural explanation for this diffuse γ-ray background involves the contribution of numerous extragalactic high-energy sources, unresolved by current instruments. In the low-energy range (<300 keV), the observed γ-ray background spectrum can be explained by the collective emission of Seyfert galaxies (a class of AGNs), while in the high-energy range (∼100 MeV) it is the collective emission of blazars that prevails; indeed, in both cases, the observed spectra of resolved individual sources compare favourably with the observed diffuse spectrum. Finally, the CGRO found no excess diffuse emission in the intermediate range of ∼1 MeV, contrary to all previous reports (obtained by instruments with underestimated and difficult to assess instrumental 'noise'). Despite that, this intermediate spectral region is not satisfactorily explained by the combination of Seyferts and blazars and another source seems to be required; the superposition of redshifted nuclear γ-ray lines from remote SNIa (mostly from the decay of ^{56}Co) seems the most promising candidate.

Bibliography

Most of the results of γ-ray astronomy in the 1990s can be found in the Proceedings of the CGRO and INTEGRAL symposia, held alternatively every two years.

A good recent overview (in French!) is:

Paul J and Laurent Ph 1998 *Astronomie Gamma Spatiale* (London: Gordon and Breach)

Some other selected references are as follows.

• The Sun:

Murphy R, Ramaty R, Kozlovsky B and Reames D 1991 *Astrophys. J.*, **371** 793

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

• Accreting compact objects:
Mandrou P *et al* 1994 *Astrophys. J. Suppl.* **92** 343
• GRBs:
Fishman G and Meagan C 1995 *Ann. Rev. Astron. Astrophys.* **33** 415
Meszaros P and Rees M 1993 *Astrophys. J.* **405** 278
• Radioactivity from stellar explosions:
Diehl R and Timmes F 1998 *Publ. Astron. Soc. Pacific* **110** 637
• Diffuse γ-ray line emission in the Galaxy:
Prantzos N and Diehl R 1996 *Phys. Rep.* **261** 1
• Diffuse γ-ray continuum emission in the Galaxy:
Strong A and Mattox J 1996 *Astron. & Astrophys.* **308** L21
• AGNs:
Dermer C 1995 *The Gamma-Ray Sky with CGRO and SIGMA* ed M Signore, P Salati and G Vedrenne (Dordrecht: Kluwer) p 39
• Diffuse extragalactic γ-ray background:
Zdziarski A 1996 *Mon. Not. Royal Astron. Soc.* **281** L9

Nikos Prantzos

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamma-Ray Bursts

The serendipitous discovery of cosmic gamma-ray bursts (GRBs) was announced in 1973, based on the data of the military Vela spacecraft, whose purpose was to monitor the limited nuclear test ban treaty. However, it was not until the late 1990s that a clear idea of their origin was obtained. A GRB is a short-duration pulse of gamma radiation (<20 keV to energies up to tens of GeV). With instruments of sufficient sensitivity, about one burst per day can be observed. GRB durations range from 10 ms to 1000 s or more, but an average is about 10 s. For this short time, they may be the brightest objects in the gamma-ray sky. An example is shown in figure 1. Each burst comes from a unique, random direction; no two events have ever been detected from exactly the same direction. The distribution of burst sources on the sky is completely isotropic (figure 2). These properties, which are quite unlike those of typical gamma-ray sources (see gamma-ray astronomy) explain, in part, why progress in understanding the origin of bursts has been slow: Gamma-ray telescopes tend to have small fields of view and provide relatively coarse location information, making it impractical to monitor the entire sky and provide precise directional information with a single detector. The main detection techniques to date have been interplanetary networks and arrays of large-area detectors such as the Burst and Transient Source Experiment aboard the Compton Gamma-Ray Observatory. However, the inherent delays in the former, and the relatively coarse localizations provided by the latter, hampered efforts to detect burst sources at other wavelengths.

A breakthrough was achieved in 1997 by the Italian–Dutch BeppoSAX spacecraft, which carried both a wide-field x-ray camera and narrow-field x-ray telescopes. The wide-field camera made it possible to detect and locate bursts in the 2–10 keV energy range with a precision of several arcminutes. The spacecraft could then be slewed within hours to view the position of the burst source with the narrow-field telescopes. It was discovered that, long after the burst had disappeared in gamma-rays, a fading x-ray afterglow was present (figure 3). Rapid optical observations reveal similar afterglows at visible wavelengths, and radio observations show that slowly fading radio afterglows are also present. Deeper optical observations usually result in the detection of a faint host galaxy at the site of the burst with a typical redshift of 1 (implying distances of billions of light-years); it is often the case that the optical afterglow outshines its host galaxy for at least several days. Figure 4 shows an example of an optical afterglow and its host galaxy. These observations have established that the distance scale of cosmic GRBs is cosmological. This explains their perfectly isotropic sky distribution, which had been a controversial clue to their origin for many years. About two dozen bursts now have well-established redshift measurements (the largest being z=4.5).

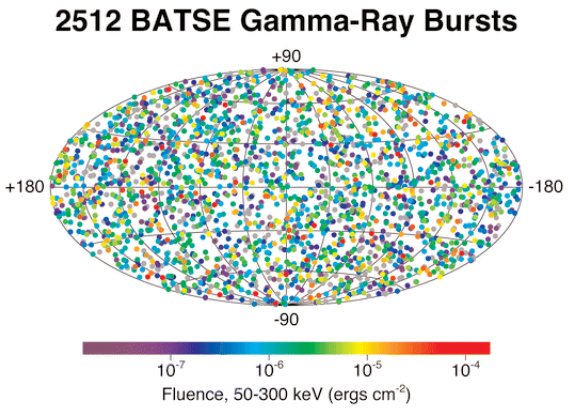


Figure 2. Locations of gamma-ray bursts detected by the Burst and Transient Source Experiment aboard the Compton Gamma-Ray Observatory. A galactic coordinate system has been used, and the bursts have been color coded according to their intensities. The precisions of the locations are typically several degrees. From www.batse.msfc.nasa.gov/batse/.

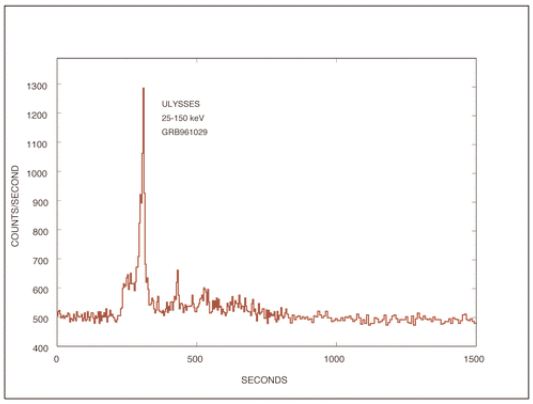


Figure 1. The time history of a particularly long, intense gamma-ray burst on 29 October 1996, as measured by the GRB detector aboard the Ulysses spacecraft.

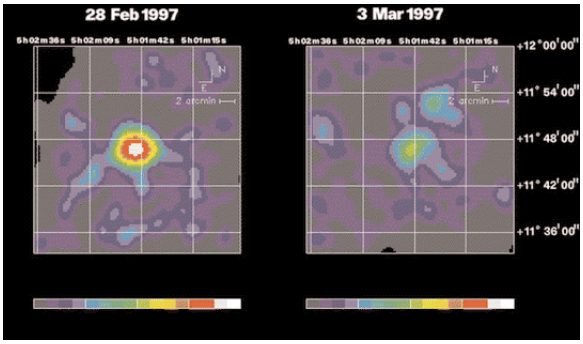


Figure 3. The first x-ray afterglow. The images show a fading 2–10 keV x-ray source, detected by the Narrow Field Instruments of the BeppoSAX spacecraft. From www.asdc.asi.it/bepposax/.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

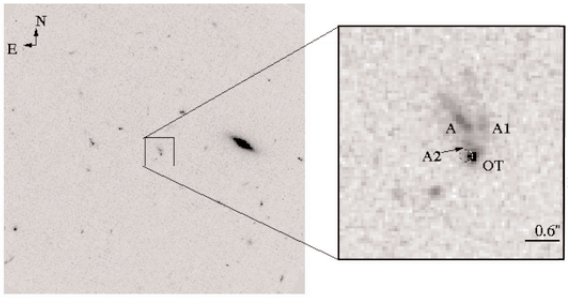


Figure 4. Hubble Space Telescope images of GRB990123. OT is the optical transient associated with the burst, A is the galaxy and A1 is a bright knot of emission associated with the galaxy. From www.astro.caltech.edu/~jsb/GRB/grb990123/index.html.

If GRB sources emitted their radiation isotropically, their total energy output would be well above 1053 erg and it would be difficult to find viable sources for the energy. However, it seems more likely, both from a theoretical and from an observational viewpoint, that at least the gamma radiation is beamed into a relatively small cone, reducing the energy output by some two orders of magnitude. Two possible energy sources are collapsars (also known as hypernovae) and the merger of two compact objects in a binary system (such as two neutron stars). Observational evidence suggests that the former phenomenon is the one most likely to be responsible for the bursts whose optical counterparts and host galaxies have been studied.

In the collapsar or hypernova scenario, a massive star (approximately 30 solar masses) exhausts its nuclear fuel supply and collapses, forming a black hole surrounded by an accretion disk. Matter from the disk falls into the black hole and an intense magnetic field helps to extract energy from the system. Shells of matter are driven outwards at velocities very close to the speed of light. Collisions between these shells form shock waves which accelerate electrons to high energies and the electrons radiate the gamma-rays which constitute the GRB. As the shells continue to move outward, they eventually encounter a denser medium (either the interstellar medium or possibly a region filled with matter which was ejected from the progenitor star in earlier phases). Here they again form shock waves which accelerate electrons, and these electrons radiate the longer-lived radio, optical, and x-ray afterglows. This model is often referred to as the 'standard fireball model'.

The light curve of the optical afterglow decays as a power law with time. However, in several cases, there is clear evidence for a temporary departure from a power law (a 'bump' in the light curve). In most of these cases, a possible explanation is the emergence of a supernova-like component in the light curve, which is initially obscured by the intense optical light immediately following the burst. (In fact, in one instance, a gamma-ray burst was found to be spatially and temporally consistent with a known supernova, 1998bw.) Such a component might be expected in the hypernova scenario. In one case, however, an alternative explanation is gravitational lensing of the light by a foreground galaxy.

For all the bursts whose optical afterglows and host galaxies have been detected, the burst source lies well within the galaxy, and it often appears to be located within the galaxy's star-forming regions. The galaxies themselves appear to be normal galaxies, indistinguishable from their neighbors at similar distances; there is no indication that any are active galaxies.

For almost all the bursts with x-ray follow-up measurements, fading x-ray counterparts have been detected. However, only about 50% of the bursts display radio or optical counterparts, despite sensitive searches. The reasons for this are not well understood, but possible causes include heavily obscured sources, very high redshift sources and sources which are outside of their host galaxies. The last explanation might be appropriate for the merging compact object model; these binary systems can be ejected from their host galaxies long before they merge and, lacking a surrounding interstellar medium, might not generate the shocks needed to produce long wavelength afterglows.

Although some of the basic physics of GRBs is now understood, the field is in many respects in its infancy. Many more observations are needed to sort out the following questions (among others): 1. Can GRBs be used to trace star formation in the early universe? 2. Can GRBs be detected at redshifts corresponding to the reionization epoch? 3. What is the relation between GRBs and supernovae? 4. Why do some bursts display no radio or optical afterglows? A major GRB mission, Swift, will address these and other questions. Swift is a NASA Midex mission scheduled for launch in October 2003. It comprises three instruments: a burst alert telescope, an x-ray telescope and a UV/optical telescope. The burst alert telescope will survey a wide field for GRBs, which it will localize with several arcminute accuracy. The spacecraft will then slew to the position of the burst to observe it with its x-ray and UV/optical cameras, providing information on the long-wavelength afterglows and on the source redshift. Swift is expected to detect over 100 bursts per year and last for a minimum of 2 years.

Web update (30 April 2003)

On December 11, 2002, scientists obtained an image of an unusual type of GRB event one minute after the explosion. They captured a particularly fast-fading type of dark burst, which comprises about half of all GRBs. These dark bursts are so named because until this event, they have had no detectable optical afterglow. NASA's High Energy Transient Explorer (HETE) discovered this burst on December 11, 2002. The burst originated six billion light-years away, and HETE relayed its location to observatories worldwide in 22 seconds. The ground-based RAPTOR (RAPid Telescopes for Optical Response) optical telescope, operated by the Los Alamos National Laboratory in New Mexico, was the first on the scene,

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

observing the afterglow at 65 seconds. Other telescopes observed the event in the minutes that followed.

Bibliography

The discovery paper is

Klebesadel R, Strong I and Olson R 1973 Observations of gamma ray bursts of cosmic origin *Astrophysical Letters* **182** L85

The two 'breakthrough' papers are

Costa E, Frontera F, Heise J *et al.* 1997 Discovery of an x-ray afterglow associated with the gamma-ray burst of 28 February 1997 *Nature* **387**: 783

van Paradijs J, Groot P, Galama T *et al.* 1997 Transient optical emission from the error box of the gamma-ray burst of 28 February 1997 *Nature* **386**: 686

Progress over the years has been reported at a series of biennial conferences, whose proceedings have been published in *AIP Conference Proceedings* (265, 307, 384, 428 and 526), in *Astrophysics and Space Science* 231 (1/2) (1995) and in the *ESO Astrophysics Symposia Series 'Gamma Ray Bursts in the Afterglow Era'* (2001)

Kevin Hurley

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Gamma-Ray Telescopes

Gamma-rays are the highest-energy photons in the ELECTROMAGNETIC SPECTRUM and their detection presents unique challenges. On one hand it is easy to detect γ-rays. The interaction cross-sections are large and above a few MeV the pair production interaction, the dominant γ-ray interaction with matter, is easily recognized. Gamma-ray detectors were far advanced when the concept of 'γ-ray astronomy' was first raised in Phillip Morrison's seminal paper in 1958. Indeed it was the expected ease of detection and the early promise of strong sources that led to the large concentration of effort in the field, even before the development of X-RAY ASTRONOMY. Today the number of known γ-ray sources is well under a few hundred whereas there are hundreds of thousands of x-ray sources. What went wrong?

The answer is simple: the detection of cosmic γ-rays was not as easy as expected and the early predictions of fluxes were hopelessly optimistic. Below we consider the physical phenomenon whereby γ-rays interact with matter, the basic detector configurations and the detectors which have provided the bulk of the observational results; we conclude with a description of the techniques now under development.

First we provide a few definitions and a few caveats. The term 'γ-ray' is a generic one and is used to describe photons of energy from 100 keV (10^5 eV) to >100 EeV (10^{20} eV). A range of 15 decades is more than all the rest of the known electromagnetic spectrum. A wide variety of detection techniques is therefore necessary to cover this huge range. We will concentrate on the telescopes in the somewhat restricted range from 1 MeV to 100 TeV. The lower definition comes from the knowledge that the lower energies are better covered by the techniques used to cover hard x-ray astronomy and that many authors would define the γ-ray regime as starting at 0.51 MeV (the rest mass energy of the electron). There are no credible detections of γ-rays at energies much beyond 50 TeV and the 'γ-ray telescope' techniques used beyond these energies are really the same as those used to study charged COSMIC RAYS and hence are best discussed under that heading. We are left with some eight decades which we will, somewhat arbitrarily, define as the medium-energy (ME) range from 1 to 30 MeV, the high-energy (HE) range from 30 MeV to 100 GeV and the very-high-energy (VHE) range from 100 GeV to 100 TeV. These ranges are not defined by the physics of their production but by the interaction phenomena and techniques employed in their detection. Thus Compton scattering dominates the ME range necessitating the use of satellite-borne Compton telescopes. The HE and VHE ranges use the pair production interaction but in very different ways; HE telescopes identify the electron pair in balloon- or satellite-borne detectors, whereas VHE detectors detect the electromagnetic cascade that develops in the Earth's atmosphere.

GAMMA-RAY ASTRONOMY is still an observation-dominated discipline and the observations have been driven not so much by the astrophysical expectations (which have often been wrong) as by the experimental techniques, which have permitted significant advances to be made in particular energy ranges. Hence the most fruitful observations have come at energies of 100 MeV; these were originally inspired by the prediction of the strong bump in the spectra expected from the decay of π^0s that are created in hadron interactions. This energy region was exploited primarily because the detection techniques were simpler and more sensitive. In contrast the ME region has the potential for very interesting astrophysics with the predicted existence of nuclear emission lines but the development of the field has been slow because the techniques are so difficult.

Peculiarities of γ-ray telescopes

There are several peculiarities that uniquely pertain to astronomy in the γ-ray energy regime. These factors make γ-ray astronomy particularly difficult and have resulted in the slow development of the discipline.

In every band of the electromagnetic spectrum astronomical telescopes make use of the fact that the cosmic rain of photons can be concentrated by reflection or refraction, so that the dimensions of the actual photon detector are a small fraction of the telescope aperture. How limited would have been our early knowledge of the universe if the optical astronomer had not been aided by the simple refracting telescope which so increased the sensitivity of the human eye. The radio astronomer, the infrared astronomer, even the x-ray astronomer, depends on the ability of a solid surface to reflect and, with suitable geometry, to concentrate the photon beam.

Above a few MeV there is no efficient way of reflecting γ-rays and hence the dimensions of the γ-ray detector are effectively the dimensions of the γ-ray telescope. (As we shall see in a later section this is not the case for ground-based VHE telescopes.) In practice to identify the γ-ray events from the charged particle background it is necessary to use detectors whose efficiency is often quite low. Hence at any energy the effective aperture of a space-borne γ-ray telescope is seldom greater than 1 m^2 and often only a few cm^2, even though the physical size is much larger. For instance the COMPTON GAMMA RAY OBSERVATORY (CGRO) was one of the largest and heaviest scientific satellites ever launched; however, its ME and HE telescopes had effective apertures of 1600 cm^2 and 5 cm^2 respectively. Beam concentration is particularly important when the background scales with detector area. This is the case with γ-ray detectors which must operate in an environment dominated by charged cosmic rays.

The problem of a small aperture is compounded by the fact that the flux of cosmic γ-rays is always small. At energies of 100 MeV the strongest source (the Vela pulsar) gives a flux of only 1 photon min^{-1}. With weak sources long exposures are necessary and one is still dealing with the statistics of small numbers. It is small wonder that γ-ray astronomers have been frequent pioneers in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the development of statistical methods and that γ-ray conferences are often dominated by arguments over real statistical significance.

As it is to photons in many bands of the electromagnetic spectrum the Earth's atmosphere is opaque to all γ-rays. The radiation length is 37.1 g cm^{-2} and the total vertical thickness at sea level is 1030 g cm^{-2}. Even the highest mountain is many radiation lengths below the top of the atmosphere so that it is virtually impossible to consider the direct detection of cosmic γ-rays without the use of a space platform. Large balloons can carry the bulky detectors to near the top of the atmosphere and much of the pioneering work in the field was done in this way. However, the charged cosmic rays constitute a significant background and limit the sensitivity of such measurements.

The background can take many forms. In deep space it is the primary cosmic radiation itself, mostly protons, heavier nuclei and electrons. This background can be accentuated by secondary interactions in the spacecraft itself. Careful design and shielding can reduce this effect, as can active anti-coincidence shields. However, at low energies induced radioactivity in the detector and its surrounds can be a serious problem. In balloons the secondary cosmic radiation from the cosmic ray interactions above the detector seriously limit the sensitivity and were the initial reason for the slow development of the field. Huge balloons that carry the telescopes to within a few grams of residual atmosphere are a partial solution, but it is still impossible to trust the measurement of absolute diffuse fluxes.

Physics of γ-ray interactions

In the energy region above 1 keV there are basically three processes by which the γ-rays can interact with the matter in the detector: these are the photoelectric effect, Compton scattering and pair production. The relative cross-sections, or, more practically, the mass absorption coefficients peak in different energy ranges. The cross-section has the same functional form for all materials, although the actual values and relative strengths of the three processes may differ somewhat. As the photoelectric effect, the interaction of the γ-ray with bound electrons in atoms, is only important at low energies (<1 MeV), we shall not consider it here.

Compton scattering

This is the most complex of the three interactions, and hence detectors which depend on it are the most complicated and the most difficult to interpret. It is important primarily in the energy range from 100 keV to 30 MeV and hence γ-ray astronomy in this range has been the slowest to develop.

In Compton scattering the electron is unbound and it alone must take up the energy and momentum of the γ-ray. In practice the γ-ray may undergo one or more Compton scatterings, losing energy to the electron in each case, until eventually it may undergo a photoelectric reaction. In each scattering the electron will take up some of the energy as kinetic energy and the γ-ray will change direction.

The relationship between the incident energy of the γ-ray, E, and the resultant energy, E', where both are expressed in units of E_e, the rest mass of the electron, is given by

$$E' = E/(1 + E[1 - \cos\theta)]$$

where θ is the scattering angle of the γ-ray. The maximum energy loss comes for backward scattering at which the scattered energy is

$$E' = E/(1 + 2E).$$

For $E > 0.51$ MeV, E' is close to 0.5 (or 0.255 MeV) which is the typical energy loss per scattering, independent of E.

At low energies the cross-section is given by the Thomson cross-section

$$\sigma_c = \sigma_T = (8\pi/3)e^2/m_e c^2 = 0.665 \text{ barns}$$

and at high energies

$$\sigma_c = (3/8)E(\ln 2E + 1/2)\sigma_T.$$

Hence Compton scattering is almost independent of the incident energy so that the cross-section decreases only slowly with energy.

Pair production

In the pair production interaction, the γ-ray is completely annihilated with its energy transferred to an electron–positron pair which is created, i.e.

$$h\nu \rightarrow e^+ + e^-.$$

The interaction takes place in the electric field of a nucleus which takes up some of the momentum. Obviously the threshold for the interaction must be >1.02 MeV (i.e. $2m_e$). The interaction can also occur in the field of an electron, but the cross-section is much less and the threshold is higher. The energy of the γ-ray is taken up by the electron–positron pair as rest mass and kinetic energy. The positron will generally annihilate with an electron to produce two γ-rays, which can Compton scatter or suffer photoelectric absorption. These secondary products at moderate energies can be totally absorbed in the detector; by measuring their energy plus the rest mass of the pair, the energy of the γ-ray is estimated.

The cross-section for pair production rises rapidly and becomes dominant above energies of about 30 MeV, after which it rises slowly to an asymptotic value. This value is given by

$$\sigma_{pp} = \sigma_0 Z^2[(28/9)\log(183/Z^{1/3}) - 2/27] \text{ cm}^2/\text{atom}$$

where $\sigma_0 = (1/137)e^4/m_e^2c^4$.

The mean distance that a γ-ray travels before it undergoes pair production is given by

$$\lambda_{pp} = 1/N\sigma_{pp}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where N is the number of target nuclei per unit volume.

The mean free path for pair production is related to the radiation length, X_0:

$$\lambda_{pp} = (9/7)X_0.$$

A critical factor is the degree to which the electron–positron pair retain the trajectory of the γ-ray. The root mean square angle between the trajectory of the secondary electron of energy E_e and that of the primary γ-ray of energy E_γ is about 4° at $E_\gamma = 30$ MeV, 1.5° at $E_\gamma = 100$ MeV and 0.2° at $E_\gamma = 1$ GeV. The value of observations at high energies is thus apparent.

Unfortunately it is not possible to measure the electron trajectory precisely as it will inevitably undergo Coulomb scattering as it passes through the material of the detector.

Measuring the energy of the γ-ray is essentially measuring the energy of the electron–positron pair and their secondary products. This requires that the detector have sufficient mass to absorb all these products. In practice this usually calls for a calorimeter with as much absorber as the spacecraft can carry.

ME and HE telescopes

Compton telescopes

The basic detector of γ-rays in the difficult 100 keV–10 MeV energy range is the scintillation detector, which consists of a solid or liquid material, in which light is produced by charged secondary particles resulting from the photoelectric or Compton scattering γ-ray interaction, and a phototube, in which light is converted into an electrical signal. A common material is thallium-activated sodium iodide, NaI(Tl). Charged particles are rejected by surrounding the detector by another plastic scintillator detector. If the outer detector is shaped like a well with small opening then it can serve as a collimator with crude angular resolution. Many of the early detectors worked on this principle.

A more sophisticated detector is the Compton telescope in which two detectors are operated in series (figure 1). In the front detector a primary γ-ray, which Compton scatters in the forward downward direction, is selected (based on the energy registered by the first detector from the recoil electron); the γ-ray is then absorbed in another Compton scatter in the lower detector. The lower detector is surrounded by an anti-coincidence scintillator to veto charged particles coming up from below. Of necessity because of the wide range of angles that the scattering may have, the efficiency of these simple detectors is poor, typically less than 1%. However, the energy and angular resolutions are improved over the simple one-stage detector.

The upper detector should have a large cross-section for Compton scattering over the desired energy range. In the 1–10 MeV region the best material is one with low Z; hence the detector should be a relatively thin liquid or plastic scintillator so that a single Compton scattering occurs with good efficiency. The second, lower, detector should totally absorb the product of the second Compton scatter and hence should be thicker and composed of high-Z material.

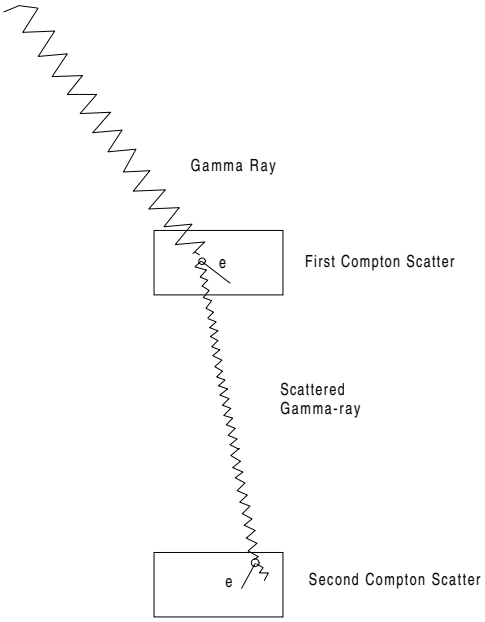


Figure 1. Schematic of double Compton scatter. In the upper scintillator the primary γ-ray Compton scatters in the downward direction; the γ-ray is then absorbed in another Compton scatter in the lower detector. In each case the energy of the recoil electron is measured and thence the energy and arrival direction of the primary are determined.

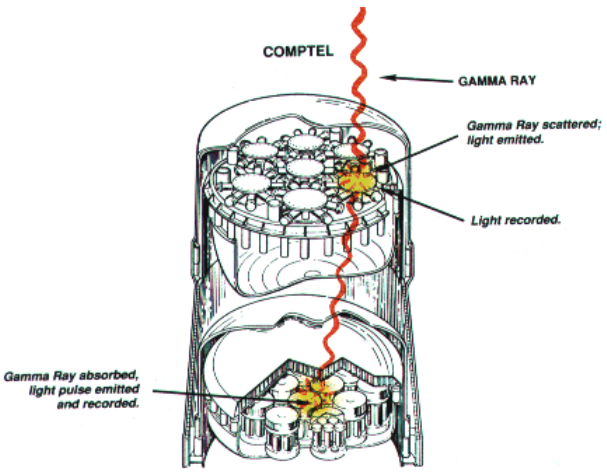


Figure 2. An example of a Compton telescope: COMPTEL. It is sensitive from 1 to 30 MeV. The angular resolution is 3–5°.

A double Compton scattering is also the basic principle used in the most sophisticated 1–30 MeV telescope flown to date, COMPTEL on the CGRO (figure 2). The primary γ-ray incident within ±40° of the telescope axis first Compton scatters in the upper detector which is a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

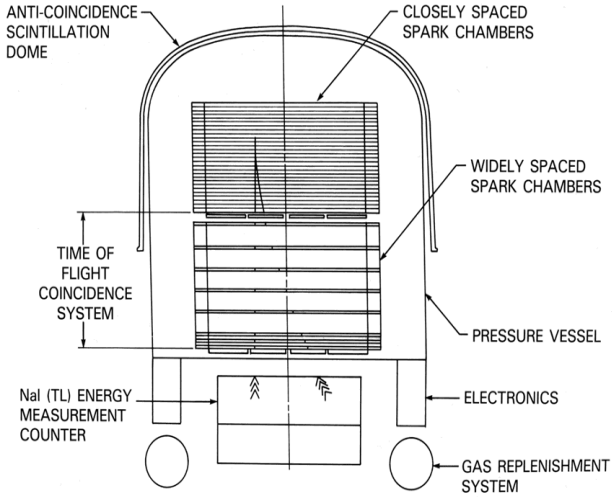


Figure 3. Example of a spark chamber telescope: EGRET. The telescope is sensitive from 30 MeV to 30 GeV. The field of view $\pm 20^\circ$ and the energy resolution is about 20%.

low-Z liquid scintillator; the second scattering takes place in the lower detector which is a high-Z NaI(Tl) scintillator. Each detector actually consists of seven modules and the separation between the two layers is 1.5 m. Hence time of flight can be used to discriminate against upward-going particles. In addition, all of the detectors are surrounded by thin plastic anti-coincidence scintillators which respond to charged particles. If the energy deposited in the upper and lower modules is measured, then the direction of the incident γ-ray can be determined to be within a narrow ring on the sky and its energy estimated to about 5–10%. A source then is apparent as the locus of intersections of a number of such rings.

Pair production telescopes

The spark chamber, long obsolete for HE physics experiments, has been the workhorse detector for γ-ray astronomy in the energy range 30 MeV–10 GeV from the early 1960s through the end of the century. The three experiments, which provided almost all the results during this period, all used the spark chamber as their principal detector. These were the USA's SAS-II (1973), Europe's COS-B (1975–82) and the joint European– USA EGRET on the Compton Gamma Ray Observatory (1991–).

Although the basic principles of the pair production telescope are simple, the detailed design is complex and accounts for the fact that the effective collection area is far smaller than the geometrical cross-section of the telescope. This is illustrated by EGRET, the pair production telescope on the CGRO. As with most pair production spark chamber telescopes, this consists of four distinct components which are shown schematically in figure 3.

(i) A spark chamber usually consists of a series of parallel metal plates in a closed container; the alternate plates are connected together electrically with one set permanently connected to ground. On an indication that a charged particle has passed through the chamber, a high voltage is applied to the second set of plates. The chamber contains a gas at a pressure such that the ionization left behind by the passage of the charged particle causes an electric spark discharge between the plates. The gas is a mixture of neon and ethane. An electron–positron pair created by a γ-ray interaction in one of the plates is then readily apparent as a pair of sets of sparks that delineate the path of the electron and positron. In practice the tracks are disjointed as the electrons suffer multiple scattering within the plates of the chamber. This limits the thickness of the plates (which should be as thick as possible to ensure that the γ-rays interact effectively, but not so thick that the electrons undergo excessive Coulomb scattering in the plate material). Multiple plates ensure that the tracks are effectively mapped. The collection area and angular resolution of the telescope are determined by the spark chamber geometry. In EGRET the spark chamber 'plates' consist of 28 wire grids interleaved with interaction plates of 0.02 radiation lengths thickness in which the γ-ray interacts. Each wire is threaded through a magnetic core memory, which is read out and reset after each event.

(ii) At least one electron must emerge from the spark chamber to ensure that it initiates a trigger that causes the application of the high-voltage pulse to the second set of plates to activate the spark chamber. A permanent high-voltage difference cannot be maintained between the plates, as the spark discharges will take place spontaneously. EGRET is triggered by a coincidence between two thin sheets of plastic scintillator with a 60 cm separation (sufficient to recognize and reject upward-going charged particles). It is the need for this trigger which limits the lower-energy threshold of the spark chamber telescope. The trigger detection system effectively defines the field of view of the telescope.

(iii) The electrons must be completely absorbed if their energy is to be measured; to achieve this there must be a calorimeter that is some radiation lengths thick. In EGRET, as in most spark chamber telescopes, this is an NaI(Tl) crystal, whose sole function is to measure the total energy deposited. At the low end of the sensitivity range the energy of the electrons can also be determined by the amount of Coulomb scattering in the plates of the spark chamber.

(iv) Finally the entire assembly is surrounded by an anti-coincidence detector which signals the arrival of a charged particle, but which has a small interaction cross-section for γ-rays. This consists of a very thin outer shell of plastic scintillator viewed by photomultipliers.

EGRET is the largest and most sensitive HE γ-ray telescope flown to date; it is the flagship instrument on

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CGRO. Approximately the size of a compact car and with a total mass of 1900 kg, the telescope has an effective collection area of 1600 cm^2. The characteristics of the telescope are listed in table 1.

The telescope was designed for a 5 yr lifetime. The gas which fills the chamber gradually becomes poisoned and must be replenished. It was expected that a filling would last 1 yr. Hence only four gas canisters were attached to the instrument for replenishment at yearly intervals. In practice the unprecedented and unexpected success of CGRO meant that the mission was extended as were the replenishment intervals so that, for a considerable fraction of the 9 yr lifetime of the mission, EGRET operated at less than optimum efficiency.

VHE telescopes

Air Čerenkov telescopes

When an HE γ-ray strikes the upper atmosphere, it produces an electron–positron pair (as it does in a spark chamber). However, if the energy of the γ-ray, and hence of the electron–positron pair, is large enough, an electromagnetic cascade will result which will continue down through the atmosphere with secondary γ-ray and electron production by bremsstrahlung and pair production. The cascade will continue along the axis of the trajectory of the original γ-ray and the total energy of the secondary particles will be a good representation of its energy.

For γ-rays of energy 100 TeV and above, sufficient particles can reach ground level for the shower to be detected by arrays of particle detectors spread over areas of 0.1 km^2. As the secondary particles all move at nearly the speed of light and retain the original trajectory of the primary γ-ray, the shower front arrives as a disk which is only a meter thick. Differential timing between the detectors can then determine the arrival direction and hence the source of the γ radiation.

At lower energies the cascade will die out as the average energy of the secondary particles drops to the point that ionization losses become the major loss process (figure 4). For a primary γ-ray of energy 1 TeV, few secondary particles will reach even mountain altitude. However, as the relativistic particles traverse the atmosphere, they excite the atmosphere to radiate Čerenkov light with high efficiency. Although the fraction of energy that goes into this mode is small (less than 10^{-6} of the primary energy), it permits a very easy way to detect the cascade and thence the γ-ray. A simple light detector (mirror, plus phototube, plus fast pulse counting electronics) provides an easy way of detecting the cascade. Early telescopes consisted of ex-World War II searchlight mirrors with phototubes at their foci, coupled to fast pulse counting electronics.

The observations are best made from a dark mountain top observatory. Since the Čerenkov angle in air is about 1° and the amount of light is proportional to the number of particles in the cascade (and hence to the energy of the γ-ray), the measurement of the atmospheric Čerenkov

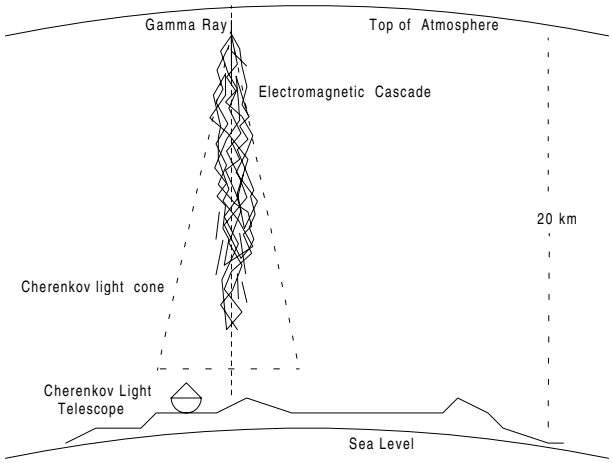


Figure 4. Schematic of atmospheric air shower detection. The γ-ray interacts with an air molecule high in the atmosphere and produces an electron–positron pair. These electrons subsequently interact and produce an electromagnetic cascade. The passage of relativistic particles through the atmosphere causes the emission of Čerenkov light which reaches ground level as a disk of diameter 200 m and thickness 1 m. The light can easily be detected by optical Čerenkov light telescopes (not to scale).

component provides a good measure of the energy and arrival direction of the γ-ray. Because the light spreads out as it traverses the atmosphere, the collection area for γ-ray detection is as large as the lateral dimensions of the light pool at detector altitude; this can be as much as 50 000 m^2.

This is one of the few astronomical techniques in which the Earth's atmosphere plays an essential positive role. However, the technique has its drawbacks. Although the atmosphere comes cheap (and the gas does not need to be replenished), the observer has no control over it; the telescope is wide open to the elements and the detector is susceptible to a troublesome background of light from Sun, Moon and stars, from the airglow, from lightning and meteors, and from a variety of man-made light sources, from satellites and airplanes to airport beacons and city lights. These limit the sensitivity for γ-ray source detection. However, the most troublesome background is that from air showers generated by charged cosmic rays of similar energy to the γ-rays under study. These air showers are cascades of particles which are superficially very similar to a pure electromagnetic cascade but contain penetrating particles (mostly muons) which reach ground level (see COSMIC RAYS: EXTENSIVE AIR SHOWERS). These showers are thousands of times more numerous and their light flashes are difficult to distinguish from those from γ-rays. Because of interstellar magnetic fields, the arrival directions of the charged cosmic rays are isotropic; hence a discrete source of γ-rays can stand out only as an anisotropy in an otherwise isotropic distribution of air showers. Unfortunately a γ-ray source would have to be very strong (a few per cent of the cosmic radiation) to be detectable in this way.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

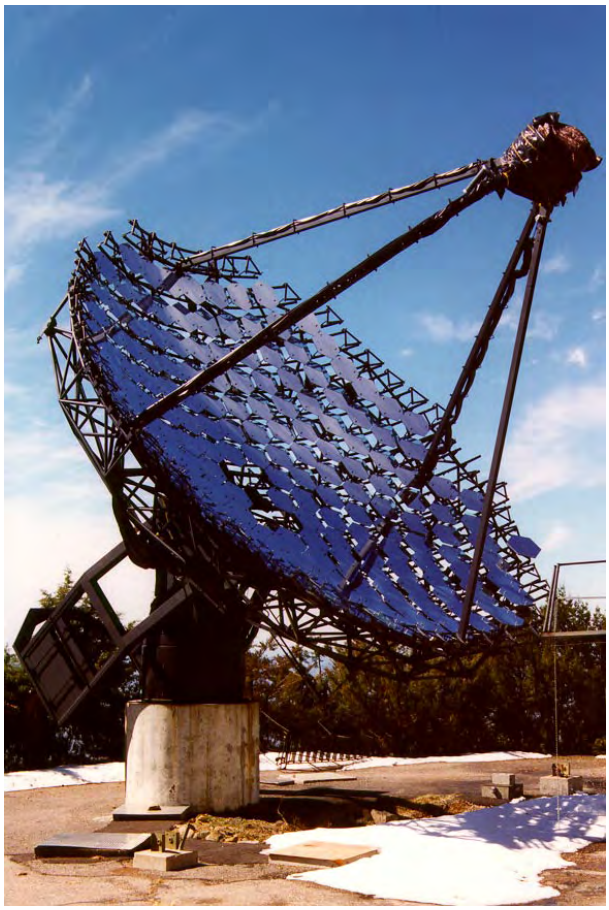


Figure 5. The Whipple 10 m γ-ray telescope. Note the '10 m' refers only to the aperture of the optical reflector; the effective collection area is >50 000 m^2 so that the γ-ray 'aperture' is 120 m.

Imaging detectors

Early attempts to discriminate the electromagnetic showers initiated by γ-rays from air showers initiated by charged particles were unsuccessful using either the ground-level arrays of particles detectors or atmospheric Čerenkov detectors. The development of the Čerenkov imaging technique gave the first effective discrimination; an array of photomultipliers in the focal plane of a large optical reflector was used to record a Čerenkov light picture of each air shower. Monte Carlo simulations of the development of air showers from photon and hadron primaries predicted that the images of the former would have somewhat smaller angular dimensions and thus could be identified. The largest optical reflector built for γ-ray astronomy is the Whipple Observatory 10 m optical reflector (built on Mount Hopkins in southern Arizona in 1968) (figure 5); in 1984 this was equipped with a photomultiplier camera with 37 pixels which was used to detect the CRAB NEBULA. This first detection led to a rapid development of the imaging technique, with significant improvements in flux sensitivity.

There are currently nine observatories around the globe using variants of this technique and the detection of some 13 sources has been claimed, both in the Galaxy and beyond. Background rejection of cosmic rays is now in excess of 99.7%, and the technique is effective from energies of 250 GeV–50 TeV. The energy resolution is typically 30–40%. A signal with significance of (5–10)σ can be detected from the Crab Nebula in just an hour of observation. Because of the very large collection area associated with the technique (>50 000 m^2), it is particularly powerful for the detection of short transients in TeV γ-ray sources (see also VERY-HIGH-ENERGY GAMMA-RAY SOURCES). In fact the shortest flare ever seen from a known γ-ray source at energies greater than 1 MeV was recorded from the blazar, Markarian 421; the duration of the TeV flare was 15–20 min. In some cases the telescope is an array of small reflectors operated in a stereo mode, e.g. the German–Spanish HEGRA experiment in the Canary Islands.

Table 1. Comparison of EGRET and GLAST.

Parameter	Units	EGRET (achieved)	GLAST (desired)
Energy range	GeV	0.02–30	0.02–300
Effective area	cm^2	1500	8000
Field of view	sr	0.5	2
Angular resolution (at 100 MeV)	deg	5.8	3.0
Energy resolution	%	10	10
Source sensitivity (>100 MeV) $\times 10^{-7}$	cm^{-2} s^{-1}	1	0.04

There are also two air shower particle detectors which have successfully detected γ-rays of a few TeV from the strongest sources. One is a large water Čerenkov detector, MILAGRO near Los Alamos, New Mexico, USA, at an elevation of 2.6 km. The other is a densely packed array of scintillation detectors in Tibet, which operates at an elevation of 4.3 km. Although these telescopes are somewhat less sensitive, they have the advantage over Čerenkov telescopes in that they can operate continuously and hence monitor a large section of the sky.

Future telescopes

INTEGRAL

The International Gamma-Ray Astrophysics Laboratory (INTEGRAL) is primarily a European mission which will be launched in 2001. It will operate in the 15 keV–10 MeV region with two prime objectives: good angular resolution (12 arcmin) and good energy resolution ($E/\delta E$) = 500). To achieve these objectives two instruments are used: the spectrometer SPI based on solid-state germanium detectors with coded aperture masks to define the field of view and the imaging IBIS which used cadmium telluride and cesium iodide detectors. These detectors are supplemented by small x-ray and optical telescopes for the monitoring of transient sources over a broad range

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of wavelengths. The detector design is based on the assumption that broad lines are emitted from point-like sources and narrow lines come from extended sources as indicated by earlier missions.

GLAST

The next-generation pair production telescope will replace the spark chamber with solid-state detectors which will be more compact, more efficient and have better angular and energy resolution. However, the general principle of the telescope will be the same, with an anti-coincidence scintillator shield, an interaction region with three-dimensional imaging system and a calorimeter; only the triggering system is unnecessary, as the imaging system can be active all the time. There are no expendables and no noisy pulsed high voltage and sparks.

NASA envisages the next-generation telescope as operating in the range 20 MeV–300 GeV, with a scheduled launch date of 2005. With the somewhat unimaginative acronym, GLAST (Gamma-Ray Large Area Space Telescope), GLAST will surpass EGRET by a factor of 10–40 in most parameters. A comparison of the two missions is given in table 1.

There are two competing technologies for the central pair production detector on GLAST. One (fiber GLAST) uses crossed planes of silicon fibers coupled to multi-anode photomultipliers, separated by thin layers of high-Z converter plates. The calorimeter uses the same kind of detector but with thicker plates. The fibers are 1.3 m long and the whole detector is 1.8 m high; they have a square cross-section of side <1 mm. This technology has already been used in cosmic ray particle experiments and is thus favored by space scientists. The other technology (silicon GLAST) uses the silicon strip technology that has been used in HE particle accelerator experiments for a number of years; it has not so far been used in space science applications. Again the layers of ionizing-particle-sensitive detectors are alternated with thin layers of lead converter. The calorimeter will be made of bars of cesium iodide, with individual read-outs to give spatial resolution.

Both technologies seem to address equally well the physical demands of GLAST, so it will be a difficult choice to select just one of them. Remarkably both technologies can achieve the dramatic improvement over EGRET, outlined in table 1, with an instrument that will only be twice as heavy.

Next-generation VHE telescopes

The success of the atmospheric Čerenkov imaging technique has led to plans for next-generation VHE telescopes (NGVHETs) of increased collection area and complexity. One of the most ambitious of these is the Very Energetic Radiation Imaging Telescope Array System (VERITAS), first proposed to the Smithsonian Institution in 1996. An international collaboration has been formed to build this array of 10 m aperture Čerenkov telescopes in southern Arizona.

Similar NGVHETs are now under construction in the southern hemisphere: the High Energy Stereoscopic System (HESS) is a European collaboration based in Heidelberg, Germany, which plans to build an array of four 12 m telescopes in Namibia and the Collaboration of Australian and Nippon for a Gamma-Ray Observatory in the Outback-IV (CANGAROO-III) is an Australian–Japanese collaboration which will add three more 10 m telescopes to their existing 10 m telescope in Woomera, Australia.

VERITAS will consist of six telescopes located at the corners of a hexagon of side 80 m with a seventh at the center. The telescopes will be similar to the design of the Whipple 10 m reflector, which is the most sensitive telescope of its kind. By employing largely existing technology in the first instance and stereoscopic imaging, VERITAS will achieve the following:

- effective area >0.1 km^2 at 1 TeV;
- effective energy threshold <100 GeV with significant sensitivity at 50 GeV;
- energy resolution 10–15% for events in the range 0.2–10 TeV;
- angular resolution $<0.05^\circ$ for individual photons and source location to better than 0.005°.

Other designs are also under consideration. In the northern hemisphere a European collaboration centered in Munich, Germany, is building an advanced technology 17 m telescope called MAGIC. This will effectively bridge the energy gap between space- and ground-based γ-ray telescopes. There are also three groups using large solar arrays in an attempt to significantly lower the threshold of the technique. Although their optics are crude, such arrays can have very large mirror areas (>1000 m^2) and attain energy thresholds below 50 GeV. Two such experiments, STACEE in the USA and CELESTE in France, have already reported the detection of a signal from the Crab Nebula.

Bibliography

Catanese M and Weekes T C 1999 Very high energy gamma-ray astronomy *Publ. Astron. Soc. Pac.* **111** 1193–222

Fichtel C E and Trombka J I 1997 *Gamma-Ray Astrophysics: New Insight into the Universe* NASA Reference Publication 1386

Gehrels N and Michelson P 1999 GLAST: the next generation high energy gamma-ray astronomy mission *Astropart. Phys.* **11** 277–82

Morrison P 1958 On gamma-ray astronomy *Il Nuovo Cimento* **7** 858–66

Trevor Weekes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gamow, George (1904–68)

Physicist, born in Odessa, Russia, researched on radioactivity and atomic fission, emigrated to the USA. Worked with EDWARD TELLER on beta decay, worked out the theory of stellar nuclear reactions in terms of the tunneling of a colliding nucleus through another nucleus's potential barrier. Even though there is only a small probability that a nucleon in the Sun would tunnel through the electrostatic repulsion barrier of another (say once every 10 000 million milllion million collisions), collisions in the solar core are so frequent that enough nuclei do get close enough to fuse and release solar energy. Gamow developed a Big Bang cosmology in which a primordial matter (ylem) existed at the origin of the universe. He supported the LEMAÎTRE Big Bang model, predicting that its consequence was a residual cosmic fireball of background radiation with a temperature of 10 kelvin (PENZIAS and WILSON later discovered this radiation at 3 kelvin, surprisingly close to Gamow's prediction). He was the prime mover for a theory for the origin of the elements in the Big Bang in which neutrons were sequentially added to hydrogen to build up the periodic table of elements one by one. This was the alpha–beta–*gamma* theory, named in part for the sequence and in part for the authors, ALPHER, BETHE and Gamow himself. Outside astronomy, he correctly theorized that the DNA structure forms a code that directs protein synthesis. He was a successful popularizer of science, inventing a cartoon character, Mr Tompkins, who traveled through a world of scientific adventure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ganymede

Ganymede, one of Jupiter's icy moons and one of the Galilean satellites, is the largest planetary satellite in the solar system (see figure 1). This moon, which is larger than the planet Mercury (Ganymede's radius is 2638 km, Mercury's radius is 2439 km), generates its own magnetic field and has large polar caps composed of thin frosts which reach an average latitude of 40 degrees. On a global scale the surface of Ganymede can be subdivided into two basic terrains: dark terrain and bright terrain (see figure 2).

Dark terrains, which make up approximately one-quarter of Ganymede's surface area, are older, more heavily cratered regions that are generally crosscut by areas of the younger bright terrain. The low ALBEDO of the dark terrain is thought to be due to mixtures of rocky material with water ice on the surface. The origin of the dark material is unknown, but it is believed to have resulted in part from the non-ice components of the many meteors and comets that have hit Ganymede over geologic time. Although dark terrain was originally characterized using global VOYAGER images, moderate to high-resolution Galileo images show that dark terrain is made up of discrete regions with a range of albedos (e.g. bright hummocks interspersed with smooth dark patches; see GALILEO MISSION TO JUPITER). The relatively higher-albedo regions within dark terrain are brightened by both tectonic and impact processes. Dark material also appears to have moved down slopes, exposing bright icy walls, and collecting in valleys and depressions.

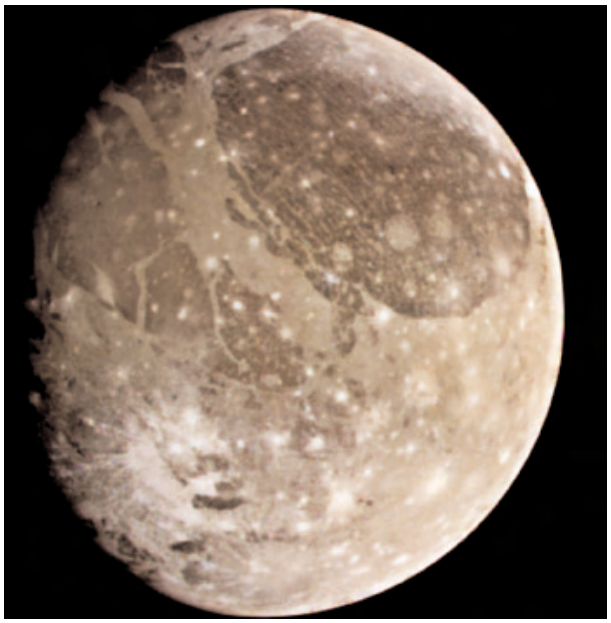


Figure 1. View of Ganymede from the Galileo spacecraft during its first encounter with the satellite. North is to the top of the picture and the Sun illuminates the surface from the right. The dark areas are the older, more heavily cratered regions and the light areas are younger, tectonically deformed regions. The brownish-gray color is due to mixtures of rocky materials and ice. Bright spots are geologically recent impact craters and their ejecta. The finest details that can be discerned in this picture are about 13.4 kilometers across.

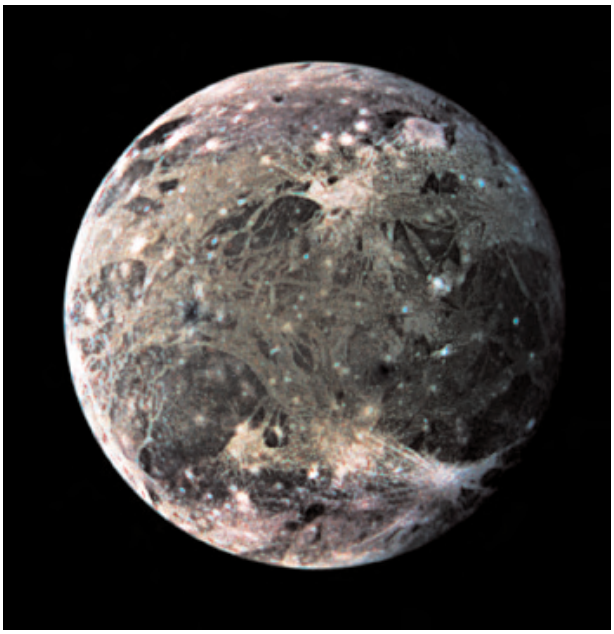


Figure 2. Global view of Ganymede's trailing side, enhanced to reveal frosty polar caps in addition to the two predominant terrains on Ganymede—bright, grooved terrain and older, dark furrowed areas. Many craters with diameters up to several dozen kilometers are visible. The violet hues at the poles may be the result of small particles of frost which would scatter more light at shorter wavelengths (the violet end of the spectrum). Ganymede's magnetic field, which was detected by the magnetometer on NASA's Galileo spacecraft in 1996, may be partly responsible for the appearance of the polar terrain.

Voyager and Galileo images show that this heavily cratered terrain, which is almost as old as the solar system itself, was fractured by tectonic activity as well as by large impact craters. Concentric furrow systems, similar to multiring structures on neighboring CALLISTO, are interpreted to have been formed when large asteroids or meteors impacted Ganymede early in its history.

Bright terrain, based on Voyager images, is divided into grooved and smooth subunits. Prior to the Galileo mission the smooth bright terrain was interpreted to have been emplaced by ice vulcanism. However, Galileo data not only reveal a paucity of volcanic features associated with bright terrain, but high-resolution data also show that many of these 'smooth' regions are in fact grooved or striated by tectonic activity.

Two wavelengths of troughs and grooves have been identified on Ganymede using Voyager and Galileo data. The average spacing between parallel grooves, based on

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

low-resolution Voyager data, is about 8.5 kilometers. However, high-resolution Galileo images reveal that smaller grooves, with an average spacing of about 1 kilometer or less, are superimposed on the longer-wavelength ridges and troughs. This two-wavelength style of deformation is interpreted to have formed by brittle deformation of an upper layer due to extension and thinning (producing the ridges/troughs with ~1 km spacing) combined with upwarping of regions by underlying warm and buoyant ductile ice (producing the longer wavelength deformation).

TECTONIC activity is now recognized as the dominant mechanism that has operated on Ganymede's surface. Tectonic resurfacing is thought to be responsible for transforming dark terrain into bright terrain. In one scenario, for example, a region that began as dark terrain is cut by numerous parallel to subparallel faults; the fault-blocks tilt with continued extension, and dark material slides down into the grooves and furrows, exposing bright, icy slopes and resulting in the formation of bright grooved terrain. While tectonic resurfacing appears to be the dominant process that forms grooved terrain, icy vulcanism cannot be completely ruled out. A depression and associated lobate material seen in Sippar Sulcus (an area of grooved terrain) are interpreted to be a volcanic caldera and water-ice volcanic flow. Additional ice-vulcanism may have occurred on Ganymede, but if it had, it has been subsequently masked by tectonic features and impact craters.

Ganymede is known as an icy satellite because its relatively low mean density (1926.8 kg m^{-3}) requires that it is made of about equal amounts, by mass, of water ice and rock-metal (actually, about 60% rock and 40% ice). Ganymede's mean density and normalized axial moment of inertia ($C/MR^2 = 0.3105 \pm 0.0028$, C is the axial moment of inertia, M is the satellite's mass and R is the moon's radius), known from radio Doppler measurements by the Galileo spacecraft, require that it be differentiated into a metallic core surrounded by a silicate mantle. The deep metal-rock interior of Ganymede is covered by a thick (800 km) ice-liquid water shell. Ganymede thus has a three-layer internal structure consisting of an inner metal core (radius 400–1300 km) depending on the composition and hence the density of the metal with the smallest radius corresponding to a pure iron core and the largest radius to an Fe–FeS core), a middle rock shell and an outer water (ice-liquid) shell.

Ganymede's magnetic field was first detected by the Galileo spacecraft. The magnetic field is large enough to carve out its own magnetosphere within the Jovian magnetosphere. The source of Ganymede's magnetic field is probably dynamo action within a still liquid part of Ganymede's metallic core. This is the same process, i.e. the geodynamo, that produces the Earth's magnetic field. The existence of a large Ganymedean magnetic field supports the inference from the gravity (Doppler) measurements by the Galileo spacecraft that Ganymede has a metallic core. The internal differentiation of Ganymede into its three-layer structure supports the inference of tectonism as the predominant surface modification process since differentiation is indicative of sufficient internal energy for endogenic activity.

Subsolidus convection in the ice-dominated outer layer of Ganymede plays a major role in controlling the evolution and present internal state of the satellite. Convection through the ice layer transfers Ganymede's internal heat to its surface. The source of this heat is mainly the present radioactive heat generation in the silicate component of the satellite. Secular cooling, i.e. the gradual cooling of Ganymede from an initially hot state established early in its history by accretion, differentiation and radioactivity, also contributes to the present heat loss from the satellite. Tidal dissipation in Ganymede is an additional potential heat source, since Ganymede is in an orbital resonance with EUROPA and Io (see IO: VOLCANISM AND GEOPHYSICS). While tidal heating is demonstrably important in Io, as evidenced by the satellite's sustained and ubiquitous volcanic activity, there is no evidence that tidal heating is important in Ganymede at present. However, tidal dissipation could have been significant in Ganymede in the past as the satellite evolved through other resonances in its orbital history. Perhaps this heat source played a role in differentiating Ganymede. Ganymede cannot have cooled so thoroughly over geologic time as to have completely frozen its metallic core since the satellite's present magnetic field requires dynamo action in a still liquid part of the core.

Web Update (31 July 2002)

New observations of Jupiter's satellites have been made during the Cassini spacecraft's flyby by Cassini, Galileo, the Hubble Space Telescope, the Chandra x-ray Observatory and Earth-based radio and optical telescopes. The Galileo spacecraft has been in orbit around Jupiter since December 1995 while Cassini observed Jupiter from October 1 2000 to March 31 2001. Cassini made its closest approach on December 30, 2000, at a distance of about 9.7 million kilometers (6 million miles). Observational results include data on interactions of the Galilean moons with Jupiter's magnetic field. Persistent, faint, far ultraviolet emission from the magnetic footprints on Jupiter's upper atmosphere from Ganymede has been observed. From these data it is inferred that Ganymede has persistent interactions with Jupiter's magnetic field despite its thin atmosphere.

Web Update reference

Clarke J T *et al* February 2002 Ultraviolet emissions from the magnetic footprints of Io, Ganymede and Europa on Jupiter *Nature* **415** 997–1000

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Bibliography

Anderson J D, Lau E L, Sjogren W L, Schubert G and Moore W B 1996 Gravitational constraints on the internal structure of Ganymede *Nature* **384** 541–3

Kivelson M G, Khurana K K, Russell C T, Walker R J, Warnecke J, Coroniti F V, Polanskey C, Southwood D J and Schubert G 1996 Discovery of Ganymede's magnetic field by the Galileo spacecraft *Nature* **384** 537–41

Schubert G 1997 Inside the solid planets and moons *Phys. World* **10** 45–9

Schubert G, Zhang K, Kivelson M G and Anderson J D 1996 The magnetic field and internal structure of Ganymede *Nature* **384** 544–5

G Schubert

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

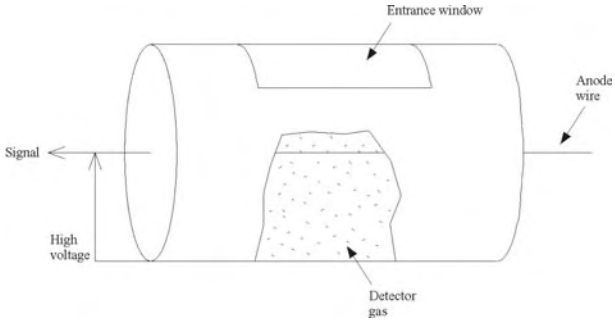


Figure 1. Schematic of a simple single-wire cylindrical proportional counter.

Gas Detectors

Gas detectors have been one of the principal methods of detecting cosmic X-RAYS since the early 1960s. They function by detecting the charge liberated when a photon interacts with the filling gas via the photoelectric process. Their principal performance advantage is that they can offer different combinations of high temporal and spatial resolutions, large area and bandwidth, good background rejection efficiency and high throughput. In addition, their basic design is well established, being a development of the Geiger counter. Only recently has the use of these devices in many areas of X-RAY ASTRONOMY and high-energy physics been replaced by CHARGE COUPLED DEVICES (CCDs), which can offer significantly better spectral resolution. Here the method of operation, the principal types and how different detector configurations are used to optimize different parameters such as energy and spatial resolutions and collecting area are discussed.

Method of operation

In its simplest form, shown in figure 1, a gas detector consists of an enclosed volume sealed with an x-ray transmitting entrance window, an anode at a typical potential of +2 kV and associated electronics. In front of the detector there is normally a grazing incidence x-ray optic, a coded mask or a mechanical collimator to focus, modulate or restrict the field of view of the incident x-rays, respectively. When an x-ray photon with an energy of ∼0.1–50 keV enters the detector it may ionize (i.e. remove one or more electrons from) one of the atoms contained in the detector volume. Under the influence of the electric field, the free electrons are then accelerated towards the anode where they cause further ionizations by colliding with other atoms. This produces an avalanche of electrons close to the anode which is detected as a pulse of charge by the associated electronics.

The number of initial electrons produced is proportional to the energy of the incident x-ray and depends on the characteristics of the filling gas, which often consists primarily of a noble gas such as Ar or Xe. Typically, a mean energy, W, of 20–25 eV is needed to liberate each electron. This means that a 6 keV x-ray liberates ∼300 electrons. If the creation of each electron–ion pair was an independent event, then Poisson statistics would apply, and the standard deviation would be the square root of the number of electron–ion pairs. However, this is not the case and this process usually gives better energy resolution than predicted in this way. The Fano factor, F, is used to quantify this difference and the FWHM energy resolution of the initial signal can be expressed as $\Delta E/E = 2.35(F/N)^{0.5}$, where N is the number of electron–ion pairs produced. F typically has values in the range 0.2–0.3 for gas-filled detectors.

Types of gas detectors

There are two principal ways of detecting the charge produced by the interaction of an x-ray with the detector gas. In a proportional counter the charge that arrives at the anode is measured directly by the associated electronics. With a suitable choice of the anode voltage the amount of charge read out is proportional to the energy of the incident x-ray—hence the name. The overall energy resolution of a proportional counter consists of two components. The first is derived from the statistics of the initial ion–electron pair production (see above) and the second from the charge multiplication process. The overall FWHM energy resolution at energy E can be expressed as:

$$\Delta E/E = 2.35[W(F + A)/E]^{0.5}$$

where $A \approx 0.6$ is the relative variance of the multiplication process. This component is the larger of the two and therefore dominates the overall energy resolution. This results in a typical energy resolution ($\Delta E/E$) of 20% at 6 keV. Examples of this type of detectors include the large area Proportional Counter Array on the ROSSI X-RAY TIMING EXPLORER satellite and the imaging Position Sensitive Proportional Counter (PSPC) on the ROSAT satellite.

In a gas scintillation proportional counter (GSPC) the primary charge cloud normally drifts into a region where the electric field strength is such that the electrons gain, on average, enough energy to excite the gas atoms to higher energy levels, but not enough to ionize them. No avalanche occurs and instead the UV light produced by secondary interactions is converted to an electrical signal by a photomultiplier tube (PMT). A schematic of a GSPC is shown in figure 2. The intensity of the UV pulse is proportional to the input x-ray energy. GSPCs provide an energy resolution of ∼10% at 6 keV, a factor ∼2 better than standard proportional counters, but at the expense of extra complexity and poorer background rejection efficiency. In addition, because of the low electric field strength in the drift region, some of the electrons produced by x-rays that interact close to the entrance window will not undergo amplification, limiting the sensitivity of these devices at energies $\lesssim$1 keV. An example of this type of detector is the Gas Imaging Spectrometer on the ASCA satellite. The Low-Energy Concentrator Spectrometer (LECS) on the BEPPOSAX satellite dispenses with separate 'drift' and 'scintillation' regions in order to extend the low-energy response down

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

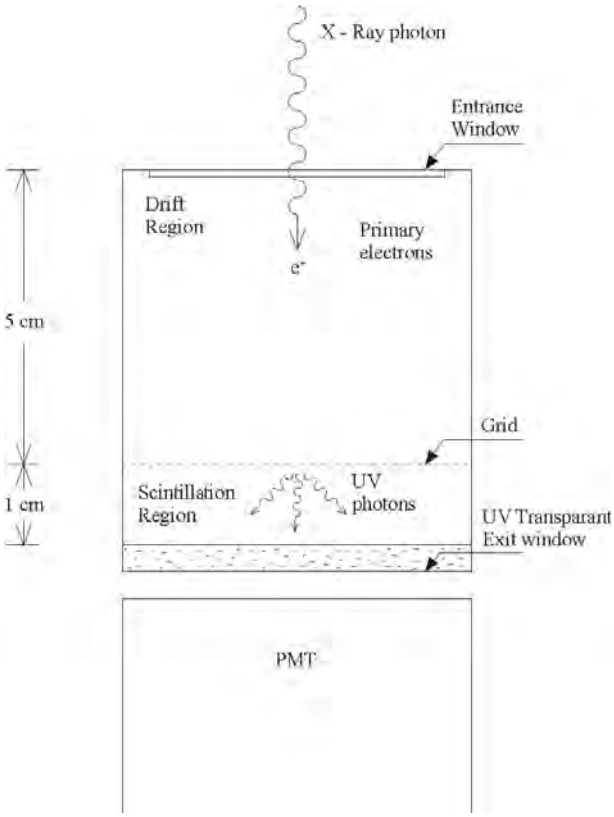


Figure 2. Schematic of a conventional GSPC with separate drift and scintillation regions.

to 0.1 keV, but at the expense of making the higher-energy ($\gtrsim$4 keV) signals no longer directly proportional to energy.

Space operation

As well as being sensitive to x-rays, gas detectors are sensitive to the energetic particles that pervade space. Many such particles have high enough energies that they can pass directly through the detector and the satellite. In doing so they deposit a line of charge in the counter gas, which is then detected in the same way as an x-ray. In an unshielded detector, the counting rate from such particles may be many times that from the target cosmic x-ray source. Improvements in gas detector design have resulted in ways of discriminating between particle and x-ray induced events. With proportional counters there are three main techniques.

1. Events that have energies between lower and upper boundaries are selected.
2. A second anode, called a guard counter, is placed in a region of the detector unreachable by x-rays. If an event is detected simultaneously on both wires then it is almost certainly produced by a charged particle. More complex arrangements of anode and guard wires divide the detector gas into many separate regions; simultaneous event detection in more than one region results in the event being rejected.
3. Since an x-ray photon is stopped by the detector gas in a very small region it yields a short pulse. A charged particle produces a long trail of charge, which will normally arrive at the anode over a much longer interval. The different timescales of these types of events can be measured and only short events accepted.

In the case of GSPCs energy and rise time thresholds are also used, supplemented, in the case of imaging systems, by analyzing the spatial distribution of light observed by the PMT.

Imaging detectors

Both proportional counters and gas scintillation proportional counters can be designed to measure the location in the detector where an x-ray was detected, and hence its likely location on the sky if a suitable x-ray optic or coded mask is used. The spatial resolution of the best imaging proportional counters is very good (the PSPC has a spatial resolution of around 0.3 mm at 1 keV), which has only been bettered by CCD and MICROCHANNEL PLATE DETECTORS. The PSPC contains two sets of thin cathode wires located at right-angles to each other interleaved with a set of anode wires. The wires are grouped together to minimize the number of pre-amplifiers required. The interaction site of an x-ray can then be recovered by measuring the relative strengths of the signals induced on the cathodes using the so-called 'center of gravity' technique. Simpler arrangements include interconnecting the sensing wires with resistors allowing only two signal processing chains per dimension.

The spatial resolution of imaging GSPCs is not as good as with proportional counters and is typically around 1 mm. In imaging GSPCs the PMT normally contains a small number of separate anodes which are read out and operated like an Anger camera. In the case of the LECS, the anodes are arranged in a square 3 × 3 configuration and measuring the distribution of light in each direction gives a measure of the centroid of the x-ray position.

Arvind Parmar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gas Giants

An alternative name for the giant planets, reflecting the fact that they are predominantly gaseous.

See: giant planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gas in Galaxies

The interstellar medium (ISM) can be thought of as the galactic atmosphere which fills the space between stars. When clouds within the ISM collapse, stars are born. When the stars die, they return their matter to the surrounding gas. Therefore the ISM plays a vital role in galactic evolution.

The medium includes starlight, gas, dust, planets, comets, asteroids, fast-moving charged particles (cosmic rays) and magnetic fields. The gas can be further divided into hot, warm and cold components, each of which appears to exist over a range of densities, and therefore pressures. Remarkably, the diverse gas components, cosmic rays, magnetic fields and starlight all have very roughly the same energy density of about 1 eV cm^{-3}. All the major constituents (or phases) of the ISM appear to be identified now, although complete multi-phase studies are extremely difficult beyond a few thousand parsecs from the Sun. The ISM is a highly complex environment which does not lend itself to simple analysis. However, this has not stopped astrophysicists from producing basic models of the ISM in order to make sense of the great wealth of data coming in from ground-based telescopes and satellites.

The study of the ISM began around 1927 with the publication of Edward Emerson Barnard's photographic atlas of the Milky Way. The atlas shows dark clouds silhouetted against the background starlight. At about the same time, spectra by John Plaskett and Otto Struve established the existence of clouds containing ionized calcium. By number of nuclei, about 90% of INTERSTELLAR MATTER is hydrogen, 10% is helium. All of the elements heavier than helium constitute about 0.1% of the interstellar nuclei, or about 2% by mass. Roughly half of the heavier elements are in the gas phase. But most of the refractory elements (Si, Ca, Fe) are depleted from the gas phase, and are locked up in small dust grains mixed in with the gas. Clouds only account for about half the mass and 2% of the interstellar volume. A far more pervasive 'intercloud' component was not identified until the discovery of PULSARS and the invention of ULTRAVIOLET/X-RAY ASTRONOMY in the mid to late 1960s.

The ISM's properties generally depend on the type of galaxy, and its distribution shows clear radial trends for a given galaxy. In DISK GALAXIES, the gas piles up into spiral arms (figure 1); this is where most of the young stars and SUPERNOVAE are to be found.

The ISM in galaxies is constantly evolving. Stellar winds and supernova explosions enrich the gas with metals over the course of billions of years. In the context of the widely accepted cosmological model of hierarchical GALAXY FORMATION, this may be compensated by the accretion of primordial gas in the outer parts of galaxies. Stars are the principal source of energy for the ISM. Starlight photons produce photoelectric emission from dust grains; these photoelectrons help to heat the neutral gas. Ultraviolet photons from the youngest stars ionize atoms and dissociate molecules. The main source of kinetic energy is the supernovae: these drive shock waves into the surrounding ISM and are largely responsible for its complexity.

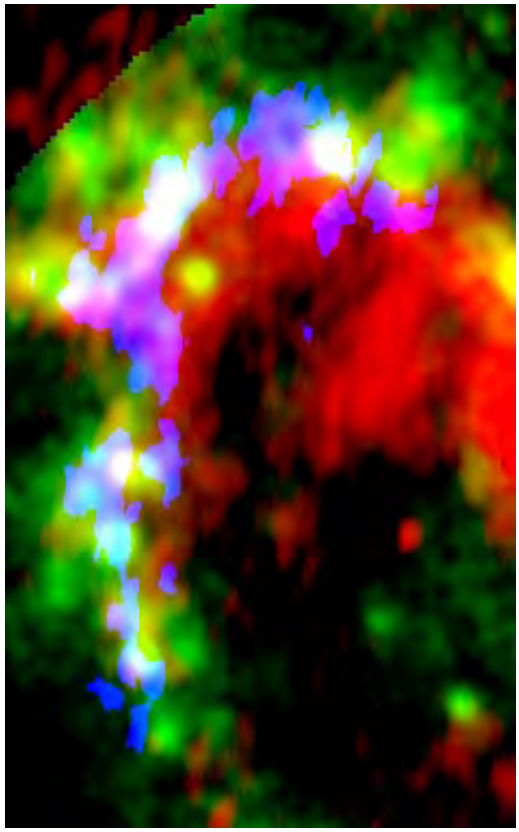


Figure 1. A color composite of the northern spiral arm in M83. The molecular gas is shown in blue, the 20 cm radio continuum in red and the ionized gas in green (courtesy of R Rand, University of New Mexico). Note that the cold gas, warm gas and dust pile up into spiral arms. Young stars form within the cold gas and then warm up the gas and dust through photoionization. Eventually, the young stars evolve to become supernovae which interact violently with the gas and dust (see figure 2). **This figure is reproduced as Color Plate 13.**

Atomic and molecular clouds

Half of the neutral atomic hydrogen and all of the molecular hydrogen in the ISM are concentrated into relatively high-density and low-temperature regions called 'clouds'. The properties of the atomic hydrogen (H I) clouds have been determined primarily from radio observations of the hyperfine ground state transitions of hydrogen at 1420 MHz (21 cm); interstellar absorption lines of trace elements such as Ca^+ also continue to play a key role in the study of these clouds. For the most common clouds, where the 21 cm radiation escapes freely, the brightness of the emission provides a direct measurement of the H I column density $N_{\rm H} = \int n_{\rm H}\,{\rm d}s$ where $n_{\rm H}$ is the atomic hydrogen density. When an H I cloud lies in front of a bright source of radio continuum emission,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

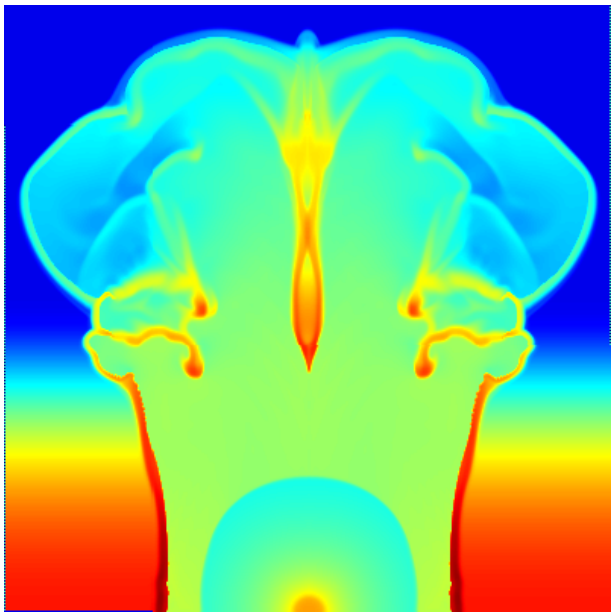


Figure 2. A simulation of the breakout from a galactic disk of a superbubble driven by multiple supernovae. The densities of gas in the disk and superbubble are shown in cross-section, with only the upper half of the disk shown. Red is high-density gas, while blue is low-density gas, with other colors of the rainbow intermediate. Each side of the image is approximately 800 pc long. (Courtesy of M-M Mac Low, American Museum of Natural History.) **This figure is reproduced as Color Plate 14.**

the decrease in the brightness of the background source at 21 cm is proportional to $N_{\rm H} = \int (n_{\rm H}/T_{\rm H})\,{\rm d}s$, where $T_{\rm H}$ is the temperature of the H I cloud. Thus, observations of H I clouds at 21 cm in emission and absorption provide direct information about cloud temperatures and column densities. Table 1 summarizes basic properties of the cold and warm neutral media. Maps of the sky show that H I clouds have complex shapes resembling thin extended sheets or filaments with embedded small clumps.

Molecular hydrogen is confined to the interiors of the densest and most massive clouds, the dark clouds, where starlight capable of dissociating molecules cannot penetrate. These clouds constitute the active star-forming component of the ISM. Because H_2 has no electric dipole moment, radiative transitions of H_2 are greatly suppressed. Therefore, most of the structural information about molecular clouds in the ISM is obtained through observations of the rotational transitions of the trace molecule CO at 115 GHz (2.6 mm). In addition, a wide variety of other molecules, including complex hydrocarbon chains, have been detected within the H_2 clouds. Molecular clouds are small (40 pc), dense (200 cm^{-3}), with structure on scales of less than 0.1 pc (see table 1 and also INTERSTELLAR MOLECULAR CLOUDS). Some of the small condensations can have densities as high as 10^5 cm^{-3}. In disk galaxies like our own, the cold neutral gas and molecular gas are confined to a disk which is much thinner than the stellar disk.

The intercloud medium

Astrophysicists have long wondered as to what confines gas clouds within the Galaxy. In 1956, Spitzer speculated that a rarefied ($\sim 10^{-3}$ cm^{-3}) hot gas ($\sim 10^6$ K), extending a kiloparsec or more above the Galactic plane, would confine the diffuse gas clouds observed far above the plane and would prevent their expansion and dissipation. Confirmation of the predicted corona came 17 yr later with the advent of space astronomy (table 1). Present-day satellites allow for the direct detection of diffuse x-ray emission and ultraviolet–x-ray spectral lines from the highly ionized trace elements within the hot gas (see the review by Spitzer 1990). The hot gas is thought to arise from the action of supernovae as we discuss below (see figure 4).

The existence of widespread hot gas in the disk of the Galaxy comes from observations of O^{5+} ions at ultraviolet wavelengths and the direct detection of soft x-ray emission. A clear demonstration of coronal halo gas far from the disk has been harder to come by. A wide range of ions is observed (Si^{3+}, C^{3+}, N^{4+}) towards the halo. Current models suggest that ultraviolet light from disk stars can only account for some of the ionization. The N V absorption lines appear to require collisional heating from a pervasive hot corona (Sembach and Savage 1992).

From a theoretical standpoint, the expectation is that hot young stars and supernovae punch holes or blow bubbles (figure 2) in the surrounding gas, and the diffuse hot component escapes into the halo through buoyancy. In fact, there are spectacular examples of bubbles seen at 21 cm in the Galactic ISM (see figure 3). However, there are many outstanding problems with these models, not least complications imposed by the magnetic field. Since the gas is thermally unstable, it is equally probable that the gas undergoes a 'cooling flow' or 'fountain flow' back towards the disk.

Filling the space between clouds are two additional components of the intercloud medium. By mass, most of the intercloud medium is in the form of a 'warm neutral' or a 'warm ionized' medium. These phases extend far beyond the thin disk of cold gas (table 1). The existence of the warm ionized medium was firmly established by 1973 from three independent observations: (i) low-frequency radio observations by Hoyle and Ellis, (ii) time delays in radio pulses from pulsars (see below) and (iii) through direct observation of H^+ recombination emission by wide-field Fabry–Perot interferometers. The gas has a density of roughly 0.1 cm^{-3} and a temperature near 10^4 K. The dominant source of ionization appears to be dilute ionizing flux from young stars in the disk, although some models suggest that the cooling radiation in old SUPERNOVA REMNANTS can be important. The deepest optical spectra to date show that the ionized gas extends to at least 5 kpc into the halo in some cases, and extends even further in radius than the H I disk.

Roughly half of the interstellar H I appears to be located in the 'warm neutral' component of the intercloud medium. This intercloud H I was first identified in 1965

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The component properties of the interstellar medium.

Component	Temperature (K)	Midplane density (cm^{-3})	Filling fraction (%)	Average height (pc)
Clouds				
H_2	15	200	0.1	75
H I	120	25	2	100
Intercloud				
Warm H I	8000	0.3[a]	35[a]	500
Warm H II	8000	0.15	20	1000
Hot H II	$\sim 10^6$	0.002	43[a]	3000[a]

[a] Value uncertain by at least a factor of 2.

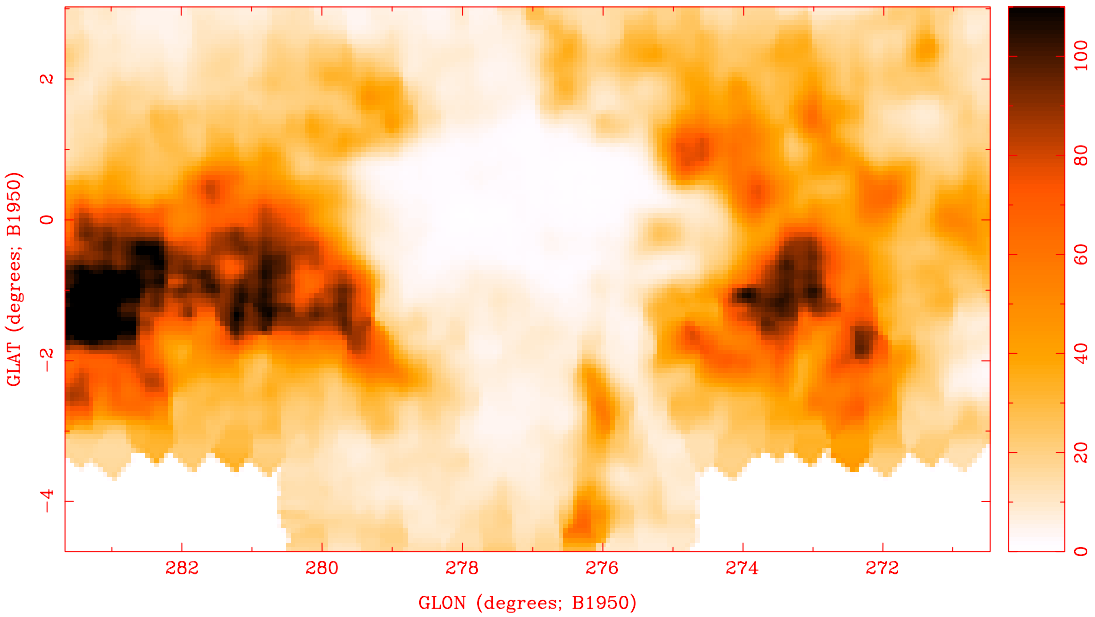


Figure 3. An image of a large Galactic H I supershell (white region at center). The empty supershell has a central brightness temperature of about 3 K; the shell edges have a brightness temperature around 60 K (black). The shell also shows narrow channels which appear to extend to the Galactic halo, forming a 'chimney' above and below the plane. The shell lies at a distance of about 6.5 kpc, has a diameter of roughly 600 pc and extends more than 1.1 kpc above the Galactic plane. The data were obtained at the Parkes Radio Telescope as part of the Southern Galactic Plane Survey. (Courtesy of N M McLure-Griffiths and J R Dickey, University of Minnesota.) **This figure is reproduced as Color Plate 15.**

as the source of the ubiquitous, relatively broad (velocity dispersion ~9 km s^{-1}) 21 cm emission features that had no corresponding absorption when viewed against bright background radio sources. The large velocity dispersions and the absence of absorption imply temperatures of 5000–10 000 K. Observations of the Ly α absorption line of H I toward bright stars show that this gas has a mean extent from the midplane of 500 pc, i.e. much thicker than the cold neutral disk. If the warm neutral medium is in pressure equilibrium with the cold component, then it would be clumped into regions occupying 35% of the intercloud volume with a density of 0.3 cm^{-3} at the midplane (table 1), although these numbers are highly uncertain.

The solar neighborhood

We now consider the interplanetary medium (heliosphere) within the solar system as distinct from the local interstellar medium. The Sun moves with a velocity of about 20 km s^{-1} relative to the local ISM. The solar wind produces a bow shock ahead of the Sun. This discontinuity (heliopause) defines the extent of the heliosphere in the upstream direction. It is expected that the Pioneer 11 or Voyager deep space probes will eventually confirm the existence of this boundary. The neutral interstellar gas is largely unaffected by the heliopause, so that the local neutral medium presumably streams relatively freely through the solar system. In contrast, ions and charged dust grains are probably deflected by the advancing heliosphere. COSMIC RAYS are relativistic particles and therefore penetrate the heliosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

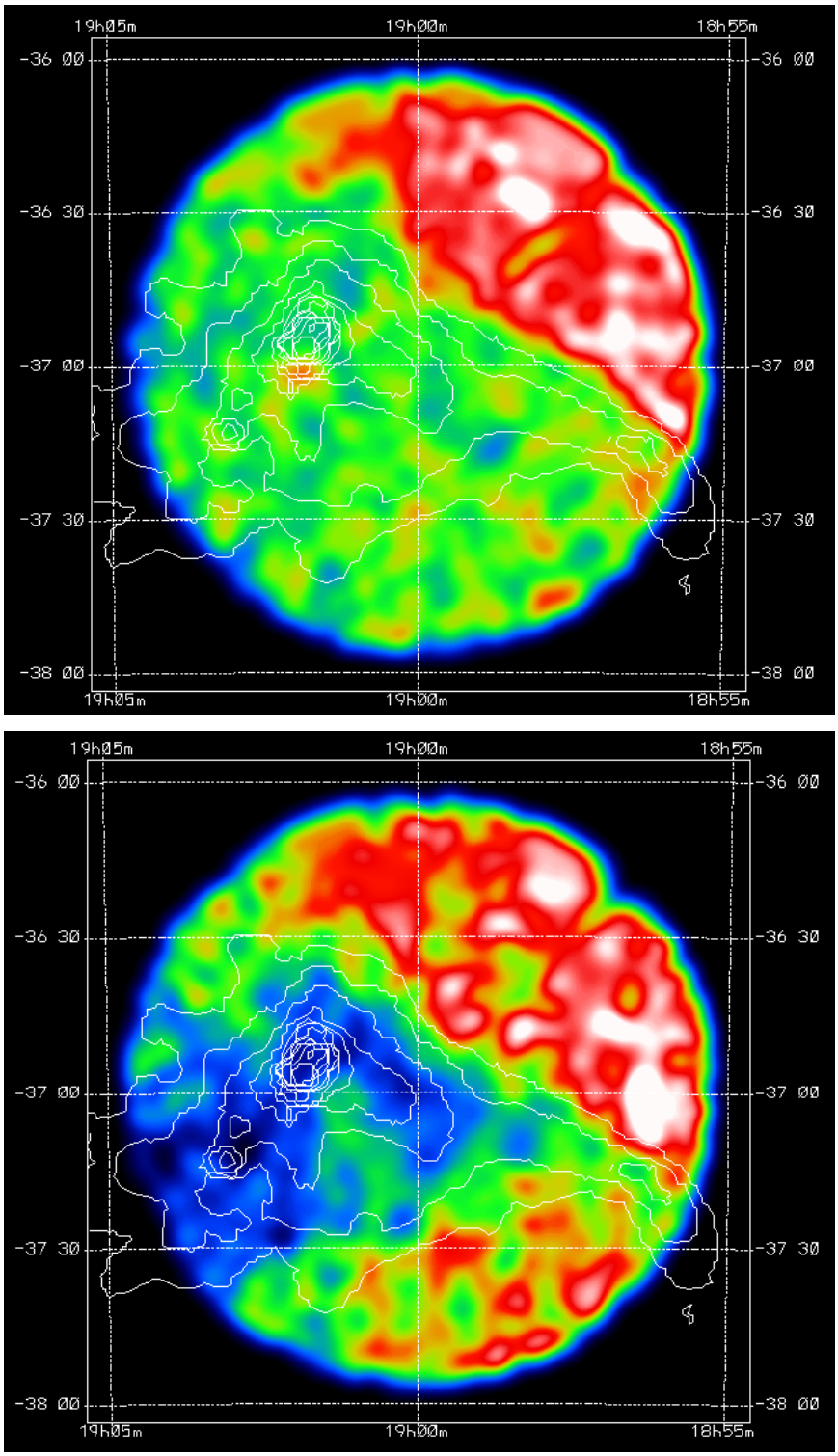


Figure 4. ROSAT false-color images of the Corona Australis dark molecular cloud. The contours show the 100 μm emission from dust in this cloud measured by the IRAS infrared satellite. The self-scaled images are for two different energy bands: (a) 100 eV and 300 eV (C band) and (b) 500–1100 eV (M band). These soft (low-energy) x-rays are absorbed by interstellar dust and gas. Because the M band x-rays are more penetrating than the C band x-rays, they are absorbed more strongly in the core of the cloud than in the periphery, while the C band x-rays from beyond the cloud are completely absorbed over the entire cloud. These images demonstrate that much of the x-ray flux originates from beyond the cloud. (Courtesy of D N Burrows, Penn State University.) **This figure is reproduced as Color Plate 16.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The most local gas we can associate with the local ISM has the properties of the warm neutral medium (see also SOLAR WIND: INTERACTION WITH LOCAL INTERSTELLAR MEDIUM). The solar radiation is backscattered by the local medium. Studies which exploit the resonance lines of hydrogen and helium show that the gas has a temperature of 8000 K, a hydrogen density of 0.25 cm^{-3} and a helium density of 0.02 cm^{-3}. Absorption line measurements toward nearby stars indicate that this warm gas is only a few parsecs in extent with hotter 10^6 K gas occupying most of the volume within 100 pc of the Sun.

Theories of the ISM

The two-phase model

Field, Goldsmith and Habing (FGH) (Field *et al* 1969) developed the first quasi-static theory of a multiphase ISM. The FGH model contained two stable phases, the 'cold' neutral phase at a temperature of ∼100 K and a 'warm' phase at $\sim 10^4$ K where about 10% of the gas is ionized. The model uses cosmic ray heating to balance the cooling of these two phases. The cooling is dominated by fine-structure excitations from C^+ in the cool gas, and collisional excitations (Lyα) in the warm gas. The model predicts an ISM where the majority of the volume is occupied by the warm ($\sim 10^4$ K) intercloud medium while the majority of the mass is contained within the dense cold clouds.

However, the FGH model indicated that the ISM should be stratified perpendicularly to the Galactic disk, which was known to violate the observed properties of the ISM. They therefore appealed to the collective explosive effects of H II regions and supernovae, which impart sufficient turbulent motions to destroy any gravitationally induced stratification. The FGH model allows the cold clouds to have a large range of sizes. The smallest clouds are destroyed by thermal conduction within the warm layer, while the largest are gravitationally unstable and prone to collapse. Otherwise, there is no limit on the size or shape of the expected neutral structures.

FGH also raised the possible existence of another stable phase at a temperature of $\sim 10^6$ K, dominated by bremsstrahlung cooling. However, without observations of the soft x-ray background or high-ionization UV spectra, there was no need to invoke the actual existence of this medium.

In the FGH model, supernova explosions were required to destroy the gravitationally induced vertical stratification of the ISM. This inspired theoretical models of the expansion and thermal evolution of supernova remnants (SNRs), and it was then that the full importance of supernovae came to be realized. Supernovae generate blast waves with speeds of 10 000 km s^{-1} which shock the ambient material, producing gas at $\sim 10^6$ K which then slowly cools through bremsstrahlung emission *and* line emission from highly ionized ions. The remnant structure is a hot bubble surrounded by a thin, dense shell of cool $\sim 10^3$ K gas. This situation proves to be quite stable; the hot bubble cannot expand through the shell expanding ahead of it. After a million years, the radius of the hot bubble is roughly 80 pc. At this point, the interior pressure of the bubble reaches the ambient pressure of the ISM. This static situation remains until the hot bubble cools and contracts after four million years or so.

Cox and Smith (1974) were the first to recognize the importance of this evolutionary sequence for the dynamical and thermal balance of the ISM. If an expanding SNR happens to intersect the static hot cavity of another remnant, the expanding remnant 'breaks out' through the old, hot remnant. Because of the low density within the static bubble, the expanding shock wave propagates preferentially into it, reheating the matter inside. Depending on the rate of Galactic SNe, the ISM could therefore evolve towards an interconnected tunnel network filled predominantly by hot gas.

To estimate the importance of such a tunnel network, a porosity parameter was introduced such that

$$q = r\tau V_{\rm sn} \tag{1}$$

where r is the average SN rate per unit volume, τ is the lifetime of an isolated bubble and $V_{\rm sn}$ is the final volume of the average SN-generated bubble. For $q \ll 1$, q is the proportion of the total interstellar volume that would be filled with hot bubbles. For an SN frequency of 1 per 80 yr, they found $q \sim 0.1$. This implies that 10% of interstellar space should be filled by unconnected, hot, SN-generated bubbles. Therefore, the probability of any new SN occurring within a preexisting cavity was 0.1, but the probability that any new SNR would expand into and reheat another older remnant was 0.55. The consequences of the overlap and phase structure of the ISM were radical.

This chain of reasoning rests heavily on supernovae occurring randomly throughout the Galactic volume. The simple conclusion prompted many to believe in the 'porosity imperative', that is the need for a hot 10^6 K phase to occupy the majority of the interstellar volume. The cold and warm ($\sim 10^4$ K) phases were relegated to the walls between the pervasive bubbles. Large filamentary structures in the ISM seem to support this, except that the overall pervasiveness and smoothness of the ISM are in conflict with this picture.

The three-phase model

This paved the way for a true theoretical tour de force, the McKee–Ostriker (MO) model of the supernova-dominated, three-phase ISM. In their model, supernovae produce the 'hot ionized medium' (HIM), the $\sim 10^6$ K component of the ISM in their bubble interiors, as well as enhancing the formation of the 'cold neutral' (CNM), 'warm neutral' (WNM) and 'warm ionized' (WIM) media along the compressed edges of remnants. There are still only two fundamental phases (CNM, HIM) in the MO theory. The WNM and WIM are restricted to the interface regions of the neutral clouds, and the WIM is in direct contact with the HIM and photoionized by thermal emission from it.

The model attempts to balance the thermal and mass exchange between the different phases. The energy input

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

from supernovae is offset by radiation cooling from the four media. The mass lost by cloud evaporation into the HIM is balanced by dense shell formation of swept-up interstellar material. The model is known to be incorrect in many details. There are numerous observations of local, dense clouds that are highly overpressured with respect to their environment. The model does not treat the influence of magnetic fields which are known to thread throughout the ISM and are likely to suppress thermal evaporation. The magnetic fields are of the right strength to play an important role in equilibrating the pressure balance. However, in the absence of a better working model, the MO theory remains the dominant conceptual framework. In fact, it remains unclear whether thermal pressure balance between the phases is an essential feature of the MO picture.

Pressure equilibrium should exist between the cosmic rays, the magnetic field and the kinetic interstellar component. However, the latter need not be the thermal pressure of the gas. Random, turbulent 'bulk' motions of the gas have a high enough energy density to provide the required effective pressure. The minimum interstellar pressure, $P = nT$, is probably about $\sim$25 000 cm^{-3} K and may even be higher.

The influence of supernovae

After a burst of STAR FORMATION has taken place, the first supernovae appear after about 10 million years. These type II supernovae arise from collapsing massive stars which have exhausted most of their nuclear fuel. Type II supernovae are distributed like the young stars in a disk galaxy, i.e. concentrated along spiral arms. A billion years after the initial starburst, the type I supernovae appear. Type Ia supernovae are thought to be due to accretion of matter by a white dwarf in a binary pair. These are to be found among the older STELLAR POPULATIONS and therefore have an exponential distribution with radius like the underlying disk.

The increase in type Ia supernovae towards the Galactic centre implies that here the ISM becomes supernova dominated, resulting in a two-phase ISM with only an HIM and a CNM. The warm phases are disrupted as diffuse clouds are shocked and heated to high temperatures. The molecular clouds survive the assault of supernovae and remain essntially intact, having only their outer warm layers stripped away. The scale height of the oldest stellar populations is higher ($\sim$300 pc) than for the atomic hydrogen ($\sim$100 pc), such that half of all type Ia supernovae explode above the gas and therefore deposit much of their energy directly to the halo. This picture produces a volume filling fraction of about 25% for the HIM, and presumably explains the H I holes seen in external galaxies.

The correlated distribution of type II supernovae has an important consequence. The time interval between successive supernovae is less than the bubble lifetime. This can result in a large-scale wind of energy into the Galactic halo. Heiles (1987) derives a two-dimensional porosity parameter

$$q_{2D} = 8.3\sigma s N^{-1} \tag{2}$$

where σ is the supernova rate in kpc^{-2} Myr^{-1}, s ($\sim$2) is a correlation factor and N ($\sim$40) is the number of type II supernovae that occur in a single association. However, this leads to a volume filling factor of more than 95% along the spiral arms, and about 80% outside of the arms, contrary to observation. Furthermore, the type II energy flow is expected to break out and produce a mass flow rate of $\sim 20 M_\odot$ per year, which would deplete the total gas content of the disk in 10^9 yr. How are we to reconcile this?

The role of pulsars

Discovered in 1967, pulsars are the rapidly rotating cores left behind by exploding type II supernovae. More than a thousand have now been observed in our Galaxy, in halo globular clusters, and in the LMC and SMC. Radio observations of these rapidly pulsating stars provide information on the plasma densities and magnetic field strengths of the intervening interstellar medium over baselines of tens of kiloparsecs.

The electrodynamic properties of a plasma are functions of the frequency of the electromagnetic wave that traverses the plasma. The group velocity of a wave in a plasma is

$$v_g = c\left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2} \tag{3}$$

where ω is the angular frequency and ω_p is the plasma frequency defined by $\omega_p^2 = 4\pi n e^2/m$, for which n, e and m are the number density, charge and mass of the free electrons.

Because the group velocity of a wave depends on its frequency, the Fourier components of a pulse will traverse a total distance d through a plasma in a time t_ω given by

$$t_\omega = \int_0^d \frac{ds}{v_g} \tag{4}$$

where s defines a small increment of distance through the plasma. Plasma frequencies in interstellar space are typically very low so we can expand equation (3) and substitute into equation (5) such that the time taken for the Fourier component of a pulse to traverse a plasma is

$$t_\omega \approx \frac{d}{c} + (2c\omega)^{-1}\int_0^d \omega_p^2\,\mathrm{d}s. \tag{5}$$

The first term is the time taken to traverse a distance d *in vacuo* and the second term is the plasma correction.

Studies of the arrival times of the various Fourier components indicate that the highest frequencies arrive ahead of the low-frequency components. What is actually measured is the derivative

$$\frac{\mathrm{d}t_\omega}{\mathrm{d}\omega} = -\frac{4\pi e^2}{cm\omega^3} D_m. \tag{6}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

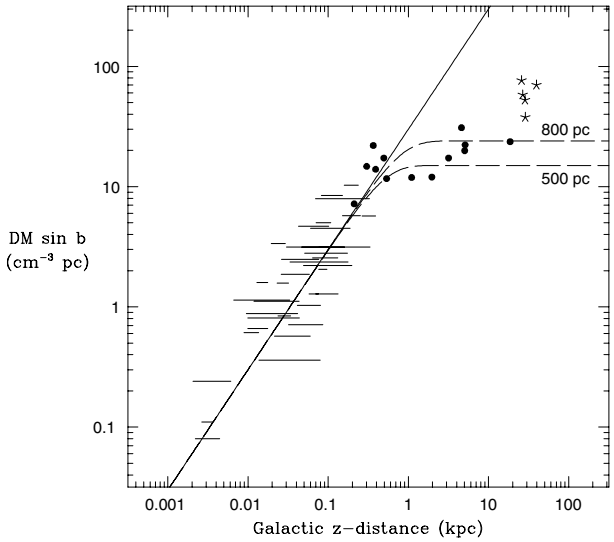


Figure 5. Dispersion measures from pulsars, $D_{\rm m} \sin b$ where b is galactic latitude, versus z-distance above the Galactic plane. The horizontal lines show the distance uncertainty for different pulsars. The black circles and stars refer to pulsars in globular clusters and the Magellanic clouds respectively. The sloping line corresponds to a model electron distribution which is uniform in density 0.03 cm^{-3}, and the two dashed lines are for models in which the electron layer has the same density at $z = 0$ but falls off with increasing z as $\mathrm{sech}^2(z/h)$ where h is 500 pc and 800 pc. (Courtesy of R N Manchester, Australia Telescope National Facility and J Taylor, Princeton University.)

The dispersion measure, defined as $D_{\rm m} = \int_0^d n_{\rm e}\,{\rm d}s$, is a measure of the total column of free electrons along the path to the pulsar in units of pc cm^{-3}. A column of 10^{20} electrons results in a $D_{\rm m}$ value of 30 pc cm^{-3} and a delay of 12 s for a signal at 100 MHz relative to infinite frequency. Dispersion measures were first obtained by the pulsar discovery team. Since then, we have come to learn (figure 5) from pulsars in distant globular clusters and the Magellanic Clouds that the warm atmosphere in the Galaxy extends to about a kiloparsec above the Galactic plane.

The radio signals reveal other important properties about the diffuse gas. In the presence of a magnetic field along the path to the pulsar, $B_\parallel$, the plane of polarization of the propagating wave will rotate by an angle χ equal to the phase delay between the ordinary and extraordinary components of the electric field. The so-called Faraday rotation angle is given by

$$\chi = \frac{\pi}{c\omega^2}\int_0^d \omega_{\rm p}^2\omega_{\rm B}\,{\rm d}s \qquad (7)$$

for which $\omega_{\rm B}$ is the cyclotron frequency. We define a quantity called the rotation measure,

$$R_{\rm m} = 0.81\int_0^d n_{\rm e}B_\parallel\,{\rm d}s \qquad (8)$$

in the traditional units of rad m^{-2}, and where $B_\parallel$ is in units of microgauss (μG). The ratio of $R_{\rm m}$ to $D_{\rm m}$ is the average galactic magnetic field strength along the path,

$$\langle B_\parallel\rangle = 1.232\frac{R_{\rm m}}{D_{\rm m}}. \qquad (9)$$

The influence of magnetic fields

Radio observations towards 500 pulsars show that there is a random and ordered magnetic field dispersed throughout the Galaxy. The ordered component has a field strength of 1–2 μG, while the random component is 5–6 μG with a cell size of about 50 pc.

The random field has a dramatic effect on the evolution of supernovae. While it can be neglected in the early stages of the explosion, as the remnant expands, the trapped magnetic field in the thin, cool shell resists the expansion, and this results in a much thicker shell with much lower compression. Now, after a million years, the hot bubble has an inner edge at 60 pc and an outer edge of 90 pc. Over the next four million years, the combined thermal and magnetic pressure of the thick shell forces the hot bubble to decrease in size. After $\sim 5\times10^6$ yr, the bubble radius is only 10 pc or so. Supernova bubbles can still overlap, but generally at a time when they are significantly weaker disturbances. This suggests that, consistent with the observations, the HIM produced by type II supernovae should be much smaller than predicted by Heiles.

Dependence on galaxy type

Surveys of several hundred galaxies show a systematic trend in the fraction of molecular gas to neutral gas for galaxies along the 'Hubble sequence' (see figure 6). This relationship is not fully understood although possible answers include the effect of the large-scale gravitational field in the formation or disruption of molecular clouds. For example, the stronger gravitational field in the bulge-dominated (early) galaxies may encourage fragmentation of the gas as a first step to forming dense clouds through collapse. It appears that there is also an enhanced fraction of molecules to neutral atoms in merging galaxies and cluster galaxies. In the latter case (e.g. Virgo), it is thought that the diffuse hydrogen has been swept away by the intracluster gas as the galaxy moves through the cluster.

Unanswered questions

On the subject of the ISM, there are vastly more questions than answers. There are few topics in astrophysics which are so well served by the new and planned generation of space-borne and ground-based observatories. These highly technological and expensive facilities provide the necessary impetus to encourage progress from theorists and computational analysts alike.

Although the principal components of the ISM have been identified and many of their properties measured, there is very little understanding of how they fit together into a dynamic system. This lack of progress has begun to stifle progress in other astrophysical fields. What are the feedback mechanisms that determine the stellar and interstellar properties of a galaxy? What regulates the star

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

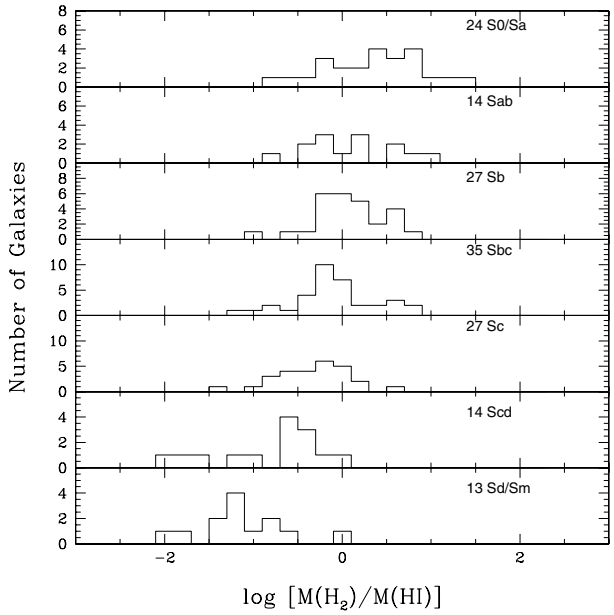


Figure 6. The ratio of molecular to atomic hydrogen for different 'Hubble types' of spiral galaxies. The histograms are presented in order of declining bulge to disk ratios, with the S0–Sa galaxies having the largest bulges. The increasing fraction of dense molecular clouds towards earlier Hubble types may reflect a higher cloud formation rate in the presence of a stronger gravitational field. (Courtesy of J S Young, University of Massachusetts.)

formation rate? Why do the many components of the ISM have very roughly the same energy density? What are the volume filling fraction and topology of the various gas phases? What is the large-scale and fine-scale topology of the magnetic field for each of the gas phases? How do the properties of the ISM change with cosmic time as more and more of the gas becomes locked up in stars?

To quote from Spitzer (1990), 'Understanding the processes that occur as the hot interstellar gas evolves in our Galaxy is an ambitious goal that we are far from achieving'. The dynamics of a compressible gas bombarded by photons and cosmic rays is a highly complex problem. Progress has been made through idealized models. In these models, one commonly recognizes three phases. Initially, the supernova ejects a rapidly expanding envelope whose interaction heats the ambient gas to high temperatures. Second, the hot gas expands and compresses or destroys clouds in its wake. Finally, the hot gas may escape or fall back down to the galactic plane. The problem is that, in practice, these three stages overlap so much that their mutual interactions are crucial. Since we can expect major gains in the computational power of supercomputers, we can look forward to the development of more realistic models in the decades ahead.

Bibliography

Cox D P and Smith B W 1974 Large-scale effects of supernova remnants on the Galaxy: generation and maintenance of a hot network of tunnels *Astrophys. J.* **189** L105–8

Field G B, Goldsmith D W and Habing H J 1969 Cosmic ray heating of the interstellar gas *Astrophys. J.* **155** L149–54

Heiles C 1987 Supernovae versus models of the interstellar medium and the gaseous halo *Astrophys. J.* **315** 555–66

Mac Low M-M, McCray R and Norman M L 1989 *Astrophys. J.* **337** 141

McKee C F and Ostriker J P 1977 A theory of the interstellar medium—three components regulated by supernova explosions in an inhomogeneous substrate *Astrophys. J.* **218** 148–69

Sembach K and Savage B 1992 Observations of highly ionized gas in the Galactic halo *Astrophys. J. Suppl.* **83** 147–201

Spitzer L 1990 Theories of the hot interstellar gas *Ann. Rev. Astron. Astrophys.* **28** 71–102

Joss Bland-Hawthorn and Ron Reynolds

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gas Tail, Cometary

The part of a comet consisting of ions and electrons formed from gas molecules expelled from the comet during a close approach to the Sun; also known as an *ion tail* or *plasma tail*. The molecules are ionized by ultraviolet light from the Sun and drawn away from the head of the comet by the solar wind. Not all comets develop a gas tail (or a dust tail), but both types of tail always point away from the Sun. Gas tails are straight, but may appear kinked or discontinuous as a result of a disconnection event—a consequence of a change in the polarity of the magnetic field entrained in the solar wind. They are predominantly blue, the characteristic color of emission by ionized carbon monoxide (CO^+) at 420 nm. Gas tails begin to form typically at 2 AU from the Sun and can extend to 10^8 km.

See also: cometary atmospheres, disconnection event.

Compare: dust tail, cometary.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gascoigne, Sidney Charles Bartholemew ('Ben') (1915–)

Astronomer, born in Napier, New Zealand, professor at the Australian National University, Canberra. Worked on stellar evolution, the distance scale of the universe and photometry. He was the project scientist for the Anglo-Australian Telescope at Siding Spring, New South Wales.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gascoigne, William (1621–44)

English astronomer and instrument-maker, died on the battlefield of Marston Moor, Yorkshire. Invented the eyepiece micrometer, using a screw to measure the distance between two wires inside the eyepiece, so as to measure small angles with precision, and made telescopes, including one with two convex lenses.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gaseous Nebulae

Orion treads the nightly sky, his jeweled sword hanging from his belt, its centerpiece a small bright cloud, the ORION NEBULA. Though under ideal conditions visible to the naked eye, the nebula was discovered telescopically by the French scientist Nicolas de Pieresc in 1610 at the same time that Galileo was probing the heavens. The Orion Nebula was but one of what would become a huge number of unresolved 'nebulae'. Some nebulae yielded to better instruments and were revealed as star clusters. In our own time, EDWIN HUBBLE began to resolve yet others as galaxies, M31 still being known as the Great Nebula in Andromeda.

Still other nebulae could not be so resolved because they were not made of stars. Turning his primitive spectroscope in 1864 to one of Herschel's 'planetary nebulae', SIR WILLIAM HUGGINS found it to radiate emission lines characteristic of a hot gas. Though some astronomers had contended that they had indeed seen the Orion Nebula break into myriad stars, Huggins found that it too displayed the characteristic three emission lines: hydrogen at 4861 Å and the mysterious green 'nebulium' lines at 4959 Å and 5007 Å that, in 1928, were identified by I S Bowen as 'forbidden lines' of doubly ionized oxygen. The Orion Nebula—and large numbers of similar objects—must also be made of hot ionized gas.

Three main kinds of 'emission nebulae'—those that radiate emission lines—inhabit the Galaxy. The PLANETARY NEBULAE, found everywhere from the galactic disk to the halo, are small shells of gas around, and illuminated by, hot dying intermediate-mass stars. SUPERNOVA REMNANTS are the heated remains of high-mass exploding stars. To distinguish between them, the Orion Nebula and its kind are known as 'gaseous nebulae' (in spite of the others being gaseous as well), as 'diffuse nebulae' from their fuzzy appearances, and as 'H II regions' because they are largely made of ionized hydrogen (figure 1).

The diffuse nebulae are distinguished by their relation to new stars and star formation. Vastly larger than planetary nebulae, they are strictly confined to the inner plane of the Galaxy where we find the giant molecular clouds (GMCs), and specifically are either inside or contiguous to the molecular clouds themselves, where they are commonly related to young T Tauri stars and to the bipolar outflows that make Herbig–Haro objects. The nebulae are local portions of the molecular clouds that have been ionized by new hot massive stars that the molecular clouds have recently spawned. More than any other galactic feature, the diffuse nebulae lead our eye to the Galaxy's newest birthing grounds.

The diffuse nebulae are closely associated with a fourth kind of galactic object, 'REFLECTION NEBULAE' that glow in a similar fashion but whose spectra are the same as those of the surrounding stars. Though they too are gas clouds, their gas is neutral and largely unilluminated, the visual radiance coming from the scattering of starlight by embedded dust grains, returning us again to the molecular clouds, which are cold because of the dust they contain.

Figure 1. No diffuse nebula can be 'typical', but the Rosette Nebula gives a fine sense of their general natures, a huge cloud of ionized gas illuminated by hot stars within. Set in the foreground are numerous dust clouds including some classic 'elephant trunks' that show the intimate relation between interstellar gas and dust (National Optical Astronomy Observatories).

Reflection nebulae are illuminated by stars too cool to ionize the gas. Turn up the stellar heat and they glow like Orion.

A tour

Several diffuse nebulae are fine showpieces that decorate the Milky Way. Though the brighter nebulae, like Orion, look rather greenish to the eye (as a result of the forbidden oxygen lines), photography reveals most as red, the color caused by the bright red H_α emission line at 6563 Å. Reflection nebulae are quickly revealed by their blue colors, as blue starlight is scattered much more efficiently than red.

Orion was featured as M42 in Charles Messier's famed 1780s list that warned comet hunters of celestial mimics. Just to the north, also in Orion's Sword, lies small M43, part of the same complex. To the east in dim Monoceros is the Rosette Nebula, which surrounds the open cluster NGC 2237, a circular galactic flower well over a degree wide. Then go north along the Milky Way into Perseus, where glows the California Nebula, faint but three degrees long. Swing near the North Pole into Cygnus and there next to bright Deneb is the immense North America Nebula, three degrees across and visible in binoculars.

Plunging south we hit the motherlode in and around Sagittarius, where diffuse nebulae abound. Here, admire the naked-eye Lagoon Nebula, which nearly harbors the Winter Solstice. Above it is the smaller Trifid Nebula, a red patch with a blue reflection border. A bit to the north flies the Eagle Nebula in Serpens, between them the Omega. Farther south the wonders yet improve, the rich southern Milky Way capped by the Carina Nebula, a massive complex over 2.5 degrees across.

Between these treasures are hundreds more, some small and compact, some great glowing clouds so faint

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

they are difficult to see. Many are known by their NGC numbers. Others are included in the 'Sharpless' catalogue of 1959, which contains 313 objects identified on the red plates of the Palomar Sky Survey (which highlight H_α radiation). Prefixed 'S', nearly all are true diffuse nebulae, with Orion called S281, and the North America S 117. Diffuse nebulae are also powerful radio emitters. The 'Westerhout' list of 1958, compiled from a long-wave radio survey of the Milky Way, includes over 50 gaseous nebulae (and several supernova remnants). Though it overlaps with the brighter Sharpless nebulae, it has the distinct advantage of including nebulae that are optically obscured by interstellar dust, radio waves traveling the Galaxy unimpeded. Indeed, W49, one of the largest diffuse nebulae in the Galaxy, is invisible optically, hidden behind the thick dust in the Great Rift. Numerous other catalogs simply designate the nebulae by galactic coordinates (G).

Other galaxies with active star formation are also filled with diffuse nebulae, some—'giant H II regions'—vastly larger than our own classic objects. The LARGE MAGELLANIC CLOUD, outlined by vast numbers of glowing red nebulae, contains the magnificent Tarantula Nebula, so big and bright that even at a distance of 52 kpc it acquired a naked-eye star name, 30 Doradus. It is similar in scale to NGC 604, which lies within a spiral arm of the Triangulum Galaxy M33. Indeed, a mere glance at the distribution of diffuse nebulae in an external galaxy shows clearly that they lie strictly along the spiral arms, and in fact help identify the arms. Since diffuse nebulae are closely identified with new stars, their locations show that the spiral arms are the galaxies' star-formation factories (figure 2).

Factoring in the distances of the known diffuse nebulae reveals an immense range in size and structure. The centerpiece Orion Nebula, considered small, is about 5 pc across and contains about 100 solar masses of gas. At the low end are compact diffuse nebulae whose diameters are measured in tenths of a parsec, while at the high end the giant diffuse nebulae 30 Doradus and NGC 604 are 400 pc across and contain thousands of solar masses lit by whole clusters of hot stars. If placed at the Orion Nebula, they would overfill the entire constellation.

The nebulae range in structure from ionized bubbles set within dark molecular clouds, through blisters on the clouds' surfaces that are heated from the outside, to extended thin sheets of ionized matter that have undefined boundaries, pervade the Galaxy, are uncataloged, and with no specific integrity push the limit of the definition of 'diffuse nebula'. This ionized background belongs more to discussions of the general interstellar medium.

The illuminating stars

Like the planetary nebulae, diffuse nebulae, predominantly made of hydrogen, are ionized by stellar radiation shortward of the hydrogen Lyman limit at 912 Å. For a star to radiate this Lyman continuum, it must be hotter than about 25 000 K, which corresponds to spectral class B1, a limit found empirically by Edwin Hubble in the 1920s.

Figure 2. The spiral arms of the galaxy M101 are outlined by great numbers of giant diffuse nebulae that dwarf the Orion Nebula, are lit by clusters of O stars, and reveal the locations of active star formation (National Optical Astronomy Observatories).

Earlier (hotter) spectral classes ionize their surroundings, while later (cooler) ones do not, their radiation producing blue reflection nebulae when the starlight is scattered from the dust embedded in the gas. Thus the diffuse nebulae are the realm of the O stars (B0 included here with them).

O stars are massive. At 25 000 K, a zero-age main sequence star has a mass around 10 $M_\odot$, the minimum mass that can ionize a diffuse nebula. Because typical B1 stars are evolved from higher masses, the average bottom-end mass is more like 12 $M_\odot$. Orion's ionizing 'Trapezium' quartet consists of B0.5, B0, O7, and O6 stars. Because mass and luminosity climb so quickly with earlier spectral class, the O6 star (with about 20 $M_\odot$) dominates, it alone effectively producing the nebula. Herschel 36, the O7 luminary of the Lagoon nebula, does the same. At the high end are 120 $M_\odot$ O3 stars whose radiation dwarfs that of the mere Trapezium. Exceedingly rare in our Galaxy, O3 stars abound in the immense cluster R136 that ionizes 30 Doradus, over 50 packed into a volume only 2 pc across.

O stars do not live very long, beginning to evolve through the main sequence band as soon as they are born. A new O7 star at 20 $M_\odot$ has but 10 million years to ionize its surroundings and make a diffuse nebula before it cools below class B1, where it makes a reflection nebula. At

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

40 $M_\odot$ we are down to 5 million years and at 120 $M_\odot$ to 3 million. At typical stellar velocities, the stars can move only a few tens of parsecs before they evolve and cool below the limit. Most O stars are thus quite attached to the nebulae that mark their places.

Above about 12 $M_\odot$, core hydrogen fusion ceases before the stellar surface cools below the limit. Diffuse nebulae can therefore be illuminated by evolved O and B stars, O giants, and even supergiants. At the extreme, the most massive stars can lose their envelopes through powerful winds and convert themselves to WOLF–RAYET STARS, which abound within the illuminating clusters of extragalactic giant H II regions. The winds of such evolving stars can have dramatic structural effects on their surroundings.

At the low limit near 10 $M_\odot$, the stars may evolve to oxygen–neon white dwarfs. Much above the limit, however, stars seem fated only to explode as Type II (core-collapse) SUPERNOVAE. The diffuse nebulae therefore mark the spots of future exploding stars, though by that time most will have left their surroundings. Supernovae within large diffuse nebula complexes can blow huge bubbles that shape the interstellar medium for new generations of stars to illuminate as diffuse nebulae.

Distances, dimensions, and distribution

Even with HIPPARCOS, diffuse nebulae are too distant for parallax measurements of their exciting stars, Orion being about 400 pc away. The distances to optically visible nebulae thus rely primarily on spectroscopic distances. These contain significant errors, as the main sequence becomes vertical with earlier spectral class, O7 stars for example having a range in luminosity that spans more than an order of magnitude. One must therefore have some idea of the evolutionary state of the star to pin down its luminosity. Much more accurate are distances from main sequence fitting of embedded clusters, the distance of the Lagoon Nebula coming from the open cluster NGC 6530 rather than from Herschel 36 alone, and that of the Rosette derived from NGC 2237.

Distances are also obtainable from nebular radial velocities found from Doppler shifts of the emission lines. Diffuse nebulae are the most extreme of disk Population I, allowing the assumption of circular motion around the galactic center. Adoption of a rotation curve for the Galaxy gives radial velocity as a function of distance along any line of sight, from which a distance can readily be found. This technique is the only one available for nebulae hidden optically by interstellar dust, for which radio emission lines are used. Unfortunately, along lines of sight toward the inner Galaxy, between longitudes 270° and 90°, two distances on either side of the point nearest the galactic center have the same velocities, so there is often an ambiguity. If distances are available from spectroscopic parallaxes, the nebulae can be used to derive the rotation curve of the Galaxy, a technique used to establish the flatness of the curve to great distances from the galactic center where dust obscuration is low. As with the planetary nebulae, a good recourse is to study the diffuse nebulae in other galaxies whose distances we already know, first among them being the MAGELLANIC CLOUDS, in which the nebulae can be readily resolved.

From a compilation of all information, astronomers loosely classify nebulae—H II REGIONS—into several groups. Smallest are the 'ultracompact H II regions' under 0.1 pc, which are found around the newest stars embedded in GMCs. Among them is the 'BN/KL complex' that lies hidden within the dusty cloud in back of the Orion Nebula and visible only through its radio radiation and the infrared that arises from dust heated by its embedded stars. They are often spatially associated with molecular masers whose sources lie within the enclosing molecular clouds. At the next scale are the 'compact H II regions' with dimensions in the realms of tenths of a parsec. The Orion Nebula, with a dense core about 0.5 pc across, could qualify, though its extended halo, ten times larger, places it more into the next category in the hierarchy: the 'classical' H II regions. Above these are 'large' structures like the North America Nebula, and at the top the 'extragalactic giant H II regions' like 30 Doradus and NGC 604, which abound in the arms of open spiral galaxies such as M101. Once the dimensions of the largest of the diffuse nebulae are measured, we can use them as distance indicators for yet more distant galaxies, a technique that has met with some limited success.

Diffuse nebulae and their associated stars are the most extreme of Population I, more so even than the GMCs, as they indicate where new star formation is actually taking place. The thickness (full width at half maximum) of the distribution in the Galaxy is only about 100 pc. Radially, diffuse nebulae track molecular hydrogen (and the GMCs) rather well, both peaking around 5 kpc from the galactic center, the galactic surface density there being over 10 times that found in the solar neighborhood (figure 3).

Spectra

The spectra of gaseous nebulae are similar to those of low excitation planetary nebulae and are formed by the same processes. The Zanstra mechanism, first applied to the planetary nebulae, produces the hydrogen emission lines, including the prominent optical Balmer lines and easily observed infrared Paschen and Brackett lines. Ignoring helium and the other elements, each Lyman continuum photon from the central star causes one hydrogen ionization, which results in one recapture of an electron by a proton, which in turn results in one Balmer photon. Ionization and recombination are responsible for the helium lines as well, and for a variety of oxygen, nitrogen, carbon, and neon emissions. The freed electrons thermally equilibrate by collision, and before they recombine can collisionally excite low-lying metastable states of lighter ions, particularly those of nitrogen, oxygen, and neon, to produce the forbidden 'nebulium' lines, all of these emissions, recombination and forbidden, extending from the ultraviolet through the infrared and even into the radio.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

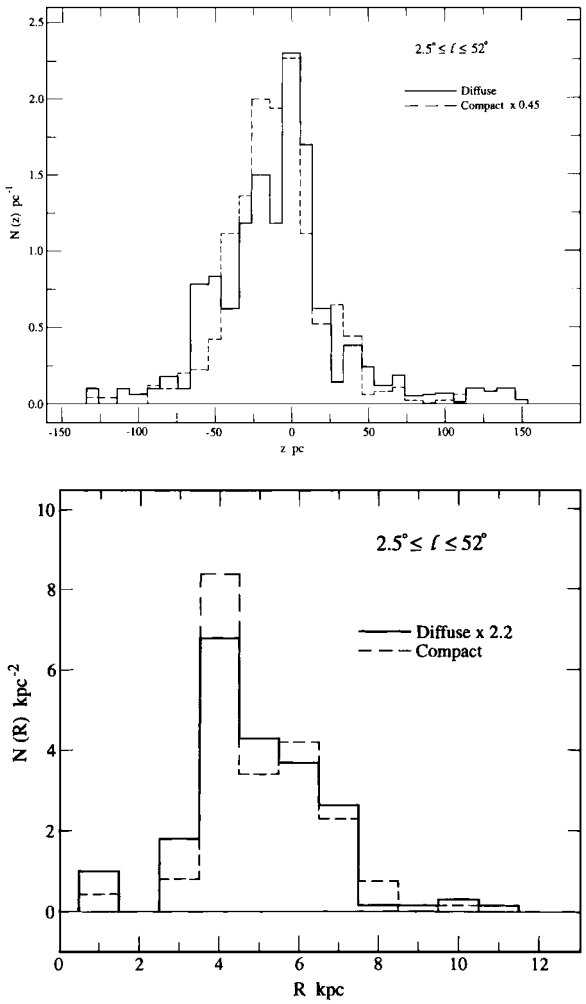


Figure 3. Diffuse nebulae of all kinds are tightly concentrated to the galactic plane (top), where z is the distance perpendicular to the plane. The number peaks about midway between the galactic center and the Sun (bottom), R the distance from the galactic center (from Lockman F J, Pisano D J and Howard G J 1996 *Astrophys. J.* **472** 173).

Spectra of diffuse nebulae differ from the average planetary nebula because their exciting stars can be no hotter than delivered by the main sequence, about 50 000 K for a zero-age O3 star, whereas 50 000 K is only modest for an evolved planetary nucleus. As a result, diffuse nebulae do not generally exhibit lines of ionized helium (He II), which arise from recombination from the doubly ionized state. Since most exciting stars are cooler, helium may not even be fully singly ionized within a nebula. Lower star temperatures also produce relatively weak double ionized oxygen lines, as O^+ may be the dominant ionization state. The exciting stars also commonly have absorption edges in the ultraviolet that allow even less ionizing ultraviolet radiation to emerge, yielding yet lower effective ionizing temperatures.

The great masses of the diffuse nebulae make the radio continuum and line spectra prominent. There are three principal sources of nebular continuum. An electron in hydrogen's second orbit can emit a pair of photons instead of Ly-α, the only constraint being that the energies of the two add to that of Ly-α. Such two-quantum emission provides most of the optical continuum emission (for both diffuse and planetary nebulae). Continuous emission is also produced by recaptures of free electrons onto bound states, producing jumps in continuum strength at the Balmer and Paschen limits at 3646 Å and 8202 Å and providing a general background. In the diffuse nebulae, these two continuum sources are added to by a stellar continuum scattered by internal dust.

Continuous emission is also radiated by free electrons that make close passes to protons but without recombination. The interaction changes one hyperbolic orbit to another, lost energy radiated as 'free–free' emission ('bremsstrahlung') that is particularly important at radio energies. The characteristic spectrum of free–free radiation readily allows the identification of diffuse nebulae as 'thermal' emitters (radiating as a result of thermally heated gas), as opposed to the non-thermal radiation of supernova remnants.

Though in principle there is no limit to the orbits of the hydrogen atom, the maximum principal quantum numbers are limited to under 1000 by interactions with neighboring atoms. Electron recombinations to high orbits n and the transitions to $n-1$ ($n\alpha$), $n-2$ ($n\beta$), etc., emit radio lines that are superimposed on the radio continuum, huge numbers of them being observed. At high quantum numbers, helium and other elements behave much like one-electron atoms. As a result, helium radio lines are observed to the high-frequency sides of their hydrogen counterparts. At slightly higher frequencies yet are similar emissions for 'carbon', a mixture of high-order lines from heavier atoms.

Because radio emissions are unimpeded by dust, they allow the discovery of nebulae that are hidden behind and within the thick dust of galactic GMCs, the emission lines also allowing measures of radial velocity and distance. Somewhat less obscured nebulae (as well as those optically visible) are also studied in the infrared through their hydrogen Brackett emissions and from the lines of a variety of other species.

While not hidden, optically visible diffuse nebulae are still dimmed and reddened by the dust (the problem affecting the planetary nebulae as well). The most noticeable effect is the steepening of the Balmer decrement, which describes the progressive weakening of higher-order lines within the series. After adopting an extinction curve (extinction versus wavelength), a coefficient C equal to the logarithmic extinction at H_β is found by forcing the Balmer line intensities to fit theory. C is readily converted to $E(B-V)$ color excess. Application of C to the extinction curve then allows the correction of the reddened line intensities. Theoretical forbidden-line ratios that span a large wavelength base serve as well. A major problem is that the diffuse nebulae are surrounded by a great deal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of circumstellar dust that may—and usually does—have a non-standard extinction curve as a result of grains that have been modified in the local GMC environment, which contributes errors to the corrected relative line intensities.

Physical parameters

The determination of the physical parameters of diffuse nebulae from optical observations involves much the same techniques as for planetary nebulae. Electron temperatures are easily derived from forbidden-line ratios, particularly of the 4363 Å and 5007 Å lines of [O III] and of the 5754 Å and 6584 Å lines of [N II]. Temperature fluctuations within the filamentary gas can cause serious distortion of the result, as the forbidden lines from the higher metastable states are preferentially produced at higher temperature, a problem difficult to account for. For both optically visible and obscured nebulae, electron temperatures are also derivable from the ratios of the fluxes in radio recombination lines to the flux densities in the neighboring continua. Account must be taken of natural masering in the radio lines and of the nebular geometry, as the observed radio flux is the average of that produced over a wide range of conditions, accurate results requiring a nebular model. Emission lines are also Doppler broadened as a result of 'turbulence' (a vague concept at best) and by thermal effects, the linewidth therefore giving an upper limit to electron temperature. The two effects can be deconvolved by observing emissions from ions of differing atomic weight. Line widths of hydrogen lines are all that is available for large numbers of distant, weakly radiating sources.

Electron temperatures depend on the balance between the heating rate, which depends on the temperature and ultraviolet spectrum of the central star, and the cooling rate, which effectively depends on the oxygen-to-hydrogen ratio (as the forbidden lines are collisionally excited by free electrons). They start notably lower than for planetary nebulae and range upward from around 4500 K for weak sources toward the center of the Galaxy through around 9000 K for Orion and up to 20 000 K for some extragalactic objects with low O/H. Since metallicity drops with distance from the galactic center, the electron temperature rises from its central low of 4500 K up to 10 000 K or so 2 kpc outside the solar orbit, the gradient about 400 K kpc^{-1}.

Densities fall equally well to optical and infrared forbidden lines, particularly to the close doublets of [O II] at 3727 Å and 3729 Å and [S II] at 6717 Å and 6731 Å. These ratios give densities within radiating sheets and filaments, which may occupy only a small portion of the nebular volume. Since the recombination rate depends on the product of the electron and proton densities, which are about equal, the fluxes in the recombination lines and continuum yield a mean electron density. Electron densities range from over 10^6 cm^{-3} for ultracompact H II regions to 5000 cm^{-3} or so for the filaments in the central regions of Orion to under 100 cm^{-3} for the huge, spread-out giant H II regions.

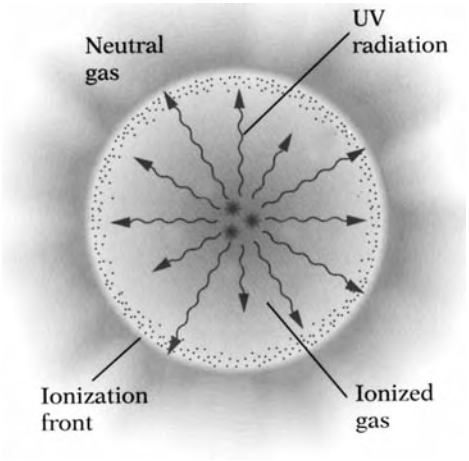


Figure 4. In the classic Strömgren sphere, hot stars ionize a bubble within a cold neutral molecular medium, the gas fully ionized out to a sharp edge (art from Kaler J B 1997 *Cosmic Clouds* Scientific American Library (NewYork: Freeman).

The mass of a diffuse nebula appears simple to find just by multiplying the electron density found from forbidden lines (essentially the proton density) by the mass of the proton and the volume of the nebula. Nebulae are highly filamentary, however, and the electron density is that of the radiating filaments. To find the mean density requires that we know the volume occupied by the filaments relative to the volume of the nebula, or the 'filling factor', which is generally set around 0.1, but is highly uncertain. The volume of the nebula is also uncertain, as the distance may not be well known and the three-dimensional structure may not be obvious; what may appear to be a rough sphere may be a sheet, or a blister on the edge of a GMC, or vice versa. The range of nebular masses is enormous, as they depend on the density and extent of the local molecular cloud (the amount of matter available to be ionized), the temperature and radius of the star (which controls the number of ionizing photons), the age of the star (how much matter it has had a chance to ionize), and on the number of sufficiently hot stars within the diffuse nebula. Masses go from under a thousandth that of the Sun for young ultracompact H II regions through 50–100 $M_\odot$ for modest objects like Orion to millions of solar masses for the giant diffuse nebulae like 30 Doradus and NGC 604 that are lit by clusters of O3 stars. The terms 'diffuse nebula' or 'H II region' define not one kind of object but a great many.

Chemical compositions stem directly from the theories of recombination and of the collisional excitation of the forbidden lines in both the optical and infrared, helium abundances for obscured nebulae also found from the radio recombination lines of high-order states. Major problems involve the incomplete ionization of helium, which gives a lower limit to the helium abundance, and severe filamentation, which can distort electron temperatures and densities. Detailed models of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

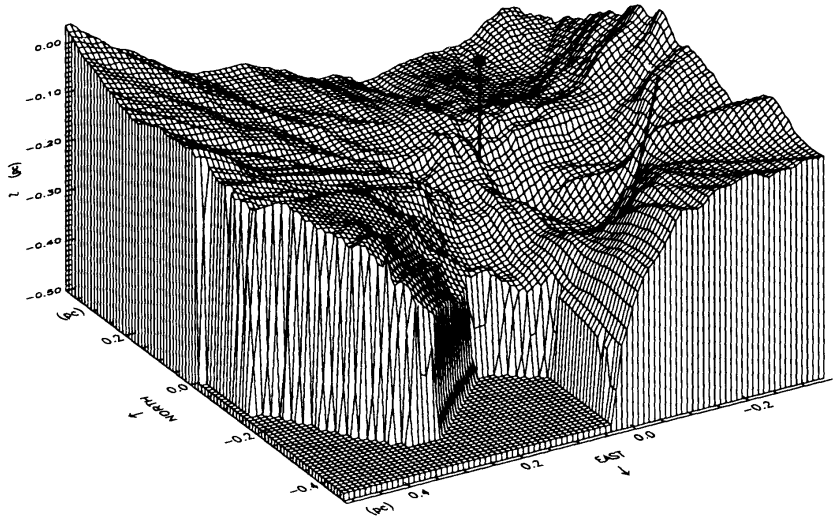


Figure 5. A three-dimensional model of the Orion Nebula shows its ionization front working its way into a dark molecular cloud, the ionizing radiation coming mostly from $\theta 1$ Orionis C, marked by the dark circle. The dark line to the right indicates the curved surface that makes the well known 'bright bar' (from Wen Z and O'Dell C R 1995 *Astrophys. J.* **438** 784).

nebulae can account for these difficulties as well as for incompleteness in the ions observed.

The resulting compositions are similar to those found in the Sun and in Population I in general. Indeed, more than any other kind of body, the nebulae define the composition of Population I. O/H, a general indicator of metallicity, averages from around double that found in the Sun in the inner Galaxy to about a third solar at distances half again the solar galactocentric distance. The N/H ratio drops in a similar way, though there are indications that it flattens out twice as far from the center as the Sun, the N/O ratio thus being anomalously high. Diffuse nebulae in other disk galaxies show similar, though controversial, relations.

Orion and other diffuse nebulae effectively at the solar circle have O/H ratios significantly lower (about 30%) than the Sun, not expected since O/H should have been built up in the interstellar medium to greater than solar in the five billion years since the Sun was born. One explanation posits that the metallicity of the Galaxy has declined as a result of infusion of low-metal gas through collisions with other systems.

Structures

A basic structure for a diffuse nebulae was defined by BENGT STRÖMGREN in 1939. His 'Strömgren sphere' is a bubble formed within a neutral cloud (a GMC) by a hot, ionizing star. The idea builds on the Zanstra mechanism, in which the number of ionizing photons radiated per second equals the number of recombinations that take place in the nebula per second. A diffuse nebula is illuminated by N stars (N_*) of (assumed similar) temperatures T_*. If the stars are black bodies, the rate of Lyman photon production depends on N_*, the integral of the Planck function from the Lyman limit to infinity (which approaches a T_*^4 proportionality for a very hot star), and the square of the stellar radius, R_*^2. The total recombination rate in a homogeneous medium depends on the product of the electron and proton density—effectively the electron density squared (N_e^2)—multiplied by a rate coefficient computed from recombination theory times the nebular volume, which depends on the nebular radius cubed ($R_{\rm neb}^3$). Equating the two yields

$$R_{\rm neb} = (\text{constant}) \times T_* R_*^{2/3} N_{\rm e}^{-2/3} N_*^{1/3} f(T_*).$$

The (constant) contains a combination of numerical and atomic constants and the recombination coefficient (which is slowly dependent on electron temperature), and $f(T_*)$ is the $\frac{1}{3}$ power of an integral of a function of $h\nu/kT$ taken from the Lyman limit to infinity (figure 4).

The photons within the Strömgren sphere fully ionize the nebula (ionizing those protons that have just captured electrons to create the neutral state) out to the radius $R_{\rm neb}$, at which point the gas suddenly becomes neutral. Photons with energies lower than the Lyman limit will pass through the Strömgren sphere unhindered, will ionize atoms with low ionization potentials, and will photodissociate molecules, producing a 'partially dissociated region' (a PDR) that is a layered transition to the fully neutral molecular gas in the GMC.

As appealing as it is, the Strömgren sphere is not a very realistic model for optically visible nebulae. By definition, the sphere must be inside a dusty GMC, and by their very visibility, the classic nebulae are not inside GMCs, as they otherwise would not be visible. The concept works better for obscured nebulae that are detectable only through their radio and infrared radiation.

Strömgren spheres, indeed diffuse nebulae in general, are dynamic. The heated gas in the Strömgren

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sphere produces a pressure at the boundary—the 'ionization front'—that expands the ionized zone into the surrounding GMC, shocking the gas in front of it, the nebula expanding, the electron density dropping as the pressures in the two zones go toward equilibration. If the nebula becomes large enough, it can break through the outer boundary of the GMC, the high pressure gas erupting in an ionized fountain that would readily be optically visible. If O stars form near the edge of the GMC, they can also erode the molecular layers separating them from open space. A nebula might then appear as a blister on the molecular cloud's surface, the stars in front of it, the nebula behind.

Though the Orion Nebula appears more or less spherical to the eye, its luminosity relative to its density and its internal velocities as measured by Doppler shifts clearly show that it is not an expanding Strömgren sphere, but in fact is such a blister that has dense ionized gas behind the star and a much lower density plasma in front of it. The star (mostly $\theta 1$ Orionis C) is eating its way into the dense dark cloud at the same time that the hidden KL object is eating its way out. The curvature of the blister relative to the line of sight gives the nebula considerable two-dimensional structure, the obvious 'bright bar' feature, for example, an ionization front seen more or less edge on, a PDR behind it. These structures make the planetary nebulae, for all their complexity, look simple (figure 5). The immense power of the ionizing O6 star is readily seen in the photoerosion of remnant disks that surround lower-mass stars seen in relief against the nebula. Superimposed on such complex nebular structures are the ever-present foreground dust clouds that take on a multiplicity of shapes that range from tiny black circular patches to elongated 'elephant trunks' in which stars are still being formed.

Gaseous (diffuse) nebulae thus allow a look into both local and global stellar evolution and into the evolution of the Galaxy as a whole. We use them to find how massive stars are born and how the Galaxy changes its chemical composition with time. Similar studies of other galaxies allow the comparison of their evolution with ours, yielding insights back into the development of our own. Equally important, they provide us with grand sights, shimmering within the Milky Way, allowing us to see the grandeur of star birth before our very eyes.

Bibliography

Aller L H 1984 *Physics of Thermal Gaseous Nebulae* (Dordrecht: D Reidel)

Kennicutt R C Jr 1984 Structural properties of giant H II regions in nearby galaxies *Astrophys. J.* **287** 116

Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Sausalito: University Science Books)

Sharpless S 1959 A catalogue of H II regions *Astrophys. J. Suppl.* **4** 257

Shaver P A *et al* 1983 The galactic abundance gradient *Mon. Not. R. Astron. Soc.* **204** 53

Shields G A 1990 Extragalactic H II regions *Ann. Rev. Astron. Astrophys.* **28** 525

James B Kaler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gaspra

An asteroid discovered by Grigorii Neujmin in 1916, designated (951) Gaspra. In October 1991 it became the first asteroid to be imaged from close by when the Galileo probe passed it at a distance of about 16 000 km on its way to Jupiter. Images showed a highly angular body measuring 18 × 11 × 9 km, pitted with small craters but with few large craters, so either Gaspra has a strong, metal-rich surface or it is a young object. Galileo also detected a significant interaction with the local interplanetary magnetic field, suggesting that there is remnant magnetism. Gaspra orbits the Sun at a mean distance of 2.21 AU (331 million km), near the inner edge of the main asteroid belt, in a period of 3.28 years; the orbital inclination is 4° and the eccentricity 0.17. It is an S-type asteroid, with a reflection spectrum similar to that of ordinary chondrite and stony-iron meteorites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gassendi [Gassend], Pierre (1592–1655)

Theologian, born in Champtercier, Provence, France. Gassendi rejected DESCARTES' philosophy, emphasizing the inductive method, atomism and a mechanistic view of nature. Gassendi observed the first recorded transit of Mercury, that KEPLER had predicted to occur in 1631.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gauge Bosons

Particles that transmit the fundamental forces. Photons convey the electromagnetic force between charged particles, intermediate vector bosons (the W and Z particles) transmit the weak nuclear force, which governs the radioactive decay of atomic nuclei, and gluons carry the strong nuclear interaction (or 'color force') between the quarks that comprise hadrons and mesons. The existence of an equivalent force carrier for gravitation, the graviton, remains unproven.

The term 'gauge', meaning 'measure', is used essentially for historical reasons, while the term 'boson' applies to any particle or atomic nucleus that has an integral spin and which, therefore, obeys Bose–Einstein statistics (the statistics that govern the behavior of systems in which an unlimited number of particles can occupy the same quantum state); bosons are not constrained by the Pauli exclusion principle.

Gauge bosons have integral values of spin (expressed in units of the Planck constant, h, divided by 2π). The photon has zero mass and charge. W particles have masses (expressed in energy terms) of 83 GeV (88 times that of the proton), the W^+ having a charge of +1 (in units of the charge on the electron) and the W^-, a charge of -1. The electrically neutral Z particles have masses of 93 GeV (nearly a hundred times that of the proton). Gluons have zero mass and charge but carry another property, analogous in some respects to electrical charge, which is called 'color'. Because the lifetime of a force-carrying particle is inversely proportional to the amount of energy that is required to make it, the range over which a gauge boson can transmit a force is inversely proportional to its mass. Because the photon and hypothesized graviton have zero mass, they can transmit the electromagnetic and gravitational forces over infinite distances. Because the heavy W and Z particles have lifetimes of about 10^{-25} s the range of the weak interaction is restricted to about 10^{-17} m. Although gluons have zero mass, and might therefore be expected to have infinite range, their 'color' properties confine their range of influence to within the diameter of the proton. The strong nuclear force between the protons and neutrons in an atomic nucleus is conveyed by pions (particles that are composed of quark–antiquark pairs).

See also: electronvolt, elementary particles, fundamental forces, fundamental particles, graviton, hadrons, mesons, neutron, Pauli exclusion principle, photon, proton, quantum mechanics, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gauss, Carl Friedrich (1777–1855)

Scientist, born in Brunswick (now in Germany), worked in mathematics and physics including number theory, analysis, differential geometry, geodesy, magnetism, astronomy and optics. A child prodigy, at an early age Gauss independently discovered BODE's law, the binomial theorem and the prime number theorem, as well as the method of the construction of a regular 17-gon by ruler and compasses. In 1801, Ceres, a minor planet discovered by G PIAZZI, was lost behind the Sun. Several predictions of its position on reappearance were made from the rather few discovery observations, including a prediction by Gauss which differed greatly from the others. When Ceres was recovered by ZACH on its reappearance it was almost exactly where Gauss had predicted. Although he did not disclose his methods at the time, Gauss had used his least squares approximation method. A *Gaussian distribution* is the name of the normal distribution of residuals in statistical data, solved by the least squares method, which Gauss refined in 1828. Gauss became director of the Göttingen observatory and published a book, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. This was a treatise on the motion of celestial bodies, in which he showed how to estimate and then to refine the estimation of a planet's orbit, following this with further work on mechanics. He carried out a geodesic survey of Hanover, inventing a surveying instrument, the heliotrope, that worked by reflecting the Sun's rays. With Weber he investigated terrestrial magnetism, making magnetic surveys. To mark Gauss' theoretical treatment of magnetism, a unit of magnetism, the *gauss*, was named after him (superseded by the SI unit, the tesla).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gellibrand, Henry (1597–1636)

Born in London, England, became professor of astronomy in Gresham College, London, and discovered the change over time of magnetic declination, showing the shift of the magnetic pole and Earth's magnetic field.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geminga

A gamma-ray source in the constellation of Gemini, its name deriving from 'Gemini gamma-ray source'. Located at a distance of about 300 light-years, Geminga is a pulsar with a period of 0.237 s and, as such, is one of the nearest known neutron stars. Its optical counterpart is a low-luminosity star of apparent magnitude 25 with an annual proper motion (angular motion across the sky) of about 0.2 arcsec. Because its luminosity outside of the gamma-ray region is so low, the nature of this source remained a mystery until the discovery of pulsed emission, by the ROSAT satellite in 1992, showed that it is a pulsar.

See also: gamma-ray astronomy, pulsars, ROSAT.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gemini

(the Twins; abbrev. Gem, gen. Geminorum; area 514 sq. deg.) A northern zodiacal constellation which lies between Auriga and Canis Minor, and culminates at midnight in early January. It represents Castor and Pollux, the twin sons of Leda, Queen of Sparta, in Greek mythology, whose brotherly love was rewarded by a place among the stars. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

The two brightest stars in Gemini, α Geminorum (Castor), (combined) magnitude 1.6, and β Geminorum (Pollux), magnitude 1.2, are about 4.5° apart and easily recognizable. The former is a multiple system, consisting of a pair of bluish-white (A1 and A2) components, magnitudes 1.9 and 3.0, separation 3.1″, period about 467 years, and a third, red (M1) component (the variable star YY Geminorum), magnitude 8.9–9.6, separation 73″, each of which has an unseen companion, periods 9.21, 2.93 and 0.81 days, making this a sextuple system. There are 11 other stars of magnitude 4.0 or brighter, including ζ Geminorum, a bright Cepheid variable (range 3.6–4.2, period 10.15 days). Other interesting variables include the Mira-type star R Geminorum (range 6.0–14.0, period about 370 days) and the dwarf nova U Geminorum (range 8.2–14.9, mean period about 105 days).

Other interesting objects include M35 (NGC 2168), an open cluster of a couple of hundred stars fainter than eighth magnitude, and NGC 2392 (the Eskimo Nebula or Clown Face Nebula), a ninth-magnitude planetary nebula.

The Geminid meteor shower appears to radiate from a point just to the west of Castor.

See also: Castor, Geminids, Pollux.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gemini Observatory

The twin 8 m Gemini telescopes represent a new generation ground-based observatory incorporating many innovative technologies that will provide astronomers in the seven-country Gemini partnership with some of the best optical/infrared observations ever obtained from Earth or space.

The Gemini Observatory began as an effort by the United States National Science Foundation to build a new generation large astronomical telescope for the entire US astronomical community. However, it was quickly realized that a partnership with the United Kingdom and Canada (who had similar observatory plans) could produce twin 8 m telescopes more economically while providing the same amount of total telescope time for US astronomers. Chile, Argentina, Brazil and Australia later joined the partnership. Each partner is allotted observing time for their astronomical communities in proportion to their financial contribution.

In order to provide complete sky coverage of both the northern and southern skies, one of the Gemini telescopes is located on Hawaii's Mauna Kea (Gemini North) and the other on central Chile's Cerro Pachón (Gemini South). Both telescopes are expected to be fully operational by mid-2001 employing a staff of approximately 100. The Gemini telescopes utilize many new technologies, engineering innovations and operational philosophies that will allow them to excel in the study of stellar nurseries, galactic evolution and distant high-redshift galaxies.

Gemini North's first adaptive optics images, during commissioning in 1999, revealed the remarkable potential of the telescopes. Over integrations of several minutes, star images of 0.08 arcsec FWHM were obtained in the infrared.

For further information see
http://www.gemini.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geminids

One of the year's most prolific meteor showers, taking place in December. The radiant lies in the constellation Gemini. Uniquely, the parent body of the Geminids is not a comet but the asteroid (3200) Phaethon, discovered in 1983. The perihelion distance of Phaethon is a mere 0.14 AU. It may be that the Geminid meteor stream consists of particles that crumble away from the asteroid's surface under the intense heat of its perihelion passage. An asteroidal source is consistent with the Geminid meteoroids' high density of around 2 g cm^{-3}, about an order of magnitude higher than other meteoroids, which are cometary dust particles. These dense particles tend to produce quite slow, bright meteors of long duration.

See also: meteor shower, Phaethon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geminus (*c*. 130–*c*. 70 BC)

Astronomer and mathematician, born in Rhodes (?), Greece, believed that 'not all the stars lie on one surface, but rather that some of them are higher and some are lower'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gemma Frisius, Reiner [Regner, Regnier] (1508–55)

Born in Dokkum, Friesland, Netherlands, became professor of medicine and mathematics at the University of Louvain. In 1530 wrote *De Principiis Astronomiae Cosmographicae* in which he describes, for the first time, how the longitude of a place may be found using a clock ('it must be a very finely made clock which does not vary with change of air') to determine the difference in local and absolute times. He described the theory of trigonometric surveying as a method of accurately locating places.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

General Relativity and Gravitation

The General Theory of Relativity (GR), created by Albert Einstein between 1907 and 1915, is a theory both of gravitation and of spacetime structure. It is based on the assumption that matter, via its energy-momentum, interacts with the metric of spacetime, which is considered (in contrast to Newtonian physics and SPECIAL RELATIVITY) as a dynamical field having degrees of freedom of its own (GRAVITATIONAL RADIATION). In brief, 'matter tells spacetime how to curve, and spacetime tells matter how to move' (J A Wheeler). GR is a generalization of Newton's theory of gravitation and of Special Relativity, which are both approximations to GR under appropriate conditions (see below). More generally, GR provides a common, coherent basis for classical (as opposed to quantum), macroscopic physics. Its relation to quantum physics is not yet well understood (QUANTUM GRAVITY). While this poses an important problem of principle, it does not present a practical difficulty to astrophysics, since the atomic and subatomic scales, on which quantum laws operate, are very small compared to macroscopic scales, where gravity dominates.

According to present fundamental physics, the only one among the four basic interactions which acts between all kinds of matter is the gravitational interaction. Since it always acts as an attraction and thus cannot be shielded, and because its range is unbounded, it dominates the behavior of matter on large scales in spite of its extreme weakness. Therefore, gravity plays a significant role in nearly all parts of astronomy and astrophysics. (In the atomic and subatomic domain gravity is unmeasurably weak and totally negligible, as is highlighted by the 'fact' that the gravitational force between a proton and an electron is weaker than the Coulomb force by the factor 5×10^{-38}.)

Relativistic gravity is important in the following areas of research: high-precision astrometry (REFERENCE FRAMES AND TIME SCALES IN GR, SPACE AND TIME REFERENCES: CONCEPTS), compact objects such as NEUTRON STARS and BLACK HOLES and systems thereof (binary PULSARS, low-mass X-RAY BINARY STARS), active galactic nuclei and quasars, supernovae and gravitational collapse, identification of dark stars and planets (MICROLENSING), cosmology (COSMOLOGY: STANDARD MODEL, DARK MATTER, COSMOLOGICAL CONSTANT, distribution of luminous and DARK MATTER via weak GRAVITATIONAL LENSING). In the near future a new window onto the universe is expected to open when GRAVITATIONAL WAVES become observable.

Though GR covers a larger range of phenomena than Newton's theory, the latter provides a satisfactory description whenever the system under study is (nearly) isolated, the relevant velocities are small compared to the speed of light, $\frac{v}{c} \ll 1$, and bodies are not too compact, $\frac{2GM}{Rc^2} \ll 1$. Moreover, Newton's theory is used to describe the basic facts of gravity; it often serves as a point of departure for approximations in GR, and its shortcomings motivate the introduction of GR and help to understand and highlight GR's achievements. For these reasons, the Newtonian theory of spacetime and gravity is reviewed before GR is considered.

Newton's theory of gravity

Can be summarized as follows.

1(a) On spacetime, the set of events 'here-now', there exists an absolute time, measurable by means of arbitrarily moving, undisturbed clocks. Distances between bodies at one instant of time satisfy the laws of Euclidean geometry. Hence, motions of bodies can be referred to rigid frames of reference associated with coordinates (t, x^a) specifying times t and positions $(x^a) = \vec{x}$.

1(b) There exist preferred reference frames, inertial frames, with respect to which the motions of N bodies which form a nearly isolated system (like the solar system or a double star), are given by Newton's gravitational equations of motion

$$\ddot{\vec{x}}_{\rm A} = \sum_{B \neq A} \frac{M_{\rm B}}{r_{\rm AB}^2} \vec{e}_{\rm AB} \qquad (1)$$

which form the basis of classical celestial mechanics. The positive constants $M_{\rm A}$ are called the gravitational masses of the bodies, $r_{\rm AB}$ denotes the distance between the simultaneous positions of the bodies A and B, $\vec{e}_{\rm AB}$ indicates the unit vector pointing from $\vec{x}_{\rm A}$ to $\vec{x}_{\rm B}$, and dots signify time differentiation. Equation (1) characterizes an inertial frame as a rigid frame with respect to which the equations of motion do not contain velocity dependent (Coriolis acceleration) terms and relative to which the center of mass of such a system moves uniformly in a straight line.

1(c) If a body, or a part of a body, is subjected not only to gravity, its motion relative to an inertial frame is governed by the equation

$$m(\ddot{\vec{x}} - \vec{g}(t, \vec{x})) = \vec{F}. \qquad (2)$$

Here, the positive constant m is the inertial mass of the body, $\vec{g}$ denotes the acceleration the particle would have if subjected to gravity only, and $\vec{F}$ is the non-gravitational force acting on the particle. The product $m\vec{g}$ is traditionally called the gravitational force, although it is not needed to describe purely gravitational motions, see equation (1). In contrast to other forces, the gravitational force depends on the particle solely through the latter's inertial mass.

1(d) Measurements indicate that the ratio gravitational mass/inertial mass for laboratory-size bodies has a universal value. This ratio,

$$\frac{M}{m} = G = 6.673(10) \times 10^{-8}\ {\rm cm}^3\ {\rm g}^{-1}\ {\rm s}^{-1} \qquad (3)$$

though far less precisely known than most other physical constants, is assumed to have a universal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

value, the constant of gravitation. Accordingly, M and m are considered as different measures of one physical quantity, 'mass'. Note, however, that the gravitational mass of the Sun defined in accordance with equation (1) is known about two orders of magnitude more accurately than G: $M_\odot = 1.32704(2) \times 10^{26}$ cm^3 s^{-2}.

1(e) If matter is described as continuously distributed, with (inertial) mass density ρ, velocity $\vec{v}$ and pressure p, the effect of gravity on motion is expressed, in generalization of (2), as

$$\rho\left(\frac{\partial\vec{v}}{\partial t} + \vec{v}\cdot\vec{\nabla}\vec{v} - \vec{g}(t,\vec{x})\right) = -\vec{\nabla}p \qquad (4)$$

and the gravitational acceleration $\vec{g}$ can be obtained from the gravitational potential

$$\phi(t,\vec{x}) = -\int \frac{G\rho(t,\vec{x}')}{|\vec{x}-\vec{x}'|}\,\mathrm{d}^3x', \qquad (5)$$

by taking its gradient,

$$\vec{g} = -\vec{\nabla}\phi. \qquad (6)$$

Equations (5) and (6) generalize the expression for the acceleration of equation (1). For a spatially bounded mass distribution, the law (5) can be replaced by Poisson's equation

$$\Delta\phi = 4\pi G\rho \qquad (7)$$

together with the boundary condition

$$\lim_{|\vec{x}|\to\infty}\phi(t,\vec{x}) = 0. \qquad (8)$$

1(f) The laws (1) and (2) for particles, or (4)–(8) for spatially bounded matter distributions, determine the motions of all matter elements uniquely if its positions and velocities at one time are specified. Without the boundary condition (8), the laws (6) and (7) do not determine $\vec{g}$, and thus do not determine the motion; Newton's theory applies only to systems which can be treated as isolated, but not to the apparently spatially unbounded cosmic matter distribution.

All laws stated above take the same forms in all inertial frames, they are invariant under Galilean transformations which express symmetries of Newtonian spacetime.

Shortcomings of the Newtonian theory

Which led, first, to the special relativity theory (SR) and, later, to GR, include the following facts and considerations.

2(a) The assumptions about time and space listed above under 1(a) are incompatible with electrodynamical and optical phenomena involving moving bodies, with well established facts about moving clocks and with many other experimental results. Those facts for which gravity can be neglected or, more precisely, where the temporal and spatial scales on which the gravitational acceleration changes, are negligible compared to the relevant scales, are well accounted for by Special Relativity. According to SR, there is no universal, absolute time. Rather, each inertial frame has its own measure of time and also of distance, and the coordinates of inertial frames are related not by Galilean, but by Poincaré transformations. The metric of spacetime determines, for each pair of events $(t_1,\vec{x}_1)$ and $(t_2,\vec{x}_2)$, whether there exists an inertial frame with respect to which the events: (a) are simultaneous; or (b) happen at the same point of space; or (c) whether the events can be connected by a light signal. Formally, the metric is given, in any inertial frame, by the interval

$$\Delta s^2 = -c^2(t_2-t_1)^2 + (\vec{x}_2-\vec{x}_1)^2 \qquad (9)$$

and the alternatives just listed correspond to

(a) $\Delta s^2 > 0$
(b) $\Delta s^2 < 0$
(c) $\Delta s^2 = 0$.

According to SR, no interaction or signal can be transmitted faster than with the speed of light in vacuum, c.

2(b) In Newtonian theory, inertial frames can only be identified globally for isolated systems. 'Real' inertial frames associated with spatially separated, nearly isolated systems such as galaxies, are relatively accelerated and cannot be fitted into a single, 'truly inertial' frame of reference. Also, freely falling test particles which can be observed locally, exhibit relative accelerations in any real, inhomogeneous gravitational field; 'free' particles exactly obeying the law of inertia do not exist since all particles 'feel' gravity in the same way, see equation (2) with $\vec{F} = 0$.

2(c) As stated in 1(f), Newtonian theory does not provide predictive dynamical laws for spatially unbounded matter distributions.

2(d) Newtonian celestial mechanics is unable to account for the 'anomalous' advance of the perihelion of Mercury's orbit, noticed first by Leverrier in 1859.

General Relativity Theory overcame these shortcomings and predicted additional phenomena.

Basic concepts and laws of General Relativity

3(a) In order to preserve locally the approximate validity of SR, it is assumed that, for any event E of spacetime, there exists a coordinate system (t, x^a) such that the statements made in connection with equation (9) are valid for spacetime points P which are 'close to E', in the sense that

$$\Delta s^2 = -c^2\Delta t^2 + \Delta\vec{x}^2 + 0(4). \qquad (10)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(The symbol 0(4) indicates that additional terms are at least as small as fourth powers of the leading ones, e.g., $c^2\Delta t^2 \cdot \Delta\vec{x}^2$, and thus are negligible for small displacements $\Delta t, \Delta\vec{x}$ connecting E to P.) Such coordinate systems are called locally inertial at E. They may be vizualized/realized as connected with a small, non-rotating, freely falling cabin, like a space shuttle. Two such coordinate systems, referring to different 'central' events E_1 and E_2, represent approximately inertial systems near E_1 and E_2, respectively, but in contrast to Newtonian or SR inertial systems they are nonlinearly related, and thus exhibit relative accelerations between those reference frames. The basic assumption, symbolized in equation (10), at the same time preserves SR as a local approximation and weakens the concept of an inertial frame as suggested by the considerations in 2(b). Mathematically this assumption means that spacetime carries a Lorentz metric whose components in a general coordinate system, $g_{\alpha\beta}(x^\gamma)$, are functions of those coordinates ($\alpha, \beta = 0, 1, 2, 3$). The fact that these functions in general cannot be reduced to constants in a finite coordinate domain means geometrically that spacetime is curved.

3(b) The most direct physical interpretation of the metric is that it determines the times shown by undisturbed clocks or natural time carriers like atoms. The motion, or 'history', of a clock is represented by a time-like curve (figure 1), and the proper time τ it measures is given by the 'arc length'

$$\tau = \int |g_{\alpha\beta}\, \mathrm{d}x^\alpha\, \mathrm{d}x^\beta|^{1/2} \tag{11}$$

along that world line. This statement is the GR-substitute for Newtons's axiom about absolute time.

3(c) According to Einstein's equivalence principle all local, non-gravitational laws of physics take the same mathematical form at the origin of a local inertial coordinate system. This principle restricts the relativity principle on which SR is based since it implies that the laws of SR hold only approximately in small domains; at the same time it generalizes its approximate validity to arbitrary gravitational fields. The principle is used to generalize laws from Special Relativity to General Relativity. Applied to the law of inertia of SR, it leads to the statement that the spacetime curve, or world line, of a freely falling particle is a straightest curve, or (timelike) geodesic, in spacetime. Applied to Maxwell's laws of electromagnetism it provides the corresponding laws in curved spacetime, i.e., in an arbitrary gravitational field. In the geometrical optics approximation this implies that light rays (≡ photon world lines) are null geodesics, i.e., geodesics everywhere tangent to the local light cones (figure 1). Analogously, one obtains the equations of motion for fluids or gases in gravitational fields. While in these and similar cases of classical physics Einstein's equivalence principle works perfectly, it proved to be complicated to generalize quantum field theory to curved spacetimes. For reasons given in the introduction this is, however, so far of no practical importance to astrophysics.

The mathematical forms of the general-relativistic laws for particles and fields show that in GR the metric $g_{\alpha\beta}$ of spacetime also plays the role of the gravitational potential which has ten components that replace the one Newtonian potential ϕ.

3(d) The third, and most important, ingredient of General Relativity is Einstein's gravitational field equation. It relates the metric field of spacetime to the distribution and motion of matter and energy. More precisely, the metric tensor $g_{\alpha\beta}$ and its first and second derivatives determine the so-called curvature tensor (which measures the tidal field) and a tensor derived from it, called the Einstein tensor $G^{\alpha\beta}$. Moreover, the energy density, energy current density and stresses of matter and fields are combined into a ten-component quantity $T^{\alpha\beta}$, the stress-energy-momentum tensor. These fields are required to satisfy the (tensor) equation

$$G^{\alpha\beta} = \frac{8\pi G}{c^4} T^{\alpha\beta}. \tag{12}$$

This system of nonlinear partial differential equations replaces Poisson's equation (7) and does for gravity what Maxwell's equations do for electromagnetism.

3(e) In contrast to SR, the topology of spacetime is not given once and for all. Rather, the problem of finding a spacetime model involves to determine a four-dimensional manifold and, on it, physical fields satisfying equation (12). This aspect of GR is particularly relevant for the study of black holes and cosmology.

3(f) The unification of special relativity and Newtonian gravitation theory brought about by Einstein's GR has solved an old conceptual problem emphasized especially by Ernst Mach. In Newtonian theory (and in special relativity), the structure of spacetime as exhibited in the set of inertial frames, acts on matter, but it cannot be observed independently of the motion of matter, and it is not influenced by matter. It is, in other words, an absolute, unobservable cause. To assume such a cause contradicts the scientific mode of thinking ('widerspricht dem wissenschaftlichen Verstande') according to Mach, and Einstein consented. According to GR, the metric of spacetime is not given once and for all, but is a dynamical field interacting with matter, acting and being acted upon; in this respect GR overcomes Mach's objection. The dynamical nature of the metric is the most important insight of principle of GR.

Since the metric depends on the circumstances, one cannot introduce coordinates with a definite physical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

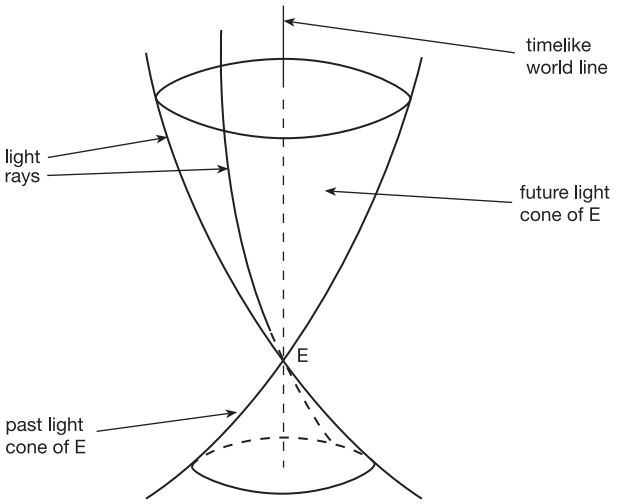


Figure 1. The future and past light cones of event E, generated by light rays, and a timelike world line which could represent a clock's motion.

meaning independently of the physical situation. Therefore, coordinates only serve to 'name' and distinguish spacetime points and to express their 'nearness' in a qualitative (topological) sense. For the same reason, the laws of the theory have to be formulated so as to hold in arbitrary coordinate systems; they have to be generally covariant. This requirement, together with the principle of equivalence, guided Einstein on his way towards GR.

Some consequences of GR

Causality

The mathematical analysis of the field equation (12) has shown that the metric (≡ gravitational field) propagates like a wave. Its value in the vicinity of an event A can influence the field at a later event B only if B is contained in the interior of or on the future light cone of A, i.e. the field propagates not faster than light. To determine the field at B, it suffices to know the field on a 'spacelike', three-dimensional subspace S of spacetime earlier than B, on that part C of S which is contained in the past light cone of B. This is illustrated, with one space dimension suppressed, in figure 2. These statements also hold for the combined gravitational and matter fields. They generalize corresponding laws of special relativity (e.g. for electromagnetic fields), and contrast sharply with the instantaneous action-at-a-distance of Newtonian forces.

Violation of local energy-momentum conservation

The field equation (12) implies that the energy tensor satisfies an equation which generalizes the laws of local conservation of energy and momentum from SR to GR, in accordance with the equivalence principle,

$$\nabla_\beta T^{\alpha\beta} = 0. \qquad (13)$$

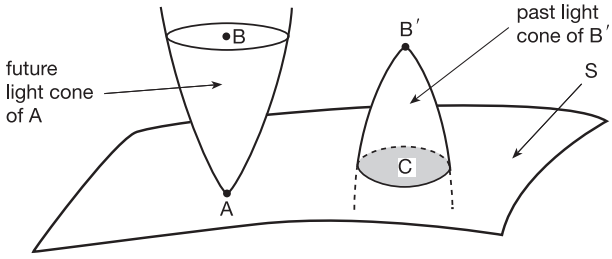


Figure 2. An event B in the interior of the future light cone of A and the region C of S which causally determines what happens at B′.

Special cases of this equation provide GR–generalizations of the equations of motion (1), (2) and (4) of Newtonian theory. The derivative operator ∇ depends on the metric and expresses that the localizable energy and momentum of matter and of non-gravitational fields are not strictly conserved, due to the action of gravity. For an isolated system, however, a global conservation law holds which includes a contribution from the gravitational field, see section Isolated systems, global energy-momentum below.

Equations of motion and gravitational waves

A fundamental task for any theory of gravitation is the prediction of the motions of bodies such as planets and stars. Approximation methods based on the gravitational field equation (12) and its consequence (13) have been developed to solve that problem. Solutions which represent the gravitational fields of well-separated, nearly spherical, slowly moving ($v \ll c$) and weakly stressed ($p \ll \rho c^2$) bodies forming an isolated system reproduce, in lowest order, the results of Newtonian dynamics and provide higher 'Post–Newtonian' order, general-relativistic corrections, see section Tests of the field equation below. These methods have been successful in describing effects in the solar system like the perihelion advance, and even the effect of the emission of gravitational waves on the motions of the binary pulsar are in agreement with the theory. We do not yet understand in detail what happens if, under the damping action of gravitational radiation, two neutron stars and/or black holes forming a binary system approach each other and then merge to form a single compact object. Such a process is expected to produce a strong gravitational wave signal, gravitational waves.

At large distances from the source, gravitational waves have properties similar to electromagnetic waves; they propagate with the fundamental speed c, are transverse, and have two independent states of polarization. However, while the patterns of two basic states of linear polarization of an electromagnetic wave form an angle of 90° as indicated symbolically by | and —, in the gravitational case they form an angle of 45° only as indicated by + and ×. This corresponds, in quantum language, to the fact that photons have spin 1 (in units of $\hbar$), while gravitons are expected to have spin 2.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitomagnetism
It follows from equations (12) and (13) that, in addition to Newtonian-type gravitational forces depending on the mass distribution only, there exist 'gravitomagnetic' forces depending on the flow of mass, analogous to magnetic forces, which depend on the motions of the charges. Such forces are taken into account in post–Newtonian approximations, but have not yet been measured in isolation, see Tests of the field equation below.

Isolated systems, global energy-momentum
The spacetime of an isolated system in GR is required to be such that its light cones behave, at large distances from matter sources, asymptotically like those of the spacetime of special relativity. It is then possible to define a conserved total energy momentum of such a system and to show that the total energy is positive, except if there is no true gravitational field, i.e., no curvature (Yau, Witten). It is also possible to give an exact expression for parts of the total energy-momentum which are radiated out of the system between two future light cones (Bondi, Sachs).

Horizons and black holes
The causal structure of a curved spacetime, indicated in the section Causality above, sometimes has the consequence that part of spacetime can never be seen by some observers. If a spherical star collapses, that particular future light cone of the center of the star which passes the star's surface at the instant when the ratio mass/radius of the star reaches the 'Schwarzschild limit' $c^2/2G$, will never reach an outside observer (figure 3). This so-called event horizon separates those events, which can be seen from outside, from those which are hidden from sight. That part of spacetime which is thus hidden is called a black hole; the horizon is its surface. Although according to (classical) GR no radiation and no particles can be emitted from a black hole, it does have a mass and attracts other bodies accordingly. For additional information on black holes see also QUANTUM GRAVITY, where singularities of gravitational fields are also briefly discussed.

Cosmology
General Relativity allows to treat spatially unbounded matter distributions and thus provides a basis for cosmology. The basic reason for this is that the solutions to equations (12) and (13) are determined, like those of Maxwell's equations, by initial data whose influence on the evolving metric and matter propagates at most with the fundamental speed c, as explained in the section Causality. Therefore, the field at some spacetime point never depends on a spatially infinite amount of matter, as it would according to the old equations (5) or (7). (Boundary conditions can be imposed in General Relativity only on initial data specifying a solution of the field equation (12), not on the field which evolves from these data.)

To allow for a static and spatially closed world model, Einstein in 1917 supplemented his field equation (12) of 1915 by a cosmological term $-\Lambda g^{\alpha\beta}$ on the right-hand side. This contribution has later been interpreted as the expectation value of the energy-momentum of the field fluctuations of the quantum vacuum. The appropriateness of this interpretation depends on a better understanding of the quantum vacuum and its relation to gravity. So far, this interpretation has not led to a computation of the value of the COSMOLOGICAL CONSTANT Λ.

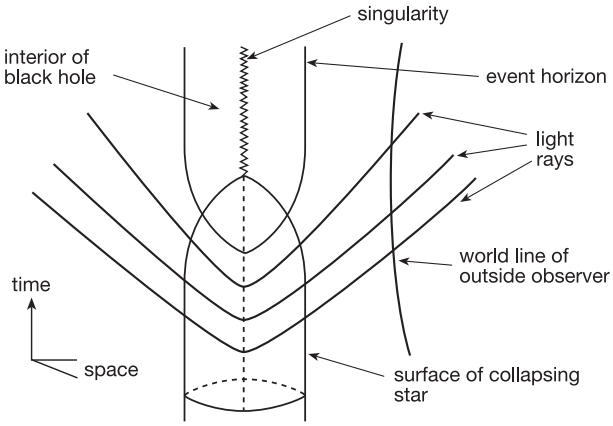


Figure 3. Collapsing star with event horizon, outside observer and light rays from the center of the star. The singularity is hidden behind the event horizon.

Experimental and observational tests of general relativity

Experiments and measurements supporting GR can be classified into those which test the Einstein equivalence principle and those which probe the gravitational field equation.

Tests of the Einstein equivalence principle
Tests of the first kind probe the approximate local validity of Special Relativity, i.e. the existence of local inertial frames or, equivalently, the possibility to assign a unique Lorentz metric (see 3(a) and (b)) to spacetime. From the viewpoint of GR these experiments serve to identify, relative to some arbitrary reference frame, a local inertial frame. Free fall experiments and Foucault pendulum experiments carried out in a terrestrial laboratory, for example, measure the acceleration and the angular velocity of the laboratory relative to a local inertial frame. Optical experiments and quantum interference experiments with neutron waves accomplish the same, and the concordance of the results establishes, up to experimental uncertainties, the uniqueness of the spacetime metric. The experimental fact that local inertial frames are accelerated relative to each other shows that spacetime is curved.

Tests of the field equation
Approximation methods to solve the field equation (12) show that, for a nearly isolated system of slowly, not too dense bodies, the metric differs from that of the flat spacetime of special relativity only in the time component;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the latter is determined by the Newtonian potential ϕ according to

$$-g_{00} = 1 + 2\frac{\phi}{c^2}. \tag{14}$$

This formula specifies how the gravitational potential affects clock rates and determines frequency shifts of electromagnetic waves; it has been tested for the Earth's gravitational field to within a measurement uncertainty of 2×10^{-4} (Vessot–Levine, gravity probe A). This formula is not specific to GR; it has to hold in any 'metric' theory of gravity to ensure the approximate validity of Newtonian theory. Tests which can distinguish between Einstein's theory, equation (12), and alternative theories of gravity require more accurate approximations to the metric. To obtain the light deflection, the metric

$$\mathrm{d}s^2 = -\left(1 + 2\frac{\phi}{c^2}\right)c^2\,\mathrm{d}t^2 + \left(1 - 2\frac{\phi}{c^2}\right)\mathrm{d}\vec{x}^2 \tag{15}$$

which exhibits space curvature, is sufficient. The validity of this approximate metric has also been confirmed by the retardation of electromagnetic waves in gravitational fields (fourth test, Shapiro effect) and by the geodetic precession (de Sitter effect) of the angular momentum of the Earth–Moon system. Additional terms are needed to measure the Schiff precession (mixed $\mathrm{d}t\,\mathrm{d}x^a$–terms, to be tested in the gravity probe B experiment in 2000), as well as the Lense–Thirring precession of the line of nodes of satellites orbiting the Earth, due to the Earth's angular momentum. This effect appears to be observable through the orbits of the satellites Lageos I and III (Ciufolini). The famous perihelion motion of Mercury's orbit requires the additional term $+2\phi^2/c^4$ in the g_{00} term of (15). The tests here summarized have verified Einstein's field equation up to relative measurement uncertainties of a few parts in 10^3. For more details, see GR: SOLAR SYSTEM EXPERIMENTS. Of still higher order of approximation is the effect of gravitational radiation reaction on the orbits of bodies; it has been verified in the case of binary pulsars (see GR TESTS: BINARY PULSARS). The major challenge of experimental gravity is the direct measurement of gravitational radiation which is expected to open a new window onto the universe.

Bibliography

The state of knowledge in each main area of general relativity theory is covered in at least one of the following reviews:

Hawking S W and Israel W (ed) 1997 *General Relativity: An Einstein Centenary Survey* (Cambridge: Cambridge University Press)

Held A 1980 *General Relativity and Gravitation* vols I and II (New York: Plenum)

Hawking S W and Israel W (ed) 1987 *Three Hundred Years of Gravitation* (Cambridge: Cambridge University Press)

A text book emphasizing relativistic astrophysics:

Straumann N 1984 *General Relativity and Relativistic Astrophysics* (Berlin: Springer)

Books concentrating on experimental and observational aspects:

Will C M 1993 *Theory and Experiment in Gravitational Physics* (Cambridge: Cambridge University Press)

Ciufolini I and Wheeler J A 1995 *Gravitation and Inertia* (Princeton, NJ: Princeton University Press)

J Ehlers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

General Relativity Tests: Binary Pulsars

PULSARS are rapidly spinning NEUTRON STARS emitting a beam of radio waves, like lighthouses. Experiment tells us that they are very stable clocks when isolated. A pulsar in a binary system is thus a moving clock, the best tool that one could dream of to test a relativistic theory. Indeed, many orbital data can be extracted from the analysis of the pulse arrival times, then compared with the predictions of the theory. Currently, general relativity passes all the available tests with flying colors.

Binary-pulsar data analysis

As the pulsar moves outwards or towards the observer, during its orbit, the Doppler effect modifies the pulse frequency like that of a fire-truck siren, and the time between two maxima of this frequency is thus a measure of the orbital period P_b.

More precisely, one can fit the pulse arrival times to a general formula taking into account any possible timing effect and depending on several phenomenological parameters. Five of them describe the approximate Keplerian orbit of the pulsar: the orbital period P_b, the epoch of periastron passage T_0, the eccentricity e, the longitude of periastron ω and the projected semimajor axis $xc \equiv a_1 \sin i$ (where a_1 is the semimajor axis of the pulsar orbit and i the inclination angle of the orbit with respect to the plane of the sky). The possible relativistic corrections to this Keplerian orbit are described by eight 'post-Keplerian' parameters:

(i) the Einstein timing parameter γ_T, which combines the second-order Doppler effect ($\propto v_1^2/2c^2$, where v_1 is the pulsar's velocity) together with the redshift due to the companion ($\propto Gm_2/r_{12}c^2$, where m_2 is the companion's mass and r_{12} the pulsar–companion distance);
(ii) the rate of periastron advance $\dot{\omega}$, i.e. the analogue of Mercury's perihelion shift for the present binary system (a relativistic effect of order v^2/c^2);
(iii) the rate of change of orbital period $\dot{P}_b$, caused by gravitational radiation damping, i.e. resulting from the fact that the system loses energy as it emits gravitational waves (an effect of order v^5/c^5 in general relativity);
(iv) the 'range' r and
(v) the 'shape' s of the Shapiro time delay, r being a global factor for this effect ($=Gm_2/c^3$ in general relativity), while s can be interpreted as the sine, $s = \sin i$, of the inclination angle;
(vi) three other parameters which have not yet been accurately measured in any system.

A similar analysis of the pulse structure also gives access to 11 extra post-Keplerian parameters, although they have not yet been determined precisely enough to be used to test relativistic gravity. (For further details, see also SOFTWARE: RADIO PULSAR SOFTWARE.)

The interest of such a phenomenological analysis of binary-pulsar data is that it is theory independent. Once the parameters have been obtained by a least-squares fit, they can be compared with the predictions of a given theory, which may be derived independently. In fact, the eight post-Keplerian timing parameters can be predicted in a relativistic theory of gravity only if one knows the two masses of the pulsar (m_1) and its companion (m_2), which are not directly measured. Moreover, the 11 pulse-structure parameters depend also on two unknown angles defining the pulsar spin axis. Therefore, if all the post-Keplerian parameters could be measured precisely enough, each binary pulsar would give $8 + 11 - 2 - 2 = 15$ tests of gravity. As shown below, only a few of them are currently available, in different binary-pulsar systems, but they are extremely constraining anyway. Moreover, even if one cannot determine enough post-Keplerian parameters to obtain a clean test, statistical arguments can still be used to constrain the possible theories of gravity.

The two most precise tests

PSR B1913 + 16

The Hulse–Taylor binary pulsar, PSR B1913 + 16, has been continuously observed since its discovery in 1974. This is a rapid ($P_b \approx 8$ h) and eccentric ($e \approx 0.6$) system, and three post-Keplerian parameters have been determined with great accuracy: γ_T, $\dot{\omega}$ and $\dot{P}_b$. The measure of $\dot{P}_b$ is in itself a great experimental achievement, as it confirms the existence of gravitational radiation. It is impressive to note that this quantity, of order -2×10^{-12}, is presently known with a relative precision of 0.35%.

All three observed parameters are nevertheless necessary to provide an actual test of gravity. In a given theory, they can be predicted as functions of the two unknown masses of the pulsar and its companion, and one obtains thus $3 - 2 = 1$ test. The equations $\gamma_T^{th}(m_1, m_2) = \gamma_T^{obs}$, $\dot{\omega}^{th}(m_1, m_2) = \dot{\omega}^{obs}$, $\dot{P}_b^{th}(m_1, m_2) = \dot{P}_b^{obs}$ define three curves in the two-dimensional plane of the masses (m_1, m_2). In fact they define three strips if one takes into account the experimental uncertainties. If these strips meet in a small region, there exists a pair of masses (m_1, m_2) which is consistent with all three observables, and the theory is thus consistent with the binary-pulsar data. If they do not meet, the theory is ruled out. Figure 1 displays these strips in the case of general relativity, which passes the test with flying colors. Note that, once the test is passed, it also gives a very accurate measure of the two masses (within the considered theory). A convenient way to illustrate the precision of this test is to use the values of the masses corresponding to the intersection of the very thin γ_T and $\dot{\omega}$ strips and to compute the prediction of general relativity for $\dot{P}_b$. One finds that it is consistent with the observed value at the 0.3% level (i.e. within 1σ error bars).

It should be noted that alternative theories of gravity may pass this test even if they are significantly different from general relativity. In this case, the three curves are very deformed with respect to figure 1, but they still meet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

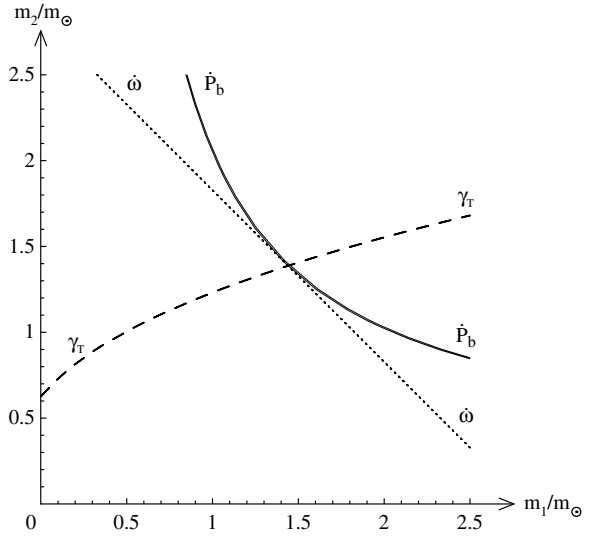


Figure 1. General relativity passes the $(\gamma_T$–$\dot{\omega}$–$\dot{P}_b)_{1913+16}$ test. The masses are given in units of the solar mass $m_\odot$, and the widths of the lines are larger than 1σ standard deviations.

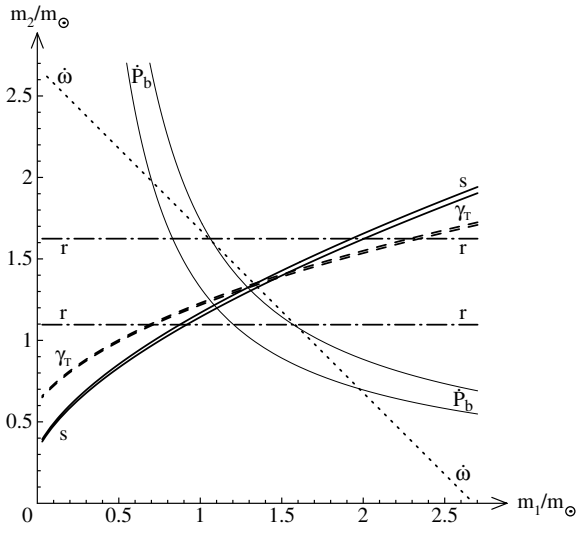


Figure 2. General relativity passes the $(\gamma_T$–$\dot{\omega}$–s–$r)_{1534+12}$ tests. The widths of the strips correspond to 1σ standard deviations. In particular, the region between the two lines labeled r is consistent with the observed value of this parameter.

in one point (corresponding usually to values of m_1 and m_2 which are different from those obtained in general relativity). However, some fine tuning is necessary to construct such alternative theories passing the test.

On the other hand, this binary-pulsar test is extremely constraining if one considers only the most natural and best-motivated alternatives to general relativity: 'scalar–tensor' theories, in which the gravitational interaction is mediated not only by a spin-2 graviton, as in Einstein's theory, but also by one or several spin-0 scalar fields carrying positive energy. Indeed, a neutron star has a small radius R, and its gravitational binding energy is proportional to $Gm/Rc^2 \approx 0.2$ (whereas the same ratio is of order 10^{-6} for the Sun). Because of nonperturbative strong-field effects, scalar–tensor theories can predict large deviations from general relativity in systems involving such compact bodies, even if they are very close to it in the weak-field conditions of the solar system. It has been shown that the PSR B1913 + 16 test rules out a wide class of these theories, including a subclass which is strictly indistinguishable from general relativity in solar system experiments. Binary-pulsar tests are thus qualitatively different from weak-field experiments.

To illustrate this difference, it is convenient to work within a generic class of scalar–tensor theories depending only on the two 'parametrized post-Newtonian' parameters β and γ (see GENERAL RELATIVITY: SOLAR SYSTEM EXPERIMENTS; the parameter γ should not be confused with the timing observable γ_T discussed above). Solar system experiments tell us that $\beta - 1$ and $\gamma - 1$ are both consistent with zero, whereas the PSR B1913 + 16 test imposes a bound on their ratio: $(\beta - 1)/(\gamma - 1) < 1.1$.

PSR B1534 + 12

The binary pulsar PSR B1534+12 is also a rapid ($P_b \approx 10$ h) and eccentric ($e \approx 0.3$) system. It has been observed only since 1991, but it is much closer to the Earth than PSR B1913 + 16, and its pulses are both stronger and narrower. It has thus been possible to measure rapidly the post-Keplerian parameters γ_T and $\dot{\omega}$ with good precision. Moreover, the orbit is nearly edge on relative to the line of sight ($i \approx 80^\circ$), and the Shapiro time delay parameters r and s have also been determined (r with a lower precision). As shown in figure 2, the strips corresponding to these four quantities are nicely consistent within general relativity.

The time derivative of the orbital period $\dot{P}_b$ is also measured at present, but it is spoiled by a poorly known Doppler contribution due to the acceleration of the system towards the center of the Galaxy. The corresponding strip is consistent with the other four observables at the 1.3σ level, but it can be brought in even better agreement if the Earth–pulsar distance is slightly larger than its estimate from dispersion measures.

Out of these $5-2 = 3$ tests, only the one corresponding to the three parameters γ_T, $\dot{\omega}$ and s is very precise. Contrary to the PSR B1913 + 16 test, it does not use the $\dot{P}_b$ observable, and therefore does not depend on the radiative structure of the theory. These two binary pulsars give thus two complementary tests of relativistic gravity. For instance, if one considers a fine-tuned theory passing one of these tests while being very different from general relativity, it usually does not pass the second one.

On the other hand, in the most natural class of scalar–tensor theories, the PSR B1534 + 12 test has also the capability of probing models which are perturbatively equivalent to general relativity in weak-field conditions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

At present, this test is slightly less constraining than the PSR B1913 + 16 one.

'Null' tests of general relativity

For many binary pulsars, it has not yet been possible to measure enough post-Keplerian parameters, and the masses of the pulsar and its companion cannot be determined as in figures 1 and 2. However, a particular combination of them and the inclination angle i, called the 'mass function', is related through Kepler's third law to the observables P_b and x:

$$\frac{(m_2 \sin i)^3}{(m_1 + m_2)^2} = \left(\frac{2\pi}{P_b}\right)^2 \frac{(xc)^3}{G}.$$

Moreover, all the neutron star masses that have been determined precisely happen to be close to $1.35 m_\odot$, with a small statistical spread. Since a pulsar is a neutron star, one can use a statistical argument on its mass m_1 and on the inclination angle i to predict a probable value for the companion's mass m_2. Usually, this information suffices to determine whether the companion may be itself a neutron star or not. When it is very probably a white dwarf, the asymmetry of the system can be used to test quite precisely some predictions of gravity theories. Actually, general relativity does not predict special effects for such asymmetric systems, and this is why they can be called 'null' tests. However, they happen to be very constraining for alternative theories of gravity.

Dipolar gravitational waves

General relativity predicts that the gravitational waves emitted by a binary system start at the quadrupolar order. If v denotes the average orbital velocity of the bodies, this means the radiated energy is of order $O(v^5/c^5)$. On the contrary, alternative theories of gravity generically predict also the emission of dipolar gravitational waves, of order $O(v^3/c^3)$. This emission almost vanishes if the two bodies are similar, but it can become much larger than the general relativistic quadrupole in asymmetric systems. Several NEUTRON STAR–WHITE DWARF binaries, such as PSR B0655 + 64, happen to have a very small observed value of $\dot{P}_b$, consistent with general relativity ($\sim -10^{-14}$) but not with a typical dipolar radiation ($\sim -10^{-11}$). This is one of the most constraining tests for alternative theories of gravity, even if one chooses the masses of the pulsar and its companion to obtain the most conservative bounds. For the most natural scalar–tensor theories of gravity, this test is almost as constraining as the PSR B1913 + 16 one discussed above.

Gravitational Stark effect

General relativity predicts that different self-gravitating bodies fall with the same acceleration in an external gravitational field (see GENERAL RELATIVITY AND GRAVITATION). On the contrary, alternative theories generically predict a violation of this 'strong equivalence principle': the acceleration of a neutron star towards the center of the Galaxy differs from that of a white dwarf. This causes a polarization of the orbit of a neutron star–white dwarf system towards the Galaxy center, analogous to the Stark effect in electromagnetism. (In the case of the Earth–Moon system in the gravitational field of the Sun, this is called the Nordtvedt effect; see GENERAL RELATIVITY: SOLAR SYSTEM EXPERIMENTS.) More precisely, the eccentricity vector $\boldsymbol{e}$ of the orbit is the sum of a fixed vector $\boldsymbol{e}_F$ directed towards the Galaxy center (proportional to the difference of the accelerations of the bodies) and of a rotating vector $\boldsymbol{e}_R(t)$ corresponding to the usual relativistic periastron advance at angular velocity $\dot{\omega}_R$. Several asymmetric systems of this kind (such as PSRs 1713 + 0747, 2229 + 2643, 1455 − 3330) happen to have a very small eccentricity. The only explanation would be that the rotating vector $\boldsymbol{e}_R(t)$ is precisely canceling the fixed contribution $\boldsymbol{e}_F$ at the time of our observation: $\boldsymbol{e}_F + \boldsymbol{e}_R(t) \approx \boldsymbol{0}$. However, this is very improbable if the system is old enough, and one can use a statistical argument to constrain the space of theories. Moreover, by considering several such systems, the probability that they have simultaneously a small eccentricity is the product of the already small individual probabilities. If m_g denotes the gravitational mass of a neutron star and m_i its inertial mass, this statistical analysis gives the bound $|1 - m_g/m_i| < 10^{-2}$. This rules out a wide class of alternative theories of gravity. For the most natural scalar–tensor theories, this test is, however, less constraining than the ones discussed above, because it is less sensitive to strong-field effects: if such a theory is close to general relativity in the solar system, it also predicts that $m_g \approx m_i$ for neutron stars.

Local Lorentz invariance and conservation laws

Theories of gravity which do not respect the symmetries of general relativity can also be tested very precisely with binary pulsars. If local Lorentz invariance of gravitational physics is not respected, there exists a preferred frame in the universe, and a binary system undergoes extra forces if it is moving with respect to this frame. This causes again a polarization of the orbit in a particular direction, as in the previous section, and statistical arguments can again be used to constrain the theories. In the 'parametrized post-Newtonian' formalism, the various possible symmetry violations of general relativity have been parametrized by a set of real numbers. At present, PSRs 1012 + 5307 and 2317 + 1439 give the tightest bounds on one of them: $|\alpha_1| < 1.2 \times 10^{-4}$. The statistical analysis of several systems also gives a very impressive bound on another post-Newtonian parameter: $|\alpha_3| < 1.5 \times 10^{-19}$. This is one of the most precise null experiments of physics. The remaining parameters are currently better constrained by solar-system experiments or isolated pulsar observations.

If the gravitational constant G varies with time, the orbital evolution of a binary system is modified, and one can prove that $\dot{P}_b/P_b$ has an extra contribution proportional to $-\dot{G}/G$. Binary-pulsar data can thus be used to constrain also the time derivative of G, but solar system experiments give currently tighter bounds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The most complete reference about the phenomenological analysis of binary-pulsar data is

Damour T and Taylor J H 1992 *Phys. Rev.* D **45** 1840–68

A detailed description of the PSR B1913 + 16 test, together with shorter ones on other binary pulsars, can be found in chapters 12 and 14.6 of

Will C M 1993 *Theory and Experiment in Gravitational Physics* (Cambridge: Cambridge University Press) pp 283–309 and 343–52

More recent data about PSRs 1913 + 16 and 1534 + 12 are given in

Taylor J H 1993 *Class. Quantum Gravity* **10** S167–74
Stairs I H *et al* 1998 *Astrophys. J.* **505** 352–7

The gravitational Stark effect and other 'null' tests of gravity are discussed notably in

Wex N 1997 *Astron. Astrophys.* **317** 976
Wex N 2000 *Pulsar Astronomy—2000 and Beyond (ASP Conf. Ser. 202)* eds Kramer, Wex and Wielebinski

The various experimental constraints on a generic class of alternative theories of gravity are compared in

Damour T and Esposito-Farèse G 1998 *Phys. Rev.* D **58** 042001

Gilles Esposito-Farèse

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

General Relativity: Solar System Experiments

Modern gravitational theories such as general relativity produce, among their outputs, dynamical equations for clock rates, for the speed of light propagation and for the motion of bodies, with each of these equations containing relativistic modifications of their Newtonian versions. Such dynamical equations are the chief ingredients for modeling the key solar system experiments which essentially 'map out' the gravitational interaction. The GENERAL THEORY OF RELATIVITY predicts novel phenomena not present in Newtonian gravity and which also differ in detail from what emerges from plausible alternative relativistic theories of gravity. The experimental mapping of the gravitational interaction strengthens verification of general relativity as one theory within a broad class of possible gravitational theories and from that verification places stringent constraints on the existence of any additional long-range force fields in physical law which might supplement the tensor interaction of general relativity. (See also GENERAL RELATIVITY AND GRAVITATION.)

Equivalence principle

Einstein's first revolutionary innovation in gravitational physics was based on the Equivalence Principle which he formulated in 1907, a decade before construction of his specific general relativity theory. Generalizing from the apparent fact that differently composed bodies fall at identical rates in a gravitational field, Einstein hypothesized that gravitational phenomena were, when tidal effects could be neglected, locally equivalent in all respects to phenomena experienced in an accelerated frame of reference. From this equivalence he concluded that clock rates should be reduced near gravitating matter when compared with rates of more distantly located clocks. Applying universally to all clocks located at heights z and $z+h$ in a gravitational field $g(z)$, the fractional rate shift between them is (to lowest order in $1/c^2$) expected to be

$$\frac{f(z+h)-f(z)}{f(z)} \simeq \frac{g(z)h}{c^2} \tag{1}$$

or, for large clock separations, the rate shift is better stated in terms of the cumulative difference between Newtonian gravitational potentials of the clocks:

$$\frac{f(\vec{r}_i)-f(\vec{r}_j)}{f(\vec{r}_i)} \simeq \frac{U(\vec{r}_i)-U(\vec{r}_j)}{c^2}. \tag{2}$$

This frequency shift was subsequently shown, however, to simply follow from energy conservation considerations, so its observation does not so much confirm general relativity theory specifically as support a broader, often called *metric*, class of gravitational theories. In 1977 a hydrogen maser clock was put on a suborbital rocket flight, and its rate was compared with an Earth-located clock by means of one- and two-way phase-coherent radio links, and the shift given by equation (2) was confirmed with precision of almost 1 part in 10^4.

Using his Equivalence Principle, Einstein also predicted that light would be deflected by the component of gravitational acceleration perpendicular to the light trajectory

$$\frac{\mathrm{d}\hat{c}}{\mathrm{d}s} = \frac{\vec{g}-\vec{g}\cdot\hat{c}\hat{c}}{c^2} \tag{3}$$

with $\hat{c}$ being the light ray's unit propagation vector. However, this was a less dramatic prediction; a century earlier Laplace and others had reached such a conclusion based on a finite velocity, corpuscular model of light. The real novelty concerning this phenomenon later emerged from Einstein's complete theory of 1916 which he based on a tensor field of gravitational interaction; it predicts twice as large a rate of deflection. Solar system experiments which have specifically focused on measuring the effects of gravity on propagation of light yield some of the most precise tests of the general relativity theory available today, since they have been able to measure the dynamics of light rays which passed close to the Sun where the solar system's strongest gravitational fields reside.

Theories consistent with the traditional form of the Equivalence Principle can be constructed using a combination of tensor and scalar fields of interaction. In these constructions, the static-limit source for the tensor field is proportional to matter's total energy density T_0^0 (which is the normally dominant component of matter's entire *stress–energy tensor* T_μ^ν), while the source for the scalar field is proportional to the trace of that tensor T. The body integral of these two sources can be shown to be identical for a general source body which is in stationary internal equilibrium, is free of significant external force disturbances and has negligible internal gravitational binding energy:

$$\int T_0^0\,\mathrm{d}V = \int T\,\mathrm{d}V \equiv Mc^2 \qquad \text{with } T \equiv \sum_\mu T_\mu^\mu. \tag{4}$$

Such conditions are rather closely fulfilled by the sources in the solar system. In such scalar–tensor theories, therefore, the observational differences between the hybrid scalar–tensor interaction and the pure tensor interaction of general relativity are suppressed until the solar system's very weak relativistic corrections to the static limit are relevant.

Dynamical equations for light and bodies

There is a rich collection of perturbations in the orbits of solar system bodies which theory suggests may result when Newtonian theory is generally modified but which are not, in fact, observed in experiment. These often overlooked *null experimental results* are nevertheless important in establishing several properties of gravity, e.g. (1) that local gravitational physics is Lorentz invariant so that the speed of the solar system through the cosmos cannot be detected from the dynamics within that system, (2) that conservation laws of energy, momentum and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

angular momentum are fulfilled in the solar system's total motion and (3) that the Newtonian-level interaction is spatially isotropic.

Accepting those foundations and then considering the class of theoretical possibilities consistent with these empirical findings, some general dynamical equations for light propagation and for motion of bodies are formed which represent the broad class of theoretical alternatives. These equations, which are then used to model some remaining key confrontations of general relativity and its alternatives with experiment, take forms as follows (expressed here to order $1/c^2$ which is sufficient for solar system applications): the coordinate speed of light near the Sun or other body (mass M and distance R) is slowed, as measured from afar, owing to the body's proximity

$$c(\vec{R}) = c_\infty \left(1 - (1+\gamma)\frac{GM}{c^2 R} + \cdots\right). \qquad (5)$$

γ is a dimensionless, theory-dependent parameter (equal to 1 in general relativity) which was first introduced by Eddington to represent general theoretical possibilities for the static, spherically symmetric tensor field exterior to a source, and which was later found by Will and Nordtvedt to have more wide-ranging significance for the general structure of the gravitational field. Here it scales the additional effect of gravity on light propagation which is not derivable from the Equivalence Principle. This location-dependent speed function leads to both deflection and retardation of light trajectories. (The equations of motion for matter and light depend on choice of coordinates for space and time. For solar system applications, coordinates in which the speed of light is isotropic to $1/c^2$ order are usually employed. Experimental observables must, of course, be independent of this choice of coordinate system.) For a ray making a complete passage by the Sun with distance of closest approach D, it produces a deflection $\delta\theta$ of light's direction of travel toward the Sun by an amount

$$\delta\theta = 2(1+\gamma)\frac{GM}{c^2 D} \qquad (6)$$

and it adds an amount δt to light's time of flight between two sites located at distances R_1 and R_2 from the Sun

$$\delta t \simeq 2(1+\gamma)\frac{GM}{c^3}\ln(4R_1R_2/D^2) \qquad (7)$$

with D again being the ray's distance of closest approach to the Sun.

Changes of the apparent angular positions of cosmologically distant radio sources, numbering in the many thousands, have been measured using very-long-baseline interferometry (VLBI) techniques in which the signals from each source recorded at two or more spatially separated radio antennas are cross-correlated as a function of time delay; the general relativistic prediction for the light deflection—double that given by the equivalence principle—has recently been confirmed by this means with precision of better than 1 part in 10^3. The time-of-flight delays of radio signals were earlier confirmed with slightly less precision using the Viking mission's ranging data to Mars accumulated during 1977–1982. Radar transponders placed on the Martian surface and on spacecraft orbiting that planet permitted measurement of the round trip time of travel for radio signals originated and detected at Earth, with the time delay signal in equation (7) being particularly important in the data produced during the periods when the Earth–Mars line of sight passed close to the Sun.

The equation of motion for planets such as Mercury which move with respect to an essentially fixed Sun is, to sufficient approximation, given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(1+\frac{1}{2}\frac{V^2}{c^2}\right)\vec{V} = -\frac{GM}{R^3}\vec{R}\left(1+\frac{2\gamma+1}{2}\frac{V^2}{c^2}\right) - (2\gamma+1)\frac{\mathrm{d}}{\mathrm{d}t}\frac{GM}{c^2 R}\vec{V} - \frac{1-2\beta}{2}\frac{G^2M^2}{c^2R^4}\vec{R} + \cdots \qquad (8)$$

with $\vec{V} = \mathrm{d}\vec{R}/\mathrm{d}t$. Eddington's γ parameter appears in this equation as a scale factor for the motional corrections to Newton's static-limit interaction, while his second parameter β scales the novel nonlinear feature of gravity. β also has value unity in general relativity, but both parameters will generally have different values in alternative theories which have a scalar field augmenting Einstein's tensor field in transmitting the gravitational interaction between matter. In these latter cases $1-\gamma$ quantifies the fractional participation or strength of the scalar field in producing the total phenomenon of gravity.

The expression from SPECIAL RELATIVITY for a body's momentum on the left-hand side of equation (8) (approximated to the $1/c^2$ order) plus both the motional and nonlinear modifications scaled by γ and β on the force side of this equation of motion together produce a precession of the Keplerian-like orbit around the central body of mass M. The precessional rate is

$$\frac{\delta\Omega}{\Omega} = (2+2\gamma-\beta)\frac{GM}{c^2(1-e^2)a} \qquad (9)$$

with Ω, e and a being the orbit's frequency, eccentricity and semimajor axis, respectively. For the general relativistic case, $\gamma = \beta = 1$, this precession amounts to 43 arcsec/century for Mercury's orbit and adds to a Newtonian precession of that planet's orbit caused by accelerations from other planets. A discrepancy of this size between predictions of Newtonian theory and the astronomical observations of Mercury's angular positions had already been established by the end of the 19th century, and the natural remedy for this anomaly was the first and immediate empirical triumph of Einstein's general relativity theory in 1916. Today, radar ranging between Mercury's surface and Earth, a somewhat less precise procedure than the ranging to specific transponder sites at Mars in the Viking mission, nevertheless confirms the predicted precession to about 1 part in 10^3 precision. This

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

serves at this time as a somewhat weaker test of general relativity than do observations of light propagation or lunar laser ranging, and further improvements in ranging precision to Mercury will not so much yield a better test of gravitational theory but will instead measure the Sun's quadrupole moment (a measure of the Sun's rotational flattening) whose Newtonian gravitational field also contributes to Mercury's orbital precession.

Revisiting a classic system—the orbit of the Moon—which has for centuries played a central role in studying details of Newtonian gravitational dynamics, it is found, thanks to implementation of the modern laser ranging technique, to also be a remarkably sensitive testing ground for measuring relativistic features of gravity. To new levels of precision the Equivalence Principle can be tested, the Eddington parameters, especially β, can be measured in novel ways and to high precision, and a sensitive search for time variation in the gravitational coupling parameter G can be made. The scientifically operable part of the Moon's equation of motion relative to Earth is remarkably Newtonian like:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = -G(t)\frac{m}{r^3}\vec{r} + \delta_{\mathrm{em}}\frac{GM}{R^3}\vec{R} + (2\gamma+1)\frac{GM}{c^2R^3}(\vec{R}\times\vec{V})\times\vec{v} + \cdots \qquad (10)$$

with m being the total Earth–Moon system mass, $\vec{v} = \mathrm{d}\vec{r}/\mathrm{d}t$ and $+\cdots$ indicating a host of Newtonian solar tidal accelerations and small relativistic accelerations of lesser interest for our purposes here. δ_{em} is the measure of a possible failure, from whatever cause, of equality of gravitational free-fall rates of Earth and Moon toward the Sun (breakdown of the previously mentioned Equivalence Principle of Einstein):

$$\delta_{\mathrm{em}} = \left|\frac{\vec{a}_{\mathrm{Earth}} - \vec{a}_{\mathrm{Moon}}}{\vec{g}_{\mathrm{Sun}}}\right|. \qquad (11)$$

This difference between free-fall rates can occur as a result of an unusual atomic composition dependence of long-range forces (discussed later) and it can also result from anomalous coupling between the external gravity of the Sun and the internal gravity within the Earth and Moon. The latter possibility is controlled by the same Eddington parameters γ and β introduced previously in equations (5) and (8); as auxiliary calculations show, the gravitational-to-inertial mass ratios of celestial bodies must differ among themselves in proportion to the fractional gravitational internal binding energies of the bodies if the Eddington parameters are anomalous:

$$\frac{|\vec{a}|}{|\vec{g}|} = \frac{M(G)}{M(I)} = 1-(4\beta-3-\gamma)\frac{1}{2Mc^2}\int\frac{G\rho(\vec{x})\rho(\vec{x'})}{|\vec{x}-\vec{x'}|}\,\mathrm{d}^3x\,\mathrm{d}^3x'. \qquad (12)$$

$\vec{a}$ is the acceleration of the body when in external gravitational field $\vec{g}$: $\rho(\vec{x})$ is the body's mass density, so the exhibited integral represents minus twice the gravitational potential energy within the body.

Universality of gravitational free fall

Thirty years of laser ranging to several passive reflectors placed on the lunar surface by some of the Apollo missions to the Moon have confirmed that the Earth and Moon fall at equal rates toward the Sun with a precision of 3×10^{-13} (a recent 1998 fit has even pushed this equality to 1×10^{-13}). The Earth has a substantial iron core while the Moon is composed of mantle-like silicate materials, so the lunar laser ranging (LLR) result confirms the composition independence of gravitational free fall to precision better than 10^{-12}, which compares quite favorably with the performance of the latest laboratory tests of the universality of gravitational free-fall rates for different materials. However, because the Earth possesses gravitational binding energy equal to about 5×10^{-10} of its total mass–energy, this experimental result permits a second interesting scientific interpretation; the factor $4\beta - 3 - \gamma$ in equation (12) is zero to precision of a few parts in 10^4, serving as the premier measurement of general relativity's nonlinearity quantified by the parameter β. This confirmation of general relativity is accomplished by measuring, in particular, the amplitude of the synodic month period (new moon to new moon) oscillation in Earth–Moon range with precision of a few millimeters; if δ_{em} were non-zero in equation (10), the anomaly in this synodic amplitude would be $2.9 \times 10^{12}\delta_{\mathrm{em}}$ cm.

Time variation of Newton's G

From analysis of the LLR data one also tests for a possible time variation in the gravity coupling parameter of Newton—$G \rightarrow G(t)$? There have been general speculations by Dirac and others that this coupling parameter might evolve in time as the universe expands, and theories alternative to general relativity such as the scalar–tensor theories usually predict such time variation with magnitude of the order of the Hubble rate:

$$\frac{\mathrm{d}G}{G\,\mathrm{d}t} \sim (4\beta - 3 - \gamma)H. \qquad (13)$$

$H \sim 7 \times 10^{-11}$ yr^{-1} is the observed Hubble expansion rate of the universe. The fit of the LLR data places the limit $\mathrm{d}G/G\,\mathrm{d}t \leq 2 \times 10^{-12}$ yr^{-1} on such variation (although a conservative allowance for possible unmodeled systematic secular effects is added by some to this result). Comparable limitations have been obtained by analysis of the radar ranging data between Earth and the planets Mars and Mercury and the arrival times of the pulsed signals from certain binary pulsar systems (see GENERAL RELATIVITY TESTS: BINARY PULSARS).

Geodetic precession of local inertial frames

The Coriolis-like acceleration term in the Moon's equation of motion is the basis of a third fundamental confirmation of general relativistic predictions achieved from LLR. Because the Earth–Moon system is not simply falling radially toward the Sun, but rather moves across the Sun's field, the local inertial frame rotates relative to distant inertial frames of reference at the average rate

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of 19.2 mas yr^{-1} (mas = milliarcsecond), as first derived by de Sitter in the context of general relativity, but later generalized to other metric theories to be

$$\vec{\Omega}_{\rm dS} \;=\; (\gamma + 1/2)\frac{GM}{c^2 R^3}\,\vec{R}\times\vec{V}. \qquad (14)$$

As a result of this acceleration, both the lunar orbit's major axis (its perigee) and its inclined orbital plane (line of nodes) precess at about this additional rate when referenced to the distant inertial frames. LLR confirms this precession to better than 0.1 mas yr^{-1} precision. NASA's GRAVITY PROBE-B mission, scheduled for launch in May 2002, will place gyroscopes in Earth orbit and measure their precessions with respect to the line of sight toward a distant star. This mission has the potential to measure the analogous de Sitter precession of the gyroscopes due to motion through Earth's gravitational field and to do so with precision approaching 1 part in 10^5, thereby determining the fundamental Eddington parameter γ appearing in equation (14) with a level of precision which goes well beyond present knowledge.

LLR is an on-going experiment, and the performance of the passive reflectors on the Moon shows no sign of diminishing with age. The three gravitational measurements here discussed can be further improved in precision as we go into the future, especially the latter two measurements whose intrinsic sensitivities grow decisively with total time span of observing the lunar orbit.

Gravitomagnetism

The full, testable content of the $1/c^2$ corrections to the Newtonian equations of motion only appears in situations of bodies in mutual motion. As in electromagnetism, these more general N-body dynamical configurations include what have become called *gravitomagnetic* forces, which are proportional to velocities of both interacting bodies, and *inductive* forces, in which the acceleration of one body accelerates the other. Successful modelings of astronomical binary pulsar systems, such as PSR 1913+16, which consist of bodies of comparable mass moving about their joint center of mass–energy in highly eccentric orbits, have necessarily brought these additional interactions into play as part of the modeling process which achieves precision fit of the pulse-arrival-time data. These interactions are particularly required for understanding the observed 4.2° yr^{-1} periastron precession rate of the PSR 1913+16 system's orbit. Closer to home, the computer computation of the Moon's position relative to Earth is routinely performed in the solar system's barycentric rest frame, and therefore the gravitomagnetic and inductive accelerations of the moving Moon by the moving Earth are indispensable components of that computation, which as a whole achieves the recent fits of the range data with precision at the subcentimeter level. Discrepancies of more than 1000 cm would otherwise be present. In fact, once local Lorentz invariance and fulfillment of the several conservation laws in gravitational physics have been experimentally confirmed, as they have, then such interactions as gravitomagnetism are necessarily specified in terms of the already discussed Eddington parameter γ.

Nevertheless, explicit and isolated experimental expression of the gravitomagnetic interaction has been sought. Marginal detection of this force has been achieved through measurement and modeling of the nodal precession rates of the near-Earth LAGEOS satellites, and the previously mentioned Gravity Probe-B mission is designed to also measure to better than 1% the precession of gyroscopes due to the gravitomagnetic torque generated by the spinning Earth.

The gravitational radiation reaction force in the binary pulsar

The binary pulsar system PSR1913+16 has also provided the pioneering, although indirect, evidence that the tensor gravitational field, like other dynamical fields, carries waves of energy, momentum and angular momentum away from its time-varying gravitational sources. Imprinted in the pulse arrival time data is, with better than 1% agreement with theory, the orbital rate speed-up which results from the order $1/c^5$ radiation reaction force derived from general relativity theory. Both ground and space-based laboratories for directly detecting and studying gravitational radiation arriving from distant astronomical sources are at present under development and construction.

Equivalence Principle violating interactions

A failure of the Equivalence Principle at some level of precision would probably be indicating the presence of an additional very weak, long-range force acting between bodies, and for which the coupling strength to sources is proportional to attributes of those sources other than total mass–energy content. The static-limit force between bulk matter resulting from most such additional interactions will then be a gradient of a potential with the form

$$\vec{F}_i = K_i\vec{\nabla}_i\sum_{j\neq i}\frac{K_j\,{\rm e}^{-\mu R_{ij}}}{R_{ij}}. \qquad (15)$$

$\mu \neq 0$ indicates the possibility of a finite-range, Yukawa-like interaction due to non-zero mass $m = \hbar\mu/c$ of the transmitting field's quanta ($\mu \to 0$ returns this interaction to inverse square). This force will produce differences in the acceleration rates of Earth and Moon toward the Sun:

$$\delta_{\rm em} = \left(\frac{K}{GM}\right)_{\rm Sun}\left[\left(\frac{K}{M}\right)_{\rm e} - \left(\frac{K}{M}\right)_{\rm m}\right](1+\mu R)\,{\rm e}^{-\mu R}. \quad (16)$$

If, for example, the coupling strengths K_i are proportional to matter attributes such as baryon number or nuclear electrostatic energy content, then the ratios K_i/M_i vary fractionally by about a part in a thousand among the different elements of nature. The mentioned experiments which have confirmed composition independence of free-fall rates to 10^{-12} then place strength limits of order $K^2/GM^2 \leq 10^{-9}$ on the long-range limit $1/\mu > 1$ AU

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for such interactions. Experiments which exploit the advantages of an Earth-orbiting satellite environment in order to achieve several orders of magnitude higher precision in testing the universality of free fall are under development.

For any atomic clock whose frequency f is determined by the difference of two energy levels

$$f = \frac{E_i - E_j}{2\pi\hbar} \tag{17}$$

the potential energy assumed in equation (15) produces, in the vicinity of source body s, an additional clock frequency shift which depends on details of the clock's atomic mechanism; adding this to the previously discussed Equivalence Principle's shift, the total rate shift δf of this particular clock when near a source body s becomes

$$\frac{\delta f}{f} = \frac{GM_s}{c^2 R}\left[1 + \left(\frac{K}{GM}\right)_s \frac{K_i - K_j}{M_i - M_j}\,\mathrm{e}^{-\mu R}\right]. \tag{18}$$

However, GM_s in this expression can no longer be identified with the Sun's total attracting strength; in interaction with planet p it is now

$$\Gamma = GM_s\left(1 + \frac{K_s}{GM_s}\frac{K_p}{M_p}\,\mathrm{e}^{-\mu R}(1+\mu R)\right). \tag{19}$$

The Sun's total attracting strength Γ is the number measured to better than 1 part in 10^{10} precision in the solar system and is the standard by which clock gravitational rate shifts can be absolutely calibrated. From the above, it is then seen that clocks will in general show anomalous rate shifts both *absolutely* measurable with respect to a standard and to a differently located clock and *relatively and locally* measurable with respect to clocks of other physical mechanisms.

The difference between the expressions in equations (16) and (18) confirms that clock experiments and free-fall experiments are, in principle, independent probes of the Equivalence Principle. However, comparing these equations quantitatively reveals that the former can rarely compete in precision as fundamental tests. For example, an on-board measurement of the relative rates of two physically different atomic clocks sent together on a spacecraft to within 4 solar radii of the Sun could probe for an Equivalence Principle violating interaction with only slightly higher precision than today's best free fall observations already achieve.

Periastron precession and finite-range interactions

If the force given by equation (15) has finite Yukawa range ($\mu \neq 0$), it contributes to the precession rate of orbit major axes. In the case of orbits with small eccentricity this added precession $\delta\omega$ is

$$\frac{\delta\omega}{\omega} = \frac{K_i K_j}{2GM_i M_j}(\mu R)^2\,\mathrm{e}^{-\mu R}. \tag{20}$$

Near-Earth satellite, lunar and planetary orbit precession rates as measured by laser and radio ranging all strongly constrain the presence of any such additional interactions with the appropriately matched ranges. For instance, LLR constrains the strength ratio $K_\mathrm{e}K_\mathrm{m}/GM_\mathrm{e}M_\mathrm{m}$ of any new interaction to be less than 10^{-11} for range $1/\mu \simeq 2\times10^{10}$ cm, and similar constraints result from using data from near-Earth satellite and planetary orbits. These powerful results hold whether or not the coupling parameters K_i of the additional interaction are proportional to total mass–energies M_i. Periastron precession rate measurements have been underappreciated observations in the search for long-range Yukawa-like interactions, and they deserve more attention in the design of future ranging experiments.

Concluding remarks

During most of the period since publication of the general relativity theory of gravity, the prevailing interpretation of Einstein's revolutionary theory has not been the measured and restrained observation—that gravity modifies physical phenomena *as if* the geometry of space and time is non-Euclidean—but instead it has generally been the stronger ontological assertion—that gravity *is* space–time curvature. Experiment provides no support for this stronger interpretative stance, it played no part in this summary of the present-day experimental situation and indeed it tends to be a divisive viewpoint which suggests a difference in nature between the gravitational force and the other forces in physical law. The inverse assertion better describes the actual relationship—that the curved SPACE–TIME arena of our world is a consequence of the gravitational field.

We are left at century's end with a wide variety of experimental evidence, which considered as a whole indicates that a dynamical tensor field, governed by specific field equations and having a unique universal coupling to the other fields and particles of physical law, and nonlinearly to itself, all as formulated by Einstein's general relativity, provides the currently measured long-range interaction between bodies. The empirically confirmed, N-body gravitational equation of motion, applicable and valid for the most general dynamical configurations, is found to be Lorentz invariant and Lagrangian based

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \vec{v}_i} - \frac{\partial L}{\partial \vec{r}_i} = 0 \tag{21}$$

with the $1/c^2$ order Lagrangian being

$$L = \sum_i m_i\left(-c^2 + \frac{1}{2}v_i^2\right) + \frac{G}{2}\sum_{i,j\neq i}\frac{m_i m_j}{r_{ij}} \tag{22}$$

$$+ \frac{1}{8c^2}\sum_i m_i v_i^4 + \frac{G}{4c^2}\sum_{i,j\neq i}\frac{m_i m_j}{r_{ij}}$$

$$\times\left(3v_i^2 + 3v_j^2 - 7\vec{v}_i\cdot\vec{v}_j - \vec{v}_i\cdot\hat{r}_{ij}\hat{r}_{ij}\cdot\vec{v}_j - 2G\sum_{k\neq i}\frac{m_k}{r_{ik}}\right) \tag{23}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and possessing the conserved system properties of energy E, momentum $\vec{P}$ and angular momentum $\vec{J}$

$$E = \sum_i \vec{v}_i \cdot \frac{\partial L}{\partial \vec{v}_i} - L \qquad \vec{P} = \sum_i \frac{\partial L}{\partial \vec{v}_i} \qquad \vec{J} = \sum_i \vec{r}_i \times \frac{\partial L}{\partial \vec{v}_i} \tag{24}$$

and a system center of energy $\vec{R}$

$$\vec{R} = \frac{1}{E} \sum_i m_i \left(c^2 + \frac{1}{2} v_i^2 - \frac{G}{2} \sum_{j \neq i} \frac{m_j}{r_{ij}} \right) \vec{r}_i \tag{25}$$

which moves at constant velocity $\vec{P}/E$.

From solar system observations, there appears room for less than a part in a thousand presence of any additional inverse square, metrically structured interaction and room for less than 1 part in 10^9 presence of any inverse square, non-metrically structured interaction (and less than 1 part in 10^{11} Yukawa-like, solar system range interaction of either structure) to modify this exhibited dynamics derivable from general relativity theory.

Bibliography

The most comprehensive book on the interface of theory with the experimental foundations of the gravitational interaction is

Will C M 1993 *Theory and Experiment in Gravitational Physics* revised edn (Cambridge: Cambridge University Press)

although the reader should be cautioned that decisive improvements in experimental results and other innovations have occurred since publication of this edition.

A general article on the history of the Moon's role in probing gravity, including the modern lunar laser ranging experiment, is

Nordtvedt K 1996 From Newton's Moon to Einstein's Moon *Phys. Today* **49**(5)

The most recent comprehensive overview of the lunar laser ranging experiment including its total scientific output is

Dickey J O *et al* 1994 Lunar laser ranging: a continuing legacy of the Apollo program *Science* **265**

The modern version of the parameterized post-Newtonian expression of the tensor gravitational field used to calculate and compare the various experimental effects within the general class of metric theories of gravity (including general relativity) was developed in

Will C and Nordtvedt K 1972 *Astrophys. J.* **177**

Kenneth Nordtvedt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

General Theory of Relativity

The relativistic theory of gravitation that was devised in 1915 by Albert Einstein (1879–1955). The general theory extended Einstein's previous work on special relativity to deal with observers in accelerated frames of reference.

A central postulate of the general theory is the principle of equivalence, that 'all freely-falling, non-rotating, laboratories are fully equivalent for the performance of all physical experiments'. It follows from this that an observer in a small closed box (e.g. an elevator, or a windowless spacecraft) cannot tell whether that box is subject to uniform acceleration or is at rest in a uniform gravitational field; both situations would give the observer a feeling of 'weight'. Conversely, if a test particle floats freely inside a closed box, an observer cannot tell whether the box is in the depths of space, far from a gravitating body, or falling freely in the gravitational field of a massive body (for example, an observer inside an elevator falling freely down its shaft would have no sensation of weight).

General relativity embodies the concept that the three dimensions of space (length, breadth, height) and the dimension of time are linked together into a four-dimensional spacetime. The effect of a distribution of mass (or energy) is to induce curvature into (i.e. to bend) spacetime in its vicinity. Conversely, particles and rays of light follow paths that are determined by the curvature of spacetime in their locality. In general relativity, gravitation is regarded as a geometric property of spacetime rather than, as in Newtonian gravitation, a force acting directly between individual massive bodies.

Although Newtonian theory is perfectly satisfactory for most applications, general relativity is capable of dealing with circumstances in which Newtonian theory is inadequate. In particular, the perihelion position of the planet Mercury's orbit advances by an amount that exceeds the Newtonian prediction by 43 arcsec per century; general relativity predicts the correct rate of advance. Other important consequences of the theory, which have been tested to a high degree of precision by experiment and observation, include: the bending of light in a gravitational field (rays of light passing close to the edge of the Sun, for example, are deflected by an amount that is consistent with general relativity); gravitational time dilation (clocks run more slowly in strong gravitational fields than in weak ones); and gravitational redshift (light emerging from a strong gravitational field is stretched to longer wavelengths).

General relativity has important applications in many areas of astrophysics and cosmology, including describing the behavior and properties of very close binaries, binary pulsars, black holes, gravitational lensing and the universe as a whole.

See also: binary system, black hole, cosmology, cosmological model, gravitation, gravitational lensing, gravitational waves, redshift, spacetime, special relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Genesis

NASA Discovery mission to collect samples of charged particles in the solar wind. Genesis carries four instruments: solar wind collector arrays, made of materials such as diamond, gold, silicon and sapphire, and designed to entrap solar wind particles; an ion monitor, to record the speed, density, temperature and approximate composition of the solar wind ions; an electron monitor, to make similar measurements of electrons in the solar wind; and an ion concentrator, to separate and focus elements like oxygen and nitrogen in the solar wind into a special collector tile. Isotopes of oxygen, nitrogen, the noble gases and other elements of the solar wind will be returned to Earth in a capsule (planned for September 2004).

At 1903 UT November 16, 2001, the Genesis spacecraft entered into orbit around the balanced-gravity point Lagrange 1. Its instruments became fully operational on December 3, 2001.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geocentric Coordinate System

Coordinates as measured from the center of the Earth.

See also: heliocentric coordinate system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geodynamo

The existence of a large-scale magnetic field in the Earth has been known for over two millennia. The invention of the compass and the observation of the angle between magnetic and true north (declination), followed much later (16th century) by the discovery of inclination (angle of dip), led to the identification of the interior of the Earth as the prime source of magnetic activity. GILBERT believed that the Earth acted as a giant permanent magnet, but the realization that the Earth possesses a liquid metallic core, far above the Curie point, led to a search for generation mechanisms. In spite of the efforts of a number of eminent physicists to promote rival theories the only mechanism now considered viable is dynamo action, caused by electric currents driven in the core by Faraday induction. This article deals with the main theoretical considerations underlying the MAGNETOHYDRODYNAMICS of the core.

Manifestation of the magnetic field

For the purpose of this article, the Earth may be considered as a solid insulator (the mantle), surrounding a metallic core. In fact, the mantle is a conductor, although a poor one. There is some evidence that the region of the mantle nearest the core is relatively highly conducting; nonetheless, the assumption of an insulator is a good one for many purposes. The core itself has an outer, liquid part and an inner solid part (see EARTH'S INTERIOR). It is thought that the inner core is the result of freezing of heavy material from the mixed outer core. The buoyancy generated by the freezing process is an important ingredient of the energy source for the magnetic field, which dissipates energy by Joule heating. Other sources of energy that have been investigated include thermal buoyancy (for example due to internal radioactive heating in the core) and the energy that can be derived from the relative precession of core and mantle. While all these possible mechanisms can in principle provide enough energy to sustain the main field, details of the way in which precessional forces could sustain the dynamo are poorly understood. We can only observe the potential field at the Earth's surface and in the atmosphere, driven by the electric currents in the core (figure 1). Very short timescale phenomena are filtered out by the weakly conducting mantle. The field is relatively weak at the surface of the mantle (less than 1 G) but because of the algebraic decay of potential fields with distance fields in the core are rather stronger. In addition there is the invisible toroidal part of the magnetic field that is confined to the core (see below). The field is predominantly dipolar at the present epoch, but the declination changes in time as has been known since 1634. Many surface features of the field also migrate, mostly to the west, with a typical timescale of 10^2 yr. This is known as the secular variation. In addition we know from studies of remanent magnetism of rocks that the main dipole field has reversed direction many times over the Earth's history, at irregular intervals but with a typical interval of 10^5–10^6 yr (figure 2). Thus the dynamo process that sustains the field must be regarded as a fully dynamic phenomenon, requiring for its understanding a proper description of core dynamics as well as the kinematics of dynamo action.

Kinematic models

In the article on DYNAMO THEORY we have seen that the magnetic field $\boldsymbol{B}$ obeys the induction equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{\nabla} \times (\boldsymbol{v} \times \boldsymbol{B}) - \boldsymbol{\nabla} \times (\eta \boldsymbol{\nabla} \times \boldsymbol{B}) \qquad \boldsymbol{\nabla} \cdot \boldsymbol{B} = 0. \quad (1)$$

Kinematic theories of the geodynamo suppose that the velocity $\boldsymbol{v}$ is given and seek eigenfunctions of the induction equation. Because of the near axisymmetry of the main field, it is natural to seek axisymmetric solutions. This is, however, prohibited by Cowling's theorem (see article on dynamo theory). There is a remedy which preserves the possibility of two-dimensional models. In mean-field (α effect) dynamos the velocity is supposed to exist on two length scales. The inductive effect of the small-scale motions on the large scales can be modelled by the inclusion of an extra term $\boldsymbol{\nabla} \times [\alpha(\boldsymbol{x})\boldsymbol{B}]$ in the induction equation. Because of this extra driving term Cowling's theorem may be avoided and axisymmetric solutions sought. Braginsky has developed another model based on the limit of small diffusivities and rapid zonal flows. Although completely different in spirit from the mean-field approach, the method yields equations that are essentially the same. Models of this type are discussed in the dynamo theory article. Because of the apparent size of the zonal flows as evidenced by the secular variation the more realistic model is of $\alpha\omega$ type, although models of the α^2 variety are more likely to lead to steady solutions. Depending on the choices of α and the differential rotation the preferred (fastest-growing) solutions may either have dipolar or quadrupolar symmetry. In this picture reversals may be explained by a change in the form of internal motion that leads to a change in the preferred symmetry of the solution. However, this quasi-kinematic approach cannot give any information on the dynamic details either of reversals or of secular variation, and so recent work has attempted to introduce a treatment of the fluid motion and the dynamical effects of the Lorentz (magnetic) force.

Core dynamics

The velocity field $\boldsymbol{v}$ can be supposed to obey the momentum equation (expressed in a frame rotating with angular velocity Ω)

$$\frac{D\boldsymbol{v}}{Dt} + 2\Omega \times \boldsymbol{v} = -\boldsymbol{\nabla} P + H\hat{\boldsymbol{r}} + \frac{1}{\mu\rho}(\boldsymbol{\nabla} \times \boldsymbol{B} \times \boldsymbol{B}) + \nu\nabla^2\boldsymbol{v} \quad (2)$$

where ν is the kinematic viscosity and ρ is the density (supposed constant, although this is a simplification). H is a source of buoyancy either due to concentration gradients associated with inner-core freezing or due to internal heat sources. This equation is to be solved together with the induction equation above and an equation for H, to give a dynamically self-consistent picture of the dynamo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

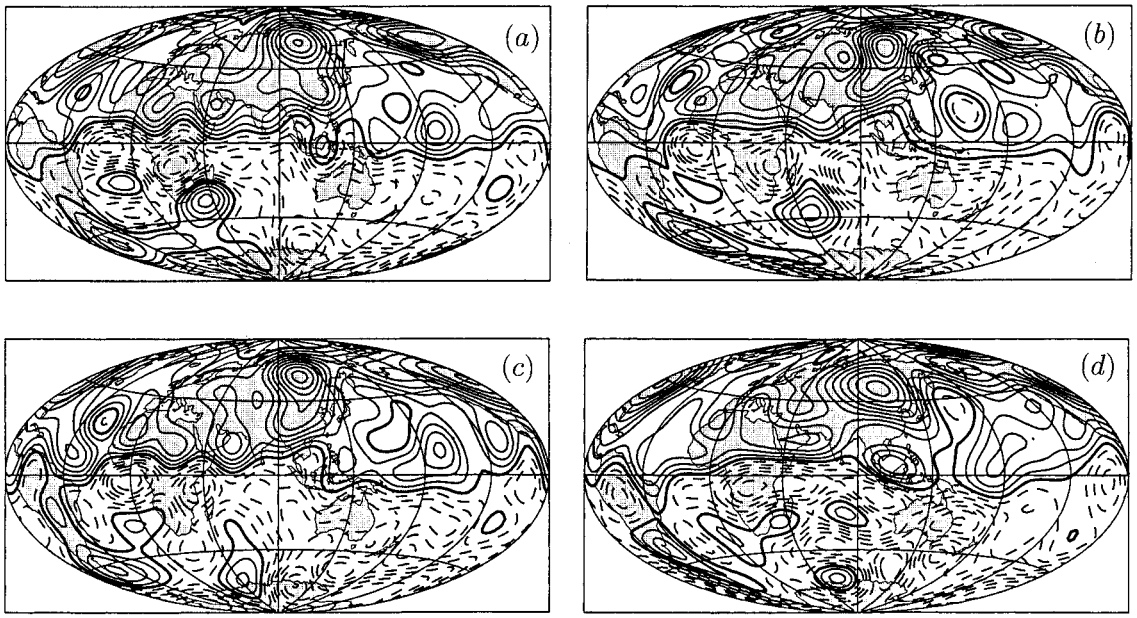


Figure 1. Contour plots of the vertical component of the magnetic field at the core–mantle boundary (from Bloxham *et al* 1989) for (*a*) 1980, (*b*) 1945, (*c*) 1915, (*d*) 1842. Solid curves are positive contours, broken curves are negative contours, bold curves are zero contours

Because of the marginal (although fully dynamic) nature of the geodynamo, we expect that $\boldsymbol{v}$ should be scaled with η/L where L is an appropriate lengthscale (presumably the core radius). Thus R_m, the magnetic Reynolds number based on this radius, is $O(1)$. For completeness we should distinguish between the poloidal part $\boldsymbol{B}_P$ and toroidal part $\boldsymbol{B}_T$ of $\boldsymbol{B}$, defined by $\boldsymbol{B}_P = \boldsymbol{\nabla} \times \boldsymbol{\nabla} \times S\boldsymbol{r}$, $\boldsymbol{B}_T = \boldsymbol{\nabla} \times T\boldsymbol{r}$. The toroidal field, which is predominantly zonal, cannot be directly observed and may be much larger than $\boldsymbol{B}_P$. Observations and stability considerations for large zonal fields suggest that $|\boldsymbol{B}_T| \sim |\boldsymbol{B}_P|$. Of course a different scaling might apply to the zonal flows, which are less constrained by rotation, but if there were to be very large zonal flows then $\boldsymbol{B}_T$ would be so large that the resulting field structure would be unstable.

The structure of the geomagnetic field is thus determined from the dimensionless constants that are implicit in the above system. These are:

the Elsasser number $\Lambda = B^2/2\Omega\mu\rho\eta$
the Ekman number $E = \nu/2\Omega L^2$
the Rossby number $R_o = \eta/2\Omega L^2$

Observations suggest that Λ is of order unity (corresponding to a field of about 20 G) while $E \ll R_o \ll 1$. The principal force balance would therefore appear to be between the Coriolis and Lorentz forces and the pressure, with viscosity and inertia playing a negligible role. Before examining the consequences of this so-called magnetostrophic balance, we consider the alternative of a weak magnetic field that plays a secondary role in any balance. Such weak-field dynamos have certainly been found in numerical simulations, but they are implausible candidates for the geomagnetic field. They are, firstly, too weak to be readily compatible with observations. Secondly, convection in a spherical shell at low Ekman number is characterized by very short ($O(E^{\frac{1}{3}})$) azimuthal length scales, which would dissipate too much energy to be a plausible dynamo. Finally in the absence of strong magnetic fields the primary force balance would be geostrophic. This causes severe difficulties at the cylinder aligned with Ω which is tangent to the inner core, causing large discontinuities in the velocity. For all these reasons the prevailing balance is likely to be magnetostrophic.

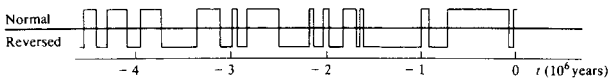


Figure 2. Record of reversals of the Earth's dipole polarity over the last 4×10^6 yr (from Moffatt 1978). Only the direction of the dipole, not its intensity, is shown

The role of viscosity

It turns out that viscosity (unlike inertia) cannot simply be ignored as pure magnetostrophic balance does not exist in general. The reduced equation

$$2\Omega \times \boldsymbol{v} = -\boldsymbol{\nabla} P + H\hat{\boldsymbol{r}} + \frac{1}{\mu\rho}(\boldsymbol{\nabla} \times \boldsymbol{B} \times \boldsymbol{B}) \qquad (3)$$

with the inviscid condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ at the core–mantle boundary, has no solution unless the Lorentz force obeys Taylor's condition

$$\int_{C(s)} (\boldsymbol{\nabla} \times \boldsymbol{B} \times \boldsymbol{B})_\phi \, \mathrm{d}z \, \mathrm{d}\phi \equiv T(s) \equiv 0 \qquad (4)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

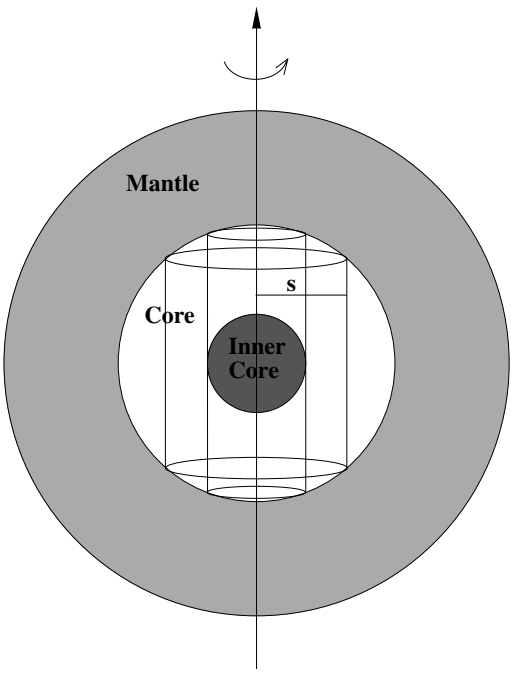


Figure 3. Sketch of the Earth. Shown are the mantle, core and inner core. Also sketched are a typical $C(s)$ and the tangent cylinder with the inner core.

where (s, ϕ, z) are cylindrical polar coordinates and $C(s)$ is the cylinder of radius s aligned with Ω and inscribed in the sphere (figure 3). (Note that the value of this integral is in general discontinuous at the tangent cylinder referred to above.) This very severe constraint is associated with a free geostrophic solution for the velocity of the form $\boldsymbol{v} = V_G(s)\hat{\boldsymbol{\phi}}$; while there might be hope that there is a choice of V_G that allows the Taylor integral to vanish, no proof is known and numerical simulations are far from providing conclusive results. Indeed, they suggest instead that the neglect of viscosity can lead to unacceptable radial singularities in the solutions, as well as removing the dynamical connection between core and mantle.

Thus many models of the geodynamo either retain the viscous term (which makes the problem extremely stiff and hard to solve numerically) or else note that viscosity is likely to be important only in thin Ekman boundary layers at the core–mantle boundary. The boundary layer equations can then be solved to produce instead of the Taylor condition $T(s) = 0$ a formula for V_G of the form $LV_G/\eta = \Lambda E^{\frac{1}{2}} f(s)T(s)$, where $f(s)$ is universal. This definition, like others, encounters problems at the tangent cylinder with the inner core, where further boundary layers may be important. Leaving the inner core aside, there remains the question of the behavior of fields and flows for very small E (which according to some estimates may be as low as 10^{-14}). Either Taylor's condition will be approximately met (a 'Taylor state') or else the zonal flows will become very large and severely limit the field growth (an 'Ekman state'). Braginsky has proposed an intermediate alternative in which the zonal flows are moderately large but also the meridional field is almost aligned with the rotation axis. The dominant balance in this case is between viscous damping and the input of energy from buoyancy forces, whereas in a Taylor state viscosity plays no role. While the idea of this balance remains appealing, it has proved hard to obtain satisfactory numerical models. While the Ekman regime can readily be found numerically, the magnetic field strength given by Λ scales as $E^{-\frac{1}{2}}$ and so this balance is unlikely to be relevant to the geodynamo.

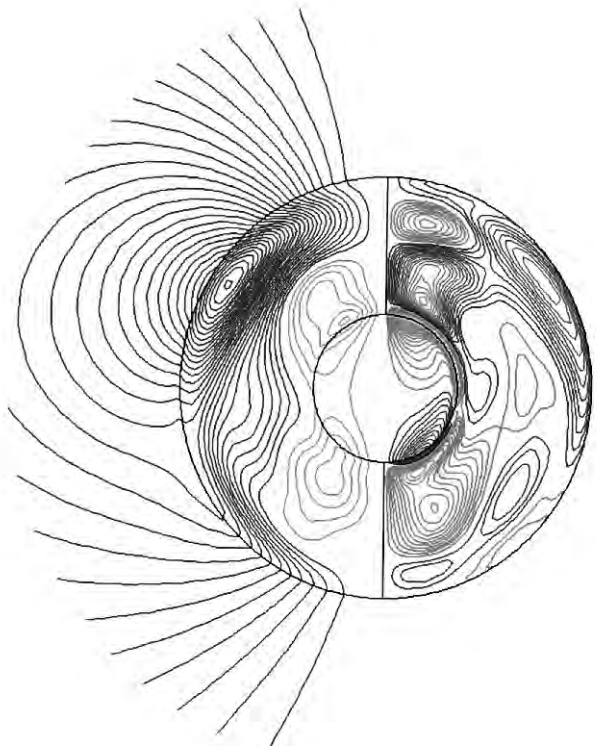
Numerical models

While the above notions have proved challenging for theorists, the advent of fast computers has led to much progress via direct numerical simulation. Cowling's theorem, that no axisymmetric field can be generated by dynamo action, necessitates either a full 3D calculation or else a mean-field model in which only the axisymmetric mean fields and flows are calculated, with the presumed non-axisymmetric small-scale convection being parametrized by the α effect. (The convection having been parametrized, the heat equation can be dropped.) An intermediate ('2.5D') method sets up the full 3D framework but performs severe truncation in the azimuthal direction, retaining the axisymmetric modes and just one or two other zonal wavenumbers. Models of the last two types can be investigated in considerable detail. Viscosity is usually retained in the equations, and E made as small as possible. Geophysically realistic values have as yet not been attained, however. Solutions show irregular oscillatory behavior, but the field strengths are of the right order for the Earth.

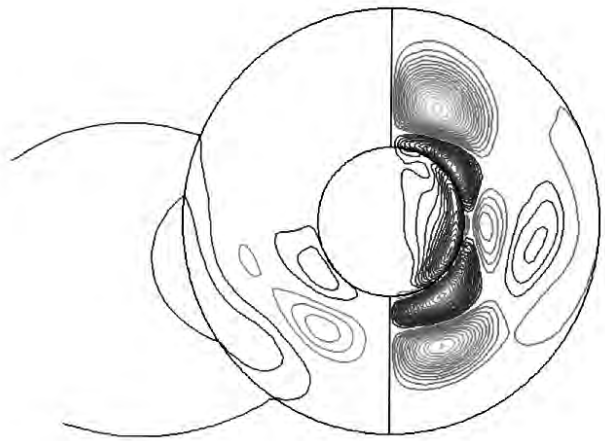
The role of the inner core is particularly important in these simulations, although it was neglected in early computational work. Its influence is manifested in two ways; the tangent cylinder separates regions containing quite different types of flow, with the most vigorous activity within. More importantly, because the inner core is a conductor (with conductivity comparable with that of the outer core) any magnetic field lines penetrating it can only evolve on the diffusion timescale (as the inner core is solid). This helps to stabilize the otherwise very wild oscillatory behavior in the models and to stop frequent reversals of the main dipole field on unrealistically short timescales.

Glatzmaier and Roberts have recently carried out a series of ever more sophisticated numerical simulations of a fully three-dimensional dynamo model (figure 4). In essence the models are the same as the 2.5D simulations, but now with full azimuthal resolution. Ekman numbers as low as 10^{-6} are reported for some runs, but these used an artificial hyperviscosity, which has the effect of mimicking the effect of much larger values of E. Nonetheless the dynamics in the core is impressively detailed (although the field as observed at the Earth's surface is much less regular in the simulation). The runs have also uncovered

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



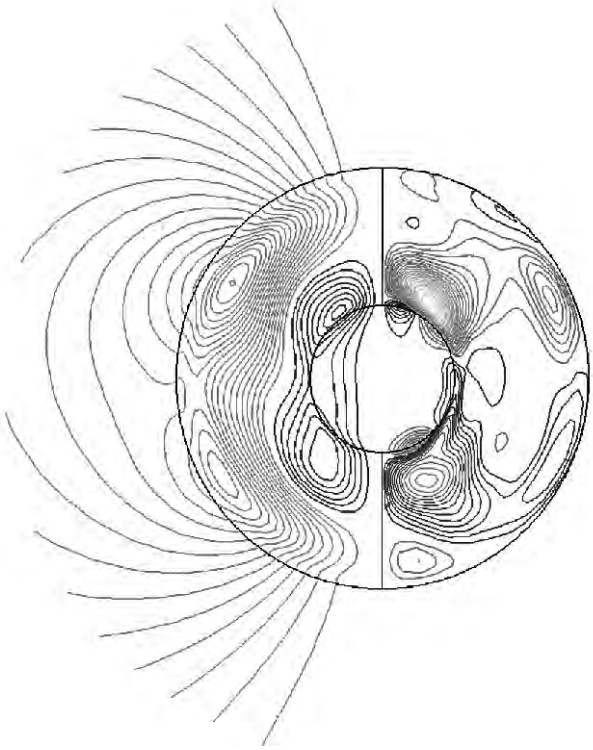
(a)



(b)

Figure 4. Lines of zonally averaged poloidal field (left halves) and contours of zonally averaged toroidal field (right halves) for a reversal in the model of Glatzmaier and Roberts (1995). Black–grey lines denote opposite directions–signs of the corresponding fields.

several reversal events, as well as a number of 'aborted reversals' whose signature is a large excursion of the virtual geomagnetic pole (defined as the axis of a mapping of the surface field onto a supposed dipole) without a complete change of sign. As before, the inner core seems to play a large role in the dynamics of reversals.



(c)

Figure 4. (Continued)

Calculations done by other groups also emphasize the crucial importance of the mechanical boundary conditions to be applied at the core–mantle boundary.

Sadly, even these massive computations, using hundreds of hours of supercomputer time, do not seem to be likely to be able to handle geophysically realistic parameters for the immediate future. For theorists the primary challenges over the next decade are likely to be an effort to understand the nature of the convective turbulence in the core, so as to produce better-parametrized models of the diffusion coefficients. Further understanding of the mechanisms associated with the solidification of the inner core will also bring better descriptions of the buoyancy forces that drive the motion. There can be little doubt that the recent computational successes, far from closing the subject down, will provide fresh impetus for further theoretical developments.

Bibliography

Bloxham J, Gubbins D and Jackson A 1989 Geomagnetic secular variation *Philos. Trans. R.Soc.* A **329** 415–502

Fearn D R 1996 The geodynamo *Earth's Deep Interior* ed D Crossley (London: Gordon and Breach) pp 79–114

Glatzmaier G A and Roberts P H 1995 A three-dimensional self-consistent computer simulation of a geomagnetic field reversal *Nature* **377** 203–9

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moffatt H K 1978 *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge: Cambridge University Press)

Proctor M R E and Gilbert A D (eds) 1994 *Lectures on Solar and Planetary Dynamos* (Cambridge: Cambridge University Press)

Michael R E Proctor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geomagnetic Storm

A fluctuation in the Earth's magnetic field caused by the arrival of a disturbance in the solar wind. Also known as a magnetic storm. The sequence of events in a typical major storm is as follows. The shock wave associated with a cloud of plasma ejected by the Sun, as a result of a solar flare or coronal mass ejection, compresses the Earth's magnetosphere, producing a rapid intensification of the geomagnetic field at ground level. This phase, which takes place over a few minutes, is called the 'sudden storm commencement'. This is followed by the 'initial phase', which lasts for up to a few hours, during which the plasma cloud flows past the Earth and the geomagnetic field strength remains higher than normal. During the next stage, the 'main phase', charged particles flow through the inner magnetosphere, thereby creating an electrical current, the 'ring current', which flows around the Earth. The ring current generates a magnetic field opposite to that of the Earth and causes a pronounced drop in the geomagnetic field lasting from a few hours to about a day. As the current declines, the field strength gradually returns to normal, this recovery phase lasting typically for a few days. Geomagnetic storms are often accompanied by displays of the aurora.

Geomagnetic storms induce currents and surges in telephone lines and power cables that sometimes cause significant damage. For example, circuit breakers in transformer stations may be caused to trip out, thereby giving rise to power blackouts.

See also: aurora, flare, magnetosphere of Earth, solar coronal mass ejection, solar wind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Geotail

Japanese–US mission to study the dynamics of the Earth's magnetotail over distances ranging from 8 Earth radii to 200 Earth radii. Part of the International Solar-Terrestrial Physics (ISTP) programme. Launched in July 1992. During the first two years, the spacecraft spent most of its time in the distant magnetotail. Its orbit apogee was kept on the night side of the Earth by using the Moon's gravity in a series of double-lunar-swingby maneouvres. In February 1995, the apogee was reduced to 30 Earth radii to study near-Earth magnetotail processes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gerard of Cremona (*c.* 1114–87)

Italian scholar, born in Cremona, Italy, flourished in Toledo, a centre of Islamic culture, learned Arabic to be able to translate lost classical works. He translated 92 works in all, including the *Almagest*, *Elements*, AL-KHWARIZMI'S works, the Hippocratic writers, and the physical works of ARISTOTLE. Unfortunately, no mathematician, he could not understand the *Almagest* after translating it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

German Astronomy

The first larger observatories in Germany started their work at the beginning of the nineteenth century. The earliest was founded in Gotha in 1791, followed by Göttingen in 1803, where C F GAUSS became director in 1818 after the observatory had been completed. The observatories in Königsberg and Munich were established in 1813 and 1816 respectively, with F W BESSEL and J G Soldner as the first directors. The cities of Hamburg (J G Repsold) and Berlin (W W Encke) followed in 1825 and 1832 and the Observatory of Bonn (F W Argelander) was founded in 1845.

Spectroscopy has been one of the great traditions in German astronomy. J FRAUNHOFER (Munich) detected hundreds of absorption lines in the solar spectrum in 1823. Together with Soldner he investigated the spectra of the brightest stars and planets and reported morphological differences of the stellar spectra. In 1838, the second director of the Munich Observatory, J LAMONT, used the large 'Fraunhofer Refractor' to perform spectroscopy of stars down to fourth magnitude. G R KIRCHHOFF and R W Bunsen (Heidelberg) developed the concept of spectral analysis in 1859 and gave a first qualitative interpretation of the solar spectrum in 1860. H C VOGEL started his pioneering work on the radial velocities of stars in Bothkamp/Kiel in 1870. Four years later, he moved to the newly founded observatory in Potsdam and detected the line shifts and binary nature of Algol in 1889. Together with F Scheiner he published the first radial velocities of stars in 1890. With more refined methods they reached an accuracy of 3 km s^{-1} per photographic spectrum. F Küstner in Bonn was able to improve this accuracy at the turn of the century. Observing F- and M-stars he obtained an accuracy of 1.4 km s^{-1}. Inspired by K SCHWARZSCHILD in Göttingen, E HERTZSPRUNG started his spectroscopic work on color–magnitude diagrams in 1910 in Potsdam. K Schwarzschild, A Unsöld, L Biermann and W Grotrian contributed fundamentally to the development of the quantitative spectroscopic analysis of stellar spectra. UNSÖLD's book '*Physik der Sternatmosphären*' (first edition 1938, second edition 1955) became the 'Bible' for generations of stellar spectroscopists.

Parallel to spectroscopy, photometry developed as another important branch of German astronomy. The *Potsdamer Durchmusterung* (1905) by G Müller and P Kempf provided magnitudes for 14 195 stars based on a photometer developed by ZÖLLNER (1861). It was the most precise photometric catalog of the time. In 1896, K Schwarzschild (then in Vienna) transformed photometry with photographic plates into a quantitative science. J Wilsing and J Scheiner (Potsdam) performed the first determinations of stellar temperatures by measuring intensities in five spectral bands in 1909. K Schwarzschild (now in Göttingen) published the *Göttinger Aktinometrie* in 1910, a catalog of photometric magnitudes of 3522 stars down to magnitude 7.5 .

All early research on the photoelectric effect was done in Germany (H Hertz 1887, W Hallwachs 1888 and J Elstner and H Geitel 1889). Subsequently, P Guthnick (Berlin) and H Rosenberg and E Meyer (Tübingen) were the first astronomers to use a photoelectric cell for astronomical measurements in 1912. A year later Guthnick published his photoelectric lightcurve of the short-period (4.6 h) low-amplitude (0.086 mag) variable star β Cephei obtained with a sodium cell. In the following years Guthnick, R Prager and K Bottlinger experimented with color filters and published the colors of more than 400 stars using a potassium hybrid cell.

The founder of astrometry in Germany was F W Bessel (Königsberg). In 1838 he determined the parallax of 61 Cygni, the first direct determination of a distance to a star. J G GALLE (Potsdam) detected the planet Neptune in 1846 based on the theoretical predictions by U Leverrier, who analysed the orbit perturbations of Uranus. The *Bonner Durchmusterung* organized by F W ARGELANDER and collaborators from 1852 to 1862 provided positions and finding charts for 73 500 stars in the northern hemisphere down to ninth magnitude. In 1867 the ASTRONOMISCHE GESELLSCHAFT organized an international collaboration of 16 observatories to produce the *Fundamentalkatalog*, a catalog with precise stellar positions of northern hemisphere stars. This work has been extended and improved through to the present day under the responsibility of the Astronomisches Recheninstitut in Heidelberg. It formed an important basis to constrain models for the dynamics of our galaxy.

The investigation of variable stars as a tool to understand stellar evolution started with the work of Argelander in 1840 and 80 years later led to the systematic search for variables by C HOFFMEISTER carried out in the observatories of Sonneberg and Bamberg. More than 4500 variable stars were detected by Hoffmeister. The photographic plate archive of the two observatories contains more than 200 000 sky exposures.

MAX WOLF in Heidelberg used his new astrophotographic methods to investigate the structure of the Milky Way and to detect many minor planets during the years from 1887 to 1932.

There has also been a remarkable tradition in the development of telescopes, optics and instrumentation. At the beginning of the nineteenth century Fraunhofer invented new technologies to produce large lenses of excellent optical quality. He also invented the grating spectrograph. Together with Metz and Utzschneider he constructed the largest and best refractors of the time. Repsold (Hamburg) became famous for the quality of his meridian circle telescopes. In 1905, K Schwarzschild formulated the complete third-order aberration theory of large telescopes with one or two mirrors, which became the basis for the development of all modern reflecting telescopes. In 1930, B SCHMIDT (Hamburg) used an aspheric refracting plate to built his famous wide-field telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The present situation of German astronomy

German astronomy deals with practically all aspects of modern astrophysical research in theory and observation. Stellar astronomy (including the investigation of the Sun) through all stages of stellar evolution is a field of great activity. This includes the important areas of star formation and the formation of planets and proto-planetary disks and, of course, the solar system. The investigation of the physics of the interstellar medium and of the structure and evolution of our galaxy is another very important area. Observational extragalactic astronomy and cosmology had difficulties in the early phases of their development because of the very limited access to large telescopes. However, after the completion of the large ESO telescopes on La Silla, the national CALAR ALTO OBSERVATORY in Spain, the EFFELSBERG and IRAM RADIO TELESCOPES and space observatories such as the ROSAT x-ray satellite this situation has improved dramatically. There are now many competetive groups, both in observation and theory, working in extragalactic astronomy and cosmology.

The development of telescopes and instrumentation has again become a very important part of present German astronomy. This comprises radio, mm and sub-mm telescopes (Effelsberg, IRAM and the Heinrich Hertz Telescope in Arizona), IR astronomy from the ground and in space (instrumentation for ESO La Silla and the VLT, the Calar Alto and the ISO and FIRST space observatories), optical astronomy (instrumentation for the ESO VLT and ESO La Silla, the Calar Alto, the LARGE BINOCULAR TELESCOPE (LBT) in Arizona and the Hobby–Eberly Telescope (HET) in Texas), UV astronomy (ORFEUS), x-ray astronomy (ROSAT, ABRIXAS, Chandra, XMM, XEUS) and gamma-ray astronomy (COMPTEL on GRO, INTEGRAL). The HEGRA EXPERIMENT on Tenerife provides an observatory for very high energies and there is also development of a gravitational wave telescope. In solar physics a new vacuum tower telescope has been constructed as the most important part of the solar observatory on Tenerife by three German institutes in collaboration with Spanish colleagues.

At present, astrophysical research in Germany is performed at 36 institutes distributed over the reunified country. Of these, 22 are university institutes (including the observatories of Bonn (with the observatory Hoher List at Daun), Hamburg, Göttingen and Munich (with the observatory at Mount Wendelstein)). They are integrated parts of the physics faculties of their universities, since in Germany an astronomer is basically a physicist with specialization in astrophysics during the last semesters before the diploma (masters degree) and in PhD thesis work. Though they have most of the teaching load and responsibility for the education of astronomy students, the university institutes are usually small (between one to five professors and three to fifteen scientific staff) and have a very low scientific budget provided by their universities. For the funding of graduate students, post-docs and instrument development they rely completely on grants, which are distributed by the funding agencies in a competitive system. Used to being continuously forced to survive in this competition, several university institutes do extremely well. Most astronomers at the universities make observations as guests at ground-based or space facilities or do theoretical work. Only a few of them at a handful of institutes are involved in instrument development. German universities are state universities, which means that they are funded through the science budgets of the individual states ('Länder') of Germany.

The funding situation is much more positive at the seven MAX-PLANCK INSTITUTES (MPIs) dealing with astronomy. These institutes have usually several (up to four) directors, more than 30 staff scientists and an excellent infrastructure, which allows them to engage in large-scale instrumentation projects, ground-based and in space, and to run their own observatories (Calar Alto Observatory, IRAM, Effelsberg, ROSAT). They are also able to form large theoretical groups and perform observational astronomy as guest observers. All these institutes are located close to universities with astronomical institutes. However, the amount of interaction between MPIs and universities in research and the education of students varies strongly from place to place. The funding of the MPIs is based on equal contributions from the federal science budget and from the science budget of the states in which the institutes are located. There is also a contribution from German industry.

The situation at the Institutes of the Wilhelm-Gottfried-Leibniz-Gesellschaft (WGL), of which there are two, and the two Institutes of the Deutsches Zentrum für Luft- und Raumfahrt (DLR) dealing with astronomy is similar to that of the MPIs. There are also four state institutes in the states of Baden-Württemberg and Thüringen, of which two are active observatories (the 'Landesternwarten' (LSW) in Heidelberg and Tautenburg). Their sizes and resources are comparable to university institutes.

The number of professional astronomers in Germany with a PhD is estimated to be 500. The ratio of tenured to non-tenured staff is roughly 1:1. Most of them are members of the Astronomische Gesellschaft (AG), which organizes scientific meetings every year. Many amateur astronomers and friends of astronomy have also joined the AG.

The extramural funding comes from separate budgets of the federal ministry for education and research (BMBF). For astronomy there are two important funding agencies—the DLR which sponsors a national program of space science and the Deutsche Forschungsgemeinschaft (DFG) which supports almost all branches of research in Germany. The DLR has contributed or contributes to space missions such as ROSAT, ORFEUS, ABRIXAS, SOFIA etc and also provides funding for focal instrumentation for ESA/NASA space missions such as Giotto, Chandra, XMM, ISO, Pathfinder etc when German groups are involved. The DFG has basically three different programs, the normal program, where tenured scientists from universities, WGL and state institutes can apply to obtain

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

funding for all kinds of individual smaller science projects involving PhD students and post-docs, special programs concentrating on selected topics such as 'cosmic plasmas' or 'star formation', which are open to all institutes, and local programs, where different institutions in and around one or two universities focus on special subjects such as 'the evolution of galaxies' or 'particle astrophysics'.

A new program is the 'Verbundforschung Astronomie' financed directly through the BMBF and the DLR, which supports instrumentation proposals by university groups for ground-based national and international telescopes (to compensate for the lack of internal resources of these groups) and provides research funds for those astronomers who succeeded in obtaining observing time on international space missions such as HST, ISO and ROSAT. This new program has turned out to be a tremendous success and has improved the competitiveness of German astronomy significantly.

The German contributions to international organizations like ESO and ESA are funded directly through the BMBF. Membership in these organizations is fundamental for German astronomy. It has widened the scope of astronomical research in Germany significantly and has internationalized the working style, in particular of the university institutes, substantially. Most importantly, it has enabled German astronomy to participate in instrumentation projects clearly beyond the means of national or bi-national collaborations.

The German system is characterized by an enormous independence of each research institute. While this certainly has significant advantages with regard to scientific originality and creativity, it is also the cause of difficult problems. Collaboration between universities of different states in larger projects is hampered by obstacles intrinsic to the German federal system. The same is true for the collaboration between universities and MPIs. This is one of the reasons why the German national observatory on Calar Alto in Spain is managed by a MPI, the Max-Planck-Institut für Astronomie in Heidelberg, and why it is difficult to organize direct participation of German universities in large telescope projects such as the LBT in Arizona, where so far only three MPIs (MPIA Heidelberg, MPIE Garching, MPIfR Bonn), one WGL institute (Astrophysikalisches Institut Potsdam) and one state observatory (LSW Heidelberg) are involved. Two university observatories (Göttingen and Munich) have, however, managed to become partners of American universities in the HET project in Texas. They have also embarked on a collaboration together with the LSW Heidelberg to build the faint object spectrographs FORS 1 and 2 for the ESO VLT.

In view of the complication caused by the federal system and the independence of the individual institutes an element of coordination is needed. This is the task of the 'Rat Deutscher Sternwarten' (RDS, Council of German Observatories). It was founded in 1959 as the 'Rat Westdeutscher Sternwarten' to nationally and internationally represent the joint interests of the astronomical institutes in West Germany. In 1990, shortly after the reunification of Germany, the renamed 'Rat Deutscher Sternwarten' became the representative organization for all German astronomical institutes. At present 31 such institutes represented by their directors are members of this important committee, which coordinates the main directions of German astronomical research, defines priorities for future large projects, and interacts with the funding agencies, state and federal ministries. It also represents German astronomy within the IAU and in the interaction with ESO. The RDS is headed by a chairman, who is elected for three years and can be re-elected for one additional term and a General Secretary.

Rolf-Peter Kudritzki and Reinhold Häfner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ghost of Jupiter (NGC 3242)

A planetary nebula in the constellation Hydra, position RA $10^{h}\,24.8^{m}$, dec. $-18^{\circ}\,38'$. It is of eighth magnitude and has a twelfth-magnitude central star. It is bluish, and with its size of $16''$ it can appear as a faint version of Jupiter.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giacconi, Riccardo (1931–)

Astrophysicist, born in Genoa, Italy, American astronomer and international project director. Worked on x-ray astronomy, and with a rocket-borne x-ray telescope intended to study the composition of the Moon, discovered Scorpius X-1, the first known x-ray source outside the solar system. Built the Uhuru x-ray satellite that made the first surveys of the x-ray sky. Led the construction and successful operation of the powerful x-ray satellite Einstein. Directed the Space Telescope Science Institute, laying the foundation for the impact of the Hubble Space Telescope and directed the EUROPEAN SOUTHERN OBSERVATORY, bringing to a successful completion its Very Large Telescope.

In 2003, Giacconi received the Nobel Prize in physics for 'pioneering contributions to astrophysics', which led to the discovery of cosmic x-ray sources. Giaconni, currently president of the Associated Universities Inc., in Washington, and Research Professor of Physics and Astronomy at Johns Hopkins University, Baltimore, discovered the first x-ray stars and the x-ray background in the 1960s and conceived of and led the implementation of the UHURU and High Energy Astronomy Observatory-2 (HEAO-2) x-ray observatories in the 1970s. He also detected sources of x-rays that most astronomers now consider to contain black holes.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Giacobinids

A periodic meteor shower associated with the comet 21P/Giacobini–Zinner which takes place in October. The radiant lies in the constellation Draco, and the shower is also known as the *Draconids*. Debris from the comet remains in a swarm close to the nucleus, and shower activity is possible only when the Earth passes the descending node of the comet's orbit shortly after the comet has passed through it. Zenithal hourly rates were high in 1933 (up to 450), 1946 and 1985. Radio-echo observations in 1946 detected a peak rate of 170 per minute.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giant Impactor Theory

The theory, also known as the *big splash* theory, that a planetoid about the size of Mars collided with the newly formed Earth, and ejecta from the collision accreted in orbit around the Earth to form the Moon. Support for the theory comes from the very close correspondence of oxygen isotope ratios in lunar and terrestrial material, and the similar mineralogies of the two bodies. In addition, the theory explains the large amount of angular momentum in the Earth–Moon system.

See: Moon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giant Planets

A collective term for the major planets Jupiter, Saturn, Uranus and Neptune, which are the four largest planets in the solar system. They are characterized by their size (of the order of 100 000 km in diameter), their density (of the order of 1000 kg m^{-3}), their largely gaseous nature, with no visible solid surface, and their extensive families of satellites. They also all possess ring systems. Of the giant planets Jupiter is the largest and most massive, and rotates the most rapidly. Saturn is the most oblate and the least dense, and has the most extensive ring system. Saturn and Uranus tie for the largest retinues of confirmed satellites, at 18. The most distant, Neptune, is the densest and has the largest measured atmospheric wind speed. The majority of exoplanets—objects orbiting other stars—are more massive than Jupiter, and may be classed as giant planets.

Compare: terrestrial planets.

See also: exoplanets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gibbous

The phase of a body in the solar system when more than half of its sunlit side is visible, but not all of it. From the Earth, the inferior planets, Mercury and Venus, and the Moon show gibbous phases. Of the superior planets, Mars is the only one that can appear markedly gibbous (with a minimum of 84% of its disk illuminated); of the outer planets, Jupiter can show a slight gibbous phase but the planets beyond are too distant to do so.

See also: phases of the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gilbert [Gilberd], William (1544–1603)

Doctor and scientist, born in Colchester, England, wrote *De Magnete* (*On the Magnet*), published in 1600. The magnetic compass was one of most useful the navigational instruments before the chronometer, but little was known about the lodestone (magnetic iron ore). Gilbert made his own experiments, such as testing the folk-belief that garlic destroys the magnetic effect of the compass needle. He draws the first clear distinction between magnetic effects and static electricity, reviewing everything about magnetism. His theory of terrestrial magnetism was curiously animistic: magnetism was the soul of the Earth and a spherical lodestone, aligned with the Earth's poles, would spin on its axis like the Earth. KEPLER and GALILEO were very interested in this aspect of Gilbert's magnetic researches, since they were seeking arguments that the Earth rotated.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gill, Sir David (1843–1914)

Astronomer, born in Aberdeen, trained as a watchmaker, had a private observatory near Aberdeen, became Her Majesty's Astronomer at the Cape of Good Hope. He redetermined the solar parallax by observing the close approach to the Earth in 1889 of the minor planets Iris, Victoria and Sappho (his was the standard value for nearly a century). His measurement was repeated by HAROLD SPENCER JONES during the close approach of Eros in 1931. Inspired by the unexpected discovery in 1882 of stellar images on a photograph of a comet, he pioneered astronomical photography. He photographed the southern sky for the *Cape Photographic Durchmusterung*, CPD, adding to ARGELANDER'S *Bonner Durchmusterung*. This initiated the international Carte du Ciel project and the subject of photographic astrometry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ginga (Astro-C)

Japanese x-ray satellite, launched February 1987. Carried three instruments: a large-area proportional counter array developed with the UK, an all-sky monitor and a US gamma ray burst detector. Observed supernova 1987A soon after its appearance, and discovered several pulsars and other transient sources. Re-entered the atmosphere in November 1991. Ginga means 'galaxy'. (See also ASTRO.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giotto

The purpose of the Giotto mission was to study the coma and nucleus of Comet P/Halley at close quarters during its 1986 perihelion passage. It was the EUROPEAN SPACE AGENCY's first deep space mission and was named after the famous medieval Italian artist, Giotto di Bondone, who, inspired by the reappearance of COMET HALLEY in 1301, had portrayed the Star of Bethlehem as a golden comet in his fresco 'Adoration of the Magi'.

Giotto was originally proposed as part of a joint US–European mission to Halley's Comet, but after NASA withdrew, ESA decided to proceed alone. It was later included in an international program to investigate the comet. Five space probes—two Soviet, two Japanese and one European—were launched towards the comet on its return to the inner solar system. While the Japanese spacecraft (Sakigake and Suisei) made long-distance measurements, the Soviet VEGAS acted as pathfinders for Giotto by accurately locating the comet's nucleus.

The cylindrical spacecraft weighed 960 kg. Its design was based on the GEOS Earth-orbiting research satellites, which were built by British Aerospace at Bristol (UK). The most significant modification was the addition of an aluminum–Kevlar buffer to protect it from high-speed dust particles during the comet encounter. The spacecraft was eventually launched by an Ariane 1 rocket (flight V14) on 2 July 1985.

Giotto carried ten scientific instruments which included

- a narrow-angle, multicolor CCD camera to obtain pictures of the nucleus,
- three mass spectrometers to measure gas and dust composition,
- a dust impact detector to measure the mass of dust particles striking the shield,
- two plasma experiments to study the solar wind and charged particles,
- an energetic particles analyser to study electrons, protons and alpha-particles,
- a magnetometer to study changes in the magnetic field,
- an optical probe to study the brightness of the coma and
- a radio science experiment to investigate the electron environment by comparing signals sent at different frequencies from the spacecraft.

The first comet data arrived at 21.00 GMT on 12 March 1986, when the spacecraft's instruments detected hydrogen ions 7.8 million km from the nucleus. Giotto crossed the bow shock, where the solar wind slowed to subsonic speed, 22 h later and entered the coma (see COMA, COMETARY).

Excitement rose as the stream of pictures and data came in to the European Space Operations Centre in Darmstadt, Germany, at a rate of 46 000 bits s^{-1}. The first of 12 000 dust impacts was recorded 62 min before closest approach (290 000 km out), but the rate of impacts rose sharply as the spacecraft passed through a jet of material from the nucleus.

Only 7.6 s before closest approach, the spacecraft was sent spinning by an impact from a 'large' (c 1 g) particle. Although contact with the Earth was lost, the spacecraft's thrusters were able to stabilize its motion and contact was restored 32 min later. By then Giotto had passed within 596 km of the nucleus on the sunward side and was heading back into interplanetary space.

The remarkably resilient spacecraft continued to return scientific data for another 24 h. The last dust impact was detected 49 min after closest approach. The historic encounter ended at 02.00 GMT on 15 March when Giotto's experiments were switched off.

Altogether 2112 images were returned before communications were disrupted. A unique series of images sent back by the spacecraft revealed the comet nucleus to be a dark, peanut-shaped body, about 15 km long and 7–10 km wide. Its surface was very irregular, with hills and depressions, while its density (0.3 g cm^{-3}) indicated a 'fluffy', porous texture. The black surface was composed of organic compounds (see also COMETARY NUCLEI).

Seven jets that were identified on the warmer sunlit side threw out 3 t s^{-1} of material, but only about 10% of the surface was active. These jets gave the comet a strange, wobbling rotation which seems to be stable on a time scale of centuries or even millennia.

Water accounted for about 80% by volume of all of the material ejected by the comet. There were also substantial amounts of carbon monoxide (10%), carbon dioxide (2.5%) methane and ammonia. Traces of other hydrocarbons, iron and sodium were also found. The surface of the nucleus had an albedo of 4–5%—blacker than coal.

Most of the dust was no larger than specks of cigarette smoke. Two major classes of dust particles were found. One was dominated by the light 'CHON' elements—carbon, hydrogen, oxygen and nitrogen. The other was rich in mineral-forming elements—sodium, magnesium, silicon, iron and calcium. All of the comet's light elements (except nitrogen) were found in the same relative abundance as for the Sun, confirming that Comet Halley consists of primitive material derived from the original solar nebula.

Giotto also discovered a magnetic cavity—a region where the magnetic field was essentially zero—around Halley's Comet. This is the only time that such a feature has been recorded.

Despite the damage sustained during the flyby, eight instruments remained operational and the mission was extended to allow an unprecedented second encounter with a comet—P/Grigg–Skjellerup. After an Earth flyby and a long hibernation, Giotto was reawakened on 4 May 1992, 12 days before the comet reached perihelion. Sweeping past at a distance of only 100–200 km on 10 July, it was the closest ever cometary encounter.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Conditions were very different from those during the Halley flyby. Grigg–Skjellerup approached Giotto at an angle of 68° instead of head on, so the relative approach speed was 14 km s^{-1} (as opposed to 68 km s^{-1} for Halley).

Giotto crossed the bow shock and entered the coma about 17 000 km from the nucleus. Grigg–Skjellerup's dust production rate was about 0.5% that of Halley, and only three sizeable particles were detected, the largest with a mass of about 30 μg.

The first cometary ions were detected 440 000 km from the nucleus, about 12 h before the closest approach. Two peaks were seen in the brightness of the coma, due to scattering of sunlight by cometary dust particles. One peak coincided with the closest approach to the nucleus. The second peak a little over 1 min later remains a mystery.

Scientists were surprised when an abrupt shock wave—caused when the supersonic solar wind slammed into plasma around the comet—was detected on Giotto's outbound leg, but not clearly identified on the inward journey (see also SOLAR WIND: INTERACTION WITH COMETS).

Magnetic field strength was slightly higher than at Halley, but no magnetic cavity was detected around the much smaller nucleus. Strangest of all was the discovery of unusual magnetic waves, each about 1000 km apart, out to at least 300 000 km from the comet. Activity rose and fell over a period of about 70 s and increased in amplitude as time went by.

The science experiments were switched off for the last time at 03.00 GMT on 11 July. On 23 July 1992, Giotto operations were officially terminated.

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giovanelli, Ronald Gordon (1915–84)

Solar physicist, born in Grafton, New South Wales, studied the nature of solar flares as electromagnetic phenomena in which electrons are accelerated by electric fields induced by changing magnetic fields. Established solar observatories at Fleurs, near Sydney, and at Culgoora in north-eastern New South Wales, adjacent to a solar radio observatory. His work on sunspots, prominences and magnetic reconnection was firmly based on physics, with a constant sense of how much there was to learn.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Glazenap [Glasenapp], Sergej Pavlovich [von] (1848–1937)

Russian astronomer who measured the light travel time across the solar system as evidenced by time delays in the times of eclipses of Jupiter's satellites (RØMER'S method). Determined a value for the light travel time between the Sun and the Earth of between 496 and 501 s (actual value is 499 s).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Global Oscillation Network Group

The Global Oscillation Network Group (GONG) is an international, community-based project, operated by the NATIONAL SOLAR OBSERVATORY for the US National Science Foundation, to conduct a detailed study of the internal structure and dynamics of the Sun over an 11 year solar cycle using helioseismology. 1024^2 velocity images are obtained by a six-station network located at Big Bear Solar Observatory (California), Mauna Loa Observatory (Hawaii), Learmonth Solar Observatory (Western Australia), Udaipur Solar Observatory (India), Observatorio del Teide (Canary Islands) and the Cerro Tololo Interamerican Observatory (Chile), providing better than 90% coverage.

For further information see http://www.gong.noao.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globular Cluster Systems in Interacting Galaxies

GLOBULAR CLUSTERS are dynamically bound and dense collections of large numbers of coeval stars. Typical globular clusters have roughly one million stars within a radius of a few parsecs. They are also usually close to spherical, hence the name globular. By virtue of their rich, isolated population of stars they provide an important laboratory for studies of STELLAR EVOLUTION. Moreover, because of their high densities, globular clusters are prime regions for the study of STELLAR DYNAMICS. Globular clusters are also ubiquitous, being found around nearly all galaxies. This suggests that the formation of globular clusters is closely connected to the formation of the galaxies to which they belong.

At young ages, globular clusters are expected to be very bright and blue because of the concentration of many young stars in a compact region. Their high luminosities, blue colors and compact sizes make young globular clusters easy to detect and identify in imaging surveys. Therefore, the absence of known examples of young globular clusters in the Milky Way and M31 is strong evidence that the current STAR FORMATION in these galaxies is somehow not conducive to the formation of globular clusters. This was not always the case, as these two major SPIRAL GALAXIES in the LOCAL GROUP both have substantial populations of old globular clusters. Similarly, ELLIPTICAL GALAXIES, which have no ongoing star formation, also typically have rich populations of globular clusters. These observations led to the general idea that the formation of star clusters as massive and dense as globular clusters was somehow favored at early times. However, the same absence of nearby examples of the process of globular cluster formation strongly limited the understanding of the key physical processes involved in globular cluster formation.

The most nearby example of a galaxy in which there is recent formation of young stellar systems with the high luminosities, blue colors and compact sizes expected of young globular clusters is the LARGE MAGELLANIC CLOUD (LMC). The LMC has several STAR CLUSTERS with ages ranging from a few Myr to several Gyr which have the properties expected for globular clusters at these ages, as well as about a dozen old (10+ Gyr) globular clusters. However, in the LMC the handful of massive, dense young clusters are strongly outnumbered by many more low-mass young star clusters that are not likely to survive processes such as stellar mass loss, tidal shocks and evaporation. Prior to the launch of the HUBBLE SPACE TELESCOPE (HST), there were a few other potential examples of similar objects in more distant galaxies, but existing observations did not have sufficient spatial resolution to determine whether objects beyond the Local Group had compact sizes characteristic of a globular cluster.

The subject of globular cluster formation in galaxy mergers has been revolutionized by HST observations, which have the image quality to resolve the starbursts in galaxy mergers and reveal large numbers of highly luminous, compact, blue star clusters (see also GALAXIES: INTERACTIONS AND MERGERS). These were first discovered by Holtzman (1992) and Whitmore (1993) and their collaborators and were immediately recognized as having the properties of globular clusters at young ages predicted by Ashman and Zepf (1992) and by Schweizer (1987). Specifically, HST images of galaxies previously recognized as the merger of two disk galaxies based on morphological and dynamical information show that these systems have large numbers of bright, blue, compact objects. These objects are striking with their bright magnitudes ($M_B < -10$) and compact sizes ($r < 10$ pc). The colors, magnitudes, and sizes of these objects are consistent with those expected for globular clusters in our own Galaxy when they were young. Subsequent moderate-resolution spectroscopy of some of the brighter examples has shown strong Balmer absorption lines, confirming the identification of these objects as young star clusters.

These candidate young globular clusters have colors, luminosities and sizes that globular clusters like those in the Milky Way are believed to have had when they were young. An important further step for testing the connection between the candidate young globular clusters and old globular clusters is to compare the masses of the clusters in the two populations. In nearly all cases, the mass of a candidate young globular cluster is determined by estimating the age of the cluster from its color or spectral characteristics, using that age and a stellar populations model to estimate the mass-to-light ratio, and then combining the mass-to-light ratio with the observed luminosity to estimate the mass. While all of these calculations are standard, they do depend on parameters that are not directly observed, including the metallicity and, most importantly, the initial stellar mass function in the young cluster. Fortunately for this purpose, the initial mass function appears to be similar in old globular clusters and young star-forming regions, so it is difficult to imagine that all of the candidate young globular clusters formed with dramatically different stellar initial mass functions. Morever, the ages of some of the young globular clusters are several hundred Myr, which is many crossing times. Thus their compact sizes suggest that they are bound, which would be unlikely if the clusters are deficient in low-mass stars, which play an important role in the mass budget of a cluster with a standard initial mass function. Finally, in the one or two cases in which a mass has been estimated dynamically from the widths of absorption lines, the masses are broadly consistent with those given by stellar population models with standard initial mass functions. It is therefore very likely that the inferred masses for the young star clusters observed in galaxy starbursts and mergers are roughly correct. This then implies that many of these objects then have masses and densities like those of Galactic globular clusters.

All globular clusters were once bound clusters of typically $(10^5$–$10^6)M_\odot$ of young stars. However, not all bound young star clusters of similar masses and densities

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

will evolve to become old globular clusters because there are a number of important dynamical processes that act to destroy globular clusters. These include evaporation, tidal shocking and dynamical friction. These dynamical processes that destroy clusters depend on the properties of the young globular clusters such as mass, radius and concentration, and also on the orbit of the cluster and the mass distribution of its host galaxy. Therefore, on an individual basis it is not possible to predict the effect of these processes on a candidate young globular cluster, as this would require detailed knowledge of its orbit, and the mass distribution of the host galaxy, even if the cluster's mass profile was well constrained. However, it is possible to address the question of the role of dynamical evolution by testing predictions for the global properties of globular cluster systems. Comparisons of systems of different ages and in different galaxies may be particularly useful in this regard (see also GLOBULAR CLUSTER SYSTEMS IN NORMAL GALAXIES).

Understanding the effects of these dynamical processes on the evolution of globular cluster systems is important for determining the relationship between the systems of massive, dense young star clusters in mergers and starbursts and systems of old globular clusters such as the Galactic globular cluster system. One critical example is that the young cluster systems appear to form with a power-law mass function. This may not be particularly surprising since the giant molecular clouds out of which they probably form also have a power-law mass function. However, it is in strong contrast to the log-normal mass function of old globular cluster systems, such as that in the Milky Way and M87, which has a peak at $1–2 \times 10^5 M_\odot$. If the young cluster systems are to be identified as globular cluster systems at young ages, then dynamical processes must preferentially destroy lower-mass clusters so that the mass function of the cluster system evolves a Hubble time from a power law to log-normal with a peak at slighly more than $10^5 M_\odot$.

Theoretical analyses suggest that dynamical processes, particularly evaporation and tidal shocking, do preferentially destroy low-mass clusters. This work has been most thoroughly pursued for the Milky Way, where it has been shown that it is plausible that the Galactic globular cluster system was once much more numerous and had a power-law mass function extending to low masses. A significant role for dynamical destruction is also indicated by the study of the structural parameters of Galactic globular clusters, which generally fall only within the bounds allowed by models of evaporation and tidal shocking. Although there are obvious uncertainties in running the clock backwards to infer the initial conditions of the Milky Way gobular cluster system, the consistency of these data with significant dynamical evolution leading to a log-normal mass function is suggestive that similar processes may apply to the candidate young globular cluster systems seen in galaxy starbursts and mergers.

Some observational constraints can be placed on the evolution of young globular cluster systems by studying systems of different ages and by looking at joint constraints on the luminosities, color and radii within a given system. Perhaps the most stringent constraints come from the studies of the gobular clusters of NGC 1275 and NGC 7252, which are about 500 Myr old based on colors and moderate-resolution spectroscopy. Both of these systems appear to have power-law mass functions to the limit of the observations at inferred masses of a few times $10^4 M_\odot$. Thus, any dynamical alteration to the mass function at masses above this lower limit must happen at later ages. This is plausible, as long as the dependence of radius on mass is not too steep. Of course, many clusters can be lost at younger ages, but the strong preferential loss of clusters with masses between this limit and the peak of the old globular cluster mass function must happen at later times. The alternative is that the young clusters are not the same type of objects as the older clusters, even though they have the same masses and densities. A critical challenge for observers in this field is to test these ideas about the dynamical evolution of systems of globular clusters.

A striking characteristic of the systems of compact young star clusters is that they make up an appreciable fraction of the recent star formation in some starbursts and mergers, as measured by the fraction of blue light in the young clusters as compared with the galaxy as a whole. In the most energetic starbursts observed, this fraction of blue and ultraviolet light in bright, compact young star clusters is up to 20%, with smaller fractions being more typical of systems not at the peak of the starburst. This result suggests that the formation of compact, massive star clusters is not a negligible mode of star formation. It also suggests that the formation of massive, compact star clusters may be efficient in starburst and mergers. This has important implications for the understanding of the globular clusters systems around early-type galaxies and bulges, which the mergers and starbursts are believed to evolve into. Specifically, if the formation of young globular clusters is efficient, then it is possible to account for much of the globular cluster populations of bulges and ellipticals through similar processes. The roughly solar metallicity of about half of the globular clusters in ellipticals is consistent with this idea. It is also interesting to note that if 20% of the star formation in a starburst is in compact clusters, then most of these clusters must not survive for a Hubble time, as old globular clusters make up only about 1% of the light in old stellar populations such as the Galactic halo and in giant ellipticals.

Bibliography

Ashman K M and Zepf S E 1998 *Globular Cluster Systems (Astrophysics Series No 30)* (Cambridge: Cambridge University Press)

is a monograph on all aspects of globular cluster systems, including the formation of globular clusters in starbursts and mergers. Schweizer's lectures in

Kennicutt R C *et al* (ed) 1998 *26th Saas–Fee Course on Galaxies: Interactions and Induced Star Formation* (Berlin: Springer-Verlag)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

contain an excellent discussion of galaxy interaction and mergers and the properties of the candidate young clusters in these systems. Kennicutt's contribution in the same volume provides a view of the formation of massive young star clusters in the broader context of star formation in galaxies. Earlier work on the subject includes

Ashman K M and Zepf S E 1993 *The Globular Cluster–Galaxy Connection* (San Francisco, CA: ASP) p 776
Larson R B 1993 *The Globular Cluster–Galaxy Connection* (San Francisco, CA: ASP) p 675

The first predictions of the existence of young globular clusters in gas-rich mergers are in

Ashman K M and Zepf S E 1992 *Astrophys. J.* **384** 50
Schweizer F 1987 *Nearly Normal Galaxies* ed S M Faber (Berlin: Springer) p 60

The first HST observations of these objects were by

Holtzman J *et al* 1992 *Astron. J.* **103** 691
Whitmore B C *et al* 1993 *Astron. J.* **106** 1354

Stephen E Zepf

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globular Cluster Systems in Normal Galaxies

All but the smallest dwarf galaxies possess a family of GLOBULAR CLUSTERS (GCs). They are among the oldest stellar systems in the universe. As such these 'astronomical fossils' provide valuable clues about the formation and evolution of their host galaxy. They offer a unique probe of the metallicity and kinematics at large radii in galactic halos, including our own Milky Way Galaxy.

For LOCAL GROUP galaxies, spectroscopic studies of GCs are well advanced. However the bulk of our knowledge about extragalactic GCs comes from imaging. Imaging studies received a boost from wide-field-of-view CCDs in the late 1980s and from the HUBBLE SPACE TELESCOPE (HST) in the 1990s. It is now possible to study the magnitude, color and spatial distribution of GC systems at the distance of the COMA CLUSTER and beyond. Imaging can be hundreds of times more efficient than spectroscopy for studying GC systems but it supplies little information about age, independent of metallicity, and nothing about kinematics or abundance anomalies.

Perhaps the most basic property of a GC system is the total number of GCs associated with a galaxy. This number varies from dwarf galaxies with a single GC to over 10 000 in the case of some cD galaxies. The number of GCs associated with a galaxy correlates roughly with the host galaxy luminosity (i.e. $N \sim L^2$). In order to facilitate comparison between galaxies, the number of GCs is normalized to $M_V = -15$. This is termed the *specific frequency*,

$$S_N = N10^{0.4(M_V+15)}.$$

The S_N value is affected by environment, with cluster ellipticals having $S_N \sim 5$ and their field counterparts $S_N \sim 3$. Some of the most luminous galaxies, typically cD galaxies, have $S_N \sim 15$, i.e. several times more GCs per unit starlight than other early-type galaxies. Therefore high-S_N galaxies have either formed or acquired GCs more efficiently than they have field stars. It appears that high-S_N galaxies have a high S_N value at all radii throughout the galaxy. Although suggestions to explain the high-S_N phenomenon have been proposed, none is, as yet, universally accepted. Spiral galaxies have $S_N \sim 1$, but if we take account of the lower mass-to-light ratio of the young stellar populations in spiral disks then this rises to $S_N \sim 2$.

Local Group

Magellanic Clouds

Both Magellanic Clouds appear to contain young (<1 Gyr), intermediate (∼5 Gyr) and old (∼15 Gyr) GCs. At a distance of only 50 kpc, this would make the Magellanic Clouds ideal laboratories to study GCs with a range of ages (see also LARGE MAGELLANIC CLOUD, SMALL MAGELLANIC CLOUD). However, there are also some disadvantages of studying the Magellanic Cloud GC systems. First, assuming that GCs have been correctly associated with each Magellanic Cloud, the GC systems are small. The LMC has ∼13 *bona fide* (i.e. old) GCs and the SMC may only have one (NGC 221). Second, it is difficult to distinguish GCs from open clusters (the ratio of open to globular clusters is about 20:1). The LMC GCs have a similar mean magnitude and metallicity to the Milky Way halo ones. They differ in their shapes (slightly more extended and flattened) and kinematics (both young and old GCs reveal evidence for a bulk rotation about the LMC).

M31

The GC system of our sister galaxy, M31, reveals many similarities to our own, with the advantage of greater numbers (i.e. $N \sim 400$ compared with $N \sim 160$ for the Milky Way). For example, the GC luminosity function shows a similar peak and width. The metallicity distribution, although not as obvious, shows a similar metal-poor ([Fe/H] ~ -1.5) and a metal-rich ([Fe/H] ~ -0.5) dichotomy. Associating these with halo and disk populations via kinematics is complicated by the nearly face-on nature of M31. Nevertheless there is some evidence that the metal-rich GCs form a disk-like rotating system. As with the Milky Way and other galaxies, there is a deficit of GCs in the very inner galactic regions. An interesting difference in the M31 GCs is the presence of abundance anomalies. In particular, the CN and Ca H+K lines have enhanced line strengths relative to Milky Way GCs. It has been suggested that this difference may have a primordial origin.

M33 and other Local Group galaxies

The late-type spiral M33 contains ∼25 old GCs, and perhaps as many again which are young or intermediate aged. The magnitude distribution, mean metallicity and kinematics of the old GCs appear similar to the Milky Way halo system. Most other Local Group galaxies are dwarf spheroidals, many of which possess a handful of GCs. These GCs tend to be metal-poor and show similarities in their color–magnitude diagrams to the young halo GCs in our own Galaxy. Direct evidence for accretion of such GCs comes from the Sgr dwarf (and its 3–4 GCs) which is in the process of merging with our Galaxy. Most dwarf galaxies have relatively normal S_N values but the Fornax dwarf has $S_N \sim 20$ indicating that GC formation can be very efficient in some small galaxies.

Luminosity function

The GC luminosity function (GCLF) is simply the number of GCs per luminosity interval (see also LUMINOSITY FUNCTION OF GALAXIES). If the number is plotted against magnitude then it is well approximated by a Gaussian distribution, i.e.

$$N = A\exp[-(M - M^0)^2/2\sigma^2]$$

where σ is the width of the Gaussian, M^0 is the turnover or peak magnitude and A is the number of GCs at the peak. An example of a GCLF for a giant elliptical is shown in figure 1. What is remarkable about the GCLF is that the peak is nearly universal, i.e. all well-studied systems

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

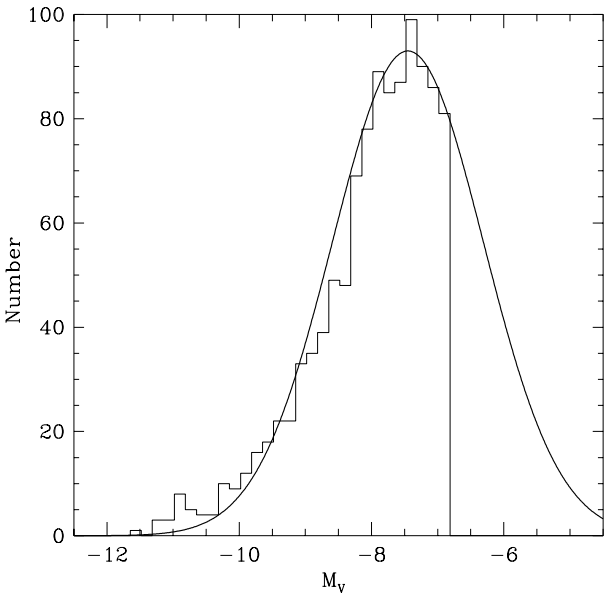


Figure 1. GCLF for the giant elliptical galaxy NGC 5846. The histogram shows the binned GC counts and the solid curve shows the best-fit Gaussian with a peak magnitude of $M_V^0 = -7.4$.

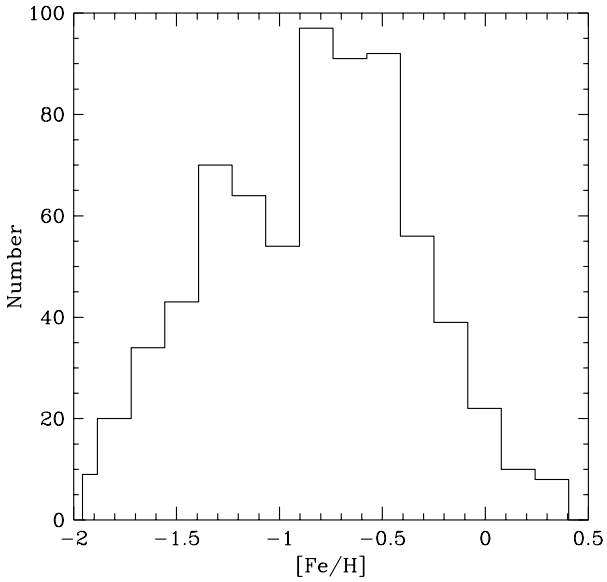


Figure 2. GC metallicity histogram for the giant elliptical galaxy NGC 5846. The histogram reveals clear evidence of a bimodal metallicity distribution indicating two distinct populations of GCs. Metallicities are estimated from photometry assuming a standard color-to-metallicity transformation.

have a peak at $M_V^0 = -7.4 \pm 0.2$. For a V band mass-to-light ratio of 2 this corresponds to a mean mass of $1.6 \times 10^5 M_\odot$. The small deviations from a universal M_V^0 value do exist and various causes have been suggested, namely luminosity, Hubble type, color (metallicity) or environmental effects. This issue has yet to be settled. However, the mass function of GCs seems to be even more robust, and closely resembles that of giant molecular clouds (GMCs). Prompted by this observation, it has been suggested that GCs form in the cores of large GMCs.

If GCs do indeed have a nearly universal mean absolute magnitude, then this can be used as a standard candle for distance determination. This has many advantages over other distance determination methods: GCs are available in large numbers around all large galaxies, they do not require repeat measurements (e.g. the study of variable stars) and the method can work for a variety of Hubble types. Its main disadvantage is perhaps a lack of a well-established theory to explain the preferred mass of GCs. The Hubble Space Telescope has made a significant impact in this area with its ability to obtain high-precision photometric data several magnitudes fainter than the peak. The GCLF method now shows excellent agreement with surface brightness fluctuations and Cepheid variable determined distances, giving values of the Hubble constant around 75 km s^{-1} Mpc^{-1}.

Metallicity distribution

In the last few years it has become clear that several, perhaps all, massive early-type galaxies have two distinct populations of GCs. These subpopulations are apparent in high-quality photometric observations as a bimodal color distribution. It is generally assumed that the color separation in the two peaks largely reflects differences in metallicity, and to a lesser extent age. The typical separation between the relatively metal-poor and metal-rich peaks is ∼1 dex in [Fe/H]. Breaking the age–metallicity degeneracy for GCs with spectroscopy and near-IR photometry is an active area of research. For the two cD galaxies studied to date (i.e. M87 and NGC 1399), their GC (metal-poor and metal-rich) populations appear to be quite old (∼10–15 Gyrs).

The relative numbers in each population are typically similar, although there is some evidence that the ratio of metal-poor to metal-rich GCs scales with S_N. In other words, high S_N galaxies have relatively more metal-poor GCs. An example of a bimodal metallicity distribution, for an elliptical with a low S_N value, is shown in figure 2.

The presence of two GC populations appears to explain the long-standing observation that the mean metallicity of the GC system is ∼0.5 dex lower in [Fe/H] than the stars in massive ellipticals. This simply arises because the metal-rich GCs have a similar metallicity to the host galaxy stars while the metal-poor ones are ∼1 dex lower in [Fe/H]. Radial metallicity gradients for the entire GC system are due to the changing relative mix of the two populations with galactocentric radius (i.e. the metal-poor to metal-rich ratio increases with radius). Thus there may be no true radial metallicity gradient, suggesting a dissipationless formation process.

The existence of multimodal color distributions indicates that GCs formed in distinct star formation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

episodes from different metallicity gas, perhaps at different times. A simple monolithic collapse (expected to result in a unimodal color distribution) appears to be ruled out by such data. Interestingly, unimodal distributions may be present in low-luminosity ellipticals although the smaller number of GCs present makes any bimodality more difficult to detect.

Metallicity–luminosity relation

The correlation between the mean metallicity of the GC system and the luminosity of the host galaxy has been subject to much debate. As the number of systems studied has increased, it now seems clear that some relationship exists from dwarf to cD galaxies. The type of relation, contribution of observational errors and second-parameter effects are the subject of ongoing debate. For a linear fit, the slope of the GC system metallicity–galaxy luminosity relation is $Z \propto L^{0.4}$. This is similar to the slope of the galaxy metallicity–galaxy luminosity relation. Taken together these indicate that the GC system has had a similar chemical enrichment history to its host galaxy.

If we subdivide the GC system into the mean metallicity of the metal-poor and metal-rich peaks, then we find that it is the metal-rich peak which correlates with galaxy luminosity. This provides further evidence that the metal-rich GCs are closely coupled to the formation of the galaxy itself. The metal-poor peak shows little or no correlation.

Spatial distribution

Radial distribution

When plotted logarithmically, the radial surface density distribution for a GC system typically shows a flat 'core' region (i.e. approaching zero slope) in the central few kpc and a roughly power-law-like decline to larger galactocentric radii. The GC system surface density is often more extended than that of the galaxy starlight. Low-S_N galaxies cover a range of power-law slopes but high-S_N galaxies only have relatively flat slopes (i.e. an extended distribution). For cD galaxies there is a change in slope at the radius of the cD envelope indicating that the outer GCs and the stellar envelope have a similar origin.

Azimuthal distribution

Less is known about the azimuthal distribution of GCs than their radial profile. However, it appears that GCs generally track the position angle and ellipticity variations of the underlying galaxy starlight. In the future, a number of GC systems may have been studied in enough detail to make comparisons with the spatial properties of x-ray-emitting gas. This will provide a unique probe of the dark matter distribution at large galactocentric radii.

Kinematic properties

The kinematics of large samples of GCs has been determined for only a few galaxies to date, although the combination of multiobject spectrographs on large telescopes is rapidly improving this situation. Studying the kinematics of the subpopulations promises to yield important clues as to their origin and ultimately place new constraints on the formation and dynamical evolution of their host galaxy.

An additional use of GC kinematic studies is their application to mass estimates of galaxies. They act like test particles in the gravitational potential and have the advantage of extending well beyond the optical limit of the galaxy. Masses are typically determined using either the projected-mass method or the virial theorem. For a system of N GCs the projected-mass is

$$M_{PM} = \frac{f_p}{G(N-\alpha)}(\Sigma_i^N V_i^2 r_i)$$

where f_p is a numerical factor which describes the type of GC orbits, α is a small correction factor, V_i is the velocity difference between the ith GC and the system mean velocity and r_i is the radial separation from the system centroid position. For small samples of GCs, the mass is likely to be overestimated. The mass from the virial theorem is

$$M_{VT} = \frac{3\pi N}{2G}\frac{\Sigma_i^N V_i^2}{\Sigma_{i<j} 1/r_{ij}}$$

where r_{ij} is the radial separation between the ith and jth GC.

Only a few galaxies have mass estimates from GC motions, although this is likely to be an active area in the coming years. Such estimates can be compared with those from x-ray satellites for massive ellipticals. However, for small ellipticals, without extended hot gas halos, GC-based estimates may be the only feasible method for estimating mass at large galactocentric radii.

Formation and destruction

The origin of GC systems is not currently known. Traditionally models fell into three main categories: primary (formation before the galaxy), secondary (formation of GCs and galaxy contemporaneously) and tertiary (formation after galaxies). The current formation ideas tend to be more hybrid in nature, motivated by the recent discoveries of two GC populations in massive ellipticals. Ideas include creation of new GCs in the gaseous merger that forms the elliptical, *in situ* formation plus accretion from smaller galaxies and *in situ* formation in two distinct star formation phases. The reality may be a combination of these basic processes (i.e. *in situ* formation, tidal accretion and the creation of new GCs), operating at varying degrees in different systems.

Once formed, a GC can be destroyed by a variety of internal and external mechanisms. In the case of protoglobular clusters, they will disrupt themselves if supernovae inject more energy than is binding the cluster together. Low-mass protoglobular clusters may evaporate or be tidally shocked. Either way low-mass clusters must be destroyed if their observed power-law luminosity function is to eventually resemble the log-normal distribution of old GCs. Other long-term

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

dynamical destruction effects include bulge and disk shocking. Disk shocking may destroy a GC as it passes through the increased density of a galactic disk. Bulge shocking is strongest in galaxies with high central densities, i.e. low-luminosity ellipticals. Only recently have predictions for GC destruction rates in different types of galaxies become available.

Bibliography

The literature on GCs in external galaxies is growing rapidly. Below is a brief, and incomplete, selection.

Reviews of GC systems can be found in

Ashman K M and Zepf S E 1998 *Globular Cluster Systems* (Cambridge: Cambridge University Press)
Harris W E 1991 *Ann. Rev. Astron. Astrophys.* **29** 543

A review of GC luminosity functions as distance indicators is given by

Whitmore B C 1996 *The Extragalactic Distance Scale* ed M Livio, M Donahue and N Panagia (Cambridge: Cambridge University Press)

The first photometric studies of GCs beyond the Local Group include

Hanes D A 1979 *Mon. Not. Astron. Soc.* **188** 901
Racine R 1968 *Pub. Astron. Soc. Pac.* **80** 326
Strom S E *et al* 1981 *Astrophys. J.* **245** 416

Photometric studies revealing two GC subpopulations include

Harris W E *et al* 1992 *Astron. J.* **104** 613

for NGC 5128,

Lee M G and Geisler D 1993 *Astron. J.* **106** 493

for M87,

Secker J *et al* 1995 *Astron. J.* **109** 1019

for NGC 3311,

Zepf S E and Ashman K M 1995 *Astrophys. J.* **443** 570

for NGC 3923,

Geisler D, Lee M G and Kim E 1996 *Astron. J.* **111** 1529

for NGC 4472,

Forbes D A, Brodie J P and Huchra J 1997 *Astron. J.* **113** 887

for NGC 5846 and

Forbes D A *et al* 1998 *Mon. Not. R. Astron. Soc.* **293** 325

for NGC 1399 and 1404.

Spectroscopic studies of extragalactic GCs include

Brodie J P and Huchra J P 1991 *Astrophys. J.* **379** 157
Grillmair C J *et al* 1994 *Astrophys. J.* **422** L9
Kissler-Patig M *et al* 1998 *Astron. J.* **115** 105
Mould J R, Oke J B and Nemec J M 1987 *Astron. J.* **92** 53

Recent attempts to explain the presence of two GC subpopulations in massive ellipticals include

Ashman K M and Zepf S E 1992 *Astrophys. J.* **384** 50
Cote P, Marzke R O and West M J 1998 *Astrophys. J.* **501** 554
Forbes D A, Brodie J P and Grillmair C J 1997 *Astron. J.* **113** 1652

The formation of GCs from large giant molecular clouds is described in

Harris W E and Pudritz R E 1994 *Astrophys. J.* **429** 177

and subsequent papers.

Duncan A Forbes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globular Cluster X-ray Sources

Twelve of the 147 globular clusters associated with our Galaxy contain a bright ($L_x \gtrsim 10^{29}$ W) x-ray source; many more contain dim ($L_x \lesssim 10^{28}$ W) x-ray sources, of which more than 40 are now known. X-ray sources have also been detected in globular clusters of the ANDROMEDA GALAXY M31.

The x-ray properties of the bright cluster sources in our Galaxy are similar to those of the low-mass x-ray binaries in the galactic disk (see also X-RAY BINARY STARS). The bright x-ray sources in globular clusters are therefore thought to be binaries in which a NEUTRON STAR accretes matter from a low-mass companion.

The nature of the dim x-ray sources in globular clusters is more diverse. A single cluster can contain several dim sources. Soft x-ray transients, cataclysmic variables, recycled radio pulsars, and RS CVN binaries all can have x-ray luminosities in the range observed for the dim sources, and are known to exist in globular clusters.

Twelve bright cluster sources correspond to about 10% of the bright x-ray sources of our Galaxy, a surprisingly large fraction if one considers that the globular cluster system holds less than 0.1% of the stars in our Galaxy. The efficient formation of x-ray binaries in globular clusters is ascribed to two processes: tidal capture and exchange encounters. Tidal capture occurs when a neutron star transfers some of its kinetic energy to tides in another star during a close passage, and enough tidal energy is dissipated to bind the neutron star in orbit around its captor. An exchange encounter occurs when a neutron star ejects one of the stars in a binary in a close encounter, and takes its place. These processes are efficient especially in globular clusters with dense cores, where the average distance between stars is relatively small.

Table 1. Table of the bright x-ray sources in globular clusters, giving the position of the source, its (approximate) x-ray luminosity, and where known the binary period. All sources except Terzan 6 are x-ray bursters. For transients, marked with T, the luminosity at maximum is given.

Cluster	Position	$\log L_x$ (W)	P_b
NGC1851	0512 − 40	29.2	
NGC6440	1745 − 20	T30.4	
NGC6441	1746 − 37	29.6	5.7 h
NGC6624	1820 − 30	30.8	11.4 min
NGC6652	1836 − 33	T29.4	
NGC6712	1850 − 09	29.1	20.6 min
NGC7078	2127 + 12	29.4	17.1 h
Terzan 1	1732 − 30	29.3	
Terzan 2[a]	1724 − 31	29.5	
Terzan 5	1745 − 25	28.8	
Terzan 6	1751 − 31	T29.7	12.4 h
Liller 1[a]	1730 − 33	T30.0	

[a] X-ray source discovered before the globular cluster!

The bright x-ray sources

The x-ray luminosities of the bright x-ray sources in globular clusters are between $L_x \simeq 10^{29} - 10^{31}$ W (see table 1). The color temperature of the x-ray spectrum in the 2–10 keV energy range is $kT_c \leq 5$ keV. At these luminosities, the x-ray source must be either a neutron star or a BLACK HOLE, that accretes matter from a companion. Eleven of these twelve cluster sources have shown x-ray bursts. (An x-ray burst is a sudden rise in the x-ray luminosity followed by a more gradual decline (see X-RAY BURSTERS), and is thought to be caused by explosive fusion into carbon of a thin helium layer on the surface of a neutron star, see figure 1.) These eleven sources are accreting neutron stars. The remaining one—in Terzan 6—may be a neutron star as well; if not, it is a black hole.

Some of the bright sources are transients, i.e. bright only occasionally. The source in NGC6440, for example, was detected in December 1971/January 1972 and in August 1998; between these dates its flux level decreased to less than 0.001 of the outburst level.

The first ever measurement of the radius of a neutron star was made from an x-ray burst of the source in the globular cluster Terzan 2 (figure 1). The spectrum can be fitted with a black body of temperature $T_{\rm bb}$; this temperature decreases during the burst. The luminosity of the burst is $L_x = 4\pi R^2 \sigma T_{\rm bb}^4$, where R is the radius of the neutron star. Assuming that the luminosity is emitted isotropically, and that the globular cluster is at a distance D (which can be determined in the optical), we can determine R from the flux f_x measured on Earth via the equation

$$f_x = \left(\frac{L_x}{4\pi D^2}\right) = \left(\frac{R}{D}\right)^2 \sigma T_{\rm bb}^4. \qquad (1)$$

For the burster in Terzan 2, $R \sim 10$ km. The uncertainty in the radius is appreciable, due mainly to deviations of the x-ray spectrum from a black body, and also to uncertainty of the distance to the cluster, and to uncertainty in the absorption by the interstellar medium. Nonetheless, x-ray bursts prove that the x-ray sources are very small, compatible with the theoretically predicted size of a neutron star.

The hypothesis that the neutron star is accreting from a binary companion has been confirmed by the determination of the orbital period for five sources; for three of these sources a counterpart has been identified at optical and ultraviolet wavelengths. Surprisingly, the orbital periods of two sources are the shortest known for x-ray binaries, 685 s in NGC6624 and 1236 s (or 792 s) in NGC6712; in these cases the mass donor must be an undermassive ($M < 0.1 M_\odot$) white dwarf. The 12.4 and 17.1 h orbital periods of the sources in Terzan 6 and NGC7078 (M15) point to a subgiant donor; the (less secure) 5.7 h period in NGC6441 suggests a donor on the main sequence.

The x-ray source in the globular cluster Liller 1 is a recurrent transient, showing outbursts lasting a month at irregular intervals of roughly half a year; when bright,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

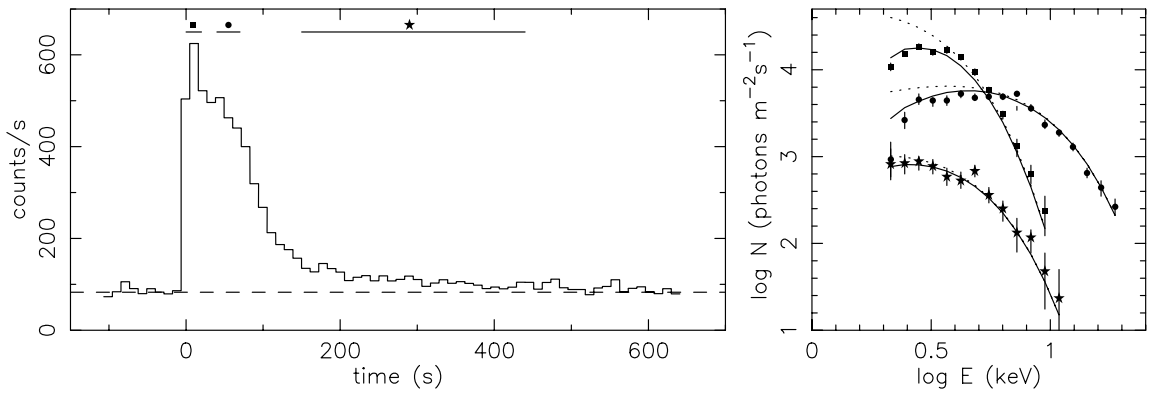


Figure 1. Left: An x-ray burst of the source in the globular cluster Terzan 2, observed with the OSO-8 satellite. Right: spectra of the x-rays at subsequent phases of the burst, with black body fits (full curve: as observed; dotted curve: corrected for interstellar absorption). From the fit of the spectrum 40–70 s into the burst, a radius of ∼10 km is derived for the neutron star.

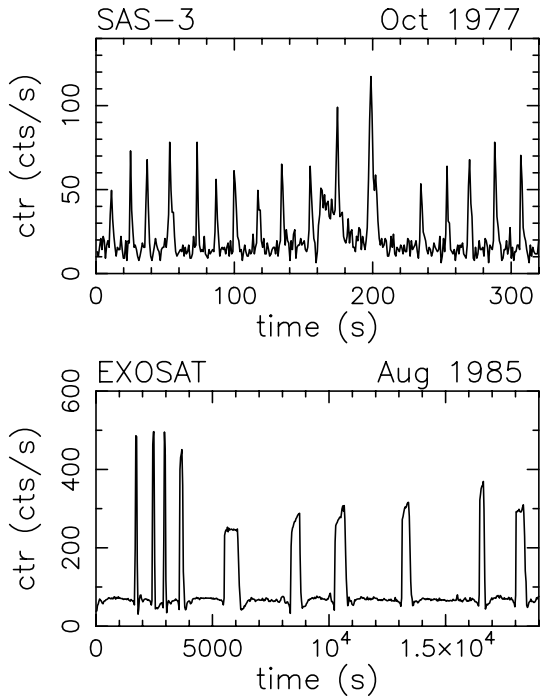


Figure 2. The Rapid Burster in Liller 1 emits its x-rays mainly in discrete bursts. The burst rate was high when the source was discovered in 1977 (top), but rather lower during some later observations (below). A thermonuclear burst starts in the 1977 observation at $t = 160$ s.

this source is remarkable because it emits all its x-rays in discrete bursts (figure 2). When first discovered in 1977 these bursts lasted a few seconds and occurred at a rate of four per minute, and the source was named the Rapid Burster. Later observations have found the source emitting its x-rays in longer bursts (up to $\sim 10^3$ s) at longer intervals (up to about an hour); and occasionally emitting a steady x-ray flux. The nature of these x-ray bursts is not clear; they are different from the thermonuclear bursts observed in many other systems in having a different time scale and in the absence of temperature evolution. Maybe they correspond to discrete accretion events onto the neutron star. The rapid bursts are occasionally interspersed by ordinary, thermonuclear bursts. Thermonuclear bursts are often called type I, and accretion bursts type II.

Some 30 x-ray sources in M31 have been identified with globular clusters of that galaxy. Their x-ray luminosities are comparable with those of low-mass x-ray binaries in (the globular clusters of) our own Galaxy.

The dim sources

At lower flux levels ($f_x \sim 10^{-17}$ W m^{-2}, say) fore- or background sources must be considered; as a result, dim sources can be safely asserted—in a statistical sense—to be related to globular clusters only when they are found in or close to the cluster cores. Currently some 40 of such dim ($L_x \simeq 10^{24} - 10^{28}$ W) sources are known in some 15 clusters, i.e. several clusters contain multiple sources.

The positional accuracies of the x-ray sources hamper identification with optical or ultraviolet counterparts, even if one limits onself to blue or variable objects (figure 3). The situation will improve when observations from the CHANDRA X-RAY OBSERVATORY become available. The only secure counterpart is the recycled pulsar of which the rotation period of 3.054 ms is present in the x-ray flux from NGC6626 (M28). Cataclysmic variables have been found with the Hubble Space Telescope as plausible counterparts for dim x-ray sources in NGC6397, 47 Tuc and NGC6752. In NGC5904 (M5) a dwarf nova discovered well outside the core in optical observations in 1941, has been detected in x-rays. For most dim x-ray sources in globular clusters no optical counterpart has as yet been found.

The x-ray flux of the dim source in NGC5272 varies by at least a factor 30, and is also remarkable because of its very soft spectrum, with a characteristic temperature $\lesssim$50 eV, when the source is bright.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

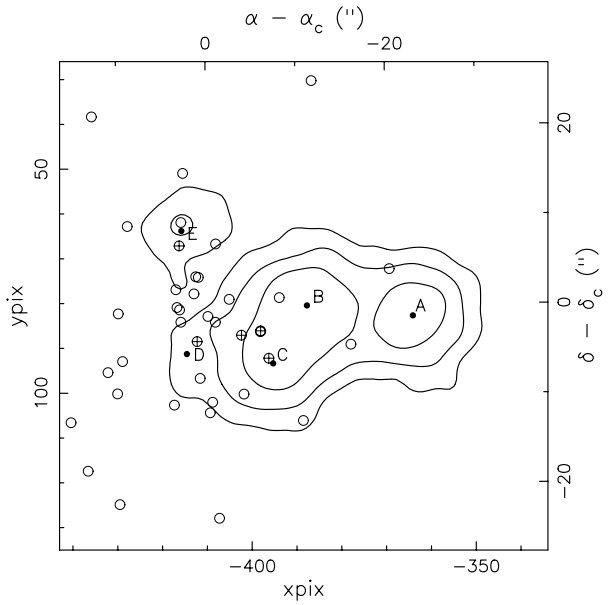


Figure 3. X-ray image of the core of the globular cluster 47 Tuc (contours at 0.4, 0.16 and 0.064 times the maximum level observed, best positions for five x-ray sources A–E indicated with •). Blue and variable stars in the core are indicated with ◦. The limited accuracy of the reconstruction of the satellite pointing leaves an uncertainty of about 2 arcsec between the x-ray coordinate system (lower and left axes, in pixels of the ROSAT HRI detector) and optical coordinates (upper and right axes, with respect to the center of 47 Tuc). Of the suggested counterparts (indicated with ⊕) two, both cataclysmic variables, are good candidates for x-ray sources C and D.

The detection of dim sources in globular clusters is subject to strong selection effects: they can be found only in relatively nearby clusters, which are not highly reddened. Taking these effects into account, one finds that the number of x-ray sources $\mathrm{d}N$ in a luminosity interval $\mathrm{d}L_x$ depends on the luminosity L_x and on the number density n of stars in the cluster core as $\mathrm{d}N \propto n^{1.5} {L_x}^{-1.5}\,\mathrm{d}L_x$. This scaling is in agreement with the observation that the total x-ray luminosity of a globular cluster is dominated by individual sources.

Theory of formation mechanisms

The fraction of neutron stars or WHITE DWARFS expected to have experienced a close stellar encounter can be several per cent in a globular cluster with a dense core. In the disk of the Galaxy—where close stellar encounters don't occur, and where binaries evolve unperturbed by third bodies—only a tiny fraction of neutron stars is found in a binary. A neutron star binary in a globular cluster is therefore much more likely to be the result of a close stellar encounter than of the unperturbed evolution of a primordial binary in the cluster. The fraction of white dwarfs in a binary in the disk of the Galaxy is much higher: a white dwarf in a binary in a cluster can therefore be the result either of a close stellar encounter or of ordinary binary evolution.

The basic principle of tidal capture of a neutron star can be understood with a simple calculation: consider a neutron star with mass m and velocity v at large distance ('infinity') relative to a main-sequence star with mass M and radius R. The relative kinetic energy E_k of the two stars is given by

$$E_\mathrm{k} = \frac{1}{2}\frac{mM}{m+M}v^2. \tag{2}$$

As the neutron star closes in, it deforms the other star. The height h of the bulge, and its mass m_t can be estimated for distance d with

$$h \simeq \frac{m}{M}\frac{R^4}{d^3};\; m_t \simeq k\frac{h}{R}M \simeq k\left(\frac{R}{d}\right)^3 m \tag{3}$$

where k is the apsidal motion constant, which depends on the central condensation of the star, and indicates how easy it is to deform the star. For a star with a deep convective envelope $k \simeq 0.14$. Thus the energy E_t in the tidal deformation is of order

$$E_t \simeq m_t\frac{GM}{R^2}h \simeq k\frac{Gm^2}{R}\left(\frac{R}{d}\right)^6. \tag{4}$$

If $E_t > E_\mathrm{k}$, the two stars cannot escape from one another anymore, and a binary is formed. This condition can be written:

$$d \lesssim 3R\left(\frac{m}{M}\frac{m+M}{2M_\odot}\frac{R_\odot}{R}\right)^{1/6}\left(\frac{10\ \mathrm{km\ s^{-1}}}{v}\right)^{1/3}. \tag{5}$$

Because of the strong dependence of E_t on d, this rough estimate is in fact pretty accurate, as more detailed calculations confirm.

The initial binary orbit is highly eccentric, $e \lesssim 1$, with a velocity at PERIASTRON close to the escape velocity. Tidal forces are expected to circularize the orbit, during which process angular momentum is conserved. The semimajor axis a_c of the circularized orbit is given by

$$a_c \leq 2d \ll a_e \tag{6}$$

where a_e is the semimajor axis immediately after capture. Thus the final orbit after circularization can have a semimajor axis of up to twice the capture distance.

The energy difference ΔE between the initial, highly eccentric, and the final circularized orbit is comparable to the binding energy E_* of the main-sequence star.

$$\frac{\Delta E}{E_*} = \left(\frac{-GMm}{2a_e} - \frac{-GMm}{2a_c}\right)\left(\frac{3GM^2}{5R}\right)^{-1} \simeq \frac{5}{6}\frac{m}{M}\frac{R}{a_c}. \tag{7}$$

The circularization process provides enough energy to destroy the main-sequence star. If the energy has to be dissipated very rapidly, it is hard to avoid destruction of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the donor. It has been suggested, however, that the energy present in the tidally induced oscillations of the star can be fed back into the orbit; in a very eccentric orbit, the energy exchange between (oscillations in) the star and the orbit is chaotic. This would mean that the star has much more time to dissipate, and thus has a much better chance of surviving.

The cross section σ for closest passage within distance d follows from conservation of energy and angular momentum in a Keplerian orbit:

$$\sigma = \pi d^2 \left(1 + \frac{2G(m+M)}{v^2 d}\right) \simeq \pi d \, \frac{2G(m+M)}{v^2}. \quad (8)$$

The second term within brackets gives the effects of gravitational focusing. This term dominates for the small relative velocities between stars in globular clusters, which justifies the subsequent approximation.

With number densities n_c and n for the neutron and target stars respectively, the capture rate of neutron stars per unit volume can be written:

$$\Gamma = n_c n v \sigma. \quad (9)$$

To obtain the formation rate in a cluster, one must integrate equation (9) over the cluster volume. Because the density is strongly peaked near the cluster center, virtually all close encounters occur in the cluster core. In a relatively dense core of a globular cluster, $n_c \sim 100$ pc^{-3} and $n \sim 10^5$ pc^{-3}. With a characteristic core volume of $\sim$1 pc^3 and stellar velocities of $\sim$10 km s^{-1} it follows from that a close binary with a neutron star is formed every 10^9 yr. For an average lifetime of a bright source of 10^9 yr, we then expect to see of order 1 x-ray source in such a cluster, in accordance with observations.

If a neutron star directly hits a giant, then the core of the giant and the neutron star may spiral in towards one another, leaving a close binary of an undermassive white dwarf and a neutron star. Such an encounter may explain the x-ray sources in NGC6624 and NGC6712.

Another way of getting a neutron star in a binary is an exchange collision: when a neutron star approaches a binary to a distance comparable with the semimajor axis of the binary, the three stars may temporarily move in complicated orbits around one another. Such a three-body system is not stable, and one star is ejected at the end, usually the lightest star. In this way, a neutron star may take the place of an ordinary star in a binary. The frequency of such encounters is proportional to the number density of binaries $n_{\rm bin}$ in the cluster core, and to the semimajor axes a of these binaries. The capture rate Γ_3 from these three-body processes may be compared with tidal capture, roughly according to:

$$\frac{\Gamma_3}{\Gamma} \simeq \frac{n_{\rm bin}}{n} \frac{a}{R} \quad (10)$$

when all three stars have equal mass. Thus the smaller number of binaries is offset by their larger size, as compared to single stars. The actual importance of exchange encounters is not clear, because the number of binaries in cluster cores is not known.

The formation of a close neutron-star binary in a globular cluster may also arise from a tidal capture between stars involved in a three-body encounter. Two-body or three-body encounters can also cause direct collisions between stars, a process which may be important for the formation of single recycled pulsars in globular clusters.

Bibliography

Bailyn C D 1995 Blue stragglers and other stellar anomalies: implications for the dynamics of globular clusters *Ann. Rev. Astron. Astrophys.* **33** 133–62

Hut P *et al* 1992 Binaries in globular clusters *Publ. Astron. Soc. Pacific* **104** 981–1034

Johnston H M and Verbunt F 1996 The globular cluster population of low-luminosity x-ray sources *Astron. Astrophys.* **312** 80–7

Supper R *et al* 1997 ROSAT PSPC survey of M31 *Astron. Astrophys.* **317** 328–49

F Verbunt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globular Clusters

Globular clusters (GCs) are aggregates of approximately 10^4–10^6 gravitationally bound stars, highly concentrated to the center, spread over a volume ranging from a few dozen up to more than 300 light-years (ly) in diameter. They resemble shining, old islands orbiting the Milky Way. As the name indicates, GCs show a largely spherical symmetry about their centers. A picture of the classic GC ω Centauri is shown in figure 1.

The stellar density in the cluster's center is so high (up to a few 10^3 stars ly^{-3}) that it is generally impossible to separate the individual stars from ground-based observations. Only recently has the refurbished Hubble Space Telescope (HST) allowed astronomers to dig into the very central regions of many Galactic globulars, where members (sometimes peculiar or even exotic) move randomly like molecules of gas, interacting according to the basic laws of gravity.

Early studies of GCs date back to the birth of modern astronomy. Since then, GCs have continued to offer excitement to both professional astronomers and sky-lovers with surprising results, and they constitute a basic benchmark for our astrophysical understanding.

The Milky Way hosts about 200 GCs. They form a halo of roughly spherical shape which is highly concentrated around the Galactic center, in the Sagittarius–Scorpius–Ophiuchus region. The most distant Galactic globulars (such as NGC 2419) are located far beyond the edge of the Galactic disk, at distances out to 300 000 ly.

RADIAL VELOCITY measurements have shown that most of the GCs are orbiting the Galaxy in highly eccentric elliptical orbits (see figure 2), with orbital periods of about 10^8 yr or even longer.

While following their orbits around the Galactic center, GCs are subject to a variety of perturbations (tidal forces from the parent galaxy, passage through the Galactic plane, star escape, internal dynamical evolution, etc) which make the existing GCs perhaps just the survivors of a much wider population, partially disrupted and spread out throughout the Galactic halo and far beyond. In this respect, it has been estimated that, within the next ten billion years or so, most of the present Galactic GCs could disappear. On the other hand, we know today that four clusters in Sagittarius (M54 in particular) are likely members of the Sagittarius Dwarf Elliptical Galaxy (discovered in 1994), currently merging into the central regions of the Milky Way.

A large majority of the galactic GCs have high relative velocities (100–300 km s^{-1}) with respect to the Sun, as they do not participate in the Galactic disk rotation. There is, however, a subsample, commonly referred to as 'disk globulars', which show properties closely connected to the disk.

Spectroscopic observations of stars in Galactic GCs have revealed that their chemical composition differs from that of the Sun in heavy elements content. GC stars are in fact typically metal poor and old. This is a signature that they were presumably born during the early stages of the Galaxy's formation and thus represent a sort of archeo-astronomical site where the universe in its youth can be studied.

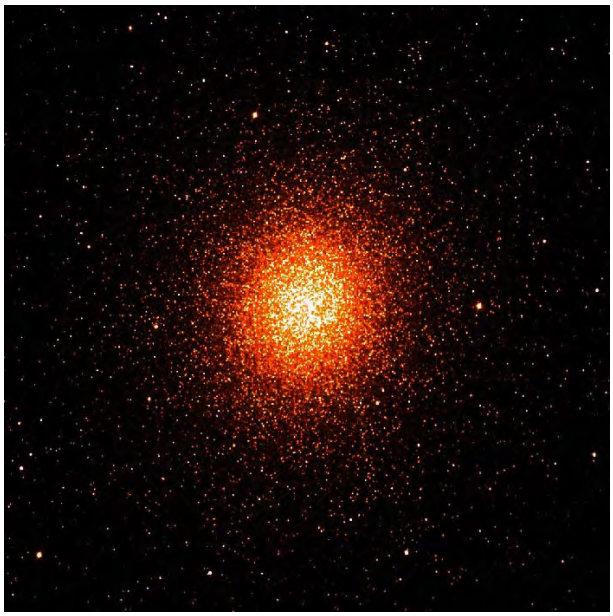


Figure 1. The GC ω Centauri.

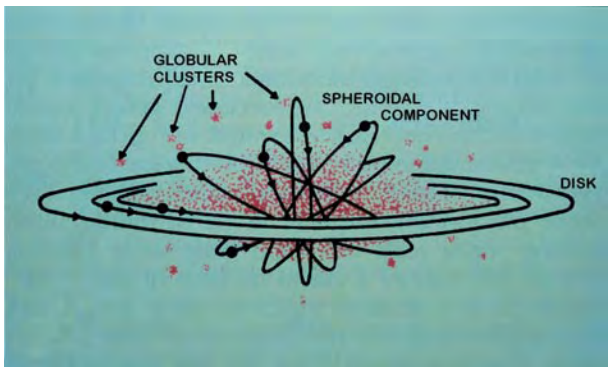


Figure 2. Orbits of GCs in the Galaxy.

Globular clusters seem to be ubiquitous. Edwin Hubble pionereed the search for globulars in the galaxies of the LOCAL GROUP with the detection of about 100 GCs in M31, the Andromeda Nebula (more than 350 GCs are known nowadays). However, it was only in the 1970s that the identification of any significant number of GC candidates became feasible around galaxies beyond the Local Group. The main reason for this difficulty in the search is that, with increasing distance, a typical cluster becomes progressively indistinguishable in shape from foreground stars or distant background galaxies.

It is now fairly well established that almost all galaxies have GC systems, in some cases (e.g. M87) containing several thousands of globulars (see GLOBULAR

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CLUSTER SYSTEMS IN NORMAL GALAXIES). There are, however, important differences. While all the globulars in our Galaxy and in M31 are old (ages of about 10 Gyr, at least), there are galaxies, such as the two Magellanic Clouds and M33 (the Triangulum Galaxy), hosting much younger GCs (ages of a few Gyr, or less).

The latest GC searches also reveal that dense, massive star clusters seem to be currently forming in the halos of some interacting galaxies (see GLOBULAR CLUSTER SYSTEMS IN INTERACTING GALAXIES). These objects are commonly interpreted as young and metal-rich GCs. This idea is not universally accepted, however. In fact, observational evidence, still quite meager, needs to be confirmed. Furthermore, astronomers seem somehow reluctant to change their traditional view of the GCs and to admit that massive, young and metal-rich GCs could possibly be even more frequent in the universe than the 'classic', ancient and very metal-poor ones.

Historical background

Perhaps the first historical detection of a globular cluster goes back to the mists of time, when human eyes first saw ω Centauri, the biggest Galactic GC, barely visible in the southern hemisphere.

The first 'astronomical' detection dates back to the 18th century. John Herschel, in the 1830s, realized that a large number of these clusters are concentrated in a relatively small portion of the sky in the direction of Sagittarius. Later on, HARLOW SHAPLEY detected variable stars in several GCs and, on the assumption they were Cepheids of known (calibrated) absolute magnitude, derived distances to them and to the Galactic center. Doing this, in 1917, Shapley understood that the Galactic center is located very far away from the Sun, in the direction of Sagittarius, and was also able to estimate the size of the Milky Way.

Today we know that Shapley significantly overestimated (by a factor of 2 or so) the size of the GCs system and of the Milky Way as a whole, mainly because the cluster variables he identified as Cepheids are actually RR Lyraes, whose absolute magnitude is about 2–4 mag fainter than Cepheids.

A key role in astronomy

Individual clusters as well as GC systems are of great worth as specific targets but also represent a powerful tool to obtain a deep insight into a large variety of astrophysical and cosmological problems (see table 1). Their study still represents a benchmark and a major field of interest in the international astronomical community.

Members of any given GC share a common history and differ one from another only in their initial mass. Consisting of a 'simple' STELLAR POPULATION (i.e. coeval, chemically homogeneous and isolated), GCs are ideal laboratories for testing the theories of stellar structure and evolution. In fact, thanks to the very large number of stars, almost every evolutionary stage (even those with very short lifetimes, down to a few 10^4 yr) is present at the appropriate statistical significance among the GC stars. This allows a direct check on the validity of the detailed evolution theory. When GCs are considered just as a million or so pointlike masses in a small volume, subject to internal and external dynamical interactions, they represent an ideal workbench to study STELLAR DYNAMICS and to test most exquisite theoretical dynamical models.

If studied as a global system, GCs constitute fossil tracers of the dynamical and chemical evolution of the parent galaxy and can be used as test particles to evaluate both the galaxy's total mass and its radial distribution.

GCs contain a variety of exciting objects by themselves worth a continuous investigation, for instance strong and weak x-ray sources, neutron stars and millisecond pulsars, white dwarfs, cataclysmic variables, binaries, blue stragglers, planetary nebulae, etc. Moreover, they contain one of the most popular intrinsic variable stars, the so-called RR LYRAE STARS. These stars have light variation amplitudes less than a couple of magnitudes and periods ranging from 0.2 to 1.1 days. Since their mean absolute magnitude is constant and fairly independent of metallicity (to within 0.3 mag), the RR Lyrae variables and the GCs, in turn, are ideal standard candles to measure distances.

Perhaps one of the most remarkable impacts of GC reasearch on other fields of astronomy is provided by the estimate of the ages of the Milky Way's globulars. GCs are, in fact, among the few objects in the Galaxy for which relatively precise ages can be derived. Since they are the oldest objects observed in the Milky Way so far, and were born during the very early stages of the Galaxy's formation, they provide a very stringent lower limit to the age of the universe. On the other hand, their age distribution and how ages vary with varying metallicity, spatial location in the Galaxy and kinematic properties make these systems direct tracers of the chronology of the first epoch of star formation in the Galactic halo and may help in understanding the whole process of galaxy formation.

Color–magnitude and Hertzsprung–Russell diagrams, and stellar evolution

The color–magnitude diagram (CMD) as well as its twin, the HERTZPRUNG–RUSSELL DIAGRAM (H–R diagram), substantially plot the temperature of a star on the X-axis, with increasing temperature to the left, and the star brightness on the Y-axis, with brighter stars at the top. Traditionally, the observational CMD reports on the X-axis the color of the stars (generally $B–V$) and on the Y-axis the observed apparent magnitude (generally V) or, if distance is known, the absolute magnitude (M_V).

The CMD is a basic, very powerful tool which allows a direct calibration of the observables in terms of fundamental intrinsic parameters (e.g. metallicity, age) as well as stringent comparisons to be made with theoretical model predictions (after 'tricky' transformations from the theoretical into the observational plane).

Figure 3(*a*) shows the observed CMD of the GC M3, and figure 3(*b*) displays the schematic CMD of a typical GC.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Importance of GCs.

Subject	Reasons for importance
Witnesses of the early Galactic evolution	• First to form • Chemically uncontaminated
Stellar Evolution Laboratories	• Simple stellar populations • Test of the 'stellar clock'
Distance indicators	• Standard candles: the RR Lyrae stars • GC system integrated luminosity function
Age indicators	• The turn-off luminosity = 'the clock' absolute ages: lower limit to the age of the universe relative ages: 'second parameter' and Galaxy formation and evolution
Dynamics probes	• Dense environment core collapse evaporation collisions merging–surviving segregation • Test particle of the galactic gravitational field
Containers of peculiar objects	• X sources (strong–weak–diffuse) • Blue stragglers • Binaries • Planetary nebulae • White dwarfs • Cataclysmic variables • Millisecond pulsars • Neutron stars

Labels indicate the main branches of the diagram. The modern STELLAR EVOLUTION theory is able to predict quite precisely the physical processes undergone by stars which evolve along the CMD, and the whole evolutionary path of a Population II star is nicely described from the early to the final stages. Each specific evolutionary phase is labeled in figure 3(*b*), along with the corresponding basic nuclear burnings. A brief description of these main evolutionary stages is presented below.

The evolutionary tracks drawn in the CMD by GC stars of given initial mass (below $1M_{\odot}$, actually $\sim 0.8M_{\odot}$) and chemical composition closely resemble the observed main ridge lines shown in figure 3(*b*). However, there is a fundamental difference. Each point of the evolutionary track is the locus reached by the same star at different ages during its evolution; conversely, each point on the observed CMD does indeed correspond to the locus of stars with same age and chemical composition, but different masses.

Given a collection of stars having the same chemical composition but different masses, each star will evolve along its evolutionary track, at its own evolutionary rate (depending on the star's mass). It will thus be possible to define loci of constant time along the various evolutionary tracks, which will yield constant-age sequences. These sequences are generally referred to as 'isochrones'. The isochrones thus represent the loci of stars with the same age and chemical composition, but with different masses. Since members of a GC can be thought of as being born from the same cloud at the same time, but with different masses, the comparison of theoretical isochrones (transformed into the observational plane) with the observed CMD is the key procedure to obtain information on the evolutionary status of GC stars.

The main sequence and the turn-off

After a free-fall phase, the gravitational contraction leads the proto-star to increase its central temperature until the ignition of the first nuclear reactions in the stellar core takes place. The nuclear energy production slows down and eventually stops the contraction. The star reaches a nice equilibrium, balancing energy production and transfer and global brightness, and enters into the main sequence (MS). This equilibrium holds on the MS for more than 70% of the total stellar lifetime ($\sim 10^{10}$ yr), until the nuclear fuel (hydrogen, i.e. 1 proton) in the very central core is completely burnt out. It should be recalled that there is a theoretical lower limit to the stellar mass ($\sim 0.08M_{\odot}$) below which H burning does not take place in the stellar core. Stars smaller than this threshold are usually called very low-mass (VLM) stars and brown dwarfs (BDs), and for masses far below this limit ($\sim 0.001M_{\odot}$) one has the giant planets.

Stars spend most of their lifetime quietly burning hydrogen in their core via nuclear fusion. Our Sun is still on the MS. As hydrogen gradually runs out in the center, a nucleus of helium (produced by the fusion of four protons) grows up, and the star core begins to contract owing to gravity, while the outer layers progressively expand and cool down. The star leaves the MS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

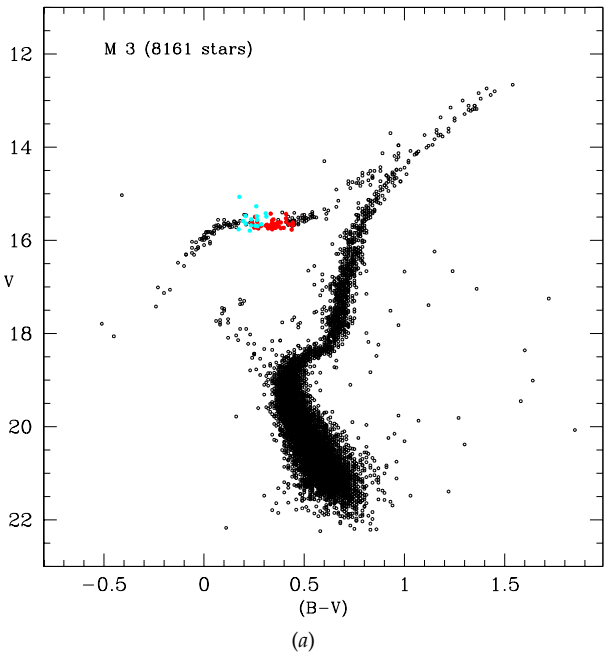


Figure 3. (*a*) Color–magnitude diagram of the globular cluster M3 and (*b*) schematic color–magnitude diagram of a typical globular cluster.

The central H exhaustion corresponds quite precisely to the bluest and hottest point on the MS, usually called the turn-off (TO) point. The precise location of the TO depends on the stellar initial mass, and theoretical models show that both luminosity and temperature of the TO increase with increasing stellar mass. Since, in turn, the initial mass depends on the age (stars burn out central H more and more slowly with decreasing mass), the TO is actually the 'stellar clock'. Turning to the isochrones, since less massive stars are still in the MS phase, while more massive ones are in more advanced evolutionary stages, the upper MS becomes more and more depleted with increasing age, while the TO becomes fainter and cooler. A straightforward relation can thus be derived between luminosity of the TO point and age. This relation provides a unique, powerful tool for determining the age of the GCs.

It should be pointed out that (a) the engine of the 'stellar clock' relies on theoretical models and (b) any method for dating stellar populations is finally founded on this 'stellar clock'. Were our understanding and running of this 'clock' incorrect, our description of the universe, as derived from the stars, would be in error. It should also be recalled that recent model calculations based on improved equation of state and radiative opacity, as well as the inclusion of element diffusion, have significant effects on both the luminosity of the TO, which is increased, and the MS lifetime, which is decreased with respect to previous 'classic' models.

Finally, the sharp MS TO observed in most clusters indicates that stars within individual GCs all formed roughly at the same time. The narrowness of the MS indicates that stars in a given GC all have very similar chemical composition, and also it constrains the fraction of binary stars within GCs. Unresolved binaries are in fact expected to produce a population just above the MS since the combined luminosity of the two stars exceeds that of a single star at the same color (see BINARY STARS IN GLOBULAR CLUSTERS).

The red giant branch

When the central fuel is exhausted, hydrogen starts burning in a thick shell which surrounds the growing helium core. A complex balancing between energy production and transfer, substantially driven by opacity, takes place. Core contraction and heating is accompanied by a progressive expansion of the outer envelope. The star cools down and brightens, owing to the larger and larger increase of the total radius, climbing along the red giant branch (RGB).

During this phase the deepening of the surface convection may reach the internal He-enriched zone and bring some extra helium to the stellar surface. All canonical computations confirm that, because of convective mixing phenomena (dredge-up) during the RGB phase, the surface He abundance exceeds the original He abundance. This increase of the He abundance may have important consequences in the subsequent evolutionary stages.

The comparison of the giant branches of different clusters has revealed that GCs of higher metallicity exhibit giant branches that are shallower and redder than those of low-metallicity clusters. The detailed astrophysics of the stars on the RGB is very complex, and the exact location of the RGB is dependent on details of the convective

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

processes within such stars commonly treated using the so-called mixing length (l) theory.

At the tip of the RGB the temperature in the He-rich core has grown up enough (10^8 K) to ignite the central He burning, via the 3α reaction ($1\alpha = 1$ He nucleus $= 2$ protons + 2 neutrons). Contrary to most of the other evolutionary stages, this ignition occurs on a very short time-scale (a few years or even less), and it is thus often referred to as the 'He flash'.

The horizontal branch

After the He flash, the star quickly rearranges its structure and starts a quiescent He core burning phase, which will last for about 10^8 yr or so. In the CMD, the star jumps on the horizontal branch (HB) where the helium in the core is converted into carbon (C) and oxygen (O). Slightly afterwards, H burning re-ignites in a shell immediately outside the He core. The relative balancing between central He burning and H shell burning has important effects on both HB lifetime and star behavior.

The precise location in luminosity and temperature of a star which arrives on the so-called zero-age HB (ZAHB) is driven by (at least) three parameters: the chemical composition of the envelope, the size of the He core mass at the He flash and the mass of the envelope (which was generally affected by mass loss during the previous RGB phase).

At given metal abundance—the so-called 'first parameter'—and helium content, the ZAHB becomes bluer and bluer (i.e. hotter and hotter) with decreasing stellar envelope mass. In turn, the position taken on the ZAHB by a given GC star will depend on the amount of mass lost during the RGB phase; the larger the mass lost, the bluer the ZAHB location.

Given the absence of quantitative observational estimates and with the persistent lack of any reliable theory about mechanisms driving the mass loss, it is difficult to evaluate the amount of envelope actually lost by stars evolving along the RGB. To reproduce the observed HB morphologies (i.e. the color distribution of HB stars), it has become common procedure (a) to introduce a free parameter, which accounts for the whole mass loss process, and (b) to adopt an arbitrary dispersion in the total mass lost by individual stars. In fact, if no mass loss (or the same amount of mass loss) is assumed for stars of equal initial mass, the whole HB would collapse to a single clump, contrary to any observational evidence. Indeed, in order to reconcile theoretical He burning models with the color distribution of their observational counterparts, i.e. the HB stars in GCs, it seems conceivable that during the RGB phase stars lose up to $0.2 M_{\odot}$ of their envelopes. However, mass loss affects only the very external layers of an RGB star, without altering the physical conditions in the core, where nuclear energy generation takes place.

Theoretical models show that both the ZAHB luminosity and the temperature location of the star on the ZAHB increase with increasing helium abundance Y, while both quantities decrease with increasing metal content Z. On the other hand, since with increasing age the initial mass decreases, at fixed total mass loss the HB distribution will become bluer and bluer with aging GCs. This property has a very deep impact on the description of formation and evolution of the Galaxy as in many studies the HB morphology is actually adopted as a 'secondary clock' (see below) for dating the GCs.

HB stars of appropriate mass may fall inside the so-called 'instability strip', also known as the 'RR Lyrae gap'. The envelopes of stars living inside or simply passing through this narrow region of the CMD (see figure 3(*b*)) become unstable for radial pulsation. The HB star thus becomes a variable of RR Lyrae type, so named from the historical prototype of this class of variables. The frequency of such stars in GCs is so high that the presence of RR Lyrae variables within the 'instability strip' of the HB is a well-defining feature of the population II stars. According to the periods of their RR Lyraes, Galactic globulars can be separated into two distinct subsets, the two Oosterhoff groups. It has been shown that the Oosterhoff separation, often referred to as the 'Oosterhoff effect', reflects a difference in the metal abundance of the host clusters, with variables in metal-poor clusters having longer mean periods than those in the metal-rich ones.

Since the HB is a bright and 'horizontal' feature of the CMD (hence, all its members have a very similar absolute magnitude), and since HB stars spend most of their HB lifetime at luminosities very close (0.1–0.2 mag) to their ZAHB luminosities, they are 'standard candles' to measure distances. RR Lyrae variables, in particular, are among the most used 'classic' standard candles of population II stellar systems. (See also HORIZONTAL-BRANCH STARS.)

The asymptotic giant branch and the white dwarfs

Eventually, when helium also vanishes in the central regions of the He core, the core contracts and He burning starts in a shell around it. The stellar structure, characterized at this stage by a carbon–oxygen core surrounded by a first He burning shell and by a second, more external, H burning shell, evolves upward in the CMD, asymptotically merging into the RGB during the asymptotic giant branch (AGB) phase.

Most AGB stars exhibit significant mass loss in the form of stellar winds, exactly in the same vein as shown while evolving along the brightest part of the RGB. Other peculiarities include He shell flashes, believed to be a consequence of the He burning shell being spatially thin, which can cause the star to migrate briefly to the instability strip and to mix burnt material into the outer layers. This results in significant peculiarities in the abundance of some chemical elements, whose detection and measurement offer important tests of the global evolutionary theory, including mass loss.

The normal AGB stage culminates, after the onset of thermal pulses and a phase of rapid mass loss or 'superwind', with the star blowing off its outer layers until all that is left is a central hot WHITE DWARF (WD) in the middle of a PLANETARY NEBULA.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Eventually, the white dwarf becomes faint and cools down, ending its evolution along the WD cooling line, a sequence almost parallel but about 4 mag fainter than the MS.

Blue stragglers

Discovered first in M3 more than 40 yr ago by Allan Sandage, and until recently just peculiar objects detected in only one or two clusters, a large number of BLUE STRAGGLERS (BSs) are being discovered nowadays by the HST observations of the central regions of almost any GC. Although all stars above a certain mass and luminosity in a given cluster are expected to have evolved off the MS, yet the BSs seem to sit nodding on the bright extension of the MS (see figure 3(*b*)), as if they were loitering on their evolutionary path compared with the other members of the cluster. Since BSs always seem to be more centrally concentrated than all other visible stars in a given cluster, their masses are expected to be significantly higher than either RGB or normal MS stars, and this would explain their brighter location along the extension of the MS. Since they are still not evolved, although more massive than the TO stars, one has to explain their origin, which is commonly associated with binaries. BSs could originate from collisions of individual stars in the very crowded central regions of the cluster or could be the merger product of primordial binaries. Some BSs have actually been found to be members of eclipsing binary systems.

Chemical composition

Galactic GCs are old, metal-poor population II stellar objects with metal abundances usually in the range -0.7 dex $<$ [Fe/H] $<$ -2.5 dex (where [Fe/H] $= \log(N_{\rm Fe}/N_{\rm H}) - \log(N_{\rm Fe}/N_{\rm H})_{\odot}$ and N is the number of atoms of the given element; our Sun has thus [Fe/H] $=$ 0.0). Although clusters almost as metal rich as the Sun (and perhaps even more) have been recently discovered near the Galactic center, the bulk of the GC population has metallicities around [Fe/H] $= -1.3$. The average metallicity of the GC system generally correlates with the total mass of the host galaxy; the bigger the galaxy, the richer the average GC.

A commonly accepted basic property of the stars in a given GC is that they all have about same age and chemical composition (at least primordial).

While no indication has ever been put forward for any age variation among stars of a single cluster, there is now gathering observational evidence that small differences in chemical composition are perhaps present. It is still a matter of debate whether they are primordial or due to evolution.

Besides the use of spectroscopic observations at intermediate and high resolution, direct information on the chemical properties of a given cluster can also be derived from its CMD. In fact, the location and intrinsic width of the main branches in the CMD both depend on chemical abundance. In particular, as a general rule, an increase in the helium fraction leads to bluer colors, while an increase of the metal content leads to redder colors. Moreover, any significant abundance spread yields a spread-out of the stars along the various branches with respect to the mean ridge line, which is different from one branch to another and increases with increasing abundance variation.

The observed color dispersion of the MS in the best-studied clusters implies upper limits to star-to-star variations in helium abundance of 0.03 dex (if dispersion is attributed to He variations only, but differential reddening, metallicity spreads, blending and intrinsic binarity may also contribute). This figure compares with an overall helium abundance of about 0.23–0.25 (mass fraction).

Similar reasoning can be applied to the giant branch. Here, since elements with low ionization potential such as Fe, Si and Mg significantly affect the opacity in the stellar envelopes, star-to-star variations would result in a color dispersion of the giants. Observations show that, with the exception of ω Cen which clearly displays a wide giant branch, compatible with abundance variations up to a factor of 10 at least, the vast majority of the Galactic GCs have quite narrow giant branches and, in turn, small intrinsic variations in the abundance of iron-peak elements.

Spectroscopy of giants is potentially a more powerful technique to detect metallicity variations. However, such observations require high resolution and are highly time consuming on large telescopes. The still poor samples already available actually confirm the general indications of the CMDs. Only ω Cen (the biggest and most elliptical Galactic GC) and perhaps M22 show a dispersion in [Fe/H].

One reason why metal homogeneity within GCs is so remarkable is that galaxies, including dwarf spheroidals of integrated luminosities comparable with GCs, all show quite wide dispersions in chemical abundances, suggesting that star formation probably occurred in a different way and that self-enrichment was at work much more efficiently.

Despite the homegeneity in [Fe/H] within stars of individual clusters, there are significant star-to-star variations in the abundances of other heavy elements. The best famous are the CN variations, known among GC stars for many years. The central question is whether these variations are primordial star-to-star differences built in when GCs formed or whether they result from stellar evolutionary processes.

Stellar evolution is expected to produce some abundance variations among stars at different evolutionary phases. If the convective envelope of an RGB star reaches sufficiently deep inside the star, it mixes material which underwent CNO processing up to the surface of the giant. This theoretical prediction is now supported by many observational results, although observational evidence would require more mixing than currently predicted by theory.

Other elements also exhibit abundance variations in the GCs. The [O/Fe] ratio in cluster giants sometimes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

shows considerable depletion compared with halo field stars, which are enhanced by a factor of several relative to the solar abundance. However, it is still not clear whether there is a real systematic ehnancement of [O/Fe] in GCs with respect to the solar fraction, and, if any, how it varies with varying total metal abundance. At present, metal-poor clusters with [Fe/H] $= -2.0$ dex seem to have approximately [O/Fe] $= +0.3$ dex, while this figure perhaps decreases to about 0 at [Fe/H] $= -0.7$ dex.

Finally, it should be mentioned that Al, Na and Mg show variations which seem to be hardly compatible with just evolutionary processes. In particular, Al and Na tend to be overabundant in GCs, show wide variations and are correlated with each other and anti-correlated with [O/Fe]. If these Al and Na variations are primordial, their anticorrelation with [O/Fe] suggests that CNO variations are also primordial, and this may have a deep impact on the description and understanding of star and cluster formation.

The horizontal branch morphology and the 'second parameter' problem

The color (temperature) distribution of the cluster stars along the HB is often referred to as the 'HB morphology' of the cluster. The number of stars within (V) and on both red (R) and blue (B) sides of the 'RR Lyrae gap' describes the overall HB morphology of a given cluster through the parameter

$$C = \frac{B - R}{B + V + R}.$$

Observations show that Galactic GCs exhibit a fairly broad variety of HB morphologies, with the C parameter ranging from -1.0, for clusters where only the red portion of the HB is populated, to $+1.0$, for clusters with stars only to the blue of the instability strip.

Since the color (temperature) of HB stars depends primarily on the stellar metal content—HB stars in metal-rich clusters are redder than those in metal-poor clusters as a result of the higher opacity in their envelopes and of higher initial masses (at fixed age)—metallicity is commonly considered the 'first parameter' driving the HB morphology.

However, metallicity variations alone cannot account for the extension in effective temperature and for the differences observed between HBs. In fact, were the HB morphology strictly driven just by metallicity, both the high chemical homegeneity among cluster members and the fine tuning of the TO masses would lead to a highly peaked clump of HB stars. The common evidence that most of the observed HBs display a wide coverage in color (only very metal-rich clusters show in fact a red stubby HB) suggests that mass loss affects in rather different ways GC stars of identical initial mass, no matter whatever mechanism(s) drives it.

On the other hand, although the HB morphology–metallicity correlation accounts for a large part of the observed HBs and most clusters exhibit a 'blueing' of their HBs with decreasing metal abundance, several other clusters do not obey this general rule. This is the essence of the long-standing historical 'second parameter problem': another parameter besides metallicity must be at work on the HB of the GCs.

Age, helium abundance and metal abundance ratios are the 'classical' second parameter candidates most often invoked to explain the observed HB morphologies.

Age has long been claimed to be the second parameter, but with no unanimous consensus among the astronomical community as not all second parameter clusters seem to be explicable in terms of age variations. Theoretical calculation shows that the average mass of a stars evolving onto the HB decreases with increasing cluster age. Since the He core mass is roughly constant for ages older than $\sim$10 Gyr, the mass of the envelope decreases with age. Hence, at fixed metallicity, older cluster ages would imply less massive HB stars, with increasingly thin evelopes and, in turn, bluer HBs. If age is the second parameter, the HB morphology could be used as a 'secondary clock', besides the TO luminosity, to test the existence of possible age gradients with varying distance from the Galactic center and the metallicity. This approach has been widely used in recent years to show that the GC system in the Milky Way is perhaps formed by different subgroups: (a) the halo GCs, possibly further subdivided into young and old, metal poor and spherically distributed throughout the wide halo; (b) the disk GCs, of intermediate metallicity, somehow kinematically related to the disk; (c) the bulge GCs, in general fairly metal rich (almost solar) and preferentially located in the central regions of the Galaxy; (d) the (possibly) captured GCs, originally members of disrupted or tidally interacting satellites of the Milky Way.

Variations in the helium content, with increasing helium abundance producing bluer HBs, could possibly explain some of the observed HB morphologies. On the other hand, also a variation in the CNO abundance (at fixed [Fe/H] content) would affect the HB morphology, with increasing CNO abundance yielding redder HBs, since elements of the CNO group directly control the opacity and energy generation within HB stars.

Also many other physical phenomena could affect the HB star properties. For instance, the stellar core rotation, if increased, would probably lead the star (a) to have a higher He core mass, consequently (b) to climb up brighter portions of the RGB and (c) if mass loss increases with increasing luminosity, as it appears to, to suffer in turn a larger mass loss, yielding bluer HBs.

Environmental conditions have also been shown to have some impact on the evolution of GC stellar populations. In particular, there are observational indications suggesting that higher cluster densities and concentrations would favor bluer HBs since stellar mass loss could be enhanced, because of the increased stellar interactions.

In conclusion, there is growing observational evidence that just a single 'second parameter' may be insufficient to explain the observed HBs and that a comprehensive understanding of the full range of the observed HB

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

morphologies may require a 'third parameter', at least, although age is perhaps a dominant parameter in most GCs.

In this respect, a distinction could be made between a 'global' second parameter ruling the HB morphology and second parameters most likely determining the HB morphology of individual clusters. The former deals with the Milky Way globulars as a system and aims at describing the 'global' properties of the HB morphology compared with the other clusters of similar metallicity. In particular, it has been found that HB morphology correlates with Galactocentric distance, suggesting that the intrinsic quantity or phenomenon which causes the anomalous HB star distribution at that metallicity is also varying with varying cluster location within the Galaxy. Several authors have concluded that, of the various second parameter candidates, only age differences can reproduce the global second parameter effect and, at the same time, successfully reproduce the observed properties of RR Lyraes and the MS TOs. On the other hand, most observed HBs present special features (gaps, clumps, distortions) which naturally require the existence of 'non-global' effects, which relevance could also vary from star to star in a cluster.

In conclusion, any choice is uncertain at this stage, and the whole issue of the 'second parameter' effect is still matter of hot discussion and tireless study.

Absolute and relative ages

One of the most frequently asked questions in astronomy is, how old is the universe we live in? Since GCs (the most metal-poor ones, in particular) were formed during the very early times of the Galaxy, and represent the oldest stellar systems for which age can be estimated fairly accurately, their absolute ages set a very stringent lower limit on the universe's age itself, since the universe cannot be younger than the oldest objects it contains.

The observational parameters required to estimate the absolute age of a given GC are the apparent magnitude and color of the TO, along with the distance to the cluster, its helium and metal content and the interstellar reddening along the line of sight. This allows the calculation of the absolute magnitude of the MS TO (the 'observational stellar clock'). Once the absolute magnitude of the MS TO has been determined, ages can be estimated by comparison with appropriate model isochrones (the 'theoretical clock') properly transformed into the observational plane.

The primary source of uncertainty in the absolute ages derived from the luminosity of the MS TO is our poor knowledge about distances to clusters, although uncertainties in quantities such as metallicity, reddening and elemental abundance ratios must also be taken into account. It should be remembered that errors of 0.07 mag in the absolute luminosity of the MS TO or 0.01 mag in its color bring a corresponding error in the derived age of about 1 billion years.

Several methods exist to measure distances to GCs. Until recently, all of them were uncertain at about a 20% level. The effects of model uncertainties on the derived age estimates are more difficult to assess as they are dependent (with different weights) on parameters and calibrations which are not directly observable (reaction rates, opacities, treatments of convection and mixing, basic input physics, etc).

A recent attempt to quantitatively evaluate the global uncertainty of both observations and stellar models concluded that the age of the oldest clusters lies in the range 11–21 Gyr (Chaboyer *et al* 1996). These figures do not allow us to discriminate among different cosmological models.

Nearby subdwarfs, whose distances can be determined to high accuracy through trigonometric parallax, provide the best standard candles to estimate GC distances, especially after the results yielded by the ESA's astrometrical satellite HIPPARCOS. The magnitude offset between the observed cluster MS and the absolute-calibrated MS of the local subdwarfs (assumed to be 'bona fide' representatives of GC subdwarfs of the same metal abundance) eventually gives the distance to the cluster.

Before HIPPARCOS, the available sample of local subdwarfs with known metallicity and accurate PARALLAXES was extremely poor. In early 1997, HIPPARCOS substantially increased the observational database, bringing the number of subdwarfs with very accurate parallaxes to about 900 stars, 30 of which have metallicities comparable with those of the GCs ([Fe/H] ≤ -1.0). The enlarged sample of subdwarfs and the use of new metal abundances in the same metallicity scale as used for GC stars have led to an almost halving of the uncertainty in derived cluster distances and, in turn, ages.

The discussion on the ages of the galactic GCs has been stirred up because of the global modifications of the astronomical distance scale implied by the HIPPARCOS results, which suggest that many Galactic objects, including the GCs, may be at distances about 10% larger from us than previously thought. The intrinsic luminosity of their TO stars should thus be about 20% brighter and they should also be roughly 15% younger, given the quoted relationship between TO luminosity and age.

The latest estimates yield therefore an age of 12–14 billion years for most of the GCs, and 13–14 billion years is perhaps the age of the most metal-poor and (presumably) oldest GC: M92.

An independent check of the distances derived from the subdwarf technique is absolutely necessary. A promising alternative is offered by the use of the WD cooling sequence. This technique relies on the matching of local WDs to WDs observed in GCs, as a matter of fact so exploiting, to a much fainter level, the same technique used with the subdwarfs. With the refurbished HST, it has become technically feasible to observe WDs in GCs. The first result from this technique for the GC NGC 6752 was obtained by Renzini *et al* in 1996. They find a distance consistent with other estimates, and derived ages of around 15 Gyr.

Although not totally incompatible, absolute ages derived for GCs in the pre-HIPPARCOS era were poorly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

consistent with most cosmological age determinations, such as those obtained through measurements of the Hubble constant. However, if the latest estimates of the GC ages obtained via HIPPARCOS parallaxes for the local subdwarfs are correct, the problems of the compatibility between the so-called 'stellar route' to cosmology and alternative routes are reduced, and would possibly disappear.

Relative ages of GCs are more reliable than absolute ages, primarily because determinations of age differences are less sensitive to stellar models and, second, and most important, because they do not depend on the assumed cluster distance. Their use allows one to address outstanding questions concerning the process of formation of our Galaxy. In particular, if we could define a reliable age–metallicity relation, as well as determine a variation of age with Galactocentric distance, we would be able to make a choice between the scenario suggested first by Eggen, Lynden-Bell and Sandage (1962), who think that the Galaxy was formed in a rapid, monolithic collapse (in a few 10^8 yr), and that proposed by Searle and Zinn (1978), who rather believe that the Galaxy formed much more gradually through the accretion of independent fragments (with masses ranging from $10^7 M_\odot$ up to $10^9 M_\odot$), over a few billion years or so. The detection of age differences among the Milky Way's GCs would in fact provide stringent constraints on the time scale and mechanism of formation of the Galactic halo.

Several methods have been employed so far to derive GC relative ages.

The most widely used technique exploits the difference in luminosity between MS TO and HB. Since the core mass of GC stars on the HB is roughly constant, the HB luminosity is almost independent of cluster age. On the contrary, the MS TO luminosity decreases with increasing the age, so the magnitude difference between the MS TO and the HB provides an estimate of the age of the GCs.

This method is elegant and simple in principle; however, it does suffer from several problems. The basic observational limitation is that while it works nicely for clusters having HBs well populated on both sides of the instability strip, it can hardly be applied when only red stubby HBs or just blue HB tails are present in the CMD. The latter case is particularly worrying since, if age is the second parameter, metal-poor clusters with blue HBs could be the oldest ones. A second basic problem is the still poor accuracy in our knowledge of the absolute magnitude of RR Lyrae stars (and thus of the HB (to yield absolute ages)) and of its dependence on metallicity.

The magnitude versus metallicity relation for RR Lyraes is usually described as: $M_V(\mathrm{RR}) = a[\mathrm{Fe/H}] + b$. The values currently obtained for a cover the range 0.15–0.35 mag dex^{-1}. This uncertainty is large enough to prevent any reliable determination of the GC age distribution since an error in the absolute magnitude of $\sim$0.20 mag implies a 2.5 Gyr error in the age. Furthermore, an estimate of the absolute age, with an error less than 1 Gyr, based on the RR Lyrae variables as absolute standard candles, requires knowledge of b to better than 0.07 mag.

The detailed comparison of the CMDs offers an alternative technique to deriving relative ages for GCs of similar metallicity. In fact, the relative distance of the clusters can be measured by simply superimposing the ridge lines of the main branches in the observed CMDs, after taking into account any differential interstellar reddening. If clusters of different metallicity are compared, further corrections (often uncertain) must be drawn from stellar atmosphere models. Once the best matching is achieved, however, the difference in the magnitude of the MS TO between the two compared clusters yields a straightforward measure of the age difference, if any.

Recent analyses of the relative ages of the best-observed Galactic GCs seem to indicate in general the existence of a small (less than 1 Gyr) spread. However, there is now also compelling evidence for larger age differences between some globulars. Probably the best-established cases are Ruprecht 106, Palomar 12, Terzan 7 and Arp 2. They (especially the first two) appear to be around 3 Gyr younger than other GCs of similar metallicity. It is unknown whether they are truly exceptions or whether other 'young' clusters lie still undetected. They could perhaps originally belong to a satellite galaxy tidally disrupted and captured by the Milky Way, but further confirmation is necessary.

In conclusion, the current status of our knowledge about the age of the Galactic GCs is consistent with an age distribution in which most Milky Way globulars are quite uniformally old (with ages of about 13 Gyr), and the oldest ones as old as 14 Gyr, but also with a few globulars significantly younger (by about 2–3 Gyr).

Bibliography

Chaboyer B, Demarque P, Kerman P J and Krauss M 1996 *Science* **271** 957

Eggen O J, Lynden-Bell D and Sandage A R 1962 *Astrophys. J.* **136** 748

Renzini A *et al* 1996 *Astrophys. J.* **465** L23

Searle L and Zinn R J 1978 *Astrophys. J.* **225** 357

Flavio Fusi-Pecci and Gisella Clementini

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globule

A small, round dark nebula thought to represent a stage in the collapse of a concentration in a molecular cloud toward a protostar. They range in mass from about 1 to 1000 or more solar masses, and in size from about 10 000 AU to 3 light-years, and are visible in silhouette against bright nebulae such as the Rosette Nebula, or against bright Milky Way starfields. Bart Bok was the first to suggest that they might be the precursors of protostars; they are sometimes called *Bok globules*. What appear to be protostars have been detected within some globules by orbiting infrared telescopes.

See also: Cometary globule.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Goddard Space Flight Center

The Goddard Space Flight Center (GSFC), in Greenbelt, MD, named for Dr Robert H Goddard, a pioneer in rocket research, was established in 1959. Since that time, GSFC has played a major role in space and Earth science.

Goddard's fundamental mission is to expand knowledge of the Earth and its environment, the solar system and the universe through observations from space. The center is committed to excellence in scientific research and investigation, the development of space systems and the advancement of essential technologies.

Approximately 11 740 people work at the Goddard Space Flight Center at all of its sites. Goddard has personnel and facilities capable of creating, building, testing, launching and operating various satellite projects in support of Earth science, space science and advanced technology programs.

Goddard's Greenbelt campus includes the Earth Observing System Data and Information System (EOSDIS) facility built in 1995, and the Earth Systems Science Building built in 1998.

The EOSDIS facility serves as a key data-retrieval node in the Earth Observing System communications system, as well as a distribution center for Earth data from numerous spacecraft and instrument sources, such as the Total Ozone Mapping Spectrometer and Tropical Rainfall Measuring Mission.

The Earth Systems Science Building houses the laboratories of Goddard's Earth Sciences Directorate: the Laboratory for Atmospheres, Laboratory for Terrestrial Physics and Laboratory for Hydrospheric Processes.

To support NASA's Space Science Enterprise, Goddard leads the mission of space-based physics and astronomy to create opportunities for conducting research through a broad variety of flight opportunities. Goddard promotes the development of advanced technology designed to enhance scientific capabilities at an affordable cost.

Goddard provides agency leadership to advance next generation spacecraft, sensor and instrument technology. Goddard scientists and engineers have gained significant expertise in areas such as optics development, cryogenics, microelectronics, x-ray astronomy, Earth observation and information systems.

For further information see
http://www.gsfc.nasa.gov.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Goddard, Robert Hutchings (1882–1945)

American rocket pioneer, whose achievements in rocketry, sometimes in the face of official disapproval or ridicule, form a long list. He was the first to prove, by static test, that a rocket will indeed work in a vacuum—that it needs no air to push against. He received a US patent on his idea for a multistage rocket (1914). In the document (1920) which laid out his life's work, entitled 'A method of reaching extreme altitudes', he outlined methods of raising weather-recording instruments higher than balloons and developed the mathematical theories of rocket propulsion to reach high altitudes. He foresaw the possibility of a rocket reaching the Moon and suggested that it could explode flash powder there to signify its arrival. This was the most dramatic, eye-catching part of the report, and was widely (and sceptically) reported in both the USA and Germany, where it coincided with the surge of interest in rockets there. He developed and fired the world's first liquid fuel rocket (gasoline/petrol and liquid-oxygen fuelled) in 1926, and was the first to launch a scientific payload (barometer and camera) in a rocket flight (1929). He made various technical advances in the use of pumps, gyros and other rocket mechanisms, as evidenced by the 200 patents in his name. NASA's Goddard Space Flight Center, located near Washington, DC, is named after him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gold, Thomas (1920–)

Astronomer born in Vienna, Austria, emigrated to Britain and then the USA. Noted for provocative, imaginative and sometimes wholly incorrect theories, he proposed the now discredited steady-state theory of cosmology (with BONDI and HOYLE), in which matter was continuously created to maintain the density of the Universe while it expanded. At Cornell University, he directed the Center for Radio Physics and Space Research, where he named the magnetosphere and studied pulsars, proposing that they were rotating magnetic neutron stars, as now thought to be correct. Gold caused some concern for the safety of US astronauts when he incorrectly theorized that the Apollo lunar lander would sink into meters of lunar dust. He is at the present time offering an unconventional view of terrestrial carbon fuel deposits (coal, oil), that they are not of biological origin but are cosmic carbon welling up from the Earth's central regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Goldberg, Leo (1918–87)

Astronomer, born in New York City of Polish origin. Became director of the Kitt Peak National Observatory in Tucson.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Goodricke, John (1764–86)

Born of English parents in the Netherlands, a deaf-mute, he discovered in 1782 that Algol, identified as a variable star by G MONTANARI in 1669, was an eclipsing binary and estimated its period. He also discovered the variability of Delta Cephei.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gould Belt

Most stars in the solar neighborhood younger than about 60 million years are located in a flattened structure of about 700 parsec (pc) in size, surrounding the Sun. This structure also contains much interstellar atomic and molecular gas, and is called Gould's Belt. It is inclined by nearly 20° relative to the general stellar population of the Galactic plane, which has an average age of several billion years. The properties of the Gould Belt can be used to trace the recent star formation history in our part of the Galaxy.

Discovery

Sir John Herschel noted in 1847 that the brightest stars in the sky are not arranged symmetrically about the Galactic plane. From observations made in South Africa (see HERSCHEL FAMILY) he concluded that the Galactic equator is crossed by a zone of bright stars which is delineated by the southern constellations Orion, Canis Major, Carina, Puppis, Vela, Crux (the Southern Cross), Centaurus, Lupus and Scorpius. The great circle which connects the stars ε Orionis (Alnilam) and α Crucis (Acrux) defines the axis of this zone of stars. Benjamin Gould pointed out in 1874 that the feature observed by Herschel is traced in the northern sky by the constellations of Taurus, Perseus, Cassiopeia, Cepheus, Cygnus and Lyra, although less conspicuously. Gould studied the distribution of these stars in detail and concluded that together they define a great circle with an inclination of 20° to the Galactic equator (see figure 1). This apparent distribution on the sky implies the presence of a flattened structure, with the Sun located in its interior, which is nowadays known as the Gould Belt.

Stars

The stellar spectral classification scheme which was developed shortly after the discovery of Gould's Belt showed that most visible stars in the Belt were of early spectral type (O and B type stars). When the principles of STELLAR EVOLUTION were established in the middle of the 20th century, it became clear that early-type stars are massive and young (less than a few tens of Myr). At the same time the Russian astronomer Ambartsumian recognized that groups of O and B stars, OB ASSOCIATIONS, are gravitationally unbound. These groups therefore expand and quickly dissolve into the Galactic disk star population, which led Ambartsumian to postulate that the observed associations must be young. Thus, both O and B stars in associations and the associations themselves were formed recently (∼10–100 Myr ago).

The realization that most O and B stars reside in OB associations immediately led to the conclusion that many of the nearby OB associations are connected with Gould's Belt. Figure 2 shows that the Scorpio–Centaurus–Lupus–Crux (Sco OB2), Orion (Ori OB1), Perseus (Per OB2), and Lacerta (Lac OB1) associations lie in the Gould Belt. These groups have typical ages of less than ∼15 Myr.

Many detailed photometric, radial velocity and astrometric studies of stars in the Solar neighborhood have been carried out in order to investigate the existence and characteristics of the Gould Belt. It was found that stars within 500–1000 pc of the Sun older than ∼60 Myr generally tend to be distributed symmetrically with respect to the Galactic plane, while ∼90% of the stars of spectral type O, B and A younger than 30 Myr, and ∼70% of early-type stars with ages between 30 and 60 Myr, define a distinct spatial and kinematical feature within the MILKY WAY GALAXY. The Gould Belt stars are distributed in a flattened elliptic structure with a semi-major axis of ∼360 pc and a semiminor axis of ∼210 pc, which is tilted by ∼20° with respect to the Galactic plane. The distances of the Gould Belt stars indicate that its three-dimensional structure is not a thin ring but rather resembles a 'disk with a hole'. The center of the Gould Belt is located ∼170 pc from the Sun in the direction $(\ell, b) \sim (130^\circ, -15^\circ)$, so that the Sun is located inside the Gould Belt. The ascending node of the ring lies in the direction of $\ell \sim 280^\circ$. The kinematics of the Gould Belt stars, which was recently much improved with the addition of astrometric data from the HIPPARCOS satellite, cannot be explained by differential galactic rotation alone but shows evidence for an additional expansion and rotation of the system as a whole.

Recently it was found that Gould's Belt is also associated with young low-mass stars. These stars are bright in X-RAYS, and ROSAT satellite observations made in 1998 show that they outline the same inclined flattened structure as observed in the early-type stars. Figure 3 shows the distribution of these low-mass stars on the sky; the Gould Belt as determined from the early-type stars is also indicated in this figure. The total stellar mass of the Gould Belt is estimated to be at least a few times 10^5 solar masses.

Interstellar matter

STAR FORMATION is an inefficient process; less than half of the available gas is generally converted into stars. This implies that recently formed stars should still be associated with copious amounts of interstellar matter. Indeed, many nearby molecular cloud complexes (e.g. the Great Rift stretching from Ophiuchus through Aquila to Cygnus, and the Orion and Perseus molecular clouds) are related to the recent formation of Gould's Belt (see also figure 2); Herschel already recognized the relation of the Orion Nebula to the Gould Belt. The molecular gas in Gould's Belt interacts with the recently formed massive stars: (1) these stars illuminate nearby gas and dust resulting in REFLECTION NEBULAE, and (2) COSMIC RAYS from supernovae produce high-energy GAMMA RAYS through the interaction with the gas. Measurements by the ESA COS-B satellite in the 1970s and the IRAS infrared satellite in the 1980s indeed revealed that the distribution of bright stars in Gould's Belt is correlated with high-energy gamma rays and nearby reflection nebulae, respectively. The latter correlation had already been noted in the 1920s by Hubble, who found, unaware of their true nature, that nearby

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

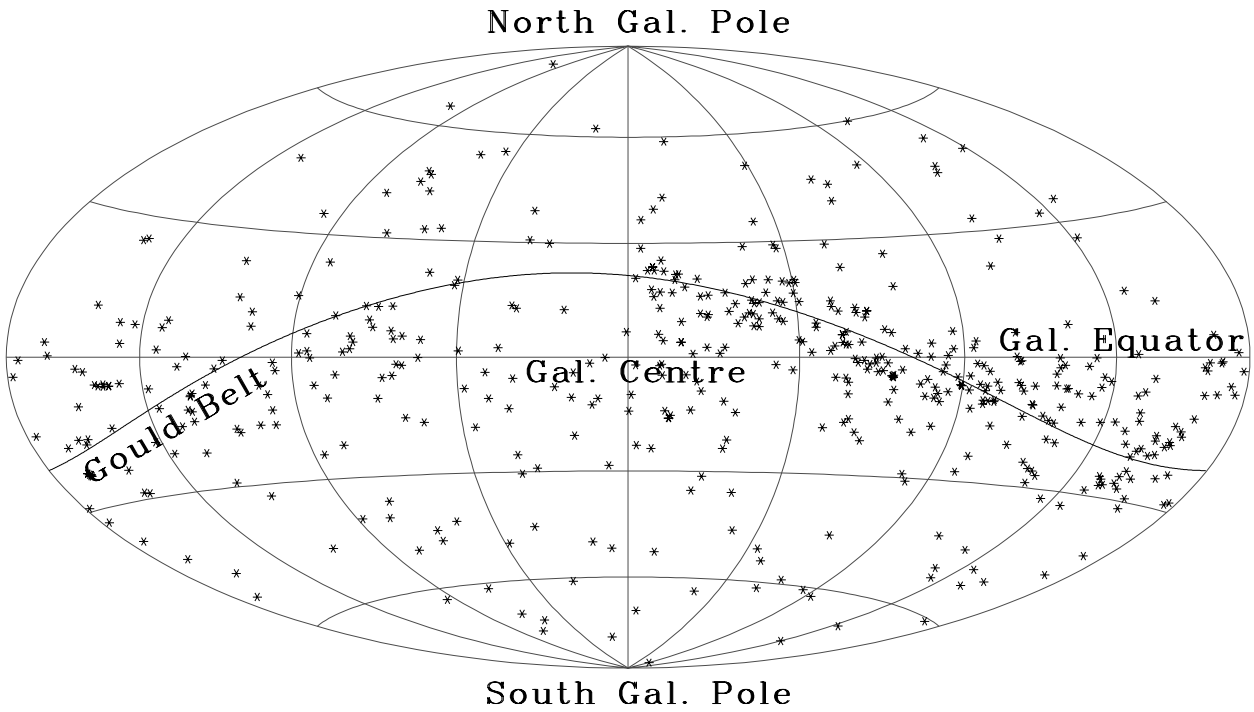


Figure 1. The distribution in GALACTIC COORDINATES of young stars (spectral type B) brighter than visual magnitude 5.25. These stars are visible to the naked eye. The curved solid line labeled 'Gould Belt' depicts a great circle inclined at 20° to the Galactic equator. The Belt is most conspicuous at its ascending node, which is in the southern celestial sky.

reflection nebulae (at the time simply called 'large diffuse nebulae') are distributed on the sky along the Gould Belt.

There is also ample evidence for atomic gas associated with the Gould Belt. Soon after the prediction of the hydrogen 21 cm line transition by Van de Hulst in 1944, and the first detection of interstellar neutral hydrogen (H I) in 1951 by Ewen and Purcell, Heeschen and Lilley observed the galactic center and anti-center regions. In the late 1920s Oort and B Lindblad had proposed a model which assumed that gas follows circular orbits around the Galactic center in a thin disk. Therefore, Heeschen and Lilley expected to find H I gas only at low Galactic latitudes ($|b| < 5^\circ$) and with radial velocities equal to zero km s^{-1} with respect to the local standard of rest. To their surprise, they found significant amounts of H I gas at $(\ell, b) \sim (0^\circ, 20^\circ)$ and $(\ell, b) \sim (180^\circ, -20^\circ)$ moving with radial velocities of $\sim +5$ km s^{-1}. This result provided the first evidence that H I gas is associated with the early-type stars in Gould's Belt (cf figure 1).

Firm evidence for the association of neutral gas with Gould's Belt was presented in 1967 by P Lindblad. Using the Dwingeloo Telescope he identified several distinct low-velocity-dispersion components in the low-latitude H I distribution in the direction towards the Galactic anticenter. Inspired by Blaauw's idea that the early-type stars in Gould's Belt form an expanding system, Lindblad showed that the velocity–longitude data for one of these local peculiar H I components (feature A) could roughly be modeled as a ballistically expanding gas ring (Lindblad's Ring) with the Sun located off-center in its interior. Later investigations expanded Lindblad's data set to include other longitude ranges and improved the quality of the data. These studies confirmed Lindblad's original hypothesis. The total mass of the interstellar matter associated with the Belt is estimated to be a few million solar masses.

Origin and evolution

In order to understand the stellar and gaseous content of Gould's Belt as one physical system all observational facts presented in the previous sections must be accounted for by one consistent formation and evolution scenario. It is thought that Gould's Belt is the result of a large-scale star formation event which started with the passage of the Carina SPIRAL ARM. This arm is currently at a distance of ∼4 kpc from the Sun along the Solar circle at $\ell \sim 280^\circ$ and passed a large gaseous condensation in the Solar neighborhood (a giant molecular cloud) roughly 60 Myr ago. This passage triggered star formation, resulting in the birth of the Cassiopeia–Taurus OB association and α Persei cluster (figure 4, panel 1). Both groups have similar ages and kinematics.

Based on the present-day stellar content of the Cas–Tau association Blaauw estimated that roughly 15 semi-simultaneous supernova explosions must have occurred shortly after its formation. The energy output of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

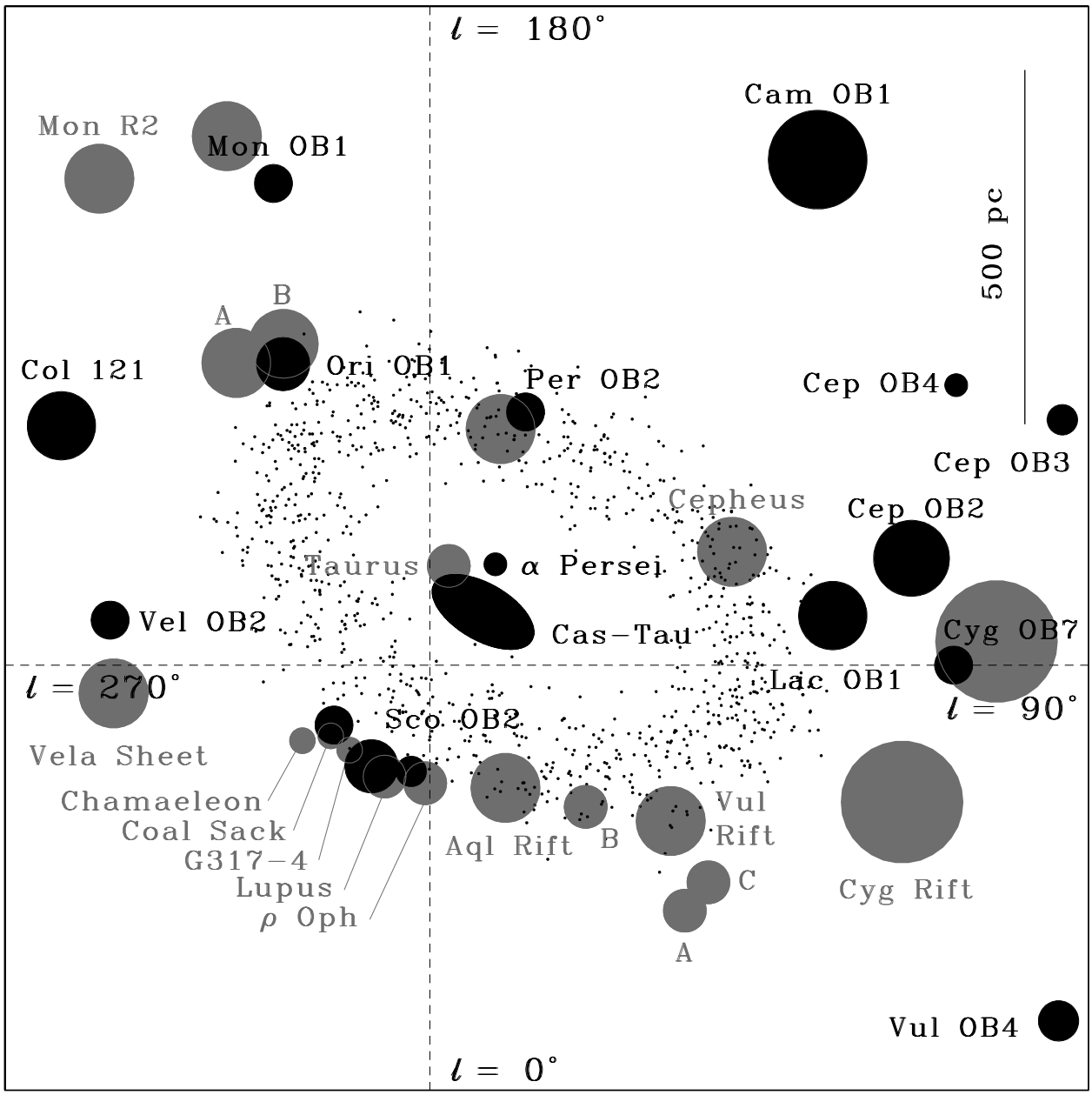


Figure 2. OB associations and molecular clouds near the Sun, projected onto the Galactic plane. The Sun is at the center of the broken lines which give the principal directions in Galactic longitude, ℓ. The size of the black-filled circles represents the projected dimension of the associations. The gray-filled circles represent molecular clouds; the circle radii are proportional to the cube roots of the cloud masses and are close to the clouds' actual radii. The distribution of small dots indicates a schematic and simple model of the gas distribution at low latitudes in Gould's Belt. Adapted from de Zeeuw *et al* (1999 *Astron. J.* **117** 354–99) and Dame *et al* (1987 *Astrophys. J.* **322** 706–20).

SUPERNOVAE, combined with the stellar winds of other massive stars, was sufficient to clear the association of its parental gas and create a rapidly expanding hydrogen shell (figure 4, panel 2). This initially spherical shell traveled through the interstellar medium while sweeping up and compressing gas and dust near the Galactic plane (Lindblad's Ring), leaving behind a cavity in the molecular cloud. The interaction between the shell and the interstellar medium caused the expansion velocity of the shell to decrease slowly. The effects of the Galactic tidal field gradually deformed the shell's circular shape into an elliptical one (figure 4, panel 3). In the direction perpendicular to the Galactic plane, the shell did not encounter much resistance, with the result that it broke

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

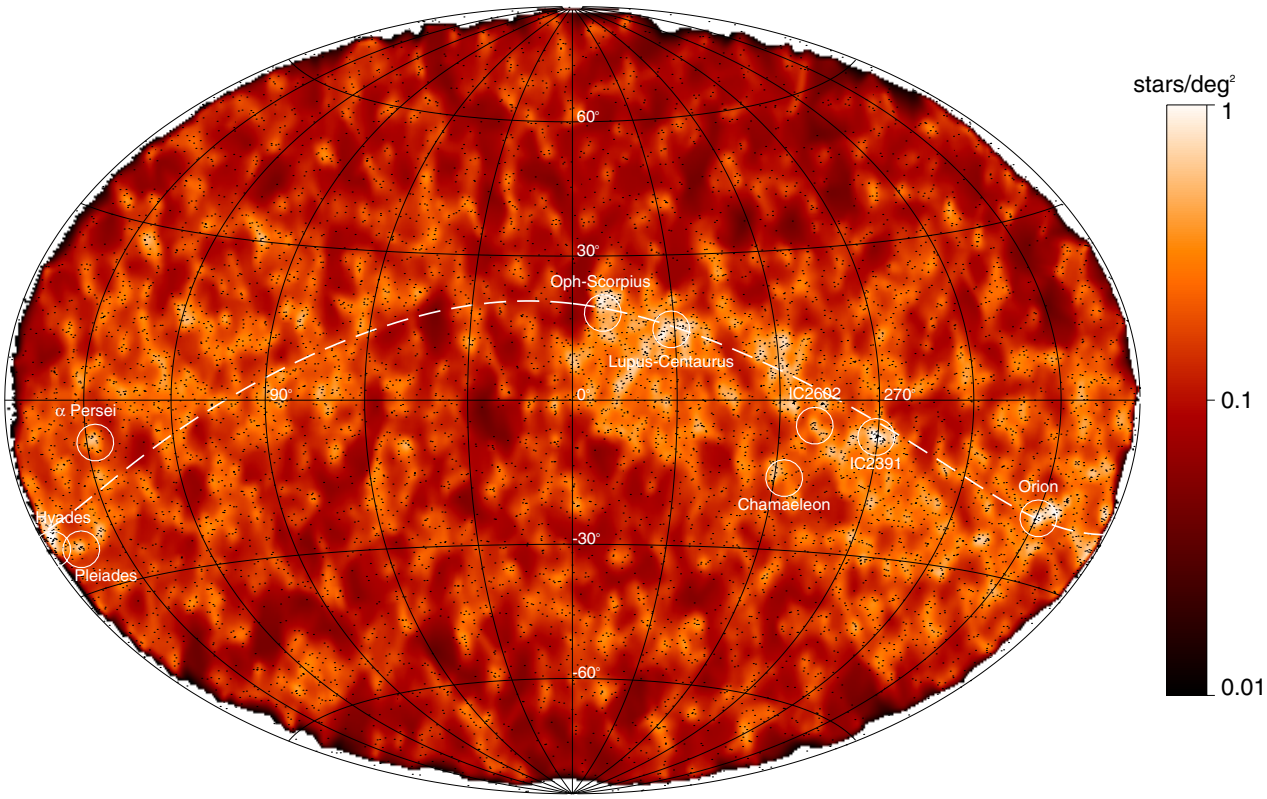


Figure 3. The all-sky distribution in Galactic coordinates of ∼8600 x-ray-bright low-mass young stars identified by the ROSAT satellite (small black dots). The gray scale represents the surface density of these stars on a logarithmic scale from 0.01 to 1 star per square degree (see gray-scale coding bar). The white broken line indicates the Gould Belt (see also figure 1). Circles denote some nearby young Galactic clusters and star forming regions. Adapted from Guillout *et al* (1998 *Astron. Astrophys.* **334** 540–4). **A color version of this figure is reproduced as Color Plate 19.**

open allowing the hot gas in its interior to blow out of the plane into the lower galactic halo through a so-called Galactic fountain.

After ∼45 Myr the physical conditions of the gas and dust (density and temperature) in the shell were suitable for the formation of stars. Low-mass stars formed throughout the entire shell (figure 4, panel 3) while STAR CLUSTERS and OB associations formed in dense clumps. These clumps are the result of fragmentation of the hydrogen shell. The new (second) generation of stars currently defines the Gould Belt (figure 4, panel 4, and figure 2). The observed motions of these stars still bear the signature of the expanding parental gas shell. Whereas the stars continued their expansion unobstructed since their birth, the gas shell continued to slow down owing to braking forces. Observations indeed show that the Gould Belt stars have radial velocities which are slightly larger than the +5 km s^{-1} gas velocities.

The most conspicuous stellar complexes in Gould's Belt are the OB associations Sco OB2, Ori OB1, Per OB2 and Lac OB1. These aggregates themselves have already induced the formation of a third generation of stars (e.g. ρ Ophiuchus, Orion Nebula Cluster, IC 348).

Open issues

The evolutionary history of Gould's Belt presented in the previous section is a simplified representation of reality and some complications remain that have not been addressed.

- Star formation in the expanding shell influences its density and velocity structure because of, e.g., local bubbles blown by the supernovae and stellar winds of the second generation of massive stars. Examples of such local bubbles are the Orion–Eridanus superbubble and the shells around Sco OB2, Col 121, Per OB2 and Lac OB1 (see, e.g., figure 5 and figure 3.15 in Pöppel (1997)).
- High-galactic-latitude observations show infalling gas near the Galactic Poles. This gas was originally blown out of the plane of the Galaxy (Galactic fountain) and is now falling back on a collision course with the Gould Belt itself. Such collisions can produce disturbances of the expanding shell, e.g. the North Celestial Pole Loop.
- The presence of several large dark cloud complexes, notably the Taurus molecular cloud and Pleiades cloud, inside Lindblad's Ring cannot be explained if the expanding shell encountered these clouds in the past.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

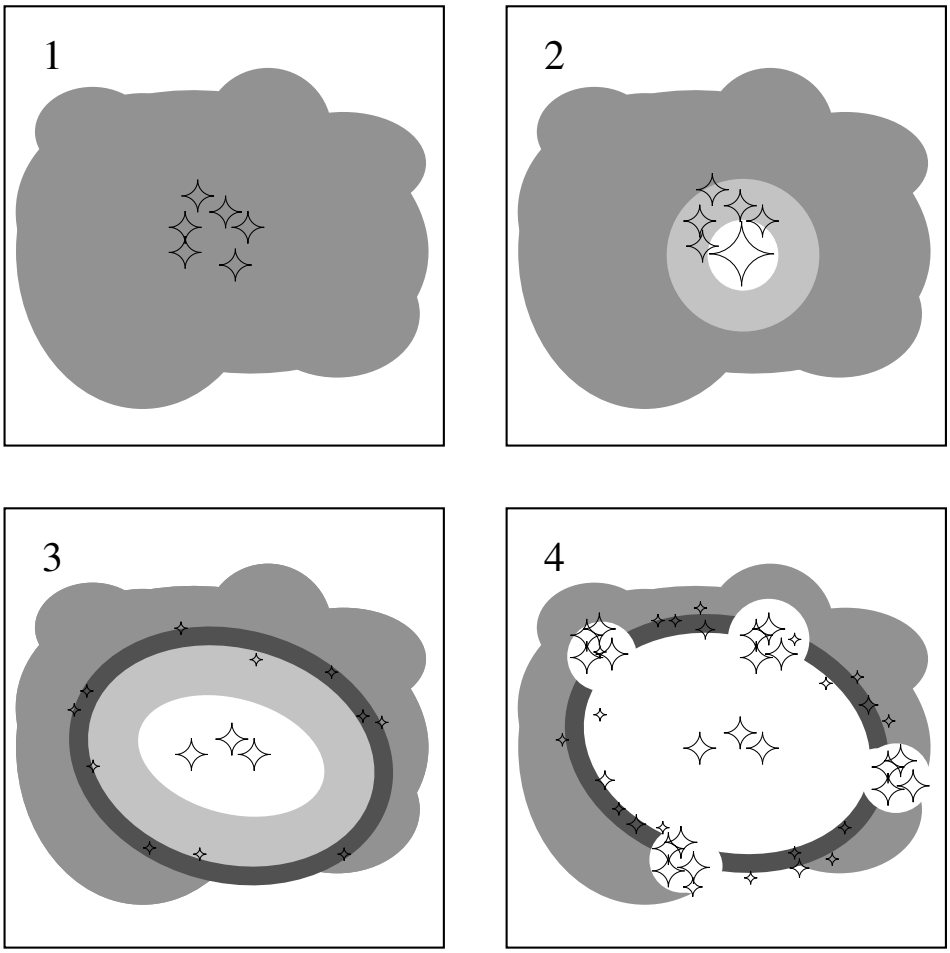


Figure 4. Schematic representation of the formation of Gould's Belt. Sixty Myr ago the Cassiopeia–Taurus (Cas–Tau) association formed inside a giant molecular cloud (panel 1). The stellar winds of the massive stars and the supernovae in Cas–Tau cleared the association from its primordial gas creating an expanding shell sweeping up material of the molecular cloud (panels 2 and 3). Star formation occurred throughout the entire shell and approximately 15 Myr ago several OB associations formed where the shell fragmented into dense condensations (panel 4). The expanding shell of gas and the newly formed stars are at present denoted as Gould's Belt.

These complexes may have been ejected inwards from the shell in the past.

- One of the most noticeable and remarkable properties of Gould's Belt is its 20° tilt with respect to the Galactic plane. There are three possible explanations for this tilt. (1) The tilt is the coincidental result of the fragmentation of the superbubble blown by Cas–Tau (this would be consistent with the fact that Gould's Belt is dominated by two associations Sco OB2 and Ori OB1). (2) The shell blown by Cas–Tau expanded into a pre-existing tilted gas layer (shingle). (3) The original burst of star formation was caused by the oblique impact of a HIGH-VELOCITY CLOUD. This scenario naturally explains the tilt of Gould's Belt.

Expectations for the future

The distribution and kinematics of the atomic and molecular gas associated with the Gould Belt have been studied by a number of authors, although it is clear that the details remain an open question; observations and analyses need to be pursued. By contrast, we still lack much information on the properties of the stars in the Belt, despite the major contributions by the Hipparcos satellite for the massive stars, and by ROSAT for low-mass stars. This situation can be considerably improved by the proposed GAIA satellite, which is currently under study by ESA. This satellite will provide accurate distances, space motions, and photometry throughout the Gould Belt region. At a limiting magnitude of ∼20, GAIA will even observe many low-mass stars. This will allow a full understanding of the Gould Belt and a reconstruction of the sequence of star formation in the Solar neighborhood

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

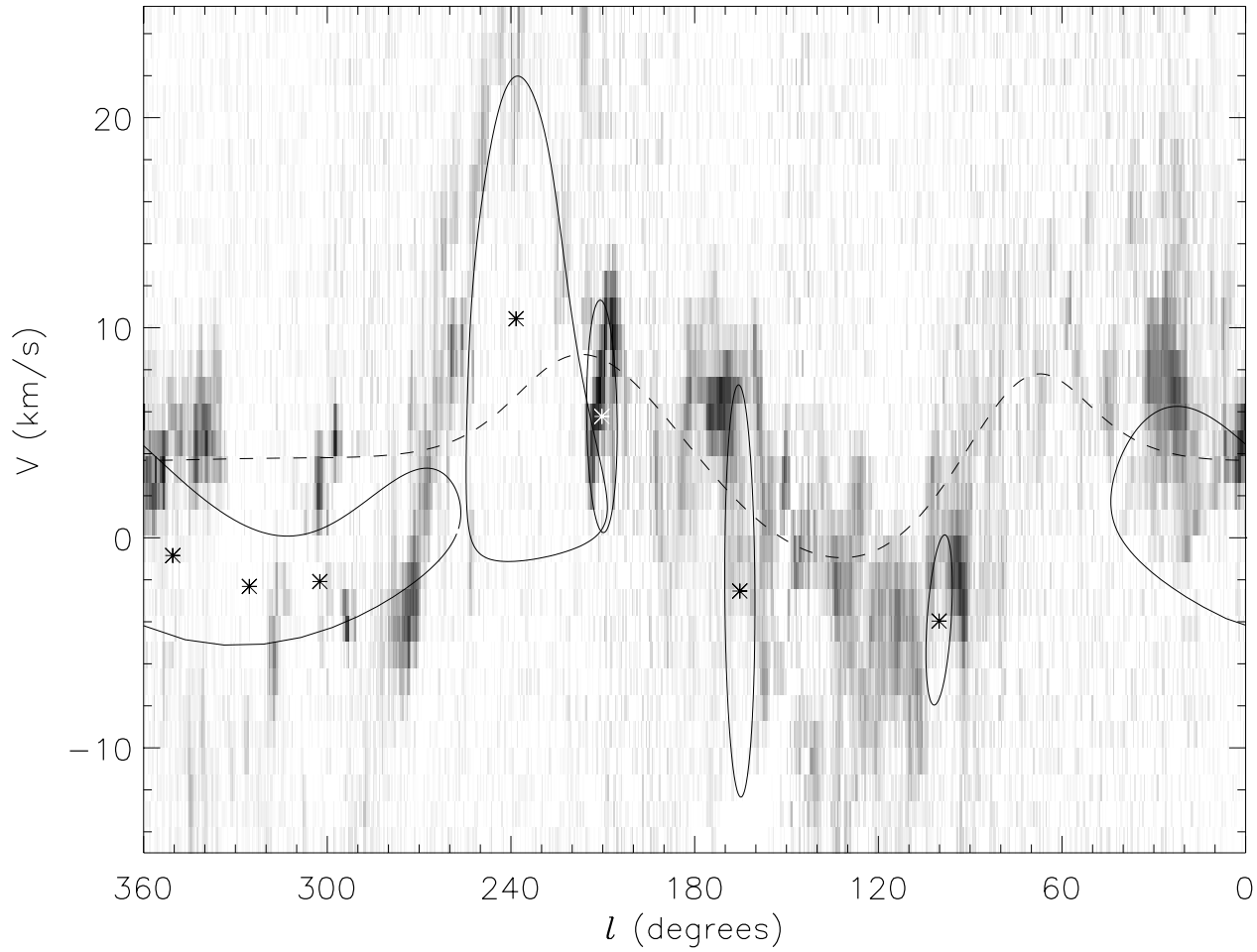


Figure 5. Position–velocity diagram of local CO gas: radial velocity of the gas versus Galactic longitude. The gray scale represents the brightness temperature of the CO clouds. The nearby OB associations are indicated by asterisks (the three subgroups of Sco OB2 at $\ell = 300^\circ$–360°, Collinder 121 at $\ell = 240^\circ$, Ori OB1 at $\ell = 210^\circ$, Per OB2 at $\ell = 165^\circ$ and Lac OB1 at $\ell = 100^\circ$). The curves (solid or broken) show the expected velocity fields for gas shells assuming that the shell velocities reflect both differential Galactic rotation and expansion powered by the energy released by the early-type association members. The broken curve denotes the supershell powered by the at present fossil Cas–Tau association (Lindblad's Ring); the solid curves denote local bubbles powered by the young OB associations in the Solar neighborhood (see above). Note that the association Collinder 121 is not part of Gould's Belt although its shell does interact with Lindblad's Ring. Adapted from Bally, Theil and Sutherland (2000 Orion's superbubble *The Orion Complex Revisited (ASP Conf. Ser.)* ed M J McCaughrean and A Burkert).

in the past 60 Myr. Detailed knowledge of the local star formation history will ultimately be of great value in the study of extragalactic Gould Belt analogs (i.e. massive-star formation in spiral DISK GALAXIES) and STARBURST GALAXIES.

Bibliography

A comprehensive study of Gould's Belt is presented by

Pöppel W 1997 The Gould Belt system and the local interstellar medium *Fundam. Cosm. Phys.* **18**, 1–271

Information on the Hipparcos and GAIA satellite projects can be found at

http://astro.estec.esa.nl/hipparcos/

http://astro.estec.esa.nl/gaia/

P T de Zeeuw, R Hoogerwerf, J H J de Bruijne, A G A Brown and A Blaauw

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gould, Benjamin Apthorp (1824–96)

Astronomer, born in Boston, MA, his early work in Germany was on the observation and motion of comets and asteroids. His greatest work was his mapping of the stars of the southern skies. He helped found the National Observatory in Cordoba, Argentina, and was its director as it compiled a catalog of stars using recently developed photometric methods. On returning to the USA, he spent his final years measuring the star clusters he had photographed in Argentina, and published the results as *Uranometry of the Southern Heavens*. He founded the *Astronomical Journal*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Graham, George (*c.* 1674–1751)

Instrument-maker, born in Hethersgill, Cumberland, England. Employee, partner, and by virtue of marriage to the niece of clock-maker THOMAS TOMPION, his heir. With TOMPION he made an *orrery*, the original one from which the name comes, for the Earl of Orrery; in fact he may have invented the instrument (which is a mechanical model for demonstrating the relative positions and motions of the planets). He invented improvements to precision clock mechanisms such as the deadbeat escapement, the mercury compensated pendulum and the cylinder escapement for watches, and he used clocks he had made to measure the exact shape of the Earth, correcting NEWTON'S figures. He made instruments for HALLEY (an 8 foot quadrant with vernier) and BRADLEY (a 12.5 foot zenith sector with a micrometer screw with which Bradley discovered the aberration of light). He also made the apparatus used for the measurement of a degree of the meridian by the French Académie des Sciences in its northern expedition.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Granat

Soviet gamma ray mission, launched December 1989. Carried the French SIGMA coded mask telescope, designed for high resolution imaging in the range 30 keV–2 MeV. Also carried instruments for x-ray imaging and spectroscopy, and x-ray/gamma ray burst detectors. Made detailed studies of the Galactic centre. Operated for more than five years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grand Unified Theory

A theory that attempts to show that the strong nuclear force, the weak nuclear force and the electromagnetic force are different manifestations of a single, unified, fundamental force.

Each of these forces may be described by quantum field theories, which imply that forces are communicated ('mediated') between real particles by means of virtual particles (particles that cannot be observed directly but which are permitted brief existence by the Heisenberg uncertainty principle). The electromagnetic force is conveyed by virtual photons, the weak nuclear force by massive 'W' and 'Z' particles (intermediate vector bosons) and the strong nuclear force between nucleons, by virtual mesons, and between quarks, by 'gluons'. Particles that are acted on by the strong nuclear force are called hadrons (hadrons, of which protons and neutrons are examples, are composed of quarks), those which do not experience the strong force are called leptons. In the present, low-energy ('cool') phase of the universe, the four fundamental forces (gravitation is the fourth force) are separate and distinct, and have markedly different strengths. Grand unified theories (or 'GUTs') attempt to unify the weak, strong and electromagnetic interactions. Many physicists believe that a more comprehensive theory (a 'theory of everything') that unifies all four forces will eventually be found.

The first step towards unification was achieved by the Weinberg–Salam–Glashow theory of the electroweak force, published in 1967, which predicted that at particle energies greater than 100 GeV (10^{11} eV), the electromagnetic and weak nuclear interactions become indistinguishable from each other (force-carrying W and Z particles would behave in the same way as virtual photons) and act as a unified force. The theory received strong confirmation through the discovery, in 1983, of the predicted W and Z bosons. Grand unified theories predict that at much higher energies, around 10^{15} GeV (10^{24} eV), the electroweak and strong interactions lose their separate identities and behave as a single, unified 'superforce'. Such energies far exceed the capabilities of terrestrial particle accelerators but, according to the Big Bang model, would have existed throughout the universe during the first 10^{-35} s of its existence, when the temperature everywhere was in excess of 10^{28} K. Thereafter, as the universe expanded and cooled, the unified force would have split into the strong and electroweak forces (the 'symmetry' between the forces was broken). About 10^{-11} s later, when the temperature had dropped to around 10^{15} K, the electroweak force would have separated out into the electromagnetic and weak nuclear forces.

The simplest grand unified theory, proposed in 1974 by Sheldon Glashow and Howard Georgi, requires the existence of a set of supermassive particles ('X-particles') that can transform quarks into leptons and leptons into quarks (the unified force then acts equally on both kinds of particle). Because quarks are the basic constituents of protons (and of all the hadrons), the theory predicts that protons themselves must eventually decay into leptons and photons (proton decay would occur if two of the constituent quarks inside a proton came sufficiently close to enable a virtual X-particle to be exchanged between them). The mean lifetime of a proton is expected to be at least 10^{32} years but, given a large enough sample of protons, the occasional decay ought to be detectable at the present time. This is being attempted by the Kamio Kande detector. Experiments have not yet detected such an event, but if the theory is confirmed, then it implies the eventual demise of matter as we know it.

See also: Big Bang theory, electroweak force, fundamental forces, gauge bosons, hadrons, leptons, quark, string theory/superstring theory, supersymmetry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Granulation

The mottled structure exhibited by the visible surface of the Sun, in other words the photosphere, when it is observed under good conditions. Due to convection hot gas is rising and cooler gas descending in the outer layers of the Sun, and this gives rise to the visible granules, typically 1200 km in diameter, which are about 100 K hotter than the duller regions which separate them.

See also: solar photosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitation

One of the four fundamental forces, gravitation is the force of mutual attraction that is exerted between massive bodies and between particles that have mass. Although gravitation is far weaker than the other three forces (the electromagnetic force and the weak and strong nuclear forces) over short ranges, it is the dominant force on large scales because its range of influence is far greater than that of the nuclear forces and because, unlike electrical charges (which can be positive or negative), all mass is positive (and mutually attractive). Because like charges repel and opposite charges attract, and matter on the large scale is electrically neutral (there are equal numbers of negative and positive charges in the universe), the electromagnetic force does not play a major role in the overall dynamics of the universe. Consequently, gravitation alone determines the motions of, and mutual interactions between, planets, stars and galaxies, and dominates the overall dynamics of the universe.

In 1687 Isaac Newton (1642–1727) published his law of universal gravitation in his book, *De Philosophiae Naturalis Principia Mathematica* (The Mathematical Principles of Natural Philosophy). It stated that the force of attraction between any two masses is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. For two bodies, with masses m_1 and m_2, separated by distance, r, the mutual force of attraction, F, is given by $F = Gm_1m_2/r^2$, where G is the universal gravitational constant ($=6.670 \times 10^{-11}$ N m^2 kg^{-2}). An equal and opposite force acts on each of the masses; thus the force exerted on m_1 is equal in magnitude but opposite in direction to the force exerted on m_2.

According to Newtonian theory, gravity is a force that acts, instantaneously, directly between individual bodies and particles (a process that is called 'action at a distance'). Although, for most purposes, Newtonian gravitation gives a perfectly satisfactory description of the motions of, for example, projectiles, planets, stars and galaxies, in certain situations (for example, where gravitation is very strong, as in the vicinity of a black hole, or when dealing with the structure and dynamics of the universe as a whole), the theory is inadequate. The best current theory of gravitation is Einstein's general theory of relativity, according to which mass distorts the geometry of space (or, strictly, four-dimensional spacetime) and the paths followed by material particles, or rays of light, in the neighborhood of massive bodies are determined by these local distortions of space.

See also: fundamental forces, general theory of relativity, graviton, mass, spacetime, supersymmetry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitational Lensing

Gravity bends light rays in a way analogous to, but quantitatively different from, the way it bends trajectories of passing particles. If light from some bright object passes close enough to some foreground mass, that object's image will be altered. The effect is more like a piece of bathroom glass in the sky than a precision-ground and well-focused lens, but the terms 'gravitational lensing' or simply 'lensing' have stuck. The observable effect varies, depending on the mass of the lens and the relative positions of lens and background source. 'Weak lensing' produces a shape change in the image, typically a stretching along one direction; for very weak lensing such stretching is imperceptible in images of individual sources but statistically measurable from images of many sources close together on the sky. 'Strong lensing' produces multiple images of the same source, one or more being brighter than the unlensed source would have been, and typically all having very distorted shapes.

The bending of light by the Sun can be measured directly, but apart from this all known examples of lensing involve very distant objects. (A peculiar feature of the physics of lensing is that lensing becomes easier to find as one observes more distant objects.) Lensing of stars in the Milky Way and its environs by foreground stellar-sized masses is now routinely observed; this requires very large surveys because of order 1 in 10^6 stars is being strongly lensed. Lensing by foreground galaxies of quasars is proportionately much more common—of order 1 in 10^3 quasars is strongly lensed. GRAVITATIONAL LENSING BY CLUSTERS OF GALAXIES is ubiquitous—a long-exposure image of any rich cluster of galaxies at cosmological distance will show obvious weak lensing and possibly also strong lensing.

What makes lensing particularly attractive is that it depends only on gravity and geometrical things such as source and lens positions; the effects of lensing do not come mixed up with other (possibly poorly understood) physical effects. Thus astrophysical inferences made from lensing will hopefully be particularly robust. Lensing is astrophysically most important as a probe of dark matter (see DARK MATTER: ITS NATURE). However, it has other, sometimes surprising, uses too.

The main part of this article is based on physical arguments. Mathematical material is confined to sections indicated between ♢ and ♦, and these may be skipped.

Basic physics of lensing

Photons are affected by a gravitational field, but not in the same way as massive particles are. For the details one needs general relativity, and because of this lensing was considered a very difficult subject for a long time (see GENERAL RELATIVITY AND GRAVITATION). However, in astrophysical situations some useful approximations apply, and these make the physics very much simpler than in full general relativity. If anything, lensing is easier than the astrophysical applications of Newtonian dynamics.

There are several ways of expressing the effect of gravity on light, in the context of lensing. Perhaps the most intuitive is through the wavefront. A wavefront is the locus of points with a given light travel time from the source. When a wavefront crosses an observer, they see an image of the source in the direction normal to the wavefront. Normally a wavefront is an expanding sphere, and light travels in straight lines. The effect of a gravitational field is to delay the part of the wavefront that passes through it. This naturally changes the shape of the wavefront and hence the direction normal to it, i.e. the position of the image, which is equivalent to saying that the light ray is bent. Figure 1 illustrates this. Note that if the gravitational delay is large enough, the wavefront can double in on itself, producing multiple images (strong lensing). A good way to gain some intuition about the wavefront is to take a plastic transparency with a blank piece of paper behind it and look at the reflections of a lightbulb. The shape of the plastic produces delays in the lightbulb's wavefront analogous to the gravitational time delays. Notice how changing the shape of the plastic causes images to merge and split, and the appearance of grotesquely stretched images at these transitions. Images of galaxies lensed by clusters show just such effects.

The wavefront formulation is not in practice the most useful because it works in terms of a single source and many observers, whereas in astrophysics one is usually interested in many sources and a single observer. So lensing work generally uses two other formulations and some standard approximations.

The standard approximations are the following. First, the gravitational fields are weak enough that massive particles simply follow Newtonian dynamics. (Weak gravitational fields should not be confused with weak lensing—strong lensing can still happen.) Second, the extent of the lens masses along the line of sight is negligible compared with the distances between source and lens and observer. Then a lens behaves as if its mass had been squashed along the line of sight into a sheet of mass. The mass distribution transverse to the line of sight is still important, however. Third, the lens and source are nearly aligned along the line of sight, in the sense that the angular separation (say θ) on the sky between the lens and source is such that $\sin\theta \simeq \theta$. Fourth, diffraction effects are negligible, because even for radio waves the wavelengths are too small.

With these approximations, the effect of gravity on photons can be described by adopting the following result from general relativity. A point mass M at perpendicular distance R from a light ray bends the light towards it by an angular amount

$$\frac{4GM}{c^2R}$$

where G is the gravitational constant and c the speed of light. For the weak field approximation to apply R must be much larger than $2GM/c^2$ (the Schwarzschild radius), but this condition is easily satisfied because Schwarzschild radii are at most a few km for stars, <1 pc

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

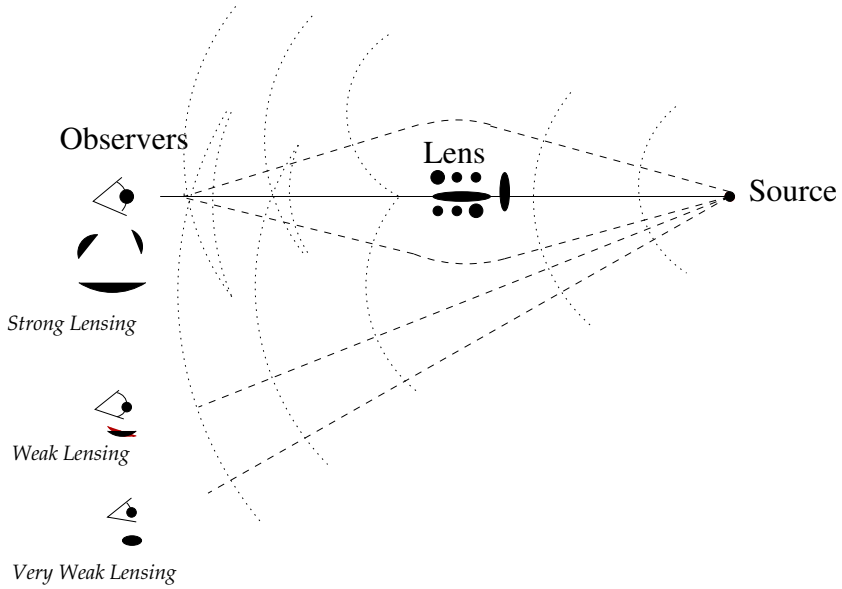


Figure 1. Schematic illustration of the wavefront. Observers see images of the source in the direction normal to the wavefront where the wavefront crosses them. If delayed enough by the lens, part of the wavefront can double in on itself, producing multiple images for some observers (strong lensing); other observers see single images (weak lensing). Images usually experience size and shape changes.

for a galaxy and $\lesssim 10^3$ pc for a cluster of galaxies. (The bending angle of $4GM/c^2R$ was the first prediction of general relativity; Newtonian dynamics would predict half that value. Measurement of the bending angle for the Sun demonstrated that the dynamics of photons is not Newtonian, and was the first test of general relativity.)

Alternatively, one can proceed from another result from general relativity, which is that a photon passing by a point mass M and then reaching an observer from a direction θ relative to the mass will have had its arrival time been delayed by

$$-\frac{4GM}{c^3}\ln\theta$$

by the mass. (Here θ will be very small, so the logarithm will be negative; hence the minus sign.) Now Fermat's principle is that photons take paths which make the arrival time a maximum, minimum or saddle point. In the absence of special delays photons simply minimize the arrival time, i.e. they follow straight lines. However, with a gravitational time delay, photons will follow a different path; this path will be such that the new arrival time (considering both the increased path length and the gravitational delay) is still a maximum, minimum or saddle point. It is often useful to consider the arrival time as a function of the arrival direction of photons from a particular source past some lens: this is known as the 'arrival time surface'. It is related to the wavefront, but is not quite the same thing, for the wavefront is a surface in real space whereas the arrival time surface is a more abstract entity.

The bending angle and arrival time formulations are equivalent. In both cases the effect of a distributed mass (rather than a point mass) is simply the sum of contributions from all points making up the mass distribution.

◊ *Derivation of the arrival time*

Consider a situation as in figure 2, where an observer is viewing a source at distance $D_{\rm S}$, with a lens (a mass screen) intervening at distance $D_{\rm L}$; $D_{\rm LS}$ is the distance from lens to the source. On galactic scales $D_{\rm L}$, $D_{\rm S}$, $D_{\rm LS}$ are ordinary distances, but on cosmological scales they must be understood as angular diameter distances, and $D_{\rm S} \neq D_{\rm L} + D_{\rm LS}$. (The reason for this complication is that the universe expands significantly over the light travel time.) We use angular coordinates for the transverse position; $\boldsymbol{\theta}_{\rm S}$ is the position of the source and $\boldsymbol{\theta}$ is its observed position after being lensed. Let $\Sigma(\boldsymbol{\theta})$ be the lens's surface mass density, i.e. mass per unit solid angle. (An important notational point: $\Sigma(\boldsymbol{\theta})$ is often defined as mass per unit physical area, i.e. $D_{\rm L}^{-2}$ times the convention in this article.) Let $\boldsymbol{\alpha}(\boldsymbol{\theta})$ be the deflection angle. Then, comparing vectors in the source plane, we obtain

$$D_{\rm S}\boldsymbol{\theta} = D_{\rm S}\boldsymbol{\theta}_{\rm S} + D_{\rm LS}\boldsymbol{\alpha}. \tag{1}$$

(By convention, $\boldsymbol{\alpha}$ is directed outwards from the deflecting mass rather than towards it.) Using $4GM/c^2R$ for the deflection from a point mass, we obtain

$$\boldsymbol{\theta} = \boldsymbol{\theta}_{\rm S} + \frac{D_{\rm LS}}{D_{\rm S}}\boldsymbol{\alpha}(\boldsymbol{\theta}) \qquad \boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{4G}{c^2 D_{\rm L}}\int \frac{\Sigma(\boldsymbol{\theta}')(\boldsymbol{\theta}-\boldsymbol{\theta}')\,{\rm d}^2\boldsymbol{\theta}'}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|^2} \tag{2}$$

This is known as the lens equation. It gives $\boldsymbol{\theta}_{\rm S}$ as an explicit function of $\boldsymbol{\theta}$, but $\boldsymbol{\theta}$ as an implicit function of $\boldsymbol{\theta}_{\rm S}$. Moreover,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

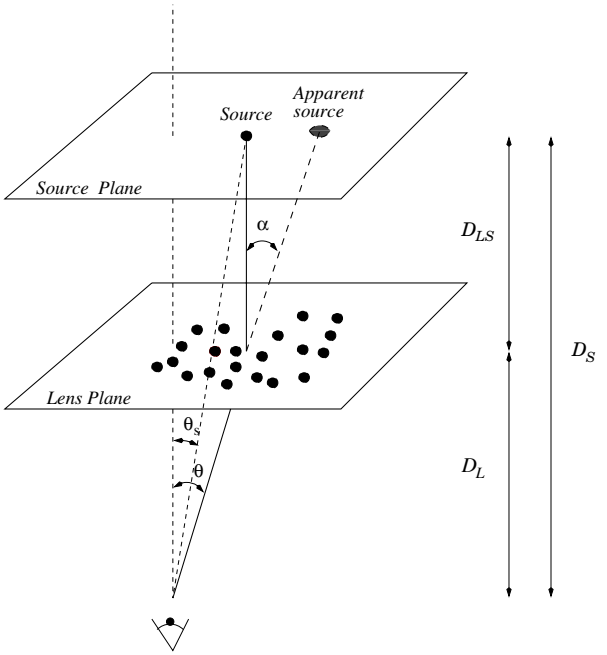


Figure 2. The geometrical meanings of the various distances and angles used in lensing work.

$\boldsymbol{\theta}(\boldsymbol{\theta}_{\rm S})$ need not be single valued, so sources can be multiply imaged.

We can change from this bending-angle formalism to the arrival time formalism by noting that the lens equation 2 amounts to equating a gradient to zero:

$$\boldsymbol{\nabla} T(\boldsymbol{\theta}) = 0 \qquad T(\boldsymbol{\theta}) = \tfrac{1}{2} T_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_{\rm S})^2 - \Psi(\boldsymbol{\theta})$$

$$\Psi(\boldsymbol{\theta}) = (1 + z_{\rm L}) \frac{4G}{c^3} \int \Sigma(\boldsymbol{\theta}') \ln |\boldsymbol{\theta} - \boldsymbol{\theta}'| \, {\rm d}^2\boldsymbol{\theta}'$$

$$T_0 = (1 + z_{\rm L}) \frac{D_{\rm L} D_{\rm S}}{c D_{\rm LS}}. \tag{3}$$

Here $T(\boldsymbol{\theta})$ is of course the arrival time, up to an irrelevant constant, for arbitrary deflections: the first term in it is the change in light travel from geometrical considerations alone; the second term is the gravitational time delay (the factor of $1 + z_{\rm L}$ comes because everything that happens in the lens plane is observed redshifted, and that goes for time delays too). The deflections that actually occur are those that make $T(\boldsymbol{\theta})$ stationary; this is Fermat's principle. Later on, we will rearrange equation (3) to show the dimensions involved a little better. ♦

The Einstein radius

In the absence of lensing the wavefront is spherical. However, in lensing we are always only interested in very small regions of the wavefront, and over a small enough angular region a sphere is indistinguishable from a parabola. Similarly, in the absence of lensing the relevant part of the arrival time surface is also parabolic, with the image of course at the bottom.

Now consider what happens if a point lens is interposed exactly along the line of sight to the source. The arrival time surface will be raised according to the lens's time delay, and if the lens is strong enough the bottom of the parabola will become a dimple. The old minimum will now become a maximum, and the whole edge of the dimple will become a minimum. Observationally, the dimple edge corresponds to a spectacular ring image, called an Einstein ring. Its radius is known as the Einstein radius, say $\theta_{\rm E}$; if the source is much further away than the lens

$$\theta_{\rm E} \simeq 0.1 \text{ arcsec} \times \left(\frac{M \text{ in } M_\odot}{\text{lens distance in parsecs}} \right)^{1/2}$$

where $M_\odot$ denotes the solar mass.

Actually the lens need not be a point mass. Just as in Newtonian gravity a sphere attracts externally as if it were a point mass, a circular lens affects light outside the circle as if it were a point mass. So any circular mass can produce an Einstein ring, provided that the Einstein ring is 'outside the mass'. In other words the mass needs to be smaller than a length that projects to $\theta_{\rm E}$; this length, which we may denote by $R_{\rm E}$ ($=\theta_{\rm E} \times$ lens distance) is also called the Einstein radius. (Whether one means $\theta_{\rm E}$ or $R_{\rm E}$ will always be clear from context.) It turns out that $R_{\rm E}$ is roughly the geometric mean of the Schwarzschild radius and the lens distance. Thus for the ANDROMEDA GALAXY, $R_{\rm E}$ is some tens of parsecs—much smaller than the galaxy itself. However, a similar galaxy at cosmological distances will have $R_{\rm E}$ comparable with or bigger than its own size, and thus can produce strong lensing. Masses that are smaller than their $R_{\rm E}$ are said to be 'compact' to lensing.

In practice, Einstein rings are rarely observed. The reason is that the above scenario for forming one requires a perfectly circular lens and a perfectly aligned source. These improbable requirements relax a little if the source is very broad, however, and a few examples have been observed. Nevertheless, the *concept* of an Einstein ring is very important, because even if there is no Einstein ring the interesting strong lensing behavior all happens over a scale of $\theta_{\rm E}$. The larger $\theta_{\rm E}$, the likelier it is that there are some background sources to strongly lens. Thus from the above expression for $\theta_{\rm E}$ one can easily infer that lensing by massive galaxies ($M \sim 10^{12} M_\odot$ and $\theta_{\rm E}$ perhaps $\sim$1 arcsec) is much likelier than lensing by nearby stars (where $\theta_{\rm E} \ll 1$ arcsec), while rich clusters of galaxies (with $M \gtrsim 10^{14} M_\odot$ and $\theta_{\rm E} \sim 10$ arcsec) are even more favorable for lensing.

◊ *The Einstein radius and other scales*

Expressions for $\theta_{\rm E}$ and $R_{\rm E}$ are easily derived by considering the lens equation for a collinear point mass and source, i.e. $\boldsymbol{\theta}_{\rm S} = 0$ and $\Sigma(\boldsymbol{\theta}) = M\delta(\boldsymbol{\theta})$. The lens equation is trivially solved by $\theta = \theta_{\rm E}$, with

$$\theta_{\rm E}^2 = \frac{4GM}{c^2} \frac{D_{\rm LS}}{D_{\rm L} D_{\rm S}} \qquad R_{\rm E}^2 = \frac{4GM}{c^2} \frac{D_{\rm L} D_{\rm LS}}{D_{\rm S}}. \tag{4}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

As mentioned in the previous section

$$R_{\rm E} \sim (\text{Schwarzschild radius} \times D_{\rm L})^{1/2}. \qquad (5)$$

By a Gauss law type argument, for any circular mass distribution $\Sigma(\theta_r)$, $\Psi(\boldsymbol{\theta})$ and $\boldsymbol{\alpha}(\boldsymbol{\theta})$ will be influenced only by interior mass. So any circular distribution of the mass M, *provided that it fits within an Einstein radius*, can produce an Einstein ring.

For given $D_{\rm L}$, $D_{\rm S}$, to obtain a compact object one needs to pack a mass (in projection) into a circle of radius $\theta_{\rm E}$, but the area of the circle is proportional to the mass. It follows that there is a critical density, say $\Sigma_{\rm crit}$, such that if $\Sigma \geq \Sigma_{\rm crit}$ somewhere then there is a compact (sub)object. Working out the algebra we easily find

$$\Sigma_{\rm crit} = \frac{D_{\rm L} D_{\rm S}}{D_{\rm LS}} \frac{c^2}{4\pi G}. \qquad (6)$$

If now we define κ as the projected mass density in units of the critical density, equation (3) can be rewritten as

$$\boldsymbol{\nabla}\tau(\boldsymbol{\theta}) = 0 \qquad \tau(\boldsymbol{\theta}) = \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\rm S})^2 - \psi(\boldsymbol{\theta})$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{\pi} \int \kappa(\boldsymbol{\theta}') \ln|\boldsymbol{\theta} - \boldsymbol{\theta}'| \, {\rm d}^2\boldsymbol{\theta}'$$

$$T(\boldsymbol{\theta}) = (1 + z_{\rm L}) \frac{D_{\rm L} D_{\rm S}}{c D_{\rm LS}} \tau(\boldsymbol{\theta}). \qquad (7)$$

From the second line of equation (7) it should be evident that ψ satisfies a two-dimensional Poisson equation

$$\nabla^2 \psi = 2\kappa. \qquad (8)$$

Note that the first two lines in equation (7) involve only dimensionless quantities and the scale of $T(\boldsymbol{\theta})$ is set entirely by the distances and lens redshift. ♦

On images and magnification

Let us now see what can happen to the arrival time surface from masses of arbitrary shape. For small masses, the shape changes slightly from being a parabola and the mininum moves a little. However, for large enough mass, a qualitative change occurs, in that a contour of constant arrival time becomes self-crossing. There are two ways in which a self-crossing can develop: as a kink on the outside of a contour line or a kink on the inside. These are illustrated in figure 3. The outer-kink type is topologically a lemniscate and the inner-kink type a limaçon. If the original contour loop enclosed a minimum then a lemniscate produces another minimum, plus a saddle-point at the self-crossing, while a limaçon produces a new maximum plus a saddle point. (Interchange maximum and minimum in the previous sentence if the original loop enclosed a maximum.) The process of contour self-crossing can then repeat around any of the new maxima and minima, producing more and more new images, but always satisfying

$$\text{maxima} + \text{minima} = \text{saddle points} + 1.$$

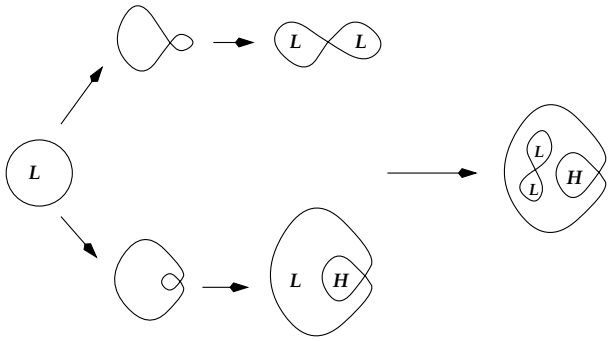


Figure 3. Multiple images via self-crossing contours in the arrival time surface. Here L marks minima and H marks maxima.

The rightmost part of figure 3 shows a typical configuration in observed systems.

Multiple images are not all equally bright. It turns out that each image is magnified by an amount equal to the inverse of the curvature of the arrival time surface. This curvature need not be the same in all directions. Very stretched images (of which an Einstein ring is an extreme example) correspond to very small curvature in one direction with larger curvature in the perpendicular direction. The standard way of quantifying lensing magnification (also called amplification in the literature) is through the 'convergence' which is total change in brightness and the 'shear' which is the amount and direction of stretching.

If the lens's mass density is infinite anywhere (as for any point mass, or say a galaxy with a central BLACK HOLE) the arrival time surface is going to have a needle-sharp but infinitely high maximum at that point, and, while this will formally be an image, because of the needle-sharp curvature the image will be (de)magnified to nothing. For example the five-image configuration in figure 3 is always in practice a four-image system, with the maximum completely demagnified by a very high central density in the lens.

Surface brightness (photon flux per unit sky area per unit telescope area) is conserved by lensing. Magnification changes only angular sizes on the sky. Thus a constant surface brightness sheet stays a constant brightness sheet when lensed. If this were not the case, the microwave background would be wildly lensed by large-scale structure. Polarization of light is also unaffected by lensing.

For unresolved sources, magnification is not observed directly; only their total luminosity changes under lensing. This leads to an interesting effect known as magnification (or amplification) bias. Imagine an unlensed patch of sky that has been surveyed down to some brightness limit for some type of unresolved objects. Now imagine the same patch lensed with some magnification factor A, for definiteness say $A > 1$. This magnification will increase the area of the patch by A, and if the survey area stays

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the same the source area will be reduced by A. Thus the number of objects available will also be reduced by A. However, at the same time all the objects will be brightened by A, so some objects that were originally below the survey brightness limit will now climb above it. The net effect depends on the luminosity function: if going fainter leads to enough new objects (a steep luminosity function) then the net effect will be an increase in the number of sources detected, otherwise there will be a net decrease.

Magnification

Magnification is formalized as the derivative of the image position with respect to the source position. It is a tensor, which we denote by $\mathbf{M}$. We have

$$\mathbf{M}^{-1} = \frac{\partial \boldsymbol{\theta}_{\rm S}}{\partial \boldsymbol{\theta}} = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} T(\boldsymbol{\theta}). \tag{9}$$

In Cartesian coordinates

$$\mathbf{M}^{-1} = \begin{bmatrix} 1 - \dfrac{\partial^2 \psi}{\partial \theta_x^2} & \dfrac{\partial^2 \psi}{\partial \theta_x \partial \theta_y} \\ \dfrac{\partial^2 \psi}{\partial \theta_y \partial \theta_x} & 1 - \dfrac{\partial^2 \psi}{\partial \theta_y^2} \end{bmatrix} \tag{10}$$

It is helpful to write $\mathbf{M}^{-1}$ in terms of its eigenvalues, and the usual form is like

$$\mathbf{M}^{-1} = (1-\kappa) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \gamma \begin{pmatrix} \cos 2\phi & \sin 2\phi \\ \sin 2\phi & -\cos 2\phi \end{pmatrix}. \tag{11}$$

The first term in equation (1) is the trace part—and comparing equations (10) and (8) shows that it must be κ—while the second term is traceless. The κ part produces an isotropic expansion or contraction, while the γ part produces a stretching in the ϕ direction and a shrinking in the perpendicular direction; κ is known as 'convergence' and γ as 'shear'.

The determinant of $\mathbf{M}$ can be thought of as a scalar magnification:

$$|\mathbf{M}| = [(1-\kappa)^2 - \gamma^2]^{-1}. \tag{12}$$

The places where one of the eigenvalues of $\mathbf{M}^{-1}$ becomes zero (and in consequence $|\mathbf{M}|$ is infinite) are in general curves and are known as critical curves. When critical curves are mapped onto the source plane through the lens equation, they give caustics; a point source lying on a caustic is infinitely magnified. Some straightforward algebra shows that for a point mass

$$|\mathbf{M}|^{-1} = 1 - \frac{\theta_{\rm E}^4}{\theta_r^4} \tag{13}$$

and that for a circular mass distribution $\Sigma \propto \theta_r^{-1}$

$$|\mathbf{M}|^{-1} = 1 - \frac{\theta_{\rm E}}{\theta_r}. \tag{14}$$

The latter case is known as the 'isothermal lens' after the isothermal sphere in stellar dynamics.

One way of proving that surface brightness is conserved by lensing is by setting up a correspondence with Hamiltonian dynamics. Let us write z for the axial coordinate in figure 2, and $\boldsymbol{w}$ for the transverse coordinates, and allow for observers at arbitrary $\boldsymbol{w}$, z. A photon trajectory can be expressed as $\boldsymbol{w}(z)$, and an observer at $\boldsymbol{w}$, z sees the photon coming from direction $\boldsymbol{\theta} = \mathrm{d}\boldsymbol{w}/\mathrm{d}z$ (in the small-angle regime). In these coordinates the lens equation $\boldsymbol{\theta}_{\rm S} = \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta})$ becomes

$$\begin{aligned} &\boldsymbol{\theta}(\boldsymbol{w}_{\rm obs}, z_{\rm obs}) - \boldsymbol{\theta}(\boldsymbol{w}_{\rm source}, z_{\rm source}) \\ &\qquad = (z_{\rm obs} - z_{\rm lens}) \frac{\partial}{\partial \boldsymbol{w}} \psi \left(\frac{\boldsymbol{w}_{\rm lens}}{z_{\rm obs} - z_{\rm lens}} \right). \end{aligned} \tag{15}$$

However, this is a solution of Hamilton's equations for the Hamiltonian

$$H = \tfrac{1}{2}\boldsymbol{\theta}\cdot\boldsymbol{\theta} - (z_{\rm obs} - z_{\rm lens}) \psi \left(\frac{\boldsymbol{w}_{\rm lens}}{z_{\rm obs} - z_{\rm lens}} \right) \delta(z - z_{\rm lens}) \tag{16}$$

with z as a formal time variable and $\boldsymbol{\theta} = \mathrm{d}\boldsymbol{w}/\mathrm{d}z$ identified as the momentum. By Liouville's theorem (or equivalently the collisionless Boltzmann equation), the photon density (say f) in the phase space of ($\boldsymbol{w}$, $\boldsymbol{\theta}$) space is conserved along photon trajectories. This f must be conserved by the act of placing the lens there too—think of f before and after going through the lens. In fact f is nothing but the surface brightness, because θ corresponds to sky area and $\boldsymbol{w}$ to telescope area. This proves surface brightness conservation, although we must be careful with two things about interpretation. First, because the result is 'along photon trajectories' we must always be looking at photons from the same source; so if the image is moved in the sky by lensing we must follow it when we measure surface brightness. Second, z here is just a formal variable and is neither time nor REDSHIFT, so the above does not imply that surface brightness is independent of redshift.

Lensing in the Milky Way

The Einstein radius for lensing of a Milky Way star by another star or a brown dwarf is, for typical distances, only ~1 AU for $R_{\rm E}$ and <1 marcsec for $\theta_{\rm E}$. Nevertheless, because of stellar motions, stars do occasionally manage to come within an Einstein radius (in projection of course) of lenses in the Milky Way. The fraction of source stars in a field that are doing this at a given time is known as the microlensing optical depth, and it is of order 10^{-6} in the Milky Way. ('Microlensing' is a term used when stellar-mass lenses are important.) As the source enters and then exits the Einstein radius of the lens, the image brightens and then dims again in a characteristic way, over timescales of weeks to months. Such lensing is now being detected by large surveys that monitor ~10^7 stars, searching for the characteristic behavior of the brightness. The aim of these surveys is to measure the optical depth in different directions and thence infer something about the mass density in compact objects in the Milky Way.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Microlensing optical depth

Approximating a star or a brown dwarf as a point lens, there will be two images, at

$$|\boldsymbol{\theta}| = \tfrac{1}{2}[|\boldsymbol{\theta}_S| \pm (|\boldsymbol{\theta}_S|^2 + 4\theta_E^2)^{1/2}]. \quad (17)$$

The image separation is too small to resolve. What will be observed is a brightening equal to the combined magnification of both images. Using equation (13) for the magnification, and adding the absolute values of $|\mathbf{M}|$ at the two image positions, we obtain the total brightness amplification

$$A_{\rm tot} = \frac{u^2+2}{u(u^2+4)^{1/2}} \qquad u = \frac{|\boldsymbol{\theta}_S|}{\theta_E}. \quad (18)$$

If the background source star has impact parameter b and velocity v (projected onto the lens plane) with respect to the lens, then

$$u = \frac{(b^2 + v^2t^2)^{1/2}}{R_E}. \quad (19)$$

Inserting equation (19) into (18) gives us $A_{\rm tot}(u)$, i.e. the light curve, plotted for three different b in figure 4. The height of a measured light curve immediately gives R_E/b, and the width gives R_E/v.

We can express the microlensing optical depth (usually denoted by τ) as follows. Using equation (4) for R_E, the total area covered by the Einstein rings of lenses at distances between D_L and $D_L + \mathrm{d}D_L$ in a patch of sky is

$$\frac{4\pi GM}{c^2 D_S} D_L D_{LS} \rho(D_L)\,\mathrm{d}D_L \times \langle \text{area of patch} \rangle. \quad (20)$$

To obtain the fraction of sky covered by Einstein rings, we divide equation (20) by the area and integrate over D_L, giving

$$\tau = \frac{4\pi G}{c^2 D_S} \int_0^{D_S} D_L D_{LS} \rho(D_L)\,\mathrm{d}D_L. \quad (21)$$

The really nice thing about this formula is that it does not depend on the mass distribution of the lenses, as long as each mass fits within its own Einstein radius (diffuse gas clouds do not count, nor does any kind of diffuse dark matter). So τ estimated from light curve monitoring could be used to make inferences about ρ.

Although trying to resolve the images in microlensing seems hopeless with foreseeable technology, it may be possible to track the moving double image indirectly. By combining the positions and magnifications of the two images, we have for the centroid

$$\theta_{\rm cen} = \frac{u(3+u^2)}{2+u^2}\theta_E. \quad (22)$$

Since images can be centroided much better (10^2–10^3 times more precisely) than they can be resolved, measuring the proper motion of the centroid seems a realistic prospect.

Lensing by galaxy clusters

At the other end of the lensing scale, we have clusters of galaxies, which are the most effective lensers of all. A rich cluster at cosmological distance will be effective for lensing over a large enough patch of sky to include plenty of background objects (faint blue galaxies) to lens. A number of clusters show all the regimes of lensing, with multiple-image systems in the inner region, arcs and arclets (highly stretched single images) outside of that and statistical stretching in the outer regions. Each of these regimes contributes different types of information about the mass distribution. Now galaxy clusters are generally not in dynamical equilibrium (there have not been enough crossing times since they formed), so their mass distributions tend to be more complicated than those of single galaxies. This makes it all the more important to have good ways of mapping cluster masses, and makes clusters particularly interesting to researchers in lensing.

Lensing constraints on clusters

There are several approaches in the literature to mass-mapping clusters from lensing data. The following sketches one approach.

Cluster lensing data provide three types of information:

(i) multiple image positions $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and so on;
(ii) relative brightnesses at $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and so on;
(iii) elongated images of sources may be obvious (arcs and arclets) or statistical (requiring a large number of background galaxies to measure).

Going back to equation (7), these can be translated into three types of constraints on $\kappa(\boldsymbol{\theta})$.

(i) Image positions imply $\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_1) = \mathbf{0}$, $\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_2) = 0$ and so on. From equation (7) these are clearly dimensionless linear constraints on $\boldsymbol{\theta}_S$ and $\kappa(\boldsymbol{\theta})$. (If there is more than one multiple-image system, there will be a separate $\boldsymbol{\theta}_S$ for each of these.)
(ii) Relative brightnesses are constraints of the type

$$\boldsymbol{\nabla}\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_1) = \langle \text{measured number} \rangle \times \boldsymbol{\nabla}\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_2).$$

(iii) Elongated images (whether arclets or statistical elongations) provide constraints of the type

$$k\frac{\partial^2}{\partial\theta_{x'}^2}\tau(\boldsymbol{\theta}_1) \le \frac{\partial^2}{\partial\theta_{y'}^2}\tau(\boldsymbol{\theta}_1)$$

where $\theta_{x'}$ and $\theta_{y'}$ are a coordinate system rotated to align with the observed direction of the elongation. This equation expresses the statement that the magnification along $\theta_{x'}$ is at least k times the magnification along the $\theta_{y'}$ direction. Provided that the elongation direction is well-measured, the constraint is linear in $\kappa(\boldsymbol{\theta})$.

Types (i) and (ii) are strong lensing constraints and available only in the inner regions of rich clusters. Type (iii)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

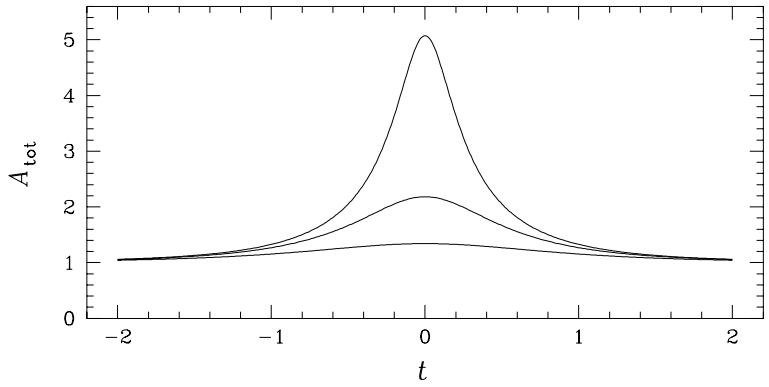


Figure 4. Light curves for impact parameters of $R_{\rm E}$ (lowest), $0.5R_{\rm E}$ and $0.2R_{\rm E}$. The unit of time is how long it takes the source to move a distance $R_{\rm E}$.

constraints apply also in the the weak lensing regime and can be extracted over a much larger area (∼5 times further out).

Expressed as above, cluster mass reconstruction becomes an underdetermined problem with linear, plus possibly some quadratic, constraints in both strong and weak lensing regimes. It is thus a relative of image reconstruction problems. The standard way of proceeding to derive an image (or a mass map in this case) would be to pixellate $\kappa(\boldsymbol{\theta})$, select some figure of merit on κ such as smoothness or entropy and then optimize it subject to the data constraints.

Multiple-image quasars

These occur when a QUASAR (or other small but very bright object) is strongly lensed by a foreground galaxy and produces two or four images with arcsec order separations. (A third or fifth image is invariably demagnified to undetectability by the lens's high-density core.) These are of some interest for constraining the masses of the lensing galaxies, but they are perhaps much more interesting to astrophysicists for two other reasons.

The first is that since quasars are often very time variable and the different images have different arrival times, the images will show the same time variability, but with offsets. These offsets, or differences in arrival times, are unique among lensing observables, because they are dimensional. All other lensing observables, i.e. image positions and magnifications, are dimensionless. Measuring a dimensional quantity immediately sets a scale for the system. That scale is the distance scale, which in a cosmological context amounts to HUBBLE'S CONSTANT. Unfortunately, an accurate measurement of variability offsets does not immediately give an accurate determination of Hubble's constant, because the offsets also depend on the distribution of mass in the lens plane; the latter still needs to be determined or modelled somehow. Still, the prospect of measuring Hubble's constant from lensing is attractive because it involves no extra physics apart from gravitation, thus hopefully leaving less room for the sort of systematic errors that plague Hubble constant estimates.

The second has to do with the extremely small size of quasars in optical continuum. Now the mass distribution of a galaxy is not perfectly smooth; it becomes granular on the scale of individual stars. This produces a very complicated network of critical lines (in the lens plane), and a corresponding complicated network of caustics in the source plane (like the pattern at the bottom of a swimming pool). The optical continuum emitting regions of quasars are small enough to fit between the caustics, but the line emitting regions straddle several caustics. As proper motions move the caustic network, the continuum region will sometimes cross a caustic, and show a sudden change in brightness; the time taken for the brightness to change is the time it take to cross the caustic. This is the phenomenon of quasar microlensing: continuum shows it but lines do not. (It is just the gravitational version of stars twinkling and planets not twinkling.) This has been observed, and modelling the caustic network and putting in plausible values for the proper motion leads to an estimate of the intrinsic size of the continuum source. The conclusion is that the optical continuum regions of a quasar—the heart that emits much more light than the entire host galaxy—cannot be much larger than the solar system.

Arrival times and Hubble's constant

In equation (7) $\tau(\boldsymbol{\theta})$ is dimensionless while $T(\boldsymbol{\theta})$ has dimensions of time, and the factor relating them is $\propto H_0^{-1}$. (This factor actually depends on Ω and Λ as well, but a case such as $z_{\rm L} < 0.5$, $z_{\rm S} > 2$, the Ω and Λ dependences are very weak.) So if we can measure the arrival time difference $T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2)$ between two images (and accurate measurements are now starting to appear) and we know $\kappa(\boldsymbol{\theta})$, we can determine H_0.

The difficulty here is that we need to know $\kappa(\boldsymbol{\theta})$, because for a given $T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2)$ the inferred H_0 can vary by a factor of several depending on the $\kappa(\boldsymbol{\theta})$ assumed. However, naturally, we can use the lensing data

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

themselves to help constrain $\kappa(\boldsymbol{\theta})$. The constraints (i) and (ii) from the cluster lensing case are available here also, since we have a multiple-image system. Constraints of the type (iii) may possibly be available. If there are more than two images then a fourth type of constraint exists, which comes from the ratio of arrival time differences between different pairs of images, i.e.

$$T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2) = \langle \text{measured number} \rangle \times [T(\boldsymbol{\theta}_2) - T(\boldsymbol{\theta}_3)].$$

Some algebra from equation (7) shows that this is a dimensionless linear constraint on $\boldsymbol{\theta}_S$ and $\kappa(\boldsymbol{\theta})$.

That the lensing constraints are mostly linear and at worst quadratic is a useful simplification, but still the number of constraints possible is small and leaves $\kappa(\boldsymbol{\theta})$ underdetermined, implying some residual uncertainty in H_0 even for perfect lensing data. Hopefully this uncertainty can be reduced by well-motivated modelling or other observations constraining $\kappa(\boldsymbol{\theta})$. However, whether the residual uncertainty in H_0 for any system can be made small enough to be interesting is at present an unanswered question.

Bibliography

The lensing literature being as large as it now is, the following cannot hope to be comprehensive. However, here are pointers to a number of key papers in the development of the field, and to some recent surveys of research in the field.

The first person to take gravitational lensing seriously as an observational prospect appears to have been Zwicky, in

Zwicky F 1937 *Phys. Rev.* **51** 290, 679

It is interesting to compare these papers with

Einstein A 1936 *Science* **84** 506

Einstein sounds almost apologetic for publishing the results for a point lens, because stellar lenses seemed improbable to observe. Zwicky, writing the following year, points out that for galaxies and clusters of galaxies the situation is very different, and lensing becomes 'virtually a certainty'.

Discoveries of the first two lensed systems (both multiply imaged quasars) appeared in

Walsh D, Carswell R F and Weymann R J 1979 *Nature* **279** 381
Weymann R J, Latham D, Angel J R P, Green R F, Liebert J W, Turnshek D A, Turnshek D E and Tyson J A 1980 *Nature* **285** 641

The first arcs in clusters were reported in

Soucail G, Fort B, Mellier Y and Picat J P 1987 *Astron. Astrophys.* **172** L14
Lynds R and Petrosian V 1987 *Bull. Am. Astron. Soc.* **18** 1014

The study of statistical weak lensing by clusters was initiated by

Tyson J A, Wenk R A and Valdes F 1990 *Astrophys. J.* **349** L1

The first quasar microlensing event is in

Irwin M J, Webster R L, Hewett P C, Corrigan R T and Jedrzejewski R I 1989 *Astron. J.* **98** 1989

The first galactic microlenses were published in

Alcock C *et al* 1993 *Nature* **365** 621
Aubourg E *et al* 1993 *Nature* **365** 623
Udalski A *et al* 1993 *Acta Astron.* **43** 289

Our present understanding of lensing theory is due very largely to Sjur Refsdal. The now-standard approximations and basic methods of calculation, and the relation between time delays and Hubble's constant, are set out in

Refsdal S 1964 *Mon. Not. R. Astron. Soc.* **128** 295
Refsdal S 1964 *Mon. Not. R. Astron. Soc.* **128** 307

These long preceded the discovery of the first lens. Quickly following the discovery of the first lensed system

Chang K and Refsdal S 1979 *Nature* **282** 561

predicted microlensing in lensed quasars. Further developments of the formalism appeared in

Kayser R and Refsdal S 1983 *Astron. Astrophys.* **128** 156

introducing the wavefront picture, and

Blandford R D and Narayan R 1986 *Astrophys. J.* **310** 568
Schneider P 1985 *Astron. Astrophys.* **143** 413

which introduce Fermat's principle by very different routes.

The most comprehensive book on lensing is

Schneider P, Ehlers J A and Falco E E 1992 *Gravitational Lenses* (Berlin: Springer)

The most recent conference proceedings covering all aspects of lensing is

Brainerd T G and Kochanek C S (ed) 2000 Gravitational lensing: recent progress and future goals *ASP Conference Series*

For eight unsolved problems in gravitational lensing, see

Blandford R D 1997 *Unsolved Problems in Astrophysics* ed J N Bahcall and J P Ostriker (Princeton, NJ: Princeton University Press)

P Saha

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitational Lensing by Clusters of Galaxies

CLUSTERS OF GALAXIES are massive and relatively rare objects containing hundreds of galaxies. Their huge mass—dominated by DARK MATTER—bends light from all background objects, systematically distorting the images of thousands of distant galaxies (shear). This observed gravitational lens distortion can be inverted to produce an 'image' of the mass in the foreground cluster of galaxies. Most of the mass in clusters of galaxies is dark matter, not the stars, gas and dust associated with the galaxies in the cluster. This dark matter outweighs ordinary matter by at least 10:1, and GRAVITATIONAL LENSING studies of the dark matter distribution around clusters promise a deeper understanding of the nature of dark matter and the scenario of structure formation.

The most direct method currently available to determine the entire mass distribution of a galaxy cluster is gravitational lensing. How does one get from the observed distorted forest of background galaxies to a map of the mass responsible for the distortion? The optimal method of inverting the observed galaxy distortions to obtain the projected mass map depends on the number of observational constraints. In turn, the resolution of the mass map will be proportional to the density of constraints on the plane of the sky. Strong lensing (multiple resolved images of a single source) can offer high mass resolution near the source image positions, but the resolution degrades far from these positions. Weak lensing (single distorted images of each resolved source) is ubiquitous since there are nearly a million background galaxies per square degree anywhere on the sky, and can yield moderate resolution maps of projected mass over wide areas. Figure 1 shows a good example of gravitational lensing (both strong and weak) by a cluster of galaxies.

The properties of clusters are functions both of the dynamical processes operating within them and of the underlying cosmology. While the present epoch mass function gives the parameters of the mass fluctuation spectrum on cluster scales, changes measured at higher REDSHIFT depend on cosmological parameters. Previous work has used the temperature (or equivalently the cluster velocity dispersion) distribution as a proxy for the mass. The average matter density of the universe Ω_m has been estimated by extrapolating observations of cluster luminosities to the rest of the universe and assuming the mass-to-luminosity ratio is the same for clusters and the universe as a whole. If we could explore mass structure and pin down Ω_m via direct observations of mass, this would significantly clarify cosmology. Recently, the first steps have been taken toward addressing three fundamental questions: what is the total mass of clusters, how is the mass distributed within clusters, and how does the cluster's mass evolve over time? Comparisons of the cluster mass spectrum, particularly its evolution over cosmic time, with N-body simulations of dark matter structure formation for different cosmological models can in principle be used to distinguish among these models.

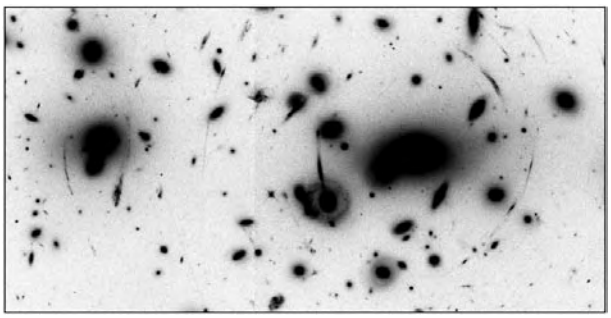


Figure 1. A Hubble Space Telescope image of the cluster of galaxies A2218, showing multiple 'arcs': lens-distorted images of background galaxies. The mass of the dark matter in the cluster is far greater than the mass of all the stars. This mass bends light rays from distant galaxies, causing a tangential 'stretching' of the images of the background galaxies. Courtesy W Couch (University of New South Wales), R Ellis (Cambridge University), and NASA.

Background sources

Observations of distant GALAXIES AT HIGH REDSHIFT suggest that we are seeing these galaxies at the epoch of formation of much of their stellar content. The resulting UV-bright spectrum, when redshifted to redshifts of 1–3, would produce the observed blue spectral shape. This backdrop of distant galaxies may be used as a tool for studying foreground mass concentrations. Their unusual spectral shape makes it particularly easy to distinguish them from red cluster galaxies which have been swept of most of their gas and thus are not undergoing STAR FORMATION. Light from the forest of distant galaxies passes by and through a foreground cluster of galaxies. For small impact parameters, close to the cluster center, gravitational lensing of a chance projection of one of these distant galaxies is strong enough to produce multiple images (strong lensing). Farther out, lensing distorts the shape of all distant galaxies (weak lensing).

Gravitational lensing studies of clusters of galaxies present an opportunity to learn about the properties of dark matter via its clustering properties on a variety of spatial scales. Of interest is the mass profile near the center (from strong lensing observations) to probe dark matter at peak densities, the mass profile at large distance from the cluster to probe the nature of structure formation (from weak-lensing observations), the correlation with luminosity, and mass segregation between galaxies and the totally dark component. Using mosaics of CCDs, wide-field weak-lensing studies can now map the mass out towards the linear regime beyond 2 Mpc radius. Both observations and N-body cold dark matter (CDM) simulations show a steep mass drop-off at large radii in clusters.

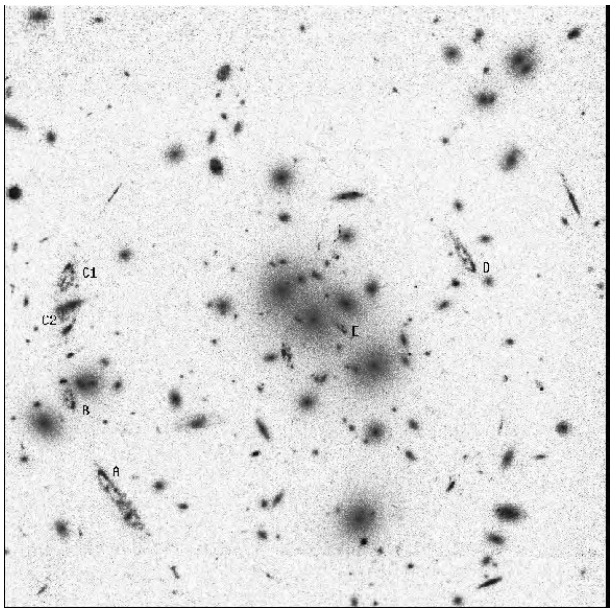
Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Strong lensing

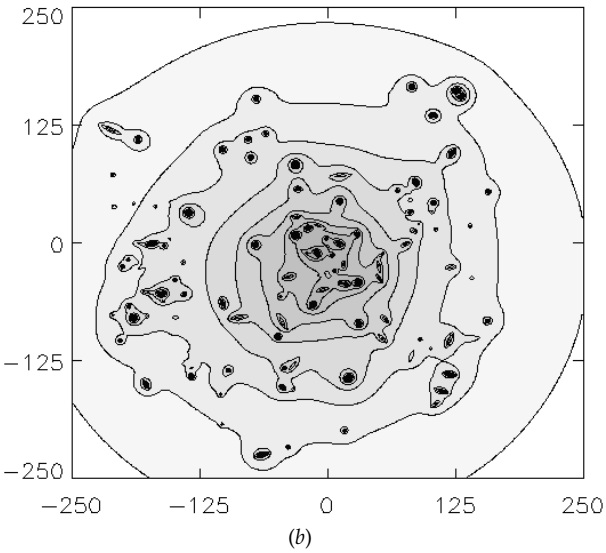
A single background galaxy, if placed by chance directly behind the cluster lens, is heavily distorted into one or more long arcs concentric with the mass centroid of the cluster. This type of strong lensing forms the basis of the highest resolution mass maps. In cases where multiple images of a source are created by the lens, the details of the position and distortion of these sub-images are highly sensitive to the projected 2-dimensional mass distribution within the lens. Parametric models of the lens mass can then be used to first unlens the sub-images to get an image of the source, and then ray trace light from this source past the lens in an iterative fit in the image plane for the lens mass parameters.

As an example, figure 2 shows a high-resolution mass map of the cluster 0024 + 1654, some three billion light-years distant, based on parametric inversion of the associated gravitational lens. This lens creates eight well resolved whole or partial sub-images of a background galaxy, seen in deep imaging with the HUBBLE SPACE TELESCOPE. Excluding mass concentrations centered on visible galaxies, more than 98% of the remaining mass is represented by a smooth concentration of dark matter centered near the brightest cluster galaxies, with a 50 kpc soft core (throughout we use Hubble constant $H_0 = 70$ km s^{-1} Mpc^{-1}). The dark matter distribution we observe in CL0024 is far more smooth, symmetric, and nonsingular than in typical simulated clusters using the CDM model. Inside 160 kpc radius, the rest-frame mass-to-light ratio is 170 times the solar value, rising with radius. Because galaxies were brighter in the past, this translates to a current mass-to-light ratio of 280. For scale, we would need around 1400 for an average cluster to extrapolate to a closed universe.

The mass core radius is smaller than most observed x-ray core radii in nearby clusters, suggesting that the x-ray gas may be less relaxed dynamically than the dark matter. Recent high resolution N-body simulations of dark matter clustering develop mass cores much smaller than the observed mass cores. Perhaps there are other physical mechanisms at work in the cores of clusters, beyond gravitational instability and merging of dark matter halos, which are capable of damping the buildup of the mass at the center. But strong lensing events like this are rare. The large mass associated with rich clusters of galaxies distorts background galaxy images over a large area of the sky. To accurately measure this effect, deep multi-wavelength imaging to 29th magnitude over a wide field must be obtained.



(*a*)



(*b*)

Figure 2. (*a*) A Hubble Space Telescope image of the redshift 0.4 cluster CL0024+1654. Multiple images (A–E) of a $z = 1.6$ background galaxy may be seen. **This figure is reproduced as Color Plate 17.** (*b*) A contour map of the projected mass density in CL0024+1654, obtained by fitting the observed distorted images of the source galaxy. The concentrated dark matter halos of individual cluster galaxies can be seen. However, most of the mass in the cluster is in a smooth non-singular distribution.

Statistical weak lensing

Foreground GALAXY CLUSTERS at redshifts 0.2–0.5 with radial velocity dispersions above 700 km s^{-1} have sufficient mass density to significantly distort background galaxies of redshift greater than 0.4–1. Lensing preserves the surface brightness and spectrum of the source, so that most lensed sources have the very faint surface brightness and blue color of the faint blue galaxies.

Since there are roughly 100 independent resolved sources per square arcminute, this enables a different kind of solution to the inverse problem, called 'statistical weak lensing'. This method has the advantage that it is generally applicable, and gives good signal-to-noise inversions in regions of low shear which are much more representative of random regions. In statistical weak

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

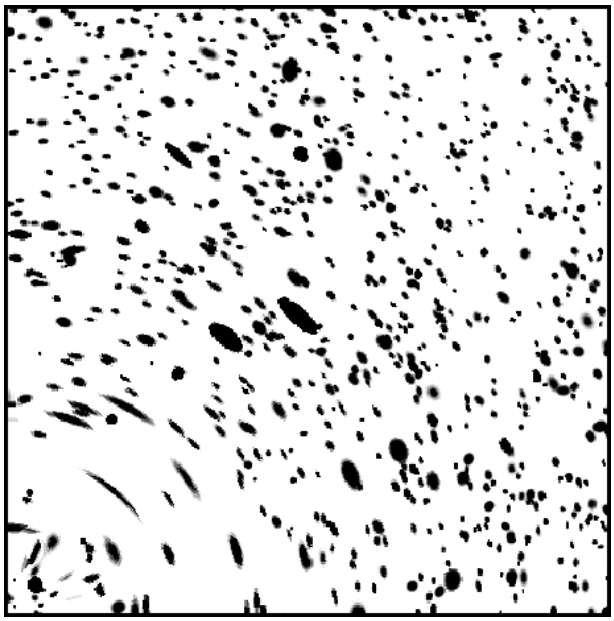


Figure 3. A simulated field, several arcminutes wide, looking through the forest of galaxies out to redshift 3. The effects of a dark matter concentration of the type found in a massive cluster centered at lower left are shown.

lensing it is necessary to average over the apparent orientation of many sources for each mass resolution element; this is why weak-lens statistical inversion uses tens of thousands of distorted background galaxies over a wide field. The largest source of noise in weak-lens inversion of deep optical images is the ellipticity noise of the source population itself; galaxies are not round and they come in all orientations. A number of non-parametric, regularized algorithms have been developed for inverting the distortion field represented by the observed distorted source galaxy images. A simulation of both weak and strong lensing is shown in figure 3. First, a long cone is populated with galaxies of appropriate size and luminosity, out to redshift 3. Then a cluster mass is placed at redshift 0.3, distorting the images of these background galaxies.

To construct a map of the gravitational lens projected mass distribution, we need to estimate the distortion at each point in the field. Start with the (r, θ) principal-axis transformed second moments of the intensity I (the tangential and radial components of the major axis a and minor axis b) for each background galaxy image g about some point on the sky (x, y). The excess tangential alignment about that point is then given by the distortion statistic

$$T_g(x, y) \;=\; (a^2 - b^2)/(a^2 + b^2)$$

For example, about position (0,0) in figure 3 this tangential alignment is high. In the weak-lensing limit it can be shown that the integral of the tangential alignment T around a circle of radius r about the point (x_0, y_0) is a measure of the mass contrast interior:

$$\int_c T(\vec{r}, x_0, y_0)\, \mathrm{d}^2 r \;=\; 2\,[\Sigma_{\mathrm{av}}(< r) - \Sigma(r)]\;\Sigma_c^{-1},$$

where Σ_c is the critical mass density for lensing the source onto the observer. A typical value for Σ_c for lenses at moderate redshift is 1 g cm^{-2}. To construct an image of the gravitational lens projected mass distribution, the distortion statistic $T(x, y)$ may be computed over a grid of positions as candidate lens centers. The average projected mass density at point (x, y) on the sky, with resolution rmin, is given by

$$\bar{\Sigma}(x, y) \;=\; \Sigma_0 \int \frac{G(r)}{r^2} T(\vec{r}, x, y)\, \mathrm{d}^2 r \;+\; \bar{\Sigma}_{\mathrm{outer}},$$

where Σ_0(rmin, rmax, zs, zl) is a function of the source and lens redshifts and is calibrated with strong lensing cluster test masses, G(r, rmin, rmax) is an apodizing filter which smoothly cuts on at rmin and off at rmax, and $\bar{\Sigma}_{\mathrm{outer}}$(rmin, rmax) is the average density in the annulus between rmin and rmax. $\bar{\Sigma}(x, y)$ is a direct map of the projected mass density.

In practice, a catalog of faint blue galaxy positions, total magnitudes, and second moments of flux is passed to the lens arclet inversion software. Currently there are more than 30 clusters with weak-lensing mass estimates. The center of (lensing) mass corresponds well to the center of optical light. The lens projected mass overdensity (see second equation above) as a function of radius from the centroid of the mass distribution may be compared to the lens cluster light (baryonic) distribution. The smoothed optical light distribution traces the (lensing) mass well. Despite earlier indications that the x-ray derived mass disagreed with the lensing mass inside 200–300 kpc, recent measurements of both quantities out to 1 Mpc radius show that the two methods are in closer agreement inside this larger radius. This suggests that the assumptions of pressure support and isothermality for the x-ray gas break down inside $\sim$ 300 kpc. This has been resolved, in large part due to wide-field imaging of cluster lenses. Until recently, observations of lensing, x-rays, and galaxy velocities (i.e. virial mass) were often made on very different physical scales and this appears to have contributed to the confusion in the literature as to which of the three independent mass estimates for a given cluster was correct. As multiband deep imaging becomes more common in lensing work, color-redshifts will enable weak-lens studies of clusters at high redshift.

Mass and light in typical clusters

Many clusters studied in this way have been chosen for their high x-ray luminosity. This may bias the statistics of mass in clusters. A recent weak-lensing survey studied ten x-ray selected 'normal' clusters of galaxies with a wide-field CCD mosaic camera. Mass and light in a sample of clusters of galaxies of typical x-ray luminosity at redshifts ($z = 0.15$–0.25) were studied. These clusters

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are more representative of cluster mass distribution than the rare x-ray luminous rich clusters studied so far. Out to 3 Mpc radius, mass generally traces light. Some of the cluster mass profiles are flatter than isothermal in a central region, with downturns at radii beyond 2 Mpc. Outside the core, these profiles are well-fit by N-body cold dark matter clustering simulations. Mass-to-light ratios at red wavelengths in the cluster rest frame scatter around 350, with no systematic trend with mass. Scaling the average M/L by the luminosity density leads to an estimate for the mean mass density of $\Omega_{\rm m} = 0.2$. But there may be low mass density regions in the huge volume outside clusters. Taking all the clusters observed to date (both high and low x-ray luminosity), lensing-inferred mass to light ratios bracket a broad range, from $200h$ to $800h$ (in solar units), with most of the clusters being in the low to middle part of this range (Here, $h = H_0/100$.)

Beyond clusters

Weak lensing is a tool which may be applied anywhere on the sky. One may even search for mass overdensities comparable to clusters without first knowing a cluster is there. Some clusters have already been found via their mass in this way. More importantly for our understanding of cosmology, weak lensing may be used to directly investigate the evolution of mass structures over cosmic time, and surveys are currently under way. Figure 4 shows a mass map of a 0.8 degree wide field, revealing several mass concentrations. If the shear pattern of sufficiently distant galaxies is measured over a wide field, then intervening mass overdensities over a wide range of redshift are revealed.

In different cosmologies mass structure develops differently; this leads to a weak-lens cosmological diagnostic. Generally, the net effect of all intervening mass overdensities is a correlation of orientations of faint galaxies separated by some angle on the sky. A direct probe of the development of mass structures, this 'cosmic shear' has recently been detected. Meaningful tests of cosmological models will result from the comparison of what is found for this cosmic shear (as well as counts of discrete clusters found via weak lensing) with theoretical models for the evolution of large scale structure. Comparing mass maps from deep weak-lensing observations with data from upcoming satellite COSMIC MICROWAVE BACKGROUND anisotropy missions will lead to a precision test of the nature of cosmic mass-energy, probing the basic theory of structure formation in our universe.

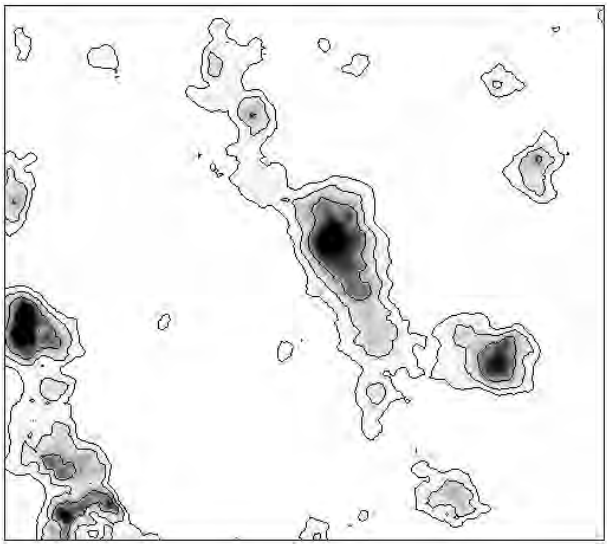


Figure 4. The ellipticities of over 50 000 faint galaxies can be inverted to reconstruct a map of the intervening mass distribution. Several projected mass concentrations are seen in this weak-lensing mass map covering 0.8 degrees of the sky. One of these structures spans 6 Mpc, twelve times the size of the map shown in figure 2. At the center of the mass structure is a cluster of galaxies.

Bibliography

Blandford R D and Narayan R 1992 *Ann. Rev. Astron. Astrophys.* 311

Kaiser N 1998 *Astrophys. J.* **498** 26

Mellier Y 1999 *Ann. Rev. Astron. Astrophys.* **37** 127

Tyson J A, Kochanski G and Dell'Antonio I P 1998 *Astrophys. J.* **498** 107

J A Tyson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitational Radiation

Gravity is one of the fundamental forces of nature, and it is the dominant force in most astronomical systems. In common with all other phenomena, gravity must obey the principles of SPECIAL RELATIVITY. In particular, gravitational forces must not be transmitted or communicated faster than light. This means that, when the gravitational field of an object changes, the changes ripple outwards through space and take a finite time to reach other objects. These ripples are called gravitational radiation or gravitational waves[1].

In Einstein's theory of gravitation (see GENERAL RELATIVITY AND GRAVITATION), as in many other modern theories of gravity, gravitational waves travel at exactly the speed of light. Different theories make different predictions, however, about details, such as their strength and polarization. There is strong indirect observational evidence (see GENERAL RELATIVITY TESTS: BINARY PULSARS) that gravitational waves follow the predictions of general relativity, and instruments now under construction are expected to make the first direct detections of them in the first years of the 21st century.

These instruments and plans for future instruments in space are described in the article GRAVITATIONAL RADIATION DETECTORS ON EARTH AND IN SPACE. Detectors must look for gravitational radiation from astronomical systems, because it is not possible to generate detectable levels of radiation in the laboratory. It follows that gravitational wave detection is also a branch of observational astronomy.

The most striking aspect of gravitational waves is their weakness. A comparison with the energy in light will illustrate this. The human eye has no trouble sensing the light from the planet Jupiter: the amount of energy that passes through the iris of the eye is far more than the minimum the eye can detect. Yet several times a week a gravitational wave, generated in a far distant galaxy, carries a similar amount of energy into the eye, and we do not notice it.

The reason is that gravity is the weakest of the fundamental forces, and the disturbance created by even such an energetic wave is so tiny that no man-made instrument has so far registered it. While all the energy in the light from Jupiter that enters the eye is absorbed in the eye, the gravitational wave passes right through, leaving behind almost none of its energy. All the matter in the present universe is similarly transparent to gravitational waves.

Gravitational radiation is today one of the last unopened windows into the universe. There are at least five reasons motivating scientists to develop gravitational wave astronomy.

- The weakness with which gravitational waves interact with matter is a great advantage for astronomy. It means that gravitational waves arrive unaffected by any intervening matter they may have encountered since being generated. There is no significant scattering or absorption, although they will be deflected by GRAVITATIONAL LENSING in the same way as light. Gravitational waves carry uncorrupted information even if they come from the most distant parts of the universe or from its most hidden regions, like the interiors of SUPERNOVAE.
- Gravitational waves are emitted by the bulk motions of their sources, not by individual atoms or electrons, as is normally the case for electromagnetic waves. They therefore carry a completely different kind of information about their sources from that which is normally available in observations of binary stars (see BINARY STARS: OVERVIEW), supernovae and NEUTRON STARS.
- Gravitational waves can be emitted by black holes, which are described in the article GRAVITATIONAL RADIATION DETECTORS ON EARTH AND IN SPACE. Indeed, gravitational waves provide the only way to make direct observations of these objects. Since there is now strong indirect evidence that giant black holes inhabit the centers of many (or even most) galaxies (see SUPERMASSIVE BLACK HOLES IN AGN), and since smaller ones are common in the Galaxy (see BLACK HOLE CANDIDATES IN X-RAY BINARIES), there is great interest in making direct observations of them.
- Gravitational waves can come from extraordinarily early in the history of the universe. The electromagnetic radiation from the big bang is called the COSMIC MICROWAVE BACKGROUND. Observations of it describe the universe at it was about 10^5 yr after the big bang. Studies of cosmological NUCLEOSYNTHESIS give information about what the universe was like as little as 3 min after the big bang. Gravitational waves, if they can be detected, would picture the universe when it was only perhaps 10^{-24} s old, just at the end of INFLATION.
- Gravitational radiation is the last fundamental prediction of Einstein's general relativity that has not yet been directly verified. If another theory of gravity is correct, then differences could in principle show up in the properties of gravitational waves, such as their polarization. In principle, there must be a better theory of gravity, since general relativity is not a quantum theory, a deficiency that theoretical physicists today are working hard to remedy. The majority belief today is that there should be a unified theory of the fundamental forces, in which gravitation is related to the other forces. Evidence for the nature of this relation could show up in observations of gravitational waves, particularly those from the big bang.

These motivations and their implications are developed in the following sections. Each section begins with an introduction to the physical ideas and then develops some of the mathematical details.

[1] They are also sometimes referred to as *gravity waves*, but since this term has a different meaning in meteorology and stellar hydrodynamics, we will avoid it here. See SOLAR INTERIOR: HELIOSEISMOLOGY.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The physics of gravitational radiation: weakness and strength

The starting point for understanding gravitational radiation is Newtonian gravity. The weakness of gravity is evident. If a child picks up a book, she defeats the cumulative gravitational pull of the entire planet Earth on the book. The strength to do this comes from the chemical forces in her muscles, which come from electromagnetic interactions.

In fact, the electromagnetic force between the electron and the proton in a hydrogen atom is 2×10^{39} times bigger than the gravitational force between them. The reason that gravity can nevertheless dominate on the cosmic scale is that opposite electrical charges cancel each other, while the gravitational forces of all the particles add.

Another fact about gravity that was known to Newton is what is now called the equivalence principle. This is the principle that all bodies accelerate in the same way in a gravitational field, so that the trajectory that a freely falling body (a body influenced only by gravity) follows in a given gravitational field depends only on its starting position and velocity, not on what it is made of.

Imagine now a machine made in some way to detect gravitational waves. Whatever the method of detection, a wave needs somehow to alter the internal state of the detector. If the wave carries a gravitational field that is completely uniform across the detector, then by the equivalence principle all of the parts of the detector will accelerate together, and its state will not change at all. To detect a gravitational wave, the machine must measure the non-uniformities of the gravitational field across a detector.

These non-uniformities are called tidal forces, because they produce the stretching effects that raise tides on the Earth. Gravitational waves are traveling tidal forces.

Newton's theory of gravity had no gravitational waves. For Newton, if a gravitational field changed in some way, that change took place instantaneously everywhere in space. This is not a wave. Let us consider what we mean by the term 'wave' in ordinary language.

Imagine a child's rubber duck floating in a bath tub half full of water. If a child presses down on the duck very gently, until is is nearly submerged, then the level of the water will rise everywhere in a nearly uniform way, and this is not called a wave. If instead he drops the duck, then the disturbance rises around the base of the duck rapidly, moves away from it and eventually reaches the walls. This is a wave. Wave motion requires a finite speed for the propagation of disturbances. If the disturbance is very slow, as for the floating duck, then the wavelength is very long, and near the site of the disturbance the wave motion is not noticeable. We say we are in the 'near zone'. However, when we are more than a wavelength away, then we see waves, and this is the 'wave zone' or 'far zone'. For the dropped duck, we see the waves because their wavelengths are shorter than the size of the bath. In general relativity, the speed of gravity is the speed of light. Because of this finite speed, gravity must exhibit wave effects.

Many of Newton's contemporaries were unhappy that his theory of gravity was based on instantaneous 'action at a distance', but Newton's theory fitted the observational facts. If gravity had a finite speed of propagation, there was no evidence for it in the solar system. Interestingly, the brilliant 18th century French mathematician and physicist Laplace tried out a variation on Newton's theory in which gravity was represented by something 'flowing' out of its source with a finite speed. He reasoned that a planet like the Earth, moving through this fluid of gravity, would experience friction and gradually spiral in towards the Sun.

Laplace could show that the observational limits on this in-spiral even in his day were so stringent that the speed of gravity in his model needed to be huge compared with the speed of light. He did not find this result attractive and took the theory no further. (Laplace also explored the notion of what we now call a black hole, which for him was a region where gravity was strong enough to trap light.)

It is interesting that, today, observations of the two neutron stars in the binary system PSR1913 + 16 spiraling together as they orbit one another provide the most convincing evidence that gravitational waves exist and are as described in general relativity (see below). Laplace had the right effect, but the wrong theory. This evidence is described in the next section.

In general relativity, Einstein used the principle of equivalence as the basis for a geometrical description of gravity. In the four-dimensional world of space–time, the trajectory of a particle falling freely in a gravitational field is a certain fixed curve. Its direction at any point depends on the velocity of the particle. The equivalence principle implies that there is a preferred set of curves in space–time: at any point, pick any direction, and there is a unique curve in that direction that will be the trajectory of any particle starting with that velocity. These trajectories are thus properties of space–time itself.

Moreover, if there were no gravitational field, the trajectories would be simple straight lines. Even in a gravitational field, a small freely falling particle does not 'feel' any acceleration: its internal state is the same as if there were no gravity. Therefore Einstein postulated that a gravitational field made space–time curved, and that the preferred trajectories were locally straight lines that simply changed direction as they moved through the curved space–time, in much the same way as a great circle on a sphere changes direction relative to other great circles as one goes along it. For weak gravitational fields of slowly moving bodies, Einstein's theory reduces to Newton's in the first approximation.

For gravitational waves, one could make a very simple detector just by monitoring the distance between two nearby freely falling particles. If they are genuinely free, then any changes in their separations would indicate the passage of a gravitational wave. Because this measures a tidal effect, the bigger the separation of the particles,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the bigger will be the change in their separation, at least for particles that are separated by less than a gravitational wavelength. Most modern gravitational wave detectors are designed to be as big as cost and practicality allows.

Although gravitational radiation is well understood in theoretical terms in general relativity, the complexity and non-linearity of Einstein's equations means that calculations are often difficult. In the historical development of general relativity, between 1915 and the 1950s and 1960s, these mathematical difficulties created confusion over the physical nature of gravitational radiation, and in particular over whether it carried energy away from the source. Improved mathematical techniques finally resolved the matter in favor of the simple physical picture presented here, but this picture would not be complete without the strong mathematical underpinning that now exists.

The question of energy in gravitational waves is still a delicate one. There is no question that waves carry energy (and momentum) away from their sources. Nevertheless, it is not possible in general relativity to localize the energy in the radiation to regions smaller than about a wavelength. Indeed the equivalence principle shows that 'point' particles feel nothing, no matter how strong the wave. The wave only acts by stretching space–time, producing a tidal distortion in the separations between particles (see the discussion of polarization below).

For this reason, energy is localized only in regions, not at points. It is nevertheless real energy: the non-linearity of general relativity allows waves to create gravitation themselves. Recent numerical simulations have shown that focussed gravitational waves can actually form black holes, trapping themselves. If the waves are weak, they enter the focussing region and re-emerge. If they are strong enough, they enter and never leave.

Gravitational waves in a quasi-Newtonian model

It is possible to calculate the approximate size of the effect of a given gravitational wave by beginning with Newtonian gravity and adding waves to it. In Newtonian gravity the gravitational field produced by a mass M at a distance r is given by

$$\phi = -GM/r \tag{1}$$

where G is Newton's gravitational constant. The field of a gravitational wave must be a ripple on this, which means a small change that oscillates in space and time. A suitable form for a change that propagates at the speed of light in the z-direction with an angular frequency ω is:

$$\delta\phi = -\epsilon\frac{GM}{r}\sin[\omega(z/c - t)] \tag{2}$$

where ϵ is a dimensionless number that would be expected to be small compared with 1. Its size is the subject of the next main section.

The field $\delta\phi$ produces an acceleration in the z-direction that depends on its z-derivative. Both $1/r$ and the sin() term depend on z. The derivative of $1/r$ will be proportional to $1/r^2$, which is how the acceleration falls off in Newton's theory (where ϕ is the only field). However, the derivative of the sin() term does not change the $1/r$; rather, it essentially just multiplies $\delta\phi$ by ω/c. At sufficiently large distances from the source, this term will dominate the $1/r^2$ term and the acceleration produced by the wave will be

$$\delta a_z = -\epsilon\omega\frac{GM}{rc}\cos[\omega(z/c - t)]. \tag{3}$$

Note that this term would not be present in the x- and y-derivatives, so these components of the acceleration are much smaller in this quasi-Newtonian model of a gravitational wave.

Effect on a simple detector

The tidal part of this acceleration, for a detector that has size ℓ in the z-direction, is to a first approximation

$$\ell\frac{\mathrm{d}}{\mathrm{d}z}\delta a_z = \epsilon\omega^2\ell\frac{GM}{rc^2}\sin[\omega(z/c - t)]. \tag{4}$$

If a detector consists of two freely falling particles with this relative acceleration, the equation of motion for their separation ℓ will be

$$\frac{\mathrm{d}^2\ell}{\mathrm{d}t^2}2 = \epsilon\omega^2\ell\frac{GM}{rc^2}\sin[\omega(z/c - t)]. \tag{5}$$

The dimensionless coefficient $\epsilon GM/rc^2$ is typically very small. Even if ϵ is of order 1, the other number is, with reasonable values for M and R,

$$\frac{GM}{rc^2} = 2.4\times10^{-21}\frac{M}{M_\odot}\left(\frac{r}{20\ \mathrm{Mpc}}\right)^{-1}.$$

The mass and distance scales here are those appropriate to a neutron star in the nearest large cluster of galaxies, the Virgo Cluster. (The distance unit is based on the astronomers' parsec, denoted pc, which is about 3×10^{16} m. The unit Mpc is a megaparsec.) It is believed that several neutron stars are formed in the Virgo Cluster each year in supernova explosions. Ever since the beginning of the development of gravitational wave detectors such events have been high on the list of possible sources of gravitational waves.

To solve equation (5), one takes r and z as constants and uses the smallness of the right-hand side, which implies that the changes in ℓ are tiny compared with ℓ itself. On the right-hand side one can therefore replace ℓ by ℓ_0, the initial value of ℓ, and then simply integrate twice in time to obtain (for an initial value $\mathrm{d}\ell/\mathrm{d}t = 0$)

$$\frac{\ell(t) - \ell_0}{\ell_0} = -\epsilon\frac{GM}{rc^2}\sin[\omega(z/c - t)]. \tag{6}$$

The right-hand side of equation (6) is identical to that of equation (2). This is an important conclusion which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

fits neatly with Einstein's geometrical conception of gravity: the size of a gravitational wave gives directly the stretching of the distance between nearby free particles. It is conventional to call this $h/2$ and refer to h as the gravitational wave potential:

$$h := 2\frac{\ell(t) - \ell_0}{\ell_0} = 2\delta\phi. \tag{7}$$

The amplitude of the oscillations of h is

$$h \sim 2\epsilon\frac{GM}{rc^2}. \tag{8}$$

It is evident from this that a detector must be able to measure changes in its own size that are smaller than 1 part in 10^{21} to have a reasonable chance of making astronomical observations. The extraordinary smallness of this effect also explains why ordinary objects in the universe are transparent to gravitational waves. As the waves pass through them, they disturb them so little (parts per 10^{21} typically) that the transfer of energy to the object and any back-reaction effects of this on the wave are negligibly small.

Energy flux carried by waves

The *energy* in the waves can also be estimated from these equations and general physical principles. Quite generally, in classical field theories, the energy flux of a propagating sinusoidal plane wave is proportional to the square of the time-derivative of the fundamental field. In electromagnetism, the Poynting flux is proportional to the square of the time-derivative of the vector potential.

In general relativity, the flux is therefore proportional to the square of the time-derivative of $h(t)$. The proportionality constant must be built only out of c, G and pure numbers. To obtain the right units, it must be proportional to c^3/G: to obtain the pure number, a calculation in general relativity is required; $1/32\pi$ for a linearly polarized wave. (Polarization is described later.) This gives

$$F_{\rm gw} = \frac{1}{32\pi}\frac{c^3}{G}\left(\frac{{\rm d}h}{{\rm d}t}\right)^2$$

$$= 1.6\times10^{-5}\left(\frac{f}{100\ {\rm Hz}}\right)^2\left(\frac{h}{10^{-22}}\right)^2\ {\rm W\ m^{-2}} \tag{9}$$

for a wave with frequency $f = \omega/2\pi$. For comparison, reflected sunlight from Jupiter has a flux on Earth of 2.3×10^{-7} W m^{-2}, almost 100 times smaller than that of a gravitational wave with an amplitude of 10^{-22}.

Deficiencies of the quasi-Newtonian model

The calculation and equations in this section have been framed within a modified Newtonian model of gravity with a propagation speed of c, and one would expect some differences from general relativity. The most important difference is in the direction in which the tidal forces act. In the simple model, wave accelerations act in the z-direction, which was the direction of propagation of the wave. This is called a longitudinal wave.

In general relativity, gravitational waves are transverse waves: if the wave propagates in the z-direction then the tidal forces act only in the x–y-plane. We will discuss later the exact form that their action in this plane takes. Remarkably, the rest of the formulas above are good approximations even in general relativity, provided that ϵ is calculated correctly, as described in the next section.

The emission of gravitational waves

The previous section described the propagation of gravitational waves, their interaction with detectors and the energy they carry. This section deals with the strength with which waves are emitted by astronomical bodies.

In Newtonian gravity there is a fundamental theorem, proved by Newton, that the gravitational field outside a spherical body is not only spherical but the same as that of a point mass located at the origin of the body. It has the form given in equation (1). In particular, the field is independent of the size of the body, as long as we consider only points outside it. This is true even if the star pulsates in a spherical manner.

This theorem is essentially the same in general relativity, and is known as Birkhoff's theorem. Outside a spherical body the field is the same as that of a black hole of the same mass as the body (the Schwarzschild metric), even if the body is pulsating spherically. However, if the pulsation is non-spherical, then the outside field will change. In general relativity the changes generally propagate as a wave. So gravitational waves will be emitted by non-spherical motions.

In general the calculation of the emitted waves is extremely difficult, since the field equations of general relativity are a system of many coupled, non-linear, partial differential equations. However, in four circumstances the emission mechanisms are understood in some detail:

- Small-amplitude pulsations of relativistic stars and black holes. Normally gravitational radiation carries away energy and damps pulsations away, but in rotating stars the opposite may happen: the radiated loss of angular momentum may allow the star to spin down to an energetically more favored state, in which case the perturbation will grow, at least until non-linear effects intervene. Discovered by S Chandrasekhar and now called the Chandrasekhar–Friedman–Schutz (CFS) instability, it is thought to limit the rotation speed of *young* neutron stars (see below). Black holes also emit gravitational radiation when they are disturbed, *e.g.* by something falling into them, but they are not unstable: they always settle down into a steady state again.
- Radiation from 'test' objects orbiting black holes. If the mass of the object is small enough then the total gravitational field may be treated as a linear perturbation of the exactly known field of a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

black hole (the Schwarzschild or, with rotation, the Kerr solution). These studies give insight into the general problem of gravitational radiation, and they also predict gravitational waveforms that might be observed by space-based observatories looking at compact stars falling into the giant black holes in the centers of galaxies (see below).

- Weak gravitational fields and slow motion. Such weakly relativistic sources are studied in the *post-Newtonian* approximation, which includes higher-order corrections to Newtonian gravity from general relativity. This is analogous to the slow-motion multipole approximation that is so powerful in the study of electromagnetic radiation. Most realistic gravitational-wave sources can be studied to some approximation this way.
- Collisions of black holes and neutron stars. These events, which are expected to be observed by gravitational wave detectors (see below), must be modeled by solving the full set of Einstein equations on a powerful computer. Techniques to do this are advancing rapidly, and simulations of realistic mergers of stars and black holes from in-spiraling orbits can be expected to yield useful results in the first years of the 21st century.

Quadrupole approximation

The post-Newtonian approximation has so far been the most powerful of these methods, and it yields the most insight into the emission mechanisms. Its fundamental result is the quadrupole formula, which gives the first approximation to the radiation emitted by a weakly relativistic system.

The quadrupole formula is analogous to the dipole formula of electromagnetism. In this language, monopole means spherical, which emits no radiation. This is also true in electromagnetism, where it is linked to conservation of charge. The 'monopole moment' in electromagnetism is the total charge of a system, and since that does not change, there can be no spherical radiation.

Again in electromagnetism, the dipole moment is defined as the integral

$$d_i = \int \rho x_i \, \mathrm{d}^3 x$$

where ρ is the charge density and x_i is a Cartesian coordinate. If this integral is time dependent, then the amplitude of the electromagnetic waves will be proportional to its first time-derivative $\mathrm{d}d_i/\mathrm{d}t$, and the radiated energy will be proportional (as we remarked earlier) to the square of the time derivative of this amplitude, *i.e.* to $\sum_i |\mathrm{d}^2 d_i/\mathrm{d}t^2|^2$.

In the post-Newtonian approximation to general relativity, the calculation goes remarkably similarly. The monopole moment is now the total mass–energy, which is the dominant source of the gravitational field for non-relativistic bodies, and which is constant as long as the radiation is weak. (Radiation will carry away energy, but in the post-Newtonian approximation that is a higher-order effect.) The dipole moment is given by the same equation as above, but with ρ interpreted as the density of mass–energy.

However, here general relativity departs from electromagnetism. The time-derivative of the dipole moment is, since the mass–energy is conserved, just the integral of the velocity v_i:

$$\dot{d}_i = \int \rho v_i \, \mathrm{d}^3 x. \tag{10}$$

However, this is the total momentum in the system, and (to lowest order) this is constant. Therefore, there is no energy radiated as a result of dipole effects in general relativity. The gravitational field far from the source does contain a dipole piece if $\dot{d}_i$ is non-zero, but this is constant because it reflects the fact that the source has non-zero total momentum and is therefore moving through space.

To find genuine radiation in general relativity one must go one step beyond the dipole approximation to the quadrupole terms. These are also studied in electromagnetism, and the analogy with relativity again is close. The fundamental quantity is the spatial tensor (matrix) Q_{jk}, the second moment of the mass (or charge) distribution:

$$Q_{jk} = \int \rho x_j x_k \, \mathrm{d}^3 x. \tag{11}$$

A gravitational wave in general relativity is represented by a matrix h_{jk} rather than a single scalar h, and its source (in the quadrupole approximation) is Q_{jk}.

As in electromagnetism, the amplitude of the radiation is proportional to the second time-derivative of Q_{jk}, and it falls off inversely with the distance r from the source. A factor of G/c^4 is needed in order to obtain a dimensionless amplitude h, and a factor of 2 to be consistent with the definition in equation (8). The result for h_{jk} is

$$h_{jk} = \frac{2G}{rc^4} \frac{\mathrm{d}^2 Q_{jk}}{\mathrm{d}t^2}. \tag{12}$$

General relativity describes waves with a matrix because gravity is geometry, and the effects of gravity are represented by the stretching of space–time. This matrix contains that distortion information. Here is the information about the transverse action of the waves that the quasi-Newtonian model of the last section did not get right.

Simple estimates

If the motion inside the source is highly non-spherical, then a typical component of $\mathrm{d}^2 Q_{jk}/\mathrm{d}t^2$ will (from equation (11)) have magnitude $M v_{\mathrm{NS}}^2$, where v_{NS}^2 is the non-spherical part of the squared velocity inside the source. So one way of approximating any component of equation (12) is

$$h \sim \frac{2GMv_{\mathrm{NS}}^2}{rc^4}. \tag{13}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Comparing this with equation (8) we see that the ratio ϵ of the wave to the Newtonian potential is simply

$$\epsilon \sim \frac{v_{\rm NS}^2}{c^2}.$$

By the virial theorem for self-gravitating bodies, this will not be larger than

$$\epsilon < \phi_{\rm int}/c^2 \tag{14}$$

where $\phi_{\rm int}$ is the maximum value of the Newtonian gravitational potential *inside* the system. This provides a convenient bound in practice. It should not be taken to be more accurate than that.

For a neutron star source one has $\phi_{\rm int} \sim 0.2c^2$. If the star is in the Virgo cluster, then the upper limit on the amplitude of the radiation from such a source is 5×10^{-22}. This has been the goal of detector development for decades, to make detectors that can observe waves at or below an amplitude of 10^{-21}.

Polarization of gravitational waves

The matrix nature of the wave amplitude comes from general relativity and has no Newtonian analog. In order to find the effect of the waves on the separation of two free particles (the idealized detector), one has to start with h_{jk} as given by equation (12) or by any other calculation, and then do three things.

1. Project the matrix h_{jk} onto a plane perpendicular to the direction of travel of the wave. In the simple case considered above, where the wave was traveling in the z-direction, this means leaving the components $\{h_{xx}, h_{xy}, h_{yy}\}$ alone and setting the remaining components to zero. It is then a two-dimensional matrix in the transverse plane.
2. Remove the two-dimensional trace of the projected matrix. Call the resulting matrix $h_{jk}^{\rm TT}$, where TT stands for transverse–traceless. In the example this means subtracting $(h_{xx}+h_{yy})/2$ from both h_{xx} and h_{yy}. Then there are only two independent components left, $h_{xy}^{\rm TT} = h_{yx}^{\rm TT}$ and $h_{xx}^{\rm TT} = -h_{yy}^{\rm TT}$.
3. To find the change in the separation of two particles that have an initial separation given by the vector ℓ_k, let the matrix $h_{jk}^{\rm TT}$ act on it:

$$\delta\ell_j = \sum_k h_{jk}^{\rm TT}\ell_k. \tag{15}$$

It is clear that any longitudinal component of the separation ℓ_j between the particles is unaffected by the wave (in the example, this is the z-separation), and that there are two degrees of freedom (the two independent components of $h_{jk}^{\rm TT}$) to move particles in the plane perpendicular to the propagation direction. These two degrees of freedom are the two polarizations of the wave.

Figure 1 shows the conventional definition of the two independent polarizations, from which any other can be made by superposition. What is shown is the effect of a wave on a ring of free particles in a plane transverse to the wave. The first line shows a wave with $h_{xy} = 0$, conventionally called the '+' polarization. The bottom line shows a wave with $h_{xx} = 0$, the '×' polarization.

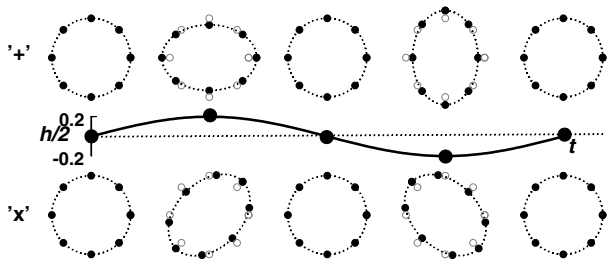


Figure 1. Polarization of gravitational waves. The center line gives the wave as a function of time, with an amplitude of $h = 0.2$, and the top and bottom lines show to scale the distortions produced by two polarizations with this amplitude.

Luminosity in gravitational waves

The energy carried by the gravitational wave must be proportional to the square of the time-derivative of the wave amplitude, so it will depend on the sum of the squares of the components d^3Q_{jk}/dt^3. The energy flux falls off as $1/r^2$, but when integrated over a sphere of radius r to obtain the total luminosity, the dependence on r goes away, as it should. The luminosity contains a factor G/c^5 on dimensional grounds, and a further factor of $1/5$ comes from a careful calculation in general relativity. The result is the gravitational wave luminosity in the quadrupole approximation:

$$L_{\rm gw} = \frac{G}{5c^5}\left(\sum_{j,k} \dddot{Q}_{jk}\,\dddot{Q}_{jk} - \frac{1}{3}\,\dddot{Q}^2\right) \tag{16}$$

where Q is the trace of the matrix Q_{jk}. Its squared third derivative must be subtracted in order to ensure that spherical motions do not radiate.

This equation will be used in the next section to estimate the back-reaction effect on a system that emits gravitational radiation.

Emission estimates

Until observations of gravitational waves are successfully made, one can only make intelligent guesses about most of the sources that will be seen. There are many that could be strong enough to be seen by the early detectors: binary stars, supernova explosions, neutron stars, the early universe. The estimates in this section are accurate only to within factors of order 1. The estimates correctly show how the important observables scale with the properties of the systems.

Man-made gravitational waves

One source can be ruled out: man-made gravitational radiation. Imagine creating a wave generator with the following extreme properties. It consists of two masses of 10^3 kg each (a small car) at opposite ends of a beam 10 m long. At its center the beam pivots about an axis. This centrifuge rotates 10 times per second. All the velocity is non-spherical, so v_{NS}^2 in equation (13) is about 10^5 m^2 s^{-2}. The frequency of the waves will actually be 20 Hz, since the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mass distribution of the system is periodic with a period of 0.05 s, only half the rotation period. The wavelength of the waves will therefore be 1.5×10^7 m, about the diameter of the earth. In order to detect gravitational waves, not near-zone Newtonian gravity, the detector must be at least one wavelength from the source. Then the amplitude h can be deduced from equation (13): $h \sim 5 \times 10^{-43}$. This is far too small to contemplate detecting.

Radiation from a spinning neutron star

Some likely gravitational wave sources behave like the centrifuge, only on a grander scale. Suppose a neutron star of radius R spins with a frequency f and has an irregularity, a bump of mass m on its otherwise axially symmetric shape. Then the bump will emit gravitational radiation (again at frequency $2f$ because it spins about its center of mass, so it actually has mass excesses on two sides of the star), and the non-spherical velocity will be just $v_{NS} = 2\pi Rf$. The radiation amplitude will be, from equation (13),

$$h_{\text{bump}} \sim 2(2\pi Rf/c)^2 Gm/rc^2 \tag{17}$$

and the luminosity, from equation (16) (assuming that roughly four comparable components of Q_{jk} contribute to the sum),

$$L_{\text{bump}} \sim (G/5c^5)(2\pi f)^6 m^2 R^4.$$

The radiated energy would presumably come from the rotational energy of the star. This would lead to a spindown of the star on a timescale

$$t_{\text{spindown}} = \frac{1}{2}mv^2/L_{\text{bump}} \sim \frac{5}{4\pi} f^{-1} \left(\frac{Gm}{Rc^2}\right)^{-1} \left(\frac{v}{c}\right)^{-3}.$$

It is felt that neutron star crusts are not strong enough to support asymmetries with a mass of more than about $m \sim 10^{-5} M_\odot$, and from this one can estimate the likelihood that the observed spindown timescales of PULSARS are due to gravitational radiation. In most cases, it seems that gravitational wave losses cannot be the main spindown mechanism.

However, lower levels of radiation would still be observable by detectors under construction, and this may be coming from a number of stars. In particular, there is a class of neutron stars in X-RAY BINARY STARS. They are accreting, and it is possible that accretion will create some kind of mass asymmetry (bump) or else lead to a rotational instability of the CFS type in the r-modes (see below). In either case, the stars could turn out be long-lived sources of gravitational waves.

Radiation from a binary star system

Another 'centrifuge' is a binary star system. Two stars of the same mass M in a circular orbit of radius R have $v_{\text{NS}}^2 = GM/4R$. The gravitational-wave amplitude from equation (13) can then be written

$$h_{\text{binary}} \sim \frac{1}{2}\frac{GM}{rc^2}\frac{GM}{Rc^2}. \tag{18}$$

Compare this with the implications of putting equation (14) into equation (8).

The gravitational-wave luminosity of such a system is, by a calculation analogous to that for bumps on neutron stars,

$$L_{\text{binary}} \sim \frac{1}{80}\frac{c^5}{G}\left(\frac{GM}{Rc^2}\right)^5.$$

In this equation there appears the important constant $c^5/G = 3.6\times^{52}$ W, a number with the dimensions of luminosity built only from fundamental constants. By comparison, the luminosity of the Sun is only 3.8×10^{26} W. Close binaries can therefore radiate more energy in gravitational waves than in light.

The radiation of energy by the orbital motion causes the orbit to shrink. The shrinking will make any observed gravitational waves increase in frequency with time. This is called a *chirp*. The timescale for this is

$$t_{\text{chirp}} = Mv^2/L_{\text{binary}} \sim \frac{20GM}{c^3}\left(\frac{GM}{Rc^2}\right)^{-4}. \tag{19}$$

The binary pulsar system—verifying gravitational waves

This orbital shrinking has already been observed in the HULSE–TAYLOR PULSAR system, containing the radio pulsar PSR1913 + 16 and an unseen neutron star in a binary orbit. Discovered in 1974 by R Hulse and J Taylor, it has established that gravitational radiation is correctly described by general relativity. For their discovery, Hulse and Taylor received the 1993 Nobel Prize for Physics.

The key to the importance of this binary system is that all of the important parameters of the system can be measured before one takes account of the orbital shrinking. This is because a number of post-Newtonian effects on the arrival time of pulses at the Earth, such as the precession of the position of the periastron and the time-dependent gravitational redshift of the pulsar period as it approaches and recedes from its companion, are measured in this system. They fully determine the masses and separation of the stars and the inclination and eccentricity of their orbit. From these numbers, without any free parameters, it is possible to compute the shrinking timescale predicted by general relativity. The observed rate matches the predicted rate to within the observational errors of less than 1%.

The stars are in an eccentric orbit ($e = 0.615$) and both have masses of $1.4M_\odot$. The orbital semimajor axis is about 7×8 m. Equation (19) assumes a circular orbit and gives a shrinking timescale of 6×10^{10} yr. This is an overestimate, however, partly because it is in any case a rough approximation, and partly because the timescale is very sensitive to eccentricity. With the observed eccentricity, a careful calculation shows that the expected shrinking timescale is around 4×10^8 yr, consistent with observations.

Chirping binaries

For a circular equal-mass binary, the orbital shrinking timescale and the frequency of the orbit determine both

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M and R. If in addition a gravitational wave detector measures the wave amplitude $h_{\rm binary}$, then the *distance* r to the binary system can be determined.

Remarkably, this conclusion holds even for binaries with unequal masses. In such a case, the measurable mass is the chirp mass of the binary, defined as $\mathcal{M} := \mu^{3/5} M_{\rm T}^{2/5}$, where μ is the reduced mass of the binary system and $M_{\rm T}$ its total mass. Then the distance r is still measurable from the chirp rate, frequency, and amplitude. In other words, a chirping binary is a standard candle in astronomy. Post-Newtonian corrections to the orbit, if observed in the waveform, can determine the individual masses of the stars and even their spins.

Recognizing weak signals

For ground-based detectors, all expected signals have amplitudes that are close to or even below the instrumental noise level in the detector output. Such signals can nevertheless be detected with confidence if their waveform matches an expected waveform. The pattern recognition technique that will be used by detector scientists is called matched filtering.

Matched filtering works by multiplying the output of the detector by a function of time (called the template) that represents an expected waveform, and summing (integrating) the result. If there is a signal matching the waveform buried in the noise then the output of the filter will be higher than expected for pure noise.

A simple example of such a filter is the Fourier transform, which is a matched filter for a constant-frequency signal. The noise power in the data stream is spread out over the spectrum, while the power in the signal is concentrated in a single frequency. This makes the signal easier to recognize. The improvement of the signal-to-noise ratio for the amplitude of the signal is proportional to the square root of the number of cycles of the wave contained in the data. This is well known for the Fourier transform, and it is generally true for matched filtering.

Matched filtering can make big demands on computation, for several reasons. First, the arrival time of a short-duration signal is generally not known, so the template has to be multiplied into the data stream at each distinguishable arrival time. This is then a correlation of the template with the data stream. Normally this is done efficiently using fast Fourier transform methods.

Second, the expected signal usually depends on a number of unknown parameters. For example, the radiation from a binary system depends on the chirp mass $\mathcal{M}$, and it might arrive with an arbitrary phase. Therefore, many related templates must be separately applied to the data to cover the whole family of signals.

Third, matched filtering enhances the signal only if the template stays in phase with the signal for the whole data set. If they go out of phase, the method begins to reduce the signal-to-noise ratio. For long-duration signals, such as for low-mass neutron star coalescing binaries or continuous-wave signals from neutron stars (see below), this requires the analysis of large data sets and often forces the introduction of additional parameters to allow for small effects that can make the signal drift out of phase with the template. It also means that the method works well only if there is a good prediction of the form of the signal.

Because the first signals will be weak, matched filtering will be used wherever possible. As a simple rule of thumb, the detectability of a signal depends on its *effective* amplitude $h_{\rm eff}$, defined as

$$h_{\rm eff} = h N_{\rm cycles}^{1/2} \tag{20}$$

where $N_{\rm cycles}$ is the number of cycles in the waveform that are matched by the template.

For example, the effective amplitude of the radiation from a bump on a neutron star (equation (17)) will be $h_{\rm bump}(2fT_{\rm obs})^{1/2}$, where $T_{\rm obs}$ is the observation time. In order to detect this radiation, detectors may need to observe for long periods, say four months, during which they accumulate billions of cycles of the waveform. During this time, the star may spin down by a detectable amount, and the motion of the Earth introduces large changes in the apparent frequency of the signal, so matched filtering needs to be done with care and precision.

Another example is a binary system followed to coalescence, i.e. where the chirp time in equation (19) is less than the observing time. For neutron star binaries observed by ground-based detectors this will always be the case (see the next section), so the effective amplitude is roughly

$$h_{\rm chirp} \sim h_{\rm bin}(f_{\rm gw} t_{\rm chirp})^{1/2} \sim \frac{GM}{rc^2}\left(\frac{GM}{Rc^2}\right)^{-1/4} \tag{21}$$

where for $f_{\rm gw}$ one must use twice the orbital frequency $(GM/R^3)^{1/2}/4\pi$. This may seem a puzzling result, because it says that the effective amplitude of the signal gets *smaller* as the stars become closer. However, this just means that the signal will be more detectable if it is picked up earlier, since equation (21) assumes that the signal is followed right to coalescence. If one picks up the signal at earlier times, then there are more cycles of the waveform to filter for, and this naturally gives a better signal-to-noise ratio. This gives an advantage to detectors that can operate at lower frequencies. This has been an important consideration in the design of modern detectors.

In general, the sensitivity of detectors will be limited not just by detector technology but also by the duration of the observation, the quality of the signal predictions and even the availability of computer processing power for the data analysis.

Astronomical sources of gravitational waves

Estimating the frequency

The signals for which the best waveform predictions are available have narrowly defined frequencies. In some cases the frequency is dominated by an existing motion, such as the spin of a pulsar. However, in most cases the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

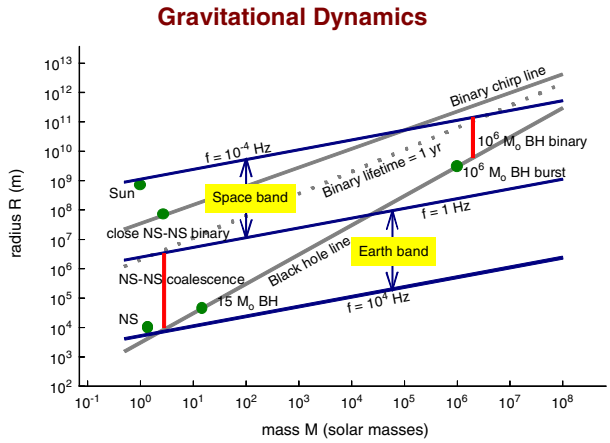


Figure 2. Mass–radius plot for gravitational wave sources.

frequency will be related to the natural frequency for a self-gravitating body, defined as

$$f_0 = (G\bar{\rho}/4\pi)^{1/2} \tag{22}$$

where $\bar{\rho}$ is the mean density of mass–energy in the source. This is of the same order as the binary orbital frequency and the fundamental pulsation frequency of the body.

The frequency is determined by the size R and mass M of the source, taking $\bar{\rho} = 3M/4\pi R^3$. For a neutron star of mass $1.4M_\odot$ and radius 10 km, the natural frequency is $f_0 = 1.9$ kHz. For a black hole of mass $10M_\odot$ and radius $2GM/c^2 = 30$ km, it is $f_0 = 1$ kHz. For a large black hole of mass $2.5 \times 10^6 M_\odot$, such as the one at the center of our Galaxy, this goes down in inverse proportion to the mass to $f_0 = 4$ mHz.

Figure 2 shows the mass–radius diagram for likely sources of gravitational waves. Three lines of constant natural frequency are plotted: $f_0 = 10^4$ Hz, $f_0 = 1$ Hz and $f_0 = 10^{-4}$ Hz. These are interesting frequencies from the point of view of observing techniques: gravitational waves between 1 and 10^4 Hz are accessible to ground-based detectors, while lower frequencies are observable only from space. Also shown is the line marking the black hole boundary. This has the equation $R = 2GM/c^2$. There are no objects below this line. This line cuts through the ground-based frequency band in such a way as to restrict ground-based instruments to looking at stellar-mass objects. Nothing over a mass of about $10^4 M_\odot$ can radiate above 1 Hz.

A number of typical relativistic objects are placed in the diagram: a neutron star, a binary pair of neutron stars that spiral together as they orbit, some black holes. Two other interesting lines are drawn. The lower (dotted) line is the 1 yr coalescence line, where the orbital shrinking timescale in equation (19) is less than one year. The upper (solid) line is the 1 yr chirp line: if a binary lies below this line then its orbit will shrink enough to make its orbital frequency increase by a measurable amount in 1 yr. (In a 1 yr observation one can in principle measure changes in frequency of 1 yr^{-1}, or 3×10^{-8} Hz.)

It is clear from the figure that any binary system that is observed from the ground will coalesce within an observing time of 1 yr. Since pulsar statistics suggest that this happens less than once every 10^5 yr in our Galaxy, ground-based detectors must be able to register these events in a volume of space containing at least 10^6 galaxies in order to have a hope of seeing occasional coalescences. When detectors reach this sensitivity (sometime in the first decade of the 21st century), then astronomers will be able to use the observed chirping binaries as standard candles to measure distance scales in the universe.

Radiation from neutron star normal modes

In figure 2 there is a dot for the typical neutron star. The corresponding frequency is the fundamental vibrational frequency of such an object. In fact, neutron stars have a rich spectrum of non-radial normal modes, which fall into several families: f-, g-, p-, w- and r-modes have all been studied. If their gravitational wave emissions can be detected, then the details of their spectra would be a sensitive probe of their structure and of the equation of state of neutron stars, in much the same way that HELIOSEISMOLOGY probes the interior of the Sun (see HIGH-ENERGY EQUATION OF STATE OF NEUTRON STARS). This is a challenge to ground-based detectors, which cannot yet make sensitive observations as high as 10 kHz.

Radiation from gravitational collapse

The event that forms a neutron star is the gravitational collapse that produces a supernova. It is difficult to predict the waveform or amplitude expected from this event, because we have no observational evidence about how non-spherical the collapse event might be in a typical supernova: the collapse is hidden deep within the star. So we can only guess. For example, a gravitational wave burst might be broad band, centered on 1 kHz, or it might be a few cycles of radiation at a frequency anywhere between 100 Hz and 10 kHz, chirping up or down. The amplitude could be large, in which a good fraction of the energy released by the collapse is radiated in gravitational waves, or it could be negligibly small. It is indeed ironic that, although detecting supernovae was the initial goal of detector development when it started four decades ago, little more is known today about what to expect than scientists knew then.

Radiation from r-modes

Hot neutron stars that rotate faster than about 100–200 Hz appear to be unstable to the emission of gravitational radiation through amplification of their r-modes by the CFS mechanism. In stars colder than about 10^8 K, viscosity may be strong enough to damp out this instability. This instability may explain why only old, recycled, cold pulsars are seen at higher rotation rates. It also suggests that the formation of a rapidly rotating neutron star may be followed by a period of steady gravitational radiation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

as the star emits angular momentum and spins down to its stability limit. If as few as 10% of all the neutron stars formed since STAR FORMATION began (at a redshift of perhaps 4) went through such a spindown, then they may have produced a detectable random background of gravitational radiation.

Interestingly, the r-modes are disturbances primarily of the fluid velocity; they have little density perturbation. Their name comes from their similarity to the Rossby waves of oceanography. The gravitational radiation they emit is not primarily mass quadrupole (as in equation (12)), but rather mass-current quadrupole, the analog of magnetic quadrupole radiation in electromagnetism. This is the wave counterpart of what is called gravitomagnetism, which is responsible for the Lense–Thirring effect: an extra precession of a spinning gyroscope as it orbits a rotating body such as the Earth caused by the spin–spin gravitational coupling of the gyroscope to the Earth.

Black holes and gravitational waves

Black holes are regions of space–time within which everything is trapped: light cannot escape, nor can anything else that moves slower than light. The boundary of this region is called the event horizon. This boundary is a dynamical surface. If any mass–energy falls into the hole, the area of the horizon increases. In addition, the horizon will generally wobble when this happens. These wobbles settle down quickly, emitting gravitational waves, and leaving a smooth (and slightly larger) horizon afterwards.

Undisturbed black holes are time independent and smooth. In fact, according to general relativity the external gravitational field of such a black hole and the size and shape of its horizon are fully determined by only three numbers: the total mass, electric charge and angular momentum of the black hole. This black hole uniqueness theorem is remarkable, considering how much variety there can be in the material that collapsed to form the black hole and that may have subsequently fallen in.

Observations of the gravitational waves emitted by a wobbling horizon or by a particle in orbit around a black hole have the potential to test the uniqueness theorem and thereby to verify the predictions of general relativity about the strongest possible gravitational fields.

Astronomers now recognize that there is an abundance of black holes in the universe. Observations of various kinds have located black holes in x-ray binary systems in the Galaxy and in the centers of galaxies.

These two classes of black holes have very different masses. Stellar black holes typically have masses of around $10M_\odot$ and are thought to have been formed by the gravitational collapse of the center of a large, evolved RED GIANT STAR, perhaps in a supernova explosion. Massive black holes in galactic centers seem to have masses between $10^6 M_\odot$ and $10^{10} M_\odot$, but their history and method of formation are not yet understood.

Both kinds of black hole can radiate gravitational waves. According to figure 2, stellar black hole radiation will be in the ground-based frequency range, while galactic holes are detectable only from space. The radiation from a black hole typically is strongly damped, lasting only a few cycles about the frequency, which for a spherical black hole is given by equation (22) with $R = 2GM/c^2$:

$$f_{\rm BH} \sim 10 \left(\frac{M}{M_\odot} \right)^{-1} \ {\rm Hz}.$$

Stellar-mass black holes

Radiation from stellar black holes is expected mainly from coalescing binary systems, when one or both of the components are a black hole. Although such systems are thought to be rarer than systems of two neutron stars, the larger mass of the black hole makes the system visible from a greater distance. By measuring the chirp mass (as discussed above) observers will recognize that they have a black hole system. It is very possible that the first observations of binaries by interferometers will be of black holes.

When a two-black-hole binary coalesces, there should be a burst of gravitational radiation that will depend in detail on the masses and spins of the objects. Numerical simulations of such events will be needed to interpret this signal and possibly even to extract it from the instrumental noise of the detector. The research field of numerical relativity is making rapid progress, and it can be expected to produce informative simulations in the first few years of the 21st century, using the largest and fastest computers available at that time.

Massive and supermassive black holes

Gravitational radiation is expected from supermassive black holes in two ways. In one scenario, two massive black holes spiral together in a much more powerful version of the coalescence we have just discussed. The frequency is much lower, but the amplitude is higher. equation (21) implies that the effective signal amplitude is almost linear in the masses of the holes, so that a signal from two $10^6 M_\odot$ black holes will have an amplitude 10^5 times bigger than the signal from two $10M_\odot$ holes at the same distance. Even allowing for differences in technology, space-based detectors will be able to study such events with a very high signal-to-noise ratio no matter where in the universe they occur.

Observations of coalescing massive black hole binaries will therefore provide strong tests of the validity of general relativity in the regime of strong gravitational fields, provided that numerical simulations can match the accuracy of the observations by that time.

The event rate for such coalescences is not easy to predict: it could be zero, but it may be large. It seems that the central core of most galaxies may contain a black hole of at least $10^6 M_\odot$. This is known to be true for our Galaxy and for a number of others nearby. Supermassive black holes (up to a few times $10^9 M_\odot$) are believed to power quasistellar objects and active galaxies (see QUASISTELLAR OBJECTS: OVERVIEW, ACTIVE GALAXIES: OVERVIEW). There is some evidence that the mass of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

central black hole is proportional to the mass of the core of the host galaxy.

If black holes are formed with their galaxies, in a single spherical gravitational collapse event, and if nothing happens to them after that, then coalescences will never be seen. However, it is believed that GALAXY FORMATION probably occurred through the merger of smaller units, subgalaxies of masses upwards of $10^6 M_\odot$. If these units had their own black holes, then the mergers would have resulted in the coalescence of many of the holes on a timescale shorter than the present age of the universe. This would give an event rate of several per year. If the supermassive black holes were formed from smaller holes in a hierarchical merger scenario, then the event rate could be hundreds or thousands per year. It is likely that only space-based observations of gravitational waves will answer these questions.

A second scenario for the production of radiation by massive black holes is the swallowing of a stellar-mass black hole or a neutron star by the large hole. Massive black holes exist in the middle of dense star clusters. The tidal disruption of main-sequence or giant stars that stray too close to the hole is thought to provide the gas that powers the quasar phenomenon. These clusters will also contain a good number of neutron stars and stellar-mass black holes. They are too compact to be disrupted by the hole even if they fall directly into it.

Such captures therefore emit a gravitational wave signal that may be approximated by studying the motion of a 'point mass' near a black hole. It will again emit a chirp of radiation, but in this case the orbit may be very eccentric. The details of the waveform encode information about the geometry of space–time near the hole. In particular, it may be possible to measure the mass and spin of the hole and thereby to test the uniqueness theorem for black holes. The event rate is not very dependent on the details of galaxy formation, and is probably high enough for many detections per year from a space-based detector.

Gravitational waves from the big bang

Gravitational waves have traveled almost unimpeded through the universe since they were generated at times as early as 10^{-24} s after the big bang. Observing them would provide important constraints on theories of inflation and high-energy physics.

Inflation is an attractive scenario for the early universe because it makes the large-scale homogeneity of the universe easier to understand. It also provides a mechanism for producing initial density perturbations large enough to evolve into galaxies as the universe expands. These perturbations are accompanied by gravitational-field perturbations that travel through the universe, redshifting in the same way that photons do. Today these perturbations should form a random background of gravitational radiation.

The perturbations arise by parametric amplification of quantum fluctuations in the gravitational wave field that existed before inflation began. The huge expansion associated with inflation puts energy into these fluctuations, converting them into real gravitational waves with classical amplitudes.

If inflation did not occur, then the perturbations that led to galaxies must have arisen in some other way, and it is possible that this alternative mechanism also produced gravitational waves. One candidate is cosmic defects, including cosmic strings and cosmic texture. (See TOPOLOGICAL DEFECTS IN COSMOLOGY.) Although observations at present seem to rule cosmic defects out as a candidate for galaxy formation, cosmic strings may nevertheless have produced observable gravitational waves.

If inflation did not occur, there could also be a thermal background of gravitational waves at a temperature similar to that of the cosmological microwave background, but this radiation would have such a high frequency that it would not be detectable by any known or proposed technique.

The random background will be detectable as a noise in the detector that competes with instrumental noise. In a single detector, such as the first space-based detector, this noise must be larger than the instrumental noise to be detected, and one must have great confidence in the detector in order to claim that the observed noise is external. This is how the cosmic microwave background was originally discovered in a radio telescope.

If there are two detectors, then one can multiply the outputs of the two detectors together and sum (integrate). In this way, the random wave field in one detector acts like a matched filtering template, matching the random field in the other detector. This allows the detection of noise that is below the instrumental noise of the individual detectors. For this to work, the two detectors must be close enough together to experience the same random wave field. In practice, the sensitivity of this method falls off rapidly with separation if the detectors are more than a wavelength apart.

Measure of the strength of random gravitational waves

When describing the strength of a random wave field, it is not appropriate to measure the amplitude of any single component. Rather, the rms amplitude of the field is the observable quantity. It is common to use an equivalent measure, the energy density $\rho_{\rm gw}(f)$ in the radiation field as a function of frequency f. For a cosmological field, what is relevant is to normalize this energy density to the critical density $\rho_{\rm c}$ required to close the universe. It is thus conventional to define

$$\Omega_{\rm gw} := \frac{{\rm d}\rho_{\rm gw}/\rho_{\rm c}}{{\rm d}\ln f}. \tag{23}$$

This is roughly the fraction of the closure energy density in random gravitational waves between the frequency f and $2.718f$.

Current and planned detectors may reach a sensitivity of $\Omega_{\rm gw} \sim 10^{-9}$ at 1 mHz and 10^{-10} at 40 Hz, but there is a possibility that backgrounds due to other sources (binary white dwarf systems and r-mode spindown, as discussed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

above) could obscure a cosmological background at these levels.

Predicted spectrum of cosmological radiation

The simplest models of inflation suggest that the spectrum of the gravitational wave background should be flat, so that Ω_{gw} is independent of frequency over a very large range of frequencies. In this case, the observed fluctuations in the cosmic microwave background radiation set a limit on gravitational radiation at ultra-low frequencies, and this constrains the energy density in the observable range (0.1 mHz–10 kHz) to below about 10^{-13} of closure. This will be too small to be seen by the current and planned detectors on the ground or in space.

However, there is a great deal of room in these models for other spectra. The period before inflation may produce initial conditions for the phase of parametric amplification that give large amounts of radiation in the observable frequency range. One family of models based in superstring theory has a spectrum that rises at high frequencies. If a cosmological background from inflation or from cosmic defects can be observed, it will contain important clues to the nature of the theory that unifies gravitation with the rest of quantum physics.

Conclusions

The first few years of the 21st century should see the first direct detections of gravitational radiation and the opening of the field of gravitational wave astronomy. Beyond that, over a period of a decade or more, one may expect observations to yield important and useful information about binary systems, stellar evolution, neutron stars, black holes, strong gravitational fields and cosmology.

If gravitational wave astronomy follows the example of other fields, such as X-RAY ASTRONOMY and RADIO ASTRONOMY, then at some level of sensitivity it will begin to discover sources that are completely unexpected. Many scientists think the chance of this happening early is very good, since the processes that produce gravitational waves are so different from those that produce the electromagnetic radiation on which most present knowledge of the universe is based, and since more than 90% of the matter in the universe is dark and interacts with visible matter only through gravitation.

Present and planned detectors are known not to be ideal for some kinds of gravitational wave sources. Sensitive measurements of a cosmological background of radiation from the big bang may not be possible with these instruments if the spectrum follows the predictions of 'standard' inflation theory. Most of the normal mode oscillations of neutron stars will be very hard to detect, because the radiation is weak and at a high frequency, but the science there is compelling: neutron star seismology may be the only way to probe the interiors of neutron stars and understand these complex and fascinating objects. Detector technology will continually improve, and these sources provide important long-term goals for this field. There will clearly be much to do after the first observations are successfully made.

Bibliography

Folkner W M (ed) 1998 *Laser Interferometer Space Antenna (AIP Conf. Proc. 456)* (Woodbury, NY: American Institute of Physics)

Marck J-A and Lasota J-P (ed) 1996 *Relativistic Gravitation and Gravitational Radiation* (Cambridge: Cambridge University Press)

Thorne K S 1994 *Black Holes and Time Warps: Einstein's Outrageous Legacy* (New York: Norton)

Wald R M (ed) 1998 *Black Holes and Relativistic Stars* (Chicago, IL: University of Chicago Press)

Bernard Schutz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitational Radiation Detectors on Earth and in Space

GRAVITATIONAL WAVES, one of the more exotic predictions of Einstein's general theory of relativity, may, after 80 yr of controversy over their existence, be detected within the next decade.

Sources such as interacting black holes, coalescing compact binary systems, stellar collapses and pulsars are all possible candidates for detection; observing signals from them will significantly boost our understanding of the universe. New unexpected sources will almost certainly be found and time will tell what new information such discoveries will bring. A full review of sources of GRAVITATIONAL RADIATION is given by Bernard Schutz.

Gravitational waves are ripples in the curvature of SPACE–TIME and manifest themselves as fluctuating tidal forces on pieces of mass placed a distance apart. The first gravitational wave detectors were based on the effect of these forces on the fundamental resonant mode of aluminum bars at room temperature. Following the lack of confirmed detection of signals, aluminum bar systems operated at and below the temperature of liquid helium were developed and work in this area is still underway. However, the most promising design of gravitational wave detector, offering the possibility of very high sensitivities over a wide range of frequency, uses test masses a long distance apart and freely suspended as pendulums on Earth or in drag-free craft in space; laser interferometry provides a means of sensing the motion of the masses produced as they interact with gravitational waves.

Ground-based detectors will be used to observe sources whose radiation is emitted at frequencies above a few Hz and space-borne detectors will be developed for implementation at lower frequencies. Already gravitational wave detectors of long baseline are being built in a number of places around the world: in the USA (LIGO project led by a Caltech–MIT consortium), in Italy (VIRGO project, a joint Italian–French venture), in Germany (GEO 600 project being built by a team of German and British research groups) and in Japan (TAMA 300 project). A space-borne detector, LISA—proposed by a collaboration of European and US research groups—has been adopted by ESA as a future Cornerstone Mission.

The observation of gravitational waves will not only provide unique information on testing aspects of general relativity but will also open up a new field of astronomy.

Background

Some early relativists were skeptical about the existence of gravitational waves. However, the 1993 Nobel Prize in Physics was awarded to Hulse and Taylor for their experimental observations and subsequent interpretations of the evolution of the orbit of the binary PULSAR, PSR 1913 +16. The decay of the binary orbit is consistent with angular momentum and energy being carried away from this system by gravitational waves.

Gravitational waves are produced when matter is accelerated in an asymmetrical way; however, owing to the nature of the gravitational interaction, significant levels of radiation are produced only when very large masses are accelerated in very strong gravitational fields. Such a situation cannot be found on Earth but is found in a variety of astrophysical systems. Gravitational wave signals are expected over a wide range of frequencies; from $\sim 10^{-17}$ Hz in the case of ripples in the cosmological background to $\sim 10^{3}$ Hz from the formation of neutron stars in supernova explosions. The most predictable sources are BINARY STAR systems. However, there are many sources of much greater astrophysical interest associated with black hole interactions and coalescences, neutron star coalescences, stellar collapses to neutron stars and black holes (supernova explosions), pulsars and the physics of the early universe.

Why is there currently such interest worldwide in the detection of gravitational waves? This is partly because observation of the velocity and polarization states of the signals will allow a direct experimental check of the wave predictions of GENERAL RELATIVITY, but more importantly because the detection of the signals should provide observers with new and unique information about astrophysical processes (see also GENERAL RELATIVITY TESTS: BINARY PULSARS).

Detection of gravitational waves

Gravitational waves are most simply thought of as ripples in the curvature of space–time, their effect being to change the separation of adjacent masses on Earth or in space; this tidal effect is the basis of all present detectors. The problem for the experimental physicist is that the predicted magnitudes of the strains in space caused by gravitational waves are of the order of 10^{-21} or lower. Indeed, current theoretical models suggest that in order to detect a few events per year—from coalescing neutron star binary systems for example—a sensitivity close to 10^{-22} is required. Signal strengths at the Earth for a number of sources are shown in figure 1.

The small signal levels mean that limiting noise sources resulting from the thermal motion of molecules in the detector (thermal noise), from seismic or other mechanical disturbances, and from noise associated with the detector readout, whether electronic or optical, must be reduced to a very low level. For signals above $\sim$10 Hz ground-based experiments are possible, but for lower frequencies where local fluctuating gravitational gradients and seismic noise on Earth become a problem, it is best to consider developing detectors to be used in space.

Initial detectors and their development

The earliest experiments in the field were ground based and were carried out by Joseph Weber of the University of Maryland about 30 yr ago. Having looked for evidence of excitation of the normal modes of the Earth by very low-frequency gravitational waves, Weber then moved on to look for tidal strains in aluminum bars which were at

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

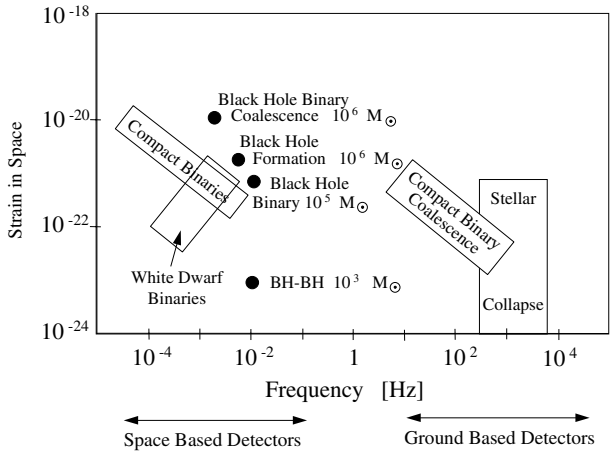


Figure 1. Some possible sources for ground-based and space-borne detectors.

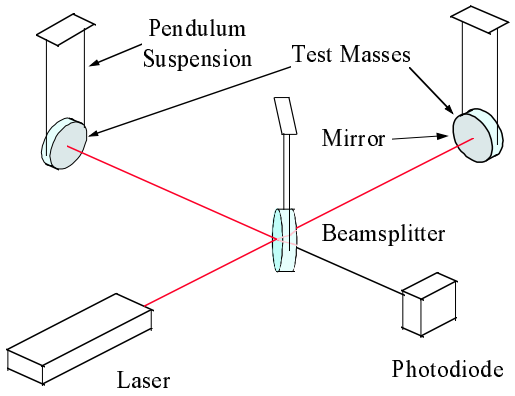


Figure 2. Schematic of gravitational wave detector using laser interferometry.

room temperature and were well isolated from ground vibrations and acoustic noise in the laboratory. The bars were resonant at 1600 Hz, a frequency where the energy spectrum of the signals from collapsing stars was predicted to peak. Despite the fact that Weber observed coincident excitations of his detectors placed up to 1000 km apart, at a rate of approximately one event per day, his results were not substantiated by similar experiments carried out in several other laboratories in the USA, Germany, Britain and Russia. It seems unlikely that Weber was observing gravitational wave signals because, although his detectors were very sensitive, being able to detect strains of around 10^{-15} over millisecond timescales, their sensitivity was far away from what was predicted to be required theoretically.

Development of Weber bar type detectors has continued with the emphasis being on cooling to reduce the noise levels, and currently systems at the Universities of Rome, Padua, Louisiana and Perth (Western Australia) are achieving sensitivity levels better than 10^{-18} for millisecond pulses. Bar detectors have a disadvantage, however, of being sensitive only to signals that have significant spectral energy in a narrow band around their resonant frequency.

Long-baseline detectors on Earth

An alternative design of gravitational wave detector offers the possibility of very high sensitivities over a wide range of frequency. This uses test masses placed a long distance apart and freely suspended as pendulums to isolate against seismic noise and reduce the effects of thermal noise; laser interferometry provides a means of sensing the motion of the masses produced as they interact with a gravitational wave (figure 2).

This technique is based on the Michelson interferometer and is particularly suited to the detection of gravitational waves as they have a quadrupole nature. Waves propagating perpendicular to the plane of the interferometer will result in one arm of the interferometer being increased in length while the other arm is decreased and vice versa. Gravitational wave strengths are characterized by the gravitational wave amplitude h, given by

$$h = 2\Delta L/L \tag{1}$$

where ΔL is the change in the length L of one arm of the detector. Thus h is a measure of the strain in space induced by a gravitational wave.

The induced change in the length of the interferometer arms results in a small change in the interference pattern of the light observed at the interferometer output. A typical design specification to allow a reasonable probability for detecting sources requires a noise floor in strain smaller than 2×10^{-23} Hz$^{-1/2}$ to be achieved. The distance between test masses possible on Earth is limited to a few km by geographical and cost factors. If we assume an arm length of 3 km the above specification sets the requirement that the residual motion of each test mass is smaller than 3×10^{-20} m Hz$^{-1/2}$ over the operating range of the detector, which may be from ~10 Hz to a few kHz. It requires that the optical detection system at the output of the interferometer must be good enough to detect such small motions.

Main noise sources relevant to interferometric gravitational wave detectors

Fundamentally it should be possible to build systems using laser interferometry to monitor strains in space which are limited only by the Heisenberg UNCERTAINTY PRINCIPLE; however, there are other practical issues which must be taken into account. Fluctuating gravitational gradients pose one limitation to the interferometer sensitivity achievable at low frequencies and it is the level of noise from this source that dictates that experiments to look for sub-Hz gravitational wave signals have to be carried out in space.

In general for ground-based detectors the most important limitations to sensitivity result from the effects of seismic and other ground-borne mechanical noise, thermal noise associated with the test masses and their

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

suspensions, and shot noise in the photocurrent from the photodiode which detects the interference pattern.

The significance of each of these sources will be briefly reviewed.

Seismic noise

Seismic noise at a reasonably quiet site on the Earth follows a spectrum in all three dimensions close to $10^{-7}/f^2$ m Hz$^{1/2}$. Thus if the motion of each test mass has to be less than 3×10^{-20} m Hz$^{-1/2}$ at a frequency such as 30 Hz then the level of seismic isolation required at 30 Hz, for example, in the horizontal direction is greater than 10^9. Since there is liable to be some coupling of vertical noise through to horizontal a significant level of isolation has to be provided in the vertical direction also.

Isolation can be achieved in a relatively simple way by realizing that for a simple pendulum system the transfer function to the pendulum mass of the horizontal motion of the suspension point falls off as 1/(frequency)^2 above the pendulum resonance. In a similar way isolation can be achieved in the vertical direction by suspending a mass on a spring. In the case of the VIRGO detector system the design allows operation to below 10 Hz and here a seven-stage horizontal pendulum arrangement is adopted with six of the upper stages being suspended with cantilever springs to provide vertical isolation. Similar systems are being developed in Australia and at Caltech. For the GEO 600 detector, where operation down to 50 Hz is sought, a triple-pendulum system is used with the first two stages being hung from cantilever springs to provide the vertical isolation necessary to achieve the desired performance. This arrangement is then hung from a plate mounted on passive 'rubber' isolation mounts and on an active (electromechanical) anti-vibration system.

In order to cut down motions at the pendulum frequencies, active damping of the pendulum modes has to be incorporated, and to reduce excess motions at low frequencies around the microseismic peak low-frequency isolators have to be incorporated. These low-frequency isolators can take different forms—tall inverted pendulums in the horizontal direction and magnetic antisprings for the vertical direction in the case of the VIRGO system; Scott Russell mechanical linkages in the horizontal and torsion bar arrangements in the vertical for an Australian design; a seismometer–actuator system for GEO 600.

Thermal noise

Thermal noise associated with the mirror masses and the last stage of their suspensions is likely to be the most significant noise source at the lower end of the operating range of long-baseline gravitational wave detectors. The operating range of a detector lies between the resonances of the test masses and their pendulum suspensions, and thus it is the thermal noise in the tails of the resonant modes which is important. For any simple harmonic oscillator, such as a mass hung as a pendulum or the internal resonance of a test mass, the spectral density of thermal motion of the mass at an angular frequency ω can be expressed as

$$\overline{x^2}(\omega) = \frac{4k_{\mathrm{b}}T\omega_0^2\phi(\omega)}{\omega m[(\omega_0^2-\omega^2)^2+\omega_0^4\phi^2(\omega)]} \qquad (2)$$

where k_{b} is Boltmann's constant, T is the temperature, m is the mass and $\phi(\omega)$ is the loss angle or factor of the oscillator of resonant angular frequency ω_0. Inspection of this equation tells us that the thermal noise away from resonance becomes smaller as $\phi(\omega)$ is reduced. For a pendulum, most of the energy is stored in the lossless gravitational field and thus the loss factor is lower than that of the material of the wires or fibers used to suspend the pendulum. Unfortunately this is not the case for the internal resonances of the test masses.

In order to keep the background thermal noise between the pendulum resonances and the internal resonances of the test masses as low as possible, the material loss factors should be as low as possible or the mechanical quality factors of the masses and pendulum resonances should be as high as possible. To achieve the level of sensitivity discussed earlier the loss factors of the test masses (of order 10 kg) must be $\sim 3 \times 10^{-8}$ or lower and the loss factor of the pendulum resonances should be smaller than 10^{-8}. Obtaining these values puts significant constraints on the choice of material for the test masses and the suspending fibers. One viable solution is to use fused silica masses hung by fused silica fibers as the intrinsic loss factor in samples of synthetic fused silica has been measured at around 2×10^{-8}. The use of other materials such as sapphire may be possible. Work in this area is being carried out by a number of universities, predominantly at Moscow and Glasgow (GEO 600 group), and more recently at Perugia, Syracuse and Stanford.

Photoelectron shot noise

For gravitational wave signals to be detected, the output of the interferometer must be held at a particular point on an interference fringe. An obvious point to choose is halfway up an interference fringe since the change in photon number produced by a given differential change in arm length is greatest at this point. The interferometer may be stabilized to this point by sensing any changes in intensity at the interferometer output with a photodiode and feeding the resulting signal back, with suitable phase, to a transducer capable of changing the position of one of the interferometer mirrors. Information about changes in the length of the interferometer arms can then be obtained by monitoring the signal fed back to the transducer.

As mentioned earlier it is very important that the system used for sensing the optical fringe movement on the output of the interferometer can sense strains in space of 2×10^{-23} Hz$^{-1/2}$ or differences in the lengths of the two arms of less than 10^{-19} m Hz$^{-1/2}$, a minute displacement compared with the wavelength of light, $\sim 10^{-6}$ m. A limitation to the sensitivity of the optical readout scheme is set by shot noise in the detected photocurrent, and from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

consideration of the number of photoelectrons measured in a time τ it can be shown that the detectable strain sensitivity depends on the level of laser power P of wavelength λ used to illuminate the interferometer of arm length L, and on the time τ, such that

$$\text{detectable strain in time } \tau = \frac{1}{L}\left(\frac{\lambda hc}{2\pi^2 P\tau}\right)^{1/2} \quad (3)$$

or

$$\text{detectable strain/Hz}^{1/2} = \frac{1}{L}\left(\frac{\lambda hc}{\pi^2 P}\right)^{1/2} \quad (4)$$

where c is the velocity of light and h is Planck's constant. Thus to achieve the required strain sensitivity level requires a laser, operating at a wavelength of $\sim 10^{-6}$ m, to provide 6×10^6 W power at the input to the interferometer. This is a formidable requirement; however, there are a number of techniques that allow a large reduction in this power requirement and these will be discussed in the next section.

Interferometric techniques for gravitational wave detectors

The situation can be helped greatly if a multipass arrangement is used in the arms of the interferometer as this multiplies up the apparent movement by the number of bounces the light makes in the arms. The multiple beams either can be separate as in an optical delay line or may lie on top of each other as in a Fabry–Perot resonant cavity.

The light should not stay in the arms longer than the characteristic timescale of the signal—otherwise some cancellation of the detected signal may occur. There is thus an optimum number of bounces. If signals of characteristic timescale 1 ms are to be searched for, the number of bounces should not exceed 50 for an arm length of 3 km. With 50 bounces the required laser power is reduced to 2.4×10^3 W, still a formidable requirement.

Power recycling

It can be shown that the optimum signal-to-noise ratio in a Michelson interferometer is obtained when the arm lengths are such that the output light is very close to a minimum. Thus rather than lock the interferometer to the side of a fringe as discussed above, it is usual to make use of a modulation technique to operate the interferometer close to a null in the interference pattern. An electro-optic phase modulator placed in front of the interferometer can be used to phase modulate the input laser light. If the arms of the interferometer are arranged to have a slight mismatch in length, this results in a detected signal which, when demodulated, is zero with the cavity exactly on a null fringe. The detected signal changes sign on different sides of the null, providing a bipolar error signal; this can be fed back to the transducer controlling the interferometer mirror to hold the interferometer locked near to a null fringe.

In this situation, if the mirrors are of very low optical loss, nearly all of the light supplied to the interferometer is reflected back towards the laser. In other words the laser is not properly impedance matched to the interferometer. The impedance matching can be improved by placing another mirror of carefully chosen transmission—a power recycling mirror—between the laser and the interferometer so that a resonant cavity is formed between this mirror and the rest of the interferometer and no light comes back towards the laser. There is then a power build-up inside the interferometer, a build-up which can be high enough to create the required laser power at the beamsplitter from $\sim$10 W or so out of the laser.

Signal recycling

To enhance further the sensitivity of an interferometric detector and to allow some narrowing of the detection bandwidth, which may be valuable in searches for continuous wave (CW) sources of gravitational radiation, another technique known as signal recycling was invented in Glasgow by the late Brian Meers. This relies on the fact that sidebands created on the light by gravitational wave signals interacting with the arms do not interfere destructively and so do appear at the output of the interferometer. If a mirror of suitably chosen reflectivity is put at the output of the system as shown in figure 3, then the sidebands can be recycled back into the interferometer where they resonate and hence the signal size over a given bandwidth (set by the mirror reflectivity) is enhanced.

Application of these techniques

Using appropriate optical configurations the required laser power may thus be reduced to a level ($\sim$10 W) where laser sources are available. It should be noted, however, that trade-offs between the different techniques are possible. For example, the GEO 600 detector uses a four-bounce delay line in each arm with power recycling and signal recycling, whereas the initial designs of the LIGO and VIRGO detectors have Fabry–Perot cavities in the arms and use power recycling but not signal recycling. It is expected that both these experiments will be upgraded to use signal recycling at a later stage of their development.

There are few lasers currently available that can produce CW laser light of enough power and with high enough frequency stability to allow the levels of sensitivity discussed above. Indeed there is only one realistic candidate—the diode-pumped Nd:YAG laser capable of producing several tens of watts of single-frequency light at 1.06 μm wavelength.

Long-baseline detectors under construction on Earth

Prototype detectors using laser interferometry have been constructed by various research groups around the world—at the Max-Planck-Institut für Quantenoptik in Garching, at the University of Glasgow, at California Institute of Technology, at the Massachusetts Institute of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

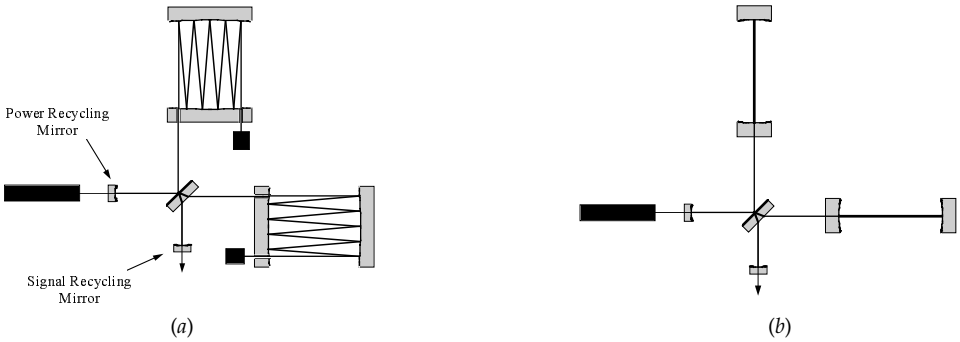


Figure 3. The implementation of power and signal recycling on Michelson interferometers with arms containing (*a*) delay lines and (*b*) Fabry–Perot cavities respectively.

Technology, at the Institute of Space and Astronautical Science in Tokyo and at the astronomical observatory in Tokyo. These detectors have arm lengths varying from a few m to 100 m and have or had either multibeam delay lines or resonant Fabry–Perot cavities in their arms. Several years ago the sensitivities of some of these detectors reached a level—better than 10^{-18} for millisecond pulses—where it was sensible to decide to build detectors of much longer baseline which should be capable of reaching the performance required to have a real possibility of detecting gravitational waves. Thus an international network of gravitational wave detectors is now under construction.

The American LIGO project comprises the building of two detector systems with arms of 4 km length, one near Hanford, Washington State, and one near Livingston, Louisiana. The French–Italian VIRGO detector of 3 km arm length is being built at Cascina near Pisa, and the TAMA 300 detector, which has arms of length 300 m, is under construction at the astronomical observatory in Tokyo. All the systems mentioned above are designed to use resonant cavities in the arms of the detectors and use standard wire sling techniques for suspending the test masses. The German–British detector GEO 600 with arms of length 600 m is somewhat different. It makes use of a four-pass delay line system with power and signal recycling, utilizes very low-loss fused silica suspensions for the test masses and should have a sensitivity at frequencies above a few hundred Hz comparable with the first phases of VIRGO and LIGO when they are in operation at the beginning of next century. A bird's-eye view of the LIGO detector is shown in figure 4.

Initial operation of these detectors is expected to commence in the early years of the next century and we can expect some very interesting coincidence searches for gravitational waves, at a sensitivity level of approximately 10^{-21} for pulses of several milliseconds duration. Later upgrades of the LIGO and VIRGO detectors should allow operation at sensitivities an order of magnitude better and at this point these detectors may be joined by other instruments being proposed in Japan and Australia.

Figure 4. A bird's-eye view of the LIGO detector under construction near Hanford in Washington State, USA.

Longer-baseline detectors in space

Perhaps the most interesting sources of gravitational waves—those resulting from BLACK HOLE formation and coalescence—lie in the region from 10^{-4} Hz to 10^{-1} Hz and a detector whose strain sensitivity is approximately 10^{-23} over relevant timescales is required to search for these. The most promising way of looking for such signals is to fly a laser interferometer in space, i.e. to launch a number of drag-free spacecraft into orbit and to compare the distances between test masses in these craft along

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

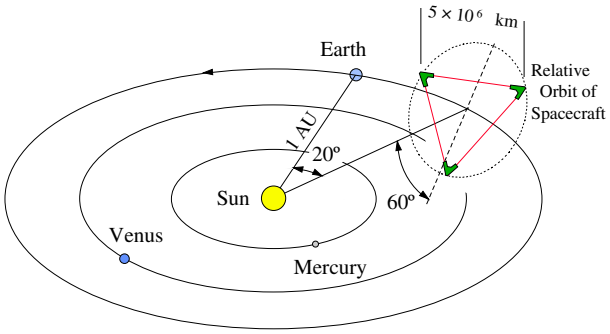


Figure 5. The proposed LISA detector.

arms making significant angles with each other using laser interferometry.

A particularly exciting experiment is being proposed by an American–European team; it consists of an array of three drag-free spacecraft at the vertices of an equilateral triangle of length of side 5×10^6 km. This cluster is placed in an Earth-like orbit at a distance of 1 AU from the Sun and 20° behind the Earth. Proof masses inside the spacecraft (two in each spacecraft) form the end points of three separate but not independent interferometers. Each single two-arm Michelson-type interferometer is formed from a vertex (actually consisting of the proof masses in a 'central' spacecraft) and the masses in two remote spacecraft as indicated in figure 5.

The three-interferometer configuration provides redundancy against component failure, gives better detection probability and allows the determination of the polarization of the incoming radiation. The spacecraft in which the interferometers are accommodated shields each pair of proof masses from external disturbances (e.g. solar radiation pressure). Drag-free control servos enable the spacecraft to follow the proof masses to a high level of precision, the drag compensation being effected using proportional electric thrusters. Illumination of the interferometers is by highly stabilized laser light from compact Nd:YAG lasers at a wavelength of 1.064 μm. For each interferometer—consisting of a central spacecraft and two distant spacecraft—two lasers in the central spacecraft, each pointing along one of the arms, are phase locked together so they effectively behave as a single laser.

For LISA to achieve its design performance, adjacent arm lengths have to be sensed to an accuracy of better than 30 pm $\mathrm{Hz}^{-1/2}$. Because of the long distances involved and the spatial extent of the laser beams, the low photon fluxes make it impossible to use standard mirrors for reflection; thus active mirrors with phase-locked laser transponders on the spacecraft will be implemented. Telescope mirrors will be used to reduce diffraction losses on transmission of the beams and to increase the collecting area for reception of the beams. Given that the available laser power in each arm is of the order of 1 W, and that sensitivity arguments similar to those already discussed for ground-based detectors can be made, photoelectron shot noise considerations suggest that the diameters of the transmitting and receiving mirrors on the space craft need to be ∼30 cm.

LISA has been adopted by ESA as a Cornerstone Project in their post Horizon 2000 program. However, because of financial uncertainties the timescale of this program is somewhat long and the possibility of an earlier launch as a NASA-led or ESA-led collaborative medium-scale mission is being enthusiastically addressed at present.

Conclusion

A large amount of effort worldwide is now being invested in the development of both ground and spaced based searches for gravitational radiation and we are entering a new era where the signals from neutron star and black hole interactions will widen our understanding of the Universe. Beyond this, however, there is the very exciting prospect that gravitational wave astronomy will, like radio astronomy and x-ray astronomy, allow the discovery of very active sources currently unknown to us.

Bibliography

The existence of gravitational waves was postulated by Albert Einstein:

Einstein A 1916 *Preuss. Akad. Wiss. Berlin, Sitzungsber. Phys.-Math. Klasse* 688

Einstein A 1918 *Preuss. Akad. Wiss. Berlin, Sitzungsber. Phys.-Math. Klasse* 154

First experiments with apparently positive results were carried out by Joseph Weber of the University of Maryland:

Weber J 1969 *Phys. Rev. Lett.* **22** 1320

These results were not confirmed by a number of experimental groups—for a discussion of this see for example

Kafka P and Schnupp L 1978 *Astron. Astrophys.* **70** 97

Tyson J A and Gifford R P 1978 *Ann. Rev. Astron. Astrophys.* **16** 521

A good review of bar detectors and developments towards laser interferometer detectors is well described in

Saulson P R 1994 *Fundamentals of Interferometric Gravitational Wave Detectors* (Singapore: World Scientific)

For a popular account of the development of the gravitational wave field the reader should consult

Thorne K S 1994 *Black Holes and Time Warps* (New York: Norton) ch 10

The challenge of reducing thermal noise in interferometric gravitational wave detectors is highlighted in

Braginsky V B, Mitrofanov V P and Panov V I 1985 *Systems with Small Dissipation* ed K S Thorne and C Eller (Chicago, IL: University of Chicago Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rowan S, Twyford S M, Hutchins R, Kovalik J, Logan J E, McLaren A C, Robertson N A and Hough J 1997 *Phys. Lett.* A **233** 303

Saulson P R 1990 *Phys. Rev.* D **42** 2437

Recent progress in interferometric detectors is outlined in

Ciufolini I and Fidecaro F (ed) 1997 *Gravitational Waves, Sources and Detectors (Edoardo Amaldi Foundation Series vol 2)* (Singapore: World Scientific)

Developments towards LISA are well documented in

1997 *1st Int. LISA Symp., Class. Quantum Gravity* **14**

Jim Hough and Sheila Rowan

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravitational Waves

Wavelike disturbances in the geometry of space and time. Just as an accelerating or oscillating charged particle will radiate electromagnetic waves, so an accelerating, oscillating or violently disturbed mass, or system of masses, is expected to radiate wave-like gravitational disturbances (i.e. gravitational waves). According to the general theory of relativity, any change in a gravitational field should propagate through space at the speed of light. For example, if the Sun were suddenly to be annihilated, 8.3 min (the light-travel time from the Sun to the Earth) would elapse before the Earth ceased to 'feel' the gravitational influence of the Sun. Likewise, gravitational waves in general should propagate through space at the speed of light.

A gravitational wave vibrates in planes perpendicular to the direction in which it is propagating. The effect of a gravitational wave passing through a solid body (or through a distribution of individual particles) is to stretch the body first along one direction (while compressing it at right angles to this direction) and then, similarly, in the plane perpendicular to this direction. In principle, gravitational waves may be detected by measuring the periodic distortions that they produce in solid bodies or in the separations between freely suspended masses or freely orbiting spacecraft. However, because gravitation is by far the weakest of the fundamental forces, gravitational waves are expected to be exceedingly feeble and difficult to detect (for example, because the gravitational attraction between two electrons is more than 10^{40} times weaker than the electromagnetic interaction, the gravitational waves radiated by a pair of oscillating electrons will also be more than 10^{40} times weaker than the electromagnetic waves that they emit). As yet no man-made device has made a confirmed detection of gravitational radiation from any source.

The most likely sources of potentially detectable gravitational waves are processes involving dramatic changes in strong gravitational fields, for example the collapse of the core of a massive star during a supernova explosion, the formation of a black hole, very close and rapidly rotating binary systems, or collisions and mergers involving neutron stars or black holes. Although no source of gravitational radiation has yet been detected directly, strong confirmation of the existence of gravitational radiation has been provided by the behavior of pulsar PSR 1913+16, which is a member of a binary system consisting of two neutron stars that orbit around each other in 7.75 h. The orbital period of this system is decreasing at a rate that is precisely consistent with what would be expected if the two stars were spiralling in towards each other as a consequence of radiating away rotational energy in the form of gravitational waves.

See also: fundamental forces, general theory of relativity, gravitation, neutron stars, pulsars, supernovae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Graviton

The hypothetical force-carrying particle (gauge boson) for gravitation. Although there is, as yet, no generally accepted quantum theory of gravity, it is widely believed that gravitation, like the other fundamental forces, should be amenable to being formulated in quantum terms and that the gravitational interaction between particles of matter should be conveyed by gravitons. The hypothesized graviton has zero mass, zero charge and a spin value of 2 (expressed in units of the Planck constant divided by 2π).

See also: fundamental forces, gauge bosons, gravitation, quantum mechanics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gravity Probe B

A relativity gyroscope experiment being developed by NASA and Stanford University to test two unverified predictions of Einstein's general theory of relativity. Due for launch into polar orbit in May 2002. It will very precisely monitor tiny changes in the direction of spin of four gyroscopes on the satellite. Free from disturbance, the gyroscopes will measure how much space and time are warped by the presence of the Earth, and how the Earth's rotation drags space-time around.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gray, Stephen (1666–1736)

Astronomer and scientist, born in Canterbury, Kent, England. Something of a scientific enigma, Gray at first devoted his energy to astronomy, observing eclipses, sunspots, the satellites of Jupiter, and communicating his results to the first Astronomer Royal, JOHN FLAMSTEED. He was interested in instrumentation and invented a microscope in which a drop of water was the lens. In his later years, he devoted himself to electricity, discovering that electricity could be conducted. He was apparently not allowed to communicate the results of his electrical experiments to the Royal Society by ISAAC NEWTON, while Newton organized its meetings and concentrated on gravity, but was twice awarded its Copley Medal as Newton's influence declined.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grazing Incidence Optics

When a beam of light is reflected from a mirror or DIFFRACTION GRATING the angle of incidence (θ) is measured from the surface normal (perpendicular to the surface) to the light ray. If this angle is greater than about 80° the reflection is said to be at grazing incidence and the angle between the surface and the ray is called the grazing angle. The grazing angle is therefore $\theta_g = 90^\circ - \theta$. Optical systems in which grazing incidence is employed are called grazing incidence optics.

Optics in which the angles of incidence are near to 90° are called normal incidence optics. Imaging systems use normal incidence, when possible, because ABERRATIONS or image imperfections introduced by the reflection geometry (rather than by diffraction or scattering) become more severe as the incidence angles increase. However, for x-rays there is a compelling reason to tolerate the severe drawbacks of grazing incidence devices.

X-RAYS, electromagnetic radiation with photon energies in the range 0.1–100 keV or with wavelength in the range 124–0.124 Å, will only reflect if the surface is highly polished and at grazing incidence. The higher the photon energy the smaller the grazing angle must be to give a good reflection. At normal incidence all x-rays are absorbed.

The only way to focus x-rays with a broad energy spectrum from a large aperture is to use grazing incidence reflections and so X-RAY TELESCOPES and devices used to focus or concentrate x-rays from laboratory sources such as synchrotron rings are grazing incidence optics.

X-ray reflectivities

The complex refractive index for x-rays in materials is given by $n = (1 - \delta) - i\beta$ where both δ and β are positive and small. The imaginary part describes absorption and is directly related to the mass absorption coefficient $\mu = 4\pi\beta/\lambda\rho$ where ρ is the mass density of the material. The mass absorption is, in turn, directly related to photoelectric absorption.

The real part of n is a little less than unity and therefore the phase velocity of x-rays in materials is slightly higher than the velocity of light in vacuum. Thus when x-rays pass from vacuum into a material the refraction angle (measured from the normal within the material) is slightly larger than the incidence angle. At a critical grazing angle given by $\cos\theta_t = 1 - \delta$ total external reflection occurs, analagous to total internal reflection of visible light at a glass–vacuum boundary. Because of the imaginary part of the refractive index, β, the reflection is not total but for grazing angles smaller than $\theta_t \approx (2\delta)^{1/2}$ the reflection coefficient for x-rays can be high. To obtain good reflectivities the surface must be coated with dense, heavy metals such as nickel, gold, platinum or iridium. Ignoring absorption edges the critical angle in degrees for heavy elements of density ρ (g cm^{-3}) at x-ray wavelength λ measured in Å is given by $\theta_t \approx 0.1\lambda\rho^{1/2}$. For gold the critical angle is approximately $\theta_t = 0.4\lambda$.

The decrements δ and β are functions of wavelength and the specular reflection coefficient for x-rays of a given wavelength can be calculated accurately using Fresnel's equations in the same way as for longer-wavelength electromagnetic radiation. However, since the reflection is only appreciable at grazing angles there is very little difference between the reflection efficiency of the parallel and perpendicular polarizations.

The value of β can be calculated theoretically using numerical electronic wavefunctions or obtained from measurements of photoelectric absorption. The real decrement δ can then be calculated using the Kramers–Kronig relation from dispersion theory. Alternatively δ can be calculated directly using Kramers–Kallmann–Mark anomalous dispersion theory. For x-ray wavelengths sufficiently far from absorption edges ($|\Delta E| > 30$ eV) the refractive index can be predicted accurately. In the vicinity of absorption edges the calculations are limited by the details of the electronic energy structure included in the theory and by extended x-ray absorption fine structure (EXAFS), which is a correction introduced by a second scattering of the out-going primary scattered wave by nearby atoms in the surface. Tabulations of the refractive index decrements for all atomic types over the entire x-ray energy band are available and therefore the Fresnel reflectivity of x-rays from mirror surfaces can be calculated in a straightforward manner.

Figure 1 shows the calculated reflectivity for pure gold. The left-hand panel shows how the critical grazing angle becomes smaller as the photon energy increases. The right-hand panel shows the reduction in reflectivity introduced by absorption edges. There is a series of M-edges in gold starting at 2.1 keV.

Bragg reflections from crystals or multilayers can be used for narrow band imaging and spectroscopy or for enhancing the high-energy response of grazing incidence surfaces but they are inefficient for broad-band use below 15 keV.

Refractive x-ray lenses are not possible because the real part of the refractive index in materials, $1 - \delta$, is too close to unity and the penetration depth is too small because β is relatively large.

Scattering of x-rays at grazing incidence

The wavelength range for x-rays is 0.1–100 Å and therefore roughness and imperfections in the surface chemistry and geometry on the atomic scale are expected to influence specular reflection from the surface. Roughness or chemical contamination introduces small phase and amplitude changes in the reflected x-ray wavefronts and some of the light is scattered (or diffracted) away from the specular direction.

The complete theory of scattering of electromagnetic radiation from surface roughness was originally developed to describe scattering of radar from land and sea but the results apply equally well to the much shorter-wavelength regime of x-rays. Surface roughness is modeled as a two-dimensional aperiodic diffraction grating

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

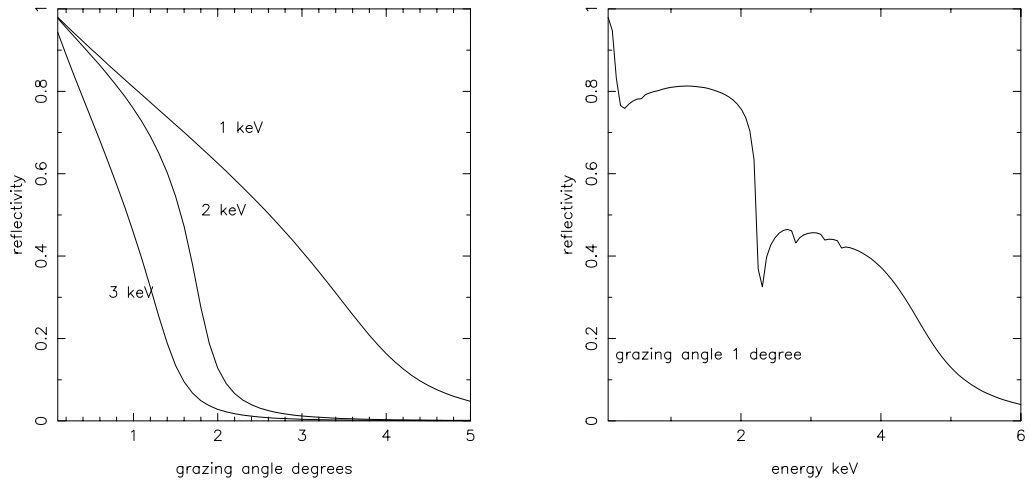


Figure 1. The x-ray reflectivity of gold.

and each Fourier component of the surface height profile, period d, produces diffraction orders m at angles given by the grating equation $\sin\theta_s - \sin\theta = m\lambda/d$ where θ is the angle of incidence and θ_s is the angle of scatter measured with respect to the surface normal. The total integrated scatter (TIS) is the fraction of the incident intensity which is scattered out of the specular beam. If the surface is reasonably smooth and first-order scattering dominates then the TIS $= R_\tau(\theta,\lambda)(4\pi\sigma\cos\theta/\lambda)^2$ where R_τ is the Fresnel reflectivity for polarization τ and σ is the root mean square (rms) of the surface height fluctuations. Typically σ must be $\leq$5 Å rms to keep the TIS at an acceptable level. Surface contamination which is distributed unevenly over the surface can be modeled in a similar way and acts like a phase grating. Any particle or dust contamination will further block or scatter the x-rays.

Surface contamination by hydrocarbons or similar compounds often forms a smooth layer or layers. A Fresnel reflection will result from each interface and the total reflection must be calculated by summing the component wave amplitudes taking into account the phase difference introduced by the thickness of the layers. The contamination acts in the same way as a multilayer coating on a conventional lens. A single monolayer of hydrocarbon contamination can severely compromise or indeed enhance the x-ray reflectivity of a grazing incidence mirror.

Geometries for imaging at grazing incidence

A nearly stigmatic telescope system must have a principal surface that is a sphere centered on the axial focus. If the F-ratio is small enough (small aperture diameter/focal length) then this sphere is well approximated by a plane perpendicular to the optical axis. For an imaging system which uses just one reflection the principal surface of the optic must necessarily be the reflecting surface itself and therefore a grazing incidence system employing one reflection is very astigmatic. A parabolic mirror with high curvature and large enough diameter such that the outer edges are reflecting at grazing incidence will bring rays parallel to the axis into a point focus. However, parallel rays which are slightly off-axis will form a highly comatic image consisting of a circle centered on the axis.

It is possible to produce a nearly stigmatic imaging system using an even number of grazing incidence reflections. The first such system devised was described by P Kirkpatrick and A V Baez in 1948. This comprises two orthogonally mounted cylinders each of which produces a line focus as illustrated in figure 2. The axially symmetric solution to the problem was provided by H Wolter in 1952. The two Wolter geometries which can be used for telescopes are called Wolter type I and type II and are illustrated in figure 3. The solid lines are the physical mirrors, the dashed lines indicate the path of on-axis rays and the dotted lines show the principal surface and surface generators. The first reflecting surface has a parabolic profile while the second is hyperbolic. In fact the surfaces are essentially the same as a CASSEGRAIN TELESCOPE except that grazing incidence rather than normal incidence is used. For the type I design the two physical surfaces meet at a join plane and the distance from this plane to the focus is the focal length. For type II the join plane is in front of the physical surfaces and therefore the focal length is larger than the instrument length. This design is best suited to larger grazing angles and is therefore only useful for longer wavelengths or lower energies in the soft x-ray and extreme ultraviolet bands. Wolter type II telescopes were used on the EXTREME ULTRAVIOLET EXPLORER satellite launched 1992.

Both the Kirkpatrick–Baez (K–B) and Wolter type I geometries can be nested. In the K–B case a large number of reflecting plates can be stacked together such that they have a common line focus. For Wolter I systems a series of surfaces of revolution are placed one inside the other with a common axis. For a Wolter type I aperture of radius R the grazing angle is given by $\theta_g = \tan^{-1}(R/F)/4$ where F

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

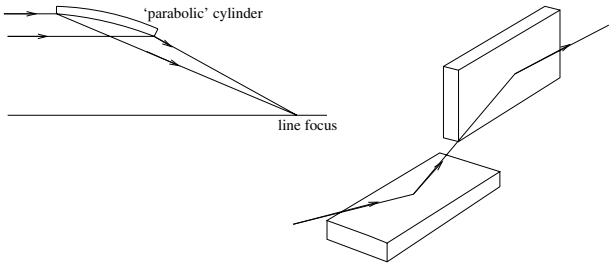


Figure 2. The Kirkpatrick–Baez geometry.

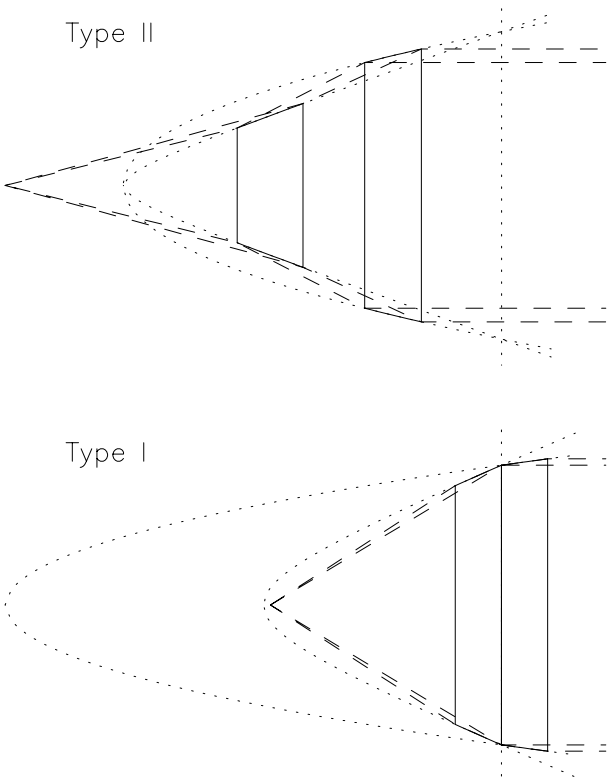


Figure 3. Wolter type I and II geometries.

is the focal length. The F-ratio therefore determines the largest grazing angles in the nest. Efficient nesting can be achieved over radii from R to $R/3$. Further surfaces added at smaller radii will increase the high-energy collecting area but have little effect on the lower-energy collecting area.

Small square pores or channels can be packed together to approximate a K–B or Wolter I geometry. If the channels are packed on a Cartesian grid then the resulting optic works in the same way as the eye of a lobster. The image consists of a true focus produced by an even number of reflections, a cruciform focus produced by a odd number of reflections and a diffuse background produced by no reflections. The Wolter I geometry can be approximated by packing the square pores in a radial pattern and using two successive plates of channels.

Designing grazing incidence telescopes

All the large focusing x-ray telescopes which have been built to date have used grazing incidence reflection in the Wolter type I geometry. The main problem with grazing incidence optics is low surface utilization. The effective area of a K–B or Wolter type I system for wavelength λ is given by $A_{\rm eff} = \frac{1}{2}R^2(\theta_{\rm g}, \lambda)\sin\theta_{\rm g}\, A_{\rm geom}$ where $A_{\rm geom}$ is the geometric area and $R(\theta_{\rm g}, \lambda)$ is the reflectivity from one reflection. The grazing angle must be small so that the square of the reflectivity is high, typically $\theta_{\rm g} \leq 1.5^\circ$, and so the ratio of the polished surface area to the effective area is very large. To give a large collecting area the surfaces must be long in the axial direction and/or they must be nested. The maximum two-reflection reflectivity is usually achieved if the grazing angles on the two surfaces are equal. The surfaces look like cylinders with a slight taper. The individual surface elements are often referred to as shells and it may be advantageous to manufacture a pair of parabolic and hyperbolic surfaces in a single monolithic shell.

The figuring errors must be significantly less than the angular resolution required and the gradient errors must be kept small over the entire axial length of the surfaces. The mirror shells must be thick enough to support the figure but aperture area is inevitably lost at the front ends of each shell. There is therefore a direct trade-off between angular resolution (shell stiffness) and effective area (aperture utilization).

The radius of the field of view is necessarily less than the grazing angle but as the grazing angles increase the upper energy cutoff of the telescope decreases so there is a direct trade-off between the maximum field of view and the broad band response.

Manufacturing grazing incidence telescopes

The traditional way of producing Wolter type I x-ray mirrors is as monolithic shells. Wolter I blanks are drilled out of a block of Zerodur or similar glass ceramic. The blanks are then ground to the correct thickness and figure. Superpolish is achieved by a very fine grinding–polishing process. Provided that the shells are thick enough very high figure tolerances and a superpolish of 3 Å rms can be met giving 0.5 arcsec angular resolution. Using monolithic shells the aperture utilization is limited to $\sim$0.1 by the shell thickness required and the shells are very massive.

Wolter type I surfaces can be approximated by thin conical aluminum foil sectors with a foil thickness $\sim$130 μm. A very large number of nested sections can give a high aperture utilization of up to $\sim$0.6. A high-quality surface finish is produced by laquering or epoxy replication of a smooth mandrel former and high x-ray reflectivity is achieved by coating the surfaces with gold or some other high-Z material. Unfortunately figuring errors limit the angular resolution to $\sim$2 arcmin. The x-ray telescopes on the Japanese satellite ASCA launched in 1993 employed foil mirrors.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Large x-ray telescopes.

	Einstein	ROSAT	Chandra	XMM
Focal length (mm)	3440	2400	10 069	7500
Diameter (mm)	580	835	1200	700
Axial length (mm)	1022	1000	1660	600
Number of shells	4	4	4	58
Shell material	Fused quartz	Zerodur	Zerodur	Nickel
Surface coating	chrome–nickel	Gold	Iridium	Gold
HEW (arcsec)	9	5	1	16
FWHM (arcsec)	3	3	0.5	8
Area, 1 keV (cm^2)	400	400	800	1475
Area, 8.0 keV (cm^2)	–	–	100	580

Wolter type I shells can be produced by a replication process. The surfaces are machined on a very stiff, stable mandrel in a similar way to the production of monolithic shells except that the surfaces are external rather than internal to the cylinder. The mandrel is then coated with ∼1500 Å of gold and a thin layer of nickel, ∼1.0 mm thick, is electroformed onto the mandrel surface. Luckily the gold acts as a release agent for the electroforming process and is conveniently transferred from the mandrel onto the electroformed nickel shell. After release the shell has a superpolished gold surface ready for use as an x-ray mirror. Shells can be produced with a thickness as little as 0.3 mm and so the aperture utilization can be ∼0.4. After release residual stress in the electroformed nickel may cause the shells to distort and the shells must be handled very carefully to avoid introducing figure errors. The first replicated Wolter type I mirrors were flown on EXOSAT launched in 1983.

Microchannel plate technology can be used to produce square pores in either the K–B (lobster eye) or the Wolter I approximation (see MICROCHANNEL PLATE DETECTORS). Very small grazing angles arise for the channels near the center of the plates giving a hard x-ray response and the geometry can provide a very wide field of view. The channel walls are thin, the aperture utilization is high and the optics have very low mass. However, it is difficult to align the channels, coat the channel walls and reduce the surface roughness.

Whatever the production process the reflecting surfaces must be kept very clean and free from dust and contamination, during manufacture, through launch to deployment in space.

Large x-ray telescopes

The brightest cosmic x-ray source Sco X-1 (SCORPIUS X-1) gives a flux at the Earth of about 150 photons cm^{-2} s^{-1} in the energy band 0.1–10 keV while typical sources are at least 10 000 times fainter. To detect the weaker sources a collecting area of a few hundred cm^2 is sufficient but to detect distant objects and perform accurate photometry or spectroscopy thousands of cm^2 are required. As the sensitivity increases with larger areas so the angular resolution must improve to avoid confusion and reveal detail in diffuse and extended objects. Successive generations of x-ray telescopes have gradually improved in collecting area and angular resolution.

Figure 4. XMM mirror module number 1.

The first large imaging x-ray telescope was the EINSTEIN Observatory, launched by NASA in 1978. The first x-ray all-sky survey produced with an imaging telescope was taken by ROSAT which was built in Germany and launched in 1990. Both these instruments used a Wolter type I design and monolithic shell technology.

XMM[1] and CHANDRA are the two largest x-ray observatories built to date. Chandra was launched by NASA on 23 July 1999 and XMM was launched by ESA on 10 December 1999. They are complementary in terms of both technology and performance. XMM contains three mirror modules produced using replication technology which provide over 4000 cm^2 of collecting area. Chandra is a single telescope constructed from monolithic shells with an angular resolution of 0.5 arcsec. Both XMM and Chandra are remarkable feats of optical engineering.

Table 1 summarizes the characteristics of these telescopes. Figure 4 shows the complete XMM mirror module number 1.

[1] XMM was renamed XMM–Newton after launch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Future goals for grazing incidence optics

To achieve a sensitivity in the x-ray energy band comparable with the largest optical, infrared and radio telescopes the collecting area must be $\sim$10 m^2. It is unlikely that monolithic Wolter I shells could ever be produced to fill such an aperture area. However, it is possible to split the aperture into sectors and stack sector plates of Wolter I geometry into mirror modules which could then be integrated into a complete x-ray primary mirror, focusing all the x-ray light into a single focal plane. The XEUS concept under study at ESA is working towards this goal.

The diffraction limit of the angular resolution in x-ray astronomy is potentially very small, $\leq$0.02 arcsec for a modest 10 cm aperture. At present x-ray telescopes are nowhere near this limit. Very high angular resolution is possible if the surfaces can be manufactured and aberrations can be reduced by using multiple reflections. W Cash has suggested this can probably be realized using very-high-quality spherical mirrors in a K–B geometry. Prototype systems with $\sim$0.1 arcsec resolution have already been built. Unfortunately a high level of nesting must be used to give a useful area and the surface elements must be aligned to extremely high precision to reach the diffraction limited performance.

Bibliography

Beckmann P and Spizzichino A 1963 *The Scattering of Electromagnetic Waves from Rough Surfaces* (New York: Pergamon)

Henke B L, Lee P, Tanaka T J, Shimabukuro R L and Fujikawa B K 1982 Low-energy x-ray interaction coefficients: photoabsorption, scattering, and reflection *At. Data Nucl. Data Tables* **27** 1–144

Kirkpatrick P and Baez A V 1948 *J. Opt. Soc. Am.* **38** 766

Wolter H 1952 *Ann. Phys.* **10** 94, 286

Richard Willingale

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Great Circle

The circle obtained on the surface of a sphere by the intersection with the surface of a plane which passes through the center of the sphere. Any plane which intersects the sphere but does not pass through the center meets the sphere in a small circle. A great circle is, in fact, the largest possible circle that can be drawn on the surface of a sphere. The shortest distance between two points on a sphere is an arc of a great circle passing though those points. Examples of great circles on the Earth are the equator and lines of longitude; on the celestial sphere, the celestial equator, the meridian, ecliptic, etc.

See also: celestial coordinates, celestial latitude, celestial longitude, ecliptic, meridian.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Great Red Spot (GRS)

A huge permanent anticyclone in Jupiter's southern hemisphere, visible as a reddish oval at just over 20 °S. The earliest unequivocal observation was by Heinrich Schwabe in 1831 (the often-quoted sighting by Robert Hooke in 1664 now seems to have been of a similar but different spot). The GRS became a striking feature around 1880, when it developed a deep red coloration. It was also prominent in the early 1970s, but on other occasions the color has been less pronounced, fading on occasion to a pale buff or salmon pink and sometimes disappearing completely, to leave what is known as the *Red Spot hollow*. Its size has varied too: from about 10 000 to 14 000 km (north–south) by 24 000 to 40 000 km (east–west). The GRS drifts back and forth in longitude, and over the course of the twentieth century it completed about three circuits back and forth around the planet. There is also a 90-day oscillation in latitude that takes it nearly 2000 km either side of its mean position. In addition, it rotates counter-clockwise with a period of about 7 days. That the GRS is anticyclonic in nature is well accepted. What is not clear is what continues to drive it; one theory is that it is powered by the energy released as latent heat from the condensation of gases below. Its color variations may be caused by variations in the concentartion of phosphine (PH_3), or even of organic compounds.

See also: Jupiter: atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Great Rift

A prominent division of the Milky Way consisting of a string of dark molecular clouds. It stretches from Cygnus (a section known as the Cygnus Rift), down through Aquila, to Sagittarius.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Great Wall

A vast sheet-like aggregation of galaxies that is about 800 million light-years long, 280 million light-years 'high', but only about 15–20 million light-years 'thick'. One of the largest known structures in the universe, it lies at a distance of about 300 million light-years. The coma cluster, one of the nearest great clusters of galaxies, forms part of the 'wall'.

As with other remote structures in the universe, the distance and dimensions of the Great Wall depend on the value of the Hubble constant, H_0, the above figures being based on a value of 70 km s^{-1} Mpc^{-1}. In general terms, the size of the Great Wall, expressed in megaparsecs, is about $60/h$ by $170/h$, and its distance about $65/h$, where $h = H_0/100$. For example, if $H_0 = 70$ km s^{-1} Mpc^{-1}, then $h = 70/100 = 0.7$, and the length of the 'Wall' is $170/0.7 \approx 250$ Mpc ≈ 800 million light-years.

See also: cluster of galaxies, Coma cluster, Hubble constant, parsec, superclusters and the Local Supercluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Green Bank (National Radio Astronomical Observatory)

Located in Green Bank, Pocahontas County, West Virginia. The site of the world's largest fully steerable radio telescope, which was under construction during 1999 and 2000. The National Radio Astronomy Observatory (NRAO) is a facility of the US National Science Foundation, and is operated under a cooperative agreement by Associated Universities, Inc. (AUI).

Funding for the new 100 m (330 ft) telescope was provided by the NSF's Division of Math and Physical Sciences. The telescope has actual surface dimensions of 100 by 110 m. Its structure allows the telescope to view the entire sky above 5° elevation. Operations are expected to begin during 2000.

Of the other instruments at Green Bank, the 42 m telescope closed on 19 July 1999, while the 91 m telescope collapsed on 15 November 1988. The collapse was due to the sudden failure of a key structural element in the box girder assembly that formed the main support for the antenna. The loss of the 91 m telescope resulted in the Green Bank Telescope Project.

The Green Bank Interferometer (GBI) is also operated by the NRAO. It includes three radio telescopes of 26 m (85 ft) diameter, designated 85-1, 85-3 and 85-2 (85-1 is also known as the Tatel Telescope). 85-1 and 85-2 form a one-baseline interferometer, funded by NASA, which is engaged in monitoring variable radio sources. This GBI–NASA monitoring program observes x-ray binary stars and active galactic nuclei.

The VLBI tracking station in Green Bank is one of four such NASA facilities dedicated to the support of very long baseline interferometry satellites.

For further information see
http://info.gb.nrao.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Greenhouse Effect

The greenhouse effect is the radiative influence exerted by the atmosphere of a planet which causes the temperature at the surface to rise above the value it would normally reach if it were in direct equilibrium with sunlight (taking into account the planetary ALBEDO). This effect stems from the fact that certain atmospheric gases have the ability to transmit most of the solar radiation and to absorb the infrared emission from the surface. The thermal (i.e. infrared) radiation intercepted by the atmosphere is then partially re-emitted towards the surface, thus contributing additional heating of the surface.

Although the analogy is not entirely satisfactory in terms of the physical processes involved, it is easy to see the parallels between the greenhouse effect in the atmosphere–surface system of a planet and a horticultural greenhouse: the PLANETARY ATMOSPHERE plays the role of the glass cover that lets sunshine through to heat the soil while partly retaining the heat that escapes from the ground. The analogy goes even further, since an atmosphere may present opacity 'windows' allowing infrared radiation from the surface to escape, the equivalent of actual windows that help regulate the temperature inside a domestic greenhouse.

Physical description

A more physical description is given by the simple two-stream approximation for the solution of radiative equilibrium in a planetary atmosphere. In the simplest case of a gray atmosphere (the opacity to thermal radiation is independent of wavelength) and in the absence of absorption of the solar radiation, the thermal emission from the surface has the following expression:

$$B(\tau_g) = \frac{\sigma\theta_g^4}{\pi} = \frac{W_S}{\pi}\left(1 + \frac{3}{4}\tau_g\right) \tag{1}$$

where B is the integral of the Planck function over all wavelengths, σ is Stefan's constant, θ_g is the surface temperature, W_S is the constant downward solar energy flux and τ_g is the infrared optical depth at the surface. From this expression, one notes that the surface temperature is greater than it would be if the atmosphere were totally transparent to thermal radiation (i.e. when $\tau_g = 0$). This reference value, defined as the equilibrium effective temperature, is given by

$$\theta_{eq} = \left(\frac{W_S}{\sigma}\right)^{1/4} \tag{2}$$

and the surface temperature may then be expressed as

$$\theta_g = \theta_{eq}\left(1 + \frac{3\tau_g}{4}\right)^{1/4}. \tag{3}$$

The atmosphere being in radiative equilibrium, the fact that the surface temperature is larger than the equilibrium temperature implies that, at higher levels, the temperature must decrease below that value. Since the atmospheric temperature, expressed as a function of the optical depth τ, is

$$\theta(\tau) = \theta_{eq}\left(\frac{1}{2} + \frac{3\tau}{4}\right)^{1/4} \tag{4}$$

one notes that, for values of τ smaller than $\frac{2}{3}$, temperatures are indeed below θ_{eq}.

A comparison between equations (3) and (4) reveals a temperature discontinuity near the surface, the surface being somewhat warmer than the layer just above it. This corresponds to a difference $\delta B = W_S/2\pi$ in the radiative budgets of the two systems: the surface is assumed to only radiate upwards within 2π, whereas an atmospheric layer radiates both upwards and downwards. Formally, the amplitude of that discontinuity decreases as τ_g increases (if $\tau_g = 1$ and 100, then $\delta\theta/\theta_{eq} = 3.1\%$ and 0.5%, respectively). Physically, however, this temperature discontinuity is hydrostatically unstable and convection will tend to counter it. In this simple approximation, the radiative lapse rate is

$$-\frac{\partial\theta}{\partial z} = \frac{3}{8}\frac{(\theta/H)\tau}{1 + \frac{3}{2}\tau} \tag{5}$$

where H is a scale height characteristic of the vertical distribution of the infrared absorbing gas(es). The temperature lapse rate varies from zero (i.e. isothermal) at the top of the atmosphere to a maximum value at the surface. At some point, it may become superior to the adiabatic lapse rate, in which case convection sets in.

The case of a semi-infinite gray atmosphere opaque to both solar and thermal radiation can also be solved in the two-stream approximation and leads to the following reduced source function $B^*(\tau) = \pi B(\tau)/W_S(0)$:

$$2B^*(\tau) = \left(1 + \frac{\alpha}{2\mu_0}\right) + \left(\frac{3\mu_0}{2\alpha} - \frac{\alpha}{2\mu_0}\right)(1 - e^{-\alpha\tau/\mu_0}) \tag{6}$$

where $W_S(0)$ is the SOLAR CONSTANT at the top of the atmosphere, α is the ratio (assumed to be constant with altitude) of the extinction coefficient for solar radiation to that for thermal radiation, μ_0 is the cosine of the zenith solar angle and τ is the optical depth for thermal radiation. The boundary conditions are

$$2B^*(0) = 1 + \frac{\alpha}{2\mu_0} \tag{7}$$

$$2B^*(\infty) = 1 + \frac{3\mu_0}{2\alpha} \tag{8}$$

and their ratio is

$$\frac{B^*(\infty)}{B^*(0)} = 4\frac{x+1}{(x+2)(x-2)} \tag{9}$$

with $x = 2(1 + \alpha/\mu_0)$ varying from 2 to ∞. For values of $\alpha/\mu_0 \leq \frac{3}{2}$ (weak to moderate solar extinction versus thermal) the solution corresponds to a greenhouse

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

atmosphere in which the temperature increases with increasing optical depth, reaching values larger than the equilibrium effective temperature at some depth. For larger values, the solution is the reverse, i.e. the temperature decreases with optical depth because solar radiation is deposited in the upper atmospheric layers instead. In all cases, the temperature varies monotonically with altitude.

In real non-gray atmospheres, the extinction (of either solar or thermal radiation) varies both with altitude and with wavelength within each spectral domain. Then, an important additional feature may be the existence of a high-altitude temperature inversion if absorption of solar photons occurs preferentially in the upper atmospheric layers. The temperature in the stratosphere increases with altitude above a temperature minimum overlying a tropospheric greenhouse, provided that solar radiation is not completely blocked above these levels. In terms of the parameter α, these conditions are met if $\alpha \gg 1$ over a reasonably large fraction of the solar spectrum, and, at the same time, $\alpha \ll 1$ over most of the rest of the spectrum.

One may quantify the amplitude of an atmospheric greenhouse in at least two ways. The simplest estimator is just the relative increase in surface temperature compared with the equilibrium effective temperature $\Delta\theta/\theta = (\theta_g - \theta_{eq})/\theta_{eq}$. A more elaborate estimator is the effective value of α that can be evaluated from the radiative properties of the atmosphere. This represents a global value of α characterizing the whole atmosphere over the whole spectrum. From equation (6), one derives the following equation:

$$\frac{\pi[B(0) - B(\tau_g)]}{W_S(0) - W_S(\tau_g)} = \frac{\mu_0}{4\alpha}\left(3 - \frac{\alpha^2}{\mu_0^2}\right) \tag{10}$$

which may be solved for α:

$$\frac{\alpha}{\mu_0} = (\delta^2 + 3)^{1/2} - \delta \tag{11}$$

where

$$\delta = \beta - \frac{\chi}{2}[(\beta^2 + 3)^{1/2} - \beta + 2] \tag{12}$$

$$\beta = \frac{2\pi B(\tau_g)}{W_S(0) - W_S(\tau_g)} \tag{13}$$

and

$$\chi = \frac{W_S(0) - W_S(\tau_g)}{W_S(0)}. \tag{14}$$

These quantities are listed in table 1 for the terrestrial planets and TITAN. By all accounts, VENUS is the champion of the greenhouse effect, with the EARTH and Titan coming in second and third. MARS definitely shows the weakest case. The following section provides an attempt to explain these striking differences in terms of the climatic evolution of each planet.

Actually, as inferred from equations (6)–(9), it is not necessary that there be a surface boundary for a greenhouse effect to take place. For instance, the tropospheric thermal structure of the giant gaseous planets (Jupiter, Saturn, Uranus, and Neptune) is partly due to a greenhouse effect exerted by molecular hydrogen, which is a fairly efficient infrared absorber. Because there is no solid surface, however, this effect is spread out over a large altitude range. Furthermore, with the possible exception of Uranus, these planets possess an internal source of heat that strongly modifies the lower boundary condition and tends to overwhelm the greenhouse effect.

As noted above, the conditions that lead to the establishment of a greenhouse effect can also be reversed, thus leading to an anti-greenhouse effect. For instance, this occurs when a high-altitude haze absorbs or scatters visible sunlight and lets infrared radiation escape. On Earth, it has been predicted that extended fires resulting from a global nuclear conflict would inject large amounts of soot particles in the stratosphere that would act in such a way and provoke a dramatic surface cooling (so-called 'nuclear winter').

The above description focused on the atmospheric greenhouse effect. It is also possible for a greenhouse to occur in solid phase, for instance in the upper layers of ice deposits. The same conditions apply, that is the icy material needs to be somewhat transparent to sunlight and opaque to infrared radiation. It has been suggested that this mechanism is currently active at the surface of icy satellites.

Greenhouses in the solar system

From measurements collected in the solar system over the last 40 yr, it appears that all planets with a substantial atmosphere exhibit a greenhouse effect. The amplitude of this effect varies greatly from one planet to another, depending on the relative strength of the physical parameters involved (heliocentric distance, surface pressure, composition, feedback mechanisms, etc). In itself, this variety provides very useful clues for a better understanding of the impact of the various parameters. There are four cases of atmospheric greenhouse for bodies with a solid surface: Venus, Earth, Mars and Titan. Recent studies suggest that the greenhouse effect has played a major role in the climatic evolution of these planets. In addition, there are two cases of a suspected solid-state greenhouse, on IO and TRITON.

Atmospheric greenhouses

(*a*) *Venus*. Venus represents the archetype of the atmospheric greenhouse. Observationally, the first piece of evidence came in the late 1950s when measurements of the Venusian emission at centimetric wavelengths revealed a surface temperature on the order of 730 K. This value is about 500 K larger than the equilibrium effective temperature and could not readily be explained by the existing models of the Cytherean atmosphere. In 1960, Carl Sagan was the first to recognize that two of the main atmospheric constituents—CO_2 and H_2O—were probably responsible for a strong greenhouse effect. The Pioneer Venus mission later confirmed his hypothesis by providing a detailed account of the contribution of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Physical quantities characterizing the greenhouses on the terrestrial planets and Titan.

Descriptor	Quantity	Venus	Earth	Mars	Titan
Surface temperature	θ_g (K)	731	288	218	94
Equilibrium temperature	θ_{eq} (K)	231	255	212	82
Relative increase of surface temperature	$\Delta\theta/\theta = (\theta_g - \theta_{eq})/\theta_{eq}$	2.16	0.13	0.03	0.15
Surface emission	$\pi B(\tau_g)$ (W m^{-2})	16 170	326	50	4.12
Solar flux at the top of the atmosphere	$W_S(0)$ (W m^{-2})	937	882	512	10.3
Solar flux reaching the surface	$W_S(\tau_g)$ (W m^{-2})	20	162	102	0.35
Relative extinction of atmosphere for light and IR	α ($\mu_0 = \frac{1}{2}$)	0.02	0.71	1.62	1.29

the various greenhouse agents: in decreasing order of importance, they are CO_2, H_2O, H_2SO_4 clouds, SO_2, CO and HCl (see VENUS: ATMOSPHERE).

One of the enigmas regarding the present-day Venus is the very low abundance of water. If Venus once possessed as much water as the Earth, what process made it disappear almost completely? In answer to this question, it was suggested that Venus has experienced a 'runaway greenhouse' in its early stage. In the original models (1960–84), it was considered that all of the water outgassed by the crust accumulated in the atmosphere as steam because initial conditions were already too warm for water to condense at the surface. Then, a runaway developed from the positive feedback between the accumulation of steam in the atmosphere and the greenhouse effect induced by that steam. This mechanism led to a high abundance of water vapor in the stratosphere, where it was photodissociated by solar UV photons. Hydrogen atoms hydrodynamically escaped from the atmosphere, thus sealing the fate of water on Venus. These early models, however, did not produce a complete loss of water because of a stratospheric cold trap that set in when the partial pressure of water vapor was still about 20 bar. Subsequent modeling (1988), taking into account moist convection as well as the 30% reduced luminosity of the fainter young Sun, suggested instead that an ocean may have existed at the surface, where temperatures reached the boiling point of water. Through the same feedback and loss mechanisms, this 'moist greenhouse' scenario leads to a complete dry-out of the atmosphere since the stratospheric cold trap does not set in before nearly all of the water has been lost.

(*b*) *Earth*. Earth's greenhouse is very mild in comparison with that of Venus. The surface temperature is only 33 K warmer than the equilibrium temperature. Yet, it has a profound significance for the existence and the development of life on our planet. As on Venus, the main greenhouse gases in EARTH'S ATMOSPHERE are CO_2 and H_2O. Besides a larger heliocentric distance, the case of the Earth reflects three fundamental differences with respect to Venus: (a) a smaller surface pressure (1/90th that of Venus) which limits the total infrared optical depth; (b) the existence of an atmospheric opacity window between 8 and 12 μm, an interval which contains a significant portion of the thermal emission from the surface; (c) a regulation mechanism for the natural abundance of atmospheric CO_2. Under the influence of weathering alone, the amount of CO_2 in the atmosphere tends to decrease owing to its dissolution in pluvial and surface waters and the formation of calcium carbonate in ocean waters. (In present-day oceans, the precipitation of calcium carbonate on the seafloor is performed by shelled organisms, mainly calcareous plankton. In the early stages, it may have been performed by bacterial mats.) On the other hand, tectonic and volcanic activity releases some of the CO_2 sequestered in seafloor carbonates. The two processes counterbalance each other and result in a stable equilibrium state for the Earth climate. This balancing mechanism has in fact been invoked as a way of resolving the 'faint young Sun' paradox: given the lower luminosity of the early Sun, the Earth should have experienced a global ice age during its first 2–3 billion years. The fossil record shows that this was not the case. Indeed, with a surface entirely covered by glaciers, the CO_2 sink due to the weathering cycle would have been much reduced and the amount of atmospheric CO_2 would have increased rapidly. This would have led to a reinforced greenhouse effect that would have raised the surface temperature above the melting point of water ice.

Since the beginning of the industrial era, human activity has resulted in a steady increase of the amount of atmospheric CO_2 and the climatic implications of that have now become a subject of concern. Present modeling suggests that a doubling of the CO_2 abundance would result in a global warming of about 2.5 K (other minor species such as CH_4 and CFCs may also have a significant impact by closing the 8–12 μm window). This estimate takes into account the feedback effect due to water vapor but assumes that the influence of clouds remains constant. The influence of an enhanced cloud cover is two-fold: it increases the albedo, therefore reducing the greenhouse effect, but it also increases the infrared opacity of the atmosphere, thereby increasing the greenhouse effect. It is generally thought that a moderating influence would prevail, although there are also 'worst-case' scenarios in which an enhanced cloud cover leads to a mean surface temperature increase of 4–5 K. In any case, because of the increased melting of the polar ice sheets and the subsequent rise of the mean sea level, a significant warming would have dramatic consequences for the habitability of the coastal areas in many countries.

(*c*) *Mars*. Mars presents a weak greenhouse effect, with a surface temperature only 6 K warmer than the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

equilibrium value. This may be explained by two key factors: (a) the surface pressure is 7/1000th that of the Earth and (b) the atmosphere is too cold to contain a significant amount of water vapor. Even though CO_2 constitutes 95% of the Martian atmosphere—which incidently contains more CO_2 than that of the Earth—the low surface pressure severely limits the infrared optical thickness. This effect is compounded by the lack of water vapor. Furthermore, the Martian greenhouse is regularly perturbed by quasi-periodic global dust storms which are triggered in the southern hemisphere. In the case of storms encircling the planet and lasting one season, calculations show that the mean seasonal daily insolation is dramatically reduced (by factors of tens to thousands depending on the latitude and the season). This may explain, for instance, the observed 5 K decrease of the mean diurnal surface temperature at the Viking 1 landing site after the 1976–7 global dust storm. However, the influence of dust loading in the atmosphere is not a one-way street. Evidence from Viking and Mars Global Surveyor also indicates atmospheric heating near the surface, even for large amounts of dust (see MARS: ATMOSPHERE).

The Red Planet has not always been so cold and dry. Surface morphological features observed by various orbiting spacecraft can only be explained by the presence of liquid water at some remote stage of the Martian evolution. It is currently assumed that a greenhouse effect generated by a thick CO_2 atmosphere (1000–5000 hPa) created a much warmer and wetter climate on the early Mars. This stage—during which life may have developed—probably lasted a few million years, until most of the CO_2 was converted into carbonates by the weathering cycle. The lack of tectonic activity on Mars prevented any recycling of these carbonates into carbon dioxide, thus quenching the atmospheric greenhouse. In principle, it would be possible to transform Mars into a habitable planet (terraforming) provided that most of the carbonates are converted back into atmospheric CO_2. Other necessary conditions for life would be production of O_2 by plants, protection from CO_2 toxicity by a buffer gas such as N_2 and shielding from solar UV radiation by CFCs.

(*d*) *Titan*. Titan, the largest satellite of Saturn and the only satellite in the solar system with a substantial atmosphere (1500 hPa surface pressure), represents an interesting case of greenhouse effect which is intermediate between those of Venus and the Earth. Despite the large distance from the Sun, the amplitude of the atmospheric greenhouse effect is relatively large. The main greenhouse gases are N_2 (constituting more than 94% of the atmosphere), CH_4 (less than 6%) and H_2 (0.1%). The abundance of the last of these constituents is critical since its absorption affects the transparency of the unique opacity window located between 16 and 25 μm. It is worth noting that the warming of the surface would be about 10 K greater were it not for the anti-greenhouse effect caused by a thick stratospheric haze.

Solid-state greenhouses

Io and Triton are very different in terms of surface composition and internal structure. One feature they may have in common, however, is the presence of a solid-state greenhouse at their surface. Io's surface is almost entirely covered with a SO_2 porous frost. From the positions of the near-infrared absorption bands due to this frost, it has been argued that a reverse temperature gradient exists at the subsolar point, i.e. the subsurface at 10–20 cm depth would be some 40 K warmer than the surface (at 85 K). However, other sulfur bands (SO_2 and S_8) measured by Voyager in the mid-infrared would argue against that conclusion. On Triton, the evidence is more indirect. Voyager images showing plumes of dark material extending several kilometers into the tenuous atmosphere suggest that local solid-state greenhouse warming of the subsurface CH_4–N_2 ice leads to rapid sublimation, resulting in a geyser-like phenomenon.

Bibliography

Goody R M and Yung Y L 1989 *Atmospheric Radiation—Theoretical Basis* (Oxford: Oxford University Press)

Kasting J F 1988 Runaway and moist greenhouse atmospheres and the evolution of Earth and Venus *Icarus* **74** 472–94

Matson D L and Brown R H 1989 Solid-state greenhouses and their implications for icy satellites *Icarus* **77** 67–81

McKay C P, Pollack J B and Courtin R 1991 The greenhouse and antigreenhouse effects on Titan *Science* **253** 1118–21

McKay C P, Lorenz R D and Lunine J I 1999 Analytic solutions for the antigreenhouse effect: Titan and the early Earth *Icarus* **137** 56–61

McKay C P, Toon O B and Kasting J F 1991 Making Mars habitable *Nature* **352** 489–96

Régis Courtin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Greenstein, Jesse Leonard (1909–)

Born in New York City, worked at the California Institute of Technology at the time of the start of the 200 in telescope, and seized the chance to apply his training in spectroscopic theory to observation, studying the interstellar medium, abundances of the elements and peculiar stars. He has made detailed studies of white dwarf stars, determining their masses, luminosities, temperatures, compositions, gravitational redshifts and motions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gregorian Telescope

A type of reflecting telescope originally designed in 1663 by the Scottish mathematician, James Gregory. It utilizes a concave paraboloidal primary mirror and a concave ellipsoidal secondary mirror that is located outside the focus of the primary. Light reflected from the secondary travels back down the telescope tube, through a central hole in the primary, to the eyepiece. Because the secondary mirror causes light rays to converge at a smaller angle than the rays reflected directly from the primary, it increases the effective focal length of the instrument, so enabling a telescope of long effective focal length to be contained within a relatively short tube. The Gregorian optical system produces an upright image but suffers from having a very small field of view. Gregorian instruments, constructed by instrument-makers such as James Short (1710–68), became popular during the eighteenth century largely because their concave secondaries were easier to make than were the convex secondaries of the broadly similar Cassegrain instruments. Several modern solar telescopes are based on the Gregorian design, since it produces a real image of the Sun which can be occulted by a disk (to reduce scattered light or extract heat).

See also: Cassegrain telescope, field of view, focal length, mirror, reflector, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gregory [Gregorie], James (1638–75)

Scottish mathematician and optician, born in Aberdeen. Gregory described in *Optica Promota* a design (which he never realized) for the first practical reflecting telescope in which a perforated primary concave parabolic mirror converges the light to the focus of a concave ellipsoidal secondary mirror. The light is reflected back to the ellipsoid's second focus behind the main mirror. A real image is formed at the common focus of the two mirrors (by contrast to the virtual image in a Cassegrain telescope). The *Gregorian telescope* (as the design came to be called) is used in solar instruments where the light and heat of the Sun can be drawn out of the instrument by a blockage placed at this position. (For the first Gregorian telescope made, *see* JOHN HADLEY, ROBERT HOOKE.) After a time in Padua, Gregory returned to Scotland (St Andrews, Edinburgh) and worked on pure mathematics. He invented a brilliant method to determine the scale of the stellar universe, by comparing Sirius (as an assumed typical star) with the Sun (assumed the same). Waiting until a planet became the same brightness as Sirius, he calculated the fraction of light that, at that time, the planet intercepted and reflected from the Sun. He was able to estimate Sirius's distance in units of the Sun's distance, arriving at a number of 83 190 AU, which he realized was a lower limit because of limitations in the scale of the solar system then known (1668). When in 1685 NEWTON repeated Gregory's calculation he concluded that Sirius was at 1 million AU.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grimaldi, Francesco Maria (1618–63)

Italian astronomer and optician, born in Bologna, became a Jesuit. Prepared a map of the Moon that Riccioli used to assign the currently used names to its principal features. Discovered diffraction of light at small apertures.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Groombridge 1830

This star (abbr. Gr 1830), despite its being at the limit of naked-eye visibility (apparent magnitude 6.42), is very well known due to its large proper motion. It is identified by its number in the *Catalogue of Circumpolar Stars*, compiled by the wealthy amateur astronomer Stephen Groombridge (1755–1832) from observations made at his private observatory in Blackheath, approximately one kilometer from the Royal Observatory at Greenwich. Groombridge used a fine transit circle of 3.5 inches aperture with circles 4 feet in diameter, made for him in 1806 by Edward Troughton (1753–1835). This excellent instrument was the first successful transit circle used in Britain, and was one of the inspirations of G B Airy (1801–92), the seventh Astronomer Royal, to design his own great transit circle which came into use in 1851. Groombridge suffered a stroke in 1827 and was unable to complete the reductions of his observations. These were completed under Airy's supervision and the catalog, containing the positions of 4243 stars in the declination band 38–90°, was published in 1838. 'Groombridge' remained the standard reference for circumpolar stars for over 150 years; a revised edition was published in 1905.

Groombridge 1830 is a yellow main sequence star, spectral type G8Vp, apparent magnitude 6.42, situated about 16° south of γ Ursae Majoris in the Plough. It lies at a distance of 30 light-years, and has a parallax of 0.109″. Its absolute magnitude is 6.6. The significance of Groombridge 1830 was first recognized in 1842 by F W A Argelander (1799–1875), Director of the Bonn Observatory, who analyzed observations of its position made by various observers since 1794 and demonstrated that it had the then largest-known proper motion of any star. Even today its proper motion of 7.058″ per annum is the third largest known, surpassed only by Barnard's Star and Kapteyn's Star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Groups of Galaxies

Gravity creates a huge variety of cosmic structure, but most galaxies are found in groups of no more than a few dozen members. Groups range from the satellite systems of giant galaxies to loose associations a few times denser than their surroundings to compact cores of rich galaxy clusters. The common thread linking these examples is that the galaxies making up a group interact more with each other than they do with the rest of the universe; thus a group is a dynamical unit.

Groups are important for the evolution of galaxies and of large-scale structures. GALAXY FORMATION appears to be a drawn-out process, involving the collapse of primordial perturbations, accretion of gas and dark matter, outright merging of distinct objects and outflows of gas enriched by supernovae. Most galaxies conduct these transactions in group environments. However, group environments are unstable; while the galaxies in a group are forming, the group itself may be separating out from the cosmic expansion, collapsing under the influence of gravity, accreting new members and finally merging with other groups to build clusters and superclusters.

We can study the present outcome of these ongoing processes in our immediate vicinity and glimpse some earlier stages at high REDSHIFT. However, reconstructing the evolution of groups is a difficult problem. Two powerful tools which complement optical studies are computer simulations and multiwavelength observations. Simulations model the key dynamical ingredients of groups, including the DARK MATTER IN GALAXIES which is otherwise inaccessible. Observations at radio and x-ray wavelengths show how intergalactic gas responds as groups evolve.

Evolutionary stages

Figures 1 through 4 show four groups, ordered by stage of development. Group evolution begins when a bound configuration of several galaxies collapses out of the Hubble flow; the LOCAL GROUP (LG) is now at this stage, with the Milky Way (MW) and M31 approaching each other for the first time (figure 1). One of our nearest neighbors, the M81 group (figure 2), illustrates a more advanced stage; as the HI image shows, three galaxies are linked by a complex structure indicating that they have already undergone at least one passage. STEPHAN'S QUINTET (figure 3), a compact group, contains an extended, possibly tidal feature in neutral hydrogen and a central cloud of hot gas visible in x-rays; this gas may have been heated by a fast, interpenetrating encounter between two galaxies. Finally, V Zw 311 (figure 4) is actually the core of cluster Abell 407; it is likely that several of these galaxies will merge over the next $\sim$1 Gyr.

These four examples represent not only different evolutionary stages but also different evolutionary *tracks*. For example, the LG will become more compact as the MW and M31 draw closer, possibly evolving into a system like the M81 group, but probably not attaining the dramatic status of Stephan's Quintet, let alone V Zw 311.

Collapse

The LG (figure 1) is probably the best place to study a collapsing group. Most of its luminosity is associated with the MW and M31, which are currently separated by $\sim$0.73 Mpc and approaching each other at 120 km s^{-1}. A simple model treats the main galaxies as point masses moving along a linear orbit—an ellipse of zero width. In this approximation, the two galaxies coincided at the big bang $t_0 \simeq 13$ Gyr ago, separated smoothly with the Hubble flow, reached a maximum separation of $\sim$1 Mpc some 5 Gyr ago and are now falling together. To account for this history, the LG's total mass must be about $4 \times 10^{12} M_\odot$ and its mass-to-light ratio $M/L_V \simeq 100$. This is several times most estimates of the MW's M/L ratio, a sign that unseen dark matter is important in extragalactic dynamics.

Thus it seems that much of the LG's mass resides in structures larger than individual galaxies. Cosmological N-body simulations show that gravity deforms a smooth distribution of dark matter with a plausible spectrum of density fluctuations into a complex web of 'pancakes' and filaments; this web constantly evolves as ever-larger scales break away from the Hubble expansion. Halos of dark matter form at the intersections of filaments and grow as mass flows along the filaments. In this picture, the MW and M31 are presumably linked by a major filament and are 'sucked together' as this filament collapses.

Spins and swings

In gravitational clustering, galaxies acquire their spins from tidal torques. At early times, the material which eventually forms a galaxy has an irregular shape, and as density fluctuations grow this material feels a torque due to the tides of other proto-galaxies. Groups seem natural places to apply this scenario; for example, it has been argued that the MW and M31 are mutually more or less edge-on because each provided the tidal torque which spun up the other. The further assumption that the total spin and orbital angular momentum of the LG is zero then yields a definite prediction for the M31's transverse motion with respect to the MW.

This simple calculation is less plausible on closer inspection. First, in a tally of angular momentum, spins are less important than satellite orbits; for example, the LARGE MAGELLANIC CLOUD's orbit about the MW has several times the angular momentum of the MW's spin, and the same may be true for M31 and M33. Second, the external torque on the MW–M31 binary due to the nearby galaxies shown in figure 1 is sufficient to give the pair a transverse velocity of at least $\sim$40 km s^{-1}. This orbital swing completely dominates the total angular momentum of the LG; clearly the LG's dynamical isolation is very imperfect. It is possible that MW and M31 set each other spinning, but the reflex torque on their relative orbit is tiny compared with the torques due to surrounding galaxies. Tests of the tidal torque hypothesis based on auditing the angular momentum of the LG seem doomed to founder on this complication.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

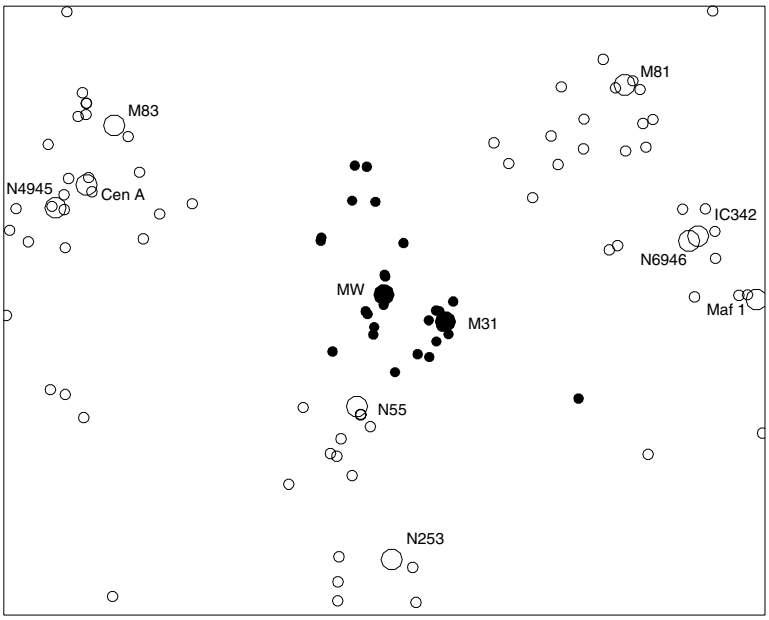


Figure 1. The LG (filled circles) and other nearby galaxies (open circles). Luminous galaxies (large circles) are labeled. (Courtesy B Tully.)

Least-action models

To investigate the effects of surrounding galaxies in greater detail requires more than two free-falling bodies. An elegant approach minimizes the action of a system of point-like galaxies subject to mixed boundary conditions: peculiar velocities vanish as $t \to 0$, and all galaxies reach their observed positions at the present. This approach assumes that each galaxy's trajectory and gravitational field are adequately reproduced by a single particle. However, it also admits a rather strong test; given the present masses and redshift-independent distances of galaxies, it predicts their redshifts.

When applied to the LG this method yields impressive agreement between predicted and observed redshifts. Plausible solutions, in which the MW and M31 approach each other for the first time, assign the LG a total mass consistent with the simple two-galaxy estimate above. However, the transverse velocity of M31 depends sensitively on the volume simulated, growing to several hundred km s^{-1} when the other galaxies in figure 1 are included; indeed, it is possible that yet more distant galaxies may further modify the solutions. These results, if correct, completely rewrite the past and future of the LG; the MW and M31 are no longer partners in a dance lasting aeons but 'ships passing in the night' at several times their escape velocity. PROPER MOTIONS for M31 and other LG galaxies would settle the issue.

Virial equilibrium

Simple models assume that bound groups reach a quasi-equilibrium state in which galaxies have isotropic orbits with random phases. This state of randomized grace results from scattering of galaxies by each other and by masses outside the group. However, when do these scatterings occur? If a group is isolated and remains fairly spherical then its constituent galaxies are not deflected from their radial trajectories until the group has collapsed to a small fraction of its maximum radius. In this case the collapse is fairly violent, and the new-born group first reaches equilibrium at ∼200 times the mean cosmic density. At the opposite extreme in which a nascent group is strongly influenced by surrounding objects the collapse is gentle and the group attains equilibrium at only ∼30 times the mean density. This is basically the same set of alternatives considered in the discussion of the LG.

Once a group is in equilibrium, a time-averaged version of the *virial theorem* applies: $2\langle T\rangle + \langle U\rangle = 0$, where angle brackets denote time averages and T and U are the kinetic and potential energy of the group, respectively. Of course, these time-averaged quantities cannot be determined observationally; instead, we observe an ensemble of groups and use orbital randomness to justify taking ensemble averages. Observable quantities include the mean harmonic radius of the group, $r_{\rm h}$, and the line-of-sight velocity dispersion, σ. These parameters may be used to check that group has actually had time to reach virial equilibrium. The collapse time for an isolated, spherical group is approximately the Keplerian period of a particle at the edge of the group, or $\sim 2.9 t_{\rm c}$, where

$$t_{\rm c} = \frac{2 r_{\rm h}}{\sqrt{3}\sigma} \qquad (1)$$

is the *crossing time*. Starting from the big bang, it takes a group about one collapse time to reach equilibrium, so a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

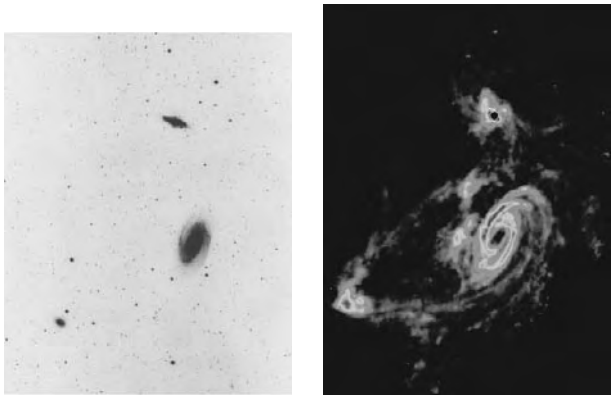


Figure 2. The M81 group in (left) visible light and (right) neutral hydrogen. (Courtesy M Yun.)

reasonable criterion might be $3t_c < t_0$. This limit is stricter than necessary for groups which are strongly influenced by neighboring masses, since such groups attain equilibrium with little or no collapse.

In view of the evidence for dark matter in the LG and in rich galaxy clusters, much effort has been devoted to measuring the masses of bound groups. For groups in equilibrium, the virial theorem provides an estimate of the dynamical mass,

$$M_{\rm d} = \frac{6\sigma^2 r_{\rm h}}{G}. \qquad (2)$$

Group mass-to-light ratios obtained using this approach generally show a wide scatter about a median value of several hundred. Some of this scatter represents fluctuations around virial equilibrium. However, a far larger and more systematic uncertainty arises in assigning galaxies to groups; if $M_{\rm d}$ is to provide meaningful group masses, the procedures used to identify groups must be carefully calibrated.

Catalogs and parameters

In the days before redshift surveys, groups were identified by noting galaxies with similar positions, apparent distances, and other signs of association. This subjective process produced accurate membership lists for nearby groups (e.g. figure 2), but contamination by foreground and background galaxies became much more serious at larger distances. As more redshifts became available it became clear that some 'traditional' groups extend along the line of sight for many Mpc; in such cases, a virial analysis is useless since σ reflects Hubble expansion instead of internal dynamics.

Redshift surveys down to a fixed limiting magnitude are probably the best available starting point in constructing group catalogs. However, redshifts reflect peculiar motions as well as Hubble expansion, and this distorts the galaxy distribution deduced from redshift surveys. Consequently, algorithms for group finding cannot treat radial and transverse directions on the same footing. Most group-finders use a 'friends of friends' algorithm which links two galaxies if their projected separation and line-of-sight velocity difference are less than some cutoff values; each set of galaxies linked either directly or via intermediaries defines a group.

The velocity cutoff strongly influences the resulting catalogs; if it is too high then unrelated galaxies along the line of sight may be linked together, while if it is too low then *bona fide* groups may be broken up. The same velocity information is used to define groups and to calculate their velocity dispersions, so it is no surprise that dispersions are also sensitive to the velocity cutoff. Tests run on simulated redshift surveys derived from cosmological N-body experiments offer some guidance in selecting distance and velocity cutoffs that produce reasonable group catalogs.

A typical catalog generated in this manner links about half of all galaxies into groups. Most groups have three or four members, while a small fraction are much richer. The median density of such groups is about 100 times the background density, and about half the groups have crossing times short enough to be virialized. The mass-to-light ratios derived for these groups are roughly consistent with the value obtained for the LG.

Dynamical evolution

It has been convenient so far to treat groups as systems of point-like galaxies. However, in reality galaxies are not points, and groups may have a good deal of dark matter not associated with individual galaxies. This can have dramatic consequences. For example, virial equilibrium is at best a temporary condition; in the long run, an isolated group cannot attain true equilibrium until all its galaxies coalesce to form a single group remnant.

Galactic encounters

Thanks to their finite sizes, galaxies are subject to tides. Galaxies in close proximity are simultaneously stretched and squeezed by tidal forces. DISK GALAXIES respond by forming dramatic bridges and tails of stars and gas which reflect the effects of tides on rotationally supported disks; ELLIPTICAL GALAXIES, supported by random motions, may exhibit distended envelopes and off-center isophotes. The explanation of such features as relics of tidal interactions was a key step in understanding the peculiar morphologies of colliding galaxies.

Groups of galaxies are likely sites for tidal interactions. Simple estimates and detailed simulations indicate that in a typical group there is roughly one close encounter per crossing time. A curious feature of tidal interactions is that slower encounters do more damage; groups have relatively low velocity dispersions, so most encounters in groups are slow enough to produce rather spectacular effects.

One example of tidal interactions is shown in figure 2. Even before the HI observations there were signs that the two brightest galaxies in this group were interacting; evidence included the 'grand design' spiral of M81 and an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

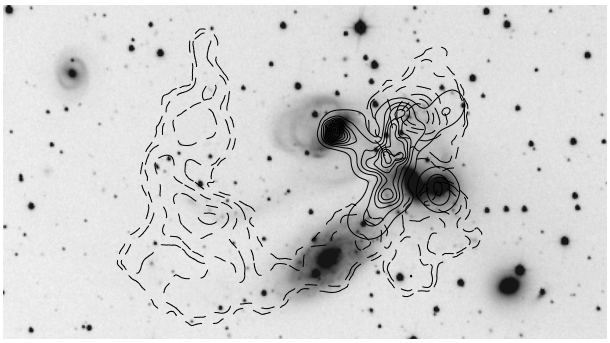


Figure 3. Stephan's Quintet, a compact group of galaxies. Plotted on top of the optical image (courtesy M Moles) are thin contours of HI emission (courtesy B Williams) and thick contours of x-ray emission (courtesy H Ebeling). The large spiral galaxy at the bottom is a foreground object.

'explosion', now recognized as a burst of star formation, in M82. The HI maps reveal complex filaments linking these galaxies as well as the third most luminous member, NGC 3077. These structures appear to be bridges and tails extracted from the extended gas disks of the individual galaxies by tidal forces during a fairly recent triple encounter; this encounter may represent the initial collapse of the M81 group.

Tidal interactions are common in COMPACT GROUPS OF GALAXIES where three or more galaxies are seen in near proximity. Besides tidal signatures similar to those already described, many disk galaxies in such groups have asymmetric rotation curves or other unusual kinematic features. Such kinematic peculiarities are among the first signs of a tidal interaction; bridges and tails develop later as perturbed stars drift away from their parent disks.

A compact group with a rather complicated history is shown in figure 3. This is Stephan's Quintet, discovered in 1877. As often happens, not all the galaxies are the same distance; in this case, the largest spiral is a foreground object, while the other five galaxies, including the small spiral on the left, are associated. Of these five, two exhibit tidal tails and other signs of interactions. However, tides are not the only forces at work here—velocity fields of both neutral and ionized hydrogen reveal *two* components separated by at least 700 km s^{-1}, while x-rays show a cloud of hot gas at the center of the group. The data suggest that gas is heated by a high-velocity shock between two distinct streams of tidal material, with further energy input from localized formation of massive stars. If so, at least two tidal encounters are required, since the debris of a single passage generally separate at roughly the encounter velocity. Moreover, the large shock velocity implies a deep potential well associated with one or more massive dark halos. Finally, the hot gas responsible for the x-ray emission points to a long-term change, since such gas, once heated, is slow to cool off.

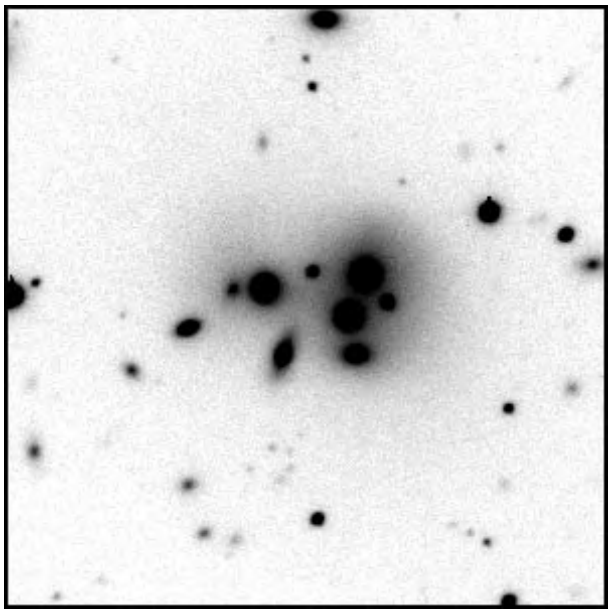


Figure 4. 'The once and future cD', V Zw 311, is a complex of galaxies at the center of a poor cluster (courtesy D Schneider).

Orbital decay

Another consequence of extended mass distributions in groups is the decay of galactic orbits. Tidal encounters between galaxies cause significant orbital decay—roughly speaking, the work required to raise tidal features of visible stars and of dark matter comes from the relative motion of the participants. Two equally matched galaxies whose halos interpenetrate after falling from a great distance will never return to anything like their maximum separation, but will pass even closer the next time around, and merge after a few orbits. This decay is slower in unequally matched encounters, but no less inexorable; it represents a one-way process, a randomization of the ordered relative motions of two galaxies. A similar process occurs when a galaxy orbits within the common halo of a larger group. In the limit where the group halo is much more massive than the galaxy this process can be described by several analytic approximations, none perfect yet all predicting that the galaxy will spiral inward after some M/m orbits, where M is the mass of the group and m is the mass of the galaxy.

Orbital decay can have dramatic consequences in extremely compact groups like the one in the center of V Zw 311 (figure 4). Here we see nearly a dozen early-type galaxies embedded in a common envelope of comparable total luminosity. The velocity dispersion of this group is $\sigma \simeq 600$ km s^{-1}, while the brighter group members have velocities within $\sim$200 km s^{-1} of the mean. Assuming this system is bound, its virial mass is about 10 times the total mass of the visible galaxies, and the unseen material is probably more extended than the luminous envelope. If the brighter galaxies are now on roughly circular orbits,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

they will spiral toward the center and presumably merge in ∼1 Gyr.

While objects as extreme as V Zw 311 are fairly rare, orbital decay occurs in both loose and compact groups. Numerical simulations in which galaxies and dark matter are represented using many thousands of particles are the most reliable way to follow the decay of galactic orbits. Such simulations show that galaxies spiral inward about as fast as predicted by the analytic models. The orbital decay timescale depends on the initial set up, being shorter if most of dark mass is associated with individual galaxy halos and longer if most of the dark mass is invested in a common group halo. However, unless the initial conditions are somehow fine-tuned, the orbits of massive galaxies shrink quite noticeably with each revolution.

In view of this relentless evolution, one might wonder whether it is correct to estimate group masses using the virial theorem. Numerical studies show that the systematic non-equilibrium effects are fairly modest; on the average, galaxies move about as fast as they would if their orbits were not decaying, and the virial theorem yields consistent mass estimates. However, as galaxies spiral in they explore less and less of the group's potential well, and any mass which lies beyond the region they sample will not be detected by a virial analysis.

Mergers

When galaxies spiral toward the center of a group it also becomes more and more likely that they will capture each other and merge. Merging can be described as a combination of orbital decay and tidal disruption. As two or more galaxies undergo successively closer passages, they are tidally stripped from the 'outside in'; most of the stripped material remains bound to the merging system and forms the body of the resulting merger remnant.

The outcome of a merger depends on the participants. Much attention has focused on the hypothesis that elliptical galaxies form by mergers of spiral galaxies. This suggestion is supported by observations of 'twin-tailed' merger remnants which have luminosity profiles and kinematic properties resembling bright elliptical galaxies and by numerical experiments which detail the mechanics of such catastrophic transformations. Together with the companion hypothesis that violent encounters can trigger bursts of star formation, merging succeeds in linking normal and peculiar galaxies into a coherent evolutionary scheme (see GALAXIES: INTERACTIONS AND MERGERS). Yet it is unlikely that all ellipticals resulted from mergers of equal-mass spirals. Some elliptical galaxies may have had multiple bursts of star formation, while others seem more thoroughly 'scrambled' than typical spiral–spiral mergers. If these objects originated in multiple mergers, they probably did so in group environments.

There are many groups with recent or ongoing mergers. One example is the loose, spiral-rich group containing Cen A, a peculiar elliptical galaxy with an active nucleus (see NGC 5128/CENTAURUS A). The dark bisecting dust lane, subtle 'shells' superimposed on the luminosity

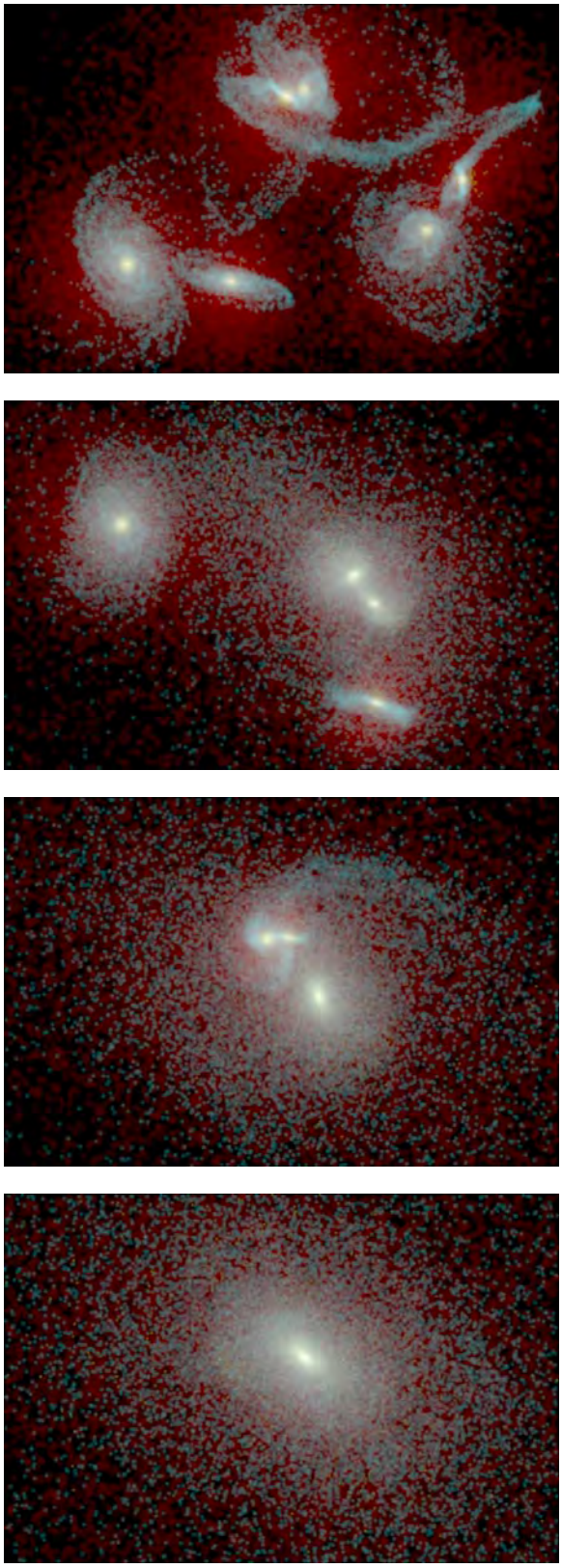


Figure 5. Numerical simulation of a compact group. These frames span an interval of ∼3 Gyr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

profile, unusual kinematics revealed by planetary nebulae and irregular structure of its extended envelope all suggest that Cen A swallowed a good-sized gas-rich companion ∼1 Gyr ago. Another example is the small group of four galaxies including NGC 3921, a classic twin-tailed remnant of a recent merger between two disk galaxies of comparable mass. A third example is Hickson Compact Group 95, in which two smaller spirals appear to be simultaneously merging with a larger elliptical galaxy.

In the case of V Zw 311 (figure 4), the extended envelope surrounding the galaxies probably represents material stripped in the early stages of the merging process. Further stripping as the galaxies sink toward the center of the group will liberate more material, building the merger remnant from the outside in. Once the larger galaxies finish merging, V Zw 311 will contain a very luminous central galaxy with an extended envelope; the expected parameters of this object suggest that we are witnessing the formation of a 'cD' galaxy similar to those which dominate the central regions of many galaxy clusters.

A numerical simulation illustrating the later stages of orbit decay and merging is presented in figure 5. This experiment started with a compact near-equilibrium configuration of six disk galaxies, each surrounded by a modest dark halo. The first frame shows the system shortly before two pairs merge; note the twin-tailed morphology of the pair at the top. Later frames show a further merger of the two remnants, a violent interaction between the composite remnant and the two remaining disk galaxies, and the final remnant, which resembles an elliptical galaxy. Scaling the initial disk galaxies to the MW, the group starts with a diameter of ∼120 kpc and takes about 3 Gyr to completely merge.

In absolute terms, the merger rate in a loose group is a good deal lower than in a compact group. However, this difference is largely due to the long crossing times of loose groups; within broad limits, there is about one merger per crossing time in loose and compact groups alike. Since loose groups are a good deal more common than compact ones, they probably dominate the net merger rate. Demographic estimates show that while galaxies in compact groups are ∼20 times more likely to undergo violent interactions and mergers, only one merger in four occurs in a compact group. The timescale for a typical loose group to merge is several times the age of the universe; nonetheless, a modest fraction of galaxies in loose groups have probably undergone multiple mergers.

Where are the relics of group mergers to be found? Some compact groups seem to be truly isolated, and the remnants of such groups should appear as bright, isolated elliptical galaxies. A small number of possible fossils have now been identified; these galaxies typically have extended halos of hot, x-ray-emitting gas similar to those seen in compact groups and poor clusters. However, while ongoing merging in groups has been building up the population of merger remnants, gravitational clustering has been incorporating these remnants into new groups and larger structures. Even today we find groups 'caught in the act' as they fall together and coalesce to form clusters. Once incorporated into clusters, galaxies are less likely to merge since their encounter velocities are now too high for tidal capture; if cluster ellipticals formed by merging, they can only have done so before the clusters themselves had collapsed. Elliptical galaxies in clusters may have formed in dense groups at redshifts $z \sim 2$; if so, compact groups are the low-redshift analogs of the birthplaces of cluster ellipticals.

Bibliography

Barnes J E 1989 Evolution of compact groups and the formation of elliptical galaxies *Nature* **338** 123–6

Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton: Princeton University Press)

Nolthenius R and White S D M 1987 Groups of galaxies in the CfA survey and in cold dark matter universes *Mon. Not. R. Astron. Soc.* **235** 505–30

Peebles P J E 1994 Orbits of the nearby galaxies *Astrophys. J.* **429** 43–65

Raychaudhury S and Lynden-Bell D 1989 Tides, torques and the timing argument *Mon. Not. R. Astron. Soc.* **240** 195–218

Joshua E Barnes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grubb, Thomas (1800–78) and Grubb, Howard (1844–1931)

Irish telescope makers, based in Dublin, then in St Albans, England. Thomas was born in Waterford and set up a machinery business. After making a 9 in telescope for himself, he made many giant reflecting telescopes commercially. One of his first contracts was for a Cassegrain reflector for ROMNEY ROBINSON of Armagh Observatory, which introduced mechanical design features of significance for the future, including the viewing position of the user (through a hole in the center of the main mirror, not from the top of the telescope), the mirror support system and its clock-driven equatorial mounting. With this successful start, the firm made telescopes that were exported all over the world, to Madras, Madrid, Mecca, Melbourne and Mississippi, to list just one letter of the alphabet. Howard, born in Dublin, joined the business and diversified to a range of optics. In 1925 the firm was taken over by CHARLES PARSONS, becoming the firm of Grubb-Parsons, and moving to Newcastle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gruithuisen, Franz von Paula (1744–1852)

German astronomer, became professsor at Munich, asserted to a skeptical astronomical community that he had seen a city, a forest and artificial structures on the Moon, near Sinus Aestuum. Proposed the impact theory of lunar craters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grus

(the Crane; abbrev. Gru, gen. Gruis; area 366 sq. deg.) A southern constellation which lies between Piscis Austrinus and Tucana, and culminates at midnight in late August. It was first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7. On later globes Plancius gave it the now obsolete name of Phoenicopterus (the Flamingo).

A small but not inconspicuous constellation, the brightest stars in Grus are α Gruis (Al Na'ir), magnitude 1.7, β Gruis, magnitude 2.1, and γ Gruis, magnitude 3.0. There are four other stars brighter than fourth magnitude. Interesting objects include the naked-eye optical double stars δ Gruis, which has yellow (G6–8) and red (M4.5) components, magnitudes 4.0 and 4.1, separation about 13′, each of which is a binary, with components of magnitude 12.8 and 8.8, separation 5.6″ and 60.6″, respectively, and μ Gruis, which has yellow (G8) components, magnitudes 4.8 and 5.1, separation about 17′, the former of which is a close binary with a component of magnitude 6.7, separation 0.15″.

There are no bright star clusters, nebulae or galaxies in Grus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Guide Telescope

A second telescope, attached to, and aligned parallel to, the main telescope, which is used to monitor the position of the object of interest in the field of view while the main telescope is being used to obtain a long-exposure image of the object. Because of imperfections in the telescope mounting, flexure of the telescope or its mounting and imperfections in the drive mechanism, the object that is being observed will tend, to some extent, to wander around the field of view during a long exposure. To compensate for this, the telescope must be guided (i.e. it must be adjusted, continually or at regular intervals, to ensure that the object remains in the center of its field of view). Traditionally, this is done by the observer adjusting the telescope controls to keep the object at the center of the field of view of the guide telescope—a tedious and time-consuming process. A widely used alternative is the autoguider, an electronic device that senses when the object drifts from its correct position and generates a signal that causes the slow-motion drives to restore the main telescope to the correct position. The autoguider may be attached directly to the guide telescope or be used to monitor a region of the field of view of the main telescope close to the object of interest. If the object of interest is too faint to be seen in the guide telescope, the guide telescope may be adjusted to focus on a nearby brighter object (guide star) while the main telescope remains pointed at the object being imaged. This process is called off-axis guiding.

In smaller sizes of telescope the role of guide telescope may have to be fulfilled by the finder.

See also: field of view, finder, telescope mountings.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gum Nebula

A very large, near-circular emission nebula, approximately 36° in diameter, in the constellations Puppis and Vela. The largest known nebula in the sky, it was discovered by the Australian astronomer C S Gum (1924–60), and is believed to be an ancient supernova remnant, with an age exceeding a million years. It is a convoluted mass of nebular wisps and loops, many of them very faint, but there are also numerous brighter parts. Its distance has been estimated at 1300 light-years, indicating that the nebulosity is approximately 840 light-years across. Within one of its brightest regions both the brightest-known O-type star ζ Puppis (spectral type O5f) and the brightest Wolf–Rayet star γ^2 Velorum (type WC8), are found.

The much more recent Vela pulsar and supernova remnant also lie within the Gum Nebula, which for many years has rivalled the Crab Nebula in interest for astrophysicists.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gum, Colin (1924–60)

Australian radioastronomer, mapped the southern sky for radio sources and emission nebulae, and discovered the *Gum Nebula* in the Vela Puppis region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

H I Region

A cloud of interstellar matter consisting predominantly of neutral hydrogen atoms. 'Neutral' means not ionized—each hydrogen atom consists of a proton and its lone electron; this is what is implied by 'H I' (pronounced 'H one'). In H II ('H two') regions, the hydrogen is ionized, each atom having lost its electron. In mass and volume, neutral hydrogen accounts for about half of all interstellar matter. A typical H I region is 15–20 light-years across, and contains 50 solar masses of hydrogen at a temperature of around 100 K. H I regions emit no visible light, but are detectable by their radio emission at a wavelength of 21 cm (frequency about 1420 GHz). This so-called 21 cm radiation is what first allowed the form of our Galaxy to be mapped in the mid-twentieth century: neutral hydrogen is present in the disk of the Galaxy, concentrated in the spiral arms, and 21 cm radiation is little affected by intervening interstellar matter. H I regions have also been detected in other spiral galaxies, and in irregular galaxies, though they are largely absent from elliptical galaxies.

Compare: H II regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

H II Regions

The birth of stars occurs in the densest parts of molecular cloud complexes. The radiation fields and stellar winds of newly formed stars excite and stir the surrounding gas. Low-mass stars ($M \leq 5M_{\odot}$) affect only relatively small volumes of the cloud. The stronger radiation fields and winds of more massive stars ($M \geq 10M_{\odot}$) can excite, ionize and excavate holes in the parent molecular cloud. A few massive stars can clear out the dense gas from the star-forming regions, exposing the recently formed group of stars. The ultraviolet stellar photon flux ionizes the surrounding gas, creating the beautiful and complex NEBULAE known as H II regions (or regions of ionized hydrogen).

The stars that create them are among the brightest, most massive and shortest-lived stars in galaxies. They have surface temperatures above 3×10^4 K and are brighter than 10^4 Suns, although their masses are only 20–100 times as great. They end their lives by exploding as SUPERNOVAE less than about 10 million years after birth, a time short compared with the age (about 4 500 million years) of our solar system. These massive stars are generically designated as OB stars and are excellent tracers of regions undergoing active STAR FORMATION both in our Galaxy and in external galaxies.

Physical properties

The spectra of H II regions at optical wavelengths are easy to recognize and contain strong emission lines of H, C, N, O and S. These lines provide us with information about the chemical abundances, densities and temperatures of the photoionized plasma. Additionally, the shapes and widths of these emission lines provide information about the spatial distributions and velocities of the emitting ions. This plasma has elemental abundances similar to those found in the solar system, and by terrestrial standards is very hot and dilute. The gas temperature is about (8–10) $\times$ 10^3 K and the average number densities are in the range 10^2–10^4 particles cm^{-3}, far below the densities in the best vacuum ever achieved on Earth. However, H II regions by terrestrial standards are huge and can contain far more material than in the Sun (cf table 1).

Not all H II regions are visible to optical telescopes. A fraction of these objects is buried deep in the parent molecular clouds and can be detected only at long wavelengths using RADIO TELESCOPES or infrared detectors. They are called ultracompact H II regions (UCHIIRs). The fraction of H II regions which these represent is generally believed to be about 20% of the total number of H II regions. They were discovered as recently as the 1960s and more than 200 are at present known. Thanks to the fast growth in radio and IR techniques, our knowledge of these objects has increased enormously in recent years. This is also true for optically observed regions, particularly after the launch of the Hubble Space Telescope and the advent of the new generation of large optical telescopes. Advances in optical, IR and radio studies together have provided us with a wealth of unprecedented detail about these fascinating objects.

Classification

H II regions in the Milky Way are classified into three main categories by their average sizes. In order of increasing size they are (i) UCHIIRs, (ii) compact H II regions and (iii) extended or 'classical' H II regions. There is a recently defined class, hypercompact (or superultracompact) H II (SUCHII) regions. They are detected at radio wavelengths and only a handful of them are at present known. Their observed properties indicate that they are denser and more compact than UCHIIRs and may be the youngest H II regions yet found.

The larger H II regions can be found in external galaxies. These include giant and supergiant H II regions. Table 1 lists these six categories along with their characteristic physical parameters including N_*, which is a (very) rough guide to the number of massive stars which excite them. This has been estimated by assuming that each star produces 10^{48} ionizing photons s^{-1}, a number we will use subsequently.

In external galaxies, the population of H II regions delineates the distribution of star-forming regions. Giant and supergiant H II regions are not single objects but a conglomerate of many individual H II regions occupying a large fraction of the volume of their parental giant molecular clouds. These objects trace the locations where vigorous star-formation activity, or a 'starburst', is taking place in gaseous (spiral and irregular) galaxies. Giant H II regions can be used as distance indicators for their host galaxies.

In SPIRAL GALAXIES, H II regions are located preferentially either along the spiral arms or in ring-like distributions at/or surrounding the dynamical centres of the host galaxies (these are called nuclear and circumnuclear starbursts). A beautiful example of a spiral arm distribution is found in the grand-design spiral galaxy M51. In the case of the nuclear and circumnuclear starbursts, the gas that is fuelling the star-forming activity was probably driven towards the dynamical center by a series of different perturbations in the disk, including interactions with other galaxies and stellar bars. Similar flows may also fuel the intense power generated in ACTIVE GALACTIC NUCLEI (AGN). This possible connection suggests that the starburst and AGN phenomena can coexist in the central regions of many galaxies. An intriguing example of an active galaxy with circumnuclear regions is the face-on SEYFERT 2 GALAXY, NGC7742.

In the case of IRREGULAR GALAXIES, which are smaller and more gas-rich, the population of H II regions is randomly distributed and does not follow a regular pattern. For instance, the brightest H II region complex in the Large Magellanic Cloud is called the 'TARANTULA' NEBULA and represents a violent starburst. It does not seem to be coincident with any particularly significant point of this galaxy. The most dramatic examples of small galaxies undergoing starburst phenomena are the so-called 'H II

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Types of H II regions.

Type of region	Size (pc)	Number density (cm^{-3})	Ionized mass ($M_\odot$)	Number of ionizing stars N_*
SUCHII	≈0.003	$\geq 10^6$	$\approx 10^{-3}$	1
UCHII	≤0.1	$\geq 10^4$	$\approx 10^{-2}$	1
Compact	≤0.5	$\geq 5 \times 10^3$	≈1	1
Classical	≈10	≈100	$\approx 10^5$	Few
Giant	≈100	≈30	10^3–10^6	$\approx 10^5$
Supergiant	>100	≈10	10^6–10^8	$\approx 10^5$

galaxies'. These are very small and dim galaxies that are difficult to detect other than in the emission lines of their H II regions.

Physical processes in H II regions

Stellar photons with energies greater than the ionization potential of hydrogen ($I_H = 13.6$ eV) ionize hydrogen and other elements. The excess energy goes into kinetic energy of the ejected electron, and this is shared with other particles by collisions, heating the gas. Photoionization takes place from the ground level of hydrogen (photoionization from higher levels is rarely important). This is balanced by proton and electron recombination into any level. The resulting cascade downwards gives the hydrogen recombination line spectrum. Lines are produced over a wide range of wavelengths from the UV to the radio region.

The degree to which the hydrogen is ionized depends on the balance between the rates of photoionization and recombination. If the characteristic time for a photon and an electron to recombine is much greater than the characteristic ionization time of an atom, the hydrogen will be essentially fully ionized. This indeed is the case in the interiors of H II regions and far less than 1% of the hydrogen is in the neutral form.

The gas temperature in H II regions is remarkably constant because of a thermostatting mechanism. The heating rate per unit volume is proportional to the number density of neutral hydrogen, which decreases with increasing gas temperature. The cooling of the plasma is largely due to the radiative decay of electron-impact-excited levels of ions such as O^{2+}. The lines produced are called 'forbidden' lines because of their low transition probabilities. The rate of emission of radiation in these lines is proportional to the populations of the excited levels from which they originate which are essentially proportional to a Boltzmann factor which increases as the temperature increases. Thus, an increased gas temperature results in decreased heating and increased cooling while the converse is true for a lowered temperature. This thermostat regulates the gas temperature to the value quoted.

The forbidden lines carry information about the physical structure of H II regions. The gas temperature and density determine the ratios of line intensities. The ratio of the intensity of the 436.3 nm line of O^{2+} to the sum of the intensities of the 500.7 nm and 495.9 nm lines is a good probe of temperature, whereas the ratio of the intensities of the 372.6 nm and 372.9 nm lines of O^+ probes the density. Many other lines can be used. Care is needed in the interpretation of such data. The ionization state of any particular element depends on the energy distribution of the photons to which the parent element is exposed. This energy distribution is strongly influenced by photoabsorption by helium. Hard photons are absorbed nearest the star and so ions with the highest parent ion ionization potential are found there (e.g. O^{2+}, N^{2+}). The radiation field is softer further out and lower parent ionization ions (e.g. O^+, N^+) predominate. The phenomenon is called 'ionization stratification'. Consequently, measurements of temperature from O^{2+} lines do not probe the same regions as the density measurements from O^+ lines. The lines can be used to obtain elemental abundances relative to hydrogen. In fact, the use of H II regions as probes of abundance variations in distant galaxies has proved to be one of the most reliable methods of obtaining this information.

Emission lines enable astronomers to probe the dynamics of H II regions. The DOPPLER EFFECT, to which there are two contributions, determines the wavelength (or frequency) width of a spectral line. Lines are broadened by thermal velocities as well as by internal bulk nebular velocities. These two velocities are about equal for hydrogen and a hydrogen line has two comparable contributions to its width. The thermal velocity of a heavy ion (e.g. O^+) is much smaller than that of the hydrogen so forbidden lines are broadened by bulk motions. Comparison of forbidden line and hydrogen line widths enables these two contributions to be distinguished. In addition to their effects on the line widths, bulk motions also affect the wavelength of the line maximum intensity. Thus line widths and peak wavelengths map velocity fields.

Important information on nebular structures can also be obtained at radio wavelengths from e.g. hydrogen radiofrequency recombination lines. Nebulae also have a strong radio continuum spectrum produced by bremsstrahlung. This gives information on the total recombination rate in the ionized gas and is one means of estimating the number of stars which excite the nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Because photons are continuously absorbed in balancing recombination, the stars do not ionize an indefinitely large region. When a star 'switches on', the gas is initially fully ionized out to the 'Strömgren radius' $R_S = (3S_*/4\pi n_0^2\beta)^{1/3}$ where n_0 is the hydrogen number density of the surrounding gas, S_* is the number of ionizing photons produced per second by the ionizing star(s) and $\beta(\approx 2.6 \times 10^{-13}$ cm^3 s$^{-1})$ is the hydrogen recombination coefficient. We find $R_S \approx 31/n_0^{2/3}$ pc. (This is an overestimate since dust in nebulae also absorbs photons.)

The transition zone between neutral and ionized gas is termed an 'ionization front' (IF) and has a thickness about equal to the mean-free-path of an ionizing photon in the neutral gas. This distance is approximately $0.05/n_0$ pc, far smaller than R_s. Consequently, nebulae bounded by ionized–neutral zones—'ionization bounded' H II regions—are sharp edged. If the nebula has insufficient hydrogen to absorb all the stellar UV there is no transition zone, all the gas is ionized and the region is called 'density bounded'.

H II region dynamics

The ionizing photons propagate outwards from the star to set up the Strömgren sphere. Conditions across the IF are related by equations analogous to the well-known Rankine–Hugoniot (R–H) jump conditions, which hold across shock fronts. However, the principle that the specific entropy increases across a shock usually allows only one solution to the R–H conditions, namely a compressive shock. The situation across IFs is more complicated since radiative heating and cooling of the gas means that this entropy condition is no longer relevant. There are in fact four allowed transitions whose nomenclature was first introduced in a classic paper by F D Kahn in 1954.

Across 'R-type' IFs, the density increases. They move supersonically into the neutral gas and divide into weak-R types and strong-R types according to whether the ionized gas moves supersonically or subsonically respectively away from the IF. Strong R-type fronts demand special circumstances for their existence and are of no astrophysical significance. Across 'D-type' fronts the density decreases. These fronts move subsonically into the neutral gas and again divide into weak D-type fronts and strong D-type fronts according to whether the ionized gas moves subsonically or supersonically with regard to the IF, the reverse of the R-type nomenclature. There is a 'forbidden' gap in IF velocity, roughly between $2c_I$ and $c_N^2/2c_I$, where $c_I(\approx 10$ km s$^{-1})$ and $c_N(\approx 1$ km s$^{-1})$ are respectively the sound speeds in ionized and the neutral gas. The internal structures and propagation of IFs were examined in classic papers by W I Axford and F A Goldsworthy in 1961. These studies (and that of Kahn) show that if an initially fast IF has a velocity which tries to decrease into the forbidden gap, a shock wave moves ahead of the IF into the neutral gas. The increased gas density behind the shock makes the IF velocity drop to a value below that of the lower limit of the gap.

If a star is suddenly switched on in a uniform region of neutral hydrogen (see H I REGION), a weak R-type front sweeps across the neutral gas, ionizing it but leaving its density (and velocity) essentially unchanged. The speed of the ionization front V_F is given by $V_F = J/n_0$, where J is the ionizing photon flux at the IF and n_0 is the neutral gas number density. As the IF moves outwards from the star, the flux is reduced by geometrical dilution (since the IF is a spherical surface centered on the star) and absorption by hydrogen atoms to maintain ionization balance. Clearly V_F decreases as the IF moves away from the star. The ionized gas has a pressure that is a factor $2T_i/T_0$ greater than that in the neutral gas (the factor of 2 arises because each hydrogen atom produces two particles when ionized). The ionized gas temperature typically is $T_i \approx 8000$ K and the neutral gas temperature is $T_0 \approx 100$ K. The ionized gas is overpressured with respect to the neutral gas by a factor of over 100 and will expand at a speed of approximately c_I. As long as $V_F \gg c_I$, this is of no significance. However, when the IF velocity drops to about $2c_I$, a shock is driven ahead of the IF and it switches from weak R-type to weak D-type. The H II region acts as a spherical piston pushing a shock front into the neutral gas. This occurs after a time approximately equal to $(n_0\beta)^{-1}$, which is roughly the recombination time in the ionized gas. The radius of the IF when this occurs is about equal to R_S.

The expansion of the H II region continues with the IF remaining weak D-type and the shock gradually weakening with time. In principle, the expansion continues until the H II region is in pressure equilibrium with its surroundings when its radius would be $(2T_i/T_0)^{2/3}R_s(\approx 30R_S)$ and the mass of ionized gas would be a factor $2T_i/T_0(\approx 160)$ greater than that in the initial Strömgren sphere. The density in the final equilibrium state is $n_f \approx (T_n/2T_i)n_0(\approx 0.006n_0)$, far lower than the original neutral density. A lower bound to the time taken to reach this final pressure equilibrium is $t_f \approx 8.8 \times 10^8/n_0^{2/3}$ yr.

It is instructive to ask whether a final pressure equilibrium state is likely. Let us first take average conditions in the general interstellar medium. The interstellar gas pressure at the solar circle (the distance of the Sun from the center of our Galaxy) is $P_{ISM} \cong 10^{-12}$ dyn cm^{-2}. The H II region reaches pressure equilibrium at a radius of about 30 pc and a density of about 0.4 cm^{-3}. However, the time to reach this equilibrium state is very long ($\cong 10^7$ yr) and most massive stars would have evolved significantly off the main sequence before equilibrium could be achieved. On the other hand, the H II region set up by a star buried deep in a high-density molecular core where the pressure is much greater ($\sim 10^{-7}$ dyn cm^{-2}) could reach equilibrium in a time $\sim 10^4$ yr with a radius of about 0.1 pc and a density $\sim 4 \times 10^4$ cm^{-3}. Such values are similar to these determined observationally for UCHIIRs. We will discuss these similarities later.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

During the expansion process, both the ionized gas and the shocked neutral gas acquire kinetic energy. Once the shock has been established, that in the neutral gas dominates. The velocity in the ionized gas rapidly falls below the sound speed and its energy is largely thermal. This comes from the heating of the gas by the stellar radiation field. The kinetic energy of the neutral gas arises from the expansion work done on it by the hot gas. Comparison of the kinetic energy in the neutral gas at any time with the total radiant energy absorbed up to that time shows that the former is much less than 1% of the latter. Nearly all the radiant energy absorbed in the ionized gas is radiated away in the forbidden lines.

Observations of regions of star formation show that the gas around young stars is far from homogenous. If an IF encounters a strong negative density gradient and overruns it, the gas expansion is accelerated. For stars located near the boundary of the dense molecular core in which they formed, the ionized gas accelerates out of this core and creates an extended flow. The evolution in a decreasing-density environment displays a rich variety of hydrodynamic phenomena including shocks in the ionized gas, receding ionization fronts, rapid evaporation of the dense core and the creation of new clumps in the expanding shocked gas.

Two factors influence the dynamical evolution of flows set up in a dense core. The first is the position of the star in the core. If the pressure equilibrium Strömgren sphere is contained within the densest part of the core for a star located there, a long-lived pressure equilibrium region can be formed. By contrast, if the star is located towards the core boundary, the IF can move rapidly outside the core region into the envelope and a transient high-pressure region is set up. The second factor is the density variation in the envelope that surrounds the core.

As an illustration, detailed investigations of the evolution of an H II region produced in an envelope with a power law density distribution with distance r, $n \propto r^{-\omega}$, show the following. If $\omega < 3/2$, the interface between the IF and the leading shock accumulates neutral gas (exactly as in the uniform-density case described above). If $\omega = 3/2$, the IF and the shock move together without a neutral interface. If $\omega > 3/2$, the IF overtakes the shock and moves out of the cloud: in this case, steep enough density distributions can lead to velocities as high as 100 km s^{-1}. Such velocities are not produced in uniform-density media when the expansion velocity never exceeds the sound speed in the ionized gas.

Another important dynamical process needs to be included. Massive stars lose material from their surfaces in a continuous stream, a 'stellar wind'. This wind is driven by radiation pressure and the typical speed ($\approx$2000 km s^{-1}) is about three times the escape velocity of the star. The mass-loss rate for a supergiant star is about $10^{-6} M_\odot$ yr^{-1}, and that for a main sequence star about $10^{-7} M_\odot$ yr^{-1}. The wind kinetic energy output rate (known as the 'mechanical luminosity') ranges from about 10^{36} to 10^{34} erg s^{-1}, far less the rate at which the star pumps radiant energy into the ionized gas. When the wind impacts on the nebular gas, it acts as a piston and pushes a shock wave into the gas, sweeping it up into a shell. The wind itself is slowed down by another shock wave. The wind gas is shocked to a high temperature ($\approx 10^7$–10^8 K) and the pressure of this hot gas drives the shell. Remarkably, the fraction of the mechanical energy of the wind transformed into kinetic energy of swept-up gas is constant (about one-fifth), far higher than the radiation conversion efficiency, and consequently, stellar winds play an important role in the dynamics of H II regions.

The evolution and dynamics of H II regions where wind activity is important are different from those where the winds are weak. The most immediately obvious difference is that a hole is blown around the star. A spectacular example of this is in the ROSETTE NEBULA where the hole is blown by the combined winds of several stars. Additionally, the velocities of the ionized shells can be several times the sound speed, much higher than in uniformly expanding H II regions. The compression of the nebular gas by the shock increases the recombination rate considerably and the IF, which initially outruns the shell, may become trapped inside the shell.

A simple evolutionary scheme

A simple expansion scheme may connect the different classes of H II region. SUCHII regions and UCHIIRs are buried inside the inner, high-pressure, parts of the parent molecular clouds and are invisible to optical telescopes. Compact H II regions are bright and visible and are located next to the opaque regions. They therefore seem to be more evolved objects. Extended optically visible H II regions represent the more mature expanded state of these objects.

Even this simple scheme meets with difficulties. The fraction of objects in each class should be roughly the time that it takes the region to expand from one class into another, divided by the typical lifetime of the star. The number of UCHIIRs known is larger than expected if they really do expand to become compact regions. This has been termed the 'lifetime problem'. There is no clear answer to this problem at present although recent surveys have found faint extended emission in several samples of UCHIIRs. This implies that some of these objects are larger, partially mitigating the problem. The simplest way around the lifetime problem is to assume pressure confinement.

Structures in H II regions

H II regions show complex structure on large and small scales. The 'Pillars of Creation' in M16, the EAGLE NEBULA, are a spectacular example. Many H II regions display such dark clouds of neutral material silhouetted against the nebula. An important question is whether the structure is primordial (i.e. reflects conditions in the neutral gas prior to ionization) or is produced by dynamical processes. Opinion at present favors the former because observational data on the structure of star-forming regions show that the neutral gas is fragmented on all

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scales. However, it is far from clear that the latter do not contribute. There are a variety of ways in which this could happen. For example, a wind-driven shell can fragment by a combination of Rayleigh–Taylor and thin-shell dynamical instabilities if the shell accelerates because of, for example, a steep drop in density ahead of the shell. On the other hand, primordial clumps may actually be destroyed by photoionization or by the passage of the shock ahead of the IF.

If the neutral material in the star-forming cloud is clumpy, considerable revision has to be made to our description of H II region dynamics. If a neutral clump survives inside the H II region, it acts a mass injection source. Mass injection occurs either by photoionization of the clump or by hydrodynamic ablation as the H II region gas moves around it. Which process is dominant depends on factors including the location of the clump with respect to the star and the strength of any stellar wind. The injected gas may move subsonically or supersonically with respect to the flow in the H II region (in the latter case, localized shocks are set up in the H II region around the clumps, generating localized supersonic velocities). The injection of mass can lead to a situation where matter has to flow out of the H II region and become neutral. In this case, the H II region is bounded by a recombination front through which ionized gas flows, recombines and becomes neutral (the opposite of an IF). Such a region can be in pressure equilibrium and models of this nature have been proposed for UCHIIRs.

Stellar nurseries revealed

The ORION NEBULA is just one of several star-forming regions in the Orion molecular cloud complex. In this complex, both low- and high-mass star formation is taking place. It has been known for about three decades that in the Orion nebula there exist compact knots that have similar spectra to H II regions. These knots are extremely small, characteristically about 500 AU in size, and are therefore not stars. An early suggestion was that the radiation came from the ionized surfaces of neutral clumps embedded in the H II region. A second suggestion was that the clumps contained stars and that the bright knots were externally ionized envelopes around these very young stars. It was, however, the first Hubble Space Telescope images of the objects obtained in 1993 which revealed their true natures. They are actually externally ionized disks of material around young low-mass stars and are known as PROPLYDs (PROTOPLANETARY DISKS). They are seen silhouetted against the nebular backgrounds and, in several, the stars are clearly visible. Thus massive star and consequent H II region formation has provided a new window for the study of low-mass stars. They also have provided new laboratories for the study of the interaction of stellar UV radiation and fast stellar winds with neutral material. Many of these objects have a 'teardrop' shape with tails pointing away from θ^1 Ori C, the brightest star in the TRAPEZIUM cluster which is responsible for the excitation of the Orion nebula. Although there are some difficulties yet to be resolved, it is likely that the morphology can be accounted for by models where an ionized flow of gas off the disk interacts with the fast stellar wind from θ^1 Ori C.

Some important problems

Most theoretical studies of H II region evolution assume that the stellar winds and radiation fields turn on suddenly with full main-sequence mechanical and radiative luminosities. The actual way this happens is unclear and has consequences for the subsequent ionization and dynamics. Moreover, massive stars commence nuclear burning when they are still in the process of accreting matter and are therefore surrounded by dense dusty cocoons when this turn-on occurs. It is therefore likely that there is an initial phase of combined inflow and outflow. There has been little work so far on the propagation of IFs into clumpy dusty media, in particular any studies of the possible transition from an H II region headed by an IF to one headed by a recombination front.

Gas in regions where active star formation is continuously taking place can be compressed by wind- and supernova-driven shock waves. Any magnetic field in the unshocked gas can be greatly magnified provided that the shocked gas cools well. Under such circumstances, IFs may move into gas where the magnetic pressure dominates the gas pressure. Magnetohydrodynamic IFs behave in a much more complex fashion than do the 'classical' hydrodynamic IFs described above; in particular there are many more jump possibilities.

The morphologies of H II regions have been the subject of many theoretical studies. For example, UCHIIRs have a whole range of morphologies ranging from simple spherical, through core–halo, irregular and cometary. It is likely that many of the differences in appearance are the result of viewing aspect. In particular, cometary UCHIIRs have been widely studied. The cause of this particular morphology has been variously ascribed to the generation of umbrella-shaped stellar-wind-supported envelopes around moving stars, flows produced at the edges of dense clouds ('champagne' flows) and flows produced when a stellar wind from a stationary star blows into a spatially varying distribution of mass-injecting clumps. Ultimately, the confrontation of theory with very detailed observational data will resolve these uncertainties.

Bibliography

There are several good books describing the physics of H II regions.

Dyson J E and Williams D A 1997 *The Physics of the Interstellar Medium* 2nd edn (Bristol: Institute of Physics Publishing)

Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Mill Valley: University Science Books)

Spitzer L 1978 *Physical Processes in the Interstellar Medium* (New York: Wiley)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Review papers:

Churchwell E 1990 Ultracompact H II regions: the impact of newly formed stars on their environment *Astron. Astrophy. Rev.* **2** 79

Franco J, Kurtz S, Garcia-Segura G and Hofner P 2000 The evolution of H II regions *Astrophys. Space Sci.* at press

Garay G and Lizano S 1999 Massive stars: their environment and formation *Publ. Astron. Soc. Pacific* **111** 1049

Habing H J and Israel F P 1979 H II regions and OB star formation *Ann. Rev. Astron. Astrophys.* **17** 345

Yorke H W 1986 The dynamical evolution of H II regions *Ann. Rev. Astron. Astrophys.* **24** 36

Specialist papers:

Axford W I 1961 Ionization fronts in interstellar gas: the structure of ionization fronts *Phil. Trans. R. Soc.* A **253** 301

Franco J, Tenorio-Tagle G and Bodenheimer P 1990 On the formation and expansion of H II regions *Astrophys. J.* **349** 126

Goldsworthy F A 1961 Ionization fronts in interstellar gas and the expansion of H II regions *Phil. Trans. R. Soc.* A **253** 277

Kahn F D 1954 The acceleration of interstellar clouds *Bull. Astron. Inst. Neth.* **12** 187

Redman M P, Williams R J R, Dyson J E, Hartquist T W and Fernandez B R 1998 Magnetic ionization fronts *Astron. Astrophys.* **331** 1099

J E Dyson and J Franco

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Högbom, Jan (twentieth century)

Swedish astronomer, pioneer in radioastronomy and interferometry, invented the *Högbom CLEAN algorithm* for deconvolving an interferometric image from the effects of incomplete sampling. In general the raw map produced by an interferometer is too dirty for scientific analysis. CLEAN uses the sampling function of the radio telescope interferometer (the 'dirty beam') to match the brightest peaks on the observed brightness distribution ('dirty map') of a radio source. It subtracts (deconvolves) the effect of the bright radio sources seen by the dirty beam from the map and looks in the residual map for less bright peaks. The position of each source is saved in a 'CLEAN component' table in the CLEAN map file, and the algorithm stops when the residual brightness distribution consists mainly of noise. A CLEAN radio map is then reconstructed from the CLEAN components. There are many variants on Högbom's basic CLEAN algorithm (associated with the names of Clark, Cornwell, Cotton–Schwab, Gerchberg–Saxton, Steer, Steer–Dewdney–Ito, and van Cittert).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hadley, John (1682–1744)

Mathematician and instrument designer, born in Enfield Chase, Hertfordshire (now in Greater London), credited with ROBERT HOOKE in building the first Gregorian reflector telescope, making it into an accurate instrument for use in astronomy. Invented a double-reflecting quadrant with bubble-level to measure the altitude of the Sun or of a star. It was used to determine position at sea and evolved into the sextant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hadrons

Subatomic particles that are composed of quarks and which are acted on by the strong nuclear force. Hadrons may be subdivided into mesons and baryons, a meson consisting of a quark–antiquark pair and a baryon of three quarks. Most hadrons, with the exception of protons and neutrons (which are baryons), have very short lifetimes.

See also: baryons, fundamental forces, mesons, neutron, proton, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hagen, Johann Georg (1847–1930)

Became a Jesuit. Director of the Georgetown College Observatory, Washington DC, and of the Roman College Observatory. Hagen is noted for his observations of light and dark nebulae and his unproven theories about them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

HALCA (Highly Advanced Laboratory for Communications and Astronomy/Muses-B)

Japanese–US radio astronomy observatory, the first very long baseline interferometry satellite. Launched in February 1997 into a highly elliptical ($1000 \times 10\,000$ km) Earth orbit. Equipped with an 8 m antenna operating at frequencies near 1.6, 4.8 and 22 GHz. Returned high resolution images of quasars and distant active galaxies. An alternative name is Haruka, the Japanese pronunciation of the HALCA acronym. Haruka is Japanese for 'far away'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hale, George Ellery (1868–1938)

American astrophysicist, born in Chicago, IL, as an undergraduate at MIT invented the spectroheliograph, which made it possible to photograph the Sun's prominences in daylight. At the University of Chicago he used a 40 in lens as the basis for the telescope of the Yerkes Observatory, completed 1897. Founded Mount Wilson Observatory, where he discovered the magnetic fields in sunspots. Planned and completed the 60, 100 and 200 in reflectors, the last, on Mount Palomar, named for him after his death. He was the first astronomer to be officially called an astrophysicist and started the *Astrophysical Journal*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hall, Asaph (1829–1907)

Carpenter, computer and astronomer; born in Goshen, CT. While at the US Naval Observatory in Washington, DC, he discovered the two satellites of Mars, naming them Deimos and Phobos.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Halley, Edmond (1656–1742)

Astronomer and scientist, born in Hagerston, Shoreditch (near London), became professor at Oxford and second Astronomer Royal at the Royal Observatory, Greenwich. From the island of St Helena, he cataloged the positions of about 350 southern hemisphere stars (*Catalogus Stellarum Australium*), including his own discovery of the globular cluster Omega Centauri; he also discovered the globular cluster M13. He dedicated a planisphere of the southern hemisphere stars to King Charles II. The planisphere flatteringly included a now disused constellation (in Carina) which Halley named 'Robur Carolinum', the oak tree in which Charles hid after his defeat by Cromwell after the battle of Worcester. He observed a transit of Mercury and invented the idea to use transits of Mercury and Venus to determine the distance of the Sun. Using PTOLEMY's catalog, he deduced that the stars had moved relative to each other and detected this 'proper motion' in three. Apart from his own scientific contributions Halley was a friend of NEWTON. He supported NEWTON in his controversy with Leibniz over who invented the calculus, persuading him to distribute his work in the *Principia*, whose publication Halley paid for, and helping him access FLAMSTEED's observations of the Moon before they were completed to Flamsteed's satisfaction. This last was thought by the two friends to be necessary for Newton to be able to test his theory of gravitation, but it was Halley who provided the more successful and dramatic proof. Using Newton's theory of cometary orbits, Halley calculated that the comet of 1682 was periodic, and predicted (*A Synopsis of the Astronomy of Comets*) that it would return in 76 years. It did (after Halley's death), and it is now called Halley's Comet. Halley is considered the founder of geophysics, especially for his work on trade winds, the tides and the magnetism of the Earth, and with the *Breslau Table of Mortality*, he laid the actuarial foundations for life insurance and annuities.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Halo, Galactic

A spherical or spheroidal distribution of globular star clusters and old stars that surrounds a spiral galaxy. The diameter of a galactic halo is comparable to, or larger than, the overall diameter of the disk within which the spiral arms are located (about 100 000 light-years in the case of the Milky Way Galaxy, the galaxy of which the Sun is a member).

The globular clusters are composed of 'Population II' stars—old, metal-poor stars that formed at an early stage in the evolution of galaxies from clouds of hydrogen and helium that contained very little in the way of heavier elements. The halo of our own galaxy also contains individual Population II stars that circulate around the galactic center in orbits that are tilted at random angles to the galactic plane. Because these stars do not share in the orderly motion of stars in the galactic disk (all of which are proceeding in the same direction around the galactic center, in the same plane), their velocities relative to the Sun are much higher than those of stars in the disk population. Consequently, they are known as 'high velocity' stars.

Studies of motions of distant globular clusters and of the rates at which stars and gas clouds in the galactic disk revolve around the galactic center show that the total mass of the Milky Way Galaxy is about 10^{12} solar masses, about ten times greater than the combined mass of all the stars, gas and dust that can be detected directly, and that most of this mass is contained in the halo. Similar results have also been obtained for other spiral galaxies. These observations imply that up to 90% of the mass of a typical spiral galaxy is made up of dark matter. While part of this total may be provided by objects with extremely low luminosities, such as brown dwarfs (stars with masses too low to enable nuclear energy generation to occur in their cores) or black holes, much of it may exist in the form of exotic elementary particles.

See also: dark matter, globular clusters, Milky Way Galaxy, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hamilton, James Archibald (1748–1815)

Irish astronomer, born in Athlone, on the recommendation of NEVIL MASKELYNE became first director of Armagh Observatory, where he supervised the erection of the equatorial telescope by EDWARD TROUGHTON. The benefits of his efforts were reaped in science not by himself but by the third director of Armagh Observatory ROMNEY ROBINSON.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Harding, Karl Ludwig (1765–1834)

German astronomer, became professor of astronomy at Göttingen,where he advised HEINRICH SCHWABE on solar observations. Discovered Juno, the third asteroid, from J H SCHRÖTER's observatory at Liliethal, where he was an assistant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Haro, Guillermo (1913–88)

Mexican astronomer, director of the Mexican Institute of Astronomy, and the Tonantzintla Observatory. Independently of George Herbig, he identified small emission-line nebulae in the Orion nebula, associated with but differing from T Tauri stars. These objects, now known as *Herbig–Haro objects*, are thought to be shock-excited impact regions with their energy derived from outflowing jets from young stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Harriot, Thomas (*c*. 1560–1621)

Teacher, probably born in Oxfordshire, Harriot travelled with Sir Walter Raleigh to Virginia as cartographer and navigator, writing on his return *A Briefe and True Report of the New Found Land of Virginia*. He was Ralegh's tutor and mathematics teacher and introduced simplified notation for algebra, inventing the symbols $<$ for 'less than' and $>$ for 'greater than', as well as . for multiplication. When he died he left a large number of manuscripts, and his scientific genius emerged. He had corresponded with JOHANNES KEPLER on optics and had discovered what is now known as Snell's law of refraction before Snell. He observed Halley's Comet in 1607 (Kepler had discovered the comet six days earlier) and a second comet in 1618. From 1609 to 1613 he made numerous telescopic astronomical observations. He observed the moons of Jupiter, unaware of GALILEO's discovery. His drawing of the Moon, the first recorded, preceding Galileo by several months, he was also first to record sunspots. These were not isolated events—he made 199 observations of the Sun over three years and deduced the Sun's rotation period. In contrast to Galileo, who published his book on his telescopic discoveries within weeks of making them, Harriot published none of his discoveries and therefore had no influence on the development of optics or astronomy (publish, or the ideas perish), although his influence as a teacher on the development of mathematics in England was profound.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Harrison, John (1693–1776)

Clockmaker, born in Foulby, West Yorkshire, England. In 1713 the British Government offered a valuable prize for the invention of a method to determine longitude accurately. Harrison developed a series of clocks through wooden models, and in 1726 invented the bimetallic pendulum that compensated for thermal expansion due to the variations of climate expected on long sea voyages, as well as several other mechanical devices to minimize friction. He made three spring-pendulum chronometers (H1, H2, H3) which worked regularly even during the rolling of a ship. They were finally surpassed in accuracy by a watch-like marine chronometer (H4), which, in a voyage to Jamaica (1761–2) determined longitude with an accuracy of two geographical miles. He was shamefully treated by the Board of Longitude, which made up conditions to delay awarding him the prize, and was finally awarded it in his old age by the intervention of King George III. Harrison's restored chronometers are displayed in the Royal Observatory, Greenwich.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hartebeesthoek Radio Astronomy Observatory

The Hartebeesthoek Radio Astronomy Observatory (HartRAO) is located in the Gauteng Province of South Africa, 60 km NW of Johannesburg. It is one of four national research facilities operated by the National Research Foundation. The Observatory was established in 1975 at the former NASA Deep Space Station 51, which was founded in 1961 to support unmanned US space probes to the Moon and planets, a function it fulfilled from 1961–74.

HartRAO operates a 26 m radio telescope which is equipped with cryogenic receivers for 1.6, 2.3, 5.0, 6–6.7 and 8.5 GHz. It has instrumentation for continuum, spectroscopic and pulsar observations. It also participates in a variety of networks using the technique of very long baseline interferometry (VLBI) for high-resolution imaging of compact radio sources and for geodesy. As the only operational radio telescope in Africa, it plays a cardinal role providing both NS extensions to northern networks and EW extensions to southern networks. It is a key element in the ground network of VLBI stations that support the Japanese space VLBI satellite, Halca.

HartRAO operates a number of global positioning system (GPS) receivers in Southern Africa as part of the International GPS Service (IGS) and serves as a regional IGS data center. From early 2000 HartRAO will also operate a satellite laser ranging system, making the Observatory a fiducial geodesy station for the African continent. The ITRF94 position of the 26 m radio telescope provides the datum for the South African WGS94 National Survey Network.

For further information see
http://www.hartrao.ac.za.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hartmann, Johannes Franz (1865–1936)

Born in Erfurt, Germany, became professor in Potsdam where he developed astrophysics. During this time an 80 cm refracting telescope was installed in the observatory for photographic use. Its performance was disappointing, and Hartmann developed the method of testing optics today named after him. The *Hartmann test* consists of partly obscuring the lens with a *Hartmann shutter*, and revealing differences of optical focus in the various zones of the lens. Hartmann refigured the telescope's lens and was able to use it for spectroscopy. After a spell at Göttingen as director of the observatory, he went to La Plata, Argentina, where he developed a theory on novae and discovered from its light curve that the minor planet Eros is a rotating non-spherical body.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Harvard–Smithsonian Center for Astrophysics

The Harvard–Smithsonian Center for Astrophysics (CfA) is a joint facility of the Smithsonian Institution and Harvard University that combines the staff and resources of the Smithsonian Astrophysical Observatory and the Harvard College Observatory under a single director to pursue research in nearly every branch of modern astrophysics—from laboratory measurements of the structure of atoms and molecules to observations of the large-scale structure of the universe. Investigations are organized broadly in seven divisions: atomic and molecular physics, high-energy astrophysics, optical and infrared astronomy, planetary sciences, radio and geoastronomy, solar and stellar physics and theoretical astrophysics.

Observational facilities include the multipurpose Fred Lawrence Whipple Observatory (FLWO) on Mount Hopkins in Arizona and the Oak Ridge Observatory in Massachusetts, as well as a 1.2 m radio telescope at the CfA headquarters in Cambridge, MA. The major instrument on Mount Hopkins, the Multiple Mirror Telescope (MMT), is operated jointly with the University of Arizona. Also located at FLWO is a 10 m diameter reflector to detect gamma rays; a 1.2 m imaging optical/infrared telescope; a 1.5 m spectroscopic telescope; an optical and infrared interferometer operated in collaboration with the University of Massachusetts, the Paris Observatory at Meudon, NASA Ames and MIT's Lincoln Laboratory; and, a 1.3 m infrared telescope operated by a consortium for a 2 micron all-sky survey.

Major support facilities in Cambridge offer a host of computers connected by a local-area network. Special laboratories are maintained for petrologic and mineralogic studies of meteorites and lunar samples, for spectroscopy of atoms and molecules, and for development of instrumentation (e.g. advanced optical and x-ray detectors). Major projects include development of a submillimeter telescope array in Hawaii, conversion of the MMT to a single-mirror telescope 6.5 m in diameter, and participation in the Magellan Project to build two 6.5 m diameter telescopes at Las Campanas, Chile.

CfA instrumentation operating in space includes a high-resolution camera on the Chandra X-ray Observatory, ultraviolet coronagraph spectrometers on the Spartan and Solar and Heliospheric Observatory (SOHO) spacecraft, special mirrors on the Transition Region and Coronal Explorer (TRACE), and detectors on the Submillimeter Wave Astronomy Satellite (SWAS). Development of new instrumentation for the Space Infrared Telescope Facility (SIRTF) continues.

Numerous facilities serving the general astronomical community are located at the CfA: the Institute for Theoretical Atomic and Molecular Physics; the International Astronomical Union's Central Bureau for Astronomical Telegrams and the Minor Planet Center; the US gateway for SIMBAD, an international astronomical computer database; and Harvard's extensive collection of astronomical photographic plates. In addition, the Astrophysics Data System (ADS), operated on behalf of the National Aeronautics and Space Administration (NASA), Control Center is also based here.

For further information see http://cfa-www.harvard.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hat Creek Radio Observatory

Located 400 km north of Berkeley, California. It is the site of the BIMA Millimeter Array, a 10-antenna aperture synthesis telescope which operates at wavelengths of 3 mm (70–116 GHz) and 1 mm (210–270 GHz).

The array is operated by the BIMA consortium, consisting of the Radio Astronomy Laboratory of the University of California at Berkeley, the Laboratory for Astronomical Imaging of the University of Illinois, and the Laboratory for Millimeter Astronomy of the University of Maryland, with support from the National Science Foundation.

The BIMA telescopes are 6.1 m in diameter. The antennas may be located at various stations along an approximately T-shaped track to allow antenna separations ranging from 7 m to 2 km. Normally, the antennas are deployed in one of four standard configurations which provide angular resolutions of roughly 0.4, 2, 6 or 14″ at 100 GHz.

For further information see
http://astron.berkeley.edu/ plambeck/technical.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hawking, Stephen W (1942–)

Cosmologist and theoretical astrophysicist, born in Oxford, England, where he studied physics at University College. Moved to Cambridge to take up research in general relativity and cosmology, became Lucasian professor (an appointment earlier held by ISAAC NEWTON, with whom Hawking has been compared). Hawking worked to develop a valid mathematical treatment of the 'singularities' in the theory of general relativity, that is the central regions in black holes, where conventional mathematics breaks down. Uniting quantum theory and general relativity, he showed that black holes can emit energy, now known as *Hawking radiation*. He investigated the conditions in the energy released in the Big Bang at the creation of the universe, and predicted that many mini-black holes would be created, perhaps as heavy as 10 tons but only the size of a proton. Hawking found that the smaller mass black holes would lose energy by Hawking radiation, and would have evaporated in the lifetime of the universe. They therefore have had an effect on the formation of structure early in the universe's history but have now disappeared. Another remarkable proposal by Hawking is his 'no-boundary proposal' explained by him as meaning '... that both time and space are finite in extent, but they don't have any boundary or edge .... There would be no singularities, and the laws of science would hold everywhere, including at the beginning of the universe.' Hawking suffers from motor neurone disease and, confined to a wheelchair, he speaks through a voice-synthesizer. Nevertheless, he has taken up an arduous and highly successful career as a popularizer of science through lectures, TV and other public appearances, and a best-selling popular book on cosmology *A Brief History of Time*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hawkins, Gerald

British-born astronomer, professor at Boston University, claimed to have 'decoded' Stonehenge. He found a pattern of alignments with 12 major lunar and solar events, finding a way to predict eclipses of the Moon, the summer and winter solstice, by use of the 56 Aubrey holes, the Heel Stone, and the four Station Stones. His work was supported by astronomer FRED HOYLE and savagely criticised by archaeologists, (e.g. 'Moonshine on Stonehenge' by R J C Atkinson). Undeterred, he turned to a similar problem in Peru, the possible astronomical alignments and associations found in the mysterious lines and gigantic geometric figures at Nazca, mapped by MARIA REICHE.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hay, William Thompson [Will Hay] (1888–1949)

British comedian and amateur astronomer, discovered the white spot on Saturn in 1933.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Haystack Observatory

The Haystack Observatory is a multidisciplinary research center of the Massachusetts Institute of Technology, and its primary mission in astronomy is the advancement of scientific research and technical developments in radio interferometry. Current emphasis is on Very Long Baseline Interferometry (VLBI) observations at 3 mm wavelength aimed at high-resolution imaging of quasars, and on the application of VLBI geodetic techniques to measurements of the Earth's orientation parameters. The facilities of the Observatory include 37 m and 18 m diameter radio telescopes, and a multi-station VLBI correlator. The Observatory's mission includes training students in instrumentation technology and providing opportunities for students to link their education to research through radio astronomy using the Observatory's telescopes. The radio astronomy program at Haystack Observatory is guided by the Northeast Radio Observatory Corporation, a consortium of twelve educational institutions.

For further information see
http://www.haystack.mit.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Heavy Elements

Chemical elements of relatively high atomic weight. Because less than 2% of the mass of a solar-type star is made up of elements heavier than helium, the term 'heavy element', in an astrophysical situation, is applied to all elements that have mass numbers greater than that of helium (4).

See also: atom, cosmic abundance of elements, element (chemical), nucleosynthesis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Heckmann, Otto Hermann Leopold (1901–83)

Astrometrist and cosmologist, became director of the Hamburg Observatory, and of the European Southern Observatory in Chile. By determining the motions of the stars in the direction of several star clusters (including the Praesepe, or Beehive cluster), he was able to see which stars were moving together in the cluster, and which, although in the same direction, were moving differently. Eliminating the latter from consideration, he was able to identify the true members of each cluster and clarify their H–R diagrams. He organized the international effort to produce the third *Astronomische Gesellschaft* to show the proper motions of 180 000 stars. As a cosmologist, he rigorously used general relativity to find open and flat solutions to the EINSTEIN field equations for a homogeneous, isotropic, possibly rotating universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

HEGRA Experiment

The La Palma cosmic-ray observatory HEGRA (High-Energy Gamma-Ray Astronomy) is an air shower experiment, located at the OBSERVATORIO DEL ROQUE DE LOS MUCHACHOS (2200 m above sea level, 28.75°N, 17.89°W) on the Canary island of La Palma, and is operated by institutes from Germany, Spain and Yerevan.

The main scientific goal is gamma-ray astronomy in the energy range from 500 GeV to 100 TeV using the technique of ground-based imaging atmospheric Cerenkov telescopes. A unique feature of the HEGRA experiment is the use of five Cerenkov telescopes to simultaneously measure the same air shower (stereoscopic observation). This technique yielded an energy spectrum with unprecedented precision of up to 24 TeV for the Blazar Markarian 501 during flaring states in 1997.

From 20 TeV up to 10 PeV air showers are also detected with the showerfront sampling technique using arrays of 243 scintillator counters, 99 wide-angle Cerenkov counters and 17 Geiger towers spread over an area of 200 m $\times$ 200 m.

For further information see
http://wpos6.physik.uni-wuppertal.de:8080/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hektor

A Trojan asteroid discovered by August Kopff in 1907, designated (624) Hektor. It belongs to the largest group of Trojans, orbiting ahead of Jupiter around the L_4 Lagrangian point. Hektor orbits the Sun at a mean distance of 5.17 AU (774 million km) in a period of 11.76 years; the inclination is 18° and the eccentricity 0.02. Its rotation period, as determined from its light curve, is 6.92 h. Hektor is the largest and brightest of the Trojans. Variations in its light curve suggest that it is elongated, measuring about 300 × 150 km; alternatively it could be shaped like a dumbbell, or a contact or close binary asteroid. Like the majority of Trojans, Hektor is D-type, with a reddish reflectance spectrum indicating a carbon-rich surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Heliocentric Coordinate System

Coordinates specifying the position of an object as seen from the center of the Sun. Heliocentric coordinates are usually referred to the plane of the ecliptic—thus the position of a planet in its orbit at any instant is often specified in terms of heliocentric ecliptic coordinates.

See also: ecliptic, geocentric coordinate system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helios

Two German–US satellites designed to study the solar wind, interplanetary dust and Galactic cosmic rays in the inner Solar System. Helios 1, launched in December 1974, passed within 48 million km of the Sun. Helios 2, launched in January 1976, approached to within 45 million km of the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helioseismic Observations

This article describes how wave motions of the Sun are observed and utilized to learn about the properties of the solar interior.

Most of what we know about the Sun, and the rest of the universe, is based on observations of electromagnetic (e.g. light) waves. We cannot see more than about 100 km below the visible surface of the Sun because the material rapidly becomes opaque to electromagnetic waves. To 'see' deeper we observe, at the surface, wave motions of the solar material itself that easily travel large distances through the solar interior. HELIOSEISMOLOGY is the science of using these motions to study the Sun's internal structure and dynamics.

Figure 1. Line-of-sight component of the velocity of the solar photosphere shown as light (dark) shades for away (toward) motions. Solar rotation causes the large variation from left to right, while the oscillations are the fine-grained texture seen most clearly at disk center. (Courtesy of SOHO/SOI consortium. SOHO is a project of international cooperation between ESA and NASA.)

Wave motions

When most physical bodies are disturbed from an equilibrium state they tend to vibrate as they return to equilibrium. The Sun is no exception. In the case of the Sun, the disturbances that make it vibrate are believed to be violent convective motions (SOLAR INTERIOR: CONVECTION ZONE) beneath and within a few hundred km of the visible surface, and to a much lesser extent, rare, powerful explosions (solar flares) above the surface (SOLAR FLARE OBSERVATIONS). The convective disturbances occur continually, so the surface of the Sun is always oscillating. The goal of helioseismic observations is to detect and measure as many of the oscillations as accurately as possible.

Once produced, several types of wave motion can propagate inside the Sun (HELIOSEISMOLOGY: THEORY). Just as with light waves, these waves can be refracted, reflected, attenuated, scattered, dispersed, and absorbed, depending on the characteristics of the medium through which they travel. It is these phenomena that enable detailed probing of the interior. The waves are more than 10^{12} times larger than visible light, so using them to study the Sun is akin to using light waves to study a 1 mm diameter marble. One simple type of wave motion occurs when material is displaced from its equilibrium position and then moves back toward its original position under the combined forces of gravity and pressure from the immediate surroundings, producing what is usually called buoyancy. It oscillates about this position because of inertia until the wave energy is dissipated. Such buoyancy-driven motions are called gravity waves. Another wave motion is driven by pressure gradients and consists of alternating regions of changing pressure. These are the well-known acoustic or sound waves. Other types of wave motion are possible within the Sun, but they have not yet become significant in helioseismology.

Waves moving within the Sun may be confined to natural cavities by reflection and refraction. If the wave motions last long enough, standing wave patterns may form. The resonant frequencies of the standing wave oscillations are determined by propagation conditions (e.g. sound speed) within the entire cavity. Cavities that trap acoustic waves produce resonant pressure (p) modes. Cavities that trap gravity waves produce gravity (g) modes. The strong density gradient at the surface of the Sun supports surface gravity waves called fundamental (f) modes. High-frequency waves are not trapped at the solar surface. These waves may travel from their point of origin once through the solar interior and then leak outward in the solar atmosphere. Both trapped and untrapped waves are important for helioseismology. The former are used to learn about average properties within the trapping cavities, and the latter are used for diagnosing local, and possibly short-lived, conditions in the solar interior.

Basic observations

There is scant evidence or expectation that the type of waves that propagate within the Sun reach the Earth directly. Therefore we detect them indirectly, observing their effect at the surface of the Sun using traditional astronomical techniques.

Solar oscillations were first detected by R Leighton in 1960 by measuring the line-of-sight component of the motion of the solar photosphere using the Doppler effect. This remains the most powerful observational method. Figure 1 is a picture of the Sun's line-of-sight motions at a single instant of time. Images like this are the major (but not the only) source of helioseismic data.

If we isolate a few arcsec of the solar surface and measure its Doppler velocity or brightness as a function

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of time, we find characteristic oscillations with periods of 5 ± 2 min superimposed on other, non-periodic variations. The Doppler velocity oscillation is strongest at the center of the solar disk with an amplitude of about 300 m s^{-1}, and decreases roughly proportional to the cosine of the angle away from disk center. This shows that the motions are nearly radial to the solar surface. If we make measurements higher in the solar atmosphere, we find a greater amplitude and a tendency for the average period to shift to shorter values. The oscillations have been detected above the chromosphere (SOLAR CHROMOSPHERIC OSCILLATIONS) but have not been used much for helioseismology. Intensity oscillations mimic the velocity oscillations in most respects except that they are superimposed on a noisier background and they are phaseshifted relative to the velocity oscillation by roughly one-quarter of a period. The sense of the phaseshift is that brightest emission occurs when the atmosphere is most compressed downward, consistent with a nearly adiabatic compression. The amplitude of intensity oscillations is about 1% (depending on spectral wavelength), and does not decrease strongly away from the disk center.

If we enlarge the measured area of the Sun, the oscillation signal decreases with the square root of the area. This demonstrates that the signals are in phase and coherent over only a small spatial distance. This finding led early observers to conclude that the oscillations were a local, atmospheric phenomenon. The picture changed as simultaneous observations at many spatial locations became possible, and as theoretical work progressed. The birth of helioseismology occurred when it was realized that the oscillations seen at a single point on the Sun are the incoherent addition of millions of larger-scale oscillations. Shortly afterward it was also recognized that some of these oscillations are global in scale and sample the average properties over large portions of the solar interior.

Helioseismic observations are made using a wide range of solar instruments. Conceptually the simplest is just a stable detector exposed to sunlight. Since the amplitude of the 5 min intensity oscillation averaged over the entire solar disk is about 10^{-6}, as long as this instrumental set-up can accurately detect tinier intensity fluctuations in less than a minute, it can be used for helioseismology. The Earth's atmosphere is generally too noisy to permit this to succeed from the ground, but it has been done from space and from high-altitude balloons. Adding optical filters to restrict which wavelengths reach the detector enables more useful observations. These types of observation have also been done in space. Next in complexity are intensity measurements made with small telescopes that form images of the Sun's disk on detector arrays (segmented diodes or charge coupled device arrays). The primary requirement for using such instruments is that the noise from all sources at a single detector element be less than the amplitude of the solar oscillations, and preferably less than the non-oscillatory solar noise sources as well. This enforces stringent requirements on guiding stability, but such observations are relatively easy to do from the ground and have been made since the mid-1980s.

The next jump in instrumental complexity is to observe the Doppler shift of a single solar spectral line. Here the simplest instrument consists of a detector preceded by a spectrometer that isolates the red and blue wings of a spectral line so that the relative intensities of unimaged sunlight can be interpreted as a Doppler shift. Since the two isolated wavelengths are so close together, many noise sources affect both intensities almost identically. This enables high-quality, ground-based observations. However, since the amplitude of the 5 min oscillation averaged over the solar disk is only about 0.7 m s^{-1}, the spectrometer must be sensitive and stable. The spectrometer of choice is a cell filled with heated sodium or potassium vapor in a magnetic field. Such cells can be arranged to scatter or transmit light preferentially at discrete wavelengths in the wings of a resonance line of the element. These instruments have been in use since the mid-1970s. Vapor cells are also used in imaging instruments, but by sacrificing some stability (the required noise level at an element of an imaging instrument is about 10 m s^{-1}), more spectral lines can be observed with other spectrometers. The current instrument of choice is a temperature-stabilized, narrow-band birefringent filter, often combined with a rapidly tunable Michelson interferometer. These instruments allow simultaneous measurements of Doppler shift to be made over a well-resolved image of the Sun. Filter systems have been in use since the mid-1980s. The number of elements that can simultaneously be observed across the solar disk has increased by about an order of magnitude per decade as detector technology has improved. Current observations use up to 10^6 individual elements.

Trapped oscillations and global helioseismology

For a nearly perfect sphere such as the Sun it is appropriate to use spherical harmonic functions to represent the trapped modes. A single mode can be represented as a single spherical harmonic using a set of three integers l, m, n. The degree, l, is the number of nodal lines on the solar surface, and for p modes ranges from zero to thousands; it is a measure of the spatial wavenumber of the mode—from the scale of the entire Sun to the scale of granulation. The azimuthal order, m, is a measure of the number of nodal lines that cross the equator, and ranges from zero to plus and minus the degree number; it represents the concentration of a mode toward the equator. Roughly speaking, the radial order, n, is the number of nodal surfaces (in the vertical component of the velocity) encountered in the radial direction, and is seen in observations to range from zero to nearly 50. Theory predicts, and observations confirm, that the Sun is oscillating in millions of different p modes simultaneously. Of these, about 3×10^5 are long lived enough to remain coherent around a solar circumference and are the foundation of global helioseismology. The remaining modes, of higher degrees, are not globally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

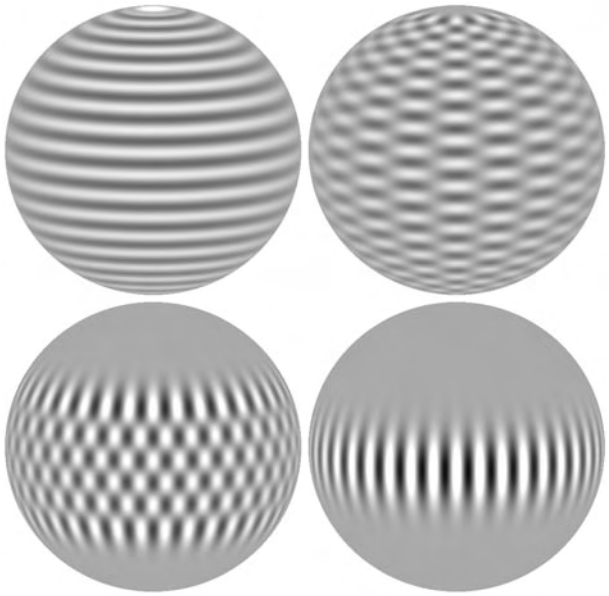


Figure 2. Spherical harmonic representations of p-mode oscillations of degree 40. The azimuthal orders in sequence are 0, 10, 35 and 40. The line-of-sight component of material motion is represented as light and dark shades of gray. Seen from a latitude of 10 degrees.

coherent and are used mainly for local helioseismology (SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY).

Different modes sample different volumes of the solar interior. That is, they are confined within resonant cavities that extend over different ranges in depth and latitude. Figure 2 represents four different global p modes showing their extent in latitude. Since p-mode frequencies are controlled by the sound speed within the sampled volumes, it is possible to deduce the radial variation of sound speed by measuring the frequencies of many different p modes. P-mode frequencies are also affected by the average rotation within the sampled volume, and the frequency variations allow the rotation profile with depth and latitude to be unraveled (SOLAR INTERIOR: ROTATION). Thus, the essence of global helioseismology is to measure the frequencies of p-mode (and f-mode) oscillations as accurately as possible to deduce interior conditions (SOLAR INTERIOR: INVERSIONS OF HELIOSEISMIC DATA).

The major task for global observational helioseismology is to produce an accurate spectrum of solar oscillations in terms of spherical harmonic integers and cyclic frequency, ν. In such a spectrum, the global p modes appear as peaks of power at specific values of l and m at specific frequencies that depend on n. Displaying a three-dimensional spectrum is difficult, and it is common to make two-dimensional projections. A frequent projection is to select an l value and show the spectrum as a function of m and ν (figure 3). This projection reveals that the frequencies of p modes vary with m value, and this provides the basis for determining the internal rotation of the Sun

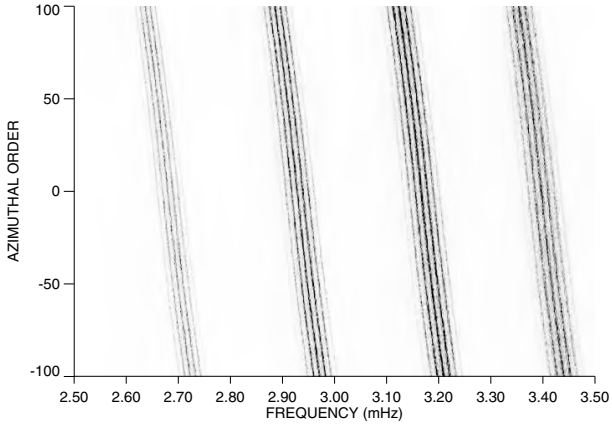


Figure 3. A section of the observed spectrum of p-mode oscillations for spherical harmonic degree 100 and four values of radial order. Dark indicates more power. The frequency variations are caused by solar rotation. Spatial leakage produces the finely spaced patterns. (Data from the GONG project, managed for the National Science Foundation by the National Solar Observatory.)

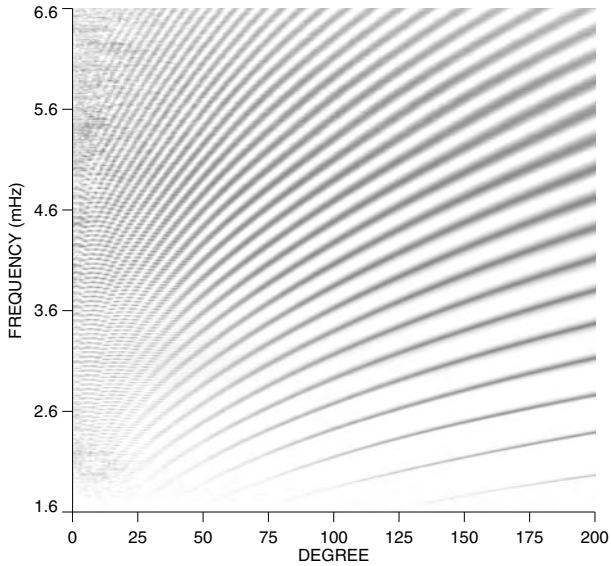


Figure 4. A portion of the observed spectrum of solar oscillations averaged over m after a correction for solar rotation. Dark indicates more power. The spectrum has been processed to reduce the dynamic range so that oscillations over most of the span can be seen. Note individually resolved modes to the left. (From observations at the geographical South Pole by an NSF-supported Bartol/NASA/NSO team.)

as well as departures from spherical symmetry. An even more popular projection removes the effects of solar rotation and then averages the result over m. The resulting spectrum, displayed as a function of l and ν, represents the spectrum that the Sun would produce if it were not rotating. An example of this kind of spectrum is figure 4.

To help define the observational requirements needed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to obtain the best possible spectrum of solar oscillations, we note that a time series of images and the l, m, ν spectrum are transform pairs of each other. Thus, if we want excellent frequency (ν) resolution, we need to acquire images for a long period of time. Similarly, to get the best possible resolution in l and m, the entire surface of the Sun should be observed. At present this is technically impossible and sets limits to the accuracy of a spectrum of solar oscillations. Seeking guidance from theory, we expect that p-mode oscillations should range in l, m from zero (the entire surface of the Sun oscillating in unison) to about 4000 (a horizontal scale of about $1/4000$ of the solar circumference or 1.1 Mm). The highest frequency p-mode oscillations that are expected to be reflected by the strong gradients in the solar atmosphere are about 5.6 mHz (180 seconds period). This means that we should sample the oscillations at least as frequently as every 90 seconds. One of the surprises of helioseismology was that mode-like oscillations are observed at higher frequencies. Some of these high-frequency oscillations may be caused by reflections high in the solar atmosphere (TRANSITION REGION MODELS), but most are thought to result from interference of outward traveling waves from a source with waves from the source that travel once through the solar interior. These waves have been clearly seen to frequencies of 11 mHz and there is evidence that they exist to a much higher frequency. Their amplitude weakens rapidly with increasing frequency and, as a practical matter, most helioseismic observations are made once per minute with little problem from inadequately sampled, weak, high-frequency waves.

A few low-frequency, low-degree p modes are expected to last without significant change for several years. This means that observations should continue for that long in order to measure the frequencies of long-lived modes accurately. Data gaps generate noise and other spectral artefacts, and observers strive to make their data as continuous as possible to produce the cleanest spectra.

Since we cannot at present observe oscillations simultaneously over the entire solar surface, there is a spread in the range of l, m values that are close matches to an oscillation pattern observed only over the visible hemisphere of the solar surface. This ambiguity causes smearing and artefacts in the spectrum. The best that can be done is to observe the full solar disk and try to use knowledge of the leakage to estimate the spectrum as if there had been no leakage.

Several steps are needed to convert a time series of Doppler velocity or intensity measurements to a helioseismic spectrum. For unimaged measurements, gaps in the time series are filled (generally using autoregressive interpolation), and a Fourier transform of the time series is computed. The result is a spectrum of the low-degree oscillations ($l = 0$ to 3). Imaged observations undergo the same steps, but each image is first decomposed into a set of spherical harmonic coefficients. It is the time series of these coefficients that are Fourier transformed to produce the oscillation spectrum. There are numerous variants of these basic steps.

Once a spectrum is produced, the frequencies of the modes must be determined. This is not trivial because the spectra are noisy, there is a varying background beneath the peaks, the shape of the peaks is intrinsically asymmetric, and there is leakage of spectral features from nearby values of l, m and ν.

Local helioseismology and short-lived oscillations

Several techniques have been developed to study the local structure and dynamics of the solar interior. The various methods use observations that are essentially the same as those used for global helioseismology. The primary difference is in how they are analysed. The simplest method consists of isolating a region of the solar disk and producing a Fourier transform spectrum of the oscillations of the region in terms of spatial wavenumber and temporal frequency. Such a spectrum is very similar to that shown in figure 4, but is usually projected into a k_x, k_y plane at a given frequency. It has the appearance of rings of power. The rings are distorted if the region is moving, or if it has a structural asymmetry. Analysis usually involves finding the flow and temperature distributions that best fit the observed distortions.

Another method is based on a Hankel spatial transform and Fourier temporal transform of p-mode oscillation patterns. This produces a spectrum in which inward and outward pattern propagation directions are separated with respect to a single point on the solar surface. If the central point is a sunspot or strong magnetic region, it is found that far more power moves inward than outward. Thus, sunspots are acoustic sinks (SUNSPOT OSCILLATIONS AND SEISMOLOGY). A spatial map of this difference can be built up by successively moving the center of the analysis to different positions.

A completely different method is closely related to terrestrial earthquake seismology. The time is measured for a particular oscillation to travel from one point on the solar surface, through the interior, to another point. Time differences can be related to variations of the properties of the material through which the wave motion passed. Figure 5 shows a so-called time–distance plot averaged over a large part of the solar disk. Plots like this are constructed by correlating the oscillation signal at a point with the signal at varying distances away as a function of time lag. Usually, the distant signal is averaged over a complete or partial annulus centered on the target point. The analysis can also be done backward to measure travel time differences for waves moving in the opposite direction. The time–distance technique is sensitive to geometrical image distortion and timing errors, so these potential observational problems require close attention. It is also necessary to carefully isolate just those waves that penetrate to a desired depth and have a particular frequency range since travel time is a strong function of frequency owing to surface effects. There are also tricky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

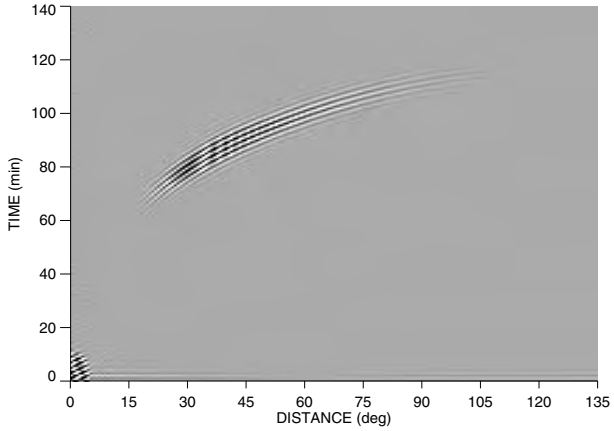


Figure 5. Observed correlation between high-frequency untrapped oscillations at a point and oscillations at different times and distances. The curved lines are prominent at the times and distances that correspond to one passage of the waves from near the surface, through the interior, to the surface again. Only oscillations with frequencies of 8.7±1 mHz and that reached a depth of about 150 Mm were used to compute this diagram. (Observations obtained at the South Pole by an NSF-supported Bartol/NASA/NSO team.)

problems of interpretation when the length of the wave is comparable to the scale of variation of the solar interior.

The newest technique for local helioseismology uses observations of the wave field at the solar surface and mathematically propagates them through the equivalent of a lens. The 'lens' can be focused at various depths to make images of features that perturb the normal progress of the wave motions. This method is called helioseismic holography because of its similarity to laser holography

The state of the art

In the 1970s, helioseismic observations were made using instruments designed for other purposes and located at single ground-based observatories. By the mid-1980s, the capabilities of such observations were fully exploited. Time resolution and continuity were improved by deploying simple, inexpensive instruments around the world in networks, and by mounting expeditions to the geographical South Pole to take advantage of continuous sunlight during the austral summer. Observations were also made from space serendipitously using an instrument designed to study solar irradiance. As solid state detector arrays became available, instruments were built to exploit them for helioseismic observations.

By the early 1990s, instrumental designs had matured, and funding agencies were convinced of the scientific merit of helioseismology, so that today, helioseismic observations have been done by a variety of space and ground-based projects. This complementarity is vital since it is the only way to detect systematic errors that can seriously affect helioseismic measurements. Space provides the great advantage of continuous sunlight unaffected by the vagaries of weather and atmospheric distortions. On the other hand, ground-based facilities can be repaired and upgraded, and are not limited by weight, power and data bandwidth constraints. Both types of observation are needed.

The major helioseismic space instruments commenced observations early in 1996 on board SOHO, the Solar and Heliospheric Observatory, a joint project of ESA and NASA. There are three helioseismic instruments on SOHO: GOLF, VIRGO and SOI/MDI. GOLF (Global Oscillations at Low Frequencies) is a non-imaging instrument that uses a sodium vapor resonance spectrometer to detect low-degree Doppler velocity oscillations. VIRGO (Variability of solar IRradiance and Gravity Oscillations) measures solar intensity oscillations with non-imaging detectors operating at three wavelengths, and also provides 12-element resolution of the solar disk intensity. SOI-MDI (Solar Oscillations Investigation—Michelson Doppler Imager) resolves the solar disk using a 1024^2 element CCD camera and a narrow-band tunable filter that provides Doppler velocity, continuum and spectral line intensity, and longitudinal magnetic field. Because of data bandwidth restrictions, only subsets of the data can be provided continuously. One such product is a reduced resolution Doppler image.

Future space helioseismic missions may seek higher angular resolution and higher time cadence (both important for local helioseismology), and helioseismology from a direction away from the Earth (important for studies of the solar poles and for deep probing of the solar core). A space mission may also be warranted if a promising way to detect g modes can be devised.

Current ground-based helioseismic observations are primarily produced by international networks established to obtain high observational duty cycles. Duty cycles of up to 90% can be obtained using six sites well distributed in longitude. The longest operating network is BiSON (Birmingham Solar Oscillation Network) organized by the University of Birmingham and operating continuously since 1981. It consists of six non-imaging potassium vapor resonance spectrometers. A similar set of instruments, except using sodium vapor, is organized in the IRIS (International Research on the Interior of the Sun) project. This project was started by the Université de Nice and has been taking continuous data since 1989. It has recently joined efforts with other projects to enhance the quality of non-imaged helioseismic data.

Ground-based imaging networks date back to 1994 when the TON (Taiwan Oscillation Network, managed by Tsing Hua University) started taking data continuously at three sites. The data are intensity images of the Sun filtered by a 1 nm filter centered on the Ca II K line. These intensity images have a diameter of about 1000 pixels and are obtained once per minute. In 1995, the GONG (Global Oscillation Network Group) started continuous Doppler velocity observations at six sites. This project is managed by the National Solar Observatory for the National Science Foundation. The GONG instruments use a Michelson interferometer to produce rapid modulation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of 220 pixel diameter images from which the Doppler shift can be determined. The data also include intensity and line strength images once per minute and line-of-sight component magnetograms less frequently. It is planned to continue the operation of GONG through a complete solar cycle and to increase the resolution to an 800 pixel diameter image size.

Several smaller networks are operational. The HdHN (High Degree Helioseismology Network) is a cooperation between the University of Southern California and the Crimean Astrophysical Observatory. The two sites operate resonance vapor spectrometers arranged to provide Doppler velocity images with diameters between 500 and 1000 pixels. The RODOMA (Roma Doppler and Magnetic) network intends to provide Doppler velocity and line-of-sight magnetograms from sites in California, Austria and Uzbekistan. These instruments will use potassium resonance spectrometers in an imaging mode with an image diameter of about 400 pixels. The ECHO (Experiment for Coordinated Helioseismic Observations) network is a cooperation between the High Altitude Observatory and the Instituto Astrofisica de Canarias, and will have sites in Hawaii and the Canary Islands to provide roughly 200 pixel diameter Doppler images, also using potassium resonance spectrometers.

Although non-imaged helioseismic measurements have been available more or less continuously since the early 1980s, imaged data were very sporadic until the mid-1990s. Notable exceptions were observations made at Big Bear Solar Observatory, at the South Pole, and a collaboration between the High Altitude Observatory and National Solar Observatory. Many of the basic discoveries of helioseismology were made using these observations. A revolution has occurred in the last few years with the availability of nearly continuous, high-quality imaged and non-imaged observations from space and ground-based networks.

The quality of helioseismic observations is currently very good. Signal-to-noise ratios are limited mainly by the Sun itself. The major lack is continuous intensity and Doppler velocity measurements at a resolution of the order of 1000 pixels per diameter. It would also be beneficial to have simultaneous observations at several heights in both intensity and Doppler velocity. A more rapid cadence of observation would reduce small effects of aliasing of high-frequency oscillations into low frequencies. Almost all imaged observations suffer from aliasing of spatial information. The only cure is to ensure that no spatial frequencies are present above the Nyquist sampling frequency.

Spectral leakage of modes is a serious problem for global helioseismology. This corrupts the oscillation spectrum so that when many modes overlap it is difficult to assign a unique frequency to each one. The situation is even worse when the individual modes cannot be resolved. An additional complication is an asymmetry in the spectral line profiles which varies with frequency, degree, and whether intensity or velocity is observed. Several explanations for this phenomenon have been suggested, but a consensus has not yet been achieved.

The most outstanding observational problem is to detect g modes. Many detections have been claimed in the past, but none has survived careful scrutiny. The problem is that noise from non-oscillatory processes is strong in the frequency range where g modes are expected, and, evidently, the g modes are weak. It seems likely that some new observational approach is required to cancel the effects of background noise while preserving the coherent g-mode signals.

The major growth area in observational helioseismology is to detect time variability of previously inaccessible regions of the Sun. Global helioseismology techniques have recently revealed zonal velocity flows near the base of the convection zone that vary with a 1.3 year period. Helioseismic holography has recently been used to make images showing the eruption and evolution of solar active regions on the far side of the Sun, many days before they rotate onto the side facing Earth. These results are initial contributions of helioseismology to a better understanding of the mystery of the solar activity cycle (SOLAR CYCLE). The new ability to make images of the far side of the Sun should become a powerful tool for the study and prediction of space weather.

Since the birth of helioseismology, the confrontation of solar models with observations has stimulated improvements in the models so that current agreement is at the one per cent level. In other words, the theoretical solar model is basically correct. Thus, a solar physics explanation for the deficit of neutrinos observed from the Sun was ruled out several years ago on the basis of helioseismology (SOLAR INTERIOR: NEUTRINOS). However, the disagreement between observed and predicted frequencies is far larger than the errors of measurement. Until recently, using improved physics in the models produced better agreement. Currently we have the odd situation that improving the physics in the models causes larger discrepancies. Evidently there is a significant missing element in the solar models. Tracking that down should yield another in the long series of surprises produced by helioseismology.

Starting in 1975, helioseismic observations have given us a continually improving picture of the nature of the solar interior. There have been many surprises. Perhaps the most sobering one is that many of our preconceptions about the nature of the solar internal structure and dynamics were found to be incorrect—sometimes spectacularly so. While much has been learned, much remains to be done.

Bibliography

Gough D and Toomre J 1991 Seismic observations of the solar interior *Ann. Rev. Astron. Astrophys.* **29** 627–85

Harvey J 1995 Helioseismology *Phys. Today* **48**(10) 32–8

Hill F, Deubner F-L and Isaak G 1991 Oscillation observations *Solar Interior and Atmosphere* ed A

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cox, W Livingston and M Matthews (Tucson, AZ: University of Arizona Press) pp 329–400
Stix M 1989 *The Sun—an Introduction* (Berlin: Springer)
Unno W, Osaki Y, Ando H, Saio H and Shibahashi H 1989 *Nonradial Oscillations of Stars* 2nd edn (Tokyo: University of Tokyo Press)

John Harvey

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helioseismology

The study of the internal structure and dynamics of the Sun through the analysis of solar oscillations. The Sun vibrates like a gong with periods ranging from minutes to hours, but principally in the range 3 to 20 min. These oscillations can be detected as small periodic Doppler shifts in the wavelengths of spectral lines emitted by localized regions of the solar surface as they rise toward, and fall away from, the observer. The oscillations are produced by sound waves (pressure waves) that propagate through the solar globe. The speed of sound depends on various factors, including temperature and density, both of which increase with increasing depth below the solar surface. Consequently, a wave moving downward from a point on the surface is refracted and eventually curves back to meet the surface at another point. The sharp change in density at the surface then reflects the wave back down into the solar interior, enabling it to bounce repeatedly around the Sun and thereby produce standing waves that cause different parts of the solar surface to vibrate up and down in a systematic fashion. The deeper the wave penetrates, the fewer the points at which it meets the surface. By analyzing the millions of different modes of oscillation, and separating out those that penetrate to different depths, solar physicists can study the structure of the solar interior in the same sort of way as geophysicists use seismic waves to study the interior of the Earth. Furthermore, by comparing the speeds at which waves travel in the same direction as, and in the opposite direction to, the rotation of the Sun itself, it is possible to determine how the rotation rate of the solar interior varies with depth and latitude.

Helioseismology provides information on the variation with depth of temperature, density, pressure, chemical composition and rotational velocity. Observations made by ground-based projects such as the Global Oscillation Network Group (GONG) and instruments carried on spacecraft such as SOHO, have shown, for example, that the boundary between the radiative and convective zones occurs 71.3% of the way from the center to the surface and that the differential rotation that is observed in the photosphere extends down to the base of the convective zone.

See also: differential rotation, Doppler effect, SOHO (Solar and Heliospheric Observatory), Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helioseismology: Theory

Observations have shown that the solar surface is in a state of continual oscillation, involving a very large number of different modes (see HELIOSEISMIC OBSERVATIONS). The observed frequencies depend in a relatively simple manner on the structure and rotation of the solar interior. To use this information to probe the Sun we need to understand how the properties of the oscillations reflect conditions in the solar interior.

The Seismic Sun

The observed solar oscillations share many similarities with the waves used in geoseismology to probe the interior of the Earth. However, while geoseismology is mostly based on travel times of waves travelling, e.g. from an Earthquake to a seismograph, global helioseismology uses frequencies of standing waves in the Sun (recently, analyses more akin to geoseismology have also been developed: see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY). Even so, the observations can be discussed conveniently in terms of propagation of waves through the Sun. The waves relevant in the observed frequency range are acoustic waves, whose properties depend primarily on the sound speed, which is roughly proportional to the square root of temperature. Thus, as one moves to greater depth within the Sun, the sound speed increases. The acoustic waves are believed to be generated immediately below the solar surface, by the intense turbulence in the uppermost part of the convection zone (see SOLAR INTERIOR: CONVECTION THEORY). With the exception of those travelling radially, the waves are refracted by the increasing sound speed (cf figure 1), as they propagate towards the solar interior; they undergo total internal refraction at the so-called lower turning point, at a distance r_t from the solar center. When the wave reaches the surface again, it is reflected by the rapidly shrinking scale of variation of the density (at least for frequencies below about 5 mHz). Thus the wave travels in a series of 'bounces' between the surface and the turning point.

Interference between such bouncing acoustic waves gives rise to the observed modes; since the dominant restoring force for acoustic waves is pressure, the modes are known as p modes. Conditions of regularity for waves propagating in a sphere enforce a specific pattern to these standing waves: as a function of position on spherical surfaces they behave as spherical harmonics (see also HELIOSEISMIC OBSERVATIONS, figure 2). This defines two of the three wavenumbers characterizing a given mode: the degree l and the azimuthal order m. The degree measures the total number of nodal lines on the solar surface; very roughly, the 'average horizontal wavelength' on the surface is given by $\lambda_h \simeq 2\pi R/l$ where R is the solar surface radius. For modes with $l = 0$ (the so-called *radial modes*) the Sun preserves its spherical shape while pulsating. The azimuthal order m is between $-l$ and l; it provides a measure of the number of nodal lines crossing the equator. For each (l, m) the Sun has a set of modes distinguished by the third wave number, viz. the radial order n which essentially measures the number of nodal surfaces in the radial direction.

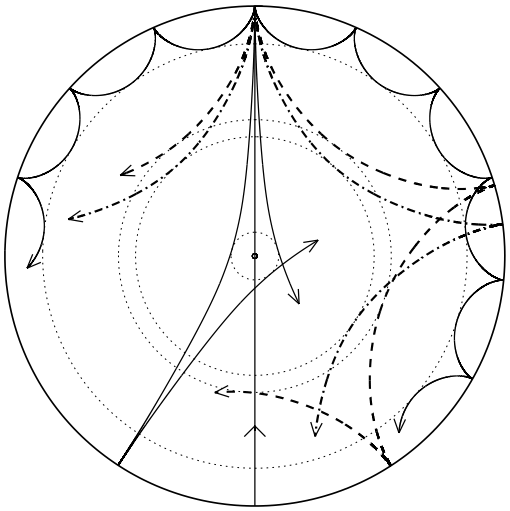


Figure 1. Propagation of rays of sound in a cross-section of the solar interior. The ray paths are bend by the increase in sound speed with depth until they reach *the inner turning point* (indicated by the dotted circles) where they undergo total internal refraction. At the surface the waves are reflected by the rapid decrease in density.

As is clear from figure 1 there is a relation between the initial direction of propagation of the waves, the depth of the turning point and the separation between successive bounces at the surface: modes propagating more nearly vertically reach a greater depth and returns to the surface further from the start. More precisely, the location of the turning point is determined by the relation $c(r_t)/r_t = 2\pi\nu/(l + 1/2)$, where c is the speed of sound and ν is the (cyclic) oscillation frequency, $\nu = 1/\Pi$ where Π is the pulsation period. The observed modes cover a range in degree from zero to more than one thousand. Correspondingly, the turning points range from very near the center to a few thousand kilometres below the solar surface. This is what allows inverse analyses (see SOLAR INTERIOR: INVERSIONS OF HELIOSEISMIC DATA) to infer local properties, in particular the sound speed, within the Sun. The frequency of a given mode depends essentially only on the structure outside the turning point. Hence, by combining frequencies of two modes with slightly differing turning points (see the dashed and dash-dotted rays in figure 1), information is obtained about conditions between the turning points. Needless to say, the actual analysis is substantially more complex and involves a large number of modes; but the principle, corresponding essentially to peeling off layer after layer of solar structure as the layers of an onion, still holds. In particular, the solar internal sound speed has been determined with high precision and resolution; comparison with the sound speed in solar models shows both generally good

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

agreement and some very intriguing differences (see SOLAR INTERIOR: STANDARD MODELS).

The oscillation frequencies also depend on the dynamics of the solar interior. Particularly important is the effect of rotation. The spherical-harmonic pattern is advected by rotation, increasing the observed frequency of modes travelling in the same direction of rotation and decreasing it for modes travelling in the opposite direction. The resulting so-called *rotational splitting* essentially reflects an average of the solar angular velocity over that part of the Sun which is occupied by the given mode. In addition to the variation in radial extent, there is a variation of the extent in latitude with the azimuthal order, modes with $|m| = l$ being concentrated near the equator, while modes with low m extending over the entire range in latitude. Thus, similarly to the case of structure, discussed above, it is possible carry out inversions to infer the angular velocity as a function of both latitude and depth. The results reveal a complex pattern of rotation which is so far not understood (see SOLAR INTERIOR: ROTATION).

It should finally be noticed that, in addition to the acoustic modes discussed so far, other types of modes are possible in the Sun. At relatively high degree one observes the so-called f modes, which are essentially surface gravity waves. Furthermore, solar models show the possible existence of internal gravity (or g) modes, with amplitudes strongly concentrated near the core of the model. Such modes would therefore provide excellent diagnostics of the structure and rotation of the solar core; unfortunately, however, they have so far not been definitely observed.

Properties of solar oscillations

The goal of the theoretical treatment of helioseismology is to relate the observable properties of the oscillations, in particular the frequencies, to the properties of the solar interior. In principle, the oscillations should be described by the complete set of magnetohydrodynamical equations governing the behavior of the solar interior. The description should then also include the convective motions in the outer approximately 28% of the solar radius, as well as the coupling to the radiation field in the solar atmosphere.

In practice, many simplifications are possible, making the problem tractable. The amplitudes of the observed modes are so small that the oscillations can be regarded as small perturbations around a state of equilibrium. Also, the effects on the oscillations of rotation and magnetic fields can be regarded as small perturbations, relative to oscillations of a spherically symmetric equilibrium model; the latter corresponds to a normal model of the solar interior. Finally, the characteristic thermal time scale in the most of the Sun is far longer than the pulsation period, and as a result the modes can to a good approximation be regarded as adiabatic, occurring without heat loss or gain; also, microscopic viscosity can certainly be neglected. Although some of these simplifications become questionable near the solar surface, they indicate that the basic problem of helioseismic theory is that of small-amplitude adiabatic oscillations around a spherically symmetric equilibrium structure. This is sufficiently simple that the frequencies of a given solar model can be computed with high precision; consequently, the observed frequencies provide a clean measure of solar structure which is relatively simple to interpret.

We now turn to the properties of such adiabatic oscillations.

Equations of linear oscillations

We assume that the amplitudes are so small that quantities of higher order than one in the perturbations can be neglected. It is convenient to work in terms of the displacement $\boldsymbol{\delta r}$ induced by the oscillation, related to the oscillation velocity $\boldsymbol{v}$ by

$$\boldsymbol{v} = \frac{\partial \boldsymbol{\delta r}}{\partial t} \, .$$

Two descriptions of the perturbations are used: the Eulerian perturbations, indicated by a prime, are at a given location, while the Lagrangian perturbations, indicated by δ, correspond to following the motion. The two are closely related: for example, the Eulerian and Lagrangian pressure perturbations p' and δp satisfy

$$\delta p = p' + \boldsymbol{\delta r} \cdot \nabla p_0 \, , \tag{1}$$

where p_0 is the pressure in the equilibrium model.

It is now straightforward to write down the linearized equations of hydrodynamics:
Equation of continuity:

$$\rho' + \mathrm{div}(\rho_0 \boldsymbol{\delta r}) = 0 \, ; \tag{2}$$

equations of motion:

$$\rho_0 \frac{\partial^2 \boldsymbol{\delta r}}{\partial t^2} = -\nabla p' + \rho_0 \nabla \Phi' + \rho' \nabla \Phi_0 \, ; \tag{3}$$

and Poisson's equation:

$$\nabla^2 \Phi' = -4\pi G \rho' \, . \tag{4}$$

Here ρ is density, Φ is the gravitational potential, and G is the gravitational constant. In addition, the perturbations in pressure and density are related by the adiabatic relation,

$$\delta p = \frac{\Gamma_{1,0} p_0}{\rho_0} \delta \rho \tag{5}$$

where $\Gamma_1 = (\partial \ln p / \partial \ln \rho)_s$, the derivative being at constant specific entropy s.

The analysis is greatly simplified by the fact that the equilibrium structure is spherically symmetric and time independent. We introduce a spherical polar coordinate system (r, θ, ϕ), where r is the distance to the center, θ is co-latitude, i.e. the angle from the polar axis, and ϕ is longitude. Then the solution is separable in time, and in the angular coordinates (θ, ϕ). The time dependence is naturally expressed as a harmonic function, characterized

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by an angular frequency ω. The dependence on (θ, ϕ) may be obtained from an analysis of the way in which the angular coordinates enter into the equations. The result is that the expansion function in (θ, ϕ) must be an eigenfunction of the tangential part $\nabla_{\rm h}^2$ of the Laplace operator. This can be taken to be a spherical harmonic,

$$(-1)^m c_{lm} P_l^m(\cos\theta) \exp(\mathrm{i}m\phi) \equiv Y_l^m(\theta, \phi) \tag{6}$$

where P_l^m is a Legendre function and c_{lm} is a normalization constant, such that the integral of $|Y_l^m|^2$ over the sphere is unity. Here l and m are integers, such that $-l \le m \le l$; Y_l^m satisfies

$$\nabla_{\rm h}^2 Y_l^m = -\frac{l(l+1)}{r^2} Y_l^m \ . \tag{7}$$

Some properties of the spherical harmonics were discussed above. Also, it is useful to note that if the oscillation is approximated locally by a plane wave, with horizontal wave number $k_{\rm h}$, then

$$\frac{l(l+1)}{r^2} \simeq k_{\rm h}^2 \tag{8}$$

according to equation (7). This identification is very useful in the asymptotic analysis of the oscillations.

It is convenient, in the mathematical analysis, to express the solution on complex form as is indeed implied by equation (6). The physical variables are obtained by taking the real part of the complex solution. Thus after the separation of variables the pressure perturbation, for example, can be expressed as

$$p'(r, \theta, \phi, t) = \Re[\tilde{p}'(r) Y_l^m(\theta, \phi) \exp(-\mathrm{i}\omega t)]$$

in terms of an amplitude function $\tilde{p}'$ which depends just on r; here $\Re$ denotes the real part. Also, it follows from the equations of motion that the displacement vector can be written as

$$\boldsymbol{\delta r} = \sqrt{4\pi}\,\Re\left\{\left[\tilde{\xi}_r(r) Y_l^m(\theta, \phi)\boldsymbol{a}_r + \frac{\tilde{\eta}(r)}{L}\left(\frac{\partial Y_l^m}{\partial\theta}\boldsymbol{a}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_l^m}{\partial\phi}\boldsymbol{a}_\phi\right)\right]\exp(-\mathrm{i}\omega t)\right\} \tag{9}$$

where

$$\tilde{\eta}(r) = \frac{L}{r\omega^2}\left(\frac{1}{\rho_0}\tilde{p}' - \tilde{\Phi}'\right) \tag{10}$$

and $L = \sqrt{l(l+1)}$; in equation (9) $\boldsymbol{a}_r$, $\boldsymbol{a}_\theta$ and $\boldsymbol{a}_\phi$ are unit vectors in the r, θ and ϕ directions. With this definition $\tilde{\xi}_r$ and $\tilde{\eta}$ are essentially the root-mean-square radial and horizontal displacements.

Given the separation of variables, the equations of stellar pulsation are reduced to a set of four ordinary differential equations for the amplitude functions $\{\tilde{\xi}_r, \tilde{\eta}, \tilde{\Phi}', \mathrm{d}\tilde{\Phi}'/\mathrm{d}r\}$. These must be combined with boundary conditions: two of these ensure regularity at the center, $r = 0$, which is a regular singular point of the equations. One condition enforces continuity of Φ' and its gradient at the surface, $r = R$. Finally, a dynamical condition must be imposed on the surface pressure perturbation. This in its most simple form imposes zero pressure perturbation on the perturbed surface, i.e.

$$\delta p = 0 \qquad \text{at} \qquad r = R \ . \tag{11}$$

Under the assumption of spherical symmetry, the frequencies are independent of m. Formally, m does not appear in the separated oscillation equations; physically, m is related to the equator in the coordinate system chosen, and this choice can have no effect in a spherically symmetric star.

The modes observed in the Sun are of high radial order or degree. In this case it is often possible to neglect the perturbation in the gravitational potential, at least in simplified analyses (Φ' should in general be included in numerical computations). In this so-called Cowling approximation, the order of the equations is reduced to two, greatly simplifying the analysis. In this case, furthermore, equation (10) directly relates the surface pressure perturbation and horizontal displacement. Using also equations (11) and (1), we obtain

$$\frac{\tilde{\eta}(R)}{\tilde{\xi}_r(R)} \simeq \frac{GM}{R^3}\frac{L}{\omega^2}$$

where M is the total mass; from this it follows that at the observed frequencies, and low or moderate degree, the oscillations are predominantly in the radial direction.

What do the frequencies depend on?

Adiabatic oscillations are essentially mechanical phenomena. Thus they depend on the forces in the star, i.e. pressure and gravity. In turn, the pressure is determined by the mass distribution, through the condition of hydrostatic equilibrium in the equilibrium model. Thus, as far as the oscillations are concerned, stellar structure is essentially characterized by the dependence $\rho(r)$ of density on distance to the center. In addition, the oscillations depend on the relation between the pressure and density perturbations, as characterized by Γ_1 (cf equation (5)). It follows that, given $[\rho(r), \Gamma_1(r)]$ for a stellar model, its adiabatic oscillation frequencies can be computed. This pair of variables can be replaced by other, equivalent pairs; since the observed solar oscillations are essentially acoustic modes, a natural choice is (c, ρ) where

$$c = \left(\frac{\Gamma_{1,0} p_0}{\rho_0}\right)^{1/2}$$

is the adiabatic sound speed. Conversely, to the extent that the adiabatic approximation is adequate, observed frequencies provide information only about such mechanical quantities.

Asymptotics of acoustic modes

A great deal of insight into the properties of the observed solar oscillations can be obtained from a simple analysis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of acoustic waves. The starting point is the dispersion relation for plane sound waves,

$$\omega^2 = c^2|\boldsymbol{k}|^2$$

where $\boldsymbol{k}$ is the wave vector. Separating $\boldsymbol{k}$ into radial and horizontal components k_r and k_h and using equation (8), we obtain

$$k_r^2 = \frac{\omega^2}{c^2} - \frac{L^2}{r^2} \tag{12}$$

where again $L = \sqrt{l(l+1)}$[1]. The interpretation of this equation is entirely in accordance with the discussion accompanying figure 1: with increasing depth and hence increasing sound speed, k_r decreases and hence the wave travels more and more horizontally. The lower turning point $r = r_\mathrm{t}$, where the wave travels horizontally, is reached where $k_r = 0$, i.e.

$$\frac{c(r_\mathrm{t})}{r_\mathrm{t}} = \frac{\omega}{L} \,. \tag{13}$$

For $r < r_\mathrm{t}$, k_r is imaginary and the wave decays exponentially. It follows from equation (13) that the lower turning point is closer to the center, the lower is the degree or the higher is the frequency. Radial modes, with $l = 0$, penetrate the center, whereas the modes of highest degree observed, with $l \gtrsim 1000$, are trapped in the outer 0.2 % of the solar radius.

To obtain a standing wave, i.e. a mode of oscillation, suitable conditions must be satisfied at the surface and at $r = r_\mathrm{t}$. These essentially induce phase changes in the wave which, together with the phase change between r_t and R, must equal an integer times π. It follows that the condition for a standing wave can be expressed as

$$\int_{r_\mathrm{t}}^{R} k_r \mathrm{d}r = (n+\alpha)\pi$$

or, using equation (12),

$$\int_{r_\mathrm{t}}^{R} \left(1 - \frac{c^2L^2}{r^2\omega^2}\right)^{1/2} \frac{\mathrm{d}r}{c} = \frac{\pi(n+\alpha)}{\omega} \,. \tag{14}$$

Here n is the mode order and α (which in general is a function of frequency) accounts for the phase changes at $r = r_\mathrm{t}$ and $r = R$. The observed frequencies of solar oscillation satisfy a relation of this general functional form with reasonable precision. It should be pointed out that equation (14) can be given a more solid justification through a formal asymptotic analysis of the oscillation equations.

The observed frequencies are most often analyzed by comparing them with frequencies computed for solar models. Thus we need to relate differences in frequencies to differences between the properties of the Sun and the model. Let $\delta\omega = \omega_\odot - \omega_\mathrm{mod}$ be the difference between the observed and the model frequencies, $\delta_r c(r) = c_\odot(r) - c_\mathrm{mod}(r)$ be the difference between the solar and model sound speeds, and $\delta\alpha(\omega) = \alpha_\odot(\omega) - \alpha_\mathrm{mod}(\omega)$ be the difference between phase changes in the Sun and model; then it may be shown from equation (14) that

$$S\frac{\delta\omega}{\omega} \simeq \mathcal{H}_1\left(\frac{\omega}{L}\right) + \mathcal{H}_2(\omega) \tag{15}$$

where

$$S = \int_{r_\mathrm{t}}^{R} \left(1 - \frac{L^2c^2}{r^2\omega^2}\right)^{-1/2} \frac{\mathrm{d}r}{c} - \pi\frac{\mathrm{d}\alpha}{\mathrm{d}\omega} \tag{16}$$

$$\mathcal{H}_1(w) = \int_{r_\mathrm{t}}^{R} \left(1 - \frac{c^2}{r^2w^2}\right)^{-1/2} \frac{\delta_r c}{c}\frac{\mathrm{d}r}{c} \tag{17}$$

and

$$\mathcal{H}_2(\omega) = \frac{\pi}{\omega}\delta\alpha(\omega) \,. \tag{18}$$

Thus, apart from the term in $\mathcal{H}_2$, the frequency difference is essentially a weighted average of $\delta_r c/c$. The weight factor has a simple physical meaning: $\mathrm{d}r/c$ is the sound travel time over the distance $\mathrm{d}r$; multiplying by the factor $(\ldots)^{-1/2}$ converts this into the sound travel time, corresponding to the radial distance $\mathrm{d}r$, along the relevant ray. Thus the average is weighted by the local sound travel time along the ray corresponding to the mode, as might indeed have been expected. The term in $\mathcal{H}_2$ is dominated by the change in the near-surface phase shift; thus it contains contributions from the treatment of the near-surface properties of the model and oscillations, as discussed below.

Equation (15) may be fitted to differences between observed and model frequencies. Even this simple type of analysis provides excellent insight into the likely sources of error in the model. An example is illustrated in figure 2. Panel A clearly shows that the scaled differences are predominantly a function of frequency; as discussed above, this indicates that the dominant sources of error in the model are in the near-surface region. However, as shown in panel B, there is a very significant contribution from the term in $\mathcal{H}_1(\omega/L)$, resulting from differences in the internal sound speed between the Sun and the model. Particularly striking is the variation near $r_\mathrm{t} \simeq 0.7R$, i.e. for modes which just penetrate beyond the convection zone, suggesting a relatively substantial and rather localized difference in sound speed in this region (cf equation (17)); this inference is indeed confirmed by more detailed analyses (see the article SOLAR INTERIOR: STANDARD MODELS).

Near-surface problems

The simple assumption of adiabatic oscillations breaks down near the surface, where the local thermal time scale is comparable to the oscillation period. Although it is possible in principle to solve the full set of nonadiabatic oscillation equations, in practice this is greatly complicated by the uncertainty about how to treat the perturbation, induced by the oscillations, in the convective flux. Furthermore, there are other uncertainties in the physics

[1] A more careful asymptotic analysis shows that this should be replaced by $L = l + 1/2$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

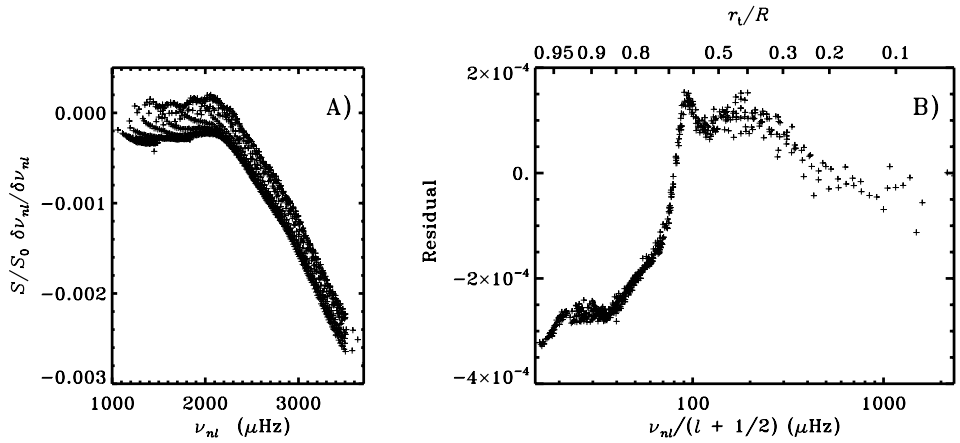


Figure 2. Panel A shows relative differences between observed and model frequencies, scaled by S/S_0, where S is defined in equation (16) and S_0 is its value in the limit $r_t \to 0$. The results are plotted against the cyclic frequency $\nu = \omega/(2\pi)$. Panel B shows the residuals after subtracting the function $\mathcal{H}_2(\omega)$, obtained by fitting the expression (15) to the data shown in panel A; these residuals essentially correspond to $\mathcal{H}_1(\omega/L)$. Here the abscissa is ν/L which is related to the location of the lower turning point, shown on the upper abscissa, by equation (13).

of the near-surface region. The model structure depends on the details of the treatment of the convective flux, usually grossly simplified, and of the convectively induced turbulent pressure which is often neglected. Perturbations in turbulent pressure are also likely to be significant, yet are very uncertain. Finally, the processes responsible for the excitation of the solar oscillations, which may also affect the oscillation frequencies, are likely to be concentrated near the solar surface.

Numerical modelling of this region, including a detailed treatment of convection and its interaction with the oscillations, is becoming computationally feasible, although not yet at a level which allows sufficiently precise predictions of the influence on the frequencies. However, it is still possible under many circumstances to suppress the frequency errors induced by our incomplete understanding of the near-surface layers. Except at high degree, the modes essentially propagate vertically in this region, with a horizontal wavelength that is much larger than the vertical wavelength (cf equation (12)). Consequently, the behavior of the modes, and the direct effect of the near-surface errors, is essentially independent of degree and hence is a function of frequency alone. The influence on the frequency does depend on degree, through the dependence of the penetration depth and hence the fraction of the solar mass which is involved in the oscillation (cf figure 1): compared with low-degree modes at the same frequency, high-degree modes are trapped close to the surface, involving little mass and hence their frequencies are more strongly affected by the near-surface errors. Quantitatively, the effect is a change $\delta_{\rm surf}\omega$ in the frequency, of the form $\delta_{\rm surf}\omega = E^{-1}\mathcal{F}_{\rm surf}(\omega)$, where

$$E = \frac{\int \rho|\boldsymbol{\delta r}|^2 \mathrm{d}V}{M|\boldsymbol{\delta r}(R)|^2} \qquad (19)$$

is a normalized mode inertia, which measures the fraction of the solar mass taking part in the oscillation; $\mathcal{F}_{\rm surf}$ is a function of frequency which depends on the properties of the near-surface layers. This scaling of the near-surface effects is largely equivalent to the asymptotic scaling by S of the frequency shifts in equation (15); in fact, the function $\mathcal{H}_2(\omega)$ in that equation contains a contribution from the near-surface errors, as described by $\mathcal{F}_{\rm surf}$.

In the analysis of observations, it is assumed that the differences between the observed solar frequencies and those computed for a solar model contain a contribution of the form given by $\delta_{\rm surf}\omega$, with an unknown function $\mathcal{F}_{\rm surf}(\omega)$. On the other hand, determination of $\mathcal{F}_{\rm surf}(\omega)$ serves as a diagnostics of the physics of the outermost layers, including possible effects of convection. In fact, as already seen in figure 2, the differences between observed frequencies and frequencies of current models are dominated by a contribution of this form; its physical origin is not fully understood, however.

Variational principle

Equation (15) still suffers from being based on a highly simplified asymptotic analysis. To obtain the full benefit of the very accurate observed frequencies, we need more precise relations between differences in frequencies and differences in solar structure; these may be obtained from the basic equations of adiabatic oscillation.

The starting point is equation (3), which we write as

$$\omega^2\rho_0\boldsymbol{\delta r} = \rho_0\mathcal{L}(\boldsymbol{\delta r}) \ . \qquad (20)$$

Here $\mathcal{L}(\boldsymbol{\delta r})$ is defined by the right-hand side of equation (3); note that ρ', p' and Φ' can in fact be obtained as functions of $\boldsymbol{\delta r}$, from the remaining linearized equations. From equation (20), ω^2 can be expressed as

$$\omega^2 = \frac{\int \boldsymbol{\delta r}\cdot\mathcal{L}(\boldsymbol{\delta r})\rho_0 \,\mathrm{d}V}{\int |\boldsymbol{\delta r}|^2\rho_0 \,\mathrm{d}V} \ . \qquad (21)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It may be shown that this is a variational expression: it is stationary to small changes in $\boldsymbol{\delta r}$. This provides the basis for analyzing the effects on the frequencies of changes in the equilibrium model or other aspects of the oscillation equations. Such changes can be expressed formally as a perturbation to the operator $\mathcal{L}$; it follows from the stationarity of equation (21) that the effect on the frequencies can be evaluated from that equation, using the eigenfunctions of the unperturbed operator.

A simple example of such a perturbation is a change in the structure of the equilibrium model. As discussed above, the helioseismic structure of the model can be specified by (c, ρ); consequently, the changes in the model, and hence in the operator $\mathcal{L}$, are characterized by the changes $\delta_r c$ and $\delta_r \rho$ in c and ρ, at fixed r. After some manipulation, the resulting change in the frequencies can be expressed as

$$\frac{\delta\omega}{\omega} = \int_0^R \left[K_{c,\rho}(r)\frac{\delta_r c}{c}(r) + K_{\rho,c}(r)\frac{\delta_r \rho}{\rho}(r) \right] \mathrm{d}r + E^{-1}\mathcal{F}_{\mathrm{surf}}(\omega) \ . \qquad (22)$$

Here the kernels $K_{c,\rho}$ and $K_{\rho,c}$ can be determined from the reference model, corresponding to the unperturbed $\mathcal{L}$; also, in accordance with the earlier discussion, we have included a term in $\mathcal{F}_{\mathrm{surf}}$ accounting for the near-surface errors. Equation (22) forms the basis for inverse analyses to determine corrections to the model sound speed and density. Note that although equation (22) was expressed in terms of (c, ρ), other variables may be used; in particular, if the equation of state of solar matter is assumed to be known, the frequency differences can be expressed in terms of, say, ρ and the abundance Y of helium, providing the basis for determination of the solar envelope helium abundance.

Effect of rotation

So far, we have considered only oscillations of a spherically symmetric star; in this case, the frequencies are independent of the azimuthal order m. Departures from spherical symmetry lift this degeneracy, causing a frequency splitting according to m.

The most obvious, and most important, such departure is rotation. Effects of rotation were already discussed in the introduction. The description provided there can be made somewhat more precise by noting that, according to equations (6) and (9), the oscillations depend on longitude ϕ and time t as $\cos(m\phi - \omega t)$, i.e. as a wave running around the equator. We now consider a star rotating with angular velocity Ω and a mode of oscillation with frequency ω_0 in a frame rotating with the star (see figure 3). Letting ϕ' denote longitude in this frame, the oscillation therefore behaves as $\cos(m\phi' - \omega_0 t)$. The longitude ϕ in an inertial frame is related to ϕ' by $\phi' = \phi - \Omega t$; consequently, the oscillation as observed from the inertial frame depends on ϕ and t as

$$\cos(m\phi - m\Omega t - \omega_0 t) \equiv \cos(m\phi - \omega_m t)$$

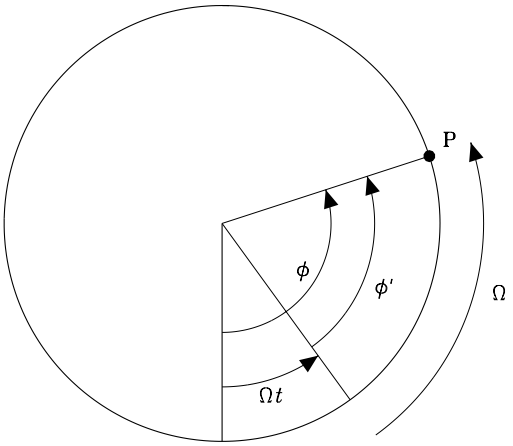


Figure 3. Geometry of rotational splitting, in a star rotating with angular velocity Ω. The point **P** has longitude ϕ' in the system rotating with the star and longitude $\phi = \phi' + \Omega t$ in the inertial system.

where

$$\omega_m = \omega_0 + m\Omega \ . \qquad (23)$$

Thus the frequencies are split according to m, the separation between adjacent values of m being simply the angular velocity.

This simple description contains the dominant physical effects of rotation on the observed modes of oscillation, but it suffers from two problems: it assumes solid-body rotation, whereas the Sun in fact rotates differentially (see DIFFERENTIAL ROTATION); and it neglects the effects, such as the Coriolis and centrifugal forces, in the rotating frame. The effects of differential rotation can be accounted for, roughly speaking, by replacing Ω in equation (23) by a suitable average; a natural expectation is that the appropriate weight is the energy density $\rho|\boldsymbol{\delta r}|^2$ in the mode. To obtain a full description, including also the additional forces in the rotating frame, we must modify equation (3) appropriately. Since the Sun rotates slowly, the centrifugal force and other effects of second or higher order in Ω can be neglected; this includes also the distortion of the equilibrium structure caused by rotation. Then it is relatively simple to write down the required modification to equations (3) and (20), as a perturbation to the operator $\mathcal{L}$. Using again the stationarity of equation (21) we finally obtain the perturbed frequencies on the form

$$\omega_{nlm} = \omega_{nl0} + m\int_0^R \int_0^\pi K_{nlm}(r,\theta)\Omega(r,\theta) r \, \mathrm{d}r\mathrm{d}\theta \qquad (24)$$

where the kernels K_{nlm} can be calculated from the eigenfunctions for the non-rotating model. It might be noted that the rotational splitting $\omega_{nlm} - \omega_{nl0}$ is an odd function of m.

The general expression for the kernels is quite complicated and will not be presented here. However,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

its basic properties follow the simple discussion given above. In particular, except for modes of low order and low degree, we have approximately that

$$K_{nlm} \propto \rho_0 r |\boldsymbol{\delta r}|^2 \ ;$$

in this limit, the splitting is essentially given by an energy-weighted average of the angular velocity, and hence is dominated by the advection of the mode by the average angular velocity, as in equation (23); the contribution from the Coriolis force is small, in comparison. This also illustrates quite simply how different modes sense the angular velocity: in the radial direction the sensitivity is restricted to the region outside the turning point r_t, whereas in latitude the sensitivity reflects the behavior of the Legendre function P_l^m. It may be shown that P_l^m is essentially confined to the latitude band $\pm \sin^{-1}[(1 - m^2/L^2)^{1/2}]$; thus modes with $m \simeq l$ have substantial amplitude only close to the equator, whereas modes with $m \simeq 0$ extend over all latitudes (see HELIOSEISMIC OBSERVATIONS, figure 2). As already discussed in the introduction, this variation in sensitivity over the range of observed modes permits inversion to infer the angular velocity as a function of r and θ.

Perhaps the most interesting inference concerns the rotation of the solar core. Unfortunately, this is also the most difficult region to probe: only the lowest-degree modes reach the core, and for these the range in m is small, making the observational determination of the rotational splitting difficult. Furthermore, the p modes have relatively small amplitude in the core, reducing their sensitivity to the core rotation, and the effect of rotation is further reduced by geometry, as a result of the small distance from the rotation axis. A similar geometrical reduction in sensitivity holds in the, otherwise dynamically very interesting, region near the polar axis throughout the Sun.

Mode excitation

We have so far discussed only the diagnostic potential of observed frequencies of solar oscillation. However, a condition for observing the modes is that they are excited. Thus we should also consider the possible causes for the oscillations.

'Classical' pulsating stars (see, for example, VARIABLE STARS) owe their pulsations to an intrinsic instability, whereby the star acts as a heat engine, extracting mechanical energy from the radiative or convective flow of energy through the star. In the solar case, such mechanisms appear not to be effective: the modes are intrinsically damped. Instead they are forced stochastically by the vigorous turbulent convection near the top of the convection zone.

Turbulence is a source of acoustic noise, the emission of sound increasing as a high power of the mach number $\mathcal{M} = v_{\rm turb}/c$, where $v_{\rm turb}$ is a typical convective velocity. In models of solar convection, $\mathcal{M}$ is of order unity in the uppermost parts of the convection zone, with intermittent supersonic flows being possible (see SOLAR INTERIOR: CONVECTION THEORY). In particular, very strong flows are associated with the cool downdrafts. Such motion acts as sources of sound waves, which excite the normal modes of the Sun. The resulting mode amplitudes are determined by a balance between the energy input from the forcing and the energy loss through radiative and convective damping. Very roughly, the outcome is that the energy of a single mode of oscillation is approximately equal to the energy of a single convective eddy with a timescale corresponding to the pulsation period. More detailed calculations, based either on simple models of convection or on hydrodynamical simulations, give results that are approximately in agreement with the observed amplitudes.

The stochastic nature of the excitation, and the resulting fluctuations in the mode amplitudes, must be taken into account in fits to determine frequencies and amplitudes from observed time strings of oscillation data. Also, it appears that the localized nature of the source causes asymmetry in the peaks in the power spectra which further complicates the fits. On the other hand, these properties of the observed power obviously carry information about the forcing. The distribution of power is in reasonable agreement with the predictions of stochastic-excitation models. Furthermore, the line asymmetries and properties of high-frequency resonances carry information about the location of the source of the forcing, placing it within the upper few hundred kilometer of the convection zone; this again is not inconsistent with models of the convection zone.

Other aspects

The oscillation frequencies, and other aspects of the observations, carry additional information, some of which is discussed elsewhere. As discussed above, rotation induces a frequency splitting which is an odd function of the azimuthal order m. Departures of solar structure from spherical symmetry gives rise to a frequency splitting which is an even function of m. Such departures might arise, for example, from latitudinal variations in the convective energy transport, or from effects of magnetic fields. More local effects of magnetic fields are observed near sunspots, which absorb and scatter the waves, and in addition are hosts to a separate class of modes (see SUNSPOT OSCILLATIONS AND SEISMOLOGY). Information about local properties of the Sun, including three-dimensional and time-dependent inhomogeneities and flow fields, are becoming available from analyses of the waves that do not assume a modal structure. Also, it has been possible to detect large sunspots on the far side of the Sun; such detections have potentially substantial practical value, in predicting possible solar storms affecting the Earth's environment, once the active region rotates to the near side.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Christensen-Dalsgaard J and Berthomieu G 1991 Theory of solar oscillations *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M Matthews (Space Science Series) (University of Arizona Press) pp 401–78

Christensen-Dalsgaard J, Däppen W, Dziembowski W A and Guzik J A 2000 An introduction to helioseismology *Variable Stars as Essential Astrophysical Tools* ed C İbanoğlu (Dordrecht: Kluwer Academic) pp 59–167

Christensen-Dalsgaard J and Dziembowski W A 2000 Basic aspects of stellar structure and pulsation *Variable Stars as Essential Astrophysical Tools* ed C İbanoğlu (Dordrecht: Kluwer Academic) pp 1–57

Gough D O 1993 Course 7. Linear adiabatic stellar pulsation *Astrophysical Fluid Dynamics, Les Houches Session XLVII* ed J-P Zahn and J Zinn-Justin (Amsterdam: Elsevier) pp 399–560

Gough D O and Toomre J 1991 Seismic observations of the solar interior *Ann. Rev. Astron. Astrophys.* **29** 627–85

Unno W, Osaki Y, Ando H, Saio H and Shibahashi H 1989 *Nonradial Oscillations of Stars* 2nd edn (University of Tokyo Press)

Jørgen Christensen-Dalsgaard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helium (He)

The second lightest element which, after hydrogen, is the second most abundant in the universe. The most abundant isotope of the neutral helium atom consists of a nucleus made up of two protons and two neutrons around which revolve two electrons. The atomic mass is 4, and the atomic number (i.e. charge on the nucleus) is 2. The nucleus of such a helium atom is also known as an alpha particle. Helium nuclei with only one neutron also exist but are more rare.

In terms of number of atoms, nearly 8%, and in terms of mass, about 25%, of all the directly observed matter in the universe is helium. If, as is generally agreed, the heavier elements have been built up from the lighter ones (starting with hydrogen) in a chain of nuclear reactions, the problem is to account for the observed abundance of helium. Helium is being produced from hydrogen by means of nuclear reactions inside stars, but this process is not adequate to account for the amount of helium actually observed. If the Big Bang theory of the origin of the universe is correct, then the observed quantity of helium could have been produced in the first few minutes of the existence of the universe, while the temperature lay in the region of 10^9 kelvins. The observed abundance of helium is one of many factors that favour the Big Bang theory.

See also: cosmic abundance of elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helix Nebula (NGC 7293)

A planetary nebula in the constellation Aquarius, position RA $22^{h}\ 29.6^{m}$, dec. $-20^{\circ}\ 48'$. It is the nearest planetary nebula to Earth, it has the largest angular size, at $15'$ by $12'$, and its magnitude of 6.5 makes it the brightest (though like all extended objects it does not have a high surface brightness). It is illuminated by its hot thirteenth-magnitude central star. The Helix gets its name from its appearance of a spiral viewed from above; an alternative name is the Sunflower Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hell [Höll], Maximilian [Miksa] (1720–92)

Born in Schemnitz, Hungary, became a Jesuit, and worked in Leutschau, Klausenburg and in Vienna where he set up and directed an observatory for Maria Theresa of Austria and Hungary. Somewhat to his own surprise, Hell was invited by Christian VII, king of Denmark, to observe the 1769 transit of Venus from the then Danish island Vardø within the Arctic Circle off the coast of Lapland, all expenses paid. The location was of considerable importance to the accuracy of the ultimate goal of the observations, since Vardø was at an extremity of the distribution of the other observing sites in Manila, Batavia, California, Peking, and Tahiti. J LALANDE was collecting the observations to combine them all into a single accurate value for the distance of the Sun and asked for Hell's data, but Hell reserved their publication until they had been seen by his sponsoring royal patron. LALANDE's suspicions were aroused, and when Hell's observations proved to show inconsistencies with others, he accused Hell of falsifying the data. Rumors circulated that Hell had not even seen the transit and had made up the data while he was in Copenhagen. Hell's scientific reputation was destroyed, although before he died in 1792 he built an observatory in Erlau and worked on the (imaginary) healing properties of magnetism with Franz Mesmer (from whose name the word *mesmerism* is derived). This story forms part of the events of the suppression of the Jesuit Society, in a power struggle in Europe in the eighteenth century between civil, royal and religious forces. In 1835 C L Littrow discovered parts of the original journal that Hell wrote during the event on Vardø, and noted that Hell had indeed altered records of the timing, according to Littrow writing new times in ink of a different color. Hell's low reputation in scientific history was apparently confirmed. In 1883, however, SIMON NEWCOMB visited Vienna, where Hell's manuscripts were kept, and came to a different conclusion. The times had indeed been altered, not after the event but while the observations were being made, correcting smudged ink, and difficulties in writing with a defective pen in the Arctic cold. Nor could Newcomb see changes of color, only a change of shade on drying. Newcomb noted that Littrow's color sensitivity was weak and that he couldn't distinguish the color of Aldebaran. Newcomb fully rehabilitated Hell in a paper called 'On Hell's alleged falsification . . .' in the *Monthly Notices of the Royal Astronomical Society*. There is a crater on the Moon named after Hell; it is not particularly deep.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hellas Planitia

The official name for the Hellas Basin, the largest impact feature on Mars and indeed the largest in the solar system, centered at 44.3 °S, 66.2 °W. The inner part of Hellas Planitia is around 2100 km across and is surrounded by a rim 2 km high. At its deepest it is 9 km below Mars's average surface level—the greatest depth on the planet. The outer part of the basin extends for about 4000 km. The impact that formed Hellas was probably the last big impact event on Mars, and may have created the uplift of the Tharsis Bulge on the opposite side of the planet.

See also: Mars: surface, Tharsis Montes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Helmholtz, Hermann Ludwig Ferdinand von (1821–94)

Born in Potsdam, Germany, had a varied career as an army doctor, professor of anatomy and physiology, and professor of physics at Berlin. At the same time as, but independently of, Lord Kelvin, he studied (1854) the source of energy of the Sun. He suggested that instead of radiating by consuming chemical energy, which would be used up in a lifetime of a few thousand years, the Sun was gradually settling under gravity so that the source of its radiation was gravitational potential energy. There was enough potential energy in the Sun to supply its radiation for several million years. Even this was shorter than the age of the Earth as deduced from geological evidence, and the problem was not resolved until the discovery of solar nuclear energy. However, Helmholtz's conjecture about a source of stellar energy has proved to be applicable at the time when stars first settle from interstellar clouds—the cloud's gravitational potential energy heats the star which it forms and ignites its nuclear reactions. The duration of this settling process is called the *Kelvin–Helmholtz timescale*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Henderson, Thomas (1798–1844)

Born in Dundee, after a career as a lawyer's clerk, appointed as HM Astronomer at the Cape of Good Hope, but soon left because of poor health. Henderson returned to Scotland as the first Scottish Astronomer Royal and, from observations made in South Africa, he was the first person to measure the distance of a star, namely Alpha Centauri, a member of the nearest star system to the Sun, at 4.5 light-years. BESSEL's independent measurements of the star 61 Cygni were, however, published three months earlier.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Henry, Paul (1848–1905) and Henry, Prosper (1849–1903) [known as Frères Henry]

Born in Nancy, the brothers worked together at the Paris Observatory searching for minor planets. Decided to exploit photography to facilitate the search and thus built telescopes at home for the purpose. With the encouragement of the then director of the observatory (Admiral Mouchez), built (1885) a 34 cm photographic refractor; this work was the foundation of the *Carte du Ciel* project to photograph the sky in its entirety. The Frères Henry made half the telescopes used in the international project, led by France. The project took the attention of French astronomy for decades, at a time when astrophysics was developing in America, Britain and Germany. The two brothers, and their lifetime of work together, are commemorated by the single name, Frères Henry, of a close pair of craters on the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Heracleitus of Ephesus (*c*. 540–*c*. 475 BC)

Greek philosopher, born in Ephesus, thought of the universe as a place of ceaseless change ('becoming'), going so far as to suggest that the Sun was created anew each day. He believed the fundamental element was fire, because it was always changing, and that the Sun and Moon were bowls of fire. So too were stars, at great distances. Eclipses occurred when the bottom of the bowl was turned towards us.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Heraclides of Pontus (*c*. 388–315 BC)

Greek philosopher, born in Heraclea, Pontus, Heraclides was taught by PLATO and ARISTOTLE, and the Pythagoreans in Athens. Although his written works are lost, it is said that he considered the planets to be individual worlds. He was the first philosopher to suggest that the rotation of the Earth would account for the apparent rotation of the stars, even distinguishing between the solar day of rotation and the sidereal day.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herbig, George Howard (1920–)

American astrophysicist, Herbig was educated and worked in California, mostly at the Lick Observatory, and Hawaii. He is known for his spectroscopic studies of recently formed stars and the interstellar medium, from which he studied the still mysterious diffuse interstellar lines. He found and investigated many Hα emission stars, T Tauri stars, and peculiar stars, all of them different stages of the early history of stars. He and GUILLERMO HARO independently discovered the eponymous *Herbig–Haro objects*, gas clouds now thought to be the result of shocks in the interstellar medium caused by ejected material from unstable young stars. Herbig showed that the lithium abundance is correlated with age in young stars, and he investigated the rotation speeds of stars of different spectral class. From his data it is hypothesized that the slow rotation of the more numerous, less massive stars is associated with the almost universal existence of planetary systems around them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herbig–Haro Objects and Exciting Stars

Herbig–Haro (HH) objects correspond to high-velocity material ejected from young stars, which interacts with the environmental gas and with previous ejections, forming a complex shock wave structure. The shock waves heat and ionize the outflowing and the environmental material, which then emits a rich emission line spectrum that is observed over a wide wavelength range.

The sources of HH objects

The stars that eject HH objects in some cases are optically detected, T TAURI STARS and in other cases are infrared sources (not detected optically). Some of the sources are not directly detected even in the infrared, as the infrared emission seen in the source region is spatially offset from a radio emission peak, which is believed to indicate the true position of the source. In such cases, the infrared emission observed close to (but not coinciding with) the RADIO SOURCE is believed to correspond to stellar light which is being scattered by dusty gas which surrounds the star.

It is tempting to think about a possible evolutionary sequence, with the more embedded, optically invisible objects corresponding to earlier stages in the evolution of the young star plus outflow system, and the optically visible T Tauri stars to later evolutionary stages (see also STELLAR EVOLUTION). However, whether or not the central star is visible also depends on the nature of the surrounding environment, and possibly also on the orientation of the system with respect to the plane of the sky.

The production of the jet

One of the main uncertainties in our current picture of HH objects and their sources is the mechanism by which the source produces the observed high-velocity ejections. Figure 1 shows a schematic diagram of the elements which are believed to be present. In the center of the system, there is a star, which is possibly surrounded by a magnetosphere (similar to the MAGNETOSPHERE OF EARTH) extending out to a few times the radius of the star (central circle of the diagram). Around the star there is an orbiting ACCRETION DISK, the material of which slowly drifts in towards the star, eventually becoming incorporated into the star (through a quite complex flow involving the interaction of the stellar magnetosphere with the inner edge of the accretion disk). The star + accretion disk system is surrounded by a flattened, infalling envelope of material belonging to the region of the interstellar medium which is collapsing to form the central star. The material of this infalling envelope is first incorporated into the accretion disk, and then into the central star.

The high velocity ejections (which we see as HH objects) could come directly from the surface of the star (as a stellar wind), but it has still not been possible to come up with an explanation of how to produce a stellar wind with appropriate parameters for explaining the observed HH objects. The currently more favored possibility is that the ejected material might actually come from the inner region of the accretion disk or from the region of interaction between the disk and the stellar magnetosphere.

The high-velocity wind escapes through two oppositely directed channels in the surrounding envelope, forming what is called a bipolar outflow (composed of two jets traveling away from the star in opposite directions). It is still unclear whether these channels are ploughed up by the passage of the jets or whether they are directly produced by the process of gravitational collapse which has formed the star (see also ASTROPHYSICAL JETS).

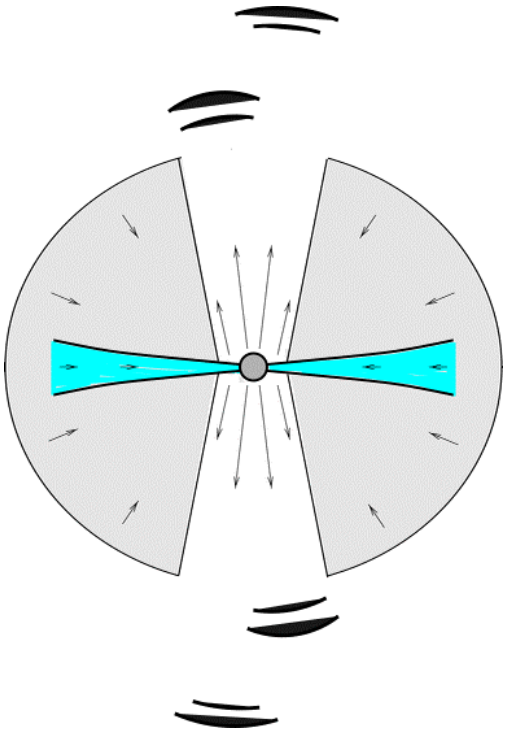


Figure 1. Schematic diagram of a young star–HH object system. The HH objects are the pairs of black arcs at the top and bottom of the diagram. The star + magnetosphere system (central circle) is surrounded by an accretion disk, which is fed by the material from a surrounding, infalling envelope. The star and/or the inner region of the accretion disk produce a wind, which escapes from the system through two oppositely directed 'channels' drilled through the infalling envelope. This collimated wind interacts with previously ejected material and with the surrounding environment, producing a system of shock waves, which we observe as HH objects.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The observational characteristics of HH objects

The region of interaction between the collimated winds ejected from young stars and the surrounding environment (or with previous ejecta) has a number of shock waves that produce an emission line and continuum spectrum. This spectrum has clear differences with the spectra of photoionized regions (showing strong emission lines of neutrals as well as of ionized species) and has been used to identify the class of objects known as HH objects.

HH objects have now been detected over a wide wavelength range, ranging from radio (centimeter) wavelengths to the ultraviolet. The observed continuum emission includes the free–free, two-photon and recombination continua of hydrogen. Emission lines of neutral, singly and doubly ionized species of many of the most abundant elements have also been detected. Ultraviolet and infrared molecular hydrogen emission lines (which are clearly associated with the HH objects) have also been detected. In 2001, observations with the Advanced CCD Imaging Spectrometer on NASA's Chandra X–ray Observatory have revealed x–ray emissions from HH 2, one of the brightest and closest Herbig Haro objects in the Orion Nebula. Previous instruments lacked the resolution and sensitivity necessary to observe these x–rays.

Traditionally, HH objects have been divided into low- and high-excitation objects, depending on the range of ionization states implied by their emission line spectra. High-excitation objects show strong lines of, e.g., singly and doubly ionized oxygen, while the emission of low-excitation objects is dominated by lines of neutral oxygen, neutral and singly ionized carbon, etc. Both the emission line and the continuum spectra have been modeled with reasonable success in terms of the excitation produced by the passage of shock waves with velocities ranging from 20 to 200 km s^{-1} (with the lower velocities corresponding to the low-excitation objects and the higher velocities to the high-excitation objects).

Some of the most interesting characteristics of HH objects are their kinematical properties. The emission line profiles have widths indicating line-of-sight velocities of up to 500 km s^{-1} (in the high-excitation objects). Given the magnitude of these velocities and the fact that most of the detected HH objects are clearly within 1 kpc of the Sun, it is easy to detect the PROPER MOTIONS of these objects.

The proper motions of approximately 10% of all of the known HH objects have now been measured. The first measurements were done with photographic plates (about 20 yr ago), using sequences of plates obtained over a time baseline of decades. In the last few years, proper motions have been measured using higher angular resolution CCD frames, obtained over a period of only a few years. In the recent images obtained with adaptive optics (in ground based telescopes) or with the HST, the proper motion of HH objects can already be detected using images obtained over a period of only a few months. This recent increase in the attainable angular resolution will produce a detailed picture of the time evolution of HH objects. During 2000–2001, spatially resolved spectroscopic observations of a few HH jets were carried out using the Hubble Space Telescope. These observations have unprecedented angular resolution and show details of the jet collimation region and of the shock structures present along the jets.

Even though HH objects have spectra that are not so diverse (allowing a clear identification of the objects) they do show a wide range of quite different morphologies. Many objects show groups of compact emission knots, located in two diametrically opposed regions (or 'lobes') with respect to the central source. The two lobes show proper motions directed away from the source, one of them showing redshifted radial velocities, and the other one being blueshifted.

In other objects, only one of the two lobes is observed. In most of these cases, the observed lobe is blueshifted, indicating that the invisible lobe is the one directed away from the observer. This is probably a result of the fact that the lobe directed away from the observer is hidden by the dense, high-extinction region surrounding the central star.

Many HH objects show only 1–3 knots in each lobe, while others show larger numbers of emission line knots. A few objects show chains of well-aligned knots, extending radially away from the central star. Even though only few good examples of such 'HH jets' are found (approximately 10% of the known HH objects), they have received a lot of attention from both observers and theoreticians, as their organized structures appear to be more straightforward to interpret than the more chaotic structures of other HH objects.

The interpretation of the jet-like HH objects

HH jets have very high Mach numbers (their velocities are 10–100 times higher than their sound speed), and interact with the surrounding environment forming a number of shocks. The leading edge (or 'head') of the jet produces a bow shock, similar to the shock formed around the nose of a supersonic airplane or a bullet.

Actually, in the case of the head of a jet, two shocks are produced: the bow shock, which pushes aside the surrounding environment to make way for the jet material, and the jet shock (also called the 'Mach disk'), which is induced on the jet by the surrounding material. This pair of shocks is called a 'working surface'.

Working surfaces can also be formed in the body of the outflow, and not only at the leading edge. If the velocity with which the material is ejected from the central source varies with time, faster material (ejected at later times) catches up with slower material ejected

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

earlier. This catching up process leads to the formation of two-shock working surfaces similar to the ones formed at the leading edge of the jet.

In HH objects, we observe a series of aligned knots. An example of this kind of structure is shown in figure 2, in which an image of the HH 34 object is shown. This image has been taken with the HST in the light of the red lines of singly ionized sulfur.

From this figure, we see that the successive knots are quite different from each other. Some of them are small, appearing just barely resolved in the HST image, and some of them are quite large, showing a clear, bow-like shape. In larger field of view images of this outflow (obtained with other telescopes), a number of knots at larger distances from the source have also been detected.

The knots along the HH 34 jet are observed to move away from the source at velocities of about 300 km s^{-1}. This motion is observed as a wavelength shift of the emission lines of the object (as a result of the DOPPLER EFFECT), and also directly as a displacement of the positions of the knots in the plane of the sky.

It is interesting that the structure of HH 34 can be reproduced with a numerical solution to the gas-dynamic equations (which describe the motion of a diffuse gas moving at supersonic velocities), if one allows oneself to choose an arbitrary time variability for the velocity with which the material is ejected from the source. For example, one can choose a variability with several sinusoidal modes. Modes with large amplitudes and long periods produce well-separated, large, bright working surfaces traveling down the outflow. Modes with low amplitudes and short periods produce little chains of low-intensity knots.

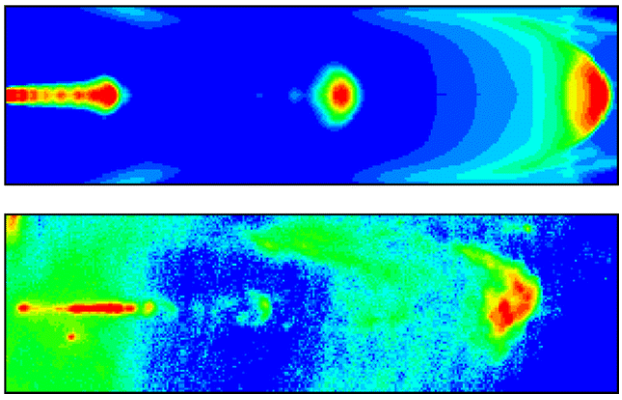


Figure 2. Bottom: HST image of the HH 34 jet, obtained through a filter which isolates the red forbidden lines of singly ionized sulfur ([S II] 6717+6731). The source of the jet is the compact knot close to the left edge of the image. Top: the corresponding intensity map predicted from a numerical model of a jet with a time-dependent ejection velocity with three sinusoidal modes of different periods and amplitudes.

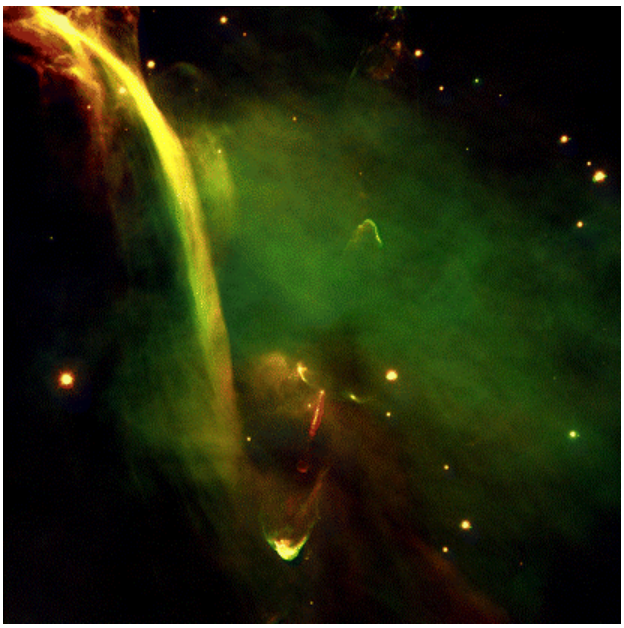


Figure 3. Composite image of HH 34 obtained with the FORSZ instrument of the European Southern Observatory's Very Large Telescope.

Picking three modes of appropriate periods and amplitudes, one can compute models with knots that have morphologies and kinematical properties that strongly resemble the knots along the HH 34 jet. An image (in the red lines of singly ionized sulfur) predicted from such a model is shown in figure 2, from which one can see that structures that resemble HH 34 can indeed be obtained from a model of a jet with a variable ejection velocity.

A recent image of HH 34 is shown in figure 3.

Molecular outflows

In many systems of HH objects, one also observes emission in molecular lines (for example, in lines of CO and NH_3 in radio and millimeter wavelengths), coming from a region that surrounds the HH objects. The molecular emission is also distributed in two lobes, one blue- and one redshifted, with velocities that in general are lower than the ones measured for the associated HH objects. It is possible that this molecular emission corresponds to material from the molecular cloud out of which the young stars are forming, which is being pushed by the stellar outflows. It has also been suggested that the molecular outflows could correspond to a slower, molecular wind, ejected from the outer regions of the accretion disk around the central star.

Recent, high angular resolution observations of molecular outflows obtained with millimeter interferometers show that the situation is really more complex. Some outflows are found to have a broad, low-velocity emission component (which probably does correspond to environmental gas which is being pushed

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by the HH flow) but also show a high-velocity, central component.

These high-velocity components of molecular outflows have dynamical properties and morphologies which are very similar to the ones of HH jets. It is possible that this molecular emission traces small quantities of environmental gas which has been entrained into the HH jet ejected by the star. Also, the molecular emission might be indicating that the gas associated with the HH ejection is at least partially molecular.

At the present time, the relation between HH objects and the molecular outflows is not clear. The ongoing debate on this subject will have interesting implications for our ideas about the formation and propagation of outflows from young stars.

Bibliography

It is straightforward to find the up-to-date literature on HH objects, as there is an on-line catalog of HH objects ('A general catalogue of Herbig–Haro objects', by B Reipurth), which can be downloaded from the web address:
http://casa.colorado.edu/hhcat/
This catalog gives an extensive list of the literature on the observations of all of the known HH objects. A more or less complete compilation of the literature on the theory of HH objects is given in the paper:

Reipurth and Raga 1999 *The Origin of Stars and Planetary Systems* ed C Lada and N D Kylafis (Dordrecht: Reidel) p 267

Also, there is a monthly 'Star formation newsletter' (edited by B Reipurth) which gives the abstracts of most of the recent papers in the fields of star formation and outflows from young stars. This newsletter is available at the web address:
http://casa.colorado.edu/reipurth/

Alejandro C Raga

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hercules

(abbrev. Her, gen. Herculis; area 1225 sq. deg.) A northern constellation which lies between Draco and Ophiuchus, and culminates at midnight in mid-June. The origin of the constellation figure is uncertain, though its name dates back to ancient Greece, where it was identified with the strong man and hero of Greek mythology. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A large but rather inconspicuous constellation, the brightest stars in Hercules are α Herculis (Rasalgethi), a multiple system, consisting of a red (M5) primary, which is a semi-regular variable, magnitude range 2.7–4.0, and a yellow (G5) secondary, magnitude 5.4, separation 4.8″, both of which have a much fainter companion, β Herculis (Kornephoros), magnitude 2.8, which also has an unseen companion, and ζ Herculis, a close binary with yellow (F9 and G7) components, magnitudes 3.0 and 5.7, separation 1.55″, period 34.5 years. There are 12 other stars brighter than fourth magnitude. Other interesting stars include the Mira-type variable S Herculis (range 6.4–13.8, period about 307 days) and the eclipsing binary DQ Herculis (Nova Herculis 1934), one of the brightest novae of the twentieth century, which attainted about magnitude 1.4 before fading to magnitude 14–15.

Other interesting objects include the Great Cluster in Hercules (M13, NGC 6205), the brightest globular cluster in the northern sky, which lies between η and ζ Herculis and, at magnitude 5.9, is just visible to the unaided eye as a faint misty patch, M92 (NGC 6341), another fine globular cluster, magnitude 6.5, and NGC 6210, a ninth-magnitude planetary nebula.

Also in Hercules are Hercules X-1, an x-ray binary, and Hercules A, a powerful extragalactic radio source.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hercules X-1

Hercules X-1 (Her X-1) is a pulsating x-ray source detected with the first x-ray satellite, UHURU, in 1971. It is an accreting X-RAY BINARY system consisting of a NEUTRON STAR that rotates every 1.24 s while orbiting a 'normal' stellar companion every 1.7 days; the misalignment of the magnetic axis relative to the rotation axis results in a distinctive pulse profile in x-rays. The 1.7 day orbital motion shows up clearly in Doppler shifts of the 1.24 s pulse; the x-ray source shows distinct eclipses every 1.7 days; and a previously known optical variable star HZ Herculis (HZ Her), colocated with the x-ray source, is found to have an identical period. The fact that we see total eclipses implies that our line-of-sight is within a few degrees of the orbital plane. The x-ray source shows sharp eclipses, attesting to its compact size. Variation in optical light by 1.8 magnitudes over the 1.7 day orbital period is attributed to x-ray heating on the side of the companion facing the active neutron star.

In addition to the 1.24 s rotation period and the 1.7 day orbital period, the system has an approximately 35 day 'superorbital' period in which the x-ray emission displays so-called main and short 'on' states separated by relative x-ray quiescence. The (0.2–100 keV) x-ray fluxes are $\sim 2 \times 10^{-9}$ ergs cm^{-2} s^{-1} during main-on state which lasts from 9–12 days; $\sim 6 \times 10^{-10}$ ergs cm^{-2} s^{-1} during the short-on state (5–7 days); and $\sim 1 \times 10^{-10}$ ergs cm^{-2} s^{-1} during the low or quiescent state. However, the 1.7 day optical and UV flux variations attributed to x-ray heating of the companion star continue throughout the 35 days with the optical light curves showing less than 20% change in peak-to-peak amplitude over the 35 day cycle. The optical light curves have a power spectrum that shows power at the sum of the orbital and long-term frequencies, but not in the difference, implying that the long-term variations are due to counter-precession of a tilted, co-rotating ACCRETION DISK. However, the underlying physics that causes the disk to start precessing has not as yet been determined. Her X-1 is one of only a half-dozen known x-ray binaries that show a superorbital period.

Due to its brightness over a broad range of energies and its high inclination, with eclipses of the neutron star, Her X-1 provides an excellent opportunity to study accretion processes onto a compact object which have come to be recognized as an important energy release mechanism in the universe. In addition it allows us to study the characteristics of matter and radiation under extreme conditions: the intense gravitational and magnetic fields of the compact object.

Temporal variations

The 1.24 s pulsations have also been detected in both optical and UV continua. The rms pulsed fraction is of order 0.5% in the UV, two to four times that of the reported optical pulsed fraction. The UV and optical pulses are in phase with the hard x-rays during orbital phases when x-ray dips occur. The lack of UV/x-ray delay indicates that the dips and the UV pulses may originate in the same material near the line of sight.

Absorption dips that last nearly as long as the x-ray eclipse regularly occur before the eclipse (pre-eclipse dips) and occasionally at orbital phase 0.3–0.6 (anomalous dips). The pre-eclipse dips exhibit the characteristic of starting out with the main-on state merged with the first eclipse ingress and progressing toward an earlier orbital phase as the main-on progresses: This pattern has persisted since the discovery of Her X-1 in 1971. The recurrence interval (1.65 days) is somewhat longer than the beat period between orbital and 35 day cycles. Since the pattern of the dips does not drift significantly in orbital phase over many 35 day cycles, it must be tied to the 35 day 'clock'.

The 35 day superorbital period is not coherent; it can vary in length from 19.5 to 21.5 orbital cycles with the most favored length being 20.5 cycles (34.85 days). The turn-on of the main-on state (which defines phase 0.0 of the 35 day cycle) occurs preferentially at orbital phases 0.2 and 0.7 (where orbital phase 0.0 refers to the center of x-ray eclipse). The existence of just these two types of turn-on causes phase jumps between successive 35 day cycles to be integral multiples of 0.5 orbital cycles.

The system has undergone several 'anomalous low' states during which it shows a substantial, unexpected drop in x-ray flux, with no substantial change in absorbing column density; little or no change in UV and optical fluxes; and a noticeable increase in pulse period. With the exception of the increase in pulse period and the degree of x-ray flux reduction these characteristics are similar to those observed during 'normal' 35 day low states.

Study of archival optical data of the variable star HZ Her shows that the system also displays intervals (lasting from months to years) when the 1.7 day optical variations attributed to x-ray heating disappear—which implies that accretion does turn off. A small amplitude modulation ($\sim$0.3 magnitudes) with twice the orbital frequency is observed at this time and is interpreted as the changing light from the distorted companion.

Spectral behavior

The normal star companion fills its ROCHE LOBE and the x-rays are powered by matter spilling over the inner LAGRANGIAN POINT and transferred through an accretion disk onto the neutron star surface. Her X-1 accretes matter at a rate of a few times 10^{-9} solar masses ($M_\odot$) per year. The energy release due to accretion is given by

$$E_{\rm acc} = G M_{\rm NS} \dot{M} / R_{\rm NS} \qquad (1)$$

where $M_{\rm NS}$ and $R_{\rm NS}$ are the mass and radius of the accreting object. The x-ray luminosity may be related to the mass accretion rate, $\dot{M}$, via

$$L_x = f \frac{G M \dot{M}}{R} \qquad (2)$$

where f, of order unity, takes account of possible anisotropy of the radiation. For Her X-1, for a distance

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of 5 kpc, L_x is about 2×10^{37} erg s $^{-1}$ during the main-on state.

At maximum light during the 'normal' main-on state Her X-1/HZ Her shows a nearly constant energy flux per logarithmic frequency interval over a frequency range of 10^{15}–10^{19} Hz. A reduction in x-ray flux during the anomalous low state is not associated with an increase in flux at some other wavelength.

Figure 1 shows the x-ray spectrum of the main-on state obtained with BeppoSax. The pulse-averaged spectrum is a flat power law with a high-energy cut-off and an excess strong thermal blackbody component at low energies (a characteristic of many binary x-ray pulsars). Its unusually high galactic latitude and correspondingly low interstellar absorption ($N_{\rm H} \sim 10^{20}$ cm^{-2}) make Her X-1/HZ Her especially well suited for the study of this soft component. Spectral features around 1.0 and 6.4 keV are attributed to Fe L- and K-shell emission; a feature between 35 and 58 keV is attributed to electron–cyclotron resonance which implies a surface magnetic field of a few $\times 10^{12}$ G. The spectrum changes as a function of pulse phase: the Fe line and thermal emission are strongest during the off-peak of the pulse. The pulse profile changes as a function of energy: the pulse peak shifts and shows less structure below 1 keV. The pulse phaseshift suggests that the soft x-rays are likely reprocessed hard x-rays and the Fe-line emission originates at the reprocessing site.

The ultraviolet spectra of Her X-1 show strong emission lines due to N V, C IV, and He II (figure 2). The Goddard High-Resolution Spectrograph on the Hubble Space Telescope (HST) resolved two moving emission components of the N V emission doublet: a broad emission line that can be attributed to the x-ray illuminated surface of a Keplerian accretion disk and a narrow line associated with the x-ray heated atmosphere of the companion star, HZ Her. The ultraviolet continuum is flatter than normal stellar spectra, presumably due to x-ray heating effects.

Optical spectra of HZ Her show Balmer absorption lines down to the series limit (figure 3). Absorption lines are seen in the spectrum both at minimum and maximum light. Emission lines at He II λ4686 and N III/C III around λ4400 are seen during maximum; those due to He II persist during the minimum. At minimum light the optical spectrum of HZ Her corresponds to a spectral type of A7 ($T_{\rm eff} = 8100$ K). At maximum light the spectral type ranges from B3 to B6 depending on the lines used. The dramatic change in spectral type from minimum to maximum is also attributed to x-ray heating of the side facing Her X-1; the lack of a good match to a single spectral type is attributed to x-ray heating effects. The optical flux is mainly from the x-ray heated star surface with the disk surface contributing less than 15% of the total optical flux.

Pulse and orbital period history

Overall the pulse period of Her X-1 shows a secular decrease or spin-up as is expected for a system accreting mass (and hence angular momentum). However, during anomalous low states the system shows a noticeable spin-down. Correlation between times of low hard x-ray luminosity and episodes of spin-down is predicted by models which suggest that a *decrease* in mass accretion rate causes the Alfvén radius at which material joins the magnetosphere to move out, thereby *increasing* the drag-producing torques of the field lines threading the disk outside the co-rotation radius, leading to spin-down.

The orbital period has been shown to be decreasing at a rate of 2×10^{-8} days yr^{-1}. If all the mass lost by the companion is accreted by the neutron star the mass transfer rate needed to explain the decrease in orbital period is $1.4 \times 10^{-8} M_\odot$ yr^{-1}, nearly an order of magnitude greater than the mass loss rate required to provide the observed x-ray luminosity. Mass loss through a simple self-excited wind also cannot explain the observed period decrease. Braking of the companion star by a magnetically-channelled stellar wind provides a mechanism for the angular momentum loss rate. UV spectra do not show the presence of either P Cygni profiles or modest line shifts indicative of wind flows; however, it is possible that the high x-ray luminosity ionizes the material in the wind flows and thereby precludes spectroscopic detection of a wind.

Open questions

Evolutionary status

X-ray binaries containing neutron stars are classified according to the mass of the 'normal' stellar companion. The high-mass systems (HMXBs) contain massive early-type stars with short lifetimes ($<10^7$ yr). More than half of HMXBs exhibit regular x-ray pulsations (implying magnetic fields of the of order 10^{12} G) with periods from 0.1 to 1000 s. The x-ray emission is caused by matter accreted from the stellar wind of the massive companion. The optical fluxes, due mainly to emission from the massive companion, are high, giving x-ray to optical luminosity ratios in the range 10^{-5} to 10. The low-mass x-ray binaries (LMXBs) contain faint optical companions with long lifetimes of 10^8–10^9 years. The x-ray emission is caused by matter overflowing from the Roche lobe of the companion. The optical light is due predominantly to x-ray heating of the accretion disk, giving x-ray to optical luminosity ratios from 10 to 10^4. With few exceptions, strong pulsations are not seen, implying that LMXBs have relatively weak magnetic fields ($\leq 10^{10}$ G). The Her X-1 system is unusual in that, with an approximately $2M_\odot$ companion, it is the only known 'intermediate'-mass system. It is like the HMXBs in that it exhibits pulses and its optical light is dominated by emission from the companion, but it accretes matter via Roche lobe overflow through an accretion disk like the LMXBs. The location of Her X-1, more than 3 kpc above the Galactic plane, suggests that it is a rather old system like the LMXBs, or that, if it originated in the Galactic plane, the neutron star must have obtained a significant kick from the SUPERNOVA that formed it. The pulses and the cyclotron line feature which attest to the high magnetic field of Her X-1 are difficult to reconcile with theories of STELLAR EVOLUTION which require that the magnetic fields of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

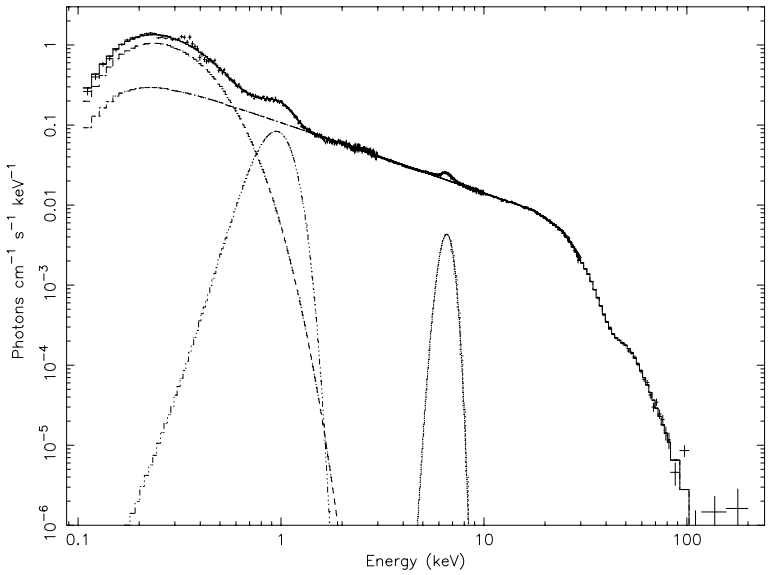


Figure 1. x-ray photon spectrum of Her X-1 taken by BeppoSax with spectral components (blackbody, power-law, Fe L line and Fe K line) indicated by broken curves. Averaged over orbital phases, excluding eclipses and dips.

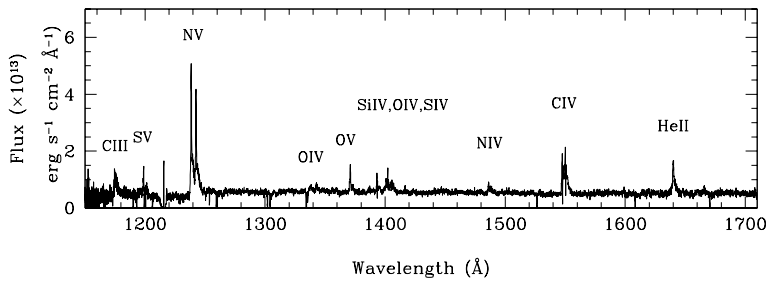


Figure 2. Ultraviolet spectrum of Her X-1 taken with HST/STIS averaged over orbital phases 0.69–0.76.

neutron stars decay over time. The large orbital decay rate is also difficult to reconcile with the relatively low mass accretion rate and lack of evidence for mass loss through a wind.

System mass

The Her X-1 system is considered one of about seven (out of several hundred known) x-ray binaries for which reliable estimates of the component masses exist. However, both methods used to determine the component masses of the system rely on assumptions that have not been independently verified. The mass determinations made utilizing x-ray and optical pulse timing assume that the companion star fills its Roche lobe and that the optical pulses are reprocessed at the inner Lagrangian point of HZ Her. RADIAL VELOCITY measurements are hampered by the uneven heating of the star (the center of light is not the center of mass) and the presence of substantial amounts of gas in the system. Lower limits to the mass of the neutron star ($M_x \geq 1.3 \pm 0.1 M_\odot$) obtained by radial velocity measurements are consistent with the pulse-timing determination ($M_x = 1.3 \pm 0.1 M_\odot$).

The superorbital period

The interpretation of superorbital periods in terms of disk precession has been questioned, because if the disk precession is controlled by the gravitational fields of the neutron and companion stars any induced precession would be damped rapidly because of differential precession within the disk. The preferential turn-ons at orbital phases 0.2 and 0.7 have been attributed to the interaction of a tilted accretion disk with the time variation of the torque induced by the companion star leading to a wobble or 'nodding' motion in the disk. Again the physics that causes the disk to tilt in the first place is not known. Several models have been put forward that attempt to tilt and precess the disk without invoking precession of the neutron star and/or its companion. One viable suggestion is that the behavior attributed to precession is maintained by the influence of the x-ray emission on the structure of the disk; that strong central illumination can maintain disk warping has been shown analytically. Another possibility

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

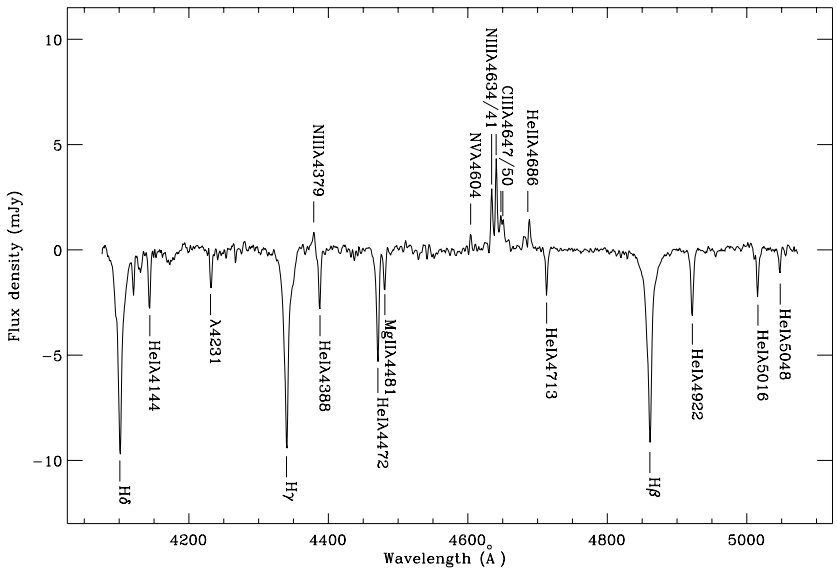


Figure 3. Optical spectrum of HZ Her obtained with the Isaac Newton Telescope on La Palma at orbital phases 0.45–0.55.

is that shadowing of the companion star may lead to transfer of matter with misaligned angular momentum.

Table 1. Properties of Her X-1.

Apparent magnitude m_V	14.7–13.0
Orbital epoch T^0_{orb}	JD2443805.019980(14)
Orbital period, P	1.700 167 720(10)d
Pulse period	1.237 759 653(3)s
$\dot{P}_{orb}$	$-2.0(0.4) \times 10^{-8}$ days yr^{-1}
NS* mass	$1.3(0.1)M_\odot$
CS mass	$2.2(0.1)M_\odot$
CS spectral type	A7-B3/6
Distance, D	5–7 kpc
Inclination of orbit to LOS, i	$87(2)^\circ$
Orbital eccentricity, e	≤ 0.0003
$a_X \sin i$	13 light seconds
x-ray eclipse duration	0.2320(3) days

* NS stands for neutron star; CS stands for companion star. LOS for line-of-sight.

Emission lines

The optical N III and C III emission lines at λ4640 and λ4650 disappear during mid-eclipse as do the O III λ3133 lines in the UV, suggesting that they can be attributed to Bowen fluorescence from the x-ray heated side of the companion star. However, the origin of the high excitation emission line of He II (λ4686) is not clear: the line persists at orbital phase 0.0 when neither the heated side of HZ Her nor the accretion disk are visible. This suggests either an accretion disk corona or an extended source of hot gas in the system.

Bibliography

Over 300 papers deal with direct observations or interpretations of the Hercules X-1/HZ Herculis system. The first discovery paper:

Tananbaum H, Gursky H, Kellogg E M, Levinson R, Schrier E and Giacconi R 1972 Discovery of a periodic pulsating binary x-ray source in Hercules from UHURU *Astrophys. J. Lett.* **174** L143

A follow-up paper:

Giaconni R, Gursky H, Kellogg E, Levinson R, Schreier E and Tananbaum H 1973 Further x-ray observations of Hercules X-1 from UHURU *Astrophys. J.* **184** 227

captured the most prominent behavior and gave the initial interpretation in terms of an accretion-powered, pulsating binary x-ray source.

The twisted, tilted accretion disk was first suggested as an explanation of the 35-d variations in:

Katz 1973 *Nature Phys. Sci.* **246** 87

A comprehensive summary of research on x-ray binaries has been compiled by:

Lewin W H G, van Paradijs J and van den Heuvel E P J 1995 *X-Ray Binaries* (Cambridge: Cambridge University Press)

Saeqa Dil Vrtilek

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hermes

An Apollo asteroid discovered by Karl Reinmuth in 1937, provisionally designated 1937 UB. Hermes is the only asteroid to have been given a name without first having received a permanent number and having had its orbit accurately calculated. On 30 October 1937 it passed Earth at a distance of 733 000 km, a close-approach record that stood for over 50 years. The parameters of its orbit are approximate only: mean distance from the Sun 1.64 AU (245 million km), perihelion 0.62 AU, aphelion 2.66 AU, period 2.11 years, inclination 16°, eccentricity 0.62. It has never been seen again, but with an estimated diameter of 900 m and a potential closest-approach distance of 0.003 AU (450 000 km) it remains near the top of the list of potentially hazardous asteroids.

See also: near-Earth asteroids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herrick, Edward Claudius (1811–62)

Bookseller and librarian, born in New Haven, CT, self-educated by association with Yale University academics, and observer of the 9 August 1837 meteor shower, which he identified as the Perseids. With the Leonids in mind (after a brilliant display in November 1833 they proved to be an annual shower, the first identified), Herrick found other August records of meteor showers. In 1839 he confirmed the radiant of the shower as being in Perseus. As a result of his library research, he also identified a third annual meteor shower around 30 April (now called the Lyrids) and, later, a fourth annual shower, in December, the Andromedids (or Bielids, named for their association with Biela's Comet). Herrick had in fact been preceded in the identification of the Perseids as an annual meteor shower by ADOLPHE QUETELET. The two of them had been preceded by anonymous English and German farmers who for generations had referred to the Perseids as the Tears of St Lawrence (martyred on 10 August 258). Feted locally for his discoveries, Herrick became the Yale University librarian, where he wrote obituaries of graduates and faculty, including his own, which he completed a few days before he died in 1862.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herrick, Samuel (1911–74)

Professor of astronomy and engineering at UCLA, founder of the field of astrodynamics, which applied the disciplines of celestial mechanics and mathematics to the problems of space flight.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herschel Family

William Herschel (1738–1822), his sister Caroline (1750–1848) and his son John Frederick William (1792–1871) freed astronomy from its preoccupation with the solar system and helped to make the study of the stars, the nebulae and the cosmos as a whole part of mainstream science.

William and Caroline Herschel

William was the third surviving child, and Caroline the fifth, of a humble bandsman in the Hanoverian Guards. William was given as good an education as the family could afford, after which in 1753 he joined his father in the military band, but Caroline's illiterate mother prevented her from learning more than reading and writing.

In 1757, after the defeat of the Hanoverian army by the French, William fled to England, where he scratched a living as a musician, first in the capital and then in the north. His fortunes took a turn for the better in 1766, when he was appointed organist to a fashionable chapel in the spa resort of Bath. Despite his busy life as performer, conductor and teacher of music, William's financial security now allowed him to develop other interests, including astronomy.

In 1772 William visited the family home in Hanover and was dismayed to find Caroline a household drudge. Their father was now dead, and William persuaded his mother and eldest brother to release Caroline so that she could join him in Bath, to pursue a career as a singer. Before long, however, Caroline found William devoting every leisure minute to astronomy. Being self-taught, he did not know that astronomers studied the nature and movements of the familiar bodies of the solar system; instead, from the start he set himself the goal of understanding 'the construction of the heavens'. He realized that to see distant and therefore faint objects he would need telescopes as large as possible, but his funds were limited, and before long he resorted to making his own reflectors with the help of some second-hand tools and Robert Smith's classic treatise on *Opticks*. These early experiments were to culminate in 1781 in two unsuccessful attempts to cast a 3 ft diameter mirror in a foundry set up in the basement of their home. He did, however, successfully build a Newtonian REFLECTOR of 20 ft focal length with 12 in mirrors (see NEWTONIAN TELESCOPE, REFLECTING TELESCOPES); but it was simply slung from a pole and used by an observer perched precariously on a ladder. In 1778 he succeeded in polishing a mirror of 7 ft focal length to a shape of high precision, and soon he was using this systematically to familiarize himself with the night sky.

In 1774 William had opened his first observing book, and on the first page he sketched the ORION NEBULA, remarking that it had changed shape since the drawing reproduced in the book by Smith. In prophetic words he commented that 'perhaps from a careful observation of this Spot something might be concluded concerning the Nature of it'. But he realized that first he must familiarize himself with the brighter stars. While thus engaged he began to collect stars that were double, hoping that they might prove of use in the detection of annual parallax, for GALILEO had suggested long ago that the more distant star of a pair might be used as a quasi-fixed reference point against which to measure the annual movements of the nearer star. On 13 March 1781, while doing this, William came to examine a 'star' in Taurus, and saw at a glance that there was something unusual about its appearance. Re-examining it four days later, he found that it had moved, and was therefore no star but a member of the solar system. Speculations about undiscovered planets were in fact commonplace at the time, but this was unknown to William, who therefore took the object to be a comet.

By this time word of this extraordinary musician had spread in scientific circles, and William's home had been visited by a number of astronomers, including NEVIL MASKELYNE, the Astronomer Royal. William did not know how properly to describe the position of the supposed comet, and Maskelyne's Greenwich instruments were inferior to William's 7 ft reflector and showed nothing unusual about any object in the region he indicated. Eventually Maskelyne identified the object by its motion. It proved to be not a comet, but a primary planet of the solar system (later named Uranus), the first to be discovered since the dawn of history.

William became famous overnight, and his allies (notably Sir Joseph Banks, President of the Royal Society) lobbied King George III—himself a Hanoverian—to make available funds to enable William to give up music and devote himself to astronomy. William was invited to take his 7 ft reflector to Greenwich to be compared with Maskelyne's instruments, and then to the court in London and afterwards to Windsor Castle. Suitably impressed, King George awarded William a modest but adequate pension, if he would settle near Windsor and show the heavens to the royal family when asked.

William accepted, without consulting Caroline, who thus changed overnight from a singer to an assistant astronomer. By the autumn of 1783 William had built a new and larger 20 ft reflector with a stable platform, and with this he began to 'sweep' the sky for NEBULAE, in a campaign that was to take two decades and increase the number of known nebulae from about 100 to 2500. Caroline shared the night watches, sitting at a desk at an open window and recording their observations. Next day she would write up a fair copy, and she later arranged their results for publication. On her spare nights she searched for comets, and she was to discover no fewer than eight. In addition, William had found that Flamsteed's British Catalogue of stars was riddled with errors, and so Caroline prepared an index that would take the user from the catalogued star back to the observations on which the catalogued data were based. This Index was published in 1798 by the Royal Society at its own expense.

In the 1770s William several times observed the Orion Nebula and thought it had altered shape and was therefore small, and so near and truly nebulous. Indeed, any nebula of a uniform milky appearance was, he thought,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

truly nebulous, while a distant star cluster would appear mottled. In 1784, however, he encountered nebulae that displayed both forms of nebulosity, and this led him to change his mind and conclude that all nebulae were star CLUSTERS, the differences in appearance being in part a reflection of the differences in the distances of the stars concerned.

This being so, the nebulae were clusters of stars, some of them scattered while others were tightly packed. Clustering implied the action of an attractive power or powers (presumably, but not necessarily, gravity), and with time a scattered cluster would become more compressed. This led William to introduce into astronomy the concept of development with time, a scattered cluster being younger than a compressed one. Furthermore, the need to classify the great numbers of specimens of nebulae that he was collecting led him to play the natural historian, a methodology wholly alien to traditional astronomy. This, plus the fact that few other astronomers could so much as glimpse the nebulae about which William was theorizing so freely, made his contemporaries uncertain of how to react to this gifted maverick.

In 1790 William came across what is now classified as the planetary nebula NGC 1514, but which to him was a star surrounded by a shell of nebulosity out of which it was surely condensing. This caused him to change his stance once more and to accept the existence of nebulosity that had not yet condensed into stars, but would in time do so.

William made many other contributions to astronomy, including the discovery of satellites of both Saturn and Uranus. Re-examining DOUBLE STARS after a long interval, he found that in some the two stars had moved in orbit around each other, though he did not have enough information to demonstrate that the force binding them together was gravitational. He published long lists in which stars of almost equal brightness were compared with each other, the intention being that a variation in the brightness of a particular star would reveal itself by disturbing the published comparisons. He investigated the shape of the Galaxy on the assumptions that his telescopes could penetrate to the borders in all directions, and that within the Galaxy the stars were distributed with fair uniformity, so that the number of stars he could see in a given direction was an index of the distance to the border in that direction. He carried out the star counts for a great circle of the sky and published an influential plan of this cross-section of the Galaxy; but in later life he abandoned both the assumptions on which it was based. In 1783 he published the first analysis of proper motions and concluded that the solar system is moving in the direction of Lambda Herculis, though the similarity between his result and that of modern analyses was partly a matter of luck. His experiments with sunlight led him in 1800 to the discovery of infrared radiation.

William also supplemented his income by making telescopes for sale, and many of the crowned heads of Europe were among his clients. For himself he completed in 1789 a monster 40 ft reflector with 4 ft mirrors. But it was cumbersome, the mirrors tarnished easily, and perhaps his own conversion to belief in nebulosity removed the reflector's primary *raison d'être*.

John Herschel

In 1788 William married, with unwelcome consequences for Caroline who found herself thereby displaced in his priorities. In 1792 his son John was born. Where William had been self-taught in astronomy, John was to enjoy a Cambridge education, and by 1815 was embarked on a career as a Cambridge don. His father, however, now in failing health, still had ambitions in astronomy that he could fulfil only through his son, and in 1816 he prevailed upon John to return home and become his astronomical apprentice. The 20 ft with which William had swept for nebulae was now in a poor state, and John was to build another to similar design. In 1821, Caroline came out of retirement to act as an amanuensis to her nephew in his trial sweeps for nebulae.

John's first task, however, was to re-examine his father's double stars, in collaboration with James South, another wealthy amateur who owned two fine equatorials ideal for the purpose. A few years later John was to be one of several astronomers who showed that the attractive force binding together the two members of a binary system was indeed gravitational.

John next used his 20 ft to re-examine his father's nebulae. In William's catalogues the nebulae were classified by type and this made the catalogues unsuitable for systematic re-examination. The problem was solved by Caroline, who had retired to Hanover after her brother's death in 1822. She rearranged the entire 2500 nebulae into a form convenient for John to use when sweeping, a work for which she was awarded the medal of the ROYAL ASTRONOMICAL SOCIETY (in whose foundation John had played a leading part, and of which William had been the nominal first President).

John's re-examination resulted in his 1833 catalogue of 2306 nebulae. It was now time to complete his father's work by surveying the southern skies invisible from England. Taking the 20 ft, an equatorial purchased from South, and one of Caroline's 'comet sweeper' telescopes, he sailed to the Cape of Good Hope where from 1834 to 1838 he engaged in an observational campaign of unparalleled intensity. His results were published in a sumptuous volume in 1847, 'being a completion of a telescopic survey of the whole surface of the visible heavens', for John had examined the entire heavenly sphere with his 20 ft. His return in 1838 marked the end of his career as an observer, and the end too of the Herschels' virtual monopoly of observations of nebulae, for in 1839 the future Lord Rosse at Birr in Ireland completed a reflector with 3 ft mirrors, followed in 1845 by 'the Leviathan of Parsonstown' with its 6 ft mirrors. John did, however, advance astronomy by working at his desk. In particular, *Philosophical Transactions* for 1864 contained his catalogue of over 5000 nebulae and star clusters; the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

great majority had been discovered by him or by his father, and the enlarged version of the catalogue published by J L E Dreyer in 1888 is the NEW GENERAL CATALOGUE OF NEBULAE AND CLUSTERS OF STARS or NGC in widespread use today.

As a theorist John lacked the daring of his father, and he was inhibited too by the obligation he felt always to portray his late father's work in a favorable light, even when he himself had come to a different view. Also, by the mid-nineteenth century the daring, even shocking theorizing of his father had become components of mainstream astronomy. William, Caroline and John had enlarged the science from the study of the solar system to the study of the cosmos.

Alexander Stewart Herschel and John Herschel II

Two of John's sons, Alexander Stewart and John II (both born in South Africa), made their own contributions to astronomy, the former particularly in cometary and meteoric studies. The later members of the family carefully preserved the manuscripts of their forebears, which are now largely in permanent collections, notably the Library of the Royal Astronomical Society.

Bibliography

Buttmann G 1974 *The Shadow of the Telescope: A Biography of John Herschel* (Guildford: Lutterworth)

Hoskin M 1982 *Stellar Astronomy: Historical Studies* (Chalfont St Giles: Science History Publications)

Hoskin M (ed) 1997 *The Cambridge Illustrated History of Astronomy* (Cambridge: Cambridge University Press) ch 7

Hoskin M (ed) 1999 *The Cambridge Concise History of Astronomy* (Cambridge: Cambridge University Press) ch 7

Lubbock C A 1933 *The Herschel Chronicle* (Cambridge: Cambridge University Press)

Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hertzsprung, Ejnar (1873–1967)

Danish astronomer, studied chemical engineering and worked as a chemist in St Petersburg before returning to Denmark to take up astronomy, worked in Göttingen and Potsdam with KARL SCHWARZSCHILD, and then at the Leiden Observatory (Netherlands) where he became director. Hertzsprung determined the proper motions of stars, and about a million positions of binary stars, as well as the brightnesses of stars, determining by photography not only their brightnesses, but also their colors (expressed as the ratio of a star's brightness through two different colored pieces of glass). A star's color is connected to its temperature, and its brightness relative to others in a cluster all at the same distance is connected to its absolute magnitude. In 1911 he plotted these quantities in what became known as a color–magnitude diagram for the Hyades star cluster, and discovered the distinction between giant stars and 'main sequence' stars. Two years later HENRY NORRIS RUSSELL plotted an equivalent diagram for stars whose spectral classes were known and which were not in clusters but so nearby that their distances and thus their absolute magnitudes could be determined. A star's spectral class is, like color, connected to its temperature. The spectral class–absolute magnitude diagram also showed the distinction between giants and main sequence stars like the color–magnitude diagram. The two diagrams have the same topography, and collectively are known as the *Hertzsprung–Russell diagram*. (By an accident of history, an apparently independent parallel study of the 'relation between brightness and spectral type in the Pleiades' was undertaken by Hans Rosenberg (1879–1940) of Göttingen, whose overlooked paper contained the first published Hertzsprung–Russell diagram (June 1910).) Hertzsprung also determined the distance to the Small Magellanic Cloud, by use of the statistical parallax of a group of Cepheid stars (triangulating onto the stars over the baseline of the Sun's journey through space over several years). Although he made a mistake, putting the stars a factor of ten too close, possibly through a slip of the pen, it is interesting that the distance, the largest distance of anything known at the time, was so large that a factor of ten too small did not immediately stand out.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hertzsprung–Russell Diagram

The Hertzsprung–Russell diagram (HR-diagram), pioneered independently by EJNAR HERTZSPRUNG and HENRY NORRIS RUSSELL, is a plot of the star luminosity versus the surface temperature. It stems from the basic relation for an object emitting thermal radiation as a black body:

$$L = 4\pi\sigma R^2 T_{\rm e}^4 \qquad (1)$$

where L is the luminosity (energy emitted per second by the whole radiating surface), σ is the Stefan–Boltzmann constant, R is the radius and $T_{\rm e}$ is defined as the temperature of a hypothetical black body of the same radius R radiating the same total luminosity L. In general, physical temperatures in stellar surface regions closely approximate $T_{\rm e}$, and the place where a photon has 50% probability of radiating directly into space—the photosphere—is well defined. Thus R in equation (1) is also well defined. In red giant and supergiant stars the radius is more uncertain, because they have very extended atmospheres partly out of hydrostatic balance.

The luminosity is customarily expressed in solar units $L_\odot = 3.90 \times 10^{33}$ erg s^{-1} and plotted on a magnitude scale $M_{\rm b} = -2.5 \log L/L_\odot + M_{\rm b,\odot}$, where $M_{\rm b}$ and $M_{\rm b,\odot} = 4.77$ are the bolometric magnitudes of the star and the Sun, respectively.

Every star at every instant of its life has a given L and R and therefore $T_{\rm e}$ which are dictated by the global structure of the star and the physical processes taking place in it. Since the structure of a star is not constant in time because of the continuous energy leakage from the surface (and nuclear fusion), L and $T_{\rm e}$ will change as the star evolves. Therefore a star will describe a path in the HR-diagram which is a function of its mass and chemical composition. It follows that the HR-diagram is a powerful diagnostic of the star's structure and evolution during its whole life.

Different populations of stars have different HR-diagrams: they go from the well-behaved ones of globular clusters, in which almost all evolutionary phases are detectable, to the poor ones of open clusters and associations, where only very few stars beyond the main sequence are visible, to the patchy HR-diagrams of composite stellar populations, in which at least one more dimension is added, namely the past history of stellar formation.

By studying the path in the HR-diagram of single stars and/or the morphology of the HR-diagram of stellar aggregates of different complexities (from clusters to galaxies) we will learn very much not only about how stars are made and evolve but also about many physical properties of the host system.

Passing from observations to theory and vice versa

The observational data are apparent magnitudes and colors in a given photometric system (for instance the classical Johnson UBV system). The colors are defined as the difference $(m_{\Delta\lambda} - m_{\Delta\lambda'})$ between any two magnitudes $m_{\Delta\lambda}$ and $m_{\Delta\lambda'}$ corresponding to the fluxes in the adopted pass-bands $\Delta\lambda$ and $\Delta\lambda'$. Going from apparent magnitudes $m_{\Delta\lambda}$ and colors to luminosities and effective temperatures is not trivial as a number of steps are needed. Determining the absolute luminosity requires distance, color excess, and bolometric corrections. Determining $T_{\rm e}$ is even more indirect. For the few cases for which the radius R is known, $T_{\rm e}$ is derived from its definition via the black body law. In all other cases some measure of the temperature is estimated by at least three different methods: (i) from the entire spectral energy distribution matched by a black body of given temperature; (ii) from the Wien law $\lambda_{\rm max} T = \cos t$, where $\lambda_{\rm max}$ is the wavelength of the maximum intensity of the spectrum; (iii) from suitable relationships between colors and temperature based on preliminary comparison with the black body emission. Each method is affected by some degree of uncertainty. In the same star, these different methods can yield different values of temperature. These are generally close to each other and to the idealized $T_{\rm e}$ defined by equation (1).

From an operational point of view, we start from apparent magnitudes and colors, correct them for extinction if possible, translate apparent magnitudes $m_{\Delta\lambda}$ into absolute magnitudes $M_{\Delta\lambda}$ if the distance is known, derive from $M_{\Delta\lambda}$ the absolute bolometric magnitudes $M_{\rm b}$ applying the so-called bolometric correction (BC), transfer colors into $T_{\rm e}$, and finally compare the resulting HR-diagram with the theoretical counterpart obtained from the theory of stellar structure and evolution. The BC is required to account for the fraction of the flux not detected by the pass-band in use. It is worth recalling here that BCs and color–$T_{\rm e}$ relationships are complicated functions of spectral type, luminosity class and chemical composition X, Y and Z, i.e. the fractional mass abundances of hydrogen, helium and all heavier elements lumped together. The whole procedure can be repeated backward, i.e. starting from theoretical L and $T_{\rm e}$ to predict the corresponding observational magnitudes and colors.

HR-diagram of cluster and field stars

Figure 1 shows the apparent HR-diagrams of two Galactic GLOBULAR CLUSTERS, the metal-rich 47 Tucanae and the metal-poor M55. Each HR-diagram clearly illustrates the properties of assemblies of stars with the same age, chemical composition and distance.

Most of the stars fall along the main sequence (MS) band, running from low to high magnitudes and colors. The bending portion of the top of the main sequence defines the so-called turn-off (TO). Along the main sequence are stars in the core H-burning phase (see below). The remaining stars are distributed in four different loci.

(i) The sub-giant branch (SGB) and the red giant branch (RGB) are so named for the large dimensions of the stars. The sub-giant branch runs nearly horizontally from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

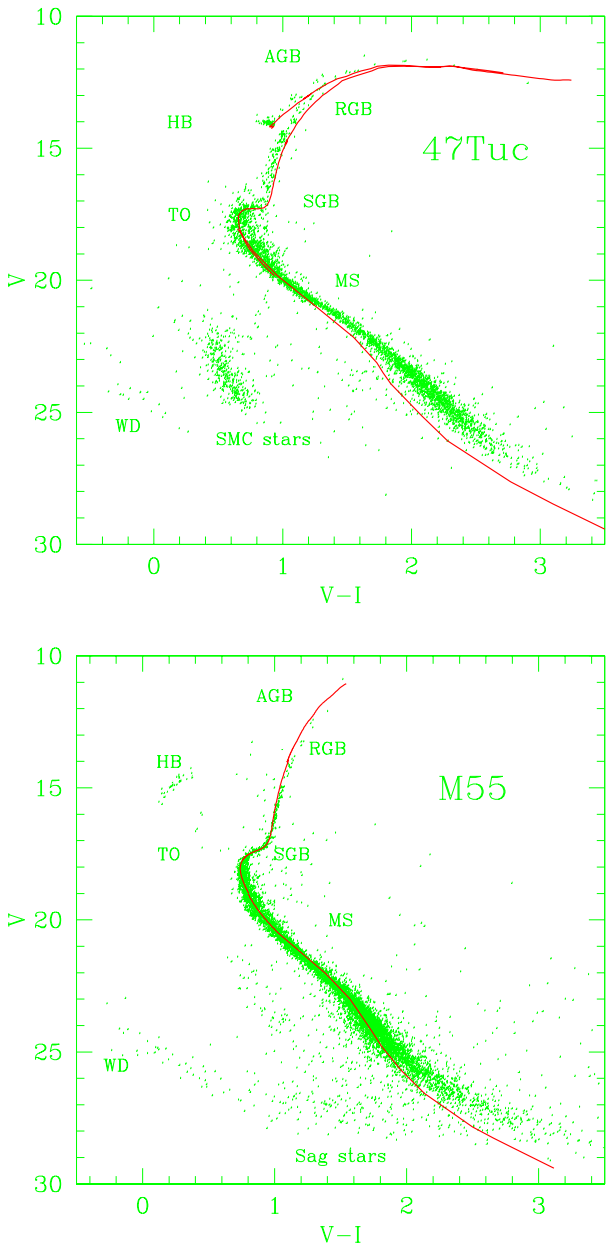


Figure 1. Upper panel: the HR-diagram of the metal-rich Galactic globular cluster 47 Tucanae. Superposed is the isochrone with $Z = 0.008$ and age of 10 Gyr. Lower panel: the same but for the metal-poor Galactic globular cluster M55. The superposed isochrone has metallicity $Z = 0.0004$ and age of 14 Gyr. In these figures, V is visual and I an infrared bandpass. Courtesy of M Zoccali.

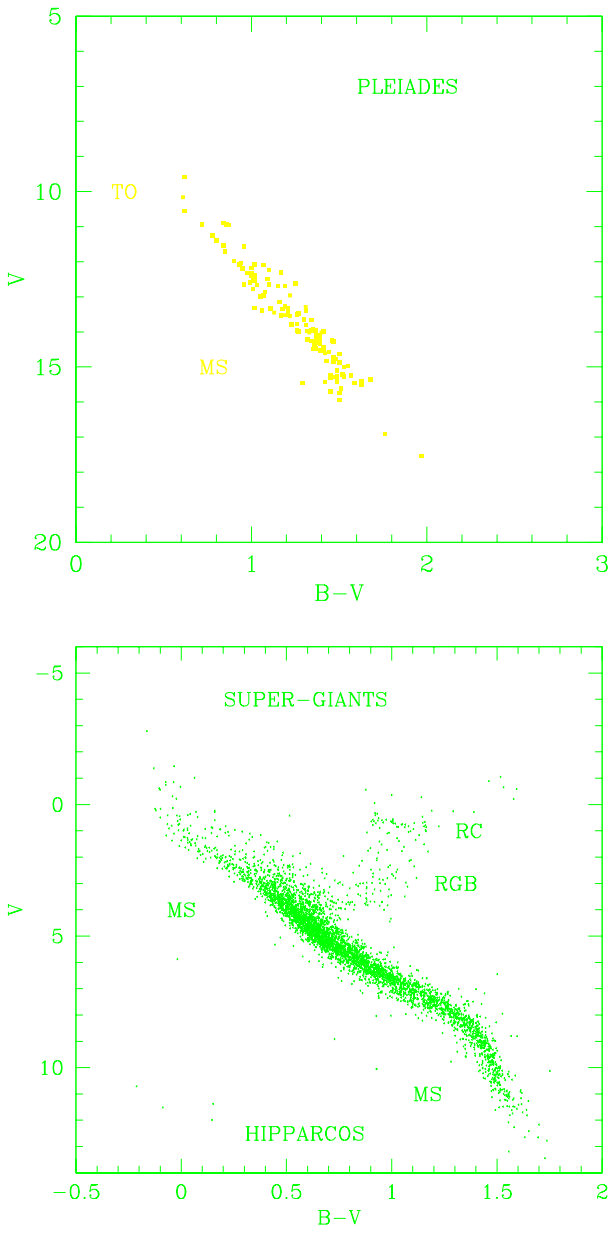


Figure 2. Upper panel: the HR-diagram of the Pleiades. Lower panel: field stars in the solar vicinity (within 50 pc of the Sun) after Hipparcos.

the turn-off to the bottom of the red giant branch. This is nearly vertical for metal-poor clusters (M55), whereas it has a pronounced curvature for metal-rich clusters (47 Tucanae). For the latter, the brightest red giant branch stars in the V-passband are not the reddest ones. The subgiant and red giant branch stars are in the shell H-burning phase prior to core He ignition.

(ii) The group of stars slightly bluer than the red giant branch and crowding into a clump in the case of 47 Tucanae, or significantly bluer and more dispersed (a sort of blue tail) in the case of M55, both correspond to the so-called horizontal branch (HB). This was so named originally because, in many intermediate-metal clusters, the corresponding feature defines a broad, essentially horizontal distribution in V. As it happens, with due allowance for the BC, even dropping blue

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

horizontal branches like that of M55 are indeed remarkably horizontal. Horizontal branch stars are burning helium in the core.

(iii) Just above the red giant branch and with a merging path are the few asymptotic giant branch (AGB) stars. They are in post core He-burning stages ('double shell source' stage) prior to the ejection of their envelopes to become first planetary nebulae (PNe) and later white dwarfs (WDs).

(iv) Finally, there is the fainter sequence of white dwarf stars that are the end product of globular cluster stars' evolution. The white dwarf sequence is nearly parallel to the main sequence, but about 12 mag fainter, and nicely corresponds to a line of constant radius (see below). The white dwarf sequence was rarely visible in the past and became clearly detected only with the advent of modern ground and space telescopes.

To complete the star census, it is worth recalling that in both panels of figure 1 contamination by fore- and background stars is clearly visible. In the case of 47 Tucanae the blob of stars fainter than the main sequence but brighter than the white dwarf sequence are stars of the Small Magellanic Cloud (SMC), displaced in this part of the HR-diagram because of the large distance. In the case of M55, the stars blurring the main sequence band on both sides are those of the Sagittarius galaxy against which M55 is observed. Situations like this are common to other HR-diagrams and must be kept in mind for a correct interpretation to be obtained.

The upper panel of figure 2 shows the HR-diagram of the Pleiades, a sparsely populated open cluster, in which only the main sequence is visible. Finally, the lower panel of figure 2 shows the recent HR-diagram of field stars in the solar vicinity (within 50 pc of the Sun) observed by the Hipparcos satellite. The novelty of this diagram with respect to many previous versions of it is that for the first time all these stars have distances measured with the greatest accuracy. One may easily identify the main sequence band, the sub-giant branch, the red giant branch, a clump of red stars hereafter known as the 'red clump' (RC), and finally the sequence of white dwarfs.

Why do we have such different HR-diagrams? What are the physical causes driving the evolution of stars and determining the morphology of the HR-diagram? Can we infer from their properties important parameters such as the age of the stars in order to set limits on the age of the universe? What can we learn from HR-diagrams about the past history of galaxies? Answering these questions is the aim of the following sections.

Basic physics of stellar interiors

Looking at the Sun as a prototype star, convincing arguments indicate that it has been radiating energy at nearly the same rate for several billion years (at least 4.6 Gyr), thus making clear that several conditions have always been satisfied.

Hydrostatic equilibrium

If the complex structure of the Sun (or a star in general) is reduced to a self-gravitating isotropic sphere of gas of constant mass, this means that gravitational pull is balanced by differential gas pressure. Neglecting all other forces but those due to gravity and pressure, at each layer of radial distance r the conditions of hydrostatic equilibrium and mass conservation are

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM_\mathrm{r}}{r^2}\rho \qquad \text{and} \qquad \frac{\mathrm{d}M_\mathrm{r}}{\mathrm{d}r} = 4\pi r^2 \rho \tag{2}$$

where M_r is the mass contained in the sphere of radius r and ρ is the local density. It can easily be checked that even a very small departure from hydrostatic equilibrium would imply a variation of the star radius over the free-fall time scale

$$t_\mathrm{ff} = 2\left(\frac{R^3}{GM}\right)^{1/2} = 3.2\times 10^3 \left(\frac{R}{R_\odot}\right)^{3/2}\left(\frac{M}{M_\odot}\right)^{1/2} \tag{3}$$

which in the case of the Sun is about 1 h. We will assume hydrostatic or quasi-hydrostatic equilibrium from now on.

Thermal equilibrium

This demands that the energy leakage from any spherical surface within the star is balanced by continuous supply from an internal reservoir. In other words the following equation must be satisfied once we have a quasi-static star:

$$\frac{\mathrm{d}L_\mathrm{r}}{\mathrm{d}r} = 4\pi r^2 \rho \epsilon \tag{4}$$

where L_r is the luminosity emitted by the sphere of radius r and ϵ is the amount of energy per unit mass and time generated inside it by all sources. What can produce the 3.90×10^{33} erg s^{-1} of the Sun's luminosity and do it for about 4.6 Gyr, or, in general, the luminosities of any other star over its whole life? These questions, taken together, limit the candidate mechanisms to gravitational contraction and nuclear fusion.

Gravitational energy

According to the virial theorem for a system in hydrostatic equilibrium, any variation of the gravitational potential by contraction goes to increasing the internal energy and liberating (radiating) an equal amount of energy. An estimate of the thermal energy of a star is soon derived from the virial theorem and gravitational potential energy Ω:

$$E_\mathrm{T} \simeq -\frac{1}{2}\Omega \simeq \frac{1}{2} q \frac{GM^2}{R} \tag{5}$$

with $q = 3/2$ (a numerical factor accounting for the mass distribution). Supposing that the whole internal energy is radiated away at the rate L (luminosity), the duration of this source is

$$t_\mathrm{K} = \frac{3}{4}\frac{GM^2}{RL} = 2\times 10^7 \left(\frac{M}{M_\odot}\right)^2 \frac{L_\odot}{L}\frac{R_\odot}{R}\ \mathrm{yr}. \tag{6}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thus, gravitational contraction and hence internal energy could power the Sun for about 20 million years. This is too short compared with the time scales of biological and geological evolution on the Earth. Similar considerations apply to any other type of star.

Nuclear fusion

Thermonuclear reactions in stellar interiors are indeed the main source of energy (see STELLAR INTERIORS: THERMODYNAMICS). The fusion of light nuclei into more massive nuclei (for instance $4 \times {}^1\mathrm{H}$ into $1 \times {}^4\mathrm{He}$ in main sequence stars) provides the required amount of energy for a very long time. To get positively charged nuclei to fuse together despite their electrical repulsion, they must approach each other closer than about 10^{-13} cm, which is the range of the strongly attractive nuclear forces (roughly the nuclear radius R_N). This means that the positive Coulomb potential energy $V_\mathrm{cou} = Z_1 Z_2 e^2/r$ for $r \geq R_\mathrm{N}$ suddenly drops to large negative values for $r < R_\mathrm{N}$. At R_N V_cou is at a maximum and of the order of 1 MeV (peak of the so-called potential barrier). At the temperatures and densities of stellar interiors, the relative kinetic energy of random thermal motions is only $E_\mathrm{k} \simeq 10^3$ eV and therefore no reactions should occur. However, thanks to QUANTUM MECHANICS, even in this case there is a finite albeit small probability of barrier crossing and reactions take place. The energy liberated by each fusion event is

$$\Delta E = \Delta M c^2 \tag{7}$$

where ΔM is the mass difference between the combined mass of the original nuclei and that of the fused nucleus. In the case of four ${}^1\mathrm{H}$ fusing into one nucleus of ${}^4\mathrm{He}$, the ${}^4\mathrm{He}$ nucleus has a mass of $3.97 m_\mathrm{p}$ (m_p is the proton mass). Therefore an amount of mass equal to $0.03 m_\mathrm{p}$ has been converted into energy and liberated to form the compound nucleus. This energy amounts to 6.3×10^{18} erg (g of matter)$^{-1}$. Similarly, when three ${}^4\mathrm{He}$ fuse to form a nucleus of ${}^{12}\mathrm{C}$, the energy liberated per gram is about 0.61×10^{18} erg. Remarkably, H fusion liberates roughly 10 times more energy than He fusion. We will see that this corresponds to the observed fact that main sequence stars are about 10 times more numerous than all post-main sequence stars. Fusion of light nuclei into heavier ones liberates energy up to the synthesis of ${}^{56}\mathrm{Fe}$, which has the largest binding energy per nucleon $\Delta M\, c^2/A$ (A is the mass number). Above this, not only is the electrical repulsion between nuclei so strong that in practice no nucleus has enough energy to cross the barrier but also the binding energy starts decreasing. This means that energy must be supplied to nuclei from the thermal reservoir in order to make them fuse. This fact limits the maximum number of large-scale fusion steps that may occur inside a star. Each of these requires a minimum temperature (and density) to occur, which means a minimum mass of the star (or its inner core).

H burning

Fusion of ${}^1\mathrm{H}$ into ${}^4\mathrm{He}$ occurs through two channels called the PP chain and the CNO cycle. The PP chain in its simplest version is a sequence of reactions: first ${}^1\mathrm{H}({}^1\mathrm{H}, \mathrm{e}^+\nu){}^2\mathrm{H}$ where two protons react to form deuterium plus a positron and a neutrino, followed by ${}^2\mathrm{H}({}^1\mathrm{H}, \gamma){}^3\mathrm{He}$. The first two reactions must occur twice in order to produce two ${}^3\mathrm{He}$ nuclei. Finally, these fuse through ${}^3\mathrm{He}({}^3\mathrm{He}, 2\,{}^1\mathrm{H}){}^4\mathrm{He}$. In total, four protons are fused into one nucleus of ${}^4\mathrm{He}$. The initial reaction can occur at about 1×10^6 K, but the last one requires temperatures greater than about 10×10^6 K. The individual reactions are very slow and a very small fraction of nuclei have enough energy to overcome the electrical repulsion. Even so, in a star like the Sun every second 6×10^{14} g of H are converted into 5.96×10^{14} g of He and 4×10^{12} g of matter are transformed into energy. The CNO cycle proceeds through a different mechanism. The initial step is a nucleus of ${}^{12}\mathrm{C}$ reacting with a proton to give rise to a series of reactions whose net result is the transformation of $4 \times {}^1\mathrm{H} \rightarrow 1 \times {}^4\mathrm{He}$, leaving all other species unchanged. The CNO cycle requires significantly higher temperatures and it is much more sensitive to temperature than the PP chain. In the CNO cycle $\epsilon_\mathrm{CNO} \propto T^{20}$; in the PP chain $\epsilon_\mathrm{PP} \propto T^5$. Main sequence stars more massive than about $1.5 M_\odot$ generate their nuclear energy via the CNO cycle at temperatures larger than 17×10^6 K.

He burning

The next step is the non-simultaneous fusion of three ${}^4\mathrm{He}$ according to ${}^4\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^8\mathrm{Be} + {}^4\mathrm{He} \rightarrow {}^{12}\mathrm{C} + \gamma$. It is accompanied by the reactions ${}^{12}\mathrm{C}({}^4\mathrm{He}, \gamma){}^{16}\mathrm{O}$ and ${}^{12}\mathrm{O}({}^4\mathrm{He}, \gamma){}^{20}\mathrm{Ne}$. He burning, called the 3α process, requires high temperatures ($T > 100 \times 10^6$ K) and takes place only after stars have exhausted the central H content and formed a central core of pure He. The threshold mass of the He core depends on the physical state of the gas. In the electron degenerate material of the central regions of low mass stars (up to about $2.2 M_\odot$) the minimum mass is about $0.5 M_\odot$, which immediately sets also a gross lower limit for a star to be able to ignite He in the core. All low mass stars in the range $0.5 < M/M_\odot \leq 2.2$ undergo explosive He ignition because core contraction proceeds to degeneracy before the ignition temperature is reached. Stars more massive than $2.2 M_\odot$ ignite helium without becoming electron degenerate and do it when the He core mass is only $\simeq 0.3 M_\odot$.

Later burnings stages

The following principles influence the occurrence of later burning stages beyond the 3α process. (i) Higher and higher threshold temperatures and densities are required; this implies that minimum values of the core mass must be reached by the stars. (ii) The amount of liberated energy decreases so that the duration of later burning stages becomes shorter and shorter. (iii) Nuclear burning can only proceed up to the formation of an iron core. (iv) With increasing core density the gas becomes more and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

more electron degenerate. (v) Ignition of a nuclear burning stage in a degenerate gas is potentially explosive and may lead to complete disruption of the star. (vi) External causes (for instance mass loss by stellar wind) may prevent a star from reaching the required core mass for a certain burning stage to occur.

Following core He exhaustion, in stars more massive than about $10M_\odot$ carbon is ignited at temperatures of about 500×10^6 K. Two nuclei of ^{12}C are fused into a variety of heavy elements, for instance $^{12}C(^{12}C, \gamma)^{24}Mg$, or $^{12}C(^{12}C, n)^{23}Mg$, or $^{12}C(^{12}C, ^4He)^{20}Ne$ etc. After core C exhaustion, at temperatures of about 1×10^9 K, high-energy thermal photons begin to break up ^{20}Ne by the reaction $^{20}Ne(\gamma, ^4He)^{16}O$, followed by $^{20}Ne(^4He, \gamma)^{24}Mg$. The core of the star consists now mainly of ^{16}O and ^{24}Mg. When the temperature reaches 2×10^9 K, oxygen burning begins. Although it is schematically indicated by $^{16}O(^{16}O, ^4He)^{28}Si$, because the most important product is ^{28}Si, many other elements are generated such as Mg, S and P and copious γ, p, n and α particles are liberated. Finally, at temperatures of about 3×10^9 K, Si burning occurs through a complex network of nuclear reactions briefly indicated by $^{28}Si + ^{28}Si \rightarrow ^{56}Ni$. Although these advanced stages involve the same physics of thermonuclear reactions encountered in previous stages, they are characterized by photo-dissociation and radiative captures that transform the nuclear material into its most stable form. The process is started by $^{28}Si(\gamma, ^4He)^{24}Mg$. The liberated ^{4}He nuclei induce a long series of reactions such as $^{28}Si(^4He, \gamma)^{32}S$, $^{32}S(^4He, \gamma)^{36}Ar$, $^{36}Ar(^4He, \gamma)^{40}Ca$, ..., $^{52}Fe(^4He, \gamma)^{56}Ni$. The nuclear sequence is terminated with Si-burning and the formation of the most stable nuclei.

Lifetime of a nuclear phase

Let f_i be the mass fraction of the star in which the temperature is high enough to burn a certain nuclear fuel i (hydrogen, helium, carbon, etc), X_i the fractional mass abundance of it and E_i the energy released per gram of matter by the corresponding complex of nuclear reactions. The duration t_i of the phase is given by

$$t_i \simeq E_i \, f_i \, X_i \frac{M}{L}. \tag{8}$$

In the case of H burning, inserting $E_H = 6.0 \times 10^{18}$ erg g^{-1}, $f_H = 0.1$, $X_H = 0.7$ (typical values) and the mass and luminosity of the Sun we obtain $t_H \simeq 7$ Gyr. In the case of He burning for which $E_{He} = 0.6 \times 10^{18}$ erg g^{-1}, $f_{He} = 0.2$, $X_{He} \simeq 1$, and a star with the same mass and luminosity as those in the horizontal branch phase (see case F in table 2), we obtain $t_{He} \simeq 10^8$ yr.

Main sequence stars obey the mass–luminosity relationship $L \propto M^\alpha$ with $\alpha = 3.5$–4 (in reality a core mass–luminosity relationship applies to all evolutionary phases). Therefore, the nuclear lifetime of any evolutionary phase depends on the star mass or core mass. If we limit ourselves to the case of main sequence stars for which the lifetime is the longest, $f_H \simeq 0.1$ and $X_H = 0.7$ we obtain the approximate relation

$$t_H \simeq 10^{10} \left(\frac{M}{M_\odot} \right)^{-\alpha+1} \text{ yr}. \tag{9}$$

More precise H-burning lifetimes are given in table 1. They vary from a few million years for very massive stars ($100M_\odot$) to many hundreds of billions of years for stars of very low mass ($0.1M_\odot$). For stars like the Sun the H-burning lifetime is about 10 Gyr.

The energy production by nuclear sources is regulated by the stellar luminosity (and ultimately by the opacity of the material); therefore the whole history of a star can be viewed as a series of slow nuclear burning stages alternating with periods of fast gravitational contractions causing the increase in the central (mean) temperature to the threshold value for the next nuclear step to occur. As a consequence t_K and t_N are also the time scales over which a star changes its properties (L, T_e, etc).

Equation of state

To a high degree of precision, stellar interiors are in conditions of local thermodynamic equilibrium so that radiation is described by the Planck law, particles obey the appropriate law of statistical mechanics and the interaction between matter and radiation is governed by the Kirchhoff law.

The structure and evolution of a star are particularly sensitive to the equation of state governing the pressure P as a function of density ρ, temperature T and molecular weight μ of the particles. Matter in stellar interiors is composed of ions and/or nuclei (complete ionization) and electrons, immersed in a bath of photons. Each component contributes to the total pressure $P = P_r + P_i + P_e$.

The functional dependence $P(\rho, T, \mu)$ for each component depends on the quantum number density of particles with momentum p (an isotropic distribution of momenta is supposed). Quantum statistics predicts that fermions (protons, neutrons, electrons etc) and bosons (photons) have different number-momentum distributions $n(p)$. However, for all particles but photons, $n(p)$ reduces to the classical Maxwell–Boltzmann law when the mean inter-particle separation is much greater than the de Broglie wavelength $\lambda_B = h/p$ (h is the Planck constant). Therefore, there will be a certain density at which the quantum behavior shows up. This condition is known as the onset of degeneracy. Because of their small mass and hence long λ_B, electrons are the first particles to become degenerate. Protons, neutrons and nuclei in general reach degeneracy only at densities comparable with nuclear densities (10^{15} g cm^{-3}), which occurs only in very peculiar situations that are ignored here. In the case of photons, considered as relativistic, zero-mass particles, the quantum $n(p)$ must always be used. Based on these considerations, the equation of state for the various components is found by integrating over the corresponding functions $n(p)$. The results are as follows.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Typical parameters of main sequence stars and core H-burning lifetimes.

$M/M_\odot$	$\log L/L_\odot$	$\log T_e$	$R/R_\odot$	T_c	ρ_c	M_c/M	M_e/M	Energy	t_H
30.0	5.11	4.60	6.6	37	3.3	0.61	0	CNO	5.7×10^6
15.0	4.29	4.49	4.7	33	5.8	0.49	0	CNO	1.2×10^7
9.0	3.60	4.37	3.5	31	9.9	0.41	0	CNO	3.0×10^7
5.0	2.76	4.23	2.3	27	20.9	0.32	0	CNO	1.1×10^8
3.0	1.91	4.09	1.7	24	40.3	0.25	0	CNO	3.7×10^8
2.0	1.25	3.36	1.4	21	65.7	0.19	0	CNO	1.1×10^9
1.0	−0.16	3.75	0.9	14	97.5	0	0.1	PP	9.7×10^9
0.5	−1.39	3.59	0.4	9	70.1	0	0.3	PP	8.7×10^{10}
0.3	−1.90	3.56	0.3	8	106.4	1	1.0	PP	1.7×10^{12}

The surface temperature T_e is in K; the central temperature T_c is units of 10^6 K; the central density ρ_c is in g/cm^3. The fractional masses M_c/M and M_e/M show the extension of the convective core and external convective envelope respectively. t_H is the lifetime in years of the core H-burning phase.

The photon pressure P_r depends only on the gas temperature:

$$P_r = \frac{1}{3}aT^4 \qquad \text{where} \quad a = \frac{8\pi^5 k^4}{15c^3h^3} = 7.564 \times 10^{-15} \text{ erg cm}^{-3} \text{ K}^{-4}. \tag{10}$$

There is no density dependence, because the number of photons is not conserved owing to continuous emission and absorption, a result already predicted by classical thermodynamics.

At ordinary low densities (no electron degeneracy) ions and electrons have the equation of state

$$P_g = P_i + P_e = \frac{k}{\mu_i m_p}\rho T + \frac{k}{\mu_e m_p}\rho T = \frac{k}{\mu m_p}\rho T \tag{11}$$

where k is the Boltzmann constant, m_p is the proton mass, and μ_i, μ_e, μ are the mean molecular weights of just ions, just electrons and the mixture of the two, respectively. The mean molecular weights of a fully ionized gas with abundance mass fractions X, Y and Z are

$$\mu_i = \frac{4}{4X+Y} \qquad \mu_e = \frac{2}{1+X} \qquad \mu = \frac{1}{2X + 0.75Y + 0.5Z}. \tag{12}$$

At increasing density (and decreasing temperature), electrons become degenerate. The equation of state is very complicated. It acquires a simple form only in the extreme cases of fully degenerate, non-relativistic electrons ($kT \ll m_e c^2$, thermal energy much smaller than the electron rest energy) and fully degenerate, relativistic electrons ($kT \gg m_e c^2$), for which we have

$$P_e = \frac{h^2}{20m_e}\left(\frac{3}{\pi}\right)^{2/3}\left(\frac{\rho}{\mu_e H}\right)^{5/3}$$

or

$$P_e = \frac{ch}{4}\left(\frac{3}{8\pi}\right)^{1/3}\left(\frac{\rho}{\mu_e H}\right)^{4/3}. \tag{13}$$

The first and second equations are for non-relativistic and relativistic electrons, respectively. The lower power of ρ in the latter expression is due to the electrons' velocities saturating at the speed of light c. In both cases the pressure of degenerate electrons does not depend on temperature but only on density.

Studying the equation of state in many stellar environments we learn that P_r is important in massive stars, and among main sequence stars electron degeneracy is important only in the low-mass ones. In more advanced evolutionary phases, electron degeneracy progressively becomes the dominant central source of pressure.

The lack of temperature dependence in the equation of state of fully degenerate electrons bears very much on stellar structure. If for any reason the temperature starts increasing, for instance by deposition of energy from nuclear reactions commencing in a degenerate gas, the equation of state does not provide a compensating expansion to quench the instability, and the temperature increases even further, thus making nuclear energy release even stronger. Thermal runaway soon occurs. In many circumstances a thermonuclear explosion occurs, tearing the star apart.

Energy transport

There are three mechanisms for the transport of heat inside a star: (i) the random thermal motion of individual electrons and/or ions that interacting with each other transport energy from hot to cold regions (thermal conduction); (ii) The continuous absorption, emission, scattering of photons with the same effect (radiative transfer). In both cases the link between flux of energy and temperature gradient is derived from the same principles. (iii) The third is the large-scale motions of fluid elements, in which rising blobs of hot, less dense gas and sinking blobs of cool, denser gas transfer energy from hot to cold regions of a star. This mechanism is called convection.

Radiative transfer

Based on simple dimensional arguments, the flux density of heat along the radial direction r due to propagation of photons with velocity c, frequency-averaged mean free path $\tilde{l}$ and heat capacity per unit volume $C = \mathrm{d}U/\mathrm{d}T =$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$4aT^3$, where U is the internal energy density aT^4, is given by

$$\frac{L_r}{4\pi r^2} = -\frac{1}{3}c\tilde{l}\frac{dU}{dT}\frac{dT}{dr} = -\frac{1}{3}c\tilde{l}4aT^3\frac{dT}{dr}$$

or

$$\left(\frac{dT}{dr}\right)_R = -\frac{3\rho\kappa}{4acT^3}\frac{L_r}{4\pi r^2} \qquad (14)$$

where the mean free path $\tilde{l}$ has been averaged over all frequencies of the black body spectrum to account for the frequency dependence of the absorption, emission and diffusion processes. In doing so, the frequency-averaged mean free path can be expressed in terms of the so-called Rosseland opacity $\kappa = 1/\rho\langle\tilde{l}\rangle$ and the fundamental equation for radiative transport is obtained.

The above equation can be interpreted in the following way: given a certain temperature gradient (and opacity κ) implied by the mass and pressure balance, the luminosity L_r is determined. The steeper the temperature gradient and the lower the opacity, the higher is the luminosity.

Convective transport

Can the radiative temperature gradient be arbitrarily steep? The answer is no, because when the gradient becomes steeper than the local adiabatic value, convection sets in and a different temperature gradient is built up. However, for the very outer layers in which a more complicated treatment is required, thanks to the high thermal capacity of convective elements in all but distended giant envelopes, the real temperature gradient is nearly identical to the adiabatic one and the following approximation holds:

$$\left(\frac{dT}{dr}\right)_C = \frac{\gamma - 1}{\gamma}\frac{T}{P}\frac{dP}{dr} \qquad (15)$$

where $\gamma = c_p/c_v$, i.e. the ratio of the specific heats at constant pressure and constant volume for the gas under consideration.

Where is convection?

Performing some algebraic manipulations, we may express the radiative gradient as $(d\ln T/d\ln P)_R \propto \kappa L_r P/M_r T^4$, where the luminosity–mass ratio L_r/M_r has been introduced. Neither $\kappa P/T^4$ nor L_r/M_r is constant throughout a star. It can be shown that the ratio $P/T^4 \propto P/p_r$ is nearly constant, whereas κ greatly varies throughout a star. In the outer layers κ is very large because of ionization, while L_r/M_r reaches a nearly constant value. Therefore we expect the external layers of a star to be often in convective conditions. In contrast, in the internal regions, while the opacity decreases to a minimum (electron scattering), the ratio L_r/M_r increases in a way that depends on the temperature sensitivity of the nuclear energy source. It follows that low-mass stars with $\epsilon_{PP} \propto T^5$ have radiative cores, whereas massive stars with $\epsilon_{CNO} \propto T^{20}$ have convective cores. Later burning stages (helium, carbon, etc) will always occur in convective conditions. Table 1 contains a few characteristic parameters for main sequence stars of different mass.

Central values

Useful analytic expressions for the central density, pressure and temperature of main sequence stars in which the equation of state is that of a perfect gas (no radiation pressure and no degeneracy) are

$$\rho_c = 1.4\frac{M}{R^3}\ \text{g cm}^{-3} \qquad P_c = 1.1\times10^{16}\frac{M^2}{R^4}\ \text{dyn cm}^{-2}$$

$$T_c = 2.3\times10^7\mu\frac{M}{R}\ \text{K} \qquad (16)$$

where mass and radius are in solar units. Along the main sequence, the density decreases and the temperature and pressure increase with the star's mass.

Sketching the evolution of solar-type stars

Stars in the mass range $(0.7–1.5)M_\odot$ are those we see today in old globular clusters and old open clusters and therefore they are of primary interest in interpreting the HR-diagram. Table 2 contains some data for the various phases we are going to describe below, limited to the case of low-mass stars such as the Sun. Representative post-main sequence paths in the HR-diagram are shown in the upper panel of figure 3. The paths illustrated are for low-metal stars with abundances more like those in globular clusters. In broad terms their evolution is similar to that of stars of solar-like compositions; in detail, they differ.

Protostars

STAR FORMATION is by far the most uncertain aspect of stellar theory. Based on observations, we know that stars form in dense, cold molecular clouds, in which various kinds of instability combine to trigger fragmentation and collapse of gas into small entities of stellar size (protostars). This stage is reached when trapping of energy by increasing opacity supplies pressure support against gravity and the condition of hydrostatic equilibrium is established. During this very short-lived phase the sole source of energy is gravitational contraction. The star is cool and red but very extended in radius (large luminosity). To avoid confusion with the later stages of evolution, neither this nor the next, pre-main sequence phase is illustrated in figure 3.

Pre-main-sequence

Once hydrostatic equilibrium is established, the contraction slows down but it continues until the central temperature is high enough to initiate nuclear burning. The star, fully convective, slides down the Hayashi line, i.e. the locus in the HR-diagram of fully convective objects in hydrostatic equilibrium. No star cooler than this limit can exist. On the Hayashi line and prior to main sequence stage, light elements such as ^{2}H, ^{7}Li, etc are all burned. T-TAURI STARS located at the right of the main sequence are associated with the pre-main sequence phase. Their phenomenology is quite complicated, because there is simultaneous infall of material onto the star (disk accretion) but also strong ejection of material in the form of winds (bi-polar flows). It takes another several million years before the star settles down on the main sequence where quiet H burning takes place.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Evolution of a $1M_\odot$ star with $X = 0.698$, $Y = 0.282$ and $Z = 0.019$.

Phase	log (age)	$\log L/L_\odot$	$\log T_e$	$R/R_\odot$	$\log T_c$	$\log \rho_c$	M_c/M	M_e/M	Energy
A1		−4.00	1.0	2×10^6	1.0	−18.0	0	1.00	E_G
A2	5.92	−0.20	3.64	2.1	6.60	0.17	0	1.00	E_G
B1	6.94	−0.30	3.64	1.6	6.77	1.04	0	1.00	E_G
B2	7.39	0.04	3.77	1.0	7.11	1.92	0	0.30	E_N
C1	7.54	−0.16	3.75	0.9	7.14	1.99	0	0.06	PP
C2	9.98	0.25	3.76	0.9	7.28	2.89	0	0.74	H shell
D	10.05	0.33	3.72	1.8	7.38	4.06	0	0.62	H shell
E	10.08	3.37	3.49	170.6	7.89	5.94	0	0.46	H shell
F	10.08	1.66	3.66	10.9	8.10	4.24	0.21	0.50	He burning
G	10.08	3.32	3.50	153.8	8.02	6.16	0	0.52	H, He shell
H	>10.08	−2.60	4.0	0.02	6.60	7.44	0	0	Thermal energy

A1, collapse; A2, protostar in hydrostatic equilibrium; B1, pre-MS (minimum luminosity on the Hayashi line); B2, pre-MS (maximum luminosity before the MS); C1, start of the MS phase; C2, end of the MS phase; D, post-MS (bottom RGB); E, tip of RGB and He flash; F, core He-Burning and HB (mid-stage); G, AGB phase; H, WD. Age in years, temperature in kelvins, density in g cm^{-3}, radius in solar units.

Main sequence

During this phase H burning produces enough energy to balance the luminosity (the star is said to be in thermal equilibrium). The nuclear production is a strong function of temperature (PP chain versus CNO cycle, the first one in the case of the Sun). Gravity is balanced by differential gas pressure (the ideal gas law in this case). The H content in the center decreases until it eventually reaches zero. Heat is carried out by radiative transport.

Post-main-sequence

When the H content becomes zero in a small central sphere, the tiny He core contracts and H burning begins in a shell surrounding the core. Gravitational contraction in the core makes it electron degenerate. The star leaves the main sequence and moves toward low T_e at nearly constant L. This phase corresponds to the sub-giant branch in the HR-diagram. The external convection penetrates very deeply into the star. The energy generated by the H-burning shell and the increased opacity force the star to expand to large radii and once the Hayashi line is reached to climb along it at nearly constant T_e but increasing luminosity. The star is in the red giant branch phase.

Red giant branch and He flash

The outward migration of the H-burning shell makes the He core grow in mass and it slowly increases temperature until the threshold limit for He burning is reached. Since the core is electron degenerate, (i) this occurs when the core has acquired a certain value of the mass and (ii) He ignition is explosive but does not disrupt the star (He flash). Core degeneracy is removed and the star begins to burn helium relatively quietly in a convective core. The mass of the He core at the He ignition stage is nearly independent of the initial mass and chemical composition of the star. The He core mass at the He flash stage is $M_{\rm Hefl} \simeq 0.5M_\odot$.

Core He burning and horizontal branch

During the brightest stages along the red giant branch mass loss by stellar wind occurs; this removes part of the envelope's mass. According to current estimates of the efficiency of mass loss and theoretical lifetimes during the brightest stages of the red giant branch, little mass is lost by stars more massive than about $1M_\odot$, whereas about 20% of the mass is lost by stars of about $0.8M_\odot$. While burning helium in the core, the star settles on the horizontal branch. The luminosity is now much lower than that at the tip of the red giant branch and T_e is higher than the corresponding value on the Hayashi line. The range in T_e covered by HORIZONTAL BRANCH STARS depends on the amount of mass ($\Delta M_{\rm RGB}$) lost at the tip of the red giant branch and several structural details that cannot be described here. Suffice it to recall that at given $\Delta M_{\rm RGB}$ and initial mass, T_e of horizontal branch models is higher for decreasing metallicity and increasing age. Since core H ignition occurs in He cores of nearly the same mass, the horizontal branch luminosity is nearly constant and the duration of central He burning is also nearly constant (a few 10^8 yr).

Asymptotic giant branch phase and planetary nebula ejection

When helium is exhausted, the C–O core contracts again and becomes strongly electron degenerate. In order to ignite carbon in such conditions, the core mass has to grow to about $1.4M_\odot$, which is not possible because the total mass of these stars is smaller than this. The star continues to evolve thanks to the He- and H-burning shells that force it to expand and to settle again on the Hayashi line along which it moves, with the luminosity increasing and mass being lost by the stellar wind. The novelty with this phase is that the He-burning shell is thermally unstable due to the high temperature dependence of the 3α reactions and its very small thickness (the so-called Schwarzschild-Harm instability). The He shell goes through a series of paroxysmal activity alternated with periods of quiescence, which causes periodic expansions and contractions of the overlying envelope. This likely enhances the efficiency of stellar winds till the whole envelope is lost. In total an additional 10% of the star mass is stripped away, revealing the hot inner regions with temperatures in excess of about 10^5 K. The overall duration of the asymptotic giant branch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

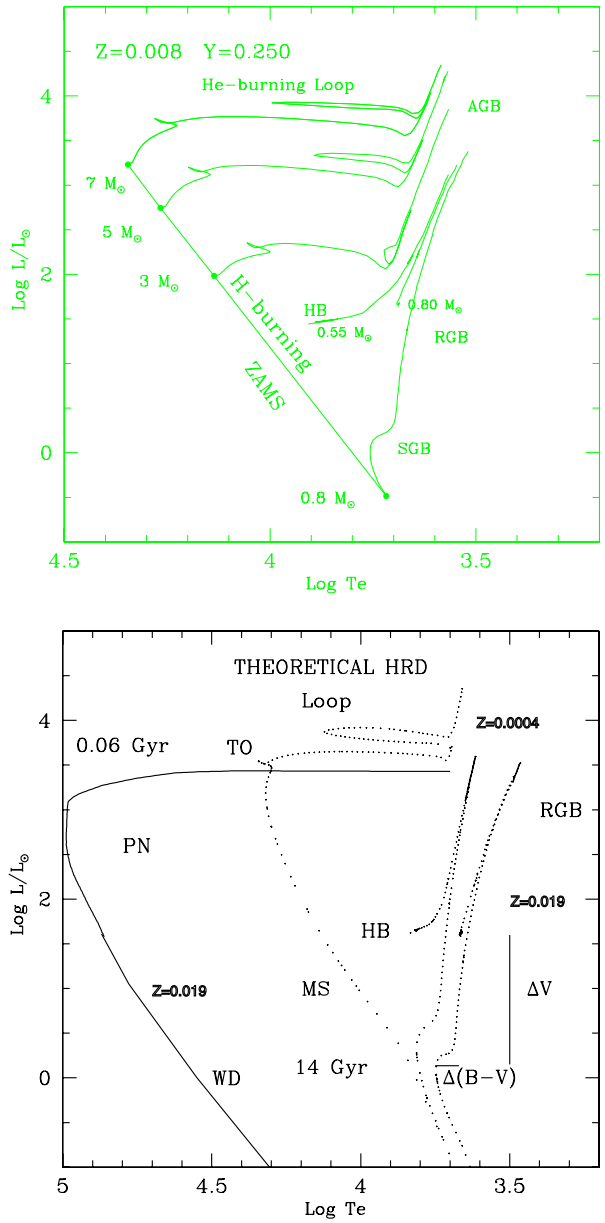


Figure 3. Upper panel: evolutionary tracks of different mass. The composition is $Y = 0.250$ and $Z = 0.008$. Lower panel: Isochrones of different age and metallicity Z as indicated. Note that a large range of T_e is covered in this panel.

phase is only a few 10^6 years. The newly expelled gas illuminated by the very hot central objects interacts with the surrounding material lost in previous stages. This is the origin of the so-called PLANETARY NEBULAE (PN), which takes about 10^4 years to disperse into the interstellar medium.

White dwarfs

As the planetary nebula disperses, the He- and H-burning shells surrounding the inert C–O core are extinguished, leaving a WHITE DWARF. This remnant is supported by electron degeneracy. Gravitational contraction is almost ineffective as a source of heat and therefore no nuclear fuel (C and/or O) can be ignited. The internal temperature is a few 10^8 K. The radius of the star is small, only a few thousand km (the radius of the Earth). The luminosity rapidly decreases to small values. The source of energy is mainly the thermal energy stored in the ions (the kinetic energy of degenerate electrons cannot be tapped). After an initial phase of rapid cooling, the white dwarf will take a long time to radiate stored energy away. As a consequence of the strong degeneracy, the white dwarf has a fixed radius determined by its mass. Decreasing its luminosity by losing energy, the white dwarf will move towards lower T_e, sliding down its line of constant radius (see the lower panel of figure 3.

A counterintuitive rule

In general, exhaustion of central fuel tends to make a star brighter and cooler (shift in the HR-diagram to the right and upward), whereas ignition of a new central fuel tends to make it fainter and hotter (shift in the HR-diagram to the left and downward).

What about other stars?

We deliberately omit here objects with masses comparable with those of giant planets ($M < 0.01M_\odot$) and brown dwarfs ($0.01M_\odot < M < 0.08M_\odot$) that never become hot enough in their cores to ignite the PP chain. In contrast, stars more massive than $0.08M_\odot$ all ignite core H burning on the main sequence.

$0.08M_\odot \leq M \leq 0.5M_\odot$

These stars are very long lived, lasting several hundred billions of years. Those with mass $\leq 0.3\,M_\odot$ are also fully convective. In any case they will never become hot enough to ignite helium in the core. If no other cause intervenes they will become helium white dwarfs.

$0.5M_\odot \leq M \leq 2.2M_\odot$

Their evolution is that thoroughly described in the previous section so that no other details are given here, but for mentioning the transition from PP chain to CNO cycle, which occurs at about $1.3M_\odot$ and is accompanied by the transition from radiative to convective core H burning.

$2.2M_\odot \leq M \leq 8M_\odot$

The major novelty here is that central He ignition takes place in a non-degenerate core. Therefore, there is no red giant branch to climb and no He flash. After core H exhaustion the stars go back to the Hayashi line, ignite He burning in the core, spend some time there and then go back towards high T_e (going around a loop) to settle on the core He-burning band. This is nearly parallel to but brighter and cooler than the MS band (see the isochrone of 0.06 Gyr, whose turn-off mass is about $7M_\odot$, shown in the lower panel of figure 3). Following exhaustion of central helium, the stars, now made of a C–O core surrounded

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by He- and H-burning shells, expand once more to the Hayashi line, experiencing the asymptotic giant branch phase on the way. Stars with mass up to about $6M_\odot$ will become C–O white dwarfs after the planetary nebula phase. Stars more massive than $6M_\odot$ may reach the stage of C ignition in the extremely degenerate C–O core while in the asymptotic giant branch phase. The consequent thermonuclear runaway will disrupt the whole star in a SUPERNOVA EXPLOSION. No remnant is left.

$8\,M_\odot \leq M \leq 100M_\odot$

This is the range of massive stars that are able to proceed through the whole series of nuclear burning stages up to the formation of the iron core. While the evolution of their cores is relatively straightforward, the path in the HR-diagram is more uncertain because these stars are heavily affected by mass loss by stellar wind. In brief, stars less massive than about (30–40)$M_\odot$ evolve in the HR-diagram in a way quite similar to that of stars of lower mass ($\simeq 7M_\odot$) as described in the previous section. These stars should eventually explode as type II supernovae while shining as red super-giants. The situation is uncertain for more massive objects for which all evolutionary phases seem to occur at high T_e, the whole H-rich envelope being lost by mass loss. They probably explode as another type of supernova. In any case the final fate of all massive stars is core collapse triggered by photodissociation of iron, electron capture and neutronization of matter in the central core. The core collapse induces a supernova explosion which leaves a compact remnant: either a NEUTRON STAR of $1.4M_\odot$ or a BLACK HOLE of larger mass depending on details of the collapsing core's physical structure.

Tracks or isochrones?

In cluster HR-diagrams, stars lie on isochrones (loci of constant age but varying mass). How large is the mass range $\Delta M_{\rm iso}$ spanned by stars in all post-main sequence phases of an isochrone or, equivalently, how much does an isochrone depart from the evolutionary sequence with the turn-off mass $M_{\rm TO}$? The answer lies in relation (8) and the short duration of all post-main squence phases, a time of the order of $0.1t_{\rm H}$ or so. Therefore, for any practical purpose the total lifetime can be approximated by $t = t_{\rm H} + t_{\rm He} = (1 + 0.1)t_{\rm H}$. Differentiating this relation and performing a little algebra, we obtain the approximate relation (to be used only for illustrative purposes)

$$\Delta M_{\rm iso} \simeq 0.045 M_{\rm TO}. \tag{17}$$

Except for massive stars, say above $10M_\odot$ (where $\Delta M_{\rm iso}$ is significant), this mass spread is small. Therefore, over ample ranges of $M_{\rm TO}$ and hence ages, isochrones are almost identical to the evolutionary track of the star just leaving the main sequence: compare the upper and lower panels of figure 3. In any case, accurate age determinations are ultimately possible only by using detailed isochrones.

Interpreting the HR-diagram of star clusters with single stellar populations

STAR CLUSTERS are assemblies of stars for which one can reasonably assume the same initial composition $[X, Y, Z]$, same age (within a narrow range), same distance and finally the same initial mass function (IMF). Therefore, they constitute useful examples of a single STELLAR POPULATION whose properties bear very much on understanding the HR-diagram of any type of stellar population from clusters to galaxies, the latter conceived as convolutions of many single stellar populations.

Single stellar populations are governed by the so-called fuel consumption theorem: the contribution of stars in any post-main sequence phase to the integrated luminosity of a single stellar population is directly proportional to the amount of fuel burnt during that stage. In other words the fuel consumption theorem explains how and why stars distribute themselves along a given isochrone with number frequencies that vary with the position (evolutionary phase). In other words it explains why clusters of different age and chemical composition have different HR-diagrams.

The turn-off mass $M_{\rm TO}$ of a single stellar population with age t is defined as the mass of the star at the stage of central H exhaustion. Stars with mass $M_i < M_{\rm TO}$ are still on the main sequence, whereas those with $M_i > M_{\rm TO}$ are in later evolutionary stages. In other words, the turn-off is the stage at which stars leave the main sequence, venture towards later stages and eventually die. The rate at which this occurs is given by

$$b(t) = \Phi(M_{\rm TO})|\dot{M}_{\rm TO}| \text{ stars yr}^{-1} \tag{18}$$

where $\Phi(M_i) = AM_i^{-1+x}$ is the IMF and $\dot{M}_{\rm TO}$ is rate at which the turn-off mass decreases at increasing age. Therefore, $b(t)$ is the evolutionary flux of the single stellar population and does not depend on the particular evolutionary stage. It is the speed at which stars enter and leave any evolutionary phase. In other words, the number of stars N_j in the jth post-main sequence phase is

$$N_j = b(t)t_j \tag{19}$$

where t_j is the lifetime of the phase.

The integrated bolometric luminosity $L_{\rm T}$ can be separated into the contribution from main sequence stars and from all the others:

$$L_{\rm T} = \sum_i^{\rm MS} L_i + \sum_j^{\rm PMS} N_j \langle L \rangle_j \tag{20}$$

where L_i is the luminosity of single MS stars and $\langle L \rangle_j$ is the mean luminosity of the post-main sequence jth phase. Denoting by M_i the lower mass limit of the IMF, and using the mass–luminosity relation $L(M) = \beta M^\alpha$, the contribution from main sequence stars becomes

$$\sum_i^{\rm MS} L_i = \int_{M_i}^{M_{\rm TO}} L(M)\Phi(M)\,{\rm d}M \sim \frac{A\beta}{\alpha - x}(M_{\rm TO}^{\alpha-x} - M_i^{\alpha-x})$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\sim \frac{A\beta}{\alpha - x} M_{\rm TO}^{\alpha - x} \tag{21}$$

and that of post-main sequence stars

$$L_{\rm PMS} = \sum_{j}^{\rm PMS} b(t) t_j \langle L \rangle_j = b(t) \sum_{j}^{\rm PMS} F_j \tag{22}$$

where $F_j = t_j \langle L \rangle_j$ is the nuclear consumption during phase j, after which the theorem is named. $F_j = m_j^{\rm H} + 0.1 m_j^{\rm He}$ where $m_j^{\rm H}$ and $m_j^{\rm He}$ are the amounts of hydrogen and helium burnt (in solar units). It follows from the above equations that the contribution to $L_{\rm T}$ by post-main sequence stars increases with the number of stars in the phase and its duration.

Given a cluster with age t, total bolometric luminosity $L_{\rm T}(t)$ (in solar units) and total number of stars $N_{\rm T}$, the theory of single stellar populations predicts–explains the number of stars expected–observed in each post-main sequence evolutionary phase. Indeed relation (17) can be rewritten as

$$N_j = \frac{b(t)}{L_{\rm T}(t)} L_{\rm T}(t)\, t_j = B(t) L_{\rm T}(t) t_j. \tag{23}$$

It can be shown that $B(t)$ is nearly independent of the IMF and constant over a large range of ages. It goes from 5×10^{-10} to 2×10^{-11} for ages lying between 10^7 and 10^{10} yr. To fully appreciate the potential of this result let us consider the asymptotic giant branch stars, for which theory predicts a lifetime of a few 10^6 yr (say 5×10^6 yr in round numbers). It follows that $N_j \simeq 10 \times 10^{-5} L_{\rm T}(t)$, which means that 10 asymptotic giant branch stars are expected in a cluster with $L_{\rm T} = 10^5\, L_\odot$, in agreement with real clusters. If the total luminosity is a factor of 10 less, almost no asymptotic giant branch stars are expected to be seen. The asymptotic giant branch phase is still there, but the corresponding stars could be easily missed for statistical reasons.

Age of star clusters

Globular clusters

Determining the age of globular clusters is a complex game which requires a knowledge of many parameters, such as the helium content Y, metallicity [M/H], CNO abundance [CNO/H], distance modulus and reddening.

Abundance of helium

Direct spectroscopic measurements of helium abundance are not possible because the stars of globular clusters are too cool. The abundance is derived from several indirect methods which yield the value $Y = 0.235 \pm 0.005$.

Metallicity

By this quantity one usually refers solely to the iron content expressed in the standard notation $[{\rm Fe/H}] = \log(N_{\rm Fe}/N_{\rm H}) - \log[(N_{\rm Fe}/N_{\rm H})_\odot]$, where N_x denotes the number density of atoms of any species. The majority of globular clusters have [Fe/H] ranging from -1.0 to -2.3 dex with a typical uncertainty of 0.15 dex. The observation of HR-diagrams showing sequences of virtually undetectable width indicates a uniform abundance of heavy elements within the stars of a particular cluster.

Abundances of α-elements

An important question that has arisen in recent years is whether the abundances of elements such as C, N, O, Mg with respect to Fe have always been like those observed in the Sun, i.e. $[{\rm X}_i/{\rm Fe}] = 0$ in the standard notation, or whether instead they had supersolar values $[{\rm X}_i/{\rm Fe}] > 0$ for $[{\rm Fe/H}] \leq -1$. Current observations yield $[{\rm X}_i/{\rm Fe}]$ from 0.3 to 1.0 for low metallicities. The overabundance is different for different elements. We will refer to it as the [C,N, O, Mg/Fe] versus [Fe/H] relation.

Reddening

Since there are many globular clusters with low color excess ($E_{B-V} \leq 0.1$) spanning a broad range of metallicities (up to $[{\rm Fe/H}] = -1$), reddening is not a serious problem in finding the intrinsic color of the turn-off.

Distances

The distance scale of globular clusters is perhaps the topic of major controversy. There are three ways to derive globular cluster distances.

(1) *Subdwarfs*. The nearby subdwarfs (metal-poor stars with halo kinematics located on a sequence parallel to but fainter than the main sequence) close enough to have measurable parallaxes allow us (i) to derive absolute M_V so that accurate testing of theoretical zero-age main sequences at varying [Fe/H] in the low-metallicity regime is possible; and (ii) to tie the relative cluster distances to the local distance scale, under the assumption that these stars are similar to unevolved stars in globular clusters.

(2) *RR Lyrae.* Modern determinations of the distance modulus reduce to comparing the apparent magnitudes of the RR LYRAE STARS or equivalently horizontal branch stars with the corresponding absolute visual magnitudes. There are several independent methods to obtain the absolute visual magnitudes of RR Lyrae stars, M_V(RR), which ultimately lead one to assess whether or not a correlation between M_V(RR) and [Fe/H] exists and try to fix the slope and zero point of this relation. The zero point is of critical importance for determining the age of the oldest clusters, whereas the slope $\Delta M_V({\rm RR})/\Delta[{\rm Fe/H}]$ bears very much on the age–metallicity relation, i.e. whether or not the age of globular clusters becomes younger at increasing [Fe/H].

(3) *White dwarfs.* This method rests on fitting the theoretical white dwarf-cooling sequences to the observational position of white dwarfs in the HR-diagram.

The second parameter

The HR-diagram of globular clusters can be ranked as a function of the metallicity [Fe/H]. Generally, metal-rich globular clusters have cool RGBs and red horizontal branches, while the opposite is true in metal-poor globular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

clusters, in which very extended, blue horizontal branches are observed. This rule is, however, broken by a number of globular clusters whose metallicities are identical (as far as we can tell) but whose horizontal branches are different. As a matter of fact, we have metal-poor globular clusters with red horizontal branches. This opens the gate to the so-called second parameter problem: in addition to metallicity another important parameter helps to determine the morphology of HR-diagrams. Several second parameters have been suggested, among which are the helium content, the [C, N, O, Mg/Fe] versus [Fe/H] relation, the age or finally a subtle combination of all of them.

Methods for determining the age

An essential prerequisite is that HR-diagrams of extremely good photometric quality with precision down to a tenth of a magnitude are available. If so, ages can be derived from the following.

(1) In the isochrone fitting method all the above ingredients are simultaneously necessary. These ages are therefore subject to the highest uncertainty.

(2) The ΔV method rests on the fact that the turn-off magnitude becomes fainter as a cluster ages, while the horizontal branch luminosity is virtually constant. ΔV is the magnitude difference between the turn-off and the horizontal branch at the turn-off color. The method is independent of reddening and to some extent of distance, but it requires an assumption for the helium abundance. The disadvantage is that not all globular clusters possess RR Lyrae stars, and some horizontal branches are not horizontal. Furthermore, the turn-off is often almost vertical, thus making the identification of the turn-off magnitude very uncertain. Finally, there is the effect of the controversial relations $M_V(\mathrm{RR}) - [\mathrm{Fe/H}]$ and $[\mathrm{C, N, O, Mg/Fe}] - [\mathrm{Fe/H}]$.

(3) The $\Delta(B - V)$ method is based on the color difference between the turn-off and the base of the red giant branch. This color difference decreases as the cluster age increases. Within a certain approximation, the method is independent of distance, reddening, photometric zero point, helium abundance and, to first order, variations in [Fe/H]. Among others, the major uncertainties are with the transformations from T_e to colors, and the [C, N, O, . . . /Fe] ratios, all these affecting the turn-off color.

Ages, age spread, and age–metallicity relation

The absolute ages, the age spread at given metallicity and the mean age–metallicity relation of globular clusters are significant for COSMOLOGY and GALAXY FORMATION. The oldest globular clusters set a firm lower limit on the age of the universe, the age spread and age-metallicity relation constraining the time scale and mechanism of halo formation. Because of this, an impressive body of literature exists on the age subject. Since even a short summary of the vivid debate is not possible here, we limit ourselves to stating that over the years the following alternative scenarios have emerged.

(1) The age of globular clusters peaks at about 16 ($\pm$2) Gyr, with wings going down to 10 Gyr and up to 20 Gyr. The spread in age is large, and the most metal-rich globular clusters are also the youngest objects. It follows that the age is the second parameter and halo formation occurred on a long time scale.

(2) Globular clusters are younger than above. Only the most metal-poor globular clusters (presumably the oldest objects) may have ages near 15 Gyr, and, finally, ages below 12 Gyr and above 20 Gyr appear to be unlikely. Furthermore, there is no substantial body of evidence that the age is the dominant second parameter nor for a significant spread in age among clusters of given metallicity.

(3) Globular clusters are even younger. Their age is about 12$\pm$2 Gyr. This result is ascribed to improvements in the equation of state and other details of the input physics that cannot be discussed here.

(4) Remarkably, most of the ages for globular clusters estimated before the Hipparcos mission fell in the range 14–18 Gyr (with exceptions), whereas most of those after Hipparcos are in the range 10–14 Gyr. The age spread and mean age–metallicity relation are still debated.

(5) Going deeper into the reason for the discordant results, this can always be reduced to knowing the distances of globular clusters. Most of the uncertainty stems from a difference in distance modulus of $\simeq$0.25 mag, which is the typical uncertainty with which the magnitudes of horizontal branch, turn-off, main sequence and white dwarf stars are known (from both theory and observations). Amazingly, an uncertainty of about 0.05 mag would immediately imply an uncertainty of $\simeq$1 Gyr. Although we cannot firmly choose between 10–14 and 14–18 Gyr, the first option is certainly to be preferred.

Old open clusters

Old open clusters, whose ages range from say 1 to 7–8 Gyr, trace most of the history of the Galactic disk. Therefore, the correct ranking as a function of age, chemical composition and kinematic properties, is of paramount importance for understanding the process of star formation in the Galactic disk. Furthermore, old open clusters have turn-off masses between $1M_\odot$ and $2M_\odot$. Therefore, they probe the stellar structure in that mass range, in which during the main sequence phase the transitions from radiative to convective cores, i.e. from PP chain to CNO cycle, occur. Later they see the transition from bright red giant branches as in M67 to much less evident red giant branches as in the Hyades. The techniques used to interpret the HR-diagrams of open clusters are similar to those of GCs. However, the much smaller number of stars present in each open cluster and their sparse spatial distribution make the identification of the main sequence, red giant branch, He-burning band—here called the red clump to distinguish it from the classical horizontal branches—much more difficult. The small total number of stars and the fuel consumption theorem cause the red giant branch, red clump, and asymptotic giant branch to be populated by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

only by a handful of stars (if any at all). The contamination by foreground objects blurs the main sequence and turn-off. In reality the whole HR-diagram is often immersed in a bath of contaminating stars. Despite these uncertainties, ages and the age–metallicity relation are derived. The ages go from 0.9 Gyr for NGC 2477 to 8 Gyr for NGC 6791. As far as the age–metallicity relation is concerned, it runs at nearly constant metallicity close to the solar value.

Young, rich clusters of the Magellanic Clouds
Young clusters with physical properties (luminosity, total mass, shape, etc...) comparable with those of globular clusters are no longer formed in our own Galaxy so that testing the theory of STELLAR EVOLUTION with the same accuracy as for globular clusters is not feasible. In the Milky Way, young clusters such as the Hyades, h+χ Persei and many young associations contain so few stars that only the main sequence is visible. One is forced to combine HR-diagrams for a chosen age class, but spurious inhomogeneities due to different chemical composition are easily introduced. Luckily, the MAGELLANIC CLOUDS (satellites of the Milky Way) are rich in objects of this type. Because of the large number of stars contained in these clusters, it is possible to make meaningful comparisons even for the shortest-lived evolutionary phases. The classical workbench for this kind of study is the cluster NGC 1866 in the LMC. This cluster, whose total mass is estimated to be in the range (3.6–5) $\times$ $10^5 M_\odot$, is well populated throughout the various evolutionary phases, exhibits an extended loop of giant stars and is rich in Cepheids, for which membership is less of a problem. The classical estimates of the turn-off mass and age of NGC 1866 are $5M_\odot$ and 70 $\times$ 10^6 yr, respectively. The existence of Cepheids among the other giant stars of the cluster allows us to check the mutual consistency between properties derived from stellar models and pulsational theory. Corresponding attempts in our own galaxy have always been hampered by the uncertainty in the cluster membership of the Cepheid stars (always the distance problem).

Star formation history of galaxies from their HR-diagrams

Figure 4 shows two HR-diagrams for field stars in different astrophysical environments: (i) the Galactic bulge (upper panel), and (ii) an external galaxy (lower panel). Although they seem to greatly differ from each other, they share many points in common. In HR-diagrams such as these, the interpretation by means of single stellar populations we have used for single clusters, no matter whether old or young, no longer applies. Indeed, these HR-diagrams are the results of many stellar populations spanning ample ranges of ages, chemical compositions and distances, suffering different amounts of extinction and reddening, and maybe having different IMFs. Finally, we remind the reader that owing to the complex nature of these HR-diagrams, in addition to the key parameters characterizing the stellar populations, the number density of stars in each

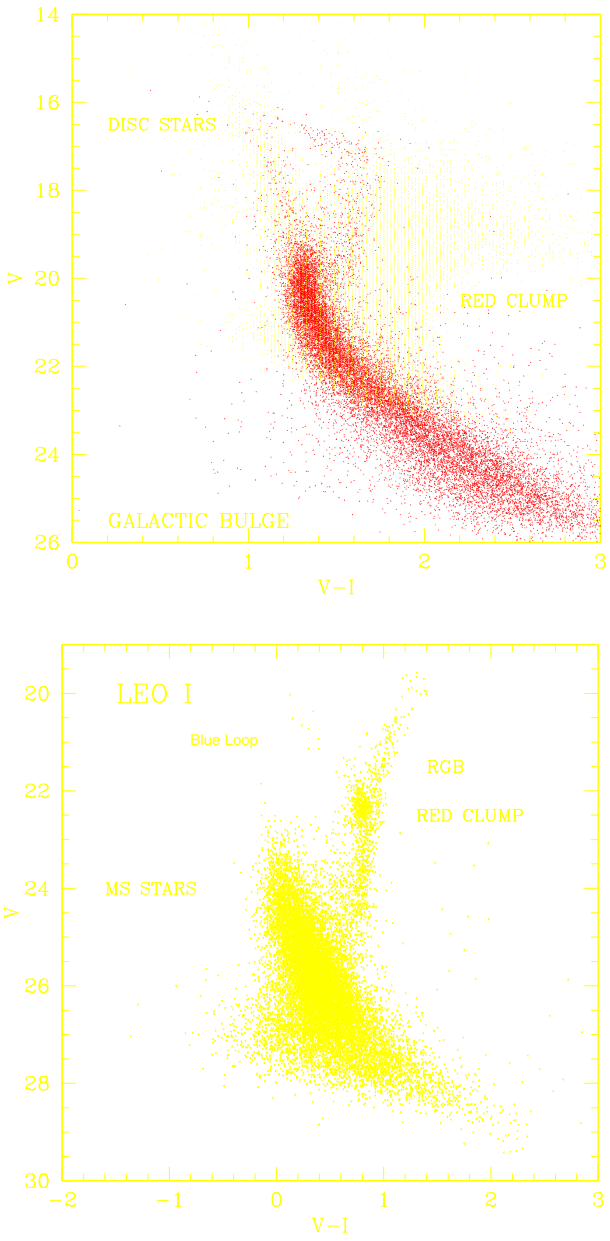


Figure 4. Upper panel: the HR-diagram toward the Galactic center as observed through one of the Baade windows. The dark areas are HST data, whereas the gray areas are ground-based data. Lower panel: the HR-diagram of the dwarf spheroidal galaxy Leo I in the Local Group.

subarea is related to the past history of star formation and structural properties of the host galaxy. Because of this, much interest has recently been given to deciphering the past star formation and chemical enrichment and structure of the population generating such complex HR-diagrams. Various sophisticated techniques aimed at unraveling the parameters in question have been designed. In the following we will briefly analyze the two HR-diagrams

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

shown.

Galactic bulge

This is the HR-diagram of stars observed towards the Galactic center through one of the Baade windows. In this HR-diagram there are stars coming from different regions along the line of sight: the stars in the bulge itself (the intrinsic depth and extinction of the bulge combine to blur their magnitudes and colors) and the stars belonging to different zones of the disk, each of these at a different distance, affected by a different degree of extinction and containing a different total mass in stars. The only easy thing to say is that main sequence stars of the disk are responsible for the blue plume along the main sequence band, whereas disentangling all the others requires an iterative comparison between simulations of the HR-diagram based on a model mass distribution, star formation rate, and chemical enrichment in both the bulge and the disk, the extinction law along the line of sight and the observed HR-diagram. Neglecting all details of the techniques used, the study of a number of such HR-diagrams has led to the following provisional picture about the various components.

(i) A two-component system (disk and spheroid) is sufficient, with no compelling evidence for the so-called thick disc.

(ii) The majority of stars belong to the disk. They are rather metal rich ($0.008 \leq Z \leq 0.02$) and relatively old (7–4.5 Gyr). The star formation rate exponentially declined with a time scale of about 2.5 Gyr. The Disc has a mass distribution with vertical scale height $z = 230$–250 pc and radial exponential profile with scale length $h_0 = 3.5$–4.5 kpc.

(iii) There is another much younger and thinner component ($z_0 = 100$ pc; $h_0 = 4.5$ kpc) with $Z = 0.015$–0.020, age from 5.0 to 2.0 Gyr and constant star formation rate.

(iv) The spheroid seems to contain two populations: a very old, metal-poor halo component ($Z = 0.0004$–0.005), age of 16–10 Gyr), with a star formation rate exponentially decreasing over a time scale of about 3.0 Gyr, and a younger metal-rich bulge component with $Z = 0.005$–0.02, an age of 10–5.0 Gyr and star formation rate exponentially decreasing with time scale of 2.5 Gyr.

(v) Ratios of the components are halo: intermediate disk = 1:1, bulge: intermediate disk = 7:9, young disk:intermediate disc = 1:10.

(vi) There seems to be a sharp increase in the number star density, a sort of stellar ring, centered at about 3.5 kpc from the galactic center with $\sigma = 0.5$ kpc.

(vii)The mass distributions along three lines toward the bulge are not the same. Does this suggest that the bulge has a bar?

(viii) Finally, the bulge, in addition to the main component of old stars, also seems to contain younger stars (age of about 8–9 Gyr and $Z = 0.005$–0.03), perhaps associated with formation of the bar.

External galaxy: Leo I

This is the HR-diagram of the LOCAL GROUP dwarf spheroidal (dSph) galaxy Leo I. In this case the distance and extinction are less of a problem as they are the same for all stars. Details of the galaxy's structure are not considered here, because the data refer to the galaxy as a whole. Leo I is a metal-poor galaxy (estimated [Fe/H] $= -2$) at a distance of about 300 kpc. Its HR-diagram shows very broad main sequence–turn-off and subgiant regions with no evidence of discontinuities, a compact broad red clump, a large red giant branch, an inconspicuous horizontal branch if any at all and a V-shaped structure above the red clump. The majority of stars in the red arm of the V-shaped structure are red giant branch objects (low mass) with a smaller fraction made of more massive core He-burning objects in the red stages of the loop. The blue arm is stars of the same evolutionary phase but on the blue side of the loop. All these features together suggest that star formation in Leo I extended from 10–15 Gyr ago to less than 1 Gyr ago and proceeded in a rather smooth fashion. The lack of a prominent horizontal branch and the low metallicity suggest that very little star formation took place earlier than $\simeq$10 Gyr ago.

Field stars have been studied in other galaxies of the Local Group going from the Magellanic Clouds, M31 and M32 to many dwarf irregulars (dIrr) and dSph, and their star formation and chemical enrichment histories have been derived. Several interesting preliminary conclusions are possible. (i) No two galaxies of the Local Group have the same star formation history. (ii) The LMC and SMC had dominant episodes of star formation starting only several Gyr ago. An old component is, however, present. (iii) Many dIrr have old populations (NGC 6822). (iv) No galaxy is composed exclusively of stars older than 10 Gyr. (v) Some galaxies have only a small fraction of stars older than 10 Gyr (Leo I). (vi) Some galaxies had nearly constant star formation rate up to the recent past followed by strong activity (Sextans A). (vii) Others had recurrent episodes of varying intensity separated by periods of quiescence (Carina). (viii) The second parameter problem is common throughout the Local Group. Many low-metallicity galaxies have red horizontal branches. This suggests that their oldest stars are indeed younger than the oldest globular clusters.

> Motion in the HR-diagram is the outward and visible sign of an inward nuclear turmoil
>
> John Faulkner to his students

Bibliography

The literature on the various subjects related to the HR-diagram being as large as now it is, the following cannot be comprehensive at all. However, we indicate some basic textbooks and some surveys of research in the field that may guide the reader to extend his/her interest.

Useful textbooks on stellar structure and evolution are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Clayton D D 1983 *Principles of Stellar Evolution and Nucleosynthesis* (Chicago, IL: University of Chicago Press)
Cox J P and Giuli R T 1968 *Principles of Stellar Structure* vol 1, 2 (New York: Gordon and Breach)
Huang R Q and Yu K N 1998 *Stellar Astrophysics* (Singapore: Springer)
Kippenhahn R and Weigert A 1990 *Stellar Structure and Evolution (Astronomy and Astrophysics Library)* (Berlin: Springer)
Phillips A C 1994 *The Physics of the Stars* (London: Wiley)
Schwarzschild M 1958 *Structure and Evolution of the Stars* (New York: Dover)

Review and research articles on stellar structure, evolution and cluster ages are

Chaboyer B, Demarque P, Kernan P J and Krauss L M 1998 *Astrophys. J.* **494** 96
Chiosi C, Bertelli G and Bressan A 1992 *Ann. Rev. Astron. Astrophys.* **30** 235
Demarque P 1996 *The Extra-galactic Distance Scale* ed M Livio (Cambridge: Cambridge University Press) p 30
Fusi-Pecci F and Cacciari C 1991 *New Windows to the Universe*, ed F Sanchez and M Vasquez (Cambridge: Cambridge University Press) p 364
Iben I Jr 1974 *Ann. Rev. Astron. Astrophys.* **12** 215
Iben I Jr and Renzini A 1983 *Ann. Rev. Astron. Astrophys.* **21** 271
Iben I Jr and Renzini A 1984 *Phys. Rep.* **105(6)** 329
Renzini A and Fusi-Pecci F 1988 *Ann. Rev. Astron. Astrophys.* **26** 199
Salaris M and Weiss A 1998 *Astron. Astrophys.* **335** 943
Salaris M, Degl'Innocenti S and Weiss A 1997 *Astrophys. J.* **479** 665
Sarajedini A, Chaboyer B and Demarque P 1997 *Publ. Astron. Soc. Pac.* **109** 1321
Vandenberg D A 1998 *The Interaction between Observation and Theory*, ed T R Bedding, A J Booth and J Davis (Dordrecht: Kluwer) p 439

The fuel consumption theorem is by

Renzini A and Buzzoni A 1986 *Spectral Evolution of Galaxies* ed C Chiosi and A Renzini (Dordrecht: Reidel) p 135

Review and research articles on dwarf galaxies and the Galactic bulge are

Mateo M 1998 *Ann. Rev. Astron. Astrophys.* **36** 435
Ng Y K, Bertelli G, Chiosi C and Bressan A 1996 *Astron. Astrophys.* **310** 771
Wyse R F G, Gilmore G and Franx M 1996 *Ann. Rev. Astron. Astrophys.* **35** 637

Cesare Chiosi

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herzberg Institute of Astrophysics

The Herzberg Institute of Astrophysics (HIA) is the Institute within the NATIONAL RESEARCH COUNCIL of Canada responsible for providing astronomical facilities, and developing related instrumentation and software for Canadian researchers. The Institute was established in 1975, and now operates 1.8 m and 1.2 m optical telescopes at the DOMINION ASTROPHYSICAL OBSERVATORY close to Victoria, BC, as well as a 26 m radio telescope, a seven-antenna synthesis telescope and two telescopes dedicated to monitoring the solar radio flux at the DOMINION RADIO ASTROPHYSICAL OBSERVATORY near Penticton, BC. HIA also supports NRC's international partnerships in the 3.6 m optical infrared CANADA–FRANCE–HAWAII TELESCOPE and the 15 m submillimeter JAMES CLERK MAXWELL TELESCOPE on Mauna Kea, Hawaii, and the Gemini twin 8 m optical infrared telescopes on Mauna Kea and on Cerro Pachon in Chile. The Institute has about 100 staff.

For further information see
http://www.hia.nrc.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Herzberg, Gerhard (1904–)

Molecular spectroscopist, born in Hamburg, Germany, winner of the Nobel prize for chemistry (1971) 'for his contributions to the knowledge of electronic structure and geometry of molecules, particularly free radicals', specifically methylene, a molecule having a carbon with a hydrogen on either side, since identified in nebulae and other interstellar objects. Herzberg's first application for a job as an astronomer, to Germany's Hamburg Observatory, was turned down with the advice that 'there is no point in thinking of a career in astronomy unless one has private means of support', and he turned to engineering, and atomic and molecular physics. At Göttingen, Herzberg worked under Max Born and James Franck, and at Bristol, on the quantum mechanics of molecules, work that he continued in Darmstadt. In 1934, he was dismissed from his job there because his wife was Jewish, and emigrated to Saskatoon in Canada. After a brief period at the University of Chicago and the Yerkes Observatory, he returned to Canada to establish a laboratory for fundamental research in spectroscopy, now known as the Herzberg Institute. He identified the water in Comet Kohoutek, and in 1980 discovered triatomic hydrogen.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hesiod (eighth century BC)

Greek epic poet who described his cosmology in *Theogeny*, the birth of the gods, which he derived from Babylonian mythology. In his family tree of the gods, Chaos (the yawning gap between Earth and Heaven) produced Erebus (the Abyss) and Night, who produced Ether and Day. At the same time, Earth produced the Heavens, Mountains, and Sea. The Heavens and Earth then produced Cronos. This family tree was mimicked by his structural model of the universe, divided into Heaven, Ether, Chaos, Earth, Erebus, Tartarus and surrounded by the Ocean. The cosmology is completely pre-scientific but Hesiod's is the earliest surviving reference to the Pleiades.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hess, Harry Hammond (1906–69)

Geophysicist, born in New York City, professor of geology at Princeton, led Project Mohole, the first expedition to drill through the Earth's oceanic crust to the mantle beneath, theorized that spreading of mid-ocean ridges was the source of new mantle-derived continental material. Also a lunar geologist.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hess, Victor Franz (1883–1964)

Physicist, born in Waldstein Castle, near Peggau in Steiermark, Austria, Nobel prizewinner in physics (1936, shared with CARL ANDERSON) 'for his discovery of cosmic radiation'. In a distinguished career in atomic, nuclear and atmospheric physics, Hess held positions in Vienna, Graz, New Jersey, Washington, DC, and Innsbruck. In 1911–13 he discovered 'ultra-radiation' (cosmic radiation), and later founded the station at the Hafelekar mountain (altitude 2300 m) near Innsbruck for observing and studying cosmic rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hevelius, Johannes [Jan Hewelcke] (1611–87)

German astronomer and instrument-maker, born in Danzig, now Gdansk, Poland. Hevelius was an accomplished instrument-maker and engraver. Influenced by TYCHO BRAHE in subject and instrumentation, he erected in Danzig what must have been then the world's finest observatory Stellaburgum (cf Brahe's Uraniborg), destroyed by fire in 1679. He made several large telescopes and some of the last large, open-sighted instruments, to which he introduced the vernier scale for reading the scales and determining star positions. He wrote a book on instruments, the *Machina Coelestis* (1673) and his widow Elisabeth posthumously published his catalog of star positions, *Prodromus Astronomiae* and his celestial atlas *Firmamentum Sobiescianum* (1687, 1690). The atlas introduced the constellations Canes Venatici, Lacerta, Leo Minor, Lynx and Vulpecula cum Ansere (the fox and the goose, now known simply as Vulpecula), and named Scutum Sobiescianum (now simply Scutum, the shield) after the King of Poland, John Sobieski. Two constellations that Hevelius introduced did not survive—Mons Maenalus and Cerberus. *Selenographia*, his atlas of the Moon, dubbed the gray flat regions as maria but his crater names were superseded by those of RICCIOLI. He discovered four comets, and was one of the first to observe the transit of Mercury.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hewish, Antony (1924–)

Radioastronomer, born in Fowey, Cornwall, Nobel prizewinner for physics (1974) for pioneering research in radio astrophysics, and the discovery of pulsars. After working in war-time radar development, he was educated and worked at Cambridge where he became professor of radioastronomy and led the Mullard Radio Astronomy Observatory from 1982–8. First worked on the scintillation or twinkling of radio 'stars', which he used to investigate the properties of the plasma clouds in the ionosphere. Extended this to identify scintillation due to the solar wind and built a radio telescope to use this technique to obtain very high angular resolution observations of radio sources. By a stroke of good fortune the observational requirements were precisely those needed to detect pulsars and it was with this telescope that Hewish's graduate student, JOCELYN BELL, noticed them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hey, James Stanley (1909–90)

English radioastronomy pioneer, made three fundamental discoveries in radio astronomy. Working on radar during the second World War, he discovered radio interference from solar storms (at first thought to be enemy 'jamming' of radar). It is surprising that he should have found this emission serendipitously, since solar radio emission had been sought without success by several scientists since Hertz discovered radio waves in 1888 (Thomas Edison, Oliver Lodge, J Scheiner and C Nordham). Then in attempting to detect and predict V-2 rockets he saw the radar reflections from meteor trails. Finally, from Richmond Park (a government military research laboratory) he detected Cygnus A as interference behind meteor trail reflections, the first recognized discrete source of radio waves. Became Chief Scientific Officer at the Royal Radar Establishment at Malvern, and constructed one of the two first variable-spacing radio interferometers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hidalgo

An asteroid discovered by Walter Baade in 1920, designated (944) Hidalgo. It was the first known Jupiter-crossing asteroid, having an orbit that takes it from 2.01 AU, near the inner edge of the main belt, to 9.69 AU, just beyond the mean distance of Saturn, in a period of 14.15 years. Hidalgo's mean distance from the Sun is 5.85 AU (875 million km); its diameter is uncertain but is about 40 km. Its highly inclined (42°) and eccentric (0.66) orbit suggests that it may be a large extinct cometary nucleus. It is a D-type object, indicating a dark surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

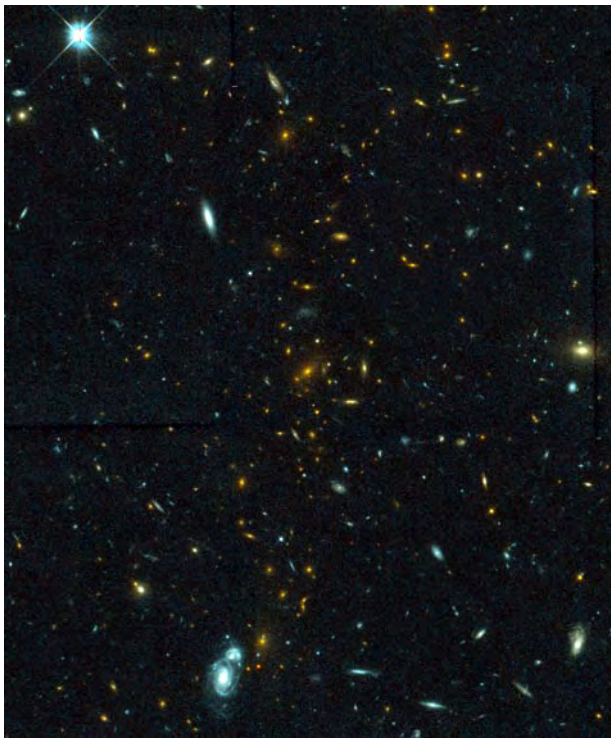


Figure 1. The distant cluster MS1054–03 at a redshift of 0.83. This is the most distant cluster found with the Rosat satellite. The image was taken with the Hubble Space Telescope (courtesy van Dokkum, Franx, ESA and NASA). The cluster is easily recognized by the high density of red galaxies. These galaxies appear red because of the high redshift of the cluster. **This figure is reproduced as Color Plate 7.**

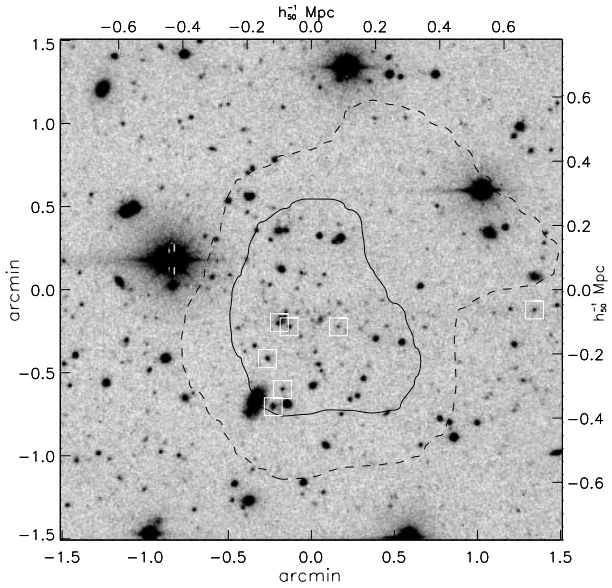


Figure 2. The most distant x-ray cluster RXJ0848.9+4452 at a redshift of 1.26. The contours show the x-ray emission, overlaid over an optical image of the cluster (courtesy P Rosati). The confirmed cluster members are indicated with boxes. Many more galaxies are expected to be member of this cluster.

High-z Clusters

CLUSTERS OF GALAXIES are the largest and most massive structures in the universe, held together by gravity. Their density is more than a thousand times the average density of the universe, and they are filled with galaxies, hot gas and dark matter.

It is very interesting to study clusters of galaxies out to the largest distances and highest REDSHIFTS. The most distant clusters are so far away that the light has traveled for a large fraction of the age of the universe, and we are therefore looking back in time when we study these clusters.

It is very hard to find the high redshift clusters. The oldest method relies on deep optical images taken of the sky. The clusters can be recognized as small concentrations of very faint galaxies. The Palomar Sky Survey was used to find clusters, but it is very hard to find clusters beyond a redshift of 0.4. Other techniques have been more efficient at finding higher redshift clusters. Powerful RADIO GALAXIES can reside in clusters, although most of the clusters found in this way are not very rich.

One of the most efficient ways of finding high redshift clusters is based on their x-ray emission. The hot gas that fills clusters emits strongly in x-rays. This x-ray emission is so bright that it can be seen out to redshifts of 1 and higher. An example of such an x-ray cluster is shown in figure 1. This cluster, MS1054–03, was discovered with the Einstein satellite in a deep survey on part of the sky. All x-ray sources in the survey were carefully identified in the optical by Isabella Gioia and collaborators. Finally, optical spectroscopy with the 10 m W M Keck telecope showed that this was the most distant cluster discovered with the Einstein satellite. It has a redshift of 0.83, and the universe was less than half its present age when the cluster emitted the light we observe.

The most distant cluster found in this way is RXJ0848.9+4452, which was first detected by the Rosat satellite, and identified as a cluster at a redshift of 1.26 by Rosati and collaborators. The image is shown in figure 2. The cluster is as massive as nearby ABELL CLUSTERS. The optical emission is quite faint because the light of the galaxies has been shifted to the infrared by the high redshift. Images in the infrared are necessary to find such high redshift clusters. Surprisingly, this cluster has been found to have a very close neighbor, at the same redshift. The second cluster had been found serendipitously by an infrared survey of the sky. This pair of clusters forms a 'supercluster'.

Now that large samples of clusters at high redshifts have been found, we can determine whether the clusters change with time, and whether the galaxies inside clusters change with time. The first question to answer is whether the number of clusters is different at high redshift. Many theoretical models predict that clusters

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

form 'hierarchically', which means that large clusters grow by the accumulation of smaller clusters. Some of these theories predict that the most massive clusters should not exist, or be very rare at redshifts of 1. Interestingly enough, the cluster MS1054–03 is so massive that its existence at a redshift of 0.83 poses problems for some theories. In general, there is no strong evidence that the number of clusters changes with redshift. The newest results indicate, however, that the most massive clusters are somewhat rarer at high redshift.

The high redshift clusters are also very suitable for studying the evolution of the galaxies inside the clusters. In the nearby universe, rich clusters are dominated by S0 galaxies and elliptical galaxies, and have only a small fraction of spiral galaxies at their center. Only recently has it become possible to obtain images of high redshift clusters with a quality good enough to undertake such studies. The very sharp images taken with the Hubble Space Telescope are of comparable quality to the earlier images taken of low redshift clusters.

Studies by Dressler and co-workers have demonstrated that the fraction of SPIRAL GALAXIES is significantly higher in clusters at a redshift of 0.5. Figure 3 shows the center of the cluster CL0939+4713 at a redshift of 0.41. The high fraction of spiral galaxies in the center of this cluster is obvious. Generally, high redshift clusters have far fewer S0 galaxies than low redshift clusters. These galaxies are the most common type of galaxy in low redshift clusters. They have probably been added to the low redshift clusters from the outside. Before they fall into the clusters, they look like spiral galaxies. When they fall in, they lose their gas, and appear as S0 galaxies soon after.

At even higher redshifts, the evidence for merging between galaxies becomes stronger. It has been speculated for a long time that ELLIPTICAL GALAXIES form by collisions of less massive galaxies. After two galaxies collide, they quickly merge together and form a new galaxy. If the two galaxies contain no or little gas, they will form a regular elliptical after the mergers.

It has always been difficult to find such mergers in clusters. However, a recent study of the rich cluster MS1054–03 shows that 18% of the member galaxies are mergers. These colliding galaxies are shown in figure 4. It is likely that the mergers evolve into elliptical galaxies or S0 galaxies. This is direct evidence that galaxies evolve through merging. (See also GALAXIES: INTERACTIONS AND MERGERS.)

The mergers and the conversion of spiral galaxies into S0's may very well explain why galaxies in clusters are very different from galaxies outside clusters. Understanding the origin of this effect is one of the goals of studies of distant cluster galaxies. By observing these clusters, astronomers can directly see what happened to the galaxies in the distant past.

Marijn Franx

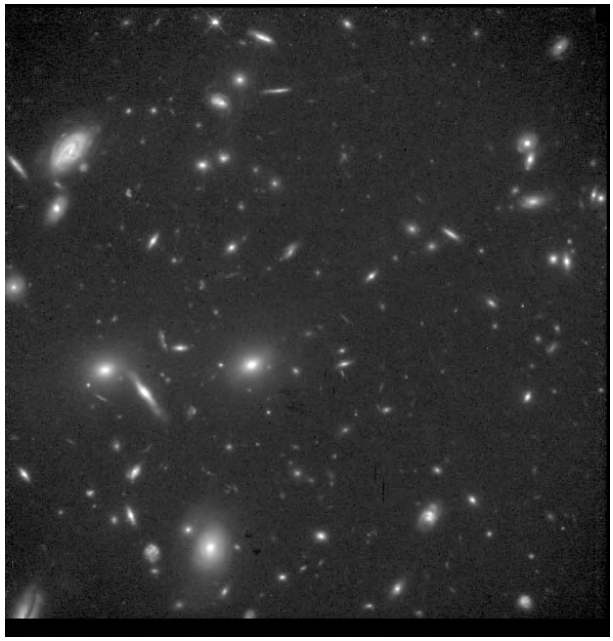


Figure 3. The central part of the distant cluster CL 0939+4713, imaged with the Hubble Space Telescope (courtesy A Dressler and NASA). The center contains many spiral galaxies, unlike clusters nearby. Many of these spiral galaxies look disturbed.

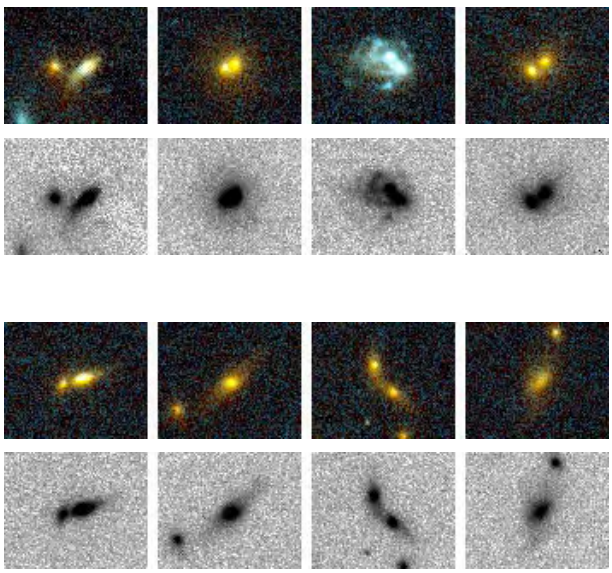


Figure 4. Mergers in the distant cluster MS1054–03 (courtesy van Dokkum, Franx, ESA and NASA). These galaxies have been confirmed to be cluster members. They are classified as mergers. Some show tidal arms, others double nuclei, or faint extensions due to the collisions of galaxies. Such mergers may have been much more prevalent in the past. **This figure is reproduced as Color Plate 8.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High Altitude Observatory

The High Altitude Observatory (HAO), in Boulder, CO, established in 1940, is dedicated to research in solar and solar-terrestrial physics with emphasis on solar–variability and its impact on the Earth. HAO is a division of the National Center for Atmospheric Research, sponsored by the National Science Foundation and operated by the University Corporation for Atmospheric Research, Boulder, CO. HAO's program includes visiting scientists, advanced education, workshops, development of ground- and space-based instruments, operation of observing facilities, distribution of data to the scientific community and numerical modeling. HAO conducts its research in strong cooperation with the university community, NASA and NOAA, and with many research centers worldwide.

For further information see
http://www.hao.ucar.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High Energy Astrophysical Observatory (HEAO)

Series of three NASA orbital observatories. HEAO-1, launched in August 1977, successfully completed the most accurate all-sky survey of x-ray sources up to that time. Discovered the 'Cygnus Superbubble' created by a series of supernovae. HEAO-2 (later known as EINSTEIN), launched in 1978, was the first true x-ray astronomy satellite. HEAO-3, launched in September 1979, carried a gamma ray spectrometer and two cosmic ray detectors. It observed spectral lines produced by electron–positron annihilation from the direction of the Galactic centre.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High Energy Transient Experiment (HETE) 2

A US–French satellite launch in October 2000. Its prime objective is to carry out a multiwavelength study of gamma ray bursts with ultraviolet, x-ray and gamma ray instruments. It will be able to localize bursts to within several arcseconds. The first HETE was lost due to a Pegasus launch failure in 1996.

On September 21, 2001, HETE detected a rare optical afterglow of a gamma-ray burst in the constellation Lacerta. The source was relatively close, only about 5 billion l.y. from Earth. The opportunity to see the afterglow in optical light provides crucial information about what is triggering these mysterious bursts, which scientists speculate to be the explosion of massive stars, the merging of neutron stars and black holes, or possibly both.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

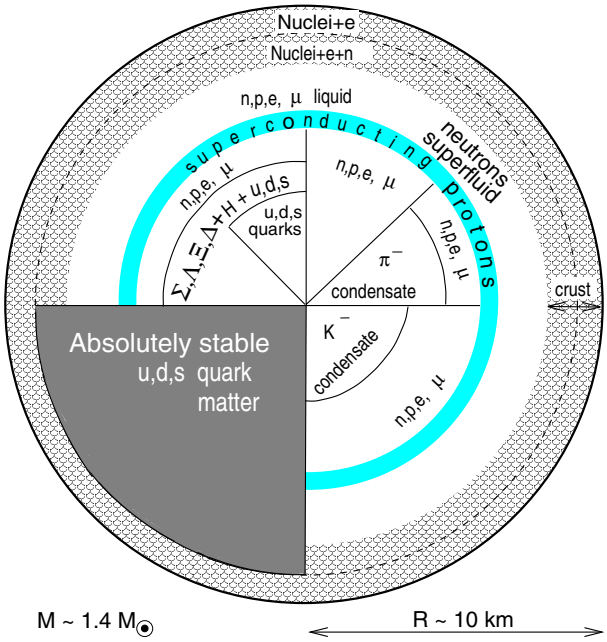


Figure 1. Competing structures and novel phases of high-density matter predicted by theory to make their appearances in the cores ($R \lesssim 8$ km) of neutron stars.

High-density Equation of State of Neutron Stars

The enormous gravitational pull that binds NEUTRON STARS compresses most of their matter to densities that are up to an order of magnitude higher than the mass density of atomic nuclei, 2.5×10^{14} g cm^{-3}. This provides a high pressure environment in which numerous subatomic particle processes, ranging from the generation of new BARYON particles ($\Sigma, \Lambda, \Xi, \Delta$), to QUARK deconfinement, to the formation of (π^- or K^-) boson condensates and H (dibaryon) matter, are expected to compete with one another, and novel phases of matter, like the quark–gluon plasma state being sought at the most powerful terrestrial particle colliders, could exist (figure 1). There are theoretical suggestions of even more exotic processes at supernuclear densities such as the formation of absolutely stable strange quark matter, a configuration of matter even more stable than the most stable atomic nucleus, iron.

No matter which physical processes are actually realized in the dense interiors of neutron stars, each one manifests itself in the high-density equation of state (EoS) that is pressure as a function of total energy density, $P(\epsilon)$, of neutron stars (figure 2).

Neutron stars are objects of highly compressed matter so that the geometry of SPACETIME is changed considerably from flat space. Thus models of such stars are to be constructed in the framework of Einstein's GENERAL THEORY OF RELATIVITY combined with theories of high-density matter. The connection between both is provided by Einstein's field equations (G denotes the gravitational constant, and $\mu, \nu = 0, 1, 2, 3$),

$$G^{\mu\nu} = 8\pi G T^{\mu\nu}(\epsilon, P(\epsilon)), \tag{1}$$

which couples the Einstein curvature tensor, $G^{\mu\nu}$, to the energy–momentum density tensor, $T^{\mu\nu}$, of neutron star matter. The tensor $T^{\mu\nu}$ contains the EoS of neutron star matter, which is derivable from a given stellar matter Lagrangian $\mathcal{L}_m(\{\chi\})$. In general, $\mathcal{L}_m$ is a complicated function of the numerous baryon ($\chi = \mathrm{p, n}, \Sigma, \Lambda, \Xi, \Delta$), meson ($\chi = \pi^-, \mathrm{K}^-$) and quark ($\chi = \mathrm{u, d, s}$) fields predicted to make their appearance in the cores of neutron stars (figure 1). Once $\mathcal{L}_m$ is specified, models for the EoS then follow according to the scheme

$$\frac{\partial \mathcal{L}_m}{\partial \chi} - \partial_\mu \frac{\partial \mathcal{L}_m}{\partial(\partial_\mu \chi)} = 0 \Rightarrow P(\epsilon). \tag{2}$$

In general, equations (1) and (2) are to be solved simultaneously since the particles move in curved spacetime whose geometry, determined by Einstein's field equations, is coupled to the total energy density of the matter fields. For neutron stars, however, the deviation from flat spacetime over the length scale of the nuclear force, ~1 fm, is practically zero up to the highest densities reached in the cores of such stars (some 10^{15} g cm^{-3}). This circumstance divides the construction of models of neutron stars into two distinct problems. Firstly, the effects of the short-range nuclear forces on the properties of neutron star matter are described in a local inertial frame, where spacetime is flat, by the parameters and laws of many-body physics. Secondly, the coupling between the long-range gravitational field and the matter is then taken into account by solving Einstein's field equations for the gravitational field, which leads to the global properties of neutron stars.

Hyperon population

In the most primitive conception, high-density neutron star matter is constituted of only neutrons. Such matter, however, constitutes a highly excited state relative to neutron star matter which will therefore quickly transform, via the weak reaction

$$\mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu}_\mathrm{e}, \tag{3}$$

into its chemically equilibrated state, made up of neutrons (n), protons (p) and electrons (e^-). NEUTRINOS do not accumulate (i.e. $\mu^{\bar{\nu}_\mathrm{e}} = 0$), because of their extremely small cross sections in neutron star matter. In this way the stellar matter develops Fermi seas of degenerate neutrons, protons and electrons, each one being filled up to the highest possible energy level determined by

$$\mu^\mathrm{n} = \mu^\mathrm{p} + \mu^{\mathrm{e}^-}. \tag{4}$$

Equation (4) is a special case of the general relation

$$\mu^\chi = B^\chi \mu^\mathrm{n} - q^\chi \mu^{\mathrm{e}^-}, \tag{5}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

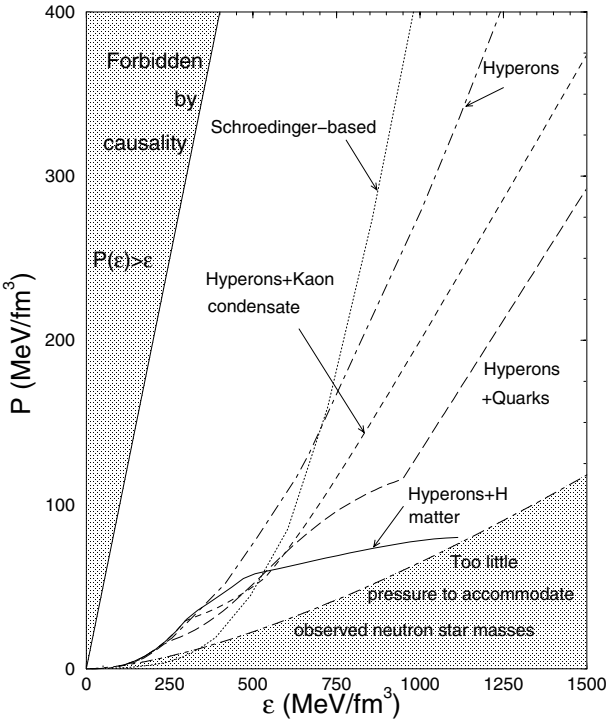


Figure 2. Models for the EoS of high-density neutron star matter. Constraints from causality and observed neutron star masses are indicated.

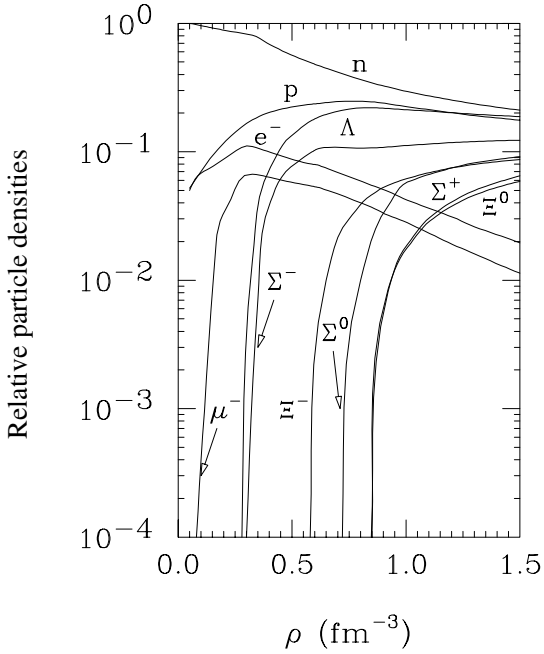


Figure 3. Baryon composition of neutron star matter.

which holds in any system characterized by two conserved charges. These are, in the case of neutron star matter, electric charge q^χ, and baryon number charge B^χ. Applied to the Λ hyperon ($B^\Lambda = 1$, $q^\Lambda = 0$), for instance, equation (5) leads to

$$\mu^\Lambda = \mu^{\rm n}. \tag{6}$$

Ignoring particle interactions, the chemical potential of a given particle χ follows from

$$\mu^\chi = \omega(k_{F_\chi}) \equiv \sqrt{m_\chi^2 + k_{F_\chi}^2}, \tag{7}$$

where $\omega(k_{F_\chi})$ is the energy–momentum relation of a free relativistic particle moving with Fermi momentum k_{F_χ}. Substituting (7) into (6) leads to ($m_\Lambda = 1116$ MeV, $m_{\rm n} = 939$ MeV)

$$k_{F_{\rm n}} > \sqrt{m_\Lambda^2 - m_{\rm n}^2} \sim 3\ {\rm fm}^{-1} \Rightarrow \rho_{\rm n} \equiv \frac{k_{F_{\rm n}}^3}{3\pi^2} \sim 2\rho_0. \tag{8}$$

That is, neutrons begin to transform into Λs at baryon number densities around twice the density of nuclear matter, $\rho_0 = 0.16\ {\rm fm}^{-1}$. This value is only slightly altered by the inclusion of particle interactions, as can be seen from figure 3.

Aside from chemical equilibrium, the condition of electric charge neutrality constitutes another characteristic feature of neutron star matter, as any net charge per particle Z/A in excess of around 10^{-36} would prevent gravity from binding neutron star matter. This too leads to decisive constraints on the particle Fermi momenta, which follow from the charge neutrality condition

$$\sum_{\chi={\rm p},\Sigma,\ldots;{\rm e}^-,\mu^-} q^\chi \frac{k_{F_\chi}^3}{3\pi^2} + \rho_M \Theta(\mu^M - m_M) \equiv 0. \tag{9}$$

The last term in equation (9) accounts for the electric charges carried by condensed mesons of type M, of which only the π^- or, alternatively, the currently more favored $\rm K^-$ come into question.

Figure 3 shows a representative sample composition of neutron star matter assumed to consist of only neutrons, protons and hyperons. The interactions are computed, following the scheme in equation (2), in the framework of relativistic nuclear field theory, where the motion of the particles is described by the coupled field equations

$$\begin{aligned}
({\rm i}\gamma^\mu \partial_\mu - m_\chi)\psi_\chi(x) &= \sum_{M=\sigma,\omega,\pi,\ldots} g_{M\chi} M(x)\psi_\chi(x), \\
(\partial^\mu \partial_\mu + m_\sigma^2)\sigma(x) &= \sum_{\chi=p,n,\Sigma,\ldots} g_{\sigma\chi}\bar{\psi}_\chi(x)\psi_\chi(x),
\end{aligned} \tag{10}$$

plus additional equations for the other meson fields ($M = \omega, \pi, \rho, \ldots$). Evidently, the free-particle case is recovered by setting the meson–baryon coupling constants $g_{M\chi} = 0$. Equations (10) are to be solved in combination with conditions (5) and (9) for the fields $\psi_\chi(x)$ and $M(x)$. This can be accomplished in the framework of different many-body approximations, of which the relativistic Hartree

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(mean-field) approximation is the simplest approach, followed, with increasing level of complexity, by the relativistic Hartree–Fock (RHF) and relativistic Brueckner–Hartree–Fock (RBHF) approximations. In the Hartree and RHF approximations the coupling constants $g_{M\chi}$ are chosen such that the theory accommodates the bulk properties of nuclear matter at saturation density ρ_0. These are the binding energy $E/A = -16$ MeV, effective nucleon mass $m_{\rm N}^*/m_{\rm N} = 0.79$, incompressibility $K \simeq 265$ MeV, and the symmetry energy $a_{\rm s} = 32$ MeV. Of the five, the value for the incompressibility of nuclear matter carries the largest uncertainty. Its value is currently believed to lie in the range between about 180 and 280 MeV. In sharp contrast to Hartree and RHF, the RBHF approximation makes use of meson-exchange models for the nucleon–nucleon interaction, whose parameters are adjusted to the nucleon–nucleon scattering phase shifts in free space and the deuteron's properties, so that all nuclear matter properties are computed self-consistently.

Once the fields $\psi_\chi(x)$ and $M(x)$ are computed, models for the EoS of high-density neutron star matter follow from the energy–momentum tensor $T^{\mu\nu} = T^{\mu\nu}(\psi_\chi)$,

$$\epsilon = T^{00}, \qquad P = \frac{1}{3}\sum_{i=1}^{3} T^{ii}. \tag{11}$$

The neutron star matter composition shown in figure 3 is computed for the relativistic Hartree approximation, with the associated EoS illustrated in figure 2. Three general physical features emerge. (1) The electric charges add up to zero at each density, as required by electric charge neutrality. (2) The generation of μ^- and Σ^- replaces high-energy electrons at the top of the electron Fermi sea. (3) The generation of baryons with isospin orientations opposite to the neutron's (e.g. Σ^+, Ξ^0) renders neutron star matter more isospin symmetric, that is closer equality between the number of neutrons and protons. All three features are essential for neutron star matter to settle down in the lowest energy state.

Another access to the derivation of the EoS is provided by Schrödinger-based treatments. These are based on phenomenological models for the two (v_{ij}) and three (v_{ijk}) nucleon interaction, eventually supplemented with relativistic corrections. The Hamiltonian thus is of the form

$$H = \sum_i^N \left(\frac{k_i^2}{2m}\right) + \sum_{i<j}^N v_{ij} + \sum_{i<j<k}^N v_{ijk}, \tag{12}$$

with N ($\sim 10^{57}$) the number of nucleons. The method adopted to solve the many-body Schrödinger equation $H|\Psi\rangle = E|\Psi\rangle$ is based on the variational approach, where the variational parameters contained in the wave function $|\Psi\rangle$ are varied to minimize the energy per baryon for a given density, i.e.

$$E(\rho) = \min \frac{\langle\Psi|H|\Psi\rangle}{\langle\Psi|\Psi\rangle}. \tag{13}$$

The energy density, $\epsilon(\rho)$, and pressure, $P(\rho)$, are then given by

$$\epsilon(\rho) = \rho(E(\rho) + m_{\rm N}), \quad P(\rho) = \rho^2 \frac{\partial}{\partial\rho} E(\rho), \tag{14}$$

which can be combined to the EoS in the form $P(\epsilon)$. The dotted curve in figure 2 is an example of an EoS derived in this framework.

Condensation of K^- mesons
Once the reaction

$$e^- \rightarrow K^- + \nu_e \tag{15}$$

becomes possible in neutron star matter, it is energetically advantageous for the matter to replace the fermionic electrons with the bosonic K^- mesons. Whether or not this actually happens depends on the effective mass of the K^- and the value of μ^{e^-} in neutron star matter, as can be seen from the threshold condition

$$m_{K^-}^* \equiv \mu^{K^-} = \mu^{e^-}, \tag{16}$$

which follows from equation (5). Calculations indicate that $m_{K^-}^*$ may become as small as $m_{K^-}^* \simeq 200$ MeV in neutron star matter above a few times ρ_0. Such mass values lie in the vicinity of μ^{e^-} so that the threshold condition for the onset of K^- condensation could indeed be fulfilled. A sample EoS of K^- condensed neutron star matter is shown in figure 2. Reaction (15) is followed by

$$n + e^- \rightarrow p + K^- + \nu_e, \tag{17}$$

with the neutrinos leaving the star as described above. By this conversion the nucleons in the cores of newly formed neutron stars can become half neutrons and half protons. The relatively isospin symmetric composition of neutron star matter achieved in this way resembles the one of atomic nuclei, which are made up of roughly equal numbers of neutrons and protons. Neutron star matter is therefore referred to, in this picture, as nucleon star matter.

H-dibaryons
Another boson that could make its appearance in high-density neutron star matter is the H-dibaryon, a doubly strange six-quark composite with baryon number two. Since its first prediction in 1977, the H-dibaryon has been the subject of many theoretical and experimental studies as a possible candidate for a strongly bound exotic state. In neutron star matter, which, as known from figure 3, may contain a significant fraction of Λ hyperons, pairs of them could combine to form H-dibaryons, which could give way to the formation of H-matter at densities somewhere between $3\rho_0$ and $6\rho_0$, depending on the in-medium properties of the H-dibaryon. The condensation could be accompanied by a pronounced softening of the EoS (solid curve in figure 2) of such matter. If indeed formed, H-matter may not form a permanent component of neutron star matter but, because of its instability against compression, could trigger the conversion of neutron star matter into hypothetical strange star matter (see below).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

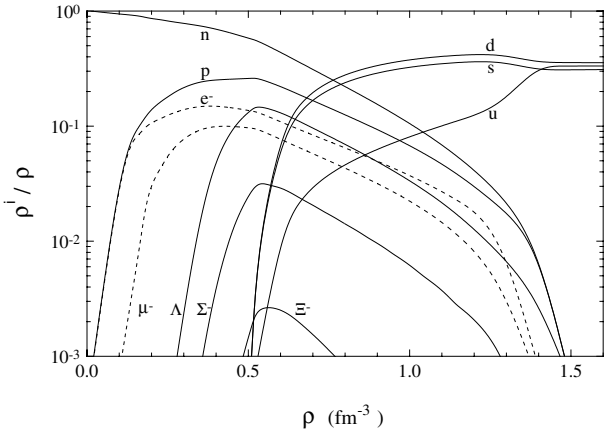


Figure 4. Sample composition of neutron star matter accounting for quark deconfinement.

Quark deconfinement

It has been suggested already back in the 1970s that, because of the extreme densities reached in the cores of neutron stars, neutrons and protons may melt, creating the quark–gluon plasma state being sought at the most powerful terrestrial heavy-ion colliders. At present one does not know from experiment at what density the expected phase transition to quark matter occurs, and one has no conclusive guide from lattice QCD simulations either. From simple geometrical considerations it follows that neutrons and protons begin to touch each other at densities of $\sim(4\pi r_{\rm N}^3/3)^{-1} \simeq 0.24$ fm^{-3}, which, for a characteristic nucleon radius of $r_{\rm N} \sim 1$ fm, is less than twice ρ_0. Above this density, therefore, it appears plausible that the nuclear boundaries of neutrons and protons dissolve and that the quarks, formerly confined inside neutrons and protons, begin to populate free states outside of the nucleons. More than that, since the mass of the strange quark is only $m_{\rm s} \sim 150$ MeV, high-energetic up and down quarks are expected to readily transform to strange quarks at about the same density at which up and down quark deconfinement sets in, giving way to the formation of three flavor (u,d,s) quark matter at densities just a few times ρ_0. A sample composition of such matter is shown in figure 4, and its associated EoS is illustrated in figure 2.

Diquark condensation and color superconductivity

Very recently it was discovered that strong correlations between up and down quarks may exist which could lead to the formation of colored diquark pairs in high-density matter. These pairs, carrying integer spin, could form a Bose condensate in cold ($T < 50$ MeV) and dense ($\rho > 3\rho_0$) quark matter which ought to exhibit color superconductivity. Both the magnitude of the gap and the critical temperature associated with the color superconductive phase were estimated to be on the order of ∼100 MeV, about two orders of magnitude larger than the gaps associated with neutron and proton superfluids. Future theoretical studies may reveal whether these features will have their correspondence in a more complete treatment of QCD at finite baryon number density.

Absolutely stable strange quark matter

So far it was assumed that quark matter forms a state of matter higher in energy than atomic nuclei. This most plausible assumption, however, may be quite deceiving because, for a collection of more than a few hundred u,d,s quarks, the energy per baryon (E/A) of quark matter can be equally well below the energy of the most stable atomic nucleus, ^{56}Fe, whose energy per baryon is $M(^{56}\mathrm{Fe})c^2/56 = 930.4$ MeV, with $M(^{56}\mathrm{Fe})$ the mass of the ^{56}Fe atom. The EoS of strange quark matter can be represented mathematically as $P = (\epsilon - 4B)/3$ (bag constants $B \sim 80$ MeV fm^{-3}). This implies a surface density of $\epsilon = 4B \sim 6 \times 10^{14}$ g cm^{-3} for stars made up of such matter which distinguishes them from neutron stars made up of conventional matter.

Bibliography

Comprehensive overviews of modern research on the EoS of neutron star matter and stellar structure are compiled by:

Glendenning N K 1997 *Compact Stars, Nuclear Physics, Particle Physics, and General Relativity* (New York: Springer)

Weber F 1999 *Pulsars as Astrophysical Laboratories for Nuclear and Particle Physics* (Bristol: Institute of Physics)

Fridolin Weber

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

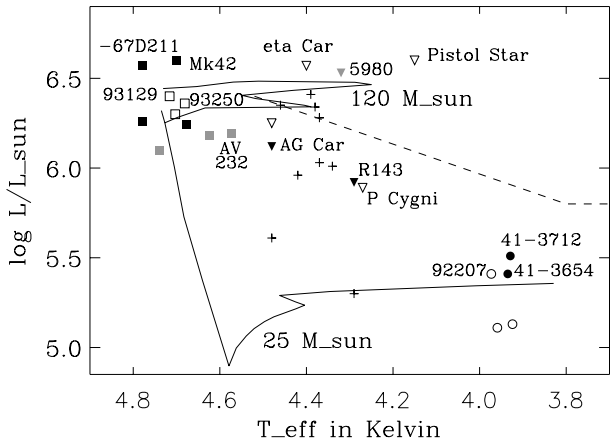


Figure 1. The Hertzsprung–Russell diagram of the most luminous stars. Stars of spectral type O and luminous blue variables are represented by squares and triangles, respectively. Open symbols correspond to the Galaxy, solid symbols to the Large Magellanic Cloud and grey symbols to the Small Magellanic Cloud. To avoid confusion the diagram does not include other hot and very luminous stars of spectral type WN detected recently in the 30 Doradus region of the LMC, which are similar to Mk42. The crosses represent the most luminous stars of the young stellar cluster in the very center of our galaxy. Circles correspond to A supergiants (open symbols, galaxy; solid, M31). Although they are a factor of 10 fainter in bolometric luminosity, objects of this kind are the brightest 'normal' stars in visual light. The theoretical main sequence and stellar evolutionary tracks (labelled by the stellar mass) are given as solid curves. The dashed curve is the 'Humphreys–Davidson limit' discussed in the text.

High-luminosity Stars

The most luminous stars in galaxies are of extraordinary astronomical importance. With the new generation of 10 m class ground-based telescopes and with the large observatories in space such as the HST they can be studied as individuals in galaxies out to 20 Mpc distance including the VIRGO and FORNAX GALAXY CLUSTERS. The quantitative analysis of their spectra yields important information about distances, chemical composition and evolution of galaxies.

In recent years galaxies at extremely large distances, billions of light-years distant, have been detected. Because of their large REDSHIFTS ($z \simeq 3$–4) these galaxies show their ultraviolet light shifted to optical wavelengths. Spectroscopy with 10 m class telescopes reveals that the radiation of these galaxies is dominated by the contribution of the most luminous stars which are usually very hot and emit mostly in the ultraviolet. Using the knowledge obtained about this extreme population of stars in our own Galaxy and the galaxies of the LOCAL GROUP it is, at least in principle, possible to investigate the physics of GALAXIES AT HIGH REDSHIFT belonging to an early phase of the evolution of the universe.

The spectral analysis of the most luminous stars is difficult and complex. The physical reason for the complexity is precisely this prodigious luminosity and the correspondingly large energy and momentum density of their intrinsic radiation field, which dominates the thermodynamics and hydrodynamics of the interior structure and the entire atmospheres. Extreme departures from local thermodynamic equilibrium (LTE), which require the extensive effort of non-LTE radiative transfer and model atmospheres for the spectral analysis, result from this dominating role of the radiation field. In addition, very strong stellar winds are initiated and maintained by the absorption of photon momentum through ionic species of the atmospheric plasma. Through these winds a significant amount of the stellar mass is lost and recycled back to the interstellar medium during the lifetime of these stars.

Figure 1 shows the HERTZSPRUNG–RUSSELL DIAGRAM (HRD) of the most luminous stars known so far in the Milky Way and the nearby galaxies of the Local Group.

The spectral type O3—the most luminous hydrogen burning stars

Luminous stars are usually formed in giant and dense molecular clouds as the most massive objects of very young stellar clusters. Typical examples are the clusters in the CARINA Nebula of the Milky Way and in the TARANTULA NEBULA of the Large Magellanic Cloud. After birth these objects are very hot. Their atmospheric effective temperatures are about 50 000 K. With masses of about a hundred solar masses or more their luminosities are in the range between one and three million solar luminosities. The energy production leading to these enormous luminosities is central nuclear hydrogen burning in the CNO cycle. The interior hydrostatic equilibrium is kept mostly through radiation pressure which may be the cause for pulsational instability. The nuclear lifetime in this stage is very short, roughly 2.5 million years. It is almost independent of the stellar mass in these extreme conditions dominated by electron scattering in the STELLAR INTERIOR.

It is not so easy to detect these most luminous objects in our Milky Way by conventional photometric techniques. Interstellar reddening and extinction and the fact that stellar colors measured in the optical part of the electromagnetic spectrum depend only very weakly on temperature as soon as $T_{\rm eff}$ is above 35 000 K make it impossible to identify these stars by photometry only. The only reliable way is spectroscopy. For optical spectroscopy the criterion for the most luminous and most massive very young stars in the stage of central hydrogen burning is the extreme weakness or complete absence of neutral helium lines relative to strong lines of singly ionized helium. The corresponding spectral type is O3. O3 stars are the hottest, most luminous and most massive stars known. Figure 2 shows how the relative strength of neutral to ionized helium lines changes between spectral types O3, O5 and O6.5 corresponding to a temperature range from 50 000 K to 41 000 K. However, there might be a significant intrinsic spread in effective temperature and luminosity among the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

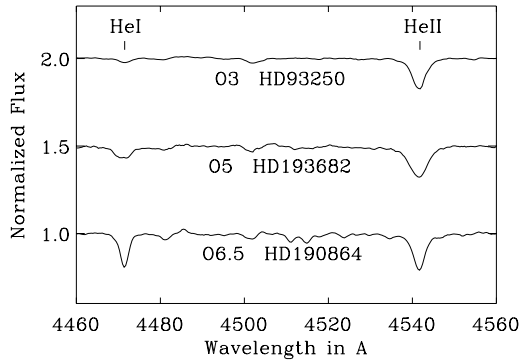


Figure 2. He I and He II spectral lines in the optical spectra of O stars. At spectral type O3 the He I lines are very weak.

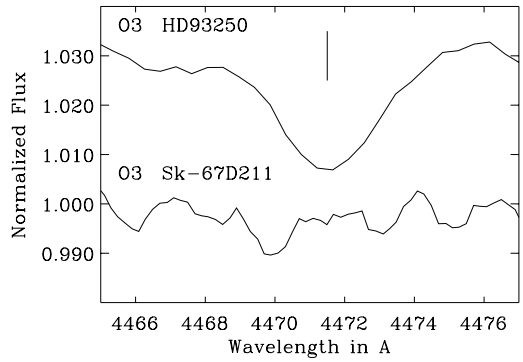


Figure 3. Enlarged plot of the He I 4471 line in two O3 stars (the vertical bar indicates a line depression of 0.01 in units of the continuum). While the line can still be recognized in the galactic object HD93250 ($T_{\rm eff} \simeq 50\,000$ K, $\log L/L_\odot \simeq 6.3$), it is absent in the spectrum of Sk-67D211, one of the hottest, most luminous and most massive stars in the Large Magellanic Cloud ($T_{\rm eff} \simeq 60\,000$ K, $\log L/L_\odot \simeq 6.6$). (See Puls *et al* 1996 *Astron. Astrophys.* **305** 171.)

O3 stars. This is indicated by figure 3 which compares high signal-to-noise spectra of two O3 stars.

The ultraviolet spectra of O3 stars are dominated by strong and broad stellar wind lines superimposed on a forest of narrow photospheric lines mostly of highly ionized iron. Figure 4 shows the UV spectrum of the possibly most massive star in our galaxy.

The signatures of the same UV stellar wind lines are also found in the optical spectra of galaxies at high redshift (figure 5), indicating that the ultraviolet light of these objects is dominated by the emission of the population of the most massive stars.

The evolution of the most luminous stars—the Humphreys–Davidson limit

During central hydrogen burning O stars evolve towards lower effective temperature in the HRD. However, how far they evolve to the right in the HRD is obviously a function of luminosity. The corresponding observational border line in the HRD is called the 'Humphreys–Davidson limit'. The physical reason for the occurrence of this limit is the strong stellar mass loss which peals off the outer stellar layers and leads to a more or less bare contracting stellar core evolving to the helium burning main sequence at very high temperatures before the star has the chance to become a red supergiant (see RED GIANT STARS). Since the strength of mass loss is a function of luminosity (see below), the Humphreys–Davidson limit has a certain slope in the HRD for luminosities larger than 6×10^5 solar luminosities. Objects with lower luminosity pass through the stage of the red supergiants and return to the blue on their way to the helium main sequence.

Close to the Humphreys–Davidson limit the most luminous stars enter an extremely unstable phase of their evolution. They become 'LUMINOUS BLUE VARIABLES' (LBVs). These objects show very strong stellar winds, a significant variability of their brightness and sometimes dramatic eruptions, where significant amounts of stellar mass are ejected and the stellar properties change dramatically. The most famous examples of this group are η Carinae and P Cygni. During such eruptions the luminosity stays constant, but the radius increases substantially and the effective temperature drops. As a result of this process LBVs in outburst become much brighter in visual light. η Carinae, for instance, suddenly became one of the brightest stars in the southern sky during the late 1830s. Something similar happened with P Cygni in the 17th century.

Before and after the LBV phase or when evolving towards or away from the red supergiant phase the most luminous blue stars are spectroscopically identified as blue supergiants with characteristic emission lines in their optical and IR spectra caused by their stellar winds. These 'normal supergiants' are extremely useful as extragalactic distance indicators (see below). The evolution of the most luminous stars comes to an end when they have lost most of their hydrogen envelope through winds and eruptions and appear as bare helium stars on the helium burning main sequence. Still suffering from extremely strong stellar winds these objects show very broad pronounced emission lines of helium, nitrogen or carbon in their optical spectra. They appear as 'WOLF–RAYET STARS'. After this evolutionary stage they explode as SUPERNOVAE.

The cluster of luminous stars in the Galactic center

There is a long-standing debate as to what is the source of luminosity of the central few parsecs of our Galaxy. Using adaptive optics combined with imaging spectroscopy in the infrared it has become clear that the center of the Milky Way is populated by young stellar clusters of very luminous and very massive stars which are only a few million years old and fully account for the bolometric Lyman ionizing luminosities observed. The most luminous objects are evolved blue supergiants close to the evolutionary stage of Wolf–Rayet stars. Their striking spectroscopic characteristic is the strong emission in the He I 2.06 μm line transition in the IR (see figure 6).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

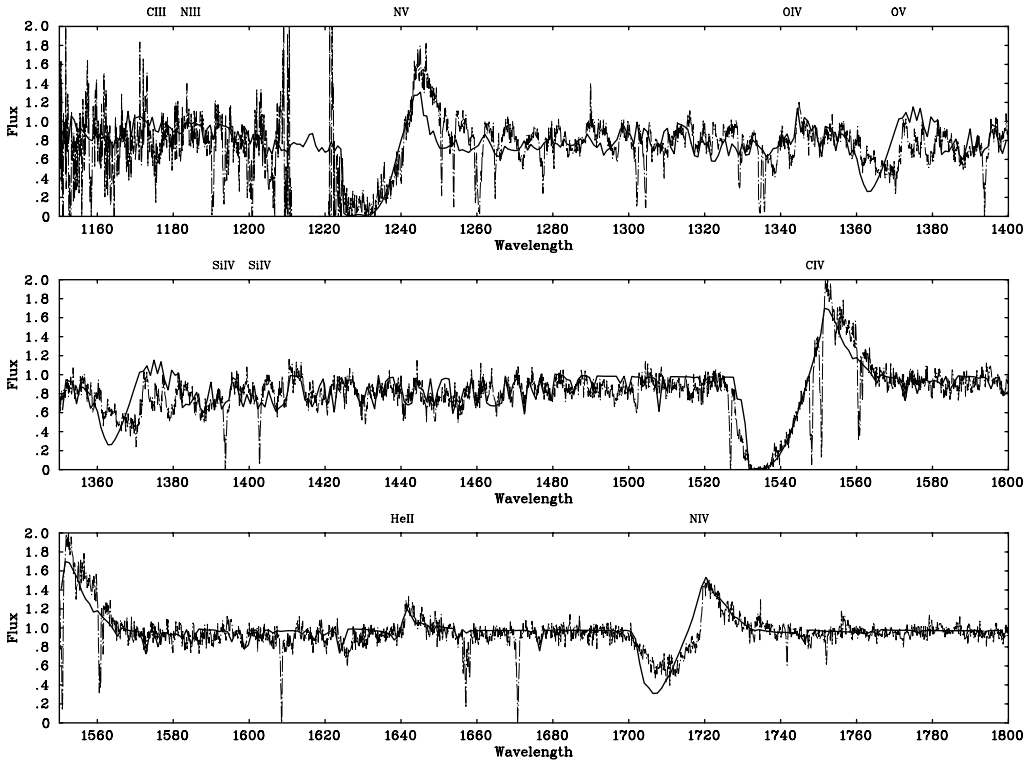


Figure 4. The observed UV spectrum (dotted) of the extremely luminous O3 star HD93129A in the Carina Nebula. The spectrum shows broad stellar wind lines of N IV, N V, O IV, O V and He II and many pseudophotospheric lines of Fe IV and V. The solid spectrum is the result of a non-LTE model atmosphere computation used for a detailed spectral analysis. (From Taresch *et al* 1997 *Astron. Astrophys.* **321** 531.)

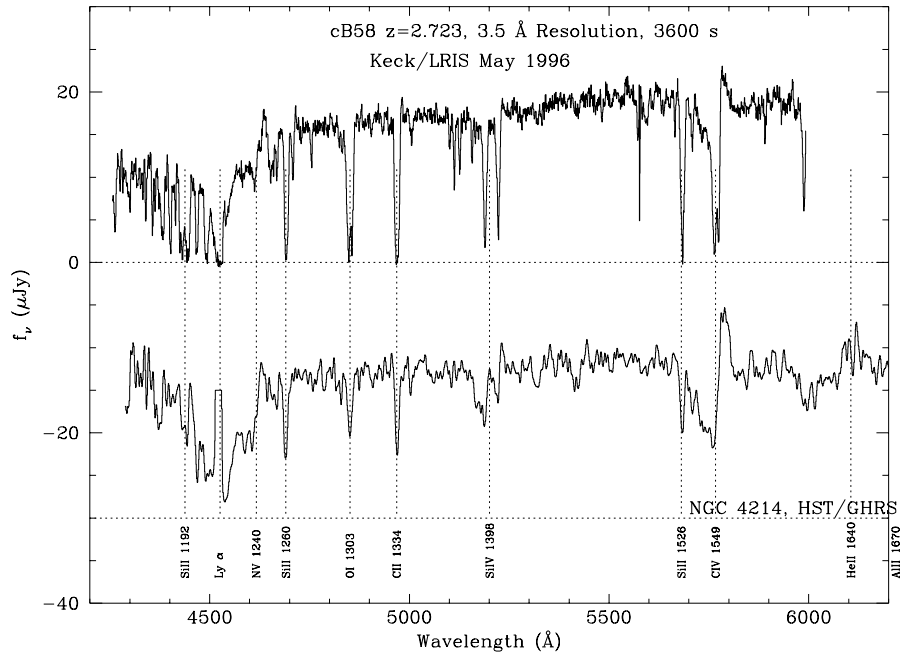


Figure 5. Optical spectrum of the gravitationally lensed galaxy cB58 at redshift $z = 2.7$ compared with a spectrum of a 'local' starburst galaxy. The broad stellar wind lines and interstellar lines can easily be recognized. (From Pettini *et al* 2000 *Astrophys. J.* **528** 96.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

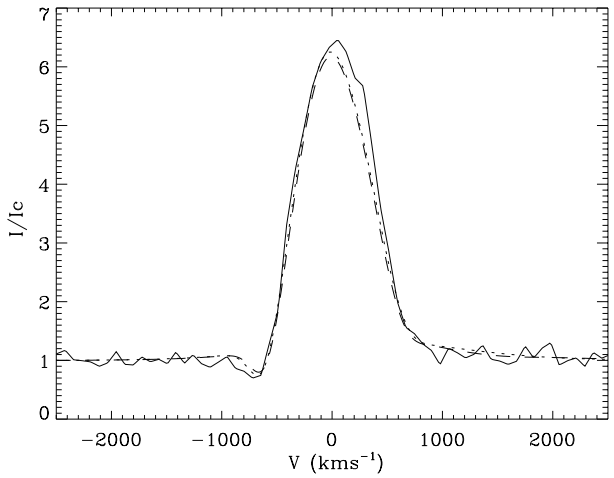


Figure 6. Profile of the K-band infrared He I 2.06 μm emission line of one of the blue supergiants in the Galactic center compared with model atmosphere calculations including stellar winds. (See Najarro *et al* 1997 *Astron. Astrophys.* **325** 700.)

Luminosities and temperatures of the brightest objects are indicated in figure 1.

The optically brightest 'normal stars', the wind momentum–luminosity relationship and the determination of extragalactic distances

Massive stars with luminosities below the Humphreys–Davidson limit evolve at (roughly) constant luminosity towards the red supergiant stage and then back towards hotter temperatures. These blue supergiants are the brightest 'normal' stars in visual light, where 'normal' means that they do not suffer from strong eruptions such as those of the LBV. The reason for their enormous visual brightness is that with decreasing atmospheric temperature during the evolution away from the O star domain an increasing fraction of the total luminosity is emitted in visual light as a consequence of Wien's law of radiation. The A supergiants shown in figure 1 have absolute magnitudes M_v of about -9 mag. Only LBV in outburst or supernovae can become brighter in visual light. Consequently, such objects are ideal for study as individuals in distant galaxies to provide information about chemical composition and distance.

One of the most interesting aspects of the recent work on normal blue supergiants is the detection that the strengths of their stellar winds are strongly correlated with stellar luminosity (see figure 7). A careful analysis of the hydrodynamics of radiation-driven winds leads to the prediction that for each supergiant spectral type the mechanical momentum of the stellar wind flow, which is the product of the mass-loss rate $\dot{M}$ and the maximum velocity v_∞, should be strongly dependent on stellar luminosity. This prediction has been confirmed by the detailed spectral analysis of stellar winds of O, B and A supergiants. Figure 8 gives an example for A supergiants (the figure shows wind momenta multiplied by the square root of the stellar radius in units of the solar radius, since this is exactly the quantity predicted to depend on luminosity or—for A supergiants—on absolute visual magnitude). The fact that the relationship holds in this range of extremely high absolute visual magnitudes is very encouraging for the use of A supergiants as extragalactic standard candles. The idea is to determine their wind momenta from spectroscopy and then to use the wind momentum–luminosity relationship observed and calibrated in the Milky Way and in galaxies of the Local Group to determine extragalactic distances. First estimates come to the very encouraging conclusion that with the use of spectrographs attached to 10 m class telescopes the method will allow accurate distance determinations out to galaxies as distant as 20 Mpc.

Bibliography

A comprehensive description of the physics of atmospheres and winds as well as the observed properties of

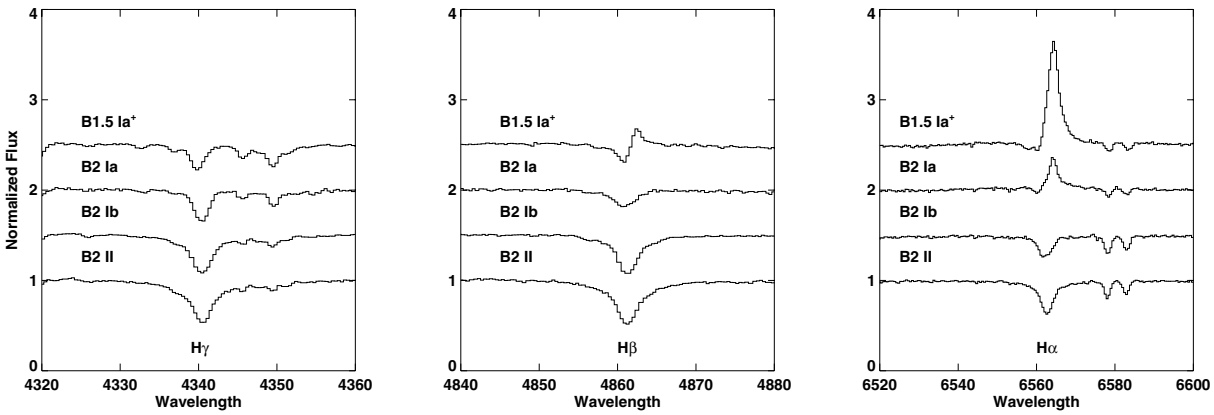


Figure 7. Hydrogen Balmer line profiles of galactic B2 supergiants of different luminosity classes showing the increasing contamination with stellar wind emission in Hα and Hβ.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the most luminous stars is given in a book which appeared recently in the series of Cambridge Contemporary Astrophysics:

Aparicio A *et al* (ed) 1998 *Stellar Astrophysics for the Local Group* (Cambridge: Cambridge University Press)

This book (the proceedings of the VIII Canary Islands Winter School) contains three longer articles by P Massey, 'Observations of the most luminous stars in Local Group galaxies' (pp 95–148), R P Kudritzki, 'Quantitative spectroscopy of the brightest blue supergiant stars in galaxies' (pp 149–262), and C Leitherer, 'Populations of massive stars and the interstellar medium' (pp 527–607) dealing with role of massive stars in galaxies. An additional article by C Chiosi (pp 1–94) deals with stellar evolution. Alternatively, a comprehensive article describing the evolution of massive and very luminous stars is the paper

Schaller G, Schaerer D, Meynet G and Maeder A 1992 New grid of stellar models from 0.8 to 120 solar masses at $Z = 0.02$ and $Z = 0.001$ *Astron. Astrophys. Suppl.* **96** 269–331

A description of our knowledge about LBV is found in

Humphreys R H and Davidson K 1994 The luminous blue variables: astrophysical geysers *Publ. Astron. Soc. Pac.* **106** 1025–51

The physical properties of the luminous blue supergiants in the central parsec of the Galaxy are discussed by

Najarro F, Krabbe E, Genzel R, Lutz D, Kudritzki R P and Hillier D J 1997 Quantitative spectroscopy of the HeI cluster in the Galactic center *Astron. Astrophys.* **325** 700–13

The article

Figer D, Najarro F, Morris M, McLean I S, Geballe T R, Ghez A M and Langer N 1998 The pistol star *Astrophys. J.* **506** 384

describes the IR diagnostics of one of the most luminous stars in the Galaxy.

The observed wind momentum–luminosity relationship of 'normal' blue supergiants is discussed in

Puls J *et al* 1996 O-stars mass-loss and wind momentum rates in the Galaxy and the Magellanic Clouds. Observations and theoretical predictions *Astron. Astrophys.* **305** 171

and in

Kudritzki R P, Puls J, Lennon D J, Venn K A, Reetz J, Najarro F, McCarthy J K and Herrero A 1999 The wind momentum–luminosity relationship of galactic A- and B-supergiants *Astron. Astrophys.* **350** 970–84

Rolf-Peter Kudritzki

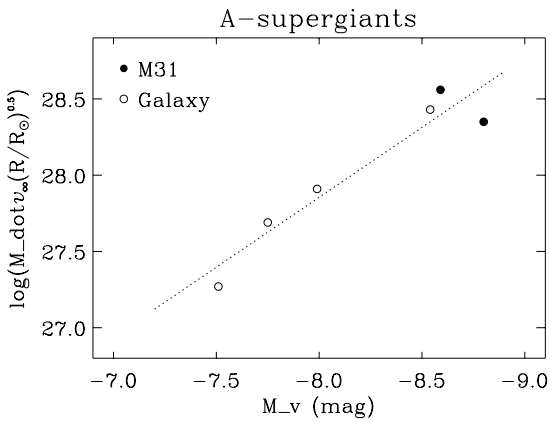


Figure 8. Modified stellar wind momenta of A supergiants in the Milky Way and the Andromeda Galaxy M31 as a function of absolute visual magnitude. This wind momentum–luminosity relationship can be used to determine distances out to the Virgo and Fornax cluster of galaxies. (From Kudritzki *et al* 1999 *Astron. Astrophys.* **350** 970.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High-redshift Quasars

The enormous luminosities of QUASARS allow their detection at great distances; the highest-redshift quasars have excited particular interest because they reveal information about the conditions in the universe at a very early epoch. In a single generation the volume of the universe encompassed by the most distant known object has increased by a factor of approximately 40. We have now observed quasars that existed within about a billion years of the Big Bang.

Historical perspective

Quasars were first identified as a discrete class of objects in 1960. Their combination of strong radio emission, substantial brightness variations, and unusual optical spectra (a series of broad, unidentified emission lines superposed on a smooth continuum) was unprecedented; these 'radio stars' were thought to represent an exotic stage of STELLAR EVOLUTION. Maarten Schmidt's decoding of the spectrum of 3C 273 in 1963 (see BRIGHT QUASAR 3C 273) revealed that the redshift[1] (z) of the source was 0.158; his interpretation of the spectrum removed quasars from the realm of our Galaxy and placed them in the distant regions of the universe, and made quasars, by a substantial factor, the most luminous known objects. Within a year the largest quasar redshift exceeded the highest measured galaxy redshift[2]. For more than 30 years the largest redshift belonged to quasars; their reign as the most distant known objects ended in 1997 with the Hubble Space Telescope-aided discovery of a $z = 4.92$ gravitationally lensed galaxy.

Table 1 lists the objects that have at one time possessed the largest known quasar redshift. It is clear that the study of the distant universe has not proceeded at a steady rate; there was a rapid increase in the mid-1960s (note the spectacular discovery of 3C 9), a hiatus of 9 years from 1973–1982, and a spate of discoveries of $z > 4$ quasars in the late-1980s. A particularly striking aspect is the change in the discovery technique for high-redshift quasars; before 1985 all the 'highest-redshift' quasar identifications were initially based upon radio detections, while all subsequent listings in the table were found purely from optical observations. The definition of what constitutes a high-redshift quasar obviously changes with time. For the past decade this term has been applied to quasars with redshifts larger than four; in July 1998 61 quasars have been identified with $z > 4$.

[1] The redshift of a photon is defined by

$$z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}} - 1.$$

For an object at redshift 1, the observed wavelengths of all the radiation have been increased by a factor of two over the emitted wavelengths.

[2] In 1961 R Minkowski capped a distinguished career by obtaining a spectrum of the radio galaxy 3C 295 ($z = 0.46$) on his last night of observing.

Table 1. Highest known quasar redshift.

Quasar	z	Month–year†	Discovery
3C 273	0.16	03–1963	Radio
3C 48	0.37	03–1963	Radio
3C 147	0.55	02–1964	Radio
3C 9	2.01	04–1965	Radio
PKS 0106+01	2.11	02–1966	Radio
PKS 1116+12	2.12	04–1966	Radio
PKS 0237−23	2.22	02–1967	Radio
4C 25.05	2.36	06–1968	Radio
5C 02.56	2.39	12–1968	Radio
4C 05.34	2.88	05–1970	Radio
OH 471	3.40	04–1973	Radio
OQ 172	3.53	06–1973	Radio
PKS 2000−330	3.78	09–1982	Radio
Q 1208+1011	3.80	07–1986	Optical‡
Q 0046−293	4.01	01–1987	Optical§
PC 0910+5625	4.04	10–1987	Optical‡
Q 0051−279	4.43	12–1987	Optical§
PC 1158+4635	4.73	12–1989	Optical‡
PC 1247+3406	4.90	09–1991	Optical§

† Date of publication of discovery.
‡ Slitless Spectroscopy.
§ Multicolor.

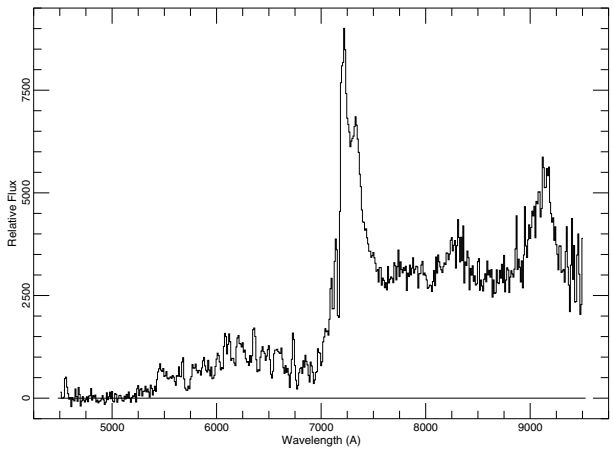


Figure 1. The spectrum of PC 1247+3406, the quasar with the highest known redshift ($z = 4.90$). The strongest emission line is Lyman α; the line at $\lambda = 9150$ Å is produced by C IV. The sharp drop in the continuum at wavelengths below the Lyman α emission line is absorption produced by intervening material.

Figure 1 displays the spectrum of PC1247+3406, the quasar with the highest known redshift ($z = 4.90$); this observation is a typical example of a high-redshift quasar spectrum. The most prominent feature is the Lyman α emission line; it is produced by the transition from the first excited state to the ground state in hydrogen, and has a rest wavelength of 1216 Å. At a redshift of 4 this feature is shifted from the far-ultraviolet to the red region of the optical spectrum. The second strongest emission line is that of thrice-ionized carbon (a doublet with rest wavelengths of 1548 and 1550 Å).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Redward of the Lyman α line the spectrum is relatively smooth and flat, but as one traverses the line to the blue, there is a dramatic decrease in the flux. At a redshift of 4 the level of the 'red continuum' is about twice that of the 'blue continuum'; in the highest-redshift quasars this drop becomes nearly a factor of four. High spectral resolution data show that the apparent decrease in the continuum level is actually due to a multitude of discrete absorption lines. This absorption is produced by individual 'clouds' of material along the line of sight; the strongest transition is Lyman α, and thus this region of the spectrum is referred to as the 'LYMAN α FOREST'. A second discontinuity frequently appears in the spectra of high-redshift quasars at a rest wavelength of ≈900 Å; in figure 1 it occurs at an observed wavelength near 5400 Å. This feature is the Lyman series limit, and is caused by a path through a neutral hydrogen column density of $\geq 10^{17}$ cm^{-2}.

Surveys for high-redshift quasars

The identification of high-redshift quasars presents a difficult challenge, as one can see from the nearly quarter-century lag between the initial quasar redshift and the discovery of the first $z > 4$ source. It is usually impossible to determine whether an image of an object, taken with a ground-based telescope, is a star or a quasar (remember that the word 'quasar' is derived from 'quasi-stellar'). Attempts to locate $z > 4$ quasars using radio observations were not successful (until quite recently). To discover high-redshift quasars with optical telescopes, a number of obstacles must be overcome (see also QUASISTELLAR OBJECTS: SURVEYS).

- High-redshift quasars are rare. On average there is one $z > 4$ quasar brighter than an R magnitude of 20 for every ≈15 square degrees of sky. Even in the regions of lowest stellar density (near the galactic poles), this amount of sky would contain approximately 10 000 stars of similar brightness.
- High-redshift quasars are red. At a redshift of 4, most of a quasar's light below 6000 Å is absorbed by intervening material. Photographic emulsions, the primary detector for wide area surveys, tend to be most sensitive in the blue. As one moves to higher redshifts, the bulk of a quasar's luminosity is observed in the near-infrared, where the background produced by the night sky is much higher than the AIRGLOW in the blue and visible regions of the spectrum.
- High-redshift quasars are faint. Only a small fraction of $z > 4$ quasars can be detected in the Palomar Sky Survey, which has served as the primary optical database for the second half of this century.

This sounds like a hopeless task indeed, until one notes that the optical spectrum of a high-redshift quasar is very different from the principal source of confusion, stars. In contrast to stars, whose spectral energy distribution is generally relatively smooth, high-redshift quasars possess a series of strong emission lines as well as broad and deep absorption features.

The explosion in the level of computational ability available to individual investigators in the 1980s led to a revolution in high-redshift quasar research. Using data obtained from either red-sensitive photographic plates or electronic detectors such as CCDs, it became possible to digitize large-scale surveys, find and characterize all objects using automated software packages (see SOFTWARE: DIGITAL SKY SURVEYS), and sift through the myriads of stars for the 1-in-10 000 source.

The vast majority of high-redshift quasars have been identified with optical telescopes. Discovery techniques include

- **Multicolor imaging.** The colors of all objects in a given region of sky are determined from images taken through a set of filters. The colors of stars fall along a well-defined locus of points in multidimensional color space, whereas high-redshift quasars lie significantly (often more than a magnitude) away from the stars. For example, if one performs a survey with the BRI filter set, the Lyman α emission line in a $z = 4.4$ quasar will fall near the center of the R filter. The $(B - R)$ color of the quasar will be extremely red (line-enhanced R emission compared to the Lyman α forest/Lyman-limit system depressed B flux) while the $(R - I)$ color will be moderately blue. Stephen Warren and his collaborators employed a multicolor technique to identify the first $z > 4$ quasar. To date about 60% of the $z > 4$ quasars have been found using multicolor photometry, including PC 1247+3406.
- **Slitless spectroscopy.** Multicolor imaging can be thought of as very low resolution spectrophotometry (spectral resolution of ≈5). Observing a large region of sky through a dispersive element (commonly referred to as the 'objective prism' technique) will produce spectra of all sources in the field with the spectral resolution defined by the instrumental configuration and the atmospheric seeing. This has the advantage over mulitcolor imaging in that all of the data are obtained at once as well as at considerably higher (≈50) spectral resolution at the cost of reduced sensitivity and higher source confusion (the spectrum from an individual object covers a much larger region of sky than does an image). Approximately one-sixth of the known high-redshift quasars have been found via this technique.
- **Variability.** It has been known since the discovery of the first quasars that they could exhibit spectacular variations in brightness; it is now believed that virtually all quasars display fluctuations of the 10–30% level on a timescale of years. Several high-redshift quasars have been identified by monitoring programs.
- **Serendipity.** In the summer of 1987 Patrick McCarthy and collaborators discovered a redshift 4.41 quasar (then the second largest redshift) while obtaining a spectrum

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a RADIO GALAXY (the quasar was near the radio source and was in the spectrograph entrance aperture). This supposedly once-in-a-lifetime event was repeated a few years later with the identification of a $z = 4.21$ quasar.

Although optical investigations have produced the vast majority of high-redshift quasar identifications, other regions of the electromagnetic spectrum may provide substantial contributions in the near future. In 1994 Patrick Henry and collaborators identified a high-redshift quasar based on x-ray observations with the ROSAT satellite (this was the first $z > 4$ object that was not discovered through optical observations), and in 1998 the x-ray discovery redshift limit was raised to 4.45 with the identification of an $R \approx 23$ quasar. The latter observation was made possible because of the high accuracy (2 arcsec) of the x-ray positions; this level of positional accuracy will be met or exceeded by future x-ray missions.

RADIO ASTRONOMY, which was the prime source of the first two decades of high-redshift quasar work, has for the most part contributed very little to the discovery of $z > 4$ quasars, but this too may be changing. In 1997 Isabel Hook and Richard McMahon found a $z = 4.72$ quasar in a radio survey; this object was later shown to be a strong x-ray source as well. Large-scale radio surveys (e.g. the FIRST project with the Very Large Array) currently underway will within the next few years undoubtedly produce a multitude of high-redshift quasar candidates.

The SLOAN DIGITAL SKY SURVEY, which will image one-quarter of the sky (the North Galactic Cap) in five optical filters, should have an enormous impact on our understanding of high-redshift quasars. The filters were selected so that a multicolor identification technique will identify high-redshift quasars brighter than an R magnitude of ≈22; this should produce a sample of more than 1000 $z > 4$ quasars, an increase of nearly a factor of 20 over the number of currently known high-redshift quasars.

Scientific programs

The next decade will see an enormous increase in the observational database on high-redshift quasars; what are the important scientific issues that will be addressed? Since these objects are seen at such large redshifts, the radiation we detect was created billions of years ago when the state of the universe was quite different from the current conditions; standard cosmological models predict that the most distant quasars are seen when the universe was considerably less than 10% of its current age, the density was approximately 200 times that of the current value, and the temperature of the cosmic background radiation field was 16 K.

- **Constraints on cosmological models.** Soon after the discovery of the $z > 4$ population, George Efstathiou, Martin Rees and Edwin Turner pointed out the consequences of the existence of black holes of a billion solar masses within a billion years of the Big Bang. These objects provide one of the key constraints on theories of galaxy formation.
- **Quasar evolution.** Within a few years of the discovery of quasars, it was shown that they were a rapidly evolving population of objects; their co-moving number density out to a redshift of one was proportional to $\approx (1 + z)^6$ (or, from a chronological viewpoint, the number density of quasars has dropped by a factor of 50–100 in the past several billion years). Patrick Osmer published a study in 1982 suggesting that this rise with redshift did not continue indefinitely; it appeared that beyond a redshift of 3.5 the number density of quasars actually declined. A number of recent investigations, using both photographic plates and CCDs and a variety of detection methods, indicate that the number density of luminous quasars peaked near redshift 3 (see figure 2).
- **The properties of intervening material.** One of the most important (if not the most important) areas of quasar research does not involve the properties of the quasars themselves but the ability to investigate a wide variety of types of non-luminous matter through absorption features in the quasar spectrum produced by material that lies along the line of sight. Since the quasar must obviously be at a greater distance than the material for absorption lines to be created, the higher the quasar redshift the further back in time we can investigate non-luminous material; we can now trace the properties of a number of types of features (e.g. the Lyman α forest, haloes of galaxies) throughout nearly the entire history of the universe.
- **Evolution of abundances of the elements.** Quasars possess emission lines of elements produced only in stellar nucleosynthesis (e.g. carbon, oxygen); one might expect that the metal abundances of quasars would

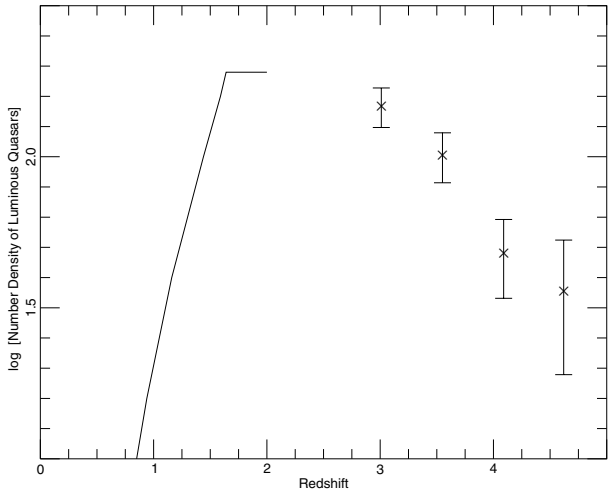


Figure 2. The number density of luminous quasars as a function of redshift based on low-redshift studies containing several hundred quasars (straight line, $z < 2$) and a $z > 2.7$ slitless survey with 90 quasars (points).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

decrease with increasing redshift. An interesting aspect of quasar spectra is that, if one ignores the region shortward of the Lyman α emission line, where the high-redshift Lyman α forest removes a majority of the radiation, it is impossible to make any estimate of the quasar redshift if the spectrum is translated to the rest frame; in other words, there appears to be no evolution in the strengths of the emission lines. Quasar abundance determinations are notoriously difficult, but the best estimates suggest that even in the highest-redshift quasars the metal abundances are approximately ten times the solar value.

- **Detection of distant clusters of galaxies.** If high-redshift quasars mark the sites of substantial density perturbations, as seems plausible, it may be possible, with substantial efforts of observing time, to detect 'normal' galaxies that are physically associated with the quasar. Some preliminary studies have already yielded intriguing results.
- **Gravitational lenses.** The longer the path length to an object, the greater its chance of being gravitationally lensed; hence high-redshift quasars are attractive candidates for gravitational lenses. If a multiply imaged high-redshift quasar is found where the lens can be identified and characterized, one will have a test of General Relativity on the largest imaginable scale!

The first decade of $z > 4$ research has revealed much about the conditions of the universe in its first billion years. Nature has been able to assemble massive objects in that time, but we have clearly seen past the peak of quasar activity (at least in the rest frame optical and ultraviolet). At redshifts above 4 the absorption due to the Lyman α forest removes more than half of the incident radiation, and this opacity continues to increase at the limits of our measurements. These discoveries, and other results summarized in the bibliography, have been produced by a sample of a few dozen objects. The next ten years should witness a dramatic increase (more than a factor of ten!) in the number of known high-redshift quasars given the capabilities of planned optical surveys and the recent demonstration that both x-ray and radio observations can be effective methods for detecting $z > 4$ objects.

Bibliography

Excellent technical reviews of the first years of the $z > 4$ era are presented by

Turner E L 1991 *Astron. J.* **101** 5
Warren S J and Hewitt P C 1990 *Rep. Prog. Phys.* **53** 1095

Papers describing some milestones in detection of distant quasars include

Henry J P *et al* 1995 *Astron. J.* **107** 1720
Hook I M and McMahon R G 1998 *Mon. Not. R. Astron. Soc.* **294** L7
Schmidt M 1963 *Nature* **197** 1040
Schmidt M 1965 *Astrophys. J.* **141** 1295
Schneider D P, Schmidt M and Gunn J E 1991 *Astron. J.* **101** 837
Warren S J *et al* 1987 *Nature* **325** 131

Technical References for all objects in table 1 are given in

Schneider D P, van Gorkom J H, Schmidt M and Gunn J E 1992 *Astron. J.* **103** 1451

The two serendipitous discoveries are described in

McCarthy P J *et al* 1988 *Astrophys. J. Lett.* **328** 29
Schneider D P, Schmidt M and Gunn J E 1994 *Astron. J.* **107** 880

Cosmological implications of high-redshift quasars are discussed in

Efstathiou G and Rees M 1988 *Mon. Not. R. Astron. Soc.* **230** 5P
Turner E L 1990 *Astrophys. J. Lett.* **365** 43

A sample of investigations of the high-redshift quasar luminosity function (using a variety of techniques)

Hawkins M R S and Véron P 1996 *Mon. Not. R. Astron. Soc.* **281** 348
Kennefick J D, Djorgovski S G and de Carvalho R R 1995 *Astron. J.* **110** 2553
Osmer P S 1982 *Astrophys. J.* **253** 28
Schmidt M, Schneider D P and Gunn J E 1995 *Astron. J.* **110** 68
Warren S J, Hewitt P C and Osmer P S 1994 *Astrophys. J.* **421** 412

Some representative articles that discuss absorption in high-redshift quasars are

Lu L, Sargent W L W, Womble D S and Takada-Hidai M 1996 *Astrophys. J.* **472** 509
Lynds C R 1971 *Astrophys. J. Lett.* **164** 73
Schneider D P, Schmidt M and Gunn J E 1991 *Astron. J.* **101** 2004
Storrie-Lombardi L J, McMahon R G, Irwin M J and Hazard C 1996 *Astrophys. J.* **468** 121

A study of the abundances in high-redshift quasars can be found in

Hamann F and Ferland G 1993 *Astrophys. J.* **418** 11

Recent work on galaxies surrounding quasars is presented by

Giallongo E *et al* 1998 *Astron. J.* **115** 2169

Donald P Schneider

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High-velocity Cloud (HVC)

An interstellar cloud of neutral hydrogen whose velocity exceeds, typically by 100–200 km s^{-1}, the rotational speed of the Galaxy (which is 220 km s^{-1}). Some HVCs are thought to have been produced by gravitational interactions—most notably the Magellanic Stream, a huge bridge of material between the Galaxy and the Magellanic Clouds, presumed to have been drawn out during a close passage in the past. Other HVCs may be clouds of intergalactic hydrogen falling into the Galaxy, or material expelled out of the galactic plane as supernova remnants and now falling back.

See: Magellanic Stream.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

High-velocity Star

A star that is moving with very high velocity relative to the Sun, usually 65 km s^{-1} or more, faster than the average velocity of stars in the solar neighborhood. High-velocity stars are usually part of the galactic halo and move in very elliptical orbits which are highly inclined to the galactic plane. Hence they do not share the revolution of the Sun and other stars in the spiral arms of the Galaxy around the galactic center and are being left behind as the Sun races ahead of them. They may be remnants of a very early stage in the evolution of the Galaxy, so high-velocity stars are runaway stars.

See also: Runaway star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hill, George William (1838–1914)

Celestial mechanician, staff member at the Nautical Almanac Office then in Cambridge, MA, computed the orbit of the Moon and, under SIMON NEWCOMB's direction, solved the difficult problem of the orbits of Jupiter and Saturn.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hind, John R (1823-1895)

English astronomer, discovered in 1852 a small nebula in Taurus which, in 1861, was found by HEINRICH D'ARREST to have disappeared. By the end of the year, D'Arrest and OTTO WILHELM STRUVE had recovered it. *Hind's variable nebula*, as it came to be called, demonstrated that at least some nebulae were small, as nothing larger than a light year in dimension can disappear in a year. The nebula is a reflection nebula which comes and goes as dark clouds intercept the light from its illuminating star. Hind discovered 11 asteroids, Nova Ophiuchus 1848, and became Superintendent of the Nautical Almanac Office.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hinotori (Astro-A)

Japanese satellite, launched in February 1981 to observe the Sun during the peak of its 11-year cycle. Carried instruments to image x-rays from solar flares and detect x-ray emission lines from highly ionized iron in solar flares. Hinotori means 'firebird'. (See also ASTRO.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hipparchus

Hipparchus (2nd century BC) was the most highly regarded Greek astronomer in antiquity before Ptolemy. In a series of writings, only one of which has come down to us, he investigated a wide range of problems in observational and mathematical astronomy. In his most important works he studied the periodicities of the motions of the heavenly bodies and used observations to test and quantify geometrical models for those motions.

Biography and writings

Hipparchus was born at Nicaea in Bithynia (Iznik, Turkey) within the first half of the 2nd century BC. While living in Bithynia, he recorded observations of weather correlated with the solar year and the risings and settings of fixed stars. (Such weather calendars were produced by many Greek astronomers from the 5th century BC onwards.) The earliest dated observations that he appears to have made—there is some disagreement about whether they should be attributed to him—are determinations of the date of the autumnal equinox in 162–158. After this there is an 11 yr gap, and then about 25 observation reports spread through the years 147–127. These reports are known to us only because they are cited in Ptolemy's *Almagest* (see PTOLEMY). Ptolemy informs us that Hipparchus was on the island of Rhodes when he made observations in 141 and 128–127, and there are plausible arguments for supposing that he resided there through most of his working life. He was in contact with astronomers in Alexandria, and very likely also with those of Babylon (see BABYLONIAN ASTRONOMY). It is not known how long he lived after 127 BC.

Late in his career Hipparchus published an annotated list of his writings, which must therefore have been numerous; most, however, seem to have been comparatively short and dealt with specialized topics. In addition to the surviving *Commentary on the Phainomena of Aratus and Eudoxus* we know titles of about 14 books. Almost all the works that we know something about were connected with astronomy, although there exist tantalizing references to his investigations of falling bodies and combinatoric mathematics. A highly critical, even polemical, tone characterized certain of his writings (the *Commentary on the Phainomena*, the *Against the Geography of Eratosthenes* and the work on planetary models). Ptolemy, who among all the classical authors who praise Hipparchus was the most competent to judge him, singles out Hipparchus for his 'love of truth' and diligence.

Models and periodicities

In his studies of solar and lunar motion, Hipparchus assumed that each of these bodies has only one periodic anomaly and travels according to a simple kinematic model of either the epicyclic or eccentric variety (see PLANETARY THEORY FROM EUDOXUS TO COPERNICUS). In the epicyclic model, the body revolves uniformly about an epicycle, that is, a circle the center of which itself revolves uniformly on a larger deferent circle concentric with the Earth. In the eccentric model, the body revolves uniformly about an eccenter, that is, a circle surrounding the Earth but with its center at a fixed distance from the Earth's center; the direction from the Earth's center to the eccenter's center may be either fixed or uniformly rotating. The two kinds of model are, as Hipparchus knew, interchangeable. In various works Hipparchus attempted to determine the periodicities of the several circular motions presumed in these models and the relative sizes and orientations of the radii.

The approach that Hipparchus used to establish periodicities can be illustrated by his determination of the length of the mean synodic month and the Moon's period of anomaly. Hipparchus started out with a relation that was embedded in the Babylonian lunar theory known as System B:

$$251 \text{ synodic months} = 269 \text{ anomalistic periods.} \quad (1)$$

This relation can be read as saying that any interval of 251 synodic months bounded by a pair of conjunctions or oppositions has exactly the same length. Since the only direct way of observing the precise moment of a syzygy was at a lunar eclipse, Hipparchus found the smallest multiple of 251 synodic months that also brings the Moon roughly from one of its nodes back to one of its nodes, which is 4267 synodic months (or just short of 345 yr).

Conveniently, Hipparchus had access to records of observations of lunar eclipses made at Babylon this long before his time. To compare them with his own eclipse observations, however, would not have been easy. The Babylonian records would have been dated according to the regnal years of the Persian kings who ruled Mesopotamia, with dates in an observational lunar calendar and probably only the time of the beginning of obscuration noted. Hipparchus must have possessed a king list giving lengths of reigns, but he cannot have known the precise sequence of 29 and 30 day Babylonian lunar months over the several intervening centuries. He did, however, have a value for the length of the mean synodic month, again from the Babylonian system B theory:

$$1 \text{ mean synodic month} = 29;31,50,8,20 \text{ days} \quad (2)$$

(the numerals after the semicolon are to be read as sixtieths, sixtieths of sixtieths, etc). Hipparchus evidently had enough confidence in equation (2) to conclude that the resulting equation

$$4267 \text{ mean synodic months} = 126\,007 \text{ days} + 1 \text{ h} (+ \text{ less than } 4 \text{ min}) \quad (3)$$

was at least accurate to the nearest day, so that it would only be necessary to compare the times of observed eclipses 345 yr apart. To test for the constancy of the time interval, he found two pairs of observed eclipses such that in one pair the Moon was both times close to its maximum

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

apparent speed, and in the other pair close to its minimum apparent speed, so as to maximize any discrepancy. In comparing the times, Hipparchus had to estimate the duration of the eclipse observed in Babylon to find the mid-eclipse and take account of the geographical situations of Babylon and Rhodes to synchronize the times. He even introduced a correction for the small influence of solar anomaly. The result of this carefully thought out analysis was a confirmation of the accuracy of the Babylonian period relations (1) and (2). By a similar investigation of the magnitudes of widely separated eclipses, Hipparchus also confirmed a Babylonian value for the Moon's mean period of latitude.

To measure the length of the tropical year, Hipparchus attempted to determine the intervals separating three observations of the summer solstice, by Meton and Euctemon in 432 BC, by Aristarchus in 280 BC, and by himself in 135 BC. For the oldest of these observations, which was originally dated according to a lunar calendar, it is probable that Hipparchus had to determine the date in his own calendrical system by assuming a first approximation of the length of the year, and the date that he found looks as if it is a day too early. The interval that he found between Aristarchus' observation and his own was $52\,960\frac{3}{4}$ days, implying a tropical year of approximately 365;14,47,35 days, whereas between the Metonic observation and his own he appears to have counted 108 478 days, implying a tropical year of approximately 365;14,44,51 days. Remarkably, the latter value, which can only have arisen from comparison of observations at precisely the 297 year interval, is attested in a Babylonian tablet (which thus is strong evidence for Babylonian contacts with Hipparchus or someone who knew his work). On the other hand, Ptolemy cites several passages in writings of Hipparchus where he elects for the value 365;14,48 days. Ptolemy also reports that Hipparchus was led by his observations to suspect that the tropical year was subject to small short-term fluctuations.

Hipparchus' most famous accomplishment, the discovery of the PRECESSION of the EQUINOXES, was made in connection with these last measurements. During lunar eclipses, Hipparchus observed the elongation of the Moon from certain fixed stars close to the ecliptic, and from this he was able to deduce the elongation of the stars from the equinoctial points. As well as small inconsistencies, which he was ready to attribute to a wobble in the equinoctial points, he also found that the elongations were about 2° less in his own time than about 150 yr earlier, when Timocharis had performed similar observations. Hipparchus was uncertain of the nature of this displacement, hypothesizing among other things that the fixed stars close to the ecliptic might exhibit a very slow quasi-planetary motion of their own. Ptolemy was able to find passages in which Hipparchus seemed to maintain the model that Ptolemy himself adopted, that the fixed stars make a slow revolution about the poles of the ecliptic relative to the equinoctial points.

Hipparchus was the first known astronomer to deduce the eccentricity of a hypothetical orbit from selected observations. His method required three observations of the longitude of a heavenly body at known times; with approximate values for the periods of longitude and anomaly of the object in question, he could reduce the problem to the geometrical problem of finding the point from which the sides of a given triangle are seen as subtending given angles. A numerical solution further necessitated some sort of trigonometry, and Hipparchus is believed to have constructed a table of the lengths of chords of angles in a standard circle, perhaps the first such table.

In the case of the Sun, observations of the dates of two equinoxes and a solstice provided the required three positions, since, if the Sun travels on a great circle (the ecliptic), the longitudes of the tropical and equinoctial points must be separated by right angles. Taking the interval from vernal equinox to summer solstice as $94\frac{1}{2}$ days, and the interval from summer solstice to autumnal equinox as $92\frac{1}{2}$ days, Hipparchus calculated that the eccentricity of the Sun's hypothetical circular orbit was 1/24 of its radius, with the apogee $65\frac{1}{2}^\circ$ east of the vernal equinoctial point. This is the version of the solar model that Ptolemy later adopted. There is good reason to believe, however, that Hipparchus also performed the same calculations with at least one other pair of intervals ($94\frac{1}{4}$ days from vernal equinox to summer solstice, and $92\frac{3}{4}$ days to the autumnal equinox), hence obtaining a different eccentricity and apogee. What Hipparchus made of this is not known.

Similarly, Hipparchus measured the lunar anomaly twice, with conflicting results. In this instance, he used sets of three observed lunar eclipses, for which it was possible to estimate the Moon's longitude by calculating the diametrically opposite position of the Sun. Ptolemy later showed that the discrepancy arose from inaccuracies in Hipparchus' solar longitudes and time intervals. Again, we have no knowledge of how Hipparchus reconciled his results.

Ptolemy knew of no writings by Hipparchus on the planets comparable with his works on the solar and lunar models. He does, however, describe a book in which Hipparchus gave observational arguments for the inadequacy of the planetary models proposed in his time. According to Ptolemy, the critical defect Hipparchus exposed in these models was that they assumed a single periodic anomaly for each planet, so that, for example, their retrogradations were always of the same size and duration.

Other astronomical research

Hipparchus devised two methods of estimating the distance in Earth radii of the Moon from the Earth's center. The first method assumed that the distance of the Sun is effectively infinite, and finds the distance of the Moon by comparing observations of the magnitude of a solar eclipse seen in two localities on the same

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

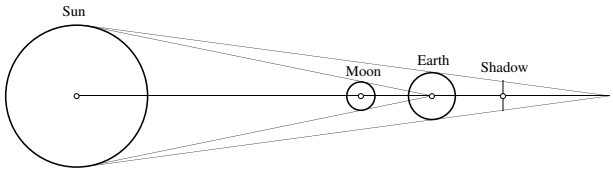


Figure 1. Hipparchus' eclipse diagram relating the Sun's and Moon's distance from the Earth.

meridian. The second method uses the eclipse diagram made familiar by Ptolemy to relate the distance of the Moon to a hypothesized solar distance. In this diagram (figure 1) it is assumed that the Sun and Moon subtend the same angle from the Earth's center and that the ratio of the diameter of the Moon to the diameter of the Earth's shadow at the Moon's distance is known. From these relations it is theoretically possible to derive the solar distance from the lunar distance and vice versa, although in practice small variations in the lunar distance correspond to extremely large variations in the solar distance. Assuming that the Sun is 490 Earth radii from the Earth's center (corresponding to a 'least perceptible' parallax of 7 arcmin), Hipparchus found the mean distance of the Moon as $67\frac{1}{3}$ Earth radii, while from the solar eclipse he found it to be 77 Earth radii.

With an approximate value for the distance of the Moon, Hipparchus was able to determine the magnitude of lunar parallax and correct observed positions accordingly. Regrettably little is known of the details of Hipparchus' handling of this topic. Of his work on the theory of eclipses, our most substantial information concerns his examination of the possible time intervals separating lunar and solar eclipses.

The most controversial part of Hipparchus' astronomical work is his star catalogue. His one surviving book, the *Commentary on the Phainomena of Aratus and Eudoxus*, is devoted to a detailed criticism of the detailed nonmathematical description of the constellations given by Eudoxus in the 4th century BC, and versified by Aratus a century later. The *Commentary* closes with a presentation of data concerning the risings, culminations, and settings of constellations, all of which presupposes numerous observations of stellar positions. Classical authors, particularly Pliny, make vague references to a Hipparchian star catalogue containing positions of the stars along with some indication of their brightnesses. Since Tycho, arguments have repeatedly been put forward, and repeatedly disputed, that the catalogue in Ptolemy's *Almagest* is an unacknowledged substantial reproduction of Hipparchus', with the longitudes adjusted to reflect the precessional motion between Hipparchus' time and Ptolemy's. At present the balance of evidence seems to preponderate in favor of the hypothesis that Ptolemy depended heavily on Hipparchus' catalogue.

Bibliography

Neugebauer O 1975 *A History of Ancient Mathematical Astronomy* (Berlin: Springer)

Toomer G J 1978 Hipparchus *Dict. Sci. Biogr.* **15** 207–24

Toomer G J 1984 *Ptolemy's Almagest* (London: Duckworth)

Alexander Jones

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hipparcos

The Hipparcos space astronomy mission was dedicated to the very accurate measurement of star positions, and their distances and motions in space. The satellite was designed, built and operated under the responsibility of the EUROPEAN SPACE AGENCY (ESA), with the guidance of an international scientific advisory committee. Launched in August 1989 and operated until mid-1993, two resulting star catalogs were completed and published in 1997: the Hipparcos Catalog is a high-accuracy compilation of nearly 120 000 star positions, distances and space motions; the Tycho Catalog is a lower-accuracy catalog of slightly more than one million stars. Together they represent astronomers' best understanding of the space distribution, space motion, and corresponding physical properties of stars in the solar neighborhood. The experiment was carried out within a large collaboration of about 200 mainly European scientists.

Background

Catalogs of star positions have been compiled for more than two thousand years. The name of the ESA mission, as well as being an acronym for 'high precision parallax collecting satellite', reflects the contribution of the ancient Greek astronomer HIPPARCHUS who prepared a naked-eye catalog of about 1000 stars around the first century BC.

Over the past few hundred years, improved instruments and observational techniques have led to star catalogs of continuously improving accuracy and size. This astronomical discipline, technically referred to as ASTROMETRY, has led to a greatly improved picture and comprehension of our place in our Galaxy. Directly and indirectly the accurate measurement of star positions has provided profound insights into the content and structure of the universe as a whole.

Star position measurements in the 17th century were directed at compiling catalogs to assist sea navigation and the determination of longitude, but measurements were also aimed at furthering scientific understanding of the Earth's motion in space. During the 18th and 19th centuries positional improvements led to the detection of small changes in the relative positions of stars, attributed to the motions of the stars themselves through space. The realization that stars lie at finite, albeit very large, distances from our own solar system led to intensified efforts at measuring the effect of PARALLAX, which is the apparent displacement of nearby stars compared to more distant stars as the Earth moves in its orbit around the Sun. Because of their enormous distances, the resulting angular displacements are very small, and even the nearest stars have parallaxes of only about 1 arcsec ($1/3600^\circ$). For more distant stars the effect is much smaller. Consequently, to measure stellar distances even within our Galaxy, angular accuracies very much better than 1 arcsec are required.

Similar considerations apply to the measurements of PROPER MOTION, which are the changes in angular position on the sky with time resulting from the projection, on the plane of the sky, of the star's motion through space. While the largest proper motions reach about 10 arcsec per year, typical stellar motions within our Galaxy, of the order of several tens of km s^{-1}, combine with their considerable distances to result in typical angular positional motions of only a few hundredths of an arcsec per year. Ground-based proper motion measurements can often discern these tiny motions, in particular by extending the observational baseline over decades. Nevertheless, reaching very high proper motion accuracies, with minimal systematic errors, faces problems similar to those encountered for high-accuracy parallax work.

Because of the size of the effects involved, the measurement of stellar parallaxes from the ground has always been very difficult, and complicated primarily by the perturbing effects of the Earth's atmosphere. Measurements are made relative to an average parallax of more distant background stars, and an estimate of this background parallax is accounted for in the correction of the measured (relative) parallax to an absolute value.

In the 1960s an alternative approach for measuring parallaxes was developed, with particular emphasis on obtaining absolute rather than relative values. The basic idea was to combine the images from two areas of sky onto one detector, and to do this at different times of the year. The parallactic displacements for the two fields would then be uncorrelated, and relative measurements would therefore circumvent the parallax zero-point problem. It was soon realized that such technology could not be employed from the ground, as it required a stable and precisely known angle between the two viewing directions, which was affected by variable atmospheric refraction, along with gravitational and thermal effects. The first ideas for implementing this technique in a space mission were submitted by P Lacroute in France in 1966. The implementation appeared too complex at that time, in particular with respect to the construction of the mirror necessary to combine the two viewing directions, and in 1970 further studies for a space mission were suspended.

The basic idea, however, survived and with the input of several new ideas, the concept of a space astrometry mission became more realistic. A meeting on space astrometry in Frascati in 1974 marked the start of what was to develop into the Hipparcos project. Its official status as an ESA study was obtained in 1976, and growing recognition of the fundamental importance of being able to obtain greatly improved stellar distances and space motions led to the acceptance of the Hipparcos mission within ESA's scientific programme in 1980.

For the 120 000 stars observed and contained within the Hipparcos Catalog, mean errors in positions, parallaxes, and annual proper motions are around 0.001 arcsec, a value which yields distance accuracies of 10% for stars at a distance of 100 pc. These numbers can be compared with the relatively small number (about 8000 stars in the 1995 Yale Parallax Catalog) of pre-Hipparcos parallaxes accumulated through a century of astrometric work, and with the lower precision

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(5–10 milliarcsec at best) of the classical photographic parallaxes. Modern ground-based electronic parallax observations (e.g. from the US Naval and Allegheny observatories) are comparable to Hipparcos both in precision and accuracy, but only a few hundred stars, mostly quite faint, have been observed to date. Hipparcos therefore provides the unique source of large numbers of high-quality parallaxes for bright stars; these are the data needed for most luminosity calibration and distance-scale problems.

As the first satellite devoted to star position measurements, the scientific interest in the Hipparcos results has led to plans for future space missions aiming at significantly higher angular accuracies, of a few millionths of a second of arc, and for a very much larger number of stars.

Operational principle

The Hipparcos satellite was built by a large European industrial consortium, led by Matra Marconi Space and Alenia Aerospazio, under contract to ESA. The satellite comprised a spacecraft platform, including the power, telecommunications, attitude and thermal control subsystems, and the payload containing the optical telescope and detector chains (table 1). The satellite was launched by an Ariane 4 rocket, flight V33, on 8 August 1989, and was operated from ESA's satellite operations center at ESOC (Darmstadt, Germany) through ground stations at Odenwald (Germany), Perth (Australia), Goldstone (USA), and Kourou (French Guyana).

Table 1. Instrument characteristics.

Optics	
Telescope configuration	All-reflective Schmidt
Field of view	$0.9^\circ \times 0.9^\circ$
Separation between fields	58°
Primary mirror diameter	0.29 m
Focal length	1.4 m
Scale at focal surface	6.8 μm arcsec^{-1}
Mirror surface accuracy	$\lambda/60$ rms at 550 nm
Primary detection system:	
Modulating grid	2688 slits
Slit period	1.208 arcsec (8.2 μm)
Detector	Image dissector tube
Photocathode	S20
Scale at photocathode	3.0 μm arcsec^{-1}
Sensitive field of view	38 arcsec diameter
Spectral range	375–750 nm
Sampling frequency	1200 Hz
Star mapper (Tycho) system:	
Modulating grid	4 slits at 90° to scan
	4 slits at $\pm 45^\circ$ to scan
Detectors	Photomultiplier tubes
Photocathode	Bi-alkali
Spectral range (B_T)	$\lambda_{\rm eff}/\Delta\lambda = 430/90$ nm
Spectral range (V_T)	$\lambda_{\rm eff}/\Delta\lambda = 530/100$ nm
Sampling frequency	600 Hz

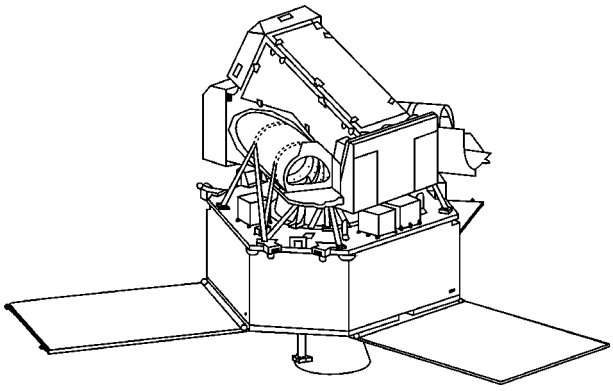


Figure 1. The Hipparcos satellite configuration with the three solar array panels deployed (seen at bottom). The hemispherical apertures define the two viewing directions. The satellite rotated about the vertical axis.

Originally intended to be stationed in a geostationary orbit, the apogee boost motor intended to circularize the orbit failed to function, and the mission was conducted from a highly elliptical transfer orbit. Apart from a considerable increase in operational complexity and a reduced fraction of useful observing time, the unplanned orbit had a relatively small effect on the data quality, and all of the mission's original goals were comfortably surpassed. Nevertheless, the high particle radiation levels in the operational orbit led to a progressive deterioration of certain satellite elements (gyroscopes, thermal control, telecommunications). Useful observations ceased in March 1993, and satellite operations were terminated on 15 August 1993. More than three years of useful data had been acquired, exceeding the 2.5 yr design goals.

The satellite comprised a small but highly accurate optical telescope, with a primary mirror diameter of about 0.3 m, and with two distinct viewing directions separated by about 58° (figure 1). A beam combiner mirror brought the light from the two viewing directions to a common focal surface. A highly accurate grid system in the focal surface modulated the star images, and the modulated light was sensed by a detector located behind the grid. The measurement principle was similar to that used in ground-based meridian instruments, the purpose of both types of instrument being the determination of accurate transit times. In the case of meridian telescopes the stellar images pass through a set of slits as a result of the Earth's rotation. In the case of Hipparcos, the passages were the results of the satellite rotation, which was maintained at a roughly constant spin rate of approximately 168 arcsec s^{-1} through the periodic firing of 20 mN gas jet thrusters. These were activated in order to maintain the required attitude, itself sensed by the passage of bright stars across the satellite's star mappers. The near simultaneity of measurements obtained in the two fields of view, together with the satellite scanning, made it possible to obtain a very accurate calibration of the angle between the two viewing directions, using closure conditions on the observed great

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

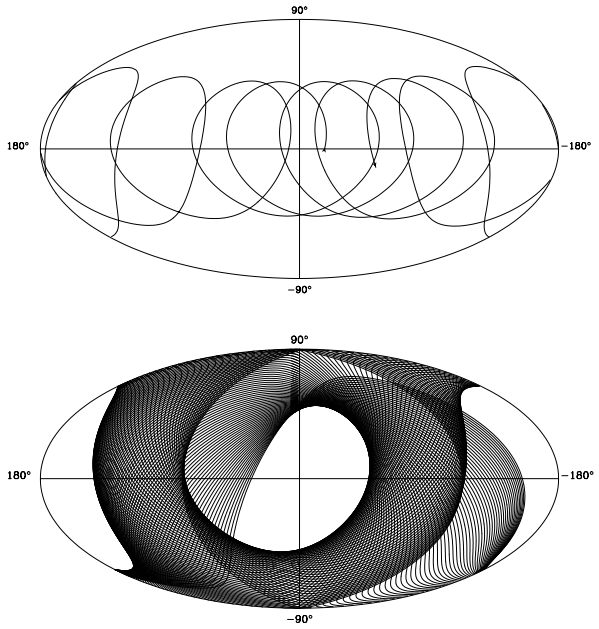


Figure 2. The scanning motion of the Hipparcos satellite on the celestial sphere, shown in ecliptic coordinates. Top: the path of the spin axis between 22 May 1990 and 23 September 1991. Bottom: the two viewing directions swept out great circles on the sky: one out of five of every great circle scan is shown for the interval 22 May 1990 to 14 July 1990 (courtesy F van Leeuwen).

circle. As a result, the very precise relative separations measured in the $0.9^\circ \times 0.9^\circ$ combined field of view could be transformed into very precise large angular separations on the sky.

Scanning the entire sky in a continuous and regular manner, whilst maintaining a constant inclination of the spin axis to the Sun direction for thermal control reasons, the observations were used to build up a network of one-dimensional angular measurements between the stars (figure 2). Scanned at different orientations and at many different epochs throughout an operational lifetime of about 3 yr, the on-ground processing then led to the reconstruction of the positions, distances, and proper motions of each star.

The two widely separated viewing directions led to the construction of a very rigid positional frame, and to the possibility of determining absolute parallaxes, in which the stellar distance determinations are independent of the assumed distance to other stars. The resulting network of star positions and proper motions was, however, not directly connected to other independent reference points, but was attached to a quasi-inertial system through the linking of the resulting star positions to those of extragalactic objects, especially extragalactic radio sources. As a result, the Hipparcos reference frame is the optical realization of the International Celestial Reference System (ICRS): the catalog positions are on the same reference system as VLBI radio measurements, and the proper motions are inertial.

Satellite observations were conducted on the basis of an 'input catalog', the compilation of target stars to be observed throughout the mission. Selected stars were chosen on the basis of their scientific importance (as perceived during the years before launch), taking into account operational and data reduction requirements of a reasonably uniform distribution of stars on the sky, with a surface density of about 3 stars per square degree. The magnitude distribution down to the observational limiting magnitude of about 12 mag also had to be consistent with the available observing time as a result of the regular scanning motion of the satellite.

Stellar images modulated by the grid in the focal surface were sampled by an image dissector tube's sensitive area, pointed sequentially to the star positions defined by the input catalog, and sampled at 1200 Hz. A data rate of 24 kb s^{-1} yielded a data stream distributed to and analysed by three European scientific consortia. The data analysis tasks were lengthy and complex, and proceeded from the raw satellite data, using the attitude determination on the basis of the satellite's sky mappers, to yield one-dimensional positions of the stars along a reference great circle at the epoch of observation.

Data analysis

The interconnection of all of the reference great circle data obtained throughout the mission, during which the sky was scanned at many different epochs, and at many different position angles, eventually yielded the two-dimensional positions of each star (in right ascension and declination), the time-dependent proper motion of each star due to its motion through space, and the additional annual elliptical component of motion due to the Earth's motion around the Sun, which provides the stellar distances. The radial component of the space velocity of each star (which can be obtained spectroscopically from the Doppler shift of the star's spectrum), needed to complete the full six-dimensional Cartesian coordinate location of each star (the three positional components and the three velocity components), was not measured by the Hipparcos satellite.

The entire data analysis chain, from the determination of one-dimensional great circle coordinates and their interconnection through the 'sphere reconstruction' step, was iterated a number of times to achieve full consistency and convergence. Results were verified through extensive cross-checks of the results of two parallel reduction efforts before publication. Median accuracies, around 1 milliarcsec (mas), depend primarily on the star magnitude (the Hipparcos measurements are in a broad-band system, referred to as Hp), and on its ecliptic latitude, due to the form of the scanning law (figure 3).

The more complex observational behavior of double and multiple stars and astrometric binaries were subjected to specific analysis procedures. In the case of resolved double or multiple systems, the results yielded accurate angular separations, position angles, and magnitude differences at the various epochs of observation, and these

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

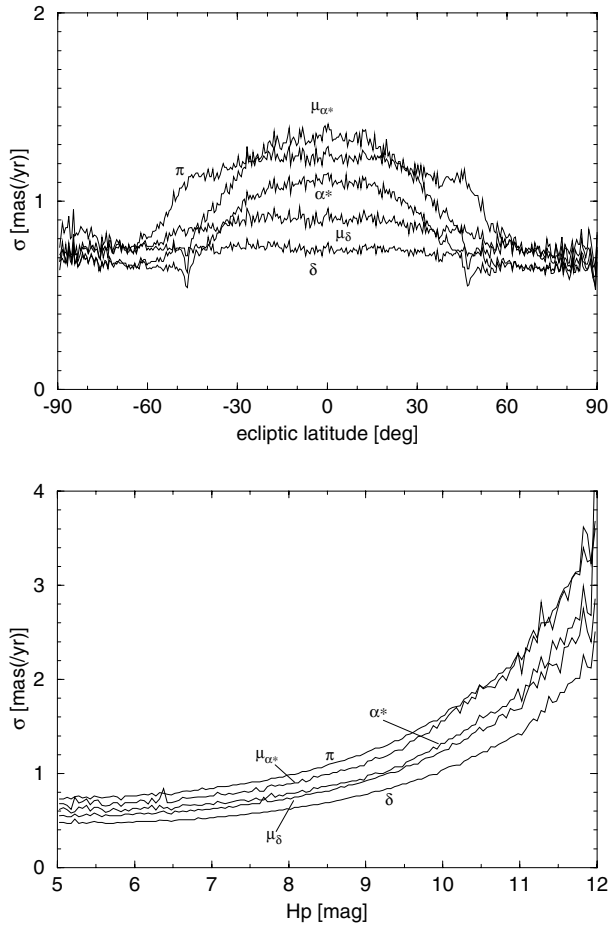


Figure 3. Hipparcos Catalog errors as a function of position and magnitude. Top: median standard errors of the five astrometric parameters as a function of Hp magnitude. The unit of the standard error is milliarcsec (mas) for the positional components (α, δ) and parallax (π), and mas yr^{-1} for the proper motion components ($\mu_{\alpha*} = \mu_\alpha \cos\delta$, μ_δ). Bottom: median standard errors of the astrometric parameters as a function of ecliptic latitude. The dependency on ecliptic latitude is a consequence of the ecliptic-based scanning law. The errors are given at the catalog epoch, J1991.25 (courtesy H Schrijver).

results were compiled into a specific annex of components of 12 195 double and multiple systems. A further 235 orbital systems, 2910 astrometric binaries (recognized through the nonlinear motion of their photocenters), and 8542 suspected non-single objects make up the total of 23 882 solved or suspected double or multiple systems classified in the Hipparcos Catalog.

Although the primary objective of the mission was the determination of accurate astrometric data, the positional results were obtained from a modulated stellar signal whose careful analysis and calibration yielded a substantial database of homogeneous photometric data which is contained in annexes to the main Hipparcos Catalog. Each star was observed an average of about 110 times throughout the 3 yr mission, yielding a median broad-band photometric precision of about 0.0015 mag for stars brighter than 9 mag. The results provide extensive data on stellar variability. As a result of the photometric analysis undertaken in advance of the catalog publication, 11 597 stars were classified as variable, of which 8237 were newly identified as such. Of these, 2712 stars were classified as periodic variables, with light curves provided in the catalog annexes, and 5542 non-periodic and unsolved variables were classified, of which 4145 were newly identified. These results have underlined the quality of photometric data which can be acquired above the Earth's atmosphere, and have led to greatly improved categorization of stellar luminosity and variability as a function of position in the HERTZSPRUNG–RUSSELL DIAGRAM.

The Tycho Catalog of more than one million stars, of lower astrometric accuracy than that of the main catalog, was constructed from the data derived from the satellite's sky mappers. It provides a positional reference catalog with about 25 stars per square degree, with a limiting magnitude of about $V = 11.5$ mag, completeness at about $V = 10.5$ mag, and a median precision at the catalog reference epoch of about 25 milliarcsec. The median astrometric precision is about 7 milliarcsec for stars brighter than 9 mag, and systematic errors are expected to be below about 1 milliarcsec.

Although the majority of objects contained within the Hipparcos Catalog are stars within our Galaxy, about 45 are bright stars in the Magellanic Clouds. In addition, a particular effort was devoted to obtaining observations of all solar system objects which could be observed by the satellite, taking into account the observing constraints of magnitude, angular size, and proximity to nearby bright objects (planets). In total, 48 minor planets and three planetary satellites (J II Europa, S VI Titan, S VIII Iapetus) were observed by the main experiment. Observations of solar system objects were complicated by their non-point-like nature, which affects the phase estimates (due to the varying solar aspect angle) and their resulting relation to stellar positions. Additionally, their rapid motion on the sky called for specific treatment of the data during their crossing of the instrument's focal plane.

Following the scientific analysis phase, the European Space Agency published the results of the project on behalf of the scientific collaborators in June 1997: the Hipparcos Catalog is a 17-volume publication containing the positions, distances, proper motions, and photometric, variability, and multiplicity properties of the more than one hundred thousand stars brighter than about 12 mag observed by the satellite (see HIPPARCOS AND TYCHO CATALOGUES). Volume 1 is an introduction and guide to the catalog data, including sky distributions illustrating statistical properties of the data. Volumes 2–4 document the satellite measurement and operational principles, and the construction of the Hipparcos and Tycho Catalogs respectively. Volumes 5–9 contain the main Hipparcos Catalog, with annexes of double and multiple star data

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

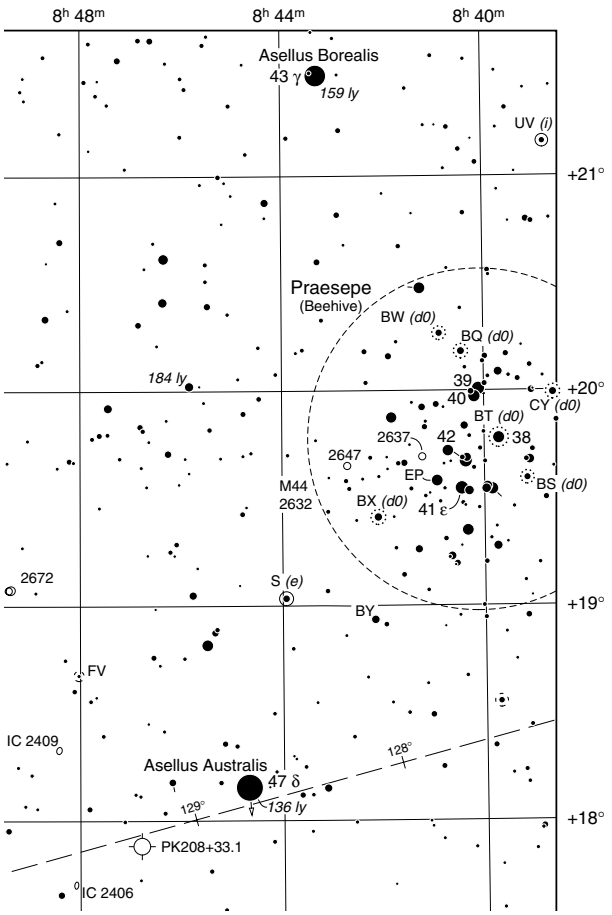


Figure 4. The Hipparcos and Tycho Catalogs are the basis for Sky Publishing Corporation's three-volume *Millennium Star Atlas*, published in 1997. Each square represents an area of 1 square degree. Graphical representation of magnitudes and variability, and of nearby and high-proper motion stars, is included in the Atlas. This example is from Chart 712, and includes the Praesepe star cluster (courtesy Sky Publishing Corporation and ESA).

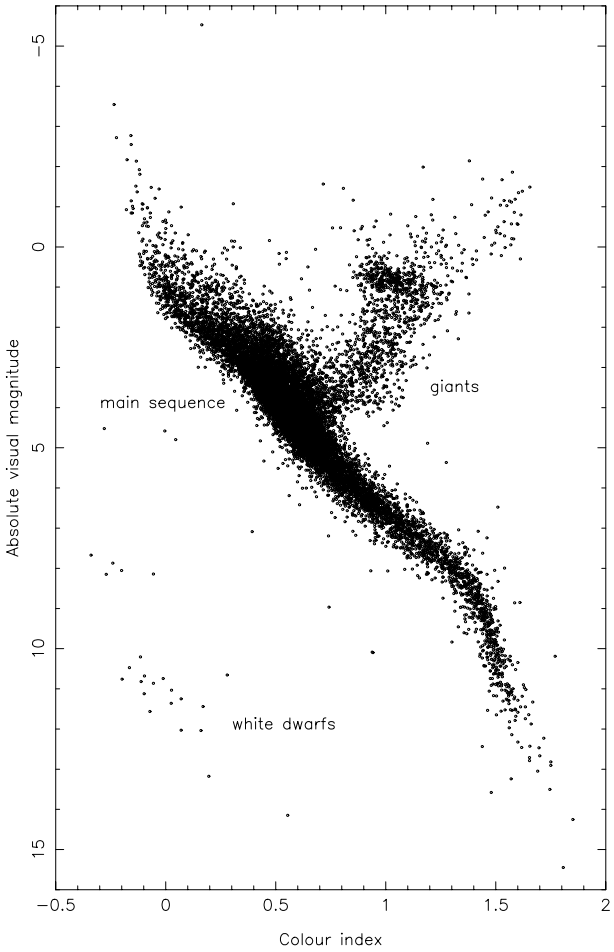


Figure 5. The Hertzsprung–Russell diagram (absolute visual magnitude, M_V, versus color index $B - V$) for about 14 000 single stars within a sphere of radius 100 pc centered on the Sun, taken from the Hipparcos Catalog. The few objects lying between the main sequence and the white dwarf sequence mostly arise from color index errors in the catalog.

(Volume 10), variability data and light curves (Volumes 11–12), and identification charts (Volume 13). Volumes 14–16 are sky charts of the Hipparcos and Tycho Catalog, comprising the Millenium Star Atlas, and indicating the most significant motions, distances, duplicity and variability properties (figure 4). All results from the mission, including intermediate astrometric data and individual photometric measurements, are included in the 6 CD-ROMs of Volume 17. The Hipparcos results can also be accessed via the WWW.

The scientific effort involved in the preparation of the Hipparcos mission and in its data analysis was substantial: some 200 scientists were involved in various aspects of the mission, in many cases for the 18 yr project's duration, from acceptance by ESA (or before) to the catalog publication.

Catalog details

The basic astrometric data in the Hipparcos and Tycho Catalogs include directions (positions), specified by celestial coordinates, their rate of change (proper motions), and trigonometric parallaxes. They are the first observational catalogs to contain parallaxes in addition to positions and proper motions for all stars in the observing programme. Positions and proper motions are given within the International Celestial Reference System (ICRS), which is consistent with the conventional equatorial system for the mean equinox and equator of J2000, previously realized by the FK5 Catalog. Timing of events is consistently based on terrestrial time (TT).

The standard astrometric model for a single star assumes uniform rectilinear space motion relative to the solar system barycenter. At some reference epoch, T_0, the motion of a star is described by the following six

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

parameters: barycentric coordinate direction, specified as right ascension, α, and declination, δ; annual parallax, π, from which the coordinate distance is $(\sin\pi)^{-1}$ astronomical units (to sufficient approximation, the distance in parsec is given by $1000/\pi$, where π is expressed in milliarcsec); the rate of change of the barycentric coordinate direction expressed as proper motion components $\mu_\alpha \cos\delta$ and μ_δ, in angular measure per unit time; radial (i.e. line of sight) velocity, V_R, in linear measure per unit time. The radial velocity is normally obtained from spectroscopic observations. The remaining five parameters are determined from analysis of the Hipparcos observations and are given for almost all stars in the catalog.

The standard astrometric model was adequate for the majority of stars, including the components of most resolved double and multiple stars. For several thousand apparently single stars, however, the standard model did not give an acceptable fit to the observations. These are probably binary stars, in which the center of light describes a curved path on the sky due to the orbital motions of the components around their common center of mass. For such 'astrometric binaries' additional parameters were introduced to describe the motion in accordance with the observations. In most cases a good empirical fit was obtained by including one or two additional terms in the Taylor expansion of the variation of the barycentric coordinate direction with time. In a few hundred cases even this was not sufficient, and Keplerian elements (sometimes partially based on existing ground-based data) were introduced to describe the orbital motion of the photocenter.

The catalog reference epoch is T_0 = J1991.25(TT), close to the central epoch of observations (the distinction between the epoch which specifies the equinox and equator of a coordinate system, typically B1950 or J2000, and the epoch at which the star passed through the given position, should be noted). The International Celestial Reference System (ICRS), as represented by the Hipparcos and Tycho Catalogs, replaces the FK5 system as the practical definition of celestial coordinates in the optical region. The construction of the ICRS ensures that no discontinuity larger than the uncertainty of the FK5 system occurs in the transition from FK5 (mean equinox and equator J2000) to ICRS. Thus, from the viewpoint of optical astrometry, the Hipparcos and Tycho Catalogs can be regarded as an extension and improvement of the J2000(FK5) system, retaining approximately the global orientation of that system but without its regional errors. However, the Hipparcos and Tycho Catalogs have not simply been tied to the FK5 system, but to the ICRS through the extragalactic radio frame. Providing positions within the ICRS means that they are consistent with present knowledge of the (extragalactic) radio reference system. Providing proper motions within the ICRS means that they are consistent with the best present realizations of a quasi-inertial reference system (i.e. non-rotating with respect to distant galaxies).

Since parallax is explicitly derived in the data reduction, the directions represented by the right ascension and declination at epoch T_0 are strictly barycentric. In previous astrometric catalogs, the unknown displacement due to parallax for each observation perturbed the final catalog positions.

At the accuracy achieved by Hipparcos, general relativity effects cannot be ignored. In the relativistic framework, the observed directions are 'proper directions' referred to the Lorentzian frame moving with the observer. In order to relate the observed directions at different epochs and of different objects to a simple geometrical model of the stars and their motions, the 'local' effects of the moving observer (stellar aberration) and space curvature (gravitational light deflection by the Sun) are removed by transforming the proper directions to 'coordinate directions' corresponding to the coordinate differences between the object and observer, expressed in the (solar system) barycentric coordinate system.

Stellar aberration was computed from the barycentric motion of the satellite, using the full relativistic formula and combining the geocentric satellite velocities accurate to about 0.2 m s^{-1}, determined by the ground stations, with the appropriate ephemeris for the barycentric motion of the Earth. Gravitational light deflection was computed in the heliocentric metric, assuming spherical symmetry. The treatment of aberration is consistent with the IAU (1984) recommendation that the total barycentric velocity of the Earth should be used; prior to that date the elliptic component of velocity, which changes very slowly with time, was conventionally ignored.

For the propagation of the astrometric parameters and their covariances to arbitrary epochs, normally only the changes in α and δ, due to the proper motion, are appreciable. In rigorous treatment the time dependence of all six parameters must however be explicitly considered, as described in the published catalog.

The proper motions are instantaneous rates of change of the barycentric coordinates at the epoch T_0. For stars with large parallax, large proper motion, and large radial velocity, the apparent acceleration of coordinates due to perspective effects may be significant. These can be calculated on the assumption of uniform space motion relative to the barycenter, and were accounted for, in the case of the few most significantly affected stars, in reducing the individual observations to T_0.

Scientific applications

The resulting stellar distances and motions impact on all areas of astronomy and astrophysics, but most particularly on the understanding of stellar structure and evolution, and the structure and dynamics of our own Galaxy. More than twenty thousand stars out to distances of about 100 pc now have their distances known to better than 10%, and another thirty thousand are known to better than 20%. Table 2 lists the 20 stars closest to the Sun, as deduced from the Hipparcos parallaxes, and supplemented by Yale

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. The 20 known stars closest to the Sun. The parallax is given in milliarcsec (mas). The distance in pc is given by $1/\pi$, where π is the parallax in arcsec. The data are from the Hipparcos Catalog, Volume 1, Table 3.6.1, supplemented by four objects from the Yale Parallax Catalog too faint to be observed by Hipparcos.

HIP	V	M_V	π	Name
70890	11.01	15.45	772.33	α Cen C
71681	1.35	5.70	742.12	α^2 Cen
71683	−0.01	4.34	742.12	α^1 Cen
87937	9.54	13.24	549.01	Barnard's star
—	13.46	16.57	419.1	Wolf 359
54035	7.49	10.46	392.40	BD +36 2147
32349	−1.44	1.45	379.21	α CMa
—	12.56	15.42	373.7	Luyten 726–8 A
92403	10.37	13.00	336.48	Ross 154
—	12.27	14.77	316.0	Ross 248
16537	3.72	6.18	310.75	ϵ Eri
114046	7.35	9.76	303.90	CD –36 15693
57548	11.12	13.50	299.58	Ross 128
—	12.32	14.63	289.5	Luyten 789–6
104214	5.20	7.49	287.13	61 Cyg A
37279	0.40	2.68	285.93	α CMi
104217	6.05	8.33	285.42	61 Cyg B
91772	9.70	11.97	284.48	BD +59 1915 B
91768	8.94	11.18	280.28	BD +59 1915 A
1475	8.09	10.33	280.27	BD +43 44 A

Parallax Catalog data for nearby stars too faint to be observed by Hipparcos.

Stars differ from each other in many important respects: in their mass, chemical composition, size, rotation rate, age, and in many other details. Understanding their composition and structure, how and where they were born, how they are evolving, and how they will eventually die, means determining their individual properties, and comparing them with detailed models of their initial conditions and subsequent evolution, involving a very comprehensive understanding of many areas of physics. Specifically, knowing the luminosity of a star requires knowledge of the star's distance, in order to convert its apparent magnitude into an absolute luminosity. The Hertzsprung–Russell diagram is the distribution of absolute magnitude versus color index. In its observational form (figure 5) it reveals one of the most important correlations of stellar properties known. The Hipparcos results have made significant contributions to the understanding of stellar properties, through the improved definition of their luminosities.

Comparing theoretical models with measurements of the brightnesses, temperatures, masses, and ages of the stars provides direct information on whether present understanding of the star formation process and subsequent evolution are correct. Ultimately, it leads to an understanding of the internal structure and other 'hidden' properties of the stars, since only when the correct physical ingredients and processes are included will the models agree with ever-improving observations. In the context of the relatively new discipline of asteroseismology, in which stars other than our Sun are scrutinized for observational signatures of solar-like oscillations, accurate distance measurements yield precise stellar luminosities, and thus provide a direct check on models of internal stellar structure.

The collective positional and kinematical distribution of the one hundred billion stars comprising our own Galaxy are linked through their gravitational environment, and via the history of star formation throughout the Galaxy. The initial properties are subsequently modified, often substantially, by small- and large-scale dynamical processes. Thus present-day star positions and motions encode the details of the initial formation and subsequent evolution of our Galaxy.

The Hipparcos Catalog results have led to important progress in understanding the space and velocity distribution of nearby stars. This knowledge is used to define the distribution of stellar luminosities, the distribution of stellar masses, and the kinematical behavior of stars in the solar neighborhood. Additionally, by observing the distances and motions of stars perpendicular to the plane of our Galaxy, the results have also been used to probe the distribution of 'dark matter' in the Universe, which is inferred to exist from cosmological and other arguments. This is possible because stars act as kinematic tracers, probing the gravitational signature of material irrespective of its physical manifestation. The results support a growing consensus concerning the main structural parameters of the galactic disk populations of stars, and on the possible distribution of dark matter in our Galaxy.

The Hyades is the nearest moderately rich open CLUSTER, whose distance has been the subject of intense interest and investigation over a very long period, due to its central role in defining the entire distance scale used in astronomy. More generally, clusters are important because they represent a gravitationally bound collection of typically a few hundred stars, born at the same time, and with a given chemical composition. The development of such a coeval system, both dynamically and with respect to the individual stellar properties such as temperature and luminosity, therefore provides an ideal laboratory for testing models of stellar evolution. Open star clusters are also subject to 'evaporation' and eventual disruption over timescales of tens or hundreds of millions of years due to gravitational encounters between stars within the cluster, or between the cluster and the galactic disk or giant molecular clouds. Hipparcos has provided a reliable direct distance measurement to about 200 stars in the Hyades cluster, allowing its internal structure, internal motion, mass segregation, and disintegration processes to be observed.

Distance estimates for clusters out to several hundred parsec have also been measured directly with Hipparcos. These are being used to extend the distance scale calibrations, and to verify models of stellar evolution according to age and chemical composition.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The problem of deriving accurate distances to objects throughout our Galaxy, beyond distances which are directly accessible by means of trigonometric parallax determinations, is one of the central themes in present day astronomy. On the largest scales, distances to other galaxies affect our knowledge of the HUBBLE CONSTANT. While most astronomers do not question the fact that the universe is expanding, the actual expansion rate is still vigorously debated. The precise value, as given by the Hubble constant, is of more than quantitative interest, since theories seeking to explain the formation and subsequent evolution of the early universe now predict properties of the present day universe which are becoming ever more stringently constrained by observations. Cepheid variables have played a very crucial role in these discussions, due to their regular pulsations with a period considered to depend primarily only on their luminosity. The Hipparcos results on these and other objects is leading to more secure distance scales within and beyond our Galaxy.

Hipparcos results are facilitating efforts to determine stellar ages and, in particular, to date the oldest stellar populations in our Galaxy, the GLOBULAR CLUSTERS. These ages therefore provide a lower limit on the age of the universe. Pre-Hipparcos estimates had put some globular cluster ages as high as 18 Gyr, older than current cosmological age estimates of the universe itself. Revised ages, derived from the new distance measurements, of perhaps 12–14 Gyr may have alleviated this age paradox.

Future developments

A number of proposed space missions are set to exploit the possibilities of measuring significantly more accurate stellar positions from space. The SPACE INTERFEROMETRY MISSION (SIM) is a US mission which will employ a medium-sized pointed interferometer to establish measurements of several tens of thousands of stars with microarcsec precision. GAIA is a mission under consideration by the European Space Agency which will combine the basic principles of Hipparcos with numerous innovative features, and which will result in microarcsec positional accuracy for approximately one billion stars. In addition to the satellite's two main astrometric lines of sight, a third viewing direction is devoted to the acquisition of radial velocities of the brighter 10^8 or so stars, down to about 17 mag. Sampling roughly 1% of the stellar population of our Galaxy, GAIA is designed to provide a six-dimensional spatial and velocity mapping extending throughout the Galaxy. The motivation is to rectify the shortcoming of classical two-dimensional star measurements by providing the third space component of the star's velocity vector, important for understanding and interpreting stellar kinematics and dynamics.

GAIA will be positioned and operated at the Sun–Earth LAGRANGE POINT L2, roughly 1.5×10^6 km from Earth, which provides a rather benign thermal environment, and an absence of eclipses and occultations. It will demand data transmission rates of about 1 Mb s^{-1} throughout the planned 5 yr lifetime. The payload will employ larger optics, CCD detectors giving higher efficiency and parallel observation of many more stars, and innovative approaches to the problem of structural and thermal stability, yielding positions nearly a thousand times better than Hipparcos. GAIA will circumvent the complexities of an 'input' or pre-defined observing catalog, and select its targets autonomously on-board. The goal is to observe essentially every star down to 20 mag, thereby largely eliminating the problems of 'selection effects' which plague the interpretation of catalogs relying on the pre-selection of objects. At an average of around 25 000 stars per square degree GAIA will dwarf the Hipparcos Catalog's 3 stars per square degree, or even the Tycho Catalog's 25.

Bibliography

http://astro.estec.esa.nl/Hipparcos is the ESA site with details of Hipparcos, access to the catalogs and annexes, and links to related scientific papers

ESA 1997 *The Hipparcos and Tycho Catalogues* European Space Agency ESA SP–1200 (17 volumes)

Høg E *et al* 1997 The Tycho Catalogue *Astron. Astrophys.* **323** L57–L60

van Leeuwen F 1997 The Hipparcos Mission *Space Sci. Rev.* **81** (3–4) 201–409

Perryman M A C 1998 The Hipparcos Astrometry Mission *Phys. Today* June 38–43

Perryman M A C *et al* 1997 The Hipparcos Catalogue *Astron. Astrophys.* **323** L49–L52

Turon C *et al* 1997 From Hipparchus to Hipparcos *Sky & Telescope* July 28–34

Michael Perryman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hipparcos and Tycho Catalogues

The decision of the International Astronomical Union to adopt the new celestial reference system ICRS (see EXTRAGALACTIC REFERENCE FRAMES), and its realization in the radio domain (ICRF, see SPACE AND TIME REFERENCES: CONCEPTS), was contingent upon the completion of a secondary catalog of star positions and proper motions in optical wavelengths. The only suitable catalog was the one obtained from the data of the Hipparcos space astrometry mission (see HIPPARCOS).

Reference frame of the Hipparcos Catalogue

Hipparcos measured the angles between the objects in its observing list. From these and their variations with time, positions, proper motions and parallaxes were calculated in a single coordinate system. However, because the angles are invariant with respect to a rigid rotation of the coordinates, the axes of this preliminary Hipparcos Catalogue were arbitrarily placed and had an arbitrary linearly time-dependent rotation. The necessary link to ICRS could then be expressed as a rotation vector R that would bring the coordinates and proper motions into the ICRS:

$$R = \epsilon + \omega(T - T_0).$$

where ϵ is the rotation vector to be applied at a reference time T_0 and ω is the yearly spin vector. The components of ϵ and ω were obtained by comparing the positions and/or proper motions of stars in the preliminary Hipparcos Catalogue with those observed with respect to extragalactic radio sources of the ICRF or more generally to galaxies which were assumed to be remote enough in order to be considered as fixed. Ten different programs were organized to provide the necessary relations to determine R. They included observation of radio stars with respect to ICRF objects by VLBI or connected radio interferometers, special large-scale photographs including Hipparcos stars and galaxies taken at two largely separated epochs, the use of large proper motion surveys made with respect to galaxies and the comparison of Earth rotation parameters obtained by VLBI and optical astrometry using Hipparcos positions. The synthesis of all these results provided the components of ϵ for 1991.25 with an uncertainty of 0.6 mas (milliseconds of arc) and of ω with an uncertainty of 0.25 mas yr^{-1}.

Contents of the Hipparcos Catalogue

The catalog contains astrometric data for 117 955 stars. It includes a basic list of 52 000 bright stars, a survey to a limiting visual magnitude V that is a function of the star's spectral type and galactic latitude b: for spectral types earlier than or equal to G5,

$$V \leq 7.9 + 1.1 \sin |b|$$

for spectral types later than G5,

$$V \leq 7.3 + 1.1 \sin |b|.$$

In addition, there are 66 000 stars of magnitudes up to $V = 12.4$ selected for their astrophysical or astronomical interest, or chosen in order to fill up empty regions of the sky. The mean astrometric characteristics of the catalog for stars of magnitude $V \leq 9$ are the following:

> Median uncertainty of positions: 0.77/0.64 mas (RA/dec)
> Median uncertainty of proper motions: 0.88/0.74 mas yr^{-1} (RA/dec)
> Median uncertainty in parallaxes: 0.97 mas
> Ten per cent of these parameters have uncertainties less than 0.47/0.66 mas (RA/dec)

In addition, complete or partial information on 14 340 double or multiple systems is provided, while 8542 other stars are marked as suspected double stars.

Hipparcos was also a wide-band photometer. Results for 118 204 stars are given, in the mean, from 110 observations of each object. The median photometric precision for bright stars (Hipparcos magnitude up to 9) is 0.0015 mag. Within the whole material, there are 11 597 variable stars, 70% of which have been first identified as variables by Hipparcos.

Contents of the Tycho Catalogue

The Tycho Catalogue was constructed from the data collected by the star-mapper of the Hipparcos mission (see also TYCHO STAR CATALOGS). In contrast with the main mission in which stars are observed roughly during 4 s per transit, all stars in the field of view of the star-mapper gave a signal, but during only a quarter of a second. This explains why many more stars have been observed and are present in the Tycho Catalogue, but that the uncertainties are much larger. Actually for most stars only positions at the mean epoch are really useful. The total number of entries is 1 058 332. This includes practically all stars up to V magnitude 10.5 and many others up to $V = 11.5$. The mean median uncertainty in position for all stars is 25 mas. It is 7 mas for stars with $V < 9$.

During the construction of the catalog, the coordinate axes were identical to those of the preliminary Hipparcos Catalogue. Then, the catalog was turned by the same rotations as the Hipparcos Catalogue so that the reference system of the final Tycho Catalogue is also the ICRS. This means that, although the Hipparcos Catalogue is the best optical reference frame realizing the ICRS, the Tycho Catalogue is also a realization of the same system.

Tycho measured magnitudes in two colors close to the V and B magnitudes of the Johnson system. The mean uncertainties for all stars are 0.07 and 0.06 mag respectively for the Tycho B and V magnitudes. For stars with $V < 9$, these numbers should be divided by 5.

Bibliography

European Space Agency 1997 *The Hipparcos and Tycho Catalogues (ESA Special Publication SP-1200)* vol 1
Kovalevsky J, Lindegren L, Perryman M A C *et al* 1997 The Hipparcos Catalogue as a realization of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

extragalactic reference system *Astron. Astrophys.* **323** 620–33
van Leeuwen F 1997 The Hipparcos mission *Space Sci. Rev.* **81** 201–409

Jean Kovalevsky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hirayama Family

Any of several groups of asteroids each of whose members share similar orbital elements, in particular a closely similar semi-major axis and orbital inclination. These similarities are surmised to indicate that the members of a particular Hirayama family share a common origin in the collisional break-up of a large parent body. The existence of such groups were first pointed out by Kiyotsugu Hirayama in 1918. Several dozen Hirayama families have been identified with a fair degree of certainty. Spectral similarities within some families support the idea of a common origin. This is true of the most populous families. The Themis family, for example, orbits at 3.13 AU with an inclination of 1°. It has several main members, including the 228 km diameter (24) Themis, and many minor members, which are all low-albedo C-type (carbonaceous chondrite) objects. However, there are families which contain a variety of types: the Alexandra family, for example, orbits at between 2.6 and 2.7 AU with inclinations of 11° to 12°, but has C-, G- and T-type members—difficult to reconcile with an origin as a single object.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hiten (Muses-A)

First Japanese Moon mission. Launched January 1990. Named Hiten after a Buddhist angel who plays music in heaven. Used to verify the swingby technique by utilizing lunar gravity. Returned engineering data, detected cosmic dust, and released a 12 kg orbiter called Haroromo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hoba Meteorite

The largest known single meteorite mass, weighing an estimated 60 tonnes; it is sometimes called the Hoba West meteorite. It still lies where it was discovered in 1920, at Hoba Farm, near Grootfontein in Namibia. It produced no impact crater, but sits partially embedded in a hollow in the ground. The Hoba meteorite is an ataxite, a variety of iron meteorite, and measures about 2.7 m in its longest dimension. It is considerably weathered, and the surrounding surface is covered by a rusty deposit, the amount of which suggests that when the meteorite landed it may have weighed 75 tonnes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hoffman, Jeffrey A (1944–)

X-ray astronomer and astronaut, born in Brooklyn, New York, educated as an astrophysicist. Researched cosmic gamma ray and x-ray astronomy, flying a balloon-borne, low-energy, gamma ray telescope as a PhD project, worked at Leicester University on the x-ray astronomy rocket payload, and was project scientist for the medium-energy x-ray experiment on the EXOSAT satellite. At MIT worked on HEAO-1. Became an astronaut in 1979 and made five flights, during the fourth of which in 1993, on the shuttle Endeavor, he serviced and restored to full capacity the Hubble Space Telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hoffmeister, Cuno (1892–1968)

German astronomer, founder of the Sonnenberg Observatory. Discovered thousands of variable stars through repeated photography of the sky and his technique of 'fly-spanking', comparing the size of the stellar images to identify changes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hogg [Sawyer Hogg, Hogg-Priestley, Sawyer-Hogg-Priestley], Helen, née Battles (1905-93)

Astronomer, born in Lowell, MA, and planned a career as a chemist. However, a total eclipse of the Sun in 1925, as she said, 'tied me to astronomy for life'. After meeting ANNIE CANNON of the Harvard College Observatory, she worked at Harvard with HARLOW SHAPLEY on star clusters. She moved to Victoria, British Columbia, and started an observing program with the 72 in telescope to study variable stars in globular clusters, chaperoned by her astronomer husband and accompanied to the dome at night by her new-born baby girl in her basket. The work continued in Toronto's David Dunlap Observatory, where she became professor at the University of Toronto. She took over 2000 photographs during her research on variable stars in globular clusters, discovering hundreds of them, and compiling her work and others into the *Catalogs of Variable Stars in Globular Clusters*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hogg, Arthur Robert (1903–66)

Australian atmospheric physicist, turned astronomer, born at Creswick, Victoria. Hogg joined the Mount Stromlo Observatory and became an astronomer after the Second World War, carrying out photoelectric photometry on eclipsing variables. The most important system that he observed was zeta Phoenicis, one of the 13 for which 'first-order' masses and radii can be determined. He determined standard magnitudes, and measured the integrated magnitudes and colors of the Magellanic Clouds, finding that the Small Cloud was appreciably bluer in its bright nuclear regions than in its outskirts. He worked on galactic clusters (he measured color–magnitude diagrams for five galactic clusters, and published a photographic atlas of southern galactic clusters). He played a leading part in the setting up and testing of the new 74 in telescope. Directed by BART BOK, he surveyed outback Australia for alternative astronomical sites when the sky quality at Mount Stromlo was compromised by the growth of Canberra, as result of which Siding Spring Mountain near Coonabarabran was chosen, becoming the site for the Anglo-Australian 150 in telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Holden, Edward Singleton (1846–1914)

Astronomer, born in St Louis, MO, designed and directed the Lick Observatory in California.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Holmberg, Erik [Eric] Bertil (1908–)

Swedish astronomer who investigated the photometric properties of galaxies. He identified a characteristic size for a galaxy in terms of its brightness (the *Holmberg radius*) and by comparing elliptical galaxies tilted at different angles to the line of sight found that galaxies do not differ significantly in brightness, concluding that they were relatively free of dust obscuration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Home Observatories

Sooner or later—generally sooner, rather than later—the average amateur astronomer is going to feel the need for a telescope. Binoculars are of immense use (see BINOCULAR ASTRONOMY), but their low magnification means that they are bound to be limited. Most amateur societies either have observatories or include members with telescopes of their own (see AMATEUR ASTRONOMICAL SOCIETIES) but it is much more satisfying to be able to go straight outdoors and make ready for observing without the need to go on a journey.

Small telescopes are portable, and this is almost essential for the would-be observer who lives in a town; light pollution will probably make the seeing conditions intolerable, and the only solution is to have a telescope which can be put into the boot of a car and moved around. There are also many observers who keep their telescopes indoors and take them out when needed, but this too has its disadvantages; moving a telescope (particularly a Newtonian REFLECTOR) may well disturb the alignment of the optics, and in any case it is only a matter of time before something is dropped, with disastrous results. All in all, a permanent observatory is very much to be recommended.

The first essential is to select a site, and this is never easy, particularly if the choice is restricted. Let it be said at once that setting up an observatory on the roof of a dwelling house is unwise. Heated air rising from below will swirl around and destroy the SEEING. It is necessary to build the observatory well away from the house—within reason, the further away the better. Also, with a permanent building, planning permission has to be watched, at least in the UK. If you are faced with an awkward Council, the way round this problem is to put down a concrete base and then place the observatory on top, without actually fastening it. It then comes under the heading of a portable structure, and the Council will be powerless unless the building interferes with a neighbor's view.

Houses and trees are always inclined to cut off part of the view of the sky, and it is axiomatic that if a tall tree can be in the most awkward possible position, it is. Remember, too, that heated air will always rise from a dwelling house, so if possible site the observatory so that there is an unobstructed view of the horizon—or, more precisely, that part of the horizon in which you are particularly interested. Lunar and planetary observers will need the southern sky, while the variable-star enthusiast or the nova hunter may well prefer the north, and those who search for comets will try to ensure that there is nothing to obstruct the western sky after sunset or the eastern sky before dawn. Nearly always, of course, it is a matter of compromise. Street lights can be an unmitigated nuisance, particularly if they remain on all night (as many of them now do), but there is no real solution here. The airgun remedy is, unfortunately, illegal.

The very simplest form of observatory, suitable for any kind of telescope, is the run-off shed. This is simply an ordinary shed mounted on rails; when the telescope is to be used, the shed is simply rolled back by its wheels. It is wise to make the shed in two sections which join in the middle, rather than having a single shed with a door. Any door has to be fixed at one end of the shed and must be either hinged or removable. If hinged, it flaps; if removable, there may be problems in replacing it during a dark, possibly windy night, because it will tend to act as a powerful sail.

Anyone who is reasonably good with their hands can make a run-off shed, and there is very little to go wrong, but it does mean that the observer is in the open when actually observing, so that the telescope may be affected by wind-shake and the observer may feel decidedly chilly. Certainly it is much more comfortable to carry out one's observational program from inside a structure.

Another simple pattern is the run-off roof. Here there are permanent walls, and the roof—or even the entire top part of the structure—is rolled back on rails, which means of course that the rails have to protrude and must be supported on an adequate mounting. Again there is not much to go wrong, and the observer is shielded from the wind as well as being screened, at least to some extent, from obtrusive lights. The run-off roof is better suited to refractors, because a Newtonian reflector must be mounted lower down and will have a more restricted view of the sky, but there are in general no insurmountable difficulties, and again an observatory of this kind is comparatively easy to build.

The material for the observatory may be wood, metal, plastic, brick or fiberglass; metal has the disadvantage of shutting in the daytime heat, so that it is essential to open the observatory and stabilize the temperature some time before observing is due to begin. One point is worth noting (it should be obvious, but can sometimes be overlooked): make sure that the roof is not perfectly level. If it is, the first shower of rain will convert it into a pond.

The popular picture of an observatory is that of a graceful dome, with a slit which can be opened; the dome must be rotatable, so that full access can be given to any part of the sky. The main problem here is that making a dome is by no means easy, and it is also essential that the ring upon which it revolves is absolutely true; otherwise the dome will stick. In some cases the entire observatory rotates, and this does solve some of the difficulties, but a complete dome is heavy to push around, and it too is apt to stick unless all the components have been made with real precision.

Fiberglass domes can now be bought commercially, and many amateurs are now making use of them; purchasing a dome and then building the supporting structure is much cheaper than paying for a complete observatory. A 'skeleton' dome, with roofing felt, can also be used, although it is less durable and cannot be expected to last for very long without major repairs.

One variant which is quite popular is what can be referred to as the 'wedding-cake' pattern. Here there is a basic structure, with a ring upon which the upper section rotates, but the shape is not that of a dome; there are vertical sides and a roof made in sections, of which one

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can be swung back to provide access to the sky. In this case the design should not be perfectly symmetrical; if it were, the telescope would never be able to point to the zenith—and although one seldom wants to look 'straight upward', it can be necessary occasionally, to observe a comet or perhaps an interesting variable star.

Probably the first decision to be made is a straightforward one: buy an observatory complete, buy parts, such as the dome, and make the rest or construct the entire observatory oneself? A run-off shed, or even a run-off roof pattern, is within the scope of most people, but it is true that making a dome, or even a wedding-cake, is more of a problem.

Some observatories are designed for special needs. The solar observer will work by day and may not be much concerned with anything but the Sun; he or she may well need an observatory with a coelostat which catches the Sun's rays and sends them in a fixed direction, so that there is no need for a rotating dome or roof, and even the observatory itself can remain fixed. One leading British solar observer, Commander Henry Hatfield, has taken matters to the limit—he designed his spectrohelioscope, and then literally had his house built round it. It works excellently, but not many people, however enthusiastic, will be able to follow his example. Amateur radio astronomers will have different requirements, and will be concerned largely with avoiding electronic 'noise' rather than light pollution.

Various other points are worth noting here. First, any driven telescope will need a power supply. Falling weights are convenient and self-sufficient, but in most cases the drive will be electrical, and connecting the observatory to a mains power supply can be awkward. Secondly, there is the question of security. If the observatory is to be set up in the owner's garden this should not be a problem, at least in most areas, but if it is to be built on a site well away from the house there must be adequate protection; there have been many cases of observatories being vandalized. It is also wise to have proper insurance cover. Moreover, there is always a potential threat from the forces of nature. A local tornado, for example, can have devastating results.

Finally, always make sure that your observatory is large enough for your needs. For much of the time it will be used in the dark, and it is only too easy to damage oneself by blundering into a telescope or some other piece of equipment.

Not every amateur astronomer can own an observatory, and there is certainly no point in erecting any permanent structure in a small garden which is surrounded by trees or houses and is flooded with artificial lights nearby. However, there is no doubt that there is tremendous satisfaction in being able to walk through one's door into an observatory which can be opened and made ready for use in a matter of minutes.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Homunculus Nebula

A cloud of dust surrounding the unstable star η Carinae, position RA $10^{\,h}\ 45.1^{\,m}$, dec. $-59^\circ\ 41'$. The cloud has been shed from the star during outbursts that began in 1843, and now measures about $17''$ by $12''$. In shape it resembles a human figure.

See: Eta Carinae Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hooke, Robert (1635–1703)

Scientist, born in Freshwater, Isle of Wight, England, educated at Christ Church College, Oxford where he met Boyle and was employed by him to construct his air pump, became professor of geometry at Gresham College, London. Discovered *Hooke's law of elasticity*. Worked on optics, simple harmonic motion and elasticity in strings. Published a monograph called *Micrographia*, containing beautiful pictures of objects Hooke had studied through a microscope he had made himself. He coined the use of the word 'cell' (meaning a tiny bare room) for the structures in cork and other living organisms. Hooke invented the conical pendulum and is credited with HADLEY as the first to build a reflecting telescope to GREGORY's design. He made important astronomical observations including the fact that Jupiter revolves on its axis. He determined this from the Great Red Spot, which he was first to observe. His drawings of Mars were later used to determine its period of rotation, since they recognizably showed Syrtis Major, and were the earliest drawings to fix the phase of Mars. He wrote a book on comets which was read by NEWTON. In 1672 Hooke attempted to show that the Earth moves in an elliptical orbit and in 1678 that an inverse square law of gravitation explained planetary motions, but seemed unable to give a rigorous proof. He communicated these ideas to Newton, and, in a bad tempered dispute continued to this day by his descendants, claimed priority in the discovery of the inverse square law. Newton, as a consequence of the dispute, expunged all references to Hooke from the *Principia*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Horizon

A plane perpendicular to the line from an observer to the zenith. The astronomical horizon is the great circle formed by the intersection of the plane of the horizon with the celestial sphere.

See also: celestial sphere, great circle, zenith.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Horizontal-branch Stars

This article deals with the subject of the horizontal branch and horizontal-branch stars, those stars which are found on the horizontal part of the Hertzsprung–Russell diagram of a globular cluster. Figure 1 shows the location of this type of star. After an introduction there are sections on studies of globular clusters and the horizontal branch, the second parameter problem and field horizontal-branch stars.

In the 19th century, astronomers thought that stars of different spectral types were composed of different elements. In the 1920s a major step forward in stellar analysis was made by Cecilia Payne Gaposchkin who determined that stars were essentially the same, i.e., they are made up mostly of hydrogen with a lesser amount of helium and very small amounts of all the other elements in the periodic table. Later on in the 20th century, with an increasing precision in the ability to measure stellar abundances, a major advance in stellar analysis was to show how stars are different. Stars have been divided into many groups depending on their detailed spectroscopic abundances. This article will discuss one important group of old stars, the horizontal-branch (HB) stars. HB stars are most easily detected in GLOBULAR CLUSTERS and open clusters, but have also been detected in the general field of the Galaxy.

Astronomers have an impressive tool to use in the study of stellar evolution, the HERTZSPRUNG–RUSSELL DIAGRAM (HRD), in which the absolute magnitudes of stars are plotted versus their colors (or spectral type or temperature). For a group of stars at the same distance, the apparent magnitude can be used (i.e., for a globular or open cluster) in place of the absolute magnitude. In the original HRDs the majority of stars fell in an inclined band, named the main sequence, running from the upper left to the lower right (from high temperature and luminosity to low temperature and luminosity). These were the normal Population I (Pop I) stars like the Sun, with spectral types running from O (very hot) to M (cool). In the upper right side of the diagram there are the red giants, which are cool, more luminous stars that have evolved from the main sequence and have become stars of a much larger radius.

Globular clusters are important objects for the study of stellar evolution. They are presumed to be among the first objects formed in our galaxy and thus give us information about the galaxy's very early history. The stars in a cluster are all at the same distance from the Sun and they (in almost all cases) are formed coevally out of interstellar material with the same chemical composition. Thus, they are experimental 'test beds' to be used by astronomers in their studies of STELLAR EVOLUTION.

HARLOW SHAPLEY investigated the magnitudes of stars in globular clusters, starting about 1915 at MOUNT WILSON OBSERVATORY. He found that the HRDs for globular clusters differed from the diagrams for the stars near the Sun in that the red stars (red giants) were brighter than the blue stars in the cluster. In the diagrams of stars near the Sun, the blue stars were the brightest stars in the diagram. The significance of this different distribution of stars in the HRD was not understood until the work of Walter Baade. He took pictures of the Andromeda galaxy (M31) in which, for the first time, individual stars in the central regions of the galaxy were resolved. These pictures were taken at Mt Wilson on the 100 inch telescope during WWII blackouts under good seeing conditions. The stars in the center of M31 had an HRD similar to that of globular clusters; the stars in the outer regions had an HRD similar to the distribution of stars in Russell's diagram for the nearby stars. This was the start of the idea of STELLAR POPULATIONS; that there are young stars with relatively larger abundances of elements other than hydrogen and helium, distributed mainly in the spiral arms of our galaxy, and that there are old stars with very small abundances of elements other than hydrogen and helium, distributed in a spherical halo around the galaxy. In time this simple picture of two populations was modified with the addition of other population types.

In 1952 at a Symposium on Stellar Populations held at the Vatican Observatory, Jan Oort presented a table in his summary in which stars were divided into five different populations. At one extreme were the Halo Pop II stars, then Intermediate Pop II, Disk Population, Intermediate Pop I and at the other extreme were the Extreme Pop I stars. We know now that Pop II stars are old, metal poor and have high velocities while the Pop I stars are young, have a higher metal abundance and have small velocities. Globular clusters were members of the Halo Pop II and it is in these clusters that we find a distribution of stars known as HB stars. In figure 1, at the top, a color–magnitude diagram (CMD) is presented for the globular cluster M92 (by Peter Stetson, Dominion Astrophysical Observatory). It is the result of CCD (charge coupled device) photometry of more than 50 000 stars. Below, a CMD is shown for the globular cluster M3 by Roberto Buonanno and his collaborators (University of Bologna) with the various areas of the diagram labeled.

In the lower middle of each diagram, a band of points can be seen, extending up to the left. These points represent the main sequence of the cluster, which is made up of stars burning hydrogen in their cores. In the early 1900s stellar energy was thought to be produced by stellar contraction. However, contraction models did not provide a long time scale for stellar evolution. In the late 1930s Bethe and von Weizäcker showed that nuclear fusion takes place in the solar core and this understanding provided the means to calculate stellar models showing how stars evolve over time. Stellar evolutionary theories are tested mainly by comparing theoretical models with observations made of stars in clusters. For stars like the Sun, protons are transformed into helium nuclei in the stellar core, liberating vast amounts of energy. This is a long-lived stage in a star's life, and during it the star remains at almost constant temperature and luminosity, on the main sequence. The more massive stars form on the upper left of the main sequence and evolve rapidly;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

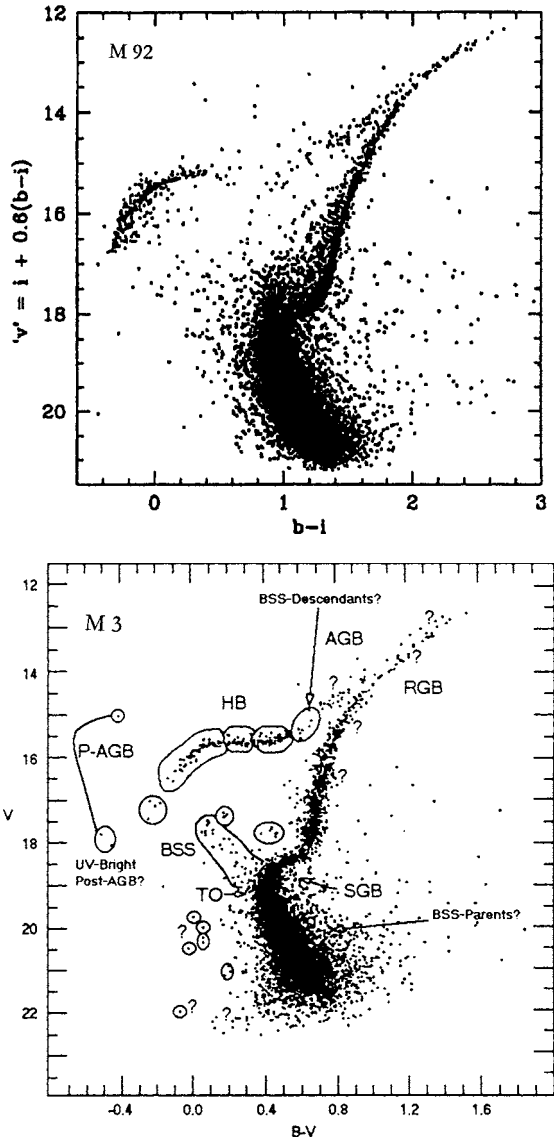


Figure 1. Top: a CMD for the globular cluster M92, showing the distribution of more than 50 000 stars (Peter Stetson). Bottom: a CMD for M3 with the various regions labeled (Roberto Buonanno).

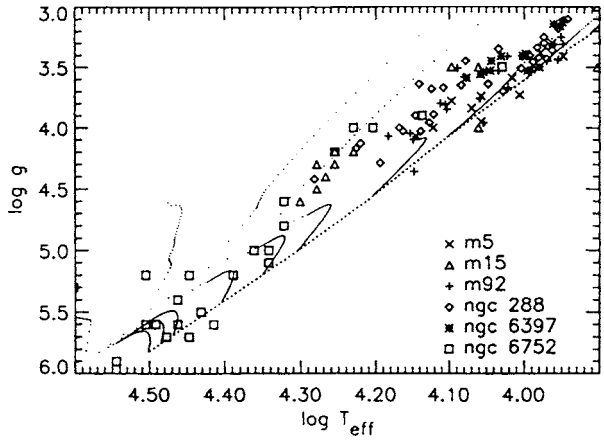


Figure 2. Sweigart models of the horizontal branch. Data points from observations in globular clusters are also shown.

the smaller mass stars sit on the lower right of the main sequence and evolve very slowly. Eventually all the hydrogen is converted to helium in the center of the star. However, the helium is at too low a temperature to burn and the core contracts, raising the temperature in the central regions of the star. Hydrogen burns in a shell outside the core, and the star leaves the main sequence and starts moving up and to the right in the diagram, mapping out the red giant branch (RGB). As the star becomes more extended on its way up the giant branch, it becomes more difficult for it to retain the gas on its surface and there is envelope mass loss. It is thought that perhaps 0.2 solar masses are lost before the star reaches the horizontal branch (HB). At the tip of the RGB, the helium in the core of the star ignites, in the so-called helium flash. The star then moves rapidly to the HB, the group of stars on the far left in the range $v = 15$–17 in the M92 diagram. If the mass of the star is small, it lands on the left extension of the HB; if the mass of the star is larger, it will fall further to the right on the HB. The locus of points that is defined by stars arriving on the HB, as a function of mass, is called the zero-age horizontal branch (ZAHB). During this stage of its life the star is burning helium in the core and hydrogen in a shell surrounding the core.

According to HB stellar models calculated by Allen Sweigart (NASA/GSVC), the evolutionary path of a classical HB star starts by brightening a bit and then moves to the blue (left), paralleling the zero-age horizontal branch (ZAHB). Then the star brightens some more and turns back toward the red (right), climbing up the asymptotic giant branch (AGB), which is the locus of the points scattered in a band just above the giant branch in the CMD (figure 2). These evolutionary tracks depend on the specific parameters of the HB star under investigation. On the AGB a star has two burning shells: hydrogen and helium.

It is interesting to note that in the 1960s, some textbooks had diagrams showing that stars evolved down the asymptotic giant branch and finally reached the HB. In modern theories of stellar evolution, the stars evolve to the ZAHB and then evolve to the red, up the asymptotic giant branch in the opposite direction.

In studies of certain variable stars (called RR LYRAE STARS) on the HB of globular clusters, Oosterhoff found in 1939 that there were two groups determined by their average period. One group had average periods of 0.55 days, the other had average periods of 0.65 days. The long-period group was found to be more metal poor than the short-period group. Work by ALLAN SANDAGE (Carnegie Observatory) and others showed that there was a difference in their periods, in that for stars at constant effective temperature the more metal-poor stars

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

had longer periods than the stars of higher metallicity. This is an illustration that the parameters concerning stars in globular clusters are more complex than originally thought.

Studies of globular clusters and the horizontal branch

In 1919 Shapley noted that the LUMINOSITY FUNCTIONS OF GALAXIES for M3 and M13 were different. M3 had many cluster-type variables but was poor in the number of blue stars, and M13 had few variables but had many blue stars. He commented that there might be significance in this difference. This was one of the first statements concerning differences between globular clusters.

In the following decades observational evidence, combined with theoretical studies, showed that globular clusters were old. Shapley had already placed the center of the globular cluster distribution at the center of the Milky Way. It was assumed that each globular cluster is formed of stars created out of the same interstellar material, at the same time and at the same distance from the solar system. Thus, if the chemical compositions, ages and kinematics could be determined for clusters as a function of galactocentric distance, one could learn much about the early chemical and dynamical history of the Galaxy.

In 1955 Halton Arp (now at the Max Planck Institute) published the results of an investigation of seven globular clusters (M2, M3, M5, M10, M13, M15 and M92). The CMDs of these seven clusters were quite similar. The [Fe/H] values (the so-called metal-abundance parameter which is the logarithm of the iron abundance divided by the hydrogen abundance, all relative to the Sun) varied from [Fe/H] $= -2.02$ for M15 to [Fe/H] $= -1.34$ for M5. It is now known that for the most metal-rich clusters [Fe/H] reaches values of -0.3, close to the solar value of zero, but in Arp's original sample there were no clusters this metal-rich. The clusters in the list all had CMDs with strong, blue HBs and there was not much variation from one cluster to the next. Even in the days of photographic photometry, the narrow sequences found in the CMDs implied that the stars in a globular cluster were chemically homogeneous. Changes in the chemical abundances, and other parameters, affect the position of the star in the HRD and thus, if the sequence is narrow, imply that the stars are very similar. Modern CCD diagrams have sequences which are much more sharply defined, reinforcing this conclusion.

The CMDs provide several methods of determining ages of globular clusters. One of the first methods was to measure the magnitude difference between the bluest point on the turnoff (TO in figure 1, bottom diagram) and the magnitude of the HB at the same color. The TO magnitude gets fainter by about 0.01 mag for each 1% increase in the age. Cluster model builders, like Don VandenBerg (University of Victoria) create isochrones which predict how a CMD of a cluster should appear, given a set of parameters (such as helium abundance, [Fe/H], [CNO/Fe], age). These isochrones can then be fitted to observed CMDs to obtain estimates of cluster ages.

In 1976 Philip, Cullen and White published a compendium of all the galactic globular clusters then available, all drawn to the same scale and with [Fe/H] and reddening values listed for each. This presentation made it easier to intercompare clusters. The first CMDs were made photographically by taking pictures through V (visual) and B (blue filters). When the magnitudes were reduced to the UBV SYSTEM then one could plot V versus $(B–V)$ and so arrive at a CMD. The accuracy of the magnitude determination for a single star was of the order of $\sim$0.03 mag for the brightest stars in a globular cluster and decreased in precision as one went to fainter magnitudes. Photoelectric photometry was much more accurate but in the 1970s stars were measured with a single-channel photometer, one at a time, and so it was difficult to measure a large number of stars in any given cluster. In the 1980s the CCD was introduced and soon became the instrument of choice at observatory telescopes to take images, measure magnitudes and record spectra. For the first time, stars in the central areas of globular clusters could be measured, and as can be seen in figure 1, tens of thousands of stars could be recorded with a single exposure. The CCD allows one to measure stars much fainter than could be measured with photography and the precision of the measurements is much higher. One per cent photometry is now possible for faint stars and this allows the various regions of the HRD (or a CMD) to be much more precisely defined. Some recent technological advances have made stellar photometry even more effective. The HUBBLE SPACE TELESCOPE has allowed extremely faint objects to be measured in very crowded areas with high accuracy. On the ground ADAPTIVE OPTICS allows a telescope to be kept in the best possible focus during observations and thus increases the accuracy of the measurements. CCDs also allow high resolution spectra with signal/noise ratios of 150 to be obtained for stars even in globular clusters and equivalent widths of lines can be measured with errors as small as $\sim$2 mÅ. This means that better tests can be made of the predictions of stellar evolution theories.

Studies of HB and other stars in clusters have been greatly aided by the increased photometric accuracy possible due to the new software which has been created to measure images on CCD frames. The most important part of this process is the fitting of the point spread function (PSF), which is a three-dimensional model of the shape of the intensity distribution of each image. This method replaces the older aperture photometry, which just measured the total luminosity of the image. At the Dominion Astrophysical Observatory Peter Stetson has written DAOPHOT, at MIT Paul Schecter has written DoPhot and in Italy Roberto Buonanno and collaborators have written ROMAFOT. These programs allow one to create a mean PSF (by measuring the profiles of isolated stars well spaced over the frame) which is then fitted to all the images to find the magnitude of each star. If

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the PSF varies across the frame (which happens as the frames become larger) the programs can apply a variable PSF as a function of position on the frame. Without the development of modern computer technology these kinds of reduction programs for magnitude determination would not be possible.

As more globular clusters were studied, it was found that not all clusters had the same type of HB. Some had long, extended blue tails; others had only red HB stars, to the right of the RR Lyrae gap. The main factor that seemed to determine the morphology of the HB was metallicity, [Fe/H], but investigators found some clusters with the same metallicity which had very different distributions of stars along the HB, suggesting the influence of a 'second parameter'.

In the early 1970s Robert Rood (University of Virginia) constructed synthetic HB models and he studied what parameters would produce the observed HBs. He confirmed the need for a mass loss of $0.2M_{\odot}$ between the main sequence and the HB and for a 10% dispersion in the amount of mass loss for stars on the HB to reproduce the observed distribution of stars. He produced a list of second parameter candidates and showed how it was possible to produce any desired HB morphology.

Gaps have been found in the distribution of stars along the HB, in field HB stars and for HB stars in globular clusters. One of the first people to discuss the gaps was Barry Newell (Canberra, Australia) while he was at Yale in the late 1970s. He found two gaps at effective temperatures $\log T_{\rm eff} = 4.11$ and 4.33 in the distribution of galactic halo blue stars and suggested that these gaps were an HB phenomenon. In the CMD these gaps occur at $(U-B), (B-V) = -0.4, -0.14$ and $-0.9, -0.24$. Another possibility, discussed by others later to explain stars on the blue end of the HB, is that some stars of a different type end up in positions near the HB.

The second parameter problem

The changes found in the morphology of the CMD of the HB were first attributed to a change in the metal abundance. As the metal abundance of the cluster increases, the giant branch in the CMD shifts to the red. For high metal abundance the blue HB stars do not appear and the red part of the HB becomes a short red stub. The most metal-poor clusters showed strong, blue HBs, and more modern observations which reach stars of fainter magnitudes show long, drooping blue tails for blue HBs of metal-poor clusters. In figure 3 two pairs of CMDs are shown. The first pair is NGC 1904 and NGC 6637 which show the change in HB morphology from a [Fe/H] of -1.5 to -0.2 in a $V, (B-V)$ diagram. The HB changes from one with a very strong, blue component to a cluster with only a short red stub of a HB. The second pair, on the right, shows a $V, (U-V)$ diagram for M3 and M13 and is a clear example of the influence of a second parameter acting. At the bottom are three CMD, reported on by Jim Hesser, quoting photographic observations by Roberto Buonanno of the clusters M15, M92 and NGC 5466, each of which have [Fe/H] near -2.1. Again there is a dramatic change in the morphology of the HB. This is in spite of the fact that these three clusters have the same metal abundance.

As more clusters were investigated it was found that there were some clusters with HBs too blue or too red for their metallicity. Astronomers studying CMDs searched for another parameter that could cause such an effect. The effect was found to be stronger for globular clusters furthest from the galactic center and is very weak for clusters in the central regions of the Milky Way. One of the first papers on the second parameter problem was by Sidney van den Bergh (Dominion Astrophysical Observatory) in 1967, in which he stated: 'It is shown that at least two parameters (one of which is metal abundance) are required to describe globular clusters. ... a second parameter which might be either age or the helium to metals ratio'. Also in 1967 Allan Sandage and Robert Wildey discussed the globular cluster NGC 7006 and pointed out that, although its metal abundance was low, it had a distribution of stars similar to the HB of a cluster of much higher metallicity.

Some additional examples of globular clusters with 'second parameter' problems are M3 and M13; M15, M92 and NGC 5466; Arp 2 and Ruprecht 106. M3 and M13 (top right in figure 3) have [Fe/H] values of about -1.6, but M3 has a redder HB while M13 has a very strong blue HB. (Note that this diagram is a $V, (U-V)$ diagram and not the usual $V, (B-V)$ diagram). Their helium abundances seem to be the same and the differences in magnitude between the main-sequence turnoff and the HB (a measurement of the cluster age) in a $V, (B-V)$ diagram are similar, implying that the two clusters have the same age.

Many parameters have been suggested as the prime second parameter. One of the first suggested was the helium abundance, but later evidence showed that the helium abundance does not vary very much from globular cluster to globular cluster. The abundances of the elements carbon, nitrogen and oxygen (CNO) have been proposed as parameters affecting HB morphology. The density of the stellar environment in which the cluster was formed may have an effect. The core mass and the rate of rotation have an influence on the morphology of the HB. A strong parameter is the age of the cluster. Young-Wook Lee, in his 1989 thesis at Yale University, has presented the results of many models of the HB. Some of these are shown in figure 4. These diagrams help to explain the effects of various parameters on the shape of the HB. The five sets of diagrams are explained in the caption.

A ratio involving the numbers of stars on various parts of the HB has been constructed by Zinn (Yale University) and Lee. It is $c = (B-R)/(B+V+R)$ where B is the number of stars bluer than the RR Lyrae gap, V is the number of variables and R is the number of stars redder than the RR Lyrae gap. It is a good way of assigning a number to the HB distribution, but it does not work well for globular clusters without blue HB stars or without red HB stars.

There are some groups of astronomers (notably the Yale group) who argue that age is the second parameter.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

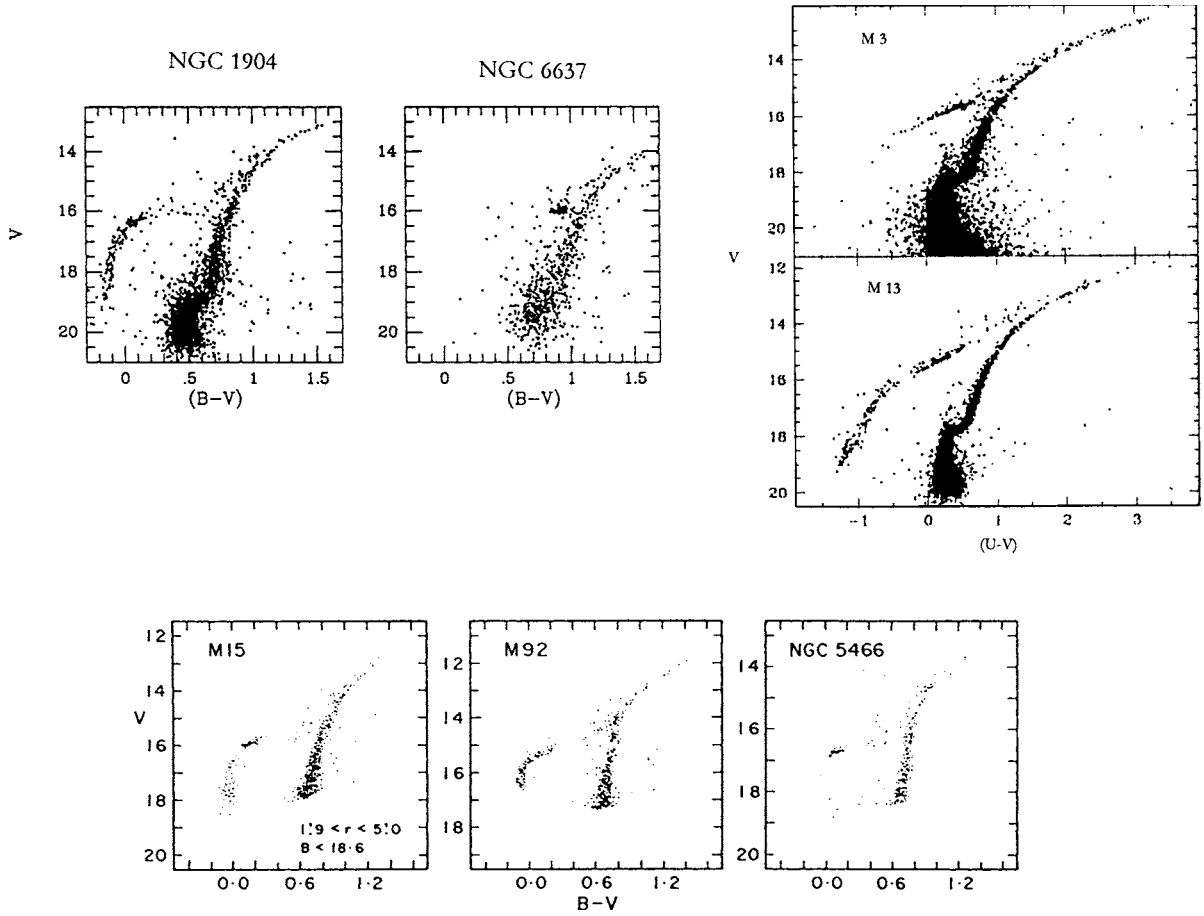


Figure 3. On the upper left is a pair of CMD of two globular clusters showing the normal differences in HB morphology for clusters with different metal abundances. On the upper right are two CMD (V versus $U-V$) of two clusters with [Fe/H] $= -1.6$ but with very different HB morphology. On the bottom are three CMD of clusters with [Fe/H] $= -2.2$, again showing very different HB morphology.

Age variations of several 10^9 years could explain the CMD differences found for globular clusters with the same [Fe/H] but different morphology. However, as we saw in figure 4, there are other possibilities. The Italian group (Flavio Fusi-Pecci, Roberto Buonanno and collaborators) contend that, although age sometimes is a second parameter, the morphology of the HB is the result of the global combination of age and other parameters. The problem has become more complex with the stream of increasingly accurate new data based on the CCD observations being made all over the world. As more spectral observations are made of HB stars in second parameter clusters with the new, very large telescopes already working and those still in construction, there will be many further discussions on the second parameter problem.

Field horizontal-branch stars

In 1933 Albitzky obtained a spectrum of HD 161817 and noted that it had a high radial velocity (-363 km s^{-1}). In 1952 Arne Slettebak noted that the hydrogen lines in HD 161817 were slightly weaker than normal for a star of its type. In 1966 Bev Oke, Jesse Greenstein and James Gunn proposed that four stars (HD 2857, 86986, 109995 and 161817) were field horizontal-branch (FHB) stars on the basis of photoelectric spectrum scans. These are stars which are supposed to be identical to the blue horizontal-branch stars in globular clusters, but which are found in the field and not associated with any cluster. They have the characteristics of low surface gravities, low metal abundance and a high velocity dispersion, all of which place them as members of Pop II.

Further observations of faint blue stars were made by Barry Newell, using his special filter system, and he identified three additional field HB candidates (HD 60778, 74721 and 130095). In 1964 the First Conference on Faint Blue Stars was held in Strasbourg, where reports were made on numerous surveys for faint blue stars and follow-up observations of selected stars. The proceedings contain a bibliography of the important papers in the field and a set of finding charts for the Humason–Zwicky (HZ) stars (a search for faint blue stars that was published in 1947).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

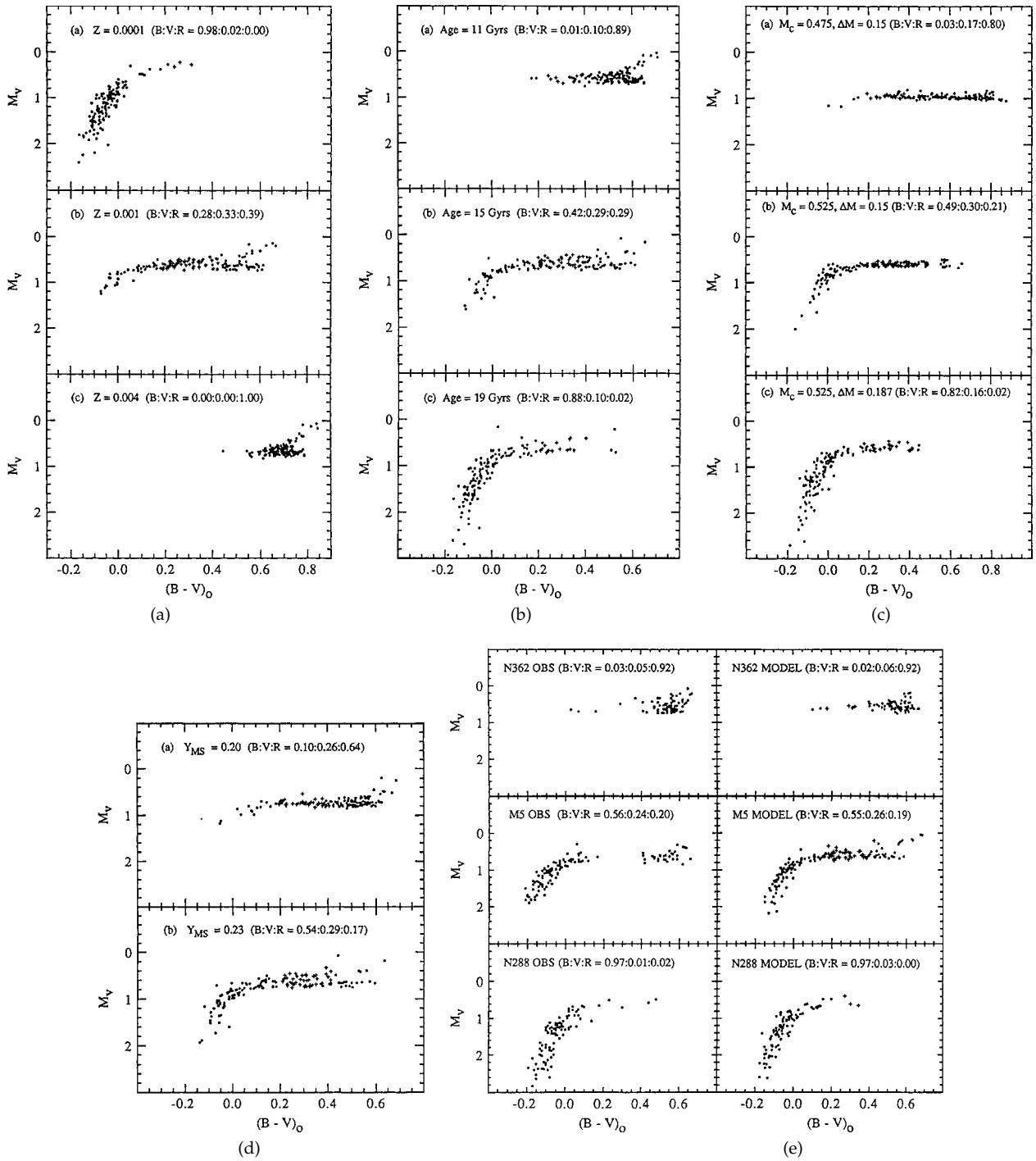


Figure 4. Models published in Y-W Lee's thesis (1989) from Yale University. The upper set of diagrams on the left show the effect of changing [Fe/H]. As the metal abundance approaches the solar value the blue HB does not appear and there is only a short red HB. The middle diagram shows the effect of changing the age of the cluster. Younger clusters have red HBs and the older clusters have blue HBs. The upper right diagram shows the effect of stellar rotation, which affects the core mass. As the rotation increases the CMD becomes bluer. On the lower left the effect of changing the surface helium abundance is shown. HBs become bluer as the helium abundance increases. On the lower right three sets of diagrams compare observational data for the globular clusters NGC 362, M5 and NGC 288 with models. The agreement is quite good.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

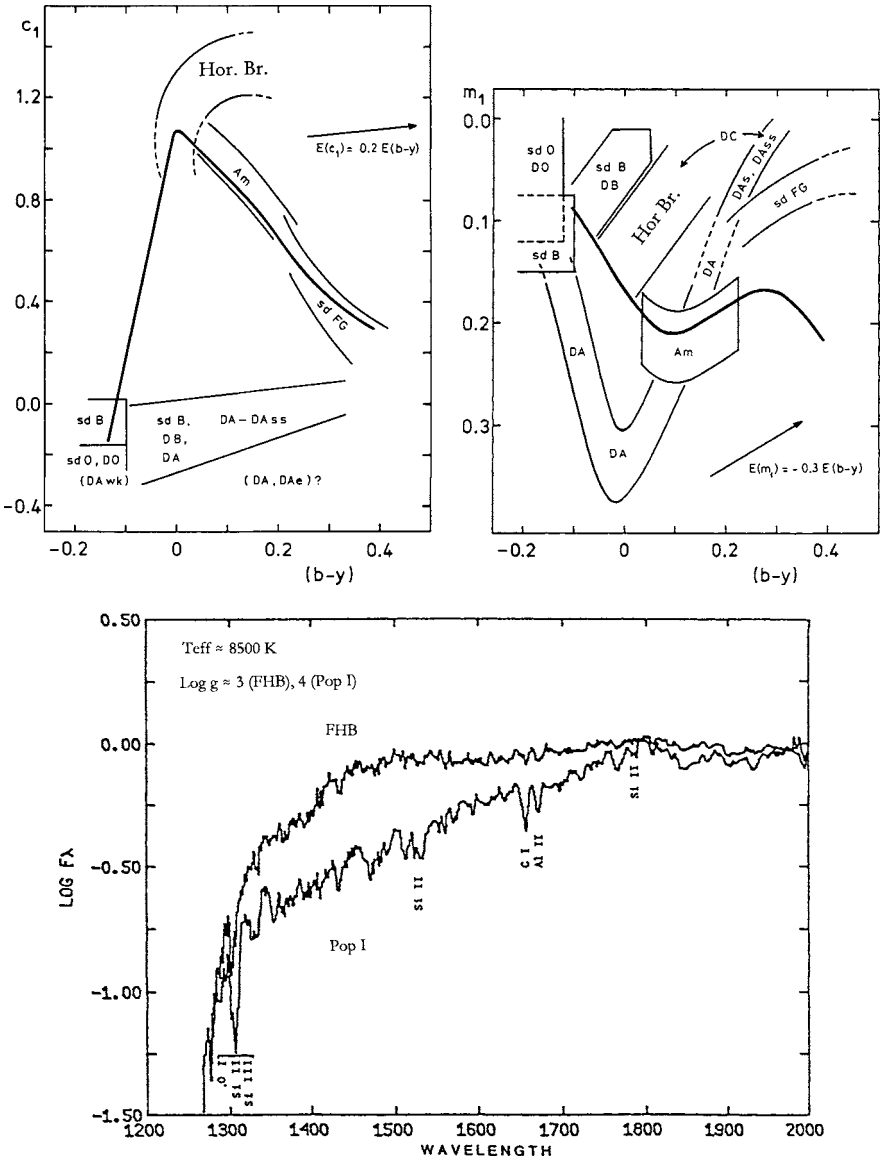


Figure 5. Ways of classifying early-type stars as field HB stars. The two diagrams at the top are from Kilkenny and Hill, showing the location of field HB stars in the Strömgren four-color system. The diagram at the bottom shows mean ultraviolet spectra of an average Pop I A-type star compared to an average Pop II A-type star. The field HB spectrum has a higher luminosity in the region around 1500 Å and does not show the absorption line features seen in the Pop I spectrum.

In the mid 1960s Tom Kinman (KPNO) made a survey for RR Lyrae stars in a field at the North Galactic Pole and a field towards the Galactic center.

In the 1960s Bengt Strömgren set up a new system of photometry, the four-color system (uvby). David Crawford was one of the people who helped calibrate the new system and Davis Philip started using it in the 1970s. John Graham (Carnegie Institution) and Katy Doremus (University of Colorado) made the first four-color observations of blue HB stars in a globular cluster, in NGC 6397. David Kilkenny (SAAO) and Graham Hill (University of St Andrews) published two plots in 1975 showing how the system could be used to classify early-type stars. They are reproduced at the top of figure 5. The leftmost plot shows the c_1 (Balmer Jump index) versus $(b-y)$ and the right plot shows m_1 (the metal index) versus $(b-y)$. The Strömgren four-color system does very well in identifying field HB stars because of the high c_1 index and lower than normal m_1 indices found for these stars. Using this system to measure early-type stars in the High Latitude Field (HLF) survey project lists of field HB stars were published and then summarized by Philip in *Finding Charts for FHB Stars* for 71 stars. As shown at the bottom of figure 5, ultraviolet spectroscopy is a good way to separate

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

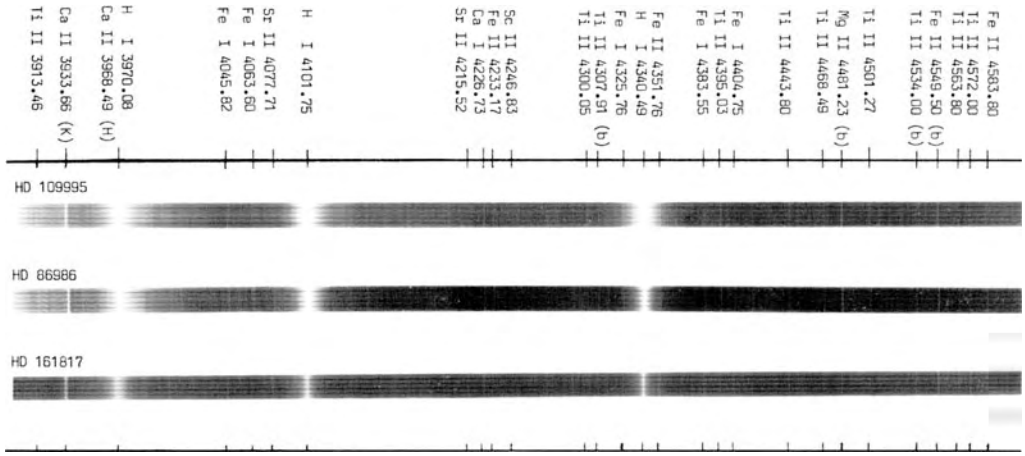


Figure 6. 6 m spectra, at 9 Å mm^{-1}, of three of the prototype FHB stars obtained by Philip and Samus in 1980.

field HB stars from normal Pop I A-type stars.

George Preston, Stephen Shectman (both Carnegie Observatory) and Timothy Beers (Michigan State University) have made a Schmidt objective-prism survey using narrow interference filters centered on the Ca II H and K lines. This HK survey identifies low-metallicity stars and a good proportion of these stars turn out to be field HB stars. They are following up with photometry and spectroscopy and there is now a catalog of thousands of field HB candidates. The new Sloan Survey is now underway and it is also turning up large numbers of faint, metal-poor early-type stars. Another survey, the Hamburg/ESO survey, has identified tens of thousands of field HB candidates. Follow-up spectral studies of these stars are being made with the CTIO/KPNO 4 m telescopes. These investigations will reach stars out to 25 kpc. Future studies of these stars will enable the galactic halo to be studied out to many tens of kiloparsecs.

Figure 6 shows the spectra of three of the original field HB stars obtained with the 6 m telescope in Russia by Philip and Samus. Spectroscopic work of Chris Corbally (Vatican Observatory) and Richard Gray (Appalacian State University), who measured stars from the Philip catalog of field HB stars, found that only two of the stars had the characteristic sharp-line features expected. The rest of the stars had broader hydrogen lines than expected.

The field HB stars are like the blue HB stars in globular clusters in that they have similar four-color indices, metallicities, photoelectric scans, velocities and stellar rotation. However, there are some differences. The field star metallicities reach much lower values of [Fe/H] (to −3.8) than do the cluster stars. Robert Kraft, in a review article, has pointed out that some field stars have lower nitrogen abundance than the cluster stars. The field HB stars could be stars that escaped from globular clusters or stars left in the halo after a globular cluster is broken up or stars captured in the halo from incoming fragments. Now that there are large databases of field HB stars and field HB candidates available, combined with the substantial number of large telescopes online and coming online, much progress in the spectroscopy of these stars should be made.

Jesper Sommer-Larsen (Copenhagen, Denmark) and his collaborators have been analysing the kinematics of the HB stars in the galactic halo and they have found that the radial component of the stellar halo's velocity dispersion decreases fairly rapidly beyond the solar circle. At the same time the tangential velocity dispersion increases with increasing radius. They conclude that the Galaxy may have formed from the merging of smaller subsystems, a formation theory favored by many theorists. For example, Leonard Searle (Carnegie Observatory) and Robert Zinn (Yale University) have proposed such a model from studies of globular clusters.

Ronald Wilhelm (University of Texas) and his collaborators have been making photometric and spectroscopic observations of the A-type stars in the galactic halo. They obtain effective temperatures, surface gravities, [Fe/H] values and radial velocities for approximately 1000 stars. They find that 40% of the stars in the sample were FHB stars, 37% were main-sequence A-type stars, 12% were Am or Ap stars and 11% were unclassifiable.

The masses of field HB stars have been investigated by Klaas de Boer (Bonn University) using Hipparcos data and he finds masses from 0.32 to $0.48 M_{\odot}$ for five field HB stars. Their absolute magnitudes follow the theoretical HB calculated by Ben Dorman (NASA/GSFC). Plans are underway to use the Space Interferometry Mission (SIM) to measure the parallaxes of the 100 or so brightest field HB stars, which will then give us better data to determine their masses and absolute magnitudes.

The HB is a remarkable feature and it is fortunate that nature has produced a situation where the HB covers such a wide range of the CMD. Different initial conditions (such as a higher average [Fe/H], could have given us only short, red HBs and we would not have the opportunity to study the effects of variations of the parameters mentioned above. At present, as more data arrive in the database,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

things are getting ever more complex. The possibilities of gaining a better understanding of the HB will improve with time and new, accurate observations. With this improvement, we will arrive at a better understanding of globular clusters and their formation.

Bibliography

Philip A G D, Liebert J W and Saffer R A (eds) 1999 *The Third Conference on Faint Blue Stars* (Schenectady: L Davis Press)

especially the articles by

de Boer *et al* 1999 *The Third Conference on Faint Blue Stars* ed A G D Philip, J W Liebert and R A Saffer (Schenectady: L Davis Press) p 515

Fusi-Pecci and Bellazzini 1999 *The Third Conference on Faint Blue Stars* ed A G D Philip, J W Liebert and R A Saffer (Schenectady: L Davis Press) p 255

Also:

Adelman S J, Upgren A R and Adelman C J (eds) 1994 *Hot Stars in the Galactic Halo* (Cambridge: Cambridge University Press 1994)

Grindlay J E and Philip A G D (eds) 1988 *IAU Symposium No. 126, The Harlow Shapley Symposium on Globular Cluster Systems in Galaxies* (Dordrecht: Kluwer)

Lee Y-W 1989 The evolution of HB stars and the calibration of globular cluster ages *Thesis* Yale University

Smith G H and Brodie J P (eds) 1993 *The Globular Cluster Galaxy Connection* (Astronomical Society of the Pacific Conference Series, 48)

A G Davis Philip

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Horologium

(the clock; abbrev. Hor, gen. Horologii; area 249 sq. deg.) A southern constellation which lies between Eridanus and Reticulum, and culminates at midnight in early November. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2.

A small, inconspicuous constellation, the brightest star in Horologium is α Horologii, magnitude 3.9. There are no other stars brighter than fourth magnitude. Interesting objects include R Horologii, a Mira-type variable star (range 4.7–14.3, period about 408 days), and NGC 1261, an eighth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Horrocks, Jeremiah (1619–41)

Schoolmaster (?) and astronomer, born in Toxteth, near Liverpool, educated at Cambridge, lived for a mere 22 years. His colleague WILLIAM CRABTREE lived near Manchester, and the group that formed around them constituted Britain's first Keplarians, *nos Keplari* as they called themselves. When he found that Ptolemaic predictions of the positions of the planets were incorrect, Horrocks developed a celestial dynamics, related to Kepler's but not identical to it, which employed concepts from terrestrial dynamics. Horrocks predicted a transit of Venus, and with Crabtree observed it by a telescopic projection method. From the observation, Horrocks calculated the solar parallax, finding a much greater distance of the Sun than anyone before him. The Keplari were the first people to realize the solar system's size and their figure was incredible to some of their contemporaries. When in 1662 the Dutch astronomer Helvetius published some material by Horrocks (*Venus in Sole Visa*), he scaled down Horrock's estimate of the size of the solar system. With his theory of celestial mechanics, Horrocks calculated the motion of the Moon, which FLAMSTEED admired as '... questionless the finest of his monuments'. NEWTON's theory of the motion of the Moon, according to one historian, was 'Newton's interpretation of HALLEY's variation of Flamsteed's version of Crabtree's account of Horrock's lunar theory'. Apparently Horrocks made his own telescope, and developed projection techniques to view the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Horsehead Nebula (B33)

A dark nebula in the constellation Orion, position RA 05^{h} 41.0^{m}, dec. $-02^{\circ}\,24'$. Though small ($6'$ by $4'$), it has the distinctive shape of a chess knight, and is one of the best known of all astronomical images, projected in silhouette against the emission nebula IC 434.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hot Dark Matter

Dark matter made of light neutrinos, with masses of a few electron-volts (eV) or less, is called 'hot dark matter' (HDM) by cosmologists. Light neutrinos would have been 'hot', moving at nearly the speed of light, in the early universe. For a few years in the late 1970s and early 1980s, hot dark matter looked like the best dark-matter candidate. However, HDM models of cosmological structure formation led to a 'top-down' formation scenario, in which superclusters of galaxies are the first objects to form after the big bang, with galaxies and clusters forming through a subsequent process of fragmentation. Such models were abandoned by the mid-1980s after cosmologists realized that if galaxies had formed early enough to agree with observations, their distribution would be much more inhomogeneous than is the case.

Since 1984, the most successful structure-formation models have been those in which most of the mass in the universe comes in the form of cold dark matter (CDM)—particles that were moving sluggishly in the early universe. For a while in the mid-1990s it appeared that a mixture of mostly CDM with 20–30% HDM gave a better fit to the observations than either one or the other. This 'cold plus hot dark matter' (CHDM) theory fitted data on nearby galaxies and clusters only if the average density of matter in the universe were at or close to the critical density ($\Omega_m = 1$). However, like all such critical-density models, CHDM required that galaxies and clusters must have formed fairly recently. It is now clear that this disagrees with observations. The evidence now increasingly favors ΛCDM models, in which CDM and a little ordinary matter make up about a third of the critical density, with a cosmological constant or some other form of 'dark energy' contributing the remainder. This model also helped resolve a crisis regarding the age of the universe. The HDM question has now become how much room is left for neutrinos in such a universe.

To understand the possible role of neutrinos as dark matter, consider how structures such as galaxies formed as the universe expanded. The expansion itself is described by our modern theory of gravity and spacetime, Einstein's theory of general relativity. In order for structure to form, there must have been some small fluctuations in the initial density of matter. The only alternative is that some mechanism generated such fluctuations after the big bang, but the only such mechanisms that have been investigated are 'cosmic defects' such as cosmic strings, and the pattern of fluctuations produced by such defects is inconsistent with the temperature fluctuations observed in the cosmic microwave background (CMB) radiation. However, 'adiabatic' fluctuations—in which all components of matter and energy fluctuate together—occur naturally in the simplest cosmic inflation models and are in excellent agreement with the latest CMB results. The evolution of adiabatic fluctuations into galaxies and clusters is easy to understand: the rich always become richer and the poor become poorer. A 'rich' region of the universe is one that has more matter than average. Although the average density of the universe steadily decreases owing to its expansion, those regions that start out with a little higher density than average expand a little slower than average and become relatively more dense, while those with lower density expand a little faster and become relatively less dense. When any region has attained a density about twice the average, it stops expanding and begins to collapse—typically first in one direction, forming a pancake-shaped structure, and then in the other two directions.

The first HDM boom occurred about two decades ago. By about 1980, improving upper limits on CMB anisotropies were ruling out the previously favored cosmological model, which included only ordinary matter. There was also evidence from a Moscow experiment suggesting an electron neutrino mass of about 20 eV, which would have corresponded to a nearly-critical-density universe in which neutrinos would have constituted most of the matter. In such a cosmology, the primordial fluctuations on galaxy scales are erased by 'free streaming' of the relativistic neutrinos in the early universe. About 1 yr after the big bang, a region about 1 ly across contained the amount of matter in a large galaxy like our own Milky Way. However, the temperature was then about 100 million kelvins (10^4 eV), so each particle had a thermal energy far higher than the rest energy of light neutrinos. As they would therefore have been moving at nearly the speed of light, these neutrinos would have rapidly spread out, and any fluctuations in density on the scale of galaxies would soon have been smoothed back to the average density. The first scales to collapse in such a HDM scenario would therefore correspond to the mass inside the cosmic horizon when the temperature dropped to a few eV and the neutrinos inside it became non-relativistic. This mass turns out to be about 10 000 times the mass of our galaxy, including its dark halo. Evidence was just then becoming available from the first large-scale galaxy surveys that the largest cosmic structures—superclusters—have masses of approximately this size, which at first glance appeared to be a success for the HDM scenario. Superclusters of roughly pancake shape were observed to surround roughly spherical voids (regions where few galaxies are found), in agreement with the first cosmological computer simulations, which were run for the HDM model. In this picture superclusters should have formed first, since any smaller-scale fluctuations in the dominant HDM would have been erased by free streaming. Galaxies then had to form by fragmentation of the superclusters. However, it was already becoming clear from observations that galaxies

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

are much older than superclusters, contrary to what the HDM scenario implies. Also, the apparent detection of electron neutrino mass by the Moscow experiment was soon contradicted by results from other laboratories. Hot dark matter fell into decline.

The CDM model originated in 1982–1984, just as the problems with the HDM model were becoming clear. Proto-galaxies form first in a CDM cosmology, and galaxies and larger-scale objects form by aggregation of these smaller lumps—although the cross-talk between smaller and larger scales in the CDM theory naturally leads to galaxies forming earlier in clusters than in lower-density regions. In this and other respects, CDM models appeared to fit observations much better than HDM. The first great triumph of CDM was that it successfully predicted (to within a normalization uncertainty factor of about 2) the magnitude of the CMB temperature fluctuations, which were discovered in 1992 using the COBE satellite. However, the simplest CDM model, standard CDM (SCDM) with the matter density equal to the critical value ($\Omega_m = 1$), had already begun to run into trouble.

Cosmological theories predict statistical properties of the universe—for example, the size of density fluctuations on various scales, described mathematically by a power spectrum. Sound or other fluctuation phenomena can be described in the same way—for example, low frequencies might be loud, corresponding to relatively high power at long wavelengths. With a given amount of fluctuation power on the large scales probed by COBE (billions of light-years), SCDM has a little too much power on small scales relevant to galaxies and clusters (millions of light-years and less). However, the fact that the SCDM theory could work fairly well across such a wide range of size scales suggested that it had a kernel of truth. Cosmologists began to examine whether some variant of SCDM might work better. For example, in the late 1980s, my former student Jon Holtzman calculated detailed expectations for 96 variants of CDM. When we compared these predictions with the data available in early 1992, it was clear that the best bets were CHDM and ΛCDM, each of which could fit the data on small scales better than SCDM. Both of these variants had been proposed in 1984, when CDM was still a new idea, but their detailed consequences were not worked out until the problems with SCDM began to surface.

Even if most of the dark matter is cold, a little HDM can still have dramatic effects on the small scales relevant to the formation and distribution of galaxies. In the early universe, the free streaming of fast-moving neutrinos would have washed out any spatial inhomogeneities in the neutrino component on the scales that later became galaxies, just as in the HDM scenario. Consequently, fluctuations in the CDM component grew more slowly on these scales, and at the relatively late times when galaxies formed there was less fluctuation power on small scales in CHDM models. Adding a little HDM appeared to be just what was needed to solve the SCDM problem that the galaxy-scale inhomogeneities are too big. Also, there was even a hint from an accelerator experiment that neutrino mass might lie in the relevant range. This was the Liquid Scintillator Neutrino Detector (LSND) experiment at Los Alamos National Laboratory, which recorded a number of events that appear to be $\nu_\mu \rightarrow \nu_e$ neutrino oscillations. Comparison of the LSND data with results from other neutrino experiments allows two discrete values of $| m^2(\nu_\mu) - m^2(\nu_3) |$, around 10.5 and 5.5 eV^2, or a range of values between 0.2 and 2 eV^2. If true, this means that at least one neutrino has a mass greater than 0.5 eV, which would imply that the contribution of HDM to the cosmological density is much greater than that of all the visible stars. Such an important conclusion requires independent confirmation. The Karlsruhe Rutherford Medium Energy Neutrino (KARMEN) experiment results exclude a significant portion (but not all) of the LSND parameter space, and the numbers quoted above take into account the current KARMEN limits. The Booster Neutrino Experiment (BooNE) at Fermilab should attain greater sensitivity and help to resolve this issue.

By 1995 simulation techniques and supercomputer technology had advanced to the point where it was possible to do reasonably high-resolution cosmological-scale simulations including the random velocities of an HDM component. The results at first appeared very favorable to CHDM. Indeed, as late as 1998 a CHDM model with Hubble parameter $h = 0.5$, mass density $\Omega_m = 1$ and neutrino density $\Omega_\nu = 0.2$ was found to be the best fit of any cosmological model to the galaxy distribution in the nearby universe. However, cosmological data were steadily improving, and even by 1998 it had become clear that $h = 0.5$ and $\Omega_m = 1$ were increasingly inconsistent with observations, and that $h \sim 0.7$ and $\Omega_m \sim 1/3$ worked much better. For example, CHDM predicts that galaxies and clusters formed relatively recently, but around 1998 increasing numbers of galaxies were discovered to have formed in the first few billion years after the big bang. Also, the fraction of baryons found in clusters, together with the reasonable assumption that this fraction is representative of the universe as a whole, again gives $\Omega_m \sim 1/3$. That there is a large cosmological constant (or some other form of dark energy) yielding $\Omega\Lambda \sim 2/3$ then follows from any two of the following three results: (1) $\Omega_m \sim 0/3$, (2) CMB anisotropy data implying that $\Omega_m + \Omega\Lambda = 1$, and (3) high-redshift supernova data implying that $\Omega\Lambda - \Omega_m \sim 0.4$. The abundance of galaxies and clusters in the early universe agrees well with the predictions of the ΛCDM model. However, the highest-resolution simulations of this model that were possible in the mid-1990s gave a dark matter spectrum that had more power on scales of a few million light-years than did the observed galaxy power

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

spectrum, although the simulations and data agreed on larger scales. This result was inconsistent with the expectations that galaxies would be more clustered than the dark matter on small scales, not less. When it became possible to do even higher-resolution simulations that allowed the identification of the dark matter halos of individual galaxies, however, their power spectrum turned out to be in excellent agreement with that of galaxies. The galaxies were less clustered than dark matter because galaxies had merged or were destroyed in very dense regions owing to interactions with each other and with cluster centers. This explanation turned a troubling discrepancy into a triumph for ΛCDM.

Thus ΛCDM is the favorite theory today. However, we know from the Super-Kamiokande evidence for atmospheric neutrino oscillations that there is enough neutrino mass to correspond to some HDM, at least $\Omega_\nu \approx 10^{-3}$, about one-fourth as much as the visible stars. How much room remains for a little HDM in ΛCDM cosmologies? The reason there is any upper limit at all from cosmology is because the free streaming of neutrinos in the early universe must have slowed the growth of the remaining CDM fluctuations on small scales. Thus, to have the galaxy structures we see today, there must be much more cold than hot dark matter. From the shape of the galaxy power spectrum observed by the 2dF galaxy redshift survey, the limit on the sum of the neutrino masses is $m(\nu) < 2.2$ eV, corresponding to $\Omega_\nu < 0.05$. This limit on $m(\mu)$ is much stronger than the best current laboratory limit. Astronomical observations that may soon lead to stronger upper limits on neutrino mass—or perhaps a detection—include more precise measurements of the power spectrum from galaxy redshift surveys, data on the distribution of low-density clouds of hydrogen (the 'Lyman-alpha forest') at high redshifts, large-scale weak gravitational lensing data and improved measurements of the cosmic background radiation temperature fluctuations on small angular scales. These types of data can be used to probe for the effects of any free streaming of neutrinos in the early universe, which can lead to less power on small scales depending on the values of the neutrino masses.

Bibliography

For a more technical article on the present subject with extensive references see

Primack J R and Gross M A K 2001 Hot dark matter in cosmology *Current Aspects of Neutrino Physics* ed Caldwell D O (Berlin: Springer)

This is also available online at astro-ph/0007165 and interactively as

http://nedwww.ipac.caltech.edu/level5/Primack4/frames.html

A positive assessment of the old HDM cosmology

Silk J, Szalay A S and Zel'dovich Y B 1983 The large-scale structure of the universe *Sci. Am.* **249**(4) 56

Problems revealed by simulations of the old HDM cosmology

White S D M, Frenk C S and Davis M 1983 Clustering in a neutrino-dominated universe *Astrophys. J.* **274** L1

Predictions for many variants of CDM

Holtzman J 1989 Microwave background anisotropies and large-scale structure in universes with cold dark matter, baryons, radiation, and massive and massless neutrinos *Astrophys. J. Suppl.* **71** 1

Comparison with data

Holtzman J and Primack J R 1993 Cluster correlations for cold and hot dark matter and other models *Astrophys. J.* **405** 428

CHDM simulations

Klypin A, Holtzman J, Primack J R and Regos E 1993 Structure formation with cold plus hot dark matter *Astrophys. J.* **416** 1

Primack J R, Holtzman J, Klypin A and Caldwell D O 1995 Cold and hot dark matter cosmology with $m(\nu_\mu) \approx m(\nu_\tau) \approx 2.4$ eV *Phys. Rev. Lett.* **74** 2160–3

Paper concluding that CHDM was the best match to data

Gawiser E and Silk J 1998 Extracting primordial density fluctuations *Science* **280** 1405–11

The latest published upper limit on the sum of neutrino masses from cosmology (also available online at astro-ph/0204152) is

Elgaroy O, *et al* 2002 New upper limit on the total neutrino mass from the 2 degree field galaxy redshift survey *Phys. Rev. Lett.* **89** 061301

Joel R. Primack

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Hot Stellar Populations in Elliptical Galaxies

The term 'hot stellar populations in elliptical galaxies' refers to the presence of a population component that produces a thermal far-ultraviolet continuum in spectra taken of the centers of elliptical galaxies. This phenomenon, called the 'UV rising branch', 'UV upturn' or 'UVX'[1] is also found in the spectra of S0 galaxy nuclei, and in those of the bulges of spiral galaxies (see GALAXIES: CLASSIFICATION).

The UVX was first discovered by the OAO-2 (see ORBITING ASTRONOMICAL OBSERVATORY). Since then, it has been found to some measure in all of the above systems observed to date. Additional data have been collected with the IUE (see INTERNATIONAL ULTRAVIOLET EXPLORER), the Astro Shuttle missions (Hopkins Ultraviolet Telescope and Ultraviolet Imaging Telescope) and with the HST (see HUBBLE SPACE TELESCOPE).

The most striking characteristic of the UVX phenomenon is its high variability among galaxies, ranging by a factor of 10 in intensity as measured by the color $(15 - V)$, the ratio of a broadband magnitude centered at 1500 Å to the optical light (see figure 1). Arguably the single most important observation, which has dominated work on the far-ultraviolet radiation from galaxies, is the apparent correlation with galaxy abundances first noted by Faber (1983). The variability and apparent relation to galactic abundances, as well as evidence for relationship to dynamic properties, have made the UV upturn an important issue for the study of the origin and formation of galaxies. It represents a discriminant among objects that is not available from the optical record.

The thermal nature of the radiation—corresponding to a blackbody spectrum with $T_{\rm eff} \sim 20\,000$ K—pointed to an origin in starlight, rather than one in galactic nuclear activity, for this ultraviolet flux component. It was also noted that the old stellar populations expected to constitute the ellipticals were likely to contain post-asymptotic giant branch stars (see STELLAR EVOLUTION, RED GIANT STARS, PLANETARY NEBULAE), which become very hot during their evolution from the red giant branch (RGB) to become white dwarfs. The other obvious candidate for the origin of the flux was a population of normal early B stars, products of recent star formation.

It was later realized that hot (blue) horizontal branch (HB) stars (see HORIZONTAL BRANCH STARS) were a copious source of ultraviolet radiation. HB stars are differentiated according to the size of their hydrogen-rich envelope that remains after the RGB phase, where mass loss takes place. Blue horizontal branch stars (with $8000 < T_{\rm eff} < 20\,000$ K) are abundant in the Galactic globular clusters (see GLOBULAR CLUSTERS) of low metal abundance. The hottest HB stars have extremely thin hydrogen envelopes, less than 0.05 solar masses, and during central helium burning evolution have surface temperatures up to 36 000 K. Their thin envelope cannot support a return to the giant branch after the central helium is exhausted, and thus they continue to emit copiously in the far-ultraviolet until their evolution to the white dwarf phase. These are termed extreme horizontal branch (EHB) stars.

However, the appearance of hot HB stars in metal-rich stellar populations was thought to be rare since, until about 1995, they had never been observed in globular clusters with metal abundances approaching that of the Sun. Indeed, until the last few years, the existence of EHB stars was thought to be rare even among the metal-poor globulars, whereas the elliptical galaxies have spectra implying a dominant population of stars exceeding the solar abundance (see ELLIPTICAL GALAXIES).

On the other hand, as was known from studies in the early 1970s, the field population of our own galaxy contains many hot subdwarf (sdB and sdO) stars. Their location on the Hertzsprung–Russell diagram coincides with the EHB stars and their post-HB progeny. However, the possible connection between the UVX phenomenon and the field hot subdwarf population was not realized until much later.

In the last few years, great strides have been made in observing hot populations in ellipticals, and these have all but ruled out an explanation in terms of radiation from young, massive stars, as is detailed in the next section. The widely accepted model is now of a hot, evolved stellar population of EHB stars whose numbers and spatial distribution may reflect both galactic enrichment history and dynamical properties. The study of the UVX has therefore concentrated on determining not only the origin of the flux but its relation to other known properties of galaxies, both of population and of structure.

Summary of data

A number of different observations have provided constraints on the originating population of the UV upturn. To summarize, the main points are as follows.

- It is observed in all elliptical, S0, and spiral bulges that have signs of neither nuclear activity nor recent star formation.
- Its amplitude, measured by the color $(15 - V)$, varies by 2.5 magnitudes and attains $15 - V \sim 2$ in the brightest example, NGC1399 (Burstein *et al* 1988) (figure 2).

[1] It is also sometimes confusingly referred to as the 'UV excess', which has a separate meaning in stellar photometry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

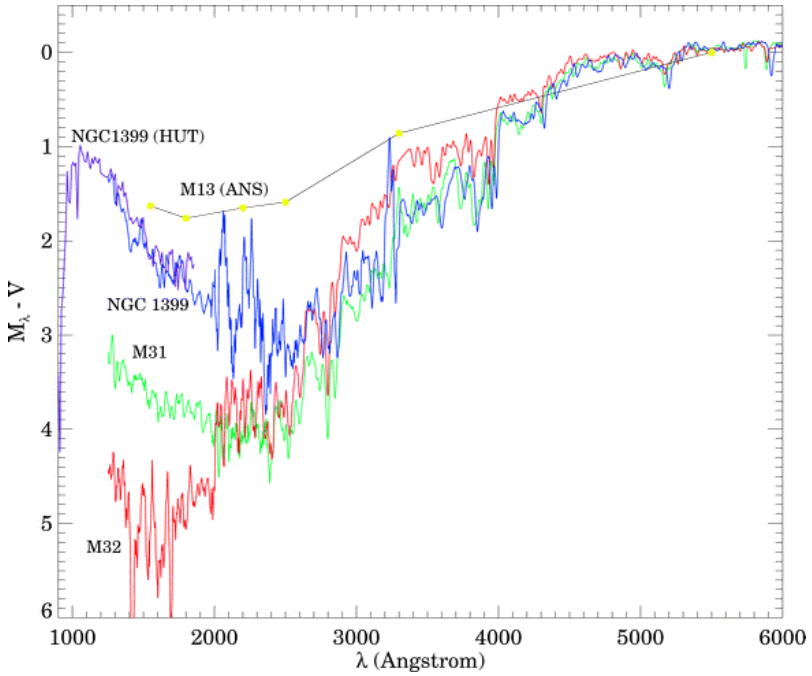


Figure 1. The UVX phenomenon as seen with IUE and HUT data. The spectra are of NGC1399 (uppermost), M31 (middle) and M32 (lowermost). For comparison is shown also the flux distribution of the Galactic globular cluster M13 taken with the Astronomical Netherlands Satellite (ANS); its ultraviolet radiation is thought to emanate from similar stars (see text). The galaxy spectrum is, however, harder (emanating from hotter stars) and has a much steeper descent from the optical to the mid-ultraviolet wavelengths, which is a consequence of the higher abundance of the background population.

- It is positively correlated with the optical magnesium index Mg_2 (Burstein *et al* 1988)[2], which in turn correlates with the galaxy central velocity dispersion σ_0 (a measure of the galaxy mass). It is also correlated with other light element indices. The UVX is not, however, correlated with indicators of iron-peak element abundance. Burstein *et al*, noting both relationships, suggested that the abundance correlation was tighter than that in σ_0 and therefore more likely to be the causal factor. Additionally, there is not yet a theory linking variations in the stellar content (such as the production of hot stars) with galactic dynamical properties, whereas the effect of abundance on stellar populations can be studied using stellar evolution theory.
- Hopkins Ultraviolet Telescope (HUT) spectral data, reaching wavelengths beyond the Lyman break at 912 Å, give an effective temperature of 20 000–25 000 K for the far-UV flux in a sample of galaxies (Ferguson and Davidsen 1993). The narrow range of characteristic temperatures argues against an interpretation in terms of young massive stars, as did the lack of strong metal absorption lines characteristic of the spectra of such objects. The characteristic temperature also disfavors an explanation purely in terms of post-AGB stars, which radiate much of their energy while at higher temperatures.

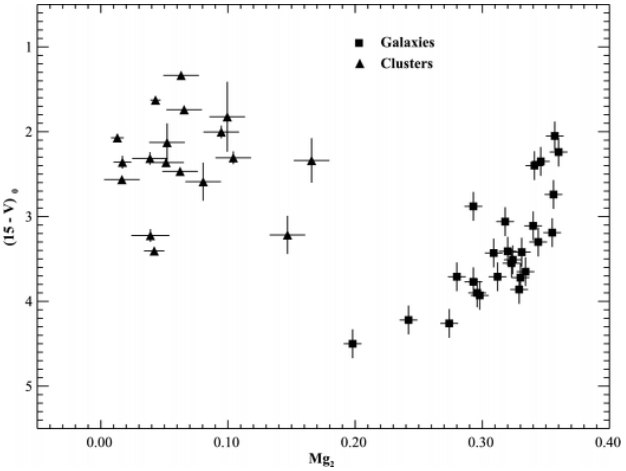


Figure 2. The 15 – *V* versus Mg_2 (UV upturn versus abundance) diagram. To the right are the galaxies, with strong absorption lines. For comparison plotted to the left are the Galactic globular clusters whose UV bright populations are resolved and studied in detail. They are weaker-lined (low-abundance) systems.

[2] Mg_2 is an abundance indicator that is strongly correlated with iron-peak abundances in (metal-poor) Galactic globular clusters, and is widely used as an abundance indicator in galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- HUT data also demonstrate qualitative differences in the spectral slopes between two systems, M31 in Andromeda and NGC1399 in Fornax (Ferguson and Davidsen 1993), implying that the UVX might emanate from a blend of sources with different properties.
- The spatially resolved data show the UV flux to be distributed in a fashion that is morphologically similar to that of old stellar populations, i.e. smooth and centrally concentrated (see figure 3). This is in contrast to the appearance of young star associations, which are clumped and dispersed from the center (O'Connell *et al* 1992). The data also show the 15 – *V* measure of UVX intensity to possess a radial gradient that is usually steeper than that of any optical index (O'Connell *et al* 1992).
- Ultraviolet photometry of Local Group galaxies, in which individual sources can be resolved and measured, confirms the presence of hot, old stellar populations in those systems, and a lack of evidence for massive stars (O'Connell *et al* 1992, Brown *et al* 2000).
- Radial gradients in the 15 – *V* color do not strongly correlate with radial gradients in the Mg_2 index (Ohl *et al* 1998). Thus, abundance cannot be the sole factor causing the strength of the UV upturn.
- The UVX has been observed with Hubble Space Telescope in galaxy clusters at redshifts $z = 0.375$ and 0.55, corresponding to lookback times of 4 and 6 Gyr. These two datasets give the first evidence that the phenomenon evolves with time, and may in future give evidence that constrains the age of galaxy formation.

Fuller reviews of the data and their implications are given in Burstein *et al* (1988), Greggio and Renzini (1990), Dorman *et al* (1995) and O'Connell (1999).

Hot evolved stellar populations

The extensive theoretical review of Greggio and Renzini (1990) named as possibilities evolved helium burning objects and their descendants, post-RGB remnants (stars losing so much mass on the red giant branch that they proceed directly to the white dwarf cooling phase), and various interacting binary star scenarios. However, the subdwarf population of our Galaxy dominates the detections of blue sources in field surveys looking for hot objects. Objects of this type, even while many of them have companions, have luminosities that imply a single- rather than multiple-star origin. This suggests that a population of hot horizontal stars may be a normal constituent of evolved stellar populations. Perhaps this also provides the strongest plausibility argument for the prevailing hypothesis that hot, helium burning stars are indeed responsible for the UVX phenomenon, rather than any of the other more exotic candidates that have yet to be fully investigated.

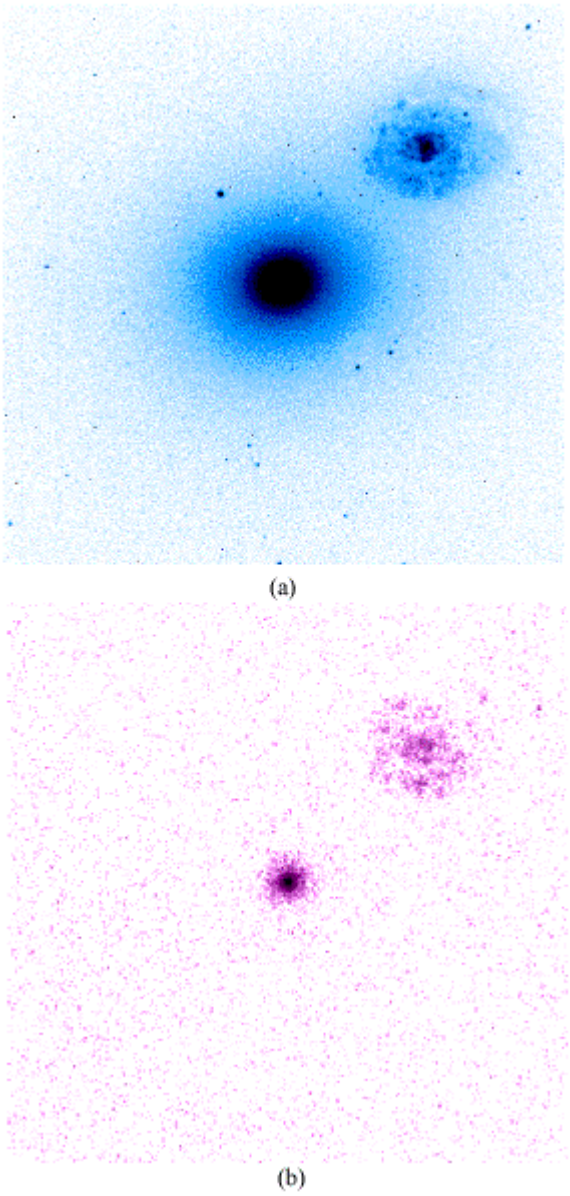


Figure 3. Elliptical and spiral galaxy optical–far-ultraviolet comparison. (*a*) B-band exposure of the same field obtained at Palomar Observatory. (*b*) Far-ultraviolet image of the elliptical M60 (NGC 4649) and Sc spiral galaxy NGC 4647, taken from the Ultraviolet Imaging Telescope. Note the contrast in the distribution of ultraviolet radiation between the elliptical and the spiral galaxy. From Ohl R G 2000 *PhD Thesis* John Hopkins University.

Much is known about hot evolved stars from the Galactic globular cluster system in which large HB populations can be resolved and studied in detail. The properties of stars in the field hot subdwarf population smoothly mesh with the hottest stars in these clusters. Additionally, the field population of hot subdwarfs is identified kinematically with the old disk population of our Galaxy which is thought to have an abundance distribution similar to that of the galaxies. Hence, our

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

understanding of the potential source of the UVX is augmented by analogy with both old Galactic hot stellar populations.

From the fuel consumption theorem (see next section), sources of far-UV radiation in old stellar populations that are most significant are those which are powered by stable nuclear reactions. The evolved hot stellar population components comprise stars in the post-asymptotic giant phase, hot and extreme horizontal branch stars, and the post-horizontal branch phases of extreme HB stars (see figure 4 for a schematic). Detailed evolutionary sequences and discussion can be found in Dorman *et al* (1993) and references therein.

Post-asymptotic giant stars are the central stars of planetary nebulae. When the hydrogen envelope of an asymptotic giant branch star is largely consumed, the star evolves rapidly towards higher temperatures. During this evolution, ending on the white dwarf cooling curve (see WHITE DWARFS) they span a temperature range of about 3000 to in excess of 100 000 K at a fixed luminosity that exceeds $2500L_{\odot}$. The lifetime in this phase, however, is short (since the fuel is rapidly consumed), generally a few times 10^4 yr to only a few thousand years. It decreases strongly with increasing luminosity, and the total energy radiated decreases with it. The spectrum integrated over this phase has a high characteristic temperature.

Horizontal branch stars are formed after a star reaches the tip of the red giant core, undergoes an off-center helium flash and subsequently develops a stable helium fusion source. The HB lifetime is substantial, about 100–150 Myr, at a luminosity of $(10–50)L_{\odot}$, so that stars that are sufficiently hot are important UV emitters. The key parameter determining ultraviolet radiation from HB stars is the size of the hydrogen-rich envelope, denoted $M_{\rm env}$, at the onset of stable central helium burning. Because the helium flash takes place at a fixed temperature in red giant interiors, the size of the helium core inside new HB stars varies relatively little with envelope mass (or, for that matter, abundance and age), being about $0.5M_{\odot}$. The temperature of an HB star increases with decreasing $M_{\rm env}$, which also corresponds to decreasing hydrogen shell luminosity. The most extreme HB stars (see below) are, initially, almost identical to pure helium burning stars.

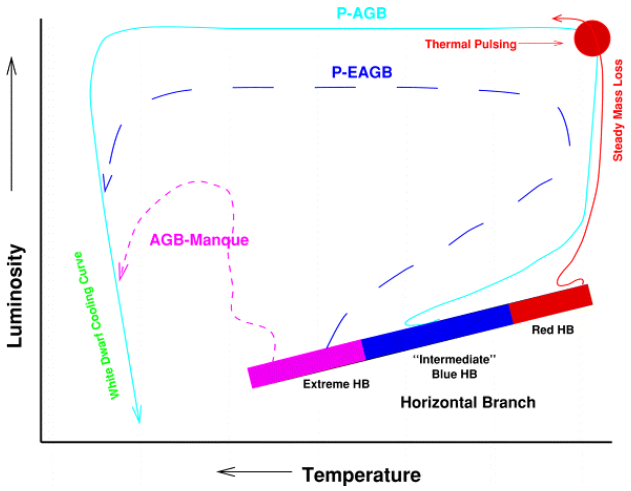


Figure 4. Schematic of the various types of hot, evolved star shown in the theoretical plane. From Dorman *et al* (1993).

Observations of Galactic globular clusters demonstrate that in real stellar populations $M_{\rm env}$ for the HB stars is often scattered, for reasons not properly understood. The distribution arises because stars in the population lose varying amounts of mass in the preceding red giant branch stage of evolution. The observational record shows that, with a few notable exceptions, stars in clusters that have relatively high heavy element abundance are usually cool ($T_{\rm eff} \sim 4000–5000$ K) and are therefore dark in the UV. In contrast, HB stars in clusters with low abundance are found to span a large range in temperature, $T_{\rm eff} \sim 6000–30\,000$ K.

Numerical models of HB stars show that stars sufficiently hot to radiate in the ultraviolet can be formed for almost any chosen abundance. However, the amount of mass that must be lost to make HB stars that are strong UV emitters is dependent on their heavy element abundance Z and also their helium abundance Y. For stars with high abundance, nearly the entire mass of the envelope must be stripped in order to produce a hot star. At low Z, models with substantial hydrogen-rich envelopes can be hot enough to radiate strongly in the far-UV.

After the core helium is exhausted, helium shell-burning is initiated. If the remaining envelope at this stage is sufficiently large (greater than $0.05M_{\odot}$) the hot HB stars return to the AGB (i.e. they develop a deep, cool convective exterior) as do their more massive red HB counterparts. They evolve through the thermally pulsing stage, finally becoming P-AGB stars and white dwarfs.

The designation ‘EHB star’ is used to refer to stars that do not reach the P-AGB stage of evolution. These stars have envelope masses $\geq 0.05M_{\odot}$, have weak or insignificant hydrogen shell burning during helium core burning and effective temperatures that range from about 20 000 to 37 000 K depending on both envelope mass and abundance. After core exhaustion they become brighter and develop significant hydrogen shell energy sources and are called ‘post-early AGB’ stars or ‘AGB manqué’ (‘failed AGB’) stars depending on whether or not they reach the early part of the AGB evolutionary sequence.

The later evolutionary stages of the EHB stars do not have enough hydrogen to develop a cool outer convection zone after the central helium source is exhausted. These stars remain hot in the post-HB phase and emit similar amounts of UV radiation over their post-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

HB lifetime to those during core helium burning. These are the AGB manqué stars. They have lifetimes typically 20–40 Myr (similar to the AGB itself), with $L \sim (100–500)L_{\odot}$. They reach effective temperatures of 80–10 000 K before settling on the white dwarf cooling sequence. The continuum of possible evolutionary paths includes also a transitional type, the 'post-early AGB' stars, that reach the earlier but not the later stages of evolution on the asymptotic giant branch. These stars resemble the behavior of AGB stars, but are slightly less luminous and somewhat longer lived.

Observations indicate that there is a strong 'stellar populations' component to the mass loss during the RGB phase. That is, the mass loss phenomenon varies among populations, so that systems that are apparently similar in the gross properties that govern stellar evolution do not produce the same HB stellar distribution ('HB morphology')[3]. This complicates the explanation of HB populations in general, and the UVX in particular.

For example, one simple reason why HB morphology might be different between two populations is age. As a population ages, the mass of the stars at the tip of the red giant branch decreases. All things being equal—in particular, the abundance of the stars and distribution of total mass lost among stars of that abundance—one would expect the distribution of $M_{\rm env}$ on the HB to be smaller for older stars and thus the probability of forming more hot stars increases. However, the observational record only supports this explanation for HB morphology differences in a limited way on a subset of the Galactic globular cluster system, and thus the observations of a particular $15 - V$ color cannot be used to give a definitive age to a stellar population.

Nevertheless, for a sufficiently 'young' evolved population, one does not expect to see far-UV radiation. For example at age 5 Gyr, the mass of red giants in metal-abundant galaxies is about $(1.4–1.5)M_{\odot}$. The amount of mass loss required to produce hot stars, about $1M_{\odot}$, is much greater than in populations of age 14 Gyr, whose giants have mass $\sim(0.8–0.9)M_{\odot}$. If mass loss processes act in similar fashion in both cases, the number of hot stars in the younger population will be smaller. Thus it is predicted that the UVX arises later in galaxy evolution, or that the restframe UV radiation should fade with increasing redshifts of distant galaxy clusters. The observations of galaxies at moderate redshift are consistent with this prediction.

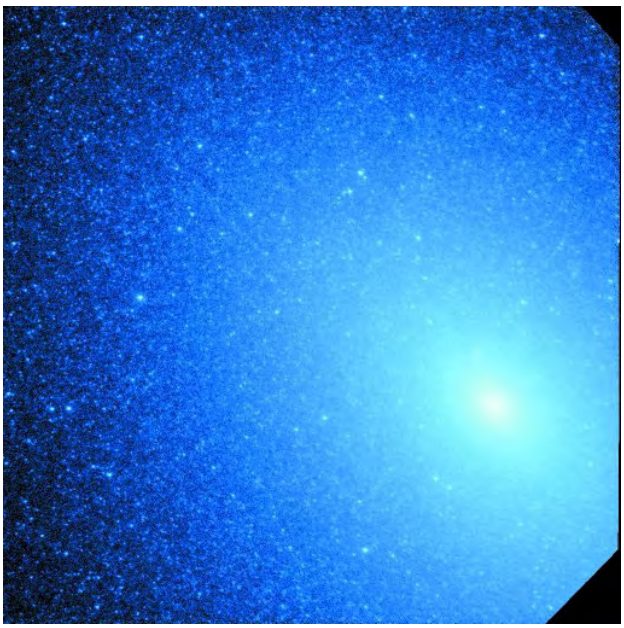


Figure 5. The local group galaxy M32, imaged in the UV (1450–3500 Å) with the Space Telescope Imaging Spectrograph of the Hubble Space Telescope. The individual stars in the picture have magnitudes consistent with being hot HB or post-HB stars. From Brown *et al* 2000, © *The Astrophysical Journal*, reproduced with permission.

The fuel consumption theorem and synthetic populations

The UVX phenomenon can be studied using the methods of evolutionary population synthesis (see SYNTHETIC STELLAR POPULATIONS). Briefly, such methods build a population model from the constituents of the population under study using results from stellar evolution calculations. Then, integrated spectra are computed using either synthetic stellar fluxes computed from model atmospheres and synthetic stellar flux calculations, or a library of empirical fluxes derived from observations. Finally, colors such as the $15 - V$ measure can be derived as a function of any input population model by integrating a synthetic spectrum over desired filter bandpasses.

To compute the synthetic $15 - V$ color of a stellar population, it is necessary to select the following ingredients for the model:

(a) The population responsible for the far-UV radiation.
(b) The constituents of the 'background' population. For elliptical galaxies the bulk of the population is unambiguously composed of older stars ($T_{\rm eff} < 7000$ K) on the main sequence and red giant branch, none of which emits significant far-UV radiation (see figure 5). This part will consist of a set of one or more isochrones representing populations of a set of ages and abundances and the assumption of a mass function.

[3] This problem is known as the Second Parameter Effect, the 'first parameter' being metal abundance: see GLOBULAR CLUSTERS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(c) A measure of the input rate of stars into the late evolutionary stages. This can found empirically through relative number counts or theoretically by differentiation performed on stellar isochrone sequences (see Dorman *et al* 1995).

The fuel consumption theorem (see e.g. Renzini and Buzzoni 1986) states that the total bolometric flux produced in an advanced phase of stellar evolution is proportional to the fuel consumed during that phase. It is instructive to sketch the derivation of this result and to show how it may be extended to synthesize stellar population properties.

The energy radiated by a star during any evolutionary phase is

$$E = \int_{\text{phase}} L \, \mathrm{d}t.$$

This energy is, to a good approximation, produced by nuclear burning and therefore proportional to the fuel consumption.

In a large stellar system, the number of stars evolving off the RGB to the HB is a Poisson process, with the input rate to the HB having a finite mean. Since this mean varies on a timescale larger than the length of any subsequent phase, the numbers of stars in that phase will approach a steady state.

If we denote this mean input rate by $\dot{N}$, and divide the evolutionary path from the beginning of the HB phase to its end by a set of time intervals $\{t\}_i$, then the number of stars in the jth time interval $\Delta t_j = t_j - t_{j-1}$ is

$$\Delta N_j = \dot{N} \, \Delta t_j.$$

The total luminosity of all the stars along the track is then

$$L = \dot{N} \sum_j L_j \, \Delta t_j.$$

In the limit where the population is large,

$$\dot{N} \, \Delta t_j \gg 1$$

(i.e. there are expected to be many stars in each interval j), this equation becomes an integral, with

$$L = \dot{N} E.$$

This is the fuel consumption result. Derived in this form, one may also see that the luminosity of a stellar population in a given passband denoted by λ is proportional to the total energy radiated in that band given by their integrated luminosity at λ over their lifetime multiplied by their rate of formation, i.e.

$$L_\lambda = \dot{N} \int_{\text{birth}}^{\text{death}} L_\lambda \, \mathrm{d}t = \dot{N} E_\lambda.$$

The terms in these equations can be estimated from fits to stellar evolutionary sequences (Renzini and Buzzoni 1986), or computed directly from the stellar evolution calculations (Dorman *et al* 1995) if the bolometric correction at λ, $\alpha_\lambda = L_\lambda/L$, is also known. Bolometric corrections are derived from the stellar flux library adopted for the analysis.

Colors can then be computed for the synthetic stellar population and compared with observations. For example, to compute the $15 - V$ color one needs to calculate both L_{1500}, which arises purely from the hot population, and L_v, the visual brightness, which consists of

$$L_v = L_{\text{evolved}} + L_{\text{RGB}} + L_{\text{MS}}$$

since all of the evolved, red giant branch and main sequence stars radiate at optical wavelengths. The contribution from the two earlier stages is evaluated at fixed age and abundance by assuming a stellar isochrone[4] and a mass function describing the number of stars forming of a given mass.

The synthetic color of the model is then given by

$$15 - V = -2.5 \log(L_{1500}/L_v) + \text{constant}$$

where the constant is a normalization chosen by convention (-21.1 on the monochromatic scale), and similarly for other colors.

Quantitative analyses of the UVX

The object of quantitative analysis is to understand how the variation of UVX strengths in galaxies might arise, how that might account for the apparent variation of the amplitude with abundance, and what range of model parameters are consistent with the behavior of spectra such as are illustrated in figure 1.

Of particular utility is a measure of the lifetime-integrated ultraviolet energy, e.g. E_{1500}, the energy radiated in a bandpass centered at 1500 Å. One may evaluate this function for models of various potential contributors to the UV upturn and thus quantitatively estimate under which circumstances each will dominate the observed radiation.

For horizontal branch models, E_{1500} is a function of M_{env}, Z, Y and age. By varying the assumed mass distribution of HB stars the effect of various hypotheses concerning its behavior can be studied. In a population

[4] An isochrone is a curve constructed from stellar evolution calculations defining the locus on the theoretical Hertzsprung–Russell diagram joining stars at a chosen age.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

model, for any given abundance E_{1500} (M_{env}) must be integrated over a mass distribution $P(M_{env})$, with unit integral over the range of the available horizontal branch masses. That is,

$$E_{\lambda}^{HB} = \int_{blue}^{red} P(M') E_{\lambda}(M') \, dM'.$$

A detailed study of the stellar models gives the following conclusions.

- The peak contribution from extreme HB stars is 20–30 times greater than for post-AGB stars, so that if the former exist in significant numbers they will dominate the UVX.
- The maximum possible post-AGB flux in a population cannot produce the strongest UV upturns that have been observed.
- E_{1500} itself is insensitive to age: it is the 'population' function $P(M)$ that gives rise to any age variation.
- E_{1500} is also relatively insensitive to the abundances of stellar populations, varying by some tens of percentage points rather than by orders of magnitude. This suggests that the variations in the UVX we see are due to variations in the number of sources that are present.
- If all of the red giants in an old stellar population turned into EHB stars, the resulting UV flux would be about 6 times stronger than in the strongest UVX galaxy. Thus, in the most extreme UVX object, NGC 1399, about 16% of the HB stars are hot.
- The range of masses for which E_{1500} is significant at high Z is somewhat smaller than at low abundance, and thus abundance alone cannot drive the UVX: one needs to invoke a variation in mass loss too.
- Surprisingly, the range of masses for which E_{1500} is significant increases with increasing Y, so that it is easier to make EHB stars in helium-enriched populations.

Figure 6, showing samples of galaxy and globular cluster data, supports the hypothesis that variability in the UVX is caused by a variation in numbers of sources. In figure 6(*a*), the globular clusters, which have known variation in the HB morphology (and therefore in numbers of UV emitting sources) and low abundance, are compared with the UVX data for galaxies, which have a spread in UVX strength and high abundance. The sequences of different types of objects separate into almost parallel lines. Figure 6(*b*) shows that the reason for the separation is the relative weakness of the galaxies in the mid-ultraviolet, at 2500 Å. This is the direct result of the steeper slope of the galaxy spectra compared with the cluster spectrum seen in figure 1.

Stellar population models constructed using the methods briefly described in the previous section can indeed reproduce the location of the data on these two diagrams. Hence, despite the uncertain relationship between the (far-ultraviolet) UVX and galaxy abundance, the mid-UV shows some promise as a population abundance indicator (but see SYNTHETIC STELLAR POPULATIONS for other issues arising from population indicators). Although the mid-UV fluxes of galaxies are strongly affected by the presence of a strong UV upturn component, the relative invariance of its temperature yields simple methods for subtracting this effect if it has been quantified by far-UV observations.

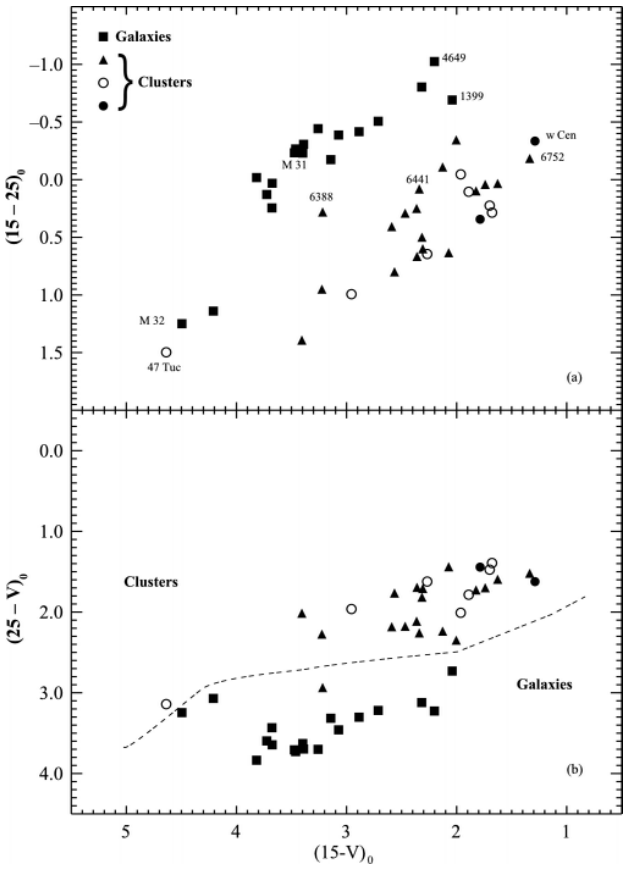


Figure 6. The UV characteristics of galaxy and globular cluster data compared (from Dorman *et al* 1995). The globular clusters (triangles and circles), with known variation in numbers of HB stars and low abundance, form sequences parallel to the galaxies (squares) with varying UVX strength and high abundance in mid-UV color–magnitude and color–color diagrams.

Various lines of argument have led researchers to adopt three distinct explanations for the galaxy population from which the UVX stars are drawn. All of these assume that the elliptical galaxies have an abundance spread which is similar to that found in our own Galaxy, and all appeal to different abundance ranges to explain the phenomenon.

(a) The UVX arises from the metal-poor tail of the galactic abundance distribution, from stars similar to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

those found in the Galactic globular clusters. By extension of the age effect on HB morphology, the high temperature of the UVX relative to most globular clusters implies that the phenomenon implies an age for the galaxies greater than the globular clusters.

(b) Assuming that the helium content of high-Z stars continues to increase for $Z > Z_{\odot}$ as is found for $Z < Z_{\odot}$, (i.e. $\Delta Y/\Delta Z > 2.5$, so that for $Z = 3Z_{\odot}$, $Y \sim 0.4$) the UVX phenomenon will arise naturally from the high end of the galactic abundance distribution, also at great age.

(c) No inference can be drawn concerning the abundances of the UVX stars since the intrinsic properties of the HB vary only weakly with abundance parameters compared with the (undetermined) nature of mass loss. The most plausible assumption is that the UVX stars are most likely to be drawn from the peak of the abundance distribution since it is by definition the most numerous background population. According to this scenario, the UVX is not directly related to abundance but indicates something yet to be understood about the history of stellar populations in galaxies.

Of these hypotheses, the first two both interpret the UVX phenomenon as an age effect based on the argument given earlier for HB morphology as largely an age effect. The 'low-abundance' hypothesis (a) appears however, to be inconsistent with the observations, at least for the strongest UVX galaxies. Since these require about 15% of all HB stars to be of EHB type, a low-Z subcomponent of this size would radiate more strongly in the mid-UV than the observations allow. The 'high-Z' abundance argument depends on an uncertain extrapolation of the helium enrichment function for these stars. Since it appears both that the galaxy abundances are in fact more strongly enhanced in lighter elements and that the UVX is not correlated with Z, helium content in stars increases with elements such as magnesium. Finally, the third, most conservative, hypothesis leaves the direct interpretation of the UVX inconclusive.

While it is common in astronomy to appeal to the need for more observational data, this is particularly true in the ultraviolet. Observations require space-based astronomy and the number of observing platforms that have been able to collect sufficiently high-quality data of distant galaxies has been small.

A final point is that the UVX is observed to be correlated with the light element index Mg_2 and not with indices measuring the iron-peak elements (E_{1500} should be relatively insensitive to light element variations, since it is sensitive to the deep structure of the star which is unaffected by light element abundance). Further, resolved observations of the UVX indicate that it does not vary radially the same way that abundance measurements do. Hence, despite the initial optimism for understanding the UVX in terms of a variation in population abundance, later evidence implies that the relationship is far from straightforward.

There are a couple of lines of evidence tentatively linking the UVX to dynamical properties of galaxies rather than abundance, a link for which there is currently no explanation. The data suggest a relationship between the shape of galactic isophotes (curves of constant brightness) such that galaxies with 'boxy' isophotes associated with possible merger products are also galaxies with stronger UV upturns. Also, there is the possibility that the UVX–Mg_2 correlation actually consists of several distinct groups rather than a continuous relationship. In this view the brightest UVX galaxies, which correspond to giant ellipticals with the largest central velocity dispersions, form a group with distinct ultraviolet properties. The UV upturn may thus be an as yet inscrutable indicator of a relationship between galaxy dynamics and stellar populations.

Bibliography

Brown T M, Bowers C W, Kimble R A and Sweigart A V 2000 *Astrophys. J.* **532** 308

Burstein D, Bertola F, Buson L, Faber S M and Lauer T R 1988 *Astrophys. J.* **328** 440

Dorman B, O'Connell R W and Rood R T 1995 *Astrophys. J.* **442** 105

Dorman B, Rood R T and O'Connell R W 1993 *Astrophys. J.* **419** 596

Faber S M 1983 *Highlights of Astronomy* **6** 165

Ferguson H C and Davidsen A F 1993 *Astrophys. J.* **408** 92

Greggio L and Renzini A 1990 *Astrophys. J.* **364** 35

O'Connell R W 1999 *Annu. Rev. Astron. Astrophys.* **37** 603

O'Connell R W, Bohlin R C, Collins N R, Cornett R H, Hill J K, Hill R S, Landsman W B, Roberts M S, Smith A M and Stecher T P 1992 *Astrophys. J. Lett.* **395** 45

Ohl R G, O'Connell R W, Bohlin R C, Collins N R, Dorman B, Fanelli M N, Neff S G, Roberts M S, Smith A M and Stecher T P 1998 *Astrophys. J. Lett.* **505** 11

Renzini A and Buzzoni A 1986 *Spectral Evolution of Galaxies* (Dordrecht: Reidel) p 232

Ben Dorman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hour Angle

The angle between an observer's meridian and a celestial body measured westwards (i.e. clockwise) from the meridian in a direction parallel to the celestial equator. Hour angle is usually expressed in terms of time units (hours, minutes, and seconds, where 1 h is equivalent to 15°) since the value of hour angle increases due to the apparent daily rotation of the celestial sphere. Thus the hour angle of a star is zero when it crosses the meridian (at upper transit); 6 h later, the celestial sphere has rotated through 90°, and the value of hour angle is 6 h. Thus, the hour angle is equal to the time which has elapsed since the last occasion on which the star crossed the meridian. After 24 h, the star returns to the meridian once more.

See also: celestial sphere, diurnal motion, meridian.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hourglass Nebula (MyCn 18)

A planetary nebula in the constellation Musca, position RA $13^{\mathrm{h}}\ 40^{\mathrm{m}}$, dec. $-67^{\circ}\ 23'$. It is only $4''$ in size, and was first imaged well by the Hubble Space Telescope. The twin lobes of ejected matter from the central star form a distinctive hourglass shape, while recently expelled gas at the center strongly resembles a human eye.

See also: Lagoon Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Houzeau de Lehaie, Jean-Charles-Hippolyte-Joseph (1820–88)

Belgian astronomer, became director of the Royal Observatory of Belgium. With Albert Lancester, Houzeau wrote *Bibliographie Generale de L'astronomie* (1887), listing all the publications in astronomy since the invention of writing.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hoyle, Sir Fred (1915–)

Astrophysicist, controversialist and science fiction writer, born in Bingley, West Yorkshire, England, became professor of astronomy at Cambridge where he founded the Institute of Theoretical Astronomy (IOTA), later merged with the Cambridge Observatory as the Institute of Astronomy. Motivated by anti-religious sentiments, and wishing to avoid the issue of cause of the start of the universe, Hoyle joined HERMANN BONDI and THOMAS GOLD in developing steady state cosmology, providing the general relativistic formulation of the theory. He became the most visible proponent of the new theory, and it was he who coined the term 'Big Bang' for the explosive event at the start of the universe, intending it to be derisive about evolutionary theories of cosmology. Since space was expanding but wanting to maintain the overall density of the universe in steady state, it was necessary to hypothesize the phenomenon of continuous creation of matter in space. Finding that this, contrary to intention, was regarded as evidence of a creator in action, Hoyle published the polemical *Nature of the Universe* about the steady state theory. The steady state theory is generally regarded as disproved (by the existence of the cosmic microwave background), although it survives in a version in which a large steady state universe in which there is a continuous sequence of randomly occurring Big Bangs. It is undisputable, however, that the development of the steady state theory led to an epoch-making discovery. It had been thought that the Big Bang created all the chemical elements (the alpha-beta-gamma theory), and if there was no Big Bang, then the elements must have been made elsewhere and at other times. This led Hoyle to work on the origin of chemical elements in stars. He collaborated through visits to the California Institute of Technology and the Mount Wilson and Palomar Observatories with MARGARET BURBIDGE, Geoffrey Burbidge and WILLIAM FOWLER, working on the synthesis of the elements, and became an author of the landmark paper on this topic, known from its authors' initials as B^2FH. Hoyle has also been, with N Wickramasinghe, a modern proponent of ARRHENIUS's panspermia hypothesis, that life on Earth comes from space, in particular in the form of viruses delivered by comets. Hoyle has published fourteen novels, a play, and numerous works for the popularization of science.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Classification

A classification scheme for galaxies, devised in its original form in 1925 by Edwin P Hubble (1889–1953), and still widely used today. The Hubble classification recognizes four principal types of galaxy—elliptical, spiral, barred spiral and irregular—and arranges these in a sequence that is called the tuning-fork diagram.

Elliptical galaxies are denoted by the letter 'E' followed by a number from 0 to 7 to indicate the degree of flattening of the observed elliptical shape. For an *En* galaxy, the number, n, is obtained from $n = 10[(a - b)/a]$ where a is the maximum diameter and b the minimum. Thus E0 galaxies appear spherical ($n = 0$ because $a = b$) while E7 galaxies are markedly flattened with a maximum diameter more than three times greater than the minimum diameter.

Spiral galaxies, denoted by the letter 'S', have a central bulge surrounded by a flattened disk within which stars, nebulae and clouds of gas and dust are concentrated into a pattern of spiral arms. Spirals are subdivided into three categories, Sa, Sb and Sc, an Sa galaxy having a large central bulge surrounded by tightly wound arms, an Sb galaxy having a somewhat smaller bulge and less tight arms, and an Sc galaxy having a relatively small bulge and loose, open arms. In barred spiral galaxies, denoted by 'SB', the arms emerge from the ends of what looks like a rigid bar, or elongated ellipsoid of stars and luminous material that straddles the central nucleus. Barred spirals are subdivided into types SBa, SBb and SBc according to similar criteria to ordinary spirals. Intermediate categories are labelled appropriately; for example, a galaxy mid-way in appearance between categories Sa and Sb would be labelled Sab. Other peculiarities are identified by additional letters such as 'p' (for 'peculiar') or 'r' (to denote the presence of a ring around the nucleus). The Milky Way Galaxy, of which the Sun is a member, used to be classified as an ordinary spiral of type Sbc, but because recent observations indicate that its central bulge is elongated, it may be more appropriate to regard it as an SBbc barred spiral.

Intermediate between the ellipticals and spirals are the lenticular (lens shaped) galaxies, denoted by 'S0' or 'SB0'. Although a lenticular possesses a central bulge surrounded by a disk, it displays no sign of a spiral structure.

Irregular galaxies, which have no obvious nucleus or ordered structure, are denoted by 'Irr' and are broadly subdivided into 'Irr I' and 'Irr II'. Irr I galaxies display evidence of ongoing star formation (for example bunches, or 'associations', of hot young stars and bright HII regions). Irr II galaxies have a disturbed appearance, their shapes seeming to have been distorted by violent internal activity or as a result of collisions or close encounters with other galaxies.

Ellipticals contain little, if any, gas and are dominated by the light of old red giant stars. Spirals and barred spirals contain significant quantities of gas and dust (about 5–15% of the galaxy's mass), Sc and SBc galaxies, on average, containing more than Sa and Sba galaxies. Some, but not all, irregulars contain higher proportions of gas and dust (up to 20–30% by mass).

See also: barred spiral galaxy, elliptical galaxies, galaxies: classification, galaxy, galaxy evolution, irregular galaxy, lenticular galaxy, Milky Way Galaxy, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble constant

The constant of proportionality, denoted by H, that relates the velocities at which galaxies are receding to their distances. According to the Hubble law, velocity of recession (V) is directly proportional to distance (D), a relationship which may be expressed as follows: $V = H_0 D$, where H_0 is the value of the Hubble constant at the present epoch in the history of the universe.

The Hubble constant, or Hubble parameter, is a measure of the rate of expansion of the universe and is, in effect, a measure of the fractional increase in the scale of the universe in unit time. Its value is usually expressed in units of velocity, expressed in kilometers per second, divided by distance, expressed in megaparsecs (i.e. km s^{-1} Mpc^{-1}), but it may also be expressed in units of s^{-1}. There is still a significant degree of uncertainty about the precise value of H_0, although there is wide general agreement that it lies between 50 km s^{-1} Mpc^{-1} and 100 km s^{-1} Mpc^{-1}. The uncertainty arises primarily because it is hard to obtain reliable measurements of the distances of galaxies. Many recent research programs using the apparent brightnesses of Cepheid variables or Type Ia supernovae as distance indicators have yielded values in the range 57–80 km s^{-1} Mpc^{-1}, with a concentration of values in the range 60–70 km s^{-1} Mpc^{-1}. A value of 65 km s^{-1} Mpc^{-1}, for example, would imply, in principle, that a galaxy at a distance of 1 megaparsec should be receding at a speed of 65 km s^{-1}, a galaxy at a distance of 1000 megaparsecs at 65 000 km s^{-1}, and so on.

The Hubble time is the time that it would take for the universe to double its size (i.e. for each galaxy to recede to twice its present distance) at a constant rate of expansion equal to the present rate. Equivalently, it is the time that the universe would have taken to expand to its present size (i.e. for each galaxy to recede to its present distance) at a constant rate of expansion. The Hubble time is equal to the reciprocal of the Hubble constant (i.e. $1/H_0$). The uncertainty in the value of the Hubble constant is reflected in a similar uncertainty in the value of the Hubble time, and hence in estimates of the age of the universe. If H_0 lies in the range 50–100 km s^{-1} Mpc^{-1}, values of the Hubble time lie between 20 and 10 billion years, respectively. If, for example, $H_0 = 65$ km s^{-1} Mpc^{-1}, then the value of the Hubble time is 15 billion years. The Hubble time is related to the age of the universe through a model of its expansion. In the standard model, the age of the universe is two-thirds the Hubble time.

See also: cosmological model, cosmology, deceleration parameter, expanding universe, Hubble law, redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Deep Fields

The Hubble Deep Fields are two small areas of the sky that were carefully selected for deep observations by the HUBBLE SPACE TELESCOPE (HST). They represent the deepest optical observations to date and reveal galaxies as faint as $V = 30$, 4 billion times fainter than can be seen with the unaided eye.

Ten-day observing campaigns were carried out by HST in December 1995 on a field in the northern hemisphere (HDF-N) and in October 1998 on a field in the southern hemisphere (HDF-S). The fields were chosen to avoid bright stars, dust or nearby galaxies interrupting the view of more distant galaxies. To make the most efficient use of HST, the fields were also required to be in the 'continuous viewing zone', at about $\pm 60^\circ$ declination. Other than that, the fields are essentially random 'core samples' of the universe. The observations were released to the public within weeks of being obtained and have been the subject of extensive research and followup observations with other telescopes. The Hubble Deep Fields represent the deepest observations not only of HST but also of some the largest ground-based telescopes and other space observatories.

Observational details

The coordinates of HDF-N are $12^{\rm h}36^{\rm m}49.4^{\rm s}$ + $62^\circ 12' 58''$ (J2000). The HDF-N has been observed at optical wavelengths by the Wide Field Planetary Camera-2 (WFPC-2) (see also ASTROMETRY WITH HUBBLE SPACE TELESCOPE), which has a field of view roughly 3 arcmin on a side, and in the infrared with the NICMOS camera. The coordinates of the HDF-S field are $22^{\rm h}32^{\rm m}56.2^{\rm s} - 60^\circ 33' 03''$. During the HDF-S campaign, observations were carried out in parallel on adjacent fields with NICMOS and STIS. The STIS field was centered on QSO J2233-606. The WFPC-2 camera observed both fields through four broadband filters, centered at 300, 450, 606, and 814 nm. The STIS observations of the southern field included UV spectroscopy of the QSO and optical and UV observations of the field around the QSO. The NICMOS observations of both the southern and northern fields were taken through broadband filters centered at 1.1 μm and 1.6 μm.

Census of the HDF-N

The WFPC-2 image of HDF-N covers an area of 5 arcmin^2. Within this area it contains approximately

3000	galaxies brighter than magnitude $V = 29$,
130	galaxies with measured spectroscopic redshifts,
20	galaxies with spectroscopic redshifts greater than $z = 2.5$,
17	sources brighter than 65 μJy at at 6.7 μm or 200 μJy at 15 μm detected by the Infrared Space Observatory,
10	stars along the line of sight through the Milky Way,
9	radio sources brighter than 9 μJy at 8.5 GHz or 40 μJy at 1.4 GHz,
4	sources brighter than 2 m Jy at 850 μm detected by the SCUBA bolometer array at the James Clerk Maxwell Telescope,
0	stars redder than $V - I = 1.8$ and
3	sources redder than $V - K = 7$.

The numbers are approximate because it is difficult to differentiate true sources from noise at the limits of the observations and because among the brighter galaxies it is sometimes difficult to decide whether to count neighboring objects separately or as part of a larger galaxy. The HDF-S has a similar number of galaxies but roughly five times as many stars owing to the different line of sight through the Milky Way.

Galactic dark matter

The lack of faint red stars in the HDF-N was used to rule out faint hydrogen-burning stars as the source of Galactic DARK MATTER. The HDF constraint is most concisely expressed as a limit on the I band absolute magnitude of the constituents of the Galactic dark halo:

$$M_I > 15.9 + \frac{5}{3}\log\left(f\frac{0.5M_\odot}{M}\right) \qquad (1)$$

where M is the mass of the objects and f is their density as a fraction of the local halo density. Faint RED DWARFS account for less than 1% of the Galactic dark halo for $M_I < 14$ and less than 6% for $M_I < 15$. The limits are even more stringent on the contribution from extremely metal-poor dwarfs.

Galaxy counts

Figure 1 shows the number of galaxies as a function of brightness in HDF-N and HDF-S, together with results from shallower ground-based surveys. The counts are compared with a simple model that includes the effect of cosmological curvature but does not include galaxy evolution. A variety of evolutionary models have been shown to match the counts, but work is still underway to determine which if any of the models match the full distribution of colors, sizes and morphologies of the HDF galaxies.

Spectroscopic and photometric redshifts

The spectroscopic survey of the HDF with the W M Keck telescope is among the deepest ever attempted. The survey is nearly complete to a magnitude $R = 23.5$ and extends more than a magnitude fainter for selected objects. Although this spectroscopy at this depth is limited to the brightest 5% of the HDF galaxies, it has served to confirm the existence of a population of galaxies at redshifts $z > 2.5$ and has validated the use of broadband photometric data as a means for estimating redshifts. These 'photometric redshifts' are determined either by fitting template spectra to the photometry of galaxies through different filters or by using the spectroscopic sample as an empirical 'training set' to determine the relation between colors and redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

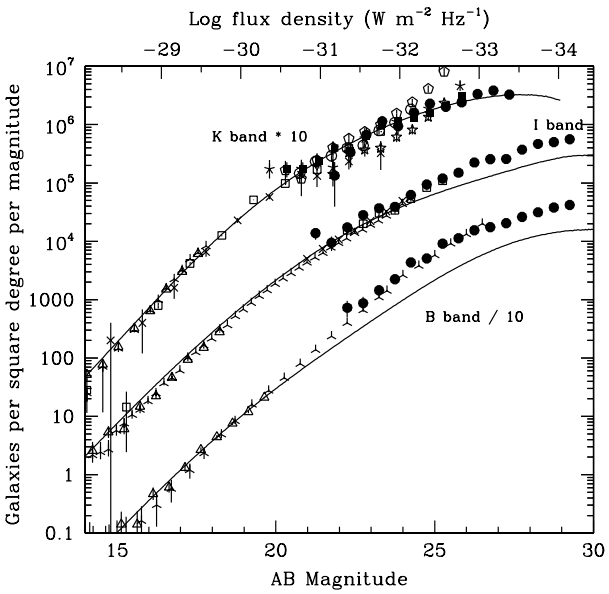


Figure 1. Counts of galaxies in the B, I and K bands (centered roughly at 0.4, 0.8 and 2.2 μm, respectively). The full symbols show the counts from the HDF observations. The HDF counts in the K band are from observations at 1.6 μm, corrected by 0.4 mag in mean color to match the brighter K band observations. The other symbols show measurements from various ground-based surveys. The lines show a no-evolution model with the following cosmological parameters: $H_0 = 65$ km s^{-1} Mpc^{-1}, $\Omega_{\rm tot} = 1$, $\Omega_{\rm matter} = 0.3$ and $\Omega_\Lambda = 0.7$. Galaxies in this model are distributed with the brightnesses and colors typical of present-day galaxies, and these are held constant with lookback time. The deviation of the counts from this model (a factor of ∼3 at faint magnitudes in the B band) indicates the extent to which galaxy evolution contributes to the modeling problem.

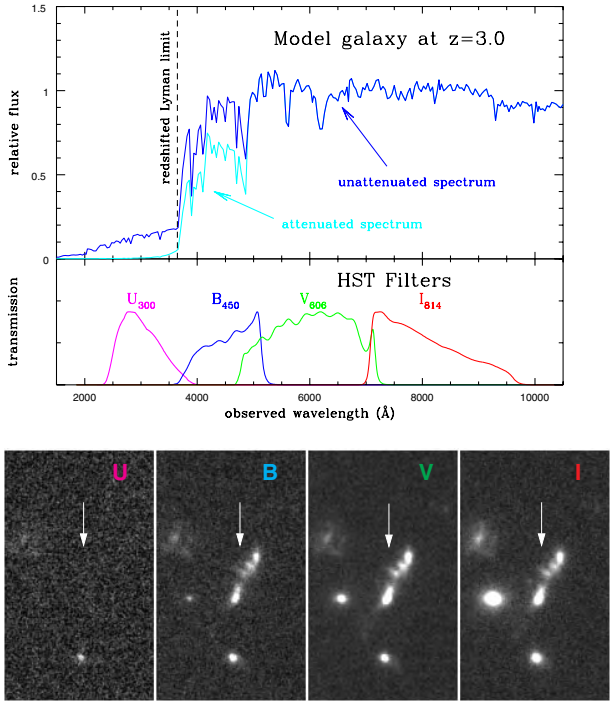


Figure 2. Illustration of the Lyman-break technique for identifying high-redshift galaxies. The top figure shows the theoretical spectrum of a star-forming galaxy at a redshift $z = 3$, with and without attenuation from intergalactic hydrogen along the line of sight. The curves below the theoretical spectra show the bandpasses of the HST filters used for the HDF observations. The signature of a high-redshift galaxy is that it 'drops out' of the ultraviolet F300W band (being either undetected or much fainter than at the next-longer wavelength). The image below shows the brightest of the Lyman-break galaxies identified in the HDF-N. **This figure is reproduced as Color Plate 11.**

Lyman-break galaxies

The most robust feature in the spectra of star-forming galaxies is the Lyman break. This opacity edge is produced when hydrogen atoms in the ground state are ionized by photons of energy greater than 13.6 eV (corresponding to wavelengths shorter than 912 Å). Photons with wavelengths slightly shorter than this are absorbed by hydrogen in stellar atmospheres, the interstellar medium of a galaxy and the intergalactic medium. In addition, as the light travels through the intergalactic medium it is partially absorbed by the Lyman-series lines of hydrogen at rest-frame wavelengths between 912 and 1216 Å. The combined effect produces a discontinuity in the spectra of high-redshift galaxies that is easy to detect. Figure 2 shows an example.

The four filters of the HDF offered the first opportunity with HST to search for Lyman-break galaxies. While the exact number of galaxies found depends on the color selection criteria, roughly 100 candidate objects were found with redshifts likely to be in the range $2.5 < z < 3.5$ and roughly 15 with redshifts likely to be in the range $3.5 < z < 4.5$. The brightest of these objects were confirmed spectroscopically. In addition, several objects were noted with colors consistent with redshifts higher than $z = 4.5$. Spectroscopy at the Keck telescope confirmed one of these candidates to be at redshift $z = 5.34$.

The Lyman-break galaxies are generally compact, with half-light radii 1 kpc $< r <$ 7 kpc. Those that are well resolved generally appear morphologically peculiar. There are several prominent examples in the HDF of Lyman-break galaxies made up of multiple components spaced irregularly within a region of less than 2 arcsecs, suggesting that we are witnessing the merging of several smaller components into a larger galaxy.

See HIGH-REDSHIFT GALAXIES for more detail.

Cosmic star-formation history

The photometric redshift estimates (including in particular those from the Lyman break) can be used to estimate the ultraviolet luminosity density of the universe as a function of redshift. This is a particularly useful quantity as it is closely tied to the history of heavy-element formation: the massive stars that emit most of the UV radiation detected in the HDF (redshifted to optical wavelengths) also produce most of the elements of atomic number $Z > 6$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In the approximation of instantaneous recycling, the metal ejection rate can be written as

$$\dot{\rho}_z = \psi \int m p_{zm} \phi(m)\, \mathrm{d}m \qquad (2)$$

where ψ is the star formation rate per unit co-moving volume, $\phi(m)$ is the stellar initial mass function (IMF) and p_{zm} is the stellar yield, i.e., the fraction of a star of mass m that is converted to metals and ejected. At short wavelengths, the luminosity density radiated per unit frequency during the main-sequence phase of stellar evolution is related to the star-formation rate ψ by

$$\rho_n u = 0.007 c^2 \psi \int m f_{\mathrm{He}}(m)\, f_\nu(m) \phi(m)\, \mathrm{d}m \qquad (3)$$

where $f_{\mathrm{He}}(m)$ is the mass fraction of hydrogen burned into helium and $f_\nu(m)$ is the normalized spectrum of stars of mass m on the main sequence. Evolutionary synthesis models give the following approximate relation between luminosity density at a rest-frame wavelength of 1500 Å and star-formation rate ψ in solar masses per year:

$$\rho_{1500} = 8.0 \times 10^{27} \psi \ \mathrm{erg\ s^{-1}\ Hz^{-1}\ Mpc^{-3}}. \qquad (4)$$

Figure 3 shows the star-formation or metal-formation rate as a function of redshift as derived from the HDF (north and south) and various ground-based surveys. Two major uncertainties affect these estimates: the effects of dust and the amount of light coming from galaxies too faint to be detected individually. Dust in star-forming galaxies absorbs ultraviolet light and reradiates it at longer wavelengths. The attempt to quantify the amount of dust-enshrouded starlight motivated studies of the HDF with the Infrared Space Observatory and with the SCUBA submillimeter bolometer array on the James Clerk Maxwell Telescope. These observations, together with measurements of the spectral-energy distributions and optical emission-line strengths for Lyman-break galaxies, suggest that the star-formation rates derived purely from the UV luminosities should be corrected upwards by a factor of 3–10.

Background measurements

To derive the luminosity density, the counts of galaxies in a particular redshift interval are extrapolated to magnitudes below the detection limit of the images. This extrapolation can be tested by measuring the background light between the detected galaxies. Measuring the extragalactic component of this background light is difficult because it is only small percentage of the zodiacal background. Estimates of the fraction of the total extragalactic background accounted for by the detected galaxies range from 30% to 90%. An analysis of the autocorrelation of the residual fluctuations in the HDF image suggests that the total extragalactic background at $\lambda = 8100$ Å lies in the range 7.8×10^{-6} erg s^{-1} cm^{-2} sr^{-1} $\leq \nu I_\nu < 1.2 \times 10^{-5}$ erg s^{-1} cm^{-2} sr^{-1}, while the analysis of the total flux left after subtracting galaxies and zodiacal light gives somewhat higher values. See also EXTRAGALACTIC BACKGROUND LIGHT.

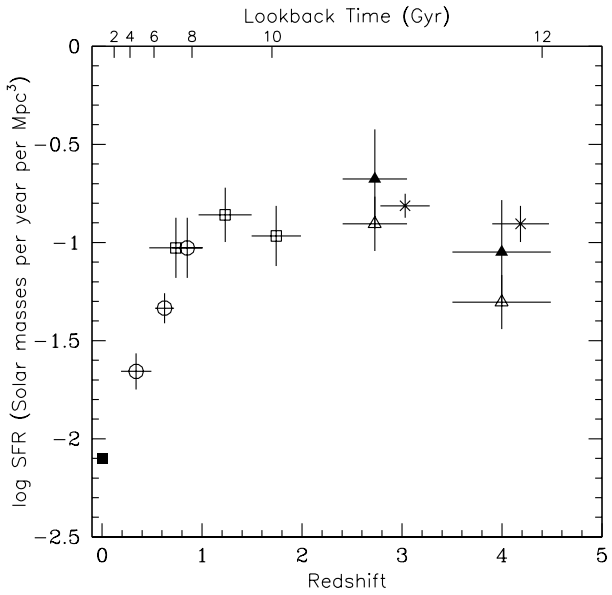


Figure 3. The star-formation rate per unit volume versus redshift (bottom axis) or lookback time (top axis). The assumed cosmology has $H_0 = 50$ km s^{-1} Mpc^{-1} and $\Omega_M = 1$. The points are derived from the total integrated UV fluxes of galaxies measured in various surveys, including the HDF-N (open triangles) and the HDF-S (solid triangles). A simple correction for dust extinction has been applied to the data, but there are large uncertainties in this correction.

Galaxy clustering

The combination of deep photometry and photometric redshifts makes the HDF images a useful tool for studying the evolution of clustering in the universe (see GALAXY CLUSTERS). However, these studies are limited by the small field size, which subtends no more than about 2 Mpc at any redshift. Nevertheless, the analyses of galaxies in different redshift bins suggests that the amplitude A_w of the angular correlation function decreases with redshift out to $z = 1$ and then rises again to higher redshift. A plausible explanation is that the galaxies at higher redshift trace higher peaks in the underlying density field and are hence biased relative to the underlying mass distribution. Another possible contribution to the stronger clustering at high redshift is the tendency for high-redshift galaxies to have multiple components, suggesting that in many cases the structures are part of the same underlying galaxy.

Bibliography

The HDF observations are described by

Williams R E *et al* 1996 *Astron. J.* **112** 1335

The use of Lyman-break galaxies to estimate the star-formation rate as a function of redshift was first presented by

Madau *et al* 1996 *Mon. Not. R. Astron. Soc.* **283** 1388

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A useful summary of the various areas of research using the HDF images can be found in

Livio M, Fall S M and Madau P (ed) 1998 *The Hubble Deep Field* (Cambridge: Cambridge Univeristy Press)

Henry C Ferguson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Diagram

Initially introduced as a way to demonstrate the expansion of the universe, and subsequently to determine the expansion rate (the HUBBLE CONSTANT H_0), the Hubble diagram is one of the classical cosmological tests. It is a plot of apparent fluxes (usually expressed as magnitudes) of some types of objects at cosmological distances, against their REDSHIFTS. It is used as a tool to measure the global geometry of the universe, and as a probe of galactic evolution.

As in most other cosmological tests, the observables used in this diagram are really proxies for the expansion factor of the universe (measured as the cosmological redshift) and some independently determined proper distance to the objects used in the test. The observed trend of distance-related quantities (e.g. apparent fluxes, angular diameters, etc) against the redshift can thus, at least in principle, be mapped to the expansion history of the universe, which in turn uniquely determines the COSMOLOGICAL MODEL. The *slope* of the trend at low redshifts determines the present normalized expansion rate, i.e. the Hubble constant. The *shape* of the trend at large redshifts determines the global geometry of the universe.

In the case of the Hubble diagram, a relativistic version of the inverse square law is used to determine the relative distances to the light sources used. Simply put, the more distant objects of a fixed intrinsic brightness should be fainter, by a factor determined exactly by the geometry of space. Redshifts are measured spectroscopically, and do not depend on the complicating factors which may affect the distance measurements, such as the evolution of sources used, sample selection effects, etc. The distances are much more difficult to determine, and this is where most of the practical problems lie.

The objects used as the test particles in a Hubble diagram are often called 'standard candles', thus expressing the hope of their constancy in time and at different locations. Yet this is hardly ever the case, and much of the effort involved goes into their standardization—removal of the various second-order parameters which cause the scatter around the ostensible universal mean luminosity of a given type of object.

Hubble (1929) used this diagram to demonstrate the expansion of the universe: fainter and thus on average more distant galaxies were receding from us with larger apparent speeds. Once the universal expansion was generally established, the test acquired other cosmological uses. There are two regimes in which the Hubble diagram is used today, each with its own goals, methods and problems.

In the low-redshift regime, the Hubble diagram is used in determinations of the Hubble constant. The local geometry is very close to being Euclidean, and thus the effects of the global curvature can be safely ignored. The look-back times to the objects used are small enough so that evolutionary effects can also be neglected. The local universal expansion is practically linear, producing the well-known HUBBLE'S LAW: $cz = H_0 D$, where c is the speed of light in a vacuum, z is the observed redshift (assuming that the peculiar velocity of the object is negligible), H_0 is the Hubble constant and D is the proper distance. Since distances are not directly measurable, even for the nearest galaxies, a proxy quantity is used; in the case of the Hubble diagram, this is often the magnitude of an object of given class, e.g. a Cepheid variable (or, more accurately, the intercept of the CEPHEID PERIOD–LUMINOSITY RELATION), a SUPERNOVA at its peak brightness, etc. If one knows the actual absolute luminosities of the objects used in the test, obtained from some independent calibration procedure, then the slope of the observed relation gives directly the local normalized expansion rate, the Hubble constant H_0. The difficulties in establishing the absolute brightness scale of the objects used for such purposes, and other problems involved in the measurements of H_0, are beyond the scope of this article.

In the more distant universe, the global geometry comes into play, causing the trend to deviate from linearity. The *shape* of the relation becomes a probe of the cosmological parameters; for the standard Friedmann–Lemaitre models, these are the density parameter Ω_0 and the cosmological constant Λ_0. (We note in passing that for many years the value of the COSMOLOGICAL CONSTANT was declared to be zero on purely ideological grounds, and only a single parameter, Ω_0, was used to characterize the global geometry, or equivalently, the deceleration parameter q_0.) Since the shape of the trend does not depend on the absolute value of the luminosities, the normalization and the Hubble constant factor out. The curvature of the trend reflects directly the global kinematics of the universe as observed over cosmological time-scales for true 'standard candles'.

Consider the effect of an increase in the mean density: it causes the expansion to slow down more, and thus the objects at a given redshift are closer and appear brighter. The same effect is accomplished by the negative values of the cosmological constant (which correspond to a positive energy density), and the opposite effect by the positive values of this parameter, which then acts as a repulsive force, accelerating the expansion. These effects become detectable only over very large look-back times, i.e. at relatively high redshifts, requiring the use of objects bright enough to be detected at such large cosmological distances. Moreover, there is a considerable parameter coupling between Ω_0 and Λ_0, as they can have qualitatively similar effects. A large baseline in redshift is also necessary in order to separate their contributions, since they may dominate the changes in the expansion rate at different epochs.

For a source with an intrinsic luminosity L, at a cosmological redshift z, the observed flux is:

$$F = \frac{L}{4\pi D^2 (1+z)^2} = \frac{L}{4\pi D_L^2}$$

where D is the proper distance and z is the source redshift. The quantity $D_L \equiv D(1+z)$ is called the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

luminosity distance. The formula differs from the classical inverse square law by the two powers of the expansion factor, $(1 + z)$: one is due to the energy loss of photons as their wavelengths are redshifted, and one is due to the relativistic time dilation in the photon emission rate, since the source is moving away from us. In the simple Friedmann cosmological models, with a vanishing cosmological constant, $\Lambda_0 = 0$, the proper distance is given by the following formulae:

$$D = \frac{c}{H_0}\frac{z(1+z/2)}{1+z} \qquad \text{for } \Omega_0 = 0$$

$$D = \frac{2c}{H_0}(1 - 1/\sqrt{1+z}) \qquad \text{for } \Omega_0 = 1$$

$$D = \frac{2c}{H_0}\frac{2 - \Omega_0 + z\Omega_0 - (2-\Omega_0)\sqrt{1+z\Omega_0}}{\Omega_0^2(1+z)} \qquad \text{otherwise}$$

where c is the speed of light in a vacuum, H_0 is the Hubble constant and Ω_0 is the density parameter of the universe, the ratio of the mean density to the critical density at the present epoch.

HUBBLE's pioneering vision of determining the global geometry through this and other tests was frustrated by the technology available to him at the time, by not being able to reach out to sufficiently large redshifts where the cosmological effects may be detectable. His quest was then carried out over many years by A SANDAGE and collaborators, as well as many others; for a fascinating historical account, see, e.g., Sandage (1995). Typically, magnitudes of the brightest cluster ELLIPTICAL GALAXIES were used as the standard candles, and the quest was on for finding the ever more distant ones. At low redshifts, suitably defined magnitudes of these objects show an intrinsic scatter of only about 30%, which was considered constant enough. Unfortunately, no conclusive cosmological results were achieved in these studies despite much effort and observing time spent, in part because they did not reach out to sufficiently large redshifts.

QUASARS, discovered in 1963, provided luminous probes of even larger redshifts, but had too much of a spread in intrinsic luminosities and too much variability to be viable for this purpose. Other candidate standard candles tried at cosmological distances include supernovae, the characteristic luminosity (L_*) of the galactic luminosity function in distant clusters, etc.

The observed apparent magnitude of a source is given by

$$m_{\rm obs} = M + 5 - 5\log D_L + K + \Delta m_{\rm sc}$$

where M is the absolute magnitude of the source, which may be affected by evolution effects, and D_L is the luminosity distance defined above. K is the so-called K correction, which accounts for the difference in the detector bandpass in the observer's frame and in the source restframe. It is given by the ratio of the fluxes of the source, with its spectrum integrated over the observed bandpass and its redshifted counterpart, and expressed in the magnitude form. This of course requires the knowledge of the source's spectrum at $z = 0$; any possible evolutionary effects are *not* included. $\Delta m_{\rm sc}$ is an empirical correction used to standardize the candle, i.e., remove statistically some other observable effects which may be correlated with M. In the case of the brightest cluster ellipticals, such corrections include subtle dependences on the cluster richness and morphology; in the case of supernovae, shapes of their light curves are used to derive such corrections.

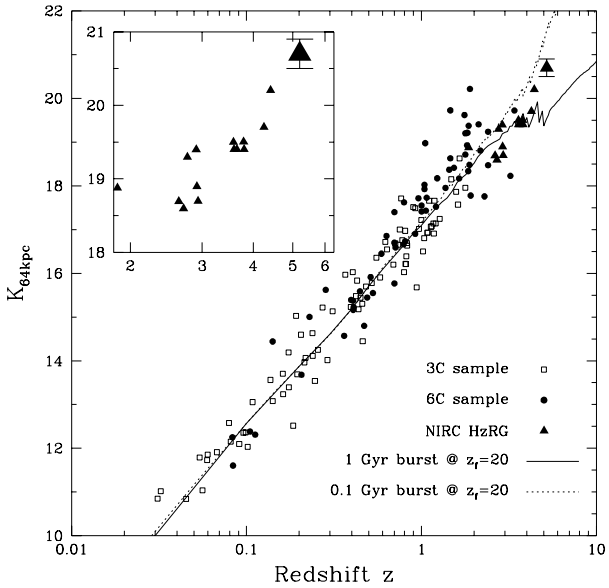


Figure 1. The infrared Hubble diagram (in the K band, 2.2 μm effective observed wavelength) for a set of powerful radio galaxies, reaching out to $z = 5.19$ (indicated with the large triangle). The insert is a zoom-in on the high-z portion of the diagram. Two galactic evolution models are shown, both corresponding to a single burst of star formation starting at $z = 20$, and assuming a cosmology with $H_0 = 65$ km s^{-1} Mpc^{-1}. $\Omega_0 = 0.3$ and $\Lambda_0 = 0$. The diagram shows a remarkably small scatter at any redshift, even though every relevant effect we know of would increase the scatter: the variations in star formation histories and formation redshifts, in the contributions from active nuclei, in extinction, line emission, merging rates, etc. While no one is yet brave enough to try to use this diagram to measure the cosmological parameters, the small observed scatter presents a challenge for theoretical models of powerful radio galaxies and their evolution. (Reprinted with permission from van Breugel W *et al* 1999 *Astrophys. J. Lett.* **518** L61 © The American Astronomical Society.)

As in the case of most other cosmological tests, the use of the Hubble diagram as a probe of the global geometry was severely undermined by the inevitable cosmic evolution of whatever objects are used as the test particles. The light from galaxies comes from their STELLAR POPULATIONS, which must evolve in time, in brightness and in color. By the late 1970s it became clear that the effects of galactic evolution will dominate over the cosmological effects for any type of galaxy at sufficiently large redshifts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

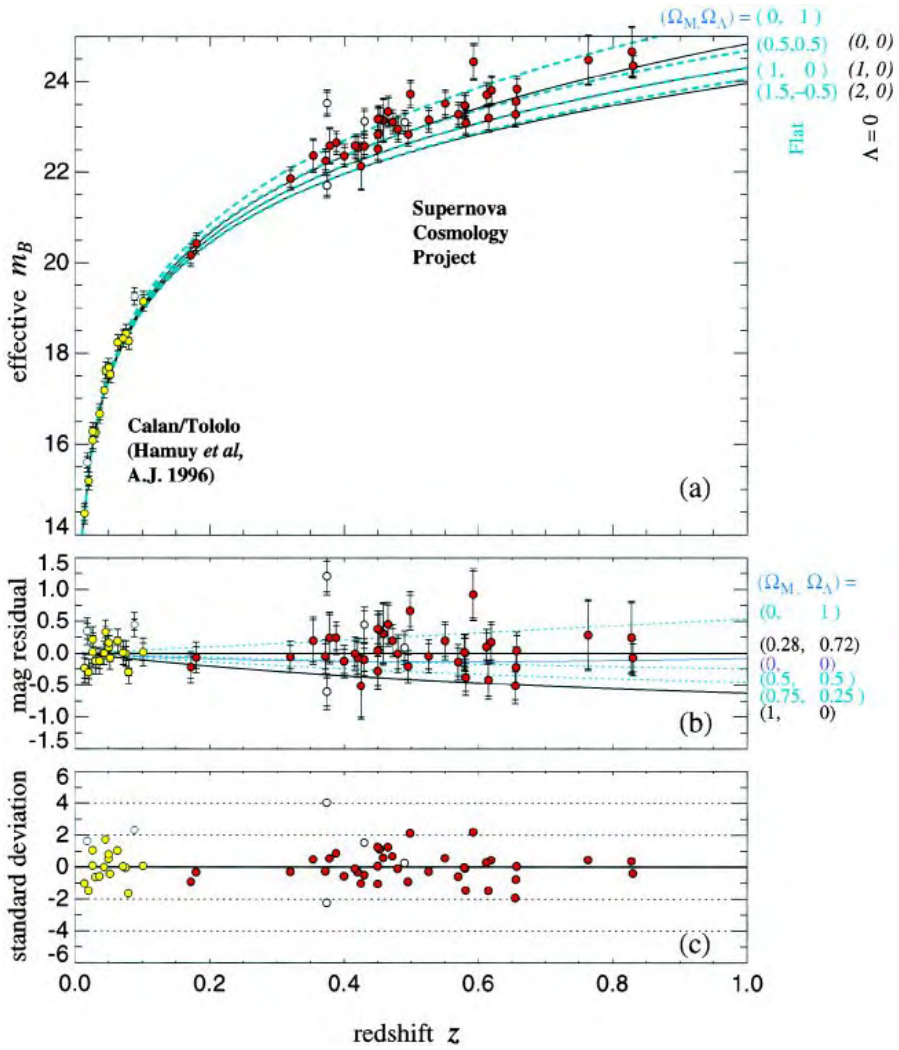


Figure 2. An example of a Hubble diagram for type Ia supernovae, with peak magnitudes corrected using the light curve shapes. Several cosmological models are drawn through the points, with the corresponding values of the parameters Ω_0 and Λ_0 as indicated. The best fit to these data indicates a low-density universe with a positive cosmological constant. The top panel shows the diagram itself. The bottom panel shows the residuals from the best-fit model, and the bottom panel shows the photometric deviations from the model in units of the standard deviations. (Reprinted with permission from Perlmutter S *et al* 1999 *Astrophys. J.* **517** 565 © The American Astronomical Society.)

This is especially true at the observed visual wavelengths, which correspond to the ultraviolet light restframe, which may be dominated by a few luminous, young and short-lived stars. Moreover, galactic merging and interactions, which must have been more frequent at higher redshifts, further complicate the interpretation of observations: galaxies may grow in time, and undergo merger-induced bursts of star formation (see GALAXIES: INTERACTIONS AND MERGERS).

The Hubble diagrams thus became a probe of galactic evolution, where a set of cosmological parameters was often assumed out of necessity in order to measure the rate of galactic evolution. More recently, the Hubble diagram has made a comeback as a potential cosmological tool, as described below.

Even if the evolutionary effects can be controlled, an additional difficulty is the possible (likely?) presence of sample selection effects. Typically, brighter representatives of any class of objects would be found at larger redshifts, as the fainter ones may not make it into a flux-limited sample. Such biases are hard to correct, since one never knows how many objects are not even detected, let alone what is their brightness distribution. Reducing the scatter of observed luminosities at a given redshift (i.e. using a more standard candle) helps, but some bias is always present. Understanding the sample completeness is essential.

In addition, it is sometimes hard to guarantee even

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the homogeneity of the samples, i.e. that the same type of objects are being selected both at high redshifts, and as their ostensible low-redshift progeny. Additional problems include possible effects of GRAVITATIONAL LENSING by the intervening masses, and interstellar extinction corrections, both in our Galaxy and in the distant galaxies used in the test.

Some hope came with technological advances in near-infrared measurements (see INFRARED TELESCOPES). The advantages of infrared observations are considerable. They probe the restframe wavelengths of normal galaxies which are much less affected by a possible flicker of massive star formation. The infrared light is typically dominated by the relatively stable and slowly changing stellar subpopulations such as the red giants. Thus, the effects of galactic evolution effects should be minimized. Infrared observations are also much less sensitive to interstellar extinction.

One example is the infrared Hubble diagram for powerful RADIO GALAXIES at high redshifts (figure 1). The advantages of radio galaxies are that they can be identified at very large redshifts (as of this writing, at $z > 5$), and that they are generally believed to be progenitors of at least some giant ellipticals today. Their chief disadvantage is the presence of a powerful active nucleus, which may be contaminating the observed brightness and affecting the host galaxy in uncertain ways. Perhaps the infrared observations of normal, quiescent cluster ellipticals (known out to $z \approx 1.3$, as of mid-1999), or magnitude intercepts of their fundamental plane correlations may be more promising.

There has been much excitement in recent years about the use of distant supernovae as possible 'standard candles' to measure the global geometry of the universe. At their peak, supernovae of type Ia reach an absolute magnitude $M_B \approx -19.5$, and are thus sufficiently luminous to be detected at redshifts approaching unity, where the effects of cosmic geometry should be readily detectable. Furthermore, the shape of the supernova light curve is found to correlate with the peak luminosity, thus providing a handy observational way to lower the scatter of the peak magnitudes by correcting for this 'second parameter'. Scatter as low as 15% or even less may be achieved (this made type Ia supernovae also a viable tool in the efforts to measure H_0 at low redshifts). Since the physics of the supernova explosions should be universal, this novel standard candle clearly has a lot of promise as a cosmological tool. Note, however, that some evolution in the observable properties of supernovae is possible, e.g. due to the slowly changing metallicities of the stellar populations from which these supernovae originate.

Using this approach, two groups have recently measured the values of the cosmological parameters Ω_0 and Λ_0 (figure 2), with mutually consistent results (Schmidt *et al* 1998, Perlmutter *et al* 1999). Both groups found that the best fit to their observations is consistent with a zero-curvature universe, preferred by the theoretical prejudice of inflation models, but with a positive cosmological constant, which was something of a surprise. If this result is confirmed by future observations and tests, it would be of profound cosmological significance.

It remains to be seen if any other types of luminous objects turn out to be good standard candles, or if new ways of standardizing the old ones are invented. In any case, the Hubble diagrams will remain a useful probe of galactic evolution, and possibly even of the global geometry of the universe.

Bibliography

The first-hand historical account of the initial uses of the Hubble Diagrams can be found in

Hubble E 1929 *Proc. Natl Acad. Sci., USA* **15** 169

This work was summarized in a much more accessible and easier to find book

Hubble E 1936 *The Realm of the Nebulae* (New Haven, CT: Yale University Press) (reprinted in 1982)

An excellent scientific and historical outline of the use of the Hubble Diagrams from the early days until about early 1990s is given in

Sandage A 1995 *The Deep Universe, (Saas-Fee Advanced Course 23)* (Berlin: Springer)

Two modern papers describing the use of the Hubble diagrams of type Ia supernovae, along with the more modern references are

Perlmutter S *et al* 1999 *Astrophys. J.* **517** 565
Schmidt B *et al* 1998 *Astrophys. J.* **507** 46

S George Djorgovski

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Law

The relationship that states that the recessional velocities of distant galaxies are directly proportional to their distances. It is named after the American astronomer Edwin P Hubble (1889–1953) who, in 1929, published his discovery that the redshifts in the spectra of galaxies are proportional to their distances. This was one of the key discoveries in cosmology for it showed that the galaxies are receding and the universe as a whole is expanding.

The relationship between velocity of recession, V, and distance D, may be written as $V = H_0 D$, where H_0 is the present-day value of a constant of proportionality known as the Hubble constant, or Hubble parameter. The law is consistent with a universe undergoing uniform expansion whereby each galaxy (or cluster of galaxies) is receding from every other one with a speed proportional to the separation between them. In such a universe each observer, no matter on which galaxy he is located, sees all other galaxies receding with speeds proportional to their distances, and no one galaxy is the unique center of this expansion. The smooth, idealized, recession of the galaxies from each other (ignoring local variations caused, for example, by the gravitational influence of clusters and superclusters of galaxies) in accordance with the Hubble law is called the 'Hubble flow'.

See also: Doppler effect, Hubble constant, Hubble diagram, redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Space Telescope

Since its deployment from the Space Shuttle Discovery on 25 April 1990, the NASA–ESA Hubble Space Telescope (HST) has revolutionized observational and theoretical astronomy with a stream of remarkable observations and beautiful images which have stunned astronomers and public alike. Yet, during its first weeks of operation, these dramatic successes seemed a world away.

A number of serious issues came to the fore almost immediately. Warping and deformation of the observatory's huge solar arrays occurred each time that the HST crossed the day–night terminator. The result was a shaking which could not be counteracted by the onboard attitude control system. On occasions, the vibration was so severe that the satellite lost its lock on the guide stars which were expected to provide pointing to an accuracy of a fraction of an arcsecond.

Even worse was the revelation that the initial images captured by different instruments revealed a mysterious halo of light around the stars. It soon became clear that the 2.4 m diameter primary mirror was badly flawed. Investigations revealed that the SPHERICAL ABERRATION was caused by a mirror which had been ground into the wrong shape—it was 2 μm too flat near its edge. The result was that, while the spatial resolution of around 0.1 arcsec was close to specification, the telescope's sensitivity was seriously compromised, since it was only focusing 15% of the incoming light. The flaw had not been noticed prior to launch because the primary and secondary mirrors had never been tested together.

Instead of inaugurating a new golden era of space exploration and discovery, the much-heralded arrival of the world's first large orbital observatory rapidly turned into a public relations fiasco as the myopic telescope and NASA management in general came under heavy criticism.

The immediate answer was to remove the 'fogging' by means of computer processing. However, in the long term, the only solution was to compensate for the faulty optics by installation of a corrective optics space telescope axial replacement (COSTAR) system. This comprised ten coin-sized, adjustable mirrors which were to be inserted into the optical train of three of HST's instruments in order to improve their focusing capability.

The COSTAR was eventually installed during the first servicing mission (SM-1) in December 1993. With the HST captured by the remote manipulator arm and located upright in the Shuttle's payload bay, the astronauts spent $35\frac{1}{2}$ hours of extravehicular activity fitting the COSTAR, replacing the warped solar arrays and installing a modified Wide-Field Planetary Camera.

On 13 January 1994, NASA triumphantly released the first images from the revamped observatory, including a view which showed the central regions of spiral galaxy M100 in unprecedented detail. Senator Barbara Mikulski proudly announced, 'I am delighted to be able to announce today, after the launch in 1990 and the early disappointments, the trouble with Hubble is over!'

The telescope

The EDWIN P HUBBLE Space Telescope is named in honour of the American astronomer who made two major breakthroughs in cosmology during the 1920s. By measuring the distance to the Andromeda galaxy, he became the first person to establish that spiral and elliptical galaxies existed far beyond our own Galaxy. Then, in 1929, he analysed the spectra of galaxies and found a linear relationship between their distances and their REDSHIFTS which suggested that they were all moving away from us, with the furthest galaxies receding at the fastest rate.

The HST was intended to extend astronomical knowledge of the universe by observing from an altitude of about 600 km (375 miles) above the turbulent atmosphere. The idea for such a space telescope was first proposed in 1946 by American astronomer Lyman Spitzer, and was later revived in the 1970s as the Large Space Telescope project. The fate of the project was closely tied to the development of NASA's reusable Space Shuttle, the only launch vehicle which could place the giant telescope into low-Earth orbit and deliver astronauts for regular servicing missions.

Although it began as a NASA project, the EUROPEAN SPACE AGENCY was also keen to participate. This resulted in a memorandum of understanding in 1977 which provided ESA with approximately 15% of the available observing time in return for providing the Faint Object Camera and a pair of solar arrays. This, in turn, resulted in positions for ESA scientists being established at the SPACE TELESCOPE SCIENCE INSTITUTE in Baltimore and the creation of the SPACE TELESCOPE EUROPEAN COORDINATING FACILITY in Garching, Germany.

At the time of its launch, the HST was the largest optical telescope ever placed into orbit. The observatory weighs approximately 11.5 tonnes, and its main body is 13.3 m (43.5 ft) long and 4.3 m (14 ft) in diameter. Two 12 m (40 ft) long solar panels attached on either side of the observatory provide a maximum power output of 4.4 kW. A large aperture door is closed to protect the telescope during servicing missions.

The HST has a Ritchey–Chrétien design, which uses a 2.4 m (94 in) primary mirror made of ultra-low-expansion glass coated with a thin film of aluminum. The mirror collects radiation across the range of wavelengths from ultraviolet to near-infrared. Light is reflected from the primary mirror to a 33 cm (13 in) secondary located 4.6 m (15 ft) in front of it. This reflects the light back through a 66 cm (26 in) hole in the primary mirror to the scientific instruments behind it in the aft shroud section.

The observatory was one of the first satellites to be designed with in-orbit maintenance in mind. In order to aid astronauts during extravenicular activity, the HST has 68 m (225 ft) of handrails and 31 foot restraints on its surface. It also contains 70 orbital replacement units—parts which are designed to be replaced in space.

Apart from providing the option to upgrade the scientific instruments and hardware such as the computer, data recorder or attitude control systems, this also allowed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for replacement of the two giant solar arrays which charge the HST's batteries when they are in sunlight. This provided a scientific bonus when one of the first set of arrays was brought back to Earth for analysis after $3\frac{1}{2}$ years in orbit. Subsequent studies of thousands of impact craters on the surface of the array revealed that those with diameters in the range 100–1000 μm were predominantly micrometeoroid in origin, while craters below 100 μm in size were mainly created by space debris.

The HST was originally equipped with five instruments: two cameras, two spectrographs and a photometer.

- The Wide Field and Planetary Camera, developed by the California Institute of Technology and the Jet Propulsion Laboratory.
- The European Space Agency's Faint Object Camera.
- The Goddard High-Resolution Spectrograph for ultraviolet observations, developed at NASA's Goddard Space Flight Center.
- The Faint Object Spectrograph, developed at the University of California, San Diego.
- The High Speed Photometer, developed at the University of Wisconsin.

Over the years, this complement of instruments has been modified due to the initial flawed optics, changing observational requirements and development of improved technology. The High Speed Photometer was replaced during the December 1993 servicing mission in order to make room for the COSTAR. At the same time, the Wide Field and Planetary Camera was also replaced by a version with modified optics (WFPC-2) to offset the HST's spherical abberation. In its wide field mode, this instrument can image an area up to 2.6 arc minutes across, while in its planetary mode, although the field of view is reduced to 1.1 arc minutes, the spatial resolution is improved to 0.04 arcsec.

In February 1997, during a second servicing mission, the Goddard High-Resolution Spectrograph and the Faint Object Spectrograph were replaced by the Space Telescope Imaging Spectrograph, which is designed to take spectra over a wider wavelength range (1050–11 000 Å) than the previous instruments combined. In addition, the Near-Infrared Camera and Multi-Object Spectrometer (NICMOS), gave the observatory the capability of observing in the near-infrared (0.8–2.5 μm) spectral region.

Unfortunately, a higher than expected rate of evaporation by the solid nitrogen coolant in NICMOS resulted in a reduced lifetime for the instrument. Its cameras were switched off after the instrument reached cryogen exhaustion on 3 January 1999. However, the science capabilities of NICMOS are expected to be fully restored after a closed-loop cooling system is installed during the fourth servicing mission (3B) in the summer of 2001.

The next generation of HST instruments includes the Advanced Camera for Surveys (ACS), which will replace the Faint Object Camera. The ACS, which is due to be installed during servicing mission 3B, will operate in three different modes. Although it can be used as a wide field camera, in its high resolution mode it will have ten times the spatial resolution of the WFPC: −2. A third mode, the Solar Blind Camera, will be able to take high-resolution ultraviolet images.

Since all of the scientific instruments will then incorporate their own internal correcting optical system, the redundant COSTAR system will be removed during a fifth servicing mission, around 2004. At the same time, the HST will receive two more instruments—the Wide Field Camera 3, which reuses parts of the original WFPC, and the Cosmic Origins Spectrograph, which will greatly expand the ultraviolet spectroscopic capability of the HST, especially in the far-UV.

Control and communication

Since the telescope is in such demand, and maneuvring in space from one target to another is both difficult and time-consuming, detailed long-term planning is essential. Computer programmers are employed to write the second-by-second instructions that point the telescope, keep it on target and command the scientific instruments.

Programmed pointing instructions are made more complicated by the fact that the observatory follows a circular orbit inclined at 28.5 degrees to the equator. Although its orbital period is approximately 95 minutes, the observatory's motion around the planet means that a particular target above the equator will be occulted by the Earth for up to 45 minutes. The onboard computer overcomes this handicap by instructing the HST to reacquire its target as the OCCULTATION ends. In this way, multiple exposures lasting many hours can be obtained. This problem does not arise for objects above the polar regions, since they are never hidden behind the planet.

The HST uses a number of systems for attitude control and accurate pointing. It has three rate sensor units, each containing two gyroscopes. Frequent failures of these gyros have caused a few headaches for mission controllers, particularly in November 1999, when the observatory went into an automatic safe mode and had to be shut down for two months after four of its gyros stopped functioning. The HST needs at least three operational gyros for accurate alignment of the instruments on its astronomical targets.

The HST also carries four reaction wheels, which move the telescope from one target to another. These only allow the telescope to slew very slowly across the sky—about 90 degrees every 18 minutes. Three fixed head star trackers refine the telescope's position by commanding the reaction wheels to make minor adjustments. At all times, the HST must be orientated away from the Sun so that its sensitive optics are not damaged.

Once in position, the observatory's set of fine guidance sensors are used to pinpoint the particular target by fixing on pairs of guide stars in the telescope's field of view. After locking on to the guide stars, these sensors are able to aim the telescope with an accuracy of 0.01 arcsec

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for up to 24 hours. The Fine Guidance Sensors also have a secondary scientific role since they can be used for astrometry, measuring star positions to within 0.005 arcsec (see ASTROMETRY WITH HUBBLE SPACE TELESCOPE). The Guide Star Catalog used by the HST to lock onto celestial targets includes the precise locations of 15 million stars down to magnitude 14.5.

Instructions from the control centre at Goddard Space Flight Center are relayed to the HST via one of NASA's Tracking and Data Relay Satellites (TDRS). The HST receives the data by means of two high-gain antennae, which usually receive programmed instructions from the Earth three times a day.

Until December 1999, scientific data were recorded on a mechanical reel-to-reel data recorder until it could be downloaded to the ground, but this was replaced by a solid state digital version during servicing mission 3A. The observational data are eventually transmitted back to a ground station at White Sands, New Mexico, via a TDRS. From there, the data are sent by satellite relay to Goddard Space Flight Center. After checking that the data have not been lost or corrupted, controllers at Goddard pass them on to the Space Telescope Science Institute in Baltimore where they are stored and processed for scientific analysis.

Responsibility for conducting and coordinating the science operations of the HST rests with the Space Telescope Science Institute, which is operated for NASA by the ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY, Inc. (AURA).

A revolution in astronomy

Apart from a brief hiatus from November 1999 to January 2000, when the telescope went into safe mode after the failure of a fourth gyro, the HST has been sending back a stream of data almost continuously for over a decade.

During this period, the observatory has led to an unprecedented series of discoveries that have invigorated all areas of astronomical research. The HST has helped confirm some astronomical theories, challenged others, and, more often than not, come up with complete surprises which leave theorists scrambling for answers.

Apart from the Sun and Earth, only Mercury is inaccessible to the HST's gaze—the planet is always so close to the Sun in the sky that it would be far too risky to point the telescope's sensitive optics in its direction. Many thousands of celestial objects have been studied with the orbiting observatory since its inauguration.

In the solar system, the HST has observed seven of the nine planets and many of their larger satellites, searched for Kuiper Belt objects, studied comets such as Hyakutake and Hale–Bopp, and imaged numerous asteroids. Major advances include the first maps of Pluto, its moon Charon and the large asteroid Vesta, and detailed views of the fragmented Comet Shoemaker–Levy 9 as it collided with Jupiter. Regular monitoring of atmospheric phenomena has revealed cyclonic systems and dust storms on Mars, and chaotic changes on the outer gas giants, including ultraviolet aurorae on Jupiter and Saturn, equatorial storms on Saturn and a transient Great Dark Spot on Neptune.

Within the Milky Way, the HST has unveiled a wondrous menagerie of intricate planetary nebulae; found evidence for brown dwarfs and solitary black holes; studied the changing environment around supernova 1987A as its encircling rings of glowing gas expand outwards and interact; discovered circumstellar disks in which planets may be forming; probed the inner reaches of star forming regions such as the Orion Nebula; found stars incubating inside tall pillars of gas and dust which are being eroded by the radiation from nearby hot stars; and investigated ancient globular clusters.

Nearby galaxies have also come under the HST's scrutiny. In the Andromeda galaxy and the ultraluminous infrared galaxy Arp 220, it revealed not one, but two, luminous core regions. Evidence for massive black holes has been found in the cores of many galaxies, and jets of material have been imaged spewing away from these central power sources. The most massive black hole so far detected, located in the core of elliptical galaxy M87, appears to be equivalent to 3 billion solar masses.

By means of extremely long exposures of the same regions of sky, the observatory has been able to obtain the deepest views yet of the distant universe, looking back more than 90% of the age of the universe, to the era when galaxies were first being formed. The first groundbreaking HUBBLE DEEP FIELD image of the northern hemisphere was assembled from 342 separate exposures by the WFPC-2 over 10 consecutive days in December 1995. A similar deep field view was obtained for the southern hemisphere in October 1998. Both images contained about 2500 galaxies with apparent brightness down to thirtieth magnitude. A subsequent 'long exposure' infrared image of a portion of the deep field—north, taken with NICMOS, uncovered the faintest galaxies ever seen.

Images such as these have provided astronomers with the first clear optical views of the host galaxies in which quasars are embedded. By studying the highly variable shapes and sizes of galaxies in the early universe, new theories have evolved concerning the frequency of collisions and the evolution of spiral and elliptical galaxies. However, while galaxies started to form quite soon after the Big Bang, studies with the HST support the idea that star formation only reached a peak much later, some 7 billion years ago.

Apart from the deep field images, astronomers have also been able to use the HST to probe the distant universe by means of gravitational lenses. Other cosmic puzzles, such as the origin of gamma-ray bursts, the age of the universe and value of the Hubble constant, have also been attacked using the space telescope.

By monitoring the visible afterglow of gamma-ray explosions, astronomers have been able to confirm that at least some of these events occur at cosmic distances, far beyond our Milky Way. Studies of Cepheid variable stars in galaxies up to 65 million light years away have pinned down the Hubble constant to 70 km s^{-1} Mpc^{-1}, with an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

uncertainty of 10%. Based on these results, astronomers have calculated that the Universe was born about 12 billion years ago.

Discoveries such as these should keep on coming. Although the HST was originally intended to have an operational lifetime of 15 years, after the successful servicing mission 3A, the telescope's life expectancy has been extended until at least 2010, by which time the Next Generation Space Telescope may be in orbit.

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble Space Telescope: Technical Specifications

The Hubble Space Telescope (HST) (figure 1) is a cooperative project between the US National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA). Designed to cover the wavelengths between approximately 0.1 and 3μm, from the near-ultraviolet through the visible to the near-infrared portions of the spectrum, it was launched on 25 April 1990, the first of NASA's 'Great Observatories'. It is operated as a general-purpose observatory capable of observing almost all celestial objects from the solar system to the distant universe. Access to the observing time is granted using peer review of proposals solicited competitively on a yearly basis without restriction on who can apply.

HST contains an $f/24$ Ritchey–Chretien reflecting telescope with a 2.4 m diameter primary mirror and five astronomical instruments that are designed to be replaced in orbit. The entire spacecraft has a cylindrical shape, 13.1 m long and 4.3 m in diameter at its widest point, not including the solar arrays, and has a mass of 11.6 t.

HST's instruments are modules that reside in separate compartments at the telescope focal plane (figure 2). The *radial instrument* is perpendicular to the optical axis with a pick-off mirror on the optical axis. The *axial instruments* are in four bays surrounding the optical axis. There are three additional sectors outside the optical axis that contain the fine guidance sensors (FGSs), which are used to point and stabilize the telescope by measuring the precise positions of field stars. The high accuracy of the FGSs for measurement of relative

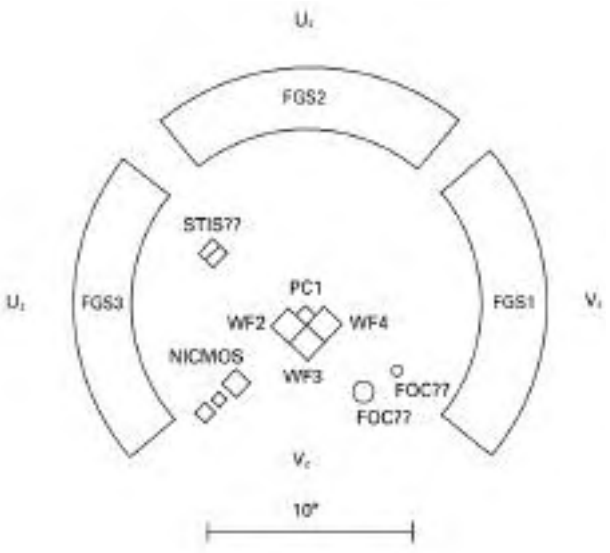


Figure 2. The different shapes show where the different instruments observe the sky relative to one another in the focal plane of the telescope. The three FGSs can observe over three-quarters of a ring surrounding the optic axis. Because the other instruments sample different parts of the focal plane, several instruments can be used simultaneously for parallel observing. The bar at the bottom of the figure is 10minutes of arc in length.

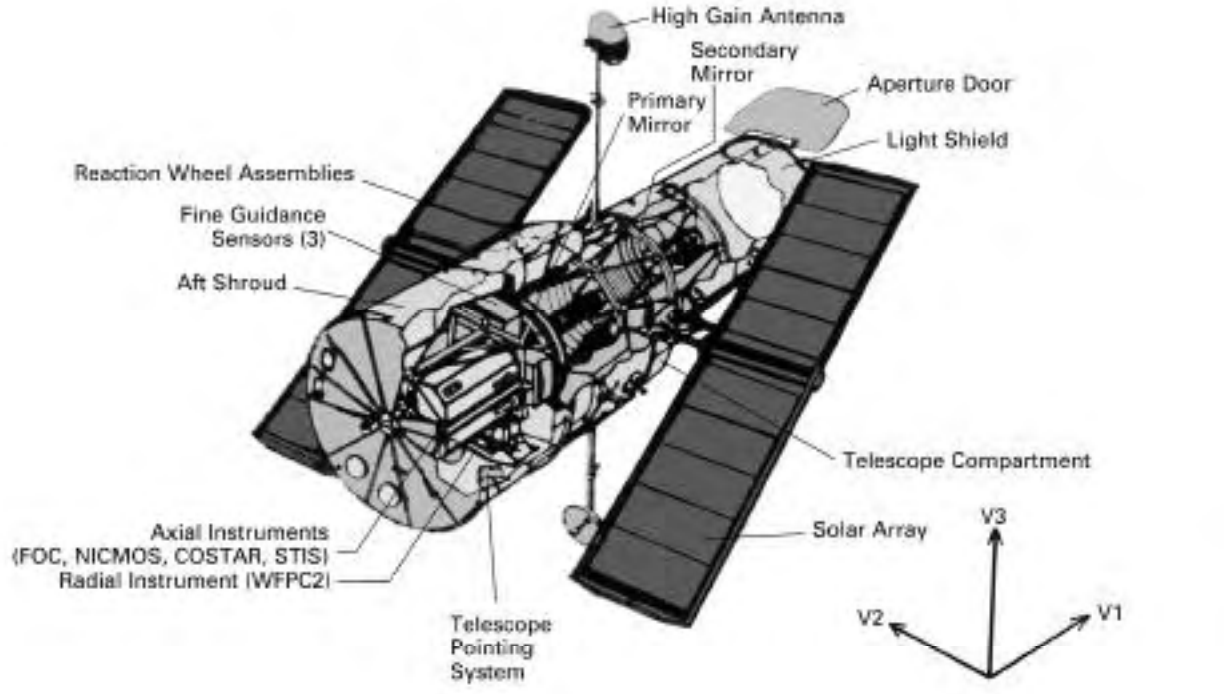


Figure 1. This schematic diagram of HST shows the key elements of the spacecraft.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

positions of stars allows them to be used as a sixth instrument for astrometry.

HST is in a circular orbit approximately 600 km above the Earth, at an inclination of 28.5° relative to the Earth's equator. It has an orbital period of 96.4 min. Its low orbit makes it accessible to visits by astronauts in *servicing missions*. During a servicing mission, the astronauts can replace the instruments, upgrade spacecraft components such as computers, repair faulty hardware and electronics, and boost the telescope to a higher orbit to offset its rate of descent of 0.4 km/yr. There have been servicing missions in 1993, 1997 and 1999, and two more are planned for 2002 and 2004.

The current complement of instruments in addition to the fine guidance sensors includes the Wide Field and Planetary Camera 2 (WFPC2, radial), the Space Telescope Imaging Spectrometer (STIS, axial), the Faint Object Camera (FOC, axial), and the Near Infrared Camera and Multi-Object Spectrometer (NICMOS, axial). The fourth axial instrument bay contains the Corrective Optics for Space Telescope Axial Replacement (COSTAR), a set of optics that correct aberrations of the primary mirror for the FOC. NICMOS stopped functioning in 1998 following the exhaustion of its solid nitrogen cryogen used to cool the detector and optics. In the servicing mission planned for 2002, the FOC will be replaced by the Advanced Camera for Surveys (ACS, axial), and the NICMOS Cooling System (NCS), a mechanical cooler, will be installed to restore functioning of the NICMOS. In the 2004 mission, the WFPC2 will be replaced by the Wide Field Camera 3 (WFC3, radial), and the COSTAR will be replaced by the Cosmic Origins Spectrograph (COS, axial).

HST was launched with a primary mirror figured to the wrong focal length, so the images in the initial instruments were severely degraded by spherical aberration. The original solar panels providing power for the spacecraft vibrated at day or night entry, making it difficult to maintain accurate pointing for precise observations at those times. The first servicing mission replaced the main axial instrument, the Wide Field and Planetary Camera (WFPC), with WFPC2, an improved version fitted with optics to correct the spherical aberration of the primary mirror. COSTAR replaced one of the original radial instruments, the High Speed Photometer, and corrected the images for the other radial instruments. The corrections restored HST to its full imaging capability, with image quality limited only by diffraction from the primary mirror. A new set of solar panels eliminated the vibrations that kept HST from maintaining its pointing accuracy.

HST maintains pointing stabilization to better than 7 milliseconds of arc and absolute pointing better than 1 second of arc. It uses an elaborate system for attitude control to achieve this accuracy. Six gyroscopes provide an inertial frame of reference that allows HST to make an initial pointing accurate to a few tens of seconds of arc. Subsequently the fine guidance sensors lock onto stars and provide very precise pointing information to the control system to stabilize the spacecraft to its highest accuracy. Reaction wheels change the orientation of the spacecraft in space in response to commands from the attitude control system. HST also contains magnetometers that sense the direction of the Earth's magnetic field; they can be used to orient the spacecraft very crudely to keep it pointed away from the Sun in case the gyroscopes fail.

Following the correction of the optics during the 1993 servicing mission, HST enjoyed success as NASA and ESA's most important mission measured by scientific and public impact. It has made discoveries in almost every field of astrophysics from the nearest to the most distant celestial objects. Free from the distorting effects of the Earth's atmosphere, HST yields images of unrivalled resolution and clarity, many of which have become public icons.

To learn more details about HST, visit http://www.stsci.edu/hst/.

Steven Beckwith

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Hubble, Edwin Powell (1889–1953)

Astronomer, born in Marshfield, MO, trained as a lawyer, with a Rhodes scholarship to Oxford University. Switched to astronomy at the University of Chicago, and became a staff astronomer at Mount Wilson Observatory where he had access to the best telescope in the world at that time—the 100 in telescope. Hubble was regarded as an aloof scientist, and, by contrast say to FRED HOYLE, did not speculate about the meaning of his discoveries, about the evolution of the universe or its origin in a 'creation event'. In fact he could be regarded as unimaginative, his work being characterized by a gathering of information from which he drew conclusions, as Francis Bacon would have done. Perhaps for this very reason the four discoveries that he made between 1922 and 1936 not only changed our view of the universe, but also endured. In the first part of this period, he proposed a classification system for nebulae, both what we would now describe as galactic and extragalactic nebulae. He recognized both emission nebulae (shining by their intrinsic emission) and reflection nebulae (redirecting the light of an associated star). He divided extragalactic nebulae into the now-standard classification system of *Hubble galaxy types*—elliptical, spiral and irregular galaxies arranged in a scheme known as the *Hubble tuning fork diagram*. In fact, with his discovery of Cepheid variable stars in NGC 6822, M33 and M31, he demonstrated the nature of extragalactic nebulae as collections of stars comparable to our own Milky Way (what we now call galaxies) at great distances. This settled a question that had been argued between CURTIS, Lundmark, SHAPLEY and ÖPIK, to name only four. He then showed that the distribution of galaxies was uniform, and they occupy a space that is cosmological. Finally, and most importantly, working with HUMASON, Hubble showed in 1929 that galaxies are moving away from us with a speed proportional to their distance (*Hubble's law*). The explanation led to the discovery of the expanding Universe. The slope of the *Hubble diagram* of the velocities of galaxies plotted against their distances is known as the *Hubble constant*, and its reciprocal, the *Hubble time*, is an estimate of the age of the universe. Hubble summarized his discoveries in a successful popular book *The Realm of the Nebulae*. He is commemorated by the name of the Hubble Space Telescope, which continues his investigations with modern space equipment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hubble's Variable Nebula (NGC 2261)

A reflection nebula in the constellation Monoceros, position RA $06^{h}\ 39.2^{m}$, dec. $+08^{\circ}\ 44'$. It is small ($2'$ by $1'$) but of quite high surface brightness. The nebula's average magnitude is 10, but, as Edwin Hubble discovered in 1916, it varies in brightness, mirroring the variability of its illuminating star, R Monocerotis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Huggins, Margaret Lindsay Murray (1848–1915)

Astronomer, whose early interest in astronomy, taught to her by her grandfather, led to a meeting with and marriage to WILLIAM HUGGINS, a spectroscopist. Together they produced some of the earliest spectra of astronomical objects, particularly the Orion nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Huggins, Sir William (1824–1910)

Astronomer, born in Stoke Newington, London, a wealthy amateur who built an observatory where he invented the stellar spectroscope. Building on KIRCHHOFF's and Bunsen's discoveries about terrestrial and solar spectral emission and absorption lines, and their conclusions that spectroscopy could reveal the composition of the source of the light, Huggins compared laboratory spectra with those of stars. At first he examined the spectra of astronomical sources visually. In 1875, after many experiments frustrated by the faintness of the stars and the insensitivity of the available photographic material, he recorded the spectrum of Vega on a dry gelatine plate. He explored the spectroscopic nature of stars, nebulae and comets. He showed that some nebulae, including the Orion nebula, have pure emission-line spectra and thus must be truly gaseous, while others, such as that in Andromeda, have continuous spectra characteristic of stars. He observed the spectrum of a nova and discovered that comets emit the light of luminescent carbon gas. He imaged solar prominences in monochromatic Hα emission by aligning a high-dispersion spectrograph slit tangent to the Sun's disc and gradually opening the slit. With Miller in 1868 he made the first measurement of the Doppler shift in a star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hulse, Russell A (1950–)

Radioastronomer and plasma physicist, born in New York City, Nobel prizewinner in physics with JOSEPH H TAYLOR (1993) 'for the discovery of a new type of pulsar, a discovery that has opened up new possibilities for the study of gravitation'. Using the 300 m radio telescope at Arecibo in Puerto Rico, Taylor and his research student Hulse searched systematically for pulsars, discovering the one now known as PSR 1913 + 16 and the *Hulse–Taylor pulsar*. The pulsar, it transpired, is in orbit around another neutron star. Each star is the mass of the Sun but very small (ten kilometers in dimension), and they are a short distance from each other, only several times the Moon's distance from the Earth. Since their mutual gravitational pull is large, their orbit shows deviations from Newton's gravitational laws. In particular, it was found that the period of the orbit is declining at the rate of about 75 millionths of a second per year, because the pulsar system loses energy by emitting gravitational waves. The rate is fully consistent with the predictions of general relativity. This was the first, albeit indirect, proof of the existence of gravitational waves.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hulst, Hendrik van de (1918–)

Dutch astronomer, born in Utrecht, became professor at Leiden. During the Second World War, because of the shortage of equipment, Dutch astronomers were confined to work in theoretical topics. JAN OORT (Leiden Observatory) realized that radio waves would be able to penetrate the dust of interstellar space and would reveal the structure of the Galaxy. He therefore set a student, van de Hulst, to determine theoretically what radio emission would come from hydrogen, the most abundant element in interstellar space, indeed in the universe. Van de Hulst showed that the strongest emission from hydrogen in the radio spectral region would arise as a result of a spontaneous change of direction of spin from the electron orbiting a neutral hydrogen atom. In any given atom the change would occur once in millions of years, but there are so many neutral hydrogen atoms in space that the cumulative effect is very significant. The prediction of radio emission at a wavelength of 21 cm was confirmed by van der Hulst, Jan Oort and C Alex Muller in 1951, and the discovery matched Oort's expectations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Humason, Milton La Salle (1891–1972)

American astronomer, born in Dodge Center, MN, who began his astronomical career as a muleteer and janitor at the Mount Wilson Observatory, and became a night assistant. He learnt photography and became HUBBLE's research assistant and then an astronomer in his own right, measuring the speeds of faint galaxies, and the properties of the supernovae in them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Humboldt, Alexander von (1769–1859)

Naturalist, geographer, explorer, born in Tegel Palace, Berlin. The *Humboldt current* off the west coast of South America is named after him. He named the gegenschein. He set in train an international project organized by GAUSS to study terrestrial magnetism. His major work, *Kosmos*, based on lectures at Berlin University, endeavoured to provide a comprehensive physical picture of the universe. In it he described the results of the survey of terrestrial magnetism, and, in a particularly influential section, he drew attention to the observations of the Sun made by HEINRICH SCHWABE of the sunspot cycle. Humboldt's description of Schwabe's work provoked numerous follow-up studies (*see* JOHANN WOLF, EDWARD SABINE).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Huygens, Christiaan (1629–95)

Born in The Hague, Netherlands, to a prominent Dutch family, tutored at home, and came under the influence of DESCARTES who was an occasional visitor at the Huygens' home. Studied law and mathematics at Leiden, turned to scientific studies of nature. Made microscopes and telescopes, and in 1655, using one of his own lenses, Huygens detected Titan, the first moon of Saturn. He discovered the true shape of the rings of Saturn with one of his telescopes with a focal length of 7 m, and in *Systema Saturnium* (1659), explained the changes in the apparent shape of the rings. His telescopes, of enormous focal length, with the objective lenses supported on masts, were so far in advance of the rest of the field that at first his observations were not believed. In 1656 Huygens patented the first pendulum clock, testing models at sea to determine longitude. He also derived the law of centrifugal force for uniform circular motion, which in combination with Kepler's third law led WREN, HALLEY and others to the inverse square law of gravitation. Huygens proved by experiment the conservation of momentum, such that the momentum before the collision of two bodies is equal to the momentum after. In his *Traité de la Lumiere* Huygens argued in favor of a wave theory of light, stating that an expanding sphere of light behaves as if each point on the wave front were a new source of radiation of the same frequency and phase. In the final years of his life Huygens composed his book *Cosmotheoros*, published posthumously, apparently regarding it as a more scientific version of FONTENELLE's discussion of extraterrestrial life. It is a discussion that on the whole a modern astronomer can identify with. The Huygens probe that will eject from the Cassini spacecraft and parachute to the surface of Titan in 2004 is named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hydra

(the Water-snake; abbrev. Hya, gen. Hydrae; area 1303 sq. deg.) A predominantly southern constellation which winds its way around almost one-third of the sky between Canis Minor and Libra. The center of the constellation culminates at midnight in mid-March, though part of it is on the meridian from late January to early May. It represents, in Greek mythology, either the multi-headed Lernaean Hydra which was killed by Hercules as the second of his 12 labors, or the water-snake that prevents Corvus, the Crow, from drinking from Crater, the cup of the god Apollo. The brightest stars of Hydra were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*. Owing to its great extent, some early celestial cartographers divided Hydra into four divisions: Hydra, Hydra et Crater, Hydra et Corvus and Continuatio Hydrae.

Hydra is the largest of the 88 constellations but is not particularly prominent. Its brightest stars are α Hydrae (Alphard), magnitude 2.0, γ Hydrae, magnitude 3.0, and ζ Hydrae and ν Hydrae, both magnitude 3.1. There are eight other stars brighter than fourth magnitude. Interesting stars include ε Hydrae, a multiple system, consisting of yellow (G5 and F7) components, magnitudes 3.8 and 7.8, separation 3″, the former of which has a close (white, A8) companion, magnitude 4.7, separation 0.2″, period 15 years, the latter of which has an unseen companion, period 9.9 days, and a fifth component, magnitude 12.7, separation 19″, and R Hydrae, a Mira-type variable (range 3.5–10.9, period about 389 days).

Other interesting objects in Hydra include M48 (NGC 2548), an open cluster of several dozen stars between ninth and thirteenth magnitude, and M68 (NGC 4590), a globular cluster, NGC 3242 (the 'Ghost of Jupiter'), a bluish planetary nebula, and M83 (NGC 5236), a spiral galaxy, all of eighth magnitude.

See also: Corvus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hydrodynamic Instabilities

Most of the universe is made of gas, over all scales ranging from the intergalactic medium to stars and giant planets. This fluid rarely remains in static equilibrium or in stationary motion—many forces and processes conspire to render it unstable. When it is unstable, it may oscillate more or less regularly around its equilibrium state, as do variable stars. Alternatively, it may reach a state of turbulent motion, which then plays a crucial role in transporting momentum, heat and chemicals and in amplifying a magnetic field when the gas is ionized. In some cases there is direct observational evidence for such turbulence: in the interstellar medium, at the surface of the Sun and of the giant planets, in the solar wind. More often only indirect proof of turbulence is available, as in accretion disks or in stellar photospheres and interiors.

To test the stability of any fluid medium in the universe, one applies to it an arbitrary perturbation—that is, one applies this perturbation to a mathematical model built with the equations of fluid dynamics (see FLUID DYNAMICS). Once perturbed, the fluid may settle back into its initial state, which is then deemed to be stable, or else the perturbation grows to finite amplitude, and the medium becomes turbulent or oscillatory. In the mathematical treatment, one makes a distinction between a linear instability, which is triggered by an exceedingly small perturbation, and a finite-amplitude instability, which requires a perturbation of a certain strength (depending on the degree of instability). However, for the description of the final state it does not much matter whether the instability is of either type.

Non-linear effects are required to prevent the infinite growth of any hydrodynamic instability. For example, turbulence acts always to suppress—at least to reduce—the cause of instability, in modifying for instance the mean flow profile (shear instabilities) or the temperature stratification (thermal convection). Numerical simulations have greatly helped to describe these finite amplitude regimes, but, even with present-day computers, the spatial resolution is often too limited for a realistic description.

Here we shall briefly review the main hydrodynamic instabilities encountered in astrophysics; we shall classify them according to their cause, describe their major properties and mention the objects where they presumably play an important role. We shall put more emphasis on recent results; for the others we refer the reader to classical treatises. Finally, we shall not discuss here the various instabilities involving a magnetic field, which are dealt with in MAGNETOHYDRODYNAMICS and ACCRETION DISKS.

Testing the linear stability

When the medium or the flow under consideration is submitted to a small perturbation, say a displacement $\delta\boldsymbol{r}(\boldsymbol{r},t)$ ($\boldsymbol{r}$ designates the spatial coordinate and t the time), then only the linear terms in $\delta\boldsymbol{r}$ may be kept in the equations describing the evolution of that perturbation. After some manipulations, the result may be cast in the general form

$$\left(\frac{\partial}{\partial t}-\mathcal{L}_1\right)\left(\frac{\partial}{\partial t}-\mathcal{L}_2\right)\cdots\left(\frac{\partial}{\partial t}-\mathcal{L}_n\right)\delta\boldsymbol{r}=0 \qquad (1)$$

where the $\mathcal{L}_i$ are linear spatial operators. Equation (1) admits separable solutions of the type

$$\delta\boldsymbol{r}(\boldsymbol{r},t)=\boldsymbol{A}(\boldsymbol{r})\exp(\sigma t)$$

and the vector function $\boldsymbol{A}$ obeys the partial differential equation (in space)

$$(\sigma-\mathcal{L}_1)(\sigma-\mathcal{L}_2)\ldots(\sigma-\mathcal{L}_n)\boldsymbol{A}(\boldsymbol{r})=\boldsymbol{0} \qquad (2)$$

which, after imposing suitable boundary conditions, constitutes an eigenvalue problem for σ. These eigenvalues depend on the spatial characteristics of the perturbation, such as its wavelength or the number of nodes. In simple cases, they are solutions of an algebraic equation of degree n, which is then called the dispersion relation. Examples of such relations are given below in equations (3) and (4).

In many cases, the eigenvalues σ are complex quantities; it is their real part $\mathcal{R}(\sigma)$ which determines the linear stability of the considered medium or flow. Instability arises whenever one of the eigenvalues has a positive real part.

Gravitational instabilities

Jeans instability

In most astrophysical objects, pressure forces balance the gravitational attraction. However, if the density (hence the mass) is too large, the medium may become unstable to infinitesimal perturbations. In a uniform medium, or locally in a non-uniform medium, the criterion for this instability takes the form

$$\sigma^2=4\pi G\rho-c_{\rm s}^2\left(\frac{2\pi}{\lambda}\right)^2\geq 0 \qquad (3)$$

where σ^2 is the exponential growth rate, $c_{\rm s}$ the isothermal sound speed ($c_{\rm s}^2=P/\rho$ in terms of pressure and density), G the gravitational constant and λ the wavelength of the perturbation. Setting $\sigma=0$ one obtains a critical wavelength $\lambda_{\rm J}$ and a critical mass $M_{\rm J}\approx\rho\lambda_{\rm J}^3$. If the medium is already turbulent for another reason, the turbulent pressure acts to increase the critical mass. So does also a magnetic field, when present.

In a uniformly rotating flow, the criterion includes a Coriolis term:

$$\sigma^2=4\pi G\rho-c_{\rm s}^2\left(\frac{2\pi}{\lambda}\right)^2-4\Omega^2\geq 0.$$

With differential rotation, as in a galactic disk, and assuming that all the mass is concentrated in the equatorial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

plane, the instability criterion involves the epicyclic frequency Ω_{ec} (Toomre 1964):

$$\sigma^2 = 2\pi G\Sigma\frac{2\pi}{\lambda} - 4\Omega_{ec}^2 \geq 0 \qquad \Omega_{ec}^2 = \frac{1}{s^3}\frac{d}{ds}(s^2\Omega)^2 \quad (4)$$

with Σ being the surface density and s the distance from the rotation axis.

The Jeans instability is probably the main cause of star formation in the interstellar medium; the dynamical phase of the contraction ends when the gas becomes opaque enough to build up a temperature gradient: the associated pressure gradient is then able to balance the gravitational attraction.

Rayleigh–Taylor instability
This instability arises when denser fluid is superposed on lighter fluid in a gravitational field or an accelerating medium. It plays a crucial role in mixing chemical species during supernova explosions. A detailed description may be found in FLUID DYNAMICS.

Shear instabilities

Kelvin–Helmholtz instability
In this prototype of all shear instabilities, two parallel flows of different velocities V_1 and V_2 are separated by a sharp interface: if the interface is perturbed by an infinitesimal amount, all downstream wavelengths are unstable. When the transition layer between the two flows has a finite width H in z, only the wavelengths larger than $\approx H$ are unstable. This result holds also for a smooth velocity profile $V_x(z)$, containing an inflection point at which the vorticity dV_x/dz has a local extremum of width H (see figure 1); this instability condition is called the Rayleigh-Fjørtøft criterion. The instability domain is further reduced when the viscosity (ν) is taken into account; the instability occurs if the Reynolds number is larger than unity, with this number being defined as

$$\mathrm{Re} = \frac{H^2}{\nu}\left|\frac{dV_x}{dz}\right|_{max} \quad (5)$$

where $|dV_x/dz|_{max}$ measures the vorticity contrast. This plain Kelvin–Helmholtz instability is likely to occur in ASTROPHYSICAL JETS.

Even in the absence of a such a vorticity extremum, a parallel shear flow becomes unstable to perturbations of finite amplitude, but at much higher Reynolds number, as demonstrated by many laboratory experiments. In channel flows, the critical Reynolds number is found around 1000, with H being the channel width and replacing the vorticity contrast in the definition above by a measure of the vorticity itself.

Shear instability in a stratified medium
When an incompressible shear flow is stratified in density (in z, the cross-stream direction), buoyancy acts to inhibit the instability. Neglecting dissipative effects, instability sets in only if the Richardson criterion is met:

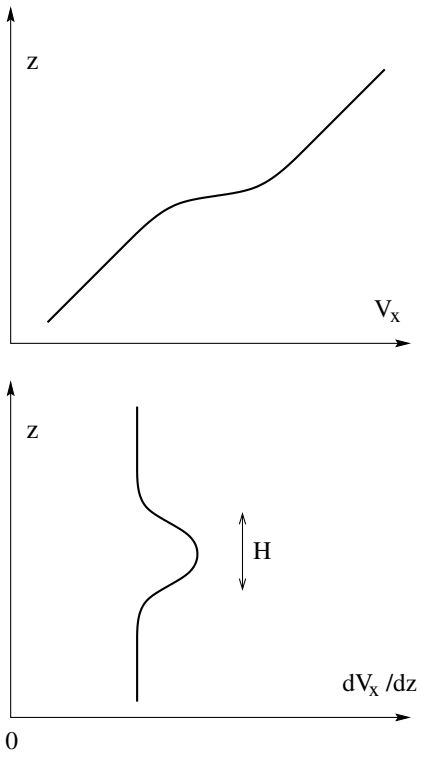


Figure 1. Plane parallel shear flow: the velocity vector is parallel to the x-axis and it varies in the (span-wise) z-direction. In this example, the velocity profile has an inflection point where the vorticity dV_x/dz is extremum: according to the Rayleigh-Fjørtøft criterion, this flow is linearly unstable.

$$\mathrm{Ri} = \frac{N^2}{|dV_x/dz|^2} \leq \mathrm{Ri}_c \qquad N^2 = -g\frac{d\ln\rho}{dz} \quad (6)$$

where N is the buoyancy frequency, which measures the stability of the density stratification (g is the acceleration due to gravity). The critical Richardson number $\mathrm{Ri_c}$ is of order unity.

Astrophysical fluids are compressible in general, and there the entropy stratification plays the role of the density stratification for incompressible fluids. The buoyancy frequency is then defined as

$$N_T^2 = \frac{g\delta}{H_P}\left[\frac{d\ln T}{d\ln P} - \left(\frac{\partial\ln T}{\partial\ln P}\right)_{ad}\right] = \frac{g\delta}{H_P}(\nabla - \nabla_{ad}) \quad (7)$$

with $\delta = -(\partial\ln\rho/\partial\ln T)_P$, H_P being the pressure scale height, and the notations used in stellar structure theory for the actual (∇) and adiabatic (∇_{ad}) temperature gradients having been introduced. Note that the stabilizing effect of the entropy stratification is felt only in the vertical direction: it does not prevent a horizontal shear from becoming unstable. In the latter case the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

turbulent motions are anisotropic, i.e. much stronger in the horizontal than in the vertical direction.

However, this instability criterion (6) applies only for adiabatic perturbations. In the presence of thermal diffusion (due to conduction or radiation), the buoyancy force weakens, and therefore also its stabilizing effect. The outcome is that the flow may sustain turbulence as soon as

$$\frac{N_T^2}{|\mathrm{d}V_x/\mathrm{d}z|^2}\frac{v\ell}{K} \leq \mathrm{Ri_c} \tag{8}$$

where v is the root mean square velocity, ℓ the dimension of the largest turbulent eddies and K the thermal diffusivity. These motions will carry heat and chemicals across the flow in the z-direction, with a turbulent diffusivity

$$D_\mathrm{t} \approx K\frac{|\mathrm{d}V_x/\mathrm{d}z|^2}{N_T^2}.$$

So far we have assumed that the fluid is homogeneous. A chemical composition gradient, with molecular weight μ increasing with depth, will also act to stabilize the flow, and one retrieves the original Richardson criterion (6) with

$$N^2 = N_\mu^2 = \frac{g\varphi}{H_P}\frac{\mathrm{d}\ln\mu}{\mathrm{d}\ln P} \qquad \varphi = \left(\frac{\partial\ln\rho}{\partial\ln\mu}\right)_{P,T}.$$

However, the stabilizing effect of this μ gradient may be drastically lowered if the horizontal shear is unstable, in which case the anisotropic turbulence acts to smooth out the horizontal inhomogeneities. The resultant turbulent diffusivity in the vertical direction then becomes (Talon and Zahn 1997)

$$D_\mathrm{t} \approx \frac{|\mathrm{d}V_x/\mathrm{d}z|^2}{N_T^2/(K+D_\mathrm{h})+N_\mu^2/D_\mathrm{h}}$$

where N_T^2 is the thermal component of the stratification as defined in equation (7) and D_h the horizontal diffusivity generated by the horizontal shear.

These shear instabilities certainly play an important role in the mixing of stably stratified stellar interiors.

Shear instability of differential rotation

Most of the results above can be transposed to differentially rotating flows. In cylindrical rotation, where the rotation rate depends only on the distance s from the rotation axis, an extremum of vorticity again renders the flow linearly unstable. Another linear instability arises when the angular momentum decreases with the distance from the rotation axis: a slight axisymmetric perturbation then grows because the pressure gradient no longer balances the centrifugal force (Rayleigh 1916). This instability has been studied in the laboratory with the famous Couette–Taylor experiment, where liquid is sheared between two concentric cylinders of mean radius s rotating at different speeds Ω_1 and Ω_2. Taking the gap Δs between cylinders as the characteristic length, one defines the Reynolds number as

$$\mathrm{Re} = \frac{s|\Omega_2-\Omega_1|\Delta s}{\nu}.$$

In the narrow gap limit, the critical Reynolds number for this linear instability is $\mathrm{Re_c} = 41.2(s/\Delta s)^{1/2}$, provided that $\Omega_1 s_1^2 > \Omega_2 s_2^2$.

However, even when $\Omega_1 s_1^2 < \Omega_2 s_2^2$, where the flow is linearly stable, turbulence has been observed for various rotation regimes, implying that the instability occurs at finite amplitude. There is still debate on whether a positive gradient of angular momentum is able to suppress hydrodynamic turbulence for certain flow profiles, such as in the Keplerian rotation of accretion disks. However, laboratory experiments seem to indicate that turbulence occurs whenever the gradient Reynolds number Re^* exceeds some critical value:

$$\mathrm{Re}^* = \frac{s^3}{\nu}\frac{\Delta\Omega}{\Delta s} \geq \mathrm{Re_c^*} \approx 6\times 10^5.$$

For a recent discussion of this problem we refer to Richard and Zahn (1999).

In objects such as stars and planets, a differential rotation in latitude becomes linearly unstable when the rotation profile meets certain conditions. For instance, when Ω varies in proportion to $1-a\cos^2\theta$ (θ being the colatitude), the flow is unstable to infinitesimal horizontal perturbations for $a > 0.29$ (Watson 1981).

Another instability, called baroclinic instability, may arise in stars whenever the rotation departs from cylindrical rotation, because the level surfaces (or isobares) do not coincide with the surfaces of equal density, and hence of equal entropy, in a homogeneous star. One is then tempted to conclude that an axisymmetric displacement directed sideways, in the wedge between these two surfaces, will escape the restoring action of the buoyancy force, and therefore lead to instability. However, such a displacement cannot be entirely arbitrary: it has to conserve angular momentum. When this is taken in account, one finds that the baroclinic instability occurs only when the specific angular momentum $\Omega r^2\sin^2\theta$ increases towards the equator on an isentropic surface.

However, this condition has been derived by assuming that the perturbations were adiabatic. Thermal diffusion will here again weaken the buoyancy force, and since in stellar interiors the viscosity ν is much smaller than the thermal conductivity K, axisymmetric perturbations will become unstable as soon as either

$$\frac{1}{s^3}\frac{\partial}{\partial s}(\Omega s^2)^2 < \frac{\nu}{K}N_T^2 \qquad \text{or} \qquad \left|s\frac{\partial\Omega}{\partial z}\right| > \frac{\nu}{K}N_T^2.$$

This so-called Goldreich–Schubert–Fricke instability belongs to the class of multidiffusive instabilities, which arise because two agents (here heat and momentum) diffuse at sufficiently different rates. Until now it has still not been

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

clarified whether this instability will lead to turbulence and mixing once it reaches its finite-amplitude regime.

On the other hand, any differential rotation is prone to shear instabilities of finite amplitude, as has been mentioned earlier: the situation is then very much like that of plane parallel shear, and the results quoted above apply. The horizontal shear of differential rotation in latitude will easily become unstable, since the Reynolds number characterizing such flows is extremely high ($\mathrm{Re}^* \approx 10^{15}$ at the base of the solar convection zone). A stable entropy stratification will inhibit the vertical shear instabilities, but only to some extent, as has already been explained: it suffices to express the vertical shear $|\mathrm{d}V_x/\mathrm{d}z|$ by $r \sin\theta \, \mathrm{d}\Omega/\mathrm{d}r$. These instabilities certainly play an important role in the mixing of stably stratified stellar interiors (Zahn 1992).

Thermal instabilities

Convection

A temperature gradient is required to transport heat through a conductive or optically thick medium, and often this gradient is parallel to the gravity force, as illustrated in stars and accretion disks. In terms of the heat flux F and of the thermal diffusivity K, this temperature gradient is given by

$$F = K \frac{\rho C_P T}{H_P} \left(\frac{\mathrm{d} \ln T}{\mathrm{d} \ln P} \right) \tag{9}$$

(C_P is the heat capacity at constant pressure). In such a stratification, a vertical perturbation will grow if it rises in colder surroundings, and this will lead to a familiar instability called thermal convection. This (linear) instability sets in when the Schwarzschild criterion is satisfied:

$$\frac{\mathrm{d} \ln T}{\mathrm{d} \ln P} > \left(\frac{\partial \ln T}{\partial \ln P} \right)_{\mathrm{ad}}$$

or in the classical notation of stellar structure theory, $\nabla > \nabla_{\mathrm{ad}}$ (small corrections apply when including the effect of thermal and viscous dissipation). ∇_{ad} measures the temperature increase in an adiabatic compression. The temperature gradient then tends to become adiabatic, and a fraction of the heat flux is transported through the convective motions. In astrophysical settings these motions are turbulent, and at sufficient optical depth the energy flux is given by

$$F_{\mathrm{conv}} \approx v\ell \frac{\rho C_P T}{H_P} (\nabla - \nabla_{\mathrm{ad}}).$$

Defining the radiative gradient ∇_{rad} as the gradient which would be needed to carry the total flux F_{tot} through conduction or radiation (cf equation (9)), the conservation of flux $F_{\mathrm{tot}} = F_{\mathrm{conv}} + F_{\mathrm{rad}}$ translates into

$$\frac{\nabla_{\mathrm{rad}} - \nabla}{\nabla - \nabla_{\mathrm{ad}}} = \frac{v\ell}{K}.$$

The higher the Péclet number $v\ell/K$, the closer the temperature gradient comes to being adiabatic (at the base of the solar convection zone, $v\ell/K \approx 10^6$). A prescription is needed to estimate the turbulent diffusivity $v\ell$, and the classical mixing-length treatment is still widely used, for lack of something better: one assumes that the convective eddies are accelerated over a distance Λ (the mixing length), after which they dissolve in the surrounding medium where they release their heat content. For most purposes, the mixing length is taken as proportional to the local pressure scale height.

Great progress has been achieved already with three-dimensional numerical simulations, but the range of scales which has to be spanned is challenging, and these simulations are necessarily restricted to a small physical domain (near the surface of the star for instance) or to the largest motion scales (when the goal is to encompass a whole convection zone).

κ-Mechanism

In stellar interiors, the opacity peaks in the ionization zones of the most abundant species; there radiative diffusion is then no longer a dissipative process, and oscillations of the star grow owing to a valve mechanism. This thermal oscillatory instability is the cause of the pulsation of most variable stars (see STELLAR PULSATION).

Eddington–Sweet circulation

The centrifugal distortion makes it impossible, in general, to achieve radiative equilibrium in a rotating star, and the resulting thermal imbalance induces a large-scale meridian flow throughout the radiation zone. The time-scale for this Eddington–Sweet circulation is:

$$t_{\mathrm{ES}} = t_{\mathrm{KH}} \left(\frac{\Omega^2 R^3}{GM} \right)^{-1} \qquad t_{\mathrm{KH}} = \frac{GM^2}{RL}$$

where M, R and L are the mass, radius and luminosity of the star and t_{KH} is the Kelvin–Helmholtz time governing the global thermal adjustment. This time-scale is very large: only rapidly rotating stars will feel the effect of that circulation, which is then potentially able to mix them. This circulation also transports angular momentum, and it tends to establish a rotation profile for which there would be no circulation at all, and hence no mixing. However, for several reasons this regime of complete rest is never attained: (i) most often the vertical shear is unstable, and turbulence generates its own transport of angular momentum; (ii) the star loses angular momentum carried away by a wind; (iii) while evolving, and hence contracting or expanding, the star alters its internal angular momentum distribution (cf Zahn 1992).

More seriously, the vertical composition gradient which builds up through nuclear burning interferes with the circulation and acts to slow it down. Presumably it will not succeed in suppressing it altogether (Maeder and Zahn 1998). However, we are still in great need of a realistic description of the turbulence arising in such a stratified, differentially rotating star to properly assess the amount of mixing.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Chandrasekhar S 1961 *Hydrodynamic and Hydromagnetic Stability* (Oxford: Clarendon)

Drazin P G and Reid W H 1981 *Hydrodynamic Stability* (Cambridge: Cambridge University Press)

Landau L D and Lifshitz E M 1987 *Fluid Mechanics* (Oxford: Pergamon)

Maeder A and Zahn J-P 1998 Stellar evolution with rotation: III. Meridional circulation with μ-gradients and non-stationarity *Astron. Astrophys.* **334** 1000

Pinsonneault M 1997 Mixing in stars *Ann. Rev. Astron. Astrophys.* **35** 557

Richard D and Zahn J-P 1999 Turbulence in differentially rotating flows: what can be learned from the Couette–Taylor experiment *Astron. Astrophys.* **347** 734

Talon S and Zahn J-P 1997 Anisotropic diffusion and shear instabilities *Astron. Astrophys.* **317** 749

Toomre A 1964 On the gravitational instability of a disk of stars *Astrophys. J.* **139** 1217

Watson M 1981 Shear instability of differential rotation in stars *Geophys. Astrophys. Fluid Dyn.* **16** 285

Zahn J-P 1992 Circulation and turbulence in rotating stars *Astron. Astrophys.* **265** 115

Zahn J-P and Zinn-Justin J (ed) 1993 *Astrophysical Fluid Dynamics (Les Houches LXVII)* (Amsterdam: North-Holland)

Jean-Paul Zahn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hydrogen (H)

The lightest and most common element in the universe. About 92% by number of atoms, and about 75% by mass, of all the directly observed matter in the universe is hydrogen. The most common form of hydrogen atom consists of one positively charged particle, the proton, and one negatively charged particle, the electron, which orbits the proton. The atomic mass is 1 and the atomic number (i.e. charge on the nucleus) is also 1. Hydrogen nuclei exist with one or two neutrons in addition to the one proton. These forms of 'heavy hydrogen' are called deuterium and tritium, respectively. Regions of neutral hydrogen are referred to as HI regions. Regions of ionized hydrogen are HII regions.

See also: cosmic abundance of elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hydrogen Spectrum

The series of absorption or emission lines that are characteristic of the hydrogen atom. According to the Bohr theory of the hydrogen atom, devised by Danish physicist Neils Bohr (1885–1962) in 1913, the hydrogen atom can be envisaged as consisting of a central nucleus (a proton) around which a single electron revolves. The electron is located in one of a number of possible permitted orbits, each of which corresponds to a different energy level. If an electron drops down from a higher to a lower level (makes a 'downward transition'), energy is emitted in the form of a photon, the energy of which is equal to the difference in energy between the two levels. Conversely, if an electron absorbs a photon of the correct energy, it can make an upward transition from a lower to a higher level. Because a photon of a particular energy corresponds to light of a particular wavelength, the various downward, or upward, permitted transitions correspond to the emission, or absorption, of particular wavelengths of light. This mechanism is responsible for the production of the various lines in the hydrogen spectrum.

In the hydrogen atom, the various energy levels are denoted by a number (a quantum number), n; $n = 1$ corresponding to the lowest level, $n = 2$ to the second level (the first 'excited' level), $n = 3$ to the third level (the second excited level), and so on. The wavelength, λ, of the line that is produced by a transition between a lower level, n, and a higher level, m, is given, in nanometers, by the following formula: $\lambda = 91.16/(\frac{1}{n^2} - \frac{1}{m^2})$. For example, a transition between $n = 2$ and $n = 3$ gives rise to a line of wavelength $\lambda = 91.16/(\frac{1}{2^2} - \frac{1}{3^2}) = 91.16/(\frac{1}{4} - \frac{1}{9}) = 656.3$ nm. This line, which is called the hydrogen alpha (Hα) line, lies towards the red end of the visible spectrum and is the first in a series of lines, known as the Balmer series, which is produced by the various possible transitions between the $n = 2$ level and higher levels. The second line, Hβ, which occurs at a wavelength of 486.2 nm corresponds to transitions between the second and fourth levels ($n = 2$, $m = 4$), the third, Hγ, at 434.1 nm, and so on. The series comes to an end at 364.6 nm ($n = 2$, $m = \infty$), which corresponds to the electron being completely removed from the atom.

Transitions between the ground level ($n = 1$) and higher levels involve larger energy gaps and, because the energies of photons are inversely proportional to their wavelengths, give rise to lines of shorter wavelengths than those of the Balmer series. The resulting series of lines, which occur in the ultraviolet region of the spectrum, is called the Lyman series. The series begins with Lyman alpha, at 121.5 nm and ends at the series limit of 91.16 nm. Transitions between the third level ($n = 3$) and higher levels correspond to small energy gaps and give rise to the Paschen series, in the infrared region of the spectrum.

See also: absorption spectrum, atom, emission spectrum, photon, quantum theory, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hydrus

(the Little Water-snake; abbrev. Hyi, gen. Hydri; area 243 sq. deg.) A southern constellation which lies between Tucana and Horologium, and culminates at midnight in late October. It was first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, which lies between the Large and Small Magellanic Clouds, the brightest stars in Hydrus are β Hydri, magnitude 2.8, which is the closest conspicuous star to the south celestial pole, about 12° distant, α Hydri, magnitude 2.9, and γ Hydri, magnitude 3.3. There are no other stars brighter than fourth magnitude. Interesting objects include NGC 1466, an eleventh-magnitude globular cluster of the Large Magellanic Cloud, and NGC 1511, a small edge-on spiral galaxy, also eleventh magnitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hygiea

The tenth asteroid to be discovered, by Annibale De Gasparis in 1849, and so designated (10) Hygiea. It is the fourth-largest asteroid, with a diameter of 408 km. It orbits the Sun at a mean distance of 3.14 AU (470 million km), in the outer part of the main asteroid belt, in a period of 5.55 years. The orbital inclination is 4, the eccentricity 0.12 and the rotation period an estimated 27.6 h. It is a C-type asteroid, with a reflectance spectrum very similar to that of carbonaceous chondrites, and has an albedo of 0.07.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hypatia of Alexandria (370–415)

Born in Alexandria, Egypt, daughter of the mathematician and philosopher Theon, and the first woman mathematician of note. As head of the Platonist school at Alexandria, she symbolized the learning that early Christians identified with paganism, and became a focal point of riots during which she was brutally murdered, perhaps under the orders of Cyril, the fanatical Christian Patriarch of Alexandria. She wrote commentaries on algebra, arithmetic, conics and on PTOLEMY's astronomical works, all lost. She developed apparatus, including an astrolabe (for measuring the positions of the stars, planets and Sun).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hyperbola

The conic section obtained when a right circular cone is cut by a plane which makes an angle with the base greater than that made by the side of the cone. It is an *open* curve, i.e. it does not close on itself like an ellipse.

See also: conic sections, ellipse, parabola.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hyperion

A mid-sized icy satellite of Saturn, discovered by W C Bond and, independently, William Lassell, in 1848. It orbits at a distance of 1 481 000 km, and is an irregular object, measuring 330 by 260 by 220 km. Hyperion was probably once a larger object which suffered a number of near-catastrophic collisions. Most of the impact debris would have been swept up by Titan, but some would have peppered Rhea, giving the latter its heavily cratered surface. Hyperion has chaotic rotation: it tumbles along in its orbit instead of rotating about a single axis; this adds credence to the collision hypothesis. Its surface is heavily cratered, with a scarp over 200 km long named Bond–Lassell Dorsum in honor of the satellite's discoverers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Iapetus

A mid-sized icy satellite of Saturn, discovered by Giovanni Cassini in 1671. It is 1440 km in diameter, and orbits at a distance of 3 561 000 km. Iapetus is unique in the solar system in its harlequin appearance: the leading hemisphere—facing the direction of orbital motion—is dark, with an albedo of around 0.05, while the trailing hemisphere is bright, with an albedo of 0.5. Its magnitude thus changes in the course of its orbit, as Cassini himself noticed, and he guessed correctly at the reason. It has been suggested that dust from Saturn's dark outer satellite Phoebe, thrown off by impacts, is spiraling inward and is swept up by Iapetus, which, like most of Saturn's satellites, has captured rotation, so that the dust would accumulate on the same (leading) hemisphere. It may be that Iapetus' surface ices contain methane, and when they are shock-heated by the impact of the dust particles, dark tarry organic substances are formed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ibn Rushd, Abu'l Waleed Muhammad Ibn Ahmad Ibn Muhammad [known as Averroes] (1128–98)

Islamic philosopher, lawyer, physician and astronomer, born in Cordoba, Spain, became physician to the Caliph of Morocco. Ibn Rushd's controversial views on philosophy—he was the greatest Islamic proponent of ARISTOTLE—led to the burning of his books and banishment by the Caliph. He wrote numerous books (87 survive) dealing with philosophy, logic, medicine, jurisprudence and astronomy, which were translated (his name westernized to Averroes, The Commentator) and were very influential on European medieval thought. In astronomy he wrote a *Treatise on the Motion of the Sphere*, *Kitab fi-Harakat al-Falak*, and a summary and commentary on PTOLEMY's *Almagest*. His Aristotelianism was retrogressive in that, admitting that developments of the Ptolemaic system such as epicyclic motion of the planets gave numerically more accurate results, he nevertheless held that only concentric spheres were perfect enough to make up the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ibn Sina, Abu Ali [known as Avicenna] (980–1037)

Islamic philosopher and scientist, born in Kharmaithen, near Bukhara (now Uzbekistan). Avicenna's life was marked by swings of fortune, from physician and favored scholar of a Samanid prince, to peripatetic country doctor when the Samanid regime fell. He became court physician in Hamadan (Iran), where he was for a time imprisoned as a political prisoner. He ended his life in a peaceful, contemplative retirement in Isfaham, where he completed his two most important works, *The Canon of Medicine* (the most influential single book in the history of medicine, identifying the manner of the spread of epidemics, including tuberculosis) and *The Book of Healing* (an encyclopedia covering the whole of the sciences). Avicenna observed Venus as a spot against the surface of the Sun, inferring that Venus must be closer to the Earth than the Sun. He was one of several Islamic authorities who stated, on no particular grounds, that the velocity of light is finite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ibrahim, ibn Sinan ibn Thabit ibn Qurra (908–46)

Geometrist and astronomer, born in Baghdad (now Iraq), grandson of THABIT IBN QURRA, worked on conic sections, studied the motions of the Sun, particularly as revealed through the geometry of shadows, and summarized his work in a book which analyses the data in terms of PTOLEMY's and his own theory of the solar motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Icarus

An Apollo asteroid discovered by Walter Baade in 1949, designated (1566) Icarus. In 1968, when it passed just 0.04 AU from the Earth, it was observed by radar and found to be about 1 km in diameter and to rotate in 2.27 h. Icarus is one of only a handful of asteroids with a perihelion distance (0.19 AU) within the orbit of Mercury; aphelion is at 1.97 AU, and its mean distance from the Sun is 1.08 AU (161 million km). The inclination is 23° and the eccentricity 0.83. Icarus is classed as a potentially hazardous asteroid.

See also: near-Earth asteroids

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ida

An asteroid discovered by Johann Palisa in 1884, designated (243) Ida. In August 1993 it became the second asteroid to be imaged in close-up from a spacecraft when the Galileo probe passed within 2400 km *en route* to Jupiter. Ida was found to be a heavily cratered object measuring $56 \times 24 \times 21$ km, possessing a small companion: the first confirmed asteroidal satellite. Named Dactyl, the satellite measures $1.6 \times 1.3 \times 1.0$ km. From Dactyl's orbital motion, Ida's density was found to be between 2.0 and 3.1 g cm^{-3}, consistent with a composition similar to that of ordinary chondrites, as suggested by its S-type reflectance spectrum. It orbits the Sun in the main asteroid belt at a mean distance of 2.86 AU (428 million km) in a period of 4.84 years; the orbital inclination is 1°, the eccentricity 0.04 and the rotation period 4.6 h.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Imaging Near-Earth Space

Space in the Earth's neighbourhood was studied remotely from the ground long before the space age began, mostly through observation of the AURORA, but also by radio investigation of the ionosphere (see PLANETARY IONOSPHERES) and the plasmasphere (the low-energy PLASMA region around the Earth) (see MAGNETOSPHERE OF EARTH: PLASMASPHERE). More recently, it has been studied mainly through direct *in situ* measurements of charged particles and electromagnetic fields. A major problem for remote sensing of plasmas is that protons neither emit nor scatter photons. The first views of Earth from space showed little evidence for the SOLAR WIND, radiation belts (see MAGNETOSPHERE OF EARTH: RADIATION BELTS), or plasmasphere that were discovered from early spacecraft measurements. When the Earth was viewed in Lyman-α ultraviolet light by the Apollo 16 mission, an extended hydrogen gas geocorona was evident, and there was even a hint of day–night asymmetry suggestive of an interaction with solar radiation. However, space views of telltale auroral displays made it clear that something fascinating was going on in space around the Earth. At the time, there was no simple means of obtaining images of the space plasmas responsible for these phenomena.

In the 1970s, measurements showed the existence of solar helium ion emissions scattered from helium ions in the plasmasphere. Analysis showed that the fluxes were sufficient to image the helium in this region using cameras of suitable response and sensitivity. In the 1980s, it was discovered that the energetic plasma regions around the Earth emit a glow of energetic neutral atoms (ENAs) (see ATOM). This results from charge exchange interactions between energetic ions and low-energy neutral gas atoms. At that point, scientists began to realize that, since fast neutral atoms radiate in nearly straight lines, they can serve much as photons can to form a flux image of their source regions. Together, these discoveries offered the opportunity to remotely sense and image diffuse space plasmas that are otherwise invisible. The ENA imaging concept was proven using instruments designed for charged particle measurements on the ISEE mission in the 1980s and the POLAR mission in the 1990s. A plan emerged for the development of the NASA Imager for Magnetopause to Aurora Global Exploration (IMAGE) in the late 1990s. Using all of the known plasma imaging techniques, IMAGE completed its first year of operations in May 2001, changing forever the way we view space plasma storms.

The SOHO mission's spectacular imagery of the Sun has provided dramatic evidence that it is anything but a quiescent and constant star. SOHO showed huge explosions launching billion-tonne coronal mass ejections (see SOLAR CORONAL MASS EJECTION) into the solar system at millions of kilometers per hour, some of which impact the Earth on their way to the heliopause. IMAGE has provided corresponding images of plasmas around the Earth and their response to the solar wind buffeting that they receive. Multispectral movies show the development of electron and ion auroras, as the solar wind scours the outer boundaries of the magnetosphere (see PLANETARY MAGNETOSPHERES) and enters into the dayside cusp regions. The ionosphere is seen to respond to auroral energy inputs by flowing out into space. Clouds of hot plasma form near midnight and drift rapidly as they encircle the Earth, inflating the geomagnetic field. The cold plasma near the Earth is swept sunward by the return flow through the centre of the magnetosphere. Finally, one sees the galactic gas flowing through the solar system, the remnant nebular dust disk (see NEBULA) and their interactions with the solar wind.

Instrumentation

The ENA glow from near-Earth space extends over several orders of magnitude in particle energy, requiring multiple imaging techniques. Energy and mass analysis of the neutral atoms is most effectively done using electrostatic deflection and time-of-flight measurements. Detection of secondary electrons produced on particle passage through thin carbon foils is an effective technique for recording the passage of a particle, yielding a time-of-flight start signal.

At energies greater than about 20 keV (greater for heavier species), neutral atoms will penetrate carbon foils that are thick enough to stop solar UV radiation, a significant contaminant signal. From 20 keV down to about 1 keV, neutral atoms can penetrate the thinnest carbon foils, but these foils do not stop UV radiation, so it is necessary to use sophisticated diffraction grating filters that stop very nearly all of the UV without stopping so large a fraction of the fast particles. In both higher-energy ranges, the subject of interest is the hot plasma that forms around the Earth, particularly during magnetospheric storms. This plasma carries the largest and longest-lived electrical current of such storms, known as the ring current because of its eventual encirclement of the Earth. Cameras for this energy range must have wide-angle fields of view, 120° or wider.

Below 1 keV, neutral atoms cannot penetrate even the thinnest carbon foils, so it is necessary first to convert the atoms into ions using interactions with other atoms, ideally to produce corresponding negative ions via electron attachment. Once charged, the atoms can be manipulated with electrostatic fields to measure their energy and accelerate them, and then time-of-flight methods can be used to determine their velocity or, knowing their energy, their mass. In this lowest energy range, the target of greatest interest may be the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

ionosphere itself, where plasmas are initially heated to a few tens of eV or higher, causing them to escape promptly from the Earth's gravity and flow into the magnetosphere. Cameras for this energy range may have somewhat smaller fields of view, but 90° or wider is highly desirable.

The scattered solar extreme UV (EUV) glow from near-Earth space originates from the atomically active cold helium ions of the ionosphere and its dense extension into the magnetosphere, known as the plasmasphere. This glow had been noted by some sounding rocket instruments and crudely imaged during Apollo mission flights to the Moon. The plasmasphere is generally contained within geomagnetic flux tubes extending at most about 4–5 Earth radii, requiring a wide-angle camera when viewed even from high-altitude Earth-orbiting spacecraft. Here again, a 120° field of view or larger is essential to encompass the object being imaged. Spectral coverage must be confined to the 30.4 nm line of He^+, which produces a substantial isotropic glow from the corresponding solar emission feature. The aurora is also active in this wavelength range, but, because of its very different morphology, will not be confused with plasmaspheric emissions.

The auroral far ultraviolet (FUV) glow is preferable to visible emissions for auroral imaging because it is relatively uncontaminated by light scattered from the sunlit parts of the auroral zone. The most useful wavelength as a measure of electron precipitation is the 135.6 nm line of oxygen, because it is scattered only very slightly and can be simply related to the energy flux of electron precipitation. Another bright line at 130.4 nm must be discriminated against because it is severely scattered and contaminated with airglow. In addition, energetic protons precipitating into the atmosphere in what is called the 'proton aurora' also emit FUV photons. These precipitating protons spend part of their time as fast neutral atoms as they begin to charge exchange with atmospheric atoms, some of which radiate in Doppler-shifted Lyman-α. These photons can be distinguished from others emitted by the cold atmospheric hydrogen by a space-based spectrographic imager capable of resolving the red shift of the precipitating keV protons as they move rapidly away from the instrument into the atmosphere. Imagers for auroral emissions can have much narrower fields of view (~15°), since the aurora is a relatively small feature of the magnetosphere, contained for the most part within the Earth's disk.

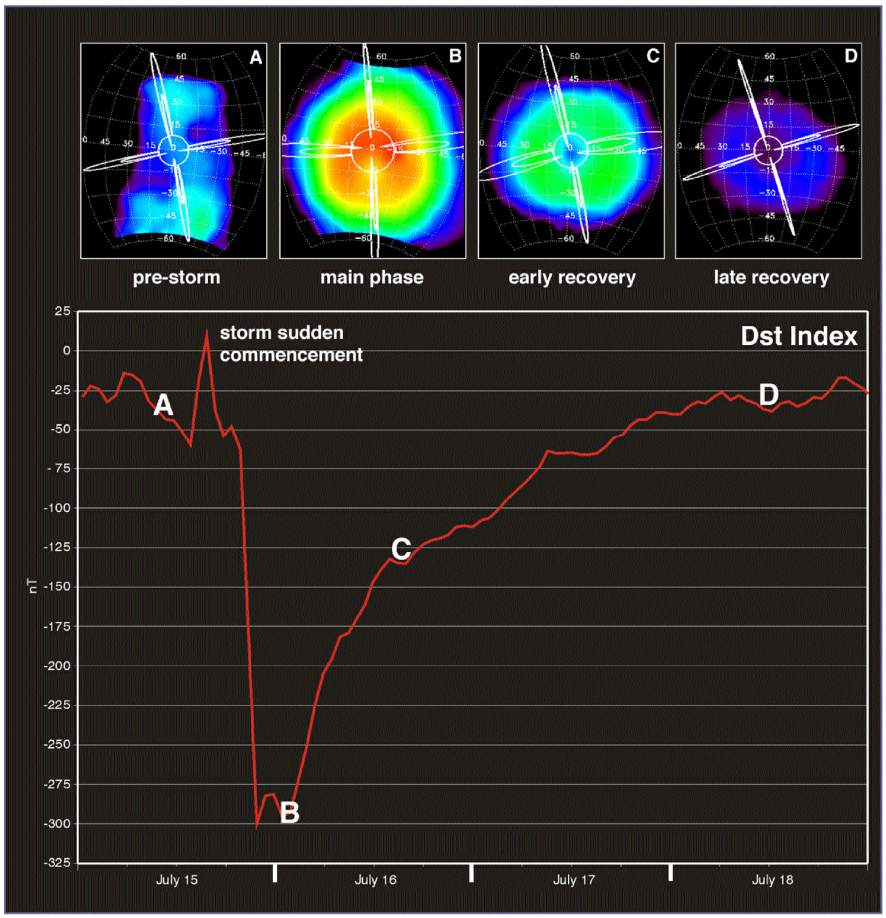


Figure 1. The development of ring current neutral atom glow during the overall development of a space plasma storm, as indicated by the concurrent variation of the equatorial geomagnetic field intensity (known as the storm time deflection, Dst).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

An additional technique — strictly speaking a remote sensing technique — has proved useful for establishing quantitative calibrations of the EUV images, for detection of relatively fine-scale features and for sensing plasma density gradients at densities below the detection limit for EUV imaging. This is the radio plasma imaging (RPI) method, using extremely long wire antennas (hundreds of metres) to broadcast and receive radio wave pulses (3 kHz–3 MHz) reflected or refracted from plasma density gradients. This technique is closely related to ground-based ionospheric sounding observations and indeed represents a generalized topside-sounder method useful for sensing the plasmasphere and any other suitably oriented plasma density gradient structures in the magnetosphere.

Observations

Since its launch in late March 2000, IMAGE has provided a large quantity of new imaging data for near-Earth space, including magnetospheric dynamics as driven by solar storm events and some observations of the heliospheric environment including interstellar gas (see INTERSTELLAR MATTER) and remnant dust.

The ENAglow of the Earth contains many expected features and several new features that are still being interpreted and modeled. Figure 1 shows an example of the development of a relatively symmetric ENA glow encircling the Earth, representative of the corresponding cloud of energetic ions. Initially, the source of new particle injection fills the magnetosphere from the antisunward side by means of a relatively strong sunward plasma flow. Hot plasma particles that were injected then find themselves on stably trapped drift orbits that similarly encircle the Earth. With time, this hot plasma cloud decays away through a combination of processes, including losses into the atmosphere as well as the losses to charge exchange that permit these images to be recorded.

The Sun and solar wind are also sources of neutral atom glow. On its way to Earth, the solar wind passes through interstellar gas and an even smaller amount of gas emitted by dust particles in the inner solar system. As it arrives near Earth, it encounters a third gaseous charge exchange medium, the geocorona. Changes in solar wind intensity can be viewed from inside the magnetosphere as a brightening of the ENA glow from the solar wind interaction with the geocorona. Ionospheric plasma heating responds promptly to solar wind energy, causing the Earth to emit low-energy neutral atoms (LENAs), while the global storm response develops over somewhat longer time scales involving the magnetic tail. Figure 2 shows an example of LENA data illustrating simultaneous observation of LENA from the Sun, from the Earth and from the interstellar medium. Long-term variations serve as a probe of the interplanetary and interstellar gas medium.

In EUV, the plasmasphere was found to be detectable above a threshold He^+ density of about 30 cm^{-3} (where He^+ density is about 10% of total density). As shown in figure 3, this provides very useful imagery of the main body of the plasmasphere. This image, taken from a disturbed period, clearly shows the formation of a drainage plume of plasma extending from the dusk sector sunward toward noon. Such a feature was predicted over 30 yr ago, but *in situ* observations from single spacecraft were somewhat ambiguous regarding the presence of this feature and could also be interpreted as evidence for 'detached plasma regions' with no connection to the main body of the plasmasphere. A large number of other

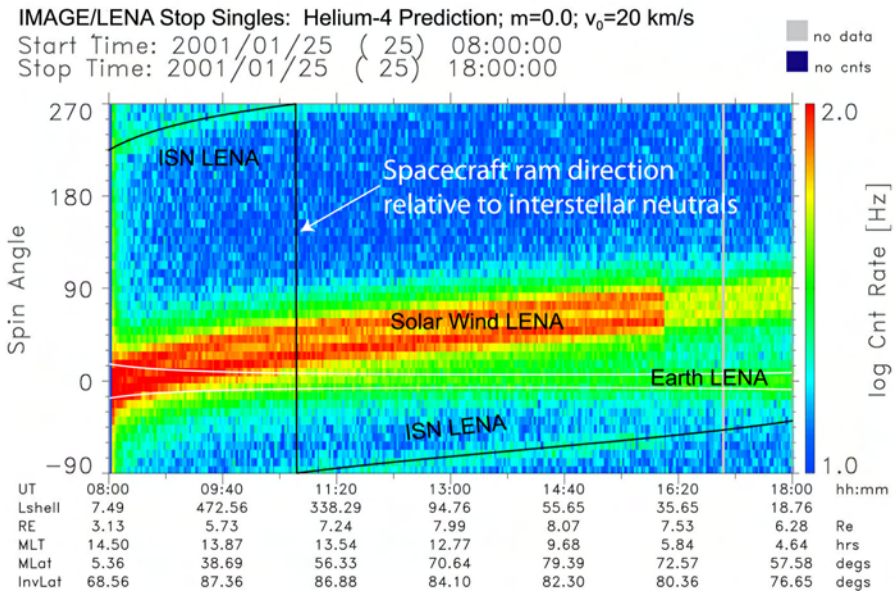


Figure 2. LENA glow from the Sun from the Earth, and from the interstellar gas, during several hours of the IMAGE orbit. Each vertical strip is formed by collapsing a 90° x 360° image in the 90° polar angle. White traces indicate the limbs of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

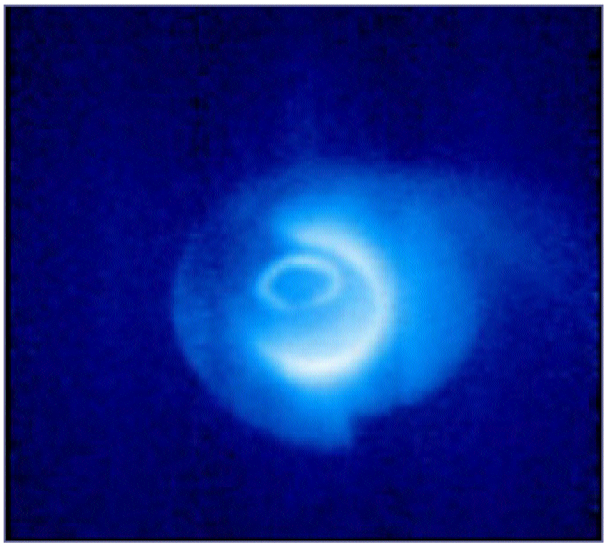


Figure 3. The EUV plasmasphere, illustrating drainage plume, shoulder and trough. The Sun's shadow provides a directional reference (toward upper left). EUV aurora and dayside ionospheric limb also appear in the image.

features have been identified in EUV imagery of the plasmasphere, including the shoulder and trough features visible in figure 3.

In FUV, IMAGE has provided valuable electron precipitation information that establishes storm context and connects with earlier auroral imaging efforts. In addition, the achievement of global Doppler-shifted proton imaging provides the first clear imagery of dayside solar wind interaction regions in which solar plasma entering the magnetospheric cusp regions is shown to be organized by interplanetary magnetic field orientation. Figure 4 illustrates the simultaneous electron / ion imaging capability and shows how different the morphologies of the two types of particle precipitation can be.

RPI obtains discrete echoes from the plasma of the topside ionosphere in the polar regions and near the plasmapause at lower latitudes. As illustrated in figure 5, echoes just inside the plasmasphere show that magnetic-field-aligned ducts exist and extend for distances of several Earth radii. Diffuse echoes from the plasmapause show that this boundary has a very irregular structure. When suitably inverted, these echo traces yield the distribution of plasma electron density for distances of 2–3 Earth radii along field lines linking the IMAGE spacecraft. These quantitative density measurements complement and serve to calibrate the EUV imagery.

Conclusions

The space environment of Earth has become an astronomical object in the sense that its component plasmas can now be remotely sensed and imaged as dynamic global structures. In another sense, the space environment contains dynamic meteorology and is an increasingly well-understood medium subject to global modeling and prediction. It can then be said that near-Earth space is the first plasma medium to be simultaneously studied using both plasma imaging and *in situ* techniques. Direct *in situ* measurements provide great detail in limited regions of space, with global structure emerging only over orbital evolution time scales. Plasma imaging produces a global perspective in the image accumulation time scale, but with modest spatial resolution. The best capabilities of the two observing techniques can now be exploited and validated against each other so as to strengthen the conclusions that can be drawn from either. The first year of operations of the IMAGE mission, in conjunction with other NASA spacecraft making *in situ* measurements, has confirmed the expected appearance of many features of near-Earth space, while revealing a number of new features that were either unexpected or previously undocumented.

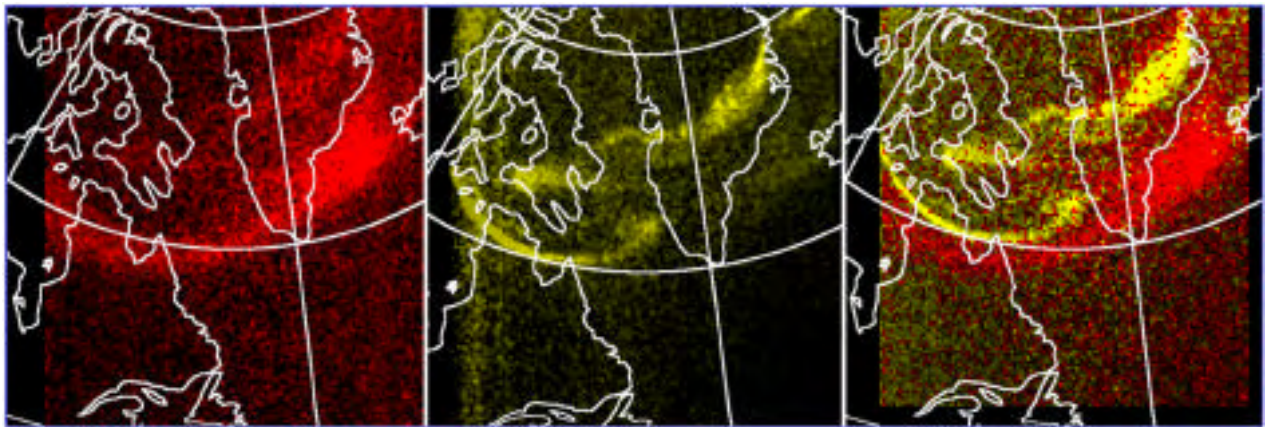


Figure 4. The electron and proton aurora in FUV, showing differences in distribution, with proton auroral emissions generally somewhat equatorward of the electron aurora.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

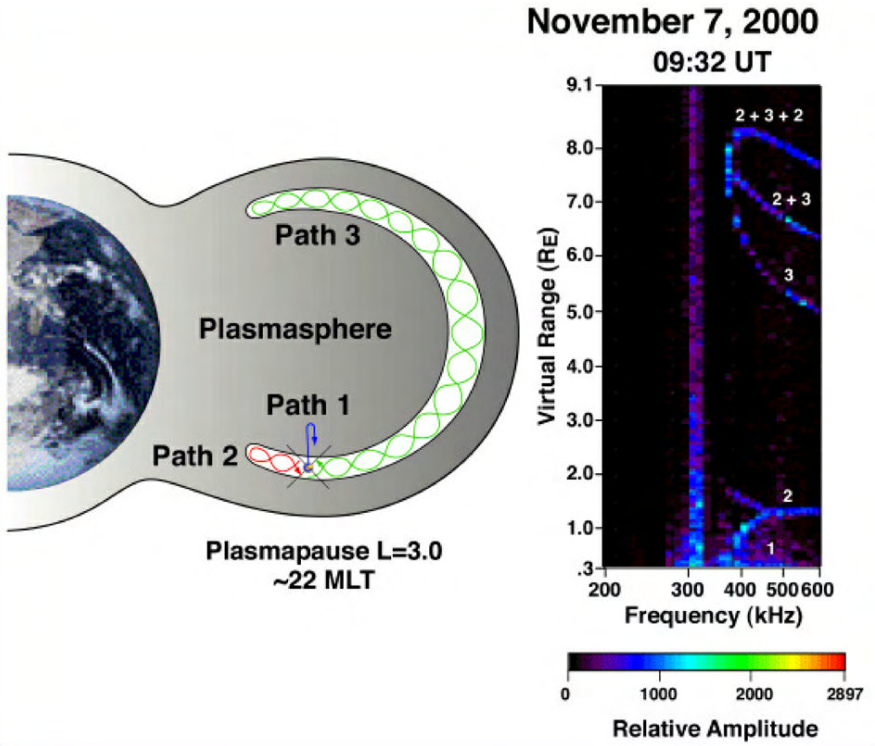


Figure 5. Radio plasma sounder traces showing the presence of ducts (magnetic flux tubes with distinct properties) in the outer plasmasphere. Traces can be interpreted to give the electron density as a function of distance along the magnetic flux tube away from the spacecraft.

Bibliography

2001 Imaging the inner magnetosphere *Geophys. Res. Lett.* **28**(6, 8)

Burch J L *et al* 2001 Views of Earth's magnetosphere with the IMAGE satellite *Science* **291** 619

Carruthers G R, Page T and Meier R R 1976 Apollo 16 Lyman alpha imagery of the hydrogen geocorona *J. Geophys. Res.* **81**(10) 1664

Holzer T E 1977 Neutral hydrogen in interplanetary space *Rev. Geophys.* **15** 467

Williams D J, Roelof E C and Mitchell D G 1982 Global magnetospheric imaging *Rev. Geophys.* **30** 183

Thomas E Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Imbrium Basin

A large multiringed impact structure on the nearside of the Moon, measuring 1123 km at its largest diameter, centered at 32.8 °N, 15.6 °W. It is visible as the lava floodplain known as the Mare Imbrium ('Sea of Showers'). Stratigraphic dating shows the Imbrium Basin to be the second youngest of the Moon's large impact features; had the impact been any greater it would most likely have shattered the Moon completely. The impact took place around 3.85 billion years ago, near the end of the period of intense cratering in the solar system known as late heavy bombardment. Ejecta from the impact covers much of the Moon's nearside. Lava flooding took place over the next 600 million years. Only the outermost ring is fairly well defined by mountain ranges, to the east by Montes Caucasus, to the southeast by Montes Apenninus, and to the south by Montes Carpatus. The Montes Alpes form part of an ill-defined inner ring. To the southwest, Mare Imbrium opens into Oceanus Procellarum. To the northwest, a more recent impact crater has formed Sinus Iridum ('Bay of Rainbows'). Two other prominent craters lie at the eastern edge: Archimedes and Plato. The basaltic lava floodplain of the mare itself is punctuated by Montes Spitzbergen and the isolated peaks Mons Pico and Mons Piton.

See also: Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inclination

The angle at which one plane is tilted to another. Taking three examples, first, the inclination of a planetary orbit is normally taken to mean the angle between the plane of the orbit and the plane of the Earth's orbit. The closest and most distant planets have the greatest orbital inclinations—Mercury's orbit is inclined at 7° and Pluto's at 17°. Second, equatorial inclination of a planet is the angle between the plane of its equator and its orbit plane. Third, the inclination of the orbit of an artificial Earth satellite is the angle at which its orbit plane intersects the plane of the equator.

See also: node, orbital elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Index Catalog (IC)

A catalog of nebulae and clusters prepared by Danish astronomer J L E Dreyer (1852–1926) as a sequel to his New General Catalog (NGC), which had been published in 1888. The Index Catalogue was published in two parts—'Index Catalog of Nebulae Found in the Years 1888 to 1894 with Notes and Corrections to the New General Catalog' (published in 1895), and 'Second Index Catalog of Nebulae Found in the Years 1895 to 1907; with Notes and Corrections to the New General Catalog and to the Index Catalog for 1888 to 1894' (published in 1908). Objects in this catalog are denoted by the letters 'IC' followed by a number, for example, 'IC 2220'.

A revised edition of the NGC and the two Index Catalogs, edited by R W Sinnott and updated to epoch 2000.0, was published as a single volume in 1988.

See also: epoch, New General Catalog of Nebulae and Clusters of Stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Indian Astronomy: History of

From the time of Āryabhaṭa (ca AD 500) there appeared in India a series of Sanskrit treatises on astronomy. Written always in verse, and normally accompanied by prose commentaries, these served to create an Indian tradition of mathematical astronomy which continued into the 18th century. There are as well texts from earlier centuries, grouped under the name *Jyotishavedāṅga*, that is, supposed to belong the latest stage of Vedic literature. These present, somewhat obscurely, a calendrical system governed by a five year cycle.

The general characteristics of the principal treatises, those written by Āryabhaṭa and his successors, seem calculated to induce a certain skepticism on the part of modern readers, who would be inclined to doubt whether any scientific content could be found in these verses. The verses are concise in the extreme, so that in general the prose commentaries are essential for any understanding. The mean motions are expressed by giving the number of revolutions in a huge period of time, typically the *mahāyuga* of 4 320 000 years. There is no tabular presentation, not even in the commentaries, in contrast to Babylonian and Greek texts, so that even the values of the trigonometrical function are given in a verse. The presentation of the mean motions is, as a rule, so closely bound up with the luni-solar calendar, itself dependent on the motions of the Sun and Moon, that it is not clear at first which is being defined. Although the texts show a clear understanding that the Earth is a sphere, and of the prime meridian, together with the need to make corrections for a change of MERIDIAN, that prime meridian is defined as a line passing through 'Lāṅka' (modern Sri Lanka, Ceylon), rather than to the actual site of observation; the commentaries, however, make it clear that Ujjain in Madhya Pradesh was that site. All these difficulties and obscurities have led Europeans to doubt whether the Indians ever observed at all. Modern studies of these texts have produced a wide range of opinions concerning the 'scientific' quality of Indian work, but recently Roger Billard has largely settled the matter by showing that the parameters were closely based on observations; see the section below on modern studies.

The bibliography of the subject is vast, but it seems to be true that the major Sanskrit texts have now been published. There are nevertheless innumerable manuscripts certainly worthy of study; see Pingree (1970 and subsequent volumes) for a survey of the manuscripts and a full bibliography of researches. In recent years important editions and reeditions have been brought out by Indian scholars, notably K S Shukla (Lucknow) and K V Sharma (Madras).

Āryabhaṭa

Āryabhaṭa may fairly be considered as the founder of Indian scientific astronomy. Unusually, the year of his birth, AD 476, is known; however the region of India in which he lived is not known. He is the author of two scientific works, one of which is extant; this has been passed down to us under the name *Āryabhaṭīya*, a name given to the material by later scholars. The other, and earlier, work may be identified as the *Sūrya Siddhānta* as described by Varāhamihira in his *Pañcasiddhāntikā* (ca AD 580). The *Āryabhaṭīya* is known from a number of manuscripts from which good editions and translations have been made; the first European edition was made by H Kern in 1873. Āryabhaṭa's works were referred to as *tantras* in his own life time, *tantra* being a general term for a treatise. Since then the tendency has been to denote the formal canons as *siddhāntas*, and the less formal, intended for easy application with simpler arithmetic, as *karaṇas*.

In the work of Āryabhaṭa one finds, at least in germinal form, a large part of the astronomical science that will develop in India over the centuries that follow, so much does it define the whole paradigm of the subject. The mathematical section is notable for a number of reasons, for an original numerical notation, for the introduction of the sine function essentially as we know it, for the value of $\pi = 3.1416$ and for the procedure of solving indeterminate equations of the form $ax - by = 1$. The astronomical parameters and procedures define completely the mean and true longitudes for the Sun, Moon and planets, as well as various procedures that would be classed as spherical astronomy. It is noteworthy that Āryabhaṭa held that the Earth revolved daily, so accounting for the daily rotation of the heavens; this view was, however, rejected by all later Indian writers.

That the work is indebted in part to Greek astronomy is evident from the use of trigonometry. However, neither the trigonometrical nor the astronomical procedures follow Ptolemaic usage closely. It may be inferred that the transmission of Greek knowledge to India included only material from pre-Ptolemaic stages, even though the transmission must have occurred much later than the time of PTOLEMY (2nd century AD).

The procedures for the computation of the longitudes of Sun, Moon and planets require that first of all the mean longitudes are calculated, always as linear functions of time. The practical epoch employed by Āryabhaṭa is noon 499 March 21 but, for theoretical purposes, he takes a moment 3600 years previous, where the 'year' is that defined by the system. In the *Sūrya Siddhānta* this comes to midnight –3101 February 17/18, while in the *Āryabhaṭīya* it comes to –3101 February 18 at 6 am, because the mean year is slightly shorter in the latter. In either case this defines the epoch of 'Kaliyuga'. All the mean motions are assumed to be zero at that epoch, and the motions are defined by a number of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

revolutions in 4 320 000 years, which is 1 577 917 800 days for the *Sūrya Siddhānta* and 1 577 917 500 days for the *Āryabhaṭīya*. Although the long period 4 320 000 years is found in speculation from much earlier centuries, and which presumably links up with Babylonian schemes, because of its obvious sexagesimal character (20×60^3) (see BABYLONIAN ASTRONOMY), the Kaliyuga as such is not ancient, but originates only with Āryabhaṭa.

An equation (or correction) is added to the mean longitude to give the true longitude. In Indian astronomy such procedures were somewhat simpler than those developed by Ptolemy. For the Sun and Moon, for example, the equation is of the form (in modern notation)

$$\mu(\theta) = \arcsin[\sin(\mu_m)\sin(\theta)]$$

where θ is angle from the apogee to the mean longitude and μ_m is a constant. For the planetary equations a second function is also needed,

$$\sigma(\theta) = \arctan\left(\frac{\sin(\sigma_m)\sin(\theta)}{1+\sin(\sigma_m)\cos(\theta)}\right).$$

These two functions (the so-called *manda* and *śīghra* corrections) were used throughout the Indian procedures for calculating the true longitude. The procedures varied from one author to another, and, indeed, Āryabhaṭa proceeds differently in his two works. In the *Sūrya Siddhānta*, which is simpler in this respect, the true longitude *L* of the outer planets depends on the mean Sun λ_s, the mean planet λ and the planet's apogee Γ; it is fixed after four steps, where α, β and γ are intermediate quantities:

$$\alpha = \Gamma - \tfrac{1}{2}\sigma(\lambda_s - \lambda)$$
$$\beta = \alpha + \tfrac{1}{2}\mu(\lambda - \alpha)$$
$$\gamma = \lambda - \mu(\lambda - \beta)$$
$$L = \gamma + \sigma(\lambda_s - \gamma).$$

The same steps are used for the inner planets, provided only that λ_s and λ are interchanged.

In the *Āryabhaṭīya* the steps differ somewhat, and moreover the 'constants' μ_m, σ_m now vary periodically by small amounts, as expressed by $\mu_1 + \mu_2|\sin\theta|$ and $\sigma_1 + \sigma_2|\sin\theta|$.

It should be noted that all the celestial longitudes are sidereal, not tropical, and are measured from a point on the ecliptic where the equinoctial point was located around the year AD 514. Precession as such plays no role in the *Āryabhatīya*, apparently a deliberate choice, although it occurs in other contemporary work. In any case, with this origin, the precessional correction in Āryabhaṭa's time was very small.

Brahmagupta

After the time of Āryabhaṭa the next figure of note is BRAHMAGUPTA, who flourished during the 7th century. He is noted for two works, the voluminous *Brāhmasphuṭasiddhānta*, and the *Khaṇḍakhādyaka*. The former is a substantial treatise on mathematics as well as astronomy, the latter a short astronomical tract, in which the parameters not only differ from those in the earlier work but agree with the *Sūrya Siddhānta* of Āryabhaṭa. It is interesting that although Āryabhaṭa is heavily criticized in the *Brāhmasphuṭasiddhānta*, the parameters given there are in poor agreement with observations compared with the work of Āryabhaṭa.

Both these works played a significant role in the transmission of Indian astronomy to the Islamic world, where they were known as *Sindhind* and *Arkand*; from the former there was derived, after some modification, the zīj (astronomical handbook) of AL-KHWĀRIZMĪ, ca AD 830 (see also ISLAMIC ASTRONOMY). It is in this period that Indian astronomy was also passed to South East Asia (Burma, Thailand, Cambodia, Indonesia), Nepal, Tibet, Central Asia and China; in South East Asia it was dependent on a source that used the epoch AD 638, an epoch still found in Burma. In fact, the analysis of a Thai calendrical treatise by Giovanni Cassini in 1691 marks the true beginning of the European interest in the subject.

The later centuries

Over the centuries, from the 10th to the 18th, many more treatises and commentaries were written. Under Moghul rule this Sanskrit tradition was deflected from its traditional course by the influence of Islamic work, for example the Persian treatise of ULUGH-BEG. By the early 18th century one curious byproduct of this influence was a Sanskrit translation of Ptolemy's *Almagest*, via a Persian version. At the same time European work became known, especially the Astronomical Tables of Philipp de la Hire (1727), transcribed into Persian as the Zīj-i Muḥammed Shāhī at the court of Jai Singh ca 1730.

Modern study

Once the obscure conventions of the Indian texts are grasped, it is not difficult to make precise calculations of the mean and true positions of the Sun, Moon and planets according to any of these treatises, so that one may compare the results with modern calculations. Comparisons have been made in this way since the early 19th century, most notably by John Bentley (1825) but also by William Dwight Whitney (1860). Bentley's efforts were handicapped both by the shortage of Sanskrit sources then known and by the fact that the parameters of 'modern' astronomy were then inadequate for the purpose. In the end his findings were held up to ridicule by indologists. In recent years, however, this approach has been pursued with singular success by Billard (1971), from whose work it is clearer than anyone might have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

imagined that the Indian parameters from the time of Āryabhaṭa are based on serious and close, often very exact, observation. Billard's *L'Astronomie indienne* marks a uniquely profound advance in our understanding of the history of Indian astronomy.

Billard studied the mean longitudes provided by the Indian canons. These are sidereal longitudes, deviating therefore from the corresponding modern values by the rate of precession. Taking all of the deviations (Sun, Moon, lunar apogee, lunar node, five planets) one finds that, for most canons, all such deviations agree at only one time, clearly the time when the parameters were fixed to agree with observation. Such a bundle of deviation curves is shown in figure 1, calculated from the *Āryabhaṭīya*. The curves in the bundle diverge widely at other times, because the rates of motion have not been made to fit observations of an earlier date and are indeed controlled to some extent by *a priori* arguments. It is clear that the actual rates of mean motions were not fixed over a long time base but were only accurately fitted to observations in Āryabhaṭa's time. While this is a defect from the astronomical point of view, it also means that the time of that agreement is very securely fixed, in one brief interval near AD 500, when the agreement with observation is excellent. When Billard's determination of the optimum year is extended to find the optimum meridian, we find that that meridian passes through central India, so reinforcing the conclusion that the observations were carried out not only in the time of Āryabhaṭa, but certainly in India, as shown in figure 2 (Mercier 1987). Thus, while his procedures clearly owe something to Greek material, we see that Āryabhaṭa's work is thoroughly Indianized and based on observations made during the lifetime of the author.

Generally speaking the Indian determinations of mean longitudes were not established over a long time base, but there is an outstanding exception in some treatises that appear in the 9th century. These are known partly through the work of Lalla, the *Śishya dhīvṛddhida tantra*. It is not that this work or its author was so outstanding, but it happens that we find here a set of modifications of the received parameters which amount to a set of mean longitudes established accurately over the period AD 500–850 and which so clearly took into account observations made at the earlier date, or if not earlier observations, at least the values of mean longitudes determined at that earlier date. Much later, Acyuta (1550–1620) of Kerala (South India), in his *Sputanirṇayana-tantra*, gave mean longitudes which also agree extremely well together over a very long time base.

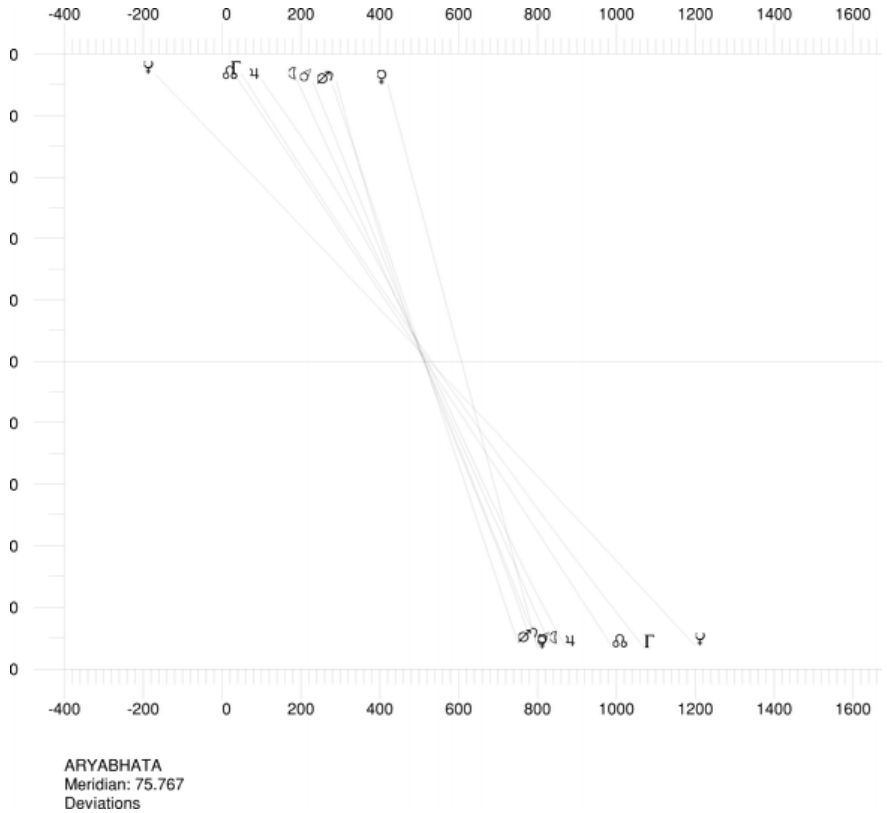


Figure 1. Deviation curves of the mean longitudes according to the *Āryabhaṭīya*. The deviation is the excess over the modern mean longitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

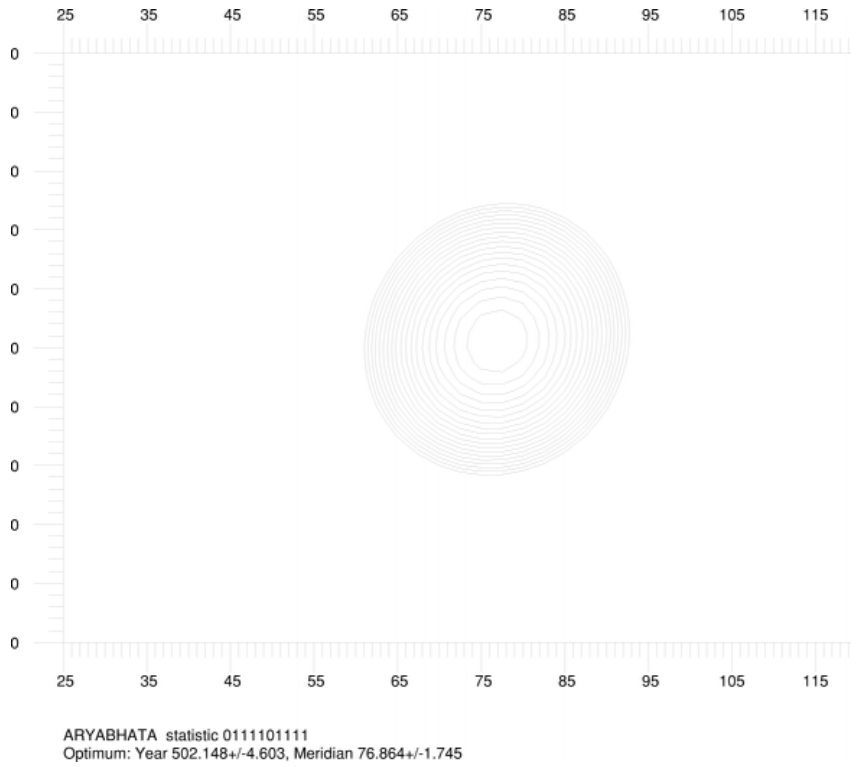


Figure 2. Contour curves showing the optimum meridian and date for the mean longitudes of the *Āryabhaṭīya* for Sun, Moon and all planets except Mercury. The optimum values are 76.86±1.76, 502.15±4.6.

Generally speaking, however, the lesson was not learnt that astronomy is natural history and requires a long time base for the determination of its parameters. There was in Indian astronomy the inbuilt assumption that a certain *a priori* framework based on the long periods, the *yugas*, could always be safely assumed, so that observations made at the time of the composition of the treatise would suffice.

Calendar

The calendar in India is luni-solar, that is it depends on the sequence of lunations, with an intercalated month every second or third year. The lunation is determined by the mean or true conjunction, but never by the visibility of the crescent, in contrast to the Babylonian or Islamic calendar (see CALENDAR IN THE MIDDLE EAST AND EUROPE).

At some time in the period following Āryabhaṭa the calendar came to be calculated according to one of the many treatises on astronomy. This had the effect of introducing certain formalities into the dating formula, in particular the 'lunar day' (*tithi*; see technical terms), in place of the ordinary civil day. In other respects the dating formula remained similar to that found in earlier periods.

There is a wealth of inscriptions in India, going back at least to the time of Aśoka (3rd century BC), and many are dated. Manuscripts also carry dates in their colophons. This means that there is much material from which to determine the application of the various astronomical procedures. Conversely, the historian calls upon expert understanding of the astronomical reckoning in order to interpret the date. Although a number of differing treatises have been used in practice to fix dates, the results rarely differ, and then by at most one day, or rather *tithi*. The era used in the date must also be established, and there are indeed many found in India. The era used in astronomy is the Śaka era, reckoned from AD 78, while the Vikrama era, from 57 BC, is widely used elsewhere.

The calendrical scheme

The calendrical scheme may be briefly sketched. The names of the months are as follows.

Caitra	Āśvina
Vaiśākha	Kārttika
Jyeshṭa	Mārgaśirsha
Āshāḍha	Pausha
Śrāvana	Māgha
Bhādrapada	Phālguna

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

These names are taken from the names of the 27 *nakshatras* (see the technical terms). Each month begins at sunrise following the moment of lunation. The Indian names for the zodiacal signs are given in the following table, together with the corresponding Western names. The Indian names are essentially translations of the Western names, so that, Mesha, for example, means 'ram'.

Aries	Mesha
Taurus	Vṛshabha
Gemini	Mithuna
Cancer	Karka
Leo	Siṃha
Virgo	Kanyā
Libra	Tulā
Scorpio	Vṛścika
Sagittarius	Dhanus
Capricorn	Makara
Aquarius	Kumbha
Pisces	Mīna

The coordination of the month and zodiacal names is so arranged that the new Moon at which Caitra begins is the last one before the Sun enters Mesha. The same monthly names are used, in addition, to specify 12 solar months. In this system, Solar Caitra begins when the Sun enters Mīna, and so on. In some inscriptions the date is given in terms of the solar months, rather than lunar months.

As in any luni-solar calendar there are intercalary months, but the Indian system is quite complex, and is such that any month might be repeated in order to keep the lunar months in step with the seasons. It may happen that the Sun does not enter a new sign from one lunation to the next, in which case the month of that name is repeated. In the same way it may enter two successive signs between two lunations, in which case the month name corresponding to the second sign is omitted. If mean lunations are used it is always the case that the interval of time between successive entries into signs is less than the mean interval between lunations, so that there are no omitted months. In either system there are always repeated months.

If the year begins in the spring the new year begins with the month Caitra, or if that month is omitted, with the month Vaiśākha. If the year begins in the autumn the new year begins with the month Kārttika, or if that month is omitted, with the month Mārgaśirsha.

Indian dates do not refer to the ordinary civil day, but to the so-called *tithi*, the lunar day. This is defined as the interval of time in which the luni-solar elongation increases by 12º. If mean lunations are used, this interval is 1/30 of the mean synodic month. The precise length of the mean *tithi* depends on the canon, but it is very nearly 0.984 of the ordinary civil day. The first *tithi* of the month begins at the moment of lunation (true or mean).

The *tithis* are counted from 1 to 15, in the first or 'bright' fortnight, starting with the new Moon, and from 1 to 15 again in the second or 'dark' fortnight, starting with the full Moon. The *tithi* in the 'bright' fortnight is marked *sudi*, an acronym for *shukla paksha dina*, and the *tithi* in the 'dark' fortnight is marked *badi*, an acronym for *bahula paksha dina*. In most cases only one sunrise occurs during the interval of the *tithi*, and that *tithi* occurs, and once only, in the sequence. However, in other cases, a *tithi* may be omitted or repeated. If it is omitted, it is because it begins after one sunrise, and ends before the next. If it is repeated, it is because it begins before one sunrise and ends after the next.

The course of the Moon is plotted against a division of the sky into 27 *nakshatras*, so-called 'lunar mansions'. These are each related to certain stars, and if their positions were strictly recognized the *nakshatras* would have varying widths. It is more common to assume a constant width of 13º 20' for each.

Almanacs

The Indian ALMANAC in the form of the *pañcāṅga* is widely available at a popular level. The almanacs are written in one of the Indian languages, Hindi, Gujarati, etc, not in English. These are popular works, intended essentially for the astrology market, which flourishes in India. Nevertheless, they are no less interesting to the historian of astronomy, continuing as they do the tradition of the medieval astronomical handbook, the zīj of Arabic–Persian tradition, itself based on the *Handy Tables* of Ptolemy. The *pañcāṅga*, as the name indicates, consists of five elements, namely *tithi*, *vāra*, *nakshatra*, *yoga*, *karaṇa* (see the technical terms). These concern only the Sun and Moon, and are the principal elements of an Indian date. Typically the almanac will go further to give the actual positions of the Sun, Moon and planets (always sidereal positions), and much besides, including a profusion of articles, many concerning astrology. It is in some ways a popular and accessible equivalent of the Astronomical Ephemeris, in contrast to the official Indian equivalent, the *Indian Ephemeris and Nautical Almanac*.

It is likely that, at the present time, most if not all of these calendars are founded on modern elements as represented in the *Indian Ephemeris*, which in turn agrees entirely with its Western counterparts, both in style and content. Indeed the positions of Sun, Moon and planets are in some almanacs not given for the Indian central meridian (82º 40' E of Greenwich) but for the Greenwich meridian.

The more ambitious of these almanacs give far more than the five basic elements and provide almost as much as the official Nautical Almanac, but in a somewhat more accessible form. These works suffer from the absence of any full acknowledgment of the methods of calculation,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or their source. However, the survey carried out for the Calendar Reform Committee (1955) included a survey of more than 50 of these almanacs, and the results showed whether traditional or modern sources of parameters had been employed, along with other technical details, including the *ayanāṃśa*.

Technical terms

It is convenient to collect here a number of the technical terms used in the Indian almanacs.

ayanāṃśa. Effectively a measure of precession, applied to the sidereal longitudes computed from the canon.
karaṇa. A sort of half-*tithi*, the interval during which the luni-solar elongation increases by 6º.
nakshatra. One of the 27 divisions of the ecliptic, named after the 27 (or 28) small constellations known collectively as the *nakshatras*.
pañcāṅga. 'Five elements', namely *tithi*, *vāra*, *nakshatra*, *yoga*, *karaṇa*.
tithi. Interval of time counted from luni-solar conjunction, representing increments of 12º in the elongation. Mean and true *tithis* are recognized, according to whether the positions of the Sun and Moon are mean or true.
yoga. The interval during which the sum of the solar and lunar positions increase by 12º.
vāra. Weekday.

Bibliography

Billard R 1971 *L'Astronomie indienne* (Paris: École Française d'Extrême Orient)
Calendar Reform Committee 1955 *Report* (New Delhi: Government of India, Council of Scientific and Industrial Research)
Indian Ephemeris and Nautical Almanac (annual) (New Delhi: Government of India)
Mercier R 1987 The meridians of reference of Indian astronomical canons *History of Oriental Astronomy: Proc. IAU Colloquium 91 (New Delhi, 13–16 November 1985)* ed G Swarup, A K Bag and K S Shukla (Cambridge: Cambridge University Press) pp 97–107
Pingree D 1970 *Census of the Exact Sciences in Sanskrit. Memoirs* vol 41 (Philadelphia: American Philosophical Society)
Sewell R and Dikshit S B 1897 *The Indian Calendar* (New Delhi: Motilal Banarsidas)
Subbarayappa B V and Sarma K V 1985 *Indian Astronomy, A Source Book* (Bombay: Nehru Centre)

Raymond Mercier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Indus

(the Indian—usually represented as a Native American; abbrev. Ind., gen. Indi; area 294 sq. deg.) A southern constellation which lies between Telescopium and Grus, and culminates at midnight in mid-August. It was first shown on Petrus Plancius' celestial globe of *c.* 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, the brightest stars in Indus are α Indi, magnitude 3.1, and β Indi, magnitude 3.7. There are no other stars brighter than fourth magnitude. ε Indi, magnitude 4.7, distance 11.8 light-years, has a proper motion of 3.96″ per year, which will bring it close to the south celestial pole in about 50 000 years. Interesting objects include θ Indi, a binary star with bluish-white (A5) and yellow (G0) components, magnitudes 4.5 and 7.2, separation 6.7″, and NGC 7049, an eleventh-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inertia

The resistance of a massive body to any change in its velocity. Any body will continue in a state of rest or of uniform straight-line motion unless acted upon by a force. Thus a spacecraft, when its rocket motors are switched off, follows a ballistic trajectory because of its inertia. In Newton's second law of motion, the relationship between acceleration, a, of a body and the applied force, F, is $F = ma$, where m is the inertial mass of the body. As far as can be determined, the inertial mass of a body is exactly equal to its gravitational mass, which determines its gravitational attraction. This is called the Equivalence Principle. There does not seem to be any *a priori* reason why these two masses should be exactly equal, and there is still considerable debate on the question of the nature of inertia.

See also: acceleration, energy, Newton's laws of motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inferior Planets

A collective term for the major planets Mercury and Venus, whose orbits lie within that of the Earth, as opposed to the superior planets, those major planets whose orbits lie outside the Earth's.

Compare: inner planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inflation

According to inflationary theory, there was a stage of exponentially rapid expansion of the early universe. The existence of this stage resolves many long-standing problems of the BIG BANG THEORY and changes our understanding of the origin of the universe and of its global structure.

Inflation and the big bang theory

Inflationary cosmology describes the very early stages of the evolution of the universe and its structure at extremely large distances from us. For many years, cosmologists believed that the universe from the very beginning looked like an expanding ball of fire. This explosive beginning of the universe was called the big bang. In the end of the 1970s a different scenario for the universe's evolution was proposed. According to this scenario, the universe at the very early stages of its evolution came through the stage of inflation, an exponentially rapid expansion in a kind of unstable vacuum-like state (a state with large energy density, but without elementary particles). This stage could be very short, but the universe within this time became exponentially large. At the end of inflation the vacuum-like state decayed, the universe became hot and its subsequent evolution could be described by the standard big bang theory (see also UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION, UNIVERSE: THERMAL HISTORY).

The existence of the inflationary stage is necessary to resolve many difficult problems which arise when one tries to unify the modern theory of elementary particles and the standard big bang theory. The standard big bang theory asserts that the universe was born at some moment $t = 0$ about 15 billion years ago, in a state of extremely high temperature T and density ρ (cosmological singularity). It is usually assumed that the standard description of the universe in terms of space and time becomes possible at the Planck time $t \sim 10^{-43}$ s after the singularity, when its density drops below the so-called Planck density $\rho_p \sim 10^{94}$ g cm^{-3} (see UNIVERSE: DENSITY). (At greater density, quantum fluctuations of space–time are so large that an alternative description of space-time is required.) The temperature of the expanding universe gradually decreased, as the reciprocal of its size, and finally the universe evolved into the relatively cold universe where we live now. Remnants of this primordial cosmic fire still surround us in the form of the microwave background radiation with a temperature of 2.7 K.

This theory was extremely successful in explaining various features of our world. However, in combination with the modern theory of elementary particles, it predicts the existence of a large amount of superheavy stable particles carrying magnetic charge: magnetic monopoles. These objects have a typical mass 10^{16} times that of the proton. According to the standard big bang theory, monopoles should appear at the very early stages of the evolution of the universe, and they should now be as abundant as protons. In that case the mean density of matter in the universe would be about 15 orders of magnitude higher than its present value of about 10^{-29} g cm^{-3}.

In addition to this problem, the standard big bang theory was unable to answer many other questions. For example, the universe is closed if $\Omega = \rho/\rho_c > 1$, it is flat if $\Omega = 1$ and it is open if $\Omega < 1$. Here ρ is the density of the universe, ρ_c is the critical density, which at the present epoch is of the order 10^{-29} g cm^{-3}. The value of the parameter Ω changes in an expanding universe. Observational data suggest that now $0.3 < \Omega < 1.2$, which does not seem unnatural. However, according to the big bang theory, this requires that at the Planck time the value of Ω should coincide with 1 with an accuracy of 10^{-60}. A slightly higher value of Ω would lead to a closed universe which would immediately collapse; a slightly smaller value of Ω would lead to an open universe, which at present would be practically empty. The need for this incredible fine tuning of initial conditions in the standard big bang theory is the essence of the flatness problem.

One may also wonder, where did the universe come from? How could it happen that its different, causally disconnected parts could begin their expansion simultaneously? Why it is so big, homogeneous and isotropic? If it must be homogeneous for whatever reason, then what is the origin of galaxies? Why is our space three-dimensional? Is it possible that our universe could be created differently?

All these problems are extremely difficult, and some of them sound almost metaphysical. That is why it is so encouraging that most of the problems mentioned above can be resolved in the context of one simple scenario of the evolution of the universe—the inflationary universe scenario.

New theories of elementary particles and the notion of a scalar field

Inflationary theory became possible owing to the rapid progress of the theory of elementary particles during the last two decades, after physicists have found a way to unify weak, strong and electromagnetic interactions (see STANDARD MODEL OF PARTICLE PHYSICS).

It is well known that electrically charged particles interact with each other by creating an electromagnetic field around them. Small excitations of this field are called photons. Photons do not have any mass, which is the main reason why electrically charged particles can easily interact with each other at a very large distance. Weak and strong interactions are mediated by similar particles. For example, weak interactions are mediated by particles called W and Z. However, whereas photons are massless particles, the particles W and Z are extremely heavy; it is very difficult to produce them. That is why weak interactions are so weak. In order to obtain a unified description of weak and electromagnetic interactions despite the obvious difference in properties of photons and the W and Z particles, physicists introduced scalar fields ϕ, which will play the central role in our discussion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The closest analog of a scalar field is the electrostatic potential Φ. Electric and magnetic fields appear only if this potential is inhomogeneous or if it changes in time. If the whole universe were to have the same electrostatic potential, then nobody would notice it; it would be just another vacuum state. Similarly, a constant scalar field ϕ looks like a vacuum state; we do not see it even if we are surrounded by it.

The main difference is that the constant electrostatic potential Φ does not have its own energy, whereas the scalar field ϕ may have potential energy density $V(\phi)$. If $V(\phi)$ has one minimum at $\phi = \phi_0$, then the whole universe eventually becomes filled by the field ϕ_0. This field is invisible, but, if it interacts with the W and Z particles, they become heavy. Meanwhile, if photons do not interact with the scalar field, they remain light. Therefore we may begin with a theory in which all particles initially are light and there is no fundamental difference between weak and electromagnetic interactions. This difference appears later, when the universe becomes filled by the scalar field ϕ. At this moment the symmetry between different types of fundamental interactions becomes broken. This is the basic idea of all unified theories of weak, strong and electromagnetic interactions. A scalar field ϕ of a similar type is the main ingredient of inflationary theory.

Development of inflationary theory

According to the big bang theory, the rate of expansion of the universe given by the HUBBLE 'CONSTANT' $H(t)$ is (approximately) proportional to the square root of its density. If the universe is filled by ordinary matter, then its density rapidly decreases as the universe expands. Therefore expansion of the universe rapidly slows down as its density decreases. This rapid decrease of the rate of the universe's expansion is the main reason for all our problems with the standard big bang theory. However, because of the equivalence of mass and energy established by Einstein ($E = mc^2$), the potential energy density $V(\phi)$ of the scalar field ϕ also contributes to the rate of expansion of the universe. In certain cases the energy density $V(\phi)$ decreases much more slowly than the density of ordinary matter. This may lead to the existence of a stage of extremely rapid expansion (inflation) of the universe.

The first version of inflationary theory was proposed in 1979 by Alexei Starobinsky of the Landau Institute, Moscow. Starobinsky's model created a sensation among Russian astrophysicists, and for two years it remained the main topic of discussion at all conferences on cosmology in the Soviet Union. However, his model was rather complicated, being based on the theory of anomalies in QUANTUM GRAVITY; it was not well known outside of Russia.

The crucial step in the development of inflationary theory was made in 1981 by Alan Guth of MIT. He suggested that the hot universe at some intermediate stage could expand exponentially, being in an unstable supercooled state called the 'false vacuum'. He called this stage of exponential expansion 'inflation'. The main idea of his model was based on the theory of cosmological phase transitions with supercooling, which was developed in the mid-1970s by David Kirzhnits and Andrei Linde at the Lebedev Institute, Moscow. This scenario was so attractive that even now many people associate inflation with expansion in a false vacuum. Unfortunately, as Guth has found himself, the universe after inflation in his scenario becomes extremely inhomogeneous. After a year of investigation, Guth, in his paper with Eric Weinberg of Columbia University, concluded that his model could not be improved.

In 1982 Linde introduced the so-called new inflationary universe scenario, which later was also discovered by Andreas Albrecht and Paul Steinhardt at the University of Pennsylvania. Inflation in this scenario occurs not only in the false vacuum $\phi = 0$ but also when the scalar field ϕ begins slowly moving towards the minimum of $V(\phi)$. This scenario was free of the main problems of the model suggested by Guth, but it was rather complicated. It assumed that the universe from the very beginning was extremely large and hot, that it was in a state of thermal equilibrium although the field ϕ was practically decoupled from all other particles. It also required the potential $V(\phi)$ to have a very specific, fine-tuned form. No realistic versions of this scenario have been proposed so far.

Finally, in 1983 Linde has found that inflation can naturally occur in many theories of elementary particles. There is no need for quantum gravity effects, phase transitions and supercooling. There is no need to assume that the whole universe was created at the same moment, that it was in a state of thermal equilibrium and that the scalar field from the very beginning was at the minimum of its potential energy density. One should consider all possible chaotic distributions of the scalar field in the early universe, and then check whether some of them lead to inflation. Those places where inflation does not occur remain small, whereas those domains where inflation takes place become exponentially large and give the main contribution to the total volume of the universe. This scenario was called 'chaotic inflation'. Its simplest version was described in the previous section. In the context of this scenario inflation is not an exotic phenomenon invoked by theorists for solving their problems. It is a very general regime which occurs in a wide class of theories including many models with polynomial and exponential potentials $V(\phi)$.

Chaotic inflation

To explain the basic idea of chaotic inflation, let us consider the simplest model of a scalar field ϕ with a mass m and with the potential energy density $V(\phi) = (m^2/2)\phi^2$ (figure 1). Since this function has a minimum at $\phi = 0$, one may expect that the scalar field ϕ should oscillate near this minimum. This is indeed the case if the universe does not expand. However, one can show that in a rapidly expanding universe the scalar field moves down very slowly, as a ball in a viscous liquid, viscosity being proportional to the speed of expansion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

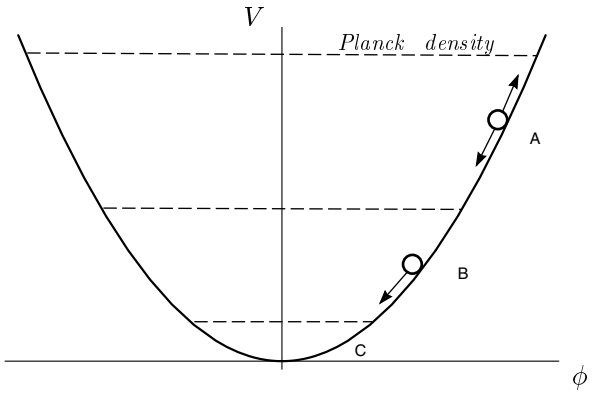


Figure 1. Motion of the scalar field in the theory with $V(\phi) = (m^2/2)\phi^2$. Several different regimes are possible, depending on the value of the field ϕ. If the potential energy density of the field is greater than the Planck density 10^{94} g cm^{-3}, quantum fluctuations of space–time are so strong that one cannot describe it in usual terms. Such a state is often called space-time foam. At a somewhat smaller energy density (region A) quantum fluctuations of space–time are small, but quantum fluctuations of the scalar field ϕ may be large. Jumps of the scalar field due to quantum fluctuations lead to a process of eternal self-reproduction of inflationary universe which we are going to discuss later. At even smaller values of $V(\phi)$ (region B) fluctuations of the field ϕ are small; it slowly moves down as a ball in a viscous liquid. Inflation occurs in both the region A and region B. Finally, near the minimum of $V(\phi)$ (region C) the scalar field rapidly oscillates, creates pairs of elementary particles, and the universe becomes hot.

There are two equations which describe the evolution of a homogeneous scalar field in our model:

$$H^2 + \frac{k}{a^2} = \frac{4\pi G}{3}(\dot\phi^2 + m^2\phi^2)$$
$$\ddot\phi + 3H\dot\phi = -m^2\phi.$$

Here $H = \dot a/a$ is the Hubble parameter in the universe with a scale factor $a(t)$, $k = -1, 0, 1$ for an open, flat or closed universe respectively, G is the gravitational constant. The first of these two equations is the Einstein equation for a homogeneous universe containing a scalar field ϕ with energy density $\rho = \dot\phi^2/2 + m^2\phi^2/2$. The second equation is the standard Klein–Gordon equation for a homogeneous scalar field. The only difference (but a very important one) is the presence of the term $3H\dot\phi$, which appears owing to the expansion of the universe. This equation is similar to the equation of motion for a harmonic oscillator, where instead of $x(t)$ we have $\phi(t)$, with a friction term $3H\dot\phi$.

If the scalar field ϕ initially was large, the Hubble parameter H was large too, according to the first equation. This means that the friction term $3H\dot\phi$ in the second equation was very large, and therefore the scalar field was moving very slowly, as a ball in a viscous liquid. Therefore at this stage the energy density of the scalar field, unlike the density of ordinary matter, remained almost constant, and expansion of the universe continued with a much greater speed than in the old cosmological theory. Because of the rapid growth of the scale of the universe and the slow motion of the field ϕ, soon after the beginning of this regime one has $\ddot\phi \ll 3H\dot\phi$, $H^2 \gg k/a^2$, $\dot\phi^2 \ll m^2\phi^2$, so the system of equations can be simplified:

$$\frac{\dot a}{a} = 2m\phi\left(\frac{\pi G}{3}\right)^{1/2} \qquad 3\frac{\dot a}{a}\dot\phi = -m^2\phi. \tag{1}$$

The first equation shows that if the field ϕ changes very slowly, then the size of the universe grows approximately as e^{Ht}, where $H = (8\pi G V(\phi)/3)^{1/2} = 2m\phi(\pi G/3)^{1/2}$.

This stage of exponentially rapid expansion of the universe is called inflation. In realistic versions of inflationary theory its duration could be as short as 10^{-35} s. When the field ϕ becomes sufficiently small, viscosity becomes small, inflation ends and the scalar field ϕ begins to oscillate near the minimum of $V(\phi)$. As any rapidly oscillating classical field, it loses its energy by creating pairs of elementary particles. These particles interact with each other and come to a state of thermal equilibrium with some temperature T. From this time on, the corresponding part of the universe can be described by the standard hot universe theory.

The main difference between inflationary theory and the old cosmology becomes clear when one calculates the size of a typical inflationary domain at the end of inflation. Investigation of this question shows that even if the size of inflationary universe at the beginning of inflation in our model was as small as the Plank size $l_{\rm P} \sim 10^{-33}$ cm, after 10^{-35} s of inflation the universe acquires a huge size of $l \sim 10^{10^{12}}$ cm. These numbers are model dependent, but in all realistic models this size appears to be many orders of magnitude greater than the size of the part of the universe which we can see now, $l \sim 10^{28}$ cm. This immediately solves most of the problems of the old cosmological theory.

Our universe is so homogeneous because all inhomogeneities were stretched $10^{10^{12}}$ times. The density of primordial monopoles and other undesirable 'defects' becomes exponentially diluted by inflation. The universe becomes enormously large. Even if it was a closed universe of a size $\sim 10^{-33}$ cm, after inflation the distance between its 'south' and 'north' poles becomes many orders of magnitude greater than 10^{28} cm. We see only a tiny part of the huge cosmic balloon. That is why the universe looks so flat. That is why we do not need expansion of the universe to begin simultaneously in many different causally disconnected domains of a Planck size.

If one considers a universe which initially consisted of many domains with chaotically distributed scalar field ϕ (or if one considers different universes with different values of the field), then domains in which the scalar field was too small never inflated. The main contribution to the total volume of the universe will be given by those domains which originally contained a large scalar field ϕ. Inflation of such domains creates huge homogeneous islands out of initial chaos. Each inflationary domain in this scenario is much greater than the size of the observable

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

part of the universe. The properties of space inside each such domain after inflation practically do not depend on initial conditions at the beginning of inflation.

Quantum fluctuations as a source for galaxy formation

Solving many difficult cosmological problems simultaneously by a rapid stretching of the universe may seem too good to be true. Indeed, if all inhomogeneities were stretched away, there will be no inhomogeneities left, and these are necessary for the creation of galaxies (see GALAXY FORMATION). Fortunately, while removing previously existing inhomogeneities, inflation at the same time created new ones. The basic mechanism can be understood as follows.

According to quantum field theory, empty space is not entirely empty. It is filled with quantum fluctuations of all types of physical fields. These fluctuations can be regarded as waves of physical fields with all possible wavelengths. If the values of these fields, averaged over some macroscopically large time, vanish, then the space filled with these fields seems to us empty and can be called the vacuum.

In the inflationary universe the vacuum structure is more complicated. Those waves which have very short wavelength 'do not know' that the universe is curved; they move in all directions with a speed approaching the speed of light. However, inflation very rapidly stretches these waves. Once their wavelengths become sufficiently large, these waves begin feeling that the universe is curved. At this moment they stop moving because of the effective viscosity of the expanding universe with respect to the scalar field.

The first quantum fluctuations to freeze are those with large wavelengths. The amplitude of the frozen waves does not change later, whereas their wavelengths grow exponentially. In the course of expansion of the universe newer and newer fluctuations become stretched and freeze on the top of each other. At that stage one cannot call these waves 'quantum fluctuations' anymore. Most of them have exponentially large wavelengths. These waves do not move and do not disappear on being averaged over large periods of time. What we obtain is an inhomogeneous distribution of the classical scalar field ϕ which does not oscillate. Perturbations of the scalar field create Gaussian adiabatic perturbations of the metric (density perturbations). It is these perturbations that are responsible for the subsequent appearance of galaxies and formation of the large-scale structure of the universe.

Note that the amplitude of inflationary perturbations depends on the Hubble parameter H, which changes only very slowly during inflation. At the same time, the wavelength of the perturbations grows exponentially. As a result, the amplitude of the density perturbations is nearly scale independent; they have flat spectrum. The fact that inflation produces adiabatic Gaussian perturbations with a nearly flat spectrum practically does not depend on the choice of the inflationary potential $V(\phi)$, even though the amplitude of the perturbations does depend on $V(\phi)$. In addition to adiabatic perturbations, inflation produces gravitational waves, which also have a flat spectrum.

Eternal inflation

The possibility that galaxies appear as a result of quantum fluctuations may seem very unusual. However, a further pursuit of this idea has led to an even less expected conclusion: quantum fluctuations may lead to an eternal process of self-reproduction of the universe. This effect is rather general, but it looks especially surprising and leads to most dramatic consequences in the context of the chaotic inflation scenario.

As we already mentioned, one can visualize quantum fluctuations of the scalar field in inflationary universe as waves, which move in all possible directions and then freeze on the top of each other. Each freezing wave slightly increased the value of the scalar field in some parts of the universe and slightly decreased this field in other parts of the universe.

Consider those rare parts of the universe where these freezing waves always increased the value of the scalar field ϕ, persistently pushing the scalar field uphill, to the larger values of its potential energy $V(\phi)$. This is an unusual and obviously improbable regime. Indeed, the probability that the field ϕ will make one jump up (instead of jump down) is equal to $1/2$ and the probability that the next time it also jumps up is also $1/2$, so that the probability that the field ϕ without any special reason will make N consecutive jumps in the same direction is extremely small; it will be proportional to $1/2^N$.

Normally one neglects such fluctuations. However, in inflationary cosmology they can be extremely important. Indeed, those rare domains of the universe where the field jumps high enough begin exponentially expanding with ever-increasing speed. Remember that the inflationary universe expands as e^{Ht}, where the Hubble constant is proportional to the square root of the energy density of the field ϕ. In the simple model with $V(\phi) = (m^2/2)\phi^2$ the Hubble constant H will be simply proportional to ϕ. Thus, the higher the field ϕ jumps, the faster the universe expands. Very soon those rare domains, where the field ϕ persistently climbs up the wall, will acquire a much greater volume than those domains which keep sliding to the minimum of $V(\phi)$ in accordance with the laws of classical physics.

This implies that, if the universe contains at least one inflationary domain of a sufficiently large size, it begins unceasingly producing newer and new inflationary domains. Inflation at each particular point may stop very quickly, but there will be many other places which will continue expanding exponentially. The total volume of all inflationary domains will grow without end. This picture differs considerably from the standard big bang cosmology which assumed that our universe is similar to one huge homogeneous expanding ball. According to the new theory, our universe is not a ball but a fractal,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which looks like a tree consisting of an infinite number of inflationary balloons producing new balloons.

In this scenario the universe as a whole is immortal. Each particular part of the universe may appear from a singularity somewhere in the past, and it may end up in a singularity somewhere in the future. However, there is no end to the evolution of the whole universe. The situation with the beginning is less certain. It is most probable that each part of the inflationary universe has originated from some singularity in the past. However, at present we do not have any proof that all parts of the universe were created simultaneously in a general cosmological singularity, before which there was no space and time at all. Moreover, the total number of inflationary bubbles on our 'cosmic tree' exponentially grows in time. Therefore most of the bubbles (including our own part of the universe) grow indefinitely far away from the root of this tree. This does not necessarily eliminate the big bang, but it removes the possible beginning of the whole universe to the indefinite past.

Until now we have considered the simplest inflationary model with only one scalar field ϕ. However, in realistic models of elementary particles there are many different scalar fields. For example, in the unified theories of weak and electromagnetic interactions there exist at least two other scalar fields, Φ and H. In some versions of these theories the potential energy density of these fields has about a dozen different minima of the same depth. During inflation these fields, just as the field ϕ, jump in all possible directions owing to quantum fluctuations. After inflation they fall down to different minima of their energy density in different exponentially large parts of the universe. Scalar fields change properties of elementary particles and the laws of their interaction. This means that after inflation the universe becomes divided into exponentially large domains with different laws of low-energy physics.

Note that, if fluctuations are not too strong, scalar fields cannot jump from one minimum of their energy density to another. In this case the new parts of the inflationary universe remember the 'genetic code' of their parents. However, if fluctuations are sufficiently large, mutations occur. In some inflationary models quantum fluctuations become so strong that even the effective number of dimensions of space and time can change. According to these models, we find ourselves inside a four-dimensional domain with our kind of physical laws not because domains with different dimensionality and with different particle properties are impossible or improbable, but simply because our kind of life cannot exist in other domains.

Inflation, observational data and particle physics

In addition to explaining many features of our world, inflationary theory makes several important predictions. First of all, most of the inflationary models predict that the universe should be extremely flat, $\Omega = 1 \pm 10^{-4}$ [1]. It is possible to have inflationary models with $\Omega < 1$ (open inflation). So far all models of this type are much more complicated than the models which lead to $\Omega = 1$.

Also, as we already mentioned, inflation generates Gaussian adiabatic perturbations of metric with a nearly scale-independent (flat) spectrum. This is one of the most important and robust predictions of inflationary cosmology that can be experimentally tested. We are unaware of any non-inflationary mechanism that would produce such perturbations.

Flatness of the universe can be experimentally verified, since the density of a flat universe is related in a very simple way to the speed of its expansion given by the Hubble constant. So far observational data are consistent with the simplest possibility $\Omega = 1$. If it happens that the total energy density is smaller than ρ_c, so that the universe is open, it will be necessary to make substantial adjustments to inflationary theory and consider models of open inflation.

Properties of density perturbations produced during inflation can be tested not only by looking at the distribution of matter in the universe. These perturbations make their imprint on the COSMIC MICROWAVE BACKGROUND radiation. They make the temperature of this radiation slightly different in different places in the sky. This is exactly what was found by the Cosmic Background Explorer (COBE) and later confirmed by several other groups of observers.

A new generation of satellites measuring the anisotropy of the microwave background radiation will bring us much more detailed information about the spectrum of perturbations of the metric in the early universe. This information, combined with independent investigation of Ω by other methods, will make it possible to perform a thorough test of various versions of inflationary theory. It may rule out many particular versions of inflationary theory. However, even though each particular version of inflationary theory can be tested, it may be difficult or even impossible to rule out the main idea of inflationary cosmology. For example, as we already mentioned, if observations show that the universe is not flat but open, we will need to consider more complicated versions of inflationary theory. However, it is important that these versions do exist. Meanwhile, at present we do not have any non-inflationary theory of an open universe which would explain why the universe is so homogeneous and predict the small inhomogeneities ('ripples in space') discovered by COBE.

It is too early to claim that the recent observational data have confirmed inflationary theory, but it is certainly true that these results already at their level of accuracy could definitively disprove most of the inflationary models, and this did not happen. Recent observational data give certain evidence in favor of the simplest

[1] Note that, from the point of view of the geometry of the universe, the flatness condition $\Omega = 1$ takes into account the contributions to ρ from all sources, including vacuum energy, or cosmological constant. However, astronomers often exclude the contribution to Ω from the vacuum energy, which sometimes leads to terminological misunderstandings.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

versions of inflationary theory predicting a flat universe with a flat spectrum of Gaussian adiabatic perturbations. Meanwhile, during the last few years the observational data practically ruled out the only alternative mechanism of formation of the large-scale structure of the universe, based on the theory of topological defects. Note also that the theory of topological defects did not address any other problems of the standard big bang theory that have been solved by inflation. Thus, although we still do not know whether inflationary theory is true, perhaps it is fair to say that it is the only internally consistent cosmological theory that is at present available.

Inflationary theory is changing together with the rapid development of elementary particle theory. The greatest challenge is to obtain a natural realization of the inflationary universe scenario in the context of supergravity, string theory or M-theory. This appears to be a very difficult task. In order to accomplish it, many different inflationary models have been suggested, some of which are based on investigation of more complicated theories with several different scalar fields. The list of new models includes 'extended inflation', 'natural inflation', 'topological inflation,' 'hybrid inflation,' etc. Most of these models are based on the same physical principles as the simplest chaotic inflation model discussed above, but each of them has its own unique features which can be experimentally tested.

Conclusion

For many centuries scientists believed that the universe is static. Later, they learned that the universe expands. This was a psychological shock, which provoked many attempts to restore stationarity either by modifying the Einstein equations or by invoking permanent matter creation in a steady-state universe. Eventually all these attempts failed. Discovery of the cosmic microwave background radiation and successful prediction of abundances of hydrogen, helium and other elements in the universe confirmed the big bang theory. Finally, the inflationary scenario appeared, which solved many of the intrinsic problems of the big bang cosmology and apparently removed the last doubts concerning its validity.

However, the development of the inflationary theory is gradually giving rise to a completely new cosmological paradigm, which is considerably different from the old big bang theory and even from the first versions of inflationary cosmology. The universe appears to be both chaotic and homogeneous, expanding and stationary. It may consist of many exponentially large domains with different laws of low-energy physics and perhaps even with different dimensionality. Our cosmic home grows, fluctuates and eternally reproduces itself in all its possible forms, as if adjusting itself for all possible types of life which it can support. Some parts of the new theory, it is hoped, will stay with us for years to come, some others may be considerably modified in the nearby future. However, it seems that the last 20 yr of development of cosmology have already changed in an irreversible way our understanding of the structure and the fate of our universe and of our own place in the world.

Bibliography

Kolb E W and Turner M S 1990 *The Early Universe* (New York: Addison-Wesley)

Liddle A R and Lyth D H 1993 The cold dark matter density perturbations *Phys. Rep.* **231** 1-105

Linde A D 1990 *Particle Physics and Inflationary Cosmology* (Chur: Harwood Academic)

Linde A D 1994 The self-reproducing inflationary universe *Sci. Am.* **271** (5) 48–55

Andrei Linde

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inflationary Universe

A modification to the standard Big Bang model that incorporates a period of exponential (accelerating) expansion, called 'inflation' into the early evolution of the universe. In a typical inflationary scenario, the exponential expansion would have started about 10^{-34} s after the beginning of time and would have ended when the universe was, say, a hundred times older (i.e. after about 10^{-32} s). During this time interval (the 'inflationary epoch'), all distances in the universe would have increased by a factor of about 10^{50}. An inflationary factor of 10^{50} corresponds to a region of space 10^{-34} light-seconds in radius and would expand to a radius of about 10^{16} light-seconds (about 300 million light years). Although this appears to imply expansion 'faster than light', because inflation involves the expansion of space, rather than the motion of particles through space, it does not contradict Einstein's assertion that nothing can exceed the velocity of light.

According to grand unified theories of elementary particles and forces, inflation was driven by a change that occurred, about 10^{-34} s after the beginning of time, when the grand unified force separated into the strong and electroweak forces and immense quantities of energy were released that previously had been stored in the vacuum of spacetime. This 'phase change' in the state of the universe has been likened to the change that occurs when water turns to ice and heat ('latent heat') is released.

The inflationary hypothesis was introduced in 1981 by American mathematical physicist Alan Guth, in order to tackle a number of outstanding issues in cosmology, including the 'flatness' problem and the 'horizon problem'. The flatness problem arises from the fact that observations indicate that the mean density of the universe is within a factor of ten of the critical density, the density of a so-called 'flat' universe (one in which the net overall curvature of space is zero). Standard Big Bang models imply that had the mean density differed even microscopically from the critical value in the first instants of the history of the universe, it would differ from the critical value by a huge factor now. The horizon problem relates to the question of how the universe on the large scale came to be so homogeneous and isotropic (the same everywhere and in all directions). For example, when we look at the microwave background in opposite regions of the sky, that radiation has been traveling through space for virtually the entire lifetime of the universe (about 15 billion years). Because the diameter of the observable universe is twice as great as this (about 30 billion light years), there would not have been enough time, in standard Big Bang models, for these widely separated regions to have had any kind of contact with each other (no signal, traveling at the speed of light, could have traveled from one to the other). How, then, could these regions be identical?

By blowing up the universe by such a huge factor, inflation ensures that the curvature of space becomes indistinguishably close to zero (the 'flat' case). If the inflationary model is correct, the observable universe is a tiny portion of the whole, and space appears flat in much the same way as the surface of the planet Earth appears flat if we examine only a tiny portion of it. The huge inflationary factor also offers a solution to the horizon problem. Prior to the epoch of inflation, there would have been time for a sufficiently small volume of space (smaller, say, than 10^{-34} light-seconds across) to iron out any internal differences and become uniform throughout. By blowing up this microscopic volume to become larger than the observable universe, inflation would ensure that the entire observable universe, on the large scale, looks exactly the same.

Inflation also resolves a number of problems relating to particle physics and, by inflating tiny quantum fluctuations to enormous scales, provides a possible mechanism for generating the 'seeds' from which galaxies and larger-scale structures subsequently developed.

See also: Big Bang theory, cosmological constant, cosmological model, grand unified theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Information Handling in Astronomy

In a science such as astronomy, information handling encompasses data collection, analysis and dissemination, as well as the way astronomers publish, interact and communicate, including with other communities, with amateur astronomers and with the public at large.

The main aim of astronomers is to contribute to a better understanding of the universe (as well as of its past and future) and consequently to a better comprehension of the place and role of humans in it.

To this end, together with theoretical investigations, they carry out observations to obtain data that subsequently undergo treatment and studies leading to the publication of results. The whole procedure can include several iterations between the various steps as well as with external fields, non-scientific disciplines, instrumental technologies and information handling methodologies.

In the following, the concept of information will cover the observational material, the more or less reduced data extracted from it, the scientific results and the accessory material used by scientists in their work (bibliographical resources, yellow-page services, and so on), as well as the communications and publications of all kinds.

It should also be noted that, contrary to other scientific disciplines, astronomy has the peculiarity of not being able to interact with the objects it investigates (except for very few bodies of the solar system where *in situ* studies can be carried out by spacecraft). Consequently all our knowledge of the universe has so far depended almost exclusively on the photons reaching us from outer space.

Astronomy has thus to rely on the ingenuity of instrumentalists to conceive and design a whole range of observing tools exploiting at best the latest technologies and the most sensitive detectors to obtain the most relevant and most varied information allowing progress of astronomical knowledge. The current trend is also towards panchromatic astronomy, i.e. combining information from the various wavelength ranges of the ELECTROMAGNETIC SPECTRUM (radio, infrared, visible, ultraviolet, x, γ, …), instead of restricting oneself to specific ranges as used to be the case in the past.

Compared with the past too, ever larger amounts of data are being collected, and the rate will probably continue to accelerate. Instruments such as NASA/ESA's Hubble Space Telescope (HST) or ESO's Very Large Telescope (VLT) are generating or will generate annually a volume of information of the order of terabytes of data. The rate of increase of observations is also matched by the diversification.

Until not so long ago, an astronomer could also work individually from the conception of a project through to the collection and analysis of data. Nowadays, as instrumentation has become more complex, teams of researchers, most often international ones, have become necessary and they quite naturally include technologists or instrumentalists.

At the other end of the chain, the teams are more and more including methodologists, i.e. specialists of information handling. If image processing is a natural consequence of a sophisticated technology, information handling rather qualifies whatever happens to already well-reduced data.

Astronomy and related space sciences have always been at the leading edge of new information technologies, often testing, contributing to and pushing for new developments, be it for the access to databases, the usage of networks or the initial explosion of world-wide web (WWW) sites, as well as for remote observing (spacecraft and ground-based telescopes remotely operated).

The recent dramatic evolution of communication and information technologies had a deep impact on the way the astronomy community interacts and works, in other words on its dynamics, and one can certainly expect more evolution in the future as the corresponding technologies will bring in new potentialities.

Information handling in astronomy thus reflects the way astronomers work and ensures progress of the astronomical knowledge which is then shared with colleague scientists as well as, on a less specialized level, with the significant community of amateur astronomers round the world (also a phenomenon proper to astronomy; see AMATEUR–PROFESSIONAL COLLABORATION IN ASTRONOMY) and with the public at large since cosmic perceptions have always been a fundamental component of human culture and philosophy. Astronomers are of course also deeply involved in EDUCATION at all levels.

As described in the following and as illustrated by figure 1, the information flow in astronomy is far from being a simple linear one. More and more of the processes and corresponding exchanges are performed electronically.

Observing and collecting data

Carrying out professional astronomical observations implies having access to ground-based and space-borne instruments which is a highly, sometimes fiercely, competitive process. Candidates have to submit proposals requesting instrument or spacecraft time with specifications for secondary instrumentation, observing modes and configurations.

The cases must be made not only on the basis of the technical capabilities requested but also with strong

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

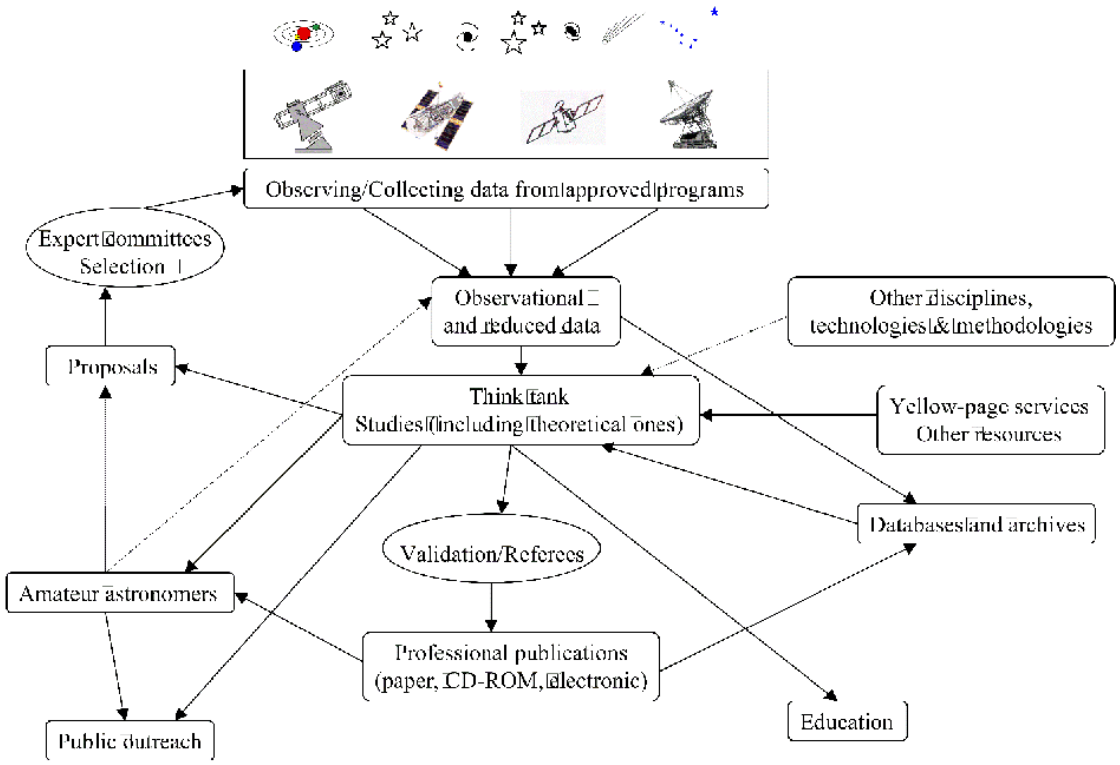


Figure 1. A schematic view of the information flow in astronomy.

references to the scientific achievements aimed at in the light, not only of what has already been achieved, but also of the expertise of the proposers themselves.

Expert committees review the technical feasibility of the proposals and the relevance of the scientific cases. They refer to the available literature (a reason for publishing, see below) and to the past observations carried out with the instrument requested in order to avoid duplication of activities of sometimes highly oversubscribed equipment (hence the need for observing logs, see below).

Oversubscribing factors vary with specific instruments or spacecraft (the new ones and the most efficient ones being generally more sought for), but it is not rare that between five and ten times more observing time is requested than is available. The task of the time allocating committees is therefore not an easy one.

Often astronomers, or even entire teams of astronomers, are encouraged to collaborate with each other, thus to share observing time and the resulting data. Astronomy is by essence an international science and it is quite common to find scientists from all over the world joining efforts in the same team (hence a need for yellow-page services, see below).

Incidentally, service observing (i.e. observations carried out, not by the proposers themselves, but by a team of trained resident astronomers and technicians) is becoming increasingly common in order to maximize return of heavy investments, to optimize instrumental pointing by combining lists of objects ('targets') of several programs and to minimize effects of some adverse factors such as seeing, weather, avoidance zones, solar activity, eclipses, radiation belts and so on.

Service observing also makes it easier to deal with contingency time and with targets of opportunity (i.e. unscheduled events such as discoveries of novae, supernovae, comets and so on). In such cases, quick and efficient interactions with teams of specialists are critical for making the best of such opportunities.

Nomenclature of celestial objects

Cataloging objects properly is a fundamental issue of all astronomical observations and studies. However, when dealing with a study field as populated as the universe, to unequivocally identify the objects at hand is something that is made complicated by several factors.

First of all, the relative positions of objects in the universe—thus in the sky—are not fixed, even if those relative movements are imperceptible to the unassisted eye and often on a human time scale (except again for the solar-system objects).

Second, and all amateur photographers know that effect, the sky looks differently if observed in different wavelength ranges: the infrared sky is not quite the same as the visible one which in turn differs from the radio one, and so on. This results from the fact that celestial bodies radiate the peak of their energy in different

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

wavelength ranges (some stars are red, others are blue; some bodies are strong emitters in the infrared, others in the x range; and so on).

Finally, and maybe most importantly, it all depends at which sensitivity and resolution the sky is observed in a specific wavelength range. The more sensitive is the tool, the more crowded is the sky 'seen' in that range, and the more precise must be the corresponding identification of a given object.

It is not rare either that an object (star, nebula, galaxy,…) become resolved in several, sometimes many, components or elements when observed at a higher resolution. Hence the need for establishing some hierarchical relationships between the designations of those objects as seen in different conditions.

In an excellent environment (i.e. away from the light pollution of densely populated areas), the unassisted human eye can see about 9000 stars brighter than magnitude 6.5 for the whole sky. This corresponds roughly to the Bright Star Catalog (BSC) that provides basic astronomical information. The Henry Draper (HD) Catalog, that goes down to about magnitude 9 (remember that the larger the magnitude, the fainter the object), lists already more than 272 000 stars.

One of the first astronomical space experiments, the Belgian–UK Ultraviolet Sky Survey Telescope (S2/68) on board the TD1 satellite, led to a catalog of ultraviolet fluxes for 31 215 stars collected at the beginning of the 1970s. A decade later, the IRAS Catalog of Point Sources (alone) gathered together some 250 000 well-confirmed infrared point sources observed by the INFRARED ASTRONOMY SATELLITE.

As to the Guide Star Catalog (GSC) for the HUBBLE SPACE TELESCOPE (HST), it contains nearly 19 million objects brighter than magnitude 16, of which more than 15 million are classified as stars. A HUBBLE DEEP FIELD (HDF) typically reveals about 1500 galaxies down to nearly magnitude 30 (i.e. nearly four billion times fainter than the human eye can see) in an area of a few arcminutes (equivalent to a dime seen at a distance of 75 feet). One thus realizes the challenge of unequivocally identifying objects in the sky as the instrumental sensitivity and resolution increase so dramatically.

The nomenclature of astronomical objects follows a number of rules that differ with the objects at hand (see CONSTELLATIONS, STARS, AND DEEP-SKY OBJECTS: NOMENCLATURE).

Solar-system objects, including natural satellites, minor planets and comets, as well as planetary features, are named following recommendations of *ad hoc* professional committees confirmed by the INTERNATIONAL ASTRONOMICAL UNION (IAU).

Discoveries of new bodies are handled by the Central Bureau of Astronomical Telegrams (CBAT) and the Minor Planet Center (MPC) that assign provisional designations until ratification by the IAU (for example, lunar craters named after outstanding scientists, asteroids bearing names selected by their discoverers, and so on). Comets are nowadays the only bodies still receiving the names (up to three) of their discoverers, together with an alphanumerical identification. The CBAT also informs on discoveries of supernovae as well as other remarkable discoveries, events and celestial objects.

A whole book would be necessary to describe in details all the systems used to designate stars and non-stellar objects. If a few bright ones have ancient historical names (Sirius), other designations include constellation memberships (α CMa), memberships of declination zones (–18°1345), sequential discovery numberings (V341 Sco), references to wavelength ranges (Sco X-1), to object natures (PSR 0031-007), to observatories, instruments or spacecraft (Lick Hα 52), to larger objects (NGC 125-6), to astronomer names (Barnard's star), and so on, not to forget the most common identifications: sequence numberings in hundreds of astronomical catalogs and observing logs available. In other words, most celestial objects have just numbers and positions on the sky.

A few words are in order here on a fashion that has been developed by non-official organizations on naming celestial objects and selling corresponding certificates. The fact is that such names have no formal or official validity whatever. Astronomers and their representative organizations dissociate themselves entirely from such commercial practices of selling fictitious names of surface features, of stars and of celestial bodies.

The IAU is the sole internationally recognized authority for naming them. Those names are not sold, but assigned according to internationally accepted rules, recognized and used by scientists, space agencies and authorities worldwide.

Catalogs, surveys, observing logs and archives

As introduced above, catalogs can be organized according to object types (e.g. planetary nebulae, quasars, pulsars, …), according to the type of data they offer (e.g. proper motions, photometric indices, spectral types, …), or both. So many catalogs are available nowadays that it is out of the scope of this article to detail them. Refer, however, to the services maintained by the data centers (see below and the list of URLs at the end of the article).

Surveys refer more to systematic coverages of the sky with specific instruments, such as the famous Palomar Observatory Sky Survey (POSS), which was in fact a professional photographic atlas of the sky as seen in two colors by the Palomar Schmidt telescope.

Important catalogs nowadays are also what is called the observing logs gathering together the details of the observations carried out with a specific instrument (objects observed; date, time, duration of observations;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

instrumental specifications and/or configurations; and so on). They are in general linked to instrumental archives.

Archiving is also a critical issue in astronomy, basically because practically all celestial objects are variable with time in a way or another and also, as seen earlier, because we are essentially passive observers of the universe and not experimenters who would be able to recreate at any time situations for further investigations.

The memory of past observations is thus fundamental for future studies since furthermore it is not known for what purposes the next generations of astronomers would use the data. For a couple of decades now, the funding of all big new projects has included provisions for comprehensive image processing, adequate archiving and further exploitation of the archive after the possible termination of the mission. This might sound an obvious policy, but it is in fact a rather recent approach.

The challenge here is of a technical nature as the average lifetime of storage media is about 4 yr and about 3 yr for the user interfaces. Thus any long-term archiving policy has to make provisions for regularly transfering data on new material.

Dababases and information hubs

Beyond catalogs and collections of catalogs, extremely powerful databases are now available. The data centers themselves have become information hubs providing a whole spectrum of world-wide services.

The Strasbourg astronomical Data Center (CDS) (figure 2) has been a long-time pioneer and is nowadays recognized as the world leader in these matters.

In the early 1970s, a number of European institutions decided indeed to collaborate and to create a data center at Strasbourg Observatory with the daunting task of establishing an enormous table of synonyms between all catalog identifications. This was definitely aimed at avoiding the repetition of a couple of situations where two qualified astronomers studied and published papers on the same object, but under different identifications and without realizing it.

However, more interestingly, this was allowing, with a single object name, access to all other identifiers, plus to the data listed in the various catalogs. As CDS also set up a comprehensive and very successful bibliographical object-oriented database, the references of the papers dealing with that object were immediately available by the same token. The popular Simbad database was born.

Simbad holds today more than 2 200 000 objects under about 5 500 000 identifiers, together with more than 105 000 bibliographical references including about 3 000 000 object citations. Anyone starting studying seriously an astronomical object must pay a visit to Simbad first.

Figure 2 shows other components of CDS's information hub. VizieR is a search-and-shop individual-catalog service, with also access to large tables published in professional journals. Aladin is an interactive digitized sky atlas. Besides the references of Simbad, the CDS bibliographical service provides also access to abstracts of several major journals (see also ADS hereafter). The dictionary of nomenclature gives references and details on usage for more than 4000 different catalog acronyms. AstroGlu is a discovery tool helping to locate database servers providing relevant information.

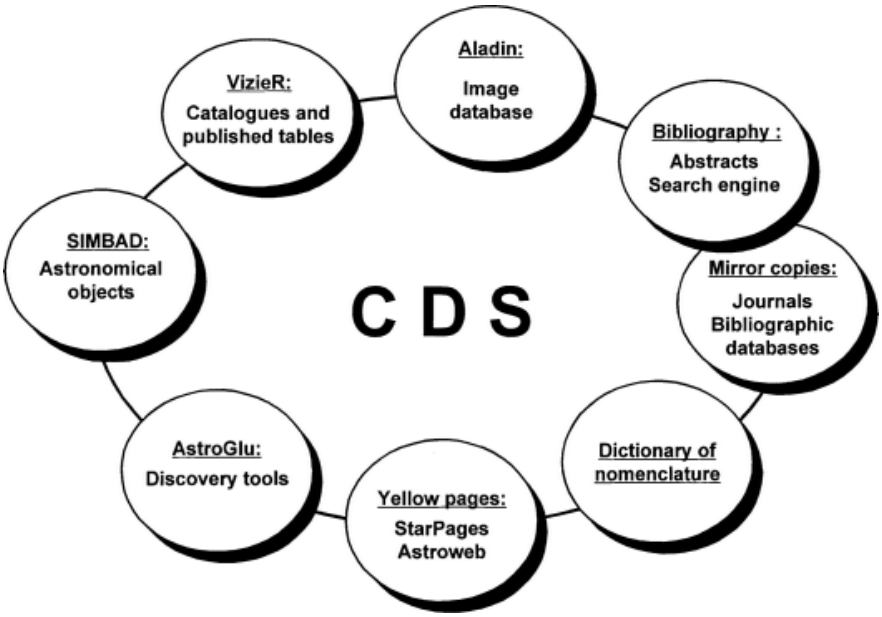


Figure 2. A schematic view of the CDS information hub (courtesy Centre de Données de Strasbourg, CDS).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Last, but not least, the StarPages are a set of permanently updated yellow-page resources providing detailed information on astronomy-related organizations (StarWorlds: about 6300 entries, more than 5500 WWW links), as well as on individual astronomers and related scientists (StarHeads: more than 5200 WWW home pages). They offer also access to an enormous dictionary (StarBits: about 140 000 entries) of abbreviations and acronyms related to astronomy and associated space sciences. The master files for StarWorlds have been used to produce figure 4 (see below).

A few other multipurpose resources deserve also a mention in this section.

The NASA/IPAC Extragalactic Database (NED) has been built around a master list of extragalactic objects for which cross-identifications of names have been established, accurate positions and redshifts entered to the extent possible, and some basic data collected. Bibliographic references relevant to individual objects have also been compiled.

The Astrophysics Data System (ADS) is a NASA-funded project whose main resource is an abstract service (about half a million abstracts for astronomy and astrophysics only), together with links to scanned images of over 40 000 journal articles. ADS provides also access to astronomical data catalogs and data archives, thereby making data collected by NASA space missions available to astronomers. It offers also access to the StarPages.

The National Space Science Data Center (NSSDC) provides access to a wide variety of astrophysics, space physics, solar physics, lunar and planetary data from NASA missions, together with some complementary data.

The Canadian Astronomy Data Centre (CADC) and the Astronomical Data Analysis Center (ADAC) at the National Astronomical Observatory of Japan (NAOJ) are national resources offering access to a number of catalogs, databases and archives.

Data processing

Until not so long ago, data processing consisted essentially in photographic-plate scanning and in reduction of raw photometric data from punched papertapes.

Nowadays, with the omnipresent digitization and multidimensionality of collected data, one rather speaks of image processing and a whole treatise could be devoted to it.

Indeed the time seems definitely gone, at least with the big instruments, when astronomers were coming themselves to the telescope (or ground observatory for a spacecraft) and later returning home with unprocessed data. The advanced experiments have developed their own specific image processing which is now part of the projects themselves since the development phase.

Several comprehensive image handling systems have however been developed such as ESO's Munich Image Data Analysis System (MIDAS) and NOAO's Image Reduction and Analysis Facility (IRAF). They can be considered as general-purpose software systems for the reduction and analysis of astronomical data and are regularly upgraded.

Complementary resources would include specific software packages and libraries (such as statistical ones), spectral line compilations, abundance libraries, opacity tables, and so on. A network such as Starlink helps UK-based astronomers to reduce and analyze their observations.

It is certainly worthwhile to mention here that astronomers have introduced the Flexible Image Transport System (FITS) which is a general way to encode both definitions of data and the data themselves and which is machine independent. The FITS format has been quickly recommended for interchange of image data between all observatories—and is now in use even outside astronomy.

Publishing and information sharing

Publishing is not only motivated by the noble aims of educating and information sharing but also strongly conditioned by career constraints involving recognition, a necessity that should not be underestimated. Recognition is sought for getting positions (i.e. grants and salaries), for obtaining acceptance of proposals (e.g. leading to data collection) and for achieving funding of projects (allowing materialization of ideas).

The pressure for recognition has contributed to the strong increase of professional publications (see figure 3), together with other factors such as the expansion of the astronomy community itself (especially after the beginning of the space age), the multiplication of large instruments and spacecraft equipped with always faster, more diversified and more efficient detectors, and so on. Commercial publishers have also put on the market more journals which were as many additional communication outlets.

The major professional journals use the peer-review procedure ('refereeing') for accepting, amending or rejecting submitted contributions. Albeit a matter of regular debates (on the principle itself as well as on the way it is conducted), the refereeing process has been so far the best one (or the less questionable one) to publish contributions with validated content, i.e. an assurance of good-quality, novel results obtained by reproducible experiments, calculations or analyses on which enough details are provided.

The most important general professional journals include the *Astrophysical Journal* and the *Astronomical Journal* published by the *American Astronomical Society*, the *Monthly Notices of the Royal Astronomical Society*,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

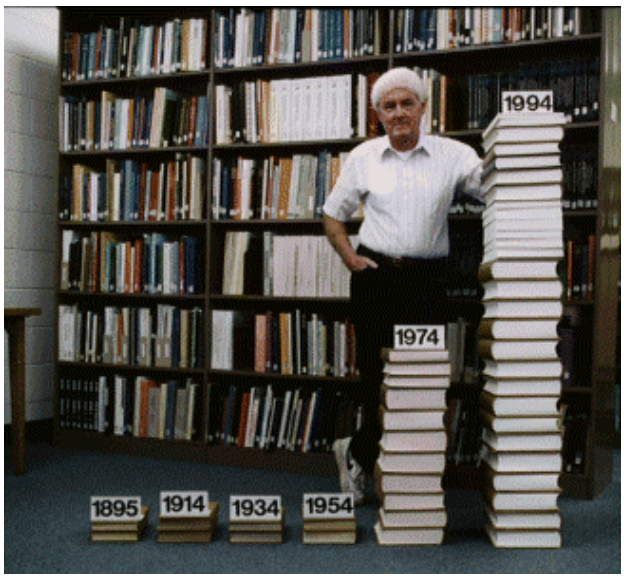


Figure 3. An illustration of the dramatic increase of astronomical literature over the past decades. Helmut A Abt, Editor-in-Chief of the *Astrophysical Journal*, is standing next to stacks of that leading professional publication (courtesy the *National Optical Astronomy Observatories, NOAO*).

and *Astronomy and Astrophysics* resulting from the merging in 1969 of a number of European professional journals (see also ASTRONOMICAL SOCIETIES AND PUBLICATIONS).

Astronomers communicate also via a whole spectrum of publications ranging from informal newsletters to books gathering together review papers by the best specialists on specific topics. Conferences, colloquiums, workshops and meetings of all kinds provide also efficient ways of exposing oneself to both excellent review talks and presentations of works in progress. The corresponding proceedings are published by commercial publishers, by learned societies, by research institutions or even by individuals, reasonably soon after the events.

As described in the following section, publishing is nowadays increasingly done electronically, or, better said, there is more and more of diversified publishing, i.e. of information available on different media (paper, CD-ROM, web sites, and so on). These media are not excluding, but complementing, each other. Several journals have an electronic counterpart.

As mentioned earlier, professional astronomers are also contributing substantially to less specialized publications, mainly directed towards amateur astronomers and the public at large. Many countries have their own such national journal, but *Sky & Telescope* is probably the magazine with the largest audience world-wide (see also MEDIA).

Electronic astronomy

As mentioned already, more and more of the information exchanges in astronomy are done electronically, both dynamically (e-mail) and passively (web sites). Electronic handling is particularly well adapted to the fluid and living nature of today's information material, to the digitized nature of most data and to the efficiency of contacts between collaborators spread over the world, as well as to the retrieval of information from the information hubs. It also makes easy the various interactions upstream and downstream, and allows diversified publishing.

Figure 4 gives an idea of the distribution, at the time of writing this article, of the astronomy-related organizations using e-mail and/or having a web site. Together with the general world distribution, more detailed views are displayed for North America and Western Europe.

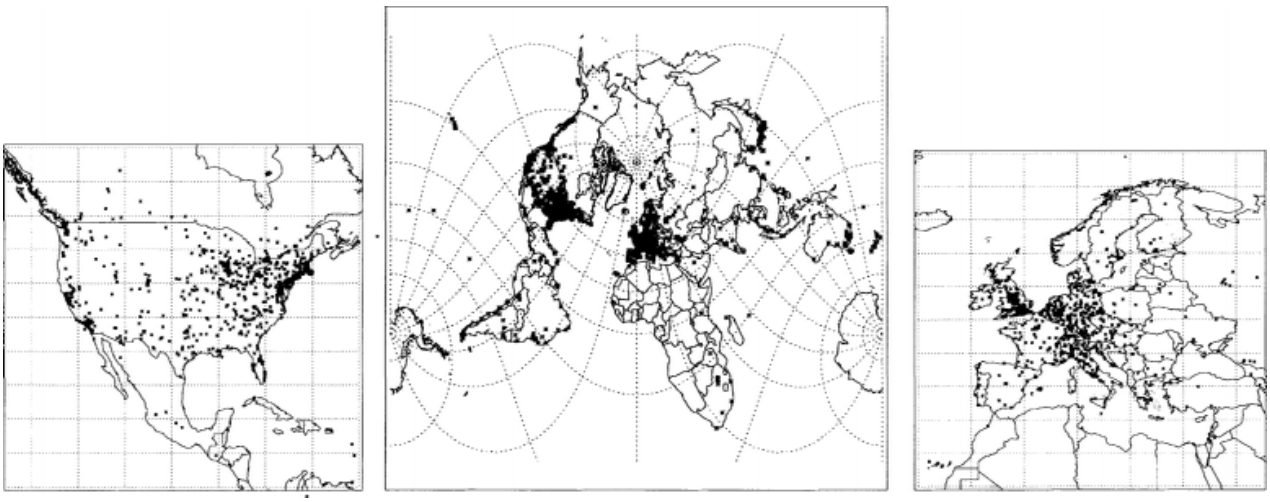


Figure 4. Geographical distributions (North America, World, Western Europe) of astronomy-related organizations (all categories) with electronic facilities (e-mail and/or web pages) in April 1999.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The largest concentrations are located in Europe and in the USA (Northeast and California), with a few nuclei in Australia, India, Japan, New Zealand, as well as a few spots in South America. A striking feature of the central map is the desperate emptiness of the African continent. A similar comment is also applicable to quite a number of the so-called third-world countries.

In Western Europe, it is striking how France, Portugal and Spain have significantly much lower densities than their European neighbors, obviously lagging behind as to the penetration of the internet and the WWW, a few years after the electronic medium started spreading quickly over the world.

The 'electronization' of astronomical information handling is, however, not a bed of roses. New facilities and new possibilities bring in naturally new questions, new challenges and new problems that have to be faced, especially at the ethical level (proper credits of downloaded material, for instance), as well as at the legal (financial and copyright policies, electronic signatures, for instance) and educational (training of young and not-so-young people) ones, without forgetting the security and the fragility of the material delivered via the electronic medium—a very worrying factor for many people.

The most demanding challenge, however, arises probably from the enormous quantity of information easily reachable (and providable) by everybody nowadays, as discussed in the following section.

The information retrieval challenge

At ACM97, the conference celebrating the 50th anniversary of the Association for Computing Machinery, the Nobel Prize laureate Murray Gell-Mann called attention to the fact that, with the digital age producing an 'immense sea of data that threatens to drown humanity', people needed to adapt how they think so that true knowledge can be distilled from the deluge.

> 'We hear, in this dawn of the so-called information age, a great deal of talk about the explosion of information and new methods for its dissemination. It is important to realize, however, that most of what is disseminated is misinformation, badly organized information or irrelevant information. How can we establish a reward system such that many competing but skillful processors of information, acting as intermediaries, will arise to interpret for us this mass of unorganized, partially false material?'

That challenge is probably the most critical one currently in the field of information handling—and astronomy is facing it too.

We shall have to rely, indeed, on the wisdom of providers and users of electronic information, as well as on the various intermediaries (compilers, information hubs,…) and on the learned societies, committee experts and so on to take this into account, to put, more than ever, the emphasis on validated and authenticated electronic information, and to reward appropriately the scientists who are or will dedicate a full-time job to such activities.

The quality of resources as well as their maintenance must be continuously improved from lessons learned with time and using the most adequate tools. Generally speaking, information has to be collected, verified, de-biased, homogenized and made available not only in an efficient way but also through reliable channels. Sophisticated techniques cannot save the extensive background, unrewarding and very careful work which is indispensable for the compilation of a valuable resource. One could never stress enough the importance of this obscure daily work consisting of patiently collecting data, checking information and updating it. This has also to be carried out by knowledgable scientists or documentalists and cannot be delegated to inexperienced clerks.

Efficient search engines working on validated and authenticated material must enable finding information looked for, if available at all, and whether it relates to celestial objects, to data, to bibliography or to the vast coverage by yellow-page services. That information must be of good quality, relevant and on target. The best search engines available today and the corresponding organizations have been mentioned in this article.

As technology leaders agree that changes are the only sure thing about computing and communications in the next decades, we can certainly expect further modifications on the information handling in astronomy and consequently on the sociodynamics of the astronomy community itself.

Useful URLs

For WWW access, here are—in alphabetical order—the URLs of the most relevant organizations and resources mentioned in the article. Most of them are collaborating with each other and are hosting mirror pages of their partners. Complementary resources are also generally available from the sites.

- American Astronomical Society (http://www.aas.org/)
- Association for Computing Machinery (http://www.acm.org/)
- Astrophysics Data System (http://adsabs.harvard.edu/)
- Canadian Astronomy Data Centre (http://cadcwww.dao.nrc.ca/CADC-homepage.html)
- Central Bureau of Astronomical Telegrams (http://cfa-www.harvard.edu/cfa/ps/cbat.html)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- International Astronomical Union (http://www.iau.org/)
- Minor Planet Center (http://cfa-www.harvard.edu/iau/mpc.html)
- NAOJ Astronomical Data Analysis Center (http://adac.mtk.nao.ac.jp/)
- NASA/IPAC Extragalactic Database (http://nedwww.ipac.caltech.edu/)
- National Space Science Data Center (http://nssdc.gsfc.nasa.gov/)
- Royal Astronomical Society (http://www.ras.org.uk/ras/)
- Simbad (http://simbad.u-strasbg.fr/Simbad)
- Sky & Telescope (http://www.skypub.com/)
- Starlink (http://star-www.rl.ac.uk/)
- StarPages (http://vizier.u-strasbg.fr/starpages.html)
- Strasbourg Data Center (http://cdsweb.u-strasbg.fr/CDS.html)

Bibliography

Because of the quick evolution of information technologies nowadays, and because of the impact they have on the dynamics of the astronomy community, some aspects of the most recent compilations and reviews could become rather quickly outdated. The basic principles will largerly remain unchanged, however, and the following books could be recommended as possible further, more technical, reading.

Boroson T, Davies J and Robson I (ed) 1996 *New Observing Modes for the Next Century* (San Francisco, CA: Astronomical Society of the Pacific)

Egret D and Albrecht M A (eds) 1995 *Information & On-Line Data in Astronomy* (Dordrecht: Kluwer)

Heck A (ed) 1997 *Electronic Publishing for Physics and Astronomy* (Dordrecht: Kluwer)

Heck A and Murtagh F (ed) 1993 *Intelligent Information Retrieval: The Case of Astronomy and Related Space Sciences* (Dordrecht: Kluwer)

Jaschek C 1989 *Data in Astronomy* (Cambridge: Cambridge University Press)

Series of specialized conferences provide also sources for advanced material. See for instance the following as well as the numerous references quoted therein:

- the Astronomical Data Analysis Software and Systems conferences (proceedings published by the Astronomical Society of the Pacific, San Francisco),
- the Astronomy from Large Databases conferences (proceedings published by the European Southern Observatory, Garching),
- the Data Analysis in Astronomy ('Erice') workshhops (proceedings published by Plenum Press, New York, and World Scientific, Singapore),
- the Converging Computing Methodologies in Astronomy workshops of the European Science Foundation (proceedings published in Vistas in Astronomy).

André Heck

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared Arrays

Infrared arrays are small electronic imaging devices subdivided into a grid or 'array' of picture elements, or pixels, each of which is made of a material sensitive to photons (ELECTROMAGNETIC RADIATION) with wavelengths much longer than normal visible light. Typical dimensions of currently available devices are about 27–36 mm square, and formats now range from 2048×2048 pixels for the near-infrared detectors down to about 128 × 128 pixels for the mid- and far-infrared regions. These imaging devices have completely revolutionized INFRARED ASTRONOMY since their introduction in the late 1980s because they provide images and spectroscopic information about regions of space that are not accessible to visible light detectors (see also INFRARED IMAGERS). Such regions include the dusty, gaseous clouds where star formation occurs, the obscured center of the Milky Way some 28 000 light-years from the solar system, as well as the most distant galaxies in the universe. Distant objects must be studied at infrared wavelengths because the Hubble expansion of the universe has shifted their spectra far into the red.

While there is some similarity between infrared arrays and silicon CHARGE-COUPLED DEVICES (CCDs), there are also fundamental differences. Infrared arrays cannot use pure silicon as the detector material because it is not sensitive to light of wavelengths longer than 1.1 μm, no charge-coupling is used to extract the signals, and much lower cryogenic temperatures are needed to operate these devices. Infrared arrays require cooling to temperatures between liquid nitrogen (77 K) and liquid helium (4 K), depending on their wavelength range of operation.

To generate an image, all infrared array detectors must convert radiation into electrical charge by absorbing a photon. The material used to detect the photon is a semiconductor, that is, a crystalline material with electrical conductivity properties intermediate between that of a good conductor like copper and an insulator like glass. Silicon is the best known semiconductor. Other examples include germanium (Ge), gallium arsenide (GaAs) and indium antimonide (InSb). Each is characterized by a crystal structure with an energy bandgap which separates a conduction band from the normal valence band of the atoms. The absorption of a photon of sufficient energy excites an electron from the valence band into the conduction band. The smaller the bandgap, the lower the photon energy needed and the longer the wavelength of radiation that can be detected by the material. If the bandgap energy (EG) is expressed in electron-volts, then the longest wavelength that can be detected is given by 1.24 μm/EG. The bandgap varies slightly with temperature, but typical values are 1.1 eV for Si, 0.67 eV for Ge, 0.18 eV for InSb

There are two classes of semiconductors used in infrared photon detection; intrinsic and extrinsic. In an intrinsic semiconductor crystal, a photon with sufficient energy creates an electron–hole pair when the electron is excited to the conduction band, leaving a positively charged site or 'hole' in the valence band; this is the same physical process that operates in a silicon CCD to detect photons. Common infrared intrinsic materials are germanium (Ge), lead sulfide (PbS), indium antimonide (InSb) and mercury cadmium telluride (HgCdTe). As already shown above, these have bandgap energies less than silicon and therefore have a longer cut-off wavelength. Extrinsic semiconductors are intrinsic materials, typically silicon and germanium, which have been 'doped' with impurity atoms so that a very low energy photon, with insufficient energy to excite an electron–hole pair directly, can still cause an excitation from an energy level associated with the impurity atom. In effect, a stepping stone is placed in the energy bandgap of the intrinsic material. Examples of extrinsic semiconductors used in astronomy applications are arsenic-doped silicon (Si:As) which can detect photon wavelengths as long as 23 μm, and gallium-doped germanium (Ge:Ga) which goes all the way to 115 μm.

Another kind of infrared array device is found frequently in industrial surveillance applications, this is the platinum silicide (PtSi) array. A PtSi array is based on the Schottky barrier principle. When a metal such as platinum silicide is brought into contact with p-type silicon, the result is an asymmetric potential barrier. Conduction holes in the metal must now overcome a slightly higher barrier to enter the valence band of the semiconductor. The barrier height is determined by the contact potential and can be less than the semiconductor bandgap; for PtSi the cut-off wavelength is 5.6 μm. Although more easily manufactured into large-format arrays, because it is an all-silicon process, the photon detection efficiency of these detectors is rather low and their use in astronomy has been quite limited so far.

All of the detectors mentioned so far are photon detectors. For completeness, a fundamentally different kind of detector should be mentioned. It is the bolometer, which belongs to a class called thermal detectors. These devices absorb photons and convert their energy to heat and the heat energy is detected through some change in physical properties such as conduction of the material. Only the energy deposited by the photons is important, not their wavelength. Semiconductor bolometers are the most highly developed, but bolometer arrays have tended to be very small (30 × 30 pixels) so far, compared with near-infrared arrays.

Any infrared array detector must perform three functions:

- store the accumulated electrical charge at the site of generation, i.e. in a pixel;
- transfer the charge on each pixel to an outlet; this task is called multiplexing;
- convert charge to a voltage which can then be readily digitized into a number for storage in a computer.

All infrared arrays separate the functions of detecting infrared radiation from that of multiplexing the resulting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

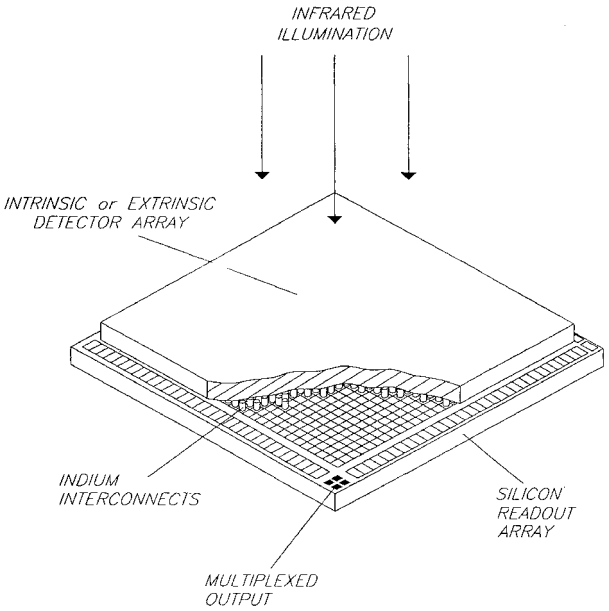


Figure 1. A schematic diagram illustrating the hybrid nature of infrared arrays. The upper slab is the infrared detector and the lower slab is the silicon-based readout device. Interconnections are made with 'bump' bonds of indium.

electrical signal, with the latter task going to silicon. Infrared arrays are like 'sandwiches' in which the upper slab is the infrared sensor (e.g. InSb, HgCdTe; Si:XX, Ge:XX, where XX represents the element used for doping, or PtSi), and the lower slab is a silicon multiplexer of some sort. A special array of 'switches' made from silicon field effect transistors (FETs) in microscopic form is typically used in the lower slab to access the signal from each IR pixel above.

The infrared-sensitive part of an infrared array is a tightly packed grid of individual pixels with minimum dead space between them (figure 1). Both slabs are provided with a grid of electrical connections in the form of tiny raised sections, referred to as 'bumps', of a soft electrical conductor called indium; indium remains soft at low temperatures. The slabs are literally pressed together to enable the indium bumps to mate. Gaps between the bumps are usually back-filled with an epoxy to help maintain the integrity of the sandwich, especially through thermal cycles to low temperatures. Charge storage may occur on the junction capacitance of the IR sensor itself or on a separate storage capacitor associated with the silicon circuitry. Infrared arrays are harder to manufacture than silicon CCDs and consequently more expensive. The leading suppliers of astronomical devices are Rockwell International, Boeing and Raytheon/SBRC in the United States. Several other companies provide infrared arrays suitable for surveillance applications.

Either the upper slab is constructed on an IR-transparent substrate, such as the sapphire substrate used for the HgCdTe arrays, or the bulk semiconductor is physically thinned to enable photons to penetrate the backside to the pixel locations on the inner face, as in the case of the InSb arrays. When the absorption of a photon with a wavelength shorter than the cut-off wavelength generates an electron–hole pair in the semiconductor, at the location of the pixel, the pair is immediately separated by an electric field which can be applied externally or internally or both. For example, in near-infrared arrays of InSb and HgCdTe, the electric field results from a reverse-biased pn junction or photodiode. The depletion region produced by the reversed-bias junction acts like a capacitor (C) which is discharged from its initial state (the full reverse-bias voltage V) by the migration of the electrons and holes; in effect, photo-generated charges are being 'stored' locally at each pixel. The maximum amount of charge (expressed as a number of electrons) which can be stored is given by $Q = CV/e$, where e is the charge on the electron (1.6×10^{-19} C). Of course, there must be a small perimeter or boundary around a pixel to define each individual photodiode, and this region cannot be too large otherwise there is a serious light loss, nor can it be too small because then there will be cross-talk between pixels, i.e. the charge from one pixel can be collected by an adjacent pixel. When the capacitor is completely discharged, the storage well is full and the pixel is said to be 'saturated'. No damage occurs, but charge accumulation ceases and the signal is no longer proportional to the number of photons detected.

Charge on each pixel is said to be 'read out' by sequentially connecting its signal to a common output line, or bus, leading to one or more output amplifiers. Multiple readout amplifiers are common in infrared arrays because thermal emission of radiation by the telescope optics and the sky itself introduces a significant 'background' light at wavelengths longer than about 2.5 μm. This background photon emission fills up the storage wells sufficiently rapidly that it helps to expedite the readout process by having more than one amplifier. On-chip circuits called 'shift registers' perform the readout and multiplex sequencing. In the most common and successful multiplexer designs, the 'unit cell' of an infrared array contains a silicon field-effect transistor (FET) used as a source-follower amplifier, essentially providing a buffer for the accumulated charge in the infrared pixel. The term used to describe this structure is SFD (source follower per detector). There is also an electrical noise associated with this circuitry which results in an unwanted 'readout' noise.

The voltage change across the detector capacitance is applied directly to the input gate of the silicon FET (across the indium bump interconnect), which in turns relays the voltage change to one or more output lines when it is switched on or 'addressed' to do so. The output voltage is amplified and digitized by external circuits and a number for the signal in that pixel is stored in the corresponding location in a computer file.

Infrared array detectors are not based on the charge-coupling principle of the silicon CCD. This is an important distinction which has some practical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

implications. For example, IR arrays do not 'bleed' along columns when a pixel saturates, and 'bad' pixels do not block off others in the same column. Also, since the pixel charge does not move, 'non-destructive' readout schemes are possible and are very effective in reducing unwanted electrical noise. Some near-infrared arrays manufactured by Rockwell International using HgCdTe detectors have achieved readout noise levels as low as two to three electrons (equivalent) using multiple non-destructive measurements to average down the noise, but noise of 15–30 electrons is more common. On the other hand, CCD features such as on-chip charge binning, charge-shifting and drift-scanning are not possible with infrared arrays.

Most long-wavelength (i.e. wavelengths >8 μm) arrays use extrinsic silicon detectors employing a technology called impurity band conduction (IBC). In an IBC device, a heavily doped infrared-active layer is placed in contact with a pure (undoped) epitaxial (thin) layer, the blocking layer, which is isolated by an oxide (insulating) layer from metal contact pads and the device is usually back-illuminated. These arrays are also called blocked impurity band (BIB) devices. Because of the high donor density, the applied (bias) electric field causes a migration of holes toward the (negative) metal contact forming a depletion region similar to that in the photodiode (pn junction) devices used at shorter wavelengths.

For wavelengths longer than 40 μm there are no appropriate shallow dopants for silicon and therefore extrinsic germanium must be used. There are a number of problems with the use of germanium. For example, to control dark current the material must be relatively lightly doped and therefore absorption lengths become long (3–5 mm). Since the diffusion lengths are also large (250–300 μm), then pixel dimensions of 500–700 μm are required to minimize 'cross-talk' of the signals between pixels. Large pixels imply higher hit rates for cosmic rays, especially for space applications, which in turn means that the readout device must be very low noise so that the background limit is reached in the shortest possible exposure time. But a large detector pixel means a large capacitance and more noise. Also, the photoconductive gain is inversely proportional to the interelectrode spacing, resulting in poor quantum efficiency unless side-illuminated detectors with transverse contacts are used. Finally, because of the very small energy bandgaps, these detectors must operate at liquid helium temperatures well below the silicon 'freeze-out' range. Despite these challenges, small format arrays with response out to beyond 200 μm have been constructed for infrared space missions such as SIRTF.

Since very narrow bandgap semiconductor crystals are harder to grow, it is tempting to seek a method of 'engineering' small bandgaps in wide-bandgap material such as gallium arsenide (GaAs). Using GaAs and AlGaAs junctions it is possible to create a 'quantum well' or potential associated with either the conduction or the valence band. The quantum well is equivalent to the particle-in-a-box problem in quantum mechanics and so the well will contain energy levels or sub-bands. The energy difference between these levels is very small compared with the normal bandgap. Transitions between the sub-bands provide the infrared detection process. Such detectors are currently under development.

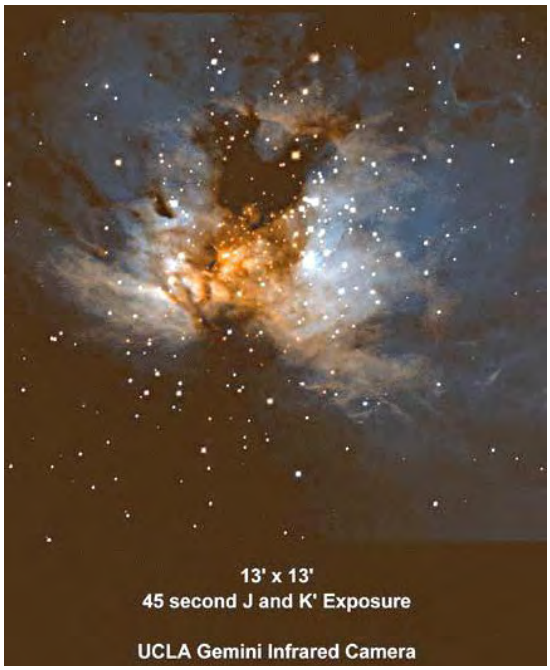


Figure 2. The star-forming region NGC2024 near Zeta Orionis in Orion. The image is a three-color composite of near-infrared bands at 1.2, 1.65 and 2.2 μm represented as blue, green and red respectively. The central region of this nebula appears black in optical photographs due to absorption by dust grains. Infrared images reveal a dense cluster of very young stars.

As an illustration of their use, figure 2 shows an image obtained with an infrared array camera.

Bibliography

Fowler A (ed) 1998 Infrared astronomical instrumentation *Proc. SPIE* **3354**

McLean I S 1997 *Electronic Imaging in Astronomy: Detectors and Instrumentation* (Chichester: Praxis)

Rieke G H 1994 *Detection of Light: from the Ultraviolet to the Submillimeter* (Cambridge: Cambridge University Press)

Ian S McLean

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared Astronomy

The infrared part of the electromagnetic spectrum covers the wavelength range from about 1 mm (300 GHz) to 1 μm (300 THz). The infrared (IR) band is usually subdivided into the sub-millimeter (1 mm to 0.3 mm), the far- (0.3 mm to 50 μm, FIR), mid- (50 μm to 10 μm, MIR) and near-infrared (10 μm to 1 μm, NIR). At the long-wavelength end it connects to the radio domain and at the short end to the optical. In the FIR and most of the MIR domain astronomical sources cannot be observed from the ground. The other domains exhibit atmospheric windows of varying depth. Astronomical sources observed in the IR exhibit radiation temperatures of a few thousand to a few degrees kelvin. The most prominent sources are PLANETS, late-type stars, interstellar dust and gas clouds (see INTERSTELLAR MATTER), infrared luminous galaxies, the 3 K COSMIC MICROWAVE BACKGROUND and non-thermal extragalactic radio sources (QUASARS). Important reasons to carry out observations in the infrared are that the effect of dust extinction is minimized in this wavelength domain, the determination of the total luminosity of dust-enshrouded objects, and the fact that the visible to ultraviolet spectra of highly redshifted galaxies and quasars are shifted into the infrared while these objects become unaccessible in the optical due to extinction by hydrogen below a rest wavelength of 91 nm.

History

In 1800 Sir John HERSCHEL dispersed solar light through a glass prism onto several thermometers. Thomas Edison observed a total solar eclipse in 1878 in the infrared. In the early 1900s the Moon, planets and bright stars were observed in the infrared by S Pettit and E Nicholson. This work was continued by the systematic and extensive infrared studies of planets and stars by G KUIPER and H L JOHNSON in the 1950s (see Low and Rieke 1974). Frank J Low fully opened the field to astronomy in the early 1960s by inventing a low-noise gallium-doped germanium bolometer. In the following years the development of array detectors based on InSb and HgCdTe (up to 1024 $\times$ 1024 pixels) as radiation-sensitive substances determined the progress in the near-infrared. Balloon, airborne and satellite based observatories allowed measurements even at far-infrared wavelengths. Dedicated telescopes and sensitive bolometers and heterodyne receivers based on Schottky and SIS diodes opened the submillimeter domain.

Telescopes and observational techniques

The short NIR domain is contaminated by variable non-thermal atmospheric line emissions and with the exception of the NIR shortward of 2 μm the infrared is dominated by the thermal emission of the local background that includes the partially transparent atmosphere and all optical elements at the ambient temperature. Sensitive measurements are only possible from high mountains, balloons, airplanes or satellites using cooled instruments, detectors or (in space) even cooled telescopes. To further suppress the thermal background, secondary telescope mirror sizes are kept as small as possible and in the MIR the solid angle on the sky that is encompassed by individual detector elements is minimized in order to suppress the atmospheric thermal radiation. A novel technique in the short NIR is 'OH background suppression' which involves instrumental blanking of the OH line emission at a high spectral resolution. Stable background suppressed observations include observational techniques such as 'nodding' (measurements at alternating positions by moving the telescope back and forth) and 'chopping' (measurements at alternating positions by moving, for example, the secondary mirror at a frequency $\geq$1 Hz). The flux calibration is usually performed via a comparison with local radiation sources (black-bodies) at ambient ($\sim$300 K) or liquid nitrogen ($\sim$77 K) temperature or astronomical standards like planets, quasars or reference stars.

Usually the angular resolution of an optical/infrared telescope is limited to about one arcsecond due to the turbulent atmosphere. In order to obtain diffraction-limited images from the ground at several telescopes 'adaptive optics' has been employed (Alloin and Mariotti 1994, Tyson 1991). Traveling through the Earth's atmosphere the initially plane wave front from the object is distorted and measured at the telescope in the optical with a Shack–Hartmann or curvature wavefront sensor (see Tyson 1991). This signal is then used to control a deformable mirror that can correct the wavefront distortions in the NIR. The resulting images have a diffraction-limited part that contains typically up to 60% of the light. The angular resolution A in arcseconds that can be achieved depends on the observing wavelength λ in micrometers and the diameter D of the telescope in meters and can be calculated via $A(\text{arcsec}) \sim 0.206 \times \lambda/D$. For an 8 m class telescope it is of the order of 0.06 arcseconds at a wavelength of 2.2 μm. Combining different telescopes interferometrically subarcsecond angular resolutions can be obtained in the submillimeter domain (Downes 1989, Watt and Webster 1990). Interferometric measurements can also be applied in the MIR and NIR domain. They provide high-accuracy astrometry and information on stellar diameters and orbital elements of multiple stars.

Telescopes

A large amount of information has been gained with INFRARED TELESCOPES in the 2 to 4 m class (e.g. the Infrared Telescope Facility, IRTF, of NASA or the New Technology Telescope, NTT, of the EUROPEAN SOUTHERN OBSERVATORY, ESO). Large-aperture telescopes with dedicated NIR and MIR facilities are the Keck Telescope in Hawaii (10 m diameter) the ESO Very Large Telescope (VLT) on Paranal in Chile (four 8 m telescopes). Others like the LARGE BINOCULAR TELESCOPE (LBT; two mechanically coupled 8.4 m telescopes) at the MT GRAHAM INTERNATIONAL OBSERVATORY in Arizona, the multinational GEMINI OBSERVATORY twin 8.1 m telescopes in Chile and Hawaii, or the 8.3 m diameter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SUBARU TELESCOPE of Japan in Hawaii will soon come into operation. In addition to space missions that are summarized below, MIR and FIR observations have been carried out using the Kuiper Airborne Observatory (KAO) with a 36 in telescope on board a C-141 airplane at a height of up to about 14 km. This observatory will be followed by SOFIA (the USA/German Stratospheric Observatory for Infrared Astronomy), a 2.5 m mirror on board a Boeing 747. In the submillimeter domain large single-dish antennas like the NOBEYAMA RADIO OBSERVATORY 40 m dish in Japan and the 30 m telescope of IRAM (INSTITUT DE RADIO ASTRONOMIE MILLIMÉTRIQUE) in Europe and the JAMES CLERK MAXWELL TELESCOPE (JCMT) in Hawaii (15 m), as well as several millimeter wavelength interferometry arrays in the USA, Europe and Japan have contributed to the investigation of line and continuum emission of gas and dust in our own and external galaxies in the local universe and at cosmological redshifts. Further progress will be achieved in the near future via the infrared optimized NGST (Next Generation Space Telescope), interferometric arrays with a very large collecting area like the ALMA (the ATACAMA LARGE MILLIMETER ARRAY in northern Chile), as well as large bolometer and spectrometer focal plane array receivers for single dish telescopes.

Detectors

A broad variety of detector techniques is required in the infrared (e.g. Rieke 1994). In the NIR and MIR range large-format focal plane detector arrays are used. Arrays that use HgCdTe as IR sensitive material are hybrid structures. They have been built using molecular beam epitaxy and liquid phase epitaxy techniques to grow the infrared sensitive HgCdTe material. The spectral response for HgCdTe focal plane arrays typically used in astronomy ranges from from 0.8 μm to 2.4 μm. The backside illuminated material is connected to multiplexers of different formats (e.g. 1024×1024 pixels). For arrays that use InSb as a detector material the spectral response goes from 0.5 to 5.4 μm. Typical pixel sizes range between about 10 and 40 μm. About 60% to 80% of all incident photons are detected. Very low dark current and low read noise values make these arrays suited for long exposures under low-background conditions as realized in most astronomical applications. On the other hand, frame rates of above 10 Hz can be reached as well. A similarly high potential for low-background observations in the MIR have BIB (blocked impurity band) detectors.

Focal plane array detectors have also been used in the FIR. A 5×5 array of stressed Ge:Ga photoconductors was flown on the KAO in an imaging spectrometer. Such arrays have also successfully been operated in space (ISO-LWS and ISOPHOT, see below). Larger-format arrays will soon be used on SRTF (Space Infrared Telescope Facility) and FIRST (Far-Infrared and Submillimeter Telescope) and SOFIA. The wavelength range between approximately 80 and 210 μm can be covered with spectral resolutions ($R=\lambda/\Delta\lambda$) of the order of a few thousand.

In the submillimeter and FIR domain ^{3}He-cooled bolometers can be used to perform sensitive measurements of cold dust and faint continuum sources in general (e.g. SCUBA: Submillimetre Common-User Bolometer Array at the JCMT and the IRAM bolometer arrays). Heterodyne receivers for low-noise high-spectral-resolution measurements in the submillimeter range employ Schottky diodes (semiconductor–metal contacts) and SIS diodes (supraconductor/insulator/supraconductor tunneling contacts).

Physical mechanisms

The most prominent infrared continuum radiation mechanism in astrophysical sources is the BLACK-BODY RADIATION that follows a Planck law (see below). This radiation is emitted by all bodies, from dust to planets and stars, and peaks for temperatures below a few thousand kelvin in the infrared. For the ambient temperature of about 300 K the spectral peak is at a wavelength of 10 μm. Other important continuum radiation mechanisms are the free–free and synchrotron radiation emitted by thermal and relativistic plasmas respectively, in or near star-forming regions in our own galaxy and in the nuclei of Seyfert galaxies and quasars. The line emission is dominated by rotational lines in the millimeter and submillimeter domain, fine structure lines of atoms and ions in the FIR and rotational vibrational as well as recombination lines in the MIR and NIR domain. Due to large variations of optical depths and large column densities of emitting and absorbing material, radiative transport of energy is of special interest for both continuum and line radiation (e.g. Osterbrock 1974, Spitzer 1978).

Radiative transport

In a plane-parallel cloud, radiative transport at a frequency ν can be treated simply. The change of specific line intensity I_ν (erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}) along an infinitesimal element dz along the line of sight through a molecular cloud is given by:

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}z}=\frac{h\nu}{4\pi}n_u A_{ul}-\frac{h\nu}{4\pi}\Phi(\nu)I_\nu(n_u B_{ul}-n_l B_{lu})=\epsilon_\nu-\kappa_\nu I_\nu. \quad (1)$$

Here n_u, n_l are the volume densities of molecules in the upper and lower states of the transition, with the corresponding statistical weights g_u and g_l. The Einstein coefficient A_{ul} (s^{-1}) is the rate of decay of n_u by spontaneous radiative transitions. The line shape is described by $\Phi(\nu)$, normalized to $\int\Phi(\nu)\,\mathrm{d}\nu=1$. The volume emissivity and absorption coefficient are denoted by ϵ_ν and κ_ν. With these coefficients the sources function Σ_ν can be defined that is equivalent to Planck black-body radiation law and is the most important source function for electromagnetic radiation in the infrared:

$$\Sigma_\nu=\frac{\epsilon_\nu}{\kappa_\nu}=\frac{2h\nu^3}{c^2}\frac{1}{\exp(\frac{h\nu}{kT_{\mathrm{ex}}})-1}. \quad (2)$$

This gives the spectral energy density of a black-body as a function of the excitation temperature T_{ex}

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(an equivalent Planck temperature) that describes the population of states l and u through a thermal Boltzmann population. Using Σ_ν and an optical depth through the cloud $\tau_\nu = -\int_0^z \kappa_\nu \, dz$ the radiative transport equation from above can be simplified to:

$$I_{\nu,\text{observed}} = I_{\nu,\text{background}} \, e^{-\tau_\nu} + \Sigma_\nu (1 - e^{-\tau_\nu}). \quad (3)$$

This equation shows that the absorption in a molecular cloud is proportional to $e^{-\tau_\nu}$ and the emission proportional to $(1 - e^{-\tau_\nu})$.

Line emission

Besides recombination lines of hydrogen and helium the IR exhibits a broad variety of fine structure lines from atoms and ions and rotational and vibrational lines of molecules. The line emission in the submillimeter and long-wavelength FIR is dominated by rotational transitions of simple diatomic and linear multiatomic molecules. They have a simple ladder of rotational energy states E_J depending on the rotational quantum number J and the rotational constant B(Hz) as (to first order):

$$E_J = hBJ(J+1). \quad (4)$$

Electric dipole transitions in molecules with permanent electric dipole moments occur at a frequency $\nu_J = 2BJ$. The maximum emission in the ladder populated at a temperature T_{rot} occurs approximately at $E_J = hBJ(J+1) \approx kT_{\text{rot}}$, or $J_{\max} \approx \sqrt{(kT_{\text{rot}})/(hB)}$. One of the most important molecules for millimeter and infrared astronomy is carbon monoxide (CO) with its lowest $J = 1$–0 rotational transition at 115 GHz (2.6 mm wavelength). In the submillimeter and FIR higher rotational transitions from warm (30 to 100 K) and hot (several 100 K) molecular gas can be observed. These transitions are excited via collisions of CO with H_2 molecules. The dipole line radiation of CO therefore indirectly traces the spatial distribution, density and temperature of molecular gas. Only a small fraction (in star-forming regions of entire galaxies typically one part in a million or less) of very hot molecular gas (a few 1000 K) can be observed via quadrupole radiation in ro(tational)-vibrational NIR lines from the H_2 molecule.

In case of ro-vibrational transitions of linear molecules that occur in the NIR domain the energy states are labeled by the vibrational (v) and rotational (J) quantum numbers

$$E(v, J) = \frac{h}{2\pi}\omega_0 \left(v + \frac{1}{2}\right) + hBJ(J+1). \quad (5)$$

The correction terms are due to centrifugal stretching and Coriolis coupling of the rotational and vibrational motions. The ro-vibrational lines in electric dipole transitions with $v \rightarrow v - 1$ and $J \rightarrow J \pm 1$ can be approximated via

$$\nu_{v,J\rightarrow J\pm1} = \omega_0 \pm 2hBJ(J+1). \quad (6)$$

An important band head of ro-vibrational CO line absorption is found in ~3000 K atmospheres of late-type stars in the NIR at a wavelength of 2.31 μm. It is used for instance to analyse stellar populations and kinematics in the nuclei of galaxies.

The line emission in the MIR and FIR is dominated by fine structure line transitions. In the optical thin limit the integrated line flux (erg s^{-1} cm^{-2}) of a fine structure line can be written as:

$$F_{\text{line}} = \frac{h\nu}{4\pi} \int_{\text{Beam}} \int_{\text{LOS}} n_u \, dl \, d\Omega. \quad (7)$$

Here LOS is the line of sight and $d\Omega$ the differential solid angle used to integrate the line emission over the telescope beam. An important fine structure line is the [CII] line at a wavelength of 158 μm. It is one of the brightest cooling lines of the interstellar medium and can contain up to almost 1% of the overall infrared luminosity of an entire galaxy. In figure 1 the infrared spectrum of the starburst galaxy M82 is shown.

Space missions

In the following some of the recent larger INFRARED SPACE MISSIONS are summarized briefly, both in terms of their technical definitions and scientific results.

IRAS

IRAS was launched on 25 January in 1983 with a telescope mounted in a liquid helium cooled cryostat. The telescope was a Ritchey–Chrétien design with a 0.57 m aperture. The mirrors were made of beryllium and cooled to approximately 4 K. During its 10 months of operation, and in addition to numerous pointed observations, IRAS surveyed more than 96% of the sky up to four times at four infrared bands centered at wavelengths of 12, 25, 60 and 100 μm with its detectors cooled to 3 K. The survey array consisted of 62 rectangular infrared detectors arranged in staggered rows such that any real point source crossing the focal plane as the satellite scanned the sky would be seen by at least two detectors in each wavelength band. The survey covered the sky in overlapping strips as the satellite was orbiting in its near-polar orbit, which precessed by about a degree each day. During some pointed observations, individual sources were mapped simultaneously at 50 and 100 μm using a cold internal chopper for flux reference. IRAS also contained a low-resolution spectrometer, which produced a relevant spectral atlas.

With 350 000 detected objects the IRAS mission increased the number of cataloged astronomical sources by almost 70% and has had a major impact on almost every area of astronomy. IRAS extended our knowledge on the evolution of external galaxies. The instruments on board the satellite detected ~75 000 starburst galaxies—galaxies which are extremely bright in the infrared due to their intense star formation. Many of these starburst galaxies showed signs of interaction with companion galaxies and have 'super winds' (galactic scale gaseous outflows) emerging from their centers due to the large number of supernova explosions which occur in these galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

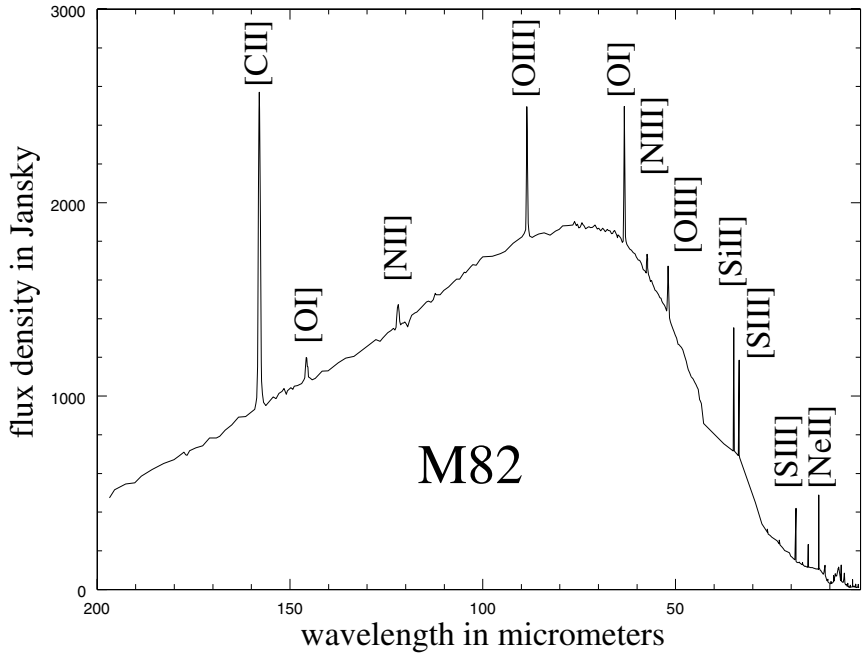


Figure 1. The infrared spectrum of the starburst galaxy M82 as taken with the ISO SWS and LWS spectrometer. In addition to the peak in the dust continuum emission, strong fine structure lines are indicated.

IRAS also found for the first time infrared cirrus: warm dust associated with diffuse clouds mostly at large galactic latitudes. IRAS discovered six new comets and found that comets in general are dustier than previously thought. The FIR observations provided evidence of zodiacal dust bands around other stars (e.g. Vega), and detected several probable protostars embedded in clouds of gas and dust.

ISO

The Infrared Space Observatory (ISO), (see figure 2) (Kessler *et al* 1996) of the European Space Agency (ESA) was an astronomical satellite, operating at wavelengths from 2.5–240 μm. Essentially, ISO consisted of a large cryostat which contained at launch about 2300 l of superfluid helium to maintain the Ritchey–Chrétien 60 cm diameter telescope, the scientific instruments and the optical baffles at temperatures of 2–8 K. ISO was launched by an Ariane 4 on 17 November 1995 and kept in operation for 29 months. The satellite spent roughly 16 h per day outside the radiation belts and during this time the sensitive FIR detectors could be operated. The four focal plane instruments on board ISO consisted of one short- and one long-wavelength spectrograph (SWS and LWS) and imaging experiments in the MIR (ISOCAM) and FIR (ISOPHOT) domain.

The SWS instrument consisted of two grating spectrometers, together covering the wavelength range between 2.38 and 45.2 μm. The overall spectral resolution covered a range of $R = 1000$ (300 km s^{-1}) to $R = 30\,000$ (10 km s^{-1}) by inserting Fabry–Pérot (FP) filters. In its grating mode, the LWS operated from 45 μm to 180 μm with a resolving power between a few 100 and 10 000 using FPs. The ISOCAM infrared imaging camera used

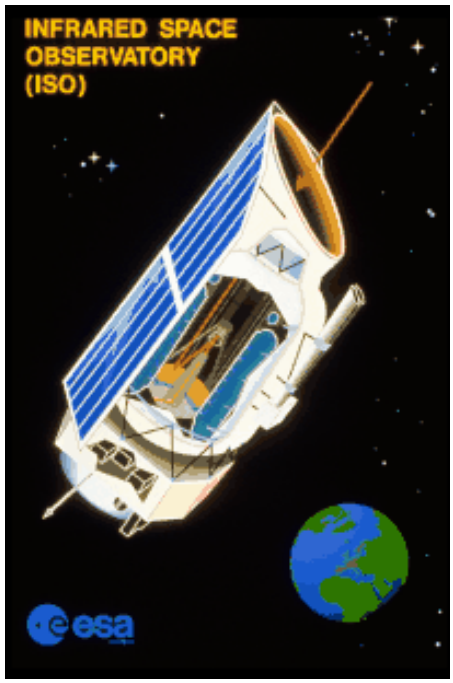


Figure 2. The Infrared Space Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

two 32 × 32 pixel detectors, one devoted to the 2.5–5 μm range and the other to the 4–17 μm range. Images in direct and polarized light at several spatial resolutions through a variety of spectral filters could be taken. ISOPHOT was an imaging photopolarimeter for the wavelength range 2.5 to 240 μm and included a low-resolution spectrometer for the 2.5 to 12 μm range.

ISO observations resulted in the first measurements of the lowest rotational transition of molecular hydrogen H_2 at 28.1 μm which is important in determining the distribution and abundance of cold molecular material directly. ISOCAM observations determined the distribution of warm dust near star-formation regions in external galaxies. Via 200 μm imaging, ISO detected 5 to 10 times more 15 to 21 K cold dust than IRAS. LWS and SWS observations of luminous galaxies allowed the determination of the relative contribution of starburst and non-thermal activity in their nuclei.

COBE

The COBE satellite was developed by NASA's GODDARD SPACE FLIGHT CENTER to measure the diffuse infrared and microwave 2.7 K cosmic background radiation from the early universe (due to the Big Bang). It was launched on 18 November 1989 and carried three instruments, a far-infrared absolute spectrophotometer (FIRAS) to compare the spectrum of the microwave background with that of a black-body, a differential microwave radiometer (DMR) to map the cosmic radiation sensitively, and a diffuse infrared background experiment (DIRBE) to search for the cosmic infrared background radiation. The cosmic microwave background spectrum was measured with a precision of 0.005%; for the first time the background was found to have an intrinsic anisotropy, at a level of 1 part in 100 000. Absolute sky brightness maps from 1.25 μm to 240 μm were obtained and the background radiation was detected in the two longest DIRBE wavelength bands at 140 and 240 μm wavelengths.

NICMOS

NICMOS (near-infrared camera and multiobject spectrometer) was an instrument which has been installed in the HUBBLE SPACE TELESCOPE (HST) in orbit and allowed it to observe in the near-infrared. The camera was a second-generation Hubble instrument installed aboard the satellite during a servicing mission in February 1997. NICMOS provided the capability of both imaging and spectroscopy at wavelengths between 0.8 and 2.5 μm. NICMOS contained three cameras designed for simultaneous operation. The optics presented the detectors with three adjacent, but not spatially contiguous, fields of view of different image scales. NICMOS employ three low-noise, high-quantum-efficiency, HgCdTe 256 × 256 element focal plane arrays in a passive dewar using solid nitrogen as a coolant. The camera contained a variety of filters, grisms (a combination of a grid and a prism to allow for dispersion on the optical axis) and polarizers selected through three independent filter wheel mechanisms. Slitless imaging spectroscopy in the widest field camera provided a multiobject spectrographic capability with a resolution of ∼200. NICMOS achieved diffraction-limited performance at 1.0 and 1.75 μm.

Sensitive near-infrared images have been obtained of solar system planets, the central regions of nearby galaxies with ACTIVE GALACTIC NUCLEI (AGN)—like the nearest radio galaxy Centaurus A—or violent star formation activity as in Arp 220. NICMOS also provided a detailed stellar census of the stars in obscured or embedded star clusters like the ORION NEBULA and the galactic center. A comparison of long exposures taken toward selected regions of the so-called HUBBLE DEEP FIELD (HDF) allowed us to improve our knowledge on the spectral energy distribution of the most distant galaxies in the universe.

Sources

In general, dust-embedded sources or dense molecular clouds heated by stars or active galactic nuclei are most luminous in line and continuum radiation in the infrared. In galaxies the luminosity is concentrated toward the nuclei and some active star-forming regions mostly within spiral arms or in barred spirals at the tips of the bars. Ultraluminous galaxies belong to the most luminous objects in the universe and can exhibit more than 10^{12} (a million million) solar luminosities in the infrared. As typical examples of infrared bright objects the nuclear region of the Milky Way (the galactic center), ultraluminous galaxies and star-forming regions are discussed in the following.

Galactic center

Photospheric emission of stars in the galactic center at a distance of 8 kpc (26 400 light years; one parsec (1 pc) equals 3.3 light years) can only be observed in the near-infrared. The visible light from the central stellar cluster is obscured by about 30 magnitudes (one magnitude is a factor of 2.5 in intensity) due to the dust and gas in the galactic plane. High angular resolution, near-infrared imaging and spectroscopy have made it possible in the last few years to measure stellar proper motion and radial velocities in the galactic center down to separations of less than five light days from the compact radio source SgrA* (in the constellation Sagittarius) at the dynamic center of the Milky Way (Eckart and Genzel 1997, Genzel *et al* 1997, Ghez *et al* 1999). These measurements make a compelling case for the presence of a compact, central dark mass of 2.6×10^6 solar masses which is most likely in the form of a supermassive black hole.

The first indications for a central mass concentration in the Milky Way (for a summary see Genzel *et al* 1994) emerged in the late 1970s from spectroscopic observations of a mid-infrared fine-structure line of Ne^+. As radio interferometric observations had discovered a compact, non-thermal radio source, SgrA*, in the same region, a plausible interpretation—by analogy to quasars—was that the large gas velocities indicate orbital motions in the vicinity of a million solar mass black hole, coincident with SgrA*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

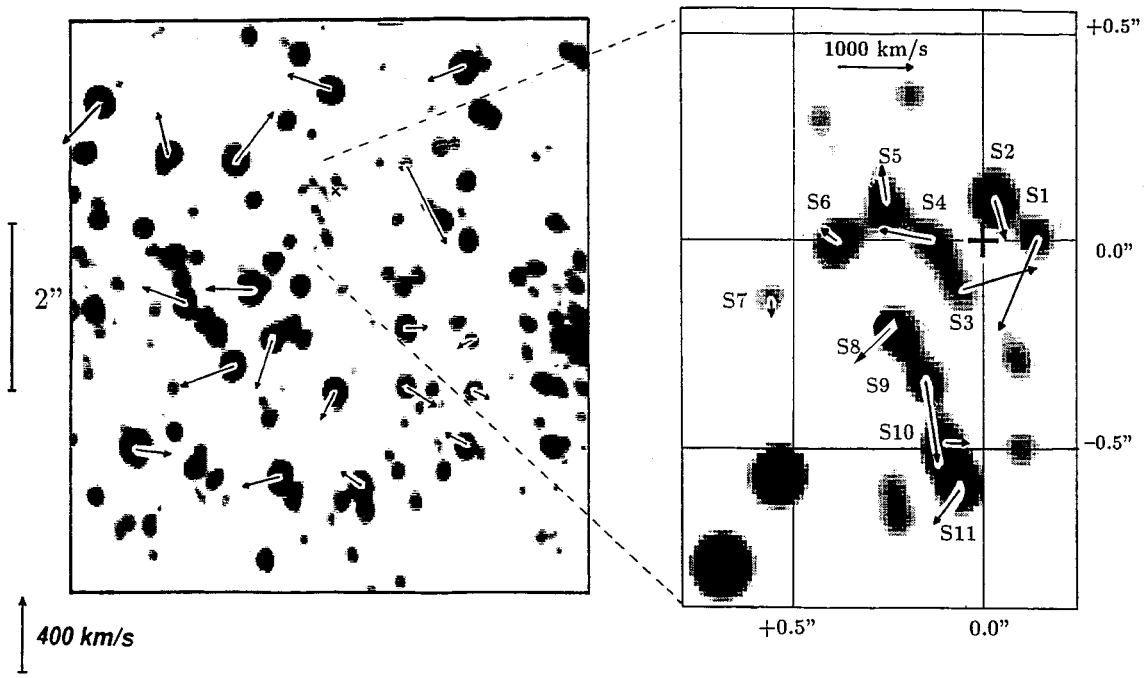


Figure 3. Proper motion vectors of the central sources close to Sgr A* at the galactic center.

In the following years further infrared (and radio) spectroscopic data including radial velocities of stars strengthened the gas dynamic evidence for a dark 1 to 3×10^6 $M_\odot$ central mass. Speckle interferometry revealed a compact 1″ diameter cluster of stars at the position of Sgr A*. Several of these stars in this so-called SgrA* cluster show proper motions in excess of 1000 km s^{-1} (see figure 3). The analysis shows that the projected velocity dispersion of stars increases towards the center as expected in the potential of a large central point mass (a 'Kepler law'). The location of the largest stellar velocities (the dynamic center), the stellar density maximum and the position of SgrA* all agree to within ±0.004 pc. The final distribution of enclosed mass as a function of true radial separation from SgrA* is shown in figure 5 of the article on SUPERMASSIVE BLACK HOLES IN INACTIVE GALAXIES. Simple physical considerations show that this central object with such a high mass and density cannot be stable and must be present in the form of a supermassive black hole. Together with the nucleus of NGC4258 (Greenhill *et al* 1995, Myoshi *et al* 1995)—in which very long baseline radio interferometric observations discovered a rapidly rotation circumnuclear disk—the galactic center currently is the very best case for a supermassive black hole.

Ultraluminous infrared galaxies

The extragalactic infrared background radiation and the identification of individual FIR galaxies (results from COBE, ISO, SCUBA) is starting to tell the history of star formation in the early universe. The IRAS satellite discovered a class of ultraluminous infrared galaxies which have the luminosity of quasars and show obvious signs of nuclear activity and advanced tidal interactions (Sanders *et al* 1988). Ultraluminous infrared galaxies (ULIRG) represent an important stage in the evolution of a wide variety of extragalactic objects including powerful nuclear starbursts, radio galaxies and quasars. They may also represent a primary stage in the formation of elliptical galaxy cores via the merger of two gas-rich spirals. ISO could contribute significantly to the investigation of ULIRGs. The ISO SWS and ISOPHOT mid-infrared spectroscopic survey of ULIRGs showed that 70–80% of them are predominantly powered by recently formed massive stars (Genzel *et al* 1998). However, about 20–30% are powered by a central AGN. These conclusions are based on an infrared 'diagnostic diagram' involving the ratio of high- to low-excitation mid-IR emission lines and the strength of the 7.7 μm PAH (polycyclic aromatic hydrocarbon) feature. At least half of the sources probably have simultaneously an active nucleus and recent or on-going starburst activity in a 1–2 kpc diameter circumnuclear disk or ring. No obvious trend for the AGN component to dominate the most compact, and thus most advanced, mergers was found. Instead, at any given time during the merger evolution, the time-dependent compression of the circumnuclear interstellar gas, the accretion rate onto the central black hole and the associated radiation efficiency may determine whether star formation or AGN activity dominates the luminosity of the system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star formation process

The investigation of the star-formation process profits significantly from infrared and submillimeter observations (Beckwith and Sargent 1996). Although the solid particles probably make up no more than about 1% of the disk mass they produce almost 100% of the radiation between 2 μm and $\sim$1 mm wavelength. The range of strong emission from the NIR to the radio is the result of a wide variation of temperatures, from $\sim$1000 K very close to the stars to $\sim$30 K near the outer edge of the disk at several hundred AU distance from the center. This makes it relatively easy to find disks at infrared wavelengths. The silhouettes of these disks are seen by the HST against the bright background of the Orion nebula. They show flattened, elliptical shapes and demonstrate for the first time that disks are truncated sharply at their outer radii.

Disks are very common surrounding many young stars. Half of the classical T Tauri-type stars in the Taurus–Auriga and Ophiuchus dark clouds show evidence for disks. An important new finding is that disks are common in dense clusters of young stars such as at the center of the Orion nebula (Lada *et al* 1993).

In general one discriminates between four stages of star formation (Shu *et al* 1987): first dense cores form in molecular clouds while magnetic and turbulent support is lost through diffusion of neutral particles. Then a protostar with a surrounding nuclear disk is formed via a collapse of the dense cores from inside out. Following this stage a stellar wind breaks out along the rotational axis of the system and creats a bipolar outflow. Finally the infall is terminated and a young star with a circumstellar disk is revealed.

Star-forming regions

Infrared observations are especially well suited for observations of gas clouds in active star-forming regions within our own galaxy such as the Trapezium region in the Orion nebula, M17 or M49 (see e.g. Yun and Liseau 1998). The UV radiation of clusters of recently formed O- and B-stars interacts in so-called photon-dominated regions (PDR) with the surrounding dense molecular gas at a distance of only a fraction of a parsec. By heating the gas, about 0.5% of the strong UV radiation is converted into bright infrared line emission. The investigation of these lines allows us to determine the density and temperature structure as well as abundances and chemical processes in these interface regions.

The ISO SWS observations allowed detailed investigations of interstellar ices and for the first time an unbiased overview of their compositions. The results include observations of solid CO_2, solid CH_4 and HCOOH. In the gas phase important information has been obtained on the abundance of water in warm star-forming regions, through the study of rotation vibrational lines with SWS and pure rotational lines with LWS. The gas to solid state ratios are important for investigations of the gas-grain chemistry in dense clouds in star-forming regions.

Bibliography

Alloin D M and Mariotti J-M 1994 *Adaptive Optics for Astronomy* (Dordrecht: Kluwer)

Beckwith V W and Sargent A I 1996 *Nature* **383** 139

Downes D 1989 Radio Astronomy Techniques *Evolution of Galaxies—Astronomical Observations* ed I Appenzeller, H Habing and P Léna (Heidelberg: Springer)

Eckart A and Genzel R 1997 Stellar proper motions in the central 0.1 parsec of the galaxy *Mon. Not. R. Astron. Soc.* **284** 576

Genzel *et al* 1998 What powers ultra-luminous IRAS galaxies? *Astrophys. J.* **498** 579

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 On the nature of the dark mass in the center of the Milky Way *Mon. Not. R. Astron. Soc.* **291** 219–34

Genzel R, Hollenbach D J and Townes C H 1994 *Rep. Prog. Phys.* **57** 417

Ghez A, Klein B, Morris M and Becklin E 1998 *Astrophys. J.* **509** 678

Kessler M F, Steinz J A, Anderegg M E *et al* 1996 *Astron. Astrophys.* **315** L27

Greenhill L *et al* 1995 *Astrophys. J.* **440** 619

Lada C J, Young E T and Greene T P 1993 *Astrophys. J.* **408** 471

Low F J and Rieke G H 1974 *Methods Exp. Phys.* **12A** 415

Myoshi M *et al* 1995 *Nature* **373** 127

Osterbrock D E *Astrophysics of Gaseous Nebulae* 1974 (San Francisco: Freeman)

Rieke G H 1994 *Detection of Light: From the Ultraviolet to the Submillimeter* (Cambridge: Cambridge University Press)

Sanders D B, Soifer B T, Elias J H, Madore B F, Matthews K, Neugebauer G and Scoville N 1988 *Astrophys. J.* **325** 74

Shu F H, Adams F C and Lizano S 1987 *Ann. Rev. Astron. Astrophys.* **25** 23

Spitzer L 1978 *Physical Processes in the Interstellar Medium* (New York: Wiley)

Tyson R K 1991 *Principles of Adaptive Optics* (Academic: San Diego)

Watt G D and Webster A S 1990 *Submillimetre Astronomy* (Dordrecht: Kluwer)

Wall J V and Boksenberg A 1990 *Modern Technology and its Influence on Astronomy* (Cambridge: Cambridge University Press)

Yun J L and Liseau R 1998 *Star Formation with the Infrared Space Observatory (ISO)* (San Fransisco: Astronomical Society of the Pacific)

Andreas Eckart

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared Astronomy Satellite

The Infrared Astronomical Satellite (IRAS), a collaboration between the United States, the United Kingdom and the Netherlands, was the first space mission aimed at a systematic all-sky survey at infrared wavelengths inaccessible from the ground. IRAS surveyed more than 96% of the sky at 12, 25, 60 and 100 μm (micrometers) with unprecedented sensitivity, revolutionized INFRARED ASTRONOMY and made significant contributions to almost every area of astrophysical research from solar system studies to cosmology.

The IRAS telescope (figure 1, made of beryllium, had a 57 cm primary mirror and was kept at about 4 K inside a cryostat cooled by liquid helium. The focal plane of the telescope carried 62 detectors, divided about equally among the four survey wavelengths, and arranged in a staggered array to collect data for the all-sky survey as the telescope scanned the sky. The detectors subtended from 0.5 to 5 arcminutes each on the sky and allowed detections down to 0.1–1 Jy depending on wavelength. The focal plane also carried a Low Resolution Spectrometer (LRS) covering from 7.5 to 23 μm with a spectral resolution of about 20 and a Chopped Photometric Channel (CPC) with one detector at each of 50 and 100 μm, as well as visible star sensors used for pointing information.

IRAS was launched on 25 January 1983 into a 900 km, 103 min Earth orbit which remained in a plane nearly normal to the Earth–Sun line. As the satellite traveled on its Sun-synchronous orbit, the telescope pointed away from the Earth and scanned methodically so the image of the sky traveled across the focal plane, and sources were sensed by the detectors. The resulting data were calibrated and processed at IPAC to catalog over 370000 infrared sources and to generate all-sky images at a resolution of a few arcminutes and spectra of almost 3000 infrared sources (see figure 2). Part of the mission was also used to conduct deeper observations of preselected targets (see figure 3). The IRAS cryostat ran out of helium in November 1983, rendering the focal plane useless and ending the mission.

IRAS detected six new comets, measured the emission from 2000 asteroids and discovered the zodiacal dust bands which trace the dust generated in collisions among asteroids. IRAS also detected more infrared emission from stars than could be attributed to the stars themselves, evidence of dust around those stars analogous to zodiacal dust around the Sun and potentially an indica-

Figure 1. An artist's rendering of IRAS in its 560-mile-high, near-polar orbit above the Earth. Credit: Infrared Processing and Analysis Center (IPAC), Caltech/JPL. IPAC is NASA's Infrared Astrophysics Science and Data Center.

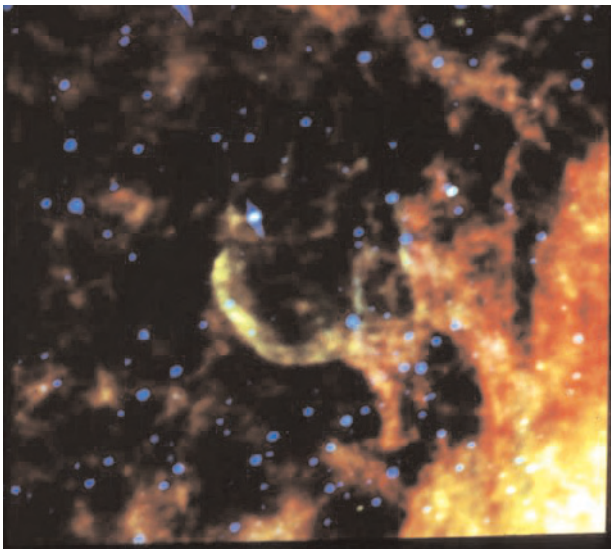


Figure 3. The Loop (or Veil) Nebula in Cygnus is thought to be the remnant of a supernova outburst that occurred over 15 000 yr ago. This IRAS image shows the infrared emission from dust grains in the expanding shell of the remnant. The image has been color coded to display relatively warm infrared emission as blue, cooler emission as green and the coldest emission as red. The blue dots are stars, and the diffuse emission to the right of the Loop is from interstellar dust grains heated by starlight. Credit: IPAC, Caltech/JPL.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

tion of planetary systems; VEGA and BETA PICTORIS are especially well-studied cases. IRAS mapped in detail the density and temperature in molecular clouds where stars form, revealed their full energy balance and found thousands of cores hosting protostars. IRAS revealed a full view of the bulge of the Milky Way and discovered diffuse infrared emission from dust in our galaxy, bright and warm towards the plane and wispy and cold above the plane, hence the name 'infrared cirrus'. Cirrus emission is a valuable source of information on the interstellar medium and a potential impediment to infrared source measurement in certain lines of sight. It was used to estimate the amount of visible light extinction by dust for each direction on the sky and to find the best windows for observing outside the Milky Way at all wavelengths.

IRAS detected a total of 75 000 galaxies and quasars whereas less than a hundred had been observed as infrared sources before. This was the first truly unbiased, unobscured survey of extragalactic objects, allowing for the analysis of their infrared colors, of their distribution in space and of their evolution over the past few billion years. IRAS discovered a new class of galaxies with tens of times more luminosity emerging in the infrared than in the visible and with overall luminosities comparable with quasars. These 'ultraluminous infrared galaxies' arise most often from collisions or mergers between normal disk galaxies and may turn into quasars after the infrared-emitting dust has dissipated; Arp 220 and IRAS F10214+4724 are especially well studied cases. At lower luminosities, IRAS data allowed detailed analysis of the STARBURST phenomenon, due to intense star formation in extremely dense gas at the centers of galaxies. IRAS has also made certain types of objects famous, for example Arp220 and F10214.

Bibliography

The IRAS home page can be found at
http://www.ipac.caltech.edu/ipac/iras/iras.html

G. Helou

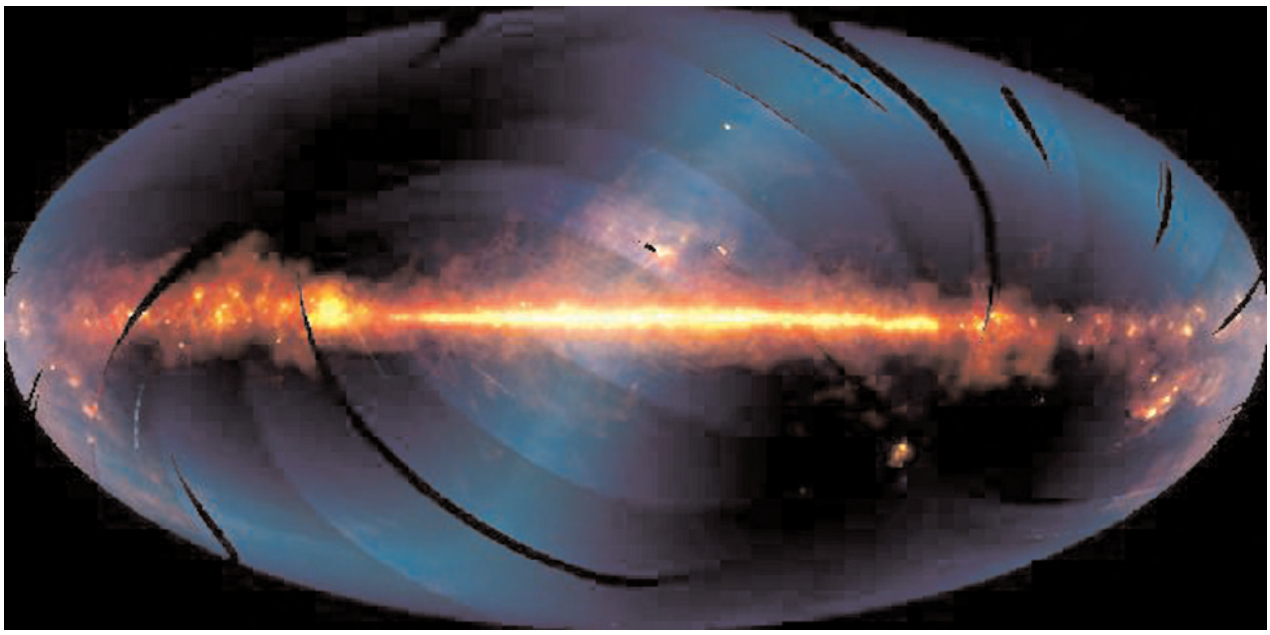


Figure 2. Nearly the entire sky, as seen in infrared wavelengths and projected at 0.5° resolution, is shown in this image, assembled from 6 months of data from IRAS. The bright horizontal band is the plane of the Milky Way, with the center of the Galaxy located at the center of the picture. The colors represent infrared emission detected in three of the telescope's four wavelength bands (blue is 12 μm; green is 60 μm and red is 100 μm). Hotter material appears blue or white while the cooler material appears red. The hazy, horizontal S-shaped feature that crosses the image is faint heat emitted by dust in the plane of the solar system. Celestial objects visible in the photograph are regions of star formation in the constellation Ophiucus (directly above the galactic center) and Orion (the two brightest spots below the plane, far right). The Large Magellanic Cloud is the relatively isolated spot located below the plane, right of center. Black stripes are regions of the sky that were not scanned by the telescope in its first six months of operation. Credit: IPAC, Caltech/JPL.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Infrared Imagers

In contrast to visual observations at wavelengths of 0.35–0.7 μm with their long history of astronomical discoveries, the only celestial source detectable in the infrared region ($\lambda > 1$ μm) by unaided human senses is the Sun. Consequently, the development of INFRARED ASTRONOMY had to wait for technical detection methods. The existence of infrared radiation was first demonstrated by Herschel, using thermometers to show that energy is being received in the invisible parts of the solar spectrum. Sophisticated versions of thermometers, so-called bolometers, are still the detector of choice at very long (submillimeter) wavelengths. At shorter infrared wavelengths, photon-detecting semiconductor detectors became available shortly after WWII. Detector systems built around these devices were only single-point photometers, and spatial images had to be obtained by sequentially measuring the signal at each position in the image. The first true infrared imagers were built in the 1980s when INFRARED ARRAYS became available for astronomy.

Problems of ground-based infrared observations

Several factors combine to make ground-based infrared observations more difficult than observations in the optical (0.3–1.0 μm) region. While the atmosphere in the optical is of uniformly high transmission, the infrared region is subdivided into windows of atmospheric transmission, separated by spectral bands where absorption by atmospheric trace gases such as CO_2 and H_2O renders the atmosphere opaque. Since the water vapor content of the atmosphere strongly depends on the climatic conditions and since the scale height of water in the atmosphere is small, infrared astronomical observations are best carried out from dry, high-altitude sites.

Ambient temperature ($T \approx 300$ K) surfaces emit copious photons at wavelengths longer than about 2 μm. Therefore, the emissivity (e) of the telescope optics must be minimized to reduce the photon background contributed by the telescope itself. While high reflectivity (r) of the mirrors, and thereby low emissivity ($e = 1-r$), is desirable for any telescope, high emissivity reduces the signal-to-noise ratio in the infrared much faster than poor reflectivity does in the optical. Dedicated INFRARED TELESCOPES are designed to minimize the area of obstructions by black, highly emitting surfaces in the lightpath (e.g. secondary support spiders, the secondary mirror mount and the hole in the primary mirror). The technologies have been developed to coat at least the secondary mirror, if not also the primary, with silver, to significantly reduce the emissivity compared with the more traditional aluminum coatings.

Infrared detectors are sensitive to the radiation emitted by ambient temperature surfaces. Simply enclosing the detector with black surfaces does therefore not eliminate the background radiation. In addition to cooling the detector itself to its operating temperature, an infrared camera requires all surfaces surrounding the detector, including the optics, to be cooled, using liquid cryogens such as liquid nitrogen or liquid helium, or by closed-cycle cooler systems. As a consequence, infrared astronomical instruments tend to be far more complicated and costly than their optical counterparts.

The design of infrared imagers

Various optical designs are possible for an infrared imager. The choice of a specific design is driven by the requirements of field of view, pixel scale, image quality, throughput and wavelength range as scientific criteria, but also by budget and schedule considerations. Further, a distinction is often made between investigator-class instruments, which are designed to primarily serve the specific scientific interest of their designer, and facility instruments, which are intended to serve a wide user community for a variety of scientific projects. Facility instruments typically offer a larger number of operating modes, have more elaborate software to allow occasional users to operate them efficiently, are generally built to higher standards of reliability and serviceability, and are better documented. Not surprisingly, they also tend to be far more expensive than investigator-class instruments and take much longer to design and build.

A specific example

Figure 1 illustrates the very basic elements of a simple infrared imager design and shows the optical path for three field points. Figure 2 gives more details of the specific technical implementation chosen for the Quick Infrared Camera (QUIRC) built at the University of Hawaii for their 2.2 m telescope. When this camera was built in 1994, it was the first to use a detector array of 1024×1024 format.

A dewar window of infrared-grade fused silica seals the inside of the cryostat and allows it to be evacuated in preparation for cooling of the interior components of the infrared camera with liquid nitrogen. The dewar window is the last optical element of an infrared camera that remains at ambient temperature. Therefore, its emissivity must be kept low and scattering of thermal radiation into the lightpath of the camera must be minimized. These requirements dictate the choice of a high-transmission material and good surface polish.

The first cooled optical element is a field lens located in or close to the focal plane of the telescope. The field lens produces an image of the telescope pupil, which is the optical surface defining the aperture of the telescope. Many infrared telescopes are designed with undersized secondary mirrors, making the secondary the pupil in the system. However, the UH 2.2 m telescope is a classic optical wide-field telescope, so the $f/10$ secondary mirror is oversized and therefore the primary mirror is the pupil of the telescope. The field lens thus produces an image of the telescope primary, seen reflected in the secondary mirror, on a surface inside the cryostat where a cold stop can be placed. Since a field lens is located close to the focal plane of the telescope its optical aberrations do not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

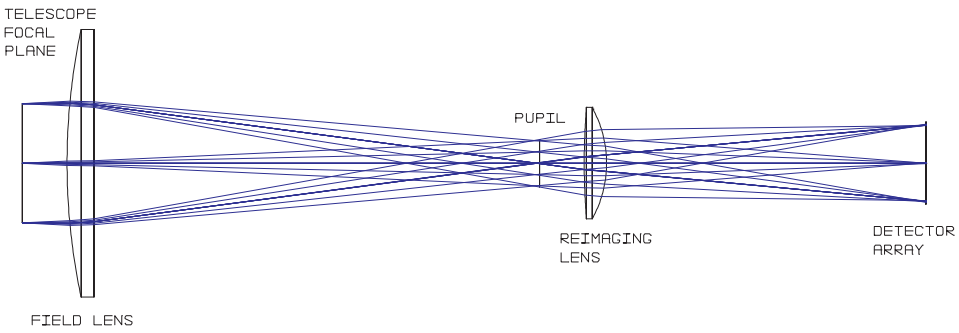


Figure 1. Basic elements of a simple infrared imager design.

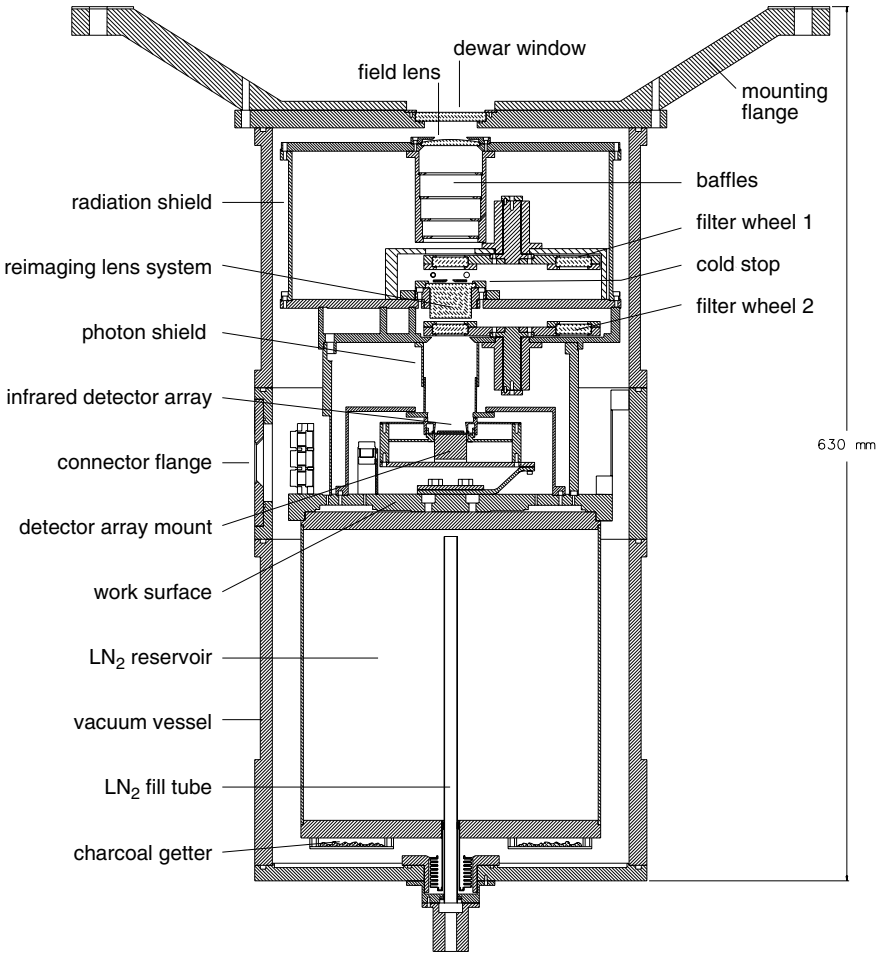


Figure 2. Infrared imager implementation for QUIRC.

significantly affect the image formed on the detector. Field lens aberrations do, however, affect the quality of the pupil image and thereby our ability to mask off unwanted thermal radiation. In a simple camera like QUIRC, the field lens is usually a single spherical lens made of a high refractive index material (ZnSe or ZnS), so that spherical aberration is minimized. A better, but more expensive, solution would be to use an aspheric lens.

The next important element is the pupil cold stop. Its purpose is to block all radiation that does not come from the telescope pupil (the primary mirror in our case) from reaching the detector. In particular, radiation coming from the mirror mount just outside of the mirror's reflective surface and radiation emitted by the central hole in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cassegrain primary mirror must be blocked. The cold stops are typically designed as an aperture slightly smaller than the pupil image, with a central occulting disk masking the image of the hole in the primary. The central occulting disk is usually suspended by wires that also act to mask the support 'spider' arms holding the telescope secondary mirror. By undersizing the hole in the cold stop, the cold stop effectively becomes the pupil of the combined telescope and camera system.

Infrared imagers use interference filters to select the spectral region of interest. Usually one or more filter wheels are located near the pupil stop of the optical system. When placed near the pupil (figure 1), light from all field points passes through the same area of the filter so that non-uniformities in the filter affect all objects in the field in the same way and can therefore be more easily corrected.

The main optics of the infrared camera are located just behind the pupil stop. Their purpose is to reimage the telescope focal plane onto the detector array. By choosing the focal length and the distances, these reimaging optics can be used to magnify or demagnify the original focal plane image, depending on the desired pixel scale. Usually, combinations of two or more lens materials are used to correct chromatic aberrations. For example combinations of BaF_2 and Schott SF6 glass work well in the near infrared, while combinations of BaF_2 and LiF, sometimes combined with other, higher-index materials such as ZnS and ZnSe work well at wavelengths out to 5.5 μm.

The final and most important element in the infrared camera is the focal plane detector array that collects the photons and allows the signal to be measured. Detector arrays of up to 1024 $\times$ 1024 pixels are now available for near-infrared observations (1.0–5.5 μm) and arrays in 2048 $\times$ 2048 format are under development. Infrared arrays for mid-infrared wavelengths typically have smaller formats (see MID-INFRARED INSTRUMENTATION).

The refractive optical design outlined above is by no means the only possible solution. Purely reflective designs have been developed, in particular for mid-infrared cameras where the choice of suitable refractive materials is even more restricted than in the near infrared. One particular, surprisingly simple reflective design is the Offner relay (figure 3), consisting of a large, concave spherical mirror and a small, convex spherical mirror of double the curvature of the large mirror. The two mirrors are mounted concentrically. In this configuration, the Offner relay reimages the focal plane of the telescope onto the detector while forming a pupil image on the small secondary mirror. This simple optical arrangement is free of all first-order aberrations and offers high throughput, but it is limited to exactly reproducing the image scale in the telescope focal plane. Further, the quality of the pupil image is not very good, but can be corrected with additional optical elements.

Additional capabilities

Most infrared imagers, in particular facility-class instruments, include additional optical elements to expand their capabilities beyond pure imaging.

Polarimetry

Infrared imaging polarimeters usually use a fixed, cooled analyser in the instrument and rotate the polarization plane by an external warm rotatable half-wave plate that emits very little thermal radiation owing to its low emissivity. Wire-grid analysers transmit one polarization plane and absorb the other. Their advantage is that they can be used in the same way as filters, and therefore allow the implementation of polarimetric capability at very low system cost. Beamsplitting elements such as Wollaston prisms split the light into two orthogonal polarization planes, allowing the simultaneous detection of both planes as two separate images of the object. This is obviously a more efficient use of the photons coming from the object and has the added advantage that any instabilities in the atmospheric transmission during the exposure time affect both polarization planes in the same way, making the observation immune to such atmospheric effects. The disadvantage is that Wollaston prisms require a collimated light path and thus dictate a slightly more complex optical design.

Spectroscopy

Spectroscopic capability can be implemented in an infrared imager by using a grism (an integrated grating and prism combination) in a collimated light path near the pupil of the system. Except for rare slitless use, spectroscopy usually requires a slit to be placed in the focal plane of the telescope.

Coronagraphic observations

A special case of imaging observations are coronagraphic observations. In this context these are not observations of the solar corona but of faint features and objects near bright stars, such as faint binary companions, including brown dwarfs, faint dust-scattering features such as remnants of protoplanetary disks, and faint emission line features associated with stellar jets. What these observations have in common with observations of the solar corona is that careful attention must be paid to reducing the scattered light from the bright central object. Infrared cameras already have a pupil stop (called a Lyot stop in the context of coronagraphs). With the addition of an occulting mask in the telescope focal plane, proper attention to stray light issues in the optical design (avoiding ghost reflections) and potentially the use of special cold stop masks, infrared imagers can become very capable CORONAGRAPHS.

Data reduction

Most near-infrared detector arrays used today provide the option of non-destructive reading, i.e. reading the accumulated image information in the array without destroying this information. After resetting a detector

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

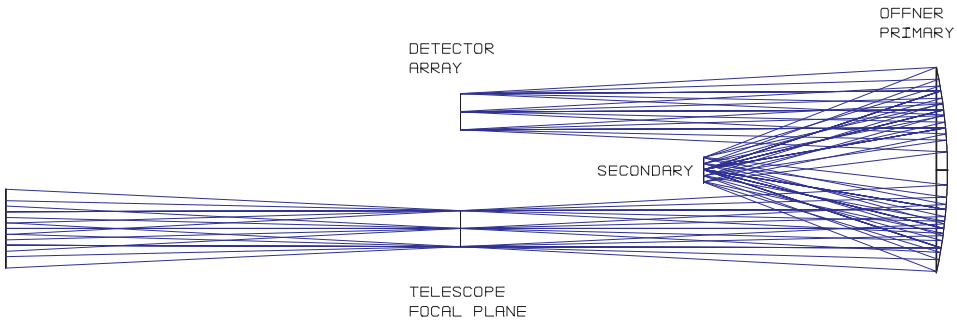


Figure 3. The Offner relay.

array to erase all previous imaging information, a typical double-correlated frame consists of two reads of the detector array: one read at the beginning of the integration time and then one read at the end. The difference of these two reads represents the signal accumulated during the integration time. While this double-correlation technique eliminates many of the artifacts from the detector array and the readout electronics, such a single frame is usually not a useful astronomical image. Two additional steps are required, at minimum, to extract and calibrate the astronomical information.

The first is usually called 'sky subtraction', but it removes more than just uniform background from the sky. For example, a dust particle on the cryostat window emits infrared radiation that is clearly not coming from the sky but is superposed on the astronomical signal. The most straightforward way of subtracting the sky would be to take an exposure of a blank piece of sky close to the object and subtract that from the object frame. While this method was commonly used in the early days of infrared imaging, the sensitivity of today's infrared imagers is so high that there is virtually no place in the sky that is really 'blank', i.e. devoid of detectable objects. Therefore, a sky frame needs to be computed from a large number of images taken at different telescope positions close to and including the position of the object of interest. All features that remain stationary in such a set of images are instrumental artifacts, such as the dust particle on the window discussed earlier. Computing a median through such a stack of frames eliminates all objects that changed their position between exposures, i.e. all astronomical sources, and preserves all stationary features. The resulting 'median sky' can then be subtracted from each frame individually, to remove instrumental artifact emission and sky background from the frames.

The second step is to correct for non-uniformities in the pixel response across the array. One way to measure this so-called 'flat field' is to take images of a uniformly illuminated surface. The inside of a telescope dome, which is quite uniformly painted and out of focus to the telescope, is usually suitable for this calibration measurement. The difference between an exposure of the dome with incandescent lights on and a second exposure with the lights off gives the response of the detector array to radiation of the few thousand kelvin temperature and eliminates artifacts from ambient temperature emission from the telescope optics. A flat 'field frame' obtained this way is then used to divide the individual astronomical frames by correcting the non-uniformities in the detector array. The resulting sky-subtracted and flat-fielded frames are then added together, compensating for the telescope motions between frames to form a 'mosaic' image.

Bibliography

The infared camera QUIRC used here as an example is described in more detail in

Hodapp K W *et al* 1996 *New Astron.* **1** 177

A good introductory textbook for modern instrumentation technology, including infrared imagers, is

McLean I S 1997 *Electronic Imaging in Astronomy* (New York: Wiley)

The technology of infrared imagers is evolving rapidly and so is the literature describing this progress. The largest body of information is found in the Proceedings of the International Society for Optical Engineering (SPIE). A number of papers in the latest SPIE Proceedings devoted mainly to astronomical instrumentation describe the current state of the art in infrared imager design:

Hodapp K W *et al* 1998 Gemini near-infrared imager *Proc. SPIE* **3354** 545
Lenzen R *et al* 1998 CONICA: the high-resolution near-infrared camera for the ESO VLT *Proc. SPIE* **3354** 606
Tokunaga A T *et al* 1998 Infrared camera and spectrograph for the Subaru telescope *Proc. SPIE* **3354** 512
Wishnow E H *et al* 1998 Long-wavelength infrared camera (LWIRC): a 10 μm camera for the Keck telescope *Proc. SPIE* **3354** 591

Klaus Werner Hodapp

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared Space Missions

Infrared radiation occurs at longer wavelengths than visible light, in the region 1–250 μm. Although cosmic near-infrared radiation can penetrate Earth's atmosphere through narrow 'windows', making ground-based observations possible, longer wavelengths are absorbed by the air. Observations at these wavelengths are possible only through space-based observatories.

Preliminary studies using instruments on sounding rockets, the Soviet Salyut 4 space station and the Kuiper Airborne Observatory took place in the 1970s and paved the way for the first infrared sky survey satellite.

The US–Netherlands–UK INFRARED ASTRONOMY SATELLITE (IRAS) was launched into a 900 km Sun-synchronous orbit on 25 January 1983. It carried a Ritchey-Chrétien telescope, with a 57 cm primary mirror, which was cooled to 2 K by liquid helium. During its operational lifetime of 300 days, IRAS surveyed 98% of the sky at 12, 25, 60 and 100 μm. Some 250 000 infrared sources were cataloged as the result of IRAS observations, compared with approximately 5000 previously known sources.

IRAS made many new discoveries within the solar system. Six new comets were found, including IRAS–Araki–Alcock, which passed close to the Earth in May 1983. Among the new asteroid finds was an unusual object, 3200 Phaethon, which may be an extinct comet. IRAS also revealed hitherto unsuspected dust trails associated with comets, as well as a torus of warm dust produced by collisions between Main Belt asteroids and associated with the Zodiacal Light.

Further afield, IRAS discovered that a number of stars, including Vega, which appeared much brighter than expected at infrared wavelengths, were surrounded by dust disks. Astronomers were also able to study cool interstellar dust clouds, newborn stars embedded inside nebulae and some 50 000 distant galaxies.

Infrared observations were made subsequently during two shortlived space missions. A 15 cm infrared telescope was flown on the Spacelab 2 mission (29 July–6 August 1985), but results were disappointing, largely due to high infrared background radiation from the Space Shuttle.

An instrument of similar size, known as the Infrared Telescope in Space, was placed on the Japanese–US Space Flyer Unit, a small orbital platform which was launched by an H-2 rocket on 17 March 1995. It operated for a few weeks until its supply of helium was exhausted, then remained dormant until the other onboard experiments were completed and the satellite could be recovered by the Space Shuttle in January 1996.

An infrared mission of a very different kind operated in the period between these two brief flights. NASA's Cosmic Background Explorer (COBE) carried the Diffuse Infrared Background Experiment, which studied the entire sky at ten wavelengths between 1 and 300 μm. Its 20 cm Gregorian telescope produced some of the best maps ever made of the infrared sky's overall structure, as well as the first map of the cumulative infrared emission from primordial galaxies which formed soon after the the Big Bang.

Also located inside the satellite's helium cryostat was the Far-Infrared Absolute Spectrophotometer, which studied the spectrum of the cosmic background radiation. Two months after launch, researchers announced that the spectrum of the Big Bang's afterglow matched that of a 2.73 K black body.

The true successor to IRAS was the European Space Agency's Infrared Space Observatory (ISO), though its mission was to focus on specific targets rather than carry out an all-sky survey. ISO was launched into a highly elliptical orbit on 17 November 1995.

Much larger than its predecessor, ISO carried a 60 cm Ritchey–Chrétien telescope, four scientific instruments and sufficient helium coolant to operate for 28 months. The camera, photometer and two spectrometers between them covered the full range of infrared wavelengths from 2 to 200 μm.

Data from ISO are still being analysed, but a number of significant initial results have been announced, including the discovery that water is much more widespread throughout the Galaxy than expected. ISO observations show that water is present in the atmospheres of Saturn's moon Titan and the outer planets, in planetary nebulae and in giant molecular clouds. The satellite was also well placed to study the prodigious dust and gas emissions from Comet Hale–Bopp.

The observatory has also revealed the first glows from concentric rings of cold dust in the Andromeda galaxy, the existence of complex organic molecules around young stars, and the first known infrared-bright gravitational arcs. Dozens of distant galaxies which contribute to the Cosmic Infrared Background have been detected.

Two more infrared observatories are currently planned. NASA's SPACE INFRARED TELESCOPE FACILITY (SIRTF) has undergone many changes since it was first proposed in the 1970s. Although its mass has been reduced to about 900 kg, and the amount of liquid helium it will carry is only 360 litres, its mission lifetime requirement remains 2.5 years.

In late 2001 or early 2002, SIRTF will be placed into an Earth-trailing heliocentric orbit which causes it to drift away from Earth at the rate of $\sim$ 0.1 AU per year. This choice of orbit has been preferred over a geocentric orbit for its more benign thermal environment. A 'warm-launch' scenario has also been chosen, whereby SIRTF's telescope assembly is launched at ambient temperature and allowed to cool passively in space. Only the focal-plane instruments and the compact cryostat are enclosed in a vacuum shell, resulting in a dramatic reduction in the volume of liquid helium required. The 85 cm diameter primary mirror is fabricated entirely of beryllium, and the telescope is a Ritchey–Chrétien design.

The Infrared Imaging Surveyor (IRIS)—originally designated as ASTRO-F—is the second infrared astronomy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mission of the Japanese Institute of Space and Astronautical Science (ISAS). Scheduled to be launched in February 2003, IRIS will be placed into a Sun-synchronous polar orbit at an altitude of 750 km by the ISAS M-V rocket. It will carry out an infrared sky survey with much better sensitivity than IRAS, and is expected to add significant information on many astrophysical problems such as the evolution of galaxies, formation of stars and planets, and the contribution of brown dwarfs to dark matter.

IRIS has a 70 cm Ritchey-Chrétien type telescope cooled to 6 K with liquid helium and Stirling-cycle coolers. It will carry two focal-plane instruments: a Far-Infrared Surveyor, which will survey the entire sky in the wavelength range from 50 to 200 μm with angular resolutions of 30–50 arcsec; and an Infrared Camera that will take deep images of selected sky regions in the near- and mid-infrared range.

Meanwhile, observational capabilities in the near-infrared will also be enhanced by the launch of the 8 m Next Generation Space Telescope (NGST) in 2007–8. Following on from the successful installation of the Near Infrared Camera and Multi-Object Spectrometer (NICMOS) on the HUBBLE SPACE TELESCOPE in 1997, NASA and ESA are looking towards building an observatory that will detect radiation at wavelengths of 0.6 to 20 μm (optimized for the 1–5 μm region). NGST will enable astronomers to study extremely distant objects whose radiation has been greatly redshifted when observed in the current epoch.

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared Telescopes

The infrared region of the spectrum covers a wide range of wavelengths—from 0.7 to 1000 μm. However, astronomers usually restrict the term 'infrared' to the more limited region of 1 to 350 μm—though even this is over 100 times wider than the visible region of the spectrum.

Infrared radiation was discovered in 1800 by William HERSCHEL, during his investigations of the visible light spectrum. When he placed a thermometer outside the red end of the rainbow spectrum, Herschel found that an invisible source of radiation caused the instrument to record a rise in temperature. He called the invisible rays 'infrared'.

We now know that the strongest infrared sources are relatively cool objects, such as red giant stars, nebulae, planets and comets, whose temperatures are between about 70 K and 3000 K. In contrast, the Sun is a fairly weak infrared emitter. With its surface temperature of about 5500 K, our nearest star emits most of its radiation at visible wavelengths. Herschel's discovery was only made possible because of the Sun's close proximity.

Little progress towards understanding the infrared universe was made until the invention of more sensitive, electronic detectors in the 1950s. Some notable exceptions were the detection of infrared emission from the Moon by Charles Piazzi SMYTH in 1856, and infrared observations of the Moon through its cycle of phases by the 4th Earl of ROSSE, who attempted to calculate its surface temperature. The first infrared observation of an object outside the solar system seems to have been made in 1878 by Thomas Edison, who observed the star ARCTURUS during a solar eclipse.

Not until 1969 was the first infrared sky survey undertaken by Gerry Neugebauer and Bob Leighton from Caltech. With the aid of lead sulphide detectors, their relatively crude 1.5 m (60 in) diameter telescope scanned 70% of the sky and detected 5612 cosmic infrared sources at 2.2 μm. More recently, detectors have been made from germanium doped with copper, mercury or gallium. Unfortunately, while these are able to probe to longer wavelengths, they require cooling to very low temperatures (2 K) in order to eliminate background interference.

Another problem to be overcome by ground-based observatories was the absorption of infrared radiation by gases such as water vapor and carbon dioxide in the Earth's atmosphere. Fortunately, in the near-infrared and mid-infrared regions, from 1 to 10 μm, there are some clear atmospheric 'windows'. From observatories on high mountain peaks, astronomers are able to use these 'windows' to investigate the infrared sky at certain wavelengths.

Most observations are obtained by means of near-infrared detectors attached to optical telescopes. For example, the 5 m (200 in) Hale telescope on Mount PALOMAR has recently been fitted with a near-infrared camera which operates at 1–2.5 μm, while the 8 m (300 in) units of the Very Large Telescope in Chile utilize a cryogenic infrared spectrometer and array camera (ISAAC) which observes in the spectral region 0.9–5 μm.

Modern instruments attached to smaller telescopes can also return data of major importance. The latest census of the near-infrared sky, known as the Two Micron All Sky Survey (2MASS) project, began in June 1997. 2MASS uses two new, highly automated 1.3 m (51 in) telescopes, one at Mount Hopkins in Arizona, and one at CERRO TOLOLO in Chile. Both telescopes are equipped with a three-color infrared camera, each channel consisting of a 256 $\times$ 256 array of detectors which is capable of observing the sky simultaneously at 1.25, 1.65 and 2.17 μm.

The development of large-format infrared detector arrays, funded by the US Department of Defense and NASA, has made the new survey possible. These mercury–cadmium–tellurium detectors, similar to those developed for the Near Infrared Camera and Multi-Object Spectrometer on the HUBBLE SPACE TELESCOPE, can now detect astronomical objects over 100 million times fainter than those seen by the original Two Micron Survey.

When completed, the 2MASS survey will contain accurate positions and fluxes for some 300 million stars and other unresolved objects, together with positions and total magnitudes for more than a million galaxies and nebulae, including galaxies in the 60° wide 'zone of avoidance', where dust within the Milky Way renders optical galaxy surveys incomplete. Rare objects such as extremely low-luminosity stars and brown dwarfs, or heavily dust-obscured AGNs and globular clusters located in the galactic plane, should also be discovered. (For further information see http://www.ipac.caltech.edu/2mass/)

Despite such advances, the best site for ground-based infrared observations remains the 4200 m (13 800 ft) high Mauna Kea in Hawaii. Located on its summit are the two largest purpose-built infrared telescopes in the world, the 3 m (120 in) NASA INFRARED TELESCOPE FACILITY and the 3.8 m (150 in) UNITED KINGDOM INFRARED TELESCOPE. Nearby are the twin 10 m (400 in) Keck telescopes and the 3.6 m (140 in) CANADA–FRANCE–HAWAII TELESCOPE, which are used for both optical and infrared observations.

However, even the Mauna Kea site is not high enough to allow far-infrared observations. In order to rise above the bulk of the water vapor and the atmosphere, astronomers have turned to placing telescopes on balloons, sounding rockets or high-flying aircraft.

In 1972 and 1974, the US Air Force Cambridge Research Laboratories (later known as the Air Force Geophysics Laboratory) carried out project HISTAR to survey the sky at various wavelengths between 4 and 27 μm. Using a liquid helium-cooled 16.5 cm (6.5 in) telescope on nine suborbital rocket flights, the project cataloged over 2000 infrared sources.

This was soon followed by the introduction of NASA's Kuiper Airborne Observatory (KAO), a modified Lockheed C-141 military aircraft equipped with a 90 cm (36 in) infrared telescope. After 20 years and more than

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1500 flights, budget cuts brought the KAO to an end on 13 October 1995.

However, a much larger, more powerful successor is scheduled to become operational in 2002. NASA's Stratospheric Observatory for Infrared Astronomy (SOFIA) will fly on board a modified Boeing 747-SP aircraft at an operational height of 12.5 km (nearly 41 000 ft). From this altitude, SOFIA will be above 99.9% of the infrared-absorbing atmospheric water vapor that limits ground-based infrared observations.

With its European-made 2.7 m (106 in) primary mirror, SOFIA will be one of the most capable infrared observatories in the world. It will operate primarily at infrared and submillimeter wavelengths, from 0.3 to 1600 μm, though it can also be used for visible observations. Despite its size, SOFIA has been designed to meet the stringent weight requirements of an airborne telescope, and the complete system is expected to weigh only about 20 tonnes (44 000 lb).

The observatory is expected to provide unprecedented views of regions of star formation in interstellar dust clouds, circumstellar dust disks, distant comets and moons, as well as spectral information on complex molecules in space, planetary atmospheres and nearby galaxies.

SOFIA is being developed by the American Universities Space Research Association (USRA), which will also operate the telescope for NASA once it becomes operational. Under an international agreement between the United States and the German government, the German Aerospace Research Centre (DLR) is responsible for the design and construction of the SOFIA telescope. In return for supplying the telescope and part of the funding for its operation, German astronomers will receive 20% of SOFIA's observing time.

see also: INFRARED SPACE MISSIONS

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ingalls, Albert (1888–1958)

American amateur astronomer and amateur telescope maker.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Inner Planets

A collective term for the major planets Mercury, Venus, Earth and Mars: those whose orbits lie closest to the Sun, inside the main asteroid belt, as opposed to the outer planets, orbiting beyond the main asteroid belt.

Compare: inferior planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Institut de Radio Astronomie Millimétrique

The Institut de Radio Astronomie Millimétrique (IRAM) is an international scientific institute covering all aspects of radio astronomy at mm wavelengths. In this spectral range thousands of molecular emission lines allow the study of physics and chemistry in the interstellar medium, and the coldest parts of the universe can be observed through the continuum emission of dust.

The IRAM headquarters are located in Grenoble, France. The institute operates two unique observatories which are available for the astronomical community. One is a radio interferometer, consisting of currently five (soon six) 15 m diameter antennas on the Plateau de Bure in the French Alps, with imaging capabilities at an angular resolution below one arcsec. The other is a 30 m diameter telescope, located at an altitude of 2900 m on the Pico Veleta in southern Spain, with a support base in Granada.

IRAM was founded in 1979 by the German Max-Planck-Gesellschaft and the French Centre National de la Recherche Scientifique. The Spanish Instituto Geográfico Nacional, initially an associate member, became a full member in 1990. About 130 people are currently working at IRAM, most of them engineers and technicians.

The objectives of IRAM include the development and production of instrumentation for mm wave astronomy, such as extremely low-noise receivers, superconducting (SIS) detectors, and fast digital spectrometers. They also include the development of data reduction software.

In addition to supporting the user community, the IRAM astronomers perform forefront research in several fields, from the solar system to cosmological distances. Among the scientific highlights from the IRAM observatories are the detection of molecular gas in high-redshift objects, the mapping of CO and dust emission in nearby galaxies, the investigation of circumbinary disks and proto-planetary disks around pre-main sequence stars, and also observations of molecular line and continuum emission of comets and other solar-system objects.

For further information see
http://iram.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Institut National des Sciences de l'Univers

The Institut National des Sciences de l'Univers (INSU) was created in 1985, under the joint authority of the French Ministry of Higher Education and the CNRS. INSU is responsible for the major facilities and programs in the fields of astronomy and Earth sciences, including the Universe Sciences Observatories.

INSU operates major research facilities, both for collecting data (by satellite, oceanographic ship, atmospheric research aircraft or multisite field campaigns), and for the computer processing and interpretation of the collected data with data bases, models and numerical simulations.

It also has administrative and technical divisions designed to help implement policies. The mission of the technical division is to design projects for the necessary equipment and instruments and to operate the French facilities.

Sciences of the Universe are also represented as a department within the CNRS, and both the CNRS department and INSU are headed by the same director. The department has responsibility for CNRS-associated laboratories and personnel and, with the help of the National Committee, for the analysis of overall scientific trends.

For further information see
http://www.insu.cnrs-dir.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Institute of Space and Astronautical Science

The Institute of Space and Astronautical Science (ISAS) is situated in Sagamihara, Japan. ISAS is a central institute for space science research belonging to the Ministry of Education, and functions as an inter-university research organization. It was established in April 1981 by reorganizing the Institute of Space and Aeronautical Science, University of Tokyo. ISAS consists of around 300 staff, which include professors, associate professors, research associates, technicians and administrative staff. The major facilities are: Kagoshima Space Center (rocket launch), Noshiro Testing Center (static firing test of rocket motors), Sanriku Balloon Center (scientific balloon launch) and Usuda Deep Space Center (deep space communication). ISAS is one of the leading agencies in the world in some fields of space science, e.g. x-ray astronomy, space plasma physics, solar physics and radio astronomy. The latest version of its satellite launch vehicle is the M-V rocket, which launches almost one satellite per year into space.

For further information see http://www.isas.ac.jp/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Instituto Argentino de Radioastronomía

The Instituto Argentino de Radioastronomía (IAR) is an astrophysics research center and radio astronomical observatory located 40 km from Buenos Aires, Argentina. It was built during 1962–65 and officially inaugurated on 26 March 1966.

The primary instruments of the institute are two 30 m single dish radio telescopes which provide a coverage of the southern sky from declinations of -90° to -10°. The maximum tracking time on a given source is about four hours. The telescopes are used with receivers that operate at 1420 MHz, 1670 MHz and 3.3 GHz. The neutral hydrogen line (21 cm) can be observed with a helium-cooled HEMT receiver and a 1008-channel autocorrelator at the backend. The HPBW of the telescopes is $\sim$15 and $\sim$30 arcmins at 3.3 GHz and 1.4 GHz, respectively.

IAR's observing facilities have been used to survey the extended H I line emission in the southern sky and, in recent years, to study the variability of extragalactic radio sources on short time-scales. Many new supernova remnants and radio counterparts of gamma-ray sources have been also discovered with these telescopes.

In addition to the astronomical instruments, the IAR has a laboratory of electronics, a department of mechanics, and a well-stocked library. There is an open policy for granting telescope time on the basis of refereed proposals.

For further information see
http://www.iar.unlp.edu.ar/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Instituto de Astrofisica de Canarias

The Instituto de Astrofisica de Canarias (IAC) and its Observatories—the OBSERVATORIO DEL TEIDE on Tenerife and the OBSERVATORIO DEL ROQUE DE LOS MUCHACHOS on La Palma—are a Spanish research center open, from 1979, to the international scientific community through its Agreements on Cooperation in Astrophysics. The IAC Observatories in fact form the European Northern Observatory (ENO), in which instruments have been installed from over thirty scientific institutions belonging to Belgium, Denmark, Finland, France, Germany, Holland, Ireland, Italy, Norway, Spain, Sweden and the United Kingdom. The Observatories were inaugurated in 1985 by the Spanish Royal Family in the presence of six other European heads of state.

Although it was created in 1982 as a Public Consortium in partnership with the Spanish State, the Autonomous Canarian Government, the University of La Laguna and the Higher Council of Scientific Research (CSIC), the IAC inherited the Observatorio del Teide, which was created in 1959.

The IAC's activities include astrophysical research, the development of ground-based and space-borne instrumentation, the training of researchers, university education and popularization. Its central headquarters are a scientific-technical institute, whose resources include a library, a computer center, workshops and laboratories. Over 300 people work at the IAC, including scientific, technical and administrative personnel.

The IAC's research program consists of projects covering the universe and cosmology, stars, the Sun, interstellar matter, planetary systems, atmospheric optics, high spatial resolution, the design and construction of telescopes, and space-borne optical and infrared instrumentation. Among these projects should be highlighted the 10 m large telescope (the Gran Telescopio de Canarias, GTC), which will become operative at the Observatorio del Roque de los Muchachos at the end of 2003.

For further information see
http://www.iac.es.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Instrumentation: Adaptive Optics

Turbulence in the Earth's atmosphere produces inhomogeneities in the refractive index of the air, which affect the image quality of ground-based telescopes (see SEEING). ADAPTIVE OPTICS (AO) is a means for the real-time compensation of image degradation. It consists of using an active optical element such as a deformable mirror to correct the instantaneous wave-front distortions. These are measured by a device called a wave-front sensor which delivers the signals necessary to drive the correcting element.

First proposed in 1953 by the astronomer Horace Babcock, the technique was developed only in the 1970s, mainly for defense applications (optical surveillance). The first operational astronomical AO system was built in France and successfully tested in October 1989 in OBSERVATOIRE D'HAUTE PROVENCE. It was later installed on the 3.6 m telescope at the EUROPEAN SOUTHERN OBSERVATORY. Since 1994, a growing number of observatories have been equipped with AO systems, mostly in the United States of America.

Figure 1 shows a schematic diagram of a typical astronomical AO system. It consists of three main components, a deformable mirror, a wave-front sensor and control electronics. They operate in a closed feedback loop. The deformable mirror first compensates the distortions of the incoming wave-front. Then a beam splitter diverts part of the light toward the wave-front sensor to estimate the residual wave-front distortions. The control system uses the wave-front sensor information to update the deformable mirror figure. As the incoming wave-front evolves, these operations are repeated indefinitely at a rate of the order of a thousand times per second.

Wave-front sensors require the use of a so-called 'guide' source. They work best on a point source. Fortunately, nature provides astronomers with many point sources, in the form of stars. However, they are quite faint. With current systems, observations are limited to the vicinity of bright stars, that is a small fraction of the sky. Wave-front sensing is also possible on extended sources if they are sufficiently bright and compact. Examples are solar system objects such as small planets, satellites or asteroids, but also a few galaxy cores or even small nebulosities. Solar astronomers are now able to do wave-front sensing on the solar surface. To improve the ability to sense the wave-front, a number of researchers are also producing artificial light sources by illuminating the atmosphere with a powerful laser beacon. Unfortunately, the laser beacon itself is affected by atmospheric turbulence, and stellar sources are still needed to complement the information. Today, laser systems are costly and have not yet fully come up to expectations.

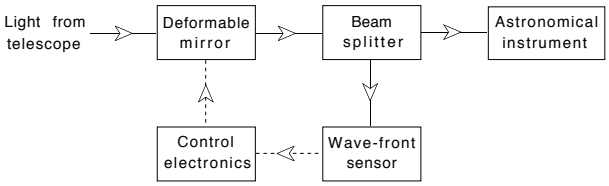


Figure 1. Schematic diagram of an AO system.

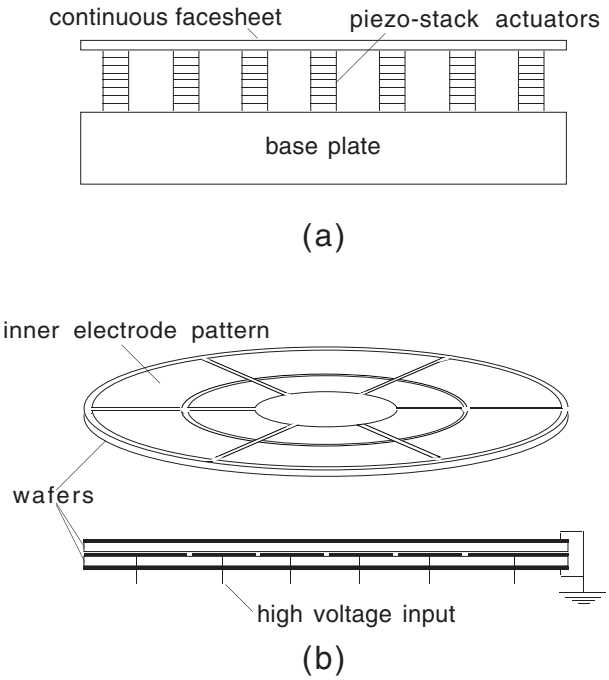


Figure 2. Deformable mirrors: (*a*) conventional, (*b*) bimorph.

Current technology

Currently, there are two competing technologies referred to as 'conventional' versus 'curvature' AO systems. Conventional systems are based on components developed in the industry: Shack–Hartmann wavefront sensors with charge coupled devices (CCD) detectors, and continuous facesheet mirrors with piezo-stack actuators. Curvature systems are based on a new approach pioneered at the Institute for Astronomy in the University of Hawaii (UH), and first used at the CANADA–FRANCE–HAWAII TELESCOPE (CFHT) on Mauna Kea. A novel wave-front 'curvature' sensor is used together with an array of discrete photon counting avalanche photodiodes (APDs) for light detection. Also a new deformable mirror technology is used based on piezo bimorph wafers.

Figure 2 shows the two main types of deformable mirrors used in astronomy. Conventional mirrors consist of a thin plate of glass or silicon supported by an array of stacked piezoelectric or electrostrictive actuators. Actuators push or pull on the plate producing local bumps or dips (figure 2(*a*)). Bimorph mirrors consist of a pair of piezoelectric wafers glued together with electrodes in between and outside. The wafers are polarized perpendicular to the surface. When voltages are applied to the electrodes, electric fields are produced inside the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

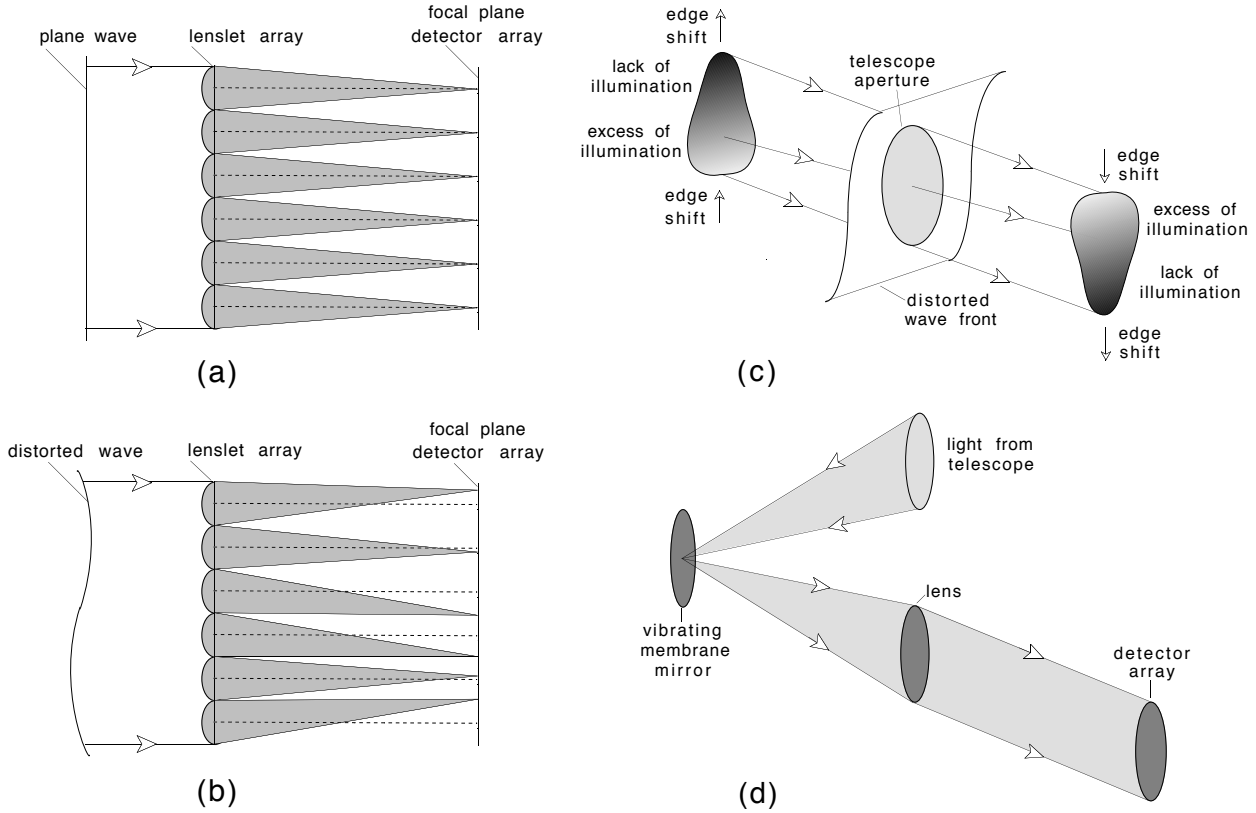


Figure 3. Wave-front sensors: (a, b) Shack–Hartmann sensor, (c, d) curvature sensor.

wafers so that one wafer contracts whereas the other one expands, locally bending the mirror (figure 2(b)). In all cases, digital to analog converters and high-voltage amplifiers are used to produce the voltages necessary to deform the mirror (typically a few hundred volts).

Figure 3 shows the two main types of wave-front sensors. A Shack–Hartmann sensor consists of an array of small lenses set in a plane conjugate to the telescope aperture. Each lenslet forms an image of the guide source on a CCD detector array (figure 3(a)). By comparison with a reference plane wave, a distorted wave-front produces shifts in the location of each of these images (figure 3(b)). A measure of the shift gives an estimate of the local wave-front slope. Mathematically, the wave-front surface can be reconstructed by integrating these wave-front slopes along two orthogonal directions. It can be shown that a least-square solution is obtained by computing the wave-front Laplacian (total curvature) from the slopes. The Laplacian is then integrated by solving a Poisson equation with radial edge slopes as boundary conditions.

A curvature sensor directly gives estimates of the wave-front Laplacian and radial edge slopes. This is done by comparing the illuminations in two symmetrically defocused images of the telescope aperture (figure 3(c)). Differences in the illumination give a measure of the local convergence or divergence of the beam, that is of the wave-front total curvature (Laplacian). Differences in the position of the aperture edges gives a measure of the local edge ray inclination, that is of the wave-front edge slope. In practice a vibrating membrane mirror is used to defocus back and forth an image of the telescope aperture (figure 3(d)). Any deviation of the wave-front surface from a plane wave produces a modulation of the light intensity at the frequency of the membrane drum oscillations (kHz rate). A phase-locked demodulation produces the sensor signals.

Real-time computers are used to control the feedback loop. Conventional systems use a wave-front computer to calculate the centroids of the Shack–Hartmann subaperture images. The coordinates of the centroids are compared with reference coordinates, and the differences form the sensor signals. Curvature systems use a two-channel counter to count photons during each half cycle of the membrane vibration. The differences in photon counts form the sensor signals. No reference is needed. In all cases, a control computer is used to convert the sensor signals into the voltages one must apply to the deformable mirror. This is done by multiplying the sensor signals with a control matrix. The control matrix is often determined experimentally with a calibration light source. The deformable mirror actuators are excited sequentially and the corresponding sensor responses are recorded. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

resulting interaction matrix is then inverted through a singular value decomposition to produce a control matrix. The eigenmodes of the control matrix are called the system modes. The eigenvalues are the loop gains of each mode. These gains can be individually modified to minimize the effect of sensor noise. The procedure is called modal control.

Performance and limitations

Figure 4 shows an example of three different profiles of the intensity distribution in a stellar image, also called the point spread function (PSF). The left profile is for an uncompensated image, and is often referred to as the seeing disk profile. The center profile is for a partially compensated image obtained with an AO system turned on. The right profile, often referred to as the AIRY DISK profile, is the theoretically ideal profile of the diffraction-limited image one would obtain with perfect optics and no atmosphere. Astronomers are used to measuring image quality in terms of the angular full width at half maximum (FWHM) of a stellar image. Expressed in radians, the FWHM of an Airy disk profile is very close to λ/D, where λ is the wavelength of light and D is the telescope diameter. That of a seeing disk profile is very close to λ/r_0, where r_0 is a seeing parameter first introduced by David Fried (1966) and now called the Fried parameter. Figure 4 shows that the FWHM of a partially compensated PSF is much closer to that of the Airy disk than to that of the seeing disk. This is because a compensated PSF consists of a narrow (almost diffraction-limited) core, on top of broad halo, barely narrower than the seeing disk. The halo is due to the light diffracted by the small-scale wave-front distortions which have not been compensated by the AO system.

A consequence is that the FWHM is a poor quality criterion for an AO-compensated image. A more widely used criterion is the Strehl ratio, which is the ratio I_c/I_d of the maximum intensity in the compensated PSF to that in the diffraction-limited PSF obtained with the same telescope (see figure 4). Visually, an image appears fairly sharp if its Strehl ratio is greater than 0.3. It appears blurry if it is less than 0.3. Following Fried, AO specialists also use another criterion called the normalized Strehl ratio or normalized resolution, which is the Strehl ratio of the compensated PSF divided by that of the pure seeing disk. If one neglects diffraction in the uncompensated PSF, then the normalized resolution is simply I_c/I_u (see figure 4). This criterion has the advantage of being telescope independent.

Figure 5(*a*) shows theoretical curves for the normalized resolution as a function of D/r_0. These are for zonal compensation over N subapertures assuming no wave-front sensing error. The curve labeled (0) is for uncompensated images. Atmospheric turbulence is supposed to follow Kolmogorov statistics. The performance of real AO systems follows very similar curves except that the actual number N_a of actuators is typically two to five times larger. The ratio N/N_a is a measure of the system compensation

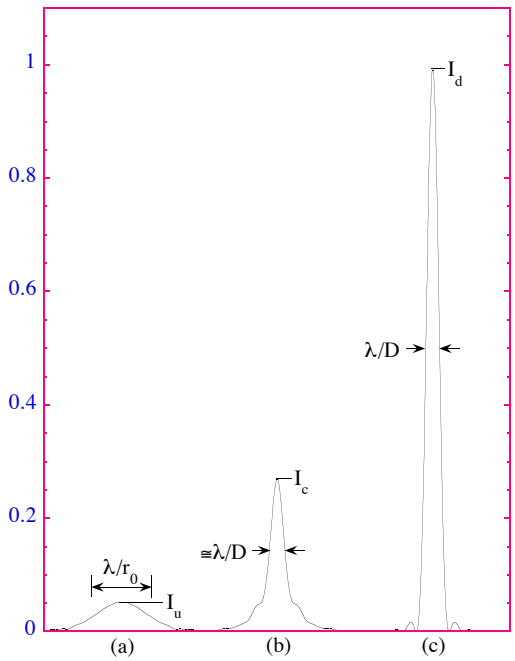


Figure 4. Intensity profiles for three different point source images: (*a*) uncompensated, (*b*) compensated, (*c*) diffraction-limited. For the clarity of the figure we have plotted only results for image motion compensation under very good seeing conditions ($D/r_0 = 4$).

efficiency. It is limited by errors due to finite wave-front sampling both in space and in time. It is interesting to note that the Strehl ratio improvement brought by adaptive optics goes to a maximum for $D/r_0 = 2.3\sqrt{N}$. At this point the gain is by a factor of $1.6N$, and the Strehl ratio of the compensated image is about 0.3.

When a faint source is used to sense the wave-front, there is an additional loss of Strehl ratio (or compensation efficiency) due to measurement noise, including photon shot noise, read-out noise and sky background noise. Curvature AO systems with photon counting APDs have constantly outperformed other systems on faint guide sources. Figure 5(*b*) shows the actual performance of three curvature systems which have been used on the CFHT: the 13-actuator UH prototype, the 19-actuator CFHT user instrument (Pueo), and the 36-actuator UH system called Hokupa'a. The broken curve indicates the projected performance for an 85-actuator curvature AO system. It is interesting to note that the limiting magnitude is independent of the number of actuators as one expects from pure photon shot noise.

Another loss of Strehl ratio occurs when the guide source is away from the observed object. This is because the two wave-front distortions differ, a property called anisoplanatism. On a 3.6 m telescope such as the CFHT, the observed loss of Strehl ratio is typically 10% over a distance of 10 arcseconds. This number depends heavily on seeing conditions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

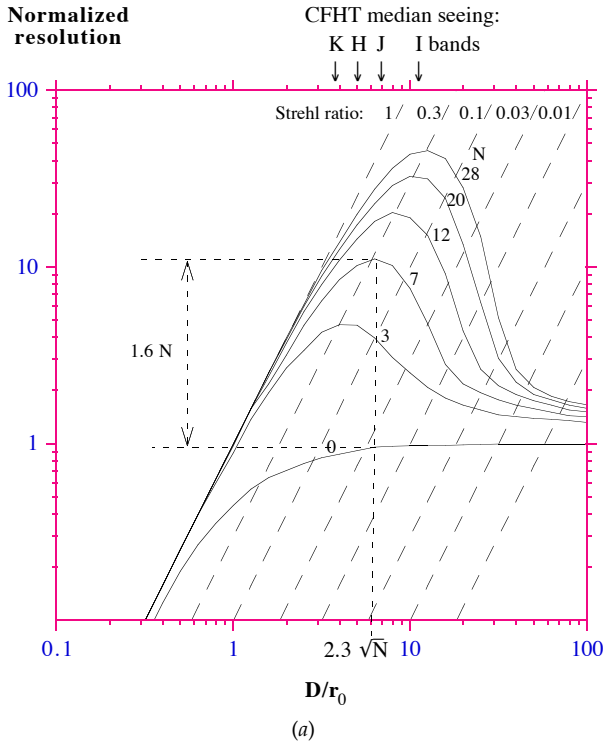


Figure 5. (*a*) The theoretical compensation performance of an ideal AO system. (*b*) The actual performance of curvature AO systems as a function of the guide source magnitude.

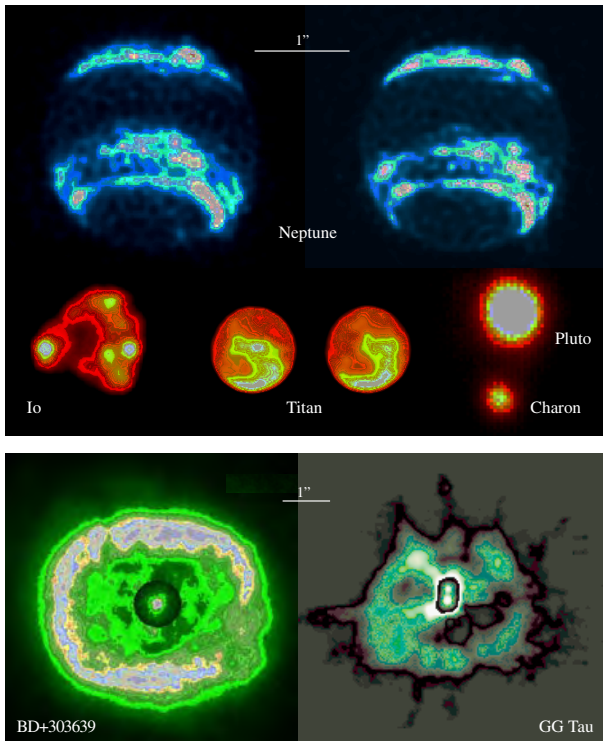


Figure 6. Examples of astronomical images recorded with adaptive optics. **This figure is reproduced as Color Plate 32.**

Examples of astronomical images

Figure 6 shows examples of infrared astronomical images obtained with adaptive optics. The upper part shows solar system objects all at the same scale. From top to bottom and left to right are two images of NEPTUNE showing rotating cloud bands, an image of IO showing glowing lava flows during an eclipse in the shadow of JUPITER, two images of TITAN showing a rotating surface feature, and an image of PLUTO AND CHARON. The bottom part shows circumstellar structures at a scale twice smaller: a PLANETARY NEBULA around an evolved star BD+303639 (bottom left) and a DUSTY CIRCUMSTELLAR DISK around a young star GG Tau (bottom right). In both cases the central source is displayed with an attenuated intensity scale. The images of Io, Pluto–Charon and GG Tau were obtained at the CFHT with the UH 13-actuator prototype AO system. The other images were obtained with the UH 36-actuator Hokupa'a system either at the CFHT (Neptune and Titan) or on the 8-m GEMINI telescope (BD+303639). In all cases the object itself was used as a guide source except for Io where nearby EUROPA was used.

Bibliography

Hardy J W 1998 *Adaptive Optics for Astronomical Telescopes* (Oxford: Oxford University Press)

Roddier F (ed) 1999 *Adaptive Optics in Astronomy* (Cambridge: Cambridge University Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tyson R K 1997 *Principles of Adaptive Optics* 2nd edn (New York: Academic)

François Roddier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Integral (International Gamma Ray Astrophysics Laboratory)

European Space Agency gamma ray astronomy mission, due for launch in 2001. Instruments include a gamma ray imager, gamma ray spectrometer, x-ray monitor and optical camera. These will pinpoint the locations of gamma ray sources to within a few arcminutes and measure their radiation energy with unprecedented accuracy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interacting Galaxies

Galaxies that interact with each other during close encounters or collisions. In a close encounter between two galaxies, both are distorted by gravitational tidal forces. Typically, matter will be pulled into bulges on both sides of each galaxy and, depending on the angular momentum (a quantity that depends on the distribution of mass and rotational velocity within each galaxy) of each galaxy and on the minimum separation and relative velocity of the two galaxies, stars and gas clouds may be flung out into long, looping 'tails' and/or material may flow into bridges that link the galaxies together. The pair of galaxies, NGC 4038 and 4039, which is known as the 'Antennae' because of the huge pair of antenna-like arms that curve away from the main bodies of the two colliding galaxies, appears to be a classic example of this process in action. Close encounters may also be responsible for initiating spiral structure in some galaxies.

When two galaxies collide, their individual member stars do not themselves collide (the average separation between stars is so great that the chance of any direct stellar collisions taking place is microscopic), but their constituent gas clouds do. Through mutual interactions, some stars will gain energy and be ejected from the member galaxies; others will lose energy and fall towards the center of mass of the two galaxies. A possible outcome of this process is that the two galaxies will merge into one. In some situations, collisions between their constituent gas clouds may strip one or both galaxies of their gas. This gas, heated by the violence of the collision, is then added to the tenuous distribution of high-temperature intergalactic gas that is known to be present in many clusters of galaxies. In other cases, gas clouds that have been compressed through collisions or tidal interactions collapse on themselves and trigger dramatic bursts of star formation. This process may account for the existence of starburst galaxies, highly luminous galaxies that shine with the light of large numbers of newborn stars. Where a compact galaxy passes through a larger spiral galaxy, it can initiate star formation in a ring within the larger galaxy's disk. The 'Cartwheel' Galaxy, for example, appears to be the result of an event of this kind.

Because most galaxies are members of groups and clusters within which the average separation of galaxies is ten to a hundred times the average diameter of a galaxy, close encounters, collisions and mergers are common events that play a major role in the evolution of galaxies and clusters. If small galaxies make close encounters with, or collide with, larger, more massive, ones, the smaller galaxies may be completely disrupted and absorbed by the larger ones. This process of 'galactic cannibalism' causes large galaxies to grow at the expense of smaller ones and is believed to be responsible for the existence of giant elliptical galaxies in the centers of richly populated clusters of galaxies.

See also: Cartwheel galaxy, cluster of galaxies, elliptical galaxies, galaxies: interactions and mergers, galaxy, spiral galaxy, starburst galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interball

Solar–terrestrial study involving two Russian PROGNOZ spacecraft and two small subsatellites made in Czechoslovakia. Designed to study plasma processes in the Earth's magnetosphere by using two pairs of spacecraft, one above the polar aurora and one in the magnetospheric tail. The Tail Probe and its subsatellite S2-X were launched in August 1995. The Auroral Probe and S2-A were launched in August 1996.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interference

The pattern of light and dark obtained when two beams of light combine or interfere with each other. If two waves of the same wavelength and amplitude have the same phase (i.e. their crests and troughs are in step), then the crests will enhance each other to produce a bright fringe light. If they are exactly out of phase, then the crests of one wave and the troughs of the other will cancel each other out, so producing a dark fringe. These patterns are also referred to as interference fringes. Such patterns may be formed with any type of electromagnetic radiation (or with other types of waves, water waves, for example), and the interpretation of these interference patterns forms the basis of the technique of interferometry.

An interferometer makes use of the phenomenon of the interference of electromagnetic wavefronts arriving at two separate detectors to attain more precise position measurements or better resolving power than could be achieved with a single telescope or detector.

For example, a typical radio interferometer consists of two telescopes, separated by a known distance. Both telescopes are used to study the same source. Because of the separation of the two receivers, there is a small delay between the arrival of a given radio wave at the first and second radio telescope (i.e. the wavefronts arriving at any instant at the two receivers will be out of phase, in general, and will interfere with each other if they are combined). The combined signal is analyzed and, for example, as the source moves across the sky, the analysis of the changing interference pattern will yield a precise position of the source.

For a given wavelength of radiation, the greater the separation of the telescopes the higher the resolving power achieved. The most precise position measurements in radioastronomy have been made by long-baseline interferometry, combining signals from two radio telescopes separated by thousands of kilometers. The longest baseline so far achieved is between California and Australia, some 10 600 km, by means of which a resolving power of 0.001 seconds of arc has been achieved at a wavelength of 13 cm.

See also: diffraction, electromagnetic radiation, polarization, aperture synthesis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interferometer

A device that utilizes interference effects between two or more light or other types of waves in order to attain higher angular resolutions than can be attained by a single telescope or other device.

The underlying principle of interference is as follows. If two light waves of the same wavelength are combined in phase (i.e. the wavecrests of one wave are in step with the wavecrests of the other) they add together to produce a wave of greater amplitude (this process is called constructive interference). Conversely, if two waves are 180° out of phase (the crests of one coinciding with the 'troughs' of the other), they cancel out, so producing a wave of zero amplitude (destructive interference). If two beams of light, originating from a single source, follow paths of different lengths and are then recombined, they will interfere with each other, producing bright fringes where they arrive in phase and dark fringes where they arrive out of phase. Likewise, beams of light originating from different parts of the same source, or from separated sources, will interfere when combined. The pattern of fringes, or interference pattern, that is produced in this way can be analyzed to yield information about the different path lengths that the beams have followed and hence to provide information about the angular size of a source, the angular separation between different parts of the same source, the angular separation between two separate point sources or the location of the source on the sky.

The interferometer was invented by Albert A Michelson (1852–1931) in 1881 in order to measure differences in path length for light beams, and has since had a wide variety of applications. A subsequent version, the Michelson stellar interferometer, consisted of two separate apertures, or slits, placed in front of a large telescope. In principle, when light from a single star, or a double star, passed through the two slits and recombined at the focus of the telescope, analysis of the resulting interference pattern allowed the angular diameter of a large star or the angular separation between the components of the binary to be measured. The resolving power of an interferometer of this kind depends on the separation between the apertures or slits; the larger the separation the better the resolution ($R = \lambda/2D$, where R = angular resolution in radians, λ = wavelength and D = the separation of the apertures). Later versions extended the separation between the apertures by means of mirrors placed on a beam located across, and extending beyond, the aperture of the telescope itself.

The radio interferometer combines the signals received by two (or more) separate radio dishes. As the source (or sources) is carried across the sky by the Earth's rotation, the advancing wavecrests have different distances to travel before reaching each of the two dishes and move in and out of phase with each other, so producing an interference pattern, which can then be analyzed. The greater the separation of the dishes (the baseline), the better the resolution. High resolution is obtained only along a direction parallel to the baseline. However, if use is made of the Earth's rotation (to change the orientation of the baseline relative to the sky), and the separation between dishes can be varied, then a series of observations can yield an image equivalent to that which would be produced by a single dish equal in aperture to the maximum separation between dishes. This technique is called aperture synthesis. Because light waves are many orders of magnitude (typically a million times) shorter than radio waves, the tolerances necessary to use separated optical telescopes as interferometers or aperture synthesis systems are orders of magnitude finer than those that radioastronomers have to achieve. Consequently, although optical fringes produced by two separate telescopes were first observed by French astronomer Antoine Labeyrie in 1975, it is only in very recent years that optical astronomers have begun to be able to use separated telescopes to good effect as interferometers and aperture synthesis systems.

The COAST (Cambridge Optical Aperture Synthesis Telescope) system achieved optical resolutions of around 0.01 arcsec in 1995. The twin 10 m Keck telescopes (on Mauna Kea, Hawaii) and the four 8 m mirrors of the Very Large Telescope (on Mount Paranal, Chile) are each expected to be able to achieve milliarc second resolutions when they become operational in interferometer and aperture synthesis modes.

See also: aperture synthesis, radio interferometer, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interferometry: Ground

Astrometry is the measurement of angular positions of stars. Historically, this was done by measuring the position of a stellar image in the focal plane of a telescope using for example a CCD, a photographic plate or the observer's eye. When changes in position over a small field of view are required—for PARALLAXES or astrometric orbits of BINARY STARS—the star's position is measured relative to other stars within the telescope's field of view. This is called narrow-angle astrometry. Global positions needed for reference frame problems require wide-angle astrometry. Here the telescope's orientation is known relative to the Earth and is called a transit circle.

The astrometric precision is limited by the size of the star's image in the telescope's focal plane, which in turn is limited by diffraction of light at the edges of the telescope's aperture. The larger the aperture, the smaller the image and the more precise the astrometric measurement. The angular resolution can be increased by combining light from multiple telescopes. Done properly, this is called optical interferometry and the resolution now depends on the baseline, the distance between the two telescopes instead of on the diameter of the telescope apertures. With this technique, the resolution, hence the precision of the measurement, can be improved simply by increasing the baseline. Baselines much longer than the diameter of the largest telescope are possible. (See also ASTROMETRY: TELESCOPES AND TECHNIQUES.)

The astrometric accuracy, however, is limited by systematic effects and not by the precision of the measurements. The atmosphere produces the largest systematics: anomalous refraction, image blurring and motion. These effects favor the use of an INTERFEROMETER.

Anomalous refraction

The Earth's atmosphere bends light, making stars appear higher in the sky than they actually are. This is because air has a larger index of refraction than empty space. As light enters a region of higher index of refraction it is bent toward the direction of the gradient. Since the air's index of refraction depends on both its temperature and pressure and therefore, on the local meteorological conditions, the stellar refraction will, in general, depart from that caused by the mean atmosphere. This departure is anomalous refraction and limits the accuracy of astrometric measurements. Although monitoring the atmosphere with the hope of correcting the data for anomalous refraction could be done, the required accuracy of the measurements is greater than what can be obtained. For example, at a zenith angle of 45°, an error in the mean temperature along the line-of-sight of only 0.5° will result in an astrometric error of 0.1 arcseconds.

For wide-angle astrometry, an interferometric measurement should be of higher accuracy because it is inherently differential; it compares the difference between the optical paths from the star to where the light is combined from the two telescopes. To the extent that the atmosphere is the same over both telescopes, its effect cancels. This cancellation is a strong argument for using an interferometer for ground-based astrometric measurements.

Image blur and motion

The weather patterns that give rise to anomalous refraction can be relatively constant over periods of hours. There are, however, random fluctuations in temperature and pressure which occur on much faster time scales. These are the same fluctuations responsible for shimmer seen when looking toward the horizon on a hot day; and for the scintillation (twinkling) of stars at night. Although scintillation does not directly affect the astrometry, it is accompanied by a distortion of the wavefront that results in image blur in a large telescope and image motion in a small telescope. These effects directly impact both the accuracy and precision of the astrometric measurement.

As with most optical media, the index of refraction of air increases with decreasing wavelength. As a result, atmospheric refraction is larger at shorter wavelengths. In principal, this shift of the stellar image with wavelength can be measured and used to estimate the total atmospheric refraction. In practice, the wavelength dependence is too small to be practical for telescopic observations. This situation is different for the interferometer which maintains its high precision looking through the atmosphere; for an interferometer, this technique is feasible and has been demonstrated.

Background

The workings of an OPTICAL INTERFEROMETER can be described in two ways. In the previous section, we described an optical interferometer as giving a higher resolution image than that obtained with a single telescope. Although this is true, an astrometric interferometer usually does not form an image; it is better to think of its operation in terms of interference fringes.

Interference fringes

The schematic of a simple optical interferometer is shown in figure 1. Two telescopes, pointing at the same star, intercept sections of a single wavefront of light. Those sections of wavefront are combined at the beam splitter, BS. The detectors D_1 and D_2 each see half the light from each section of wavefront. The interferometer is aligned when the wavefronts entering the detectors are parallel to each other and the path length from the star to the beam splitter is the same regardless of which telescope intercepted the light. The path lengths are equalized with delay lines which consist of mirrors on a cart that can be moved to change the delay internal to the interferometer.

A change of the internal delay shifts the relative phase of the two waves reaching a detector. When the waves are exactly in phase, they superimpose, reinforcing each other. They result in a larger amplitude and more light. This is called constructive interference. When the wave are 180°, or half a wavelength, out of phase, they cancel each other with one at a maximum while the other is at a minimum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

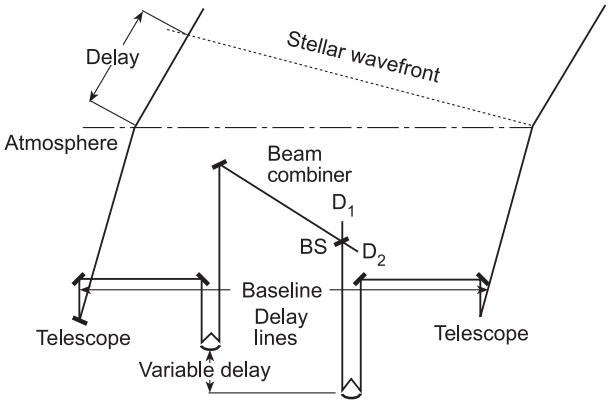


Figure 1. Layout of a simple Michelson interferometer.

This case results in no light at the detector. This is destructive interference. In an astrometric interferometer, the fringe is detected as a variation of intensity as a function of delay in the combined beams.

To look at this from another perspective, consider a perfect beam splitter for which the reflected beam is retarded by a quarter wave (90° of phase) relative to the transmitted beam. When the internal delay exactly matches the external delay, the light from one telescope is a quarter wave retarded relative to the light from the other telescope when it reaches detector D_1 but is a quarter wave advanced when it reaches detector D_2. In both cases, there is quarter wave phase shift between the two waves so the two detectors see equal light. This is exactly what is expected since the system is completely symmetric. When the internal delay changes, the phase difference increases at one detector and decreases at the other. The output of the detectors varies sinusoidally with phase but since they are 180° out of phase, the total amount of light is constant.

The fringe packet and temporal coherence

We have seen how interference fringes are formed. As the delay changes, the outputs of the beam combiner oscillate. These intensity oscillations have a period of one wavelength. Now, consider what happens when the interferometer passes light over a range of wavelengths. When the path lengths are equal, all wavelengths interfere constructively. But since the blue fringes are shorter than the red fringes, as the delay difference increases, successive intensity maxima are broadened by the blue light reaching maximum earlier than the red light. This smearing of the colors across the peak results in a lower maximum. The same process occurs at the minimum; the blue minimum occurs before the red and the minimum at one wavelength is partially filled in by the other wavelengths. The result is reduced fringe contrast or visibility[1]. The result is each successive fringe has lower visibility. The central fringe is the highest visibility fringe and is also called the white light fringe because it is the only fringe that occurs nearly simultaneously for all wavelengths. For a bandpass of $\Delta\lambda$, centered at wavelength, λ, the visibility drops off as

$$V = V_0 \frac{\sin(\pi d \Delta\lambda/\lambda^2)}{\pi d \Delta\lambda/\lambda^2} \tag{1}$$

where V_0 is the visibility of the white light fringe, and d is the delay offset from that fringe. Thus the fringes are not visible at all delays; they are confined to a region of delays of width $\lambda^2/\Delta\lambda$ and form a fringe packet.

The path length mismatch is referred to as the delay because it can be thought of as a mismatch in the time of arrival of the wavefronts between the arms of the interferometer. For the same reason, the width of the fringe packet is referred to as the temporal coherence because it is related to the maximum timing error that can be allowed for interference to occur.

Astrometry

Because the white light fringe occurs when the internal delay equals the external (or geometric) delay, the delay can be written as

$$d_{\rm int} + C = \boldsymbol{B} \cdot \boldsymbol{s} \tag{2}$$

where $\boldsymbol{B}$ is the baseline, the vector distance between the two telescopes, $\boldsymbol{s}$ is the unit vector pointing toward the star and C is the delay constant, a zero point offset in the internal delay measurement. This equation contains four parameters that describe the instrument, the three components of the baseline and the delay constant. It also contains two unknowns for each star, the right ascension and declination. Including these parameters explicitly in equation (2) gives

$$\begin{aligned} d_{\rm int} = {} & B_{\rm x}\cos(\delta)\cos(t-\alpha) \\ & - B_{\rm y}\cos(\delta)\sin(t-\alpha) \\ & + B_{\rm z}\sin(\delta) - C \end{aligned} \tag{3}$$

where t is the local sidereal time. A right-handed coordinate system has been used with z-axis parallel to the Earth's rotation axis, y-axis pointing toward the east point on the horizon, and x-axis pointing toward the point where the celestial equator crosses the meridian.

An observation with an astrometric interferometer consists of pointing the telescopes at the same star and moving the delay lines until the white light fringe is located. The time and internal delay of the white light fringe are recorded. Figure 2 shows what a typical data set looks like for a north–south baseline. It can be seen from equation (4) and figure 2, that the delay for a star is a sinusoidal variation with a 24 h period. For two stars, the difference in the phase of this oscillation is equal to the difference in the stars' right ascensions. Moving from the equator to the pole, the amplitude of the oscillation

[1] To be precise, the fringe contrast should be called the visibility amplitude, since properly defined, the visibility is a complex number. The visibility phase depends on source structure but since it is usually zero for astrometric sources, we ignore it here.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

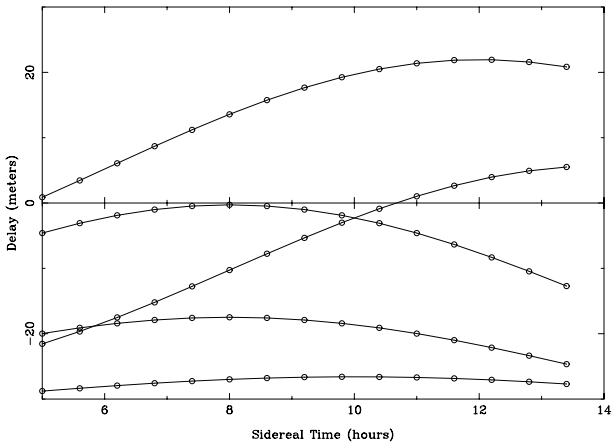


Figure 2. Delay as a function of time for a north–south baseline. Data for the same star are connected by a solid curve. Stars at higher declination show a smaller amplitude variation and the mean delay is more negative. For two stars, the difference in time between the maxima is the difference in the star's right ascensions.

decreases and its offset from zero delay increases. With enough data, the four instrumental parameters, all the declinations, and the relative right ascensions can be determined.

Although this method will work, it requires nearly horizon to horizon coverage in order to determine the amplitude, phase, and offset of the delay variations, each with errors comparable to a single measurement. Thus it is difficult to exploit partial nights of data and the observations are susceptible to slow instrumental drifts. A more robust interferometer has at least three independent baselines (four telescopes). With three baselines, three delays are measured simultaneously and the number of data increases faster than the number of unknowns as stars are added to the list of observations. In this way, the baselines and relative angles between the stars can be determined with a short set of observations. Multiple observations of the stars through the night are needed only to tie the coordinate system to the Earth's rotation axis. This type of instrument is less susceptible to subtle instrumental effects corrupting the results since the instrumental parameters can be determined from data spanning a shorter time.

Practical considerations

To achieve milliarcsecond accuracy with 100 meter class baselines, the delay must be determined with an accuracy of better than 1 μm. This implies that the internal delay, or at least changes in the internal delay needs to be measured with similar accuracy. This is while those delays are changing by several tens of meters as the observations are switched between different parts of the sky. Unmonitored changes produce noise on the astrometric data. The most important parts of the interferometer to monitor are the delay lines and the telescopes. The delay lines allow large, rapid variations of the internal delay without tilting the beam or introducing path length vibrations. A laser metrology beam, injected into the system parallel to but displaced from the stellar beam, is used for monitoring the position of the delay line cart; hence, the change in the internal delay.

At high frequencies, the signal from the laser metrology system is used for controling vibrations. At lower frequencies, it becomes the primary astrometric data.

The only moving parts in the interferometer are the delay lines and the telescopes. Since the delay lines are monitored, most of the astrometric noise is due to the telescopes; either due to bearing errors or to an offset between its mechanical and optical pivots. Bearing errors can be kept down to a few tenths of a micrometer. The pivot offsets can be controled during manufacture. Instead of zeroing the pivot-point offsets, it is more important to make the telescopes identical so that the delays introduced are the same for the telescopes at both ends of the baseline and can cancel. Alternatively, these motions can be monitored with a three-dimensional metrology system tying the position of the telescope to local bedrock.

Most of the optical paths through the system should be in vacuum. This is certainly true for the delay lines because the paths from the telescopes to the star all contain the same amount of air. Since the external delay is the difference between these path lengths, the air cancels out and the external delay can be thought of as occurring in vacuum above the Earth's atmosphere. Therefore the change in internal delay must also occur in vacuum in order to keep the air paths matched in all arms of the interferometer. If this is not done, the dispersion due to propagation through air will severely limit the observing bandpass. The rest of the path lengths do not have to be in vacuum as long as the amount of air is matched between the arms. However, the less air in the path the better since small temperature gradients will refract the light and air turbulence will degrade the stellar wavefronts, as previously described.

Finally, the mirrors in the feed optics between the telescope and the beam combination laboratory can move, introducing delay errors. Because these mirrors are exposed to the elements and can experience large temperature variations, they can be unstable and produce delay variations that may limit the performance of the system. These variations can be minimized either by monitoring the path length variations with a metrology system, or by controling and monitoring the environment so that the variations are kept to a tolerable level.

Atmospheric effects

For wide-angle astrometry, traditional measurements with a transit circle are limited in accuracy by our inability to correct the data for anomalous refraction. An interferometer measures the difference in optical path between its arms. To the extent that the telescopes sample the same atmosphere, refraction affects all the light paths

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

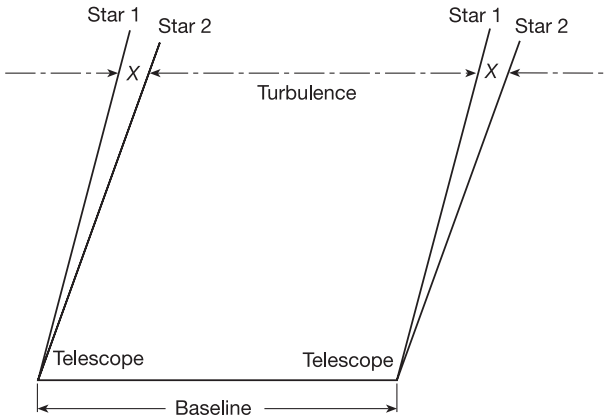


Figure 3. Diagram of the observation of two stars separated by a small angle.

the same and atmospheric refraction has no effect on the interferometer data. The major systematic error in transit circle measurements, anomalous refraction, is missing from the interferometer data.

For an interferometer, the largest systematic effect is due to the zenith angle of the star being different at the two ends of the baseline. Since the air path varies as the secant of the zenith angle, this introduces a delay error that is zero for an observation at the zenith but grows toward the horizon in the direction of the baseline. The delay error is proportional to the length of the baseline. For a 100 meter baseline, the difference between the optical and geometric paths is 2 meters at a 45° zenith angle, but the delay error is never more than 30 μm corresponding to an astrometric error of 60 milliarcseconds. By contrast, the angles measured with a transit circle are directly affected by atmospheric refraction. Thus the correction has to be calculated and applied with 1000 times the accuracy in order to achieve the same performance as the interferometer.

The atmosphere also adds random noise to the systematics caused by refraction. The optical path length through the atmosphere depends not only on the geometric path length, but also on the index of refraction of the air along that path. Fluctuations in the temperature, pressure and composition (primarily water vapor) of the air result in changes of the index of refraction and hence the delay. These variations in the index of refraction distort, as well as delay the stellar wavefront. These distortions are the cause of the SEEING which blurs the images made with a conventional telescope. The change in delay adds noise to the interferometric measurement.

This atmospheric noise is poorly behaved. It increases both with increasing baseline length and with increasing interval between observations. For extended observations of a single source, the accuracy of the measurement improves as the sixth root of the integration time, instead of the square root dependence typical of white noise.

The first consideration in dealing with this problem is the choice of baseline. For short baselines, the delay variations increase almost linearly with baseline length so the astrometric accuracy should improve very slowly with increasing baseline length. However, the size of the delay variations saturates when the baseline length reaches the outer scale of atmospheric turbulence. When the current flurry of optical interferometer construction began in the late 1980s, it was generally thought that the outer scale was very large so there was little advantage to using long baselines. However, since then, evidence has been steadily accumulating for outer scales at interferometric sites on the order of 30–100 meters. Longer baselines may indeed improve the accuracy of ground baseline astrometry.

If the air mismatch between the arms of the interferometer can be estimated, the observed delays can be corrected for the mismatch. This effectively reduces or eliminates the atmospheric noise. The correction is possible because the index of refraction of air increases toward shorter wavelengths with a known functional form. The delay variations are larger at blue wavelengths than at red wavelengths. This dispersion can be measured and is used to correct the observed delays for the atmospheric contribution. Specifically, over the visible spectrum, the index of refraction of air can be represented to high accuracy as a quadratic function of wavelength:

$$(n-1)_{\rm air} = n_0 + n_1\lambda + n_2\lambda^2. \qquad (4)$$

The delay can be represented by two terms; a vacuum delay, $d_{\rm v}$, which is the delay that would have occurred if there were no air; and the air mismatch, $L_{\rm air}$. Even though it is the index of refraction of air that is varying with time, not the actual air path, equation (4) provides a sufficiently good approximation to the problem. The optical path difference, OPD, is now a function of wavelength where $OPD = d_{\rm v} + (n-1)L_{\rm air}$. The fringe phase is given by

$$\begin{aligned}\phi &= \frac{2\pi}{\lambda} OPD \\ &= 2\pi[(d_{\rm v} + n_0 L_{\rm air})/\lambda + n_1 L_{\rm air} + n_2 L_{\rm air}\lambda]. \end{aligned} \qquad (5)$$

The first term depends on wavelength and represents the offset between the internal delay and the center of the fringe packet. The second term represents an offset between the center of the fringe packet and the white light fringe. The third term, the one containing n_2 causes a decrease in visibility for large air mismatches. As $L_{\rm air}$ increases, the central fringe is displaced from where it would have been in the absence of the air mismatch. The fringe packet is displaced by a different amount so a fringe is no longer centered in the fringe packet. By measuring both the position of the fringe packet and the offset of the fringe from the center of the packet, it is possible to solve for the air mismatch and the vacuum delay.

Since the water vapor content along the line-of-sight is not known, the coefficients used in equation (4) will not be exact. Nevertheless, they are known well enough and improvements in accuracy of a factor of 5 to 10 are routinely obtained. The wavelength dependence of the index of refraction produces a corresponding change in angle for observations with a transit circle, but this effect is too small to be exploited in a practical instrument.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Narrow angle astrometry

Restricting astrometric observations to a small portion of the sky results in higher precision astrometry since some systematic effects are common to all of the observations. These common effects shift the measured position for all the stars, but do not affect the measurement accuracy of the angles between the stars.

Delay fluctuations between two paths through the atmosphere increase as the separation of the paths increases. Figure 3 shows the observation of two stars separated by a small angle. If the separation, X, of the lines-of-sight at the distance of the turbulence is much smaller than the baseline, then the delay fluctuations between stars will be much smaller than the delay variations of either star separately. For this to work, the atmosphere must not change significantly between the observations of the two stars. For practical purposes, this means the stars must be observed simultaneously. For baselines on the order of 100 meters and stars separated by less than an arc minute, the measurements should be dramatically more accurate than wide angle measurements. Accuracies better than 100 microarcseconds, and possibly as good as 10 microarcseconds should be possible.

To achieve 10 microarcsecond accuracy on a 100 meter baseline, the differential delay between the two stars has to be measured with 5 nm accuracy—ten to one hundred times the accuracies required for wide angle astrometry. As a result, all optical paths in the system have to be measured with a metrology system.

Very narrow angle astrometry

The data in narrow angle astrometry is the same as it is in wide angle astrometry: the delay of the fringe as a function of time. When the separation of the two stars is so small that the fringe packets of the two stars observed on the same baseline overlap, a single detector can be used to observe both sets of fringes. In this case, the fringe amplitude becomes the observable. If the difference in delay between the two stars is zero, or an even number of wavelengths, the fringes from the two stars are in phase and strong fringe contrast is observed. However, if the delay difference is an odd number of half wavelengths, the fringes from the two stars are out of phase. The intensity from one star will be at a maximum when the other star will be at a minimum and the fringes will cancel, resulting in low fringe contrast. This is a simple example of measuring the spatial coherence to determine an image of the source. This type of observation is straightforward and is readily applied to measuring the orbits of binary star systems. Much has been published with the separations of the stars determined to an accuracy of 50 to 100 microarcseconds.

Bibliography

Detecting fringe phase with an optical interferometer and its usefulness for astrometry was demonstrated with a series of optical interferometers built on Mt Wilson, California. These were the Mark I, the Mark II, and the Mark III interferometers:

Shao M and Staelin D H 1980 *J. Appl. Opt.* **19** 1519
Shao M *et al* 1988 *Astron. Astrophys.* **193** 357

For the first demonstration of dispersion compensation for improving astrometric accuracy:

Colavita M M, Shao M and Staelin D H 1987 *Appl. Opt.* **26** 4113

Narrow angle astrometry has been demonstrated with the Palomar Testbed Interferometer. Interferometers are being built to combine light from the two Keck telescopes and the four large telescopes of the VLT.

Colavita M M *et al* 1999 The Palomar Testbed Interferometer *Astrophys. J.* **510** 505–21
Armstrong J T, Hutter D J, Johnston K J and Mozurkewich D 1995 *Physics Today* **42**

The highest resolution images obtained to date are with the Navy Prototype Optical Interferometer (NPOI). The NPOI was built to do wide angle astrometry, but very narrow astrometry of binary star systems is a major portion of their science program.

Benson J A *et al* 1997 Multichannel optical aperture synthesis imaging of Zeta 1 Ursae Majoris with the navy prototype optical interferometer *Astron. J.* **114** 1221

David Mozurkewich

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Intergalactic Medium

About half a million years after the big bang, the ever-fading cosmic blackbody radiation cooled below 3000 K and shifted first into the infrared and then into the radio, and the smooth baryonic plasma that filled the universe became neutral. The universe then entered a 'dark age' which persisted until the first cosmic structures collapsed into gravitationally bound systems and evolved into stars, galaxies and black holes that lit up the universe again. Some time between REDSHIFTS of 7 and 15, stars within PROTOGALAXIES created the first heavy elements; these systems, together perhaps with an early population of QUASARS, generated the ultraviolet radiation that reheated and reionized the cosmos. The history of the universe during and soon after these crucial formative stages is recorded in the all-pervading intergalactic medium (IGM), which is believed to contain most of the ordinary baryonic material left over from the big bang. Throughout the epoch of structure formation, the IGM became clumpy and acquired peculiar motions under the influence of gravity and acted as a source for the gas that becomes accreted, cools and forms stars within galaxies and as a sink for the metal-enriched material, energy and radiation which they eject. Observations of absorption lines in quasar spectra at redshifts up to 5 provide invaluable insight into the chemical composition of the IGM and the primordial density fluctuation spectrum of some of the earliest formed cosmological structures, as well as of the ultraviolet background radiation that ionizes them.

Cosmological reionization

At epochs corresponding to z~1000, the IGM is expected to recombine and remain neutral until sources of radiation develop that are capable of reionizing it. The detection of transmitted flux shortward of the Lyα wavelength in the spectra of sources at z~5 implies that the hydrogen component of this IGM was ionized at even higher redshifts. There is some evidence that the double reionization of helium may have occurred later, but this is still controversial. It appears then that substantial sources of ultraviolet photons were already present when the universe was less than 7% of its current age, perhaps quasars and/or young star-forming galaxies: an episode of pregalactic STAR FORMATION may provide a possible explanation for the widespread existence of heavy elements (such as carbon, oxygen and silicon) in the IGM, while the integrated radiation emitted from quasars is probably responsible for the reionization of intergalactic helium. Establishing the epoch of reionization and reheating is crucial for determining its impact on several key cosmological issues, from the role reionization plays in allowing protogalactic objects to cool and make stars to determining the small-scale structure in the temperature fluctuations of the cosmic background radiation. Conversely, probing the reionization epoch may provide a means for constraining competing models for the formation of cosmic structures and for detecting the onset of the first generation of stars, galaxies and black holes in the universe.

Intergalactic hydrogen density

The proper mean density of hydrogen nuclei at redshift z may be expressed in standard cosmological terms as

$$\begin{aligned}\bar{n}_{\rm H} &= (\rho_{\rm crit}/m_{\rm H})(1-Y)\Omega_{\rm b}(1+z)^3 \\ &= (1.1\times10^{-5}\ {\rm cm}^{-3})(1-Y)\Omega_{\rm b}h^2(1+z)^3\end{aligned}$$

where Y is the primordial He abundance by mass, $\rho_{\rm crit}=3H_0{}^2/8\pi G$ is the critical density, $\Omega_{\rm b}=\rho^{\rm b}/\rho^{\rm crit}$ is the current baryonic density parameter and $H_0=100h$ km s^{-1} Mpc^{-1} is the present-day Hubble constant. Standard NUCLEOSYNTHESIS models together with recent observations of deuterium yield Y=0.247±0.02 and $\Omega_{\rm b}h^2$=0.0193±0.0014. Thus,

$$\bar{n}_{\rm H} = (1.6\times10^{-7}\ {\rm cm}^{-3})\frac{\Omega_{\rm b}h^2}{0.019}(1+z)^3.$$

As some of the baryons had already collapsed into galaxies at z=2–5, the value of $\Omega_{\rm b}h^2$=0.019 should strictly be considered as an upper limit to the intergalactic density parameter.

Because of the overwhelming abundance of hydrogen, the IONIZATION of this element is of great importance for determining the physical state of the IGM. Popular cosmological models predict that most of the intergalactic hydrogen was reionized by the first generation of stars or quasars at z=7–15. The case that has received the most theoretical studies is one where hydrogen is ionized by the absorption of photons, H+$\gamma\rightarrow$p+e (as opposite to collisional ionization H+e$\rightarrow$p+e+e) shortward of 912 Å; that is, with energies exceeding 13.6 eV, the energy of the Lyman edge. The process of reionization began as individual sources started to generate expanding H II REGIONS in the surrounding IGM; throughout an H II region, H is ionized and He is either singly or doubly ionized. As more and more sources of ultraviolet radiation switched on, the ionized volume grew in size. The reionization ended when the cosmological H II regions overlapped and filled the intergalactic space.

Photoionization equilibrium

At every point in an optically thin, pure hydrogen medium of neutral density $n_{\rm HI}$, the photoionization rate per unit volume is

$$n_{\rm HI}\int_{\nu_{\rm L}}^{\infty}\frac{4\pi J_\nu\sigma_{\rm H}(\nu)}{h_{\rm P}\nu}\,{\rm d}\nu$$

where J_ν is the mean intensity of the ionizing radiation (in energy units per unit area, time, solid angle and

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

frequency interval) and h_P is the Planck constant. The photoionization cross-section for hydrogen in the ground state by photons with energy $h_{P\nu}$ (above the threshold $h_{P\nu}{}^L$=13.6 eV) can be usefully approximated by

$$\sigma_H(\nu) = \sigma_L(\nu/\nu_L)^{-3} \qquad \sigma_L = 6.3 \times 10^{-18}\ \mathrm{cm}^2.$$

At equilibrium, this is balanced by the rate of radiative recombinations p+e→H+γ per unit volume,

$$n_e n_p \alpha_A(T)$$

where n_e and n_p are the number densities of electrons and protons and $\alpha_A=\sum<\sigma^n v_e>$ is the radiative recombination coefficient, i.e. the product of the electron capture cross-section σ^n and the electron velocity ve, averaged over a thermal distribution and summed over all atomic levels n. At the commonly encountered gas temperature of 10^4 K, α_A=4.2×10–13 cm^3 s^{-1}.

Consider, as an illustrative example, a point in an intergalactic H II region at (say) z=6, with density $\bar{n}_H$=(1.6×10–7 cm^{-3})(1+z)^3=5.5×10^{-5} cm^{-3}. The H II region surrounds a putative quasar with specific luminosity $L_\nu=10^{30}(\nu_L/\nu)^2$ erg s^{-1} Hz^{-1}, and the point in question is at a distance of r=3 Mpc from the quasar. To a first approximation, the mean intensity is simply the radiation emitted by the quasar reduced by geometrical dilution,

$$4\pi J_\nu = \frac{L_\nu}{4\pi r^2}.$$

We then have for the photoionization timescale

$$t_{ion} = \left[\int_{\nu_L}^{\infty} \frac{4\pi J_\nu \sigma_H(\nu)}{h_P \nu}\, d\nu\right]^{-1} = 5 \times 10^{12}\ \mathrm{s}$$

and for the recombination timescale

$$t_{rec} = \frac{1}{n_e \alpha_A} = 5 \times 10^{16}\ \mathrm{s} \times \frac{\bar{n}_H}{n_e}.$$

As in photoionization equilibrium $n_{HI}/t^{ion}=n^p/t^{rec}$, these values imply $n_{HI}/n^p\approx10^{-4}$, that is hydrogen is very nearly completely ionized.

A source radiating ultraviolet photons at a finite rate cannot ionize an infinite region of space, and therefore there must be an outer edge to the ionized volume (this is true unless, of course, there is a population of UV emitters and all individual H II regions have already overlapped). One fundamental characteristic of the problem is the very small value of the mean free path for an ionizing photon if the hydrogen is neutral $(\sigma_L n^H)^{-1}$=0.9 kpc at threshold, much smaller than the radius of the ionized region. If the source spectrum is steep enough that little energy is carried out by more penetrating, soft x-ray photons, we have one nearly completely ionized H II region, separated from the outer neutral IGM by a thin transition layer or 'ionization front'. The inhomogeneity of the IGM is of primary importance for understanding the ionization history of the universe, as denser gas recombines faster and is therefore ionized at later times than the tenuous gas in underdense regions. An approximate way to study the effect of inhomogeneity is to write the mean rate of recombinations as

$$\langle n_e n_p \rangle \alpha_A(T) = C \langle n_e \rangle^2 \alpha_A(T)$$

(assuming T_A is constant in space), where the brackets are the space average of the product of the local proton and electron number densities and the factor C>1 takes into account the degree of clumpiness of the IGM. If ionized gas with electron density n_e filled uniformly a fraction $1/C$ of the available volume, the rest being empty space, the mean square density would be $<n_e{}^2>=n_e{}^2/C=<n_e>^2C$.

The IGM is completely reionized when one ionizing photon has been emitted for each H atom by the radiation sources and when the rate of emission of UV photons per unit (comoving) volume balances the radiative recombination rate, so that hydrogen atoms are photoionized faster than they can recombine. The complete reionization of the universe manifests itself in the absence of an absorption trough in the spectra of galaxies and quasars at high redshifts. If the IGM along the line of sight to a distant source were neutral, the resonant scattering at the wavelength of the Lyα (2p→1s; $h_P\nu_\alpha$=10.2 eV) transition of atomic hydrogen would remove all photons blueward of Lyα off the line of sight. For any reasonable density of the IGM, the scattering optical depth is so large that detectable absorption will be produced by relatively small column (or surface) densities of intergalactic neutral hydrogen.

Gunn–Peterson effect

Consider radiation emitted at some frequency ν_e that lies blueward of Lyα by a source at redshift z_e and observed at Earth at frequency $\nu_0=\nu_e(1+z^e)^{-1}$. At a redshift z such that $1+z=(1+z_e)\nu_\alpha/\nu_e$, the emitted photons pass through the local Lyα resonance as they propagate towards us through a smoothly distributed sea of neutral hydrogen atoms and are scattered off the line of sight with a cross-section (neglecting stimulated emission) of

$$\sigma[\nu_0(1+z)] = \frac{\pi e^2}{m_e c} f\ \phi[\nu_0(1+z)]$$

where f=0.4162 is the upward oscillator strength for the transition, φ is the line profile function (with normalization $\varphi\int\varphi(\nu)\, d\nu=1$), c is the speed of light and e and m_e are the electron charge and mass, respectively. The total optical depth for resonant scattering at the observed frequency is given by the line integral of this cross-section multiplied by the neutral hydrogen proper density $n_{HI}(z)$,

$$\tau_{GP} = \int_0^{z_e} \sigma[\nu_0(1+z)] n_{HI}(z) \frac{d\ell}{dz}\, dz$$

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

where $d\mathscr{l}/dz = cH_0^{-1}(1+z)^{-1}[\Omega_M(1+z)^3+\Omega_K(1+z)^2+\Omega_\Lambda]^{-1/2}$ is the proper line element in a Friedmann–Robertson–Walker metric and Ω_M, Ω_Λ and $\Omega_K=1-\Omega^M-\Omega_\Lambda$ are the matter, vacuum and curvature contributions to the present density parameter. As the scattering cross-section is sharply peaked around ν_α, we can write

$$\tau_{GP} = \int_0^{z_e} \sigma[\nu_0(1+z)] n_{HI}(z) \frac{d\ell}{dz}\, dz$$

In an Einstein–de Sitter ($\Omega_M=1$, $\Omega_\Lambda=0$) universe, this becomes

$$\tau_{GP}(z) = \frac{\pi e^2 f}{m_e H_0 \nu_\alpha} \frac{n_{HI}}{(1+z)^{3/2}} = 6.6\times 10^3 h^{-1} \frac{\Omega_b h^2}{0.019} \frac{n_{HI}}{\bar{n}_H} (1+z)^{3/2}.$$

The same expression for the opacity is also valid in the case of optically thin (to Lyα scattering) discrete clouds as long as n_{HI} is replaced with the average neutral density of individual clouds multiplied by their volume filling factor.

In an expanding universe homogeneously filled with neutral hydrogen, the above equations apply to all parts of the source spectrum to the blue of Lyα. An absorption trough should then be detected in the level of the rest-frame UV continuum of the quasar; this is the so-called 'Gunn–Peterson effect'. Between the discrete absorption lines of the Lyα forest clouds, quasar spectra do not show a pronounced Gunn–Peterson absorption trough. The current upper limit at $z_e\approx5$ is $\tau_{GP}<0.1$ in the region of minimum opacity, implying from equation (14) a neutral fraction of $n_{HI}/\bar{n}^H<10^{-6}h$. Even if 99% of all the cosmic baryons fragment at these epochs into structures that can be identified with quasar absorption systems, with only 1% remaining in a smoothly distributed component, the implication is a diffuse IGM which is ionized to better than 1 part in 10^4.

In modern interpretations of the IGM, it is difficult to use the Gunn–Peterson effect to quantify the amount of ionizing radiation that is necessary to keep the neutral hydrogen absorption below the detection limits. This is because, in hierarchical clustering scenarios for the formation of cosmic structures (the cold DARK MATTER model being the most studied example), the accumulation of matter in overdense regions under the influence of gravity reduces the optical depth for Lyα scattering considerably below the average in most of the volume of the universe, and regions of minimum opacity occur in the most underdense areas (expanding 'cosmic minivoids').

A clumpy IGM

Owing to the non-linear collapse of cosmic structures, the IGM is well known to be highly inhomogeneous. The discrete gaseous systems detected in absorption in the spectra of HIGH-REDSHIFT QUASARS blueward of the Lyα emission line are assigned different names based on the appearance of their absorption features (see figure 1).

The term 'Lyα forest' (LYMAN ALPHA FOREST) is used to denote the plethora of narrow absorption lines whose measured equivalent widths imply H I column densities ranging from 10^{16} cm^{-2} down to 10^{12} cm^{-2}. These systems, observed to evolve rapidly with redshift between $2<z<4$, have traditionally been interpreted as intergalactic gas clouds associated with the era of baryonic infall and galaxy formation, photoionized (to less than a neutral atom in 10^4) and photoheated (to temperatures close to 20 000 K) by an ultraviolet background close to the one inferred from the integrated contribution from quasars. Recent spectra at high resolution and high signal-to-noise ratio obtained with the *Keck* telescope have shown that most Lyα forest clouds at $z\sim3$ down to the detection limit of the data have undergone some chemical enrichment, as evidenced by weak, but measurable, C IV lines. The typical inferred metallicities range from 0.3% to 1% of solar values, subject to uncertainties of photoionization models. Clearly, these metals were produced in stars that formed in a denser environment; the metal-enriched gas was then expelled from the regions of star formation into the IGM.

An intervening absorber at redshift z having a neutral hydrogen column density exceeding 2×10^{17} cm^{-2} is optically thick to photons having energy greater than 13.6 eV and produces a discontinuity at the hydrogen Lyman limit, i.e. at an observed wavelength of $912(1+z)$ Å. These scarcer LLSs are associated with the extended gaseous haloes of bright galaxies near the line of sight and have metallicities which appear to be similar to that of Lyα forest clouds.

In 'damped Lyα systems' the H I column is so large ($N_{HI}\lesssim10^{20}$ cm^{-2}, comparable with the interstellar surface

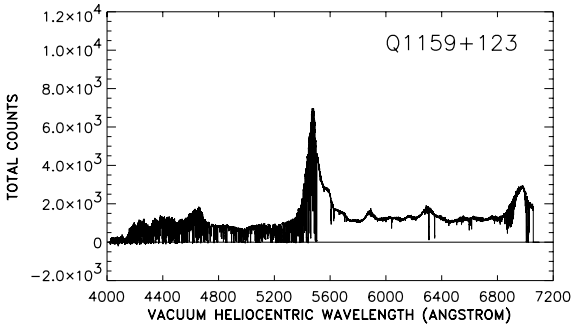


Figure 1. High resolution ($\lambda/\Delta\lambda=37\,000$) spectrum of the $z_e=3.50$ quasar Q1159+123, taken with the *Keck* High Resolution Spectrograph (exposure time 8 h). The data are taken from Songaila (1998). The Lyα forest is clearly seen in absorption blueward of the atomic hydrogen Lyα emission line from the quasar (the broad peak at 5470 Å) and is produced by resonant Lyα scattering in gas clouds along the line of sight between us and the quasar. The spectrum shows a Lyman-limit system (LLS) just shortward of 4150 Å, i.e. close to the quasar emission redshift. Most of the features between the Lyα and C IV emission (the other broad peak just below 7000 Å) are C IV intergalactic absorption lines.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

density of spiral galaxies today) that the radiation damping wings of the Lyα line profile become detectable (see LYMAN ALPHA ABSORPTION: THE DAMPED SYSTEMS). While relatively rare, damped systems account for most of the neutral hydrogen seen at high redshifts. The typical metallicities are about 10% of solar and do not evolve significantly over a redshift interval $0.5<z<4$ during which most of today's stars were actually formed.

Except at the highest column densities, discrete absorbers are inferred to be strongly photoionized. From quasar absorption studies we also know that neutral hydrogen accounts for only a small fraction, ~10%, of the nucleosynthetic baryons at early epochs.

Distribution of column densities and evolution

The bivariate distribution $f(N_{\rm HI},z)$ of H I column densities and redshifts is defined by the probability dP that a line of sight intersects a cloud with column density $N_{\rm HI}$ in the range $dN_{\rm HI}$, at redshift z in the range dz,

$$\mathrm{d}P = f(N_{\rm HI}, z)\,\mathrm{d}N_{\rm HI}\,\mathrm{d}z.$$

As a function of column density, a single power law with slope –1.5 appears to provide at high redshift a surprisingly good description over 9 decades in $N_{\rm HI}$, i.e. from 10^{12} to 10^{21} cm^{-2}. It is a reasonable approximation to use for the distribution of absorbers along the line of sight:

$$f(N_{\rm HI}, z) = AN_{\rm HI}^{-1.5}(1+z)^{\gamma}.$$

Lyα forest clouds and LLSs appear to evolve at slightly different rates, with $\gamma=1.5\pm0.4$ for the LLSs and $\gamma=2.8\pm0.7$ for the forest lines. Let us assume, for simplicity, a single redshift exponent, $\gamma=2$, for the entire range in column densities. In the power-law model (16) the number N of absorbers with column densities greater than $N_{\rm HI}$ per unit increment of redshift is

$$\frac{\mathrm{d}N}{\mathrm{d}z} = \int_{N_{\rm HI}}^{\infty} f(N'_{\rm HI}, z)\,\mathrm{d}N'_{\rm HI} = 2AN_{\rm HI}^{-0.5}(1+z)^2.$$

A normalization value of $A=4.0\times10^7$ produces then ~3 LLSs per unit redshift at $z=3$ and, at the same epoch, ~150 forest lines above $N_{\rm HI}=10^{13.8}$ cm^{-2}, in reasonable agreement with the observations.

If absorbers at a given surface density are conserved, with fixed comoving space number density $n=n_0(1+z)^3$ and geometric cross-section Σ, then the intersection probability per unit redshift interval is

$$\frac{\mathrm{d}P}{\mathrm{d}z} = \Sigma n\frac{\mathrm{d}\ell}{\mathrm{d}z} = \Sigma n_0(1+z)^3\frac{\mathrm{d}\ell}{\mathrm{d}z}.$$

If the universe is cosmologically flat, the expansion rate at early epochs is close to the Einstein–de Sitter limit, and the redshift distribution for conserved clouds is predicted to be

$$\frac{\mathrm{d}P}{\mathrm{d}z} \propto (1+z)^3\frac{\mathrm{d}\ell}{\mathrm{d}z} \propto (1+z)^{1/2}.$$

The rate of increase of $f(N_{\rm HI},z)$ with z in both the Lyα forest and LLSs is considerably faster than this, indicating rapid evolution. The mean proper distance between absorbers along the line of sight with column densities greater than $N_{\rm HI}$ is

$$L = \frac{\mathrm{d}\ell}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}N} \approx \frac{cN_{\rm HI}^{1/2}}{H_0\Omega_{\rm M}^{1/2}2A(1+z)^{4.5}}.$$

For clouds with $N_{\rm HI}>10^{14}$ cm^{-2}, this amounts to $L\sim0.7h^{-1}\Omega_{\rm M}^{-1/2}$ Mpc at $z=3$. At the same epoch, the mean proper distance between LLSs is $L\sim30h^{-1}\Omega_{\rm M}^{-1/2}$ Mpc.

Intergalactic continuum opacity

Even if the bulk of the baryons in the universe are fairly well ionized at all redshifts $z\lesssim5$, the residual neutral hydrogen still present in the Lyα forest clouds and Lyman limit systems significantly attenuates the ionizing flux from cosmological distant sources. To quantify the degree of attenuation we have to introduce the concept of an effective continuum optical depth $\tau_{\rm eff}$ along the line of sight to redshift z,

$$\langle \mathrm{e}^{-\tau}\rangle \equiv \mathrm{e}^{-\tau_{\rm eff}}$$

where the average is taken over all lines of sight. Neglecting absorption due to helium, if we characterize the Lyα forest clouds and LLSs as a random distribution of absorbers in column density and redshift space, then the effective continuum optical depth of a clumpy IGM at the observed frequency ν_0 for an observer at redshift z_0 is

$$\tau_{\rm eff}(\nu_0, z_0, z) = \int_{z_0}^{z}\mathrm{d}z'\int_0^{\infty}\mathrm{d}N_{\rm HI}\,f(N_{\rm HI}, z)(1-\mathrm{e}^{-\tau})$$

where $\tau=N_{\rm HI}\sigma^{\rm H}(\nu)$ is the hydrogen Lyman continuum optical depth through an individual cloud at frequency $\nu=\nu_0(1+z)/(1+z_0)$. This formula can be easily understood if we consider a situation in which all absorbers have the same optical depth τ_o independent of redshift and the mean number of systems along the path is $\Delta N=\int dz\, dN/dz$. In this case the Poissonian probability of encountering a total optical depth $k\tau_o$ along the line of sight (with k integer) is $p(k\tau_o)=e^{-\Delta N}\Delta N_k/\tau_o k!$, and $\langle e^{-\tau}\rangle=e^{-k\tau_o}p(k\tau_o)=e[-\Delta N(1-e^{-\tau o})]$.

If we extrapolate the $N_{\rm HI}^{-1.5}$ power law in equation (16) to very small and large column densities, the effective optical depth becomes an analytical function of redshift and wavelength,

$$\tau_{\rm eff}(\nu_0, z_0, z) = \frac{4}{3}(\pi\sigma_{\rm L})^{0.5}A\left(\frac{\nu_0}{\nu_{\rm L}}\right)^{-1.5} \times (1+z_0)^{1.5}[(1+z)^{1.5}-(1+z_0)^{1.5}].$$

Because of the rapid increase with lookback time of the number of absorbers, the mean free path of photons at 912 Å becomes so small beyond a redshift of 2 that the radiation field is largely 'local'. Expanding equation (23) around z, one obtains $\tau_{\rm eff}(\nu^{\rm L})\approx0.36(1+z)^2\,\Delta z$. This

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

means that at z=3, for example, the mean free path for a photon near threshold is only Δz=0.18, and sources of ionizing radiation at higher redshifts are severely attenuated.

Recent observational results

NASA's Far Ultraviolet Spectroscopic Explorer (FUSE) satellite has given astronomers a glimpse of the ghostly cobweb of helium gas left over from the Big Bang, which underlies the universe's structure. The helium is not found in galaxies or stars but spread thinly through space. The observations help confirm theoretical models of how matter in the expanding universe condensed into a web-like structure pervading all the space between galaxies. The helium traces the architecture of the universe back to very early times. This structure arose from small gravitational instabilities seeded in the chaos just after the Big Bang.

Bibliography

For a recent deuterium abundance measurement and its implications for the baryon density parameter see

Burles S and Tytler D 1999 *Astrophys. J.* **499** 699

A short and hardly comprehensive list of references on issues related to the reionization of the IGM includes

Arons J and Wingert D W 1972 *Astrophys. J.* **177** 1
Shapiro P R and Giroux M L 1987 *Astrophys. J.* **321** L107
Meiksin A and Madau P 1993 *Astrophys. J.* **412** 34
Tegmark M, Silk J, Rees M J, Blanchard A, Abel T and Palla F 1997 *Astrophys. J.* **474** 1
Gnedin N Y and Ostriker J P 1997 *Astrophys. J.* **486** 581
Haiman Z and Loeb A 1997 *Astrophys. J.* **483** 21
Miralda-Escudé J 1998 *Astrophys. J.* **501** 15
Madau P, Haardt F and Rees M J 1999 *Astrophys. J.* **514** 648

The use of Lyα resonant absorption as a sensitive probe of intergalactic neutral hydrogen was predicted independently by

Gunn J E and Peterson B A 1965 *Astrophys. J.* **142** 1633
Shklovsky I S 1965 *Sov. Astron.* **8** 638
Scheuer P A G 1965 *Nature* **207** 963

The quoted upper limit to the Gunn–Peterson optical depth at $z\approx5$ is from

Songaila A, Hu E M, Cowie L L and McMahon R G 1999 *Astrophys. J.* **525** L5

The high-resolution quasar spectrum shown in figure 1 was taken from

Songaila A 1998 *Astron. J.* **115** 2184

who also discusses the evolution of metal abundances in the Lyα forest. An extensive discussion of the physics of the intergalactic medium can be found in

Peebles P J E 1993 *Principles of Physical Cosmology* (Princeton, NJ: Princeton University Press) ch 23

The physics of an ionized hydrogen gas is covered in

Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Mill Valley, CA: University Science Books) ch 2 and 3

Our present understanding of the Lyα forest is summarized in

Rauch M 1998 *Annu. Rev. Astron. Astrophys.* **36** 267

The use of hydrodynamic cosmological simulations to make quantitative prediction of the physical state of the IGM was pioneered by

Cen R, Miralda-Escude' J, Ostriker J P and Rauch M 1994 *Astrophys. J.* **437** L9

Piero Madau

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

International Astronomical Union

The mission of the International Astronomical Union (IAU), founded in 1919, is to promote and safeguard the science of astronomy in all its aspects. Through its 11 scientific divisions and 40 commissions, the IAU plays a key role in promoting and coordinating worldwide cooperation in astronomy. Strong emphasis is placed on personal involvement of its over 8300 individual members from 67 member countries. The IAU is integrated into the worldwide scientific community through its membership of the International Council for Science (ICSU).

The tasks of the IAU range from the definition of fundamental astronomical and physical constants and unambiguous nomenclature, rapid dissemination of new discoveries, organization of international observing campaigns, and promotion of educational activities in astronomy, to early informal discussions of possible future international large-scale facilities.

The IAU is also the sole internationally recognized authority for giving designations and names to celestial bodies and their surface features.

The organization of scientific meetings is a key activity. The triennial General Assemblies feature a rich scientific program, recorded in the 'Highlights of Astronomy'. The IAU also sponsors about a dozen carefully selected Symposia and Colloquia each year. Proceedings of these meetings are published under the auspices of the IAU as important records of the status of their scientific fields.

The IAU promotes astronomical education and research in developing countries through the International Schools for Young Astronomers and Teaching for Astronomy Development program, and through joint initiatives with other ICSU and UN organizations.

An IAU 'Information Bulletin' is published twice per year and distributed free of charge to individual members, adhering bodies, and astronomical institutes and libraries throughout the world. The IAU homepage on the Internet provides information and news from the Union and links to Divisions, Commissions, Working Groups and related organizations. The official languages of the IAU are English and French.

For further information see
http://www.iau.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Atomic Time

The international reference scale of time based on the continuous intercomparison of atomic times registered on numerous atomic clocks in laboratories and institutions around the world. Since 1971, international atomic time, which is denoted by the abbreviation TAI, has been provided by the Bureau International de l'Heure in Paris.

An atomic clock is a device that is regulated by reference to the natural frequency of a specific transition between two levels of a particular atom, usually cesium-133. The basis of atomic time is the SI (Système International) second, which is defined as 9192 631 770 cycles (periods of vibration) of the radiation corresponding to the hyperfine transition between two levels of the ground state of cesium-133. Because of small irregularities and changes in the rotation period of the Earth, time systems such as Universal Time (UT), which are based on the rotation of the Earth, differ very slightly from the much more uniform time-scale provided by atomic time. Since 1972, the time signals disseminated for civil use have been offset by whole numbers of seconds from TAI in order to keep these signals within 1 s of UT. The offset is changed in steps of 1 s when required. The adjusted atomic time-scale given by these signals is called Coordinated Universal Time and is denoted by UTC.

See also: atom, radiation, Timescales: Terrestrial, Universal Time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Cometary Explorer (ICE)

NASA spacecraft, originally known as ISEE-3 (International Sun–Earth Explorer). First spacecraft to encounter a comet. Launched in August 1978. After completing its original mission, it was reactivated and diverted to pass through the tail of Comet Giacobini–Zinner on 11 September 1985. It also observed Halley's Comet from a distance of 28 million km in March 1986. (See also INTERNATIONAL SUN–EARTH EXPLORER and EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Dark-Sky Association

The International Dark-Sky Association is a membership-based non-profit organization whose goals are to be effective in stopping the adverse environmental impact on dark skies—by building awareness of the problem of light pollution and of the solutions—and to educate everyone about the value and effectiveness of quality night-time lighting.

For further information see http://www.darksky.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Space Science Institute

The International Space Science Institute (ISSI) was established in 1995 in Bern, Switzerland. Its main purpose is to provide a forum for the multidisciplinary analysis and interpretation of data from space research missions. The principal scientific focus is on solar system science and its implication for astrophysics and cosmology on one hand and for the Earth sciences on the other. The scientific activities are carried out through the organization of workshops on important scientific issues (twelve from its inception until 1999) and through international scientific teams of three to fifteen scientists, assembling at ISSI repeatedly for various periods of time. The scientific output is published in the ISSI Space Science Series of books (published by Kluwer Academic, Dordrecht), ISSI scientific reports (European Space Agency) and papers in scientific journals. Financial support is received in matching amounts from ESA and Switzerland.

For further information see
http://www.issi.unibe.ch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Sun–Earth Explorer (ISEE)

Series of three US satellites designed to study the solar wind and its interaction with the Earth's magnetosphere. ISEE-1 and 2 were placed into highly elliptical Earth orbits. ISEE-3 was placed in a halo orbit at the L1 Lagrangian point between the Sun and Earth. It gave advance warning of solar storms heading towards Earth. (See also INTERNATIONAL COMETARY EXPLORER and EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

International Ultraviolet Explorer (IUE)

NASA–ESA–UK ultraviolet observatory, launched in January 1978. Operated for 24 hours a day from geosynchronous orbit. Short and long wavelength spectrographic cameras covered ultraviolet wavelengths from about 1200 to 3400 Å. IUE observed planets, comets, novae, supernovae and galaxies. In July 1994 it observed the collision of Comet Shoemaker–Levy 9 with Jupiter. IUE provided confirmation of the nature of the precursor star which exploded as supernova 1987A, and revealed the chemical elements present in the star Nova Cygni 1978. Also took part in a major campaign to determine the nature of the mini-quasar at the centre of Seyfert galaxy NGC 7469. Provided a Final Data Archive containing more than 100 000 spectra. Operations ceased in September 1996. (See also EXPLORER)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interplanetary Dust

The birth of a solar system is enshrouded in gas and dust. A collapsing rotating cloud increases in density until a critical point is reached and a YOUNG STELLAR OBJECT begins its rapid accretion at the center, still surrounded by a disk of gas and dust (figure 1). As stars evolve through a T-Tauri phase, much of this primordial dust cloud is removed by strong stellar winds and radiation forces. What remains may further accrete into planets and smaller bodies. Dust not accreted into larger bodies may find itself spiralling into the star by radiation drag or colliding with other particles, breaking up, and being ejected by radiation pressure.

By the time a star is as old as our Sun, no dust remains from those early years. Yet we find ourselves in a sea of dust, which we refer to as the zodiacal cloud or interplanetary dust complex. About one-third of other nearby stars are also surrounded by dust disks, though only one—Beta Pictoris—has been successfully imaged at the close of the 20th century (figure 2).

The zodiacal cloud is not primordial dust. Rather it is dust that has been generated, in the case of our own solar system, primarily by asteroid collisions and comet mass loss. Studies of the origin and nature of the zodiacal cloud provide insights into post-primordial dust production processes around other stars. Indeed, dust around other stars like ours is evidence for the existence of asteroidal and cometary bodies that would be too small to detect directly. The number of such systems is an indication of how common planet formation may be.

What is interplanetary dust?

This is a broad category of material which may range in size from submicron particles to automobiles. Basically, anything in orbit about the Sun which is too small to be a likely candidate for eventual individual designation as an asteroid or comet could be considered 'dust'—or better, interplanetary 'debris'. The particles can be of any composition as long as they are not artificial. The focus, however, is on the smaller members of this population.

History

Interplanetary dust can be seen with the naked eye in the morning and evening sky as a faint glow, brightest towards the Sun, extending and tapering along the ecliptic (figure 3). It can be comparable in brightness to the Milky Way. Since the zodiacal constellations lie along the ecliptic, this gave rise to the term 'ZODIACAL LIGHT' to describe the phenomenon. The 11th century Persian poet, astronomer and mathemetician, Omar Khayyam alluded to this 'false dawn' in his Rubaiyat. In 1683, Italian-born astronomer Gian Domenico Cassini 'discovered' the light in Europe (where inexplicably it had previously received no obvious note) and later ascribed it to a cloud of dust flattened on the solar equator, becoming brighter as one approached the Sun or its plane of symmetry. Under very good conditions, the zodiacal light can be traced completely

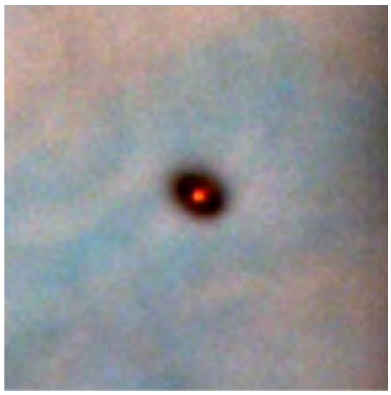


Figure 1. A protoplanetary disk and imbedded young star near the center of the Orion nebula, imaged by the Hubble Space Telescope.

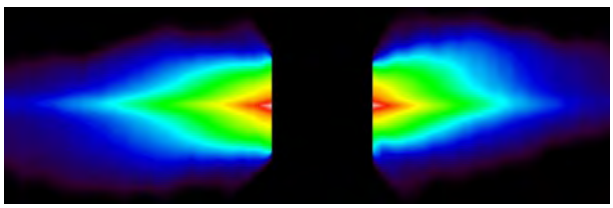


Figure 2. Beta Pictoris disk, imaged by the Hubble Space Telescope.

Figure 3. The zodiacal light, seen from the top of Mauna Kea, Hawaii. Courtesy M Ishiguro, Kobe University group.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 4. Engraving from *Bible Readings for the Home Circle*, 1889.

across the sky, with a small increase in brightness at the anti-Sun position, which was named the 'Gegenschein' by the German naturalist and explorer Alexander von Humboldt in the early 19th century.

The existence of interplanetary dust having been established in one context (zodiacal light), it was in the late 18th and early 19th centuries when it was realized that this dust might be showering the Earth. The father of acoustical science, Ernst Chladni, argued for the extraterrestrial origin of METEORS, FIREBALLS and METEORITES in 1794. The spectacular Leonid meteor shower of 1833 (figure 4), and the observation that the meteors appeared to emerge from a stationary point in the constellation Leo, led many scientists of the day to independently conclude that these meteors were of extraterrestrial origin.

Near the end of the 19th century 'cosmic spherules' were collected from deep ocean sites and found to be chemically similar to meteorites. Analysis by E J Öpik of deep ocean samples collected in the 1930s suggested that the Earth was accumulating 8×10^9 kg of meteoritic material per year (or about 1.5 grams of material per square centimeter per year). If packed into a rock-like cube with a density of 3.5 g cm^{-3}, its dimension would be 132 meters. Dust collection experiments were conducted in ernest in the first half of the 20th century—often in scientists' back yards or on their rooftops. Rain and snowmelts were studied, generally involving magnetic separation of particles, and yielding nickel–iron particles which were presumed to be meteoritic. Estimates of the amount of material accumulated by the Earth over the course of a year varied by orders of magnitude, ranging from around 10^4 to 10^8 kg. Methods depended on meteor observations and meteorite falls as well as dust collection experiments. It was recognized that these estimates were necessarily incomplete. In the early 1950s, Warren Thomsen laid out several gallon cans lined with plastic on a farm several miles outside of Iowa City, and over a six month period examined the collected contents under a microscope. He estimated that magnetic, spherical, meteoritic dust fell at a rate of 2.0×10^9 kg/year over the entire Earth. This result (though ultimately within a factor of 2–4 of the later consensus value) sparked criticism which gave focus to concerns about the high potential for contamination of dust samples collected on the Earth's surface.

An effort was made to link dust collected on the Earth to dust observed in space. In 1947, van de Hulst estimated the mean mass density of interplanetary dust particles (IDPs) giving rise to the zodiacal light, taking into account light scattering properties of particles. Assuming this dust is swept up at the Earth's orbital velocity, the annual infall of dust would be 6×10^8 kg.

As new technologies became available, they were applied to the problem of identifying infalling interplanetary dust and to determine its properties. In the 1940s, radar was used to determine meteor velocities. Converted V2 rockets were launched with acoustical detectors to detect the sounds of small particles bouncing off the surface. The problems of contamination near the Earth's surface were solved with the advent of aircraft collectors by 1960 and the use of high-altitude balloons. Zodiacal light was increasingly studied photographically and with photoelectric photometers and polarimeters, but airglow, extinction and other atmospheric problems made this difficult. By the 1960s the zodiacal light could be observed from spacecraft. With better measurements of zodiacal light and better collections of extraterrestrial particles in the upper atmosphere of the Earth, the consensus value for the amount of dust accreted by the Earth each year became around 10^9 kg.

Table 1. Spacecraft studying interplanetary dust.

	Launch date	Heliocentric distance (AU)
Pioneer 8	1967	0.97–1.09
Pioneer 9	1968	0.75–0.99
HEOS 2	1972	1.0
Pioneer 10	1972	1–18
Pioneer 11	1973	1–10
Helios 1	1974	0.3–1
Helios 2	1976	0.3–1
Galileo	1989	0.7–5.3
Hiten	1990	1.0
Ulysses	1990	1.0–5.4
Cassini	1997	1.0–10

In the last third of the 20th century, spacecraft and orbiting infrared telescopes have dramatically increased our knowledge of the complexity of the interplanetary dust cloud. Beginning in the late 1960s, these particles were sampled directly in space via impact detectors (see table 1). The Pioneer 8 and 9 spacecraft detected small particles which seemed to come straight from the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

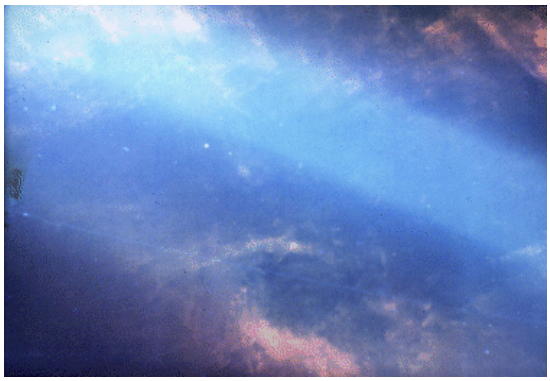


Figure 5. The sky in the infrared as seen from the Infrared Astronomical Satellite. The clouds in the upper right and lower middle are interstellar cirrus. The zodiacal cloud contains broad banded structures associated with asteroid collisions and contrail-like features associated with short-period comets. M Sykes, University of Arizona.

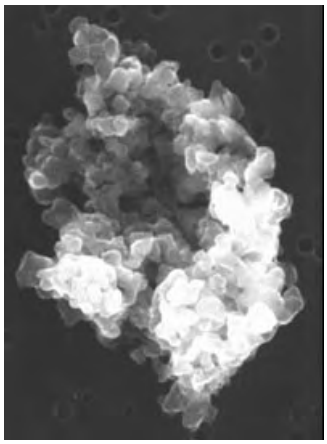


Figure 6. A stratospheric interplanetary dust particle, approximately 11 μm in diameter. Courtesy of NASA.

Sun. These were designated β-meteoroids. In the 1970s, Pioneers 10 and 11 extended measurements of dust densities beyond the orbit of Jupiter. In the 1980s, the Infrared Astronomical Satellite, an orbiting telescope sensitive to thermal emission, discovered numerous spatial structures within the zodiacal cloud in the inner solar system (figure 5). In the 1990s, impact detectors on the Galileo and Ulysses spacecraft detected streams of dust orginating from the Jupiter system. In 1993, using information from the Ulysses dust experiment, Eberhard Grün and colleagues announced the first direct detection of interstellar dust grains within the solar system.

These observations have greatly expanded the issues that can be addressed through studies of interplanetary dust (and debris). Specific dust production mechanisms, such as asteroid collisions and cometary mass loss, can be constrained. The birth of METEOR STREAMS is observed. The dynamical evolution of dust from its source regions can be modeled and the effects of radiation forces and gravitational perturbations from other planets explored. The study of interplanetary dust now offers increased insights into the nature and evolution of ASTEROIDS, COMETS and other source objects.

To date we have been restricted to a vantage point wholely within the cloud, and mostly interior to the dust production regions of the solar system. This makes it impossible to directly observe the edges of the cloud, to determine its full extent and more easily study the interactions between the cloud and the interstellar environment surrounding it. It is also difficult to observe structures which overlap each other along our lines of sight. In the 21st century, small telescopes might be sent beyond the orbit of Jupiter to observe from the outside what appears to be primary dust production regions of the cloud we see from Earth and to measure the extended cloud beyond. In the meantime, the Earth continues to collect dust, meteors and meteorites, from which more science is extracted as their context continues to broaden, and in 2006 the Stardust mission will return to the Earth a sample of dust taken from the coma of a comet.

Composition and morphology

Unmelted interplanetary dust, collected in the upper atmosphere by aircraft and balloons and collected from Antarctic ice, exhibits a range of porosities from delicate 'fairycastle' structures to more compact forms (figure 6). Large particles are more likely to be heated to melting upon entering the Earth's atmosphere. These melted 'cosmic spherules' characterize particles extracted from lake bottoms and soil—primarily because the unmelted particles are difficult to distinguish from terrestrial materials.

Typical compositions of IDPs collected at the Earth are similar to primitive carbonaceous chondritic meteorites (e.g., CI, CM, CR), many with relative excesses of carbon. It is interesting that these meteorites are rare, representing less than 3% of recent meteorite falls. However, details of IDP minerology (particularly evidence of interactions between minerals and solar nebula gas) in the larger (50–500 μm) MICROMETEORITES collected from polar ice suggest that their origin may be more primordial than that of the CARBONACEOUS CHONDRITE parent bodies. This may or may not be the case for the smaller (10 μm) particles collected in the stratosphere.

Dynamics and evolution

To understand the origin of particles collected at the Earth's orbit or detected by spacecraft elsewhere in the solar system we need to understand the forces acting on particle orbits. It is the evolution of these orbits under these forces that gives rise to what we perceive as the zodiacal cloud.

There are four general mechanisms which can affect the orbit of a particle: radiation forces, gravitational perturbations by planets, the interplanetary magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

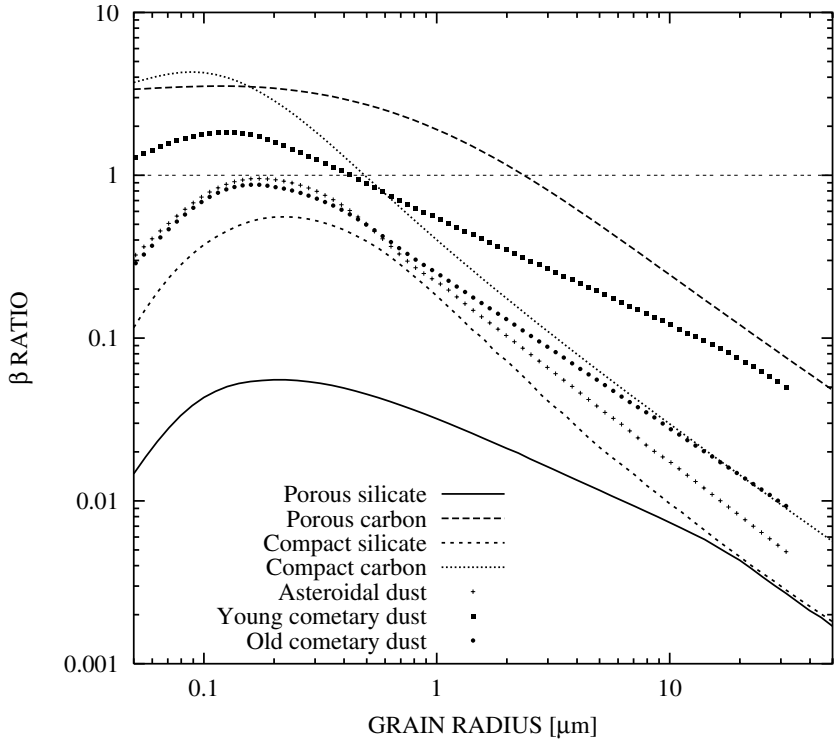


Figure 7. A log–log plot of the ratio β of radiation force to gravitational force as a function of particle radius for materials of different composition and density. Courtesy A Krivov and I Mann, MPI für Aeronomie.

field and collisions. With the exception of gravity, the other forces depend on particle size and mass, with the result that the ultimate distribution of particles in the interplanetary dust cloud will depend upon the size and mass of the particles being discussed.

Radiation forces

As a particle orbits the Sun, it will absorb or scatter solar photons, thereby picking up a component of the photon's momentum away from the Sun. This radiation pressure diminishes the effect of the Sun's gravitational attraction. Its strength is generally described by the ratio β of radiation force to gravitational force. Since both forces diminish as the square of a particle's distance from the Sun, their ratio is the same throughout the solar system, and depends solely on physical properties of the particle: $\beta \sim Q_{\rm pr}/\rho s$, where $Q_{\rm pr}$ is a radiation scattering efficiency factor, ρ is the particle mass density, and s is its radius. Smaller particles are generally more sensitive to radiation pressure than larger particles because their surface-area-to-mass ratio is greater for a given density. However, there comes a point when smaller particles become transparent to a given wavelength of light, and the effect of radiation pressure is diminished. For particles in orbit around the Sun, this transparency begins to occur near a radius of 0.1 μm. Figure 7 shows β as a function of size for spherical particles of various compositions.

If an object is large enough, the effect of radiation pressure is minimal and β is near zero. A small particle, with non-negligible β, chipped off or emitted from a big object ($\beta \sim 0$) will move into a larger orbit in response to the effectively weaker pull of the Sun. If its β is greater than 1, it will be repelled by the Sun and escape the solar system. However, presuming that the particle is ejected from an object already in orbit about the Sun, a much lower β will be required for escape. Figure 8 shows the orbital paths of particles of different β's which are released from a circular orbit. Particles with $\beta < 0.5$ remain in the solar system.

Because of the finite speed of light, particles scattering or absorbing a solar photon will pick up a component of momentum which acts to slow them down that is proportional to v/c, where c is the speed of light and v is the component of particle motion perpendicular to the direction to the Sun. This is known as Poynting–Robertson (P–R) drag. Its effect is maximum at perihelion, where the particle's velocity is maximum. The result is that a particle's orbit will become more circular as it slowly spirals into the Sun. Both radiation pressure and P–R drag act only in the orbital plane of the particle. The timescale (in years) for P–R decay from a circular orbit at a heliocentric distance, R, is $T_{\rm pr} = 400R^2/\beta$. So even a fairly large object having a radius of 1 cm and density of 3 g cm^{-3} (hence $Q_{\rm pr} \sim 1$ and $\beta = 1.9 \times 10^{-5}$) that starts out in the asteroid belt at a heliocentric distance of 2.5 AU in a circular orbit will have fallen into the Sun after 130 000 000 years—a short time in comparison with the age

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

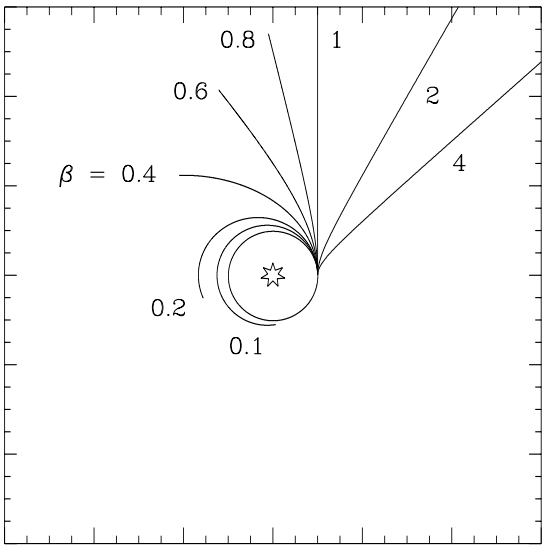


Figure 8. The orbital paths of particles having different values of β, ejected from a large ($\beta \sim 0$) object in circular orbit of radius 1 AU about the Sun. Particles having $\beta > 0.5$ escape the solar system. Particles in repulsive orbits ($\beta > 1$) are typical of comet tail particles. Courtesy of D Lien.

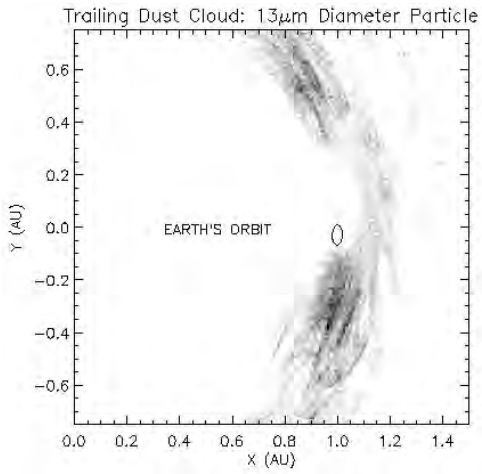


Figure 9. A simulated image of the Earth's resonant ring in a rotating reference frame with the Earth (the Earth's path shown as a small ellipse). Figure courtesy of S Jayaraman, Vanguard Research.

of the solar system. A 10 μm radius particle would take only 130 000 years. Thus, from radiation forces alone, a star's primordial interplanetary dust cloud would be lost on a fairly short timescale.

An interesting radiation force which acts primarily on particles large enough to sustain a temperature gradient from night to day is the Yarkovsky effect. The day side of an object will be warmer and hence emit a greater amount of thermal energy than the dark night side. Since radiation imparts momentum, the day side will feel a greater push away from the Sun than the night side feels a push towards the Sun. The net effect of this radial force on the object's orbit will be negligible. However, because of rotation and the fact that a warm surface takes a while to cool when it is no longer being heated, the evening terminator of the object will be warmer than the dawn terminator. This results in a force component which tends to either decrease or increase the orbital speed of the object, depending on its sense of rotation relative to its direction of motion. As a consequence of the Yarkovsky effect, an object between centimeters and tens of meters in diameter can be moved a significant fraction of an AU over many millions of years. This becomes even more important when combined with the effects of gravitational perturbations by planets.

Solar wind

The SOLAR WIND can also modify dust particle orbits analogously to radiation pressure and P–R drag. The solar wind is a fully ionized, magnetized plasma flowing nearly radially outward from the Sun. It is composed primarily of protons, electrons and alpha particles. The impact of the solar wind on dust particles has an effect that is negligible when compared to radiation pressure, but the 'corpuscular drag' force is about 30% that of P–R drag. This force becomes relatively more important as particles become small enough to be transparent to the solar radiation field. The particle cross sections to the solar wind continue to be described by geometric optics.

Gravitational perturbations by planets

Newton's law of gravity states that each object in the universe attracts each and every other body. The gravitational attraction of a dust particle is negligible, but the attraction of planets can be significant as a particle moves around the Sun. The effects of planets on particle orbits are greatest when a particle comes close to a planet and when it passes through a gravitational resonance. Resonances arise when the orbital motion of a particle is commensurable to that of a perturbing planet, causing an effect on the particle's orbit analogous to a forced harmonic oscillator. In some cases the particle's orbit becomes chaotic, changing wildly in eccentricity and inclination. Sometimes it can be 'trapped', its inward spiral interrupted for a time. This occurs in the vicinity of the Earth where dust is temporarily trapped in a resonant ring extending completely around the Earth's orbit (figure 9). The ring is not symmetric, having a region of higher density which trails the Earth as it moves around the Sun.

Dust interaction with the interplanetary magnetic field

UV radiation from the Sun, incident on a dust particle's surface, will release photoelectrons, resulting in the dust particle becoming positively charged. Equilibrium between the photoelectric current and the currents of the solar wind electrons and ions induces a surface potential of the order of a few volts positive. This causes the particle to interact with the magnetic field being carried away from the Sun with the solar wind. A particle orbiting the Sun

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

near the ecliptic will feel the resultant Lorentz force up or down, generally perpendicular to its orbital plane, as it travels through regions of different polarity, called sectors.

Collisions

In the asteroid belt, objects move with relative speeds of $\sim$5 km s^{-1}, when objects up to tens or hundreds of kilometers in diameter collide 'head on', they will be destroyed or 'catastrophically disrupted'. The mass of the original objects is redistributed into smaller fragments, dust and a small amount of vapor. When smaller dust particles collide the results are the same—mass is redistributed into smaller sizes. Some of the energy of collision goes into fragmentation and some into kinetic energy of the fragments. The latter causes some dispersion in the orbital elements of the fragments. The importance of this mechanism with respect to dust is that mass may be moved from sizes that are relatively insensitive to radiation forces to sizes that are very sensitive. Some amount of mass in 1 μm-sized particles will find itself on escape trajectories from the solar system. Also, some fraction of the fragments of debris having experienced numerous collisions may migrate from a relatively stable orbital position in the asteroid belt to a resonance region which puts them into Earth-crossing orbits.

Sources

The existence of a zodiacal dust cloud and the fact that particle orbits will ultimately decay under P–R drag necessitates a source or sources which replenishes the cloud we observe today. In the 20th century there was much discussion about potential sources of this dust. As more and more asteroids were being discovered, it was thought that the dust giving rise to the zodiacal light as well as METEOROIDS were just the small end of the asteroid size distribution. However, as methods were developed to determine the orbits of meteoroids prior to entering the Earth's atmosphere, it was discovered that many had high inclinations or large eccentricities, more similar to cometary orbits. Only four of 20 particles impacting early experiments on Pioneers 8 and 9 were determined to have asteroidal orbits. Consensus shifted to cometary sources. At the same time, however, estimates of dust produced by all comets fell far short of that felt necessary to generate the observed cloud. The situation was not improved until space-based observations in the thermal infrared revealed significant dust production directly related to both comets and asteroids.

Comets

Comets are known agglomerations of ice and dust. However, the dust we see in COMETARY TAILS is only a fraction of the refractory material comets release. The greatest mass of dust is lost in large particles. Unlike the micron-sized tail particles which stream away from the Sun by radiation pressure, millimeter to centimeter sized particles have low β values and move slowly away from the parent comet, tracing out a portion of the comet's orbit ahead of and behind the comet's position. This is a dust *trail*, the first stage in the evolution of a meteor stream (figure 10). These are not easily detected at visible wavelengths because comet refractory material is very dark. Because dark particles are warmer than lighter particles, the dark particles will be brighter in the thermal infrared—which is how cometary dust trails were discovered. One conclusion from studying these large comet particles is that comets are far rockier than previously thought, by factors of several. This gives comets dust-to-gas mass ratios comparable to that estimated for Pluto, at which distance these dust sources are thought to have formed.

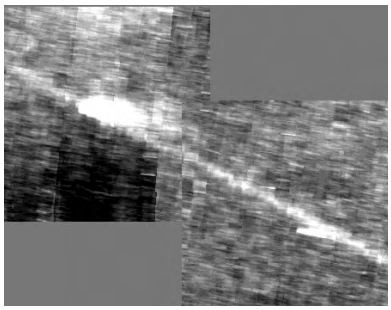


Figure 10. The most prominent dust trail in 1983 was associated with Tempel 2, appearing as a long narrow feature tracing out a portion of its orbit, ahead of and (primarily) behind the comet. Trails are the first stage in the dynamical evolution of meteor streams. They are the principal source of cometary dust contributed to the interplanetary dust complex. M Sykes, University of Arizona.

The cometary contribution to the zodiacal cloud is primarily due to short-period comets. Contributions to the zodiacal cloud from comets newly entering from the OORT CLOUD are small. Most dust particles released will escape the solar system, unless strongly perturbed by Jupiter or another planet into orbits which will retain them in the inner solar system.

Asteroids

Dust production by asteroids was first observed by IRAS (figure 11). IRAS observed bands of dust within the zodiacal cloud, which were associated with the major Hirayama asteroid families, Themis, Koronis and Eros. These families were sites of large disruptive events in the past in which a parent asteroid was shattered into a large number of smaller fragments. Continued disruptions of the smaller of these fragments by collisions with background asteroidal and cometary particles generate the dust we observe today. This process must be repeated throughout the asteroid belt, giving rise to the asteroidal contribution to the zodiacal cloud. While it is thought that the innermost bands, associated with the Themis and Koronis families, may be generating only 5% of the dust in the zodiacal cloud at present, their low inclinations result in their being a significant source of interplanetary dust particles at the Earth's orbit!

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 11. A map of the sky within 30° of the ecliptic plane constructed from IRAS scans high-pass filtered in latitude, showing dust bands (horizontal) encircling the inner solar system. These are associated with dust production within prominent asteroid families—sites of large past catastrophic collisions. Dark vertical bands indicate the areas not scanned during the third and final survey by IRAS which covered 75% of the sky. The angled lines and other noise-like features are associated with the plane of our Milky Way galaxy. M Sykes, University of Arizona.

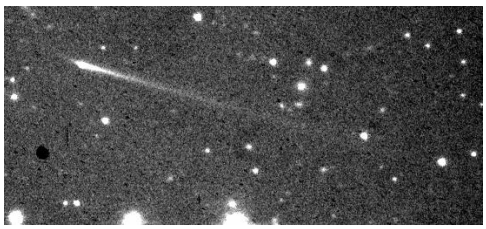


Figure 12. Analysis of the dynamical evolution of dust associated with the 'comet' Elst–Pizarro has led some to hypothesize that it is actually debris from a recent asteroid collision. C Hergenrother, University of Arizona.

Figure 13. An external view of the solar system in the thermal infrared with a spatial resolution of 0.2 AU at an angle of 30° above the ecliptic plane. The zodiacal cloud is simulated using a parametric model constrained by IRAS observations. To enhance the cloud, the image is logarithmically stretched. Jupiter is to the upper left. Earth and Mars are distinguishable to the left and right of the inner cloud, respectively. M Sykes, University of Arizona.

While the dust bands indicate the sites of ancient large collisional events, smaller dust-producing events are occurring in the asteroid belt all the time. Initially, debris from an asteroid collision would assume a distribution in space very similar to that of a comet dust trail. Differences in detail over time would arise due to the fact that the former is the result of a single event, while the latter is a more continuous process (by comparison). The discovery of 'comet' Elst–Pizarro in 1996 may have been one such example (figure 12). It had been detected in 1979, but was stellar in appearance with an asteroid-like orbit (low inclination and low eccentricity), so it was given a provisional asteroidal designation. Analysis of the dust seen in 1996 indicated a short outburst. Some analyses suggest a single event as would be expected for an asteroid collision.

One might hypothesize that if all asteroids are engaged in dust production via collisions then the radial distribution of dust through the asteroid belt should reflect the density of asteroids. To some extent this is true, and is reflected in the observations of the asteroidal dust bands. On the other hand, the orbits of the dust being generated decay under P–R drag, which acts to significantly smooth out local variations. This has been seen by the Ulysses spacecraft as it traveled through the asteroid belt, and swung around Jupiter back towards the Sun. The dust impact experiment on Ulysses indicated a relatively smooth distribution of dust flux from about the orbit of Mars to that of Jupiter. If this dust were dominated by asteroidal sources, it suggests that even the TROJAN ASTEROIDS (in a 1:1 resonance with Jupiter's orbit) are significant generators of dust.

Io

As the Ulysses spacecraft approached within approximately 1 AU of Jupiter, it began to detect streams of tiny particles ($< 0.1\ \mu$m) traveling at extremely high velocities ($> 200\ \mathrm{km\,s^{-1}}$). Analysis of these observations and later observations by the dust detector aboard the Galileo spacecraft showed that these particles originated from the volcanoes on Jupiter's moon Io (see IO: VOLCANISM AND GEOPHYSICS).

Interstellar space

Our solar system is traveling through the Milky Way in a direction that is different from the flow of a local interstellar cloud of gas and dust. As a consequence, dust particles penetrate the solar system and add to the interplanetary dust population, if only temporarily. These particles have been detected by the Ulysses, Galileo and possibly Hiten spacecraft. Their path through the solar system is complicated by gravitational focusing, radiation pressure and Lorentz forces, but there is still a discernable flow from the upstream direction of the interstellar cloud.

The shape of the zodiacal cloud

From the Earth, we are limited in our ability to discern the boundaries and large scale distributions of dust within our solar system, because we are looking from the inside

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

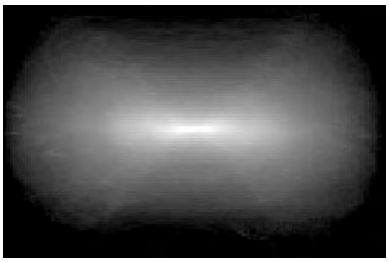


Figure 14. An external view of the zodiacal cloud along the ecliptic plane. A physical model of 75% asteroidal and 25% cometary dust particles is used. Unlike the parametric model, the physical model has a distinctly toroidal shape (planets are not shown). The image is logarithmically stretched. Image courtesy of Keith Grogan, Goddard Space Flight Center.

out. Two kinds of models have been developed to provide a picture of the cloud: parametric and physical. In a parametric model, we assume we know the functional form of the spatial distribution of particles. For instance, for a family of particles spiraling into the Sun by P–R drag, perhaps complicated by mutual collisions and Lorentz forces, the number density of particles would be expected to vary as r^{-k} where r is heliocentric distance. Some collision models and data from the Helios spacecraft suggest that the distribution of dust perpendicular to the ecliptic plane may be describable by a formula that looks like $\exp[-w(|z|/r)^t]$ where z is the distance from the ecliptic plane. The values of the parameters k, w and t are then determined by comparing predictions made by such a model with observations. Complexity can be increased by considering other reasonable conditions, such as the orientation of the cloud with respect to the ecliptic (after all, why should everything be exactly symmetric about the plane of the Earth's orbit?). Mathematically, the parametric model cloud stretches out to infinity in all directions (with decreasing density, of course). Viewed from outside the solar system such models generally provide uniform clouds, increasing in brightness rapidly as one approaches the Sun (figure 13).

With the discovery of the sources of zodiacal dust, the formation of the cloud itself can be simulated by considering those sources and modeling the dynamical and collisional evolution of those particles (figure 14). This is more realistic inasmuch as it deals directly with processes. We are still constrained, however, by the limitations of our observations within the inner solar system. Dust production is expected even beyond Neptune in the KUIPER BELT, through collisions. It is difficult to constrain such mechanisms without direct observations without the foreground contamination of local interplanetary dust.

A complex picture

At the beginning of the 21st century, one can no longer speak of 'interplanetary dust' in a monolithic fashion. The pieces of the puzzle that have been collected to date suggest that the composition and distribution of dust vary with size and location in the solar system. Throughout the inner solar system, the largest debris is probably asteroidal, but as we consider dust below a millimeter to centimeter in size, we would expect to see a significant increase in the contribution due to comets. This may in part offer an explanation for the difference in compositions between collected dust and meteorites at the Earth. However, the much larger fraction of 'primitive' carbonaceous material in micrometeorites may have more than one explanation. Such carbonaceous material is physically weaker than stony material, and asteroids composed of such weaker material may more easily generate dust through collisions in the main belt, which is then removed via P–R drag. Further, observations by Ulysses suggests significant dust production is occurring within the dark and carbonaceous Trojan asteroid population. There is mass fractionation of dust in the inner solar system, because small particles (not on escape trajectories) will spiral into the Sun more quickly than large particles. Because orbital decay times increase with increasing particle size, eventually collision timescales will be shorter. The result is that the largest particles or debris will tend to be found near their source locations in the asteroid belt and near comet orbits, and changes in their spatial distribution will be dominated by mechanisms such as gravitational perturbations—the mechanism by which meteorites are transported to Earth-crossing orbits. As we move away from the Earth's orbit we cross the resonant ring of dust which locally enhances a narrow range of particle sizes. As we move away from the plane of the ecliptic we would find that the composition of the asteroidal component of the dust reflects the composition of higher latitude asteroids.

The advent of space-based observations from telescopes and spacecraft has resulted in linking interplanetary dust directly to its sources, which include both asteroids and comets. The study of dust now provides a new means by which the nature and evolution of those sources can be explored. Better estimates of the physical composition of comets, and hence their formation locations, have been determined. Asteroid collision models now have a new constraint to reproduce. Dust production in our own solar system is the one bit of 'ground truth' we have in understanding how dust around other older stars is being produced today. Ultimately, it will provide us with our best insights into the formation and existence of small bodies in other solar systems.

Bibliography

An overview of dust collection experiments up to 1950 may be found in:

Buddhue J D 1950 Meteoritic dust *Meteoritics* (University of New Mexico)

An overview of interplanetary dust studies through the 1970s may be found in:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hodge P W 1981 *Interplanetary Dust* (New York: Gordon and Breach)

A collection of interplanetary dust studies through the mid-1990s may be found in:

Gustafson B and Hanner M (ed) 1996 *Physics, Chemistry, and Dynamics of Interplanetary Dust (Astronomical Society of the Pacific Conference Series, Volume 104)*

Mark Sykes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interplanetary Network

An interplanetary network is a collection of spacecraft, separated by large distances, equipped with γ-ray burst detectors. By timing the arrival of a burst at each pair of spacecraft, an annulus of possible arrival directions for the burst is derived. With three spacecraft, two possible arrival directions are obtained, corresponding to the two intersections of two annuli (figure 1). With four non-coplanar spacecraft, a unique arrival direction can be determined. The accuracy of the determination depends, among other things, on the separations between the spacecraft. Separations of the order of an astronomical unit or more can give accuracies in the tens of arcseconds range. In many cases, the individual detectors have no direction-finding capability. However, if one or more detectors in the network can determine even a coarse arrival direction, it is often possible to derive a unique location with a three-spacecraft network by eliminating one of the possible intersections.

Interplanetary networks have been in operation since 1977. The current (third) network began in 1991 and at present consists of the Ulysses, Mars Odyssey, Wind and HETE missions. In general the detectors on interplanetary spacecraft are small, and the burst detection rate is about one per week. However, the good location accuracy allows successful searches for γ-ray burst counterparts to take place.

Kevin Hurley

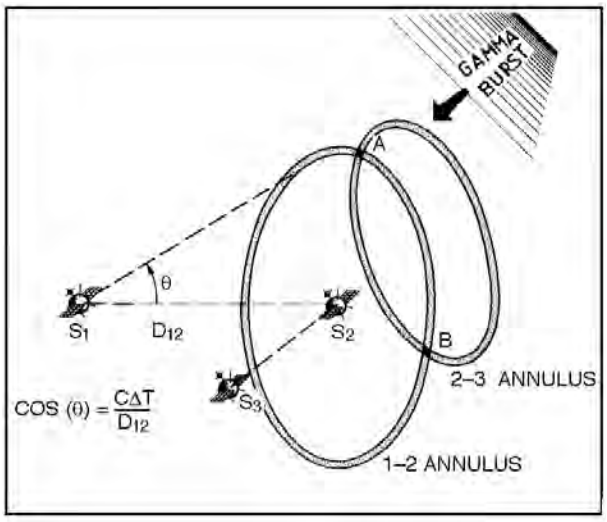


Figure 1. The triangulation method.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Interstellar Absorption Lines

Various constituents of the gaseous interstellar medium absorb radiation from a background source at specific combinations of wavelengths. As a result, there may be one or more localized reductions of intensity (or outright gaps), often referred to as 'lines' (images of a spectrograph's entrance slit—see figure 1), when one views a spectrum of the source. This effect arises from the ability of individual atoms, ions and simple molecules in space to assume unique, well-defined states of excitation (see also SOLAR SPECTROSCOPY: ATOMIC PROCESSES). These states are created by specific combinations of the electron wavefunctions, the coupling of spins or angular momenta of subatomic particles and, for molecules, the additional excitation levels arising from various allowed modes of vibration and rotation of bonded atoms. A transition from a state of low excitation (for interstellar species, usually the lowest possible one) to a more energetic state is made possible by the absorption of a photon having an energy very close to the difference in energy of two levels. A change in the reverse direction results in the emission of a photon. Observations of either of these processes can reveal the amount, composition, general physical state and movement of the interstellar gaseous medium.

Our earliest indication of the existence of gases in space arose from evidence that some absorption lines in the spectra of stars were not produced by the outer parts of the stellar atmospheres, as is usually the case for most of the spectral features that can be seen (e.g. the FRAUNHOFER LINES in the spectrum of our Sun). This finding was first brought out by observations of the BINARY STAR δ Orionis reported in 1904 by J HARTMANN. He found that the absorption features from singly ionized calcium atoms (Ca II) did not exhibit the periodically changing Doppler shifts resulting from the stars' orbital motions (as revealed by other spectral features). Later, other observers noted that these lines were generally stronger for more distant stars, thus ruling out the explanation that the lines arose from gaseous envelopes around the binary systems or the extended atmospheres of unseen companion stars.

Figures 1 and 2 show an example of absorption lines from the doublet of neutral sodium (Na I) at wavelengths of 589.0 and 589.6 nm in the spectrum of the star 6 Cas. The multiple lines within each feature illustrate that different clouds of gas have different Doppler shifts, indicating that they are moving at slightly different velocities projected along the line of sight.

Atoms and molecules have many transitions between excited electronic levels well above the ground state, but such features never arise from interstellar gases because all of the species are in levels with the lowest possible energy[1].

[1] An exception is that some atoms can be found in excited fine-structure states within the lowest electronic level. The relative populations in these states can give important information on local physical conditions in the gas, since they are populated by collisions with other atoms, even at the low density of the interstellar medium; see figure 7.

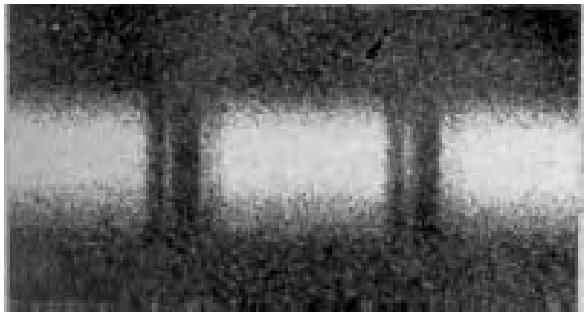


Figure 1. A print of the original photograph of a small portion of the spectrum of the star 6 Cas in the vicinity of the two D lines of neutral sodium. The intensity of starlight at successively longer wavelengths was recorded as a bright band spread from left to right, with vertical interruptions ('lines') caused by absorption from the intervening sodium atoms in space. (From Münch G 1957 *Astrophys. J.* **125** 42–65 © American Astronomical Society.)

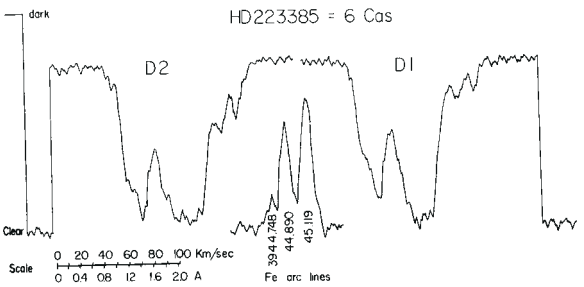


Figure 2. A densitometer tracing of the photographic exposure (hence 'dark' means a strong exposure level and 'clear' means little or no exposure) depicted in figure 1. The D2 atomic transition is twice as strong as the one for D1, and they are separated by 0.6 nm ($\equiv$6 Å on the scale shown at the bottom). Peaks and valleys within the D2 and D1 features are caused by absorbing clouds that have different Doppler velocity shifts. Emission lines from an iron arc are used to establish an accurate wavelength scale. (From the same source as for figure 1 © American Astronomical Society.)

By contrast, the atmosphere of a star has a large enough density and high enough temperature to create frequent collisional excitations with enough energy to populate the excited levels, thus leading to a much richer collection of spectral features.

A fundamental challenge in the investigation of interstellar lines is to separate them from the plentiful stellar lines. The easiest way to accomplish this is to observe only those stars that are rotating rapidly enough to make their lines wash out, owing to the superposition of many Doppler shifts from points on the star's surface that are moving at different radial velocities.

Constituents that can be observed at visible wavelengths and thus are available to ground-based, optical observatories include Ti II, Na I, Ca I, Ca II, K I, NH, CN, CH and C_2. In addition to these features, there are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

more than 100 broad absorptions called diffuse interstellar bands that, for the most part, have origins that are poorly understood. These bands are detected only through significant accumulations of interstellar material, and it is generally acknowledged that they probably arise from complex, polyatomic molecules. It is clear that many types of molecules are involved, because different lines of sight exhibit features that have dissimilar relative strengths.

Unfortunately, all of the species identified with transitions in the visible part of the spectrum are minor ones, and, moreover, the atomic transitions (with one exception, Ti II) apply to stages of ionization that are not generally favored for most of the interstellar medium. Thus, while revealing the presence of gaseous matter and its motions, the lines in the visible part of the spectrum do not tell us much about the composition of the gas. Fortunately, other spectral regions offer much of the missing information. The ultraviolet part of the spectrum (91.2 nm–310 nm; see ULTRAVIOLET ASTRONOMY)[2] has a great abundance of strong lines from important atomic and molecular species, with many elements offering choices of features from several different degrees of ionization. The infrared and submillimeter regions (3 μm–1 mm) offer insights on the presence of many different molecules in dense interstellar clouds, with a number of transitions at wavelengths in the tens of centimeters (see INFRARED ASTRONOMY and MILLIMETER AND SUBMILLIMETER ASTRONOMY). These molecules can be seen in emission or, if there is a background continuum source, also in absorption. Finally, there is the very important (but very weak) transition of neutral hydrogen at a wavelength of 21 cm (or frequency = 1420.406 MHz) caused by a reversal in the alignment of the proton and electron spins (see RADIOASTRONOMY).

An overview of the composition of the interstellar medium is presented in INTERSTELLAR MATTER. Briefly, the most abundant chemical element in space is hydrogen, making up about 91% of all the atoms. Helium is the second most abundant element, and it accounts for nearly all of the remaining 9%. The heavier elements in the periodic table add up to only 0.13% of the total (by number of atoms, not by mass). Large fractions of many of the heavier elements have condensed into solids, in the form of small dust grains (see INTERSTELLAR GRAINS).

Within the disk of our Galaxy (the Milky Way) the average density of the gas is about 1 atom cm^{-3}, but there are enormous contrasts that range from about 10^{-3} atom cm^{-3}, where all of the atoms are ionized and very hot, to 10^4–10^5 atom cm^{-3} in cold clouds that are known from their emission lines in the radio to have complex molecules (see INTERSTELLAR MOLECULAR CLOUDS), or to even higher densities in regions that are collapsing to form stars. Over the distances to typical hot, luminous stars that can be seen by the unaided eye, a representative column of intervening atoms (mostly hydrogen) corresponds to a surface density in projection of about 3×10^{20} atom cm^{-2}. Put in terms of our everyday experience, this is equivalent to the number of air molecules at sea-level pressure and 0 °C in a column that is 11 cm long. For minor constituents that can be detected in optical or ultraviolet absorption lines, a typical detection limit for a stellar spectrum taken with a modern spectrograph that can give a high signal-to-noise ratio and good spectral resolution is of order 10^{10} atom cm^{-2}. Again using a terrestrial example, this is equivalent to the surface density of gold atoms contained within a column of sea water that is 1 cm long.

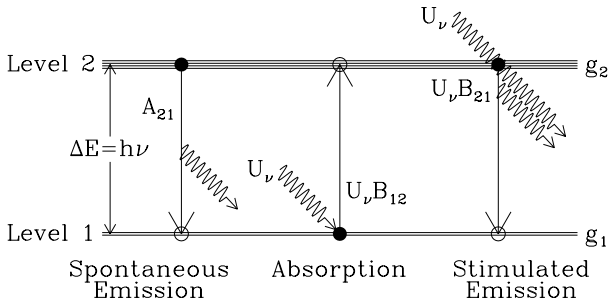


Figure 3. Three types of radiative interactions driven at rates A_{21}, $U_\nu B_{12}$ and $U_\nu B_{21}$ between two atomic levels separated by an energy difference ΔE. The statistical weights (number of degenerate quantum states within the levels) for levels 1 and 2 depicted here are $g_1 = 2$ and $g_2 = 4$. The Einstein coefficients A_{21}, B_{12} and B_{21} are related to each other in the manner shown by equation (1).

Fundamentals of radiative interactions

A single atom[3] interacts with radiation in three fundamental ways that were originally defined by EINSTEIN and are illustrated in figure 3. First, an atom in an excited level can decay to the lower state spontaneously and emit a photon with an energy $h\nu$ very nearly equal to ΔE, the energy difference between the two levels. The probability of this happening per unit time is given by the quantity A_{21}, a property of the radiative transition between the levels called the Einstein A coefficient (or transition probability). For transitions that are observed in interstellar studies, values for A_{21} range from 2.9×10^{-15} s^{-1} for the 21 cm transition of neutral hydrogen to a few $\times\, 10^9$ s^{-1} for the strongest UV lines between electronic excitation levels.

The other two modes of interaction, regulated by the rate coefficients B_{12} and B_{21} called the Einstein B coefficients, occur when the atom is exposed to a radiation

[2] Because the Earth's atmosphere is opaque to the UV radiation, these wavelengths must be observed by telescopes in orbit. The region below the Lyman limit at 91.2 nm is also useful in principle, but the intensities of all but the nearest stars in the sky are strongly attenuated by interstellar hydrogen atoms as they become ionized by the radiation. In the soft x-ray region, the opacity from hydrogen ionization is much lower. However, spectroscopic findings based on x-ray observations have been limited by the shortcomings of currently available instruments.

[3] Hereafter, for brevity, the term 'atom' is understood to mean a neutral atom, an ion or a molecule.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

field. If photons with an energy $h\nu = \Delta E$ are present with an energy density U_ν (per unit frequency), transitions can be induced at a rate $U_\nu B_{12}$ from level 1 up to level 2. The absorption of the photon supplies the necessary energy for this to happen. The reverse of this process is also possible, i.e. a photon can induce an atom in level 2 to decay to level 1, and in so doing, a new photon is created that has the same frequency, phase and direction of the one that caused the event. This process, called stimulated emission, occurs at a rate $U_\nu B_{21}$. There are instances where ΔE is small and, as a result, there are enough atoms in the upper level to create enough stimulated emission to act as a negative absorption that partly cancels the normal absorption. This effect is significant for many transitions in the radio region.

The Einstein A and B coefficients are related to each other by the equation

$$g_1 B_{12} = g_2 B_{21} = \frac{c^3}{8\pi h\nu^3} g_2 A_{21} \tag{1}$$

where g_1 and g_2 are the statistical weights of the lower and upper levels, respectively. This relationship arises from the requirement that if the atoms were totally immersed in a hollow volume that had an interior radiation field with a Planck black body energy density distribution

$$U_\nu = \frac{8\pi h\nu^3}{c^3(e^{h\nu/kT} - 1)} \tag{2}$$

with a temperature T, they must establish level populations n_1 and n_2 consistent with a Boltzmann distribution:

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} \exp\left(-\frac{\Delta E}{kT}\right). \tag{3}$$

That is, without any coupling to other process, the atoms must be in thermal equilibrium with the radiation field. This can only happen when the coefficients for the transitions satisfy equation (1)[4].

Since the lifetime of the upper state is finite, the uncertainty principle allows radiative processes to occur with photons that have energies $h\nu$ slightly different to ΔE. The relative response as a function of frequency ν is given by the Lorentz profile

$$\phi(\nu) = \frac{\gamma}{4\pi^2[(\nu - \nu_0)^2 + (\gamma/4\pi)^2]} \tag{4}$$

where ν_0 is the central frequency of the transition having a reciprocal mean upper state lifetime γ (not necessarily equal to A_{21} because other processes, such as collisions or other transitions, may shorten the state's lifetime even more).

Instead of having an atom inside a radiation bath with an energy density U_ν, we may think of an equivalent situation where the atom is exposed to a highly directed beam with a photon flux per unit area and frequency $N_\nu = cU_\nu/h\nu$. The atom's transition rate from level 1 to level 2 then becomes $(h\nu N_\nu/c)B_{12}$. If N_ν varies with ν, the rate becomes $(h\nu/c)B_{12}\int N_\nu\phi(\nu)\,d\nu$. It follows that the effective geometrical cross section of an atom to remove photons from a beam is simply

$$\sigma(\nu) = \frac{h\nu}{c}\phi(\nu)B_{12} = \frac{g_2}{g_1}\frac{c^2}{8\pi\nu^2}\phi(\nu)A_{21}. \tag{5}$$

Interpretations of the observations

By custom, representations of the strengths of transitions seem to differ for investigators who work with different spectral regions or types of transitions. Many radio transitions are expressed in terms of A_{21} defined above, but most of the molecular transitions have their strengths defined by the squared modulus of the dipole matrix element $|\mu|^2$ used for calculating transition probabilities of electric dipole transitions:

$$A_{21} = \frac{g_1}{g_2}\frac{64\pi^4\nu^3}{3hc^3}|\mu|^2. \tag{6}$$

For transitions in the visible and ultraviolet, it is customary to speak in terms of a quantity f, called the oscillator strength or sometimes just the f value. This quantity refers to the cross section of an atom (equation (5)), but expressed in terms of the outcome from a classical derivation of $\int\sigma(\nu)\,d\nu = \pi e^2/m_e c$ for an electron with a charge e and mass m_e that is bound in a harmonic potential. For most discussions in a context of absorption lines, the applicable f value is the one that goes from the lower to the upper state f_{12}, but the subscript is usually omitted. The alternative forms of equation (5) are

$$\sigma(\nu) = \frac{8\pi^3\nu}{3hc}|\mu|^2\phi(\nu) \tag{7}$$

and

$$\sigma(\nu) = \frac{\pi e^2}{m_e c} f\phi(\nu). \tag{8}$$

A collection of atoms having a differential number distribution $N(v)$ as a function of velocity v projected along the viewing direction will present a combined cross section

$$\sigma(v) = \frac{\pi e^2}{m_e c} f\lambda \int_0^\infty \phi(\nu)N[v + \lambda(\nu_0 - \nu)]\,d\nu \tag{9}$$

if all of them are in the lower level. (λ is the central wavelength of the transition, c/ν_0.) Generally, the Doppler shifts in frequency caused by velocities over the extent of $N(v)$ are very much larger than the line's natural width γ. Given that $\int_{-\infty}^{\infty}\phi(\nu)\,d\nu = 1$, we can simplify equation (9) to the form

$$\sigma(v) = \frac{\pi e^2}{m_e c} f\lambda N(v) = 0.026\,54\, f\lambda N(v) \tag{10}$$

if we ignore the small contribution to $\phi(\nu)$ from the low-level, broad Lorentz wings that are well beyond the

[4] The B coefficients are sometimes defined in terms of a radiation energy flux J_ν instead of energy density U_ν. These coefficients are equal to $4\pi/c$ times the ones described here.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

frequency shifts due to Doppler motions. The explicit details of equation (9) are important only for very strong lines in the ultraviolet, which will be discussed later. In other applications, cross sections can simply be expressed in terms of either equation (10) or equations (5) and (7) with $\phi(v)$ replaced by $\lambda N(v)$.

If some of the atoms are excited to an upper level, the stimulated emission from this level reduces the cross section by a factor $1 - \exp(-\Delta E/kT_e)$, where T_e is the excitation temperature of the levels defined by the Boltzmann equation (equation (3)). An extreme example of this is with the 21 cm transition of neutral hydrogen. The levels have a separation $\Delta E/k = 0.068$ K, and the populations are very strongly coupled to the gas's kinetic temperature (of order 20–100 K) because collisions populate the upper level much more frequently than its radiative decay rate. Thus the net absorption is dependent on the very small difference between absorption and stimulated emission, giving an absorption coefficient that is proportional to $N(v)/T$. (An even more extreme example is sometimes observed for transitions of the OH radical ($\lambda = 18$ cm) or H_2O ($\lambda = 1.35$ cm), where some source of energy populates the upper level so that $g_1 n_2 > g_2 n_1$, making T_e negative, and the region becomes a maser radio source; see RADIO SOURCES.) For visible and UV transitions, n_2 is almost always equal to zero because the states undergo radiative decay so rapidly. When this is true, we can neglect stimulated emission.

So far, we have not yet addressed the issue of how the collective absorption by the N atoms that contribute to $\sigma(v)$ produces an observable attenuation of a beam of radiation. Consider that these atoms are distributed with a uniform density and confined within some volume that has a cross sectional area A perpendicular to the beam. As the radiation starts to traverse this volume, it encounters an infinitesimal fraction p of the atoms which block the beam and reduce its intensity $I(v)$ by a relative factor $p\sigma(v)/A$, i.e. $\mathrm{d}I(v)/I(v) = -[\sigma(v)/A]\,\mathrm{d}p$. The cumulative effect of successive encounters with the remaining fractions of atoms is found by integrating the expression

$$\int_{I_0(v)}^{I(v)} \frac{\mathrm{d}I'(v)}{I'(v)} = -\frac{\sigma(v)}{A}\int_0^1 \mathrm{d}p \tag{11}$$

which, in turn, leads to an exponential absorption law for the ultimate fraction of the radiant energy that is transmitted,

$$\frac{I(v)}{I_0(v)} = \mathrm{e}^{-\tau(v)}. \tag{12}$$

In this expression, the optical depth $\tau(v)$ is identical to $\sigma(v)$ with the understanding that $N(v)$ now denotes the number of atoms per unit velocity and per unit area in projection. (Henceforth, we will adopt this definition of $N(v)$ and work with the quantity $\tau(v)$.)

Generally speaking, $N(v)$ can be a complicated function that produces absorption profiles with fairly arbitrary shapes, such as those shown in figure 2. If the resolving power of a spectrograph is so high that no detail in the structure of the line is lost by blurring, one can derive $N(v)$ from an observation by inverting equations (10) and (12),

$$N(v) = \frac{m_e c}{\pi e^2 f \lambda}\ln\left[\frac{I_0(v)}{I(v)}\right] \tag{13}$$

and estimating $I_0(v)$ by interpolation of the flux on either side of the line[5]. This method is the ideal way to interpret an absorption line observation, but its accuracy can be compromised, sometimes seriously, if (1) the velocity structure in the line is not fully resolved by the instrument, (2) the intensity in some part of the profile is very near zero (as it is in the example shown in figure 2) or (3) either the reading that corresponds to a zero intensity or the level of the continuum intensity $I_0(v)$ is hard to measure.

For either of the first two reasons stated above, one may have an observation of generally good quality but not be able to obtain a direct measure of $N(v)$. We consider first the simplest case where the details in the velocity structure of a line are degraded, or in extreme cases completely lost, by an instrument with limited wavelength resolution. Here, the intensity profile I_λ (not $\tau(\lambda)$) is smoothed by a convolution with the instrument's response function $S(\lambda)$.

If the absorption at all wavelengths within the feature is so weak that the approximation

$$\tau(\lambda) \approx \frac{I_0(\lambda) - I(\lambda)}{I_0(\lambda)} \tag{14}$$

is valid to a needed level of accuracy, we can determine the total number of absorbing particles in projection along the line of sight, the column density $N_{\mathrm{total}} \equiv \int N(v)\,\mathrm{d}v$, by the relation

$$\begin{aligned} N_{\mathrm{total}} &= \frac{m_e c^2}{\pi e^2 f \lambda^2}\int \frac{I_0(\lambda) - I(\lambda)}{I_0(\lambda)}\,\mathrm{d}\lambda \\ &= \frac{1.13\times 10^{19}}{f\lambda_{\mathrm{nm}}}\frac{W_\lambda}{\lambda}. \end{aligned} \tag{15}$$

(This equation is an integral form of equation (13) with the approximation of equation (14) and the transformation $\mathrm{d}v = c\,\mathrm{d}\lambda/\lambda$.) The integral in equation (15) is called the equivalent width of a line and is usually expressed in wavelength units, i.e. W_λ. (λ_{nm} is the wavelength expressed in the units of nanometers and N_{total} is in atoms cm^{-2}.) A geometrical interpretation of W_λ is that it represents the width of a box that extends from the continuum level to zero intensity and has the same area as the space enclosed by the reduced (absorbed) intensity $I(\lambda)$ and the continuum level $I_0(\lambda)$. This concept is illustrated by figure 4 (but, in this depiction, for a line that is too strong for equation (15) to be valid). The value

[5] Almost all of the modern methods of detecting intensity use detectors that give a linear response. However, for a densitometer tracing of a photographic exposure, such as the example shown in figure 2, this is not true. One must measure the response of the photographic film to varying levels of illumination and then use this information to apply a correction to each of the film density (darkness) values in the spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

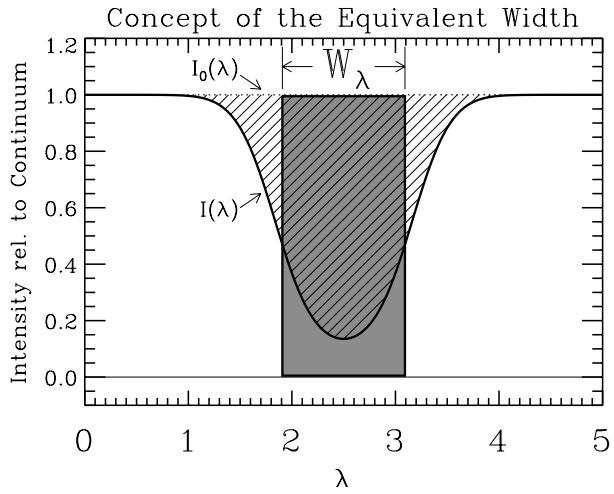


Figure 4. An illustration that demonstrates the concept of an equivalent width, as discussed in the text.

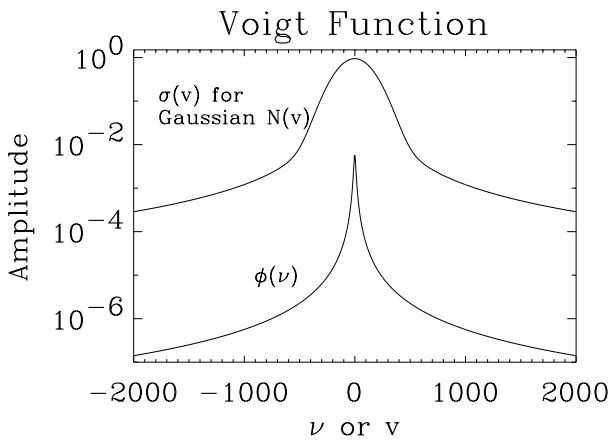


Figure 5. A semilogarithmic display of Voigt profiles for (lower curve) an atom at rest (giving the simple profile form $\phi(\nu)$ of equation (4)), compared with the combined effect of many atoms (equation (9)) having a Gaussian distribution of velocities (upper curve), as described in equation (16). Both scales in this representation are in arbitrary units.

of W_λ is unchanged by instrumental smoothing, making it a useful general property of a line and one for which we can obtain an answer for $N_{\rm total}$ that is not dependent on the instrument's $S(\lambda)$.

When a line becomes strong enough that it starts to become saturated, the approximation in equation (14) is no longer valid, and the linear relationship between N and W_λ expressed in equation (15) breaks down. A line with moderate saturation is illustrated in figure 4. If the line is well resolved by the instrument, we can go back to equation (13) and integrate it over velocity to obtain $N_{\rm total}$. If one has reason to believe that the line is not well resolved and is saturated, the value of N cannot be determined unless some assumptions are made or some independent information is available about the shape of the line.

A common practice is to assume that the line is formed by atoms that have a Maxwellian speed distribution, resulting in a velocity projected along the line of sight that follows the familiar Gaussian distribution centered at a macroscopic radial velocity v_0,

$$N(v) = \frac{N_{\rm total}}{\sqrt{\pi}b} \exp\left(-\frac{(v-v_0)^2}{b^2}\right) \tag{16}$$

where b is a velocity dispersion parameter (often called a b value or Doppler parameter). This assumption has its roots in the early, but usually mistaken, belief that the Doppler shifts are always dominated by the thermal motions of the atoms with $b = (2kT/m)^{1/2}$ (or perhaps turbulent motions that resemble thermal ones), rather than unpredictable macroscopic dynamical processes that could lead to profiles with very different shapes. Moreover, it overlooks the fact that one may be viewing a profile composed of multiple components, perhaps even superposed on each other (figure 2 is a good illustration of this situation). In spite of these shortcomings, the formalism is still very useful.

When $N(v)$ obeys the Gaussian form given in equation (16), the outcome for $\sigma(v)$ (or $\tau(v)$ for a column density representation of $N(v)$) in equation (9) is a Voigt profile. One example of this profile is depicted by the upper curve in figure 5. Its upper, parabolic portion (above 10^{-2} in this particular representation) gives optical depths that conform to equation (10) and is the only portion that can be detected for weak or moderately saturated lines. Only when the lines are exceptionally strong and saturated does one have a chance of seeing the effect of the broad Lorentz wings, as with the Lα and Lβ profiles in figure 6.

When the Voigt profile starts to become saturated, its equivalent width falls short of an extrapolation of the linear trend shown in equation (15). It now takes on the form

$$\frac{W_\lambda}{\lambda} = \frac{2bF(\tau_0)}{c} \tag{17}$$

where $F(\tau_0)$ is a special function of the line's central optical depth

$$\tau_0 = \frac{\pi^{1/2}e^2 f\lambda}{m_{\rm e}c}\frac{N_{\rm total}}{b} \tag{18}$$

given by

$$\begin{aligned} F(\tau_0) &= \int_0^\infty [1-\exp(-\tau_0\,{\rm e}^{-x^2})]\,{\rm d}x \\ &= \frac{\pi^{1/2}}{2}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}\tau_0^n}{n!n^{1/2}}. \end{aligned} \tag{19}$$

(In the limit $\tau_0 \ll 1$, $F(\tau_0) = \sqrt{\pi}\tau_0/2$ and equations (17)–(19) reduce to equation (15).)

If one can measure W_λ/λ for two or more lines of different strengths from a particular atom, it is then possible to derive within this formalism the values of $N_{\rm total}$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

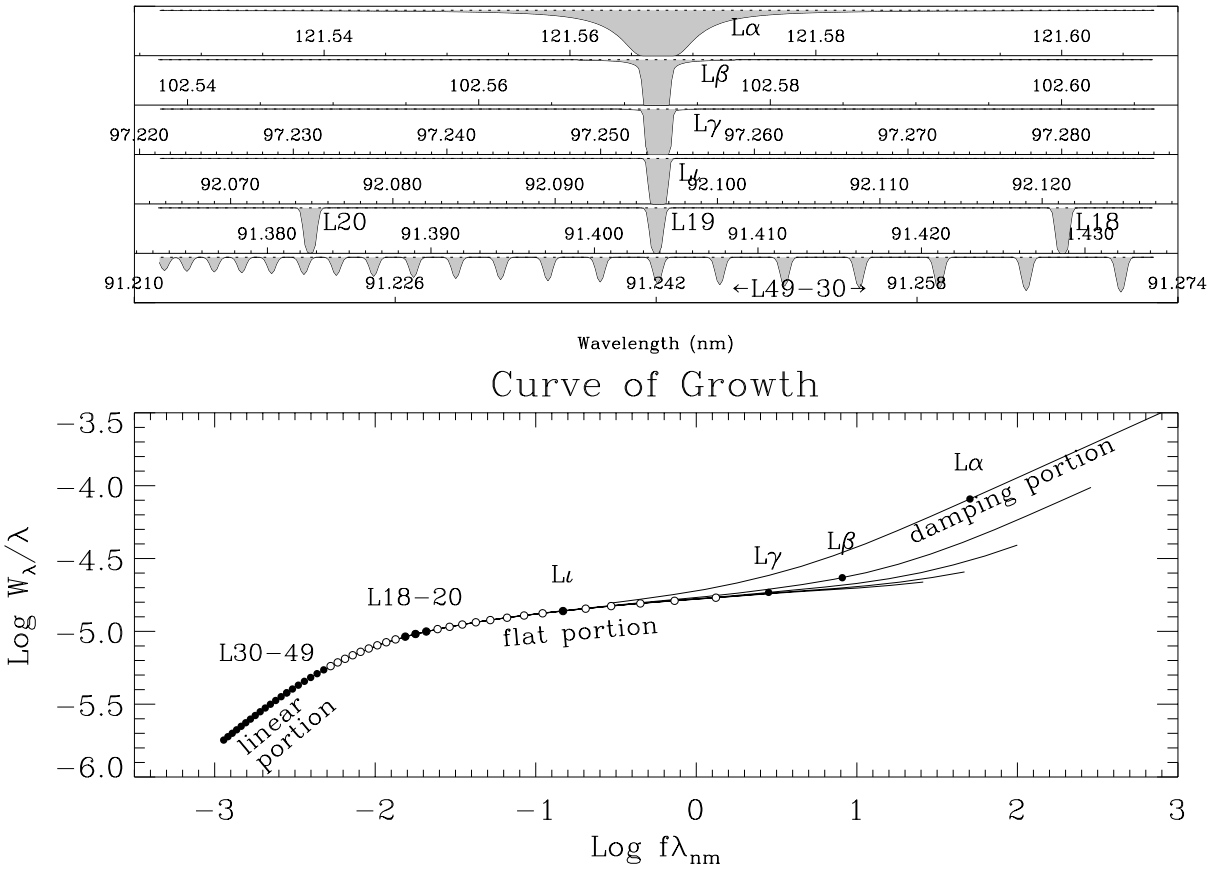


Figure 6. Selected line profiles in the Lyman series of hydrogen displayed over a radial velocity range of ±100 km s^{-1} (top panel) and a plot that shows where their equivalent widths occur on a curve of growth (bottom panel). In this example, $N_{\rm total}$(H I) = 2 × 10^{16} atom cm^{-2} and the velocity profile is a pure Gaussian (equation (16)) with b = 1 km s^{-1}. Different members of the Lyman series have different damping curves because their γ values are different (see equation (21)). (The condition that b = 1 km s^{-1} for hydrogen is an artificial one that was imposed to give the best demonstration of the principle of the curve of growth. Generally, interstellar or intergalactic hydrogen is observed to have $b > 10$ km s^{-1} when $N_{\rm total}$ is as low as 2 × 10^{16} atom cm^{-2}.)

and b by finding the best fit with a curve of growth[6]—a curve that takes on a universal shape for $\log(W_\lambda/\lambda) - \log(b)$ as a function $\log(N_{\rm total} f\lambda) - \log(b)$ for a Gaussian velocity distribution (equation (16)). Opportunities to perform such an analysis arise either from transitions that are completely unrelated or, alternatively, from a group of spectral features, called a multiplet, between a pair of electronic levels. The transitions in a multiplet are separated from each other by the slightly different energies that arise from different couplings of the electronic spins and angular momenta.

A curve of growth is illustrated in figure 6 for the Lyman series of atomic hydrogen[7]. Lines weaker than about L30 (only L30–L49 are shown) exhibit a nearly linear trend between W_λ/λ and $f\lambda$. The intensity profiles of lines that are found on the linear portion of the curve of growth (bottom sequence of features in the top panel of figure 6) approximately reflect an upside down representation of the shape of $N(v)$. As the transitions become stronger, their equivalent widths are described by equations (17)–(19) and their shapes become distorted into a more box-like profile. Lines in this regime are said to lie on the flat portion of the curve of growth. Finally, as the lines become exceptionally strong, the broad Lorentz (or damping) wings of the Voigt profile become evident and the equivalent widths reflect this additional contribution. Lines that exhibit this

[6] A misapplication of the curve of growth fitting with strongly saturated lines can lead to egregious errors in $N_{\rm total}$ (usually resulting in significant underestimates). This is especially true if narrow absorption features are hidden in a broad line whose equivalent width is dominated by a small amount of gas with a large value of b. In less extreme situations, such as a blended ensemble of lines with only mildly different bs and strengths, the underestimate in $N_{\rm total}$ that results from the oversimplification that there is a single component is relatively mild: it is usually of order 20% or less.

[7] The Lyman series is a progression of transitions out of the innermost (and lowest-energy) electron orbit of the atom. They lie between the ionization limit of hydrogen at λ = 91.25 nm and the Lα feature at 121.57 nm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

property are said to lie on the damping portion of the curve of growth.

If the lines are completely dominated by the damping wings, the optical depth is given by the product of $N_{\rm total}$ and each atom's $\sigma(\nu)$ (equation (4) in the limit that $\nu - \nu_0 \gg \gamma/4\pi$ substituted into equation (8)):

$$\tau(\nu) = N_{\rm total} \frac{e^2 f \gamma}{4\pi m_{\rm e} c(\nu - \nu_0)^2}. \qquad (20)$$

By integrating this expression over all ν, it follows that

$$N_{\rm total} = \frac{m_{\rm e} c^3}{\lambda^2 e^2 f \gamma} \left(\frac{W_\lambda}{\lambda}\right)^2. \qquad (21)$$

If possible, one should try to identify the damping wing shapes to confirm that indeed equation (21) is the valid one to use, instead of equations (17)–(19) that might, instead, apply to a portion of the curve of growth that has about the same slope.

When some of the velocity structure in a profile is partly resolved by the instrument so that more information is present than just W_λ, one can invoke an approach that is more sophisticated than the curve of growth analysis. A model for $N(v)$ is constructed on the assumption that it can be approximated as the sum of separate Gaussian components, each with its own $N_{\rm total}$, v_0 and b. One then solves (or searches) for the best fit between the recorded data and the expectation

$$\frac{I(v)}{I_0(v)} = \exp\left(-\sum_{i=1}^{n} \tau_i(v)\right) * S\left(\frac{v\lambda}{c}\right) \qquad (22)$$

an expression that substitutes $\tau(v)$ for $\sigma(v)$ in equation (9) for each of the n components and uses the definition for $N(v)$ given in equation (16) ($*$ denotes convolution). This is usually done by varying the free parameters $N_{{\rm total},i}$, $v_{0,i}$ and b_i until a minimum χ^2 is found, indicating that the best fit has been achieved. A shortcoming of this method is that one may be forced to make an arbitrary choice between two or more solutions that have equally good fits with the data (and, likewise, the number n of possible components is a matter of judgment).

Insights derived from the lines

Lines in the visible part of the spectrum have been used to explore the kinematics of the interstellar material as well as to bring out the general statistical properties of small clumps of gas and how they are distributed in space. These studies have been facilitated by spectrographs with very high resolving powers, some being able to differentiate Doppler shifts that change the wavelength by only one part in a million. Observations of molecular features at visible wavelengths have given insights on the presence of a few simple diatomic species. Measurements in the infrared and radio parts of the spectrum have completed the picture for more complex molecules, usually residing in clouds that are too dense to view at ultraviolet and visible wavelengths.

A particularly noteworthy observation was the discovery by A McKellar in 1941 that CN molecules in front of the star ζ Ophiuchi had their excited rotational levels populated according to a Boltzmann distribution (equation (3)) for a temperature slightly below 3 K. The significance of this was not fully appreciated until the discovery of the cosmic microwave background radiation some 22 yr later (see COSMIC MICROWAVE BACKGROUND).

The UV spectrometer and telescope on the COPERNICUS (OAO-3) satellite that operated from 1972 to 1981 opened up an important spectral range that contained the Lyman series of hydrogen, strong lines of elements ranging from carbon to zinc in the periodic table and, perhaps most significant, the only means (at the time) for detecting molecular hydrogen—the most abundant molecule in the universe. From spectra obtained by Copernicus, we learned from the absorption features of five-times-ionized oxygen (O VI) about the presence of gases at temperatures of around 3×10^5 K (hotter gases with $T \sim 10^6$ K are revealed by a diffuse, soft x-ray background emission). We also discovered that the gas-phase abundances of various heavy elements are significantly reduced as they condense into solid form to create the dust grains. The relative completeness of this depletion is increased for lines of sight where the density is large, and elements that can form refractory chemical substances tend to be more depleted than ones that form more volatile compounds.

The launch of the HUBBLE SPACE TELESCOPE (HST) in 1990 started a new era in the investigation of absorption lines in the ultraviolet. Vast increases in collecting area and spectrograph resolution over those of previous orbiting observatories have significantly improved our ability to record spectra with better precision and over much greater distances (toward fainter objects). We can now gather many spectral elements simultaneously. This has been facilitated by advances in light detection technology, starting from zero-dimensional detectors on board Copernicus (scanning photomultiplier tubes), moving on to one-dimensional ones (linear arrays) on the first-generation instruments of HST and finally arriving at two-dimensional sensors on a second-generation instrument called the Space Telescope Imaging Spectrograph (STIS). Figure 7 shows a small segment of the spectrum of a star (one echelle spectrum order out of 53 obtained in a single exposure). The wavelength resolving power of this spectrum is 1 part in 200 000, a figure that allows us to discriminate Doppler velocity differences of only 1.5 km s^{-1}. To appreciate the the better information content provided by higher resolutions, one can compare the fine details and the velocity scales shown in this figure with those of figure 2. With the added capabilities of HST, we have been able to survey the abundances of elements with very low abundances, probe the interstellar abundances and conditions toward stars in the halo of the Galaxy and (at lower resolution) examine absorption features from gas systems associated with other galaxies, using some of the brightest quasars and active galactic nuclei as the background light sources.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

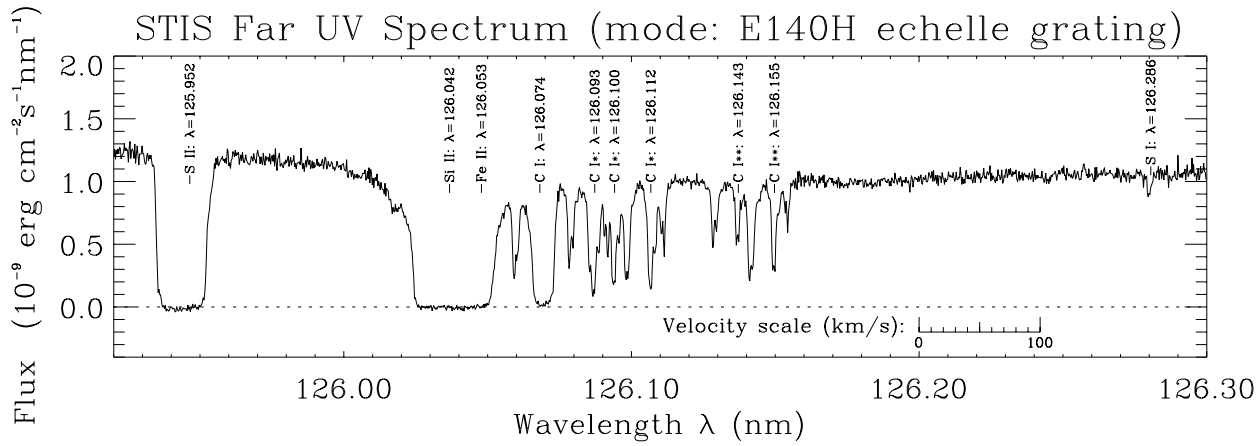


Figure 7. A small portion of the spectrum of the star λ Cep, recorded by the STIS on the HST, using a grating configuration that gave the highest resolution. Transitions are identified with markers aligned with the strongest velocity component at a velocity of -13 km s^{-1} in a heliocentric reference frame. For neutral carbon (C I), there are three main velocity components at -33, -13 and -3 km s^{-1}. Five out of the six C I transitions arise from two excited fine-structure levels in the ground electronic state (C I* is the first excited level and C I** the second one). The strengths of these features, relative to that out of the lowest level, indicate the local densities and temperatures of the gas where the carbon atoms are located. The wavelengths of the Si II and Fe II transitions are so close to each other that their absorption features overlap.

By virtue of the very large (Doppler) redshifts produced by the Hubble expansion of the universe, we can observe with large-aperture telescopes on the ground the UV transitions from distant gas clouds that are in front of even more distant quasars (see INTERGALACTIC MEDIUM and QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES). Intergalactic gas clouds of almost pure hydrogen were discovered by the numerous Lyman α absorption features that appear at wavelengths shortward of the quasar's red-shifted Lyman α emission feature (see LYMAN ALPHA FOREST). These clouds are mostly ionized, but they still contain enough neutral hydrogen to create conspicuous absorption features, including also the higher members of the Lyman series (shown in the very simplistic representations of the top panel of figure 6). Some sight lines to quasars penetrate foreground galaxies or their outermost fringes (see GALAXIES AT HIGH REDSHIFT). Such cases have given us opportunities to explore the character and composition of the gas in these distant systems, many of which cannot be seen in direct images.

Telescopes on the ground with apertures that range from 8 to 10 m in diameter are beginning to flourish. Working with modern, high-resolution spectrographs and sensitive electronic imaging devices (see SPECTROGRAPHS: HIGH-RESOLUTION SPECTROGRAPHS, CHARGE-COUPLED DEVICE), these new facilities will bring forth an abundance of superior observations of absorption lines that appear in quasar spectra and will tell us much about the evolution of gas systems through cosmic time.

Bibliography

A very simplified development of the ideas presented here, with an accounting of the principal results from the Copernicus satellite:

Spitzer L Jr 1982 *Searching Between the Stars* (New Haven, CT: Yale University Press)

A review of recent conclusions about atomic abundances in our Galaxy from spectroscopic observations by the Hubble Space Telescope:

Savage B D and Sembach K R 1996 Interstellar abundances from absorption line observations with the Hubble Space Telescope *Ann. Rev. Astron. Astrophys.* **34** 279–329

A review of diffuse interstellar bands:

Herbig G H 1995 The diffuse interstellar bands *Ann. Rev. Astron. Astrophys.* **33** 19–73

Edward Jenkins

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interstellar Chemistry

Molecules are not limited to the solar system but appear to exist throughout the universe. Interstellar chemistry is the study of molecular matter in interstellar space. It seeks to determine what the molecules can tell us about the physical conditions in the sources where they are detected and how the molecular matter is produced.

More than 25 yr of observational studies have shown that a rich variety of gas-phase molecules is associated with INTERSTELLAR MOLECULAR CLOUDS, which are vast regions of gas and tiny dust particles found among the stars in our Galaxy and others. A considerable number of molecules are also found in the extended atmospheres of certain classes of older stars. Molecules are detected in astronomical environments by comparing spectroscopic lines, in absorption or in emission, with known spectra in the laboratory. From the strengths and shapes of spectral lines, astronomers can deduce the concentrations of the molecules as well as the heterogeneous physical conditions prevailing in the sources. The existence of a solid interstellar phase consisting of dust particles of size ≈ 0.1 μm covered with a mantle of assorted molecular ices is inferred from the wavelength dependence of the scattering and absorption of visible light as well as from broad spectroscopic features. It is thought that interstellar molecules are formed locally, i.e. where they are detected, by chemical processes that can occur efficiently under physical conditions—very low temperatures and very low densities—quite different from those on Earth. Although our understanding is far from complete, the current state of knowledge of these chemical processes is sufficient to calculate correctly the abundances of many well-known interstellar molecules and to make accurate predictions for the abundances of others still unseen.

Interstellar molecules

The first gas-phase interstellar molecules—CH, CH^+, and CN—were discovered in the years 1937–1941 using the technique of optical absorption spectroscopy, a method which is only possible for so-called 'diffuse' interstellar clouds, which are sufficiently rarefied that stellar radiation can penetrate them. A far greater extent of molecular development is found in dense interstellar clouds, which have sufficient matter to extinguish the radiation of background starlight. Although the gas phase of such clouds contains approximately 99% of the matter, it is the dust particles that are responsible for most of the extinction. Dense clouds, which are sites of newly formed stars, are typically found in 'giant' complexes of material with higher-density objects (known as cores) embedded in lower-density material. The cores can contain protostars, which consist of warming, infrared-emitting, collapsing material in the act of forming new stars. Figure 1 shows a schematic view of a dense cloud complex.

Dense clouds have been studied primarily via RADIOASTRONOMY and, to an increasing extent, via INFRARED ASTRONOMY. Both techniques can be employed since wavelengths longer than visible wavelengths are extinguished to a diminished degree. In the gas phase, individual molecules undergo rotational and vibrational motions; the rotational–vibrational energies are quantized according to the laws of quantum mechanics. Transitions between rotational levels lead to spectra from the radio through the submillimeter-wave portion of the electromagnetic spectrum while transitions involving vibrational levels lead to spectra in the infrared region. Astronomers sometimes use the word 'radio' to apply to all wavelengths larger than infrared.

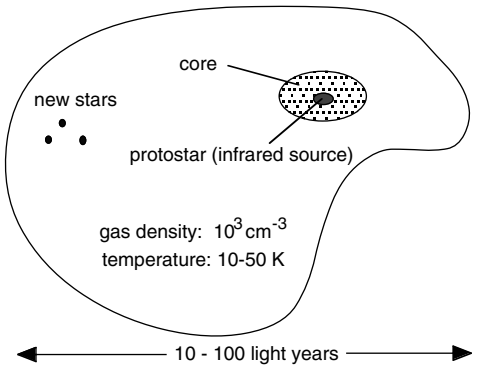


Figure 1. Schematic view of a dense interstellar cloud complex with some associated phenomena.

The first molecule detected by radioastronomical means was the diatomic species OH, which was found in 1963. The first polyatomic molecules—ammonia (NH_3) and water (H_2O)—were discovered in 1968 and the first organic (carbon-containing) polyatomic molecule—formaldehyde (H_2CO)—was detected soon thereafter. Since then, new molecules have been detected regularly every year, and there currently are known to exist over 100 gas-phase molecules consisting of up to 13 constituent atoms. These molecules are listed in table 1. (Some molecules in this table, classified as circumstellar, have been detected only in extended stellar atmospheres or both in these sources and in interstellar clouds.) In addition to these species, molecules containing isotopes such as deuterium, ^{13}C, and ^{18}O have been detected. In the last decade, infrared techniques have become increasingly powerful despite severe atmospheric interference, and a significant fraction of the molecules in table 1 have been detected in the infrared. Infrared detection is especially useful for non-polar molecules since such species do not possess strong rotational spectra. Although most 'radio' detections are observed over wide regions of clouds in emission, infrared detections are seen in absorption against background sources of radiation or against protostars forming in the clouds themselves. Much recent work in the infrared has been accomplished with the European infrared satellite called ISO (see INFRARED SPACE MISSIONS (INCLUDING ISO)).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The molecules listed in table 1 have all been detected under high spectral resolution. This fact, coupled with the detection of at least several lines for each species, makes the identification unambiguous except for a few species such as the simplest amino acid glycine (NH_2CH_2COOH), which are followed by question marks in table 1. The certainty of the assignments is important since the nature of the gas-phase molecules seen in interstellar clouds is very different from what chemists are used to on Earth. A glance at table 1 shows that most of the large molecules are organic in nature; this observation mirrors the relative numbers of inorganic and organic molecules known on Earth. On the other hand, the carbon-containing molecules in astronomical sources are often very unsaturated (hydrogen-poor), unlike naturally occurring species on Earth. Moreover, a large fraction of the interstellar molecules are not stable in the terrestrial laboratory because they are quite reactive. The unstable molecules include molecular ions (all but one positively charged), fragments known as 'radicals,' and metastable isomers, which are unusual geometrical arrangements of atoms. For example, HNC is an isomer of HCN, and c-C_3H is a cyclic isomer of linear C_3H, itself a rather reactive radical. It is because of the ingenuity of molecular spectroscopists, who have managed to take the spectra of many unusual species in the laboratory, that this witches' brew of interstellar molecules has been characterized. Although most of the molecules shown in table 1 have been found in clouds in our own Galaxy only, an increasing number have been found in external galaxies. In fact, simple molecules have been found in sources as distant as quasars and radiogalaxies with REDSHIFT $z > 4$.

Molecular spectra can be used for purposes other than the identification of individual molecules. From the relative intensities of the spectral peaks, relative concentrations, or 'abundances', can be determined. The abundances reflect both the relative amounts of elements available for incorporation into molecules and the chemical processes occurring in interstellar clouds. From studies of diffuse clouds, where the gas phase is mainly atomic, it is possible to determine elemental abundances that pertain to the gaseous interstellar medium. In general, these show that, compared with stellar atmospheres, heavy elements are partially depleted from the gas onto the dust grains. The depletion may or may not be greater in dense clouds. The abundances of heavy elements are also affected by the metallicities of galaxies. In the gas phase of a 'typical' interstellar cloud in our Galaxy, hydrogen is the dominant element, while the elements carbon, nitrogen and oxygen are lower by a factor of $\approx 10^{-4}$, with oxygen more abundant than carbon ('oxygen-rich' conditions). 'Carbon-rich' conditions, on the other hand, exist in certain stellar atmospheres.

Not surprisingly, the dominant molecule in dense interstellar clouds is molecular hydrogen, H_2. Whereas in diffuse clouds there is also a significant amount of hydrogen in atomic form, in dense clouds the hydrogen is almost completely molecular. One exception to this statement occurs in ionized regions around newly formed high-mass stars, where the hydrogen is mainly in its ionized atomic form. The second most abundant species in the gas—carbon monoxide (CO)—has a fractional abundance with respect to H_2 of $\approx 10^{-4}$ by number for typical dense clouds in our Galaxy, implying that this molecule consumes much of the available elemental carbon and oxygen. The fractional abundances of all polyatomic organic molecules with respect to H_2 are rather small; for example, that of methanol ranges from $\approx 10^{-7}$ to 10^{-10}. Despite their low fractional abundances, the total amounts of organic molecules in interstellar clouds are greater than corresponding amounts on the Earth. The molecular abundances in interstellar clouds do not reflect any equilibrium but are determined by rate processes. The positively charged species suggest an ionization source, and this is universally acknowledged to be COSMIC RAYS, which are chiefly high-energy protons. Such particles can ionize material directly or indirectly through the generation of ionizing photons.

From the point of view of many astronomers, the major use of molecular spectra from interstellar sources is the determination of the physical conditions and their heterogeneities in the sources. The rotational temperature of a molecule can be found from the relative intensities of differing spectral lines, and this temperature is related to the kinetic temperature of the source. The overall density of a source can be found from an analysis of the minimum density necessary to collisionally excite molecules which subsequently radiate; weakly polar molecules such as CO are easily excited and so appear at all densities above a low value whereas strongly polar molecules such as CS and HCN can only be detected in high-density cores where the collisional excitation is more rapid than the rapid rotational emission. In this manner, the low-density portions of dense cloud complexes have been determined to have gas densities $\approx 10^3$ cm^{-3} while the densities in the cloud cores range from $\approx 10^4$ to 10^8 cm^{-3}. Temperatures in quiescent cores, which show no dramatic signs of star formation, can be as low as 10 K, while temperatures in the more diffuse regions can rise to 50 K.

In addition to the narrow spectra, there also exist broad spectra from interstellar sources, the interpretation of which can be ambiguous. One such class of spectra derives from ices constituting INTERSTELLAR GRAIN mantles. The ices form in dense clouds as molecules condense onto the cold dust particles. Condensed-phase species tend not to show rotational spectra in the radio region because rotation is hindered by intermolecular forces. Condensed-phase species do undergo the same vibrational motions as gas-phase species. The vibrational spectra of molecules in the solid state are both broadened and shifted compared with the gas-phase values. The broadening and the shifting are sensitive to the local environment, so that these quantities can be used to characterize the nature of interstellar grain mantles. The most abundant molecule in the ice is water, which has a fractional abundance relative to gas-phase hydrogen of $\approx 10^{-4}$ by number. This high

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Gas-phase interstellar and circumstellar molecules.

2 atoms				
H_2	CH	CH^+	NH	OH
HF	HCl	C_2	CN	CO
CSi	CP	CS	NO	NS
OS	NaCl	KCl	AlF	AlCl
PN	SiN	SiO	SiS	CO^+
SO^+				
3 atoms				
H_3^+	CH_2	NH_2	H_2O	H_2S
C_2H	HCN	HNC	HCO	HCO^+
HOC^+	HN_2^+	HNO	HCS^+	C_3
C_2O	C_2S	SiC_2	CO_2	N_2O
SO_2	OCS	MgCN	MgNC	NaCN
4 atoms				
NH_3	H_3O^+	C_2H_2	H_2CO	H_2CN
$HCNH^+$	H_2CS	C_3H	c-C_3H	HCCN
HNCO	$HOCO^+$	HNCS	C_3N	C_3O
C_3S				
5 atoms				
CH_4	SiH_4	H_2C_3	c-C_3H_2	CH_2NH
H_2COH^+	CH_2CN	NH_2CN	CH_2CO	HCOOH
C_4H	HC_3N	HCCNC	HNCCC	C_5
C_4Si				
6 atoms				
C_2H_4	CH_3OH	CH_3SH	CH_3CN	CH_3NC
NH_2CHO	H_2C_4	HC_3NH^+	HC_2CHO	C_5H
C_5N				
7 atoms				
CH_3NH_2	CH_3CCH	CH_3CHO	c-CH_2OCH_2	CH_2CHCN
C_6H	HC_5N	C_7^- (?)		
8 atoms				
$HCOOCH_3$	CH_3COOH	CH_3C_3N	H_2C_6 9 atoms	
	C_7H			
C_2H_5OH	CH_3OCH_3	C_2H_5CN	CH_3C_4H	C_8H
HC_7N				
10 atoms				
CH_3COCH_3	CH_3C_5N(?)	NH_2CH_2COOH(?)		
11 atoms				
HC_9N				
13 atoms				
$HC_{11}N$				

abundance far exceeds what is found in the gas. Carbon dioxide (CO_2) is another well-characterized abundant surface species. Other molecules include CO, which is far less abundant than in the gas, as well as methane (CH_4), methanol (CH_3OH), ammonia (NH_3) and H_2S. The shape of the CO spectral line has been used to show that this species is found in two types of mantle environments—a polar and a non-polar one. Broad spectra also arise from a population of grain cores that consists of silicates, much like beach sand. A population of carbonaceous grains is inferred from a broad bump in the extinction of star light through diffuse clouds at 2200 Å, although whether this bump is due to graphite or amorphous carbon particles remains to be determined. Finally, emission from a planetary-nebula-type source known as the 'Red Rectangle' has been used to infer the existence of much smaller dust particles, known as nanoparticles, which consist of either carbonaceous or silicon-based material.

Not all broad spectral features originate from dust particles. A series of previously unidentified emission features throughout the infrared, known as the 'UIBs' or 'UIR' bands (standing for unidentified infrared bands), most probably arises from a collection of large molecules ($\geq$30 atoms in size) known as polycyclic aromatic hydrocarbons (PAHs), which are found terrestrially in automobile exhaust. Although no specific interstellar features can be identified with particular PAH species, it is likely that the broad interstellar spectra arise from a collection of PAHs with slightly differing vibrational frequencies and an overall fractional abundance with respect to H_2 of $\approx 10^{-7}$. An alternative interpretation in terms of small amorphous carbon particles exists.

Another series of broad features is seen in the visible, typically against the light of background stars, and presumably arises from diffuse interstellar clouds. These so-called 'diffuse interstellar bands' (DIBs; the word 'diffuse' is used to mean 'broad') have been known for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

much of the century and have remained unassigned except for a few tentative possibilities. Following Sherlock Holmes' dictum that if one rejects the impossible, the improbable must be truth, investigators have considered the possibility that large gas-phase molecules can account for the lines and their large widths. The astronomical problem with this hypothesis is that small molecules are attacked by radiation, which is one reason that polyatomic molecules are not detected in diffuse clouds. It is possible, however, that very large molecules are relatively immune to photodestruction since they can store excess energy much like a heat bath. The spectroscopic problem with the hypothesis is that the molecules must heretofore have eluded laboratory detection. Most recently, five DIBs have been assigned to the negatively charged molecular ion C_7^-, which has been studied in detail in the laboratory within the last year. Another two lines may be due to the gaseous 'fullerene' ion C_{60}^+, although in the laboratory this species has only been studied in a solid matrix phase in which the vibrational spectral frequencies are displaced from the gas phase values. The remaining DIB problem is a large one; its eventual solution may yield surprising inferences about the state of interstellar matter.

Low-temperature chemistry

Although some molecules are formed in extended stellar atmospheres and blown out into space, the radiation field present in the unshielded interstellar medium is sufficient to destroy them long before they are incorporated into a dense molecular cloud along with other stellar detritus by the force of gravity. Large molecules such as PAHs, assumed to be formed in carbon-rich stellar atmospheres, and dust particle cores, observed to be formed in both oxygen-rich and carbon-rich stellar atmospheres, probably withstand photodestruction to a considerable extent. Thus, as interstellar clouds form, they are composed of a gas phase consisting mainly of atoms and a solid phase consisting of dust particles without ices. To produce the interstellar molecules of table 1 from atomic precursors requires an active local chemistry, which can occur both in the gas phase and on the surfaces of the dust particles.

The chemistry is severely constrained by the low density and temperature of the cloud, however dense it may be. The low density requires the gas-phase chemistry to consist of two-body collisions, since three-body collisions, by which is meant the collision of a third body with the temporary molecule formed by the collision of two other particles, are improbable. Temporary gas-phase molecules last for periods far less than 1 s while collisions in a gas of density $\approx 10^4$ cm^{-3} occur every $\approx 10^6$ s, or 1 week. Unlike the situation in the gas phase, dust particles can act as third bodies for reaction because atoms and molecules can stick to them. Consider a gas-phase species that lands on a dust particle and sticks to it with a certain efficiency. A second species then lands and sticks on the same grain, either atop the first species (Eley–Rideal mechanism) or elsewhere, in which case the two species must diffuse on the grain surface to find one another (Langmuir–Hinshelwood mechanism).

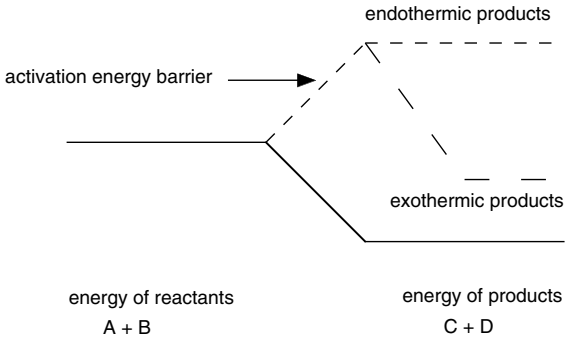


Figure 2. The energy profiles of three types of chemical reactions: endothermic, exothermic with activation energy and exothermic without activation energy.

The low temperatures remove from consideration both gas-phase and grain-surface reactions that require significant amounts of energy. These processes include all 'endothermic' reactions, in which the products lie at higher energy than the reactants, and those 'exothermic' reactions that require energy to overcome a potential barrier (the so-called 'activation' energy) occurring during the course of reaction even though the energy of products lies lower than that of reactants. See figure 2.

The rates of gas-phase chemical reactions are normally expressed in terms of thermal rate coefficients $k(T)$ and concentrations. For example, the rate of the reaction

$$\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C} + \mathrm{D} \tag{1}$$

is written as

$$\mathrm{d}n(\mathrm{A})/\mathrm{d}t = -\mathrm{d}n(\mathrm{C})/\mathrm{d}t = -k(T)n(\mathrm{A})n(\mathrm{B}) \tag{2}$$

where $n(I)$ stands for the concentration of species I. The thermal rate coefficient is related to the reaction cross section σ by

$$k(T) = \langle v\sigma \rangle \tag{3}$$

where v is the relative collision velocity between the reactants and the symbols $\langle \cdots \rangle$ stand for a thermal average over all degrees of freedom. In the limit of large numbers of reactive species on a grain surface, a rate equation treatment is appropriate for the diffusive (Langmuir–Hinshelwood) mechanism although the rate coefficients must incorporate a random walk or quantum mechanical tunneling motion rather than a simple relative velocity. In the limit of small numbers of reactive species on a grain surface, the rate equation method is problematic.

For endothermic reactions or for exothermic reactions with potential barriers, it is well known that

$$k(T) = A(T)\exp(-E_a/T) \tag{4}$$

where $A(T)$, the pre-exponential factor, can have a weak direct or inverse temperature dependence depending, for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

gas-phase processes, mainly on the nature of the long-range potential and E_a, the activation energy, is written here in units of temperature. For most chemical reactions involving neutral species $E_a \geq 10\,000$ K so that, at typical interstellar cloud temperatures under 100 K, the rate of reaction is negligible. For reactions with smaller activation energies and light reactants, tunneling processes, not included in the above expression, can increase the rate coefficient at low temperatures, especially for processes on grain surfaces. Even so, interstellar chemistry is effectively limited to reactions with zero or small activation energy. In the gas phase, these include the following classes of exothermic processes: (i) reactions involving charged and neutral species, known as ion–molecule reactions, and (ii) reactions between neutral species in which at least one of the reactants is an atom or a 'radical', which is a particularly reactive molecule with an odd number of electrons. On the surfaces of grains, reactions of the second type must dominate since positive ions striking grains are thought merely to be neutralized by the excess negative charge present.

Ion–molecule reactions have been studied extensively in the laboratory down to low temperatures; they are known to occur with temperature-independent rate coefficients $k(T) \approx 10^{-9}$ cm^3 s^{-1} for non-polar neutrals and, for polar neutrals, with somewhat larger rate coefficients that increase weakly with decreasing temperatures. Gas-phase neutral–neutral reactions involving a radical are somewhat slower at room temperature, with a weak temperature dependence that is not fully understood. Very few studies of these systems at low temperatures have been undertaken. The products of all classes of reactions cannot be easily guessed at; rather, they must be measured or calculated via detailed quantum mechanical methods. For diffusive surface reactions, very few reactive systems have been studied in any detail, and estimates are often used for the rates of reaction.

The formation of interstellar molecules

Starting from atomic matter, it is difficult to form diatomic molecules in a low-density gas. Consider the collision of two hydrogen atoms to form a temporary hydrogen molecule, known as a 'complex'. This complex typically flies apart long before a third body approaches. There is a small chance that the complex can stabilize itself via the emission of a photon in a process known as 'radiative association', but this chance has been calculated to be remote for the association of two hydrogen atoms. Indeed, radiative association becomes efficient only for larger species. Although there do exist more efficient gas-phase syntheses of H_2 involving H^+ or H^-, it is now generally thought that interstellar H_2 is formed by diffusive recombination on dust grains and released back into the gas via both non-thermal and thermal processes, as depicted in figure 3. The recombination has been studied in the laboratory.

Once H_2 is formed, a complex gas-phase chemistry ensues in dense interstellar clouds; this chemistry

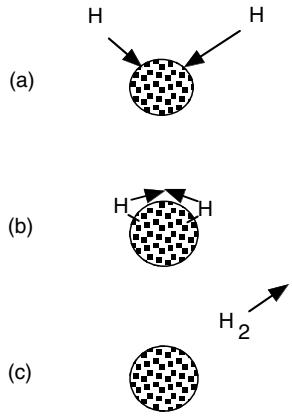


Figure 3. The formation of interstellar H_2 by a diffusive (Langmuir–Hinshelwood) process on an interstellar grain.

is dominated by positive ion–molecule reactions, but contains radical–neutral reactions as well. The ionization is typically provided by cosmic rays, which can penetrate through a total column density (concentration × path length) of more than 10^{25} cm^{-2}. The flux of cosmic rays, which are presumably formed in supernovae, is probably dependent on position within a galaxy and varies from one galaxy to another. Attempts to estimate this flux based on molecular observations are uncertain to an order of magnitude. The ionization provided is not great in a quantitative sense; both theory and observations of ions indicate that the fractional ionization is perhaps 10^{-7} with the largest uncertainty being the unmeasured metallic ion concentration. The small ionization is sufficient, however, to drive the ion–molecule chemistry. In diffuse clouds, the strong interstellar radiation field penetrates and provides a more potent source of ionization. The UV radiation is sufficiently energetic to ionize carbon, which possesses a rather low ionization potential, into C^+, leading to a factional ionization of $\approx 10^{-4}$. Even in diffuse clouds, however, ionization of hydrogen and helium requires cosmic rays.

Let us see how the initial stages of molecular synthesis work by considering the formation of water and simple hydrocarbons (molecules containing only carbon and hydrogen).

When a hydrogen molecule collides with a cosmic ray proton (CRP), the major products are the molecular ion and an electron:

$$H_2 + \mathrm{CRP} \longrightarrow H_2^+ + e^- + \mathrm{CRP}. \qquad (5)$$

The newly formed H_2^+ 'quickly' recombines with a neutral hydrogen molecule to form the simple polytomic ion H_3^+:

$$H_2^+ + H_2 \longrightarrow H_3^+ + H. \qquad (6)$$

This ion can react with different atoms present in the initial gas. The reaction with oxygen atoms is followed by a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sequence of reactions with ubiquitous H_2 leading to the rather stable protonated water ion:

$$O + H_3^+ \longrightarrow OH^+ + H_2 \quad (7)$$

$$OH^+ + H_2 \longrightarrow H_2O^+ + H \quad (8)$$

$$H_2O^+ + H_2 \longrightarrow H_3O^+ + H. \quad (9)$$

Not reactive with H_2, H_3O^+ eventually recombines with free electrons; recombination reactions are known to be dissociative in nature but only recently have experiments been able to determine the branching fractions among the various possible products. For the reaction between H_3O^+ and electrons, both the radical OH and the normal molecule water (H_2O) are products.

A similar sequence of reactions starting with atomic carbon leads to the methyl (CH_3^+) ion; this ion cannot react with H_2 in the normal manner but does undergo a rather slow but competitive radiative association reaction with hydrogen,

$$CH_3^+ + H_2 \longrightarrow CH_5^+ + h\nu \quad (10)$$

to lead to CH_5^+, which is a precursor of methane (CH_4) and methyl (CH_3).

Once simple molecules are synthesized, they are depleted via reactions with atomic and molecular ions and, in the case of radicals, neutrals. Photodestruction also plays a significant role, especially in diffuse clouds. Some of the depletion reactions lead to more complex species; others do not. Sequences of reactions leading to the production and depletion of most of the species in table 1 have been written down, although laboratory evidence for many of these processes is lacking. Occasionally, neutral–neutral reactions play important roles in the synthesis of molecules. Two examples, both studied in the laboratory at low temperature, are

$$C + C_2H_2 \longrightarrow C_3H + H \quad (11)$$

$$CN + C_2H_2 \longrightarrow HC_3N + H \quad (12)$$

these reactions produce the interstellar molecules C_3H and cyanoacetylene (HC_3N) more rapidly than do ion–molecule and dissociative recombination pathways.

Clearly, gas-phase chemistry does not represent the complete story. If molecular hydrogen is formed on grain surfaces, so must other species be. The production of molecules other than H_2 on grain surfaces is not well understood. It does appear that the dominant process is hydrogenation via hydrogen atoms which, although much less abundant than H_2 molecules in the dense cloud gas, are more reactive on surfaces and are able to hydrogenate species into rather saturated (hydrogen-rich) forms. The hydrogenation of atoms such as C, N and O into the saturated forms CH_4, NH_3 and H_2O via successive 'three-body' reactions with H is non-controversial, but the hydrogenation of molecular species such as CO into more hydrogenated molecules—H_2CO (formaldehyde) and CH_3OH (methanol)—is indeed controversial since small activation energies of uncertain amounts are involved. Although large amounts of hydrogenated species can probably be produced on grains, and are detected via infrared absorption studies, it is still unclear how some of these molecules get back into the gas phase since evaporation is too slow for all but the lightest species. Because the time scale for sticking to a dust grain is short compared with cloud lifetimes and because cold interstellar clouds are not devoid of gaseous material, there must exist non-thermal 'desorption' mechanisms such as heat generation by cosmic ray bombardment and by exothermic chemical reactions occurring on the grain surfaces.

Models of dense interstellar clouds

Chemical models of interstellar clouds date from the early 1970s. Such models are used to calculate concentrations of molecules as a function of time. To produce a time-dependent model, one integrates coupled differential rate equations, of the type shown in equation (2) but with both formation and depletion processes, for all species in the model. The initial concentrations and the elemental abundances to be used are constraints. The approach is only approximate for grain chemistry. Early models were steady state in nature, used only gas-phase reactions and did not involve the solution of simultaneous differential equations; rather, the concentration derivatives were set to zero so that the coupled equations could become algebraic in nature. Models of this type are able to explain the observed abundances of many of the smaller molecules.

After a decade of work and advances in computer technology, gas-phase pseudo-time-dependent models became available. In these models, coupled rate equations are solved but the physical conditions (temperature, density) remain fixed and uniform. The initial abundances are mainly atomic except for H_2. Such homogeneous models still represent the dominant approach today; these now contain thousands of gas-phase reactions connecting hundreds of molecular species. For the best-studied dense molecular cloud core, a quiescent core with gas density 10^4 cm^{-3} known as the Taurus Molecular Cloud 1 (TMC-1), pseudo-time-dependent models can produce order-of-magnitude agreement with observed abundances for up to 80% of the more than 50 detected molecules in the source. The time of best agreement depends on the gas-phase elemental abundances chosen, most specifically on the gas-phase carbon-to-oxygen abundance ratio. For oxygen-rich gas, the most complex of detected molecules are predicted to be abundant only at pre-steady-state times of $\approx 10^5$ yr, after which they tend to 'combust' into CO, much like the incomplete combustion that occurs in an old furnace.

Since gas-phase models ignore the contribution of grain chemistry to gas-phase abundances (other than to the formation of H_2), it is tempting to suggest that the synthesis of molecules with abundances not reproduced by such models can be ascribed to grain chemistry. Chemical models with both gas-phase and grain-surface reactions have been reported, but these

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are still beset with uncertainties, including the correct mathematical approach and the role of surface chemistry in the production of species returned to the gas. Besides the absence of grain chemistry, other possible sources of error in gas-phase pseudo-time-dependent models include (i) uncertainty in the rates and products of gas-phase reactions, especially neutral–neutral processes and (ii) the assumption of homogeneous and time-independent physical conditions. More advanced gas-phase and gas-grain models, incorporating heterogeneous and changing physical conditions, have been reported. One particular approach is to utilize a number of slabs of different densities and temperatures; as one goes from slabs representing outer layers to slabs representing inner portions of the clouds, the densities become higher and the temperatures lower. Reality, unfortunately, is still more complex than this. There is good evidence that dense clouds have time-dependent self-similar structures on many distance scales down to clumps of stellar masses.

Recently, an interesting feature of gas-phase models, known as 'bistability', has been found. For selected choices of the parameters (density, temperature, elemental abundances), there exist not one but at least two solutions to the rate equations, labelled the 'low-ionization' and 'high-ionization' phases. Whether this bistability is an artifact of the models or whether it actually occurs in interstellar clouds remains to be determined.

Given the physical complexity of dense cloud complexes, it is important to note the major successes of simple gas-phase chemical models of cloud cores: (i) they predict the abundances of the major species (e.g. CO) and most minor species correctly, (ii) they account for the witches' brew (radicals, ions, metastable isomers) of unusual molecules present and the fact that the chemistry is mainly unsaturated (hydrogen-poor) despite the large abundance of H_2 and (iii) they account for very strong isotopic fractionation, particularly regarding deuterium, in which singly deuterated isotopomers (e.g. DCN) can have abundances approaching 10% of the normal species (HCN) despite the fact that the D/H elemental abundance ratio is only 10^{-5}.

The fractionation of deuterium is a thermodynamic effect made possible by the low temperatures. Consider the reaction of the ion H_3^+ with HD, the major repository of deuterium:

$$H_3^+ + HD \longrightarrow H_2D^+ + H_2. \qquad (13)$$

The reaction is exothermic by approximately 200 K owing partially to differences in the so-called zero-point energies of vibration between products and reactants. The reverse reaction, which is endothermic by this amount, cannot occur rapidly (see equation (4)) at temperatures significantly under 200 K because of the Boltzmann factor in the expression for the rate coefficient. If we neglect other reactions, equate the rates of the forward and backward reactions (as is appropriate at steady state) and make the assumption that the pre-exponential factors are the same, we obtain the following expression:

$$\frac{[H_2D^+]}{[H_3^+]} \approx \frac{[HD]}{[H_2]} \exp(200/T) \qquad (14)$$

where T is the temperature and the symbols $[\cdots]$ refer to concentration. Although not quantitative owing to the exclusion of side reactions and the assumption concerning the pre-exponential factors, the relation shows how the ratio $[H_2D^+]/[H_3^+]$ can become quite large at low temperatures. The high abundance ratio between the deuterated isotopomer and the normal species for H_3^+ and other molecular ions (e.g. HCO^+) then leads to the production of large abundances of deuterated neutral species through ion–molecule reaction chains. Isotopic fractionation also occurs on grain surfaces when high abundances of atomic D, generated from the dissociative recombination reactions of deuterated ions such as H_2D^+ and DCO^+ with electrons, adsorb onto the dust.

Current and future prospects

Much current work in interstellar chemistry pertains to distinctive regions surrounding sites of STAR FORMATION in dense interstellar clouds. Star formation is a complex process which gives rise to many physical manifestations in the surrounding material. The unusual physical conditions in turn give rise to chemical manifestations which are quite distinct from the ambient interstellar medium.

Sites of high-mass star formation are associated with shock waves and more quiescent dense regions called hot cores. The physics of shock waves is rather complex and is highly dependent on the size and direction of any magnetic field. In the limit of low magnetic field, there is a well-defined shock front which induces a rapid jump in temperature, depending on the square of the shock velocity. The temperature can rise to thousands of kelvins before subsequent cooling, so that a high-temperature chemistry can occur with endothermic reactions plus those exothermic reactions that contain activation energy. At sufficiently high shock velocities, molecular dissociation is manifest. Sputtering also occurs off dust grains, a process which is probably the only one to place refractory elements such as silicon into the gas phase. With a sufficiently large magnetic field, the temperature of the shock increases continuously in a more modest manner, and ions stream relative to the neutral species, powering otherwise slow ion–molecule reactions. In addition to their association with high-mass star formation, shocks are also found in the diffuse interstellar medium. Modelers have attempted to understand the shock chemistry in both diffuse and dense media, and have focused on those molecules the abundances of which constitute 'shock signatures'. In the diffuse medium, the molecule CH^+ has long been associated with shocks, where it can be formed by the strongly endothermic reaction

$$C^+ + H_2 \longrightarrow CH^+ + H. \qquad (15)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Still, after much work, the importance of shock chemistry remains unclear. In diffuse regions, more attention is now being given to interstellar intermittent turbulence as a source for high-temperature chemistry.

Perhaps more successful have been chemical models of hot cores, which are quiescent regions of high density ($\approx 10^{6-8}$ cm^{-3}) and elevated temperatures ($\approx$100–300 K) in the vicinity of newly forming high-mass stars. In chemical models of hot cores, it is assumed that the rising temperatures drive material off grain mantles. Based on current ideas of grain chemistry, the desorbed molecules are thought to be mainly saturated (hydrogen rich) in nature, in good agreement with observations that show hot cores to have high abundances of simple species such as water and ammonia as well as more complex species such as the alcohols methanol (CH_3OH) and ethanol (C_2H_5OH), dimethyl ether (CH_3OCH_3) and methyl formate ($HCOOCH_3$). The original desorbed species undergo a gas-phase chemistry which takes $\approx 10^{4-5}$ yr to change the chemical composition and allows investigators to determine the ages of particular hot cores.

Sites of low-mass star formation are associated with protostellar disks, the chemistry of which is a current topic of considerable interest since radioastronomers are beginning to use interferometric methods to resolve such disks spatially and since the chemistry of these objects can be related to the chemistry of comets and planets in the initial stages of formation. These disks have higher densities than do ambient interstellar clouds but, unlike high-mass star formation regions, are characterized by low temperatures ($\leq$50 K) to within 10 AU (1 AU = mean distance between the Sun and the Earth) of the protostar. The high densities and low temperatures enhance the roles of grains and grain surface chemistry. Still, ion–molecule reactions are now known to be crucial because cosmic rays can penetrate through the plane of the disk at distances greater than a few AU from the protostar. Consequently, the chemistry bears some relation to that of the normal interstellar medium. Complications include the existence of an additional mode for ionization—radioactive decay—and the need to include three-body reactions for that portion of the disk closer than $\approx$10 AU to the newly formed star because of the especially high densities in the inner regions.

Although most observational and theoretical work on interstellar molecules has concerned our own Milky Way Galaxy, increasing emphasis in future work will shift to extragalactic studies. One can expect the chemistry occurring in different galaxies to have subtle and not-so-subtle differences from the chemistry in the Milky Way owing to factors such as varying ionization rates and metallicities as well as differing rates of star formation. Another field with potential for the future is the laboratory detection of unusual carbon-containing molecules that are likely candidates for interstellar detection. Several laboratories are currently working on both microwave and visible detection of such species.

In addition to increasing emphasis on the role of molecules as probes of diverse physical conditions and the synthesis of new and unusual molecules likely to be present in the interstellar gas, much research—both experimental and theoretical—will continue to be concerned with improving our understanding of the basic chemistry. For example, new experiments are being undertaken to simulate low-temperature grain chemistry, currently a large uncertainty in our understanding of the interstellar medium, and to simulate how large molecules such as PAHs can form in the gas phase. New theoretical techniques are being used to probe the chemistry involved in the formation of dust particles, and new approaches are being tried out to incorporate grain chemistry into chemical models. These and other efforts testify to a vigorous area of current research. Interstellar chemistry has been a major concern in astronomy for only 25–30 yr, but the role of molecules as probes of the cosmos and of our place in it will continue to intrigue many investigators in the future.

Bibliography

Duley W W and Williams D A 1984 *Interstellar Chemistry* (London: Academic)

Faraday Division, Royal Society of Chemistry 1998 *Chemistry and physics of molecules and grains in space Faraday Discuss.* **109**

Herbst E 1995 Chemistry in the interstellar medium *Ann. Rev. Phys. Chem.* **46** 27–53

Herbst E and Klemperer W 1973 The formation and depletion of molecules in dense interstellar clouds *Astrophys. J.* **185** 505–33

van Dishoeck E F (ed) 1996 *Molecules in Astrophysics: Probes and Processes* (Dordrecht: Kluwer)

Eric Herbst

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interstellar Grains

Submicron solid particles are dispersed through interstellar gas. These *interstellar grains* absorb and scatter light, thus shielding some regions from ultraviolet radiation, but also limiting our ability to detect photons which have been emitted by astronomical objects. Dust grains re-radiate absorbed energy in the infrared, thus contributing to the overall emission spectrum of astronomical systems ranging from dusty disks around stars to ultraluminous starburst galaxies.

A naked-eye view of the sky from a dark site on a clear summer night reveals dramatic dark patches in the Milky Way. These dark regions are not due to a deficiency of stars—they are instead the result of obscuration by dust clouds interposed between the Earth and distant stars. The obscuration tends to be greater at shorter wavelengths; as a result, the light reaching us from distant, obscured stars is 'reddened'. This reddening by interstellar dust can be understood as arising from scattering and absorption by a population of interstellar submicron dust grains.

The grain population spans a range of sizes, from molecules[1] containing only tens of atoms to particles as large as ~ 0.3 μm, containing $\sim 10^{10}$ atoms. Most of the grain mass appears to be due to two types of solid, in approximately equal amounts: an amorphous silicate mineral and carbonaceous material. A number of elements—including silicon and iron—are primarily in solid form in the interstellar medium. Approximately two-thirds of the interstellar carbon in diffuse clouds is in solid form (see INTERSTELLAR ABSORPTION LINES).

It is of course important to characterize the wavelength-dependent interstellar extinction so that astronomical observations can be 'corrected' for the obscuring effects of dust. In addition, the infrared emission from dust grains provides a valuable probe of dense regions, and the dust grains themselves play important roles in interstellar chemistry (shielding from ultraviolet radiation, and catalysing the formation of H_2), interstellar gas dynamics (radiation pressure forces on dust grains, and coupling of charged dust grains to magnetic fields) and heating and cooling of interstellar gas. Dust grains are central to many problems in modern astrophysics.

Observational evidence: summary

There are many different astronomical phenomena which both reveal the existence of interstellar dust grains and provide information allowing us to infer the properties of this dust. Some of the information is quite direct, due to absorption, scattering, or emission of light by the grains:

(a) Wavelength-dependent extinction–attenuation and 'reddening' of the light from distant stars due to intervening dust (see figure 1).

(b) Spectroscopic features in the extinction. There are a number of extinction features, including: a strong

[1] Since all grains are 'molecules', it is natural to consider small molecules as the small-size end of the overall grain population.

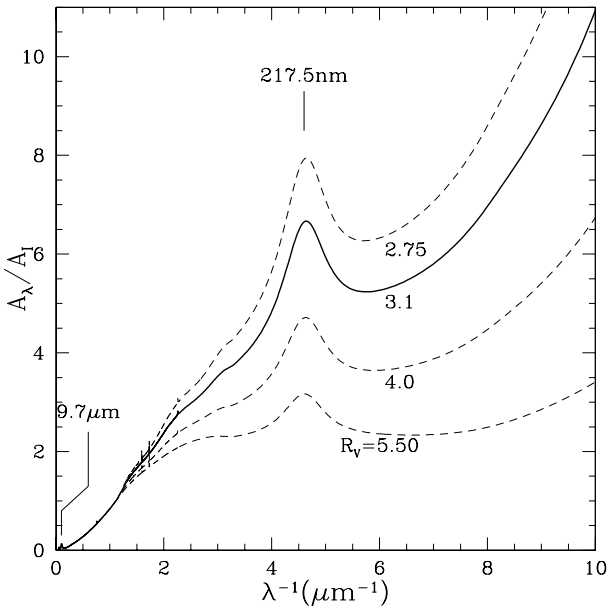


Figure 1. Wavelength-dependent extinction, normalized to the extinction at $I = 900$ nm, for different types of clouds, identified by the value of $R_V \equiv A_V/(A_B - A_V)$, where $V = 550$ nm and $B = 440$ nm. The average extinction for diffuse clouds is characterized by $R_V \approx 3.1$. Dense gas near the surfaces of molecular clouds can have R_V as large as 5.5. The extinction at I is approximately proportional to $N_{\rm H} = N({\rm H}) + 2N({\rm H_2}) + N({\rm H^+})$, with $A_I/N_{\rm H} \approx 2.6 \times 10^{-26}$ m^2/H.

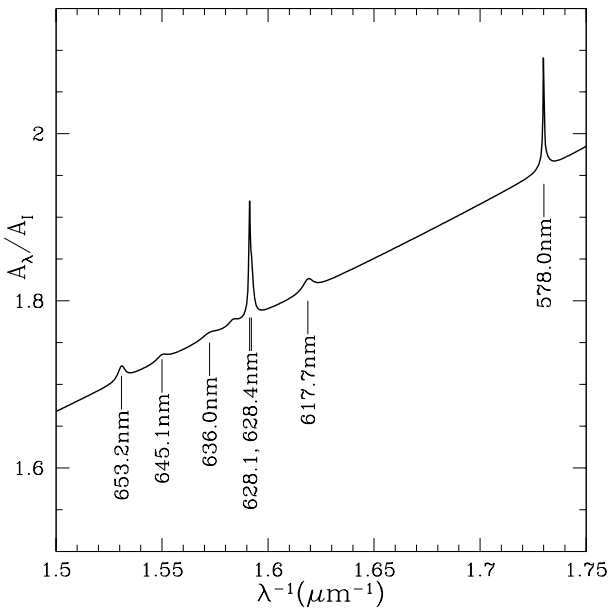


Figure 2. A portion of the extinction curve showing some of the 'diffuse interstellar band' extinction features, labelled with their respective wavelengths. These diffuse bands may be due to impurities in grains, or to 'free-flying' large molecules/ultrasmall grains.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and very broad extinction 'bump' at 217.5 nm (see figure 1), probably due to carbonaceous material, perhaps graphite; infrared extinction features at 9.7 μm and 18 μm, almost certainly due to silicates; a number of weaker 'diffuse interstellar bands' (see figure 2), the strongest of which are at 443 nm and 578 nm, and which remain generally unidentified; an absorption feature at 3.4 μm, seen in diffuse clouds, presumably due to the C-H stretching mode in aliphatic hydrocarbons; and a number of absorption features, seen only in molecular clouds, due to ice mantles which apparently coat the grains in these regions. The strongest such feature is a 3.1 μm feature attributed to H_2O ice.

(c) Polarization of starlight—preferential attenuation of one linear polarization over another by aligned interstellar dust grains, so that initially unpolarized light from a star is partially polarized by the time it reaches the Earth (see POLARIZATION).

(d) Reflection nebulae—dust clouds which are relatively close to bright stars, so that the starlight reflected by dust grains near the cloud surface renders the cloud visible (see REFLECTION NEBULAE).

(e) X-ray haloes around x-ray point sources located behind interstellar dust clouds. The haloes result from small-angle scattering of x-rays by interstellar dust grains.

(f) Infrared emission from dust grains heated by interstellar starlight (see figure 3).

(g) Infrared emission features indicative of aromatic hydrocarbons. The strongest features from the general interstellar medium are at 11.3 μm, and 7.7 μm (see figure 3).

(h) Emission features at 18 μm and 9.7 μm, generally thought to be due to silicates, from dust in regions with radiation fields $\geq 10^4$ times stronger than the average starlight background.

(i) There is strong evidence for emission from interstellar grains in the far red, presumably fluorescence following absorption of a shorter-wavelength photon.

(j) In the local interstellar medium, the dust-to-gas ratio appears to be approximately constant—the dust follows the gas.

Although limited, we also have some quite direct evidence concerning interstellar grains:

(a) Relic interstellar grains found in primitive 'carbonaceous chondrite' meteorites.

(b) Impacts with interplanetary probes of interstellar grains passing through the solar system.

In addition, there are a number of astrophysical phenomena which provide indirect information about interstellar grains:

(a) Underabundances, relative to solar abundances, of certain elements in interstellar gas—including iron, silicon, magnesium, and carbon. These elements are presumed to be underabundant in the gas phase because a large fraction of the atoms are locked up in interstellar grains.

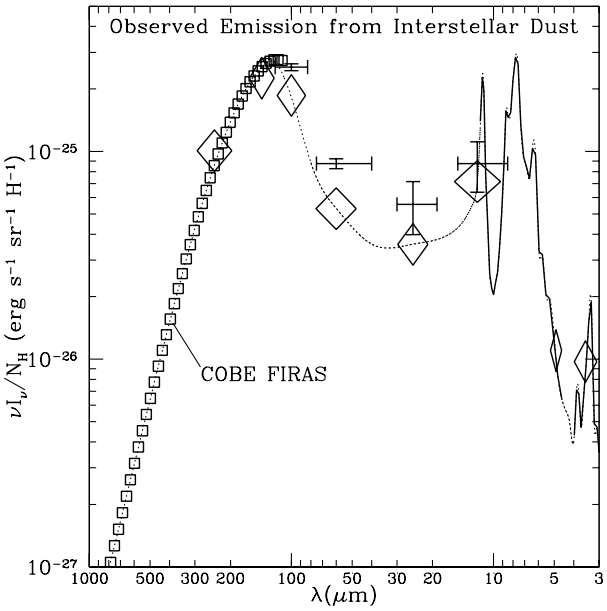


Figure 3. Infrared emission from interstellar dust grains from dust in diffuse clouds, per H nucleon. Crosses indicate data from the Infrared Astronomy Satellite (IRAS) at 100, 60, 25 and 12 μm. Squares indicate data from the FIRAS instrument on the Cosmic Background Explorer (COBE). Diamonds indicate data from the DIRBE instrument on COBE. The heavy full line at 12–5 μm and 4.6–3 μm is the spectrum measured by the Infrared Telescope in Space (IRTS).

(b) The presence in different parts of the interstellar medium of molecular hydrogen (H_2) with abundances far exceeding what could be produced by purely gas-phase processes. The required rate of H_2 catalysis on interstellar grain surfaces provides information on interstellar grain properties, including total surface area.

Extinction by interstellar dust grains

The 'extinction curves' shown in figure 1 represent the wavelength-dependent extinction A_λ at wavelength λ observed on different kinds of sightlines. Careful studies of the extinction on many lines of sight reveal that the extinction curves can be approximately described by a one-parameter family of extinction curves, and the curves shown in figure 1 are obtained from fitting functions originally developed by Cardelli *et al* (1989) and others. It is convenient to take $R_V \equiv A_V/(A_B - A_V)$ as the parameter, where $V = 550$ nm and $B = 440$ nm. Values of R_V as small as 2.75 and as large as 5.5 are observed in different regions.

Grains in diffuse regions tend to have $R_V \approx 3.1$, while larger values of R_V tend to be seen when studying the extinction by dust in dense clouds. Larger values of R_V—corresponding to 'greyer' extinction—are indicative of larger grains. The observed tendency for larger values of R_V to be found in denser regions strongly suggests that characteristic grain sizes are larger in these regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It can be shown that the increased values of R_V cannot be explained simply by accretion of atoms and molecules from the gas phase—because of the very large amount of surface area contributed by the smaller grains, accretion from the gas phase would result in only a very small increase Δa in the radii of all grains, with minimal effect on the extinction at B and V. Instead, small grains must coagulate with larger grains to convert diffuse cloud dust to dense cloud dust.

Extinction features

The strongest spectroscopic feature in the extinction curve is a conspicuous 'bump' at 217.5 nm (see figure 1). This feature has approximately the wavelength and width expected for small (radius $a \leq 15$ nm) particles of graphite. While identification of the feature is still not certain, it appears highly likely that it is due to small carbonaceous particles with sp^2 carbon–carbon bonds as in graphite and polycyclic aromatic hydrocarbons. The strength of the 217.5 nm feature requires that ~15% of the solar abundance of carbon be present in small ($a \leq 15$ nm) particles.

Between 1.32 μm and 400 nm there are many weak diffuse extinction features, known as the 'diffuse interstellar bands', or DIBs. Approximately 300 such features have been identified, with full width at half-maximum (FWHM) ranging from ~0.05 nm up to ~4 nm. The strongest DIB is at 442.9 nm.

The first DIBs were recognized in 1934 by Merrill, yet they remain essentially unidentified to this date. They must be due to either 'impurities' within dust grains, or to small free-flying molecules (i.e. ultrasmall dust grains). A few of the weaker features have been identified recently as electronic transitions of C_7^-, and it seems likely that many other of the DIBs are also produced by small hydrocarbon molecules, either neutral or charged (either positively or negatively).

A broad extinction feature at 3.4 μm is attributed to the C–H stretching mode in aliphatic (chainlike) hydrocarbons. This feature is present in the diffuse interstellar medium.

There are strong infrared extinction features peaking at 9.7 μm and 18 μm which are almost certainly due to amorphous silicates with a composition approximating that of olivine ($Mg_xFe_{2-x}SiO_4$). In regions where the dust is hot (e.g. circumstellar dust, or the dust near the TRAPEZIUM in the ORION NEBULA), these features appear in emission. In some circumstellar dust shells and disks (e.g. the dusty disk around the Herbig Ae/Be star HD 100546), sharp features characteristic of crystalline silicates appear in emission, but these features have not been detected in either emission or absorption in the interstellar medium, indicating that the bulk of interstellar silicates are amorphous.

In dark clouds, a number of additional features appear in the infrared extinction, presumably due to growth of molecular ice mantles on the refractory dust grain cores. The strongest feature is at 3.08 μm and is due to amorphous H_2O ice. Additional features have been identified as frozen CH_3OH (3.53 μm), CO (4.67 μm), CH_4 (7.65 μm) and CO_2 (15.2 μm).

Alignment of dust grains

When initially unpolarized starlight passes through the dusty interstellar medium, it acquires both linear and circular polarization. The linear polarization is due to preferential attenuation of one linear polarization over the other, due to a population of non-spherical dust grains which are somehow aligned. The linear polarization peaks near $V = 550$ nm, with peak polarization $P \approx 0.03 A_V$. The circular polarization, generally quite small, is due to conversion of linear to circular polarization when the alignment direction of the dust grains undergoes a twist along the line of sight to the source.

Spinning dust grains have magnetic moments antiparallel to their angular velocities, and the angular momentum therefore precesses around the local galactic magnetic field. As a result, the observed direction of starlight linear polarization must be either parallel or perpendicular to the projection of the magnetic field on the sky. Theories of grain alignment lead us to expect the grain angular momentum to tend to align with the magnetic field, and the 'long' axis of the grains to tend to be perpendicular to the angular momentum. The long axis of the grain therefore tends to be perpendicular to the magnetic field, and as a result light becomes linearly polarized *parallel* to the projection of the magnetic field on the plane of the sky.

Our understanding of the physics of dust grain alignment is not yet complete, but it appears that the observed alignment in diffuse clouds is produced mainly by a combination of paramagnetic dissipation (originally proposed by Davis and Greenstein in 1951) and radiative torques on irregular dust grains due to anisotropic starlight. In some cases the alignment may result from drift of the dust grains through the gas cloud.

Scattering of light by dust grains

Reflection nebulae such as NGC2023 or the beautiful filamentary structures near the Pleiades show that dust grains scatter starlight. While most conspicuous when a bright star is located near the surface of a dense clouds, reflected starlight also manifests itself as the 'diffuse galactic light'—reflected starlight seen in all directions in the sky where there is dust. Measurements of the surface brightness of reflection nebulae, or of the diffuse galactic light, provide constraints on the scattering properties of interstellar dust grains—both the total 'albedo' $\omega = C_{\rm sca}/C_{\rm ext}$, and the angular dependence of the scattering, often characterized by $g \equiv \langle \cos\theta \rangle$, where θ is the scattering angle. At visible wavelengths, the albedo $\omega \approx 0.5$ and the grains are moderately forward-scattering, $g \approx 0.5$, consistent with current models for interstellar grains which reproduce the wavelength-dependent extinction as well as the infrared emission from dust.

Scattering by dust grains can also be observed at x-ray wavelengths. A grain is essentially transparent to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$h\nu \geq 0.5$ keV x-ray photons. Since the refractive index of the grain material is very close to 1, the scattering can be calculated in the 'Rayleigh–Gans' approximation, and one finds that only small-angle scattering is expected. X-ray haloes around compact x-ray sources have been imaged by the EINSTEIN (HEAO-2) and ROSAT observatories; the observed x-ray haloes appear to be approximately consistent with the scattering expected for a grain model developed to account for the optical-UV extinction curve.

Dust grain luminescence

Observations of REFLECTION NEBULAE (e.g. NGC 7023) as well as the general interstellar diffuse clouds or 'cirrus' appear to show evidence of far-red continuum radiation in excess of what is expected from simple scattering of starlight by dust grains. The far-red emission peaks near ∼700 nm, and has a spectrum resembling the luminescence from hydrogenated amorphous carbon illuminated by $\lambda \leq$ 550 nm radiation. This suggests that some of the grain material may resemble hydrogenated amorphous carbon, although ultrasmall silicon grains have also been proposed as the source of the emission. Even if it is assumed that hydrogenated amorphous carbon is responsible for the observed emission, an accurate estimate of the quantity of material required is not possible, since luminescence efficiencies of hydrogenated amorphous carbon in the laboratory depend on the preparation of the sample.

Infrared emission from dust grains

Starlight is in part absorbed by dust grains, and the absorbed energy is re-radiated by dust grains in the infrared. The observed emission spectrum for interstellar dust is shown in figure 3. It consists of far-infrared emission peaking at $\lambda \approx 140\ \mu$m, plus substantial emission at shorter wavelengths.

The emission at $\lambda < 12\ \mu$m shows conspicuous emission features at 11.3, 8.6, 7.7, 6.2 and 3.3 μm; emission peaks at these same wavelengths have also been observed from reflection nebulae, planetary nebulae, HII regions, and circumstellar dust. The emission features have been identified as characteristic of polycyclic aromatic hydrocarbons (PAHs): C-H stretch (3.3 μm), C-C stretch (6.2 and 7.7 μm), in-plane C-H bend (8.6 μm) and out-of-plane C-H bend (11.3 μm). Variations in the relative strengths and precise wavelengths of these features from one object to another may be due to changes in the PAH mixture, including changes in the fraction which are positively or negatively charged.

The 'classical' grains with radii 0.01–0.3 μm radiate a spectrum characteristic of thermal emission at the (steady) temperature of the grain, ∼15–20 K for grains in the diffuse interstellar medium. This accounts for nearly all of the emission at $\lambda > 60\ \mu$m. For very small grains, however, absorption of a single starlight photon can appreciably change the grain temperature. For example, a single photon of energy $h\nu = 10$ eV can heat a 230 atom graphite grain to a peak temperature $T \approx 300$ K. At this temperature the grain can radiate effectively at wavelengths as short as $\lambda \approx 8\ \mu$m. The observed emission feature near 7.7 μm could therefore be due to thermal emission from grains with ∼100–300 atoms. The relatively large amount of power radiated at $\lambda < 10\ \mu$m requires that the grains with ≤300 atoms account for an appreciable fraction (≥15%) of the total absorption of starlight by interstellar grains.

Microwave emission from dust grains

Sensitive observations of the cosmic microwave background radiation have revealed 10–60 GHz emission from interstellar matter with intensities greatly exceeding what would be expected from an extrapolation of the thermal far-infrared emission to these lower frequencies. It appears likely that the observed 10–60 GHz emission is largely rotational electric dipole emission from very rapidly rotating ultrasmall grains, although a fraction of the radiation could be thermal emission from grains containing materials that are ferromagnetic (e.g. metallic Fe inclusions) or ferrimagnetic (e.g. magnetite Fe_3O_4).

Interstellar dust grains in the solar system

The solar system was formed out of interstellar gas and dust approximately 4.5 billion years ago. The formation of planetesimals and planets was accompanied in many cases by high temperatures and violent conditions, and most interstellar dust particles were destroyed. However, the class of meteorites known as carbonaceous chondrites (see METEORITES) contain small particles with unusual isotopic ratios (see ISOTOPIC ANOMALIES) which indicate that they did not form in the solar nebula, but rather must have been formed in a region with an anomalous composition (e.g. outflow from an evolved star) long before the formation of the solar system. Therefore these particles must have been part of the interstellar grain population prior to the formation of the solar nebula.

To date, five different types of presolar grains have been isolated and identified (see table 1). The evidence that the grains listed in table 1 were truly interstellar is compelling, but it is important to realize that they apparently do not include typical interstellar grains, for the simple reason that the procedures used to isolate interstellar grains in meteorites are designed to deliberately destroy silicate material (which comprises the bulk of the carbonaceous chondrite meteorite 'matrix'). These laboratory procedures are therefore *not* going to find interstellar silicate grains even if they are present.

Detectors aboard the Ulysses and Galileo probes measured impacts by interplanetary dust particles, but also observed impacts attributed to the flux of interstellar grains expected due to the 20 km s^{-1} motion of the Sun relative to the local interstellar medium. (see SOLAR WIND: ULYSSES, GALILEO MISSION) The detectors are sensitive only to collisions with the large-size end of the interstellar grain distribution, but the observed distribution of impact energies appears to be consistent with both the overall numbers and size distribution expected based on studies of interstellar extinction.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Interstellar grains in meteorites.

Composition	Diameter (μm)	Abundance†
C (diamond)	0.002	5×10^{-4}
SiC‡	0.3–20	6×10^{-6}
C (graphite)‡	1–20	1×10^{-6}
Al_2O_3 (corundum)	0.5–3	3×10^{-8}
Si_3N_4	∼1	2×10^{-9}

† Mass fraction in 'primitive' carbonaceous chondrite meteorites.

‡ SiC and graphite grains sometimes contain very small TiC, ZrC, and MoC inclusions.

The lines of evidence listed above serve to strongly constrain theoretical models for the interstellar grain population. Unfortunately, interstellar grain researchers have thus far not been able to uniquely determine the grain model from the available observational constraints, so debates continue concerning the details.

Nevertheless, there is broad consensus on a number of properties of interstellar grains:

(a) The grain population must have a broad size distribution, with grain radii[2] covering the range 0.5–300 nm—a factor of at least 10^8 in grain mass.

(b) Most of the grain mass is in particles with radii $50 \leq a \leq 300$ nm.

(c) Most of the grain area is in particles with radii $a \leq 50$ nm.

(d) Approximately 50% of the grain mass in diffuse clouds is contributed by amorphous silicate material, accounting for the broad 9.7 μm and 18 μm silicate features. Most of the silicon, iron and magnesium abundance is in solid form.

(e) Approximately 50% of the grain mass in diffuse clouds is contributed by carbonaceous material. Approximately two-thirds of the total carbon abundance is in solid form.

(f) The 217.5 nm feature is probably due to some form of carbonaceous material, possibly graphitic.

Graphite-silicate grain model

It is not possible to invert the observations to obtain a unique grain model. Instead, one makes some assumptions concerning the grain composition and the form of the size distribution, and then attempts to adjust the model to achieve a good match to the observed interstellar extinction, infrared emission and other constraints such as observed gas phase abundances.

One grain model which has proven fairly successful in conforming to observations consists of a mixture of carbon grains and silicate grains. The carbon grain material is taken to have the optical constants of crystalline graphite. A good fit to the extinction can be obtained if both graphite and silicate grains have size distributions which are approximately a power-law, $dn/da \propto a^{-3.5}$, truncated at $a_{\rm min} \approx 5$ nm and $a_{\rm max} \approx 250$ nm The 217.5 nm feature is then reproduced by the $a \leq 15$ nm graphite grains. This grain model—the combination of graphite and silicate grains, and the $dn/da \propto a^{-3.5}$ power law—was first put forward in 1977 by Mathis, Rumpl and Nordsieck, and is often referred to as the 'MRN' model. The MRN model achieves a good fit to the $R_V = 3.1$ extinction curve for diffuse clouds with essentially all of the Mg, Si and Fe in the silicate grains, and approximately two thirds of the C in the graphite grains, in reasonable agreement with observed depletions. The grains are heated to temperatures ∼18 K by starlight, and the resulting thermal emission is approximately consistent with the observed far-infrared emission at $\lambda > 60\mu$m.

[2] The particles are of course not spherical. By grain 'radius' we refer to the radius of a sphere of equal volume.

Ultra-small dust grains

Unfortunately, the original graphite-silicate grain model described above fails to reproduce the $\lambda < 50$ μm infrared emission shown in figure 3. The model lacks the ultrasmall grain component required to explain the observed $3 \leq \lambda \leq 50$ μm infrared emission from interstellar dust. The simplest modification is to allow the carbon grain distribution to extend down to very small sizes, with the smallest grains assumed to have infrared optical properties suitable to explain the emission features at 3.3, 6.2, 7.6, 8.6 and 11.2 μm when heated to the appropriate temperatures by absorption of single starlight photons. In order to have sufficient numbers of ultrasmall grains, while still reproducing the extinction curve, the size distribution of at least the carbonaceous grains can no longer be approximated by a single power law.

Other grain models

Other grain models, with different grain geometries and/or compositions, have been proposed to account for the observed interstellar extinction and infrared emission:

(a) Mathis and Whiffen proposed a grain model wherein the larger grains are porous aggregates of small graphite and silicate particles.

(b) Some authors have favored metal oxides (MgO, SiO, FeO), either in addition to or in place of silicates.

(c) Some models assume the silicate grains in diffuse clouds to be coated with a carbonaceous 'mantle' material, which might be hydrogenated amorphous carbon.

(d) Metallic Fe and FeS incorporate significant fractions of the Fe in some grain models.

(e) Small particles of crystalline silicon, with hydrogenated or oxidized surfaces, have been proposed as an explanation for the observed luminescence near ∼700 nm.

Grain models generally tend to have the bulk of the Fe, Mg, Si in dust, approximately two-thirds of the C, and about 20% of the O, resulting in dust-to-gas mass ratios of ∼0.007.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Dynamics of interstellar grains

Interstellar grains are acted on by forces due to a number of distinct physical processes, including:

(a) Gas drag forces when the grain velocity differs from that of the gas. In addition to direct collisions with atoms and ions, there is also a 'plasma drag' force on charged grains due to momentum transfer with ions which do not actually collide with the grain. For subsonic motion, the plasma drag force is typically a factor of ∼20–30 larger than the drag due to direct collisions with ions.

(b) Electromagnetic forces on charged grains. Since there is usually a 'plasma' reference frame in which the electric field is very small (interstellar plasma being a very good conductor), the electromagnetic force can be attributed to the Lorentz force $\boldsymbol{F} = Q(\boldsymbol{v}/c) \times \boldsymbol{B}$, where Q is the grain charge, $\boldsymbol{B}$ is the magnetic field and $\boldsymbol{v}$ is the grain velocity relative to the plasma.

(c) Scattering and absorption of photons ('radiation pressure') when the grain is illuminated by an anisotropic radiation field.

(d) Poynting–Robertson drag, as when a grain is moving perpendicular to a directional radiation field. While not normally important in the interstellar medium, Poynting–Robertson drag can be very important for dust grains orbiting stars (see INTERPLANETARY DUST).

(e) Recoil forces when photoelectrons or photodesorbed molecules are emitted anistropically from a grain which is illuminated by an anisotropic ultraviolet radiation field.

(f) Gravitational force. In clouds supported by gas pressure, dust grains will tend to 'sediment' toward the minimum of the gravitational potential.

Because the gas and grains are subject to different forces, the dust grains generally have a drift velocity relative to the gas. In typical diffuse clouds, anisotropic starlight can result in drift velocities of order ~ 0.1 km s^{-1}. In regions with strongly anisotropic ultraviolet radiation fields (such as photodissociation regions) the drift velocities can be larger.

The rotational dynamics of interstellar grains are also of great interest, in connection with both the problem of grain alignment (which requires alignment of the grain angular momentum vector with the local magnetic field) and with electric dipole radiation which will be radiated by very rapidly rotating ultrasmall grains. The grain angular momentum is affected by: collisions with gas atoms and molecules; recoil associated with photoelectric emission; recoil associated with H_2 formation on the grain surface; absorption and scattering of starlight by an asymmetric grain; thermal infrared emission; the interstellar magnetic field acting on the magnetic dipole moment resulting from the Barnett effect in a spinning grain; and the interstellar magnetic field acting on the magnetization induced in the spinning grain by the interstellar magnetic field (the 'Davis–Greenstein' torque associated with paramagnetic dissipation).

Because the interstellar medium is far from thermodynamic equilibrium, some of these torques can act systematically.

Grains can be driven to rotational kinetic energies much larger than $kT_{\rm gas}$ (where $T_{\rm gas}$ is the gas temperature) by collisions with gas atoms and molecules (because the grain and gas temperatures differ), by photoelectric emission or by absorption and scattering of starlight (because the starlight radiation field is not in thermodynamic equilibrium with the grain temperature), and by H_2 formation (because the H_2 abundance and gas and grain temperatures are not in thermodynamic equilibrium).

Gradual alignment of the grain angular momentum with the magnetic field can be produced by paramagnetic dissipation (because the grain rotational kinetic energy is not in thermodynamic equilibrium with the 'vibrational' temperature of the grain).

Effects of interstellar grains

Photoelectric heating

Perhaps the most important effect of interstellar dust grains is their role in heating the gas via photoelectric emission. When a dust grain absorbs an ultraviolet photon of energy $h\nu$, there is a probability $Y(h\nu)$—often referred to as the 'yield'—that a photoelectron will escape from the grain surface and thermalize with the local gas. When this happens, the kinetic energy of the emitted electron (at 'infinity', since the grain it is escaping from may be charged) acts to heat the interstellar medium. On average, grains must capture electrons as rapidly as they are ejected, and the net heating rate then is equal to the difference between the mean energy of the emitted electrons and the captured electrons. Since photoelectric yields Y may be of the order of ∼10%, and since the kinetic energy of the electron at infinity may be of the order of ∼1 eV (if the grain is not highly positively charged) the photoelectric heating mechanism, at best, converts a few per cent of the absorbed ultraviolet starlight energy into heating of the gas. While this few per cent conversion efficiency may seem small, the availability of energy in starlight is so great that photoelectric heating is generally the dominant heating mechanism for diffuse interstellar gas.

Radiative forces on dust grains

In astrophysics, anisotropic radiation fields are the rule rather than the exception, and these anisotropic radiation fields can have dynamical consequences for dust and gas. Circumstellar grains are of course subject to extremely anisotropic radiation from the star. At a typical point in the interstellar medium, the non-uniform distribution of stars in the galaxy, together with patchy obscuration by interstellar dust, results in starlight with appreciable anisotropy. The dipole component of the starlight anisotropy may be typically ∼10%.

When the radiation field is anisotropic, scattering and absorption by the grain produces a net force on the grain, commonly referred to as the 'radiation pressure' force. If the grain is coupled collisionally to the gas by gas drag and perhaps magnetic fields, then the force exerted on the grains is transmitted to the gas. These radiation pressure forces can dominate the dynamics of gas in certain regions, such as dust-forming winds from cool stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

H_2 formation on grains

Interstellar chemistry largely starts with the catalysis of H_2 on interstellar grain surfaces (see INTERSTELLAR CHEMISTRY). The inferred production rate of H_2 in the interstellar medium requires that an appreciable fraction of the H atoms arriving at interstellar grains surfaces must leave the grains as part of H_2 molecules. While the details of the kinetics remain uncertain, the overall picture is broadly as follows: when an H atom in a diffuse cloud arrives at a grain surface it has a high probability of sticking. The H atom then explores some fraction of the grain surface by either thermal diffusion or quantum tunnelling, until it either finds another H atom with which it can recombine to form H_2, or it arrives at a location on the grain surface where the binding by either van der Waals forces ('physisorption') or formation of a chemical bond ('chemisorption') is strong enough that it becomes trapped. If such trapping occurs, then the surface coverage of trapped H atoms builds up until newly arrived H atoms have an appreciable probability of reacting with a previously trapped H atom rather than becoming trapped themselves, or the adsorbed H atoms are removed by some other process (e.g. photodesorption or thermal desorption).

Coupling magnetic fields to neutral gas

Dynamically important magnetic fields are commonly present in interstellar gas (see INTERSTELLAR MAGNETIC FIELDS). In a dense molecular gas with very low fractional ionization, a large fraction of the 'free' charge resides on positively and negatively charged dust grains. Under these conditions—which prevail in high-density regions in molecular clouds, and presumably in protostellar disks—dust grains dominate the coupling of magnetic fields to the gas, since the neutral gas atoms and molecules themselves are not directly coupled to the magnetic field. The charged dust grains are coupled both to the magnetic field (by Lorentz forces) and to the neutral gas (by collisional drag). As a result, the grains will drift with a velocity which is intermediate between the velocity of the neutral gas and the velocity with which the magnetic field lines (and plasma) 'slip' through the neutral gas. Thus the dust grains determine the rate of 'ambipolar diffusion' of magnetic field lines in dense molecular regions.

Charged dust grains play a similar role in the dynamics of magnetohydrodynamic shock waves in gas of low fractional ionization but dynamically significant magnetic fields.

Formation and destruction of interstellar grains

While interstellar grains are observed to be ubiquitous, it is not obvious that this should be so. In the violent interstellar medium (see INTERSTELLAR MATTER, SUPERNOVA REMNANTS), grains can be destroyed when the gas which they are in is overtaken by a shock wave with shock speed $v_s \geq 200$ km s^{-1}, as is expected in supernova blastwaves. Estimates of the rate of occurrence of supernovae, together with models for the resulting blastwaves, lead one to estimate that the mass in interstellar grains will be returned to the gas phase on a timescale of $\sim 5 \times 10^8$ yr, short compared with the age of the Galaxy and the interstellar medium. We know that grains are injected into the interstellar medium in outflows from cool red giants and supergiants, and even in supernova ejecta, but if there were no conversion of gaseous atoms back to solid form in the interstellar medium, we would expect very little grain material to be present at any time. In particular, even for elements like Fe we would expect most of the atoms to be in the gas phase. The fact that this is not so—that most of the interstellar Fe is in fact missing from the gas phase—requires that there be efficient recondensation of Fe, Si, Ca and other elements back into solid form *in the interstellar medium*. The 'mineralogy' of interstellar grains must therefore largely reflect the surface chemistry which will occur on the surfaces of interstellar grains.

Bibliography

An excellent overall description of interstellar dust grains is provided in

Whittet D C B 1992 *Dust in the Galactic Environment* (Bristol: IOP Publishing)

The observed optical properties of dust have been recently reviewed in

Mathis J S 1990 *Ann. Rev. Astron. Astrophys.* **28** 37

The ultrasmall grain population is discussed by

Puget J L and Leger A 1989 *Ann. Rev. Astron. Astrophys.* **27** 161

An overall review of interstellar and circumstellar dust is given by

Dorschner J and Henning T 1995 *Astron. Astrophys. Rev.* **6** 271

Microwave emission from dust grains is reviewed by

Draine B T and Lazarian A 1999 *Microwave Foregrounds* ed A de Oliveira-Costa and M Tegmark (San Francisco: Astronomical Society of the Pacific) p 133

Bruce T Draine

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interstellar Magnetic Fields

The interstellar medium is permeated by magnetic fields that have important astrophysical consequences. This article describes how interstellar magnetic fields are detected, measured and mapped, the results of such observations and the role played by interstellar magnetic fields in the physics of the interstellar medium.

The intensity of SYNCHROTRON RADIATION may be used to infer the strength of the magnetic field in the plane of the sky, $B_\perp$. Synchrotron radiation is polarized perpendicular to $B_\perp$, so POLARIZATION observations map the geometry of interstellar magnetic fields. Also, the percentage polarization is a measure of the ratio B_u/B_t of uniform to total fields. Dust grains are aligned by magnetic fields, and the linear polarization of background starlight and thermal dust emission traces the geometry of magnetic fields. A new technique that also traces $B_\perp$ is mapping of linearly polarized spectral-line emission.

Faraday rotation of the position angle of linearly polarized radiation may be used to measure the strength of magnetic fields parallel to the line of sight, $B_\parallel$, if the density of interstellar electrons is known. By observing the dispersion measures of pulsars, $B_\parallel$ weighted by the electron density may be determined. Finally, detection of Zeeman splitting of spectral lines gives $B_\parallel$.

Although the various probes sample different phases and regions of the interstellar medium, a fairly consistent picture has emerged. The Galactic interstellar magnetic field has a spiral pattern, with reversals in field direction in interarm regions. Near the Sun $B_t \approx 5$–$10\ \mu$G, with a ratio of uniform to random components $B_u/B_r \sim 0.3$–0.7. The field strength increases inward in the Galaxy, with milligauss fields in the Galactic center. The Galactic magnetic field is probably maintained by a Galactic dynamo, although the details are very unclear.

The interstellar gas, cosmic ray gas and magnetic fields are an unstable dynamic system, with blast waves from SUPERNOVAE inflating the system and extending lobes of magnetic field, hot gas and COSMIC RAYS outward from the Galactic plane to form the Galactic halo.

Observations of the position angles of dust polarization generally show little small-scale variation in interstellar clouds; analysis suggests that uniform and tangled field components may be roughly comparable in strength. Magnetic field strengths increase little with gas density up to $n(H_2) \sim 10^3$ cm^{-3}, with $B \approx 10$–$20\ \mu$G in diffuse interstellar clouds and extended dark clouds; at higher densities the data are consistent with the theoretical prediction $B \propto n^{1/2}$. In some molecular cores there is evidence for the 'hourglass' pinch that would be produced by cloud contraction with the magnetic field frozen into the matter. Strengths of the uniform component of magnetic fields in molecular clouds are only slightly less than the critical values needed for support against gravitational collapse. Hence, magnetic fields play a crucial role in the STAR FORMATION process.

Observational techniques

All methods of observing magnetic fields in the interstellar medium are sensitive only either to the component of $\boldsymbol{B}$ parallel to the line of sight, $B_\parallel$, or to the component in the plane of the sky, $B_\perp$. Direct tracers involve the detection of polarized radiation emitted by or passing through the interstellar medium.

Polarization

The polarization of electromagnetic radiation may be conveniently described by the Stokes parameters I, Q, U and V, where I describes the total intensity, Q and U describe the linear polarization and V describes the circular polarization of radiation. The electric vector of an electromagnetic wave propagating along the z-axis may be resolved into two orthogonal components

$$E_x(t) = E_1 \cos(2\pi\nu t)$$

$$E_y(t) = E_2 \cos(2\pi\nu t + \phi).$$

Then

$$I = E_1^2 + E_2^2$$

$$Q = E_1^2 - E_2^2$$

$$U = 2E_1E_2 \cos(\phi)$$

$$V = 2E_1E_2 \sin(\phi)$$

and the total polarized intensity is

$$I_p = (Q^2 + U^2 + V^2)^{1/2}.$$

What are often stated for observations of linear polarization are the fractional polarization I_p/I and the position angle ψ of the major axis of the polarization ellipse,

$$\psi = \frac{1}{2} \arctan\left(\frac{U}{Q}\right).$$

If radiation is observed with a detector sensitive to linear polarization, the maximum signal will be detected when the detector is oriented at the position angle ψ. If a detector sensitive to circular polarization is used to measure the intensities I_R and I_L of the right and left circularly polarized components, Stokes V may be inferred directly from the difference. Unfortunately, there is not a standard used consistently by astronomers for the sense of circular polarization, so care must be exercised.

Tracers of $B_\perp$

Synchrotron radiation

An electron moving with velocity $\boldsymbol{v}$ at an angle θ to a magnetic field $\boldsymbol{B}$ will move in a helical path around the field lines owing to the Lorentz force $\boldsymbol{v} \times \boldsymbol{B}$ and will therefore emit electromagnetic radiation as a result of this acceleration. An ensemble of electrons is generally treated by assuming that the distribution function of the energies of the electrons is a power law with index p:

$$N(E)\,\mathrm{d}E = N_0 E^{-p}\,\mathrm{d}E.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

For an assumed uniform magnetic field in a region of extent L along the line of sight, the intensity of radiation that would be emitted is given by

$$I(\nu) = \frac{\sqrt{3}e^3 L}{8\pi mc^2}\left(\frac{3e}{4\pi m^3 c^5}\right)^n N_0 B_\perp^{n+1}\nu^{-n}a(n)$$

where $n = (p-1)/2$ and $a(n)$ is a slowly varying function of order unity. For a uniform magnetic field the fractional linear polarization is

$$\frac{I_p}{I} = \frac{n+1}{n+5/3}.$$

Hence, by measuring the position angle of the linearly polarized synchrotron radiation emitted by highly relativistic electrons ($\gamma \gg 10^3$), the direction of the magnetic field in the plane of the sky may be mapped. An estimate of the ratio of strengths of the random (B_r) and uniform (B_u) components of the magnetic field may be obtained from the degree of polarization, since for a completely random field the synchrotron radiation will be unpolarized.

By making various assumptions, the most important being equilibrium between the energy densities of the relativistic electrons and of the magnetic field, an estimate of the strength of the magnetic field may be obtained from the flux density S_ν of the synchrotron emission:

$$B_\mathrm{eq} = f(n)\frac{D^2}{V}(S_\nu \nu^n)^{2/7}$$

where D is the distance to the emitting gas, V is its volume and $f(n)$ is a function of n.

These techniques are most useful for mapping the large-scale morphologies of interstellar magnetic fields in external galaxies and in the Milky Way.

Dust polarization

Linear polarization at optical, infrared or millimeter wavelengths arises from elongated grains with their short axes aligned with the magnetic field. The Davis–Greenstein mechanism assumes that INTERSTELLAR GRAINS are aligned via paramagnetic relaxation, which damps the components of a grain's angular momentum perpendicular to $\boldsymbol{B}$, gradually aligning the spin axis with the magnetic field direction. Although the original Davis–Greenstein theory considered only thermal spins, much weaker magnetic fields can align grains because of superthermal spin energies produced by ejection of H_2 molecules from their formation sites on grains. Internal damping due to Barnett relaxation normally causes the grain to spin about its axis of greatest moment of inertia, so the long axis of the grain is oriented perpendicular to the field.

However, there are other alignment mechanisms. One which may dominate paramagnetic relaxation in protostellar jet or wind regions is Gold alignment, in which gas streaming past the grains excites their angular momenta perpendicular to the flow direction, causing the grains to tend to rotate with their long axes parallel to the flow. Unless there is reason to believe that some other process dominates, Davis–Greenstein alignment is assumed.

Grain properties and the grain alignment mechanism are not understood in sufficient detail for field strengths to be reliably inferred, but observations of linear polarization trace the position angle of the component of $\boldsymbol{B}$ in the plane of the sky.

Starlight passing through the interstellar medium will be linearly polarized owing to extinction by aligned interstellar grains along the line of sight, so long as the magnetic field is not completely random. Because Davis–Greenstein alignment produces the maximum extinction cross-section perpendicular to the magnetic field, extinction will be greater perpendicular to field than parallel to it. Hence, the position angle of the polarization ellipse of background starlight will be parallel to the magnetic field direction. On the other hand, when emission from dust grains is observed, the position angle of maximum emission will be perpendicular to the magnetic field. Thermal emission from linearly polarized dust grains at far-infrared and millimeter wavelengths may be used to map magnetic field morphologies.

Spectral-line linear polarization

Linear polarization may also arise in radio frequency spectral lines formed in the interstellar medium. This Goldreich–Kylafis effect may be used to probe magnetic field morphologies in molecular clouds. For a molecule with (say) levels with total angular momenta $F = 1$ and $F = 0$, in the presence of a magnetic field the degeneracy of magnetic sublevels will be removed and the level with $F = 1$ will be split into three sublevels. The three transitions will be polarized as usual for this situation, but so long as the three sublevels of the $F = 1$ state are populated equally, the net polarization is zero. If the line optical depth to the source is anisotropic, the radiation field will be anisotropic and the radiative rates between level $F = 0$ and the three sublevels of $F = 1$ will be unequal. Thus, unequal populations of the three sublevels can be sustained, and net linear polarization will result. Since astrophysical sources are always anisotropic at some level, linear polarization of line emission is expected. The degree of polarization depends on the degree of anisotropy, the optical depth of the spectral line, the angles between the line of sight, the magnetic field, and the axis of symmetry of the velocity field, and the ratio of the collisional excitation rate to the radiative rate. The direction of the polarization can be either parallel or perpendicular to the magnetic field; therefore, care must be taken in inferring the direction of $\boldsymbol{B}$.

Tracers of $B_\parallel$

Faraday rotation

A linearly polarized electromagnetic wave may be decomposed into two oppositely circularly polarized

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

waves. These two waves propagate through an interstellar medium that contains ions and magnetic fields with different phase velocities. The result is that the position angle of the linear polarization will rotate. This Faraday rotation produced by propagation through a path length L is given by

$$\Delta\psi = \frac{e^3}{2\pi m^2 c^2 \nu^2} \int_0^L B_\parallel(z) n_e(z)\, \mathrm{d}z.$$

Hence, observations of Faraday rotation of linearly polarized radiation (usually synchrotron radiation) may be used to infer the strength of the line-of-sight component of $\boldsymbol{B}$ if information about the electron density is known or assumed. Faraday rotation is generally quoted as a rotation measure,

$$RM = -\int_0^L B_\parallel n_e\, \mathrm{d}z.$$

Note the convention that a negative RM means $\boldsymbol{B}$ is pointed away from the observer, $B > 0$.

In the special case of observations of PULSARS, information about the electron density may be obtained from the dispersion with radio frequency in the arrival time of radio pulses. The dispersion measure is given by

$$DM = \int_0^L n_e\, \mathrm{d}z.$$

Then the mean line-of-sight component of $\boldsymbol{B}$ is

$$\overline{B}_\parallel = \frac{\mathrm{RM}}{\mathrm{DM}}.$$

This $B_\parallel$ is a vector average along the line of sight and may be much smaller than the localized fields if there are reversals in field direction.

Zeeman splitting

If a spectral line forming region is permeated by a field $\boldsymbol{B}$, the radiation is split by the normal ZEEMAN EFFECT into three separate frequencies, $\nu_{\sigma-} = \nu_0 - \nu_Z$, $\nu_\pi = \nu_0$ and $\nu_{\sigma+} = \nu_0 + \nu_Z$, where ν_0 is the line frequency, $\nu_Z = BZ$ and Z is the Zeeman sensitivity in Hz $\mu\mathrm{G}^{-1}$. For a magnetic field in the plane of the sky, the three Zeeman components are linearly polarized with the π component parallel to and the σ components perpendicular to the magnetic field direction. For a magnetic field along the line of sight, $I_\pi = 0$ and the I_σ are oppositely circularly polarized. If the magnetic field is parallel (antiparallel) to the direction of propagation of the radiation, $I_{\sigma+}$ is right (left) circularly polarized. In the general case of arbitrary angle between the line of sight and the magnetic field, all three Zeeman components will be elliptically polarized.

If B is sufficiently large, then $\nu_Z \sim \Delta\nu$ where $\Delta\nu$ is the line width. Then the Zeeman components may appear separated, and the full magnetic field strength can be derived directly from the measured ν_Z. Generally $\nu_Z \ll \Delta\nu$, and it is not possible to infer complete information about B from Zeeman observations. Although the Stokes parameters V, Q and U for the Zeeman components provide in principle full information about magnetic field strength and direction, in practice full information on $\boldsymbol{B}$ cannot be obtained owing to the extreme weakness of Q and U (typically about an order of magnitude weaker than V). The Stokes V spectra reveal the sign (i.e. direction) and magnitude of the line-of-sight component $B_\parallel$. That is, $V = (\mathrm{d}I/\mathrm{d}\nu)\nu_Z \cos(\theta)$. By fitting the frequency derivative of the Stokes parameter I spectrum $\mathrm{d}I/\mathrm{d}\nu$ to the observed V spectrum, $B_\parallel = \nu_Z \cos(\theta)/Z$ may be inferred.

For a large number of clouds whose magnetic fields are randomly oriented with respect to the observed line of sight, it is possible to infer statistically the mean correction to go from $B_\parallel$ to $|\boldsymbol{B}|$. That is,

$$\overline{B}_\parallel = \frac{\int_0^{\pi/2} |\boldsymbol{B}| \cos\theta \sin\theta\, \mathrm{d}\theta}{\int_0^{\pi/2} \sin\theta\, \mathrm{d}\theta} = \frac{1}{2}|\boldsymbol{B}|.$$

In general, only those species with an unpaired outer electron will have a magnetic moment that scales with the Bohr magneton, $M_\mathrm{B} = eh/4\pi mc = 1.4$ Hz $\mu\mathrm{G}^{-1}$. For the 1420 MHz line of H I, $Z = M_\mathrm{B}$. Molecular lines that are sensitive to the Zeeman effect have $Z \approx M_\mathrm{B}$; for example, $Z = 0.98$ Hz $\mu\mathrm{G}^{-1}$ for the 1667 MHz line of OH. Otherwise, the Zeeman effect will scale with the nuclear magneton, 1840 times smaller than M_B. Unfortunately, most molecules do not have large Zeeman splitting factors Z, so the possibilities for Zeeman observations are limited. Lines for which the Zeeman effect has been detected that have been used for measurements of interstellar magnetic fields are the 21 cm line of H I, 18 cm and 2 cm lines of OH, the 1.3 cm H_2O maser line and the 3 mm lines of CN. Other possible Zeeman molecules that might become useful probes of interstellar magnetic fields include CH, CCS and SO.

Magnetic field of the Milky Way

Polarization of starlight produced by interstellar magnetic fields was discovered accidentally about 50 yr ago during a search for polarization produced in STELLAR ATMOSPHERES. It was clear that the interstellar medium was responsible for the observed polarization because its amount correlated with interstellar extinction. Dust grains aligned by magnetic fields produce the polarization. Polarization of background stars provides a useful technique for measuring the morphology of interstellar magnetic fields near the Sun. Optical and near-infrared polarization data now exist for thousands of stars.

Figure 1 shows a plot of starlight polarizations for nearly 10 000 stars. Each line represents the polarization measurement for a star; the length is proportional to the fractional polarization and the orientation is in the direction of maximum polarization, which is parallel to the magnetic field. These data show several important characteristics of the local interstellar magnetic field. First, for Galactic latitudes b within about 10° of the Galactic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

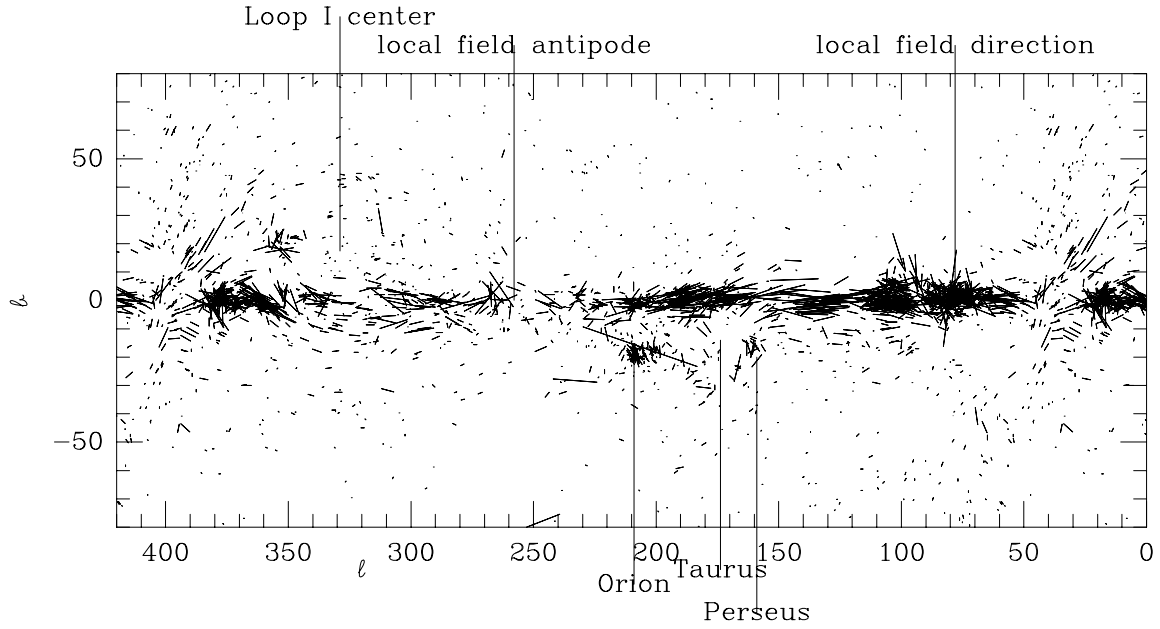


Figure 1. Starlight polarization defining the direction of the Galactic magnetic field near the Sun. The direction of the local spiral arm field, the Loop I interstellar bubble and star formation regions in Taurus, Perseus, Ophiuchus and Orion are marked.

plane, the magnetic field is generally parallel to the plane of the Galaxy. Looking directly along the local spiral arm, toward Galactic longitudes of $\ell \sim 80^\circ$ and $\ell \sim 260^\circ$, the polarization directions show convergence down the spiral arm. The polarization data suggest that the interstellar magnetic field is not heavily tangled on large spatial scales and lies in spiral arms. In principle it is possible to derive the ratio B_u/B_r from an analysis of the uniformity of the optical polarization data. The results are somewhat dependent on models that must be adopted for the analysis, but are reasonably self-consistent: $B_u/B_r \approx$ 0.7–1.0.

At positions more than $\sim 10^\circ$ off the Galactic plane there is considerable small-scale structure. Positions in Taurus, Perseus, Ophiuchus and Orion correspond to massive molecular clouds undergoing active star formation; the magnetic field lines are obviously perturbed from the regular, large-scale pattern. Finally, at high Galactic latitude several local interstellar bubbles each produced by multiple supernovae may be seen. The most prominent is Loop I, whose walls are outlined by a loop of magnetic field lines with a radius of $\sim 60^\circ$.

Synchrotron emission may be used to infer the total (uniform and random) strength B_t of the Galactic magnetic field. Maps of the Galaxy in radio synchrotron emission show spiral arms, implying that B is stronger in arm than in interarm regions. Derivation of the field strengths involves various assumptions and careful analysis—the best recent results are the following. Near the solar circle ($R = 8.5$ kpc is used here), $B_t \approx 6\ \mu$G in spiral arms and $B_t \approx 4\ \mu$G when averaged azimuthally, which includes arm and interarm regions. By scaling this azimuthal value by the observed increase in synchrotron emission with decreasing Galactic radius, at $R = 4.0$ kpc one finds $B_t \approx 8\ \mu$G. Near the solar circle the ratio of the uniform to the random magnetic field strength $B_u/B_r \approx 0.5$ from the ratio of the observed to the maximum possible synchrotron polarization.

Synchrotron emission decreases with height above the Galactic plane. When modeled as a thin and thick disk with equal emission at the Galactic plane, the scale heights are about 0.2 and 1.8 kpc at the Sun. Also, the scale height increases with Galactic radius. Because synchrotron emission depends on $B_\perp^{n+1}$, the scale height of B is significantly larger than that of the synchrotron emission.

Faraday rotation of the linearly polarized radiation from pulsars and extragalactic RADIO SOURCES provides information about the strength of the interstellar magnetic field along the line of sight. Pulsar data give the mean field $\overline{B}_\parallel$ directly from the ratio RM/DM. However, $\overline{B}_\parallel$ is a weighted mean, weighted by the electron density; this method therefore samples mainly the warm ionized component of the interstellar medium, which is believed to occupy a small fraction of the volume of interstellar space.

The pulsar data show clear evidence for reversals in the magnetic field direction at Galactocentric radii corresponding roughly to interarm regions in the Galaxy. The field in the local spiral arm points in the direction of Galactic rotation, with reversals at $R \approx 5.5$, 8.1 and 9.4 kpc. $\overline{B}_\parallel \approx 1.4\ \mu$G near the Sun, rising to $\sim 3.5\ \mu$G in the Sagittarius arm and $\sim 6.5\ \mu$G inside the second reversal radius (5.5 kpc). Note that these values are for the uniform component, B_u, only. Statistical analysis of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

deviations of $\overline{B}_\parallel$ from a uniform field or from comparisons of $\overline{B}_\parallel$ values from pulsars close together in the sky yields estimates of the strength of the random component of the magnetic field; the result is $B_r \approx 5\ \mu$G. Therefore, $B_u/B_r \approx 0.3$.

The H I Zeeman effect may also be used to measure $B_\parallel$ in diffuse interstellar gas. Extensive observations of truly randomly selected regions do not exist. One study of a large region in which RMs are consistently large and negative yielded many upper limits $B_\parallel \approx 7\ \mu$G, with detections at lower values when greater sensitivity was achieved. One position with $B_\parallel \approx 12\ \mu$G came from an H I filament probably produced by an interstellar shock which compressed both the gas and the field. Observations of the H I Zeeman effect toward well-known interstellar bubbles (Loop I, Loop II, the Eridanus bubble) yield $B_\parallel \sim 5$–$10\ \mu$G. The magnetic field has been enhanced by compression in the walls of the bubbles. When the H I Zeeman effect is detected in diffuse clouds, the observations yield typical values $B_\parallel \sim 5$–$20\ \mu$G (although there are many upper limits $B_\parallel < 5\ \mu$G), with clouds of densities up to $n(\mathrm{H\,I}) \sim$ several hundred being sampled. Hence, there is very little increase in B with gas density for $n < 10^3$ cm^{-3}.

Differences in results derived from pulsar, optical polarization, H I Zeeman and synchrotron emission studies may be attributed to the different phases of the interstellar medium being sampled. Pulsars sample the warm ionized medium, optical polarization and H I Zeeman sample primarily the neutral medium, and synchrotron emission samples the entire volume.

The central region of the MILKY WAY GALAXY has very strong magnetic fields, with strengths of several milligauss. Evidence for strong fields comes from large-scale, linear filaments of radio emission, highly polarized radio continuum emission, polarized infrared emission, Zeeman measurements of H I and OH absorption, and anisotropic scattering of radio emission from small sources. Models of the Galactic center have a circumnuclear disk threaded by a poloidal field, with the rotation of the disk producing an azimuthal component.

Magnetic fields in galaxies

A problem in studying the interstellar magnetic field in the Milky Way galaxy is that we are inside it, which allows detailed studies of the local interstellar medium but makes global study difficult. Observations of synchrotron emission from external galaxies provide a picture of the large-scale interstellar magnetic fields in SPIRAL GALAXIES. Maps of the linear polarization of synchrotron radiation are maps of the ordered component of $B_\perp$. In general magnetic field lines show a strong spiral structure. Sometimes the polarization maps define spiral structure even more clearly than tracers of material spiral arms, although often the magnetic spiral arms are found offset from the material arms. A possible reason for this is that magnetic fields in the densest parts of spiral arms are somewhat tangled or random owing to star formation or supernova activities, reducing the fractional polarization. Downstream from the spiral density-wave shock, gas will be accelerated away from the shock in streaming motions, which will act to untangle magnetic field lines and increase the polarization. Hence, the direction of $B_\perp$ in spiral galaxies may be the best technique for study of these large-scale streaming motions in spiral galaxies.

Although the magnetic fields in spiral galaxies are symmetric, there does not seem to be a single symmetry mode. In some an axisymmetric spiral arm pattern dominates with the magnetic field direction everywhere directed inwards along the spiral pattern. Others have a bisymmetric mode, with reversals in the direction of magnetic fields from one arm to the next, as seen in the Milky Way Galaxy. (It should be noted that only a few galaxies are classifiable in this way.) The strong spiral magnetic patterns are probably best explained by dynamo action in spiral galaxies. A typical value for the total magnetic field strength in galaxies is $B_t \approx 10\ \mu$G, with large uncertainties.

Magnetic fields and star formation

Magnetic fields may play two crucial roles in the star formation process—they can provide support against gravitational collapse of dense interstellar clouds and they can solve the problem of how angular momentum is transported away from collapsing protostellar cores.

Observations of the field geometry provide significant empirical input to and tests of theories. Background starlight polarization observations show that large-scale fields in clouds are smooth, with little change in the vicinity of clumps or cores. For extinction $A_V > 2$ magnitudes, near-infrared observations make it possible to probe up to $A_V \sim 40$ magnitudes. The field morphologies in dense clouds and cores observed in this way seem to be the same as in the envelope regions. However, the percentage polarization does not seem to increase much beyond about $A_V \approx 1.3$ magnitudes, which suggests that the grains in dense clouds do not produce polarization and hence cannot be used to map magnetic field morphologies at large extinctions.

Far-infrared and millimeter wavelength observations of thermal dust emission have proven to be the best technique for probing magnetic field morphology in dense interstellar clouds. Since the emission is generally optically thin, this technique samples material throughout dense molecular gas. However, sensitivity limits this technique to $A_V > 20$ magnitudes. The maps show that the polarization directions in dense molecular clouds are generally fairly uniform, at least on large spatial scales. Figure 2 shows far infrared polarization observations of $B_\perp$ in Orion Molecular Cloud 1 (OMC1). The typical percentage polarization is ∼2%, although values up to ∼10% are observed. Analysis of the OMC1 data suggests that there is an 'hourglass pinch' to the field lines, as expected toward dense cores on theoretical grounds. At the higher angular resolution available with mm-wave arrays, more structure is seen. In particular, near the dense central core there is a change in position angle of almost

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

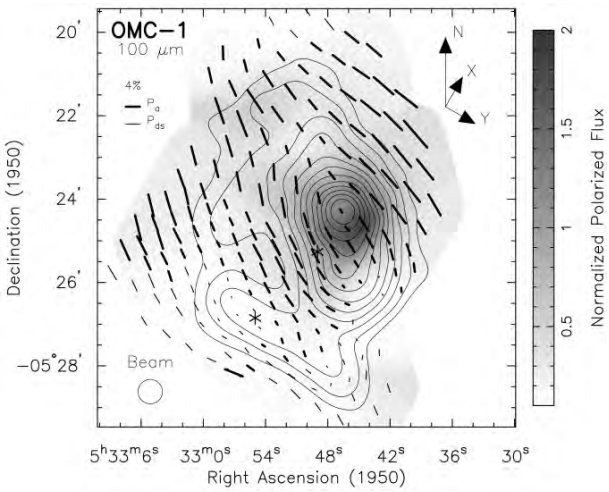


Figure 2. Geometry of the magnetic field in the OMC1 star formation region from 100 μm linearly polarized dust emission.

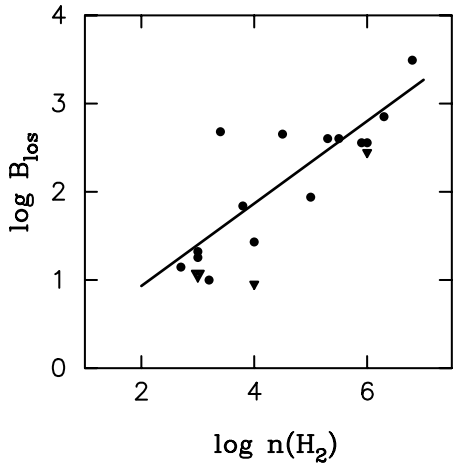


Figure 3. Relationship between magnetic field strength and gas density in self-gravitating molecular clouds.

90°, which may be due to a twisted magnetic field or to Gold alignment in the outflow from the core. However, the main geometry of the field is perpendicular to the minor axis of the cloud, which is expected if the tension of the uniform field is providing significant support against gravity.

The crucial parameter for evaluating the importance of magnetic fields in star formation is the ratio of the mass to the magnetic flux, M/Φ_B. Theoretical values for M/Φ_B (see also MAGNETIC FIELDS IN STARS) come from detailed computer simulations of interstellar clouds threaded by magnetic fields. The critical value of M/Φ_B is

$$(M/\Phi_B)_{\rm critical} \approx 1/(63\ G)^{1/2}.$$

Clouds that are initially magnetically supercritical ($M/\Phi_B > (M/\Phi_B)_{\rm critical}$) will collapse on relatively short time scales unless there is some other support mechanism. If the magnetic field was insufficient to stop the initial collapse, its compression during collapse cannot bring the cloud into equilibrium and halt the collapse. On the other hand, an initially magnetically subcritical ($M/\Phi_B < (M/\Phi_B)_{\rm critical}$) cloud will become unstable in its core owing to ambipolar diffusion. Only ions are frozen into the magnetic field, while the neutral material (by far the majority of the mass) can contract gravitationally unaffected directly by the magnetic field. However, neutrals will collide with ions in this process. This frictional force will provide support against gravity to the neutrals, but there will be a drift of neutrals into the core without a significant increase in the magnetic flux; this is ambipolar diffusion. Eventually M/Φ_B in the core becomes supercritical, and dynamical collapse and star formation can proceed.

M/Φ_B can be determined empirically from the magnetic field strength and the column density. Since

$$M = 2.8N(\mathrm{H_2})m_{\rm H} \times A \text{ and } \Phi_B = BA,$$

$$\frac{(M/\Phi_B)_{\rm observed}}{(M/\Phi_B)_{\rm critical}} = \frac{2.8N(\mathrm{H_2})m_{\rm H}(63\ \mathrm{G})^{1/2}}{B}$$

where A is the surface area of a cloud over which $N(\mathrm{H_2})$ and B are determined.

The role of magnetic fields may also be investigated with the virial theorem (ignoring possibly important terms such as surface pressure):

$$\mathcal{W} + 2\mathcal{T} + \mathcal{M} = 0$$

where the gravitational, kinetic, and magnetic virial terms respectively are:

$$\mathcal{W} = -\tfrac{3}{5}aGM^2/R$$

$$\mathcal{T} = \tfrac{3}{2}M\sigma^2$$

$$\mathcal{M} = \tfrac{1}{3}b|\boldsymbol{B}|^2R^3.$$

For the simple case of a uniform-density sphere, $a = b = 1$. A flattened, centrally condensed geometry will cause $a > 1$. For magnetically critical simulation models, $a \approx 1.2$ and $b \approx 0.3$.

Measurements of magnetic field strengths in molecular clouds depend on detection (and sometimes mapping) of the Zeeman effect in molecular lines (although when atomic hydrogen comes from photodissociation of H_2 near H II regions, the H I line may trace molecular gas). Most molecular Zeeman detections and sensitive limits have been in the 18 cm ground-state Λ doublet lines of OH. However, the OH radical probably does not trace densities above $n(\mathrm{H_2}) \sim 10^4$ cm^{-3} in most cases. An exception is Λ doublet OH lines that arise in excited rotational states, which require high gas densities for collisional excitation. Recently, detections of the Zeeman effect have been achieved toward the OMC1 and DR21(OH)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

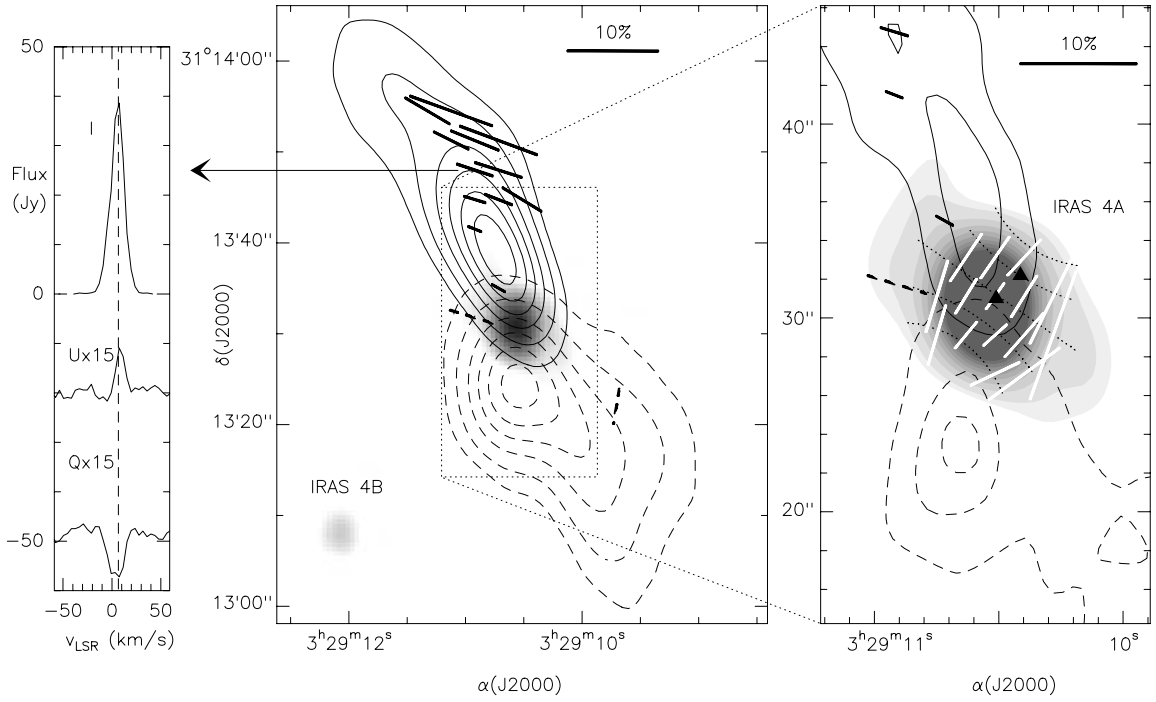


Figure 4. CO bipolar outflow in the NGC1333 IRAS4 star formation region. The geometry of the magnetic field is mapped by the linear polarization of dust continuum and CO spectral line emission.

cores in the 3 mm lines of CN, which sample densities $n(H_2) \sim 10^{5-6}$ cm^{-3}.

To date, there have been 15 measurements of $B_\parallel$ in self-gravitating molecular clouds and an additional 12 sensitive upper limits, mainly in extended dark clouds with $n(H_2) \sim 10^3$ cm^{-3}. All 27 results are shown in figure 3. The data show a clear correlation of $B_\parallel$ with $n(H_2)$; the large scatter is due partially to the fact that only one component of $\boldsymbol{B}$ is observable. For $n(H_2) > 10^3$ cm^{-3}, the data are consistent with $B \propto n^{1/2}$, a result in agreement with theory.

A statistical analysis of the Zeeman results yields $(M/\Phi_B)_{\rm observed} \approx 1 - 2(M/\Phi_B)_{\rm critical}$. Thus, magnetic fields are sufficiently strong to play a significant and perhaps crucial role in the support of molecular clouds and in their evolution. The analysis in terms of the virial theorem yields the result that kinetic and magnetic energies are approximately the same, and clouds are in approximate virial equilibrium.

Gas velocities in molecular clouds are generally supersonic by a factor of a few but are approximately equal to the Alfvén velocity $(B/(4\pi\rho)^{1/2})$, which suggests that internal motions are due to MAGNETOHYDRODYNAMIC (MHD) WAVES. Although MHD waves might be expected to organize supersonic motions and significantly lengthen the dissipation time scale, recent supercomputer simulations suggest that Alfvén waves will dissipate on a time scale not much longer than that of non-magnetic turbulence. The rate of star formation may be regulated by the ambipolar diffusion rate and by the rate of injection of mechanical energy into clouds. Photoionization could play an important role, since the ambipolar diffusion rate would depend strongly on the ionization rate. Despite the constant dissipation of the turbulent energy, it may be continuously re-generated by the winds of stars that form inside the clumps and by torsional Alfvén waves that also transport angular momentum outward.

Theories of bipolar winds from YOUNG STELLAR OBJECTS often include a significant role for the magnetic field. Figure 4 shows an example of the use of linearly polarized spectral-line emission to map the geometry of magnetic fields in the CO bipolar flow from the NGC1333 IRAS4 region. The flow is initially at a small angle to the field, but the magnetic field appears to deflect the flow so that it becomes parallel to the field.

Future work

Our understanding of the interstellar magnetic field and its role in the formation of interstellar clouds and in star formation is still very incomplete. New telescopes (such as the upgraded Arecibo telescope, the Green Bank telescope, the Atacama Large Millimeter Array and the Stratospheric Observatory For Infrared Astronomy) should produce an explosion of new and improved data during the next decade. Comparison of these data with the detailed simulations of the interstellar medium and of star formation regions that advances in computer technology are enabling should be a focus of future research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The techniques and results of polarization observations of the interstellar medium are reviewed in

Roberge W G and Whittet D C B (ed) 1996 *Polarimetry of the Interstellar Medium (ASP Conf. Ser. 97)*

A detailed description of the large-scale magnetic field of the Milky Way Galaxy is in one chapter of the above volume:

Heiles C 1996 A comprehensive view of the Galactic magnetic field, especially near the Sun *Polarimetry of the Interstellar Medium (ASP Conf. Ser. 97)* ed W G Roberge and D C B Whitter

A recent review of the methods and results of polarization observations of dust scattering and emission is

Weintraub D A, Goodman A A and Akeson R L 2000 *Protostars and Planets IV* ed V Mannings, A P Boss and S S Russell (Tucson, AZ: University of Arizona Press) in press

The study of the OMC1 cloud from which figure 2 is taken is

Schleuning D A 1998 *Astrophys. J.* **493** 811

The Zeeman effect and its application to the measurement of magnetic fields in interstellar clouds are described in

Crutcher R M 1999 *Astrophys. J.* **520** 706
Crutcher R M *et al* 1993 *Astrophys. J.* **407** 175

An article that summarizes knowledge about large-scale magnetic fields in galaxies and the intergalactic medium and discusses theories about the origin of cosmic magnetic fields is

Zweibel E G and Heiles C 1997 *Nature* **385** 131

Richard Crutcher

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interstellar Matter

The enormous volume of space between the stars in the Milky Way Galaxy is filled with interstellar matter (ISM). The ISM plays a central role in the processes of STAR FORMATION and GALAXY EVOLUTION. Stars form from the ISM in dense molecular clouds. The radiant and mechanical energy produced by stars heats, ionizes, and produces structures in the ISM. Gradual or catastrophic mass loss from stars enriches the ISM in heavy elements and causes the composition, distribution, and physical state of the ISM to change as the galaxy evolves. Current studies of the ISM are aimed toward characterizing its basic properties, determining the physical processes that control these properties, and establishing the role the ISM plays in the evolution of the Galaxy.

The existence of matter between the stars was first proposed by WILHELM STRUVE in 1847 based on an analysis of star counts which suggested that the number of stars per unit volume decreases with distance from the Sun. Struve proposed that the starlight was experiencing absorption proportional to distance. It was not until 1909 that Kapteyn realized the full significance of this interstellar extinction. Shortly later Barnard documented through direct imagery the irregular variations in the distribution of the absorbing matter. The identification of the source of the extinction with small particles of interstellar dust was finally accepted in the 1930s through the work of Trumpler and Stebbins, Huffer, and Whitford.

Diffuse emission nebulae powered by O stars with $T > 30\,000$ K were the component of the ISM first observed spectroscopically by astronomers The spectrum of the brightest example, the Orion nebula, was recorded in 1863 by Huggins. In the late 1920s Eddington, Zanstra and Bowen were able to show that such emission nebulae were clouds of ionized interstellar hydrogen containing He and trace amounts of heavier elements.

Absorption by interstellar Ca^+ in the interstellar gas was first spectroscopically recorded by Hartmann in 1904. However, it was not until the early 1930s that Plaskett and Pearce obtained extensive surveys of Ca^+ absorption in a large number of stars and convincingly associated the absorption with gas between the stars. They showed that the strength of the Ca^+ absorption increased with distance and that the absorption line radial velocity varied with distance and direction according to that expected from a simple model of gas rotating in a disk galaxy.

In more recent times the full diagnostic power of observations across the entire electromagnetic spectrum have been used to study emission and absorption produced by interstellar gas and dust in order to determine its composition, distribution, and physical state.

Components and composition

The ISM (gas + dust) in the solar vicinity has Population I element abundances. This has been established through abundance studies of the gas in diffuse emission nebulae (see GASEOUS NEBULAE), optical and ultraviolet absorption line studies of gas in neutral hydrogen regions, and by inference from studies of the composition of matter in the atmospheres of young O and B stars that have recently formed from the ISM. The ISM gas phase composition by relative number of atoms is 91% H, 9% He, and a trace abundance of the heavier elements.

Given the great abundance of H and He in the gas, much of the physics of the ISM involves interactions among these atoms, electrons, and photons. The H can exist as H^0, H_2, or H^+ while the He can exist as He^0, He^+, or He^{2+}. Absorption of radiation by H and He plays a dominant role in determining the nature of the average radiation field in interstellar space. In a neutral hydrogen region the strong H I absorption at the 13.6 eV Lyman edge results in an average radiation field resembling that of A and B type stars for $3 < E < 13.6$ eV with a greatly reduced intensity for $E > 13.6$ eV from H I absorption. At energies between ~3 and 1 eV the interstellar radiation field is dominated by the light from G and K stars.

The ionization state of the heavier elements in an H I REGION is determined by the ionization potential of the element compared to 13.6 eV. As a result, the abundant heavy elements listed in order of decreasing abundance O, C, Ne, N, Mg, Si, Fe, and S exist in the ionic states O^0, C^+, Ne^0, N^0, Mg^+, Si^+, Fe^+, and S^+. In H II REGIONS where the H I is photoionized by radiation from stars with $T > 30\,000$ K, the heavier elements are in higher states of ionization, e.g. O^+, O^{2+}, C^{2+}, C^{3+}, etc, with the actual ionic ratios controlled by the ionizing radiation field of the star. In the hot ISM where the gas is heated to $T > 10^6$ K by shock processes, the heavy elements exist in high states of ionization created by electron collisions. Therefore, elements such as O and C will exist as O^{6+}, O^{7+}, C^{6+}, and C^{7+}, with the actual ionic ratios determined by the kinetic temperature of the electrons (see INTERSTELLAR CHEMISTRY). A study of these very different gas phases requires observations in the radio, infrared, optical, ultraviolet, extreme ultraviolet and x-ray regions of the spectrum.

The gas phase abundances of HEAVY ELEMENTS in H I regions are reliably measured by ultraviolet absorption line spectroscopy toward hot stars. The line of sight column densities of H I and H_2 can be measured via the H I Lyman series absorption lines and the H_2 Lyman band absorption in the far-ultraviolet (see LYMAN ALPHA ABSORPTION). The metal line column densities can be determined from the many resonance line absorption transitions found in the ultraviolet. The results of these studies with the Copernicus Satellite (1972–1980) and more recently the Goddard High Resolution Spectrograph (1990–1997) aboard the Hubble Space Telescope (1990–) reveal that many of the heavy elements have gas phase abundances substantially less than Population I abundances because of the incorporation of the heavy elements into solid particulate matter. This effect is referred to as element depletion. For sight lines through moderately dense clouds in the galactic disk, the highly refractory elements such as Fe, Cr, and Ni have gas phase abundances reduced by factors of 100 to 200 while for other

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

elements, such as S, P, and Zn, the gas phase abundances are only reduced by factors of several. Si, Mg, and Mn exhibit intermediate reduction factors ranging from 10 to 30.

A detailed study of elements both missing and present in the gas phase of the ISM provides important indirect information about the presence or absence of those elements in interstellar dust. It is found that there is a general progression toward increasing gas-phase abundances of the depleted elements from cool neutral clouds found in the galactic disk to warm neutral clouds in the disk to warm neutral clouds in the galactic halo at distances away from the mid-plane of up to 1 kpc. This progression provides information about the destruction of dust grains in those processes that transport matter between these different galactic regions.

The INTERSTELLAR GRAINS can also be studied via their emission, scattering, and absorption properties. Approximately 1% of the mass of the ISM is in the form of interstellar dust. The most abundant particles consist of solid silicate grains and a separate carbon component which may be in the form of graphite grains. The silicates have been spectroscopically identified through infrared emission and absorption features at 9.7 and 18 μm while the graphite grains may be responsible for the 217.5 nm interstellar extinction bump. The grains likely have a distribution of sizes ranging from very small up to several μm in diameter. In the denser interstellar regions the grains can accrete atoms which form mantles on the grain cores. In the densest regions, mantles of H_2O ice are detected through an infrared absorption feature near 3.1 μm. The grains can affect the intensity of the interstellar radiation field though the extinction they produce, especially in dense interstellar regions. By locking up heavy elements from the gas the grains remove important coolants from the ISM. The grains also have significant roles in several interstellar processes including molecule formation and the heating of the gas through both molecule formation and photoelectric heating. The most abundant molecule observed in interstellar space, H_2, forms on the surfaces of grains. The dust also likely includes a population of very small grains, the PAHs or polycyclic aromatic hydrocarbons, which consist of planar molecules of combined benzene rings with attached H atoms. These particles have nm sizes and may represent the dividing line between large molecules and small particles. The PAHs are the leading candidates for a number of infrared emission bands.

High-energy (10^6–10^{20} eV) COSMIC RAYS exist in the space between the stars. Cosmic ray protons are most abundant with $n(\mathrm{p}) \sim 10^{-10}$ cm^{-3} and $n(\mathrm{e})/n(\mathrm{p}) = 10^{-2}$. The cosmic ray electrons have lower abundances than protons because they lose their energy more rapidly than the protons through the production of galactic non-thermal synchrotron radiation. Cosmic rays are $\sim 10^6$ times over abundant in various light elements because of cosmic ray spallation collisions of protons with heavier nuclei. Although the cosmic rays have a very low particle density, they can have substantial effects on the physical state of the thermal gas in the ISM through the processes of cosmic ray ionization and heating. Cosmic ray collisions with the gas also produce galactic γ-rays. Cosmic rays interacting with the galactic magnetic field may have important dynamical consequences. For example, the pressure of cosmic rays may play a role in the support of the gas found in the galactic halo.

The galactic general magnetic field in the solar region of the Galaxy has an average absolute strength of $\sim 5 \times 10^{-6}$ G. It contains an ordered component and an irregular component of roughly comparable strength. High-energy cosmic ray electrons spiraling in the galactic magnetic field produce non-thermal SYNCHROTRON RADIATION. Spinning dust grains are aligned by the magnetic field which leads to the observed interstellar polarization of starlight. The magnetic field has considerable dynamical effects since its pressure ($B^2/8\pi$) is comparable to the interstellar gas pressure. In dense interstellar regions the magnetic field strength is greatly amplified. Theories of the processes causing the collapse of an interstellar cloud into a protostar must involve a consideration of the dynamical consequences of the INTERSTELLAR MAGNETIC FIELD. On the larger scales, loop like galactic interstellar structures may be dynamically controlled by magnetic fields.

The different interacting components of the ISM discussed above have comparable energy densities (and pressures). For example, the energy density of the ISM radiation field from 91.2 nm to 8 μm is $\sim$0.5 eV cm^{-3}, the cosmic rays have $\sim$1.5 eV cm^{-3}, and $B^2/8\pi$ for a magnetic field with $B = 5 \times 10^{-6}$ G is 0.6 eV cm^{-3}. Gas motions associated with turbulence and waves in the ISM are estimated to have an energy density of $\sim$1 eV cm^{-3}. The different thermal phases of the gas discussed in the next section have $P/k = nT \sim 3 \times 10^3$ cm^{-3} K which implies an energy density of 0.3 eV cm^{-3}. The similarity of these energy densities (and pressures) is probably not a coincidence but is suggestive of a strong interaction and resulting energy and pressure equipartition among these different components of the ISM.

Phases and physical conditions of the ISM

The conditions in the ISM deviate very significantly from those found in a gas in thermal equilibrium. The mean energy density of the ISM radiation field is $\sim$0.5 eV cm^{-3} which corresponds to the energy density of a Planck black body distribution at $T \sim 3$ K. However, the spectrum of the interstellar radiation field is produced by the integrated but diluted radiation from stars with temperatures ranging from $\sim$3000 to $\sim$30 000 K. The extremely dilute nature and spectral hardness of this radiation field causes huge deviations from conditions found in thermal equilibrium. In the ISM, the populations of atomic and molecular energy levels and of different states of ionization will not be determined by the temperature of the gas but from the actual dominating processes for populating and depopulating the given state. The temperature assigned to gas in the ISM only refers to the fact the gas particles have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a distribution of velocity reliably described by a Maxwell–Boltzmann distribution at a particular temperature. The existence of such a velocity distribution is caused by the overwhelming number of elastic compared to inelastic particle collisions in a gas consisting of mostly H and He. Thus, the temperatures given for various interstellar regions only refers to the kinetic temperature of the gas.

The ISM of the Milky Way contains separate gas phases with kinetic temperatures that are cold ($T \sim 10$ K), cool (100 K), warm (10^4 K) and hot (10^6 K). The different phases of the ISM are listed in table 1 with approximate properties given for each phase including the kinetic temperature, T(K), the representative physical density, $n(\mathrm{cm}^{-3})$, the volume averaged density, $\langle n(\mathrm{cm}^{-3})\rangle$, the exponential scale height above the galactic plane, h(kpc), and the very uncertain volume filling factor, f. Most of the mass of the ISM (∼80%) is contained in the confined and extended components of the neutral gas. The extended component of the warm ionized gas contains ∼15% of the mass and the very extended hot gas contributes ∼3% to the mass.

In molecular clouds the hydrogen is mostly molecular and cold ($T \sim 10$ K). The molecular clouds are traced by emission from CO and other molecules. Although the molecular clouds comprise a small fraction of the volume of the ISM, they are estimated to contain ∼50% of the mass of the ISM. Molecular clouds are not normally considered a 'phase' of the ISM since they are not in pressure equilibrium with the more widely distributed gas.

The cold neutral medium (CNM) contains H I and H_2 and is studied via 21 cm H I emission and absorption, and ultraviolet absorption by H I, H_2, and various heavy elements. With a temperature of ∼80 K and a physical density of ∼40 atom cm^{-3} the CNM occupies only ∼3% of the volume of the ISM in the galactic disk.

The warm neutral medium (WNM) produces H I 21 cm emission but weak absorption. It can also be studied via ultraviolet absorption line spectroscopy. With $T \sim 8000$ K and $n \sim 0.4$ cm^{-3} the WNM occupies a substantial fraction (∼35%) of the volume of the ISM in the galactic disk.

The warm ionized medium (WIM) has been studied through the optical emission lines of Hα, [N II], and [S II] it produces. Observations of radio wave dispersion in the direction of pulsars yields a direct measure of the line-of-sight value of the electron column density in the WIM. With $T \sim 8000$ K and $n \sim 0.2$ cm^{-3}, the WIM occupies ∼25% of the volume of the gas in the disk of the Milky Way. The power (∼10^{41} erg s^{-1}) required to maintain the ionization of the WIM is comparable to the total power input of all galactic supernovae or ∼15% of the ionizing radiation emitted by galactic O stars.

The hot ionized medium (HIM) contains gas with $T \sim 10^6$ K and $n \sim 3 \times 10^{-3}$ cm^{-3}. It is detected through its soft x-ray emission. Somewhat cooler hot gas associated with the HIM has been detected through highly ionized atomic absorption produced by O VI, N V, and C IV and by emission from C IV. The HIM has a very uncertain volume filling factor of 40 ± 25%.

The different gas phases probably exist in rough pressure equilibrium. Various techniques have been used to estimate the thermal pressure in the neutral phase of ISM. The pressures are most often reported as $P/k = nT$ (cm^{-3} K). In neutral interstellar regions containing C^0 a typical value of P/k is $\sim 3 \times 10^3$ cm^{-3} K. However, there are large deviations from this typical value, particularly in regions overpressurized by shock waves from exploding stars or from the winds of hot stars. Pressures are substantially elevated in H II regions where relatively dense gas is rapidly heated to $T \sim 8000$ K by the O stars that have recently formed. Such overpressurized H II regions expand outward as the gas in them tries to achieve pressure equilibrium with the general ISM. In dense molecular regions pressures are found to be much larger as a result of the self-gravity of the gas. The pressure in the ionized gas at distances of ∼1 kpc from the GALACTIC PLANE has been estimated to be ∼3000 cm^{-3} K. The rate of drop off of gas pressure at larger distances away from the galactic plane is not well determined.

The total pressure produced by all the components of the ISM must be capable of supporting the galactic gas layer. At the galactic mid-plane that weight per unit area requires a total pressure, P/k, of approximately 30 000 cm^{-3} K. The combined pressure contributions from the various components of the ISM (e.g. gas thermal pressure, magnetic pressure, cosmic ray pressure, gas turbulent pressure) must add up to this number. In regions of greatly elevated pressure such as in OB associations with frequent supernova explosions, the higher pressure may drive gas into the halo in a flow pattern sometimes referred to as a 'galactic fountain'. In such a flow, high temperature ($T > 10^6$ K) gas bursts out of the galactic plane into the halo, cools by adiabatic expansion, and eventually rains back onto the galactic disk as warm and cool infalling fountain clouds.

General distribution

The stirring of matter in the galactic disk by supernova explosions, stellar winds, cloud–cloud collisions and other processes causes the various phases of the ISM to extend different distances away from the plane of the Galaxy. The stratification of this 'galactic atmosphere' is roughly described by layers with exponential density distributions away from the galactic plane with scale heights $h \sim 0.1$ kpc for the cool clouds, ∼0.4 kpc for the warm neutral gas, ∼1 kpc for the warm ionized gas, and ∼5 kpc for the hot ionized gas. These scale heights have been determined by studying the fall off of $N(\mathrm{X}) \sin |b|$versus$|z|$ where $N(\mathrm{X})$ represents the column density of species X toward stars at galactic latitude b having different distances, z, away from the galactic plane. For cold neutral gas and warm neutral gas, ultraviolet observations of H I Lyman alpha and H_2 Lyman band absorption are used to obtain $N(\mathrm{X})$. For the warm ionized medium, the observations are of pulsar dispersion measure which is proportional to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Phases of the gaseous ISM†.

Phase	State of H	$\sim T$ (K)	$\sim n$ (cm^{-3})	$\langle n(\mathrm{cm}^{-3})\rangle$	h (kpc)	$N_\perp$ (cm^{-2})	f
Cold neutral medium (CNM)	H^0, H_2	80	40	1.1	0.1	2.1×10^{20}	3%
Warm neutral medium (WNM)	H^0	8000	0.4	0.16	0.4	1.8×10^{20}	35%
Warm ionized medium (WIM)	H^+	8000	0.2	0.024	1	6.7×10^{19}	25%
Hot ionized medium (HIM)	H^+	10^6	0.003	0.001	5	10^{19}	$40 \pm 25\%$

† Molecular clouds are not listed as a phase of the ISM since they are not in pressure equilibrium with the more diffuse gas. In molecular clouds the hydrogen is mostly molecular, $T \sim 10$ K, and $n > 300$ cm^{-3}.

line of sight electron column density. The ∼5 kpc extent of the hot phase has been traced by observations of highly ionized atoms such as C^{3+} and N^{4+} toward stars in the galactic disk and halo. The thick warm neutral gas layer is often called the Lockman layer while the warm ionized gas layer is called the Reynolds layer.

The ISM gas density distribution in the Galaxy exhibits a radial dependence. The mid-plane densities of H_2 in the ISM as traced by the CO molecule are greater for regions interior to the solar circle and appear to peak at a galactocentric distance R of approximately 5 kpc. In the outer Galaxy, the H I is observed to warp and flare. The warp is to positive values of b over the region $l = 70$–140° and to negative values of b over the region $l = 240$–300°. The warping and flaring in the outer Galaxy has been observed to extend over a region ∼10 kpc thick. The radial extent of the neutral H I in the Galaxy is roughly 30 kpc or more.

Structures

The various kinds of structures found in the interstellar gas provide important information about the many physical processes affecting the ISM. In the disk of the Galaxy the cool interstellar matter has a highly irregular cloud-like structure. Information about the extents and sizes of these clouds can be inferred from observations of dust extinction toward large numbers of stars and from observations of interstellar atomic absorption lines toward stars. While clouds of many shapes and sizes probably exist, the extinction data yield information about the average properties of clouds containing dust. A summary of those properties is given in table 2. INTERSTELLAR ABSORPTION LINE studies confirm the existence of ∼7 cool diffuse clouds kpc^{-1} along the lines of sight to stars within several kpc of the Sun. Extreme examples of the large clouds are the molecular clouds which often trace regions of active star formation. Molecular clouds have masses in the range 10^5–10^6 $M_\odot$.

In addition to these clouds found in the mostly neutral gas, the ISM contains numerous structures which illustrate the direct interaction between stars and the ISM. These include the photoionized H II or diffuse nebulae which are found in star forming OB ASSOCIATIONS and the associated REFLECTION NEBULAE which are evident when the stars illuminating the gas and dust are cooler than B2 and therefore do not have adequate amounts of Lyman continuum radiation to ionize the gas. Other examples

Table 2. Average properties of neutral interstellar clouds traced by dust.

	Diffuse clouds	Large clouds
$\langle E(B-V)\rangle$	0.061 mag	0.29 mag
Number (kpc^{-1})	6.2	0.8
Contribution to $\langle n(H)\rangle$	0.7 cm^{-3}	0.4 cm^{-3}
Diameter	4 pc	70 pc
$n(H)$ in cloud	40 cm^{-3}	10 cm^{-3}
Mass for spherical cloud	50 $M_\odot$	$6 \times 10^4 M_\odot$

are the SUPERNOVA REMNANTS and PLANETARY NEBULAE which trace the processes of violent and peaceful stellar death, respectively. The shock heating of the gas in the supernova remnants can create large volumes of hot gas that expands outward and eventually joins with the general ISM. The processes of both violent and peaceful stellar death provide a means to enrich the ISM in elements created at some stage in the interiors of the dying stars.

Energetic events occurring in the disk can create H I shells and H I supershells. The extensions of these structures into the halo may produce structures called worms and galactic chimneys. Some of these structures mark the sites of OB associations where the combination of multiple stellar winds and multiple supernovae create over-pressurized regions which can in some cases feed hot gas into the halo and leave behind H I shells and supershells.

The H I observed at intermediate velocity ($|v| = 30$–100 km s^{-1}) at high latitudes in the northern galactic hemisphere traces substantial amounts of gas in arch-like features. The high velocity H I clouds ($|v| > 100$ km s^{-1}) which cover ∼20% of the sky trace gas with a number of different origins, including gas stripped out of the Magellanic Clouds through tidal interactions with the Milky Way, galactic fountain gas, and perhaps infalling intergalactic gas. The recent discovery that the extensive high velocity cloud Complex C has a metallicity of 0.09 ± 0.02 Solar implies that the galaxy may still be accreting low-metallicity material which could be a remnant of the formation of the Local Group of galaxies.

In the galactic disk the ISM has a greater density in the galactic SPIRAL ARMS which are waves of enhanced density that move through the gas and the stars of the disk. Local examples of these spiral arms in the Milky Way are clearly traced by the young stars of OB associations and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the molecular clouds from which the stars have recently formed. The three most prominent sprial arms near the Sun are the Perseus Arm, the Local Arm and the Sagittarius Arm. These three sprial arms have widths of ∼1 kpc and are separated by ∼1.5 kpc. In between the spiral arms, the interarm gas has a lower density and is more smoothly distributed.

Near the galactic center there is a $\sim 10^9$ solar mass concentration of the ISM, mostly in the molecular form in an expanding ring of ∼200 pc radius. This expanding structure may have its origins in the explosive release of energy from the galactic center or from the dynamical effects of a stellar bar. The ISM in this highly disturbed molecular ring has much more extreme properties than found elsewhere in the Galaxy.

Controlling physical processes

The physical state and distribution of the different components of the ISM are controlled by a number of physical processes. Identifying the dominant (controlling) processes is an important goal of studies of the ISM. In many interstellar situations it is assumed that the populations of ionic states or atomic and molecular energy levels are not changing with time. Under such 'steady-state' assumptions an identification of the dominating processes allows for a determination of the basic properties of the gas. However, in situations where the time evolution is rapid the steady-state assumption may be invalid, in which case it is necessary to follow the full time-dependent evolution of the gas.

The ionization state of the interstellar gas in H I and H II regions is controlled by starlight photoionization balanced by radiative and dielectronic recombination. In H I regions the radiation is produced by the integrated radiation from stars which is strongly attenuated for $E >$ 13.6 eV by H I absorption. In an H II region one or more O stars are usually the dominant sources of the ionizing radiation. In the hot ISM the ionization is controlled by electron collisions followed by radiative and dielectronic recombination.

The equilibrium abundances of molecules in the ISM are controlled by the dominating molecule production and destruction processes. The most abundant molecule, H_2, forms on the surfaces of grains and in diffuse clouds is destroyed by the absorption of far-UV radiation between 91.2 and 111 nm in the Lyman and Werner absorption bands. Following the absorption into an excited molecular electronic state, the excited H_2 molecule undergoes radiative emission into a bound state of the ground electronic state or into the unbound vibrational continuum of the ground electronic state which results in dissociation. The dissociation process occurs ∼10% of the time. Once the H_2 has formed, it can react with heavy atoms in various molecular reaction chains to form many of the ∼100 different molecules found in dense interstellar clouds. The atom–molecule reaction process is greatly accelerated by the atomic ionization occurring in dense interstellar clouds penetrated by ionizing cosmic rays. The ion–molecule chemistry occurring in diffuse and dense clouds seems capable of explaining the existence of many of the molecules found in these clouds. In very dense interstellar molecular clouds the dissociating UV radiation is absent. This allows the molecular chemistry to proceed to the creation of very complex polyatomic molecules containing as many as 10–13 atoms. Interstellar grains also likely participate in the production of molecules other than H_2. For example, in addition to molecule production processes similar to the formation of H_2 on grain surfaces, the exposure of a grain with a complex icy mantle to a harsh environment may lead to the ejection of various complex molecules from the matter in the surface layers.

The equilibrium kinetic temperature of the gas is controlled by the important cooling and heating processes. In most interstellar regions the cooling is produced by collisional excitation of atomic fine structure levels followed by radiative deexcitation with the loss of the photon from the region. In H I regions electron or hydrogen collisional excitation of low lying states in C^+ and O^0 can lead to the production of the cooling transitions of C II at 157 μm and O I at 204 and 76 μm. In hotter H II regions where the colliding particles have more energy, the important cooling transitons occur as forbidden optical emission lines from O^+, O^{2+}, and N^+. In the H II regions surrounding hot stars, the dominant heating process is the difference in the relatively large kinetic energy liberated to the gas by ejected photoelectrons (mainly from H I) as compared with the energy of the recombining electrons. The heating processes are not as well understood for H I regions. Several possibilities include heating caused by the photoelectric ejection of rapidly moving electrons from grains, or the heating occurring as high-speed H_2 molecules are ejected from grain surfaces during the molecule formation process. In this ejection process a modest fraction of the H_2 energy of formation is available to eject the molecule from the grain with a substantial energy of motion.

Theories of the physical state of the ISM

Theories of the interstellar gas and dust in the galaxy must be able to explain the co-existence of the different observed phases of the ISM. Over certain ranges of conditions in the ISM, multiple gas phases can co-exist in rough pressure equilibrium. Multi-phase models of the ISM consider the various interactions occurring between the general ISM and the stars which form from the intersellar gas and subsequently heat and ionize the gas. Before the discovery of the hot ISM in the early 1970s, a two-phase model was developed by Field, Goldsmith, and Habing to explain the co-existence of a cool neutral phase with $T \sim 10^2$ K and a warm neutral phase with $T \sim 10^4$ K. This two-phase model of the ISM demonstrated that two thermally stable gas phases could exist at the same pressure, provided a thermally unstable phase occurred at an intermediate temperatures. With the discovery of the hot ISM with $T \sim 10^6$ K, the two-phase model of the ISM was extended into a three-phase model by McKee and Ostriker. In the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

three-phase model the dynamic third hot phase is created by the collective action of stellar winds and supernova explosions that shock heat large volumes of gas to high temperatures. This shock-heated high-temperature gas tries to achieve pressure equilibrium with the other gas phases and begins to cool by evaporating the matter contained in the warm and cool clouds. The hot gas can also cool by radiative emission and by adiabatic expansion into the halo, provided it can break out of the galactic plane.

In the three-phase model of the ISM the warm ionized gas can be explained as a ionized transition region between the warm neutral phases and the hot phases. The importance of this component of the gas was not fully recognized when the three-phase models of the ISM were first introduced. Although the warm ionized medium produces quite faint Hα emission over most of the sky, it radiates a very large total power of $\sim 10^{41}$ erg s^{-1}, which is equivalent to the total power of all supernovae or $\sim$15% the ionizing radiation produced by O stars.

The three-phase model of the interstellar gas has been developed into a detailed theory involving the interactions among the different gas phases, luminous hot stars, and various types of supernova explosions. The self-gravitating molecular clouds are not considered a phase in the three-phase model of the ISM since the gas they contain is not in pressure equilibrium with the other gas phases. One of the most uncertain aspects of the three-phase model of the ISM involves our knowledge of relative filling factors and the overall importance of the different phases. Estimates of the average filling factor of the hot gas in the disk of the Milky Way unfortunately currently range from $\sim$15% to $\sim$65%. It is disturbing to conclude this overview of the ISM by saying that even in 1999 we do not know which phase of the ISM fills most of the space between the stars in the disk of the Milky Way.

Bibliography

The classic text book, Spitzer L 1978 *Physical Processes in the Interstellar Medium* (New York: John Wiley) continues to be the fundamental reference on the physics of the ISM.

The text by Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Mill Valley: University Science Books) is the standard reference for the physics of interstellar gas in gaseous nebulae ionized by starlight or by collisional processes.

Physical processes occurring in the ISM are the topic of a symposium held in Grand Teton National Park, 1987 *Interstellar Processes* ed D J Hollenbach and H A Thronson Jr (Dordrecht: D Reidel)

The evolution of the ISM is discussed in the Astronomical Society of the Pacific Symposium 1990 *The Evolution of the Interstellar Medium* ed L Blitz (San Francisco: ASP Conf. Series vol 12)

Blair D Savage

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Interstellar Molecular Clouds

A molecular cloud is a condensation of interstellar gas in which the atoms are bound together as molecules rather than existing as free atoms or ionized particles. All of interstellar and intergalactic space is filled with gas, albeit of low density compared with the best vacuums achievable in terrestrial laboratories. The state of interstellar gas is largely determined by its density and temperature, and molecular clouds are the coldest, densest phase of the interstellar medium. In a molecular cloud, the gas is typically overwhelmingly molecular with less than 1% in either the atomic or ionized states.

Molecules in the interstellar medium are identified by characteristic emission (or absorption) lines in their spectra. The frequencies of these lines can usually be measured to better than one part in a million, and the probability of a misidentification when two or more lines from a single molecule are detected is vanishingly small. Molecules can rotate around or vibrate along the interatomic axes defined by the chemical bonds; torsional modes also exist. It is the transition from one rotational or vibrational state to another of individual molecules that gives rise to the spectral lines of a molecular gas. Because each molecule consists of different atoms in a unique geometry, the spectral lines from each molecule are unique. The rotational and vibrational transitions occur primarily in the infrared and radio portions of the electromagnetic spectrum; thus molecular clouds are detected primarily by infrared and radio telescopes. Most of the molecules detected in the interstellar medium to date have been identified from lines at millimeter wavelengths, the high-frequency end of the radio portion of the spectrum.

Molecular clouds are the sites of all known STAR FORMATION in the Milky Way, and observations of other galaxies imply that star formation takes place in molecular clouds everywhere in the universe. Thus the Sun and the solar system were formed in a molecular cloud long since dissipated. Molecular clouds exist throughout the Milky Way and are found in most SPIRAL GALAXIES, but are not abundant in ELLIPTICAL GALAXIES. Molecular clouds are found to very high REDSHIFTS, in one case when the universe was less than 10% of its current age.

The state of the gas

The characteristic temperatures of molecular clouds are generally quite low, ranging typically from 5 K to 300 K. Most of the gas in molecular clouds tends to be at the low end of this range; the higher end is observed only in regions where the gas is locally heated. The temperatures can be inferred from the ratios of the strengths of the spectral lines from rotational states at different energies. The lower the temperature of a cloud, the more slowly the molecules rotate on average, and the stronger are the lines from low-lying rotational energy states. A typical median temperature for a molecular cloud in the vicinity of the Sun is about 10 K.

Molecular clouds are composed overwhelmingly of molecular hydrogen gas, H_2; 99.99% of the molecules by number in a molecular cloud are hydrogen. The density within a typical molecular cloud ranges from about 10^2 cm^{-3} to about 10^6 cm^{-3}, although higher densities are achieved in regions of active star formation. The median density of the molecular gas in a typical cloud near the Sun's position in the Milky Way is about 10^3 cm^{-3}, much greater than the 0.3 cm^{-3} median density of the ambient atomic gas. At the cold temperatures characteristic of most molecular clouds, H_2 does not emit or absorb radiation; thus its density is usually inferred indirectly from the presence of trace constituents such as carbon monoxide (CO). Observations of trace molecules indicate that it is collisions with neutral particles that excite the molecules into higher rotational states, but atomic hydrogen is too underabundant to produce the observed spectral lines; only H_2 is sufficiently abundant to produce the collisional excitation. Occasionally, a bright infrared source is observed in or behind a molecular cloud and the spectral lines of molecular hydrogen can be observed directly in absorption against the background source. In these cases, the quantity of molecular hydrogen observed is consistent with that inferred indirectly from the trace molecules.

In all molecular clouds there is an admixture of dust, typically about 1% by mass. When the dust becomes opaque to the ambient ultraviolet radiation, photons energetic enough to destroy the chemical bonds between the atoms of the gas are effectively prevented from entering a cloud. The dust radiates energy efficiently in the infrared portion of the spectrum, a wavelength regime to which molecular clouds are relatively transparent except in the densest regions. Thus, when molecules collide with the dust, the energy of the collisions heats the dust particles which then emit infrared radiation that escapes the molecular cloud. The dust therefore acts as an efficient thermostat keeping the cloud cold.

The formation of molecules in the interstellar medium

At the cold temperatures typical of molecular clouds, molecular hydrogen cannot form as the result of two hydrogen atoms directly colliding and binding together, because the excess energy of the collision cannot be radiated away. This is due to the symmetrical nature of H_2 which has no permanent electric dipole moment required for the radiation to take place. Thus H_2 is thought to form when hydrogen atoms attach themselves to the surfaces of the dust particles which are always present. The atoms migrate on the surface of the dust, bind together to form a hydrogen molecule, and the excess energy goes into heating the dust particle and ejecting the molecule from the surface. The dust thus plays a crucial role in the formation of molecular clouds: it provides the site for the formation of the H_2, it prevents the destructive ultraviolet radiation from breaking the chemical bonds

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Identified interstellar and circumstellar molecules: gas-phase species. (Credit: John Black and Lew Snyder)

Diatomic	*Triatomic*	*4 atoms*	*6 atoms*	*9 atoms*
AlF	C_3	*c*-C_3H	C_5H	CH_3C_4H
AlCl	C_2H	*l*-C_3H	C_5N	CH_3CH_2CN
C_2	C_2O	C_3N	C_5S	$(CH_3)_2O$
CH	C_2S	C_3O	C_2H_4	CH_3CH_2OH
CH^+	CH_2	C_3S	CH_3CN	C_8H
CN	CO_2	C_2H_2	CH_3NC	HC_7N
CO	H_3^+	HCCN	CH_3OH	
CO^+	HCN	$HCNH^+$	CH_3SH	
CP	HNC	HNCO	HC_3NH^+	
CS	HCO	HNCS	$HCOC_2H$	*10 atoms*
H_2	HCO^+	$HOCO^+$	$HCONH_2$	$(CH_3)_2CO$
HCl	HOC^+	H_2CO	*l*-H_2C_4	CH_3C_4CN (?)
HF	HCS^+	H_2CN		
KCl	H_2O	H_2CS		
NH	H_2S	H_3O^+	*7 atoms*	
NO	HNO	NH_3	C_6H	*11 atoms*
NS	MgCN	*c*-SiC_3	CH_2CHCN	HC_9N
NaCl	MgNC		c-CH_2OCH_2	
OH	NH_2	*5 atoms*	CH_3C_2H	
PN	N_2H^+	C_5	HC_5N	
SO	N_2O	C_4H	$HCOCH_3$	*13 atoms*
SO^+	NaCN	C_4Si	NH_2CH_3	$HC_{11}N$
SiC	OCS	*l*-C_3H_2	C_7^-	
SiN	SO_2	*c*-C_3H_2		
SiO	*c*-SiC_2	CH_2CN		
SiS		CH_4		
		HC_3N	*8 atoms*	
		HC_2NC	CH_3C_3N	
		HCOOH	C_6H_2	
		H_2CHN	C_7H	
		H_2C_2O	$HCOOCH_3$	
		H_2NCN	CH_3COOH	
		H_2COH^+		
		HNC_3		
		SiH_4		

and dissociating the molecules, and it provides an effective temperature regulator that keeps the clouds cold.

Molecular hydrogen is not the only species thought to form on the surfaces of interstellar dust. The abundances of molecules that contain many heavy atoms, such as the complex organic molecules (see table 1), are too large to have formed from gas-phase attachment of carbon, nitrogen and oxygen. An important chemical pathway is thus thought to occur on the surface of dust grains and reactions that require the presence of dust as a catalyst are known collectively as grain-surface chemistry (see also INTERSTELLAR GRAINS).

Many other species, particularly molecules with only a few atoms, are thought to form directly in the gas phase without catalysis by dust. These are primarily through ion–molecule reactions, but reactions involving neutral radicals may also be important. A key ion–molecule reaction involves ionizing H_2 by COSMIC RAYS energetic enough to penetrate a molecular cloud producing H_2^+ which can subsequently bind with another hydrogen atom. The resulting H_3^+ ion is relatively abundant, highly reactive and the first step in forming many other molecules through exchange interactions with C, N and O in the gas phase. Molecules thought to form through ion–molecule reactions are CO, CH, H_2O and CH_4. Other chemical pathways result from relatively high-temperature reactions that occur in shocks, such as those produced in supernova explosions, and from high ultraviolet radiation fields such as those that occur in photon-dominated regions close to O and early B stars and relatively transparent diffuse molecular clouds.

The dust consists of molecules in the solid state and is composed primarily of atoms heavier than hydrogen, in the form of refractory (i.e. temperature-stable) and volatile compounds. The refractory compounds are primarily oxides of magnesium and silicon, and carbon compounds such as graphite, amorphous carbon (soot) and polycyclic aromatic hydrocarbons, large benzene-ring-based structures that contain admixtures of elements other than carbon. The volatile compounds on the dust surfaces are typically ices of water, ammonia, methane and other simple molecules that are composed of C, N, O and H. The dust particles are microscopic, with sizes that range up to about 1 μm (10^{-6} m) in diameter.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

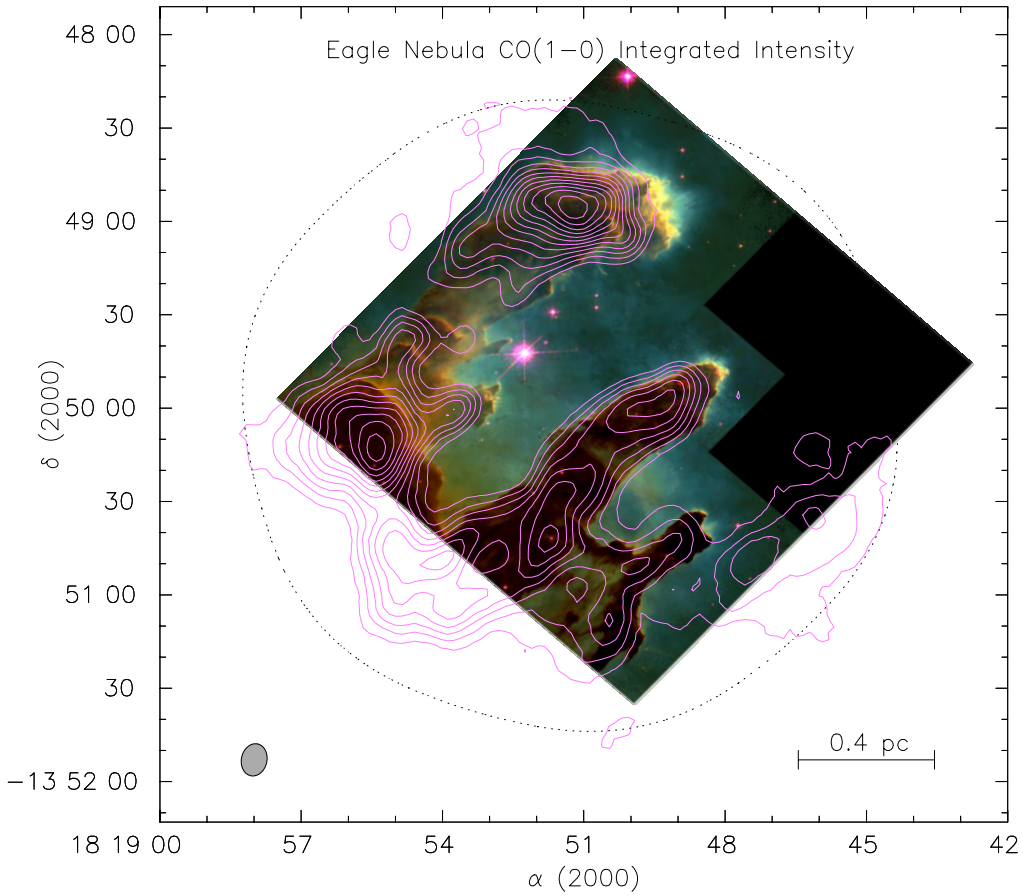


Figure 1. Image of the region of active star formation in the Eagle Nebula showing columns of dust (dark regions) made with the Hubble Space Telescope. Overlaid are contours of emission from the CO molecule observed with the BIMA Array. Note how closely the CO contours match the locations of the dust. The CO traces the more abundant but unobservable H_2 molecule. (Credit: Marc Pound.)

The composition of molecular clouds

Although molecular clouds are composed overwhelmingly of H_2, the molecule is not generally detectable for the same reason that the formation of H_2 from two gas-phase hydrogen atoms is prevented: the molecule lacks a permanent electric dipole moment. Other much less abundant molecules are, however, easily detected and the list of molecules detected in molecular clouds now totals about 120; a listing is given in table 1. These include carbon monoxide (CO), the most abundant molecule after H_2, water (H_2O) and many commonly known inorganic molecules such as ammonia (NH_3), hydrogen sulfide (H_2S), hydrogen cyanide (HCN) and common table salt (NaCl). The first organic molecule discovered in the interstellar medium was formaldehyde (H_2CO) in 1968. Since then, many organic molecules have been detected including formic acid (HCOOH), methanol (CH_3OH) and ethanol (CH_3CH_2OH), proving that the formation of organic molecules, including rather complex ones, occurs even in the relatively harsh conditions of interstellar space. Searches are now underway for glycine, the simplest of the amino acids, and buckminsterfullerene, C_{60}. Spectral lines consistent with origins in these and related molecules have already been detected, but positive identifications have not as yet been made.

Table 1 shows that the overwhelming majority of molecules are composed of H, C, N and O. This is due in part to the relatively high abundance of these elements compared with others, but also because the molecules formed from other heavy elements tend to produce refractory compounds that reside in the dust particles.

Because the density of the gas in molecular clouds is so low, highly reactive species such as radicals, molecules with unpaired electrons which have very low equilibrium abundances in terrestrial reactions, can be quite abundant in the interstellar medium. For example, OH, CH, CN are all easily detected. The radical HCO^+ was observed first in the interstellar medium and only later observed and confirmed in the laboratory. Some species such as $HC_{11}N$ have never been detected in terrestrial laboratories; their presence is inferred from the rotational constants determined from one observed transition and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the detection of other transitions at precisely predicted frequencies determined from these constants.

The abundance of CO is only 10^{-4} that of H_2, yet it produces extraordinarily bright spectral lines at wavelengths of 2.6 mm and 1.3 mm. High-energy transitions of this molecule have been detected into the far infrared portion of the spectrum from regions in which the temperature is inferred to be several hundred K. CO is a molecule that is excited at relatively modest densities and is even more tightly bound than H_2; consequently CO and H_2 are thought to be largely cospatial. The CO thus serves as a tracer for the undetectable H_2. Knowledge about the distribution of molecular clouds in the Milky Way and in other galaxies comes primarily from observations of the CO molecule (see also GAS IN GALAXIES). CO is, however, almost always optically thick, and direct inferences of the H_2 content from CO observations must be done with some care.

Masers

In most molecular clouds, the higher the energy of a rotational state, the fewer the number of molecules that occupy that state. Occasionally, conditions can arise in which there are more molecules in the higher energy state than the lower one, resulting in a condition known as a population inversion. If the cloud is transparent to radiation with an energy equal to the difference in energy between the states with a population inversion, the higher-lying molecules can be stimulated to emit radiation at a rate much higher than they would spontaneously, as they do in a laser. Such a region is known as a maser, an acronym that stands for microwave amplification by stimulated emission of radiation.

The radiation from regions within a molecular cloud that emit as masers can be extraordinarily bright; in fact the brightest known spectral lines in the radio portion of the spectrum come from interstellar masers. Many molecules have transitions that can behave as masers, but the most commonly observed ones are OH, H_2O, SiO, H_2CO and CH_3OH. The maser emission comes from regions that are typically quite small; for OH and H_2O masers, these regions are only as large as the solar system. However, because they are so bright, they can be observed to very large distances.

Masers have played an important role in astronomy because their motions make it possible to determine absolute distances with the aid of very long baseline interferometry (VLBI). For example, the only absolute determination of the distance to the center of the Milky Way has been done using statistical parallaxes of water masers in a molecular cloud, Sgr B2, thought to lie very close to the center. Similarly, the best measured absolute distance to another galaxy comes from a geometric distance measurement to the water masers in the disk of the galaxy NGC 4258. It is the extraordinary precision of VLBI positional measurements coupled with the very bright emission from the masers that make such determinations possible.

Properties of molecular clouds

Molecular clouds in the Milky Way have masses that range from a few times the mass of the Sun ($M_\odot$), to about $10^6 M_\odot$; molecular clouds with masses greater than $10^4 M_\odot$ are known as giant molecular clouds (GMCs). Along with the GLOBULAR CLUSTERS of stars, GMCs are the most massive individual objects in the Milky Way. The numbers of molecular clouds by mass can be described as a power law with an exponent of -1.5 ($\mathrm{d}N/\mathrm{d}M = N_0 M^{-1.5}$), which implies that most of the mass in molecular clouds lies at the high-mass end: more than half the total mass of the molecular gas in the Milky Way resides in clouds with masses $>10^5 M_\odot$. GMCs have typical diameters of about 50 pc, but many are quite elongated and GMCs observed in the solar vicinity often have extents as large as 100 pc.

GMCs are self-gravitating, i.e. they are held together by the mass of the molecular gas contained within them. This is inferred from the presence of star formation within GMCs, implying that at least part of a GMC must be self-gravitating, and from column density measurements that imply there is approximately enough gas to bind GMCs gravitationally. However, the best evidence that entire clouds are self-gravitating comes from measurements of the internal pressure within molecular clouds; these are invariably about 10 times the pressure of the ambient interstellar medium. If the clouds were not self-gravitating, they would rather quickly fly apart in about 10 million years.

Clouds with masses $<$ about $10^3 M_\odot$ are, however, often not self-gravitating. These clouds, known as high-latitude clouds (HLCs), were first found in abundance at large angles from the midplane of the Milky Way. They nevertheless lie at small distances from the plane as do all molecular clouds. These objects have now been found in profusion throughout the Galactic plane and, like the GMCs, have internal pressures much greater than the ambient pressure in the interstellar medium. However, unlike the GMCs, column density measurements show that the HLCs do not have enough mass to be self-gravitating, often by factors of 100 or more; and since their lifetimes must be very short their formation is a bit of a mystery. Presumably, they are formed in regions of relatively high interstellar pressure, perhaps in shocks; the ubiquity of these clouds implies that shocks are pervasive in the interstellar gas.

Molecular clouds are very inhomogeneous objects; median densities within the emitting regions are an order of magnitude larger than the density averaged over the entire volume of the cloud. Apparently, molecular clouds are organized into relatively smooth clumps of turbulent gas with higher than average density which fill only about 5–10% of the cloud volume. Stars form in the densest, most tightly gravitationally bound of these clumps, which are sometimes elongated into filaments. An alternative description suggests that the gas is organized in a fractal structure in which the distribution of density is the same regardless of the linear scale on which the cloud is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

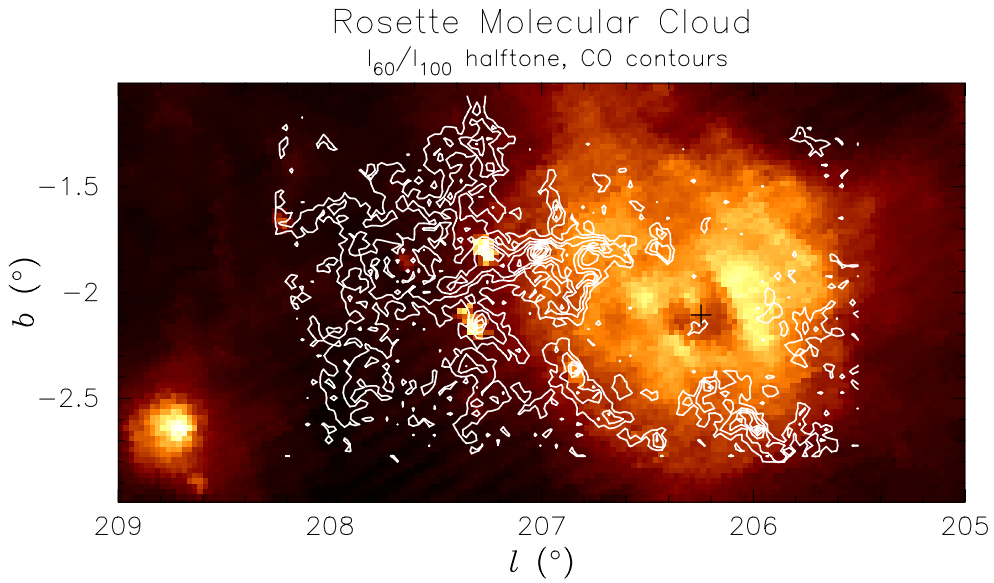


Figure 2. The Rosette Nebula molecular cloud, a good example of a Giant Molecular Cloud, is shown in the white contours, which trace the emission from CO. The infrared radiation from nebula is shown in the greyscale image as the bright region to the right of the main body of the molecular cloud. The absence of molecules toward the bright nebula is due to the dissociation of the molecules by ultraviolet radiation from hot stars in the center of the nebula (not seen in this image). (Credit: Jon Williams and Leo Blitz.)

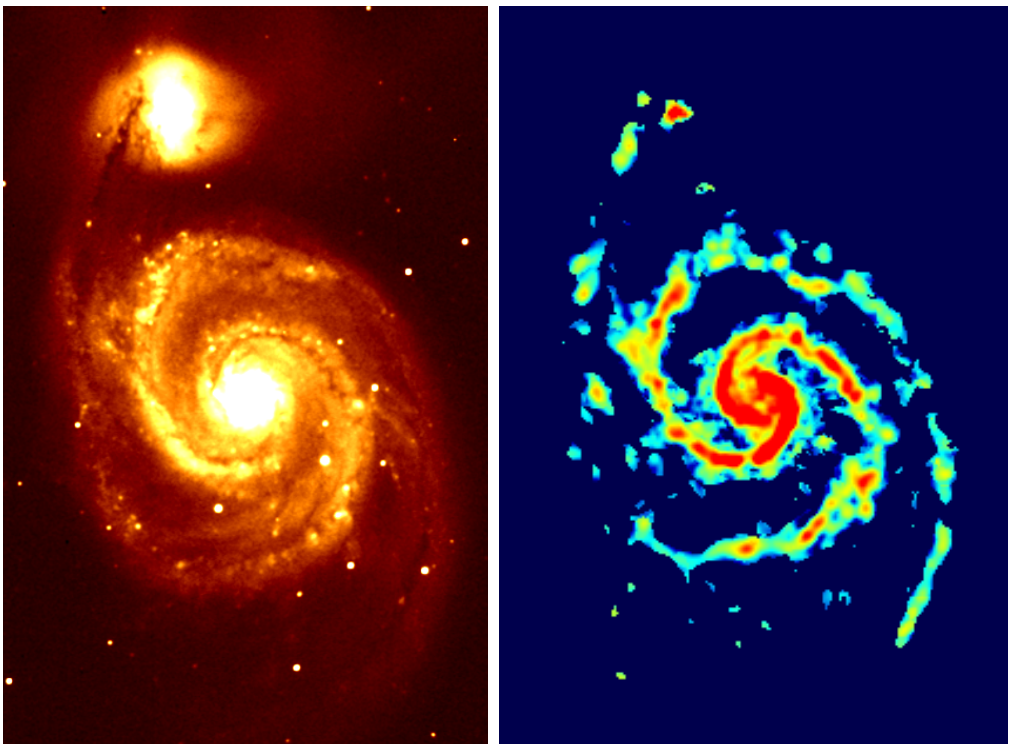


Figure 3. The left-hand panel shows an optical image of the Whirlpool Galaxy, M51, together with the small companion galaxy NGC 5195 in the upper left-hand portion of the image. The right-hand image shows the CO emission mapped with the BIMA Array from M51; the CO traces the molecular clouds in the galaxy. Note how tightly confined the CO emission is to the well-defined symmetric spiral arms in the galaxy. There is very little molecular gas between the spiral arms. (Credit: Stuart Vogel and the BIMA SONG collaboration.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

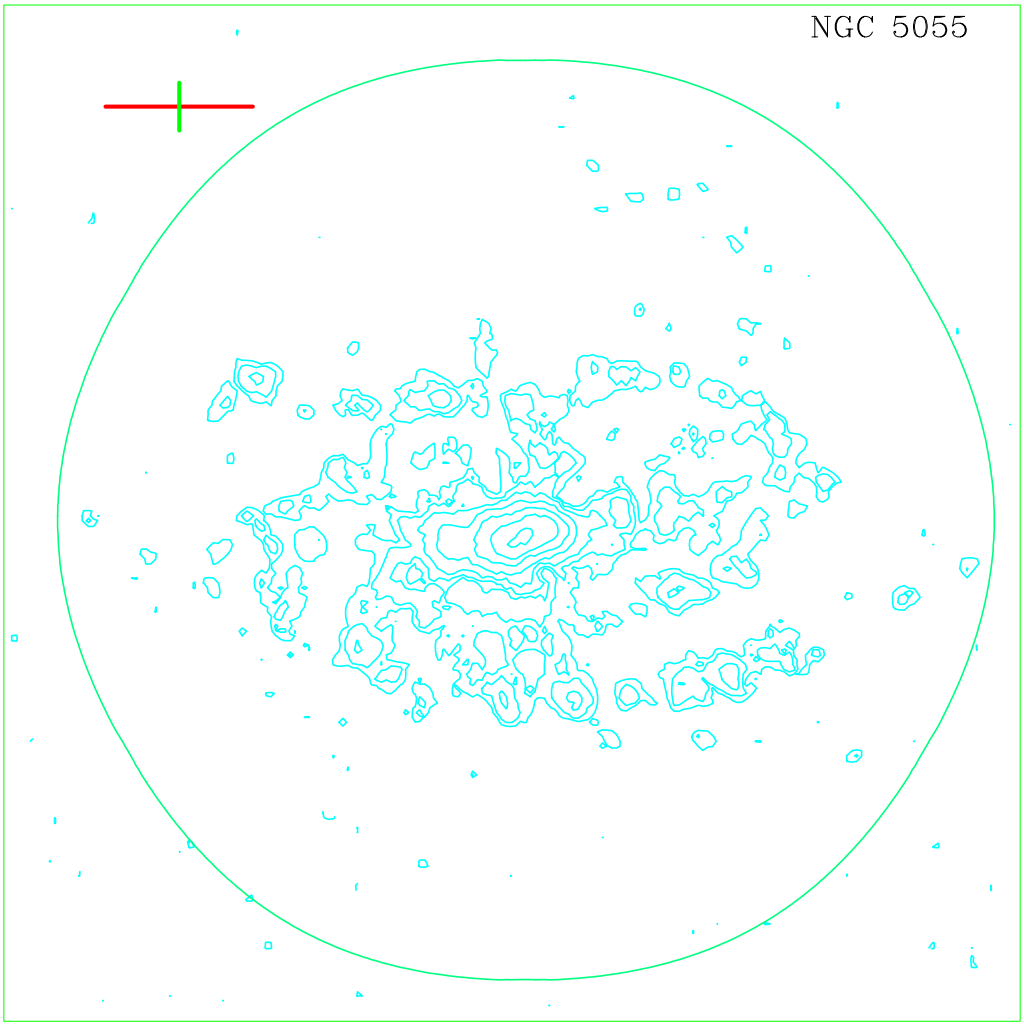


Figure 4. CO emission superimposed on an optical image of the flocculent spiral galaxy NGC 5055 mapped with the BIMA Array. The optical image shows only very short spiral arm segments. Although the CO emission shows evidence of some two-armed spiral structure, the spiral structure is much less pronounced than in M51. (Credit: Michele Thornley and the BIMA SONG Collaboration.)

observed. The fractal description appears to be more appropriate for the diffuse parts of a molecular cloud.

The formation process of GMCs is poorly understood. Although numerous mechanisms have been proposed, none has been shown to be primarily responsible for the formation of GMCs. In the disk of the Milky Way, it is thought that molecular clouds form by condensing from atomic hydrogen in regions where the atomic hydrogen has become self-gravitating and overdense; it has not been established what produces these regions of overdensity. Some other process, however, must be responsible for the formation of the small molecular clouds, the HLCs, because of their lack of self-gravity.

The molecular clouds in the nucleus of the Galaxy probably form by yet some other process. Within about 300 pc of the Galactic center, there is a strong concentration of molecular gas which has a surface density more than 100 times greater than that near the Sun. In this region, probably no more than about 1% of the neutral gas mass is atomic, and there is insufficient H I to condense into molecular clouds. The ambient gas pressure in this part of the Milky Way is almost 1000 times greater than that in the vicinity of the Sun, however, and it may be that gas clouds are stable at these pressures only if they are fully molecular.

Once formed, GMCs produce stars almost immediately. In the vicinity of the Sun only one GMC is known without intense star formation, and, in the Milky Way as a whole, no more than half of the GMCs are without massive stars forming in their midst. The star formation within a GMC sows the seeds of its own destruction because the high mass stars are intense emitters of ultraviolet radiation which dissociates the molecules, turning them back into ionized and molecular gas. The process of destruction is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

quite rapid, and GMCs are thought to be transient objects with lifetimes no greater than about (30–50)$\times 10^6$ yr. The small HLCs are thought to live for even shorter times. Another argument for the short lifetimes of molecular clouds comes from their strong confinement to SPIRAL ARMS, observed in the Milky Way and in other galaxies. The confinement implies that the clouds cannot live longer than the time it takes to traverse a spiral arm, about 50 million years. A third observation from which the youth of molecular clouds is inferred shows that clusters of stars older than about 50 million years never have molecular gas associated with them; clusters younger than about 20 million years always have associated molecular clouds.

The internal structure of molecular clouds cannot be maintained without some source of energy injection into the clouds. Like stars, molecular clouds would all collapse under their own self-gravity if they were not supported by some means. Unlike stars, this support does not come from the temperature of the gas which is too low to produce the large widths of spectral lines observed throughout molecular clouds. The large line widths suggest rather that bulk motion of the gas, such as that from turbulence, is responsible for the line widths. However, such turbulence should decay in a very short time, typically less than 10^6 yr, unless there is some sort of energy that can maintain the turbulence within a cloud.

Magnetic fields, known to be present from the polarization of the dust, from polarization in the spectral lines, and from the Zeeman splitting of the spectral lines in masers, are thought to provide the necessary support and mediate the energy injection into the clouds. However, the strengths of the magnetic fields have not yet been measured in the gas that makes up most of the mass of molecular clouds. Nevertheless, magnetic field strengths of about 10 μG are inferred, for which the measured line widths would be at or below the Alfvén velocity. The star formation in molecular clouds is accompanied by energetic outflow activity which provides enough energy to maintain the support of the molecular clouds. Furthermore, the magnetic fields within molecular clouds may be tied to the fields in the general interstellar medium, and the explosions that take place there (such as by SUPERNOVAE) may be able to transfer some of their energy to the molecular clouds along magnetic field lines.

Star formation in molecular clouds

All known star formation takes place in molecular clouds, and the overwhelming majority is known to take place in GMCs. This is not surprising since most of the molecular gas resides in GMCs, but star formation in low-mass molecular clouds also has been observed. A region of active star formation known as the Eagle Nebula is shown in figure 1; the molecular gas (traced by CO) is seen to be closely confined to the columns of dust.

All stars younger than a few million years are either embedded within molecular clouds or close by, having recently been liberated from them. The youngest stars are still embedded within their nascent molecular clouds and are optically obscured by the associated dust in the densest portions of a molecular cloud. These stars are not visible optically and can only be seen with INFRARED TELESCOPES. The most massive young stars, the O stars, form together with stars of lower mass and produce regions of highly ionized emission called H II REGIONS; the brightest of these are always found still associated with the molecular clouds from which the O stars formed. This radiation, along with ionizing winds from these O stars, can dissociate and ionize the molecular gas and is the likely cause of the dissolution of molecular clouds.

An example of a molecular cloud–H II region combination is shown in figure 2, which shows the well-known ROSETTE NEBULA and its associated GMC. In the middle of the bright nebula are five O stars that provide the ionizing radiation causing the H II region to fluoresce. The mean density of the H II region is very nearly equal to the mean density of the molecular cloud and is considerably higher than the density of the surrounding gas. The molecular cloud was, therefore, presumably much larger and was ionized by the O stars in the nebula itself. A new generation of stars is forming deep within the molecular cloud.

GMCs have substructure, with most of the mass residing in clumps of gas which are manifested as well-defined regions within a GMC that have a coherent velocity, and are separated from other clumps both spatially and in velocity. The mean density within the clumps is typically about 500–1000 cm^{-3}, and the mean density averaged over the volume of a GMC is only about 50 cm^{-3}. Thus the clumps take up only about 5–10% of the volume of a GMC. Star formation takes place only in the clumps that are the most tightly gravitationally bound and, within these, only in the densest regions in the central parts of the clump. Within the clumps are yet denser regions known as cores that give rise to individual stars or binaries. The degree to which the cores are separate entities from the clumps, and the way that the cores separate from the clumps, is a matter of current research. However, stars tend to form in groups, clusters or associations and, by implication, there must be many cores within a typical clump.

Cores have typical densities of $\sim 10^5$–10^6 cm^{-3} and can be observed in a variety of molecules. Some cores already have stars or protostars (stars still in the process of formation) within them, but some appear to be in an earlier stage of formation and are starless. The cores that form low-mass stars (stars with masses comparable with the Sun) have masses comparable with what would be produced by a Jeans instability for gas at about 10 K. For such an instability to produce a star, however, the core must lose its magnetic support. This is thought to take place by ambipolar diffusion, diffusion of the magnetic field along the field lines, but other processes, such as MAGNETIC RECONNECTION, may also play a role. Once collapse begins, it is thought to take place from the inside out, with the central regions collapsing first. Material with low angular momentum falls onto the central star, and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

material with a larger angular momentum is thought to form a disk. Observational evidence for disks around low-mass stars is quite conclusive. The disks are the material from which planets similar to our solar system form. The formation of a low-mass protostar also generally gives rise to jets which can sometimes be observed in the optical and infrared portions of the spectrum and resemble the jets seen in quasars and radio galaxies, but with much lower energies. These jets can accelerate the ambient molecular material forming bipolar outflows which are very common in regions of active star formation. These outflows may be the source of much of the support of molecular clouds.

High-mass star formation is likely to proceed differently because, once the highest-mass stars form, they rapidly produce regions of ionized gas known as H II regions, which are energetic enough to destroy a molecular cloud. Thus it is generally supposed that the highest-mass stars appear only after the low-mass stars have substantially formed. The cores that form high-mass stars have not yet been conclusively identified, but this is likely to be the result of observational bias. High-mass star formation is thought to take place very rapidly, and once a high-mass protostar has formed it almost immediately affects the region in its vicinity. It therefore becomes difficult to disentangle the properties of pre-existing cores that form high-mass stars from the effects of a previously formed star.

Molecular clouds in galaxies

Molecules have been observed in galaxies beyond the Milky Way since 1975. The great distance of these galaxies makes it difficult to resolve individual molecular clouds in any but the nearest galaxies: the MAGELLANIC CLOUDS, M31 and M33. Nevertheless, much has been learned about molecular clouds in other galaxies, largely from observations of CO and a few other molecules such as HCN, CS, H_2O and OH.

Spiral galaxies are known to be much more abundant in molecular gas than elliptical galaxies; only a handful of elliptical galaxies are known to have molecular gas. Elliptical galaxies are known to be gas poor in general, so the paucity of molecular gas is not surprising. IRREGULAR GALAXIES, although generally gas rich, generally have relatively weak CO lines, but, since these galaxies usually have a relatively low abundance of heavy elements, the paucity of CO may not be indicative of a relative absence of H_2.

Spiral galaxies exhibit a plethora of morphologies in the molecular gas, with CO sometimes clearly outlining the spiral arms, as is seen in figure 3 for the galaxy M51, and sometimes no spiral arms are evident as is seen for NGC 5055 in figure 4. The surface density of molecular gas is sometimes seen to increase right into the nucleus, but sometimes the nucleus is devoid of molecular gas as is the case in M31. Galaxies which appear to be morphologically similar in the optical portion of the spectrum may have very different morphologies in the molecular gas. These differences have only recently been recognized because the resolution and sensitivity necessary to make high-resolution images of the molecular gas have only recently been achieved through millimeter-wave interferometry.

By analogy with the Milky Way, stars in other galaxies are thought to form where the molecular gas is bright and therefore abundant. This is confirmed by the presence of H II regions and strong infrared radiation from embedded stars, both of which trace recent star formation, coincident with the molecular clouds in the galaxies. Although there are large variations of the star formation rate in the molecular gas both from galaxy to galaxy and even within a galaxy, there is generally a good correlation of indicators of star formation with molecular gas.

In the nuclei of some galaxies known as STARBURST GALAXIES, the rate of star formation is 10–100 times the rate of star formation in a normal galaxy. All starbursts are rich in molecular gas and have large quantities of molecules such as HCN, CS and HCO^+, all of which are indicators of relatively dense gas. Evidently, starbursts occur when extraordinarily large quantities of molecular gas are transported into the nucleus of a galaxy, which then becomes unstable to rapid star formation. In general, it is only possible to trace the formation of high-mass stars in these systems and the total star formation rate is generally inferred by assuming that the mix of stars that is formed is similar to that in the Milky Way. Starburst galaxies, and their spectacular cousins the ultraluminous galaxies, are usually the product of interactions and mergers of two galaxies. Simulations suggest that the molecular gas plunges into the center in the process of merging, which may explain how such a large quantity of gas is brought to the nucleus so quickly (se GALAXIES: INTERACTIONS AND MERGERS).

Galaxies also sometimes have 'active' nuclei, in which explosive events are thought to be energized by a BLACK HOLE. That is, in falling into a black hole, energy is liberated which can strongly heat gas to produce very high expansion velocities, highly excited atoms and jets of material expelled from the nucleus. One class of these galaxies is known as SEYFERT GALAXIES, after their discoverer, Carl Seyfert, the nuclei of which are thought to be surrounded by tori of molecular gas which provide the fuel for the nuclear activity. These tori, although widely expected, have not yet been observed because the resolution of the observations has not yet been available. The molecular clouds in the circumnuclear tori are probably quite different from the molecular clouds observed in the disk of the Milky Way.

Molecular gas has also been observed toward the edge of the observable universe in some of the highest-redshift galaxies yet observed. One galaxy with a redshift of 4.7 is known to contain vast amounts of CO at a time that the universe was less than 10% of its current age. A number of galaxies with redshifts between 1 and 4.7 have been detected and there is a great deal of interest in both finding more GALAXIES AT HIGH REDSHIFT and observing them in detail. The detected galaxies exhibit huge bursts of star formation and can tell us a great deal about the early formation of stars and the evolution of galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Blitz L 1993 Giant molecular clouds *Protostars and Planets III* ed Levy and Lunine (Tuscon, AZ: University of Arizona Press) p 125

Blitz L *et al* 1993 The centre of the Milky Way *Nature* **361** 417

Langer W D *et al* 2000 Chemical evolution of protostellar matter *Protostars and Planets IV* ed Mannings, Boss and Russell (Tuscon, AZ: University of Arizona Press) p 97

McKee C F 1999 The dynamical structure and evolution of giant molecular clouds *The Origin of Stars and Planetary Systems* ed Lada and Kylafis (Dordrecht: Kluwer) p 29

van Dishoek E F and Hogerheijde M R 1999 Models and observations of the chemistry near young stellar objects *The Origin of Stars and Planetary Systems* ed Lada and Kylafis (Dordrecht: Kluwer) p 97

Williams J P, Blitz L and McKee C F 2000 The structure and evolution of molecular clouds: from clumps to cores to the IMF *Protostars and Planets IV* ed Mannings, Boss and Russell (Tuscon, AZ: University of Arizona Press) p 97

Leo Blitz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Intracluster Medium

With the expression 'intracluster medium' (ICM) one indicates the material other than ordinary galaxies present in clusters of galaxies. The presence of the ICM is inferred from direct observations and indirect evidence. In 1951, a famous paper by F ZWICKY on the morphology of the COMA CLUSTER of galaxies in the northern hemisphere reported his discovery of a diffuse light in the region between galaxies belonging to the Coma cluster: '*. . . vast and often very irregular swarms of stars and other matter exists in the spaces between the conventional spiral, elliptical, and irregular galaxies*'. Following Zwicky's initial discovery of the presence of an intragalactic matter made up from stars, in 1966 a source of x-rays was detected in the VIRGO CLUSTER and was followed by detections in the Perseus cluster and in the Coma cluster itself in 1971. Since then, several x-ray satellites have observed nearly hundreds of clusters showing that indeed they are bright sources of x-rays. The observations in the optical wavelengths and the x-ray detections indicate that the clusters of galaxies contain an ICM and this material is present in two phases: a diffuse stellar component, irregularly distributed, and a hot tenuous gas that fills the cluster more or less homogeneously. Indirect evidence for the presence of an unseen component in clusters of galaxies comes from the application of the virial theorem to the observed velocity dispersion of galaxies in clusters: for the cluster to be in equilibrium an additional component other than galaxies, the hot gas, and diffuse star light, must be present to keep the clusters bound.

Properties and origin of the diffuse light

The first observations of the diffuse light associated with the ICM in clusters of galaxies were based on photographic techniques, followed recently by observations with CHARGE-COUPLED DEVICES (CCDs). Several difficulties affected these earlier studies. Photographic plates do not have an homogeneous response through them (and this behavior cannot be accounted for during the data analysis as is done for images acquired with CCDs) and typically the surface brightness of the intracluster light is less than 1% of the night sky brightness. In addition there is scattered light from bright objects, the contribution from discrete sources—dwarf or irregular galaxies—below the detection limit, and clusters extend over large areas on the sky: nearby clusters such as Virgo or FORNAX cover an area of several square degrees (the image of the Moon or the Sun covers half a square degree on the sky) while linear detectors such as CCDs cover a much smaller area (of the order of hundreds of square arcminutes) and mosaicing is required. Furthermore, at the position of cluster centers one normally finds a giant elliptical galaxy with an extended halo of stars: these objects are classified as cDs, i.e. cluster dominant galaxies. In these cases it is difficult to disentangle the diffuse light associated with the halo of the cD galaxy and the ICM diffuse stellar component. The best estimate based on these observations for the diffuse light component of the ICM is that its contribution is 10–50% of the total cluster light in galaxies, in a region between 200 h^{-1} kpc and 450 h^{-1} kpc from the central cD galaxy (here the HUBBLE CONSTANT is written as 100 h^{-1} km s^{-1} Mpc).

An alternative method for probing intracluster star light in the ICM is through the direct detection and measurements of the stars themselves. In the 1970s a line of attack to track down the debris making up the diffuse light in the ICM was to look for intergalactic supernovae, but these studies reported a negative answer for the Coma cluster, while an intergalactic supernova was detected in 1981 in the Virgo cluster. In the 1980s another suggestion was that the cD galaxies in clusters might be surrounded by a population of globular clusters stripped from their parent galaxies: there are now suggestions that the pronounced globular cluster overabundance of some galaxies at the centers of galaxy clusters is best explained by assuming the existence of a population of intergalactic globular clusters.

Very recent observations have shown that direct detection of individual stars in the ICM is possible. Studies of the ICM in the Virgo and Fornax clusters have detected several hundreds of point-like emissions from individual stars in their post-AGB phase of PLANETARY NEBULAE. The planetary nebulae are very easy to detect because their nebular outer shell encircling the central star emits nearly 15% of the total energy from the central star in the [O III] emission line at 5007 Å: by simply taking an image through a narrow-band filter centered at the planetary nebulae redshifted [O III] emission and in its adiacent continuum, it has been possible to identify nearly a hundred planetary nebulae emissions in nearby clusters. Simulations based on the overall number of detected planetary nebulae emissions and the properties of their luminosity function in early-type galaxies indicate that the diffuse stellar component in the ICM can contribute up to 50% of the total light emitted by galaxies in the Virgo and Fornax clusters.

Independent evidence for the presence of stars in the ICM comes from observations carried out with the HUBBLE SPACE TELESCOPE. These observations have detected red-giant-branch stars (stars nearly 13 Gyr old) in a Virgo cluster field centered 235 kpc east of the galaxy NGC 4486, placed at the center of this cluster. The excess star number counts in this surveyed field with respect to a control field and calibrations based on well-known luminosity profiles of elliptical galaxies (which have a population with similar age and color to the intergalactic stars detected with the Hubble Space Telescope) indicate that the diffuse light in the ICM contributes nearly 10% of the stellar mass in Virgo cluster galaxies.

The discrepancy between the detections based on the [O III] line emission from planetary nebulae, observed in the region of the Virgo cluster core, and the red number count excess may be related to a strong inhomogeneity in the distribution of the stellar component of the ICM. The detailed spatial distribution of this diffuse stellar component is not known yet on the Mpc scale (typical radius for a cluster of galaxies is nearly 3 Mpc), and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

its distribution may be related to the origin of this ICM component itself. Tidal interactions between galaxies in a cluster are expected to be frequent and cause galaxies to lose a substantial fraction, 30–70%, of their stars to the cluster potential where they become the stellar component of the ICM. Because this diffuse component would then originate from tidal tails, one expects its spatial distribution to be quite inhomogeneous. The debris can be relaxed in the cluster core but the further out one goes, the more unrelaxed the population might become.

Large CCDs, mosaiced detectors and observations with the Hubble Space Telescope will soon address these questions and give more information on the origin of this component of the ICM. The discovery of single stars in the ICM gives an opportunity to detect the local cluster mass-over-light ratio at the positions where they are identified, and have strong implications on the baryonic content of clusters and the origin of the hot gas phase in the ICM.

Properties and origin of the hot gas

The x-ray hot gas is the main '*visible*' component of clusters of galaxies: its total mass varies from 1 (*groups*) to 7 (*rich clusters*) times the stellar mass present in cluster galaxies and it can reach up to 30% of the total cluster mass—visible plus dark matter. The gas mass in the ICM is also highly correlated with the luminosity from ellipticals and lenticulars present in the cluster. The total mass in the hot-gas component of the ICM ranges from 2×10^{13} to 5.2×10^{14} solar masses within a 3 Mpc radius: this corresponds very roughly to a gas density of 10^{-3} atoms cm^{-3}. The x-ray emission from galaxy clusters is generally interpreted as due to thermal bremsstrahlung in hot gas (gas temperature within 10^7–10^8 K): the observations of the temperature profiles of the hot gas are consistent with the gas itself being in hydrostatic equilibrium, i.e. the gas pressure is balanced by the gravitational force generated by the mass distribution in the cluster. Poor clusters have low central density, low gas mass and low temperature, whereas rich clusters have high central density, high gas mass and high temperature.

The hot emitting gas is enriched with heavy elements: x-ray satellite observations done in the late 1970s revealed the presence of heavy-element emission lines in the x-ray emission associated with clusters. The inferred iron abundance of the ICM seems to have an universal value, i.e. about 0.35 of the solar iron abundance. The detection of the iron lines in the x-ray spectra has shown that the gas is enriched in metals and has been processed, at least in part, into the stellar content of cluster galaxies. The general view is that the heavy elements in the ICM come from galaxies (mainly ellipticals and lenticulars) which eject gas through gas winds which are powered by supernova explosions. In fact the iron masses detected in the ICM are correlated with the luminosity of cluster early-type galaxies. It is interesting to compare the iron yield needed to reproduce the typical observed iron mass over stellar mass ratio in clusters. For a typical cluster, the value turns out to be 5×10^{-3} which is 5 times larger than that measured for the Milky Way galaxy: the conclusion is that one cannot account for the iron mass in the ICM using standard iron mass yield, i.e. that of our Galaxy. Astronomers are led to believe that there was a high supernova activity in cluster galaxies in the past, either a much higher past average SNI rate as compared with the present value observed in ellipticals (at least a factor 10) or a high SNII rate (as generated for istance by a flat initial mass function).

Chemical evolution models, which makes use of bimodal star formation or a higher SNII rate, are able to predict the iron masses (typically of the order of 10^{12} solar masses) in the ICM, but the total ejected masses of gas are always an order of magnitude smaller than the observed ones. This would imply that the bulk of gas observed in the hot x-ray emitting component of the ICM has a primordial origin, i.e. it is pristine material produced during the big bang and was never used to form stars. The agreement between the prediction of the chemical evolution model for early-type galaxies and the observed heavy element content of the ICM is based on the assumption that there had been a complete mixing of the galactic heavy-element-enriched gas with the unprocessed ICM gas on a time scale shorter than the age of the universe.

The recent discovery of massive compact halo objects (MACHOs) and the estimate of their total mass fraction in our galaxy halo may give an alternative explanation for the origin of the hot gas in the ICM. The latest observations indicate that MACHOs may constitute 30–80% of the total galactic halo mass, and, if one assumes that they are common components of all cluster galaxy halos and they are the remnants of an early population of stars that also lost 30% of their initial mass as gas to the ICM, then this mass loss from the stellar progenitors of MACHOs is potentially a very significant source of gas in groups and clusters (see also WIMPS AND MACHOS).

Distribution of the hot diffuse gas

The hot diffuse component of the ICM peaks at the center of the cluster and falls off with distance from the center, and the x-ray emission can be traced out to a distance of 3 Mpc. Recent observations carried out with the latest generation of x-ray satellites indicate that the hot gas in the ICM traces the hierarchical distribution of dark matter in clusters. Recent observations for the nearby Fornax cluster produced a gas density map and the temperature profile of the hot gas over an area of 0.5 square degrees, centered on the galaxy NGC 1399. Using this gas density map and the temperature profile in the hydrostatic equilibrium equation, the mass distribution is computed and the result shows the presence of a more concentrated dark matter distribution associated with the galaxy and a shallower mass distribution at larger radii associated with the cluster. Similar evidence is also found for the Virgo cluster. Such observations are crucial to establish the clustering properties of the dark matter, which are strongly dependent on its intrinsic nature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Additional components of the ICM

The sum of the masses in cluster galaxies, stellar diffuse plus hot gas ICM components in clusters, can account for up to 30% of the gravitational mass needed to keep any clusters bound. Therefore an additional '*unseen*' dark component is present in the ICM. Possible candidates for dark matter constituents are MACHO-like objects and remnants of massive stars, such as black holes, neutron stars or cool white dwarfs, but, particularly in the case of the dark component of the ICM, elementary particles or non-baryonic matter are important as dark matter candidates. In clusters of galaxies the ratio between the so-called 'baryonic mass' and total mass is near the limit imposed by nucleosynthesis constraints, based on the relative abundances of light elements such as deuterium, helium and lithium observed in our universe, and a cosmological model with $\Omega_0 = 0.2$. Therefore any additional dark components in clusters must be in some sort of non-baryonic form.

Future developments

The new x-ray observations made with the x-ray ASCA satellite are indeed giving astronomers new insight into the properties of the hot-gas component of the ICM; the determination of the number counts of the stellar component of the ICM via large-format CCD images and observations with the new advanced camera on the Hubble Space Telescope will give the welcome opportunity to determine the local mass-over-light ratio of the ICM itself. These two complementary lines of work will help in understanding the mechanism which led to the formation of clusters, and therefore to galaxy and star formation within the cluster itself. The use of the most accurate temperature profiles and gas density distribution maps at high resolution will represent a powerful tool to determine the accurate cluster mass spectrum which is crucial for its implications for cosmological models and the determination of the average mass density of the universe.

Bibliography

Arnaboldi M *et al* 1996 The kinematic of planetary nebulae in the outer regions of NGC 4406 *Astrophys. J.* **472** 145

Arnaud M *et al* 1992 Some constraints on the origin of the iron enriched intra-cluster medium *Astron. Astrophys.* **254** 49

Carlberg R G *et al* 1996 Galaxy cluster virial masses and Ω_0 *Astrophys. J.* **462** 32

David L P 1997 The baryonic content of galaxies and clusters *Astrophys. J.* **484** L11

Feldmeier J J, Ciardullo R and Jacoby G H 1998 Intracluster planetary nebulae in the Virgo cluster. I. Initial results *Astrophys. J.* **503** 109

Matteucci F and Vettolani G 1988 Chemical abundances in galaxy clusters: a theoretical approach *Astron. Astrophys.* **202** 21

Moore B *et al* 1996 Galaxy harrassment and the evolution of cluster galaxies *Nature* **379** 613

Sarazin C L 1988 *X-ray Emission from Cluster of Galaxies* (Cambridge: Cambridge University Press)

White S D M, Navarro J F, Evrard A E and Frenk C S 1993 The baryon content of galaxy clusters: a challange to cosmologiocal orthodoxy *Nature* **366** 429

Magda Arnaboldi

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Io: Plasma Torus

The discovery of the four largest moons of Jupiter (Io, EUROPA, GANYMEDE and CALLISTO) by GALILEO GALILEI in 1610 was crucial to the rejection of the then prevalent theory that the Earth was at the center of the universe. In the following 350 years, several additional small satellites of Jupiter were discovered (see JUPITER: SATELLITES) but relatively little was learnt about the physical properties of what are now called the Galilean satellites and nothing was known of their interaction with Jupiter and its magnetic field. A turning point occurred in 1964 when E K Bigg showed that radio emissions from Jupiter at decametric wavelengths (previously discovered by Franklin and Burke in 1950) were modulated by the orbital position of the innermost Galilean satellite, Io. The presence of radio emissions indicated that Jupiter had a significant magnetic field and the modulation by Io suggested a complex interaction between the magnetic field and the Galilean satellite (see JUPITER: MAGNETOSPHERE, MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS).

The next important discovery was made by R A Brown in 1972 who detected optical emission from a neutral sodium gas cloud surrounding Io and accompanying it in its orbit about Jupiter. Sodium atoms have two strong lines at visible wavelengths (589.0 nm and 589.6 nm) which are excited by a resonance scattering mechanism. The Sun provides the light source in this case and the sodium neutrals scatter any light at the emission wavelengths they receive in their rest frame. Although the light source comes from only one direction (the Sun), the emission is essentially isotropic. At the rest wavelengths of neutral sodium emission, the luminosity of the Sun is actually rather low because of deep FRAUNHOFER LINES. However, Io has an orbital velocity of 17.4 km s^{-1} about Jupiter. The neutral sodium cloud accompanying Io in its orbit therefore sees the solar illumination Doppler shifted by up to about 0.02 nm (the Doppler shift being greatest when Io is at its greatest elongation from Jupiter as seen from the Sun). This is sufficient to provide a 10-fold modulation of the brightness of the sodium gas cloud surrounding Io depending upon the orbital phase of the satellite.

The neutral sodium cloud is a further product of the interaction between Io and Jupiter's magnetic field. Io orbits at a radial distance of 5.91 Jupiter radii (R_J) where 1 R_J is equivalent to 7.138×10^4 km. Jupiter's magnetic field, which defines the rotation rate of Jupiter, has an angular velocity of 870.536° day^{-1}. The magnetic field at Io's orbit is completely dominated by Jupiter (i.e. there is no influence on the magnetic field strength from the solar wind interaction—this only becomes important at distances greater than about 20 R_J from the planet). Io itself may have a small intrinsic magnetic field but this remains unproven. The magnetic field lines of Jupiter co-rotate with the rotation of the planet. Therefore, at Io's orbit, the field lines have a velocity of 74.2 km s^{-1} and have a velocity with respect to Io of 56.8 km s^{-1}. Ions produced in Jupiter's magnetic field are trapped and co-rotate with the magnetic field (this is a consequence of Faraday's law). Hence, magnetospheric ions have a high relative velocity with respect to Io. Ions which impact the surface and/or atmosphere of the satellite have considerable energy which is used to eject fresh neutrals from Io. This sputtering process, with some minor modifications, is still thought to be the principal mechanism by which material is removed from Io 25 years after it was first suggested by Matson *et al*. Recently, strong evidence has been presented suggesting that this sputtering process might be supplemented by direct removal of ions (produced by photoionization of SO_2 for example) from Io's atmosphere by the co-rotating magnetic field.

The flyby of Io by the VOYAGER 1 spacecraft in 1979 provided a huge leap in our knowledge of the Jupiter system. The main result was the discovery of active volcanism (see IO: VOLCANISM AND GEOPHYSICS) which had been predicted by Peale, Cassen and Reynolds in an article in *Nature* published just five days before the flyby. Volcanism is a possible consequence of tidal forces acting on Io produced by an orbital resonance between Io, Europa and Ganymede. The surface of Io was observed to be almost craterless, indicating rapid resurfacing of the satellite resulting from the volcanism. The estimated resurfacing rate is around 0.1 mm per year. The remarkable colors on the surface indicated a predominance of sulfur and the satellite was shown to have a tenuous SO_2 atmosphere. In addition, the plasma in Jupiter's magnetosphere near the orbit of Io was shown to be dominated by sulfur and oxygen ions. The latter provided remarkable confirmation of the ground-based detection of emission from S^+ ions in Jupiter's magnetosphere made by Kupo *et al* in 1975. This observation was controversial at the time because it appeared to contradict measurements made by the PIONEER 10 spacecraft which seemed to indicate a dense plasma dominated by protons. Ground-based observers prior to the Voyager 1 flyby consistently measured S^+ emission appearing to come from a ring about Jupiter centered roughly on the orbit of Io. Electron densities could be determined by measuring the ratio of an S^+ emission doublet at (671.6 nm and 673.1 nm) produced by collisional excitation. Values between 1000 and 10 000 cm^{-3} were consistently recorded—much lower than densities derived from Pioneer 10 measurements. Just prior to the Voyager 1 flyby, Pilcher and Morgan detected emission from O^+ through its collisionally excited doublet at a wavelength of 372.7 nm.

The Voyager results completely vindicated the results from ground-based observations and gave a clear picture of the plasma environment near Io's orbit. Plasma in Jupiter's magnetosphere sputters material

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

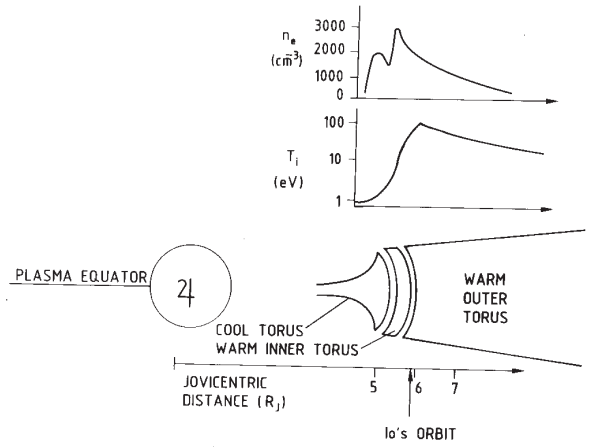


Figure 1. Schematic diagram showing a section through the torus. The 'plasma equator' indicates the equilibrium position for ions in the torus. This plane is tilted by about 7° with respect to the rotation equator. The variations of ion density and temperature in the torus with Jovicentric distance are shown in the form of plots. The density peaks just inside Io's orbit in a region referred to as the 'warm inner torus' (or sometimes 'the ribbon'). The temperature peaks just outside Io's orbit and decreases radially outwards as the plasma expands into the middle magnetosphere. Inside the warm inner torus, the plasma is cold. The regions have been widely separated in the diagram for emphasis.

from the surface and/or atmosphere of Io, producing a series of neutral clouds. Neutral sodium is the most easily detected from ground-based observation because it is a remarkably efficient scatterer of sunlight. However, it is only a trace element, making up around 5% of the neutral clouds. Sulfur and oxygen neutrals actually dominate the composition. Oxygen atoms have been detected from ground-based observations at 630.0 nm. Electrons in Jupiter's magnetosphere interact with the neutrals to produce ions by electron impact ionization. The ions are then picked up by Jupiter's co-rotating magnetic field. The ions are thus swept into a ring about Jupiter centered roughly on Io's orbit. The ions diffuse both inwards and outwards from this ring to fill a torus-shaped volume in space which is called the Io plasma torus or 'IPT'.

The density, temperature and composition of the IPT vary strongly with the radial distance from Jupiter. Figure 1 shows a schematic diagram of a section through the torus.

Inside Io's orbit the plasma is cool and the composition is dominated by singly ionized species (mostly S^+ and O^+). The electron and ion temperatures, although not in equilibrium, rise rapidly between 5.4 R_J and 6.0 R_J. The higher-energy electrons give rise to a more highly ionized plasma and S^{2+} becomes the most numerous sulfur ion around 5.8 R_J. The electron temperature continues to rise outside Io's orbit and more highly ionized species are seen as one moves towards the middle magnetosphere of Jupiter. The ion temperature, however, decreases due in part to the expansion of the plasma into the middle magnetosphere. The electron density reaches a maximum of around 3000 cm^{-3} at around 5.7 R_J—just inside the orbit of Io. The estimated composition at 5.9 R_J (at Io's orbit) is estimated to be 14% S^+, 48% O^+, 19% S^{2+}, 2% O^{2+}, 2% S^{3+}, 2% Na^+, <2% SO^+_2. A 10% mixing ratio of protons is usually assumed but this is uncertain. The estimated mass of the IPT is around 2–10^6 tonne.

The Voyager spacecraft provided only a snapshot of the conditions in the IPT during its flyby. Observations of the IPT using ground-based telescopes (e.g. the European Southern Observatory) and Earth-orbiting observatories (e.g. Hubble Space Telescope, International Ultraviolet Explorer, Extreme Ultraviolet Explorer, Infrared Space Observatory) have become the main tools in the investigation of the IPT since they provide both a synoptic view of the IPT and can be used to investigate the time variability of the IPT. All observable emissions are produced by collisional excitation and therefore the observed intensity in any emission line is proportional to the product of the electron density and the ion density integrated along the line of sight through the toroidal volume. Because Earth-based observers view the torus

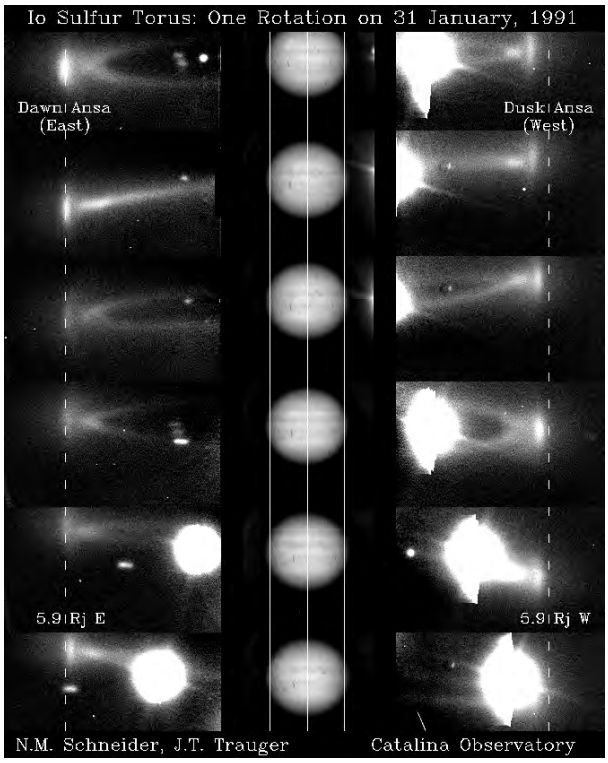


Figure 2. A time-series of ground-based observations of the Io plasma torus. Jupiter is seen in the center of the frame. Jupiter rotates in 9.925 h. The Io torus oscillates over this time scale because the magnetic axis of Jupiter is tilted by 9.8° with respect to its spin axis. The ionized sulfur emission seen here also shows an east (left)–west (right) brightness asymmetry. Image courtesy of N M Schneider (University of Colorado) and J T Trauger (JPL).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

roughly edge-on, the maximum emission is seen at the ansae (outer edges) of the IPT around 5.7 R_J from Jupiter in projection (see figure 2) and decreases rapidly as the radial distance decreases. Most of our knowledge stems from observations of the ansae of the IPT over a rather limited radial distance from Jupiter (roughly 5.0–6.5 R_J). The integrated emission from the IPT is around 4×10^{12} W with the strongest emissions being in the extreme ultraviolet. The 68.5 nm emission of S^{2+} is responsible for 5% of the emission from the IPT alone.

Figure 2 shows how the IPT appears to wobble in a Jupiter-centered reference frame. Jupiter's magnetic dipole is tilted by 9.8° with respect to the rotation axis. The equilibrium position for ions in a dipole magnetic field is the magnetic equator (where the magnetic mirror force becomes zero). However, Jupiter is rotating with high angular velocity and thus the centrifugal force on the ions has to be taken into account. As a result, the equilibrium position for the ions is at the so-called centrifugal equator which is between the rotational equator and the magnetic equator. The centrifugal equator is tilted by about 7° with respect to the rotational equator. In a reference frame connected to Jupiter's rotation axis, the IPT therefore appears to wobble with the ansae oscillating up and down with an amplitude of about 1.4 R_J (peak to peak).

Another feature of figure 2 is that the IPT on the west of Jupiter (right) appears to be brighter than on the east (left). This feature was first recognized in Voyager data and is believed to be caused by the existence of an electric field which shifts the IPT slightly towards the east as seen from the Earth (Barbosa and Kivelson 1983). IPT ions in the east are therefore slightly further away from Jupiter than their counterparts in the west. Being further away in a dipole system, the ions are less confined and therefore less dense. Collisional excitation is strongly proportional to density and thus the east appears slightly fainter.

The IPT is highly variable on several different time scales. A variability is seen in the brightness of the torus with a period equal to Jupiter's rotation period (System III). The brightness modulation is anti-correlated with the ion temperature (Schneider and Trauger 1995). The mechanism producing this phenomenon is at present unknown. A second variation with a period roughly 3% longer than System III (sometimes referred to as System IV) is also evident and may be the result of long-lived density waves in the IPT (Roesler *et al* 1984). The torus is also variable on long time scales and can vary in brightness from year to year by up to a factor of 4. There is also some evidence that the flow past Io modifies the plasma parameters so that the brightness downstream of Io is slightly enhanced.

New observations of Jupiter, its satellites and environs have been made by Cassini, Galileo, the Hubble Space Telescope, the Chandra X-ray Observatory and Earth-based radio and optical telescopes during the CASSINI spacecraft's flyby. The GALILEO spacecraft has been in orbit around Jupiter since December 1995 while Cassini observed Jupiter from 1 October 2000 to 31 March 31 2001. Cassini made its closest approach on 30 December 2000, at a distance of about 9.7 million kilometers (6 million miles). Data from the flyby observations show that Io leaves a magnetic footprint on Jupiter's upper atmosphere which appears as a spot of ultraviolet emission that remains fixed underneath Io as Jupiter rotates. The specific physical mechanisms responsible for generating these emissions are not well understood, but in general the spot seems to arise because of an electromagnetic interaction between Jupiter's magnetic field and the plasma surrounding Io, driving currents of around 1 million amperes down through Jupiter's ionosphere. Io's magnetic footprint extends well beyond the immediate vicinity of Io's flux–tube interaction with Jupiter, and much farther than predicted theoretically; the emission persists for several hours downstream.

The Cassini spacecraft has also discovered a fast (greater than 10^3 kms^{-1}) and hot magnetospheric neutral wind extending more than 0.5 au from Jupiter, and the presence of energetic neutral atoms (both hot and cold) that have been accelerated by the electric field in the solar wind. It is suggested that these atoms originate in volcanic gases from Io, undergo significant evolution through various electromagnetic interactions, escape Jupiter's magnetosphere and then populate the environment around the planet. Thus a 'nebula' is created that extends outwards over hundreds of Jovian radii.

Bibliography

Barbosa D D and Kivelson M G 1983 Dawn–dusk electric field asymmetry of the Io plasma torus *Geophys. Res. Lett.* **10** 210–3

Brown R A 1974 Optical line emission from Io *Exploration of the Planetary System* ed A Woszczyk and C Iwaniszewska (Dordrecht: Reidel) pp 527–31

Clarke J T *et al* February 2002 Ultraviolet emissions from the magnetic footprints of Io, Ganymede and Europa on Jupiter *Nature* **415** 997–1000

Dessler A J 1983 *The Physics of the Jovian Magnetosphere* (Cambridge: Cambridge University Press)

Krimigis S M *et al* February 2002 A nebula of gases from Io surrounding Jupiter *Nature* **415** 994–996

Kupo I, Mekler Y and Eviatar A 1976 Detection of ionized sulphur in the Jovian magnetosphere *Astrophys. J.* **205** L51–3

Matson D L, Johnson T V and Fanale F P 1974 Sodium D-line emission from Io: Sputtering and resonant scattering hypotheses *Astrophys. J.* **192** L43–6

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Nash D B, Carr M H, Gradie J, Hunten D M and Yoder C F 1986 Io *Satellites* ed J A Burns and M S Matthews (Tucson, AZ: University of Arizona Press)
Peale S J, Cassen P and Reynolds R T 1979 Melting of Io by tidal dissipation *Science* **203** 892–4
Pilcher C B and Morgan J S 1979 Detection of singly ionized oxygen around Jupiter *Science* **205** 297–8
Roesler F L, Scherb F and Oliversen R J 1984 Periodic intensity variation in [SIII] 9531 Å emission from the Jupiter plasma torus *Geophys. Res. Lett.* **11** 128–30
Schneider N M and Trauger J T 1995 The structure of the Io torus *Astrophys. J.* **450** 450
Spencer J R and Schneider N M 1996 Io on the eve of the Galileo mission *Annu. Rev. Earth Pl. Sci.* **24** 125–90

Nicolas Thomas

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Io: Volcanism and Geophysics

Io is the innermost of the four large Galilean satellites of Jupiter discovered by Galileo Galilei in 1610. Io's mean radius (1821 km) and bulk density (3.53 g cm^{-3}) are comparable with those of the MOON. However, long before the VOYAGER spacecraft encounters, it was apparent from Earth-based observations that Io is very different from the Moon: it has an unusual color and anomalous thermal properties, and it is surrounded by immense clouds of ions and neutral atoms. These observations were much better understood following the discovery of active volcanic plumes and hot spots on Io during the Voyager 1 spacecraft encounter in 1979. The volcanism is driven by tidal heating, from periodic deformation of Io's shape as it orbits massive JUPITER.

Many Voyager-era investigators thought that Io's active volcanism was dominated by sulfurous eruptions rather than the silicate volcanism seen in the inner solar system. More recent observations by Earth-based telescopes and the GALILEO MISSION (1995–1998+) have shown that the volcanism on Io is dominated by high-temperature eruptions of silicate lava.

Gravity measurements from tracking of the Galileo spacecraft indicate that Io has a large iron or iron–iron sulfide core, constituting up to 20% of the satellite's mass. The bulk density of Io, its rugged topography and models of satellite origin suggest that the bulk composition of the crust and mantle is silicate. However, the surface composition has been greatly altered by volcanic outgassing. Sulfur dioxide (SO_2) has been positively identified both as a gas and as a surface frost. Elemental sulfur is considered a likely surface component on the basis of (1) the similarity of Io's ultraviolet through near-infrared spectral reflectance to that of powdered sulfur and (2) the detection of sulfur ions in IO'S PLASMA TORUS. The detections of clouds of Na and K around Io indicate that these elements must be present on the surface in some form.

Tidal heating mechanisms and internal structure

The importance of tidal heating in powering the enhanced heat flow and active volcanism of Io is widely accepted. All plausible non-tidal heat sources are about 2 orders of magnitude less energetic than the power output of Io (at least 10^{14} W). Accretional heating can supply, at most, half the gravitational potential energy of Io; the power output, averaged over 4.5 billion years, is 5×10^{11} W. However, the accretional heat was probably removed by conduction or convection very early in Io's history. Core formation can supply only a fraction of the maximum accretional energy, and it also is short lived. Radiogenic heating is a continuous process and is estimated to supply 5×10^{11} W. Heating of the surface by an electrical current running between Io and Jupiter's ionosphere provides a maximum power of 2×10^{11} W, but an ionosphere is likely to divert most of the current around Io.

The basic tidal heating mechanism involves deformation of Io into a triaxial ellipsoid by Jupiter's gravitational field. Because Io's orbit is eccentric, the satellite undergoes a periodic deformation. The dissipation also decreases Io's orbital eccentricity, which is forced to remain non-zero by the orbital resonance between Io and EUROPA. The total rate of tidal dissipation within Io depends on the internal structure and material properties of both Io and Jupiter. In particular, the transfer of orbital energy from Jupiter to Io is a consequence of the bulge raised on Jupiter by Io, and the rate of such transfer depends on Jupiter's dissipation factor, but this value is poorly known.

The mechanism of heat dissipation is intimately tied to Io's internal structure. An early model, with 'runaway' melting of the interior and tidal energy dissipated within a thin, elastic lithosphere, now seems unlikely. Dissipation of all the tidal energy within the lithosphere requires that it be 8–18 km thick, and such a thin lithosphere is inconsistent with the presence of mountains more than 10 km high (see below). The model may be modified by invoking a thicker lithosphere and heat transfer by volcanism rather than conduction, perhaps overlying crystal-rich magma rather than a fully molten interior. However, runaway melting might never have occurred if subsolidus convection removes the heat faster than it is added. Most workers consider Io to be a differentiated body with an iron-rich core, a solid mantle perhaps with an upper partly molten layer, overlain by a thick (30–100 km) lithosphere. Most of the tidal energy is probably dissipated by viscous or viscoelastic deformation below the lithosphere, and most of the heat is transported to the surface by silicate magmas.

Geomorphology

The morphology and surface markings on Io are almost entirely the result of volcanic processes, with the possible exception of the high mountains. Impact craters, the dominant landform on most solid planetary bodies, have not been observed on Io, despite the comet flux concentration by Jupiter's gravity. The lack of impact craters indicates that the surface is very young, and is consistent with estimates of resurfacing by volcanic activity ($\sim$1 cm yr^{-1}). The intense volcanism has resulted in a variety of surface features of three general categories: volcanic centers, mountains and plains (figure 1).

Io is spotted by local dark markings that probably consist of recently erupted silicate lavas and marking active volcanic centers. Where picture resolution and illumination conditions permit discrimination of surface relief, many of these markings are seen to lie within depressions that resemble terrestrial calderas. Calderas have steep, scalloped walls, flat floors and smooth rims; they form by collapse over shallow magma chambers following eruption of volcanic materials. Dark markings may cover all or part of a caldera floor and adjacent regions. The calderas range from about 10 to 200 km in diameter and are as deep as $\sim$2 km. Most calderas are not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

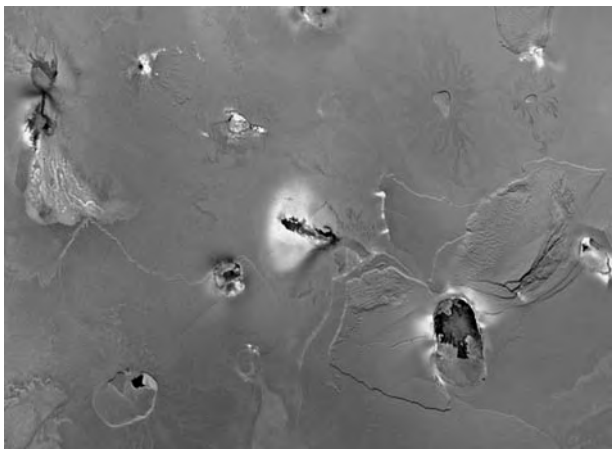


Figure 1. Mosaic of Voyager images of Io at a scale of 1 km/pixel. The landforms of Io include tall mountains, plateaus, calderas and layered plains. There are also color and albedo markings suggestive of lava flows and plume deposits.

associated with an obvious edifice, but some low shield-shaped edifices with summit calderas are present. The steep caldera walls and association with high-temperature hot spots are inconsistent with the concept of a thick sulfur crust. Io's crust is probably composed of silicate rocks, and the sulfurous materials form a relatively thin surface veneer as well as ground fluids at depth.

Elongated sinuous markings like lava flows can be seen in many areas, especially when not confined within a caldera (figures 1 and 2). (The caldera floors may be covered by coalesced flows, so the flow margins are less apparent.) Most dark markings are sources of high-temperature hot spots consistent with ongoing eruptions of silicate lava. Thus, the dark areas are the sites of active volcanic vents, and there are about 100 such active centers on Io. The spectral reflectances of the dark materials are consistent with silicates containing magnesium-rich orthopyroxene, and with varying degrees of sulfurous coatings. Diffuse bright white, yellow and red markings in and around volcanic centers are common and may be the result of sulfurous fumarolic and plume activity.

The second major class of surface feature is mountains, which typically have irregular outlines and rugged surfaces that appear tectonically disrupted. They may be as large as 600 km in basal diameter and 15 km high. The origin of these high mountains is not resolved. They may be degraded volcanoes, exposures of a silicate basement complex or uplifted (and sometimes tilted) crustal blocks. Most workers agree that the mountains must be dominantly silicate rather than sulfur, because sulfur does not have sufficient strength to support the observed relief. Although superposition relations are unclear, the mountains are generally considered to be the oldest materials exposed on Io's surface, but higher-resolution images are required to confirm this hypothesis.

Most of Io's surface is covered by plains with little apparent relief. They have complex ALBEDO and color patterns, and, where relief is discernible, they are often seen to be layered. The plains and their surface patterns may have been formed by a variety of materials and processes, including lava flows, deposition from plumes, condensation of volatiles and possibly pyroclastic flows.

Volcanic plumes

One of the most dramatic phenomena on the surface of Io is the active volcanic plumes. They erupt under conditions of pressure, temperature and gravity very different from those on Earth, and they are probably driven by SO_2 and/or sulfur in which multiple phase changes can occur, so the plume dynamics and thermodynamics are unlike those of terrestrial eruptions. The low gravity and especially the very low atmospheric pressure on Io ($\sim 10^{-9}$ bar) result in plumes up to 10^4 times larger on Io than would be produced by comparable eruptions on Earth (see VOLCANISM IN THE SOLAR SYSTEM).

Nine eruption plumes were observed during the Voyager 1 encounter; eight of these were also observed 4 months later by Voyager 2. Galileo has observed a total of 10 plumes as of mid-1998, four of which appear related to Voyager-era plumes, so there are a total of 15 volcanic centers with observed plumes. Ring-shaped surface deposits suggest that many other plumes have been recently active as well. Many of the plumes, such as that of Prometheus (figure 2), are 50–150 km high, long lived, deposit bright white material (probably SO_2 frost) and are associated with high-temperature hot spots. Prometheus and other plumes show signs of lateral migrations over time of up to 100 km. Pele's plume is very different from Prometheus-type plumes, as it is very faint at visible wavelengths, up to 460 km high and deposits bright red material. The Pele hot spot is also unusual because it maintains an intense high-temperature component at surprisingly constant levels over long periods of time.

The dominant volatiles driving explosive volcanism on Earth, H_2O and CO_2, seem to be highly depleted on Io. Discernible H_2O or carbonate absorption bands are absent, and C has not been detected in Jupiter's magnetosphere near Io (see JUPITER: MAGNETOSPHERE). H has been detected over Io's poles, which may indicate the presence of magmatic H_2O or H_2S. Nevertheless, volatiles such as SO_2 or sulfur are thought to be the major species driving the explosive volcanism.

Surface changes

About 30 large-scale (tens of kilometers) surface changes are obvious from comparison of the Galileo images with those acquired by Voyager in 1979 (figure 2). These include new plume deposits of several colors, bright and dark flows, and caldera-floor materials. There have also been significant surface changes on Io during the Galileo mission, such as a new 400 km diameter dark deposit around Pillan Patera (figure 3). While these surface changes are impressive, the number of large-scale changes observed in the 4 months between the Voyager 1 and Voyager 2 flybys in 1979 suggested that over 17 years the cumulative

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

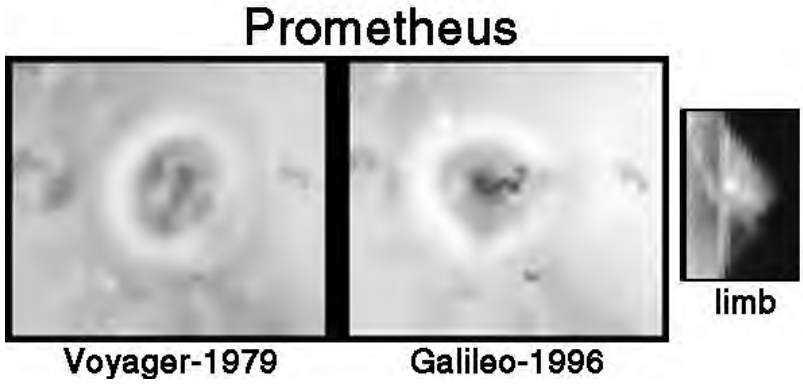


Figure 2. Three views of the active plume Prometheus. The left image shows the Prometheus region as seen by Voyager, and the middle view shows the same region seen by Galileo. There are new dark lava flows and the plume vent has moved about 70 km to the west. The right view shows the plume on the limb, against black sky, as seen by Galileo. Plume's height is 75 km, and bright ring has a diameter of about 300 km.

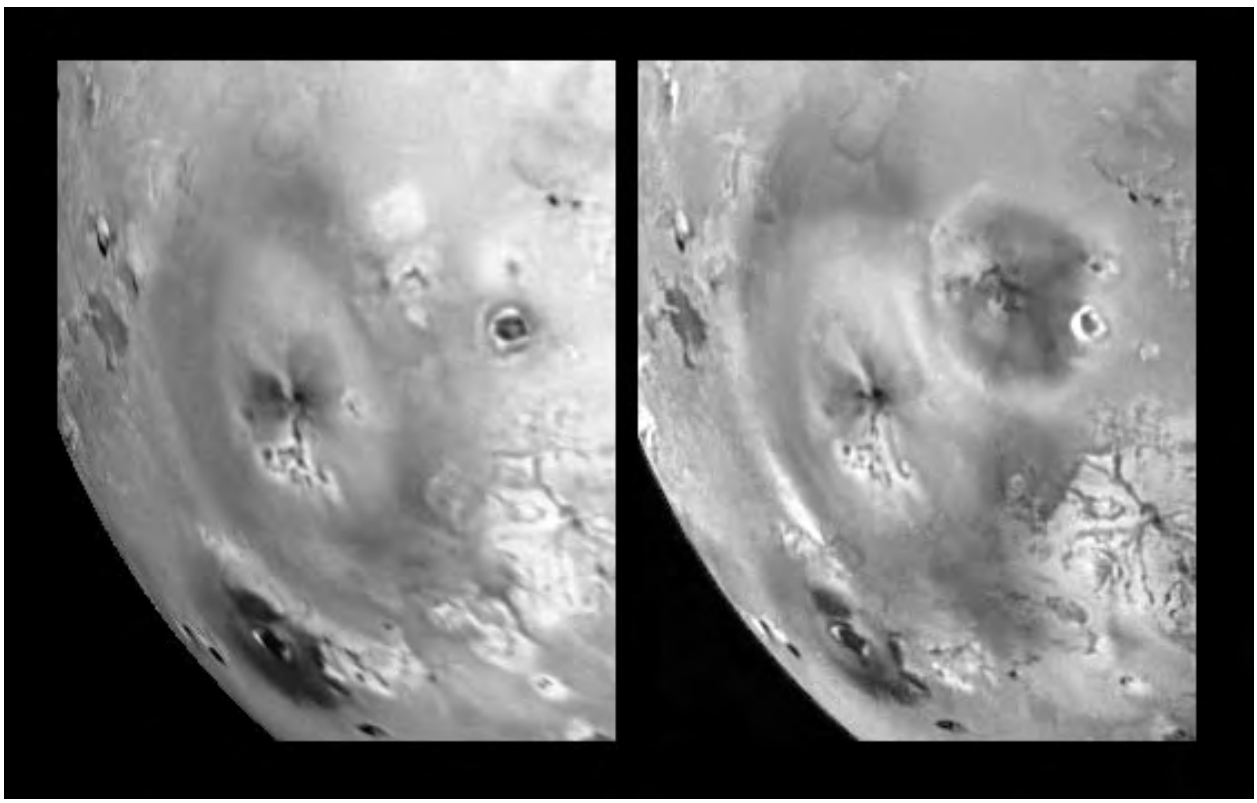


Figure 3. Two views of the Pele and Pillan volcanoes on Io, as seen by Galileo in the spring of 1997 (left) and fall of 1997 (right). The large ring is the fallout deposit of the 460 km high Pele plume. The eruption near Pillan Patera north-east of Pele produced the new dark deposits seen in the right-hand image, with a diameter of 400 km.

changes would have been much more impressive. There are two reasons why this was not actually the case. First, the most widespread plume deposits are ephemeral and seem to fade away within a few years. Second, it appears that a large fraction of the volcanic activity is confined to repeated resurfacing of dark calderas and flow fields that cover only a few per cent of Io's surface.

Hot spots

Evidence for hot spots or thermal anomalies on Io was acquired years before the Voyager encounters, but it was not understood. Thermal infrared observations

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 4. Image of Io in eclipse (when Io is in Jupiter's shadow). Small bright spots are high-temperature hot spots. The two brightest spots are due to Pele and Pillan (see figure 3). Pillan produced a very high-temperature hot spot, with lava temperatures exceeding 1700 K, probably because of the eruption of ultramafic lava from Io's mantle. Diffuse glows from electronic excitation of gases highlight Io's limb and active plumes.

had revealed two curious relations. First, when Io is eclipsed by Jupiter (figure 4), its temperature falls as expected for a surface with low thermal inertia, but its minimum temperature is too high. Second, the infrared brightness temperature of Io is significantly higher at shorter wavelengths. After the Voyager 1 encounter and the discovery of active volcanism, these observations were understood as being due to hot spots with temperatures well above 150 K, the peak temperature of the surface if heated only by the distant Sun. More than 50 hot spots have detected by Voyager, Galileo and telescopic observers. Voyager could not detect small areas at temperatures higher than ∼650 K because of limitations in sensitivity and wavelength coverage. Ground-based telescopic observers and Galileo have detected much higher temperatures, consistent with silicate volcanism.

Observations by Galileo have confirmed that high-temperature hot spots are ubiquitous on Io. There are more than 40 volcanic centers that reveal temperatures well above those expected from sulfur volcanism. Perhaps most surprising, however, has been the discovery than at least a dozen of these centers are erupting lavas with temperatures exceeding 1500 K, the peak temperature of basaltic lavas erupting on Earth. The most likely explanation is that these lavas are ultramafic (magnesium-rich) silicates.

The current eruption of ultramafic lavas raises questions about the composition of Io's crust. The intense volcanic activity, if it has persisted for at least a few per cent of geological time, should have strongly differentiated the silicates in Io. This differentiation, in the absence of remixing, should produce a ∼50 km thick, low-density crust rich in the lighter elements common in silicate rocks such as sodium, potassium, silica and aluminum. However, the high-temperature lavas suggest that dense melts from the magnesium-rich mantle are reaching the surface, so vigorous mixing between the mantle and crust may produce a crust with a mafic composition.

The best estimate of Io's global hot-spot heat flow is 2.5 W m^{-2}, and Io's total heat flow may be twice this value owing to conducted heat from recently emplaced lava flows. Therefore, the total energy loss from Io is probably from 2.5 to 5.0 W m^{-2}, orders of magnitude greater than the average heat loss from Earth (0.08 W m^{-2}) or from the Moon (0.02 W m^{-2}). The maximum energy flow theoretically possible from tidal heating is 0.8 W m^{-2} if steady state over the age of the solar system. Since Io's current heat flow clearly exceeds this limit and is emanating from many separate volcanic centers, it seems likely that the rate of tidal dissipation and/or heat flow has varied over time.

Global distribution of volcanism

Active and inactive volcanic centers appear to be almost uniformly distributed over Io, but several types of volcanic activity are non-uniform. The active plumes clearly prefer the equatorial region, with 14 out of 15 erupting from within 30° of the equator. The persistently active hot spots (which must release the most heat) are also concentrated in the equatorial region. A field of bright vents seen in eclipse occurs over the sub-Jovian region, and a region of diffuse glow seen in eclipse at low resolution, suggests that a similar field may exist over the anti-Jovian region. The region of Io from longitude 240° to 360° has distinctive color and albedo patterns and the plumes in this region display a greater range of characteristics than plumes at other longitudes.

The distribution of volcanism over Io's surface can be expected to reveal something about the nature of heating within the interior. Because of the intense tidal heating and the apparently high temperatures in Io's interior, convection should be the main internal mode of heat transfer. If tidal heating is predominantly in a partially molten layer (asthenosphere) of order 100 km thick, convection should occur globally in the asthenosphere with centers of upwelling and downwelling separated by a few hundred kilometers. Modeling indicates that the heat flow should be highest in the equatorial regions. Thus this model nicely explains the presence of active or recently active volcanic centers peppering the surface at several hundred kilometer intervals, and with an equatorial concentration of more energetic hot spots. The distribution of active plumes seems to follow the distribution of especially energetic volcanic centers, although plumes have not (yet) been seen at many of these centers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The major current exception to this uniform longitudinal pattern of heat flow is Loki Patera, an especially large dark-floored caldera which alone accounts for about a quarter of Io's heat loss. The heat being radiated from Loki is an order of magnitude larger than the heat loss from all the Earth's currently active hot spots. Tidal heating in the deep mantle would result in larger-scale convection perhaps with a relatively small number of mantle plumes spaced widely apart and carrying most of the deeply generated heat. Perhaps this can explain the Loki anomaly.

Future exploration

Many important questions about Io remain unanswered. The Galileo spacecraft reached Jupiter and made a close flyby of Io in 1995, but a tape recorder anomaly precluded the acquisition of high-resolution images or spectra of Io. With the extended mission now in progress, the plan is to return to Io in late 1999. There will be two close flybys of Io, one a near-equatorial pass providing a close look at volcanoes such as Pele and Pillan, and the second a pass under the south pole to determine whether Io has a self-sustained magnetic field. Other key issues that will be addressed are (1) emplacement mechanisms and compositions of erupting lavas, (2) composition of the crust, (3) origins of the large mountains and other landforms, (4) dynamics of the plumes and (5) polar volatile compositions.

Bibliography

Belton M J S, West R A and Rahe J (eds) 1989 *Time-Variable Phenomena in the Jovian System (NASA SP-494)* (Washington, DC: US Government Printing Office)

Burns J A and Matthews M S (eds) 1986 *Satellites* (Tucson, AZ: University of Arizona Press)

Johnson T V and Soderblom L A 1983 Io *Sci. Am.* **249** (6) 56–67

Keszthelyi L and McEwen A 1997 Magmatic differentiation of Io *Icarus* **130** 437–48

McEwen A S *et al* 1998 High-temperature silicate volcanism on Jupiter's moon Io *Science* **281** 87–90

McEwen A S *et al* 1998 Active volcanism on Io as seen by Galileo SSI *Icarus* **135** 181–219

Morrison D (ed) 1982 *Satellites of Jupiter* (Tucson, AZ: University of Arizona Press)

Spencer J R and Schneider N M 1996 Io on the eve of the Galileo mission *Ann. Rev. Earth. Planet. Sci.* **24** 125–90

Alfred S McEwen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ion

An atom that has a net electrical charge because it has lost, or gained, one or more electrons relative to the normal complement possessed by a neutral atom. The number of electrons in an ordinary, neutral, atom is the same as the number of positively charged protons in its nucleus. Compared to a neutral atom, a positive ion has fewer electrons (and a net positive charge) and a negative ion has more (and a net negative charge). The term 'ion' also applies to molecules and other groupings of atoms with a net electrical charge.

See also: atom, electron, ionization, molecule, proton.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ionization

The process whereby atoms lose one or more electrons to become positively charged ions. Ionization will occur if an atom absorbs a sufficiently energetic photon or if it suffers a sufficiently violent collision.

The minimum amount of energy needed to remove an electron to infinity from the ground state of a given atom is called the ionization potential, or ionization energy. For a hydrogen atom, with its single electron in the ground state (the lowest energy level of the atom), the ionization potential is 13.6 electronvolts (eV). If the electron is in a higher (excited) level, less energy is needed to remove it from the atom. In general, the ionization potential (IP) for an electron in excitation level n ($n = 1$ for the ground level, 2 for the first excited level, and so on) of a hydrogen atom is given by $\mathrm{IP} = 13.6/n^2$ eV. Thus, for example, the energy required to remove an electron from the first excited level $(n = 2) = 13.6/2^2 = 13.6/4 = 3.4$ eV. Where an atom has two or more electrons, the ionization potential for the second and subsequent electrons is greater than for the first electron. For example, the ionization potentials for the first and second electrons in a helium atom are 24.6 eV and 54.4 eV respectively. Where the energy (E) absorbed by an electron exceeds the ionization potential (IP), the kinetic energy of the escaping electron is given by $E - \mathrm{IP}$.

An atom that has lost one electron is identified by the roman numeral II (the neutral atom, by I) or by the superscript +. For example, neutral hydrogen is denoted by HI and ionized hydrogen by HII or H^+. Doubly ionized helium (helium that has lost both electrons) would be denoted by He III or He^{++}. In many astrophysical situations, for example, in very high-temperature gases, very high ionization stages (i.e. atoms that have lost many electrons) are encountered, and the roman convention is normally used; for example, Fe XVI represents an iron atom that has been stripped of 15 out of its total complement of 26 electrons. Ions that retain at least one bound electron can absorb or emit radiation, so producing spectral lines (emission and absorption lines) that differ in wavelength from those produced by neutral atoms. Photons are also emitted when an ion captures, or recaptures, an electron (this process being called 'recombination').

Ionized gases are encountered in a wide variety of astrophysical situations; for example, in the interiors and atmospheres of stars, in luminous nebulae (glowing clouds that contain substantial quantities of ionized hydrogen and which, for that reason, are otherwise known as HII regions), or in high-temperature intergalactic gas.

See also: absorption spectrum, atom, electron, electron-volt, emission spectrum, ion, photon, plasma, radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

IRIS (ESRO-2B)

European Space Research Organisation satellite, launched May 1968. Carried out studies of solar and cosmic x-rays, and solar particle emissions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Iris (Infrared Imaging Surveyor/Astro-F)

Japanese infrared satellite scheduled for launch in 2003. Intended to carry out a sky survey and observations of individual sources. Will carry a 70 cm Ritchey–Chretien telescope and a hybrid cooling system with liquid helium and mechanical cooling. (See also ASTRO.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Irish Astronomy

This account of national astronomy in Ireland is given under four categories: (1) national astronomical observatories, (2) university and college departmental astronomy, (3) private astronomical observatories and individuals and (4) astronomical societies and groups. A short history of each establishment is followed by a report on its current status together with present research activities. The article is partly composed from published literature.

National astronomical observatories

Dunsink Observatory

DUNSINK OBSERVATORY, the observatory of Trinity College Dublin, was founded in 1783. It was designed by the architect Graham Moyers according to principles laid down by Henry Ussher. Ussher was the first director of the new observatory, and he held office until his death in May 1790.

Initial work at the Observatory included stellar and planetary positional astronomy using transit instruments by Jesse Ramsden and astronomical regulators by Arnold and Crosthwaite. Stellar positional astronomy was begun in Dublin by Henry Ussher on 1 January 1788. Ussher maintained a weather record from July 1786 to May 1790, and it was he who encouraged and guided the young (later Admiral Sir) Francis Beaufort.

John Brinkley succeeded Ussher as Andrews' Professor in December 1790, becoming Royal Astronomer in 1792. The problem of measuring the distances of the stars occupied much of Brinkley's time. The method used was to measure the apparent motion of nearby stars against the more distant stellar background utilizing the different viewing positions afforded by the annual motion of the Earth in its orbit around the Sun.

Brinkley was followed in 1827 by Sir William Rowan Hamilton, then 21 yr of age, who held the post of Director until 1865. Internationally respected as a mathematical genius, Hamilton invented quaternions, an algebraic system with similarities to vector analysis. He scratched the following set of equations on a stone of Broome (Brougham) Bridge near Dublin on 16 October 1843:

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ijk} = -1.$$

Various cyclic properties resulting from these equations also hold true and so with four pure numbers a quaternion may be formed as follows:

$$q = a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}.$$

A scalar is simply a quarternion with $b = c = d = 0$, while a vector is one with $a = 0$. Hamilton also made significant contributions in optics and mechanics, and predicted conical refraction.

Robert Stawell Ball headed the Observatory from 1874 until 1892, his main contribution being in the theory of screws dealing with the movements of a rigid body in three dimensions.

Several later holders of the office of Director of Dunsink Observatory were accomplished mathematicians: Charles Jasper Joly, Edmund Taylor Whittaker and Henry Crozier Plummer. Joly (1897–1906) wrote many articles on quaternions and edited Hamilton's *Elements of Quaternions*. He also compiled a zodiacal catalogue, that is a catalogue of stars lying in the ecliptic or apparent path of the Sun through the heavens. Whittaker (1906–12) wrote texts on *Modern Analysis* and *Analytical Dynamics*, while the future Prime Minister and President of Ireland, Eamon de Valera, attended lectures given by Whittaker. Plummer (1912–21) commenced the study of variable stars and star clusters, as well as writing the text *An Introductory Treatise on Dynamical Astronomy*.

Following Plummer, Charles Martin was appointed acting Director in 1921 and occupied the position until 1936. Dunsink Observatory then fell into abeyance, closing in 1937. However, in 1943, de Valera obtained government approval for the establishment of the School of Cosmic Physics in the recently constituted Dublin Institute for Advanced Studies (DIAS), and by 1947 approval for the School was passed. By this time, Dunsink Observatory had been acquired by the State and, reopening in 1947 under the direction of Hermann Alexander Brück, became the astronomy section of the School of Cosmic Physics.

Soon after reopening, a proposal was put forward by Eric Mervyn Lindsay (see below) to establish a large telescope in the southern hemisphere, the Armagh–Dunsink–Harvard (ADH) Telescope—a 36 in Baker–Schmidt later equipped with a 33 in diameter objective prism. Solar studies were also commenced at Dunsink with the erection of a vertical CASSEGRAIN TELESCOPE and concave grating spectrograph. Measurements of the solar ultraviolet spectrum between 310 and 360 nm were incorporated in the second revision of Rowland's Preliminary Table. Stellar photoelectric photometry studies were initiated with the 28 in reflector in pulse-counting mode. A study of the interstellar absorption band at 443 nm was also undertaken. The scintillation, or *twinkling*, of the stars was also the subject of intensive study at this time.

Brück resigned in 1957 to take up the posts of Astronomer Royal for Scotland and Professor of Astronomy in the University of Edinburgh. Mervyn Archdall Ellison succeeded to the Dunsink post in 1958, undertaking solar flare observations with Dunsink then becoming a world data center for such studies.

Following Ellison's untimely death in 1963, Patrick Arthur Wayman was appointed Director in 1964. Investigations of the southern sky included photometry of Magellanic Cloud cepheids. Solar system and extragalactic studies also became important branches of research with the appointment of Tao Kiang to the staff. Kiang studied ancient Chinese records of Comet Halley and the structure of the main asteroid belt lying between Mars and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jupiter. Wayman retired in 1992 to be replaced by Evert J A Meurs, whose research interests include extragalactic astronomy, particularly active galactic nuclei (AGNs), starburst galaxies and stellar evolution, large-scale structure and clusters of galaxies. The observational material now studied extends beyond the visible light band into higher (x-rays) as well as lower (infrared, radio) energy bands of the electromagnetic spectrum.

The Governing Board of the School of Cosmic Physics recently decided to develop an interactive educational facility, particularly for the general public and school groups, which was opened in December 1997. The subject matter of the principal exhibits is inspired by the Junior Science Curriculum, providing an introduction to the astronomical world and highlighting the role of 'Gravity in the Universe'. There is also the long-established program of Open Nights at the Observatory.

Armagh Observatory

ARMAGH OBSERVATORY was founded in 1790 by Richard Robinson, the Anglican Archbishop of Armagh and Primate of All Ireland, as part of a greater plan for the establishment of a University of Ulster in Armagh. The Rector of the parish of Kildress in Cookstown, Co Tyrone, James Archibald Hamilton, was appointed as the first Astronomer of the new observatory. Astronomical observations began in July 1793 with the timing of a transit of α Ophiuchi using a transit instrument and a clock by Alexander Waugh of Armagh. Meteorological observations commenced in December 1794 and, with the exception of the period 1825–33, have been carried out continuously to the present day.

The early astronomical work consisted of transit observations for rating Thomas Earnshaw's regulator, and determining longitude by lunar observations. An equatorially mounted telescope was commissioned from the firm of J & E Troughton and arrived in December 1795. Hamilton died in 1815 and was succeeded by William Davenport. He concentrated on transit observations and equatorial measurements, but took his life in 1823 in his study.

A successful period in the observatory's history then ensued with the appointment of the highly respected THOMAS ROMNEY ROBINSON. Receiving his early education in Belfast, Robinson, by the age of 12, had printed a book of his own poems, the proceeds from the sale of which enabled him to commence studies at Trinity College Dublin. Robinson acquired a new transit instrument (1827) and mural circle (1831) from Thomas Jones & Son. After 26 yr of arduous labour with these telescopes, he produced a star catalogue entitled *Places of 5,345 Stars Observed at Armagh Observatory between the years 1828 and 1854.* For this catalogue, Robinson was awarded the Royal Medal of the Royal Society.

In 1835, Armagh Observatory acquired a 15 in equatorially mounted, clock-driven REFLECTOR by THOMAS GRUBB of Dublin. This instrument also employed an 18-point suspension system for supporting the primary speculum metal mirror. Robinson had a keen interest in meteorology since his college days and was also the inventor of the Robinson cup-anemometer, a wind-speed measuring instrument, an early version of which, still installed at the Observatory, dates from 1870. Robinson died in 1882 at the age of 89 yr while still in office. He had headed the Observatory for 59 yr, the longest tenure for any observatory director.

Robinson was succeeded by the Danish astronomer, JOHN LOUIS EMIL DREYER, famous historian of astronomy. Soon after arriving in Armagh, Dreyer began soliciting for the construction of a moderately large refractor as a memorial to his predecessor. The outcome was the erection of a 10 in Grubb REFRACTOR in 1885, and with it Dreyer confirmed some of his observations for his *New General Catalogue of Nebulæ and Clusters of Stars* published in 1888, and the two Index Catalogues in 1895 and 1910. Another important text was his *History of the Planetary Systems from Thales to Kepler* (1906). Dreyer also edited the works of that earlier Danish astronomer, TYCHO BRAHE, publishing most of his Tycho work during his tenure at Armagh. Dreyer resigned the directorship of the Observatory in 1916 and moved to Oxford, there to continue his work.

Joseph Alfred Hardcastle, a great-grandson of Sir William Herschel, was appointed director in 1917, but was taken ill while travelling to Armagh and had to stop the journey at his parents' home where he died at the age of 48. The following year, 1918, saw the appointment of the well-known telescope maker William Frederick Archdall Ellison. He held office until 1936. His son, Mervyn (see above), was Director of Dunsink Observatory from 1958 to 1963.

A strong programme of observational astronomy was then initiated by Eric Mervyn Lindsay, Director from 1937 until 1974. Lindsay's vision was to see Armagh Observatory have access to a moderately large telescope at a good observing site. He approached the governments in Belfast and Dublin, as well as HARLOW SHAPLEY of the Harvard College Observatory with the view of setting up a telescope in the southern hemisphere. The result, after much negotiation, was the Armagh–Dunsink–Harvard (ADH) Baker–Schmidt telescope, operating at the Boyden Observatory near Bloemfontein in South Africa from 1950 until 1976. The two nearby galaxies, the Magellanic Clouds, were intensively studied by Armagh astronomers during this period.

Lindsay was also responsible for the establishment of the Armagh Planetarium in the mid-1960s. The Planetarium was officially opened in 1968 with Patrick Moore as the first Director. It was Lindsay who, in 1948, secured a research position in Armagh for the Estonian-born astronomer ERNST JULIUS ÖPIK, who made fundamental contributions to various branches of astronomy, including solar system studies and stellar structure, and remained at Armagh until his retirement in 1981.

Following Eric Lindsay's death in 1974, Öpik became acting Director until the appointment of Martin Jan Hugo de Groot in 1976. For the next two decades the Observatory

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

focused attention on the mechanisms underlying the ejection of mass from hot stars, and the atmospheric and surface activity of the Sun and stars. Following de Groot's retirement in 1994, Mark Edward Bailey was appointed Director in 1995.

Current research at Armagh concentrates on stellar astrophysics including star formation and astrophysical jets, solar system studies, extrasolar planetary systems, and solar system–terrestrial relationships. Stellar astrophysics problems being investigated include helium stars and the late stages of stellar evolution, star-spots and stellar activity, and the heating mechanisms of the Sun's outer atmosphere corona. Spacecraft, such as YOHKOH (Japanese for Sunbeam) and the Solar and Heliospheric Observatory (SOHO), are being used in these solar studies, while solar system research includes the study of the formation and dynamics of comets and asteroids. Solar system–terrestrial studies are concentrated on how solar activity and interplanetary constituents affect the Earth's climate.

Armagh Observatory is one of the principal providers of astronomical research in Northern Ireland. In recent years, baseline support for its activities, together with those of the Armagh Planetarium, has come from the Department of Education Northern Ireland (now the Department of Culture, Arts and Leisure), while much of its research is funded through grants from the UK PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL.

University and College Departmental Astronomy

National University of Ireland

University College Cork. The Crawford Observatory of University College Cork was built in the 1870s by Howard Grubb. Its general plan is in the form of a Gothic-style two-storey central section flanked by two single-storey east–west wings. Grubb was awarded a gold medal at the Paris Exhibition of 1878 for the design and construction of the main instrument, an 8 in refractor. The mounting was modified in the 1890s to incorporate a 13 in astrograph as well as the 8 in refractor. The transit circle, probably the best preserved in Ireland today, employed a novel design in that the declination scales were inscribed on glass and read with transmitted light. The Observatory, which is undergoing some renovation, also houses Grubb's unique 4 in siderostatic telescope designed primarily for solar spectroscopic research.

Current astrophysical research includes investigations into cosmic-ray showers, cosmic-ray bursts from supernovae, cataclysmic astrophysical events, and low-mass x-ray binaries. Other fields of study are in gamma-ray and neutrino astrophysics. Infrared, optical and x-ray observations of white dwarfs, neutron stars and black hole candidates are made using ground-based (Keck, Hawaii) and space-based (ROSAT) observatories.

In addition, Hamiltonian quantum and other gravity theories, and mathematical aspects of supergravity, and super- and cosmic-string theory are investigated.

University College Galway. Atmospheric physics. The studies of the Earth's atmosphere include the greenhouse effect, ultraviolet radiation and air pollution. A field study at Mace Head, Co Galway, is part of the global environmental measurement program.

Gamma-ray astronomy. This research group uses a 10 m reflector on Mt Hopkins, Arizona, USA, for collecting faint Cherenkov light flashes produced in the upper atmosphere which have their origin in supernova remnants (SNRs) and AGNs.

Optical astronomy. Research areas include high-speed optical photometry of pulsars and globular clusters and the use of adaptive optics and other image-sharpening techniques in large telescopes including laser guide stars.

National University of Ireland at Maynooth. Stellar and planetary atmospheres. Collaborations with NASA's JET PROPULSION LABORATORY (JPL) on Lyman α emission from atomic hydrogen, excited by electrons, have provided input into stellar and planetary atmospheric models.

Submillimeter astronomy. This includes studies of interstellar clouds in the Galaxy and other, active, galaxies. This group also participates in the Planck Surveyor Cosmic Microwave Background EUROPEAN SPACE AGENCY (ESA) mission, scheduled for launch in 2005.

Gamma-ray astronomy. This team uses the 10 m telescope of the Whipple Observatory in Arizona to detect extremely energetic radiation from galactic and extragalactic sources, mainly AGNs.

Atmosphere. This involves study of the region 80–250 km, especially the mesopause (80–100 km), for nightglow emissions at optical and infrared wavelengths.

Space Technology Ireland Ltd. Among the projects in which this group is at present involved are the design and construction of spacecraft instrumentation for Gravity Probe B (NASA), Phoenix (ESA), Mars Express (ESA) and Rosetta (ESA). Experiments currently returning data to the Space Technology Ireland (STIL) group are Ireland's LION instrument on ESA's SOHO Mission which was built at STIL and the WAVES instrument on NASA's WIND Mission.

Regional technical colleges

Cork Institute of Technology. Current astrophysical research includes investigations into rapid variability in radioloud quasars and BL Lac objects and microvariability in both radioquiet and radio-intermediate quasars. Observations are taken at multifrequencies in collaboration with international partners. Other fields of study include observations of low surface brightness galaxies and blue compact galaxies. Observations for the various programs are taken at La Palma, Calar Alto, SAAO and with the ASCA satellite.

Galway–Mayo Institute of Technology. Astrophysics. This recent research initiative in astrophysics has been closely involved with the Whipple Gamma-ray Collaboration

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which operates an atmospheric Cherenkov imaging detector on Mt Hopkins, Arizona. The system comprises a 10 m optical reflector with an imager made of over 300 photomultiplier tubes and is used to detect compact sources of very-high-energy cosmic gamma-rays from pulsars and AGNs.

University College Dublin.
Astronomical research. This includes star formation and the study of blue compact dwarf galaxies and aspects of galaxy formation. Recent work includes investigations of the emission mechanisms for TeV cosmic ray sources and gamma-ray bursters (GRBs) using NASA's COMPTON GAMMA RAY OBSERVATORY, and searching for flaring and quiescent counterparts of these objects using optical and radio telescopes, such as the WESTERBORK SYNTHESIS RADIO TELESCOPE.

Dublin Institute for Advanced Studies (DIAS)
The DIAS was established in 1940 under the Institute for Advanced Studies Act of 1940 with two constituent Schools, Celtic Studies and Theoretical Physics. In 1947 a third School, Cosmic Physics, was established, itself divided into three Sections, Astronomy, Astrophysics (formerly Cosmic Ray) and Geophysics.

The central administration is located at 10 Burlington Road, Dublin 4, the Geophysics and Astrophysics Sections of the School of Cosmic Physics are at 5 Merrion Square, Dublin 2, and the Astronomy Section is located at Dunsink Observatory, Castleknock, Co Dublin. The DIAS is principally supported by an annual grant in aid from the State, and undertakes research and postgraduate student training.

Current research interests in the Astrophysics Section include plasma astrophysics, especially particle acceleration theory, and outflow phenomena associated with star formation. The Astronomy Section concentrates on advanced stages of stellar evolution and extragalactic x-ray astronomy. The School of Cosmic Physics has been involved in a number of space missions and is currently contributing to the optical monitor for INTEGRAL. Through the La Palma agreement it coordinates access to modern observational facilities.

Queen's University Belfast
Astronomical research is carried out at The Queen's University of Belfast in the Astrophysics and Planetary Science (APS) Research Division of the School of Mathematics and Physics. Two experimental programs on the upper atmosphere (led by Sir David Bates) and solar physics (Dan Bradley) combined about 30 yr ago. The APS Division was born out of this merger.

Initial activities of the group included the development of instrumentation for rocket and balloon flights and for ground-based telescopes. The current APS Division carries out observational astrophysics using ground-based optical and radio telescopes. Space observatories, such as SOHO, ISO and the Hubble Space Telescope, are also used extensively. The main thrust of the group is in astronomical spectroscopy and imaging in a number of areas of Galactic astrophysics. These include studies of the Sun and other stars, the interstellar medium and gaseous nebulae. A solar system research group investigates the origin and distribution of comets and asteroids. The APS Division is also involved in vigorous undergraduate and postgraduate teaching programmes, with some students carrying out their research projects at the Armagh Observatory. Queen's University and Armagh Observatory together form the Northern Ireland Node of the STARLINK computer network.

Private observatories and individuals

Birr Castle (Earl of Rosse)
William Parsons, Lord Oxmantown, later THIRD EARL OF ROSSE (1841), was educated at Trinity College and Magdalen College Oxford graduating in mathematics in 1822. He began telescope making in 1827, and completed a 72 in reflector in 1845 with which he discovered that year the spiral nature of the galaxy Messier 51 (M51, later named the Whirlpool galaxy) and similar structure the following year in the galaxy M99. Spectral studies of eleven nebulae were carried out using the giant reflector, known as the 'Leviathan of Parsonstown'. William's son, Laurence, later the FOURTH EARL OF ROSSE (1867), conducted lunar temperature measurements. Laurence's brother, (later Sir) Charles, was the inventor of the steam turbine. Otto Boeddicker produced an elaborate naked-eye map of the Milky Way in 1892. Sir CHARLES PARSONS purchased the equipment of Sir Howard Grubb and established the firm of Grubb-Parsons in Newcastle-upon-Tyne.

Markree Castle (E J Cooper)
This observatory was founded in Sligo in 1831 by Edward Joshua Cooper. Cooper purchased a 13 in Cauchoix objective and mounted it on a wooden stand at Markree. At the time, this was the most powerful refracting achromatic telescope in the world. Other equipment belonging to Cooper was a Troughton transit instrument (1831), a meridian circle with 7 in telescope by Ertel (1839) and a comet seeker by Ertel (1842). In a ROYAL ASTRONOMICAL SOCIETY report dated 1851, Markree Observatory was described as 'the most richly furnished private observatory then known'. Andrew Graham discovered the minor planet, (9) Metis, there in 1848, computing its orbit and analyzing its irregularities. A four-volume ecliptic star catalogue was also produced. The observatory ceased functioning in 1902.

Daramona Observatory (W E Wilson)
Founded in 1871 by William Edward Wilson in Streete, Co Westmeath, this observatory flourished until 1908. It was equipped with a 12 in Grubb reflector (1871) and a 24 in reflector with which the first photoelectric measurements of stars were made in 1895. The transits of stars were recorded by photography. A measurement of the effective temperature of the Sun was published in 1902.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Millbrook Observatory (J Birmingham)
From 1866 until 1884, the observatory of John Birmingham was in operation near Tuam, Co Galway. Equipped with a 4.5 in refractor, Birmingham compiled a catalogue of red stars. He also made observations of variable stars, sunspots and meteor showers.

Sherrington Observatory (W Erck)
The observatory was established in Co Wicklow by Wentworth Erck and operated from 1877 until 1890. Instruments included an Alvan Clark 7.5 in refractor, formerly owned by Dawes, and a 15 in reflector. Studies undertaken comprised solar disk observations and measurements of the motion of Jupiter's Great Red Spot.

J E Gore's Observatory (1879–1910)
John Ellard Gore founded his observatory first at Ballysodare, Co Sligo, then at 3 Northumberland Road, Dublin. His equipment included a 3 in refractor, and he produced a catalogue of variable stars, and a list of binary system orbits.

W S Monck's Observatory (1888–1915)
William Stanley Monck's observatory was situated at Earlsfort Terrace in Dublin. He employed a 7.5 in refractor and made the first photoelectric observations of planets. He came close to discovering the underlying nature of the diagram now named after Hertzsprung and Russell.

Edgeworth K E (1880–1972)
KENNETH ESSEX EDGEWORTH was born at Daramona House, a few years after his uncle W E Wilson had established his observatory there. Edgeworth had a military career, but he contributed to a wide variety of topics including electrical engineering, economics and astronomy. His main astronomical writings concerned the origin, structure and development of the solar system. In 1943, he predicted the existence of a large reservoir of potential comets occupying the region of space beyond the orbit of Neptune. An object, later named 1992 QB_1, was detected in this region in 1992 by David Jewitt and Jane Luu. This trans-Neptunian object was in fact the first of the Edgeworth bodies to be identified. GERARD KUIPER also made a similar postulation concerning the region at the edge of the planetary system in 1951, and this part of the solar system is now referred to as the Kuiper belt or Edgeworth–Kuiper belt.

Astronomical societies and groups

The Irish Astronomical Society (IAS)
The Irish Astronomical Society was founded in Dublin in 1937, in the same year as Dunsink Observatory was closed. On 1 October 1937, Joseph MacDermott published a letter in the *Dublin Evening Mail* suggesting the formation of a popular Dublin Astronomical Society. A group of 12 interested amateur astronomers responded to the letter and held a meeting in the Red Bank Restaurant. At that meeting, they formed the Irish Astronomical Society and appointed Mr MacDermott as Honorary Lecturer. The Society rented premises at 5 South Leinster Street and held its first Annual General Meeting at Jury's Hotel, Dame Street, on 21 March 1938. Later, during the Second World War, the Society met at 3 Burgh Quay with Uinsionn S Deiseach as resident lecturer.

From the 1950s onwards, many professional astronomers gave lectures to the Society, and by public expression in the newspapers moves were started to make the Society more widespread. After consultation with the Society in Dublin, it was agreed that various Centres should be established; the original IAS would now be the Dublin Centre of the IAS. Centres were established in Belfast (18 October 1946), Armagh (3 December 1946), Londonderry (4 November 1948), Clonmel, Cork and Galway. The Constitution and By-laws of the Society were formally adopted at the first council meeting at Armagh Observatory on 14 June 1947. A Council was in overall control of the Centres and comprised nominees from each Centre in addition to the Directors of Armagh and Dunsink Observatories.

Most of these Centres gradually declined and by the early 1970s only Dublin and Belfast remained. A revision of the Constitution was proposed in 1970 and a meeting was called to be held in Dundalk in November 1971 to consider the proposed revision. Because of the civil unrest then prevailing in Northern Ireland, the Belfast Centre felt it was unwise to travel across the border and the meeting never took place. In 1974, the Belfast Centre proposed the setting up of an essentially new Society having its secretariat based at the Armagh Planetarium. At a meeting of delegates in April 1974, the Dublin Centre succeeded in having this proposal rejected, but a short time later the Belfast Centre withdrew its affiliation with the Dublin Centre and created an independent organisation: the Irish Astronomical Association. As a result, the Dublin Centre is now, once again, the Irish Astronomical Society, holding a regular program of meetings and lectures, with a secretariat at PO Box 2547, Dublin 14 (telephone 00-353-1-298-0181). Publication: *Orbit*.

The Irish Astronomical Association (IAA)
The Irish Astronomical Association came into being when the Belfast Centre of the Irish Astronomical Society withdrew its affiliation with the IAS in the season 1973–4 (see previous section). The IAA was initially envisaged as an all-Ireland organization, with headquarters at the Armagh Planetarium and meetings to be held in both Belfast and Armagh. Nowadays, the Association is primarily active only in and around Belfast, although attracting members from the whole of Ireland, and it focuses its activities on a regular program of meetings and lectures. The IAA can be contacted c/o Mr R McLoughlin, 31 Manse Road, Ballygowan, Co Down, BT23 6HE. Publication: *Stardust*.

Astronomy Ireland
The most recent major amateur astronomical society to emerge in Ireland has been Astronomy Ireland, created

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in July 1990. The organization concentrates on astronomy and space science. Two lectures are held in Dublin each month from September to April, given by professional and amateur astronomers. The headquarters of the organization is at PO Box 2888, Dublin 1. Publication: *Astronomy & Space*.

Other societies

Cork Astronomy Club, c/o Mr Charles Coughlan, 12 Forest Ridge Crescent, Wilton, Cork (telephone 00-353-21-543669). East Antrim Astronomical Society, c/o Mr John McConnell, 56 The Meadows, Glen Road, Maghaberry, Moira (telephone 028-9261-9936). Shannonside Astronomy Club, c/o Mr Mike Murphy, Newmarket-on-Fergus, Co Clare (telephone 00-353-61-368025). Tullamore Astronomical Society, c/o Mr Sean McKenna, 145 Ardan Vale, Tullamore, Co Offaly (telephone 00-353-506-41983).

The Astronomical Science Group of Ireland (ASGI)

The Astronomical Science Group of Ireland was formed in 1974 as a collaborative forum for professional astronomers, physicists and technicians in both Northern Ireland and the Republic. The ASGI is currently an umbrella organization including members from all the observatories and university departments in which astronomical research is carried out, and also many of the smaller groups and societies. Meetings are held biannually in spring and autumn at each of the principal affiliated research institutions, in rotation. Financial support for the Group comes from an annual subscription from each of the affiliated groups or institutions, while the Group's affairs are managed by a small executive, comprising a Chairperson and a Secretary. Some of the papers presented at the ASGI have been published in *The Irish Astronomical Journal*.

Bibliography

Beesley D E 1976 Formation of the Irish Astronomical Association *Ir. Astron. J.* **12** 216

Bennett J A 1990 *Church, State and Astronomy in Ireland: 200 Years of Armagh Observatory* (The Armagh Observatory and The Institute of Irish Studies of The Queen's University of Belfast)

Butler C J 1986 Early photoelectric photometry in Dublin and Daramona *Ir. Astron. J.* **17** 373–7

Butler J 1994 Report on the Crawford Observatory to University College Cork, 11 pp

Butler J and Hoskin M 1987 The archives of Armagh Observatory. *J. Hist. Astron.* **18** 295–307

Dixon S M 1892 The photo-electric effect of star-light *Astron. Astrophys.* **11** 844

Dreyer J L E 1883 *An Historical Account of the Armagh Observatory* (Liverpool: Henry Greenwood, Printers)

Dreyer J L E 1890 New General Catalogue of Nebulæ and Clusters of Stars *Mem. R. Astron. Soc.* **49** 1–237

Dreyer J L E 1895 Index Catalogue of Nebulæ found in the years 1888–1894 *Mem. R. Astron. Soc.* **51** 185–228

Dreyer J L E 1906 *History of the Planetary Systems from Thales to Kepler* (Cambridge: Cambridge University Press) (Reprinted 1953 *A History of Astronomy from Thales to Kepler* (New York: Dover))

Dreyer J L E 1910 Second Index Catalogue of Nebulæ and Clusters of Stars *Mem. R. Astron. Soc.* **59** 105–98

Erck W 1877 Description of an observatory erected at Sherrington, Bray *Observatory* **1** 135–7

Gore J E 1888 A revised catalogue of variable stars *Proc. R. Ir. Acad.* **17** 97–150

Gore J E 1890 A catalogue of binary stars for which orbits have been computed *Proc. R. Ir. Acad.* **17** 571–99

Graham A 1848 Discovery of minor planet (9) Metis *Mon. Not. R. Astron. Soc.* **8** 146

Lindsay E M 1950 The Irish Astronomical Society *Ir. Astron. J.* **1** 1

McFarland J 1996 Kenneth Essex Edgeworth—Victorian polymath and founder of the Kuiper belt? *Vistas Astron.* **40** 343–54

McKenna S M P 1967 Astronomy in Ireland from 1790 *Vistas Astron.* **9** 283–96

Minchin G M 1895 The electrical measurement of starlight *Proc. R. Soc.* **58** 142–54

O'Connor J 1997–1998 A brief history of the Irish Astronomical Society *Orbit* **36–37**

Parsons C 1926 *The Scientific Papers of William Parsons, Third Earl of Rosse (1800–1867)* (Percy Lund, Humphries & Co Ltd. printers)

Robinson T R 1850 Description of an improved anemometer for registering the direction of the wind, and the space which it traverses in given intervals of time *Trans. R. Ir. Acad.* **22** 155–78, 3 plates

Robinson T R 1859 *Places of 5,345 Stars Observed from 1828 to 1854, at the Armagh Observatory* (Dublin: Alex. Thom & Sons)

Ussher H 1787 An account of the observatory belonging to Trinity College, Dublin *Trans. R. Ir. Acad.* **1** 3–21

Wayman P A 1987 *Dunsink Observatory 1785–1985. A Bicentennial History* (Dublin Institute for Advanced Studies and the Royal Dublin Society)

Wayman P A and de Groot M 1980 The Astronomical Science Group of Ireland 1974–1980 *Ir. Astron. J.* **14** 111–3

Wilson W E 1900 *Astronomical and Physical Researches made at Mr. Wilson's Observatory, Daramona, Westmeath*

John McFarland

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Iron Meteorite

A meteorite composed mainly of nickel–iron, with traces of other metals; also referred to simply as an *iron*, and formerly known as a *siderite*. Irons account for over 6% of all known meteorite specimens. They are the easiest type to identify, being heavy, magnetic and rust-colored; their metallic sheen tarnishes quickly on the Earth's surface, but otherwise irons show better resistance to weathering than do other meteorites. Iron meteorites are believed to have originated from the cores of asteroidal parent bodies which differentiated before being shattered by impact. They are divided into three main classes according to the crystal structure and nickel content of the nickel–iron: *hexahedrites* (4–6% nickel), *octahedrites* (6–12%) and *ataxites* (over 12%). When sectioned, etched and polished, hexahedrites reveal a fine rectangular grid of so-called *Neumann lines*; octahedrites display a hexagonal hatching known as the *Widmanstätten pattern*; while ataxites show no obvious characteristic. The largest known meteorites are irons. The two most massive are the Hoba (Namibia, ataxite, 60 tonnes) and Cape York (Greenland, octahedrite, 58 tonnes total) meteorites.

See also: Cape York meteorite, Hoba meteorite, Widmanstätten pattern.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Irregular Galaxy

A galaxy with an irregular appearance and no well-defined shape or structure. Irregular galaxies can be divided into two broad classes, designated Irr I and Irr II. Irr I galaxies contain many OB associations (clumps of hot, young, highly luminous stars) and HII regions (luminous clouds of ionized hydrogen that surround hot young stars) and significant quantities of gas (typically about 20% of the total mass is in the form of gas). Some of them contain hints of incipient spiral structure. Their masses range from about a hundred million (10^8) solar masses to several billion ($\sim 3 \times 10^9$) solar masses and their diameters from about 3000 to about 30 000 light-years. Irr II galaxies are amorphous in general appearance and often contain substantial amounts of gas and dust. Some have distorted shapes that appear to have been caused by collisions, or close encounters, with other galaxies or by violent internal activity.

See also: galaxy, Hubble classification, interacting galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Isaac Newton Group of Telescopes

The largest optical telescope for astronomy in Western Europe stands on the rim of an extinct volcano at an altitude of approximately 2350 m on the island of La Palma, part of the Canary Islands archipelago. The 4.2 m William Herschel Telescope, together with the 2.5 m Isaac Newton Telescope and the 1.0 m Jacobus Kapteyn Telescope, form the Isaac Newton Group of Telescopes (ING). First scheduled use by astronomers started in May 1984, and since then thousands of research projects have been carried out. Many ground-breaking discoveries have been made using these telescopes, such as the first detection of the optical afterglow of gamma-ray bursts, the discovery of black holes in binary star systems, and the study of supernova explosions in distant galaxies.

The Isaac Newton Group of Telescopes aims to deliver an effective telescope operation and a coherent instrument development program to facilitate the execution of world-class astronomical research. The construction, operation, and development of the ING telescopes is the result of a collaboration between the United Kingdom, the Netherlands and Eire. The telescopes are operated on behalf of the UK Particle Physics and Astronomy Research Council and the Nederlands Organization for Scientific Research. The observatory site 'Roque de los Muchachos' is provided by Spain. The ING comprises approximately 65 staff.

For further information see http://www.ing.iac.es/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ishtar Terra

One of the three large upland regions of Venus (the other two being Aphrodite Terra and Lada Terra), situated in the planet's northern hemisphere and centered at 70.4 °N, 27.5 °E. In its greatest dimension it measures 5609 km. It takes its name from a Babylonian goddess of love. The western end of Ishtar Terra is dominated by the pear-shaped volcanic plateau Lakshmi Planum, 2343 km across. The plateau contains two large calderas, Colette Patera and Sacajawea Patera, and is bounded to the northwest and southwest by mountain ranges, named Akna Montes and Danu Montes respectively, and to the north and northeast by Frejya Montes. To the east of Lakshmi lies Maxwell Montes, the highest mountain range on the planet. The eastern end of Ishtar consists of Fortuna Tessera, a less elevated area covered with a complex of intersecting troughs and ridges. Fortuna, and the other tesserae to the north and west of Lakshmi, result from lateral compression, part of the complex deformations that occurred on Venus before the lava eruptions in surrounding lowland regions.

See also: Maxwell Montes; Venus: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Islamic Astronomy

With the advent of Islam in the seventh century, and the swift creation of the largest empire known to medieval man, the civilization that it produced was the expression of the intellectual potential of that sprawling empire which stretched from the Pyrenees in the west to the Central Asian steppes in the east and the Indian subcontinent in the southeast. The very vastness of this empire required for its administration new scientific and technological resources. Some of those resources were already available from more ancient civilizations, others had to be created afresh.

The legacy from the past

The more ancient sciences that the Islamic civilization could draw upon were the Greek, the Persian, the Indian and indirectly the more ancient Babylonian ones. But the Greek legacy was by far the most important, perhaps because of its coherence as it was articulated within the Aristotelian cosmological framework. It is not surprising, therefore, to find the Greek astronomical tradition exacting the lion's share of the concerns of the Muslim scientists who began to forge their own tradition from the very dawn of their own civilization. The major astronomical texts that were resuscitated from the Greek tradition and were translated very early into Arabic—the lingua franca of the Islamic civilization—were essentially the three most important works of PTOLEMY (*fl.* AD 150): the *Almagest*, the *Planetary Hypotheses*, and the *Handy Tables*, the last being quickly superseded by the more up-to-date Islamic astronomical handbooks that pushed it out of circulation.

The other two texts of Ptolemy received the full attention that they deserved at the hands of the Muslim astronomers. But even as they were being translated they were also critically assessed. In the earliest surviving translation of the *Almagest*, that which was completed in the year AD 829, we find attempts to correct the perceived mistakes of the Greek original, as in the case of the value of the synodic lunar month.

But the *Almagest* contained other real mistakes, as we shall see. In it Ptolemy lays down the foundation of Greek astronomy starting with the basic cosmological and mathematical underpinnings of that astronomy, and proceeding to the observations of the phenomena themselves—always giving the mathematical models with which such phenomena could be saved, and the positions of the naked-eye planets that could be predicted. The most basic astronomical parameters of Greek astronomy are determined in this work.

In the *Planetary Hypotheses*, Ptolemy went a step further, by modifying some of the parameters he had determined in the *Almagest*, but most importantly by explicitly re-formulating the same Aristotelian cosmological principles of Greek astronomy, which were only assumed in the *Almagest*. The motion of the same planets could now be materially accounted for by the motion of the real spherical bodies that were supposed to carry them. Furthermore, Ptolemy even hoped to exhibit these motions with planetaria-like constructions.

The reform of astronomy

The sheer time that separated the translators of the *Almagest* from Ptolemy himself was sufficient to show some discrepancies between the original text and the observations the Muslim astronomers were conducting as they were translating. Parameters such as the rate of precession, the inclination of the ecliptic, the eccentricity of the solar orbit—and the resulting maximum solar equation, as well as the motion of the solar apogee (assumed fixed by Ptolemy)—were all found to be erroneous. To rectify the situation Muslim astronomers undertook new observation programs, built larger and more accurate instruments, and most importantly began to question the observational methods of Ptolemy that led to the errors in the first place. To give only one example, it was quickly realized, sometime during the early part of the ninth century, that Ptolemy's method for determining the solar eccentricity which depended on the observation of the Sun at the solstices was in itself flawed, because the variation of the solar declination at those points was indeed very hard to observe with the instruments of late antiquity and medieval times. Hence a new observational method was devised in which the critical observations were conducted in the middle of the seasons where the same variation is much easier to detect. The new method immediately yielded the desired fruits, and from the ninth century onward one would rarely find an Arabic astronomical source that would quote the Ptolemaic parameters. Instead one finds the improved ones that were determined during Islamic times.

On the level of instruments, in addition to building larger instruments, there were also developed new methods for perfecting older ones and for surpassing them in conception. Although the ASTROLABE was known from Greek times, new methods of projection were devised for the instrument as early as the ninth century which deployed new trigonometric techniques that were unknown from the Greek sources and are still essentially the same now. During the same period other mathematical fields such as algebra were also being re-formulated along lines that were not known from Greek times. The religious requirements that may have contributed to the re-formulation of such fields may have also given rise later on to new fields such as *mīqāt* (timekeeping) in which both spherical trigonometry and applied astronomy were brought to full maturity. The projection of astrolabes itself continued to be developed throughout Islamic times, and reached such heights that universal plates were later on devised that could be used anywhere on the globe.

The determination of basic parameters that made the Ptolemaic *Handy Tables* quickly obsolete also gave rise to the construction of new astronomical handbooks which included mainly mean motion and equation tables. They continued to develop and to evolve into more useful and 'user-friendly' tables throughout the Islamic period as well.

Faced with a Greek astronomical legacy whose observable errors became widely known from the earliest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Islamic times, the level of criticism directed at that legacy was also developed at the same pace. The same cosmological principles that were assumed in the *Almagest*, and enunciated more explicitly in the *Planetary Hypotheses*, when examined with a critical eye were also found to produce other theoretical problems of their own that were not considered in the Greek corpus. In essence those problems touched upon the very foundation of Greek astronomy, namely the principle of consistency between the accepted principles and the results derived therefrom. For example, it was quickly realized that although Ptolemy had stipulated in both the *Almagest* and the *Planetary Hypotheses* that he would follow the Aristotelian cosmology he then went on to violate it at almost every turn in those two books. These violations were noted very early on, and the first treatise that has survived in which such cosmological issues were discussed comes from the same first half of the ninth century. Other works were also motivated by the same concern. Collectively such works formed yet another new field that was not known in Greek, namely the field of *'ilm al-hay'a*, which focused mainly on the issue of consistency between the cosmological principles and the mathematical representations of those principles. For reasons that have to do with the general intellectual and religious environment of the time, the new discipline, *'ilm al-hay'a*, also distinguished itself by steering away from the religiously condemned discipline of astrology, in which Greek astronomy was heavily mired, and at the same time devoted itself only to planetary theories without much emphasis on practical astronomy. From the very beginning, therefore, we note the emergence of an Islamic astronomy which was new in its conception, scope and philosophical foundation.

Planetary theory in Islamic astronomy

Nowhere were the inadequacies of Greek astronomy more pronounced than in the planetary theories that included them (see PLANETARY THEORY FROM EUDOXUS TO COPERNICUS). To take only one example of the gross violation of the principle of consistency, Ptolemy had accepted the Aristotelian cosmological principle of uniform circular motion for the celestial spheres, but then when he came to describe the motions of the particular planets, and the spheres that carried them while they moved so that they would not create void, he resorted to mathematical models which required for all the planets, with the exception of the Sun, the uniform motions of those physical spheres around axes that did not pass through their centers. That in itself is impossible, and thus all the results derived from such models, although mathematically feasible and able to produce rather reliable predictions for the positions of planets, were nevertheless cosmologically absurd.

Beginning with the eleventh century, the criticism of these features of the Greek astronomical tradition became more focused and much more elaborate than before, thus producing books with such titles as *Doubts Regarding Ptolemy*, or *Recapitulation in Regard to Ptolemy*. In later centuries, the same violations committed by Ptolemy began to be codified in separate books called *Ishkālāt* (Problems, or Doubts). The natural result engendered by such systematic criticism was the production of alternative mathematical models that could at once be cosmologically feasible, account for the observations, i.e. save the phenomena, and still predict the positions of the planets with an accuracy as good as that predicted by Ptolemy if not better.

In the same eleventh century new attempts at constructing new models that would meet those strict requirements were first conceived but without meeting with much success. However, in the centuries that followed this activity became the most prevalent activity, and one astronomer after another would expend great efforts to produce such new mathematical models. The new discipline dealing with these planetary theories, the *'ilm al-hay'a*, was the natural home for such discussions.

Thirteenth-century development

The thirteenth century saw the establishment of one of the most elaborate astronomical institutions of medieval Islam, the Marāgha Observatory, which was founded by the astronomer and first director Naṣīr al-Dīn al-Ṭūsī (died 1274) in 1259, and his colleague who built the instruments for that observatory, Mu'ayyad al-Dīn al-'Urḍī (1266), where they were both joined later by Ṭūsī's student Quṭb al-Dīn al-Shīrāzī (1311) and collaborator Yaḥyā b. Abī al-Shukr al-Maghribī (1286) and others. As a result, the activity of rebuilding planetary theories with new mathematical models that were not plagued by the inadequacies of the Greek models began to move to more sophisticated grounds.

In their pursuit of such models, both Ṭūsī and 'Urḍī had to devise new mathematical theorems and to reconfigure their own mathematical models in order to accomplish the required motions of the celestial spheres, always working with the same Ptolemaic assumption that those spheres had the same properties stipulated by Aristotelian cosmology. In figure 1, 'Urḍī's lemma simply stipulated that if two lines, AG, DZ always formed equal angles with line AZ, either internally or externally, then the line GD joining their extremities would always be parallel to line AZ. When applied to the model of the upper planets as in figure 2, now marked by the line connections designated as 'Urḍī (Shīrāzī), that theorem will demonstrate that lines ZD and KL will always be parallel. This arrangement will then allow the epicyclic center of the planet Z to move uniformly around the center of its own sphere L, which in turn moves uniformly around the center of its own carrier K, which is placed from the observer at point Q at a distance equal to one and a half times the Ptolemaic eccentricity e. Note that in this arrangement, as required by the Aristotelian cosmology, all the spheres move in place around an axis that passes through their centers, unlike the Ptolemaic configuration, also illustrated in figure 2, where the sphere with center T moves uniformly around the equant D. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

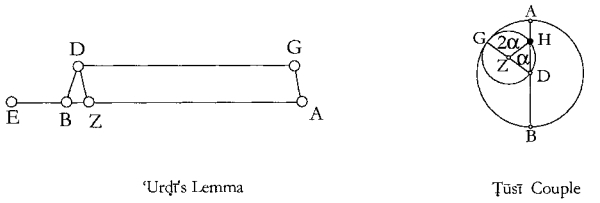


Figure 1. Mathematical theorems for the motions of the celestial spheres: (*a*) 'Urḍī's lemma; (b) Ṭūsī couple.

only violation of Aristotelian cosmology left is that the spheres proposed first by 'Urḍī, who was then followed by Shīrāzī, are still eccentric just as in the Ptolemaic model. But that is an easy hurdle to overcome for it can be assumed that the said eccentricity QK, if need be, could be substituted for by an epicycle and a concentric as was done by APOLLONIUS and Ptolemy after him when they demonstrated mathematically the equivalence between eccentric and epicyclic models. In fact, the connection designated in the same figure 2 with the name of the fourteenth century Damscene astronomer Ibn al-Shāṭir (1375) achieved just that. For, as it can be easily seen, all the connections proposed by this astronomer are strictly geocentric, as required by Aristotelian cosmology. His response to the deployment of epicycles, which could still be conceived as violating that cosmology, is that the epicycles of the largest planets are much smaller than some of the fixed stars and if Aristotle accepts that such fixed stars are made of the same element as the surrounding sphere that carries them then he would have to accept the same kind of composition in the heaven produced by the presence of such small epicycles.

Finally, when the Sun is held fixed, and the Earth is allowed to move around it, as stipulated by Copernican astronomy, we note in figure 3 that the same theorem first proposed by 'Urḍī is still deployed to assure the parallelism of lines OD, NK and their equivalents in the Copernican model. The theorem employed by NICOLAUS COPERNICUS was not enunciated as a new theorem as was done by 'Urḍī, nor was it proven by Copernicus as was done by 'Urḍī, and had to wait until KEPLER wrote to Maestlin asking him about it where he received the proof with the response.

Similarly, the theorem now called the Ṭūsī Couple (figure 1), which was first proposed in 1247 by Ṭūsī, was also used by both Ibn al-Shāṭir and Copernicus in their account for the motions of the planet Mercury. The theorem itself stipulates that if two spheres AGB, and GHD, where the diameter AD is twice the length of GD, are made to move in such a way as the bigger sphere moves in place at uniform speed and the smaller sphere moves in the opposite direction at twice that speed, then the original point of tangency H will oscillate back and forth along the diameter of the larger sphere, thereby creating linear motion out of uniform circular motion. Ṭūsī used that theorem to account for the motions of the Moon (not illustrated) and the upper planets as seen in figure 2.

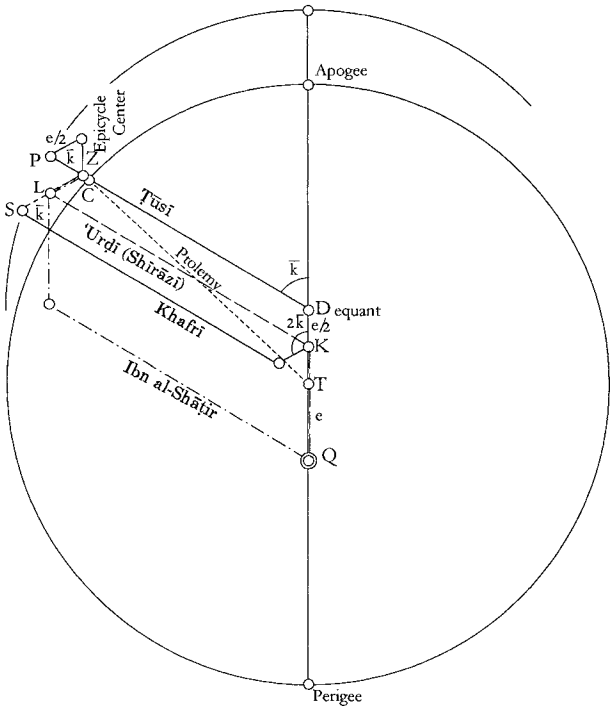


Figure 2. Mathematical models applied to the motions of the upper planets.

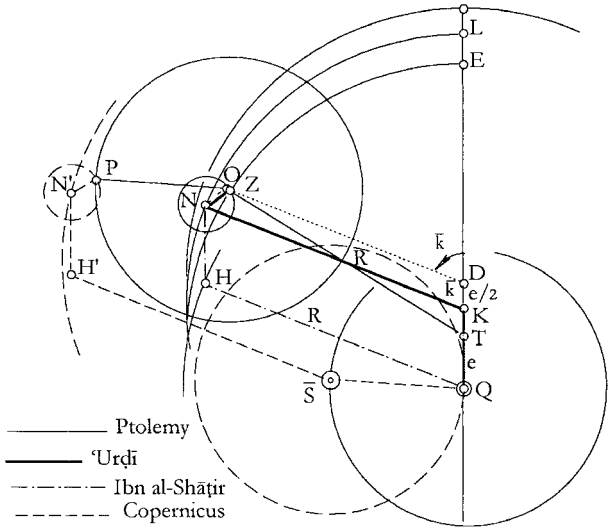


Figure 3. Comparison of the models of Ptolemy, 'Urḍī, and Ibn al-Shāṭṭir with the Copernican model.

With those two theorems being the only additions made by Copernicus to the Ptolemaic models (heliocentricity excepted) it becomes very clear that the mathematical developments that were devised in the Islamic world were themselves the same ones deployed by Copernicus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Islamic astronomy in the sixteenth century

But by the sixteenth century, Muslim astronomers had finally come to realize that all those problems, when stated in the Ptolemaic–Aristotelian fashion, do not in fact have a unique mathematical solution. Figure 2 illustrates the mathematical equivalence of all the solutions proposed, and the person who apparently came to that realization was the same Khafrī (1550) whose configuration for the upper planets is also illustrated in the same figure.

This line of research which was pursued by Muslim astronomers in order to resolve the problems posed by the Greek astronomical tradition seems to have been the same line that was followed by Copernicus in the introduction to his *Commentariolus* (again heliocentricity excepted). From that perspective, it becomes very clear that the Islamic astronomical tradition, which seems to have been passed on to Europe during the Renaissance, was itself embedded in the much larger context of rejecting the Greek astronomical tradition, rather than preserving it as is often asserted.

Bibliography

Kennedy E S 1956 Survey of Islamic astronomical tables *Trans. Am. Phil. Soc., new ser.* **46** 123–77

King D 1987 Universal solutions in Islamic astronomy *From Ancient Omens to Statistical Mechanics: Essays on the Exact Sciences Presented to Asger Aaboe (Acta Historica Scientiarum Naturalium et Medicinalium)* vol 39, ed J L Berggren and B R Goldstein (Copenhagen: Bibliotheca Unversitatis Hauniensis) pp 121–32

Saliba G 1994 *A History of Arabic Astronomy: Planetary Theories During the Golden Age of Islam* (New York: New York University Press)

Saliba G 1996 Arabic planetary theories after the eleventh century AD *Encyclopedia of the History of Arabic Science* (London: Routledge) pp 58–127

Saliba G 1997 A redeployment of mathematics in a sixteenth-century Arabic critique of Ptolemaic astronomy *Perspectives Arabes et Médiévales sur la Tradition Scientifique et Philosophique Grecque. Actes du Colloque de la Société Internationale d'Histoire des Sciences et de la Philosophie Arabe et Islamique (Paris, 31 March–3 April 1993)* ed A Hasnawi, A Elamrani-Jamal and M Aouad (Paris: Peeters) pp 105–22

Sayılı A 1960 The observatory in Islam *Publications of The Turkish Historical Society* ser. vii, no 38

Swerdlow N 1973 The derivation and first draft of Copernicus's Planetary Theory: a translation of the *Commentariolus* with commentary *Proc. Am. Phil. Soc.* **117** 423–512

George Saliba

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Isotope

Any of two or more forms of the same chemical element, the atomic nuclei of which contain the same number of protons (and therefore have the same atomic number) but different numbers of neutrons (and therefore have different mass numbers). For example, oxygen has three stable isotopes, oxygen-16, oxygen-17 and oxygen-18, denoted symbolically by ${}^{16}_{8}\mathrm{O}$, ${}^{17}_{8}\mathrm{O}$, and ${}^{18}_{8}\mathrm{O}$, respectively (the subscript denotes atomic number and the superscript, mass number). Each nucleus contains 8 protons, together with 8, 9, or 10 neutrons, in the cases of oxygen-16, oxygen-17, and oxygen-18. Oxygen 16 comprises about 99.76% of naturally occurring oxygen.

See also: atom, neutron, proton.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Isotopes in the Interstellar Medium

ISOTOPES (or isotopomers) of the same chemical element have different atomic weights. The nuclei have different numbers of neutrons, but the same number of protons. For deuterium, ^{2}H or D, the nucleus consists of a proton and neutron; for helium the two stable isotopes are ^{3}He and the much more abundant ^{4}He. NUCLEOSYNTHESIS has changed isotopic and element abundances in the big bang. Subsequently, there were further changes in STELLAR INTERIORS, NOVAE, SUPERNOVAE and COSMIC RAY interactions with interstellar (IS) material. This processing changes the abundance of light elements and isotopes; all elements heavier than lithium must have been produced in stars.

It is generally accepted that the universe began with an explosive event, the *big bang*. About 100 s after the big bang, neutrons and protons produced *primordial* D, ^{3}He, ^{4}He and ^{7}Li. In the standard big bang nucleosynthesis (SBBN) of the *light elements*, the most important parameter is the baryonic density, usually expressed as the baryon-to-photon ratio, η. After expanding and cooling, the ionized material recombined and then formed stars and galaxies. For stars of mass $M \geq 0.08 M_\odot$, nucleosynthesis products are expelled into the interstellar medium (ISM). In the solar system and perhaps in the local interstellar medium (LISM), we can measure samples directly. Accurate IS abundances cannot be deduced from cosmic ray data since the sources and interaction histories are not known in sufficient detail. For regions far from the Sun, we can measure gas phase abundances using spectroscopy, but we cannot determine the constitution of dust grains. Abundance ratios of elements in the gas phase may be affected by differing amounts of condensation onto dust grains; however, such processes will not affect the ratios of isotopes.

Measurements of ISM abundances allow us to constrain the global effect of nucleosynthesis by ensembles of stars, while abundance determinations for the circumstellar regions around individual stars allow us to investigate yields for a given object. In all stellar nucleosynthesis schemes, one must differentiate between *primary* and *secondary* products. Primary products are those produced from H or He in a short time compared with the age of the galaxy, $\sim 10^{10}$ yr. Secondaries are produced from primaries. Thus ^{12}C is a primary, while ^{13}C is a secondary. The abundance of secondaries grows with time relative to the abundance of primaries. The study of *chemical evolution* of galaxies allows us to trace the history of nucleosynthesis; models are constructed to account for the combined effects of nucleosynthesis in individual stars and the ejection of such material into the ISM as well as the conversion of ISM material into stars in the STAR FORMATION process.

In this review, we will emphasize data for selected species and sketch the measurement processes and interpretations needed to obtain abundances. The light elements deserve special treatment because of their relation to the big bang.

Results for D, He and Li

The light element abundances are especially interesting since these provide tests of big bang models and constrain baryon densities in the SBBN (figure 1). The SBBN products will be altered by stellar processing. Separating the effects of SBBN from stellar processing to determine the primordial abundances is a great challenge. We believe that stars *always* net destroy deuterium; ^{3}He and ^{7}Li may be net produced, while ^{4}He is *certainly* net produced. Measurements show that ^{4}He makes up $\geq$20% of the mass of the universe; the other big bang products constitute a mass fraction of $\leq 10^{-4}$. Our Sun has a mass fraction of $\sim$27% helium and 2% metals. Metals, designated as Z, are all elements heavier than helium. Most of the helium was produced in the big bang; when helium is produced in stars, metals are also produced (see CHEMICAL COMPOSITION OF STARS). Solar system metals were produced in earlier generations of stars, expelled into the ISM and incorporated in the protosolar nebula (see SOLAR SYSTEM: FORMATION). The solar system represents material not processed in the last 4.5×10^9 yr, while the ISM is more processed. Compared with both, high-redshift extragalactic material was not very processed since the big bang.

Deuterium

Measurements of Lyman series ultraviolet IS absorption lines in diffuse clouds toward intense background sources provide the most reliable D/H ratios. Most of the D and H in these clouds is atomic, that is D I and H I. Measurements of local gas have been obtained with the International Ultraviolet Observatory (IUE), Copernicus and Hubble Space Telescope (HST). Determinations of D/H ratios are limited by the small abundance of D relative to H and the small velocity shift caused by the different nuclear masses. The radial velocity of a D I feature must be separated from the corresponding H I feature by -82 km s^{-1} and the line widths of H and D must be related. If thermally broadened, the D I linewidth must be equal to 0.7 times that of the corresponding H I feature which is corrected for saturation; if turbulence broadened the widths must be equal. The complexities of INTERSTELLAR CHEMISTRY preclude determinations of D/H ratios from measurements of deuterated molecules such as DCO^+ or DCN. Similarly, a detection of the $\lambda = 92$ cm D I hyperfine transition in a molecular cloud will not bring information about the D/H ratio. More useful for D/H ratios might be measurements of rotational lines of the most abundant deuterated species, HD. However, one needs column densities of H_2, which can be uncertain by factors of 2 or more.

There are two interpretations of measured D/H ratios. In the first, there are significant deviations from the mean value. This interpretation is based on comparisons of H I and D I column densities for clouds along the same line of sight. In the second interpretation, there is a *single* D/H ratio. For the LISM, this is $(1.50 \pm 0.10) \times 10^{-5}$, with a 1σ error. This value and solar system ratios give lower

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

limits to the primordial D/H ratio. The optical depths of H I may be *the* limit to determinations of galactic D/H ratios. Further progress will come with the FUSE satellite, which will allow measurements of higher transitions to improve estimates of the optical depths of H I. FUSE will also extend measurements of the D/H ratios to regions of different stellar processing in the disk and halo of our galaxy and other nearby galaxies.

For clouds with REDSHIFTS of ∼2.5 or more, one can use Earth-bound optical telescopes to measure D I in the absorption line spectra of quasistellar objects. Accurate D/H ratios also require accurate H I column densities. These require accurate estimates of the optical depths of the H I lines. Large numbers of sources are needed to find unblended D I absorption, since sources with such 'clean' spectra constitute ≤2% of the total. Some observers found a ratio of $\approx 2 \times 10^{-4}$; this may be caused by blending of D I with blueshifted H I. The currently favored high-redshift ('high Z' in figure 1(*b*) and 'early universe' in table 1) ratio is $\mathrm{D/H} = (3.4 \pm 0.25) \times 10^{-5}$. This ratio is consistent with current chemical evolution models, using accepted parameters. From the SBBN relation (figure 1), this ratio allows us to estimate an upper limit to the baryon density.

Helium isotopes

The spectroscopic abundance of neutral helium is difficult to measure. The most reliable ratios are based on lines emitted from regions ionized by high-mass stars, that is regions, or gas, ionized by the extremely hot remnants of evolved stars, PLANETARY NEBULAE (PNe).

^{3}He. Outside the solar system there are two sources of data for ^{3}He/H ratios. The Solar Wind Ion Composition Spectrometer on the Ulysses spacecraft measures the abundance of ^{3}He and ^{4}He atoms. The velocities allow us to separate solar system from LISM abundances. The ratio is $^3\mathrm{He}/^4\mathrm{He} = (2.5^{+0.7}_{-0.6}) \times 10^{-4}$. The quoted statistical error is 3 times the systemic error.

The second ratio is based on radio measurements of the hyperfine transition from ^{3}He$^+$ at 3.46 cm in regions and PNe. The hyperfine line is thermalized and optically thin. There is no ^{4}He$^+$ line, since the ^{4}He$^+$ nucleus has no spin. Models which account for density structure in the ionized gas will raise the ^{3}He$^+$/H ratio; elaborate models give ratios up to a factor of 2 larger than homogeneous models.

Standard stellar models predict that $1M_\odot$ stars which are now PNe should be expelling large amounts of ^{3}He. Observations show that NGC 3242, a PNe with a $1M_\odot$ progenitor, has 100 times the ^{3}He found in the ISM. If ^{3}He were produced in *every* $1M_\odot$ star passing through the RED GIANT phase just now, the ISM should show a ratio similar to NGC 3242. The most likely process destroying ^{3}He in solar-mass stars is extra mixing in the stellar interiors during the red giant phase. This brings ^{3}He into deeper layers where temperatures are higher, enhancing destruction. To be consistent with ISM ratios, at least 70% of the ^{3}He must be destroyed before the PN stage.

In figure 2 there is a gradient in the ^{12}C/^{13}C ratio with distance from the Galactic center. If much ^{3}He were produced in stars, STELLAR EVOLUTION models predict a ^{3}He/H gradient *opposite* to the ^{12}C/^{13}C gradient. The lack of a ^{3}He/H gradient supports the contention that *not* much ^{3}He is produced in stars, so that stellar processing has little effect on the ^{3}He abundance. The average of the ISM data and Ulysses result is $^3\mathrm{He/H} = (2.0 \pm 0.7) \times 10^{-5}$. This should be primordial.

^{4}He. The present measured abundance of helium is an *upper* limit to the primordial mass ratio, y_p. Since the baryon-to-photon ratio, η, is only a weak function of the relative abundance of ^{4}He (see figure 1), even small uncertainties in y_p have *large* effects on estimates of the baryon density (figure 1). Extragalactic optical observations are best for these determinations.

Corrections of measurements to obtain primordial abundances must be empirical, since ^{4}He is produced in an ensemble of stars determined by an initial mass function, which may change with time and location. Also the star formation rate was certainly different in the past. One obtains an estimate of y_p by assuming a linear relation between helium abundance and metallicity, Z (usually the O/H abundance ratio); the traditional value of this relation is $\Delta y/\Delta Z \sim 3$. Extrapolating this relation of to zero metallicity, we can estimate y_p. Measurements have been extended to a larger group of low metallicity, blue compact galaxies (BCGs). Previous data taken have been revised because some He$^+$ spectra were affected by unsuspected line absorption and because of biases in previous fits of Y versus metallicity.

The best estimate of the primordial helium mass fraction, obtained for BCGs, is 0.242. An extrapolation to O/H = 0 gives $y_\mathrm{p} = 0.244 \pm 0.002$ while, for N/H = 0, $y_\mathrm{p} = 0.245 \pm 0.001$. From the BCG data, we adopt a value $y_\mathrm{p} = 0.244 \pm 0.002$.

Lithium

Li is very fragile, easily destroyed in deeper layers of even low-mass stars. Compared with ^{7}Li, the ^{6}Li nucleus is more loosely bound, so is more easily destroyed by nuclear burning. The accepted conclusion is that ^{6}Li is produced by collisions of cosmic rays with atoms; ^{7}Li is produced in SBBN, in spallation and in stellar nucleosynthesis.

In the ISM, the ratio of ^{7}Li/^{6}Li is free from systematic effects. Toward ζ Oph there are two velocity components in absorption with ratios of 8.6 ± 0.8 and $1.4^{+1.2}_{-0.5}$. Toward ρ Oph, the ratio is 11.1 ± 2; toward ζ Per, $5.5^{+1.3}_{-1.1}$. These ratios indicate that there are actual differences. In addition, the solar system ratio, 12.3, is significantly different from those of ζ Oph and ζ Per. This indicates that the ISM is very inhomogeneous and the solar system formed from the ISM like that toward ρ Oph.

For young low-mass stars, that is POPULATION I STARS, observations show a wide range of Li/H ratios. There is no general theory of Li processing in low-mass stars; one must use correlations based on observations. The best estimates

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of primordial Li abundances are based on observations of the older population of stars, that is POPULATION II STARS. For these stars, Li abundances are rather constant for a wide range of $T_{\rm eff}$ and are largely independent of metallicity (estimated from the abundance of iron); this has been referred to as the *Spite plateau*. The simplest explanation is that the mechanisms which destroy lithium in population I stars are absent in population II objects and that the Li in extreme population II stars is primordial. Newer results show that there is at least some Li destruction. There is also a slight increase of Li abundance with metallicity; this is the first direct evidence for stellar Li production. The average observed Li/H value, for upper and lower envelopes, is $(1.16^{+0.6}_{-0.3}) \times 10^{-10}$. If there are unsuspected sources of destruction of Li, one can fit to the *upper* envelope to obtain the ratio $\mathrm{Li_p/H} = (2.5^{+0.7}_{-0.5}) \times 10^{-10}$ (see figure 1(*b*)).

Summary and interpretation of light element abundances

We plot the best estimates of primordial abundances versus the baryon-to-photon ratio η_{10} in figure 1(*b*). For consistency, all the primordial abundances must give a single value of η_{10}. From the data, there is a consistent model for $\eta_{10} = 4.7$. One must use the ^{7}Li/H ratio determined under the assumption that the upper envelope of the measurements is the best choice. The D/H value measured for high-redshift objects must be 3.4×10^{-5}, and the ^{3}He/H ratio must be nearly at the lower limit. The most critical test of SBBN rests with the ^{4}He/H ratio; if the BCG ratios were lowered by 10%, there would be an inconsistency with the D/H ratio.

The best value of η_{10} is 4.7. Because systematic errors dominate, we can only give a range of errors; η_{10} is between 2 and 5. For $\eta_{10} = 4.7$ and a microwave background temperature of 2.73 K, the *baryonic* mass density is 3.2×10^{-31} g cm^{-3}. For a HUBBLE CONSTANT of 65 km s^{-1} Mpc^{-1}, the density needed to halt the expansion of the universe is 25 times larger. Thus, baryons alone cannot halt this expansion.

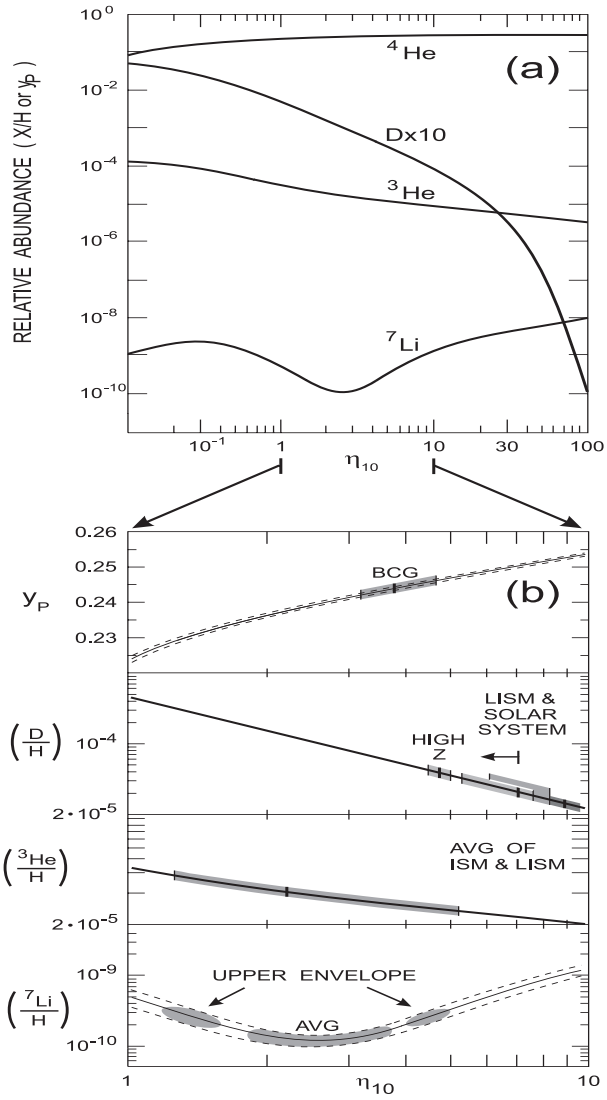


Figure 1. (*a*) The SBBN production scheme as a function of η_{10}, which is 10^{10} times the baryon-to-photon ratio. The number ratios, X, refer to D, ^{3}He and ^{7}Li, the mass fraction, y, refers to ^{4}He. The curve for D is multiplied by 10 for clarity. (*b*) An expanded version of (*a*), showing the region of parameter space where the abundances of the elements are located. The shaded regions show where the data in table 1 are located. In the panel for D/H ratios, the leftward arrow indicates that the solar system and LISM ratios are *lower* limits to the primordial values. For D/H, 'high Z' refers to gas toward high-redshift quasars; the values for lower D/H ratios are from the solar system and LISM. For $y_{\rm p}$, the BCGs are blue compact galaxies. BCGs are essentially very large regions with low metallicity; thus their y values should be close to $y_{\rm p}$. The dashed curves show the uncertainties in the theoretical relations.

Results for CNO isotopes

Isotope shifts of atomic spectra, caused by different nuclear masses, are sufficiently large for light elements, but not for heavier elements. These must be measured by molecular spectroscopy. A comparison of isotopomers might be affected by *selective dissociation* and *chemical fractionation*. Selective dissociation occurs when molecules are destroyed by spectral line radiation; the more abundant species may be self-shielded. Chemical fractionation arises from small differences in the different molecular zero-point vibration energies, related to mass and to the temperature of formation. The result is to enrich the heavier molecular species. If this were a large effect, the measured ^{12}CO/^{13}CO ratio would not represent the ^{12}C/^{13}C ratio. The most reliable carbon isotope ratios are given by CH^+ measurements, because CH^+ is formed only at high temperatures. CH^+ lines are measured in the optical, so these ratios are restricted to the LISM. The best $^{12}CH^+/^{13}CH^+$ ratio is 67 ± 3. From radio astronomy data, the average of 13 sources in the LISM is ^{12}C/^{13}C $= 69 \pm 6$ (table 2). The agreement of the CH^+ isotope ratio with these carbon isotope ratios from RADIO ASTRONOMY indicates that the combined effects of fractionation and selective

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Selected light element ratios.

Abundance ratio	Protosolar	Nearby ISM	Early universe, BCGs or population II stars
$(D/H) \times 10^5$	2.4 ± 0.5	1.50 ± 0.10	3.4 ± 0.25
$(^3He/^4He) \times 10^4$	1.5 ± 0.3	2.0 ± 0.7	–
Y	0.273–0.277	0.260 ± 0.018	0.244 ± 0.002
$(^7Li/H) \times 10^{10}$	–	–	$1.16^{+0.6}_{-0.3}$ (average of all) $2.5^{+0.6}_{-0.5}$ (upper envelope)
$^7Li/^6Li$	12.3	6.2; 1.4 to 11.1	–

dissociation are small. On this basis, radio observations of CO and H_2CO have been used to determine the $^{12}C/^{13}C$ ratio over the Galaxy.

Nitrogen and oxygen are not affected by chemical fractionation. However, determinations of $^{14}N/^{15}N$ or $^{16}O/^{18}O$ ratios may be influenced, since these ratios are frequently measured as double ratios, $H^{12}C^{15}N$ and $H^{13}C^{14}N$. Then, one uses ratios from CO and H_2CO to obtain the desired results.

Carbon

In figure 2(*a*) we plot the ISM isotope ratio $^{12}C/^{13}C$ determined from H_2CO and CO as a function of D_{GC}. These results show that nearly all the ratios obtained from H_2CO lie above those from CO. This is consistent with interstellar chemistry models. Molecular clouds near the Sun show a higher ratio than clouds in the inner part of the disk of our Galaxy, and the H_2CO ratios have a larger scatter than ratios obtained from CO data. A least-squares unweighted fit to all of the data gives

$$^{12}C/^{13}C = (7.5 \pm 1.9)\, D_{GC} + (7.6 \pm 12.9).$$

The uncertainties are RMS values. There is a large scatter between the data and the fit. Some of this is due to errors, but there is evidence for actual source-to-source differences. If the solar system ratio, 89, represents the LISM 4.5×10^9 yr ago, the ratio has decreased by ∼20%. Taking limits as actual values, optical studies of 191 circumstellar envelopes give an average ratio of 18; radio data (table 2) gives an average of 40. The stellar production of ^{13}C is diluted by a factor of >2 in the LISM.

The $^{12}C/^{13}C$ ratio is a measure of the primary to secondary processing. The ^{12}C is thought to be produced on rapid time scales primarily via He burning in massive stars. The ^{13}C is thought to be produced primarily via CNO processing of ^{12}C seeds from earlier stellar generations. This occurs on a slower time scale primarily during the red giant phase in low- and intermediate-mass stars or novae. The $^{12}C/^{13}C$ ratio for the solar system is 89 (table 2). As a function of time, ^{13}C is expected to increase with respect to ^{12}C, as with all secondary products. Thus, the ISM $^{12}C/^{13}C$ ratio should be lower than the solar system value (table 2); this ratio is only about 20% above the isotope ratio in the LISM. Thus, over the last 4.5×10^9 yr, lower-mass stars have not produced a large amount of ^{13}C in the LISM.

Nitrogen isotopes

To obtain $^{14}N/^{15}N$ ratios from $H^{13}CN$ and $HC^{15}N$ data, $^{12}C/^{13}C$ ratios must be known. We take these from the fit to figure 2(*a*). Applying these results to the HCN data, we obtain the $^{14}N/^{15}N$ ratios plotted in figure 2(*b*):

$$^{14}N/^{15}N = (19.0 \pm 8.5)\, D_{GC} + (288.6 \pm 65.1).$$

The $^{14}N/^{15}N$ ratio in the solar system is 270. Near the Sun, the average is 388 ± 32 (eight sources). From the fit, the value at 8.5 kpc is 450. Both values are consistent with an ∼70% increase in this ratio since the formation of the solar system. The Galactic center region and the CARBON STAR IRC+10216 show high metallicities and $^{14}N/^{15}N$ ratios >500 (table 2). This gives support to the notion that $^{14}N/^{15}N$ ratios are directly correlated with metallicity. If these regions are typical, one expects an increase of $^{14}N/^{15}N$ ratios in the inner Galactic disk. However, the data show a somewhat *lower* value (table 2). Studies of H II REGIONS in the inner part of the Galactic disk have a higher metallicity. Thus, the chemical evolution of ^{14}N and ^{15}N in the inner disk must have been *very* different from that of the Galactic center, the well-known carbon star CW Leo (IRC+10216) or the LISM. These results can be explained if novae are the main contributors to ^{15}N. The $^{14}N/^{15}N$ ratio in the LMC is ∼100. This galaxy is metal poor; this result indicates that the production of ^{15}N is primary and not secondary.

Oxygen

There are three stable oxygen isotopes. Usually, one obtains the $^{16}O/^{18}O$ ratio from a double ratio; for figure 2(*c*), we used $(H_2^{12}C^{18}O)/(H_2^{13}C^{16}O)$. To obtain a ratio involving only oxygen, one must use the carbon isotope ratios (figure 2(*a*)) to separate the contributions of carbon from oxygen. The isotope values are plotted in figure 2(*c*). The fit gives

$$^{16}O/^{18}O = (58.8 \pm 11.8)\, D_{GC} + (37.1 \pm 82.6).$$

The fit is surprisingly good. For five sources in the LISM, the average ratio is 557 ± 30 (table 1). From the fit, the value at 8.5 kpc is 537, with large uncertainties.

Ratios of ^{18}O to ^{17}O show no gradient with D_{GC}. The average value of this ratio is markedly below those

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Radio data for isotope ratios of heavier elements.

Isotope ratio	Galactic center	Inner Galactic disk	LISM	Solar system	Carbon stars	Nuclei of galaxies
$^{12}C/^{13}C$	∼20	53 ± 4	69 ± 6	89	>30	≥30
$^{14}N/^{15}N$	>600	320 ± 30	388 ± 32	270	>515	–
$^{16}O/^{18}O$	250	327 ± 32	557 ± 30	490	320–1260 >2700	∼200
$^{18}O/^{17}O$	3.5 ± 0.2	3.6 ± 0.2	3.6 ± 0.2	5.5	0.6–0.9 < 1	8
$^{32}S/^{34}S$	∼22	∼22	∼22	22	–	–
$^{29}Si/^{30}Si$	1.5	1.5	1.5	1.5	–	–

of the solar system, but is larger than the carbon star value. The average $^{16}O/^{18}O$ ratio in the ISM near the Sun is comparable with that found in carbon stars, to within the errors. In the inner Galactic disk, the ratio is a factor of 2 lower; the value in the Galactic center is comparable. If the isotopes are formed in stars more massive than $10M_{\odot}$, then ^{16}O is a primary nucleosynthesis product, while ^{17}O is secondary. The major production scheme for ^{18}O is not certain. One expects a radial gradient in a primary/secondary ratio but not in a secondary/secondary ratio. The gradient observed in $^{16}O/^{18}O$ combined with the lack of a gradient in $^{18}O/^{17}O$ indicates that ^{18}O is secondary. However, the observed $^{16}O/^{18}O$ ratio for the LISM, 557 ± 30 (table 2), is larger than the solar system value, 490. This indicates an *increase* with time, whereas primary/secondary ratios should *decrease* with time. Thus, this model cannot reproduce the evolution of either ratio, from comparisons of solar system and LISM ratios. To explain both the temporal and the spatial behavior of $^{16}O/^{18}O$, one must assume that there is an unknown mechanism which has inhibited the ^{18}O production in the last few 10^9 yr. Since the nucleosynthesis of ^{18}O is still not well understood, this is possible. As expected for a purely secondary/secondary ratio, there is no variation of $^{18}O/^{17}O$ with D_{GC}. Observations show a marked decrease in the $^{18}O/^{17}O$ ratio between the formation of the solar system and the present. So far, models have been based on the assumption that only massive stars produce ^{17}O. If ^{17}O is produced in low- and intermediate-mass stars, the $^{18}O/^{17}O$ ratio is a secondary/secondary result, with ^{17}O having a longer production time scale, i.e. more secondary than ^{18}O. Then in the past few 10^9 yr, the abundance of ^{17}O will grow faster than ^{18}O; this explains the decrease of the ratio with time. It is also possible that the solar system is not a representative sample of the ISM 4.5×10^9 yr ago.

Sulfur, silicon, magnesium and aluminum isotopes

For other isotopes, there are fewer data, as compared with CNO results. There is little or no variation in the $^{32}S/^{34}S$ ratio with D_{GC}. The ISM ratio agrees with the solar system value, so there seems to have been no change with time. Additional data show a $^{32}S/^{34}S$ ratio of 24.4 ± 5.0, a $^{34}S/^{33}S$ ratio of 6.27 ± 1.01 and $^{34}S/^{36}S$ of 115 ± 17. The last ratio is lower than the solar system value of 200, so ^{36}S may be a secondary product. This is expected on the basis of chemical evolution models and indicates good agreement. We take this (optimistically) as evidence for the correctness of both.

In the solar system, the $^{28}Si/^{29}Si$ ratio is 19.7. For some sources it is certain that the optical depths of ^{28}SiO lines are moderate, so there can be no direct determination of the $^{28}Si/^{29}Si$ from measurements of the rotational lines of SiO isotopomers. The isotopes ^{29}Si and ^{30}Si have been measured in the Galactic center region. To within the errors, the ratio agrees with that measured for the solar system. Thus, there appears to be no gradient in the $^{29}Si/^{30}Si$ ratio with D_{GC}.

Radioactive nuclei which decay by γ ray emission can be detected with satellites. These data refer to heavier elements ejected into the ISM by larger-mass stars, novae or supernovae. The most well-studied isotope is ^{26}Al; this might be produced during the red giant phase of lower-mass stars.

The presence of ^{26}Al in the ISM is detected by the γ ray emission at 1.806 MeV, from the decay of ^{26}Al to the first excited state of ^{26}Mg. The mean half-life is 1.05×10^6 yr. If supernovae produced *all* of the ^{26}Al, there would be an overproduction of the stable isotope ^{27}Al. It seems likely that ^{26}Al is formed in red giant stars.

An excess of solar system ^{26}Mg in meteoritic samples had been interpreted as evidence that an enrichment of ^{26}Al occurred shortly before the formation of the solar nebula. This may have been a result of triggered formation. A less likely interpretation is that ^{26}Al was produced elsewhere and transported here. The rare stable isotopes of magnesium, ^{25}Mg and ^{26}Mg, have been measured in IRC+10216 using the molecule MgNC. The ratio of ^{24}Mg:^{25}Mg:^{26}Mg was found to be 78:11:11, with a quoted error of ± 1. This result agrees with the terrestrial ratio of 79.0:10.0:11.0.

Summary of isotopomers of heavier elements

There is excellent evidence for a gradient in the $^{12}C/^{13}C$ ratio with distance from the center of the Galaxy, D_{GC}. There is evidence for source-to-source variations; for some ratios, the solar system value does not represent the average of the LISM at the time of formation. There is also

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

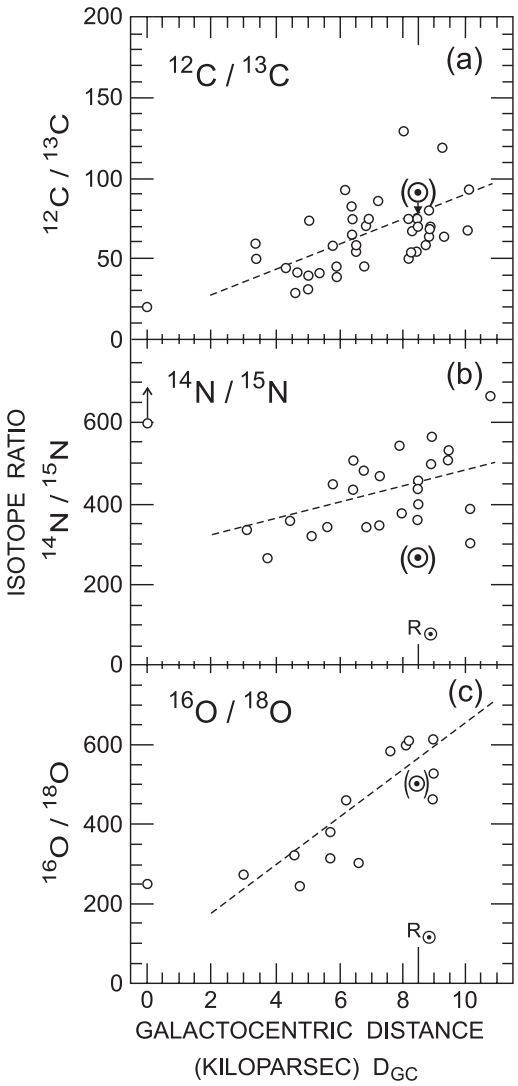


Figure 2. Plots of isotope ratios as a function of distance from the Galactic center, D_{GC}. The distance of the Sun from the center is 8.5 kpc. The lines are unweighted least-squares fits to the data. For carbon and oxygen ratios, the solar system values are shown as (⊙); for nitrogen, the value is the tip of the arrow. (*a*) Data for the $^{12}C/^{13}C$ ratio, from CO and H_2CO. The ratios from CO data lie below those from H_2CO. (*b*) Data for the $^{14}N/^{15}N$ ratio, based mostly on double isotope ratios in the HCN molecule. (*c*) Data for the $^{16}O/^{18}O$ ratio, based on double isotope ratios in the H_2CO molecule. The value at D_{GC} is the average for the galactic center; the $^{14}N/^{15}N$ value is a lower limit.

evidence for a gradient in the $^{16}O/^{18}O$ ratio and reasonable evidence for a $^{14}N/^{15}N$ gradient with D_{GC}. The values of the oxygen ratio near the Sun and the nitrogen ratio in the inner Galaxy cannot be explained on the basis of current theories. Isotopomer studies of the nuclei of external galaxies are just beginning. These results can be important for an understanding of the star formation history in such objects.

Bibliography

The study of isotopomers is driven by data, so the most recent collections are the most useful. The results of a meeting held at the International Space Science Institute, Bern, were used extensively. These proceedings are contained in

Prantzos N, Tosi M and von Steiger R (eds) 1998 *Primordial Nuclei and their Galactic Evolution* (Dordrecht: Kluwer)

Prantzos N and Diehl R 1996 *Phys. Rep.* **267** 1

Studies of ^{26}Al and related topics are in a detailed recent review of isotopes is

Wilson T L 1999 *Rep. Prog. Phys.* **62** 145

A review of the light element abundances is contained in

Reeves H 1994 *Rev. Mod. Phys.* **66** 193

A compendium of solar system isotopomer data is given by

Grevesse N *et al* 19?? *Cosmic Abundances (Astronomical Society of the Pacific Conf. Ser. 99)* ed S S Holt and G Sonneborn (San Francisco: Astron. Soc. Pacific)

Solar models with estimates of the protosolar helium mass fraction are given by

Brun *et al* 1998 *Astrophys. J.* **506** 913

Circumstellar $^{12}C/^{13}C$ ratios are collected in

Charbonnel C and do Nascimento J D 1998 *Astron. Astrophys.* **336** 915

Thomas L Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Isotopic Anomalies

Until the early 1970s, all models of solar system evolution assumed that the inner nebula underwent a high-temperature stage, destroying all grains and thus erasing the entire presolar record: all solid bodies would have accreted from the cooling of this nebula, constituting a gaseous initial reservoir essentially homogeneous in composition (see SOLAR SYSTEM: FORMATION).

This basic assumption was based on the observation that in all available material, terrestrial or extraterrestrial (lunar samples, meteorites), every element was measured with a fixed isotopic composition. Actually, some tiny differences were observed, but could be explained as resulting from known processes. Let us consider the case of the oxygen isotopes. Oxygen has three stable isotopes, ^{16}O , ^{17}O and ^{18}O , with abundance ratios 99.76%, 0.04% and 0.2% respectively. If for example one compares the O isotopic ratios in water vapor and the liquid water from which it evaporated, one will find a systematic enrichment in light oxygen isotopes (^{16}O), and accordingly a depletion in heavier isotopes (^{17}O and ^{18}O), in the gas phase (the reverse being observed in the liquid). Moreover, this mass-dependent 'fractionation' process is such that, with respect to ^{16}O, the depletion of ^{18}O is 2 times higher than that of ^{17}O. Thus, on a diagram where the $^{17}O/^{16}O$ variation, with respect to a given reference, is plotted against the $^{18}O/^{16}O$ variation, all samples fall on a line ('fractionation line') with slope 1/2, which represents the ratio of mass differences: (17–16)/(18–16) (see figure 1).

In 1969, two major events modified this view. First, the return of lunar samples from the APOLLO missions had forced all experimenters to drastically increase the sensitivity of their analytic tools, to allow measurements to be performed at a microscopic grain scale. Second, a very large 'primitive' meteorite, a carbonaceous CHONDRITE called the ALLENDE METEORITE, fell and was retrieved prior to any contamination. This meteorite contains millimeter-sized white inclusions imbedded in a dark silicate matrix. These inclusions are made of highly refractory minerals, that is very-high-temperature condensates. Although very difficult to measure on such a small scale, their O isotope contents were determined. Very surprisingly, they happened to be quite distinct from any other ever measured, including that of the Allende bulk matrix. Actually, they did not plot on the oxygen fractionation line, but rather on a line with slope 1, the proposed interpretation of which was as follows: these inclusions are mixtures of 'normal' grains ('A' on figure 1), condensed from the same isotopic reservoir as all other solar system material, and grains ('B') with oxygen atoms constituted of ^{16}O only (no ^{17}O and no ^{18}O). There is no known process, operating in the solar region, capable of producing a sample free of ^{17}O and ^{18}O if it contained originally a mixture of the three stable isotopes. This implies that the grains with pure ^{16}O condensed prior to their being incorporated in the solar nebula, in an environment where no ^{17}O and no ^{18}O was present, and have been preserved

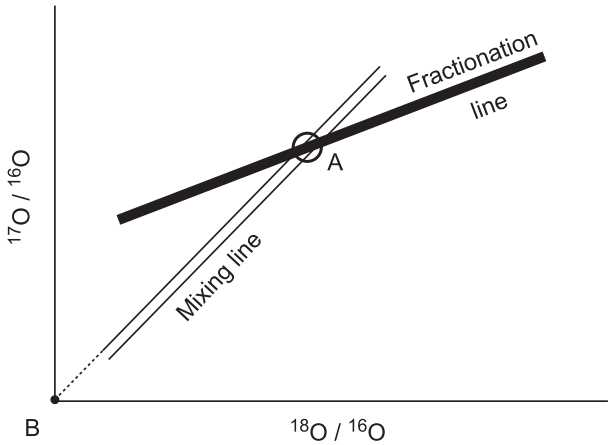


Figure 1. Isotropic diagram for oxygen. (After Clayton R N *et al* 1973 *Science* **182** 485–8 (© AAAS 1973).)

since. A major consequence was derived: the thermal evolution of the early solar system did not entirely erase the memory of presolar events; it is possible to identify grains with isotopic content indicating their presolar origin. Thus, the ^{16}O-rich phase could have been formed in a supernova envelope.

'Isotopic anomalies' refer to such abundance ratios which cannot be accounted for by solar system processes acting on samples accreted in an initial homogeneous reservoir of 'solar' composition. They are relics of stellar environments where presolar grains formed prior to their incorporation in the protosolar nebula, preserving their original isotopic composition until their accretion in solar system objects.

Another important isotopic anomaly, discovered soon after the pioneering O anomaly, is that of Mg. Mg has three stable isotopes, ^{24}Mg, ^{25}Mg and ^{26}Mg, with relative abundances 79%, 10% and 11%. In some carbonaceous chondrites, the Mg isotopes have been measured with a slight ^{26}Mg enrichment, found to be proportional to their $^{27}Al/^{24}Mg$ ratio. The interpretation is as follows: Al contains at present only one stable isotope, ^{27}Al. However, models of STELLAR EVOLUTION show that a number of stars, in particular SUPERNOVAE, also synthesize some ^{26}Al, an isotope which decays into ^{26}Mg with a half-life of 0.73 million years. The excess of ^{26}Mg in a grain would result from the *in situ* decay of ^{26}Al. The discovery of excess ^{26}Mg in a grain would thus indicate that it condensed in a gas containing both ^{27}Al and ^{26}Al.

The presence of ^{26}Al in the protosolar nebula would imply that the time between the synthesis of ^{26}Al, in a stellar environment, and its incorporation in the solar region and within the parent bodies of meteorites, is no more than a few million years: if not, all would have decayed, been homogenized in the gaseous nebula, and no anomalies would have been observed. In this very short time scale, considerably shorter than that of interstellar cloud collapse, a supernova would have exploded very

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

close to the protosolar nebula, loading it with unstable isotopes such as ^{26}Al. According to one model, this supernova, by its explosion, would have both driven the collapse of the presolar cloud, and injected some isotopes the decay of which could have served as the major energy source to initiate planetary evolution.

Actually, these pioneering and fundamental investigations triggered a huge development of experimental work, leading to the discovery of many more isotopic anomalies, sometimes identified at the level of individual micrometer-sized grains: indeed, one can state that interstellar grains have been isolated in the laboratory. Different classes of such presolar grains, extracted from all major types of carbonaceous chondrites, have been identified. They are all refractory enough in composition to have survived the inner solar system formation processes: so far, nanodiamonds (a few nm in size), SiC, graphite and corundum are the most numerous.

When measured at grain level, much larger isotopic anomalies are found, with respect to the first analyses, which were performed on mixtures of presolar and solar grains. Moreover, in addition to ^{26}Al, other short-lived isotopes have been discovered through excess of their daughter products: ^{41}K (decayed from ^{41}Ca, with a half-life of 0.1 million years) and ^{44}Ca (from ^{44}Ti, with a half-life of only 48 yr). Although ^{26}Al and ^{41}Ca could have been formed in different types of stars, such as supernovae, RED GIANT STARS (actually asymptotic giant branch) or WOLF–RAYET STARS, according to current nucleosynthesis models, ^{44}Ti seems to originate only from the explosive supernovae burning He: their carriers, when identified, are truly grains condensed in a supernova SN envelope, and preserved since.

These examples illustrate the potential that a systematic study of isotopic anomalies has to describe the geography of the presolar stellar environment and the history of solar system formation from the collapse of the interstellar cloud to the accretion of protoplanetary bodies. It emphasizes the major role that microscopic analyses of extraterrestrial samples can play in astrophysics.

Jean-Pierre Bibring

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Istituto di Radioastronomia, Bologna

The Istituto di Radioastronomia in Bologna is the main institute for research in radio astronomy in Italy. It conducts the study of the universe through radio waves and operates the Italian radio telescopes.

The institute, founded in 1970 from a previously existing university laboratory, is operated and funded by the Consiglio Nazionale delle Ricerche (CNR). The headquarters are located in Bologna, at the CNR Research Area, while the two observing stations are in Medicina (Bologna) and Noto (Siracusa, Sicily).

The staff consists of about 55 people, including astronomers, electronic engineers–physicists, software specialists, technical and administrative personnel. Additionally, 10 professors and researchers from the University of Bologna and the Observatory of Bologna, as well as many PhD students, post-doctorate fellows and undergraduate students, are affiliated with the institute.

The institute operates three radio telescopes. The 'Croce del Nord' radio telescope, in Medicina, is the largest transit instrument in the world, with its T-shaped arms of 564 m and 640 m, respectively. Since it was built, in the early 1960s, it has produced two important radio source catalogs (known as B2 and B3), with data on more than 30 000 objects.

The two identical single-dish antennas, each 32 m in diameter, located in Medicina and Noto, built in the 1980s, have been designed to work in conjunction with other foreign telescopes with the very long baseline interferometry technique (VLBI). They are members of European networks for both astronomy and geodynamics, and participate in observations with radio telescopes belonging to the American network (VLBA), to other countries in the world, and in space (HALCA). The institute is member of the JOINT INSTITUTE FOR VLBI IN EUROPE (JIVE).

A new large single-dish radio telescope, 64 m in diameter, is being built in Sardinia to increase the radio observing facilities, and to track interplanetary probes.

The research activity of the institute is the study of the physics of radio galaxies and quasars, stars, the interstellar medium and cosmology. The institute has gained a high level of expertise in the technological field related to the instrumentation required for large radio telescopes.

For further information see
http://www.ira.bo.cnr.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Italian Space Agency (ASI)

A government agency, founded in 1988, for the purpose of identifying, coordinating and managing Italian space programs. Under its five-year National Space Plan, ASI aims to promote, support and control the scientific, technological and commercial applications of space activities as well as to promote new technological capabilities in Italian aerospace industries.

Under ASI's leadership, Italy has been involved in a variety of space activities, including the International Space Station, Tethered Satellite, X-SAR (X-band Synthetic Aperture Radar) and Lageos (Laser-ranged geodetic satellite) programs. Cooperative scientific ventures include the Beppo-Sax x-ray satellite and the SAC-B satellite for the study of solar physics.

ASI also promotes and supports Italian scientific and industrial participation in EUROPEAN SPACE AGENCY programs, including Cassini–Huygens and Integral.

For further information see
http://www.asi.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ivory, Sir James (1765–1842)

Mathematician, born in Dundee, Scotland. After a business career he became professor of mathematics at the Royal Military College, Great Marlow (which became Sandhurst). He retired early because of ill-health and devoted himself to mathematical research, particularly as applied to astronomical problems such as the gravitational attraction of ellipsoids, the shape of self-gravitating rotating fluid bodies, the orbits of comets and atmospheric refraction.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

James Clerk Maxwell Telescope

The 15 m diameter James Clark Maxwell Telescope (JCMT), the largest facility in the world designed specifically to operate in the submillimeter region of the spectrum, became operational in 1987. It is situated close to the summit of Mauna Kea, Hawaii, at an altitude of 4092 m.

The JCMT is operated by the Joint Astronomy Center in Hilo, Hawaii, on behalf of PPARC in the United Kingdom, the National Research Council of Canada and The Netherlands Organization for Scientific Research.

It is used to study a wide range of astronomical objects, including the Sun, comets, molecular clouds, galaxies, quasars and cosmic background radiation. On 11 July 1991, JCMT became the largest telescope ever to observe a total solar eclipse.

Its performance has been enhanced by the addition of SCUBA—the submillimeter common-user bolometer array—a camera and photometer which operates at millimeter and submillimeter wavelengths. SCUBA is cooled to 0.1 K and is the most sensitive and versatile of a new generation of imaging devices for the submillimeter region. It has two arrays of bolometric detectors. The long-wave array operates in the 750 and 850 μm atmospheric transmission windows, while the short-wave array is used for observations at 350 and 450 μm.

SCUBA, which was designed and constructed by the ROYAL OBSERVATORY IN EDINBURGH in collaboration with Queen Mary and Westfield College, London, began observations in May 1997. Since then, discoveries include detection of a ring of dust particles around Epsilon Eridani; a distant galaxy hidden in dust; and possible new solar systems in formation around Vega and Fomalhaut.

For further information see
http://www.jach.hawaii.edu/JCMT/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jansky, Karl Guthe (1905–50)

American radio engineer, founder of radioastronomy, employed by Bell Telephone Laboratories to reduce noise ('static') in shortwave transatlantic radiotelephony. Using a crudely directional, rotating shortwave receiver array operating at 20.5 MHz and 45 kHz, and nicknamed the Merry-Go-Round, he noticed a continuous interference which changed its strength over the course of the day. At first, he believed the noise was caused by the Sun, but it transpired that the time of the peak was shifting over the course of the year, and a partial solar eclipse in 1932 did not reduce the signal. He realized that the noise had a period of a sidereal day, not a solar day, and was associated with the Milky Way, particularly the direction of Sagittarius, where the Milky Way was brightest. This static being outside the control of radio engineering, he moved on to other work. In 1973 the radioastronomical unit of flux density (10^{-26} watts per square meter per hertz) was named the *jansky* in his honor.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Janssen, Pierre Jules César (1824–1907)

Solar astronomer, born in Paris, founded and became head of the Astrophysical Observatory at Meudon, observed the bright line spectrum of a solar prominence in the total eclipse of 1868. The brightness of the emission lines suggested to him that they should be visible even outside an eclipse, and so they were (*see also* LOCKYER). Photographed the Sun (published in 1904 his *Atlas de Photographies Solaries*), discovered solar granulation (*see also* NASMYTH). 'Once celestial photography was realized, the French astronomers divided the Universe among themselves: Janssen took the Sun, Loewy and Puiseux the Moon, and the brothers HENRY the stars'. Established a solar observatory on Mont Blanc.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Japanese Astronomy

Astronomy in Japan has made rapid progress in the last two decades, placing Japan among the leaders in the field. Not only mm wave and x-ray astronomy since the 1980s but also the optical and infrared, gravitational wave and neutrino astronomy are now at the highest level in these fields. Astronomers are preparing ambitious future plans in both ground-based and space-borne astronomy, in spite of the many challenges that basic science in Japan faces.

Historical view of Japanese astronomy

Historically, astronomical research in Japan contributed little to the human understanding of the universe before World War II. In the Meiji era (1868–1912), Japan started introducing science and technology to catch up with the Western world. The government established faculties of engineering in universities as one of the earliest of such attempts in the world. However, most of the systematic and organized activities of basic sciences, including astronomy, started only after World War II.

The Norikura Solar Corona Observatory, started in 1950, and the OKAYAMA ASTROPHYSICAL OBSERVATORY with a 188 cm optical telescope (1960) of Tokyo Astronomical Observatory (TAO) and other observatories produced only a few observational results of the highest quality during the period of the 1950s to 1970s. However, these observatories established a foundation for observational astronomy in Japan. The TAO was then with the University of Tokyo and reorganized in 1988 as an interuniversity research institute, the NATIONAL ASTRONOMICAL OBSERVATORY, JAPAN (NAOJ). The progress in theoretical research in Japan, on the other hand, was remarkable in this period. The brilliant works by Chushiro Hayashi on the early universe, stellar structure and origin of the solar system, Yoshihide Kozai in celestial mechanics and Satio Hayakawa in high-energy astrophysics were among memorable contributions to the theoretical astronomy and astrophysics of humankind. In particular, the systematic works on the formation of stars and the solar system by C Hayashi and his group had built a firm foundation in the field.

In the 1960s small but extensive attempts started in radio, x-ray and infrared astronomy in Japan and grew rapidly throughout in the 1970s. It was in the 1980s that those efforts were rewarded with remarkable scientific and technological results.

Developments in radio astronomy

Solar radio observations in Japan started in the 1950s. The development of solar microwave interferometers by Haruo Tanaka was among the earliest in the field. The millimeter wave observations in Japan started in 1970 with a 6 m diameter telescope at Mitaka constructed by Kenji Akabane, Masaki Morimoto and their group. Their efforts led TAO to establish the NOBEYAMA RADIO OBSERVATORY (NRO) in 1982 with two top-level mm wave telescopes, the 45 m diameter telescope (the world's largest mm wave telescope until now) and the mm wave interferometer composed of five (a sixth added later) 10 m antennas with 570 m baseline. Norio Kaifu, Masato Ishiguro and their group led the construction and work on related developments such as high-precision antennas, acousto-optical radio spectrometer, Fourier correlation radio spectrometer, high-sensitivity SIS mixer receivers, closely cooperating with engineers from Japanese industries. Since its establishment the Nobeyama mm wave telescopes have been providing rich scientific results: detection of many exotic new molecules in interstellar clouds, protoplanetary disks, evidence of supermassive black holes in galactic nuclei, to name a few, making Japan one of the leading countries in radio astronomy.

Recently (1998) NRO succeeded in linking the 45 m telescope with the array of six 10 m diameter antennas for a seven-element mm wave interferometer called RAINBOW. It will be the most sensitive mm wave interferometer for the time before the planned international project ALMA (mentioned below). Another facility of NRO is a radioheliograph, completed in 1992 under the leadership of Shinzo Enome. The 84-element array of 80 cm diameter parabolas yields high-resolution radio images of the sun every second.

NRO and Mizusawa Astrogeodesy Observatory of NAOJ had started the VLBI observations with other Japanese institutes and organized a domestic VLBI network (J-Net). In 1997 the VLBI group led by Hisashi Hirabayashi and his group launched the world's first space VLBI satellite HARUKA as a collaboration among ISAS, NAOJ and overseas institutes. The launched 8 m diameter antenna is orbiting around the Earth to link with radio telescopes worldwide to form a 30 000 km aperture radio telescope (VSOP). It was an engineering and logistical challenge, and recently VSOP has produced clear images of many energetic galactic nuclei with extremely high resolution of 0.1–1 marcsec, approaching a linear size about 100 times that of supermassive black holes.

NRO, with its large mm wave telescopes, was the first top-level observational facility Japan had for astronomy. In the field of mm wave and sub-mm wave astronomy, university groups are now undertaking a number of small but excellent projects: a 4 m diameter mm wave telescope of Nagoya University (located at Las Campanas, Chile), a 60 cm CO molecular line mapping telescope of the University of Tokyo (La Silla, Chile), a 1.2 m remote-operation sub-mm carbon atom line survey telescope of the University of Tokyo (atop Mt Fuji) etc, yielding a large amount of observational data for specific scientific subjects.

Developments in x-ray astronomy

HAKUCHO, the first Japanese x-ray satellite, was launched in 1979 by the INSTITUTE OF SPACE AND ASTRONAUTICAL SCIENCE (ISAS) under the leadership of Minoru Oda. Subsequent series of ISAS x-ray satellites, TENMA (1983), GINGA (1989) and ASUKA (1993) under the leadership of Yasuo Tanaka and his group, with their rich scientific

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

results made Japanese x-ray astronomy one of the most productive in the world. Since the early identification of the x-ray source Sco-X1 with an evolved star using the balloon-borne x-ray telescope and the Okayama 188 cm telescope, efforts by ISAS and groups in universities in parallel with developments in space engineering have revealed the physics of dying stars and their energetic activities. ASUKA is a relatively small satellite (420 kg) but is actively continuing x-ray spectroscopic observations of various objects from star formation sites to clusters of galaxies in the very distant universe.

Solar x-ray observations have also been made by a successful collaboration of x-ray astronomy at ISAS and traditional solar physics at TAO/NAOJ. YOKO, launched in 1991 following the first Japanese solar x-ray satellite HINOTORI (1981), has been providing a tremendous number of x-ray images of the Sun and contributed to understanding of the physical process of energetic solar flares (project leaders: Yoshiaki Ogawara and Yutaka Uchida).

Recently ISAS developed the M-V rocket with a payload capability much higher than that of previous L-series rockets. The first launch of M-V was celebrated by the success of HARUKA/VSOP mentioned earlier. By using M-V, new astronomical projects such as Astro-E, IRIS and Solar-B are being promoted. Astro-E is a 1650 kg mass satellite with very high x-ray spectroscopic and imaging capabilities, to be launched in early 2000. Solar-B, a next-generation HINOTORI, is a 50 cm aperture optical solar telescope slated for 2004. IRIS is described below.

Developments in optical and IR astronomy

An 8.2 m aperture optical and IR telescope 'Subaru' atop Mauna Kea, Hawaii, started its test observations in early 1999 and will be opened to the astronomical community in late 2000. It is one of the new-generation telescopes with a large active-support primary mirror housed in a turbulence-suppressed cylindrical enclosure and has seven observational instruments that cover ultraviolet to mid-infrared ranges. Its high performance was demonstrated by stellar images with a full width at half-maximum of 0.2 arcsec and a number of spectacular scientific results through the test observation period. The construction of the SUBARU TELESCOPE, led by Keiichi Kodaira and later by Norio Kaifu and their group, started in 1991 as a 9 yr project of optical and IR astronomers of Japan.

Experience of Japanese optical astronomers has been mostly limited to the Okayama 188 cm telescope built in 1960 and the Kiso 105 cm Schmidt camera built in 1974. Infrared astronomers studied through the pioneering construction of the Agematsu 1 m IR telescope (1973) and observations with overseas telescopes led by Haruyuki Okuda and Shuji Sato. To build an engineering and technical basis for optical and infrared astronomy, the Advanced Technology Center was established at Mitaka campus of NAOJ headquarters. Close cooperation with university groups and industries was also organized for preparation, basic design and construction of the telescope and observing instruments. To overcome the lack of middle- to small-size modern telescopes and poor weather conditions in Japan, a 2 m IR telescope atop Haleakala, Hawaii, and a 1.2 m IR telescope in South Africa are being constructed by the University of Tokyo and by Nagoya University, respectively. Japan had never established a permanent governmental scientific facility abroad before the Subaru. NAOJ being a government organization, much effort was needed to overcome this hurdle and finally in 1997 the Subaru Telescope Facility was formally established in Big Island, Hawaii. Further progress toward internationalization is definitely needed for Japanese science and its future.

The Subaru Telescope, equipped with adaptive optics and versatile observational instruments, together with other new-generation telescopes such as VLT and GEMINI will embark on new frontiers including direct observations of extra-solar-system planets, detection of many small icy planetesimals orbiting beyond Pluto, observations of mysterious galactic nuclei, dark matter, distant clusters of galaxies and the youngest celestial bodies in the expanding universe.

Developments in other fields of astronomy and astrophysics

Recent major developments of experimental research in astrophysics in Japan include neutrino and gravitational wave astronomy. Kamiokande's neutrino detection from supernova SN1987A opened a new field of neutrino astrophysics. The Kamiokande (Kamioka, Japan) is a Cherenkov light detector initially built to detect proton decay (led by Masatoshi Koshiba and Yoji Totsuka). Because of the importance of neutrino observations, the Cosmic Ray Institute (CRI, University of Tokyo) constructed SUPER-KAMIOKANDE which is 10 times larger than Kamiokande with 50 kt of pure water in a tank 39 m in diameter and 41 m high, at the Kamioka site, 1 km underground. It confirmed the finite mass of the neutrino recently.

TAMA-300 is a Japanese pilot project in gravitational astronomy led by Yoshihide Kozai and joint groups of NAOJ and other institutes. The 300 m long L-shaped interferometer of TAMA-300 at Mitaka started operation in 1999. Its high sensitivity would make TAMA-300 able to detect gravitational waves for the first time, if they have fortunate events such as that for Kamiokande. As 3–4 km scale gravitational telescopes, LIGO of USA and VIRGO of France–Italy are to be commissioned within a few years; the Japanese gravitational telescope group is working on a further upgrade of TAMA-300.

AGASA, a large-aperture high-energy cosmic ray detector of CRI, is in operation at Akeno, Yamanashi, Japan. It aims to detect cosmic rays with extremely high energy ($>10^{20}$ eV) and identify their origin, which is still unknown.

Theoretical studies in Japan have kept their high level through the 1950s to the 1990s. The extensive studies of atmospheres of low-temperature stars by Takashi Tsuji,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the discovery of a mathematical solution for a rotating black hole by Akira Tominaga and Fumitaka Sato, the inflationary universe theory of Katsuhiko Sato and the theory of supernova explosions of Daiichiro Sugimoto and Kenichi Nomoto are examples of prominent contributions among many active theoretical studies in Japan. The availability of powerful computers is also providing a strong basis for theoretical work in Japan.

Future plans

Encouraged by this remarkable progress, Japanese astronomers are discussing their ambitious future facilities. An immediate project of Japanese ground-based astronomy is LMSA, which was originally planned as a Japanese large mm and sub-mm array consisting of 50 antennas each 10 m in diameter aimed at a wide range of study from planetary formation to cosmology. In 1994 LMSA was recognized as the next national project strongly supported by the astronomical community of Japan and the search for a site started in parallel in Chilean dry plateaux. The 4800 m elevation Pampa La Bola near the Atacama desert was selected as the best site. At present radio astronomers of Japan, USA and Europe are intensively discussing the consolidation of their similar projects: LMSA, MMA (USA) and LMA (Europe). The unified world array, ALMA, will be a challenging mm and sub-mm wave telescope with 100-element antennas and 10 km maximum baseline. Such an interferometer would enable humankind to obtain high-quality images of a variety of cosmic phenomena with spatial resolution of 0.01 arcsec. Compared with the current highest angular resolution of 0.1 arcsec of the Hubble Space Telescope and new-generation optical–IR telescopes, the images will be ten times sharper. The ALMA project will be an equally shared triple cooperation by Japan, USA and European countries with a total cost of the order of 1 billion dollars. Astronomers wish to start the operation before 2010.

VERA (VLBI Exploration of Radio Astrometry) is a national plan for high-accuracy radio measurements of distances to stars throughout the Galaxy to obtain a '3D map' of the Galaxy for the first time. Mizusawa and the NRO VLBI group of NAOJ, collaborating with the University of Kagoshima and others, started the construction of VERA toward its operation within a few years, with possibilities of cooperation with Asian countries. A sensitive gravitational wave telescope with 3 km baseline is being planned by groups at the University of Tokyo, NAOJ and other institutes. If coordinated with LIGO and VIRGO, it will provide information on the direction of gravitational wave sources. The Telescope Array project, a plan to extend the aperture for detecting extremely high-energy cosmic rays to 60 times that of AGASA, is being promoted by CRI under collaboration with the University of Utah, Australian groups and other universities.

Space-borne astronomy is another major frontier for the future of Japanese astronomy. In addition to the above-mentioned plans Astro-E and Solar-B, ISAS will launch the IR Imaging Surveyor (IRIS) in 2003 as the first Japanese infrared astronomy satellite. It is a survey-type 70 cm aperture telescope cooled by a hybrid system of liquid helium and a refrigerator with high-sensitivity imaging arrays covering near-IR to far-IR wavelengths. The enormous amount of high-quality IRIS images would also be effectively used by combining them with the IR observations from Subaru and equivalent ground-based IR telescopes.

Even larger astronomical satellites using HII rockets developed by the NATIONAL SPACE DEVELOPMENT AGENCY (NASDA) are being discussed. SELENE, a lunar orbiting explorer for lunar science, will be launched in 2004 as the first attempt. A second version of SELENE including pilot astronomical observations from the lunar surface is under consideration. As an extension of the HII program, astronomers have been discussing extensively various plans for a Japanese space telescope. The H2L2 project is one such plan to launch a 3–4 m aperture cooled telescope to a Lagrange point for IR observations. The future of Japanese astronomy requires close and flexible cooperation among institutions such as NAOJ, ISAS, NASDA and CRI and universities.

Challenges Japanese astronomy is facing

Recent activities in astronomy and astrophysics, backed by a high level of industrial technology and economic power, have enabled Japan to advance and achieve considerably in this field of science. However, Japanese astronomy is still facing difficult situations compared with those in the USA and European countries.

The numbers of Japanese astronomers and related institutes are extremely low. They are concentrated at two institutes (NAOJ and ISAS) and four departments at universities (Universities of Tokyo, Kyoto, Nagoya and Tohoku). The number of the INTERNATIONAL ASTRONOMICAL UNION (IAU) members per million population of Japan is only 2 and Japan is ranked roughly the 24th among the IAU member nations. Compared with about a dozen of the top-rank countries, Japan is struggling to achieve the same level of astronomy with 1/4–1/5 as many astronomers. Although astronomy is one of the most popular subjects among university students, the number of faculty positions for astronomy is very low even now and for the majority of students astronomy courses are not available. Such a situation partially comes from the inflexible system of Japanese universities.

The operation of scientific organizations is also a subject for improvement. Employment, budget and accounting are tightly controlled by governmental regulations, which are often restrictive for scientific activities. It also causes difficulties in international cooperation or activities in different cultures and traditions of various countries. In 1998 the Japanese Parliament formally proposed an extensive reform of governmental organizations including national institutes, such as NAOJ and ISAS, and universities. The proposed plan includes possible improvements with respect to the difficulties

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Japanese basic science is facing. The future of Japanese astronomy toward the 21st century will depend greatly on the outcome of this innovation.

Norio Kaifu

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jeans, Sir James Hopwood (1877–1946)

Astrophysicist, born in Ormskirk, Lancashire, England, worked at Cambridge, Princeton and Mount Wilson Observatory, and retired early to devote himself to research. Like CHANDRASEKHAR, Jeans worked on physical problems such as thermodynamics, applying the physics to astronomy, and writing lucid accounts of the whole field in books such as *The Dynamical Theory of Gases* (1904), *Theoretical Mechanics* (1906) and *Introduction to the Kinetic Theory of Gases* (1940). He was a brilliant broadcaster and writer of popular astronomy, including the best-selling *The Universe Around Us* (1929), published at the time when the scale of the expanding universe was becoming clear. He proposed a tidal theory for the creation of the solar system, based on the hypothesis of a close approach of a star to the Sun, which drew out a cigar of material, from which condensed the planets. The *Jeans length* is the characteristic size of a cloud of gas that will collapse under its own gravitational attraction, into, say, a cluster of stars or an individual star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jeffreys, Sir Harold (1891–1989)

Geophysicist and mathematician, born in Fatfield (near Durham), England, became Plumian professor at Cambridge, worked in many areas of science, including geophysics, where he studied earthquakes, and was the first to conclude from the waves that earthquakes initiated that the Earth has a liquid core. In astronomy he studied the outer planets, proposing models for their structures. He also studied the origin of the solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jet Propulsion Laboratory (JPL)

Located in Pasadena, California, JPL is a research, development and flight center operated for NASA by the California Institute of Technology. Its primary role is the exploration of the solar system through the use of robotic scientific spacecraft. JPL is also responsible to NASA for supporting research and advanced development related to flight projects.

JPL was founded in the 1930s when Caltech professor Theodore von Karman conducted pioneering research into rocket propulsion in the area adjacent to the current site. From 1944 onwards, the laboratory conducted research for the US Army. In November 1957 it was selected to design and build the first American satellite. Data from the 14 kg Explorer 1 led to the discovery of the Van Allen radiation belts.

On 3 December 1958, JPL was transferred to NASA. Since then, it has participated in almost all of the agency's planetary missions, including the Ranger and Surveyor lunar projects, the Mariner series, the Viking missions to Mars, the Voyager missions to the outer solar system, the Magellan Venus radar mapper, Galileo and Cassini. Current projects include the series of Mars Surveyor missions and proposals to send spacecraft to Europa and Pluto. JPL was also the US manager of the IRAS project to map the infrared sky.

In addition, JPL conducts studies of the Earth and its environment through satellites such as the Solar Mesosphere Explorer and Topex-Poseidon. A number of JPL experiments have been flown in the Shuttle payload bay, including several Shuttle imaging radar missions.

The laboratory also designs and operates NASA's Deep Space Network of antennas, which is used to communicate with lunar and interplanetary spacecraft.

For further information see
http://www.jpl.nasa.gov/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jodrell Bank Observatory

The Jodrell Bank Observatory is part of the University of Manchester and was founded by Bernard Lovell in December 1945. Its prime instrument, the 76 m, MK1 radio-telescope, was completed in 1957. It was given a major upgrade in 1971 and is now known as the Lovell Telescope. In its early years it pioneered the technique of long baseline interferometry which led to the discovery of quasars. A major use now is in the study of pulsars. In 1964 a second 26 × 32 m telescope, the MK2, was completed. Given a new surface in 1987, it is now used as one element of the MERLIN VLBI NATIONAL FACILITY. Observations by the Lovell and MK2 telescopes led to the discovery of the first gravitational lens, the double quasar, and their study continues to form a major part of the research efforts of the observatory. It also operates a number of telescopes, including the Very Small Array, on Mount Teide, Tenerife, to study fluctuations in the cosmic microwave background, and is building receivers for the Planck Surveyor spacecraft to extend this work in the future.

For further information see http://www.jb.man.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

John of Holywood [known as Sacrobosco] (*c*. 1200)

Thirteenth century English translator who taught in Paris and provided students with a text-book on timekeeping, calculational methods and astronomy, the latter called *The Sphere*. A watered-down version of John of Seville's translation of *The Elements* by AL-FARGHANI, it was the long-term, best-selling, standard textbook in astronomy until the seventeenth century (it was published in 200 editions in the age of printing), even if it was barely adequate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Johnson Space Center

The Johnson Space Center (JSC) in Houston is NASA's lead center for the space shuttle and the International Space Station programs and for biomedical research. Areas of study include Earth sciences and solar system exploration, astromaterials and space medicine. About 14 000 people, including 3000 civil servants, work at JSC.

For further information see http://www.jsc.nasa.gov.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Johnson, Harold Lester (1921–80)

Astronomical photometrist, born in Denver, CO, he had a peripatetic career at American observatories such as Lick, Lowell, Washburn, Yerkes, McDonald, Tucson, Tonanzintla and the National Observatory, Mexico. At Lick, Gerard Kron inspired Johnson to use his knowledge of electronics acquired in wartime service to establish the fundamental limits of photoelectric photometry. Using the newly developed photomultiplier, he established that there was an error of a magnitude (factor of two) in the standard magnitudes of the stars in Selected Areas, then used as the basis to determine the distances of galaxies. He worked with W W MORGAN, to establish new photoelectrically measured standards, now known as the *Johnson–Morgan system*, or the U, B, V system, the *de facto* international standard of stellar (and later, galactic) photometry, used to construct color–magnitude diagrams of clusters and galaxies. Johnson demonstrated how star clusters of different ages have characteristically different color–magnitude diagrams and, in particular, how the turn-off point, where stars begin to depart from the zero-age main sequence can be used to estimate fairly precisely the ages of the clusters. Johnson's design of stellar photometer became the benchmark and it was replicated all around the world. In Texas, Johnson worked with Frank Low to extend the U, B, V system to the longer wavelengths of the near infrared, the R, I, J, K, L and N bands, providing a knowledge of the relative spectral distribution of star light from 0.3 to 10 microns, which he was able to calibrate to absolute flux. He was thus able to produce absolute energy curves for all sorts of important categories of stars: sub-dwarf stars, cepheids, M stars, carbon stars, infrared objects, even circumstellar shells and quasars. This important work provided the observational challenge for the theoretical modeling of a variety of astrophysical phenomena.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Joint Institute for Laboratory Astrophysics (JILA)

An institute run jointly by the University of Colorado and the US National Institutes for Standards and Technology (NIST)—formerly the National Bureau of Standards—since 1962. Located on the main campus of the University of Colorado at Boulder.

JILA was originally established to study gaseous interactions and precision measurements. It has subsequently expanded its mission to include the development of new measurement methods and standards, the improvement of industrial competitiveness, and the education of graduate students in technology.

JILA is a center for advanced theoretical and experimental research on subjects such as atomic interactions, spectroscopy, gravitational physics, radiative transfer and stellar interiors.

Examples of cutting-edge technologies developed in the JILA laboratories include the world's most precise laser, the coldest place in the universe, tools for manipulation of some of the smallest human-made objects in the world, ultrasensitive methods for vibration isolation, and state-of-the-art computer modeling of the Sun's turbulence.

Major areas of research include atomic and molecular interactions, new states of matter (Bose–Einstein condensates), cooling and trapping of matter, optical and laser physics, nonlinear optics and optical techniques, chemical physics, materials processing and semiconductor fabrication, nanometrology, precision measurement, gravitational physics, stellar atmospheres, interstellar astrophysics, solar physics and extragalactic astronomy.

Astrophysical research within JILA covers much of modern astrophysics—the interstellar medium, the Sun, stellar atmospheres and coronas, supernovae, x-ray stars, active galaxies, extragalactic physics and cosmology.

JILA astronomers are actively involved in a wide range of observational programs using leading ground-based facilities and major orbital observatories such as the Infrared Astronomical Satellite, the Hubble Space Telescope, the Compton Gamma-Ray Observatory, the Rossi X-Ray Timing Explorer and Chandra.

Traditional areas of interest include stellar atmospheres, and clusters of luminous stars, both in the galactic plane and in nearby starburst galaxies. Other major interests are the development of theoretical models to interpret the evolution of Supernova 1987A, and studies of cosmic x-ray sources. Extragalactic research activity focuses on the physics of interstellar matter, 'starburst' galaxies, infrared galaxies, the distribution of intergalactic gas, active galactic nuclei and quasars.

For further information see
http://jilawww.colorado.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Joint Institute for VLBI in Europe

The Joint Institute for VLBI in Europe (JIVE) operates a purpose-built Data Processor for Very Long Baseline Interferometry (VLBI), comprehensively supporting the data analysis and processing requirements of the European VLBI Network (EVN). The EVN is an integrated array of radio telescopes spread throughout Europe and beyond, which conducts unique, high-resolution, radio astronomical observations of cosmic radio sources. JIVE was established in 1993 and is located in Dwingeloo, the Netherlands. JIVE is supported via national funding agencies. Its host institute is the Netherlands Foundation for Radio Astronomy.

JIVE has played a leading role in the construction and development of a new generation of VLBI correlator, the MkIV. The MkIV, produced its 'first fringes' in July 1997, was formally inaugurated in October 1998 and produced its first astronomical image in April 1999. It began production correlation in October 1999. Research interests of the staff include: faint radio sources, supernova remnants, pulsar astrometry, maser emission around stars, gravitational lenses, active galactic nuclei and VLBI techniques (incuding space VLBI and polarization VLBI).

For further information see
http://www.nfra.nl/jive.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Joy, Alfred Harrison (1882–1973)

Astrophysicist, staff member of Mount Wilson Observatory from 1915 for nearly 60 years, Joy worked on spectroscopy and the determination of the radial velocities of stars. He used the results to find the absolute magnitudes of Cepheid variable stars (as a contribution to the determination of the intergalactic distance scale), the distance of the galactic center and the Sun's rate of revolution about it. During the observational study of dark clouds he identified a population of irregular variable stars with chromospheric emission lines; these were named T Tauri stars after the bright prototype.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Julian Date (J.D.)

Also known as the Julian period, the number of days, and the fraction of a day, which have elapsed since noon on 1 January, 4713 BC. Thus 18.00 hours UT on 1 March 1980 corresponds to J.D. 2 4444 3000.25. For certain purposes, for example, when analyzing the frequency with which events occur over long periods of time, it is most convenient to use this system of dating. The Julian period was proposed by Joseph Justus Scaliger in 1583 and named by him after his father, Julius Caesar Scaliger.

See also: calendar, Universal Time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Juno

The third asteroid to be discovered, by Karl Harding in 1804, and so designated (3) Juno. Its diameter is 268 km, and it orbits the Sun in the main asteroid belt at a mean distance of 2.67 AU (399 million km) in a period of 4.36 years; the orbital inclination is 13°, and the eccentricity 0.26. Juno rotates in 7.21 h. It is an S-type asteroid, with a reflectance spectrum indicating compositional similarities to stony-iron meteorites and ordinary chondrites; the albedo is quite high at 0.24.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jupiter

JAt 5.2 AU from the Sun (table 1), Jupiter is the closest and the most massive of the GIANT PLANETS, which extend in the outer solar system up to heliocentric distances of 30 AU. Its mass amounts to 318 terrestrial masses, and its diameter is 11 times the terrestrial one (table 2). Like the other giant planets, Jupiter has a low density (1.31 g cm^{-3}) which reflects its chemical composition, mostly dominated by hydrogen and helium. Jupiter has a fast rotation around its polar axis (9 h 55 min) which induces a characteristic structure of belts and zones, parallel to the equator.

Table 1. Orbital characteristics of Jupiter.

Semi-major axis of the orbit	5.202 AU
Eccentricity	0.048
Inclination over the ecliptic	1°18′ 28″
Sidereal period	11 yr 314.84 days
Mean orbital velocity	13.06 km s^{-1}

Table 2. Physical characteristics of Jupiter.

Equatorial diameter	142 796 km
Equatorial diameter relative to the Earth	11.19
Polar diameter	133 540 km
Flattening	0.062
Mass relative to the Earth	317.9
Mean density	1.31 g cm^{-3}
Surface gravity (at P = 1 bar)	24.8 m s^{-2}
Surface gravity relative to Earth	2.64
Escape velocity	59.64 km s^{-1}
Sidereal rotation	9 h 50 min to 9 h 56 min
Inclination of equator versus orbital plane	3° 04′
Albedo	0.45
Main atmospheric composition	H_2 (89%) He (11%) CH_4 (0.2%)

Jupiter has a ring system and many satellites (see JUPITR: RINGS, JUPITER: SATELLITES). Its ring is very tenuous; the planet is surrounded by 39 satellites (discovered to date), including the four main ones known as the Galilean satellites (see figure 1). Another characteristic of Jupiter is its magnetosphere, more massive than those of the other giant planets, and comparable in some regards to the terrestrial magnetosphere (see JUPITER: MAGNETOSPHERE, MAGNETOSPHERE OF EARTH).

Our knowledge of Jupiter started with the beginning of telescopic observations, more than three centuries ago. Over the last three decades, considerable progress has come from space exploration, but also from the on-going technological developments of ground-based astronomy.

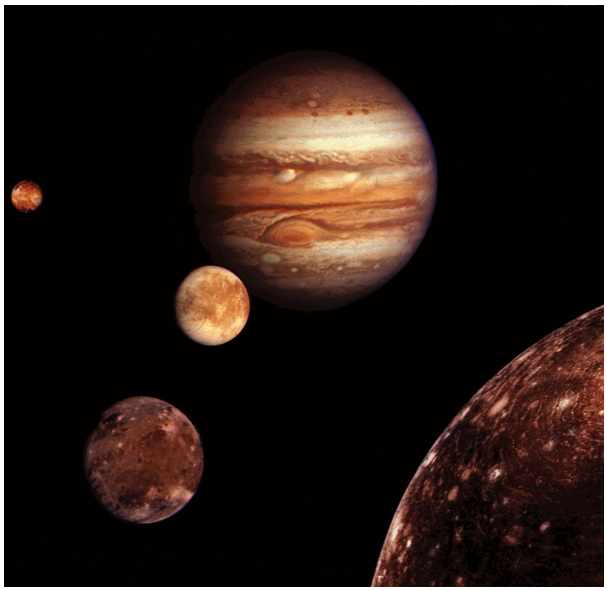


Figure 1. Jupiter and the four Galilean satellites. This picture shows a combination of images from the Voyager 1 spacecraft as it encountered the Jovian system in March 1979. The Great Red Spot appears in the southern hemisphere of the Jovian disk. Io is the red satellite on the left, Europa is the white satellite at the center, Ganymede is left and Callisto is on the front. Jupiter and the four Galilean satellites were most probably formed together; this formation scenario is needed to account for the near-circular orbits of the Galilean satellites, all very close to the equatorial plane of Jupiter.

The early observations

As one of the brightest objects in the sky (visual magnitude of –2.6 and angular diameter up to 44" at opposition), Jupiter has been known from antiquity. Its astronomical observation started almost four centuries ago with the appearance of the first refractors and telescopes. Galileo is believed to be the first astronomer to have discovered, in 1610, the four large satellites orbiting the planet, known since that time as the Galilean satellites. A few decades later, the band and zone structure was identified; in 1664, Robert Hooke discovered the GREAT RED SPOT (GRS) which was then extensively studied by Cassini. A continuous monitoring of the planet has taken place since that time, showing in particular the stability of the GRS for over three centuries.

The development of spectroscopy, early in the twentieth century, allowed the first detection of some atmospheric compounds (methane and ammonia were discovered by Wildt in 1932), but hydrogen, the most abundant atmospheric constituent, although strongly suspected on theoretical bases, was not identified until 1960 (detection by Kiess and colleagues); helium, the second gas in terms of abundance, remained undetected until the space era. Since the 1970s, however, the development of ground-based infrared spectroscopy has led to the detection of more than a dozen atmospheric constituents, including

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

phosphine, germane, water vapour, ethane, acetylene and various isotopic species (see JUPITER: ATMOSPHERE). In July 1994, with the collision of comet Shoemaker–Levy–9 with Jupiter, astronomers were given an exceptional opportunity to study in real time the response of a planetary atmosphere to a major meteoritic impact. The set of collisions induced strong stratospheric heating and the formation of new species, some of them remaining visible for the following two years (see SHOEMAKER–LEVY 9–JUPITER COLLISION).

Another major result of the ground-based exploration of Jupiter was the early discovery, in the 1950s, of a strong radio emission, in the decimetric and decametric range. This non-thermal emission was interpreted as the signature of a strong magnetic field. The decimetric emission was interpreted as synchrotron radiation of electrons orbiting along the radiation belts of Jupiter; this radiation is modulated by the motion of Jupiter's inner Galilean satellite IO. The decametric radiation, characterized by sporadic emissions, appears to be due to magnetic outbursts and storms (see figure 2).

The space exploration

The first step in the space exploration of the Jovian system was its encounter by the Pioneer 10 and 11 spacecraft, in 1973 and 1974 respectively. The PIONEER MISSIONS provided the first high-resolution images of the atmospheric structure, and an *in situ* investigation of its magnetosphere. But the space era really started in 1979 when the two VOYAGER spacecraft encountered the planet. Not only did the cameras reveal an unexpected variety and complexity in the cloud structures and their dynamics, but they also discovered the ring system of Jupiter, active volcanoes on Io (providing the first evidence for active volcanism outside the Earth), and a great diversity in the surfaces of the other Galilean satellites. The Voyager spacecraft also discovered other smaller satellites, bringing the list of known satellites to 16.

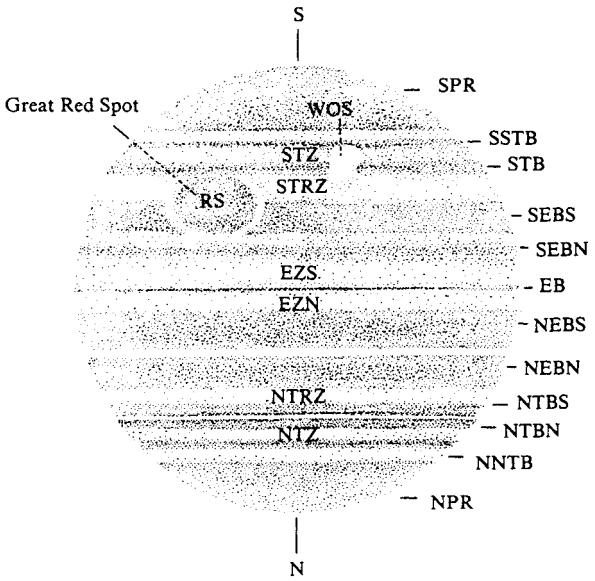


Figure 2. Belts and zones in Jupiter. Their nomenclature is given in table 3. Temperature measurements recorded in the infrared range have shown that the belts exhibit higher temperatures than the zones. It is generally believed that zones are regions of ascending motion on top of which clouds form, while the belts are are cloud-free regions of descending currents. This analysis has been confirmed by the Galileo probe, which has entered a dry and cloud-free region in one of the hot spots of the NEB; however, Galileo observations have also demonstated that the Jovian circulation is more complex by far than this simple scheme, with considerable small-scale structure.

The GALILEO MISSION, launched in 1989, arrived at Jupiter on December 7, 1995. Upon its arrival, a descent probe was sent into the Jovian atmosphere and provided data down to an atmospheric level at a pressure of 22 bars. Since that time, the Galileo orbiter has extensively monitored the Jovian atmosphere and the Galilean satellites, providing a huge amount of information about the atmospheric structure, cloud evolution and dynamics, the satellite surfaces and the magnetospheric environment. In particular, the analysis of images of EUROPA's surface suggests a possible ocean of water below the water ice crust of Europa. In order to study more deeply the surface and subsurface of Europa, the Galileo mission has been extended into the Galileo Europa Mission (GEM).

The CASSINI spacecraft, en route to Saturn, flew past Jupiter in 2000. Cassini observed Jupiter from 1 October 2000 to 31 March 2001, making its closest approach on 30 December 2000, at a distance of about 9.7 million kilometers (6 million miles). During the fly-by, new observations of Jupiter, its satellites and environs were made by Cassini, Galileo, the Hubble Space Telescope, the Chandra X-ray Observatory and Earth-based radio and optical telescopes.

Composition and structure

Why is Jupiter mostly composed of light elements? Its low density can be explained in the light of the model which accounts for the formation of the two classes of planets, the telluric ones and the giants. It is now generally believed that the solar system was formed from a protoplanetary disk which encountered the gravitational collapse of an unstable interstellar cloud submitted to accelerating rotation. The central part of the disk accreted into the proto-Sun, while planets were formed from the accretion of planetesimals made out of the disk matter.

At small heliocentric distances (R_h<2 AU), the temperature was probably such that only metals and silicates could condense, giving birth to the cores of the telluric planets. However, at larger heliocentric distances, many more abundant molecules, like CO_2, H_2O, CH_4 or NH_3, were present in solid form. These ices, added to the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

metals and silicates, contributed to solid cores which could be at least ten times more massive than the Earth. At this stage, the gravitational field of these cores became sufficient to accrete by gravitational collapse the surrounding gaseous nebula, mostly formed of hydrogen (75%) and helium (25%) as is all interstellar and cosmic material. As a result, the atmosphere of the giant planets reflects, to first order, the chemical composition of the primordial nebula which gave birth to the SOLAR SYSTEM, 4.5 billion years ago.

Because they were at further distances from the Sun, in a less dense part of the protoplanetary disk, the accretion process of the core probably took more time for URANUS and NEPTUNE than for Jupiter and Saturn. The accretion phase of the surrounding nebula may then have taken place at a time when the primordial gas was already partly swept away in the T-Tauri phase of the young Sun, and less primordial gas may have been available. This scenario could explain why Uranus and Neptune have a very large fraction of their mass in their central core (probably 60 to 80%) as compared with Jupiter and Saturn (3 to 15%).

For this reason, Jupiter, the most massive and gaseous of all giant planets, is a precious witness to the early stages of solar system history. The study of its elemental and isotopic composition can provide key diagnostics on the chemical composition of the primordial solar nebula in the outer solar system at the time of the planets' formation. The abundance of deuterium and helium are of special interest for this study. In addition, a comparative analysis of elemental and isotopic abundances on all giant planets is a valuable tool for understanding the formation and evolutionary processes of these objects.

From the similarity of the chemical composition of Jupiter with cosmic abundances, it is possible to see some analogy between the giant planet and the smallest stars; the main difference, however, is that the mass of Jupiter is only 0.1% of the solar mass, while the critical mass for star formation is about 8%. Below this threshold, the internal temperature is not sufficient to give rise to thermonuclear reactions through which hydrogen is converted into helium and later into other heavier elements. Jupiter, in this view, could be considered as a low-mass example of the class of 'brown dwarfs', those hypothetical cold objects which are currently extensively searched for, as they might contribute to a significant part of the hidden mass of the Universe.

As hydrogen is the main element of the Jovian atmosphere, most of the constituents are in reduced form: CH_4, NH_3, PH_3, GeH_4, AsH_3 and H_2O. H_2S, which had remained undetected from spectroscopic remote sensing observations, has been analysed by the mass spectrometer of the Galileo probe. Another class contains the hydrocarbons which are formed from the photodissociation of methane by the solar UV radiation (C_2H_2, C_2H_6…). A third class contains the isotopic species: HD, $^{13}CH_4$, CH_3D, $^{12}C^{13}CH_2$, $^{15}NH_3$…. The ISO satellite has recently discovered the presence of gaseous H_2O and CO_2 in the stratosphere of Jupiter, as well as in those of the other giant planets and TITAN this oxygen source might originate from an interplanetary meteoritic flux or also partly from the giant planets' icy satellites and rings.

For over two decades, there have been repeated attempts, both from ground-based and space observations, to determine the relative abundances of elements which are expected to be most abundant after hydrogen and helium, namely C, O, N, P and S. After many uncertain and sometimes contradictory results, the Galileo probe has given us important new information on the chemical composition of the lower atmosphere, at pressure levels up to 20 bars. From the present state of knowledge, it is now believed that in the deep troposphere, below the levels where condensation may take place, there is an enrichment in carbon, sulfur, phosphorus and nitrogen with respect to the solar value (there is no definite result on a possible oxygen enrichment, because the region entered by the Galileo probe was especially dry, and not representative of the whole planet). This enrichment is most likely due to the ices contained in the core, which were probably mixed in the Jovian atmosphere in the heating phase generated by the collapse of the surrounding nebula around the Jovian core. It can be seen that the scenario of formation of the giant planets mentioned above, called the 'nucleation scenario', receives strong support from the measurements of elemental abundances. It is also expected that the enrichment should increase with heliocentric distance as the relative mass of the core increases for the more distant giant planets. This is indeed the case for methane, strongly enriched on Uranus and Neptune.

The thermal vertical profile of Jupiter is characterized by two main regions: in the troposphere, down to the interior, the heat is transferred by convection (as in stars) and the temperature decreases as the altitude increases, following an adiabatic gradient. The temperature reaches a minimum of 110 K at the tropopause, where the pressure level is about 100 mbar. Above the tropopause, the temperature increases again with height, due to the absorption of the solar UV and near-IR flux by methane and aerosols; this is the statosphere where the energy is transferred by radiation.

The region sounded in the visible and infrared range is mainly the upper troposphere and the lower stratosphere, at pressures ranging mostly between a few bars and about 1 mbar; the UV radiation probes the Jovian stratosphere. In addition, the upper stratosphere is also probed at specific infrared wavelengths corresponding to ion or molecular transitions, where thermal emission probes the microbar pressure level. Some of the atmos-

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

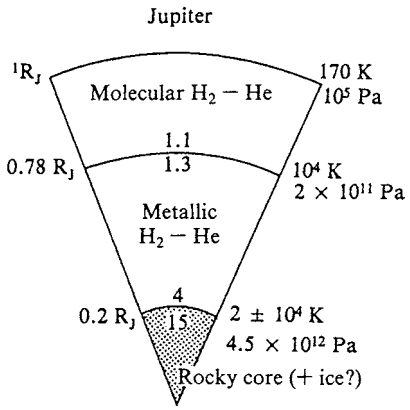


Figure 3. Model of the internal structure of Jupiter (after Stevenson 1982). Jupiter is believed to have formed from the gravitational collapse of primordial gas (mostly composed of hydrogen and helium) on a central core made of rock and ice having about 10 to 15 terrestrial masses, at an early point in the history of the solar system. The internal temperature may then have reached 10^5 K. The heat accumulated during the accretion phase has been continuously re-radiated to space since the origin in a slow contracting phase. The central temperature today is believed to be 23 000 K. Most of Jupiter's interior is probably made of convective liquid metallic hydrogen. There may be no definite transition between the metallic hydrogen and the molecular hydrogen which overlays it.

pheric species condense in the upper troposphere: from thermochemical calculations, NH_3 is expected to condense at around 0.5 bar, and NH_4SH and H_2O at lower levels; the NH_3 cloud is believed to be responsible for the white color of the planet observed in the visible range, especially in the zones.

The internal structure of Jupiter is only known from indirect measurements and from theoretical modeling (see figure 3). The known parameters are the radius, mass and density, and the gravitational moments have been measured by the Voyager spacecrafts. Another important parameter is the internal energy of the planet, which was first discovered by Low and his colleagues from ground-based infrared measurements, and was later more accurately measured by the Voyager spacecraft. Based on these observational constraints and on theoretical models, it is possible to estimate the pressure (50 to 100 Mbars) and the temperature (23 000 K) at the center of the planet. It is currently believed that the interior consists of a solid nucleus of about 10 to 15 terrestrial masses, probably made of rocks and metals at the center, surrounded by a convective ocean of metallic hydrogen, where the magnetic field of Jupiter is believed to be generated. It is likely that helium differentiation takes place in the liquid hydrogen, with helium droplets falling toward the center, as in the case of Saturn. The released energy probably contributes, with the planet's contraction, to the observed internal energy of Jupiter. Above the metallic hydrogen ocean lies the molecular hydrogen convective envelope. A powerful method for learning about the internal structure of Jupiter will be the measurement of the oscillation modes of the planet, as these oscillations are expected to probe its deep interior.

Dynamics

The internal energy source of Jupiter has important consequences for Jovian dynamics. The main features are the belt and zone structures induced by the fast rotation of a convective sphere (table 3). Spacecraft observations have shown that the zones are colder and cloudier than the belts; they are believed to be regions of ascending motion, while the belts, which are drier, are regions of subsidence. The Galileo probe, in December 1995, entered one of these dry regions, so that the atmospheric structure and composition retrieved from the probe data are probably not representative of the whole planet. Moreover, high-resolution images and spectroscopic observations from the Galileo orbiter have demonstrated that the Jovian circulation is much more complex than this simple zone/belt scheme, and that vertical motions also exist on a small scale.

Table 3. Nomenclature of the Jovian belts and zones.

Designation	*Definition*	*Approximate latitude (degrees)*
NPR	North polar region	47–90
NNTB	North-north temperate belt	43
NTZ	North temperate zone	35
NTBn	North temperate belt (northern component)	30
NTBs	North temperate belt (southern component)	23
NTRZ	North tropical zone	15–20
NEBn	North equatorial belt (northern component)	14
NEBs	North equatorial belt (southern component)	10
EZn	Equatorial zone (northern component)	+3
EB	Equatorial belt	0
EZs	Equatorial zone (southern component)	–3
SEBn	South equatorial belt (northern component)	–10
SEBs	South equatorial belt (southern component)	–19
GRS	Great Red Spot	–22
STRZ	South tropical zone	–25
STB	South temperate belt	–29
STZ	South temperate zone	–37
WOS	White oval south	–35, –37
SSTB	South-south temperate belt	-41
SPR	South polar region	45–90

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Coriolis forces induce zonal winds of opposite direction north and south of each zone, generating multiple time-variable eddies. The biggest structure is the GRS, in the southern hemisphere, which is currently interpreted as a giant anticyclonic structure, i.e. a column of ascending motion; however, its stability for over three centuries is not understood at present. There are many other smaller eddies which were first observed with the Voyager cameras; these features have been monitored with the HST and with the camera of the Galileo orbiter. Most of them evolve rapidly with a time-scale of a few months.

If the white color of the zones can be reasonably interpreted as NH_3 ice, the composition of the clouds and/or gases present in the eddies is still a puzzle, especially in the case of the GRS. Its red color suggests the presence of phosphorus or sulfur compounds, but their chemical composition has not been identified yet.

The Jovian magnetosphere

Jupiter has at least three factors which favor the presence of a large magnetosphere: the presence of a liquid metallic interior, the fastest rotation and the largest size of all planets. Consequently, Jupiter has the largest magnetosphere in the solar system.

The Jovian magnetic field is believed to be generated by the dynamo effect within the liquid metallic hydrogen ocean which extends in Jupiter's interior below about 0.75 Jovian radii down to the central core. Following the first identification from ground-based radio measurements, the planetary magnetic field was explored by Pioneer 11 in 1974, then the two Voyager spacecraft in 1979. The dipole moment, tilted by 10° with regard to the rotation axis, is almost 20 000 times the terrestrial value; the intensity of the Jovian magnetic field at the equator is about 10 times the terrestrial value at the Earth's surface; it is even about three times higher at the Jovian poles than at the equator.

Due to the strong magnetic field, the fast rotation of the planet and the large heliocentric distance of the planet, the Jovian magnetosphere is extremely extended. Its structure, however, shows considerable analogy with the Earth's magnetosphere. It is characterized, in the sunward direction, by the presence of a shockwave, a magnetopause and a magnetosheath. A major difference from the Earth configuration is the presence of Io, the orbit of which crosses the Jovian radiation belts (see MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS). In the antisolar direction, the magnetic tail extends almost up to Saturn's orbit.

Rings and satellites

It has been believed for centuries that Saturn was the only giaant planet with a ring system In 1977, however, nine rings were discovered around Uranus from an airborne stellar occultation experiment. Then, in 1979, a ring was fortuitously found around Jupiter on Voyager 1 images, very close to the planet. Improved images were obtained a few months later using the Voyager 2 camera. Finally, in 1989, Voyager 2 confirmed the existence of a ring system around Neptune, also inferred from previous ground-based occultation experiments.

The Jovian ring is located in the planet's equatorial plane. The first ring extends over about 5700 km, and its outer edge lies at about 57 000 km from the planet's limb. Like Saturn's ring, the ring is very thin, as its width does not exceed 1 km. This first ring is followed by another brighter one, which extends over 800 km, itself followed by a third ring, darker and very tenuous, made of micrometer-size particles, which extends up to three Jovian radii. The rings are likely to be made of dust and water ice.

The satellites of Jupiter fall into three categories (table 4): the four big Galilean satellites, known since their discovery by Galileo in 1610, the four inner satellites Metis, Adrastea, Amalthea and Thebe, and the high outer satellites. Of the inner satellites, only Amathea was known prior to the Voyager exploration. They are all smaller than 100 km in diameter and irregular in shape, and the two closest to Jupiter, Metis and Adrastea, are associated with its ring system. The inner satellites, as well as the Galileans, are believed to have accreted from a protoplanetary disk surrounding the protoplanet Jupiter. In contrast the outer satellites, which all have orbits with eccentricities higher than 0.1 and inclinations larger than 20°, are most likely captured bodies, originating from comets or from the family of Trojan asteroids.

Table 4. The satellites of Jupiter.

Name	*Sidereal orb. period (days)*	*Semi-major axis (plan. rad.)*	*Diameter (km)*	*Density ($g\ cm^{-3}$)*
Metis	0.294	1.80	(40)	?
Adrastea	0.297	1.80	(30)	?
Amalthea	0.489	2.52	270–170–150	?
Thebe	0.678	3.11	(80)	?
Io	1.769	5.78	3632	3.53
Europa	3.551	9.40	3138	2.97
Ganymede	7.155	14.99	5262	1.94
Callisto	16.689	26.33	4800	1.86
Leda	240.0	156.0	2–14	?
Himalia	250.6	159.8	180	?
Lysithea	260.0	163.2	6–32	?
Elara	260.1	163.4	80	?
Ananke	631(r)	290.0	6–28	?
Carme	692(r)	313.0	8–40	?
Pasiphae	735(r)	326.3	8–46	?
Sinope	758(r)	332.0	6–36	?

(r) = retrograde.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Galilean satellites

Discovered by Galileo in 1610, the four Galilean satellites have played an important role in the history of astronomy. In 1675, the Danish astronomer Ole Rømer used their mutual eclipses and occultations, at conjunction and opposition, to show that light had a finite speed and to estimate its velocity.

Together with Jupiter and its ring system, the four Galilean satellites represent a sort of replica of the solar system, on a smaller scale: they all have direct, nearly circular orbits, very close to the Jovian equatorial plane; their densities decrease as their distance to Jupiter increases because, as in the case of the solar system, the lightest elements of the closest satellites have evaporated due to heating from Jupiter. The study of the four Galilean satellites has greatly benefited from the Galileo mission, devoted to their in-depth exploration until early 2000.

While the three other Galilean satellites are covered with water ice, Io, Jupiter's innermost Galilean satellite is covered with sulfur and/or sulfur compounds which are responsible for its reddish color. One of the most important discoveries of Voyager was the first evidence for active volcanism on Io. The origin of this volcanism is internal heating caused by frictional dissipation of tidal energy generated by its orbital resonance with Europa. Its high density (3.5 g cm^{-3}) suggests an interior mostly made of silicates, as in the case of the Moon. Io has a very tenuous stable atmosphere of SO_2, with a surface pressure in the range of 4–30 nbars. In addition, material removed from Io's surface (atoms of H, O, S, Na, K) is sent to space and supplies a torus surrounding Io's orbit and containing neutral atoms and ions.

Europa, the smallest Galilean satellite, shows a surface of pure water ice, with very few topographic features. What is remarkable on this surface is the net of linear structures which covers the surface. High-resolution images of Galileo suggest that these structures could be due to the presence of liquid water below the icy crust, which would occasionally flow out as a geyser. If confirmed, Europa would be the only object apart from Earth to have an ocean of liquid water at present. Below this ocean, the interior of Europa is likely to contain silicates.

GANYMEDE, the third Galilean satellite, is the largest satellite of the solar system; it is even larger than MERCURY and PLUTO. The surface of Ganymede is dark, which implies, together with its density (1.94 g cm^{-3}), that the satellite is composed of water ice and silicates in roughly equal amounts. A major result from Galileo was the discovery of a magnetic field around Ganymede; its origin is currently unclear.

Finally, CALLISTO, Jupiter's outermost Galilean satellite, shows a heavily cratered surface, much older than the others. Being farther from Jupiter, Callisto probably benefited from a weaker internal energy source, and its surface has thus remained unchanged over its history.

Bibliography

Hill, T W February 2002 Magnetic moments at Jupiter *Nature* **415** 965–966

Atreya S K, Pollack J B and Matthews M S (ed) 1989 *Origin and Evolution of Planetary and Satellite Atmospheres* (Tucson, AZ: University of Arizona Press)

Belton M, West R and Rahe J (ed) 1989 Time-variable Phenomena in the Jovian System *NASA Special Publication* 494

Gehrels T (ed) 1976 *Jupiter* (Tucson, AZ: University of Arizona Press)

Morrison D (ed) 1988 *Satellites of Jupiter* (Tucson, AZ: University of Arizona Press)

Noll K S, Weaver H A and Feldman P D (ed) 1996 *The Collision of Comet Shoemaker–Levy 9 and Jupiter* (Cambridge: Cambridge University Press)

Stevenson D J 1982 Interiors of the Giant Plants. *Annual Review of Earth and Planetary Science* **10** 257–95

1979 Mission to Jupiter and its satellites (Voyager 1 encounter). *Science* **204** 945–1008

1996 Galileo at Jupiter: results from the probe. *Science* **272** 837–60

1996 Galileo at Jupiter: results from the orbiter. *Science* **274** 377–412

Thérèse Encrenaz

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Jupiter Atmosphere

In many aspects, JUPITER is intermediate between a planet and a star. As a star, it is mainly composed of hydrogen and helium, has an internal source of energy (although not from nuclear reactions) and has a convective interior (at least in part). As a planet, it experiences cold temperatures, receives energy from solar radiation in its atmosphere and shows a cloud structure with meteorological features different from but not dissimilar to those of the telluric planets. The scientific questions raised by investigations of Jupiter belong equally to both sides, stellar and planetary physics. On the planetary side, the comprehension of the meteorology of Jupiter remains a difficult challenge for atmospheric dynamics, for example. On the stellar side, the recent discovery of brown dwarfs, and extrasolar planets of comparable size (or even greater) (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS) shed new light on Jupiter, as the closest object to Earth in a new category of astronomical objects, which remain to be classified. Jupiter is not only the largest planet of the solar system, it is now viewed as a small object in the class of the extrasolar giant planets and brown dwarfs. It is at the bottom line of a classification starting from the gaseous planets, to the brown dwarfs and up to the stars themselves. However, contrary to the far and inaccessible stars, Jupiter is close enough to be studied by interplanetary probes. The exploration of Jupiter environment started with the PIONEER 10 and 11 spacecraft in the years 1973–4, followed by VOYAGER 1 and 2 in 1979–80. Jupiter also received a visit from the Ulysses spacecraft en route to its polar orbit around the Sun, in 1992 (see SOLAR WIND: ULYSSES). Ulysses made measurements of some plasma properties, density and temperature of the electrons, for the first time outside the equatorial plane. Then, in 1995, the Galileo probe observations gave the first *in situ* measurements in a giant planet's atmosphere, followed by remote sensing observations by the Galileo orbiter after 1995 (see GALILEO MISSION TO JUPITER). Our knowledge of Jupiter has been deeply modified by these space missions, together with the continued observations from ground-based telescopes with uninterrupted improvement in sensitivity and accuracy. Even if many unknowns remain on the giant planet, even if the reddish color of some of its clouds remains a mystery, our knowledge of Jupiter is now infinitely more detailed than before space exploration. The composition, cloud structure and temperature structure have been measured or inferred, at least in general guidelines; the circulation is accurately characterized, even if not understood globally. Also, ideas on the formation of Jupiter are now more precise, even if, as usual, the new findings have also opened new and difficult questions for investigators of the 21st century.

Meteorology

Even after the close observation of several space missions, the meteorology of Jupiter still keeps many of its mysteries. Decades of painstaking telescopic observations by amateur and professional astronomers gave a first general description of Jupiter's wind system (figure 1): large zonal winds are permanently present and roughly correlated with the banded structure of the clouds (zones and belts, being the generic names of these white and dark bands). Jets alternate in prograde and retrograde direction, producing the so-called differential rotation of the planet. The equator of Jupiter makes a full rotation in 9 h 50 min 30 s (defining system I of longitude) while the midlatitudes rotate in 9 h 55 min 40 s (system II). A more accurate system of rotation is now used, in relation to the periodicity of the radio emission of Jupiter, itself related to the magnetic field. In system III, the sideral period of rotation is 9 h 55 min 30 s, and the winds, measured with reference to this system of longitude, reach about 120 m s^{-1} at the cloud top at the equator (prograde). The differential rotation is very stable temporally: only subtle variations are seen between Voyager and Galileo in the wind profiles, mostly in the strongest jets, but the latitudes of the jets and their amplitude remain essentially constant. The depth of the wind system has long been debated, some astronomers favoring models with a thin atmospheric layer where winds are produced by the solar energy deposition close to the clouds' level, while others predicted a thick meteorology, with differential rotation extending into the interior, generated by the deep convection. The Galileo probe, sounding Jupiter at 6.57°N latitude, has observed strong winds down to the 22 bar

Figure 1. Jupiter as observed from Hubble Space Telescope (October 1995). Credit NASA.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

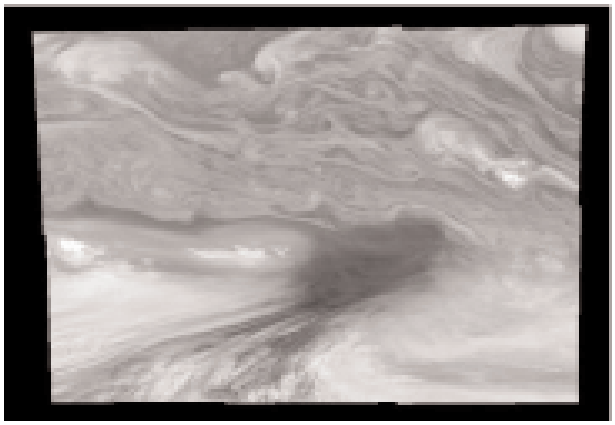


Figure 2. A hot spot of Jupiter, as observed from the Galileo Orbiter on 17 December 1996.

level, well below the upper meteorological layer: winds increase from 120 m s^{-1} at the cloud level (1 bar level) to about 170 m s^{-1} at 5 bar, and remain constant below. This observation seems to favor the models where the winds are supported by internal energy transport more than models with solar energy deposition, although the debate is still open, internal motions being possibly also driven by the deposition of energy in a thin upper boundary layer in other more recent models.

Another peculiarity of Jupiter's atmosphere is the morphology of its numerous spots. Among them, the most famous is the GREAT RED SPOT (GRS), observed during more than three centuries, since its discovery by Hooke in 1664. It is a large anticyclonic oval vortex, about 20 000 km wide in the longitudinal direction, located in the southern tropical zone. New observations from Galileo have shown that the center of the GRS is rotating slower than its boundary, with even an apparent counterrotation of the core of the GRS. The cloud structure inside the GRS, deciphered by Galileo's remote sensing instruments, shows a spiraling structure, with higher-altitude clouds at the core. Many numerical models of the GRS have been developed: the formation of large vortices is a general property often observed in turbulent rotating flows. In a recent model of turbulence, Sommeria *et al* (1988) explain the stability of large vortices by a statistical formulation of the turbulence, defining an analogy of the thermodynamics for the turbulence. This model is related to modern developments aimed at a global understanding of turbulence, one of the most difficult problems in physics. The observations of Jupiter therefore help to constrain such models, thanks to the observations of turbulent flows at scales inaccessible to laboratory experiments.

There are many other interesting features in the Jovian meteorology such as the White Oval Spots (WOSs), three anticyclonic ovals, first observed in 1938 in the South Temperate Belt. In 1998 and again in 2000, as their separation was at a minimum, two of the WOSs merged into one, leaving only one WOS, a unique event in the meteorological history of Jupiter. Galileo has measured the wind structure inside the WOSs, before and after the event, as well as the cyclonic structure between two oval spots. The most famous meteorological features in recent Jovian physics are the North Equatorial Belt Hot Spots (figure 2), because the Galileo probe unexpectedly descended directly inside one of them. These so-called 'hot spots' are regions of low cloud opacity, where thermal emission from the deeper layers is directly transmitted, which give these spots a larger flux in thermal emission at wavelengths where the atmospheric opacity is small enough for deeper soundings than in the cloud layers, such as the 5 μm spectral range. Hot spots have been observed from the ground since the beginning of infrared astronomy (Westphal, 1969). They were soon interpreted as relatively cloud-free regions. They are therefore dark in the visible, since the bright regions on Jupiter are the most cloudy. Radiative transfer calculations have shown that the atmosphere of Jupiter can be sounded as deep as 6–8 bar at 5 μm wavelength in hot spots, which is the deepest level sounded in the optical range. The dynamical formation of hot spots is still poorly understood: they seem connected to strong convective cells, the hot spots being a region of downward convection, with strong vertical descent winds. The upper atmosphere is similarly aspired downwards, and the cloud particles are vaporized in the process, owing to the higher pressure–temperature encountered in the descent, leaving a very clear atmosphere. The downward flux of the atmosphere also explains the strong desiccation in condensed molecules, such as H_2O, NH_3 and H_2S, which has been measured by the Galileo probe. The upward-convective regions of the atmosphere should, on the contrary, present high cloud opacity and a saturated abundance in condensable molecules, as the opposite processes would work. It has been suggested that the equatorial plumes, seen as large bright clouds in the equatorial region, could be identified with these upward cells: there is apparently a one-to-one association of plumes with the hot spots in pairs, as a visual inspection of the morphology of the North Equatorial Belt suggests. An interesting fact is that the plumes observed along the equator are regularly spaced, numbering 9–13, like the hot spots. The generation of the regular structure of plumes–hot spots could be due to a travelling wave in the atmosphere, called a Rossby wave (Ortiz 1999).

Composition

The composition of Jupiter in heavy elements gives important clues for the understanding of its formation. The Galileo probe measurements have shown an enrichment by a factor of 2–3, compared with the solar composition, in the following elements: C, S, Kr, Xe and

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Ar. The same is true for N, although the condensation of NH_3 introduces more difficulties in the measurements, and therefore larger uncertainties, with a deep abundance found at about 3.6 times the solar composition in nitrogen. This is also supposed to be the case for oxygen, present in H_2O on Jupiter, although the meteorological conditions in the hot spots where the probe descended were very desiccated and prevented any firm conclusion about the deep abundance of oxygen. Such enhancements are attributed to the scenario of formation of Jupiter from the solar nebula. First, the accretion of PLANETESIMALS (these small condensed objects, probably similar to comets or asteroids, are the elementary bricks of the solar system) forms a planetary core. When a core mass of about 10 Earth masses is reached, the proto-Jupiter begins to accrete the gas in the solar nebula, composed mainly of H_2 and He. Since the core does not contain the most volatile gases, degassed from the planetesimals, this mechanism can explain an enrichment in heavy elements, compared with hydrogen and helium.

The case of helium is particular, because of interior processes, at depths where hydrogen is condensed in metallic form: the lack of miscibility of He in metallic H_2 induces a phase separation of He droplets falling into the interior. As a result, He is depleted in the outer layers of the atmosphere, as is observed by Galileo, and the release of gravitational energy, contributes to the internal source. The depletion of neon is explained similarly by interior processes, by dissolution of neon in helium droplets. Depending on the temperature of the formation of planetesimals, several other elements could also be depleted, such as nitrogen (most of the nitrogen should be present in N_2, a very volatile gas). In Jupiter, a similar enhancement in nitrogen to that for the other elements is found: this implies a very low temperature for the formation of the planetesimals, of the order of 30 K only, much lower than the temperature at Jupiter's orbit, which is of the order of 160 K, and lower also than the temperature in the region of formation of the Oort cloud comets, in the Uranus–Neptune region, of the order of 55 K. Two explanations are competing: either the planetesimals forming the core of Jupiter came from the outer limits of the solar nebula, with much lower temperatures than in the Jovian vicinity, or Jupiter itself was formed at a larger distance from the Sun than its present orbit, and migrated to its present position as a result of viscous damping in the early solar nebula. The latter hypothesis is now considered seriously, after the discovery of extrasolar giant planets very close to their central star, a system which is considered to be impossible to form in a solar system formation scheme without migration of the protoplanet.

Temperature and atmospheric structure

Before the direct measurements by the Galileo probe, the temperature structure of Jupiter was measured with accuracy down to the 1 bar level only, from remote sensing ground- or space-based observations (stellar or radio occultation, infrared spectroscopy). The general features of the temperature structure were therefore already well known before Galileo, but only for the upper atmosphere: at pressures greater than 0.6 bar, the temperature increases towards the interior. The temperature lapse rate is limited by the convective instability, which transports the internal heat to the external layers. In the upper atmosphere, on the contrary, solar photons are absorbed by the atmospheric gases (H_2, CH_4) and hazes. This upper atmosphere heating generates an increase of temperature in the upper atmosphere. The heating from above (by solar radiation) and from below (by convection) induces the formation of a temperature minimum, at a level called the tropopause, which is situated on Jupiter at about 0.6 bar with a temperature of 110 K. The horizontal variations of temperature at the tropopause are usually low, of a few degrees kelvins for a given pressure level between equator and pole, except in some local areas such as the the GRS where larger temperature gradients have been found. The atmospheric structure of Jupiter has been measured accurately by the Galileo probe, at the location of the probe entry site, in a hot spot environment. Figure 3 shows the measured temperature profile. As anticipated by previous models (but never measured before), the tropospheric profile measured by Galileo is found to follow closely an adiabatic gradient, which corresponds to the maximum temperature gradient supported by an atmosphere in a convective regime.

The temperature profile in the upper atmosphere shows very distinctive and interesting features, with first strong heating above an isothermal region in the stratosphere, and oscillations in the upper part of the profile, which are attributed to gravity waves (figure 3). Gravity waves are the analogs in a gaseous atmosphere of the waves seen at the surface of a lake. The importance of

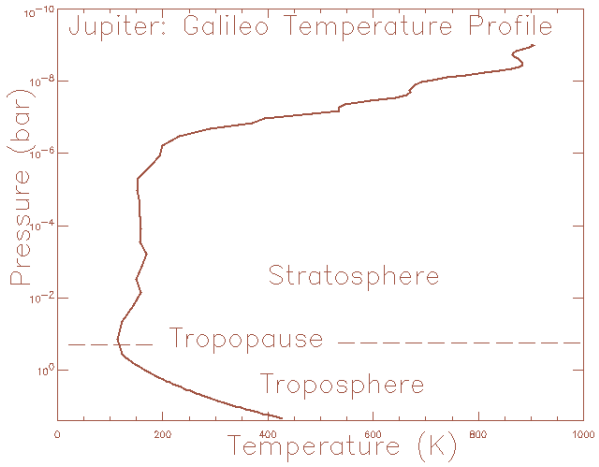


Figure 3. Temperature profile of Jupiter (after Seiff *et al* 1998 *J. Geophys. Res.* **103** 22857).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

these waves in PLANETARY ATMOSPHERES is due to the transport of heat and momentum they can dissipate in the upper stratosphere. They could provide the missing link between the transport of internal heat in the upper troposphere to the external, radiative layers. The propagation of gravity waves directly depends on the atmospheric thermal profile: they propagate only in the upper atmosphere, when the temperature gradient is lower than the adiabatic gradient, which corresponds to the region above about 0.6 bar. The strong heating seen in the thermal profile could be partly due to the dissipation of the large-amplitude gravity waves, seen in figure 3 as the oscillations in the upper part of the curve.

Clouds

Even though the origin of the colors of the clouds of Jupiter remains unknown, the Galileo measurements have given new insight into the atmospheric structure and the vertical disposition of the clouds (see also CLOUDS IN PLANETARY ATMOSPHERES). The cloud structure revealed by the Galileo probe was very specific to the particular location of the probe entry site, within a hot spot area, and it is still difficult to reconcile the models for more typical regions on Jupiter, with the measurements in a very localized and particular region. The direct measurements by the probe revealed three cloud layers, vertically separated: a first cloud around 0.5 bar, corresponding to the expected position of the ammonia cloud, a second cloud layer with a base at 1.34 bar, which could be the expected ammonium hydrosulfide cloud, and a detached cloud deeper at 1.6 bar which could be a thin water cloud consistent with the low water abundance found by the probe. To the surprise of the investigators, no deeper cloud layer was found, down to 22 bar, where the transmission from the probe stopped. There was no trace of the expected water cloud at 5 bar or below. For a solar composition (and *a fortiori* for a larger than solar abundance of oxygen), a ubiquitous H_2O cloud is expected to form on Jupiter. The observations by the orbiter have now been interpreted, and they extend the probe measurements to other places on Jupiter, with the difficulties inherent to remote sensing observations where vertical structures are not easily disentangled. Some observations by the orbiter have now convincingly proved the presence of water clouds in several places. A deep and ubiquitous water cloud is out of range of the optical observations, but in several places the water clouds are apparently transported upward by strong convection, up to observable levels. Such clouds, similar to the tropical convective clouds on Earth, are also associated with lightning and with water vapor abundance in the atmosphere larger than in the hot spots. Observations by the Galileo Orbiter continue to study the cloud structure of Jupiter, progressing towards a global understanding of cloud variations between zones and belts of Jupiter. The difference in cloud structure between belts and zones is interpreted as mostly due to a difference in opacity of the cloud layers. The composition of the cloud particles is not easily measured by remote sensing observations, and the Galileo probe had no instrument with the ability to measure the composition of cloud particles. The cloud composition is therefore mostly inferred from thermodynamics arguments and from the knowledge of the composition of the gaseous atmosphere. The bright and white clouds observed in the Jovian belts are thought to be mostly composed of ammonia particles. It is nevertheless quite plausible to have a mixture of cloud particles in a given cloud layer, owing to the vigorous convection at work in many parts of the Jovian atmosphere. Water, ammonia and ammonium hydrosulfide could all be part of the upper cloud particles. The coloration of the belts can be caused by a small amount of chromophores, sulfur or phosphorus compounds being possible candidates for their composition. The clouds of Jupiter therefore still keep many of their mysteries.

Aurorae

On Earth, a polar AURORA is observed when charged particles from the solar wind precipitate into the atmosphere, exciting the atmospheric molecules which then emit light by de-excitation processes. The charged particles are guided along the magnetic field lines. The geometry of the magnetic field therefore also governs the location of the precipitation, mainly in the polar regions. On Jupiter, similar processes have been observed regularly since the first detection of polar emissions in H and H_2 by the INTERNATIONAL ULTRAVIOLET EXPLORER satellite in 1979, followed by the Voyager observations in 1979 and 1980. The differences from the Earth's aurorae are mainly due to the difference in atmospheric composition, with hydrogen dominating Jupiter's atmosphere, instead of oxygen and nitrogen in the Earth's atmosphere. The main atmospheric effects expected from the precipitation of charged particles are threefold:

1. direct excitation of molecules, followed by de-excitation by emission of photons,
2. thermal heating of the atmosphere, followed by enhanced thermal emission, and
3. chemistry induced by charged particles in the atmosphere subject to the auroral precipitation.

The UV emission observed since 1979 is due to electronic transitions of H_2 and H, and is relevant to the first mechanism. The primary particles coming from the magnetosphere (high-energy electrons or ions, see JUPITER: MAGNETOSPHERE) decelerate into the atmosphere and produce a large number of secondary electrons, which again excite the atmospheric molecules. These processes take place in the upper atmosphere, mainly between the nanobar and millibar pressure levels.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

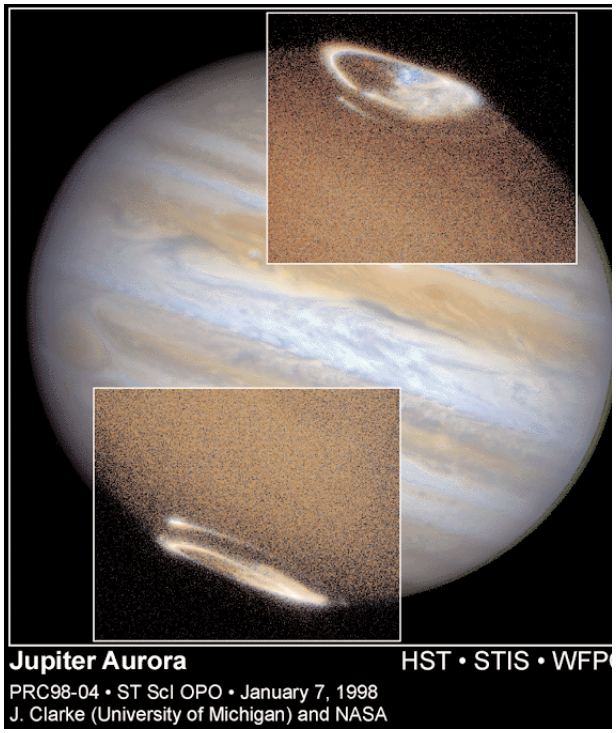


Figure 4. Auroral emissions of Jupiter, observed from Hubble Space Telescope. A visible image of Jupiter is superimposed with ultraviolet polar regions images taken with the STIS UV imager of HST.

Thermal emission of the atmosphere has also been observed in methane infrared emissions. Methane acts as the main radiator in the upper atmosphere of Jupiter. The energy deposited by charged particles is therefore mainly radiated by methane, in particular from its strong band at 7.8 μm. In the near infrared, emissions of a particular ion, H_3^+, have also been observed by ground-based spectroscopy. This molecular ion is a very peculiar molecular assembly, with three protons linked by two electrons, and, because of its high reactivity, it is an important constituent of the ionosphere of hydrogen atmospheres. Its strong infrared emissions have been detected in the mid-infrared (2–4 μm).

All these emissions have also been observed with imaging instruments, giving access to maps of the emissions and allowing us to correlate them with the geometry of the magnetic field (figure 4). Galileo observations have also observed emission in the visible wavelengths, on the night side of Jupiter, which are probably due to H_2 emissions between electronically excited bands of H_2. These observations give the most detailed maps of the aurorae to date. The geometry of the magnetic field is beautifully visualized in these images, with a north magnetic pole shifted compared with the geographic pole and an auroral oval descending toward 60° latitudes for longitudes close to 180° and tangential to the north pole on the other side.

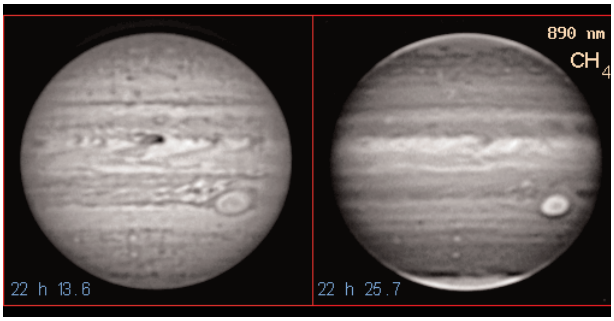


Figure 5. Jupiter observed at Pic du Midi (9 July 1996): left, image in broad red filter, showing the usual contrasts in the cloud structure; right, image in methane filter at 890 nm, enhancing the contrasts over the highest altitude clouds. The polar haze, the equatorial belt and the GRS are particularly prominent.

Auroral precipitations may also have a role in the chemistry of the atmosphere of Jupiter. The polar regions of Jupiter are covered by a high-altitude haze, which is permanently present, and limited to the regions above 60° north and south. These hazes are best visualized in images taken in filters with strong atmospheric absorption (such as methane or hydrogen band filters), which are more sensitive to the upper atmospheric contrasts (figure 5). The composition of this haze is still unknown, but a possibility is that it could be formed by heavy hydrocarbons formed from the chemistry of methane induced by charged particles. The formation of heavy hydrocarbons in a hydrogen-dominated atmosphere is a very peculiar process, unique in the solar system, and it is therefore highly interesting for planetary modeling.

New observations of Jupiter, its satellites and environs have been made by Cassini, Galileo, the Hubble Space Telescope, the Chandra X-ray Observatory and Earth-based radio and optical telescopes during the CASSINI spacecraft's flyby. The Galileo spacecraft has been in orbit around Jupiter since December 1995 while Cassini observed Jupiter from 1 October, 2000 to 31 March, 2001. Cassini made its closest approach on 30 December, 2000, at a distance of about 9.7 million kilometers (6 million miles). Data from the flyby observations show that Io leaves a magnetic footprint on Jupiter's upper atmosphere which appears as a spot of ultraviolet emission that remains fixed underneath Io as Jupiter rotates. The specific physical mechanisms responsible for generating these emissions are not well understood, but in general the spot seems to arise because of an electromagnetic interaction between Jupiter's magnetic field and the plasma surrounding Io, driving currents of around 1 million amperes down through Jupiter's ionosphere. The other Galilean satellites may also leave footprints, and the presence or absence of such footprints should illuminate the underlying physical mechanism by

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

revealing the strengths of the currents linking the satellites to Jupiter.

Jupiter's x-ray aurora has been thought to be excited by energetic sulfur and oxygen ions precipitating from the inner magnetosphere into the planet's polar regions. High spatial resolution observations demonstrate that most of Jupiter's northern auroral x-rays come from a hot spot located significantly poleward of the latitudes connected to the inner magnetosphere. The hot spot seems to be fixed in magnetic latitude and longitude and occurs in a region where anomalous infrared and ultraviolet emissions have also been observed. From the data it is inferred that the particles that excite the aurora originate in the outer magnetosphere. The hot spot x-rays pulsate with an approximately 45 min period, a period similar to that reported for high-latitude radio and energetic electron bursts observed by near-Jupiter spacecraft. These results invalidate the idea that Jovian auroral x-ray emissions are mainly excited by steady precipitation of energetic heavy ions from the inner magnetosphere. Instead, the x-rays seem to result from currently unexplained processes in the outer magnetosphere that produce highly localized and highly variable emissions over an extremely wide range of wavelengths.

From Jupiter to brown dwarfs

The recent detection of brown dwarfs has placed Jupiter in a new perspective: Jupiter is not only the largest planet of the solar system, it is also now an intermediate step between planets and stars. Low-mass stars, known as brown dwarfs, have been detected with masses as low as 10 Jupiter masses. An interesting question in the formation of stars and planets is the question of the differences in composition which are expected as a result of different formation processes. Brown dwarfs are supposed to form from a central condensation in a nebula, but Jupiter, as all the planets, was formed within the solar nebula by accretion processes, starting from solid planetesimals. Since the planetesimal composition is depleted in the most volatile elements, they induce a planetary composition different from the nebula. Indeed, the composition of Jupiter is found to be enhanced in heavy elements. In stellar terminology, Jupiter is described as having a high metallicity, compared with the solar composition. On the other hand, a brown dwarf forming from its own nebula would be expected to have a composition mostly similar to its nebula. Nevertheless, it must be emphasized that many stars are formed in multiple systems, and the theory of formation of binary systems is not fully understood yet. The distinction between star and planetary formation may therefore be more complex than this simple description. When more measurements become available on brown dwarfs, and, it is hoped, on extrasolar giant planets, the measurement of metallicity will be a major point to discriminate between possible scenarios of formation. This is probably one of the most challenging problems for an understanding of the formation of giant planets.

Bibliography

Clarke J T *et al* February 2002 Ultraviolet emissions from the magnetic footprints of Io, Ganymede and Europa on Jupiter *Nature* **415** 997–1000

Gladstone G R *et al* February 2002 A pulsating auroral X-ray hot spot on Jupiter *Nature* **415** 1000–1003

1996 *Science* **272** 837 *et seq* (Probe issue)

J. Geophys. Res. **103** (special issue, September)

1999 *Planet. Space Sci.* **47** (special issue)

The following websites are also useful:

Galileo mission: http://www.jpl.nasa.gov/galileo

Planetary Data System:
http://www.pds.nasa.gov/pds_home.html

Pierre Drossart

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Jupiter: Magnetosphere

Jupiter's magnetosphere is the cavity surrounding the planet which contains and is controlled by Jupiter's magnetic field (figure 1). It reaches to a distance of ~60 Jupiter radii on the dayside of the planet (one Jupiter radius, R_J, is taken to be 71 373 km), and extends into a long comet-like tail on the nightside which has a diameter of ~300–400 R_J and a length of at least 3000 R_J. Jupiter's field is confined to this cavity by SOLAR WIND PLASMA WAVES which flow away from the Sun throughout the solar system, and bound the magnetosphere on the outside. Because the solar wind is supermagnetosonic in the planet's rest frame, a bow shock stands in the flow upstream from the cavity, at a distance of ~75 R_J on the dayside, across which the solar wind is slowed, compressed, and heated. The location of the boundary of the magnetosphere, the magnetopause, is determined by the condition of pressure balance between the shocked solar wind plasma (termed the magnetosheath) on one side, and the magnetospheric plasma and field on the other. On the inside of the cavity, the magnetospheric magnetic field lines extend down into the ionosphere and upper atmosphere (thermosphere) of the planet (see JUPITER: ATMOSPHERE), such that the magnetosphere, ionosphere, and thermosphere are strongly coupled together. The plasma inside the magnetosphere contains contributions from both Jupiter's ionosphere and the solar wind, consisting mainly of ionized hydrogen and helium, but by far the most important plasma source is the Galilean moon IO, which orbits deep inside the magnetosphere at a radial distance of 5.9 R_J, and liberates about 1 ton s^{-1} of sulfur dioxide gas (comparable to the production rate of an active comet). Consequently, the magnetospheric plasma is dominated by the presence of ions of sulfur and oxygen. The most important source of momentum and energy for the magnetosphere is the planet's rotation at a period of 9 h 55 min, such that the most important dynamics result from the presence of the Io plasma source deep within the rapidly rotating magnetosphere. However, as in the case of the EARTH'S MAGNETOSPHERE, solar wind coupling at the magnetopause may contribute significantly to the dynamics of the outer regions, and certainly to the formation and properties of the nightside magnetic tail.

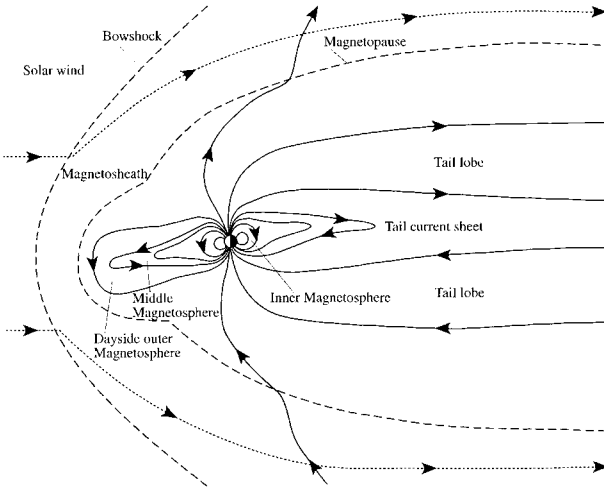


Figure 1. Sketch of Jupiter's magnetosphere in the noon–midnight meridian plane, with the Sun to the left and the solar wind blowing from left to right. The arrowed solid lines are the magnetic field lines, while the dashed lines are the magnetopause and bow shock, as indicated. The dotted lines are plasma streamlines.

Discovery and exploration

The discovery that Jupiter generates an intense magnetic field via dynamo currents flowing in its interior, was made in the 1950s from the observation that the planet is a source of radio emissions in the decametric (~10 MHz) and decimetric (~1 GHz) wave bands (see MAGNETOSPHERES: JUPITER, RADIO EMISSIONS). The decimetric emission is SYNCHROTRON RADIATION generated by gyrating energetic (~10 MeV) radiation belt electrons trapped by Jupiter's magnetic field near the equatorial plane within a few R_J of the planet. The spatial structure of this emission and its variations with planetary rotation indicated that the magnetic field is dipolar in form, like the Earth's, with the magnetic dipole axis tilted by ~10° relative to the planetary rotation axis. The POLARIZATION of the radiation, however, showed that the polarity of the field is opposite to Earth's, with field lines running from the northern hemisphere of the planet, via the equatorial regions, to the southern hemisphere. Whilst this decimetric emission is very steady in time, apart from a modulation at the planetary rotation period, the decametric radiation is instead characteristically bursty, and varies in intensity over several orders of magnitude on time scales of the order of minutes, with an averaged power around a hundred times that of the decimetric emission. Although the first to be discovered, by Burke and Franklin in 1955, the details of the decametric mechanism remain to be understood in detail. However, it is thought to involve emissions at the cyclotron frequency from ~keV electrons accelerated in the magnetosphere, which move along the magnetic field lines towards Jupiter's ionosphere. The upper cut-off frequency of 40 MHz then corresponds to the cyclotron frequency of the highest field strength accessible to these particles over the Jovian poles, equivalent to 1.4–10^{-3} T. This remarkably large field is about twenty times the strength of the Earth's polar field, despite the fact that Jupiter's radius is more than ten times that of the Earth. In 1964, Bigg also discovered that part of the decametric emission is directly influenced by the position of Io, thus providing the first evidence of a complicated electrodynamic interaction

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

between Io and the jovian magnetosphere–ionosphere system. In 1974, Brown using ground-based optical observations also discovered the existence of a large cloud of neutral sodium atoms surrounding Io, which was followed in subsequent years by optical observations of gas and plasma tori in the vicinity of Io's orbit, composed of the atoms and ions of sulfur and oxygen.

In situ exploration of Jupiter's environment began in November–December 1973 with the fly-by of the NASA PIONEER 10 spacecraft, followed by Pioneer 11 in 1974, VOYAGER 1 and 2 in 1979, the European Space Agency's ULYSSES in 1992 (en route to the Sun's polar regions; see SOLAR WIND: ULYSSES), culminating in the insertion into orbit of the NASA GALILEO MISSION spacecraft in 1995. At the time of writing only the results of the initial analyses of Galileo data are available, mainly concerning the interactions of the Galilean moons with the magnetospheric environment (see MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS). Most of the information described here, therefore, has been derived from the five earlier fly-by missions, augmented by radio and optical observations from Earth. The Pioneer spacecraft were instrumented principally to measure the jovian magnetic field and energetic particle environment at energies above ~1 MeV. The Voyager spacecraft made the first detailed *in situ* measurements of thermal plasmas at lower energies, at 10–100s eV and 10–100s keV, and of plasma waves. Similar measurements were also made by Ulysses. The trajectories of these spacecraft are shown projected onto Jupiter's orbital plane in figure 2, together with the locations of nominal magnetopause and bow shock positions. The *X* axis points towards the Sun, and the *Y* axis from dawn to dusk across the magnetosphere. It can be seen that all these spacecraft explored the pre-noon dayside magnetosphere on their inbound passes, and that Pioneers 10 and Voyagers 1 and 2 passed through the pre-dawn nightside sector outbound. All these trajectory segments were confined to the near-equatorial region. In contrast, Pioneer 11 passed out of the magnetosphere near noon at northern latitudes of ~33°, while Ulysses exited near to dusk at southerly latitudes of ~37°. Since the APOAPSIS of the Galileo spacecraft has also been confined mainly to the local time sector between midnight and dawn during the main phase of its mission, it can been seen that, while the dawn side of the magnetosphere has received considerable observational study, the dusk side remains almost unexplored.

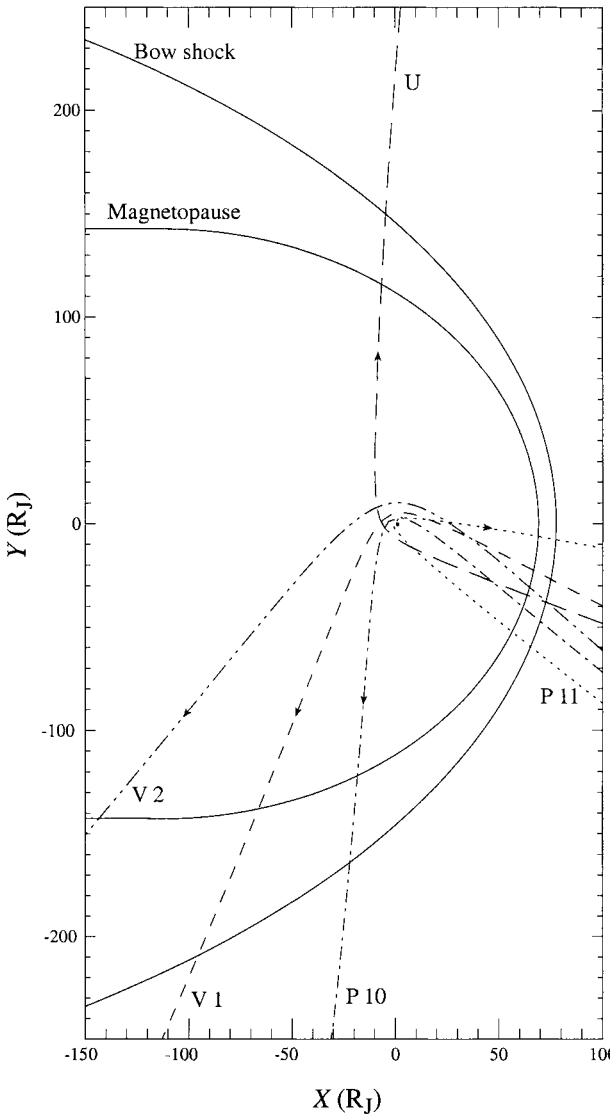


Figure 2. Trajectories of the five fly-by spacecraft relative to Jupiter, projected onto Jupiter's orbital plane. *X* points towards the Sun and *Y* from dawn to dusk. P 10 and P 11 refer to Pioneers 10 and 11, V 1 and V 2 to Voyagers 1 and 2, and U to Ulysses. Arrows are plotted in the direction of spacecraft motion on the outbound portions of the trajectories. Also plotted are the average positions of the bow shock and magnetopause (adapted from D E Huddleston *et al* 1998 *J. Geophys. Res.* **103** 20075–82).

Jupiter's magnetic field

Internal planetary field

Although the main features of the magnetic field produced by Jupiter's internal currents were inferred from radio observations prior to the first measurements by spacecraft, *in situ* data are required for detailed characterization. A recent model based on fits to the fly-by data, constrained also by Earth-based observations of jovian auroral emissions associated with the magnetic footprint of Io, indicates a best-fit centered dipole axis which is inclined at 9.5° to Jupiter's spin axis towards system III longitude 201°. The corresponding dipole moment is $4.26–10^{-4}$ T R_J^{-3}, such that the surface field is $0.4–10^{-3}$ T at the magnetic equator and approximately double that at the poles (neglecting the dynamical flattening of Jupiter's figure which increases the dipole polar surface field to ~$1.0–10^{-3}$ T). However, large quadrupole and octupole moments are also present which produce

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

significant asymmetries in the near-planet field, in particular increasing the peak surface field in the northern polar regions to ~1.5–10^{-3}, in agreement with the value obtained from the 40 MHz upper cut-off in the decametric emission.

Size of the magnetosphere

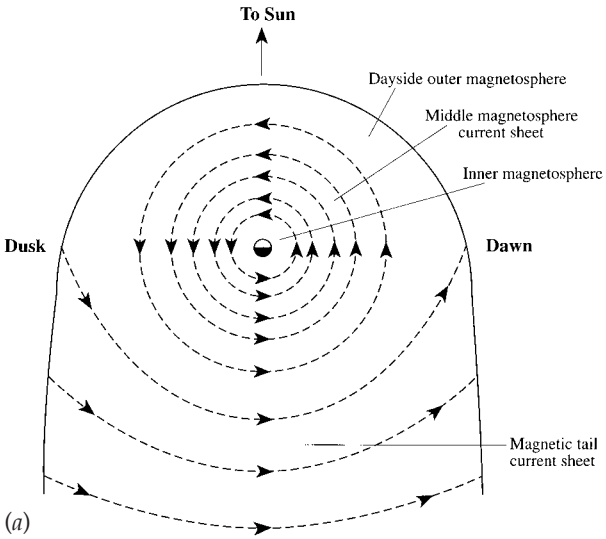
Using the above value of the dipole moment, a simple estimate can be made of the expected size of the magnetospheric cavity in the solar wind, based on pressure balance across the magnetopause. If magnetospheric plasma pressure is neglected such that the internal pressure is wholly magnetic, and taking a nominal solar wind dynamic pressure of ~0.1 nPa at Jupiter's orbit, a simple calculation shows that in the subsolar region (i.e. in the equatorial region near noon) the boundary should lie at a distance of ~35 R_J. In contrast, the fly-by observations of the dayside magnetopause, suitably adjusted to take account of the outward flaring of the boundary away from noon (see figure 2), indicate a typical distance of ~60 R_J. Furthermore, the response to variations in solar wind dynamic pressure indicates a much 'squashier' system than expected, with subsolar boundary positions varying between ~40 and ~80 R_J, indicating a position which varies inversely as the ¼ or (1/5) power of the dynamic pressure, compared with the (1/6) power expected for a dipole magnetic field. The reason for the great inflation of the magnetosphere, and for its 'squashy' nature, lies in the fact that (unlike Earth), the magnetospheric plasma makes a substantial contribution to the internal pressure, at least comparable to that made by the field.

Inner magnetosphere

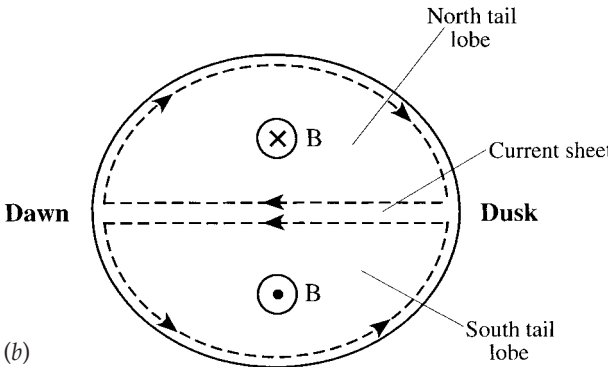
The structure of the magnetic field within the cavity as revealed by spacecraft observations is indicated schematically in figure 1, which shows a sketch of the field lines in the noon–midnight meridian plane. Four basic regions are identified, whose boundaries are defined by the magnetic field lines. The inner magnetosphere is defined by the torus-shaped region of field lines which cross the equatorial plane within ~5 R_J of Jupiter's center. The field in this region is essentially an undisturbed planetary field, which contains the radiation zone of synchrotron-emitting energetic electrons. This region is distinguished from the middle magnetosphere, which bounds it on the outside, by the lack of significant electrical currents flowing in the plasma, such that the outer boundary of the region essentially coincides with the inner boundary of the Io plasma torus where strong equatorial currents begin. However, the inner region is affected by the fringing field of the latter currents, which produce a nearly uniform northward field in the inner region of strength ~200 nT. This field is directed opposite to the southward planetary field in the equatorial plane, but is much weaker than the latter, since even at ~5 R_J the equatorial dipole field is still 3400 nT.

Middle magnetosphere

As indicated above, the middle magnetosphere is characterized by the presence of strong azimuthal (eastward) electric currents flowing in the plasma near the equatorial plane whose magnetic effect distends the field lines outwards from the planet, as shown in figure 1. In the equatorial plane this region extends from the inner edge of the Io plasma torus at ~5 R_J to an outer limit on the dayside which depends upon the degree of extension of the magnetosphere, and lies typically ~15 R_J inside the magnetopause. For a typical ~60 R_J magnetopause position at noon, therefore, the outer edge of the middle



(*a*)



(*b*)

Figure 3. (*a*) Sketch of the current system in Jupiter's magnetic equatorial plane, showing the eastward current of the middle magnetosphere region, which closes round the planet, and the dusk-to-dawn currents of the tail current sheet, which separates the tail lobes, and closes around the magnetopause. (*b*) Sketch of the field and current in a cross section through the tail, looking down the tail from the planet. The north tail lobe field points away from the planet (circled cross), while that of the southern lobe points towards the planet (circled dot).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

magnetosphere current sheet lies at ~45 R_J. On the nightside, the current sheet merges continuously into the current system of the magnetospheric tail. Figure 3(*a*) shows a sketch of the equatorial currents of the middle magnetosphere which, to a first approximation, close azimuthally around the planet, and the adjacent equatorial tail currents which flow from dusk to dawn across the system. The tail currents close north and south over the tail lobe magnetopause to form two D-shaped solenoidal current systems back-to-back, as shown in the cross section in figure 3(*b*). The full thickness of the current sheet is ~5 R_J, and the current within it, and hence the perturbation field produced, falls with distance from the planet approximately as $r^{-1.5}$. The value of the exponent depends somewhat on distance and local time, but is in any case considerably smaller than the r^{-3} dependence of the planetary dipole field. Consequently, while in the vicinity of Io's orbit the perturbation fields of the current sheet are relatively small compared with the planetary field (as indicated above), such that the field lines remain basically dipolar in form, they become dominant at distances beyond ~15 R_J. In the latter region the low latitude field is thus characterized by strong radial fields which point outwards from the planet north of the current sheet, and inwards to the planet south of the sheet, as shown in figure 1. At the center of the current sheet the field is weak and points south. The total current flowing in the annular middle magnetosphere current sheet (excluding the tail system) is ~200 MA.

Within ~30 R_J of the planet the current sheet is centered close to the planetary dipole equatorial plane, tilted at 9.5Â° to Jupiter's spin axis as indicated above. With increasing distance, however, the current sheet increasingly departs from this plane due to two effects. The first is that, as the dipole axis rotates around with the planet, the information about the rotation is communicated to the outer regions at a finite speed. Consequently, at a given distance from the planet the effective position of the dipole is increasingly retarded in azimuth relative to its true position. Within ~30 RJ the delays are insignificant, but beyond this distance the effective information propagation speed is ~40 RJ h–1 (~750 km s–1), leading to observable delays. The second effect is the 'hinging' of the current sheet in the magnetic tail, where the plane of the sheet departs from the dipole equator and becomes aligned with the tail axis (defined by the flow of the solar wind), as shown in figure 1. This effect becomes apparent at down-tail distances beyond ~35 R_J.

Outer magnetosphere

The outer magnetosphere is a dayside region observed during the inbound passes of the fly-by spacecraft of uncertain local time extent, in which there is no evidence of an equatorial current sheet. Instead, the field in this region, although rather variable, points on average to the south in the equatorial region, in the direction of the planetary equatorial field, and overlies the middle magnetosphere closer to the planet. The region is thus bounded by the magnetopause on one side, and by the middle magnetosphere on the other. A TRANSITION REGION of disordered field may occur between the outer and middle magnetosphere regions (particularly well marked on the inbound Ulysses pass), where the equatorial field undergoes sharp changes indicative of the presence of plasma current layers, but which are not ordered by magnetic latitude or planetary rotation period. The thickness of the outer magnetosphere in the equatorial region is ~15 R_J, and the average equatorial field strength varies between ~5 and ~15 nT, depending on the degree of extension of the magnetosphere. In all cases, however, the field is significantly stronger than that of the planetary dipole, due to the fringing effect of the current sheet field.

Magnetic tail

The magnetic tail on the nightside of the planet consists of two tail lobes of oppositely directed flux where the field is relatively uniform and strong, pointing away from the planet in the northern lobe and towards the planet in the southern lobe, which are separated by the thin (few R_J) equatorial current sheet mentioned above, where the plasma carries the required electric current from dusk to dawn across the tail. Overall, the tail has an approximately cylindrical shape with a diameter of ~300–400 R_J, such that the two lobes are D-shaped in cross section, as shown in figure 3(*b*). The field strength in the lobes falls with increasing distance as the tail expands in radius, reaching ~2 nT at a down-tail distance of ~150 R_J as measured by Voyager 2 (see figure 2). By analogy with Earth, it seems probable that the tail is formed by MAGNETIC RECONNECTION at the dayside magnetopause between the interplanetary and magnetospheric field, which results in the formation of open magnetic flux tubes which map from the planet's polar cap, through the magnetopause, and into interplanetary space (figure 1). The flow of the solar wind then carries them onto the nightside of the planet, where they are stretched out to form the tail lobes. As they are stretched out down-tail, the open tubes slowly sink in towards the center plane of the tail, where they reconnect again within the current sheet, returning closed magnetic flux tubes, attached to Jupiter at both ends, back towards the planet. However, where the reconnection and return flow take places in the jovian tail is at present unknown. On the basis of this model we may estimate that the lobe field lines flow towards the current sheet at a speed of a few tens of km s^{-1}, such that they remain 'open' for several Earth days. Consequently, the length of the tail (equal to the speed of the solar wind times the length of time for which lobe field lines remain open) will be several thousand R_J (i.e. a few AU). The amount of magnetic flux in each lobe is ~4–10^{11} Wb, corresponding to a region of

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

~10° latitude surrounding each magnetic pole at ionospheric heights.

Bending of the field meridian planes

For a pure dipole, field lines lie in planes of constant longitude relative to the dipole axis. However, the planetary field lines at Jupiter are noticeably bent out of such meridian planes, associated with the existence of azimuthal fields directed either eastward (in the sense of planetary rotation) or westward around the magnetic axis. The bending of meridian planes is of interest because it provides information about the transfer of magnetic forces between one plasma region and another connected by the magnetic field. Two major effects are illustrated in figure 4(*a*), which shows a view looking down onto the jovian magnetosphere from above the northern pole, and shows field lines in two regimes of latitude mapping out from Jupiter's ionosphere to the equatorial plane. The field lines in the high-latitude region mapping into the outer magnetosphere are bent away from noon and towards the tail as a consequence of the solar wind interaction, such that in the northern hemisphere (as shown) the perturbation fields are eastwards on the dusk side of the magnetosphere and westward on the dawn side (and vice versa in the southern hemisphere). The current system responsible is that of the magnetopause and tail (as sketched in figure 3(*b*)).

At lower latitudes, however, the field lines mapping to the middle magnetosphere current sheet show a consistent sense of bending independent of local time (insofar as it has been measured), with azimuthal fields which are westward above the current sheet (as shown), and eastward beneath it. This pattern of field bending is associated with a torque on the magnetospheric plasma which acts to spin it up towards the angular velocity of planetary rotation, i.e. it is associated with the transfer of angular momentum from the thermosphere/ionosphere to the magnetosphere, as will be discussed further below. The current system responsible is sketched in figure 4(*b*), and involves a radially outward current in the equatorial plane which closes along magnetic field lines through equatorward currents flowing in the ionosphere. Typically, outside the current sheet the observed azimuthal field is a small fraction of the radial field, such that the radial currents are a correspondingly small fraction of the azimuthal current. Figure 4(*a*) shows that the two effects discussed here produce similar effects which are difficult to separate in the dawn magnetosphere where most observations have been made. However, the Ulysses observations at dusk demonstrate that the two effects are separately present, one dominant in the middle magnetosphere, the other in the outer magnetosphere.

Jupiter's plasma populations and their dynamics

Sources of plasma mass and momentum

The nature of the plasma dynamics in a planetary magnetosphere depends on the nature of the plasma sources and sinks, and the nature of the transport processes which convey the plasma from the former to the latter. The plasma sources include the solar wind at the outer boundary and the planet's ionosphere at the inner boundary, together with the surfaces and atmospheres of any moons that happen to orbit within the cavity. The sources of momentum include the antisunward flow of the solar wind on the outside, and the planet's rotation

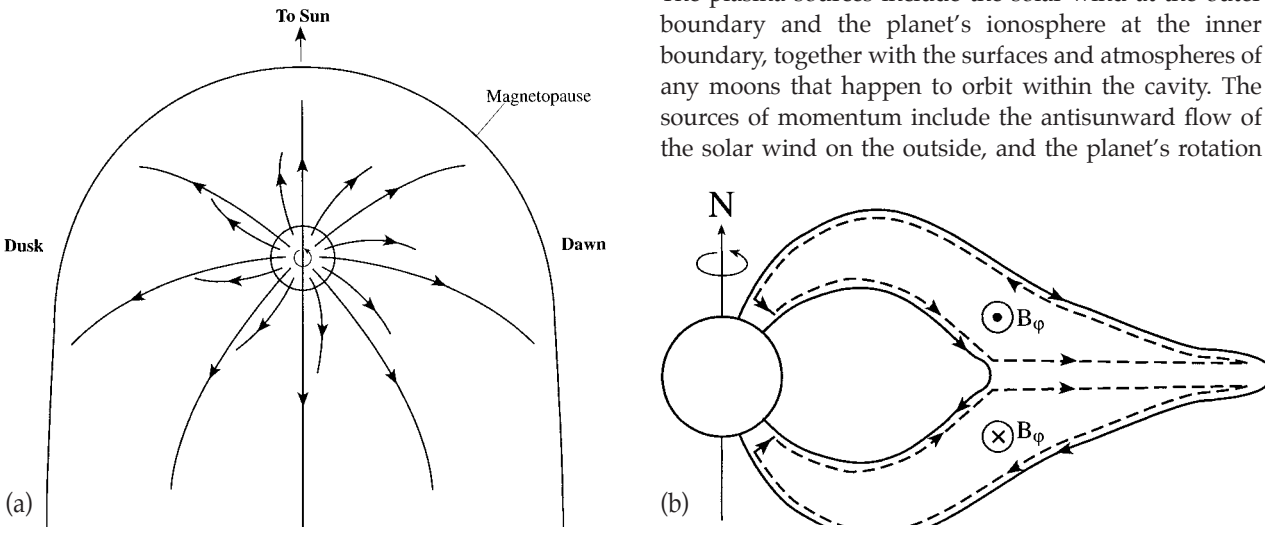


Figure 4. (*a*) Sketch of field lines emanating from the northern hemisphere of Jupiter projected onto the equatorial plane, showing the bending of the field lines out of meridian planes. High-latitude field lines mapping to the outer parts of the magnetosphere are bent away from noon by the interaction with the solar wind. The current system responsible is the magnetopause–tail system. Lower-latitude field lines mapping to the middle magnetosphere current sheet are bent consistently in the clockwise sense, associated with the transfer of anticlockwise planetary angular momentum from the thermosphere/ionosphere to the magnetosphere. (*b*) Sketch of the current system associated with planetary angular momentum transfer. The arrowed solid lines are magnetic field lines, and the dashed lines show the direction of current flow. The circled symbols marked B_φ indicate the direction of the azimuthal perturbation magnetic field produced by these currents, out of the diagram north of the current sheet, and into the diagram south of the sheet.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

on the inside. Solar wind interaction at the boundary carries magnetospheric flux tubes from the dayside to the nightside in the outer regions of the magnetosphere, as discussed above in relation to the formation of the magnetic tail, from which closed flux tubes eventually return flowing sunward through the central regions of the magnetosphere. In the absence of such flows, the magnetospheric plasma and field will rotate with the planet, the angular momentum being transferred by ion–neutral collisions at the feet of the field lines in the lower ionosphere.

Brice and Ioannidis in 1970 were the first to consider the relative importance of these two flow systems at Jupiter. Flows of plasmas and embedded magnetic fields are associated with an electric field which is transverse to both the velocity vector $\boldsymbol{V}$ and to the magnetic field $\boldsymbol{B}$, given by $\boldsymbol{E} = -\boldsymbol{V}-\boldsymbol{B}$. The overall strength of a flow system can then be measured by the voltage associated with its electric field, since by Faraday's law 1 V is equivalent to the transfer of 1 Wb s^{-1} of magnetic flux embedded in the flow. The electric field associated with the solar wind driven flow is directed from dusk to dawn across Jupiter's magnetosphere, and by analogy with the Earth, the associated voltage can be estimated to be ~1 MV (i.e. the transfer through this flow system of ~1 MWb s^{-1} from the dayside to the tail in the outer regions, and the return of the same amount, in the steady state, in the central regions). The electric field associated with rotation is directed radially outwards in the equatorial plane, and for rigid corotation with the planet (i.e. rotation with the same angular frequency as the planet) the associated voltage is ~400 MV. Rotation with the planet is thus by far the most important flow at Jupiter, although as indicated above, this statement does not preclude the dominance of solar wind driven effects in the outer regions and magnetotail.

Estimates indicate that both the solar wind and the ionosphere represent sources of a few -10^{28} ion s^{-1} for Jupiter's magnetosphere, consisting principally of hydrogen (i.e. protons, together, of course, with sufficient electrons to keep the gas electrically neutral overall). The corresponding mass sources are a few tens of kg s^{-1}. The ionospheric source also uniquely provides molecular hydrogen ions, H^+_2 and H^+_3, as minor constituents, while the solar wind provides He^{2+} and traces of heavier ions such as carbon. The identification of all these species within the jovian plasma has confirmed the presence of both sources. The major discovery of the Voyager fly-bys, however, was that the jovian system is not dominated by a hydrogen plasma as had previously been anticipated, but by a sulfur and oxygen plasma which originates from the sulfur dioxide atmosphere of the volcanic moon Io, which orbits at a distance of 5.9 R_J. The source rate is estimated to be ~3–10^{28} ion s^{-1}, similar to the solar wind and ionospheric sources, but because the ions are heavy, with a mean mass of ~21 amu, the corresponding mass source of ~1000 kg s^{-1} is overwhelmingly dominant. The sodium source at Io, though easily visible in optical emission, is less than this by a factor of around a hundred, and is thus negligible in overall terms. Recent estimates indicate that the moon EUROPA, which orbits at a radial distance of 9.4 R_J, is also a significant source of oxygen plasma originating from the surface water ice, with rates of ~2–10^{27} ion s^{-1} and corresponding mass rates of ~50 kg s^{-1}. The plasma dynamics of Jupiter's magnetosphere is therefore dominated by the consequences of the presence of strong heavy-ion sources lying deep within a rapidly rotating magnetosphere.

The Io plasma torus

The Io plasma which exists within Jupiter's magnetosphere originates principally from electron-impact ionization of the clouds of sulfur and oxygen atoms which orbit in the vicinity of Io, originating in the latter's atmosphere (see IO: PLASMA TORUS). The density of these atoms peaks at a few tens of cm^{-3} near the orbit of Io, with oxygen being the more numerous species as expected from the sulfur dioxide source, and falls off by an order of magnitude within ~1 R_J on either side. These neutral particles orbit with Io at Keplerian speeds of ~17 km s^{-1}, being influenced only by the gravitational force of Jupiter. When these atoms are ionized, however, the resulting ions and electrons suddenly sense the electromagnetic environment as well, that is to say the southward ~2000 nT magnetic field of the planet, and the ~0.1 V m^{-1}

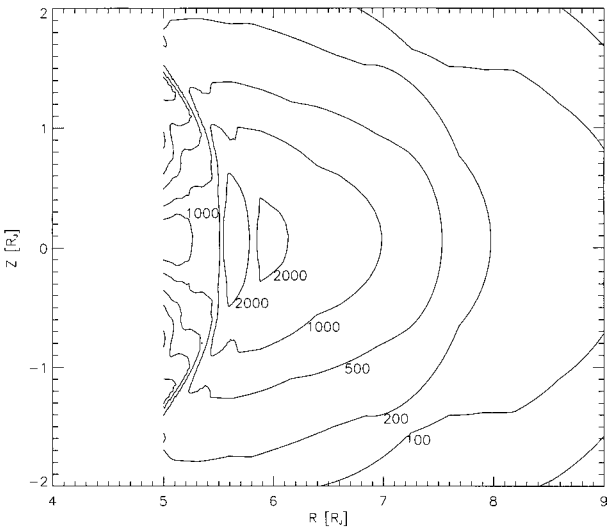


Figure 5. Contours of electron density in the Io plasma torus in the meridian plane, determined from Voyager 1 PLS data. The vertical scale is distance from the centrifugal equatorial plane, while the horizontal axis is distance from Jupiter's spin axis, both in R_J. The numbers on the contours give the electron density, equal to the ion charge density, in units of electrons cm^{-3}. (Taken from F Bagenal *et al* 1997 *Geophys. Res. Lett.* **24** 2119–22.) (Copyright 1997 by the American Geophysical Union.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

outward-directed electric field associated with the ~70 km s^{-1} flow of the near-corotating plasma. The effect of these fields is such as to cause the charged particles immediately to drift with the plasma at the corotation speed (they are 'picked up' by the plasma flow), and also to acquire a gyratory speed about the field lines equal to the difference between the corotation and Keplerian speeds, equal to ~55 km s^{-1}. This corresponds to a 'thermal' energy of ~250 eV for oxygen ions and ~500 eV for sulfur, but only 0.01 eV for electrons. Subsequently, the ions are cooled by Coulomb collisions with the electrons, and the consequently heated electrons are cooled by collisional excitation of the low-lying energy levels of the ions, thus leading to the observed optical emission from the torus.

The most detailed information about the low-energy plasma distribution which results from these processes was obtained during the inbound passage of Voyager 1. Figure 5 shows contours of electron density (equal to the ion charge density) which have been derived from these data. The principal population is the warm plasma torus which, in the equatorial plane, extends outwards from a jovicentric distance of ~5.6 R_J. The ions in this region consist of two populations, a suprathermal population of recently ionized few 100 eV particles (increasing in energy with increasing distance, up to ~2 keV at ~10 R_J), comprising ~10–20% of the population, and a cooled population with a temperature of ~60 eV (also increasing with distance up to ~300 eV at 10 R_J). The electron temperature in this region is ~10 eV. In the inner part of the warm torus, within ~7.5 R_J, there are approximately equal numbers of sulfur and oxygen ions, with the oxygen being principally O^+, while the sulfur is roughly equally divided between S^+ and S^{2+}. Outside this distance, the plasma is richer in oxygen, possibly due to the Europa source, with roughly equal numbers of O^+ and O^{2+}, and the density of S^{3+} becomes comparable with those of S^+ and S^{2+}. Inside ~5.6 R_J the plasma cools precipitately to form the cold plasma torus near the equatorial plane at distances of ~5.0–5.4 R_J. The ion and electron temperatures in this region are just a few eV, and the composition is a somewhat sulfur-rich combination of S^+ and O^+.

The distribution of the torus plasma along the field lines is determined by a balance between the plasma pressure force, the magnetic mirror force, the centrifugal force, the gravitational force due to Jupiter, and a field-parallel electric force which is required to ensure that the ion and electron charge densities are equal at all points. Apart from the latter, the most important physical effect is that the centrifugal force tends to compress the plasma at the centrifugal equator (the point of maximum distance along a field line from the planetary spin axis), while this compression is resisted by the plasma pressure. In the warm torus, equilibrium is reached for a scale height along the field lines away from the equator (the distance over which the density falls by a factor of ~2.7) of ~1 R_J. In the cold torus, however, the scale height is reduced by the lower temperatures to ~0.3 R_J.

The spatial structure of the torus plasma across the field lines reflects both the distribution of the atomic gas sources in the vicinity of the orbits of Io and Europa, together with the nature of the cross-field transport mechanism. Because the outwardly directed centrifugal force on the plasma is dominant (being much greater, certainly, than the inward force of Jupiter's gravity), flux tubes containing high-density plasma from the Io source will tend to 'fall' outwards to larger distances, restrained by the frictional drag of ion–neutral collisions at the feet of the field lines in the ionosphere. These flux tubes will be replaced by tubes containing lesser densities moving inwards, it being ultimately supposed, of course, that some mechanism (presently unknown) exists for emptying the flux tubes of their plasma content at large distances. This picture explains the basic cross-field structure shown in figure 5, which is indicative of rapid transport of warm dense plasma away from Jupiter at distances beyond the neutral gas sources. However, the inward transport from those sources towards Jupiter is much weaker, such that the plasma has time to cool by radiation, collapses onto the centrifugal equator due to the reduced pressure, and then recombines, thus explaining the existence and properties of the cold torus population.

Although the basic physical picture of outward radial transport of the torus plasma by centrifugally driven flux tube interchange thus seems clear, the details of the process are not yet determined, including basic time and spatial scales, and is simply parameterized in many theoretical models by an empirically determined spatial diffusion coefficient. Observationally, the warm torus plasma is found to pervade the equatorial current sheet out to the outer boundary of the middle magnetosphere at several tens of R_J. Due principally to the expansion of the flux tubes, the equatorial ion/electron charge density falls from at peak of ~3000 cm^{-3} (during the Voyager 1 flyby) near the inner edge of the torus at ~5.7 R_J, to ~70 cm^{-3} at ~10 R_J (as in figure 5), and down to ~0.1 cm^{-3} at several tens of R_J. The fraction of suprathermal particles in the population appears to increase with distance, however, such that the average energy also increases, rather than falling as expected for an expanding plasma. Typical values are a few 100 eV. Sporadic enhancements of low-temperature plasma are also observed in the dayside outer magnetosphere, correlated with decreases in the strength of the magnetic field. We may conjecture that these represent plasma fragments which have become detached from the middle magnetosphere current sheet.

Hot plasma population

The outwardly diffusing Io torus plasma is not, however, the only population which is present in Jupiter's

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

magnetosphere. Observations by the Voyager and Ulysses spacecraft have shown that a low-density but high-energy population is also present, consisting of roughly equal numbers of protons and heavy ions (mainly sulfur and oxygen). The number density and average energy of this population both increase on moving towards the planet, before falling near the inner edge of the warm torus. In terms of the above discussion, we may picture this low-density population to be transported radially inwards as the high-density torus plasma is transported radially outwards. Indeed, evidence has been found in Galileo data for sporadic localized inward injections of hot ions within the current sheet at distances between ~10 and ~30 R_J. These injections have aspects in common with SUBSTORMS at Earth, but at Jupiter they are not confined to the nightside but occur at all local times. Inward transport of this plasma results in compression and heating as observed, the energy required being derived ultimately from the outward 'falling' torus plasma. The presence of the hot plasma thus acts partially to suppress this transport.

The hot plasma density is much less than that of the warm torus plasma throughout the system, being ~10^{-2} to 10^{-3} cm^{-3} in the outer regions, increasing to perhaps ~1 cm^{-3} in the inner part of the Io torus. These particles thus make little contribution to the overall mass or charge density compared with the low-energy plasma, except perhaps in the dayside outer magnetosphere. However, their average energy is sufficiently large that they make the dominant contribution to the plasma pressure at all points except for the innermost part of the warm torus where the extreme warm plasma density (figure 5) and the falling hot ion temperature combine to produce comparable warm and hot plasma pressures. In the outer part of the magnetosphere the hot ion temperature is a few tens of keV, with the distribution having a non-Maxwellian high-energy tail extending above 1 MeV. The average energy then increases with decreasing distance, reaching a peak of ~2 MeV at ~7 R_J according to Voyager 1 measurements, before falling to ~100 keV inside Io's orbit at 5.9 R_J. Electrons are also present with comparable energies but significantly lower densities, such that they make a smaller contribution to the pressure. It is these energetic particles in the inner part of the warm torus which form the source of the radiation belts within the inner magnetosphere. From the vicinity of Io's orbit these particles are more slowly diffused inwards due to the presence of fluctuating electric fields driven by winds in the thermosphere, gaining further energy as they do so. This input is balanced in the steady state by particle flux and energy losses in the inner magnetosphere which are due to wave-induced particle precipitation into the atmosphere, absorption by ring material and moons, and (for electrons) synchrotron radiation.

A key feature of the hot ion population is that within the outer part of the middle magnetosphere its pressure is comparable with that of the magnetic field. As a consequence it 'inflates' the planetary field to form the current sheet structure observed in this region (figure 1), a process which also inflates the magnetosphere beyond the size expected for pressure balance between the solar wind and the planetary magnetic field alone, as discussed above. The distended field lines then provide the inward force (the $\boldsymbol{j}$–$\boldsymbol{B}$ Lorentz force of the azimuthal current) which in the steady state balances the outward pressure gradient of the hot plasma. The Io torus plasma also plays a role in current sheet formation, though a lesser one, since the field must also provide the inward force necessary to balance the centrifugal and pressure gradient forces of this population. With regard to the dominant hot plasma population, Voyager observations indicate that the equatorial pressure remains greater than that of the field throughout the outer part of the middle magnetosphere, while falling below that of the field inside ~10 R_J, due to the rapidly increasing strength of the dipole field. Consequently, as noted above, the perturbation fields produced by the plasma at and inside these distances becomes smaller than the planetary field, such that the field geometry then assumes a quasi-dipolar form.

Plasma flow and field bending

Observations of the plasma flow within the middle magnetosphere, extending outwards from the orbit of Io, generally confirm a primary plasma flow in the sense of planetary rotation as discussed above. However, departures from rigid corotation are observed which are due to two main effects. The first occurs in the main source region of the torus plasma in the vicinity of Io's orbit, where neutral atoms are ionized and 'picked-up' by the plasma. Because ANGULAR MOMENTUM is continuously provided to the newly ionized particles, the plasma in this region rotates more slowly than for rigid corotation, by an amount which is just such that the ion–neutral collisions in the lower ionosphere provide the required torque. Ground-based spectroscopic observations of the plasma, together with *in situ* data from Voyager 1, indicate that the plasma flow is slowed by ~4 km s^{-1} in a ~2 R_J-wide region centered near to Io's orbit by this effect (the rigid corotation speed is 74 km s^{-1}). Outside this region, near-rigid corotation is resumed.

As the torus plasma diffuses radially outwards, however, angular momentum must again be continuously added to maintain plasma rotation at near-rigid speeds. If no angular momentum is added, conservation of plasma angular momentum requires the azimuthal speed of the plasma to fall inversely as the distance from the spin axis, while for rigid corotation the speed must increase in direct proportion to this distance. In order to

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

maintain near-rigid corotation of an equatorial outwardly diffusing plasma, the angular momentum flux into the equatorial region must be constant, independent of distance. For radial mass fluxes corresponding to the Io source, it turns out that in the inner part of the torus, the required angular momentum flux can be supplied by ion–neutral collisions in the lower ionosphere for only minimal departures of the flow from rigid corotation. In this region, therefore, the plasma very nearly rigidly corotates with the planet outside the source region. However, with increasing distance from the planet a given area of the equatorial plane becomes connected via the magnetic field to a decreasing area of the ionosphere, which is located nearer to the rotation axis. Consequently, the ionospheric torque ultimately becomes small even for large departures from corotation. Thus beyond a certain radial distance, depending on the Io mass flux and the ion–neutral friction in the ionosphere, the azimuthal velocity is expected to break away from near-rigid corotation, to peak, and then to fall inversely with distance in the regime where the input of angular momentum becomes small. Voyager observations indicate that the flow is near to that expected for rigid corotation to equatorial distances of ~20 R_J, where the azimuthal flow speed is ~200 km s^{-1}, and falls below rigid corotation at larger equatorial distances. Voyager data for a relatively compressed magnetosphere indicate flows in the outer regions which do not fall with distance as then expected, but rather remain at values which are a factor of ~2 lower than rigid corotation speeds. When the magnetosphere is more extended during intervals of low solar wind dynamic pressure, however, the flow speeds in the outer regions are rather lower than this relative to rigid corotation, as indicated by data from both Pioneer 10 and Ulysses.

We should mention that, while in the above discussion we have described the input of angular momentum from the planet as being due to ion–neutral collisions in the lower ionosphere, this angular momentum is actually transmitted to the equatorial plasma via the bending of the field lines out of meridian planes which was shown in the inner part of figure 4(*a*), and whose current system is shown in figure 4(*b*). The $\boldsymbol{j}$–$\boldsymbol{B}$ force associated with the radially outward current in the equatorial plane (into the diagram) acts to increase the speed of a subcorotating equatorial plasma, while the $\boldsymbol{j}$–$\boldsymbol{B}$ force of the closure currents in the ionosphere acts in the opposite direction as a drag force on the rotation of the thermosphere. The corotation of the thermosphere is maintained to the extent allowed by viscous coupling to the corotating denser atmosphere beneath.

Flows in the outer magnetospheric regions are exceedingly uncertain at the present time. Outbound Voyager 2 measurements established the existence of a layer of plasma adjacent to the dawn tail magnetopause at ~150 R_J which was flowing antisunward, opposite to

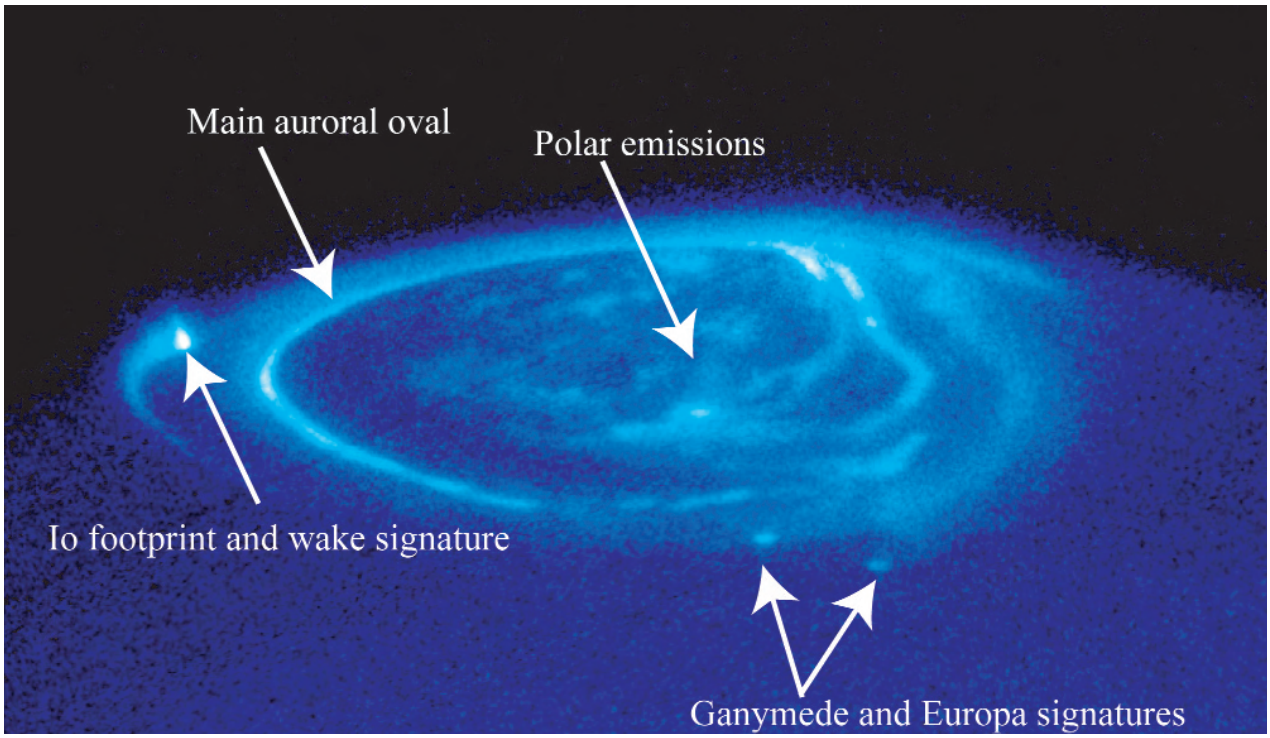


Figure 6. Ultraviolet image of the northern polar regions of Jupiter captured by the HST showing the main features of the auroral emissions as discussed in the text. Courtesy of John Clarke (University of Michigan) and NASA.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

the direction of planetary rotation. The flow speed was ~500 km s^{-1} in a layer a few tens of R_J wide. The nature of the overall dynamics, however, concerning, e.g., the ultimate loss process for Io plasma, and the interaction between planetary and solar wind driven flows in the tail, are yet to be determined.

Jovian auroral emissions

Auroral emissions from the polar regions of Jupiter were first detected in the UV by two independent spacecraft missions: the International Ultra–Violet Explorer (1979) and the Voyager spacecraft (1979 and 1980). These particular emissions result from direct excitation of atmospheric species by collisions with precipitating energetic magnetospheric electrons (see JUPITER: ATMOSPHERE). Auroral features were subsequently observed at various other wavelengths including the infrared (thermal emissions) and near–infrared (particularly from ionospheric H^{3+} ions). The IR aurora is excited by particle heating, or from Joule heating by large–scale current systems closing in the ionosphere. The auroral regions have also been associated with strong x–ray signatures, and furthermore are powerful sources of radio emissions.

Since these early measurements, data of much higher spatial and temporal resolution have become available both at visible wavelengths from the SSI instrument on the Galileo spacecraft, in the UV from the HUBBLE SPACE TELESCOPE (HST), in the infrared from Earth–based telescopes, as well as x–ray emissions from the CHANDRA x–ray observatory. These observations show that there are perhaps four main types of auroral emission. First is the Io footprint (IFT) emission, a ~1 MR emission which maps magnetically to the vicinity of the moon. There is also an auroral 'tail' stretching downstream of the IFT in the near–corotating flow. The Io related emissions are thus quite distinct from the other jovian auroral emissions, and are clearly visible in figure 6. It is understood that the corotational electric field which is developed by the relative motion between Io and the near–corotating magnetic field lines flowing past the satellite induces a potential difference of approximately 500 kV between the outer and inner faces of the moon. This potential is thoght to cause currents to flow from Io toward the ionosphere, both northward and southward, along the outer portion of the magnetic flux tube linking the satellite to the ionosphere. Return currents then flow along the inner portions of the flux tube towards Io. This circuit closed through the jovian ionosphere and also in the ionosphere of the moon, and is associated with the Io–controlled decametric radio emissions (see JUPITER: RADIO EMISSIONS). This 'unipolar inductor' model explains the spot of aurora associated with the upward directed portion of the current system (implying downward flowing electrons), i.e. the IFT, but does not explain the existence of the 'tail' of auroral brightness downstream of Io by as much as ~180°. However, recent work by Hill and Vasyliunas (2002) associates this downstream wake signature with the plasma currents which are required to accelerate the newly injected plasma from Io up to approximate corotation with Jupiter. Recent images indicate that footprint and wake auroras are also associated with Europa and Ganymede, and it seems likely that these features are due to the same downtail mass–loss process.

The second auroral component is the main auroral oval (MAO), which has been the subject of a plethora of reports recently. The MAO is the most significant emission in terms of energy output, and takes the form of circumpolar bands around both the northern and southern magnetic poles, consistently observed in all the wavebands mentioned above. Although this emission is of variable width (on average ~500–1000 km) and intensity (up to a few MR), it appears to be essentially continuous in local time, at dipole co–latitudes of ~16°, closer to the pole than the Io flux tube. It has been known for some time that this auroral region maps magnetically to the middle magnetosphere. Recently it has been suggested to be associated with the current system which maintains corotation of the equatorial plasma in the middle magnetosphere. Specifically the MAO is thought to be associated with the upward–directed field–aligned current at the inner edge of the magnetosphere–ionosphere coupling current circuit shown in figure 4(*a*). The former of the models cited above proposes an empirical model of the aurora, using estimates of the angular velocity of the plasma in the equatorial plane and a suitable magnetic field model. An interesting consequence of this model is that one would then expect the auroral brightness to anti–correlate with the dynamic pressure of the upstream solar wind. As the magnetosphere contracts, the angular velocity increases due to conservation of angular momentum, the currents go down and the aurora dims; conversely as the magnetosphere expands, the angular velocity decreases, the currents go up and the aurora brightens. It is thought that the hectometric radio emissions would also follow this modulation, although recent observations show that at least part of this radio emission exhibits impulsive increases which correlate positively with shocks in the solar wind.

The third component of the aurora is the high latitude diffuse emissions, which occur regularly and have a brightness of a few x 100 kR. These polar cap auroras are generally extended across the dusk side of the polar cap poleward of the MAO. At this time, a production mechanism for these emissions has not been suggested. More recent papers have also discussed a fourth feature which is seemingly a regular occurrence in the auroral regions. This feature appears consistently near magnetic local noon, and is reminiscent of the Earth's polar cusp. This is

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

referred to as a 'cusp–like' feature, which was reported by to be rapidly evolving, very bright (up to ~40 MR) and localised near noon. This feature lies poleward of the MAO and therefore it is conjectured that it may be controlled by pressure and/or magnetic field changes in the upstream solar wind. However, the intensity of this feature requires some form of acceleration mechanism, and could not be obtained simply by solar–wind particles flowing along open field lines (in the conventional 'cusp–like' sense).

The recent fly–by of Jupiter by the Cassini spacecraft en–route to Saturn (December, 2001), along with an intense HST campaign to complement Earth–based telescope observations, may help to unravel the response of the auroral regions to the upstream solar wind conditions monitored by Cassini.

Bibliography

Hill T W 2001 The jovian auroral oval *J. Geophys. Res.* **106** 8101

Pallier L and R Prangé 2001 More about the structure of the high latitude jovian aurorae *Planet. Space Sci.* **49** 1159

Southwood D J and M G Kivelson 2001 A new perspective concerning the influence of the solar wind on the jovian magnetosphere *J. Geophys. Res.* **106** 6123

Cowley S W H and E J Bunce 2001 Origin of the main auroral oval in Jupiter's coupled magnetosphere–ionosphere system *Planet. Space Sci.* **49** 1067

Waite J H Jr *et al* 2001 An auroral flare at Jupiter *Nature* **410** 787

Gurnett DA *et al* February 2002 Control of Jupiter's radio emission and aurorae by the solar wind *Nature* **415** 985

Results from the Pioneer, Voyager, and Ulysses Jupiter fly-bys are summarized, respectively, in the following publications:

Gehrels T 1976 *Jupiter* (Tucson, AZ: University of Arizona Press) p 1254

Dessler A J (ed) 1983 *Physics of the Jovian Magnetosphere* (Cambridge: Cambridge University Press) p 544

1992 Ulysses at Jupiter, *Science* **257** 1503–57

S W H Cowley and E J Bunce

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Jupiter: Rings

The faint Jovian ring system was discovered in images taken by the spacecraft Voyager 1 in March 1979 and imaged in more detail by its sister craft Voyager 2 four months later. The density of these rings appears to be about a million times less than that of Saturn's rings (see SATURN: RINGS). This explains why, situated so close to the bright planet, they had never been observed from the Earth. Pioneer 11, which flew by the planet five years before the VOYAGER MISSIONS, had observed sudden variations in the number of charged particles orbiting around Jupiter at certain distances from the planet. Some scientists suggested that the absorption of particles by a ring was a possibility. Five years later this hypothesis was proved to be correct!

Within the diverse class of known PLANETARY RINGS, Jupiter's ring structure represents one of the extremes. Along with Saturn's E and G rings, it is among the most tenuous systems yet discovered. Such rings are of particular interest because the processes that dominate within them are very different to those in more substantial ring systems. Mutual collisions between particles are rare, and the particles making up these tenuous rings are expected to be dominated by electromagnetic effects and radiation drag forces. Small particles should also be short-lived in the high-radiation environment of Jupiter.

In this article, the properties of the Jovian ring system are reviewed; emphasis is placed on the findings of the GALILEO MISSION, since this work refines several interpretations of Jupiter's rings.

Morphology

The Jovian rings are the archetype of ethereal planetary rings: bands of very low optical depth containing micron-sized particles. This faint and diffuse ring system, as imaged by Voyager, consisted of three separate components: (1) a relatively bright narrow ring, referred as the main ring; (2) a vertically extended halo interior to the main ring; (3) a very tenuous ring, called the gossamer ring, exterior to the main ring.

Spectacular images from the Galileo spacecraft changed this picture somewhat. Galileo has been orbiting Jupiter and its moons since 1996. The spacecraft took three dozen images of the rings and small moons during orbits of Jupiter in 1996, and as a result of much improved observations by this mission, we now know that the rings have the following properties.

(1) The thin main ring is about 7000 km wide, extends between 1.71 and 1.81 Jovian radii ($1\,R_J = 71\,399$ km). The normal optical depth τ is about 10^{-6}. This ring has patchy (20–30%) brightness, with a maximum at $1.79R_J$ and a minimum around 127 850 km in the vicinity of Metis (figure 1).

(2) The thick toroidal halo ($\tau = 10^{-6}$) extends inwards from the main rings at $1.71R_J$ down to $1.4R_J$. Roughly 50% of the halo's brightness originates from a region within 700 km of the equatorial plane, while the halo is visible up to 10 000 km above and below the plane (figure 1).

(3) The gossamer ring, previously seen only in one Voyager image, is in fact two faint rings (with $\tau = 10^{-7}$) that are bounded by the orbits of the small ring-moons, Amalthea and Thebe (figure 2). The first, Amalthea's gossamer ring, is visible between the main ring's periphery and about $2.55R_J$, i.e. inwards of Thebe's orbit. Very faint material continues past Thebe's orbit to $3.5R_J$. The gossamer rings have thicknesses that are comparable to the maximum elevations of their bounding satellites above Jupiter's equatorial plane. When viewed edge-on, the top and bottom edges of the gossamer rings are about two or three times brighter than the body of the rings.

The composition of the ring particles could not be deduced from the Voyager images alone, ground-based observations are still very difficult, and the Galileo images are being analysed. However, to first order, we are concerned only to know whether the particles are made of ice or of rock (metal particles are unlikely on cosmochemical grounds). Because of the very short lifetime of micron-sized ice grains, it seems probable that the grains are rock rather than ice. Moreover, the small satellites closest to the ring are known to be rocky by virtue of their low albedo (see JUPITER: SATELLITES). This is also suggestive of a rock ring rather than an ice ring. Relative spectral reflectivity measurements show the ring material to be reddish, indicating that the rings could consist of small silicate grains.

Nature and origin

Particle lifetimes

The particles making up the visible Jovian ring are quite tiny since the distribution of particle sizes shows a peak around 2 μm. Such micron-sized grains reside in the fierce magnetospheric environment (see JUPITER: MAGNETOSPHERE). Accordingly, sputtering by energetic ions and electrons can be effective in destroying ring particles because the ejection process is efficient and because large fluxes of energetic particles are found in Jupiter's magnetosphere. Impacts by micrometeroids and other dust can also abruptly cause the demise of the rings. Elimination trough sublimation is not competitive on the non-icy materials at the distance of Jupiter from the Sun, but could be if instead the particles were water ice. The different lifetimes of a micron-sized grain in a Jovian ring against loss are summarized in table 1.

While ring grains are eroded and destroyed, their orbits are expected to evolve secularly under the influence of drag forces (plasma drag dominates). Table 1 shows that evolution times are similar to lifetimes for erosion and destruction.

Although none of these estimates is well determined because some pertinent parameters are not known accurately, the basic finding for Jupiter's rings is inescapable: particle lifetimes are short ($\leq 10^{2-3}$ years). The rings must be continually regenerated, by collisions with parent bodies, which may be the known small ring-moons or other unseen ones.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

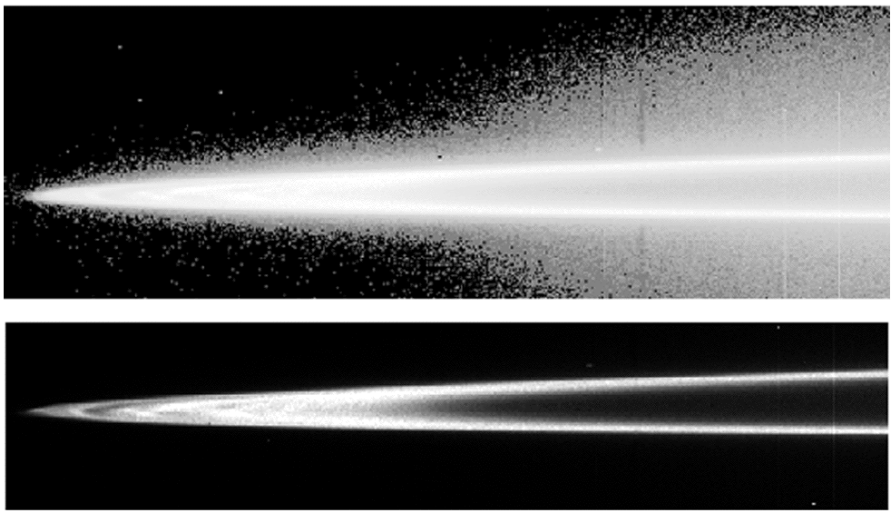


Figure 1. Main ring and halo system. Mosaic of images of Jupiter's rings taken by the Galileo spacecraft. The planet is to the right of this mosaic. The two images have been processed to show (top panel) the out-of-plane halo and (bottom panel) the flat main ring.

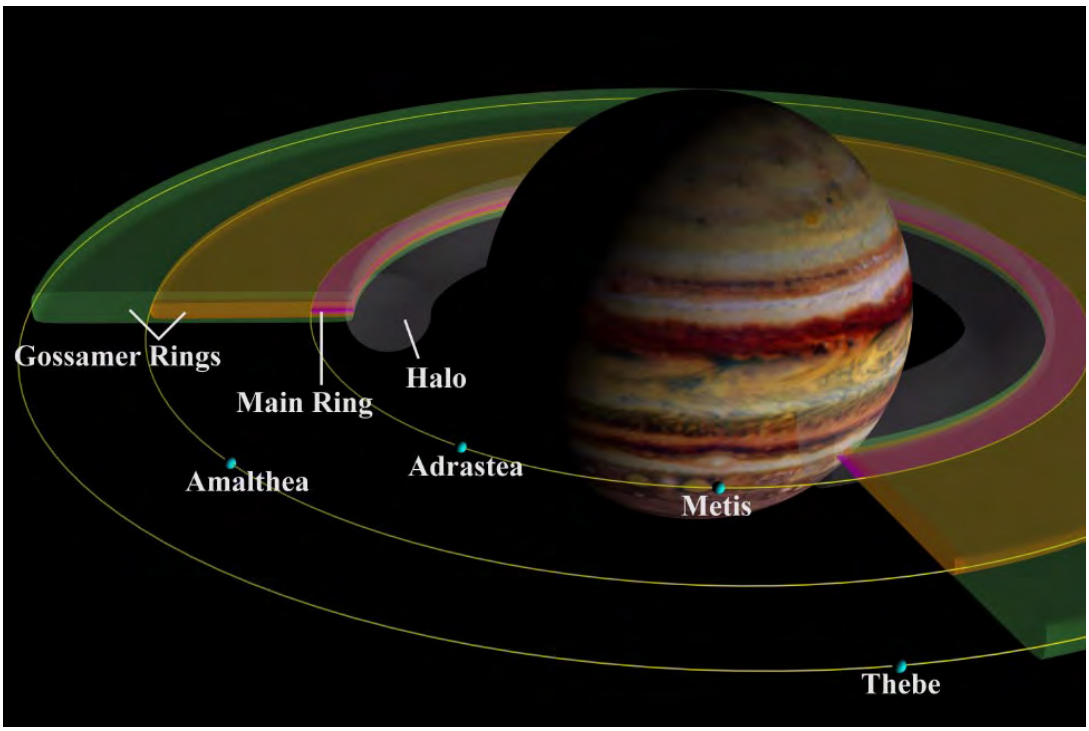


Figure 2. The entire Jovian ring system and small inner satellite orbits. This schematic drawing depicts the major components of Jupiter's rings and the four small satellites, Metis, Adrastea, Amalthea and Thebe. The relationship of the satellites' orbits to the rings can be seen. (Courtesy of Jonathan Joseph, Cornell University.)

The role of the known embedded satellites

A rich family of satellites surrounds Jupiter. In particular, there are four small inner satellites close to Jupiter: moving radially outward from Jupiter, we see Metis (mean radius $R = 20$ km), Adrastea ($R = 10$ km), Amalthea ($R = 90$ km) and Thebe ($R = 50$ km). Discovered in 1892 by Edward Barnard, Amalthea was considered to be the closest satellite to Jupiter for 87 years, until three other satellites, of which two are now the closest, were discovered by the Voyager probes in 1979.

Before the Galileo probe, it was still unclear what role, if any, was played by the satellites in the structure

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

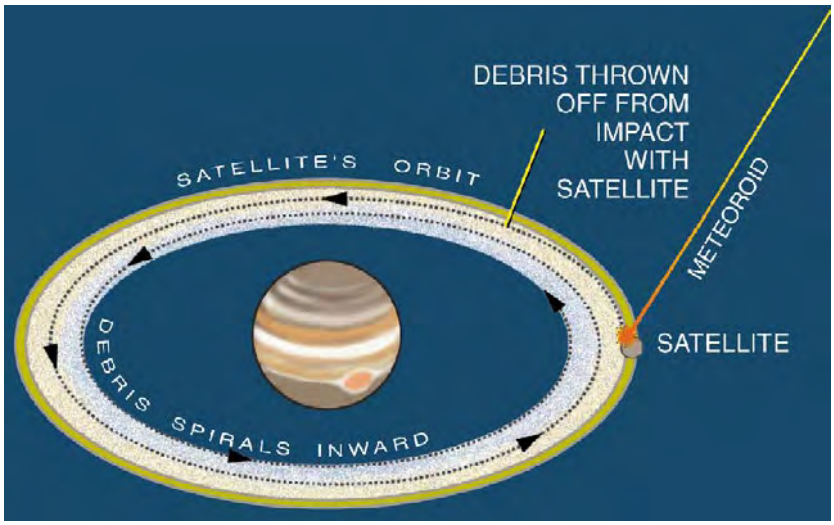


Figure 3. A graphic representation of the ring-forming process. When a meteoroid strikes one of the satellites, fine dust is thrown up, and gradually spirals inward toward the planet. (Courtesy of Jim Houghton.)

Table 1. Characteristic times for minor non-icy grains in the Jovian rings.

Process	Lifetime (years)
Loss	
Magnetospheric erosion	$10^{3\pm1}$
Meteoroid erosion	$10^{5\pm1}$
Evaporation	10^{7}
Evolution	
Radiation forces	10^{2}–10^{5}

of the Jovian rings. But the unique morphology of the gossamer rings, as imaged by Galileo, can be generated only if the ring particles are collisional ejecta evolving radially inward from the satellites Amalthea and Thebe. The source of Jupiter's ring system and how it works is described below and in figure 3.

The environment of the small satellites closest to Jupiter is very hostile. Accordingly, their surfaces are exposed to hordes of energetic particles moving within the magnetosphere of the planet and to the bombardment by gas and dust particles originating in the volcanoes of Io, as well as by interplanetary meteoroids, or fragments of comets and asteroids, at speeds greatly magnified by Jupiter's huge gravitational field. However, satellites can be either sources or sinks for rings, depending how much ejecta is created in an impact and what fraction of that ejecta has sufficient speed to escape.

If the moon is too big, dust particles will not have enough velocity to escape the moon's gravitational field. With diameters of a few tens of kilometers, the small Jovian moons are perfectly suited for the ejection of debris created by the meteoritic bombardment. Moreover, typical escape speeds from the satellites are many tens of m s^{-1} for Amalthea and Thebe, and between 0–10 m s^{-1} for Adrastea and Metis. These escape speeds are lower than the orbital speeds, so ejected particles enter orbits that are much like those of their source satellites, both in their distance from Jupiter and in their slight tilt relative to Jupiter's equatorial plane. Due to Jupiter's oblateness, the tilted orbits wobble back and forth in only a few months. After a few years, the orbits have smeared into a cylindrical band. Observed from near the equatorial plane, the band looks rectangular rather than the familiar elliptical arc seen on Saturn's rings and Jupiter's main ring. As the grains drift inward, the rectangle will brighten toward the planet, and will contain an embedded bright bow-tie shape because the drifting particles retain their initial orbital inclinations.

Since ring-moons of 10 km radius are the most efficient source of collisional ejecta, the most obvious sources of the thin main ring are meteoritic impacts onto the satellites Adrastea and Metis. Note that each ring's vertical thickness would be affected by the inclination of the orbits of the inner satellites, which are struck by meteoroids. Thebe's inclination is the largest, Amalthea has a significant inclination, and Adrastea and Metis have no inclination. Thus, the thickness of each ring increases as you travel away from Jupiter. This also explains why the main ring is a *thin* ring.

Formation of the halo

The halo is probably composed of material that is evolving inward from the main ring. Over the time of an orbital revolution, the size of the typical particles shrinks since lifetimes are similar to evolution times, as shown in table 1. This explains why the visible halo grains are simultaneously closer to the planet and smaller than those of the main ring. Since the smaller grains are more susceptible to the electromagnetic forces, the halo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

particles are some main ring particles that have been driven vertically by these electromagnetic forces. When halo orbits become highly distored, particles are lost in Jupiter.

Conclusion

In short, the Galileo probe's findings showed the intimate relationship between the Jovian rings and the satellites orbiting within them. These rings should be regarded as one example of the broad class of faint planetary rings. The G and E rings of Saturn certainly fall into this category, and the F ring shares some key characteristics. The dominant physical processes for diffuse rings are often different from those that apply to the denser rings of Saturn and Uranus (see also URANUS: RINGS). In particular, small particles last for limited times and so the rings must be continually regenerated if these rings are to be long-term members of the solar system. For the first time, scientists have identified the source and the nature of the Jovian ring material: dust blasted off the four small moons. The faint rings of the other giant planets (see e.g. NEPTUNE: RINGS) of the solar system are probably associated with many small moons, and it is expected that future spacecraft will reveal similar systems at Saturn and the other giant planets.

Bibliography

Greenberg R and Brahic A (eds) 1984 *Planetary Rings* (University of Arizona Press)

Ockert-bell M E, Burns J A, Danbar I J, Thomas P C and Veverka J 1999 The structure of Jupiter's ring system as revealed by the Galileo imaging experiment *Icarus* **138** 188

François Poulet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jupiter: Satellites

As befitting the king of the planets, JUPITER is orbited by an entourage of at least 39 natural satellites in addition to its faint rings, intense radiation belts and occasional temporary visitors from Earth and the outer solar system. Named after Zeus' lovers and other mythological companions, Jupiter's moons can be divided into four groups on the basis of their sizes and orbits (table 1). Nearest the planet, four tiny inner moons lie embedded within Jupiter's tenuous rings and are the likely sources of the dust-sized particles that make up the ring system. Farther out are the Galilean satellites, four large worlds comparable in size with the Earth's Moon which vary systematically with distance from Jupiter in composition, appearance and geological activity. Two groups of captured satellites occupy the outskirts of the Jupiter system. The first four of these outer moons orbit in the direction of Jupiter's rotation, with orbits that are inclined at angles of 25°–29° to the planet's equator. The outermost four have highly inclined retrograde orbits, opposite to Jupiter's spin direction. In 2002, a team of astronomers discovered 11 new satellites by examining images taken in December 2001 and using later follow-up observations. It is estimated the new-found natural satellites are each about two to four kilometers (one to two miles) in diameter, and were probably passing rocks captured by Jupiter's gravity long ago. The satellites were formally announced by the International Astronomical Union on Circular No. 7900 (16 May 2002).

The Galilean satellites are large enough to be easily seen with inexpensive binoculars, devices that are far superior to the primitive instruments with which the moons were discovered in 1610. GALILEO GALOLEI is credited with first finding the four moons that bear his name and cited the existence of another gravitational center of movement in the universe as evidence to support the heretical hypothesis that the Earth and other planets revolved around the Sun. The Galilean satellites were also probably discovered independently by the German astronomer SIMON MARIUS, who named the moons on the basis of a suggestion by Johannes Kepler. Decades later, OLE RØMER of Denmark realized that the irregularities in the times of the eclipses of the Galilean satellites (used by mariners to determine longitude) were due to the changing distance between Jupiter and the Earth, and used this to make the first determination of the speed of light. Apart from Amalthea, first seen by E Barnard (the discoverer of BARNARD'S STAR, the second closest system to the Sun), the remaining satellites eluded detection until the 20th century. Because of their small sizes and proximity to Jupiter, the inner satellites Metis, Adrastea and Thebe went unseen until the VOYAGER explorations in 1979. Arriving four months apart on their way to the outer solar system, the two Voyager spacecraft also discovered Jupiter's rings and gave us our first detailed views of Amalthea and the surfaces of the Galilean satellites, including glimpses of the violent volcanic activity on Io.

Since December 1995, Jupiter has been orbited by an artificial satellite sent from Earth to study the planet and its rings, moons and magnetosphere. The GALILEO MISSION was launched in 1989 from the space shuttle Atlantis and took 6 yr to arrive at Jupiter following flybys of Venus, the Earth and two main-belt asteroids, Gaspra and Ida, the first such small bodies to be imaged by a spacecraft.

Table 1.

Satellite	*Orbital distance (1000 km)*	*Orbital eccentricity*	*Orbital inclination*	*Orbital period (days)*	*Radius (km)*	*Mass (kg)*	*Discoverer*	*Date*
Metis	128	~0	~0	0.295	30×20×17	9.56×10^{16}	Synnott	1979
Adrastea	129	~0	~0	0.298	10×8×7	1.91×10^{16}	Jewitt	1979
Amalthea	181	0.003	0.40	0.498	125×73×64	7.17×10^{18}	Barnard	1892
Thebe	222	0.015	0.80	0.674	58×49×42	7.77×10^{17}	Synnott	1979
Io	422	0.004	0.04	1.769	1818	8.94×10^{22}	Galileo	1610
Europa	671	0.009	0.47	3.551	1561	4.80×10^{22}	Galileo	1610
Ganymede	1070	0.002	0.21	7.155	2634	1.48×10^{23}	Galileo	1610
Callisto	1883	0.007	0.51	16.689	2408	1.08×10^{23}	Galileo	1610
Leda	11 094	0.148	26.07	238.72	5	5.68×10^{15}	Kowal	1974
Himalia	11 480	0.158	27.63	250.566	85	9.56×10^{18}	Perrine	1904
Lysithea	11 720	0.107	29.02	259.22	12	7.77×10^{16}	Nicholson	1938
Elara	11 737	0.207	24.77	259.652	40	7.77×10^{17}	Perrine	1905
Ananke	21 200	0.167	147	631 (R)	10	3.82×10^{16}	Nicholson	1951
Carme	22 600	0.207	164	692 (R)	15	9.56×10^{16}	Nicholson	1938
Pasiphae	23 500	0.378	145	735 (R)	18	1.91×10^{17}	Melotte	1908
Sinopë	23 700	0.275	153	758 (R)	14	7.77×10^{16}	Nicholson	1914

Values for the outer moons are approximate.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The mission was made up of two components: a Jupiter probe, which entered the atmosphere of the giant planet and relayed the first direct measurements of its composition, structure and dynamics, and an orbiter equipped with four remote sensing instruments sensitive to wavelengths from the ultraviolet to the thermal infrared along with a battery of instruments for monitoring the charged particles, waves and fields of Jupiter's space environment. Galileo has produced a number of new insights about Jupiter's satellites and rings, solving some of the mysteries arising from the Voyager data and presenting new puzzles. New observations of Jupiter's satellites have also been made during the Cassini spacecraft's flyby by Cassini, Galileo, the Hubble Space Telescope, the Chandra X-ray Observatory and Earth-based radio and optical telescopes. Cassini observed Jupiter from 1 October 2000 to 31 March 2001. Observational results include data on interactions of the Galilean moons with Jupiter's magnetic field. In the following sections we will summarize what has been learned so far from the space missions.

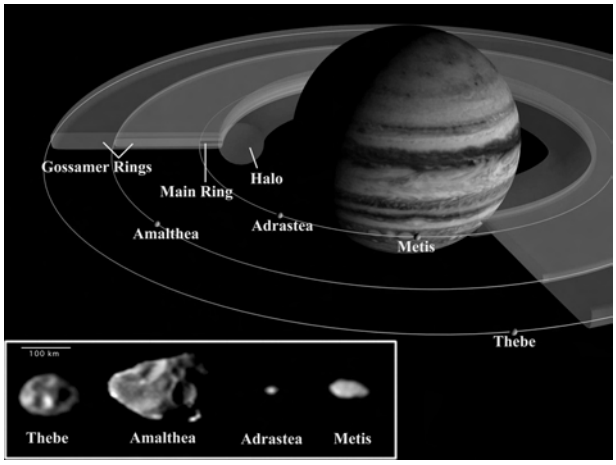


Figure 1. The inner satellites of Jupiter. (Courtesy of Cornell University and NASA.)

Inner satellites

With the exception of Amalthea, little was known about the small inner satellites of Jupiter prior to Galileo's arrival other than their orbits. The innermost pair, Metis and Adrastea, both orbit at the edge of Jupiter's brightest main ring at distances of 1.79 and 1.81 Jupiter radii (R_J) respectively from the center of the planet. Both lie within the Roche limit and would be disrupted by tidal forces if they had no internal strength. Their orbital periods are shorter (~7 h) than the period of Jupiter's rotation (~10 h); hence their orbits will gradually decay and they will ultimately fall into Jupiter. Unlike the other satellites, Metis and Adrastea orbit faster than Jupiter's plasma sheet, the dense swarm of trapped charged particles that corotates with the Jovian magnetic field. In contrast, Amalthea at $2.54R_J$ and Thebe at $3.11R_J$ have periods of about 12 and 16 h, respectively, and so they are overtaken by the faster corotating charged particles. All of the inner satellites are thought to be in synchronous rotation, having spin periods the same as their orbital periods so that they keep the same face pointed towards Jupiter at all times. Their densities and bulk compositions are unknown.

Distant imaging by Galileo allowed the shapes and sizes of the four inner satellites to be determined for the first time (see table 1 and figure 1) and also measured their colors and albedos. All are dark, red objects that become progressively redder with proximity to Jupiter. The coloration is similar to but darker than Io's, and could be due to contamination by sulfur compounds drifting inwards from Io. The orbital leading hemispheres of the inner satellites are 25–35% brighter than the trailing hemispheres, and their leading sides are also less red than their trailing sides. It was first thought that Amalthea's color dichotomy could be caused by charged particle implantation but, when the same hemispheric asymmetry was also found on Metis and Adrastea (where such implantation effects should be reversed), it was decided that the leading–trailing differences were due to impactors preferentially swept up on the moons' leading sides as they move in their orbits around Jupiter. The irregular shapes of the inner satellites suggest histories of bombardment and catastrophic disruption, and no geological features other than impact craters can be seen on their surfaces.

Jupiter's complex ring system is intimately related to the inner satellites (figure 1). Unlike the bright rings of Saturn, Jupiter's dark rings are probably made up of non-ice materials (see JUPITER: RINGS). The main, most prominent, ring extends from Metis and Adrastea inwards to $1.72R_J$ and is less than 30 km thick. Adrastea is located right at the edge of the main ring, while Metis orbits just within it. Reaching inwards from the main ring is the 'halo', a broad, diffuse cloud of charged dust particles levitated out of the equatorial plane by electromagnetic forces. Exterior to the main ring are two faint 'gossamer' rings trailing inwards from Amalthea and Thebe. The thickness of Thebe's gossamer ring is the same as the height of Thebe's orbital excursions above and below Jupiter's equatorial plane. A similar correspondence exists between Amalthea's orbit and the thickness of its gossamer ring. This confirms that the moons are the sources of the ring particles, probably debris kicked up by impacts. The ring particles are launched with orbital inclinations similar to their parent moons' and likewise oscillate above and below the equatorial plane. Because the particles spend less time on the equatorial plane than above or below, the gossamer rings have peculiar banded profiles, such that their margins appear brighter than the ring centers when viewed edge on. Metis and Adrastea appear to be the sources of the main ring. Interestingly, tiny Adrastea may contribute a greater share of ring particles than its larger companion. Adrastea is near the optimum size for shedding debris

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

from craters into space: smaller satellites present too little surface area for impactors to target, while larger satellites recapture more of their impact ejecta because of their greater gravity. The expected dynamical lifetime for ring particles is exceedingly short compared with the age of the Jupiter system, so the impact events that produced the currently observed particle population must have occurred relatively recently.

Galilean satellites

The Jupiter system is often likened to a miniature solar system because of the compositional gradation exhibited by its four largest moons (figure 2). Io (see below), the nearest to Jupiter of the four at $5.9R_J$, has a bulk density of 3.57 g cm^{-3}, corresponding to silicate rock and metal composition. EUROPA, at $9.4R_J$, has a bulk density of 2.97 g cm^{-3}, suggestive of dominantly silicate composition despite its icy exterior. GANYMEDE and CALLISTO have densities indicative of ice–rock mixtures with at least 40% ice by mass. At $15R_J$, Ganymede's density is 1.94 g cm^{-3} while Callisto, the farthest satellite ($26R_J$), has a density of only 1.86 g cm^{-3}. This compositional gradation is thought to be primordial, caused by the condensation of refractory materials in the warm, interior portions of the proto-Jovian nebula and of more volatile ices in the cooler outer reaches, in much the same way as the formation of planets in the solar system as a whole. Some of the compositional differences, particularly those seen on the surfaces of the satellites, may be due to geological activity as well. Ganymede, Europa and especially Io are heated by tidal flexing which powers their geological activity, along with modest contributions from the decay of radioisotopes and electromagnetic induction. The tidal flexing is driven by the eccentricities of their orbits, which are forced by the 1:2:4 Laplace resonance of their orbital periods: each time Ganymede completes one orbit of Jupiter, Europa orbits exactly twice, while Io undergoes four revolutions. Without this resonance to sustain it, the orbit of any of the satellites would soon circularize. The orbits of all three moons are gradually migrating outwards and will soon (i.e. in a few hundred million years) encounter a resonance with Callisto, possibly inducing activity on this geological late bloomer. Most of the energy dissipation in the satellite system takes place on Io, which as a consequence is the most active object in the solar system.

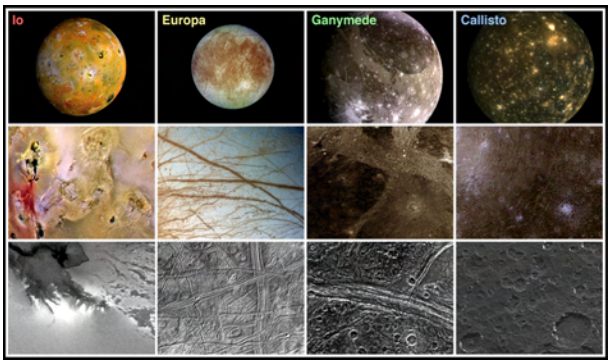


Figure 2. The four Galilean satellites. (Courtesy of DLR and NASA.) This figure is reproduced as Color Plate 27.

Io

Unique among solid objects in the solar system, Io's surface is devoid of recognizable impact craters. The amount of geological resurfacing necessary to obscure craters at their estimated present rate of formation is enormous—more than 1 cm of burial per year over the entire satellite. An even larger rate of resurfacing is required to account for the present heat flow observed from the moon (more than 2 W m^{-2}): assuming that this heat is carried by molten silicates at their liquidus temperatures, new materials must erupt at a rate of about 10 cm yr^{-1}, globally averaged. Comparisons between Io's appearance during the Voyager flybys and in Galileo pictures taken 20 yr later suggest that most of this resurfacing, over short time scales, is restricted to several dozen volcanic centers distributed more or less randomly around the globe. If so, the heat flow requires that fresh volcanic deposits tens of meters thick were emplaced in these active areas over the past two decades.

Volcanoes on Io have diverse morphologies often markedly different from their counterparts on Earth and Mars (see VOLCANISM AND GEOPHYSICS). They range from vast lava lakes such as Loki, the source of much of Io's thermal emission, to powerful plumes such as Pele that reach hundreds of kilometers in height and produce deposits large enough to be seen by the Hubble Space Telescope. Sulfur compounds such as SO_2 drive the plumes and paint Io's surface a surprising array of muted red, yellow and white hues. The lavas themselves are dark and generally lack the cones characteristic of terrestrial volcanoes. Galileo measured the thermal emission of the lavas with two different instruments during times when the moon was in Jupiter's shadow and found temperatures exceeding 1700 K, far hotter than the melting point of sulfur or even common silicate rocks such as basalt. Either the lavas are superheated prior to eruption or they are made up of materials with very high melting points. Magnesium-rich ultramafic magmas of the sort that commonly erupted on the early Earth are likely candidates. Some evidence consistent with this interpretation was found in the form of spectral absorption features at a wavelength of about 0.9 μm, which may indicate magnesium-rich silicate minerals such as orthopyroxene. The most persistent hot spots on Io seem to be located near the equator, consistent with models of tidal heating in the upper asthenosphere. Clusters of hot spots are found near the equator on both the Jupiter-facing (subjove) side and the side opposite Jupiter (antijove).

Io is a major polluter of the Jupiter system and the source of dust streams and plasma particles that diffuse throughout the Jovian environment. Sulfurous deposits from Io discolor Europa as well as the inner satellites. An

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

escaping cloud of neutral sodium atoms surrounds Io as it orbits Jupiter (along with even larger quantities of sulfur and oxygen, which are harder to see). Many of these atoms are subsequently ionized and swept up by the rotating Jovian magnetic field as it streams past Io. These electrons and ions make up a plasma torus that encircles Jupiter and wobbles up and down along with the Jovian magnetic field (which is inclined with respect to Jupiter's rotation axis). Energetic ions from the torus reimpact Io and eject even more particles, making the plasma torus self-sustaining (see IO: PLASMA TORUS). The motion of Jupiter's magnetic field past the conducting satellite also induces an electric potential across Io of some 400 000 V. Powered by this electrical generator, currents from Io modulate Jupiter's radio emissions and produce bright visible and ultraviolet glows in the polar regions of the distant planet, known as the flux-tube footprints. The footprints are electrically connected to Io along magnetic field lines running from Jupiter's high latitudes to the subjove and antijove points on Io's equator. Galileo imaged bright blue optical emissions from plume gases at these locations on Io while the satellite was eclipsed by Jupiter. These visible aurorae are probably produced by the flow of electrons through SO_2 vented by volcanoes. Other eclipse emissions at red and green wavelengths are not localized at the plumes and are likely to be due to oxygen and sodium in Io's tenuous atmosphere. In sunlight, a jet of neutral sodium can be seen escaping from Io in the direction away from Jupiter, apparently accelerated by the electric field and then neutralized by recombination.

Europa

Smallest of the Galilean moons, Europa was only distantly imaged by the Voyager spacecraft and remained the least explored of the four until the Galileo mission. Nevertheless, the satellite generated a great deal of interest because of the speculation that its icy exterior may harbor a subsurface sea of liquid water. Like Io, Europa is heated by friction as its shape is tidally deformed during each orbit of Jupiter—not enough to replicate Io's violent activity but perhaps enough to melt some of the ice and generate silicate volcanism on the sea floor. The Voyager investigators realized that its smooth, bright surface and paucity of recognizable impact craters indicated that the satellite was geologically young and ascribed its global system of lineaments (fractures, ridges and bands) to a thin ice shell overlying a ductile interior. On closer inspection by Galileo, Europa was revealed to have a variety of perplexing surface features that are unique in the solar system.

Two major terrain types dominate Europa's surface. Ridged plains are ubiquitous on the satellite and are made up of dense networks of crisscrossing lineaments with varying ages and orientations, testifying to a long history of tectonic activity. Concentrated near the equator but found sporadically at higher latitudes are the so-called chaos regions, thermally disrupted areas where the ridged plains have been erased by localized heating from below. The precise mechanisms by which these terrains formed are still hotly debated, and years of analysis and modeling will probably be needed before a consensus emerges. The most common landforms of the ridged plains are pairs of peculiar double ridges that often reach several hundred kilometers in length. They are typically 100–200 m high and less than 1 km across and are divided by a steep central valley. These ridge pairs appear to be part of a progression of tectonic features that begins with simple cracks in the ice surface. Icy material is deposited on the flanks of favorably oriented fractures by some cyclic process of extrusion and withdrawal, leaving ridge pairs with central valleys. As ridge pairs mature, their outer margins are darkened by as yet unidentified red-brown non-ice contaminants that discolor the surrounding surfaces to distances of several kilometers from the brighter medial ridges, forming striped 'triple bands'. Some ridges evolve into large, multiple-ridge sets composed of several individual ridge pairs, probably by repeated reactivation of crustal weaknesses associated with the ridges. Other lineaments have little observable relief and instead appear to be dilational bands caused by crustal spreading. Older surface features that are broken by these dilational bands can frequently be matched up through cutting and pasting the digital images, by removing the dilational bands and joining together the pre-existing crust like pieces of a puzzle. Intermediate landforms with characteristics of both ridges and bands are also observed. Many of the ridges and bands (but not fractures) are also strike-slip faults, commonly displaying left-lateral displacement in the northern hemisphere and right-lateral displacement in the southern hemisphere. A particularly puzzling class of lineaments are the cycloids or 'flexi', features resembling cracks, ridges and bands but having curved paths that abruptly change direction at ~100 km intervals, yielding arcuate patterns with semicircular sections joined by angular cusps.

Strong support for the notion of a subsurface ocean came from Galileo's first closeup pictures of chaos regions. One such site on Europa's trailing hemisphere, called Conamara, is an irregular depression about 100 km wide characterized by the thorough destruction of much of the original surface structure and replacement by an irregular jumble of lowlying ice with a red-brown stain similar to that of the triple-band margins. The remaining plate-like blocks of surviving crust have shifted, rotated and tilted from their original positions. Studies of other thermally disrupted areas suggest that such crustal blocks may gradually disaggregate in place. Two theories have been developed to explain these features. The first suggests that the chaos regions are the sites of crustal thinning and melting, leaving the crustal blocks adrift on near-surface liquid water analogous to ice floes in the polar regions of Earth. The second suggests that portions of the cold, dense surface subsided into warmer, low-density ice upwelling from below via

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

solid-state convection. Interestingly, many of the remaining crustal blocks in the Conamara chaos have large ridge sets and other heavy structures superimposed. Such heavy blocks should be among the first to founder if the surface were underlain by low-density ice; hence this observation is more consistent with the former explanation of the blocks as 'icebergs' floating on a higher-density liquid substrate.

The few impact craters scattered across Europa also suggest a thin ice crust. The largest (up to 50 km diameter) are flat, low structures surrounded by concentric fractures, tensional graben that may have formed as the ductile subsurface flowed to fill the newly created hole. The most recent large impact was Pwyll, a 26 km diameter crater on Europa's trailing side that still possesses a prominent set of bright rays. Because of the bombardment of the surface by charged particles from the plasma sheet, these crater rays are expected to be removed by sputtering erosion over time scales as short as a few million years. The continuous ejecta from Pwyll forms a dark-brown, 40 km diameter splotch similar in color to the chaos regions and the margins of the triple bands. The most fascinating aspect of Pwyll is its near total lack of relief: it is no deeper than 200 m below the level of the surrounding surface. For comparison, similarly sized impacts on Ganymede typically excavated enormous cavities 2 km deep. Either the Pwyll impact penetrated through the ice crust and the crater was flooded with liquid at the time it formed or the topography has subsequently slumped under its own weight by glacier-like flow. Europa's surface is extremely cold, however – less than 130 K on the equator at noon—and ice is as stiff as rock at such frigid temperatures. Even if the flatness of Pwyll is due to solid-state flow, the short time scale for topographic relaxation seems to require that the ice in the shallow subsurface is quite warm, perhaps because it is in close contact with liquid water below.

Theory is still struggling to keep pace with the flood of new data, but a picture of Europa as a world dominated by the tides and torques of Jupiter is emerging from the Galileo observations. Europa's tidal bulges wax and wane with an 85 h period owing to the changing distance between the moon and the planet, and also rock back and forth as they try to track Jupiter. As a result, the solid ice shell is subjected to stresses that change with both time and location as the shell stretches to accommodate Europa's changing figure. One consequence of this tidal flexing is that any cracks that penetrate through the shell to a liquid layer below will be pulled apart and slammed shut again each 85 h Europan 'day', providing a possible mechanism for building the satellite's unique ridge pairs. In some locations, the daily cycle starts as the fractures open and then the two sides begin to shear past one another in a transverse sense. Later in the day the stresses reverse, closing the cracks and favoring transverse motion in the opposite direction. This reverse shear is ineffective, however, since the cracks are under compression during this part of the cycle. The net effect is to ratchet the sides of the cracks past one another, forming strike-slip faults with mainly left-lateral displacements in the northern hemisphere and an opposite sense of motion in the south, just as observed. The most prominent example of this process on Europa may be Astypalaea Linea, an enormous strike-slip feature that rivals the San Andreas fault on Earth. The strange cycloidal lineaments are also thought to be due to the daily tides: a slowly opening crack should follow a curvilinear path across the surface as it tracks the changing tensional stress throughout the day, but the propagation ceases during the period when the local stress becomes compressional instead. Crack propagation resumes the next day when the stresses are again tensional, but by then the orientation of the local stress field has changed and the crack sets off in a new direction. The cusps of the cycloids, in this model, are the places where the crack propagation paused during the formation of these odd features.

Another consequence of Europa's eccentric orbit is that Jupiter exerts a torque on the satellite that makes Europa (or at least its outer shell) spin at a rate slightly faster than it orbits. As a result, the surface of the moon gradually shifts eastwards with respect to Jupiter over geological time. Europa's non-synchronous rotation is too slow to be measured directly, but it has left its mark on the landscape. Galileo has explored several locations where successive generations of lineaments overprint their predecessors at oblique angles, indicating systematic rotation of the local stress fields. The sense of rotation is clockwise in the northern hemisphere and anticlockwise in the south, consistent with the predictions of theory. The ice shell should stretch as it shifts over the tidal bulges, and indeed a zone of crustal extension is seen to the west of the anti-Jupiter point. (The Jupiter-facing side of Europa, where such extension is also predicted to occur, has not yet been seen as of this writing.) The model also explains why some fractures may be dormant for extended periods and reactivated later when relocated to more favorable longitudes.

These theoretical arguments, if correct, imply the existence of a subsurface liquid layer in order to raise tides sufficient to fracture the ice shell and to decouple the crust from the rocky interior and allow it to rotate freely. Together with the evidence of localized heating and the suspected infall of organic materials from such sources as comets, the possibility of an ocean on Europa could have profound significance to the search for extraterrestrial life. Recognizing this, NASA is developing a mission specifically designed to determine whether or not a subsurface sea exists on Europa. The Europa Orbiter spacecraft will directly detect any extensive ocean by taking the 'pulse' of the satellite (measuring its tidal amplitude using gravimetry and altimetry) and sounding the subsurface with radar and will also produce a global map of this intriguing moon and search for suitable landing sites and signs of present-day activity.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Ganymede

Ganymede is the largest Galilean satellite and indeed the largest moon in the solar system. It is slightly bigger than Mercury, and would be considered a planet in its own right if not for the fact that it is in orbit about a larger world. Its interior is strongly differentiated into an iron-rich core, a silicate mantle and an icy crust much thicker than Europa's. Close flybys of Ganymede by the Galileo spacecraft revealed an intrinsic magnetic field with a strength about 1% of Earth's, suggesting that the satellite's iron core is still molten. Hence we suspect that core formation may have taken place relatively recently, perhaps triggered by tidal heating within the last billion years or so. Such tidal heating is negligible today, but Ganymede may have passed through periods of intense heating as its orbit evolved prior to being captured into its present stable resonance with Europa and Io. During differentiation, the H_2O and CO_2 ices buried deep within Ganymede floated up towards the surface, displaced by the heavier silicates and metal sinking into the interior. Water ice should have expanded as it decompressed and converted from higher-density phases into its familiar hexagonal form. The volume of the satellite would have increased as a consequence and the resulting global expansion probably split Ganymede's surface, replacing some of the ancient crust with lanes of younger material. This tumultuous history seems to be reflected in the geology of the satellite's surface.

Ganymede's crust is divided into two major terrain types. Older areas such as Galileo Regio are dark and heavily cratered and may represent the moon's primordial surface. Few tectonic features are found on these ancient terrains other than vague furrows probably formed by giant impacts. Craters on Ganymede display a remarkable diversity ranging from simple bowl shapes similar to their lunar counterparts, through flat-floored basins with curious central pits and domes, to ancient, flat circular structures called palimpsests that are scarcely recognizable as impact features and evidently formed while the moon's interior was still warm and ductile. A few scalloped depressions in the ancient terrain may indicate early outpourings of liquids from the interior, the watery equivalent of volcanic magmas. Ganymede's younger regions, in contrast, are brighter and show signs of intense tectonic deformation. These younger areas consist of narrow lanes of 'grooved terrain' tens of km wide and hundreds of km in length. Once suspected to be of volcanic origin, the grooved terrains were seen in closeup Galileo pictures to consist of closely spaced domino-style faults and graben indicative of crustal extension. At least two scales of tectonic deformation can be recognized in these regions: fine fractures spaced a few hundred meters apart, indicating brittle failure of the cold surface, are superimposed on broad undulations with wavelengths of several km, probably produced by ductile necking of the crust.

Ganymede's high latitudes are decorated by bright frosty polar caps. These frosts are concentrated on the slopes that face away from the Sun and are probably derived from ice that sublimated away from the warmer equatorial areas. The polar regions may also be brightened by charged particles that spiral inwards along magnetic field lines and stir up the surface on impact. Charged particles are probably responsible for Ganymede's faint oxygen aurorae, observed near the poles at ultraviolet wavelengths. Ganymede's tenuous atmosphere is made up mostly of hydrogen atoms flowing rapidly away from the moon, amounting to a total loss of millions of kg per year. The hydrogen gas is probably derived from the dissociation of water ice on the surface of the satellite by charged particles and solar ultraviolet photons. The light H escapes, while the heavier O is left behind to form strong oxidants such as molecular oxygen, ozone and hydrogen peroxide. Ganymede's older regions are tinged by a dark non-ice contaminant that is distinct in color from Europa's enigmatic red-brown stains but similar in appearance to Callisto.

Callisto

Only one of the Galilean satellites retains a surface unmodified by internal geological activity. Callisto appears to have passively recorded a history of bombardment by impactors. Unlike Io, Europa and Ganymede, Callisto is not trapped in an orbital resonance with its neighbors that forces tidal heating. Galileo gravity measurements indicate that Callisto's interior is only slightly differentiated, largely unchanged from its initial accretionary mixture of rock and ice. Nevertheless, Galileo's magnetometer recorded an induced magnetic signature from the satellite indicating the presence of an electrically conducting layer not far below the surface, such as subterranean brines. Callisto's dark, battered surface is similar to the older terrains on Ganymede. Several of the latest large impacts took place on Callisto's orbital leading side. The greatest was Valhalla, which produced a palimpsest 600 km in diameter and a set of rings that nearly encircle an entire hemisphere. More recent impacts excavated bright, relatively clean ice from beneath the surface. Paradoxically, at subkilometer scales there are far fewer craters on Callisto than Ganymede. High-resolution pictures of Callisto show a smooth blanket of dark material instead of the pock-marked landscape expected from more distant vistas. Sublimation of CO_2 and H_2O ices may be responsible for the dearth of small craters on Callisto. In many places the landscape appears to have been degraded by sublimation, leaving behind a lag of dark non-ice material that fills the topographic lows. The icy rims of several small craters have eroded into isolated knobs, while pits have been produced in other places by the removal of ice. Some initially circular craters have been deformed into irregular depressions, and several landslides of dark lag material extend from

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

the less stable slopes of the eroding crater rims. The dark mantling material may have an external origin, perhaps deposited by impactors or debris that drifted inwards from the outer satellites.

Outer satellites

Little is known about the eight small moons that orbit beyond Callisto. At present even their sizes can be only roughly estimated, pending knowledge of their reflectivities. Their orbits fall into two groups, both at odds with Jupiter's rotation (figure 3). The innermost family (Leda, Himalia, Lysithea and Elara; names ending with the letter 'a') have dynamically young, prograde orbits that are steeply inclined with respect to Jupiter's equatorial plane (table 1). The outer family (Ananke, Carme, Pasiphae and Sinopë; ending with 'e') orbit in the retrograde direction, opposite to Jupiter's rotation, and are so far away from Jupiter that their paths are strongly perturbed by the gravity of the Sun. All of these satellites are dark, reddish objects that could conceivably contribute to the coloration of Callisto and Ganymede, dusting the icy moons with debris derived from impacts. No spectral features diagnostic of their composition have yet been recognized on these satellites. The outer family satellites are somewhat redder than the inner group, similar to the colors of the TROJAN ASTEROIDS. The distant Trojans are also gravitationally trapped by Jupiter, in stable points ahead of and behind Jupiter as it orbits about the Sun. The ranges of rotation periods and shapes (lightcurve amplitudes) of the satellites are similar to those of Trojan and main belt asteroids, suggesting comparable collisional histories.

These two satellite families may be the fragments of larger parent bodies that shattered when they were gravitationally captured by Jupiter. A vivid example of one such disruption was displayed recently by comet Shoemaker–Levy 9 when it ventured too near Jupiter and was torn to pieces by the tides (see SHOEMAKER–LEVY 9–JUPITER COLLISION). Much more violent forces are expected in the case of a satellite captured by gas drag after colliding with Jupiter's atmosphere or protoplanetary nebula. Such forces are capable of disrupting much stronger and larger objects like the parent bodies of the outer satellites. Perhaps some of the Pasiphae family fragments leaked out of the Jupiter system during the upheaval: the Trojans could be rogue satellites that escaped from Jupiter and were later recaptured. Evidence of past episodes of capture and catastrophic disruption can be seen on the surface of Callisto, which bears the scars of several impacts by Shoemaker–Levy 9-like objects as linear chains of impact craters found on the Jupiter-facing hemisphere.

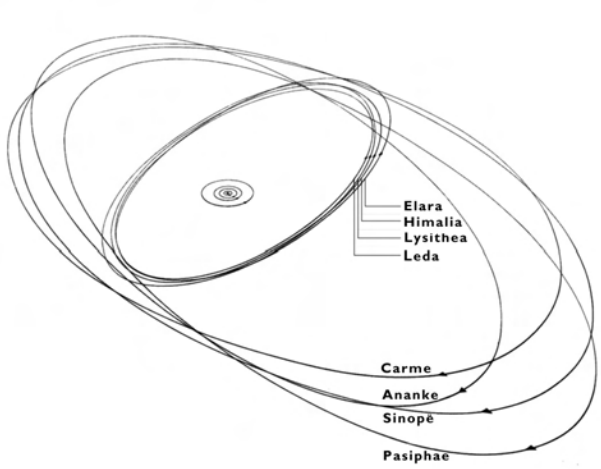


Figure 3. The outer satellites of Jupiter. (Courtesy of G Hunt and P Moore 1981 *Jupiter* (London: Mitchell Beazley).)

Web update (30 April 2003)

Recent results from the Galileo spacecraft reveal Jupiter's inner moon, Amalthea, has a very low density. Amalthea is apparently a loosely packed pile of rubble. The empty gaps between solid chunks likely take up more of the moon's total volume than the solid pieces, and even the chunks are probably material that is not dense enough to fit some theories about the origin of Jupiter's moons. Amalthea now seems more likely to be mostly rock with maybe a little ice, rather than a denser mix of rock and iron.

Bibliography

For a summary of Voyager results, see

Morrison D ed 1982 *Satellites of Jupiter* (Tucson, AZ: University of Arizona Press)

A pre-Galileo update is in

Burns J A and Mathews M S ed 1986 *Satellites* (Tucson, AZ: University of Arizona Press)

Initial analyses of the Galileo mission results are reported in

1998 *Icarus* **135** (1)

The relationship between the inner satellites and the rings is discussed by

Ockert-Bell M *et al.* 1999 *Icarus* **138** 188–213

Recent observations of the outer satellites are reported by

Luu J 1991 *Astronomical Journal* **102** 1213–25

Paul Geissler

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Kaiser, Frederik [Frederick] (1808–72)

Dutch astronomer, director of Leiden Observatory, observer of double stars, the Milky Way and Mars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kant, Immanuel (1724–1804)

German philosopher, born in Königsberg, published his view of the universe in *General History of Nature and Theory of the Heavens* (1755) in which he presented his nebular hypothesis of the formation of the solar system. It was much like the present theory that the Sun and planets formed from the condensation of a rotating disc of interstellar material. Kant identified the Milky Way as a lens-shaped collection of stars in orbit around its center like the rings of Saturn, and as one of many 'island universes' (he coined the term). He suggested that the tides raised by the Moon were the reason why the Moon always presented the same face towards us, and that tidal friction was slowing the rotation of the Earth. Kant's philosophy is described in *Critique of Pure Reason*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kanzelhöhe Solar Observatory

The Kanzelhöhe Solar Observatory was founded during World War II and belongs to the Institute of Astronomy of the University of Graz. The site is located at 1400 m altitude near Villach in Austria and is very well suited for long-term studies of the Sun. Therefore it was selected as the only ground-based support observatory in Central Europe for satellite missions (SOHO). With a newly installed magneto-optical filter, high time cadences of solar magnetograms can be obtained. Sunspot numbers and flare patrols are submitted to the scientific community.

For further information see
http://www.solobskh.ac.at/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kapteyn Astronomical Institute

The Kapteyn Astronomical Institute, University of Groningen, the Netherlands, specializes in the areas of structure of galaxies and the universe and interactions between stars and the interstellar medium, using many national and international observing facilities all over the electromagnetic spectrum. Specific areas of research include the distribution and nature of dark matter in galaxies, gravitational lenses, molecular astrophysics and the large-scale structure of the universe.

The Kapteyn Institute is located at the university campus (Zernike complex) on the outskirts of the city of Groningen in the northern part of the Netherlands. It was founded at the end of the eighteenth century by Jacobus Kapteyn as an 'astronomical laboratory', where plates taken elsewhere were measured and analysed. Kapteyn is the founder of modern research on galactic structure through star counts and the father of the famous 'Plan of Selected Areas'.

The Institute currently has about fifteen permanent scientific staff members and eight support personnel and on average about five postdoctoral fellows and twenty graduate students working on their PhD theses. In addition there are about fifty undergraduate students in astronomy. Funding comes mostly from the university with additional positions for temporary appointments (postdoctorates and graduate students) from the Netherlands Organization for Scientific Research NWO (the Research Council) and international funds such as from the European Commission.

The Kapteyn Institute is part of the Netherlands Graduate School on Astronomy (NOVA) together with the three other major astronomy institutes at the universities of Amsterdam, Leiden and Utrecht. The Kapteyn Institute shares its building with part of the Netherlands Foundation for Space Research (SRON), which is funded by NWO and specializes in the building of instrumentation for astrophysical space facilities.

Research at the Institute covers a broad range of topics, ranging from molecular astrophysics of the interstellar medium, late stages of stellar evolution, structure and dynamics of galaxies to large-scale structure of the universe. A major highlight of the work at the Institute concerns the distribution and nature of dark matter in galaxies, first through the observations of the rotation in spiral galaxies using the 12 cm spectral line of neutral hydrogen on the Westerbork Synthesis Radio Telescope and currently, at an increasing level, using various forms of gravitational lensing.

Research facilities are all on the basis of national and international levels. On the Earth these include the WESTERBORK SYNTHESIS RADIO TELESCOPE and the associated European Network for Very Long Baseline Interferometry, the (UK–NL–Canada) JAMES CLERK MAXWELL TELESCOPE at millimeter wavelengths, optical and near-infrared facilties at the UK–NL ISAAC NEWTON GROUP OF TELESCOPES at La Palma, Canary Islands and the EUROPEAN SOUTHERN OBSERVATORY in Chile. Through SRON and ESA, access is provided by a whole range of space facilities, of which the Hubble Space Telesope and the ESO Infrared Space Observatory (ISO) mission are currently the most important.

For further information see http://www.astro.rug.nl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kapteyn, Jacobus Cornelius (1851–1922)

Dutch astronomer, who, at the University of Groningen, and with the aid of convicts from a prison nearby, measured photographic plates taken by DAVID GILL at the Cape of Good Hope. He compiled a catalog of half a million southern stars, from which he derived values for the density of stars in space as a function of distance, brightness and spectral class, extending Herschel's work (see HERSCHEL FAMILY). In the *Kapteyn model* of the Galaxy, he confirmed that the Galaxy was lens-shaped. However, his conclusion that the Sun was located near the center was an error caused by interstellar absorption which limited the distance that he could sample, in the same way that someone in a fog could believe there was nothing outside the circle that he could survey. Kapteyn went on to identify Selected Areas, in which he proposed that astronomers worldwide sample the faint stars and thus map the Galaxy. This project was undermined by the First World War, and was perhaps too big to be completed even in peacetime, but stimulated many individual discoveries. He measured the velocities of stars and discovered 'star streaming', which proved to be a manifestation of the rotation of the Galaxy. He discovered *Kapteyn's star*, which has the second largest proper motion after BARNARD's star. He successfully explained the expanding halo around Nova Persei, discovered by CHARLES DILLON PERRINE as a light echo, reflecting the nova flash from surrounding dust clouds. The *Kapteyn telescope* on La Palma is named after him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kapteyn's Star

A faint red dwarf star (apparent magnitude 8.86, spectral type M0V), situated in the southern constellation of Pictor, about 12″ north-west of Canopus. Discovered by the Dutch astronomer Jacobus Kapteyn (1851–1922) in 1897, it is notable for having the second largest proper motion known: its motion of 8.67″ per annum is exceeded only by that of Barnard's Star. It is also one of the closest stars to the Sun: 12.8 light-years distant, it has a parallax of 0.255″ and its absolute magnitude is 10.9.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Karin Cluster

Group of some 40 known asteroids (including 832 Karin) which formed from the collisional breakup of a 25 km body some 6 million years ago.

Bibliography

Nesvorny D, Bottke W F, Dones L and Levison H F 2002 *Nature* **417** 720

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Keeler, James Edward (1857–1900)

Astronomer, born in La Salle, IL. In 1881 went with a pack train of mules, loaded with mirrors, telescopes and bolometers, to the 14 495 foot summit of Mount Whitney, California (the highest elevation in the contiguous United States), to measure, undisturbed by the atmosphere, the infrared spectrum of the Sun and its flux, the 'solar constant'. Became director of the Allegheny and Lick observatories, established that Saturn's rings were not solid by showing that they rotated with motions which conformed to KEPLER's law (as JAMES CLERK MAXWELL had predicted), implying that they consisted of meteoric particles in individual orbits. Founded the *Astrophysical Journal* in 1895 with GEORGE E HALE, in the first volume of which Keeler's paper on Saturn's rings appeared the month after he had taken the spectra!

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kepler

Johannes Kepler (1571–1630) was arguably the most innovative astronomical theorist in the millennium and a half from Claudius PTOLEMY's *Almagest* (c. AD 150) to Isaac NEWTON's *Principia* (1687). Before Kepler, planetary and lunar theory had consisted in combining circular motions, either strictly uniform or angularly uniform about an off-center 'equant' point, so as to 'save the appearances'. The planets were taken to be perfect beings for whom the only thinkable motion was circular. Kepler changed all this by envisaging planets and satellites as bodies moving in orbits under the action of forces issuing from the central body. In the detailed working out of this dynamical conception, he arrived, with the crucial help of TYCHO BRAHE's astronomical observations, and inspiration from WILLIAM GILBERT's magnetic philosophy, at the rules for which he is famous: elliptical paths, each radius vector from Sun to planet sweeping out equal areas in equal times, with the periodic times of the several planets proportional to the 3/2 power of their mean solar distances.

Since the eighteenth century these rules have been called 'laws'. Were they empirical laws, verified with a specifiable precision? The truth is more complicated and more interesting. Kepler's revolution can rightly be understood only in its historical context, with careful attention to his processes.

Copernicanism

From 1589 to 1594 Kepler attended the University of Tübingen, aiming to become a Lutheran pastor. An eager student, he ranged widely through the approved learning of his time. In theology he found the Lutheran doctrines of the bondage of the will and the Real Presence in the Eucharist indigestible, but kept his doubts to himself. He was attracted to Neoplatonic symbolism and to the writings of NICHOLAS OF CUSA. Through the teaching of MICHAEL MÄSTLIN, professor of mathematics and astronomy, he became enamored of Copernicus's heliocentric astronomy, though without access as yet to Copernicus's *The Revolutions of the Heavenly Spheres* (1543). The Tübingen theological faculty had prohibited the teaching of heliocentric astronomy as contrary to Holy Writ, but Mästlin privately helped a few exceptional students like Kepler to understand the kinematic economy of the heliocentric arrangement.

Copernicus's *Revolutions* had been published with a preface asserting the hypotheses of the work to be merely devices for accommodating the phenomena. This 'positivist' preface, not in fact by NICOLAUS COPERNICUS but by ANDREAS OSIANDER who had guided the book through the press, expressed a widely accepted view of astronomical hypotheses. The COPERNICAN SYSTEM contradicted both the literal sense of Scripture and the accepted theory of projectile motion. The kinematic economy of this system could be achieved otherwise, by the 'geoheliocentric' arrangement, in which the Earth remained at rest with the Moon and Sun circling about it, and the remaining five planets circled about the Sun. Tycho Brahe was the leading proponent of this alternative system.

Meanwhile, Copernicus got high marks for something separate from his heliocentrism: the replacement of Ptolemy's equant point by an epicyclic mechanism. The equant, according to Copernicus, by making motion on the circle non-uniform, violated a first principle of the astronomer's art. His insistence on strictly uniform circular motions in astronomical theory was accepted by many late sixteenth-century astronomers, including Tycho.

Kepler in his student days had not yet faced the difficulties of a detailed fitting of astronomical theory to observational data. What attracted him so powerfully to heliocentrism was the theological symbolism he saw in it. Nicholas of Cusa had portrayed the circle as an image of the Trinity: God the Father at the center, the Son at the periphery, the Holy Spirit in the relationship of center to periphery. For Kepler this Trinitarian image expanded to include the whole spherical world, with the Sun at the center and the stars at the periphery reflecting back the solar light. God in creating the world had created an image of Himself, and had placed the contemplative creature Man (also created in His image) on a moving planet, that he might, by calculating for his own displacement, discern the beautiful layout of the world. As Kepler later put it, travel is broadening.

The cosmographic mystery

Early in 1594 the position of mathematics teacher at a Protestant boys' school in Graz, Austria, fell vacant, and the senate of the University of Tübingen, asked to recommend a successor, chose the mathematically talented Kepler. This summons perturbed Kepler; he accepted it only with the proviso that he might later return to complete his theological studies.

While teaching at Graz, Kepler made an astronomical discovery that brought renown and confirmed him in an astronomical career. He had been persistently seeking, he tells us, the causes of three things: the number of the planets, their distances from the Sun, and the periods of their motions. Euclid in his *Elements* demonstrates that there are just five regular convex polyhedra: cube, tetrahedron, dodecahedron, icosahedron and octahedron. Let spheres be circumscribed about and inscribed within each of these, and let these spheres be nested so that the in-sphere of one polyhedron is the circum-sphere of the next; then there will be six spheres in all, just as many as there are planets in the Copernican system. Moreover, if the polyhedra are taken in the order just given (and Kepler provided an *a priori* explanation for this order), then the radii of the spheres prove to be in astonishingly close agreement with the radii of the planetary orbits. Uncertainty remained as to the exact values of the orbital eccentricities, and therefore as to the thickness to be assigned to each sphere. Nevertheless, Kepler's discovery of an *a priori* derivation of the number and distances of the planets impressed his teacher Mästlin, who helped him in preparing an account of it: the *Cosmographic Mystery*,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

published in March 1597. Kepler sent a copy to Tycho Brahe.

The polyhedral hypothesis was a formal or aesthetic cause, plausible to Kepler because he believed that God in creating the world had made it beautiful for the delectation of the contemplative creature, Man. In the *Cosmographic Mystery* Kepler also considered the *efficient cause* of the planetary motions. He hypothesized it to be a moving soul (*anima motrix*) situated in the Sun, the efficacy of its causal action diminishing with distance. In confirmation whereof he showed that average linear speed was less for planets farther from the Sun, and the linear speed of each individual planet diminished as its solar distance increased. Kepler did not have our principle of inertia; he was assuming that speed was proportional to force, and that an immaterial virtue, issuing from the Sun and rotating with it, pushed the planets about.

Meanwhile the Counter Reformation came to Inner Austria. In September 1598 the Protestant school in Graz was closed down and most Protestants were exiled. Kepler remained as district mathematician until August, 1600, when he too was exiled for refusing to convert to Catholicism. Earlier in 1600 he had visited Tycho in Bohemia and agreed to assist him in preparing his works for publication. In October 1600 he arrived in Prague with his family, his future now totally dependent upon Tycho's goodwill and sponsorship. In September 1601 he was granted a stipend by the Emperor Rudolph II to collaborate with Tycho in assembling new planetary tables. On 24 October Tycho died, and two days later Kepler was appointed imperial mathematician in Tycho's place. The 10 succeeding years in Prague would be among the most fruitful of his life.

The Optical Part of Astronomy and *The New Astronomy*

Kepler's *The Optical Part of Astronomy* (1604) contained major new results for both optics and astronomy: the geometrical explanation of the formation of images through small apertures, as in the *camera obscura* (used in the observation of solar eclipses); the most accurate table of astronomical refractions yet to appear, based on a law of refraction closely approximating the sine law; the first enunciation of the inverse-square law for the intensity of illumination from a point source; the discovery of the formation of inverted images on the retina of the eye. In planetary theory the revolutionary work was Kepler's *New Astronomy* (1609), a study of the motions of Mars.

This book has the form of a narrative tracing Kepler's journey in arriving at the elliptical orbit and area rule, but it is not strict history; Kepler used the narrative form in an effort to disarm hostile critics and persuade astronomers discomfited by the new ideas. Throughout he was guided by the conception of the Sun's dynamical role as articulated in the *Cosmographic Mystery*. From this conception came his hypothesis that the orbit of Mars lay in a plane passing through the Sun's center, with a constant inclination to the ecliptic; the complicated theories of the latitudes in earlier astronomy were thereby replaced by a simple, accurate predictive scheme. All earlier astronomers had reckoned the orbital eccentricities of the planets from the midpoint of the Earth's orbit (the 'mean Sun'), or in geocentric terms from the Earth's center; Kepler wanted the eccentricities to be reckoned from the true Sun, and showed that observational data were not opposed. His further study of Mars's motion in longitude yielded yet more innovative results.

To know Mars's motion about the Sun it was necessary to start from observations of Mars in opposition to the Sun, when the planet's longitude as seen from the Earth is also its heliocentric longitude. Kepler had twelve such observations to work with, ten made by Tycho, two by himself. He assumed a circular path for the planet and an equant point, then by trial and error so determined the solar distances of the equant point and circle's center as to obtain the best fit to four of the observed longitudes. With the circle's radius as unit, the solar distance of the circle's center proved to be 0.11332, and that of the equant 0.18564. Then he compared the theory's predictions with all twelve observations. The largest deviation was $2'12''$; the average deviation, $51''$. The theory's predictions were thus as accurate as the observations, but as Kepler next showed, it was false. He called it his 'Vicarious Theory': not the true theory, but accurately predicting Mars' heliocentric longitudes.

The theory was disproved by two sorts of observation of Mars near perihelion and near aphelion: Mars' latitudes when in opposition, and its longitudes when Mars, Earth and Sun were no longer aligned. From the triangles involved he calculated Mars's solar distances at aphelion and perihelion in terms of the Earth–Sun distance as unit. The midpoint of the line from perihelion to aphelion proved to be just halfway between the Sun and equant point, not 0.61 of the way as in the Vicarious Theory. If he altered the Vicarious Theory to agree with this 'bisection of the eccentricity', the altered theory erred by about $8'$ in the 'octants' (at 45° from aphelion or perihelion).

No equant theory, he concluded, and hence no Copernican epicyclic replacement for an equant theory, could yield true predictions where the orbital eccentricity was large. Therefore, Kepler argued, it was justifiable to resort to physics. Bisection of the eccentricity implies an inverse proportionality of speed to solar distance in the apsidal line. Was it not then the Sun's virtue that moved the planets in their paths, with a strength that fell off with distance?

Kepler actually spoke not of speeds but of 'delays' (*morae*). Assuming these to be proportional to solar distance, he undertook to allot times to each degree in the semicircle from aphelion to perihelion so that the times would add up to a half period. The area rule first emerged as a short cut: all the solar distances are contained in the areas swept out, so let the times be as these areas. Years later (in his *Epitome of Copernican Astronomy*, 1621) he would show that this rule meant an inverse proportionality of distance with the *component* of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the planet's speed at right angles to the radius vector. In the *New Astronomy* his justification of the rule remained confused.

Applying the area rule to the circular orbit, Kepler found it put the planet 8′ ahead at 45° from aphelion, and 8′ behind at 135° from aphelion: if area meant time, there was too much area in the middle longitudes. By triangulation Kepler verified that, indeed, the orbit came within the circle in the middle longitudes; it was oval. Concocting an epicyclic theory that yielded an oval orbit, he found the opposite kind of error: −8′ at 45°, +8′ at 135°. An elliptical orbit halving the difference would be just right. But why should the planet move in such an orbit?

An explanation had to be found for the planet's changing its solar distance. Kepler's ultimate hypothesis was a quasi-magnetic mechanism in which the Sun's surface was one pole of a magnet, and the two poles of the planet caused it to be alternately attracted and repelled, so as to librate sinusoidally along the radius vector. At first he imagined the orbit resulting from the libration would be 'puffy-cheeked', differing from the ellipse of the previous paragraph. In chapter 58 he claimed to refute the puffy-cheeked orbit from observation, and then reported that the libration and the ellipse could be made to agree, concluding that the orbit could only be a perfect ellipse. The latter statement has often been repeated by commentators. But Kepler's empirical refutation of the puffy-cheeked orbit is obscure and cannot be right. His claim to have excluded all alternatives to the ellipse cannot stand.

What was important for astronomy's future, however, was that Kepler's combination of ellipse and area rule led to the *Rudolphine Tables* (1627), planetary and lunar tables more accurate than any produced before. Kepler's *ephemerides* (daily positions), published for the years from 1617 to 1637, helped win converts to his tables. Not that Kepler's hypothetical mechanisms, or his area rule, gained a following: astronomers publishing tables in mid-century—chiefly Ismael Boulliau, Vincent Wing and Thomas Street—used substitutes for the area rule that were computationally more convenient, while recognizing that they must nearly duplicate Kepler's numbers. The elliptical orbit came to be widely accepted.

Kepler's harmonic law

Kepler discovered his 'third' or 'harmonic law' in May 1618, and announced it, without supplying empirical evidence, in his *Harmonic of the World* (1619). This latter work is Kepler's final account of the geometrical harmonies adorning God's creation. In Book IV of the *Epitome of Copernican Astronomy* (1621) he derived the harmonic law from efficient causes, assuming his pre-inertial dynamics plus monotonic variations of the volumes and densities of the planets with solar distance. By the 1630s, empirical evidence from improved telescopes had shown the monotonic variation of volume with distance to be untenable. In Kepler's *Rudolphine Tables* the planetary distances were not in strict accordance with the harmonic law; the first to make them so was JEREMIAH HORROCKS (1618?–41), and Thomas Streete followed this Horrocksian practice in his *Astronomia Carolina* (1661), from which Isaac Newton learned of it.

Kepler's other harmonic speculations were of no influence in later astronomy, yet have a certain kinship with supersymmetry and the anthropic principle in recent physics and cosmology.

The sorrows and difficulties of Kepler's last 18 years—deaths of a wife and children, money problems, excommunication from the Lutheran Church, trial of his mother for witchcraft, etc—can only increase our wonder in his remarkable journey of discovery.

Bibliography

Caspar M 1959 *Kepler* transl. C D Hellman (London: Abelard-Schuman)

—— 1990 *Kepler* revised edn transl. C D Hellman and O Gingerich (New York: Dover)

Stephenson B 1987 *Kepler's Physical Astronomy* (Berlin: Springer)

Curtis Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kepler's Laws

Johannes KEPLER published his three laws in 1609 and 1618 (table 1). Based on the very precise observations of Mars' orbit made by Tycho Brahe, and developed in the Copernican system, they describe the shape of the orbit of the planets around the Sun as an ellipse (first law), explain how the planets move around the Sun (second law) and give a relationship between the extent of the orbit and its period (third law).

Historically, these laws represented a strong and original improvement compared with the complicated system of epicycles derived by PTOLEMY. They permitted the calculation of much more precise astronomical tables. Furthermore, they introduced a revolutionary concept: a physical phenomenon can be simply expressed by a mathematical equation. Johannes Kepler stated his laws, which did not explain the physical cause of the trajectories and movements of the planets, before GALILEO GALILEI's and Isaac NEWTON's work. Actually, Kepler's laws can be inferred from Newton's laws. They apply in fact to any system including two objects in close gravitational interaction so that other gravitational fields can be neglected. That is the case for the solar system, dominated by a central massive object surrounded by smaller ones. When transposed to binary stars, to extrasolar systems or to any pairs of objects obeying the two-body interaction, the third law allows us to determine the total mass of the system. This point has therefore crucial consequences: the third law is a universal scale that makes it possible to remotely measure the mass of objects in the universe.

The physical content of Kepler's laws is explained first. For sake of simplicity, the solar system, to which Kepler applied his laws, is chosen as the standard case. Other systems are also considered, when the third law is examined. Kepler's life and ideas are then jointly presented, in order to demonstrate the major steps of his work. The revolutionary ideas he introduced are compared with the previously existing astronomical systems. Finally, the physical content of the laws is examined according to the concept of integral quantities, and their limitation is presented.

Table 1. Definition.

- First law: the path of each planet around the Sun is an ellipse, with the Sun at one of the foci
- Second law: each planet moves in such a way along its orbit that a line drawn from the Sun to the planet sweeps equal area in equal amount of time
- Third law: the ratio of the cube of the semimajor axis of orbit to the square of the orbital period is the same for all planets

Kepler's three laws

Kepler's first law

- The path of each planet around the Sun is an ellipse, with the Sun at one of the foci.

Kepler's first law, purely geometric, describes the shape of the orbit as a very simple figure: the ellipse (table 2). An ellipse being a plane figure, this already constitutes a remarkable result. Thanks to TYCHO BRAHE's observations, Kepler proved that planets moving in three-dimensional space follow a path included in a fixed plane. Astronomers before him only supposed this fact. Furthermore, Kepler was able to exhibit the exact shape of the trajectory as well as the role of the Sun. The Sun occupies one of the foci of the ellipse, the other focus and the center of the ellipse playing absolutely no part. According to the predominating position of the Sun, this first law confirmed the relevance of the COPERNICAN SYSTEM.

The elliptical shape summarized the complicated and *ad hoc* accumulation of epicycles invented by the Greeks: in the trajectory plane, three orbital elements are enough to define an ellipse. The semimajor axis a and the eccentricity e give the extent and the oblateness of the trajectory (figure 1, figure 2, table 3); the direction of the major axis of the ellipse is determined by a further constant.

Kepler's first law, when defining the trajectory of a planet, remains, however, incomplete. The trajectory gives all the possible positions, but no time indications. The time dependence, crucial for the construction of astronomical ephemerides predicting the planetary positions, is given by the second law.

Table 2. Conic sections.

Eccentricity	Conic section
$e = 0$	Circle
$0 < e < 1$	Ellipse
$e = 1$	Parabola
$e > 1$	Hyperbola

The orbit of an object in the gravitational field of another object is a conic section. The general mathematical form of conic trajectories expressed in polar coordinates (r,θ) is $r = p/(1 + e \cos \theta)$, with p the parameter of the conic section and e its eccentricity (see figure 7). The nature of the conic section depends on the value of the eccentricity.
The trajectories of the eight planets of the solar system are nearly circular. The closest point to the Sun is the perihelion and the farthest the aphelion. Perihelion and aphelion, aligned with the focus, define the line of apses, or major axis. They are separated by a distance $2a$, a being the semimajor axis, related to the parameter p and eccentricity e by $a = p/(1 - e^2)$ (figure 2).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

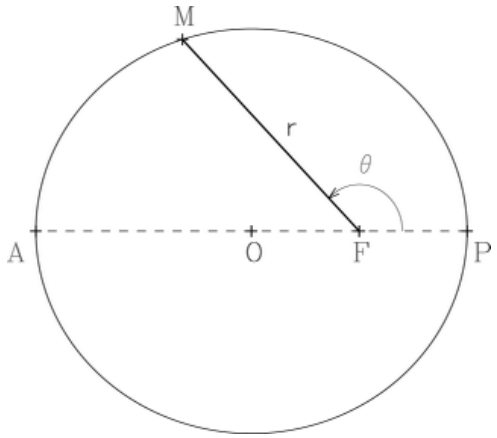


Figure 1. The trajectory of orbiting objects around the Sun is an ellipse, the Sun being located at one of the foci. The semimajor axis *a* is half of the distance between perihelion P and aphelion A; the distance between the center O and the focus F defines the eccentricity *e*: OF = *ae*. When considering objects orbiting the Earth, the closest and farthest points are respectively the perigee and apogee, and pericenter or apocenter in the general case. Small bodies in the solar system have usually greater eccentricities than the planets. For comets, the value of *e* is often very close to 1.

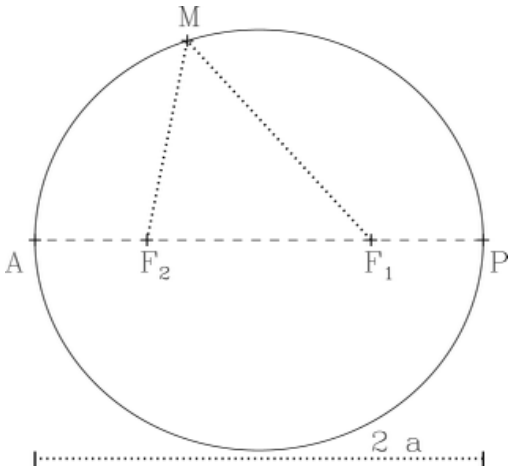


Figure 2. The sum of the distances between a given point of the ellipse and the two foci is constant and equal to 2*a*.

Kepler's second law

- Each planet moves in such a way along its orbit that a line drawn from the Sun to the planet sweeps an equal area in an equal amount of time.

The second law does not introduce the time dependence in a simple way, since, except in the particular case of a circular path, the position along the orbit cannot be expressed explicitly. It was therefore necessary for Kepler to consider not the position of the planet, but the area swept by the position vector around the Sun, in order to show that this area increases linearly with time (figure 3). The first step was to identify that a planet moves faster when closer to the Sun. The precise expression of this second law was carefully derived from the analysis of a very large amount of data concerning the position of Mars. In the first published version of this law, the discovery of the equal-area law was first derived from two false steps that by chance cancelled each other. Much later, Kepler gave an exact demonstration.

In fact, this second law establishes the possibility of astronomical prediction: it makes it possible to follow step by step the progression of a planet along its orbit. According to his first and second laws, Kepler constructed timetables for the planets, which were used for more than one century. They were much more precise than the astronomical tables based on the PTOLEMAIC SYSTEM. For the first time, predictions concerning the declinations of the planets were made possible (actually, the Ptolemaic system was unable to predict the declinations of the planets). More than the novelty of the approach, the accuracy of the predictions ensured the success of Kepler's laws.

Kepler's third law

- The ratio of the cube of the semimajor axis of orbit to the square of the orbital period is the same for all planets.

This third law introduces another concept compared with the two previous ones, which are restricted to one single object around the Sun. It gives a relation valid for all the objects orbiting the Sun. Kepler discovered the relation between the semimajor axis *a* of the orbit and the sidereal period *T*. The ratio T^2/a^3 is constant for all objects in the solar system orbiting the Sun (table 4). The introduction of sidereal rotation periods gave another argument in favor of a heliocentric view of the solar system, since sidereal periods are corrected for the Earth's motion.

Table 3. Orbital elements.

The complete description of a planetary orbit results from six integration constants. These constants can be, for a numerical approach, the initial values of the radius and velocity vectors. In terms of integral quantities, one defines the angular momentum σ and eccentricity vectors ***e***, plus the perihelion time. For a geometric view of the problems, one defines six orbital elements:
• semimajor axis *a*
• eccentricity *e*
• inclination *i* compared with the ecliptic
• longitude of the ascending node Ω (the lines of nodes is at intersection of the ecliptic plane and the planetary orbital plane)
• argument of the perihelion ω
• time of the perihelion τ

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

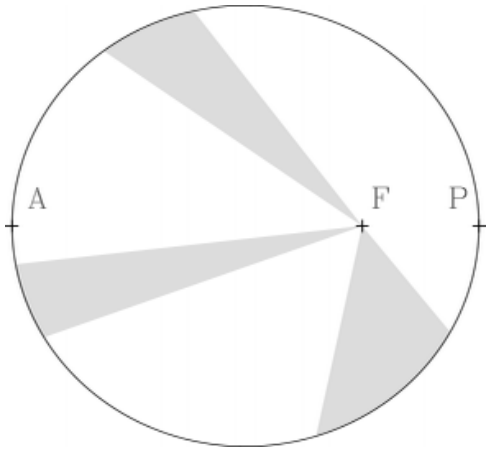


Figure 3. The body orbiting around the Sun moves in such a way that equal areas are swept out in equal amounts of time. This relation implies that the velocity decreases with increasing distance to the Sun. The maximum velocity is reached at perihelion, and the minimum one at aphelion.

Table 4. Semimajor axis and eccentricity.

Planet	a (AU)	T (yr)	e
Mercury	0.3871	0.2408	0.2061
Venus	0.7233	0.6152	0.0068
Earth	1.0000	1.0000	0.0167
Mars	1.5237	1.8809	0.0934
Jupiter	5.2026	11.862	0.0483
Saturn	9.5547	29.458	0.0560
Uranus	19.2181	84.010	0.0461
Neptune	30.1096	164.79	0.0097

The precision of the measurements of the semimajor axis a and the revolution period T is in fact limited by the planetary interactions perturbing the central solar gravitational field. With a in astronomical units and T in years, the ratio T^2/a^3 is very close to 1 for all inner planets whose masses are negligible compared with the solar mass. It is about 0.999 for Jupiter, since the mass of this planet represents one-thousandth of the solar mass, and around 0.995 for the other outer planets, owing to perturbations.

Kepler's work did not explain the physical reason for this law. The inverse square law of gravitation discovered by Isaac Newton shows that the ratio T^2/a^3 is determined by the mass of the Sun. The exact formula introduces the total mass of the system, including the mass of the planet, which is in fact negligible compared with the mass of the Sun.

The mathematical expression of Kepler's third law becomes very simple provided that the semi-major axis is expressed in astronomical units and the orbital period in years. One astronomical unit (1 AU) represents by definition the semimajor axis of the Earth's orbit ($1.495\,978\,7 \times 10^{11}$ m). Within this unit system, Kepler's third law becomes, for any object belonging to the solar system,

$$\frac{T^2}{a^3} = 1.$$

This relation allows us to determine the semimajor axis of the orbit, knowing its period, or vice versa. For example, the sidereal period of comet Halley being 76 yr on average, its semimajor axis is about 17.9 AU. According to Kepler's second law, and taking into account that the cometary perihelion is close to the Sun, this comet spends most of its time as far from the Sun as Neptune's orbit (figure 4).

Extension of Kepler's third law: mass determination

Kepler's third law offers the unique possibility to measure masses in the universe, when extended to binary systems. The constant which appears in this law is expressed as a function of the total mass $m_1 + m_2$ of both objects of the system and of the gravitational constant G (experimentally determined first by Henry Cavendish in 1798):

$$\frac{T^2}{a^3} = \frac{4\pi^2}{\mathcal{G}(m_1 + m_2)}.$$

The orbital period T and semimajor axis a can be derived by direct or indirect measurements. Therefore, their observation in a binary system may provide the measurement of the total mass of the system. Further indications, if available, are necessary to determine independently the individual masses m_1 and m_2.

As previously, with distances expressed in astronomical units, time in years, and mass M in solar mass units, the third law becomes (with $\mu = M/M_\odot$ and $M_\odot$ the mass of the Sun):

$$\frac{T^2}{a^3} = \frac{1}{\mu}.$$

As an example, the mass of the asteroid 45 Eugenia, surrounded by a small satellite discovered in 1998 at CFHT with ADAPTIVE OPTICS (figure 5) was derived from the orbital distance and period of this satellite, about respectively 1190 km (7.9×10^{-6} AU) and 4.7 days (0.0129 yr). This leads to a mass of about $3 \times 10^{-12} M_\odot$, or 6×10^{18} kg, for Eugenia.

Historical overview

Education

According to himself, Johannes Kepler was born on 27 December 1571 at 2:30 pm, after a pregnancy of 224

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

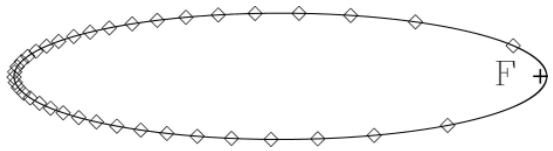


Figure 4. Illustration of an elliptical orbit with an eccentricity close to 1 (e = 0.967, as for comet Halley). The time interval between two consecutive diamonds is constant.

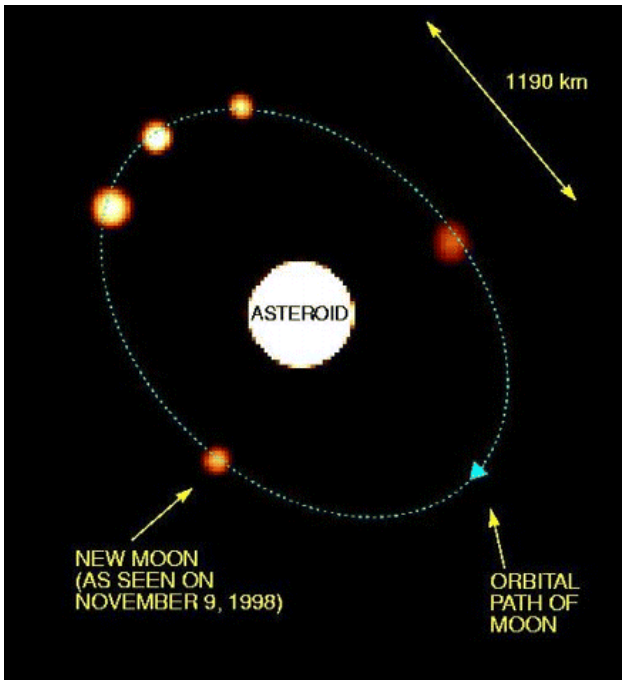


Figure 5. The asteroid 45 Eugenia and its satellite, in a composed view of images taken with adaptive optics at the CFH Telescope.

days, 9 h and 53 min, since he was conceived on 16 May at 4:37 am. This anecdote shows Kepler's close attachment to numbers. During his childhood in Weil (today in Baden-Württemberg, Germany) in a disunited family, he endured a very weak constitution, suffering from recurrent gastric problems, as well as from myopia and polyopia. His father became a mercenary and did not take care of his family; his mother narrowly missed being burnt at the stake. By chance, as he early proved himself to be a prodigious child, Kepler benefitted from the performance-based educational system founded by the Dukes of Württemberg. As Lutherians, they had promoted such a system in order to turn out the erudite clergy and the efficient administration they needed.

Johannes Kepler's education followed the classical steps: primary school, high school, seminary and finally university at Tübingen in 1589. There, he learned Greek, Hebrew, dialectics, ethics and physics. He met Michael Maestlin, who taught mathematics and converted Kepler to the Copernican system. As a young student, Kepler was already convinced about the mathematical superiority of the Copernican system compared with the Ptolemaic system. Maestlin derived from his observations of a comet in 1577 that only the Copernican system was able to define an adequate frame for the cometary orbit.

Kepler, who intended to pursue a position in religious life, was finally allowed in 1591 to devote his attention to theology. However, he did not obtain his doctorate directly, since he accepted in 1594 a position of *mathematicus* in Graz, Styria, where he started teaching the basic of astronomy. He acceded to the status of imperial mathematician in 1601, but went on practicing astrology his whole life long.

His first original work, *Mysterium Cosmographicum* (*The Sense of the Universe*), published in 1596, showed that, despite his encounter with the views of Copernicus, Kepler was bound by preconceived ideas derived from the Ptolemaic system. He considered the solar system as being governed by geometric relationships. He was convinced that all the objects moving in the sky had to be in harmony with divine secrets. Remarking that the known planets (from Mercury to Saturn, plus the Moon) were six (six being a perfect number since 6=1 + 2 + 3, namely the sum of all its divisors), he tried to find some trigonometrical relation between the distances of the planets to the Sun. Of course he failed. He also tried to organize the planetary orbits, ranged according to the heliocentric system, into the spheres inscribed in and circumscribed around the five regular polyhedra (figure 6). The agreement with observed data was quite acceptable.

Tycho Brahe's observations

In 1597, after a strong credit reduction following the death of the king, Tycho Brahe left his observatory in Denmark for Prague. Three years later, Kepler left his chair in Graz, because of the counter reformation and the persecution of Protestants. He accepted an invitation in Prague from Tycho Brahe and met him there. Their encounter proved to be very fruitful, despite tumultuous relationships because Brahe considered Kepler as an assistant instead of an equal. Tycho Brahe convinced Kepler to leave his numerological speculations. Kepler took Tycho's observations, made with very precise instruments but naked eyes, into the highest consideration. However, he did not try to follow Tycho's cosmological system with all planets orbiting the Sun but the Sun orbiting the Earth.

When Kepler started his collaboration with Tycho Brahe, he benefitted from the observations and theory of Tycho and of his disciple Longomontanus. Kepler was first supposed to work on the Moon's motion. However, he wished to work on the planetary eccentricities, and especially on Mars, since he was convinced, indeed rightly, that the orbit of Mars presented the strongest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

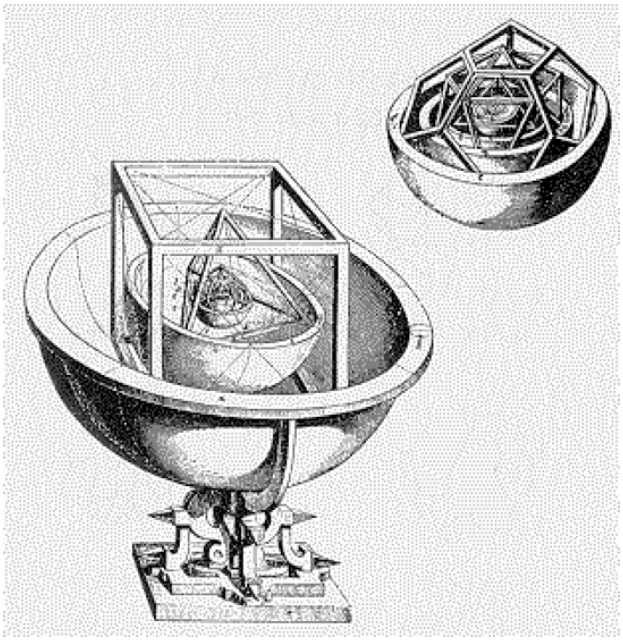


Figure 6. The five regular polyhedra and the planetary orbits. The orbit of the Earth is circumscribed by a perfect dodecahedron and a sphere containing the orbit of Mars, circumscribed by a tetrahedron and Jupiter's orbit, circumscribed by a cube and Saturn's orbit. Inscribing an icosahedron in the Earth's orbital sphere gives Venus's sphere, and then Mercury's sphere is inscribed in an octahedron.

eccentricity among the known planets—Mercury's eccentricity is actually greater, but the observations of this planet so close to the Sun cannot reach the same precision. Kepler found unacceptable the differences in longitudes and the complete disagreement in declinations between the best theoretical expectations and the observations. The differences reached more than 8 arcmin, compared with the 2 arcmin precision of Tycho's observations. Therefore, Kepler decided to abandon the Ptolemaic system and its derivatives, composed of series of epicycles. He then had to search for a new model.

Kepler's work

The fundamental novelty of his approach consists in the place devoted to the Sun. Contrary to Ptolemy, who considered the Earth as the unique point common to all planetary orbital planes, contrary to NICOLAUS COPERNICUS and to Tycho Brahe, who assigned this role, respectively, to the center of the Earth's orbit or to the center of the Sun's orbit, he calculated that the inclination of Mars' orbit was unambiguously defined when considering that the planetary orbital plane contains the Sun. He estimated this inclination to be about 1º50', comparable with the value accepted today: 1º51'. According to this observational conclusion, Kepler was reinforced in the relevance of the heliocentric system: he had in fact discovered that the Sun defined the best frame for a precise dynamical analysis.

He was aware, following the observations of Tycho Brahe, that the Earth does not move with a uniform speed, but slower or faster depending on the distance from the Sun. He recognized, and this constitutes the basis of his discovery, that the speed of the Earth at aphelion and perihelion is inversely proportional to the distance to the Sun at these points. He derived from these observations two false ideas that by chance cancelled—forces are inversely proportional to distances, but proportional to the velocity—and combined them in order to derive his second law. The second law was established first. Later, Kepler corrected his first form of the second law. He explained that not the velocity but its orthoradial component (perpendicular to the radius vector) is inversely proportional to the distance of the Sun.

In order to establish the shape of Mars' orbit, he made the assumption that the Earth orbit was circular, what is not too far from reality, despite his second law. The difference between the semimajor and semiminor axes of an ellipse is a second-order term of the eccentricity (figure 7). Nevertheless, it took him 8 yr to establish the laws, which became later Kepler's first and second laws. They were published in 1609 in the *Astronomia Nova* (*New Astronomy*). The third law, which finally needed another 9 yr of work, was published by Kepler in 1618, in *Harmonices Mundi* (*Harmony of the Universe*). In 1627, 3 yr before his death, he published the *Tabulae Rudolphinae*, in honor of his protector, which presented the planetary ephemeris based on his three laws.

A revolutionary approach

Why did the geocentric system, devised by Claudius Ptolemy, survive almost 1500 yr, since it required more than 70 simultaneous circular motions to derive the movements of the inner objects of the solar system? It gave an accurate description enough of what could be observed with the instruments at that time. It was able to predict the future paths of the planets well enough, according to the standards of those times. The planetary declinations were not predicted, but physical reasons could make explain this: the progressions of the planets, etymologically wandering objects, differ intrinsically from the stellar regularity by their annual parallax and complex motions in right ascension. Furthermore, this system completely agreed with the philosophical and cosmogonical doctrines prevailing at the time. Celestial motions had to be expressed in terms of combinations of circles, since the circle and uniform circular motion represented harmony. The careful analysis which puts the Sun at the center of the solar system and considers the Earth as one of the planets shows that orbital motion can be formulated within an adequate theoretical frame but needs to abandon the apparent evidence that the Earth holds a central position.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

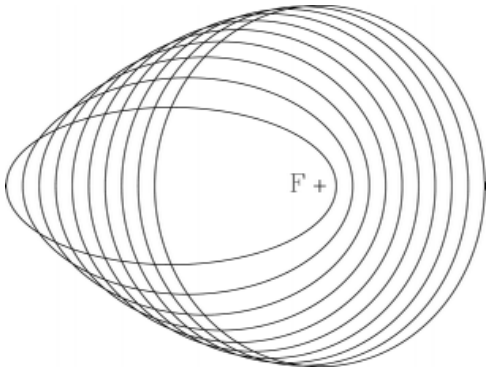


Figure 7. Set of ellipses with the same semimajor axis *a* but varying eccentricities *e* (0, 0.1, 0.2,…,0.9). Ellipses with small eccentricity remain nearly circular, but clearly decentered with respect to the focus F.

As described above, one cannot explain the role of Johannes Kepler in modern astronomy without reference to Nicolaus Copernicus and Tycho Brahe. Kepler was convinced of the superiority and perfection of the heliocentric theory, since that theory permitted in fact more harmony and simplicity than the previous view. He trusted in Tycho Brahe's observations, being of course impressed by the precision achieved, 10 times better than previously. Kepler actually had an original attitude towards observational results. He tried for a while to explain the agency of the planetary orbits with pure geometric methods. Since such theories failed to accurately fit observational data, he was to abandon this view and searched in other directions. When attempting to account for observed facts with new mathematical concepts, he succeeded, and then explored this way. Thus, Kepler admitted the ontological difference between the ideal figure—the circle—and its physical representation—the ellipse. The 9 yr separating the first two laws and the third one indicate that his new approach was not an easy way. Let us remember that physical notions such as instantaneous speed or differential calculus had not yet been invented.

Finally, Kepler's work can be considered as a cornerstone of physics and science during the 17th century. They were implicitly sustained by the following ingredients:

- observed facts prevail over *a priori* philosophical or theoretical schemes;
- the physical analysis needs to be conducted in an adequate reference frame;
- mathematical equations allow us to give physical laws in a condensed form.

These points must not induce the misinterpretation that Kepler's work was linearly developed, without error, delusion or fallacy. The greatest success of Kepler is due to his intense calculations: he was able to explore in detail numerous directions and to finally chose the best one. On the other hand, Kepler's laws remain purely kinematic: they describe the motion of planets but do not explain the cause of the movement. After Johannes Kepler, Galileo Galilei stated the essential concept of inertial frame, and Isaac Newton realized the synthesis of dynamics laws.

Beyond Kepler's laws

Implicit hypotheses

Through a modern physical approach, Kepler's laws are derived from the dynamics of a two-body system. The physics of such a system is based on the assumption that the two bodies are isolated and move under their mutual gravitational attraction. Thus, in the solar system, a given planet is supposed to only interact with the Sun, the perturbation due to the other planets being neglected. Kepler's laws also implicitly assume that the mass of the planet is negligible compared with the mass of the Sun. To a first approximation, this is true: the mass of the Sun represents 1000 masses of Jupiter. In fact, Newton's laws show that this second assumption can be relaxed. As soon as the two bodies remain isolated, their BARYCENTER furnishes a valid inertial frame. Then, instead of orbiting the Sun, the planet (the Sun as well) orbits the barycenter of the system.

Integrals

It is possible to express the content of Kepler's laws in very different ways. Let us examine the two-body interaction in terms of constant physical quantities.

- The two bodies being isolated, the total momentum of the system is constant. Then the frame based on the barycenter of the system is inertial.
- The force between the two bodies being central, and always directed towards the barycenter of the system, the angular momentum is constant. The plane of the trajectory contains the barycenter of the system and is perpendicular to the angular momentum.
- The gravitational interaction force deriving from a potential energy, the total mechanical energy of the system remains constant. The semimajor axis *a* depends only on the mechanical energy and not on the angular momentum.
- Another constant term comes from the fact that the force varies as the inverse of the square of the distance. The direction of the eccentricity vector indicates the direction of the pericenter (figure 8); its norm is the eccentricity *e*. Bound states have a negative energy and $e < 1$: they correspond to elliptical trajectories. Diffusion states have $e \geq 1$ and a positive energy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

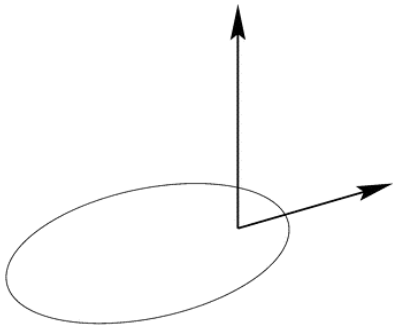


Figure 8. The trajectory plane and the direction of the major axis defined by the angular momentum and eccentricity vectors.

N-body interaction

What happens when the two-body interaction is not sufficient to account for reality? The recent case of the capture of comet Shoemaker–Levy 9 by Jupiter and the collision of the cometary fragments with Jupiter in July 1994 shows that the solar system is not only governed by the solar gravitational field. In a refined analysis, the perturbations due to the more massive bodies—especially the giant planets—are no longer negligible for long-term evolution, so that the two-body description, including only the central solar gravitational field, is invalid for an accurate description. The absence of any large object between Mars and Jupiter, the inclination of Uranus's rotation axis, the resonance between the motions of Neptune and Pluto (the ratio of their orbital periods is exactly 2:3) and the planetary rings, indicate that interactions between multiple objects (called *N*-body interaction) play a major role in the long-term evolution and stability of the solar system. New tools have been developed for such problems, such as Lagrangian mechanics. However, numerical computations are necessary to carry out the solutions of the *N*-BODY PROBLEM.

Bibliography

Crombie A C 1952 *Augustine to Galileo, the History of Science A.D 400–1650* (London: Falcon)

Dreyer J L E 1953 *A History from Thales to Kepler* (New York: Dover)

Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (New York: Freeman)

Verdet J-P 1993 *Astronomie & Astrophysique: Textes Essentiels* (Paris: Larousse)

Wilson C 1972 How did Kepler discover his first two laws? *Sci. Am.* **226** 92–106

Benoît Mosser

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kepler's Star

A supernova remnant in Ophiuchus, approximately 2.2° east of ξ Ophiuchi. The star was first seen in October 1604 by Johannes Kepler (1571–1630), and described in his book *De Stella Nova*. It reached a maximum apparent magnitude of -2.5 and remained visible to the unaided eye for about a year. The records indicate that it was a type Ia supernova. Today the visible remnants comprise just a few faint filaments and knotted fragments of nebulosity, but it is a strong source at radio wavelengths, with the identification 3C 358.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Keyhole Nebula (NGC 3324)

A dark nebula in the constellation Carina, position RA 10^{h} 44.3^{m}, dec. $-59^{\circ}\,53'$. It is seen in silhouette against the bright Eta Carinae Nebula.

See: Eta Carinae Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Khayyam [al-Khayyam], Omar [Abu ol-Fath ebn-Ebrahim ‘Omar ol-Khayyami of Nishapur] (c. 1044–c. 1123)

Poet, mathematician, astronomer, born in Nishapur, Persia (now Iran), to a family of Arabic tent makers (a translation of his name), an outstanding mathematician and astronomer. The Saljuq Sultan, Malikshah Jalal al-Din, appointed him to the observatory at Ray around 1074 to determine the solar calendar, as an aid to tax collection. Khayyam developed a calendar called *Al-Tarikh-al-Jalali*, which had an error of one day in 3770 years (superior to the Georgian calendar with an error of 1 day in 3330 years). It was based on Khayyam’s amazingly accurate determination of the length of the year as 365.242199 days. The length of the year is currently 365.242190 days, decreasing by about six units in the sixth decimal place over a century. Khayyam is best known as a result of Edward Fitzgerald’s translation in 1859 of nearly 600 short four-line poems, the *Rubaiyat of Omar Khayyam*, but versions of the verses in Persian literature pre-date Khayyam, and no-one knows which were written by him. The best known is the fateful quatrain:

> The Moving Finger writes; and, having writ,
> Moves on: nor all thy Piety nor Wit
> Shall lure it back to cancel half a Line,
> Nor all thy Tears wash out a Word of it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kiddinu (340 BC–?)

Head of an astronomical school in Babylon. He worked out the precession of the equinoxes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kiepenheuer-Institut für Sonnenphysik

The Kiepenheuer-Institut für Sonnenphysik (KIS) conducts experimental and theoretical investigations of physical processes on and within the Sun. Its headquarters are in Freiburg, Germany. The number of employees (scientific, technical and administrative) is about forty. The institute was founded in 1943 by Karl-Otto Kiepenheuer and, since 1978, it has borne the name of its founder. The KIS is a research institution of the State of Baden-Württemberg. The main observing facility of the KIS is the vacuum tower telescope within the OBSERVATORIO DEL TEIDE in Izaña on Tenerife, Spain.

For further information see http://www.kis.uni-freiburg.de.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kinetic Theory of Gases

The theory, developed in the nineteenth century, notably by Rudolf Clausius (1822–88) and James Clerk Maxwell (1831–79), that the properties of a gas (temperature, pressure, etc) could be described in terms of the motions (and kinetic energy) of the molecules comprising the gases. The theory has wide implications in astrophysics. In particular, the perfect gas law, which relates the pressure, volume, temperature, and number of molecules in a gas, and which is fundamental to theoretical models of the interior of stars, is consistent with this theory.

See also: atmospheric pressure, combustion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kirch family

Kirch, Gottfried (1639–1710)

Kirch, Maria Margarethe, née Winckelmann (1670–1720) his wife

Kirch, Christfried (1694–1740); and Kirch, Christine, his children

A family of astronomers and calendar-makers. Gottfried Kirch was born in Guben (Germany) and acquired a 4 foot quadrant and some telescopes, making observations of comets, including the comet of 1680, apparently the first comet to be discovered with a telescope. He discovered the star chi Cygni to be variable and observed Mira. He became director at the Berlin observatory in 1700 and began to prepare its calendars and ephemeredes. Christfried Kirch became director of the observatory himself in 1716. He observed star occultations, solar and lunar eclipses, Jupiter, Saturn and the comet of 1723. Maria (born in Panitsch, Germany) became interested in astronomy through Christoph Arnold of Sommerfeld, the so-called 'astronomical peasant', a self-taught astronomer. In 1692, she married Gottfried, working with him regularly, making observations and especially doing calculations. She discovered the comet of 1702. After Gottfried's death (1710), she continued to publish on her own, and when Christoph returned to Berlin, she worked with him too. Father, mother, son and daughter, Christine—the Kirch's published astronomical ephemeredes essentially as a family business.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kircher, Athanasius (1602–80)

Polymath, born in Geisa a.d. Ulster, Germany, became a Jesuit. After a succession of academic appointments, he took up independent studies. He wrote an enormous range of books and sold exclusive rights to publish them for a large sum of money, so he was perhaps the first scientist who was able to live from the sale of his books. He corresponded with prominent astronomers such as GB RICCIOLI, GD CASSINI and HEVELIUS and acted as a kind of bulletin board for their observations. In one book Kircher describes a fantastic voyage through space to discuss his theories of the solar system. Accompanied by the 'music of the spheres', the two protagonists travel to other worlds, and converse with the intelligent life forms they find there. Kircher's philosophy was a blend of science and superstition combining empirical observation with magic and religion, foreshadowing popular science fantasy today.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kirchhoff, Gustav Robert (1824–87)

Born in Königsberg, Prussia (now Kaliningrad, Russia), became professor of physics at Heidelberg where he collaborated with Bunsen, later professor in Berlin. With *Kirchhoff's laws of electricity* he extended the work of Ohm. He worked on black-body radiation and spectrum analysis. With *Kirchhoff's laws of radiation* he explained the dark lines in the Sun's spectrum as caused by absorption of particular wavelengths as light from a hot source passes through a cooler gas. He showed that a given element produced the same characteristic pattern of emission and absorption lines. This began the topic of stellar astrophysics, by which the composition, density, temperature and other physical conditions of a star's atmosphere can be found.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kirkwood Gaps

Narrow regions in the asteroid belt where few or no asteroids are found. The gaps arise from orbital resonances with Jupiter: commensurabilities of the period of Jupiter with the period of an asteroid orbiting in a Kirkwood gap. Because of Jupiter's dominant gravitational influence, an object orbiting in one of these gaps would sooner or later be perturbed into a different orbit. For example, there is a pronounced gap at 3.28 AU, because an asteroid orbiting at this distance would have a period exactly half of Jupiter's. This orbit would be unstable, and the asteroid would soon be perturbed into a different one. This particular gap corresponds to a 2:1 commensurability and is often referred to as the 2:1 resonance; it marks the boundary between the Cybele group of asteroids and the outer edge of the main asteroid belt. Other gaps are found at other simple ratios of orbital periods, for example 3:1 and 5:2. However, at distances beyond about 3.5 AU the resonances correspond not to gaps but to groupings of asteroids: the Hilda group at the 3:2 resonance, for example.

A similar mechanism is responsible for shaping planetary ring systems. For example, the Cassini division in Saturn's rings corresponds to a 2:1 resonance with the orbit of the satellite Mimas, and ring particles at the outside of the A ring are in a 7:6 resonance with the co-orbital satellites Epimetheus and Janus. The connection between gaps and resonances was first pointed out in the mid-nineteenth century by Daniel Kirkwood.

See also: asteroid belt.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kirkwood, Daniel (1814–95)

Astronomer and mathematician, born in Harford, MD, became professor of mathematics at Delaware and at Indiana. He explained the structure of the distribution of asteroids in the asteroid belt as a result of the influence of Jupiter. If an asteroid resonates with Jupiter, i.e. has an orbital period that is a simple fraction (1/2, 2/3, ...) of the period of Jupiter, the repeated perturbation over time by Jupiter on the asteroid, when they have a close approach, ejects it into an eccentric, Earth-crossing orbit. The resulting lack of asteroids with these periods in the asteroid belt became known as *Kirkwood gaps*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kitt Peak National Observatory

Kitt Peak National Observatory (KPNO) is one of the cornerstones of the US system of publicly accessible astronomical facilities. Its two 4 m class telescopes are increasingly specialized towards wide-field applications that expand the discovery space for NASA missions and ground-based large-aperture telescopes. Time is awarded on the basis of scientific merit through a national system of peer review.

The US National Science Foundation established KPNO in 1958 under the leadership of the ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY. The site is the summit of a desert mountain at 2100 m, located 85 km kilometers southwest of Tucson, Arizona, on the land of the Tohono O'odham Nation. It was selected for clear skies, good seeing and a weather pattern allowing access to the North Galactic Cap.

KPNO offers time on three major telescopes: the Mayall 4 m; the modern 3.5 m WIYN telescope operated by the partnership of the Universities of Wisconsin, Indiana, Yale and the National Observatories; and the 2.1 m. Premier scientific instruments include an 8000 pixel-square CCD Mosaic imager at the 4 m prime focus, a bench spectrograph coupled to the 1 degree focal surface by 100 optical fibers at the WIYN, and a simultaneous four-color near-infrared imager at the 2.1 m.

KPNO also acts as the site host for a number of major facilities, including telescopes of the NATIONAL SOLAR OBSERVATORY, the NATIONAL RADIO ASTRONOMY OBSERVATORY, the University of Arizona and eight university-based consortia.

For further information see
http://www.noas.edu/kpno.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kleinmann–Low Nebula (KL Nebula)

An extended, dusty region of star formation behind the Orion Nebula, position RA $05^{h}\ 35.2^{m}$, dec. $-05^{\circ}\ 22'$, discovered in 1967 by Douglas Kleinmann and Frank Low. There are several discrete sources of infrared radiation, including the Becklin–Neugebauer object, thought to be young stars or protostars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kolmogorov, Andrey Nikolayevich (1903–87)

Mathematician, born in Tambov, Tambov province, Russia, became professor in Moscow, developed probability theory, building it up from fundamental axioms, applied his work to turbulence (his theory is used in the study of stellar interiors) and dynamical systems in relation to planetary motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Konkoly Observatory

Konkoly Observatory in Budapest is equipped with a 1 m Ritchey–Chrétien–Coudé telescope, carrying out observational astronomical studies, mainly related to the physics of the variable stars, galactic structure, solar activity and the terrestrial upper atmosphere. These tasks imply operation of the internationally recognized observational network developed during the last three decades. A very important task—perhaps the most important one—is the development of the national astronomical information system, including the observatory's library.

The main areas of work include studies concerning the behavior of variable stars: investigation of multiple periodicity and period changes of pulsating variables, as well as research on stellar activity of various time-scales. Studies related to the galactic structure and physics of interstellar matter with an emphasis on the star-forming processes. Studies of the upper atmosphere of the Earth and Mars with space-borne equipment. Studies on solar activity, investigation of the problems of the problems of sunspots and prominences.

For further information see
http://www.konkoly.hu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Korff, Serge Alexander (1906–89)

Astrophysicist, born in Helsinki, Finland, fled the Bolsheviks and researched at Mount Wilson Observatory, CalTech and New York University, investigated cosmic radiation, making observations from observatories, balloons and aircraft worldwide.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Korolyov [Koroljow], Sergej [Sergey, Sergei] Pavlovich [Pawlowitsch] (1907–66)

Rocket designer, founder of the Soviet space program, born in Zhitomir, Russia. After pioneering experiments on rockets in Moscow, Korolyov fell foul of Stalin's pogroms but was recalled from the stalags to make liquid-fuel rocket boosters for military aircraft. After the war he took over captured German V-2 missiles, and developed the Soviet Union's first intercontinental ballistic missile. He directed the design, testing, construction and launching of the Vostok, Voskhod and Soyuz manned spacecraft as well as of the unmanned spacecraft in the Cosmos, Molniya and Zond series. His identity and his key work for Russia's space program, as a Soviet analog to the USA's VON BRAUN, were not known in the West until his death.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kraus, John (1910–)

Radioastronomer, born in Ann Arbor, MI, became professor of electrical engineering and astronomy at Ohio State University, and director of the Ohio State–Ohio Wesleyan Radio Observatory. Author of widely used textbooks on radio techniques, he is the inventor of many types of radio antennas. Surveyed the radio sky at centimeter wavelengths, mapped and cataloged some 20 000 radio sources, including many quasars. Instigator of an extensive search for extraterrestrial artificial radio signals. The search serendipitously discovered natural narrow-band signals, interpreted as cold, isolated neutral hydrogen clouds, and in 1977 found the 'WOW!' signal (annotation by Jerry Ehman on the computer print-out) that, being celestial, narrow-band and intermittent, suggested an intelligent origin. However, the signal never recurred from that direction and is now attributed to a signal from a secret Earth-launched space probe, rather than a star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kroto, Sir Harold (1939–)

Spectroscopist, born 'Krotoschiner' in Wisbech, Cambridgeshire. Nobel prizewinner for chemistry (1996) with chemists Robert F Curl and Richard E Smalley 'for their discovery of fullerenes'. Became professor at the University of Sussex. Worked on microwave spectroscopy, measuring and calculating the spectra of small molecules. Although many such molecules do not exist naturally on Earth, they exist in space and can be detected by radioastronomy, both in stellar atmospheres and in interstellar gas clouds. Kroto was particularly interested in carbon-rich giant stars and had identified spectral lines in their atmospheres as long-chained molecules of carbon and nitrogen, termed cyanopolyynes. To study their formation, he allied with physical chemists Smalley and Curl and they reproduced conditions in stellar atmospheres in laboratory equipment. Studying the chain of events that might produce cyanopolyynes, they discovered carbon in clusters of 60 atoms arranged in a soccer ball-like structure called buckminsterfullerene, after the American architect R Buckminster Fuller, inventor of the geodesic dome. Like graphite and diamond, C-60 fullerene is expected to be a constituent of interstellar space, where it will become hydrogenated by attached hydrogen atoms.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kuhn, Thomas S (1922–96)

Science historian, born in Cincinnati, OH. Developed a theory of scientific revolution which visualized science not as a steady progression of knowledge, but as 'a series of peaceful interludes punctuated by intellectually violent revolutions' in which one world view is suddenly replaced by another. For example, EINSTEIN's theory of relativity overthrew Newton's concepts of physics, and GALILEO's legendary experiments banished the Aristotelian theory that bodies fell at a speed proportional to their weight. But scientists were not typically objective, free thinkers, Kahn claimed. They accepted and usually related their work to a paradigm, an archetypal solution to a problem. Thus, astronomers resisted, for a long time, research that challenged the paradigm of Ptolemy's theory that the Sun revolves around the Earth, setting aside, for example, the work of ARISTARCHUS, who suggested as long ago as the third century BC that the planets revolve around the Sun. It took nearly 2000 years before the Ptolemaic paradigm was suddenly replaced by the new paradigm of the scientific revolutionary, COPERNICUS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kuiper Belt

The trans-Neptunian solar system is now known to contain a vast population of icy bodies called the Kuiper Belt, first discovered by David Jewitt and Jane Luu in August of 1992. The discovery confirmed that the region immediately beyond the planets is not merely empty space but is instead richly populated with a vast reservoir of fossil remnants from the original solar nebula. The enormous scientific potential of the Kuiper Belt was soon recognized and the belt is now the subject of intensive observational and dynamical studies.

The interest in the Kuiper Belt originated with the problem of the origin of the short-period COMETS (comets with orbital periods shorter than 200 yr). Jan OORT proved in 1950 that the long-period comets (those comets with orbital periods longer than 200 yr) come from the OORT CLOUD, a spherical cloud with radius on the order of 10 000 AU. It was thus simply assumed that the short-period comets were somehow derived from the long-period comet population, perhaps through capture by the giant planets, primarily Jupiter.

This assumption was cast in doubt in the 1970s, when Paul Joss pointed out that the probability of capture of the long-period comets by Jupiter was too small to explain the observed number of short-period comets. Because of the limited computer power at the time, Joss' result went largely unconfirmed until 1988, when the Canadian astronomers Martin Duncan, Tom Quinn and Scott Tremaine carried out computer simulations of the capture process of the long-period comets by the giant planets. They confirmed Joss' result, but also found that, during capture, the orbital inclination distribution was roughly preserved. This means that, if the original population contained comets with high inclinations, the captured population should also contain comets with high inclinations. This result agrees very well with observations of the long-period comets: the Oort cloud is a spherical cloud, implying that the long-period comet inclinations should also be roughly spherical, which is indeed the case. However, the short-period comets mostly have low-inclination orbits, inclined less than 30° from the ecliptic plane. The simulations' results thus imply that a spherical cloud like the Oort cloud could not produce a flat structure like the short-period comets; instead, a much better source would be a flat belt of comets in the outer solar system. This comet belt is now called the Kuiper Belt, after the Dutch astronomer Gerard KUIPER.

This idea of a trans-Neptunian comet belt was actually first mentioned in 1949 by an Irishman named Kenneth EDGEWORTH, then elaborated upon by Gerard Kuiper in 1951. Their basic argument was that there was no *a priori* reason why the solar system should end abruptly at Neptune or Pluto. Perhaps large planets could not form beyond Neptune, but smaller bodies might be able to. Edgeworth and Kuiper thus envisioned a population consisting of residual 'PLANETESIMALS' (small solid bodies) left over from the formation of the planetary system. Also, because of the low temperatures, it seemed likely that these distant objects would be ice rich, identical to comets.

In 1987, shortly before Duncan, Quinn and Tremaine published their results, spurred on by the dramatic contrast between the apparent emptiness of the outer solar system and the richly populated inner solar system (remember that Pluto was the only permanent solar system object known beyond Neptune at the time), Jewitt and Luu started a telescopic survey in order to address the question of the contents of the outer solar system. After five years of searching, they detected the first Kuiper Belt object (KBO), 1992 QB1, in August 1992. The Kuiper Belt has become the new frontier of the solar system.

In 2002, astronomers determined that 2002 LM60, an icy Kuiper belt object dubbed Quaoar, by its discoverers, is the largest body found in the solar system since the discovery of Pluto 72 years ago. Quaoar (pronounced kwa-whar) is about half the size of Pluto. Like Pluto, Quaoar dwells in the Kuiper belt. It is the farthest object in the solar system ever to be resolved by a telescope. Quaoar is about 4 billion miles away from Earth, well over a billion miles farther away than Pluto. Unlike Pluto, its orbit around the Sun is circular, even more so than most of the planetary-class bodies in the solar system. Although smaller than Pluto, Quaoar is greater in volume than all the asteroids combined (though probably only one-third the mass of the asteroid belt, because it is icy rather than rocky). Quaoar's composition is thought to be largely ices mixed with rock, not unlike the

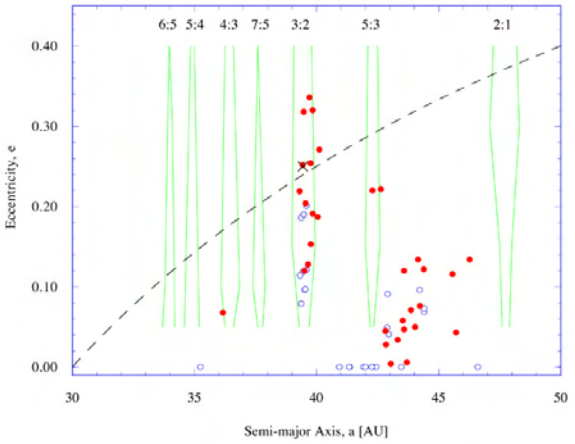


Figure 1. Semimajor axis versus eccentricity for KBOs. Solid circles denote objects observed over at least 2 yr (secure orbits), empty circles objects observed only once (probably lost). Pluto is marked by X. Neptune is at $a = 30$ AU. The dashed curve marks a perihelion distance of 30 AU; objects above the line are Neptune crossers (i.e. orbits are unstable). Orbital elements computed by Brian Marsden.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

makeup of a comet, though 100 million times greater in volume.

Structure of the Kuiper Belt

Neptune defines the inner edge of the Kuiper Belt. While the outer edge of the belt remains unknown, it is known that the belt extends at least as far as ~130 AU (see the section on scattered KBOs). Researchers have counted 346 KBOs since they were first recognized in the early 1990s, but, assuming a 4% albedo for the ones known, it is estimated that there are at least 70 000 of them with diameters larger than 60 mi (100 km). The known KBOs are sufficient to give us a general idea of the radial distribution of the belt, which is plotted in figure 1. It can be seen that the radial distribution of the belt is highly non-uniform. The region from Neptune to ~42 AU is empty owing to perturbations by the planet, except for a concentration of objects at 39 AU, the location of the 3:2 mean motion resonance of Neptune. Based on their dynamics, the known KBOs can be divided roughly into three distinct categories: the classical KBOs, the resonance KBOs and the scattered KBOs.

Classical KBOs

The classical KBOs are characterized by low-eccentricity, low-inclination orbits, suggesting that they formed directly from the solar nebula and have managed to preserve their primordial orbits.

Resonance KBOs

The Kuiper Belt contains a number of 'mean motion resonances' with Neptune, regions where periodic perturbations by the planet have induced special dynamical properties restricted to the regions. An object in an m:n resonance would make n orbits around the Sun for every m orbits completed by the planet. For example, Pluto is in the 3:2 resonance (located at 39 AU), meaning that it completes 2 orbits for every 3 orbits completed by Neptune. Renu Malhotra showed that some of these resonances are stable, which should lead to population enhancement inside the resonances, even if their immediate environs are unstable owing to Neptunian perturbations. However, the only resonance currently known to be densely populated is the 3:2 mean motion resonance. Just as the resonance protects Pluto from close encounters with Neptune, it also protects other members of the resonance (now called 'PLUTINOS' to highlight their dynamical similarity with Pluto). In view of the abundance of the Plutinos, Pluto is then best understood as the largest known KBO, rather than as a deviant planet. (It should be noted that Pluto may not be the only large body that formed in the Kuiper Belt: because of its strong similarities with Pluto, it is suspected that the Neptune satellite TRITON also originated in the Belt but has somehow been captured into an orbit around Neptune.) The apparent fraction of Plutinos among the known KBOs is roughly 35%, but this is an overestimate owing to several observational biases (including the fact that the resonance lies near the inner edge of the Kuiper Belt, rendering Plutinos easier to detect). The bias-corrected Plutino fraction is in fact closer to 10–20%, and the number of Plutinos larger than 100 km is ~30 000. The discovery of the Plutinos was completely unexpected, unpredicted by any theory.

Given that objects inside 42 AU are quickly ejected by Neptune, the question is as to how the 3:2 resonance became populated. Malhotra proposed that the resonance became populated during the radial migration of the giant planets. It has long been known that, during the final stages of planet formation, the giant planets grew large enough to scatter away nearby planetesimals. During the scattering process the planets exchanged angular momentum with the planetesimals, but, because of their different masses, the exchange is different for each planet. The net effect was that Jupiter lost angular momentum and its orbit shrank, while the orbits of the other giant planets expanded outward. Neptune may have migrated as much as 5 AU outward from its original location. As Neptune moved outwards, its mean motion resonances also moved with it, sweeping through the planetesimal disk. Along its path the 3:2 resonance could have captured PLUTO–CHARON, along with the Plutinos. The resonance capture theory also predicts clear population enhancements at other resonances with a strong capture probability, such as the 2:1 resonance at 48 AU. No 2:1 resonance KBO has been identified at this time but this may be due to the fact that the 2:1 resonance is significantly further away than the 3:2 resonance and its members would be much more difficult to detect. Malhotra pointed out that how many planetesimals these resonances succeeded in capturing depended on the migration rate of Neptune. A measurement of the relative resonance populations could thus provide an unanticipated method to calculate the migration rate of Neptune. This is a scientific question previously thought to be out of the realm of direct observations.

Scattered KBOs

The third dynamical class in the Kuiper Belt, the scattered KBOs, possesses large, eccentric orbits and is exemplified by the object 1996 TL_{66} (figure 2). Whereas the known classical KBOs are clustered inside 50 AU, with an eccentricity $e = 0.59$, and a semimajor axis $a = 85$ AU, 1996 TL_{66} reaches as far as 130 AU at aphelion. It is the only scattered KBO known thus far, but is most likely just the first detected of a much larger population of similar bodies. The scattered KBO population is estimated to contain ~10 000 objects like 1996 TL_{66}, with a total mass of ~0.05 M_{Earth} contained between ~40 and 200 AU (1 $M_{\mathrm{Earth}}=6\times10^{24}$ kg is the mass of the Earth). Although this estimate is crude, there is little doubt that the Kuiper Belt extends much further than suspected.

The origin of the scattered Kuiper Belt is unknown. One hypothesis posits that it is a natural product of the scattering process that produced the Oort cloud. The Oort cloud is believed to consist of planetesimals scattered by

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

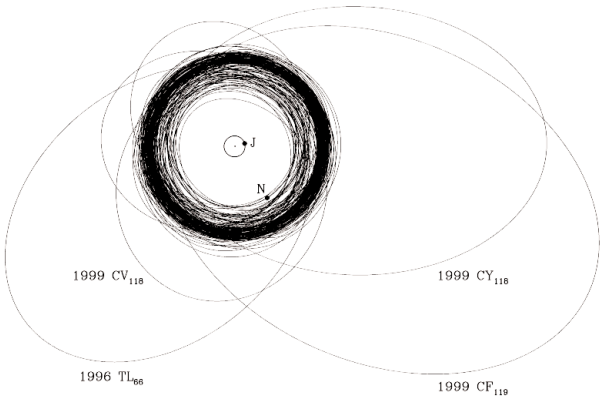


Figure 2. Planar view of the outer solar system. The dense band of orbits refers to the best observed KBOs (including Pluto), while the large, eccentric orbits of 1996 TL_{66}, 1999 CV_{118}, 1999 CY_{118} and 1999 CF_{119} extend far beyond the previously known Kuiper Belt, and identify them as scattered KBOs.

the giant planets (primarily Uranus and Neptune) when they nearly reached their full sizes. It is easy to imagine that perhaps not all the scattered planetesimals reached Oort cloud distances. Some perhaps were scattered to more intermediate distances, perhaps forming a radial distribution of material that bridges the gap from the planetary region to the Oort cloud. This scenario predicts that the outer solar system would be continuously populated, with the classical Kuiper Belt just beyond the planets, followed by the scattered Kuiper Belt which would extend all the way to the Oort cloud.

Population statistics

Thanks to ground-based surveys, the census of KBOs in the hundred kilometer range is now well established. The luminosity function of the Belt (inside the 30–50 AU region) can be obtained from the cumulative surface density measured by recent ground-based surveys. Fitting the surface density with the power law

$$\log[S(m_R)] = \alpha(m_R - m_0)$$

where m_0 is the (red) magnitude at which the surface density S reaches $1^{\circ-2}$, and 10α gives the slope of the luminosity function. Fitting the ground-based data yields α=0.58±0.05 and m_0=23.27±0.11. In other words, there is 1 KBO$^{\circ-2}$ at red magnitude m_R=23.3, increasing by ~4 per magnitude. The luminosity function outside $20<m_R<25$ may be different, and remains to be assessed.

The luminosity function provides a clue as to the size distribution in the Kuiper Belt, i.e. how the population is divided into different size ranges. Assuming that KBOs obey a differential power law size distribution of the form $n(r)\,dr : r^{-q}\,dr$, where $n(r)\,dr$ is the number of objects in the radius range from r to $\mathbb{r}+\,dr$, the ground-based observations are best fitted with $q = 4$. HUBBLE SPACE TELESCOPE observations suggest a larger q, but this result remains controversial.

Given the size distribution, it is possible to calculate the total mass provided that the minimum and maximum radii of the distribution are also known. This is not the case for the Kuiper Belt, and an exact determination of the belt's total mass has not been possible. The current assessment of the Kuiper Belt population can be summarized as follows. The total mass between 30 AU and 50 AU is estimated at ~$0.2M_{\rm Earth}$. There are no observations of the Kuiper Belt at larger distances, but the mass of the belt at larger distances has been roughly constrained by the dynamics of spacecraft and comets. If a massive belt existed at great distances, it would have affected the motion of spacecraft such as Pioneer 10 (currently at 67 AU), VOYAGER 1 (65 AU) and Voyager 2 (51 AU) as well as the motion of comet Halley. The estimated upper limit is $5M_{\rm Earth}$ within 100 AU—well above the mass estimated from observations of the scattered Kuiper Belt (see section on scattered KBOs).

This $0.2M_{\rm Earth}$ mass inside 50 AU includes ~70 000 objects larger than 100 km (in comparison, the asteroid belt contains ~230 asteroids larger than 100 km). Objects with radius <50 km are numerous but lie beyond the detection capabilities of most ground-based telescopes, while very large bodies (radius >250 km), although easily detectable from the ground, are rare. Note that the known KBOs are much larger than the known short-period comets (typical radius: 5–10 km), implying that the former cannot be the direct precursors of the latter. It has been postulated that the short-period comets were actually collision fragments that were then ejected from the Belt.

However, we must bear in mind that the Kuiper Belt has been much eroded dynamically since its formation, and the current population is but a remnant of the original population. From models that simulate the growth of KBOs, it is estimated that an original mass of ~10 $M_{\rm Earth}$ is needed to create the observed Kuiper Belt population.

Short-period comets and Centaurs

It is believed that the short-period comets originated in the Kuiper Belt, although it is not clear which part of the belt (the classical or scattered Kuiper Belt) is the main source. If comets originated as classical KBOs, these KBOs could have escaped from the belt in a slow 'leakage' via close encounters with Neptune. The KBOs would maintain their stable orbits inside the Kuiper Belt for 3–4 billion years before suffering a close encounter with Neptune. Once a Kuiper Belt object encounters Neptune, it has a 33% chance of being scattered into a short-period comet orbit. Other (less likely) alternative fates include being ejected from the solar system or colliding with the planets.

The timescale to go from a Kuiper Belt orbit to an inner solar system orbit is about 10^7 yr. Thus, if the Kuiper Belt is continuously supplying short-period comets, one would expect to see ex-KBOs making their way into the inner solar system. The prime candidates for these ex-KBOs are the 'CENTAURS', objects whose

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

orbits cross those of the giant planets. The Centaurs' orbits are highly chaotic and it is apparent that their residence between the giant planets is only temporary. The transient nature of the Centaurs, and their location between the Kuiper Belt and the inner solar system, renders them the best candidates for recent Kuiper Belt escapees that are perhaps now evolving into short-period comets. Seven Centaurs are known at the present time. Interestingly, the brightest Centaur, 2060 CHIRON, has been proven to be a comet, so there is no doubt that Centaurs are ice-rich bodies that can sublimate under the right conditions. The Centaurs are important because (1) they provide the missing dynamical link between the KBOs and the short-period comets and (2) they are much closer (brighter) than the typical KBOs, and therefore make much better targets for physical studies.

Physical properties of KBOs

The properties of KBOs are of intense interest since these bodies have changed relatively little since their time of formation in the solar nebula. Laboratory experiments predict that KBOs should have an 'irradiation mantle' produced by high-energy particle irradiation of surface ices. When carbon-containing ices (e.g. CH_4 or CO) are irradiated by the high-energy particles prevalent in interplanetary space such as cosmic rays and solar wind particles, this irradiation leads to the selective loss of hydrogen in surface materials, while encouraging the formation of carbon bonds. The irradiation thus alters the surface chemically and physically. Exactly how the color and reflectivity of the irradiated ice change with time depends on the specific composition and the irradiation flux. For example, cosmic rays are very energetic (10 MeV-10 GeV range) and can penetrate several meters, whereas solar wind particles (<1 keV) can only penetrate depths ~1 μm. The result is a dark, non-volatile solid residue that is rich in carbon compounds. After 4.5 billion years, KBOs and Centaurs are thus expected to sport a thick non-volatile, organic-rich crust, dark owing to its high carbon abundance. Note that this non-volatile crust would have a composition quite different from the interior, which is still pristine ice. The thick crust may survive on the object's first entry into the inner solar system.

Ideally, optical and infrared spectra are needed in order to assess the surface composition of KBOs and confirm or refute the irradiation mantle hypothesis. Unfortunately, physical properties of KBOs are extremely difficult to obtain because of their faintness. Until recently, instead of spectra, astrononomers have had to rely on optical and near-infrared broadband colors for compositional information. Surprisingly, KBOs and Centaurs exhibit a wide range of optical colors, ranging from nearly neutral (i.e. all wavelengths are reflected equally) to very red (red wavelengths are preferentially reflected), which is unmatched by known comets and asteroids. In fact, KBOs and Centaurs possess the most extreme red colors seen among small solar system bodies. The large spread of colors suggests considerable diversity in the materials present on the surfaces of KBOs. Interestingly, the very red optical spectrum of the Centaur 5145 Pholus has been reproduced with irradiated ices containing a mixture of N_2 and CH_4. The unique red colors suggest the presence of an indigenous surface material not seen closer to the Sun and are consistent with the irradiation mantle.

The irradiation hypothesis predicts that irradiation mantles should exist on the surfaces of all KBOs, thus resulting in similar colors for all KBOs. The observations clearly indicate otherwise. Astronomers have not conclusively identified the cause of the color diversity, but two simple explanations have been suggested for the origin of the color diversity. One possible explanation is that KBOs possess intrinsically different compositions, and that the different colors are tracers of the compositional variations. This has been observed in the asteroid belt, where asteroids show different compositions that seem to be related to their sites and temperatures of formation. However, the KBOs, as far as we know, all formed more or less where they are now seen, implying formation temperatures in the range 40–50 K. It is hard to see how strong compositional differences might arise if there is little change in the formation temperature, but they might occur, and this is a possible interpretation of the data. Another possible explanation is that collisions in the Kuiper Belt might puncture the irradiation mantles on some objects, revealing fresh interior material, as well as showering nearby regions with impact debris. The freshly exposed crater and the impact debris would be unirradiated material, and thus should be of a different composition and less red than the irradiated mantle. The overall color of a particular KBO would then depend on how much of the surface area is covered by recently excavated craters. This is analogous to the resurfacing of the Moon, where recent impacts have deposited bright rays on the darker background material. This collision resurfacing hypothesis predicts albedo variations on the surface of KBOs, and a test of this prediction may soon be feasible with the largest telescopes.

Besides this preliminary result on the surface composition, little information is available about other properties of KBOs. Attempts to measure their shapes and rotation periods have so far yielded inconclusive results.

Origin of the Kuiper Belt

From the apparent connection of the Kuiper Belt with so many solar system objects, it is clear that the belt played an important role in shaping the outer solar system. However, how did the belt form, and why did planetary accretion take the form of a large population of small bodies instead of a single large planet? The near-circular, low-inclination orbits of KBOs suggest *in situ* formation from the solar disk—this hypothesis has indeed been supported by dynamical simulations. Integrations of orbits beyond Neptune by Matt Holman, Jack Wisdom, Martin Duncan and Hal Levison showed that a belt of trans-Neptunian bodies formed at the beginning of the

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

solar system would survive for the age of the solar system. The implication is that the Kuiper Belt is simply a remnant fossil from the solar nebula that was left behind because it was not scattered away or incorporated into planets. All other solar system bodies (e.g. planets, asteroids, known comets) have been heavily modified by solar heating or collisions over the age of the solar system. Stranded far beyond the planets, little has happened to KBOs since birth. Although their top surface layers have been modified by irradiation and collisions, in this case collisions are actually beneficial to astronomers in that they reveal the pristine interior ice. In other words, in the Kuiper Belt is the solar nebula 4.5 billion years on, and with its discovery astronomers now have the chance to make direct observations of the nebula itself. This is the main scientific appeal of the Kuiper Belt.

Comparison with circumstellar disks

Are circumstellar disks extrasolar Kuiper Belts?

The recognition of our own Kuiper Belt leads us naturally to suspect that similar planetesimal structures might exist around other stars. It is now well known that dust disks are common around main-sequence stars, the most famous of which is the dust disk surrounding β Pic, suspected to be a young planetary system. Dust in disks like β Pic's is depleted by several destruction mechanisms such as radiation pressure, POYNTING–ROBERTSON drag and collisional destruction. If the dust is to be in steady state, it must be continuously replenished by some as yet unknown source. A simple and plausible source is collisional grinding between planetesimals in a Kuiper Belt.

The discovery of the Kuiper Belt proves that, indeed, planetary accretion at large distances is more likely to take the form of numerous small bodies rather than a single large planet. Recent discoveries of extrasolar planets suggest that planetary accretion is not an uncommon phenomenon, further amplifying our interest in extrasolar planetary systems. The obvious question then is whether we are observing an extrasolar Kuiper Belt in β Pic. Put in another way, does the Kuiper Belt also include a dust disk?

Kuiper Belt dust production rate

KBOs in the 100 km diameter range are relatively rare in the belt and are not likely to collide with each other. However, these large bodies are likely to collide with the much more numerous small bodies, e.g. those in the 1 km radius range. If we extrapolate the current KBO census down to 1 km radius range, there would be ~10^9–10^{10} of these small bodies. A crude estimate of the dust production rate through collisional grinding of these small projectiles would be 400–4000 kg s^{-1}.

This dust would be subjected to several destruction mechanisms. Grains with radius between 0.05 and 0.5 μm would be blown out of the solar system by radiation pressure very quickly. Larger grains would spiral toward the Sun under the effect of the Poynting-Robertson force and solar wind drag on the timescale of a few ×10^7 yr (some 20% of this dust should reach the orbit of the Earth, but would be difficult to distinguish from asteroidal dust). Finally, dust grains may also be destroyed by collisions with interstellar grains, or with other particles in the belt. Taking these destruction mechanisms into account, it would take on the order of 10^6 years to deplete the dust in the Kuiper Belt.

The non-detection of Kuiper Belt dust by the infrared satellite COBE placed an upper limit of $0.3M_{\rm Earth}$ on the amount of dust present in the Kuiper Belt. More stringent constraints for the amount of dust in the Kuiper Belt come from spacecraft that have traveled through the Kuiper Belt itself, such as Voyager 1 and 2 and PIONEER 10 (see the section on population statistics). Voyager 1 traveled north from the ecliptic, reaching an ecliptic latitude of +35°, while its twin Voyager 2 traveled southward and reached an ecliptic latitude of –34°. Both spacecraft detected numerous dust impacts that could only come from dust in the Kuiper Belt. Assuming that the Kuiper Belt is roughly 30° thick, and assuming an average dust grain size of 1 μm in radius, the Voyager results imply that there is at present ~$10^{-8}M_{\rm Earth}$ of dust in the Kuiper Belt. It is unlikely that such a small dust mass can be detected in the near future from any ground-based or near-Earth telescope. Another mass limit was obtained from Pioneer 10, which has spent more than a decade in the Kuiper Belt. Based on the fact that no dust impact damaged the spacecraft's propellant tank, scientists estimated that the Kuiper Belt could not contain more than $0.1M_{\rm Earth}$ in centimeter-size dust particles.

The dust mass in the β Pic disk is very uncertain but, roughly, the amount of dust in β Pic contained in micron-size particles is ~$3\times10^{-3}M_{\rm Earth}$, or 10 000 times more dust than is currently contained in our Kuiper Belt. However, as was pointed out earlier, if the original Kuiper Belt population is ~100 times larger than the current population, the dust content (which scales as the square of the number of bodies) must have been ~10 000 times larger. This suggests that the early Kuiper Belt (at the age of 100 million years or so) may have been much like the β Pic disk. Other indirect evidences exist which point toward this scenario. Recent HST images of the main sequence stars HD 141569 and HR 4796A clearly show the existence of dust rings (see URL http://oposite.stsci.edu/pubinfo/pr/1999/03/pr-photos.html). In particular, HR 4796A is surrounded by a ring 70 AU in radius, whose morphology is highly reminiscent of our Kuiper Belt. These observational and theoretical results support the hypothesis that many of the known disk systems can be identified with the Kuiper Belt, i.e., the Kuiper Belt is most likely a common phenomenon in stellar systems.

The future of the Kuiper Belt

What we have learned thus far about the Kuiper and its implications for the early solar system is only the tip of the iceberg. Although surveys have continuously identified new KBOs since 1992, studying KBOs is quite a different story from finding them. Until recently, there

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

was little astronomers could do besides trying to increase the sample of known KBOs. However, the imminent arrival of a new generation of large (8 m–10 m class) telescopes promises to galvanize the field by giving astronomers the capability to study these faint objects in detail. Kuiper Belt research is also given a great boost by new and on-going large sky surveys whose purpose is to find large, bright KBOs (red apparent magnitude m_R<22) suitable for detailed physical studies. For example, recent optical and near-infrared spectroscopic observations of the scattered KBO 1996 TL_{66} (the brightest KBO after Pluto and Charon) revealed a surface that is neutral in color and shows no sign of abundant ices as seen on Pluto. This rules out the possibility of a pure water ice surface on TL_{66}, although small amounts of other ices (including dirty water ice) cannot be ruled out. Obviously spectroscopic observations of many more KBOs are needed before astronomers can claim to understand the composition of the Kuiper Belt, but this goal is now in sight.

To properly interpret the spectroscopic observations in terms of the molecules and grains that make up the KBO surfaces, astronomers will also have to rely on laboratory experiments that simulate (as best as technology allows) the environment of the Kuiper Belt. For example, the experiments should simulate the irradiation and photoprocessing of a wide variety of ices by ultraviolet photons. Similar experiments have been performed in the past, but the focus in the past had been on INTERSTELLAR GRAINS, whose source of irradiation is mainly high-energy cosmic ray particles. Such experiments will have to be modified to approximate the Kuiper Belt environment. The results can then be compared with the extensive body of knowledge on interstellar grains, giving astronomers the unprecedented opportunity to observe the evolution of interstellar material as it is incorporated into a solar nebula. In short, Kuiper Belt research is in its nascent stage and its demands will push astronomical technology to the limit, but the potential reward is the answer to a question that has driven astronomers since the days of Copernicus: how the solar system formed (see also ORIGIN OF THE SOLAR SYSTEM).

Bibliography

Chiang E and Brown M 1999 *Astronomical Journa*l **118** 1411

Dones L 1997 *From Stardust to Planetesimals (Astronomical Society of the Pacific Conference Series vol 122)* ed Y J Pendleton and A G G M Tielens (San Francisco: Astronomical Society of the Pacific) pp 347–65

Jewitt D 1999 *Annual Review of Earth and Planetary Science* **27** 287

Jewitt D, Luu J and Trujillo C 1998 *Astronomical Journal* **115** 2125

Levison H and Duncan M 1997 *Icarus* **127** 13

Luu J X and Jewitt D C 1996 *Scientific American* **274** 32

Luu J X and Jewitt D C 1996 *Astronomical Journal* **112** 2310

Luu J, Marsden B, Jewitt D, Trujillo C, Hergenrother C, Chen J and Offutt W 1997 *Nature* **387** 573

Jane Luu

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Kuiper, Gerard Peter (1905–73)

First applicant of multidisciplinary techniques to the study of the solar system, i.e. the first planetary scientist, born in Harencarspel, Netherlands, educated at Leiden, and moved to the USA, becoming director of Yerkes Observatory and founding (1960) and directing the Lunar and Planetary Laboratory of the University of Arizona. Spectroscopically detected the methane atmosphere on Saturn's satellite Titan and the carbon dioxide atmosphere of Mars. Discovered Miranda (fifth moon of Uranus), and Nereid (second moon of Neptune). Identified the belt of comet-like debris at edge of solar system (*Kuiper's belt*; *see also* KENNETH EDGEWORTH). Selected the landing sites on the Moon for the Ranger probes and Apollo landings. The *Kuiper Airborne Observatory* for infrared astronomy (used extensively for remote sensing of the planets) was named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kulik, Leonid (1883–1942)

Russian meteorite researcher, trained as a forester. Led expeditions to investigate the Tunguska impact, published the accounts of flattened trees and small craters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kuo Shou-Ching (1231–1316)

Chinese astronomer, constructed at Linfen (Shansi) the 'Simplified Instrument', a torquetum to observe the positions of the stars. It was so-called because it was equatorially mounted, and was thus able to omit some rotations and calculations which followed from the original Arab design, altazimuthally mounted. The instrument survives at Purple Mountain Observatory, having been moved to Nanjing in the Ming dynasty, when it had been forgotten that the change in latitude rendered it useless. According to JOSEPH NEEDHAM, the idea of this instrument was transmitted back through Arabic astronomy, via Gemma Frisius to Tycho Brahe, and influenced his instruments and concept of equatorial astronomy. Shou-Ching developed spherical trigonometry, as well as making a clockwork-driven celestial globe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lépaute, Nicole-Reine Etable, née de la Briére (1723–88)

French astronomer, aide to JOSEPH LALANDE calculating circumstances of an annular eclipse of the Sun visible in France and Europe. She worked on the predictions for the return of HALLEY's Comet in 1759, including the perturbations by Jupiter and Saturn. She also assisted her husband, the royal clockmaker of France, making calculations such as the number of oscillations per unit time of pendulums of various lengths.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

La Condamine, Charles Marie de (1701–74)

Born in Paris, France, had an army career, was sent by the Académie Royale to Peru to measure the length of a degree of meridian at the equator, repeated (at Cayenne in Guyana) RICHER's experiments on the variation of weights at different latitudes. Although a less gifted astronomer than Godin and a less reliable mathematician than BOUGUER (both his colleagues), he often received the major part of the credit for the expedition, because of his amiable nature and his talent as a writer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

La Hire, Philippe de [Philippe I] (1640–1718) and La Hire, Gabriel-Philippe de [Philippe II] (1677–19)

Astronomer, Philippe I was born in Paris, France, educated as an artist, taught mathematics and architecture, wrote on conic sections. He installed the first transit instrument in the Paris Observatory and produced tables giving the movements of the Sun, Moon and the planets. He conducted experiments in terrestrial magnetism and meteorology, surveyed the French coastline (with PICARD) and participated with his son in the geodesic measurements carried out by JACQUES CASSINI to extend the meridian of Paris north from Amiens to Dunkerque. His son, Philippe II, a dilettante and astronomer, born in Paris, lived and was educated at the Paris Observatory, assisting his father in observations. Father and son were accused of plagiarism by LE FEVRE after publishing ephemerides to which Le Fevre claimed authorship. Philippe II published on observational and physical astronomy but also dabbled in architecture, inventions and other natural sciences.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

La Plata Observatory

The most important event in the history of astronomy in Argentina 10 years after the foundation of the OBSERVATORIO ASTRONÓMICO DE CÓRDOBA (1870), was the setting up of the Observatorio de La Plata (LPO) (see figures 1 and 2), officially founded on 22 November 1883. Its first Director was Navy Lieutenant Francisco Beuf, of French origin. Indeed, the LPO's existence was set off by an invitation from the Observatory of Paris to prepare a French–Argentine expedition to Bragado, province of Buenos Aires, to observe the transit of Venus on 6 December 1882.

This expedition was equipped with a 21.6 cm telescope, chronometers, a portable meridian circle and meteorological instrumentation (incorporated soon afterwards into the Observatory). The LPO's observational power was later increased with two transit circles (1886), a 10.5 cm altazimuth telescope, a 83 cm reflector (1890), a 33 cm astrograph (1890), a 43 cm refractor (1894), a Repsold vertical circle, two Repsold transit instruments and a Zeiss comet finder. In 1887, LPO was incorporated to the 'Carte du Ciel' project (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS) to observe the zone of the sky from –23° to –32° through a 33 cm standard astrograph which arrived in La Plata in 1890.

In 1905, LPO became part of the National University of La Plata and its specific objectives widened to cover meteorological, seismic and magnetism studies.

In 1911, Dr William J. Hussey, former director of the Ann Arbor Observatory in Michigan, became Director of LPO. He systematized research and incorporated LPO into the worldwide project of the Astronomisches Gesellschaft (1863) (see ASTRONOMICAL SOCIETIES AND PUBLICATIONS) of surveying all stars down to magnitude 9 from –52° to –82° and, later, from –42° to –82°. It took 25 years to complete this survey (positions of 33 300 stars in six catalogues, A to F). Under Hussey's directorship extrameridian astronomy began. F. Aguilar, an engineer who equipped the Observatory with a spectrograph, took over afterwards. Aguilar was then replaced by the German astronomer JOHANNES HARTMANN in 1920 who continued Hussey's work, developing a theory on novae and discovering that the minor planet Eros is a non-spherical rotating body. In 1927 Dr B. H. Dawson, a Michigan graduate, started observing visual double stars and, on 8 November 1942, he discovered Vela's nova.

Figure 1. The French-style main building of La Plata Observatory.

Figure 2. Reading room of La Plata Observatory library.

During Aguilar's second period as director (1934–1943), Dr A. Wilkens (who started celestial mechanics studies) and H. Wilkens (stellar statistics) — both German, Dr R. Cesco and Dr E. Gaviola joined LPO staff. Dr A. Wilkens and R. Platzeck together with J. Sahade and C. Cesco undertook the measurement of stellar temperatures in double systems. Dr Livio Gratton and Dr P. Sconzo — both Italian — also joined the staff. The former led research on binary, Cepheid and giant K-type stars, while the latter started an observational program to refine the orbits of asteroids, many of them having been discovered then. In 1935, LPO offered PhD degrees in astronomy for the first time in the country.

By the 1950s Dr Jorge Sahade (IAU president 1985–1988) recommended the acquisition of a 2.15 m telescope like that to be installed in Kitt Peak by the ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY. Surprisingly enough, the telescope was acquired but it saw first light in 1986 when, through an agreement among the Universities of La Plata, Córdoba and San Juan and the Research National Council (CONICET), it was set up in Calingasta Valley (province of San Juan), becoming a national observing facility: the Leoncito Astronomical Complex, CASLEO (see SOUTH AMERICAN ASTRONOMY).

Binary statistics and systematic spectroscopy in open clusters started by the end of the 1950s, conducted by Dr C. Jaschek (later on, Director of STRASBOURG OBSERVATORY and the CDS). By the 1960s the construction of a

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

photometer allowed Dr A. Feinstein to perform photometric observations on southern open clusters.

In 1971 LPO installed a Danjon astrolabe in Río Grande, Tierra del Fuego, through an agreement between the University of La Plata, CONICET, Tierra del Fuego Government and the Navy Hydrology Service. Again via agreement with the Navy Hydrology Service and the Navy Observatory of Washington, a zenithal photographic tube was also installed in the Navy Base of Punta Indio, province of Buenos Aires. LPO discontinued astronomical observations in the 1970s because of city-light pollution. Since then, observational facilities have been provided by Chilean observatories and CASLEO.

Today — as ever since 1905 — research and the teaching of astronomy are the LPO's main objectives. It plays a leading role in the Argentine part of the GEMINI project and most of its members actively collaborate with researchers all over the world. Over 55 staff members are working on a variety of topics including galactic dynamics, celestial mechanics, stellar evolution, radioastronomy and extragalactic reference systems. Recently, a vigorous GPS development has taken place. Funds for research are mainly provided by official agencies such as CONICET.

Ruben Vazquez

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Laboratory and Space Measurements of the Effects of Gravity

Early philosophers such as Aristotle tried to describe the motion of bodies in the Earth's vicinity with concepts that confused the topics of dynamics (the motion of a body under any external force) with the nature of the force (gravitation) itself. The teaching that followed this doctrine lasted until the mid 1600s when difficulties were becoming apparent in explaining, for example, the motion of arrows whose velocity continued after the force that caused their motion was removed. GALILEO GALILEI [1] began experiments in Pisa to investigate the dynamics of material bodies and the properties of motion under what is now known to be the gravitational force. Galileo was successful in showing that much of Aristotle's dogma was contrary to carefully observed experiments. This work inspired NEWTON [2] to apply the mathematics of calculus to problems of celestial mechanics such as the apparent motion of comets. Newton was able to separate the laws of dynamics from those of gravitation, both of which he was able to describe in simple mathematics with high precision.

The dynamics of a particle of mass M, moving under an external force, F, was described by the vector equation for the acceleration A as

$$\boldsymbol{F} = M\boldsymbol{A}.$$

It was Newton's conjecture that the acceleration, and not the velocity as the Aristotelians had taught, was proportional to the force. This provided the key to understanding dynamics.

Newtonian gravity was encapsulated in the expression for the gravitational force of attraction, F, between two masses, M_1, M_2 separated by a distance r:

$$F = \frac{GM_1M_2}{r^2}.$$

The constant G was determined experimentally by Cavendish [3] and others [4] to be (in modern SI units) 6.67×10^{-11} N m^2 kg^{-2}.

Both Newton and Galileo were aware that if a body were moving under the influence of a gravitational force (rather than any other kind of force, such as electromagnetism for example) then the mass of that body would appear twice in the calculation of the resulting acceleration since the gravitational force is directly proportional to the mass (the 'gravitational mass'), and the acceleration is inversely proportional to the mass (the 'inertial' mass). If every object had the property that its 'gravitational mass' is exactly proportional to its 'inertial mass', then these two values for the mass would cancel and all objects would have the same acceleration, and therefore the same velocity, in an identical gravitational field. Galileo observed this to be true in a number of experiments and Newton systematically explored the equivalence of gravitational and inertial mass for different materials in a series of experiments with pendula of different composition. There was (and is now) no explanation in Newtonian theory of why there are these two separate roles for mass, and several experiments of exquisite precision were carried out in the late 1800s by Eötvös and others to see if the equivalence broke down at some level of precision.

The mathematical statement of the 'Law' of gravitation permitted the accurate calculation of planetary positions, which observations confirmed to the highest precision available. The first deficiencies in this approach began to emerge in 1845 [5] with observations of the slightly elliptical orbit of Mercury, the planet closest to the Sun and therefore the object orbiting in the highest gravitational potential of all the planets. As measurement techniques became more accurate and the baseline for the observation of the secular change in Mercury's orbit became longer and the orbit therefore determined more precisely, the discrepancy became increasingly uncomfortable for Newtonian gravity, despite the fact that the simple inverse square law provides extremely accurate descriptions of most phenomena.

Following the success of EINSTEIN's 'Special Theory of Relativity' [6] in explaining the null result of the Michelson–Morley experiment to measure the effect of the Earth's supposed velocity through an Aether, the next step was to incorporate gravitation within a relativistic theory. This Einstein did with his 'General Theory' in 1916 [7]. At low velocities and small gravitational potentials the new theory had to be consistent with Newtonian gravity in predicting the motion of objects. Rather than leave the equivalence principle as an uncomfortable experimental fact, Einstein raised it to be the underlying concept which illuminated the radical thinking behind his new theory. He proposed that the gravitation was to be understood as a curvature of spacetime created by the presence of mass. Matter moved in this curved SPACETIME in response to the local curvature. Every object in a particular gravitational field experienced the same curvature and hence all objects moved with the same acceleration. The equivalence of inertial and gravitational mass is therefore implicit in the structure of general relativity, and the 'equivalence principle' needs no explanation or resolution.

General relativity cannot yet be understood as the final theory of gravity (see also GENERAL RELATIVITY AND GRAVITATION). The theory has had remarkable success in naturally accounting for the precession of the perihelion of Mercury and predicting the displacement of stellar images close to the Sun but the accuracy to which it has been tested is only a few parts in ten thousand. The most unsatisfactory aspect of the theory is the complete incompatibility with the, now accurately tested, theory of QUANTUM MECHANICS. It is this unresolved problem that is motivating both theorists and experimentalists to test the theoretical predictions of general relativity to higher and higher precision in the search for clues as to where progress can best be made in unifying gravitation with the other fundamental forces of nature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

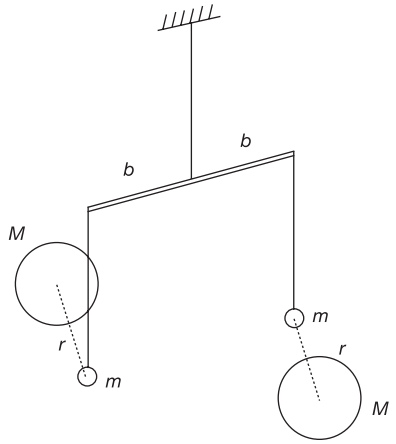


Figure 1. A torsion balance.

Expressed in terms of Newtonian theory, laboratory tests focus on the accuracy of the inverse square law, the value of the gravitational constant, G, and its time derivative, together with evidence that the equivalence principle breaks down at some level of accuracy or for some sets of materials. The dependence of the gravitational force on distance, r raised to the power -2, is appropriate for a truly long-range force, and this implies that any particle postulated as mediating the force (the 'graviton') must necessarily have zero mass, exactly similar to the case of the photon as the mediator of the electromagnetic force. Tests of the accuracy of the inverse square law therefore provide limits on the mass of the graviton. Since general relativity uses the concept of curved spacetime with metric tensors as the mathematical interface between matter and curvature, an expansion of the metric tensor can be made in terms of parameters which express the small deviations from the special relativistic metric in flat spacetime. This Parameterized Post Newtonian (PPN) metric can be used to analyze experiments whose outcome thereby places limits on the various PPN [8] parameters.

Laboratory tests

The gravitational force is extremely weak, nearly forty factors of ten weaker than the electromagnetic force. As a result the gravitational constant is very difficult to measure to high precision. Other factors, such as the inability to shield stray gravitational fields from test masses, make the design and execution of laboratory experiments extremely taxing. Astronomical measurements provide high precision in terms of the dynamics of orbits, but the quantity determined is actually the product of the gravitational constant and the mass of the central object, a number that cannot be determined independently. Most accurate determinations of G use a torsion balance [9] as shown in figure 1.

A fibre with a known torsional constant suspends a balanced pair of test masses. Two methods of executing the measurement are used, the force measurement and the period measurement. In the former, the rotational position of the beam is measured with no source masses close to the test masses which are suspended from the balance. The source masses are moved into the required position with respect to the test masses. As a result of the gravitational force the beam rotates about the torsional axis. A known external force is then used to rotate the beam back to its original position. Accurate measurements of the source masses, the test masses and their separation then permits the determination of G in terms of absolute units. The period method requires the determination of the period of torsional oscillation of the beam both when the source masses are in position and when they are absent. This method of determining G has the advantage that various parameters which are difficult to measure accurately, cancel out in comparing the two periods. This method is not so sensitive to drifts of the apparatus, and time measurements are always more accurate than measurements of angular position because nonlinearity in the detector has much less effect. The difficulties are that the long measurement times place the experiment at risk of substantial environmental changes during the data taking and the nonlinear effects of small oscillations can affect the determination of the oscillation maxima. The outcome of these measurements over a period of 150–200 years is that the constant of gravitation is known only to an accuracy of a few parts in 10 000: a very poor level of accuracy compared with other fundamental physical constants. Some part of this inaccuracy lies in the difficulty of determining in an absolute sense the mass and geometry of the test masses used as the source of the gravitational force. If the unification of physical forces occurs at the theoretical level, it may well be important to have accurate values for these fundamental constants to test the validity of any proposed unification scheme, and this motivates efforts to measure G.

Space and ground tests of general relativity and the equivalence principle

A direct prediction of general relativity is that clocks will be affected by the gravitational potential in which they are located, as measured by observers located at infinite distance from any gravitating masses. The first demonstration of this effect was published by Pound and Rebka [10, 11], using the Mössbauer effect to generate a very narrow absorption line as the detector of the frequency shift. The resulting reduction in frequency (redshift) of the ^{57}Fe gamma rays from a source at the top of a tower some 20 metres high was exactly in accordance with Einstein's prediction. This experiment is now interpreted as a demonstration of the equivalence principle, to a reasonably high degree (about 1%), though not to the accuracy already achieved by Eötvös and colleagues. In 1979 Vessot and Levine [12] repeated the gravitational redshift experiment in space using a hydrogen maser on a Scout rocket. The result again

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

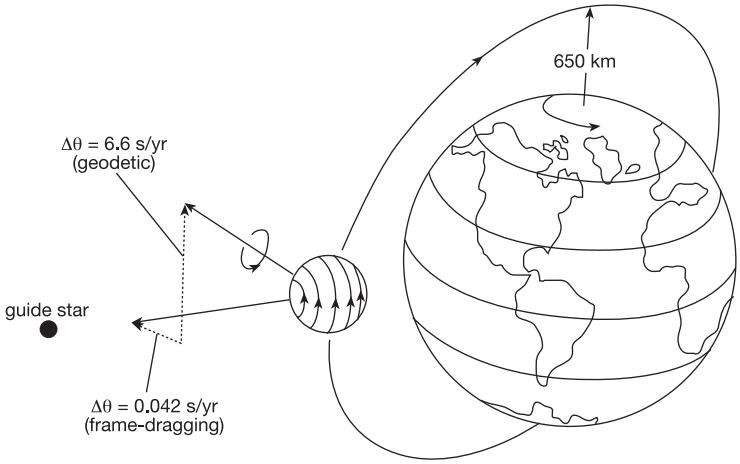


Figure 2. Schematic representation of the 'laboratory test' to be performed by Gravity Probe B.

agreed with the predictions of Einstein's general theory of relativity to about one part in 10^4.

Since the equivalence principle lies at the very foundations of general relativity, various schemes have been devised to extend the accuracy beyond that possible on the ground. Worden [13] has proposed an experiment which, in principle, repeats Galileo's famous Tower of Pisa comparison of the rate of fall of masses of different composition. The generic name for the experiment is STEP, the Satellite Test of the Equivalence Principle. The two test masses are made as concentric cylinders which can slide along, without touching, a linear superconducting bearing. The axis of the bearing is maintained in the plane of the orbit and fixed in inertial space. When the axis is pointing towards the Earth the two masses are attracted in that direction and the experimental apparatus can measure the relative motion of the masses to about one part in 10^{18}. Space provides an excellent environment for such precision experiments. The high vacuum and cryogenic requirements can be met with modern technology, and the absence of interfering gravity gradient signals is extremely important.

A number of subtle effects were predicted by the general theory of relativity [14] and these include departures of spinning particles from simple motions, and more complex behaviour of spinning particles orbiting spinning masses. Beyond description by Newtonian dynamics, these effects are related to the properties of 'parallel transport' in a curved space and the dragging of the local spacetime coordinate system by the spinning central mass. Observation of such motions would test much more of the detailed mathematical structure of general relativity than is possible simply from the motion of non-spinning test particles. The precision required to test general relativity by searching for, and measuring, these non-Newtonian effects has been beyond the available technology until the 1980s. The development of cryogenics, low noise electronics, quantum detectors and space platforms has now brought general relativity within the grasp of experimental physics although the techniques required are some of the most challenging uses of modern technology. The GRAVITY PROBE B experiment [15] sets out to perform the first 'laboratory test' of the general relativistic equations for spinning particles and is shown diagramatically in figure 2 (see also GENERAL RELATIVITY: SOLAR SYSTEM EXPERIMENTS).

This is due for launch in 2001 on a NASA rocket which will place the cryogenic payload in a polar orbit at an altitude above the Earth of 650 km. In the experimental package a spinning rotor is used as the test particle. Manufactured with exquisite precision to avoid disturbing forces created by any imperfections, the rotor is set spinning in such a way that the angle between its rotational axis and the fixed stars can be measured as the experiment proceeds. According to general relativity the spin axis should move slowly with respect to the coordinate frame defined by the stars due to two effects. The Fermi–Walker transport of the spin vector in space curved by the presence of the Earth will cause a drift of the spin axis by 6.6 arcseconds per year in the plane of the orbit. In addition to this, the spin of the Earth drags the coordinate frame around in such a way that there is a motion perpendicular to the orbit of 0.042 arcseconds per year, and the experiment is planned to measure this latter effect to an accuracy of about 1%.

Conclusions

By extending the laboratory into space, tests of our understanding of the gravitational force can be made with a precision which should, eventually, point the way to a resolution of the incompatibility of general relativity and quantum mechanics.

Bibliography

[1] Ronan C A 1974 *Galileo* (London: Weidenfeld and Nicholson)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

[2] Newton I 1686 *Philosophiae Naturalis Principia Mathematica* (London)
[3] Cavendish H 1798 Experiments to determine the density of the Earth *Phil. Trans. R. Soc.* **88** 469–526
[4] Poynting J H 1891 On the determination of the mean density of the Earth and the gravitational constant by means of a common balance *Phil. Trans. R. Soc.* A **182** 565–656
[5] Le Verrier U J J 1859 *C. R. Acad. Sci., Paris* **49** 379
[6] Einstein A 1905 *Ann. Phys., Lpz.* **17** 891
[7] Einstein A 1916 *Ann. Phys., Lpz.* **49** 769
[8] Will C M 1993 *Theory and Experiment in Gravitational Physics* (Cambridge: Cambridge University Press)
[9] Chen Y T and Cook A 1993 *Gravitational Experiments in the Laboratory* (Cambridge: Cambridge University Press)
[10] Pound R V and Rebka G A 1960 Apparent weight of photons *Phys. Rev. Lett.* **4** 337–41
[11] Pound R V, Rebka G A and Snider J L 1960 Effects of gravity on gamma radiation *Phys. Rev.* B **140** 778–803
[12] Vessot R F and Levine M W 1979 A test of the equivalence principle using a space borne clock *J. Gen. Rel. Grav.* **10** 181–204
[13] Worden P W 1978 Equivalence principle test in Earth orbit *Acta Astronautica* **5** 27–42
[14] Schiff L I 1960 Possible new experimental tests of general relativity *Phys. Rev. Lett.* **4** 215
[15] Turneaure J P *et al* 1986 The Gravity Probe B relativity gyroscope experiment: approach to a flight mission *Proc. 4th Marcel Grossman Meeting on General Relativity* ed R Ruffinin (Elsevier) pp 411–64

The following references provide more general reading:

Pais A 1982 *Subtle is the Lord: The Science and Life of Albert Einstein* (Oxford: Oxford University Press)
Kenyon I R 1990 *General Relativity* (Oxford: Oxford University Press)
Schutz B F 1985 *A First Course in General Relativity* (Cambridge: Cambridge University Press)

A M Cruise

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lacaille 9352

A red dwarf star (spectral type M2–M3V) of apparent magnitude 7.35, situated in the constellation Piscis Austrinus, close to the border with Sculptor and about 6° south of Fomalhaut. It is the ninth closest star, at a distance of 10.7 light-years, parallax 0.304″; its absolute magnitude is 9.8. It is especially significant for having the fourth largest known proper motion, 6.896″ per annum. Stars of such high proper motion are of great value in studies of stellar distribution, and especially of star streaming and galactic rotation. Unfortunately they are very few—only 45 star systems are known with annual proper motions greater than 2″, all within a radius of 96 light-years of the Sun. They range in apparent magnitude from −0.05 to 13.31; only 14 are visible with the unaided eye. They are reasonably well distributed around the sky, but this is a regrettably small number of points of reference on which to construct a model of stellar motions. Thus apparently insignificant stars such as Lacaille 9352 play a major role in astronomical research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lacaille, Abbé Nicholas [Nicolas] Louis de (1713–62)

Astronomer, born in Rumigny, France. From 1750 to 1754 he led an expedition to the Cape of Good Hope, where he was the first to measure the curvature of the Earth in South Africa, and, with a small telescope, compiled a still-used catalogue of nearly 10 000 southern stars. In the course of his survey he discovered and listed 50 nebulous objects (eight are now regarded as non-existent). He classified his objects in three categories, foreshadowing the modern astrophysical interpretation: Lacaille I: nebulae; Lacaille II: nebulous star clusters; Lacaille III; nebulous stars. In his *Atlas*, he named 15 southern constellations unseen by classical astronomers: Antlia, Caelum, Circinus, Fornax, Horologium, Mensa, Microscopium, Norma, Octans, Pictor, Pyxis, Reticulum, Sculptor, Telescopium, and renamed the constellation Musca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lacerta

(the Lizard; abbrev. Lac, gen. Lacertae; area 201 sq. deg.) A northern constellation which lies between Cygnus and Andromeda, and culminates at midnight in late August. It was introduced by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* of 1687. Hevelius also gave it the alternative name of Stellio (the Stellion—a newt with star-like dorsal spots found along the Mediterranean coast), but this was seldom used. A few years earlier, in 1679, the French architect and celestial cartographer Augustin Royer had depicted a new constellation Sceptrum (the Sceptre—representing the French sceptre and hand of justice, in honor of Louis XIV), which culminated directly overhead as seen from Paris, in approximately the same position, but this was not widely accepted and was soon replaced by Hevelius' Lacerta.

A small, inconspicuous constellation, the brightest stars in Lacerta are α Lacertae, magnitude 3.8, and 1 Lacertae, magnitude 4.1. The Milky Way cuts across the northern part of Lacerta and the constellation contains several open star clusters, including NGC 7243, which consists of several dozen stars of magnitude 8.5 or fainter. Also in Lacerta is BL Lacertae, the variable nucleus (range 12.4–17.2) of a distant elliptical galaxy and the type object of its class.

See also: BL Lacertae objects.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lada Terra

One of the three large upland regions of Venus (the other two being Aphrodite Terra and Ishtar Terra), situated in the planet's southern hemisphere and centered at 54.4 °S, 17.5 °W. In its greatest dimension it measures 8614 km. It takes its name from a Slavic goddess of love. Like the other upland regions, Lada Terra is covered with tesserae: criss-crossing troughs and ridges. Lada is less elevated than the other main upland regions, and though extensive is less clearly delineated. It lacks the prominent peaks and rifts that characterize the others.

See also: Venus: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lagoon Nebula (M8, NGC 6523)

An emission nebula in the constellation Sagittarius, position RA 18^{h} 03.8^{m}, dec. -24° $23'$. At sixth magnitude and measuring $90'$ by $40'$ it is one of the most prominent nebulae in the sky. A bright feature in the central region is known from its shape as the Hourglass Nebula. It contains the star cluster NGC 6530, and is illuminated by a pair of blue giant stars, 7 and 9 Sagittarii. The lanes of dark nebulosity that cross the Lagoon are what give it its name.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lagrange, Joseph-Louis (1736–13)

Mathematician, born in Giuseppe Lodovico Lagrangia in Turin, Sardinia-Piedmont (now Italy), in a family of modest means. As he said: 'If I had been rich, I probably would not have devoted myself to mathematics'. Contrast with GERHARD HERZBERG. As a young man, he published work on the calculus of variations, developed a theory of dynamics based on the principle of least action, fluid mechanics (where he introduced the *Lagrangian function*) and the orbits of the Moon, Jupiter, its moons and Saturn. A protégé of EULER, he succeeded him as director of mathematics at the Berlin Academy of Science, and worked with him on the three-body problem, the motion of the Moon, perturbations of the orbits of comets by the planets, the stability of the solar system and other mathematical problems. At the age of 51 he moved to Paris, where he was acclaimed as a French mathematician through the publication of *Mécanique Analytique*. This work summarized all the work done in the field of mechanics since the time of NEWTON, using differential equations to transform mechanics into the branch of mathematical analysis that it is today. He was threatened by the events of the Reign of Terror, being vulnerable as a foreigner, but was saved by the intervention of Lavoisier, who himself came under suspicion and was guillotined.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lagrangian Points

Five neutral points (points at which an object experiences no net gravitational force), in the combined gravitational field of two massive bodies which are orbiting their center of mass, at which an object of much smaller mass can exist in equilibrium. They are so named because Joseph-Louis Lagrange in 1772 was the first to find them as solutions to a restricted case of the three-body problem. Three of the points lie on a line joining the centers of the two massive bodies, M_1 and M_2, assumed for simplicity to be in circular orbits. The *inner Lagrangian point*, denoted by L_1, lies between M_1 and M_2, while L_2 and L_3, the *outer Lagrangian points*, lie on either side. The other two Lagrangian points, L_4 and L_5, are in the orbit of M_2 about M_1, respectively 60° ahead of and behind M_2 in its orbit. L_4 and L_5 therefore form equilateral triangles with M_1 and M_2. A body at L_1, L_2 or L_3 is in unstable equilibrium, very prone to gravitational perturbations, but a body at L_4 or L_5 is less prone to perturbations and has long-term orbital stability.

In practice, the situation is complicated by the ellipticity of real orbits and the presence of gravitational perturbations. In the solar system the best example of bodies in stable orbits at Lagrangian points are the Trojan asteroids at the L_4 and L_5 points of Jupiter's orbit around the Sun. Perturbations by the other planets, notably Saturn, make each Trojan oscillate back and forth between points about 45° and 80° ahead of or behind Jupiter. One Martian Trojan is known, orbiting near Mars's L_5 point. A *halo orbit* is an orbit around a Lagrangian point. The Solar and Heliospheric Observatory (SOHO) was placed in a halo orbit around the Earth's inner Lagrangian point, L_1, between the Earth and the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lalande, Joseph-Jérôme Lefrançais de (1732–1807)

Astronomer, born in Bourg-en-Bresse, France. In what was one of the first internationally coordinated scientific campaigns, to determine the Moon's parallax, he was sent in 1752 to Berlin by the French Académie des Sciences. Other astronomers took up positions at six other sites located on more or less the same meridian from Stockholm to the Cape (where Lalande's colleague LACAILLE was located). They observed the Moon (and indeed Mars) simultaneously against its background of stars, and, knowing their mutual distance (in kilometers), they were able to use the method of triangulation in order to determine the Moon's distance. He edited *La Connaissance des Temps* the French almanac, and became professor of astronomy in the Collége de France, and later director of the Paris Observatory. His chief work is *Traité d'Astronomie* (1764), and he also produced a comprehensive star catalog, recording the planet Neptune on two occasions before its discovery without recognizing that it was not a star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lambert, Johann Heinrich (1728–77)

Mathematical physicist, born in Mulhouse, Alsace, France, proved that pi is irrational, devised techniques for photometry (the *lambert*, the unit of light intensity, is named after him). He wrote on non-Euclidean geometry, the theory of cartography, and perspective in art. *Lambert's problem* in orbital dynamics is: given a planet A at a given time and planet B at a later time, with what velocity to launch a projectile to get from A to B by orbiting under the influence of the Sun? This is the fundamental question of interplanetary space travel. When asked by King Frederick II in which science he was most proficient, Lambert claimed: 'All'!

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lamont, Johann von (1805–79)

Astronomer, born in Braemar, Aberdeenshire, Scotland, took German nationality, and became professor of astronomy at Munich. Determined the orbits of moons of Saturn and Uranus, worked on terrestrial magnetism, writing a handbook on the subject (1849); discovered that the magnetic field of the Earth fluctuates in a period of 10 years. Von Lamont recorded Neptune on at least three occasions over a four-day period just before its discovery by Galle, and perhaps should have recognized its motion, but did not.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Langren [Langrenus], Michael Florent van [Michael Florentius] (c. 1600–75)

'Spherographer', i.e. astronomer, cartographer, geographer and engineer, born in Amsterdam (probably), into a prominent Belgian family of map- and globe-makers. Tried to determine longitude by predicting phenomena on the Moon (such as the illumination and shadowing of lunar mountains) that could be observed from all points of the Earth and used as clocks. In the course of this work, he prepared lunar maps, including views of the Moon on a daily basis in its cycle. In the hope of patronage, he named the lunar features with the names of prominent figures, such as Pope Innocent X, and royalty. The Infanta Isabelle Claire Eugenie appears three times: a mountain Eugeniae (in Mare Belgicum), Mare Eugenianum, and a mountain Isabellae (in Mare Eugenianum). Became 'Mathematician and Cosmographer' to King Philip IV of Spain.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lansberge, Philip van [Lansbergen, Philips] (1561–1632)

Minister, born in Ghent, Belgium, publishing books on mathematics and astronomy, in which he supported Copernican theories. However, he could not believe in non-circular planetary motions and did not accept KEPLER's theory of elliptical planetary motion, publishing astronomical tables which he hoped (in vain) would support COPERNICUS over Kepler.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Laplace, Pierre-Simon (1749–1827)

Celestial mechanician, born in Beaumont-en-Auge, Normandy, France, became professor of mathematics at the Ecole Militaire in Paris, examining the cadet Napoleon Bonaparte. This position made Laplace well known to people in positions of power, which he opportunistically exploited, becoming, under Napoleon, Minister of the Interior (Napoleon soon removed him from office 'because he brought the spirit of the infinitely small into the government'.) He researched probability theory, the calculus, and celestial mechanics. His work was interrupted by the Reign of Terror. Even though, wisely, self-exiled from Paris, Laplace, as a member of the metrication committee of the Académie des Sciences was consulted about rationalization of the calendar and metrication of angles and did not demur with the Revolutionary proposals, although he knew they were impractical and even incorrect. As more stable times returned, Laplace became head of the Paris Observatory, and was criticized for 'neglecting all the observations except those needed for his formulas'. He presented the *Laplace nebular hypothesis* in *Exposition du Systeme du Monde* (1796), that viewed the solar system as originating from a large, flattened, slowly rotating cloud of hot gas. He also expressed another remarkably modern view, of the impact of comets on the Earth: 'The small probability of collision of the Earth and a comet can become very great in adding over a long sequence of centuries. It is easy to picture the effects of this impact on the Earth. The axis and the motion of rotation have changed, the seas abandoning their old position. . . , a large part of men and animals drowned in this universal deluge, or destroyed by the violent tremor imparted to the terrestrial globe'. The *Exposition* was an introduction for Laplace's most important, five-volume work *Traité du Mécanique Céleste*, which set up the differential equations of dynamics, solving them to describe the motion, including the orbits of the planets, the rotation and shape of the Earth and the tides. The book contains what is now called *Laplace's equation* (in fact, known earlier). Laplace explained that 'I have sought to establish that the phenomena of nature can be reduced in the last analysis to actions at a distance between molecule and molecule, and that the consideration of these actions must serve as the basis of the mathematical theory of these phenomena'. In Laplace's later work *Théorie Analytique des Probabilités*, and with this philosophy in mind, he applied probability theory to errors in observations, the determination of the masses of Jupiter, Saturn and Uranus, surveying and geodesy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Large Binocular Telescope

The Large Binocular Telescope (LBT) project is a collaboration to build a twin interferometric telescope between institutions in Arizona, Germany, Italy and Ohio. Arizona includes astronomers at the University of Arizona, Arizona State University and Northern Arizona University. Germany is represented by the LBT Beteiligungsgesellschaft which is composed of the Max-Planck-Institut für Astronomie in Heidelberg, Landesstenwarte Heidelberg, the Max-Planck-Institut für Radioastronomie in Bonn, the Max-Planck-Institut für Extraterrestrische Physik in Munich and the Astrophysikalisches Institut Potsdam. The Italian astronomical community is represented by the Osservatorio Astrofisico di Arcetri in Florence. Partners at individual institutions include Ohio State University in Columbus, the University of Notre Dame and the Research Corporation in Tucson.

The two 8.4 m borosilicate honeycomb primary mirrors for LBT are being cast at the STEWARD OBSERVATORY Mirror Laboratory. The baseline optical configuration of LBT includes adaptive infrared $f/15$ secondaries of a Gregorian design. The interferometric focus combining the light from the two 8.4 m primaries will reimage the two folded Gregorian focal planes to three central locations. The telescope structure is being fabricated in Italy. After pre-erection in the factory, it will be shipped to Arizona towards the end of 2000. The enclosure is being built on Mount Graham in southeast Arizona.

For further information see
http://medusa.as.arizona.edu/lbtwww/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Large Hot Subdwarfs and Extreme Horizontal Branch Stars

The hot subdwarf stars were first identified by their extremely blue colors in early field and cluster color-select photometric surveys (Humason and Zwicky 1947). Spectroscopic studies soon followed (Münch 1958, Greenstein 1965). From the cluster studies, it was quickly recognized that hot subdwarfs must have lower luminosities than main sequence stars of the same blue color, i.e. O and B stars, and it was Greenstein who first labeled them 'hot subdwarfs'. They were divided into subdwarf O (sdO) and subdwarf B (sdB) classes based on their spectral features, with the hotter sdO stars showing strong He II λ 4686 absorption and very weak or absent He I absorption, while, in the cooler sdB stars, the He II feature was lacking or very weak, and the neutral helium lines showed a broad range of strengths. Model atmosphere analyses of their spectra have confirmed their high effective temperatures and surface gravities, and their smaller radii and luminosities.

In recent years, the sdB stars have become firmly identified with the horizontal branch (HB) phase of post-main-sequence stellar evolution, in which stars derive their energy from the burning of helium into carbon (see HORIZONTAL BRANCH STARS). The relationship is seen most clearly in metal-poor globular clusters (GC), where the cluster blue horizontal branch (BHB) stars often are seen to have an extended blue 'tail', or extended horizontal branch (EHB). However, not all metal-poor clusters exhibit this phenomenon, and the existence or absence of blue tails in GC HBs is a mystery which could be strongly connected to the process that forms sdBs. Pairs of clusters such as M3–M13 (famous so-called 'second parameter' pairs) are important for learning about the origin of sdB stars. A review is given by Moehler (1999). Photometrically and spectroscopically, the cluster EHB stars are found to be indistinguishable from their field sdB counterparts. In this article, the sdB and EHB designations are used interchangeably.

To set the stage for a discussion of the sdB stars, we first briefly review the main evolutionary stages of intermediate-mass stars, i.e. those with initial masses of about (1–8)$M_{\odot}$. These stars ultimately will end their lives as white dwarfs (WDs). We take as a representative example a star having an initial mass equal to that of our own Sun, with the reminder that initially more massive stars will evolve on shorter timescales.

Evolution of intermediate-mass stars

Stellar birth and early evolution

Stars begin their lives as contracting condensations in massive clouds of gas and dust. Their compressed central regions eventually become hot and dense enough for hydrogen to undergo nuclear fusion reactions that transform it into helium, liberating vast amounts of energy in the process. After central hydrogen ignition, the star settles into its first stable stage of STELLAR EVOLUTION, the hydrogen-burning main sequence. Our own Sun is in this stage, which for stars having a similar initial mass lasts for some 10^{10} yr. In the central 10% of the mass (the core), the temperature and density are high enough to sustain nuclear burning reactions, and hydrogen is steadily converted into helium. The energy is released both as light and as kinetic energy of the gas particles (that is, as heat), both of which serve to stabilize the collapse of the star against its own self-gravity. At the end of the main sequence phase, the available hydrogen in the core is completely converted into inert helium. The core material is not hot or dense enough to initiate helium burning, and, lacking a source of fuel, fusion reactions within the core cease.

Red giant branch evolution

Without the continuing generation of heat and pressure from the now-extinct fusion reactions, the stellar core can no longer support itself against its own gravitational pull, and it begins a slow, steady contraction, becoming even hotter and denser as it does so. For intermediate-mass stars like the Sun, the core is ultimately able to stabilize itself owing to the onset of quantum-mechanical electron degeneracy, which provides the pressure necessary to balance the star's self-gravity. Hydrogen does continue to burn in a thin shell around the now-degenerate $0.1M_{\odot}$ inert helium core, and the star enters the red giant branch (RGB) stage of evolution (see RED GIANT STARS).

The RGB is so named because the extremely high temperatures at the surface of the core drive shell hydrogen burning at an even higher luminosity than the previous stage of core hydrogen burning. The increased energy flow causes the outer, hydrogen-rich envelope to expand and cool, and hence the star moves up and to the right in the H–R diagram. As shell hydrogen burning continues to add helium to the core, the peculiar properties of the degenerate matter result in the core shrinking as the mass grows. The core grows even hotter as a result, shell hydrogen burning proceeds at an even faster rate, and the stellar luminosity rises even further. The star thus steadily 'ascends' the RGB as the core continues to grow in mass. The star remains in this stage of evolution for about 10^9 yr, i.e. about 10% of the time it spends on the main sequence. We note that the termination of this phase of evolution is not determined by the exhaustion of a nuclear fuel as it was at the end of the main sequence phase, but rather by the ignition of a new nuclear fuel, helium.

Horizontal branch evolution

When the inert, degenerate helium core mass reaches $\sim$(0.45–0.50)$M_{\odot}$, the temperature and density become high enough to initiate a second stage of nuclear burning—the so-called 'triple-alpha' (3α) process—wherein three helium nuclei are fused to a single nucleus of carbon. The re-ignition of the central 'furnace' provides sufficient heat and pressure to lift the electron degeneracy and return the core material to a 'normal' condition, where the pressure is determined through an ideal gas equation of state. The star

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

once again is able to hold gravity at bay, and it commences a second stable phase of energy production through the fusion of helium to carbon in the core. Hydrogen continues to burn in a shell surrounding the helium-burning core, with each energy source providing approximately half of the star's total luminosity. This phase of evolution has been termed the 'horizontal branch' (HB), since the V magnitude of stars in this phase is about the same for a broad range of effective temperatures ($T_{\rm eff}$). This results in a flat or horizontal distribution in the H–R diagram, at least for all but the hottest blue HB stars. HB stars remain in this stage of evolution for about 10^8 yr, i.e. about 10% of the time they spend on the RGB. At the end of the HB phase, helium is exhausted in the core, having been converted to carbon and oxygen, the latter having been produced by α capture onto carbon.

Subsequent evolution through the asymptotic red giant branch and PLANETARY NEBULA phases, and the star's final WHITE DWARF phase, are discussed elsewhere in this encyclopedia.

Observational characteristics of EHB stars

Optical colors and spectra

In EHB stars, the optical broadband colors are quite blue, $B-V \sim -0.3, U-B \sim -1.0$, so that the detection efficiency is very high in ultraviolet-excess color surveys. However, a significant fraction (15–20%) of hot subdwarfs are found in binaries with bright main sequence or subgiant companions which redden the optical colors. The optical spectra are characterized by blue continuua, with broad, shallow hydrogen and diffuse helium absorption lines (compared with main sequence B stars). As previously noted, the distinction between the sdO and sdB subclasses of hot subdwarfs has been based on the presence and strength of the H, He I and He II absorption lines.

Recently, Jeffery *et al* (1997) carried out spectroscopy of hot subdwarfs previously classified as 'helium-rich subdwarf B stars' from low-resolution spectra. They found a wider range in He II/He I and He/H ratios than previously found in the class. Their revised classification scheme for helium-rich hot subdwarfs is based on higher-resolution, high signal-to-noise ratio (S/N) spectra and can be extended to include all hot subdwarfs with spectra dominated by H, He I or He II.

Ultraviolet spectra and abundances

The ultraviolet (UV) spectra of sdB stars show numerous photospheric lines of heavy elements with high excitation, such as N III, C IV, Si IV and others, usually very sharp compared with the lines of hydrogen and helium. Additional weak, very sharp features are seen in the optical spectra at very high dispersion and S/N. The elemental abundances are modified by diffusion and, possibly, mass-loss processes in the hot, higher-gravity atmospheres, with helium typically underabundant, and with carbon and silicon underabundant. Nitrogen usually appears normal. The modified abundance patterns obscure relationships between sdBs and their progenitor stars and make their evolutionary history more difficult to establish.

The photometric and spectroscopic observational history of hot subdwarfs is rich, if brief; a summary is given by Saffer and Liebert (1995).

Astrophysical properties

Atmospheric parameters

In the first comprehensive analysis of hot subdwarfs, Greenstein and Sargent (1974) found that most sdB stars fell into a $T_{\rm eff}$ range of 25 000–35 000 K and had a range of surface gravities log g between 5 and 6. They argued that the HB and EHB stars had similar structures, with helium-burning cores and hydrogen-rich envelopes. The EHB stars, however, had such thin envelopes ($\lesssim 0.01 M_\odot$) that they could not sustain a hydrogen-burning shell, which in normal HB stars produces approximately half of the stellar luminosity.

The Palomar Green (PG) Survey (Green *et al* 1986) revealed large numbers of hot subluminous stars. The new, large sample prompted fresh analyses to determine effective temperatures, surface gravities and chemical abundances (e.g. Moehler *et al* 1990, Saffer *et al* 1994, Saffer and Liebert 1995 among others; see Saffer and Liebert 1995 for a summary). Complementary programs in the southern hemisphere, the Edinburgh–Cape (EC) and Hamburg Schmidt (HS) surveys, have also identified large numbers of hot subdwarfs at high galactic latitudes (see also BLUE STARS AT HIGH GALACTIC LATITUDES). The compilation of Kilkenny *et al* (1988) lists 1225 spectroscopically identified hot subdwarfs, many of which now have estimated temperatures and surface gravities from modern model atmosphere analyses. The addition of hot subdwarfs from the EC and HS surveys will undoubtedly significantly increase these numbers.

Evolutionary status

Modern theoretical evolutionary calculations of HB and EHB sequences make predictions of effective temperatures and surface gravities for a range of core and envelope masses (Sweigart 1987, Caloi 1989, Dorman *et al* 1993). Comparisons with observational estimates for both the field sdB stars and globular cluster EHB stars indicate that the bulk have a very narrow range of helium core masses near $0.5 M_\odot$ and very thin hydrogen-rich envelope masses $M_{\rm env} < 0.01 M_\odot$.

Figure 1 shows the distribution of a large-scale spectroscopic survey of high-latitude B stars in the $T_{\rm eff}$–log g plane, where the open symbols represent the high-gravity subdwarfs (Saffer *et al* 1997). The helium abundance is indicated by the type of symbol, being strongly depleted, mildly depleted, normal or enhanced for open circles, triangles, squares or stars, respectively. The filled symbols represent halo B-type star candidates. Proceeding from low to high surface gravity, the two solid curves are post-AGB evolutionary sequences labeled with the core mass (Schönberner 1983), the dot-dashed curve is the population I zero-age main sequence (ZAMS) (Allen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1973), the diagonally curving dashed and dotted lines are $0.5M_{\odot}$ core helium-burning HB stars with varying overlying hydrogen-rich envelope masses (Sweigart 1987, Caloi 1989), and the solid, nearly horizontal curve at $\log g \gtrsim 6.0$ is the helium ZAMS (Paczyński 1971). The latter is labeled with the mass of the helium-burning core (see HERTZSPRUNG–RUSSELL DIAGRAM, POPULATION I STARS). The undulating curves lying between the post-AGB and HB tracks are post-core helium exhaustion stars where helium burning continues in a shell around an inert carbon–oxygen core (Caloi 1989).

The theoretical sequences also show that following the exhaustion of helium in the core, EHB stars are unable to ascend the asymptotic giant branch (AGB) because their thin hydrogen envelopes cannot sustain a hydrogen-burning shell. Instead, they move to higher temperatures and luminosities for durations of about 10% of their core-helium-burning lifetimes, as they undergo helium shell flashes in the region of the H–R diagram occupied by the sdO stars. Thus, at least some of the sdO stars appear to be the direct descendents of the EHB stars. Some low-mass post-AGB evolutionary tracks also pass through the sdO region (Schönberner 1983), and mergers of pairs of He WDs also may produce sdO stars (Bailyn and Iben 1989).

Chemical abundances

Detailed calculations of radiative acceleration on helium and heavier elements suggest that some elements can be supported against gravitational settling in the hot, high-gravity atmospheres of sdB stars (Lamontagne *et al* 1985, 1987, Michaud *et al* 1985, 1989; Bergeron *et al* 1988; see also CHEMICAL COMPOSITION OF STARS). Observations generally support this picture, as hot subdwarfs do show markedly non-solar chemical abundance distributions, presumably as the result of diffusion processes (see also SOLAR ABUNDANCES). However, the observed detailed abundance patterns are generally not in agreement with theoretical predictions. For example, silicon is predicted to be supported in sdB stars but is found to be underabundant, and helium should be extremely depleted (so that helium absorption lines should not be detectable) but in general is found to be underabundant only by a factor of 10. These results suggest that other particle transport processes, such as a weak stellar wind, operate in subdwarf atmospheres. A review is given by Fontaine and Chayer (1997), and previous abundance analyses from UV spectroscopy have been reviewed by Heber (1998).

The EC 14026 stars

In 1997 a new class of pulsating star was discovered. These stars, designated the EC 14026 stars after the prototype, are pulsating subdwarf B stars with $T_{\rm eff} \sim 35\,000$ K and $\log g \sim 6.0$. The multiperiodic oscillations have periods $P \sim 120$–500 s and amplitudes typically less than 1%. Photometric surveys have thus far identified 18 pulsators. Their spectra reveal that several of the hot subdwarfs are binary, having a cooler main sequence or subgiant F or G type star. Apparently constant stars mingle with the pulsators, so that there is no well-defined instability strip. This poses a significant challenge to the Fe opacity bump mechanism proposed to drive the oscillations (Charpinet *et al* 1997). Indeed, new abundances for pulsating sdB stars derived from high-resolution Keck spectra (Heber *et al* 1999, 2000) strengthen the conclusion that the iron abundance is not the sole discriminating factor for pulsation driving. Identification of the modes of oscillation and the driving mechanism will lead to an improved asteroseismological understanding of hot subdwarfs and to better estimates of their masses, rotation rates, physical structure and chemical compositions.

EHB distances, scale height and midplane density

The distribution of EHB stars in the H–R diagram suggests a mass of $0.5M_{\odot}$, and reddening along the line of sight is estimated by comparing the observed and theoretically predicted colors. Combined with the stellar mass and reddening, the derived EHB effective temperatures and surface gravities lead to estimates of heights above the GALACTIC PLANE ($z = d \sin b$) of 1–3 kpc. Assuming an exponentially decreasing density with increasing height above the Galactic plane, early scale height estimates were made using samples with relatively bright limiting magnitudes. The derived scale height z_0 ranged from 180 to 240 pc, indicating membership in the old disk population. Saffer and Liebert (1995) used a deeper, more complete high-latitude sample to derive $z_0 = 500$ pc, implying that the population is significantly older than the bulk of the stars in the old part of the thin disk. Most recently, Mitchell (1998) used a faint ($V \lesssim 18$), pencil-beam sample of field sdB stars to show that a substantial fraction of sdB stars fainter than the limiting magnitudes of the large color-select surveys (PG, EC, HS) belong to the halo population and not that of the old disk. de Boer *et al* (1997) studied the orbits of relatively bright sdB stars ($V \lesssim 14.5$), finding many eccentric orbits. However, none of the stars could be positively halo objects, probably because of the bright limiting magnitude of their sample.

The Kitt Peak Downes Survey (KPD; Downes 1986) in the Galactic plane, similar to the PG Survey both in color-excess criteria and in limiting magnitude, also identified large numbers of hot subdwarfs. The sample is especially well suited for the determination of the EHB midplane space density. Using a $V/V_{\rm max}$ analysis, Saffer and Liebert (1995) estimated the field EHB space density to be $n_0 = 7.5 \times 10^{-7}$ pc^{-3}, compared with previous estimates of 1.5–2 $\times 10^{-6}$ and 4×10^{-6}. Assuming a core helium-burning lifetime $\tau = 1.5 \times 10^8$ yr, the EHB birthrate is then $b = n_0/\tau = 5 \times 10^{-15}$ pc^{-3} yr^{-1}. Current estimates of the local white dwarf birthrate range from 0.5×10^{-12} to 2×10^{-12} pc^{-3} yr^{-1}, that is, the contribution to the local white dwarf birthrate from the EHB channel is just 0.25–1%, compared with previous estimates of about 2%.

Origins and subsequent evolution

The suggestion that hot subdwarf stars might produce the significant amounts of far-ultraviolet (FUV) flux seen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

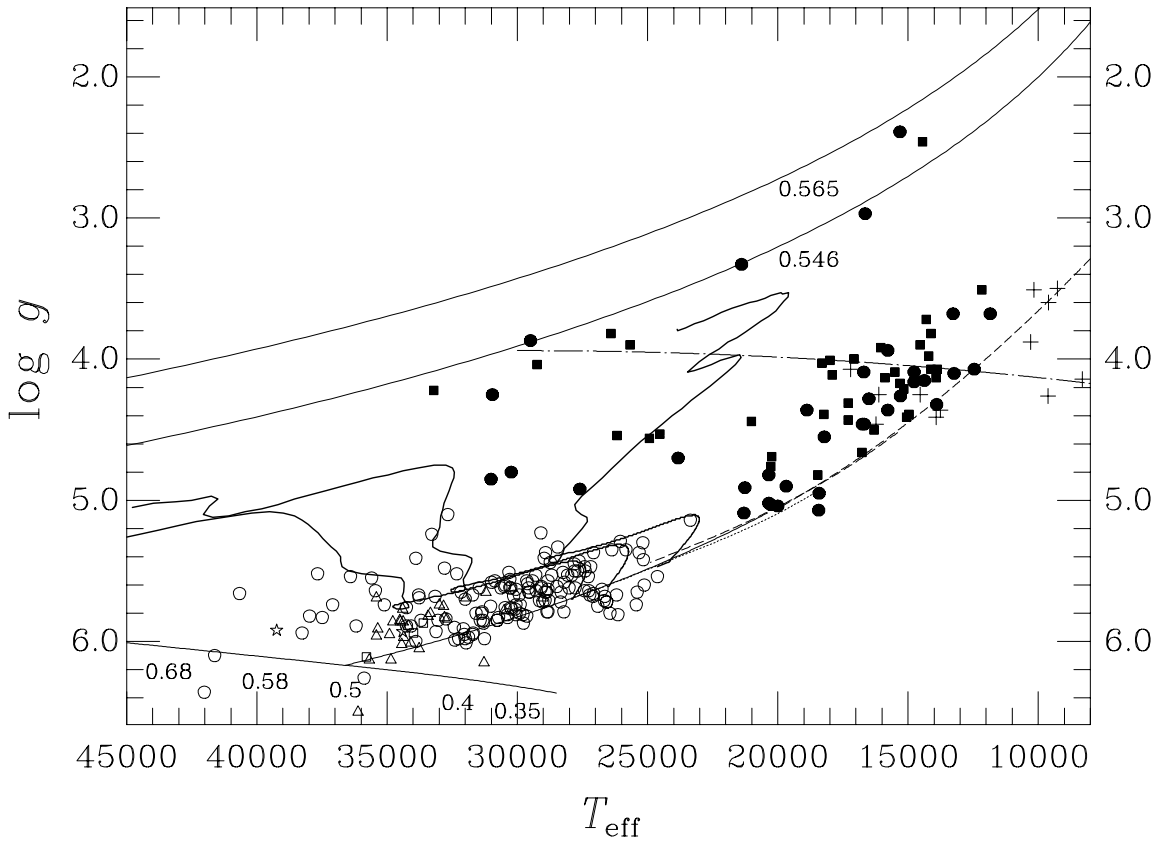


Figure 1. Distribution of a large-scale spectroscopic survey of high-latitude B stars: see text for details.

in many ELLIPTICAL GALAXIES (e.g. Dorman *et al* 1995) as well as their ubiquitous presence in the field have motivated numerous studies to determine their origins. Any proposed mechanisms must be able to reproduce the inferred physical structure of the EHB stars, namely, a helium-burning core of $\sim 0.5 M_{\odot}$ and a very thin ($<0.01 M_{\odot}$) hydrogen-rich envelope. The key problem is how to achieve the critical mass for core-helium ignition while at the same time removing virtually all of the hydrogen-rich envelope.

In metal-poor globular clusters, Reimer's mass loss in single RGB stars can plausibly remove upwards of $0.3 M_{\odot}$ from the hydrogen-rich envelope. In such a stellar population, this is sufficient to completely strip the envelope of the RGB progenitors (low-mass turn-off stars) by the time their cores grow to the critical helium-ignition mass on the RGB. This scenario is more problematic in the field and in younger, more metal-rich GLOBULAR and GALACTIC OPEN CLUSTERS, where the turn-off mass is higher. Some success has been achieved for stars with extremely high metallicities and/or helium abundances.

Alternatively, in CLOSE BINARY STARS, ROCHE LOBE overflow could easily strip the envelope. Mengel *et al* (1976) first investigated this scenario and successfully produced EHB stars, but only for a very narrow range of initial stellar masses and orbital separations. For very close initial separations, the envelope is stripped before the helium core grows to ignition mass, leaving a helium WD. For wide initial separations, helium ignition occurs while a substantial envelope mass remains, producing a normal HB star. Liebert *et al* (1994) found the range of orbital separations producing EHB stars in this scenario to be too narrow to explain the HB morphology of the metal-rich Galactic cluster NGC 6791.

Recent studies of field sdB stars imply a large binary fraction, as revealed both by composite colors and by high-precision radial velocity variations. Green and Chaboyer (1998) showed that sdB stars can be distingushed by their spectroscopic and kinematic properties. Some ∼15–20% of field sdB stars have composite spectra indicating the presence of a cool companion. These either do not vary or have small radial velocity variations indicative of binaries with wide separations (a few AU). A second group of ∼45–50% of the field sdB stars have large radial velocity variations with characteristic timescales of days. These have no obvious visible companions, which are assumed to be WDs. Rarely, the invisible companions are detected through eclipses or reflection effects. The remaining ∼35% of field sdBs are apparently single.

Provided that most field sdB stars are the counterparts of the GC EHB stars, their fate seems certain. Since they

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cannot add to their $0.5M_{\odot}$ cores in a subsequent AGB phase, after the short-lived post-EHB phase they seem destined to directly enter the WD cooling sequence, albeit with CO core masses of $\sim 0.5M_{\odot}$, about $0.1M_{\odot}$ less than the average WD. As noted previously, the sdB stars contribute only a very minor fraction to the total WD birthrate.

Unsolved problems

Much progress has been made in the last decade toward the understanding of the astrophysical properties and evolutionary histories of the sdB/EHB stars. The understanding of several key issues remains elusive.

(1) How are hot subdwarfs formed? Evidence increasingly points to a close binary evolution connection, but the details are still sparse. A comprehensive theoretical investigation of close binary Roche lobe overflow and envelope stripping over a wide range of initial stellar masses and initial separations is needed. No such investigation has been undertaken since that of Mengel *et al* (1976). Infrared observations of the apparently single population of field sdB stars would confirm or rule out the existence of low-mass secondaries hitherto undetected. These results would have significant implications for the outcome of the common-envelope phase which all sufficiently close systems must undergo.

(2) Theoretical predictions of the detailed abundance distributions in sdB photospheres remain at odds with the observations. New generations of model atmosphere codes that self-consistently take elemental diffusion and stellar winds into account in calculating the atmospheric structure are now available and should shed new light on this persistent problem. A description of these new codes is given by Dreizler (1999).

(3) The pulsating sdB stars offer a new and exciting asteroseismological window into sdB STELLAR INTERIORS. Detailed abundance distributions of the iron-group elements may help solve the current conundrum of the existence of constant stars in the same temperature and gravity ranges as known pulsators. Iron itself, the suspected opacity source for the driving mechanism, does not seem to be the discriminating factor.

Bibliography

Allen C W 1973 *Astrophysical Quantities* (London: Athlone)
Bailyn C D and Iben I Jr 1989 *Astrophys. J.* **347** L21
Bergeron P, Wesemael F, Michaud G and Fontaine G 1988 *Astrophys. J.* **332** 964
Caloi V 1989 *Astron. Astrophys.* **221** 27
Charpinet S, Fontaine G, Brassard P, Chayer P, Rogers F J, Iglesias C A and Dorman B 1997 *Astrophys. J.* **483** L123
de Boer K S, Aguilar Sanchez Y, Altmann M, Geffert M, Odenkirchen M, Schmidt J H K and Colin J 1997 *Astron. Astrophys.* **327** 577
Dorman B, O'Connell R W and Rood R T 1995 *Astrophys. J.* **442** 105
Dorman B, Rood R T and O'Connell R W 1993 *Astrophys. J.* **419** 596
Downes R A 1986 *Astrophys. J. Suppl.* **61** 569
Dreizler S 1999 *Astron. Astrophys.* **352** 632
Fontaine G and Chayer P 1997 *3rd Conf. on Faint Blue Stars* ed A G D Philip *et al* (Davis) p 169
Green E M and Chaboyer B C 1998 *Bull. Am. Astron. Soc.* **193** 102.13
Green R F, Schmidt M and Liebert J 1986 *Astrophys. J. Suppl.* **61** 305
Greenstein J L 1965 *Galactic Structure, Stars and Stellar Systems* vol V, ed A Blaauw and M Schmidt (Chicago, IL: University of Chicago Press) p 361
Greenstein J L and Sargent A I 1974 *Astrophys. J. Suppl.* **28** 157
Heber U 1998 *Ultraviolet Astrophysics Beyond the IUE Final Archive (ESA SP 413)* ed W Wamsteker, R Gonzalez Riestra and B Harris (ESA) p 195
Heber U, Reid I N and Werner K 1999 *Astron. Astrophys.* **348** L25
Heber U, Reid I N and Werner K 2000 *Acta Astronomica* **8** in press
Humason M L and Zwicky F 1947 *Astrophys. J.* **105** 85
Jeffery C S, Drilling J S, Harrison P M, Heber U and Moehler S 1997 *Astron. Astrophys. Suppl.* **125** 501
Kilkenny D, Heber U and Drilling J S 1988 *SAAO Circular 12*
Lamontagne R, Wesemael F and Fontaine G 1987 *Astrophys. J.* **318** 844
Lamontagne R, Wesemael F, Fontaine G and Sion E M 1985 *Astrophys. J.* **299** 496
Liebert J, Saffer R A and Green E M 1994 *Astron. J.* **107** 1408
Mengel J G, Norris J and Gross P G 1976 *Astrophys. J.* **204** 488
Michaud G, Bergeron P, Heber U and Wesemael F 1989 *Astrophys. J.* **338** 417
Michaud G, Bergeron P, Wesemael F and Fontaine G 1985 *Astrophys. J.* **299** 741
Mitchell K J 1998 *Astrophys. J.* **494** 256
Moehler S 1999 *Rev. Mod. Astrophys.* **12** 281
Moehler S, Heber U and de Boer K S 1990 *Astron. Astrophys.* **239** 265
Münch G 1958 *Astrophys. J.* **127** 642
Paczyński B 1971 *Acta Astronomica* **21** 1
Saffer R A and Liebert J 1995 *Proc. 9th European Workshop on White Dwarfs* ed D Koester and K Werner (Berlin: Springer) p 221
Saffer R, Bergeron P, Koester D and Liebert J 1994 *Astrophys. J.* **432** 351
Saffer R A, Keenan F P, Hambly N C, Dufton P L and Liebert J 1997 *Astrophys. J.* **491** 172
Schönberner D 1983 *Astrophys. J.* **272** 708
Sweigart A V 1987 *Astrophys. J. Suppl.* **65** 95

Rex A Saffer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Large Magellanic Cloud

The larger of two nearby companions of the Milky Way Galaxy that can be seen with the naked eye in the southern hemisphere sky and which are named after the Portuguese navigator, Ferdinand Magellan, who observed them in 1519 during his circumnavigation of the world. Located in the constellation of Dorado, at a distance of about 170 000 light-years, the Large Magellanic Cloud (LMC) has an overall diameter of about 35 000 light-years and contains about 10^{10} stars. Its overall mass is between one-tenth and one-twentieth of that of the Milky Way Galaxy. It was originally classified as an irregular galaxy (of type Irr I in the Hubble classification), but, because it has a conspicuous central bar and shows hints of what may be an incipient spiral arm, it is now considered to be an irregular barred spiral, the prototype of a class designated 'Sm'. It contains the Tarantula Nebula, a huge HII region, with a diameter of about 900 light-years, which surrounds a vigorous star-forming region, and is one of the largest HII regions known. The LMC, which contains numerous clouds of gas and dust, young star clusters and supernova remnants, appears to be experiencing a major bout of star formation that was probably triggered by the effects of a close encounter with the Milky Way a few billion years ago.

See also: barred spiral galaxy, H II regions, irregular galaxy, Local Group, Milky Way Galaxy, nebula, Small Magellanic Cloud, spiral arms, Tarantula Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Large-Scale Plasma Dynamics

Most of the universe is occupied by gas that is electrically neutral but more or less fully ionized, i.e. PLASMA, and permeated by a magnetic field $\boldsymbol{B}$ on a large scale L. The plasma, driven by the energy of the stars, is in a remarkably active, and often suprathermal, state. The coronal mass ejection from the Sun is an example of the active state, while the million degree corona of the Sun is an example of the suprathermal state, giving rise to the dynamical expansion that produces the supersonic SOLAR WIND sweeping out through the solar system. The interstellar gas is powerfully agitated by the intense heating around massive HIGH LUMINOSITY STARS. Expanding SUPERNOVA shells and the CRAB NEBULA are examples of dynamical states, whose shock fronts heat the gas throughout large volumes of interstellar space to suprathermal temperatures of 10^5 K or more, causing the gas to expand up out of the gaseous disk of the Galaxy to contribute to the galactic halo. So a proper formulation of the dynamics of a magnetized plasma is essential for understanding the activity of the Sun and other stars and of the Galaxy.

In a tenuous fully ionized gas the individual electrons and ions gyrate around the magnetic field while being scattered by electrostatic interactions with their neighbors. The cyclotron frequency of the gyration is $\Omega = qB/Mc$ for a particle with mass M and charge q. The cyclotron radius R is $w_\perp/\Omega$ where $w_\perp$ is the velocity of the particle perpendicular to the magnetic field. The electric field of an individual particle is screened off by its neighbors beyond the characteristic Debye length $D = (kT/4\pi Ne^2)^{1/2}$ in a plasma with temperature T and number density N. The effective mean free path for both electrons and ions is approximately (Spitzer 1956)

$$\lambda = \frac{2 \times 10^5 T^2}{N \ln(1.2 \times 10^4 T^{3/2}/N^{1/2})} \text{ cm}$$

This article examines the large-scale bulk dynamics of such plasmas and fields.

With rare exceptions the plasma beneath the surface of a star is collision dominated ($\lambda \ll R$), and the electric current density $\boldsymbol{j}$ required by Ampère's law,

$$4\pi \boldsymbol{j} = c\,\boldsymbol{\nabla} \times \boldsymbol{B}$$

is pushed through the plasma by the electric field $\boldsymbol{E}'$ in the local frame of reference of the moving plasma, where $\boldsymbol{j} = \sigma \boldsymbol{E}'$ and the conductivity σ is approximately 2×10^7 T$^{3/2}$ s^{-1} for ionized hydrogen. The electric field $\boldsymbol{E}'$ in the frame of the plasma moving with velocity $\boldsymbol{u}$ is related to the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ in the laboratory frame by the Lorentz transformation

$$\boldsymbol{E}' = \boldsymbol{E} + \frac{\boldsymbol{u} \times \boldsymbol{B}}{c}$$

on neglecting terms second order in u/c. The current density is automatically regulated to satisfy Ampère's law because of the rapid growth of the electric field,

$$\frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{c}\,\boldsymbol{\nabla} \times \boldsymbol{B} - 4\pi \boldsymbol{j}$$

in the direction of any current deficit, with the energy coming from the magnetic field through the induction equation $\partial \boldsymbol{B}/\partial t = -c\,\boldsymbol{\nabla} \times \boldsymbol{E}$. Thus $\boldsymbol{E}'$ continually evolves with changing $\boldsymbol{u}$ and $\boldsymbol{B}$ so that $\boldsymbol{E}' = c\,\boldsymbol{\nabla} \times \boldsymbol{B}/4\pi\sigma$. The energy flow in the frame of the large-scale bulk motion of the plasma is always from the field into the current, as a consequence of the nonvanishing resistivity, and the large-scale dynamics is described by the familiar induction and momentum equations of MAGNETOHYDRODYNAMICS, with $\boldsymbol{\nabla} \times \boldsymbol{E}'$ in the induction equation providing the resistive diffusion of $\boldsymbol{B}$.

In contrast, the mean free path λ in extended stellar atmospheres, in the solar wind, in planetary magnetospheres and in interstellar space (excluding cold dense molecular clouds) is generally large compared with the cyclotron radius R, so the electrons and ions gyrate relatively freely in the magnetic field. The large-scale bulk velocity $\boldsymbol{u}$ of the plasma represents the local mean momentum per unit mass of the individual electrons and ions. The velocity $\boldsymbol{w}$ of a particle relative to the mean motion $\boldsymbol{u}$ represents the thermal velocity, which has zero mean when summed over the local particles.

Consider, then, the bulk dynamics of a plasma and magnetic field in the limit that interparticle collisions can be neglected ($\lambda \gg R$) and the electrons and ions gyrate freely in the large-scale magnetic field $\boldsymbol{B}$. The particle pressure $p_\perp$ perpendicular to $\boldsymbol{B}$ is isotropic about the direction of $\boldsymbol{B}$ as a consequence of the cyclotron motion $w_\perp$, and

$$p_\perp = \frac{1}{2} \sum M w_\perp^2$$

where the sum is over all particles in unit volume. The pressure $p_\parallel$ parallel to $\boldsymbol{B}$ is given by

$$p_\parallel = \sum M w_\parallel^2$$

where the sum is again over unit volume, and $w_\parallel$ represents the thermal velocity parallel to $\boldsymbol{B}$. It should be noted that $p_\parallel$ and $p_\perp$ are not necessarily equal in a collisionless plasma (see COLLISIONLESS PROCESSES IN ASTROPHYSICAL PLASMAS). However, thermal anisotropy and any strong electron conduction velocity (relative to the ions) generally excite plasma oscillations that scatter the particles, pulling the thermal motions toward isotropy ($p_\parallel \sim p_\perp$). Thus, for instance, in the high-speed tenuous solar wind the anisotropic expansion maintains a measurable but limited thermal anisotropy. On the other hand, no significant anisotropy is expected in the quasi-static corona close to the Sun.

Now the bulk velocity of a tenuous plasma with mass density ρ is described by the momentum equation

$$\frac{\partial \rho u_j}{\partial t} = \frac{\partial}{\partial x_k}(-p_{jk} + M_{jk} + R_{jk})$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u_k}{\partial x_k} = 0$$

where p_{jk} is the pressure tensor, with diagonal terms $p_\parallel$ and $p_\perp$. Off-diagonal terms represent viscosity, if it is relevant. The Maxwell stress tensor M_{jk} is the usual $-\delta_{jk}B^2/8\pi + B_jB_k/4\pi$, representing the isotropic pressure $B^2/8\pi$ and the tension $B^2/4\pi$ in the direction of the field. The Reynolds stress tensor R_{jk} is $-\rho u_j u_k$, representing compressive stress in the direction of the velocity. That is to say, p_{jk} is the momentum flux of the thermal motions and $-R_{jk}$ is the momentum flux of the bulk motion. This form of the momentum equation follows directly from application of Gauss' theorem to the net momentum transport into a fixed volume V with dimensions small compared with L but sufficiently large that $NV \gg 1$.

Individual particle motions

To proceed further, consider the motion of an individual ion or electron, described by a mass M and charge q. With $L \gg R$ it is sufficient to use the guiding center approximation, in which the particle moves primarily in a circle with cyclotron radius $R = Mw_\perp c/qB \ll L$. The center of the circle—the guiding center—slides freely along the field while it drifts slowly across the field as a consequence of the large-scale variation of the field. The cyclotron motion represents a mean electric current $q\Omega/2\pi$ around the perimeter of the area πR^2, providing a diamagnetic dipole moment $\mu = Mw_\perp^2/2B$ that is a constant of the motion for each particle in the limit of large L/R. The diamagnetic moment repels the particle from regions of strong field with a force $\mu\,\partial B/\partial s$ along the field, where s represents distance along the field. Thus the motion $w_\parallel$ parallel to $\boldsymbol{B}$ is described by

$$\frac{\mathrm{d}w_\parallel}{\mathrm{d}t} = -\mu\frac{\partial B}{\partial s}.$$

It should be noted that in a plasma with isotropic thermal motions ($p_\parallel = p_\perp$) this repulsion from a converging field exactly compensates for the particle concentrating effect $(Nw_\parallel/B)\,\partial B/\partial s$ of the converging field. When $p_\perp > p_\parallel$, the net effect is to push the plasma away from the regions of strong field, whereas, when $p_\perp < p_\parallel$, the converging field lines enhance the plasma in regions of strong field. The isotropic plasma maintains a uniform density along the field, for the reason given.

Consider the force $\boldsymbol{F}$ perpendicular to $\boldsymbol{B}$ applied to the particle. The initial effect is to accelerate the particle in the direction of $\boldsymbol{F}$. That motion provides a Lorentz force and a drift velocity $\boldsymbol{W}$ in the direction $\boldsymbol{F}\times\boldsymbol{B}$, so that the Lorentz force $q\boldsymbol{W}\times\boldsymbol{B}/c$ opposes $\boldsymbol{F}$. The acceleration of the particle falls to zero when $\boldsymbol{W}$ reaches the value $c\boldsymbol{B}\times\boldsymbol{F}/qB^2$ and its Lorentz force cancels the applied $\boldsymbol{F}$. The total displacement ξ in the direction of $\boldsymbol{F}$ is $(Mc^2/q^2B^2)\boldsymbol{F}$ such that the impulse $q\xi\times\boldsymbol{B}/c$ is equal to the final transverse momentum $M\boldsymbol{W}$.

In the presence of an electric field $\boldsymbol{E}_\perp$ the force is $\boldsymbol{F} = q\boldsymbol{E}_\perp$ and the final velocity across the magnetic field is the electric drift velocity $\boldsymbol{u}_\perp = c\boldsymbol{E}\times\boldsymbol{B}/B^2$ independent of the charge and mass of the particle. Every ion and electron takes up the velocity $\boldsymbol{u}_\perp$, so that $\boldsymbol{E}_\perp = -\boldsymbol{u}\times\boldsymbol{B}/c$ and the plasma moves in the frame of reference in which there is no electric field, $\boldsymbol{E}'_\perp = 0$. Thus $\boldsymbol{u}$ represents the bulk motion of the plasma, and the electric field $\boldsymbol{E}$ in the laboratory frame arises because the electric field in the frame of reference of the plasma is zero. Thus the value of $\boldsymbol{E}$ depends on the arbitrary choice of the laboratory frame of reference relative to the plasma motion.

Consider, then, the electric field $\boldsymbol{E}_\parallel$ in the direction of $\boldsymbol{B}$. In the presence of strong unequal electron and ion thermal anisotropies, $\boldsymbol{E}_\parallel$ prevents any significant separation of ions and electrons along the field, so that potential differences as large as the thermal energy are theoretically possible. However, in the large-scale slowly evolving plasma there is generally little or no anisotropy, as a consequence of the small-scale plasma instabilities driven by any anisotropy. Therefore, generally $\boldsymbol{E}_\parallel$ can be neglected, with the result that $\boldsymbol{E} = -\boldsymbol{u}\times\boldsymbol{B}/c$ and the Faraday induction equation becomes the familiar MHD induction equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{\nabla}\times(\boldsymbol{u}\times\boldsymbol{B})$$

stating that the magnetic field is carried bodily with the bulk motion $\boldsymbol{u}$ of the plasma in the frame of reference in which there is no electric field.

On the other hand, in the presence of small scales and transient conditions, arising in shock fronts, flare current sheets, auroral filaments and sheets, and in the terrestrial magnetopause, there may be substantial $\boldsymbol{E}_\parallel$, with interesting and important dissipation and diffusion not found in the large-scale plasma and field.

Particle drifts

Returning to the guiding center approximation to investigate the individual particle motion, note that the magnetic field $\boldsymbol{B}$ is assumed to vary over the large scale L. The direction of the field may vary along the field, so that the thermal motion $w_\parallel$ of the guiding center along the field introduces a centrifugal force $F = Mw_\parallel^2 K$, where K is the curvature of the field line through the guiding center. The result is the *curvature drift*

$$\boldsymbol{u}_\mathrm{c} = \frac{Mw_\parallel^2 c}{qB^4}\boldsymbol{B}\times[(\boldsymbol{B}\cdot\boldsymbol{\nabla})\boldsymbol{B}].$$

Then the magnitude of the field may vary in a direction perpendicular to the field, so that the cyclotron motion experiences a stronger field on one side of the circle than the other, producing the *gradient drift*

$$\boldsymbol{u}_\mathrm{g} = \frac{Mw_\perp^2 c}{2qB^4}\boldsymbol{B}\times\boldsymbol{\nabla}\frac{B^2}{2}.$$

The bulk acceleration $\mathrm{d}\boldsymbol{u}/\mathrm{d}t$ provides an inertial force $\boldsymbol{F} = M\,\mathrm{d}\boldsymbol{u}/\mathrm{d}t$, producing the *polarization drift*

$$\boldsymbol{u}_\mathrm{p} = \frac{Mc}{qB^2}\boldsymbol{B}\times\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Note that u_c and u_g are small $O(R/L)$, and u_p is small $O(R/L)(u^2/w^2)$ compared with w, so that in the limit of small R/L the bulk motion reduces to $\boldsymbol{u}$.

Now all three of these drift motions depend on the sign of the charge q. Hence they provide an electric current, with density

$$\boldsymbol{j}_\perp = \frac{c}{B^2}\boldsymbol{B}\times\left[\boldsymbol{\nabla} p_\perp + (\boldsymbol{B}\cdot\boldsymbol{\nabla})\boldsymbol{B}\frac{p_\parallel - p_\perp}{B^2} + \rho\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}\right]$$

on taking into account the several geometrical factors.

The obvious question is whether $\boldsymbol{j}_\perp$ satisfies Ampère's law. Noting that $(\boldsymbol{\nabla}\times\boldsymbol{B})_\perp$ can be written as $\boldsymbol{B}\times[(\boldsymbol{\nabla}\times\boldsymbol{B})\times\boldsymbol{B}]/B^2$, Ampère's law becomes

$$\boldsymbol{B}\times\left[\rho\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}+\boldsymbol{\nabla}_\perp p_\perp+(\boldsymbol{B}\cdot\boldsymbol{\nabla})\frac{p_\parallel - p_\perp}{B^2} - \frac{(\boldsymbol{\nabla}\times\boldsymbol{B})\times\boldsymbol{B}}{4\pi}\right]=0$$

requiring that

$$\rho\frac{\mathrm{d}\boldsymbol{u}_\perp}{\mathrm{d}t} = -\boldsymbol{\nabla}_\perp\left(p_\perp+\frac{B^2}{8\pi}\right)+[(\boldsymbol{B}\cdot\boldsymbol{\nabla})\boldsymbol{B}]_\perp\left(\frac{1}{4\pi}-\frac{p_\parallel - p_\perp}{B^2}\right)$$

for any acceleration perpendicular to $\boldsymbol{B}$. This expression is recognizable as Newton's equation for the bulk motion $\boldsymbol{u}$ perpendicular to the magnetic field $\boldsymbol{B}$. The term $(p_\parallel - p_\perp)/B^2$ represents the net centrifugal force of the excess field-aligned particle motion along the curved lines of $\boldsymbol{B}$. In the common case of near thermal isotropy it can be neglected, and the momentum equations take the same form as with a classical fluid. The essential point is that Ampère's law is automatically satisfied by a plasma (Parker 1957).

The special case of bulk motion $u_\parallel$ parallel to a stationary $(\partial/\partial t = 0)$ field $\boldsymbol{B}$ is described by the equation

$$\rho\left(\frac{\partial}{\partial t}+u_\parallel\frac{\partial}{\partial s}\right)u_\parallel = -\frac{\partial p_\parallel}{\partial s}+\frac{p_\parallel - p_\perp}{B}\frac{\partial B}{\partial s}.$$

where the term in the anisotropy $p_\parallel - p_\perp$ is a consequence of the net diamagnetic moment μ of each particle and the convergence of the field lines. The equation for conservation of particles is

$$\frac{\partial\rho}{\partial t}+\frac{\partial\rho u_\parallel}{\partial s} = \frac{\rho}{B}u_\parallel\frac{\partial B}{\partial s}.$$

The electric current $\boldsymbol{j}_\parallel$ required by the torsion $(\boldsymbol{\nabla}\times\boldsymbol{B})_\parallel$ arises from a weak $\boldsymbol{E}_\parallel$ acting on the particles moving along $\boldsymbol{B}$, particularly those for which $w_\parallel > w_\perp$.

General considerations

These calculations show the mutual consistency of the equations of Newton and Maxwell, established long ago by Poynting's theorems. The MHD equations are the expression of Newton and Maxwell for the large-scale bulk motion of both dense and tenuous plasmas. The MHD induction equation applies to any fluid that cannot support an electric field $\boldsymbol{E}'$ in its own frame of reference. No other assumption is involved. Then note that at each point in space the mechanical gyrations of the electrons and ions in the inhomogeneous magnetic field automatically supply the $\boldsymbol{j}_\perp$ required by Ampère's law. Thus the electric current is a local phenomenon and does not form an electric circuit, in which an emf at any one location affects the current everywhere around the circuit. Each local region creates its own current in exact conformity to Ampère's law and is uninfluenced, except through the eventual consequences of the momentum equation, by what is happening elsewhere around the current path. Thus the large-scale bulk motion of a plasma is a dynamical contest between the inertia and pressure gradients of the plasma and the Maxwell stresses in the magnetic field carried with the plasma.

The electric current is a passive quantity, locally defined, and contains negligible energy and stress. The kinetic energy of the electron conduction velocity is small $O(c^2/L^2\omega_p^2)$ compared with the magnetic free energy associated with the electric current, where ω_p is the electron plasma frequency. Electron inertia comes into the picture only in rapid variations and in intense current sheets in fields with small scales. The MHD equations can be written in terms of $\boldsymbol{j}$ and $\boldsymbol{E}$, of course, by replacing $\boldsymbol{u}$ with the electric drift velocity $c\boldsymbol{E}\times\boldsymbol{B}/B^2$ and expressing $\boldsymbol{B}$ in terms of $\boldsymbol{j}$ through the Biot–Savart integral. The result is a largely intractable global integro-differential equation because of the physical remoteness of $\boldsymbol{j}$ and $\boldsymbol{E}$ from the primary variables $\boldsymbol{u}$ and $\boldsymbol{B}$.

Note that the divergence of the Poynting vector $c\boldsymbol{E}\times\boldsymbol{B}/4\pi$ can be written as $-(\partial/\partial t)B^2/8\pi - \boldsymbol{u}\cdot\boldsymbol{F}_L$ where $\boldsymbol{F}_L$ is the Lorentz force $(\boldsymbol{\nabla}\times\boldsymbol{B})\times\boldsymbol{B}/4\pi$. Expressed in terms of $\boldsymbol{u}$ the Poynting vector can be written $\boldsymbol{u}_\perp B^2/4\pi$, representing the convective transport of magnetic enthalpy $B^2/4\pi$. Thus we can write the energy equation

$$\frac{\partial}{\partial t}\left(\frac{B^2}{8\pi}\right)+\boldsymbol{\nabla}\cdot\left(\boldsymbol{u}_\perp\frac{B^2}{4\pi}\right)+\boldsymbol{u}\cdot\boldsymbol{F}_L = 0.$$

Now if Coulomb interactions (collisions) between particles are introduced, there is a frictional drag between the electron and ion motions, opposing their relative motion required by Ampère's law, and leading to a small electric field $\delta\boldsymbol{E}'$ in the frame of reference of the plasma. The bulk motion of the plasma is not affected by $\delta\boldsymbol{E}'$ because $\delta\boldsymbol{E}'$ exerts no average force on the particles, its effect being cancelled by the frictional drag that requires its introduction. However, $-c\,\boldsymbol{\nabla}\times\delta\boldsymbol{E}'$ must be added to the right-hand side of the induction equation where it appears as a diffusion term, allowing the field to spread out slowly relative to the bulk flow $\boldsymbol{u}$. Resistive diffusion is treated in the article MAGNETOHYDRODYNAMICS and is often negligible in large-scale flows.

There remains the problem of determining the pressure of the dynamical plasma. In the simple case that the thermal motions are approximately isotropic the pressure may vary adiabatically, $\mathrm{d}(p/\rho^\gamma)/\mathrm{d}t = 0$, unless there is significant RADIATIVE TRANSFER or strong thermal conduction. In that case a suitable energy equation must

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

be used to determine the varying temperature of the gas. The polytropic variation $\mathrm{d}(p/\rho^{\alpha})/\mathrm{d}t = 0$ is a device that is sometimes used to represent the pressure–density relation through monotonic compression or expansion, with the parameter α chosen to approximate the actual net heating or cooling.

If the variations of the anisotropic plasma are adiabatic, the Chew–Goldberger–Low approximation may be useful. It is based on the transverse invariant $\mu = Mw_{\perp}^2/2B$ and the longitudinal invariant $\Lambda = hw_{\parallel}$ where h represents the varying scale of the field in the direction of the field as a consequence of longitudinal compression. Conservation of magnetic flux in a flux bundle with cross section S requires that SB be constant, while conservation of particles in the tube requires that $Sh\rho$ be constant. Relating $p_{\perp}$ to μ and $p_{\parallel}$ to Λ, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{p_{\perp}}{\rho B}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{p_{\parallel}B^5}{p_{\perp}^3}\right) = 0$$

There are, of course, more complicated situations where a complete simultaneous solution with the Boltzmann equation may be required.

Finally, there are substantial regions of planetary ionospheres, stellar photospheres and cold dense interstellar gas clouds in which the gas is only slightly ionized, leaving a massive background of neutral atoms that is collisionally coupled to the ions and electrons. This does not alter the basic dynamical principles outlined above, but the dissipation terms become more complicated. The Hall and Pedersen effects are introduced. The ions and electrons collide mainly with neutral atoms rather than with each other. If there are few enough ions and electrons, the Lorentz force of the field, exerted on the ions and electrons, may drive the ions and electrons through the neutral atoms fairly rapidly, representing the Pedersen effect and sometimes referred to as *ambipolar diffusion*. The non-dissipative Hall effect in the large-scale bulk motion is large compared with the resistive diffusion to the degree that the mean free path of the electrons exceeds their cyclotron radius. The Pedersen effect exceeds the Hall effect to the degree that the mean free path of the ions exceeds their cyclotron radius.

Bibliography

Chew G F, Goldberger M L and Low F E 1956 The Boltzmann equation and the one fluid hydrodynamic equation in the absence of collisions *Proc. R. Soc.* A **236** 112

Parker E N 1957 Newtonian development of the dynamical properties of ionized gases of low density *Phys. Rev.* **109** 924

Parker E N 1996 The alternative paradigm in magnetospheric physics *J. Geophys. Res.* **101** 10 587

Schindler K, Hesse M and Birn J 1991 Magnetic field-aligned electric potentials in non-ideal plasma flows *Astrophys. J.* **380** 293

Spitzer L 1956 *Physics of Fully Ionized Gases* (New York: Interscience)

Eugene N Parker

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Larmor, Joseph (1857–1942)

Physicist, born in Magheragall, County Antrim, Ireland, became Lucasian professor in Cambridge, working on electricity, dynamics and thermodynamics, and especially the ether, the postulated material which pervades space and in which electromagnetic radiation travels. Finding no evidence that the ether existed, he was led to some of the results inherent in EINSTEIN's theory of relativity, giving an explanation of the FITZGERALD contraction independently of LORENTZ. He calculated the energy radiating from an accelerating electron and the splitting of spectral lines in a magnetic field. Larmor's name is used in connection with *Larmor precession*, the spinning of an electron in a strong magnetic field, with the electron said to oscillate at the *Larmor frequency*. Poised at the divide between Newtonian and Einsteinian physics, he could not quite bring himself to believe in relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Las Campanas Observatory

One of the large observatories in Chile, Las Campanas Observatory, is located at the southern edge of the Atacama Desert. The local conditions make Las Campanas an ideal site for optical and infrared observations. Las Campanas is the main observational facility of the CARNEGIE OBSERVATORIES and has been active since 1971. Funding for the operation of Las Campanas is provided by The Carnegie Institution of Washington, a private not-for-profit organization.

Currently functioning telescopes include the 2.5 m du Pont and the 1 m Swope telescopes, while the 6.5 m telescopes of the Magellan Project are currently under construction. The first of these two technologically new instruments will go into operation in early 2000, while the second telescope will go into operation in 2002. The Magellan Project is a joint enterprise with Harvard University, MIT, University of Michigan and University of Arizona. The Universities of Warsaw (Poland), Nagoya (Japan) and Birmingham (UK) operate other telescopes at Las Campanas.

Various imaging, photometric and spectroscopic instruments are available at Las Campanas and a suite of new generation instruments are being built by the partners, for the Magellan telescopes.

Among many fundamental contributions, observations at Las Campanas have provided insight to problems like the Great Attractor, the Las Campanas redshift survey, the universe at high redshift, the age and distance scale of the universe, the structure and chemical composition of our own galaxy, the Galactic Bulge, etc. The supernova 1987A was discovered on a photographic plate obtained at Las Campanas.

For further information see
http://www.ociw.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lassell, William (1799–1880)

Born in Bolton, England. A wealthy businessman (a Liverpool brewer), Lassell built progressively larger reflecting telescopes. His innovation was to have them made not in the Herschelian tradition of wood and altazimuthally mounted but with massive, strong iron frames, equatorially mounted on high-quality bearings. He commissioned engineer/astronomer JAMES NASMYTH, to make his 24 in telescope, with a fine machine-figured 500 lb speculum mirror. According to family tradition, Lassell missed using the telescope, erected at his house 'Starfield', to discover Neptune, a letter communicating ADAMS' predicted position having been accidentally destroyed by a too zealous maid. When the coordinates of 'Leverrier's planet', discovered 23 September 1864, were published in *The Times* Lassell found it and saw that not only did Neptune reveal a distinct disc, but it had a satellite, Triton. Lassell also thought it had a ring, like Saturn, but this turned out to be a feature of a maladjusted mirror support system. Lassell went on to discover two moons around Uranus (Ariel and Umbriel) and co-discovered with BOND the eighth moon of Saturn (Hyperion). He moved his 24 in telescope to the clear skies of Malta and built his 48 in telescope there, making a particularly fine study of the Orion nebula. The success of this instrument inspired later generations of astronomers to travel to prime locations (like La Palma and Hawaii) to set up their observatories. The 24 in telescope is being reconstructed in Liverpool.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lasser, David (1902–)

Science fiction editor and union organizer, born in Baltimore, MD, founded the American Interplanetary Society, the first organization in the USA to deal with space travel. Wrote an influential book, *The Conquest of Space* (1931). Derided in Congress as 'not only a radical but a crackpot, with mental delusions we can travel to the Moon'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lavanha, Joao Baptista (1550–1624)

Navigator, born in Portugal, became chief engineer to Philip II, and taught mathematics to sailors, publishing a text *Regimento Nautico*, and developing instruments for navigation, including astrolabes, quadrants and compasses.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

LDEF (Long Duration Exposure Facility)

Large NASA satellite deployed from the Space Shuttle on 7 April 1984. Intended to expose materials and experiments to open space for one year. Recovery was delayed by the Challenger disaster until 11 January 1989. Carried 57 science and technology experiments to measure the effects of atomic oxygen, space radiation, micrometeoroids, man-made debris, vacuum and other space-related phenomena on more than 10 000 test specimens. Flew in a gravity-gradient stabilized orbit, which gave unique information on the nature and direction of approach of cosmic debris. One meteoroid experiment recorded the direction and precise time of occurrence of more than 15 000 impacts. Also the first satellite to detect beta meteoroids—minute particles accelerated by solar radiation. Impacts on LDEF's trailing edge showed for the first time that debris particles exist in elliptical orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Le Fevre [Le Fèvre; LeFebvre], Jean (1652–1706)

Born in Lisieux, France, became an associate of JEAN PICARD and PHILIPPE DE LA HIRE, became a member of the Académie des Sciences and worked on rather routine work calculating ephemeredes, and surveying. Le Fevre accused La Hire of stealing his tables after Philippe de La Hire published *Tabulae Astronomicae* in 1687. In 1701, resentful at not having been named official publisher of ephemerides to the Académie, Le Fevre published a preface to *Connaissances des Temps* attacking Philippe's son, GABRIEL-PHILIPPE DE LA HIRE. The preface was replaced on government orders and Pontchartrain, La Hire's protector, led a campaign that resulted in Le Fevre being expelled from the Académie.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Le Gentil de la Galaziere, Guillaume-Joseph-Hyacinthe-Jean-Baptiste (1725–92)

Discovered M32, the Andromeda galaxy's companion galaxy, also discovered the gaseous nebula M8, the Lagoon nebula (independently of FLAMSTEED) and M36 and M38 (see ODIERNA).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leavitt, Henrietta Swan (1868–1921)

Astronomer, born in Lancaster, MA. Became a volunteer and then staff member of the Harvard College Observatory. Directed by PICKERING to work on variable stars in the Magellanic Clouds, measuring their brightness on repeated exposures on photographic plates. Set up standards of photographic magnitudes (brightness) of stars that were accepted worldwide under the name of the 'Harvard Standard', and could be used as comparisons to measure the magnitudes of other stars. Discovered the relation between the period and luminosity of classical Cepheid variable stars. This period–luminosity (P–L) relation made possible measurements of the distances of clusters and galaxies from the Earth and thus the scale of the universe. Leavitt's P–L relation was re-calibrated and used by the Hubble Space Telescope to determine the Hubble constant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Legendre, Adrien-Marie (1752–1833)

Mathematician, born in Paris (or Toulouse), France, into a rich family. Studied mathematics, and carried out independent research studying the gravitational attraction of ellipsoidal masses (by which could be modeled rotating planets or stars), during which he introduced what are now known as *Legendre functions*. This work brought him membership of the Académie des Sciences. He continued research on celestial mechanics and mathematics, and worked on a triangulation survey between the Paris and Greenwich observatories to measure the Earth. He was a member of the committee of the Académie des Sciences to standardize weights and measures. He lost his wealth in the Revolution in 1793, but nevertheless, with the support of his new wife, he continued scientific work. He published a book describing how to determine the orbit of a comet, under simplifying assumptions, and giving the least squares method of fitting a curve to the data. GAUSS reinvented the least squares method in 1809, acknowledging that it had appeared in Legendre's book, but still claiming priority, something which Legendre found hard to forgive. He made a political blunder in 1824 by not supporting the government's candidate in an election, lost his pension and died in poverty.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leiden Observatory

The Leiden Observatory (Sterrewacht Leiden) was founded in 1633 and is the oldest university astronomy department in the world. It is part of the Faculty of Mathematics and Natural Sciences of Leiden University.

The research program at Leiden concentrates on the formation and evolution of stars and galaxies. This research makes use of international optical, infrared submillimeter and radio telescopes, as well as theoretical analysis and interdisciplinary laboratory experiments. The goal of the educational program is to train students as independent researchers and problem solvers for careers both as professional astronomers and as leaders of industry. The educational program is broad and accessible to English and/or Dutch speaking students.

Leiden Observatory is an internationally oriented institute, having diverse collaborations with astronomers in Europe, the USA and elsewhere. The institute is a member of the European Association for Research in Astronomy (EARA), a network of top institutes in Cambridge, UK, Paris, Munich and Tenerife, and of the Netherlands Research School for Astronomy (NOVA). It hosts the Sackler Laboratory for Astrophysics and the NOVA-ESO Expertise Center for VLTI (a center for optical interferometry).

During the twentieth century Leiden Observatory has been associated with and produced many outstanding astronomers, including Willem De Sitter, Ejnar Hertzsprung, Jan Hendrik Oort and Hendrik van de Hulst. Leiden Observatory was the cradle of Dutch radio astronomy, where the Dwingeloo and Westerbork radio telescopes were designed and developed. In 1944 van de Hulst predicted that galactic neutral hydrogen should radiate an observable spectral line at 21 cm in the radio. Subsequently 21 cm observations became a fundamental tool in galactic and extragalactic research. The Netherlands Foundation for Radio Astronomy was founded in Leiden by Jan Oort and was located at Leiden Observatory until 1974, when it was moved to Dwingeloo.

At present Leiden Observatory has a permanent scientific staff of seventeen, four technical support staff and two administrative support staff. In addition, there are about eleven post-doctorate positions, thirty-two PhD candidates and fifty undergraduate students.

For further information see
http://www.strw.leidenuniv.nl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lemaître, Abbé Georges Edouard (1894–1966)

Priest, civil engineer and astrophysicist, born in Charleroi, Belgium, became professor of the theory of relativity at Louvain, researched on cosmic rays and the three-body problem. Proposed (1927) independently of FRIEDMANN, an evolving cosmological model in general relativity. This mathematical solution to the EINSTEIN equations, sought but not discovered by Einstein, indicated that the universe had begun in a 'Big Bang', and originated from a dense mass concentration which Lemaître envisaged as a primeval atom. HUBBLE's discovery of the expansion of the galaxies immediately gave credibility to this theory of the origin of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lemonnier, Pierre Charles (1715–99)

Astronomer, born in Paris, surveyed the figure of the Earth in the Arctic Circle, and made 12 pre-discovery observations of Uranus but didn't follow up his observations and thus missed recognizing it as a planet.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lenticular Galaxy

A lens-shaped galaxy that has a central bulge and disk but no spiral arms. Lenticular galaxies, which are denoted by 'S0' or 'SB0' in the Hubble classification scheme, appear to be intermediate in type between elliptical and spiral galaxies. They are flatter than the flattest ellipticals ('E7' ellipticals) but show no signs of spiral structure in their disks. It is possible that they may be spiral galaxies that have been stripped of their gas and dust.

See also: barred spiral galaxy, elliptical galaxies, Hubble classification, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leo

(the Lion; abbrev. Leo, gen. Leonis; area 947 sq. deg.) A northern zodiacal constellation which lies between Cancer and Virgo, and culminates at midnight in early March. It represents the Nemean lion that, in Greek mythology, Hercules killed as the first of his 12 labors. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large, conspicuous constellation, Leo is easily recognized by the asterism of the Sickle, a back-to-front question mark formed by six of its brightest stars. At the base of the Sickle is α Leonis (Regulus), magnitude 1.4. The next brightest stars in the constellation are β Leonis (Denebola), magnitude 2.1, γ Leonis (Algieba), a binary with orange (K1) and yellow (G7) components, magnitudes 2.5 and 3.6, separation 4.6″, period about 600 years, and δ Leonis (Zosma), magnitude 2.6. There are eight other stars of magnitude 4.0 or brighter. Other interesting stars include R Leonis, a Mira-type variable (range 4.4–11.3, period about 310 days), X Leonis, a U Geminorum star (range 11.1–15.7, mean period about 16.9 days), and Wolf 359 (CN Leonis—a UV Ceti star, range 11.5–17.1), which, at a distance of 7.8 light-years, is the third closest star to the Sun.

Other interesting objects in Leo include the ninth-magnitude spiral galaxies M65 (NGC 3623), M66 (NGC 3627), M95 (NGC 3351), M96 (NGC 3368) and M105 (NGC 3379), the first two and latter three of which appear close together in the sky. There are also many fainter galaxies, including the dwarf spheroidal galaxies Leo I (tenth magnitude) and Leo II (twelfth magnitude), which are members of the Local Group.

The Leonid meteor shower appears to radiate from close to γ Leonis.

See also: Leonids, Regulus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leo Group

The constellation LEO contains a rich concentration of galaxies, including five Messier objects. The Leo Group is the nearest group of galaxies with both giant ellipticals and spirals, which makes it valuable for tests of the extragalactic distance scale. The region also harbors two remarkable intergalactic features: a long stream of stars and gas that has been pulled from a SPIRAL GALAXY by gravitational tides and an enormous ring of intergalactic gas the origin of which remains a puzzle.

Nomenclature, extent and mass

The distribution of galaxies in Leo reveals a significant enhancement of bright galaxies within the boundaries shown in figure 1. These galaxies also lie within a narrow REDSHIFT range: more than 100 have measured redshifts in the range 500 km $s^{-1} < cz < 1700$ km s^{-1}, while none has a smaller redshift and only two have redshifts up to 1000 km s^{-1} higher. This concentration of galaxies between about 7 and 25 Mpc is usually called the Leo Cloud, which is a combination of the nearer Leo I Cloud (sometimes described as a 'spur' off of the supergalactic plane) and several groups constituting the more distant Leo II Cloud. The Leo I Cloud can be subdivided further into two groups of galaxies usually labeled by their brightest members as the 'M66 Group' and the 'M96 Group' (see MESSIER OBJECTS). The term 'Leo Group' sometimes refers to a combination of the M66 and M96 groups, but in more recent literature it generally refers to the M96 Group alone. The M66 Group is also frequently called the 'Leo Triplet' because of the three bright galaxies at its core.

Since the distances to most galaxies have not been measured with high precision, it is often difficult to determine the physical relationship of galaxies seen close together on the sky. Early group-identification methods had only a small number of redshifts available, so they relied heavily on visual clues such as the angular sizes of the galaxies and of the morphological features within them. These methods yielded the basic breakdown of groups in the region that is described above and which has been confirmed by independent distance measurements as described later.

Redshift information actually confused questions of membership for a period of time because of the problem of separating a galaxy's Hubble expansion redshift from its Doppler shift due to local dynamical motions. For example, percolation and hierarchical group-identification methods search for neighboring galaxies on the basis of their angular and redshift separation, building chains of associations from which groups are identified. These methods become confused by the Leo region's high density of sources because nearer galaxies with a local dynamical velocity directed away from us and more distant galaxies with a velocity toward us can have similar total redshifts.

The distance-determination problem is exacerbated in the Leo region because of its proximity to the VIRGO CLUSTER.

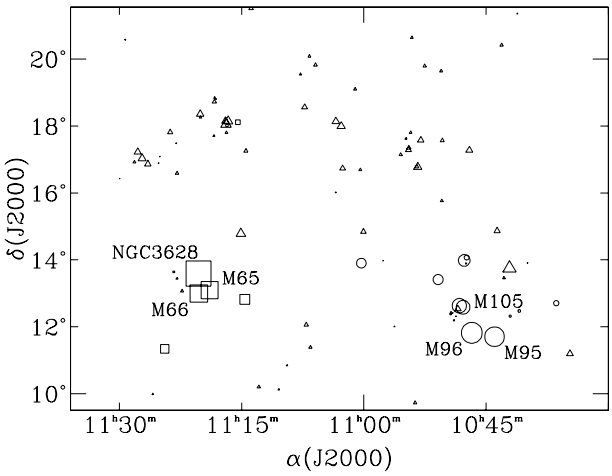


Figure 1. Galaxies in the Leo region with redshifts 500 km $s^{-1} < cz < 1700$ km s^{-1}. Circles indicate likely members of the M96 Group; squares, the M66 Group (Leo Triplet); triangles, members of the background Leo II Cloud. Symbol sizes are proportional to the angular diameter of the galaxy.

While the Leo Cloud is outside of the Virgo Cluster's turnaround radius, the cosmic expansion in the region has been slowed so that the range of redshifts is smaller than would be expected for the depth of the region.

Depending on which galaxies are assumed to be members of a group, estimates of the group mass vary dramatically. Mass-to-light ratios ranging from about $30M_\odot/L_\odot$ to $300M_\odot/L_\odot$ have been claimed for the Leo Group. The highest values result when galaxies from the Leo I and II Clouds are grouped together. However, independent distance measurements indicate that the Leo I and II Clouds are at distinctly different distances. Based on narrower definitions for inclusion in groups, the lower mass-to-light ratios appear more accurate.

Distance laboratory

Independent methods for determining distances are important in extragalactic astronomy because local dynamical motions may make expansion distances inaccurate. Groups of galaxies provide an important testing ground for distance methods, allowing examination of their reliability and accuracy. The Leo Group has been a special focus of distance studies because it includes early- and late-type galaxies and is near enough to permit detailed observations. In addition, M105 (NGC 3379) has been studied heavily as a prototype of surface-brightness distributions within elliptical galaxies, and both M95 and M96 are part of the HUBBLE SPACE TELESCOPE key project on the extragalactic distance scale. M96 even cooperated with the distance studies by producing a type Ia supernova in 1998.

There is a serious difficulty in comparing and cross-calibrating distance methods. Most are applicable to either early- or late-type galaxies, but galaxies tend to separate morphologically, with early-type galaxies residing in dense clusters, and late-type galaxies in the field. Thus the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Distances determined for the five major galaxies in the M96 group. The method is labeled in the left-hand column, and the distances in Mpc are listed under each galaxy. The galaxies' morphological types and absolute magnitudes are listed below their NGC and Messier catalog numbers.

Distance technique	M96 NGC 3368 SAB(rs)ab −20.8	M105 NGC 3379 E1 −19.8	M95 NGC 3351 SB(r)b −19.7	NGC 3384 SB(s)0$^-$ −19.2	NGC 3377 E5–6 −18.9
Cepheid variables	11.9 ± 0.8		10.1 ± 0.9		
Type Ia supernovae	10.4 ± 0.8				
SB fluctuations		9.4 ± 0.7		9.5 ± 0.8	9.7 ± 0.8
PN luminosity function	9.6 ± 0.7	10.1 ± 0.7		10.4 ± 0.7	10.6 ± 0.8
Red giant branch tip		11.0 ± 0.7			
Tully–Fisher method	10.2 ± 1.4		9.8 ± 1.4		
Faber–Jackson method		8.7 ± 1.2		9.0 ± 1.7	12.6 ± 2.8

diverse collection of galaxies in the M96 Group provides an unusual opportunity to cross-calibrate distance methods.

The distance of the M96 Group also presents a way for estimating the HUBBLE CONSTANT. Since M105 is a giant elliptical, the same distance method can be used for it as for ellipticals in distant clusters, and an accurate *relative* distance determined. The M96 Group is itself too nearby and its redshift too strongly influenced by the Virgo cluster to provide an accurate expansion/distance ratio directly. Relative distances determined for member elliptical galaxies in the COMA CLUSTER indicate that it is approximately 9.5 times farther away than M105. Thus the M96 Group distance implies a Coma Cluster distance, and the Coma Cluster redshift provides an accurate measure of the expansion. Current estimates of the relative distance of the M96 Group and Coma Cluster indicate that $H_0 \approx 750$ km s$^{-1}/d_{M96}$.

Table 1 lists some of the distance methods that have been applied to the five primary galaxies in the M96 Group. The uncertainties associated with each distance are not all determined by consistent procedures and are probably only indicative of the relative accuracy within that distance method. Overall, the distances indicate that the group is at approximately 10.2 ± 0.2 Mpc based on a weighted average of these measurements.

Cepheid variables permit some of the most accurate distance determinations using the well-calibrated periods and luminosities of Cepheids in the Local Group (see CEPHEID PERIOD–LUMINOSITY RELATION). The value of M96's distance using this method is slightly discrepant with other distance estimates. Since only seven Cepheids were measured in this galaxy and reddening corrections are uncertain, this value may not be as accurate as expected.

The type Ia supernova 1998BU in M96 provides a direct link to another high-accuracy distance standard. Type Ia SUPERNOVAE are important because their peak brightness is an excellent 'standard candle' that can be used over cosmological distances. However, few of these supernovae have been seen in nearby galaxies that have other high-quality distance estimates. One of the difficulties in estimating this type of distance is determining the extinction due to foreground dust. The reddening of this supernova is quite large, producing about 1.66 magnitudes of extinction at B-band. Assuming a peak SN Ia absolute magnitude of $B = -19.38$ yields a distance in excellent agreement with most other methods, although slightly lower than for the Cepheids in M96. It should be noted that different research groups estimate a range of peak SN Ia magnitudes both brighter and fainter than this value.

The surface-brightness (SB) fluctuation technique uses the statistical variation in the number and brightness of stars along different lines of sight through a galaxy to estimate its distance. The values in the table are determined from ground-based observations. Individual observers generally derive consistent results for galaxies within the group, but discrepancies are present between different observers and between ground-based and Hubble Space Telescope determinations. The uncertainties listed in the table partially account for these discrepancies. Recent re-examination of these and other galaxies in the M96 Group indicates that the mean group distance determined using this method is probably closer to 10.7 ± 0.3 Mpc.

The PLANETARY NEBULA (PN) luminosity function uses the observed steep cutoff in PN counts at high luminosities to estimate a distance. The luminosity of the cutoff is effectively a standard candle. The technique has been applied to the early-type galaxies in the group, as well as the bulge of M96, with excellent consistency. Another good luminosity standard has been detected in M105 with the Hubble Space Telescope: stars at the tip of the red giant branch. This provides another independent distance in excellent agreement with the other methods.

Table 1 also lists older results using less accurate methods, which are nonetheless important because of their wide applicability. The TULLY–FISHER method relates spiral galaxies' rotation speed to their luminosity; similarly, the Faber–Jackson method relates the velocity dispersion in early-type galaxies to their luminosity. These values are less accurate, but their overall agreement with the other methods is very good.

The distance to the M96 Group has also been estimated using the globular cluster luminosity function,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Other probable members of the M96 Group.

Name	Type	Absolute magnitude
NGC 3489	SAB(s)dm	−18.9
NGC 3412	SB(s)0^0	−18.7
NGC 3299	SAB(s)dm	−16.9
DDO 88	SAB(s)m	−15.9
UGC 5812	Sm	−15.2
CGCG 65–86	dS0	−14.3
Leo dw A	Im	−11.1

applied to a combined sample of globular clusters around both NGC 3377 and NGC 3379, yielding a distance of 10.7 ± 2.2 Mpc. Several other methods have also been applied, using the Leo Group to calibrate the technique, and hence the distances are not independent of the values listed here.

In addition to the five galaxies listed in table 1, several fainter galaxies currently believed to be part of the M96 Group are known at present, including one extremely faint dwarf (Leo dw A) discovered during mapping of hydrogen gas in the group. Table 2 lists these other probable members.

The angular distance between the M66 and M96 Groups indicates they are separated by no less than 1.5 Mpc, so they are probably not part of the same group. Furthermore, the M66 Group is usually estimated to be about 25% closer than the M96 Group based on comparisons of the Tully–Fisher distances; however, tidal disturbances in the Leo Triplet galaxies may degrade the accuracy of these estimates. The distance estimate for the type Ia supernova, SN1989B, in M66 (NGC 3627), suggests the distance may actually be similar to the M96 Group's.

A type Ia supernova (1967C) has also been detected in NGC 3389 in the background Leo II Cloud. This galaxy has a redshift ($cz = 1300$ km s^{-1}) only slightly larger than that of the M96 Group (mean $cz = 950$ km s^{-1}) and is separated by only a few arcminutes from M105 and NGC 3384. The angular and redshift proximities were suggestive that it might be a member of the M96 Group, but the supernova distance is more than 2 times larger than the M96 Group's distance.

Intergalactic features

The Leo I Cloud contains two remarkable intergalactic features. The M96 Group contains intergalactic hydrogen loosely distributed along an elliptical orbit around M105 and NGC 3384, which are near the center of the group (figure 2). This 'ring' of intergalactic gas is about 225 kpc in diameter and contains about $2 \times 10^9 M_\odot$ of neutral hydrogen. Searches for emission at all other wavelengths have been unsuccessful, leaving open the possibility that this gas is primordial.

The ring of gas was discovered in 1983 by chance during 'blank-sky' calibration of observations made of a galaxy in the region. Initial observations uncovered only the denser part of the ring between M105 and M96. This

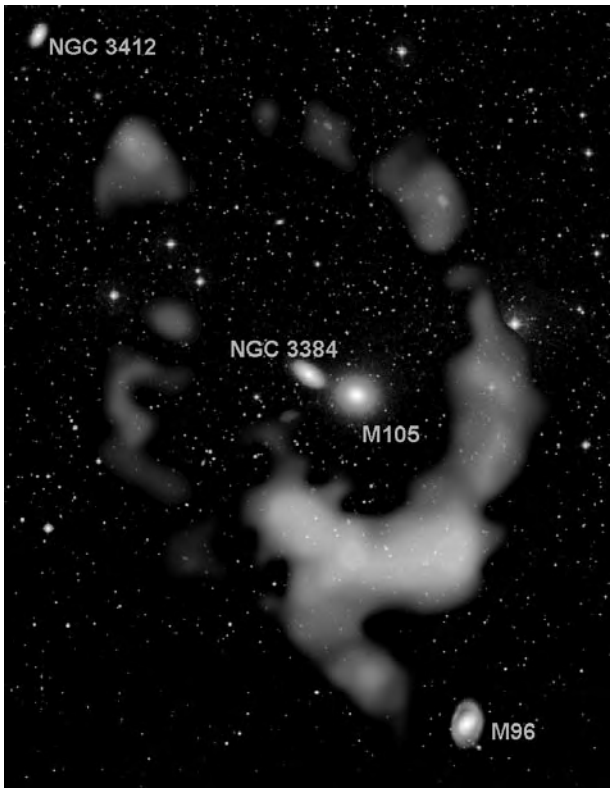


Figure 2. The ring of intergalactic gas in the M96 Group. This is a composite of the optical image from the second Palomar Sky Survey and a gray-scale image depicting the strength of neutral hydrogen emission measured at 21 cm using the Arecibo radio telescope. The intergalactic gas appears to be orbiting the pair of galaxies at the center of the image.

prompted early speculation that the gas might be the result of a Spitzer–Baade collision in which NGC 3384 and M96 collided head-on, stripping NGC 3384 of all of its gas and M96 of a fraction of its gas while the stellar disks passed through each other. Later observations revealed the ring shape of the gas distribution, and the kinematics of the gas indicates it is in an elliptical orbit around the pair of galaxies M105 and NGC 3384.

A spur of gas extends south from the ring toward M96, suggesting a tidal interaction with M96. There is no direct indication from optical or radio observations that the outer portion of M96 is tidally disturbed, so it seems unlikely that the intergalactic gas was stripped from it. Alternatively, the intergalactic gas may come from a disrupted gas-rich dwarf. However, for the gas to become distributed at points all around the ring by differential rotation, several orbits after the initial event would be required. Since the orbital period for the ring is about 4×10^9 yr, this timescale approaches the age of the universe, so it remains plausible that this gas is remnant material from the time of the galaxies' formation.

Another extragalactic feature is found in the M66 Group. In 1956, ZWICKY noted a faint plume of optical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

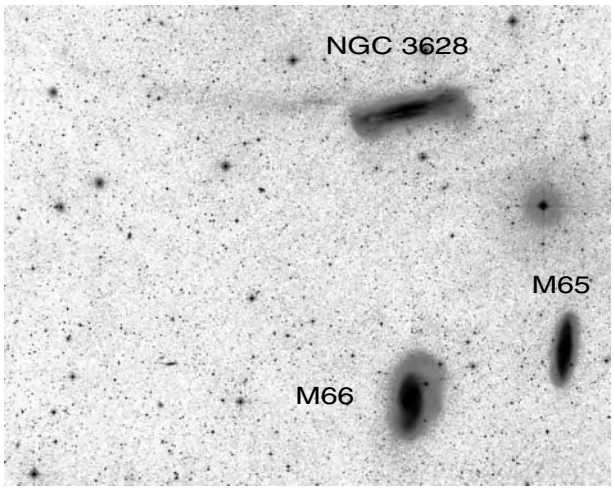


Figure 3. The Leo Triplet of galaxies. Kinematic models indicate that M66 passed by NGC 3628, causing a long tidal plume to be drawn out of NGC 3628. The plume extends east of NGC 3628 almost 100 kpc. This negative of the optical image is from the second Palomar Sky Survey; it is highly 'stretched' to bring out low surface brightness features.

emission outside NGC 3628 (figure 3). At the time, there was considerable interest in the possibility that these features were galaxies in the process of formation. Modern observations and models suggest instead that such features are produced when galaxies suffer close encounters.

The plume extends about 100 kpc east of NGC 3628 and is one of the largest tidal features known. Studies at various wavelengths indicate the presence of stars, gas and dust in similar proportions as would be found in the outer disk of a spiral galaxy. Models indicate that M66 passed close by NGC 3628, and the encounter was in a prograde direction for NGC 3628 but retrograde for M66. As a result, the outer material in NGC 3628 was sped up into much larger orbits while that in M66 became more tightly bound. Both galaxies show morphological indications of disturbance in their outer regions where the low surface brightness emission is distorted. The far end of the plume appears to have a large enough velocity to become unbound from NGC 3628 and may eventually have an appearance similar to a dwarf galaxy.

Bibliography

The groups in the Leo region were originally outlined by

de Vaucouleurs G 1975 Nearby groups of galaxies *Galaxies and the Universe* ed A Sandage *et al* (Chicago, IL: University of Chicago Press) pp 557–600

de Vaucouleurs' divisions have held up remarkably well through the years even though there was relatively little distance or redshift information when the chapter was written in 1965.

The distances of Leo Group galaxies are the subject of intensive research at present. A 1997 paper that reviews many of the techniques employed is

Graham J A *et al* 1997 The Hubble Space Telescope extragalactic distance scale key project. VII. The discovery of Cepheids in the Leo I Group galaxy NGC 3351 *Astrophys. J.* **477** 535

Work on the 1998 supernova in M96 is not yet published, but several new papers re-examining the distance scale in Leo can be expected.

Discoveries about the intergalactic ring of gas in Leo are reviewed by

Schneider S E 1989 Neutral hydrogen in the M96 Group: the galaxies and the intergalactic ring *Astrophys. J.* **343** 94

A popular article from 1984 discusses early investigations of the intergalactic ring and other intergalactic features, including the tidal feature in the Leo Triplet:

1984 Between the galaxies *Am. Sci.* **72** 574

Stephen E Schneider

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leo Minor

(the Lesser Lion; abbrev. LMi, gen. Leonis Minoris; area 232 sq.deg.) A northern constellation which lies between Ursa Major and Leo, and culminates at midnight in late February. It was introduced by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* of 1687.

A small, inconspicuous constellation, the brightest stars in Leo Minor are 46 Leonis, magnitude 3.8, and β Leonis, a very close binary with yellow (G8 and F8) components, magnitudes 4.4 and 6.1, separation 0.2″. Interesting objects include NGC 3245, an eleventh-magnitude lenticular galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leonardo da Vinci (1452–1519)

Painter, inventor and polymath, born in Vinci (near Empolia), Italy. Although astronomy does not figure large in Leonardo's works, he realized the possibility of constructing a telescope ('making glasses to see the Moon enlarged'). He suggested that '... in order to observe the nature of the planets, open the roof and bring the image of a single planet onto the base of a concave mirror. The image of the planet reflected by the base will show the surface of the planet much magnified'. He understood how the Moon shone by reflected sunlight and he correctly explained the 'old Moon in the new Moon's arms' as reflected light from the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leonids

A periodic meteor shower that takes place in November. The radiant lies in the constellation Leo. Some debris from the parent comet, 55P/Tempel–Tuttle, has spread around the orbit to produce a meteor stream that gives modest activity, with an average zenithal hourly rate of 10 in most years. But the greatest concentration of meteoroids remains near the comet. So when the comet returns to perihelion every 33 years there is potential storm activity, particularly if the Earth passes the descending node of the stream's orbit shortly after the comet has. In 1966, for example, some observers estimated that rates were as high as 200 000 per hour in one 40 min period. Not every return of Tempel–Tuttle produces storm activity. Following the great storms of 1799, 1833 and 1866, there was little more than average activity in 1900 and 1933. In 1999 the shower peaked at around 1500 per hour over a 2 h period. The main Leonid swarm (sometimes called the *ortho-Leonids*) is thought to consist of a number of dense ribbons of meteoroids in a sheaf-like configuration. A storm would then occur only if the Earth intersects one of these ribbons. Tempel–Tuttle's retrograde orbit gives the incoming meteoroids a high relative velocity, producing very fast meteors. There are records of Leonid storms going back over a thousand years.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leptons

Particles that are not acted on by the strong nuclear force and which are not built of quarks. Leptons have no internal structure and cannot be broken down into smaller component particles. The leptons are the electron, the muon, the tau (or tauon), and the three types of neutrino, together with their antiparticles. The muon and tau are negatively charged particles that appear to be heavier analogs of the electron, the mass of the muon (105.6 MeV) being 200 times that of the electron, and the mass of the tau (1.78 GeV) being about 3500 times greater than that of the electron. Neutrinos are particles with zero electrical charge and zero, or exceedingly small, mass. The electron and its associated neutrino (the electron-neutrino), together with two types of quark (the 'up' and 'down' quarks), and their antiparticles, constitute a family of particles called the 'first generation'. The muon and the muon-neutrino form part of the second generation, the tau and tau-neutrino being members of the third generation.

See also: electron, electronvolt, elementary particles, fundamental forces, fundamental particles, mass, neutrinos, quark, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lepus

(the Hare; abbrev. Lep, gen. Leporis; area 290 sq. deg.) A southern constellation which lies between Orion and Columba, and culminates at midnight in mid-December. It is sometimes said to be the quarry of Canis Major, one of the two dogs of Orion (the Hunter). The brightest stars of Lepus were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small constellation, the brightest stars in Lepus are α Leporis (Arneb), magnitude 2.6, and β Leporis (Nihal), a close binary with a yellow (G5) primary, magnitude 2.8, and a much fainter, magnitude 7.3, companion. There are six other stars brighter than fourth magnitude, including the double star γ Leporis, which has pale yellow (F6) and orange (K2) components, magnitudes 3.6 and 6.2, separation 97″, that share the same common proper motion and radial velocity. Other interesting objects include R Leporis, a Mira-type variable star (range 5.5–11.7, period about 427 days), sometimes called 'Hind's crimson star' after the British astronomer John Russell Hind (1823–95), who remarked on its deep red color, NGC 2017, a small group of six associated stars (magnitude 6.4–12.4), two of which are binaries, IC 418, a ninth-magnitude planetary nebula, and M79 (NGC 1904), an eighth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lescarbault, Edmond

French country doctor living in Orgenes. Claimed that on 26 March 1859 he had watched for an hour the passage of something across the solar disc that matched URBAIN LEVERRIER's calculations for the hypothetical planet Vulcan. Leverrier, co-predictor with ADAMS of Neptune from the deviations of Uranus from its orbit, tried to repeat his earlier success by attributing deviations of Mercury from its orbit to an intra-Mercurial planet, Vulcan, or more than one planet. MAX WOLF had found photographs of a number of circular spots on the Sun, which could also be interpreted as isolated observations of solar transits of a planet inside the orbit of the Earth. LEVERRIER investigated Lescarbault's claims and was shown some figures recorded on prescription paper, covered with grease and laudanum. Surprisingly, he accepted the observations and assimilated them into an orbit of Vulcan. He mobilized French astronomers into a search for Vulcan during the solar eclipse of 1860—no intra-Mercurial planets were found. Other round spots on the Sun were independently photographed at Greenwich and Madrid, at about the time of a possible transit of Vulcan, but when Leverrier predicted that there would be a transit of Vulcan across the Sun on 22 March 1877 nothing was seen. J C Watson and Lewis Swift (1820–1913) claimed to see small planets (at different positions) near the Sun during the total solar eclipse of 1878 but at the 1929 eclipse ERWIN FREUNDLICH found nothing. In 1916, ALBERT EINSTEIN published his general theory of relativity, which explained the deviations in the motions of Mercury as a non-Newtonian effect called the advancement of perihelion, and the need for Vulcan disappeared. The reported sightings were misidentified sunspots (Wolf, Weber), misidentified stars (Watson and Swift) and accidental sightings of small asteroids or Sun-grazing comets passing inside Earth's orbit (Lescarbault).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leucippus ($fl.$ fifth century BC)

Philosopher, born in Miletus, Asia Minor. Virtually nothing is known of his life, except that he was the founder of Atomism, developed by his disciple DEMOCRITUS and accessibly presented by LUCRETIUS in *De Rerum Natura*. Essentially Atomism is a theory to explain change. Atoms are microscopic particles and are permanent. They move in the permanent framework of space and alter their arrangements—appearances change. It is a recognizably scientific theory. Leucippus is credited with two books, *The Great World System* and *On the Mind*, but his theories and writings are not reliably separable from those of Democritus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leuschner, Armin Otto (1868–1953)

Astronomer and teacher, born in Detroit, grew up in Germany, and returned to the USA to lead the astronomy department of the University of California at Berkeley. His research was in celestial mechanics and he calculated orbits of asteroids and comets, developing a technique that succeeded even when based on observations of a rapidly moving Earth-grazing object. He was highly regarded as a teacher.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Leverrier [Le Verrier], Urbain Jean Joseph (1811–77)

Celestial mechanician, born in Saint-Lô, France. Worked at the Paris Observatory, under ARAGO. Later became director where his drive for efficiency and for total control over the staff made him unpopular, particularly his diktat that 'One should not read in the papers the names of assistant astronomers who made discoveries, when the credit belongs exclusively to the director under whose orders they work. Young astronomers should of course receive thanks and a medal for each discovery'. Staff unrest was brought to the attention of the government, and Leverrier was removed from the post. He was later reinstated under the authority of a directing council. Leverrier had found fame as a young man when Arago had suggested that he calculate the position of Neptune from irregularities in Uranus's orbit (*see also* ADAMS). Neptune was found at the predicted place by the German astronomer GALLE. Leverrier attempted to repeat his success with Neptune by examining a discrepancy in the motion in the perihelion of Mercury, and attributed this to an intra-Mercurial planet, which he called Vulcan (*see* EDMOND LESCARBAULT).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Levi ben Gerson (1288–1344)

Philosopher, therologian and astronomer, born in Bagnols, Gard, France. Invented Jacob's staff, an instrument to measure the angles between stars. It is a staff with square cross-section, about 1 m long, with perforated cross-pieces which slide along the staff. The end of the staff is held at the eye and the cross-piece slid along until the two ends fit within the two stars being measured. The staff is calibrated and engraved with the tangent of the angle of separation, so that the angle can be read from the position of the cross-piece. Levi observed solar and lunar eclipses in 1337, giving explanations, and made other astronomical observations of the Moon, Sun and planets using a camera obscura. He wrote that the Milky Way shines by the reflected light of the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lexell, Anders John (1740–84)

Finnish astronomer, born in Abö, became professor of mathematics at St Petersburg, one of the first to prove that Uranus was a planet (rather than a comet), by computing its orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ley, Willy (1906–69)

Rocket scientist, writer, born in Berlin, Germany. Inspired by reading a work by the space pioneer, HERMANN OBERTH, Ley founded the German Society for Space Travel (1927), enrolled WERNHER VON BRAUN, and helped develop the liquid-fuel rocket. Fled to the USA, and became a science writer, including science fiction and film scripts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Libra

(the Scales; abbrev. Lib, gen. Librae; area 538 sq. deg.) A southern zodiacal constellation which lies between Virgo and Ophiuchus, and culminates at midnight in early May. The brightest stars of Libra were regarded by the ancient Greeks as comprising the constellation Chelae (the Claws—of the Scorpion) and were cataloged as such by Ptolemy (*c.* AD 100–175) in the *Almagest*. The name Libra was introduced by the Romans in the first century BC to reflect the fact that, at that time, the Sun was in Libra at the autumnal equinox, when day and night are of the same length (in balance), though this idea may date back to the Sumerians of 2000 BC. Libra is the only inanimate object among the zodiacal constellations.

A rather inconspicuous constellation, the brightest stars in Libra are β Librae (Zubeneschamali, from the Arabic for 'the northern claw'—of the Scorpion), magnitude 2.6, and α Librae (Zubenelgenubi, 'the southern claw'), a wide double star with bluish-white (A3) and pale yellow (F4) components, magnitudes 2.8 and 5.2, separation 231″, that share the same proper motion, and a third, unseen companion which revolves around the former. There are four other stars brighter than fourth magnitude. Another interesting star is ι Librae, a multiple system consisting of two close binaries separated by 58″, one consisting of two bluish-white (B9) components, magnitudes 5.1 and 5.6, separation 0.1″, period 22.4 years, and a third, unseen companion, the other consisting of two much fainter components, of magnitudes 10.5 and 11.2v, separation 1.9″. Other interesting objects include the Algol-type variable star δ Librae (range 4.9–5.9, period 2.33 days) and NGC 5897, a ninth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Libration

An oscillation of a celestial object about some mean position. The term is used most often in connection with the Moon, to describe a number of effects which cause the hemisphere presented to an observer on the Earth to vary slightly as the Moon proceeds around its orbit. Over time, the Moon appears to rock slightly from side to side, and to nod slightly up and down. Although the Moon has synchronous rotation, rotating on its axis in the same time as it takes to orbit the Earth, the different types of libration make it possible for 59% of the Moon's surface to be visible over the course of a period of 30 years or so. The effect of libration is to make surface features near the Moon's limb sometimes extremely foreshortened, sometimes easier to observe.

The main types of libration are those termed *geometric librations*, which arise from the features of the Moon being observed from different angles at different times. The Moon has an elliptical orbit, in which its orbital velocity varies; this, combined with its constant rate of axial rotation, leads to *libration in longitude*. At perigee and apogee, when closest and farthest from the Earth respectively, the Moon presents its average east–west aspect to the Earth, but between these two positions we can see farther round its western or eastern limb. The difference amounts to a wobble of $\pm 7^\circ\, 45'$. *Libration in latitude* arises from the Moon's orbital plane being inclined to the plane of the ecliptic, and to a lesser extent from the slight inclination of the Moon's equator to its orbital plane. For various positions in its orbit we can therefore see more of one or other of the polar regions. This effect amounts to $\pm 5^\circ\, 9'$. A lesser effect is *diurnal libration*, a result of the Earth's rotation changing the observer's position, so that we can see a little more of the eastern or western limb when the Moon is rising or setting. This effect is greatest for an observer on the equator, where it contributes $57'$. As opposed to geometrical libration there is also *physical libration*, a consequence of small variations of the Moon's orbit caused by its slightly irregular shape. Physical libration amounts to just $2'$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lick Observatory

Lick Observatory, located on Mount Hamilton 30 km east of San Jose, California, at an elevation of 1280 m, serves astronomers from throughout the University of California system. It is administered by University of California Observatories/Lick Observatory, which has its headquarters on the Santa Cruz campus. This multi-campus research unit also serves as the University of California liaison with the W M KECK OBSERVATORY. UCO/Lick provides the technical facilities and staff to design and fabricate state-of-the-art instrumentation.

Lick has seven major research telescopes. The original observatory building houses the 110-year-old 36 in (0.9 m) refractor, the second largest in the world, as well as a 1 m reflector. Also on the mountain is a 0.5 m twin astrograph and a 0.75 robotic telescope. The largest telescope is the 3 m Shane reflector.

The research undertaken at Lick spans the diverse interests of University of California astronomers. In 1892, using the 36 in refractor, E E Barnard discovered Jupiter's fifth moon, the first found since Galileo. Today, Lick continues that tradition. The Shane reflector, coupled to the Hamilton high-dispersion echelle spectrograph, has become the world's premier planet finder. Other on-going programs include supernova searches, adaptive optics development and inertial proper motions of stars.

For further information see
http://www.ucolick.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Life on Other Worlds

Life on other worlds refers to the existence of life in any form beyond the Earth. The study of life on other worlds, known today as exobiology, bioastronomy or astrobiology, uses the concepts and techniques of modern astronomy, biology and biochemistry to illuminate the origin, evolution and distribution of life in the universe. Recent historical scholarship has shown, however, that long before the birth of modern science, the question was examined in the context of the 'plurality of worlds' tradition dating back to the ancient Greeks. An important finding of this scholarship is that the idea was often elaborated not in haphazard fashion by marginal figures but in the context of the science (or natural philosophy) of its time, by leading thinkers, who framed their ideas as a serious debate over substantive issues.

During the 20th century life on other worlds has become part of the larger concept of cosmic evolution, the theory that the entire universe is evolving, and that its ultimate end products are not merely planets, stars and galaxies, but life, mind and intelligence. The result is a 'biological universe' rather than a purely physical universe. By its very nature an answer to the question of life on other worlds is extremely difficult. However, because it bears so strongly on the problem of humanity's place in the universe, the answer is one of the most sought-after in the history of science. The debate over life on other worlds is therefore a study of science attempting to answer an extremely important question at the limits of its capabilities. This article covers the history of that debate; the current study of life in the universe is covered in the article on EXOBIOLOGY.

The classical plurality of worlds tradition

The question of life on other worlds began with Greek cosmology, was sustained by medieval commentaries on Aristotle and was transformed by the Copernican theory in the 16th century. In its original meaning the term 'world' meant not a planetary world but the entire *kosmos*, the visible universe that the Greek atomists believed to be a result of the chance collision of atoms and that Aristotle theorized was an ordered set of concentric spheres, with the Earth at the center and the sphere of the fixed stars at its outer periphery. In the Greek atomist view, expounded in the 4th and 5th centuries BC by Leucippus, Democritus and Epicurus, an infinite number of worlds (*aperoi kosmoi*) existed completely beyond the human senses. However, such worlds were not beyond human reasoning; because the infinite number of atoms could not all have been used in our finite world, the atomists deduced that other worlds, some of them inhabited, must be created in like manner. This doctrine was eventually spread throughout Europe by the Roman poet LUCRETIUS (ca 99–55 BC), who reiterated that other worlds must exist 'since there is illimitable space in every direction, and since seeds innumerable in number and unfathomable in sum are flying about in many ways driven in everlasting movement'. The concept of infinite worlds was therefore driven by the physical principles of the atomist cosmology.

In the same way, the physical principles of ARISTOTLE (383–322 BC) were diametrically opposed to an infinite number of worlds. In Aristotle's cosmology the Earth was not only the physical center of the nested spheres, it was also the center of motion. According to Aristotle's doctrine of natural motion and place as expressed in his *De caelo* (On the Heavens), everything in the cosmos moved with respect to that single center—the element earth toward Earth, the element fire away from Earth, while air and water assumed intermediate places. If there were more than one world, the elements of earth and fire would have more than one natural place toward which to move, a physical and logical contradiction. Only one world (in the sense of *kosmos*) could exist.

While the treatise of Lucretius was lost until the 15th century, Aristotle's *De caelo* was transmitted to the Latin West in the 12th century. In the context of Christianity the question of a plurality of worlds (*plures mundi*) was the subject of repeated commentary, most notably in connection with the issue of God's omnipotence. The issue for the Scholastic tradition was whether God could create other worlds if he wished. Gradually over the course of the 13th and 14th centuries the answer evolved from the assertion that God could suspend natural law to create other worlds to the affirmation of the possible existence of more than one center of motion, and therefore more than one world. In reality, however, the medieval Scholastics did not believe that God had created a plurality of worlds.

However, it was not the medieval Scholastic tradition that gave birth to the modern discussion of life on other worlds. That distinction belongs to the heliocentric theory of COPERNICUS, which made the Earth a planet and the planets potential Earths. The *De revolutionibus* (1543) altered the meaning of 'world' from cosmos to planet, where each planet was a center of motion and a world of its own. The question was exactly how similar to the Earth were the planets, an issue that could be explored more scientifically with the invention of the telescope and that has been the driving force for much planetary research ever since. Already in the first announcement of his observations with the telescope in *Siderius nuncius* (Sidereal Messenger, 1610), GALILEO noted that the surface of the Moon was 'not unlike the face of the Earth' and that its bright parts might represent land and its darker parts water. In the same year, KEPLER conjectured that one large and particularly circular lunar crater might be an artificial construction of lunar inhabitants. Bishop JOHN WILKINS, in his *Discovery of a World in the Moone* (1638), made a broader case for lunar inhabitants, and began to address the religious implications.

It did not take long for the question of other solar systems to be broached. As early as 1584 the Dominican monk GIORDANO BRUNO had argued for an infinite number of planetary systems in *De l'infinito universo e mondi* (On the Infinite Universe and Worlds). Although an avowed Copernican, Bruno based his reasoning largely

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

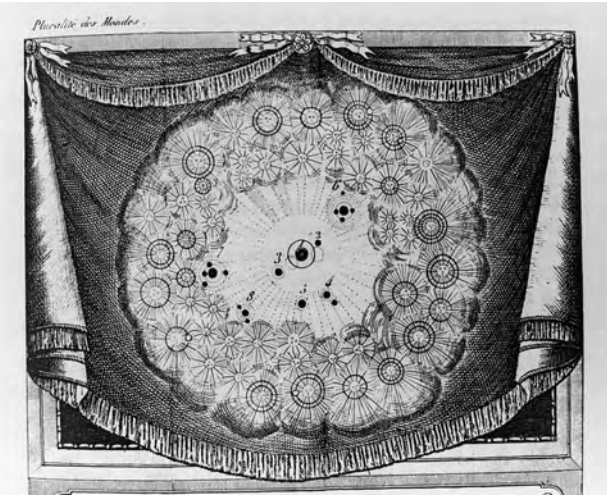


Figure 1. Frontispiece to the 1821 French edition of Fontenelle's *Entretiens sur la pluralité des mondes*, depicting the plurality of solar systems. A similar image adorned the many editions of this work, including the first edition in 1686, graphically spreading the idea of many solar systems.

on metaphysical arguments such as the unity of the universe. His burning at the stake in 1600 by the Roman Inquisition, driven not so much because of his belief in an infinite number of worlds as for his other 'heresies', nevertheless had a chilling effect on those who would speculate on life on other worlds. Thus the French philosopher RENÉ DESCARTES, greatly influenced by a revived atomism, offered in *Principia philosophiae* (Philosophical Principles, 1644) a mechanical philosophy in which a universe full of atoms necessarily formed vortices, systems analagous to our solar system, around every star. Although Descartes himself was careful to avoid speculation about inhabited planets, in his *Entretiens sur la pluralité des mondes* (Conversations on the Plurality of Worlds, 1686) the French Cartesian Bernard le Bovier de Fontenelle spread this idea of a plurality of inhabited solar systems (figure 1). It was also espoused in a similar context in the *Cosmotheoros* (1698) of the Dutch astronomer CHRISTIAAN HUYGENS. Thus it was through the Cartesian cosmology that the idea of other solar systems first became ingrained in European consciousness.

Within 50 years of its inception the Cartesian world view was swept away by the Newtonian system, set forth in NEWTON's *Principia Mathematica* (1687). In the Newtonian world view solar systems did not form by necessity; indeed under Newtonian laws of gravitation the whole question has proved to be one of the greatest complexity until the present day. Although they could not resolve the question of other solar systems by the laws of nature, however, Newtonians were anxious to keep a role for God in the universe, and latched on to the plurality of worlds doctrine as a component of natural theology, arguing that a universe full of life demonstrated the magnificence and omnipotence of God.

During the 18th and 19th centuries little progress was made in characterizing the nature of other planets, and even less in determining whether there might be other solar systems. Much of the discussion may be seen as a struggle of the idea of other worlds with Christianity; although the Newtonians had reconciled other worlds with theism via natural theology, the doctrines of Christianity were another matter. Three choices existed in this struggle: reject other worlds, reject Christianity or reconcile the two. Historically, all three of these possibilities came to pass, as described in detail in Michael Crowe's *The Extraterrestrial Life Debate, 1750–1900* (1986).

The modern tradition: cosmic evolution and life on other worlds

The modern study of life in the universe is intimately connected to the idea of cosmic evolution. Despite the tentative acceptance of Darwin's theory of evolution in the biological realm, at the beginning of the century the universe itself was believed to be static. This was certainly the position of A R Wallace, co-founder with Darwin of the theory of natural selection. His universe, as described in *Man's Place in the Universe: A Study of the Results of Scientific Research in Relation to the Unity or Plurality of Worlds* (1903), was not only static but also only 3600 light years in diameter, with the Earth still situated near the center of the Galaxy. In this view of the universe, shared by many astronomers of the day, Wallace argued that there were no other inhabited worlds and that humanity was the supreme creature. Gradually over the course of the century biological evolution was extended to cosmic evolution (figure 2), whereby the entire universe, some 15 billion light years in extent, existed in a state of constant change ending in abundant life and intelligence. This is one of the most profound changes in world view of the 20th century. The basic tenets (although not the ultimate products) of cosmic evolution have been confirmed by astrophysics, the science whereby the physical nature of celestial bodies may be determined.

The 20th century debate over other worlds focused first on the solar system, specifically the planet Mars. Mars dominated the extraterrestrial life debate for much of the century, progressing from the search for intelligence to searches for vegetation, organic molecules and finally fossils. PERCIVAL LOWELL lauched the first great Martian debate with his claim in 1895 that he had visually detected linear markings on Mars that he interpreted as artificial constructions (canals), built by Martians to channel water on their dying planet. Lowell's Mars was therefore the epitome of planetary evolution. A lengthy debate ensued among astronomers, and not until 1909 did the French astronomer ANTONIADI resolve the 'canals' into dark patches. Not everyone was satisfied even with that evidence, and we know today that very few of the canals have any physical basis. The search for vegetation was the focus from 1924 into the 1960s, based on changes in dark patches on Mars, supported by spectroscopic evidence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

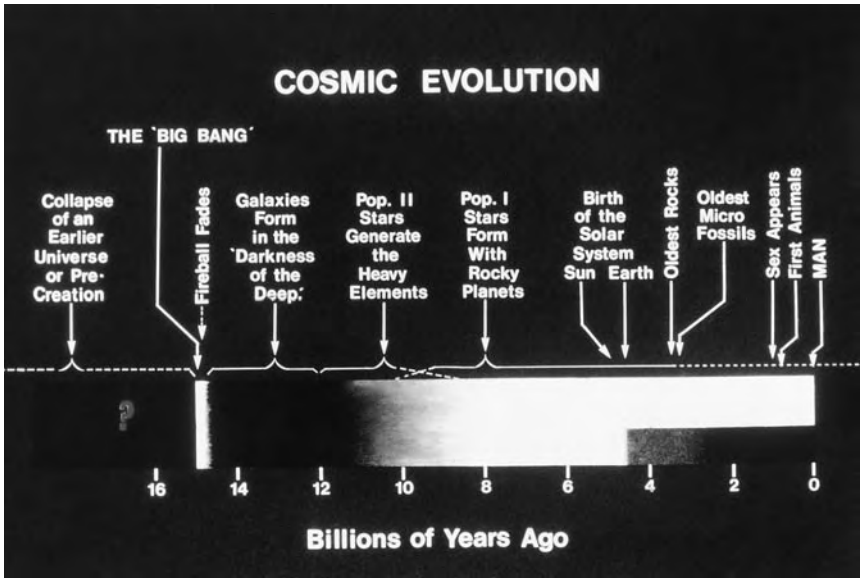


Figure 2. Cosmic evolution, as depicted in this NASA drawing, encompasses events from the Big Bang to intelligence.

in the late 1950s. Such evidence spurred on the space age exploration for life on Mars, culminating with the VIKING landers in 1976. Those landers detected no organic molecules down to parts per billion at the two landing sites. In the wake of this surprising result, the search languished for two decades until the announcement in 1996 of possible fossils in a Martian meteorite found in the Antarctic. Even here, with samples in hand, the evidence was open to interpretation and the sample subject to contamination.

The problem of evidence and interpretation, a constant theme of the 20th century extraterrestrial life debate, is also evident in the search for other solar systems. The 'nebular hypothesis' of Laplace, based on Newtonian gravitation, held that our solar system was formed by condensation from a rotating cloud of gas and dust. If so, solar systems should be a common byproduct of stellar evolution, a process in which the star forms at the center and debris coalesces into planets. This theory was challenged early in the 20th century, first by the Chamberlin–Moulton hypothesis and then the Jeans–Jeffreys hypothesis, both of which had the solar system forming as a result of a very rare encounter between two stars in which material pulled out of the stars formed the planets. However, technical difficulties doomed both theories; by the 1940s a modified nebular hypothesis came into favor, and by the 1950s Harvard astronomer Harlow Shapley could posit the existence of billions of planets. Moreover, the 1940s saw the first observational claims for extrasolar planets, by Kaj Strand and Erik Holmberg, followed by Peter van de Kamp's claims for Barnard's Star in 1963. These observations, based on extremely small gravitational perturbations of the parent star by the supposed planets, were questionable and subsequently proven spurious, and not until 1995 was there confirmation of a planet around a Sun-like star, known as 51 Pegasi. Again, the detection was indirect, using a spectroscopic method to detect the 'Doppler effect' as the star was tugged back and forth by the planet. Since then, more than a dozen planets have been confirmed outside our solar system (see EXOPLANETS). As in the case for Mars, an important lesson is that, despite the difficulties, in the end progress is made.

The existence of planets is only a necessary, not a sufficient, condition for life. Planets must have the right conditions for life to originate, and the origins of life are therefore very important to the question of life on other worlds (see LIFE: ORIGINS AND POSSIBLE DISTRIBUTION IN THE UNIVERSE). It is hampered by our lack of knowledge of how life originated on Earth. During the century three distinct hypotheses have been proposed. Following Darwin's suggestion of life originating in a 'warm little pond', in the 1920s A I Oparin, J B S Haldane and their followers proposed life's origin in a soup of dilute chemicals. The Miller–Urey experiment in 1953, and numerous experiments on prebiotic chemistry under assumed primitive Earth conditions, have been the mainstay of origins of life research during the century. An alternative theory, the panspermia hypothesis, was championed by Arrhenius at the beginning of the century and refined by others at its end. It claimed that life came in some form from outer space, a claim reinforced by the discovery of possible fossil life in the Martian meteorite. At the very end of the century, with the discovery of bacteria existing under extreme conditions of temperature, pressure and acidity, a third theory gained ground: the idea of life originating underground in a 'deep biosphere', and gradually emerging on the surface as conditions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

improved. With such new discoveries, the field of origin of life was poised to make good progress in the 21st century, especially with NASA's renewed committment to astrobiology.

The Search for Extraterrestrial Intelligence (SETI) attempted to leapfrog all other arguments by detecting a signal, usually at radio wavelengths, produced by intelligence. Frank Drake undertook the first such search with an 85 foot radio telescope at Green Bank, West Virginia, in 1960. Since then numerous searches have been undertaken with increasingly sophisticated signal detectors, processing millions of channels simultaneously. Even so, most searches have only reached 100 light years and have utilized only a few of the billions of channels possible. The SETI INSTITUTE's Project Phoenix, University of California's Project SERENDIP and Harvard's BETA Project are three of the most important SETI searches now underway.

The theory that life is abundant in the universe is sufficiently comprehensive to qualify as a world view, with the same status as the Copernican and Darwinian world views. The idea of the 'biological universe', or the 'biophysical cosmology', as such a world view places the debate in proper perspective in the history of science. As with other world views, the biological universe bears heavily on humanity's place in the universe, and therefore arouses passionate debate. Like them, it deals with a basic property of the universe and like them it is testable: the Viking landers were local tests of the world view, and SETI is a more far-reaching test. Finally, like other world views the idea of a biological universe has found expression in popular culture, in this case chiefly in the form of debate over unidentified flying objects, and in science fiction, which incorporates the alien as one of its major themes. As the latest subject in the venerable search for our place in the universe, the question of the biological universe shows no signs of abating.

Bibliography

Crowe M 1986 *The Extraterrestrial Life Debate, 1750–1900: The Idea of a Plurality of Worlds from Kant to Lowell* (Cambridge: Cambridge University Press)

Dick S J 1982 *Plurality of Worlds: The Origins of the Extraterrestrial Life Debate from Democritus to Kant* (Cambridge: Cambridge University Press)

Dick S J 1996 *The Biological Universe: The Twentieth Century Extraterrestrial Life Debate and the Limits of Science* (Cambridge: Cambridge University Press)

Dick S J 1998 *Life on Other Worlds* (Cambridge: Cambridge University Press)

Guthke K S 1990 *The Last Frontier: Imagining Other Worlds, from the Copernican Revolution to Modern Science Fiction* (Ithaca, NY: Cornell University Press)

Shklovskii I S and Sagan C 1966 *Intelligent Life in the Universe* (San Francisco, CA: Holden Day)

Steven Dick

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Life: Origins and Possible Distribution in the Universe

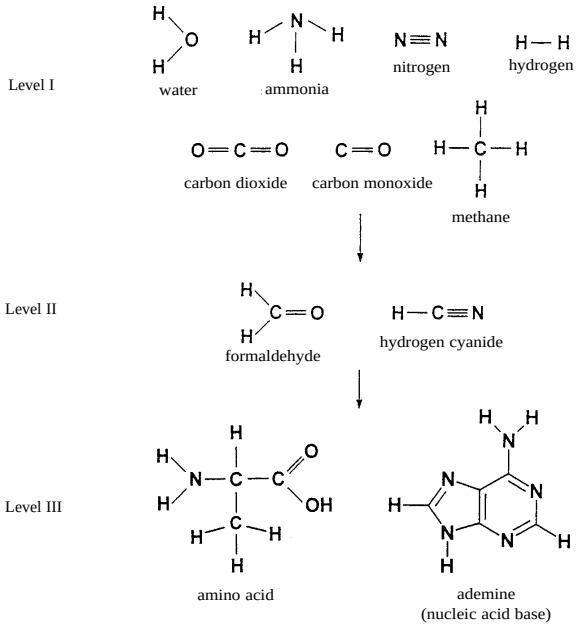
On Earth, life probably originated in liquid water with chemical systems capable of transferring their molecular information via self-replication and also to evolve by mutation. Unfortunately, the direct clues which may help chemists to identify the molecules which participated in the emergence of life on Earth about 4 billion years ago have been erased by plate tectonics, the permanent presence of running water, the unshielded solar ultraviolet radiation and by oxygen produced by life. By analogy with contemporary life, it is generally believed that primitive life originated from the processing by liquid water of reduced organic molecules mainly built of carbon, hydrogen, oxygen and nitrogen atoms. Organic matter might have been formed in the primitive Earth atmosphere or in hydrothermal vents. A large fraction of prebiotic organic molecules might have been brought by extraterrestrial meteoritic and cometary dust grains decelerated by the atmosphere. Any celestial body harbouring permanent liquid water may therefore accumulate the ingredients which generated life on the primitive Earth. The possibility that life might have evolved on early Mars when water existed on the surface marks it as a prime candidate in a search for bacterial life beyond Earth. Europa has an icy carapace. However, cryovolcanic flows at the surface point to a possible water subsurface region which might harbor a basic life form. The atmosphere and surface components of Titan is also of interest to exobiology for the insight into a hydrocarbon-rich living world. One-handed complex molecules and preferential isotopic fractionation of carbon, common to all terrestrial life forms, can be used as basic indicators when searching for life beyond Earth.

Origins of terrestrial life

Water

It is generally thought that life on Earth emerged in liquid water. For this reason, water is considered as one of the prerequisites for life to appear and evolve on a planet. In contemporary living systems, water participates as a solvent in the organization of biopolymers in three-dimensional structures and as a chemical reactant in most of the biochemical pathways. Hydrogen bonds are formed between water molecules and organic molecules providing that the latter contain –OH, –NH, –SH groups in addition to carbon and hydrogen. As a consequence of this affinity many CHONS organic molecules are soluble in water. Hydrocarbons cannot form hydrogen bonds with water molecules and are consequently insoluble in water. Abiotic organic molecules can be divided into two families: hydrocarbons and CHONS-containing molecules. When brought into the presence of liquid water, hydrocarbons try to escape the water molecules while CHONS have some affinity for water, especially those which bear ionizable groups. When these two species are forced to coexist within the same molecules, the duality generates interesting topologies. When they are separated by long distances—such as in fatty acids or phospholipids—micelles, vesicles or liposomes are formed due to clustering of the hydrophobic groups. Over short distances, hydrophobic and ionizable groups generate chain conformations which depend strongly on the sequence i.e. the distribution of the group along the chains.

Water molecules are widespread in the universe as grains of solid ice or as very dilute water vapor. Liquid water is a fleeting substance which can persist only above 0 °C and at pressures higher than 6 mbar. Therefore, the size of a planet and its distance from the star are two basic characteristics that will determine the presence of liquid water. If a planet is too small, like Mercury or the Moon, it will not be able to retain any atmosphere and, therefore, liquid water. If the planet is too close to the star, the mean temperature rises due to starlight intensity. Any sea water present would evaporate delivering large amounts of water vapor to the atmosphere thus contributing to the GREENHOUSE EFFECT. Such a positive feedback loop could lead to a runaway greenhouse: all of the surface water would be transferred to the upper atmosphere where



Scheme 1. The build-up of prebiotic molecules. Simple gaseous molecules like H_2, N_2, H_2O, NH_3, CO, CO_2 (Level I) require a supply of energy (UV, heat, lightning, cosmic rays, shock waves) to react with each other. They generate compounds like hydrogen cyanide and formaldehyde which store chemical energy in their double and triple bonds (Level II). These multiple bonded molecules react spontaneously in liquid water to form prebiotic molecules like amino acids or nucleic acid bases (Level III).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

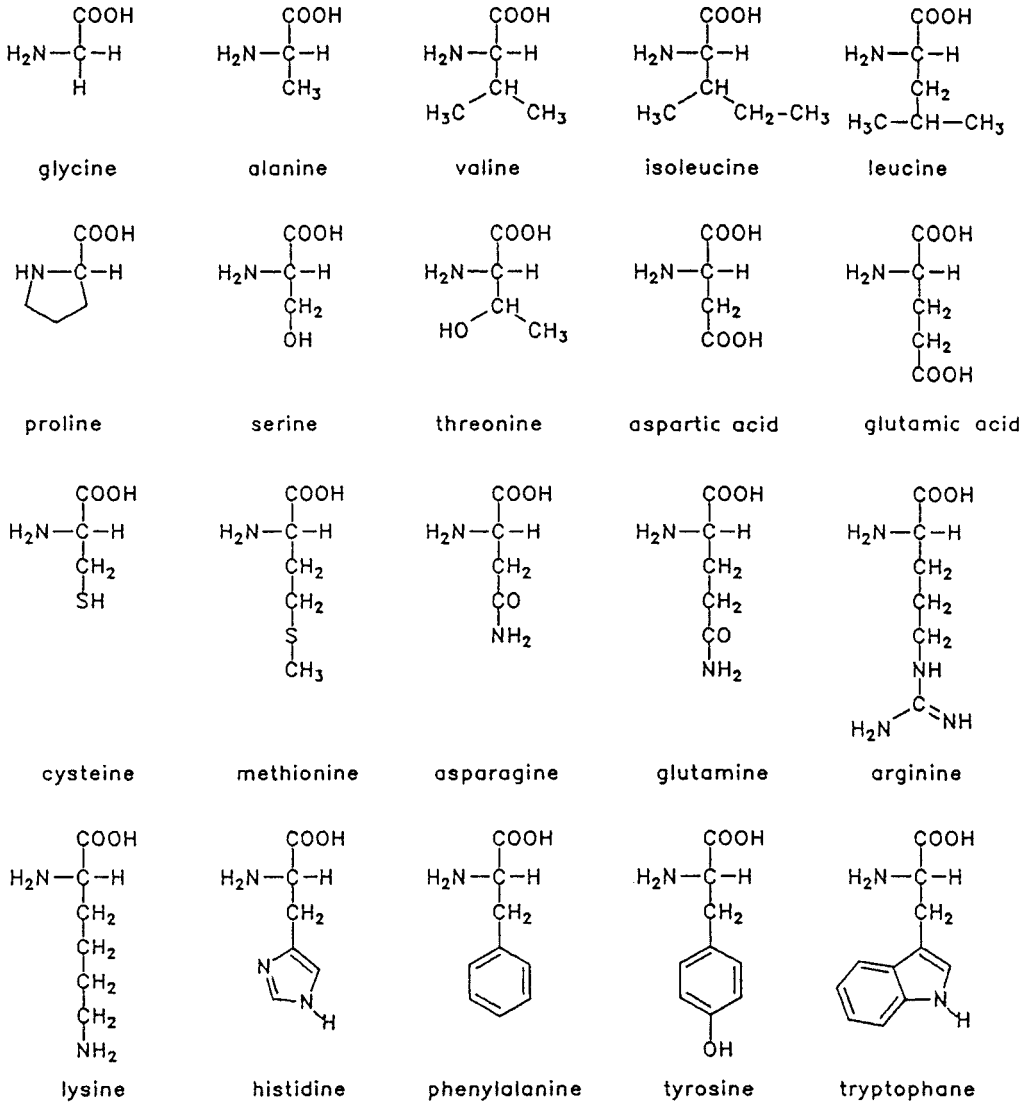
photodissociation by ultraviolet light would break the molecules down into hydrogen, escaping into space, and oxygen combining to the crust. If a planet is far from the star, it may host liquid water providing that it can maintain a constant greenhouse atmosphere. However, water would provoke its own disappearance. The atmospheric greenhouse gas, CO_2 for instance, would be dissolved in the oceans and finally trapped as insoluble carbonates by rock-weathering. This negative feedback would lower the surface pressure and the temperature to an extent where water would be largely frozen. The size of the Earth and its distance from the Sun are such that the planet never experienced either a runaway greenhouse or a divergent glaciation.

Terrestrial production of reduced organic molecules

Oparin suggested that the small reduced organic molecules needed for primitive life were formed in a primitive atmosphere dominated by methane (see scheme 1).

The idea was tested in the laboratory by Miller (1998) when he exposed a mixture of methane, ammonia, hydrogen and water to electric discharges. In his initial experiment, he obtained four of the twenty naturally occuring amino acids (table 1), via the intermediary formation of hydrogen cyanide and formaldehyde. Miller's laboratory synthesis of amino acids occurs efficiently when a reducing gas mixture containing significant amounts of hydrogen is used.

Table 1. Twenty amino acids are used by living systems to build up proteins. Seventeen have been obtained in the laboratory under primitive Earth-simulating conditions (arginine, histidine and lysine are still missing). Eight of them (glycine, alanine, valine, isoleucine, leucine, proline, aspartic acid and glutamic acid) are present in the Murchison meteorite.



Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

However, the true composition of primitive EARTH'S ATMOSPHERE remains difficult to define. The source of the primitive terrestrial atmosphere can be found in a combination of the volatiles CO_2, N_2, H_2O, trapped in the rocks that made the bulk of the planet's mass (internal reservoir) and a late accreting veneer of extraterrestrial material (external reservoir). The formation of the internal reservoir is a natural consequence of the ACCRETION of the planet. A late-accreting veneer of volatile-rich meteorites and comets may have contributed to replace the atmosphere produced by the planet's early accretion, which was subsequently blown off by the giant impact between the Earth and a Mars-sized planetary embryo which generated the Moon and caused the early atmosphere massive hydrodynamic loss. The abundance and isotopic ratios of noble gases neon, argon, krypton and xenon suggest that meteorites alone or in combination with planetary rocks could not have produced the Earth's entire volatile inventory and that a significant contribution from icy PLANETESIMALS was required (Owen 1998).

The discovery of sugar and several related organic compounds in two carbonaceous meteorites (which contain carbon as one of their important constituents) in 2002 suggests another fundamental building block of life on Earth may have come from outer space. Other organic, carbon–based compounds that play major roles in life on Earth, such as amino acids and carboxylic acids, have previously been found in meteorites, but no sugars. Sugars and the closely related compounds recently discovered collectively called polyols, are critical to all known life forms. They act as components of the nucleic acids RNA and DNA, constituents of cell membranes and cellular energy sources. A small sugar called dihydroxyacetone and several sugar–like substances, known as sugar acids and sugar alcohols, have been identified in a study of the Murchison and Murray meteorites. All these are important for life today. One sugar alcohol, glycerol (also known as glycerin), that is used by all contemporary cells to build cell walls has also been found. In addition, preliminary evidence of other compounds that may contain larger sugars critical in cellular metabolism, such as glucose have been discovered.

Water in contact with the olivine of ultramafic rock is reduced to molecular hydrogen (H_2) with the concomitant oxidation of Fe^{II}. The molecular hydrogen formed may be combined at high temperature with CO_2 for the abiotic formation of organic compounds such as hydrocarbons and fatty acids through Fischer–Tropsch type (FTT) synthesis. The analyses of fluids from the peridotite–hosted Rainbow hydrothermal field on the Mid–Atlantic Ridge indicate de novo synthesis of linear saturated hydrocarbons. The chain length of the hydrocarbons is between 16 and 29 carbon atoms. The discovery of FTT reactions in ultramafic hydrothermal systems on Earth provides an alternative pathway for the formation of early membranes and the origin of life.

The surface temperature of the early Earth was very much dependent upon the partial pressure of CO_2 in the atmosphere. The dominant view in recent years has been that the early atmosphere was a weakly reducing mixture of CO_2, N_2, and H_2O, combined with lesser amounts of CO and H_2 (Kasting and Brown 1998). An efficient greenhouse effect induced by a high CO_2 partial pressure was the most likely mechanism capable of compensating for the faint young Sun, the luminosity of which was about 30% lower than today. The dominant sink for atmospheric CO_2 was silicate weathering on land and subsequent formation and deposition of carbonate sediments in the ocean. However, a negative feedback system probably maintained the Earth's surface temperature above freezing: if the temperature were to drop below freezing, silicate weathering would slow down and volcanic CO_2 would accumulate in the atmosphere. The CO_2 partial pressure could have been as high as 10 bar prior to the emergence of the continents because of limited silicate weathering and continental carbonate rocks storage. Under these conditions, the early Earth could have been as hot as 85 °C. Recent analysis of samarium/neodymium ratios in the ancient zircons indicate that the continent grew rapidly, so a dense CO_2 atmosphere may have been restricted to the first few hundred millions of years of the Earth's history.

In a weakly reducing gas mixture of CO_2, N_2, and H_2O, the production of amino acids appears to be very limited. More recently, Wächterhäuser (1998) suggested that the carbon source for life was carbon dioxide. The energy source required to reduce carbon dioxide was provided by the oxidative formation of pyrite (FeS_2) from troilite (FeS) and hydrogen sulfide. Pyrite has positive surface charges and bonds the products of carbon dioxide reduction, giving rise to a two-dimensional reaction system, a surface metabolism. Experimental laboratory work is about to legitimate this new hypothesis.

Deep-sea hydrothermal systems may also represent likely environments for the synthesis of prebiotic organic molecules and perhaps for primitive life (Stetter 1998). Experiments have been carried out in order to test whether amino acids can be formed under conditions simulating the hydrothermal vents (Holm and Andersson 1998).

Extraterrestrial delivery of organic molecules to Earth

COMETS show substantial organic material. According to Delsemme's analysis (1998), Comet Halley dust particles ejected from the nucleus contains 14% organic carbon by mass. About 30% of cometary grains are dominated by light elements C, H, O, N and 35% are close in composition to carbonaceous chondrites. Among the molecules identified in comets are hydrogen cyanide and formaldehyde.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

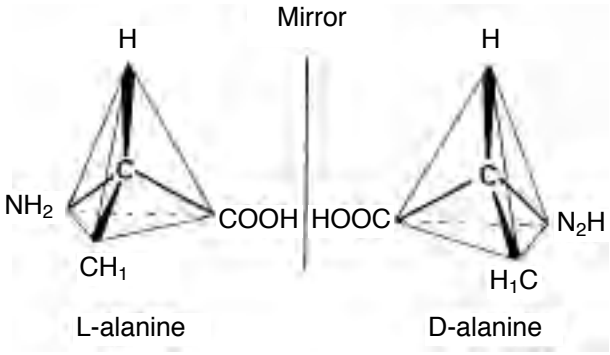


Figure 1. Each amino acid—with the exception of glycine—is asymmetrical and exists under two non-superimposable mirror images called enantiomers L and D. Only L enantiomers are found in proteins.

The presence of purines, pyrimidines and formaldehyde polymers has also been infered from the fragments analyzed by Giotto Picca and Vega Puma mass spectrometers. However, there is no direct identification of the complex organic molecules present in the dust grains and in the cometary nucleus. Many chemical species of interest for exobiology have been detected in Comet Hyakutake in 1996, including ammonia, methane, acetylene (C_2H_2), acetonitrile (CH_3CN) and hydrogen isocyanide (HNC) (Irvine *et al* 1996). In addition to hydrogen cyanide and formaldehyde (H_2CO), seen in several earlier comets, Comet Hale–Bopp was also shown to contain methane, acetylene, formic acid (HCOOH), acetonitrile, hydrogen isocyanide, isocyanic acid (HNCO), cyanoacetylene (HC_3N_2) and thioformaldehyde (H_2CS). Cometary grains might, therefore, have been an important source of organic molecules delivered to the primitive Earth.

The study of METEORITES, particularly the carbonaceous chondrites that contain up to 5% by weight of organic matter, has allowed close examination of extraterrestrial organic material delivered to the Earth. Eight proteinaceous amino acids have been identified in the Murchison meteorite among more than 70 amino acids (Cronin 1998). These amino acids are asymmetric and the two-handedness L and D are generally found in equal proportions.

However, Cronin and Pizzarello recently found excesses of L enantiomers of about 9% for non-protein amino. The presence of L-enantiomeric excesses in the Murchison meteorite suggests an extraterrestrial asymmetric synthesis of amino acids, asymmetry which is preserved inside the meteorite. These excesses may help to understand the emergence of one-handed (homochiral) life. For example, proteins are built up with twenty different amino acids (table 1). Each amino acid, with the exception of glycine, exists in two enantiomeric forms L and D but proteins use only L ones (see figure 1). Proteins adopt asymmetrical rigid geometries, α-helices and β-sheets, which play a key role in the catalytic activity. Homochirality is now believed to be not just a consequence of life, but also a prerequisite for life, because stereoregular structures such as protein β-sheets do not form with mixtures of amino acids of both handedness. The excess of one-handed amino acids found in the Murchison meteorite may result from the processing of the organic mantles of interstellar grains by circularly polarized synchrotron radiation from a neutron star remnant of a supernova. Strong infrared circular polarization resulting from dust scattering in reflection nebulae in the Orion OMC-1 star-formation region has ben recently observed. Circular polarization at shorter wavelengths might have been important in inducing chiral asymmetry in interstellar organic molecules that could be subsequently delivered to the early Earth. Whether the extraterrestrial organic material collected on Earth bear witness to the interstellar organic chemistry is still unknown and we really need to gain more information about the different steps between interstellar medium (ISM)—where more than 100 molecules, the bulk of them organic, have been detected so far by radioastronomers (Irvine 1998)—and the Earth in order to better understand the global ISM–comets–planets ecosystem from an exobiology point of view.

Dust collection, both above the terrestrial atmosphere and in the Greenland and Antarctica ice sheet (Maurette 1998), shows that the Earth captures interplanetary dust as MICROMETEORITES at a rate of about 50–100 tons per day. About 99% of this mass is carried by micrometeorites in the 50–500 μm size range. This value is about 2000 times higher than the most reliable estimate of the meteorite flux, i.e. about 0.03 tons per day. This amazing dominance

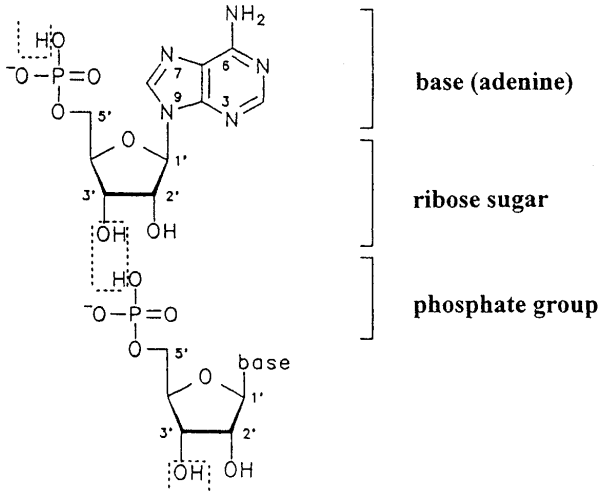


Figure 2. Nucleotides are the building blocks of nucleic acids. Each nucleotide is composed of a phosphate group, four different heterocyclic bases and a ribose sugar. The bases can be reconstructed in the laboratory under primitive Earth-simulating conditions. They are also present in the Murchison meteorites. The spontaneous formation of the ribose sugar is still an unsolved challenge.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 2. Interstellar and circumstellar molecules. (From W M Irvine 1998 *Origins Life Evol. Biosphere* **28** 365–83 updated June 1999.)

2 atoms	*3 atoms*	*4 atoms*	*5 atoms*	*6 atoms*	*7 atoms*	*≥8 atoms*
H_2	C_2H	C_2H_2	C_4H	C_2H_4*	C_6H	CH_3COOH
C_2	CH_2	l-C_3H	C_3H_2	H_2CCCC	HC_5N	$HCOOCH_3$
CH	HCN	c-C_3H	H_2CCC	CH_3OH	CH_2CHCN	CH_3C_3N
CH^+	HNC	NH_3	HCOOH	CH_3CN	CH_3C_2H	C_7H*
CN	HCO	HNCO	CH_2CO	CH_3NC	CH_3CHO	CH_3C_4H
CO	HCO^+	$HOCO^+$	HC_3N	CH_3SH	CH_3NH_2	CH_3OCH_3
CS	HOC^+	$HCNH^+$	CH_2CN	NH_2CHO	c-CH_2OCH_2	CH_3CH_2CN
OH	N_2H^+	HNCS	NH_2CN	C_5H		CH_3CH_2OH
NH	NH_2	C_3N	CH_2NH	HC_3HO		HC_7N
NO	H_2O	C_3O	CH_4	HC_3NH^+		CH_3C_4CN?
NS	HCS^+	H_2CS	SiH_4*	C_5N		CH_3COCH_3
SiC*	H_2S	C_3S	C_4Si*			HC_9N
SiO	OCS	HCCN	C_5*			C_8H
SiS	N_2O	H_3O+	HCCNC			H_2C_6
SiN*	SO_2	H_2CN	HNCCC			$HC_{11}N$
SO	SiC_2*	H_2CO	HCO_2H^+			$(C_2H_5)_2O$?
HCl	C_2S	c-SiC_3				
CP*	C_2O					
SO^+	C_3*					
NaCl*	MgNC*					
AlCl*	MgCN*					
KCl*	NaCN*					
AlF*	HNO					
PN	H_3^+					
CO^+						
SiH?						
HF?						

* Detected only in stellar envelopes.
? Tentative detection.
c- and l-: cyclic and linear molecules, respectively.

of micrometeorites already suggests their possible role in delivering complex organics to the early Earth 4.2–3.9 Gyr ago when the micrometeorite flux was enhanced by a factor of 1000. In the Antarctic micrometeorites, a high percentage of unmelted chondritic micrometeorites from 50 to 100 μm in size has been observed, indicating that a large fraction crossed the terrestrial atmosphere without drastic thermal treatment. In this size range, the carbonaceous micrometeorites represent 80% of the samples and contain 2% of carbon, on average. They might have brought about 10^{20} g of carbon over a period of 300 million years corresponding to the late terrestrial bombardment phase. This delivery represents more carbon than that engaged in the surficial biomass, i.e. about 10^{18} g. Amino acids such as α-amino isobutyric acid have been recently identified in these Antarctic micrometeorites. These grains also contain a high proportion of metallic sulfides, oxides and clay minerals which belong to various classes of catalysts. In addition to the carbonaceous matter, micrometeorites might also have delivered a rich variety of catalysts, having perhaps acquired specific crystallographic properties during their synthesis in the microgravity environment of the early solar nebula. They may have functioned as tiny chondritic chemical reactors when reaching oceanic water.

Primitive cellular life

By analogy with contemporary living systems, it was first generally believed that primitive life emerged as a cell-like organized system thus requiring, at least, boundary molecules able to isolate the system from the aqueous environment (membrane), catalytic molecules providing the basic chemical work of the cell (enzymes) and informative molecules allowing the storage and the transfer of the information needed for replication (RNA).

Boundary molecules

Fatty acids are known to form vesicles when the hydrocarbon chains contain more than ten carbon atoms. Such vesicle-forming fatty acids have been identified in the Murchison meteorite. However, the membranes obtained with these simple amphiphiles are not stable over a broad range of conditions. Stable lipids can be obtained by condensing fatty acids with glycerol or with glycerol phosphate, thus mimicking the sf contemporary phospholipids. Primitive membranes could initially have also been formed by simple isoprene derivatives (Deamer 1998).

Catalytic molecules

Most of the catalytic chemical reactions in a living cell are achieved by proteinaceous enzymes made of 20 different

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

homochiral L-amino acids. Amino acids were most likely available on the primitive Earth as complex mixtures. The number of condensing agents capable to assemble amino acids into peptides in water is restricted, especially when looking for prebiotically plausible compounds. Carbodiimides, R–N=C=N–R, can be used in water providing a careful choice of the substituent R and of experimental conditions. The simplest carbodiimide, H–N=C=N–H, can be considered as a transposed form of cyanamide NH_2–CN which is present in interstellar medium. In water, cyanamide forms a dimer, dicyandiamide, which is as active as carbodiimides to form peptides. However, the reactions are very slow and did not proceed beyond the tetrapeptide. Diaminomaleonitrile, the tetramer of HCN, has been found to form diglycine in 3.1% yield.

Clays and salts can also be used to condense amino acids in water. Amino acid adenylate anhydrides condense readily in the presence of montmorillonite. When subjecting mixtures of glycine and kaolinite to wet–dry and 25 °C–94 °C temperature fluctuations, the formation of oligopeptides up to five glycines long has been observed. The drying/wetting cycle procedure was extended to alanine in the presence of silica, alumina and smectic clays. Mutual catalysis of amino acids in the salt-induced peptide formation reaction has been demonstrated for the case of glycine/alanine. Efficient mineral-catalyzed (hydroxylapatite, illite clay) condensation of amino acids into long peptides have been recently reported (Ferris 1998).

Thermal condensation of amino acids has been described by Fox. He has shown that dry mixtures of amino acids polymerize when heated at 130 °C to give 'proteinoids'. In the presence of polyphosphates, the temperature can be decreased to 60 °C. High molecular weights were obtained when an excess of acidic or basic amino acids are present. In aqueous solutions, the proteinoids aggregate spontaneously in microspheres of 1–2 μm, presenting an interface resembling the lipid bilayers of living cells. The microspheres increase slowly in size from dissolved proteinoids and are sometimes able to bud and to divide. These microspheres are described as catalyzing the decomposition of glucose and work as esterases and peroxydases. The main advantage of proteinoids is their organization into particles but they represent a dramatic increase in complexity. When heating a mixture of selected L-amino acids, the resulting condensate is only 50% peptidic, the peptidic fraction is racemized, the peptide linkages are ambiguous since they include the α, β and γ functions of the dicarboxylic amino acids and the sequences are multiple although not completely random.

According to De Duve, the first peptides appeared via thioesters. Chemical reactions capable to selectively condense the protein amino acids at the expenses of the non-protein ones have been described. Helical and sheet structures can be modeled with the aid of only two different amino acids, one hydrophobic, the second hydrophilic. Polypeptides with alternating hydrophobic and hydrophilic residues adopt a water-soluble β-sheet geometry because of hydrophobic side-chain clustering. Due to the formation of β sheet, alternating sequences gain a good resistance toward chemical degradation. Aggregation of alternating sequences into β sheets is possible only with all-L or all-D polypeptides. When hydrophobic and hydrophilic amino acids coexist within the same polypeptide chain, the duality generates interesting topologies such as stereoselective and thermostable β-sheet structures. Short peptides have also been shown to exhibit catalytic properties (De Duve 1998).

Informative molecules

In contemporary living systems, hereditary memory is stored in nucleic acids built up with bases (purine and pyrimidine), sugars (ribose for RNA, deoxyribose for DNA) and phosphate groups.

The accumulation of significant quantities of natural RNA nucleotides does nor appear as a plausible chemical event on the primitive Earth (Schwartz 1998). Purines are easily obtained from hydrogen cyanide or by submitting reduced gas mixtures to electric discharges. No pyrimidine synthesis from electric discharges has been published so far whereas hydrogen cyanide affords only very small amounts of the base. Condensation of formaldehyde leads to ribose among a large number of other sugars. The synthesis of purine nucleosides (the covalent combination of purine and ribose) and of nucleotides have been achieved by heating the components at the solid state. The yields are very low and the reactions are not regioselective. Interestingly, efficient clay-catalyzed condensation of nucleotides into oligomers up to 55 monomers long have been recently reported (Ferris 1998).

The synthesis of oligonucleotides is much more efficient in the presence of a preformed pyrimidine-rich polynucleotide acting as a template. Non-enzymatic replication has been demonstrated by Orgel and his coworkers. The preformed chains align the nucleotides by base pairing to form helical structures which bring the reacting groups in close vicinity.

The prebiotic synthesis of oligonucleotides remains an unsolved challenge. Each step of the chemical pathway has been achieved in the laboratory but the desired product is never dominant. It is always contaminated with by-products which are wastes or even poisons. Chemists have considered early living systems using simpler informative molecules. The ribose sugar has been replaced by glycerol for the backbone build up. The chemistry of these simplified informative molecules does

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

not appear much easier. The straightforward and selective formation of ribose diphosphates from glycolaldehyde phosphate suggests that pyranosyl–RNA may have been involved in the early forms of life. Intense experimental work is presently run in Eschenmoser's laboratory along this new avenue (Schwartz 1998).

Primitive life only based on RNA?

Cech found that some RNA, the ribozymes, are able to act as catalytic molecules (James and Ellington 1998). For example, they increase the rate of hydrolysis of oligoribonucleotides. They act also as polymerization templates since chains up to 30 monomers long can be obtained starting from a pentanucleotide. Since this primary discovery, the catalytic spectrum of these ribozymes have been considerably enlarged by directed test tube molecular evolution experiments initiated in Joyce's and Szostak's laboratories. Since RNAs have been shown to be able to act simultaneously as informative and catalytic molecules, RNA has been considered as the first living system on the primitive Earth (RNA world) because they can be at the same time genotypes and phenotypes and can fulfill the basic cycle of life consisting of self-replication/mutation/selection. One should, however, remember that their synthesis under prebiotic conditions remains an unsolved challenge. It seems unlikely that life started with RNA molecules because these molecules are not simple enough. The RNA world appears as an episode in the evolution of life before the appearance of cellular microbes rather than as the birth of life. Chemists are now tempted to consider that primitive self-replicating systems must have used simpler informational molecules than biological nucleic acids or their analogs. Since self-replication is, by definition, autocatalysis, they are searching for simple autocatalytic molecules capable of mutation selection. Rebek and co-workers have used nucleic base-like A–B templates able to form hydrogen-bonded ternary complexes with individual A and B subunits and to catalyse the synthesis of new A–B molecules (Burmeister 1998). These experiments are illuminating examples of autocatalysis but they do not yet address the problem of self-replication of long linear sequences. Von Kiedrowski showed that simple A–B molecules unrelated to nucleotides can provide exponentially replicating autocatalytic models. Luisi has demonstrated beautiful examples of autocatalytic micelle growth. However, these autocatalytic systems do not really store hereditary memory and cannot therefore evolve by natural selection.

Geological records of primitive life

The geological record also provides important information. The isotopic signatures of the organic carbon of the Greenland metasediments bring indirect evidence that life may be 3.85 billion years old. Schidlowski has compiled the carbon isotopic composition of over 1600 samples of fossil kerogen (a complex organic macromolecule produced from the debris of biological matter) with carbonates in the same sedimentary rocks. The organic matter has an average $\delta^{13}C$ of $-25\pm7‰$, whereas the carbonate varies by no more than about ‰ around a mean of ‰ ($\delta^{13}C$ expresses the permil deviation of the $^{13}C/^{12}C$ ratio of the sample relative to that of a conventional standard, the carbonate skeleton of a fossil cephalopod which has a $^{13}C/^{12}C$ ratio of 0.0112). The *ca* 25–30% shift is taken as an indication that biosynthesis by photosynthetic organisms was involved. In fact, this offset is now taken to be one of the most powerful indications that life on Earth was active nearly 3.9 Gyr ago because the sample suite encompases specimens right across the geologic timescale. Some organic matter in ancient sediments has been measured as being even more enriched in the light isotope of carbon which would suggest involvement of methane-utilizing organisms. This conclusion is fully consistent with the remarkable diversity of the 3.465 billion year old fossilized microflora (Schopf 1998).

Were the first replicators based on autocatalysis, on thioester chemistry, on organic molecules adsorbed on pyrite? This exciting question is still open. A major problem lies in the fact that we remain ignorant of the true historical facts on Earth from the time when life started.

Life beyond Earth

The main constituents of terrestrial life, i.e. water and organic molecules, are widespread in the universe. Primary organic chemistry is very active in interstellar space. So far, 83 different organic molecules have been identified by RADIOASTRONOMY.

The most important identifications for prebiotic chemistry are probably are hydrogen cyanide, formaldehyde and cyanamide. Although there is little basis for speculating that interstellar organic molecules arrived intact at the prebiotic Earth's surface, there is growing interest in the possibility that interstellar molecules may be preserved intact in comets and in altered form in carbonaceous meteorites.

The molecules which allowed the emergence of terrestrial life appear to be universal. On Earth, they were subjected to a given set of pressure and temperature constraints. Do similar constraints exist elsewhere?

Life on Mars?

MARS mapping by Mariner 9 and Viking 1 and 2 revealed channels resembling dry river beds. Three major classes of channels were identified: (i) dendritic runoff channels and valleys primarily associated with cratered terrain older than 3.8 Gyr and which are generally associated with fluvial water erosion; (ii) outflow channels suggestive of large-volume flows in cataclysmic events; (iii) fretted channels which are steep-walled with smooth flat floors suggestive of erosion by

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

debris flows (McKay 1998). The inventory of the total amount of water that may have existed on the surface of Mars is difficult to estimate and varies from some meters to several hundred meters. The climatic conditions required for the formation of the valley networks are also poorly understood. The most plausible explanation is a relatively thick CO_2 atmosphere (1–5 bar) in the early history of Mars, giving rise to a greenhouse effect which would have allowed for mean global surface temperature to remain above the freezing point of water. Carr suggested recently that the valley networks, like the fretted channels, formed mainly by mass wasting, aided by groundwater seepage into the mass-wasted debris, between 3.8 and 3.7 Gyr ago. Therefore, it is generally considered that liquid water has been restricted to the very early stages of Martian history. By about 3.8 Gyr ago, most of the atmospheric CO_2 was probably irreversibly lost due to carbonate formation and pressure and temperature declined. However, the presence of extensive coastlines at the border of the northern lowlands where the Valles flow supports the presence of a large ocean in the northern hemisphere. This northern ocean could possibly have been active in epochs as young as Middle Amazonian, 500 million years ago. On the other hand, several features suggest the intermittent presence of water on the surface of Mars. Outflow channels have been active in Maja and Ares Valles (Late Hesperian), in Kasei and Mangala Valles (Early Amazonian) and at Elysim (Middle Amazonian). Lacustrine deposits show the same range of age.

The early histories of Mars and Earth clearly show similarities. Geological observations collected from Martian orbiters suggest that liquid water was once stable on the Mars surface, attesting to the presence of an atmosphere capable of deccelerating carbonaceous micrometeorites. Therefore, chemical evolution may have been possible on Mars. The Viking 1 and 2 lander missions were designed to address the question of extant life on Mars. Three experiments were selected to detect metabolic activity such as photosynthesis, nutrition and respiration of potential microbial soil communities. The results were ambiguous since, although 'positive' results were obtained, no organic carbon was found in the Martian soil by gas chromatography–mass spectrometry. It was concluded that the most plausible explanation for these results was the presence, at the Martian surface, of highly reactive oxidants like H_2O_2 which would have been photochemically produced in the atmosphere. The Viking lander could not sample soils below 6 cm and therefore the depth of this apparently organic-free and oxidizing layer is unknown. It has been estimated that the depth of diffusion for H_2O_2 is less than 3 m. Direct photolytic processes can also be responsible for the dearth of organics at the Martian surface.

The alpha proton x-ray spectrometer (APXS) on board the rover of the Mars Pathfinder mission in 1997 measured the chemical composition of six soils and five rocks at the Ares Vallis landing site. Some rocks are similar to terrestrial andesites but it is not certain that these rocks are igneous. The texture of other rocks are difficult to interpret and might be sedimentary or metamorphic.

Although the Viking missions were disappointing in the first instance for EXOBIOLOGY, in the long run the programme has proven extremely beneficial for the investigation of the possibility of life on Mars. Prior to Viking it had been apparent that there was a small group of meteorites, all of igneous origin, called SNC METEORITES (after their type specimens Shergotty, Nakhla and Chassigny) that had comparatively young crystallization ages equal to or less than 1.3 Gyr. One of these meteorites, EETA 79001, was found on Antarctica in 1979. It had, trapped within glass pockets, gas which both compositionally and isotopically matched, in all respects, the make-up of the Martian atmosphere as measured by the mass spectrometer utilized for assessing the soil for the presence of organic compounds. The data provide a very strong argument that at least that particular SNC meteorite comes for Mars, thus supporting the circumstantial conclusion that materials of a young age must have derived from a planetary-sized body. There are now twelve SNC meteorites known in total; others were found more recently on the Antarctic Ice Cap including one of much older age, ALH 84001. They can be all shown to be related by comparing their oxygen isotopic compositions. Only these twelve specimens (out of a total of *ca* 20 000 meteorites) define a correlation line of slope one half on a plot of $\delta^{18}O$ versus $\delta^{17}O$ with $\delta^{17}O = 0.321 \pm 0.013‰$, i.e. displaced from the Earth reference line.

The two SNC meteorites EETA79001 and ALH84001 supply new and highly interesting information. A subsample of EETA79001 excavated from deep within the meteorite has been subjected to stepped-combustion. The CO_2 release from 200 °C to 400 °C suggests the presence of organic molecules. The carbon is enriched in ^{12}C with $\delta^{13}C$ of about –27‰. The difference between organic matter and carbonates in Martian meteorites is greater than the shift seen on Earth which could also be indicative of biosynthesis but some as yet unknown reason cannot be ruled out. McKay *et al* (1996) reported the presence of PAHs, carbonates globules and ovoid features which may represent a signature of relic biogenic activity on Mars but today this biological interpretation has almost been abandoned.

Even if the evidence for ancient life in ALH84001 is not established, the two Martian meteorites show the presence of organic molecules, suggesting that the ingredients required for the emergence of a primitive life may have been present on the surface of Mars. Therefore, it is tempting to consider that microorganisms may have

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

developed on Mars until liquid water disappeared. Since Mars probably had no plate tectonics and since liquid water seems to have disappeared from the Mars surface very early, the Martian sub-surface perhaps keeps a frozen record of the very early forms of a terrestrial-like life. American, Japanese and European space agencies have planned a very intensive exploration of Mars. Exobiology interests are included especially in the analysis of samples from sites where the environmental conditions may have been favorable for the preservation of evidence of possible prebiotic or biotic processes. The NASA program included Mars Polar Lander (launched in January 1999, but failed in December 1999) which was intended to land near the southern polar ice cap and analyze some soil samples. Mars 2001 Lander/Rover, to be launched in April 2001, will deliver a small instrumented rover 'Marie-Curie' similar to Sojourner for *in situ* analyses. Mars 2003 Lander/Rover (Spring 2003) will deliver a more sophisticated rover for sample analysis and acquisition as part of the NASA–CNES Mars 05 Sample Return mission. The Martian samples collected both in 2003 and 2005 will be sent into low Martian orbit by two distinct mini-rockets and finally brought back to Earth in 2008. ESA has led a study to design a multi-user integrated suite of instruments for the search for evidence of life on Mars. The priority has been given to the *in situ* organic and isotopic analysis of samples obtained by subsurface core drilling. The multi-user package is expected to be mounted on the 2005 Mars Sample return lander. A first version of the ESA Mars Exobiology integrated package could be launched in June 2003 as part of the ESA Mars Express mission.

Life within Europa's ocean?

EUROPA appears as one of the most enigmatic of the Galilean satellites. With a mean density of about 3.0 g cm^{-3}, the Jovian satellite should be dominated by rocks. Ground-based spectroscopy, combined with gravity data, suggests that the satellite has an icy crust roughly 150 km thick and a rocky interior. The Voyager images showed very few impact craters on Europa's surface, indicating recent, and probably continuing, resurfacing by cryovolcanic and tectonic processes. The Voyager spacecraft also revealed that Europa's surface is crossed by numerous intersecting ridges and dark bands. It has been suggested that Europa's thin outer ice shell might be separated from the silicate interior by a liquid water layer, prevented from freezing by tidal heating as the result of the variation of the gravitational field across the body of the satellite. Heat transfer from the core to the bottom of the ocean, similar to thermal vents in terrestrial oceans, is another possible source of thermal energy. Although the existence of such an ocean is still uncertain, the last images from Galileo showing evidence for mobile icebergs support the presence of liquid water at shallow depths below the surface, either today, or at some time in the past (Carr *et al* 1998). If liquid water is present within Europa, it is quite possible that it includes organic matter derived from thermal vents. Terrestrial-like prebiotic organic chemistry and primitive life may therefore have developed in Europa's ocean. If Europa maintained tidal and/or hydrothermal activity in its subsurface until now, it is possible that bacterial activity is still present. Thus, the possibility of an extraterrestrial life present in the hypothetical ocean of Europa must be seriously considered. The most likely sites for extant life would be at hydrothermal vents below the most recently resurfaced area. To study this directly would require making a borehole through the ice in order to deploy a robotic submersible. On the other hand, biological processes in and around hydrothermal vents could produce biomarkers that would be erupted as traces in cryovolcanic eruptions and thereby be available at the surface for *in situ* analysis or sample return. Mineral nutrients delivered through cryovolcanic eruption would make the same locations the best candidates for photosynthetic life.

What about Titan?

TITAN's atmosphere was revealed mainly by the Voyager 1 mission in 1980, which yielded the bulk composition (90% molecular nitrogen and about 1–8% methane). Also, a great number of trace constituents were observed in the form of hydrocarbons (acetylene, ethane, ethylene, etc), nitriles (HCN, C_2N_2, HC_3N) and oxygen compounds (CO and CO_2). Titan is the only other object in our solar system to bear a resemblance to our own planet in terms of atmospheric pressure (1.5 bar) and carbon/nitrogen chemistry (Raulin 1998). It represents, therefore, a natural laboratory to study the formation of complex organic molecules on a planetary scale and over geological times. The ISO satellite has recently detected tiny amounts of water vapour in the higher atmosphere, but Titan's surface temperatures (94 K) are much too low to allow the presence of liquid water. Although liquid water is totally absent, the satellite provides a unique milieu to study, *in situ*, the products of the fundamental physical and chemical interactions driving a planetary organic chemistry. Titan also serves as a reference laboratory to study, by default, the role of liquid water in exobiology.

The Cassini–Huygens spacecraft launched in October 1997 will arrive in the vicinity of Saturn in 2004 and perform several flybys of Titan making spectroscopic, imaging, radar and other measurements. A descent probe will penetrate the atmosphere and will systematically study the organic chemistry in Titan's geofluid. During 150 min, *in situ* measurements will provide detailed analysis of the organics present in the air, in the aerosols and at the surface.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Arguments in favor of universal life

Arguments can be taken from the study of the origins of terrestrial life. Terrestrial life probably appeared between 4 and 3.8 billion years ago when some organic molecules processed by liquid water began to transfer their chemical information and to evolve by making a few accidental transfer errors. Schematically, the prebiotic condition can be compared to spare pieces of a robot. By chance, some pieces assembled to form a robot capable of assembling other spare pieces to form a second identical robot, etc. Sometimes, a minor error in the building generated more efficient robots. The number of pieces required for the first robot is still unknown. The pieces have been erased by plate tectonics, the permanent presence of liquid water, the unshielded solar ultraviolet radiation and by life itself. If the number was small, life must arise on any body presenting environmental conditions similar to those which prevailed on the primitive Earth because simple chemistry is reproducible. If the number of pieces was large, then life is probably a very rare event, perhaps restricted to our planet. The idea of a primordial simplicity is supported by chemists who found that the spontaneous formation of the RNA sugar ribose under prebiotic conditions is very difficult. They consider now that primitive life was probably simpler than a cell and even than RNA. A simple self-replicating system was probably more robust and offered better chances to resist cataclysmic impacts which might have periodically sterilized the Earth during heavy bombardment. A simple primitive life would be more easily extrapolatable to another celestial body.

Biologists have shown that bacterial life can survive under extreme conditions. Life has continued to develop very well in water which is very acidic, alkaline, or is a strong brine solution. It has also survived and flourished in water at high pressure and at temperatures above 100 °C (Stetter 1998). A flourishing biosphere has been discovered a kilometer deep below the surface of the Earth. Even the space environment is not so harmful. The idea that life could migrate in space, known as panspermia, was originally put forward by Richter and Arrhenius. Recent discoveries have given new support to this idea. These are: (i) the identification of meteorites of lunar and probably also of Martian origin; (ii) the probability of small particles reaching escape velocities by the impact of large meteorites on a planet; (iii) the ability of bacterial spores to survive the shock waves of a simulated meteorite impact; (iv) the high UV resistance of microorganisms at the low temperature of deep space; (v) the high survival rate of bacterial spores over extended periods in space, provided they are shielded against the intense solar ultraviolet radiation or are coated by a mantle of absorbing material which attenuates the solar UV radiation (Horneck 1995). Panspermia will not tell us how life started in another star system but may contribute to distribute life in the universe.

Delivery of extraterrestrial organic molecules is of special interest since it can still be studied today. Such an import process in the early Earth only required an atmosphere to decelerate the particles. Such an atmosphere existed 3.8 billion years ago as documented by geological clues including the Greenland sediments. We have clear evidence that water existed in substantial amounts on the surface of Mars at some earlier epoch, attesting to the presence of an atmosphere capable of decelerating micrometeorites. Therefore, primitive life might also have developed on Mars. How long water was present at the surface is unknown. Nor do we know if water still exists in subsurface aquifers, although there are clear indications of the existence of large permafrost regions and of old water flows out of associated areas. Given the discovery of a flourishing biosphere even a kilometer deep below the surface of the Earth, it would be possible that a similar vast microbial community could still be present below the surface of Mars, having long ago retreated into that ecological niche following the disappearance of the surface water environment. The possibility that life may have evolved on Mars during an early period when water existed on the surface and that life may still exist deep below the surface, marks it as a prime candidate in a search for life beyond the Earth. Finding an extinct terrestrial-like life on Mars would be a major discovery. It would indicate that terrestrial life started with a rather simple set of autocatalytic molecules. If, by a stroke of good fortune, it could be shown that the Martian life used D-amino acids, for example, this would exclude any ancient panspermia from Mars to the Earth and would leave the sign of terrestrial homochirality to a pure chance event. Within the solar system, Europa may have an ocean of liquid water beneath its icy crust as suggested by data and theory. If submarine volcanism exists on Europa, the question arises of whether such activity could support life as in volcano–hydrothermal sites on the Earth's seafloor. New planets have been recently discovered beyond the solar system (see EXOPLANETS by Jean Schneider). How to detect extrasolar life on these planets will probably be a formidable challenge for astronomers and radioastronomers in the next century. The detection of water and ozone (an easy detectable telltale signature of oxygen) in the atmosphere will be a strong indication but not an absolute proof. The detection of an electromagnetic signal (the SETI programme) would be more convincing but probably more problematic. Such discoveries would overjoy chemists, who suspect that life is universal because organic molecules and water are universal, but who need experimental proof to claim it.

Bibliography

Holm N Gand Charlou J L Initial indications of abiotic formation of hydrocarbons in the Rainbow ultramafic hydrothermal system, Mid-Atlantic Ridge *Earth Planet. Sci. Lett.* **191** 1–8

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Cooper *et al* December, 2001 Carbonaceous Meteorites as a Source of Sugar-related Organic Compounds for the Early Earth *Nature*

Burmeister J 1998 Self-replication and autocatalysis *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 295–312

Carr M H *et al* 1998 Evidence for a subsurface ocean on Europa *Nature* **39** 363–5

Cronin J R 1998 Organic molecules on the early Earth. Clues from the origin of the Solar System: meteorites *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 119–46

Deamer D W 1998 Membrane compartments in prebiotic evolution *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 189–205

DeDuve C 1998 Possible starts for primitive life. Clues from present-day biology: the thioester world *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 219–36

Delsemme A H 1998 Cosmic origin of the biosphere *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 100–18

Ferris J P 1998 Catalyzed RNA synthesis for the RNA world *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 255–68

Holm N G and Andersson E M 1998 Organic molecules on the early Earth: hydrothermal systems *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 86–99

Horneck G 1995 Exobiology, the study of the origin, evolution and distribution of life within the context of cosmic evolution: a review *Planet. Space Sci.* **43** 189–217

Irvine W M 1998 Extraterrestrial organic matter: a review *Origins Life Evol. Biosphere* **28** 365–83

Irvine W M *et al* 1996 Spectroscopic evidence for interstellar ices in Comet Hyakutake *Nature* **383** 418–20

James K D and Ellington A D 1998 Catalysis in the RNA world *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 269–94

Kasting J F and L L Brown 1998 The early atmosphere as a source of biogenic compounds *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 35–56

McKay C P 1998 Life on Mars *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 386–406

McKay D S, Gibson E K, Thomas-Keprta K L, Vali H, Romanek C S, Clemett S J, Chellier X D F, Maechling C R and Zare R N 1996 Search for past life on Mars: possible relic biogenic activity in martian meteorite ALH 84001 *Science* **273** 924–30

Maurette M 1998 Micrometeorites *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 147–86

Miller S L 1998 The endogenous synthesis of organic compounds *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 59–85

Owen T C 1998 The origin of the atmosphere *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 13–34

Raulin F 1998 Titan *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 365–85

Schopf J W 1998 Tracing the roots of the universal tree of life *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 336–62

Schwartz A W 1998 Origins of the RNA world *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 237–54

Stetter K O 1998 Hyperthermophiles and their possible role as ancestors of modern life *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 315–35

Wächterhäuser G 1998 Origin of life in an iron-sulfur world *The Molecular Origins of Life: Assembling Pieces of the Puzzle* ed A Brack (Cambridge: Cambridge University Press) pp 206–18

A Brack

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Light

Electromagnetic radiation of those wavelengths to which the human eye responds, i.e. from just below 400 nanometers (4×10^{-7} m, 0.4 microns, or 4000 angstroms) to just over 700 nm. Different wavelengths within this range correspond to different colors as perceived by the eye. In order of decreasing wavelength are red, orange, yellow, green, blue, indigo, and violet. Wavelengths shorter than visible are referred to as ultraviolet, wavelengths longer than red are infrared. Light may also be regarded as particles, or photons, the longer-wavelength radiation corresponding to lower-energy photons.

In common with other types of electromagnetic radiation, light travels in a vacuum at a speed of about 300 000 km s^{-1}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Light Element Nucleosynthesis

The early universe was hot and dense behaving as a cosmic nuclear reactor during the first 20 min of its evolution. It was, however, a 'defective' nuclear reactor, expanding and cooling very rapidly. As a result, only a handful of the lightest nuclides were synthesized before the density and temperature dropped too low for the nuclear reaction rates to compete with the universal expansion rate (see UNIVERSE: THERMAL HISTORY). After hydrogen (^{1}H $\equiv$ protons) the next most abundant element to emerge from the Big Bang is helium (^{4}He $\equiv$ alpha particles). Isotopes of these nuclides (deuterium and helium-3) are the next most abundant primordially. Then there is a large gap to the much lower abundance of lithium-7. The relative abundances of all other primordially-produced nuclei are very low, much smaller than their locally observed (or, currently observable) abundances. After a brief description of the early evolution of the universe emphasizing those aspects most relevant to primordial, or 'big bang' nucleosynthesis (BBN), the predicted abundances of the light nuclides will be presented as a function of the one 'free' parameter (in the simplest, 'standard' model: SBBN), the nucleon (or 'baryon') abundance. Then, each element will be considered in turn in a confrontation between the predictions of SBBN and the observational data. At present (summer 1999) there is remarkable agreement between the SBBN predictions of the abundances of four nuclides (D, ^{3}He, ^{4}He and ^{7}Li) and their primordial abundances inferred from the observations. However, there are some hints that this concordance of the hot big bang model may be imperfect, so we will also explore some variations on the theme of the standard model with regard to their modifications of the predicted primordial abundances of the light elements.

In the simplest, standard, hot big bang model the currently observed large-scale isotropy and homogeneity of the universe is assumed to apply during earlier epochs in its evolution (see COSMOLOGY: STANDARD MODEL). Given the currently observed universal expansion and the matter and radiation (CBR: 'cosmic background radiation', the 2.7 K 'black body radiation') content, it is a straightforward application of classical physics to extrapolate back to earlier epochs in the history of the universe (see COSMIC MICROWAVE BACKGROUND). At a time of order 0.1 s after the expansion began the universe was filled with a hot, dense plasma of particles. The most abundant were photons, electron–positron pairs, particle–antiparticle pairs of all known 'flavors' of neutrinos (ν_e, ν_μ and ν_τ) and trace amounts of neutrons and protons ('nucleons' or 'baryons'). At such early times the thermal energy of these particles was very high, of order a few MeV. With the exception of the nucleons, it is known or assumed that all the other particles present were extremely relativistic at this time. Given their high energies (and velocities close to, or exactly equal to, the speed of light) and high densities, the electroweak interactions among these particles were sufficiently rapid to have established thermal equilibrium. As a result, the numbers and distributions (of momentum and energy) of all these particles are accurately predicted by well-known physics.

Nucleosynthesis in the early universe

The primordial yields of light elements are determined by the competition between the expansion rate of the universe (the Hubble parameter H) and the rates of the weak and nuclear reactions (see HUBBLE CONSTANT). It is the weak interaction, interconverting neutrons and protons, that largely determines the amount of ^{4}He which may be synthesized, while detailed nuclear reaction rates regulate the production (and destruction) of the other light elements. In the standard model of cosmology the early expansion rate is fixed by the total energy density ρ,

$$H^2 = 8\pi G\rho/3 \tag{1}$$

where G is Newton's gravitational constant. In the standard model of particle physics the early energy density is dominated by the lightest, relativistic particles. For the epoch when the universe is a few tenths of a second old and older, and the temperature is less than a few MeV,

$$\rho = \rho_\gamma + \rho_e + N_\nu \rho_\nu \tag{2}$$

where ρ_γ, ρ_e and ρ_ν are the energy densities in photons, electrons and positrons, and massless neutrinos and antineutrinos (one species), respectively; N_ν is the number of massless (or, very light: $m_\nu \ll 1$ MeV) neutrino species which, in standard BBN, is exactly 3. In considering variations on the theme of the standard model, it is useful to allow N_ν to differ from 3 to account for the presence of 'new' particles and/or any suppression of the standard particles (e.g. if the τ neutrino should have a large mass). Since the energy density in relativistic particles scales as the fourth power of the temperature, the early expansion rate scales as the square of the temperature with a coefficient that depends on the number of different relativistic species. The more such species, the faster the universe expands, the earlier (higher temperature) will the weak and nuclear reactions drop out of equilibrium. It is useful to write the total energy density in terms of the photon energy density and g, the equivalent number of relativistic degrees of freedom (i.e. helicity states, modulo the different contributions to the energy density from fermions and bosons),

$$\rho \equiv (g/2)\rho_\gamma . \tag{3}$$

In the standard model at $T \sim 1$ MeV, $g_{SM} = 43/4$. Account may be taken of additional degrees of freedom by comparing their contribution to ρ with that of one additional light neutrino species:

$$\Delta\rho \equiv \rho_{TOT} - \rho_{SM} \equiv \Delta N_\nu \rho_\nu . \tag{4}$$

If the early energy density deviates from that of the standard model, the early expansion rate (or, equivalently,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the age at a fixed temperature) will change as well. The 'speed-up' factor $\xi \equiv H/H_{\rm SM}$ may be related to ΔN_ν by

$$\xi = (\rho/\rho_{\rm SM})^{1/2} = (1 + 7\,\Delta N_\nu/43)^{1/2}. \qquad (5)$$

As we will see shortly, the ^{4}He abundance is very sensitive to the early expansion rate while the abundances of the other light nuclides depend mainly on the nuclear reaction rates which scale with the nucleon (baryon) density. Since the baryon density is always changing as the universe expands, it is convenient to distinguish between models with different baryon densities using a dimensionless parameter which either is conserved or, changes in a known and calculable fashion. From the very early universe until now the number of baryons in a comoving volume has been preserved and the same is roughly true for photons since the end of BBN. Therefore, the ratio of number densities of baryons ($n_{\rm B}$) and photons (n_γ) provides just such a measure of the universal baryon abundance:

$$\eta \equiv (n_{\rm B}/n_\gamma)_0 \qquad \eta_{10} \equiv 10^{10}\eta. \qquad (6)$$

The universal density of photons at present (throughout this article the present epoch is indicated by the subscript '0') is dominated by those in the CBR (for $T_0 = 2.73$ K, $n_{\gamma 0} = 412$ cm^{-3}) so that the baryon density parameter $\Omega_{\rm B} \equiv (\rho_{\rm B}/\rho_{\rm c})_0$, the ratio of the present baryon density ($\rho_{\rm B}$) to the present critical density ($\rho_{\rm c}$), may be related to η and the present value of the Hubble parameter $H_0 \equiv 100h$ km s^{-1} Mpc^{-1},

$$\eta_{10} = 273\Omega_{\rm B}h^2. \qquad (7)$$

It should be noted that prior to electron–positron annihilation there were fewer photons in every comoving volume (by a factor very close to 4/11); this is automatically accounted for in all numerical BBN codes. It is simply a matter of consensus and convenience that the baryon abundance is quoted in terms of its present value.

In SBBN (i.e. $N_\nu = 3$) the abundances of the light nuclides synthesized primordially depend on only one 'free' parameter, η. SBBN is thus 'overconstrained' since one value (or, a narrow range of values set by the observational and theoretical reaction rate uncertainties) of η must account consistently for the primordial abundances of D, ^{3}He, ^{4}He and ^{7}Li. At the same time this value or range of η must be consistent with current estimates of (or, bounds to) the present baryon density. For these reasons BBN is one of the key pillars supporting the edifice of the standard model of cosmology and it is the only one which offers a glimpse of the earliest evolution of the universe. In the following we will first identify the key landmarks in the first 20 min in the evolution of the universe in order to identify the physical processes responsible for determining the primordial abundances of the light nuclides. Then, after presenting the SBBN predictions (as a function of η; see figure 1) we will review the current status of the observational data, as well as the steps necessary in order to go from 'here and now' to 'there and then' when using the data to infer the true primordial abundances. Then we will be in a position to assess the consistency of the standard model.

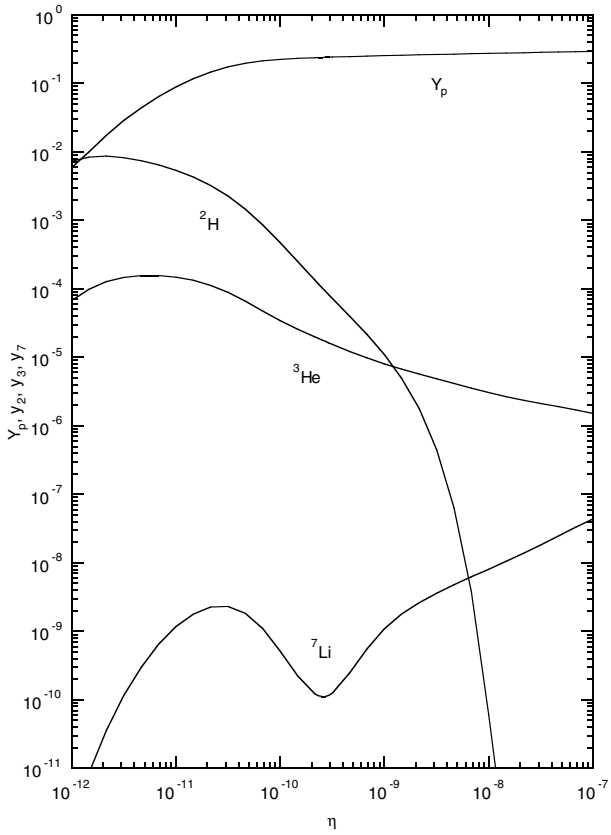


Figure 1. The predicted primordial abundances as a function of η. $Y_{\rm p}$ is the ^{4}He mass fraction while $y_{2\rm P}$, $y_{3\rm P}$, $y_{7\rm P}$ are the number density ratios to hydrogen of D, ^{3}He and ^{4}He respectively.

Weak equilibrium and the ^{4}He abundance

Consider now those early epochs when the universe was only a few tenths of a second old and the radiation filling it was at a temperature (thermal energy) of a few MeV. According to the standard model, at those early times the universe was a hot, dense 'soup' of relativistic particles (photons, $e^\pm$ pairs, three 'flavors' (e, μ, τ) of neutrino–antineutrino pairs) along with a trace amount (at the level of a few parts in 10^{10}) of neutrons and protons. At such high temperatures and densities both the weak and the nuclear reaction rates are sufficiently rapid (compared with the early universe expansion rate) that all particles have come to equilibrium. A key consequence of equilibrium is that the earlier history of the evolution of the universe is irrelevant for an understanding of BBN. When the temperature drops below a few MeV the weakly interacting neutrinos effectively decouple from the photons and $e^\pm$ pairs, but they still play an important role in regulating the neutron-to-proton ratio.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

At high temperatures, neutrons and protons are continuously interconverting via the weak interactions: $n + e^+ \leftrightarrow p + \bar{\nu}_e$, $n + \nu_e \leftrightarrow p + e^-$, and $n \leftrightarrow p + e^- + \bar{\nu}_e$. When the interconversion rate is faster than the expansion rate, the neutron-to-proton ratio tracks its equilibrium value, decreasing exponentially with temperature ($n/p = e^{-\Delta m/T}$, where $\Delta m = 1.29$ MeV is the neutron–proton mass difference). A comparison of the weak rates with the universal expansion rate reveals that equilibrium may be maintained until the temperature drops below ~0.8 MeV. When the interconversion rate becomes less than the expansion rate, the n/p ratio effectively 'freezes out' (at a value of ≈1/6), thereafter decreasing slowly, mainly as a result of free neutron decay.

Although n/p freeze-out occurs at a temperature below the deuterium binding energy, $E_B = 2.2$ MeV, the first link in the nucleosynthetic chain, $p + n \rightarrow D + \gamma$, is ineffective in jump-starting BBN since the photodestruction rate of deuterium ($\propto n_\gamma e^{-E_B/T}$) is much larger than the deuterium production rate ($\propto n_B$) owing to the very large universal photon-to-baryon ratio ($\gtrsim 10^9$). Thus, the universe must 'wait' until there are so few sufficiently energetic photons that deuterium becomes effectively stable against photodissociation. This occurs for temperatures $\gtrsim$80 keV, at which time neutrons are rapidly incorporated into ^{4}He with an efficiency of 99.99%. This efficiency is driven by the tight binding of the ^{4}He nucleus, along with the roadblock to further nucleosynthesis imposed by the absence of a stable nucleus at mass 5. By this time ($T \gtrsim 80$ keV), the n/p ratio has dropped to ~1/7, and simple counting (2 neutrons in every ^{4}He nucleus) yields an estimated primordial ^{4}He mass fraction

$$Y_P \approx \frac{2(n/p)}{1+n/p} \approx 0.25. \qquad (8)$$

As a result of its large binding energy and the gap at mass 5, the primordial abundance of ^{4}He is relatively insensitive to the nuclear reaction rates and, therefore, to the baryon abundance (η). As may be seen in figure 1, while η varies by orders of magnitude, the predicted ^{4}He mass fraction, Y_P, changes by factors of only a few. Indeed, for $1 \leq \eta_{10} \leq 10$, $0.22 \leq Y_P \leq 0.25$. As may be seen in figures 1 and 2, there is a very slight increase in Y_p with η. This is mainly due to BBN beginning earlier, when there are more neutrons available to form ^{4}He, if the baryon-to-photon ratio is higher. The increase in Y_p with η is logarithmic; over most of the interesting range in η, $\Delta Y_p \approx 0.01\,\Delta\eta/\eta$.

The ^{4}He abundance is, however, sensitive to the competition between the universal expansion rate (H) and the weak interaction rate (interconverting neutrons and protons). If the early universe should expand faster than predicted for the standard model, the weak interactions will drop out of equilibrium earlier, at a higher temperature, when the n/p ratio is higher. In this case, more neutrons will be available to be incorporated into ^{4}He and Y_p will increase. Numerical calculations show that, for a modest speed-up ($\Delta N_\nu \lesssim 1$), $\Delta Y_p \approx 0.013\,\Delta N_\nu$. Hence, constraints on Y_p (and η) lead directly to bounds on ΔN_ν and on particle physics beyond the standard model [1].

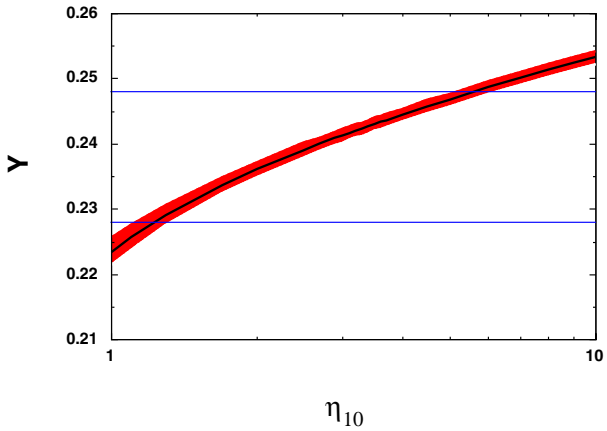


Figure 2. The predicted ^{4}He abundance (solid curve) and the 2σ theoretical uncertainty [2]. The horizontal lines show the range indicated by the observational data.

It should be noted that the uncertainty in the BBN-predicted mass fraction of ^{4}He is very small and almost entirely dominated by the (small) uncertainty in the n–p interconversion rates. These rates may be 'normalized' through the neutron lifetime, τ_n, whose current standard value is 887 ± 2 s (actually, 886.7 ± 1.9 s). To very good accuracy, a 1 s uncertainty in τ_n corresponds to an uncertainty in Y_p of order 2×10^{-4}. At this tiny level of uncertainty it is important to include finite-mass, zero- and finite-temperature radiative corrections, and Coulomb corrections to the weak rates. However, within the last few years it has emerged that the largest error in the BBN-prediction of Y_p was due to a too large time-step in the numerical code. With this now under control, it is estimated that the residual theoretical uncertainty (in addition to that from the uncertainty in τ_n) is of the order of 2 parts in 10^4. Indeed, a comparison of two major, independent BBN codes reveals agreement in the predicted values of Y_p to 0.0001 ± 0.0001 over the entire range $1 \leq \eta_{10} \leq 10$. In figure 2 is shown the BBN-predicted ^{4}He mass fraction, Y_p, as a function of η; the thickness of the band is the $\pm 2\sigma$ theoretical uncertainty. For $\eta_{10} \gtrsim 2$ the 1σ theoretical uncertainty in Y_p is $\lesssim 6 \times 10^{-4}$. As we will soon see, the current observational uncertainties in Y_p are much larger (see, also, figure 2).

Deuterium—the ideal baryometer

As may be seen in figure 1, the deuterium abundance (the ratio, by number, of deuterium to hydrogen: hereinafter, $(D/H)_P \equiv y_{2P}$) is a monotonic, rapidly decreasing function of the baryon abundance η. The reason for this behavior is easily understood. Once BBN begins in earnest, when the temperature drops below ~80 keV, D is rapidly burned to ^{3}H, ^{3}He and ^{4}He. The higher the baryon abundance, the faster the burning and the less D survives. For η_{10} in the 'interesting' range 1–10, y_{2P} decreases with the ~1.6

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

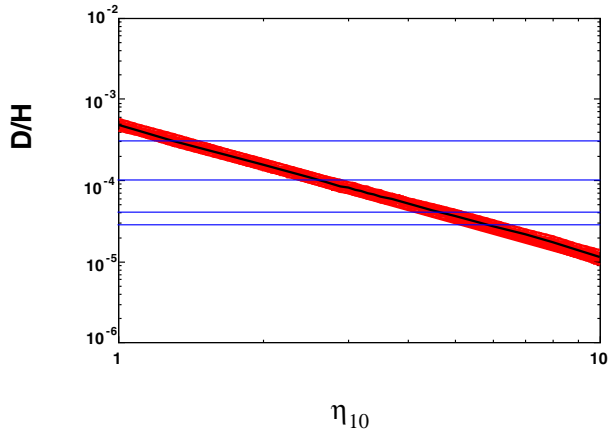


Figure 3. The predicted D/H abundance (solid curve) and the 2σ theoretical uncertainty [2]. The horizontal lines show the range indicated by the observational data for both high D/H (upper two lines) and low D/H (lower two lines).

power of η. As a result, a 10% error in y_{2P} corresponds to only a 6% error in η. This strong dependence of y_{2P} on η_{10}, combined with the simplicity of the evolution of D/H in the epochs following BBN, is responsible for the unique role of deuterium as a baryometer [3]. Because almost all the relevant reaction cross sections are measured in the laboratory at energies comparable with those of BBN, the theoretical uncertainties in the BBN-predicted abundance of deuterium is quite small, 8–10% for most of the interesting η range shown in figure 3.

Deuterium and helium-4 are complementary, forming the crucial link in testing the consistency of BBN in the standard model. While the primordial D abundance is very sensitive to the baryon density, the primordial ^{4}He abundance is relatively insensitive to η. Deuterium provides a bound on the universal baryon density while helium-4 constrains the early expansion rate of the universe, offering bounds on particle physics beyond the standard model.

Helium-3—complicated evolution

As may be seen in figure 1, the predicted primordial abundance of ^{3}He behaves similarly to that of D, decreasing monotonically with η. Again, the reason is the same: ^{3}He is being burned to the more tightly bound nucleus ^{4}He. The higher the nucleon abundance, the faster the burning and the less ^{3}He survives. In contrast to the behavior of D/H versus η, the decrease of ^{3}He/H with increasing η is much slower. This is simply a reflection of the much tighter binding of ^{3}He compared with D. Although the BBN predictions for the abundance of ^{3}He have similarly small uncertainties (8–10%) to those of D, it is much more difficult to exploit the observations of ^{3}He to test and constrain BBN. The complication is the evolutionary history of the ^{3}He abundance since BBN.

Although any deuterium cycled through stars is burned to ^{3}He during the stars' pre-main-sequence evolution, ^{3}He will survive in the cooler stellar exteriors while being destroyed in the hotter interiors. For the more abundant lower-mass stars which are cooler, a larger fraction of prestellar ^{3}He (along with the ^{3}He produced from prestellar D) survives. Indeed, for sufficiently low-mass stars (less than a few solar masses) incomplete burning actually leads to a buildup of newly synthesized ^{3}He (to be contrasted with the prestellar D and ^{3}He) which may—or may not—be returned to the interstellar medium. In fact, some PLANETARY NEBULAE are observed to be highly enriched in ^{3}He. So, the evolution of ^{3}He is complex with stellar destruction competing with primordial and stellar production. Indeed, if all low-mass stars were as prolific producers of ^{3}He as indicated by some planetary nebulae, the solar system and local interstellar medium abundances of ^{3}He should far exceed those inferred from observations. Thus, at least some low-mass stars must be net destroyers of ^{3}He. Given this necessarily complex and uncertain picture of production, destruction and survival, it is difficult to use current observational data to infer the primordial abundance of ^{3}He. Unless and until ^{3}He is observed in high-redshift (i.e. early universe), low-metallicity (i.e. nearly unevolved) systems, it will provide only a weak check on the consistency of BBN.

Lithium-7– the lithium valley

The trend of the BBN-predicted primordial abundance of lithium (almost entirely ^{7}Li) with η is more 'interesting' than that of the other light nuclides (see figures 1 and 4). The 'lithium valley', centered near $\eta_{10} \approx 2$–3, is the result of the competition between production and destruction in the two paths to mass 7 synthesis. At relatively low baryon abundance ($\eta_{10} \lesssim 2$) mass 7 is mainly synthesized as ^{7}Li via $^3\mathrm{H} + {}^4\mathrm{He} \rightarrow {}^7\mathrm{Li} + \gamma$. As the baryon abundance increases at low η, ^{7}Li is destroyed rapidly by $(\mathrm{p}, \gamma)2\alpha$ reactions, hence the decrease in $(\mathrm{Li/H})_P \equiv y_{7P}$ with increasing η seen (at low η) in figures 1 and 4. Were this the only route to primordial synthesis of mass 7, this monotonic trend would continue, similar to those for D and ^{3}He. However, mass 7 may also be synthesized via $^3\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^7\mathrm{Be} + \gamma$. The ^{7}Be will later capture an electron to become ^{7}Li. This channel is very important because it is much easier to destroy ^{7}Li than ^{7}Be. As a result, for relatively high baryon abundance ($\eta_{10} \gtrsim 3$) the latter channel dominates mass 7 production and y_{7P} increases with increasing η.

As may been seen in figure 4, the BBN-predicted uncertainties for y_{7P} are much larger than those for D, ^{3}He or ^{4}He. In the interval $1 \leq \eta_{10} \leq 10$ the 1σ uncertainties are typically ∼20%, although in a narrow range of η near the bottom of the 'valley' they are somewhat smaller (∼12–15%).

From here and now to there and then

To test the consistency of SBBN requires that we confront the predictions with the primordial abundances of the light nuclides which, however, are not observed but, rather, are inferred from observations. The path from the observational data to the primordial abundances is long

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

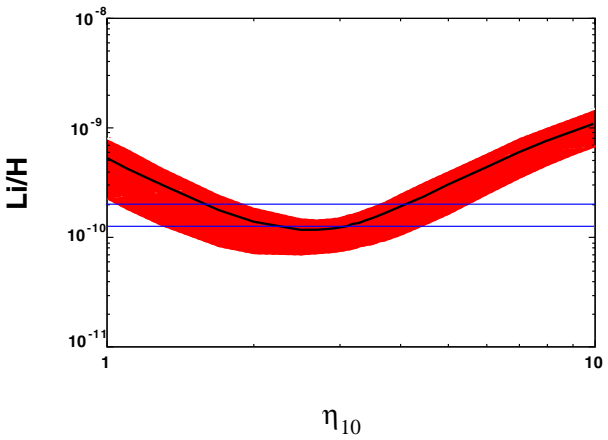


Figure 4. The predicted ^{7}Li abundance (solid curve) and the 2σ theoretical uncertainty [2]. The horizontal lines show the range indicated by the observational data.

and twisted and often fraught with peril. In addition to the usual statistical and systematic uncertainties, it is crucial to forge a connection from 'here and now' to 'there and then', i.e. to relate the derived abundances to their primordial values. It is indeed fortunate that each of the key elements is observed in different astrophysical sites using very different astronomical techniques. Also, the corrections for chemical evolution differ among them and, even more important, they can be minimized. For example, deuterium (and hydrogen) is mainly observed in cool, neutral gas (so-called H I REGIONS) via UV absorption from the atomic ground state (the Lyman series), while radio telescopes allow helium-3 to be studied utilizing the analog of the hydrogen 21 cm line for singly ionized ^{3}He in regions of hot, ionized gas (so called H II regions). Helium-4 is probed using the emission from its optical recombination lines formed in H II regions. In contrast, lithium is observed in the absorption spectra of warm, low-mass halo stars. With such different sites, with the mix of absorption–emission, and with the variety of telescopes and different detectors involved, the possibility of correlated errors biasing the comparison with the predictions of BBN is unlikely. This favorable situation extends to the obligatory evolutionary corrections. For example, although until recently observations of deuterium were limited to the solar system and the Galaxy, mandating uncertain evolutionary corrections to infer the pregalactic abundance, the Keck and HUBBLE SPACE TELESCOPES have begun to open the window to deuterium in high-redshift, low-metallicity, nearly primordial regions (Lyman-α clouds). Observations of ^{4}He in chemically unevolved, low-metallicity ($\sim$1/50 of solar) extragalactic H II regions permit the evolutionary correction to be reduced to the level of the statistical uncertainties. The abundances of lithium inferred from observations of the very metal-poor halo stars (one-thousandth of solar abundance and even lower) require almost no correction for chemical evolution. On the other hand, as noted earlier, the status of helium-3 is in contrast to that of the other light elements. For this reason, ^{3}He will not be used quantitatively in this article.

The currently very favorable observational and evolutionary situation for the nuclides produced during BBN is counterbalanced by the likely presence of systematic errors in the path from observations to primordial abundances. By their very nature, such errors are difficult—if not impossible—to quantify. In the key case of deuterium there is a very limited set of the most useful data. As a result, and although cosmological abundance determinations have taken their place in the current 'precision' era of cosmology, it is far from clear that the present abundance determinations are truly 'accurate'. Thus, the usual *caveat emptor* applies to any conclusions drawn from the subsequent comparison between the predictions and the data. With this caution in mind the current status of the data will be surveyed in order to infer 'reasonable' ranges for the primordial abundances of the key light elements.

Deuterium

Since deuterium is completely burned away whenever it is cycled through stars, and there are no astrophysical sites capable of producing deuterium in anywhere near its observed abundance [3], any D abundance derived from observational data provides a lower bound to its primordial abundance. Thus, without having to correct for Galactic evolution, the 'here-and-now' deuterium abundance inferred from UV observations along 12 lines of sight in the local interstellar medium (LISM) bounds the 'there-and-then' primordial abundance from below $((\mathrm{D/H})_\mathrm{P} \geq (\mathrm{D/H})_\mathrm{LISM} = (1.5 \pm 0.1) \times 10^{-5})$. As may be seen from figures 1 and 2, any lower bound to primordial D will provide an upper bound to the baryon-to-photon ratio [4].

Solar system observations of ^{3}He permit an independent, albeit indirect, determination of the pre-solar-system deuterium abundance [5]. This estimate of the Galactic abundance some 4.5 Gyr ago, while somewhat higher than the LISM value, has a larger uncertainty ($\mathrm{D/H} = (2.1 \pm 0.5) \times 10^{-5}$ [6]). Within the uncertainty it is consistent with the LISM value suggesting there has been only modest destruction of deuterium in the last 4.5 Gyr. There is also a recent measurement of deuterium in the atmosphere of Jupiter using the Galileo Probe Mass Spectrometer [7] ($\mathrm{D/H} = (2.6 \pm 0.7) \times 10^{-5}$ (see GALILEO MISSION TO JUPITER)).

To further exploit the solar system and/or LISM deuterium determinations to constrain or estimate the primordial abundance would require corrections for the Galactic evolution of D; see GALACTIC EVOLUTION. Although the simplicity of the evolution of deuterium (it is only destroyed) suggests that such correction might be very nearly independent of the details of specific chemical evolution models, large differences remain between different estimates. It is therefore fortunate that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

data on D/H in high-redshift (nearly primordial), low-metallicity (nearly unevolved) Lyman-α absorbers have become available in recent years [8, 9]. It is expected that such systems still retain their original, primordial deuterium, undiluted by the deuterium-depleted debris of any significant stellar evolution. That is the good news. The bad news is that, at present, D abundance determinations are claimed for only three such systems, and that the abundances inferred for two of them (along with a limit to the abundance for a third) appear to be inconsistent with the abundance estimated for the remaining one. Here we have a prime illustration of 'precise', but possibly inaccurate, cosmological data. Indeed, there is a serious obstacle inherent to using absorption spectra to measure the deuterium abundance since the isotope-shifted deuterium absorption lines are indistinguishable from the corresponding lines in suitably velocity-shifted ($-81\,\mathrm{km\,s^{-1}}$) hydrogen. Such 'interlopers' may have been responsible for some of the early claims [8] of a 'high' deuterium abundance [10]. Indeed, an interloper may be responsible for the one surviving high-D claim [11]. At present it seems that only three good candidates for nearly primordial deuterium have emerged from ground- and space-based observations. It may be premature to constrain cosmology on the basis of such sparse data. Nonetheless, the two 'low-D' systems suggest a primordial deuterium abundance consistent with estimates of the pre-Galactic value inferred from LISM and solar system data ($(\mathrm{D/H})_\mathrm{P} = (3.4 \pm 0.25) \times 10^{-5}$ [12]). To illustrate the confrontation of cosmological theory with observational data, this D abundance will be adopted in the following. However, the consequences of choosing the 'high-D' abundance ($(\mathrm{D/H})_\mathrm{P} = (20 \pm 5) \times 10^{-5}$ [11]) will also be discussed.

Helium-4

As the second most abundant nuclide in the universe (after hydrogen), the abundance of ^{4}He can be determined to high accuracy at sites throughout the universe. However, as stars evolve they burn hydrogen to helium and the ^{4}He in the debris of STELLAR EVOLUTION contaminates any primordial ^{4}He. Since any attempt to correct for stellar evolution will be inherently uncertain, it is sensible to concentrate on the data from low-metallicity sites. Extragalactic regions of hot, ionized gas (H II regions) provide such sites, where the helium is revealed via emission lines formed when singly and doubly ionized helium recombines. As with deuterium, current data provide ambiguous estimates of the primordial helium abundance. Since the differences ($\Delta Y = 0.010$) are larger than the statistical uncertainties ($\lesssim \pm 0.003$), systematic errors probably dominate. Among the currently most likely sources of such errors are uncertain corrections for collisional excitation in helium, uncertain corrections for unseen neutral helium and/or hydrogen, and underlying stellar absorption (leading to an underestimate of the true strength of the helium emission lines). In contrast, since the most metal poor of the observed regions have metallicities of order 1/50–1/30 of solar, the extrapolation from the lowest-metallicity regions to truly primordial introduces an uncertainty no larger than the statistical error.

Using published data [13, 14] for 40 low-metallicity regions, one group [15] finds $Y_\mathrm{P} = 0.234 \pm 0.003$. In contrast, from an independent data set of 45 low-metallicity regions another group [16] infers $Y_\mathrm{P} = 0.244 \pm 0.002$. Clearly, these results are statistically inconsistent. It is crucial that high priority be assigned to further H II region observations to estimate or avoid the systematic errors. Until then, since the error budget for Y_P is probably dominated by systematic rather than statistical uncertainties, in what follows a generous range for Y_P will be adopted: $0.228 \leq Y_\mathrm{P} \leq 0.248$.

Lithium-7

Cosmologically interesting lithium is observed in the spectra of nearly 100 very metal poor halo stars [17, 18] (see GALACTIC METAL-POOR HALO). These stars are so metal poor they provide a sample of more nearly primordial material than anything observed anywhere else in the universe; the most metal-poor among them have heavy-element abundances less than one-thousandth of the solar metallicity. However, these halo stars are also the oldest objects in the Galaxy and, as such, have had the most time to modify their surface abundances. So, even though any correction for Galactic evolution modifying their lithium abundances may be smaller than the statistical uncertainties of a given measurement, the systematic uncertainty associated with the dilution and/or destruction of surface lithium in these very old stars could dominate the error budget. There could be additional errors associated with the modeling of the surface layers of these cool, low-metallicity, low-mass stars needed to derive abundances from absorption-line spectra. It is also possible that some of the Li observed in these stars is non-primordial (e.g. that some of the observed Li may have been produced post-BBN by spallation reactions (the breakup of C, N and O nuclei into nuclei of Li, Be and B) or fusion reactions ($\alpha + \alpha$ to form ^{6}Li and ^{7}Li) in cosmic-ray collisions with gas in the ISM). In a recent analysis [19], it is argued that as much as $\sim$0.2 dex of the observed lithium abundance, $A(\mathrm{Li}) \equiv 12 + \log(\mathrm{Li/H})$, could be post-primordial in origin.

The very large data set of lithium abundances measured in the warmer ($T > 5800$ K), more metal-poor ([Fe/H] < -1.3) halo stars defines a plateau (the 'Spite-plateau' [17]) in the lithium abundance–metallicity plane. Depending on the choice of stellar-temperature scale and stellar atmosphere model, the abundance level of the plateau is $A(\mathrm{Li}) = 2.2 \pm 0.1$, with very little dispersion in abundances around this plateau value. The small dispersion provides an important constraint on models which attempt to connect the present surface lithium abundances in these stars to the original lithium abundance in the gas out of which these stars were formed some 10–15 Gyr ago. 'Standard' (i.e. non-rotating) stellar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

models predict almost no lithium depletion and, therefore, are consistent with no dispersion about the Spite plateau. Although early work on mixing in models of rotating stars was very uncertain, recently stellar models have been constructed which reproduce the angular momentum evolution observed for the much younger, low-mass open cluster stars. These models have been applied to the study of lithium depletion in main sequence halo stars. A well-defined lithium plateau with modest scatter and a small population of 'outliers' (overdepleted stars), consistent with the current data, is predicted for depletion factors between 0.2 dex and 0.4 dex [20].

To err on the side of caution, a generous range for the plateau abundance, $2.1 \leq A(\mathrm{Li}) \leq 2.3$, is adopted. If depletion is absent, this range is consistent with the primordial lithium 'valley' (see figure 4). For depletion $\geq$0.2 dex, the consistent primordial lithium abundances bifurcate and move up into the 'foothills', although a non-negligible contribution from post-BBN lithium could move the primordial abundance back down again (figure 4).

Confrontation of theory with data

In the context of the 'standard' model (three families of massless, or light, two-component neutrinos), the predictions of BBN (SBBN) depend on only one free parameter, the nucleon-to-photon ratio η. The key test of the standard, hot, big bang cosmology is to assess whether there exists a unique value or range of η for which the predictions of the primordial abundances are consistent with the light-element abundances inferred from the observational data. From a statistical point of view it might be preferable to perform a simultaneous fit of the inferred primordial abundances of D, ^{3}He, ^{4}He and ^{7}Li to the SBBN predictions. In this manner the 'best fit' η, along with its probability distribution may be found, and the 'goodness of fit' assessed [21]. However, since systematic uncertainties most likely dominate observational errors at present, the value of this approach is compromised. An alternative approach is adopted here.

As emphasized earlier, deuterium is an ideal baryometer. As a first step the primordial abundance of deuterium inferred from observations at high redshift will be compared with the SBBN prediction to identify a consistent range for η. Then, given this range, the SBBN abundances of ^{4}He and ^{7}Li are predicted and these are compared with the corresponding primordial abundances derived from the observational data. The challenge is to see whether the D-identified range for η leads to consistent predictions for ^{4}He and ^{7}Li. Recall that, because of its complicated evolutionary history, it is difficult to use ^{3}He to test and constrain SBBN. Furthermore, another consistency test is to compare the SBBN-inferred η range with the present baryon density derived from non-BBN observations and theory. Is our model for the very early evolution of the universe consistent with the present universe?

From the two well-observed, high-redshift absorption line systems with 'low D', the estimate adopted for the primordial D abundance is $(\mathrm{D/H})_\mathrm{P} = (2.9 - 4.0) \times 10^{-5}$ (see figure 3). Also shown for comparison in figure 3 is the allowed range of the primordial deuterium abundance suggested by the 'high-D' abundance inferred from observations of one lower-redshift absorption-line system. With allowance for the ~8% uncertainty in the theoretically predicted abundance, the favored range (low D) for η is quite narrow: $\eta_{10} = 5.1 \pm 0.36$. It is clear from figure 4 that, for the baryon abundance in this range, the BBN-predicted lithium abundance is entirely consistent with the Spite plateau value, even if the plateau were raised by ~0.2 dex to allow for modest stellar destruction–dilution or lowered by a similar amount owing to post-BBN production. For this narrow range in η the predicted ^{4}He mass fraction varies very little. For $\eta_{10} \approx 5$, $\Delta Y_\mathrm{P} \approx 0.010\,\Delta\eta/\eta$, so that, including the error in the predicted abundance, $Y_\mathrm{P} = 0.247 \pm 0.001$. As may be verified from figure 2, this is within (albeit at the high end of) the range allowed by the data from the low-metallicity, extragalactic H II regions. Given the current uncertainties in the primordial abundances, SBBN is consistent with 'low-D, high-η'.

The significance of this concordance cannot be underestimated. A glance at figure 1 provides a reminder of the enormous ranges for the predicted primordial abundances. That the simplest hot, big bang cosmological model can account for ('predict') three independent abundances (four with ^{3}He; although ^{3}He has not been employed in this comparison, its predicted abundance is consistent with extant observational data) by adjusting only one free parameter (η) is a striking success. The theory, which is in principle falsifiable, has passed the test. It need not have. Indeed, future observational data coupled to better understanding of systematic errors may provide new challenges. For example, if in the future it should be determined that the primordial helium mass fraction were lower than $Y_\mathrm{P} = 0.245$, this would be inconsistent (within the errors) with the 'low-D, high-η' range derived above. Similarly, if the best estimate for the D-determined η range changed, the comparison performed above should be repeated. With this in mind, what of the 'high-D, low-η' range which has been set aside in the current comparison?

If, in contrast to the deuterium abundance adopted above, the true value were higher, $(\mathrm{D/H})_\mathrm{P} = (10\text{–}30) \times 10^{-5}$, the SBBN-favored range in η would be lower (see figure 3). Accounting for the ~8% uncertainty in the theoretically predicted abundance, $\eta_{10} = 1.7 \pm 0.28$. Inspection of figures 2–4 reveals that, as long as $\eta_{10} \gtrsim 1.1\text{–}1.3$, consistency with ^{4}He and ^{7}Li can be obtained. Hence, for 'high D' as well, the standard model passes the key cosmological test.

Comparison of BBN with non-BBN baryon density estimates

Having established the internal consistency of primordial nucleosynthesis in the standard model, it is necessary to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

proceed to the next key test. Does the nucleon abundance inferred from processes which occurred during the first thousand seconds of the evolution of the universe agree with estimates–bounds to the nucleon density in the present universe?

It is a daunting task to attempt to inventory the baryons in the universe. Since many (most?) baryons may be 'dark', such approaches can best set lower bounds to the present ratio of baryons to photons. One such estimate [22] suggests a very weak lower bound on η of $\eta_{10} \geq 0.25$, entirely consistent with the BBN estimates above. Others [23] have used more subjective (although cautious) estimates of the uncertainties, finding a much higher lower bound to the global budget of baryons: $\eta_{10} \geq 1.5$, which is still consistent with the 'low-η' range identified using the high-D results.

A possible challenge to the 'low-η' case comes from an analysis [24] which employed observational constraints on the Hubble parameter, the age of the universe, the structure-formation 'shape' parameter and the x-ray cluster gas fraction to provide non-BBN constraints on the present density of baryons, finding that $\eta_{10} \geq 5$ may be favored over $\eta_{10} \leq 2$. Even so, a significant low-η, high-D range still survives.

Cosmology constrains particle physics

Limits on particle physics beyond its standard model are mostly sensitive to the bounds imposed on the ^{4}He abundance (see also STANDARD MODEL OF PARTICLE PHYSICS). As described earlier, the ^{4}He abundance is predominantly determined by the neutron-to-proton ratio just prior to nucleosynthesis. This ratio is determined by the competition between the weak interaction rates and the universal expansion rate. The latter can be modified from its standard model prediction by the presence of 'new' particles beyond those known or expected on the basis of the standard model of particle physics. For example, additional neutrino 'flavors' ($\Delta N_\nu > 0$), or other new particles, would increase the total energy density of the universe, thus increasing the expansion rate (see equations (4) and (5)), leaving more neutrons to form more ^{4}He. For ΔN_ν sufficiently small, the predicted primordial helium abundance scales nearly linearly with ΔN_ν: $\Delta Y \approx 0.013\,\Delta N_\nu$. As a result, an upper bound to $Y_{\rm p}$ coupled with a lower bound to η (since $Y_{\rm p}$ increases with increasing baryon abundance) will lead to an upper bound to ΔN_ν and a constraint on particle physics [1].

The constraints on $N_\nu = 3 + \Delta N_\nu$ as a function of the baryon-to-photon ratio η are shown in figure 5. For 'low D, high η' there is little room for any 'extra' particles: $\Delta N_\nu \lesssim 0.3$. This would eliminate a new neutrino flavor ($\Delta N_\nu = 1$) or a new scalar particle ($\Delta N_\nu = 4/7$) provided they were massless or light ($m \ll 1$ MeV) and interacted at least as strongly as the 'ordinary' neutrinos. In contrast, much weaker constraints are found for 'high D, low η' where N_ν as large as 5 ($\Delta N_\nu \leq 2$) may be allowed (see figure 5).

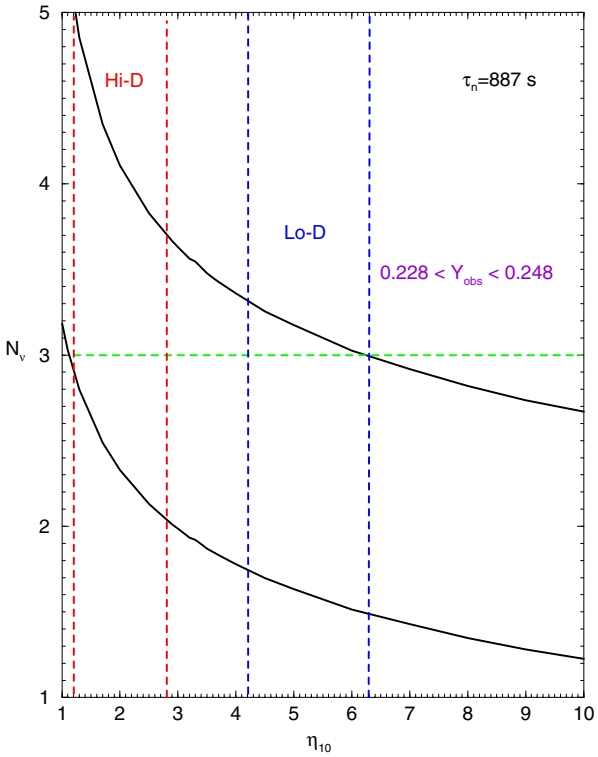


Figure 5. The number of equivalent massless neutrinos $N_\nu \equiv 3 + \Delta N_\nu$ (see equation (4)) as a function of η. The region allowed by an assumed primordial ^{4}He mass fraction in the range 0.228–0.248 lies between the two solid curves. The vertical bands bounded by the dashed lines show generous bounds on η from 'high' and 'low' deuterium.

Summary

The standard, hot, big bang cosmological model is simple (assuming isotropy, homogeneity, Newtonian–Einsteinian gravity, standard particle physics, etc) and, probably, simplistic. In broad brush it offers a remarkably successful framework for understanding observations of the present and recent universe. It may seem hubris to expect that this model could provide a realistic description of the universe during the first 20 min or so in its evolution when the temperature and density were enormously larger than today. According to the standard model, during these first 20 min the entire universe was a nuclear reactor, turning neutrons and protons into the light nuclides. This prediction presents the opportunity to use observations here and now to test the theory there and then. As described in this article, SBBN predicts observable primordial abundances for just four light nuclides, D, ^{3}He, ^{4}He and ^{7}Li, as a function of only one adjustable parameter, η, the nucleon-to-photon ratio, which is a measure of the universal baryon abundance. Even though it is currently difficult to use the extant observational data to bound the primordial abundance of ^{3}He, SBBN is still overconstrained, yielding three predicted abundances for one free parameter. Furthermore, the baryon density

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

inferred from SBBN must be consistent with that derived from observations of the present universe. Given the many possibilities that SBBN could be falsified by the empirical data, it is a remarkable success of the standard model that there is consistency between theory and observations provided that there are a few billion photons (most of them in the 2.7 K CBR) for every neutron or proton (nucleon) in the universe.

This success establishes primordial nucleosynthesis as one of the main pillars of our standard model of cosmology, providing the only probe of the physical universe during its very early evolution. Alternative theories of gravity and/or particle physics must now be tested against the impressive success of SBBN.

Bibliography

[1] Steigman G, Schramm D N and Gunn J 1977 *Phys. Lett.* B **66** 202
[2] Hata N, Scherrer R J, Steigman G, Thomas D and Walker T P 1996 *Astrophys. J.* **458** 637
[3] Epstein R, Lattimer J and Schramm D N 1976 *Nature* **263** 198
[4] Reeves H, Audouze J, Fowler W and Schramm D N 1976 *Astrophys. J.* **179** 909
[5] Geiss J and Reeves H 1972 *Astron. Astrophys.* **18** 126
[6] Geiss J and Gloeckler G 1998 *Space Sci. Rev.* **84** 239
[7] Mahaffy P R *et al* 1998 *Space Sci. Rev.* **84** 251
[8] Carswell R F, Rauch M, Weymann R J, Cooke A J and Webb J K 1994 *Mon. Not. R. Astron. Soc.* **268** L1
Songaila A, Cowie L L, Hogan C and Rugers M 1994 *Nature* **368** 599
[9] Tytler D, Fan X-M and Burles S 1996 *Nature* **381** 207
Burles S and Tytler D 1996 *Astrophys. J.* **460** 584
Burles S and Tytler D 1998 *Astrophys. J.* **499** 699
Burles S and Tytler D 1998 *Astrophys. J.* **507** 732
Burles S, Kirkman D and Tytler D 1999 *Astrophys. J.* **519** 18
[10] Steigman G 1994 *Mon. Not. R. Astron. Soc.* **269** 53
[11] Webb J K, Carswell R F, Lanzetta K M, Ferlet R, Lemoine M, Vidal-Madjar A and Bowen D V 1997 *Nature* **388** 250
Tytler D, Burles S, Lu L, Fan X M, Wolfe A and Savage B D 1999 *Astron. J.* **117** 63
[12] Burles S and Tytler D 1998 *Proc. 2nd Oak Ridge Symp. Atomic and Nuclear Astrophysics* ed P Mezzacappa (Bristol: Institute of Physics Publishing) p 113
[13] Pagel B E J, Simonson E A, Terlevich R J and Edmunds M 1992 *Mon. Not. R. Astron. Soc.* **255** 325
[14] Skillman E and Kennicutt R C 1993 *Astrophys. J.* **411** 655
Skillman E, Terlevich R J, Kennicutt R C, Garnett D R and Terlevich E 1994 *Astrophys. J.* **431** 172
[15] Olive K A and Steigman G 1995 *Astrophys. J. Suppl.* **97** 49
[16] Izotov Y I and Thuan T X 1998 *Astrophys. J.* **500** 188
[17] Spite F and Spite M 1982 *Astron. Astrophys.* **115** 357
Spite M, Maillard J P and Spite F 1984 *Astron. Astrophys.* **141** 56
Spite F and Spite M 1986 *Astron. Astrophys.* **163** 140
Hobbs L M and Duncan D K 1987 *Astrophys. J.* **317** 796
Rebolo R, Molaro P and Beckman J E 1988 *Astron. Astrophys.* **192** 192
Spite M, Spite F, Peterson R C and Chaffee F H Jr 1987 *Astron. Astrophys.* **172** L9
Rebolo R, Beckman J E and Molaro P 1987 *Astron. Astrophys.* **172** L17
Hobbs L M and Pilachowski C 1988 *Astrophys. J.* **326** L23
Hobbs L M and Thorburn J A 1991 *Astrophys. J.* **375** 116
Thorburn J A 1992 *Astrophys. J.* **399** L83
Pilachowski C A, Sneden C and Booth J 1993 *Astrophys. J.* **407** 699
Hobbs L and Thorburn J 1994 *Astrophys. J.* **428** L25
Thorburn J A and Beers T C 1993 *Astrophys. J.* **404** L13
Spite F and Spite M 1993 *Astron. Astrophys.* **279** L9
Norris J E, Ryan S G and Stringfellow G S 1994 *Astrophys. J.* **423** 386
[18] Ryan S, Norris J and Beers T 1999 *Astrophys. J.* **523** 654
[19] Ryan S, Beers T, Olive K A, Fields B D and Norris J 2000 *Astrophys. J.* **530** L57
[20] Pinsonneault M H, Walker T P, Steigman G and Narayanan V K 1999 *Astrophys. J* **527** 180
[21] Hata N, Scherrer R J, Steigman G, Thomas D and Walker T P 1996 *Astrophys. J.* **458** 637
[22] Persic M and Salucci P 1992 *Mon. Not. R. Astron. Soc.* **258** 14P
[23] Fukugita M, Hogan C J and Peebles P J E 1998 *Astrophys. J.* **503** 518
[24] Steigman G, Hata N and Felten J E 1999 *Astrophys. J.* **510** 564

Gary Steigman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Light Year (l.y.)

A unit of distance measurement equal to the distance traveled by a ray of light *in vacuo* in one year. Since light travels at a speed of 299 792 458 m s^{-1} (186 282 miles s^{-1}), this distance is equivalent to 9.46×10^{12} km (i.e. 9.46 million million km) or 63 240 astronomical units. The unit is commonly used to express large distances in the universe and represents the time taken for a ray of light to cross these distances. Approximately 3.262 light-years equal one parsec.

See also: light.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Light-Gathering Power (Light Grasp)

A measure of the amount of light collected and brought to a focus by a telescope. The amount of light collected is proportional to the surface area of the collector (objective or primary mirror) and so is proportional to the square of the aperture. For example, a telescope with an aperture of 2 m will collect four times as much light as a telescope with an aperture of 1 m.

The pupil of the dark-adapted human eye has an aperture of about 7 mm (0.007 m). Relative to the dark-adapted human eye, a telescope of aperture D (expressed in meters), has a light grasp, $G = D^2/(0.007)^2 \approx 20\,000D^2$. For example, each of the two Keck telescopes (the largest single-mirror telescopes in the world) has an aperture of 10 m and a light grasp of $20\,000 \times (10)^2 = 2000\,000$.

See also: aperture, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Limb

The edge of a celestial body which appears as a disk, either to the naked eye or telescopically. As the body moves across the sky, the 'leading' part of the limb is termed the *preceding limb*, and the opposite, 'lagging' part the *following limb*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lindblad, Bertil (1895–1965)

Swedish astronomer, became director of the Stockholm Observatory. He studied the velocities of stars, including dynamics of star clusters. He realized that if the Sun and stars were in orbit around the Galaxy, it must be rotating and the stars nearer to the galactic center must be rotating faster than those away from it, just as, in the solar system, the inner planets move faster than the outer ones. Stars that were not rotating around the Galaxy must be plunging in and out through the middle of the Galaxy. Thus he showed that the Galaxy was divided into several subsystems rotating about the same axis, each subsystem having its own rotation rate and degree of flattening. The 'high-velocity' stars are actually moving slowly, not around the Galaxy but through its center, and the Sun is moving quickly through them, at right angles to their radial motion. By considering the stars in orbit, Lindblad was able to determine an approximate distance to the center of the Galaxy, estimate the galactic mass and the period of the Sun's orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lipperhey [Lippershey], Hans [Jan] (1570?–1619)

Spectacle-maker, born in Wesel (Germany), settled in Middelburg, the capital of Zeeland, Netherlands, apparently the inventor of the telescope. Lipperhey applied for a patent. A letter of the government of Zeeland to its delegation to the States General of the Netherlands, dated 25 September 1608, instructs them to help the bearer, 'who claims to have a certain device by means of which all things at a very great distance can be seen as if they were nearby, by looking through glasses, which he claims to be a new invention'. The patent application, the first record of a telescope (although not the first of an idea of a telescope), was turned down because the device could not be kept a secret. Other claimants to the invention appeared immediately afterwards in the Netherlands, including Jacob Metius of Alkmaar, and Sacharias Janssen, like Lipperhey also a spectacle-maker in Middelburg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lithuanian Astronomy

Lithuanian folklore, archaic calendars and terminology show that Lithuanians were interested in astronomy from ancient times. A lot of celestial bodies have names of Lithuanian origin that are not related to widely accepted ancient Greek mythology. For example, the Milky Way is named '*Pauksciu Takas*' (literally the way of birds), the constellation of the Great Bear '*Didieji Grizulo Ratai*' (literally the great pole cart), the open cluster of Pleiades '*Sietynas*' (literally the sieve), the planet Jupiter '*Indraja*', etc.

Development of modern science in Lithuania began with establishment of Vilnius University in 1579, 10 years after formation of the Polish–Lithuanian Commonwealth. The beginning of astronomical research could be related to mathematician Oswald Krüger who brought the first telescope and started the first astronomical observations in Vilnius. The book '*Illustriora theoremata et problemata mathematica ex opticis geometria astronomia* …' by Joanes Rudomina Dusiatski (1633), Krüger's student, was one of the university's first publications devoted to astronomy. In this book Galileo's telescope and observations of planets made in Vilnius were described. However, internal conflicts and foreign invasion in the middle of the 17th century ruined the country's economy and suspended further development of science. The revival in the study of science including astronomy began in the middle of the 18th century. In 1753 the Astronomical Observatory of Vilnius University was established by Thomas Zebrowski. Premises for the observatory were built atop of the three-floor university building. The golden age of astronomy began when Marcin Odlianicki Poczobut was appointed the director of Astronomical Observatory in 1764 and remained in this position for more than 40 years. He considerably enlarged the observatory and acquired a lot of new books and instruments. Among them we should mention a transit instrument, a mural quadrant, a heliometer and three achromatic refractors. Poczobut and his students were diligent and skilful observers. They obtained a lot of observational data. They observed lunar and solar eclipses, positions of asteroids, planets and comets. The most important data of that period were observations of Mercury that Jerome J. de Lalande used for calculations of Mercury's orbit. These observational programs were continued by Poczobut's successors up to the 19th century.

The Polish–Lithuanian Commonwealth declined as a political power during the 18th century. Following three partitions the old state ceased to exist in 1795. The largest part of ethnic Lithuania including Vilnius passed to Russia. The university continued its activities without much interference. However, repressions followed the uprising of 1830–1831 against the Russian regime. In 1832 Vilnius University was closed. The observatory was entrusted to the Academy of Sciences of St Petersburg and continued its activities. Eventually, positions of astronomers at the observatory were taken over by Russian astronomers, mainly from Pulkovo Observatory. The new staff were not interested in astrometric observations and made a progressive decision to go over to astrophysics, a new field of astronomy at that time. For the astrophysical research new instruments were purchased, among them were a Merz spectroscope, a visual Schwerd's photometer and a photoheliograph. The first successful photographs of the Sun obtained in 1868 stimulated further observations of the Sun and monitoring of the solar activity. In addition, spectroscopic studies of the Sun and photometric observations of stars were carried out. Unfortunately, in 1876 a fire broke out at the observatory, causing heavy damage. The observatory did not receive means for refurbishment and five years later it was closed. The library and instruments were transferred to Pulkovo Observatory and other institutions of Russia. Revival of astronomy was possible only in the period between World Wars I and II. Then the larger part of ethnic Lithuania became independent with a temporary capital Kaunas. Poland occupied the smaller part of it including the old capital Vilnius. By 1920 Vilnius University was reopened and the Astronomical Observatory began its activities. In the following decade new observatory buildings at the outskirts of Vilnius were built and new instruments were purchased. Among these were two 15 cm Zeiss astrographs and a 48 cm telescope. The research was mainly concentrated on investigation of variable stars and on stellar statistics.

An astronomical observatory was established in Kaunas University in 1928. It was equipped with several small telescopes. The observatory planned to build a large telescope and purchased a 63 cm mirror for that purpose. However, the beginning of World War II and subsequent political alterations discontinued activities of the observatory. By the beginning of World War II Lithuania was incorporated into the Soviet Union until recovery of its independence in 1990.

After World War II activities of astronomers resumed at Vilnius University. Eventually, a new astronomical center at the Institute of Physics and Mathematics of the Academy of Sciences emerged. The postwar generation of astronomers concentrated on applications of stellar photometry for investigation of physical properties of stars and galactic structure. They studied the properties of the UBV photometric system and made some proposals for its revision. Later, they devised a new seven-color photometric system under Vytautas Straizys. The colors and the mean wavelengths of the Vilnius photometric system are as follows: U, 345 nm; P, 374 nm; X, 405 nm; Y, 466 nm; Z, 516 nm; V, 544 nm; S, 655 nm. This system enables temperatures, luminosities, metallicities and reddenings of various kinds of stars to be derived.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Recently, the Vilnius photometric system was combined with colors of the Stromgren system and a new Stromvil photometric system proposed. In the period 1968–1980 two observing stations were built, one at Moletai (Lithuania) and another at Maidanak Mountain (Central Asia, Uzbekistan). Moletai was equipped with 1.65 m and 63 cm reflectors and a 35 cm Maksutov telescope. At the Maidanak observing site 1 m and 48 cm telescopes were installed . The telescopes were mainly used for stellar photometry. After the disintegration of the Soviet Union and re-establishment of independent Lithuania the telescopes of the Maidanak Observatory were taken over by the Academy of Science of Uzbekistan. At the present time astronomical research is carried out in two universities, Vilnius University and Vilnius Pedagogical University, as well as in two institutes, the Institute of Theoretical Physics and Astronomy and the Institute of Physics. The main research fields are physical properties of stars and interstellar matter, the structure and evolution of the galaxy, structure and evolution of stellar systems, and design of astronomical instrumentation. The *Bulletin of Vilnius Astronomical Observatory* was published in 1957–1992. The international quarterly journal *Baltic Astronomy* has been published since 1992.

Bibliography

http://www.astro.lt/

Jokubas Sudzius

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Little Dumbbell Nebula (M76, NGC 650–51)

A planetary nebula in the constellation Perseus, position RA $01^{h}\ 42.3^{m}$, dec. $+51^{\circ}\ 34'$. At eleventh magnitude and measuring $2'$ by $1'$ it looks something like a miniature version of the Dumbbell Nebula (M27). Each of the two lobes has a separate NGC number.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Liu Xin (first century BC–first century AD)

Astronomer to the Chinese war-lord Wang Mang. A hundred years before Ptolemy he constructed astronomical tables, and listed 1080 stars in six magnitude categories. He determined the year to be 365.2502 days. In AD 4, he organized a national congress on the modern scale, attended by 1000 scientists.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Liverpool John Moores University Astrophysics Research Institute

Astrophysics at Liverpool John Moores University continues a tradition of astronomical endeavor on Merseyside that stretches back nearly 400 years. The Astrophysics Research Unit (ARI) is now located in a new, purpose-built building on the Mersey riverfront. Its primary objective is to undertake world-class research, allied to innovative teaching and wider educational projects. It is supported by University, Research Council and European funds and is the home of the Liverpool Telescope (LT) project.

The LT is the world's largest robotic telescope and will be fully operational as a UK national research facility in the Canaries in 2000. It has been designed and constructed by the university-owned company, Telescope Technologies Ltd, who are sited next to the ARI, and who are currently building several other large telescopes. The ARI runs a unique joint degree program in astrophysics with the University of Liverpool; a pioneering distance-learning course (with students all over the world), and a schools' program in science, technology and IT focused on access to the Liverpool Telescope.

For further information see www.livjm.ac.uk/astro/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lobate Scarp

A long, sinuous cliff-like feature consisting of a series of connected lobes. They are found chiefly on Mercury, where they are interpreted as thrust faults caused by compressive forces. They vary from 20 to 500 km in length and from a few hundred meters to 2 km in height. Their planet-wide distribution is taken as evidence of a global shrinking early in Mercury's history when the mantle and core cooled and contracted, and the crust 'puckered'. One of the longest is Discovery Rupes, which crosses craters; others are found in the Caloris Basin. The global contraction must therefore have happened after the era of major cratering, which ended about 3.9 billion years ago.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Local Group

Not long after EDWIN HUBBLE established that galaxies are 'island universes' similar to our home galaxy, the MILKY WAY, he realized that a few of these external galaxies are considerably closer to us than any others. In 1936 he first coined the term 'Local Group' in his famous book *The Realm of the Nebulae* to identify our nearest galactic neighbors. More than 60 yr later, the galaxies of the Local Group remain particularly important to astronomers because their proximity allows us to obtain our most detailed views of the properties of normal galaxies beyond our own. These nearby systems also provide our clearest views of how galaxies interact with one another in the relatively small volume of space of the Local Group.

The brightest members of the Local Group are so close to us that on a clear, dark night away from city lights it is possible to see them with the unaided eye: in the southern hemisphere the LARGE MAGELLANIC CLOUD (LMC) and SMALL MAGELLANIC CLOUD (SMC) shine brightly, while in the north the ANDROMEDA and TRIANGULUM GALAXY can be seen as faint smudges of light in the sky. These two galaxies are, in fact, the most distant objects visible with the naked eye. From both hemispheres, the gossamer glow of the Milky Way reveals the presence of billions of stars spread throughout the thin disk of our home galaxy. These five galaxies constitute the most luminous and massive members of the Local Group.

Although the existence, if not the true nature, of the five naked-eye Local Group galaxies and the Milky Way has been known to humans since antiquity, the first member identified telescopically was Messier 32 (more typically referred to as M32) by G-J Le Gentil in 1749. Since then, astronomers have steadily identified additional members of the Local Group. By the time Hubble first introduced the concept of the Local Group in 1936, he was able to list 11 galaxies that he considered to be members of the group. At present (1999), 43 galaxies can be catalogued as probable members of the Local Group; these systems are listed in table 1 along with the dates of discovery for each. Remarkably, more Local Group members have been found in the past 30 yr than in all previous human history. Also, the era of discovery is almost certainly not over as future surveys uncover more members or as new nearby galaxies are found serendipitously.

Why are the galaxies of the Local Group so difficult to identify? The principal reason is that, apart from our Milky Way Galaxy and the large Andromeda and Triangulum galaxies (known also as M31 and M33, respectively), the known members of the Local Group are DWARF GALAXIES. By definition, these systems have low intrinsic luminosities. They usually also exhibit very low surface brightness (see LOW SURFACE BRIGHTNESSES GALAXIES). This property is a measure of how spread out the galaxy's light is on the sky. In the case of nearly every dwarf galaxy member of the Local Group, the surface brightnesses are lower than that of the night sky, lending them a ghostly appearance and making them very difficult to detect, even at close range[1].

Much of the recent success in finding new Local Group members is due to the availability of the many large-scale photographic surveys of the sky carried out since the seminal Palomar Sky Survey of the 1950s. Soon after these surveys were begun, visual searches of the photographic plates identified new nearby galaxies. Starting in the 1970s, automated measurements and analyses of the plates from these surveys helped uncover nearly all of the most recently discovered Local Group members. However, even the most complete optical survey cannot find all of the galaxies in the sky. For example, searches for galaxies near the bright band of the Milky Way itself are severely hindered by the high stellar density in this part of the sky and by the clouds of gas and dust within the plane of our Galaxy. This INTERSTELLAR MATTER effectively blocks all optical light from distant objects, making it impossible to find galaxies lurking in the background. Ongoing and planned surveys in the infrared and radio wavelengths can penetrate the haze of the Milky Way by detecting radiation that is unaffected by dust obscuration. These searches are almost certain to reveal several new Local Group members in coming years.

Another complication in producing a complete census of the Local Group is the uncertainty involved with defining the group's boundary. The best way to establish this is to determine which local galaxies are gravitationally bound to one another. Since M31 and the Milky Way dominate the mass of all probable Local Group members, this process requires a good estimate of the masses of these two giant galaxies (see below). In addition, we need accurate information on the distances and motions of individual candidate Local Group galaxies to determine whether they are physically bound to the M31–Milky Way system. Table 1 lists the 43 galaxies that appear to be likely members of the Local Group based on this approach.

Although well defined, this method of identifying and counting members of the Local Group is highly uncertain. For example, apart from a few of the nearest galaxies, we cannot measure the PROPER MOTION—the angular movement across the sky—of external galaxies. A galaxy that may be moving towards or away from us at a moderate speed may be moving very rapidly across our line of sight. Thus, some of the galaxies in the table that we believe are bound to the Local Group may actually only be 'passing through the neighborhood'. A second problem is that distances to local galaxies are notoriously difficult to determine reliably. Methods that work for the Magellanic Clouds may not be applicable to other nearby galaxies, and vice

[1] Unlike the apparent brightness of a galaxy, the surface brightness of an extended object does not change as a function of distance—at least for distances up to a few hundred million light-years. This makes low-surface-brightness galaxies difficult to detect anywhere. Consequently, a very large number of low-surface-brightness galaxies may still remain hidden throughout the universe, enough possibly to fundamentally change our views of the distribution and numbers of galaxies in the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Galaxies of the Local Group.

Galaxy	Other name	Year of discovery	RA (2000)	Declination (2000)	Type	Subgroup	Distance (Mly)	V_T	Luminosity ($10^6 L_\odot$)	Mass ($10^6 M_\odot$)
M31	NGC 224	–	$00^h42.7^m$	$+41^\circ\ 16'$	SbI–II	M31	2.5	3.4	25 000	700 000
Milky Way		–	$17^h45.7^m$	$-29^\circ\ 01'$	Sbc	MW	0.03	–	8 300	350 000
M33	NGC 598	–	$01^h33.9^m$	$+30^\circ\ 40'$	ScII–III	M31	2.7	5.9	3 000	30 000
LMC		–	$05^h23.6^m$	$-69^\circ\ 45'$	IrrIII–IV	MW	0.16	0.4	2 100	20 000
SMC	NGC 292	–	$00^h52.7^m$	$-72^\circ\ 50'$	IrrIV–V	MW	0.19	2.0	580	1 000
WLM	DDO 221	1923	$00^h02.0^m$	$-15^\circ\ 28'$	IrrIV–V	LGC	3.0	10.4	500	150
M32	NGC 221	1749	$00^h42.7^m$	$+40^\circ\ 52'$	E2	M31	2.6	8.1	380	2 120
NGC 205	M110	1864	$00^h40.4^m$	$+41^\circ\ 41'$	E5p/dSph–N	M31	2.6	8.1	370	740
NGC 3109	DDO 236	1864	$10^h03.1^m$	$-26^\circ\ 10'$	IrrIV–V	N3109	4.1	9.9	160	6 550
IC 10	UGC 192	1895	$00^h20.4^m$	$+59^\circ\ 18'$	dIrr	M31	2.7	11.6	160	1 580
NGC 185	UGC 396	1864	$00^h39.0^m$	$+48^\circ\ 20'$	dSph/dE3p	M31	2.0	9.1	130	130
NGC 147	DDO 3	1864	$00^h33.2^m$	$+48^\circ\ 31'$	dSph/dE5	M31	2.3	9.4	130	110
NGC 6822	DDO 209	1864	$19^h44.9^m$	$-14^\circ\ 48'$	IrrIV–V	LGC	1.6	9.1	94	1 640
IC 5152		1895	$22^h02.7^m$	$-51^\circ\ 18'$	dIrr	LGC	5.2	11.2	70	400
IC 1613	DDO 8	1906	$01^h04.9^m$	$+02^\circ\ 08'$	IrrV	M31	2.3	9.6	64	795
Sextans A	DDO 75	1942	$10^h11.1^m$	$-04^\circ\ 43'$	dIrr	N3109	4.7	11.3	56	395
Sextans B	DDO 70	1955	$10^h00.0^m$	$+05^\circ\ 20'$	dIrr	N3109	4.4	11.4	41	885
Sagittarius		1994	$18^h55.1^m$	$-30^\circ\ 29'$	dSph-N	MW	0.08	4.0	18	–
Fornax		1938	$02^h40.0^m$	$-34^\circ\ 27'$	dSph	MW	0.45	7.6	16	68
Pegasus	DDO 216	1958	$23^h28.6^m$	$+14^\circ\ 45'$	dIrr/dSph	LGC	3.1	12.0	12	58
EGB 0427+63	UGCA 92	1984	$04^h32.0^m$	$+63^\circ\ 36'$	dIrr	M31	4.2	13.9	9.1	–
SagDIG	UKS1927-177	1977	$19^h30.0^m$	$-17^\circ\ 41'$	dIrr	LGC	3.4	13.5	6.9	9.6
And VII	Cassiopeia	1998	$23^h26.5^m$	$+50^\circ\ 42'$	dSph	M31	2.5	15.2	5.7	–
UKS2323-326	UGCA 438	1978	$23^h26.5^m$	$-32^\circ\ 23'$	dIrr	LGC	4.3	13.8	5.3	–
Leo I	DDO 74	1955	$10^h08.5^m$	$+12^\circ\ 19'$	dSph	MW	0.81	10.1	4.8	22
And I		1972	$00^h45.7^m$	$+38^\circ\ 00'$	dSph	M31	2.6	12.8	4.7	–
GR 8	DDO 155	1956	$12^h58.7^m$	$+14^\circ\ 13'$	dIrr	GR8	4.9	14.4	3.4	7.6
Leo A	DDO 69	1942	$09^h59.4^m$	$+30^\circ\ 45'$	dIrr	MW	2.2	12.8	3.0	11
And II		1972	$01^h16.5^m$	$+33^\circ\ 26'$	dSph	M31	1.7	12.7	2.4	–
Sculptor		1938	$01^h00.2^m$	$-33^\circ\ 43'$	dSph	MW	0.26	8.5	2.2	6.4
Antlia		1985	$10^h04.1^m$	$-27^\circ\ 20'$	dIrr/dSph	N3109	4.1	14.8	1.7	12
And VI	Peg dSph	1998	$23^h51.7^m$	$+24^\circ\ 36'$	dSph	M31	2.7	14.1	1.4	–
LGS 3	Pisces	1978	$01^h03.9^m$	$+21^\circ\ 53'$	dIrr/dSph	M31	2.6	14.3	1.3	13
And III		1972	$00^h35.3^m$	$+36^\circ\ 31'$	dSph	M31	2.5	14.2	1.1	–
And V		1998	$01^h10.3^m$	$+47^\circ\ 38'$	dSph	M31	2.6	15.0	1.0	–
Phoenix		1976	$01^h51.1^m$	$-44^\circ\ 27'$	dIrr/dSph	MW	1.4	13.2	0.9	33
DDO 210	Aquarius	1959	$20^h46.8^m$	$-12^\circ\ 51'$	dIrr/dSph	LGC	2.6	14.7	0.8	5.4
Tucana		1985	$22^h41.8^m$	$-64^\circ\ 25'$	dSph	LGC	2.9	15.2	0.6	–
Leo II	DDO 93	1950	$11^h13.5^m$	$+22^\circ\ 09'$	dSph	MW	0.66	12.0	0.6	9.7
Sextans		1990	$10^h13.1^m$	$-01^\circ\ 37'$	dSph	MW	0.28	10.3	0.5	19
Carina		1977	$06^h41.6^m$	$-50^\circ\ 58'$	dSph	MW	0.33	10.9	0.4	13
Ursa Minor	DDO 199	1955	$15^h09.2^m$	$+67^\circ\ 13'$	dSph	MW	0.21	10.3	0.3	23
Draco	DDO 208	1955	$17^h20.3^m$	$+57^\circ\ 55'$	dSph	MW	0.27	10.9	0.3	22

versa. It is quite common that galaxies once believed to be Local Group members are later removed from the list of members as we refine our estimates of their distances and motions. Indeed, two of the galaxies Hubble first proposed to be Local Group members are now known to be located much further away. Thus we can not only expect additions to the table of Local Group members as new galaxies are discovered but also subtractions from the list as we learn that some of the galaxies are not members after all.

Global properties of the galaxies of the Local Group

The individual galaxies of the Local Group span a large range of basic properties. The luminosities of Local Group galaxies range from a minimum of about 250 000 times the luminosity of the SUN, to a maximum of more than 20 billion times the luminosity of the Sun, a range of a factor of 75 000. It is also clear that the galaxies of the Local Group span a very large range in size even though their extreme dimensions are impossible to determine precisely. The smallest systems in the Local Group are approximately 1000 ly in diameter, while the luminous parts of the giant galaxies M31 and the Milky Way span over 100 000 ly from end to end. This range is comparable with the size range exhibited by mammals, from the tiny bumblebee bat to the blue whale. Figure 1 illustrates the relative sizes of the visible portions of most of the Local Group galaxies.

Virtually every major galaxy type is represented in the Local Group. M31, the Milky Way and M33 are all examples of SPIRAL GALAXIES, but each represents a slightly different subclass of this family of galaxies. M31, for example, exhibits a prominent central bulge and well-defined spiral arms throughout the thin disk of the galaxy; it is classified as an Sb galaxy (see GALAXIES: CLASSIFICATION). M33 has a very weak, possibly non-existent central bulge and very poorly defined spiral arms; it is an Sc galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. A family portrait of many of the members of the Local Group. The galaxies are shown to scale and to roughly the correct range of relative brightness. The Milky Way is shown as it may appear to a viewer located outside the galaxy. This montage was produced by Bruno Binggeli of the University of Basel.

Although it is clear that the Milky Way is a spiral galaxy—the thin structure of the Milky Way and the existence of a concentration of stars and GLOBULAR CLUSTERS towards the constellation Sagittarius all confirm this classification—the specific spiral subtype is extremely difficult to determine reliably from our unfavorable vantage point inside the Galaxy. Various indirect indicators suggest that our Galaxy can be classified as an Sbc galaxy, intermediate between the properties described for M31 and M33. Some recent studies at radio, infrared and optical wavelengths also suggest that our Galaxy contains an elongated central bar composed of old, metal-rich stars. If true, then the Milky Way is an example of a BARRED SPIRAL GALAXY, and its specific subtype is SBbc, where the upper-case 'B' denotes a barred system. All three galaxies are typically considered 'giant' spirals, despite the fact that M33 is only about 10% as luminous or massive as M31. By comparison, our Galaxy is about half as luminous and massive as M31.

All the remaining galaxies of the Local Group are dwarf systems of various types. The galaxies that exhibit the most variation in appearance and in their global properties are the DWARF IRREGULAR GALAXIES. The most massive example of this type of galaxy in the Local Group is the LMC, one of the satellites of the Milky Way and the second closest external galaxy to the Sun. The LMC is a massive dwarf, and its global properties place it near the ill-defined boundary separating dwarf irregular galaxies from small spirals. Careful studies reveal features in the LMC normally found in spiral galaxies, such as a central bar-like concentration of older stars, and some evidence of indistinct spiral arms. The LMC's companion, the SMC, is a true irregular with little sign of large-scale structure. The remaining irregular galaxies of the Local Group are identified in the table. All of these systems are substantially smaller and less luminous than the LMC.

The principal reason that dIrr galaxies appear to have such chaotic structures is that they typically form stars in small clumps called associations embedded within the galaxies. Because young stars are typically very luminous, they are often the most prominent stars seen in optical images of these types of dwarf galaxies. If the star-forming regions are irregularly distributed—and they usually are—they give a strong impression that the entire galaxy has a highly distorted, chaotic overall structure. Figure 2 illustrates the appearance of a dwarf irregular galaxy Sextans A, both from a ground-based image and from an image taken with the HUBBLE SPACE TELESCOPE. The luminous young blue and red stars are distributed non-uniformly, lending this galaxy its highly disorganized appearance.

This apparent lack of large-scale organization in dIrr galaxies is somewhat of an illusion produced by the luminous young stars that generally make up a minority of the total stellar population of such galaxies. Observations in the red and infrared are most sensitive to the light output of the dominant population of old and middle-aged stars in these galaxies. When astronomers try to determine the distribution of these oldest stars, even the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

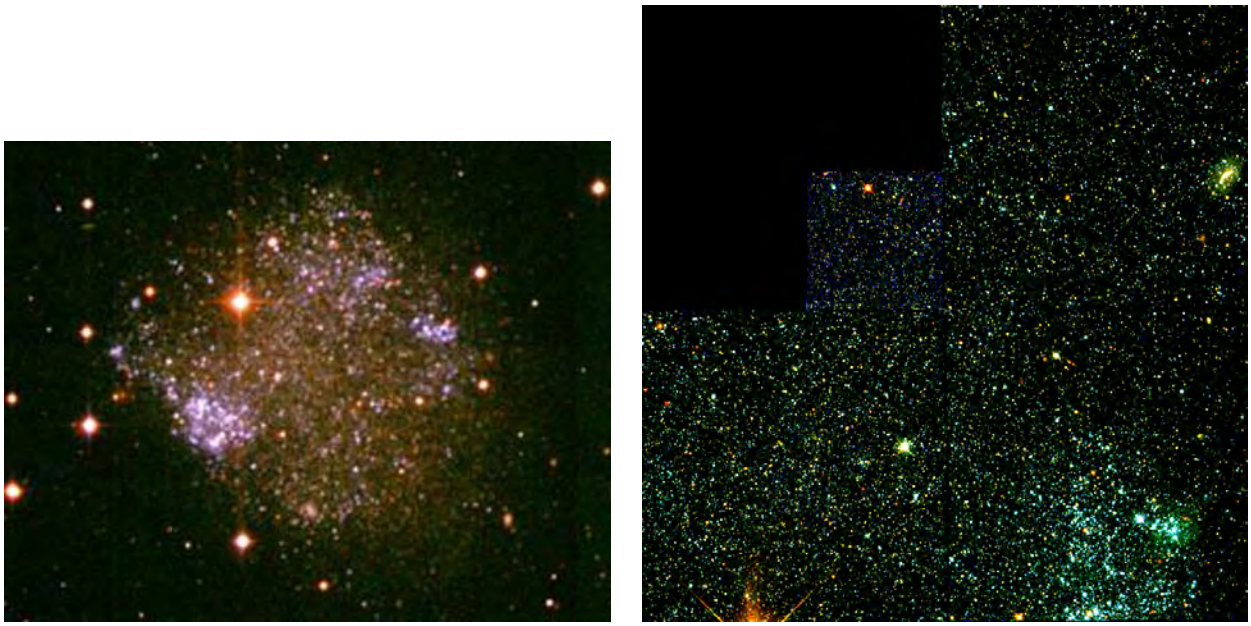


Figure 2. (Left panel) A ground-based color image of the Local Group irregular galaxy Sextans A. The entire central region of the galaxy is included. Notice how the bright clumps in the body of the galaxy are dominated by blue–white stars. Such stars are very hot and young objects indicating that these clumps are regions of relatively recent star formation in the galaxy. This image was taken by Diedre Hunter of Lowell Observatory. **This figure is reproduced as Color Plate 9.** (Right panel) An image of the central region of Sextans A obtained with the Hubble Space Telescope. In this image the individual stars in the clump located to the left of the galaxy's center in the left-hand panel are now easily apparent. Note too how red stars are more common away away from the bright clump of blue stars; these redder stars trace star formation in the galaxy long ago and show clearly that star formation has gone on over an extended period in Sextans A. This complex spatial and temporal star formation history is common for most of the dwarfs of the Local Group, especially the irregular galaxies. The HST photograph was provided by Robbie Dohm-Palmer of the University of Minnesota and the University of Michigan. **This figure is reproduced as Color Plate 10.**

most irregular dwarfs exhibit a much smoother, more symmetric appearance. An analogy is the surface of a pot of slowly boiling soup: although bubbles erupt at different locations at different times, the overall distribution of the soup is relatively uniform. The active star-forming regions of irregulars correspond to the bubbles: very prominent when active, but short lived and all interspersed in a more uniform medium.

With the striking exception of the Magellanic Clouds, the dIrr galaxies of the Local Group tend to be found far from the two large galaxies (M31 and the Milky Way) of the Local Group. A few of the least luminous dwarf irregulars are also among the most metal-poor galaxies known. Because stars produce heavy elements which they then eject back into the interstellar medium at the ends of their lives, the low abundances of such elements in the smaller dwarfs suggest that these galaxies are relatively pristine, unevolved systems. These galaxies may be forming stars in large quantities for the first time in their lives. As such, these galaxies are invaluable 'living fossils' that can tell us of the properties of gas and stars in the early universe.

One particularly interesting Local Group irregular galaxy is IC 10. This galaxy is forming stars at an unsustainably rapid rate. If it were to continue it would soon exhaust its raw materials for making stars (gas and dust) in only a few million years. The implication is that unless it started out with an astoundingly large reservoir of gas from which to form stars, IC 10 has probably been caught during a particularly active—but short-lived—phase in its star-formation history. Such galaxies that appear to form stars at unsustainably high rates are known as STARBURST GALAXIES. IC 10 is probably the closest example of a true starburst galaxy, although one active region of star formation in the LMC—the 30 Doradus Region—exhibits similar characteristics on a subgalactic scale.

The remaining dwarfs of the Local Group are ellipsoidal systems. These galaxies are characterized by a roughly circular or elliptical outline on the sky, and by a smooth, centrally concentrated distribution of light. One of these, M32, is considered to be an example of a true DWARF ELLIPTICAL GALAXY. As such, it represents the low-luminosity end of the very large family of elliptical galaxies which includes some of the largest, most luminous and most massive individual galaxies known. M32 is also noteworthy because it appears to harbor a massive BLACK HOLE in its extremely bright nucleus, it may be a unique local example of a galaxy with no ancient stars and, as a companion of M31, it shows distortions that indicate a strong gravitational interaction with its massive parent.

The remaining ellipsoidal Local Group galaxies are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

known as DWARF SPHEROIDAL GALAXIES (or dSph galaxies). About half of all galaxies in the Local Group are of this type and they are apparently similarly common in other groups and clusters. Consequently, these dim, unassuming galaxies probably represent the most common type of galaxy in the entire universe. Within the Local Group, dSph galaxies are typically found in the company of a larger parent galaxy. For example, of the 13 close companions of the Milky Way, nine are dSph satellites. Virtually all the remaining dSph galaxies of the Local Group are found near M31. The most luminous dSph galaxy, NGC205, is a highly distorted companion of M31 (figure 3). The two lowest-luminosity galaxies known are dSph companions of the Milky Way: the Draco and Ursa Minor systems. These galaxies emit less light than some individual globular clusters—massive, compact star clusters typically found within extended halos surrounding elliptical, spiral and larger dwarf galaxies (see HALO, GALACTIC). Nevertheless, the large dimensions and large masses (most in the form of matter that we cannot see directly) of galaxies such as Draco and Ursa Minor distinguish them from the more compact clusters.

Five of the dwarf galaxies of the Local Group are difficult to classify as ellipsoidal or irregular systems because they exhibit some of the properties of both. These 'transition' galaxies may represent the late stages of star-formation episodes in dwarf irregulars fortuitously caught during the period when the youngest stars begin to fade from prominence. This picture is consistent with the star-formation histories we measure for transition systems and with the growing evidence that dSph and dIrr galaxies have undergone complicated star-formation histories over the entire lifetime of the universe. A few hundred million years after periods of active star formation, a dSph galaxy that may initially have looked like a dIrr system could become a transition galaxy. This intimate relation between dIrr and dSph galaxies in which 'transition' systems act as a 'missing' evolutionary link remains controversial. Nonetheless, there is little doubt that some dSph galaxies looked a lot like low-luminosity dIrr systems at some point(s) in the past.

Galaxies are not the only inhabitants of intergalactic space within the Local Group. There is growing evidence of the existence of isolated clouds of gas, usually in the form of neutral hydrogen, distributed throughout the group. This material may represent gas expelled from other galaxies or may correspond to primordial matter that has yet to collapse into small stellar systems such as globular clusters or dwarf galaxies. Some remote globular star clusters could plausibly be 'free-floating' members of the Local Group that were ejected from their parent galaxies during past interactions of individual galaxies within the group.

One prominent class of normal galaxy is not found within the Local Group: giant elliptical galaxies. This is not entirely surprising; since large ellipticals are relatively rare in the local universe, we would have been 'lucky' to find one closeby. Moreover, giant ellipticals tend to be found in regions with a high density of galaxies, a manifestation of the so-called GALAXY MORPHOLOGY–DENSITY RELATION. Because the Local Group is a loose, low-density collection of galaxies, it would have been unusual—although not impossible—for it to contain one or more big elliptical galaxies. The nearest giant elliptical galaxy to us is probably MAFFEI 1. The slightly more distant Centaurus A (NGC 5128) is somewhat easier to study because it is not located so close to the Milky Way in the sky and suffers far less obscuration by interstellar dust in the plane of our Galaxy.

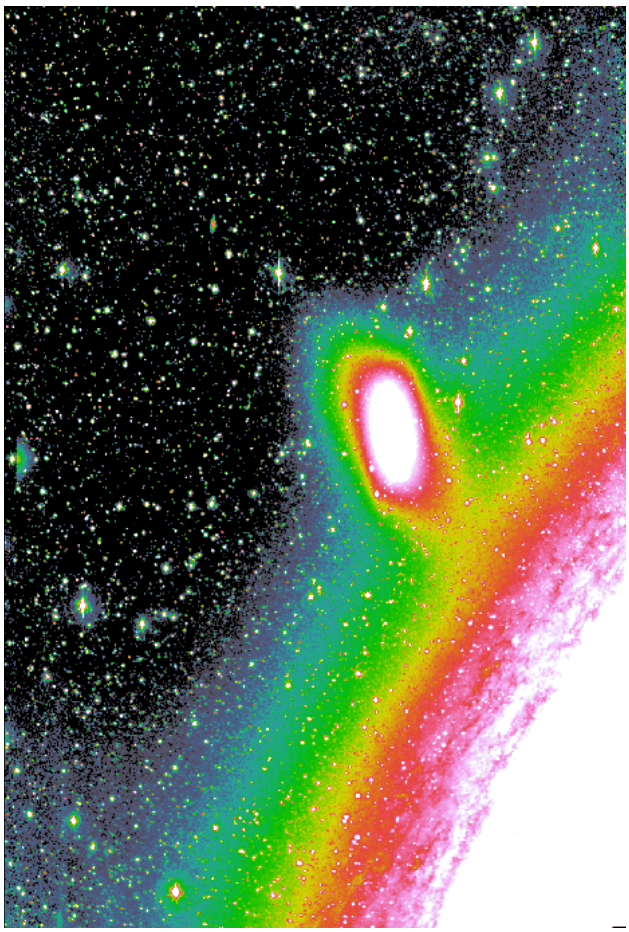


Figure 3. An image of NGC 205, a dwarf spheroidal companion of M31. The bright galaxy to the lower right of the figure is M31 itself. Note how close to the larger galaxy NGC 205 appears to be and how the outer extent of NGC 205 appears distorted from the smooth elliptical shape apparent near the galaxy core. This distortion is due to the strong tides that are induced in NGC 205 during its close passage to its much larger parent, M31. M32 and the Sagittarius dwarf also show evidence of tidal disruption by M31 and the Milky Way, respectively. The Magellanic Clouds show clear evidence of tidal distortion due to mutual interactions over the past few billion years. This image was obtained by Paul Harding, Heather Morrison and Anne Fry of Case Western Reserve University.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The motions of galaxies within the Local Group

The motions of galaxies in the Local Group appear to violate HUBBLE'S LAW that the universe is uniformly expanding. The reason for this has to do with the definition of the Local Group as the collection of nearby galaxies that are gravitationally bound to one another. Just as the planets of the solar system do not appear to be expanding away from us because they are bound to the Sun, the mutual attraction of galaxies within the Local Group has overcome the universal expansion in our immediate neighborhood. The most striking example of this is M31, which is currently approaching our galaxy at nearly 50 km s^{-1}: if on a true collision course, the two galaxies will meet in about 8–10 billion years. More likely, M31 and the Milky Way make up a binary system in which the two galaxies orbit their common center of gravity. Inevitably, tidal effects will cause the two galaxies to merge into one giant system at some time in the distant future. Many of the other galaxies of the Local Group also exhibit motions towards us, signifying that they too are either in orbit about the Milky Way or currently moving along orbits about M31 or other Local Group galaxies that cause them to move towards us at the present time. In all of these cases, the orbital motions are larger than the universal expansion velocities we would expect to measure for such nearby galaxies based on Hubble's law.

The mutual attraction of M31 and the Milky Way can be used to estimate the mass of the Local Group. Much as a ball thrown in the air first rises, stops and then falls, M31 and the Milky Way are now falling towards each other after their initial movement apart after the BIG BANG. If one measures the relative velocities and locations of the two galaxies and one estimates how long it has been since they were together—essentially the age of the universe estimated from the Hubble constant or from the ages of the oldest stars—it is possible to estimate the combined mass of M31 and the Milky Way. This sort of analysis was first carried out by F D Kahn and L Woltjer in 1959. Modern applications of this technique reveal that the combined mass of M31 and the Milky Way is in the range between 500 billion and 2 trillion times the mass of the Sun. Because virtually all the matter of the Local Group is located in these two giant spiral galaxies, this is also our best estimate of the total mass of the group.

As massive as the Local Group is, the total light output of all of the galaxies in the group M31 and the Milky Way is equivalent to 'only' 30 billion times the luminosity of the Sun. This is unusual because for normal stars the ratio of their mass (in units of the mass of Sun, or 2×10^{30} kg) and their luminosity (also in units of the luminosity of the Sun, or 4×10^{26} W) is around 1. For the Local Group the ratio is much higher, approximately 50. This suggests that the group is dominated by 'DARK MATTER' which contributes to the local gravitational field but remains invisible at any wavelength of ELECTROMAGNETIC RADIATION. The large mass-to-light ratio of the Local Group implies that only 2–5% of the total matter in the group is visible to us through emitted or reflected radiation.

We can also measure the masses of Local Group galaxies individually by determining the range of velocities of individual stars and gas clouds that orbit within individual systems. So long as Newton's law of gravity remains valid over the physical scales of the galaxies, astronomers find that Local Group galaxies typically exhibit mass-to-light ratios in the range 3–100 (see table 1). Thus, the dark matter within the Local Group is not spread out uniformly throughout the group, but is instead concentrated about the individual galaxies we detect optically. This suggests that possibility that the Local Group may contain some nearby examples of 'dark' galaxies consisting of only dark matter with no luminous material. No such systems have been detected yet. The dwarf galaxies of the Local Group do offer one important clue about the nature of dark matter. Certain types of dark matter that have been postulated in the past—in particular NEUTRINOS—cannot account for the surprisingly high mass densities required to account for the total masses of these small galaxies.

The distribution of galaxies within the Local Group

Figure 4 shows a stereoscopic picture of the Local Group. It is quite clear from this three-dimensional image of the Local Group that the group's volume is not uniformly filled with galaxies. Instead, most of the galaxies appear to congregate into three clumps; the remaining galaxies are spread out into a larger 'cloud' of galaxies that occupies a large fraction of the total volume of the Local Group. The smaller 'subgroups' correspond to M31 and its satellites, the Milky Way and its companions, and the 'NGC3109 subgroup' comprising five small galaxies that are well separated from M31 and the Milky Way. The galaxies of the more extended cloud—the so-called 'Local Group Cloud'—are not obviously associated with any single larger galaxy. Table 1 lists the subgroup with which each Local Group galaxy is most likely associated.

In general, each subgroup corresponds to a set of gravitationally bound galaxies. The 13 satellites of the Milky Way, for example, are all likely to be in orbit about our Galaxy while M31 maintains a similar stable of small companion galaxies. This situation is analogous to the moons that orbit individual planets of the Solar System: although Jupiter and Saturn orbit the Sun, both planets possess their own large families of bound satellites. In the case of the Local Group, the subgroups are weakly bound together, and there is evidence that some satellites may be occasionally 'swapped' from one subgroup to another. For example, some models of the dynamical evolution of the Local Group suggest that the Magellanic Clouds and perhaps the Leo I dSph galaxy may have first formed near M31, but are now satellites of our Galaxy.

Some Local Group galaxies show unmistakable signs of strong mutual gravitational interactions. As noted above, the Milky Way and M31 have overcome the initial expansion and are now falling in towards one another. The close companions of M31—NGC 205 and M32—both

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

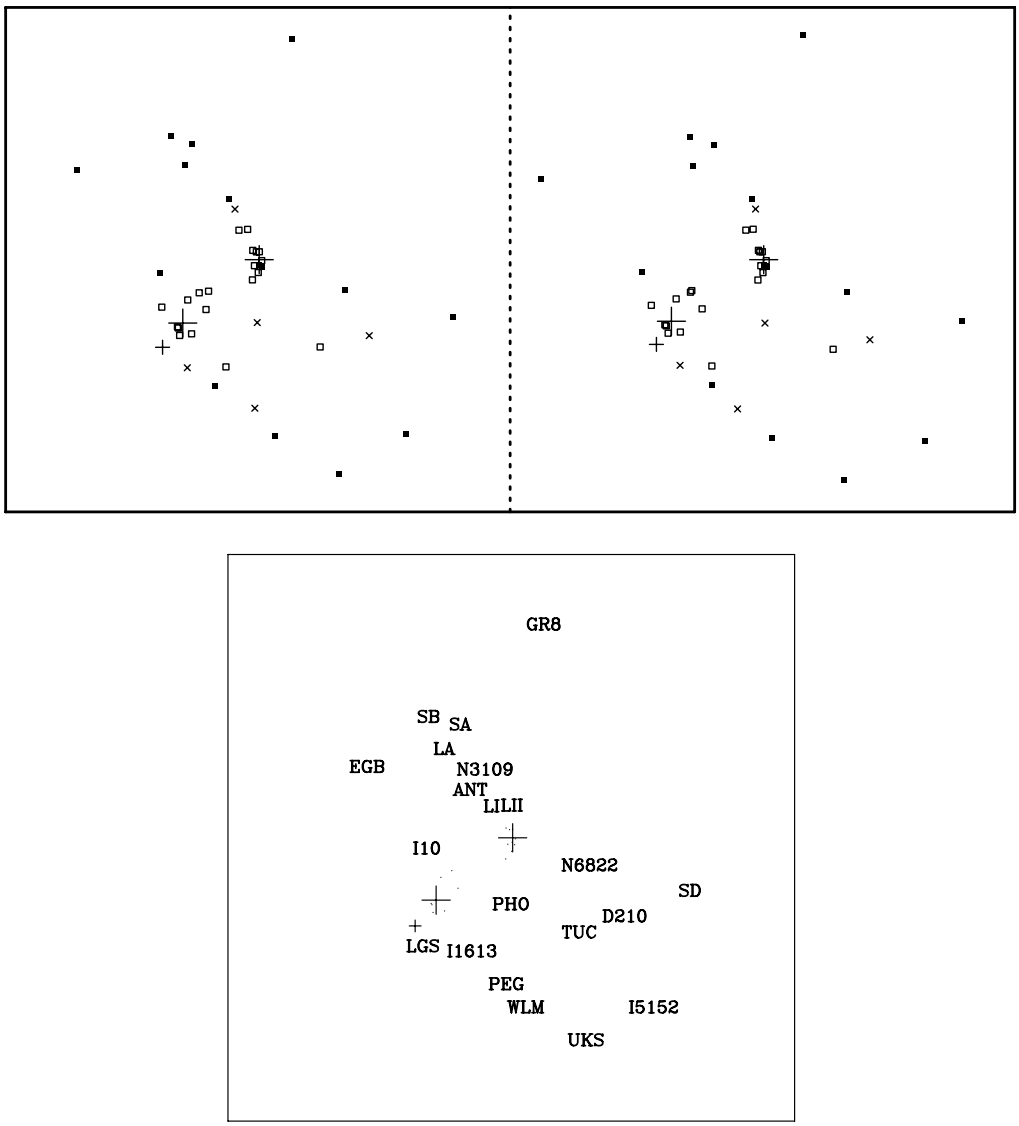


Figure 4. (Top panel) A stereo map of the Local Group showing its approximate three-dimensional structure. The view is from a point located about 7 million ly directly in front of the Sun's current motion about the center of the Milky Way. The large cross near the center of each panel denotes the position of the Milky Way, while the large off-center cross is M31. The smaller cross represents M33. The open squares show the locations of the dwarf ellipticals and dwarf spheroidal galaxies; the solid squares denote the dwarf irregulars; and the small ×s are the transition galaxies. The MW, M31 and NGC 3109 groups are easily visible; the latter is on the 'far' side of the figure. The large diffuse 'bubble' of galaxies extending to the lower right and towards the viewer is the Local Group Cloud. (Lower panel) A guide to the identities of individual galaxies in the stereo image. The systems are denoted with abbreviations that should—in conjunction with the names listed in table 1—uniquely define each galaxy. The galaxies closest to M31 and the Milky Way are not identified in order to minimize confusion in those crowded regions of the figure.

show global distortions that may be due to strong tidal effects induced by their close passage to their parent galaxy (figure 3 shows the distortions induced in NGC 205 by its close passage by M31). The Magellanic Clouds reveal considerable evidence that they too have interacted in the past. For example, the MAGELLANIC STREAM is a long arc of neutral gas that was probably ejected from one or both of the Clouds as they passed close to one another and passed the Milky Way. Because the SMC is the least massive component of the interacting pair, it has suffered the most. One clear sign of this is the fact that the SMC is elongated significantly along our line of sight, resembling a cigar whose long axis is roughly pointed towards us. This elongation reflects the stretching of the galaxy as it is literally pulled apart by its abusive neighbor. Detailed models of the dynamical evolution of the Magellanic Clouds indicate that they will both fall into the Milky Way in the next few billion years. Although the Clouds will be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

completely disrupted in the process, even the much-larger Milky Way will be affected as its disk is puffed up by the energy injected by the infalling dwarfs.

Perhaps the most spectacular example of an interaction within the Local Group involves the Sagittarius dSph galaxy. Discovered only in 1994, the SAGITTARIUS DWARF GALAXY is now known to be the closest galaxy to the Milky Way. Sadly for Sagittarius, this is too close, and the dwarf is being severely torn apart by the gravitational tidal forces exerted on it by our much more massive Galaxy. A clear indication that this process has already begun is the fact that Sagittarius has been stretched along a long arc that currently is known to extend nearly all of the way around the sky. Like the Magellanic Clouds, Sagittarius does not have long to live as a separate galaxy before it too is completely disrupted and merges into the main body of the Milky Way.

Cosmological implications of the Local Group

Although most studies of the origin of the universe focus on radiation coming from extremely distant galaxies, the Local Group also plays a critical role in cosmological studies. For example, with the advent of large ground-based and space-based telescopes and the use of high-sensitivity electronic detectors, it has become possible to measure the complete fossil record of star formation in many of the galaxies of the Local Group. These investigations provide a critical point of comparison with studies aimed at understanding how galaxies formed in the first place by studying objects detectable at the edge of the visible universe. So far, observations of Local Group systems reveal that normal galaxies have formed stars in unexpectedly complex and varied ways. Some nearby dwarf spheroidal galaxies are composed almost entirely of ancient stars formed soon after the big bang, but most of these galaxies formed large numbers of stars over long periods of time extending as recently as the past few hundred million years. The dwarf irregulars continue to form stars today, in some cases perhaps for the first time. Astronomers are just now beginning to try to reconcile these detailed observations of Local Group galaxies with studies of larger numbers of extremely remote galaxies.

The Local Group also offers insights into the question of whether small galaxies have merged over time to form larger systems such as M31, the Milky Way or the Magellanic Clouds. Some of the mergers can be studied today such as the disruption of the Sagittarius dwarf and the interactions of the Magellanic Clouds and our Galaxy described above. However, most mergers that led to the formation of larger galaxies must have occurred long ago when the universe was still quite young; we are only now beginning to understand how to disentangle evidence of these past encounters within our own Galaxy. The merger histories of Local Group galaxies will eventually shed new light on the conditions of the early universe and on the nature of the dark matter that helped drive the formation of galaxies in the first place.

On an even larger scale, the Local Group itself seems to interact significantly with other nearby groups of galaxies. One of these, the Sculptor Group, appears to be strongly elongated along a line projecting back to the Local Group. Because our group is considerably more massive, it is probable that the gravitational force of the Local Group has significantly distorted the Sculptor Group. One result of this interaction is the lack of a clear boundary between the Sculptor Group and our own. Some galaxies traditionally associated with Sculptor are occasionally catalogued with the Local Group and vice versa. Unlike the Local Group, Sculptor lacks any giant galaxies, but does contain a number of lower-luminosity spirals and many dwarf galaxies. On roughly the opposite side of the sky the M81–Maffei Group represents another nearby concentration of galaxies that may have a significant effect on the evolution and internal motions of the Local Group. Unlike the Sculptor Group, the M81–Maffei Group contains some giant galaxies, including the large spiral M81 and the elliptical galaxy Maffei 1. As with the Sculptor Group, some of the galaxies assigned to the M81–Maffei Group have been considered at times to be Local Group members and vice versa. On top of the interactions with neighboring groups, the Local Group is also 'falling' into the nearby VIRGO CLUSTER of galaxies. The attempt to understand the details and implications of these local interactions and large scale motions remains a very active field of modern research.

Bibliography

Hodge P 200 *An Atlas of Local Group Galaxies* (Dordrecht: Kluwer) (an atlas of photographs of all Local Group galaxies)

Mateo M 1998 Dwarf galaxies of the Local Group *Ann. Rev. Astron. Astrophys.* **36** 435–506 (a recent technical review of the Local Group)

Sandage A and Bedke J 1994 *The Carnegie Atlas of Galaxies* (Washington, DC: Carnegie Institution) (stunning photographs of many galaxies including most of the brighter members of the Local Group)

van den Bergh S 2000 *The Galaxies of the Local Group* (Cambridge: Cambridge University Press) (a galaxy-by-galaxy description of the Local Group; many historical comments included)

All of these contain many additional references.

Mario L Mateo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lockyer, Joseph Norman (1836–1920)

Civil servant, professor at the Solar Physics Observatory, Kensington (later moved to Cambridge), born in Rugby, Warwickshire, England. Attracted to astronomy by an observation of the annular solar eclipse of 1858, erected an observatory at his home in Hampstead, observed Mars. Lockyer showed spectroscopically that phenomena seen during a total eclipse (such as the prominences) could be seen in daytime on the Sun and were not lunar phenomena. These conclusions were simultaneous with conclusions by JULES JANSSEN, and were reported to the same meeting of the Académie des Sciences, which offered the pair a medal (in contrast to the squabble over precedence in Neptune's discovery, *see* JOHN ADAMS). Lockyer identified all the emissions in the solar spectrum except for one spectral line near the sodium emission. He suggested that the line was due to a new element which he named 'helium' after the Greek, helios, for Sun. Helium was isolated on the Earth 25 years later by William Ramsay (1895), and Lockyer was knighted. Lockyer founded *Nature*, a new journal dedicated to the natural sciences and continuing to this day. While on holiday in Greece he noticed that temples were oriented east–west and connected this alignment to sunrise at their foundation. He extended this thought to temples in Egypt and found an alignment corresponding to midsummer sunset in the main axis of the Great Temple of Amon-Re at Karnak. Other alignments were apparently connected with Sirius. At Stonehenge he hypothesized that the midsummer Sun rose originally over Heel stone, and calculated back to determine when it would have done so, to establish the date when Stonehenge was built. Although his work is regarded as incorrect, Lockyer's belief in the astronomical purposes of Stonehenge and other stone circles was the impetus behind this kind of research in the twentieth century (see GERALD HAWKINS) and he was thus the founder of astroarchaeology. He built the Hill Observatory, later the *Norman Lockyer Observatory* (Sidmouth, Devon), just before his death; it is still operational, with an active society of astronomers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Logan, James (1674–1751)

American civil servant, born in Lurgan, County Armagh, Ireland, emigrated (1699) to Pennsylvania to work for William Penn, and became one of the first men of science in the colonies. Took up mathematics, hybridized corn, invented the Conestoga wagon, got his own telescope, worked on the aberration of lenses, observed the heavens, and wrote a commentary on HALLEY's lunar tables.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lohrmann, Wilhelm Gotthelf (1796–1840)

Surveyor and amateur astronomer, born in Dresden, he was the second selenographer (after TOBIAS MAYER) to make and publish a lunar map based on quantitative measurements and geometric calculations, rather than merging eye-sketches. Lohmann's map was based on a diameter of the Moon of about a meter, over five times the diameter of Mayer's. He was helped and encouraged by CARL GAUSS and J ENCKE. Publication of the atlas (*Topography of the Visible Surface of the Moon*), including over 7000 craters, started in 1824 but was a protracted process, slowed by Lohrmann's failing eyesight, and not completed (supervised by Julius Schmidt) until 38 years after his death.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lomonosov, Mikhail V (1711–65)

Russian chemist and astronomer. Studied aurorae, and in 1761 observed the transit of Venus, concluding from the back-illuminated edge of the planetary disk that it had an atmosphere. Introduced Copernicanism and Newtonianism into Russia.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Long-period Comet

A comet whose period is more than 200 years. Comets with periods of less than 200 years are termed *periodic comets* or *short-period comets*. Long-period comets are also 'periodic', but a comet can be incontrovertibly classified as periodic only when it has been observed on at least two returns; the choice of 200 years as a cut-off is traditional, but arbitrary. It happens that no comet calculated to have a period of over 200 years has yet been observed to return since its discovery. The first to do so will be Comet Peters (C/1857 O1), expected to return in 2092. In the system of cometary designations introduced in 1995, 'C/' is the prefix assigned to long-period comets.

Long-period comets show a much greater spread of orbital inclinations than do short-period comets. Many have very long periods indeed, and some appear to be on open, hyperbolic orbits. However, when the apparent orbits of long-period comets round the Sun (osculating heliocentric orbits) are referred instead to the center of mass of the solar system (to yield barycentric orbits), they are all found to be elliptical, once allowance is made for planetary perturbations and non-gravitational forces. About one-third of all long-period comets are found to have periods close to a million years, corresponding to aphelia of rather less than 50 000 AU, while the periods of the other two-thirds are fairly evenly spread between a million years and 200 years. The million-year peak points to the origin of these comets in the Oort Cloud.

See also: comets, non-gravitational force, Oort Cloud, short-period comet, solar system: nomenclature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lorentz, Hendrik Antoon (1853–1928)

Born in Arnhem, Netherlands, became professor of mathematical physics at Leiden University. Nobel prizewinner 1902, jointly with PIETER ZEEMAN, for his mathematical theory of the electron demonstrating the effect of a strong magnetic field on wavelength of the light produced by an atom (this was before the discovery of the electron). Lorentz's name is commemorated in the *FitzGerald–Lorentz contraction*, which is a reduction in the length of an object at relativistic speeds, and formed the basis of EINSTEIN's special theory of relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lovell, Alfred Charles Bernard [Sir Bernard] (1913–)

Physicist and astronomer, born in Oldland Common, Gloucestershire, England. Worked on cosmic rays at Manchester, researched radar in the Second World War, and afterwards with J S HEY procured an ex-army mobile radar unit used to detect V-2 rockets and attempted to detect cosmic ray showers with it. Interference from the electric trams at Manchester displaced the work to the university's botanical research station at Jodrell Bank (now the Nuffield Radio Astronomy Laboratories), where he found that echoes came from the ionized trails of meteors, particularly the Giacobinid display of October 1946. With JA Clegg, he laboriously built a 218 foot diameter wire-bowl parabolic transit telescope, used by Hanbury Brown and Hazard to detect M31. This showed the advantage of a large fully steerable telescope, and Lovell became the driving force behind the construction of the 250 foot diameter Mark 1 Jodrell Bank Telescope, completed by engineer HC Husband in 1957. Subject to painful allegations about its cost-overruns during its construction, the Mark 1 telescope was thrust into public view and rescued from ignominy that year by receiving signals from Sputnik, and American space vehicles. Lovell was knighted in 1961. He used the telescope to track and monitor flaring variable radio stars. Lovell has written many popular books on astronomy, radio astronomy in particular.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Low Surface Brightness Galaxies

Low surface brightness (LSB) galaxies are galaxies that emit much less light per area than normal galaxies. Because of their lower contrast with the night sky they are hard to find, and hence their contribution to the general galaxy population has long been underestimated.

Galaxies with low surface brightnesses are hard to find because the night sky is never black. Scattered sunlight coming from dust inside and outside our atmosphere, air-glow, street lights, faint Galactic and extra-Galactic sources all illuminate the sky. This foreground and background glare acts as a filter that only lets us see a part of the total galaxy population.

Compare this with the example of a person sitting inside a brightly lit room at night. This person will have trouble seeing what is happening outside because of the bright reflections of the room in the windows. Only the brightest objects (e.g. headlights of cars) will be easily visible. Headlights of bicycles, or pedestrians will be very hard to make out, even though they may dominate the traffic outside. The bright reflections thus filter out large parts of the total population.

In more practical terms, selection effects caused by the brightness of the night sky have made astronomers only give attention to those galaxies that were most easily observed. The underlying reason for these selection effects is that galaxy catalogs are usually created using certain isophotal diameter and/or magnitude limits (with an additional implicit surface brightness limit). The brightness of the night sky ensures that galaxies of a certain fixed luminosity look largest, and are thus most easily seen, if their central surface brightness has some intermediate 'optimum' value. If they have higher surface brightnesses they look too compact and resemble stars; with lower surface brightnesses they are difficult to distinguish from the night sky.

Diameter- or magnitude-limited catalogs thus mainly contain galaxies with a narrow range in central surface brightnesses. They will be biased and incomplete for galaxies with central surface brightnesses other than the 'optimum' value.

In the following, galaxies that are not affected by selection effects and that have an optimum value for detection are called 'high surface brightness' (HSB) galaxies. For all practical purposes these can be equated to the well-known SPIRAL GALAXIES that define the Hubble sequence. The dimmer galaxies that are affected by selection effects are called LSB galaxies. Examples are shown in figure 1.

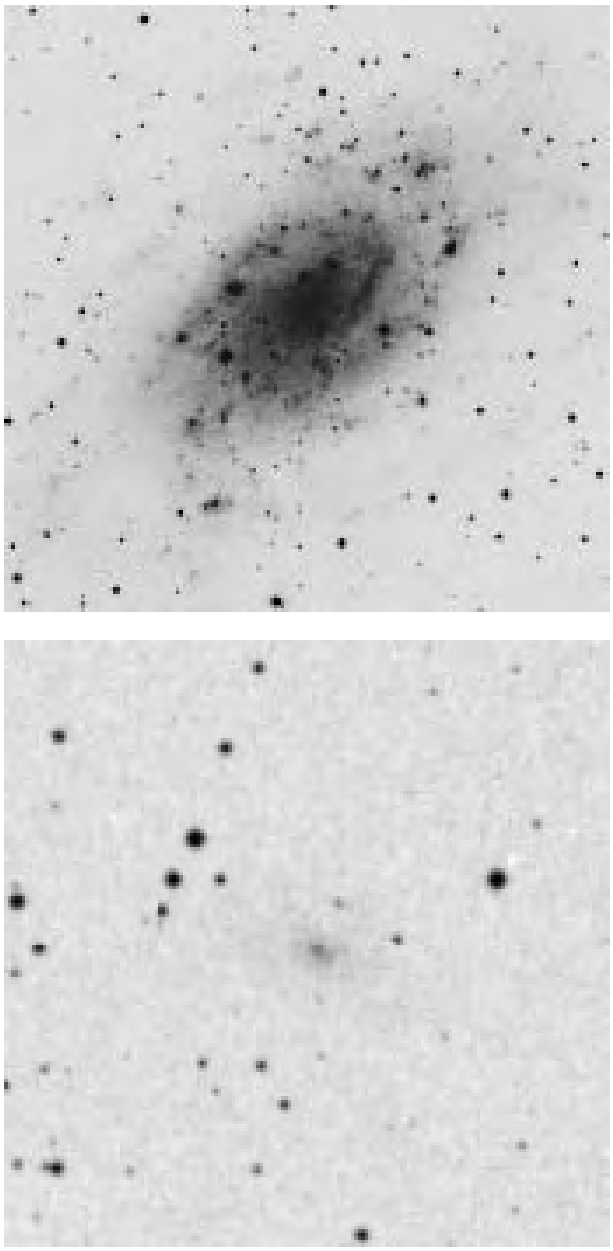


Figure 1. An HSB galaxy (NGC 2403, left panel) and an LSB galaxy (UGC 128, right panel) with identical luminosities, shown at the same physical scale, and observed with the same instrument. As the light of the LSB galaxy is much more spread out it is harder to distinguish from the glare of the night sky.

Selection effects

The best way to get a feeling for the severity of selection effects is by discussing a practical example. We will discuss five different types of model galaxies, which approximately span the currently known range in structural properties of DISK GALAXIES. All are assumed to be pure exponential disks, with central surface brightnesses μ_0 and scale lengths h as given in table 1. Four of these galaxies (A, B, C and D) have identical luminosities $M_B = -21$. The fifth one, E, is equal in size to galaxy B, but 2.5 mag or a factor of 10 fainter. It is furthermore assumed that all five types are spread through space homogeneously and in equal numbers. In each volume element there are therefore an equal number of different types of galaxies.

We will now 'observe' these galaxies and create a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Contribution of various galaxy types in diameter-limited catalog.

Type[a]	M_B[b]	h[c]	μ_0[d]	$D_{\max}$[e]	Fraction	
A	−21.0	0.5	16.0	60	0.09	Compact HSB
B	−21.0	5.0	21.0	125	0.72	Normal HSB
C	−21.0	25.0	24.5	76	0.18	Big LSB
D	−21.0	50.0	26.0	—	0.00	Giant LSB
E	−18.5	5.0	23.5	30	0.01	Normal LSB

[a] Type of galaxy. [b] Absolute magnitude. [c] Scale length in kpc. [d] Central surface brightness in mag arcsec^{-2}. [e] Maximum distance in Mpc where galaxy is included in catalog. [f] Fraction of cataloged galaxies that will be of this type.

diameter-limited catalog. The condition for inclusion in the catalog will be that the diameter of the 25 mag arcsec^{-2} isophote D_{25} must be larger than 1 arcmin. It is obvious that galaxies at large distances will look smaller than those nearby. Therefore, as we examine each type of galaxy and move it to increasingly larger distances, its D_{25} will at some point become less than 1 arcmin, and consequently it will no longer be included in the catalog. The distance where galaxies of a certain type drop out of the catalog is given in table 1: galaxy A will drop out of the sample at 60 Mpc; galaxy B will be in the sample out to distances of 125 Mpc; galaxy C out to 76 Mpc. Galaxy D has no D_{25} diameter and will never be taken up in the catalog. Finally, galaxy E will be seen only out to 30 Mpc.

Galaxies of type B correspond to normal HSB galaxies, while type E represents typical LSB galaxies. Type C is a big LSB galaxy. Types A and D are in practice rare. Cataloged HSB galaxies (such as B) can be detected in a volume 72 times larger than the volume in which LSB galaxies (such as E) can be detected. Assuming that all types of galaxies are spread equally through space, the corresponding volumes cataloged galaxies occupy show that the catalog will be severely dominated by HSB galaxies of type B and will lead to the conclusion that almost three-quarters of the total galaxy population consists of type B. Only an apparently insignificant 1% of the galaxies in the catalog are of type E. Furthermore, we would not even know that giant galaxies, such as type D, existed. On the basis of these results we would conclude that three-quarters of the observed galaxy population had a constant central surface brightness.

To correct for these effects we must therefore realize that every LSB galaxy that *is* cataloged represents a large number of uncataloged LSB galaxies. This means that we must assign to every galaxy a weight inversely proportional to the volume in which it is included in the catalog. Hence, if we try to reconstruct the true galaxy population from this diameter-limited catalog, every LSB galaxy of type E must receive a weight 72 times larger than a galaxy of type B. Note though that we still would not be able to say anything about type D. These volume-correction factors are crucial for our understanding of LSB galaxies.

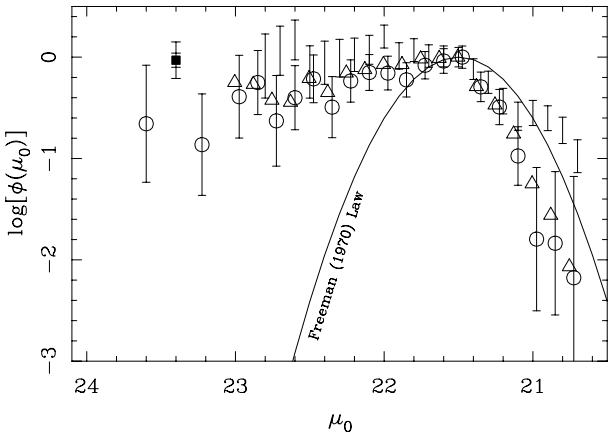


Figure 2. The number density of galaxies plotted versus central surface brightness. There is a large excess of galaxies at the LSB side of Freeman's value.

History

Selection effects and their possible consequences did not receive much attention until 3 decades ago. Before that only incidental references appeared in the literature. HUBBLE noted in 1936 that intrinsically faint galaxies will be detectable to much smaller distances than intrinsically bright galaxies and are consequently distributed through a much smaller volume of space. Thus 'intrinsically bright nebulæ greatly outnumber the faint ones among nebulæ with a given *apparent* luminosity'. In 1938 SHAPLEY announced the discovery of two intrinsically faint LSB galaxies, and in 1965 Arp showed that most galaxies then investigated occupied only a very narrow strip in a luminosity–diameter diagram; implying that they have almost identical surface brightnesses.

This last conclusion was explicitly derived by Freeman in 1970 who noted that 28 galaxies out of his sample of 36 had disks with central surface brightnesses in the range $\mu_0^B = 21.65 \pm 0.3$ mag arcsec^{-2}. Taken at face-value, this result had important implications for theories of GALAXY STRUCTURE and EVOLUTION. Freeman's law requires that either all galaxies have identical mass surface densities coupled with just a small spread in their mass-to-light ratios or that STAR FORMATION history, age, angular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

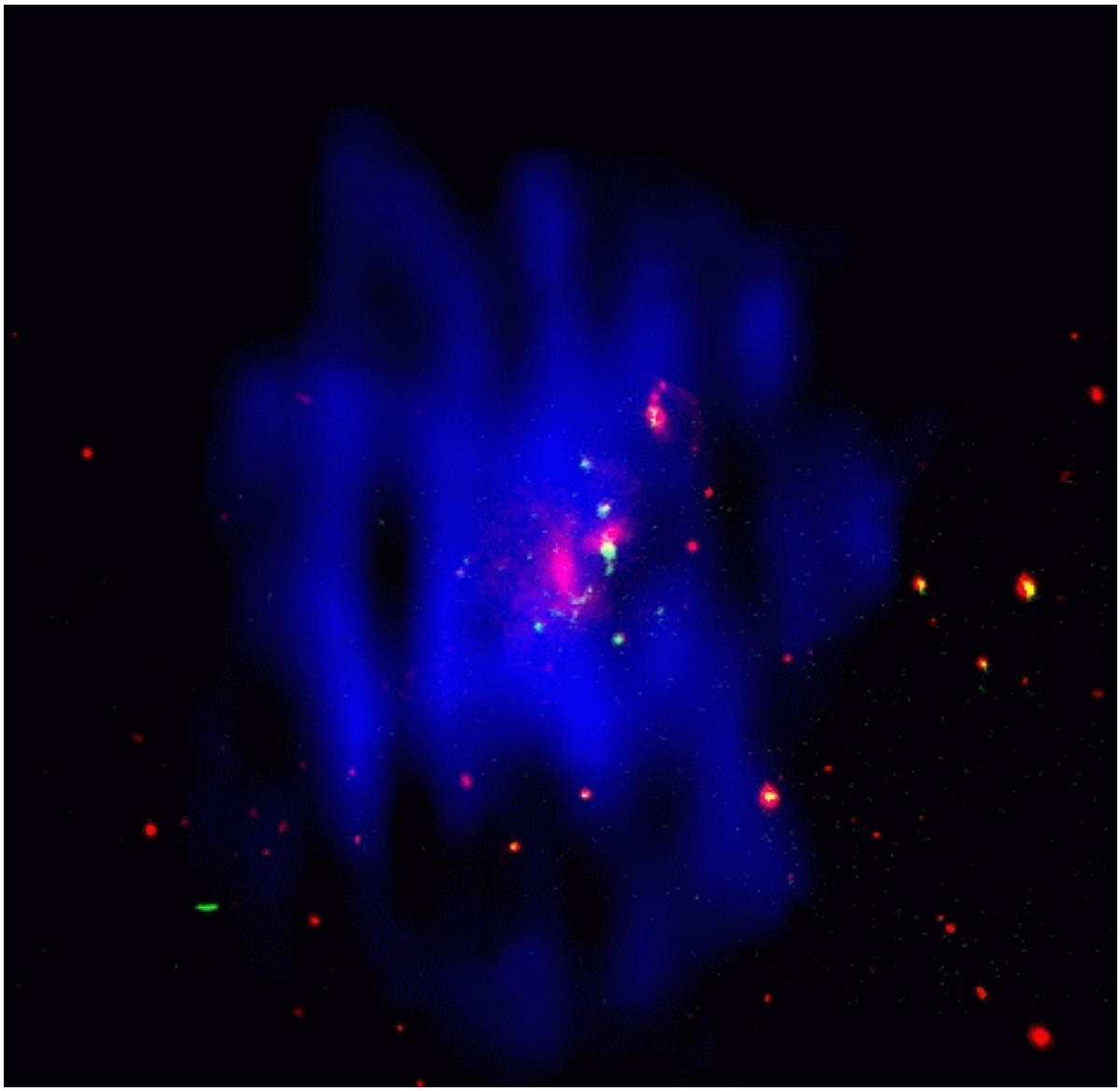


Figure 3. LSB galaxy F563-1. This false-color image is a combination of optical, 21 cm radio and 6563 Å Hα spectral-line observations. The red false colors show the stars in LSB galaxy F563-1 observed though a red filter. Blue false colors show the extent of the large and diffuse disk of neutral hydrogen surrounding the optical galaxy. The green–yellow dots in the center are observed in the light of the Hα spectral line at 6563 Å, showing the few star formation regions in this galaxy. The red and yellow dots surrounding the galaxy are foreground stars. The gas and stars in this galaxy represent a mass of a few billion solar masses, yet this is only a small fraction of the mass of the huge dark matter halo that surrounds this galaxy. This galaxy is at a distance of 45 Mpc or 147 million light-years. The diameter of the neutral hydrogen disk is 43 kpc or some 140 thousand light-years. **This figure is reproduced as Color Plate 12.**

momentum and mass all conspire to produce a constant central surface brightness

Disney in 1976 showed that selection effects could cause Freeman's law and suggested that there might be many galaxies of both high and low surface brightness ('crouching giants') hidden below the night sky. The discovery in 1987 of the giant LSB galaxy Malin-1 put LSB galaxies in the limelight. This extremely large and massive LSB galaxy was discovered accidentally during a survey for small LSB galaxies in the VIRGO CLUSTER. Malin-1 has a central disk surface brightness 4 mag arcsec^{-2} fainter than Freeman's value, yet it has a scale length of tens of kpc, a luminosity $M_B \simeq -22$ and an H I mass of slightly less than $10^{11} M_\odot$. This did show that LSB galaxies were not necessarily small and faint dwarf galaxies.

Careful searches on photographic survey plates then started turning up LSB galaxies in large numbers. These galaxies were mostly of late type and rich in hydrogen.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

After correction for various selection effects one of the conclusions was that the space density of LSB galaxies at $\mu_0 = 23$ is approximately 10^5 times higher than predicted by Freeman's law (figure 2).

The currently known LSB galaxies ($25 < \mu_0(B) < 23$) may make up 30–50% of the total galaxy population.

Properties

The properties of LSB galaxies can be summarized as follows.

They are more gas rich than HSB galaxies. One of the probable reasons they have lower surface brightnesses is because their gas surface densities are well below the critical threshold density for star formation. LSB galaxies thus have very low star formation rates. Although following the large-scale structure of galaxies, they are locally more isolated than other galaxies.

Investigations of the colors of LSB galaxies showed that they were bluer than normal late-type galaxies. Combined with the low metallicities found in H II REGIONS this implies a sporadic star formation rate, ruling out the notion of LSB galaxies as 'faded disks'. LSB galaxies are not normal galaxies which have stopped forming stars and are now fading away. Instead, they are likely to be slowly evolving. The contrast of the few young stars with the old background population, or, in other words, the rate of current star formation rate and average past star formation rate are much higher than in normal HSB late-type galaxies. A small amount of recent star formation (as observed) is already enough to make the colors significantly bluer.

Synthesis observation in the neutral hydrogen 21 cm line (see figure 3) confirmed this unevolved picture: LSB galaxies have extended gas disks with low gas surface densities, and high $M_{\rm HI}/L$ ratios. Despite the low surface densities the gas component is usually dynamically more important than the stellar component.

LSB galaxies are on the same TULLY–FISHER RELATION as HSB galaxies. This requires a subtle interplay between the central luminosity surface density Σ_0 and the total mass-to-light ratio M/L. As the Tully–Fisher relation can be written as $L[\Sigma_0(M/L)] \propto V^4$, the product $\Sigma_0(M/L)$ must be constant for all galaxies on the Tully–Fisher relation. LSB galaxies must have higher mass-to-light ratios than normal galaxies.

The rotation curves of LSB galaxies showed that this was indeed the case. LSB galaxies are DARK MATTER dominated almost all the way into their centers. The rotation curves rise less steeply than those of HSB galaxies. The halos of LSB galaxies are likely to be of lower density than those of HSB galaxies. The large dark matter dominance makes the stellar disks of LSB galaxies extremely stable, making it possible for the disk to exist at such low surface densities. The low metallicities make cooling difficult, so that a molecular component, where stars are formed, is difficult to realize. LSB galaxies can be said to be trapped in their current evolutionary state. Even external influences, like interactions, might not be enough to significantly speed up the evolution of LSB galaxies.

The above results apply to blue LSB galaxies that have been discovered in photographic plates. Recent CCD surveys (see also DETECTION OF FAINT OBJECTS) have started turning up a class of red LSB galaxies, showing that even the deep photographic plates suffered from selection effects. However, not enough is known about these red LSB galaxies to be able to say whether they are a cosmologically significant component of faded galaxies.

Importance

The single most important reason for studying LSB galaxies is that no representative sample of nearby galaxies has yet been compiled, cataloged and investigated. If a large population of hitherto unseen objects exists, the true extent of galaxy properties may be much larger than assumed.

Furthermore, investigations of individual LSB galaxies show that they form an alternative track of galaxy evolution, free from the instabilities and interactions that have shaped the Hubble sequence. They give us the opportunity to study unevolved galaxies in great detail.

Bibliography

Two good reviews on which this article has largely been based and which contain all relevant references are

Bothun G D, Impey C and McGaugh S S 1997 Low-surface-brightness galaxies: hidden galaxies revealed *Publ. Astron. Soc. Pac.* **109** 745–58

Impey C and Bothun G D 1997 Low surface brightness galaxies *Ann. Rev. Astron. Astrophys.* **35** 267–307

Erwin de Blok

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lowell Observatory

Lowell Observatory, founded in 1894 by Percival Lowell, is one of the largest independent, privately managed observatories in the world. Endowed by its founder, Lowell Observatory continues today as a private research institution dedicated to the study of astronomy. The large redshifts of galaxies were discovered by Lowell astronomer Vesto M Slipher. In 1930, Clyde Tombaugh discovered Pluto at Lowell Observatory.

Today, Lowell Observatory has a staff of 50, including 20 astronomers. The Observatory operates eight telescopes: four at its original campus on Mars Hill and four at a dark-sky site on Anderson Mesa, southeast of Flagstaff, Arizona. Lowell is a partner with the US NAVAL OBSERVATORY and the Naval Research Laboratory in the Navy Prototype Optical Interferometer (NPOI) at Anderson Mesa and with Boston University in the shared use and development of Lowell's 72 in (1.8 m) Perkins Telescope.

The Observatory is engaged in research spanning many areas in modern astronomy and astrophysics, while maintaining its traditional emphasis on study of the solar system. In addition to funds from the original endowment, a significant component of the Observatory's support is from federal grants and contracts.

For further information see http://www.lowell.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lowell, Percival (1855–1916)

Businessman, traveler, diplomat and, eventually, astronomer, born in Boston, MA, into a distinguished family (younger brother a president of Harvard University, sister a writer). His boyhood interest in astronomy revived at the age of 37, when after correspondence with WILLIAM H PICKERING they set up an expedition to observe the 1894 Mars opposition from the clear skies of Arizona. They founded the *Lowell Observatory* at Flagstaff, seen as a 'speculative, highly sensational and idiosyncratic project' to study the solar system as 'an investigation into the condition of life on other worlds, including . . . their habitability by beings like or unlike man'. Lowell had, unobjectively, already made up his mind about the implications of SCHIAPARELLI's identification of a network of canali on Mars: 'The most self-evident explanation from the markings themselves is probably the true one; namely, that in them we are looking upon the result of the work of some sort of intelligent beings'. Lowell studied Mars for 15 years, noting the spring-time dark band around the south pole, described as a 'ring of antarctic ocean', which fed the green 'vegetation' of the dark areas. He drew Mars in exquisite and imaginary detail, covering its map with a network of hundreds of straight 'canals' intersecting in 'oases'. According to JAMES KEELER, writing to GEORGE HALE (co-editors of the *Astrophysical Journal*) Lowell's papers were written in a 'dogmatic and amateurish style', and 'he draws no line between what he sees and what he infers', an opinion which led to the *Astrophysical Journal* refusing to publish any of Lowell's submissions. Lowell used the last eight years of his life to search for a trans-Neptunian planet, having analysed the orbit of Uranus and found discrepancies even after allowance for Neptune. He was unable to find the planet, but his search was continued after his death by CLYDE TOMBAUGH at Flagstaff, with Pluto discovered in 1930. The first two letters of Pluto, and the symbol for the planet, are Percival Lowell's initials.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lucretius (95–55 BC)

Roman poet and philosopher. His major work is the atomist poem *De Rerum Natura* (*On the Nature of Things*), popularizing the work of DEMOCRITUS and EPICURUS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Luminosity Function of Galaxies

The definition of the (optical) luminosity function (LF) of galaxies follows those of other astronomical objects such as stars or globular clusters. It is the number of galaxies that exist in a given volume of space having a given luminosity. Thereby, 'luminosity' generally means a total magnitude in any photometric passband U, B, V, R, I, etc. However, because of the spectral sensitivity of photographic plates which were classically and still are used for LF studies most of our knowledge is based on B- and R-band photometry. More recent work utilizes modern wide-field CCD cameras and infrared detectors to explore the LF of galaxies at other wavelengths.

Galaxies span a large range in total visual luminosity. At the extreme ends we find huge stellar systems with $L_B \sim 5 \times 10^{10} L_\odot$ ($M_B \sim -22$) such as the giant elliptical galaxy Messier 87 in the nearby VIRGO CLUSTER and the 5×10^5 times fainter dwarf elliptical galaxies Draco and Ursa Minor in the LOCAL GROUP with integrated blue light emission $M_B \sim -7.5$ comparable with a single massive main-sequence star.

The LF of galaxies is an important observational ingredient for cosmology as well as for galaxy formation and evolution. It holds fundamental information about (i) the power spectrum of the primordial density fluctuations, (ii) the physical processes that convert mass into light, e.g. gravitational collapse, cooling and star formation, and (iii) the mechanisms that destroy–generate galaxies or change their morphology such as tidal interaction, merging and ram pressure stripping. As such the LF is essential to interpret the huge apparent population of faint blue galaxies discovered at intermediate redshift ($z \approx 0.5$), to estimate the luminosity and baryonic densities of the universe and to test models of galaxy formation and evolution, e.g. the popular concept of 'biased galaxy formation'. Often the quality of a theory is assessed on how well the shape of the observed LF can be explained.

In the following we give a short historical overview about the discovery of the LF of galaxies and provide details of various observing techniques. We further explain the main features of a typical LF and discuss why nature makes LFs drawn from different parts of the universe not always look the same. We conclude with some remarks on crucial issues that are important for a deeper understanding of the LF of galaxies in the future.

A historical retrospect

The first result on the LF was published by Edwin Hubble in the mid-1930s initiated by the evidence that the velocity–apparent magnitude relation for the Shapley–Ames galaxies studied by Milton Humason exhibit only a small scatter. Interpreted in the context of a linear expansion of the universe, i.e. velocity proportional to distances (HUBBLE LAW), these observations meant that the width of the absolute magnitude distribution, i.e the LF, of these galaxies was also small. Hubble's results, although they lacked precision, suggested a Gaussian profile centered at $M_B \sim -18$ and $\sigma \sim 0.9$ mag. There seemed to be upper and lower limits for galaxy luminosities. In 1942, the validity of a bell-shaped form for the LF was called in question by Fritz Zwicky who predicted from thermodynamical considerations (principle of conservation of energy, virial theorem and Boltzmann principle) the existence of a large number of low-mass, faint dwarf galaxies in the universe and consequently a steeply rising LF towards fainter luminosities. Even though most of these dwarfs remained undetected at that time, Zwicky's idea received some support from discoveries of very faint dwarf galaxies in the Local Group.

These apparently controversial historical results exemplify the problems LF research encounters: selection bias and incompleteness. Hubble's primary goal was to measure redshifts and hence he had to concentrate on high surface brightness galaxies: ELLIPTICALS (E), LENTICULARS (S0) and SPIRALS (Sp). In this way the numerically dominant galaxy types in the universe, which we know today are the hardly visible low surface brightness DWARF IRREGULAR (dIrr) and DWARF ELLIPTICAL GALAXIES (dE), did not contribute to his consideration. Zwicky's theoretical result, on the other hand, found assistance from a spatially well-defined galaxy aggregate where the number of known dwarf galaxies was small but steadily increasing with time. Hence the Local Group catalogue reflected the actual situation much better.

Cluster versus field

When searching for galaxies across a photographic plate the eye is quickly caught by galaxy clusters which exhibit a high density contrast against the general field. Because of their nature clusters are the best places to perform LF studies. They inherit many hundreds of galaxies of all different morphological types which are all located in a small volume of space. Hence the cluster distance D, averaged over a fair number of individual known galaxy distances, can be used to convert apparent magnitudes m of all cluster members into luminosities ($M = m - 5 \log[D\,(\mathrm{Mpc})] - 25$). The identification of cluster members against background and foreground objects requires the cluster to be well isolated in real and velocity space to keep contamination by non-cluster galaxies minimal. The bright E, S0 and Sp galaxies are identified based on velocity data and/or morphological criteria. At fainter luminosities most galaxies are dwarfs for which velocities are not easily accessible. The numbers of these galaxies are estimated statistically by comparing galaxy counts in a cluster with an adjacent reference field. The excess in the measured cluster LF is assigned to the cluster population. However, the heterogeneous distribution of field galaxies makes this procedure difficult and a matter of dispute. Alternatively, the unique morphological appearance of dwarf galaxies, in particular dE galaxies, provides an excellent tool to decide which object belongs to a cluster. The drawback with this method

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is the requirement for galaxy images of high angular resolution, which technically restricts the application to the closest clusters.

The field LF is much more difficult to measure. About 80% of all galaxies reside outside of rich clusters where galaxy densities are down by a factor of 100–1000. Large volumes must be surveyed to obtain representative samples of objects. Panoramic and pencil-beam surveys are carried out for this purpose using long integration times to reach low surface brightness galaxies and bright stellar systems at larger distances. Redshifts (velocities) are the backbone for field LF studies as they provide the important distance information. Employing the velocity as a distance indicator becomes possible because gravitationally induced peculiar velocities are mostly absent in the field and thus observed velocities are closely linked to the Hubble flow (the general expansion of the universe), although corrections for large-scale bulk motions are still required. Field surveys have generally an apparent magnitude limit owing to the sensitivity limits of the spectrographs. This means that intrinsically bright galaxies always predominate in galaxy samples as they are visible over large distances. To fainter luminosities the surveyed volumes shrink substantially and the galaxy numbers drop significantly. For instance the detection volume for the Small Magellanic Cloud is 4000 times smaller than for M87 just because of the difference in their total magnitudes. At that level inhomogeneities in the three-dimensional galaxy distribution do not always average out and can cause biases in the data (see DISTRIBUTION OF GALAXIES, CLUSTERS AND SUPERCLUSTERS). Various sophisticated methods exist to estimate the LF from a magnitude-limited galaxy sample (the $1/V_{\max}$, C and STY methods to name the most important ones), however the substantial corrections at the faint end of the LF make the shape in any case highly uncertain.

Overall and type-specific LFs

Since 1970 more than 100 cluster and field studies contributed to reveal the overall signature of the LF (see figure 1): after an exponential rise at $M_B \sim -22$ ($H_0 = 50$ km s^{-1} hereafter) the LF turns off at a characteristic luminosity L^* (total magnitude M^*), which is sometimes referred to as the 'knee', and follows a power law at fainter luminosities. The most popular parameterization to fit the data was proposed by Schechter in 1976:

$$\phi(L)\,\mathrm{d}L = \phi^*(L/L^*)^{\alpha}\,\exp(-L/L^*)\,\mathrm{d}(L/L^*) \qquad (1)$$

with the normalization parameter ϕ^*, or equivalently in magnitudes

$$\phi(M)\,\mathrm{d}M = (0.4\ln 10)\phi^* 10^{0.4(M^*-M)(1+\alpha)} \times \exp[-10^{0.4(M^*-M)}]\,\mathrm{d}M.$$

Thereby the faint-end slope is given by $-(1+\alpha)$, i.e. decreasing for $\alpha > -1$, flat for $\alpha = -1$ and increasing for $\alpha < -1$. The integral of equation (1) yields the total number of galaxies in a given volume in space whose luminosities are brighter than L: $N(>L) = \int_L^\infty \phi(L')\,\mathrm{d}L' = \phi^*\gamma(1+\alpha, L/L^*)$. Their total luminosity is $L_{tot}(>L) = \int_L^\infty L'\phi(L')\,\mathrm{d}L' = \phi^* L^*\gamma(2+\alpha, L/L^*)$, where γ is the incomplete gamma function.

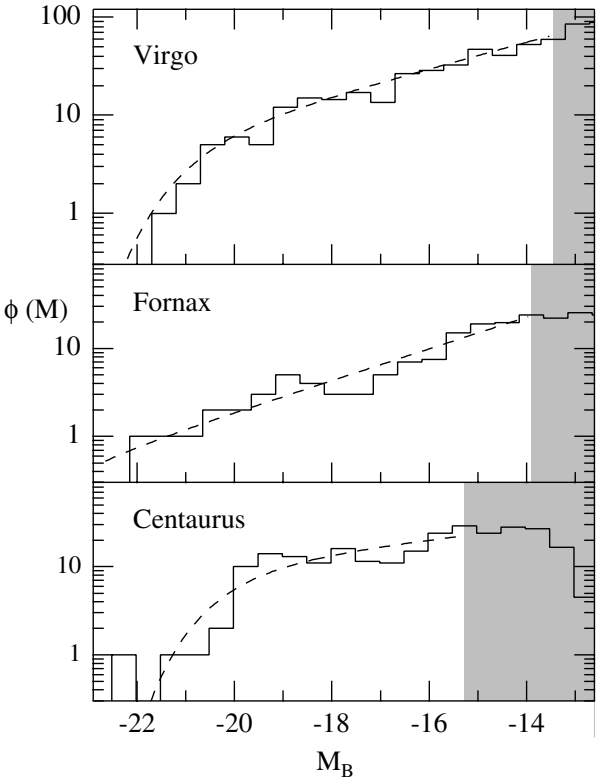


Figure 1. Data from the three clusters Virgo, Fornax and Centaurus (histograms) illustrate the diversity in observed cluster LFs. By chance Virgo is exceptionally well fitted by the Schechter form (dashed curve), whereas the local features in the LFs of Fornax and Centaurus remain unresolved. The shaded areas indicate the sample incompleteness intervals.

The Schechter function describes reasonably well the observed cluster LFs (see figure 1). Mismatches occur at the bright end owing to the presence of overluminous cluster members or if the LF is rich in local features. Nevertheless, the apparent good agreement led to the suggestion that the values $M_B^* \sim -21.0$ (about the total magnitude of the Andromeda galaxy) and $\alpha \sim -1.25$ might be of physical significance. However, this idea was not further supported by subsequent deeper surveys of clusters of various densities, richnesses, morphological classes or evolutionary stages which found a wide range of faint-end slopes ($-2.2 < \alpha < -0.9$). It was also noticed that the Schechter parameters are strongly correlated and vary significantly with the depth of a survey. The latter means that the same cluster population analyzed to different magnitude limits will not necessarily give the same Schechter parameters. Robust solutions cannot be expected above $M_B \sim -14$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In contrast to clusters where the signature of the environment is still imprinted on the data, the LF for field populations (including groups) gave quite a consistent picture with $M_B^* \sim -21.0$ and a flat LF down to $M_B \sim -17$. However, it is not very surprising after the previous remark that this result has been revised in recent years by deeper redshift surveys ($M_{B,lim} \sim -15$) reporting on prominent deviations from a flat LF ($-1.3 < \alpha < -0.7$). The disturbing point now is the observed scatter in α. Wouldn't we expect a well defined, unique asymptotic behavior of the field LF if averaged over a cosmological meaningful volume? This discrepancy has to be resolved in order to understand possible selection biases in the data, limitations of incompleteness corrections and eventually real differences in the dwarf-to-giant ratio for a given survey area.

After learning about all the differences between the LFs of high- and low-density regions and even among various clusters hope is dwindling that the observed spectrum of LFs can be reduced to a common denominator. However, on looking more carefully one immediate explanation comes into mind: the galaxy type mixture which intimately correlates with the environmental density. This phenomenon encompasses (1) the well-known GALAXY MORPHOLOGY–DENSITY RELATION: early-type galaxies (E, S0, dE) preponderate in the high-density regions whereas late-type galaxies (Sp, dIrr) dominate the intercluster medium, i.e. groups and field, and (2) the observation that the dwarf-to-giant ratio is correlated with the local density. These trends automatically open the question on the LFs of individual morphological galaxy types.

Our knowledge on the type-specific LFs is still in its infancy. Work started in the 1980s with first results on the Coma cluster, followed by extensive studies of the three nearby clusters Virgo, Fornax and Centaurus. Based on detailed morphological information the above-mentioned five main galaxy families could be studied individually. The results suggest that the type-specific LFs are very different but that each, except the dIrr LF, show little or no variation from one cluster to the other. The classical Hubble types E, S0 and Sp exhibit bell-shaped LFs; they exist only above a certain threshold luminosity. Dwarf galaxies are less luminous than the giants and govern completely the overall LF for galaxies fainter than $M_B \sim -16$. The large populations of dE galaxies in clusters show a steep Schechter-like LF with a mean α of -1.4. The most puzzling LF is that of the dIrrs with α values between -1.3 and -0.3. The large scatter may have its origin in the definition of the dIrr class that encompasses hardly visible low surface brightness galaxies (Im) with no or very little recent star formation activity and bright compact starburst galaxies (BCD). Difficulties in detecting the faintest members of the family can introduce selection biases to the data. However, density-dependent mechanisms that trigger star formation would also affect the LF of these gas-rich dwarfs.

The morphological resolution of field galaxies is not as accurate as for nearby cluster galaxies owing to the larger distances involved. Spectra are used instead to subdivide galaxies according to their level of star formation, a quantity that is closely related to the basic branches of early-type (E, S0, dE) and late-type (Sp, dIrr) galaxies at the present epoch. Recent deep-redshift surveys of many thousand galaxies in the redshift range $0 < z < 1$ sampled the type-specific LFs in the field as faint as $M_B = -15$. The results are qualitatively consistent with Gaussian profiles for the E–S0, and Sp galaxies and with steep LFs for dEs and dIrrs. It is interesting to note that an apparent steep dIrr LF at that magnitude limit is not necessarily in contradiction with an asymptotic value $\alpha = -1$ as illustrated in figure 2 (see 'Extreme Field'). The large number of dIrrs scales a flat dIrr LF upwards leading to a steep slope at $M_B = -15$.

The evidence for invariant type-specific LFs offers for the first time a semi-empirical tool to explore the behavior of the overall LF as a function of the environment. For this purpose the five type-specific LFs as found in three clusters (table 1) are weighted with observed and, in the case of the 'Extreme Field', hypothetic type mixtures to generate synthetic overall LFs. Similarly to working with the same ingredients but using different recipes so that the result is either a cake or a custard, this approach produces the whole range of observed LFs (see figure 2) from rich, dense clusters like Coma and loosely concentrated systems like Virgo to groups and the field. The very different characteristics of LFs from clusters to the field are explained satisfactorily with this method, which clearly indicates the importance of the type mixture and type-specific LFs for the understanding of the overall LF.

Table 1. Analytical functions and parameters that represent good first-order approximations for observed type-specific LFs in galaxy clusters. The faint-end slope for dIrrs is assumed to be -1 owing to the lack of conclusive empirical results.

E	Gauss	$\mu_B = -18.3$	$\begin{cases}\sigma_{M<\mu_B} = 2.2\\ \sigma_{M>\mu_B} = 1.3\end{cases}$
S0	Gauss	$\mu_B = -18.9$	$\sigma = 1.1$
Sp	Gauss	$\mu_B = -18.3$	$\sigma = 1.4$
dIrr	Schechter	$M_B^* = -16.2$	$\alpha = -1$
dE	Schechter	$M_B^* = -17.8$	$\alpha = -1.4$

Summary and prospects

Generally speaking there is a good understanding of the optical LF of galaxies down to a luminosity $M_B \sim -16$ but unfortunately this is still far away from the complete picture. The most basic detail is that the LF is the sum over separate LFs for the individual galaxy types. There are two major components in the overall LF which have been effectively discovered by Hubble and Zwicky. The first component consists of the three families of high surface brightness galaxies E, S0 and Sp. The second component is made up by the dwarfs, low surface brightness galaxies which are intrinsically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

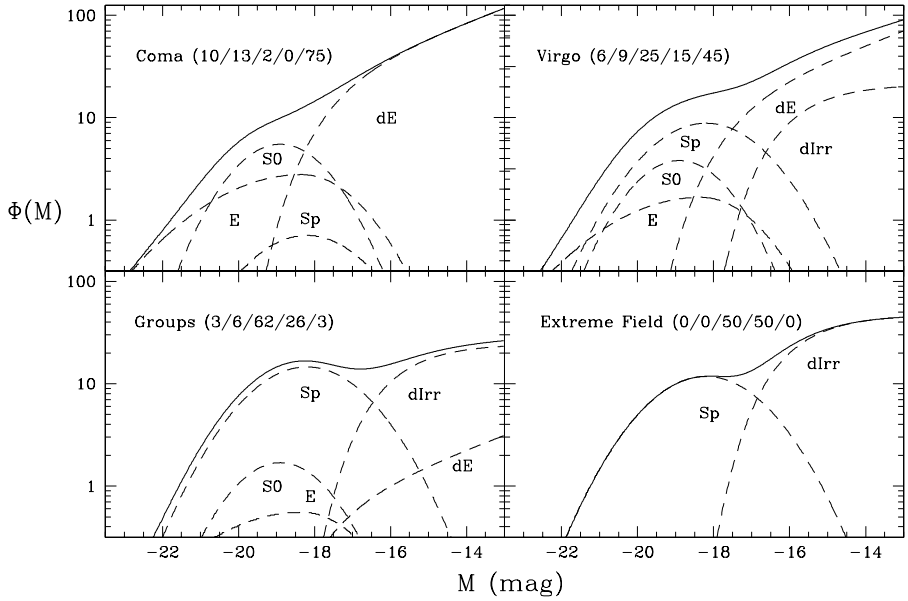


Figure 2. Synthetic overall LFs (solid curves) for four different environments with the galaxy type mixtures (E:S0:Sp:dIrr:dE) indicated as percentages. The contributions of the five main morphological types are the dashed curves.

fainter than the giants and fully control the asymptotic behaviour of the LF in all environments. This can be either a steep or moderate–flat power law depending on the local galaxy density. Due to the very different nature of the two components, many clusters (e.g. Fornax, Centaurus, Coma, Abell 963), some groups (e.g. Antlia) and field populations exhibit a prominent dip in their LFs at the transition luminosity $M_B \sim -17.5$.

Fundamentally we would like to fully understand the formation and evolution of galaxies. Results on the LF clearly indicate that this goal cannot be achieved by treating the phenomenon 'galaxy' as unity. Rather, the various morphological types have to be studied individually. From cluster work first evidence emerges that the LFs for E, S0 and Sp at the present epoch have well-defined Gaussian profiles which are unaffected by the environment. The same appears to be true for the LF of dE galaxies which follows a steep Schechter profile to all known luminosity limits. For the dIrr LF we are still lacking conclusive results as the situation is far more complex. There are some physical arguments against a single LF for dIrrs in clusters. Overall, more detailed LF studies in the local universe and at different redshifts, i.e. time epochs, are desperately needed, focusing on the separation of the morphological components in a consistent way. These results will reveal the significance of time and environmentally induced physical processes for the evolution of galaxies.

Another key issue is the variation of the morphological mixture with galaxy density and the dwarf-to-giant ratio in particular. Most current cosmological theories predict that, as the universe expands, galaxies clump together to form groups, which in turn merge together to form clusters. These 'bottom-up' scenarios have difficulties in explaining the high dwarf-to-giant ratio in clusters compared with the lower fraction in groups and the field. Where is this 'excess' of cluster dwarfs coming from? That question is fundamentally related to the space density of dIrrs which is not very well known as the faintest members of these dwarfs may remain undetected in the optical. However, this situation is going to change with systematic surveys of the local universe in neutral hydrogen (H I). dIrrs contain a substantial amount of H I gas and thus become 'visible' by their radio emission at 21 cm. Another advantage of H I surveys over optical work is that each galaxy is automatically tagged with its distance via the H I velocity. Preliminary results for nearby groups from ongoing H I surveys (e.g. HIPASS) are very promising. Many new group members have been discovered in H I but all have faint optical counterparts. Firstly this means that the average number of dIrrs and the dwarf-to-giant ratio in groups are higher than estimated to date. Secondly, the H I mass-to-light ratio for dIrrs is not increasing with fainter luminosities. The missed fraction of dIrrs per magnitude unit appears to be small which gives us greater confidence in the optical results.

Bibliography
Books:

Binney J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)
Silk J 1989 *The Big Bang* (New York: Freeman)

Journal articles:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Binggeli B, Sandage A and Tammann G A 1988 The luminosity function of galaxies *Ann. Rev. Astron. Astrophys.* **26** 509

Dressler A 1984 The evolution of galaxies in clusters *Ann. Rev. Astron. Astrophys.* **22** 185

Helmut Jerjen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Luminous Blue Variables

Luminous Blue Variable stars (LBVs) are among the most luminous hot stars in galaxies. They show very irregular and unpredictable brightness variations. The typical timescale of the variations is of the order of a few years and the visual brightness of the star can vary by about a magnitude. Some LBVs have suffered very large eruptions. LBVs represent a short critical phase in the evolution of very massive stars (see also STELLAR EVOLUTION).

LBVs are highly unstable, very luminous stars. They are recognized by their strange variability on many timescales. The most characteristic variations show brightenings or fadings by a factor of two or three over a period of years. Some LBVs are changing all the time; others go through phases of about constant brightness that may last ten years before they start changing again. Observations with satellites have shown that when these stars are faint in visual light, they are bright in UV radiation and vice versa. This shows that the variations are due to changes in the radius and surface temperature of the stars. During the visually faint phases, the stars are relatively hot and small. When they are visually bright their surface temperature drops but their radius increases by a factor between two and ten. The reason for these changes is still unknown.

Several LBVs have been observed to go through large eruptions, when they brighten by as much as a factor of 100. Such eruptions are rare and may occur only once every few thousand years. During an eruption the LBV ejects a large amount of mass, of the order of the total mass of the Sun. The most famous LBV is the star η Carinae (ETA CARINAE), which had an eruption in late 1830s, when it suddenly became one of the brightest stars in the Southern sky. The star is now surrounded by a magnificant bipolar NEBULA consisting of the gas ejected during the eruption.

LBVs are very rare: there are only five confirmed LBVs in our Galaxy and a few tens in other galaxies. This is because they represent a relatively short phase in the life of the most massive stars. Yet, this phase is critical for the evolution of these stars: the large mass loss of the LBVs prevents the massive stars from becoming red supergiants. After the LBV phase they have lost so much mass that they evolve directly into very hot helium-rich stars (the WOLF–RAYET STARS).

Two special LBVs

P Cygni

The star P Cygni was discovered in the constellation of Cygnus the swan in August 1600 by the Dutch cartographer Bleau, when it suddenly appeared at a place where no star had been visible before. The star remained bright for many years and then slowly faded untill it was invisible to the naked eye. In 1655 it brightened again for a few years but was invisible during the late 1600s. Now the star can be seen with the naked eye (at magnitude 4.9). From the distance of the star, 6000 light-years away, and from the brightness in visual and ultraviolet light, it is concluded that the star has a radius of 76 solar radii, a surface temperature of 19 000 K and a mass of about 30 times that of the Sun.

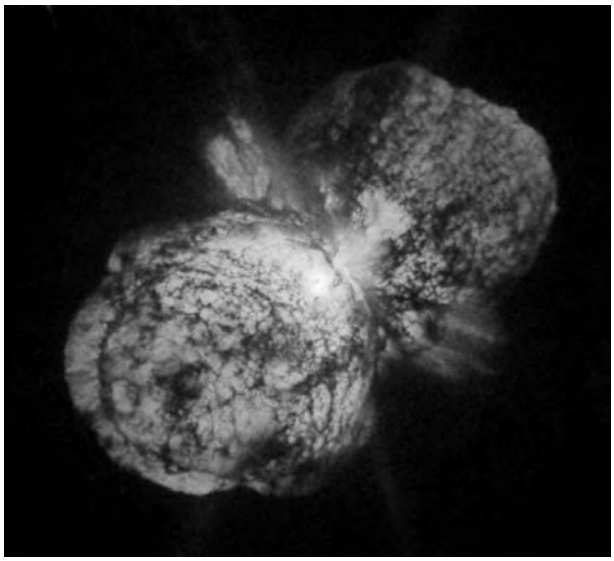


Figure 1. The nebula around the Luminous Blue Variable η Carinae is the result of the large eruption in the mid 1800s. This image was taken with the Hubble Space Telescope.

The historic light curve shows that since 1700 the star has been slowly brightening at a rate of 15% per century. This is interpreted as a very slow expansion of the star due to its evolution from blue to yellow supergiant. P Cygni is surrounded by a very faint nebula, which is the result of the ejection during the large eruption in 1600. The nebula was ejected wih a velocity of 140 km s^{-1}.

η Carinae

The most dramatic brightness variations of any LBV is shown by η Car. During the eruption in the period of 1837 to 1860 the star was the brightest star in the Southern sky, despite its large distance of about 8000 light-years. At maximum, the star reached a luminosity of 30 million times that of the Sun. It faded rapidly after the eruption, because it was obscured by the dust that had been formed in the ejected material. Since the mid-1900s the star has gradually brightened in visual light as the extinction by dust decreases.

η Car is surrounded by a spectacular nebula consisting of gas and dust ejected during the large eruption (figure 1). The nebula has a double-lobed structure, with an expansion velocity of about 30 km s^{-1} near the center to 1000 km $^{-1}$ at the outer edge. The total amount of mass in the nebula is about ten times the mass of the Sun. The typical bipolar structure is not explained: it could be due to the fact that η Car may have a companion star nearby, or due to rapid rotation of the star. Unfortunately the star itself is hidden in the central region of this nebula. The star can only be studied through its light reflected from the dust cloud surrounding it. This has shown that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

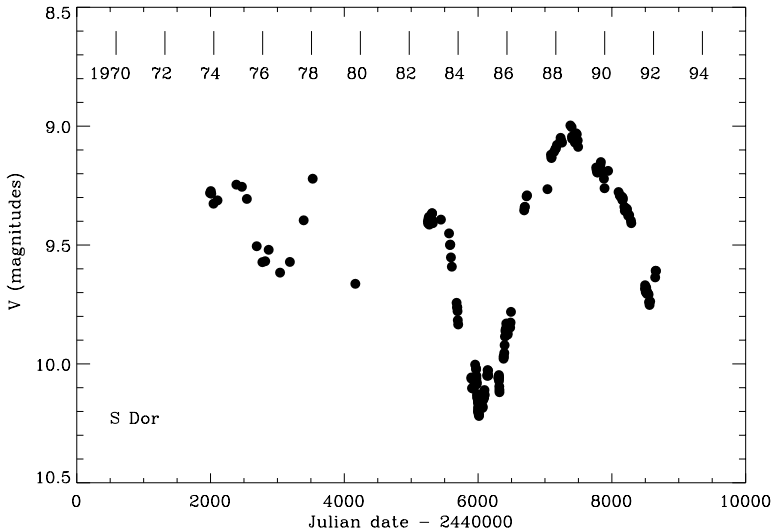


Figure 2. The light curve of the LBV S Doradus in visual light between 1975 and 1992. Bright is up, faint is down. An increase of one magnitude corresponds to a decrease in brightness by a factor 2.51.

central object varies in emission lines, especially those of hydrogen and neutral helium, with a period of 5.52 years. This suggests that the central object is in fact a BINARY SYSTEM consisting of two stars of about 80 solar masses orbiting each other at a distance of about 20 astronomical units. The binarity may have been responsible for shaping the bipolar nebula.

The light variations of LBVs

The light curves of LBVs are highly irregular and unpredictable. A typical light curve of the LBV star S Doradus is shown in figure 2.

Three types of light variations can de distinguished in the light curves of LBVs:

(a) Short time variations on a timescale of months with an amplitude of 0.1 to 0.3 magnitudes (the little wiggles near the deep point of the curve of figure 2).
(b) Typical LBV variations on a timescale of years with an amplitude of about 0.5 to 1 magnitude (the large dips and peaks in figure 2).
(c) Large eruptions that have been observed in P Cygni and η Carinae, but not in S Doradus. The presence of nebulae around LBVs indicates that most, if not all, LBVs go through large eruptions, and that these may occur every few thousand years.

The short time variations are caused by NON-RADIAL STELLAR PULSATIONS. Normal hot supergiants show the same type of pulsations (sometimes called α Cygni variations). The much larger typical LBV variations (sometimes called S Doradus variations) are due to some unknown instability inside the star that makes the radius grow and shrink in an irregular way by up to a factor of ten. During these variations the total luminosity of the star remains almost constant but the light is redistributed in the visual and the ultraviolet. This means that as the radius expands, the surface of the star gets cooler, and so the color of the star gets redder and the spectral type changes from early-B to A-type. These changes are indeed observed during the light variations. The origin of the large eruptions is unknown. It may be related to the fact that the stars are so luminous that the outward directed force due to radiation pressure is almost as large as the inward directed gravitational force. In that case the net attracting force inside the star can be so small that even minor disturbances can create a large instability.

Basic properties

The basic properties of well studied LBVs are listed in table 1. This list contains the five confirmed Galactic LBVs and five LBVs in the LARGE MAGELLANIC CLOUD.

Luminosities

LBVs are very luminous with a luminosity, L, between 260 000 and 3 400 000 times that of the Sun. The luminosity of the LBVs remains almost constant during their variations in visual brightness and in color.

Surface temperatures

The effective temperatures, $T_{\rm eff}$, of LBVs vary with the visual brightness. At visual maximum the stars have a temperature of about 8000–9000 K. At visual minimum the LBVs are much hotter, from 30 000 K for the most luminous ones to 10 000 K for the least luminous LBVs. The changes in temperature are correlated with the visual brightness variations. The larger the visual brightness variations, the larger the temperature differences. This is because the decrease in visual light is compensated by an increase in the ultraviolet light as the star becomes hotter. The stars η Carinae and P Cygni, which do not show typical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The basic properties of LBVs in the Galaxy and in the Large Magellanic Cloud.

Name	System	Distance (pc)	Luminosity ($L_\odot$)	$T_{\rm eff}$ (K)	Radius ($R_\odot$)	Nebula
η Carinae	Galaxy	2600	3 400 000	20 000–30 000	150–70	yes
AG Carinae	Galaxy	6100	1 200 000	9 000–30 000	450–40	yes
R 127	LMC	51000	1 200 000	8 500–30 000	500–40	yes
R 143	LMC	51000	790 000	8 500–20 000	400–70	no
P Cygni	Galaxy	1800	720 000	19 000	80	yes
S Doradus	LMC	51000	650 000	8 000–22 000	400–60	no
HR Carinae	Galaxy	5000	290 000	14 000	90	yes
HD 160529	Galaxy		290 000	8 000–11 000	280–150	no
R 110	LMC	51000	290 000	7 600–10 000	300–180	no
R 71	LMC	51000	260 000	9 000–14 000	200–90	yes

LBV variations with amplitudes of about a magnitude, also show little temperature variation, except during their historic large eruptions.

Radii

The radius of LBVs varies with the effective temperature. High temperatures correspond to small radii and vice versa. During the hot and visually faint phase the radius is typically between 40 and 100 solar radii (except for HD 160529 and R 110). During the cool and visually bright phase the radius is typically a few hundred solar radii. This is in agreement with the almost constant luminosity of the LBVs during their variations, which implies that the radius is approximately proportional to $T_{\rm eff}^{-2}$.

Masses

The masses of LBVs are not well known. The high luminosity of the stars indicates that their original mass is in the range from 40 solar masses for the least luminous LBVs, like HD 160529, R 110 and R 71, to 100 solar masses for the most luminous ones like AG Carinae and R 127. The star η Carinae is by far the most luminous and most massive one with an estimated initial mass of 150 solar masses. The actual (present day) mass of the LBVs is considerably smaller than their initial mass. The nebulae around many the LBVs show that they have lost a significant fraction of their initial mass already. Evolutionary calculations suggest that the masses of LBVs are between 40% and 75% of their original mass.

Spectra

The spectra of LBVs change with their photometric variations and their temperature variations. When the stars are visually faint the spectrum is typically that of an early-B supergiant. When the stars are visually bright their spectrum is that of an A-type supergiant. The studies of the spectral lines show that the atmospheres of LBVs are enriched in helium and nitrogen and depleted in carbon and oxygen, compared with normal stars.

The visual spectrum of LBVs is dominated by spectral lines with 'P Cygni profiles' consisting of an emission component and a blue-shifted absorption component. The wavelength shift of the absorption component is due to the Doppler effect and indicates an outflow velocity of several hundred km s^{-1}. The strength of the lines and the outflow velocity show that the LBVs are losing mass at a rate of about 2×10^{-5} to 6×10^{-5} solar masses per year. The mass loss rate of η Carinae is much higher and of the order of 10^{-3} solar masses per year.

Nebulae around LBVs

Almost all LBVs are surrounded by nebulae which are due to large eruptions in the past. The diameters of the nebulae are typically of the order of 0.2–2 parsec and the expansion velocity is between 25 and 100 km s^{-1}. The LBV η Carinae is an exception, as it shows different velocities from about 100 to 1000 km s^{-1} in different knots. The dynamical age $\tau_{\rm dyn}$ of LBV nebulae can be derived from their radius and outflow velocity: $\tau_{\rm dyn} = r_{\rm neb}/v$. The dynamical ages of the LBV nebulae are in the range 5000–50 000 years. The nebula of η Carinae has a dynamical age of 150 years and that of P Cygni has an age of 400 years, both in agreement with the observed historic large eruptions.

The evolutionary connection

The evolutionary stage of the LBVs can be derived from their location in the HERTZSPRUNG–RUSSELL DIAGRAM shown in figure 3. The stars make horizontal excursions in this diagram, due to the variations in radius and temperature at constant luminosity. The stars are close to the Humphreys–Davidson limit. This is the empirically determined luminosity upper limit in the Hertzsprung–Russell diagram. No stars brighter than this limit (except LBVs) have been found in the Milky Way or in other galaxies (see HIGH-LUMINOSITY STARS).

The location of the LBVs in the Hertzsprung–Russell diagram shows that the stars are in a post-main sequence evolutionary phase. The stars get their energy from helium fusion in the center. The increased helium and nitrogen abundance in their atmospheres is due to the fact that the original hydrogen-rich layers have been expelled, partly by mass loss when the star was still on the main sequence, and partly in the large eruptions. The helium and nitrogen enriched layers were formed by hydrogen fusion via the CNO cycle in the convective core of the star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

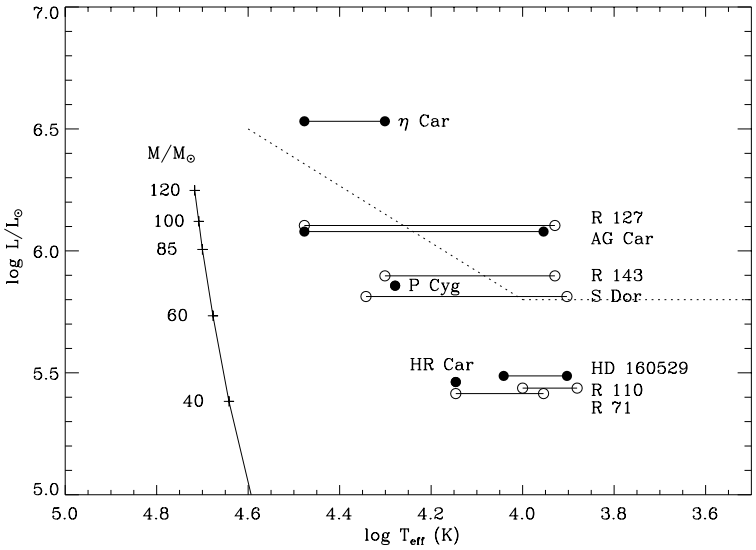


Figure 3. The location of the LBVs in the Hertzsprung–Russell diagram. Because the stars are variable they cover a range of effective temperatures. The full curve is the main sequence with the masses predicted by evolution theory indicated. The dashed line is the Humphreys–Davidson limit. Filled and open circles indicate LBVs in our Galaxy and in the Large Magellanic Cloud respectively.

The average duration of the LBV phase has been derived in a statistical way from a comparison between the numbers of LBVs and the numbers of other stars in the same range of initial masses. This shows that the LBV phase typically lasts a few times 10^4 years. With a typical mass loss rate of about 4×10^{-5} solar masses per year and the ejection of a few solar masses per large eruption every few thousand years, the total mass lost during the LBV phase can amount to a few tens of solar masses. The LBV phase ends when the star has ejected almost all of its hydrogen-rich material. The star then evolves into a hot helium-rich star, i.e. a Wolf–Rayet star.

Unsolved problems

Since the discovery of the large eruptions of P Cygni and η Carinae, their cause has been a great mystery. Several suggestions have been made:

(a) an instability due to a possibly fast rotating core,
(b) a geyser-type of instability due to the ionization of hydrogen in a critical layer,
(c) a 'strange-mode instability' due to the coupling of several pulsation modes, or
(d) large radiation pressure at a critical temperature zone in the star due to the iron opacity.

None of these theories can correctly predict the properties such as temperature, radius and luminosity of the stars for which the eruptions have been observed, nor the ejected amount of mass or the shapes of the nebulae.

The origin of the more common LBV variations with a size of a few magnitudes is also not understood. It is known that they are due to large variations in stellar radius by about a factor of ten or so. Originally it was assumed that the increase in radius is due to a sudden increase in the mass loss rate, which results in an opaque wind that makes the star appear to be larger. However, recent studies have shown that the winds of LBVs are not opaque and that the changes in radius must be due to some instability below the photosphere of the star. The instability is probably due to the large radiation pressure and the resulting small effective attraction in the outer layers of the star. The stellar layers are then only loosely bound to the star so that even small disturbances can have large effects. It has been suggested that the large eruptions are simply the extremes of the typical LBV variations.

Bibliography

The observations and properties of LBVs have been discussed in the review paper:

Humphreys R M and Davidson K 1994 The Luminous Blue Variables: Astrophysical Geysers *Publ. Astron. Soc. Pacific* **106** 1025

Reviews on specific properties of LBVs, their evolution and their variability can be found in the proceedings of conferences that were specially devoted to these stars:

Davidson K, Moffat A and Lamers H J G L M (ed) 1989 *Physics of Luminous Blue Variables* (Dordrecht: Kluwer)

Nota A and Lamers H J G L M (ed) 1997 *Luminous Blue Variables: Massive Stars in Transition* (Astron. Soc. Pacific Conf. Series, vol 120)

Henny J G L M Lamers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Luna

Series of Soviet automated Moon missions launched 1959–76. Lunas 16, 20 and 24 were successful sample return missions. Lunas 17 and 21 carried the first automated Moon rovers, known as Lunokhods 1 and 2. (See also LUNOKHOD.) Luna 3 returned the first pictures of the lunar far side. Lunar 9 sent back the first pictures from the Moon's surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lunar Orbiter

Series of five NASA Moon-orbiting spacecraft, launched 1966–67. Designed to map potential APOLLO landing sites. Perturbations in the orbit of Lunar Orbiter 1 provided the first data on lunar mass concentrations (mascons) and associated gravity anomalies. Used to create the first detailed lunar atlas, with spatial resolution down to 1 m.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lunar Prospector

NASA Discovery mission, launched in January 1998. Designed to spend one year in orbit around the Moon. Carried a neutron spectrometer which gave strong indications of water ice in shadowed craters at both lunar poles. Also gave high resolution gravity data and returned information on the composition of the lunar crust. Lowered into a 30 km altitude orbit in January 1999 on completion of its primary mission. The extended mission was due to end in July 1999.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lunar Transient Phenomenon (LTP)

A purported localized and short-lived change in the appearance of a feature on the surface of the Moon. LTPs tend to be reported by amateur observers of the Moon. They take various forms, including temporary colorations, usually red; bright flashes, visible especially in shadows or on the night side; extended cloudy patches; and obscurations of normally visible features.

Permanent changes to lunar features reported before the days of close-range photography from spacecraft are now discounted, being ascribed to deficiencies in mapping, observational error, or wishful thinking. A celebrated instance was the announcement in 1866 by Julius Schmidt that the crater Linné, described by previous observers as a small, deep pit, had vanished, to be replaced by a white patch. Modern photographs show Linné as a small crater of standard morphology, surrounded by light-colored ejecta.

Temporary changes reported more recently may be genuine, but they remain controversial. In 1958 the Soviet astronomer Nikolai Kozyrev spotted a red glow at the center of the crater Alphonsus, and obtained a spectrum which showed emissions of molecular carbon which he claimed was evidence of a volcanic emission. The observation was not corroborated, and evaluation was clouded by Cold War mistrust. Lack of corroboration remains a problem, as LTPs are usually seen by a single observer, and not photographed.

Unequivocal sightings of LTPs could be evidence that the Moon is not the geologically inert body it is generally held to be. There is a connection between LTPs and moonquakes—seismic tremors detected by instruments left on the Moon during the Apollo missions that originate near the crust/mantle boundary. Both phenomena are more common when the Moon is at perigee. The suggestion is that tidal flexing is triggering the release of gas or dust. It is possible that brief LTPs are caused by the impact of small meteorites.

See also: Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lunar-A

Japanese lunar orbiter, scheduled for launch in 2003. Intended to fire three penetrators 1–3 m into the lunar surface to detect Moonquakes and measure soil temperature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lund Observatory

Lund Observatory is part of Lund University, Lund, Sweden, active in astronomy since 1668. The past century featured stellar statistics (C V L Charlier) and (extra)galactic studies (K Lundmark, who produced a Milky-Way panorama in the 1950s). Current projects include stellar astrophysics, space astrometry, design of large optical and radio telescopes, and involvement in various international observatories.

For further information see
http://www.astro.lu.se/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lunokhod

Two Soviet Moon rovers, the first automated roving vehicles to operate on another world. Operated from Earth by remote control. Carried stereo cameras, laser reflector, magnetometer, x-ray spectrometer and cosmic ray detector. Lunokhod 1 was launched on board Luna 17 in November 1970. It landed on Mare Imbrium and operated for 322 days, covering 10.5 km. Lunokhod 2 was launched on board Luna 21 in January 1973. It landed to the east of Mare Serenitatis and operated for four months, covering 37 km. (See also LUNA.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lupus

(the Wolf; abbrev. Lup, gen. Lupi; area 334 sq. deg.) A southern constellation which lies between Centaurus and Scorpius, and culminates at midnight in early May. It has long been associated with a wild animal, though the identification with a wolf dates from comparatively recent times. It is usually shown on early celestial charts as a wolf impaled on a thyrsus (staff) held by Chiron, the centaur, represented by the neighboring constellation of Centaurus. The brightest stars of Lupus were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A small but not inconspicuous constellation, overshadowed by Centaurus to the west, the brightest stars in Lupus are α Lupi, magnitude 2.3, β Lupi, magnitude 2.7, and γ Lupi, a very close binary with bluish-white (B2) components, magnitudes 3.5 and 3.6 (combined magnitude 2.8), separation 0.8″. There are ten other stars of magnitude 4.0 or brighter. Other interesting binaries include ξ Lupi, which has bluish-white (A3 and B9) components, magnitudes 5.1 and 5.6, separation 10.4″, and μ Lupi, a triple system consisting of a close pair of bluish-white (B7) components, magnitudes 5.0 and 5.1 (combined magnitude 4.3), separation 1.1″, and a third, bluish-white (A2) component, magnitude 7.2, separation 23″. Other interesting objects include NGC 5822, an open cluster of more than 100 stars between ninth and twelfth magnitude, and NGC 5986, a ninth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Luyten, Willem Jacob (1899–1994)

American astronomer, worked at the University of Minnesota, determining proper motions of more than 120 000 stars, as a means of finding nearby or high-velocity stars, and determining their distances. He repeated the Palomar Observatory Sky Survey, building an automated computerized plate-measuring machine to compare it with the earlier survey, and thus determined the proper motions of 400 000 more. With these data he discovered the great majority of the white dwarfs known.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyman Alpha Absorption: The Damped Systems

Damped Lyα absorption systems are high column density gaseous objects that are detected in the spectra of background QUASARS on the basis of neutral hydrogen Lyα absorption lines broadened by radiation damping to large rest-frame equivalent widths. The damped Lyα absorption systems are of particular interest because (1) they can be identified to very large REDSHIFTS, (2) they dominate the mass density of neutral gas in the universe, containing at redshifts $z \approx 3$–5 a comoving mass density of neutral gas roughly comparable to the mass density of stars in present-day galaxies, and (3) they are predominantly neutral, allowing for heavy element abundances to be measured without the need of applying large ionization corrections. Many, most, or all damped Lyα absorption systems probably arise in the central parts of ordinary galaxies, although the exact relationship between damped Lyα absorption systems and galaxies is not yet firmly established.

It has been known since the late 1960s that the rest-frame ultraviolet spectra of HIGH-REDSHIFT QUASARS exhibit large numbers of discrete, isolated, narrow absorption lines. It was very quickly noted that the rate of incidence of these absorption lines is much larger at wavelengths shortward of the quasar Lyα emission lines than at wavelengths longward of the quasar Lyα emission lines. The reason is that almost all of the absorption lines at wavelengths shortward of the quasar Lyα emission lines are Lyα (or perhaps Lyβ, Lyγ, or higher-order Lyman transition) absorption lines produced by cosmologically distributed gaseous material that by chance intervenes along the lines of sight to the background quasars, at redshifts less than those of the background quasars. Collectively, the absorption lines that occur at wavelengths shortward of the quasar Lyα emission lines are known as the 'LYMAN ALPHA FOREST', because they give the impression of a thick forest of absorption that blankets the rest-frame ultraviolet spectra of high-redshift quasars.

The vast majority of the Lyα-forest absorption lines are of extremely low neutral hydrogen column density—typically $N = 10^{13}$–10^{15} cm^{-2}. But a very small minority of the Lyα-forest absorption lines are of much higher neutral hydrogen column density—ranging up to nearly $N = 10^{22}$ cm^{-2}. At neutral hydrogen column densities exceeding $N \approx 2 \times 10^{20}$ cm^{-2}, the Lyα absorption line profiles are optically thick in the radiation damping wings, and the Lyα absorption lines occur on the damped part of the curve of growth. For this reason, the highest neutral hydrogen column density Lyα-forest absorption lines are known as 'damped Lyα absorption lines', and the redshift complexes of absorption lines (consisting of Lyα and higher-order Lyman transitions as well as various heavy element transitions) associated with damped Lyα absorption lines are known as 'damped Lyα absorption systems'.

The column density threshold that distinguishes high column density damped Lyα absorption systems from lower column density Lyα-forest absorption systems is usually taken to be $N = 2 \times 10^{20}$ cm^{-2}. Although there is no particular physical significance that can be ascribed to this value, it is nevertheless a convenient choice because (1) absorption lines arising from such large neutral hydrogen column densities occur on the damped part of the curve of growth for any value of the velocity dispersion satisfying $\sigma < 100$ km s^{-1} (which in practice includes all quasar absorption line systems) and (2) it corresponds to the isophotal sensitivity threshold of the 21 cm neutral hydrogen survey of nearby galaxies conducted by Bosma in 1981, thus allowing for straightforward comparison of properties of damped Lyα absorption systems with properties of present-day galaxies. For present-day SPIRAL GALAXIES, for example, the Bosma radius (or the radius at which the gaseous disk reaches a neutral hydrogen column density threshold of $N = 2 \times 10^{20}$ cm^{-2}) extends to roughly 1.8 times the Holmberg radius (or the radius at which the stellar disk reaches the surface brightness of the night sky).

For Lyα absorption lines on the damped part of the curve of growth, the relationship between neutral hydrogen column density N and rest-frame Lyα equivalent width W is approximately given by

$$\frac{N}{2 \times 10^{20}\ \mathrm{cm}^{-2}} \approx \left(\frac{W}{10\ \text{Å}}\right)^2. \qquad (1)$$

A column density threshold $N = 2 \times 10^{20}$ cm^{-2} thus corresponds to a rest-frame Lyα equivalent width threshold $W = 10$ Å. In practice, damped Lyα absorption systems are identified on the basis of observed-frame equivalent width limited surveys for damped Lyα absorption lines. At high redshifts, even the lowest column density damped Lyα absorption systems give rise to Lyα absorption lines of observed-frame equivalent widths of many tens of Å and are easily detected in spectra of relatively low spectral resolution and signal-to-noise ratio. Roughly 1000 quasars have so far been surveyed for damped Lyα absorption lines, from which roughly 100 damped Lyα absorption systems have so far been identified, at redshifts ranging from $z \approx 0$ through 5.

At high redshifts, the observed rate of incidence of damped Lyα absorption systems is significantly larger than the predicted rate of incidence of damped Lyα absorption systems, under the assumptions that the absorption systems arise within the Bosma radii of spiral galaxies and that galaxies are fixed in comoving coordinates. The exact factor by which the observed rate of incidence exceeds the predicted rate of incidence depends on the choice of cosmological model, but could be as large as ≈ 2 for plausible cosmological parameters. This important result has been interpreted as suggesting that the gaseous disks of spiral galaxies were larger or more numerous in the past.

In more or less all cases that have so far been examined, the damped Lyα absorption systems exhibit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

heavy element absorption lines from abundant low-ionization species, including transitions of C II, O I, Mg II, Si II and Fe II. In many (but not all) cases that have so far been examined, the damped Lyα absorption systems also exhibit heavy element absorption lines from abundant high-ionization species, including transitions of C IV and Si IV. The low- and high-ionization line profiles appear to be kinematically distinct, suggesting a multiple phase structure of the absorbing regions. In spectra of high spectral resolution and signal-to-noise ratio, damped Lyα absorption systems also exhibit heavy element absorption lines from low-abundance species, including transitions of Ti II, Cr II, Ni II and Zn II. These low-abundance species are of particular importance for measuring heavy element abundances.

Neutral gas content

The most important property of the damped Lyα absorption systems is that they dominate the mass density of neutral gas in the universe, containing at redshifts $z \approx 3$–5 a comoving mass density of neutral gas roughly comparable to the mass density of stars in present-day galaxies. Because stars form from neutral (rather than ionized) gas, this has the important implication that the damped Lyα absorption systems trace the bulk of material available for STAR FORMATION to high redshifts.

It is straightforward to show that the comoving mass density of neutral gas follows directly from observations of damped Lyα absorption systems, without regard to the geometry of the absorbing regions. Consider a population of absorption systems of some given neutral hydrogen column density N_0 and rate of incidence $n(z)$ per line of sight intercepted at redshift z over some redshift interval $\mathrm{d}z$ or cosmic time interval $\mathrm{d}t$. The proper mass density of neutral hydrogen contributed by the absorbers is obtained by taking the product $n(z)\,\mathrm{d}z$ (which gives the mean number of absorption systems intercepted by a line of sight over the redshift interval $\mathrm{d}z$), multiplying by the column density N_0 (which then gives the mean column density traversed by a line of sight over the redshift interval $\mathrm{d}z$), multiplying by the mass of the hydrogen atom m_{H} (which then gives the mean mass column density of a line of sight over the redshift interval $\mathrm{d}z$), and dividing by $c\,\mathrm{d}t$, where c is the speed of light.

The comoving mass density of neutral gas is then formed by multiplying by a factor $\mu = 1/(1-Y) \approx 0.33$ to allow for a helium abundance $Y = 0.25$ (by mass), dividing by $(1+z)^3$ to allow for the expansion of the universe, and dividing by the current critical density ρ_{crit} to yield a dimensionless mass density parameter Ω_{g}. The result is the comoving mass density of neutral gas in units of the current critical density.

In practice, the comoving mass density of neutral gas is determined not for absorption systems of some given neutral hydrogen column density N_0 but for absorption systems that span a range of neutral hydrogen column densities. The distribution of neutral hydrogen column densities is characterized by the column density distribution function $f(N)$, which is defined in such a way that the $f(N)\mathrm{d}N\mathrm{d}X$ is the number of absorption systems intercepted per line of sight with neutral hydrogen column density in the interval N to $N + \mathrm{d}N$ and in the absorption distance interval $\mathrm{d}X$. In terms of the column density distribution function, the mass density parameter is given by

$$\Omega_{\mathrm{g}} = \frac{H_0}{c}\frac{\mu m_{\mathrm{H}}}{\rho_{\mathrm{crit}}}\int_{N_{\mathrm{min}}}^{N_{\mathrm{max}}} N f(N)\mathrm{d}N \qquad (2)$$

where N_{min} and N_{max} are the minimum and maximum neutral hydrogen column density limits, respectively, and H_0 is the HUBBLE CONSTANT.

Over the range of neutral hydrogen column densities that have yet been measured, the column density distribution function of the damped Lyα absorption systems is well represented by a power-law form as

$$f(N) = BN^{-\beta} \qquad (3)$$

where the power-law index is measured to be $\beta \approx 1.6$. For such a shallow column density distribution function, the mass density parameter of equation (2) formally diverges as $N_{\mathrm{max}} \rightarrow \infty$, indicating that $f(N)$ must steepen at some sufficiently large neutral hydrogen column density. This steepening has not yet been measured. To obtain a convergent result, equation (2) is usually evaluated with N_{max} set equal to the largest observed neutral hydrogen column density, which at high redshifts approaches $N_{\mathrm{max}} = 10^{22}$ cm^{-2}. Uncertainty of the shape of the high column density end of the column density distribution function is the major uncertainty of Ω_{g}.

At redshifts $z \approx 3$–5, evaluation of equation (2) yields a mass density parameter $\Omega_{\mathrm{g}} \approx 2 \times 10^{-3}\ h^{-1}$, where h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1} and where the exact value depends on the choice of cosmological model. This is equal to within observational uncertainty to the mass density parameter of stars in present-day galaxies, which is determined by multiplying the luminosity density of the nearby universe by the mean mass-to-light ratio of the stars that dominate the luminosities of galaxies. This very important result has been interpreted as indicating that high-redshift damped Lyα absorption systems are the progenitors of present-day galaxies, i.e. that the gaseous content of high-redshift damped Lyα absorption systems is gradually converted into the stellar content of present-day galaxies. It is unclear whether a systematic decrease in Ω_{g} with decreasing redshift—reflecting the conversion of gas into stars with time—has yet been measured.

Heavy element abundances

Another important property of the damped Lyα absorption systems is that they are predominantly neutral, allowing for heavy element abundances to be measured without the need of applying large ionization corrections. In particular, because the absorbing regions are self-shielded against ionizing radiation by high column density gas, neutral hydrogen and singly-ionized heavy elements (of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

elements with first ionization potentials less than one Rydberg) are the dominant ionization species. Gas-phase heavy element abundances are therefore obtained simply by dividing column densities of singly-ionized heavy element species by column densities of neutral hydrogen.

Absorption lines from low-abundance singly-ionized species, including transitions of Ti II, Cr II, Ni II and Zn II, are of particular importance because they are usually unsaturated (or only mildly saturated), allowing accurate column densities to be measured. For several reasons, most attention has so far been focused on transitions of Zn II and Cr II. In the interstellar medium of the Galaxy, Zn is not highly depleted onto dust grains, suggesting that the gas-phase abundance of Zn indicates the true abundance of Zn. In contrast, in the interstellar medium of the Galaxy, Cr is highly depleted onto dust grains, suggesting that the gas-phase abundance ratio of Zn to Cr is an indicator of the presence or absence of dust. Hence a single measurement of Zn II and Cr II column densities can simultaneously provide information about heavy element abundances and dust content.

At high redshifts ($z \approx 3$), Zn II and Cr II column densities of several dozen damped Lyα absorption systems have so far been measured. Results of these measurements indicate that the mean heavy element abundances are ≈ 0.1 times solar values, with a dispersion of a factor of two or three about this value. (Heavy element abundances of high-redshift damped Lyα absorption systems range from ≈ 0.01 times solar values to ≈ 0.5 times solar values.) The Zn to Cr abundance ratios suggest the presence of modest amounts of dust, although it is not exactly clear what values the undepleted abundances should take on—especially of Zn.

At low redshifts ($z < 1.6$), Zn II and Cr II column densities of only a few damped Lyα absorption systems have so far been measured. It is especially striking that there is no pronounced trend for heavy element abundances to increase with decreasing redshift, and in particular no trend for heavy element abundances to approach solar values at redshifts $z \approx 0$, as would be expected under scenarios of 'cosmic' chemical evolution due to the conversion of gas into stars. It is not yet clear how to interpret these results.

Relationship with galaxies

The most important outstanding issue regarding the damped Lyα absorption systems concerns their relationship with galaxies. Although the circumstantial evidence linking damped Lyα absorption systems with galaxies is substantial, attempts to observationally establish a direct relationship between damped Lyα absorption systems and galaxies have met with only limited success. Galaxies responsible for a small number of damped Lyα absorption systems have been identified, at both low and high redshifts. But galaxies responsible for other damped Lyα absorption systems have eluded identification, some despite very substantial effort. So while it is known that at least some damped Lyα absorption systems arise in galaxies, it is not known what range of properties (in terms, for example, of morphological type, luminosity, and surface brightness) are spanned by the absorbing galaxies, or even whether all damped Lyα absorption systems arise in galaxies.

Part of the difficulty is that the high neutral hydrogen column densities of damped Lyα absorption systems imply small impact parameters to the lines of sight, and it is difficult to identify faint galaxies at small angular separations to bright quasars. Recent advances in ADAPTIVE OPTICS technology may finally allow progress in this area.

Understanding the relationship between damped Lyα absorption systems and galaxies is probably the key to interpreting the heavy element abundance results. Measurements of damped Lyα absorption systems probe truly cosmic quantities in a way that measurements of few other classes of objects can do. The neutral hydrogen and heavy element content of the high-redshift universe—which is established on the basis of a uniform collection of measurements of damped Lyα absorption systems—is far more completely and homogeneously determined than the neutral hydrogen and heavy element content of the nearby universe—which is established on the basis of a diverse collection of visible- and radio-wavelength observations of gas, stars, and H II regions. Establishing the relationship between damped Lyα absorption systems and galaxies could well lead to surprises about where much of the neutral gas and heavy element content of galaxies reside.

Bibliography

Lanzetta K M, Wolfe A M and Turnshek D A 1995 The *IUE* survey for damped Lyα and Lyman-limit absorption systems: evolution of the gaseous content of the universe *Astrophys. J.* **440** 435

Lanzetta K M, Wolfe A M, Turnshek D A, Lu L, McMahon R G and Hazard C 1991 A new spectroscopic survey for damped Lyα absorption lines from high-redshift galaxies *Astrophys. J.* **375** 1

Prochaska J X and Wolfe A M 2000 Metallicity evolution in the early universe *Astrophys. J.* **533** 5

Storrie-Lombardi L J, McMahon R G and Irwin M J 1996 Evolution of neutral gas at high redshift: implications for the epoch of galaxy formation *Mon. Not. R. Astron. Soc.* **283** 79

Wolfe A M, Turnshek D A, Smith H E and Cohen R D 1986 *Astrophys. J. Suppl.* **61** 249

Kenneth M Lanzetta

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyman Alpha Forest

The Lyman alpha forest is an absorption phenomenon seen in the spectra of high redshift QSOs and galaxies (figure 1). It is the only direct observational evidence we have of the existence and properties of the general INTERGALACTIC MEDIUM, and, as we have reason to believe, of most of the baryonic matter contents of the universe.

On its way to us the light of a bright, distant QSO passes through intervening intergalactic gas and through gas clouds associated with foreground galaxies. Absorption by the gas modifies the spectra of the background objects and imprints a record of the gas clouds' physical and chemical states on the observed background QSO and galaxy spectra. The whole arrangement is reminiscent of a giant cosmic slide projector, where a QSO plays the role of the light bulb, and the intervening gas clouds are the slides, changing the colors of the light source by absorbing parts of the (white) spectrum.

The name 'Lyman α forest' refers to the appearance of the optical QSO spectra, which show a forest of hundreds of sharp absorption lines, mostly from the neutral hydrogen (H I) Lyman α line, superimposed on the more smoothly varying QSO continuum (figure 2). Almost all the lines in the Lyman α forest correspond to the same atomic transition (which at 1215.67 Å is in the ultraviolet wavelength region). The phenomenon was first observed in the optical waveband ($\sim$ 4000–9000 Å) implying that the gas clouds causing the absorption are highly redshifted by the Hubble expansion. The absorption systems appear spread out into a 'forest' of lines because each line is redshifted by a different amount in proportion to the absorbing cloud's distance from us. The stronger ones among the absorption systems do show further spectral signatures in addition to their Lyα line: higher-order Lyman series lines begin to be detectable for absorption systems with Lyα close to saturation. Clouds with H I column densities larger than $N \sim 10^{17}$ cm^{-2} start showing a discontinuity due to continuous absorption at a rest frame wavelength 912 Å, beyond the limit of the Lyman series. These 'Lyman limit systems' occupy a column density regime where a gas cloud starts shielding itself against ionizing radiation from the outside. Clouds with even higher column densities ($N > 10^{19}$ cm^{-2}) exhibit the damping wings caused by the internal finite lifetime of the Lyα transition. The gas in these 'damped Lyα' systems (see LYMAN ALPHA ABSORPTION: THE DAMPED SYSTEMS) is almost completely self-shielded and mostly neutral. Most absorption systems with $N > 10^{14.5}$ cm^{-2} also show metal absorption lines (triply ionized carbon and silicon, and some other common elements and ionization stages). For that reason the higher column density systems are usually referred to as 'metal' or 'heavy element' systems. Here we are concerned only with the low column density gas, i.e., those absorption systems where the Lyα line is not saturated, which we will refer to as the Lyα forest proper. There definitely is an overlap between metal systems and Lyα forest systems in this restricted definition, but we have theoretical reasons to believe that there is a genuine dichotomy between intergalactic gas (represented by the Lyα forest), even if partly polluted by metals, and the invariably metal-enriched higher column density systems, thought to be related to galaxies.

Lyα forest absorption systems have now been observed from REDSHIFT zero (with UV satellites) up to the highest redshifts at which background light sources (QSOs and galaxies) can still be found (currently $z \sim$ 5–6).

The Gunn–Peterson effect: where does the absorption come from?

Lyα forest absorption in a QSO spectrum was predicted and first detected by Gunn and Peterson (1965). The basic idea is as follows: going back in time an increasing fraction of the total baryonic mass of the universe must be in the form of gas. The absorption cross-section of the Lyα line of neutral hydrogen is large enough that even if only a small fraction of the total mass of the universe were in the form of H I the redshifted Lyα lines should completely absorb a part of the spectrum of any background light source. The absorption should essentially assume the shape of an absorption trough in a QSO spectrum, extending blueward from the Lyα emission of the QSO. This particular absorption pattern is referred to in the literature as the 'Gunn–Peterson effect'. Gunn and Peterson did detect such a trough but the light of the QSO was not completely absorbed and there was some residual light left in the spectral region in question. The relative weakness of the absorption could mean two things: (a) there is little hydrogen left in intergalactic space and by the time the Lyα forest is observed most of the matter has already condensed into galaxies. Or (b) most of the hydrogen is not in neutral form, where it can produce Lyα absorption, but is fully ionized.

QSO surveys (see QUASISTELLAR OBJECTS: SURVEYS) later showed that the second possibility is more important: the combined ionizing radiation output from all known QSOs at high redshift amounts to a UV radiation field probably strong enough to keep most of the baryonic matter in the universe highly ionized—i.e., if the baryons are predominantly in the form of a more or less homogeneously distributed gas.

This conclusion is based on the assumption that the gas is in approximate photoionization equilibrium with the cosmic UV background field, i.e., the rate of recombinations of electrons with protons to form neutral hydrogen balances the rate of ionizations from the ground state of H I,

$$n_{\rm e} n_{\rm p} \alpha(T) = n_{\rm HI} \Gamma. \qquad (1)$$

Here Γ is the rate of photoionizations per neutral hydrogen atom, caused by the hydrogen-ionizing portion of the UV background field. The quantities $n_{\rm e}$, $n_{\rm HI}$ and $n_{\rm p}$ refer to the number densities of electrons, neutral hydrogen atoms, and protons (= ionized hydrogen), respectively.

Later observations of QSOs with higher spectral resolution (< a few hundred km s^{-1}) showed that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

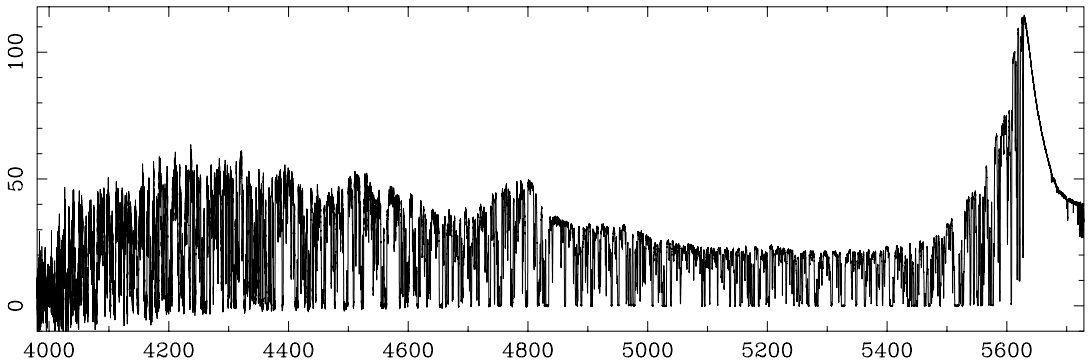


Figure 1. High resolution spectrum of the Lyα forest part of a redshift 3.63 QSO, taken with the HIRES spectrograph on the Keck 10 m telescope in Hawaii. The plot shows the flux of the QSO in arbitrary units versus the observed wavelength in units of Angstroms. The noise level can be judged from the longer wavelength wing of the broad, intrinsic Lyα emission line of the QSO (near 5650 Å). All of the ragged features are high redshift absorption lines. Most of the lines between the QSO's Lyα and Lyβ emission lines (the humps at 5650 and 4750 Å) are due to Lyα absorption by intervening gas. The actual rate of incidence of absorbers decreases towards shorter wavelengths (lower redshifts). Nevertheless, the line density increases to the blue because higher order absorption lines from the Lyman series appear and overlap randomly with Lyα lines of systems at lower redshift.

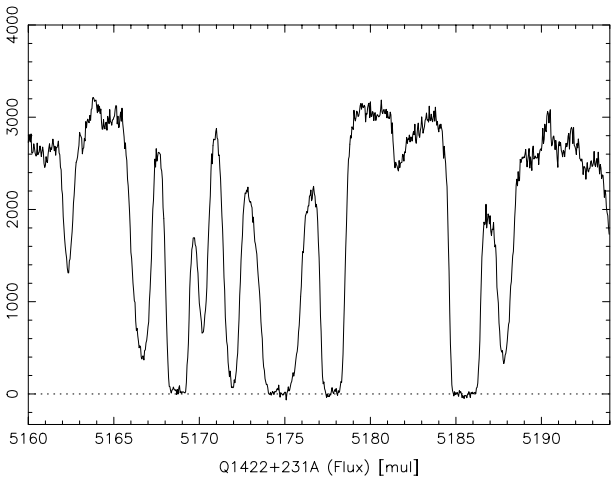


Figure 2. Detailed section of the previous spectrum. The image shows a number of absorption lines all corresponding to the neutral hydrogen (H I) Lyα 1215.67 Å transition. Lines close to saturation (= zero flux in the line center) have neutral hydrogen column density typically around $N \sim 10^{14.2}$ cm^{-2}, corresponding to a gas density enhanced by roughly an order of magnitude with respect to the mean density of the universe. The mean redshift of the stretch shown is $z = 3.248$. The spectral region extends over 3480 km s^{-1}. For a flat $\Omega = 1$ universe this corresponds to a spatial extent of approximately 9.6 h^{-1} Mpc along the line-of-sight.

what appeared as a smooth absorption 'trough' to earlier observers is in fact a ragged 'forest' of hundreds of individual absorption lines. In other words, the distribution of neutral hydrogen in the universe is inhomogeneous on scales down to the width of a typical Lyα line (see below). The degree of clumpiness appears magnified by the absorption pattern in the QSO spectra, because the residual (= unabsorbed) portion of the QSO's flux $I \propto e^{-\tau}$ depends exponentially on the Lyα optical depth τ, which itself depends almost quadratically on the gas density (or the electron density n_e; for a highly ionized gas at constant temperature). Thus small density fluctuations produce enhanced fluctuations in the optical depth. The Lyα forest absorption is observed in velocity space, and a convergent velocity field (e.g., a collapsing gas cloud) could also produce absorption 'lines'. Caustics in velocity space may form if several gas volume elements are moving at the same velocity relative to the observer. Indeed, if Lyα clouds are produced by gravitational collapse, both overdense regions and infall should contribute to an absorption line.

Spectroscopy of the Lyα line is an incredibly sensitive method to detect baryonic matter at any redshift. The photoionization cross section of neutral hydrogen is so large that an extremely tenuous gas at or below the mean density of the universe can be detected easily in absorption. The method of choice for studying the Lyα forest is optical high resolution spectroscopy, with a spectral resolution $\Delta\lambda/\lambda > 30\,000$ sufficient to resolve Lyα lines thermally broadened by the photoionization heating from the UV background. With 8 m class telescopes, a spectrum of a QSO suitable for further analysis of the Lyα forest absorption can be obtained within a few hours of observing time.

Basic observational properties of Lyα absorbers

Early models of Lyα absorption systems envisaged the absorption lines as arising from discrete 'clouds' of gas in intergalactic space. The clouds would be analoguous to galaxies, but the gas densities appeared too tenuous and too highly ionized to give rise to star formation. Moreover, the average clustering of the lines in velocity space was too weak for the clouds to be related directly to galaxies. Thus the gas giving rise to Lyα absorption systems came

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to be referred to as 'Lyα clouds' or 'intergalactic clouds', to distinguish it from gas associated with galaxies. It it is worth keeping in mind that even the term 'cloud' does already imply a prejudice as to the spatial distribution of the gas.

Before the advent of cosmogonies dominated by DARK MATTER (which could have supplied the gravitational attraction necessary to hold together the gas), it was thought more likely that the intergalactic clouds would be confined by the pressure of an even hotter and more tenuous intercloud medium. In this model, a cloud would be homogeneous, static or slowly expanding and, at least for some of its lifetime, in thermal photoionization equilibrium with the UV background.

Such an object has a simple observational signature in absorption. With only the intrinsic atomic line width of the Lyα transition and purely thermal motions contributing to the absorption line broadening, an individual line profile consists of a convolution of a Lorentzian resonance curve and a Maxwell–Boltzmann velocity distribution. The result is a so-called Voigt profile. Each such profile is fully specified by only 3 parameters: its position in velocity space (redshift), its column density (i.e. the number of neutral hydrogen atoms per unit area), and its line width (traditionally expressed in terms of the Doppler parameter $b = \sqrt{2}\sigma$, where σ is the standard deviation of the Maxwellian velocity distribution). The Lyα forest as a whole can then be characterized in terms of the distribution functions of these three quantities.

Column densities

Let us define $\mathrm{d}\mathcal{N}/\mathrm{d}N$ as the number of absorption lines, $\mathrm{d}\mathcal{N}$, per unit redshift with an H I column density between N and $N + \mathrm{d}N$. This function tells us how likely it is for our line-of-sight to a background QSO to intersect a cloud with a given H I column density N. Observationally, for H I column densities N spanning the amazing range from 10^{12} to 10^{22} cm^{-2} the distribution was found to be well-parametrized by a single power law in column density,

$$\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}N} \propto N^{-1.5}. \qquad (2)$$

To give an idea of the normalization, there are hundreds of unsaturated lines ($N < 10^{14.5}$ cm^{-2}) per unit redshift but, typically, less than one damped Lyα system ($N > 10^{21}$ cm^{-2}). For comparison, a QSO spectrum at redshift 3 covers $\Delta z \sim 0.6$ of the Lyα forest (between the QSO redshift and the redshift corresponding to the onset of the Lyβ series).

The distribution of absorption systems in velocity space

We may consider two limiting cases. (a) The distribution of absorbers over very large (Gigaparsec) scales; this is is equivalent to a distribution in time or redshift, as the universe changes considerably during the time it takes a light ray to traverse a Gigaparsec. (b) The distribution on small (Megaparsec) scales; here we expect local astrophysics (gravity, galactic outflows, nearby sources of ionizing radiation) to affect the clustering of the lines in velocity space.

(a) The evolution of the number of absorption lines $\mathcal{N}$ with time or redshift z (and with a column density above a certain threshold) traditionally has been approximated by a power-law in $(1 + z)$,

$$\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}z} \propto (1 + z)^{\gamma}. \qquad (3)$$

For a universe where Lyα clouds are non-evolving, $0 < \gamma < 1/2$, depending on the cosmological model. With the observational baseline now extending from the local universe out to redshift five it has become clear that γ is not constant with time. In the local universe ($z \sim 0$) up to redshift ~ 1.5, γ is consistent with the above no-evolution values. However, beyond redshift $z \sim 1.5$ there is a sharp increase in the number of absorption lines, with γ rising up to a value $\gamma \sim 5$ by redshift $z \sim 4$.

There is a general trend for the number of absorption lines to increase with redshift, but in any individual QSO spectrum there is a relative lack of absorption systems close to the redshift of the QSO. This so-called 'proximity effect' has been ascribed to the ionizing radiation produced by the QSO itself, which reduces the neutral hydrogen fraction for gas close to the QSO. The effect can be used to measure the strength of the ionizing background radiation: consider a point at a distance from the QSO, where the number of Lyα systems has declined by half as compared to the average line density in the Lyα forest. At this point the intensity of ionizing radiation from the QSO must equal the intensity of the general UV background. Knowing the luminosity of the QSO we can compute the flux at that point, and thus, the intensity of the ionizing background. This method, though fraught with many uncertainties, has yielded the first measurement of the ionizing UV background.

(b) The small-scale distribution of absorption lines in velocity space along the line-of-sight, often referred to as 'clustering', has yielded only limited information about the nature of the Lyα forest systems. If the clustering is measured with the same methods used for galaxy surveys, namely by applying the two-point-correlation function to discrete absorption lines, there is a significant signal only for the highest column density ($N > 10^{15}$ cm^{-2}) systems. These stronger absorption lines are invariably accompanied by absorption from heavier elements (carbon, silicon, oxygen, iron in various stages of ionization), so the high column density clouds must in some way be more closely associated with galaxies as the production sites of the metals. The observed stronger clustering of this gas may arise when gas clouds move in a galactic gravitational potential well, or they may just reflect stellar ejecta in the interstellar medium of those galaxies. However, most of the Lyα forest consists of weaker lines, which do not cluster along the line-of-sight appreciably. Gas densities inferred for the weaker lines are likely to be much lower than for galaxies. If gravity is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the main structure forming agent it is not very surprising that lower density regions are clustered more weakly.

Searches for voids in the Lyα forest similar to those seen in the spatial distribution of galaxies have been equally unsuccessful, implying again that most Lyα absorption systems are more homogeneously distributed in space than galaxies.

Absorption line widths

The width of an absorption line is a measure of the total velocity distribution in the gas. Both microscopic (thermal motion) and macroscopic processes (turbulence, bulk flows, broadening of an extended object by the Hubble expansion) may contribute to the line profile. Without an *a priori* theory about line formation the total width b of the line profile can only be used as an upper limit to the width due to any individual process. As an example, the Doppler parameter of a thermally broadened absorption line with an additional Gaussian contribution describing turbulence in the gas is given exactly by

$$b = \sqrt{\frac{2kT}{m} + b_{\rm turb}^2} \tag{4}$$

where T is the temperature of the gas, k is Boltzmann's constant, m is the mass of the hydrogen atom, and $b_{\rm turb}$ is the turbulent contribution to the Doppler parameter.

At redshift $\sim$ 3 observations show the Doppler parameters b to be distributed according to a Gaussian with a mean $\overline{b}$ between 25 and 30 km s^{-1}, and a cutoff $b_{\rm cut}$ below about 19 km s^{-1}, i.e.,

$$\frac{{\rm d}\mathcal{N}}{{\rm d}b} \propto \exp\left(\frac{(\overline{b}-b)^2}{2\sigma_{\rm b}^2}\right) \quad \text{for } b \geq b_{\rm cut} \tag{5}$$

$$= 0 \qquad \text{for } b < b_{\rm cut}, \tag{6}$$

This analytic relation gives a good description of the actual distribution. Doppler parameters below the cutoff are very rare, but there appears to be a weak non-Gaussian tail towards larger Doppler parameters. The parameters of the distribution quoted vary somewhat among different researchers, but there is agreement in that the line widths at higher redshifts (3.5–4.5) are lower by perhaps 25% .

If the broadening were purely thermal then typical b values as discussed above would correspond to upper limits on the gas temperature $T \leq 3$–5×10^4 K.

Size, density, and ionization state of Lyα clouds

Absorption line studies suffer from the limitation that in general only one-dimensional information along the line-of-sight is available. From the absorption line itself we measure the temperature and column density of the gas, and from observations of the QSO luminosity function we can compute an estimate of the ionizing background radiation, but to obtain the density and the ionized fraction of the gas an additional independent estimate of the size of the absorbing gas cloud is needed.

Observations of close lines-of-sight to groups of QSOs or to multiple images of gravitationally lensed QSOs can restore to some degree the missing second dimension and provide a measurement of the cloud sizes, at least as projected in the plane of the sky. Imagine that we had two lines of sight with a known separation. For Lyα 'clouds' with a given size, some of the clouds will intersect both lines-of-sight, others only one of them. In a statistical sense, the typical size of Lyα forest absorbers can be determined from the numbers of 'hits' and 'misses', i.e., from knowing how often an absorption system seen in one of the lines-of-sight also appears in the second one. Such estimates have yielded astonishingly large transverse sizes on the order of 0.1–1 Mpc proper separation. These sizes enable us to compute rough estimates of the ionization fraction and density of the gas.

Assuming that the neutral fraction x of hydrogen is determined by photoionization equilibrium (cf equation (1)), we have

$$x = 3.9 \times 10^{-6} \left(\frac{T}{3\times10^4}\right)^{-0.35} \left(\frac{I}{10^{-21}}\right)^{-0.5} \times \left(\frac{N}{10^{14}}\right)^{0.5} \left(\frac{D}{100\ {\rm kpc}}\right)^{-0.5}, \tag{7}$$

where T is the gas temperature in K, I is the intensity of the ionizing UV background in units of ergs Hz^{-1} sr^{-1} s^{-1} cm^{-2}, N the H I column density in cm^{-2}, and D the thickness of the cloud (or the path length of our line of sight through it).

Likewise, the total number density of the gas (the number of protons per cm^3) is given by

$$n = x^{-1} N D^{-1} = 8.3 \times 10^{-5} \left(\frac{T}{3\times10^4}\right)^{0.35} \left(\frac{I}{10^{-21}}\right)^{0.5} \times \left(\frac{N}{10^{14}}\right)^{0.5} \left(\frac{D}{100\ {\rm kpc}}\right)^{-0.5}. \tag{8}$$

This is about four orders of magnitude less dense than the gas in the disk of our Galaxy. Even if the clouds are not spherical but flattened (so they would appear more extended across than along the line-of-sight) the conclusion is inescapable that the gas is highly ionized and the density in the typical cloud is within a factor of a few from the mean density of the universe, far less than the average density in a galaxy.

The intergalactic medium as a cosmic fluid

So far we have used only simple astrophysical arguments to infer the basic physical properties of the Lyα absorbers, without explicit reference to a model of structure formation. The decomposition into Voigt profiles had originally been justified by the assumption that Lyα absorption systems are discrete, pressure-confined gas clouds in hydrostatic, thermal and ionization equilibrium, floating in intergalactic space. This picture provided an analytically tractable model which made a host of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observationally testable predictions, several of which unfortunately contradicted the observational evidence. For example, the large range in column densities observed, and the evolution of the number of absorption systems with redshift are difficult to reproduce in that model. Moreover, observations of the spectral shape of the cosmic microwave background radiation with the COBE satellite have provided arguments against the existence of a hot intercloud medium necessary to provide the confinement pressure for the Lyα clouds.

The pressure-confinement model finally fell from grace because of its lack of a convincing theory of the formation of the cloudlets, and the rise (in popularity) of structure formation cosmogonies dominated by the presence of large amounts of weakly interacting dark matter. The gravitational collapse of dark matter would have trapped large amounts of baryonic gas as well. The most popular, cold dark matter (CDM) structure formation scenario predicts a large abundance of collapsed CDM halos with individual masses too small to form stars and turn into galaxies. Warm photoionized intergalactic gas sinks into these 'mini'-halos or accretes onto dark matter filaments and sheets. The thermal gas pressure prevents the gas from further collapse; relatively stable gaseous configurations are formed. These structures are visible only in absorption since there are no stars which could produce any light.

A breakthrough in cosmology occured in the early 1990s when hydrodynamic cosmological simulations became available. As far as the comparison with data is concerned the hydrodynamics simulations represent an advance over pure dark matter (so-called N-body) simulations, as they attempt to directly predict observable astrophysical quantities. The hydro-codes include, in a simplified way, gas dynamics, elementary radiative processes, heating and cooling, and some schematic stellar feedback, all of which are essential to an understanding of the formation of structure for ordinary (baryonic) matter like galaxies, stars and intergalactic gas clouds. The ultimate hope of the hydro-simulations is to reproduce realistic galaxies, but at the current level of detail and spatial resolution possible the intergalactic medium with its simple physics is perhaps the most promising target for quantitative modelling.

The cosmic web

If the underlying cosmological picture (a universe dominated by cold dark matter) is correct, then the hydro-simulations are telling us that the spatial distribution and physical state of the Lyα forest gas is more complex than previously thought. The gas is arranged in filaments and sheets, in what has been called the 'cosmic web', closely tracing the dark matter distribution on large scales (figure 3). Low column density absorption systems ($N \lesssim 10^{14}$ cm^{-2}) are associated with sheet-like structures or pancakes of gas (length scale $\sim$ a few hundred kpc to 1 Mpc proper). The gas accretes through weak shocks (developing a double humped temperature profile), and

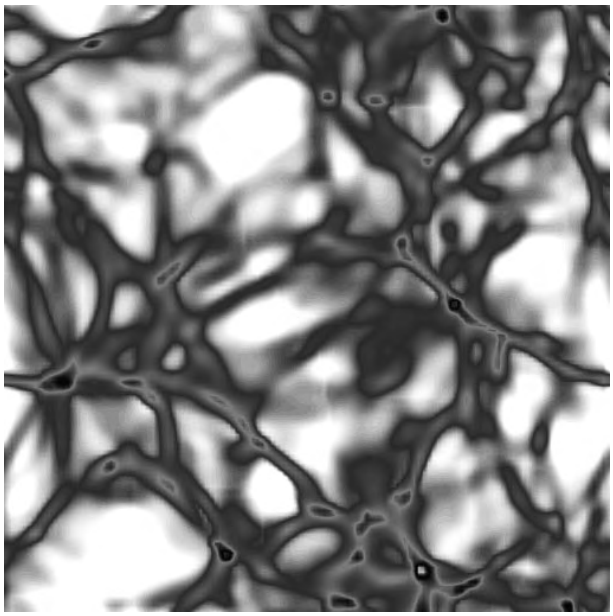


Figure 3. Cosmological hydro-simulation: projection of the baryonic density distribution in a simulated box at $z = 3$ (same spatial extent as for figure 2) showing the 'cosmic web' filamentary structure of the Lyα forest (image courtesy Michael Norman, University of Illinois).

settles in a dense, central cooling layer, presumably to form stars in some of the denser regions. At the lowest column densities gas remains unshocked and just bounces back because of the hydrostatic pressure. The gas is partly confined by dark matter gravity and partly by ram-pressure. Higher column density clouds arise in more filamentary structures, with column density contours of $N \sim 10^{14}$ cm^{-2} extending continuously and at relatively constant thickness ($\sim$ 40–100 kpc proper) over Mpc distances. With increasing column density the absorber geometry becomes rounder; column density contours at $N \gtrsim 10^{16}$cm^{-2} invariably are spherical. Such absorbers more closely correspond to the aforementioned minihalos; there the enclosed gas column is high enough to make the absorption system appear as a Lyman limit or damped Lyα system. Looking at the higher column density, optically thick gas on scales of several Mpc one gets a somewhat different impression of chains of mini- or larger halos, lining up like pearls on a string, quite similar to the structure seen in N-body simulations of the dark matter distribution. To produce as much absorption as observed, a large fraction of all baryons (80–90%) is required to reside in the low column density Lyα forest, mostly in the column density range $10^{14} < N < 10^{15.5}$ cm^{-2}.

In general this theoretical picture very well reproduces the observational properties mentioned earlier. The column density distribution is in excellent agreement, as are the large transverse sizes measured in projection, and the clustering along the line-of-sight. In this model the rapid evolution of the number of the absorption systems

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with redshift is largely a consequence of the expansion of the gas with the Hubble flow. The statistical distribution of the flux level $I = e^{-\tau}$ in the Lyα forest is extremely well matched by these models, as can be seen from a comparison of the observed distribution with one from a simulated ΛCDM universe (figure 4). The spatial correlations as evident from the flux power-spectrum are equally well reproduced. The most complex piece of observational information, the distribution of Doppler parameters is qualitatively understood, but as it depends on a variety of initial conditions (epoch of reionization, baryonic density of the universe, cosmological model) and possible stellar energy feedback, several of these parameters need to be tuned carefully to get acceptable quantitative agreement.

Perhaps most importantly, the successes of the hydro-models show that the majority of the Lyα absorbers are consistent with being low density condensations formed by gravitational collapse of the intergalactic medium. In this picture the gas seen as Lyα forest *is* the original reservoir of matter from which galaxies are condensing.

Conversely, the success in reproducing the observations of the Lyα forest provides perhaps the best observational evidence for *hierarchical* structure formation we have to date.

Cosmology with the Lyα forest

Most of the intergalactic gas observed at high redshift has experienced only mild gravitational collapse. A typical region of space has undergone little processing other than heating by photoionization and compression, in competition with adiabatic cooling by the Hubble expansion. It turns out that the weaker, unsaturated absorption lines ($N < 10^{14}$ cm^{-2}) are still on or near the linear regime of gravitational collapse. Even for lines close to saturation the overdensities with respect to the mean density of the universe are less than about a factor $\delta \sim 10$–15. On spatial scales on the order of one Mpc and larger where the thermal pressure of the gas is not important the intergalactic medium traces the underlying mass distribution much more closely than the stellar light of galaxies observed in emission. This fact and the simple astrophysics involved should make the intergalactic medium an ideal cosmological laboratory.

Recently there have been various attempts at tapping the cosmological potential of the Lyα forest. The link between the observable appearance of the Lyα forest and the various cosmological input parameters can be described approximately by the Gunn–Peterson relation for the H I optical depth τ, generalized to include an inhomogenous density and velocity field. As long as the gas is highly ionized and in photoionization equilibrium (not necessarily thermal equilibrium), and the gas is unshocked, the optical depth for Lyα absorption at redshift z is proportional to

$$\tau(z) \propto \frac{(\Omega_b h_{50}^2)^2}{\Gamma} \left(\frac{H(0)}{h_{50} H(z)}\right) T^{-0.7} \times \left(\frac{\rho}{\overline{\rho}}\right)^{\alpha} (1+z)^6 \left(1 + \frac{\mathrm{d}v_{\mathrm{pec}}}{H(z)\mathrm{d}r}\right)^{-1} \quad (9)$$

This equation relates the optical depth to the mean baryonic density (in gas) in units of the critical density, Ω_b, the Hubble constant at redshift z, $H(z)$, the average gas temperature T, the proper baryon density ρ, the photoionization rate Γ, and the gradient of the local peculiar velocity $\mathrm{d}v_{\mathrm{pec}}/\mathrm{d}r$ along the line-of-sight. A further convolution with a Voigt profile is necessary to include the proper thermal velocity broadening. The exponent α ($\alpha = 2$ for an isothermal gas) takes account of the fact that in denser regions of the universe the gas is typically warmer because it is more effectively heated by photoionization, but α also depends on the reionization history of the gas and the amount of adiabatic expansion/compression. Values of $\alpha \approx 1.6$–1.8 are given in the literature.

To turn this relation into a complete description of the observed Lyα forest, cosmology has to predict the cosmic density and velocity fields, the fraction of the closure density in the form of gas, the equation of state of the intergalactic medium, and the ionizing radiation field. Measuring cosmological parameters then boils down to creating Lyα forest spectra according to a given cosmological prescription, and iterating with varying input parameters until good agreement between observed and predicted properties is obtained. At the time of writing these techniques are just beginning to be explored, but it has become obvious that there are at least three main areas to benefit from such studies.

From equation (9) the optical depth is directly proportional to the ratio $(\Omega_b h_{50}^2)^2/\Gamma$. In other words, a higher density produces stronger absorption, but a higher photoionization rate reduces the neutral fraction of the gas and decreases the absorption. With a suitable hypothesis for or an independent measurement of the photoionization rate Γ (e.g. estimating the density of ionizing photons from QSO surveys) the baryon density Ω_b of the universe may be constrained.

The 'equation of state' of the intergalactic medium, as the statistical distribution of the gas volume elements in temperature-density ($T - \rho$) space has become known, has a more complex and subtle influence on the optical depth distribution: a change in the slope $\mathrm{d}T/\mathrm{d}\rho$ changes the value of the temperature of a given volume element and thus the recombination coefficient $\alpha(T)$, the neutral fraction and column density, and the absorption line width. It appears that the distribution of the Lyα forest lines in Doppler parameter–column density ($b - N$) space is a distorted map of the density–temperature relation, and can be used to constrain the latter. $T(\rho)$, in turn, contains information on the epoch of reionization and reheating and on the sources of ionizing radiation.

Finally, it is clear that the fluctuation properties (amplitude, spatial correlation) of the density ρ and peculiar velocity v_{pec} fields propagate through to the optical depth (cf equation (9)). At least in the regime where the Lyα lines are strong enough to be detected but not yet too strong to be saturated, the spectra provide a record of the initial conditions of gravitational structure

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

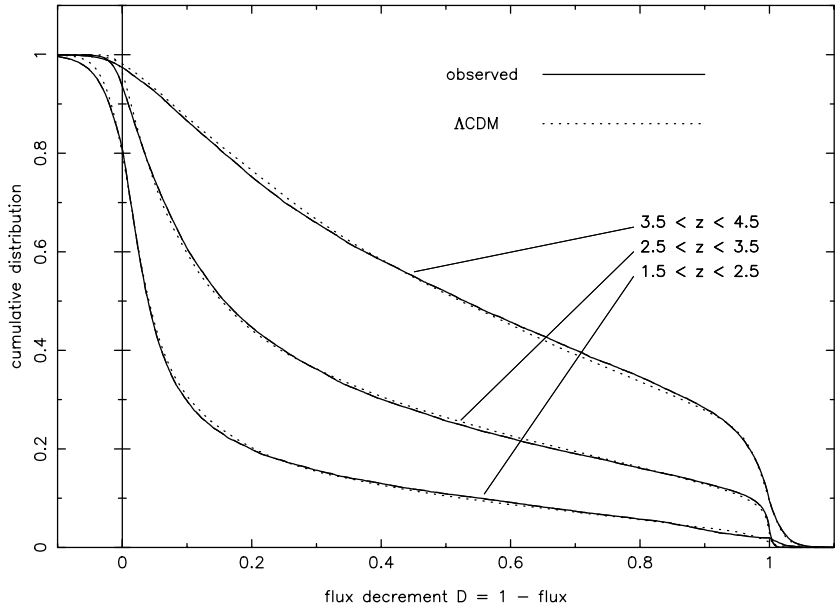


Figure 4. Observed and simulated statistical distributions of the absorbed flux level in the Lyα forest, for three different mean redshifts. The solid lines show the distribution for a sample of QSOs observed with the Keck telescope, the dotted lines are from a simulation of a ΛCDM universe (performed by Renyue Cen *et al* at Princeton University). The good agreement in the shape of the distributions lends observational support to the Lyα forest being a by-product of hierarchical structure formation.

formation. In principle the Lyα forest spectra can be inverted to provide a measurement of the power spectrum shape and amplitude of the initial fluctuations.

The helium Lyα forest

We conclude with a glance at one of the most intriguing new topics of research, the Lyα forest absorption due to singly ionized helium (He II). The He II forest has only recently become acessible as the transition is in the far ultraviolet (304 Å) and even if redshifted out to $z \sim 3$ still needs to be observed with a UV spectrograph from space. In addition, it is very difficult to find a QSO with its UV continuum intact, as the Lyman limit continuum absorption from higher redshift systems tends to obliterate the far UV in most QSOs. Nevertheless, there are potentially big scientific rewards to be gained: (1) since He II and H I have different ionization potentials, looking at the ratio of the column densities from their respective Lyα forests one can constrain the shape of the UV spectrum ionizing the two species. (2) Helium and hydrogen atomic masses differ sufficiently that a comparison of the Doppler parameters of both can measure the amount of non-thermal line broadening. (3) The first observations of He II forests are showing that He II ionization appears to be patchy and not fully developed by redshift 3–4. This observation could help us to understand how and when the universe was first ionized.

There are many other topics related to the rapidly growing field of QSO absorption lines for which we refer the interested reader to the literature given below (see also QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES; INTRINSIC AGN ABSORPTION LINES). These include metal absorption systems in general, and the correspondence between absorption systems and various galactic or interstellar environments, as seen in local galaxies. Damped Lyα systems, the absorbers most relevant to high redshift galaxy formation, are treated in the separate article LYMAN ALPHA ABSORPTION: THE DAMPED SYSTEMS.

Bibliography

Reviews of the subject at a level accessible to students, and detailed references to the literature are given in

Weymann R J, Carswell R F and Smith M G 1981 Absorption lines in the spectra of quasistellar objects *Ann. Rev. Astron. Astrophys.* **19** 41

Rauch M 1998 The Lyman Alpha Forest in the spectra of QSOs *Ann. Rev. Astron. Astrophys.* **36** 267

The scientific literature on the subject begins with the first measurement of the Gunn–Peterson effect, published as

Gunn J E and Peterson B A 1965 On the density of neutral hydrogen in intergalactic space *Astrophys. J.* **142** 1633

A number of conference proceedings give very useful overviews at a somewhat more technical level than the reviews; they are indispensible as guides to the many topics related to QSO absorption lines which could not be treated here:

Blades J C, Turnshek D A and Norman C (eds) 1988 *QSO Absorption Lines: Probing the Universe (Proc. QSO*

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Absorption Line Meeting, Baltimore 1987) (Cambridge: Cambridge University Press)

Meylan G (ed) 1995 *QSO Absorption Lines (Proc. ESO Workshop, November 21–24, 1994)* (Berlin: Springer)

Petitjean P and Charlot S 1997 *Structure and Evolution of the Intergalactic Medium from QSO Absorption Lines (Proc. 13th IAP Astrophysics Colloquium)* (Paris: Editions Frontières)

Recent work on cosmological applications of the Lyα forest is discussed in articles by Weinberg *et al*, Hui, Haehnelt and Nusser in

Banday A J, Sheth R K and Da Costa L N (eds) 1999 *Evolution of Large Scale Structure from Recombination to Garching (Proc. MPA-Garching Cosmology Conference)* (Enschede: PrintPartners Ipskamp)

Michael Rauch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lynx

(the Lynx; abbrev. Lyn, gen. Lyncis; area 545 sq. deg.) A northern constellation which lies between Ursa Major and Auriga, and culminates at midnight in late January. It was introduced by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* of 1687. An inconspicuous constellation, Hevelius (who distrusted telescopic sights for stellar observations) called it Lynx because, he wrote, one would have to be lynx-eyed to see it!

The brightest stars in Lynx are α Lyncis, magnitude 3.1, 38 Lyncis, a binary with bluish-white (A3 and A4) components, magnitudes 3.9 and 6.6, separation 2.7″, the latter of which has an unseen companion, and HR 3579, magnitude 4.0. Another interesting multiple star system is 12 Lyncis, which consists of a close pair of bluish-white (A2) components, magnitudes 5.4 and 6.0, separation 1.7″, period about 700 years, and a third component, magnitude 7.1, separation 8″. Other interesting objects include NGC 2419, a tenth-magnitude globular cluster remarkable for its remoteness from the halo of globular clusters surrounding the Galaxy (its being further away than the Magellanic Clouds), and NGC 2683, a tenth-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyot, Bernard Ferdinand (1897–1952)

Astronomer, born in Paris, worked at the Paris Observatory at Meudon, invented the coronagraph, a device that creates an artificial eclipse in a telescope and allows the Sun's corona to be observed at any time. With it he recorded the spectrum of the corona in new detail, and the first time-lapse pictures of solar prominences. Invented the *Lyot filter*, a birefringent interference filter made of alternating layers of polaroid and calcite plates. With a photoelectric polarimeter of unprecedented accuracy, he also pioneered the astronomical study of the polarization of light reflected from the surface of the Moon and of the planets, finding that the lunar surface behaves like volcanic dust and that Mars has sandstorms. Lyot's name is commemorated in the *Télescope National Bernard Lyot* of the Pic du Midi Observatory in the Pyrenees.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyra

(the Lyre; abbrev. Lyr, gen. Lyrae; area 286 sq. deg.) A northern constellation which lies between Hercules and Cygnus, and culminates at midnight in early July. It is an ancient constellation pattern, which was associated with an eagle or vulture in the Indian subcontinent and Arab countries, and with the mythical lyre invented by Hermes and given by Apollo to Orpheus in ancient Greece. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small but conspicuous constellation, Lyra is easily recognized by virtue of α Lyrae (Vega), which at magnitude 0.0 is the fifth brightest star in the sky and from northern temperate latitudes is almost immediately overhead on summer evenings, when together with α Aquilae (Altair) and α Cygni (Deneb) it makes up the asterism of the Summer Triangle. Other bright stars include γ Lyrae (Sulafat), magnitude 3.3, and β Lyrae (Sheliak), a multiple system consisting of an eclipsing binary (range 3.25–4.36, period 12.91 days) and two widely separated fainter stars, magnitudes 7.2 and 9.9v, separations 46″ and 86″, the brighter component of which also has an unseen companion. There are no other stars brighter than fourth magnitude. Other interesting stars include ε Lyrae, the so-called 'Double Double', which consists of two close binaries, ε^1, with bluish-white (A4 and F1) components, magnitudes 5.1 and 6.0, separation 2.5″, and ε^2, also with bluish-white (A8 and F0) components, magnitudes 5.1 and 5.4, separation 2.3″, themselves separated by 209″, and RR Lyrae, the brightest known member of a class of pulsating variables (range 7.06–8.12, period 0.57 day).

Other interesting objects in Lyra include M57 (NGC 6720, the Ring Nebula), a ninth-magnitude planetary nebula, and M56 (NGC 6779), an eighth-magnitude globular cluster.

The Lyrid meteor shower (sometimes called the April Lyrids to distinguish it from a lesser shower that occurs in June) appears to radiate from a point to the south-west of Vega.

See also: Epsilon Lyrae, Lyrids, Ring Nebula, RR Lyrae stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lyrids

A meteor shower that takes place in April, sometimes known as the *April Lyrids*. The radiant lies in the constellation Lyra, close to the border with Hercules, near the star Vega. There are occasional outbursts, most recently in 1982, where the usual peak zenithal hourly rate of 10–15 increases to 100 or more. The parent comet, C/1861 G1 Thatcher, has the longest period (415 years) of any comet known to be associated with a meteor shower. Its high orbital inclination of 80° means that the meteor stream is little affected by planetary perturbations, and Lyrid activity therefore shows a long-term constancy.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mästlin, Michael (1550–1631)

Born in Göppingen, Germany, taught by, and an assistant to, PETER APIAN, became professor of astronomy at Tubingen, where he taught the Copernican as well as the Ptolemaic system. He taught KEPLER, and helped him to publish his first work, *Mysterium Cosmographicum* (1596). He made his own astronomical instruments. He was one of the first three or four astronomers to see in Cassiopeia what is now called TYCHO's supernova, the new star of 1572. He determined that it had no motion over the night by holding a thread between two stars, with the nova on the same line, and was not therefore a comet.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Möbius, August Ferdinand (1790–1868)

Mathematician, born in Schulpforta, Saxony (now Germany), studied astronomy under GAUSS, and wrote his doctoral thesis on 'The occultation of fixed stars by planets'. Became professor of astronomy at Leipzig, rebuilt its observatory, and wrote on astronomy and celestial mechanics, this work being overshadowed by his imaginative mathematical work on mapmaking and analytical geometry (eponymous mathematical items include a Möbius net, a Möbius function, and the one-sided surface called the Möbius strip).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mädler, Johann Heinrich von (1794–1874)

German astronomer, observed the Moon and drew with WILHELM BEER a map of the Moon (1837). Became director of the Dorpat Observatory (Estonia).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mégantic Astronomical Observatory

The Mont Mégantic Observatory is operated by the Université de Montréal. It is located 250 km east of Montreal, on the summit of the highest mountain in Quebec, at an altitude of 1111 m. Its 1.6 m diameter Ritchey–Chrétien telescope is used for observations at visible and near-infrared wavelengths.

The telescope is used for imaging, spectroscopy and polarimetry at visible wavelengths. An infrared camera is also available, though it is often 'borrowed' by astronomers who observe with the CANADA–FRANCE–HAWAII TELESCOPE. There is also a meteorological station on the site.

For further information see
http://www.astro.umontreal.ca/omm/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M1

M1 is the famous CRAB NEBULA in Taurus, one of the most celebrated objects in the sky, the remains of a supernova seen to explode in AD 1054. Messier stumbled across it in 1758 while following a comet that passed through Taurus; he described the nebula as 'elongated like the flame of a candle'. However, it had already been discovered by an English astronomer, John Bevis, in 1731. M1 lies just over a degree northwest of Zeta Tauri and is an eighth-magnitude patch measuring about 8 by 6 arc minutes in apparent size; its true dimensions are about 11 by 7.5 light-years. Its popular name comes from a description by Lord Rosse, who observed filaments extending from it like the pincers of a crab. At its center is a sixteenth-magnitude neutron star, the Crab pulsar; this is the compressed core of the star that exploded. M1 lies about 6500 light-years away.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M10

M10 is a globular cluster in the constellation Ophiuchus (the star 30 Ophiuchi is just 1 degree to the east-southeast). It was discovered by Messier who did not resolve any stars because it is fairly compressed. It has a diameter of 70 light-years but its bright core is only half that size. Its visual brighness is 6.7 mag and it displays a slight flattening (ellipticity E1). The distance to the Sun is about 13 400 light-years (with some sources quoting values up to 16 300 light-years).

M10 was investigated for variable stars by Baade but only contains three to four of them and has hence not been investigated in great detail by astronomers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M100

M100 is one of the largest spiral galaxies in the Virgo Cluster, its true size being similar to that of the Andromeda Galaxy, M31. It lies in Coma Berenices, completing a triangle around the star 6 Comae Berenices with M98 and M99, all of which were discovered by Mechain on the same fruitful night in March 1781. Although M100 is cataloged as of ninth magnitude, roughly twice as bright as M98 and M99, it is no easier to see visually than those galaxies because its light is spread over a greater area, resulting in a low surface brightness. It is classified as type Sc. Lord Rosse was the first to see its spiral arms, in 1850. Allan Sandage has concluded that the two main arms are about 3 000 light-years thick, twice the thickness of the arms in our own Galaxy. An asymmetry in the shape of the spiral arms seems to be to gravitational disturbance by a nearby dwarf elliptical galaxy, NGC 4322, in a smaller-scale version of the interaction between the Whirlpool Galaxy, M51, and its companion. Four supernovae have been seen here, in 1901, 1914, 1959 and 1979. The 1979 event was unusually bright, reaching a peak of eleventh magnitude; the following year, radio emission from its remnant was picked up by astronomers using the Very Large Array, the first time an extragalactic supernova had been detected at radio wavelengths. Infrared images of M100 taken in 1995 showed the existence of a short bar across the galaxy's center, ending in two short arms that curve in the opposite direction to the outer spiral arms. At the tips of these arms are areas of star formation, part of ring of star formation about 1500 light-years from the galaxy's core that shows up on images taken at ultraviolet wavelengths. This finding is consistent with photographs which show that the color of the galaxy's core is similar to that of the spiral arms,suggesting recent star formation in the region which may be another consequence of the interaction with the nearby dwarf. M100 has figured prominently in attempts to measure the distance scale of the universe, on account of its eminence in the Virgo cluster. Observations of 20 Cepheid variables in its spiral arms with the Hubble Space Telescope led to an accurate determination of its distance (and hence that of the Virgo Cluster) at 56 ± 6 million light-years, announced in 1994.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M101

M101 is a grand spiral galaxy in Ursa Major, located 5.5 degrees north of Eta Ursae Majoris, the star at the end of the handle of the Plough (or Big Dipper). It was discovered in March 1781 by Mechain and added by Messier to his list at the last minute without having checked it. M101 is classified as of type Sc, meaning that its arms are loosely wound, and they show beautifully on long-exposure photographs. Surprisingly, the galaxy can be detected as a pale, rounded disk with binoculars but no details will be visible without a telescope. The brightest clouds of hydrogen in the spiral arms have their own NGC numbers, notably NGC 5461 and 5462 in the longest arm on the galaxy's western side. Three supernovae have been detected, in 1909, 1951 and 1970. M101 is surrounded by a family of about eight small satellite galaxies. Ultraviolet observations in 1981 showed ripples of star formation extending eastwards from M101 towards one of these satellites, the irregular dwarf galaxy NGC 5477, another example of interactions affecting a galaxy's size and structure. From observation of Cepheid variables, M101 is around 25 million light-years away.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M102

M102 is a duplicate observation of M101, an error due to Mechain. He originally reported it in 1781 to Messier who included it on trust in his third and final list without checking, as he also did with M101 and M103. However, Mechain admitted his mistake in a letter to Bernoulli two years later, and so M102 is now regarded as non-existent.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M103

M103 is an open cluster in Cassiopeia, about 1 degree northeast of Delta Cassiopeiae, discovered by Mechain in 1781. As with M101 and M102 Messier did not check it before inclusion; had he done so, it seems likely that he would also have spotted the larger and far more comet-like open cluster NGC 663 1.5 degrees to the northeast. Both these clusters are members of the Cassiopeia OB8 association. Admiral Smyth described M103 as a fan-shaped group, and estimated it to have about 20 members easily detectable in his telescope, the four brightest ranging from seventh to ninth magnitude. The brightest of these, which is actually a neat double divisible in small apertures, is now known to lie in the foreground and hence is not a true member at all. According to modern measurements, the cluster lies just over 8000 light-years away and has an age of around 30 million years. M103 was the last object in Messier's third and final list of objects; the subsequent objects were added later by others.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M104

M104 is a spiral galaxy on the border of Virgo and Corvus, popularly termed the SOMBRERO GALAXY because of its resemblance to a wide-brimmed Mexican hat. M104 is a spiral galaxy of type Sa (or possibly intermediate between Sa and Sb), seen only about 6 degrees from edge-on. It was discovered by Mechain, and added by hand to Messier's own copy of the printed version of his catalog. In 1921 the French astronomer C Flammarion suggested adding it officially to the Messier list as M104, and the suggestion has been generally accepted. Photographs show a large central bulge surrounded by a swarm of globular clusters and crossed by a dark lane of dust in the plane of the galaxy; this dark lane seems to have been first detected by William Herschel. Although lying in Virgo, the Sombrero is probably not a member of the Virgo Cluster but is somewhat closer to us, about 35 million light-years away.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M105

M105 is an elliptical galaxy in Leo about 1 degree to the north of M95 and M96 with which it forms a triangle; it is also physically associated with that pair, all three being part of a sparse cluster known as the Leo group of galaxies. It was discovered in 1781 by Mechain and was mentioned in his letter to J Bernoulli in which he admitted the duplication of M101 and M102. Although Messier apparently never saw this object for himself, in 1947 it was identified as NGC 3379 by Helen Sawyer Hogg who proposed its adoption as M105. It is classified as type E1, meaning it is almost perfectly round. M105 has two smaller and fainter companions: NGC 3384, an elliptical, and NGC 3389, a spiral. Radio observations at 21 cm wavelength published in 1989 showed the existence of a ring of hydrogen gas over 600 000 light-years wide orbiting M105 and NGC 3384. There are no signs of any stars or even dust in this ring, suggesting that it is composed entirely of primordial hydrogen left over from the Big Bang that never became part of any galaxy. In addition, nearby M96 seems to be interacting with the ring and capturing some of the gas.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M106

M106 is a spiral galaxy in Canes Venatici just over a degree and a half south of the fifth-magnitude star 3 Canum Venaticorum. At eighth magnitude it is fairly easily within reach of small telescopes. Mechain discovered it in 1781 and mentioned it in the same letter to Bernoulli as his discovery of the objects now known as M105 and M107. In 1947 it was identified as NGC 4258 by Helen Sawyer Hogg, who proposed its adoption as M106. It is classed as type Sb but shows signs of distortion in its spiral arms. Since it is tilted steeply to our line of sight, the full extent of the disturbance is difficult to judge. From radio observations of the velocities of water masers orbiting the galaxy's nucleus, astronomers have deduced the existence of a central body less than a light year across with a mass of around 36 million Suns—in other words, a probable massive black hole. One supernova, 1981K, has so far been seen to erupt in M106. Estimates of the galaxy's distance range from about 23 to nearly 30 million light-years, depending on the technique used.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M107

M107 is a globular cluster in Ophiuchus, about 2.75 degrees south of Zeta Ophiuchi. Like M105 and M106 it was discovered by Mechain but was apparently never seen by Messier. In 1947 Helen Sawyer Hogg identified it as being NGC 6171 and suggested it as an addition to the Messier catalog with the designation M107. Visually it appears relatively sparse by comparison with other globulars, covering about one-third the apparent diameter of the full Moon but appearing only of eighth magnitude. Its distance is around 19 000 light years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M108

M108 is a spiral galaxy in Ursa Major, located one-third of the way from M97 (the Owl Nebula) to Beta Ursae Majoris. Messier referred to it in his catalog in a note on M97, saying that it lay nearby, but since its position had not been determined he never included it as a separate entry. Like so many objects, it seems to have been first pointed out to him by Mechain. In 1960 the American historian O Gingerich identified this object as NGC 3556 and proposed calling it M108. It is usually classified as type Sc but G de Vaucouleurs regarded it as a barred spiral. In fact its true shape is difficult to ascertain since it is seen nearly edge-on and has a patchy appearance caused by dark clouds of dust. Radio observations, made with the Very Large Array and announced in 1997, have revealed the existence of an invisible shell of hydrogen gas 21 000 light-years across extending from the eastern side of the galaxy. The origin of this shell remains a puzzle; one possibility is that it is the shock wave from the eruption of a huge cluster of supernovae millions of years ago. In recent times it has been the site of only one observed supernova, 1969B.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M109

M109 is a barred spiral galaxy in Ursa Major, lying 0.66 degrees east of Gamma Ursae Majoris in the bowl of the Plough (or Big Dipper). As with M108, it was mentioned by Messier in his note on M97 but was not identified with a known galaxy until 1960, when O Gingerich concluded that it was the same as NGC 3992. M109 is classified as midway in type between SBb and SBc. Much of its central region shows an emission-line spectrum due to ionized gas, evidence of unusual activity in its nucleus and placing it among the type of galaxies known as 'liners'. M109 has been the site of one observed supernova, 1956A.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M11

M11 is an open or galactic cluster in Scutum (2 degrees west and a bit south of η Scuti). It has a very interesting arrow-shaped outline in its densest part and a V-shaped group of stars ahead of the arrow. This prompted Admiral Smyth to describe it as a 'flight of wild ducks' and it is therefore sometimes known as the 'wild duck' cluster. The German astronomer Gottfried Kirch first discovered the cluster in 1681.

M11 contains 2900 stars within a diameter of about 15 light-years. For an open cluster it is very rich and compact, with stars in the central regions separated on average by just 1 light-year. Five hundred of these stars are brighter than 14 mag and the overall brightness of the cluster is 6.3 mag. The distance to the Sun is about 6000 light-years, i.e. it is probably not a condensation in the Scutum star cloud against which it appears but closer to us. The cluster is sometimes thought to be comparatively young at 250 million years but some sources also quote double that value. Support for a higher age comes from the relative abundance of yellow and red giants of absolute magnitude of -1 (i.e. comparatively faint old stars). In fact, M11 has more yellow giants than most open clusters and is similar to M67 in that respect.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M110

M110 is the second of two small elliptical galaxies near the Andromeda Galaxy, both of which can be seen through small telescopes. Messier cataloged the first companion of M31 as M32, but strangely never included the second on his list, even though he showed M31 with both companions in a drawing in 1807 and claimed to have discovered it in 1773. Until 1966 it was known solely as NGC 205, when the English amateur astronomer K Glyn Jones proposed adding it as M110, the final Messier object. M110 lies over 1 degree northwest of M31 and although it is larger than M32 it is more difficult to see, having a lower surface brightness. It is classified as type E5 or E6, meaning that it is considerably elongated in shape. Two unusual aspects of M110 are the existence of dust clouds and clusters of young stars near its center, not normally encountered in elliptical galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M12

M12 is a globular cluster in Ophiuchus close to M10 but slightly larger, fainter and much less concentrated towards the center. It was discovered by Charles Messier, but William Herschel was the first to resolve it into stars in 1783.

The cluster has a diameter of about 100 light-years and is at a distance of about 19 000 light-years from the Sun. It is a very loose cluster and was at one point even thought to be an intermediate type between globular and galactic cluster (this was before it was discovered that the two types of cluster are completely different in age and stellar population). It has a visual brightness of 6.7 mag and a diameter of 100 light-years. Allan Sandage investigated the cluster for variables and found only 13.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M13

M13 is a globular cluster in Hercules (2.5 degrees south of η Herculis), the 'Great Hercules Globular Cluster' discovered by Edmund Halley in 1714 who noted that 'it shows itself to the naked eye when the sky is serene and the Moon absent'. It was first resolved by William Herschel and has since been the target of much professional and amateur interest, mainly because it is very bright and easily observable from the northern hemisphere.

M13 is a fairly open cluster, almost perfectly circular in outline, with a diameter of 150 light-years, a visual brightness of 5.8 mag and over half a million stars. At its center the stars are 500 times more concentrated than they are in the solar neighborhood. Its age was estimated by Arp to be 14 billion years (1962). The high latitude of the cluster leads to a large major axis of its orbit which means that it spent most of its long life at a great distance from the disruptive tidal effects of the Galactic Center—hence the great concentration of stars and its almost perfect shape. Its distance to the Sun is 22 800 light-years as measured from its variables (slightly higher from radial velocity measurements). Like M3 it contains one hot, young, blue star (Barnard no 29) whose presence in this old cluster cannot easily be explained. It also contains 15 variable stars and four apparently star-poor regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M14

M14 is a galactic cluster in Ophiuchus, discovered by Messier and first resolved into component stars by William Herschel in 1783. It has a diameter of 55 light-years and lacks a dense central concentration. The outline is slightly flattened (ellipticity type E1) and its visual brightness is 7.6 mag. The distance to the Sun is approximately 27 400 light-years.

M14 is relatively rich in variable stars, 70 having been discovered so far. Of particular interest is a nova (16 mag) that appeared in this cluster between 21 and 28 June 1938 but was not discovered until 1964 when Amelia Wehlau of the University of Western Ontario found it during a search of photographic plates taken by Helen Sawyer Hogg between 1932 and 1963. It was only the second nova to appear in a globular cluster (the first being in M80 in 1860) and the first ever to be photographed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M15

M15 is a globular cluster in Pegasus (close to ϵ Pegasi), first discovered by Jean Dominique Maraldi on 7 September 1746. It was resolved into stars by William Herschel.

This is another very rich and old globular cluster. It has a diameter of 88 light-years and its superdense core is undergoing collapse. This often features in the dynamic evolution of globular clusters: of the 147 in the Milky Way, 21 have a collapsed core (such as, e.g., M30 and M70) and there are eight more candidates. In the case of M15 it is not entirely clear whether the core has collapsed simply under its own gravity or whether there is a supermassive object at the center resembling the objects thought to be at the centers of galactic nuclei. The Hubble Space Telescope has therefore recently taken pictures of the dense core. By measuring the velocities of the central stars it will be possible to resolve this question.

The age of the cluster was estimated at 12 billion years. It has a visual brightness of 6.2 mag and is at a distance of around 32 000 light-years.

M15 is the third richest cluster in terms of variables: 112 have been discovered so far, one of which is a classical (type II) Cepheid. It also contains nine pulsars, one of which seems a component of a neutron star binary. Systems of neutron-plus-neutron star binaries are particularly interesting because they exhibit strongly general relativistic effects, such as significant perihelion shifts, effects on passing light and gravitational waves which carry away rotational energy and lead to a slowing-down of the pulsar frequency as well as of the orbital period. It is also the first cluster in which a planetary nebula could be identified, and in 1988 an x-ray source was found to be a normal-plus-neutron star binary system with a period of 9 h.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M16

M16 is an open cluster in the constellation Serpens, associated with the EAGLE NEBULA (1 degree north and 2.5 degrees west of γ Scuti). It was discovered by de Chéseaux in 1746 as a 'cluster of stars' and sits on the next inner spiral arm of the galaxy away from us next to M17.

M16 is in many ways similar to the Lagoon Nebula M8. The cluster formed from the gas of the Eagle Nebula and its bright young stars now light it up. The nebula exhibits ongoing star formation in the dark 'elephant trunks' visible to the north. It also contains the black globules that are collapsing protostellar clouds (mainly to the southeast). The Hubble Space Telescope has taken many detailed pictures of this particularly beautiful object.

The star cluster is one of the intrinsically most luminous with an absolute magnitude of -8.21. It has a diameter of 7 arc min at a distance of 7000 light-years and a visual brightness of 6.4 magnitudes. Its age has been estimated at 5.5 million years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M17

M17 is a diffuse nebula in Sagittarius variously known as the Omega, Swan, Horseshoe or Lobster Nebula. It was discovered by de Chéseaux in 1746 and is a neighbor to M16.

The diameter of its brightest part is about 15 light-years but the low-luminosity and dark material extends to a size of at least 40 light-years. Like many nebulae, it consists of bright filaments and dark, obscuring matter and star formation has either only recently stopped or is still proceeding in its center. A cluster of about 35 bright but obscured stars is embedded in it and the nebula shines by re-emitting the light absorbed from these stars. The mass of the brightest gas has been estimated to be about 800 solar masses (much more than in the otherwise comparable Orion Nebula, M42). The main difference between this nebula and M8 and M16 is the difference in its absorption regions: M17 contains a large amount of dark, obscuring material, more so than either of the other two nebulae.

Its distance to the Sun is between 5 and 6000 light-years. The nebula is a radio source in 9.4 cm and has been shown to have a 29%-polarized infrared emission. This might indicate the existence of 'bipolar nebulae' or lobes of scattering material in the polar regions. A twin-lobed radio source extending from the central infrared cluster has also been detected.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M18

M18 is an open cluster in Sagittarius discovered by Messier as a 'cluster of small stars'. It lies between M17 and M24.

The cluster contains about 20 members and has an angular diameter of 9 arcmin. Its visual magnitude is 7.5 which makes it one of the looser and poorer clusters. Correspondingly, it has not received much attention from scientists to date. Its distance from the Sun has been estimated at 4900 light-years but that value is disputed with some sources quoting up to 6000 or less than 4000 light-years. M18 is generally considered to be quite young (32 million years). Its hottest stars are of type B3 (i.e. very hot and young) but it contains bright yellow and orange as well as blue stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M19

M19 is a globular cluster in Ophiuchus, discovered by Messier but first resolved by William Herschel in 1784. It can easily be located at 8 degrees due east of Antares and embedded in the Milky Way.

It is at a distance of 27 000 light years (4600 light-years from the Galactic Center), has a diameter of 25 light-years and a visual brightness of 6.8 mag. M19 is the most oblate of the known globular clusters with an ellipticity E3–E4.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M2

M2 is a bright globular cluster in Aquarius forming an almost right-angled triangle to the north with α and β Aquarii. It was first mentioned by Jean Dominique Maraldi in 1746 but independently rediscovered by Charles Messier who described it as a nebula without stars. William Herschel was the first to resolve the stars. M2 is a rich and compact globular cluster with 150 000 stars within a diameter of about 150 light-years and with a slight ellipticity (type E1, meaning it is almost perfectly round). Its visual brightness is 6.5 mag and its distance from the Sun is 36 000 light-years (i.e. well beyond the Galactic Center)—but there is some doubt about this number, with some sources setting the distance at around 50 000 light-years. The age of the cluster has been determined at 13 billion years.

Compared with other globular clusters, M2 has a relatively low number of variable stars: so far, 21 variables have been detected, most of which are RR Lyrae stars with periods of less than a day, but there are also three classical (type II) Cepheids with periods between 15 and 19 days and one RV Tauri star with a period of 67 days. The cluster's brightest stars are red and yellow giants of 14.1 mag (absolute magnitude −3.0).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M20

M20 is a diffuse nebula in Sagittarius known as the 'TRIFID NEBULA' because of its three-lobed appearance. It is situated roughly 2 degrees northwest of M8. Messier described a 'cluster of stars', William Herschel assigned different numbers to different parts of the object and John Herschel was the first to use the term 'trifid'.

M20 is a typical emission nebula where young hot stars at the center excite the gas of the surrounding nebula which then de-excites by emitting a reddish glow. It is itself surrounded by a blue reflection nebula particularly conspicuous to the north. The three-part division is due to dark obscuring filaments between the bright nebulae. It has a diameter of 30 light-years and a distance of 2200–7600 light-years from the Sun. Its visual brightness is 9 to 6.8 mag—the range of values deriving from the fact that the overall brightness is hard to detect in this object whose central region is dominated by a few very bright stars. The brightest of these are a triple system consisting of three very hot and four fainter members. Like M8 and M42, M20 is a radio source at 9.4 cm. This relatively short-wavelength radiation indicates that the source must be thermal emission and the temperature of the nebula has been calculated from this to be about 10 000 K.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M21

M21 is a young open cluster in Sagittarius, close to but not associated with the Trifid Nebula (M20). It was discovered in 1764 by Messier who described it as a 'star cluster near M20'.

The cluster has about 100 probable member stars, some of which are early B-type stars and the reddening through interstellar dust is comparatively low. For that reason, M21 lends itself to star formation studies and to the calibration of distance indicators (not having to worry about the reddening is always an advantage for that). However, research interest in this cluster has been quite low and it is one of the few young open clusters still lacking a detailed study. Its age is commonly thought to be about 8 million years and its distance is set at 4300 light-years. There is some controversy associated with this number, however, and it is not entirely clear whether M20 or M21 is closer to us.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M22

M22 is a rich, old globular cluster in Sagittarius, very bright (5.1 mag) and fairly close (about 10 000 light-years) but only 1 degree out of the ecliptic. Its vincinity to the ecliptic plane is both an advantage and a disadvantage: it often appears in conjunction with planets and was discovered in 1665 (the first globular cluster recorded) by the German astronomer Abraham Ihle while he was observing Saturn. But since it is viewed along a line of sight passing through the galactic disk and bulge, the cluster is heavily reddened (i.e. the light from the cluster is shifted into the red end of the spectrum by scattering off dust grains) and so basic parameters like age and metallicity are difficult to determine exactly. In the case of M22, they are still somewhat uncertain, with the age being normally set at about 12 billion years.

Because of its brightness, the cluster has had a rich observational history, appearing in the lists of Halley, Lacaille, le Gentil and de Chéseaux before being included in Messier's catalog. William Herschel was the first to resolve it into its constituent stars and in 1959, M22 was one of the very first clusters for which a color–magnitude diagram was obtained. More recently, it has been investigated for variable and other exotic types of star: 32 variables and one planetary nebula have been found so far and since 1992 three possible CH stars (a very exotic type of binary, rare in clusters, see also M14) have been identified. The cluster is potentially very interesting because it is indicative of conditions in the very early protogalaxy, but because of the problems mentioned above (mainly the reddening) it has not yet been extensively studied.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M23

M23 is a very poor, loose open cluster in Sagittarius (one of Messier's original discoveries) with boundaries that are difficult to define. So far, 150 stars within a linear diameter of about 15 light-years have been shown to be members. The cluster is old for an open cluster (about 300 million years) and is thought to have spent a large fraction of its lifetime in the gas-poor regions between the spiral arms of our Galaxy. It can therefore be used to test theories about the effect of gas-rich and gas-poor environments on stellar evolution and a recent study on M23 indicates that the environment does not greatly influence the metallicity of stars after they have formed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M24

M24 was discovered by Messier in 1764 as a 'star cluster'. It is not, however, a real cluster but simply a patch of the Milky Way that is seen through a tunnel in the otherwise obscuring interstellar dust and therefore looks like a cluster. This patch is visible to the naked eye two-thirds of the way between γ Scuti and μ Sagittarii. It has an average magnitude of 4.6.

A dim open cluster can be detected inside the star cloud (NGC 6603) which in some cases is identified with the Messier object. This is wrong, however, since Messier clearly described the cloud rather than the dim (11 mag) cluster.

The interstellar medium has a tendency to accumulate in huge clouds about 25 light-years across. Small windows in these clouds are important for our understanding of galactic structure since they allow us to study material that would normally be obscured.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M25

M25 is an intermediate-age (90 million years) open cluster in Sagittarius that was discovered by de Chéseaux in 1746. It has a visual magnitude of 6.5 and is conspicuous even with normal binoculars. Because of its relatively small distance (about 2000 light-years) and the comparative ease with which it can be observed, it was one of the first used for studies of variable stars. In 1960, Allan Sandage identified its one Cepheid variable, U Sagittarii, and used it to calibrate the period–magnitude relationships for Cepheids. Since then, however, clusters richer in variables have been identified and interest in M25 has waned.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M26

M26 is an open cluster in the constellation Scutum, an apparent neighbor to M11. It was possibly seen by Le Gentil before 1750, but Messier discovered it independently in 1764 and noted that it was 'not distinguished in 3.5 foot telescope'.

Of the 90 confirmed members of this cluster, 25 are very bright but no extensive studies into the stellar population or the range of masses have been conducted so far. In fact, M26 is one of the least observed objects in the Messier catalog with its latest research paper published in 1940 (when its age was determined to be about 90 million years and its distance 5000 light-years). Its most distinguishing feature is an apparent ring of low stellar density around the center. In a region of diameter 3 arcmin (cluster diameter 15 arcmin) the density of stars is about 13% less than in the adjacent areas. This is now thought to be caused by dust extinction.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M27

M27 is a planetary nebula in Vulpecula also known as the 'DUMBBELL NEBULA' because of its shape. It was the first planetary nebula to be discovered (Messier, 12 July 1764), and even though it is neither the largest nor the brightest, it is the easiest to observe: larger planetary nebulae tend to be fainter and brighter ones tend to be small.

The nebula was created about 3000 to 4000 years ago when a giant star blew off its outer layers. It is therefore still expanding at a rate of 6.8 arc seconds per year and the central star (the remaining core of the giant) is a hot, blue subdwarf dwarf with a temperature of 85 000 K (spectral type O7) with possibly a faint yellow companion. This star emits light mostly in the non-visible parts of the spectrum, which is then absorbed and re-emitted by the nebula in the visible (predominantly in one green spectral line). The distance to the planetary nebula is difficult to determine (a common problem for these objects): values proposed range from 400 to 3500 light-years. But if the distance is taken to be 1200 light-years then the central star has an intrinsic luminosity of one-third of the solar luminosity and the nebula as a whole emits 100 solar luminosities entirely in the visible part of the spectrum. This gives it an apparent magnitude of 7.4 (i.e. it is not visible to the naked eye). The angular diameter of the nebula's brightest part is 6 arcmin and the dumbbell shape is due to us observing the nebula from its equatorial plane. If we saw it from one of the poles, it would look ring-shaped, more like M57.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M28

M28 is a rich, dense globular cluster in Sagittarius. It was discovered by Messier as a 'nebula containing no star' and first resolved into its component stars by William Herschel.

As for most of the dense globular clusters, M28 is interesting mainly because of its variable stars and because of its binary star systems. The evolution of binaries in rich environments has received a lot of attention recently, as they seem less common in these clusters than they are predicted to be. On the other hand, the more exotic varieties of binary systems (such as neutron star binaries) seem particularly frequent in the cores of rich clusters.

In 1987 a millisecond pulsar was discovered in M28 (the second to be found in a globular cluster after the discovery of the pulsar in M4). It rotates over 2000 times faster than the Crab pulsar (once every 11 milliseconds). Periods as short as this are thought to be produced by interaction with a binary companion, and in 1997 an x-ray burst in M28 was interpreted to be evidence for the neutron star slowly accreting material from that companion. Interest in M28 since then has mainly centered on its x-ray activity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M29

M29 is a very young, open cluster in Cygnus. It was discovered by Messier in 1764 and is particularly highly obscured by interstellar material. It has been estimated that in the line of sight to M29, the interstellar dust is 1000 times more dense than usual. Scattering of the cluster light leads to an overall reddening and dimming together with a high polarization. These effects make it difficult to gauge how much light and which color is really emitted by the member stars, and hence their age and distance is difficult to determine.

Within correspondingly large error bars, M29 is thought to be young (i.e. below 10 million years) and only about 4000 light-years away. It contains one variable of γ Cas type and several more suspected variables, and some investigation into its three binary systems and the hot stars in the cluster have been done, but these studies are all rendered difficult by the uncertainties associated with the fundamental parameters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M3

M3 is a very rich globular cluster in the constellation Canes Venatici. It was the first original discovery by Messier and probably the object that sparked his systematic search for nebulae and clusters. It has approximately 500 000 member stars within a diameter of 300 light-years and is with a visual brightness of 6.5 mag one of the brightest clusters in the northern hemisphere. Its distance to the Sun is about 30 600 light-years and its age has been estimated at 6.5 billion years.

M3 is a very well-studied globular cluster partly because it contains more variable stars than any other. So far, 212 have been counted and 186 periods determined. The brighter stars are of remarkably uniform magnitude and there is one very hot, young, blue star (discovered by Allan Sandage) whose presence in this old cluster continues to be a mystery (see also M13). Current theories propose that the young star might have been captured relatively recently by the gravitational attraction of the cluster as a whole.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M30

M30, like M15, M70 and possibly M62, is an old globular cluster (in Capricornus) whose central core has undergone gravitational collapse. It was discovered by Messier in 1764 but has only been very actively studied since about 1994 when it was discovered that the luminosity function of the cluster does not agree with the function shape derived from standard stellar evolution models. M30 has therefore become a test case for the study of the modification of stellar populations in dense collapsed regions.

Overall the light emitted by the cluster is much more blue than would be expected for an object this age. It has a higher frequency of blue stragglers (blue, hot main sequence stars which seem too young to be in such an old cluster) in the core than other globular clusters (48 have been identified so far), which supports the theory that they are formed by star or binary mergers. Additionally, the core is thought to be unusually poor in red giants, but so far no theory has been able to successfully explain this.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M31: The Old Stellar Populations

M31 (the Andromeda nebula) is a large highly inclined spiral galaxy about 1.4 times as luminous as our Galaxy. The Milky Way and M31 are the two dominant members of the Local Group of galaxies. The proximity of M31 (about 770 kpc from the Sun) allows its stellar populations to be studied in some detail with the Hubble Space Telescope and large ground-based telescopes. M31 contains all the usual ingredients of a spiral of intermediate (Sb) Hubble type: atomic, molecular and ionized gas, dust, dark matter and stars of all ages. For the study of stellar populations, however, M31's oldest stars are of particular interest, because they show some very significant differences, as a population, from the oldest stars of our Galaxy. Although we do not yet fully understand the reasons for these differences, they surely came about through differences in the processes by which these two large galaxies formed in the early universe.

The bulges of M31 and the Milky Way

In the Milky Way, the disk stars are all younger than about 10 Gyr. The oldest stars (ages up to about 14 Gyr) are found in the spheroidal components of the Galaxy, i.e. the bulge and the very diffuse stellar halo which extends out to at least 100 kpc from the galactic center. These old non-disk components, the bulge and the stellar halo, are often discussed together, although their origin may be quite different.

The structure of the spheroidal component was established by dynamical processes, like the merging of clumps of gas and dark matter and the dynamical instabilities of flat disks, which began at the same time as the oldest stars were forming. So its structure and its old stellar population are intimately related. It is interesting therefore to contrast the structure of the spheroidal components of M31 and the Galaxy before we go on to discuss their stellar content.

The bulge of M31 is relatively large, about 25% of the mass of the disk, while for our Galaxy the corresponding fraction is about 15%. Although these fractions are not so different, the bulges of M31 and our Galaxy are structurally different and typical of large and small bulges respectively. The M31 bulge has the spheroidal shape and the characteristic radial dependence of surface brightness that is seen in most large bulges. Its surface brightness $I(r)$ follows the empirical de Vaucouleurs law $\log I(r) \propto -r^{1/4}$, from a radius of 200 pc out to beyond 20 kpc. From dynamical theory, this $r^{1/4}$ law is associated with the violent relaxation process, in which rapid changes of the gravitational field statistically redistribute the orbital energies and angular momenta of the stars. The stellar system that emerges from this redistribution has the characteristic $r^{1/4}$ surface brightness distribution. Such rapid changes of the gravitational field come about, for example, as clumps of matter interact and merge to form the early bulge. The compression of gas in these same mergers is believed to set off the first bursts of star formation that produced the stars which we now see as the old stellar population of M31.

The bulge of the Milky Way is very different. Its smaller bulge is seen edge-on and shows a box-like shape, with a surface brightness distribution that is exponential with radius. This kind of structure is common for small edge-on bulges. From studying the motions of gas in our Galaxy and in other galaxies with bulges that appear boxy when seen edge-on, we know that these bulges would not appear round and symmetric when viewed from above. The boxy shape is dynamically associated with an asymmetric bar-like structure. These small bar-bulges are believed to form from the disks of the parent galaxies, through the instability of the rotating disk. This is likely to be a much more quiescent process than the violent merging that led to the formation of the larger $r^{1/4}$ bulges.

In comparing the bulges of M31 and the Galaxy, it seems likely that the different dynamical histories of these two bulges would lead to different star formation histories. We might expect the violent early merger processes in M31 to be associated with rapid star formation and chemical evolution, with chemically enriched stars flung out to large distances by the violent dynamical relaxation. On the other hand, in the inner regions of our Galaxy, the star formation and bulge formation appear to have occurred in a more sedate disk-like environment; the bulge itself formed out of the disk, and the stars of the bulge would remain confined to the inner regions of the Galaxy.

The differences between the old stellar populations of M31 and the Galaxy are most readily studied by observing individual stars in globular clusters and in the outer regions of the diffuse stellar populations in these two galaxies.

The halos of M31 and the Galaxy

First we compare the stellar populations of the globular star clusters in M31 and the Galaxy which have now been studied in great detail. Their stellar color–magnitude diagrams (CMDs) and integrated spectra give estimates of chemical abundances and ages. In most of their properties, the clusters in the two galaxies are alike. The stellar populations within individual clusters of the same chemical abundance are similar, as are the distributions of luminosity of the clusters in the two cluster systems and the radial distributions of the clusters within their parent galaxy. The M31 and Galactic globular clusters cover a similar range of chemical abundance, with the outer clusters of M31 being marginally more metal-rich than their Galactic counterparts[1].

Despite the similarities in the properties of the globular clusters of M31 and the Galaxy, the metallicity distributions in the diffuse (i.e. non-cluster) stellar halos

[1] Here 'metal' means all elements heavier than He, and is often denoted generically as Fe: metal abundance is usually expressed in terms of the (Fe/H) ratio as $[\mathrm{Fe/H}] = \log_{10}$ (stellar Fe/H)/(solar Fe/H). For example, a low metal abundance of $[\mathrm{Fe/H}] = -2$ corresponds to 0.01 of the solar value.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

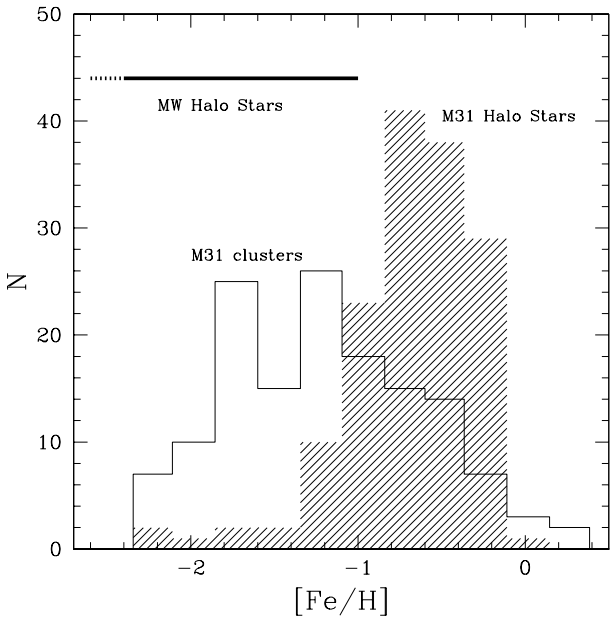


Figure 1. Comparison of metallicity distributions of the M31 halo field stars (hatched) and the M31 globular clusters (adapted from Durrell *et al* (1994)). The bar at the top left indicates the metallicity range of the halo field stars in the Galaxy. The M31 halo stars are much more metal-rich than the Galactic halo stars, although the metallicity distributions for the globular clusters in M31 and in the Galaxy are similar.

of the Galaxy and M31 are strikingly different. For our Galaxy, most halo stars have metallicities [Fe/H] < -1, with a mean of about -1.6. For the outer bulge or halo of M31, CMDs are now available for many fields at distances between about 7 and 40 kpc from the center. All of these studies show that the M31 halo stars cover a very wide range of metallicities, from near-solar down to metallicities [Fe/H] < -2, but the mean metallicity is about -0.6, much higher than for the halo of the Galaxy. This holds true even at the largest radius in M31. Figure 1 shows the marked difference between the metallicity distributions of the stellar halo and the globular clusters of M31. For comparison, the range of metallicities for the Milky Way halo stars is also shown.

Conclusion

What are we to make of the similarities and differences in the stellar populations illustrated in figure 1? The globular clusters in M31 and the Galaxy have similar distributions of metallicity, while the outer stellar halo of M31 is much more metal rich than its Galactic counterpart. These are all old stellar populations, so the origin of the differences will be found in the details of how these galaxies formed, and this is still poorly understood.

Some useful clues are emerging about the nature of the globular clusters. The violent environment of present-day merging galaxies is seen to produce large numbers of young globular-like star clusters. This suggests that the old clusters of our Galaxy and M31 formed in the very early phase of galaxy formation, as clumps of matter were merging together to build the galaxy and the chemical enrichment was proceeding rapidly through massive star formation triggered by the merging.

When we compare the diffuse halos of M31 and the Galaxy, the similarity ends. The two halos are chemically very different, and we must conclude that they formed along quite different paths. At this time, it seems likely that in M31's outer regions we are seeing chemically enriched stars that were scattered out to large radii as part of the violent formation of its $r^{1/4}$ bulge. On the other hand, the halo and bulge of our Galaxy appear to be chemically and dynamically distinct. The Galactic bulge rotates and is relatively metal-rich, like the M31 halo stars in figure 1, while the halo shows no significant rotation and is metal-poor. Following the seminal work of Searle and Zinn (1978), it is now widely believed that the metal-poor stellar halo of our Galaxy is just the debris of small metal-poor satellite galaxies that were accreted and tidally disrupted by our Galaxy. The tidally disrupting Sagittarius dwarf galaxy shows that this process of accretion and halo-building is still going on in our Galaxy.

For more background on M31, see the articles on the ANDROMEDA GALAXY and on the LOCAL GROUP in this encyclopedia.

See also the article on the GALACTIC METAL-POOR HALO.

For current opinions on mergers and the formation and properties of globular clusters, see the articles on GALAXIES: INTERACTIONS AND MERGERS, on GLOBULAR CLUSTERS, on GLOBULAR CLUSTER SYSTEMS IN NORMAL GALAXIES, and on GLOBULAR CLUSTER SYSTEMS IN INTERACTING GALAXIES.

Bibliography

For more details on the chemical abundance distribution in the halo of M31, see

Durrell P *et al* 1994 *Astron.J.* **108** 2114

For the article that started the current view of the formation of the galactic halo, see

Searle L and Zinn R 1978 *Astrophys.J.* **225** 357

Ken Freeman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M32

M32 is an elliptical galaxy (type E2) in Andromeda, a dwarf companion to the Andromeda galaxy M31. It was the first elliptical galaxy to be discovered (Le Gentil, 1749). Its proximity (725 kpc) allowed its outer regions to be resolved into stars by the first generation of large optical telescopes (Baade 1944). Dynamically, it is thought to be dominated by a central supermassive object (similar to the nucleus of M31).

M32 is in many ways a very unusual object. Compared with other dwarf galaxies of the same absolute brightness, its central surface brightness is four orders of magnitude higher and its core radius three orders of magnitude smaller. It is therefore sometimes referred to as a 'compact' elliptical to distinguish it from 'dwarf' ellipticals. Only a very few similar objects have been found, and none is as extreme. Additionally, its nucleus seems to be several billion years younger and more metal-rich than the outer regions: excess radiation in the ultraviolet indicates a hot, young stellar component in the core (figure 1). This is unexpected since elliptical galaxies are generally thought to be older than spirals and to have finished star formation long ago.

These peculiarities, coupled to the fact that M32 and M110 are the closest elliptical galaxies and therefore our best chance to study such objects, have led to M32 being the target for a variety of studies. So far, no single theory has been advanced that could explain all of the galaxy's puzzling characteristics.

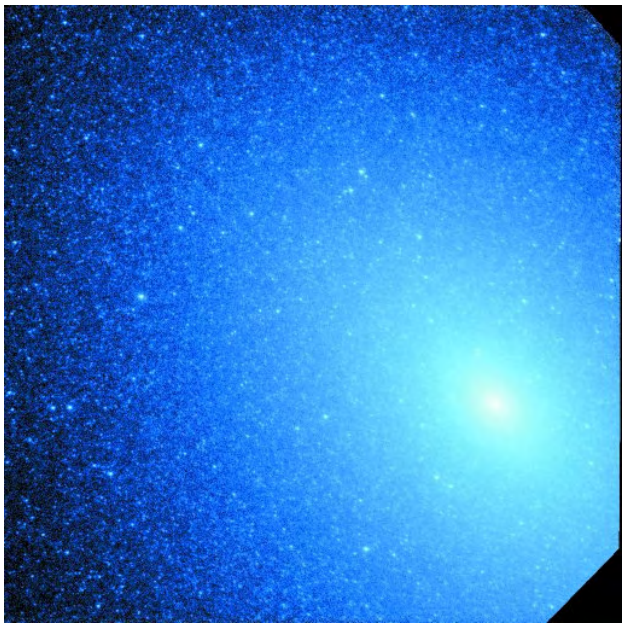


Figure 1. The local group galaxy M32, imaged in the UV (1450–3500 Å) with the Space Telescope Imaging Spectrograph of the Hubble Space Telescope. From Brown *et al* 2000, © *The Astrophysical Journal*, reproduced with permission.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M33

M33 is a spiral galaxy in Triangulum, an apparent neighbor to and possibly distant companion of M31. It is of type Scd, i.e. a late (advanced in its evolution) spiral without a bar or a central black hole, and small compared with the Milky Way and M31 (as all galaxies in the Local Group are small compared with these). The Milky Way and M31 are giant spirals whereas M33 is closer to what is thought to be the average size for a spiral galaxy.

It was discovered by Hodierna before 1654 and independently rediscovered by Messier in 1764. M32 is visible to the naked eye under exceptionally good conditions just south of the line joining α Trianguli to β Andromedae.

The distance to this galaxy is 3 million light-years. It has a diameter of about 50 000 light-years (half of that of the Milky Way) although faint outliers are supposed to extend out much further (more than 60 000 light-years). Its mass is between 10 and 14 billion solar masses (luminous matter only—the size of the dark matter halo of all close galaxies is still a matter of much research and controversy).

The most striking feature of M33 are the pronounced spiral arms which include globular clusters and population II stars as well as large regions of ionized hydrogen (H II). In particular there is one H II region which is roughly 1500 light-years in diameter and so clearly visible as a knot in the spiral arm that it has its own number in Herschel's catalog. It is the largest known H II region, has a spectrum similar to the Orion nebula (M42) and contains over 200 hot, massive, young stars. In addition, there are over 80 diffuse emission nebulae in the spiral arms (visible as knots) and some ongoing star formation in the main disk. M33 is therefore a good target for the study of star formation in an entire galaxy.

The globular clusters in its halo are quite different from the Milky Way globular clusters and seem to be several billion years younger. So far, there is no satisfactory explanation for this phenomenon. Its overall stellar population, on the other hand, does not show any obvious irregularities. No supernovae have yet been observed but several supernova remnants. One hundred and twelve variables have been identified: amongst them four novae, 25 Cepheids and one eclipsing x-ray pulsar (i.e. a neutron star with some binary companion of possibly quite considerable mass). The rest of the variables are mostly RR Lyrae stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M34

M34 is an open cluster in Perseus just north of the straight line between Algol (β Persei) to γ Andromedae. Because of its great angular diameter (35 arcmin, larger than the full Moon) and its relatively high apparent magnitude (5.5) it is easily visible to the naked eye. It was first discovered by Hodierna before 1654 and later independently rediscovered by Messier.

The cluster is fairly loose with only about 100 stars inside the diameter (translating to 14 light-years at a distance of 14 000 light-years). Its age has been estimated at about 250 million years, which puts M34 just between the Pleiades (70 million years) and the Hyades (800 million years)—two of the best studied open clusters in the sky. It can therefore be used for comparison studies of the evolution of chromospheric activity, rotational velocity and lithium abundance in solar-type stars which are common in all three clusters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M35

M35 is a rich open cluster in Gemini (in the direction of the galactic anticenter). It is of roughly intermediate (but not very well determined) age and is visible to the naked eye under good conditions (5.3 mag).

M35 was probably first discovered by de Chéseaux in 1746: he mentioned a 'star cluster above the northern feet of Gemini'. More recently, because of its high luminosity, the large angular diameter (28 arcmin) and its small distance (about 800 pc), it has been one of the most frequently photographed objects in the catalog. Pictures of this cluster taken over the last 100 years can be used for extensive study of the cluster members' proper motions and hence a lot of recent research has centered on the comparison of its internal motion to the predictions for the dynamical evolution of open clusters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M36

M36 is an open cluster in Auriga and the first of a set of three (true neighbors) in the southern part of that constellation (the other two are M37 and M38). All of these were first discovered by Hodierna before 1654.

Amongst the roughly 60 members of M36 are many rapidly rotating stars very similar to the stars found in the Pleiades cluster M45. The cluster does not contain any red giants (unlike its neighbors M37 and M38) and is therefore thought to be very young, i.e. only about 25 million years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M37

M37 is the brightest of the open clusters forming the group of three in Auriga (see also M36, M38). It has a visible magnitude of 6.2, is therefore not visible to the naked eye. Its diameter is about 20–25 light-years and with over 500 members it is one of the richer open clusters. It contains at least 12 red giants and is estimated to be about 300 million years old, i.e. it is a pretty evolved type of open cluster. There is some suggestion from radial velocity measurements of the member stars, that it is contracting slowly.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M38

M38 is the faintest of the three open clusters in Auriga. It forms a distinctive π- or cross-shape and contains approximately 150 stars. Of these, only one is a yellow giant and the age of the cluster has been estimated at about 220 million years. Because of this age and its proper motion it has been suggested that this cluster is in a binary association with the open cluster NGC 1907 but this has yet to be conclusively proven.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M39

M39 is a sparse open cluster in Cygnus, and even though Messier independently discovered it in 1764, its visual magnitude of 5.2 and extended size (32 arcmin, larger than the full Moon) makes it likely that it had been observed previously. There are claims that Le Gentil identified it in 1750 but they are doubtful, since he did not leave a good position, and there are suggestions that Aristotle described the cluster as early as 325 BC.

M39 is thought to have about 81 member stars, but they are difficult to identify since the same area is very rich in faint Milky Way stars. These members are all main sequence with no stars that have spectral type earlier (i.e. hotter) than A, and the cluster has been dated to about 230 to 300 million years. Its small distance (about 300 pc) and large size make it an easy object to study, and recent interest has concentrated on the cluster's dynamical evolution and on comparing its stellar population to other clusters of the same age and metallicity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M4

M4 is a very loose globular cluster in Scorpius and the nearest globular cluster to the Sun. It was discovered around 1745 by Philippe de Chéseaux but remained unpublished (de Chéseaux included a list of nebulous objects in a letter to a friend; eight of these objects were original discoveries but he never published them and so they were not generally known to his contemporaries). It is a very open (or loose) cluster with a distinct bar-like structure, about 10 300 member stars, and a diameter of about 55 light-years. At its distance of 7000 light-years that means that the angular diameter is close to that of the full Moon. It is not as conspicuous as could be expected from this, however, since it is heavily obscured by dust (the ρ Ophiuchi dust complex is nearby). Nevertheless, it can be detected with the naked eye about 1.3 degrees west of Antares.

M4 includes 43 variable stars and in 1987 the first millisecond pulsar was discovered amongst its members. It has a period of 3 ms and is therefore 10 times faster than the Crab pulsar (see M1)—periods as short as this are now thought to be produced by some interaction with a binary companion. The Hubble Space Telescope has taken pictures of some of the 40 000 white dwarves that the cluster is predicted to contain. Analysis of these pictures will allow important conclusions to be drawn about the cooling rate of white dwarves which is crucial for the determination of the Galaxy's age from white dwarf temperatures.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M40

M40 is a fluke in Messier's list of nebulous objects since it is a double star in Ursa Major, also known as Winnecke 4. Messier noted its position when he was looking for a nebula reported by Hevelius in the same region (probably because he thought Hevelius might have mistaken this double star for a nebula). Hevelius had described a 'supra tergum nebulosa', a nebula above the back (of Ursa Major), but this is likely to be a mistake, since none has ever been found in the region, and even the double star is some way from the position described.

The double star has two components, a primary of spectral type G0 (same type as the Sun but four times brighter) and a secondary of spectral type F8 (slightly hotter). Its distance has been recently determined by the Hipparcos satellite to be 510 light-years. Many published versions of Messier's catalog leave out this object since it does not fit with the rest of the list.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M41

M41 is a poor, nearby open cluster in Canis Maior. It is, with a visible magnitude of 4.6, comparatively bright, and there have been suggestions that it was first recorded by Aristotle in 325 BC. If this were true, M41 would be the faintest object recorded in antiquity. However, the reference is not entirely clear. It is clear that the cluster was both included in Hodierna's list of nebulous objects (16654) and recorded by Flamsteed in 1702.

For a poor open cluster, M41 has a remarkably rich population of unusual stars. The percentage of binary systems is thought to be as high as 40% from spectroscopic evidence. Additionally, there are several luminous red giants, some of which are suspected of low-level variability, and two confirmed white dwarfs. Research has very much concentrated on these exotic stars but studies of the cluster are still hampered by the fact that its basic parameters (age and distance) have not so far been satisfactorily determined. The literature quotes age determinations between 30 and 200 million years with the uncertainty due to the fact that the exact amount of reddening for this cluster has been difficult to pin down. Starlight will be scattered off the interstellar dust and thus be shifted into the red part of the spectrum, making the emitting stars look both older and fainter than they really are. To correct for this effect, it is important to have some idea of the amount of interstellar dust in the line of sight and this is not always easy to determine.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M42

M42 is better known as the ORION NEBULA. It is a huge (four times the area of the full Moon), bright (4 mag) diffuse nebula which constitutes the main part of a still larger cloud extending well over half the constellation of Orion. This cloud also contains M43, Barnard's Loop, the Horsehead Nebula and the reflection nebula around M87.

As a whole, the nebula is similar to other diffuse nebulae but it is the brightest and the star formation region closest to Earth. It harbors at its center one of the youngest known clusters, the Trapezium cluster. Its stars light up the nebula (consisting of pure ionized hydrogen in the central regions and some gas and dust in the outer parts) with the starlight partly reflected and partly absorbed and re-emitted. Because of its great luminosity, rich structure can be discerned in the nebula even with conventional telescopes, and detailed investigation with the Hubble Space Telescope has recently uncovered such details as disks of dust and gas forming around and interacting with the bright stars. These are thought to be the beginnings of planetary systems.

At its northern end, the nebula is divided into two parts by a prominent dust lane. Messier assigned an extra number (M43) to the smaller part.

As would be expected for such a bright object, M42 has a rich history in recorded astronomy from Ptolemy (who cataloged the Trapezium cluster if not the nebula itself) via the French lawyer and amateur astronomer Nicholas-Claude Fabri de Peiresc (1610), Christian Huygens (1684), Hodierna (before 1654, he produced the first drawing) and Charles Messier.

Considering that Messier's list concentrated mainly on faint, nebulous, not well known or even unknown objects, it is surprising that he chose to include the objects M42 to M45 which are all bright and were well known at his time. He recorded them in a single night (4 March 1769) and it has been suggested that their only function was to bring the number of objects in the first published version of his catalog (submitted 1771, published 1774) up to 45—possibly simply to beat the number (42) included in Lacaille's 1755 catalog of southern objects.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M43

M43 is a part of the Orion nebula (M42) separated from the main body by a dust lane and sometimes referred to as 'de Mairan's nebula'. It was first discovered in 1733 by de Mairan as a 'brilliance surrounding a star' and consists primarily of a relatively dust free region of ionized hydrogen (H II) and its centrally located exciting star (NU Orionis). Dust mixed with gas lying along or outside the near circular boundary of M43 is important for the energy balance in the region. The Orion A complex consisting of M42 and M43 is the best studied H II region in our galaxy, but work has mainly focused on the central parts and the conditions in the outer gas are not well understood.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M44

M44 is the 'Beehive' or 'Praesepe' (= manger) open cluster in Cancer. Like M42 and M43, this is a very bright (3.5 mag) object known since antiquity. It was first resolved into its component stars by Galileo and contains approximately 350 members. In many ways, such as direction of motion and age (about 700 million years), it is strikingly similar to the nearby Hyades cluster and it has been suggested that both clusters originated in one gaseous cloud. Their x-ray properties are very different, however, possibly because of differences in their binary populations.

Praesepe contains at least four blue stragglers and several δ Scuti stars. The cluster is particularly suited for studies of these stars because of its well-known age and metallicity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M45

M45 is better known as the Pleiades, a young open cluster in Taurus. Again, this is a very bright (1.6 mag) object known since antiquity. The Pleiades are sometimes referred to as the 'Seven Sisters' since that is the number of stars normally visible to the naked eye. In Greek mythology these represent Pleione and her daughters with Atlas: Alcyone, Asterope (a double star), Electra, Maia, Merope, Taygeta and Celaeno. On a very good night more than 12 of the over 500 mostly faint stars that constitute the cluster can be visible.

The Pleiades were first mentioned by Hesiod (1000 BC) but there are also references in the *Odyssey* (Homer) the Bible and countless other recordings. The name could either derive from Pleione, one of the brighter stars and the mythological mother of the seven sisters, or from the Greek word for 'to sail', since it was during the months that the Pleiades were visible in the sky that the Mediterranean was safe for Greek sailors.

The cluster itself is fairly young (70 million years), consists of a few bright and many faint stars but has such low density that its expected lifetime is not much more than another 250 million years. The bright stars are embedded in nebulous material that reflects the light from them. The brightest of these reflection nebulae is around Merope but they were only discovered from the second half of the nineteenth century onwards, when long-exposure photography became available. More recently still, the bright stars were found to rotate rapidly. Pleione ejected a gas shell in 1938–52, an event that was predicted from the rotation by O Struwe.

Since then, the cluster has been shown to contain a very high number of dwarf members: amongst these are 22 binary systems, three triples and an unusual number of white dwarfs. Their presence in such a young cluster has still not been explained. Additionally, brown dwarfs were found in 1995: these are objects with masses between those of stars and planets and had only been theoretically predicted to exist before they were observed in the Pleiades cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M46

M46 is a moderately rich open cluster in Puppis, in a direction where interstellar distinction is unusually small. It was the first object discovered by Messier after he presented the original version of his list (M1 to M45) to the Academy. It has a total population of over 500 stars and its age has been estimated at 300 million years. It contains several red giants and three eclipsing binaries but has not been at the center of much research so far.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M47

M47 is an open cluster in Puppis, first recorded (before 1654) by Hodierna. It was independently rediscovered by Messier on 19 February 1771, but he made a sign error when computing its position and so it was considered a 'missing object' (i.e. possibly spurious) until Oswald Thomas clarified Messier's mistake in 1934. It has to be emphasized again that Messier's list, unlike many of the earlier compilations, contains very few mistakes and no spurious objects. Consequently, the confidence in Messier was such that this wrong position survived into the General Catalog by John Herschel who remarked that: 'This cluster has not since been observed. It is probably a very loose and poor one'.

Under good conditions, M47 is visible to the naked eye (5.2 mag). It is a loose, young (78 million years) cluster of bright stars that resembles the Pleiades in its overall stellar population and it can therefore be used for studies comparing the stellar evolution as a function of mass in different clusters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M48

M48 is another 'missing object' (see also M47). The open cluster in Hydra was recorded by Messier on 19 February 1771 but an error in the data reduction resulted in a wrong position. The error was cleared up and the correct object identified by T F Morris in 1959. In the meantime, the cluster was independently rediscovered by Caroline Herschel. It has an angular diameter of 54 arcmin and a visual magnitude of 5.5 and can therefore be observed with the naked eye under good conditions. The cluster contains about 80 stars but membership is still not entirely clear and it is thought to be about 300 million years old.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M49

M49 is the brightest elliptical galaxy in Virgo. It was discovered by Messier on 19 February 1771 and was the first member of the huge Virgo cluster of galaxies to be recorded, later followed by two more giant ellipticals (M60, M87). It is less dense than M87 and more yellow overall than most galaxies in the cluster.

Ellipticals are galaxies thought to be in a late stage of their evolution and M49 contains no gas, dust, or star-forming systems but around 6000 globular clusters. These are particularly interesting because they can be shown to fall very strictly into two different age groups. Some theories of galactic evolution predict a large fraction of ellipticals to have been formed by the merging of spiral galaxies, which would naturally produce a younger and an older population of globular clusters within the same elliptical. Ongoing studies of the dynamics attempt to determine whether they are consistent with this picture.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M5

M5 is a globular cluster in Serpens Caput halfway between Arcturus and Antares and 0.5 degrees north of 5 Serpentis. It was discovered by the German astronomer Gottfried Kirch in May 1702 but independently rediscovered by Messier in 1764 (Kirch did not publish his discovery). With a diameter of 130 light-years it is the largest of Messier's globular clusters (although larger ones have since been discovered) and unusually asymmetrical. Its visual brightness is 5.6 mag and its age has been estimated at 13 billion years. Its distance from the Sun is 22 800 light-years.

M5 contains 105 variables the brightest of which can be seen 3 arc minutes southwest of the cluster's centre with simple 25×100 binoculars. A dwarf nova was discovered in 1981. Dwarf novae are small stars that ignite due to the accretion of material from a binary companion. These events are more rare than supernovae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M50

M50 is a young, open cluster in Monoceros. It was discovered in 1772 by Charles Messier but there is a possible earlier recording by G D Cassini.

The cluster has an age of about 100 million years and about 200 stars in the main body. Some of these are red and yellow giants but studies of the evolution of high-mass stars have not so far included this cluster. The distance has been estimated to be about 940 pc.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M51

M51, the WHIRLPOOL GALAXY in the constellation Canes Venatici, was discovered by Charles Messier on 13 October 1773, when observing a comet. It has a visual magnitude of only 8.4 and is hence difficult to observe as light pollution makes it easily fade into the background. Nevertheless, Lord Rosse in 1845 recognized its spiral structure and it is therefore the first spiral galaxy that was identified as such. M51 has been classified as Sc, a so-called 'grand design' spiral whose spiral arms can be traced through almost three revolutions. It is the dominating member of a small group of galaxies (also including M63) at a distance of about 9.6 Mpc (37 million light-years). Interactions with its close companion galaxy NGC 5195 have possibly caused the strong spiral structure.

The nucleus of this galaxy is heavily obscured by dust but the presence of an AGN (active galactic nucleus) is strongly suggested by the point source of hard x-rays within it. On the evidence of this and on the basis of optical emission lines, the galaxy was classified as Seyfert 2 or LINER (low-ionization nuclear emission-line region—a type of AGN) and most current research on the galaxy centers on the properties of its AGN.

Additionally, the denser parts of the galaxy have been shown by the ISO satellite to harbor star-forming regions and in 1994 a supernova of type Ic was found by the amateur astronomers Jerry Armstrong and Tim Puckett.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M52

M52 is a comparatively rich but poorly studied open cluster in Cassiopeia. Its visual magnitude is 7.3 and it extends over 13′ but is situated in a region of high interestellar absorption which causes a non-uniform reddending across the cluster and renders determination of its age and distance very difficult. Its distance has been variously given as 25 or 150 Mpc and its age is thought to be around 35 million years (or much older with another distance). The cluster contains 193 probable member stars amongst which there are two potential δ-Scuti variables (one of which was confirmed in 1998), three slowly pulsating B stars and one eclipsing binary.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M53

M53 is a metal-poor (i.e. old), moderately compact globular cluster in Coma Berenices, belonging to the outer halo of our galaxy. It is—within uncertainties—identical in age to M92 and very rich in variable stars. M53 has a visible magnitude of 7.6 and a distance of about 56 000 light-years from us or 60 000 light-years from the Galactic Center. There is little or no foreground reddening so the distance is reasonably well determined. The cluster contains 47 known RR Lyrae stars and about 114 blue stragglers in a strikingly bimodal radial distribution similar to M3.

Clusters in the outer halo are particularly interesting for the study of the formation of our galaxy since they are believed in some cases to be older than the galaxy itself (see for example M54).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M54

The globular cluster M54 in the constellation Sagittarius was discovered by Charles Messier on 24 July 1778 and is hence the first extragalactic globular cluster ever to be recorded. It was not until 1994, however, that it was shown not to be a member of our Galaxy but of a very nearby dwarf galaxy SagDEG (the Sagittarius Dwarf Elliptical Galaxy), one of four globular clusters in that galaxy and possibly its nucleus. It has a visual magnitude of 7.6 and since it is at a distance of slightly over 80 000 light-years this makes it one of the most luminous globular clusters known, second only to ω Centauri. It has a metal-poor blue horizontal branch typical of old globular clusters and contains 82 known variables, the majority of which are RR Lyrae. An accurate age analysis has so far not been possible, but it is thought to be between 0.5 and 1.5 billion years older than M68 and M5 (i.e. about 11 billion years) which would make it significantly older than other member clusters and around 3 billion years older than the field (i.e. the non-cluster stars of our Galaxy).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M55

M55 is a large, loose globular cluster in the constellation Sagittarius with a visual magnitude of 6.3 and a diameter of 19′. It was discovered by Lacaille between 1751 and 1752 when he was observing in South Africa but it is difficult to see from the northern hemisphere. Although Messier first looked for it in 1764 he only found it in 1778. M55 is a halo cluster with a high galactic latitude, which means there is little extinction from dust in the galactic plane and its size and luminosity therefore make it a good object for photometric study. It has an extremely blue horizontal branch and its age is thought to be around 14.5 billion years. If this age determination is correct it can be used to put a lower limit on the age of the universe which is thought to be between 10 and 20 billion years. Particularly interesting is its population of 74 blue stragglers, some of which are thought to be members of binary systems. To confirm this would be an important test for current theories of blue straggler formation, most of which involve binary systems.

M55 also contains a low-luminosity x-ray source discovered by the ROSAT satellite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M56

M56 is a globular cluster in Lyra in a good position for northern hemisphere observers. It has a visual magnitude of 8.3 and an angular diameter of 7.1′ (about 60 light-years at a distance of about 31 600 light-years). The cluster is not particularly rich and its low galactic latitude means that the region is strongly contaminated by field stars (i.e. membership of stars to the cluster is somewhat difficult to establish). It contains two RR Lyrae variable stars, one BL Her, one RV Tauri and three red variables (note that this is the latest count: Helen Sawyer Hogg in the 1940s found 12 variables, among which was a Cepheid that might not be a member).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M57

M57, the 'RING NEBULA', is a planetary nebula in the constellation Lyra. It was discovered by Antoine Darquier de Pellepax and is now considered the prototype planetary nebula in the northern summer sky, even though William Herschel, who invented the term, did not consider M57 a planetary nebula. It consists of a hot (100 000 K) central star surrounded by a ring of dust and gas which is probably some sort of torus and would look similar to the dumbbell nebula M27 if it could be viewed from its equatorial plane. Planetary nebulae are thought to form when an old giant blows off its outer layers, which in this case must have happened in some sort of polar or bimodal form since we are seeing the torus rather than a shell. Recent Hubble images of M57 show subarcsecond dust globules—probably bits of the old star's shell. The hot core left over after the explosion ionizes the gas around it which re-emits the energy in a very few lines (this is very common for planetary nebulae, see also, for example, the description of M27): from the innermost parts of the nebula outwards these are UV radiation, forbidden lines of ionized O and N, and red hydrogen lines. The ongoing chemistry in the nebula is likely to be dominated by photodissociation, shocks and ion–molecule reactions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M58

M58 is a bright, barred spiral galaxy in Virgo, one of the brightest in the Virgo cluster and one of only four barred spirals in Messier's catalog (the others are M91, M95 and M109). It has excited considerable recent research interest mainly because of indications of an active galactic nucleus (AGN). It was classified as a Seyfert 1.9 or a LINER (low-ionization nuclear emission-line regions—a type of low-luminosity AGN) since it has a flat-spectrum radio core (i.e. the core emits equally across all radio frequencies—this indicates some very high-energy processes) and a point-like x-ray source at its nucleus. However, the LINER ionization source is not clear and the energetics of the nucleus are difficult to explain by current models. The latest attempts center around the possible existence of some sort of ionized accretion disk around the x-ray source.

The galaxy is at a distance of 60 million light-years, has visual magnitude 9.7 and contains two recent supernovae (18 January 1988, type II; 28 June 1989, type I) which represent the other big focus of research in this galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M59

M59 is an early-type (i.e. old) elliptical galaxy in the constellation Virgo and a member of the Virgo cluster (the nearest cluster of galaxies, much larger than our own Local Group). It is considerably less luminous than M49, M60 and M87 but still one of the brighter elliptical galaxies in the cluster. M59 contains around 2000 globular clusters, easily 10 times more than the Milky Way but much fewer than the three giants above. The galaxy's most interesting feature is its disk component: there are two basic types of elliptical galaxies, one disky, more regular and one irregular, boxy, usually radio-loud with x-ray activity indicating recent mergers. Disk components are thought to be caused by flattening through rotational velocity and hence the recent work on M59 has concentrated on its internal dynamics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M6

M6 is one of two prominent open clusters in Scorpius positioned in the 'stinger' at the end of the scorpion's tail. Unlike nearby M7, which is seen against a backdrop of rich Milky Way starfields, the surroundings of M6 are relatively star-free. Both clusters are visible to the naked eye as hazy spots but it seems that the first person to establish the true nature of M6 was the Swiss astronomer de Cheśeaux around 1745 before Messier rediscovered it in 1764. The stars of M6 form a shape reminiscent of the outline of a butterfly, from which comes its popular name the Butterfly Cluster. In one of the 'wings' lies its brightest member, BM Scorpii, an orange giant that varies between about fifth and seventh magnitude every 27 months or so. M6 appears somewhat smaller and fainter than M7 since it lies at roughly twice the distance, 1 600 light-years. Its age, though, is only about half as great, some 80 million years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M60

M60 is a giant elliptical galaxy in the Virgo cluster (see also the mention under M59). It has a diameter of 120 000 light-years and an intrinsic luminosity of about 60 million solar luminosities. M60 contains around 5100 globular clusters (for comparison the Milky Way has about 150). These clusters have been studied intensively since they are very interesting for the formation history of elliptical galaxies: the color of the clusters and their positions in the halo gives some clues as to the heavy-element enrichment process during formation.

M60 is also interesting dynamically: arguments from internal dynamics require a central black hole of 10^9 solar masses. This fits with theories of galaxy formation which predict that all large ellipticals should host black holes left over from an earlier quasar phase.

In the radio part of the electromagnetic spectrum M60 possesses weak jets and radio lobes also indicating an active nucleus. Additionally, its spectrum possibly indicates that star formation was going on until fairly recently which would be highly unusual for such an old galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M61

M61 is a barred spiral galaxy in Virgo—another large member of the Virgo cluster. It was discovered by Oriani on 5 May 1779 when following a comet. Incidentally, Messier also saw it that night but mistook it for the comet (he was to make the same mistake twice more before he realized that the object did not move and included it in his catalog six years later). The galaxy is at a distance of about 60 million light-years and has a visual magnitude of 9.7. Its core remains unresolved but recent HST pictures showed a nuclear spiral structure of the massive star-forming regions. It is thought that the bar accumulates gas in the nuclear regions, produces the star-forming rings and might eventually generate or feed an AGN.

Three supernovae have been observed in M61: 1926A, 1961I and 1964F.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M62

M62 is a globular cluster in Ophiuchus which Messier discovered in 1771 but for which he determined an accurate position only in 1779 (hence the relatively high number in the catalog). It is at a distance of 21 500 light-years from the Sun, and has a visual magnitude of 6.5. Its distance from the galactic center is only about 6100 light-years and hence its irregular shape is probably due to deformation by galactic tides. It is very similar to M19, its apparent neighbor, in size and magnitude. Research interest in this cluster has centered on its 89 variable stars. These are mostly RR Lyrae and 74 periods have been determined. It is interesting to note that M19, in spite of its similarity, only has four variable stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M63

The 'Sunflower' spiral galaxy M63 in Canes Venatici was discovered by Messier's collaborator Pierre Mechain, on 14 June 1779. It is in a physical group with M51 and several smaller galaxies and represents the archetypical example of the class of galaxies known as 'flocculent spirals'. In these galaxies, the visible spiral structure is patchy and localized, formed by low-amplitude density waves. It is not clear whether they are simply weak counterparts to the grand-design spirals or whether they are dynamically different. M63 exhibits a weak two-arm structure in the near-infrared and structures similar to the giant molecular associations of the grand design spirals M51 and M100 have been detected.

M63 also has a mildly active nucleus and a supernova was observed in May 1971.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M64

M64 is a spiral galaxy in Coma Berenices sometimes called the 'BLACK EYE' or ('Sleeping Beauty') galaxy, because of the conspicuous dark dust structure near its center. It is at a distance of roughly 12 million light-years but close to the Virgo cluster, which makes its recession velocity almost certainly dominated by infall (and hence renders the distance uncertain). M64 has attracted a lot of interest because of its very unusual dynamic structure: it has an inner disk of gas and stars and a counter-rotating outer disk which is mostly gas. The dust lanes of the 'black eye' are in the inner disk. The two disks are coplanar to about 7 degrees. The transition region between them is not in a steady state and contains vigorous star formation. Originally, it was thought that this bizarre two-disk system was produced by the accretion of a companion galaxy, but even that model does not fully explain the dynamic behavior and it also fails to explain why systems like these are so very rare (only three galaxies with counter-rotating disks have been discovered so far).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M65

M65 is a spiral galaxy in Leo, part of the Leo triplet which consists of M65, M66 and NGC 3628. Its distance has variously been determined as 6.7 Mpc (de Vaucouleurs 1975) and 23.4 Mpc (Sandage and Tamman 1975) but the lower value is generally considered more likely since it seems more consistent with H I studies. The galaxy shows a prominent dust lane on the facing edge with some possible star-formation regions close to it. The stellar population in the disk is smooth and old. The triplet system is very tightly gravitationally bound (see also the description of M66) but that does not seem to have had any effect on the star formation rates of any of the member galaxies. Equally, the spiral arms of M65 have remained tightly wound so far, without any indication of tidal disruption.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M66

M66 is a spiral galaxy in Leo and also part of the Leo triplet consisting of M65, M66 and NGC 3628. The close association between this galaxy and NGC 3628 could be responsible for the disturbed nature of the entire system. M66 is larger than M65 without a well-defined central bulge and very deformed spiral arms. A tidal tail extends from NGC 3628 into the direction of M65 and the two galaxies are linked by a bridge of neutral hydrogen. This clearly indicates a recent close encounter which is thought to have triggered some unusual dynamic processes in M66, such as the central concentration of stellar and gas mass and the formation of a gaseous bar. This bar causes an efficient radial mass accretion across the entire disk and the galaxy is thought by some to be evolving into an AGN (active galactic nucleus). It has an H_1/H_2 mass ratio that is close to Seyfert galaxies and seems to be currently undergoing at least a weak form of nuclear starburst.

Three recent supernovae have been observed in M66: 1973R, 1989B and 1997bs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M67

M67 is an old open cluster in Cancer, the oldest in Messier's catalog at about 4 billion years and still one of the oldest known. It is well studied since it is comparatively rich for an open cluster, at a distance of only 4000 light-years from the Sun and with a high galactic latitude (i.e. there is little reddening or dust extinction). The population of M67 (around 500 stars, at least 38% of which are binaries) serves as a paradigm for studies of structure and evolution of population I solar-age stars, just as Hyades and Pleiades do for younger objects. It has a very well developed population of red giants and some blue stragglers (a further indication that the phenomenon of blue stragglers—hot, young stars in a predominantly old population—is in some way connected to a high incidence of binary associations).

M67 has also been used for extensive comparisons of the evolution of cluster and field stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M68

M68 is a very metal-poor globular cluster in Hydra, very similar to M30. It is a large and loosely spread object at high galactic latitude and has been studied mainly for its age and its metallicity. It is thought to be around 13 billion years old (i.e. identical in age, within the errors, to M15), but the exact metallicity is difficult to determine since the indicators lose sensitivity at the low end. However, it is not quite metal-poor enough to consist of population III (i.e. primordial) stars.

The stellar population of M68 is predominantly old with a gap in the horizontal branch of the HR-diagram very similar to that in M15: the eight bluest stars are clearly separated from the remainder of the horizontal branch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M69

M69 is a metal-rich, globular cluster in Sagittarius, which was originally thought to be a 'disk' cluster, i.e. a member of the subsystem of globular clusters in the galactic disk. However, the distinction between disk and halo clusters is controversial and never very clear. M69 can also be regarded as a halo cluster with a very eccentric orbit. Additionally, it has the age and metallicity of a typical halo cluster.

The cluster was discovered by Lacaille. Messier first missed it in 1764 (the object is small and faint and becomes more difficult to observe further north) but found it with a better telescope in 1780. It has a bright and compact core with eight known variables. A possible (but still unconfirmed) cataclysmic binary was discovered in 1986.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M7

M7 is an open or galactic cluster in the constellation Scorpius. It is easily visible with the unaided eye some degrees northeast of Shaula (λ Scorpii) and was first mentioned by Ptolemy in AD 130 (his mention might also include the nearby M6). Hence the cluster is often referred to as 'Ptolemy's Cluster'. It was included in the published lists of nebulous and stellar objects by Hodierna (around 1654—counted 30 stars), Halley (1678) and by the Abbé de la Caille (1755—counted '15 or 20' stars).

M7 contains about 80 stars of magnitudes greater than 10 (the brightest is a yellow giant). Its overall visual brightness is 4 mag and its diameter is 18 light-years. The age was estimated at 220 million years (note that Koelbloed gives a much lower value of 70 million years) and the distance to the Sun as 800 light-years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M70

M70 is a globular cluster in Sagittarius at a distance of about 28 000 light-years. It has a southern declination and is therefore difficult to observe from the northern hemisphere but it is a bit more luminous and bigger than its apparent neighbor M69. For a globular cluster it is of only intermediate metallicity and it has a very pronounced blue straggler sequence. The far-UV light is smoothly distributed and not dominated by a single bright star, it is therefore not entirely clear what causes the strong emission in that region. There are only 10 variable stars, all of which are RR Lyrae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M71

M71 is an unusual globular cluster in Sagitta in that it is very metal rich (only M69 is richer) and very sparse. For a long time its classification was undetermined between a sparse globular cluster and a very dense open cluster. It is, however, now generally thought to be a globular cluster. Its second unusual feature is its very low-eccentricity orbit. Globular clusters usually move around the galactic plane in the halo, but there is a population of metal-rich clusters which are concentrated close to the plane. M71 belongs to these and has been studied carefully in order to gain more insight into the process of cluster formation. Interestingly, its stars seem to have a very varied abundance of some of the heavier elements which might indicate that the gas from which they were formed was not properly mixed. Unlike most globular clusters, M71 has only a very few variables. Eight have been identified so far, none of which are of the common RR Lyrae type.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M72

M72 is a globular cluster in Aquarius, not very concentrated, quite distant but intrinsically very luminous. It hosts 42 known variable stars and is part of a small group of globular clusters together with NGC 6584, NGC 6864, M75 and others. This group has been the focus of some interest since it has a high apparent retrograde motion around the galactic center. The significance of a group of apparent retrograde clusters lies in the hypothesis that clusters of the galactic halo formed in relatively large satellite galaxies which subsequently merged with the Galaxy to become the galactic halo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M73

M73 is an asterism of four stars which was recorded by Messier only as a help for the determination of the position of M72 (which is 1.5 degrees to the west). It is not entirely clear whether the group is a real cluster rather than a mere asterism. E M Arnal classified it as an open cluster in 1979, but this assertion does not seem to have been checked.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M74

M74 is a spiral galaxy in Pisces with a linear diameter of 95 000 light-years (i.e. of roughly the size of the Milky Way) and a visual magnitude of 9.4. In many ways, this is a prototypical 'grand design' spiral. Its arms are about 1000 light-years broad, it has a very small and bright nucleus and is particularly beautiful on photographs because it is face-on. As well as its normal appearance, however, it has been shown to have a very extended disk of neutral hydrogen with a complex velocity structure (i.e. it is slightly warped). The velocity field in the outer disk is inconsistent with differential rotation in a flat disk. The galaxy also exhibits some irregularities in its otherwise smooth and regular spiral arms. A recent merger event (or recent merger events) has been postulated to explain this behavior but no suitable candidate can be discerned in the vicinity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M75

M75 is a globular cluster in Sagittarius at a distance of 57 700 light-years (i.e. beyond the Galactic Center). It is compact and has a very high luminosity (about 160 000 solar luminosities). It is a member of the same group as M72 and therefore has an apparently retrograde orbit around the Galactic Center. Like most globular clusters with a high degree of concentration and a relatively high metallicity, it has a fair number of variable stars (10 RR Lyrae, six with known periods) and some radio sources.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M76

M76 is a planetary nebula in Perseus, often called the 'LITTLE DUMBBELL NEBULA' (or the 'Cork' or 'Butterfly' Nebula). It was assigned two NGC numbers since it was originally thought to be two distinct objects. Like most nebulae, it has a very ill-defined distance: estimates range from 1700 to 15 000 light-years. The bright part consists of a slightly elliptical ring that is pointing towards the observer (i.e. it is viewed nearly edge-on). Attached to that ring are two inner lobes (the wings of the butterfly) and fainter outer lobes which move at a much lower expansion velocity. Additionally, there is a polar cap on the southeast side, which has a higher expansion velocity again. The nature of these outer lobes and the polar cap remains unknown, but it is generally thought that the nebula was probably formed by some interaction of stellar winds. The central star has been identified and has a temperature of about 60 000 K.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M77

M77 is a spiral galaxy in Cetus, at a distance of about 60 million light-years and with a diameter of between 120 000 and 170 000 light-years. It belongs to the class of galaxies known as 'Seyfert galaxies', i.e. it has a brilliant nucleus with intense star-forming activity in a disk around it, an emission-line spectrum and fairly faint spiral arms. Seyfert galaxies themselves are a particular type of AGN (active galactic nucleus) galaxy. Most of these have (or are thought to have) a supermassive accreting object at the center that powers the intense luminosity of the object. In the case of M77 this object is thought to have about 10^7 solar masses. Recent research interest has mostly centered around the central object and the star-formation activity in the disk. M77 has the most luminous water-vapour maser yet reported (350 solar luminosities) which indicates an ongoing burst of star formation. But because of its brightness, proximity and favorable orientation, the galaxy in general provides a good laboratory for studying the Seyfert nucleus, inner disk and circumnuclear ring.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M78

M78 is a reflection nebulosity seen against the Lynds 1630 dark cloud and appears to be associated with the Orion complex. It contains many bright, hot B stars which are probably very young and about 45 variables have been identified in or near the nebula. For this reason, M78 is also sometimes referred to as a 'young embedded cluster' and has been important for the study of stellar evolution. Its visual structure is dominated by a band of dark obscuration that delineates the northern edge of the nebula. It resembles a faint comet. Its distance has been estimated at 1600 light-years and its visual magnitude is 8.3.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M79

M79 is a globular cluster in Lepus, at a distance of about 41 100 light-years. It is moderately metal-poor (metallicity comparable to M13) but has an extremely blue horizontal branch and generally shows more UV bright objects than would be expected for such a cluster. It has been speculated that there could be kinematic processes changing the stellar evolution and M79 is sometimes regarded as a possible instance of core collapse because of the irregularity of its profile and the fairly high concentration. It is slightly elliptical and contains seven known variables.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M8

M8 is a diffuse nebula also known as the 'LAGOON NEBULA' because of the dark band separating it into two bright parts. The nebula is associated with an open star cluster (NGC 6530) which can be seen with the naked eye in the constellation Sagittarius to the southeast of the bright nebula M20. The cluster was first discovered by Flamsteed around 1680 (he was the first Astronomer Royal and published a star catalog *Historia Coelestis Britannica* in 1712). The French Astronomer Le Gentil described the nebula in 1747.

The open cluster was created by star formation by the collapsing gas in the nebula and most of the interest in this object centers on the fact that star formation can still be seen to occur. The nebula is rich in substructures, notably the dark globules (which are collapsing protostellar clouds) and the recently discovered twisted-rope structures which are possibly tornado-like phenomena caused by the large difference in temperature between the hot surface and the cold interior of the clouds combined with the pressure of starlight. Within its brightest part, there is a region known as the 'HOURGLASS NEBULA' where star formation appears to take place currently. The bright emission of this part (and the emission in the rest of the nebula) is caused by the gas re-emitting radiation that was absorbed from the very hot young stars. The open cluster contains between 50 and 100 of these with a large number of irregular variable stars such as 'flare' stars (type UV Ceti). Its age is estimated to be just 2 million years. The nebula has a diameter of 140 light-years and is about 5 200 light-years away from the Sun. Its visual brightness is 6.0 mag and it also emits radio waves in the 9.4 cm band.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M80

M80 is a globular cluster in Scorpius that is unusual for two reasons (which may be related). In May 1860 a dwarf nova occurred in the cluster possibly caused by a cataclysmic variable. Dwarf novae are rare events in globular clusters (see also M14). M80 also has the largest and most concentrated population of blue straggler stars (which are now thought to be a late stage in the evolution of a close binary system) ever observed in a galactic globular cluster. This may suggest that it is in a transient dynamical state during which stellar interactions are delaying the core-collapse process.

The distance to M80 is not very well determined and varies between 27 000 and 36 000 light-years. Its visual magnitude is 7.3 and its diameter is 8.9′.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M81

M81 is an early-type spiral galaxy with a prominent bulge and well-defined arms. It is sometimes also known as 'Bode's Galaxy'; it is very bright, and can under good conditions be seen with the naked eye in the constellation Ursa Major. Together with M82 it forms a small group and there are some indications that the two have had a close encounter in the past (M82 seems to have been affected more—its shape is more irregular than that of M81). With a distance of only 12 million light-years, M81 is a fairly nearby galaxy and has therefore been extensively observed. It has a dense and unresolved core which emits strongly in the radio waveband as well as hosting an x-ray source. This, together with other indications, such as low-ionization emission lines, has led to the conclusion that the core of M81 must also host a low-luminosity AGN (active galactic nucleus). As is usual with these phenomena, the details of the object generating the power in the center are not well understood. But it is clear that M81 shows features similar to a weak Seyfert galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M82: The Starburst Galaxy

The nearby irregular galaxy M82 drew the attention of observers decades ago owing to its spectacular appearance, which suggested that it was an 'exploding' galaxy. The appearance of an explosion derives from the consequences of the enormous amounts of star formation activity in this galaxy. Intense star formation events in the central regions of galaxies, known as the starburst phenomenon, may be a common part of galaxy evolution, so the study of M82, the prototypical STARBURST GALAXY, is of particular importance. The STAR FORMATION is associated with a very rich interstellar medium with large amounts of molecular gas, dust and ionized gas as well as x-ray-emitting plasma. These are extremely unusual conditions in a galaxy, which raises the question of its origin as well as of its evolution and eventual fate.

The galaxy M82 is a disk galaxy several times smaller than the Milky Way and in one of the closest groups (3.6 Mpc), whose three brightest members are M81, M82 and NGC 3077. There appears to have been a tidal interaction between the members of the group that has torn neutral hydrogen from the outer parts of M81 and M82. This interaction, which took place about 10^8 yr ago, has perturbed the inner part of M82, probably producing the stellar bar in the bulge region. The presence of the bar and the tidal disturbance of the gaseous disk have led to the inflow of gas from the disk to the central 1 kpc of the galaxy. This buildup of gas occurred fairly quickly and it led to a period of rapid star formation that has occurred in a relatively small region. The energy released by the multitude of massive stars produced a superbubble of hot gas that has broken out of the disk and is flowing rapidly away from the galaxy.

Central star formation activity

The most direct evidence for star formation comes from the identification of the many young STAR CLUSTERS in the central region. Many of these star formation regions are obscured by dust, but, despite this limitation, HUBBLE SPACE TELESCOPE imaging reveals nearly 100 star clusters which have a typical size of 3 pc and a typical luminosity of several million solar luminosities, which exceeds that of any of the famous star formation regions in the Milky Way and even the most luminous star-forming cluster in the Large Magellanic Clouds. These star clusters are so dense that they are most similar to GLOBULAR CLUSTERS, although in these environments they will most likely become tidally disrupted rather than becoming long-lived gravitationally bound units. These clusters can occur in associations, with at least 45 such clusters found in one region alone. The optical observations set a lower limit to the actual number of clusters since dust extinction hides many clusters. Observations at some wavelengths are not affected by dust extinction, such as radio continuum and line observations. The continuum observations are sensitive to some of the SUPERNOVAE that occur, which should be associated with the young star clusters, but none of the radio supernovae coincides with an optical cluster. If the radio supernovae are randomly distributed in young star clusters, then the lack of coincidence with the optical clusters implies that they must be associated with optically obscured star clusters. These obscured star clusters would have to be 20–40 times more common than the optically visible clusters in order to satisfy the statistical result that the supernovae do not coincide with visible clusters.

The rate of star formation and the length of time over which the starburst has occurred are not determined by direct studies of the star clusters, but rather from the studies of the emission by the interstellar medium. Of the total luminosity produced by M82, 84% is emitted at wavelengths longward of 3 μm, and this is dominated by reemission from dust. The dust absorbs the ultraviolet light from the young stars and it is heated to a temperature of about 45 K. At this temperature, most dust grains emit like black bodies or gray bodies with a peak emissivity at about 60 μm. Some smaller dust grains are heated to higher temperatures owing to non-equilibrium effects, emitting at wavelengths as short as a few microns. However, the smallest dust components, the poly-aromatic hydrocarbons, are probably destroyed by the intense ultraviolet radiation field, since their infrared spectral signatures are absent.

Because of the heavy obscuration in the optical region, the infrared lines are the most useful diagnostics of the conditions within the star-forming region because the optical depth in the 50–157 μm range is less than 0.1. Both moderate-ionization lines ([O III] λ 51.8 μm, λ 88.4 μm, [N III] λ 57.3 μm) and low-ionization lines ([O I] λ 63.2 μm, λ 145.5 μm, [C II] λ 157.7 μm) have been measured by the Kuiper Airborne Observatory as well as by the Infrared Space Observatory. The moderate ionization lines are typical of H II REGIONS, so their luminosities and the line ratios give direct insight into the number of ionizing photons as well as the photon-to-gas density ratio. Strong low-ionization lines are not found in classical H II regions, but they are typical of photodissociation regions, where the ionizing radiation is dissociating the molecular gas and heating it, eventually becoming part of the H II region. Models for this emission indicate a star formation rate of $10 M_\odot$ yr^{-1} and a relatively young age for the starburst, (3–5) $\times$ 10^6 yr, and a mass of young stars of $10^8 M_\odot$. Several factors, such as the dynamics within the core and the superwind of the starburst, point to an age several times greater ((1–2) $\times$ 10^7 yr), and this longer age is probably the more reliable one. This longer age can be reconciled with the age inferred from the star formation rate if the young star clusters have an excess of high-mass stars (a flatter initial mass function), although this point is controversial. The star formation rate emphasizes the extreme nature of the starburst phenomenon: about 3 times the total star formation rate of the Milky Way occurs within just a few hundred parsecs of the center of M82, 10^{-4} of the surface area of the Milky Way.

The stars are forming from the plentiful molecular gas in the core, which has been the subject of numerous

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

studies, both with single dish and synthesis array telescopes and in several transitions of a variety of molecules. These studies indicate that there is about $2 \times 10^8 M_\odot$ of molecular gas in an irregular ring that is in circular or elliptical orbit about the center of the galaxy. The star formation efficiency for this gas has been quite high, about 30%, given the estimate for the mass of young stars, $10^8 M_\odot$, and roughly 30 times higher than the typical star formation efficiency in the disk of the Milky Way.

The ensemble of young stars has several effects on the environment from which they formed. In addition to the photodissociation and H II regions produced by the ultraviolet photons, stellar winds and supernovae are extremely important in shaping the interstellar gas. Fast stellar winds from massive stars will create a hot bubble in a uniform medium along the lines of the classic superbubble scenario. At the end of the life of a massive star, the supernova event deposits about 3 times as much energy as the fast wind, eventually dominating the dynamics. The supernova rate is estimated to be about one supernova every 10 yr, 3–10 times greater than that for the entire Milky Way, but confined to this relatively small region. The frequency of supernovae is large enough that it is appropriate to treat the energy deposition as an input of energy that is smooth in time. To follow the response of the gas to this energy input, numerical hydrodynamic models have been used, assuming a locally uniform medium with a decreasing density out of the disk. These models show that, in the early stages, a spherical hot bubble develops, but, as the bubble size becomes comparable with the disk thickness, it begins to elongate perpendicularly to the disk, along the direction of the steepest density gradient. As the shock proceeds along the density gradient, it changes from a decelerating shock to an accelerating one. For the usual decelerating shock, the dense shell that has been formed is stable, but, with an accelerating shock, the dense shell is subject to the growth of Rayleigh–Taylor instabilities and it breaks up. This leads to the rapid outflow of material from the galaxy, discussed below.

Various aspects of this hydrodynamic phenomenon are seen at several wavelengths, both in the active star-forming region as well as in the flow extending from the disk. In the position–velocity diagram for ^{12}CO(1–0, 2–1), ^{13}CO(1–0) and ^{18}CO(1–0), there is an absence of emission at ± 45 km s^{-1} for a region about 80 pc in diameter (figure 1). These observations indicate the presence of a ring of expanding molecular gas of radius 65 pc around a region where the molecular gas has been destroyed. The destruction of the molecular gas was probably due to a combination of ionizing photons, fast stellar winds and supernovae, but the expansion velocity of the molecular gas is probably due to the supernova shocks. There is also an absence of radio continuum emission in a region roughly coincident with the CO hole, although there is the possibility that free–free absorption by the ionized gas has dimmed the radio continuum.

The x-ray emission reveals only three point sources in the central region, none of which is coincident with any of the radio point sources or the optical clusters. However, x-rays are dominated by more diffuse emission, which is brightest in the central region and most likely fills the central hole in the molecular gas. This emission appears to consist of multiple components with emission lines, and one decomposition consists of a hard, a medium and a soft component. The soft and possibly the medium component are probably emission from a hot plasma of temperature 0.3 keV and 1.0 keV. The presence of the line emission from eight different elements (from ASCA) plus the diffuse nature of the emission (from ROSAT) indicate that the radiation is produced by hot dilute gas (density of $\sim$0.1 cm^{-3}, temperature of 10^7 K), although of subsolar metallicity. The metallicity of the stars in M82 is subsolar, as expected for a modest-size galaxy, but the metallicity of the gas was expected to be considerably greater owing to the many supernovae that have occurred. Determination of the metallicity in the x-ray-emitting gas with ASCA is not simple because of multiple temperature components and unresolved point sources. Yet it is unlikely that the metallicities are grossly in error, with the largest metallicities occurring for S and Si (0.4–0.5 of the solar value), while Ne and Mg have abundances near 0.2 of the solar value and O, N and Fe have metallicities less than 0.1 of the solar value. These abundance ratios are perplexing, since they do not correspond to those expected from either type of supernova, and it will be interesting to see whether the combination of high spatial resolution and good spectroscopy (from the CHANDRA X-RAY OBSERVATORY) will confirm the abundances.

Finally, the x-ray observations require a hard component (approximately 14 keV if it is a thermal plasma), which is found to vary on a timescale of days and has a luminosity that can approach 10^{41} erg s^{-1}. This seemed unlikely to have a stellar origin, and this led to the suggestion that a supermassive compact object was

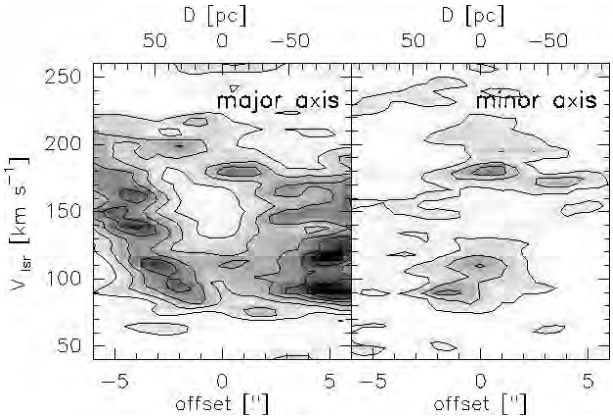


Figure 1. Position–velocity diagrams for the ^{12}CO($J = 1$–0) transition along the major and minor axes of M82. The intensity in the line is proportional to the darkness, so the light region in the center indicates a hole in the molecular distribution surrounded by an expanding molecular ring (from Weiss *et al* 1999 *Astron. Astrophys.* **345** L23).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

present, such as a black hole, with a mass in excess of $500M_{\odot}$ if it emits below the Eddington luminosity. This might be thought of as a tiny ACTIVE GALACTIC NUCLEUS, possibly activated by the abundant gas in the central region. From energetic considerations, this nuclear source has little influence on the development of the starburst, which has a bolometric luminosity that is 3 orders of magnitude greater (8×10^{43} erg s^{-1}). The infrared line ratios also are sensitive to the presence of an active galactic nucleus, whose hard radiation will produce high excitation lines. However, the observed ratio of the high- to medium-excitation lines is particularly low in M82 (there is only an upper limit to [Ne V]/[Ne II]), confirming that an active galactic nucleus is either absent or unimportant. Insight into the presence of a central black hole will be provided by the new generation of x-ray telescopes, and Chandra observed M82 during the first few months of operation. X-ray emission from a strong central component was not discovered, confirming the absence of a strongly emitting active galactic nucleus and indicating that the hard emission is most likely caused by another mechanism.

The superwind

The outflow of gas from the star-forming region is observed in several wavebands and it was expected from models as well. The first indication of an outflow came from the study of colors and emission line images in the 1960s, where filaments were seen aligned with the minor axis and extending to at least 3 kpc from the center of the galaxy. These initial studies included spectroscopy of the emission line filaments, which clearly showed recession velocities from the galaxy of hundreds of km s^{-1}, with the exact value depending on the adopted inclination correction, which is substantial since the minor axis of the galaxy is only 8°–9° from the plane of the sky. Whereas many of the basic elements were identified in these observations, more detailed optical emission line studies along with x-ray investigations have given greater insight into the phenomenon.

Optical emission line studies have been carried out with Fabry–Pérot observations, long-slit spectroscopy and narrow-band imaging. These efforts reveal that the outflow on either side of the galaxy is not quite aligned with the minor axis and has a deprojected outflow velocity of 500–800 km s^{-1}. For lines of sight through these outflow regions, the emission lines are bimodal, which is interpreted as the near and far sides of the outflow region (lobes or cones) expanding laterally with a velocity of about 150 km s^{-1}. Close to the galaxy, the emission lines are generally of low ionization, consistent with photoionization from the hot stars in the galaxy. However, in the outer region (beyond 1 kpc), the relative strength of high-ionization lines increases, such as that from [O III], which is unexpected for photoionization, but consistent with collisional ionization from shocks. Beyond 3 kpc, the optical emission line gas becomes generally undetectable, with the exception of a filament 11 kpc from the galaxy and

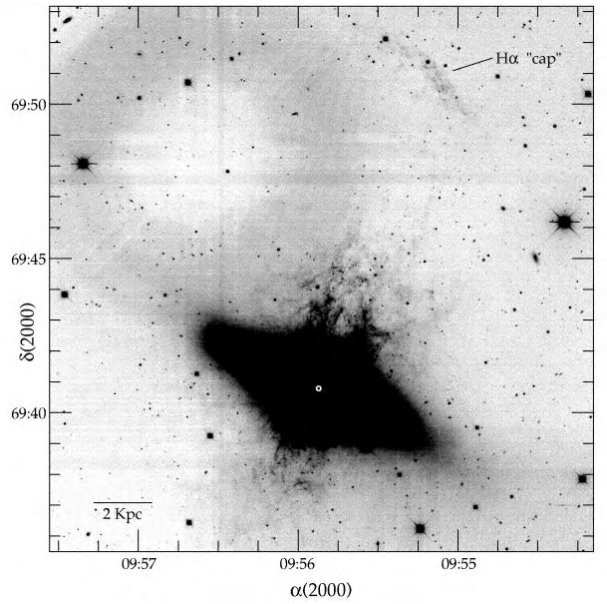


Figure 2. An image of M82 taken in the light of the hydrogen line Hα. The 'donut' on the upper left is an artifact of the corrective optics. The Hα filaments that extend upward from the center trace the breakout and the outflow of the superwind. The Hα cap to the top of the figure is 11 kpc from the galaxy and is probably produced by an impact between the superwind and tidal debris, present before the starburst event (from Devine and Bally 1999 *Astrophys. J.* **510** 197).

with an orientation that is nearly at right-angles to the flow direction (see figure 2). The most likely explanation for this filament is that it represents the interaction of the outflow of material with the cold gas that has been torn from M82 and M81 in their tidal interaction. The strongest support for this interpretation comes from the x-ray observations (the PSPC on ROSAT), which are coincident with the optical emission. The shock interaction between the outflowing gas and slow-moving tidal debris would lead to both the production of soft x-rays and optical emission line excitation. From the temperature of the x-ray emission (10^7 K), a velocity of 800 km s^{-1} is inferred, which is similar to that inferred from the optical emission lines closer to the galaxy. With this velocity and the projected distance, a minimum timescale of 14×10^6 yr is inferred for the time since the breakout of gas from the nuclear region.

The x-ray emission (from ROSAT) is coincident with the optical emission line gas in the inner region, but it extends beyond the optical emission, being visible to about 6 kpc on both sides of the galaxy (plus the additional separated emission 11 kpc from the galaxy on one side) (see figure 3). The temperature of the extended flow regions is approximately 0.5 keV (6×10^6 K), with only a weak decrease in the temperature with distance along the outflow (from 0.6 keV to 0.4 keV). This is a critical measurement for understanding the flow, since, if adiabatic expansion controlled the temperature of the gas as it flowed outward, the temperature would drop far

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

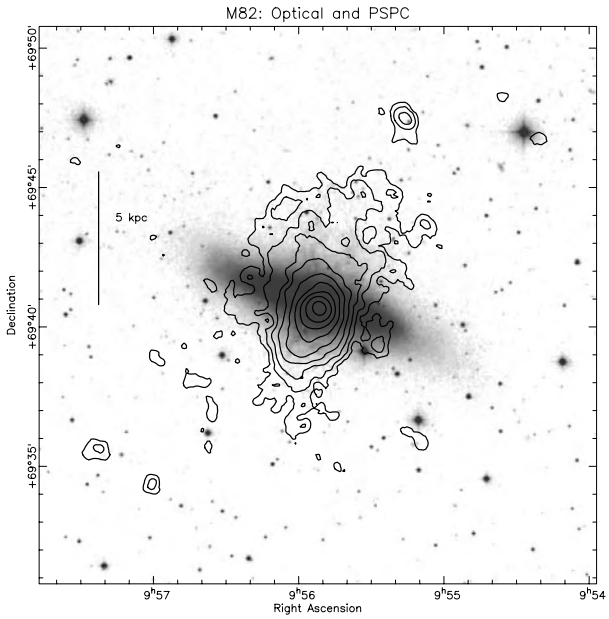


Figure 3. Contours of the x-ray intensity, obtained with the PSPC on ROSAT (0.1–2.4 keV, 30″ resolution), is superimposed upon an optical image of M82. The intense emission from the central region is from the hot gas, while the extended emission perpendicular to the disk, due to the superwind, probably results from an interaction between the superwind and halo gas (from Strickland, Ponman and Stevens 1997 *Astron. Astrophys.* **320** 387).

more rapidly than is observed. The alternative model that has developed posits that a gaseous halo surrounds M82 prior to the starburst event. Following the breakout of the hot gas from the disk of the galaxy, the superwind of hot gas shocks with the halo gas, converting kinetic energy to thermal energy. In this picture, the thermal energy carried by the outflowing gas is unimportant compared with its kinetic energy, which is not subject to adiabatic losses. This also helps to solve the metallicity problem, since most of the emission originates in gas that is characteristic of the pre-starburst supernova, which, for a low-mass galaxy like M82, is subsolar.

In addition to the interaction of the hot gas with a halo environment is the issue of whether the outflow entrains gas, such as for the region close to the galaxy. A determination of the mass flux of the superwind places it at several times greater than that expected from the ensemble of supernovae, suggesting that there has been 'mass loading' of the flow. In this context, mass loading is the process whereby mass is added to the superwind by processes other than the normal shock heating of gas by supernova blast waves. Mass loading is most likely to occur in one of two ways: by the conductive evaporation of cold clouds or by turbulent mixing of cold gas into the hot flow. Conditions are also favorable for conductive evaporation to occur because we know that plentiful molecular gas is close to the hot x-ray-emitting gas. However, efficient conductive evaporation requires magnetic fields that are simply connected between the hot and cold phases, and whether this criterion is met is unknown. Conditions are suitable for turbulent mixing because there is a large differential velocity between the cold gas and the hot superwind. This will lead to the growth of Kelvin–Helmholtz instabilities, producing tongues of cold gas that will mix with the hot gas (shocks will accompany this process). In this case, it remains to be seen whether the rate at which gas can be mixed into the flow is adequate.

A final issue is whether the superwind will eventually fall back to the galaxy or whether it will escape the gravitational field of M82 and become part of the intergalactic medium. The characteristic rotational velocity of M82 is only 100 km s^{-1}, and it is declining with radius, so the escape velocity is unlikely to be greater than 200–300 km s^{-1}, which is significantly less than the velocity of the superwind, 500–800 km s^{-1}. In the outermost region of the wind, 11 kpc from M82, the escape velocity is probably only 50–150 km s^{-1}, so even the cold shocked material is unlikely to be bound to the galaxy. These comparisons indicate that the superwind will remain unbound to the galaxy and will pollute the surrounding medium. If starburst events are common in galaxies, this will be an important means of enriching the intergalactic medium.

Future directions

The nearest starburst galaxies, M82 and NGC 253, have been fundamental to our understanding of the starburst phenomenon, and we look to further studies to settle the outstanding questions. One of the central issues is how the star formation is initiated. Although star formation is occurring in the disk of M82 on kpc scale, it is extremely elevated in the central region, where it has evacuated a hole in the molecular disk. This star formation may propagate outward by creating high-pressure hot x-ray-emitting gas that compresses the cold molecular gas. Current observations show that conditions exist for this to take place, in which case we should see a progression in the age of star clusters as a function of radius in the nuclear region. The detection of these star clusters will require high-angular observations in the infrared region in order to reduce the extinction. This will require observations from NICMOS on the Hubble Space Telescope and subsequent observations with the Next Generation Space Telescope (see INSTRUMENTATION: NEXT GENERATION SPACE TELESCOPE). Also, these observations will identify star clusters that are obscured at optical wavelengths, permitting us to determine whether the radio supernovae are indeed associated with star clusters. These observations should help resolve the issue of whether the star formation produces relatively high numbers of high-mass stars.

Regarding the superwind, there are several outstanding problems that will be addressed with the x-ray telescopes Chandra and the X-RAY MULTI-MIRROR MISSION (XMM).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The central region is tremendously complex, containing point sources, line and continuum emission from hot gas, along with partial absorption by cold gas of widely varying column densities. The high angular resolution of Chandra will be essential in separating these aspects of the emission and determining the basic conditions in the hot gas, such as the temperature, pressure and metallicity. These observations will help us to understand the role that the hot gas has in triggering further star formation in the center and in transforming the cold molecular environment into a hot x-ray emitting medium. Another issue is the degree of entrainment of material in the superwind flow, which would produce a metallicity gradient along the flow, and this will probably best be determined by XMM observations, since sensitivity rather than angular resolution is paramount. Finally, the extent of the wind and its effect on polluting the local environment is an important but challenging issue. The maximum observable extent of the x-ray superwind will be measured by XMM, but that determination may be limited by the low surface brightness of the emission relative to the background. Nevertheless, the future of the field will be very exciting and major advances should occur in the next few years.

Bibliography

The literature is extremely extensive for M82, so I have chosen some of the most recent publications in their areas, which reference the important works that preceded them.

The star clusters that are seen in the central region are discussed by

O'Connell *et al* 1995 *Astrophys. J.* **446** L1

The molecular content in the galaxy and the molecular superbubble are discussed by

Weiss *et al* 1999 *Astron. Astrophys.* **345** L23

The x-ray imaging and temperature studies are thoroughly described by

Strickland D K, Ponman T J and Stevens I R 1997 *Astron. Astrophys.* **320** 378

Models for the breakout and the superwind may be found in

Suchkov *et al* 1996 *Astrophys. J.* **463** 528

Joel N Bregman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M83

M83 is a spiral galaxy of type SABc (i.e. it is an intermediate type between a barred spiral and a normal spiral galaxy) in Hydra (sometimes referred to as the 'Pinwheel Galaxy'). It was fist discovered by Lacaille during the time he spent at the Cape of Good Hope (1751–2) and is therefore the first galaxy beyond the Local Group to be discovered. Its distance is estimated at 15 million light-years, which makes it also the nearest barred spiral.

M83 is a typical example of what is known as a 'starburst galaxy'. Starbursts are areas having very high star formation rates caused by the fast transformation of gas, the violent interaction of stars with the interstellar medium through ionizing fronts, stellar winds and the high rate of supernovae associated with a young stellar population. The origin of the phenomenon is not well understood but it is important for galaxy evolution. Most starbursts are associated with AGN (active galactic nuclei) as well as high-density gas and a non-symmetric gravitational potential usually due to a bar or an interacting galaxy. M83 is typical in that it has a bar as well as a suspected AGN.

Together with NGC 5128 and NGC 5253, M83 forms a small group. Both of these other galaxies are peculiar, which may indicate a previous interaction. Six supernovae have been reported so far—more than in any other Messier object.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M84

M84 is a lenticular galaxy in Virgo at a distance of approximately 60 million light-years. It is populated by old yellow stars and has a system of globular clusters (much less dense than that of M87). It also has a rapidly rotating nuclear gas disk and hosts an AGN (active galactic nucleus) that is most likely powered by a supermassive (310^8 solar masses) central object. This object emits two small but conspicuous jets in the radio wavelength. Three supernovae have been reported so far.

Probably because of its greater distance, M84 is not particularly well studied. Most recent research interest has been concerned with the study of the supernovae rather than with the AGN.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M85

M85 is an early-type lenticular galaxy in Coma Berenices, the northernmost member of the Virgo cluster. It has a distance of approximately 60 million light-years and is very similar to M84 in many respects: it also has an old yellow population and a fairly luminous disk.

The most interesting aspect of M85 is its x-ray properties. Unlike most of the other early-type Virgo galaxies, it is very faint in x-rays and its spectrum seems to consist of two parts—a hard (i.e. more energetic) part possibly caused by a population of evolved stellar sources such as low-mass x-ray binaries, and a soft (i.e. less energetic) part that may be due to hot interstellar medium. This is unusual because the x-ray spectrum of a galaxy like this is normally dominated by emission from the interstellar medium.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M86

M86 is a lenticular or elliptical (classification is not entirely clear) galaxy in Virgo, one of the largest in the Virgo cluster. It has a conspicuous system of faint globular clusters and a very small dwarf companion. Since it lies at the heart of the Virgo cluster, its light is actually blue-shifted with respect to us (i.e. the galaxy is moving towards us), probably due to infall into the Virgo group from the far side.

M86 has a faint plume of x-ray emission extending out from the optical shape of the galaxy. It has been speculated that this is due to ram pressure of the intercluster medium which is stripping the galaxy of its gas. In addition to the plume, a faint optical distortion has been detected, probably due to star formation caused by cooling in the hot gas that is being stripped. This has caused a great amount of interest in the galaxy since it offers a rare chance to observe star formation in a cooling flow environment. M86 forms a small group together with M84.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M87: The Massive Galaxy

Messier 87 (NGC 4486) is a giant elliptical galaxy which dominates the core of the nearby VIRGO CLUSTER of galaxies at a distance of about 16 Mpc from the Sun. It is the archetypal example of the class of brightest cluster galaxies which dominate the centers of many rich concentrated clusters and, with a total visual magnitude of about 9.6, it is about 4 times brighter than our own Milky Way Galaxy would appear at the same distance. In addition to a striking visual appearance (see figure 1), it is also famous as the host galaxy of the bright radio source VIRGO A and the spectacular jet associated with its active galactic nucleus (see M87: THE NUCLEUS AND JET). M87 is in fact not the most luminous galaxy in the Virgo cluster; the bright elliptical Messier 49 (NGC 4472) located some 4.5° away is ∼0.2 magnitudes (22%) brighter. However, M87 is the centrally dominant galaxy, and it is this privileged position which has made it the subject of many different studies.

Like most ELLIPTICAL GALAXIES, M87 exhibits a smooth distribution of light decreasing rapidly away from the center. Near the middle it appears almost round, but on very deep images it looks noticeably flattened with an apparent axial ratio of about 0.6. The intrinsic shapes of elliptical galaxies are difficult to determine and, unlike spiral galaxies, most do not contain extended disks of rotationally supported cold hydrogen gas with which to determine their masses. From an observational point of view, therefore, the question of whether elliptical galaxies are surrounded by massive halos of DARK MATTER like their disk galaxy counterparts is still open to debate (see also DARK MATTER IN GALAXIES).

There are three main methods which have been used to constrain the extended mass distributions in elliptical galaxies such as M87. These are (i) analysis of the mean streaming motions and velocity dispersions of the integrated starlight, (ii) the use of globular clusters as tracers of the gravitational potential and (iii) the analysis of temperature and brightness profiles for hot x-ray emitting gas surounding the galaxy.

Stellar kinematics

Spectroscopic studies of the integrated light from galaxies can yield important information on the motions of stars within the galaxy, from analysis of the positions, widths and shapes of absorption features in the spectra. For nearly spherical galaxies such as M87, the interpretation of these kinematical data in terms of a dynamical model is usually carried out within the framework of the spherically symmetric Jeans equation:

$$\frac{1}{\rho}\frac{\mathrm{d}(\rho\sigma_\mathrm{r}^2)}{\mathrm{d}r} + 2\frac{\beta\sigma_\mathrm{r}^2}{r} = -\frac{GM(r)}{r^2}$$

where $\rho(r)$ is the three-dimensional density of stars at a distance r from the centre of the galaxy, $\sigma_\mathrm{r}(r)$ is the root mean square random velocity (known as the velocity dispersion) in the radial direction, $\beta(r) = 1 - \sigma_\mathrm{t}^2/\sigma_\mathrm{r}^2$ is the velocity anisotropy which measures the relative importance of radial and tangential random motions and $M(r)$ is the total (luminous and dark) mass enclosed within radius r. Unfortunately this problem is underdetermined, given the number of observable properties, and some assumptions need to be made. The simplest of these is to assume that the random motions at any point within the galaxy are isotropic, when $\beta = 0$. Early observations of the core of M87 revealed a velocity dispersion profile which rose toward the center, a feature that was initially attributed to the presence of a massive BLACK HOLE. However, it was soon demonstrated that models without a central point mass were equally acceptable if the velocity dispersions became radially elongated towards the center. In fact, the most recent stellar kinematical data reveal that the velocity dispersion increases only slowly within the inner few arcseconds of M87. To reconcile these observations with the $3 \times 10^9 M_\odot$ black hole inferred from Hubble Space Telescope studies of ionized gas within 1″ of the center requires that the stellar orbits in this region are in fact nearly circular. This has been interpreted as evidence that the center of M87 once contained a binary black hole which preferentially depletes stars on low-angular-momentum orbits via three-body encounters, leaving a largely tangential velocity dispersion.

Outside the core, the velocity dispersion continues to fall steadily with increasing radial distance. However, the stellar absorption-line spectra are very hard to observe

Figure 1. The massive galaxy M87. The swarm of pointlike objects surrounding the galaxy are its rich population of globular clusters. **This figure is reproduced as Color Plate 36.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

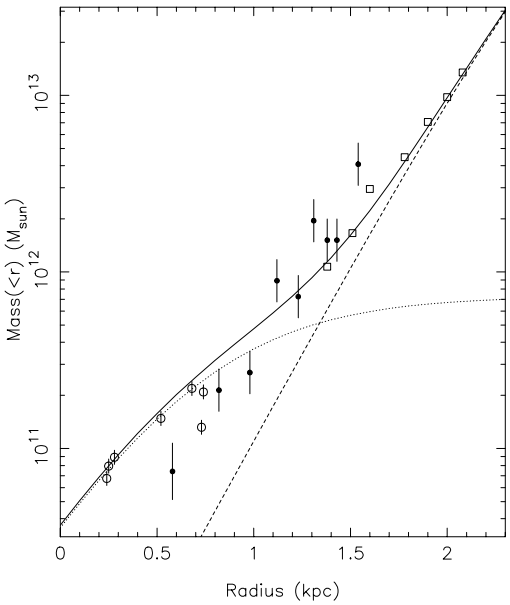


Figure 2. Enclosed mass as a function of radius from the center of M87. Open circles show the estimates from integrated light studies, filled circles from the globular clusters and open squares from the hot x-ray gas. The dotted curve shows a constant M/L model for the mass distribution due to the stellar halo. The dashed line shows a model fit to the extended dark halo whose density profile has been derived from cosmological N-body simulations. The total mass profile is shown by the solid curve.

with useful signal-to-noise beyond $r \sim 2'$ (10 kpc for a Virgo distance of 16 Mpc), even though the integrated light profile can be followed to beyond 100 kpc. The best estimate of the total stellar mass in M87 is $M \sim 4 \times 10^{11} M_\odot$ within 10 kpc of the center.

A useful measure of the presence of dark matter in galaxies is the 'mass-to-light' (M/L) ratio between the luminosity due to stars (at some particular wavelength) and the inferred mass. For M87 the V-band mass-to-light ratio derived from integrated light studies is $M/L \sim 10$ times that of the Sun. This is close to the upper end of the range of M/L values seen in other elliptical galaxies; extrapolating out to the limits of the stellar halo assuming a constant M/L ratio (i.e. that the mass in the halo traces the starlight) would imply a total mass for the M87 galaxy of $M \sim 9 \times 10^{11} M_\odot$ (figure 2). While the central black hole has a mass which is clearly negligible on these scales, it can still affect the global structure of the galaxy via the scattering and redistribution of stellar orbits on rare passages close to the galaxy nucleus.

Globular clusters

At radii beyond about 10 kpc, there are two independent methods to constrain the global mass distribution of M87. The first uses kinematical studies of the large population of GLOBULAR CLUSTERS which surrounds M87 (see figure 1) as test particles to probe the associated gravitational potential. Globular clusters are compact star clusters of $\sim 10^5$ stars which are found ubiquitously in all types of galaxy, but M87 is atypical in having an unusually large population of more than 10 000 clusters. Although their combined masses contribute only a few per cent of the total mass of the halo, because of their relatively bright magnitudes (more than 50% of the M87 clusters have an apparent magnitude brighter than $m_v \sim 23.5$) their radial velocities can be obtained using large ground-based telescopes. Spectroscopic studies have now been made of more than 200 clusters around M87. The most striking result of these measurements is that the velocity dispersion of the globular clusters *increases* with radius. Dynamical analyses of the kinematics of the globular cluster system can be made using simple projected mass estimators or using the Jeans equation. Both of these methods indicate that M87 is surrounded by an extended dark halo with an enclosed mass which increases approximately as $M(< r) \sim r^{1.7}$. The inferred mass-to-light ratio in solar units ranges from $M/L \sim 10$ at $r = 10$ kpc to $M/L \sim 50$ at $r = 30$ kpc.

X-ray gas

Very extended x-ray emission from hot $T \sim 10^7$ K gas around M87 has been detected with the Position Sensitive Proportional Counter (PSPC) on the ROSAT x-ray satellite. Assuming that the gas is a single-phase medium in hydrostatic equilibrium, and that the underlying gravitational potential is spherically symmetric, the mass distribution in the halo can be obtained uniquely from the run of x-ray surface brightness and temperature with radius. Although the x-ray emission can be detected out to over 300 kpc from the centre of M87, temperature estimates can only be derived for the central 200 kpc. The density and temperature of the gas are related to the mass distribution by

$$M(< r) = -\frac{krT}{G\mu m_p}\left(\frac{\mathrm{d}\log n_e}{\mathrm{d}\log r} + \frac{\mathrm{d}\log T}{\mathrm{d}\log r}\right)$$

where μ is the mean molecular weight of the gas which depends weakly on its chemical composition, n_e is the electron density and the physical constants have their usual meanings. The electron density can be obtained straightforwardly from the x-ray surface brightness distribution. Obtaining the temperature profile is more problematic, but useful limits can be obtained from the energy sensitivity of the PSPC detectors. Single-temperature fits to $T(r)$ indicate that the temperature rises from $kT \sim 1.5$ keV at the center to $kT \sim 4$ keV at 200 kpc. The implied mass distribution is shown in figure 2 and agrees well with that determined from the globular cluster kinematics where they overlap, rising to $M(< r) \sim 10^{13} M_\odot$ within $r = 100$ kpc. Although more recent temperature measurements with the ASCA satellite indicate that the x-ray halo probably contains more than one temperature component, the close agreement between the mass distributions derived from these two methods,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which rely on different physical assumptions, suggests that the uncertainties are unlikely to be large. Less than 5% of this mass is due to the x-ray emitting gas itself, the remainder residing in a dark matter halo.

Conclusions

M87 is one of the best-studied elliptical galaxies and a good test case for formation theories of brightest cluster galaxies. The rise in the globular cluster velocity dispersion profile and the temperature and density profiles of the x-ray gas both indicate that the mass density is dominated by a dark halo component beyond $r \sim 10$ kpc. Whether such extensive dark matter halos are ubiquitous in elliptical galaxies, particular those outside of rich clusters, is an open question, since the position of M87 at the center of the Virgo cluster makes it difficult to separate the galaxy halo potential from that of the Virgo cluster as a whole. Studies of the global gravitational potential of the Virgo cluster using galaxy radial velocities suggest that the mass distribution continues to rise steeply as shown in figure 2 out to at least $r \sim 1$ Mpc.

Bibliography

Cohen J G and Rhyzov A 1997 The dynamics of the M87 globular cluster system *Astrophys. J.* **486** 230–41

McLaughlin D E 1999 Evidence in Virgo for the Universal Dark Matter Halo *Astrophys. J.* **512** L9–12

Nulsen P E J and Böhringer H 1995 A ROSAT determination of the mass of the Central Virgo Cluster *Mon. Not. R. Astron. Soc.* **274** 1093–106

Ray M Sharples

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M87: The Nucleus and Jet

The giant elliptical galaxy M87 at the center of the VIRGO CLUSTER is one of the clearest and nearest examples of an ACTIVE GALACTIC NUCLEUS (AGN) and therefore has been the subject of a large number of studies over the last three decades (see also M87: THE MASSIVE GALAXY). The discovery by Curtis (1918) of an extended prominence emanating from what appeared to be an otherwise normal ELLIPTICAL GALAXY remained little more than a curiosity until Baade and Minkowski (1954) carried out the first thorough studies of what is now known as the M87 optical jet. The proposal by Blandford and Rees (1974) that jets must play a fundamental role in transporting energy from the nuclear energy source to the extended and distant radio lobes established a sound physical underpinning to further investigations of the AGN phenomenon. The presence of massive BLACK HOLES at the center of galaxies is widely believed to be the common origin of the AGN phenomenon. The black hole model is very appealing because it provides an efficient mechanism that converts gravitational energy, via accretion, into radiation within a very small volume as required by the rapid variability of the large energy output of AGNs (e.g. Blandford 1991).

The AGN standard model comprises a central black hole, with mass in the range $\simeq(10^6$–$10^9)M_\odot$, surrounded by an ACCRETION DISK that releases gravitational energy. The radiation is emitted thermally at the local blackbody temperature and is identified with the 'blue bump', which accounts for the majority of the bolometric luminosity in the AGNs. The disk possesses an active corona, where infrared SYNCHROTRON RADIATION is emitted along with thermal bremsstrahlung x-rays. The host galaxy supplies this disk with gas at a rate that reflects its star formation history and, possibly, its overall mass, thereby accounting for the observed luminosity evolution. Broad emission lines originate homogeneously in small gas clouds of density $\simeq 10^9$ cm^{-3} and size $\simeq 1$ AU in random virial orbits about the central continuum source. Plasma jets are emitted perpendicular to the disk. At larger radii, the material forms an obscuring torus of cold molecular gas. Orientation effects of this torus to the line of sight naturally account for the differences between some of the different classes of AGNs. While this broad picture has been supported and refined by a number of observations, direct evidence for the existence of accretion disks around SUPERMASSIVE BLACK HOLES is sparse and detailed measurements of their physical characteristics are conspicuous by their absence.

The black hole in M87

Ground-based observations of M87 first revealed the presence of a cusplike region in its radial light profile accompanied by a rapid rise in the stellar velocity dispersion and led to the suggestion that it contained a massive black hole (Young *et al* 1978, Sargent *et al* 1978). Stellar dynamical models of elliptical galaxies showed, however, that these velocity dispersion rises did not necessarily imply the presence of a black hole but could instead be a consequence of an anisotropic velocity dispersion tensor in the central 100 pc of a triaxial elliptical potential. Considerable controversy has surrounded this and numerous other attempts to verify the existence of the black hole in M87 and other nearby giant ellipticals using ground-based stellar dynamical studies.

One of the major goals of the HUBBLE SPACE TELESCOPE (HST) has been to establish or refute the existence of black holes in active galaxies by probing the dynamics of AGNs at much smaller radii than can be achieved from the ground. HST emission line imagery of M87 has led to the discovery of a small-scale disk of ionized gas surrounding its nucleus which is oriented approximately perpendicularly to the synchrotron jet. This disk is also observed in both the optical and the UV continuum. Similar gaseous disks have also been found in the nuclei of a number of other massive galaxies, e.g. Cen A. Work by Macchetto and collaborators (1997) using the Faint Object Camera (FOC) on the HST has shown unambiguously for the first time that there is a thin Keplerian disk extending to a distance of at least 1″ or 73 pc from the center. By measuring the rotation curve from 1″ (73 pc) down to 0.″06 (3.5 pc) they could derive a number of key parameters for the disk, notably a central mass concentration of $(3.2 \pm 0.9) \times 10^9 M_\odot$ within a sphere of less than 3.5 pc radius and an inclination of the disk to the line of sight of 41°, and showed that the most likely and simplest explanation for this huge mass is that it is a supermassive black hole.

The extended disk

At distances of 1″–2″, or 73–146 pc, the disk morphology shows a three-arm spiral pattern and at larger distances yet the gaseous structure becomes less and less organized until it appears to connect with an extended (17″ or 1200 pc) system of filaments which show prominent emission notably in Hα. This prompted Sparks and collaborators (1993) to suggest a causal relationship between the extended filaments and the inner disk and propose that the filaments are streamers of gas flowing away from the nucleus, rather than falling into it. To date this scenario has had some supporting evidence from measurements of blueshifted broad absorption lines in the nuclear spectrum, which can be interpreted as implying both turbulence and an outflow of material away from the nucleus. A possible model to explain these lines and the observed emission line filaments is that they are the result of a bi-directional wind generated in the accretion disk; this wind also removes angular momentum from the disk to allow accretion through the disk onto the black hole.

The jet

Emanating from the nucleus is the jet which shows considerable structure, with a sequence of bright features reaching ∼25 arcsec from the galaxy center. A counterjet is not detected either at optical or at radio wavelengths, although imaging polarization measurements of the south-eastern radio lobe (Sparks *et al* 1991) found a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

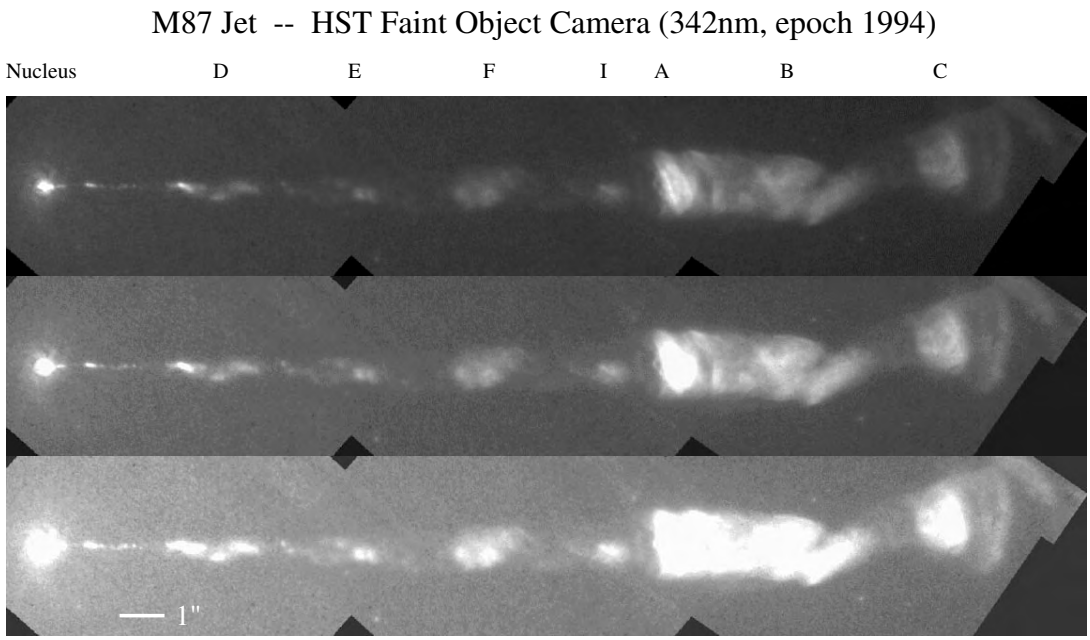


Figure 1. HST FOC image of the jet observed at wavelength 342 nm in 1994 with 0.03 arcsec resolution. Each image is a mosaic of three different FOC pointings. The three panels show the same image at different contrast settings. Knots are labeled with usual letter designations.

polarized counterlobe, thus providing evidence for the existence of an invisible counterjet. Within observational limitations it has been recognized that the detailed morphology of the jet is similar in the radio and optical bands and consistent with those in the infrared and x-ray bands. VLBI mapping shows the nucleus to have an asymmetric, complex 'core–jet' structure extending ∼0.2 arcsec and subluminal motion with velocity $\sim 0.3c$ has been detected for a bright component in the nuclear jet. The jet is highly polarized and it is generally accepted that it emits through the synchrotron process. A fundamental question posed by the observations of the jet is where the synchrotron-radiating electrons are accelerated. Are they accelerated in the nucleus and transported to the sites of the emission? The very short lifetimes of the electrons emitting in the optical to x-ray regions in a uniformly filled jet compared with the light travel time from the nucleus to the end of the observed jet appears to rule this out and suggest that *in situ* particle reacceleration occurs in locations along the jet.

High-resolution images of the jet taken in the UV with the FOC reveal a wealth of details. Figure 1 shows the entire jet as seen by the FOC at 0.03 arcsec resolution (∼2.4 pc). Many structural details of the classical knots are readily apparent, as are many new compact structures. The first 200 pc of the jet—roughly the region between the nucleus and knot D—appears as a narrow chain of numerous bright, unresolved knots. The opening angle of the jet as defined by this chain of features is extremely narrow, less than about 1°, whereas a value near 6.5° is apparent for the rest of the jet beyond knot D. There are additional compact features between knots D and E, as well as bright, slightly resolved condensations (scale ∼0.1 arcsec) throughout the knots D, E, F and I of the 'inner' jet.

The FOC images also reveal many fine structures within the knots that are apparent for the first time. Knot A, in particular, is the brightest knot in the jet and is thought to represent a shock in flow. The eastern edge of knot A also shows several faint structures which are symmetric about the jet axis and which must be related to the impending 'shock' in the bulk flow at knot A. There is a faint 'bar' of emission centered on the jet axis and very nearly perpendicular to it; this is a shock in the high-speed flow at the jet center. There are also faint 'fans' of emission that start at the jet edge adjacent to the bar and then trail inwards towards the jet axis and a faint loop or 'cap' of emission precisely on the jet axis before the brightest region of knot A. These features are all related to shocking of the bulk flow at knot A, and represent conical sheets seen in projection, and are similar to those predicted in numerical models.

A comparison of the FOC and VLA (VERY LARGE ARRAY) images taken at nearly the same epoch in 1994 reveals systematic differences all along the jet. In the radio image, there is faint diffuse emission which appears to fill the jet from the nucleus to knot A and defines the edges of a cone with an opening angle of about 6.5°. While traces of interknot emission are also present in the FOC image, it is much fainter relative to the knot emission. In addition, the optical emission in the knots is more condensed along the jet axis. There are other detailed differences in knot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

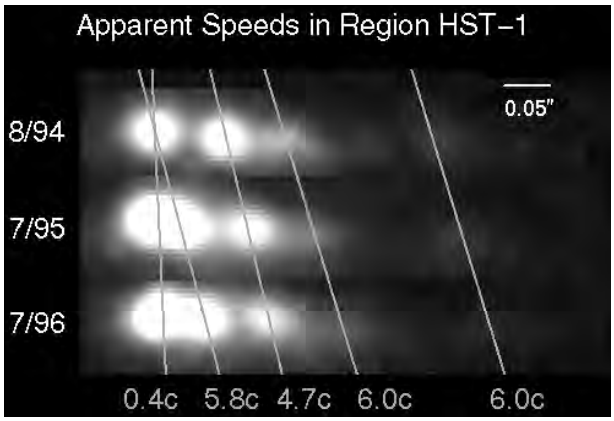


Figure 2. Sequence of HST FOC images of jet region HST-1 which is located about 1 arcsec from the nucleus. The images are observed at $\lambda = 342$ nm at epochs August 1994, July 1995, and July 1996. The slanting lines connect features between epochs, with apparent speeds given in units in c, the velocity of light.

structure, especially in and around knot E. The optical images show a bright pair of condensations in knot E, but these are only a relatively faint region in the radio image. Similarly there is a pair of unresolved features in the optical image between knots D and E, which again are not apparent in the radio image.

Polarimetry

High levels of polarization are a key characteristic of synchrotron radiation. POLARIMETRY provides clues to the direction of the magnetic field within the jet and hence is one of the most important diagnostics for deciphering its structure. The radio polarization has been well studied and it was found that the field is primarily parallel to the jet axis, except in the bright transverse features in knots A and C, where the B field runs along the transverse feature (i.e. normal to the jet axis). The parallel field in most of the jet can be understood as arising from shearing of a random field, while the perpendicular fields could arise from compression at shocks.

Optical polarimetry is especially interesting, because it can potentially tell us about the three-dimensional structure of the jet (Perlman *et al* 1999). The optical emission is more concentrated on the jet axis, and hence it provides an opportunity to obtain information about the magnetic structure at different depths within the jet. We observed typical fractional polarizations that are in the range from 30% to 50% and are comparable with those in the radio band. The magnetic field is generally aligned along the jet, except in the bright transverse features in knots A and C, and at the upstream ends of HST-1 and knot D, where the field runs perpendicular to the jet axis. There is evidence for radio–optical polarization differences in knot E and in the knot A 'bar' region which are attributed to radial structure in the jet.

Proper motion

A 5 yr search for secular variations in the jet—including structural changes, flux variations, proper motions and changes in polarization—is under way (Biretta *et al* 1999). This work can give important clues to the velocity field within jets, about which very little is known from direct observation. Monitoring with the VLA at radio frequencies has detected motion in much of the jet with apparent speeds averaging near $0.5c$. The very fastest regions of these wavelengths are in knot D and appear to move outward with speeds near $2.5c$. At optical wavelengths one of the most active regions in the jet is the complex of compact knots about $1''$ (73 pc) from the nucleus designated HST-1. Figure 2 shows a sequence of FOC images from 1994 to 1996. The complex contains a bright eastern component which moves outward slowly ($0.4c$) and which appears to emit bright compact knots at speeds in the range from $4c$ to $6c$. The emitted components rise and fade within 1 yr. The eastern component is a slow-moving disturbance or shock in the flow, while the fast components are representative of the jet fluid speed. Similar speeds are seen in knot D (figure 3), with measured outward velocities in the range from $2.6c$ to $5.0c$, and there is rapid fading (half-life <1 yr) in features FOC3 and FOC4, which are located just east of knot D. The large speeds have a number of interesting implications within the context of the standard relativistic beaming model. The largest speeds give the strongest constraints, and the observed speed of up to $6c$ implies motion with typical Lorentz factors $\theta = 18^\circ$ and $\gamma = 12$, which are quite consistent with those predicted by the unified models for FR-I sources.

Conclusions

The study of the jet and nuclear regions of M87 has provided an unprecedented laboratory to test many of the recurrent theories of AGN and will continue to give us important clues for the understanding of the relevant physical processes.

Bibliography

Baade W and Minkowski R 1954 *Astrophys. J.* **119** 215
Biretta J A, Sparks W B and Macchetto F D 1999 *Astrophys. J.* **520** 621
Blandford R D 1991 *Proc. Heidelberg Conf. Physics of AGN* ed W J Duschi and S J Wagner (Berlin: Springer) p 3
Blandford R D and Rees M J 1974 *Mon. Not. R. Astron. Soc.* **169** 395
Curtis H D 1918 *Publ. Lick Observatory* **13** 31
Macchetto F R, Marconi A, Axon D E, Capetti A, Sparks W B and Crane 1997 *Astrophys. J.* **489** 579
Perlman E S, Biretta J A, Zhou F, Sparks W B and Macchetto F D 1990 *Astron. J.* **117** 2185
Sargent W L W, Young P J, Boksenberg A, Shortridge K, Lynds C R and Hartwick F D A 1978 *Astrophys. J.* **221** 731
Sparks W B, Ford H C and Kinney A L 1993 *Astrophys. J.* **413** 531

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

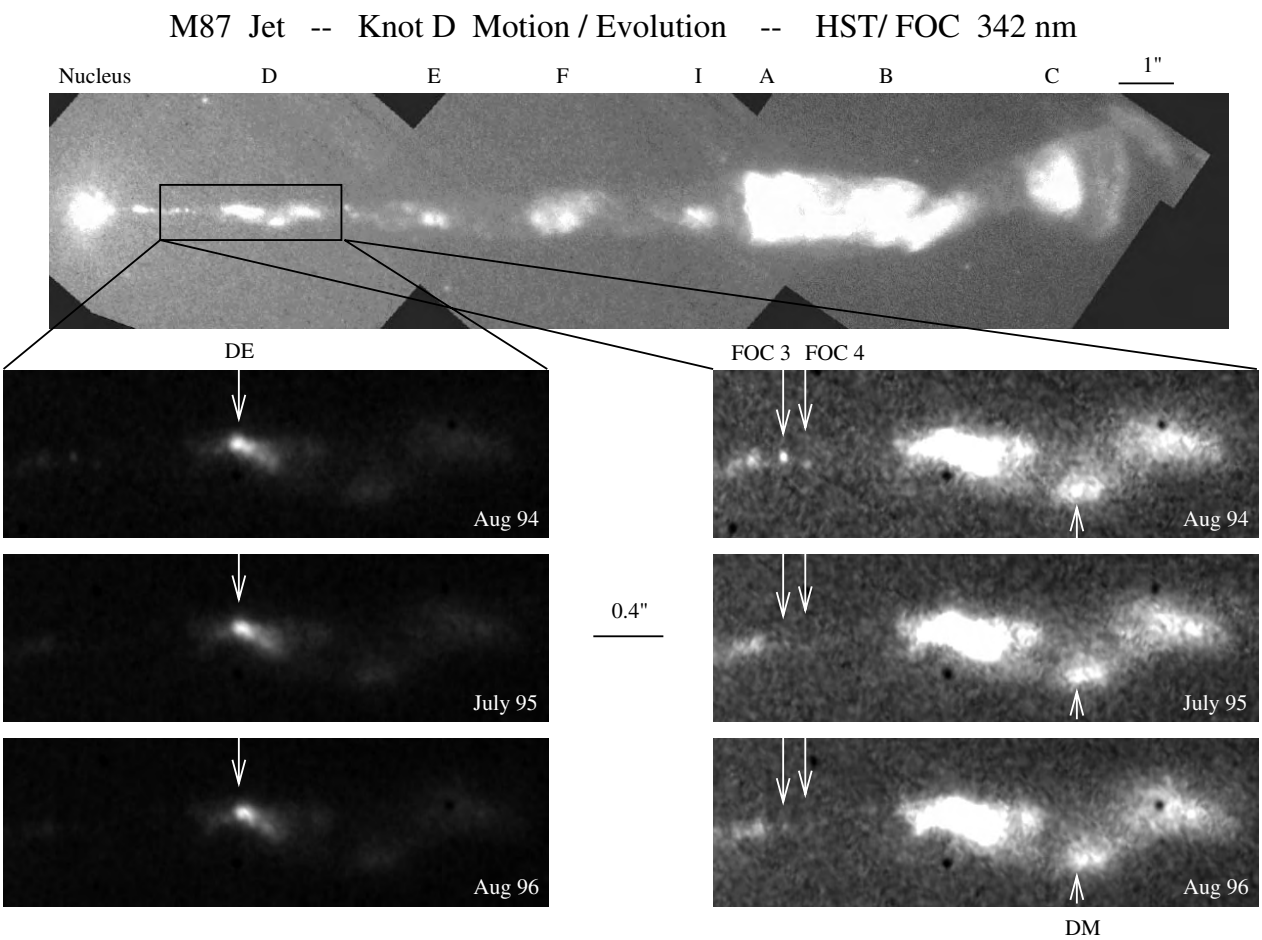


Figure 3. Knot D region at three epochs. Top panel shows entire jet to provide orientation; lower panels show knot D at three epochs with two different contrast settings. Arrows are at fixed locations relative to galaxy nucleus and can be used to judge motions. Features DE and DM (labeled) move outward at 4*c*–5*c*, while FOC3 and FOC4 fade from view.

Sparks W B, Fraix-Burnet D, Macchetto F D and Owen D N 1991 *Nature* **355** 804

Young P J, Westphal J A, Kristian J, Wilson G P and Landauer F T 1978 *Astrophys. J.* **221** 721

D Macchetto

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M88

M88 is a spiral galaxy in Coma Berenices at a distance of between 12 and 17 Mpc. It is also a member of the Virgo cluster. The galaxy falls into the category of flocculent spirals, which means that it has a stochastic spiral structure in its stellar disk with a faint two-arm structure in the infrared indicating a possible underlying density wave. Its most interesting feature is the chemically decoupled nucleus, i.e. the galaxy has a chemically distinct, metal-rich nucleus of a distinctly different age from the bulge. Additionally, the observation of secondary star-formation bursts may indicate the existence of several more chemically distinct stellar subsystems. The origin of this phenomenon is not known.

In 1999 a supernova was observed in M88.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M89

M89 is an elliptical galaxy in Virgo at a distance of about 60 million light-years. It is a member of the Virgo cluster.

In the optical, M89 does not show any unusual features but in the radio wavelengths it has a faint enveloping structure as well as a jetlike extension reaching about 10′ out from the center. The radio core is of low luminosity but variable. It also emits strongly in the UV—UV light is usually associated with direct non-thermal radiation or dust heated by stars, but both of these are absent from the optical data. So far, no theory has been proposed that could explain all of these features, but the Hubble Space Telescope recently discovered a very small AGN (active galactic nucleus) at the center of this galaxy. It is thought to be a modest accretion event onto a black hole and represents the faintest AGN known so far.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M9

M9 is a globular cluster in the constellation of Ophiuchus (3 degrees south and 2 degrees east from η Ophiuchi), only 5500 light years from the Galactic Center. It has a visual brightness of 7.3 mag which is partly because of its great distance from the Sun (26 000 light-years) but also because it lies at the edge of a dark nebula and is hence dimmed by dust to the north and west. The cluster is slightly elliptical in outline and has a diameter of 49 light-years.

Walter Baade investigated M9 for variable stars and found just 13, 10 of which have periods around 2.4 days.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M90

M90 is one of the larger spirals in the Virgo cluster but it has little mass. It also has one of the largest velocities of approach of all the Messier objects and may possibly be in the process of escaping from the Virgo cluster. Its spiral arms are tightly wound and smooth but, unusually for a spiral, there is no ongoing star formation in the arms but only in the inner disk region. The galaxy may currently be evolving into a state similar to M64.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M91

M91 in Coma Berenices was long regarded as one of the 'missing' Messier objects, since nothing nebulous could be seen near its listed position. It was even suggested that Messier had really seen a comet at this position in March 1781 but not recognized its true nature. In 1969 a Texan amateur, William C Williams, gave what is now accepted as the probable answer: Messier originally measured the position of this object through his telescope by reference to M89, but when he came to write up his observations he mistakenly applied the correction to the coordinates of M58, another galaxy in this area. The resulting position coincides with that of NGC 4548, a tenth-magnitude barred spiral which would have been within reach of Messier's telescope. If M91 is indeed the same as NGC 4548, it is a member of the Virgo cluster, but is fairly unremarkable visually in small amateur instruments.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M92

M92 is an exceptionally ancient globular cluster in Hercules which would be better known if it were not overshadowed by M13. It was discovered by Johann Bode in 1777 and rediscovered in 1781 by Messier, although since it is bright enough to be just visible to the naked eye under ideal conditions it is surprising that it was not recognized earlier. As in most globulars, almost all the known variables in M92 are short-period pulsating variables of the RR Lyrae type, although in this case there is also a rarity: an eclipsing binary of W Ursae Majoris type. A study of M92 by H Arp, W Baum and A Sandage found that the most luminous members of the cluster are red giants with absolute magnitudes of around −3. Visually, in small instruments, M92 has a very condensed center with a surrounding coma like the head of a comet. Analysis of results from the Hipparcos satellite gives a distance for M92 of about 29 000 light-years, slightly farther than M13, and an estimated age of 14 billion years, making it the oldest globular known.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M93

M93 is an open cluster situated almost exactly on the galactic equator in Puppis. Of sixth magnitude, it is at the limit of naked-eye visibility and is easily detectable with binoculars, near third-magnitude Xi Puppis. It is elongated, and through small telescopes appears shaped somewhat like the tip of a knife, although Admiral Smyth called it starfish-shaped. M93 was discovered by Messier in March 1781, and described by him as a 'cluster of small [i.e. faint] stars without nebulosity'. Modern measurements identify the brightest stars as blue giants of eighth magnitude, among a total of some 60 stars covering a diameter of about 18 arc minutes. However, it is difficult to determine how far the outer reaches really do extend; published diameters for the cluster range from only 8 arc minutes (Messier's own estimate) up to 24 arc minutes. Modern distance determinations put the cluster at 3600 light-years from us, giving it a true diameter of about 20 light-years, and its age is some 300 million years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M94

M94 is a spiral galaxy in Canes Venatici, lying just north of a line between Alpha and Beta Canum Venaticorum. The galaxy is presented almost face-on to us and in small telescopes appears almost perfectly rounded with a bright nucleus; Admiral Smyth understandably called it 'comet-like'. It is classified as midway between type Sa and Sb, meaning that its arms are quite tightly wound. Small telescopes reveal traces of its spiral arms,which contain many patchy dust clouds in their inner regions. Ultra-violet observations have revealed a ring of star formation over 7000 light-years in diameter around this galaxy's core. Long-exposure photographs with large instruments reveal an outer ring of light as well, both rings providing evidence for past activity in the galaxy. M94 was discovered in March 1781 by Mechain, as were all the subsequent objects in Messier's catalog. M94 is around 15 million light years away.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M95

M95 is a barred spiral galaxy in Leo, under the body of the lion about a third of the way from Alpha to Beta Leonis. It was discovered in March 1781 by Mechain but at tenth magnitude is faint and difficult for small telescopes. M95 is presented almost face-on to us, displaying a central bar and tightly wrapped spiral arms that together create the shape of a Greek theta. M95 forms a pair with nearby M96, usually regarded as lying around 25 million light-years away, although observations with the Hubble Space Telescope in 1995 suggest a distance about 50% greater.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M96

M96 is a spiral galaxy in Leo, lying about two-thirds of a degree to the east of M95. Like that galaxy, it was discovered in March 1781 by Mechain and was then checked by Messier before inclusion in his catalog. Visually, it appears rounded with a bright center and traces of spiral arms; it is classified as type Sa. A supernova of type Ia erupted within it in 1998. M96 lies at a similar distance to M95, although estimates of the distances are discrepant. A figure of around 25 million light-years has usually been adopted, but observations of Cepheid variables in M96 with the Hubble Space Telescope have yielded a distance closer to 40 million light-years. Whatever the true distance, these galaxies, and M105 to the north, are seemingly part of a small cluster, the Leo Galaxy Group, that also includes M65 and M66.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M97

M97 is a planetary nebula in Ursa Major, lying just south of the bowl of the Big Dipper, 2.25 degrees from Beta Ursae Majoris. This nebula was discovered in February 1781 by Mechain, who noted it was difficult to see; modern observers estimate it as either tenth or eleventh magnitude, placing it among the faintest objects in Messier's catalog. Lord Rosse in 1848 named it the OWL NEBULA because its rounded outline with two dark patches gave the impression of an owl's face. Rosse claimed at the time that a star was visible in each cavity, like a pair of eyes, but the fainter of these subsequently vanished; perhaps it was not a true star at all but a brighter knot of nebulosity. There is, though, a central star, listed variously as fourteenth or sixteenth magnitude, which is the source of the nebula's gas. Visually through small telescopes the Owl appears as a pale, grey disk larger than Jupiter, although larger instruments (and long-exposure photographs) reveal a greenish colour. As with all planetary nebulae its distance is uncertain, but is probably in the region of 1500 light-years, somewhat more distant than M27 but still among the closest planetaries to us.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M98

M98 is a spiral galaxy in southern Coma Berenices, just 0.5 degrees from the fifth-magnitude star 6 Comae Berenices. It was a Mechain discovery, in March 1781. M98 is a spiral galaxy of intermediate type lying nearly edge-on to us, as a result of which it appears extremely elongated in shape, about four times longer than it is wide. William Herschel in 1783 declared that 'its feeble branches extend above a quarter of a degree', although only large telescopes will show it this big; in amateur instruments its length will be about half as large as Herschel saw it, and the nucleus appears star-like. Although lying in Coma, M98 is actually a member of the Virgo Cluster. One surprising aspect is that it shows not a redshift but a blueshift, indicating it is moving towards us at a relative speed of 140 km s^{-1}. In fact, it is not the only Virgo cluster member to show a blueshift, nor does it have the largest value; rather,these discordant figures demonstrate that there is a wide range of motions among the individual members of large clusters such as that in Virgo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

M99

M99 is a spiral galaxy in Coma Berenices. It was another Mechain discovery in March 1781, on the opposite side of the fifth-magnitude star 6 Comae Berenices from M98, but the two galaxies are of quite different appearance. Whereas M98 is edge-on to us, M99 is turned almost face-on so that it appears rounded in small instruments. It is classified as type Sc. Unusually, it seems to possess three spiral arms (a symmetrical two is the most common arrangement), although these will be visible only in larger apertures or on long-exposure photographs. They were first seen by Lord Rosse in 1848, making this the second galaxy in which spiral structure was definitely detected (M51 was the first), although of course at that time the true nature of galaxies was not recognized. M99 has been the site of three known supernovae, in 1967, 1972 and 1986. In contrast with M98, which shows a small blueshift, M99 has one of the largest redshifts of any galaxy in the Virgo Cluster, 2400 km s^{-1}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mach, Ernst (1838–1916)

Moravian/Austrian physicist, scientist/philosopher, the basis of whose natural philosophy was that all knowledge is a matter of experiments, indeed sensations, so that 'laws of nature' are summaries of experience provided by fallible senses. EINSTEIN, who had been taught by Mach, was very influenced by this perspective and developed the theory of relativity, expressing much the same thing by incorporating the speed of light into concepts of simultaneity, etc. *Mach's Principle* is a philosophical statement that in some way the properties of a body are influenced by the properties of all the other bodies of the universe. This is not easy to interpret scientifically, but the principle influenced Einstein's concept of space-time. Contrary to Einstein's belief that he had produced the theory of general relativity in conformance with this principle, WILLEM DE SITTER found solutions to Einstein's field equations in the absence of matter, i.e. no matter exists other than the test body, but it does have inertia. Mach also worked on acoustics, and the ratio of a body's speed to the speed of sound is known as the *Mach number*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maclaurin, Colin (1698–1746)

Born in Kilmodan, Argyllshire, Scotland, became professor of mathematics at the University of Edinburgh and worked on the theory of tides, the impact of bodies, and in 1742 published his *Treatise of Fluxions*, the first systematic exposition of NEWTON's methods. Its influence was negative in that it prolonged the currency of Newton's geometric approach to calculus rather than the algebraic analysis developed on the continent. In the book, he used the special case of Taylor's series now named the *Maclaurin series*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maclear, Thomas (1794–1879)

Irish astronomer, verified Henderson's parallax of alpha Centauri. Trained as a doctor in England, Maclear was an amateur astronomer who became an FRS on the basis of his contributions to astronomy, and then was appointed to the Royal Observatory at the Cape of Good Hope, where he was Director from 1834 to 1870 (see SOUTH AFRICAN ASTRONOMY).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maffei 1 and 2

Two relatively nearby galaxies which lie so close to the plane of the Milky Way that they are almost completely obscured by interstellar dust. They were discovered in 1968 as a result of infrared observations conducted by the Italian astronomer, Paulo Maffei. Maffei I appears to be an elliptical galaxy, located at a distance of about 330 000 light-years. Maffei 2 is a spiral galaxy that lies at a distance of some 15 million light-years.

See also: elliptical galaxies, infrared astronomy, interstellar absorption lines, Milky Way Galaxy, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magellan

Highly successful NASA Venus orbiter, launched from Space Shuttle Atlantis in May 1989. Arrived at Venus in August 1990. Used synthetic aperture radar to map 98% of the surface of Venus at 120–300 m resolution and provided global altimetry data. Revealed more than 1000 impact craters and 1100 volcanic features. Also discovered sinuous valleys, and unique geological structures known as coronae and arachnoids. Precision radio tracking of the spacecraft measured Venus' gravitational field. Entered the planet's atmosphere in October 1994.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magellan, Ferdinand (1480?–1521)

Portuguese explorer, whose crew brought back from his voyages into the southern hemisphere news of the existence of what became known as the *Magellanic clouds*. At first, they were conceived as two fragments of the Milky Way that had somehow come loose, but now are known to be the two nearest galaxies to our own. The Magellan spacecraft to Venus commemorates his explorations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magellanic Clouds: Interstellar Phenomena

The LARGE and SMALL MAGELLANIC CLOUDS are two naked-eye IRREGULAR GALAXIES in the southern sky. Extending over $\sim 7^\circ \times 7^\circ$, the Large Magellanic Cloud has a prominent stellar bar superposed on a disk, which is tilted by about 40° with respect to the sky. The Small Magellanic Cloud, $\sim 4^\circ \times 3^\circ$, has a bar-like structure, with its long axis stretched along the line of sight. The Large and Small Magellanic Clouds are at distances of 50 kpc and 60 kpc, respectively, and are believed to be satellite galaxies of our MILKY WAY. Their masses, $2 \times 10^{10} M_\odot$ and $2 \times 10^9 M_\odot$, are much smaller than the mass of the Milky Way, a few times $10^{11}\ M_\odot$.

The Magellanic Clouds are close enough to each other and close enough to the Milky Way that tidal interactions have significantly altered the distribution of their masses. Tidal interactions are responsible for the formation of the stellar bar in the Large Magellanic Cloud as well as the elongated geometry of the Small Magellanic Cloud. A large-scale neutral hydrogen (H I) map of the Magellanic Clouds (figure 1) shows that tidal interactions have produced not only the Magellanic Bridge between the two galaxies, but also the MAGELLANIC STREAM trailing behind the Small Magellanic Cloud and a Leading Arm in the opposite direction. The Magellanic Clouds cannot be considered as independent galaxies; they form a 'Magellanic System'.

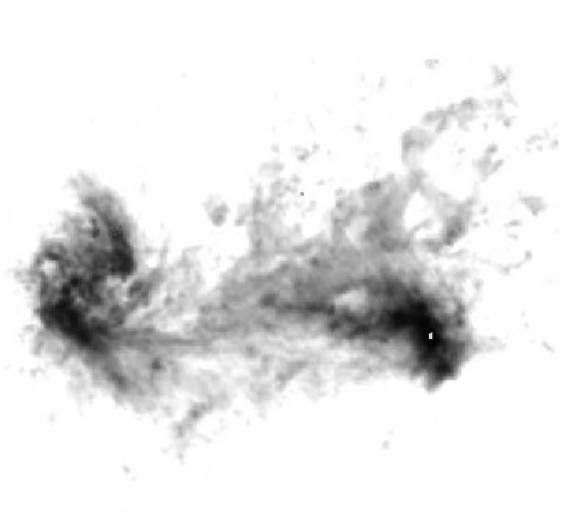


Figure 1. A neutral hydrogen view of the Large Magellanic Cloud (on the left) and the Small Magellanic Cloud (on the right). The gas connecting the two galaxies forms the Magellanic Bridge. The start of the Magellanic Stream can be seen above the Small Magnellanic Cloud. This image was obtained with the Parkes multibeam receiver of the Australia Telescope. Image courtesy of R Haynes *et al*.

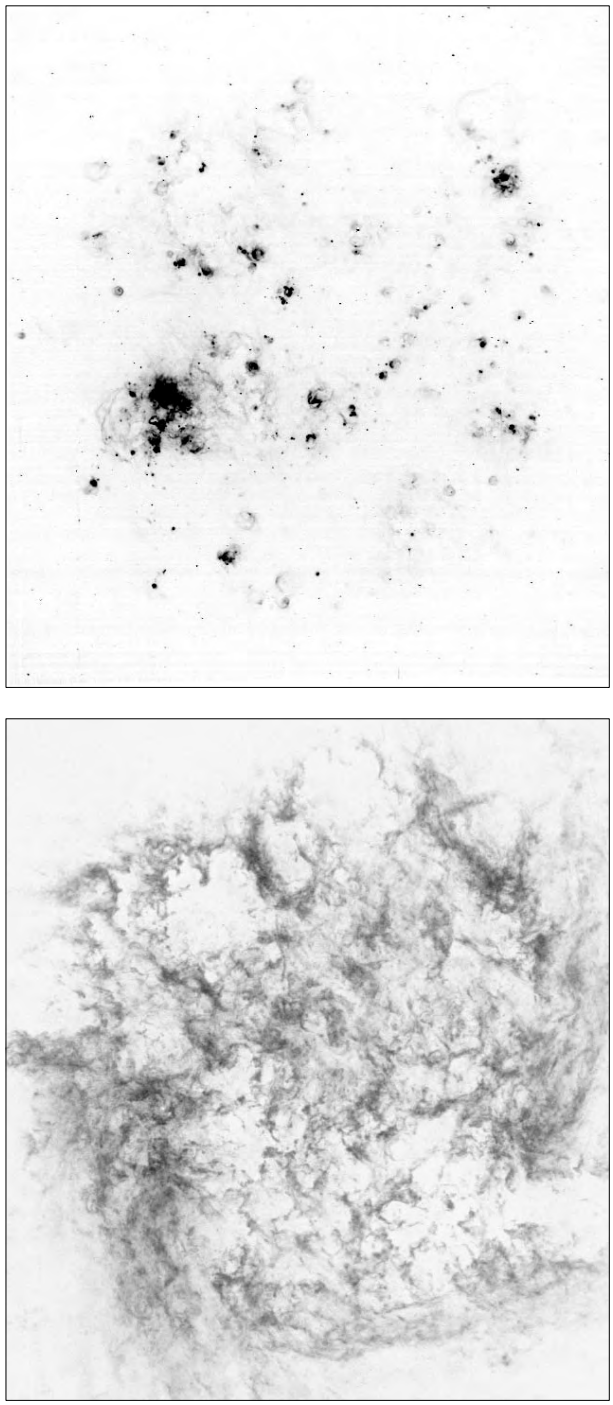


Figure 2. (Top) Hα image of the Large Magellanic Cloud showing the 10^4 K ionized interstellar gas. The main body of the Large Magellanic Cloud is $7^\circ \times 7^\circ$. The brightest nebula is the giant H II region 30 Doradus. Shells of various sizes are clearly seen. (Bottom) H I 21 cm line image of the Large Magellanic Cloud with the same field of view. The Hα image was taken with M Bessell's CCD camera at Siding Spring Observatory. The H I map was obtained with the Australia Telescope Compact Array and the Parkes 64 m Radio Telesope. Image courtesy of S Kim *et al*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

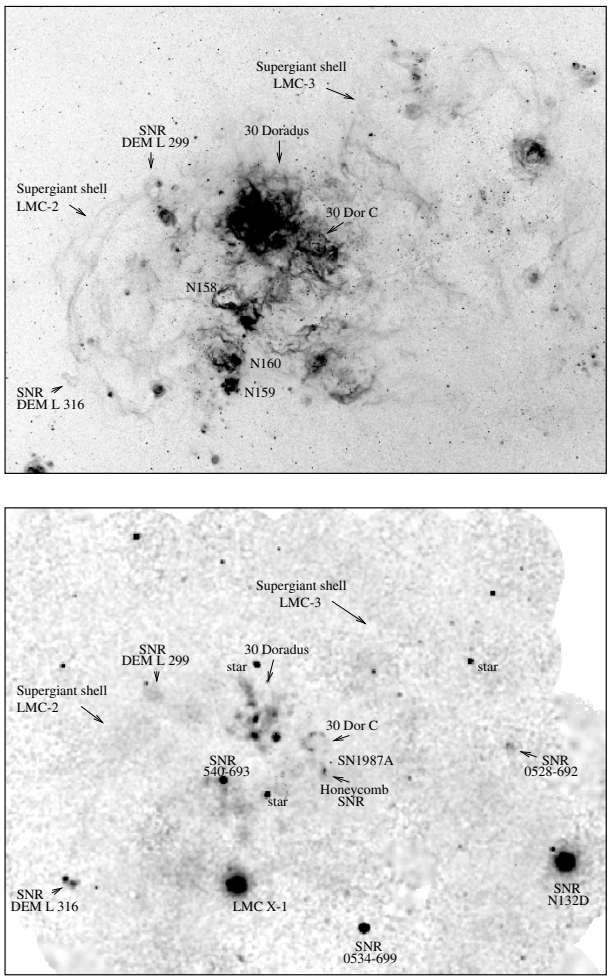


Figure 3. (Top) Hα image of a portion of the Large Magellanic Cloud centered on the 30 Doradus complex. (Bottom) Soft x-ray image of the same region in the 0.1–2.0 keV band. The field of view is $140' \times 110'$. Diffuse x-ray emission is clearly detected in supernova remants, some superbubbles, the giant H II region 30 Dor and two supergiant shells. The Hα image was taken with the Curtis Schmidt Telescope at Cerro Tololo Inter-American Observatory. The x-ray image was a mosaic of observations made with the High Resolution Imager on board the ROSAT x-ray satellite (Röentgen Satellite).

Interstellar shells

The Magellanic Clouds are active in STAR FORMATION and contain large numbers of massive stars. Massive stars ionize the ambient interstellar medium via their ultraviolet radiation and dynamically interact with the ambient medium via fast stellar winds and supernova blasts. The dynamical interactions sweep the ambient interstellar gas into dense shells that are best seen in Hα images (figure 2). The wide variety of interstellar shells in the Magellanic Clouds are formed by the following mechanisms. (1) Single massive stars blow 'bubbles' during their lifetime and form SUPERNOVA REMNANTS eventually. These shells usually have sizes of a few to a few tens of parsecs. (2) Massive stars in an OB ASSOCIATION collectively produce 'superbubbles' in the interstellar medium, and their sizes are a few tens of to a couple of hundred parsecs. (3) When multiple generations of star formation have taken place close to one another, an interstellar 'hole' or a 'supergiant shell' with a diameter as large as 1000 pc may form.

The Large Magellanic Cloud has ∼40 known supernova remnants, including the young remnant of SN 1987A, ∼20 superbubbles around OB associations and 10 supergiant shells. The Small Magellanic Cloud has fewer interstellar shells; for example, it has ∼15 known supernova remnants, and only one supergiant shell has been identified. The Magellanic Clouds are the only galaxies where superbubbles and supergiant shells can be studied comprehensively, as both the physical conditions of the shells and their underlying massive stellar content can be resolved and studied in great detail (see also MAGELLANIC CLOUDS: STAR POPULATIONS).

Multiphase interstellar medium

The interaction between massive stars and the interstellar medium produces shock-heated gas at a few 10^6 K. The mixture of hot ionized gas and cooler ionized and neutral gas forms a multiple-phase structure within the interstellar medium. The Magellanic Clouds provide ideal laboratories to study the multiphase interstellar medium because of their proximity, low extinction and small inclination angle.

The Hα image and the H I 21 cm line map of the Large Magellanic Cloud show vastly different morphologies (figure 2). The ionized warm gas is distributed near recent star-forming regions, while the neutral atomic gas shows fossil shell structures accumulated from previous epochs of star formation.

The distribution of the hot ionized gas relative to the cooler gas is revealed by comparisons of soft x-ray images with Hα images (figure 3). The distribution of the hot ionized gas has a good correspondence to that of the cooler ionized gas, or star formation regions. Hot ionized gas is present in supernova remnants, the giant H II region 30 Doradus, superbubbles, supergiant shells and the unbounded fields. It is conceivable that the hot ionized gas was initially produced and confined in the central cavities of interstellar shells. The hot gas can be injected into the halo, if a shell is larger than the scale height of the gas disk, e.g. a supergiant shell. The hot gas may escape the shell interior if the ambient medium is inhomogeneous and a breakout occurs. These mechanisms can satisfactorily explain the distribution of the hot ionized gas with respect to the cooler gas.

Most of the mass of the interstellar medium is in the neutral components. The Large and Small Magellanic Clouds have H I masses of $7 \times 10^8 M_\odot$ and $6.5 \times 10^8 M_\odot$, and molecular H_2 masses of $1.4 \times 10^8 M_\odot$ and $3 \times 10^7 M_\odot$, respectively. The molecular H_2 masses may be somewhat uncertain because they are derived from CO observations and the conversion factors from CO to H_2 for the Magellanic Clouds are not well known because of their lower metallicities and higher ultraviolet radiation field.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Chu Y-H, Suntzeff N, Hesser J and Bohlender D (eds) 1999 *New Views of the Magellanic Clouds (IAU Symposium No 190)* (Publications of the Astronomical Society of the Pacific)

Westerlund B E 1997 *The Magellanic Clouds* (Cambridge: Cambridge University Press)

You-Hua Chu

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magellanic Clouds: Stellar Populations

The Magellanic Clouds (figure 1) have long been seen as the prototypical young STELLAR POPULATION. The presence of young GLOBULAR CLUSTERS in the Clouds spoke to southern hemisphere observers of the opportunity to study close up processes which have not occurred in the Milky Way for a long time. Young globulars are also seen in other gas-rich, highly disturbed environments, such as merging galaxies (Holtzman *et al* 1996; Brodie *et al* 1998).

Figure 1. The Large and Small Magellanic Clouds imaged by M Bessell with a 3 in camera mounted on the 16 in telescope at Siding Spring Observatory.

The history of star formation

The Hubble Space Telescope has opened up the main sequence in the Magellanic Clouds for detailed study. The pioneering work of Gallagher *et al* (1996) (see figure 2), followed by Holtzman *et al* (1997), yields a good fit with a three-epoch STAR FORMATION history, an initial burst, middle age and a recent burst of star formation. Sufficient main sequence is exposed by the WFPC2 to show that a power law initial mass function (IMF) is a satisfactory fit with $\mathrm{d}n/\mathrm{d}m = m^{-s}$ and $2 < s < 3$, similar to the closest fit in the solar neighborhood. The upper end of the IMF and its high-mass cutoff has been investigated by Massey (1999).

The recent burst began approximately 3 Gyr ago, which may be compared with the orbital timescale of the LMC, 1.3 Gyr (Gardiner *et al* 1994). Multiple LMC fields have now been studied (Geha *et al* 1998) and broadly similar results are seen in all. The star formation history of the LMC as a whole is being studied by Harris *et al* (1999). Olsen (1999) finds that the star formation history of the inner region of the LMC may be more steady. The situation in the SMC is discussed by Hatzidimitriou (1999).

In its cluster formation history the LMC seems to have experienced a quieter middle age according to Da Costa (1999). However, possibly the middle-aged clusters (Sarajedini 1998) were formed and have simply failed to survive. A catalog of Magellanic Cloud associations is provided by Bica *et al* (1999). Comparing the LMC with IC 1613, van den Bergh (1998) draws a distinction between the cluster and field star formation histories.

The Magellanic Clouds have long been the best laboratory for the study of young STAR CLUSTERS. An update on this work is provided by Grebel *et al* (1999). A natural connection can be made between the young clusters and large molecular clouds which reach 2×10^6 solar masses in the NANTEN CO survey (Fukui *et al* 1999).

Kinematics

Details of the stellar kinematics of the Magellanic Clouds pale before the wealth of detail now available on the kinematics of the gas by Kim *et al* (1998, 1999) from the Australia Telescope Compact Array and from the Parkes telescope (see also Putman *et al* 1998). However, a tracer of the intermediate-age population is provided by the CARBON STARS (Kunkel *et al* 1997), which show a cool rotating disk with a circular velocity of approximately 60 km s^{-1}. The cluster kinematics is discussed by Schommer *et al* (1992) and the cluster also forms a simple disk. The stars are too sparse to provide useful age resolution.

Chemistry

The star formation history of the LMC which we have just described together with the assumptions of the simple model of chemical enrichment predict $Z(t)$, the time evolution of the metallicity, which is in quite satisfactory agreement with observations (Dopita *et al* 1997). The present-day stellar abundances are analyzed for Cepheids by Luck *et al* (1998) and by Smith (1999). Da Costa and Hatzidimitriou (1998) trace the chemical evolution of the SMC and find that the simple model is satisfactory. A detailed chemical evolution model is presented by de Freitas Pacheco (1998). A bursting model by Pagel and Tautvaisiene (1998) fits the data very well. An added complication has been introduced by Da Costa (1999), who notes a radial abundance gradient in the LMC's old star clusters.

Surveys

Recent surveys have enriched our knowledge of the Magellanic Clouds. The principal stellar surveys include the EROS and MACHO MICROLENSING surveys, which have yielded a particularly fine harvest of VARIABLE STARS (Alcock *et al* 1999; Beaulieu *et al* 1999), and the photometric survey of Zaritsky (1999).

Bibliography

Alcock C *et al* 1999 *Astron. J.* **117** 920
Beaulieu J P *et al* 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 313
Bica E *et al* 1999 *Astron. J.* **117** 238
Brodie J *et al* 1998 *Astron. J.* **116** 691
Da Costa G S 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 397
Da Costa G and Hatzidimitriou D 1998 *Astron. J.* **115** 1934

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

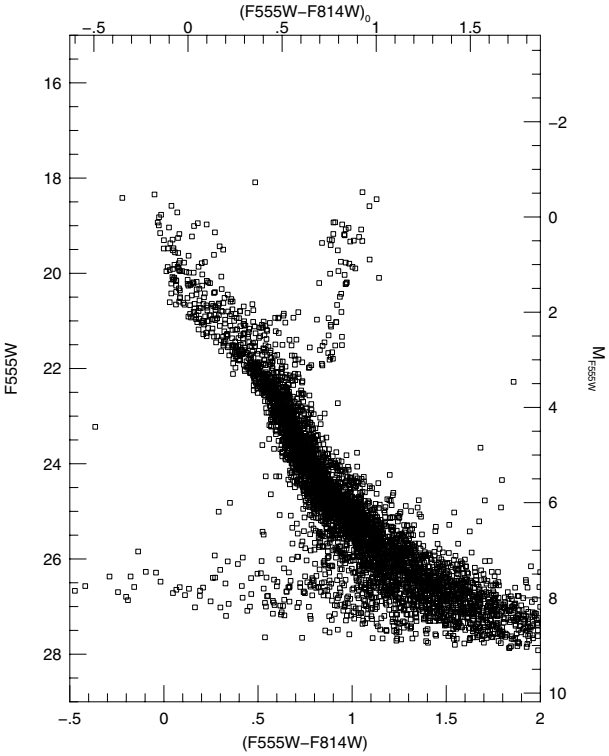


Figure 2. Color–magnitude diagram of a WFPC2 field near NGC 1866 in the LMC from Holtzman *et al* (1997). Stars in the Hertzsprung gap (i.e. between the main sequence and the giant branch) between magnitudes 20.5 and 22 show the history of star formation from 1 to 10 billion years ago.

de Freitas Pacheco J 1998 *Astron. J.* **116** 1701
Dopita M *et al* 1997 *Astrophys. J.* **474** 188
Fukui Y *et al* 1999 *Publ. Astron. Soc. Japan* **51** (NANTEN special issue)
Gallagher J *et al* 1996 *Astrophys. J.* **466** 732
Gardiner L *et al* 1994 *Mon. Not. R. Astron. Soc.* **266** 567
Geha M *et al* 1998 *Astron. J.* **115** 1045
Grebel E, Zaritsky D, Harris J and Thompson I 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 405
Harris J, Zaritsky D, Grebel E and Thompson I 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 347
Hatzidimitriou D 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 299
Holtzman J *et al* 1996 *Astron. J.* **112** 416
——1997 *Astron. J.* **113** 656
Kim S *et al* 1998 *Astrophys. J.* **503** 674
——1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 101
Kunkel W *et al* 1997 *Astrophys. J.* **488** L129
Luck E *et al* 1998 *Astron. J.* **115** 605
Massey P 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 173
Olsen K 1999 *Astron. J.* **117** 2244
Pagel B and Tautvaisiene G 1998 *Mon. Not. R. Astron. Soc.* **299** 535
Putman M *et al* 1998 *Nature* **394** 752
Sarajedini A 1998 *Astron. J.* **116** 738
Schommer R *et al* 1992 *Astron. J.* **103** 447
Smith V 1999 *New Views of the Magellanic Clouds: IAU Symp. 190 (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff, p 259
van den Bergh S 1998 *Astrophys. J.* **507** L39
Zaritsky D 1999 *Astron. J.* **118** 2824

For additional reading see

Chu Y H, Hesser J and Suntzeff N (ed) 1999 *New Views of the Magellanic Clouds: IAU Symp. 109 (ASP Conf. Ser.)*

Jeremy Mould

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magellanic Stream

The Magellanic Stream (MS), which covers an arc of more than 90° across the sky, is a thin band of neutral hydrogen (H I). It is basically continuous, although there are at least six separate regions within the stream where the particle density is visibly enhanced. One end of the MS appears to be connected with the LARGE and SMALL MAGELLANIC CLOUDS, with the other end trailing away from them (see also MAGELLANIC CLOUDS: INTERSTELLAR PHENOMENA). Even though the MS is very close to our Galaxy, at least as far as extragalactic distances are concerned, the process through which it originated is still widely debated.

While it seems most likely that the MS is the result of gravitational interactions (see GALAXIES: INTERACTIONS AND MERGERS), several viable mechanisms exist which could theoretically produce such a feature. One of the major difficulties in determining the origin of the MS is that the exact distance is not known. Another is that while some theories are capable of explaining certain aspects of the MS, there is no theory that can completely reproduce all observed features. Two proposed mechanisms that have each been studied in detail are *ram-pressure stripping* and *tidal stripping*.

There are observed properties of the MS and its surroundings that support the ram-pressure model. First, the edges of the Large and Small Magellanic Clouds on the opposite side of the stream are well defined, as if they are interacting with a gaseous medium through which they are passing. Also, the end of the MS most distant from the Magellanic Clouds has a RADIAL VELOCITY component 200 kilometers per second different from that of the clouds, while the end apparently connected to the clouds does not show this discrepancy. The radial velocity component for the MS, for the most part, shows a smooth and continuous variance between its two ends, as it should if it were the result of ram-pressure stripping. Finally, it has been proposed that the observed clumpy regions within the stream represent regions of gas that lost varying amounts of ANGULAR MOMENTUM through the stripping process.

On the other hand, computer simulations have been performed to examine the tidal stripping model that bear striking resemblance to observations. This simulated stream shows positional agreement with the observed stream, as well as appropriate values for the total mass and radial velocity profile. One difficulty with this model, however, is that theoretically a second stream which leads the Magellanic Clouds should also be produced, but is not observed. While additional regions of neutral hydrogen have been detected in the vicinity of the Magellanic Clouds, it is unclear that any of these should be interpreted as a leading arm. Another difficulty is that stars as well as gas should be affected by tidal forces, while no obvious stellar component has been detected in the MS.

Bibliography

Gardiner L T and Noguchi M 1996 *N*-body simulations of the Small Magellanic Cloud and the Magellanic Stream *Mon. Not. R. Astron. Soc.* **278** 191–208

Moore B and Davis M 1994 The origin of the Magellanic Stream *Mon. Not. R. Astron. Soc.* **270** 209–21

Robert Mohr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnenus [Magnen, Magnien], Johann Chrysostom (*c*. 1590–*c*. 1679)

Physician, born in Luxeuil, Burgundy, became professor of medicine and philosophy at Pavia, Italy, an astrologer, who revived atomism, establishing a viable alternative to Aristotelian philosophy and paving the way for the modern scientific view as expressed by GALILEO.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetars

A magnetar (contraction of 'magnetic star') is a NEUTRON STAR for which the magnetic field energy exceeds all other sources of energy, including rotation. In most cases, this implies a magnetic field strength of approximately 10^{15} G, which would make magnetars the objects with the most intense magnetic fields known. Magnetars were initially proposed to explain the properties of the soft gamma repeaters (SGRs) (Thompson and Duncan, 1995). SGRs are sources of short, repeating, super-Eddington bursts of γ-rays (up to ~100 keV). Three are known to exist in our Galaxy (SGR1900+14, SGR1806–20, and SGR1627–41), and a fourth is apparently in the Large Magellanic Cloud (SGR0525–66). (These designations come from their initially determined positions in right ascension and declination; in some cases, however, subsequent refinements of the positions have taken place. By convention, the designations have not been changed to reflect this.) Their locations and possible associations with supernova remnants suggest that they are neutron stars. The SGRs undergo sporadic periods of bursting activity, punctuated by long quiescent intervals. Occasionally (perhaps every century) they emit very intense giant bursts of γ-radiation up to energies of several MeV, whose intensities at Earth are greater than those of any other cosmic source (Hurley *et al*, 1999). The magnetar model postulates that the energy source for all of these outbursts is an intense magnetic field which stresses the neutron star crust and causes it to crack, exciting Alfvén waves and energizing electrons in the neutron star's magnetosphere. Small cracks explain the frequent, less intense bursts, while global cracking and magnetic field reconnection are responsible for the giant bursts. The SGRs are all quiescent, steady sources of low-energy x-radiation (2–10 keV). In several cases, clear evidence is seen for periodic emission, either of quiescent x-rays, or of the bursting γ-rays. The periods are in the range 5–8 s and, in the magnetar model, are explained by the presence of hot spots on the neutron star surface, with magnetic field dissipation and neutron star crustal stress as the heating source. Experimental evidence for the existence of superstrong magnetic fields comes from observations of the spin-down of the SGRs, deduced from the low-energy x-rays (Kouveliotou *et al*, 1998). The interpretation of these observations is that, similarly to the radio pulsars, the neutron star is losing energy by dipole radiation from its magnetic field.

Bibliography

Hurley K, Cline T, Mazets E *et al* 1999 A giant periodic flare from the soft gamma-ray repeater SGR1900+14 *Nature* **397** 41

Kouveliotou C, Dieters S, Strohmayer T *et al* 1998 An X-ray pulsar with a superstrong magnetic field in the soft gamma-ray repeater SGR1806–20 *Nature* **393** 235

Paczyński B 1992 GB 790305 as a very strongly magnetized neutron star *Acta Astronomica* **42** 145

Thompson C and Duncan R 1995 The soft gamma repeaters as very strongly magnetized neutron stars. I. Radiative mechanism for outbursts *Mon. Not. R. Astron. Soc.* **275** 255

Kevin Hurley

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Magnetic Buoyancy

'Magnetic buoyancy' is a term that is applied to two related, though definitely distinct, phenomena. This article is concerned with magnetic buoyancy as a mechanism for destabilizing atmospheres containing magnetic fields that are in mechanical equilibrium. The term is also applied to the buoyancy of an isolated tube of magnetic field; it is however important to note that this latter phenomenon is not a manifestation of an instability, but rather of a lack of equilibrium. Instability driven by magnetic buoyancy is likely to be an important mechanism in the disruption of any large-scale magnetic field. In particular, it is the most probable cause of the break-up of the Sun's large-scale toroidal field, posited to lie at the base of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE).

The basic instability mechanism

The simplest way of understanding magnetic buoyancy as an instability mechanism is to consider an atmosphere in hydrostatic equilibrium with a unidirectional, horizontal magnetic field that depends only on height, z. The equation of magnetohydrostatic balance is then

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(p+\frac{B^2}{2\mu_0}\right)=-\rho g. \tag{1}$$

Suppose that a small parcel of gas (and magnetic field) is raised slowly, without any bending of the field lines, an infinitesimal distance from a height z to $z+\mathrm{d}z$; denote changes within the gas parcel by 'δ' and changes in the atmosphere external to the parcel by 'd'.

Conservation of mass and magnetic flux of the parcel lead to the relation

$$\frac{\delta B}{B}=\frac{\delta\rho}{\rho}. \tag{2}$$

Conservation of specific entropy of the parcel gives

$$\frac{\delta p}{p}=\gamma\frac{\delta\rho}{\rho} \tag{3}$$

where γ is the ratio of specific heats. Maintaining the total pressure balance between the parcel and its surroundings implies that

$$\delta p+\frac{B\delta B}{\mu_0}=\mathrm{d}p+\frac{B\mathrm{d}B}{\mu_0}. \tag{4}$$

Instability occurs if, on displacement, the parcel is less dense than its surroundings (i.e. $\delta\rho<\mathrm{d}\rho$) and hence continues to rise. Manipulation of equations (2), (3) and (4) leads to the following local criterion for magnetic buoyancy instability:

$$-\frac{B^2}{\mu_0 p}\frac{\mathrm{d}}{\mathrm{d}z}\ln\left(\frac{B}{\rho}\right)>\frac{\mathrm{d}}{\mathrm{d}z}\ln(p\rho^{-\gamma}). \tag{5}$$

The left-hand side represents the potentially destabilizing influence of the magnetic field, fields that decrease with height making the gas top-heavy to some extent; the right-hand side is a measure of the convective stability (or instability) of the atmosphere. From the point of view of magnetic buoyancy instabilities, the interesting case is when the atmosphere is convectively stable (i.e., the right-hand side of inequality (5) is positive), any instability thus being driven by the magnetic field.

Further effects

The instability criterion (5) was derived under the most simplifying assumptions possible—no bending of the field lines, no diffusion and no rotation. In this section we shall consider how the instability mechanism is affected by such considerations, all of which are important in an astrophysical context.

Three-dimensional instability

When fully three-dimensional motions are considered, the change in the instability mechanism is quite subtle. On the one hand, bending of the field lines entails work against magnetic tension; this though becomes arbitrarily small as the wavelength in the direction of the imposed field becomes arbitrarily long. On the other hand—and not so obviously—three-dimensional motions can reduce the work done against magnetic pressure. The basic idea is that instability occurs by the release of gravitational potential energy, this being induced by horizontal variations in the density, and hence, also, in the gas pressure. For the two-dimensional modes considered above, in which the field lines remain straight (so-called interchange modes), gas pressure and magnetic pressure are inextricably linked; it is impossible to do work against gas pressure, necessary for producing density variations, without incurring the penalty of doing work against magnetic pressure. However, this constraint is relaxed by irrotational flows that, though compressible, are essentially solenoidal in the plane perpendicular to the imposed field; such motions can generate density fluctuations without having to do work against magnetic pressure. As a consequence, the local criterion for instability to three-dimensional modes is

$$-\frac{B^2}{\mu_0 p}\frac{\mathrm{d}}{\mathrm{d}z}\ln(B)>\frac{\mathrm{d}}{\mathrm{d}z}\ln(p\rho^{-\gamma}). \tag{6}$$

Thus it is the gradient in B that is now the key destabilizing ingredient, compared to a gradient in B/ρ for interchange modes (cf (5)). For two-dimensional motions in the plane of the field (the 'opposite' case to interchange modes), geometrical constraints lead to a more stringent instability criterion than that given by (6).

The role of diffusion

We have so far considered the instability for ideal (diffusionless) MAGNETOHYDRODYNAMICS. In reality, however, the instability will be influenced by the diffusion of magnetic field and heat, and by viscosity (denoted by η, κ and ν respectively). In the SOLAR INTERIOR the molecular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

values of the diffusivities satisfy the inequalities $\nu \ll \eta \ll \kappa$. In the absence of rotation, viscosity acts merely as a damping mechanism; we shall therefore concentrate here on the role of η and κ and set $\nu = 0$. For simplicity we shall only consider the instability of interchange modes. The diffusive modification to the instability criterion (5) is given by

$$-\frac{B^2}{\mu_0 p}\frac{\mathrm{d}}{\mathrm{d}z}\ln\left(\frac{B}{\rho}\right) > \frac{1}{\gamma}\frac{\eta}{\kappa}\frac{\mathrm{d}}{\mathrm{d}z}\ln(p\rho^{-\gamma}). \qquad (7)$$

It can be seen clearly how instability is facilitated by small values of η (that help maintain the destabilizing magnetic field gradient) and large values of κ (that act to erode the stabilizing entropy gradient).

Diffusive effects also lead to qualitatively new behavior in the form of oscillatory instabilities (i.e. the growth rate $s = s_r + is_i$ with $s_r > 0$ and $s_i \neq 0$). The criterion for such an instability (again with $\nu = 0$) is given by

$$-\frac{B^2}{\mu_0 p}\left(1-\gamma+\frac{\eta}{\kappa}\right)\frac{\mathrm{d}}{\mathrm{d}z}\ln\left(\frac{B}{\rho}\right) > \frac{\gamma}{g}\eta(\eta+\kappa)\frac{k^6}{k_H^2}+\frac{\mathrm{d}}{\mathrm{d}z}\ln(p\rho^{-\gamma}) \qquad (8)$$

where k_H and k are the horizontal and total wavenumbers. The most interesting feature of (8) is the occurrence of instability, for $\eta/\kappa < \gamma - 1$, with $-\mathrm{d}\ln(B/\rho)/\mathrm{d}z$ *negative*, a situation that at first sight one would imagine to be stable. Unsurprisingly, the mechanism behind the instability is very different to that discussed above. However, it is again instructive to consider a small upward displacement of a tube of magnetic flux. From the analysis leading to (5) it follows that if $-\mathrm{d}\ln(B/\rho)/\mathrm{d}z < 0$ then a displaced parcel will be denser than its surroundings and will fall. The crucial element though is that the raised parcel may be hotter than its surroundings; thermal diffusion will then allow the passage of heat from parcel to surroundings, with the possibility that when the tube returns to its initial height it is cooler, and hence denser, than it was originally. If this is the case, it will thus 'overshoot' on the downward side, with repetition of this process leading to an oscillatory instability.

Incorporating rotation

Since magnetic buoyancy is envisaged as being the mechanism responsible for the instability of the solar toroidal field, it is obviously important to consider the role played by rotation. Interchange modes—which may be regarded as the cartesian analogue of axisymmetric modes in a spherical system—are strongly stabilized by the angular momentum constraint imposed by rotation. Thus, in systems where rotation is of any significance, three-dimensional modes are strongly favored. Furthermore, rotation provides a distinct orientation to the system, with the consequence that instability occurs in the form of travelling waves. In general, for a rotating, diffusive system it is not possible to extract the stability criteria analytically; however, various limiting cases do provide explicit results. For example, when rotation is rapid, and when gravity, rotation and the imposed magnetic field are mutually orthogonal (i.e. a toroidal field at the equator) then ideal (diffusionless) three-dimensional travelling waves are unstable when

$$-\frac{B^2}{\mu_0 p}\frac{\mathrm{d}}{\mathrm{d}z}\ln\left(\frac{B}{\rho}\right) > \frac{\mathrm{d}}{\mathrm{d}z}\ln(p\rho^{-\gamma}). \qquad (9)$$

Coincidentally, this is identical to (5), the criterion for *interchange* modes in the *absence* of rotation.

Discontinuous fields

All of the instability criteria above have addressed the issue of the instability of a smoothly varying magnetic field. A related problem—and one that shows magnetic buoyancy instability in its starkest form—is that of the instability of *discontinuous* magnetic fields. If a slab of magnetic field is embedded in an otherwise field-free atmosphere then the magnetic pressure, and hence also the gas pressure, must be discontinuous at the boundaries of the field. If the temperature is assumed to be continuous (though this is not crucial), then the density at the boundaries of the slab must also be discontinuous. At the upper interface the density jump is destabilizing, with (lighter) magnetic gas supporting (heavier) non-magnetic gas; the gas here is genuinely top-heavy and the resulting instability is of a Rayleigh–Taylor type. (By contrast, at the lower boundary the density jump is stabilizing.) There is no instability criterion as such, all locally top-heavy equilibria being unstable, at least in the absence of diffusion. For a conventional (non-magnetic) Rayleigh–Taylor instability, the preferred mode (in the sense of maximizing the growth rate) has, in the absence of diffusion, infinitesimal horizontal scales. The presence of a magnetic field changes this picture; small scales are preferred in the direction perpendicular to the field, longer scales in the direction along the field. Indeed, once the field is sufficiently strong, interchange modes are preferred.

The nonlinear evolution

The most exciting recent work on magnetic buoyancy instability has been in the study of its nonlinear evolution. In the context of the SOLAR MAGNETIC FIELD, one of the most important questions is whether a large-scale ordered field can break up into smaller-scale field concentrations of the form that might be conveyed to the surface to produce ACTIVE REGIONS. Much of the work on the nonlinear development of magnetic buoyancy instabilities has considered an initial state of a slab of uniform, unidirectional, horizontal magnetic field in a field-free atmosphere (leading to magnetic Rayleigh–Taylor type instabilities) and we shall only discuss this case here. It is instructive first to consider the constrained two-dimensional (interchange) evolution before considering the fully three-dimensional case.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

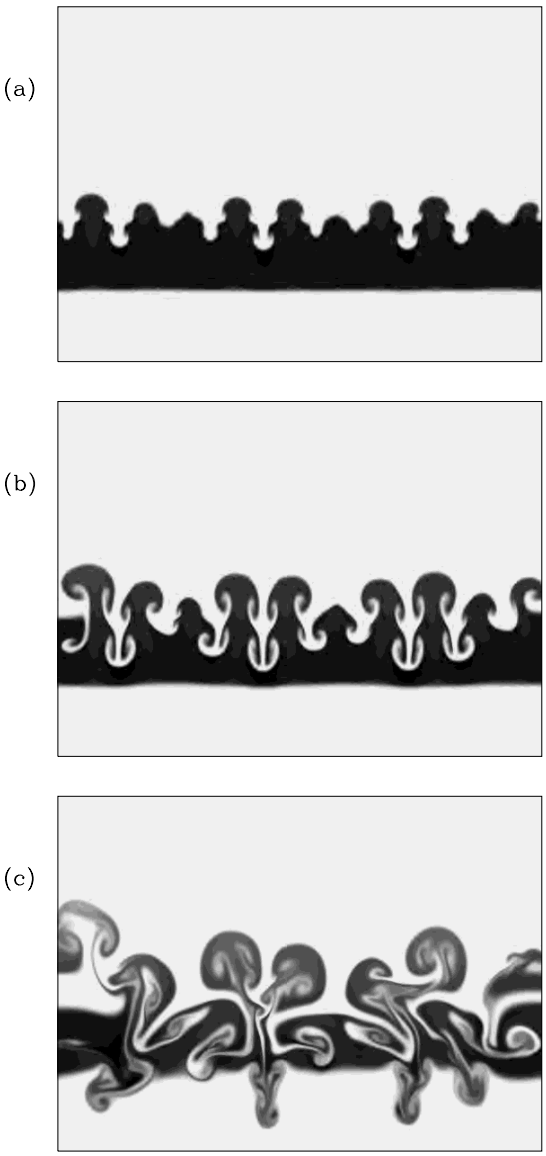


Figure 1. Evolution of the magnetic field for an interchange mode at three diferent times. The field lines are directed into the page. Initially the field occupies the region $0.2 < z < 0.4$.

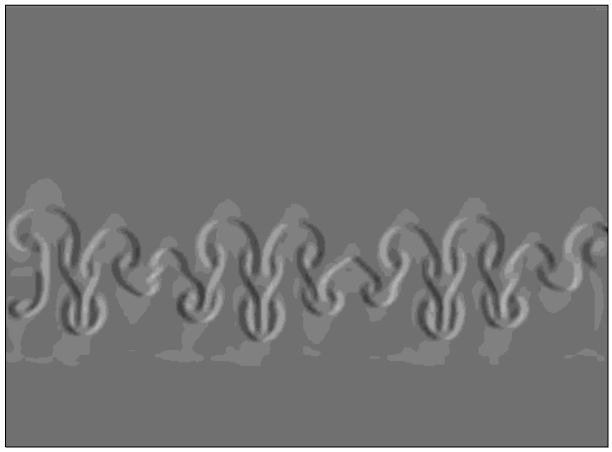


Figure 2. The vorticity corresponding to figure 1(b). Dark (light) shades depict positive (negative) vorticity, with a clockwise (anticlockwise) circulation. It can be seen that the vorticity is concentrated at the mushroom boundaries.

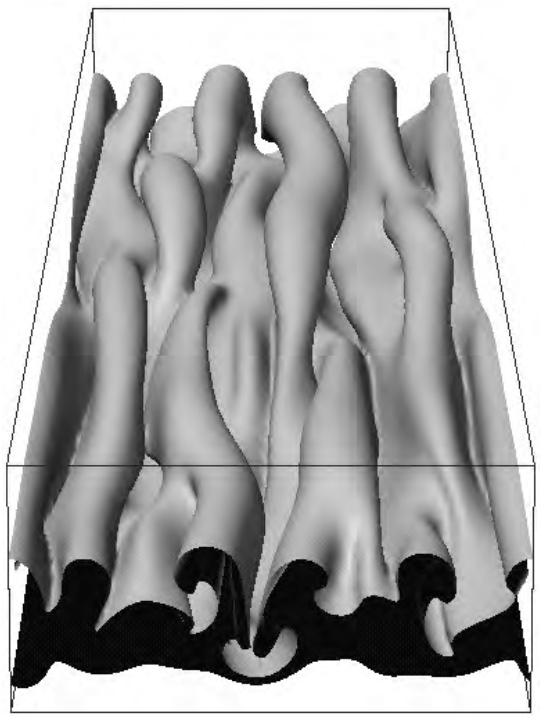


Figure 3. Isosurface of the magnetic field strength for the three-dimensional evolution. Initially the field lines are solely into the page. The cross-sectional cut reveals the mushroom structure perpendicular to the initial field direction. In the third direction, the field has developed undulations as a result of the Crow instability of parallel vortices of opposite sign.

Interchange modes

As mentioned above, instability results from the destabilizing density jump at the upper magnetic interface. The rise of the buoyant magnetic gas and the fall of the overlying denser, non-magnetic gas, results in a shear flow centered on the (moving) magnetic interface. This in turn leads to a secondary Kelvin–Helmholtz instability, which wraps the gas into concentrated patches of vorticity and gives the magnetic field the mushroom shape characteristic of certain Rayleigh–Taylor instabilities (see figures 1 and 2). The subsequent development of the instability is governed, to a large extent, by the interaction of these vortices. The key interactions turn out to be between vortices on neighboring mushrooms—rather than those between vortices on the same mushroom—with the result that the vortices, and their associated magnetic field, are pulled downwards. What is surprising is that the vortex inter-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

David W Hughes

actions are sufficiently strong as to be able to overcome the intrinsic buoyancy of the magnetic field. As a consequence, there is no wholesale rise of the magnetic field; indeed, pockets of fairly strong field are carried down even below the original lower boundary of the field (figure 1(c)).

Three-dimensional modes

Even when the motions are completely unconstrained, the instability initially evolves in an essentially two-dimensional manner, as described above. Strong vortices are again formed, as a result of a secondary Kelvin–Helmholtz instability, with the vorticity predominantly parallel (or antiparallel) to the initial field. It is in the subsequent vorticity-driven phase that the two- and three-dimensional cases differ strongly. The dominant interactions are again between vortices on neighboring mushrooms. However, whereas for interchange modes this leads to a predominantly downward advection of the vortices, in three dimensions the interactions are such as to drive a further instability, resulting in undulations *along* the vortices. This instability mechanism (known as the crow instability) is precisely that which can sometimes be observed in the interacting trailing vortices of aircraft. Since the magnetic field is closely tied to the fluid, the buckling of the vortices causes a corresponding arched structure in the magnetic field (figure 3). Such a mechanism may be of significance in the development of longitudinal variation in the emerging solar magnetic field.

Future research

This article has concentrated on the fundamental properties of magnetic buoyancy instabilities, an understanding of which has come from the study of fairly straightforward equilibrium configurations. Future research, which will undoubtedly concentrate predominantly on the nonlinear behavior, will seek to explain the role of magnetic buoyancy instability in a somewhat broader context. For example, in the solar interior, where magnetic buoyancy is believed to be responsible for the disruption and escape of the predominantly toroidal field, the influence on the instability of a poloidal field component, of a strong shear, and of the overlying convection, must all be considered.

Bibliography

Acheson D J 1979 Instability by magnetic buoyancy *Solar Physics* **62** 23–50

Hughes D W 1991 Magnetic buoyancy *Advances in Solar System Magnetohydrodynamics* ed E R Priest and A W Hood (Cambridge: CUP) pp 77–104

Hughes D W and Proctor M R E 1988 Magnetic fields in the solar convection zone: magnetoconvection and magnetic buoyancy *Ann. Rev. Fluid Mech.* **20** 187–223

Matthews P C, Hughes D W and Proctor M R E 1995 Magnetic buoyancy, vorticity, and three-dimensional flux-tube formation *Astrophys. J.* **448** 938–41

Parker E N 1979 *Cosmical Magnetic Fields* (Oxford: Clarendon)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetic Fields in Stars

Magnetism—the force that deflects the needle of a compass—and magnetic fields have been found in some hundreds of stars during the past 50 yr. Magnetic fields have been detected in T Tauri stars and other pre-main-sequence stars, several types of main sequence stars, white dwarfs and neutron stars. We now know a number of methods by which such magnetic fields may be detected, we are in the process of learning more about how they are distributed over stellar surfaces, and we understand some of the ways in which these fields reflect—and sometimes influence—the evolution of the stars which possess them.

The first stellar magnetic fields were detected in sunspots (see SUNSPOT MAGNETIC FIELDS) in the Sun by George Ellery Hale in 1908. Almost forty years later, in 1947, the first magnetic field in a star other than the Sun was found by Horace W Babcock, who discovered a magnetic field in the star 78 Virginis, a 'chemically peculiar' main sequence star (see STELLAR EVOLUTION) about twice as massive as the Sun. Magnetic fields are now known in perhaps 200 other A and B stars of the middle main sequence, all of which are, like 78 Vir, chemically peculiar (which means that they have very unusual surface chemical compositions). In such stars the fields are generally found to be roughly dipolar in form; in other words, they have an overall structure reminiscent of that of a simple bar magnet or of the Earth's magnetic field, with a north and a south pole, between which the the magnetic force (as represented by magnetic lines of force) points along loops connecting one pole to the other. The typical field strength of such stars—over the whole stellar surface—is of the order of 1000 G (0.1 T), some 3000 times greater than the strength of the Earth's surface field, and about as strong as the magnetic field of a good permanent horseshoe magnet.

The discovery of PULSARS in 1967 by Jocelyn Bell Burnell and Anthony Hewish was soon recognized to be both the discovery of NEUTRON STARS and of magnetic fields in such stars. The pulsed radio radiation emitted by these spinning, magnetized neutron stars is still almost the only means for detecting single neutron stars. The roughly dipolar fields of neutron stars are initially of the order of 10^{12}–10^{13} G (10^8–10^9 T) and then seem to decay in strength by about a factor of 100. Most or all neutron stars thus appear to be formed initially with fields about 10^{10} times stronger than are found in the magnetic middle main sequence stars.

Three years later the first magnetic field was detected in a WHITE DWARF by James Kemp, John Swedlund, John Landstreet and Roger Angel. Fields are now known in about 50 other white dwarfs. These fields range from about 10^5 to 10^9 G (10–10^5 T) in strength, roughly a factor of 10^4 stronger than those of middle main sequence stars. The white dwarf fields also appear to be approximately dipolar in structure. Unlike neutron stars, only a few per cent of all white dwarfs have detectable magnetic fields.

The first magnetic fields in stars of the lower main sequence were detected in 1980 by Richard Robinson, Pete Worden and Jack Harvey. Fields are now known in more than 50 cool stars, mostly rather young, active stars, or in stars in which a companion enforces rapid rotation (see the SOLAR–STELLAR CONNECTION). Recently, fields have been found in a few PRE-MAIN-SEQUENCE STARS and T TAURI STARS. In cool stars, the fields detected have very different distributions over the stellar surface from those of the stellar types already described. Instead of simple, roughly dipolar structure, these fields seem to occur in forms more like giant sunspots, or in patches on the stellar surface that may resemble solar active regions. The field strengths are typically of the order of 10^3 G (0.1 T), while the fraction of the stellar surface covered is typically only of order 20–50%.

Unlike the fields of middle main sequence stars, white dwarfs and neutron stars, all of which are observed to change in structure only very slowly with time, the distribution of magnetic flux on the surface of a lower main sequence star usually changes substantially in a period of weeks or months. It is generally supposed that, in the cool stars, the fields observed are generated by a dynamo process operating in the convective outer envelope of the star, while the more stable fields of middle main sequence stars, white dwarfs and neutron stars are 'fossil fields'—large-scale fields produced during an earlier stage of evolution, and subsequently frozen into the highly electrically conductive matter of the star. In this article we will focus on the fields of middle main sequence stars and of white dwarfs, leaving those of solar-type stars and of neutron stars to other articles.

Methods of detecting stellar magnetic fields

Basic physics

Most of the magnetic fields detected in main sequence and white dwarf stars are found by detecting the ZEEMAN EFFECT in the stellar spectrum. This effect splits each energy level of an atom in a magnetic field into several magnetic substates, leading to a number of effects that can in favorable cases be detected.

When placed in an external magnetic field, a state i of an atom with energy E_i and total angular momentum quantum number J splits into $2J + 1$ magnetic substates equally spaced in energy, with a spacing which varies from one atomic level to another. Transitions between level i and another level f of energy E_f are characterized by the frequency $\nu_{if} = (E_f - E_i)/h$ when no magnetic field is present. When a field B is applied, the splitting of the lower and upper energy levels by the magnetic field leads to the splitting of the spectral line associated with this transition into three closely spaced group of lines. These groups arise because most atomic transitions allow the magnetic quantum number M to change by -1, 0, or 1. The $\Delta M = M_f - M_i = 0$ group, called π components, are distributed symmetrically about ν_{if}. The two groups of lines with $\Delta M = \pm 1$, called σ components, are shifted systematically to frequencies above and below ν_{if}, with the $\Delta M = +1$ group on one side and the $\Delta M = -1$ group

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on the other. The typical separation between the π and one of the σ groups is

$$\Delta\lambda_B = \bar{g}eB\lambda^2/4\pi mc^2$$

where $\bar{g}$ is a number of order 1 which varies from one transition to another. In familiar units, the splitting is

$$\Delta\lambda_B(\text{nm}) = 4.67 \times 10^{-3}\bar{g}B(\text{kG})\lambda(\mu\text{m})^2.$$

Thus a field of 1 kG (0.1 T) leads to a π–σ separation of the order of 0.001 nm at 500 nm = 0.5 μm.

The π and σ groups of lines are *polarized*. If the magnetic field is transverse to the line of sight, the π components (in emission) are linearly polarized parallel to the applied field and the σ components are linearly polarized in the orthogonal direction. If the field is parallel to the line of sight, the π components are suppressed and the two groups of σ components have opposite circular polarizations.

When the magnetic field strength is strong enough (of order 10^5 G) that the perturbation of the atom by the field is larger than the spin–orbit perturbation, the relatively complex Zeeman effect is superseded by the Paschen–Back effect, which leads essentially to splitting of all lines into simple triplets. A field of order 10^6 G results in a significant quadratic Zeeman effect, which systematically shifts lines of large upper principal quantum number n_f relative to lines of smaller n_f. As a field of order 10^7 G is reached and the magnetic interaction energy becomes comparable with the Coulomb energy of the atomic electrical field, the atomic spectrum of any atom, even H, becomes extremely complicated.

At fields above about 10^6 G, another magnetic effect occurs that is very useful for detection of fields in white dwarfs: polarization of *continuum* radiation. The broad-band light from a star with a field strength of this order is circularly polarized, essentially because the field forces electrons to spiral about field lines in a preferred direction. For still larger fields, broad-band linear polarization can also occur.

Field measurement methods

The splitting, shifting and polarization of spectral lines by the Zeeman, Paschen–Back and quadratic Zeeman effects, and the occurrence of continuum polarization for sufficiently large fields, have provided astrophysicists with a number of methods of detecting and measuring magnetic fields in stars.

The most straightforward of these methods is useful for stars that have a very small projected equatorial rotational velocity $v_{\text{eq}} \sin i$ and hence sharp spectral lines. In this case, one can directly observe the splitting of spectral lines into components if the field is of the order of a few kilogauss or more in main sequence stars, or about 10^6 G in white dwarfs. The separation of the observed line components provides a direct measurement of the modulus of the magnetic field averaged over the surface of the star, a quantity called the mean field modulus, B_s, as shown in figure 1.

The polarization introduced into spectral lines by magnetic splitting provides a second powerful method of field detection. This method depends on the fact that, in a magnetic field with a significant component along the line of sight, the σ components on one side of the line center absorb circularly polarized light of one sense of polarization, while the σ components on the other side of line center absorb the opposite circular polarization. If we observe the stellar spectrum through polarizers that pass each of the two senses of circular polarization, the absorption lines in one of the two circularly polarized spectra are not at precisely the same wavelengths as the same lines in the other polarized spectrum, because of the small wavelength difference between the two sets of σ components. The presence of a field can be detected by this shift in the position of spectral lines between spectra observed in right- and left-circularly polarized light, or equivalently in the spectrum formed by the difference of the two spectra divided by their sum, as shown in the left-hand panel in figure 2.

The fact that, in the presence of a magnetic field transverse to the line of sight, the absorption by the π components is orthogonally polarized with respect to the polarization of the σ components leads to a similar effect when a spectrum is observed through linear polarizers oriented parallel to and perpendicular to the field direction, as seen in the right-hand panel of figure 2. In general, the observable effect in linear polarization is considerably smaller than in circular polarization.

Detection of the polarization effects from a stellar magnetic field is possible only if there are substantial regions on the stellar surface over which the field does not change direction too much. Clearly, a field structure in which many small tubes of magnetic flux directed out of the star are closely mixed in with small tubes of inward-directed flux will lead to no net polarization, as the effects of adjacent oppositely directed flux tubes will cancel. On the other hand, the detection of circular polarization, which is not readily produced in line profiles by other mechanisms, is a very robust indicator of the presence of a field. Furthermore, very small levels of polarization (0.1% or even 0.01%) can be measured reliably. The result is that circular polarization methods, which measure what is called the 'mean longitudinal field' B_ℓ, provide much more sensitivity to weak but geometrically simple magnetic fields than methods that depend on studying line profiles. In the best cases, B_ℓ values as small as tens of gauss can presently be detected.

Magnetic fields in middle main sequence ('peculiar A') stars

The oblique rotator

Magnetic fields are detected in middle main sequence stars (stars of between about 2 and 10 times the mass of the Sun, which are burning hydrogen in their cores) both by the detection of circular polarization in spectral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

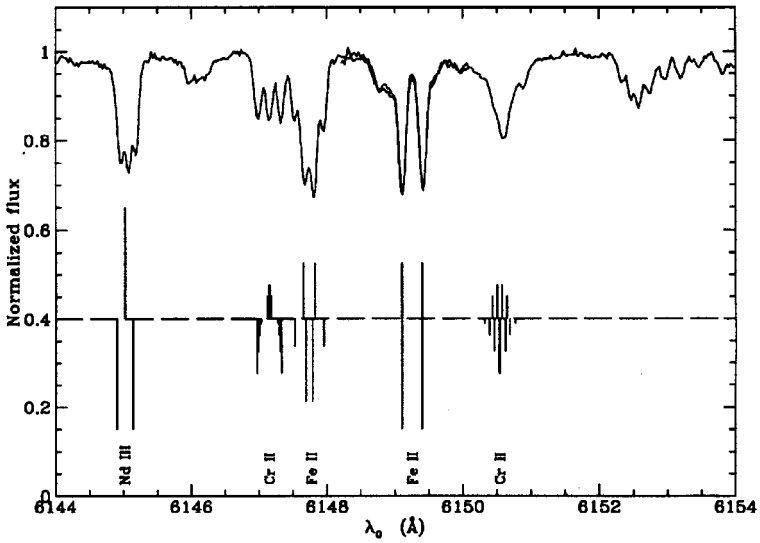


Figure 1. A portion of the spectrum of the very slowly rotating magnetic Ap star HD 94660 (the spectrum at the top of the figure), showing splitting of spectral lines by a magnetic field of $B_s \approx 6400$ G. Below the stellar spectrum is a schematic diagram of the Zeeman splitting pattern of several strong stellar spectral lines. Each of the vertical lines above and below the line at intensity 0.4 (respectively π and σ components) represents one Zeeman component; the position of each component is calculated for a field of 6400 G, and the height of each line is proportional to the strength of that component. The ion responsible for each strong line is identified at the bottom of the figure. (From G Mathys 1990 *Astron. Astrophys.* **232** 151, reproduced by permission of *Astronomy and Astrophysics*.)

lines and by observing Zeeman splitting of spectral lines in stars with small $v_{\rm eq} \sin i$ values. All well-confirmed detections of fields in middle main sequence stars are in members of a class of stars known as 'chemically peculiar A stars', often called Ap (or Bp) stars (see also PULSATING AND CHEMICALLY PECULIAR UPPER MAIN SEQUENCE STARS). These stars have long been known to have (sometimes very) anomalous atmospheric chemical composition compared with the Sun. Their chemistry is typically characterized by unusually large amounts of chemical elements such as Sr, Cr, Eu and other rare earths, Si and (only in the most massive Bp stars) He. Most of these stars also have unusually small amounts of a few elements as well, often He and O. The elements which are anomalous depend systematically on the mass of the star. It appears that all chemically peculiar stars having the same general chemical anomalies as the known magnetic stars probably have fields.

A first question about the magnetic middle main sequence stars (I will call them 'magnetic Ap stars') is that of determining the geometry of the magnetic field over the stellar surface. This cannot be determined by directly observing the disks of such stars, of course; they are much too small in angular size for direct imaging. We must use other kinds of information to deduce the magnetic geometry.

The magnetic Ap stars are in most cases observed to vary periodically in apparent brightness, in the strengths and profiles of some or most spectral lines, and in the measured components of the magnetic field. All these quantities are generally observed to vary with the same period. The period of variations is typically in the range 1–10 days, although periods as short as 0.5 days and as long as some decades or known. An example is shown in figure 3.

An important clue to the origin of the variations is furnished by the observed fact that the value of the projected rotational velocity $v_{\rm eq} \sin i$ is closely correlated with the period of variation. Large values of $v_{\rm eq} \sin i$ are *only* found for stars with short periods and, the longer the period, the smaller the values of $v_{\rm eq} \sin i$ are. The facts that the observed periods are found with an enormous range of values and that the periods are closely related to the projected rotational velocities clearly indicate that the observed variations are due to the *rotation* of the magnetic Ap star. Variation in the average magnetic field indicates that the star must have a magnetic field that varies from one place to another at the surface, either in strength or in inclination or both. Thus, when we see mainly magnetic field lines directed towards the observer, we measure a large value of B_ℓ, but, when the field lines are mainly perpendicular to the line of sight, B_ℓ is small. The variations in spectral line intensity and shape indicate that the relative abundances of various chemical elements vary from one place on the star to another. When we are looking at a part of the star in which some element (such as He) is relatively abundant, the spectral lines are strong and deep; when we look at a different part of the star which has relatively less He, the spectral lines are weaker. The variations in surface chemistry in turn influence the amount of light emitted at various wavelengths and lead to the variations observed in brightness as the star rotates.

We observe that the value of B_ℓ of a magnetic Ap star generally varies in a fairly sinusoidal fashion. When we

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

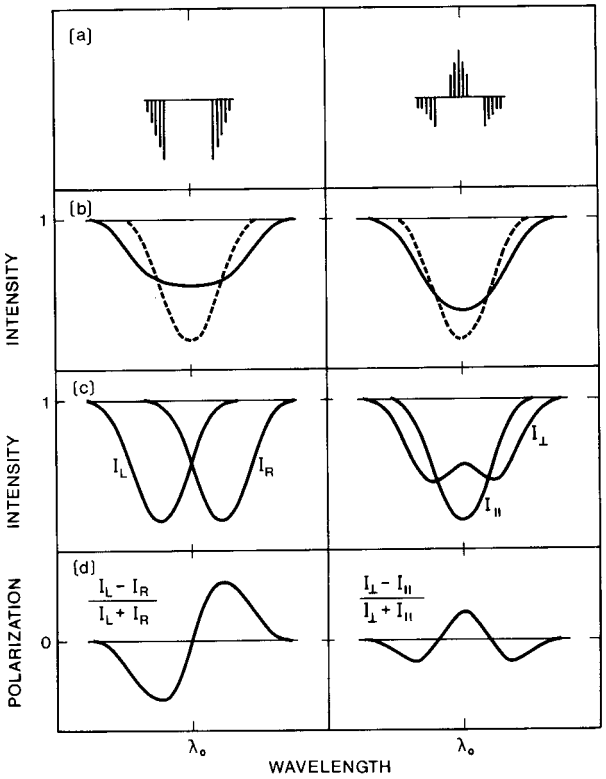


Figure 2. Effects of a magnetic field on the line profile and polarization of stellar spectral lines. The panels show schematically the effect of a magnetic field with the Zeeman pattern displayed in (*a*) for a field parallel to the line of sight (left) and perpendicular to it (right). (*b*) The change in the line from the non-magnetic case (dotted) to the magnetic case (solid) for Zeeman splitting comparable with the Doppler width of the spectral line. (*c*) The absorption line in left- and right-circular polarization (left) and in linear polarization parallel to and perpendicular to the stellar field (right). (*d*) The net circular (left) and linear (right) polarization in the line. (From D J Landstreet 1980 *Astron. J.* **85** 611, reproduced by permission of the American Astronomical Society.)

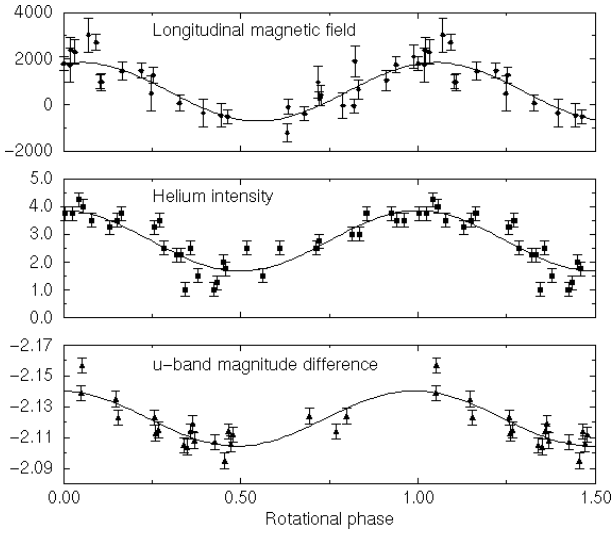


Figure 3. Variations of the massive magnetic Bp star HD 184927 (whose most striking abundance peculiarity is a considerable excess of atmospheric He) with a period of 9.53 days. The horizontal axis gives time in phase units (fractions of one cycle from the periodically repeating time when the helium spectral line intensities are strongest). The three panels show (from top to bottom) the mean longitudinal field in kilogauss, the strength of one helium line in arbitrary units and the brightness of the star (in magnitudes) seen through a narrow (Strömgren) ultraviolet filter. (From G A Wade *et al* 1997 *Astron. Astrophys.* **320** 172, reproduced by permission of *Astronomy and Astrophysics*.)

try to reproduce this behavior by calculating the variations expected from various simple models of the field geometry, we find that this observation is consistent with the idea that the structure of the magnetic field over the surface of the star is in the general form of a dipole, typically inclined (oblique) to the rotation axis of the star by some fairly large angle. As the star rotates, we thus usually see one pole of the dipolar distribution, and then the other. This model is known as the 'oblique rotator model'.

Astronomers are also actively working to use observed spectrum variations to deduce the distributions of different chemical elements over the surface of the star. Often the models that fit the observations have a distribution of the elements that is roughly axisymmetric around the axis of the magnetic dipole. It is not uncommon to find very large differences in the fractional abundances of some elements over the stellar surface; in some cases there may be more than 100 times more atoms of some elements per unit volume of gas in one part of the star's atmosphere than in another part.

Origin of the observed magnetic fields and of the chemical anomalies

Astronomers generally accept two possible origins for observed stellar fields. One is that a field may be generated by the interplay between convection (boiling motions of the gas) in the outer layers of a star and the overall rotation of the star. These motions may act as a dynamo in the highly conducting outer layers of a star, producing a complex and time-varying field. This mechanism seems to be the cause of the magnetic field observed in the Sun. In spite of much theoretical effort, such dynamo fields are still not well understood.

The second possible origin is that the field is the result of the collapse of a huge gas cloud to form a tiny star, trapping in the partly ionized, electrically conducting gas some small fraction of the weak galactic magnetic field. As the field lines of the galactic field are squeezed together, the strength of the entrained field is amplified by a very large factor. Thus, the observed fields of the magnetic Ap stars may be 'fossil' magnetic fields. This is not as unreasonable an idea as may at first appear. Owing to the very large bulk and high electrical conductivity of a star, a field formed in this way would take a very long time, of the order of 10^{10} yr, to decay. This is longer than the main sequence lifetime

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a middle main sequence star, and so a fossil field could easily persist throughout the life of such a star.

The absence of any long-term changes in the observed magnetic field strengths of magnetic Ap stars (such as the solar activity cycle), other than the periodic variations caused by stellar rotation, and the simple overall geometry deduced for these fields, suggests that they are probably fossil fields rather than dynamo-generated ones. This view is supported by the lack of any obvious means of generating a large dynamo field in a star that may be rotating 100 times slower than the Sun and that has almost no convection in its outer layers.

The presence of remarkable chemical peculiarities in these magnetic stars also requires an explanation. In fact, this is not quite as anomalous as it seems. The lower-mass stars of the main sequence ($M \leq 2M_{\odot}$) show quite homogeneous abundance patterns, with overall content of 'heavy elements' (everything heavier than He) that depends only on the age of the star. (Older low-mass stars, formed when the universe was less rich in the elements synthesized in many generations of supernova explosions, have less of all heavy elements than younger stars.) Very massive main sequence stars ($M \geq 10M_{\odot}$) lose mass rapidly, and this ensures that their surfaces reveal their bulk composition which, because of their very short main sequence lifetimes, is essentially that of the contemporary interstellar material from which they form. The intermediate-mass stars, among which the magnetic Ap stars are found, almost all exhibit some degree of chemical individuality. In most A and B stars the variations are no more than some tens of per cent of excess or deficiency compared with other similar stars, but several other families of stars (mostly slowly rotating and apparently non-magnetic stars) are known in which certain deficiencies or enhancements can reach much larger values. 'Metallic-line A' (Am) stars often have 10 times less Ca and Sc than other main sequence stars of similar mass, and 10 times more of some rare earths. 'Mercury–manganese' (HgMn) stars have enhancements of some elements by factors of from order 10^2 (V, Mn, Ga) up to 10^5 or more (Eu, Pt, Hg). The abundance anomalies of the magnetic Ap stars are simply some of the most spectacular types of anomaly in a mass range filled with variety.

These chemical abundance anomalies are generally believed to be confined to the atmospheres and outer envelopes of intermediate-mass stars, rather than being representative of the bulk chemical composition of these stars, for several reasons. First, the wide variety of observed compositions, in stars all of which formed relatively recently in galactic history, does not correspond to any similar variety of compositions in the interstellar clouds which form stars, or in other young stars of low or high mass. Furthermore, the extremes of anomaly are so great (factors of 10^5 or more) that it is not possible to imagine any way in which star formation could have led to gas clouds with such peculiar composition. Instead, we believe that the observed chemical anomalies are essentially surface phenomena, due to powerful processes that *separate* elements, raising some into the atmosphere while others sink out of sight.

The main sorting process leading to chemical anomalies is microscopic diffusion of atoms of low-abundance elements, relative to the dominant hydrogen of the stellar gas. Under the influence of gravity, elements with higher atomic mass than hydrogen tend to sink into the interior of the star. In a sufficiently stable atmosphere, this process would eventually lead to an exterior layer made up only of hydrogen, as is actually observed in many white dwarfs. However, there are competing processes. One of the most important is the outward force felt by atoms and ions which can absorb photons of many wavelengths from the outward flow of radiation through the star. This absorption imparts an outward acceleration to such ions and lifts them up to higher levels in the stellar envelope. Thus, the overall effect of diffusion is to allow some elements to sink in the atmosphere under the dominant influence of gravity, while others are lifted towards the surface by radiation.

These sorting processes compete with various mixing processes such as convection. Thus, because the outer layers of low-mass main sequence stars are strongly convective, all sorting processes are strongly inhibited, and these Sun-like stars exhibit very similar compositions.

In contrast, the main sequence stars of intermediate mass are precisely the stars with sufficiently stable atmospheres to allow diffusion to sort the chemical elements, at least to some extent. Rapid rotation is capable of generating slow mixing currents, and so the more rapidly rotating A and B stars have only modestly sorted surface chemistries. Most of the more peculiar middle main sequence stars are slowly rotating. The magnetic Ap stars have the additional feature that the presence of the magnetic field rather strongly inhibits mixing motions in the outer layers. The chemical peculiarities of the magnetic Ap stars are simply a particularly strongly developed aspect of a characteristic found in all stars in this mass range.

Magnetic fields in white dwarfs

Observations and modelling

Magnetic fields are detected in white dwarfs by the same methods used for magnetic Ap stars, namely by direct observation of magnetic splitting of spectral lines and by observation of circular polarization in line wings. Fields are also detected by means of the continuum polarization produced by fields of more than about 10^6 G. The deduced fields range in strength from about 10^5 up to 10^9 G. At the low end of this range, the spectrum of a white dwarf is hardly perturbed at all by the field. For fields in the range from about 10^6 to 3×10^7 G, splitting of familiar spectral lines is easily seen. For still larger fields, the wavelengths and shapes of spectral lines are so strongly altered by the field that the spectrum is not recognizably related to that of any non-magnetic white dwarf. The spectrum of one magnetic white dwarf is shown in figure 4.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

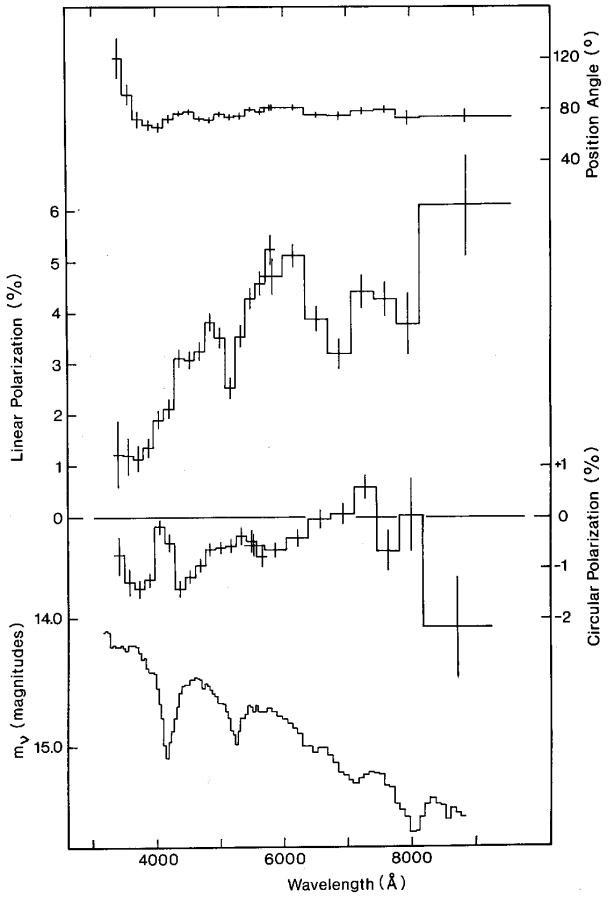


Figure 4. The flux and polarization spectrum of the magnetic white dwarf GD 229, which has a field of order 10^9 G. The lowest curve shows the wavelength variation of the flux, the second lowest of circular polarization, the third lowest the percentage linear polarization and the top the position angle of linear polarization.

The polarization and/or line splitting is observed to be variable in about one-quarter of the known magnetic white dwarfs. Observed variations are periodic, with periods in the range from about 1 h to 20 days. These periods are so long compared with any reasonable oscillation period of a white dwarf that they must be rotation periods, and so we are again quickly led to the oblique rotator model for the variations. The observed variations in the magnetic field strength and in spectral line shapes are again interpreted as simply being due to the fact that we see a magnetic field that is inclined to the stellar rotation axis from different directions as the star rotates. The fact that most magnetic white dwarfs do *not* vary may imply that, in most magnetic white dwarfs, the magnetic field is axisymmetric about the rotation axis or possibly that most magnetic white dwarfs rotate with periods of decades or more.

Modelling of observed spectra and their variations is possible if the fields are not too large, say not much more than about 3×10^7 G. Modelling of such fields leads to deduced magnetic field geometries that are roughly dipolar, like the fields of magnetic Ap stars. For stronger fields, the main difficulties come from the complex and uncertain behavior of even simple atoms such as H in the presence of fields so strong that they influence the motion of the electron(s) as strongly as the central Coulomb attraction of the atomic nucleus. The observations of spectra of white dwarfs with fields of 10^8 G or more are in reasonable accord with the results of atomic physics calculations, but no detailed modelling has yet been possible.

Fields as small as about 10 kG could be detected in most white dwarfs. Surveys of many white dwarfs to about this level of precision have shown that the great majority of white dwarfs do not possess detectable magnetic fields. About 4% of the total population of white dwarfs have fields, with about equal probabilities per decade of field strength over the range 10^5–10^9 G.

Origin of white dwarf fields

Trying to understand the origin of the fields observed in white dwarfs presents us with substantial challenges. There are no obvious mechanisms for producing large-scale, ordered, static fields in either magnetic Ap stars or white dwarfs after they are formed. We observe that a small fraction of middle main sequence stars, and of white dwarfs, have magnetic fields large enough to detect, in the range 10^2–10^5 G on the main sequence and 10^5–10^9 G in white dwarfs. The observed magnetic Ap fields may be due to magnetic flux retention during star formation, and the fields of white dwarfs could be due to the further retention of that same flux as magnetic Ap stars collapse to become white dwarfs. This hypothesis is consistent to some extent with the relative values of observed field strength, since, if the magnetic flux Φ threading a star's equator is retained during a collapse, the magnetic field strength will increase as

$$B \propto \Phi / R^2$$

where R is the stellar radius. Thus the decrease in radius by a factor of 10^2 as a star becomes a white dwarf could lead to a field strength increase by a factor of 10^4, about the difference observed between the ranges of field strength on the main sequence and among white dwarfs. However, this does not explain how that magnetic flux is retained in the evolution stages between the main sequence and white dwarf stages; the intervening giant state is expected to be largely convective, which might be expected to expel much of the magnetic flux in a star. Furthermore, this idea does not explain why the largest (10^8–10^9 G) fields are as common as fields 10^3 times smaller; on the main sequence the largest fields are a modest tail on a distribution that is very strongly peaked around fields of less than 10^3 G.

Magnetic fields in neutron stars present us with further challenges. It appears that almost all neutron stars have fields of the order of 10^{10}–10^{13} G. Again, these are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

about the fields that would be expected by magnetic flux retention from the main sequence. However, if this is the origin of magnetic fields in neutron stars, why do almost *all* neutron stars have large fields, while only a small fraction of white dwarfs have large fields? This question is made still more puzzling by the fact that virtually no magnetic main sequence stars are known in the mass range that is expected to evolve eventually to neutron stars.

Bibliography

Chanmugam G 1992 Magnetic fields of degenerate stars *Ann. Rev. Astron. Astrophys.* **30** 143–84

Dworetsky M M, Castelli F and Faraggiana R 1993 *Peculiar versus Normal Phenomena in A-type and Related Stars* (San Francisco, CA: Astronomical Society of the Pacific)

Landstreet J D 1992 Magnetic fields at the surfaces of stars *Astron. Astrophys. Rev.* **4** 35–77

North P, Schnell J and Žižňovský J 1998 *Proc. 26th Meet. and Workshop of the European Working Group on CP Stars, Contrib. Astron. Obs. Skalnaté Pleso* **27** (3)

Schmidt G D 1995 White dwarfs as magnetic stars *Rev. Mod. Astron.* **8** 147–62

John D Landstreet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetic Helicity

Magnetic helicity measures structural properties of a magnetic field such as twist, shear, linking and braiding. It is conserved in the limit of infinite electrical conductivity. In a PLASMA with high conductivity (magnetic Reynolds number $R_m \gg 1$) reconnection preserves magnetic helicity to an excellent approximation.

Magnetic helicity has proven to be a valuable tool for both observers and theorists. It provides a common measure of topological structure. Thus the net twist of field lines within a CORONAL LOOP and the net shear of a coronal arcade can be directly compared. Its conservation provides a constraint on the evolution of a magnetic field. For example, suppose a field undergoes a series of MAGNETIC RECONNECTIONS. Each reconnection takes energy away from the field. If this process continues until no more energy is available, then the final state will have the minimum possible energy given the magnetic helicity and the boundary conditions. This process is known as Taylor relaxation; the final state is generally a linear FORCE-FREE MAGNETIC FIELD.

A field region with large magnetic helicity, whether or not it is force free, stores substantial amounts of energy and exerts pressure on its surroundings. If a region of the solar CORONA begins to build up excessive helicity, then the surrounding fields and plasmas may not be able to confine it. A SOLAR CORONAL MASS EJECTION may then release the excess helicity and energy into the SOLAR WIND.

Magnetic helicity and field topology

Early in the 19th century Gauss discovered an integral which measures the linking of two curves. He employed this integral first to describe the paths of asteroids whose orbits link the Earth's orbit and later to study linked electrical circuits. For two closed directional curves the linking number L counts how many times one curve passes through the other (see figure 1). One way of calculating L without integration involves looking at a drawing of the curves, i.e. a projection onto a plane, and examining the places where one curve crosses over the other. For each crossing, orient the drawing so that the overlying curve points to the right. Assign the crossing the number +1 if the underlying curve points upwards and -1 if it points downwards. The linking number L equals one-half the sum of these numbers over all crossings.

The magnetic helicity inside a region $\mathcal{V}$ is, in essence, the sum of linking numbers over all pairs of field lines within $\mathcal{V}$. To begin with, consider a magnetic field consisting of N tubes of magnetic flux. For now, assume that each tube closes upon itself within $\mathcal{V}$. Each tube has flux $F_i, i = 1, \ldots, N$, and tubes i and j have linking number L_{ij}. Then the helicity

$$K = \sum_{i=1}^{N} \sum_{j=1}^{N} L_{ij} F_i F_j. \qquad (1)$$

The diagonal terms $L_{ii} F_i^2$ measure linking of field lines within the same tube. If the field lines in tube i all twist about a central axis then a typical field line makes L_{ii} turns about the axis.

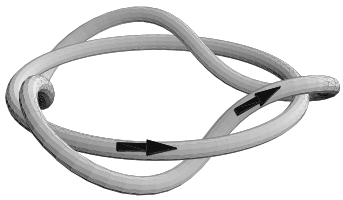


Figure 1. Two linked magnetic tubes with linking number $L = 3$.

Note the analogy with magnetic inductance: the energy of N circuits carrying currents $I_1, \ldots, I_N$ is proportional to $\sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} I_i I_j$, where M_{ii} is a self-inductance and M_{ij} is a mutual inductance. Thus we may call $L_{ii} F_i^2$ the self-helicity of tube i and $2L_{ij} F_i F_j$ the mutual helicity between tubes i and j.

A magnetic surface is a surface $\mathcal{S}$ where the normal component of the magnetic field vanishes, $\boldsymbol{B} \cdot \hat{\boldsymbol{n}}|_{\mathcal{S}} = 0$. If $\mathcal{V}$ is bounded by such a surface, then the magnetic field is called closed: as no field line can escape through $\mathcal{S}$, each line must close upon itself or come arbitrarily close to doing so. Thus to an approximation the field can be described by a collection of N closed flux tubes. In the limit $N \to \infty$ the approximation becomes exact; meanwhile the double-sum expression for K turns into a double integral

$$K_{\mathcal{V}} = \int_{\mathcal{V}} \boldsymbol{A} \cdot \boldsymbol{B} \, \mathrm{d}V$$

where

$$\boldsymbol{A}(\boldsymbol{x}) = -\frac{1}{4\pi} \int_{\mathcal{V}} \frac{(\boldsymbol{x} - \boldsymbol{x}') \times \boldsymbol{B}(\boldsymbol{x}')}{|\boldsymbol{x}' - \boldsymbol{x}|^3} \, \mathrm{d}V'.$$

The vector potential $\boldsymbol{A}(\boldsymbol{x})$ satisfies $\boldsymbol{\nabla} \times \boldsymbol{A} = \boldsymbol{B}$; the above integral form also satisfies $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$. A gauge transformation $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\phi$ will not change K, provided that $\mathcal{V}$ is bounded by a magnetic surface. The magnetic helicity of open fields will be defined below.

Care must be taken if $\mathcal{V}$ is multiply connected, i.e. has a hole like a doughnut. In this case there are vector fields $\boldsymbol{G}$ where $\boldsymbol{\nabla} \times \boldsymbol{G} = \boldsymbol{0}$ but $\boldsymbol{G} \neq \boldsymbol{\nabla}\phi$ for any single-valued function ϕ. (For example if ϕ represents angle in polar coordinates then it is not single valued.) Gauge transformations $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{G}$ will indeed change K.

Other helicities

Given any two divergence-free vector fields $\boldsymbol{V}$ and $\boldsymbol{W}$ ($\boldsymbol{\nabla} \cdot \boldsymbol{V} = \boldsymbol{\nabla} \cdot \boldsymbol{W} = 0$) we can define a helicity integral which measures their mutual linking. Assume here that $\mathcal{V}$ is all of space or that its boundary $\mathcal{S}$ provides a closed surface for both fields ($\boldsymbol{V} \cdot \hat{\boldsymbol{n}}|_{\mathcal{S}} = \boldsymbol{W} \cdot \hat{\boldsymbol{n}}|_{\mathcal{S}} = 0$). Their helicity is then defined by

$$H_{\mathcal{V}}(\boldsymbol{V}, \boldsymbol{W}) = H_{\mathcal{V}}(\boldsymbol{W}, \boldsymbol{V}) = \int_{\mathcal{V}} \boldsymbol{A}_{\boldsymbol{V}} \cdot \boldsymbol{W} \, \mathrm{d}V \qquad (2)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

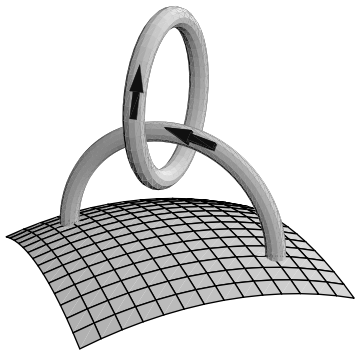


Figure 2. A loop linking a closed tube.

where $\boldsymbol{\nabla} \times \boldsymbol{A_V} = \boldsymbol{V}$. For example $K_\mathcal{V} = H_\mathcal{V}(\boldsymbol{B}, \boldsymbol{B})$.

Suppose $\mathcal{V}$ contains a fluid with flow velocity $\boldsymbol{v}$. If $\boldsymbol{V}$ and $\boldsymbol{W}$ are transported by $\boldsymbol{v}$ without diffusion (e.g. $\partial \boldsymbol{V}/\partial t = \boldsymbol{\nabla} \times (\boldsymbol{v} \times \boldsymbol{V})$) then their helicity will be conserved, $\mathrm{d}H_\mathcal{V}(\boldsymbol{V}, \boldsymbol{W})/\mathrm{d}t = 0$. Thus $K_\mathcal{V}$ is conserved in an infinitely conducting plasma. Similarly, a fluid with zero viscosity transports the vorticity $\boldsymbol{\omega} = \boldsymbol{\nabla} \times \boldsymbol{v}$ without diffusion; the flow thus conserves the kinetic helicity

$$H_\mathcal{V}(\omega, \omega) = \int_\mathcal{V} \boldsymbol{v} \cdot \boldsymbol{\omega}\, \mathrm{d}v.$$

The general helicity formula (2) is linear in each term, e.g.

$$H_\mathcal{V}(a\boldsymbol{V}_1 + \boldsymbol{V}_2, \boldsymbol{W}) = aH_\mathcal{V}(\boldsymbol{V}_1, \boldsymbol{W}) + H_\mathcal{V}(\boldsymbol{V}_2, \boldsymbol{W}).$$

Magnetic helicity in an open volume

A vector field $\boldsymbol{V}$ in a volume $\mathcal{V}$ with boundary $\mathcal{S}$ is open if some of its flux crosses the boundary, i.e. $\boldsymbol{V} \cdot \hat{n}|_\mathcal{S} \neq 0$ somewhere on $\mathcal{S}$. The concept of linking will now be extended from closed fields to open fields. We first note that a closed curve may link an open curve, as in figure 2. Similarly, if $\boldsymbol{V}$ is open, but $\boldsymbol{W}$ is closed, then we may still employ the formula (2) to measure the linking of $\boldsymbol{V}$ and $\boldsymbol{W}$. Here the vector potential in the integral must always refer to the open field to ensure gauge invariance. (By Stokes' theorem the line integral of $\boldsymbol{A_V}$ about a closed field line of $\boldsymbol{W}$, i.e. $\oint \boldsymbol{A_V} \cdot \boldsymbol{W}|\boldsymbol{W}|^{-1}\, \mathrm{d}\ell$, equals the net flux of $\boldsymbol{V}$ linked by the $\boldsymbol{W}$ line. Integrating over the $\boldsymbol{W}$ flux leads to equation(2).)

Our strategy for measuring the magnetic helicity of an open magnetic field $\boldsymbol{B}$ will be to decompose $\boldsymbol{B}$ into a closed field and an open field with minimal structure, then sum the self- and mutual helicities of these fields. Of the many possible fields inside $\mathcal{V}$ consistent with the boundary flux $\boldsymbol{B} \cdot \hat{\boldsymbol{n}}|_\mathcal{S}$, one unique field minimizes the total magnetic energy: the potential field $\boldsymbol{P}$ where $\boldsymbol{\nabla} \times \boldsymbol{P} = \boldsymbol{0}$. This field has zero current inside $\mathcal{V}$; its source currents must exist outside of $\mathcal{V}$ or as surface currents in $\mathcal{S}$. Owing to its unique properties, we choose $\boldsymbol{P}$ to be our open field.

We decompose $\boldsymbol{B}$ into the sum of the potential field $\boldsymbol{P}$ and the closed field

$$\boldsymbol{B}_{cl} = \boldsymbol{B} - \boldsymbol{P} \qquad \boldsymbol{B}_{cl} \cdot \hat{n}|_S = 0.$$

Then

$$\begin{aligned} K_\mathcal{V} &= H_\mathcal{V}(\boldsymbol{B}, \boldsymbol{B}) \\ &= H_\mathcal{V}(\boldsymbol{B}_{cl}, \boldsymbol{B}_{cl}) + 2H_\mathcal{V}(\boldsymbol{P}, \boldsymbol{B}_{cl}) + H_\mathcal{V}(\boldsymbol{P}, \boldsymbol{P}). \end{aligned}$$

The first two terms measure the self-linking of a closed field and the mutual linking between an open and a closed field. Both of these are well defined by equation (2). Unfortunately, the last term cannot be defined by equation (2) without ambiguity. The accepted procedure simply assigns the value 0 to $H_\mathcal{V}(\boldsymbol{P}, \boldsymbol{P})$. A similar procedure assigns zero voltage to Earth in an electrical circuit. As the potential field $\boldsymbol{P}$ occupies the minimum energy state of $\mathcal{V}$, the assignment of zero helicity is natural. With this definition the helicity becomes

$$K_\mathcal{V} = \int_\mathcal{V} (\boldsymbol{A} + \boldsymbol{A_P}) \cdot (\boldsymbol{B} - \boldsymbol{P})\, \mathrm{d}V. \tag{3}$$

Helicity dissipation

Suppose the boundary flux $\boldsymbol{B} \cdot \hat{\boldsymbol{n}}|_\mathcal{S}$ is fixed. Then no helicity can cross the boundary and the helicity can only change as a result of dissipation (see next section for transfer of helicity across boundaries). We will use Ohm's law to express the electric field in terms of the electric current and the magnetic diffusivity, $\boldsymbol{E} = \eta \boldsymbol{J}$.

From the definition of $K_\mathcal{V}$ (equation (3)) and Maxwell's equations

$$\frac{\mathrm{d}K_\mathcal{V}}{\mathrm{d}t} = -2 \int_\mathcal{V} \eta \boldsymbol{J} \cdot \boldsymbol{B}\, \mathrm{d}V.$$

Note that the magnetic energy is proportional to $W_\mathcal{V} = \int_\mathcal{V} B^2\, \mathrm{d}V$ with dissipation $|\mathrm{d}W_\mathcal{V}/\mathrm{d}t| = 2\int_\mathcal{V} \eta J^2\, \mathrm{d}V$. Thus, using a Schwarz inequality,

$$\left|\frac{\mathrm{d}K_\mathcal{V}}{\mathrm{d}t}\right| \leq \left(2\eta W_\mathcal{V} \left|\frac{\mathrm{d}W_\mathcal{V}}{\mathrm{d}t}\right|\right)^{1/2}. \tag{4}$$

In order to apply this inequality to reconnection events, define a length scale $L \equiv |K_\mathcal{V}|/W_\mathcal{V}$. This length is comparable with the size of the volume if the magnetic field has significant helical structure. For example, a linear force-free field inside a spherical magnetic surface of radius R has $L \approx 0.31R$. A dissipation time τ_d can now be defined as $\tau_d = L^2/\eta$. Consider an arbitrary reconnection or dissipation process occurring over a time Δt in an isolated volume. Integrating equation (4) over time gives

$$\left|\frac{\Delta K}{K}\right| \leq \left(\frac{\Delta t}{\tau_d}\right)^{1/2}. \tag{5}$$

Thus for any fast reconnection event ($\Delta t \ll \tau_d$), ΔK will be negligible. For flares in the solar corona with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

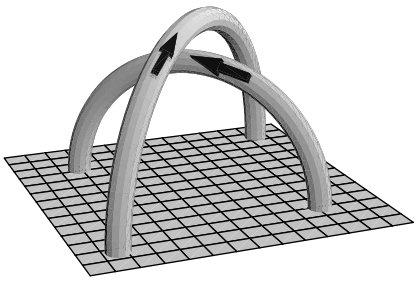


Figure 3. Two crossed loops.

$\Delta t \approx 1000$ s, $L \approx 10^3$ km and $\eta \approx 10^{-6}$ km^2 s^{-1} we find $\tau_d \approx 10^{12}$ s, and $|\Delta K/K| < 3\times10^{-5}$. In practice, even these small limits significantly overestimate ΔK, as they make no assumptions about the reconnection geometry. As a crude estimate, if we assume that reconnection takes place only in a small fraction ϵ of the total volume then we can effectively replace η by $\epsilon\eta$ and τ_d by τ_d/ϵ. Then the bound (5) on ΔK has an additional factor $\epsilon^{1/2}$.

Helicity transport

In this section we ignore dissipation and consider the transport of helicity across a boundary S due to fluid motions. Let $\boldsymbol{v}$ be the fluid velocity. Also recall that $\boldsymbol{A_P}$ is a vector potential for $\boldsymbol{P}$. Inside $\mathcal{V}$, let $\widetilde{\boldsymbol{A}}_{\boldsymbol{P}}$ be the unique vector potential satisfying $\boldsymbol{\nabla}\cdot\widetilde{\boldsymbol{A}}_{\boldsymbol{P}} = 0$, $\widetilde{\boldsymbol{A}}_{\boldsymbol{P}}\cdot\hat{n}|_S = 0$. Then one can show that

$$\frac{\mathrm{d}K_{\mathcal{V}}}{\mathrm{d}t} = 2\oint_S[(\widetilde{\boldsymbol{A}}_{\boldsymbol{P}}\cdot\boldsymbol{v})\boldsymbol{B} - (\widetilde{\boldsymbol{A}}_{\boldsymbol{P}}\cdot\boldsymbol{B})\boldsymbol{V}]\cdot\mathrm{d}\boldsymbol{S}.$$

The first term measures the effect of motions parallel to the boundary, whereas the second measures the effect of motion across the boundary.

If S is the plane $z = 0$ and $\mathcal{V} = \{z > 0\}$, then one finds

$$\widetilde{\boldsymbol{A}}_{\boldsymbol{P}}(\boldsymbol{x}) = \frac{1}{2\pi}\int_{\mathrm{S}}\frac{B_z(\boldsymbol{x}')}{r}\hat{\boldsymbol{\theta}}_{\boldsymbol{x}\boldsymbol{x}'}\,\mathrm{d}S$$

where $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}\boldsymbol{x}'} = \hat{\boldsymbol{z}}\times\hat{\boldsymbol{r}}$ is the angular direction at $\boldsymbol{x}$ about the integration point $\boldsymbol{x}'$. For example, suppose the flux through $z = 0$ consists of two flux tubes of flux F_1, F_2 centered at $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Suppose these tubes rotate with angular velocities w_1 and w_2. Also suppose that the tubes move about each other with angular velocity $\mathrm{d}\theta_{12}/\mathrm{d}t = [\boldsymbol{v}(\boldsymbol{x}_1) - \boldsymbol{v}(\boldsymbol{x}_2)]\cdot\hat{\boldsymbol{\theta}}_{\boldsymbol{x}\boldsymbol{x}'}$. Then

$$\frac{\mathrm{d}K_{\mathcal{V}}}{\mathrm{d}t} = -\frac{1}{2\pi}\left(w_1F_1^2 + w_2F_2^2 + 2\frac{\mathrm{d}\theta_{12}}{\mathrm{d}t}F_1F_2\right). \qquad (6)$$

The first two terms describe the twisting of individual tubes, while the final term describes mutual braiding.

Helicity of coronal loops

The helicity transport equations often greatly simplify calculations. The helicity of a field depends only on its present geometry and ignores its past history. To illustrate, we calculate the mutual helicity of two coronal loops (see figure 3). For simplicity, let the corona be the upper half-space $\{z > 0\}$. The photospheric ($z = 0$) ends of the loops are assumed to be small and located at the points $\boldsymbol{x}_1^+, \boldsymbol{x}_1^-$ for loop 1 and $\boldsymbol{x}_2^+$, and $\boldsymbol{x}_2^-$ for loop 2. Here $B_z > 0$ at $\boldsymbol{x}_1^+$ and $\boldsymbol{x}_2^+$. Also, if the loops cross over each other when viewed from above, then let the upper loop be loop 1. Finally, let θ_{1+2-} be the angle of the line segment from $\boldsymbol{x}_1^+$ to $\boldsymbol{x}_2^-$, measured with respect to the x-axis.

Construct the field by first placing loop 2 in its present position, but with loop 1 at infinity. This configuration has zero mutual helicity. Next move loop 1 into place, using equation (6) to measure the helicity change. Because we are moving the upper loop, none of the angles change by more than π. The result is (ignoring the self-helicities due to twist)

$$K_{\mathcal{V}} = \frac{F_1F_2}{\pi}(\theta_{1+2-} + \theta_{1-2+} - \theta_{1+2+} - \theta_{1-2-}).$$

Addition of helicities

Suppose we divide space into two regions $\mathcal{V}_1$ and $\mathcal{V}_2$ with boundary S. For example, S could be the solar photosphere, with $\mathcal{V}_1$ the solar interior and $\mathcal{V}_2$ the corona. Let the magnetic field $\boldsymbol{B} = \boldsymbol{B}_1$ in $\mathcal{V}_1$ and $\boldsymbol{B} = \boldsymbol{B}_2$ in $\mathcal{V}_2$. We will notate this as $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2]$. We wish to compare the helicity integrated over all space

$$K(\boldsymbol{B}) = K([\boldsymbol{B}_1, \boldsymbol{B}_2]) = \int \boldsymbol{A}\cdot\boldsymbol{B}\,\mathrm{d}V$$

with the helicities $K_{\mathcal{V}_1}$ and $K_{\mathcal{V}_2}$ of the regions $\mathcal{V}_1$ and $\mathcal{V}_2$.

The corresponding potential fields $\mathcal{V}_1$ and $\mathcal{V}_2$ are $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. One can show that

$$K_{\mathcal{V}_1} = K([\boldsymbol{B}_1, \boldsymbol{B}_2]) - K([\boldsymbol{P}_1, \boldsymbol{B}_2]). \qquad (7)$$

Thus $K_{\mathcal{V}_1}$ represents the helicity K of all space measured relative to that of the field with zero current in $\mathcal{V}_1$, i.e. $K([\boldsymbol{P}_1, \boldsymbol{B}_2])$. As $K_{\mathcal{V}_1}$ can also be calculated using equation (3), which only integrates over $\mathcal{V}_1$, $K_{\mathcal{V}_1}$ is independent of the exterior field $\boldsymbol{B}_2$.

Similarly, we may calculate $K_{\mathcal{V}_2}$ as

$$K_{\mathcal{V}_2} = K([\boldsymbol{P}_1, \boldsymbol{B}_2]) - K([\boldsymbol{P}_1, \boldsymbol{P}_2]) \qquad (8)$$

where we have exercised our freedom to choose the exterior field to be $\boldsymbol{P}_1$ without changing the result. Summing equations (7) and (8) gives us the addition formula

$$K_{\mathcal{V}_1} + K_{\mathcal{V}_2} = K([\boldsymbol{B}_1, \boldsymbol{B}_2]) - K([\boldsymbol{P}_1, \boldsymbol{P}_2]).$$

Note that the field $[\boldsymbol{P}_1, \boldsymbol{P}_2]$ is not potential; it has surface currents on S. The last term always vanishes when the boundary S is a plane or a sphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Berger M A and Field G B 1984 The topological properties of magnetic helicity *J. Fluid Mech.* **147** 133–48

Marsh G E 1996 *Force-Free Magnetic Fields: Solutions, Topology, and Applications* (Singapore: World Scientific)

Moffatt H K, Zaslavsky G M, Comte P and Tabor M (eds) 1992 *Topological Aspects of the Dynamics of Fluids and Plasmas (NATO ASI Series 218)* (Dordrecht: Kluwer)

Ricca R L and Berger M A 1996 Topological ideas and fluid mechanics *Phys. Today* (December) 28–34

Mitchell A Berger

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetic Reconnection

Magnetic reconnection is a fundamental dynamical process in highly conductive plasmas. It can be regarded as the process that removes the following difficulty. On typical dynamical time scales a sufficiently hot spatially extended PLASMA behaves approximately as an ideal fluid in the sense that resistive effects are ignorable. As a consequence, the magnetic field is 'frozen' to the plasma motion and magnetic topology is conserved. This sets strong limitations on the accessible dynamical states. Large-scale magnetic flux tubes, which are strongly stretched out by the plasma pressure, as for instance observed in PLANETARY MAGNETOSPHERES or in stellar CORONAS, would be unable to release large amounts of their energy and return to a correspondingly relaxed state, as long as the plasma is trapped in the flux tubes. In other words, efficient transformation of magnetic to kinetic energy would largely be ruled out in ideal plasmas. There would be no obvious process that could counteract the generation of magnetic flux by dynamo processes and the magnetic fields in many space and astrophysical situations would grow secularly. Also, plasmas with magnetic fields of different origin would not be able to mix. Beginning in the late 1950s, several authors, including P A Sweet, E N Parker, H E Petschek and J W Dungey, introduced magnetic reconnection as the central process allowing for efficient magnetic to kinetic energy conversion in SOLAR FLARES and for interaction between the magnetized interplanetary medium and the MAGNETOSPHERE OF EARTH.

How does reconnection circumvent the difficulty associated with frozen-in magnetic fields? Resistive dissipation is more effective the more the electric current is localized to regions with a small spatial scale length. Thus, in reconnection a small-scale structure is generated in some region, such that there the constraint of ideal dynamics is broken. The interesting aspect is that a *local* non-ideality can have a *global* effect. Under such circumstances highly conducting plasma structures are able to transform magnetic to kinetic energy in an efficient way and the magnetic topology can change. According to a major line of present thinking, this is what happens in solar flares or magnetospheric substorms, and possibly in many other plasma processes in the universe.

Basic model

The formal description of reconnection requires the choice of a dynamical model. Here we confine the discussion to magnetohydrodynamics, where we allow for a finite resistivity (resistive MHD or 'RMHD') as the only non-ideal transport process. The corresponding basic equations consist of a combination of fluid dynamics and electrodynamics:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0 \tag{1}$$

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} + \rho \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\nabla p + \boldsymbol{j} \times \boldsymbol{B} \tag{2}$$

$$\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B} = \boldsymbol{j}/\sigma \tag{3}$$

$$\frac{\partial e}{\partial t} + \nabla \cdot (e\boldsymbol{v}) = -p \nabla \cdot \boldsymbol{v} + j^2/\sigma \tag{4}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{5}$$

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j} \tag{6}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{7}$$

Here, ρ, $\boldsymbol{v}$, p, $\boldsymbol{j}$, $\boldsymbol{B}$, $\boldsymbol{E}$, σ^{-1}, e and μ_0 denote respectively mass density, velocity, pressure, current density, magnetic field, electric field, resistivity, plasma energy density and vacuum permeability (see the article on MAGNETOHYDRODYNAMICS). Here we mention only the fact that the equations (1)–(7) imply the conservation of energy. The balance of mechanical and electromagnetic energy, respectively, take the form

$$\frac{\partial}{\partial t}\left(\frac{\rho v^2}{2} + u\right) + \nabla \cdot \left(\left(\frac{\rho v^2}{2} + u + p\right)\boldsymbol{v}\right) = \boldsymbol{j} \cdot \boldsymbol{E} \tag{8}$$

$$\frac{\partial}{\partial t}\left(\frac{B^2}{2\mu_0}\right) + \nabla \cdot \left(\frac{1}{\mu_0}\boldsymbol{E} \times \boldsymbol{B}\right) = -\boldsymbol{j} \cdot \boldsymbol{E}. \tag{9}$$

Adding these two equations gives conservation of energy

$$\frac{\partial}{\partial t}\left(\frac{\rho v^2}{2} + u + \frac{B^2}{2\mu_0}\right) + \nabla \cdot \left(\frac{\rho v^2}{2}\boldsymbol{v} + (u + p)\boldsymbol{v} + \frac{1}{\mu_0}\boldsymbol{E} \times \boldsymbol{B}\right) = 0. \tag{10}$$

For some purposes it has proved useful to impose the condition of incompressibility on the flow velocity

$$\nabla \cdot \boldsymbol{v} = 0 \tag{11}$$

replacing (4). This simplifies the problem significantly. It should, however, be kept in mind that for an incompressible flow, an energy conservation law of the form of (10) is not available. However, mass conservation and momentum balance are still described appropriately.

In a resistive fluid the importance of resistivity is measured by the Lundquist number

$$S = \frac{v_A L}{\eta} \tag{12}$$

where $v_A = B/\sqrt{\mu_0 \rho}$ is the Alfvén velocity, $\eta = (\mu_0 \sigma)^{-1}$ the magnetic diffusivity and L a typical (global) scale length. Alternatively, one uses the magnetic reynolds number $R_m = vL/\eta$, where the Alfvén velocity is replaced by a typical plasma velocity v. (In the literature the expression 'magnetic Reynolds number' frequently is also used for the quantity S.) Large values of S or R_m, which are typical for space and astrophysical plasmas, correspond to small resistive effects. In the limit of large S or R_m, the terms involving resistivity can be neglected (unless singularities form) and equations (1)–(7) reduce to the equations of ideal magnetohydrodynamics (IMHD).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

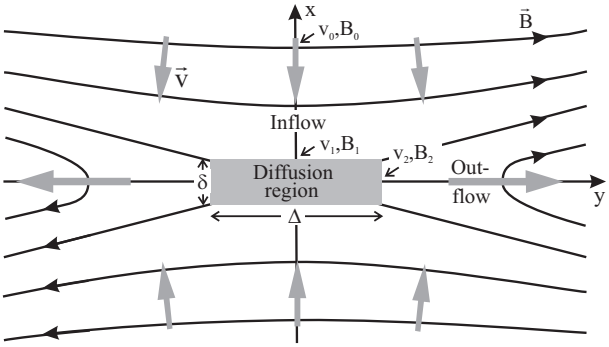


Figure 1. Qualitative pattern of two-dimensional reconnection.

Importantly, IMHD implies conservation of magnetic field line topology. In IMHD this fact can also be expressed by the property that two plasma elements that are connected by a magnetic field line at one time are connected by a magnetic field line at any later time (magnetic line conservation). Furthermore, the magnetic flux through an arbitrary contour transported by the plasma velocity field is also conserved. These properties provide the quantitative background for the dynamical constraints of IMHD mentioned above. In particular, they imply that large-scale topological reconfigurations of the magnetic field structure, as assumed to be associated with stellar and magnetospheric activity, are ruled out. In the following, we summarize in what sense magnetic reconnection resolves that dilemma and what is known at present about that process.

Two-dimensional reconnection

The simplest geometry in which reconnection may be described has two spatial dimensions, requiring the presence of an ignorable coordinate in three-dimensional physical space. In this section Cartesian coordinates x, y, z are used and it is assumed that the physical quantities are independent of z. We will first consider steady states and then introduce time dependence.

Steady-state reconnection

The basic configuration of two-dimensional steady-state reconnection is shown in figure 1. All field quantities are independent of time. Also, the magnetic field $\boldsymbol{B}$ and the plasma velocity $\boldsymbol{v}$ are assumed to lie in the x, y-plane, while for the electric field a non-vanishing z-component is admitted. The plasma is highly ideal such that the Lundquist number S (12) is much larger than 1.

To obtain an efficient conversion of magnetic to kinetic energy (along the trajectories of fluid elements) it is appropriate to assume a stagnation-type flow field $\boldsymbol{v}$ and oppositely directed magnetic fields in the upper and lower part of the inflow region (figure 1). The magnetic field vanishes at the origin (neutral point); viewed three-dimensionally a neutral line (line on which $B = 0$) extends along the z-axis.

Since S is large, for a smooth plasma flow with maximum gradients associated with the global length scale L the frozen-in condition would not allow annihilation of magnetic flux to any significant extent. This difficulty is avoided by the presence of a 'diffusion region' near the neutral line, where the resistive term j/σ in Ohm's law is much larger than in the approximately ideal environment ('external region'), typically by an enhancement of j_z. The diffusion region has length scales δ and Δ (figure 1) with $L \geq \Delta \geq \delta$. A locally defined Lundquist number, where L is replaced by δ in (12) can be considerably smaller than the global Lundquist number, indicating that in the diffusion region resistive diffusion can play an important role. There, the plasma and magnetic fields may decouple effectively, so that field annihilation along the fluid path becomes possible.

Under the present conditions (5) implies that E_z is a positive constant, say E_0. The presence of the diffusion region allows for a non-vanishing value of E_0, because otherwise (i.e. under ideal conditions with $\boldsymbol{j}/\sigma$ negligible) the z-component of equation (3) would require $E_z = 0$ at the neutral point, such that E_0 would have to vanish.

Another important property of the present geometry (shown in figure 1) is that $\partial B_y/\partial x > \partial B_x/\partial y$ or $j_z > 0$. Therefore, $\boldsymbol{E} \cdot \boldsymbol{j} = E_0 j_z > 0$ holds, which by (8) or (9) implies that magnetic energy is converted to kinetic energy. In fact, from (8) one finds

$$\frac{\partial}{\partial s}\left(\frac{v^2}{2} + \frac{u+p}{\rho}\right) > 0 \tag{13}$$

where (1) was used assuming that $\rho v \neq 0$, and s denotes the arc length of the trajectory of the plasma element (increasing in the direction of $\boldsymbol{v}$). Note that the thermal part on the left-hand side of (13) is enthalpy per unit mass rather than internal energy per unit mass, because the work done by the pressure force is included.

For a discussion of the consequences of mass and momentum conservation we specialize the resistive MHD equations (1)–(7) further, using the incompressibility condition (11) with constant density ρ_0 instead of (4). Then the resistive RMHD equations for a steady state assume the form

$$\rho \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\nabla p + \boldsymbol{j} \times \boldsymbol{B} \tag{14}$$

$$E_0 + \boldsymbol{v} \times \boldsymbol{B} \cdot \boldsymbol{e}_z = j_z/\sigma \tag{15}$$

$$\nabla \cdot \boldsymbol{v} = 0 \tag{16}$$

$$(\nabla \times \boldsymbol{B}) \cdot \boldsymbol{e}_z = \mu_0 j_z \tag{17}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{18}$$

Quantities in the outer inflow region will be characterized by their magnitudes at the point $(x_0, 0)$ where the positive x-axis crosses the boundary, and are labeled by the subscript zero, in particular (in addition to ρ_0, E_0)

$$p_0 = p(x_0, 0), \quad B_0 = B_y(x_0, 0), \quad v_0 = -v_x(x_0, 0)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(v_0 and B_0 are indicated in figure 1). Analogously, the subscripts '1' and '2' refer to the center inflow and outflow points on the boundary of the diffusion region (figure 1) and the subscript 'nl' is used for quantities on the neutral line.

Keeping the shape of the boundary fixed, except for the global scale length L, it can be expected that under the present conditions ρ_0, p_0, v_0, B_0, L and σ is a set of control parameters. Note, however, that in view of the nonlinearity of the problem a solution is not guaranteed for arbitrary parameter choices.

To disregard configurations that are merely the result of a similarity transformation, it is of interest to note that from these parameters three independent dimensionless quantities may be formed, which are conveniently chosen as

$$M_0 = \frac{v_0}{a_0}, \quad S_0 = \frac{a_0 L}{\eta}, \quad \beta_0 = \frac{2\mu_0 p_0}{{B_0}^2} \tag{19}$$

where a_0 is the (inflow) Alfvén velocity $B_0/\sqrt{\mu_0\rho_0}$. The earlier discussion of steady-state reconnection in the literature largely ignores the parameter β_0. This seems justified if β_0 is negligibly small or if pressure is constant in the external region. (It is only the gradient of the pressure that counts.) Then, reconnection is a two-parameter process, for instance described by M_0 and S_0. The parameter M_0 is regarded as of particular interest and is usually called reconnection rate. It measures the velocity with which the plasma enters the region of consideration (normalized by the local Alfvén velocity). The so-defined reconnection rate should not be confused with the rate of magnetic flux reconnection, which is defined by the rate at which flux conservation is violated in the reconnection process, which, in the present case, is given by the electric field component E_z along the neutral line, which equals $E_0 = v_0 B_0$.

There is no fully satisfactory analytical treatment of the system of equations (14)–(18). There are solutions for the external (ideal) region and solutions for the diffusion region, based on singular asymptotic expansions. However, a rigorous matching of such solutions has not yet been achieved. In this situation one introduces intuitive assumptions or simplifications. Much of the discussion in the literature is based on the following approximate picture.

Consistent with $j_z > 0$, let us assume that the aspect ratio $\kappa = \Delta/\delta$ is large compared to 1, that derivatives with respect to x are large compared with derivatives with respect to y and that $|B_x| \ll B_0$. Pressure is treated as constant in the external region. Then approximate relations are obtained in the following way:

Condition of incompressibility (11):

$$v_1 \Delta = v_2 \delta.$$

x-component of momentum balance (14) at $y = 0$:

$$p_1 + \frac{{B_1}^2}{2\mu_0} = p_{\rm nl}.$$

y-component of momentum balance at $x = 0$, ignoring B_x:

$$\frac{\rho_2 {v_2}^2}{2} + p_2 = p_{\rm nl}.$$

Ohm's law:

$$E_0 = v_1 B_1 = v_2 B_2 = j_{\rm nl}/\sigma.$$

Ampère's law (6) (replacing the derivative by a difference quotient)

$$j_{\rm nl} = \frac{B_1}{\mu_0 \delta}.$$

Combining these equations and using that, in view of the assumptions, $\rho_2 = \rho_1 = \rho_0$, $p_1 = p_2 = p_0$ one obtains

$$v_2 = a_1 \tag{20}$$

$$M_1 = \frac{1}{\kappa} = \frac{1}{\sqrt{S_1}} \tag{21}$$

$$\frac{M_1}{M_0} = \left(\frac{B_0}{B_1}\right)^2 \tag{22}$$

$$\frac{\Delta}{L} = \frac{S_1 B_0}{S_0 B_1}. \tag{23}$$

This system of equations has to be completed by an equation for the ratio B_0/B_1 which requires a more complete solution of equations (14)–(18). In the absence of such a solution one introduces an additional condition as an *ad hoc* assumption, or from the external solution alone, or on the basis of numerical computations. We give three examples.

(a) Sweet–Parker model. Here it is assumed that the diffusion region is a thin extended structure such that Δ becomes of the order of L. For simplicity, let us set $\Delta = L$. The external region is largely homogeneous such that approximately $B_1 = B_0$ and $S_1 = S_0$. Under these conditions, (21) gives the reconnection rate as

$$M_0 = \frac{1}{\sqrt{S_0}}.$$

This rate is generally regarded as too low to be relevant for typical conditions in stellar atmospheres and space plasmas because of their large Lundquist numbers.

(b) Petschek's model. In this model it is assumed that $\Delta \ll L$. In that case, it is necessary to consider the presence of slow-mode shock waves (here in the limit of incompressibility) which implies that B_1 may be considerably smaller than B_0. Approximately, one finds $B_1/B_0 = 1 - 4M_0/(\pi \ln(R_{m_0}))$. The maximum reconnection rate occurs near $B_1/B_0 = 1/2$, such that

$$M_0 < \frac{\pi}{8}\frac{1}{\ln R_{m_0}}.$$

Typically this reconnection rate is considerably larger than that of the Sweet–Parker process.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

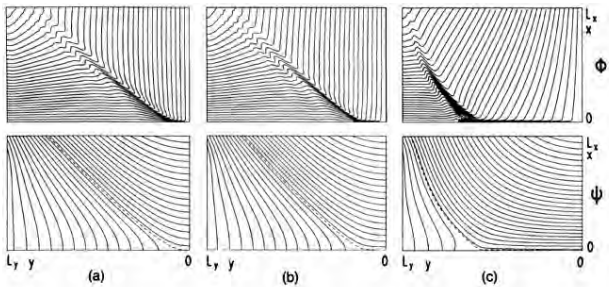


Figure 2. Numerical solutions of equations (25) and (26) with $\Psi = \hat{A}$ and $\Phi = \hat{D}$ by Biskamp. In (*b*) and (*c*) the Lundquist number of (*a*) is increased by factors of 2 and 4, respectively (from Biskamp 1986).

(c) Further reconnection models. Several authors (e.g. W I Axford, B U Ö Sonnerup, E R Priest, T G Forbes) have generalized the models by Sweet and Parker and by Petschek in various respects. The most general are the fast reconnection models of Priest and Forbes. They included electrical currents in the external region and obtained a description that contains the Sweet–Parker and Petschek models as particular cases.

For numerical studies (as for other purposes) it is convenient to represent $\boldsymbol{v}$ and $\boldsymbol{B}$ by single flux functions $D(x, y)$ and $A(x, y)$. This is possible because of the vanishing divergence of both fields and the absence of z-components,

$$\boldsymbol{B} = \nabla A \times \boldsymbol{e}_z, \quad \boldsymbol{v} = \nabla D \times \boldsymbol{e}_z. \tag{24}$$

Then, one eliminates the electric current density by using (17), and the pressure by taking the curl of the momentum equation. The remaining equations of the system (14)–(18) are usually written in non-dimensional form (here non-dimensional quantities carry the hat-label), such that A is normalized by $B_0 L$, the velocity potential D by $a_0 L$ and coordinates by L

$$[\Delta \hat{D}, \hat{D}] = [\Delta \hat{A}, \hat{A}] \tag{25}$$

$$M - [\hat{A}, \hat{D}] = -\frac{1}{S_0} \Delta \hat{A}. \tag{26}$$

For functions $f(x, y)$, $g(x, y)$ the symbol $[f, g]$ is defined by

$$[f, g] = \frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}.$$

Equations (25) and (26) have been solved numerically for a variety of boundary conditions by several groups. Figure 2 shows, for example, a result by Biskamp, demonstrating that a Sweet–Parker current sheet rapidly develops for increasing S_0.

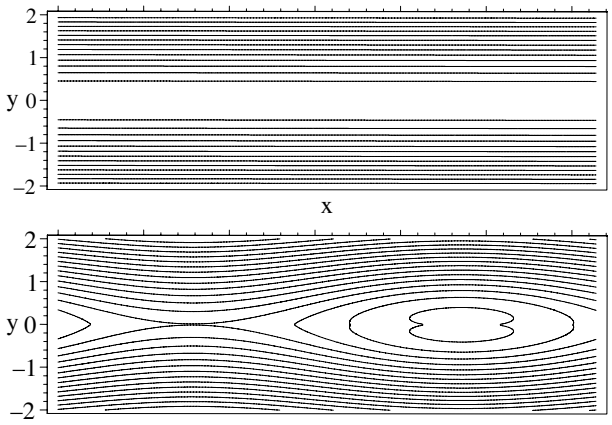


Figure 3. Magnetic field lines of the unperturbed Harris sheet (upper panel) and the linear tearing mode (lower panel).

Time-dependent reconnection

Although, historically, steady-state reconnection has been given great deal of attention, it seems that in many cases magnetic reconnection occurs as a time-dependent process. Several features of steady-state reconnection are also present in typical time-dependent (two-dimensional) cases, such as a neutral line and an associated stagnation-flow pattern. This analogy is particularly close for driven reconnection, where—as in steady states—the plasma inflow is determined by boundary conditions.

A qualitatively different case arises when reconnection occurs as an unstable process. The prototype of an instability involving reconnection is the tearing mode (suggested by H P Furth, J Killeen and M N Rosenbluth). A plane current sheet located in an infinite domain undergoes spontaneous formation of magnetic islands (figure 3). Resistivity plays a similar role as in steady states: it is important only in regions of strong current concentration. Assuming that the unperturbed configuration does not involve such concentrations, it can be described in the limit of $S \to \infty$. The classical example is the Harris sheet, where the unperturbed magnetic field $\boldsymbol{B}$, the flux function A and the plasma pressure p are given (in dimensionless form) by

$$\boldsymbol{B} = -\tanh(x)\boldsymbol{e}_y, \quad A = \ln(\cosh(x)), \quad p = \frac{1}{\cosh^2(x)}$$

which is a static solution of (25) and (26) for infinite S.

The instability generates the required current concentration spontaneously. The dynamical evolution is described by equations (25) and (26), if generalized to include time dependence. In view of the time dependence, it is appropriate to derive the electric field from the time dependence of the flux function A, rather than from an electric potential. In dimensionless form one obtains the following linearized equations for the perturbations ϕ and ψ of the velocity potential D and the flux function A

$$\Delta \frac{\partial \psi}{\partial t} = [\Delta a, A] + [\Delta A, a] \tag{27}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\frac{\partial a}{\partial t} = \frac{1}{S}\Delta a - [A, \psi]. \tag{28}$$

Choosing modes of the form

$$\psi(x, y, t) = \hat{\psi}(y)\, \mathrm{e}^{\mathrm{i}\alpha y + qt}$$

with a corresponding expression for ϕ, (27) and (28) give two ordinary differential equations for $\hat{\psi}$ and $\hat{\phi}$. These equations are solved analytically by a singular perturbation method for the regime

$$\frac{1}{S} \ll |q| \ll 1, \quad |q^2| \ll \alpha^2 < 1.$$

The essential aspect is the occurrence of a thin region around $y = 0$ of width $\epsilon = (q/(\alpha^2 S))^{1/4}$, where the current density becomes large. Using appropriate scaling in this region and in the external region, one finds explicit solutions to lowest significant order in ϵ. The matching condition determines the dispersion relation, i.e. q as a function of α

$$\hat{q} = \frac{\sqrt{\lambda}}{\pi} \frac{\Gamma\left(\frac{\lambda+1}{4}\right)}{\Gamma\left(\frac{\lambda+3}{4}\right)} (1 - \alpha^2)(1 - \lambda^2)$$

where

$$\lambda = \frac{\hat{q}^{3/2}}{\hat{\alpha}}, \quad \hat{q} = qS^{1/2}, \quad \hat{\alpha} = \alpha S^{1/4}.$$

The tearing mode develops a series of magnetic islands with corresponding X-type and O-type neutral lines (figure 3). The local structure near the X-line resembles the steady-state reconnection pattern of figure 1.

For the reconnection processes associated with solar flares and magnetospheric substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS) more realistic two- and three-dimensional models have been developed (pioneered by J Birn, A Otto, T G Forbes, Z Mikic and others). Figure 3 gives a qualitative sketch of the magnetic field structure as it develops with time. The original equilibrium configuration becomes unstable by a process which is a generalization of the tearing mode shown in figure 3. During its nonlinear evolution a plasmoid forms, which grows, becomes accelerated and eventually leaves the system, carrying a substantial amount of energy that was stored in the original equilibrium. Processes of this kind have been suggested to be relevant for magnetospheric substorms, solar flares and SOLAR CORONAL MASS EJECTIONS. For the magnetosphere it is believed that the onset of the non-ideal (e.g. resistive) process is related to the formation of a thin current sheet late in phase (a) in figure 4.

In the case of three-dimensional modeling one encounters new aspects, as compared with reconnection in two dimensions, which are discussed in the following section.

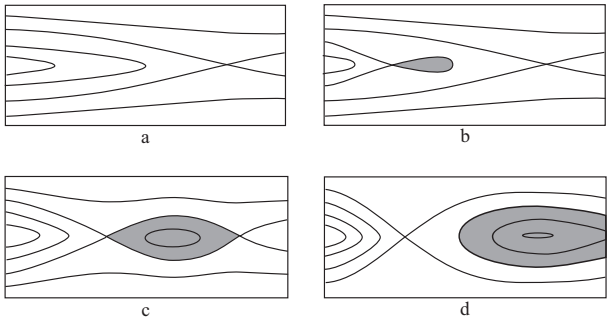


Figure 4. Plasmoid formation and ejection in a stretched magnetic field configuration.

Three-dimensional reconnection

The two-dimensional models discussed so far seem to be realistic for reconnection occurring in three-dimensional space only if the z-dependence is small and if the extent of the reconnection region along the perpendicular direction (z-direction) is large enough that effects of the edges can be neglected. Moreover, it requires that magnetic flux of exactly opposite direction is convected along the x-axis into the reconnection region. Each of these assumptions is doubtful, and so a generalization with a component of the magnetic field along the invariant direction is required which allows for magnetic flux to approach the reconnection region with a non-vanishing z-component. This is most simply realized by adding a constant B_z-component in the model given by equations (14)–(18). This requires an additional (E_x, E_y) component of the electric field, which has the form of a gradient ($\nabla(B_z D)$ for the representation of $\boldsymbol{v}$ given in equation (24)). It therefore does not destroy the stationarity of these models nor does it modify the momentum equation.

Although the additional B_z-component seems to be a minor modification, it gives rise to several fundamental questions about the notion of reconnection. In two dimensions ($B_z = 0$) reconnection is usually defined by the existence of an X-type neutral point and a flow of stagnation type which transports magnetic flux across the separatrices, i.e. the field lines which end at the neutral point and separate the magnetic flux of the inflow and outflow regions (see figure 1). With the additional B_z component, the former neutral line of the two-dimensional models now becomes an ordinary magnetic field line and the former separatrices, or separatrix surfaces, respectively, do not exist anymore or, if the notion of a separatrix is applied to the projection of the field onto the plane perpendicular to the field line, they are not unique. (The latter can be shown by the example $\boldsymbol{B} = (y, x, 1)$, where every field line possesses separatrices in this sense, i.e. has an X-type magnetic field in the plane perpendicular to the field line.) These difficulties become even more serious for fully three-dimensional magnetic fields without translational invariance. Several methods have been proposed to solve these difficulties of localizing and defining reconnection.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

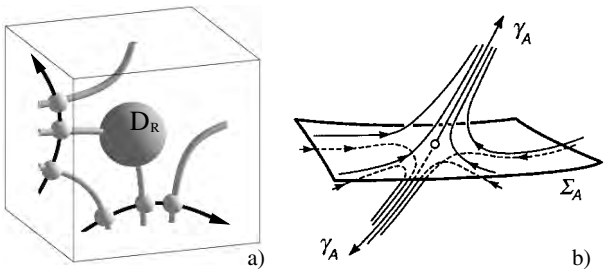


Figure 5. (*a*) Sketch of a breakdown of magnetic line conservation at a localized non-ideal region D_R. Two plasma elements (small spheres) which originally share a field line end up on different field lines. (*b*) Topology of field lines in the vicinity of an A-type generic null, showing the spine (γ_A) and the fan (Σ_A); B-type nulls have reversed field directions (from Lau and Finn 1990).

First, it is tempting to use the plasma flow in addition to the structure of the magnetic field to identify reconnection. However, this quantity is not independent of the frame of reference used, and for instance the location of the stagnation point depends on the observer. In another approach Hesse and Schindler therefore used the original meaning of reconnection, i.e. a breakdown of magnetic field line conservation (first suggested by Axford). They introduced the notion of general magnetic reconnection to occur if

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \neq 0 \tag{29}$$

where the integral is evaluated for field lines passing through a localized non-ideal region D_R embedded in an otherwise ideal plasma (see figure 5). The criterion (29) is sufficient for a breakdown of magnetic line conservation, provided all magnetic field lines start and end in the ideal region outside D_R. This is a consequence of the general form of magnetic field line conservation

$$\frac{\partial \boldsymbol{B}}{\partial t} - \nabla \times (\boldsymbol{w} \times \boldsymbol{B}) = \lambda \boldsymbol{B} \tag{30}$$

where $\boldsymbol{w}$ is the transport velocity of the field lines, which can be identified with the plasma velocity $\boldsymbol{v}$ in the ideal region but may differ from it in non-ideal processes. Equation (30) implies $\boldsymbol{B} \cdot \nabla\lambda = 0$, and therefore λ is constant on magnetic field lines. Moreover, in the ideal region we have $\boldsymbol{w} = \boldsymbol{v}$ and $\lambda = 0$ and hence $\lambda = 0$ across D_R as well. In this case equation (30) together with the induction equation implies

$$\boldsymbol{E} + \boldsymbol{w} \times \boldsymbol{B} = \nabla\Psi \tag{31}$$

and therefore $\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = 0$ along all magnetic field lines, because $\nabla\Psi$ vanishes in the ideal region so that Ψ is constant outside D_R. A non-vanishing integral (equation (29)) therefore requires a breakdown of magnetic line conservation. Vice versa, if $\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = 0$ holds for all magnetic field lines crossing D_R, the potential Ψ can be integrated within D_R from

$$\boldsymbol{E} \cdot \boldsymbol{e}_B = \nabla\Psi \cdot \boldsymbol{e}_B \tag{32}$$

and the field line velocity, given by

$$\boldsymbol{w} = (\tilde{\boldsymbol{E}} \times \boldsymbol{B})/\boldsymbol{B}^2 \tag{33}$$

with $\tilde{\boldsymbol{E}} = \boldsymbol{E} - \nabla\Psi$, exists provided there is no magnetic null within D_R. In this case (29) is also necessary for a breakdown of magnetic line conservation.

Magnetic null points

The existence of $\boldsymbol{w}$ given by (33) is critical if there are magnetic nulls within D_R. Using $\boldsymbol{E} = -\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}$ for a given evolution of an electromagnetic field, (32) can be restated by the existence of a potential $\tilde{\varphi} = \varphi + \Psi$ with

$$\boldsymbol{B} \cdot \nabla\tilde{\varphi} = -\boldsymbol{B} \cdot \frac{\partial \boldsymbol{A}}{\partial t} \tag{34}$$

where $\boldsymbol{A}$ is a vector potential for $\boldsymbol{B}$. Given the potential $\tilde{\varphi}$ on a surface crossed only by non-recurring field lines this condition defines $\tilde{\varphi}$ along these field lines. This method, called potential mapping, does not necessarily lead to a smooth potential $\tilde{\varphi}$ if field lines from separated regions join at magnetic nulls. For instance, smooth boundary conditions on $\tilde{\varphi}$ given for all field lines entering a surface enclosing the null, may lead to discontinuities of $\tilde{\varphi}$, and if the boundary is part of the ideal region the condition on $\tilde{\varphi}$ corresponds to boundary conditions on the plasma velocity $\boldsymbol{v}$. Therefore, Greene, followed by Lau and Finn, argued that in an almost-ideal plasma magnetic nulls are the site where non-ideal terms, especially the resistive term in Ohm's law, become important, and hence a breakdown of magnetic field line conservation may take place.

Magnetic nulls can be classified in terms of the eigenvalues of the tensor $\nabla\boldsymbol{B}$. They have either one real and two complex conjugated eigenvalues or three real eigenvalues. For the latter case they are called type A for $(+\,-\,-)$ signs of the eigenvalues and type B for $(-\,+\,+)$ (see figure 5). The eigenvectors of the complex conjugated eigenvalues, or of the real eigenvalues with the same sign, span a magnetic surface called the *fan* surface by Priest and Titov. The third eigenvector defines the spine as shown in figure 5. In the the case of more than one magnetic null the fan surfaces of an A-type and B-type null intersect at a structurally stable magnetic field line called separator. It can be shown that this field line is also a potential site of reconnection due to discontinuities in $\tilde{\varphi}$ or singularities of $\boldsymbol{w}$ for corresponding boundary conditions.

The topological structure of magnetic nulls led Priest and Titov to propose two additional mechanisms of reconnection called spine and fan reconnection. They showed that certain prescribed motions of the field lines on a surface enclosing the null produce singular field line velocities according to equation (33), and hence require a breakdown of field line conservation. In spine and fan reconnection the current tends to concentrate along the the spine and fan respectively.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reconnection without nulls

Magnetic nulls are not the only places where magnetic reconnection may occur. For instance if equation (34) is integrated over a closed field line with a non-vanishing contribution of the right-hand side, one finds that the potential $\tilde{\varphi}$ may not exist and therefore processes breaking the magnetic line conservation have to be present. This is also reflected by the criterion (29) which does not require nulls. For a non-vanishing magnetic field the method of potential mapping always leads to a smooth potential $\tilde{\varphi}$ and transport velocity $\boldsymbol{w}$. However, the latter might be very large, much higher than the Alfvén velocity, which excludes under realistic conditions an ideal evolution. This may happen, as noted by Priest, Forbes and Demoulin, in layer-like regions where the potential mapping or mapping of foot points of field lines shows strong gradients and which are therefore called magnetic flipping layers or quasi-separatrix layers.

While the method of potential or field line mapping aims at finding potential sites of reconnection and thus adds to the general criterion (29) certain conditions on the structure of the magnetic field, Hornig gave a more restricted definition of reconnection by generalizing the observation that in two dimension the field line velocity $\boldsymbol{w}$ has a singularity at the X-point. A covariant description shows that this singularity is a special type of null of the corresponding four-vector field W^4. This property is structurally stable in the transition from two to three dimensions, where now the site of reconnection is determined by a line of finite length along which W^4 vanishes. Within this definition it is in particular possible to distinguish a simple local slippage of plasma relative to the field lines, which also may satisfy (29) but which is not usually called reconnection, from reconnection itself.

Another aspect of reconnection is the dynamics of MAGNETIC HELICITY. While in two dimensions the source of magnetic helicity ($-2\boldsymbol{E}\cdot\boldsymbol{B}$) vanishes, this is not necessarily the case in three dimensions. Hence magnetic reconnection in three dimensions does not necessarily conserve magnetic helicity.

Collisionless reconnection

Magnetic reconnection can also occur in the absence of a collisional resistivity. Collisionless reconnection processes are based on non-ideal terms that in a more refined macroscopic picture appear on the right-hand side of Ohm's law (3) in addition to the resistive term. For instance, a current-driven microinstability may lead to fluctuations that on the macroscopic level have an effect similar to resistivity based on particle collisions. Also, resonant wave–particle interaction, off-diagonal terms of the electron pressure tensor or electron inertia have been suggested for magnetic reconnection. The final assessment of the role that each of these processes plays in reconnection requires a full three-dimensional kinetic description. Although such a kinetic point of view is crucial for the understanding of the small-scale plasma physics of reconnection, it less crucial for the overall dynamics, which in many of its features seems to be largely independent of the type and details of the non-ideal process. Thus, even highly collisionless reconnection processes, as for instance occurring in the Earth's magnetosphere, have been successfully simulated by using a simple resistive model of the form (3). Often, the resistivity is empirically adapted, for instance by spatial localization or by introducing an *ad hoc* dependence of η on the electric current density. The formation of thin current sheets in the pre-reconnection dynamics seems to play an important role in the onset of collisionless reconnection processes.

Bibliography

Biskamp D 1993 *Nonlinear Magnetohydrodynamics* (Cambridge: Cambridge University Press)

Parker E N 1979 *Cosmical Magnetic Fields* (Oxford: Clarendon Press)

Priest E R and Forbes T G 1999 *Magnetic Reconnection* (Cambridge: Cambridge University Press)

Tsinganos K C (ed) 1996 *Solar and Astrophysical Magnetohydrodynamic Flows* (Dordrecht: Kluwer)

Vasyliunas V M 1975 Theoretical models of magnetic field line merging *Rev. Geophys. Space Phys.* **13** 303

Karl Schindler and Gunnar Hornig

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamic Instabilities

Magnetohydrodynamic (MHD) stability theory determines whether a given PLASMA equilibrium configuration will either remain unchanged or evolve dynamically, in response to a general initial disturbance. If there is an initial disturbance whose amplitude subsequently grows in time, then the plasma is unstable, otherwise the plasma is stable. In addition, if the initial destabilizing disturbance has a small amplitude (normally referred to as infinitesimally small), then the plasma is linearly unstable and linear theory is appropriate. However, if the initial amplitude has a finite magnitude, then it is nonlinearly unstable. In many physical applications the plasma switches from a stable configuration to an unstable one as some parameter or property of the equilibrium exceeds a critical value. When the critical value is reached, the plasma is at marginal stability.

Many of the various solar phenomena can be understood by considering the typical timescales of the various physical processes in the solar CORONA. This is shown in table 1 and discussed in the subsequent sections.

Table 1. Typical timescales, in seconds, for the physical processes in the solar corona assuming $B = 100$ G, $l = 5 \times 10^7$ m, $\rho = 8.36 \times 10^{-13}$ kg m^{-3} and $T = 2 \times 10^6$ K. R is the gas constant.

Physical process	Symbol	Definition	Timescale (s)
Alfvén	τ_A	$l\sqrt{\mu\rho}/B$	5
Sound	τ_s	$l/\sqrt{RT}$	200
Conduction	$\tau_{\rm cond}$	$pl^2/\kappa_\parallel T$	600
Radiation	$\tau_{\rm rad}$	$p/\rho^2 Q(T)$	3000
Diffusion	τ_d	l^2/η	2.5×10^{15}

In this article a brief description of the equilibrium state is presented, followed by an introduction to the various methods used to assess the stability properties of the plasma. Next we describe ideal and non-ideal MHD instabilities before discussing some particular effects of coronal magnetic fields that penetrate the dense SOLAR PHOTOSPHERE.

MHD equilibrium

The basic MHD equilibrium equations are

$$\nabla p = \boldsymbol{j} \times \boldsymbol{B} + \rho \boldsymbol{g} \tag{1}$$

for magnetic equilibrium. A non-uniform magnetic field, i.e. $\boldsymbol{j} \neq 0$, can create current driven instabilities such as the MHD kink instability. An adverse density profile and the gravity term can drive Rayleigh–Taylor modes and the pressure gradient can drive localized modes. In addition to the force balance equation (1), the equilibrium magnetic field and plasma must satisfy

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \boldsymbol{j} = \frac{1}{\mu} \nabla \times \boldsymbol{B} \tag{2}$$

and an energy equation. For most situations this is taken as the adiabatic energy equation with

$$p = \text{const}\ \rho^\gamma .$$

For thermal instabilities a more appropriate energy equation, instead of the adiabatic equation, must be used, including thermal conduction, optically thin radiation ($\rho^2 Q(T)$) and some form of coronal heating. Hence, thermal equilibrium is described by

$$\nabla \cdot (\kappa \nabla T) = L(p, T) \equiv \rho^2 Q(T) - H. \tag{3}$$

κ is the thermal conductivity tensor, but in a magnetized plasma this is dominated by the component parallel to the magnetic field, $\kappa_\parallel$. It has a strong temperature dependence and is expressed as $\kappa_\parallel = 10^{-11} T^{5/2}$ W m^{-1} K^{-1}. L is the total loss function, including the coronal heating term, H, and $Q(T)$ is normally approximated by a piecewise continuous function of the form

$$Q(T) = \chi T^\alpha$$

where χ and α are constants dependent on the temperature. These are given in table 2. Finally, the temperature is coupled to the pressure and density through the gas law

$$p = \rho \mathcal{R} T$$

where $\mathcal{R}$ is the gas constant.

Table 2. The constants used in the radiative loss function $Q(T) = \chi T^\alpha$ W m^{-3} kg^{-2}. (From Cook *et al* (1989).)

T	χ	α
$T > 10^{6.5}$	7.15×10^{17}	0
$10^6 < T < 10^{6.5}$	1.25×10^{34}	−2.5
$10^5 < T < 10^6$	1.25×10^{19}	0
$10^4 < T < 10^5$	3.93×10^{13}	1.1

Stability methods

The energy method

The linear stability of a given equilibrium can be tested using three different methods. If it is important to know whether a plasma is either stable or unstable, an energy method (Bernstein *et al* 1954) can be used. The sign of the perturbed potential energy, δW, integrated over the volume of the plasma determines stability. δW is defined by

$$-\frac{1}{2} \int \xi \cdot \boldsymbol{F}(\xi)\, \mathrm{d}V \tag{4}$$

and

$$\begin{aligned} \delta W > 0 \text{ for all } \xi &\Rightarrow \text{stability} \\ \delta W < 0 \text{ for one allowable } \xi &\Rightarrow \text{instability.} \end{aligned} \tag{5}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$\boldsymbol{F}(\xi)$ is the linearized force given by

$$\begin{aligned}\boldsymbol{F}(\xi) = &\nabla(\xi \cdot \nabla p + \gamma p \nabla \cdot \xi) \\ &+ (\nabla \times \nabla \times (\xi \times \boldsymbol{B})) \times \boldsymbol{B}/\mu \\ &+ (\nabla \times \boldsymbol{B}) \times \nabla \times (\xi \times \boldsymbol{B})/\mu \\ &- \nabla \cdot (\rho\xi)\boldsymbol{g}. \end{aligned} \tag{6}$$

The displacement ξ must be continuous and satisfy the boundary conditions. The advantage of this method is that it is possible to prove instability by choosing a trial displacement that makes δW negative. This gives a necessary condition for stability. In addition, it is possible to rearrange the energy integral into two terms, one a positive definite term and the other a simpler integral. Minimizing this second integral and showing that it is positive gives a sufficient condition for stability. Thus, simple bounds on the stability threshold can be obtained.

The normal mode method

Assuming a time dependence of the form $e^{-i\omega t}$, the linearized MHD equations reduced to an eigenvalue problem of the form

$$-\omega^2 \xi = \boldsymbol{F}(\xi),$$

for the eigenvalue ω^2. If all values of ω^2 are positive then the plasma is stable but if there exists one $\omega^2 = -\sigma^2$ that is negative, then the plasma is unstable. Here ω is the frequency of oscillation of a stable wave and σ is the growth rate of an unstable mode. The spectrum of the system is complicated by the existence of a continuous spectrum for ω^2 as well as the discrete spectrum of modes.

The initial value method

A final approach is to treat the equations as an initial value problem and time advance the equations numerically. If the initial disturbance is sufficiently general, then the most unstable mode will grow and eventually dominate the solution. This approach is frequently used as a forerunner to a full nonlinear simulation. Whichever approach is used, it is important that a complete description of the linear stability properties of the equilibrium is obtained before detailed (and time-consuming) nonlinear calculations are carried out.

Rayleigh–Taylor instability

The Rayleigh–Taylor instability occurs when gravity interacts with an adverse density gradient. Imagine two plasmas of uniform densities ρ_1 in $z > 0$ (region 1) and ρ_2 in $z < 0$ (region 2), separated by an interface at $z = 0$. Using a normal mode analysis with disturbances of the form $f(x, z, t) = f(z)\, e^{i(\omega t - kx)}$, matching the two solutions in the separate regions, $z > 0$ and $z < 0$, and using continuity of both the normal velocity and the perturbed total pressure at the perturbed interface, the incompressible dispersion relation is obtained as

$$\omega^2 = kg\frac{\rho_2 - \rho_1}{\rho_2 + \rho_1}. \tag{7}$$

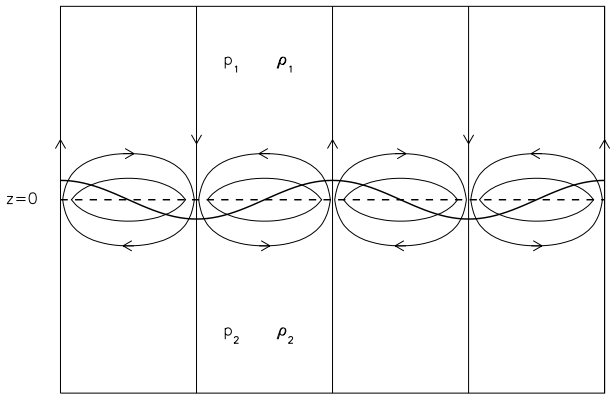


Figure 1. The streamlines for the Rayleigh–Taylor instability are shown. The unperturbed interface at $z = 0$ is shown as a thick dashed line and the perturbed interface as the thick full curve

Notice that the plasma is unstable when $\rho_1 > \rho_2$, i.e. heavy plasma overlies light, and that the fastest growing instability occurs for $k \to \infty$. Thus, the wavelengths are extremely small when the plasma is unstable. The form of the instability is shown in figure 1.

The inclusion of a uniform magnetic field at an angle θ to the x-direction modifies this result giving the dispersion relation as

$$\omega^2 = kg\frac{\rho_2 - \rho_1}{\rho_2 + \rho_1} + \frac{2}{\mu}\frac{(kB\cos\theta)^2}{\rho_2 + \rho_1}. \tag{8}$$

The second term on the right-hand side represents the effect of magnetic tension that is introduced when the perturbation bends the fieldlines. When $\theta = 0$, then the magnetic field stabilizes the Rayleigh–Taylor mode if the wavenumber satisfies

$$k > \frac{g\mu}{2B^2}(\rho_1 - \rho_2). \tag{9}$$

Note that if $\theta = \pi/2$ so that

$$\boldsymbol{k} \cdot \boldsymbol{B} = kB\cos\theta = 0$$

the magnetic field does not influence the Rayleigh–Taylor mode at all. This condition appears in many stability criteria. Further information about gravitational instabilities is given in the article on MAGNETIC BUOYANCY.

Rayleigh–Taylor instabilities are expected whenever dense plasma overlies light plasma. A classic example of this type of situation occurs at the lower edge of a SOLAR PROMINENCE. Here the dense plasma is supported against gravity by the Lorentz force. There is normally a fairly clear lower edge to prominences so the Rayleigh–Taylor instability must be inhibited possibly by either magnetic shear or photospheric line-tying effects.

Magnetic instabilities

Magnetic instabilities are best described by the example of cylindrically symmetric magnetic fields. In this case, (1)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

reduces to

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(p + \frac{B^2}{2\mu}\right) = -\frac{B_\theta^2}{\mu r} \tag{10}$$

where the azimuthal and axial components of the magnetic field are $B_\theta(r)$ and $B_z(r)$ respectively and the gas pressure is $p(r)$. The left-hand side is the total pressure gradient and the right-hand side is the magnetic tension. Since the equilibrium only depends on the radial coordinate, all perturbations can be Fourier analyzed in the form

$$f(\mathbf{r}, t) = f(r)\,\mathrm{e}^{\mathrm{i}(m\theta + kz - \omega t)} \tag{11}$$

and the linearized equations reduce to a set of ordinary differential equations.

Low-m, global modes

The nature of the mode is defined by the value of the azimuthal wavenumber m. $m = 0$ is the sausage mode and $m = 1$ is the kink mode. The form of these perturbations is shown in figure 2. In both cases the regions of strong magnetic field force the plasma towards the regions of weak magnetic field and so reinforce the plasma displacement. However, it is important to note that the kink mode instability is the only mode that actually displaces the magnetic axis.

The exponential term has constant amplitude along the helix defined by $m\theta + kz =$ constant. The pitch of the disturbance matches the pitch of the equilibrium magnetic field when

$$\boldsymbol{k} \cdot \boldsymbol{B} = \frac{m}{r} B_\theta + k B_z = 0. \tag{12}$$

This condition is satisfied at a particular radius, r_s, the radius of a mode rational surface, and corresponds to a singularity in the ideal MHD, marginal stability equations when $\omega^2 = 0$.

Inertial line-tying

The stability of solar coronal magnetic fields is strongly influenced by the effect of the very dense photosphere. Imagine an Alfvén wave propagating down from the corona. When it reaches the photospheric level there is a large increase in density of a factor of 10^8 and the wave is almost totally reflected. In a similar manner, a slow MHD WAVE is also reflected. The photosphere provides a strong stabilizing effect, called inertial line-tying, and the simplest way to model this is to assume that all components of the displacement (or equivalently the velocity) are zero at the photospheric ends of the fieldlines. This means that the wavenumber along the magnetic fieldline is not only discretized but, because the components of the displacement are out of phase with each other, also couples Fourier components together. An individual Fourier mode does not satisfy the boundary conditions and so many modes must be included.

In periodic configurations the first modes to become unstable are incompressible modes. This can be understood, since if some of the available energy is used in compressing the plasma then there is less to drive the instabilities. Thus, contracting the plasma as in figure 2(*a*) means that the plasma will flow towards the expanding regions, otherwise the plasma will be compressed. However, the line-tying conditions prevent this flow and so the most unstable line-tied plasma disturbances are usually compressible.

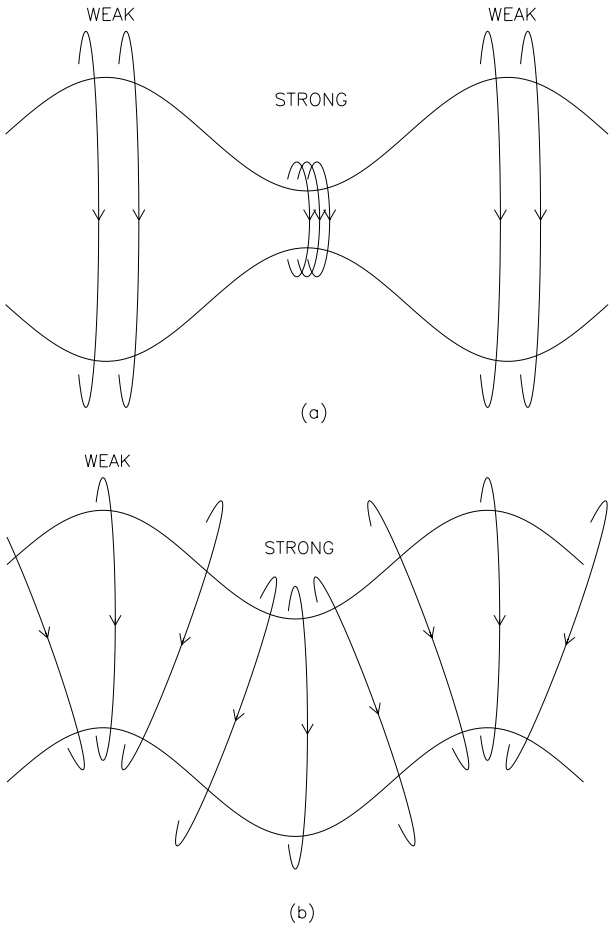


Figure 2. The displacements for (*a*) the sausage mode and (*b*) the kink mode.

High-m, localized modes

High m values for the azimuthal wavenumber produce modes that are strongly localized about a particular radius. These modes are driven by an adverse pressure gradient and are unstable, in the absence of line-tying, if Suydam's criterion holds, namely

$$\frac{B_z^2}{4}\left(\frac{q'}{q}\right)^2 + \frac{2\mu p'}{r} < 0 \tag{13}$$

where $q = r B_z / B_\theta$ gives a measure of the shear in the field. Line-tying modifies this result by adding two stabilizing effects due to field line tension and compression. The condition for instability is now

$$a_1 \frac{B_z^2}{4}\left(\frac{q'}{q}\right)^2 + a_2 \frac{2\mu p'}{r} + a_3 \frac{B^2}{r^2} + a_4 \frac{\gamma \mu p B_\theta^2}{\gamma \mu p + B^2} < 0 \tag{14}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where the constants, a_i, are defined in De Bruyne and Hood (1989).

Resistive instabilities

In ideal MHD the magnetic fieldlines are frozen into the plasma and hence retain their topological identity. However, once resistivity is included, fieldlines can diffuse through the plasma, allowing them to break and reconnect. Thus, new topological configurations can be accessed that were prohibited under ideal MHD. In cases where the new configurations have a lower magnetic energy, there are resistive instabilities that are driven by this available free magnetic energy. These new instabilities grow at a rate that is slower than the rates for ideal modes but faster than pure diffusion. In general, they have a growth rate that is related to a hybrid of the Alfvén and diffusion timescales, τ_A and τ_d as defined in table 1. The main resistive mode in the solar corona is the tearing mode with a growth rate

$$\sigma \propto \tau_A^{-2/5} \tau_d^{-3/5}. \tag{15}$$

The resistive modes are localized about the mode rational surfaces that occur in ideal MHD, marginal stability ($\omega^2 = 0$) when

$$\boldsymbol{k} \cdot \boldsymbol{B} = \frac{m}{r} B_\theta + k B_z = 0.$$

The nonlinear development of the tearing mode can lead to steady MAGNETIC RECONNECTION.

The sudden eruption of a solar FLARE suggests that the initial instability is due to an ideal MHD instability. However, the release of magnetic energy requires resistivity. Driving the field lines together in response to an ideal MHD instability means that reconnection and energy release can occur on a much faster timescale than the tearing mode timescale.

Thermal instabilities

The form of the optically thin plasma radiation in the solar corona and other astrophysical plasmas can drive thermal instabilities, also called radiative modes and condensation modes. The shape of an analytic fit to the optically thin radiation is shown in figure 3 as a function of the temperature at both constant density and constant pressure. Note that for the case of constant pressure the radiation decreases for temperatures in the range 10^4 K to 10^7 K. Thus, if the temperature in the corona decreases due to a perturbation, then the radiation will increase causing the plasma to cool further (Parker 1957). This thermal runaway will continue until the temperature drops below 10^4 K, which is incidentally the typical temperature of prominences. Thermal instabilities can be understood by examining the linearized energy equation. Two distinct forms of the thermal instability arise depending on the ratio of the sound travel time, τ_s, to the radiative timescale, $\tau_{\rm rad}$. If $\tau_s \ll \tau_{\rm rad}$, the thermal instability evolves isobarically. For a uniform equilibrium and perturbations of the form $f(x,t) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r})+\sigma t}$, the dispersion relation for the isobaric thermal mode is

$$\sigma = \frac{\gamma}{\gamma - 1}\left(-\frac{(\boldsymbol{k}\cdot\boldsymbol{B})^2}{B^2}\frac{l^2}{\tau_{\rm cond}} + \frac{(2-\alpha)}{\tau_{\rm rad}}\right) \tag{16}$$

where the timescales are defined in table 1 and α is given in table 2. The isochoric limit, when $\tau_s \gg \tau_{\rm rad}$, gives

$$\sigma = (\gamma - 1)\left(-\frac{(\boldsymbol{k}\cdot\boldsymbol{B})^2}{B^2}\frac{l^2}{\tau_{\rm cond}} - \frac{\alpha}{\tau_{\rm rad}}\right). \tag{17}$$

Physically, these modes can be understood as follows. In the isobaric limit any thermal disturbances evolve much more slowly than the propagation of pressure variations. So sound waves can propagate to equalize the pressure changes before the temperature can evolve significantly. Hence, the modes evolve at constant pressure. This is the case at coronal temperatures. In the isochoric limit the thermal instability evolves so quickly that there is no time for sound waves to propagate away from the instability region. Thus, flows cannot be set up to equalize the pressure and so the density cannot change. Thus, the instability evolves at constant density.

Thermal instabilities are one mechanism proposed for the formation of cool, dense solar prominences. The hot corona begins to condense and a siphon flow sucks plasma from the CHROMOSPHERE into the corona in order to equalize the pressure.

Onset of convection

A plasma heated from below becomes unstable to convection when the temperature gradient exceeds a critical value. Neglecting all the dissipation mechanisms initially and the magnetic field, we can analyze the linearized MHD equations for an equilibrium pressure and density satisfying hydrostatic balance, the ideal gas law and a prescribed temperature profile $T_0(z)$, where z is the vertical coordinate. The local criterion for the onset of convection occurs when the temperature gradient exceeds the adiabatic value. Thus, instability sets in when

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{p}{\rho^\gamma}\right) < 0 \tag{18}$$

so that entropy decreases with height. This may be expressed as

$$\frac{\mathrm{d}T}{\mathrm{d}z} < -\frac{\gamma - 1}{\gamma}\frac{g}{\mathcal{R}T}. \tag{19}$$

Thus, if the temperature decreases at a faster rate, then convective motions set in. (See the article on MAGNETIC BUOYANCY for a description of the instability and the inclusion of a magnetic field.)

The inclusion of dissipation, through viscosity and thermal conduction, severly complicates the stability analysis and analytical progress can be made by using the Boussinesq approximation. Here sound waves are filtered out and density variations are only included in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

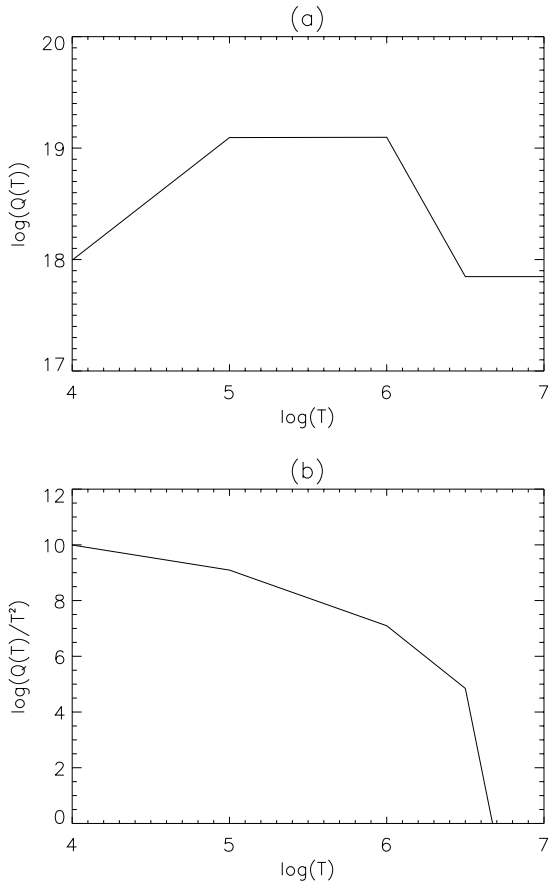


Figure 3. The optically thin radiation is approximated as a piecewise continuous function of temperature. The figures show the behavior of radiation against temperature at (*a*) constant density and (*b*) constant pressure. Note the logarithmic scales.

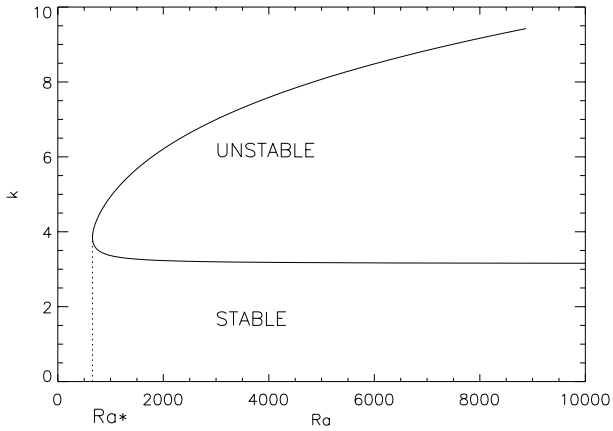


Figure 4. The marginal stability curve where k is the total wavenumber and Ra is the Rayleigh number.

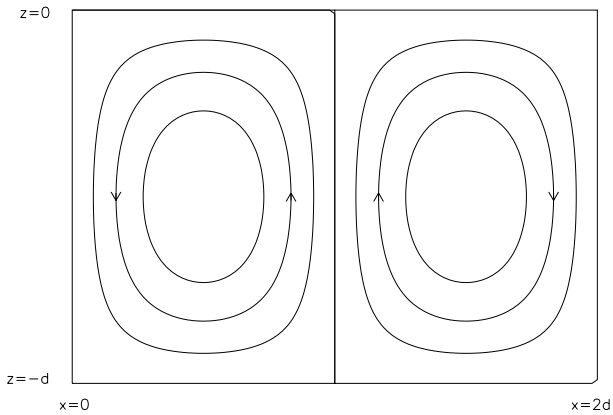


Figure 5. The streamlines for the convective motions that arises when the temperature gradient is large enough.

the gravitational term, as $\rho = \rho_0(1 + \alpha T_1)$, where T_1 is the perturbed temperature and α is the volume expansion coefficient. For the vertical velocity perturbation of the form

$$v_z(x, y, z, t) = V_z \, \mathrm{e}^{\sigma t} \, \mathrm{e}^{\mathrm{i}(k_x x + k_y y)} \sin \frac{\pi}{d} z$$

where d is the depth of the layer, stability depends on the value of the Rayleigh number

$$Ra = \frac{g\alpha \Delta T}{\kappa \nu}. \tag{20}$$

κ is the thermal conductivity, ν is the kinematic viscosity and ΔT is the temperature difference between the top and bottom of the layer. Setting $\sigma = 0$ in the dispersion relation gives the relation between Ra and the wavenumber k, defined by $k^2 = k_x^2 + k_y^2 + \pi^2/d^2$, for marginal stability. Instability occurs if Ra exceeds the critical value $Ra^* = 27\pi^4/4$. The marginal stability curve is shown in figure 4 and the streamlines, showing the cellular pattern of the plasma flow, are illustrated in figure 5.

White light images of the photosphere show granulation patterns giving clear evidence of convection. Larger scales of convection also occur, giving rise to supergranules and giant cells.

Kelvin–Helmholtz instability

The Kelvin–Helmholtz instability arises when there is a horizontal shear flow, $U(z)$, and the plasma density is stratified due to gravity. The classical Kelvin–Helmholtz instability is best illustrated for two plasmas of uniform density and uniform horizontal flows separated by an interface. The incompressible, linear equations are expressed in terms of the vertical velocity and solved. Assuming perturbations of the form

$$v_z = V_\pm \, \mathrm{e}^{\mathrm{i}(kx - \omega t) \mp kz}$$

where in region 1, $z > 0$, the upper sign is used and in region 2, $z < 0$, the lower sign is used. The two regions are matched at the perturbed interface using continuity of pressure and continuity of the normal displacement, ξ_z, where

$$\xi_z = \frac{v_z}{\omega - kU}.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The dispersion relation is

$$\rho_1(\omega - kU_1)^2 + \rho_2(\omega - kU_2)^2 = kg(\rho_2 - \rho_1) \qquad (21)$$

and the plasma is unstable if

$$(U_2 - U_1)^2 > \frac{g}{k}\frac{(\rho_2^2 - \rho_1^2)}{\rho_2\rho_1}. \qquad (22)$$

Thus, if k is sufficiently large, there is always a Kelvin–Helmholtz instability for any difference in the background flow speeds. Obviously, viscosity will stabilize these short-wavelength instabilities.

Kelvin–Helmholtz instabilities are likely to occur between the fast and slow plasma flows in the SOLAR WIND. In addition, the field lines separating the open field regions of CORONAL HOLES and the closed field regions of ACTIVE REGIONS will also experience these flow induced instabilities.

Bibliography

Bernstein I, Frieman E, Kruskal M and Kusrud R 1954 *Proc. R. Soc.* A **244** 17

Cook J W, Cheng C-C, Jacobs V L and Antiochos S K 1989 *Astrophys. J.* **338** 1176–83

De Bruyne P and Hood A W 1989 Simple tests for the ideal MHD stability of line-tied coronal magnetic fields *Solar Phys.* **119** 87–106

Parker E N 1957 *Astrophys. J.* **117** 431

Alan Hood

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamic Waves

Magnetohydrodynamic waves are found in a wide variety of astrophysical plasmas. They have been measured in plasma fusion devices and detected in the MAGNETOSPHERE OF EARTH, the SOLAR WIND and a number of magnetic structures seen in the Sun's atmosphere. In space plasmas their detection is often indirect, by matching measured properties (such as propagation speed or pressure variation) with theoretically deduced properties and correlations, although the magnetosphere and solar wind provide direct *in situ* measurements. A complication in the identification of magnetohydrodynamic waves in space plasmas has been the realization that generally such plasmas are strongly inhomogeneous, structured by magnetism, plasma density and temperature variations, or by plasma flows. This strong inhomogeneity brings about complications in the description of magnetohydrodynamic wave phenomena, with magnetic flux tubes and plasma flow tubes being of special interest. In the solar atmosphere, SUNSPOTS, SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES, CORONAL LOOPS and POLAR PLUMES are examples of flux tube structure, and oscillatory phenomena have been detected in all of them. Oscillations have also been detected in SOLAR PROMINENCES and generated by SOLAR FLARES.

Waves are the way in which information in a medium may be communicated from one region to another. In a gas communication can be achieved by sound waves, in the form of compressions and rarefactions of the medium which propagate at the sound speed c_s. For an ideal gas with equilibrium pressure p_0 and density ρ_0 the sound speed in a medium is given by

$$c_s = \left(\frac{\gamma p_0}{\rho_0}\right)^{1/2} \tag{1}$$

where γ is the ratio of specific heats. Magnetism introduces an additional speed, determined by the strength B_0 of the magnetic field and the plasma density ρ_0. This so-called Alfvén speed v_A is defined by

$$v_A = \left(\frac{B_0^2}{\mu\rho_0}\right)^{1/2} \tag{2}$$

where μ is the magnetic permeability of the medium. These two speeds determine the nature of wave propagation in magnetohydrodynamics.

The speeds c_s and v_A that characterize sound and magnetic effects are conveniently supplemented by the introduction of two further speeds, a slow speed c_t and a fast speed c_f, defined by

$$c_t^{-2} = c_s^{-2} + v_A^{-2} \qquad c_f^2 = c_s^2 + v_A^2. \tag{3}$$

The slow speed is evidently both sub-sonic and sub-Alfvénic; the fast speed c_f is super-sonic and super-Alfvénic. The slow speed c_t proves to be of particular significance in magnetic flux tubes and is accordingly sometimes referred to as the tube speed, although it is not the only speed of special significance in flux tubes (see below). The four speeds c_t, c_s, v_A and c_f arise in various ways in the description of magnetohydrodynamic wave phenomena, and their magnitudes may vary dramatically from application to application. For example, in a 2×10^6 K solar corona with a magnetic field strength of 10^2 G the sound speed is about 200 km s^{-1} and the Alfvén speed about 10^3 km s^{-1}, whereas in a solar photospheric flux tube both these speeds fall to about 10 km s^{-1}. The corresponding slow and fast speeds in the solar coronal illustration are 196 km s^{-1} and 1020 km s^{-1}, whereas in the photospheric flux tube these speeds are 7 km s^{-1} and 14 km s^{-1}, respectively.

The distinctive feature of a magnetohydrodynamic plasma is that, in addition to compressive disturbances, the magnetoacoustic waves with their affinity to sound waves, there arises the possibility of a wave which propagates without compressing the plasma, driven purely by magnetic tension forces. This is the Alfvén wave, named in honour of the 1970 Nobel Laureate Hannes Alfvén (1908–1995) who first described the wave in 1942 and made fundamental contributions to the foundations of magnetohydrodynamics. Alfvén waves are closely analogous to the familiar transverse waves of an elastic string, the magnetic field providing the tension force that determines (together with the string or plasma density) the speed of propagation of the wave. The magnetic field in a plasma is also able to provide another force, akin to the plasma pressure; this is the magnetic pressure, ${B_0}^2/2\mu$ N m^{-2}. Overall the field provides magnetic pressure forces that resist any tendency to compress the plasma (acting in addition to the plasma pressure) and magnetic tension forces which act to straighten any distortions in the magnetic field. The tension force gives rise to the Alfvén wave; the magnetic pressure force gives rise to slow and fast magnetoacoustic waves. Broadly, the fast magnetoacoustic wave results when the magnetic pressure force acts in unison with the plasma pressure force; the slow magnetoacoustic wave results when these two pressure forces act in opposition.

To describe magnetohydrodynamic waves in detail it is necessary to examine the equations of ideal MAGNETOHYDRODYNAMICS. We consider an inhomogeneous medium because it allows a description of waves in the important case of a magnetic flux tube but also includes, as a special case, the classical problem of waves in a uniform medium. Consider, then, a basic state in which there is an applied unidirectional magnetic field $\boldsymbol{B}_0 = B_0\hat{\boldsymbol{z}}$ aligned with the z-axis in an appropriate coordinate system. In equilibrium the plasma density is ρ_0, the temperature T_0, the pressure p_0 and the field strength B_0. These quantities may all vary in a direction perpendicular to the applied magnetic field; equilibrium is achieved by requiring that

$$\text{grad}\left(p_0 + \frac{B_0^2}{2\mu}\right) = 0 \tag{4}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

showing that any spatial variations in magnetic pressure are balanced by corresponding variations in plasma pressure. The case of a uniform medium corresponds to taking B_0, p_0, T_0 and ρ_0 all constant.

We consider isentropic (or adiabatic) disturbances governed by the equations of ideal magnetohydrodynamics. Denote by p and ρ the pressure and density perturbations of an ideal plasma with ratio of specific heats γ (usually taken to be 5/3). The perturbed magnetic (induction) field is $\boldsymbol{B}$ and the fluid motion is $\boldsymbol{v}$. To examine small-amplitude linear disturbances about this equilibrium, it is convenient to introduce the perturbation in total pressure, p_T:

$$p_T \equiv p + \frac{1}{\mu}\boldsymbol{B}_0\cdot\boldsymbol{B}. \tag{5}$$

Then, splitting the velocity field $\boldsymbol{v}$ into components perpendicular and parallel to the applied magnetic field, so that $\boldsymbol{v} = \boldsymbol{v}_\perp + v_z\hat{\boldsymbol{z}}$, the linearized equations of ideal magnetohydrodynamics yield

$$\rho_0\left(\frac{\partial^2}{\partial t^2} - v_{\rm A}^2\frac{\partial^2}{\partial z^2}\right)\boldsymbol{v}_\perp + \nabla_\perp\left(\frac{\partial p_{\rm T}}{\partial t}\right) = 0 \tag{6}$$

$$\rho_0\left(\frac{\partial^2}{\partial t^2} - c_{\rm t}^2\frac{\partial^2}{\partial z^2}\right)v_z + \frac{c_{\rm s}^2}{c_{\rm f}^2}\frac{\partial}{\partial z}\left(\frac{\partial p_{\rm T}}{\partial t}\right) = 0 \tag{7}$$

$$\frac{\partial p_{\rm T}}{\partial t} = \rho_0 v_{\rm A}^2\frac{\partial v_z}{\partial z} - \rho_0 c_{\rm f}^2\,\mathrm{div}\,\boldsymbol{v}. \tag{8}$$

This system of coupled partial differential equations determines the relationship between the flow $\boldsymbol{v}$ and the total pressure variation $p_{\rm T}$. The operator $\nabla_\perp$ denotes the component of the gradient operator that acts perpendicular to the applied magnetic field (i.e. perpendicular to the z-axis). Notice the occurrence of the four speeds $c_{\rm t}$, $c_{\rm s}$, $v_{\rm A}$ and $c_{\rm f}$, each a function of the coordinates perpendicular to the applied magnetic field.

It may be noted that equations (7) and (8) combine to give

$$\frac{\partial^2 v_z}{\partial t^2} = c_{\rm s}^2\frac{\partial}{\partial z}(\mathrm{div}\,\boldsymbol{v}). \tag{9}$$

The absence of magnetic effects in this equation is simply a reflection of the fact that magnetic forces act perpendicular to the magnetic field and so have no effect on motions along the field.

Uniform medium

In the case of a uniform medium, B_0, p_0, T_0 and ρ_0 are all constants and so are the basic speeds $c_{\rm t}$, $c_{\rm s}$, $v_{\rm A}$ and $c_{\rm f}$. This allows a relatively straightforward reduction of equations (6)–(8). It proves convenient to work in terms of the compressibility $\Delta \equiv \mathrm{div}\,\boldsymbol{v}$. Then Δ satisfies the fourth-order wave equation

$$\frac{\partial^4\Delta}{\partial t^4} - (c_{\rm s}^2 + v_{\rm A}^2)\frac{\partial^2}{\partial t^2}\nabla^2\Delta + c_{\rm s}^2 v_{\rm A}^2\frac{\partial^2}{\partial z^2}\nabla^2\Delta = 0 \tag{10}$$

involving the three-dimensional Laplacian ∇^2.

Since the medium is uniform and unbounded we may Fourier analyse equation (10) in a Cartesian coordinate system (with position vector $\boldsymbol{r} = (x, y, z)$), writing

$$\Delta(x, y, z, t) = \Delta_0 \exp\,\mathrm{i}(\omega t - \boldsymbol{k}\cdot\boldsymbol{r}) \tag{11}$$

where Δ_0 is the amplitude (a constant), ω the frequency and $\boldsymbol{k} = (k_x, k_y, k_z)$ the wavevector of the disturbance; other variables (e.g. flow $\boldsymbol{v}$ and field $\boldsymbol{B}$) are represented similarly. The Laplacian operator becomes

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = -k^2$$

where $k = (k_x^2 + k_y^2 + k_z^2)^{1/2}$ is the magnitude of the wavevector. Then, for $\Delta_0 \neq 0$, equation (10) yields

$$\omega^4 - (c_{\rm s}^2 + v_{\rm A}^2)k^2\omega^2 + c_{\rm s}^2 v_{\rm A}^2 k_z^2 k^2 = 0. \tag{12}$$

This is the dispersion relation for magnetoacoustic waves; it relates the wave frequency ω to the wavevector $\boldsymbol{k}$, thus describing the propagation.

The dispersion relation (12) may be solved to give the phase speed c ($\equiv \omega/k$). In terms of the angle ϕ that the wavevector $\boldsymbol{k}$ makes with the applied magnetic field $\boldsymbol{B}_0$, with $k_z = k\cos\phi$, we obtain

$$c^2 = \frac{1}{2}c_{\rm f}^2\left[1 \pm \left(1 - 4\frac{c_{\rm t}^2}{c_{\rm f}^2}\cos^2\phi\right)^{1/2}\right]. \tag{13}$$

The smaller root (corresponding to the $-$ sign) is the slow magnetoacoustic wave, and the larger root (corresponding to the + sign) is the fast magnetoacoustic wave. The waves are compressive (since $\Delta \neq 0$) and so involve density, pressure and temperature variations—as do sound waves. Unlike sound waves, magnetoacoustic waves exhibit anisotropic propagation, their phase speeds depending on the angle of propagation to the applied magnetic field; this is simply a reflection of the directionality that an applied magnetic field imposes on a plasma. When propagation is along the applied magnetic field $\boldsymbol{B}_0$, so that $\phi = 0$, then $c^2 = c_{\rm s}^2$ or $v_{\rm A}^2$. When propagation is perpendicular to the applied magnetic field, so that $\phi = \pi/2$, then $c^2 = 0$ or $c_{\rm f}^2$; in other words, the slow wave is unable to propagate across the magnetic field, whereas the fast wave propagates with the speed $c_{\rm f}$. This is the greatest speed that the fast wave can achieve; at any other angle its speed lies below $c_{\rm f}$ and above (or equal to) the greater of $c_{\rm s}$ and $v_{\rm A}$. The phase speed of the slow wave is always less than or equal to the smaller of $c_{\rm s}$ and $v_{\rm A}$.

Magnetoacoustic waves are compressive but disturbances that are incompressible ($\Delta = 0$)—producing no changes in plasma density of an element moving with the fluid—are also permitted and in fact provide a distinct form of wave motion. This is the Alfvén wave with dispersion relation

$$\omega^2 = k_z^2 v_{\rm A}^2 \tag{14}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

giving a phase speed c according to

$$c^2 = v_A^2 \cos^2 \phi. \tag{15}$$

Evidently, then, the phase speed of an Alfvén wave cannot exceed the Alfvén speed, v_A. Moreover, like the slow wave, the Alfvén wave is unable to propagate across the magnetic field, since $\omega^2 = 0$ when $\phi = \pi/2$. The Alfvén wave involves no motions along the magnetic field ($v_z = 0$) and no variations in plasma density ($\rho = 0$), pressure ($p = 0$), temperature or total pressure ($p_T = 0$). Thus the Alfvén wave is a transverse wave that propagates incompressibly, involving no changes in plasma density or pressure.

Altogether, then, we see that in the presence of an applied magnetic field there are three waves available to the medium, the Alfvén wave and the fast and slow magnetoacoustic waves. All three waves exhibit a degree of anisotropic propagation but this is most marked in the Alfvén wave and the slow wave, neither of which can propagate orthogonally to the applied magnetic field; the fast wave is only weakly anisotropic and achieves its greatest phase speed ($c = c_f$) when propagating perpendicularly to the applied magnetic field. The anisotropy exhibited by magnetohydrodynamic waves is in sharp contrast to the isotropic behavior of sound waves in an uniform medium, this simply being a reflection of the directionality imposed by a magnetic field.

The behavior of the magnetohydrodynamic wave speeds is conveniently represented in a diagram by plotting the phase speed c as a function of propagation angle ϕ, using a polar coordinate (c, ϕ) representation; see figure 1.

Related to the information provided in the phase speed plot is the behavior of the group velocity $\boldsymbol{c}_g$ of a wave, defined by

$$\boldsymbol{c}_g = \frac{\partial \omega}{\partial \boldsymbol{k}}. \tag{16}$$

The group velocity of a wave is the velocity with which energy (and information) is propagated by the wave. For the Alfvén wave equation (16) gives simply $\boldsymbol{c}_g = v_A \hat{\boldsymbol{z}}$, showing that an Alfvén wave propagates energy and information *along* the applied magnetic field, whatever the direction of the wave itself. The group velocity for magnetoacoustic waves is more complicated to obtain and is usually displayed in diagrammatic form, showing a plot of the *group speed* $c_g = |\boldsymbol{c}_g|$ as a function of propagation angle; see figure 2. Again, note the highly anistropic behavior of the slow wave which is confined to propagating energy close to the applied magnetic field, whereas the fast wave propagates energy approximately isotropically in all directions.

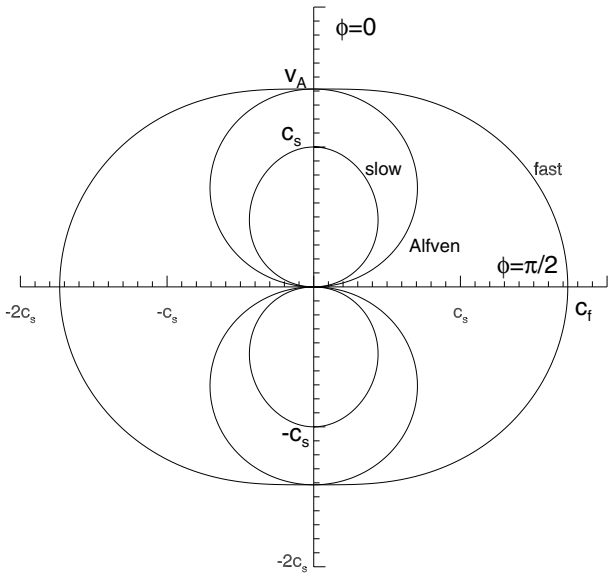


Figure 1. Phase speed diagram for magnetohydrodynamic waves, showing the phase speed $c(\phi)$ (in units of the sound speed c_s) as a function of the angle ϕ of propagation to the applied magnetic field, for the case $v_A > c_s$.

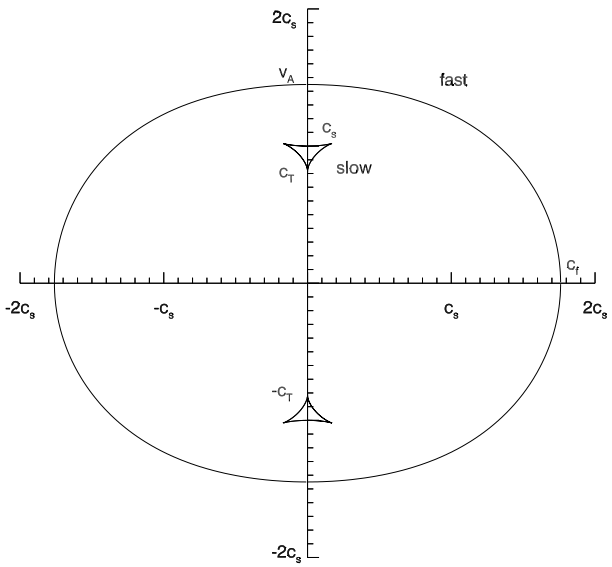


Figure 2. Group speed diagram for magnetoacoustic waves, showing the group speed $c_g(\phi)$ (in units of the sound speed c_s) as a function of the angle ϕ of propagation to the applied magnetic field, for the case $v_A > c_s$.

Magnetically structured media

The nonuniform media generally encountered in space plasmas may frequently be modeled in terms of a magnetically structured plasma. Structuring may be in the magnetic field, plasma density or temperature, or through the presence of plasma flows. Both Cartesian and cylindrical coordinate representations are possible. The cylindrical form is particularly interesting in view of the wide-spread occurrence of flux tube structures. Returning to equations (6)–(8), we express them in cylindrical coordinates (r, θ, z), writing $\boldsymbol{v} = (v_r, v_\theta, v_z)$. Equation (6)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

then yields

$$\rho_0(r)\left(\frac{\partial^2}{\partial t^2} - v_{\rm A}^2(r)\frac{\partial^2}{\partial z^2}\right)v_r + \frac{\partial^2 p_{\rm T}}{\partial r\,\partial t} = 0 \qquad (17)$$

$$\rho_0(r)\left(\frac{\partial^2}{\partial t^2} - v_{\rm A}^2(r)\frac{\partial^2}{\partial z^2}\right)v_\theta + \frac{1}{r}\frac{\partial^2 p_{\rm T}}{\partial\theta\,\partial t} = 0. \qquad (18)$$

These equations may be reduced to ordinary differential equations by writing

$$v_r(r,\theta,z,t) = v_r(r)\exp\,{\rm i}(\omega t + n\theta - k_z z) \qquad (19)$$

for mode number $n = 0, 1, 2, \ldots$. The resulting equations may be manipulated to yield the equation

$$\frac{\rm d}{{\rm d}r}\left\{\frac{\rho_0(r)[k_z^2 v_{\rm A}^2(r) - \omega^2]}{[m^2(r) + n^2/r^2]}\frac{1}{r}\frac{\rm d}{{\rm d}r}(rv_r)\right\} = \rho_0(r)(k_z^2 v_{\rm A}^2(r) - \omega^2)v_r \qquad (20)$$

for v_r, with m^2 given by

$$m^2(r) = \frac{(k_z^2 c_{\rm s}^2 - \omega^2)(k_z^2 v_{\rm A}^2 - \omega^2)}{(c_{\rm s}^2 + v_{\rm A}^2)(k_z^2 c_{\rm t}^2 - \omega^2)}.$$

Equation (20) is sometimes called the Hain-Lüst equation.

The differential equation satisfied by $p_{\rm T}$ is also of interest:

$$\rho_0(r)[k_z^2 v_{\rm A}^2(r) - \omega^2]\frac{1}{r}\frac{\rm d}{{\rm d}r}\left(\frac{1}{\rho_0(r)(k_z^2 v_{\rm A}^2(r) - \omega^2)}r\frac{{\rm d}p_{\rm T}}{{\rm d}r}\right) = \left(m^2(r) + \frac{n^2}{r^2}\right)p_{\rm T}. \qquad (21)$$

Equations (20) and (21) possess singularities at $\omega^2 = k_z^2 v_{\rm A}^2$ and $\omega^2 = k_z^2 c_{\rm t}^2$, where the leading coefficients of the differential equations vanish; these singularities generate the Alfvén continuum and the slow continuum, respectively. The presence of these singularities is an indication of a number of interesting effects connected with the phenomenon of resonant absorption, of particular interest in the question of solar CORONAL HEATING MECHANISMS and in the absorption of p-modes in sunspots.

There is a simple solution of equations (8), (17) and (18) that we may note immediately. If $\partial/\partial\theta = 0$, corresponding to symmetric disturbances of the tube, then equation (18) reduces to

$$\frac{\partial^2 v_\theta}{\partial t^2} = v_A^2(r)\frac{\partial^2 v_\theta}{\partial z^2}. \qquad (22)$$

This is a torsional Alfvén wave which propagates torsional oscillations v_θ with the local value of the Alfvén speed $v_{\rm A}(r)$. Equation (22) illustrates the phenomenon of phase mixing of Alfvén waves, whereby a wave rapidly becomes highly corrugated with large radial gradients building up in v_θ. The buildup of such large gradients is important for the damping of magnetohydrodynamic waves because ordinarily the Alfvén wave in particular is difficult to damp. In a uniform medium the Alfvén wave may propagate vast distances before it is significantly damped by ohmic or viscous processes. However, with phase mixing or resonant absorption operating damping is considerably more efficient.

We may illustrate phase mixing by noting that the solution

$$v_\theta = v_0 \sin\, k_z[z - v_{\rm A}(r)t] \qquad (23)$$

of the Alfvén wave equation (22) gives a wave of amplitude v_0, and $|v_\theta| \le v_0$ at all times. However, the radial gradient $\partial v_\theta/\partial r$ of the motion grows in amplitude linearly in time t, for

$$\frac{\partial v_\theta}{\partial r} = -v'_{\rm A} t v_0 k_z \cos\, k_z[z - v_{\rm A}(r)t] \qquad (24)$$

where v'_A denotes the slope of the Alfvén speed profile. Thus $|\partial v_\theta/\partial r|$ grows secularly on a timescale of $1/v'_{\rm A}$. This phase mixing timescale can be very short. For example, in a coronal loop where the Alfvén speed changes from say 2000 km s^{-1} in the loop to 1000 km s^{-1} in the loop's immediate environment, across a distance of 500 km, the phase mixing timescale is 0.5 s. The result of such phase mixing—and resonant absorption achieves similar results—is that wave damping may occur in significantly shorter times or distances than is the case for magnetohydrodynamic waves in a uniform medium. In fact, an estimate of wave damping for an Alfvén wave undergoing phase mixing in a medium with Alfvén speed gradient $v'_{\rm A}$ gives a damping length (the distance a wave travels before its amplitude is reduced by a factor e, or 37%, of its initial value) of order $(6v_{\rm A}^5/\omega^2 v'^2_{\rm A}\nu)^{1/3}$. The cube root dependence means that damping lengths in a structured medium are considerably shorter than those in a uniform medium. For the coronal conditions illustrated above (and taking ν to be 2×10^{11} m^2 s^{-1}), the damping length for a wave of period 100 s is about 10^4 km. Such considerations are important in the question of coronal heating.

Flux tube waves

Magnetic flux tubes, regions of magnetic field that are distinguished from one another through plasma density or temperature, magnetic field strength or plasma flow, are of particular interest. Two types of flux tube have received detailed study: the isolated flux tube of magnetic field embedded in a field-free medium and the magnetically embedded tube of dense plasma in a magnetic medium. The isolated tube finds specific application in the solar photospheric tube (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES) and the embedded tube has application to the solar wind and the solar coronal loop.

Consider again equations (20) and (21), examining the specific case of a flux tube of radius a, field strength B_0 and plasma density ρ_0 embedded in a magnetic environment with field strength $B_{\rm e}$ and plasma density $\rho_{\rm e}$:

$$B_0(r) = \begin{cases} B_0, & r < a \\ B_{\rm e}, & r > a \end{cases} \qquad \rho_0(r) = \begin{cases} \rho_0, & r < a \\ \rho_{\rm e}, & r > a. \end{cases} \qquad (25)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

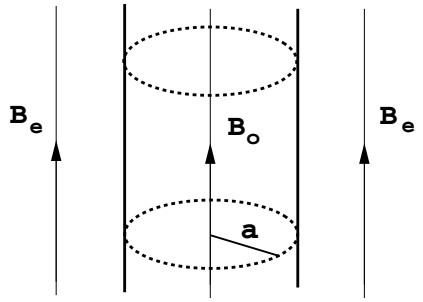


Figure 3. The equilibrium state of a magnetic flux tube of radius a, field strength B_0 and plasma density ρ_0 embedded in a magnetized plasma of field strength B_e and plasma density ρ_e.

The tube is in magnetostatic pressure balance with its surroundings (see equation (4)). The Alfvén, sound and slow speeds within the tube are v_A, c_s and c_t, and their values in the external medium are v_{Ae}, c_e and c_{te} (see figure 3). The case of an isolated tube embedded in a field-free environment corresponds to setting $B_e = 0$ (so that $v_{Ae} = 0$), but for magnetically embedded tubes in a strongly magnetized plasma (i.e. a plasma in which the Alfvén speeds exceed the sound speeds) the field strengths B_0 and B_e are generally comparable and, moreover, $v_{Ae}, v_A >> c_e, c_s$.

For the equilibrium (25) the differential equation (21) may be solved for p_T in terms of Bessel functions $J_n(n_0 r)$ in $r < a$ and $K_n(m_e r)$ in $r > a$, with the result

$$\rho_e(k_z^2 v_{Ae}^2 - \omega^2) n_0 \frac{J_n'(n_0 a)}{J_n(n_0 a)} = \rho_0(k_z^2 v_A^2 - \omega^2) m_e \frac{K_n'(m_e a)}{K_n(m_e a)} \quad (26)$$

where

$$n_0^2 = \frac{(\omega^2 - k_z^2 v_A^2)(\omega^2 - k_z^2 c_s^2)}{(c_s^2 + v_A^2)(\omega^2 - k_z^2 c_t^2)}$$

and

$$m_e^2 = \frac{(k_z^2 c_e^2 - \omega^2)(k_z^2 v_{Ae}^2 - \omega^2)}{(c_e^2 + v_{Ae}^2)(k_z^2 c_{te}^2 - \omega^2)}.$$

Equation (26) is the dispersion relation for flux tube waves, describing magnetoacoustic waves in a uniform magnetic flux tube; it applies for waves that are confined to the tube, requiring that $m_e > 0$. The integer n that arises (through the Fourier analysis in equation (19)) in the description of tube waves defines the geometry of the vibrating tube. The case $n = 0$ corresponds to a symmetric pulsation of the tube, with its axis remaining undisturbed, and is commonly referred to as the sausage wave (or mode). The case $n = 1$ describes the kink mode, which involves lateral displacements of the tube (maintaining a circular cross-section) with the axis of the tube resembling a wriggling snake; the plasma in the neighborhood of the tube is also disturbed. There are also fluting modes ($n \geq 2$), which leave the axis of the tube undisturbed but distort the tube boundary. See figure 4.

The restriction $m_e > 0$ imposed on the flux tube dispersion relation means that the amplitude of a wave

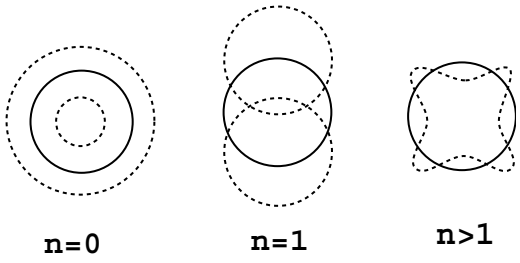


Figure 4. The modes of oscillation of a flux tube, showing the sausage ($n = 0$), kink ($n = 1$) and fluting ($n \geq 2$) waves.

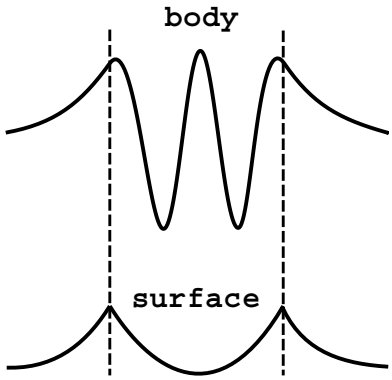


Figure 5. Surface and body waves in a flux tube. Surface waves are largely confined to the region near the boundary of the tube whereas body waves occupy the whole of the tube.

declines with radius r ($>a$), so that far from the tube there is no appreciable disturbance. Inside the tube (for $r < a$) the disturbance is oscillatory if $n_0^2 > 0$ or nonoscillatory (evanescent) if $n_0^2 < 0$. Modes that inside the tube are oscillatory in r are called body waves, because they disturb the whole of the interior of a tube. Waves that have an evanescent form inside and outside the tube are called surface waves, since they are mainly confined to the region near the boundary of the tube. See figure 5. Magnetoacoustic surface waves may also exist on a simple planar interface between two magnetized plasmas. Finally, we note that just as there are both slow and fast magnetoacoustic waves in a uniform medium, there are slow and fast tube waves, and these may be body or surface modes of the sausage, kink or fluting form. Moreover, the presence of the tube radius a in the dispersion relation means that, unlike magnetoacoustic waves in a uniform unbounded medium (equation (12)), the phase speed c ($=\omega/k_z$) of a tube wave depends on its wavelength ($2\pi/k_z$): tube waves are dispersive.

Two speeds prove to be of special significance in magnetic flux tubes: the slow speed c_t of the tube and a mean Alfvén speed c_k, defined by

$$c_k^2 = \frac{\rho_0 v_A^2 + \rho_e v_{Ae}^2}{\rho_0 + \rho_e}. \quad (27)$$

The speeds c_t and c_k are important in both isolated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and magnetically embedded (photospheric and coronal) tubes. The speed c_k is called the kink speed, and it is the speed with which the kink mode of oscillation propagates for waves that are much longer than a tube radius ($k_z a \ll 1$). The speed c_t is associated with both slow body and slow surface waves. The slow surface wave is of particular interest in isolated flux tubes. We may illustrate these speeds for a strongly magnetized plasma and for a photospheric flux tube. In a strongly magnetized plasma such as the solar corona (for which Alfvén speeds are much larger than sound speeds), the slow speed is close to the sound speed inside the tube. For a high-density ($\rho_0 \gg \rho_e$) strongly magnetized plasma (with $B_0 \approx B_e$) the kink speed is $\sqrt{2}v_A$, and so is some 40% higher than the tube's Alfvén speed. By contrast, in a photospheric tube with $\rho_0 \approx \rho_e/2$ and $v_A \approx c_s$, the tube speed c_t is about $v_A/\sqrt{2}$ and the kink speed c_k is about $v_A/\sqrt{3}$, so both speeds are well below the Alfvén speed.

Ducted waves

For a strongly magnetized plasma the dispersion relation (26) possesses two sets of modes, namely fast and slow body waves. There are no surface ($n_0^2 < 0$) waves. Both sets of modes are dispersive, with the fast waves being strongly dispersive. An interesting aspect of these waves is that the fast body waves occur only if $v_{Ae} > v_A$. Thus fast body waves occur in regions of low Alfvén speed which typically correspond to regions of high plasma density; this is the usual circumstance in coronal loops. Regions of low Alfvén speed in strongly magnetized plasmas provide wave guides for fast magnetoacoustic waves. It is interesting to report that fast kink mode oscillations in coronal loops have been identified by the TRACE spacecraft.

There are close analogies here between the behavior of fast magnetoacoustic body modes in a strongly magnetized plasma and Love waves in the Earth's crust and Pekeris sound waves in an internal ocean layer; consequently, fast sausage waves are sometimes referred to as magnetic Pekeris waves and fast kink waves as magnetic Love waves. The distinctive dispersive wave signature they produce when impulsively excited has received interest because of its possible importance in the interpretation of oscillations detected in the corona at radio wavelengths. In particular, an impulsively generated fast body wave produces a signature which consists of three parts: a low-amplitude periodic phase, followed by a larger amplitude quasi-periodic phase, and finally a decay phase. The shortest timescales in the motions are those in the quasi-periodic phase. An estimate of the timescale τ in the periodic phase is provided by

$$\tau = \frac{2\pi a}{j_0 v_A}$$

where j_0 (=2.40) denotes the first zero of the Bessel function J_0. For a tube of radius $a = 10^3$ km and Alfvén speed $v_A = 10^3$ km s^{-1} this produces a timescale of about 1 s.

Nonlinear waves

Magnetohydrodynamic waves may become nonlinear, having amplitudes that are too large to permit a linear treatment. Many of the MHD waves detected in the solar wind and waves generated by a solar flare—and seen to rapidly propagate out across the solar disk—are nonlinear. Alfvén waves provide an exact solution of the nonlinear MHD equations. Nonlinearity frequently leads to wave amplitude growth and steepening, just as an ocean wave undergoes such changes on approaching the shore line. Shock waves are nonlinear waves with profiles that are particularly sharp—discontinuities in the ideal case—and across which dissipative processes act to balance nonlinearity. Conservation relations (total energy, mass, momentum, etc) determine the character of the shock wave. In keeping with the existence of both slow and fast magnetoacoustic waves, both slow shocks and fast shocks may occur. Slow shocks may arise in photospheric flux tubes and in the solar wind, and are also invoked in MAGNETIC RECONNECTION theories; fast shocks would seem to be implicated in observed blast waves generated by a solar flare. When dispersive effects are important, as in magnetic flux tubes, solitons may occur; solitons are waves that propagate nonlinearly preserving their basic shape, even after interactions with other nonlinear waves, and achieving a balance between nonlinearity and dispersion. There are theoretical grounds for believing that solitons may occur in the Sun (see, for example, SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: THEORY).

Bibliography

Goedbloed J P 1983 *Lecture Notes in Magnetohydrodynamics (Rijnhuizen Report 83-145)* (Assoc. Euratom—FOM) pp 289

Goossens M 1991 MHD Waves and wave heating in nonuniform plasmas *Advances in Solar System Magnetohydrodynamics* ed E R Priest and A W Hood (Cambridge: Cambridge University Press) p 137

Roberts B 1991 Magnetohydrodynamic waves in the Sun *Advances in Solar System Magnetohydrodynamics* ed E R Priest and A W Hood (Cambridge: Cambridge University Press) p 105

Roberts B 1992 Magnetohydrodynamic waves in structured magnetic fields *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) p 303

Tu C-Y and Marsch E 1995 *MHD Structures, Waves and Turbulence in the Solar Wind* (Dordrecht: Kluwer) pp 210

Wright A N 1994 MHD theory of magnetic pulsations *Physical Signatures of Magnetospheric Boundary Layer Processes* ed J A Holtet and A Egeland (Norwell: Kluwer) p 329

B Roberts

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamics

Magnetohydrodynamics (or MHD for short) is the study of the interaction between a magnetic field and a plasma treated as a continuous medium (e.g. Cowling 1957, Roberts 1967, Priest 1982, 1994). Most of the universe is not a normal gas but is instead a plasma. We are all familiar on Earth with the three states of matter (solid, liquid and gas). You change from one state to another (such as ice to water) by heating, and if you raise the temperature of gas sufficiently it changes to the fourth state of matter, namely plasma. In the plasma state the atoms have split into positive ions and negative electrons, which can flow around freely, so the gas becomes electrically conducting and a current can flow. (MHD can also be used to describe the behavior of an electrically conducting liquid.)

Many dynamical processes in the universe are caused by the subtle nonlinear relationship between a magnetic field and a plasma. In a normal gas such as the air we breathe, there is virtually no interaction with a magnetic field. But in a plasma the extremely close coupling with the magnetic field means that whatever the plasma is doing intimately affects the magnetic field and vice versa. Indeed, on Earth we are in an extremely unusual part of the cosmos, a tiny island of solid, liquid and gas. But, as soon as we go up to the ionosphere, the plasma universe begins, including the region between Earth and Sun, the whole of the Sun itself, as well as the interstellar and intergalactic media and the stars and galaxies contained in them.

In MHD we are not concerned with individual particles but treat the plasma as a continuous medium. It builds partly on electromagnetism and partly on fluid mechanics. The assumption of a continuous medium is valid for length-scales much larger than the mean-free path for particle collisions

$$\lambda_{\rm mfp} \approx 300\left(\frac{T}{10^6\ {\rm K}}\right)^2\left(\frac{n}{10^{17}\ {\rm m}^{-3}}\right)^{-1}\ {\rm m},$$

which is typically 3 cm in the solar chromosphere and 30 km in the solar corona.

The magnetic field has several physical effects:

(a) It exerts a force, which may accelerate plasma or create structure.
(b) It stores energy, which may later be released as, for example, a solar eruption or a solar flare.
(c) It acts as a thermal blanket, which, when wrapped around a cool solar prominence, say, may protect it from the surrounding corona.
(d) It channels fast particles and plasma.
(e) It drives instabilities and supports waves.

MHD is important in a wide variety of cosmic phenomena. The different aspects of MHD are described in the articles that follow and the applications are discussed in many of the articles throughout the encyclopedia, notably in those about the Sun, where MHD phenomena are widespread. In this introductory article, we first describe a *magnetic flux tube* and then introduce the fundamental equations. Particular attention is given to the *induction equation* and the *Lorentz force* with which a magnetic field acts on a plasma. Then we describe the force balance of a plasma at rest, known as magnetohydrostatics, and mention briefly the possible wave modes and instabilities.

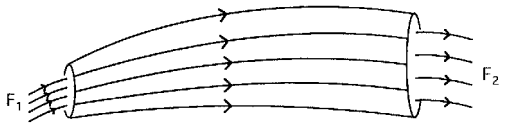


Figure 1. Segment of a magnetic flux tube.

Magnetic flux tubes

A magnetic field line is a curve such that the tangent at any point is in the direction of the magnetic field. Its equation, for a two-dimensional magnetic field having components (B_x, B_y), is

$$\frac{{\rm d}y}{{\rm d}x} = \frac{B_y}{B_x}$$

or ${\rm d}x/B_x = {\rm d}y/B_y = {\rm d}z/B_z$ in three dimensions.

A magnetic flux tube is the surface generated by the set of magnetic field lines which intersect a simple closed curve. Flux tubes are the building blocks of a magnetic configuration, but they must not be thought of as independent isolated structures.

The strength (F) of a flux tube is the amount of magnetic flux crossing a section (S), i.e.

$$F = \int_s \boldsymbol{B}\cdot{\rm d}\boldsymbol{S}.$$

Consider a finite segment of a flux tube bounded by plane sections S_1 and S_2 (figure 1). There is no flux across the walls of the tube, and so physically, if no flux is created inside the tube, the flux (F_1) entering through section S_1 equals the flux (F_2) leaving through section S_2. In other words, the strength (F) is constant along the tube. Mathematically, integrating over the whole surface (S_V) of the segment of the tube between S_1 and S_2 gives, since $\boldsymbol{B}\cdot{\rm d}\boldsymbol{S}$ vanishes on the curved part

$$\int_{S_V}\boldsymbol{B}\cdot{\rm d}\boldsymbol{S} = \int_{S_1}\boldsymbol{B}\cdot{\rm d}\boldsymbol{S} + \int_{S_2}\boldsymbol{B}\cdot{\rm d}\boldsymbol{S} = -F_1 + F_2$$

where F_1 is the flux from left to right and, on S_V, ${\rm d}\boldsymbol{S}$ is in the direction of the outwards normal. However, the Gauss divergence theorem gives

$$\int_{S_V}\boldsymbol{B}\cdot{\rm d}\boldsymbol{S} = \int_V \nabla\cdot\boldsymbol{B}\,{\rm d}V = 0$$

since $\nabla\cdot\boldsymbol{B} = 0$. Thus $F_1 = F_2$ as required, and therefore $\nabla\cdot\boldsymbol{B} = 0$ implies that no flux is indeed created and the strength is constant along the flux tube.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

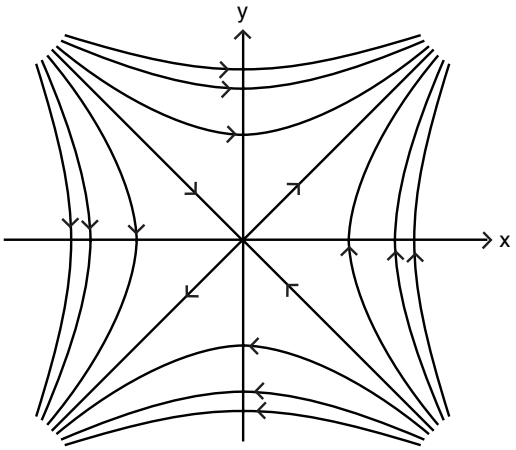


Figure 2. Field lines for the magnetic field $(B_x, B_y) = (y, x)$.

If the cross-sectional area (A) of a flux tube is small, then $F \approx BA$. Thus, as the magnetic field lines become closer together, so A becomes smaller and, since F is constant, B increases in value, and vice versa.

When sketching field lines from expressions for the field components there are three stages:

(a) Evaluate the expressions for the field lines and sketch a typical one.
(b) Decide the directions of the arrows.
(c) Decide the spacings of other field lines.

Thus, for example, if $B_x = y$ and $B_y = x$, the field lines are given by $\mathrm{d}y/\mathrm{d}x = x/y$ or $y^2 - x^2 =$ constant. When the constant is zero we obtain the two field lines $(y = \pm x)$ through the origin (called separatrices); when the constant is positive (negative) we have branches of a rectangular hyperbola intersecting the y-axis (x-axis). From the orientations of the axes in figure 2 we have decided that the positive directions are to the right and upward; thus, for instance, on the positive x-axis the field is simply $B_y = x$, which is positive and so the arrow is directed upwards. Also, as one moves out along the x-axis the magnitude of the field increases and so the field lines become more closely spaced. The origin is a special point, an X-type neutral point, where the field vanishes and the topology of nearby magnetic field lines is hyperbolic; it represents a weak spot in a configuration where magnetic energy tends to be released after the formation of a current sheet.

Fundamental equations

The MHD equations are a unification of the equations of slow electromagnetism and fluid mechanics. Maxwell's equations are

$$\nabla \times \boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t} \qquad \nabla \cdot \boldsymbol{B} = 0 \tag{1}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \qquad \nabla \cdot \boldsymbol{D} = \rho_c \tag{2}$$

where $\boldsymbol{B} = \mu \boldsymbol{H}, \boldsymbol{D} = \epsilon \boldsymbol{E}$ and $\boldsymbol{j}$ and ρ_c are the electric current and charge densities present in the plasma. (Sometimes a subscript 0 is placed on μ and ϵ.)

Here $\boldsymbol{H}$ is the *magnetic field*, $\boldsymbol{B}$ is the *magnetic induction* (although we shall loosely refer to it as the magnetic field), μ the *magnetic permeability* of free space, $\boldsymbol{E}$ the *electric field*, $\boldsymbol{D}$ the *electric displacement*, ϵ the *permittivity* of free space, ρ_c the *charge density*, $\boldsymbol{j}$ the *electric current density*. These are supplemented by Ohm's law

$$\boldsymbol{E} = \boldsymbol{j}/\sigma \tag{3}$$

where σ is the electrical conductivity.

The equations of fluid mechanics for a perfect gas, on the other hand, are

$$\rho \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\nabla p \tag{4a}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho \nabla \cdot \boldsymbol{v} = 0 \tag{4b}$$

$$p = \mathcal{R}\rho T \tag{4c}$$

and an energy equation (see below), where ρ is the *plasma density* (the mass per unit volume), $\boldsymbol{v}$ the *plasma velocity*, p the *plasma pressure*, T the *temperature*, $\mathcal{R}$ the *gas constant*. The operator $\mathrm{d}/\mathrm{d}t = \partial/\partial t + \boldsymbol{v} \cdot \nabla$ is the total (or material) derivative and represents the time rate of change moving with an element of plasma, in contrast to $\partial/\partial t$ which represents the time rate of change at a fixed point of space.

Equation (4a) is the *equation of motion*, which says that the mass times acceleration of a moving element of plasma equals the sum of the forces acting on the element. Here we have included just the pressure gradient $(-\nabla p)$. Equation (4b) is the *equation of mass continuity* and is simply a mathematical expression of the physical fact that no plasma is being created or destroyed: thus, if for example mass is flowing outwards away from an element (so that the flow is 'diverging' and the divergence $\nabla \cdot \boldsymbol{v}$ is positive), then the density of the element must be decreasing so that from (4b) $\mathrm{d}\rho/\mathrm{d}t$ must be negative. Equation (4c) is the *perfect gas law*, stating that the pressure of a plasma is proportional to its density and temperature. In principle (4a) determines the velocity $\boldsymbol{v}$, (4b) the density ρ, while (4c) and an energy equation determine the pressure p and temperature T.

The (internal) energy equation may be written

$$\rho \frac{\mathrm{d}e}{\mathrm{d}t} + p \nabla \cdot \boldsymbol{v} = \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla \boldsymbol{T}) - Q_v + Q_r$$

where

$$e = \frac{p}{(\gamma - 1)\rho}$$

is the internal energy density, $\boldsymbol{\kappa}$ is the thermal conductivity tensor, Q_v is the heating by viscous dissipation, Q_r is the radiative energy loss and γ is the ratio of specific heats. In

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

many astrophysical and solar applications the (optically thin) radiative loss term can be expressed as

$$Q_r = \rho^2 Q(T)$$

where $Q(T)$ is a function describing the temperature variation of the radiative loss.

In a neutral gas the electromagnetic (1)–(3) and fluid dynamic (4) equations are decoupled and so the electromagnetic and fluid properties are independent. However, in MHD we modify the above equations in four ways:

(a) A plasma feels an extra force, the so-called Lorentz force ($\boldsymbol{j} \times \boldsymbol{B}$), which is added on to the right-hand side of (4*a*). It is this term which couples the fluid equations to the electromagnetic equations. It is well known that an element $\mathrm{d}\boldsymbol{l}$ of wire carrying a current J in a magnetic field feels a force $J\,\mathrm{d}\boldsymbol{l} \times \boldsymbol{B}$ perpendicular to the wire and to the field, and so it is natural that a plasma element of volume $\mathrm{d}V$ carrying a current of density $\boldsymbol{j}$ per unit volume should feel a force $\boldsymbol{j}\,\mathrm{d}V \times \boldsymbol{B}$.
(b) The presence of the electric current adds an ohmic heating term (j^2/σ) to the energy equation.
(c) Ohm's law states that the electric field in a frame moving with the plasma is proportional to the current, but the total electric field on moving plasma is $\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}$, where $\boldsymbol{E}$ is the field seen in a frame where the observer is at rest, so (3) is modified by adding $\boldsymbol{v} \times \boldsymbol{B}$ to the left-hand side. This too couples the electromagnetic equations to the fluid equations.
(d) We consider processes with plasma and wave speeds much slower than the speed of light ($v \ll c$), so that the displacement current ($\partial \boldsymbol{D}/\partial t$) in (1) is negligible. This in turn means that we don't need to consider $\nabla \cdot \boldsymbol{D} = \rho_c$ since it just determines ρ_c if needed.

There are at least a dozen different approximations that are commonly used for the MHD system of equations. Some of the more well-known ones are: incompressibility (when ρ is constant following the motion or $\nabla \cdot \boldsymbol{v} = 0$), which requires that the flow speed be much smaller than the sound speed and Alfvén speed; a *steady state* ($\partial/\partial t = 0$ for all variables); the Boussinesq approximation (filtering out sound waves by including density variations only in the gravitational term in the equation of motion); an *isothermal state* (when T = constant); an ideal MHD state (when $\eta = \nu = 0$, $R_m = R_e = \infty$, see below); an inviscid state (when the viscous effects are negligible); an irrotational (potential) flow (when the vorticity, $\nabla \times \boldsymbol{v}$, vanishes); an isentropic state (when the entropy is constant); a force-free field ($\boldsymbol{j} \times \boldsymbol{B} = 0$); a potential magnetic field ($\boldsymbol{j} = 0$); a strong magnetic field ($\beta \ll 1$); a weak magnetic field ($\beta \gg 1$, where β is defined by equation (15) below); a supersonic flow regime (plasma velocity larger than sound speed); a subsonic flow regime (velocity smaller than sound speed).

Validity of MHD in collisionless plasmas

Many cosmic plasmas are collisionless, where the mean free-path for binary collisions between particles is much greater than the characteristic length-scale of the system. MHD ignores particle interactions, and so the reader may at first think that it is of little use in collisionless plasmas. In practice, however, MHD and its two-fluid variants often describe the behavior of collisionless plasmas surprisingly well. For instance, the solar wind beyond 10 solar radii is completely collisionless, but MHD models describe its global velocity, temperature and density quite well, including such time-dependent aspects as shock waves, stream interactions and turbulence. In principle, the appropriate theory for collisionless plasmas is kinetic plasma theory derived from Vlasov's equation. However, due to its mathematical complexity, it is rarely used to construct a global model of the solar wind and is instead confined mainly to calculating effective transport coefficients and modeling localized effects such as shock structure.

MHD has been successful in describing collisionless plasmas for several reasons. Firstly, conservation of mass, momentum and energy are principles common to both ideal MHD and collisionless systems. It is only when dissipation processes are considered that classical MHD becomes problematical. Secondly, ionized particles undergo a gyro-motion about the magnetic field which prevents them from traveling unimpeded in the direction perpendicular to the field. Thus, long-range interactions occur only along the field, while short-range interactions across the field may be described by MHD-like equations. Finally, many plasmas are perturbed by waves which interact with particle motion, even along the field, scattering them in all directions and so re-introducing some form of 'effective' collisionality.

In ideal MHD the magnetic field is frozen to the plasma, but this is also a natural consequence of gyro-motion in a collisionless plasma whenever the $\boldsymbol{E} \times \boldsymbol{B}$ drift of the gyro-centers dominates all other drifts. Thus, if a collisionless plasma contains particles undergoing gyro-motion and the $\boldsymbol{E} \times \boldsymbol{B}$ drift dominates, it will obey the ideal MHD Ohm's law.

The MHD Lorentz force ($\boldsymbol{j} \times \boldsymbol{B}$) is recovered in a collisionless plasma whenever the particle gyro-velocity is much smaller than the speed of light. When the Lorentz force on individual particles is added together it gives a net force per unit volume of

$$\boldsymbol{F} = \rho_c \boldsymbol{E} + \boldsymbol{j} \times \boldsymbol{B}$$

where ρ_c is the electric charge density. However, the ratio of the two terms on the right is

$$\frac{|\rho_c \boldsymbol{E}|}{|\boldsymbol{j} \times \boldsymbol{B}|} \approx \left(\frac{V}{c}\right)^2$$

which is much less than unity when the particle speed (V) is non-relativistic. In other words, $\rho_c \boldsymbol{E}$ is usually

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

negligible compared with the ($\boldsymbol{j} \times \boldsymbol{B}$) force for the low-energy particles.

In a collisionless plasma, the most important new effect arises in the pressure and viscous terms in the momentum equation, since the gas pressure is generally anisotropic relative to the magnetic field. The plasma exerts a pressure ($p_\parallel$) along the field which is usually different from the pressure ($p_\perp$) perpendicular to the field. Similarly, the viscous stress tensor is also anisotropic and dependent on the magnetic field. If $p_\parallel$ and $p_\perp$ can be expressed in terms of the local bulk properties of the plasma, then the MHD equations can still be applied by just adding the appropriate forms of the pressure and viscous stress tensors.

In a collisionless plasma equation (6*a*) in the next section is replaced by the generalized Ohm's law. For a fully ionized plasma it takes the form

$$\boldsymbol{E} = -\boldsymbol{v} \times \boldsymbol{B} + \frac{\boldsymbol{j}}{\sigma} + \frac{m_e}{ne^2}\left(\frac{\partial \boldsymbol{j}}{\partial t} + \nabla \cdot (\boldsymbol{vj} + \boldsymbol{jv})\right) - \frac{\boldsymbol{j} \times \boldsymbol{B}}{ne} - \frac{\nabla \boldsymbol{p}_e}{ne}$$

where $\boldsymbol{vj}$ and $\boldsymbol{jv}$ are dyadic tensors and $\boldsymbol{p}_e$ is the electron stress tensor. The first term on the left-hand side of this equation is the convective electric field, while the second term is the field associated with Ohmic dissipation caused by electron–ion collisions. The next three terms describe the effects of *electron inertia*, while the penultimate term is the *Hall effect* and the last term includes the *electron gyroviscosity*. For a partially ionized plasma, collisions between charged particles and neutrals lead also to ambipolar diffusion.

Induction equation

With the above assumptions, equations (1)–(3) become

$$\boldsymbol{j} = \nabla \times \boldsymbol{B}/\mu \tag{5a}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \tag{5b}$$

$$\boldsymbol{E} = -\boldsymbol{v} \times \boldsymbol{B} + \boldsymbol{j}/\sigma \tag{6a}$$

where

$$\nabla \cdot \boldsymbol{B} = 0 \tag{6b}$$

We may therefore eliminate $\boldsymbol{j}$ and $\boldsymbol{E}$ by substituting for $\boldsymbol{j}$ from (5*a*) in (6*a*) and for $\boldsymbol{E}$ from (6*a*) in (5*b*), with the result that $\partial \boldsymbol{B}/\partial t = \nabla \times (\boldsymbol{v} \times \boldsymbol{B}) - \eta \nabla \times (\nabla \times \boldsymbol{B})$, where $\eta = 1/(\mu\sigma)$ is the magnetic diffusivity, here assumed uniform. By expanding out the triple vector product in the last term and using (6*b*), we obtain finally

$$\frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times (\boldsymbol{v} \times \boldsymbol{B}) + \eta \nabla^2 \boldsymbol{B} \tag{7}$$

which is known as the induction equation.

This is the basic equation for the magnetic behavior in MHD: it determines $\boldsymbol{B}$ once $\boldsymbol{v}$ is known. In electromagnetism the electric current and electric field are primary variables, with the current driven by electric fields, and then the magnetic field is a secondary variable produced by currents. However, in MHD the basic physics is quite different, since the plasma velocity ($\boldsymbol{v}$) and magnetic field ($\boldsymbol{B}$) are the primary variables, determined by the induction equation and the equation of motion, while the resulting current ($\boldsymbol{j}$) and electric field ($\boldsymbol{E}$) are secondary and may be deduced from (5*a*) and (6*a*) if required.

If V_0, L_0 are typical velocity and length-scale, the ratio of the first to the second term on the right-hand side of (7) is, in order of magnitude, the magnetic Reynolds number

$$R_m = \frac{L_0 V_0}{\eta}.$$

Thus, for example, in a solar active region where $\eta \approx 1\ \mathrm{m^2\ s^{-1}}$, $L_0 \approx 10^5$ m, $V_0 \approx 10^4\ \mathrm{m\ s^{-1}}$ we find $R_m \approx 10^9$ and so the second term on the right of (7) is completely negligible. In turn, equation (6*a*) reduces to $\boldsymbol{E} = -\boldsymbol{v} \times \boldsymbol{B}$ to a very high degree of approximation. This is the case in almost all of the solar atmosphere, indeed in almost all of the universe—the only exception is in regions (such as current sheets) where the length-scale is extremely small, so small that $R_m \leq 1$ and the second term on the right of (7) becomes important.

If $R_m \ll 1$, the induction equation reduces to

$$\frac{\partial \boldsymbol{B}}{\partial t} = \eta \nabla^2 \boldsymbol{B} \tag{8}$$

and so $\boldsymbol{B}$ is governed by a diffusion equation, which implies that field variations (irregularities) on a scale L_0 diffuse away on a time-scale of

$$\tau_d = \frac{L_0^2}{\eta}$$

which is obtained simply by equating the orders of magnitude of both sides of (8). The corresponding speed at which they diffuse is

$$v_d = \frac{L_0}{\tau_d} = \frac{\eta}{L_0}.$$

With $\eta \approx 1\ \mathrm{m^2\ s^{-1}}$, the decay time for a sunspot is (with $L_0 = 10^6$ m) 10^{12} sec = 30 000 yr, so that the process whereby sunspots disappear in a few days cannot just be diffusion (and is probably instead decaying by the convection away from the spot of many small flux tubes). Similarly, the diffusion time for a magnetic field pervading the Sun as a whole (with $L_0 = 7 \times 10^8$ m) is 5×10^{17} s = 10^{10} yr. This is of the order of the age of the universe, so a magnetic field in a star at its formation has not had time to diffuse much.

Consider, for example, a one-dimensional magnetic field ($B(x,t)\hat{\boldsymbol{y}}$) satisfying

$$\frac{\partial B}{\partial t} = \eta \frac{\partial^2 B}{\partial x^2} \tag{9}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

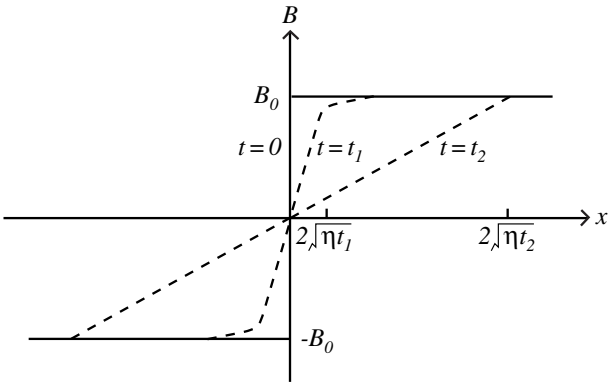


Figure 3. The magnetic field (B) in a diffusing one-dimensional current sheet as a function of distance (x) for times $t = 0, t_1, t_2$.

and suppose that initially the field is $B(x, 0) = B_0$ for $x > 0$ and $B(-x, 0) = -B(x, 0)$, as shown in figure 3. Physically, what do we expect to happen? Since (9) has the form of a heat conduction equation and we know that heat tends to flow from a hot region to a cool one and smooth out a temperature gradient, we expect the same diffusive process to occur for our magnetic field and for the initially steep magnetic gradient at $x = 0$ to smooth out, as shown in figure 3.

Mathematically, the required solution of (9) turns out to be, in terms of the *error function* (erf(ξ))

$$B(x,t) = B_0 \operatorname{erf}\left(\frac{x}{\sqrt{(4\eta t)}}\right) = \frac{2B_0}{\sqrt{\pi}} \int_0^{x/\sqrt{(4\eta t)}} \mathrm{e}^{-u^2}\, \mathrm{d}u.$$

This solution does indeed have the form shown in figure 3 and it may be verified *a posteriori* by substituting back into (9). The resulting field lines diffuse through the plasma and cancel at $x = 0$.

The main reason for variations in R_m from one phenomenon to another is variations in the appropriate length-scale (L_0). If $R_m \gg 1$, the induction equation becomes

$$\frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times (\boldsymbol{v} \times \boldsymbol{B})$$

and Ohm's law reduces to $\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B} = \boldsymbol{0}$ so that the total electric field in a frame of reference moving with the plasma vanishes. Then, if we consider a curve $\mathcal{C}$ (bounding a surface $\mathcal{S}$) which is moving with the plasma, in a time $\mathrm{d}t$ an element $\mathrm{d}s$ of $\mathcal{C}$ sweeps out an element of area $\boldsymbol{v}\,\mathrm{d}t \times \mathrm{d}\boldsymbol{s}$. The rate of change ($\mathrm{d}/\mathrm{d}t \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}$) of magnetic flux through $\mathcal{C}$ then consists of two parts, namely

$$\int_{\mathcal{S}} \frac{\partial \boldsymbol{B}}{\partial t} \cdot \mathrm{d}\boldsymbol{S} + \int_{\mathcal{C}} \boldsymbol{B} \cdot \boldsymbol{v} \times \mathrm{d}\boldsymbol{s}.$$

As $\mathcal{C}$ moves, so the flux changes partly because the magnetic field is changing with time (the first term) and partly because of the motion of the boundary (the second term). Then, by putting $\boldsymbol{B} \cdot \boldsymbol{v} \times \mathrm{d}s = -v \times B \cdot \mathrm{d}\boldsymbol{s}$ and applying Stokes' theorem to the second term, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \int_{\mathcal{S}} \left(\frac{\partial \boldsymbol{B}}{\partial t} - \nabla \times (\boldsymbol{v} \times \boldsymbol{B}) \right) \cdot \mathrm{d}\boldsymbol{S}$$

which vanishes in the present approximation.

Thus, the total magnetic flux through $\mathcal{C}$ remains constant as it moves with the plasma. In other words we have proved magnetic flux conservation (figure 4(*a*)). It follows that plasma elements that form a flux tube initially do so at all later times (figure 4(*a*)). There is also magnetic field line conservation (or conservation of magnetic connectivity), namely that, if two plasma elements lie on a field line initially they will always do so (figure 4(*b*)). At $t = t_1$, say, suppose elements P_1 and P_2 lie on a field line, which may be defined as the intersection of two flux tubes. Then, at some later time ($t = t_2$) by magnetic flux conservation P_1 and P_2 will still lie on both tubes, and so they will lie on the field line defined by their intersection. We interpret the above results by saying that the magnetic field lines move with the plasma—we say that they are frozen into the plasma—and plasma can move freely along field lines, but in motion perpendicular to them they are dragged with the plasma or vice versa.

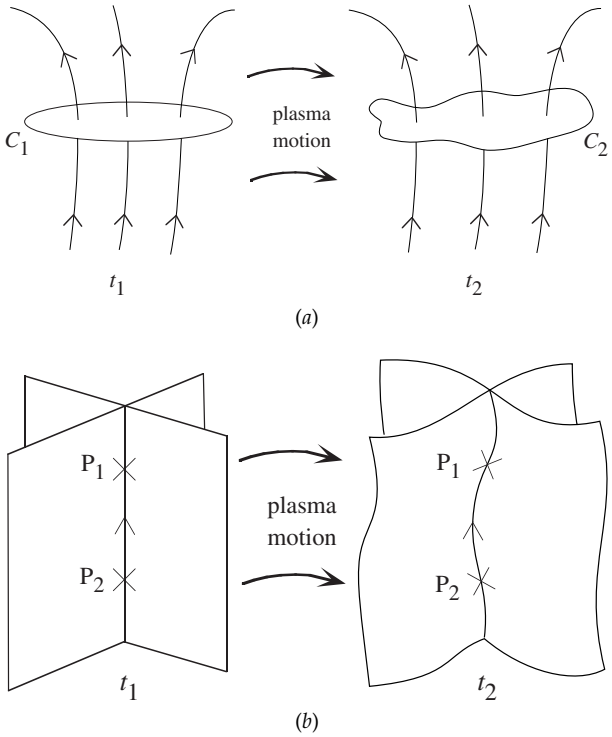


Figure 4. (*a*) Magnetic flux conservation—if a curve $\mathcal{C}_1$ is distorted into $\mathcal{C}_2$ by a plasma motion, the flux through $\mathcal{C}_1$ at t_1 equals the flux through $\mathcal{C}_2$ at t_2. (*b*) Magnetic field line conservation—if plasma elements P_1 and P_2 lie on a field line at t_1, then they will lie on the same line at t_2.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Lorentz force

One equation relating our fundamental variables $\boldsymbol{v}$ and $\boldsymbol{B}$ is the induction equation; the other is the equation of motion, whose most common form is

$$\rho \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\nabla p + \boldsymbol{j} \times \boldsymbol{B} \tag{10}$$

although other forces such as viscous or gravitational may sometimes be important too. The term $-\nabla p$ is the plasma pressure gradient: it acts from regions of high plasma pressure to low pressure and is perpendicular to the isobars (curves of constant pressure).

The magnetic force ($\boldsymbol{j} \times \boldsymbol{B}$) is perpendicular to the magnetic field, and so any plasma acceleration parallel to the magnetic field must be caused by other forces. By substituting for $\boldsymbol{j}$ from Ampère's law (5*a*) and using an identity for the triple vector product, we can write it as

$$\boldsymbol{j} \times \boldsymbol{B} = (\nabla \times \boldsymbol{B}) \times \boldsymbol{B}/\mu = (\boldsymbol{B} \cdot \nabla)\frac{\boldsymbol{B}}{\mu} - \nabla\left(\frac{B^2}{2\mu}\right).$$

The two terms on the right-hand side have important physical interpretations. Since the second term has the same form as $-\nabla p$, we can say that it represents the effect of a *magnetic pressure* of magnitude $B^2/(2\mu)$. It gives a force when B^2 varies with position, and the direction of the force is from regions of high magnetic pressure to low magnetic pressure.

The first term represents the effect of a *magnetic tension* parallel to the magnetic field of magnitude B^2/μ per unit area. It gives a force when the field lines are curved, just like an elastic band or rope. By writing $\boldsymbol{B} = B\hat{\boldsymbol{s}}$ in terms of the unit vector ($\hat{s}$) along the magnetic field, the tension term ($(\boldsymbol{B} \cdot \nabla)B/\mu$) may be written

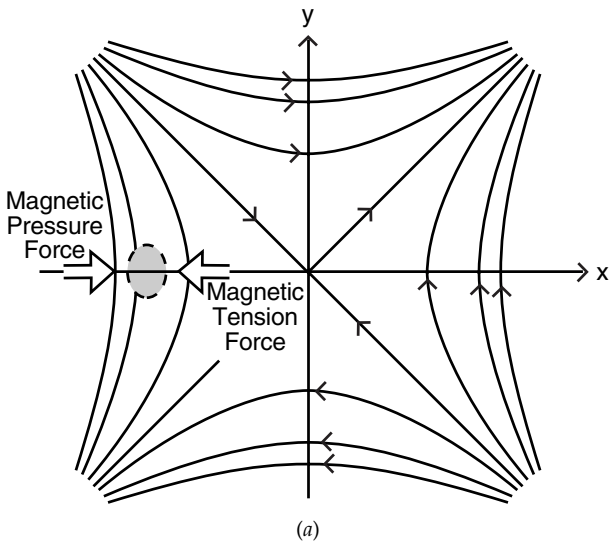
$$\frac{B}{\mu}\frac{\mathrm{d}}{\mathrm{d}s}(B\hat{\boldsymbol{s}}) = \frac{B^2}{\mu}\frac{\mathrm{d}\hat{\boldsymbol{s}}}{\mathrm{d}s} + \frac{B}{\mu}\frac{\mathrm{d}B}{\mathrm{d}s}\hat{\boldsymbol{s}} = \frac{B^2}{\mu}\frac{\hat{\boldsymbol{n}}}{\mathcal{R}} + \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{B^2}{2\mu}\right)\hat{\boldsymbol{s}}$$

where $\hat{\boldsymbol{n}}$ is the principal normal and $\mathcal{R}$ is the radius of curvature. The second term on the right of this equation is irrelevant, since it cancels with the component of $-\nabla(B^2/2\mu)$ parallel to $\boldsymbol{B}$, as it must since $\boldsymbol{j} \times \boldsymbol{B}$ is perpendicular to $\boldsymbol{B}$. However, the first term on the right is the magnetic tension term, which shows that when the radius of curvature is small the tension force is large.

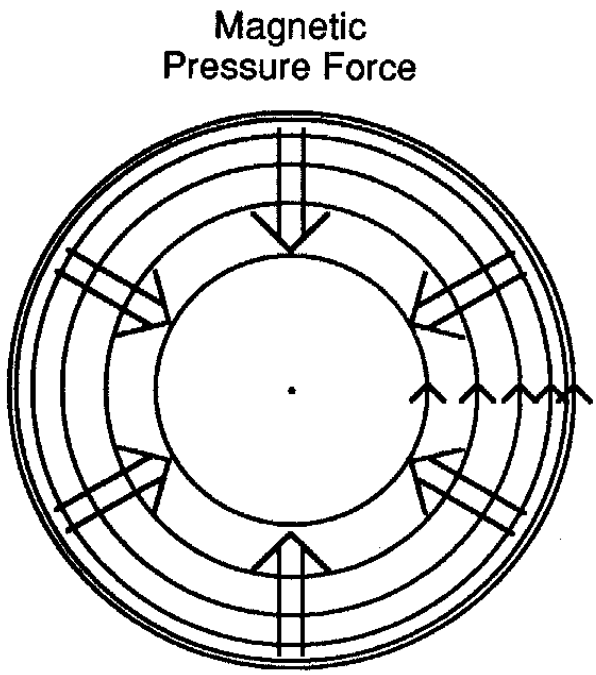
Thus, we have shown that the magnetic force has two distinct effects: the magnetic pressure tries to compress the plasma and produces a net force if B varies with position; the magnetic tension tends to make the field lines shorten themselves like elastic bands and gives a net force if the field lines are curved.

As an example, consider the field with components $B_x = y$, $B_y = x$, with X-type field lines shown in figure 5(*a*). At a point on the positive x-axis, the curvature of the field lines suggest a magnetic tension force to the right, while the fact that the magnetic pressure is increasing (like x^2) away from the origin suggests a magnetic pressure force to the left. However, from the field lines alone, it is not evident which dominates. Now

$$(\boldsymbol{B} \cdot \nabla)\frac{\boldsymbol{B}}{\mu} = \left(y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}\right)\left(\frac{y}{\mu}\hat{\boldsymbol{x}} + \frac{x}{\mu}\hat{\boldsymbol{y}}\right) = \frac{x}{\mu}\hat{\boldsymbol{x}} + \frac{y}{\mu}\hat{\boldsymbol{y}}$$



(*a*)



(*b*)

Figure 5. Magnetic forces due to the fields (*a*) $\boldsymbol{B} = y\hat{\boldsymbol{x}} + x\hat{\boldsymbol{y}}$ and (*b*) $\boldsymbol{B} = r\hat{\theta}$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which becomes $x\hat{\boldsymbol{x}}/\mu$ on the x-axis and so does indeed act away form the origin. Also, the magnetic pressure force is

$$-\nabla\left(\frac{B^2}{2\mu}\right) = -\nabla\left(\frac{x^2+y^2}{2\mu}\right) = -\frac{x}{\mu}\hat{\boldsymbol{x}} - \frac{y}{\mu}\hat{\boldsymbol{y}}$$

which becomes $-x\hat{\boldsymbol{x}}/\mu$ on the x-axis and so does act towards the origin. We find therefore that the tension and pressure forces are in precise balance, as they must be since the electric current is

$$\boldsymbol{j} = \frac{1}{\mu}\left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)\hat{\boldsymbol{z}} = 0$$

and so the Lorentz force $\boldsymbol{j}\times\boldsymbol{B}$ vanishes everywhere.

Also, consider next the field $\boldsymbol{B} = r\hat{\theta}$, for which the field lines are the circles r = constant, becoming more closely spaced as one moves outwards (figure 5(*b*)). Thus, physically the magnetic tension and pressure forces are expected both to act inwards. Mathematically, we find the magnetic pressure force is

$$-\nabla\left(\frac{B^2}{2\mu}\right) = -\nabla\left(\frac{r^2}{2\mu}\right) = -\frac{r}{\mu}\hat{\boldsymbol{r}}$$

which does indeed act inwards. Also, the magnetic tension is

$$(\boldsymbol{B}\cdot\nabla)\frac{\boldsymbol{B}}{\mu} = \left(\frac{B_\theta}{r}\frac{\partial}{\partial\theta}\right)\left(\frac{B_\theta\hat{\boldsymbol{\theta}}}{\mu}\right) = \frac{\partial}{\partial\theta}\left(\frac{r\hat{\boldsymbol{\theta}}}{\mu}\right)$$

and now r/μ does not vary with θ but $\hat{\theta}$ does! In fact $\partial\hat{\theta}/\partial\theta = -\hat{\boldsymbol{r}}$ and so $(\boldsymbol{B}\cdot\nabla)\boldsymbol{B}/\mu = -r\hat{\boldsymbol{r}}/\mu$, which again acts inwards, as expected.

Magnetohydrostatics

Consider the equation of motion when pressure, magnetic and gravitational forces are acting

$$\rho\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\nabla p + \boldsymbol{j}\times\boldsymbol{B} + \rho\boldsymbol{g} \tag{11}$$

where

$$\boldsymbol{j} = \nabla\times\boldsymbol{B}/\mu. \tag{12}$$

If L_0, v_0, L_0/v_0 are typical values for the length-scale, plasma velocity and time-scale respectively, the order of magnitude of the current from (12) is $j_0 = B_0/\mu L_0$. Then, in terms of the typical density (ρ_0), pressure (p_0) and magnetic field (B_0), the sizes of each term in equation (11) are $\rho_0 v_0^2/L_0$, p_0/L_0, $B_0^2/\mu L_0$, $\rho_0 g$, respectively.

Now, provided the magnetic term is of the same order as the largest force term, we have force balance if the first term is much smaller than the third, namely

$$v_0^2 \ll \frac{B_0^2}{\mu\rho_0} \equiv v_A^2$$

where v_A is known as the *Alfvén speed*. Then (11) reduces to the equation *for magnetohydrostatic force balance*

$$0 = -\nabla p + \boldsymbol{j}\times\boldsymbol{B} + \rho\boldsymbol{g} \tag{13}$$

where $\boldsymbol{j} = \nabla\times\boldsymbol{B}/\mu$, $\nabla\cdot\boldsymbol{B} = 0$ and $\rho = p/(\mathcal{R}T)$.

In this equation, when gravity is negligible we have magnetostatic balance

$$0 = -\nabla p + \boldsymbol{j}\times\boldsymbol{B} \tag{14}$$

if the fourth term in (11) is much smaller than the third, namely $L_0 \ll B_0^2/(\mu\rho_0 g) = 2H/\beta$, where

$$H = \frac{p_0}{\rho_0 g} = \frac{\mathcal{R}T_0}{g}$$

is the pressure scale-height and

$$\beta = \frac{p_0}{B_0^2/(2\mu)} \tag{15}$$

in the plasma beta, namely the ratio of plasma to magnetic pressure.

In turn, in (14) the magnetic force dominates if $2\beta \ll 1$, and then (14) reduces further to the equation

$$\boldsymbol{j}\times\boldsymbol{B} = \boldsymbol{0} \tag{16}$$

for a FORCE-FREE MAGNETIC FIELD, in which the magnetic field is in equilibrium with itself under a balance between magnetic pressure and magnetic tension forces.

Suppose now that gravity is directed vertically downwards in the negative z-direction ($\boldsymbol{g} = -g\hat{\boldsymbol{z}}$). Then the component of (13) parallel to a particular magnetic field line is $0 = -\mathrm{d}p/\mathrm{d}s - \rho g\cos\theta$, where s is measured along the field, or, since $\mathrm{d}s\cos\theta = \mathrm{d}z$, we have $0 = -\mathrm{d}p/\mathrm{d}z - \rho g$. After putting $\rho = p(\mathcal{R}T)^{-1}$ this becomes

$$\frac{\mathrm{d}p}{\mathrm{d}z} = -p\frac{g}{\mathcal{R}T}$$

which may be integrated to give

$$p = p_0\exp\left(-\int_0^z \frac{g}{\mathcal{R}T}\,\mathrm{d}z\right)$$

where p_0 is the pressure at the base ($z = 0$) of the field line. If the variation ($T(z)$) of the temperature with height is known, this determines the pressure and therefore the density.

If, in particular, the temperature is uniform ($T = T_0$) then

$$p = p_0\,\mathrm{e}^{-z/H}$$

so that the pressure (and density) decrease exponentially with height, with the scale-height H being the vertical distance over which the pressure falls off by a factor e.

Down in the solar photosphere, where $T_0 = 6000$ K, the scale-height is about 150 km or less and so, over for instance a vertical distance of 1.5 Mm, the pressure and density would fall off by a factor of $\mathrm{e}^{10} \approx 20\,000$. By contrast, in the solar corona where $T_0 = 2\times10^6$ K, say, the scale-height is about 100 Mm, and so the density falls off much more slowly. Indeed, for many purposes, in the corona we can neglect the effect of gravity, i.e. when the vertical scales of interest are 100 Mm or less.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Potential fields

If the pressure gradient and gravitational force are negligible, (16) holds and a particular case of interest is when the current vanishes, so that

$$\nabla \times \boldsymbol{B} = \boldsymbol{0} \tag{17}$$

where

$$\nabla \cdot \boldsymbol{B} = \boldsymbol{0}. \tag{18}$$

Equation (17) may be satisfied identically by putting $\boldsymbol{B} = \nabla\psi$ and then (18) gives Laplace's equation

$$\nabla^2\psi = 0.$$

Thus, many of the general results of potential theory may be applied and the magnetic field is said to be current-free or potential.

Magnetic reconnection

MAGNETIC RECONNECTION is a fundamental process in a plasma, whereby the magnetic connectivity of plasma elements changes and magnetic energy is converted into other forms such as heat, kinetic energy and fast-particle energy. It involves a localized breakdown of ideal MHD in a small region due to, for example, resistive effects. It is responsible for many dynamic processes in the universe, such as SOLAR FLARES and GEOMAGNETIC STORMS and probably solar coronal heating (see CORONAL HEATING MECHANISMS).

Magnetic helicity

For a perfectly conducting plasma in a closed volume (V), the MAGNETIC HELICITY is

$$K = \int_V \boldsymbol{A} \cdot \boldsymbol{B}\, \mathrm{d}V$$

where $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, and the state of minimum energy for fixed K is a linear force-free field. K is a measure of the sum of the twist and the linkage of flux tubes. Thus, for two linked tubes

$$K = \Theta_1 F_1^2 + \Theta_2 F_2^2 + 2L F_1 F_2$$

where Θ_1 and Θ_2 are the twists of the two tubes F_1, F_2 their magnetic fluxes and L the linking number. Reconnection can convert linkage helicity to twist helicity, but it produces only a small change in the total helicity in a plasma of large global magnetic Reynolds number.

Magnetohydrodynamic waves and shock waves

In a gas there are sound waves which propagate equally in all directions at the sound speed

$$c_s = \left(\frac{\gamma p_0}{\rho_0}\right)^{1/2}.$$

In a plasma there are also waves, but they are of several types. Waves are very important in the solar atmosphere and throughout the cosmos. For example, they may be seen propagating out of sunspots or away from large solar flares. They are also a prime candidate for heating the solar atmosphere.

In MHD there is an incompressible mode known as the Alfvén wave, which propagates at the Alfvén speed along the magnetic field. Also, there are two compressible modes known as slow and fast magnetoacoustic waves, which propagate slower and faster respectively than the Alfvén speed. These MHD waves have propagation characteristics that depend on the direction of propagation relative to the magnetic field.

Small-amplitude sound waves propagate without change of shape, but when the amplitude is finite the crest can move faster than its trough, causing a progressive steepening. Ultimately, the gradients become so large that dissipation becomes important, and a steady *shock wave* shape may be attained with a balance between the steepening effect of the nonlinear convective term and the broadening effect of dissipation. The dissipation inside the shock front converts the energy being carried by the wave gradually into heat. The effect of the passage of the shock is to compress and heat the gas.

In MHD the slow magnetoacoustic wave can steepen to form a slow-mode shock and the fast wave to form a fast-mode shock.

Magnetohydrodynamic instabilities

Equilibrium magnetic fields in the universe can go sometimes unstable to a wide variety of instabilities. Sometimes the effect is to create fine-scale structure and at other times it is to produce dynamic events such as solar flares and dramatic eruptions from the Sun called SOLAR CORONAL MASS EJECTIONS. The two main ways of analysing stability are by investigating the natural (or normal) modes of variation or by an energy (or variational) method.

The instabilities that commonly occur include: the *Rayleigh–Taylor instability*, when heavy fluid is supported above light fluid; *kink instability* or sausage instability of a magnetic flux tube when its twist is too large; *resistive instability* such as the tearing-mode instability of a current sheet or a sheared structure; convective instability of a plasma that is heated sufficiently from below; and flow instabilities that occur when the shear in a flow is too great (such as the *Kelvin–Helmholtz instability* and in accretion discs the *Balbus–Hawley instability*).

Conclusion

It is important to recognize that magnetohydrodynamics builds on the tools of both fluid dynamics and electromagnetism, but it possesses many new features that are present in neither. Furthermore, it is advisable to build both a physical and mathematical understanding of magnetic field behavior, since they both complement one another and together give a deeper understanding than either alone.

In the following articles, various aspects and applications of MHD theory are developed in detail. In a companion article, Parker describes the key physical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

processes involved in LARGE-SCALE PLASMA DYNAMICS. Then basic DYNAMO THEORY for the generation and maintenance of cosmical magnetic fields is developed by Moffatt and in particular for the GEODYNAMO by Proctor and for solar and stellar dynamos by Rosner. Then several important aspects of MHD theory are described, including FORCE-FREE MAGNETIC FIELDS (Low), MAGNETIC HELICITY (Berger) and MHD WAVES (Roberts). The theory for MHD INSTABILITIES is set up by Hood and applied in particular to MAGNETIC BUOYANCY by Hughes and to MHD: MAGNETOCONVECTION by Cattaneo. The key process of MAGNETIC RECONNECTION is reviewed by Schindler and Hörnig and its role in MHD: MAGNETIC RECONNECTION AND TURBULENCE is described by Matthaeus. Finally, two astrophysical environments where MHD plays a crucial role are discussed, namely the MHD OF ACCRETION DISKS by Brandenburg and the MHD OF ASTROPHYSICAL WINDS by Heyvaerts.

Acknowledgments
We are most grateful to colleagues in St Andrews and Durham and to Jean Heyvaerts and Karl Schindler for helpful comments.

Bibliography
Cowling T G 1957 *Magnetohydrodynamics* (New York: Interscience)
Priest E R 1982 *Solar Magnetohydrodynamics* (Dordrecht: Reidel)
Priest E R 1994 Magnetohydrodynamics *Plasma Astrophysics* ed J G Kirk, D B Melrose and E R Priest (Berlin: Springer) pp 1–112
Priest E R and Forbes T G 2000 *Magnetic Reconnection: MHD Theory and Applications* (Cambridge: Cambridge University Press)
Roberts P H 1967 *An Introduction to Magnetohydrodynamics* (London: Longman)

Eric R Priest and Terry G Forbes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamics of Accretion Disks

An ACCRETION DISK is a flat formation of gas and dust rotating about a central object and accreting matter inwards by transporting angular momentum outwards, so that the centrifugal support is gradually removed from the fluid parcels. There are three main classes of accretion disks: (i) disks around accreting compact stars (white dwarfs, neutron stars or black holes) in binary systems, (ii) disks around protostars and (iii) disks around active galactic nuclei. In several cases the existence of disks has been verified observationally with the Hubble Space Telescope. Images can be found at http://www.stsci.edu/pubinfo/pictures.html. An example of an image of a disk around a young stellar object together with an associated jet is given in figure 1.

Accretion disks form because matter generally has angular momentum and therefore cannot fall directly onto the central object. In the disk midplane the gravitational acceleration, GM/R^2, at distance R from the central object of mass M is balanced mainly by the centrifugal acceleration, $\Omega^2 R$. Here G is the gravitational constant and Ω is the angular velocity. Except in a few systems the disk mass is negligible and so this balance leads to the rotation law

$$\Omega(R) = (GM/R^3)^{1/2} \tag{1}$$

which is essentially Kepler's third law. In the vertical direction, i.e. parallel to the rotation axis, gravity is balanced by a pressure gradient, so that the mass is concentrated towards the midplane. There is usually some energy dissipation caused by turbulence and magnetic fields. The heat released during this process can be radiated away at the two disk surfaces. The corresponding removal of rotational energy causes matter gradually to spiral towards the central object. Large-scale fields and winds from the disk surfaces can also remove angular momentum from the disk, and would thus contribute to the accretion mechanism directly.

Accretion disks around young stellar objects are only transient phenomena because they provide essentially just a waiting queue for matter before it can fall onto the central object. The disk will disappear once all the available material has been consumed, which could be after about 10^7 yr in the case of protostellar disks.

Energy conversion in disks

The kinetic and potential energy of accreted matter in the disk is constantly being converted into heat and then radiation by dissipative processes. In disks around neutron stars or white dwarfs the luminosity of the disk can by far exceed the luminosity of the central object. The total orbital energy per unit mass is $\frac{1}{2}(R\Omega)^2 - GM/R = -\frac{1}{2}GM/R$. The mass accretion rate $\dot{M}$ is controlled by the rate of dissipation, giving a disk luminosity of

$$L = \frac{GM\dot{M}}{2R}. \tag{2}$$

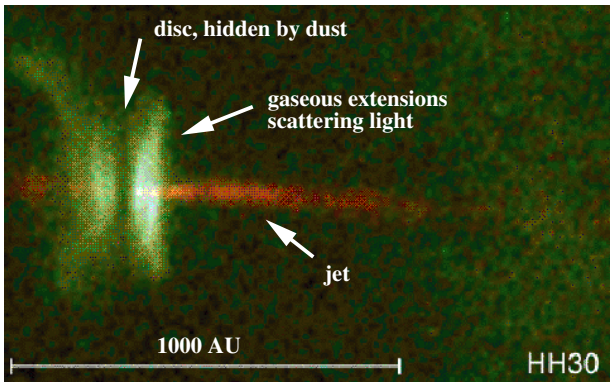


Figure 1. Image of a disk around a young stellar object together with a jet emanating from the disk center along the rotation axis. The length of the horizontal bar is 1.5×10^{11} km. (Adapted from Burrows *et al* (1996).)

This formula is essentially independent of the efficiency of the mechanism that accomplishes the energy conversion. If the central object is a white dwarf or a neutron star there will be a boundary layer at the star's surface, where as much energy can be dissipated as in the disk itself.

Equation (2) assumes that all the heat is radiated away instantaneously. In recent years it has been found that in some disks the luminosity is much lower, because significant amounts of energy can be advected radially towards the center. If the central object is a black hole the advected energy may never appear as radiation.

In the case of a black hole at the center the disk can extend down to three Schwarzschild radii, i.e. $r_{\min} = 6GM/c^2$, or less in the case of rotating black holes (c is the speed of light). The energy that is released by an accreted mass m is $\frac{1}{2}GMm/r_{\min}$. Using for $r_{\min}$ the expression above, this becomes ηmc^2, where $\eta \approx 0.1$ is an efficiency factor relative to the maximum possible value permitted by Einstein's famous formula $E = mc^2$. Note that the efficiency of hydrogen fusion in stars is only $\eta = 0.007$. For this reason the rate of energy release by accretion, $\eta\dot{M}c^2$, of a disk around a supermassive black hole of 10^8 solar masses with $\dot{M}$ of a few solar masses per year can be as large as 10^{40} W. This is believed to be the mechanism that powers quasars.

The role of magnetic fields

In the absence of turbulence a purely laminar shear motion would be totally insufficient to explain the dissipation and corresponding heat release of real disks. Thus, turbulence is necessary to produce small enough scales where microscopic viscosity and ohmic diffusion can act to dissipate kinetic and magnetic energy into heat.

It has been a long-standing debate as to what causes turbulence in disks. In the absence of magnetic fields differentially rotating disks are unstable when the specific angular momentum, ΩR^2, increases inwards (Rayleigh's criterion). Indeed, purely hydrodynamic mechanisms such as (nonlinear) instabilities and convection have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

proved unsuccessful so far. Numerical work and theoretical arguments indicate that nonlinear instabilities do not operate in astrophysical disks and that convection produces accretion torques that are probably even of the wrong sign; furthermore, this process would not be self-sustained. For a comprehensive review of those issues see Balbus and Hawley (1998). The issue of purely hydrodynamical instabilities is not fully settled, however, because this would require direct simulations at Reynolds numbers of at least 10 000, which cannot be achieved with present computers.

In the presence of magnetic fields, however, there is a powerful linear instability, the magneto-rotational instability, which prevents the gas flow in the disk from being laminar. In 1991 Balbus and Hawley pointed out its importance in driving turbulence in accretion disks. Their papers have spawned a lot of work attempting to quantify the properties of the resulting turbulence and the related angular momentum transport. The magneto-rotational (or Balbus–Hawley) instability is a consequence of the shear, which causes a destabilization of the slow magnetosonic waves. The instability exists regardless of the orientation of the magnetic field. In the special case of a magnetic field parallel to the rotation axis the instability is axisymmetric.

In the absence of magnetic fields, or when the electrical conductivity is too low to make the magnetic field important, the disk would be purely hydrodynamic, in which case rotation would have a strongly stabilizing effect. Still, variation of the angular velocity in the vertical direction would lead to a linear hydrodynamic instability, but its growth rate is much smaller than that of the magneto-rotational instability. The vertical shear instability could be important if the conductivity is low.

The magneto-rotational instability

The magneto-rotational instability exists already in an incompressible, unstratified fluid that is differentially rotating provided that the shear parameter

$$q \equiv -\partial \ln \Omega / \partial \ln R \tag{3}$$

is positive. For thin accretion disks we have $q = +3/2$; see equation (1). In that case the dispersion relation is

$$\omega^4 - \omega^2(2v_A^2 k^2 + \Omega^2) + v_A^2 k^2 (v_A^2 k^2 - 3\Omega^2) = 0 \tag{4}$$

where $v_A = B/(\mu_0 \rho)^{1/2}$ is the Alfvén speed, B is the vertical field strength, ρ is the unperturbed density, μ_0 is the permeability, k is the wavenumber along the magnetic field and ω is the frequency. There are two solutions for ω^2, an upper branch corresponding to Alfvén waves and a lower branch corresponding to slow magnetosonic waves. The fast magnetosonic waves have been filtered out by the assumption of incompressibility. On the lower branch ω^2 becomes negative when

$$v_A < \sqrt{3}\Omega / k. \tag{5}$$

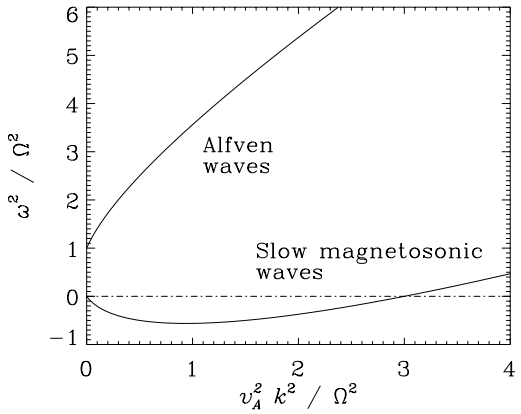


Figure 2. Dispersion relation for slow magnetosonic and Alfvén waves in an incompressible medium. For Alfvén waves ω^2 is always positive. However, for not too large values of k the slow magnetosonic waves become unstable when $v_A^2 k^2 < 3\Omega^2$, so $\omega^2 < 0$, that is ω is imaginary.

In that case ω is purely imaginary and small perturbations grow exponentially with a maximum growth rate

$$\max(\mathrm{Im}\,\omega) = \tfrac{3}{4}\Omega \tag{6}$$

at scale $\ell = 2\pi/k_{\max}$, where

$$k_{\max} \approx \Omega / v_A. \tag{7}$$

It is clear from equation (5) that the instability works only if the field is not too strong. The largest possible field strength depends on the smallest admissible k, i.e. the largest scale available to the system. One such scale would be the disk height.

The disk is formally unstable even in the limit of vanishing magnetic field strength, $v_A \to 0$. However, maximum growth would then occur for perturbations whose scale would become progressively smaller, as given by equation (7). At larger scales the growth of the instability would be so slow that viscous effects would render the instability irrelevant for driving hydromagnetic turbulence.

In figure 2 we have plotted the two branches of the dispersion relation, equation (4), corresponding to slow magnetosonic and Alfvén waves. There are also the fast magnetosonic waves that result if the assumption of incompressibility is relaxed (see Balbus and Hawley 1998). However, when the sound speed c_s is much larger than v_A, the two lower branches are nearly independent of the neglect of the fast magnetosonic branch.

There is a mechanical analogue to the magneto-rotational instability, which can be helpful in understanding the nature of the instability. The following example applies to the nonaxisymmetric case that is relevant in the presence of a toroidal magnetic field. Consider two particles, A and B, on a gravitationally bound orbit around a central object (figure 3). Assume that the two particles have the same distance from the central object but

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

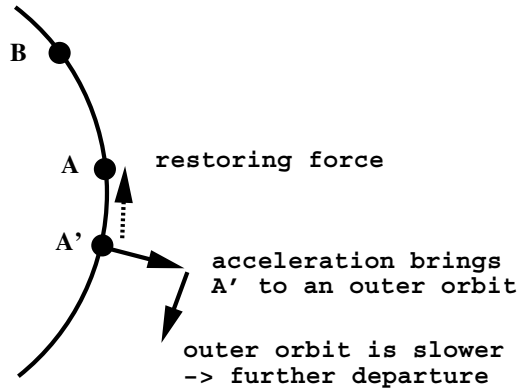


Figure 3. Mechanical analogue of the magneto-rotational instability.

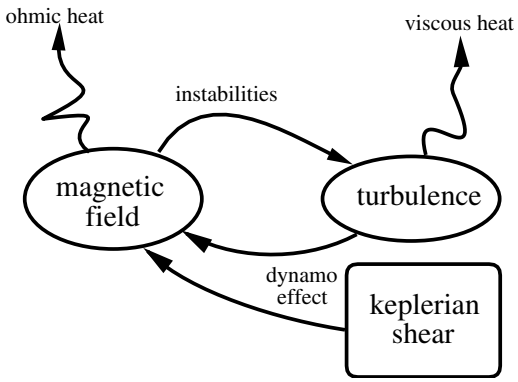


Figure 4. The energy in the Keplerian shear motion is converted into thermal energy via ohmic and viscous heating. Magnetic fields are needed as a catalyst to keep the system turbulent.

are separated azimuthally by a small amount. Interesting behavior occurs when the two particles are coupled elastically. Assume that the particle at position A is perturbed to position A′. The separation between the two particles is then increased and there is a restoring force trying to return the particle to its original position. This accelerates the particle from A′ back towards A. However, this leads to a gain in its angular momentum which then forces the particle onto an outer orbit. There, however, the orbital speed is slower, see equation (1), so the particle will separate from its original position A even further. This is the cause of an instability.

There are some parallels with the phenomenon of tidal disruption of a star passing near a black hole. In that example the restoring force is the gravitational attraction that holds the star together. In the case of the magneto-rotational instability the restoring force is the magnetic tension force. If the field is too strong, however, the fluid parcels stay in their original position, suppressing the instability.

If the conductivity in the disk is poor, there will be significant slippage between the field and the fluid. The magneto-rotational instability ceases when the collision frequency of neutral atoms with ions becomes less than the rotational frequency. This can happen in some parts of protostellar disks which have temperatures below 1000 K. At those temperatures the degree of ionization is very low and the fraction of charge carriers can be as small as 10^{-10}. In the protosolar nebula, out of which the solar system was formed, this may have been the case in a broad ring near the Earth's orbit.

Intrinsic magnetism of disks

It has long been suspected and now been confirmed by local simulations of rotating shear flow that the resulting turbulence is capable of maintaining the field by DYNAMO action. (See Schramkowski and Torkelsson (1996) for a review.) Dynamo action is a process by which kinetic energy can be converted into magnetic energy. In disks the energy comes from the kinetic energy in the shear. Most of it goes into the magnetic field which, in turn, keeps the instability going, thereby feeding the turbulence. All the energy that is removed from the shear is constantly being dissipated (see figure 4). In that sense the magnetic field acts effectively like a catalyst that enables kinetic energy to be tapped from the shear motion and then to be released as heat and radiation via viscous and ohmic heating.

The magnetic fields can contribute directly to exerting a torque that leads to an accretion flow and to angular momentum transport outwards. Any radial field component, B_R, will be sheared out by the differential rotation. This enhances the azimuthal field component, B_ϕ. The resulting torque, which is proportional to the product $B_R B_\phi$, is such that there is an inward mass flow in the disk and an outward transport of angular momentum. If the field was perfectly frozen into the gas (no magnetic diffusion) the field lines would eventually be parallel to the flow and would have no effect.

The strength of the dynamo-supported magnetic field is limited by various nonlinear feedbacks. On the one hand, when the field is too strong the magneto-rotational instability is suppressed, see equation (5), but there are other mechanisms, such as magnetic buoyancy, which can also limit the field strength.

A nondimensional measure for the total stress is

$$\alpha \equiv \langle \rho u_R u_\phi - B_R B_\phi / \mu_0 \rangle / \langle p \rangle \qquad (8)$$

which is found from numerical simulations to be typically between 0.01 and 0.1. Here, $\boldsymbol{u}$ is the velocity, and p is the pressure. The precise value of α depends on the magnetic field strength, which may vary with time. Furthermore, since the gas pressure drops faster with height than the magnetic field, α can increase away from the midplane; see equation (8). However, there are at present no global simulations that include the region far outside the disk, where the field must eventually fall off.

Knowing the value of α the radial disk structure can be calculated in closed form, by neglecting the vertical disk structure. In terms of the accretion rate $\dot{M}$ the vertically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

integrated disk density, $\Sigma = \int \rho \, dz$, is given as a function of radius by

$$\Sigma \approx 2000 \alpha_{-2}^{-4/5} \dot{M}_{13}^{7/10} M_1^{1/4} R_8^{-3/4} \text{ kg m}^{-2} \quad (9)$$

where $\alpha_{-2} = \alpha/10^{-2}$, $\dot{M}_{13} = \dot{M}/10^{13}$ kg s^{-1}, M_1 is the central mass in solar masses and $R_8 = R/10^8$ m (e.g. Campbell 1997).

The disk structure is not always steady. In fact, in some parameter regimes, for example where ionization and recombination become important, the disk can become viscously unstable and undergo limit cycle oscillations (CATACLYSMIC BINARIES). Similar phenomena are also known to occur in disks around young stars. In that case the steady solution would still be described by an equation similar to equation (9), but with different coefficient and exponents, because of the different radiative processes involved.

External magnetic fields

External magnetic fields are maintained by currents outside the disk, for example by currents in the central object (typically a neutron star or a white dwarf, but not a black hole), or in the environment in which the disk is embedded (molecular cloud or host galaxy).

If the magnetic field comes from a star at the center, angular momentum can be transferred to the star from those parts of the disk whose local angular velocity exceeds the stellar angular velocity. This is the case inside the corotation radius, i.e. the radius where the angular velocity of the disk coincides with that of the star. If the part of the disk inside the corotation radius contains sufficient angular momentum, this process can lead to a noticeable spin-up of the central star itself. Otherwise, if the disk does not extend sufficiently far inside the corotation radius most of the field lines of the star couple with the slowly rotating outer parts of the disk which then leads to spin-down of the star. It is often difficult to say whether spin-up or spin-down will occur; this depends on the field strength and field geometry that results from the presence of a central star with a magnetic field, all of which affect the precise location of the inner radius of the disk. Indeed, there are stars where spin-up and spin-down phases are observed to alternate on a timescale of months and years. This could be explained by changes of the location of the inner edge of the disk.

The star's magnetic field increases sharply towards the star (as r^{-3} for a dipole field) and the strong field in the inner parts causes the disk to disrupt. The precise disruption mechanism and hence the precise location of the inner disk radius are still controversial. Possible mechanisms include a viscous instability of the disk (Campbell 1997) or simply a loss of hydrostatic equilibrium. Somewhere near that radius the field will also become too strong for the magneto-rotational instability to operate; see equation (5). Near the inner edge of the disk gas is thought to be channelled along the stellar magnetic field lines towards the central star.

Outflows and jets

The outer layers, away from the disk midplane, are probably heated by MAGNETIC RECONNECTION (Joule dissipation). This plausibly leads to the formation of a hot CORONA. As in the Sun, this layer can then no longer be in hydrostatic equilibrium and therefore some gas must be blown off continuously in the form of a wind. Large-scale magnetic fields can also directly contribute to accelerating outflows. This may be possible if there is a large-scale poloidal field tilted away from the rotation axis by at least 30°. In that case the component of the centrifugal force along the field will dominate over the corresponding component of the gravity force and so gas can be driven along field lines away from the disk.

At larger distances from the disk surface these outflows are seen to be strongly collimated towards the rotation axis. There is at present no clear consensus as to what causes this collimation into jets (see ASTROPHYSICAL JETS). Perhaps the most plausible mechanism is based on magnetic forces, in particular the radially inward pointing component of the Lorentz force (i.e. the hoop stress), which results from the presence of a strong toroidal field. This toroidal field is partly advected from the disk by a slow wind and partly generated just outside the jet by shearing the poloidal field. Today most models assume an externally maintained magnetic field. It is still unclear whether dynamo-generated magnetic fields can be responsible for the launching and collimation of a jet. Other topics of current research concern the variability and knottedness of jets.

Bibliography

Balbus S A and Hawley J F 1991 A powerful local shear instability in weakly magnetized disks. I. Linear analysis *Astrophys. J.* **376** 214–22

Balbus S A and Hawley J F 1998 Instability, turbulence, and enhanced transport in accretion disks *Rev. Mod. Phys.* **70** 1–53

Burrows C J *et al* 1996 Hubble Space Telescope observations of the disk and jet of HH 30 *Astrophys. J.* **473** 437–51

Campbell C G 1997 *Magnetohydrodynamics in Binary Stars* (Dordrecht: Kluwer)

Schramkowski G P and Torkelsson U 1996 Magnetohydrodynamic instabilities and turbulence in accretion disks *Astron. Astrophys. Rev.* **7** 55–96

Axel Brandenburg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamics of Astrophysical Winds

Stellar mass loss became an issue when Parker (1958) first introduced the concept of the solar wind. Such a thermal wind results from coronal heating processes indirectly associated with the existence of a convection zone, also responsible, with rotation, for the star's dynamo action (see articles on SOLAR WIND: THEORY, CORONAL HEATING MECHANISMS, SOLAR INTERIOR: CONVECTION ZONE and MAGNETOHYDRODYNAMIC WAVES). So, mass loss of this type also exists for a large class of main sequence stars. Magnetized winds affect, by Lorentz torques, the angular momentum budget of the rotating objects from which they emanate. The rotation of solar-type stars at their formation is much larger than that of the Sun, although not close to the break-up velocity (at which the centrifugal force balances gravity at the equator). By this mechanism, a single star emitting an MHD wind progressively spins down. Its rotation-associated activity declines in time, as is in fact observed (see article on SOLAR–STELLAR CONNECTION). Stars of a type earlier than F2 which have no such activity remain fast rotators when aging, as observed. For active stars in synchronized binary systems, such as RS CVn, the loss of spin angular momentum by wind emission is converted into a loss of orbital angular momentum, which, unlike single stars, causes them to spin up and become more active, as observed. Newly born solar-type stars of the T Tauri class rotate typically at only one-tenth of their break-up velocity (Bertout 1989). Some spin-down mechanism must be active during the protostellar accretion phase, probably also due to winds emanating from the star and from the protostellar accretion disk. The structure of MHD winds has been first explored by *a priori* assuming some shape for the flow surfaces (Weber and Davis 1967). Pulsar winds called for relativistic extensions of these studies (Michel,1969 Goldreich and Julian 1970). In the past two decades new objects have been discovered which emit winds which are so strongly directive as to be named jets. Relativistic bulk velocities are reached in some of these flows, associated with active galactic nuclei or galactic micro-quasars (see article on JETS). Most often the jet phenomenon is associated with accretion, i.e. the capture of external material by intense gravitational pull to some central, usually compact, object. These jetting winds raised again the question of the two-dimensional structure of MHD flows and of wind acceleration mechanisms, by MHD or other forces.

Angular momentum and accretion

Accretion cannot proceed as direct fall to the accreting object because the matter to be accreted initially possesses too much angular momentum. In the absence of angular momentum loss, each plasma element would simply orbit the accreting object without ever falling on it. It is crucial to the accretion process that at least part of this gas gets rid of excess angular momentum.

In friction between neighboring orbiting gas elements one gas parcel loses and the other gains angular momentum. The one which loses descends on an orbit closer to the accreting object, while the other recedes to a broader one. Viscosity due to friction between atoms is by far insufficient to achieve observed accretion rates but instabilities in the disk, mediated by a weak magnetic field, have been shown to supply turbulence causing a large enough equivalent viscous effect (Balbus and Hawley 1991). A gaseous turbulent disk orbiting the accreting object forms. In its inner regions matter slowly spirals down to the center (Shakura and Sunyaev 1973) while the disk's outskirts slowly expand as a growing reservoir of angular momentum. Tidal torques from a companion eventually limit this growth.

An alternative possibility is that winds or jets convey the excess angular momentum of accreting matter to infinity. The process is similar to the one which torques down stars which emit a wind.

How are winds and jets propelled?

The solar wind is accelerated by the pressure of the hot solar corona. The innermost regions of some accretion disks may be very hot too and emit a wind that could be made directive by hydrodynamical interaction with an anisotropic environment, such as the disk itself. Flow refraction at oblique radiative shocks could add to this effect for forming bipolar anisotropic outflows (Mellema and Frank 1997), but the most extreme focusing is unlikely to be of a purely hydrodynamic origin.

The intense radiation field in the environment of some of these objects could cause radiative acceleration. For the radiative force on a completely ionized plasma to exceed the gravitational one, the luminosity of the radiating object should exceed a rather high threshold proportional to its mass, the Eddington luminosity. This, however, does not seem to be the actual situation for most active galactic nuclei. Intense radiation could still efficiently boost a leptonic component of the plasma to relativistic bulk velocities in jets where such motions and high-energy radiation are observed. More on this can be read in papers by O'Dell (1981), Dermer and Schlickeiser (1993) and Markowith *et al* (1995).

The Lorentz force density $\vec{j} \times \vec{B}$, $\vec{j}$ being the electric current density and $\vec{B}$ the magnetic field, could very well play a role in plasma acceleration, as discussed below. In relativistic dynamics, ignored here for simplicity, the electric force density should be considered as well.

Stationary, axisymmetric MHD flows

Magnetized plasma winds are often discussed assuming axisymmetry, stationarity and perfect MHD. The basic concepts of MHD are reviewed in the article on MAGNETOHYDRODYNAMICS. If, as is often envisaged, the wind source is an accretion disk, it should be effectively dissipative, for otherwise the magnetic flux threading the disk would be advected with the accreting matter and pile up near the accreting object or in it. The necessary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

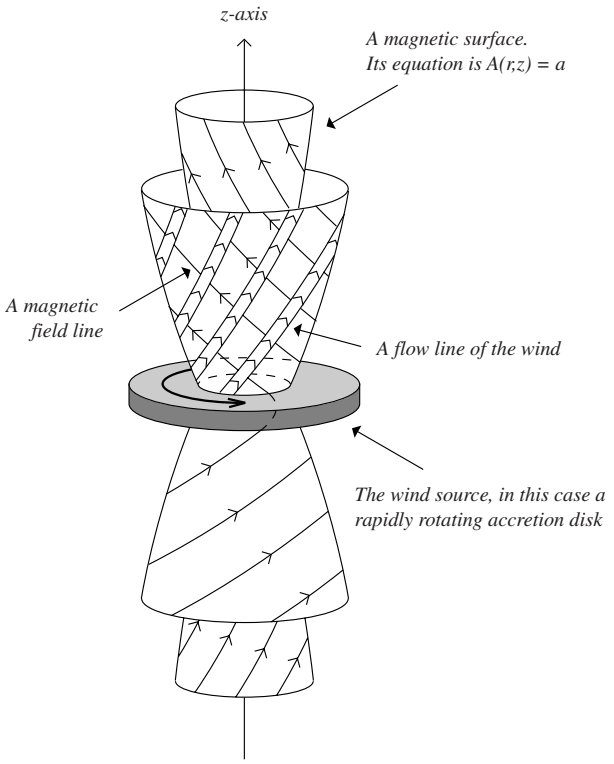


Figure 1. Geometry of a stationary, axisymmetric MHD flow.

magnetic diffusivity could be provided by some form of turbulence or, in partially ionized plasmas, by ambipolar diffusion (Königl 1989).

General properties

In axisymmetric and stationary MHD flows, field lines lie on axially symmetric surfaces, the magnetic surfaces, with equation $A(r, z) = a$ in cylindrical coordinates r, ϕ, z. $A(r, z)$ is the flux through the circle of radius r centered on the axis at position z. Each value of a defines one such surface. Stationarity and perfect MHD imply that the plasma flow lies on magnetic surfaces. This geometry is represented in figure 1. The fluid velocity is the superposition of a field-aligned motion and of a rigid-body rotation of the magnetic surface, at angular velocity $\Omega(a)$. The ratio of mass flux to magnetic flux is a constant, $\alpha(a)$, on a surface a. The motion also conserves the ratios, $E(a)$ and $L(a)$, of total energy and angular momentum flux to mass flux. There is a magnetic contribution to these quantities since the energy flux has a Poynting part, while the azimuthal component of the Lorentz force exerts a torque on the plasma. If a polytropic approximation is made to the actual thermal energy balance, the ratio $P/\rho^\gamma = Q(a)$ is also constant on each magnetic surface. The rates of mass, energy and angular momentum loss in the wind are integrals from pole to equator of $\alpha(a)$, $\alpha(a)E(a)$ and $\alpha(a)L(a)$. It is important to note that the specific angular momentum per escaping particle, $L(a)$, can be significantly larger that the specific angular momentum of the matter at the root of the magnetic surface a on the wind source, situated, say, at an axial distance $r_0(a)$, which is almost exactly Ωr_0^2. The Alfvén radius on magnetic surface a, defined by $r_{\rm A}^2(a) = L(a)/\Omega(a)$, is the effective lever arm of escaping matter, typically the distance up to which the magnetic field remains strong enough to impose corotation of escaping plasma with the wind source. For strong enough magnetization it may be larger than r_0, the loss of angular momentum of the wind source being then dominated by the torques exerted by electric currents induced in the wind. This is how single stars with a convection zone, and thus dynamo action and winds, can be spun down. A calculation of this evolution is, however, difficult because it involves the time history of the star's magnetic flux and of the distribution with a of $r_{\rm A}(a)$. There are two ways by which the magnetic field can accelerate the plasma in the wind zone. One is by centrifuging it away, the other is by the upward push of the magnetic pressure of the azimuthal field.

Centrifugally driven winds

For perfect MHD conditions, magnetic field lines behave like rigid structures wherever the magnetic energy density exceeds the kinetic and rotational energies. Regions on a field line where such conditions are met are closer to the source than the Alfvén point. In a rest frame accompanying the rotation of a given magnetic field line, the constrained plasma elements can glide on the 'rigid' field lines like beads on a wire in response to the centrifugal force and gravitational pull. The combined force density vanishes on the accretion disk surface if the latter is in Keplerian rotation and the net field-aligned force is oriented outwards if the angle of the meridional field with the direction of the axis exceeds 30°. Under these conditions a plasma flow can be centrifugally driven along the field lines threading the accretion disk (Blandford and Payne 1982). An exactly similar process at the surface of a star is not possible because this would imply that it rotates near breakup, but some gas pressure could push the plasma over the initial potential barrier, the flow becoming centrifugally driven further out. It takes adequate plasma conditions in the disk for favorable conditions for such acceleration to be realized at the disk's surface. If the field is too stiff or the magnetic diffusivity too large, the poloidal field lines cannot be sufficiently bent to meet the minimum 'launching angle'. If the field is too flexible, plasma cannot be efficiently centrifuged away. Detailed analysis has shown that, for this whole process to work, the plasma β parameter in the disk should be of order unity, and the disk magnetic Reynolds number based on radius and radial accretion velocity should be much larger than unity (Ferreira and Pelletier 1993; Li 1995). It is not yet known whether such conditions can indeed be achieved.

Coiled spring acceleration

When the magnetic field is not dynamically dominant, inertia bends the field lines backwards to rotation and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

an azimuthal field component appears. The vertical component of the associated Lorentz force is opposite to the gradient of B_ϕ^2 and can lift the plasma up if this quantity decreases with altitude. This is the 'plasma gun' force (Contopoulos 1995), which can be described also pictorially as the force exerted by a coiled spring: a field line, which is more 'coiled' where B_ϕ is larger, acts as a locally overtwisted spring which pushes on the plasma to relax its stresses. Since B_ϕ is zero, by symmetry, on the equatorial plane of the disk, this force has an unfavorable sense there, pinching the disk's plasma. Favorable conditions would usually be met further away from the source, past the Alfvén point, where this effect may provide continued acceleration relaying the centrifugal force described above. For the 'coiled spring' force to be effective close to the disk's surface, the Alfvén point must be close to it, which means that the field needs to be relatively weak, as opposed to the situation where centrifugal acceleration operates. Numerical calculations have proved the effectiveness of this process although possibly in an intrinsically time-dependent regime (Shibata and Uchida 1986; Ouyed and Pudritz 1997).

MHD torque and accretion

The Lorentz force has an azimuthal part which exerts a torque on the plasma. When B_z is an even function of z, it acts close to the disk's equatorial plane to decelerate the rotation. The angular momentum extracted from the disk's plasma is transferred to the outflowing wind which brings it to infinity. Plasma ejection thus appears as a means of getting rid of the excess angular momentum of accreting matter, a fraction of it being diverted out for ejection. Accretion and ejection then form together a self-consistent global machine in which the accretion flow causes wind ejection, while the torque exerted by the wind on the disk allows accretion to proceed. This can be achieved at the expense of a minor mass loss if the specific angular momentum of escaping wind material is much greater than the specific angular momentum of accreting material, which implies that the Alfvén radius r_A is much larger than the radial distance r_0 of the foot point of the field line in the disk. Centrifugally driven winds would meet this requirement more easily. The mechanical energy lost by accreting matter in wind-driven accretion, unlike in viscosity-driven accretion, does not reappear on the spot as radiation due to local dissipation but mostly escapes as wind kinetic energy, undetectable until its eventual dissipation at intermediate or terminal shocks.

What focuses jets?

Evidence for focusing

There is considerable direct evidence for focusing of mass ejection from active galactic nuclei as well as from galactic microquasars (see article on JETS). Young stellar objects in the protostellar evolution stage also have anisotropic outflows which come in the form of weakly collimated molecular outflows, with a much faster and more focused central stream visible in optical lines (Mundt 1985) and containing Herbig Haro objects which are cold condensations formed in these flows. Outflows from planetary nebulae also show, in a rather large subclass, considerable evidence for focused outflows (Livio 1997).

The hoop stress

By which physical mechanisms can such winds be turned into well-focused outflows? The extreme focusing observed in jets from active galactic nuclei and similar objects of stellar size calls for some explanation that could be sought in MHD forces. Indeed, one particular component of the Lorentz force, $\vec{j}_z \times \vec{B}_\phi$, which is radial, should have the required sign for focusing, for it can be expressed in terms of the total current $J(r, z)$ through the circle of radius r centered at the point z on the z-axis as $(-\mu_0/2\pi r)Jj_z$, which is indeed negative, hence pinching, if J and j_z are of the same sign. Since j_z is the surface density of J, this must necessarily be so in at least some central part of the flow.

The transfield equation

Assuming stationarity, axisymmetry and perfect MHD the shapes of the magnetic surfaces, and flow surfaces, are determined by the projection of the equation of motion perpendicular to them. The resulting transfield equation (Okamoto 1978) is a partial differential equation for the function $A(r, z)$ which also involves the density $\rho(r, z)$. All other quantities can be eliminated using the first integrals provided by the other MHD equations which integrate once, and so would it be for ρ itself if the Bernoulli equation, expressing conservation of the total specific energy $E(a)$, were explicitly solvable for ρ in terms of $|\nabla A|$ and A. A transfield–Bernoulli system can be established not only in classical MHD but also for special relativistic flows and for flows in a Kerr space–time appropriate for describing general relativistic effects near rotating black holes (Beskin 1997; Camenzind 1986). The transfield equation is of second order, linear in the second-order derivatives of $A(r, z)$ but non-linear in the function and its first derivatives. Its properties depend on the ratio of fluid velocity to characteristic propagation speeds, described in the article on magnetohydrodynamic waves. It is singular at the Alfvén surface, the locus of points where the meridional flow velocity v_P equals the meridional Alfvén speed, at the slow critical surface, the locus where v_P equals the slow-mode speed in its direction, and at the fast critical surface, the locus where v_P equals the fast-mode speed. The latter two surfaces are associated with singularities of the Bernoulli equation, regular solutions of which require special conditions to be satisfied at these surfaces. They determine those first integrals which are not simply given by boundary conditions (Weber and Davis 1967; Heyvaerts and Norman 1989). The character of the second-order derivative part of the equation changes from elliptic to hyperbolic at the cusp surface, from hyperbolic to elliptic at the slow surface and from elliptic to hyperbolic again at the fast surface. In each hyperbolic region, a double set of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

characteristics can be defined in the meridian plane. It has been shown that one such set has a limiting characteristic, a line to which characteristics of the considered family become tangent. The limiting characteristics represent the border of one region of the hyperbolic domain which cannot communicate with the other one by the considered mode. In particular, the fast limiting characteristic surface, not the fast critical surface, is the boundary at which MHD influence of the downstream regions on the upstream ones is shut off.

Asymptotics

The asymptotics of MHD flow, as expressed by the transfield–Bernoulli system, must be studied to see whether indeed the hoop stress causes its progressive alignment with the polar axis. General exact results have been obtained for polytropic flows (Heyvaerts and Norman 1989). An upper bound on the value of $r|\nabla A|$ in the asymptotic domain is obtained from the obvious requirement that the Poynting flux cannot exceed the total energy flux, which has been used to prove that, as r approaches infinity following a magnetic surface, z/r can only approach a non-zero limit or diverge. Far from the polar axis in the asymptotic domain the potentially dominant force perpendicular to magnetic surfaces is the hoop stress, except in the vicinity of exceptional magnetic surfaces where it must vanish and can therefore not dominate over other forces (Heyvaerts and Norman 1997). The asymptotic form of the transfield equation simply reduces to the statement that, except in thin boundary layers, the hoop stress vanishes. Between these boundary layers, the total asymptotic meridional electric current enclosed in the magnetic surface, $J_\infty(a)$, becomes a constant J_∞ independent of a, implying a vanishing meridional current density. If J_∞ is zero the quantity z/r diverges following magnetic surfaces which asymptotically become nested paraboloids. If, on the other hand, J_∞ is non-zero, no current can flow between pairs of magnetic surfaces on which r can approach infinity. It must flow in some central core consisting of asymptotically cylindrical magnetic surfaces on which r remains finite. The current which reaches infinity in this peri-axial core returns to the wind source through one of the boundary layers forming around those magnetic surfaces on which the meridional field, and therefore also the azimuthal field which is derived by inertial bending from it, vanishes. For a dipolar type of symmetry, the equatorial plane is one such exceptional surface. At precisely these surfaces, the hoop stress vanishes while it competes with opposing forces, such as pressure, in their close neighborhood. Such regions have the structure of a sheet pinch, as observed in the solar wind (Thomas and Smith 1981). It then seems that hoop stresses indeed have the potential to focus efficiently the rotating plasma stream and to turn asymptotically the wind into a very directive jet. This idea is also supported by numerical simulations (Sakurai 1985; Ustyugova *et al* 1995) and exact analytical solutions for several different types of self-similar flow (Blandford and Payne 1982; Chan and Henriksen 1980; Tsinganos and Sauty 1992).

Stability issues

Numerical studies have revealed that the flow regime may in fact not be stationary at all, even when the mass inflow and its rotation are kept constant at the boundary of the computing domain. Ouyed and Pudritz (1997) have shown examples of such behavior for flows propelled by the toroidal field coiling. When the back-reaction of angular momentum extraction on the wind source is taken into account, the system has been found to enter an unstable collapse behavior (Matsumoto *et al* 1996). Therefore the question of stability of such flows is particularly relevant.

Axisymmetric MHD wind flows exhibit shear motions and carry electric currents, features which are known to be potential sources of instabilities (see the article on MAGNETOHYDRODYNAMIC INSTABILITIES). This important issue has been addressed up to now in simplified geometries. Differential axial motions could drive Kelvin Helmholtz (KH) instabilities, causing friction and mixing between the focused wind and the neighboring medium. The linear stability of different geometries to this type of perturbation has been rather comprehensively discussed (Birkinshaw 1991), also with an axial uniform magnetic field. Body modes, which affect the whole plasma in the jetting wind, appear to be typical of these supersonic flows and are found to be unstable on a length-scale of order a Mach number of times the jet diameter. High density in the fast-jetting wind or a strong axial field reduce the instability which may even be quenched for small enough β parameters. In current-carrying jetting winds, MHD instabilities, such as the kink one, are likely to develop because the toroidal component tends to grow much larger than the poloidal one and the Kruskal–Shafranoff instability criterion, that the field coils more than one turn about the axis, is easily satisfied even on a relatively short length along the axis. The growth rates of these modes depend on the radial current profile and can be smaller than (Appl and Camenzind 1992) or comparable with those of KH modes. However, the most significant issue concerning instabilities is their non-linear evolution. It may either lead to small-scale turbulence, giving rise to an effectively dissipative flow which may still retain its global organized character, or cause a large-scale change of the fluid motion. It appears that the KH instability of a magnetized medium tends to smooth out the axial velocity profile forming a turbulent transition layer around the jetting wind (Malagoli *et al* 1996; Min 1997) and that reconnection phenomena take place in this layer, leading to dynamical alignment between field and flow vectors (Jones *et al* 1997). It is suspected that the development of the kink MHD mode would affect the efficiency of the confinement by the hoop stress forces, having the effect of limiting the ratio of the toroidal to poloidal field to a value of a few (Eichler 1993). Exactly how much depends on the elastic current profile. This would induce effective dissipation, the energy that would be otherwise stored in the form of azimuthal fields being ultimately transformed into pressure which opposes efficient collimation by the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hoop stress. The poloidal field could then become, in the wind acceleration zone, the most significant agent for collimation which would then be ballistically conserved (Spruit *et al* 1997).

Bibliography

Appl S and Camenzind M 1992 *Astron. Astrophys.* **256** 354–70
Balbus S and Hawley J 1991 *Astrophys. J.* **376** 214–33
Bertout C 1989 *Ann.Rev. Astron. Astrophys.* **27** 351
Beskin V 1997 *Phys. Uspekhi* **40** (7) 659
Birkinshaw M 1991 *Beams and Jets in Astrophysics* ed P A Hughes (Cambridge: Cambridge University Press) p 279
Blandford R and Payne D 1982 *Mon. Not. R. Astron. Soc.* **199** 883–903
Camenzind M 1986 *Astron. Astrophys.* **156** 137–51
Chan K and Henriksen R 1980 *Astrophys. J.* **241** 534–51
Contopoulos J 1995 *Astron. Astrophys.* **450** 616–27
Dermer C and Schlickeiser R 1993 *Astrophys. J.* **416** 458–84
Eichler D 1993 *Astrophys. J.* **419** 111–6
Ferreira J and Pelletier G 1993 *Astron. Astrophys.* **276** 625–36
Goldreich P and Julian W 1970 *Astrophys. J.* **160** 971–7
Heyvaerts J and Norman C 1989, *Astrophys. J.* **347** 1055–81
Heyvaerts J and Norman C 1997 *IAU Symp 182* ed B Reypurth and C Bertout (Dordrecht: Kluwer) pp 275–90
Jones T, Gaalaas J, Ryu D and Frank A 1997 *Astrophys. J.* **482** 230–44
Königl A 1989 *Astrophys. J.* **342** 208–23
Li Z Y 1995 *Astrophys. J.* **444** 848–60
Livio M 1997 *Space Sci. Rev.* **82** 389–406
Malagoli A, Bodo G and Rosner R 1996 *Astrophys. J.* **456** 708–16
Markowith A, Henri G and Pelletier G 1995 *Mon. Not. R. Astron. Soc.* **277** 681– 99
Matsumoto R, Uchida Y, Hirose S, Shibata K, Hayashi M, Ferrari A, Bodo G and Norman C 1996 *Astrophys. J.* **461** 115–26
Mellema G and Frank A 1997 *Mon. Not. R. Astron. Soc.* **292** 795–807
Michel F C 1969 *Astrophys. J.* **158** 727–39
Min K W 1997 *Astrophys. J.* **482** 733–46
Mundt R 1985 *Protostars and Planets II* (Tucson, AZ: University of Arizona Press) pp 414–33
O'Dell S 1981 *Astrophys. J.* **243** L147–9
Okamoto I 1978 *Mon. Not. R. Astron. Soc.* **185** 69–108
Ouyed R and Pudritz R 1997 *Astrophys. J.* **484** 794–809
Parker E 1958 *Astrophys. J.* **128** 664–76
Sakurai T 1985, *Astron. Astrophys.* **152** 121–9
Shakura N and Sunyaev R 1973 *Astron. Astrophys.* **24** 337–55
Shibata K and Uchida Y 1986 *Publ. Astron. Soc. Jap.* **38** 631–60
Spruit H, Foglizzo T and Stehle R 1997 *Mon. Not. R. Astron. Soc.* **288** 333–42
Thomas B and Smith E 1981 *J. Geophys. Res.* **86** 11 105–10
Tsinganos K and Sauty C 1992 *Astron. Astrophys.* **255** 405–19
Ustyugova G V, Koldoba A V, Romanova M M, Chechetkin V M and Lovelace R 1995 *Astrophys. J.* **439** L39–42
Weber D and Davis L 1967 *Astrophys. J.* **148** 217–27

J Heyvaerts

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamics: Magnetic Reconnection and Turbulence

Magnetic reconnection is a distinctive dynamical phenomena by which neighboring parcels of magnetized plasma communicate with one another. It is thought to to be important in the dynamics of SOLAR FLARES, and in the magnetopause and tail regions of the magnetosphere (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE; GEOMAGNETIC TAIL). Reconnection is also thought to be an important element of low frequency plasma turbulence in the SOLAR WIND, the interstellar medium and in other astrophysical plasmas. Reconnection may influence heating, plasma flows and transport of charged particles including COSMIC RAYS.

Examples of MAGNETIC RECONNECTION are typically portrayed in highly idealized contexts. For example, special assumptions are made regarding the choice of a model for plasma dynamics, e.g. MAGNETOHYDRODYNAMICS (MHD), as well as imposition of steady-state conditions. Other assumptions are often made regarding geometrical symmetries, including reduction to a two-dimensional (2D) model, or assumption of reflectional symmetry in quadrants about the X-type NEUTRAL POINT around which reconnection activity is centered. All of these are perfectly reasonable special cases to be employed in demonstrations of the essential physics of reconnection. However, reconnection can also occur in situations that are unsteady in time, and that are much less ordered than in the simplest models. Fluctuations in MHD variables (see SOLAR WIND TURBULENCE) are likely to occur in relatively messy 'real' scenarios that involve reconnection, and can be expected to have significant effects on the reconnection process. Reconnection occurs readily as one facet of complex MHD activity, and it appears to be an essential feature of MHD turbulence.

Models and broad definition

The traditional 2D models of reconnection (Sweet–Parker and Petschek models) describe a set of dynamical phenomena that occur when blobs of plasma containing oppositely directed magnetic field $\boldsymbol{B}$ are either pressed together or encounter each other spontaneously. Since there is a region of low field strength between the blobs, there is an excess magnetic pressure outside the central region that drives plasma towards the boundary between the blobs, at a point determined by a perturbation. This pressure causes the plasma to flow toward the center with a speed $V_{\rm in}$ as illustrated in figure 1. If the plasma has high electrical conductivity, magnetic flux conservation implies that steep gradients of magnetic field are produced. In the standard models the process is assumed to be laminar and steady, in which case the approach of the blobs is limited by the rate at which magnetic flux can diffuse across a thin region, called the reconnection zone. This causes annihilation of matching contributions from positive and negative flux regions. This merging process relaxes the magnetic flux constraint and allows the field lines to 'break' and 'reconnect'. New magnetic islands form as plasma flows outward at a speed $V_{\rm out}$ towards the region between the blobs, sometimes called the weak field region. The flow pattern is illustrated in figure 1. The key features of the traditional laminar picture of reconnection are: (a) the approach of blobs containing opposite-signed field, (b) production of steep magnetic gradients, i.e. high electric current densities, between the blobs, (c) inflow limited by magnetic diffusivity, and (d) steady outflow at $V_{\rm out} = V_{\rm A}$ the Alfvén speed computed in the inflow or strong-field regions.

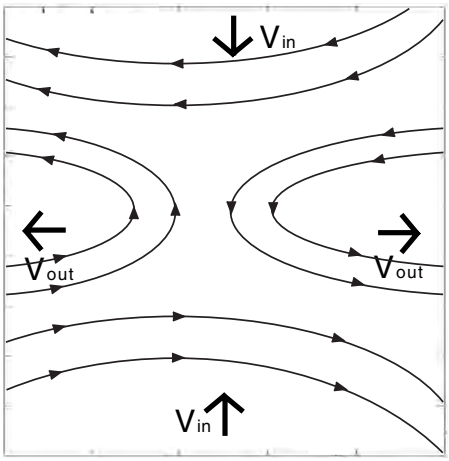


Figure 1. Diagram of two-dimensional magnetic fields in a reconnection configuration. The central reconnection zone contains an X-type neutral point. Inflow towards the reconnection zone at speed $V_{\rm in}$ is suggested by arrows in the strong field regions at the top and bottom. Outflow at $V_{\rm out}$ in weak field regions is also suggested.

Change in topology of field lines is an important by-product of the process. The rate at which topology change occurs, and the rate of Ohmic heating in the reconnection zone, is limited by the rate at which diffusivity can act to relax the limitations of flux conservation.

Most of the essential features of the traditional laminar reconnection picture remain useful in three dimensions, and when turbulent fluctuations are present. However, when reconnection is part of a complex MHD flow, carefully controlled smooth conditions far from the reconnection zone are not relevant any longer. The process may not commence from an equilibrium as is often assumed in reconnection models that invoke linear instabilities such as the tearing mode. In addition, there are many additional possibilities for types of encounters of plasma blobs in three dimensions, so a more general characterization of reconnection is needed. A widely accepted definition is that magnetic reconnection occurs in a plasma whenever an electric field $E_{\parallel}$ is present along a magnetic separator, that is, along a line of intersection of two separatrix surfaces that divide space into different magnetic cells. The separator is commonly referred to as the reconnection line, the X-line or the null line. This definition is useful for both laminar and turbulent models of reconnection.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

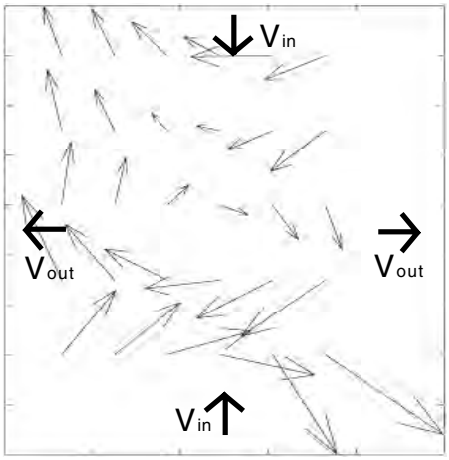


Figure 2. Map of magnetic field vectors near an active reconnection zone in the SSX experiment. Arrow length is proportional to local field strength, with longer arrows ≈1 kG. The Alfvén speed is $\approx 10^7$ cm s^{-1}. Although there is an overall X-type topology, there is evidence for multiple X-points (and at least one small island or 'bubble') near the reconnection zone. (Courtesy of M Brown and the SSX experimental team.)

Turbulent magnetic reconnection

The idealized constructs of smooth and steady magnetic islands undergoing laminar reconnection are easily extended to allow less symmetric and more active configurations. Figure 1 was intentionally constructed to be not quite fully symmetric. Simulations of MHD indicate that these types of magnetic configurations tend to experience complex undulations, involving non-steady induced electric fields that can be quite large compared with the $E_\parallel$ expected from laminar reconnection models. Fluctuations in the large-scale reconnection fields can vary in space and time in accord with the MHD equations, producing broad-band turbulence in and around the region of reconnection activity. In more complex cases in which reconnection activity is expected to occur, it may exhibit much more complex properties than in its laminar form. The reconnection geometry becomes non-steady and distorted.

Figure 2 shows an example of what is perhaps a more realistic reconnection geometry, using data taken from the SSX spheromak laboratory experiment at Swarthmore College. In SSX, two blobs of MHD scale plasma, or plasmoids, encounter each other under controlled, but not extremely symmetric or steady conditions. Experimental diagnostics indicate that reconnection occurs. Characteristic flows are detected, as suggested in figure 2, as well as changes in magnetic topology.

It may be quite a bit easier in astrophysical contexts for reconnection to arise in the presence of fluctuations than it would be in their absence. Distinctive effects that might be expected in a turbulent reconnection scenario include the possible production of multiple X-type neutral points in the reconnection zone, or the production of steep gradients and broadband fluctuations through a spectral cascade.

MHD turbulence

There is a strong similarity between MHD equations and the equations of hydrodynamics. Thus one expects that the complex dynamical motions associated with fluid turbulence at high Reynolds numbers should have a counterpart in the behavior of the fluid MHD plasma. In fact, observations of plasmas in the corona, the interstellar medium and the solar wind, as well as many laboratory plasma devices, bear out this expectation. MHD turbulence is also studied using accurate numerical simulation methods up to large-scale Reynolds numbers of a few thousand. A Reynolds number, usually defined as $R = LV_0/\nu$, with L and V_0 respectively the characteristic length and speed, and ν the kinematic viscosity, is a measure of the typical strength of nonlinear effects compared with dissipative effects. When R is large, complex nonlinear motions give rise to the irregular motions associated with turbulence, and a cascade of energy from large to small scales. In MHD the magnetic Reynolds number, $R_m = V_0 L/\eta$, for resistivity η, should be large as well, if the turbulence is to involve magnetic fluctuations as well as velocity fluctuations.

An interesting analogy exists between the dynamical behavior of the magnetic field in MHD and of the vorticity $\vec{\Omega} = \nabla \times \boldsymbol{v}$. The magnetic field obeys the induction equation. The equation for the evolution of the vorticity can be obtained formally from the induction equation under the replacements $\boldsymbol{B} \rightarrow \vec{\Omega}$ and $\eta \rightarrow \nu$. Thus the dynamics of the magnetic field in MHD are similar to the dynamics of the vorticity in ordinary hydrodynamics. The analogy is not perfect because the vorticity is related to the velocity $\boldsymbol{v}$ which appears in its dynamical equation, whereas the magnetic field is not related to the velocity $\boldsymbol{v}$ that appears in the right-hand side of its dynamical equation.

This analogy is particularly interesting from the point if view of magnetic reconnection, because it points towards an analogy with the phenomenon of 'vortex reconnection'. Vortex lines must 'break and reconnect' in order that vortices interact and merge. The vortex merger that results from this process is familiar in stirred cups of coffee, and in weather maps, and is associated with the production of large eddies from small ones. This occurs at the same time as smaller eddies are generated from large ones due to the cascade of energy to small scales.

MHD simulations indicate that reconnection, and associated magnetic flux tube merger or coalescence, are very typical in MHD turbulence. Figure 3 shows plots of magnetic field lines and intensity plots of electric current density for a simulation of two-dimensional MHD turbulence, shown at a time $t = 6$ characteristic times from the randomly generated initial state. It is apparent that of the many close magnetic islands in the top panel, many pairs are experiencing reconnection and merger. The current intensity shows the signature of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

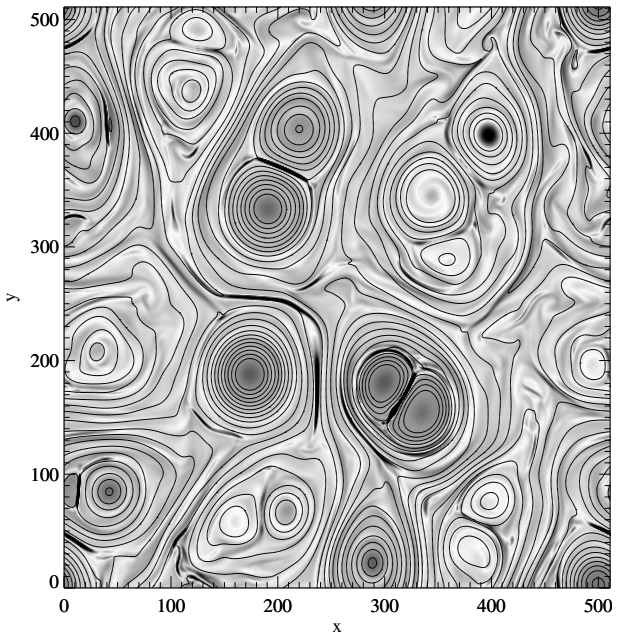


Figure 3. Magnetic field lines (contour lines) and electric current density (gray intensity scale) for a spectral method simulation of two-dimensional MHD turbulence at $R_m = 2000$, at time $t = 4$ nonlinear times from the initial state. Reconnection zones have formed between a number of adjacent islands that are coalescing, triggering localized non-steady reconnection activity throughout the periodic simulation box. Reconnection is an essential feature of the cascade process in MHD turbulence.

intense filamentation of current at X-type neutral points in the center of small, non-steady reconnection zones lying between the merging islands. Although not shown, the plasma velocity near these X-points also has the characteristic reconnection pattern of inflow and outflow near these regions, but the flow is very irregular and turbulent in general.

Turbulent dissipation and coherent structures

The effects associated with turbulence near a reconnection zone lead to a higher effective resistivity, called a turbulent resistivity or turbulent magnetic diffusivity. This is a property of MHD turbulence in general and is directly related to the analogous phenomenon of turbulent viscosity in hydrodynamic turbulence. (Note that turbulent resistivity may be supplemented by so-called *anomalous resistivity* associated with plasma instabilities of non-fluid, kinetic origin.) In general turbulence is expected to increase both the rate of reconnection and the rate of energy dissipation relative to estimates based upon laminar theories. Turbulence near an active reconnection zone also induces non-steady fluctuations in the measurable MHD variables, rendering more complex the plasma flow, electric field and magnetic field associated with reconnection.

As a consequence of turbulent reconnection, energy is directly cascaded to small scales, producing turbulent heating through a direct cascade. At the same time, larger magnetic field structures are produced through merger of magnetic islands, sometimes called an inverse cascade. The transfer of a significant amount of magnetic excitation to large scales is sometimes discussed in terms of negative temperature statistical mechanics, or the phenomenon of 'self-organization.' The characteristic structures of MHD reconnection, magnetic islands at the larger scales, and electric current filaments at small scales, can be thought of as preferred MHD coherent structures, and are possible sources of intermittency in plasma dynamics at MHD fluid scales.

Bibliography

Biskamp D 1993 *Nonlinear Magnetohydrodynamics* (New York: Cambridge University Press)

Frisch U 1995 *Turbulence* (New York: Cambridge University Press)

Matthaeus W and Lamkin S 1986 Turbulent magnetic reconnection *Phys. Fluids* **29** 2513

Parker E 1979 *Cosmical Magnetic Fields* (Oxford: Clarendon)

William Matthaeus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetohydrodynamics: Magnetoconvection

Convective motions occur naturally in layers of fluid heated from below. In the Sun turbulent convection carries most of the luminosity over the outer 30% by radius of the star. Because of the high electrical conductivity of the solar plasma convective motions interact with magnetic fields. The modified form of convection that occurs in a conducting fluid when externally imposed magnetic fields are present is referred to as magnetoconvection. The interaction of convection and magnetic fields is believed to be responsible for such diverse phenomena as the darkness of SUNSPOTS, the oscillation of SUNSPOT PENUMBRAE, the formation of the magnetic network (see CHROMOSPHERE: NETWORK), and the amplification of magnetic fields in SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES.

The nature of the interaction between convective motions and magnetic fields depends on several physical parameters. Of particular importance is the strength of magnetic forces relative to the strength of the convective driving due to buoyancy. If magnetic forces are strong the convection can be greatly inhibited or in extreme cases even suppressed. If, on the other hand, magnetic forces are weak, the magnetic field can be strongly distorted by the motions. It is customary to organize the study of magnetoconvection into different regimes defined by the vigour of the resulting motions. The *linear regime* studies the conditions under which buoyancy forces just overcome magnetic forces and the convection develops to infinitesimal amplitude. In the *weakly nonlinear regime* the amplitude of the convection is assumed to be small but still of sufficient magnitude to lead to appreciable nonlinear effects. In the *fully nonlinear regime* the motions are vigorous, possibly even turbulent, and there is a strong interaction between the convection and magnetic fields.

The linear regime

The conditions under which infinitesimal convection develops can be simply estimated by considering the forces acting on a fluid parcel displaced vertically a small distance δz from its equilibrium position in a superadiabatic atmosphere. A small fluid element will experience a destabilizing buoyancy force of magnitude $g\rho\alpha\Delta\delta z$, where g is the acceleration due to gravity, ρ the fluid density, α the coefficient of thermal expansion and Δ the local temperature gradient. If the displacement causes a distortion of wavelength ℓ in an otherwise uniform ambient magnetic field, there will be a magnetic restoring force of magnitude $B^2\delta z/\mu_0\ell^2$, where B is the field intensity and μ_0 is the magnetic permeability. If the buoyancy force exceeds the restoring force the fluid element once displaced will continue to move and a state of overturning convection will develop. If, on the other hand, the magnetic restoring force is stronger than buoyancy the parcel will eventually return to its initial equilibrium position. Both buoyancy and magnetic forces are weakened by diffusion at rates proportional to κ, the thermal diffusivity, and η, the magnetic diffusivity. If $\kappa > \eta$, buoyancy forces will be weakened more rapidly than magnetic forces and the fluid parcel will return to its equilibrium position with a greater velocity than that with which it left it, thus causing the fluid element to overshoot. In this case the convection develops as a growing oscillation. This form of instability, often called *overstability*, is characteristic of a broad class of systems—*doubly diffusive systems*—where stabilizing and restoring agents diffuse at different rates.

Some of these ideas can be made more quantitative by considering the development of convection in a layer of Boussinesq fluid (Chandrasekhar 1961). The strength of the destabilizing buoyancy force and the stabilizing magnetic force can measured by the *Rayleigh* number R and the *Chandrasekhar* number Q defined by

$$R = \frac{g\alpha\Delta d^4}{\kappa\nu} \quad \text{and} \quad Q = \frac{B_0^2 d^2}{\mu_0\rho\eta\nu} \tag{1}$$

where ν is the kinematic viscosity and d is the vertical extent of the layer. Overturning convection develops when the Rayleigh number exceed a critical value $R^{(e)}$ given by

$$R^{(e)} = R_0 + C_1 Q \tag{2}$$

where R_0 is the critical Rayleigh number for the onset of convection in the absence of magnetic effects and C_1 is a constant that depends on boundary conditions and on the geometry of the system. As we mentioned earlier, if $\eta < \kappa$ overstability is also possible provided the magnetic field is strong enough. The critical Rayleigh number for the onset of growing oscillations has the form

$$R^{(o)} = R_0(1 + C_2) + C_3 Q \tag{3}$$

where the constants C_2 and C_3 depend as before on boundary conditions and geometry, but now also on the diffusivities. In most cases $R^{(o)} \leq R^{(e)}$ so that overstability, if possible, is the first type of instability to develop in a fluid layer as the temperature gradient is gradually increased.

The weakly nonlinear regime

The weakly nonlinear regime describes magnetoconvection when the amplitude of the motions is small but finite. The existence of both overturning and oscillatory forms of convection gives rise to a rich and complex set of possible dynamical behaviors. The primary objective of weakly nonlinear theory is to classify these possibilities in terms of *amplitude equations* and their associated *bifurcation diagrams* (see, for instance, Guckenheimer and Holmes 1983).

The simplest type of amplitude equation describes the behavior of small amplitude solutions when the *bifurcation parameter*, for instance the Rayleigh number, is near its critical value for the onset of overturning convection. The dynamical behavior for typical initial conditions can be described by

$$\dot{a} = \varepsilon a + \gamma a^3 \tag{4}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

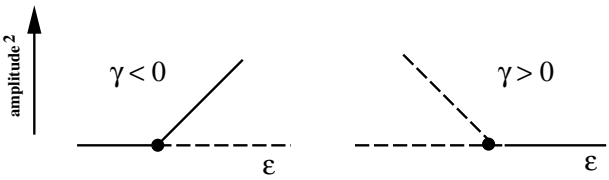


Figure 1. Bifurcation diagram describing the behavior of small amplitude solutions. The ordinate measures the (square) of the amplitude of the solutions, the abscissa the magnitude of the bifurcation parameter. Solid and dashed lines correspond to stable and unstable solutions respectively. $\varepsilon = 0$ is a bifurcation point where branches of solutions are created or destroyed.

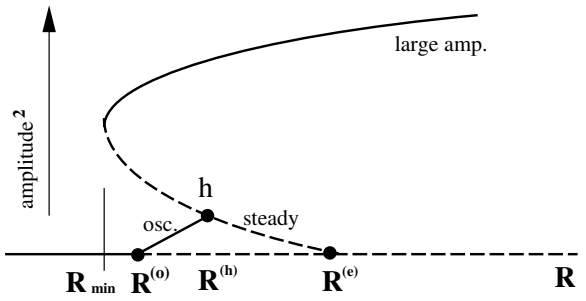


Figure 2. Bifurcation diagram describing the relationship between various branches of solutions. Solid and dashed lines correspond to stable and unstable solutions respectively.

where $\varepsilon = (R - R^{(e)})/R^{(e)}$. With this notation $\varepsilon = 0$ corresponds to a bifurcation point where branches of solutions are either created or destroyed. Here γ depends on the physical parameters specifying the problem and determines the stability of these solutions. If $\gamma < 0$ small amplitude solutions are *stable* and weak initial conditions develop into convective motions with a characteristic amplitude of $\sqrt{\varepsilon/|\gamma|}$. If $\gamma > 0$ small amplitude solutions are *unstable* and weak initial conditions develop to large amplitude. These two cases are represented schematically by the two bifurcation diagrams in figure 1.

More complicated diagrams can be used to represent more complex situations. For instance, the interplay between overturning and oscillatory convection is described by the diagram in figure 2. In this case the static solution (no motion) becomes unstable to oscillatory convection at $R = R^{(o)}$. A branch of unstable overturning convection is destroyed at $R = R^{(e)}$ so that no small amplitude solutions exist for $R > R^{(e)}$. For $R_{\min} \le R \le R^{(o)}$ convection is possible at large but not at small amplitude. As R approaches $R^{(h)}$ the period of the oscillations increases, eventually to become infinite at $R = R^{(h)}$ where the branch of oscillatory solutions collides with the unstable branch of overturning convection. In the vicinity of such points, called *homoclinic points*, solutions can become aperiodic and even chaotic. Other complex dynamical behavior is possible in (horizontally) unbounded domains where the possibility of oscillatory convection leads to solutions in the form of traveling, modulated and pulsating waves.

All these different dynamical possibilities can be described by amplitude equations analogous to, but in general more complex than, (4) (see, for instance, Proctor and Weiss 1982). Formally such equations are only valid when the amplitude of the motion is small, but in many circumstances the predictions of weakly nonlinear theory appear to remain qualitatively correct even when the solutions grow to larger amplitudes. Thus weakly nonlinear theory is often used as a starting point for the systematic study of magnetoconvection in the fully nonlinear regime.

The fully nonlinear regime

The study of the interaction of large amplitude convection with an externally imposed magnetic field divides naturally into two approaches, *kinematic* and *dynamical*. Kinematic theory can be regarded as the limiting case of magnetoconvection for very weak fields. In this limit the Lorentz force can be neglected and the momentum equation decouples from the induction equation. It is then possible to study the induction equation alone for prescribed velocity fields. In the dynamical regime the Lorentz force is important and the induction and momentum equations must be solved simultaneously. Because of the difficulties in obtaining analytical solutions in this regime, dynamical theory has particularly benefited from numerical work.

The most important parameter in kinematic theory is the *magnetic Reynolds number*, $R_{\rm m} = Ud/\eta$, where U is a characteristic amplitude of the velocity. It measures the relative efficiencies of convective to diffusive processes, i.e. the time it takes to diffuse the magnetic field over a distance d divided by the time it takes to *advect* the magnetic field over the same distance. The magnetic Reynolds number is huge in most astrophysical circumstances.

The simplest kinematic problem considers the structure of a unidirectional field near a stagnation point flow. If the velocity is axisymmetric and given in cylindrical coordinates (r, z) by the stream function $\Phi(r, z) = Ur^2 z$ then the steady-state magnetic field has a Gaussian profile of the form

$$B_z(r) = B^* \exp(-\tfrac{1}{2} R_{\rm m} r^2) \qquad B^* = \tfrac{1}{2} B_0 R_{\rm m} \qquad (5)$$

where B_0 is the amplitude of an initial uniform field. This shows that the steady state is determined by a balance between advection and diffusion which occurs when the magnetic field is concentrated on structures with characteristic scale of order $R_{\rm m}^{-1/2}$. More realistic calculations show that in general the magnetic field is amplified by straining motions and expelled from regions of closed circulation.

The above example shows that within kinematic theory magnetic fields can be amplified arbitrarily provided $R_{\rm m}$ is arbitrarily large. In a realistic situation one expects an eventual breakdown of the kinematic assumption and a modification of the velocity due to the action of the Lorentz force. Of direct physical interest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

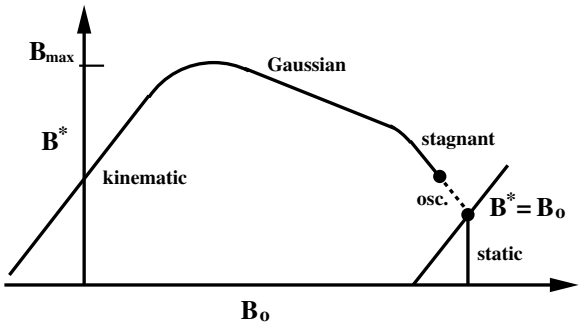


Figure 3. Limits of kinematic amplification. Peak field B^* as a function of initial filed B_0. After Galloway and Moore (1979).

is the determination of the strongest field that can be obtained by convective amplification. In particular, how it compares with an equipartition field, i.e. one whose energy density is comparable to the kinetic energy density. For very weak initial fields the above example indicates that the maximum field strength $B^* \approx B_0 R_\mathrm{m}$ for a nearly axisymmetric configuration. For very strong initial fields linear theory predicts that the convection becomes suppressed, in which case $B^* = B_0$. It is reasonable to assume that the maximum value is obtained for some intermediate value of B_0.

We envisage a sequence of experiments where convection at a fixed Rayleigh number develops in the presence of an initially uniform field of strength B_0. In each experiment B^* is the maximum field strength obtained in the final steady or statistically stationary state. The behavior of B^* as a function of B_0 is shown schematically in figure 3. For weak initial fields B^* increases linearly with B_0, the slope of the line being precisely R_m. Eventually the curve turns over when the vorticity generated by magnetic curvature forces becomes comparable to the vorticity generated by horizontal temperature gradients. At this point B^* attains its maximum value B_max given by

$$B^2_\mathrm{max} \approx \frac{U^2}{\log(R_\mathrm{m})}(\nu/\eta). \qquad (6)$$

To the right of the peak B^* decreases slowly with B_0 and the magnetic field profile remains approximately Gaussian. For stronger initial fields the convection is no longer capable of confining the magnetic field and B^* decreases more rapidly with B_0. If B_0 is increased further there is a transition to oscillatory convection terminating with the eventual suppression of the motions. Clearly the above considerations suggest that convective motions can indeed amplify magnetic fields to values exceeding the equipartition strength U provided that the ratio ν/η is sufficiently large.

If compressibility becomes important the above arguments must be modified. In particular, another critical field strength B_ev must be considered corresponding to a magnetic field whose pressure is comparable to the thermodynamic pressure of the ambient fluid. The significance of B_ev can be appreciated by noting that an isolated flux tube of this strength in thermal and mechanical equilibrium with its surroundings must be *evacuated*. There is substantial numerical evidence that partially evacuated flux structures can indeed be generated by convective processes in strongly stratified layers. It is as yet unclear whether the extreme case of a completely empty tube can be maintained by convection.

Bibliography

Chandrasekhar S 1961 *Hydrodynamic and Hydromagnetic Stability* (Oxford: Oxford University Press)

Galloway D J and Moore D R 1979 Axisymmetric convection in the presence of a magnetic field *Geophys. Astrophys. Fluid Dyn.* **12** 73–105

Guckenheimer J and Holmes P 1983 *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (New York: Springer)

Proctor M R E and Weiss N O 1982 Magnetoconvection *Rep. Prog. Phys.* **45** 1317–79

Fausto Cattaneo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Bow Shock

The bow shock (figure 1) is the nonlinear wave which stands in the SOLAR WIND flow upstream of the Earth's magnetosphere, at which the solar wind plasma is heated and decelerated in preparation for diversion around the magnetosphere. The dissipation processes at the shock depend on the properties of a collisionless plasma and lead to a rich range of energetic particles and plasma waves.

Why and where?

In an ordinary gas, a shock wave forms when an obstacle is placed in a flow which is supersonic, i.e. the relative speed between flow and obstacle is greater than the sound speed. For example, a supersonic aircraft has an associated shock wave. In a PLASMA, such as the solar wind, the situation is complicated by the existence of other wave modes in addition to sound waves. However, the principle remains true, and the solar wind flow speed is greater than the sound speed and the Alfvén speed, so that a plasma shock is formed, owing to the obstacle created by the MAGNETOSPHERE OF EARTH.

Shock waves can arise in a number of different situations in astrophysics. For example, an explosive event may generate supersonic flows which impact the surrounding gas, and this leads to a blast shock wave, where there is a limited amount of energy associated with the shock (e.g. SUPERNOVA REMNANT or SOLAR FLARE). In the case of the magnetized planets in the solar system, the magnetosphere presents an essentially impermeable obstacle (see PLANETARY MAGNETOSPHERES), and a steady shock wave is formed which stands at a relatively constant distance from the planet in the upstream flow. This is called a bow shock, in analogy to the bow wave in water ahead of a ship.

What is a shock wave? It is a wave through which the plasma flows, and in doing so it increases its characteristic wave speed (e.g. sound speed). This is accomplished by an increase in the temperature, and a corresponding decrease in the flow speed, so that the sound speed increases relative to the flow speed. The shock wave is thus the transition from supersonic flow upstream to subsonic flow downstream. The shock acts to transform some of the kinetic energy of the upstream flow into an increase in the thermal energy of the downstream flow. There is no return to the initial upstream state of the flow, so there also has to be an increase in the entropy of the flow at the shock (i.e. dissipation) so that the changes there are irreversible.

Now consider a supersonic flow. By definition sound waves produced by a source are immediately swept downstream; they cannot propagate upstream against the flow. An obstacle in a supersonic flow, without a bow shock, could never disturb the upstream flow so that it is diverted around the obstacle. The necessary deflection must be accomplished by pressure-like forces, which are communicated into the upstream flow to warn, as it were, the incoming plasma that there is an obstacle ahead. However, such forces are the same as the restoring forces that lead to the characteristic wave modes of the plasma. For example, in sound waves the restoring force is provided by the particle pressure, and hence the pressure force can only be propagated at the speed of sound. Thus, in order for the obstacle to affect the flow upstream, there must be a surrounding layer of subsonic flow (the magnetosheath), and it is this that the bow shock provides.

At the bow shock the magnetic field, density and temperature all increase as the solar wind transits the shock. This is similar to what happens in a fast magnetosonic wave (i.e. fast-mode MHD wave). Thus the bow shock is an example of a fast-mode collisionless shock, and the governing characteristic wave speed is that of the fast magnetosonic wave.

In an ordinary gas the dissipation necessary at a shock is provided by collisions between the gas molecules, so that the width of the shock is associated with the dissipational scale length which is of the order of the collision mean free path. In space plasmas the mean free path between collisions is so large that the system is essentially collisionless. Consequently the dissipation is supplied by plasma processes related to the gradients at the shock, particularly changes in the electric and magnetic field. The plasma shock processes consequently govern the width and internal structure of the shock layer. The Earth's bow shock has a width of between roughly 100 km and 1000 km, depending on the shock and plasma parameters.

The bow shock surrounds the magnetosphere on its upstream side, and is about $15R_E$ from the Earth at the 'nose' of bow shock, i.e. at the subsolar point. However, the position of the shock is highly variable (from $12R_E$ to $20R_E$) and depends on the solar wind parameters, most importantly the ram pressure. The shock position changes on the time scale of minutes in response to changes in the solar wind, and the motion of the shock can be in the range 10–100 km s^{-1}. One important consequence is that observations of the bow shock are usually due to the shock moving over a spacecraft, rather than to the motion of the spacecraft around its orbit. The motion of the shock allows a translation from a time series of measurements to a cross-section through the shock giving its spatial structure. This can obviously become impossible if the shock is unsteady.

The overall shape of the bow shock has been investigated by collating large sets of shock crossings determined from measurements of the magnetic field. Most models used to fit the observations are cylindrically symmetric conic sections (i.e. the surface generated by rotating a parabola or hyperbola around its symmetry axis), with suitable scaling to take into account variation of the solar wind ram pressure. Locating the bow shock at any given time is not yet a precise operation, since the magnetospheric obstacle is not rigid, but also responds to changes in the solar wind pressure and magnetic field orientation.

The bow shock is important for different reasons. The solar wind flow is processed by the shock before it hits the magnetosphere, so the coupling between solar wind and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

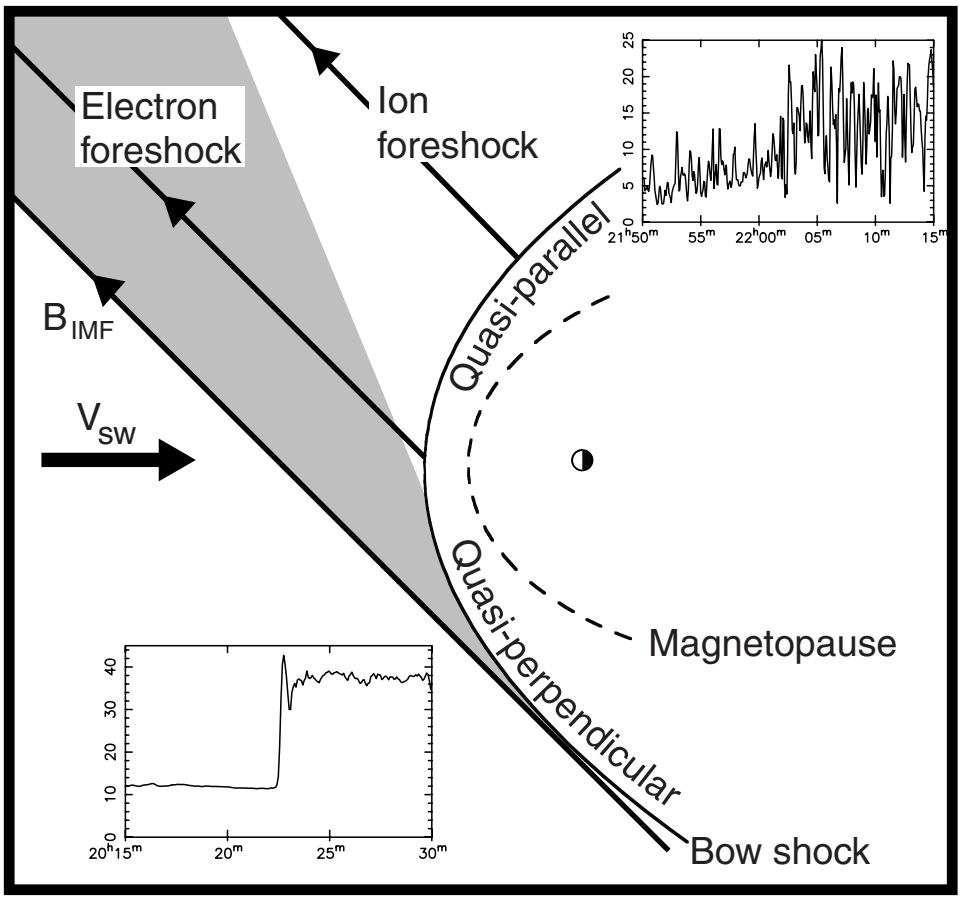


Figure 1. A schematic view of the Earth's bow shock in the equatorial plane. The direction of the upstream solar wind flow is indicated by V_{SW}. The upstream IMF direction for typical solar wind conditions is shown by three representative field lines, including the tangent field line which defines the upstream edge of the foreshock. For clarity the field lines are shown ending at the bow shock; in reality they pass into the magnetosheath and drape around the magnetosphere. The regions of quasi-perpendicular and quasi-parallel shock regimes are indicated. Examples of the corresponding different types of shock crossing are shown as time series of the magnetic field magnitude in nT (data from ISEE spacecraft—see www.igpp.ucla.edu). The typical zones of the electron and ion foreshocks are shown upstream of the bow shock.

magnetosphere is mediated by the shock and its downstream region (the magnetosheath). More importantly the bow shock is an accessible example of an astrophysical shock in a collisionless plasma. Such shocks are believed to be common in a wide range of astrophysical situations, and the Earth's bow shock allows us to study in detail shock processes such as dissipation, wave generation and particle acceleration. A spacecraft crossing the shock takes between seconds and minutes, and in that time instrumentation can record the electric and magnetic fields and particle distribution functions at high resolution. This has allowed an analysis of the plasma processes producing the necessary dissipation at the shock and how these processes, and hence the shock structure, are controlled by variations in the solar wind parameters. Observations of the bow shock show that it is a rich source of waves and energetic particles, and *in situ* observations have provided something like a plasma laboratory, so that much of the physics of collisionless shocks has been deciphered.

Different kinds of shock

From the earliest spacecraft observations it became clear that there are different types of collisionless shock, and that these correspond to different solar wind parameters (see SOLAR WIND SHOCK WAVES AND DISCONTINUITIES). These parameters include the speed, density and temperature, but also, and most importantly, the direction of the upstream interplanetary magnetic field (IMF). Observed crossings of the bow shock are generally divided into two broad classes, depending on the so-called shock normal angle θ_{Bn}, between the upstream magnetic field and the normal vector to the shock surface. If the magnetic field upstream of the shock is more nearly perpendicular to the shock normal (i.e. parallel to the shock surface) then the shock is called quasi-perpendicular. If the field is more parallel to the shock normal, then the shock is termed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

quasi-parallel. Usually the dividing line between these two classes is taken as 45°, so that a quasi-perpendicular shock has $\theta_{Bn} > 45°$, and a quasi-parallel shock has $\theta_{Bn} < 45°$. The dividing line is slightly arbitrary, it is sometimes difficult to accurately assign a value of θ_{Bn} in the case of a quasi-parallel shock. There are other divisions to be made when categorizing shock crossings. The Mach number (the ratio of the upstream flow speed normal to the shock to the characteristic wave speed) also plays an important role. For typical solar wind conditions the bow shock has a high Mach number (e.g. an Alfvén Mach number 5–10).

Quasi-perpendicular and quasi-parallel shocks have very different structures, as is evident from, for example, profiles of the magnetic field through a shock crossing. Quasi-perpendicular shocks have a single sharp gradient, relatively smooth upstream flow, although with waves in the downstream region caused by the shock. Such shocks most closely resemble the laminar flow discontinuity that would be expected at an ordinary gas shock. Quasi-parallel shocks are much more convoluted: the shock transition is embedded in large amplitude waves both upstream and downstream, and it is often very difficult to assign an exact moment for the shock crossing. The shock crossing appears much thicker than in the quasi-perpendicular shock, but even within the extended quasi-parallel shock transition one can find abrupt changes with sharp gradients similar to the quasi-perpendicular shock. The presence of waves in the solar wind upstream of the quasi-parallel shock can make it difficult to assign an accurate value for the θ_{Bn} of the shock, since the waves are convected into the shock by the flow and produce an ever-changing shock normal angle.

As discussed above, the bow shock forms a curved surface shrouding the Earth's magnetosphere, and the interplanetary field, over the scale length of the bow shock, is typically uniform. Consequently, the shock normal angle varies over the surface of the bow shock, and the locations of quasi-perpendicular and quasi-parallel portions of the bow shock are controlled by the direction of the IMF. Thus, observationally, as well as rapid traversals of the shock because of its motion in response to changing solar wind ram pressure, the type of the shock crossing can change because of changes in the IMF direction.

The fundamental differences in structure between the major shock classes can be explained by the effect of the magnetic field geometry on the motion of particles (mainly protons) at, and immediately upstream of, the shock. The two important components of particle motion are motion along and gyration perpendicular to the field direction. At a high Mach number shock, such as the bow shock, the magnetic field increases by a factor of about four, from upstream to downstream, although peaks at the shock ramp itself can be greater. The field jump at such shocks can produce reflection of some of the protons in the incoming flow. This process depends completely on the kinetic properties of forces on individual protons, and so can never be included properly in a fluid model of a plasma shock. Ion reflection seems to be a ubiquitous property of high Mach number shocks, irrespective of their θ_{Bn} value. Ion reflection is observed using particle instruments which can measure not just particle density but also the particle velocity space by simultaneously recording particle energy and arrival direction.

The ion reflection process occurs at the location of high gradient in the field and is approximately specular, i.e. the component of velocity normal to the shock is reversed. The resultant velocity has components away from the shock along the magnetic field direction, and also, importantly, perpendicular to the field. The latter produces gyrational motion with a gyroradius of the order of 500 km, which is greater than the thickness of the shock ramp.

At a quasi-perpendicular shock the field lines are near parallel to the shock surface, so that motion along the field does not lead to rapid escape from the shock. At the same time the gyromotion about the field forces the ions to return to the shock ramp, where, on their second encounter, they are likely to pass into the downstream region. On the other hand, at a quasi-parallel shock the field-aligned motion can help guide any reflected ions away from the shock ramp before their gyration brings them back. Similar arguments show that particle motion across the shock, from downstream to upstream against the flow, is much more likely to occur at quasi-parallel shocks than at quasi-perpendicular shocks.

These fundamentally different behaviors, re-encounter with the shock or escape from it, lead to different (collisionless) heating mechanisms at the different types of shock. At the quasi-perpendicular shock the reflected-gyrating ions which re-encounter the shock pass into the downstream region where they contribute most of the ion heating. In effect, the reflection process provides a direct conversion, for a small fraction (∼15%) of the incoming ions, from upstream kinetic (flow) energy to downstream thermal (gyrating) energy. As described this process is not true dissipation, but the large temperature anisotropy of the gyrating ions makes the plasma unstable to a number of different wave modes (see MAGNETOSPHERE OF EARTH: WAVES). These waves scatter and isotropize the ions leading to true dissipation.

The foreshock

Before we consider thermalization at the quasi-parallel shock, there is a unique feature of collisionless shocks to be described. The argument concerning the θ_{Bn} control of escape from the shock actually depends on the velocity of the particles concerned. If a particle has sufficiently high velocity at even the quasi-perpendicular shock, i.e. if its motion away from the shock along the field is faster than the speed at which the field line is convected into the shock, then it will be able to escape into the region upstream of the shock. By the various mechanisms of shock acceleration the bow shock does produce both energetic electrons and ions, and with sufficient velocities to stream ahead of the shock. The region upstream of the shock populated by energetic particles and their related waves, is termed the foreshock. Because the energetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

populations have unstable velocity space distributions (e.g. beam like) they can drive a wide range of plasma waves from electromagnetic waves at radio frequencies (80 kHz), through electrostatic Langmuir waves (40 kHz), down to low-frequency MHD-like waves (50 mHz).

The foreshock only exists because of the collisionless nature of the solar wind, and the various wave types and energetic particles present there can only be explained using kinetic, as opposed to fluid, theories (see SOLAR WIND: KINETIC PROPERTIES). Furthermore, the foreshock is vital to understanding several basic phenomena at the bow shock itself, such as particle acceleration and, in the case of the quasi-parallel shock, even the shock structure. It can be seen that in some ways the existence of the foreshock undermines our discussion of the shock location as the limit of an obstacle's influence in the upstream flow. The velocities of the energetic particles are greater than the wave speeds of the plasma. On the other hand, the associated energy densities within the foreshock are small enough that the solar wind flow is only slightly affected. Indeed, if the foreshock were to strongly affect the incoming flow, then the position and nature of the bow shock would be changed drastically, which would feedback to the characteristics of the shock.

The foreshock has definite regions within it characterized by different energetic particle populations and associated waves (see SOLAR WIND: ENERGETIC PARTICLES). The upstream edge of the foreshock is the surface everywhere parallel to the solar wind magnetic field and tangential to the bow shock surface. The foreshock does not extend indefinitely upstream, since particles are eventually scattered and coupled to the solar wind flow. The foreshock global morphology is organized by the IMF direction, the bow shock shape, and, crucially, particle cross-field drift relative to particle parallel velocity. Particles energized at the shock with high speeds will tend to more closely follow a field line parallel to the tangent field line. This leads to a velocity filter which organizes the foreshock spatially, so that one can distinguish an electron foreshock and an ion foreshock. In each case one expects, and sees, the fastest beams occurring on the upstream edge. Electron distributions vary from energetic beams (up to 100 keV) near the upstream edge to a backstreaming heat flux deeper in the foreshock. Ion distributions vary from field-aligned beams (with temperatures similar to that of the solar wind), through 'intermediate', to 'diffuse' ions with near-isotropic velocity space distributions, with energies up to several hundred keV.

The coupling between the shock and foreshock is vital to understanding how the quasi-parallel shock manages to heat the incoming flow. Out of this understanding has developed what is known as a 'cyclic reformation' model. Because of the quasi-parallel field geometry, and lack of collisions, particles with moderate energies can easily stream away in front of the shock. Once there they can excite plasma instabilities, creating a region of interacting waves and particles. Since the flow is, by definition, faster than the most important wave speeds, the majority of waves in the foreshock must eventually be convected by the flow into, and through, the shock itself. Based on simulations and observations, high Mach number quasi-parallel shocks have a highly dynamic structure which alternates between a sharp transition in the field (and correspondingly the thermalization) and a longer, more gradual transition. This alternation cycles in time and is linked to the convection of upstream waves (generated by foreshock ions) into the shock layer, which modifies the local shock normal angle. Furthermore, periodic bursts of reflected ions (with $\sim$20% the incident density) are observed within the shock structure. These reflected ions appear to be specularly reflected out of the incident distribution. It is the interaction of the foreshock waves, as they are swept into the shock, with the reflected ions which forms one part of the reformation cycle mechanism. This model can be taken further by admitting that that the upstream waves, which control the shock structure, are only coherent over a certain scale length. Consequently, it had been suggested that the idea of a single shock surface, in the quasi-parallel regime, has to be abandoned, and that different parts of the bow shock are at different stages in the cycle from abrupt to extended transition.

D Burgess

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Convection

Magnetospheric convection is the process whereby ions and electrons are transported throughout the *magnetosphere*, the region of space dominated by the Earth's magnetic field (see MAGNETOSPHERE OF EARTH). Throughout much of the magnetosphere, the ions and electrons travel together with the same velocity perpendicular to the magnetic field; this is called the *convection velocity* $\boldsymbol{u}$ and is given numerically by $\boldsymbol{u}_E = [\boldsymbol{E} \times \boldsymbol{B}]/B^2$, where the local magnetic and electric fields are $\boldsymbol{B}$ and $\boldsymbol{E}$, respectively. In the absence of time variations and significant magnetic-field-aligned electric fields, magnetospheric convection patterns can be deduced from electric-field patterns in the ionosphere. At high energies and in regions of large field gradients, however, particles also exhibit energy- and charge-dependent drifts, such that they move at field-perpendicular speeds significantly different from the local convection velocity.

Particle motions

Drift velocities

The motion of a charged particle in magnetic and electric fields can be resolved into several components, with fundamental time constants relevant to each. The magnetic, or *Lorentz*, force is given by $F = q\boldsymbol{v} \times \boldsymbol{B}$, where the particle velocity perpendicular to the field line is $v_\perp$ and the particle charge (including sign) is given by q. This force results in a *gyromotion* of particles around the field line: the particle moves in circles around the field line. The particle's velocity parallel to the field line $v_\parallel$ is unchanged by the Lorentz force, so that the net motion is a helix around the field line. The angular frequency of this circular motion is called the *gyrofrequency* ω_c given by qB/m, where m is the particle mass. The radius of curvature of this gyromotion is called the *gyroradius* and is given by $v_\perp/\omega_c$, where the velocity perpendicular to the field line is $v_\perp$. Electrons move in a right-hand sense around the field direction, and ions in a left-hand sense. This is typically the fastest temporal scale and shortest spatial scale for plasma motion, and, so long as the electric and magnetic fields do not change on these scales, leads to a conservation of particle *magnetic moment* $mv_\perp^2/2B$.

Because of the strength of the Lorentz force, any other external force perpendicular to the magnetic field does not result in a free acceleration along that force direction, but instead results in a particle drift u_d perpendicular to both the magnetic field and the external force: $\boldsymbol{u}_d = (\boldsymbol{F}_{ext} \times \boldsymbol{B})/qB^2$. This drift can be visualized to result from changes in the gyroradius of the particle during one cycle (figure 1(top)). During the part of the cycle where the particle is moving along the force, its perpendicular velocity and therefore its gyroradius is increasing; during the other half of the cycle, the particle moves opposite to the external force and its velocity is decreasing. This leads to a cycloidal motion that is the sum of the simple circular motion plus a steady drift given by u_d. The center of curvature of the circular motion is called the *guiding center*. Thus the particle motion can be thought of as the sum of the circular gyration plus the drift motion of the guiding center. For the case where the external force is from a large-scale electric field $\boldsymbol{E}$, the drift velocity is given by $\boldsymbol{u}_E = [\boldsymbol{E} \times \boldsymbol{B}]/B^2$.

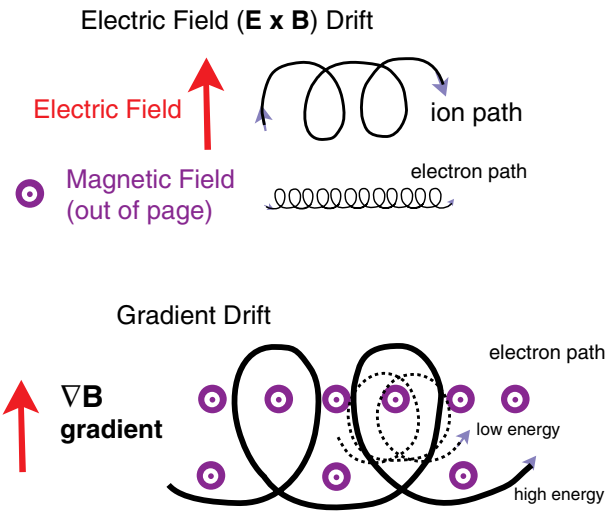


Figure 1. Top: drifts from external forces result in a drift perpendicular to the magnetic field and the force. In the case of the electric field, the drift is an $E \times B$ *drift*. Bottom: gradients in the field magnitude also result in particle drift, with the fastest drift for higher-energy particles.

Since the electric-field drift is neither mass, energy, nor charge dependent, it represents the drift of the entire low-energy plasma population and is called the *convection velocity* or $E \times B$ *drift*. A *flux tube* is a set of field lines which cross a particular area. Consider the set of particles which all have their guiding centers on the field line in the center of a flux tube. Since all of these particles drift together, their guiding centers permanently remain on the same field line as each other. Therefore, you can imagine the entire magnetic flux tube and its associated particles moving as a unit. This is the idea of *frozen-in-flux*, which was developed by HANNES ALFVÉN. This approximation breaks down if the magnetic field changes topology (as by *reconnection*, see MAGNETIC RECONNECTION) or if electric fields along the magnetic field accelerate particles along the field line, but it is a useful way to visualize the large-scale motion of low-energy plasma.

If there is no electric field aligned along the magnetic field ($E_\parallel$), and if the fields are steady state, then each magnetic-field line is an equipotential. Thus perpendicular electric fields at high altitudes are simply mapped to electric fields at low altitudes and the motion of the plasma at the high-altitude end of the field line is reflected in the motion of the plasma in the ionosphere. Each flow path is an equipotential in steady-state conditions.

The second periodic motion of a particle trapped in the Earth's magnetic field is its *bounce motion* from one

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

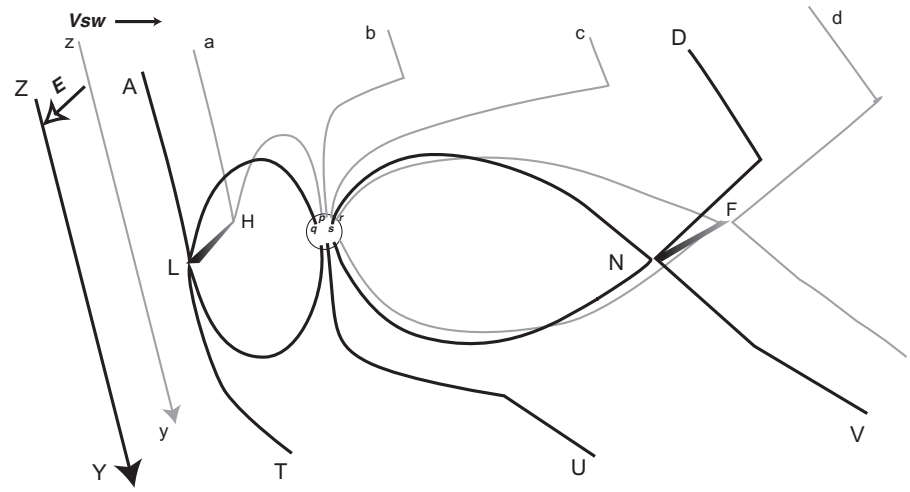


Figure 2. Schematic of magnetic reconnection. The Sun is to the left. Two southward-pointing solar-wind field lines separated in the Y direction have a potential drop across them, with line zy (gray, in background) at a higher potential than ZY (in foreground). When those field lines connect to the Earth's field at the dayside X-line (HL), the potential drop aA is mapped to HL and from thence down to the ionosphere (pq). This electric field sets the ionosphere in motion. As the solar wind convects from a to b to c to d, the ionospheric feet of those field lines move antisunward as well. The open fields reconnect at the tail X-line (FN), which maps to the ionosphere as line rs. The newly closed field lines convect back sunward (not shown) to complete the cycle.

hemisphere to the other. The particle, preserving its first adiabatic invariant (the *magnetic moment* $mv_\perp^2/2B$) and its total energy $mv^2/2$, increases its *pitch angle* $\alpha(=\sin^{-1} v_\perp/v)$ as it travels towards the Earth from the equatorial plane. When its pitch angle reaches 90°, the particle's parallel velocity goes to zero then reverses. The particle *mirrors* and returns to the other hemisphere. The integral of the parallel velocity along a bounce path leads to the second adiabatic invariant, which is also conserved if the fields change slowly compared to the bounce period.

Energy-dependent drifts

Higher-energy particles have energy-dependent drifts in addition to their convection velocity. Gradient drift is a result of a change in magnetic-field strength over the course of one gyration (figure 1(bottom)), in a manner similar to $E \times B$ drift (figure 1(top)). The effect is to move high-energy particles along a contour of constant field strength. Similarly, if the particles move along a curved field line, the centrifugal force also leads to a particle drift. In the Earth's equatorial plane, both gradient and curvature drifts lead to westward drift of ions and eastward drift of electrons. The drifts are given by $u_g = (mv_\perp^2 \boldsymbol{B}\times\nabla\boldsymbol{B})/2qB^3$ for the gradient drift, and $u_c = -(mv_\parallel^2 \boldsymbol{R} \times \boldsymbol{B})/R^2 qB^2$ (where $\boldsymbol{R}$ is the radius of curvature of the field line) for the curvature drift. These charge-dependent drifts lead to a westward current circling the Earth, the *ring current* (see MAGNETOSPHERE OF EARTH: RING CURRENT). Since the drift velocity depends on the energy of the particles, it can be shown that the intensity of the ring current is proportional to the energy density of the particles in the ring current. Since the gradient drift is perpendicular to the field gradient, energetic particles trapped in the equatorial plane move along contours of constant B. The drift cycle around the Earth leads to a third adiabatic invariant, which is less frequently conserved (because it requires field constancy on the timescale of a drift period, which is minutes to hours, depending on the energy of the particle and its distance from Earth). Again, the motion of the guiding center is a useful concept for following particle trajectories. However, because of the larger gyroradii of energetic particles, a particle can be lost from the magnetosphere even if its guiding center remains inside.

Steady-state convection

Corotation

Near the Earth, the low-energy plasma motion is dominated by *corotation*. Collisional drag between the ionosphere and the neutral atmosphere sets up currents which create an electric field which brings the less massive magnetospheric plasma into flow equilibrium with the neutral atmosphere. Mapping the electric field from the ionosphere to the magnetosphere then results in the inner-magnetosphere low-energy plasma (the *plasmasphere*) drifting around the Earth counterclockwise once per 24 h (see MAGNETOSPHERE OF EARTH: PLASMASPHERE). In rotation-dominated magnetospheres such as Jupiter's, partial corotation is observed essentially throughout the magnetosphere, although significant slippage occurs in the outer magnetosphere, and near the orbit of the satellite Io where the ionization rate is large (see JUPITER: MAGNETOSPHERE). In Earth's magnetosphere, however, corotation only dominates close to the Earth (within about 4–6 R_E). The convection electric field from corotation is given by $E_{cor} = -(\boldsymbol{\Omega} \times \boldsymbol{r}) \times \boldsymbol{B}$, where $\boldsymbol{\Omega}$ is the angular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

frequency of the Earth and $\boldsymbol{r}$ the radius vector to the location in question.

Solar-wind-driven convection

Although the motion of cold plasma trapped near the Earth is dominated by corotation, the motion of the bulk of the magnetosphere is determined by its large-scale interaction with the SOLAR WIND. The Earth's magnetic field has a defined outer boundary, the *magnetopause* (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). The solar wind flowing past the magnetosphere sets up a convective flow inside the magnetosphere, so that the regions of the magnetosphere just inside the boundary flow antisunward (i.e. in the same direction as the neighboring solar wind), and the plasmas deeper in the magnetosphere have a return sunward flow.

This antisunward motion can either be driven by quasi-viscous forces at the magnetopause (e.g. diffusion of particles across the boundary or momentum transfer by large-amplitude Kelvin–Helmholtz surface waves), or by magnetic reconnection. Magnetic reconnection is the process whereby plasma flows across a boundary between regimes of different magnetic topology. In the case of the Earth, dayside (northward pointing) closed field lines become connected to southward-pointing interplanetary magnetic field lines, with a transfer of flux across the boundary (figure 2). The mapping of the solar-wind electric field throughout the magnetosphere leads to a large-scale convection pattern with a dawn-to-dusk electric field (figure 3). The reconnection model has had spectacular successes in predicting the variation of the convection speed and direction as a function of interplanetary conditions (see below), and is considered the dominant form of solar-wind interaction. MHD models of the magnetosphere, with their numerical resistivity, naturally yield both forms of interaction process.

Combined flows

The simplest models of convection merely added the externally driven convection electric field vectorially to the corotation electric field. However, it can be shown that, in steady conditions, electric charges on the inner edge of the plasma sheet shield the inner magnetosphere from the solar-wind-driven convection. In times of changing convection, however, this external convection electric field penetrates more easily. In times of increasing convection, the outer parts of the plasmasphere can be stripped off and convect to the dayside magnetopause, creating tongues of plasma. The combined flow is shown in the bottom panel of figure 3.

Convection for southward IMF

This transport of plasma across the magnetopause boundary in an open geometry requires an electric field along the x line. The integral of $\boldsymbol{E}$ along the x line, sometimes called the dayside reconnection rate R_D, represents the rate at which closed dayside flux becomes

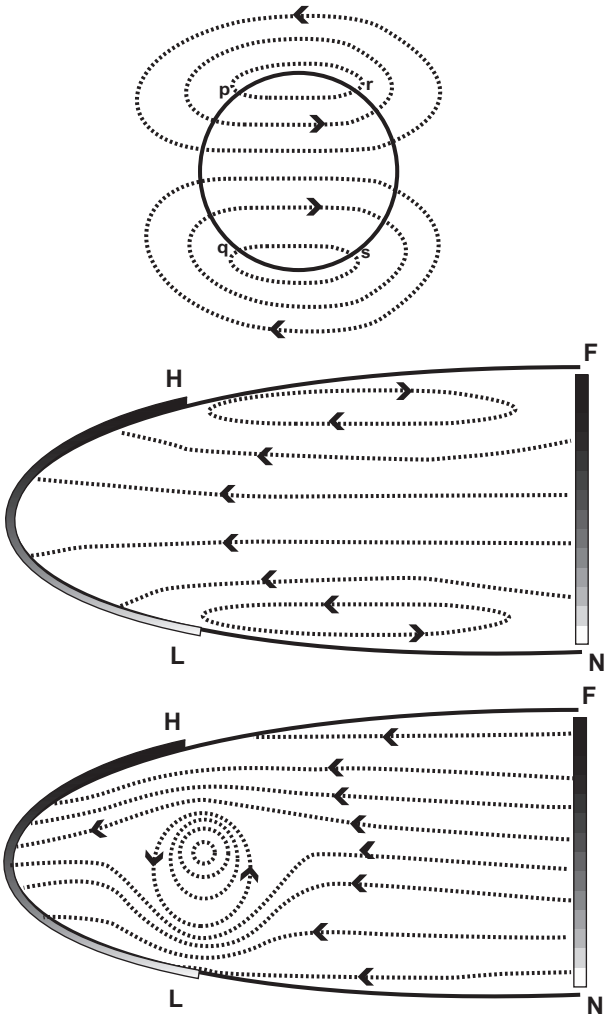


Figure 3. Schematic of convection in the ionosphere (top) and the corresponding flow paths in the equatorial plane (middle). The flow paths intersect the magnetopause at the dayside (HL) and nightside (FN) x-lines, which map to the ionosphere at pq and rs, respectively. A small convection cell from viscosity is shown on the edges of the middle panel. Adding in the corotation electric field and the plasma sheet shielding yields the bottom convection pattern.

'open', or interconnected to the interplanetary magnetic field. The dayside x line is shown as HL in figures 2 and 3, with H at the higher potential.

Similarly, the flux transfer rate across the nightside portion of the x line, denoted by R_N, represents the rate at which open magnetic flux is closed. This x line is shown as FN in figures 2 and 3. Over a long time average, $\langle R_D \rangle$ must equal $\langle RN \rangle$, but the two instantaneous rates frequently differ. When $R_D > R_N$, the amount of open flux increases and thus the size of the polar cap (the footprint of the open field line region in the ionosphere) increases. Magnetic energy stored in the tail likewise increases. When $R_N > R_D$, open flux in the magnetotail closes and the polar cap size

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

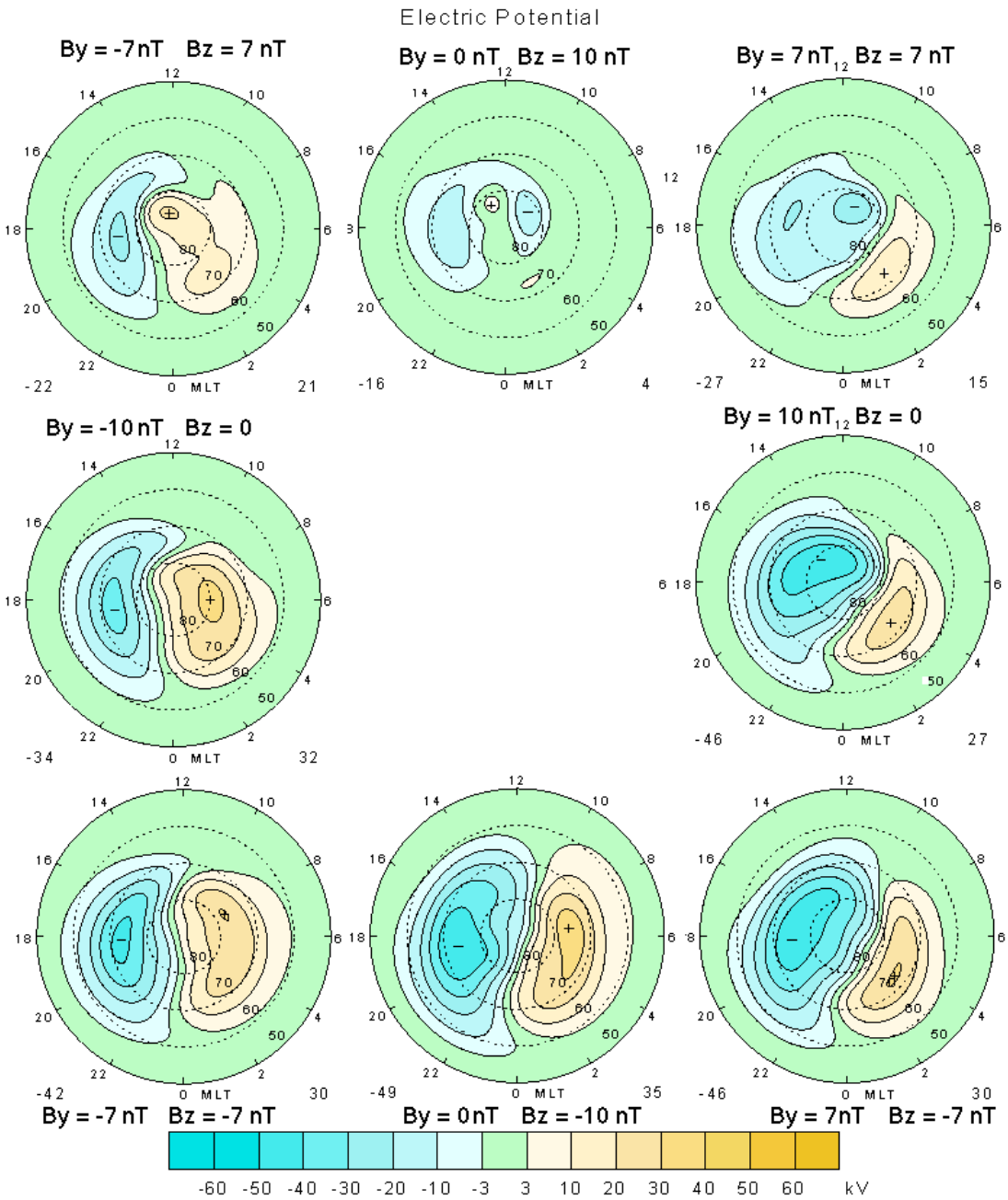


Figure 4. Ionospheric convection patterns for various orientations of the IMF (adapted from Weimer D R 1996 *Geophys. Res. Lett.* **23** 2549–52). In each case the view is from above the Earth's north magnetic pole and the Sun is towards the top. The minimum and maximum potential for each case is shown at the bottom.

decreases. The magnetic energy is released in a *substorm* (see MAGNETOSPHERE OF EARTH: SUBSTORMS). In a steady state, $R_D = R_N = \Phi_{pc}$, the electric potential drop across the antisunward flowing part of the ionosphere. This polar cap potential drop Φ_{pc} thus characterizes the global rate of convection, both in the ionosphere and in the magnetosphere. Polar cap potentials range from a minimum near 20 kV in quiet times to 200 kV, or more in very active times, with a median value near 50 kV.

In the ionosphere, the flow lines (which are equipotentials) form closed contours and a convection pattern can be discerned, with generally antisunward flow on the field lines inside the polar cap (connected to the antisunward-flowing solar wind), with sunward return

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

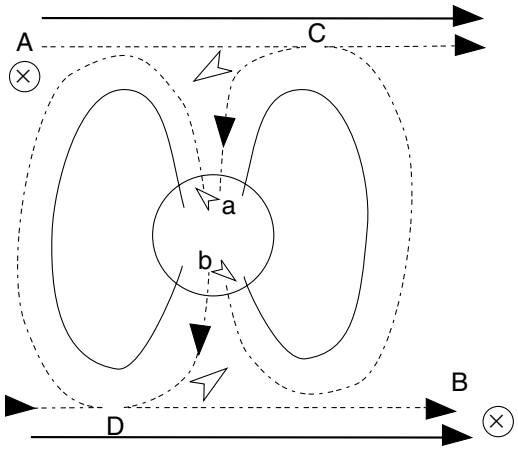


Figure 5. Schematic of reconnection of the Earth's field (light lines) with an IMF with a positive B_y component (heavy lines; view from the sun). The merging line (not shown) runs from C to D across the dayside. Newly interconnected field lines (shown dashed) are dragged tailward by the solar wind. Tension in the field lines (A–a) and (B–b) causes the flow to cross noon from dawn to dusk in the northern hemisphere and from dusk to dawn in the southern hemisphere (open arrows).

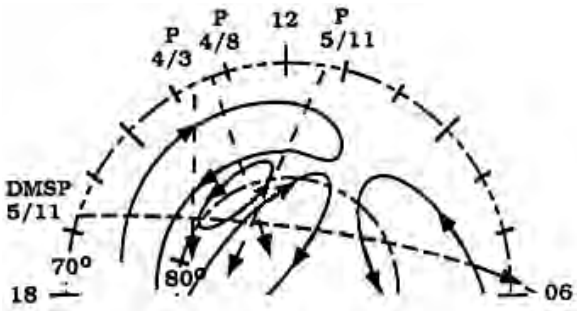


Figure 6. Four-cell convection pattern frequently observed for strongly northward IMF. In this case four spacecraft passes confirmed the location and orientation of the convection cells. (Adapted from Maynard N C *et al* 1998 *J. Geophys. Res.* **103** 29–46).

flow on closed field lines at lower latitudes (figure 3). The integral of the electric field around the polar cap boundary must (by Faraday's law) equal the negative time change of the magnetic flux enclosed. Since the Earth's magnetic field is constant, then the integral of the electric field around the boundary in the Earth's frame of reference must be zero and the convection pattern will look like closed flow contours, even in times when the merging rates differs between the dayside and the nightside. This flow pattern changes in strength and skew with the magnitude and orientation of the *interplanetary magnetic field* (IMF). The larger the southward (negative B_z) component of the IMF, the larger the polar cap potential drop. One typically uses a Geocentric Solar Magnetospheric (GSM) coordinate system for the IMF. The $+x_{GSM}$ axis is defined as pointing from the Earth to the Sun; the $+z$ axis points generally towards the ecliptic north, but is tilted so that the Earth's dipole is in the x–z_{GSM} plane; and the y_{GSM} axis completes a right-hand system, with positive y pointing generally opposite to the Earth's orbital motion.

The very strong dependence of the potential on the IMF B_z is the most compelling evidence for magnetic merging being the dominant form of momentum transfer from the solar wind to the magnetosphere. Empirical studies also indicate a background potential drop (5–15 keV) that is not related to B_z but does appear to depend on the solar-wind velocity squared. This may indicate a small diffusive or turbulent momentum transfer across the magnetopause, or it may just indicate that the best functional form for the merging momentum transfer has not yet been found.

B_y effects

The effect of the y component (dawn–dusk) of the IMF is to skew the flow pattern from dawn to dusk, with the flow asymmetry opposite in opposite hemispheres. In the northern hemisphere, for example, for IMF $B_y > 0$, the dayside flow in the polar cap crosses noon from the dusk side to the dawn side, making the antisunward flow strongest near the dawn polar cap boundary. Just below the polar cap boundary, in the closed field regions, the flow is eastward.

For $B_y < 0$ (or for $B_y > 0$ in the southern hemisphere), the flow is in the opposite sense: towards dusk below the polar cap boundary, and towards dawn above it. Convection patterns for various orientations of the IMF are shown in figure 4. These patterns, from fits to low-altitude electric-field measurements, are similar to those compiled from ionospheric flow measurements or ionospheric magnetic-field perturbations. However, being fitted functions to many passes of spacecraft data, they have smaller gradients than are typically observed in a single spacecraft pass.

The dawn–dusk asymmetry of the flow can be best explained in terms of the field tension resulting from interconnection with the IMF (figure 5). The dayside x-line is tilted along the current direction so that the component of the IMF along the x-line is constant as the magnetopause is traversed. Since the reconnected field line has one end in the duskside magnetosheath and its footpoint in the dawn ionosphere, the foot is dragged duskward.

Most published convection patterns (as in figure 4) are given with the Earth's rotation subtracted. This leads to a less-cluttered plot, since the equipotentials at low latitudes are dominated by corotation. The plasma sheet shielding effect (see above) means that convection patterns with corotation subtracted generally do not extend below about 55° invariant latitude. The lack of convection below 50° is useful in calibrating electric field and velocity measurements, which can suffer minor offsets from modest pointing errors.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

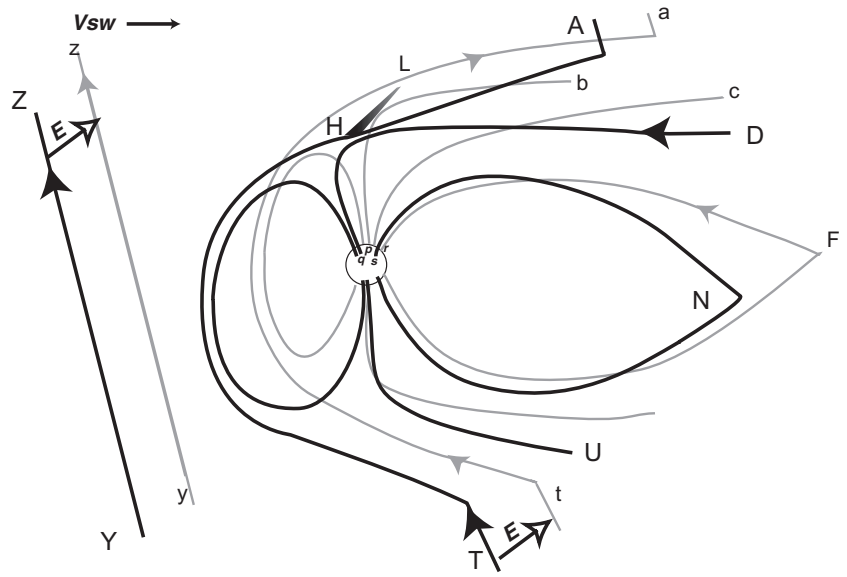


Figure 7. Reconnection of northward-pointing IMF field lines (YZ and yz) with the tail field lines tailward of the cusp. The magnetosheath flow causes the IMF to bend, having a tendency to follow the curvature of the magnetopause (TA and ta). Dipole tilt or the x-component of the IMF will favor reconnection in one hemisphere or another (here shown in the north). After reconnection at the x-line FN, dusk to dawn electric field in the magnetosphere gets mapped down to the ionosphere, creating sunward flow on the dayside. Tension then swings the flow to the dawn or dusk, depending on the sign of B_y. (Gray field lines are farther away; black field lines are closer).

Convection for northward IMF

In times of northward IMF, the polar cap convection is more complicated. Since the dayside reconnection rate R_D decreases, the polar cap has a tendency to shrink. Furthermore, since less magnetic flux connects to the solar wind, the electric potential across the polar cap also decreases. Effects of the y component of the IMF (see above) become more pronounced.

The most dramatic difference from southward IMF conditions is the appearance within the polar cap of substantial regions of sunward flow. If the IMF B_y is large, the polar cap convection can exhibit a single large convection cell with, e.g., sunward flow on the dawn side and antisunward flow on the dusk side, as occurs for B_y large and negative in the northern hemisphere (see figure 4). When the northward component of the IMF is larger than the B_y component, the convection can show a four-cell pattern: there is a central pair of cells with sunward flow in the central part of the polar cap and antisunward flow around the edges, which is similar to but opposite in sense to the normal convection pattern, and lower latitude cells of the normal sense. Although the driving mechanism for this kind of convection is still controversial, the existence of the pattern is well established, both by multiple spacecraft passes (e.g. figure 6) or from arrays of velocity measurements from ground-based radar.

The driving mechanism for the higher-latitude reverse cell is considered to be 'reverse reconnection': connection of high-latitude field lines with the IMF tailward of the magnetospheric cusps (figure 7) (see also MAGNETOSPHERE OF EARTH: DAYSIDE CUSP). Reconnection can occur in just one hemisphere (as is shown here), or in more unlikely circumstances, can occur in both hemispheres simultaneously, leading to the creation of newly closed magnetic flux. In addition, there exists a lower-latitude pair of cells in the forward sense (antisunward at higher latitudes) which persists even when the higher-latitude cells reverse direction. The driving mechanism for these lower-latitude cells could either be small-scale magnetopause reconnection, closure of open field lines (in cases where $R_N > R_D$), or quasi-viscous processes at the magnetopause.

Time-dependent convection

Increasing reconnection

The difference between the dayside and nightside merging rate results in an expansion (if $R_D > R_N$) or contraction (if $R_D < R_N$) of the open field region, while maintaining closed flow contours in the ionosphere. The merging rate is the integral of the electric field along the x-line in the *moving* frame of reference. The difference between the electric field in the x-line's frame of reference and in the Earth's frame of reference is just $\boldsymbol{v}_{\rm pc} \times \boldsymbol{B}$, where $\boldsymbol{v}_{\rm pc}$ is the speed of the polar cap boundary in the Earth's frame of reference.

The dayside convection pattern is observed to change very rapidly in response to changes in the IMF, with the information propagating to the ground at approximately the Alfvén speed (a few minutes from the magnetopause). The changes in the dayside flow get communicated to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the nightside portion of the ionosphere more slowly (5–10 min), but still significantly faster than the convection time from the dayside to the nightside along the magnetopause (30–60 min).

During times of increasing reconnection, the shielding currents do not have time to be set up and strong sunward flows can be observed at relatively low latitudes. It is in these times that plasmas in the outer portions of the plasmasphere can be stripped away.

Bursty bulk flows

Nightside reconnection results in newly-closed field lines which then shorten and move earthward. The flows in the magnetotail are observed to be unsteady, however, with short ($\sim$ 1 min) bursts of rapid flow separated by several minutes of slow flow. The sunward flow in the magnetotail may be choked off because, as the flux tube travels earthward, its volume decreases and the plasma pressure inside increases. This leads to a back pressure retarding the flow.

Decreasing reconnection

After a northward turning of the IMF, the dayside reconnection rate is smaller than the nightside rate. The flow across the nightside *X*-line need not stop immediately, however, and the polar cap boundary retreats poleward. Open magnetotail flux becomes closed. Northward turnings of the IMF have frequently been associated with the expansion onset of *substorm*, but the cause of that association has not been proven.

If the convection is decreasing, the pre-existing shielding currents are too strong for the new convection pattern, and an overshielding condition can occur. This can result in significant antisunward flows at low latitudes until a new steady state is established. This effect (plus the effect of continuing nightside reconnection) may be a significant driver of the lower-latitude cells observed during northward IMF conditions (see figure 7).

Summary

The convection of plasmas in the Earth's magnetosphere is becoming better understood, both from a theoretical standpoint and from *in situ* measurements. However, the magnetosphere is rarely steady in space or in time, and the effects of rapid time variations on the merging rate are still not well understood. By using multiple spacecraft measurements, coordinating with arrays of velocity measurements from ground-based radars, we can begin to elucidate better the changes in the convection in response to both changes in the IMF and from internal changes such as substorms.

Patricia Reiff

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Dayside Cusp

The magnetospheric cusps are the dayside, polar regions of the magnetosphere that have the first direct magnetic connection to interplanetary space as the SOLAR WIND flows past the magnetosphere. This direct magnetic connection allows solar wind and magnetosheath particles and energy to enter the magnetosphere, beginning the complex coupling between interplanetary space and Earth's environment. Because the solar wind that is continuously flowing toward Earth is highly variable, the cusps are dynamic regions, changing abruptly as the driving forces in the solar wind change. The magnetospheric cusps provide an important window to interplanetary space, and a natural laboratory to investigate the coupling of mass, momentum and energy from the flowing solar wind into Earth's magnetosphere, ionosphere and upper atmosphere.

The magnetospheric cusps extend from the dayside magnetopause into the northern and the southern polar regions of Earth (figure 1; see also figure 1(*a*) of MAGNETOSPHERE OF EARTH). The location of the cusps and their properties depend strongly on the interaction occurring on the magnetopause between the flowing magnetosheath on its outer surface and the magnetosphere that is confined within. The dominant process that facilitates the linkage across the magnetopause is MAGNETIC RECONNECTION. If magnetic reconnection were not occurring on the magnetopause, i.e. if the magnetosphere were 'closed', the magnetospheric magnetic field would lie tangent to the magnetopause everywhere along its inner surface. The magnetic field over this entire surface would trace to a vanishingly small region at the magnetic poles. Consequently, any plasma that crossed the magnetopause and flowed into the magnetosphere along the magnetic field would be focused into these narrow regions at the northern and the southern poles. The occurrence of magnetic reconnection on the magnetopause causes the magnetic field to cross the magnetopause over much of its surface, leading to a much broader volume in the polar regions into which the magnetosheath plasma will flow. The rate of reconnection and its location on the magnetopause will determine the energy transfer and the breadth of the polar regions into which magnetosheath plasma will penetrate, as well as the properties of this penetrating plasma.

Much of what we know about the cusp has been obtained from instruments on orbiting satellites, on sounding rockets and from ground instrumentation. Satellites and rockets are used to measure directly the magnetosheath and magnetospheric plasma and fields, while ground instruments measure the effects of the precipitating plasma and of the interconnection with the solar wind on the ionosphere and upper atmosphere. The observed low-altitude effects include dayside aurora, acceleration of the ionosphere and neutral atmosphere and heating of ionospheric plasma that then flows out and populates the magnetosphere.

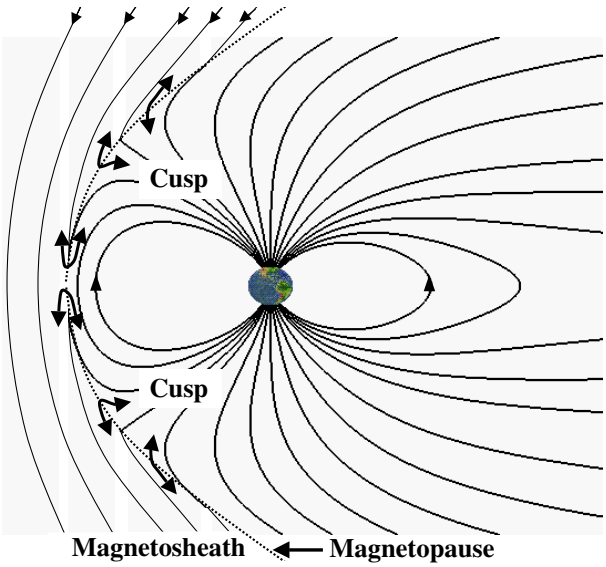


Figure 1. Location of the cusps in the magnetosphere. The cusps extend from the high-latitude ionosphere and upper atmosphere out to the dayside magnetopause. In the case when Earth's magnetic field is interconnected with interplanetary space, the magnetosheath and magnetospheric plasma freely mix across the open magnetopause (arrows crossing magnetopause).

The properties of the cusp are described here by considering the case in which the interplanetary magnetic field (IMF) is directed southward. Under these conditions, the solar wind drives the magnetospheric and ionospheric plasma mainly in the anti-Sunward direction, and the dominant source of cusp plasma is the low-latitude dayside magnetosheath. For different orientations of the solar wind magnetic field, the magnetosphere and ionosphere are pulled in different directions, and the location where the cusp is connected to the magnetopause varies. Since the IMF is typically quite variable, the cusp regions at any given time can be rather complex, reflecting a history of competing forces and magnetosheath source properties.

When the IMF is southward and reconnection occurs at low latitudes on the equatorial magnetopause, the Earth's high-latitude magnetic field becomes interconnected with the low-latitude magnetosheath (figure 1). These interconnected field lines extend from Earth out to the high-speed solar wind in interplanetary space. As a consequence of the interconnection of the field lines across the magnetopause, the magnetosheath plasma flows freely into the magnetosphere, and the magnetospheric plasma escapes out into the magnetosheath and eventually to interplanetary space. This interchange of plasma across the magnetopause is illustrated with the arrows in figure 1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observing the cusp

Inside the magnetosphere, the entering magnetosheath plasma and the escaping magnetospheric plasma can be clearly identified. This is due in part to the different densities and temperatures of the plasma on either side of the magnetopause, and in part to the different mass composition in the two regions. The plasma in the outer dayside magnetosphere typically has a low density (<1 cm^{-3}) and a high temperature ($\sim 10^8$ K). In the dayside magnetosheath, the density is considerably higher (1–100 cm^{-3}), and the temperature is lower (10^5–10^7 K). Consequently, when a satellite moving from low to high latitudes within the magnetosphere crosses the boundary between open and closed field lines, it will detect the entering magnetosheath plasma at the lower energies, below about 1 keV for the electrons and below about 10 keV for the ions. There will also be an abrupt decrease in the higher-energy magnetospheric plasma above these approximate energies. The plasma mass composition is also used to distinguish plasma of direct solar wind origin from plasma of ionospheric origin. Mass composition experiments detect the presence of helium nuclei (alpha particles) in the cusp with density of a few per cent of the hydrogen density. The relative abundance observed in the cusp matches well the relative abundance in the solar wind and magnetosheath, but is considerably higher than the relative abundance in the closed magnetic field portion of the outer magnetosphere.

An example of a crossing through the cusp made by NASA's POLAR satellite is shown in figure 2, obtained from the Hydra Electron and Ion Hot Plasma Instrument. A segment from the orbit of Polar is shown on the right of the figure. Over this 3 h time period, the satellite moved from low latitudes to high latitudes, remaining near local noon. At the end of this time period, Polar was located roughly over the geomagnetic pole. The gray intensity indicates electron (upper panel) and ion (lower panel) count rates as functions of particle energy for energies ranging from a few eV to about 20 keV. The ions measured here are primarily protons, with a small contribution coming from other higher-mass species. Dark regions indicate low count rates, and light regions indicate high count rates. The locations and energies where the count rates are below a threshold level of the instrument are indicated with white.

The transition from closed to open magnetic field lines can be clearly seen in figure 2 at approximately 16:48 UT. Prior to this time, Polar detected high-energy electrons and ions on the closed magnetospheric field lines, up to the highest energies measured. When the boundary between open and closed field lines was crossed, the dense, lower-energy electrons and ions that originated in the magnetosheath were abruptly observed, and the highest-energy particles were seen to no longer be present. The cusp is the region in the magnetosphere, observed on this crossing between approximately 16:48 UT and 18:00 UT, where the magnetosheath plasma has entered across the dayside magnetopause and filled a funnel-shaped region of the magnetosphere that extends down to the ionosphere and upper atmosphere.

Spatial and temporal variability of the cusp

The cusp, which is defined primarily by the density and average energy of its plasma, is observed to be localized in latitude and in longitude, and exhibits spatial and temporal structure. The spatial structure is primarily due to the large-scale magnetospheric convection and to the large-scale spatial characteristics of the magnetosheath and of the magnetopause. The temporal variability of the cusp is due primarily to temporal changes in the rate and location of magnetic reconnection on the dayside magnetopause.

One of the most consistently observed spatial features of the cusp is the variation of the average energy of the plasma with latitude. As can be seen in figure 2, the cusp ions are concentrated at high energies near the boundary between open and closed field lines (at lower latitudes), and the average energy decreases as the spacecraft moves to higher latitudes. This dependence of particle energy on latitude arises from the evolution of the plasma following magnetic reconnection. The spatial dependence of the cusp plasma depends on three main effects: (a) the travel time required for the entering magnetosheath plasma to penetrate into the magnetosphere after reconnection occurs; (b) the changes in the magnetosheath plasma as it flows over the dayside magnetopause; and (c) the spatial dependence of plasma acceleration that occurs when plasma crosses the magnetopause current sheet.

The large-scale magnetospheric convection driven by reconnection on the dayside magnetopause carries the newly interconnected magnetic field and plasma to higher latitudes over the polar regions and eventually deep into the magnetotail. This flow is directed perpendicular to the magnetic field, and carries all plasma along at a common convection speed. As time progresses, the magnetopause crossing point of a given magnetic field line moves to higher latitudes and eventually far down the tail.

When reconnection first occurs, the magnetosheath electrons and ions begin to flow into the magnetosphere across the magnetopause. As the particles travel along the magnetic field toward the ionosphere, the large-scale convection carries the plasma and the magnetic field to higher latitudes. Consequently, the particles that have very high velocities parallel to the magnetic field will quickly penetrate deep into the magnetosphere prior to appreciable convection to higher latitudes. Particles with low parallel velocities will penetrate more slowly into the magnetosphere, and therefore will be convected to higher latitudes in the time required for them to penetrate a given distance into the magnetosphere.

Therefore, at the low-latitude edge of the cusp, i.e. on magnetic field lines that reconnected only a brief time ago, the only magnetosheath particles that can be observed are those that have a sufficiently high parallel velocity to reach the observation locations in the time since reconnection of the field line occurred. As an observer moves to higher latitudes, i.e. onto magnetic field lines that reconnected progressively longer ago, magnetosheath plasma with progressively lower energies will be observed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

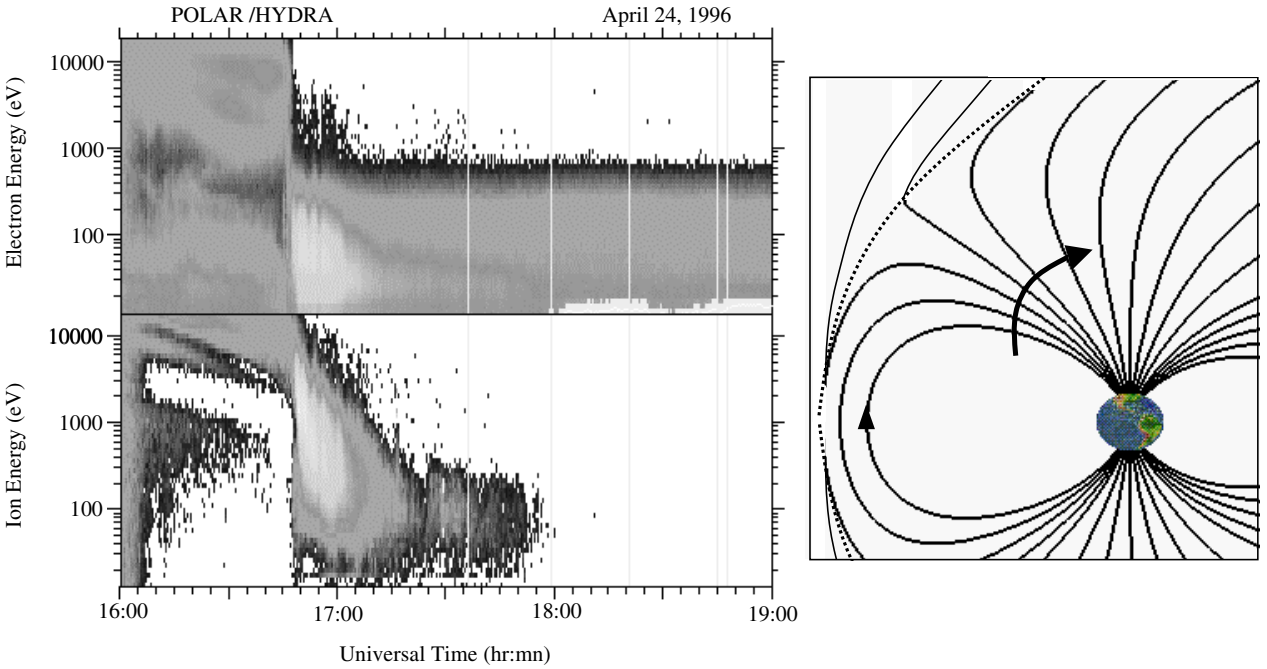


Figure 2. Segment of the orbit of the NASA Polar satellite (right) and electron (upper) and ion (lower) measurements obtained from the Hydra instrument on board the Polar satellite (left). Dark shades indicates low particle fluxes, light shades indicate high particle fluxes, and the white areas indicate where the particle fluxes are below a threshold level. The cusp is the region of intense, localized particle fluxes observed at the lower and intermediate energies for electrons and ions, respectively; in this case after 16.47 UT.

Another factor that contributes to the plasma characteristics in the cusp is the spatial structure of the magnetosheath, which is the dominant source population for the cusp plasma. At the Sunward-most point of the magnetopause, the magnetosheath is highly compressed, with a high density, a high temperature and a small flow velocity. As the magnetosheath then expands and flows over the magnetosphere, its density and temperature decrease and its velocity increases, eventually rejoining the high-speed solar wind far down tail. Therefore, when the reconnected field lines cross the magnetopause near its Sunward-most location, the magnetosheath plasma that flows into the cusp will have its highest densities and temperatures. As the plasma convects away from the reconnection site and the magnetopause crossing point of the field line moves to higher latitudes, the magnetosheath source plasma will have a lower density and temperature and higher flow velocity that eventually becomes directed away from the cusp. Consequently, the density and temperature of plasma entering the cusp will be lower at the higher latitudes, as seen in figure 2.

As the magnetosheath accelerates around the magnetosphere, it eventually becomes supersonic, with its downtail bulk flow speed becoming much larger than the thermal speed of the ions. Within the thermal distribution of the ion velocities, only a very small number of ions will then have velocities directed back toward Earth. Although the high-latitude magnetospheric field lines are still interconnected with the magnetosheath, which still has an appreciable density and temperature, no magnetosheath ions will be detected deep within the magnetosphere on these field lines. This effect contributes to the disappearance of the magnetosheath ions at high latitudes.

Finally, the spatial characteristics of the cusps are affected to some extent by the acceleration of the magnetosheath plasma as it crosses the open magnetopause. The acceleration that particles receive is related to the strength and relative orientation of the electrical current and the electric field at the magnetopause. On the low-latitude dayside magnetopause, the current is parallel to the large-scale electric field (see figure 1(*d*) of MAGNETOSPHERE OF EARTH), causing the particles to be accelerated and gain energy as they cross the magnetopause. At high latitudes, the magnetopause current reverses direction, becoming anti-parallel to the large-scale electric field. Particles that cross the magnetopause at these locations will be decelerated. This acceleration and deceleration across the magnetopause tends to enhance the entry of magnetosheath plasma in the low-latitude portion of the cusp, and inhibit magnetosheath entry at high latitudes.

Note, however, that even in the absence of magnetosheath ions, the magnetosheath electrons are detected continuously throughout the cusp and polar cap on the open field lines. Because the electrons have much higher speeds than the ions, due to their smaller mass, the electron thermal speed remains higher than the tailward magnetosheath bulk flow speed. The thermal distribution

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of magnetosheath electrons will, therefore, always contain a large fraction of the total electron density with velocities directed earthward along the interconnected magnetic field. Because the magnetosheath electrons entering the magnetosphere are flowing into a region where magnetosheath ions do not have access, the charge balance must be accomplished by a background population of magnetospheric ions, probably of ionospheric origin. Some deceleration of the electrons is likely to occur at the magnetopause to maintain charge balance with the less dense background ions. The steady entry of magnetosheath electrons throughout the entire open polar cap is referred to as polar rain. Polar rain electrons are a direct tracer of the solar wind electrons that flow rapidly along the magnetic field and freely penetrate into the open magnetosphere.

The cusp has been described so far as a steady region, whose properties depend on location within the region. The cusp, however, is a dynamic region that responds rapidly to the changing solar wind conditions and to the rate of energy coupling between the solar wind and the magnetosphere. Because of the direct link between the cusp and the dayside magnetopause, the measurements in the cusp are valuable for studying magnetic reconnection and its role in energy transfer processes in the magnetosphere as well as more generally in space and laboratory plasma.

With steady reconnection, the spatial structure measured by a satellite passing through the cusp should progress smoothly as it moves from low to high latitudes, as described above. At each new latitude, the plasma will have slightly different characteristics, reflecting the slightly different time that the magnetic field line has been open to magnetosheath plasma entry. When the reconnection rate is not steady, there will be regions in the cusp which underwent reconnection some time ago that directly neighbor regions that only recently reconnected. In this case, there will be discontinuous jumps in the observed cusp plasma between two spatially contiguous regions reflecting an abrupt change in the length of time that the magnetic field lines have been open. The changes in the plasma characteristics across these jumps have been used to estimate quantitatively the variations in rate of reconnection. It has been found that reconnection often occurs in short bursts, of order 1 min or less, separated by roughly 10 min.

Issues for the future

The cusp is a dynamic region in the magnetosphere that directly connects the flowing solar wind with Earth's ionosphere and upper atmosphere. Although the dominant processes that form the cusp and determine its interaction with neighboring regions are understood at a basic level, many important issues remain unresolved. One important feature of the cusp is that the observed plasma often shows considerably more spatial and/or temporal structure than the basic picture presented here can describe. For example, measurements of the cusp particles, such as those illustrated in figure 2, at times show multiple, overlapping bands of particles, rather than a single band. It is not known what processes at the magnetopause or within the magnetosphere cause this structuring in energy. Also, the measurements typically become more complex as a spacecraft moves away from local noon. It is often unclear if the cusp particles crossed the magnetopause on open magnetic field lines, resulting from reconnection, or if they crossed closed magnetic field lines through some other process. Finally, from measurements by a single satellite, there is always an ambiguity between temporal and spatial changes as a region is traversed. The properties of magnetic reconnection or the structure of the magnetopause and magnetosheath that are inferred from cusp measurements depend critically on our ability to differentiate temporal and spatial variability. These issues are topics for future research, where multisatellite missions, theoretical analyses and numerical simulations will play key roles in advancing our understanding of the cusps.

T G Onsager

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Geomagnetic Storms and Solar Wind Origins

In the mid-1800s extraordinary disturbances accompanied by a great decrease in the horizontal component of the Earth's magnetic field were termed 'GEOMAGNETIC STORMS', or simply 'magnetic storms'. The challenge facing modern space physics in understanding the electromagnetic structure and dynamics of the solar–terrestrial plasma environment originated historically in the study of geomagnetic storms. The effects of geomagnetic storms in the ionosphere and thermosphere are the result of energy inputs from a chain of processes involving kinetic, chemical and electrodynamic coupling between the magnetosphere, ionosphere and thermosphere. The importance of predicting magnetic storms lies not only in its 'academic' value in tracing the way the magnetosphere and ionosphere extract solar wind energy, but also in its practical aspects, influencing human and societal problems such as the effects on communications and satellite anomalies.

Geomagnetic storms are multifaceted phenomena that originate at the Sun and occur in the solar wind, the magnetosphere, the ionosphere and the thermosphere. What causes storms, how is the ring current formed and intensified during magnetic storms, what are the effects of the storm in the ionosphere, how do changes in the magnetosphere–ionosphere system affect the Earth's upper atmosphere, and how can we predict the occurrence and intensity of the storms? These are some of the major questions relating to magnetic storms (Kamide *et al* 1998, Tsurutani *et al* 1997). Magnetic storms must be understood as a chain of processes from the Sun to the Earth when considering the many different aspects of magnetic storms and testing different scenarios on the cause-and-effect relationship of a variety of magnetic storm phenomena.

What is a geomagnetic storm?

Geomagnetic storms at low latitudes

Geomagnetic storms were first detected by using ground magnetograms. The characteristic signature of a magnetic storm is a depression in the magnetic field lasting more than some tens of hours. This depression is caused by the ring current in the magnetosphere encircling the Earth and flowing westward at a geocentric distance of 2–7 R_E (where R_E is the Earth's radius) (see MAGNETOSPHERE OF EARTH: RING CURRENT). A general picture of a typical magnetic storm is shown in figure 1 in terms of variations in the *Dst* index constructed from *H* component magnetograms at low and middle latitudes, where *H* is the horizontal component of the Earth's vector magnetic field. Note that *Dst* is the average of magnetic perturbations at mid- and low-latitude stations distributed in longitude.

A geomagnetic storm customarily involves three phases, beginning with a sudden increase in the *H* component (storm sudden commencement or SSC),

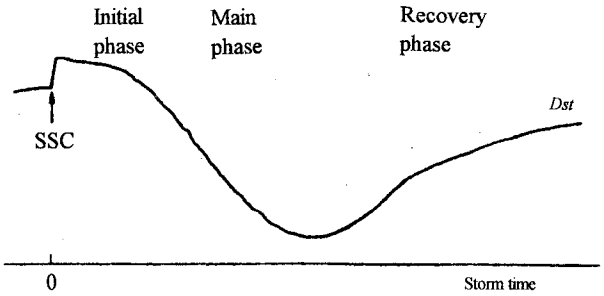


Figure 1. Schematic diagram of a typical geomagnetic storm in terms of the *Dst* index. For the definition of the initial, main and recovery phases, see text.

followed by a period of arbitrary length in which the elevated field does not change very much. This period is called the *initial phase*. The sudden increase in the geomagnetic field is caused by an interplanetary shock. The initial phase is followed by the *main phase* where the development of a depressed *H* component occurs, enduring over a period of a few to several hours. The storm typically concludes with a slow recovery toward the pre-storm level over hours to tens of hours, i.e. the *recovery phase*.

Energy of the ring current

The occurrence of a magnetic storm can most clearly be identified by the existence of the main phase, during which the magnetic field on the Earth's surface is significantly depressed. The depression has a strong latitudinal dependence, being maximum at the equator and minimum at the poles. The intensity of a magnetic storm is commonly defined by the minimum *Dst* value, or the maximum depressed *Dst* magnitude, at the main phase. The depression of the magnetic field during the main phase is explained as the effect of the ring current in the magnetosphere. The ring current is carried primarily by energetic (20–200 keV) ions (see MAGNETOSPHERE OF EARTH). During the recovery phase, the ring current decays due to charge exchange, Coulomb interaction and wave–particle interaction processes in the volume of space occupied by the ring current particles.

The magnitude of a magnetic storm, measured by the *Dst* index, is approximated by the total energy of ring current particles, *E*, through the so-called Dessler–Parker–Sckopke relationship:

$$Dst^*/B_0 = 2E/3E_m$$

where Dst^* is the *Dst* value after a correction for magnetopause currents due to solar wind pressure, B_0 is the magnetic field at the equator on the Earth's surface and $E_m (= 8 \times 10^{17}$ J) is the total magnetic energy of the geomagnetic field outside the Earth. Note that *Dst* includes the magnetic effects not only of the symmetric ring current but of other currents, such as ionospheric, field-aligned and tail currents. According to the above

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

relationship, the *Dst* value is, to a first approximation, linearly proportional to the total energy of the ring current particles. This is the reason why the *Dst* index is used practically as a measure of the intensity of geomagnetic storms.

What is the minimum level in *Dst* is required for a geomagnetic storm? In fact, assigning a threshold serves only an operational purpose in arranging 'storm-time' data and has no physical basis. Small storms can be substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS). It is understood that intense storms are those with the peak *Dst* value of −100 nT or less.

Geomagnetic storms at high latitudes

During a magnetic storm, defined by the magnetic depression at low and mid latitudes, intense substorms take place successively at high latitudes. A magnetospheric substorm is a transient phenomenon initiated on the night side of the Earth in which a significant amount of energy is deposited into the auroral ionosphere and the inner magnetosphere (Rostoker *et al* 1980). A basic question remains unanswered involving whether a magnetic storm is a superposition of intense substorms, each of which constitutes an elementary storm. That is, it is not clear whether the occurrence of substorms is a necessary condition for a magnetic storm or merely coincidental (Siscoe 1997).

Magnetic storms in interplanetary space

When spacecraft observations of solar wind properties became available, it was discovered that the solar wind magnetic field, i.e. the interplanetary magnetic field (IMF), plays a major role in generating magnetic storms. The primary cause of magnetic storms is an intense, long-duration, southward IMF. In fact it became possible to model successfully the growth of the ring current using, as input, only the dawn–dusk component of the interplanetary electric field in the ecliptic plane normal to the Sun–Earth line. By that time it was recognized that an SSC is not a necessary condition for a storm to occur, and hence the initial phase is not an essential feature. The only essential feature of a geomagnetic storm is the significant development of a ring current and its subsequent decay. Note also that if an interplanetary shock is not followed by southward magnetic fields, the sudden increase in the geomagnetic field is called an SI (sudden impulse).

Solar origins of geomagnetic storms

It has long been known that geomagnetic storms start at the Sun. In fact, geomagnetic storms begin when disturbances in the solar wind reach the Earth's magnetosphere. Solar flares and coronal mass ejections (CMEs) are the most energetic of various types of coronal disturbances, although a number of geomagnetic storms are also generated by CORONAL HOLES.

It is important to understand how CMEs are created at the Sun and how they interact with interplanetary structures. Considerable uncertainties exist, however, regarding their magnetic configuration, which must play a key role in unveiling the physics for different sizes of interplanetary CMEs (ICMEs) and such changes as the dimming of coronal features, which is known to be crucial for generating intense geomagnetic storms. One of the major problems is that it is difficult at present to visualize the three-dimensional structure of CMEs and the corresponding magnetic field lines extending into interplanetary space.

Flares and CMEs

The dominant solar events that occur near the maximum sunspot phase of the solar cycle are impulsive ejecta, often called solar flares and CMEs, although how these two elements of solar activity are interrelated is not well understood (see SOLAR CORONAL MASS EJECTION: OBSERVATIONS). CMEs, distinct particle and field structures with field orientations not generally favoring the typical spiral equatorial orientation, are most numerous near the solar maximum and have been demonstrated to cause major geomagnetic storms at that phase of the solar cycle, and at other phases as well.

These events occur on a wide variety of spatial scales and have a variety of speeds, but the ones that are most effective in creating magnetic storms are statistically shown to be rapid, with speeds exceeding the ambient wind speed, so that a forward shock is formed. The fast ejecta are in general associated with flares. There are three essential features of CMEs near the Sun: bright outer loops, a dark region and filamentary structures near the coronal base. It is not well understood, however, whether all these propagate efficiently into interplanetary space.

Coronal holes

During the declining phase of the solar cycle, another type of solar event is shown to dominate, in which the coronal holes emerge from polar regions and extend into, and sometimes far across, the equatorial regions. Coronal holes represent low-temperature regions and the areas of 'open' magnetic field lines. Fast (750–800 km s^{-1}), tenuous plasma is continuously emitted from these solar regions.

When the coronal holes expand effectively toward the equator during the declining phase of the solar cycle, high-speed streams escaping from them create geomagnetic storms. Because these hole regions are long-lived and evolve relatively slowly, they appear to 'corotate' with the Sun over a 27 day period as observed from Earth, leading to the so-called recurrent geomagnetic storms.

Solar wind causes of geomagnetic storms

The causes of geomagnetic storms are strong dawn-to-dusk electric fields in the solar wind, which directly transport solar wind energy into the magnetosphere, driving plasma convection in the magnetosphere. These electric fields are caused by a combination of two factors: solar wind velocity and the southward IMF. Of the two, the southward field is probably more important because of its greater variability, while the solar wind velocity is empirically found to be a minor factor. Solar wind ram

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

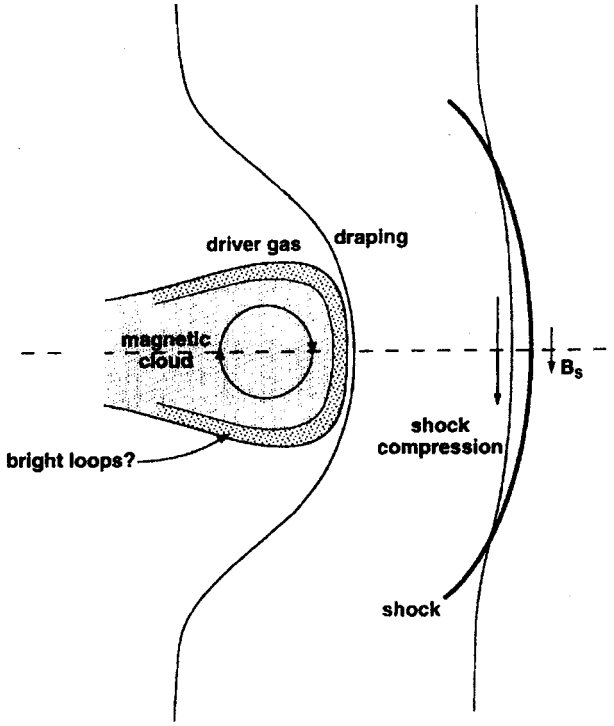


Figure 2. Schematic diagram of regions of intense interplanetary magnetic fields in solar ejecta typically seen near solar maximum. After Tsurutani and Gonzalez (*J. Atmos. Terr. Phys.* **57** 1369 (1995)).

pressure is also important in ring current energization. The specific circumstances or solar wind conditions typically leading to enhanced dawn–dusk electric fields (a B_s magnitude greater than 10 nT) of substantial duration (longer than 3 h) were identified to generate intense geomagnetic storms. There can be a variety of causes of enhanced southward IMF of long duration in the high-speed stream region.

Solar ejecta

Figure 2 shows schematically the regions of possible intense IMF of such solar ejecta, i.e. driver gas, detected at 1 AU. As a fast plasma and field structure propagates from the Sun through interplanetary space, it tends to compress the slower plasma ahead, creating a 'sheath' between the shock and the interplanetary manifestation of the ejecta. The sheath fields leading the rapid ejecta often contain substantial north–south field components, possibly due to compression and draping of the ambient IMF over the ICMEs. Both the remnant fields and plasma in the ejecta proper and those of the sheath can generate intense geomagnetic storms, if the IMF is directed southward.

Most 'geoeffective' ICMEs, giving rise to the largest geomagnetic storms, are magnetic clouds, a subset of ejecta characterized by high IMF magnitude, low field variance and large-scale coherent field rotations, often including large north–south components. This high-field region is typically a low-β ($\sim$0.1) plasma, where β is defined as the ratio of plasma pressure to magnetic pressure. The field reversals within magnetic clouds feature magnetic field reconnection with the Earth's field during the period of the southward field and a lack of reconnection during the northward field. The storm initial phase is caused by increased ram pressure behind the fast shock. The higher plasma density and higher velocity combine to form a much larger solar wind ram pressure. This pressure compresses the Earth's magnetosphere considerably, increasing the field magnitude near the equator. Since interplanetary shocks are thin discontinuities, they create an abrupt onset of the initial phase, which has been called an SSC.

If there are large-amplitude, long-duration southward field intervals in either the sheath or the ejecta proper, a main phase will follow. The storm recovery usually begins as the magnetic field in the magnetic clouds becomes less southward and gradually rotates to a northward direction. However, there is no guarantee that southward magnetic field events will follow the shock, and therefore a storm main phase may not follow the shock.

There are numerous mechanisms that can lead to north–south variations in the sheath, such as shock compression and draping around the ejecta (see figure 2). Intense southward fields can be found in each of the ejecta itself and the sheath. When they are found in both regions simultaneously, the resultant geomagnetic storm should have a two-step main phase.

Corotating interaction region

If a coronal hole is at lower latitudes, the Earth's magnetosphere will be bathed in this stream once each solar rotation. High-speed streams emanating from the coronal hole can create a geomagnetic storm. Typically the solar wind containing low-speed (less than 400 km s^{-1}) and high-density (tens $\times$ 10^6 of particles/m^3) streams lies ahead of the fast stream in interplanetary space. The interaction of the rapid stream with the slow stream ahead tends to create a plasma and field compression. This is called a corotating interaction region (CIR; see SOLAR WIND: COROTATING INTERACTION REGIONS). CIRs are bounded on the leading and trailing edges by forward- and reverse-propagating compressional waves respectively.

Figure 3 shows schematically the formation of CIRs in which magnetic field fluctuations are present in the high-speed stream proper. The magnetic fields of the lower speed tend to be more curved and those of the higher speed more radial. The stream–stream interface (IF) is formed at the boundary between the rapid and slow-speed plasmas. CIRs are characterized normally by intense magnetic fields that reach 30 nT.

A 27 day modulation in geomagnetic activity has been noted since the nineteenth century. This periodicity was later discovered to arise from high-speed solar wind streams originating in coronal holes. Although the high-speed wind from coronal holes generally carries

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

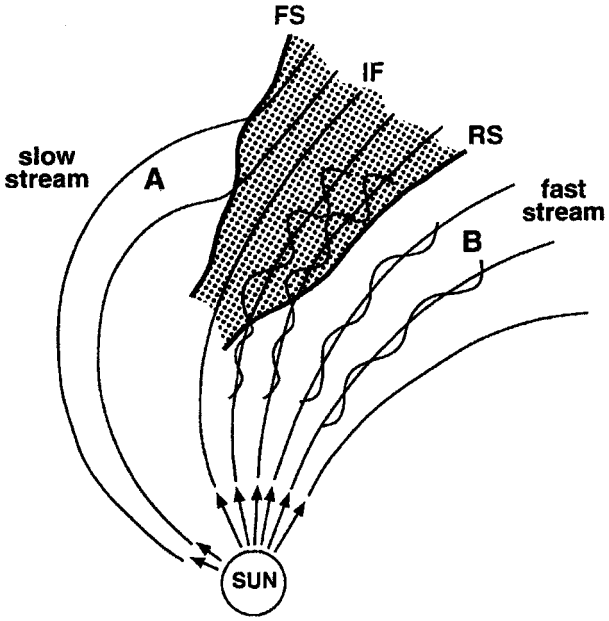


Figure 3. Schematics of the formation of corotating interaction regions (CIRs) during the declining phase of the solar cycle. The interaction between a high-speed stream (B) and a slow-speed stream (A) are shown together with the CIR (shaded). The forward shock (FS), interface surface (IF) and reverse shock (RS) are also indicated. After Tsurutani *et al* (*Geophys. Res. Lett.* **22** 3397 (1995)).

relatively low field magnitude and a radial orientation not conductive to production of steady and substantial north–south fields, CIRs resulting from the interaction of this rapid wind with slower, denser streamer wind produce geoeffective field compressions and deflections. Geomagnetic storms associated with CIR-like plasma signatures generally lack the sudden commencements. These storms are termed gradual geomagnetic storms.

Enhancement/modulation effects of geomagnetic storms

One remarkable characteristic of the solar wind escaping from coronal holes is the presence of continuous Alfvénic activity, consisting of large-amplitude, quasi-periodic fluctuations in the IMF orientation with periods from tens of minutes to a few hours. This activity is generally in phase with similar fluctuations in the solar wind flow direction. Thus, in the interplanetary medium following CIRs, the southward field components associated with these Alfvén waves cause reconnection with the Earth's magnetic field and small injections of plasma into the inner magnetosphere. Because of such persistent substorm activity, prolonged recovery phases of the storms will result. Events of this type are known as 'high-intensity, long-duration, continuous auroral electrojet (*AE*) activity' (HILDCAA) events. The average north–south (B_z) component in HILDCAA intervals is nearly zero, but the induced reconnection in the magnetospheric response results in a continuous occurrence of substorms of almost equal intensity.

Another mechanism which can modulate storm activity is based on the simple geometric mapping from the solar equatorial system, in which the prevailing IMF is ordered, into a magnetospheric system which then orders the magnetospheric response. This mapping, from GSEQ (solar equatorial) to GSM (solar magnetospheric) coordinates, was described by Russell and McPherron (1973). The result of the offsets between the solar equatorial, ecliptic and terrestrial rotational planes is to create a modulation in the rotation from the transverse solar equatorial IMF component (GSEQ B_y) into the north–south field in the magnetospheric system (GSM B_z). The tilt of the terrestrial dipole adds a diurnal component as well. The combined effect is such that negative GSEQ B_y, corresponding to the 'toward sector' IMF, maps most directly to negative GSM B_z at 2200 UT on 5 April, while positive GSEQ B_y ('away sector') maps most directly to negative GSM B_z at 1000 UT on 5 October.

This modulation, called the Russell–McPherron effect (see section on solar cycle distribution below), has been shown to increase in amplitude with increasing activity thresholds and to persist even for very disturbed geomagnetic conditions. Figure 4 shows a seasonal variation of the number of intense storms with clear maxima at the equinoxes.

Storm/substorm relationships

Successive occurrence of substorms

Intense substorms occur frequently during the main phase of magnetic storms. Many researchers believe, or tacitly assume, that magnetic storms develop as a result of the frequent occurrence of substorms. There are certainly good reasons to believe that a geomagnetic storm consists of intense substorms. During episodes of substorm activity, energy can be deposited into the inner magnetosphere, leading to the formation of the so-called partial ring current, which is connected to the substorm electrojet through field-aligned currents. If intense substorms occur successively when the effects of previous substorms still remain in terms of the partial ring current, the local time extent of that partial ring current could increase and evolve into a complete ring. In other words, an individual substorm may cause only a 'mini' ring current, but if substorms occur frequently enough, the injected ring current particles accumulate in the trapping region, forming the symmetrical ring current as a geomagnetic storm.

Is the successive occurrence of many intense substorms a necessary condition for a magnetic storm? We should refocus on the question of why some large substorms are associated with significant increases in the ring current, while other equally large substorms have little effect on it. The main question which then arises regarding the storm/substorm relationship involves the nature of the physical processes which lead to the growth of the ring current during magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

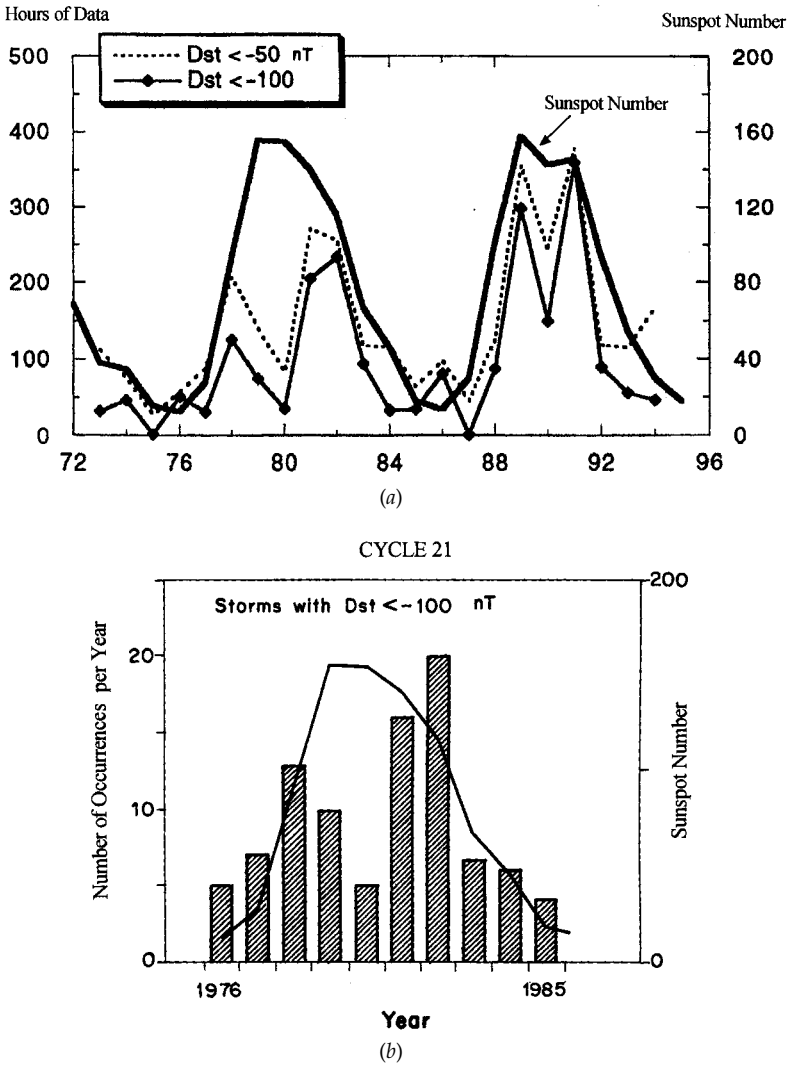


Figure 5. (*a*) Yearly averaged number of hours with *Dst* less than −100 nT (solid curve with diamonds), and with *Dst* less than −50 nT, divided by 5 (dashed curve). Yearly averaged sunspot number (heavy curve) for the interval 1972–96 is also shown. After Kamide *et al* (*J. Geophys. Res.* **103** 17 705 (1998)). (*b*) Dual-peak distribution of intense geomagnetic storms (peak *Dst* less than −100 nT) for solar cycle 21. After Gonzalez *et al* (*Planet. Space Sci.* **38** 181 (1990)).

storms. Substorms were in some way thought to be responsible for the growth of the ring current. In fact, early studies of 'injections' of energetic particles into the inner magnetosphere suggested that the occurrence of substorms led to the acceleration of particles to energies which allowed them to be effective current carriers in the ring current. However, it later became clear that the storm-time ring current was carried by energetic ions with energies typically in excess of several tens of keV. The question of how ring current particles attain their energies and whether substorm disturbances play an integral role in that process is still open.

Ring current composition

Questions continue with regard to the formation, population and composition of the ring current during magnetic storms, as well as the loss mechanism of ring current particles during the storm recovery phase that sometimes takes only several hours, but occasionally several weeks. The relative importance of the solar wind and the ionosphere as the origin of ring current ions in the 20–200 keV energy range seems to have been identified recently. The ionospheric component has been shown to be most abundant during the main phase of intense magnetic storms. This implies that ionospheric ions are accelerated upward by parallel electric fields or by perpendicular heating (involving plasma waves) leading to ion conic distributions, both of which are associated with the onset

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

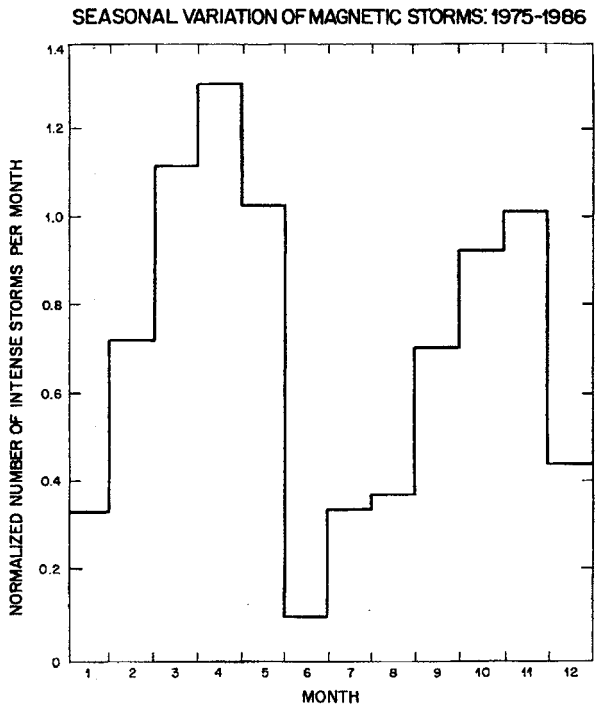


Figure 4. Seasonal distribution of intense geomagnetic storms (peak Dst less than -100 nT) for the interval 1975–86. The normalized number of these storms per month is given. After Gonzalez and Tsurutani (*Eruptive Solar Flares* pp 277–86 (Berlin: Springer) (1992)).

of substorm expansion.

Observations of field-aligned ionospheric ions indicate that injection occurs over a broad range of geocentric distances. These ions are trapped through an efficient magnetospheric process, becoming the main contributors to the storm-time ring current. Therefore, the successive occurrence of intense substorms appears to be a necessary condition for the main phase of magnetic storms. This in fact supports the classical notion that a magnetic storm consists of continual substorm activity.

On the other hand, an enhanced solar wind electric field of sufficient duration directly drives magnetospheric convection, which, in turn, enhances the ring current at the same time. Substorms are often triggered in the course of this process of extraction of power from the solar wind required to build up the storm-time ring current. In this scenario, the action of substorm activity is merely a byproduct of the overall energy process in the magnetosphere–polar ionosphere coupling.

Although the ring current as expressed by Dst can be reproduced mainly by changes in the large-scale electric field driven by southward IMF, the role of substorms is an issue suitable for further study in quantitative detail. In other words, the relationship between substorms and storms is currently poorly understood, and therefore basic questions remain unanswered regarding the hypothesis of whether a magnetic storm represents a superposition of intense substorms.

Solar cycle distribution of geomagnetic activity

ICMEs and CIRs are considered the primary causes of the enhanced duskward electric fields in the solar wind driving geomagnetic storms, while HILDCAAs and the Russell–McPherron effect can be modifiers which do not produce storms without an ICME or CIR, but enhance storms once they are generated. In reality, however, these four mechanisms couple differently from event to event, and even within a geomagnetic storm, from phase to phase. For example, the enhanced field in shocked solar wind can map to even a stronger or weaker southward IMF in GSM coordinates, depending on different seasons. This may be called the 'post-shock Russell–McPherron effect'. CIRs with a southward IMF can also be effective when the time of year is favorable.

ICMEs and CIRs as the primary drivers for strong geomagnetic storms may explain the strong dependence of geomagnetic activity on the solar cycle. Within a solar cycle, there are two peaks in geomagnetic activity, one somewhat before and the other following solar maximum. For two solar cycles, figure 5(*a*) shows the yearly averaged sunspot number and the hours when $Dst < -50$ and -100 nT. It is clear that the number of intense magnetic storms measured by the intervals of low Dst follow the sunspot cycle faithfully at the beginning and toward the ending of the solar cycle, but have pronounced dips during the years of solar maximum (near 1980 and 1990). The occurrence rate of ICME is shown to follow the sunspot cycle, while strong CIRs are most predominant during the declining phase due to the presence of large polar coronal holes extending into lower latitudes. Thus, a simple superposition of occurrence rates of rapid ICMEs and strong CIRs might well produce the double-peaked pattern of figure 5(*a*). Figure 5(*b*) is another example of such a dual-peak distribution, but only for solar cycle 21.

Another statistical trend in geomagnetic activity is its modulation in accordance with the 22-year cycle (Cliver *et al* 1996). The average levels in overall geomagnetic activity are shown to be highest during the rising phase of odd-numbered sunspot cycles and during the declining phase of even cycles. These 'double-solar-cycle' variations can have the following two sources: the first is a combination of the Russell–McPherron effect and the heliographic latitude dependence of the IMF strength, as well as the solar magnetic polarity. The second is an intrinsic difference in large-scale solar magnetic fields, manifested as changes in the solar wind between odd and even cycles. The odd- and even-numbered cycles have some differences in average solar wind speed profiles. A variety of evidence exists for systematic differences in solar properties between even and odd cycles. It must be noted, however, that studies to date have used the *aa* index due to its availability for over 100 years. The *aa* index reflects geomagnetic activity at sub-auroral latitudes. Evidence for a 22 year period in such an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

index does not necessarily guarantee similar modulation in geomagnetic storm activity.

Space weather predictions

Throughout Earth and human history, people have stood in awe of the wildly changing colors and forms of auroras, which occur most frequently during geomagnetic storms (see AURORA). 'Space weather' refers to conditions in the solar wind, magnetosphere, ionosphere and atmosphere that can influence the reliability of space and ground-based technological systems. In other words, the aurora is one of the end results of space weather. Our reliance on technological systems, such as spacecraft missions, power systems, navigation systems and communications, is growing, and these systems are shown to be vulnerable to the unpredictability of space weather conditions. In fact, the risks for disruption of satellite operations, communications and power grids on the ground could be minimized if reliable forecasts of space weather were possible. Studies have already begun aimed at predicting the sequence of extreme conditions that occur in the solar–terrestrial coupling system.

Space weather plays an ever-increasing role in our society as human beings develop more and more delicate technologies in space and on Earth (see SOLAR–TERRESTRIAL CONNECTION: SPACE WEATHER PREDICTIONS). In view of the crucial effects of geomagnetic storms on such human-societal systems as radio communications and satellite drag, there is a strong need for prediction schemes to be continually upgraded. It is essential to integrate our observation and modelling efforts in the space weather project.

The goal of space weather studies is to be able to predict the occurrence of intense geomagnetic storms. How does one monitor the intensity of geomagnetic storms? Their intensity is commonly defined by the minimum *Dst* value during the magnetic storm, which is equivalent to the maximum *Dst* magnitude at the main phase. It may be natural to assume that when a major event is observed near the Earth, something major is also occurring throughout the chain of processes from the Sun to the Earth, seriously affecting the space surrounding the Earth. Is this assumption really correct in magnetic storms? Is a large disturbance in the interplanetary structure really necessary to generate a major magnetic storm? Is it meaningful to rely on the minimum *Dst* value to define the storm intensity?

During the largest magnetic storms, the main phase often goes through two steps. That is, well before the *Dst* decrease has fully recovered to its pre-storm level, a succeeding decrease tends to occur. If this is truly the case, the largest magnetic storms consist merely of two medium-size storms. In other words, what is perhaps needed in generating an intense magnetic storm are two separate regions of southward IMF in an interplanetary structure. One of the most important factors in generating intense magnetic storms is perhaps the slow decay rate of the ring current.

Bibliography

Cliver E W, Boriakoff V and Bounar 1996 The 22-yr cycle of geomagnetic and solar wind activity *J. Geophys. Res.* **101** 27 091–109

Kamide Y *et al* 1998 Current understanding of magnetic storms: storm–substorm relationships *J Geophys. Res.* **103** 17 705–28

Rostoker G *et al* Magnetospheric substorms: definition and signatures 1980 *J. Geophys. Res.* **85** 1663–8

Russell C T and McPherron R L 1973 Semiannual variation of geomagnetic activity *J. Geophys. Res.* **78** 92–108

Siscoe G L 1997 Big storms make little storms *Nature* **390** 448–9

Tsurutani B T, Gonzalez W D, Kamide Y and Arballo J K 1997 *Magnetic Storms* (Washington, DC: American Geophysical Union)

Y Kamide

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Geomagnetic Tail

The geomagnetic tail is an elongated region of the MAGNETOSPHERE OF EARTH extending from the near-Earth space in the antisunward direction. It acts as a giant energy reservoir for the magnetosphere and is therefore an important participant in dynamic processes such as GEOMAGNETIC STORMS and substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS).

The geomagnetic tail has a roughly cylindrical shape with two D-shaped lobes which contain strong magnetic fields and cool plasma lying back to back with a region of hot plasma and weak magnetic field between. The geomagnetic tail extends far beyond the lunar orbit, at least out to 1.5 million km from the Earth, and has a diameter of about 0.3 million km. The magnetic field points toward the Earth in the northern lobe and away from the Earth in the southern lobe. The field magnitude decreases from about 100 nT at the inner edge of the geomagnetic tail (at 40 000 km from the Earth) to about 10 nT reached around 600 000 km from the Earth; tailward of that distance the field magnitude is almost constant at about 10 nT. These antiparallel fields in the tail lobes are supported by a cross-tail current sheet embedded in a thicker plasma sheet roughly in the equatorial plane (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). The average electric field is about $0.3\,\mathrm{mV\,m^{-1}}$ in magnitude and oriented in the dawn-to-dusk direction across the geomagnetic tail. The plasmas enter the tail either from the solar wind or from the Earth's ionized upper atmosphere, the ionosphere, and are in continuous large-scale convective motion under the electromagnetic forces (see SOLAR–TERRESTRIAL CONNECTION: COUPLING BETWEEN SOLAR WIND, MAGNETOSPHERE, IONOSPHERE, AND NEUTRAL ATMOSPHERE, MAGNETOSPHERE OF EARTH: CONVECTION).

The solar wind energy enters the magnetosphere at a variable rate proportional to the solar wind velocity and interplanetary magnetic field direction. When the interplanetary magnetic field points southward (opposite to the magnetospheric field), magnetic reconnection opens the magnetosphere allowing for energy, flux and momentum transfer from the solar wind to the magnetosphere. When the interplanetary field points northward, the magnetosphere is closed and the energy transfer to the magnetosphere is much reduced. During periods of southward interplanetary field, energy extracted from the solar wind is stored in the geomagnetic tail in the form of magnetic field energy and later explosively released in a sequence of events known as a magnetospheric SUBSTORM. During substorms, MAGNETIC RECONNECTION is initiated at the center of the geomagnetic tail about 150 000 km from the Earth. As a consequence, a portion of the plasma sheet tailward of this distance is ejected from the magnetosphere in a magnetically confined plasma bubble, a plasmoid (see figure 1). This sequence lasts 2–4 h during which about 10^{15} J of energy flows through the geomagnetic tail. The magnetosphere is continuously in a dynamic state, where this energy storage–release cycle is repeated on average 4–5 times per day. The substorm process is analogous to a dripping faucet where water (energy) is constantly fed to the water drop (geomagnetic tail), the drop grows until it falls (plasmoid ejection) and when more water is added the drop starts to grow again.

The formation of a geomagnetic tail is a characteristic of the planet–solar wind interaction: each of the solar system planets has a PLANETARY MAGNETOSPHERE and a tail. Furthermore, even smaller bodies with active plasma environments such as COMETS have a tail, although in the case of comets the magnetic forces are weak and the dynamics governed by the ion motion is quite different.

Structure of the geomagnetic tail

Magnetic fields

The structure of the geomagnetic tail is governed by the magnetospheric magnetic field. In the near-Earth region (inside $7R_E$, $1R_E \simeq 6400$ km), the magnetic field is dominated by the nearly dipolar internal geomagnetic field (giving about 30 000 nT at the Earth's surface at the equator). Further away, the magnetospheric magnetic field is a combination of the internal field and a magnetic field created by extraterrestrial current systems induced by the interaction of the solar wind and interplanetary magnetic field with the internal geomagnetic field. The solar wind pressure compresses the dipole field on the dayside. Magnetic reconnection alters the connectivity of the field lines in the polar caps creating a class of 'open' field lines, which are connected to the Earth only at one end. These open field lines created on the dayside are carried downstream in the solar wind flow and stretched into the antiparallel field structure of the tail lobes. The lobe field lines are separated by a cross-tail current sheet. Within the current sheet, at about $100R_E$ from the Earth in the tailward direction, there is another quasi-permanent reconnection region called the distant neutral line. Here the lobe field lines are again reconnected creating 'closed' field lines (connected to the Earth at both ends) Earthward of the neutral line and 'solar wind' field lines (completely disconnected from the Earth) tailward of the neutral line. This magnetic field topology is a permanent structure in the geomagnetic tail.

The Earth's orbital motion causes a small aberration from the Sun–Earth line in the solar wind velocity direction in an Earth-based coordinate system; the geomagnetic tail is aligned along the velocity direction, on average 4° from the Sun–Earth line. The Earth's dipole axis is tilted 11° relative to the rotation axis and the rotation axis is tilted 23° relative to the ecliptic plane normal. Therefore, the orientation of the dipole axis varies between ±35° during the year (pointing toward the Sun during the northern hemisphere summer and away from the Sun during the northern hemisphere winter) and ±11° daily owing to the Earth's rotation. Because the inner magnetosphere current systems follow the dipole motions, the center of the cross-tail current sheet is (3–4)R_E below the Sun–Earth

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

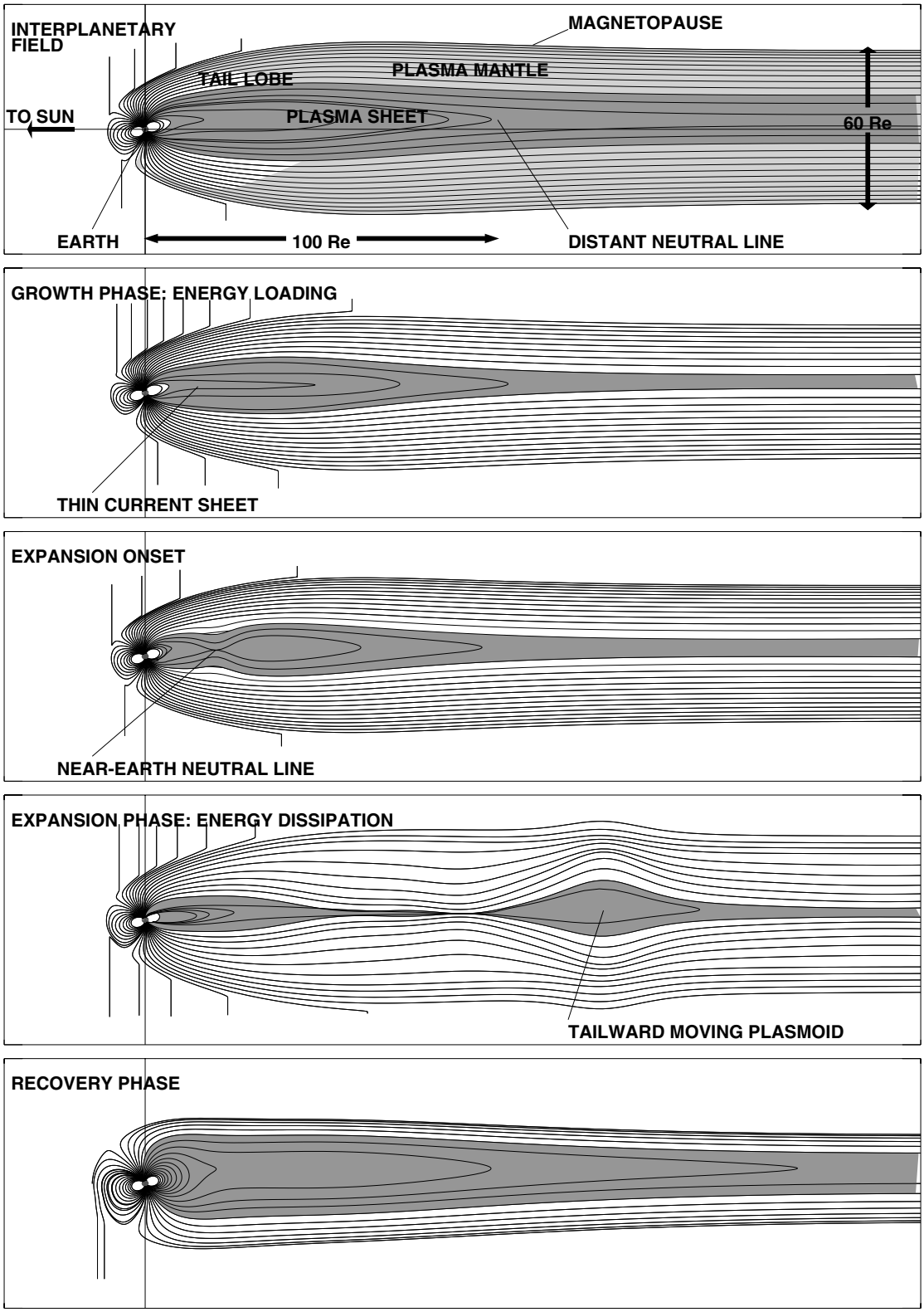


Figure 1. A noon–midnight meridian cut of the geomagnetic tail during northern hemisphere summer (dipole tilt angle toward the Sun). Reconfiguration of the magnetosphere during the substorm process is shown in the sequence. The most important plasma regions are shown in the top panel. The solid lines depict the magnetic field lines; the dark gray shading shows the plasma sheet.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

line during the northern hemisphere winter and above the Sun–Earth line during the northern summer.

During periods of northward interplanetary magnetic field, magnetic reconnection occurs at high latitudes poleward of the cusps reconfiguring already open magnetic field lines. When the interplanetary field has a component also in the cross-tail direction, reconnection occurs predominantly in two quadrants of the geomagnetic tail (northern dawn and southern dusk for dawn-to-dusk directed interplanetary field) and the reconnected solar wind field lines exert a tangential stress on the lobe field lines (proportional to the field component tangential to the magnetopause). This leads to twisting of the magnetotail such that the plasma sheet is at an angle with respect to the equatorial plane. During periods of southward interplanetary field, the twisting is weaker because the tangential stress in the reconnection-dominated quadrants is smaller. The effects of the interplanetary field are larger in the distant tail where the lobe field strength is smaller.

The large-scale extraterrestrial current systems are permanent features of the geomagnetic tail but are highly variable both in intensity and location (see figure 2). Additional, temporary current systems form during dynamic processes. The cross-tail current in the geomagnetic tail flows in the dawn-to-dusk direction and is closed by currents along the magnetospheric boundary, the magnetopause (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). Typical current densities are 10–100 nA m^{-2} in the inner tail (at $15R_E$) and 1 nA m^{-2} in the distant tail, the total current is roughly 2–3 MA. The ring current encircles the Earth near the equatorial plane in the inner magnetosphere (see MAGNETOSPHERE OF EARTH: RING CURRENT). The current density is less than 1 nA m^{-2} during quiet times and 1–10 nA m^{-2} during geomagnetic storms; the total current varies between 0.1 and 5 MA. Two large-scale magnetic-field-aligned current systems couple the magnetosphere and the ionosphere. The region 1 currents flow at high auroral latitudes, originating from magnetospheric boundaries and from the geomagnetic tail. These currents flow toward the ionosphere in the morning sector and away from the ionosphere in the evening sector. The region 2 currents flow at subauroral latitudes from the morning-sector ionosphere to the equatorial geomagnetic tail and from the tail to the evening-sector ionosphere. These currents carry roughly 1–3 MA; the current density is a few μA m^{-2} at the ionospheric altitudes and 3 orders of magnitude lower in the geomagnetic tail. The most important temporary current system is the substorm current wedge, which connects the geomagnetic tail and the auroral ionosphere during the substorm expansive phase carrying roughly 1 MA.

Electric fields

The solar wind–magnetosphere interaction induces an electric field in the dawn-to-dusk direction. The plasma at low altitudes in the plasmasphere co-rotates with the Earth, which gives rise to an electric field pointing radially inward (see MAGNETOSPHERE OF EARTH: PLASMASPHERE). The combination of these two components gives the large-scale tail electric field typically about 0.1–0.5 mV m^{-1}, being larger during dynamically active periods (southward interplanetary magnetic field) and smaller during quiet times (northward interplanetary field). In addition to that, electric fields induced by the magnetic field configuration changes (Faraday's law $\boldsymbol{\nabla} \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$), intense and strongly fluctuating small-scale electric fields, and a variety of plasma waves are present in the geomagnetic tail.

Plasmas

The geomagnetic tail has two plasma sources, the solar wind and the dense ionosphere. Both source populations are very cold: the plasma temperature in the solar wind is around 10 eV (1 eV corresponds to 11 600 K), and in the ionosphere about 0.1 eV. The solar wind plasma (or actually plasma in the magnetosheath between the magnetopause and the bow shock) enters the magnetosphere via dayside reconnection; a portion of the particles enter through the magnetospheric boundaries also along the geomagnetic tail. In this plasma mantle, the cold plasma flows tailward typically at about 200 km s^{-1} (somewhat slower than in the adjacent magnetosheath) and drifts across the tail lobes at a speed of about 20 km s^{-1} toward the plasma sheet. The ionospheric plasma rises both from the polar cap and from the auroral regions; especially the dayside cusp region is an effective source for magnetospheric plasma (see MAGNETOSPHERE OF EARTH: DAYSIDE CUSP). In the distant tail, the solar wind source dominates. The accelerated ionospheric particles form an important source for the $\geq$15 keV inner-tail population especially during geomagnetically active times.

The combination of the electric and magnetic fields in the geomagnetic tail controls the large-scale plasma motions. In the tail lobes, the dawn-to-dusk electric field and Sun–Earth-directed magnetic field cause the plasma to drift toward the plasma sheet. Plasmas in the tail lobes and plasma mantle are cool; their energies are below a few hundred eV and densities less than 0.1 particles cm^{-3}. Within the plasma sheet, the magnetic field is predominantly northward and the plasma convects Earthward. As the electric field is parallel to the cross-tail current, the plasma is energized ($\boldsymbol{j} \cdot \boldsymbol{E} > 0$) to form the plasma sheet population whose mean energy is a few keV and average density a few particles cm^{-3}. Between the lobe and the plasma sheet, the plasma sheet boundary layer is characterized by anisotropic plasma distributions and multiply streaming plasma components. In the inner magnetosphere, the outer van Allen radiation belt hosts a population of high-energy (tens of keV to several MeV) electrons, and the ring current consists of energetic ions (tens to hundreds of keV). These ions are mostly protons but also heavier elements such as ionospheric oxygen and helium and alpha particles from the solar wind are present especially during periods of high magnetospheric activity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

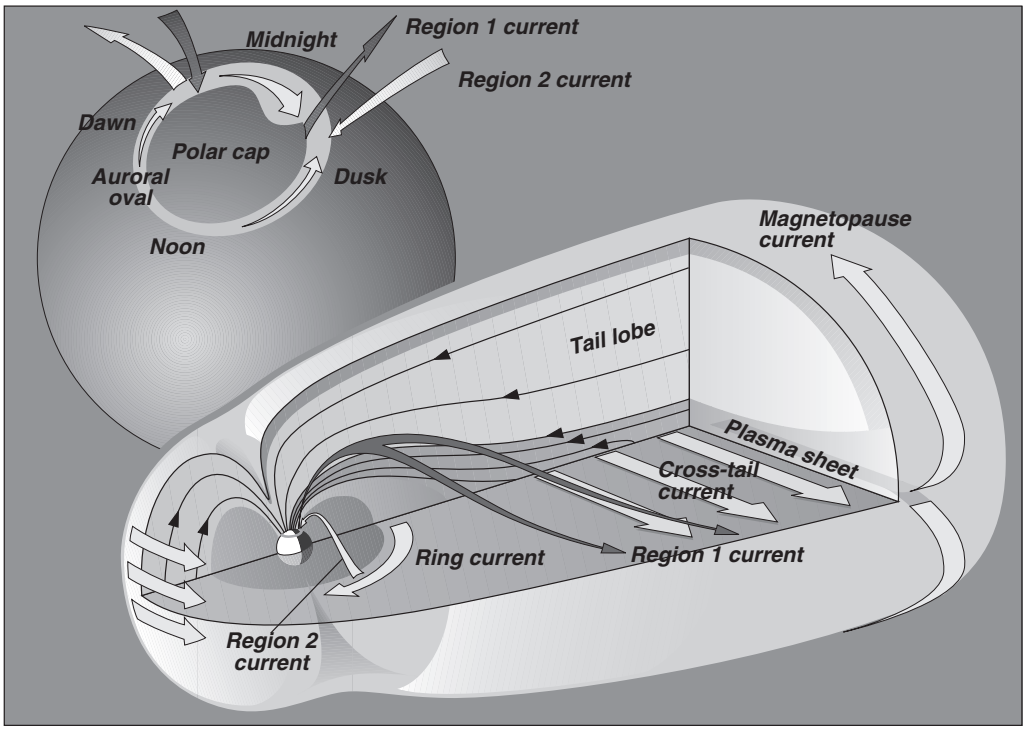


Figure 2. Magnetospheric current systems and a three-dimensional view of the inner magnetosphere and geomagnetic tail. The enlargement of the Earth shows the connectivity of the current systems to the auroral ionosphere.

At the Earthward end, the geomagnetic tail field lines all connect to the high-latitude portion of the Earth's dipolar field. The closed plasma sheet field lines connect to the auroral regions encircling the polar caps. Hence, the plasma sheet is the source region for the particles (mainly electrons) creating auroral light as they collide with the atmospheric atoms. The lobe field lines are magnetically connected to the polar caps poleward of the auroral ovals.

Dynamics of the geomagnetic tail

The geomagnetic tail is a remarkably stable and coherent structure, with distinct plasma regions identifiable from observations out to the distance of $240R_{\rm E}$ covered by *in situ* measurements. Current sheets, such as the cross-tail current sheet, are often the locations where instabilities can grow in space plasmas, and this is the case also for the geomagnetic tail. However, the magnetic field component normal to the cross-tail current sheet stabilizes the system under most circumstances so that significant energy input is required before large-scale instabilities can grow.

During the basic dynamic cycle in the magnetosphere, the magnetospheric substorm, energy stored in the geomagnetic tail is dissipated by a reconfiguration process. The process is initiated by a localized growth of a microinstability within the cross-tail current sheet and quickly expands so that its effects are observable over a large portion of the magnetosphere. Therefore, the substorm is a global reconfiguration process in the magnetosphere.

The repetition and intensity of the dynamic cycles are controlled by the rate of energy input from the solar wind roughly given by $\epsilon = 10^7 V B^2 (7R_{\rm E})^2 \sin^4(0.5\theta)$, where V is the solar wind velocity, B the interplanetary magnetic field and $\tan(\theta) = B_Y/B_Z$. Correlation analyses using solar wind velocity, interplanetary magnetic field and a magnetic activity indicator (AL) derived from ground-based magnetometer data show that the correlation between $VB_{\rm south}$ and AL has two peaks, one at 20 min corresponding to the direct solar wind driving of magnetospheric convection and another broad peak at 1 h corresponding to the substorm-associated strong energy dissipation in the geomagnetic tail.

Although the substorm energy comes from the solar wind, the location and time of the instability onset leading to the global reconfiguration are controlled by internal tail processes. The changing boundary conditions and magnetic field configuration lead to the formation of a thin current sheet that extends from the inner geomagnetic tail ($(8–10)R_{\rm E}$) out to at least $(20–30)R_{\rm E}$. This current sheet has a thickness of only a few hundred km, and the magnetic field component normal to the current sheet is very small, only a few nT. Such a thin current sheet is a required precondition before the microinstability leading to the substorm expansion phase can grow. The instability evolves like a tearing-type process breaking the magnetic topology and has a growth time of only a fraction of a minute. The details of the microphysical instability process are as yet unknown.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The substorm cycle begins with energy extraction from the solar wind via dayside reconnection at a rate of 10^{11} W and its storage in the tail magnetic energy. This substorm growth phase typically lasts about 1 h giving about 10^{15} J of energy to the tail. Magnetic reconnection at a newly formed near-Earth neutral line (20–30)R_E from the Earth initiates the reconfiguration process, which leads to disruption of the inner tail current sheet, consequent decrease in the lobe magnetic field (Ampère's law $\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j}$) and the formation of a plasmoid or, more generally, a flux rope, which is ejected from the geomagnetic tail in the antisunward direction (see figure 1). Because the reconfiguration process is limited in the cross-tail direction, the plasma jets Earthward at a speed higher than the convection velocity on either side. The flow shear creates a pair of field-aligned current sheets, where the current flows to the ionosphere in the morning-side leg and out of the ionosphere in the evening-side leg. This is the substorm current wedge. The stored energy is dissipated during these processes, which typically last about 30 min (substorm expansion phase). After that, the energy dissipation ceases and the neutral line moves tailward to replace the distant neutral line. The magnetosphere recovers its original state during a slow process of rebuilding the cross-tail current.

The structure and dynamics of the geomagnetic tail have been probed by several spacecraft since the beginning of the space age. The first measurements were made by the EXPLORER spacecraft in the early 1960s. These were followed by a series of ten Interplanetary Monitoring Platform (IMP) spacecraft during the 1960s and early 1970s. The first measurements from the distant geomagnetic tail were obtained from PIONEERS 7 and 8, which crossed the tail at distances of 1000 R_E and 500 R_E, respectively. A more detailed picture of the dynamics was given by the three INTERNATIONAL SUN–EARTH EXPLORER (ISEE) spacecraft in the late 1970s and early part of the 1980s. While ISEE 1 and ISEE 2 had apogees at 22 R_E, ISEE 3 was located for part of its lifetime in the distant geomagnetic tail probing it out to 220 R_E. The International Solar Terrestrial Physics (ISTP) program in the 1990s is a joint effort between all the major space agencies. While there are several spacecraft probing various regions of the geospace, the most important for the study of the geomagnetic tail has been the INTERBALL tail probe reaching out to 30 R_E and GEOTAIL, which has covered the tail from 220 R_E inward to about 10 R_E in great detail.

Bibliography

Nishida A, Baker D N and Cowley S W H 1998 *New Perspectives of the Earth's Magnetotail (Geophysical Monograph Series)* (Washington, DC: American Geophysical Union)

Tuija I Pulkkinen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Magnetopause

The magnetopause is the location where the external PLASMA and magnetic field pressure of the magnetosheath balance the internal plasma and magnetic field pressure of the MAGNETOSPHERE OF EARTH. It is usually considered the boundary of the magnetosphere. The magnetic field orientation changes across the boundary from its magnetosheath to its magnetospheric orientation, implying that this boundary is a current sheet. While pressure balance and changes in the orientation of the magnetic field at a current sheet are relatively simple concepts, the structure of the Earth's magnetopause is complicated because the plasmas on both sides of the discontinuity are magnetized.

For typical SOLAR WIND conditions, the distance from the center of the Earth to the magnetopause along the Earth–Sun line, called the subsolar distance, is about 10 times the radius of the Earth, or 10 Earth radii (approximately 64 000 km). The shape of this boundary can be approximated by an ellipsoid of revolution about the Earth–Sun line with the magnetopause at the terminator about 50% further from the Earth than it is at the subsolar point. In the magnetotail, this approximation is not accurate. The magnetopause does not converge in this region; rather it extends many hundreds of Earth radii in the direction away from the Sun. Figure 1 is a cut through the magnetosphere in the noon–midnight meridian plane, showing the magnetopause. As discussed below, the magnetopause shown in figure 1 is called a 'closed' magnetopause because there are no magnetic field lines crossing the boundary.

Since the magnetopause is in pressure balance, changes in the internal or external pressure will cause the location and shape of this boundary to change. The solar wind dynamic pressure (that is, the solar wind mass density times the square of the solar wind velocity) is far from constant. This dynamic pressure is converted to plasma thermal pressure and magnetic field pressure in the magnetosheath. Thus, variations in the solar wind dynamic pressure cause the magnetopause to move toward and away from the Earth. The subsolar distance can move in to about half its nominal distance for extremely high solar wind pressure and can move out to about twice its nominal distance for extremely low solar wind pressure.

A magnetopause was hypothesized long before spacecraft observed it and measured many of its properties. In 1931, Sydney Chapman and Vincenzo Ferraro suggested that a charge-neutral plasma (composed of ions and electrons) ejected from the Sun during magnetic storms would envelop the Earth. Because the medium surrounding the Earth is a good electrical conductor, the plasma from the Sun would not penetrate the Earth's magnetic field. Instead, the plasma (and magnetic field) from the Sun would pile up and be swept around the Earth's magnetic field, forming the boundary

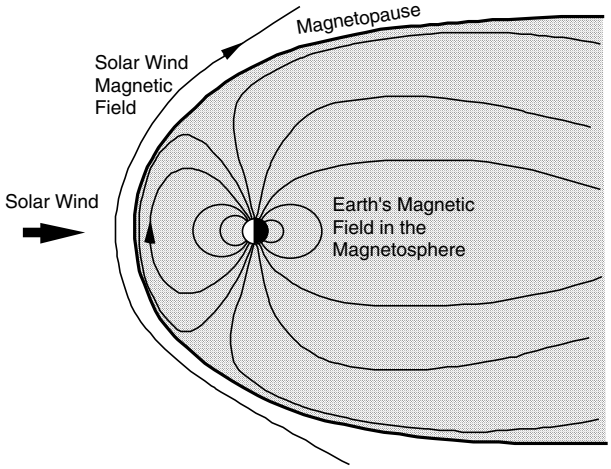


Figure 1. The Earth's magnetopause is the location where the pressure of the solar wind and the internal pressure of the Earth's magnetosphere balance. This magnetopause is closed so that no magnetic field lines cross the boundary and the plasmas on either side do not intermix.

that is now known as the magnetopause. In effect, a current system would be set up on the magnetopause surface which would shield the planetary magnetic field from the interplanetary plasma. Chapman and Ferraro were attempting to explain the initial decrease in the Earth's magnetic field during magnetic storms which were known to be associated with solar activity. They believed that the plasma ejected from the Sun was intermittent and that it did not contain a magnetic field (thus, the magnetic pressure external to the magnetopause was zero). It is now known that the charge-neutral plasma from the Sun (the solar wind) is unsteady but not intermittent and does contain a magnetic field. The continuous solar wind and the presence of the magnetic field in the solar wind (see SOLAR WIND: MAGNETIC FIELD) have important implications for the structure of the magnetopause beyond what Chapman and Ferraro envisioned. However, their pioneering work on the magnetopause was significant and is a good starting point for understanding this complex boundary.

As a first approximation, the structure of the magnetopause can be investigated by neglecting the internal plasma in the magnetosphere and neglecting the external magnetic field in the solar wind. The first assumption is a reasonable approximation because the internal plasma pressure is about 5–10 times lower than the external plasma pressure. The second assumption has more profound implications discussed below.

In this first approximation, the external, unmagnetized plasma pressure is balanced by the internal magnetic field pressure. If the external plasma has zero temperature, then the balance occurs between the external dynamic pressure (i.e. the plasma density and bulk flow velocity) and the internal magnetic field. In this case, the unmagnetized ions and electrons impinge on the magnetic field and are completely turned around. Because the ions are more

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

massive than the electrons, they penetrate slightly further into the Earth's magnetic field. This charge separation produces an electric field which self-consistently turns the ions and electrons around. Throughout this process, the external plasma pressure and the internal magnetic field pressure must balance. This relatively simple description of the magnetopause was formulated and solved analytically before the first observations of the Earth's magnetopause. It predicts a relatively thin current layer which is of the order of the geometric mean of the electron and ion gyroradii (for typical conditions in the magnetosheath, this mean is approximately the ion gyroradius or approximately 50–100 km). It also predicts that there is no transfer of plasma across this thin layer since all plasma incident on the layer is ejected.

While neglecting the presence of a plasma in the magnetosphere has an effect on the structure of the magnetopause, especially at the flanks or sides of the magnetopause, the major difference between this simple approximation and the Earth's magnetopause is that it neglects the external magnetic field in the magnetosheath. Although the magnetic field in the magnetosheath is weaker than its magnetospheric counterpart, it is the presence of this magnetic field that causes most of the interesting phenomena associated with the Earth's magnetopause.

The next simplest approximation of the magnetopause is to assume that it is a tangential discontinuity. The magnetic field may undergo an arbitrary rotation across this type of discontinuity but no magnetic field lines cross the discontinuity and the plasma on either side of this magnetic field rotation does not intermix. Plasma intermixing is not allowed across a tangential discontinuity because, under ideal MAGNETOHYDRODYNAMICS (MHD), the magnetic field is 'frozen in' the plasma and a group of ions or electrons that share a common field line never share that field line with another group of ions or electrons on an adjacent field line. Since no magnetic field lines thread the boundary and the plasma on either side of the discontinuity cannot intermix, the magnetopause (and magnetosphere) is considered 'closed'. A closed magnetopause has the magnetic topology in figure 1. Once again, the total pressure on either side of the tangential discontinuity must balance. At the location where the magnetic field rotates, plasma can move along the discontinuity by executing particle orbits in the layer. By virtue of the different orbits for ions and electrons, this particle motion constitutes a current which self-consistently maintains the layer. The thickness of this structure can be of the order of an ion gyroradius (or approximately 50–100 km, for typical conditions in the magnetosheath). Spacecraft observations indicate that, at least at times and over some limited areas, the magnetopause has the characteristics of a tangential discontinuity.

In the 1970s, as spacecraft observations became more detailed, it became clear that solar wind plasma regularly crossed the Earth's magnetopause. In particular, solar wind plasma was found in a layer approximately 3000 km thick inside the magnetopause. This layer is much thicker than a tangential discontinuity magnetopause that has a thickness of about 100 km. It thus became important to understand how this magnetized plasma crossed a magnetic boundary and what the implications were for the structure of the boundary. One suggestion was that the plasma diffused across the tangential discontinuity. In a magnetized plasma this diffusion must occur 'anomalously' through the action of waves which disrupt the regular particle motion about the magnetic field, thus breaking the 'frozen-in' condition. Another way to produce this diffusion would be growth of an instability such as the Kelvin–Helmholtz instability caused by shear flows of the plasma at the magnetopause. If the Kelvin–Helmholtz instability were to grow to saturation amplitudes and enter a non-linear phase, plasma elements on two sides of the magnetopause could intermix. Spacecraft observations of waves and shear flows at the magnetopause held promise for anomalous diffusion. This subject is still controversial, but the observed wave amplitudes now appear to be insufficient to produce substantial anomalous diffusion necessary to match the observations of solar wind plasma crossing the magnetopause. Similarly, most shear flows at the magnetopause do not appear to be large enough to drive the Kelvin–Helmholtz instability to saturation.

In 1961 James Dungey suggested that a breakdown of the frozen-in condition of ideal magnetohydrodynamics would cause the solar wind and terrestrial magnetic fields to diffuse into a region and interconnect. He developed these ideas from work on solar magnetic field interconnection (or reconnection) but realized that conditions favorable for MAGNETIC RECONNECTION could occur when (and where) the magnetic field in the magnetosheath was directed opposite the terrestrial magnetic field. Under magnetohydrodynamics (outside the small region where the diffusion and interconnection occurs), the tangential discontinuity magnetopause would decompose into a series of MAGNETOHYDRODYNAMIC WAVES, each performing a change in the plasma or magnetic field parameters that is required for the transition between the solar wind and magnetosheath conditions. Despite this complex structure, the total pressure remains constant through each wave front and the 'magnetopause' is still a balance between external and internal total pressure.

Computer simulations that consider kinetic effects in the plasma (as opposed to MHD theory which treats the plasma as a fluid or fluids) indicate that some of these waves merge or change characteristics, so that, at least to a reasonably good approximation, the magnetopause during reconnection has the properties of a rotational discontinuity. The primary differences between tangential and rotational discontinuities are that magnetic field lines thread a rotational discontinuity and the plasma on either side can cross this type of discontinuity. Across a rotational discontinuity, the mass, momentum and energy densities are conserved. These conservation laws give rise to a series of jump conditions for the density, velocity and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

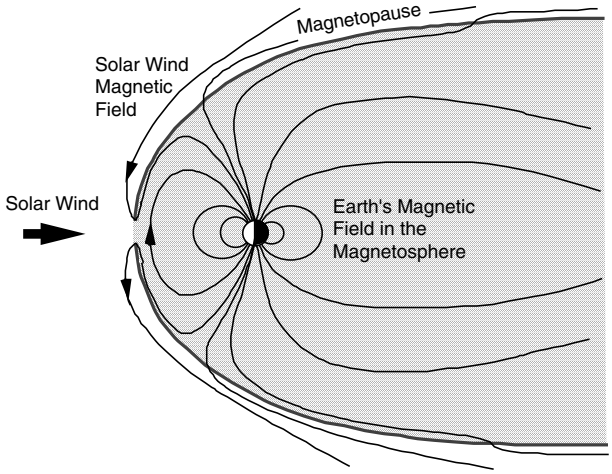


Figure 2. An open magnetopause where reconnection is occurring in the subsolar region. The magnetopause is still the location where the solar wind and magnetospheric pressures balance but magnetic field lines thread the boundary and plasmas on either side can intermix.

energy across the discontinuity. For example, the mass density and the normal velocity must be conserved across the discontinuity. Under certain assumptions, the jump conditions can be tested at the Earth's magnetopause. These tests reveal that the magnetopause is often a rotational discontinuity.

The magnetopause structure when magnetic reconnection occurs is significantly different from the magnetopause in figure 1. Figure 2 is a cut through the magnetosphere in the noon–midnight meridian plane, showing the magnetopause when reconnection is occurring in the subsolar region. Magnetosheath and magnetospheric magnetic field lines interconnect in a relatively small region (called the diffusion region) near the subsolar point. The diffusion region is very small compared with the magnetopause and, as a consequence, the probability that a spacecraft will cross through the diffusion region is also small. Even if a spacecraft crossed through the region, it might not be recognized since the distinction between this region and the magnetopause outside of the diffusion region is relatively subtle and only now beginning to be understood from a theoretical standpoint.

Under the action of the magnetosheath plasma flow around the magnetosphere and the magnetic field line tension that is released in the reconnection process, these interconnected magnetic field lines convect away from the diffusion region toward the magnetotail. More magnetic field lines convect and reconnect in the diffusion region, continuing the process. Thus, figure 2 is a snapshot in time where the magnetopause can be traced through the rotation of the magnetic field from its magnetospheric to its magnetosheath orientation. Because of the interconnection of the magnetic field and the intermixing of the plasma, the magnetopause in figure 2 is described as an 'open' magnetopause.

Although the magnetopause structures in figures 1 and 2 are significantly different, there are some features that these discontinuities have in common. The first feature is that the magnetopause is the location where the magnetic field changes from its magnetosheath orientation to its magnetospheric orientation. For the closed magnetopause in figure 1, the magnetic field rotates in a plane tangential to the magnetopause from its magnetosheath to its magnetospheric orientation and no magnetic field lines thread the boundary. For the open magnetopause in figure 2, the magnetic field undergoes the same rotation tangential to the magnetopause, but there is also a small component of the magnetic field that threads the boundary.

The second feature that remains in common with all magnetopause structures is that the total magnetic field and plasma pressure remains constant across the boundary. For the closed magnetopause in figure 1, the internal pressure of the magnetosphere deflects the incoming magnetosheath plasma around what is in effect an impervious obstacle. For the open magnetopause in figure 2, this deflection still occurs but the obstacle is no longer impervious. Only about 1% of the plasma crosses the open magnetopause in figure 2. However, this transfer of plasma has a profound effect on the structure and dynamics of the Earth's magnetosphere. Further information on the magnetopause can be obtained from the book entitled *Physics of the Magnetopause* (Song *et al* 1995).

Bibliography

Chapman S and Ferraro V C A 1931 A new theory of magnetic storms: part I—the initial phase *Terr. Magn. Atmos. Electr.* **36** 77–97

Dungey J W 1961 Interplanetary field and the auroral zones *Phys. Rev. Lett.* **6** 47

Song P, Sonnerup B U Ö and Thomsen M F (ed) 1995 *The Physics of the Magnetopause (Geophysical Monograph 90)* (Washington, DC: American Geophysical Union)

Stephen A Fuselier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Plasma Sheet

The plasma sheet is a slab of hot plasma in the equatorial region of the geomagnetic tail, and carries the cross-tail (dawn-to-dusk) current that supports the oppositely directed magnetic fields of the tail lobes (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL). The plasma sheet and the embedded current sheet are clearly present to distances of at least 1.5 million km away from the Earth.

The existence of the current sheet was inferred immediately after the discovery of the geomagnetic tail by the IMP-1 satellite in 1964; the magnetic field directions are nearly antiparallel between the northern and southern lobes. It was also inferred that the current sheet must be filled with hot plasmas to sustain the magnetic pressure in the lobe. The lobe field magnitude is typically ~20 nT at a distance of 30 R_E (Earth radii; $1R_E \sim 6378$ km), and the distance dependence is given by $\sim 124x^{-0.54}$, where x is the downtail distance (in units of R_E) from the Earth.

Later observations have revealed that the plasma temperature and number density in the central plasma sheet are typically of order ~keV and $\sim 10^6$ m^{-3}, respectively. The magnetic field inside the plasma sheet is weaker and more variable than in the lobe. However, it is still consistent with being oppositely directed in the northern ($B_x > 0$) and southern ($B_x < 0$) parts and separated by a neutral sheet ($B_x = 0$), where B_x is the magnetic field component in the Sun–Earth direction and positive toward the Sun. Towards the neutral sheet, the field magnitude decreases and plasma pressure increases, and thus the central plasma sheet is characterized by a high-β plasma ($\beta \gg 1$), where β is the ratio of thermal to magnetic pressure. Usually ions make the major contribution to the pressure and to the cross-tail current associated with the pressure gradient.

The half-thickness of plasma sheet is a few R_E near the tail center and becomes thicker towards the dawn and dusk flanks. It can be clearly identified from the adjacent regions by differences in the magnetic field and plasma properties. Across the boundary from the lobe to plasma sheet, the β value jumps from $\ll 1$ to a fraction of unity. Energetic ion beams are often observed to stream along the magnetic field lines in a layer (called the plasma sheet boundary layer) just outside of this transition. Inside the plasma sheet, plasma flow properties are highly variable with speeds ranging from <50 km s^{-1} (stagnant) to >1000 km s^{-1}. However, the flow directions are consistently associated with characteristic magnetic field variations; it is sunward in association with northward B_z, and antisunward with southward B_z. This coherent flow property represents convection in the plasma sheet. On the average, the convection is directed earthward in the near-Earth region, while it is tailward in the distant tail. This feature is well understood in terms of magnetic reconnection occurring in the neutral sheet of the geomagnetic tail.

In the near-Earth region, the plasma sheet is magnetically connected to the Earth at high latitudes, and particles that can penetrate to low altitudes lose energy through interactions with the dense atmosphere, causing auroral emissions. On the other hand, some of the plasma sheet particles are injected into the ring current region during storms and substorms. Therefore the plasma sheet is the source of the auroral particles as well as of the ring current (see also GEOMAGNETIC STORM, MAGNETOSPHERE OF EARTH: RING CURRENT, MAGNETOSPHERE OF EARTH: SUBSTORMS).

Magnetic reconnection in the plasma sheet

The structure and dynamics in the plasma sheet are basically understood in terms of MAGNETIC RECONNECTION, in which an X-line (neutral line) is formed in the dawn–dusk direction. As shown in the upper panel of figure 1, 'Petschek' reconnection leads to the formation of four distinct boundaries (slow shocks). The upper and lower regions represent the northern and southern lobes, respectively, where cold plasmas are convected toward the neutral line, while in two separate regions inside the boundaries hot plasmas flow out earthward on the earthward (left-hand) side and tailward on the tailward (right-hand) side of the neutral line. The earthward and tailward flows are accompanied by northward ($B_z > 0$) and southward ($B_z < 0$) magnetic fields, respectively. Hence the plasma flows carry magnetic flux, and the resultant convection electric field is directed dawn to dusk on both sides of the neutral line. Statistically, the neutral line is formed at distances of $\sim 140R_E$ down the tail, which is called a 'distant' neutral line. The convection electric field is of the order of 0.1 mV m^{-1} on the average. The field lines on the earthward side are closed, with both feet connected to the ionosphere at high latitudes in the northern and southern hemispheres, while on the tailward side both ends are connected to the interplanetary magnetic field (IMF).

In addition to the distant neutral line, magnetic reconnection is initiated explosively at a time close to substorm onset in the near-Earth region (20–30 R_E). The resultant neutral line is called a 'near-Earth' neutral line (NENL). The NENL formation is required to dissipate magnetic energy stored in the tail lobes during the substorm growth phase because the reconnection rate in the distant tail is generally too slow to balance the enhanced reconnection on the dayside under continuously southward IMF. In the growth phase, the magnetic field configuration becomes tail like and the current sheet becomes thinner in a wide region of the geomagnetic tail. Then, magnetic reconnection commences somewhere in an extremely thin sheet near the central axis of the tail. The plasma sheet on the tailward side of the NENL, which was originally convecting earthward with closed field lines, forms a plasmoid with a magnetic island, or flux ropes, which is expelled down the tail. On the earthward side, high-speed earthward flows transport northward magnetic flux towards the Earth, which contributes to 'dipolarization' of the magnetic field near the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

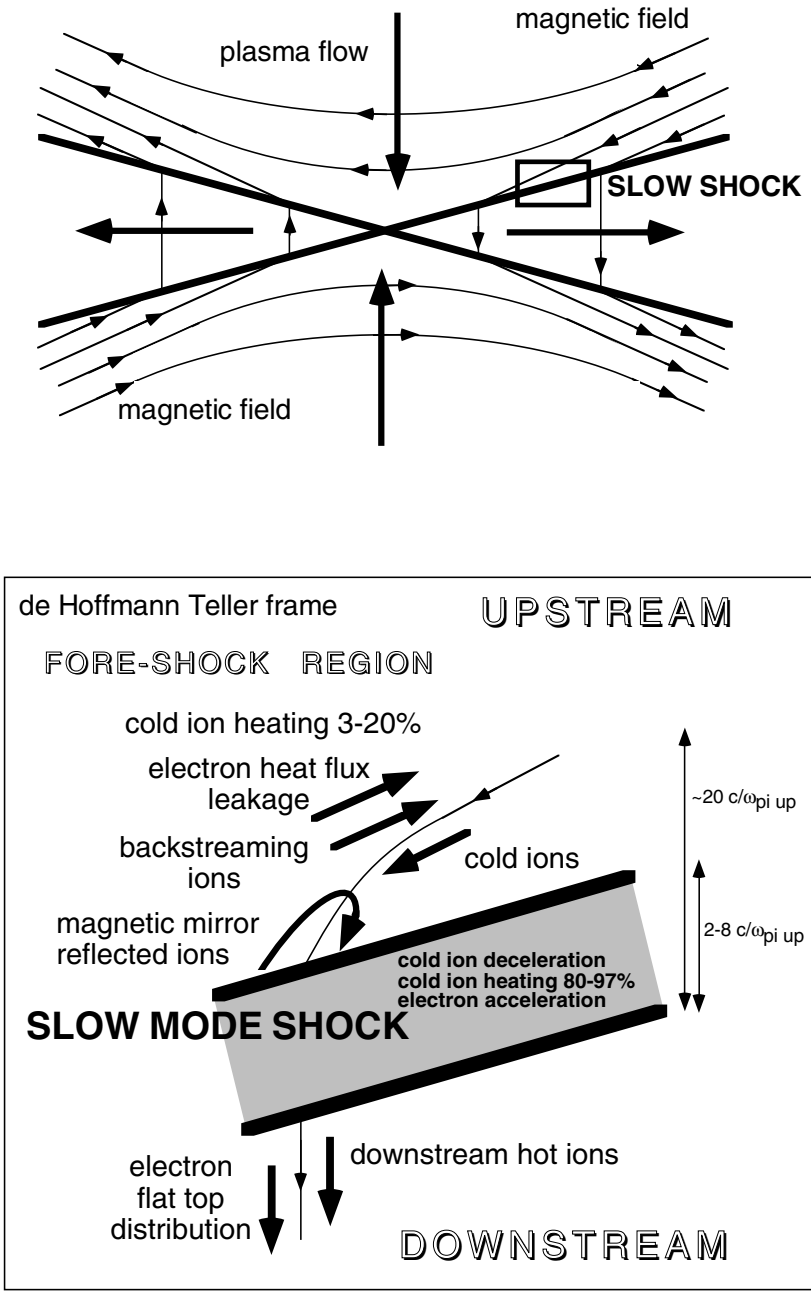


Figure 1. A schematic structure of 'Petschek' reconnection and slow-mode shocks in the Earth's magnetotail.

Recent *in situ* plasma measurements on satellites have revealed characteristic features in the particle distribution functions associated with magnetic reconnection. Comparison of the observed data with results from computer experiments has enabled us to understand them qualitatively and quantitatively. Magnetic reconnection and the associated particle heating processes are topics of current research in space physics.

Plasmoid

A plasmoid is created initially near the center of the geomagnetic tail at distances of 20–30 R_E by magnetic reconnection at a time close to substorm onset, as described above. The magnetic reconnection of open (lobe) field lines leads to a configuration change from originally closed field lines to a magnetic island, or flux ropes, on the tailward side of the neutral line. Since it is no longer connected magnetically to the Earth, the plasmoid can freely propagate down the tail (see also figure 1 in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL). The tailward propagation velocity is ∼300 km s^{-1} in the near-Earth region, subsequently accelerating to ∼700 km s^{-1} at ∼100 R_E downtail. Then, it slows down to ∼500 km s^{-1}. It is noted that the initial tailward speed is only a fraction of the lobe Alfvén velocity, but, beyond ∼100R_E, it becomes comparable with, or a little lower than, the Alfvén velocity in the ambient lobe region. The plasmoid is initially confined in a narrow region near the tail axis, but it expands during the tailward propagation, until at ∼100R_E the dawn–dusk extent reaches the magnetopause, while the length (in the Sun–Earth direction) reaches several tens of R_E.

Plasma sheet boundary layer

Energetic ion beams are often observed to stream along the magnetic field direction near the plasma sheet–lobe boundary. This region is called the plasma sheet boundary layer (PSBL). In the near-Earth region, these energetic field-aligned beams are directed earthward and precipitate into the ionosphere near the poleward boundary of the auroral oval on the night side. A part of them are reflected back to the tail by the magnetic mirror force and constitute counterstreaming beams. While it is generally interpreted in terms of current sheet acceleration for the near-Earth observations, their generation mechanism is most likely to be associated with magnetic reconnection. Beyond several tens of R_E down the tail, they are observed just outside of a region in which the β value jumps from $\ll 1$ to a fraction of unity. They often coexist with cold ions of the lobe origin. This is a fore-shock region of the slow shocks formed by magnetic reconnection.

Slow shocks

With the formation of an X-type neutral line, cold plasmas in the lobes are convected toward the plasma sheet with the $\boldsymbol{E} \times \boldsymbol{B}$ drift under the action of the dawn-to-dusk electric field. The velocity of a slow-mode wave approaches zero as the propagation direction becomes perpendicular to the magnetic field. Therefore the plasma flow becomes supersonic with respect to the slow-mode wave in the vicinity of the X-type neutral line, even when the inflow velocity of the plasma is small. As a result, the collision of plasmas flowing from the northern and southern lobes is expected to give rise to slow shocks. The lower panel in figure 1 shows a schematic view of the structure of a slow-mode shock in the Earth's magnetotail. On entering the plasma sheet, cold plasmas are heated and accelerated by the slow shock. In the fore-shock region, backstreaming energetic ions and incident cold ions coexist, and the cold ions are preheated before crossing the slow shock. These features are evidenced in the geomagnetic tail by recent spacecraft observations (although there are a considerable number of other cases in which the slow-shock condition cannot be satisfied).

Current sheet acceleration

Let us consider the motion of a single particle near the neutral sheet under the cross-tail (dawn-to-dusk) electric field in a tail-like magnetic field configuration, in which the magnetic field magnitude decreases toward the neutral sheet. As they approach the neutral sheet owing to the $\boldsymbol{E} \times \boldsymbol{B}$ drift, particles are largely accelerated by the cross-tail electric field and begin to meander in the dawn–dusk direction (duskward for ions and dawnward for electrons) near the neutral sheet. They are further accelerated to higher energies with the meandering cycles. This is called current sheet acceleration.

This process may generally occur everywhere in the neutral sheet with the dawn-to-dusk electric field, but it appears most prominently in a thin current sheet near the diffusion region of magnetic reconnection. Recent *in situ* observations from satellites have revealed the existence of non-gyrotropic ion distribution functions as evidence of the acceleration in the thin current sheet, of which the thickness is estimated as $\sim$ 1000 km ($<$ 500 km in extreme cases), that is comparable with, or smaller than, a typical ion Larmor radius.

Sources of the plasma sheet

Potential sources of the plasma sheet are the solar wind and the Earth's ionosphere. According to the reconnection–convection model, the solar wind (magnetosheath) plasma enters the magnetosphere via dayside reconnection and is transported to the mantle region in the geomagnetic tail. It also continues to enter the mantle at the open-tail magnetopause. Then a portion of the mantle plasma is convected into the plasma sheet. (See also MAGNETOSPHERE OF EARTH: CONVECTION, MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL for the reconnection–convection model.) Plasmas (especially ions) in the topside ionosphere are energized so as to escape both from the polar cap and from the auroral region. Particles escaping from the dayside cusp and the polar cap are transported along magnetic field lines to the mantle and lobe region, and trace similar trajectories to those of solar wind origin. Upflowing ion beams accelerated above the auroral oval are directly injected into the near-Earth plasma sheet. Generally the ionospheric plasma makes a significant contribution to the near-Earth plasma sheet during geomagnetically active times, while the solar wind plasma is usually the major source.

Although the structure and dynamics are reasonably interpreted in terms of the reconnection process, it is still a matter of controversy whether this process is always the most important for plasma supply to the plasma sheet population. In the plasma sheet, there is a general tendency for the density to increase and the temperature to decrease as geomagnetic activity drops to a quiescent level. This may suggest that a diffusive process in the flank magnetopause is more effectively transporting cold dense solar wind plasma directly into the plasma sheet during quiet times, in which case the reconnection process may be considered to be less effective.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It should be noted that the ion temperature in the plasma sheet is several times higher than the solar wind kinetic energy (∼1 keV) and several orders of magnitude higher than the temperature of the ionospheric plasma. The temperature of the lobe plasma just outside the plasma sheet is as low as several tens of eV, while the ion temperature in the central plasma sheet is typically around several keV, and the electron temperature is 1/4–1/7 of the ion temperature. Efficient heating processes must take place across the plasma sheet–lobe boundary and inside the plasma sheet. Possible candidates are the slow shocks at the plasma sheet boundary, the current sheet acceleration and the resistive dissipation in the magnetic diffusion region, but their quantitative assessment is yet to be done in future work.

Bibliography

See papers in the special section on magnetotail dynamics in *J. Geophys. Res.* **103** (A3). For example:

Hoshino M, Mukai T, Yamamoto T and Kokubun S 1998 Ion dynamics in magnetic reconnection: comparison between numerical simulation and geotail observations *J. Geophys. Res.* **103** 4509–30

Ieda A *et al* 1998 Statistical analysis of the plasmoid evolution with geotail observations *J. Geophys. Res.* **103** 4453–65

Nagai T *et al* 1998 Structure and dynamics of magnetic reconnection for substorm onsets with geotail observations *J. Geophys. Res.* **103** 4419–40

Nishida A, Baker D N and Cowley S W H 1998 *New Perspectives of the Earth's Magnetotail (Geophysical Monograph Series 105)* (Washington, DC: American Geophysical Union)

Nishida A, Mukai T, Yamamoto T, Kokubun S and Maezawa K 1998 A unified model of the magnetotail convection in geomagnetically quiet and active times *J. Geophys. Res.* **103** 4409–18

Toshifumi Mukai

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Plasmasphere

The plasmasphere is a torus of cold, dense plasma encircling the Earth that usually exhibits a very sharp edge at a distance ranging between 2 and about 6 Earth radii in the equatorial plane. Figure 1 is a simplified diagram of the plasmasphere as viewed from the Sun, showing some of its main features: the plasmapause, the main body of the plasmasphere, the dusk-bulge region and a detached plasma region outside the main body of the plasmasphere. The plasmasphere is essentially an upward extension of the medium-to-low-latitude ionosphere, formed as ionospheric ions flow upwards from their creation points at low altitudes to fill the relative vacuum of space surrounding the Earth. The sharp edge called the plasmapause was entirely unexpected. Here densities drop by a factor of 10–100 in a relatively short distance. The plasmapause was discovered in 1963 nearly simultaneously by Konstantin Gringauz in the Soviet Union, using lunar rocket observations of the thermal plasma, and by Donald Carpenter in the United States, observing very low frequency plasma waves called whistlers.

These whistlers, generated by lightning flashes, propagated between northern and southern hemispheres along geomagnetic field lines on both sides of the plasmapause.

The plasma density ranges from 10^4 cm^{-3} just above the topside ionosphere at ~1000 km altitude to 10^2–10^3 cm^{-3} within and 10–10^2 cm^{-3} just outside the plasmapause in the equatorial plane. The average radius of the plasmasphere varies in response to changing SOLAR-WIND conditions. In addition, the plasmasphere extends to larger radial distance in the afternoon–dusk sector in a region termed the dusk bulge. These two characteristic features of the plasmasphere have been very successfully explained by considering the influences of the magnetospheric electric fields on plasma drifts. Erosion of the plasmasphere in the dusk sector appears to produce outlying structures that may be attached plumes or high-density regions completely detached from the main body of the plasmasphere.

Plasmaspheric composition and temperature

The plasmasphere is composed mostly of protons and for this reason was originally called the protonosphere. The base of the plasmasphere is taken as the altitude (~1000 km) at which protons replace oxygen as the dominant species in the ionospheric plasma. Small amounts of helium and oxygen ions (1–5% oxygen and 2–20% helium) as well as trace amounts of singly charged nitrogen and deuterium and doubly charged helium and oxygen ions are also found in the plasmasphere. Satellite observations indicate that, on occasion, an enhanced torus of heavy ions (O^+ and O^{++}) appears in the outer plasmasphere in the region of a newly established plasmapause.

Ion temperatures in the plasmasphere are generally between 0.5 eV and several eV (1 eV = 11 700 K). In the outer plasmasphere, temperatures are, for the most part, anticorrelated with the density changes, rising sharply as the densities decline, moving toward the plasmapause and beyond. It is not clear whether ion and electron temperatures are in equilibrium in the outer plasmasphere. In fact, observations suggest that ion temperatures may exceed electron temperatures in this region. In the outer plasmasphere, the cold plasma temperature varies with magnetic and solar activity, reflecting the corresponding changes in its heat sources. One of the major heat sources is the terrestrial ring current, which is also found in this region of space (see MAGNETOSPHERE OF EARTH: RING CURRENT). The energetic plasma of the ring current overlaps the outer regions of the plasmasphere and dusk bulge, transferring energy to the thermal plasma through Coulomb collisions, and the damping of plasma waves. The time dependence of this energy transfer, its variation with solar and magnetic activity and its relationship to the changing composition and energy characteristics of the ring current are key areas of research at the present time.

The structure of the plasmasphere

Cold plasmas, such as those that form the plasmasphere, are strongly influenced by the large-scale electric field in the magnetosphere. This electric field drives a circulation of magnetic field lines and cold plasmas along trajectories that approximately coincide with electric field equipotentials. The large scale electric field in the inner magnetosphere originates from the superposition of two major sources: (1) the corotation electric field, resulting from the Earth's rotation, which points radial inward at all local times, and (2) the convection electric field, generated by the solar-wind–magnetosphere interaction, assumed to be constant and directed from dawn to dusk throughout the entire region. Figure 2 shows electric potential patterns (and thus cold plasma drift trajectories) from each of these sources along with the combined potential from their superposition. The electric potential patterns are drawn in the magnetospheric equatorial plane with the Sun to the bottom of the figure. In steady state, the plasmapause corresponds to the boundary between the closed drift paths encircling the Earth and the open drift paths moving plasma from the magnetotail to the dayside magnetopause (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). Within the plasmasphere, magnetic field lines, that confine the plasma, approximately rotate with the Earth allowing high densities to build up. Outside the plasmasphere, the circulation of magnetic field lines is influenced strongly by the convection electric field. The resultant trajectories of these external field lines no longer enclose the Earth. Magnetic field lines external to the plasmasphere convect to the magnetopause boundary on the dayside of the Earth, sweeping their burden of upflowing ionospheric plasma out of the magnetosphere. On these field lines, a buildup of upflowing plasma from the ionosphere, comparable to that achieved within the plasmasphere,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

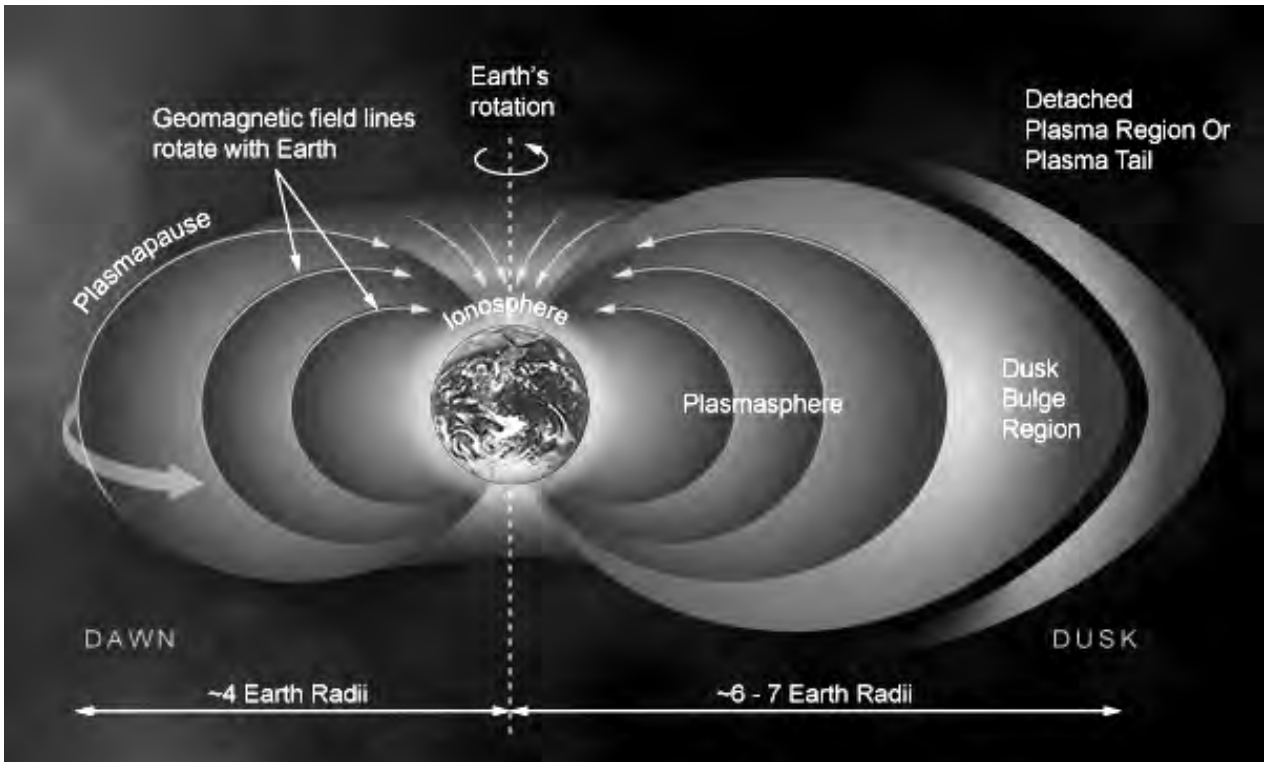


Figure 1. View from the Sun of the Earth's plasmasphere. The plasmasphere is a torus-shaped region of cold, dense plasma surrounding the Earth in the near-space environment that has a very sharp edge called the plasmapause at equatorial distances of 4–6 Earth radii. It is essentially an extension of the ionosphere to high altitudes. The plasmasphere bulges out to greater radial distances on the dusk side of the Earth due to the nature of the plasma drift paths. However the dusk bulge region appears to be a separate region from the main body of the plasmasphere. Within the dusk bulge region, the plasma contains many structures, which may by blobs of plasma detached from the main body of the plasmasphere or plumes of plasma still attached by thin tails in another local time sector.

cannot occur. This difference in field-line circulation patterns accounts for the significant density difference between the plasmasphere and the so-called plasma trough region beyond. Solar-wind-generated electric fields vary continuously in strength while the corotation electric field remains constant to a first approximation. Since the size and shape of the plasmasphere depend upon the interplay of these electric fields, the mean radius of the plasmapause tends to diminish as the strength of the convection field increases.

In steady state, the dawn–dusk asymmetry in the drift paths produces an observed extension of the plasmasphere to larger radial distances on the dusk side, called the dusk-bulge region. However steady state is rarely achieved in practice. Erosion of the outer plasmasphere occurs with timescales of the order of a day and refilling takes several to 10 d. The solar wind, the source of the convection electric field, does not remain constant for such long periods. Although large-scale plasmaspheric features are explained to some extent by the potential patterns given in figure 2, inner magnetospheric electric fields are seldom this simple and plasmaspheric features exhibit interesting smaller-scale structures. The key process that complicates this picture is the tendency for energetic ions drifting at the inner edge of the plasma sheet to create electric fields through charge separation that shield the convection electric field from regions inwards of their location (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). Recent satellite observations indicate that the inner magnetosphere electric fields have quite complicated spatial distributions as a consequence of shielding. Models indicate that shielding is time-dependent (being noticeable after 10 min but taking several hours to complete) and not always effective. Because of these properties, the electric field penetrates very close to Earth at times, and complicated electric field patterns are created in the region of the outer plasmasphere and plasmapause.

The dusk bulge region is also rather more complicated than the steady-state picture implies. Observations indicate that the main plasmasphere is nearly circular, with only a slight bulge near dusk. During periods of enhanced convection activity, erosion of the plasmasphere in the dusk sector appears to produce outlying structures. It is not clear from satellite data whether these structures are plumes, which extend from the dusk region, or are completely detached from the main plasmasphere. Meanwhile, a new plasmasphere boundary is formed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

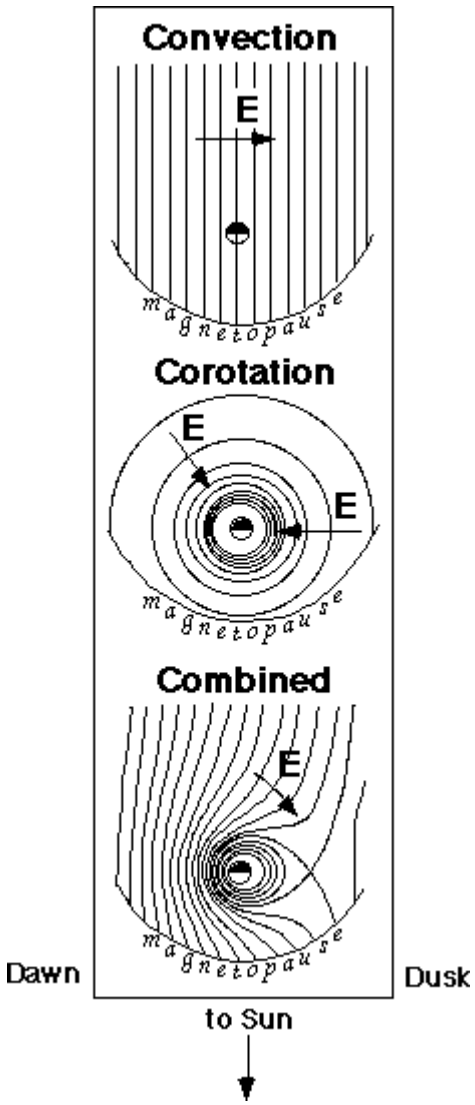


Figure 2. Lines of constant electric potential are shown looking down on the Earth in the equatorial plane with the Sun at the bottom. Since cold plasma moves along constant equipotentials in the magnetosphere, these lines also represent plasma drift paths. In this simplified representation (which nonetheless captures the most important elements of plasmaspheric structure), the pattern of the drift paths is determined by the superposition of two key large scale electric fields: the convection and corotation electric fields. The convection electric field is directed from dawn to dusk (left to right above) and has constant amplitude throughout the inner magnetosphere in this simplified model. It is produced by the solar-wind–magnetosphere interaction and drives plasma from the night side to the dayside magnetopause boundary. The corotation electric field points radially inward toward the Earth at all local times and forces cold plasma to approximately move with the Earth as it rotates. The superposition of the potentials from these two fields produces the combined pattern. In steady state, the boundary between open and closed drift paths is the plasmapause. A pronounced dawn–dusk asymmetry, seen in the drift paths, is responsible for the dusk bulge region of the plasmasphere.

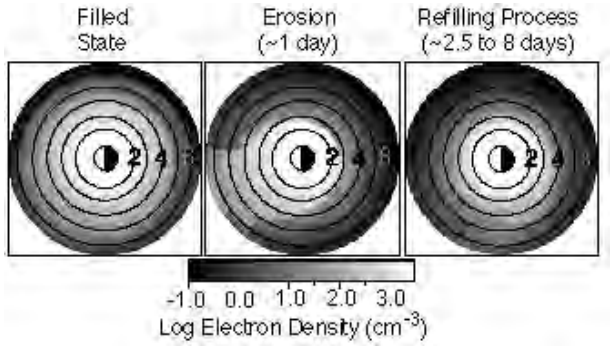


Figure 3. Model view of the plasmasphere during erosion and recovery, looking down on the Earth in the equatorial plane, with the Sun to the left. The plasmasphere is initially completely filled and in equilibrium with the underlying ionosphere (left panel). During erosion the convection electric field is enhanced which causes the boundary between closed drift paths, that encircle the Earth, and open drift paths, that move plasma to the dayside magnetopause boundary, to move inward. Plasma that was initially on closed drift paths now finds itself on open drift paths. This plasma is stripped off the plasmasphere and swept to the magnetopause boundary layers (center panel). As magnetic activity quietens, the convection electric field weakens and the open/closed drift path boundary moves outward. Newly corotating but empty plasma regions begin to refill immediately as plasma moves upward along field lines from the ionosphere (right panel). The circles mark the radial distance every 0.5 Earth radii moving outward from the center of the Earth.

closer to the Earth on the nightside under the influence of the enhanced convection electric field (by mechanisms yet to be observed). In a fashion similar to that of a lathe, the new nightside boundary radius appears to be communicated to the dayside and eventually to the dusk sector through approximate rotation of the affected nightside regions with the Earth, leading (statistically) to a roughly circular condition of the main plasmasphere. Finally, the difference in field-line circulation patterns, which accounts for the significant density difference between the plasmasphere and the so-called plasma trough region beyond, does not necessarily explain the abruptness of the density change at the plasmapause that is often observed. The mechanisms at work in plasmasphere boundary formation remain a topic of research.

Erosion and recovery of the plasmasphere

The plasmasphere goes through continuous cycles of erosion and recovery. Figure 3 is a simulation of one such cycle. When the convection electric field strengthens, the region of closed trajectories encircling the Earth decreases in size. The plasmapause, which roughly marks the outermost boundary of this region of closed trajectories, moves inward. Thermal plasma, formerly trapped on closed trajectories, now finds itself on open drift paths. It is peeled off the plasmasphere and moved along open drift paths, to be lost at the magnetopause boundary. The number of ions lost from outside a newly formed plasmapause boundary, that has moved inward from a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

distance of 5 to 3.5 Earth radii, is of the order of 1×10^{31}. When the convection electric field weakens, the plasmapause moves outward again trapping a relatively empty plasma region on corotating drift paths. Refilling of this region from the ionosphere, through magnetic-field-aligned plasma flows, begins immediately. The time required to build up high plasma densities in the outer plasmasphere through this refilling process, increases with radial distance because of the accompanying increase in the volume enclosed by the magnetic field lines. Refilling times range from ∼1 d at 2.5 Earth radii to ∼8 d at 4 Earth radii. As a result of this radial increase in refilling times, the plasmasphere tends to behave as if it were separated into two distinct regions: an inner region which is completely refilled on short timescales, and an outer region which is almost always in the process of refilling but never quite reaches equilibrium before the next erosion episode begins. Time-varying electric fields further complicate this process by producing plumes of plasma still attached to the main body of the plasmasphere that extend from the dusk plasmasphere to the dayside magnetopause boundary. In addition, numerous small-scale regions of enhanced density are observed between the plasmasphere and the magnetopause in the dusk region during times of enhanced activity. There is evidence to suggest that these cold plasma structures remain trapped for extended periods, presumably on open drift paths, implying that the processes by which plasma is swept out of the dayside magnetopause are not fully understood. There is also new evidence from whistler wave observations to suggest that plasmapause formation may partly involve a process that dumps plasma (∼50% of the amount lost at the magnetopause) from the corotating outer portions of the eroded plasmasphere into the underlying ionosphere. The roles of instabilities in enhancing losses of plasma at the plasmapause boundary and in creating irregular plasma structures have not yet been resolved.

Recent interest has focused on the response of the plasmasphere to induction electric fields that occur during magnetic substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS). These induction fields can sometimes be even stronger than convection electric fields generated by the interaction between the solar wind and the Earth's magnetosphere. They are produced by reconfigurations of the Earth's magnetotail. During the growth phase of substorms, magnetic field lines become more stretched producing induction fields that are directed eastward opposite in direction to the westward convection electric field. When substorm onset occurs, the Earth's magnetotail field lines snap back into a more dipolar configuration producing induction fields that point westward and add to the convection electric field (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL). These fluctuations in the large-scale electric field play a role in the production of detached plasma regions and turbulence in the outer plasmasphere.

There is still one additional source of electric fields in the inner magnetosphere that has a strong influence on the dynamics and structure of the plasmasphere. Neutral-atmospheric winds, through collisions, drive ions across magnetic field lines, creating electric fields. These electric fields which are generated in the ionosphere map out along magnetic field lines to high altitudes and alter the potential pattern in the magnetosphere and with it the cold plasma drift paths. There are three major sources of neutral winds that are important in altering magnetospheric electric fields, including (1) solar heating which maximizes on the dayside at the subsolar point and drives winds antisunward from the dayside to the nightside, (2) weather disturbances and topology at the Earth's surface which drive gravity waves that propagate upward to the ionosphere, and (3) auroral heating which drives gravity waves towards the equator from the polar regions. Electric fields from each of these neutral-wind systems impact the large-scale electric field in the vicinity of the outer plasmasphere and plasmapause region. The interplay between plasmaspheric erosion and refilling during disturbed periods and the neutral-wind system is not yet fully understood. Progress in realistic modeling of the plasmasphere requires a complete description of the inner magnetosphere and its coupling to the ionosphere–neutral-atmosphere system as well as to the geomagnetic tail.

Bibliography

Carpenter D L and Lemaire J 1997 Erosion and recovery of the plasmasphere in the plasmapause region *Space Sci. Rev.* **80** 153–79

Lemaire J F and Gringauz K I 1998 *The Earth's Plasmasphere* (Cambridge: Cambridge University Press)

Moldwin M B 1997 Outer plasmaspheric plasma properties: What we know from satellite data *Space Sci. Rev.* **80** 181–98

Wolf R A and Spiro R W 1997 Numerical modeling of the ring current and plasmasphere *Space Sci. Rev.* **80** 199–216

Janet Kozyra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Radiation Belts

The Van Allen radiation belts, first discovered by the EXPLORER I satellite in 1958, consist of charged particles in orbits determined by the Earth's magnetic field (see VAN ALLEN BELTS). The Earth's magnetic field is approximately a dipole, see figure 1. Both electrons and ions which make up the radiation belts have typical energies of one million electron volts (1 MeV), and execute three different types of periodic (oscillatory) motion about the magnetic field. They gyrate about magnetic field lines at the electron cyclotron frequency, $\omega = qB/m$, which depends on the local magnetic field strength, decreasing like $1/r^3$ for a dipole. Those with finite velocity parallel to the Earth's magnetic field bounce between hemispheres with a frequency which is proportional to their parallel velocity, and inversely proportional to the length of magnetic field lines. Lastly, they drift around the Earth with a frequency which is proportional to their perpendicular energy, and inversely proportional to the length of magnetic field line about which they gyrate.

Source and loss processes

There are two separate zones of electrons, called the *inner* and *outer zone*, while radiation belt protons have only an inner zone on average. Heavier ions of atmospheric and SOLAR WIND origin, notably oxygen and helium, contribute to the radiation belts; however their percentage of composition is small. Another recently discovered component comes from the trapping of COSMIC RAYS which enter the magnetosphere singly charged, become stripped of outer electrons as they interact with the neutral atmosphere, and trapped by the Earth's magnetic field. This so-called ANOMALOUS COSMIC RAY belt contains helium, nitrogen, oxygen and neon nuclei. The primary source of inner zone protons comes from the decay of neutrons produced by cosmic rays scattered off the atmosphere. This process is called *cosmic ray albedo neutron decay* (CRAND), and produces electrons as well as protons as a neutron decays with a 10.5 min half-life. Neutrons produced by scattering of energetic cosmic rays off the atmosphere are themselves energetic (MeV range), and most of their energy goes into the proton in the decay process, while the electron is emitted in a random direction with a few hundred keV energy. These electrons can, like the protons, be trapped by the Earth's magnetic field, but they are inadequate as a source of MeV electrons in either inner or outer zones. A loss process due to the interaction of electrons with electromagnetic waves, described below, is responsible for the existence of two separate zones for electrons. This loss mechanism becomes comparable to the rate at which electrons are supplied to the region around 3 R_E geocentric distance in the magnetic equatorial plane, producing a *slot region* between inner and outer zones (see review by Walt (1996)).

The source of electrons for both inner and outer zones is primarily at larger radial distances, outside the orbit of geosynchronous spacecraft which orbit the Earth at a radial distance of 6.6 R_E with a period of 24 h. A major storage region for plasma or ionized gas consisting primarily of electrons and protons in the magnetosphere is the *plasma sheet* in the geomagnetic tail (see MAGNETOSPHERE OF EARTH: PLASMA SHEET; GEOMAGNETIC TAIL). The Earth's magnetic field is stretched into a long windsock in the anti-sunward direction by momentum coupling of solar wind flow past the magnetosphere, with a reversal of the magnetic field from outward in the southern hemispheric tail lobe to inward or earthward in the northern. There is a balance of magnetic and kinetic particle pressure, described by the equations of MAGNETOHYDRODYNAMICS (MHD), which leads to confinement of plasma sheet electrons and ions in the region of magnetic field reversal where the magnetic pressure is weakest. Some of this plasma originates in the solar wind, making its way into the magnetosphere in the tail region where the magnetic field is weak, where plasma of solar wind origin is identifiable by different ion mass and charge states from protons. Another source of plasma is the Earth's ionosphere. Plasmasheet electrons have an average energy ~600 eV, with ion energies typically greater by a factor of seven. Both solar wind flow and MHD flows in the magnetosphere impart greater energy to ions because of their greater mass (scaling as $\frac{1}{2}mv^2$), with electrons and ions drifting together transverse to the magnetic field B in most regions of the magnetosphere. Both charge species are energized as they are transported earthward by the two-cell *convection* that provides return flow towards Earth in the central tail region, driven by anti-sunward flow at the boundary (see MAGNETOSPHERE OF EARTH, figure 1). In detail, processes such as *reconnection* of oppositely directed magnetic field lines in the tail impart magnetic field-aligned flow and thermal kinetic energy to the plasma, and play a role in large-scale reconfiguration of the magnetotail, from a stretched–stressed shape to a more dipolar shape, called *substorms* (see MAGNETOSPHERE OF EARTH: SUBSTORMS). Impulsive electron and ion injection at geosynchronous orbit due to substorms is well documented, and provides a major outer boundary source for energetic particles in the inner magnetosphere, along with steady, sunward convection. This is the main source for radiation belt electrons, as well as the low energy ion population in the 100 keV energy range which makes up the Earth's *ring current* (see MAGNETOSPHERE OF EARTH: RING CURRENT). The latter is well distinguished from the proton radiation belts by energy, radial location with maximum flux around 4 R_E geocentric, and an outer versus inner boundary origin.

There is an approximate constant of motion associated with each type of periodic particle motion: gyro, bounce and longitudinal drift. These are called *adiabatic invariants*, and the first conserves the magnetic moment of a particle, which is the product of the equivalent current carried by a particle gyrating about a magnetic field line times the area enclosed by the orbit. The second invariant is proportional to the parallel momentum, mass times parallel velocity,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

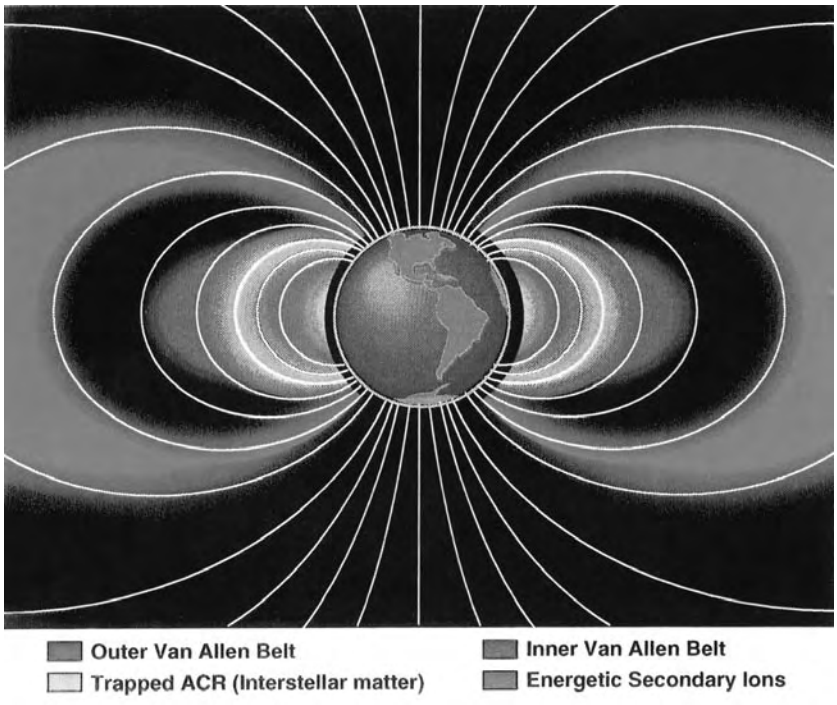


Figure 1. Dipole representation of the Earth's magnetic field showing proton inner zone, inner and outer zone electron radiation belts, and the anomalous cosmic ray component. **This figure is reproduced as Color Plate 34.**

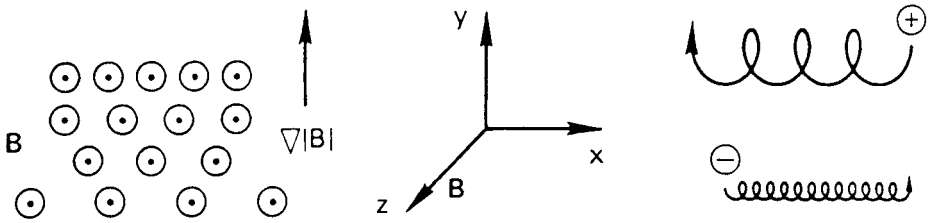


Figure 2. Drift of a gyrating particle in a non-uniform magnetic field directed out of the page. Radial (earthward) gradient points up, ions drift westward and electrons eastward due to changing magnetic field strength across their gyro orbits.

and the length of the magnetic field line along which bounce motion is executed. The third invariant is the magnetic flux enclosed by a complete longitudinal drift path, which is the integral of the magnetic field times the area enclosed, $\int B\,\mathrm{d}a$. The principal mechanism for energizing radiation belt electrons as they are transported radially inward involves conservation of the first and second adiabatic invariants. Thus, as an electron gyrating about a magnetic field line moves earthward in the convection flow indicated in figure 1, into a region of increasing magnetic field strength, its perpendicular energy must increase in order to conserve the *first adiabatic invariant*. A charged particle's longitudinal drift in the radial magnetic field gradient of a dipole field depends on the sign of the charge, as shown in figure 2, providing the source of the ring current as well as radiation belt particle longitudinal drifts, with ions drifting westward and electrons eastward. Fluctuations in electric and magnetic fields, e.g. due to low frequency MHD waves, give rise to radial diffusion of charged particles represented by their distribution f in velocity and position, which on average transports particles from their source to loss region. This process can be described by a diffusion equation of the form

$$\frac{\partial f}{\partial t} = L^2 \frac{\partial}{\partial L}\left(\frac{1}{L^2} D_{LL} \frac{\partial f}{\partial L}\right)$$

when the radial motion is so small, while the particle is interacting with a fluctuating electric or magnetic field, that one can neglect the change in its longitudinal drift velocity. Particles are assumed to be uniformly distributed in longitude, and remain so during the diffusion process. Neither of these assumptions is valid during a substorm injection, nor during the compression of the outer boundary of the magnetosphere by an interplanetary shock wave which originates from the Sun. A third implicit assumption is that, while particle distributions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can be expressed as functions of the six dimensions of velocity and position coordinates, diffusion in velocity and position are decoupled processes. The latter is often a good assumption, because *radial diffusion* occurs on the timescale of the drift period, while velocity diffusion occurs on the timescale of particle bounce and gyromotion. Under these assumptions, the following two radial diffusion coefficients have been calculated for fluctuations in the magnetic and electric field, respectively:

$$D_{LL}^{M} = \frac{\pi^2}{2}\left(\frac{5}{7}\right)^2 \frac{R_E^2 L^{10}}{B_0^2} \nu_{\text{drift}}^2 P_A(\nu_{\text{drift}}) \quad (1)$$

$$D_{LL}^{E}(L, \nu_{\text{drift}}) = \frac{L^6}{8R_E^2 B_0^2} \sum_{n=1}^{n} P_n(L, n\nu)_{\nu=\nu_{\text{drift}}}. \quad (2)$$

These coefficients are written as functions of the surface magnetic field strength at the Earth's equator, $B_0 = 0.3$ G, the geocentric radial distance at the equator L in units of earth radius R_E, the power spectrum P_A of the asymmetric part of the magnetic perturbation at the particle longitudinal *drift frequency* ν_{drift}, and the nth Fourier component of the global electric field perturbation P_n, for magnetic and electric field-driven radial diffusion, respectively. While the *magnetic diffusion coefficient* is written for equatorially confined particles, and is substantially reduced for those in bounce motion off the equator, the *electric diffusion coefficient* is independent of the equatorial *pitch angle* of a particle, which is the angle between the velocity vector and the magnetic field at the equator. Electric field-driven diffusion tends to predominate in transporting particles to the low L values of the radiation belt flux peaks inside geosynchronous orbit. Diffusion coefficients which include both theoretical calculations of the power spectral functions P, and calculations based on local measurements of P, all decrease rapidly with decreasing L because of the strong L dependence in (1) and (2). There is substantial spread, but all have the behavior that diffusion times are greater than a day into $L = 4$.

The primary loss processes for radiation belt particles are collisions with the neutral atmosphere and interaction with right (left) hand circularly polarized electromagnetic waves which propagate along the magnetic field such that an electron (ion) sees the rotating electric field vector of the wave remain constant at its gyrofrequency. This electric field can then cause particle diffusion in velocity which changes either the particle's energy, pitch angle, or both. Particles which diffuse toward more parallel motion will strike the atmosphere and be collisionally lost, e.g. via *charge exchange* with a neutral. Thus diffusion in velocity space acts to move particles away from more perpendicular pitch angles, which result from increasing their perpendicular velocity as they diffuse radially inward in the Earth's magnetic field, toward pitch angles with increased collisional interaction with the neutral atmosphere. This type of wave–particle interaction sets a theoretical limit on the flux of radiation belt particles called the *stable trapping limit*, since the velocity space diffusion coefficient is proportional to wave power near the electron and ion gyrofrequencies and harmonics, which in turn is proportional to the density of particles and gradients in velocity space away from the peak at 90° to the magnetic field. This mechanism has been generalized to nonparallel propagation, and was highly successful in explaining the electron slot region which separates the inner and outer radiation belt zones.

Non-diffusive processes

The preceding, steady-state view of radiation belt processes has been accepted for three decades, and shown to be successful in predicting particle flux levels during geomagnetically quiet periods. However, it was learned during the 1989–91 maximum in solar SUNSPOT activity that there are significant deviations from this quiescent picture. For example, on 24 March 1991, an interplanetary shock initiated by a *coronal mass ejection* (CME) at the Sun, a global scale detachment of plasma from the outer atmosphere or corona of the Sun (see SOLAR CORONAL MASS EJECTION), impacted the Earth's magnetosphere at a velocity estimated to be as high as 1400 km s^{-1} and compressed the outer boundary of the magnetosphere inside the orbit of geosynchronous spacecraft for several hours. This impact launched an MHD pulse called a *magnetosonic wave*, which spread around the flanks of the magnetosphere at a speed comparable to the drift velocity of outer zone electrons between geosynchronous orbit and $L = 9$–10. Electrons drifting eastward synchronously with the wave pulse saw an approximately constant azimuthal electric field, which follows from Faraday's law and the time-dependent compression of the Earth's dayside magnetic field. This electric field transported electrons resonant with the pulse radially inward in the direction of $E \times B$, forming a flux peak at $L = 2.5$, normally the slot region, with an energy peak at 13 MeV. Electron fluxes increased in this region by several orders of magnitude, and remained elevated for well over a year. Simultaneously, solar energetic protons which originate both at the Sun during the coronal mass ejection, and are accelerated by interaction with the *interplanetary shock* as it propagates toward the magnetosphere, formed a new, trapped proton radiation belt, when they were accelerated inward by the same azimuthal electric field which formed the new electron belt. While the March 1991 event produced the largest radiation belt perturbation of the Sunspot Cycle 22 solar maximum (1989–91), others were seen which require a similar explanation that violates the steady-state radial diffusion assumptions.

Another important type of variability in outer zone electron fluxes is associated with *high speed solar wind stream* interactions with the magnetosphere. As far back as 1965, during the first minimum in sunspot activity following the solar maximum of the International Geophysical Year, the Explorer 26 satellite observed regular monthly injections and decay in outer zone electron fluxes which were later confirmed to be correlated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with high speed solar wind flows. A correlation was also noted with southward orientation of the interplanetary magnetic field (IMF), which in figure 1 (MAGNETOSPHERE OF EARTH) is seen to favor enhanced reconnection of the solar wind magnetic field with the Earth's northward field at low latitudes on the dayside. Southward IMF enhances the coupling of solar wind flow energy which drives convection and substorm activity internally to the magnetosphere. The regularity of outer zone electron flux variation seen at *sunspot minimum* is due to the orderly structure of the *solar current sheet*, associated with solar wind outflow from the heliomagnetic equator which may cross the Earth's orbit once or twice per solar rotation. Superimposed upon such recurrent behavior are CMEs, which become more prevalent approaching *sunspot maximum* when solar activity in general increases. CMEs can produce a detached plasma region with large-scale organized helical magnetic field structure called a *magnetic cloud*, carried radially outward by the ambient solar wind flow. Magnetic clouds impacting the magnetosphere can provide an extended period of southward IMF favorable to enhanced convection and substorm injection of plasma sheet particles into the inner magnetosphere. Such a sequence of events was evident for the 6–11 January 1997 magnetic cloud interaction with the magnetosphere, which produced prolonged substorm activity, increase in the ring current which identifies a *geomagnetic storm*, and significant enhancement of outer zone electron fluxes (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS). An increase in relativistic electron fluxes by several orders of magnitude around $L = 4$ was observed on the timescale of a few hours, which cannot be explained by radial diffusion. Observations of *ultra low frequency* (*ULF*) oscillations in the mHz frequency range, which corresponds to the longitudinal drift frequency of hundred keV electrons at geosynchronous orbit and MeV electrons at $L = 4$, suggest that *drift-resonant acceleration* by the electric field associated with these ULF oscillations may play an important role in transporting the electrons radially inward on the timescale of a few electron drift periods.

A dynamic view of the radiation belts is necessary to explain the variations in particle fluxes by orders of magnitude which can occur on a timescale of minutes to hours, modifying integrated fluxes to which spacecraft are exposed over much longer periods. New multipoint spacecraft and ground-based diagnostics and computer modelling are expected to greatly alter the static view of the radiation belts shown in figure 1, and provide radiation belt environment forecasting capability (*space weather* prediction) of importance to satellite systems and manned presence in space.

Bibliography

Hudson M K, Elkington S R, Lyon J G, Marchenko V A, Roth I, Temerin M, Blake J B, Gussenhoven M S and Wygant J R 1997 Simulations of radiation belt formation during storm sudden commencements *J. Geopphys. Res.* **102** 14 087–102

Hudson M K, Elkington S R, Lyon J G, Marchenko V A, Roth I, Temerin M and Gussenhoven M S 1996 MHD/particle simulations of radiation belt formation during a storm sudden commencement *Radiation Belts Models and Standards* ed J F Lamaire, D Heynderickx and D N Baker (Washington, DC: American Geophysical Union) pp 57–62

Hudson M K, Elkington S R, Lyon J G, Goodrich G C and Rosenberg T J 1999 Simulation of radiation belt dynamics driven by solar wind variations *Sun–Earth Plasma Connections* ed J L Burch, R L Carovillano and S K Antiochos (Washington, DC: American Geophysical Union) pp 171–2

Hughes W J 1995 The magnetopause, magnetotail and magnetic reconnection *Introduction to Space Physics* ed M K Kivelson and C T Russell (Cambridge: Cambridge University Press) pp 227–87

Walt M 1996 Source and loss processes for radiation belt particles *Radiation Belts Models and Standards* (Washington, DC: American Geophysical Union) pp 1–13

Wolf R A 1995 Magnetospheric configuration *Introduction to Space Physics* (Cambridge: Cambridge University Press) pp 288–329

Mary K Hudson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Ring Current

The ring current consists of energetic positive ions that drift westward through the Earth's inner magnetosphere. That ring of westward current causes a southward perturbation in the magnetic field on the Earth's surface. The ring current is always present, but it intensifies during a magnetic storm.

The discovery of the ring current

In the nineteenth century, scientists noticed that the horizontal magnetic field strength occasionally decreases over most of the Earth, often after times when strong activity has been observed on the solar surface. It seemed as if the Earth sat within a great ring of westward current that strengthened during times of intense solar activity. Thus the idea of an Earth-centered ring current arose long before the space age and long before the development of modern plasma physics. However, in those early days, there was no sure way to determine how electrical current could flow through the space beyond the Earth's atmosphere.

A 'magnetic storm' is defined by a worldwide decrease in the horizontal magnetic field strength on the Earth. Nowadays, the effect is quantified in terms of the *Dst* (storm disturbance) magnetic index, which is an average of the northward magnetic perturbation observed at several low-latitude ground stations, with effects of quiet-time neutral winds carefully subtracted. In a typical magnetic storm, the *Dst* index decreases sharply over a period of a few hours, indicating a strengthening of the ring current. This period of decrease is called the 'main phase' of the storm. In a very large storm, the total decrease in *Dst* can be as much as 600 nT (nanoteslas), about 2% of the Earth's total field strength at low latitudes. In the 'recovery phase' of the storm, which typically lasts a day (or several days for very large storms), the index returns to its normal level (near zero).

In the first half of the twentieth century, theoretical calculations and laboratory experiments clarified how charged particles should be expected to move in the dipolar field of the Earth. The early pioneers in the field were Kristian Birkeland and Carl Størmer and later Hannes Alfvén and Sydney Chapman. They demonstrated that particles with energies below about 100 million electron volts (MeV) could easily be trapped in the Earth's nearly dipolar magnetic field. A charged particle executes a spiral motion about a field line, as shown in figure 1. As a charged particle moves toward the Earth along a dipole field line, its motion parallel to the field gradually slows as it encounters stronger and stronger fields. The field-aligned velocity eventually reverses, as if the particle were reflected from a mirror. Then the particle heads back out into the magnetosphere and towards the other pole of the Earth, where it is similarly reflected. The particle thus executes a nearly periodic motion along the field line, between 'mirror points' in the northern and southern hemispheres. It also drifts very slowly from one field line to another, with negative particles (electrons) drifting east around the Earth and positive ions drifting west. The eastward drift of negative particles and westward drift of positives leads to a net westward electrical current around the Earth. Theorists naturally identified this westward current as the ring current that had been inferred from ground measurements.

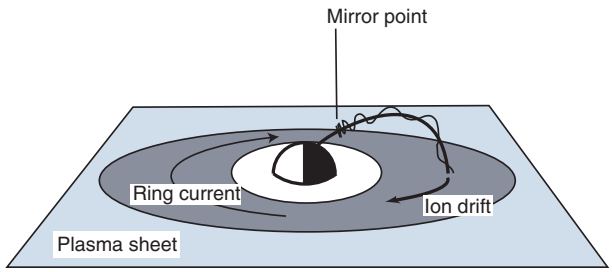


Figure 1. The geometry of the ring current. The shaded regions are in the equatorial plane. The southern hemisphere is not shown.

Early in the space age, Alexander Dessler, Eugene Parker and Norbert Sckopke demonstrated that the magnetic perturbation near the Earth caused by a charged particle moving in a dipole magnetic field is a constant times the particle energy, independent of how far the particle is from the Earth. Thus the *Dst* index provides an estimate of the total energy of all of the charged particles trapped in the magnetic field. A typical magnetic storm with $Dst = -100$ nT corresponds to a ring-current energy of 4×10^{15} J.

Spacecraft observation of the particles that carry the bulk of the ring current proved to be non-trivial, and comprehensive measurement of their properties even more so. Neither of the first two major particle populations discovered in the inner magnetosphere had enough total energy to explain the ring current. The first population to be discovered was the Van Allen radiation belt particles with energies above 1 MeV (see VAN ALLEN BELTS), but those were insufficiently numerous. The second was the plasmasphere, which includes a huge number of particles; but with less than an electron volt of energy per particle it also proved insufficient to explain the ring current (see MAGNETOSPHERE OF EARTH: PLASMASPHERE). By the late 1960s, however, spacecraft measurements made by Louis A Frank and collaborators at the University of Iowa began to detect a particle population that had sufficient energy density to explain the observed magnetic perturbations. Their observations and those of succeeding spacecraft flown in the early 1970s showed that the ring current was carried primarily by ions in the energy range from a few keV to a few hundred keV. The ions occupy field lines that cross the equatorial plane between about 2.5 and 7 Earth radii from the Earth's center. Of course, these ring current ions are accompanied by a nearly equal number of electrons, to maintain charge neutrality, but most of those neutralizing

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

electrons are less energetic and produce a current that is proportionately weaker.

The chemical composition of the ring current also held a surprise. The instruments that first discovered the ring current particles could not distinguish between different ion species, but most scientists assumed that the particles were hydrogen ions (i.e. protons), the dominant component of the solar wind (see SOLAR WIND COMPOSITION). However, when the composition of the ring current was actually measured, the truth turned out to be different. The lower-energy part of the ring current turned out to be predominantly O^+ ions, while the high-energy part was mostly H^+. The solar wind contains virtually no O^+, which, however, is a major component of the ionosphere. Thus it became clear that a large fraction of the ring current ions come from the Earth and not from the solar wind.

The source of the ring current

The most obvious source of ring current particles is the plasma sheet, the hot (kilovolt) plasma that occupies the equatorial region of the magnetosphere beyond the ring current (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). The plasma sheet is the source of the AURORA. Like the ring current, the plasma sheet contains both H^+ and O^+ ions, but the ratio of O^+ to H^+ decreases with distance from the Earth. Convection causes plasma sheet particles to drift Sun-ward from the magnetotail and divert around the Earth (see figure 2.) In the near-Earth region, westward gradient drift dominates the motion of most ions above about 10 keV, and the particles are trapped in Earth-circling orbits. The separatrix between the two types of trajectories is called the 'Alfvén layer'. The particles that lie inside the separatrix and are trapped around the Earth are generally identified as ring current particles, while those on open trajectories are identified as plasma sheet particles.

The Alfvén layer, which separates convection-dominated trajectories from gradient-drift-dominated trajectories, is not stationary in time. It moves toward the Earth in times of strong convection and away from the Earth in quiet times, as indicated in the right half of figure 2. A brief period of strong convection brings plasma sheet ions close to the Earth on the night side. If the convection weakens again by the time the particles drift to the day side, they remain on trapped orbits about the Earth. This kind of event, in which fresh plasma sheet particles are added to the trapped ring current, is called an 'injection'.

A second source of the ring current is the underlying ionosphere. Ions are observed flowing up along field lines from the ionosphere to the magnetosphere. Some of these upflowing ions have energies not much greater than the thermal energy of particles in the underlying ionosphere, i.e. less than an ELECTRON VOLT. However, intense electrical currents that flow along auroral-zone field lines generate plasma waves that heat the ions. Electric fields that accelerate electrons downward to produce bright auroral forms also accelerate ionospheric ions upward. Consequently, some ionospheric ions are accelerated to kilovolt energies on their way to the magnetosphere, and some of them become part of the ring current.

Ring current loss mechanisms

Overall, the most important loss mechanism for ring-current ions is charge exchange with neutral atoms in the Earth's exosphere, sometimes called the hydrogen geocorona, the outermost part of the neutral atmosphere. The crucial reactions are

$$H^+ + H \rightarrow H + H^+$$

$$O^+ + H \rightarrow O + H^+.$$

A kilovolt H^+ or O^+ ion captures an electron from a hydrogen atom of the Earth's exosphere, which typically has an energy much less than 1 eV, producing a kilovolt neutral atom and an ion of very low energy. The energetic neutral travels off in a straight line and, in most cases, leaves the Earth's environment entirely. The low-energy ion remains trapped in the inner magnetosphere but, because of its low energy, makes a negligible contribution to the ring current.

Charge exchange lifetimes increase rapidly with distance from the Earth, because the density of the exosphere decreases rapidly with distance from the planet. Charge-exchange lifetimes in the ring current region range from hours near the Earth to weeks in the outer region. The recovery phase of a storm, in which the *Dst* index recovers back to its pre-storm level over period of 1–3 days, is usually attributed to charge-exchange loss of ring current ions.

There are several other loss mechanisms:

(a) Convection to the dayside magnetopause. Periods of enhanced convection can sweep ring current particles Sun-ward to the magnetopause, even as they bring fresh particles in.
(b) Coulomb scattering. Ring current ions gradually lose energy by collisions with low-energy electrons, particularly within the plasmasphere where the electron density is high. This loss process is particularly important for the lowest-energy ring current ions.
(c) Pitch-angle scattering. Interactions with high-frequency waves can scatter a ring current ion. If the scattering results in a velocity that is very nearly parallel to the magnetic field, it may move the particle's mirror point to such low altitude that the ion is lost by collisions in the dense part of the atmosphere. Whether or not this loss mechanism is strong enough to have a major effect on the total loss rate for ring current ions is a matter of current debate and study.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

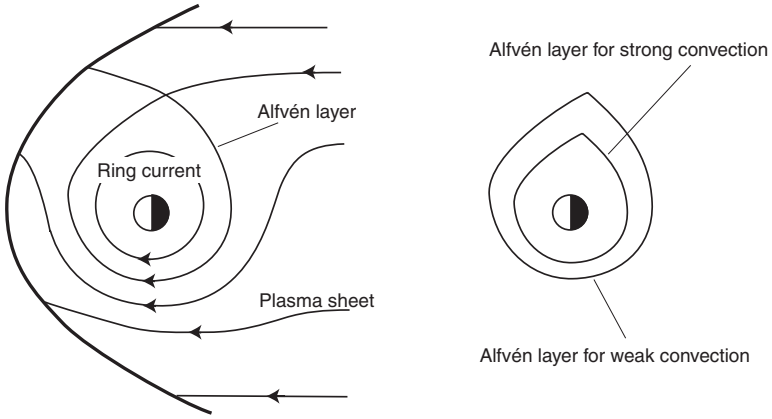


Figure 2. Drift trajectories of energetic ions in the equatorial plane of the Earth's magnetosphere. The Sun is to the left, the tail to the right. The left diagram shows drift paths. The drift motion is a combination of Sun-ward convection, which dominates far from the Earth, and westward (clockwise) gradient drift, which dominates near Earth. The right diagram shows how the regime of closed drift paths shrinks in times of strong convection.

Relationship of the ring current to the rest of the magnetosphere

The overall magnetic configuration of the magnetosphere can be substantially affected by the ring current. In a magnetic storm, the particle pressure in the ring current region becomes comparable to the magnetic pressure, and the magnetic field is substantially inflated. The magnetic connection between the ionosphere and the magnetosphere is thus altered: a field line from a given latitude in the ionosphere extends further from Earth when the ring current is strong. This change in the magnetic field allows the aurora, which normally lies on field lines connected to the plasma sheet, to be seen at lower latitudes.

The ring current also affects the distribution of the convection electric field. Particularly during times of strong convection, and thus particularly in the main phase of a magnetic storm, the ring current does not flow entirely on closed loops in the magnetosphere. There is often a partial ring centered in the pre-midnight sector, with its ends connected to the conducting ionosphere through magnetic-field-aligned currents. A current down into a region of the ionosphere causes it to charge positively, whereas a region of upward current charges negatively. Thus partial rings of current in the magnetosphere affect ionospheric currents and electric fields.

A matter of current debate is the relationship between magnetic storms and magnetospheric substorms (see MAGNETOSPHERE OF EARTH: SUBSTORMS). The substorm is a brief but intense disturbance in the magnetosphere–ionosphere system. It is centered on the night side of the Earth, in the plasma sheet and auroral ionosphere. Typically, a few large substorms occur during the main phase of any large magnetic storm.

The geosynchronous orbit, where a spacecraft orbital velocity is the same as the rotation rate of the Earth, lies at 6.6 Earth radii from the center of the Earth. It is home to many operating spacecraft, some of which carry particle detectors. That orbit typically lies close to the Alfvén layer and affords a good view of particle injections. Nearly every magnetospheric substorm injects some fresh particles into the geosynchronous orbit region. However, most substorms do not occur within magnetic storms. In other words, every substorm injects fresh particles into the inner magnetosphere, but most do not cause a lasting enhancement of the ring current. It is not clear whether the substorm is essential to the injection of the storm-time ring current.

Bibliography

Chen M W, Schulz M and Lyons L R 1997 Modeling of ring current formation and decay: a review *Magnetic Storms* ed B T Tsurutani, J K Arballo, W D Gonzalez and Y Kamide (Washington, DC: American Geophysical Union) pp 173–86

Williams D J 1987 The Earth's ring current: present situation and future thrusts *Phys. Scr.* **T18** 140–51

Richard Wolf

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Substorms

The term *magnetospheric substorm* is used to describe a host of phenomena which occur as a result of the transfer of momentum and energy from the SOLAR WIND into the MAGNETOSPHERE OF EARTH and subsequently into the ionosphere and upper atmosphere. The word 'substorm' was initially used in the early part of the 1960s to portray rapid and repeatable variations of the polar magnetic field during magnetic storms. In order to characterize the overall phenomenology of auroral disturbances, the term was modified to *auroral substorm*, before becoming more widely incorporated as the magnetospheric substorm in the 1970s. The study of magnetospheric substorms has tended to attract controversy and this remains the case currently. At the moment there are a number of aspects of magnetospheric substorms on which there is general consensus within the scientific community and this article deals with those aspects before moving to the more controversial aspects.

The AURORA is caused by the precipitation into the ionosphere of particles, with energies up to 10 keV or more, and is extremely dynamic during auroral substorms. The original description of an auroral substorm derived from the development of auroral activity based on the analysis of data from arrays of all-sky cameras deployed around the auroral oval during the International Geophysical Year (IGY) of 1957. A schematic representation of the development of an auroral substorm as portrayed in 1964 is given in figure 1. The quiet state (figure 1, panel A) consists of multiple arcs on the nightside drifting equatorward. A sudden brightening of the equatorward-most arc (figure 1, panel B) represents the onset of the *expansion phase* of the auroral substorm. Thereafter, the auroral activity moves westward, poleward and eastward (figure 1, panel C) forming a broad region of bright and dynamic auroral activity in the midnight sector, often known as the *auroral bulge*. As the auroral activity within the bulge continues to move westward, the westward edge of the bulge becomes deformed with a sharp kink developing, which is called the *westward travelling surge*, WTS (figure 1, panel D). Auroral activity continues throughout 60 min or so after the onset (figure 1, panels D and E), including the formation of auroral forms called *omega bands* at the eastern edge of the bright aurora (figure 1, panel D). Eventually, when the auroral activity has reached its most poleward extent, it starts to decline and the luminosity decreases, representing the start of the *recovery phase* (figure 1, panel F). During this time quiet auroral arcs reappear near the equatorward part of the active region, while the luminosity in the poleward part of the active region disappears. Also the high latitude edge of the substorm-disturbed region moves equatorward and recovery is then complete when the polar cap, the region within the auroral oval, has shrunk in size and auroral activity has moved to high latitudes. Figure 2 represents a series of six images taken by the VIS Earth Camera from the POLAR spacecraft on 12 January 1997. This series of images illustrates the global auroral activity associated with a magnetospheric substorm starting from the quiet auroral oval at 0658 UT to the maximum extent of the auroral emission at 0833 UT. Research, partly based upon global images of the aurora from space such as those in figure 2, has led to refinements to the picture of auroral activity described above which will be discussed later.

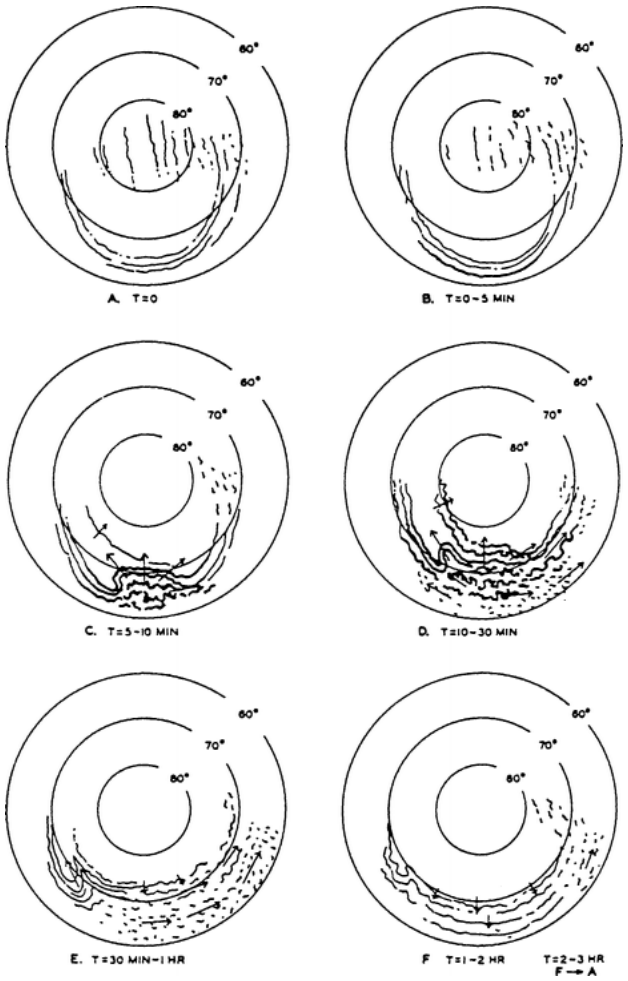


Figure 1. Schematic representation of six stages in the development of an auroral substorm as determined from all-sky camera data (from Akasofu 1964).

It is now generally accepted that a magnetospheric substorm comprises three separate phases: the growth, expansion and recovery phases. The *growth phase* is generally an interval of energy transfer from the solar wind and subsequent storage in the tail of the magnetosphere together with a small amount of dissipation in the ionosphere. This phase was not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

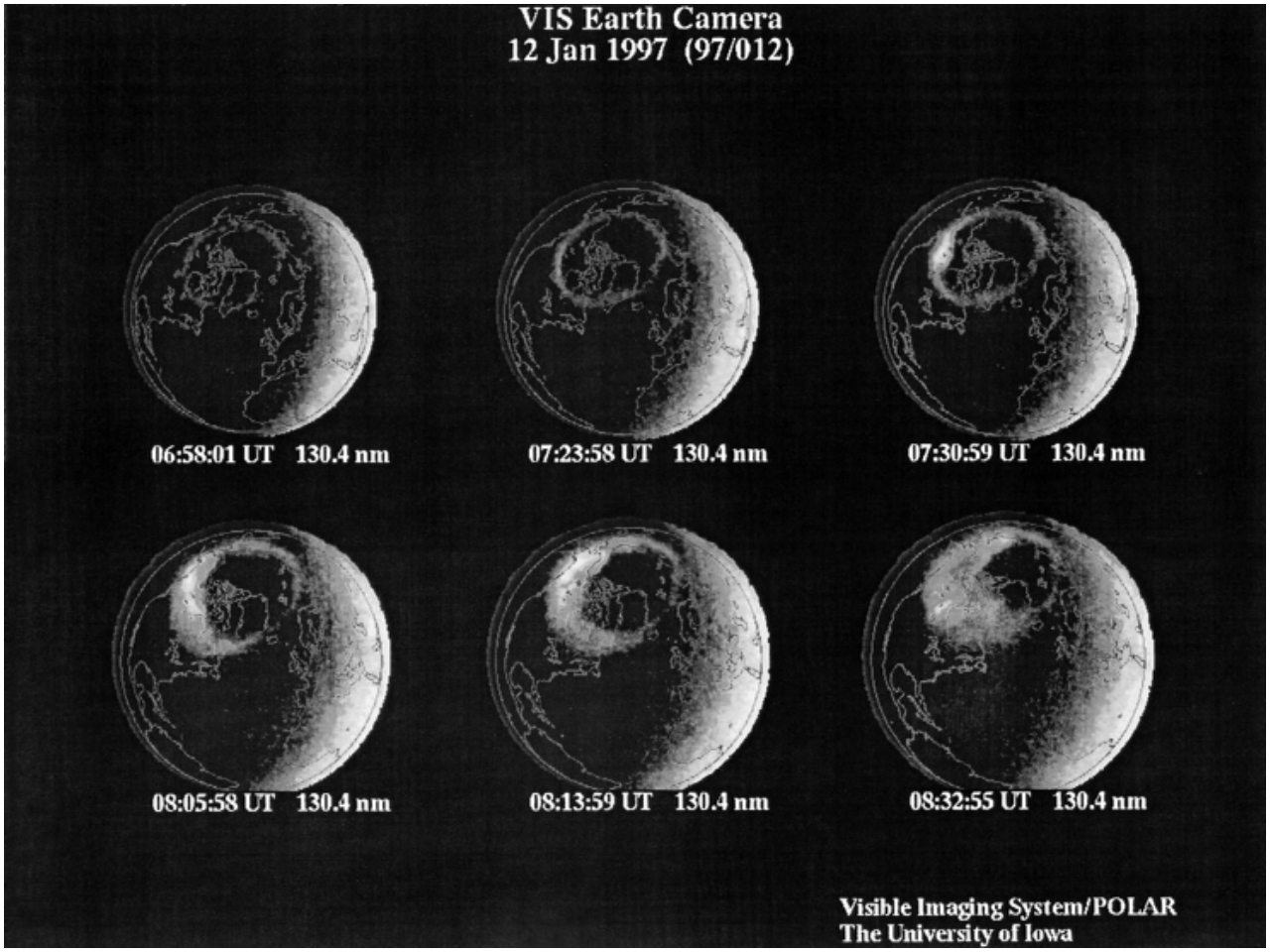


Figure 2. A series of six images from the VIS Earth Camera (Frank *et al* 1995) on the Polar spacecraft. The images have been supplied by Professor L A Frank and Dr J B Sigwarth.

considered part of the original auroral substorm, and is the interval prior to panel B in figure 1. The expansion phase is an interval when energy stored in the tail is explosively released, with some energy appearing in the ionosphere, some in the ring current and some being lost to the solar wind. The onset of the expansion phase occurs at the time of panel B in figure 1 and continues through to panel F. Finally, the recovery phase is an interval when the magnetosphere–ionosphere system relaxes back to a quiescent state. By describing the associated phenomena, we now consider each of these phases in turn.

Substorm phases

The interplanetary magnetic field (IMF) is convected away from the Sun with the solar wind, varying in magnitude and direction as it is transported. The growth phase of the magnetospheric substorm is considered to begin with the onset of southward interplanetary magnetic field (IMF) at the dayside magnetopause. This leads to MAGNETIC RECONNECTION between the IMF and the geomagnetic field, thereby creating open magnetic flux which is connected at one end to the Earth and at the other to the solar-wind plasma. This newly opened magnetic flux is transported by the motion of the solar wind into the *magnetotail* where it is added to the lobes (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL). In this way, magnetic energy, $B^2/2\mu_0$ per volume, is stored in the tail as the lobe magnetic field intensity increases. The region of the ionosphere, which is threaded by open flux, is termed the *polar cap*. Ionospheric flow is stimulated by the transport of open flux from the dayside into the tail and the polar cap also expands. If the IMF remains southward for an extended period of time, for example several tens of minutes, then ionospheric flows continue to be stimulated and the polar cap continues to expand, thereby also moving the auroral oval to lower latitudes. This global convection process is the *directly driven* part of the magnetospheric substorm and some energy is dissipated in the ionosphere, mainly via *Joule heating* of the ionospheric plasma. In the nightside ionosphere the auroral forms tend to be quiescent and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

relatively stable. There is one dynamic auroral feature which takes the form of equatorward moving arcs. These arcs begin their equatorward drift close to the border between the auroral oval and the polar cap, and several can be seen during the interval of the growth phase.

A further consequence of the addition of magnetic flux in the tail is the re-orientation of the magnetic field in the near-Earth tail, for example at geostationary orbit, where the field becomes less dipolar-like and more like the stretched mid and far tail field. This re-orientation results from the formation and subsequent intensification of a thin current sheet in the inner magnetosphere, the *cross-tail current*. Another signature at geostationary orbit is the reduction of particle fluxes, especially in the energy range 30–300 keV. This reduction in particle fluxes is caused by the 90°pitch angle particles, which drift through the nightside during this time, moving closer to the Earth. Small pitch angle particles are less affected and the resultant pitch angle distribution becomes elongated, and is often referred to as 'cigar-like'. This particular signature normally occurs in the late growth phase, some 10 min or so prior to the expansion phase onset. Further down the tail the stretching of the field caused by the addition of magnetic flux into the tail lobes results in *plasma sheet thinning* (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). The growth phase, as defined by the interval from the southward turning of the IMF to the expansion phase onset, normally lasts between 30–60 min.

The expansion phase of the magnetospheric substorm is normally accepted as starting with the brightening of the equatorward-most auroral arc in the nightside ionosphere. This brightening begins in a localized region near midnight and propagates along the length of the arc both westwards and eastwards at phase velocities of up to 10 km s^{-1}. This is then followed by a poleward expansion of the auroral luminosity at speeds of typically a few hundred m s^{-1}. This is known as the *auroral break-up* because of the rapid deformation of the original quiet arc. The poleward expansion forms the auroral feature known as the auroral bulge and the westward motion of the western edge of the bulge develops the WTS, as discussed earlier, which has a westward phase velocity of ~1 km s^{-1} and may ultimately reach the dusk meridian. This original concept of the WTS, however, seems not to apply when considered carefully. Based upon high time resolution auroral images taken by the UV imager on the Viking spacecraft, it seems that the WTS is not a single surge form which propagates but is more likely a series of new surge forms which appear to the west of previously formed regions. In fact, each individual surge form does not appear to propagate far from the meridian on which it was initiated. Eastern motion of the auroral activity at the eastern part of the substorm active region can also occur but is in general less distinctive.

Also at the expansion phase onset a wave, called a *Pi 2 pulsation*, occurs. This wave has a frequency between 5–25 mHz, equivalent to a period range of 200–40 s, and is normally of order 10 mHz, a period of 100 s, and lasts for a few cycles whilst decaying in amplitude. The Pi 2 pulsation is primarily a nightside feature which can be seen over a wide range of local times, sometimes around to noon, and at latitudes from the auroral zone to the equator although it is most well defined at mid latitudes. The 'global' nature of the wave allows it to be considered as an ideal timing indicator for substorm onset in the event of no auroral observations, at least to times of order 1 min. These waves result from an impulsive change in the geomagnetic tail which causes a fast magnetohydrodynamic (MHD) wave to propagate throughout the tail (see MAGNETOHYDRODYNAMIC WAVES). This wave couples to a cavity mode inside the plasmasphere causing the well-defined mid-latitude signature.

At the same time as the development of the enhanced luminosity associated with the auroral break-up, ground magnetic perturbations are also observed which occur as a result of enhanced currents flowing in the ionosphere, termed *auroral electrojets*. These magnetic deflections are typically several hundred nT but during particularly intense intervals of current flow, can reach up to 2000 nT. Note the magnitude of the field at Earth's surface is typically 50 000 nT. With increasing time, there may be subsequent intensifications of the auroral activity, usually to the west of the original break-up, resulting in further enhancements of the current, and a poleward motion of the luminosity and current, often apparently in a stepwise manner rather than a smoothly varying way. At the time the poleward motion of the luminosity stops and the auroral region is at its widest, the expansion phase is said to have ended and the recovery phase begins. The enhanced auroral luminosity and the enhanced electrojets are a consequence of a change in the flux and energy spectrum of the precipitating electrons into the ionosphere. There are more particles with a harder spectrum resulting not only in increased luminosity but also in enhancements in electron density at lower altitudes, sometimes down to less than 100 km. A further characteristic signature caused by this change in *particle precipitation* is the absorption of radio waves in the MHz frequency band.

Most spacecraft observations in the magnetosphere during the expansion phase have been made at geosynchronous orbit. Here there are two typical signatures of onset, one being the dispersionless, in energy, injection of electrons and ions in the energy range 30–300 keV, the other the *dipolarization* of the magnetic field. *Dispersionless injections* of particles tend to occur in a limited local time sector near midnight. All energies appear at the spacecraft at the same time. Subsequently electrons and ions are subject to forces due

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to the gradient and curvature of the near-dipolar magnetic field in the inner magnetosphere which cause the particles to drift around the Earth, electrons to the east and ions to the west. Thus a spacecraft at geosynchronous orbit located several hours or more away from midnight will observe enhancements in the particle fluxes which are dispersed in energy, with higher energy particles reaching the spacecraft before lower energy particles. The magnetic field re-orientation, called the dipolarization, is also a characteristic signature of the early expansion phase. Again this feature is initially limited in local time extent in the tail, although there is evidence for expansion to west and east during the expansion phase in a stepwise manner. A further reproducible signature of the expansion phase onset in space is the onset of a burst of AURORAL KILOMETRIC RADIATION (AKR). This radiation, which typically occurs in the frequency band 200–500 kHz, is generated on the field lines of discrete aurora typically 1.5–3R_E above the auroral ionosphere and results from the electron cyclotron maser instability which is induced by energetic electrons.

The recovery phase is perhaps the hardest to characterize of the three phases. During this interval the system returns to a quiescent state. However, there are still optical features such as omega bands which are present and moving, in general eastwards. A double oval is also seen to form when the luminosity in the centre of the oval decreases, creating two bands of enhanced luminosity at the equatorward and poleward edges of the oval. The optical luminosity is reducing along with the electrojet currents, but decay times can be relatively long, of the order of 30–60 min, since particles are still present and drifting in the magnetosphere and consequently precipitating into the ionosphere. Ultimately, however, the system returns to a state which is similar to that prior to the onset of the growth phase. The timescales for the three phases are typically 30–60 min (growth), 1–2 h (expansion) and 30–60 min (recovery). The situation may become more complex if the IMF varies on timescales which are longer than those described above.

The phenomena that are observed during each of the three phases have been described above. Typically these observations have been made on the ground or at geosynchronous orbit. There have been other space observations as well but these will be mainly discussed below as we describe various models for the expansion phase of the substorm. It is this part of the substorm process that attracts most of the current debate.

Substorm current wedge

Apart from the auroral activity described earlier which occurs during the substorm, there are also characteristic magnetic signatures which appear during the expansion phase in particular. The auroral electrojets have already been mentioned, but the westward electrojet forms part of one of the universally accepted features of substorms, the *substorm current wedge* (figure 3). At expansion phase onset the cross tail current is diverted along magnetic field lines into the ionosphere where the enhanced current flows as the westward electrojet before flowing out of the ionosphere back into the tail. The diversion of the cross tail current changes the stretched tail field configuration observed at geostationary orbit near midnight during the growth phase and results in the dipolarization of the field. The current wedge consists then of two *field aligned currents* (FAC), downward to the Earth in the east and upward from the Earth in the west. The upward current is generally believed to be colocated with the western edge of the auroral activity and is intimately related to the WTS. Each surge form which is part of the WTS as it evolves has an upward FAC at its core with a circulating Hall current in the region surrounding the core. While there is a downward FAC in the volume surrounding the surge, there is generally a preference for downward current in the east to connect to the upward FAC by the westward electrojet. This wedge of current has characteristic magnetic signatures at mid-latitudes which are termed *magnetic bays* because of the similarity of the time varying magnetic field to coastal bays on a map. The north–south component of the field has perturbations which are positive, i.e. has a northward direction, at longitudes within the current wedge, since the direction of the magnetic field perturbation due to the currents is in the northward direction. The east–west component has a positive bay, i.e. in an eastward direction, on the western half and a negative bay at the eastern half. The east–west component bays are maximum at longitudes which coincide with the field-aligned currents. The SCW is closely related to the Pi 2 pulsations which occur at expansion phase onset and the waves are seen as a high-frequency response on the lower frequency bay variation. This current system is occasionally referred to as the *DP-1 current* (disturbance polar type 1) and is an expansion

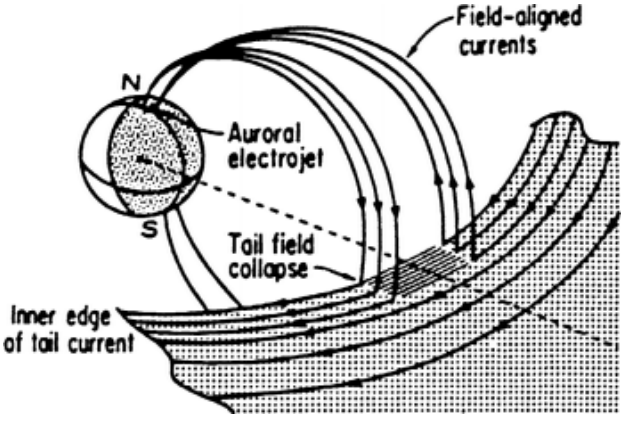


Figure 3. A schematic representation of the substorm current wedge (from Clauer and McPherron 1974). (Copyright 1974 by the American Geophysical Union.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

phase signature alone. The *DP-2 current* is the current system which develops during the growth phase and is equivalent to the ionospheric flow which is stimulated by reconnection processes at the dayside magnetopause.

Substorm models

The first model proposed to explain many of the observations of magnetospheric substorms is the *near-earth neutral line* (NENL) model. Here the cause of the expansion phase onset is magnetic reconnection of closed magnetic field lines in the tail which originally thread the plasma sheet forming a *neutral line* in the near-Earth tail. Consequently the process which powers the substorm expansion is reconnection. Figure 4 illustrates the original version of the NENL which demonstrates schematically the changes in the magnetotail occurring during substorms. The panels in figure 4 are cuts along the midnight meridian plane of the magnetotail. The onset of reconnection in the tail at a new neutral line, N' (figure 4, panel 2) follows from the growth phase thinning of the plasma sheet. Initially, the NENL is of limited azimuthal extent and, as reconnection continues, newly reconnected magnetic field lines form an island of plasma tailward of the neutral line which is disconnected from the Earth (figure 4, panels 3 and 4). Plasma flows into the neutral line from above and below and out from the neutral line in both earthward and tailward directions. Ultimately, open magnetic flux of the tail lobes begins to reconnect (figure 4, panels 5–8) and the island of plasma forms a *plasmoid* which then, free from restraining forces, travels down the geomagnetic tail and is ejected into the solar wind. The timescale for reconnection to 'eat' through the plasma sheet is believed to be of the order of 5–15 min. The formation of the NENL interrupts the cross-tail current, which had intensified during the growth phase, causing the near-Earth magnetic field to adopt a more dipolar-like orientation, the dipolarization. The current system that results is the substorm current wedge. The radial position of the onset of reconnection in the geomagnetic tail has been debated often. The key signatures that are used to identify the existence of a NENL are southward-directed magnetic field and tailward flow. The evidence from a number of studies, involving data from several different spacecraft, suggests that these signatures are not seen in the tail at radial distances less than $19R_E$ from the Earth. Recent work with the GEOTAIL spacecraft indicates that tailward flows

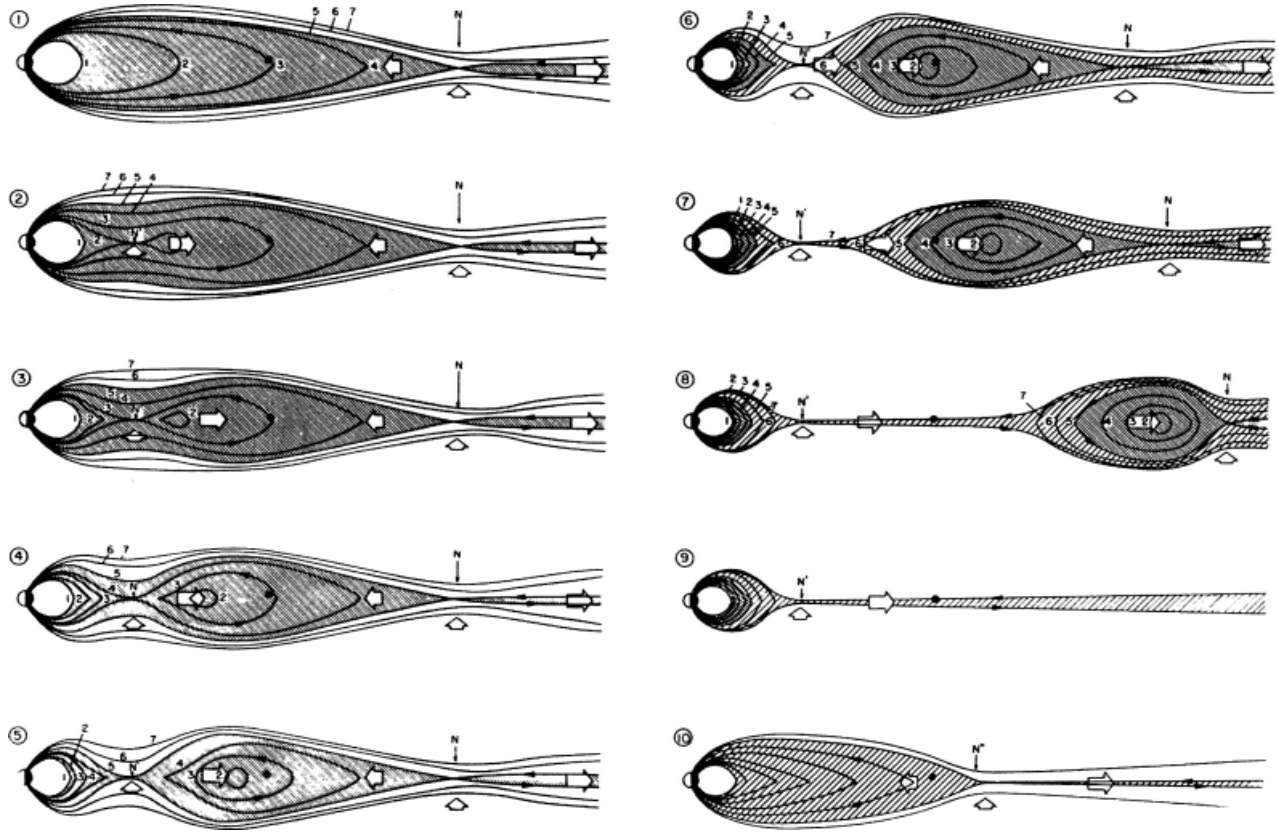


Figure 4. A schematic representation of the changes in the tail configuration during the substorm. This is the near-Earth neutral line model (see Hones 1979).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are seen predominantly between 20–30 R_E down tail and, interestingly, mainly pre-midnight. Furthermore, these flows occur within ±5 min of a Pi 2 pulsation observed on the ground and last on average between 7–9 min.

The location of the onset of reconnection at expansion phase onset some 20–30 R_E downtail does not appear to be consistent with a number of ground observations, which indicate that the auroral break-up maps to between 6–10 R_E downtail. These observations include the location of the initial disruption in the auroral forms at the equatorward-most auroral arc. Also, equatorward-drifting arcs are seen to continue to drift after the onset of the expansion phase, as identified either by Pi 2 pulsations or by auroral break-ups. Then the auroral arc which breaks up is seen to occur close to the location of latitudes where long period ULF waves are observed which imply both closed field lines and ones which will allow the formation of standing waves on them. This latter point implies that the magnetic field which maps from the auroral break-up into the tail is not as stretched as suggested in the NENL model. Thus, reconnection may not be the only process involved in the substorm expansion phase onset.

One proposal to overcome these difficulties is that the reconnection region in the tail does not connect magnetically with the auroral arc which brightens at expansion phase onset. If the onset of reconnection in the tail at some 20–30 R_E downtail is the start of expansion phase onset, earthward flows of plasma would result. This earthward flow would be reduced as it enters the near-Earth tail as a result of the increase in the magnetic field. *Convection breaking*, as this process is termed, leads to magnetic flux pile up, dipolarization, the formation of the substorm current wedge and energetic particle precipitation (see MAGNETOSPHERE OF EARTH: CONVECTION). These signatures then propagate tailward down the tail. This proposal leads to the conclusion that the auroral break-up, or expansion phase onset in the ionosphere, does not occur at the same time as the onset of the expansion phase in the tail.

The *current sheet disruption* (CSD) model is an alternative to the NENL model which is proposed to explain the expansion phase onset. In this model, the thin current sheet still develops in the inner magnetosphere during the growth phase. A combination of lower hybrid waves, produced by streaming non-adiabatic ions interacting with adiabatic electrons, and the lower hybrid drift instability at the density gradient on the boundary of the plasma sheet produces anomalous resistivity which disrupts the cross-tail current, thereby resulting in the substorm current wedge. A further key difference from the NENL model is that disruption of the tail current begins close to geostationary orbit and then expands radially earthward and tailward. In fact it has been proposed that a rarefaction wave is launched downtail reducing the B_z component in the magnetotail and inducing plasma sheet thinning. Ultimately these effects induce reconnection in the tail which leads to the subsequent formation of the plasmoid. Thus closure of open magnetic flux is not necessary for expansion phase onset in the CSD model, although it does occur at some time later.

Another class of model considers the effects of the ionosphere on the magnetotail, through the action of field aligned currents, and such models have been termed *magnetosphere–ionosphere coupling* (MIC) models. In such models, changes in the ionospheric conductivity influence the tail through positive feedback processes, leading to expansion phase onset. By including reconnection in the tail, a *globally integrated substorm* (GIS) model has been developed.

In summary, although there is good agreement on the signatures of the expansion phase, there is no such agreement on the cause of the expansion phase onset. There are a number of models for this element of the substorm process, each of which predicts certain expansion phase signatures. A general consensus, or unified theory, still eludes us however and is a major challenge over the next few years.

Multiple onsets and pseudo break-ups

The discussion above is based upon many observations, but in general relates to what are termed isolated substorms. There are two aspects of such events which have not been covered. During most substorms the expansion phase does not continue in a smooth manner which would imply a continuous release of stored energy. Rather, following the initial auroral break-up, there are a series of intensifications of auroral luminosity, enhancements of the westward electrojet and Pi 2 pulsations which imply that the stored energy is released impulsively. Such *multiple intensifications* have a typical timescale of 10–15 min. The auroral luminosity appears generally to the west and poleward of the initial intensification which is in general separate from the stepwise motion of the WTS. Furthermore, the enhanced ionospheric currents also appear poleward of the initial location. There is also evidence from mid-latitude magnetograms that the western FAC of the SCW moves westward. These intensifications continue throughout the expansion phase, ceasing only when the aurora and currents reach their poleward-most location. A second feature of substorms which has become apparent is the so-called *pseudo break-up*. Such events occur during the growth phase of the substorm. They have most of the characteristics of an auroral expansion phase, for example auroral brightening, Pi 2 pulsations, enhancements of the electrojet current, but the brightened aurora does not appear to expand poleward and westward and after a few minutes simply fades away. It appears therefore that processes which lead to expansion phase onset are in some way quenched. How this happens

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

remains uncertain, although the various paradigms discussed above require different physical mechanisms for such quenching to take place.

A typical substorm will include a growth phase, an expansion phase and a recovery phase and such a description is adequate under conditions where the IMF B_z does not remain south for longer than the typical length of those three phases, of the order of 3 h. However, it is possible that the IMF will in fact remain southwards for longer than this interval. In such a case it is entirely possible that substorms will occur when there does not appear to be a separate growth phase, i.e. the expansion phase of the second substorm will occur at some time during the recovery phase of the first substorm. Also, the auroral oval can continue to expand equatorward in a global sense during the expansion phase. Of course the main physics of the growth phase, i.e. energy storage, is continuing to take place if the IMF remains southward.

Modular approach to auroral activity

Auroral activity is central to the substorm process. The original definition of the auroral substorm was based upon records from a network of all-sky cameras based around the auroral zone. Subsequent developments in spacecraft instrumentation provided images of part of, or the whole, auroral oval at good temporal and spatial resolution. Based on data from one such satellite, Viking, certain researchers have put forward an alternative approach to considering the auroral substorm. This approach stems from the considerable variety which can be identified in auroral activity during the various phases of the substorm, such that no two substorms are exactly alike. The proposal is that certain cycles of auroral activity, modules, couple together to provide the global auroral substorm. The cycles of activity represent the response in the ionosphere to different magnetospheric processes coupling in a repeatable manner.

The approach proposed divides the auroral activity into primary and secondary auroral modules. The primary auroral modules represent the fundamental characteristics necessary for a complete description of the global auroral activity within a substorm. Such modules are also important as they represent timing indicators for other observations. Separate auroral forms occur but these secondary auroral modules do not exist independently of the primary modules. There is uncertainty in the time and location at which these secondary auroral modules appear such that they are not easily included in the framework of a magnetospheric substorm. Here we consider only the primary modules.

The first primary module is auroral growth and dayside patterns and effectively represents the auroral activity during the growth phase. This is something of a departure from the original auroral substorm which did not include any growth phase signatures. Such auroral activations are most probably directly driven by the solar wind. The next time marker or module is auroral onset. In this modular approach there is no difference accorded to break-up and pseudo break-up since the event may or may not develop into a major disturbance. Whether it does or not depends upon its coupling to other modules. However, if the third module, the auroral bulge, develops, then the onset is considered to be an expansion phase onset. The fourth module is the double oval and its variations. During the expansion phase the auroral activity moves poleward. A double oval occurs when the auroral luminosity in the centre of the broad substorm oval decays, when both the equatorward-most activity and poleward-most activity are seen to behave independently, the implication being that different parts of the magnetosphere are responding independently. The final module consists of polar arcs which were originally thought to be independent of the substorm process, since it was thought that they disappeared at expansion phase onset. However, it has been demonstrated that they can occur throughout the substorm phases, and often are seen to intensify during the expansion phase onset. Furthermore, it seems that there is evidence that such arcs can occur on both open and closed magnetic field lines.

Some of the key features of the original auroral substorm, such as the westward travelling surge and omega bands, do not appear as primary auroral modules. They are, however, secondary modules, and therefore represent refinements to the basic model.

External triggering of substorms

The models proposed above for expansion phase onset all advocate an internal instability as being responsible for the explosive release of energy. There is an alternative suggestion, however, that this need not be the case and that substorm expansion phase onset may in fact be triggered by changes in the IMF. This proposal is based upon the empirical evidence of both case studies and statistical analyses that a significant positive turning of the IMF occurs at the time of expansion phase onset. Reductions in the IMF B_y component may also apparently trigger substorms. The proposed mechanism for this triggering is a reduction in the electric field imposed on the magnetosphere by the solar wind, although the physical link is, at the moment, controversial. The delay between the northward turning and the expansions phase onset is very short—a few minutes.

Magnetic indices

The study of substorms has been aided by the use of magnetic indices. Such indices are used to provide a summary of the magnetic variations over a period of time. They can also be used to investigate physical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

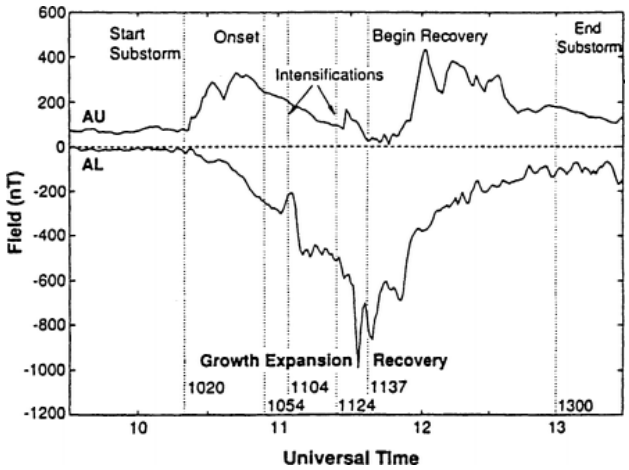


Figure 5. The AU and AL magnetic indices for a well-studied substorm demonstrating the three phases of a substorm (from McPherron and Manka 1985). (Copyright 1985 by the American Geophysical Union.)

processes and in terms of substorm behaviour the auroral electrojet indices are perhaps the most widely used. The AE indices are determined from the north–south component of 12 stations which are approximately equally spaced in longitude around the auroral regions. The AU and AL indices represent the upper and lower envelopes of the north–south components of these 12 stations, i.e. the maximum and minimum deflections at any given time. Since the magnetic variations are caused by the ionospheric currents, they should represent the behaviour of these currents during the various phases of the substorm. An example of the AU and AL indices for one particular event is given in figure 5. In this example, an isolated substorm comprising of the three phases is described. The growth phase appears to be mainly observed by the AU index, while the largest deflection in the AL index takes place in the expansion phase, during which there are intensifications of the electrojet currents. The recovery phase is represented by the return of the AL index to lower values, but there are still significant enhancements which may occur during this phase in the AU index.

Bibliography

Akasofu S I 1964 The development of the auroral substorm *Planet. Space Sci.* **12** 273

Baker D N, Pulkkinen T I, Angelopoulos V, Baumjohann W and McPherron R L 1996 Neutral line model of substorms: past results and present view *J. Geophys. Res.* **101** 12 975

Clauer C R and McPherron R L 1974 Mapping the local time-universal time development of magnetospheric substorms using mid-latitude magnetic observations *J. Geophys. Res*. **79** 2811

Frank A L *et al* 1995 The visible imaging system for the Polar spacecraft *Space Sci. Rev.* **71** 297

Hones E W Jr 1979 Plasma flow in the magnetotail and its implications for substorm theories *Dynamics of the Magnetosphere* ed S I Akasofu (New York: D Reidel) p 545

Lui A T Y 1996 Current disruption in the earth’s magnetosphere: observations and models *J. Geophys. Res.* **101** 13 067

Lyons L R 1996 Substorms: fundamental observational features, distinction from other disturbances, and external triggering *J. Geophys. Res.* **101** 13 011

McPherron R L 1995 Magnetospheric dynamics *Introduction to Space Physics* ed M G Kivelson and C T Russell (Cambridge: Cambridge University Press) p 400

McPherron R L and Manka R H 1985 Dynamics of the 1054 UT, March 22, 1979 substorm event: CDAW-6 *J. Geophys. Res.* **90** 1175

Rostoker G 1996 Phenomenology and physics of magneto-spheric substorms *J. Geophys. Res.* **101** 12 955

Mark Lester

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetosphere of Earth: Waves

Waves are oscillations with respect to an equilibrium position of a given quantity, ω being the angular frequency of the oscillation. Here we are concerned with magnetic and electric field oscillations, the electromagnetic and electrostatic waves. Two main reasons lead us to study such waves in the magnetosphere, their role in the dynamics of the particles with which they interact and the remote diagnosis of magnetospheric plasma parameters.

One important characteristic of the magnetospheric plasma is that it is collisionless. In such a medium the waves that are mostly generated by the free energy of the particle distribution functions are the means to relax this free energy. Waves, once amplified, can heat particles, permit exchange of energy between different populations of particles and precipitate magnetospheric particles into the atmosphere.

Two regions are particularly favorable for the generation of waves, the equatorial region of the magnetosphere where most of the energetic particles are concentrated owing to the convection electric field and the auroral field lines, where particles originating from the ionosphere, outer magnetosphere and the solar wind are accelerated (see MAGNETOSPHERE OF EARTH: CONVECTION, SOLAR–TERRESTRIAL CONNECTION: COUPLING BETWEEN SOLAR WIND, MAGNETOSPHERE, IONOSPHERE, AND NEUTRAL ATMOSPHERE).

Magnetospheric waves were first observed on the ground. *In situ* measurements by spacecraft have helped in understanding the origin of these waves and permitted new observations of waves that cannot propagate to the ground, in particular electrostatic waves.

The different kinds of magnetospheric waves can be organized by their frequency range, ultralow frequency (ULF, up to 5 Hz), extremely low frequency (ELF, 5 Hz–3 kHz), very low frequency (VLF, 3–30 kHz) and low frequency (LF, 30–300 kHz). We are not concerned with frequencies above this range. An overview of the location in the various regions of the magnetosphere of the different types of waves is given in figure 1.

Before describing and interpreting the different observations, a summary of the main wave modes and a short description of some wave–particle interaction processes are needed to understand how these modes can be amplified and interact with particles.

Principal wave modes in the magnetospheric plasma

The different wave modes must satisfy the Maxwell equations, the equations of motion of the particles and conservation laws. The solutions for monochromatic propagating plane waves can be written as $E = E_0 \exp[\mathrm{i}(\boldsymbol{k}\cdot \boldsymbol{r} - \omega t)]$, E_0 being the wave amplitude, ω the angular frequency, and $\boldsymbol{k}$ the wavevector in which direction the wave propagates. All linear fluctuations, or small-amplitude fluctuations with respect to the equilibrium, can be represented as superpositions of these eigenmodes. In plasmas, there can be both transverse electromagnetic and longitudinal electrostatic waves, the wave components being either mainly perpendicular or parallel to the wave propagation vector $\boldsymbol{k}$, respectively. Contrary to vacuum where a continuum of electromagnetic waves can propagate, in plasmas only some frequency windows are possible, which are represented on figure 2, in the $(\omega, \boldsymbol{k})$ plane.

MHD approach

MAGNETOHYDRODYNAMIC (MHD) theory allows the investigation of wave propagation modes in the limit $\omega \ll \Omega_\mathrm{i}$, where Ω_i is the proton gyrofrequency—typically 1 Hz, for an infinite homogeneous medium. The linearization of the plasma equations in such conditions leads to a dispersion relation that has three solutions. These three modes have different polarizations, the perturbation being either in a plane perpendicular to the geomagnetic field $\boldsymbol{B}_0$ or with a magnetic component parallel to $\boldsymbol{B}_0$ (compressional). In what follows, parallel ($\parallel$) and perpendicular ($\perp$) refer to $\boldsymbol{B}_0$.

The shear Alfvén mode or Alfvén mode (also known as the toroidal mode) has the simplest dispersion relation:

$$\omega = k_\parallel V_\mathrm{A} \tag{1}$$

where $k_\parallel$ is the parallel component of $\boldsymbol{k}$, V_A is the Alfvén velocity ($V_\mathrm{A} = B/(\mu_0\rho)^{1/2}$), μ_0 is the magnetic permittivity of free space and ρ is the mass density of the plasma.

The two other modes are the fast magnetosonic mode (the compressional or poloidal mode) and the slow magnetosonic mode (which has also a compressional component). Their dispersion relations do not have simple expressions; they depend on both the Alfvén and the sound velocities, and on their direction of propagation.

The slow mode cannot propagate in cold plasma and is rapidly damped for high temperature, and thus most often only the two other modes are considered (figure 2(*a*)). The ULF fluctuations whose frequency satisfies $\omega \ll \Omega_\mathrm{i}$ are identified with these modes.

Electromagnetic modes at higher frequencies

When considering higher frequencies, it is necessary to modify the equations. A first step is to add new terms to the Ohm's law equation which take into account the movement of both electrons and ions (i.e. a bi-fluid model), the new system of equations then being called the Hall MHD equations. The kinetic approach is more general, taking into account the velocity distribution function of particles. This is used to calculate possible wave amplification or damping and/or consider the effect of a hot plasma population. The principal wave modes thus obtained for different conditions are represented in figures 2(*b*)–2(*e*). The different modes depend on both the direction of propagation $\boldsymbol{k}$ and the corresponding values of the characteristic frequencies: the gyrofrequencies $\Omega_\mathrm{i,e}$ (linked to the magnetic field amplitude) and the plasma frequencies $\omega_\mathrm{pi,e}$ (linked to the plasma density).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

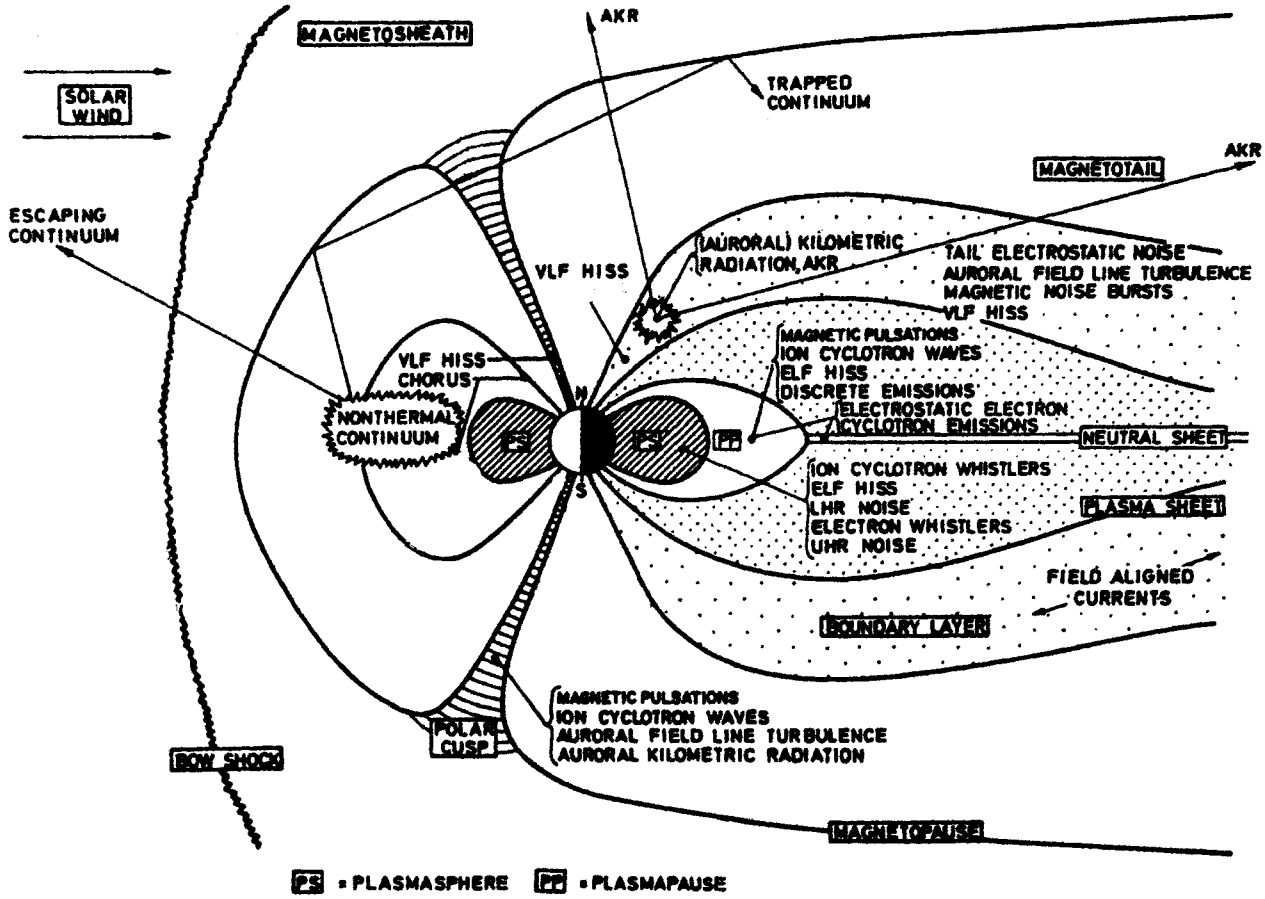


Figure 1. Types of magnetospheric plasma waves and regions of occurrence in a noon–midnight meridian plane cross section of the Earth's magnetosphere.

The most frequent situation encountered in the magnetosphere is the so-called dense plasma or weak magnetic field approximation, where $\Omega_e < \omega_{pe}$ (figures 2(*b*) and 2(*c*) for parallel and perpendicular propagation). For $\omega < \Omega_i$ there is a left-hand-polarized mode and up to Ω_e a right-handed-polarized mode (called the whistler mode), whose low-frequency limits are the shear Alfvén and magnetosonic modes respectively. At higher frequencies there are two modes ω_L and ω_R, on each side of the plasma frequency for small $k_\parallel$; at the limit $k_\parallel = 0$, it is a kind of ZEEMAN EFFECT, $\omega_{L,R} = \omega_p \pm \Omega_e/2$. The plasma frequency ω_{pe} is an electrostatic mode. Figure 2(*c*) gives the modes in the perpendicular propagation. The polarization modes in this case are either ordinary or extraordinary, meaning that the electric component of the wave is either parallel to $\boldsymbol{B}_0$ or rotating in a plane perpendicular to $\boldsymbol{B}_0$. Note in this case the presence of new characteristic frequencies or resonances, the lower hybrid and the upper hybrid resonances ($\omega_{LH} \approx (\Omega_e\Omega_i)^{1/2}$ and $\omega_{UH} = \omega_{pe}^2 + \Omega_e^2$). The forbidden frequency ranges for the different modes are indicated by dark vertical lines on the left of each plot.

In the auroral zones there are regions of very low density. The possible modes in this case are given in figures 2(*d*) and 2(*e*). The modes are the same as in the previous case, but the possible frequency windows are different.

Electrostatic modes above the electron gyrofrequency

Electrostatic waves have their electric component parallel to $\boldsymbol{k}$, being longitudinal. Among them, the plasma frequency propagates at $k_\parallel$ (parallel to $\boldsymbol{B}_0$). Using the kinetic theory and modeling the plasma by a Maxwellian distribution function it is possible to find electrostatic modes with $k_\perp$. These Bernstein modes only exist at frequencies above the electron gyrofrequency (figure 2(*f*)). The behavior of the dispersion relations (one curve between each electron gyroharmonics) is different below and above ω_{UH}. Above ω_{UH} the dispersion relations present maxima corresponding to a null group velocity ($V_g = \partial\omega/\partial k$), at frequencies denoted by F_q.

Some wave–particle interaction mechanisms

The modes described before, except the Bernstein modes, are possible propagation modes in a cold magnetoplasma.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

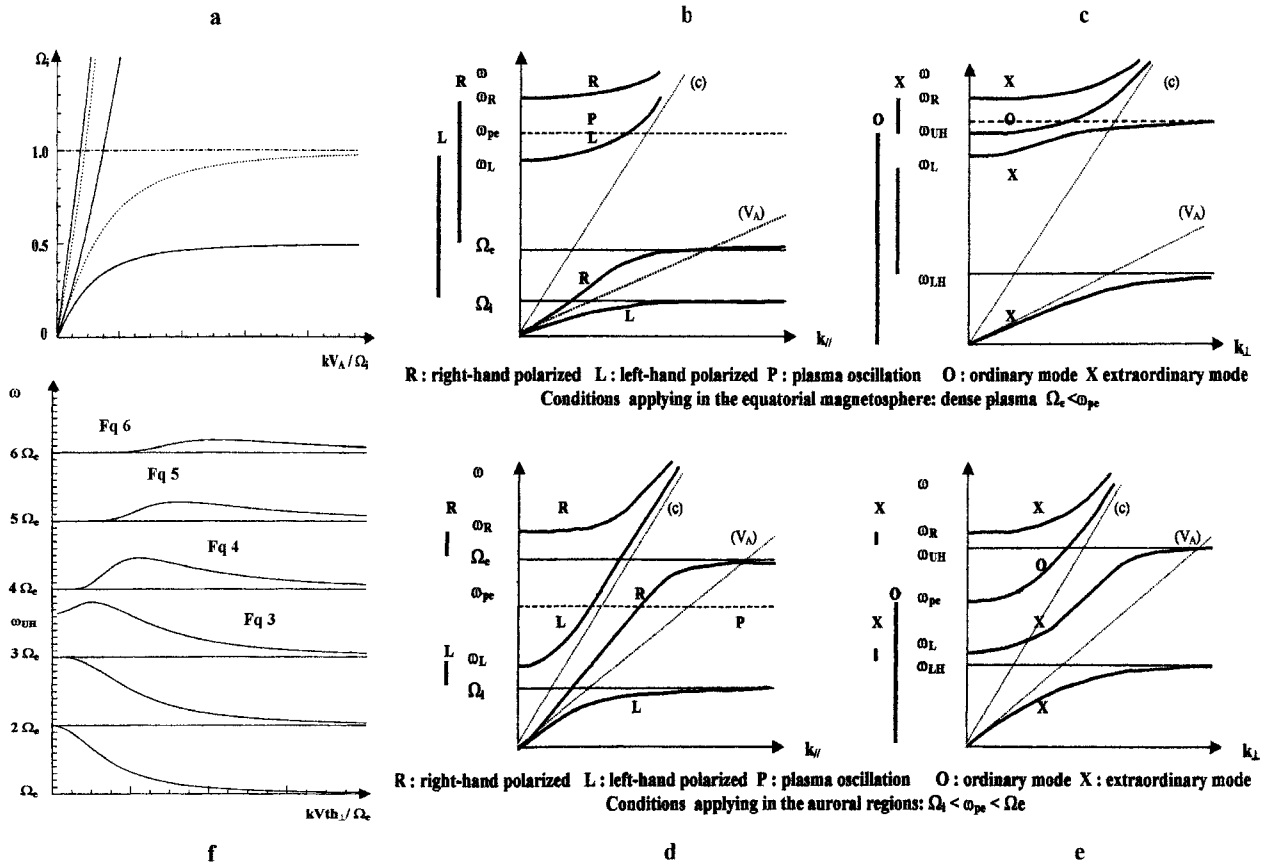


Figure 2. The different wave modes relevant to the magnetosphere in the ω, k plane (dispersion relations). (*a*) MHD modes: there are two modes for cold plasma conditions (dotted curves), fast magnetosonic and shear Alfvén modes from left to right, and three for warm plasma, fast magnetosonic, shear Alfvén and slow magnetosonic modes from left to right. The dispersion relations are given for a propagation angle between $\boldsymbol{k}$ and $\boldsymbol{B}_0$ of 60°.(*b*)–(*e*): Electromagnetic modes in cold plasma, for a single ion species. Black vertical lines correspond to forbidden propagation. (*b*), (*c*) correspond to dense plasma or weak magnetic field ($\Omega_e < (\omega_p)$, which applies in the equatorial magnetosphere, for (*b*) parallel and (*c*) perpendicular propagation. (*d*), (*e*) correspond to $\Omega_i < \omega_p < \Omega_e$ which applies in most parts of the auroral zones, for (*d*) parallel and (*e*) perpendicular propagation. (*f*) The electrostatic Bernstein modes: the frequencies of the dispersion relations' maxima above ω_{uh} are called F_q.

In practice, the cold plasma approximation is often valid, considering that the hot particles are far less numerous than the cold plasma particles, the latter sustaining the propagation of the waves. Nevertheless, in order to be amplified waves need some source of energy: that is mostly provided—at least in a linear phase—by the free energy of the particles, the 'hot' particles. To calculate the amplification of the waves it is necessary to take into account the particle distribution function, in a kinetic treatment, starting from the Vlasov equation.

The exchange of energy between given waves and particles will depend on the possible matching between characteristic particles and wave velocities. This matching is called resonance; some resonances are described below.

Landau resonance—introduction to particle trapping

The Landau resonance is a collisionless interaction between a longitudinal plasma wave and particles whose velocity is nearly equal to the wave phase velocity. The resonance condition is

$$V_{res} = \omega/k. \tag{2}$$

This was first studied by Landau for electrostatic waves propagating in a homogeneous plasma without a magnetic field. By neglecting the temperature of the plasma and considering the ions at rest, one finds the dispersion relation $\omega \approx \omega_{pe}$.

The wave growth rate γ is obtained from the linearized Vlasov equation, where the wave amplitude E is $E = E_0 \exp(\gamma t)$ and f is the particle distribution function, for $V_{res} = \omega/k > 0$:

$$\gamma = (\pi/2)(\omega_{pe}^2/k^2)\partial f/\partial v|_{v=\omega/k}.$$

The important point is that the sign of the growth rate depends of the slope of the distribution function: if there

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

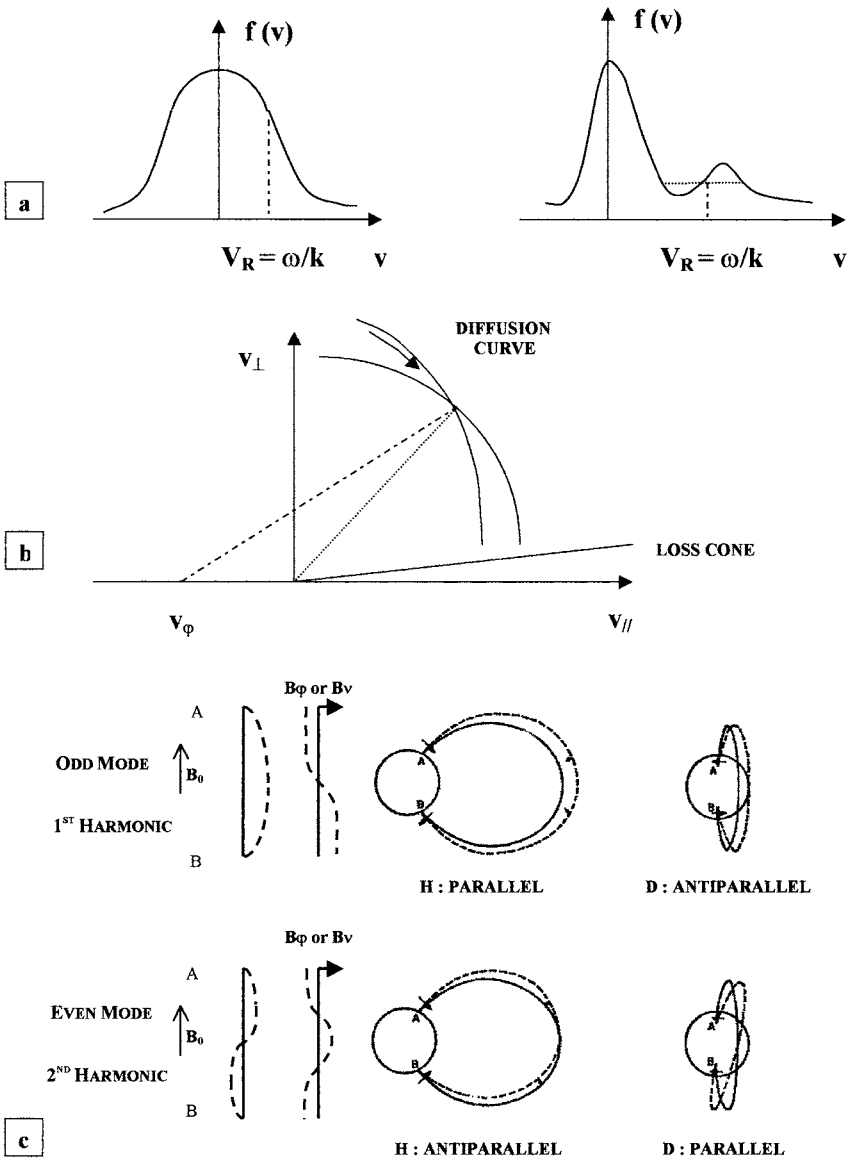


Figure 3. Schematic of wave–particle interaction processes. (*a*) Landau resonance. Landau damping at left, Landau amplification and formation of a plateau at right. V_R is the wave phase velocity. (*b*) Gyroresonance interaction. Particle diffuses along the diffusion curve where the particle energy is constant in the reference frame of the wave V_φ (see text). (*c*) Schematic of the fundamental (top part) and second harmonic (bottom part) oscillations of geomagnetic field lines. Shown from left to right are the oscillation along the field line, the amplitude along the field line of the poloidal (B_v) and toroidal (B_φ) components in a meridian plane, the projection in the meridian plane and a three-dimensional view of the azimuthal component. A and B are the ionospheric footprints of the field lines. The full curves represent the equilibrium state and the broken curves the oscillations.

are more particles at velocity $v = V_{\rm res} + \Delta v$ than particles at $v = V_{\rm res} - \Delta v$, then on average particles can give energy to the waves, the slope of the distribution function being positive (for the case of positive velocities).

In a Maxwellian plasma, $\partial f/\partial v < 0$, and the particles take energy from the wave, damping them. This is called Landau damping (figure 3(*a*), left-hand side). In the opposite case, the wave can take energy from the particles and is amplified. This is what happens when the distribution function has some warm component (e.g. a beam) superimposed on the thermal or Maxwellian distribution. The amplification can continue until the evolution of the distribution function forms a plateau, the slope then being equal to zero (figure 3(*a*), right-hand side). This effect is called quasi-linear diffusion.

The particles that are resonating, or in phase, with the wave are first accelerated and then decelerated as they move through the wave structure. In the frame of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

reference of the wave they oscillate in the wave potential well. If the particles suffer many oscillations, they are trapped. The higher the wave amplitude, the stronger is the trapping effect. The trapping frequency is $\omega_T = (qkE/m)^{1/2}$; the trapping time for one oscillation is $\tau_T = 1/\omega_T$.

The Landau effect, damping or amplification, dominates as long as $\tau_L = 1/\gamma < \tau_T$. Trapping, a non-linear phenomenon, can stop the energy exchange; on average, therefore, it cancels out the Landau effect.

Similar interactions can apply in a magnetized plasma. Trapping may exist for electromagnetic monochromatic waves with a frequency $\omega_T = (kv_\perp qb_1/m)^{1/2}$, where b_1 is the magnetic wave amplitude and $v_\perp$ the perpendicular particle velocity.

Gyroresonant interactions—an example of quasi-linear diffusion

For a plasma in a uniform magnetic field the wave–particle interaction can take place through gyroresonance. This means that, in the frame of reference linked to the parallel velocity of the particle, the particle rotates around the magnetic field at the same angular velocity as the wave, or a multiple of it. The resonance condition is

$$\omega - k_\parallel v_\parallel = N\Omega \tag{3}$$

where N is an integer. The most important cases is $N = \pm 1$; $N = 0$ is the Landau resonance.

We consider the gyroresonance condition (3) with $N = -1$, for which the particle and the wave propagate in opposite directions:

$$V_R = v_\parallel = (\omega - \Omega_{i,e})/k_\parallel. \tag{4}$$

Electrons resonate with right-hand-polarized waves and protons with left-hand-polarized waves.

Because of its resonance with the warm particles, the wave amplitude b can grow, $b = b_0 e^{\gamma t}$, where b_0 is the initial wave amplitude, γ the growth rate and t the time. To calculate the growth rate, the distribution function of the warm population susceptible to resonance with the waves has to be taken into account in the dispersion relation. The solution of the dispersion relation for the angular frequency is complex, $\omega = \omega_r + i\gamma$, where

$$\gamma \propto \eta(V_R)[A(V_R) - A_C] \tag{5}$$

$$A(V_R) \propto \int (\partial f/\partial\alpha)\, d\alpha|_{v_\parallel = V_R}$$

and

$$A_C = \omega/(\Omega_{i,e} - \omega).$$

In these expressions $\eta(V_R)$ is the relative number of resonating particles, A and A_C the particle anisotropy and the critical anisotropy respectively and α is the pitch angle (angle between the direction of the particle velocity and the magnetic field). The wave will grow if γ is positive, which implies that A has to exceed the critical anisotropy A_C. In general the anisotropy depends on V_R, so γ varies with the frequency through equation (4). A is positive if $\partial f/\partial\alpha$ is positive, i.e. if the perpendicular temperature $T_\perp$ exceeds the parallel temperature $T_\parallel$. In the special case of a bi-Maxwellian distribution function, $A = T_\perp/T_\parallel - 1$. As for Landau resonance, the slope of the distribution function is important, A representing the free energy to amplify the waves.

The above expression for the growth rate corresponds to the linear phase: the waves cannot grow indefinitely; some non-linear process will stop their amplification. One process that applies to wide-band wave spectra is quasi-linear diffusion. The waves can diffuse the particles both in pitch angle and in energy, i.e. reducing the pitch angle in order to diminish the anisotropy and diminishing the particle energy respectively, the diffusion in pitch angle being the most important effect. This is schematically shown in figure 3(*b*), where the particle constant-energy curve and the diffusion curve are drawn, the particle energy being constant in the reference frame of the wave on the diffusion curve. One can see that the pitch angle diffusion is faster than the energy diffusion. The diffusion coefficient is proportional to the wave power b^2. The quasi-linear pitch angle diffusion limits the anisotropy and precipitates the particles into the atmosphere (particles that enter the loss cone).

Bounce resonance

Until now we have considered a homogeneous magnetic field. For a slightly inhomogeneous magnetic field on the gyroradius scale, as is the case for the dipolar geomagnetic field, there are other periodicities (or characteristic frequencies) in the particle movement, τ_b (ω_b) the bounce period (frequency) and τ_d (ω_d) the drift period (frequency). The bounce period is the time for the particle to oscillate back and forth along a field line from one hemisphere to the other between magnetic mirror points, and the drift period is the time for particles to drift in azimuth around the Earth. The bounce resonance condition is

$$\omega - N\omega_d = \pm\omega_b \tag{6}$$

where N is an integer and ω/N the east–west angular phase velocity of the wave. Poloidal waves are amplified by taking energy from energetic protons. The resonance will correspond to a breakdown of one or more of the particle's adiabatic invariants, the magnetic moment invariant and the longitudinal invariant. The exchange of energy is

$$dW/dL = q\omega/N B_0 R_E^2/L^2.$$

The resonant particles will diffuse in W, L space, where W is the energy and L the McIlwain parameter.

ULF waves

Ultralow-frequency waves (from 600 to 0.2 s periods) have been studied for many years by means of ground-based magnetometers. ULF waves are separated into

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Classification of ULF pulsations.

	Period (s)
Pc1	0.2–5
Pc2	5–10
Pc3	10–45
Pc4	45–150
Pc5	150–600
Pi1	1–40
Pi2	40–150

continuous (Pc) and irregular (Pi) pulsations, classified in each category, according to their period (table 1).

ULF waves comprise propagating waves and standing oscillations. The latter, in the Pc3–5 range, are mainly resonances of the Earth's geomagnetic field lines.

Pc3–5: field line resonances; fundamental and harmonics
Pulsations in the 10–600 s period range have been observed on the ground for many decades and more recently by spacecraft. Much work has been devoted to both observations and theoretical interpretations. Because of the large spatial scale of the phenomenon, it is not easy to put together all the observations needed. While Pc3–5 waves cover the 1.5–100 mHz range, a given event has a quasi-sinusoidal behavior, with a narrow frequency bandwidth.

From combined results of many statistical and case studies of spacecraft data and ground measurements, these observations can be organized into four main classes: fundamental-mode toroidal resonances, harmonic toroidal resonances, radially polarized pulsations and compressional low-frequency pulsations.

A field line resonance can be understood as an excitation of the magnetic field line by some mechanism, the field line being considered as a string whose two ends are anchored in the ionosphere. As for strings, the oscillation can be either the fundamental or a higher harmonic. The possible resonance frequencies are

$$\omega = \pi N \langle V_A \rangle / \ell \tag{7}$$

where N is an integer, $\langle V_A \rangle$ is the average Alfvén velocity over the field line and ℓ is the length of the field line. The field line resonances are usually interpreted in terms of standing Alfvén waves. These waves generally have toroidal and poloidal components. The poloidal component is in the direction of the principal normal of the magnetic field B_ν and the toroidal component is in the azimuthal direction B_φ. In figure 3(*c*) the first harmonic (top part) and the second harmonic (bottom part) of a field line oscillation are drawn. At left, the modulus of the field oscillation is plotted; the second panel gives the amplitude along the field line of the poloidal and toroidal components. The third column gives the projection in the meridian plane and the last one a three-dimensional view of the azimuthal component. A and B are the ionospheric footprints of the field lines.

The azimuthally polarized (toroidal) Pc5 pulsations are observed predominantly at dawn and at dusk for a large range of L values. The amplitude of the azimuthal component is weak at the geomagnetic equator and increases with the latitude. The period of the oscillation increases with L value (or radial distance from the Earth). These observations are consistent with the hypothesis that they are the fundamental of the field line oscillations: the magnetic component of the odd mode is null at the equator, the amplitude of its oscillation increasing along the field line and reaching a maximum at the ionosphere, the period of the oscillation increasing as the length of the field line increases (larger L value). It is identified as a shear Alfvén mode. The facts that the maximum occurrence is on the flank of the magnetosphere and that the maximum amplitude and occurrence rate are observed for the larger L values ($L > 8$) indicate that these Alfvén waves are excited by surface waves at the magnetopause owing to the Kelvin–Helmholtz instability (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). The latter waves propagate from noon toward both dawn and dusk.

Higher-frequency regular pulsations (Pc3), also azimuthally polarized and observed both on the ground and in space, are the most common Pc activity on the day side, from 06:00 to 16:00 magnetic local time (MLT), with an abrupt cutoff at those MLTs. Spectrogram displays of such waves show equally spaced monochromatic frequencies. They are interpreted as harmonics of the field line resonances, the even modes showing a maximum amplitude in the vicinity of the geomagnetic equator. The uniformity of occurrence and amplitude of the harmonics with radial distance needs to be explained, with respect to the source of excitation. Different external sources have been invoked, linked to interplanetary activity, waves coming from the magnetosheath and global magnetospheric compressions.

Second-harmonic radially polarized oscillations of the field lines are observed in the dusk and night sides of the magnetosphere. They are believed to be amplified by the bounce resonance interaction with ions of some keV.

Storm-time Pc5 waves
The so-called storm-time Pc5 waves seem to be substorm-associated pulsations, detected mainly in the dusk sector near the equatorial region and associated with energetic particle injections (see MAGNETOSPHERE OF EARTH: SUBSTORMS). The drifting energetic ions, among other possibilities, may be responsible for their amplification.

Pc1 or pearl: a way to transfer energy from magnetospheric protons to ionospheric particles
Ground-based study of Pc1 waves intensified when magnetometer chains multiplied and geomagnetically conjugated stations started to operate. The name 'pearls' comes from their waveform time displays, looking like a succession of pearls. Pc1 waves are mostly observed during quiet to moderately active periods, at $3.5 < L < 7$, at about 1 Hz.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

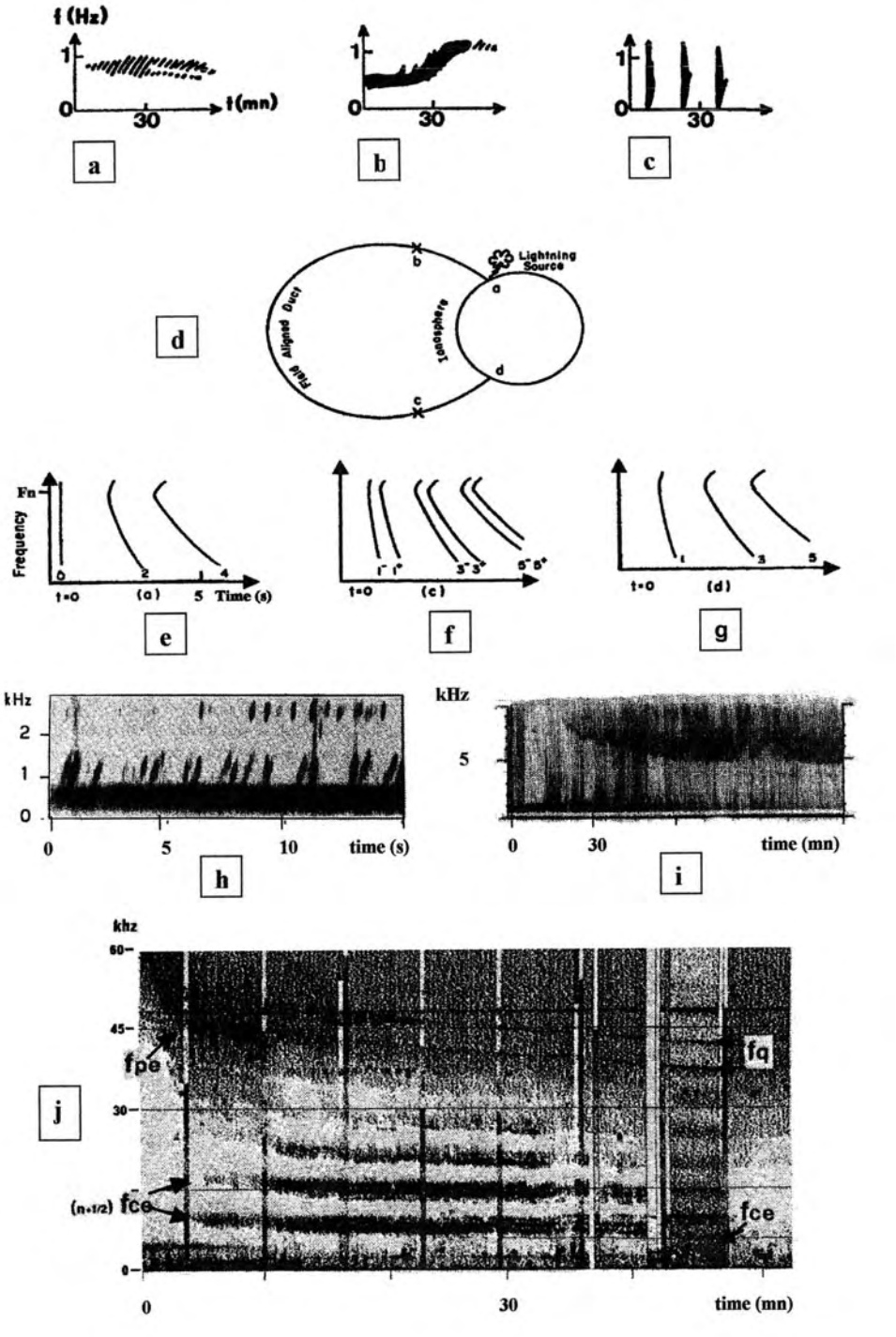


Figure 4. Different examples of dynamic spectra of magnetospheric waves. All are magnetic components, except for the bottom one which is a dynamic spectrum of wave electric components. At the top are schematics of a (*a*) pearl, (*b*) an iregular pulsation of decreasing period (IPDP) and (*c*) a short irregular pulsation (SIP). Just below (*d*) the schematic of an atmosphere and of the resulting whistlers are shown (*e*)–(*g*). (*h*) An example of hiss (broad-band noise) at the top of which are chorus elements; the gap in frequency between the two chorus series is at about half the electron gyrofrequency. (*i*) An example of continuum radiation. (*j*) An example of different electrostatic waves observed at the equator: $F_{\rm pe}$, $(n + 1/2)F_{\rm ce}$, $F_{\rm q}$ frequencies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The spectral analysis of pearls reveals an important characteristic: after a rather monochromatic start, a pearl's frequency spectrum broadens into a succession of dispersive rising tones, separated by a period T of the order of 2 min (figure 4(*a*)). The occurrence at geomagnetically conjugate stations (i.e. located at the two footprints of the same field line) is shifted in time by $T/2$. The observations are consistent with generation by gyroresonance between left-handed waves and ring current energetic protons in the equatorial magnetosphere, the waves propagating back and forth along closed magnetic field lines. T is the wave propagation time between two hemispheres. From the dispersion observed on the ground and the repetition period, the L value and cold plasma density in the generation region at the equator can be inferred. From the initial frequency of the events and the variation with time of the dispersion, combined with theoretical results on wave amplification (equation (5)), characteristics of the resonating particles have been inferred. The frequency spread is interpreted as being due to new waves generated by the trapping of protons in the potential well of the initial wave, giving rise to side bands.

In situ measurements performed in the equatorial region, around $L = 6.6$, were surprising. The expected structure separated by a well-defined repetition period was not observed. In those magnetospheric regions there is an abundance of some 1–2% of cold helium ions. This induces new characteristic frequencies at one of which the mode of propagation of Pc1 changes: the waves are focused in a restricted region of the magnetosphere and the time resolution could not resolve the fine structure, if present, for waves above the helium gyrofrequency. Waves below this frequency can propagate to the ground, and in simultaneous ground–spacecraft data the presence of fine structure has been observed at both locations.

These observations have confirmed the amplification of Pc1 waves by energetic protons and their diffusion into the atmosphere. They have also shown that, in the presence of ionospheric helium ions, the waves transfer via non-linear processes some of the proton free energy to electrons and helium ions of ionospheric origin. The waves in this collisionless plasma thermalize particles. A diagram of the different processes involved is given in figure 5.

Irregular pulsations of decreasing period: a signature of drifting protons?

IPDPs, Pi1 pulsations, are observed on the ground in the evening sector for about 1 h, after substorm onset. They consist of a rather wide band of noise whose central frequency increases with time. A schematic of their dynamic spectra is given in figure 4(*b*). The variation of the frequency has been related to the displacement of the inner boundary of the radiation belts during substorm. The association of IPDPs with proton auroras and their westward drift at a velocity compatible with the proton drift at medium energies (<100 keV) are in favor of the generation of IPDPs by protons. The combined radial and westward drifts of energetic protons under the combined action of the convection electric field, which is reinforced during substorms, and magnetic field gradient drift explain the displacement of the interaction region. One of the effects of the convection on particles is to increase their anisotropy, favoring wave generation, which in turn precipitates protons into the atmosphere. Correlated wave measurements in space only show monochromatic waves: at a given L value, only one limited frequency range is emitted, whereas waves observed on the ground come from different L shells.

From the IPDPs frequency increase, an estimate of the intensity of the convection electric field can be inferred, leading to some tens of mV m^{-1}, a value compatible with other E field measurements.

Short irregular pulsations: a field-aligned current signature

Short-time Pi1 pulsations, called SIPs, cover a rather wide frequency range (figure 4(*c*)). They are mainly observed on the ground during active periods, their intensity being at a maximum in the auroral zone. Spacecraft observations have shown that SIPs observed in the night sector were signatures of field-aligned currents moving quickly in the vicinity of the spacecraft during a substorm. The most intense are correlated with the change of configuration of the geomagnetic field, the dipolarization, signature of a substorm onset. These observations and deduced quantities (a drift velocity of $\cong$ 70 km s^{-1} in the east–west direction, and a current intensity $\cong$ a few 10^{-8} A m^{-2}) are in good agreement with the so-called current disruption model of substorm onset. The signature of the currents is sometimes accompanied by some higher-frequency turbulence.

On the ground, SIPs are also observed on the day side. Are they a signature of field-aligned current there too?

Pi2: signature of substorm onset

Pi2 magnetic pulsations, of longer period than Pi1 (40–150 s), mostly appear at substorm onset. They are interpreted as the signature of disruption of cross tail current at substorm onset, diverted down along the magnetic field lines. Current then flows in the ionosphere and returns to the tail along magnetic field lines, transported along the field lines by shear Alfvén waves. The Pi2s are more clearly seen at ground mid-latitude stations. The analysis of Pi2s simultaneously recorded at different stations permits remote determination of substorm onset local time.

Whistler mode waves in the equatorial magnetosphere

Whistler waves, described below, propagate in the right-hand mode, in the frequency range of some hundred Hz to some kHz, at sonic frequencies, below the electron gyrofrequency. Those propagating parallel to the geomagnetic field lines have been observed on the ground, over many decades, since 1930.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

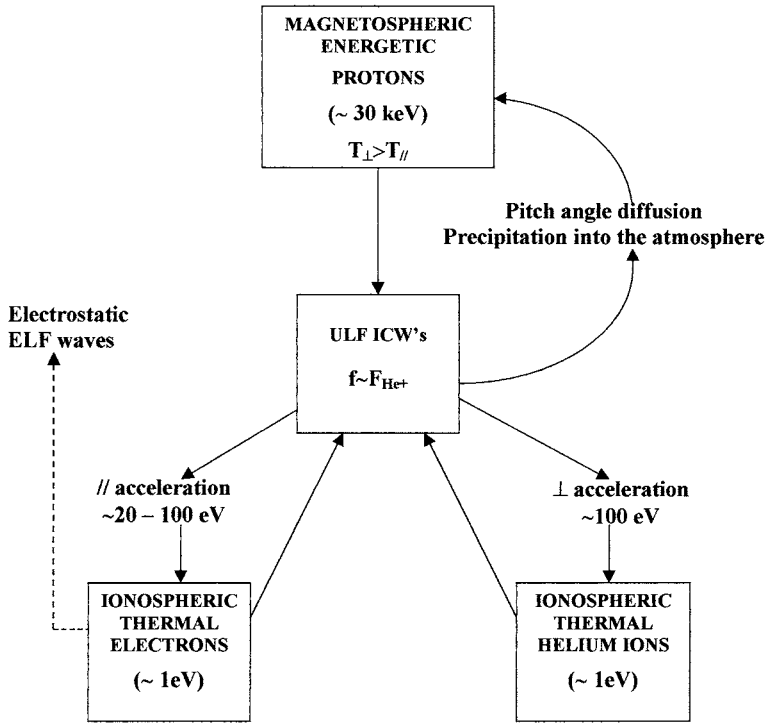


Figure 5. Schematic of the series of wave–particle interaction processes implied in relation to Pc1 waves in a multi-ion plasma: an example of the role of waves in the dynamics of magnetospheric particles. In the equatorial magnetosphere, the anisotropic energetic magnetospheric protons (some 10 keV) amplify ion cyclotron waves that precipitate those protons into the atmosphere by quasi-linear diffusion. In the presence of thermal helium ions originating from the ionosphere, new characteristic frequencies appear. This allows, by non-linear effects, heating of these thermal helium ions and thermal electrons, also of ionospheric origin. The accelerated electrons amplify new electrostatic waves.

Whistlers: a diagnostic tool for equatorial magnetosphere parameters

Lightning discharges are short impulses that excite a wide frequency range of electromagnetic waves. The induced electromagnetic waves propagate from the atmosphere into the ionosphere along a field line and then into the magnetosphere on slightly overdense ducts, in the right-hand mode (figures 4(*d*)–4(*g*); figure 4(*d*) shows their trajectory in a meridian plane). Their group velocity V_g, or propagation velocity, depends on the frequency; the frequency corresponding to the minimum group velocity is called the nose frequency, F_N (see figures 4(*e*)–4(*g*)). They can be observed on the ground in the opposite hemisphere using a radio receiver (figures 4(*e*) and 4(*g*)). The sound thus obtained, generally a falling tone, gave their name to such waves, whistlers. They can also be observed on board spacecraft (figure 4(*f*)). A given emission can propagate back and forth along a field line, as can be seen on the schematic spectrograms (figures 4(*e*)–4(*g*)). Similarly to what is done with pearls, from these dispersion characteristics and from the time delay between the successive hops information on the plasma traversed can be inferred. With some reasonable hypothesis for waves propagating inside the plasmasphere (dipolar magnetic field, cold and dense plasma), it can be shown that the nose frequency $F_N \cong 0.4F_{ce,eq}$, $F_{ce,eq}$ being the equatorial electron gyrofrequency, allowing one to determine L. The density at the equator is determined by the propagation time $t_g \propto (n_{e,eq})^{1/2}$. For frequencies above F_N the cold plasma hypothesis is no longer valid, and, from the observed deviation from this, the electron temperature can be inferred. When the dawn–dusk electric field varies, inducing a change in the plasmapause position, the electric field can be deduced from the variation of L with time.

Hiss: a way to regulate the radiation belt electron flux

Hiss is a very common phenomenon in the near-Earth magnetosphere, outside and inside the plasmasphere, predominantly inside the plasmasphere (plasmaspheric hiss). It is a rather broad-band stationary electromagnetic emission, some hundred Hz wide, in the frequency range from 100 Hz to some kHz. *In situ* observations have confirmed that hiss is the result of a gyroresonant interaction with electrons in the energy range from some keV to some tens of keV. The low-frequency part of figure 4(*h*) is an example of hiss.

Spacecraft measurements have allowed a comparison of wave spectra with the experimental growth rate γ calculated by means of the electron distribution functions. The growth rate γ (relation (5)) maximizes at the wave

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spectrum peak. Calculation of the anisotropy has confirmed that the quasi-linear diffusion of particles by the amplified waves was efficient: $A > A_C$ when hiss is present, and A is very close to A_C when the amplitude, and thus the diffusion, is strong. The role of hiss in diffusing the particles into the atmosphere, thus limiting the flux of electrons in the radiation belts, has been confirmed.

Chorus: an example of non-linear effects

The peculiarity of chorus is its spectral shape, a succession of quasi-monochromatic rising tones (or sometimes falling tones), spaced in time by some hundred of ms. Their frequency range lies from some 100 Hz to some kHz, at higher frequencies than hiss. Figure 4(*h*) gives an example of simultaneous observation of hiss and chorus; the gap between the two series of chorus elements is at about half the local gyrofrequency. They are preferentially generated after substorms, when new energetic electrons are injected and drift eastward around Earth, and are observed outside the plasmasphere at dawn and on the dayside. Chorus generation in its initial phase is due to gyroresonant interaction between whistler mode waves and electrons (some keV). Their spectral shapes are due to non-linear effects, whose starting point may be the monochromatic structures, evidenced in the concomitant hiss, observed just below the frequency of chorus elements. Interpretation of chorus fine structure is a topic of research. Trapping of the electrons by the monochromatic waves and detrapping by the inhomogeneity in the geomagnetic field are being considered. One difficulty is in measuring the electron distribution function with time resolution as chorus elements. Globally, chorus participates in the precipitation of electrons into the atmosphere.

Waves observed above the electron gyrofrequency in the equatorial magnetosphere

Above the electron gyrofrequency F_{ce}, numbers of wave types are observed in the equatorial magnetosphere (figure 4(*j*)). The full set of Bernstein mode waves is observed: electrostatic electron cyclotron harmonic (ECH) waves between the electron gyroharmonics below F_{UH}, F_{UH} waves and waves at the F_q frequencies (figure 2(*f*)). As can be seen from figures 2(*b*) and 2(*c*), there is a wide forbidden frequency range for the electromagnetic mode between F_{ce} and F_{pe}; it is not surprising that observed waves in this range are electrostatic, electromagnetic waves being at and above F_{UH}.

Electrostatic electron cyclotron harmonic waves at $(n+1/2)\,F_{ce}$: a source for the diffuse auroras?

Electrostatic ECH waves are a common feature of the equatorial magnetosphere, just outside the plasmasphere, in an L value range of field lines having their footprint in the auroral zone (∼6–7). They are confined in the equatorial region, amplified by 100 eV to some keV electrons whose distribution function is of the loss cone type, containing regions where $\partial f/\partial v_\perp > 0$, created by the convection of the plasma from the tail toward the Earth. The lowest harmonic ($(3/2)F_{ce}$) is the strongest and the most often present. From the first case studies ECH waves were thought to be responsible for the diffuse auroras as a result of pitch angle diffusion into the loss cone. Statistical studies have shown that the required amplitude to provoke strong diffusion of 1 keV electrons ($\simeq 2$ mV m^{-1}) was only reached 2% of the time, while diffuse auroras are a permanent feature. $(3/2)F_{ce}$ waves cannot be the only cause of diffuse auroras. A complementary mechanism has to be found. The higher harmonics ($n > 1$) being inefficient in diffusion, other candidates are proposed to precipitate these electrons, such as an acceleration by field-aligned currents or a diffusion by chorus. Figure 4(*j*) gives an example of multiple $(n + 1/2)F_{ce}$ waves observed in the equatorial magnetosphere.

F_{UH} and F_q

The natural lines at the upper hybrid frequency F_{UH} and at F_q frequencies are weak, the most intense being in the close vicinity of F_{UH} (or F_{pe}, as it is very difficult to experimentally differentiate F_{UH} and F_{pe} for large F_{pe}/F_{ce} values). A very weak line at F_{pe} should always be present, reflecting the fluctuation of the electrons due to their thermal agitation, which maximizes very close to F_{pe} but is often hidden either by the lack of sensitivity of the instruments or by stronger emissions. When observable it is a tool to determine the plasma properties: the spectrum peak gives the plasma frequency (and thus the density) and its width gives the thermal temperature of electrons. Sometimes the F_{UH} emission becomes very intense and broadens, amplified by a loss cone distribution function of some 100 eV electrons. Figure 4(*j*) gives an example of weak F_q's and of F_{pe} emission, first rather intense and then of weak amplitude.

Non-thermal continuum radiation

The electromagnetic non-thermal 'continuum' radiation is observed in the 5–100 kHz frequency range, above the local plasma frequency. It is observed in two forms, a series of discrete lines and a smooth continuum (figure 4(*i*)), hence its name. If emitted above the solar wind electron plasma frequency (∼ 20 kHz), it can escape out of the magnetosphere, the escaping continuum. It is highly structured in frequency, composed of narrow-band lines. If generated below the solar wind plasma frequency, it is trapped in the magnetosphere. The different emissions are mixed and superimposed after multiple reflections, forming a broad-band continuum. Sources at the plasmapause have been deduced from multispacecraft measurements. Observations in the tail suggest that the plasma sheet boundary layer and/or the magnetopause can also be source regions.

The exact generation mechanism of the non-thermal 'continuum' is still under debate. Whereas it is generally admitted that the starting point is an electrostatic emission around F_{UH} originating in density gradients, different processes are proposed to convert the electrostatic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

emission into an electromagnetic one. Linear conversion, possible in density gradients, is probably inefficient (∼1%). Non-linear coupling is more efficient, such as the coalescence of F_{UH} waves with low-frequency electrostatic waves to form the electromagnetic emission or the decay of the F_{UH} wave into the electromagnetic wave and a low-frequency wave.

Auroral emissions

Among the wave–particle interactions in the auroral zone, the most striking effects are related to the processes which accelerate ionospheric ions toward the magnetosphere, the generation of radio emissions and the possible role of auroral waves in the acceleration of electrons which cause visible auroras.

AKR: a radio source

The Earth is a strong radio source that has been discovered by high-altitude Earth-orbiting spacecraft. The emission is generated in the auroral night side region, in the 50–700 kHz frequency range, at altitudes varying between 5000 and 10 000 km. The emission is powerful (10^7–10^8 W) and is called auroral (or terrestrial) kilometric radiation (AKR or TKR). The plasma frequency of the dense ionospheric layers below this altitude is much higher than the AKR frequency and acts as a cutoff frequency for the radio emission; thus it is not observed on the ground. The occurrence of AKR is correlated with auroras and accelerated electrons, thought to be the source of AKR, by the cyclotron maser instability (see AURORAL KILOMETRIC RADIATION).

Other auroral emissions: a way to accelerate electrons and ionospheric ions

As well as AKR, many kinds of waves are observed in the auroral zone. To account for the acceleration of various particle species, a combination of different wave–particle interaction mechanisms has to be considered. The complete picture of the roles of the different waves has not yet been obtained.

From ground observations, very little was known about auroral waves as compared with subsequent progress made as a result of rockets and polar orbiting spacecraft. The auroral acceleration region of particles is located between 65° and 75° geomagnetic latitude at an altitude between 1000 and 25 000 km. The parallel electric field in this region corresponds to a potential drop of 1–10 kV. How the waves participate in the build up of the parallel electric field is one of the main questions. A field-aligned electric field can be produced by the separation between electrons and ions due to the difference in their mirroring point along a field line but also by waves, since parallel electric field are associated with kinetic Alfvén waves. Small-scale localized solitary structures can add up to produce a significant potential drop. Oxygen electromagnetic ion cyclotron waves can also accelerate electrons. All these phenomena can contribute to down-going electron acceleration. Up-going observed electron acceleration can be achieved by fluctuating parallel electric fields at about 1 Hz or by an interaction with lower hybrid resonance (LHR) waves.

The positive parallel slope in the distribution function of accelerated electrons can generate different kinds of observed waves: auroral hiss, LHR waves, plasma waves or upper hybrid resonance (UHR) waves (see figure 2(*d*) and (*e*)).

An important source of magnetospheric particles is the outflow of ionospheric ions. Before escaping, the ions are accelerated perpendicularly to the field line and then they move upward under the action of the mirror force, or conservation of the first adiabatic invariant: when the ion reaches higher altitudes its parallel energy increases at the expense of its perpendicular energy, as the field strength decreases. The heating can be significant from a few eV to about 100 eV. However, how are the ions accelerated perpendicularly to the magnetic field? Different observed waves are candidates; electrostatic ion cyclotron waves (EICWs) or LHR waves, amplified by electrons, when very intense can collapse into small-scale structures that could accelerate ions. It is also proposed that down-going Alfvén waves suffer a mode conversion to ion cyclotron waves when the Alfvén wave frequency matches the ion gyrofrequency (for either H^+ or O^+) leading to a resonant absorption that heats the ions. This conversion mode can occur in a multi-ion component plasma, which is the case in the auroral region. As these heated ions move along the field line they interact with waves at different frequencies, the local ion frequency varying as the geomagnetic field does, causing broad-band electromagnetic waves. A statistical study correlating wave and ion measurements has shown that most of the oxygen outflow is correlated with broad-band electromagnetic ULF–ELF waves for all local time sectors.

Conclusion and perspectives

Not all magnetospheric waves have been described, in particular those occurring at the magnetopause, in the cusp region or in the tail. The near-equatorial region has been rather well explored and the global role of the waves and wave–particle interactions has been pretty well understood, although some non-linear processes remain to be elucidated. The corresponding importance of different processes occurring in the auroral regions is still to be established. Furthermore, it remains for us to understand the physical processes by which solar wind particles enter into the magnetosphere and the role of waves in such processes. This is one of the aims of the CLUSTER project to be launched in 2000. Its four identical spacecraft will successively fly through the key regions that are the polar cusp, the magnetopause and the tail. Its three-dimensional configuration will allow us to separate temporal and spatial effects, derive differential quantities such as div and curl, and in particular locate wave source regions, determine $\boldsymbol{k}$ vector distribution functions and establish wave polarization.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Definitions

Alfvén velocity. $V_A = B/(\mu_0\rho)^{1/2}$; μ_0 is the magnetic permittivity of free space and ρ is the mass density of the plasma; V_A can also be written $V_A = c\Omega/\omega_p$.

Frequency. $F = \omega/2\pi$, ω being the wave pulsation or angular frequency. ω is commonly used for various relations, dispersion relations and velocities. For observed waves, in Hz, F is more often used.

Gyrofrequency. This is the angular velocity of the charged particles in the presence of a magnetic field, in the plane perpendicular to the magnetic field: $\Omega_{i,e} = q_{i,e}B/m_{i,e} - (F_{ci}, F_{ce})$, where q, m are the charge, mass of the considered species and B is the intensity of the magnetic field ; a useful relation is F_{ce} (in Hz) $= 28B$ (B in nT).

Gyroradius. Radius of gyration of the charged particle: $r_g = v_\perp/|\Omega_{i,e}| = mv_\perp/|q|B$.

Group velocity. $V_g = \partial\omega/\partial k$, gives the direction of the wave energy flow.

L-value or McIlwain parameter. L characterizes a magnetic field line $L = R/R_E$ at the equator, R being the geocentric distance and R_E the Earth radius.

Loss cone—first adiabatic invariant. $\mu = mv_\perp^2/2B$ is the magnetic moment of the particle (ratio between the perpendicular energy of the particle and the magnetic field). This quantity can often be considered as constant and is called the first adiabatic invariant. Combined with the conservation of the particle energy along its trajectory, one obtains $\sin^2\alpha_1/\sin^2\alpha_2 = B_1/B_2$. A particle mirror on the field line at the point where its pitch angle $\alpha = 90^\circ$, being then trapped on this field line. If the point is too low (in the atmosphere), the particle is lost by collision. The corresponding equatorial angle range for which the particles are lost is called the loss cone.

Lower hybrid resonance (LHR). $\omega_{LH} \approx (\Omega_e\Omega_i)^{1/2}$ for one ion species.

Phase velocity. $V_\varphi = \omega\boldsymbol{k}/k^2$, direction of propagation and speed of the wave front, parallel to $\boldsymbol{k}$

Pitch angle. Angle between the particle velocity and the direction of the geomagnetic field $\boldsymbol{B}_0$, usually called α.

Plasma frequency. $\omega_{pi,e} = (n_{i,e}q_{i,e}^2/m_{i,e}\varepsilon_0)^{1/2} - F_{pe}$, F_{pi}, i or e referring to ions or electrons and ε_0 being the vacuum dielectric constant; a useful relation is F_{pe} (in kHz) $=9n_e^{1/2}$ (n_e in particles cm^{-3}).

Spectrogram. This is a visualization in three dimensions of the time evolution of the waves in a frequency–time plane. The time is the abscissa, the ordinate is the frequency and the amplitude is given by the intensity of the signal, either in a gray scale or in color. At each time, it is the result of a Fourier transform of the waveform signal.

Upper hybrid resonance (UHR). $\omega_{UH} = (\omega_{pe}^2 + \Omega_e^2)^{1/2}$.

Bibliography

Anderson B J, Engebretson M J, Rounds S R, Zanetti L J and Potemra T A 1990 A statistical study of Pc3–5 pulsations observed by the Ampte/CCE magnetic fields experiment 1. Occurrence distributions *J. Geophys. Res.* **95** 10 495–523

André M 1997 Waves and wave–particle interactions in the auroral region *J. Atmos. Terr. Phys.* **59** 1687–712

Baumjohan W and Treuman R A 1997 *Basic Space Plasma Physics* (London: Imperial College Press)

Belmont G, Fontaine D and Canu P 1983 Are equatorial electron cyclotron waves responsible for diffuse auroral electron precipitation? *J. Geophys. Res.* **88** 9163–70

Helliwell R A 1965 *Whistlers and Related Ionospheric Phenomena* (Stanford, CA: Stanford University Press)

Kennel C F and Petschek H E 1966 Limit on stably trapped particle fluxes *J. Geophys. Res.* **771** 1–28

Kurth W S, Gurnett D A and Anderson R R 1981 Escaping nonthermal continuum *J. Geophys. Res.* **86** 5519–31

Lester M, Hughes W J and Singer H J 1983 Polarization patterns of Pi2 magnetic pulsations and the substorm current wedge *J. Geophys. Res.* **88** 7958–68

Orr D 1973 Magnetic pulsations within the magnetosphere: a review *J. Atmos. Terr. Phys.* **35** 1–50

Sazhin S S and Hayakawa M 1992 Magnetospheric chorus emissions: a review *Planet. Space Sci.* **40** 681–97

Sazhin S S, Hayakawa M and Bullough K 1992 Whistler diagnostics of magnetospheric parameters: a review *Ann. Geophys.* **10** 293–308

Solomon J, Cornilleau-Wehrlin N, Korth A and Kremser G 1989 Generation of ELF electromagnetic waves and diffusion of energetic electrons in steady and non-steady situations in the Earth's magnetosphere *Plasma Waves and Instabilities at Comets and in Magnetospheres (Geophysical Monograph 53)* ed B T Tsurutani and H Oya pp 119–33

Southwood D J 1981 Low frequency pulsation generation by energetic particles *ULF Pulsations in the Magnetosphere* ed D J Southwood (Dordrecht: Reidel)

Stix T H 1992 *Waves in Plasmas* (New York: American Institute of Physics)

Treuman R A and Baumjohan W 1997 *Advanced Space Plasma Physics* (London: Imperial College Press)

Young D T, Perraut S, Roux A, de Villedary C, Gendrin R, Korth A, Kremser G and Jones D 1981 Wave–particle interactions near Ω_{He^+} observed on GEOS 1 and 2.1. Propagation of ion cyclotron waves in He$^+$-rich plasma *J. Geophys. Res* **86** 6755–72

Nicole Cornilleau-Wehrlin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetospheres: Jupiter, Radio Emissions

The magnetosphere of Jupiter (see JUPITER: MAGNETOSPHERE) is an immense particle accelerator, allowing free electrons in it to reach kilo- to mega-electron-volt energies (keV to MeV). These energetic electrons, moving along and circling around Jovian magnetic field lines, produce intense electromagnetic emissions all over the radio range, with wavelengths ranging from millimeters to tens of kilometers.

Among the dozen or so Jovian radio components, we emphasize here the *synchrotron radiation* from energetic (MeV) particle belts in the decimeter range, and the *cyclotron radiation* from keV electrons in high magnetic latitude regions (auroral field lines, and those linked to the Io–Jupiter electrodynamic circuit (see MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS)).

Discovery and observations of Jovian radio emissions

Between 1955 and 1958, radio emissions from Jupiter were discovered using ground-based radiotelescopes, at frequencies between 10 and 5000 MHz (wavelengths between 30 m and 6 cm). These emissions are extremely intense: if they had a thermal origin, the source temperatures should be ~10^4–10^5 K at 6 cm, and 10^{18} K at 30 m! The latter are so unphysically high that these radio emissions were attributed to non-thermal mechanisms, involving the motion of free electrons in a magnetic field. This was the first inference of a magnetic field, and hence a magnetosphere, for a planet other than Earth. These interpretations were confirmed by high-frequency radio maps like that of figure 1, and by *in situ* measurements of the magnetic field and charged particles around Jupiter by the Pioneer 10 and 11 and Voyager 1 and 2 spacecraft in the 1970s. At the same time, the Earth's magnetosphere and auroral radio emissions were discovered (see MAGNETOSPHERE OF EARTH, MAGNETOSPHERE OF EARTH: WAVES and MAGNETOSPHERE: AURORAL KILOMETRIC RADIATION), and in the 1980s Voyager 1 and 2 revealed the magnetic fields, magnetospheres and radio emissions of Saturn, Uranus and Neptune (see SATURN: MAGNETOSPHERE, URANUS AND NEPTUNE: ATMOSPHERES, IONOSPHERES AND MAGNETOSPHERES).

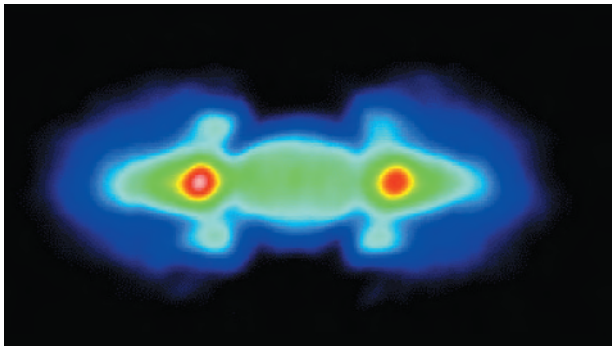


Figure 1. Map of decimeter radio emission (DIM) around Jupiter at a wavelength of 22 cm. The synchrotron emission originates from electron populations near the equator and at high latitudes, and dominates the thermal emission from the planetary disk. The brightness asymmetry between the two ansas comes from the non-dipolar character of the Jovian magnetic field. (From Leblanc Y *et al* 1997 *Astron. Astrophys.* **319** 274–81.)

Jovian radio emissions have been studied from the ground since 1955, at frequencies ≧10 MHz, with various centimeter/decimeter wavelength antennas, as well as decameter arrays (table 1). At lower frequencies, the Earth's ionosphere reflects radio waves, and observations have been performed from space with experiments on board near-Earth spacecraft (RAE 1 and 2 in low Earth orbit, and Wind at the L1 Lagrange point), space probes which flew by Jupiter (Voyager 1 and 2 in 1979 and Ulysses in 1992), and a Jupiter orbiter (Galileo, in orbit since 1995). These various instruments measured the radio emission intensity and polarization, and their variations versus time and frequency.

At frequencies below a few hundred MHz, the angular resolution (λ/D, with λ the radio wavelength, and D the antenna dimension, typically <100–1000 m) becomes very poor, and no direct information can be obtained on the radio source locations. Those must then be

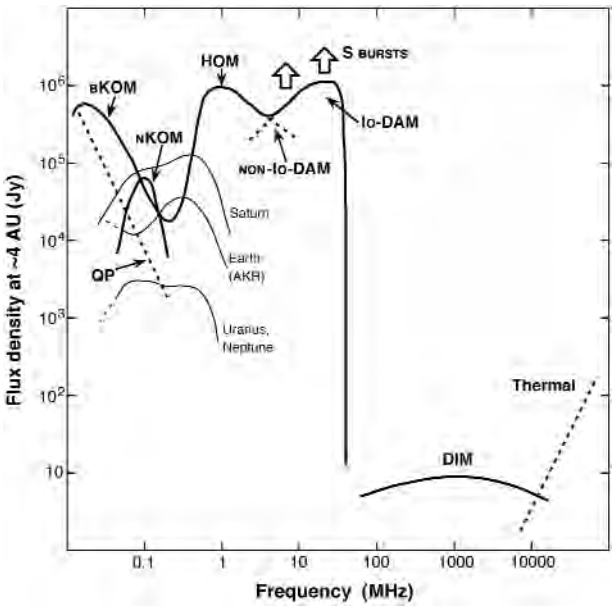


Figure 2. Spectra of Jovian radio components (bold lines) and auroral kilometric radiations from the other giant planets and the Earth (light lines). Average emission levels are sketched as they would be detected from the minimum Earth–Jupiter distance (~4.2 AU). Peak flux densities may exceed the displayed one by up to one order of magnitude, and Jovian S burst fluxes may reach 10^7 Jy (1 Jy = 1 jansky = 10^{-26} W m^{-2} Hz^{-1}). The low-frequency limit of DAM control by Io is somewhere between 1 and 3 MHz. (Adapted from Zarka 1998.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 1. Some key properties of Jovian radio magnetospheric components.

Component	*Frequency range*	*Source location*	*Beaming*	*Power*	*Emission process*	*Discovered/observed by*
Thermal	≧10 GHz	Atmosphere	Isotropic	$<10^9$ W below 10 GHz	Thermal	GB/HF
DIM	≦100 MHz →~10 GHz	Radiation belts (Van Allen)	Mostly $\perp \boldsymbol{B}$	≧10^9–10^{10} W	Synchrotron	GB/HF
Non-Io DAM	A few MHz → 40 MHz	Along auroral field lines ($\lambda_B \approx 67°$–$75°$)	Hollow beam at large angle/$\boldsymbol{B}$	10^{10}–10^{11} W	Cyclotron-maser	GB/LF Voyagers
Io DAM (L bursts/ S bursts)	A few MHz → 40 MHz	Along Io and Io's torus field lines ($\lambda\alpha_B \approx 65°$)	Hollow beam at large angle/$\boldsymbol{B}$ (70°–80°)	Up to 10^{11} W	Cyclotron-maser (?)	GB/LF Voyagers, Wind
HOM	200 kHz → a few MHz	Along auroral field lines ($\lambda_B \approx 68°$–$72°$)	Hollow beam at 30°–90°/$\boldsymbol{B}$	10^9–10^{10} W	Cyclotron-maser	Voyagers, Ulysses, Galileo
bKOM	~10 kHz →>300 kHz (up to 1 MHz)	Along auroral field lines ($\lambda_B \approx 70°$–$75°$)	Hollow beam at 30°–80°/$\boldsymbol{B}$	10^8–10^9 W	Cyclotron-maser	Voyagers, Ulysses, Galileo
nKOM	~50 kHz →200 kHz	Plasma	Broad filled inhomogeneities in Io's plasma torus pattern, around radial direction	~10^8 W	Upper hybrid or plasma frequency?	Voyagers, Ulysses, Galileo
QP bursts/ non-thermal continuum	A few kHz → 50/700 kHz	Auroral? Magnetopause?	? ?	10^7–10^8 W	Cyclotron-maser? Conversion?	Voyagers Ulysses

GB/HF = ground-based high-frequency (centimeter/decimeter wavelength) antennas, including the Very Large Array in New Mexico, the Arecibo radiotelescope in Puerto Rico, etc.

GB/LF = ground-based low-frequency (decameter) arrays in Nançay (France), Kharkov (Ukraine) or Gainesville (Florida).

λ_B = magnetic latitude (the Jovian dipole field makes a tilt of 9.6° relative to the rotation axis).

DIM = decimeter; DAM = decameter; HOM = hectometer; bKOM = broadband kilometer; nKOM = narrowband kilometer; QP = quasi-periodic.

determined indirectly. The only exception is Ulysses, whose radio experiment was able to measure the direction of arrival of incoming radio waves, although only below 1 MHz.

However, while the spectrum of thermal and decimeter emissions (figure 2 and table 1) is very stable in time, it becomes very sporadic and structured in the frequency–time plane at lower frequencies, and gains several orders of magnitude in intensity below 40 MHz. It is thus possible to build frequency-time images of intensity or polarization variations, with high signal-to-noise ratio. These 'dynamic spectra' permit us to follow the rapid variations of emission characteristics (down to a few milliseconds) with a good spectral resolution (down to a few kHz or less) over a broad spectral domain, and these constitute the main tool compensating for the dramatic lack of angular resolution at long wavelengths. Figure 3(*a*) displays a typical dynamic spectrum of low-frequency Jovian radio emissions recorded by Voyager 2 in the kilometer-to-decameter wavelength range, during one complete rotation of the planet (9 h 55.5 min). Figure 3(*b*) is a dynamic spectrum recorded from the ground (in Nançay), displaying the fine temporal structure of decameter emissions named 'S' bursts ('S' stands for 'short').

Characteristics of Jovian radio components

Almost a dozen Jovian magnetospheric radio components have been identified in some 40 years of ground-based and spacecraft observations (see Carr *et al* 1983, Zarka 1998 and references therein). This number is larger than for any other magnetized planet, and Jovian emissions also reach higher intensities. This is mainly due to the intense, quasi-dipolar Jovian magnetic field, together with the fact that most planetary radio emissions are cyclotron emissions: the larger amplitude $|\boldsymbol{B}|$ of the near-surface magnetic field at Jupiter ((4.3–14) –10^{-4} T, compared with <1–10^{-4} T for the other planets) corresponds to higher electron cyclotron frequencies ($f_{ce} = e|\boldsymbol{B}|/2\pi m_e$, with e the charge of the electron and m_e its mass). The large Jovian magnetic dipole moment also creates a huge magnetospheric cavity in the solar wind,

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

with a radius varying between 60 and 120 Jovian radii. The solar wind ram pressure (or kinetic energy flux) thus provides a large input power on this magnetospheric obstacle (~10^{15} W), a fraction of which feeds electron acceleration. Jupiter's rapid rotation is also an effective energy source. Finally, large numbers of accelerated electrons converge towards the magnetic poles along auroral field lines or are trapped within Van Allen 'radiation' (actually energetic particle) belts (see MAGNETOSPHERE OF EARTH: RADIATION BELTS), where they produce intense radio waves.

In addition, the plasma populating every other PLANETARY ATMOSPHERE comes from the solar wind or the planet's ionosphere, whereas a third important source exists at Jupiter: the satellite Io, whose volcanism produces a dense plasma torus around its orbit (see IO: PLASMA TORUS). An electrodynamic circuit forms from the interaction of Io with the Jovian magnetic field, driving strong currents and leading to electron precipitation towards Jupiter and intense radio emissions in the decameter wavelength range.

The exact number of radio components of truly distinct origin is not known because some of them may be different manifestations of the same source and/or phenomenon, for example over different frequency ranges. Table 1 is an up-to-date list of some key properties of these radio components. Their average spectra, as they would be detected from the minimum Earth–Jupiter distance (~4.2 AU), are plotted on figure 2, while figure 3(*c*) sketches how these various emissions appear in the frequency–time plane. Figure 4 sketches the corresponding radio source locations in the Jovian magnetosphere. Detailed information can be found in Rucker *et al* (1997) and Zarka (1998).

Auroral radio components

Part of the decameter emission (DAM) is strongly related to the satellite Io (the Io DAM component), i.e. it is observed only for specific configurations of Io–Jupiter–observer, and the rest is independent of Io (non-Io DAM). Both Io and non-Io DAM emissions are structured in 'arcs' in the frequency–time plane (figure 3(*a*)), and various modulations and fringes may affect them, due to diffraction by plasma inhomogeneities near Jupiter or in the interplanetary medium. Arc shapes have been explained in terms of the complex topology of the near-surface Jovian magnetic field (non-dipolar terms) and of the non-isotropic directivity of the emission. The

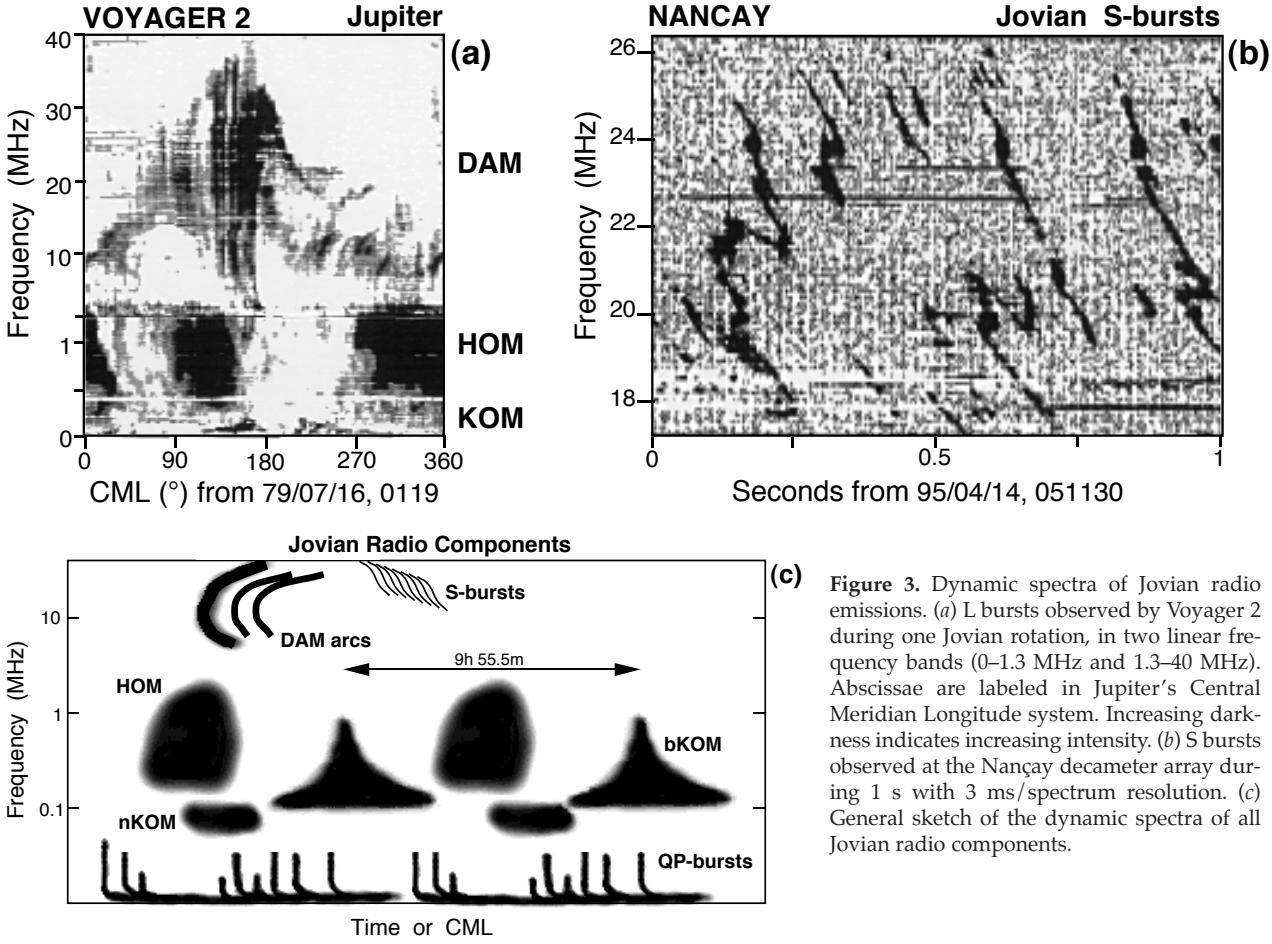


Figure 3. Dynamic spectra of Jovian radio emissions. (*a*) L bursts observed by Voyager 2 during one Jovian rotation, in two linear frequency bands (0–1.3 MHz and 1.3–40 MHz). Abscissae are labeled in Jupiter's Central Meridian Longitude system. Increasing darkness indicates increasing intensity. (*b*) S bursts observed at the Nançay decameter array during 1 s with 3 ms/spectrum resolution. (*c*) General sketch of the dynamic spectra of all Jovian radio components.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

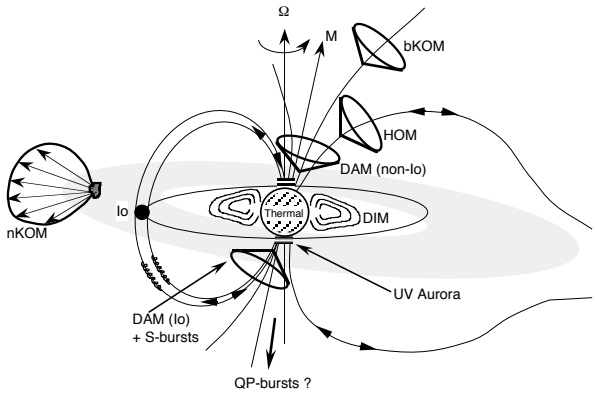


Figure 4. Radio source locations in Jupiter's internal magnetosphere and Io plasma torus. The nKOM is emitted in ~70° wide beams from quasi-point sources in the torus, while the auroral emissions (bKOM, HOM, DAM and possibly QP bursts) are produced near the local gyrofrequency and beamed in hollow cones with axes aligned with the high latitude magnetic field lines. Arrows indicate field lines along which energetic electron precipitation is known to occur.

hectometer component (HOM), independent of Io, is thought to be the low-frequency extent of non-Io DAM. Its source and that of bKOM have been directly located by Ulysses. The broadband kilometer (bKOM) source overlaps with that of HOM, but these components appear distinct on dynamic spectra (figures 3(*a*), (*c*)). They may have the same generation mechanism, driven by different sources of energetic electrons.

DAM, HOM and bKOM constitute the Jovian auroral (high magnetic latitude) emissions, in many respects similar to Earth's auroral kilometric radiation. They are generated between the Jovian surface and a few planetary radii above it, along field lines at the footprints of which the precipitated electrons generate UV, visible and IR auroras. Io DAM is linked to the Io–Jupiter current system, and can be considered as a secondary 'auroral' source, at a slightly lower latitude than non-Io DAM.

The low-frequency part of the Jovian radio spectrum, between a few kHz and a few tens of kHz, had been attributed after Voyager flybys to non-thermal continuum radiation generated at the dayside magnetopause. Ulysses observations attributed at least part of it to the merging of the low-frequency part of ubiquitous bursts recurring with quasi-periods (QP) of 15 to 40 min. Their origin is unclear, but they may be auroral too.

Time variations

DIM to bKOM components are strongly modulated by the planetary rotation (9 h 55 min 23.7 s), due to their non-isotropic directivity. Actually, magnetospheric radio emissions are the only means to determine the rotation rate of the giant planet's interior, where the magnetic field originates, through the thick cloud layers.

The narrowband kilometer component (nKOM) lags the planetary rotation by a few per cent, due to its origin in Io's torus, whose plasma is not in perfect co-rotation with the planet.

Fluctuations of the solar wind parameters (ram pressure, density, velocity and magnetic field) induce modulations of the bKOM, HOM, non-Io DAM, and possibly QP bursts, but do not affect Io DAM. This is expected because the radio sources of all these components are in 'external' regions of the magnetosphere, while that of Io DAM is along field lines connected to Io's torus, in the inner magnetosphere. But solar wind variations also modulate the occurrence of nKOM, although it originates from the torus itself. The penetration of the solar wind influence there remains to be explained.

Finally, Io DAM is observed with two strikingly different morphologies: it may consist, like non-Io DAM, of slowly varying emissions ('L' bursts on timescales of minutes; 'L' stands for 'long'; figure 3(*a*)), or consist of a series of S bursts (each lasting a few tens of milliseconds; figure 3(*b*)). These two morphologies correspond to different regimes of energy release, and possibly to different generation mechanisms.

Ganymede's radio emissions

In addition to the above Jovian components, Galileo observations revealed the existence of the magnetosphere and radio emissions of the satellite GANYMEDE (Gurnett *et al* 1996), thus providing evidence of a hierarchical organization of embedded magnetospheres in the solar system. Ganymede radio emissions are not auroral, but appear similar to the continuum radiation from planetary magnetospheres. The observed energetic particle environment of Ganymede, and the periodic reconnection of its magnetic field with that of Jupiter, suggests that it might control some Jovian radio emissions in a similar but probably less prominent way than Io.

Generation mechanisms and acceleration processes

The generation of all magnetospheric radio emissions (except nKOM) requires energetic electrons moving in a magnetic field. The intensity and polarization of the radio components are the main constraints permitting us to identify the generation mechanism (Goldstein and Goertz 1983).

The moderate intensity and linear polarization of DIM indicate a synchrotron emission process, non-thermal but incoherent (where the intensity produced by each electron adds up to form the total emission). Detailed maps as in figure 1 have been successfully modeled using an isotropic and an equatorial, trapped electron distribution. The decimeter (DIM) emission is much weaker or non-existent at the other giant planets because of energetic particle absorption by the rings. The Jovian

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

DIM was enhanced for several weeks after the collision of Comet Shoemaker–Levy 9 with Jupiter in 1994.

The circular or elliptical polarization of DAM, HOM and bKOM, and their extremely high intensities, rather require a coherent, maser-like emission mechanism. Ulysses measurements definitely proved that these radio emissions are produced close to the local electron cyclotron frequency, f_{ce}, at each point along auroral magnetic field lines, on the so-called extraordinary propagation mode. This confirmed the cyclotron-maser mechanism as the best candidate for their generation. This mechanism consists of coherent generation at the cyclotron frequency by unstable electron populations containing free energy susceptible to amplify waves, in the form of an excess of electrons with high velocity perpendicular to the local magnetic field. The elliptical polarization of DAM implies that the plasma density within the radio sources is extremely low (possibly <5 cm^{-3}), which is a favourable condition for the cyclotron-maser mechanism.

Electron acceleration processes are not yet firmly identified. Magnetospheric compressions, observed by Ulysses, may lead to the so-called 'Fermi' and 'betatron' acceleration of electrons trapped on magnetic field lines. After several Jovian rotations, electrons may thus reach the MeV energies required by the synchrotron generation of DIM. Auroral emissions require rapid acceleration processes, which include magnetic reconnection in the magnetotail, parallel electric fields (observed in Earth's auroral regions and near Io's magnetic flux tube), and heating by low-frequency plasma waves (observed by Ulysses near Io's torus, or possibly originating from the current disk in the case of QP bursts). Parallel electric fields may be electrostatic or non-stationary (so-called Alfvén waves). The latter are thought to be generated by the interaction between Io and the Jovian magnetic field (Io–Jupiter electrodynamic circuit).

Multispectral (UV/IR/radio) studies of the 'Io' auroral source permit estimates of the energy of precipitated particles and allow us to compute an energy budget: $\sim 10^{12}$ W are dissipated in each hemisphere by the Io–Jupiter circuit, mostly in the form of Alfvén waves converting into electron acceleration (keV to tens of keV); the latter precipitate into an area of $\sim 10^5$ km^2 in the Jovian ionosphere, producing (2–3)–10^{10} W in the IR and UV ranges, and up to 10^9 W above the ionosphere in the DAM radio range. The S bursts are so intense that the cyclotron-maser mechanism might be not efficient enough.

Finally, nKOM was shown by Ulysses to be generated by discrete sources in Io's torus at the local electron plasma frequency (proportional to the square root of the electron density) or upper hybrid frequency (the square root of the sum of the squares of the electron plasma and cyclotron frequencies). These sources might be density inhomogeneities, waxing and waning with time (in relation to solar wind fluctuations). The specific radio generation mechanism has not yet been identified.

Interest of the study of Jovian radio emissions and perspectives

Although representing a small fraction ($\sim 10^{-5}$) of the input solar wind power on the magnetospheric cross section, radio emissions (together with auroral UV and IR emissions) are essential for remote sensing studies. Their intensity, polarization and frequency–time structure are directly linked to the topology of the Jovian magnetic field and energetic electron distributions at the radio sources, which can thus be remotely monitored. The knowledge gained in understanding the generation of Jovian radio emissions will be applied to other distant radio sources (solar, stellar), that no spacecraft can reach. Extrapolating the fact that Jovian DAM is the most intense radio emission in the solar system, it appears interesting to search for exoplanets through their possible magnetospheric radio emissions.

Web Update (31 July 2002)

New observations of Jupiter, its satellites and environs have been made by Cassini, Galileo, the Hubble Space Telescope, the Chandra X–ray Observatory and Earth–based radio and optical telescopes during the CASSINI spacecraft's flyby . The Galileo spacecraft has been in orbit around Jupiter since December 1995 while Cassini observed Jupiter from October 1, 2000 to March 31, 2001. Cassini made its closest approach on December 30, 2000, at a distance of about 9.7 million kilometers (6 million miles).

Radio emissions from Jupiter are believed to be generated along the high–latitude magnetic field lines by the same electrons that produce the aurorae, and both the radio emission in the hectometric frequency range and the aurorae vary considerably. Simultaneous observations using the Cassini and Galileo spacecraft of hectometric radio emissions and extreme ultraviolet auroral emissions from Jupiter show that both of these emissions are triggered by interplanetary shocks propagating outward from the Sun. When such a shock arrives at Jupiter, it seems to cause a major compression and reconfiguration of the magnetosphere, which produces strong electric fields and therefore electron acceleration along the auroral field lines, similar to the processes that occur during geomagnetic storms at the Earth.

Ground–based observations have shown that Jupiter is a two–component source of microwave radio emission: thermal atmospheric emission and synchrotron emission from energetic electrons spiraling in Jupiter's magnetic field. Observations of 13.8 GHz synchrotron emission collected during the Cassini fly–by confirm the presence of electrons with energies up to 50

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

MeV. These energetic electrons may be repeatedly accelerated through an interaction with plasma waves, which can transfer energy into the electrons. Preliminary comparison of the data with model results suggests that electrons with energies of less than 20 MeV are more numerous than previously believed. The most widely accepted theory postulates that the relativistic electrons originate in the solar wind or in the outer Jovian magnetosphere and diffuse inward, gaining energy through conservation of the first and second adiabatic invariants. This can account for electrons with energies around 10 MeV near the peak of the synchrotron zone. However, additional acceleration mechanisms are required to explain higher energies.

References

Nature, 28 Feb. 2002 **415** 987–991

Web Update references

Gurnett D A *et al* February 2002 Control of Jupiter's radio emission and aurorae by the solar wind *Nature* **415** 985–987

Bolton S J *et al* February 2002 Ultra–relativistic electrons in Jupiter's radiation belts *Nature* **415** 987–991

Bibliography

Carr T D, Desch M D and Alexander J K 1983 Phenomenology of magnetospheric radio emissions *Physics of the Jovian Magnetosphere* ed A J Dessler (Cambridge: Cambridge University Press) pp 226–84

Goldstein M L and Goertz C K 1983 Theories of radio emissions and plasma waves *Physics of the Jovian Magnetosphere* ed A J Dessler (Cambridge: Cambridge University Press) pp 317–52

Gurnett D A, Kurth W S, Roux A, Bolton S J and Kennel C F 1996 Evidence for a magnetosphere at Ganymede from plasma wave observations by the Galileo spacecraft *Nature* **384** 535–7

Rucker H O, Bauer S J and Lecacheux A (ed) 1997 *Planetary Radio Emissions IV* (Vienna: Austrian Academy of Science Press)

Zarka P 1998 Auroral radio emissions at the outer planets: observations and theories *J. Geophys. Res.-Planet* **103** 20159–94

Philippe Zarka

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Magnetospheres: Jupiter, Satellite Interactions

Most of the satellites of Jupiter, notably the large Galilean satellites Io, Europa, Ganymede and Callisto (see JUPITER: SATELLITES), orbit deep inside the magnetosphere of Jupiter (see JUPITER: MAGNETOSPHERE) and are therefore immersed in the flow of magnetospheric plasma (made of a mixture of electrons and ions) and subjected to an interaction with the strong Jovian magnetic field. These interactions have dramatic consequences for the Jovian magnetosphere, the satellites of Jupiter and even the ionosphere of Jupiter, particularly in the case of the innermost large satellite Io.

Since during the first three decades of research in this area, observations of this interaction were available only for Io, ideas were initially developed for this case. The early history of this case began with the discovery of the control of radio emissions from Jupiter by Io (see MAGNETOSPHERES: JUPITER, RADIO EMISSIONS). It was followed by first attempts to physically explain these observations. The next important step was the discovery of an ionosphere of Io by the PIONEER 10 spacecraft. A more advanced understanding was developed in response to ground-based observations and the many significant results of the VOYAGER MISSION. The GALILEO MISSION has revealed that all four of the Galilean satellites show interactions with significantly different features and a challenging diversity of physical processes.

The Io effect and the unipolar inductor model

In 1964 the so-called Io effect was discovered by using data on the radio emissions from Jupiter in the decametric range of wavelengths (several tens of meters). It was found that the radio emissions are strongly controlled by the orbital position of the satellite Io with respect to the observer. It was generally concluded that an electrodynamic interaction of Io with the Jovian magnetic field was responsible for the Io effect. The simple idea of the unipolar inductor model of this interaction is shown in figure 1, where Io is shown much larger than in reality for clarity. The unipolar inductor model starts from the idea that the plasma and magnetic field are rotating rigidly with Jupiter with a velocity at the orbit of Io which exceeds the orbital velocity by about 57 km s^{-1}. An observer at rest with respect to Io then observes a motional electric field directed outward from Jupiter. Across the diameter of Io this electric field corresponds to a voltage of about 500 000 V applied to Io. This voltage arises much in the same way as in an electric generator, where electric wires are moved across a magnetic field. This voltage then drives a current system first through Io, as shown in figure 1. The current circuit is then continued along the magnetic field lines which can be viewed as highly conducting wires. The circuit is finally closed by currents through the Jovian ionosphere perpendicular to the magnetic field. When the currents pass through the Jovian ionosphere a 'hot spot' is expected to form as a consequence of the electric heating and precipitation of energetic particles.

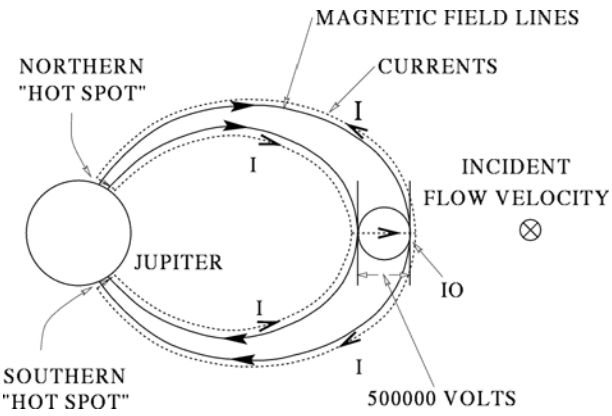


Figure 1. Principle of the unipolar inductor model (not to scale). The corotating plasma flows towards Io into the plane of the drawing. The voltage drop is based on realistic magnetic fields.

The center of this hot spot is defined by the footpoints of the magnetic field lines passing through Io in this simplified picture. Early estimates yielded currents of 1 MA in each branch of the current system based on a voltage of 700 000 V across Io using a somewhat too high magnetic field of Jupiter at that time. The unipolar inductor model also involved the idea that electrons accelerated along the magnetic field lines provide the energy for the emission of radio waves. The frequency of the radio emissions was associated with the local electron cyclotron (or gyro) frequency (the frequency of the gyro motion of electrons around the magnetic field), an idea which is basically accepted today. In this initial work the current was assumed to be closed through the conducting interior of Io. This unrealistic assumption was later replaced by current closure through a conducting ionosphere of Io, which was discovered by a radio occultation experiment on the first spacecraft that explored Jupiter, i.e. Pioneer 10 in 1973. This discovery of an ionosphere implied the existence of a neutral atmosphere around Io. Another far-reaching discovery was the ground-based observation of clouds of neutral gas like sodium atoms with Io as the source.

The Voyager encounters and the Alfvén wing model

The Voyager flybys at Jupiter in 1979 yielded a number of important new observational results, which better defined the problem, but also sparked a number of new theoretical ideas, which led to a much more mature picture of Io's electrodynamic interaction. An SO_2 atmosphere was discovered in the infrared which provided an initial definition of the neutral atmosphere as the basis of the ionosphere discovered by Pioneer 10 and confirmed much later by the Galileo spacecraft. This highly dynamic neutral atmosphere which was later investigated by the Ultraviolet Explorer and the Hubble Space Telescope (HST) is basically due to volcanic activity on Io driven by tidal friction in its interior. Another discovery by Voyager was the Io plasma torus (see IO: PLASMA TORUS),

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a population of plasma mostly composed of sulfur and oxygen ions, originating from Io and forming a torus at the radial distance of Io from Jupiter and trapped in the Jovian magnetic field. The high plasma densities in the torus with electron concentrations of several 1000 cm^{-3} at the center of the torus and the energy balance of the torus requires mass loss rates from Io of about 1 ton s^{-1}. This makes Io comparable to a comet of intermediate gas production at 1 AU. In contrast to the comets in orbit around the Sun (see SOLAR WIND: INTERACTION WITH COMETS) Io is moving through its own ion tail. This is because the torus plasma is moving at the fast rotation rate of Jupiter, thereby overtaking slowly moving Io all the time. Thus plasma leaving Io through its wake is reencountered on the front side approximately a rotation later. Io is thus the major source of plasma mass in Jupiter's magnetosphere.

The huge densities (for the magnetospheres of the outer planets) also imply a large inertia of the plasma flow incident on Io. This requires an improvement of the unipolar inductor model described above. The unipolar inductor model must be replaced by the more general and more accurate Alfvén wing model. This is based on the radiation of Alfvén waves (see MAGNETOHYDRODYNAMIC WAVES) by the satellite. This radiation of Alvén waves is similar to the establishment of a bow shock by a supersonic aircraft. In contrast to the sound waves forming a bow shock, the Alfvén waves are restricted to move parallel or antiparallel to the magnetic field for an observer in the rest frame of the magnetospheric plasma. Figure 2 schematically shows the formation of Alfvén wings in the rest frame of Io. Because of the velocity of the magnetospheric plasma the wings are inclined with respect to the unperturbed magnetic field of Jupiter. The northern and southern Alfvén wings distort the magnetic field due to their currents. During the flyby of Voyager 1 at Io the Alfvén wing signature was identified successfully, indicating currents of about 3 MA in one branch of the wing. Figure 2 also illustrates that the Alfvén wings must be reflected either by the ionosphere of Jupiter or at the torus boundary. The reflected Alfvén wave must pass behind Io on its way to the opposite hemisphere. If it passes far behind Io there is no feedback of the Alfvén wave on Io. If Io has not moved very far before the reflected Alfvén wave returns, there will be an interaction of the reflected Alfvén wave with Io. In the extreme case that the reflected Alfvén waves pass through Io's ionosphere several times, the unipolar inductor model is obtained as an extreme case of very close coupling of the Jovian ionosphere and Io's atmosphere–ionosphere system. Figure 2 also suggests that a large number of reflected wing portions may trail the satellite. These reflected wings have also been proposed as the explanation for another significant discovery by Voyager, i.e. the discovery of multiple arcs in frequency–time diagrams of the decametric radio emissions. The Alfvén wing configuration is expected to vary in response to changes in torus density and the magnetic latitude of Io describing its position in the torus. It varies between the pure Alfvén wing picture without feedback to a configuration with some features of the unipolar inductor model. The plasma flow through Io's atmosphere and its electric conductivity are determined by a number of processes of atomic and molecular physics. These are referred to as ion pick-up, charge exchange and elastic collisons. In the wake region there is an excess of plasma leaving Io's environment (relative to the incoming plasma), thereby forming an ion tail. The flow also drags neutral atoms and molecules away to feed the neutral clouds referred to above. These processes describe the loss of mass from Io to be replaced by volcanism.

Advanced Earth-based observations and the Galileo mission

The description of the interaction between Io and the Jovian magnetosphere includes a 'hot spot'. It has been described above as the footpoint of the disturbance system radiated by Io parallel and antiparallel to the magnetic field toward the Jovian ionosphere. Observations of this footpoint contain important diagnostic information on the nature of this disturbance system and also the Jovian magnetic field. First observations of the footpoint were obtained in the infrared in 1993 and with increasing sensitivity by the HST in recent years. The Io-related disturbance system consists of a 'bright spot' with a wake trailing the spot over a large part of a circle around Jupiter.

A new leap forward in our knowledge on the magnetospheric interaction of the Galilean satellites came through the many encounters of the Galileo orbiter with the satellites which is continuing at the time of writing. On 7 December 1995 the Galileo spacecraft passed through the wake of Io at an altitude of 900 km. The main features of the observations were the detection of a deep dip in magnetic field magnitude, an almost stagnant ionospheric plasma near closest approach, and a system of plasma waves which are at least partly due to the pick-up of ions of Ionian origin. The higher torus density during this encounter compared with conditions observed by Voyager in 1979 suggest a somewhat stronger interaction than at that time. It has been suggested that the observations may require the presence of an internal magnetic field although it is not clear yet whether the disturbances in the magnetoplasma around Io really require such a field. During the Galileo encounter at Io peculiar distributions of electron beams in the energy range from 15 keV to above 140 keV parallel and antiparallel to the magnetic field were observed which can be considered the first direct evidence for acceleration processes due to the magnetospheric interaction of Io. They are also of appreciable interest for the mechanisms generating radio emissions from Jupiter's environment.

Until recently there was no observational evidence for the nature of the interaction between the three Galilean satellites outside Io, i.e. Europa, Ganymede and Callisto, and the Jovian magnetosphere. The situation was changed by the HST observations of an oxygen atmosphere of Europa just before the arrival of Galileo at Jupiter. In the meantime there is also

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

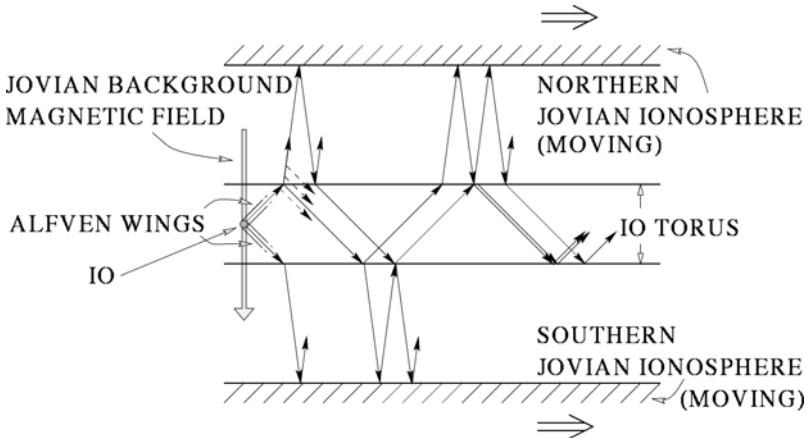


Figure 2. Idealized sketch of the Alfvén wing system generated by Io viewed in the rest frame at Io. In this frame the magnetospheric plasma and the ionosphere of Jupiter are moving toward the right. Reflected wings are shown with reflections at the torus boundary, ionosphere of Jupiter and repetitions thereof.

evidence for an oxygen atmosphere on Ganymede and a CO_2 atmosphere on Callisto. Without any further information this suggests the following scenario for these satellites. Photoionization and collisional ionization by magnetospheric electrons produce ionospheres, which interact with the magnetosphere of Jupiter via the general Alfvén wing model as explained above. A substantial difference with the case of Io exists, however. The neutral atmosphere must be maintained against the loss processes due to the magnetospheric interaction in the absence of the violent volcanism at Io. For Europa the idea has been developed that sputtering by magnetospheric ions generates the observed oxygen atmosphere. A similar idea might be applied to Ganymede and Callisto.

It came as a complete surprise that Ganymede possesses a substantial internal magnetic dipole field, possibly of dynamo (see DYNAMO THEORY) origin. With an equatorial surface field of the dipole of 750 nT roughly antiparallel to the satellite's spin axis, Ganymede is surrounded by its own magnetosphere which itself is embedded in the huge Jovian magnetosphere. Broad Alfvén wings are radiated northward and southward from the polar caps of Ganymede's magnetosphere. An idealized sketch is shown in figure 3. The magnetosphere has closed field regions at low and intermediate latitudes.

The magnetospheric interaction with the remaining satellites Europa and Callisto cannot be explained by the unmodified Alfvén wing model, as the flyby observations show. To explain the observations a magnetic field of internal origin is required again. However, the observations particularly at Callisto show that a dipole moment fixed to the satellite does not explain the observations. The only alternative is an internal field due to electromagnetic induction generating eddy currents. Such an internal field requires an electrically conducting interior of the satellite, which may be provided by subsurface oceans with high electric conductivity due to leaching of salts. The eddy currents are excited because

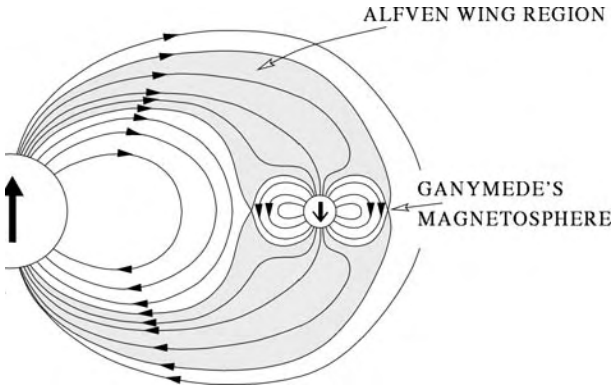


Figure 3. The magnetosphere of Ganymede inside the magnetosphere of Jupiter (not to scale)

the Jovian magnetic field has a time varying component at each satellite, which is due to the tilt of the Jovian dipole moment with respect to the spin axis of Jupiter. The induced fields provide a component of field perturbations which has in principle to be assessed at all satellites.

Bibliography

Hill T W, Dessler A J and Goertz C K 1983 The Io Jupiter interaction *Physics of the Jovian Magnetosphere* ed A J Dessler (Cambridge: CUP) pp 365–72

Khurana K K, Kivelson M G, Stevenson D J, Schubert G, Russell C T, Walker R J, Joy S and Polanskey C 1998 Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto *Nature* **395** 777–80

Kivelson M G, Khurana K K, Russell C T, Walker R J, Warnecke J, Coroniti F V, Polanskey C, Southwood D J and Schubert G 1996 Discovery of Ganymede's magnetic field by the Galileo spacecraft *Nature* **384** 537–41

Neubauer F M 1998 The sub-Alfvénic interaction of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the Galilean satellites with the Jovian ionosphere *J. Geophys. Res.* **103** 19 843–66

Fritz Neubauer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetospheres: Plasma-Ring/Dust Interactions

Dust particles in space are often immersed in plasmas and UV radiation and, consequently, carry electrostatic charges. The motion of small micron- and submicron-sized charged grains can be influenced or even dominated by electric and magnetic fields in addition to gravity, drag and radiation pressure. Dust particles in plasmas are unusual charge carriers. They are many orders of magnitude heavier than any other plasma particles (electrons and ions) and they can have time-dependent charges many orders of magnitude larger (negative or positive). Small grains act as plasma probes since their charging and dynamics are set by the plasma environment.

Figure 1 shows one of the most unexpected findings of the Voyagers at SATURN, the sporadically appearing, radially expanding dust clouds above the dense B-ring (see also SATURN: RINGS). The 'spokes' consist of charged submicron-sized dust particles that are lofted above the rings due to electromagnetic forces (Mendis *et al* 1984). Meteorites bombarding the rings generate plasma clouds that rapidly drift away from the synchronous orbit (the distance where the rotation rate of the planet matches the Kepler orbital rate) and lift off submicron-sized grains from the ring plane as they move above the rings. Each dust particle follows an approximate Kepler orbit so that grains in the cloud closer to Saturn move faster. As 'spokes' age the Kepler shear results in an increasing tilt and the contrast between the dust cloud and the ring diminishes.

The 'spokes' were the first evidence that, contrary to general belief, gravity alone cannot explain all of the complex dynamics in PLANETARY RINGS. The ring/halo region and the dust streams at Jupiter and the E-ring at Saturn are other examples where the magnetospheric plasma environment shapes the size and spatial distribution of small dust particles.

Electrostatic charging of dust in plasmas

The electrical charge of a dust grain, Q, changes due to the currents reaching its surface

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \sum_k J_k \tag{1}$$

where J_k represent various electric currents. Electron and ion collection, and secondary and photoelectron emission currents, dominate dust charging in most space plasmas (Whipple 1981). All of these currents are functions of the plasma properties (density, composition, energy distribution) and also the properties of the dust grain (size, velocity, composition, surface roughness). As a dust grain collects charges it changes the electrostatic potential distribution in its environment. A grain that initially collects more electrons than ions develops a negative potential well that enhances the ion flux and lowers the electron flux reaching its surface. The electrostatic surface potential of a grain that balances these fluxes is the equilibrium potential. For an isolated dust grain, with radius a, the surface potential can be easily related to the net charge, $\phi = Q/a$, resulting in

$$Q_\mathrm{e} \simeq 700\phi_\mathrm{v} a_\mu \tag{2}$$

where Q_e is the number of extra/missing electrons, ϕ_v is the surface potential measured in volts and a_μ is the radius of the grain in microns.

In a case where the electron and ion thermal currents are the only charging currents, the equilibrium potential of a grain in a plasma where the electron and ion temperatures ($T_\mathrm{e} = T_\mathrm{i} = T$) are the same is

$$\phi_\mathrm{eq} = -\beta\frac{kT}{e} \tag{3}$$

where, for example, $\beta = 2.5$, 3.6 and 3.9 for H^+, O^+ and S^+ plasma respectively. At this grain potential $J_\mathrm{e}(\phi_\mathrm{eq}) = J_\mathrm{i}(\phi_\mathrm{eq})$. In a $kT = 1$ eV hydrogen plasma $\phi_\mathrm{eq} = -2.5$ V, and a 1 μm radius particle will collect approximately 1800 extra electrons. Because of the statistical nature of the collection currents, charges fluctuate around their equilibrium value and the electron and ion fluxes balance only on average. The electrons and ions bombarding a grain that is in charge equilibrium might recombine on the surface and return to the plasma as neutral particles.

Electrostatic charging can change the physical characteristics of the dust particles. For example, the electrostatic tension might overcome the tensile strength of a grain, resulting in chipping off protruding fingers or even full disruption (Grün *et al* 1984).

The presence of dust can alter the properties of any plasma environment acting as a source or sink for its density, momentum and energy. In a low-density region, for example, photo- and secondary electron production might become an important plasma source. Similarly, energetic plasma particles can sputter off atoms and molecules from dust grains, changing the density as well as the composition of the plasma environment (Johnson 1990).

In fact, if the presence of dust significantly alters the plasma density the charging equations for a single dust particle discussed above are no longer valid. In the case of significant plasma depletion ($Qn_\mathrm{d} \simeq en_\mathrm{p}$, where n_d is the density of the dust grains with a mono-dispersed size distribution for simplicity, Q is the charge calculated for a single grain and n_p is the plasma density) the charge of a dust particle in a cloud of other dust particles is reduced.

Often grains traverse regions where the plasma parameters change rapidly and the grains do not accommodate quickly enough to assume their equilibrium potential. Charged dust grains in PLANETARY MAGNETOSPHERES can rapidly trade energy and angular momentum with the electric and magnetic fields due to their time-dependent electrostatic charges. Hence, the evolution of the spatial distribution of small dust particles is often shaped by the plasma environment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

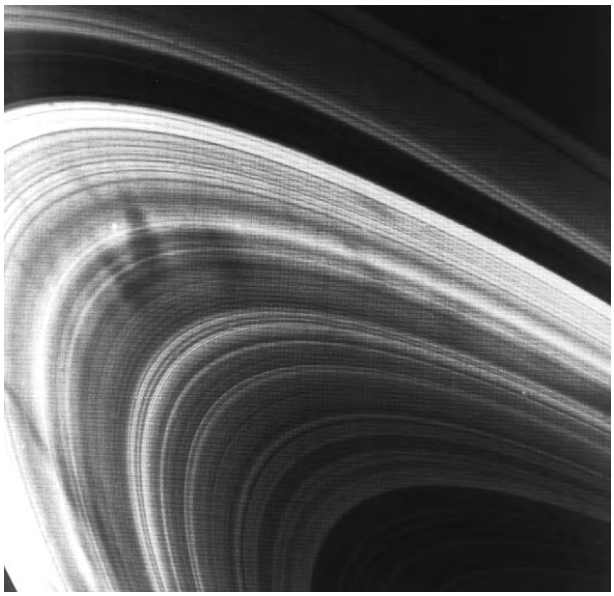


Figure 1. Image of Saturn's rings taken by Voyager 2 showing 'spokes', the dark radial markings.

The simultaneous integration of the current balance and the equation of motion represent the basic approach of 'gravito-electrodynamics', that considers the motion of charged dust grains where electrodynamic forces are just as important as gravity.

Dynamics of charged dust

The equation of motion of a charged dust grain (of mass m and charge Q) using an inertial coordinate system fixed to the planet's center, is

$$\ddot{\boldsymbol{r}} = \frac{Q}{m}(\dot{\boldsymbol{r}} \times \boldsymbol{B} + \boldsymbol{E}_c) - \frac{\mu}{r^3}\boldsymbol{r} \tag{4}$$

where $\boldsymbol{r}$ is the grain's position vector, $\boldsymbol{B}$ is the magnetic field, and μ equals the gravitational constant times the planet's mass. For an infinite conductivity magnetosphere that rigidly co-rotates with the planet with a rotation rate of Ω, $\boldsymbol{E}_c = (\boldsymbol{r} \times \Omega) \times \boldsymbol{B}$ is the co-rotational electric field. On the right-hand side there could be other terms describing forces due to satellites, higher-order terms in the planet's gravitational field, radiation pressure, and plasma and neutral drag forces, for example.

The magnetic fields of Jupiter (J) and Saturn (S) can be approximated as that of a simple dipole located at the center with its magnetic moment aligned with the rotation axis of the planet (ignoring, for the moment, the 10° tilt between the rotation axis and the dipole moment of Jupiter). In the equatorial plane of a dipole $B = B_0(R/r)^3$, where B_0 is the magnetic field at the surface and R is the radius of the planet ($B_0^{\rm J} = 4.2$, $B_0^{\rm S} = 0.21$ G and $R_{\rm J} = 7.1 \times 10^4$, $R_{\rm S} = 6 \times 10^4$ km). The magnetic field lines pierce the equatorial plane at right angles pointing anti-parallel to Ω. The resulting co-rotating electric field points radially outward with an amplitude of $E_c = E_0(R/r)^2$ ($E_0^{\rm J} = 5.8 \times 10^{-5}$, $E_0^{\rm S} = 2.4 \times 10^{-6}$ V m^{-1}). The force acting on dust particles associated with the co-rotational electric field also depends on their charge, $F_{\rm E} = QE_c$. The charge in turn, as discussed above, is a function of the plasma environment, the material properties of the grain and its charge.

For an average density of $\rho = 1$ g cm^{-3} the ratio of the electrostatic force to gravity

$$F_{\rm E}/F_{\rm G} = \alpha\phi_{\rm v}a_\mu^{-2} \tag{5}$$

($\alpha^{\rm J} \simeq 6\times10^{-3}, \alpha^{\rm S} \simeq 5\times10^{-4}$). In the typical range of $-50 < \phi_{\rm v} < +10$, for particles with $a_\mu \gg 1$, electrostatic forces represent a perturbation only and these grains follow approximate Kepler orbits. On the other hand, particles with $a_\mu \ll 1$ can be dominated by electromagnetic forces and gravity becomes a perturbation. Below we discuss examples of observations that can be best explained by recognizing electromagnetic effects coupling dust and plasma in the magnetospheres of Jupiter and Saturn (Horányi 1996) (see also JUPITER: MAGNETOSPHERE, SATURN: MAGNETOSPHERE).

Jupiter's ring

Jupiter's faint ring, similarly to Saturn's E-ring, provides an opportunity to study forces acting on single dust particles, as collisions and self-gravity do not hide single-particle dynamics.

The Jovian ring system has a complex spatial structure (see JUPITER: RINGS). The main ring is located at $1.71R_{\rm J} \le r \le 1.81R_{\rm J}$ with an optical depth $\tau \simeq 10^{-5}$. The main ring consists of 'large bodies' in the cm–km size range that are responsible for the charged particle absorption detected by Pioneer 11 and also a large population of small dust particles in the $0.1 \le a \le 100$ μm size range. At its inner edge the main ring suddenly turns into a vertically extended, toroidal halo with a maximum scale height of 3000 km. The halo comprises small dust grains with a similar size distribution to the dust population of the main ring. The size distribution of the dust in the main ring and in the halo appears to follow a power-law size distribution $n(a)\,{\rm d}a \sim a^{-\gamma}\,{\rm d}a$ with $\gamma = 2.5 \pm 0.5$. The third component of the ring system is a much fainter outward extension of the main ring that reaches near the orbit of the moon Amalthea at $2.54R_{\rm J}$.

The plasma environment of the ring is likely to be dominated by photoelectrons escaping from Jupiter's ionosphere. Close to Jupiter ($r < 3R_{\rm J}$) the expected equilibrium potential of the dust particles $\phi_{\rm eq} > 0$ and, due to the decreasing contribution of the ionospheric electrons at larger distances, $\partial\phi_{\rm eq}/\partial r > 0$. This radial dependence of $\phi_{\rm eq}$ can be responsible for much of the observed structure of the Jovian ring by causing very fast transport from the ring to the atmosphere.

An alternative model is based on resonances with the higher-order terms in Jupiter's magnetic field. This model requires slow transport of the ring material in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

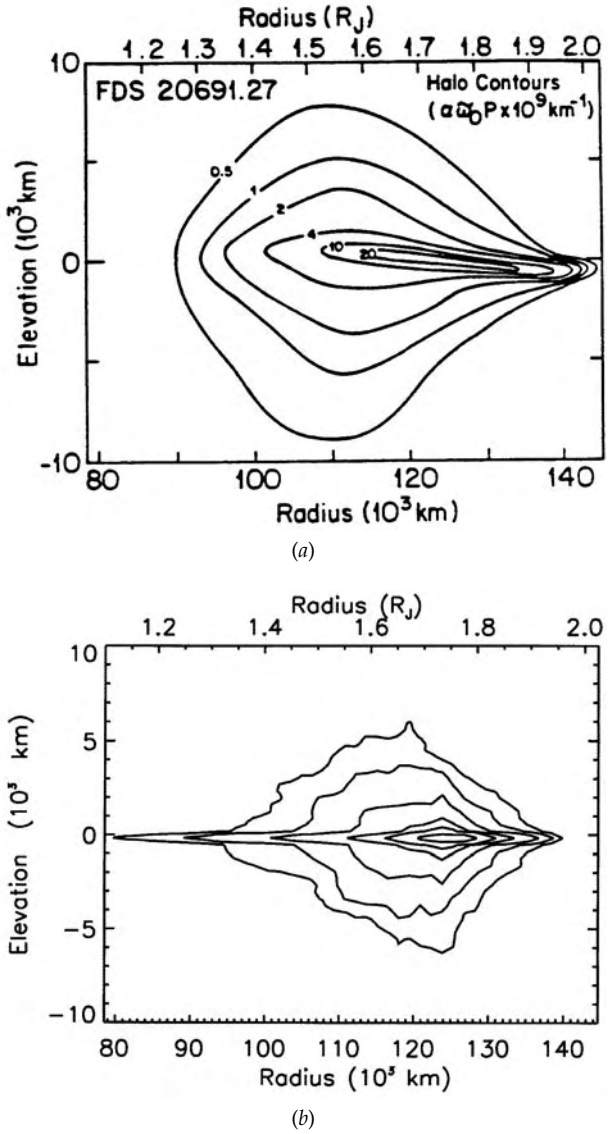


Figure 2. (*a*) The intensity contours, based on a Voyager image, of a halo slice reprojected into a rectangular coordinate frame. (*b*) The calculated contour plots of the normalized brightness distribution. The contour levels are (as in *a*) 0.025, 0.05, 0.1, 0.25 and 0.5. The model calculation reproduces the optical depth and the size and spatial distribution of dust in this region.

order to induce the observed vertical extent via trapping in resonances. The 3D spatial structure predicted by these models is quite different. However, the VOYAGER observations were not sufficient to restrict 3D models and did not allow for the detection of any time-dependent behavior.

In figure 2, the structure seen by Voyager was modeled by following a large number of particles generated via bombardment of the bigger boulders in the main ring and transported towards Jupiter's atmosphere. The 'tilted' nature of the Jovian magnetic field results in out-of-the-ring-plane forces and torques. Particles moving around the equatorial plane will 'see' a component of the magnetic field, $\boldsymbol{B}_{\parallel}$, that is parallel to the equatorial plane periodically changing its orientation. Particles on a prograde (counter-clockwise) trajectory experience a Lorentz force pushing them above (below) the ring plane as they cross the sectors where $\boldsymbol{B}_{\parallel}$ points toward (away from) Jupiter. This effect will force particles onto orbits with increasing inclinations explaining the large vertical extent of the halo.

Dust streams from Jupiter

Jupiter was first recognized as a source of dust particles during Ulysses' encounter with the planet in 1992 as very high-speed intermittent streams of very small grains were discovered (see SOLAR WIND: ULYSSES). The first estimates for the mass and velocity of these stream particles were uncertain since the detector was not calibrated in the suggested size and velocity range. Similar fluxes were seen with the identical dust detector on board the GALILEO MISSION TO JUPITER.

The suggested mechanism to eject dust particles from within the Jovian magnetosphere recognizes that grains can become positively charged to ≈+3 V near the moon Io. Once a grain charges positively it will be accelerated by the outward-pointing co-rotational electric field and leave Jupiter's magnetosphere. Outside the magnetosphere the azimuthal component of the solar wind magnetic field periodically accelerates the dust particles ejected from Jupiter toward and away from the ecliptic plane. This can explain the intermittent nature of the observed dust streams since both Ulysses and Galileo stayed in the ecliptic plane on their approach to Jupiter.

In fact, the detailed analysis of the propagation of the stream of dust particles in the solar wind led to the recognition that they must be much smaller in size and move much faster than first suggested. The best estimate for the size range of the dust grains seen by Ulysses is $0.005 \leq a_\mu \leq 0.015$ with velocity $v > 100$ km s^{-1}.

Positively charged particles escaping from the volcanoes of Io are accelerated outward and gain energy, W, from the co-rotating electric field

$$W = \int_{6R_J}^{50R_J} E\,Q\,\mathrm{d}r \simeq 2 \times 10^{-8} a_\mu \text{ J} \qquad (6)$$

where the lower limit of the integral is the location of Io and the upper limit is the approximate size of Jupiter's magnetosphere. Ignoring gravity, the work done by the electric field is the kinetic energy of the ejected dust particle; its exit velocity at the outer boundary of the magnetosphere can be estimated as

$$v_{\text{exit}} \simeq \frac{3}{a_\mu} \text{ km s}^{-1}. \qquad (7)$$

According to this model, particles in the size range of $0.005 < a_\mu < 0.015$ are expected to stream out of the Jovian magnetosphere with velocities in the range of $200 < v_{\text{exit}} < 600$ km s^{-1}, in agreement with the observations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn's E-ring

This faint ring extends from 3 to $8R_S$ with a sharp peak in its optical depth ($\tau \simeq 10^{-6}$) at the orbit of ENCELADUS at $r \simeq 4R_S$. The suggested particle size distribution to explain the wavelength dependence of the brightness in this region is surprisingly narrow, concentrated around $1 \pm 0.3\ \mu$m. The vertical extent of the ring has a minimum at Enceladus.

If the gravitational potential due to a central body changes as $1/r$ (Kepler problem) or r^2 orbits remain closed (i.e. the orientation of an eccentric orbit remains fixed in space). Any additional forces due to charging or higher-order terms in the planet's gravitational field, for example, result in orbital precession.

At Saturn the angular rate at which the pericenter (longitude of closest approach to the planet) of an orbit (with semimajor axis a_K) of a grain moves due to its electric charge is

$$\omega_\phi = -2\frac{QB_0}{mc} \simeq -5\frac{\phi_v}{a_\mu^2}\left(\frac{R_S}{a_K}\right)^3 \text{ deg day}^{-1} \qquad (8)$$

and the rate of change due to oblateness (equatorial bulge)

$$\omega_J = \frac{3}{2}\omega_K J_2 \left(\frac{R}{a_K}\right)^2 \simeq 50\left(\frac{R}{a_K}\right)^{3.5} \text{ deg day}^{-1} \qquad (9)$$

where ω_K is the Kepler angular rate and J_2 is the coefficient of the second-order term in the spherical expansion of Jupiter's gravitational field. For charged grains the net pericenter motion, $\omega_\phi + \omega_J$, can be positive, negative or even zero depending on the particle's size, charge and location in the magnetosphere.

Solar radiation pressure causes periodic changes in the angular momentum and hence orbital eccentricity of a dust particle in orbit around Saturn. The maximum eccentricity solar radiation pressure can induce is

$$e_{max} \approx \frac{3hf}{2\mu(\omega_\phi + \omega_J)} \qquad (10)$$

where h and f are the specific (i.e. per unit mass) angular momentum and acceleration due to radiation pressure.

In the plasma environment at the E-ring, the characteristic surface potential of the dust particles is expected to be in the range of $-8 < \phi_v < -4$ resulting in a near cancellation of the precession for $a_\mu \simeq 1$ grains. Due to continuous bombardment by micrometeorites and the E-ring particles themselves, Enceladus is the likely source of particles with a range of sizes. The magnetosphere selects particles with $a_\mu \simeq 1$ and forces them to follow highly eccentric orbits. Smaller or bigger particles will stay close to the orbit of their source. The vertical structure of this ring is caused by the out-of-the-ring-plane component of the radiation pressure. In figure 3 many of the observed features are captured using computer simulations.

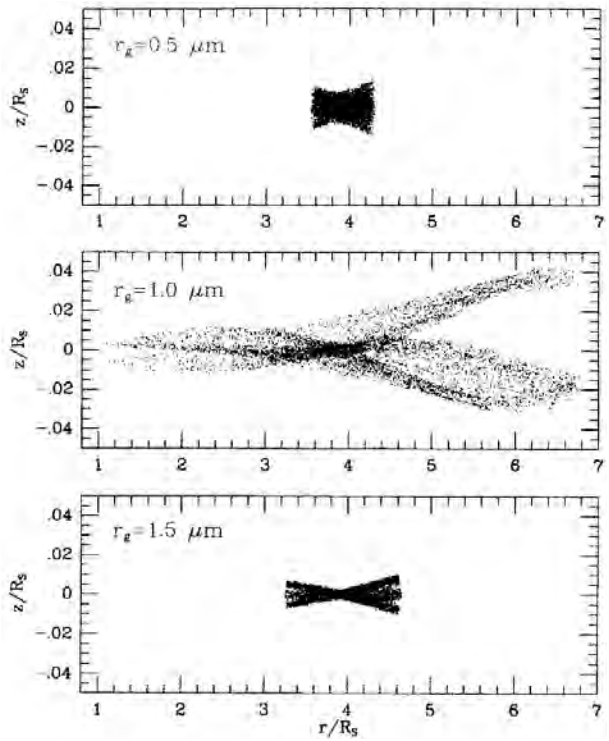


Figure 3. A scatter diagram of the simulated E-ring shown in cylindrical coordinates (r, Z), from randomly sampled orbits of 0.5 (top), 1(middle) and 1.5 μm (bottom) radius particles.

Future studies

Since December 1995 the Galileo spacecraft—now an artificial satellite of Jupiter—has been returning *in situ* and remote-sensing data from Jupiter's dusty magnetosphere. New Galileo images reveal much more detail about the ring/halo region and indicate time-dependent azimuthal asymmetries and large radial variations. Azimuthal brightness differences are likely to be related to short lifetimes, since Kepler shear would swiftly wash them away. Most likely, these features are related to the stochastic production of new material in the ring via meteoroid bombardment. The *Galileo* data—especially images taken while crossing the ring plane—are expected to show in great detail the spatial structure and size distribution of dust in the ring/halo region. New measurements in the Io plasma torus will help understanding of the charging history and subsequent energization and ejection of dust streams from the magnetosphere.

The CASSINI MISSION launched in October 1997 will reach Saturn in July 2004. Its cameras, plasma and dust detectors will monitor the activity of the 'spokes' and show in detail the spatial and size distribution of dust throughout its ring system. These and other missions to comets, the Moon or any other object in our solar system will find other examples, where observations will only be explained by recognizing dust–plasma interactions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Grün E, Morfill G E and Mendis D A 1984 Dust–magnetosphere interactions *Planetary Rings* ed R Greenberg and A Brahic (Tucson, AZ: University of Arizona Press) pp 275–332

Horányi M 1996 Charged dust dynamics in the solar system *Ann. Rev. Astron. Astrophys.* **34** 383–418

Johnson R E 1990 *Energetic Charged Particle Interactions with Atmospheres and Surfaces (Physics and Chemistry in Space 19)* (Berlin: Springer)

Mendis D A, Hill J R, Ip W H, Goertz C K, and Grün E 1984 Electromagnetic processes in the ring system of saturn *Saturn* ed T Gehrels and M S Matthews (Tucson, AZ: University of Arizona Press) pp 546–89

Whipple E C 1981 Potential of surfaces in space *Rep. Prog. Phys.* **44** 1197–250

Mihály Horányi

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnification

When a telescope is used visually, its magnification is the ratio of the angular diameter of the image as seen by the eye to the angle subtended by the object when viewed directly without the aid of the telescope. The magnification, or magnifying power, of a telescope (M) is given by the ratio of the focal length of the objective or mirror (F_o) to the focal length of the eyepiece (F_e) (i.e. $M = F_o/F_e$). For example, a telescope with a focal length of 2 m (2000 mm), used with an eyepiece of focal length 20 mm, would have a magnification of $2000/20 = 100$. The same telescope used with an eyepiece of 10 mm focal length would have a magnification of $2000/10 = 200$.

The magnification of a telescope is also equivalent to the ratio of the aperture of the collector (objective or primary mirror) to the diameter of the exit pupil (the image of the collector produced by the eyepiece). If all the light collected by the telescope is to enter the observer's eye, the exit pupil must be no larger than the diameter of the dark-adapted eye (about 0.007 m). To achieve this, with a telescope of aperture D (meters), the magnification should be at least $D/0.007 \approx 140D$ (i.e. about 140 for a telescope of 1 m aperture, or about 14 for a telescope of 0.1 m aperture). The maximum practicable magnification is limited by factors such as image contrast (the higher the magnification, the fainter the image) and by 'seeing' conditions (turbulence in the Earth's atmosphere causes sources to twinkle, shimmer and shake; this affects the quality of the image). For small and moderate-sized telescopes an approximate guide to the highest practicable magnification that is likely to give usable images is about 20 per centimeter of aperture (e.g. about 200 for a telescope of 10 cm aperture). In practice, it is seldom worthwhile to use magnifications greater than 200–300.

See also: aperture, Barlow lens, eyepiece, exit pupil, focal length, objective lens, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnitude Scales and Photometric Systems

More than any other aspects of astronomy, the subjects of magnitude scales and photometric systems are encumbered by history. The intensity of light from stars and other cosmic objects is usually expressed in magnitudes, an inverse logarithmic scale that confuses physicists who work in SI units, but that is practical for astronomers. The apparent magnitude of an object is a measure of the intensity of radiation within a particular wavelength interval received from that object at the Earth. The absolute magnitude is the magnitude that the object would have were it situated at a distance of 10 parsecs (pc, about 32.6 light-years (ly)) from the Sun. The relation between apparent magnitude m and absolute magnitude M is $m - M = 5 \log d - 5$, where the distance d is in parsecs.

The total energy, integrated over all wavelengths, received at the Earth from an object is also expressed as a magnitude, the bolometric magnitude. The difference between the bolometric magnitude $M_{\rm bol}$, and the magnitude $m_{\rm A}$ in bandpass A is called the bolometric correction $\rm BC_A$. BC without a qualifier normally refers to the correction to the visual magnitude; BC or $\rm BC_V$. The zero point of the bolometric magnitude scale is usually set by adopting $M_{\rm bol}(\rm Sun) = 4.75$, which implies $\rm BC_{Sun} = -0.07$.

Early astronomers compared star with star, a procedure that still retains great benefits. The surface temperatures of common stars range from 30 000 K down to 3000 K, and their apparent brightnesses cover a range of almost a factor of 10^{10}, from the sky background upwards (this range does not include the Sun). The majority of stars are constant in total light output and in temperature and as no laboratory lamps have energy distributions very similar to those observed in stars it is natural that astronomers seek to use the standard candles in the sky. Photometric systems represent attempts to define standard bandpasses and sets of standard sources, measured with these bandpasses, that are well distributed about the whole sky. Different photometric systems measure different wavelength bands. All photometric systems enable the measurement of relative fluxes, from which can be inferred particular properties (such as temperature) of the emitting object, but different systems claim to do it more precisely or more efficiently than other systems. Some of the systems were developed and modified by different astronomers over many years and the literature contains confusing versions and calibrations. Some people have despaired that it is too confusing and have suggested that we should start again with a well-defined ultimate system, but recent analysis has shown that modern versions of the existing photometric systems can be placed on a firm quantitative basis and that more care with passband matching will ensure that precise and astrophysically valid data can be derived from existing, although imperfect, systems.

Several recent large-scale astronomical projects are providing significant new magnitudes for large numbers of a wide range of astronomical objects. Firstly, the gravitational lensing projects, such as MACHO, EROS, OGLE and AGAPE, are identifying a range of variable stars and measuring light curves in the GALACTIC BULGE and the MAGELLANIC CLOUDS; secondly, wide-field survey projects, such as the SLOAN DIGITAL SKY SURVEY, are measuring the magnitudes of all objects above a certain brightness in the northern sky. Finally, the remarkable astrometric satellite HIPPARCOS measured extremely precise magnitudes for more than 118 000 stars, mostly brighter than 9th magnitude, over the whole sky and from outside the atmosphere. These projects are indicative of the quality and quantity of data that are becoming available and that will be a great challenge for standard magnitude calibrations.

The magnitude scale of Hipparchus; the intensity scale of Pogson

In the earliest recorded star catalog, HIPPARCHUS (2nd century BC) divided the stars in the sky into six groups. Twenty of the brightest stars that could be seen were called first-magnitude stars and those at the limit of visibility were called sixth-magnitude stars. Intermediate-brightness stars were put in intermediate magnitude classes. In the 18th century, astronomers were using telescopes and had begun to measure the light intensities of stars by closing down the telescope aperture until the image of the star under study just disappeared (the disappearance aperture). By taking the ratio of the squares of the disappearance apertures of two different stars, the relative intensity of the stars' light could be calculated. This was the beginning of astronomical visual photometry. Norman R Pogson (in 1856) at the Radcliffe Observatory compared his measurements of stellar brightness with stellar magnitudes given in contemporary star catalogs (such as those of Stephen Groombridge and the zone observations of Friedrich Argelander and Friedrich Wilhelm Bessel) and suggested the simple relationship $m = 5 \log a + 9.2$ to relate the magnitude m of a star and the disappearance aperture a (in inches). This relation implies a coefficient of 2.5 for the relation between magnitude and the logarithm of the intensity as $I \propto a^2$. Around that time, Gustav Theodor Fechner and Wilhelm Edward Weber (1859) were investigating the response of the eye to light and proposed the following psychophysical law: $m - m_0 = s \log I/I_0$, where m is a perceived brightness and the constant s defines the scale. Pogson's work implied a scale of -2.50 for astronomical visual (eye) photometry, and we thereby have the basis for the inverse logarithmic scale. There was continuing disagreement concerning the adoption of this exact scale of 2.5 and it was not until almost 30 yr later, after the Harvard photometry results were published in 1884, that adoption was ensured. The constant m_0, which defines the zero point, has undergone much refinement since Pogson's estimate and was officially set by the specified visual magnitudes of stars in the 'north polar sequence'. The early photometry catalogs are based on

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

this sequence but the magnitude scale today is established by the contemporary whole-sky photometric standard star catalogs.

New magnitudes from new detectors

Technological advances over the last 100 yr have provided a series of light detectors to supplement the eye. These detectors, in general, respond differently to light of different wavelengths from the eye; that is, they are more sensitive to blue light or to red light than is the eye. The advent of photography in the late 19th century revolutionized astronomy, as did the introduction of photomultiplier tubes with their light-sensitive photocathodes in the mid-20th century and sensors such as silicon charge-coupled devices (CCDs) and infrared detectors over the last 20 yr. Light intensities, or magnitudes, measured with these new detectors naturally differ from the visual magnitudes and depend on the color of the star. Initially, there was only the difference between visual magnitudes and 'blue' photographic magnitudes to be considered, but several factors resulted in a proliferation of different passbands and photometric systems: the extension of photographic and photocathode sensitivities to a wider wavelength range, the use of colored glass filters and interference filters to sample the starlight in narrower bands within the total wavelength sensitivity range of the detectors and, more recently, the requirements of survey instruments to provide maximum sensitivity for the detection of faint objects.

Rationale for multicolor photometry

Much photometry of astronomical objects is carried out in order to measure the apparent total brightness of objects and their relative brightnesses at different wavelengths, that is, their energy distributions. It is possible to characterize the temperatures of most objects from the overall shapes of their energy distributions. It is also possible to infer the metal content of stars from depressions in their energy distributions at particular wavelengths. These depressions (absorption lines) are due to the absorption of flux, principally by Fe and Ti (which have very rich line spectra), Ca, Mg and the molecules CN and CH (which have very strong lines in the blue–violet region of cool stars). There are many other molecular absorption bands (such as TiO, CO and H_2O) that depress the continuum in very cool stars; such molecular features are also used to provide information on the temperature, chemical composition and luminosity. The energy distributions of galaxies and star clusters can be analyzed to extract the relative numbers of different kinds of stars making up these composite objects. REDSHIFTS of very distant galaxies and quasi-stellar objects can also be measured from the positions of depressions or peaks in their energy distributions. These are called photometric redshifts and have been used very successfully with data from the Hubble Space Telescope and the Sloan Digital Sky Survey.

Multicolor photometry is best thought of as very-low-dispersion spectroscopy. The entire high-resolution spectrum of a star or other cosmic object contains a large amount of information, but, when dealing with extremely faint objects or with large numbers of objects, it is a great advantage to measure a small number of wavelength bands in as short a time as possible. Such a minimal technique is invaluable if it enables the derivation of many of the same parameters obtainable from a complete (and very redundant) description of the spectrum. A great deal of effort therefore has gone into accurately measuring and calibrating colors and depressions in terms of temperatures, metal abundances and other parameters, and investigating which of competing minimal descriptions of a star's spectrum is the most accurate or most practical.

Finally, GRAVITATIONAL LENSING results in changes in the brightness of the object independently of the color measured. Most intrinsically variable stars, however, have different amplitudes in red and blue light. Consequently, gravitational MICROLENSING surveys are efficiently carried out by splitting the light between a blue and a red channel for simultaneous direct comparison. Equal-amplitude variations in blue and red channels imply a lensing event, not a variable star.

Photometric systems: natural and standard

A light detector, a telescope, a set of filters and a method of correcting for atmospheric extinction make up a natural photometric system. Each observer therefore has their own natural system. The standard system is indirectly defined by a list of standard magnitudes and colors that have been measured for a set of typical stars, using the natural system of the originator. These are often called the primary standards. Later lists comprising more stars and fainter stars but based on the primary standards are called secondary standards. However, in the case of all photometric systems, recently published secondary standards effectively redefine the standard system because they tend to be more accurately measured than the primary lists and to represent contemporary detectors, filters and practice.

The term 'color' is an abbreviation for color index, which is the difference between the apparent magnitudes in two different spectral regions. Photometry results have generally been published as a series of colors and a single magnitude. The zero points of many color systems are set so that α Lyrae (Vega) has zero colors. In the southern hemisphere (where Vega is inaccessible) and often also in the north, the zero point is set by requiring that an ensemble of unreddened A0 stars have colors of zero magnitude. (See also STELLAR PHOTOMETRY.)

The original standard systems

The most influential of the early works of photoelectric photometry were the broadband Johnson UBVRI and Kron RI systems, which covered the wavelength region between 310 and 900 nm (3100 and 9000 Å). The natural systems of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Harold L Johnson and of Gerald E Kron and coworkers served as 'standard' systems for many other users who attempted with varying success (owing to differences in detectors, filters, telescopes and techniques) to duplicate the originators' natural systems. That is, using their own detectors and filters, astronomers measured stars from the Johnson and Kron lists and linearly transformed their natural magnitudes and colors to be the same as the Johnson and Kron colors and magnitudes. They then applied those same linear coefficients to transform the colors and magnitudes of unknown stars onto the Johnson or Kron system. The original blue and yellow filters were chosen by Johnson from readily available glasses so that when used with the 1P21 photomultiplier tube they approximated the ordinary blue (B) photographic response (~436 nm) and the visual (V) response (~545 nm). A more violet magnitude U (~367 nm), which is useful for very hot stars, was obtained by using a common violet glass. In retrospect, these choices should have been based more on astrophysics and less on glass availability, but so much work has been done in this UBV system that the weight of history ensured its continuation. Intercomparison of much of the published broadband photometry (in particular, photometry carried out more than 15 yr ago) often shows scatter of more than 0.03 magnitudes, but more recent photometry obtained using better equipment, better matched natural systems and better secondary standard stars agrees to better than 0.01 magnitude, or 1%.

The 1P21 phototube was a remarkable invention and its high blue sensitivity dominated the development of photometric systems for over 30 yr. There were red-sensitive devices available but observations were made only for bright stars because for many years these devices were much less sensitive, noisier and less reliable than the 1P21. In the mid-1970s new detector materials became available; in particular, the gallium arsenide and multialkali phototubes, which provided high ($\gg$15% quantum efficiency) sensitivity between 300 and 860 nm, and the infrared-sensitive InSb (indium antimonide) photodiodes together with low-noise preamplifiers, which revolutionized photometry between 1000 and 4000 nm. Both developments enabled photometry to be done on faint stars that had hitherto been the sole province of the blue-sensitive detectors. Photometry done with the new red-sensitive tubes was placed on either the Kron or the Johnson standard system, again with mixed success, and it has only been in the last 10 yr that A W J Cousins' RI 'near-natural' standard system (based on the Kron system) has gained widespread acceptance. It has also been very useful that the Cousins system's R (~638 nm) and I (~797 nm) bands are similar to the contemporary photographic R and I bands.

Johnson also introduced the infrared alphabetic JKLMN (approximately 1.22, 2.19, 3.45, 4.75 and 10.4 μm) system in the mid-1960s, using PbS (lead sulfide) detectors and bolometers. The water vapor in the Earth's atmosphere defines a series of wavelength bands (windows) through which observations from the ground can be made; Johnson used interference filters (and, unfortunately, the atmospheric H_2O absorption bands) to define what he called the J, K, L, M and N bands. Ian S Glass, in his early observations with an InSb detector, used the additional band H (~1.63 μm), between I and K, and in his choice of filters attempted to match the other Johnson bands. All infrared observers have proceeded in a similar fashion and have concentrated mainly on copying the Johnson K magnitude scale. Identical detectors have been used but a range of slightly different filters and observatory altitudes have produced subtly different systems. The publication of sufficient numbers of stars in common from the different natural systems has helped delineate the differences, and transformations between the systems are now quite reliable.

Passbands or response functions

The most important specifications of a photometric system are the passbands or response functions of its magnitudes. For a variety of reasons, technical and historical, the passbands of the original broadband photometric systems have not been known with certainty and this has inhibited close matching of natural systems and has prevented computation of accurate synthetic colors from theoretical spectra. The recent availability of spectrophotometry for many stars combined with the increased precision of second-generation photometric catalogs has, however, enabled the passbands to be derived indirectly by computing synthetic colors from spectrophotometry of stars with well-defined standard colors and adjusting the passbands until the computed and standard catalog colors agree. This technique has enabled the passbands of the major systems to be well defined, which in turn has permitted filters to be designed that still will result in good passband matches with a variety of detectors. In addition, when it is not possible to match passbands exactly with some detectors, such as photographic plates, it is possible to predict accurately the differences between photographic and photoelectric magnitudes by computing the synthetic magnitudes using the different passbands.

In figure 1 the normalized passbands of the Johnson-Cousins UBVRI system are shown plus an added Z band for CCD observations. The F_ν (flux per unit frequency interval) spectrum of an A0 star is shown for orientation. Table 1 lists the effective wavelengths $\lambda_{\rm eff}$, the approximate bandwidth $\Delta\lambda$, which is the full width at half-maximum of the passband, and the absolute calibration of the UBVRIJHKL system, based on the flux of Vega, for a zero-magnitude A0 star. Note that the effective wavelengths of the broad bands change with the color of the objects. The effective wavelengths listed are for an A0 star.

Other photometric systems

Real or perceived drawbacks in existing photometric systems (the UBV system in particular) stimulated the design of other photometric systems better suited for measuring temperatures, metal-line blanketing, effective

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Johnson–Cousins–Glass UBVRIJHKLM system.

	U	B	V	R	I	J	H	K	L	M
$\lambda_{\rm eff}$ (nm)	367	436	545	638	797	1220	1630	2190	3450	4750
$\Delta\lambda$ (nm)	66	94	85	160	149	213	307	39	472	460
F_ν ($V = 0$) (10^{-30} W cm^{-2} Hz^{-1})	1790	4063	3636	3064	2416	1589	1020	640	285	154

Table 2. Effective wavelengths (nm) and FWHM bandpasses (nm) for selected photoelectric systems.

	$\lambda_{\rm eff}$	$\Delta\lambda$		$\lambda_{\rm eff}$	$\Delta\lambda$		$\lambda_{\rm eff}$	$\Delta\lambda$
Geneva U	350	47	Walraven W	323.3	15.4	Washington C	391	110
B	424	76	U	361.6	22.8	M	509	105
B_1	402	38	L	383.5	21.9	T_1	633	80
B_2	448	41	B	427.7	49.0	T_2	805	150
V	551	67	V	540.6	70.3			
V_1	541	44						
G	578	47						
Strömgren u	349	30	DDO 35	349.0	38.3	Thuan–Gunn u	353	40
v	411	19	38	381.5	33.0	v	398	40
b	467	18	41	416.6	8.3	g	493	70
y	547	23	42	425.7	7.3	r	655	90
β_w	489	15	45	451.7	7.6			
β_n	486	3	48	488.6	18.6			

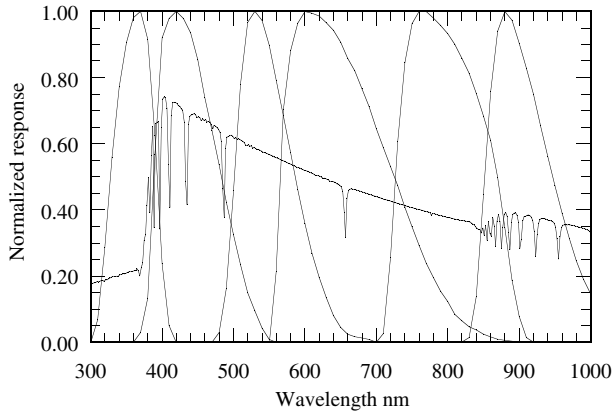


Figure 1. The passbands of the standard UBVRI system.

gravity and interstellar reddening. Some of these systems used broad bands comparable with the UBVRI system, while others used narrower bands defined by different mixes of glass filters or interference filters. Effective wavelengths and other specifications of some of the better-known systems are given in table 2 and are discussed below.

Geneva and Walraven systems

Difficulties with matching natural systems have been eliminated by the strategy employed by proponents of the Geneva ($UBB_1B_2VV_1G$) and Walraven (VBLUW) systems. The latter takes its name from Th and J H Walraven. These multiband photometric systems are supervised by small groups who control the instrumentation and supervise the data reduction and calibration. The colors have been well calibrated in terms of gravity, temperature and abundance. Such closed systems have excellent precision but not necessarily greater than that possible from the open Cousins UBVRI system with careful bandpass matching.

Washington system

This CMT_1T_2 system was devised to use the wideband sensitivity of the extended-red detectors, to improve the sensitivity of blue–violet colors to metallicity and gather more violet light in cool stars and to try to separate the effects of CN from other metal lines. We have found that the violet C band is a very useful metallicity indicator for faint K giants but that the M band contains little more information than does V; T_1 and T_2 have no advantages over R and I. The minimal CVI system is very useful for metal-weak K stars.

Strömgren four-color system

The uvby system was devised by Bengt Strömgren to measure better the Balmer discontinuity, the metallicity and the temperature of A, B and F stars. The bands are essentially separate unlike the UBV bands, which overlap. The u band is completely below the Balmer jump; v measures the flux near 400 nm, a region with much absorption due to metal hnes; b is centered near 460 nm and is affected much less than B by metal-line blanketing; y is essentially a narrower V band. The u filter is colored glass, the others are interference filters. Two special indices are derived: $m_1 = (v - b) - (b - y)$, which measures metallicity, and $c_1 = (u - v) - (v - y)$, which measures the Balmer discontinuity. The index $b - y$, like $B - V$, is used primarily as a temperature indicator. The system is capable of very high precision but, unfortunately, errors in the width of v filters manufactured some years

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ago resulted in nonstandard filters being supplied to many users. Since then, published photometry has exhibited some systematic differences in c_1 and m_1, and there are difficulties in computing these colors from theoretical spectra, particularly for cool stars. Recent standard catalogs of new and more homogeneous observations are of high precision and internal consistency and it should now be possible to define better the v band. Two additional interference filters (15 and 3 nm wide) centered on the Hβ line are often used together with the four colors. The Hβ index is used to derive luminosities in B stars and reddening in F and G stars. The Strömgren system was the first photometric system devised to measure specific stellar features. Because of the short-wavelength baseline of its four color filters, 1% photometry at least is required to utilize the system's advantages over the UBVRI system.

DDO (35, 38, 41, 42, 45, 48) system
This system (also built around the sensitivity of the 1P21 photomultiplier) was designed for the analysis of G and K dwarfs and giants. The 35 filter is the u filter of the four-color system; the 38 filter is also a glass filter and better measures metal blanketing than the v filter, being further to the violet and wider; 41 measures the CN band; 42, 45 and 48 are continuum filters. The color 35–38 (the 3538 index) measures the Balmer jump, 3842 measures the metallicity and 4245 and 4548 are used for gravity and temperature measurements. By restricting the measurements to the blue spectral region, complicated corrections for spectral line blanketing are necessary to derive temperatures and gravities. Good results, especially for faint K dwarfs, can be obtained by using $V - I$ or $R - I$ as the temperature indicator. Because of the narrow bandwidth of some of the filters, the DDO (David Dunlap Observatory) system has been mainly restricted to relatively bright stars.

Thuan–Gunn system
The uvgr system of Trinh Xuan Thuan and James E Gunn was devised in the mid-1970s from the UBVR system for use with an S20 photocathode detector and in order to avoid the strong mercury emission lines from city lights and [O I] lines in the night sky. The g and r bands are of similar width to the V and R bands whereas the u and v bands are about half the width of the U and B bands. The $g - r$ color has a longer baseline than $V - R$ but transforms well.

Photographic systems

Originally photographic emulsions were only sensitive to light blueward of 490 nm. These were the O emulsions. Different chemical sensitizing shifted the red sensitivity cutoff to longer wavelengths: G 580 nm, D 650 nm, F 700 nm and N 880 nm, approximately. By using blue-cutoff glass filters and the red cutoff of the emulsions, various photographic passbands were made. Photographic U used a violet filter for both blue and red cutoffs. Attempts were made to convert the photographic colors onto the photoelectric UBVR system but these were not often very accurate because of limitations in iris photometry and poor matches of the bandpasses. In recent years, astronomical photography has undergone a renaissance caused, first, by the development of new fine-grain emulsions (Kodak IIIaJ, IIIaF and more recently TechPan) and the utilization of methods of greatly increasing the sensitivities of the J, F and TechPan emulsions (using hydrogen gas) and of the N emulsions (using silver nitrate solution) and, second, by the use of new scanning microdensitometers and better methods of intensity calibration. Averages of several wide-field Schmidt camera plates or higher scale prime-focus plates can now produce photometry to a few per cent to very faint limits. Theoretical investigation of bandpasses enables better filter design for bandpass matching or predicts the relevant transformations and systematic differences between photoelectric and photographic photometry. Photographic photometry these days is usually restricted to attempted matches to the Johnson U and B or the Thuan–Gunn g systems using IIIaJ plates, to Cousins R or Thuan–Gunn r using IIIaF plates, and to Cousins I using IVN plates. Direct photographic calibration from step wedges is usually supplemented by direct magnitude measurements of stars in each field using a CCD array.

CCD photometric systems

The high quantum efficiency of CCDs and their inherent linearity have made them the detectors of choice in recent years for most areas of photometry. Unfortunately, the advantages of the CCDs were initially not fully attained because some users paid insufficient care to define their passbands and to standardizing their photometry. This resulted in internally precise results but an inability to relate these results with much confidence to the standard system data or to theoretically derived magnitudes and colors. Astronomers now realize the importance of matching their CCD passbands to standard passbands or deriving accurate passbands for their natural systems to enable them to be calibrated using synthetic photometry.

The standard BVRI system can easily be realized with thinned CCDs and colored glass filters but the U system is more problematical owing to the lower UV response of many CCDs. The Z band, between I and J, is also now often added to CCD UBVRI-based systems.

The HST WFPC2 photometric system

The Wide Field Planetary Camera (WFPC2) on the HUBBLE SPACE TELESCOPE has a suite of interference filters that cover both the space UV and the optical spectrum. Although not identical to the well-established UBVRI and uvby systems, there are passbands that are quite similar. A lot of attention has been given to calibrating the WFPC2 system both from actual ground-based observations where possible and from synthetic photometry so that excellent transformations are possible to the older standard systems and reliable temperature calibrations can be made from model atmosphere fluxes. Figure 2 shows some of the WFPC2 bandpasses. Table 3 lists effective wavelengths

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 3. Effective wavelengths (nm) and FWHM bandpasses (nm) for CCD-based systems.

	λ_{eff}	λ_{eff}	$\Delta\lambda$		λ_{eff}	λ_{eff}	$\Delta\lambda$		λ_{eff}	λ_{eff}	$\Delta\lambda$
HST 336	334	339	47	HIP B_T	421	439	70	SDSS u′	356	360	64
439	430	443	71	V_T	526	542	100	g′	475	500	135
450	451	477	107	H_P	517	595	230	r′	620	632	137
555	532	559	147					i′	761	772	154
675	667	675	127	UBV B	436	464	94	z′	907	907	147
814	788	805	147	V	545	558	85	EROS BE1	485	506	109
				R	641	666	160	RE1	657	679	191
MACHO B	519	543	144	I	791	799	143	BE2	539	575	190
R	682	700	178	Z	909	906	96	RE2	767	796	260

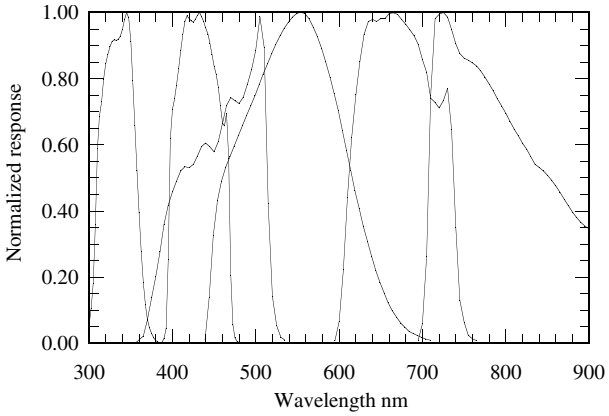


Figure 2. Some of the passbands of the HST photometric system.

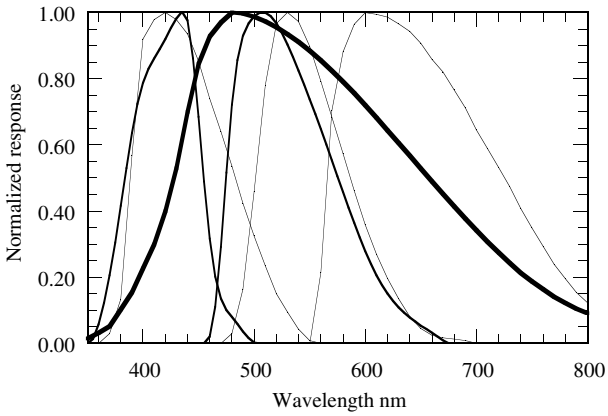


Figure 3. The Hipparcos passbands in comparison with the standard BVR passbands. H_P is shown by the thickest line, B_T and V_T are shown by the medium thick lines and BVR are shown by the thin lines.

(for an A0 and M0 star) and FWHM of some of the HST passbands and those of the other CCD systems discussed below.

The Hipparcos and Tycho photometric systems

The ESA Astrometric Satellite, Hipparcos, used three independent photometric detectors. The main Hipparcos passband (H_p) corresponded primarily to the spectral response of the S20 photocathode of the image dissector scanner combined with the transmission of the optics. The Tycho photometric data were derived from the star trackers and measured magnitudes in B_T and V_T, with passbands somewhat similar to standard B and V. Figure 3 shows the Hipparcos and Tycho passbands in relation to the standard BVR passbands. The large width of the H_p passband results in significant systematic differences between the H_p magnitudes and standard V magnitudes, depending on reddening, metallicity and luminosity. Nevertheless, the extremely high precision H_p magnitudes (∼0.0015) and the lower but still good precision (∼0.012) for the Tycho V_T magnitudes combined with the whole-sky coverage make these catalogs an invaluable resource, not only for measurements of individual stars, but for enabling intercomparisons to be made between and within ground-based photometric systems.

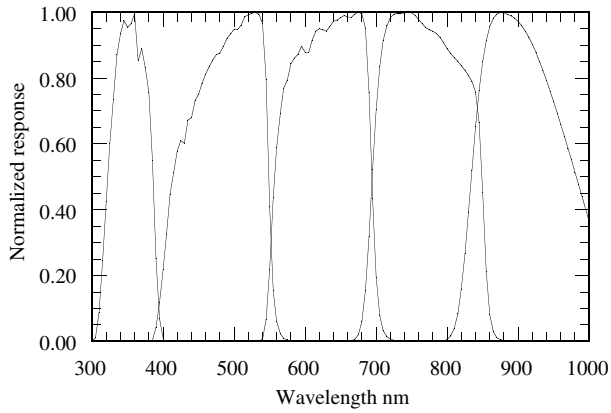


Figure 4. The Sloan Digital Sky Survey passbands.

The Sloan Digital Sky Survey photometric system

This photometric system comprises five color bands (u′, g′, r′, i′ and z′) that divide the entire CCD sensitivity range between the atmospheric UV cutoff near 300 nm and the CCD cutoff near 1100 nm. The passbands, related to those of the Thuan–Gunn system, and shown in figure 4, are essentially nonoverlapping and most are wider than those

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the UBVRI system, ensuring high efficiency for faint object detection. For ease of transformation into other systems or duplication of the system by others, it would have been better were the bands to have overlapped more with a less rectangular profile but the system itself will be very well defined by observations made with a duplicate detector and filters on a separate telescope. Unlike most other photometric systems, the zero points of the SDSS system have been placed on the spectrophotometric AB magnitude system defined by the absolute fluxes of four F subdwarfs. The passbands are essentially filter defined and in general have blue edges defined by a colored glass and red edges by a short-pass interference coating (see FILTERS).

Gravitational lensing projects

The MACHO photometric system

The MACHO project for monitoring gravitational microlensing events utilizes simultaneous CCD imaging in two passbands by sharing the light between two cameras using a dichroic beam splitter. The blue and red bands are further limited using an interference filter on the red side and the sensitivity cutoff of the thick CCDs on the blue side. The blue band approximates a broad blue-shifted V band while the red band approximates the R band. Good transformations are possible to the VR system and reliable calibrations are possible using synthetic photometry. As well as detecting many microlensing events, the MACHO project has provided unique and invaluable data on variable stars in the Magellanic Clouds and the Galactic Bulge.

The EROS photometric system

The EROS1 observations were taken consecutively through two broadband filters B_E and R_E that produced respectively passbands midway between B and V and R and I. Two different sets of filters were used during the course of the observations and the $B_E - R_E$ colors were transformed into V–I. EROS2, like MACHO, has two cameras and the light is divided using a dichroic beamsplitter. The division is made at a redder wavelength ($\sim$650 nm) than for the EROS1 system and the EROS2 blue band more resembles the Hipparcos passband.

Michael Bessell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maidanak Observatory

Located on Mount Maidanak in Uzbekistan. The most important instrument is a 1.5 m telescope designed for high-resolution imaging.

Until 1991, several republics of the former Soviet Union had their observatories at Maidanak. The Moscow Sternberg Astronomical Institute was the most powerful co-owner of Mount Maidanak.

After the fall of the USSR, the facility became the property of the Ulugh Beg Astronomical Institute in Tashkent. Economic difficulties have not allowed the observatory to be fully used and maintained. However, astronomers from the Sternberg Astronomical Institute (Moscow), the Astronomical Observatory of Kharkov University (Ukraine), and the Ulugh Astronomical Institute continue to use Maidanak Observatory and the 1.5 m telescope.

Grants from the Maidanak Foundation in Scandinavia are the main source of funds. These permit astronomers to cover their travel expenses, to make observations and to maintain the instruments.

For further information see http://www.maidanak.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Major Planets

A collective term for the largest planetary bodies in the solar system: Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto. Pluto, included in this category since its discovery in 1930, was found in the 1970s and 1980s to be smaller and less massive than seven of the solar system's planetary satellites, and just over twice the size of the largest asteroid, Ceres. In the 1980s and 1990s, calls for Pluto's demotion from major planet status led to a number of attempts to define 'major planet' more closely. Arbitrary size limits (diameters of 1000 km or 2000 km, excluding Ceres but including Pluto) or mass limits (similarly set so as to exclude Ceres but include Pluto) have been proposed. An alternative suggestion is that a body should be called a major planet if it is large enough for its own gravity to have overcome the structural strength of its material and made it roughly spherical (this definition would admit not only Pluto and Ceres, but also other large asteroids). Another criterion defines a major planet as a body whose gravitational influence extends over a particular range of distance from the Sun. By most versions of this definition, both Ceres and Pluto are denied major planet status.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maksutov Telescope

A modification of the Schmidt telescope devised in the early 1940s by D D Maksutov, in Moscow and, independently, by A A Bouwers, in Holland. The Maksutov uses a thin concave meniscus lens, located at the front of the telescope tube, to compensate for the spherical aberration (an optical defect) that would otherwise be introduced by its concave spherical primary mirror. Because the lens is thin, it has negligible chromatic aberration. The convex side of the meniscus lens, which itself has a spherical curve, faces towards the primary mirror. Because its corrector lens can be placed closer to the focus than that of a conventional Schmidt, a Maksutov is shorter and more compact than a Schmidt is. Like the Schmidt, the Maksutov has a low focal ratio and a wide field of view, which makes it particularly useful as a photographic survey instrument.

As with the Schmidt, the Maksutov has a curved focal plane, which lies inside the instrument. Consequently, in its original form, it can be used only for photography and cannot be used for visual observations. It can be adapted for visual observation by introducing a small secondary mirror that reflects light to a focus at the side of the tube (Matsutov–Newtonian) or, through a hole in the primary mirror, to a focus at the rear of the tube (Maksutov–Cassegrain). In the Maksutov–Cassegrain design, the secondary mirror can be figured as a small central reflective surface on the rear of the meniscus lens.

See also: Cassegrain telescope, field of view, focal ratio, focus, mirror, Newtonian telescope, Schmidt telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maksutov, Dmitri Dmitievich (1896–1964)

Soviet optician and telescope maker. After fighting in the Russian Revolution and the First World War, he worked on astronomical optics at Odessa, Moscow and Pulkovo, and invented the *Maksutov telescope*. The design is a development of the Schmidt telescope, replacing the aspheric lens with a negative meniscus lens with spherical surfaces which are easier to make. The telescope tube is also more compact, with consequential financial savings.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Malin, David (20th century)

Lancashire-born astrophotographer, used the Anglo-Australian Telescope and the UK Schmidt Telescope to create spectacular, beautiful and educative photographs of celestial objects, based on techniques developed to suit the fine-grain, sensitive, high-information content photographic material used for scientific pictures. His technique of unsharp masking filters out the diffuse content of the picture (to make it possible to see delicate detail within bright areas of nebulae and galaxies). His implementation of JAMES CLERK MAXWELL's technique of three-color addition combines three separate black and white pictures obtained through different colored filters (to give true color pictures revealing the broad spectral differences in the features and thus the astrophysical interest). Influential best-selling author and dynamic lecturer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Malmquist Bias and Completeness Limits

Many problems in observational astronomy concern a particular statistical distribution function of some quantity such as the number of objects with some attribute, x. Examples of x would include parallax, apparent magnitude, absolute magnitude, linear separation of binary stars, stellar mass, galaxy mass, x-ray luminosity, the occurrence frequency of novae and supernovae, distances to stars from measured parallaxes, Cepheid distances to galaxies, etc, all when the relevant data are sampled in some specified way. The precepts used in the sampling, such as limits on apparent magnitudes, or the way the parallax stars were chosen, or the period limits that were put on the observed Cepheid period–luminosity relations usually determine the form of the *observed* frequency distribution and how it differs from the true, unbiased, distribution. The problem of recovering the true underlying distribution from the various different observed distributions is a fundamental subject of statistical astronomy.

To set a basis for the discussion of observational selection bias, we mention two of the many examples that illustrate the problem:

(a) The Yale *Bright Star Catalog* in its latest edition (the fourth) lists much information on the 9110 stars brighter than $V = 6.5$ apparent magnitude over the sky. The distribution of absolute magnitudes of the entries is highly biased toward giant and supergiant stars; the limit of $V = 6.5$ apparent magnitude is so bright that no intrinsically faint late M dwarfs are listed. The reason is that for an intrinsically faint star with an absolute magnitude say of $M_V = +10$ to make the catalog, the apparent magnitude limit of $V = 6.5$ requires that its distance modulus be $m - M = -3.5$, or a distance of 2 parsecs. The number of stars with such a small distance is nil. To reach any significant number of M dwarfs would require the apparent magnitude limit of the catalog to be fainter than $V = 10$ if it were to encompass a more representative distance of greater than 10 parsecs for stars with absolute magnitude +10.

(b) Consider the opposite case of the apparent distribution of absolute magnitudes of stars in a catalog that is complete in a *fixed volume of space*, say within 25 parsecs of the Sun. One such catalog is by the staff of the Royal Observatory (Woolley *et al* 1970). The intrinsically faint stars that are missing in the *Bright Star Catalog* appear in large numbers in the *Nearby Star Catalog*, while the intrinsically bright stars are now missing. Their space density is simply so low that the volume of space within 25 parsecs is too small to include even one. The distant-limited *Nearby Star Catalog* is biased in the opposite way from the *Bright Star Catalog*.

It is obvious that to obtain an unbiased luminosity function (i.e. the true distribution of the number of stars at absolute magnitude M in an interval of $M + dM$ and $M - dM$ per unit volume) one must combine in some way both catalogs, each biased at opposite ends of the true luminosity function. This can be done by sampling the flux-limited listings in the *Bright Star Catalog* to fainter apparent magnitudes and by increasing the spatial volume sampled in the distance-limited *Nearby Star Catalog*. But even then selection effects remain, caused by the fact that there will still be limits of completeness, although widened, that must be corrected for.

These examples of the two different biases due to observational selection contain many of the elements of general bias problems in statistical astronomy. The most famous of such problems is that of the Malmquist bias. It appears as the difference in the mean absolute magnitude, $\langle M \rangle_{\mathrm{FL}}$, of a set of objects that are chosen in a flux-limited way (i.e. complete to a given apparent magnitude) compared with $\langle M \rangle_{\mathrm{DL}}$ for the same class of object chosen to be complete within a given volume of space.

The purpose of this article is to explain why the Malmquist bias exists, to describe how to identify it in any given sample of objects (stars or galaxies), to show a method of how to correct for it in statistical investigations, and to show how to identify the distance-limited subset of the complete flux-limited data that defines the true, bias free, distribution function.

An early example of how an observational limit on radio flux produced a highly biased distribution of intrinsic luminosities of radio sources

As soon as the early positions of RADIO SOURCES became accurate enough in the early 1960s to mount optical identification programs, it became evident that most of the radio sources were intrinsically bright early type (E) galaxies. Distances to these galaxies could be obtained from their optically measured REDSHIFTS using the linear Hubble expansion law with an assumed value of the HUBBLE CONSTANT. Absolute radio power could be calculated using the distances and the observed radio fluxes.

The result was the curious strong correlation of absolute radio power, L_{R}, with redshift shown in figure 1. Plotted as ordinate is the intrinsic radio power (erg s^{-1} over the frequency range of 10^7 to 10^{10} Hz). The abscissa is the distance modulus (based on a Hubble constant of 50 km s^{-1} Mpc^{-1}) calculated from the redshift of each source.

The apparent increase of L_{R} with increasing redshift has a close z^2 dependence. This drastic increase of absolute radio luminosity (by six orders of magnitude) with distance (squared) could have been interpreted as a real increase in radio power with look-back time (i.e. an evolutionary effect) until it is realized that the apparent correlation is in fact due to the flux limitation of the radio catalogs. Note the lines of constant radio flux of 1, 3 and 10 janskys.

The explanation of this being due to the flux limitation was soon proved by adding a fainter sample of the optical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

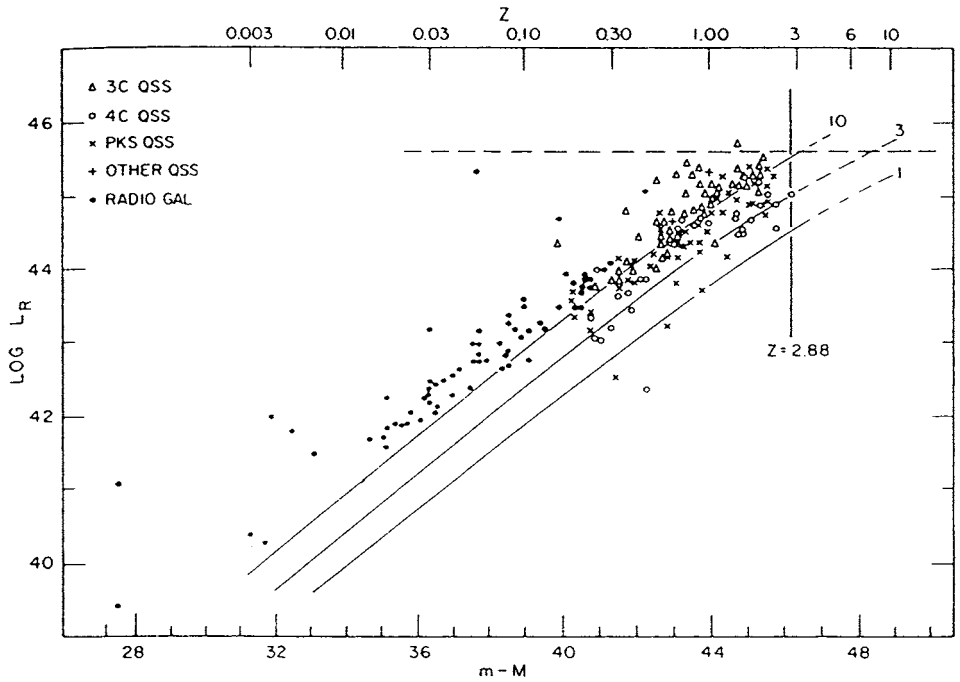


Figure 1. Relation between absolute radio power, L_R, over a band width from 10^7 to 10^{10} Hz and redshift (distance) for flux-limited data from various radio catalogs. The apparent correlation is due to the selection bias caused by the bright flux limitation in the relevant radio catalogs and the existence of a very broad intrinsic radio power luminosity function. (Diagram from *Astrophys. J.* **178** 25 1972.)

identifications to flux levels that soon reached 0.01 janskys, fully 100 times fainter in radio flux than the data in figure 1. The resulting diagram then filled the triangular region at the lower right half of figure 1. Clearly, the radio luminosity function is extremely broad, covering power limits between 10^{40} and 10^{46} ergs s^{-1}, and is also so steep that the spatial volume sampled at small redshifts is too small to include even one of the most powerful sources at such small distance moduli. Note the similarity of this selection effect to that of the *Bright Star Catalog*.

Figure 1 has the form of what is called a Spaenhauer diagram in the current bias literature. Its use is powerful, first as a diagnostic to identify the presence of bias, and then as an aid to devise methods of correction. Bias in any data sample can always be proved using such a diagram by adding a fainter sample to see if the bias disappears at the original bright limit but reappears with the same pattern near the limit of the fainter sample.

The Spaenhauer diagram approach is not the only method to discover and to correct for observational selection bias. A powerful parallel method, useful for example in the problem of bias in the TULLY–FISHER RELATION for distance determination, is that of 'normalized distances' (Teerikorpi 1997). That reference describes the development by a Franco-Finnish group on the analysis of data that are germane for the extragalactic distance scale.

The Malmquist bias

What is the bias?

Consider the properties of Spaenhauer diagrams that are used here to explain the Malmquist bias in flux-limited data sets. For the demonstration we use data for galaxies from the flux-limited revised *Shapley–Ames Catalog of Bright Galaxies* (the RSA), although the same general bias properties apply to many other types of flux-limited distributions in statistical astronomy.

It is known that the general luminosity function over all galaxy types is very broad, spanning the range of absolute magnitudes from $M_V = -23$ to -14 (see LUMINOSITY FUNCTION OF GALAXIES). A Spaenhauer diagram using such a large intrinsic dispersion would resemble the 'open' diagram of figure 1. A more 'closed' diagram is obtained by treating a subsample of the complete galaxy data by restricting the discussion to specific Hubble types and van den Bergh luminosity classes where the individual type and class-specific luminosity functions have much smaller intrinsic dispersions.

Figure 2 shows a typical Spaenhauer diagram of the 'closed' type where the intrinsic dispersions are fairly small. Absolute magnitude, calculated from the known redshifts (assuming $H_0 = 50$ km s^{-1} Mpc^{-1}) is the ordinate. Log redshift is the abscissa. The data are from the complete set of Sc galaxies of van den Bergh luminosity class I (see GALAXIES: CLASSIFICATION) as listed in the RSA. This catalog contains the all-sky sample of galaxies brighter than apparent magnitude $B = 13$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

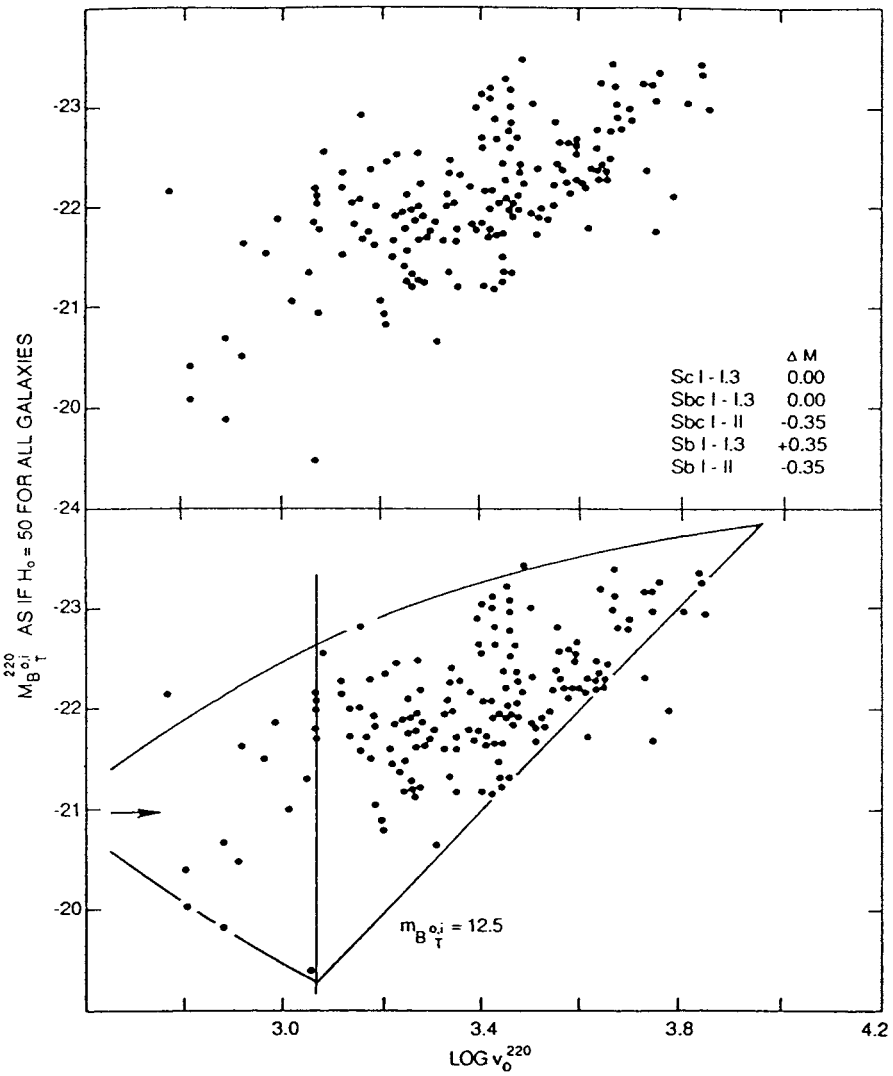


Figure 2. Top: The apparent correlation of redshift and the absolute magnitudes ($H_0 = 50$) of ScI galaxies in the flux-limited RSA catalog. Bottom: Same as top but with the limit lines of the Spaenhauer configuration superposed. The apparent correlation in the top panel is due to selection bias. The faint end of the luminosity function to the right of the vertical line is progressively cut away at increasingly large redshifts.

Because it is a flux-limited catalog, it has the same bias properties as the *Bright Star Catalog* described earlier. Intrinsically faint galaxies are underrepresented relative to an unbiased (true) distribution (per unit volume) in its absolute magnitude distribution.

The Malmquist bias is defined as *the difference in absolute magnitude between the flux-limited and distance-limited distributions*. How is it detected? How is it calculated and corrected for?

The top panel of figure 2 shows that the mean absolute magnitude of the sample becomes progressively brighter with increasing distance in the same way as for the radio sources in figure 1. As in figure 1, there is a sharp lower bound to the distribution caused in both diagrams by the flux-limit of each catalog; 1 jansky in figure 1 and near the effective apparent magnitude limit of $B = 12.5$ for the RSA (the nominal limit of $B = 13$ is made brighter here by the correction for internal absorption that has been applied to the data).

The straight-line flux limit at $B = 12.5$ is drawn in the bottom panel of figure 2. The upper and lower curved envelope lines are the theoretical expectation loci for one galaxy to be detected in the sample. These envelope lines are from a calculation using an assumed Gaussian luminosity function with a fixed dispersion of 0.7 mag, and individual volume elements enclosed within redshift (distance) limits of $v + \mathrm{d}v/2$ and $v - \mathrm{d}v/2$ for a uniform density distribution of objects. (The shape of the envelope lines is independent of the value assumed for the Hubble constant.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

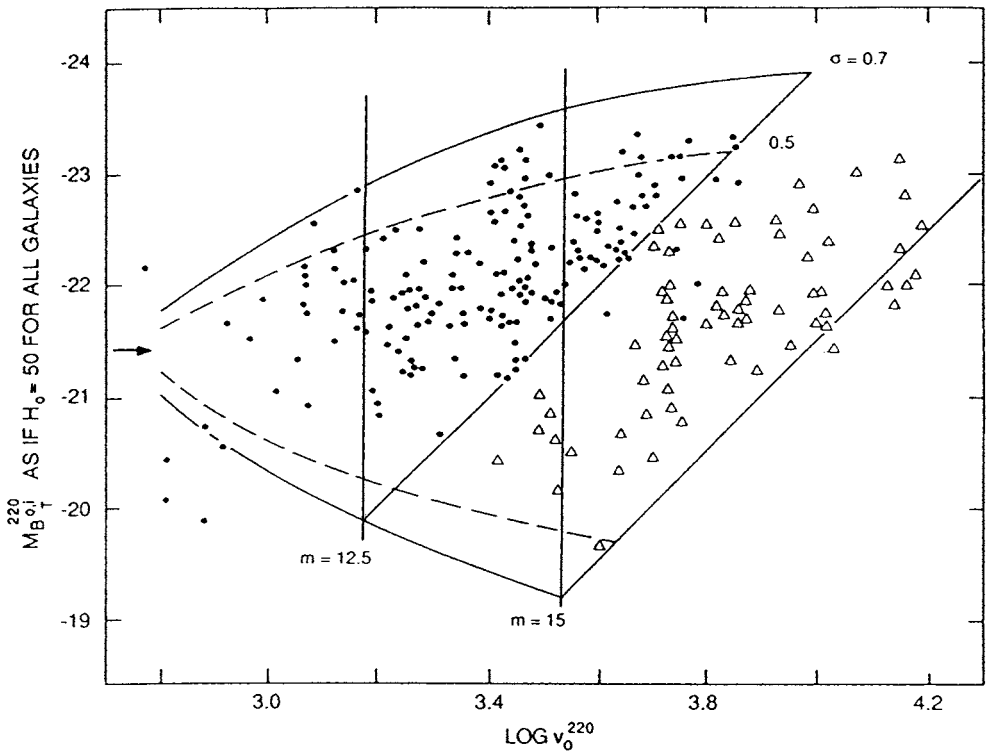


Figure 3. Same as figure 2 but with a fainter sample of ScI galaxies added to a flux limit of $B = 15$. This figure illustrates how adding a fainter sample decides between a real variation of $\langle M \rangle$ with distance versus the effect of bias in any given flux-limited data set.

A family of Spaenhauer envelopes can be computed for different dispersions. Also, different forms of luminosity functions (i.e. non-Gaussian) can be assumed to make any number of different sets of templates. The set of Gaussian templates used here is described by Sandage (1994).

The vertical line in figure 2 is put at the junction of the lower curved envelope line and the flux limit line at $B = 12.5$. The part of the Spaenhauer configuration that is to the left (smaller distances) of the vertical line is the distance-limited subset of the complete data. There is no bias in this region because the sampling of the luminosity function is complete. However, all data to the right of the vertical line are progressively biased at progressively larger distances because the luminosity function is progressively cut from the sample by the flux-limit line. The mean absolute magnitudes calculated in vertical strips of velocity width dv to the right of the vertical line become brighter with increasing distance. This is the detailed bias property of any flux-limited sample.

The classical Malmquist bias is defined as the *integral* of this detailed effect, integrated over the complete data set. Clearly, the mean absolute magnitude of the distance-limited subset (all galaxies to the left of the vertical line in figure 2) is the magnitude of the apex point shown by the arrow. This magnitude is fainter than the average magnitude of all the points in the complete data set, i.e. taken over the total sample in the top panel of figure 2. As stated above, the difference between the distance-limited mean absolute magnitude and the average over the complete data set is defined as the Malmquist bias as he originally calculated it in 1920.

The calculation of this 'integrated' Malmquist bias is a straightforward problem in statistical astronomy. Modern derivations can be found in Mihalas and Binney (1968) and Sandage (1995). The formal equation is

$$\langle M \rangle_{\mathrm{DL}} - \langle M \rangle_{\mathrm{FL}} = (\mathrm{rms})^2 A(m)^{-1}\, \mathrm{d}A(m)/\,\mathrm{d}m$$

where DL and FL refer to distance-limited and flux-limited mean absolute magnitude, $A(m)$ is the number of objects at m in interval dm, and d$A(m)$/dm is the slope of the growth of $A(m)$ with m at m. For a uniform space distribution, $\log A(m)$ grows as $0.6m$ and the Malmquist correction is clearly

$$\langle M \rangle_{\mathrm{DL}} - \langle M \rangle_{\mathrm{FL}} = 1.382\ (\mathrm{rms})^2.$$

How to detect the bias

The infallible signature of bias in any sample is an apparent brightening of the individual absolute magnitudes with increasing distance in a flux-limited sample. If the effect is due to bias, the bias can be proved by plotting a Spaenhauer diagram (i.e. M versus distance). To test that the correlation is due to bias rather than say due to 'evolution', we need only to add a fainter sample. If the effect is due to bias, the effect will disappear at bright

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

magnitude and will reappear at the limiting magnitude of the fainter sample.

This proof for bias in the ScI flux-limited distributions in figure 2 is made in figure 3 where a fainter sample to the magnitude limit $B = 15$ is added for ScI galaxies. *The extended Spaenhauer configuration also fits the added data*. The distance-limited unbiased portion of the data is now increased to a redshift of $\log v = 3.55$, but the distribution begins to become biased again (the faint part of the luminosity function drops out for larger redshifts), but now at larger redshifts in the added fainter sample.

Detailed correction for the bias must either be made at every redshift interval by accounting for the brightening at every redshift, or by restricting the sample to the distance-limited subset of the data to the left of the vertical lines in figures 2 and 3.

Malmquist bias in a more subtle form is also present in flux-limited samples that use the Tully–Fisher relation for distance determination. A large literature exists, an introduction to which can be had in the review article by Teerikorpi (1997), and in the bias series of papers of which Sandage (1994) is the first.

Other types of bias similar to Malmquist

Other important types of bias problems in statistical astronomy are central to many calibration problems. They can only be mentioned here, with literature citations given in the bibliography.

Bias in trigonometric parallax data sets

Selection procedures for sets of trigonometric parallax data determine the bias characteristics of the resulting parallax database. The problem can be seen in its progressive development by reading the following references in the order Eddington (1940), Wallerstein (1967), West (1969), Lutz and Kelker (1973), Hanson (1979). Other references to the problem are listed therein. It is an extremely important problem in using trigonometric PARALLAXES to calibrate the absolute magnitudes of many different types of objects such as subdwarfs, for example, with the resulting distances to the system of globular clusters.

Bias in Cepheid distances using the P–L relation

Because the CEPHEID PERIOD–LUMINOSITY RELATION has an intrinsic dispersion in absolute magnitude at a given period due to the finite width of the instability strip in the HR diagram, bias is present at the faint end because all discovery programs are flux-limited. The method to detect and correct for the bias is given in Sandage (1988).

Bibliography

Eddington A S 1940 The correction of statistics for accidental error *Mon. Not. R. Astron. Soc.* **100** 354

Hanson R B 1979 A practical method to improve luminosity calibrations from trigonometric parallaxes *Mon. Not. R. Astron. Soc.* **186** 875

Lutz T E and Kelker D H 1973 On the use of small trigonometric parallaxes for the calibration of luminosity systems *Publ. Astron. Soc. Pacific* **85** 573

Mihalas D and Binney J 1968 *Galactic Astronomy* (San Francisco: Freeman) ch 4

Sandage A 1988 Cepheids as distance indicators when used near their detection limit *Publ. Astron. Soc. Pacific* **100** 938

Sandage A 1994 Bias properties of extragalactic distance indicators I: the Hubble constant does not increase outward *Astrophys. J.* **430** 1 (the first of a series of papers on bias problems)

Sandage A 1995 Practical cosmology: inventing the past *The Deep Universe (Saas-Fee 23rd Advanced Course)* ed B Binggeli and R Buser (Berlin: Springer) chs 4, 10

Teerikorpi P 1997 Observational selection bias affecting the determination of the extragalactic distance scale *Ann. Rev. Astron. Astrophys.* **35** 101

Wallerstein G 1967 On the use of small parallaxes for calibration purposes *Publ. Astron. Soc. Pacific* **79** 317

West R M 1969 On the calibration of M_V (K) for giants by means of trigonometric parallaxes *Astron. Astrophys.* **3** 1

Woolley R *et al* 1970 Catalogue of stars within 25 parsecs of the Sun *Royal Observatory Annals* no 5

Allan Sandage

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Malmquist, Gunnar (1893–1982)

Swedish astronomer, worked on statistical astronomy, for example the populations of stars in the Galaxy. The *Malmquist bias* is a statistical effect by which the fainter members of a population are more represented in a brightness-limited sample than they should be—there are more of them to cross by accidental measuring error above the cut-off line than there are above the limit to fall below.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mantle

The Earth's mantle is composed of a thick layer of solid rock extending from the molten alloy of iron which is the Earth's CORE, 2891 km below the surface, to the so-called Mohorovicic discontinuity or Moho, a few kilometers below the surface. Roughly, it can be considered as a succession of spherical concentric shells which are, from the shallower to the deeper: the lithospheric mantle (from the Moho to 80 km), the upper mantle (from a low velocity zone below the lithosphere to 400 km), the transition zone (from 400 to 660 km), the lower mantle (670 to 2800 km) and the D″ (thermal and/or compositional layer above the core). The increases in temperature and pressure with depth are responsible for this layering which mainly follows the mineralogical phase changes of olivine, the main mineral component of mantle rocks. However, at the present state-of-the-art we cannot definitively preclude the possibility of compositional changes with depth. Actually, the Earth's mantle is much more complex than a simple succession of concentric layers. For short time scales, it behaves like a perfect solid, transmitting seismic elastic waves with almost no attenuation while for geological time scales it behaves like a very viscous fluid. The original accretion energy and the still active long-life radioactive elements act as a powerful energy supply which animates the mantle engine: thermal convection. The convective motions within the mantle are ultimately responsible for the surface plate tectonics, the generation of oceanic lithosphere at mid-oceanic ridges and for the return of lithospheric plates to the mantle at subduction zones. These water-rich sinking plates which have cooled at the Earth's surface for hundreds of millions of years interact with the hot mantle to generate subduction zone volcanism. Their final fate, either to be reincorporated into the upper mantle or to sink deep into the lower mantle and feed the D″ zone, is still vigorously debated within the geophysical community (see also EARTH'S INTERIOR).

A short history

The Earth's mantle is inacessible to direct observations. Our technologically sophisticated civilization remains defeated by the huge pressures and high temperatures which are prevalent even at depths as shallow as a few kilometers. Thus, the deepest drilling experiments have rapidly encountered technical problems due to the increase in temperature with depth (roughly 30 °C per kilometer within the continental crust). Until now, no drilling experiment has succeeded in penetrating all the way through the Earth's continental or oceanic CRUST. The Moho, which separates the crust from the mantle, is roughly 30 km below the surface of the continental crust but only at 6 km below the surface of the oceanic crust. This seismic discontinuity, which is due to the sudden increase in the density of the rocks, characteristic of the mantle, has taken its name from A Mohorovicic, who was the first to detect it in 1907 in the Balkans.

The only direct access that scientists can have to mantle rocks is the very rare ophiolitic or kimberlitic outcrops which have been brought to the surface by tectonic or volcanic events. The ophiolites form groups of mafic and ultra-mafic rocks ranging from basalt to peridotite which have been obducted onto continental crust. This phenomenon occurs when a piece of oceanic lithosphere is carried along during the closure of an OCEAN and is piled onto continental crust. When they are denuded by erosion, outcrops may expose, as is the case in Oman, the whole sequence of lithospheric plate rocks from the pillow lava of the oceanic crust to the upper mantle material below the Moho. The kimberlites are very different in nature. They are composed of peridotite rocks set intrusively by violent volcanic eruptions from deep upper mantle into the continental crust.

Too deep to observe directly, the interior of the Earth has been explored recently by the use of simulations which have allowed scientists to construct a model step by step.

The thickness of the Earth's mantle (2.88×10^6 m), its mass (4.043×10^{24} kg) and the ranges of the characteristic times of physical phenomena which affect it (from fractions of a second for seismic wave propagation to billions years for continental drift), make its behavior difficult for us to understand without the help of high-accuracy metrology. It has only been during the 20th century that improvements in mathematics, physics, chemical sciences and, more recently, the advent of the computer, have allowed significant progress to be made.

Eratosthenes was the first to measure the radius of the Earth around 250 BC. Newton proposed in 1687 that the Earth is an ellipsoid with flattening of 1/230. In 1736 the French Academy of Sciences organized two scientific expeditions in Laponia (directed by A Clairaut and P L de Maupertuis) and in Peru (Directed by P Bouguer and Charles de la Condamine), which confirmed the flattening of the Earth. In 1798, Lord Cavendish proposed an experiment to weigh the Earth and found a mean density of 5.5×10^3 kg m^3. This value, which is roughly twice the density of surface rocks, implied that there was an increase in density with depth. In 1838, William Hopkins, Profesor at Cambridge University, took account of the two modes of heat transfer—conduction and convection—and proposed three possibilities for the internal structure of the Earth: a thin crust overlying a molten Earth, a thick crust and a solid core separated by molten rocks and an entirely solid Earth. The thin crust model was refuted by part of the scientific community, which reasonably argued that such a liquid layer would have been subject to huge tidal effects by the Moon and the Sun, which would have broken the crust. On the other hand, astronomical considerations about the moment of inertia of the Earth also suggested an increase in the density of material near the center of the Earth. This proposal was in agreement with meteorite composition that required a silicate mantle overlying a denser iron core. However, at the end of the 19th century, the inaccessibility of the Earth resulted in some highly speculative inferences about the internal structures of the Earth and the mantle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

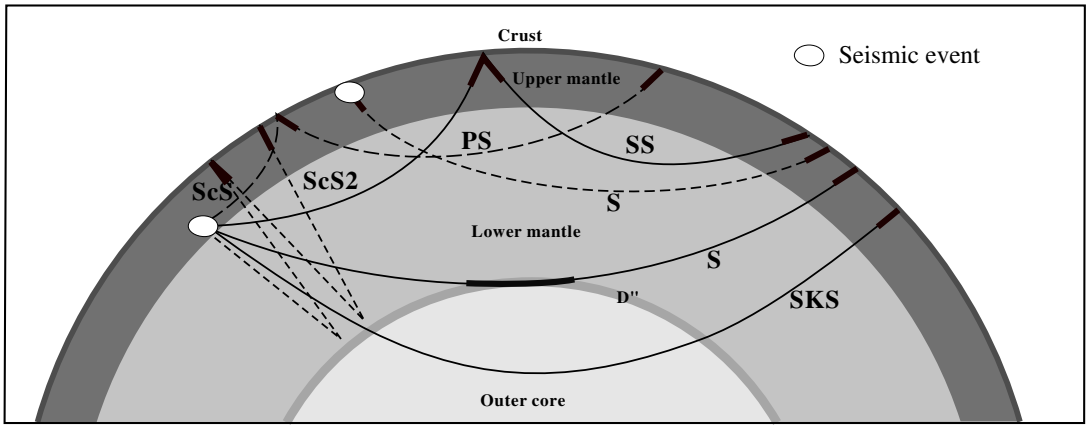


Figure 1. The energy released by seismic events propagates within the Earth's interior and therefore through the mantle as P-waves and S-waves. When encountering obstacles or discontinuities these waves may bounce, be converted from one type to the other, be transmitted through the interfaces or be split. A 'PS'-wave starts from the focus of a seismic event as a P-wave, it propagates through the mantle, bounces under the Earth's surface and continues its journey through mantle as an S-wave. The c in ScS, denotes the reflection of an S-wave on the core surface. K is used for transmission within the core, as for SKS. In this case the S-wave is converted to a P-wave at the core–mantle boundary (CMB), transmitted as a P-wave within the core and is converted back to an S-wave on reaching the mantle. The study of these different wave propagations, the arrival times of which are known to be a function of the surface angular distance of the recording station to the epicenter, allows information to be deduced about its particular path within the mantle.

The elastic mantle

The real birth of modern geophysics occurred with the fortunate detection of a Japanese earthquake in Potsdam on 18 April 1889. A new science was born that would allow scientists to explore the deepest parts of the Earth and to compare the results of observations with models.

The energy released by an earthquake propagates through the deep Earth as elastic waves for which propagation properties depend intimately on the density, compressibility and rigidity of the medium. Encountering discontinuities or obstacles, the compressive waves (P-waves) or shear waves (S-waves) can bounce, be converted from one type to the other, be transmitted or be split. During the first few decades of seismological study, scientists learned to recognize the different types of waves from the times they took to travel from the earthquake focus to recording stations. They defined a nomenclature for these waves according to their propagation within the Earth. Figure 1 shows the different paths of the simplest P- and S-waves within the mantle and their names.

The analysis of arrival times of waves at at least four independent recording stations allows retrieval (if a velocity law is assumed along the path) of the origin time and the geographical location of the epicenter (the point located at the surface above the focus of the earthquake). Conversely, the knowledge of the initial parameters of earthquakes has allowed the velocity models inside the mantle to be refined. During the first two decades of the 20th century, the core, the inner-core, and the main discontinuities of the mantle were discovered. Radial models of the interior properties of the Earth have been regularly improved, leading to the so-called preliminary reference earth model (PREM) which was published by Dziewonski and Anderson in 1981 (figure 2). This model gives the evolution of the gross Earth properties with depth (density, pressure, compressibility, P- and S-wave velocities). The propagation of S-waves in the mantle proves that it is composed of solid rock (S-waves disappear in the liquid outer core but reappear in the inner core). From the evolution of density with depth it is easy to conclude that the mantle is composed of compressible rock. However, density discontinuities (and therefore wave velocity discontinuities) do exist in the PREM, mainly at 670 and at 400 km below the crust and at the Moho level (or from more recent estimations, 660 and 410 km). The two deepest discontinuities bound the transition zone that separates the upper mantle from the lower mantle (below 660 km). A question which is still debated amongst geophysicists is: do these sharp density jumps correspond to compositional discontinuities or to phase changes in the mineralogical structure of the silicate rocks composing the Earth's mantle?

Mantle composition

The Earth is a part of the solar system in which the planets and the Sun have been built by the collapse of a nebula and the ACCRETION of planetesimals. The METEORITES, which still regularly fall on the surface of the Earth, are a surviving witness to those early times. Meteorites can be classified into three main categories: irons or siderites, which are made principally of native metal; stones or aerolites which contain mainly silicate rocks but little metal; and stony irons or siderolites which contain abundant silicates and metals. The stone meteorites (CHONDRITES) represent 86% of the meteorite population and are thought to have a chemically primitive composition. The chondrites are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

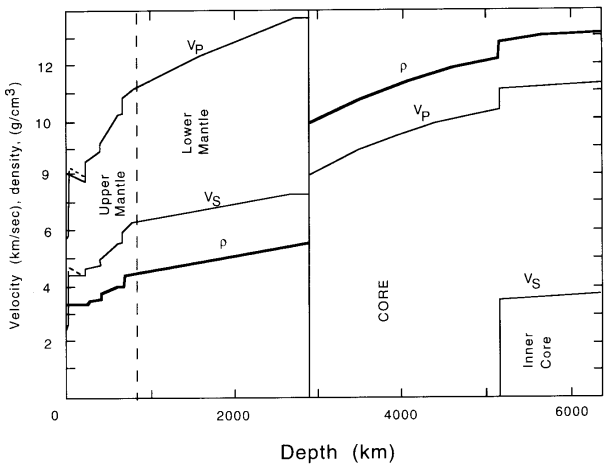


Figure 2. The PREM gives the evolution of density, pressure, compressibility and mean seismic velocity for P- and S-waves within the Earth's interior. P- and S-waves do exist within the mantle while S-waves do not propagate within the core because of its liquid state but reappear in the inner core.

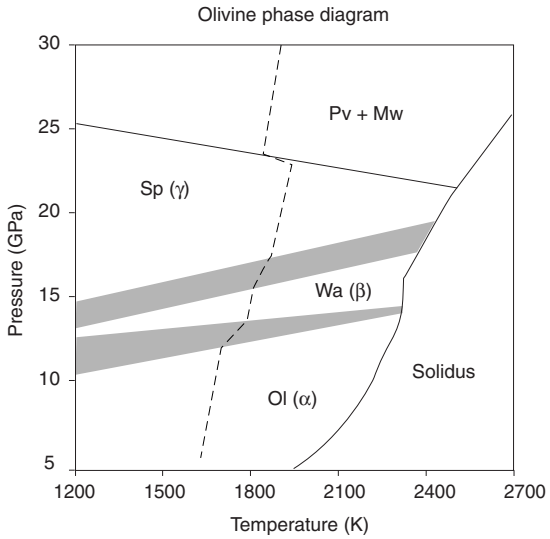


Figure 3. Stability diagram of the olivine mineral phase as a function of temperature and pressure. The Clapeyron slope is the slope of the phase stability in a temperature–pressure diagram. Transitions from α to β phases and from β to γ phases are exothermic (positive Clapeyron slope) while the transition from γ phase to perovskite and magnesiowustite is endothermic (negative Clapeyron slope).

chemically similar but differ by the presence of volatile components (CARBONACEOUS CHONDRITES) or by their iron content and the ratio of oxidized to metallic iron (ordinary or enstatite chondrites). Table 1, which gives the relative distribution of the main silicate minerals in chondrites, shows that olivine is the most abundant mineral. During the accretion phase of the Earth, the conversion of kinetic and gravitational energy into thermal energy raised the temperature above the melting point and induced the segregation of an olivine rich mantle floating above a denser Ni–Fe rich core. Thus, the composition of the Earth, and therefore of its mantle, is broadly chondritic even though it does not exactly match any one particular type of chondrite.

Table 1. Normative mineralogy of ordinary chondrites.

Species	High iron	Low iron
Olivine	36.2	47.0
Hypersthene	24.5	22.7
Diopside	4.0	4.6
Feldspar	10.0	10.7
Apatite	0.6	0.6
Chromite	0.6	0.6
Illmenite	0.2	0.2
Troilite	5.3	6.1
Ni-Fe	18.6	7.5

From these considerations, scientists have proposed dividing the Earth into three reservoirs: the bulk silicate Earth (primitive mantle), the depleted mantle and the bulk Earth (the core, the mantle and the crust). Several compositions have been estimated for the silicate portion of the Earth, based on meteorites and on samples of upper mantle and crustal rocks. The pyrolite, defined by A E Ringwood from complementary composition of melts and residual mantle material, is one of the mixtures representative of the Earth's mantle rocks which are now used to study the behavior of mantle material (see the review by Horton E Newsom in Ahrens (1995)).

Mantle mineralogy

Besides olivine, the main silicate minerals in the upper mantle are orthopyroxene, clinopyroxene, garnet, illmenite and chromite. As the stability of these minerals is only achieved at limited pressure and temperature ranges (see figure 3), the mineralogy of mantle rocks changes with depth, due to solid–solid phase changes. The elastic properties of the rocks are therefore very sensitive to the phase changes of olivine, which induce clear signatures on the seismic velocity profiles. These discontinuities at depths of around 410 and 660 km correspond to the transition from α to β forms of olivine minerals and to the transition of the spinel phase of olivine to perovskite and magnesiowustite minerals.

High-pressure experiments, which are now conducted with multi-anvil presses and with diamond-anvil presses, allow the study of the temperature–pressure stability domains of olivine. This apparatus consist of a thin sample of rock between two diamonds, heated with a laser device. It is then possible to reach pressures corresponding to the outer core pressure. The main problems encountered concern the small size of the samples and the actual control of the temperature and pressure during the experiment. Nevertheless, it has been possible to draw the Clapeyron slopes of the olivine mineral phase changes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Clapeyron slope marks the boundary of the phase existence domain in a temperature–pressure diagram. The agreement between mineralogical studies and seismically recorded discontinuities has proved the implication of the olivine phase changes in the mantle discontinuities nature of the rock mantle phase changes. However, this cannot preclude that a part of the discontinuity may also arise from a compositional change. It is important to note that the α to β olivine phase change, at a depth of 410 km, is characterized by a positive Clapeyron slope (exothermic reaction), while the transition from γ-spinel to perovskite, at a depth of 660 km, displays a negative Clapeyron slope (endothermic behavior). These values of the Clapeyron slopes will strongly influence the dynamics of convection currents within the mantle.

Thermal state of the mantle

Scientists still use direct methods to measure the temperature inside the Earth. However, the stability domains of olivine minerals give reference points to fix the temperature at phase change levels. Thus, according to the Clapeyron slope of the γ-spinel to perovskite phase change, the mean temperature at a depth of 660 km is around 1900 to 2000 K. Similarly, the temperature around 410 km should be close to 1600 K. From geochemical considerations, the rate of partial melting observed near the mid-oceanic ridge basalts leads to a temperature of 1400 K under the lithospheric plates. Temperatures can be extrapolated downward through the mantle on the basis of an adiabatic increase which should rise about 1000° above the CMB temperature. On the other hand, recent progress in high-pressure and high-temperature experiments allow the behavior of iron in physical conditions close to those prevailing within the Earth's core to be studied. At the inner core boundary, under the effects of pressure, the core material reaches a transition from liquid to solid which is proved by the reappearance of S-wave propagation. Unfortunately, uncertainties about the actual composition of the core and about the effects the alloy has on the decrease of the crystallization temperature, lead to quite important uncertainties in the determination of the inner core temperature. It may be around 6000 K for pure iron, but extrapolation towards the surface suggests a temperature of around 4000 K at the CMB, approximately 1000 K for the effect of the alloy on the crystallization temperature and 1000 K for the adiabatic decompression through the outer core. In this case, the temperature jump at the CMB could be a few hundred degrees to a thousand degrees. These results are in agreement with the presence of a low velocity layer at the CMB, the D" layer, in the seismological Earth models. The thickness of this layer varies from a few hundred kilometers in some areas to places where it seems disappear. The high thermal gradients in such a thin layer may induce partial melting for sinking material and may be a good place to generate new rising plumes through the mantle. Many in the geophysical community believe that it is the final destination of sinking slabs and the origin of the hot spot volcanism. This could explain the particular signature of the oceanic island basalts which are more primitive in character than the depleted mid ocean ridge basalts.

The fluid mantle

For short time scales such as the time taken for the passage of an earthquake wave through the Earth, the Earth's mantle rocks behave as perfect solids with a very weak attenuation of seismic energy. Thus, the most powerful earthquakes generate free oscillations and long-period surface waves which may last for several hours or even days. This perfect elastic behavior does not occur when the characteristic times of the events are comparable with geological time scales. This is well illustrated by the observation of post-glacial rebound which is occuring today following the melting of the ice sheets which covered the old Canadian and Fenno-scandien cratons during the last glacial period. The relaxation of constraints due to the disappearance of the ice loads on the continent provokes, a few tens of thousand years later, a visco-elastic rebound of the surfaces; these are currently rising by a few millimeters a year. These post-glacial rebounds can be detected from coastline evolution and from gravitational anomalies. Numerous models have been proposed to quantify these trends based both on the elastic (explaining the reaction to unloading) and the viscous behavior (explaining the time delay) of the mantle. The comparison of these models with geological observations make it possible to estimate a viscosity of 10^{23} to 10^{24} Po for the mantle, which is no longer considered as a solid but as a very viscous fluid. This viscosity is so high that it is impossible to make laboratory deformation experiments on rocks. For much longer times such as those of continental formation and drift (several million to several billion years) it is possible to consider the mantle as a perfect fluid. The continents, for which the mean density is lower, float on the mantle like icebergs in water. They display deep cold roots, which may reach several hundreds of kilometers within the mantle. They are dragged along by the deep mantle motions and have supported several episodes of accretion and dispersion of supercontinents during the history of the Earth.

The convective mantle

At the beginning of the history of the Earth, around 4.55 billion years ago, a phase of intense planetoid accretion occurred which continued for approximately 100 million years. During this phase, the kinetic energy of the accreting masses was transformed into heat that caused the whole planet to melt and the gravitational segregation of iron to create the core. A significant proportion of that heat is retained within the Earth's mantle and is currently still escaping. World-wide heat flow measurements at the surface of the Earth have shown that the flux dissipated by the mantle is approximately 42×10^{12} W. This is roughly twice the heat release which would be expected from the radioactive disintegration of the four remaining long-life radioactive elements (^{40}K, ^{235}U, ^{238}U and ^{232}Th). However, silicate rocks conduct heat badly (a simple calculation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

shows that it would take approximately 40 billion years to cross the entire mantle by conduction). Therefore, a large proportion of the heat which accumulated during the Earth's accretion, and also that which is continuously produced by RADIOACTIVITY, is available to drive convection within the mantle by inducing density anomalies through thermal expansion. The effects of these density anomalies are counteracted by the adiabatic cooling, the thermal dissipation of heat to the surrounding fluid and the viscous friction. All these effects are gathered into one dimensionless number: the Rayleigh number.

This quantity is representative of the geometrical aspects of a fluid layer but also of the thermodynamical and mechanical properties of the fluid. When the Rayleigh number is lower than a critical value, heat only escapes from the fluid layer by conduction. However, when the Rayleigh number is higher than the critical value, the fluid cannot stay at rest. Convection starts and proceeds to evacuate heat much more efficiently than conduction. It is possible to calculate the Rayleigh number of the mantle by considering the values of the physical parameters revealed by seismology, and the viscosity value obtained from post-glacial rebound. The over-adiabatic temperature and the internal heating can be estimated from composition, mineralogical studies and heat flux measurements. The value finally obtained for the Rayleigh number is around 10^6, that is to say several thousand times higher than the critical value. Therefore, the Earth's mantle cannot stay at rest but is churned by the very efficient motions of convection. The technological improvements to computers have allowed the development of more and more sophisticated numerical models of mantle convection which have recently concluded that the mantle is not only convecting but that this convection is highly chaotic.

Effects of the phase changes on mantle convection

The two main discontinuities in the Earth's mantle at depths of 410 km and 660 km will influence the structure of the convection currents which are crossing them. Indeed the density jumps, due to phase changes, are locally much stronger than the density anomalies due to thermal expansion. When a sinking convection current (i.e. a current colder than the surrounding material) encounters the endothermic phase change at a depth of 660 km, its low temperature implies that the phase change will occur deeper than the mean depth of the discontinuity, according to the negative Clapeyron slope. Then, the upper mantle material may be blocked in the light phase, and its sinking buoyancy may be too weak to reach the actual phase change level. In this case, the sinking current may be unable to cross the discontinuity that may separate upper and lower mantle into independent convective systems. The reverse situation is analogous for rising currents (i.e. a current hotter than the surrounding material) which attempt to cross an endothermic discontinuity from below. Conversely, when a sinking current encounters an exothermic discontinuity, as at a depth of 410 km, the phase change occurs at lower pressure and the fluid easily reaches the actual transition level.

Figure 4 shows the effects of such an endothermic phase change on the structure of convection. In panel (a), the Clapeyron slope is set to zero (i.e. no phase change) and the convection can cross the discontinuity without difficulty. In panel (c), the phase change is set to -4×10^6 Pa K^{-1}, the convection currents are blocked at the discontinuity which can no longer be crossed by the fluid. The most interesting case is given in panel (b); there the Clapeyron slope has been set to -2×10^6 Pa K^{-1} (this value is very close to the experimental value for the spinel-perovskite phase change). Fluid accumulates for a while at the discontinuity and when the threshold is crossed sudden and violent avalanches begin which draw huge quantities of matter from the upper to the lower mantle. This effect, which occurs for a Clapeyron slope close to that of mantle rock, has been found to be a common feature for a very broad range of geophysical and geometrical parameters (Brunet and Machetel 1998).

Evidence for mantle convection

Theoretical studies of fluid mechanics are therefore able to describe the chaotic behavior and global structure of convection within the mantle. This phenomenon is the huge engine which is responsible for the large-scale tectonic motions that are observed at the Earth's surface. Nowadays, the connection between the lithospheric plate motions and mantle convection have to be more accurately studied in order to be quantified. Indeed, the lithospheric plates are much more rigid than the underlying mantle and cannot be rigorously taken into account by the numerical models of convection which assume a fluid behavior for the mantle. The continuous generation of lithosphere at the mid-ocean ridges, its journey at the Earth's surface and its return to the mantle at subduction zones cannot be understood without the churning of matter induced by mantle convection. Similarly, continental drift and the cyclic creation and dispersion of supercontinents necessitate mechanisms of comparable amplitudes within the mantle to those which are observed at the surface.

The chaotic thermal anomalies, induced by mantle convection, result in local density and temperature anomalies within the mantle which affect the velocity of wave propagation mainly through the rigidity of the rocks. Therefore, a cold anomaly induces a positive velocity anomaly and, conversely, a high temperature creates a slow zone. On the basis of radial seismic Earth models (e.g. PREM), seismologists have been able to calculate local departure to mean values, thanks not only to the improvement in seismological networks but also to improvements in computer performance which have allowed the treatment of more numerous data and more complex computations. To reach this goal, the Earth has been decomposed into small cells for which a global inversion of all the ray paths is used to calculate local velocity. It would be tempting to consider that this procedure will give us a detailed map

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

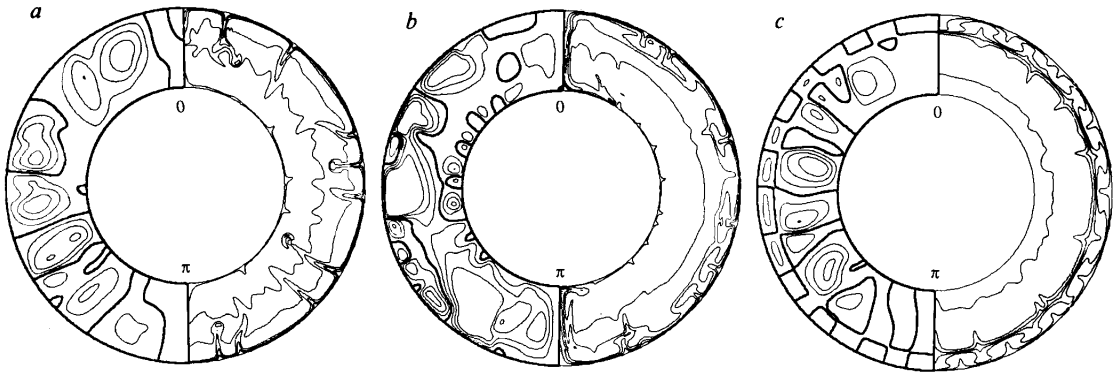


Figure 4. Isocontours of stream functions (left part of the spheres) and temperature fields (right part of the spheres) for a model mantle with endothermic phase change at 670 km. When the Clapeyron slope is set to 0, no influence of the phase change can be deciphered on the global mantle circulation (panel a). With a Clapeyron slope of -4×10^6 Pa K^{-1}, the endothermic phase change prevents convection from crossing the discontinuity that results in a two-layered convection between lower and upper-mantle (panel c). With a Clapeyron slope close to the value given par mineralogical studies (-2×10^6 Pa K^{-1}), a partly layered regime prevails with occurrence of mantle avalanches (panel b).

of the temperature field within the mantle. Unfortunately, seismic sources are not evenly distributed on the surface of the Earth; they are concentrated mainly at plate borders leaving large areas of oceanic plates without seismic sources. Furthermore, seismic stations are, for obvious commodity and maintenance reasons, mainly installed on continents that are better able to cover the northern hemisphere. Additionally, the interpretation of seismic velocity anomalies in terms of temperature is only an approximation since other effects such as anisotropy or compositional anomalies may affect the wave velocity. Therefore, the tomographical studies only give a rough (but improving continuously) idea of the thermal structure of the mantle. Up to now the results do not allow conclusions to be drawn either for two-layer convection, whole layer convection or partly layered convection within the Earth's mantle. Figure 5 gives the result of a tomographical study by Van der Hilst and Seno (1993). It seems from the most recent seismological results that certain slabs can cross the endothermic discontinuity at a depth of 660 km, as in the Mariana Trench, whilst others bend at this discontinuity without crossing it.

The convective thermal anomalies also induce perturbations of the gravity field of the Earth. Indeed, a rising convection current corresponds to the rising of hot, and therefore lighter, material that results in a negative density anomaly. However, the thermal effects of convection cannot be separated from the dynamical effects of the convection currents on the surface and discontinuity topographies. The strength of the rising currents pushes and deforms the surface leading to an opposite effect on gravity (creation of a positive density anomaly). Furthermore, for a rising convection current crossing the entire mantle, the surface would be deflected upward, resulting in a positive density anomaly; the 660 km discontinuity would also be deflected upward, but the 400 km exothermic discontinuity would be deflected

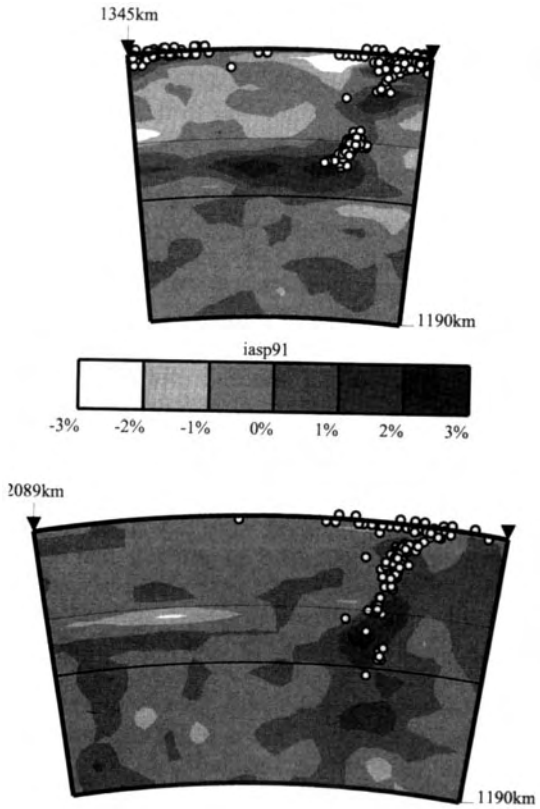


Figure 5.

downward. The compensation of these effects at different depths depends strongly on the viscosity contrasts, which are assumed between the different parts of the mantle. This quantity remains very badly constrained. The relationship between the global gravity field and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

convection will strongly depend on the global structure of convection, which may not extend over the whole thickness of the mantle.

Avalanches and large scale tectonic features

Recent numerical models of mantle convection, based on the existence of mantle avalanche phenomena, have been used to propose self-consistent mechanisms for large-scale tectonic features such as huge volcanic events, the world-wide increase of basalts generated at mid-oceanic ridges, periods of low-reversal of the magnetic fields, the pulsation of super-continents, and the brutal acceleration of continental motions (as for India during its journey toward Eurasia). The avalanches are characterized by the accumulation at the 660 km discontinuity of sinking cold material, which are not immediately able to cross the discontinuity. When the avalanche starts (figure 6), a brutal transfer of matter occurs from the upper to the lower mantle. This sinking matter is compensated by return flows from the lower mantle into the upper one. While the avalanches are localized on the sphere, these return flows imply huge world-wide arrivals of hot lower mantle near the surface that may induce faster and more efficient dorsal evacuation of heat. The cold material sinking during an avalanche is colder and therefore more viscous than the surrounding material. It quickly reaches the core surface where it sweeps the hot, less viscous material into very high unstable temperature anomalies which may start as vigorous new plumes toward the surface. These plumes cross the lower mantle to spread at shallow depth under the surface. The low-viscosity zones which are created may lubricate the motion of the converging continents above the avalanche inspiration. At the same time, the arrival of cold material at the core surface increases the heat transfer from the core to the mantle, which may increase the stability of the convective structure into the outer core and to stabilize the dynamo.

Conclusions

Our knowledge of the Earth's mantle has progressed rapidly over recent decades thanks to improvements in metrology, seismology and computing power. However, it is not yet known whether convection in the Earth's mantle occurs in cells that extend throughout the entire depth of the mantle, or if convection in the upper and lower mantle is separated. A mixed situation could prevail with partial layering of the mantle. The more and more accurate seismic tomography results describe a complex situation with no general rules but a strong sensitivity to the local constraints of global plate motion. The slabs either bend at the 660 km discontinuity or cross it deep into the lower mantle. A slow cooling has decreased the Earth's internal temperature since the Archean era. Geochemical studies of lavas show that it may have reached a few hundred degrees, which may have changed the global regime of convection from layering to whole mantle convection. This trend, if it were confirmed by future studies, would imply that at least for a period, an avalanche regime might have occurred. Such a possibility could have provided a mechanism powerful enough to have gathered and dispersed the supercontinents several times during the Earth's history.

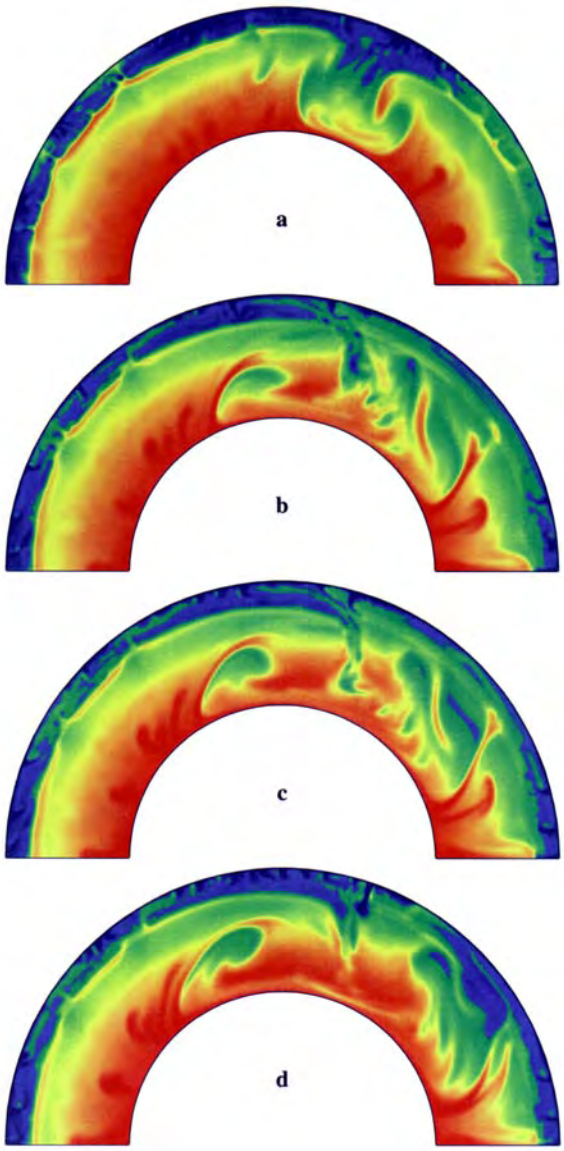


Figure 6. Evolution of the temperature field of a model mantle during an avalanche. If their existence is confirmed, such phenomena may be invoked to explain most of the large- scale tectonic features and in particular the apparent pulsation of supercontinents at the Earth's surface. **This figure is reproduced as Color Plate 35.**

Bibliography

Ahrens T J 1995 (ed) Global Earth Physics, a Handbook of Physical Constants (Washington: American Geophysics Union) p 376

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Anderson D L 1989 Theory of the Earth (Oxford: Blackwell) p 366

Brunet D and Machetel P 1998 Large-scale tectonic features induced by mantle avalanches with phase, temperature, and pressure lateral variations of viscosity *J. Geophys. Res.* **103** 4929–45

Machetel P and Weber P 1991 Intermittent layered convection in a model mantle with an endothermic phase change at 670 km *Nature* **350** 55–7

Van der Hilst R and Tetsuzo S 1993 Effect of relative plate motion on the deep structure and penetration depth of slabs below the Izu-Bonin and Mariana island arcs *Earth Planet. Sci. Lett.* **120** 395–407

Philippe Machetel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maraldi, Jean-Dominique [Giovanni, Maraldi II] (1709–88)

Italian-French astronomer. Observed the planets, discovered two globular clusters in 1746, M15 and M2.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mare (pl: maria)

A dark lunar plain. The name, which is Latin for 'sea', was first used on maps of the Moon in the seventeenth century, when astronomers believed the Moon to be Earthlike, and that the light and dark areas were land (sometimes called 'terrae') and water. In fact the maria are composed of basalt, solidified lava which erupted between 3 and 4 billion years ago after the Moon had suffered large impacts which weakened or, in some cases, penetrated its crust. 'Mare' has become so well established that it is retained in lunar nomenclature. The maria range in size from huge lava floodplains such as Oceanus Procellarum (2568 km across) and the large impact features Mare Orientale and Mare Imbrium, down to the 150 km Mare Anguis. Maria are much more prevalent on the Moon's nearside; on the farside the crust is thicker and withstood impacts better. They are smooth areas with few craters: by the time they had formed the great episode of cratering was over, and they bear the scars only of recent impacts, which in the past 3 billion years have been much less frequent.

The term 'mare' was also formerly used for dark regions on Mars (e.g., Mare Tyrrhenum, named after the Tyrrhenian Sea, between Italy and Sicily), which in the nineteenth century were also assumed to be 'seas'. However, in general they do not correspond to topographic features, and are now used only on albedo maps (showing areas that appear light and dark) of the planet.

See also: Imbrium basin, Orientale basin, solar system: nomenclature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mariner Missions

Mariner was the name given to the earliest set of American space missions to explore the planets and to the spacecraft developed to carry them out. The missions were planned and executed by the JET PROPULSION LABORATORY (JPL) of the California Institute of Technology, which had been designated by the National Aeronautics and Space Administration (NASA) as its lead center for planetary missions.

Birth Pangs of the Planetary Program—1960–1962

In January 1959, shortly after NASA had been established and JPL had become affiliated with it, the laboratory proposed a program of space missions that contemplated sending two spacecraft to VENUS and two to MARS at each opportunity starting in 1960. With some modifications, this program was adopted by NASA and titled VEGA. In the next few months, as it became clear that launch-vehicle development was falling short of expectations, Vega was canceled and replaced by less ambitious programs—RANGER for lunar and Mariner for planetary missions. The Mariner program, approved by NASA in July 1960, would be carried out by a series of increasingly capable spacecraft, beginning with Mariner A to fly by Venus in 1962 and Mariner B for missions to Venus and Mars (possibly with landers) beginning in 1964. The spacecraft were to be launched by the Atlas–Centaur launch system.

Considerable progress was made at JPL in developing the spacecraft and the missions for Mariners A and B, including the scientific payloads and the teams of scientists to implement them, but the development of the Centaur upper-stage rocket ran into continual problems so that Mariner A was canceled in August 1961. To replace it, JPL proposed to use a simpler Ranger-type spacecraft on top of the Atlas–Agena launch system for the first observation of Venus. The scientific payload would have to be severely curtailed. In recognition of its Ranger heritage, the project was called Mariner R.

Mariner 1 was launched from Cape Canaveral on 22 July 1962. A small error in the launch-guidance equations in the computer onboard the Atlas caused it to fly erratically, and it was destroyed by the range safety officer after 290 s.

The features of Mariner spacecraft are summarized in table 1.

Mariner 2 to Venus—1961–1963

The identical Mariner 2 spacecraft was launched on 27 August, and it carried out its mission by flying by Venus on 22 December at a distance of 34 762 km. It carried a payload of only six scientific instruments weighing only 18 kg, but each of them operated essentially as intended. The Microwave Radiometer measured the temperature of the planet surface to be about 425 °C and the Infrared Radiometer registered the cloud-top temperature as −35 °C.

The other experiments operated continuously during most of the flight, providing the first long-duration monitoring of conditions in interplanetary space. The Micrometeorite Detector registered the impacts of dust particles, detecting only two. The Charged-Particle Experiment monitored cosmic rays. The Magnetometer Experiment monitored the interplanetary field. The Solar-Wind Experiment determined for the first time the density, velocity, temperature and composition of the plasma streaming out from the solar corona.

The spacecraft continued to operate nearly faultlessly for 21 days beyond Venus encounter (130 days total) out to 87.4 million km from Earth. It thus demonstrated the efficiency of its three-axis stabilized design, its capability of effective two-way communication with Earth over vast distances with a transmitter power of only 3 W and the ability of the Deep Space Network for precise tracking throughout the flight. It was the most successful space mission in history up to that time.

Mariner 4 to Mars—1962–1965

The Mariner B mission was originally authorized in July 1960 for a landing on Venus or Mars in 1964. In February 1961 the Venus lander was dropped from consideration, but in April 1962 it was reinstated, and the Mars landing was dropped. In May 1963 the Mars landing was postponed until 1966 and redesignated Mariners E and F, and these were effectively canceled in July 1964. This sad history resulted from the troubles with the Atlas–Centaur.

Meanwhile, a much less ambitious Mars flyby mission, called Mariner C, with a spacecraft barely half the weight of Mariner B, was proposed in midsummer 1962, approved in March 1963 and ready for launch by the Atlas–Agena in November 1964. The launch of Mariner 3 on 5 November appeared normal, but the signal of solar-panel deployment was not received, and communication ceased as the battery ran down. It was soon determined that the nose cone (appropriately called the 'shroud') had failed to jettison. This information prompted a frantic effort to design and fabricate a new shroud in the three weeks before the launch period closed.

Mariner 4 was launched on 28 November 1964 and flew by Mars on 14 July 1965 at a range of 9844 km, taking a sequence of 21 photographs of the surface. The mission revealed two very important new facts: (1) that the surface was covered with impact craters; (2) that the atmosphere was mainly carbon dioxide at a pressure of less than 7 mbar. The latter was determined by the Radio Occultation Experiment, which had not even been accepted as part of the mission until after launch. The other scientific experiments were three to detect charged particle radiation, a Solar Plasma Probe, a Magnetometer and a Cosmic Dust Detector.

Mariner 5 to Venus—1965–1967

A project to revisit Venus with a spare Mariner 4 spacecraft was authorized in December 1965, and Mariner 5 was launched on 14 June 1967. It flew past Venus at 10 151 km

Mariner Missions © Kluwer Academic Publishers 1997

Table 1. Mariner spacecraft.

Spacecraft	Target	Mass (kg)	Science payload (kg)	Data rate (bits s^{-1})	Data storage (megabits)
Mariner 2	Venus	203	18	8.33	0
Mariner 4	Mars	261	16	33.3	5.24
Mariner 5	Venus	245	10	33.3	1
Mariners 6, 7	Mars	413	58	670	195
Mariner 9	Mars	998[a]	63	16 200	180
Mariner 10	Mercury	534	78	117 600	180

[a] This includes 466 kg of propellant to put the spacecraft into Mars orbit.

from the center on 19 October and successfully conducted seven experiments, one day after the Soviet Venera 4 had dropped on the surface a landing capsule that failed to provide any data.

The Plasma Probe (1), the Magnetometer (2) and the Energetic Particle Detectors (3) were designed to investigate the interaction between the planet and the interplanetary plasma. They determined that the magnetic field of Venus (if any) is much weaker than the Earth's and too weak to hold off the solar wind and produce an Earth-like magnetosphere. However, when the solar wind reaches the top of the atmosphere it is deflected by the ionosphere so that it flows around the planet without touching the surface. The Ultraviolet Photometer (4) found large quantities of hydrogen in the upper atmosphere at a very low temperature, but no oxygen was detectable.

Three experiments utilized the radio links to and from the spacecraft. The S-band Occultation Experiment (5), by analysing the downlink signal, obtained profiles of refractivity, temperature, density and pressure in the neutral atmosphere and of electron density in the ionosphere. These data demonstrated that the temperature was 475 °C and the pressure was 90 atm at the surface. The Dual-Frequency Occultation Experiment (6), using two radio transmissions from Earth that were analysed on the spacecraft, obtained very detailed information on the ionosphere. Range and Doppler tracking of the spacecraft provided the Celestial Mechanics Experiment (7), which determined the masses of Venus and the Moon with unprecedented accuracy.

Mariners 6 and 7 to Mars—1965–1969

For the 1969 missions to Mars, the availability of the Atlas–Centaur launch system made possible major increases in the size and sophistication of the spacecraft and their scientific payloads. Mariner 6 was launched on 25 February and its twin, Mariner 7, on 27 March; they flew by the planet on 31 July and 5 August, respectively. Each spacecraft carried two vidicon cameras—wide angle and narrow angle—and a much higher data transmission rate provided 200 times the picture data of Mariner 4. The two spacecraft acquired 143 pictures before encounter that showed the entire visible disk or a considerable fraction of it and 59 pictures near encounter that covered small areas, about half of which had resolutions of 0.2 km or better. All the close-up pictures were in the southern hemisphere and, although they covered only about 10% of the surface, they did reveal some new surface features and laid to rest the myth of the Martian canals. Later missions, however, showed that most of the interesting features of Mars had been missed or misinterpreted.

The Infrared Radiometer measured a temperature of 150 K on the south polar cap, indicating it to be carbon dioxide ice and not water ice. The Ultraviolet Spectrometers on both spacecraft detected carbon monoxide in the atmosphere but, surprisingly, saw no trace of nitrogen. The Radio Occultation Experiment confirmed the Mariner 4 finding on the atmospheric pressure and detected an ionosphere. The Celestial Mechanics Experiment measured the planet's mass with unprecedented precision.

Mariner 9 to Mars—1968–1972

For the next Mars opportunity in 1971, the plan was for a dual-spacecraft orbiting mission that promised a large increase in information over the three earlier flybys. However, on 8 May the Centaur booster failed to boost Mariner 8 into orbit. Mariner 9, successfully launched on 30 May, arrived at Mars on 14 November to find the planet shrouded by the most intense global dust storm that had ever been observed. Settling into its intended orbit, the spacecraft simply waited out the storm, which had cleared by March, and in 349 days of operation it achieved virtually all of the objectives of the dual mission.

Mariner 9 revolutionized our understanding of Mars. The camera returned 7329 photographs and 54 billion data bits—27 times as much as the three preceding missions. It discovered many volcanoes, including OLYMPUS MONS, 27 km tall, an enormous system of deep canyons, dubbed VALLES MARINERIS, that stretch about one-quarter of the way around the planet, a plethora of channels of five different types, many of them appearing to be ancient river beds, numerous evidences of eolian erosion and deposition, a variety of meteorological phenomena and many other things.

The Infrared Radiometer Experiment measured the thermal inertia of the surface over much of the planet and found the surfaces of PHOBOS AND DEIMOS to be very poor thermal conductors. The Infrared Spectrometer Experiment detected carbon dioxide, water, dust and ice crystals in the atmosphere, monitored the atmospheric

temperature and determined the atmospheric pressure at numerous points. The Ultraviolet Spectrometer Experiment measured the atmospheric pressure over much of the surface, thus determining the relative altitude in many places, and measured ozone and atomic hydrogen in the atmosphere.

The S-band Occultation Experiment, with more than 300 occultations to analyze, made many measurements of the atmospheric temperature profile, the surface pressure, and the electron density in the ionosphere. The Celestial Mechanics Experiment, by analyzing the metric information in the tracking data, determined a more accurate description of the gravity field of the planet, improved the ephemeris of Mars by an order of magnitude, and obtained topographic data with a precision of 100 m.

Mariner 10 to Venus and Mercury—1969–1975

A mission to MERCURY was authorized by NASA in December 1969. For the first time in the Mariner program, the spacecraft was not to be built by JPL in house, and the Boeing Corporation won the contract to build it. Mariner 10 was put on a course to Venus by an Atlas–Centaur rocket on 3 November 1973 and flew by the planet on 5 February 1974 at just the proper location to be deflected by the gravitational field into an orbit to Mercury. This was the first employment of the 'gravity-assist' technique of interplanetary navigation. Mariner passed Mercury on 29 March at a distance of 703 km and was deflected into a new orbit that took it around the Sun in 176 days and back to Mercury at intervals of exactly two Mercurian years. The second encounter occurred on 21 September at about 50 000 km, the larger distance being chosen to extend the photographic coverage at good viewing angles. The third, on 16 March 1975, was targeted to the dark side of the planet at 327 km altitude in order to maximize the information about the magnetic field and the magnetosphere.

Mariner 10 exhausted its attitude-control gas supply on 24 March 1975, and the spacecraft transmitter was turned off for the last time, terminating the mission after 506 days and nearly 1.6 billion kilometers of travel.

Mariner 10 was the most complex and sophisticated of all the Mariner spacecraft, and its mission achieved many space firsts. It was the first mission to visit two planets and to fly by the same planet three times. It was the first to return photographs in real time instead of storing them, as the tape recorder failed before the second Mercury encounter. Rotatable solar panels were used for the first time to maintain their temperature and power output through the large change in solar distance. 'Solar sailing' was attempted and used successfully for the first time, as the positions of the solar panels and the high-gain antenna were manipulated to maintain the orientation of the spacecraft by radiation pressure without the expenditure of attitude-control gas; this technique made the third encounter possible. Optical navigation was employed for the first time as the cameras provided the angles between Mercury and stars. The Celestial Mechanics Experiment had available, for the first time, radio signals in both X- and S-band.

The TV camera was equipped with two CASSEGRAIN TELESCOPES to provide high- and low-resolution pictures and a filter wheel to permit photography in several wavelength bands, including the ultraviolet. It returned more than 1000 pictures of Earth and Moon, 3500 pictures of Venus and 3700 pictures of Mercury, a small number of which had resolution as small as 134 m. They provided very detailed coverage of the clouds of Venus (made possible by the ultraviolet sensitivity of the cameras) and the surface of Mercury. The motion of the clouds appeared to indicate that the primary circulation pattern on Venus is vertical upward motion at the subsolar point and subsidence over the poles.

Mercury was found to look generally like the MOON, but with some distinctly non-lunar features, including large scarps or cliffs nearly 3 km high and as much as 500 km long, probably indicative of crustal shrinkage. The density of small craters was surprisingly similar to that on the Moon and Mars, implying that all these planets received similar intensities of meteorite bombardment, contrary to earlier assumptions.

A major and unexpected discovery was an intrinsic magnetic field, which, although about 60 times weaker than Earth's, was sufficient to produce a bow shock and a magnetosphere that was a miniature copy of Earth's.

The scientific payload included also a Radio Science Experiment, a Solar Plasma Experiment, a Charged Particle Telescope, an Infrared Radiometer and two Ultraviolet Spectrometers. One of these was designed for detecting planetary airglow; the other, intended for sounding Mercury's atmosphere by looking at the sun as it was occulted, also made observations on Comet Kohoutek and the Gum Nebula. It detected a very sparse atmosphere of helium on Mercury. Surface temperature extremes of 90 K and 460 K were measured by the radiometer. The Radio Science Experiment measured the diameter and mass of Mercury with unprecedented precision and detected no trace of an ionosphere.

The exploration of the solar system, ably initiated by the Mariner series, was continued by spacecraft called VIKING and VOYAGER.

Bibliography

The scientific results of each successful mission were reported during and after the mission by papers in *Science* and *Journal of Geophysical Research*.

Ezell E C and Ezell L N 1984 *On Mars: Exploration of the Red Planet, 1958–1978* Scientific Publication 4212 (Washington, DC: NASA)

Koppes C R 1982 *JPL and the American Space Program* (New Haven, CT: Yale University Press)

Snyder C W and Moroz V I 1992 Spacecraft exploration of Mars *Mars* ed H H Kieffer, B M Jakosky, C W Snyder and M S Matthews (Tucson, AZ: University of Arizona Press) ch 3, pp 71–119

Conway W Snyder

Marius [Mayr], Simon (1573–1624)

Astronomer, born in Gunzenhausen, Bavaria, Germany, competitor of GALILEO. Educated in astronomy and meteorology, he was appointed mathematician of the Margrave of Ansbach in 1601, printing an annual almanac as part of his duties. Went to Prague to study with TYCHO BRAHE, moved to Padua to study medicine. Observed the nova of 1604 with BALDESSAR CAPRA and helped Capra published a book on it. Capra lost his reputation by flagrantly plagiarizing a manuscript by Galileo and was expelled from the university. The scandal tainted Marius, who moved back to Ansbach where he learned in 1609 from an artillery officer, Freiherr Hans Philip Fuchs, that a Dutchman had tried to sell him a telescope. Marius grasped the concept and realized it with a telescope which he used to observe Jupiter. He claimed in a book of 1614, *Mundus Iovialis Anno M.DC.IX Detectus Ope Perspicilli Belgici (The Jovian World, discovered in 1609 by means of the Dutch Telescope)*, that he had first observed Jupiter's moons in December 1609, a month before Galileo. The claim infuriated Galileo, coming from someone involved in an earlier theft of Galileo's work, and especially as there were unresolved inconsistencies in the observations and dates. Whatever the truth of that matter, Marius's names for the satellites stuck. *Mundus Iovialis* contains Marius' telescopic observation of M31, the Andromeda nebula, something that he independently discovered, although it had been observed around AD 905 and documented in AD 964 by the Persian astronomer AL-SUFI, in his *Book of Fixed Stars*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars

Mars is the fourth planet from the Sun and the most Earth-like. Its radius is only half that of the Earth, it is significantly colder and has a much thinner atmosphere (table 1). However, the Martian environment is the most similar to the terrestrial one in the solar system: most atmospheric and geologic processes have their equivalent on our planet, for instance the global atmospheric circulation and the volcanic processes. One of the major differences is related to the absence of liquid water at the surface in the present conditions, although there is much evidence that liquid water existed in the past.

Initial data from NASA's Mars Odyssey gamma ray spectrometer instrument suite has been used to detect hydrogen, which indicates the presence of water ice in the upper meter (three feet) of soil in a large region surrounding the planet's south pole. The amount of hydrogen detected indicates 20 to 50% ice by mass in the lower layer. Because rock has a greater density than ice, this amount is more than 50% water ice by volume.

Mars is one of the brightest objects in the sky and therefore has been known since the earliest antiquity. Because of its distinct red color, strong brightness and strange movement in the sky, it has often been associated with death or war by the oldest civilizations; its modern name is actually that of the Roman god of war. Historically, the observation of Mars has often triggered important astronomical progress. This is still true in the modern era, since Mars has been the target of more than 30 spacecraft as of this writing. The reasons for this situation are not only that Mars is the most accessible planet but also that it has the most Earth-like environment and is probably the most favorable location for the development of life in the solar system: although the current conditions at the surface are actually very hostile to life, this was almost certainly different in a distant past. Mars is therefore likely to become the first planet where humans will go, although this still appears a rather distant issue.

The two satellites of Mars are also interesting objects. They were discovered only in the late 19th century because of their very small size and proximity to Mars. Both satellites are very dark and have a very low density, indicating that they are most probably captured asteroids (see PHOBOS AND DEIMOS). Because of their orbital position around Mars, they are the most easily accessed objects of this type, although they probably suffer processes that are not common on other asteroids, in particular higher impact fluxes.

Table 1. Mars' main characteristics.

	Mars	*Earth*
Mass	6.4185–10^{23} kg	59.75–10^{23} kg
Mean radius	3389.92 km	6 388 km
Mean density	3.9335 g cm^{-3}	5.52 g cm^{-3}
Moment of inertia	0.3662±0.0017	0.3308
Semimajor axis	227.94–10^6 km	149.6–10^6 km
Eccentricity	0.0934	0.0167
Orbital period	686.98 Earth days	365.256 Earth days
Sinodic period	779.94 Earth days	—
Rotational period	24.6229 h	23.9345 h
Obliquity	25.19°	23.45°
Equatorial surface gravity	3.727 m s^{-2}	9.78 m s^{-2}
Equatorial escape velocity	5.027 km s^{-1}	11.18 km s^{-1}
Mean atmospheric pressure	5.6 mbar	1013 mbar
Mean surface temperature	–63 Â°C	13 °C

Observations and space exploration

The physical study of Mars began with the invention of the telescope at the beginning of the 17th century. Galileo first demonstrated the spherical shape of the planet from his telescopic observations. Some 70 yr later, Christiaan Huygens and G D Cassini had recognized the main characteristics of the planet (see MARS: SURFACE). The study of Mars improved very quickly at the end of the 19th century, when the progress of optical instruments made it possible to observe the high variability of the surface and atmosphere. However, it was only in the 1950s that atmospheric studies really developed, and many incorrect ideas were widely accepted at the beginning of the space era (see MARS: ATMOSPHERE). The first synthesis of modern observations was published by G de Vaucouleurs in 1954; it was generally considered at this time that the atmosphere was mainly composed of nitrogen, with a total pressure of 85 mbar—15 times the actual value—and that, although chlorophyll plants were certainly absent, surface variations could be due to the seasonal growth of lichen at the surface. Mars is therefore a dramatic example of how space exploration completely changed our perception of the solar system (see SOLAR SYSTEM: EXPLORATION). The earliest attempt to visit Mars was made by the Soviet Union in 1960. After six failures on the US and Soviet sides, Mariner 4 (see MARINER MISSIONS) finally returned images of Mars in 1965, during the first flyby of another planet. The results of the mission were actually surprising, since the oldest terrains resembled the Moon much more than the Earth—a view confirmed during the next flybys by Mariner 6 and 7 in 1969.

Mariner 9, the first spacecraft orbiting Mars in 1971, produced the first systematic mapping and analysis of the atmosphere. These data allowed us to obtain a full view of Mars' complexity, in particular concerning the surface characteristics of the planet and its very complex geologic history. In 1976 Viking (see VIKING MISSION), a very complete mission consisting of two orbiters and two

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

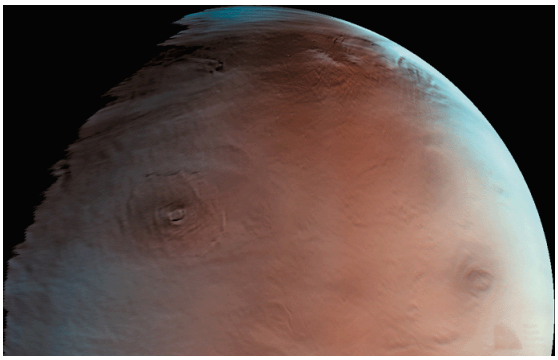


Figure 1. View of Mars acquired in June 1998 by Mars Global Surveyor, taken from low orbit, centered on northern Tharsis. Olmypus Mons is left of center and Ascraeus Mons at the lower right. Olympus Mons is about 550 km across. The image was taken during northern winter; clouds and haze can be seen to the north (up) above Olympus Mons. The morning terminator runs left of Olympus Mons. (Mars Global Surveyor image MRPS88925.)

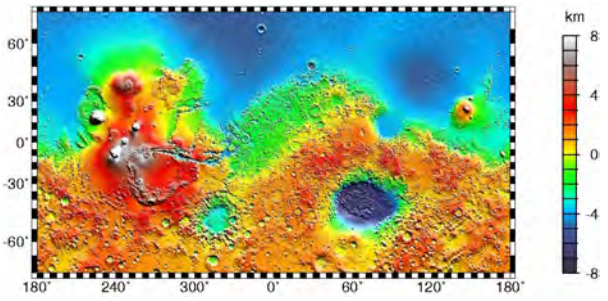


Figure 2. Topographic map of the equatorial regions of Mars, retrieved from radar measurements by Mars Global Surveyor (Mercator projection). The vertical acuracy is approximately 5 m. Obvious on the map are the hemispheric dichotomy, with cratered highlands in the south and younger low plains in the north, the Tharsis and Elysium domes at 260° and 150° longitude, the giant volcanoes of Tharsis, Olympus Mons (18°N, 228°E) and Elysium Mons (25°N, 150°E), Valles Marineris and the outflow channels (0°N, 340°E), and the two large basins in the southern hemisphere, Argyre (50°S, 320°E) and Hellas (45°N, 110°E) has been interpreted as further evidence for a possible ancient seafloor from these data. (Credit: NASA–MOLA Science Team.)

landers, was successfully delivered to Mars and performed observations until 1982. Most of our current knowledge of Mars is still derived from the analysis of these data. Other, more limited, results were acquired by Mars 2 and 3 (1971), Mars 4, 5 and 6 (1973), Phobos 2 (1989), and Mars Pathfinder (1997). After the failures of two major missions (Mars Observer in 1992 and Mars-96 in 1996), the exploration of Mars was revived with the launch of Mars Global Surveyor (NASA) (figure 1 and figure 2) in 1996 and the approval of the Mars Express mission (ESA), which should complete the phase of global characterization of the planet. The next major step will focus on *in situ* analyses, most notably with the Mars sample return project (NASA–CNES–ASI) scheduled for the period 2003–8.

Another, unexpected source of data concerning Mars was identified in the early 1980s. Several volcanic meteorites (called the SNC METEORITES) were identified as Martian in origin because of their young crystallization ages (150 My–1.3 Gy) and particular isotopic ratios similar to the Martian atmosphere. There are currently 13 SNC meteorites in collections that are thought to have been ejected during large impact events. The analysis of these rocks has helped put very strong constraints on the evolution of Mars, notably concerning its formation. They are not as useful as collected samples, however, because their source area is unknown and because we do not know to what extent they are typical of the Martian rocks.

Dynamics and climate

The movement of Mars in the sky is spectacular since, like all external planets, it has a retrograde motion when it approaches Earth at its closest distance. In the case of Mars the variation of the apparent speed is very quick, and the magnitude of this movement is much more marked than for Jupiter or Saturn. The very large variation of Mars' brightness with distance from Earth (about a factor of 50) was one of the major drawbacks of the antique geocentric system. In the modern period, the precise observations of Mars by Tycho Brahe led Kepler to state his laws of planetary motion, and in particular to reject the old assumption that planetary orbits are circular—indeed, Mars has by far the largest eccentricity among the 'regular' planets.

This orbital particularity has a deep impact on the Martian climate: there is a marked difference in seasonal length between the two hemispheres, with a longer and colder winter in the south at present, resulting in compositional differences between the polar caps, with a permanent water-ice cap exposed only in the north. This situation reverses every 25 000 yr, which may have induced large climatic variations in the past. On a time scale of 10 Myr, the periodic variations of obliquity (15°–35°) probably govern the climatic change, by changing the magnitude of seasonal effects. On longer time scales, these variations are chaotic and obliquity may range between 0° and 60°. This situation is unique among terrestrial planets, because Mars has not been stabilized by a massive satellite, like the Earth–Moon system, or by tidal dissipation like Mercury or Venus. As a consequence, Mars may have suffered very different climatic environments in the distant past, with a very different repartition of volatiles between the main reservoirs (polar caps, regolith, permafrost, atmosphere and possible water bodies). Indeed, the morphologic evidence suggests that liquid water ran at the surface not only at the end of the period of heavy bombardment when the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

atmosphere was thicker (3.8 Gyr ago) but also much more recently during several different episodes. In both cases, this would have required higher surface temperatures than at present, and therefore a GREENHOUSE EFFECT with larger atmospheric pressures. This is specially true during the early stages of the solar system (see ORIGIN OF THE SOLAR SYSTEM): according to stellar evolution models the Sun was 30% dimmer than it is today, so that the CO_2 pressure had to be at least 2–3 bar to permit such temperatures, which is consistent with estimates of planetary outgassing (see PLANETARY ATMOSPHERES). However, one of the main issues concerning the evolution of Mars is related to the stability of the primordial atmosphere and to the transition to a colder, lower-pressure environment. Another controversial topic concerns the possible existence of lakes or oceans during later episodes of climatic warming.

Internal structure and composition

Although the Viking landers were equipped with two seismometers, these experiments provided very few results about the internal structure of Mars. Most ideas in this field are derived from orbiting spacecraft and from the composition of SNC meteorites.

Mars, like the other terrestrial planets, is probably divided into a crust, mantle and core (see PLANETARY INTERIORS). The thickness and density of the core can be constrained from the knowledge of mass, radius and moment of inertia of the planet, with a rather large uncertainty. Although Mars Pathfinder measured the moment of inertia with a sufficient accuracy in 1997, two models are compatible with the current data: Mars could possess either a small core of high density (a core of iron would represent 15% of the planet's mass for a radius of 1300 km), or a large low-density core (a mixture of sulfur and iron representing 25% of the planet's mass for a radius of 2000 km). The first model is more similar to the Earth's core, but studies of the SNC meteorites' composition rather suggest a large fraction of nickel (~45%) and sulfur (~12%) in the core. The latter figure, while uncertain, is important because a high fraction of sulfur would slow down the solidification rate of the core through geologic times. The lack of strong magnetic field on a global scale implies that the core is not convecting and is therefore either completely solid or liquid. A magnetic field does exist at the surface, however, which was perhaps induced by a past planetary field, suggesting that a dynamo effect has existed, but stopped at a certain time as the core became solid and stably stratified.

The mantle thickness is estimated to be between 1500 and 2100 km, depending on the core properties. Globally, Mars is richer than Earth in Fe and poorer in Al. The mantle itself is traditionally divided into an upper and a lower mantle, the limit corresponding to the olivine–spinel phase transition. The upper mantle is estimated to be 900–1100 km thick and to be mainly composed of olivine and orthopyroxenes plus clinopyroxenes at moderate depth, like on Earth. The amount of water is estimated to be between 0.004% and 1% in mass. In the lower mantle, olivine is replaced by spinel and orthopyroxene by majorite. Other, higher-pressure phase transitions (e.g. β-spinel to γ-spinel) may occur in the lower mantle, depending on the core size and composition, and on temperature at the base of the mantle. The spinel–perovskite transition may have been present only early in the history of the planet, when mantle temperatures were higher.

The mantle is covered by a crust with significantly lower density, the thickness of which is poorly constrained (it depends on the density difference between the crust and the mantle). Estimates of the mean crustal thickness vary between 28–40 km and 150 km. Large spatial variations exist anyhow, with minimal and maximal values beneath the Hellas basin and the Tharsis dome respectively (8–10 km and up to 70 km in the first model mentioned). The hemispheric asymmetry is interpreted as related to a thickening of the crust under the southern highlands. The SNC meteorites suggest that the crust is primarily basaltic in composition, although a material similar to that of the terrestrial continents or the lunar highlands may exist.

Formation and thermal evolution

Mars has accumulated from planetesimals with composition slightly different from the Earth. Models of the primitive nebula and SNC meteorites' composition suggest that the material was richer in iron and volatile elements than the Earth. Consequently, Mars' interior would have been richer in water initially. Most outgassing probably took place very early when the heat flow was very high (prior to 4.2 Gyr ago), contributing to the early atmosphere.

The isotopic and elemental ratios measured in the SNC meteorites demonstrate that core formation happened very early and was finished just after accretion 4.6 Gyr ago. This implies that accretional heating alone was large enough to melt all the material very early, like on Earth and Venus, so that Mars was initially hot and differentiated. This is in strong contrast with previous ideas concerning the thermal history of Mars, where heating was assumed to result mostly from radiogenic activity, therefore taking place after accretion and much more slowly. Consequently, the initial thermal profile was assumed to be of the present kind, i.e. temperature decreased with increasing distance from the center. Another very large amount of energy was released during core formation itself as the denser materials sank to the center, enough to heat the whole planet by about 300 K. As a consequence, mantle convection was very active during the first few 100 Myr, producing substantial

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

volcanic activity at the surface. It has been proposed that the Martian surface was completely melted in the early times and was covered by a magma ocean like the Moon. The estimated heat flow decreased from 200–400 mW m^{-2} after formation to ~150 mW m^{-2} at the end of the heavy bombardment and to ~30 mW m^{-2} today. An important topic in future space missions is actually to measure the present heat flow to better constrain the evolution models.

During differentiation, as the core formed, metallic iron reacted with water in the interior to form iron oxides, and a massive hydrogen loss probably accompanied this process. The different models and estimates disagree on the amount of water left in the mantle and surface at the end of the heavy bombardment, i.e. at the time when the oldest terrains formed. However, the equivalent of several hundred meters of water precipitated on the surface seems required to explain the erosional features observed.

After most of the heat was dissipated to space, mantle convection slowed down and volcanic activity was restricted to local areas. The internal structure of the planet was not modified, except for a slow thickening of the lithosphere (the external and most rigid shell) as the planet cooled down; the present thickness is estimated to be ~120 km. Rapid interior cooling was accompanied by a global contraction, which produced the wrinkle ridges observed in oldest terrains. The dense atmosphere which is assumed at the time when oldest terrains were exposed probably dissipated at the same period, as a consequence of the end of widespread volcanic activity. Volcanism was then restricted to discrete areas of the surface, in particular to the domes of Tharsis and Elysium.

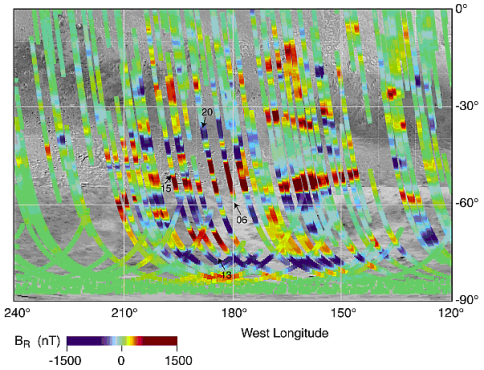


Figure 3. Map of the magnetic field in the southern hemisphere, measured by Mars Global Surveyor in 1999. The global magnetic field itself is much smaller than the Earth's, but significant magnetic anomalies were measured in the old terrains of the southern hemisphere. The strongest anomalies form stripes of alternating polarities ~1000 km long and 200 km wide, which are oriented east–west. They are thought to be fossils of a past planetary field that would have stopped early in the history of the planet. The stripes could result from occasional reversals of the magnetic field, like on Earth, suggesting lateral motions of the surface materials on early Mars. (Credit: NASA.)

Constraints on structure and evolution

After a long controversy during the 1970s and 1980s, the first detailed measurements of the magnetic field were performed by Mars Global Surveyor in 1999 (figure 3). No global field was detected, but strong magnetic anomalies were observed in the highlands of the southern hemisphere. This remnant magnetism is thought to have been induced in surface materials at a time when the planet had a global magnetic field related to a dynamo effect, and therefore had a convective core. Younger plains in the northern hemisphere show no evidence of magnetism, suggesting that the dynamo had stopped when the hemispheric dichotomy formed, and that these regions suffered major heating (either from impacts or from volcanism) that obliterated all remnant magnetic signature. In the south, magnetic anomalies display a regular pattern of stripes with alternating polarity. The stripes are ~1000 km long and 200 km wide and are oriented east–west, presenting a similarity to magnetic patterns observed on the terrestrial seafloors. It has been proposed that these anomalies are traces of successive reversals of the polarity of the early magnetic field which were trapped in surface materials drifting over the mantle. This would imply that some form of lateral plate tectonics had been active in the southern hemisphere of Mars just after surface formation and had stopped when the northern plains formed. However, the Martian surface is globally dominated by extensional tectonics rather than plate tectonics, so that this possible process, if it really occurred, stopped early in the history of the planet.

Modeling shows that the mantle convection is dominated by deep upwelling in the form of cylindrical plumes, similar to the terrestrial hot spots. In contrast, sheet-like, shallow upwelling similar to terrestrial plate tectonics cannot occur on Mars at present because of the thick lithosphere. Convection models can be run with different temperature profiles inside the planet. When most of the mantle heating comes from below (hot core), a larger number of plumes is produced (from 6 to 15 when using reasonable profiles for Mars). The presence of phase transitions of olivine appears to be an important factor in the plumes' geometry. They act to merge the plumes and reduce their number, and to concentrate their energy, and therefore models using these phase transitions give a better match of the general characteristics of the Martian surface. Some of these plumes would have a prominent surface expression in the surface topography and volcanic flux. This mechanism could explain the origin of the two major volcanic domes of Mars (Tharsis and Elysium) and of the hemispheric dichotomy. In the latter case, a very small core size (20% of the radius) is required to produce a strong asymmetry, so that the dichotomy, if related to mantle convection, should have formed very early, contemporaneously with core formation and accretion; alternatively, the dichoto-

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

my could be related to a giant impact after the end of the heavy bombardment. Concerning the two volcanic centers, they are likely to be the results of strong mantle plumes beneath the two regions. Because there are only two such regions at the surface, they probably formed early in the history of the planet when the heat flow was largest.

Biology and search for life

Although the idea that some form of life probably existed on Mars was widely accepted at the end of the 19th century, this view changed with the first studies of the atmosphere, which demonstrated that Mars is a cold planet with a rather unfavorable environment. The picture of an Earth-like Mars itself did not survive the first flyby by Mariner 4; the 22 close-up pictures transmitted by the spacecraft were acquired in the old terrains, and showed only very craterized areas, much more similar to the Moon than to the Earth. In addition, the first spectroscopic measurements at high resolution showed that the present atmosphere is too thin to allow for liquid water or to protect the surface from the ultraviolet radiation.

The situation changed again with the first complete mapping in 1971 by Mariner 9. It indicated that Mars had once been a very active planet and that water had run on the surface, possibly forming lakes or even transient oceans. In these conditions, it was possible that a complex prebiotic chemistry had developed in the distant past or even that life had originated on Mars. If so, there was a possibility that it had survived the worsening conditions, so that the search for extant life was one of the main goals of the Viking mission in 1976. Three biology experiments were embarked on the Viking landers to test whether biological activity was still detectable at the surface. All results are consistent with simple oxidation reactions in inorganic materials, although some results were considered marginally positive at first and are not completely understood today. There is a general agreement as to the results of the Viking biological experiments, that no biological process has been caught in action. These do not imply, however, that ancient Martian life could not have survived in particular places where more favorable conditions would have lasted longer, if it ever existed.

The debate was renewed in 1996 when the group of D McKay announced that it had found possible evidence of fossil life in one of the SNC meteorites, called ALH 84001 (figure 4). The possible fossils are very primitive bacteria, i.e. monocelled organisms with no nuclei, that would have lived ~3.6 Gyr ago. The meteorite was found in Antarctica in 1984 and was identified as Martian in origin because of the composition of its gaseous inclusions; however, it differs from other SNC meteorites by its age, being much older than the others (4–4.5 Gyr, versus 1.3 Gyr). It was ejected from Mars by a meteoritic impact 16 Myr ago, and fell on Earth about 13 000 yr ago.

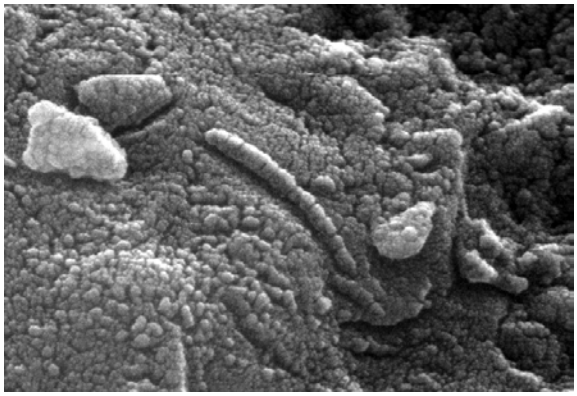


Figure 4. Electron microscope image of ALH 84001, one of the SNC meteorites. The globules of carbonates present in this essentially volcanic rock contain very small, elongated structures similar in size and shape to modern terrestrial bacteria (their sizes range from 20 to 100 nm). Although 100 times smaller than organisms that were widespread on Earth 3.5 Gyr ago, these structures have been tentatively interpreted as fossils of primitive Martian bacteria, although other teams have proposed different interpretations. (Credit: Science + NASA.)

Although essentially basaltic, it contains small globules of magnesium carbonates than were probably formed in liquid water about 3.6 Gyr ago. At this time, the Martian atmosphere was probably much thicker than today and valley networks formed in the southern highlands, suggesting abundant liquid water at the surface.

In 2001, high-resolution pictures taken by the Mars Orbiter Camera on NASA's Mars Global Surveyor spacecraft show distinct, thick layers of rock within craters and other depressions, suggesting that they may have formed in lakes or shallow seas. Such layered rock structures where there were once lakes are common on Earth. The pancake-like layers of sediment have compressed and cemented to form a rock record of the planet's history. The regions of sedimentary layers on Mars are spread out and scattered around the planet. They are most common within impact craters of Western Arabia Terra, the inter-crater plains of northern Terra Meridiani, the chasms of the Valles Marineris, and parts of the northeastern Hellas Basin rim.

D McKay's group identified three types of peculiar objects inside the carbonate grains. They consist of small structures resembling terrestrial bacteria, small mineral grains similar to those produced by bacteria on Earth and organic compounds (polycyclic aromatic hydrocarbons (PAHs)) that resemble the decay products of bacteria on Earth. All these structures could have a different origin not related to life, but taken together they were considered as strong evidence for ancient life, although not conclusive proof. Three years later, and after a large mobilization of the scientific community on this topic, the issue is very uncertain. ALH 84001 is certainly the most studied rock in the world, and all the observations

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

are on the edge of what can be done with the current techniques and instruments; in particular, the possible fossils are only ~100 nm long—this is actually from very different fossils of bacteria that lived at the same time on Earth, which were 100 times larger although similar in shape. Some of the structures observed are definitely not related to measurement artifacts, but it has not been demonstrated that the structures mentioned above are not terrestrial pollutants that entered the meteorite after its fall, although there is still no consensus as to their Martian origin. The weakest point in the demonstration is probably related to what is not observed: in particular there is no observation of bacterial growth or reproduction, and nothing looking like cell walls.

Apart from its proper interest, the study of hypothetical early Martian life would be of primordial interest to understand how life developed on Earth. The two planets were in many ways similar during their early years, so that the same processes should have acted on the two planets. There are very few traces of the early stages of life on Earth, because most of the older rocks have been recycled by plate tectonics or destroyed by erosion. Even if life has never developed on Mars, the study of prebiotic chemistry would give precious clues as to the apparition of terrestrial life.

The search for extant life is still a priority for future missions to Mars. The relatively recent discovery of life in very hostile environments on Earth, such as hot submarine springs or dry deserts in Antarctica, is actually a strong indication that life can survive in extreme conditions. Although the most favorable locations on Mars are not known, three ingredients seem required to maintain life: liquid water, protection from the solar UV flux and some source of energy. Accordingly, the most favorable locations include older volcanic vents, possible lacustrine deposits in the canyons, hydrothermal deposits, and high-latitude water-rich deposits where transient liquid water may still exist today inside the regolith. Liquid water probably also exists at the base of the permafrost, where the geothermal flux melts the ice in the ground. In all cases, living organisms would preferably dwell underground, at depths large enough to protect them from the UV flux.

Conclusion: main issues

There have been two main phases in the history of Mars. The first billion years were characterized by high internal temperatures, vigorous mantle convection, high heat flow, intense volcanic activity and perhaps a significant magnetic field. Volcanic activity probably helped to maintain a thick atmosphere of CO_2, and several hundred meters of liquid water were apparently available at the surface to shape the valley networks in the oldest terrains. Then, as accretional and differentiation heat dissipated, and as radiogenic heating decreased, the whole planet cooled down very quickly. Mantle convection slowed down, the magnetic field froze as the core probably became less convective and atmospheric recycling became inefficient, so that the total pressure became too low to sustain high temperatures and liquid water. This major change took place shortly after the end of the heavy bombardment period, 3.8 Gyr ago.

Although Mars now appears as an old, exhausted planet with only marginal geologic activity, there are numerous hints of large climate changes in a much more recent period. Variations of obliquity probably result in major climatic changes by governing the magnitude of seasonal effects. In the long run, these variations are chaotic and may have produced episodes of hot climate, when liquid water could have formed lakes or even oceans in more or less stable conditions.

NASA's Global Surveyor has revealed images which suggests the Martian climate is changing significantly, even today. The observations were made during a full Martian year, 687 Earth days and show that pits—often referred to as the 'Swiss cheese' terrain—at the southern polar ice cap of Mars have dramatically increased in diameter, indicating the material has evaporated rapidly compared to last year. Images of the south polar cap taken in 1999 were compared with images of the same locations taken in 2001, and it was discovered that pits had enlarged, mesas had shrunk, and small buttes had vanished. In all, the scarps that enclose the pits and bound the mesas and buttes retreated about 3 m in one Martian year. This rapid retreat of polar scarps can only occur if the ice is frozen carbon dioxide (dry ice). Retreat of scarps made of water ice is much slower and would not have been measurable from one Martian year to the next.

Early Mars was apparently very similar to early Earth, so that the study of the red planet can give useful hints about the Earth's evolution. In particular, the issue of prebiotic chemistry and the origin of terrestrial life could be addressed by the future Mars exploration program. Other major questions arise, especially concerning the evolution of the atmosphere and climate, the internal structure and the formation of the surface. These questions can be answered by understanding the types of rocks present at the surface and in the mantle, the absolute age of the surface regions, the distribution of volatiles, the history of volcanic activity and heat flow, the origin of the hemispheric dichotomy and Tharsis and their relationship with mantle convection, and by comparisons with other solar system bodies. These issues can be addressed by combining remote sensing observations, *in situ* analyses and laboratory studies of returned samples. Remote sensing is expected to provide a global knowledge of mineralogic and elemental composition related to morphology and a better map of the remnant magnetic field and of the gravity

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

field, so as to better constrain the geologic evolution and processes that shaped the surface; in addition, it should help identify the most promising landing sites. Concerning *in situ* measurements, a seismic network and a set of small meteorological stations are among the highest priorities, together with rock analyses and biologic experiments. Finally, only laboratory studies of Martian samples will provide the opportunity to determine the absolute ages and detailed composition. These points will constitute the bulk of the on-going Mars exploration program that will culminate with a sample return.

Bibliography

Beatty J, Petersen C and Chaikin A 1999 *The New Solar System* 4th edn (Cambridge: Cambridge University Press)

Carr M 1996 *Water on Mars* (Oxford: Oxford University Press)

Cattermole P 1992 *Mars* (London: Chapman & Hall)

Kieffer H H *et al* (eds) 1993 *Mars* (Tucson, AZ: University of Arizona Press) (review papers concerning nearly all aspects of Mars, mostly at professional level; the first chapters present a very good review of historical observations and previous syntheses up to the late 1980s)

Sheehan W 1996 *The Planet Mars: a History of Observation and Discovery* (Tucson, AZ: University of Arizona Press) (accessible to the general reader, very well documented)

Spohn T, Sohl F and Breuer D 1998 Mars *Astron. Astrophys. Rev.* **8** 181–236 (mainly focused on internal structure and bulk composition)

Weissman P, McFadden L-A and Johnson T 1999 *Encyclopedia of the Solar System* (New York: Academic)

Stéphane Erard

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mars

Series of Soviet spacecraft designed to return data about the planet Mars 1960–74. After several launch failures, Mars 1 successfully left Earth orbit in 1962 but contact was lost en route. Mars 2 and 3 arrived during the 1971 global duststorm which obscured the planet's surface, so no useful pictures were obtained. They also carried landers and the first automated planetary roving vehicles, but these either crashed or failed to operate in the duststorm conditions. Mars 4 failed to go into orbit around the planet in 1974. Some images were returned by Mars 4 and 5 orbiters. Atmospheric data were returned during descent of the Mars 6 lander, but contact was lost prior to touchdown. The Mars 7 lander missed the planet by 1300 km.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars Express

Mars orbiter proposed by the European Space Agency. Expected to launch in 2003. May carry a small British-led lander called Beagle-2 which will search for life by burrowing under rocks and analysing soil samples.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars Global Surveyor

NASA Mars orbiter. Launched November 1996, arrived at Mars in September 1997. Carries an advanced camera system to map the planet at high resolution, a thermal-emission spectrometer and two magnetometers. Discovered regional magnetic fields and mapped global topography. Prime mapping mission, delayed when a faulty solar array prevented aerobraking, began in March 1999.

Breaking News Update (30 April 2002)

NASA's Global Surveyor has revealed images which suggests the Martian climate is changing significantly, even today. The observations were made during a full Martian year, 687 Earth days and show that pits—often referred to as the 'Swiss cheese' terrain—at the southern polar ice cap of Mars have dramatically increased in diameter, indicating the material has evaporated rapidly compared to last year. Images of the south polar cap taken in 1999 were compared with images of the same locations taken in 2001, and it was discovered that pits had enlarged, mesas had shrunk, and small buttes had vanished. In all, the scarps that enclose the pits and bound the mesas and buttes retreated about 3 m in one Martian year. This rapid retreat of polar scarps can only occur if the ice is frozen carbon dioxide (dry ice). Retreat of scarps made of water ice is much slower and would not have been measurable from one Martian year to the next.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars Odyssey

Mars Odyssey is part of NASA's Mars Exploration Program, a long-term effort of robotic exploration of the red planet. 2001 Mars Odyssey was launched on 7 April 2001 and arrived at Mars on 24 October 2001, 02:30 Universal Time. Firing of the main engine slowed the spacecraft's speed and allowed it to be captured by Mars' gravity into an elliptical orbit around the planet. Aerobraking—using atmospheric drag on the spacecraft—has reduced the long, highly elliptical orbit into a shorter, 2 h circular orbit of approximately 400 km (about 250 miles) altitude for the mission's science data collection.

Mars Odyssey has begun its science mapping mission, turning its science instruments towards Mars on Monday, 18 February, 2002. Odyssey's primary science mission will continue through to July 2004. The primary mission objective is to map the amount and distribution of chemical elements and minerals that make up the Martian surface, including hydrogen (most likely in the form of water ice in the shallow subsurface of Mars). It will also record the radiation environment in low Mars orbit to determine the radiation-related risk to any future human explorers.

The three primary instruments carried by 2001 Mars Odyssey are as follows:

THEMIS (Thermal Emission Imaging System), for determining the distribution of minerals, particularly those that can only form in the presence of water;

GRS (Gamma Ray Spectrometer), for determining the presence of 20 chemical elements on the surface of Mars;

MARIE (Mars Radiation Environment Experiment), for studying the radiation environment.

During and after its science mission, the Odyssey orbiter will also support other missions in the Mars Exploration Program. It will provide the communications relay for US and international landers, including the next mission in NASA's Mars Program, the Mars Exploration Rovers to be launched in 2003. Scientists and engineers will also use Odyssey data to identify potential landing sites for future Mars missions.

One of the first images returned by the Mars Odyssey was a thermal infrared image of the Martian southern hemisphere that captured the polar carbon dioxide ice cap at a temperature of about –120 °C (–184 °F). The image, taken as part of the calibration process for the instrument, shows the nighttime temperatures of Mars, demonstrating the 'night-vision' ability of the camera system to observe Mars, even when the surface is in darkness.

The Martian Radiation Environment Experiment (MARIE) began gathering science data on 13 March, 2002, after communications were successfully re-established with the instrument. Engineers had been working since late February, 2002, trying a variety of techniques to communicate with the instrument, which stopped working in August, 2001. The results of their tests indicate the problem may be related to a memory error in the onboard software of the radiation instrument.

Initial science data from the spacecraft suggests possible identification of significant amounts of frozen water. The gamma ray spectrometer instrument suite has been used to detect hydrogen, which indicates the presence of water ice in the upper meter (three feet) of soil in a large region surrounding the planet's south pole. The amount of hydrogen detected indicates 20–50% ice by mass in the lower layer. Because rock has a greater density than ice, this amount is more than 50% water ice by volume.

Bibliography

The 2001 Mars Odyssey Home Page is at http://mars.jpl.nasa.gov/odyssey/index.html

Pam Spence

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mars Pathfinder

First of NASA's Discovery missions. Launched in December 1996 and arrived at Mars on 4 July 1997. Mainly intended as a technology demonstration mission. Used airbags to cushion the landing on Mars. The Carl Sagan Memorial station returned images of an ancient flood plain in Ares Vallis. The 10 kg Sojourner rover used an x-ray spectrometer to study the composition of rocks and travelled about 100 metres. Relayed an unprecedented 2.3 gigabits of data, including 16 500 images from the lander's camera, 550 images from the rover camera, 16 chemical analyses of rocks and soil, and 8.5 million measurements of atmospheric pressure, temperature and wind. Operations ceased after 83 days on 27 September 1997.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars Surveyor 2001

Two NASA missions to explore Mars, originally due for launch in April 2001 but now postponed. The orbiter will carry a high-resolution infrared imaging spectrometer and a gamma ray spectrometer to map surface composition and hydrogen in the soil. The lander will carry an infrared spectrometer and a small roving vehicle with a robotic arm to collect soil samples, a radiation environment experiment and a plant to make liquid oxygen from the atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars: Atmosphere

The atmosphere of MARS is much thinner than the terrestrial one. However, even the simplest visual telescopic observations show a set of atmospheric events such as seasonal exchange of material between polar caps, temporal appearance of clouds and changes of visibility of dark regions on the disk of the planet. In 1947 the prominent CO_2 bands in the near-infrared part of the Martian spectrum were identified. Later, after 1960, weak bands of H_2O, CO and O_2 were also discovered. The Doppler shift was used to separate Martian H_2O and O_2 lines from much stronger terrestrial ones. Flights of spacecraft to Mars opened new possibilities of bringing instruments close to the planet and even onto its surface. Remote sensing measurements were made initially from fly-by spacecraft (Mariner 4 in 1965, Mariner 6 and 7 in 1969) and later from orbits (Mariner 9 and Mars 3 in 1971–2, Mars 5 in 1974, Viking 1 and 2 in 1976–80, Phobos 2 in 1989, Mars Global Surveyor since 1997). *In situ* studies were fulfilled by means of descender/landers: Mars 6 in 1976, Viking 1 in 1976–82, Viking 2 in 1976–80 and Mars Pathfinder in 1997. Measurements from the orbit included TV imaging, spectrometry (by nadir and limb viewing), radio occultation (a method of sounding of planetary atmospheres using Earth based measurements of a radio signal emitted by a spacecraft when it approaches or leaves the limb of the planet; the phaseshift and attenuation due to refraction appears by the passage of radio waves through the ionosphere and troposphere), photometry and IR radiometry. *In situ* chemical composition and meteorological measurements were provided by means of instruments on the landers. All these observations not only revealed the current Martian atmosphere, but also suggested that in the long past Mars possibly had a denser atmosphere, warmer climate and open water on the surface. The history of water on Mars is one of the most interesting problems in solar system studies. It is directly connected with the possible presence of life on Mars, at least in the distant past (see EXOBIOLOGY, LIFE ON OTHER WORLDS). The book by Kieffer *et al* (1992) may be recommended as a comprehensive source for professional use.

Chemical and isotopic composition

The most abundant gaseous components of the Martian atmosphere are carbon dioxide (its volume mixing ratio—ratio of the number density of the given gas to the full number density—is 95%); nitrogen (2.7%) and argon (1.6%). The average column abundance of CO_2 is ~80 m atm, about 20 times more than in EARTH'S ATMOSPHERE. Column abundance is the amount of gas above 1 cm^2 of the surface in units of cm atm or m atm. Column abundance in cm atm equals the number of molecules in the column divided by Loschmidt's number (= 2.687×10^{19} cm^{-3}). Other identified gases are so-called minor or trace constituents: oxygen (0.13%), carbon monoxide (0.07%), water vapor (0–0.1%), ozone (0–0.2×10^{-4}%), helium (1.1×10^{-4}%), neon (2.5×10^{-4}%), krypton (0.2×10^{-4}%) and xenon (0.08×10^{-4}%). The average molecular weight is equal to 43.4 amu below the level of the homopause (~125 km on Mars). (The homopause level H_t is the height where values of two diffusion coefficients, eddy and molecular, are equal. It divides the atmosphere into two parts, the homosphere, below H_t, and the heterosphere, above H_t. Eddy diffusion dominates in the homosphere. All gases are well mixed there. The average molecular weight is constant, and its value should be used in barometric formula. In the heterosphere every gas follows its own barometric equation. Consequently, light gases like H and He are the main constituents of the outer parts of the upper atmosphere (>500 km).)

The average H_2O column abundance is ~1 cm atm. This is equivalent to a ~10 μm layer of precipitated water covering the whole surface of the planet. On recent Mars almost all water is hidden in permafrost, polar caps and hydrated minerals. The atmosphere is extremely dry. The abundance of water vapor varies with latitude, season, time of day and from place to place, mainly due to condensation/sublimation and adsorption/desorbtion processes. Latitudinal and seasonal variations according to measurements from VIKING 1 and 2 orbiters are presented in figure 1. Asymmetry between south and north polar regions is clearly visible there: during the summer the residual north polar cap sublimes and the H_2O abundance above the northern polar region reaches the maximum (~90 μm of precipitated water) in difference from the southern cap. The probable reason is that southern cap saves some quantity of CO_2 condensate even in summer time. The vertical H_2O distribution has not been thoroughly studied.

Some data were obtained in the PHOBOS mission. They showed that the mixing ratio decreases from about 0.01% at heights of 15–20 km to 3×10^{-4}% at 30–40 km. The duration of these observations was short (~1 month during northern spring) and limited by near equatorial regions.

Ozone on Mars was identified initially at the polar regions, but later it was also found at low latitudes, although with much lower abundance. These observations were made with UV spectrometers installed on board spacecraft and the ozone band near 0.27 μm was used. Recently, successful measurements have also been made from Earth in the IR band at 9.6 μm using heterodyne spectrometry, a super high-resolution spectrometric technique using a mixing of detected radiation with a laser beam at very close frequency, the radio frequency signal being measured on the output. A global average mixing ratio of 1.7×10^{-8} has been found.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

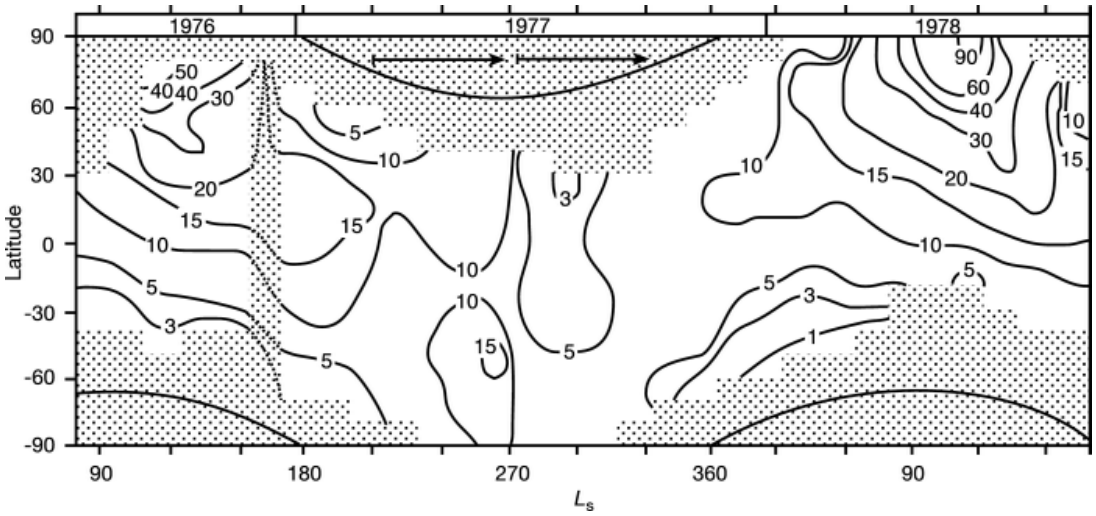


Figure 1. Water vapour cycle on Mars according to measurements from Viking orbiters: column water abundance (in μm of precipitated water) as a function of season and latitude. The absorption band at 1.38 μm was used in the observations. Seasonal longitude L_s is given along the x-axis. $L_s = 0°$ corresponds to the northern spring equinox, 90° to northern summer. These observations covered more than full Martian year (=686 terrestrials days). Local variations along areographic longitudes and daily variations also exist, but they were averaged in this presentation. Shaded areas indicate no observations, and smooth curves are the latitude poleward of which the Sun will not rise above the horizon (Jakosky and Farmer 1982).(Copyright 1982 by the American Geophysical Union.)

There are also observations from the Hubble Space Telescope in the UV.

The only *in situ* measurements of gases in the Martian atmosphere have been provided by means of mass spectrometers aboard Viking 1 and Viking 2 landers. An example of the mass spectra obtained is shown in figure 2(*a*). The most important achievement was the identification of noble gases (Ne, Ar, Kr, Xe) and their isotopes. It was found that abundance of their non-radiogenic isotopes per unit of mass is a factor of 100 less than on Earth (figure 2(*b*)). It was found also that $^{14}N/^{15}N$ and H/D ratios are lower on Mars than on Earth (twice for nitrogen and six times for hydrogen). These differences are consistent with the idea about the primordial Martian atmosphere being more dense than the recent one. The Martian isotope composition of noble gases is so specific that about dozen meteorites (SNC METEORITES) showing the same composition of captured gases are seriously believed to have originated from Mars.

Temperatures, pressures, winds

The globally/annually averaged surface temperature of Mars is ~210 K. This is only about 4 K more than Mars would have without atmosphere. So the GREENHOUSE EFFECT (a warming of the planet arising due to the lower opacity of the atmosphere for the solar radiation than for the planetary one; quantitatively it may be described by the difference $\Delta T = T_s - T_e$, where T_s is the globally averaged surface temperature, T_e is the effective temperature of the planet (= the temperature of a blackbody that would radiate the same flux as the planet)—on the Earth ΔT = 38 K, about 10 times more than on Mars) is much weaker on Mars than on Earth due to the thinner atmosphere. The minimum surface temperature is about 145 K; it is achieved in the polar regions during winter, and the atmospheric CO_2 is partially transformed there to the solid phase. The maximum temperature (in middle latitudes at summer noon) reaches ~300 K.

The globally/annually averaged pressure near the surface of Mars is ~5.6 mbar. Surface pressure varies with seasons and amplitude by ~25% (figure 3). This is mostly due to a global variation in the mass of the atmosphere resulting from seasonal condensation and sublimation processes of the carbon dioxide in polar regions and, locally, to the effect of the general circulation. The average tropospheric scale height (the vertical scale corresponding to a pressure decrease by a factor of e) on Mars is ~10 km, near to the terrestrial value (8 km). The local pressure depends drastically upon topography (see MARS: SURFACE). For instance, the highest point on the planet (the huge volcano Olympus Mons) is some 25 km above zero level, the lowest (Hellas Planitia) about 6 km below.

A typical vertical profile of the Martian atmosphere is shown in figure 4. It is colder than the terrestrial atmosphere at all heights. In general the vertical structure of both atmospheres can be divided into three main regions: low atmosphere, middle atmosphere (mesosphere) and upper atmosphere (thermosphere). The mesosphere is the region of minimum temperature. Here is the sink for the thermal energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

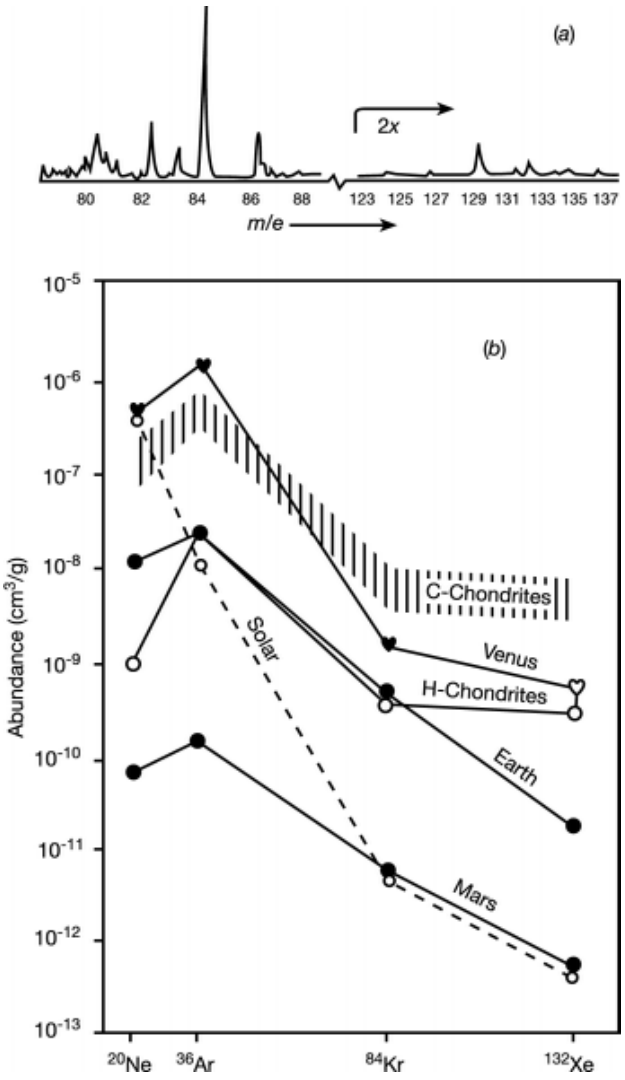


Figure 2. Noble gases on Mars. (*a*) Examples of mass spectra obtained by Viking landers, Kr (on the right) and Xe (on the left). (*b*) Comparison of abundance of non-radiogenic isotopes of noble gases in the atmosphere of Mars with the same on Earth, Venus, the Sun and in the material of chondrites (Owen and Bar-Nun 1995). Two conclusions follow from the results presented here: (a) Mars is more depleted in volatiles than the Earth and (b) the relative abundance of Kr and Xe isotopes corresponds neither to C- nor to H-chondrites.

flux coming from the surface and the troposphere (heated by solar visible and near-IR radiation) and also from the thermosphere (heated by solar extreme UV radiation). A key component of the mesospheric thermal balance is the absorption of solar near-IR radiation by CO_2. There are some important qualitative differences between Martian and terrestrial atmospheres: (*a*) the most abundant Martian atmospheric gas (CO_2) participates in condensation/sublimation processes, (*b*) diurnal temperature variations near the surface of Mars reach ~100 K, (*c*) there is no temperature maximum inside the Martian middle atmosphere (on Earth such a maximum is created by ozone), (*d*) the influence of the aerosols on atmospheric heating is stronger on Mars, (*e*) unlike on Earth, the redistribution of heat by evaporation and condensation of water is negligible on Mars. Diurnal temperature variations in the Martian atmosphere are large below a few kilometers, but relatively small higher up. The amplitude of diurnal variations in the troposphere and mesosphere depends on the aerosol content. During dust storms the upper part of the troposphere and mesosphere are warmer, and daily variations are stronger than under ordinary conditions.

Horizontal differences in temperature lead to differences in pressure, but they are smoothed by winds. Winds transfer heat together with air masses. This influences local temperature profiles. Also the dust can be lifted from the surface and change its solar heating. Condensation is another way of forming aerosols, also with a feedback effect on temperatures. Winds are strongly affected by topography. *In situ* measurements by Viking landers showed wind velocities from 1 to 10 m s^{-1} at a height of 1.6 m with prominent diurnal and seasonal variations. The daily maximum was observed often near 12 h local solar time. Maximal wind speeds were observed during periods of GDS (great dust storm—see below). Only indirect estimates are available for larger heights. They are based mainly on temperature profiles retrieved from results of IR sounding. Zonal wind velocities estimated in this way sometimes reach 20–100 m s^{-1} at heights of 5 to 40 km. The meteorological packages on Viking and MARS PATHFINDER landers recorded daily variations of surface pressure and wind well explained by atmospheric thermal tides (figure 5).

On Earth the equatorial excess of heat obtained from Sun is transferred by atmospheric motion to the poles and leaves the planet as thermal IR radiation. The Coriolis force reduces the atmospheric flow from the meridional direction to the zonal, creating a zonally symmetric or Hadley regime of circulation. The general circulation on Mars is somewhat different. First, because the atmosphere of Mars is very thin (and the thermal inertia of the surface is low, especially compared with Earth's oceans) the Martian ground is never far from from radiative equilibrium: the transport of heat by the atmosphere does not change the radiative balance as strongly as on Earth. Second, during most seasons except around the equinoxes, one hemisphere is usually warmer than the other, and the circulation tends to occur from one hemisphere to the other. In particular, around the summer solstice, mean surface temperatures are indeed highest in the summer polar region, with a monotonic decrease toward the winter pole. As the consequence, the Hadley circulation is then characterized by one unique and intense cell between summer hemisphere and winter hemisphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

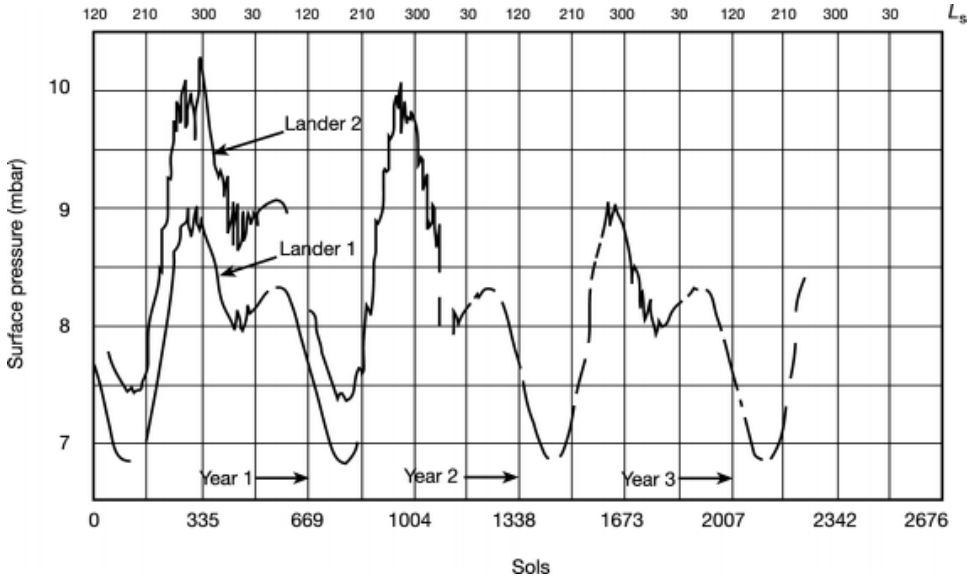


Figure 3. Seasonal changes of surface pressure on Mars. Sols (along the *x*-axis) are solar Martian days). These data were obtained from meteorological measurements by Viking landers. One of them (Viking 1) worked for about 6 Earth years and 4 months.

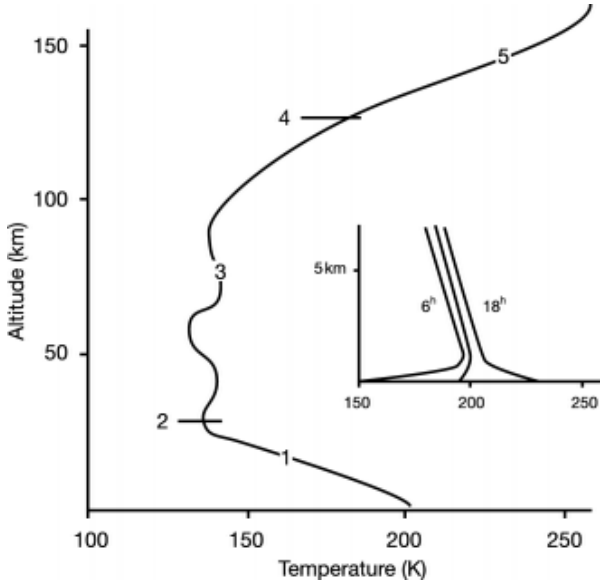


Figure 4. Schematic presentation of the vertical temperature profile in the Martian atmosphere, based mainly on *in situ* measurements during descent of Mars 6, Viking 1, Viking 2 and Pathfinder landing probes. 1 is the troposphere, a lower part of the atmosphere with large negative temperature gradient, 2 is the tropopause, 3 is the mesosphere, 4 is the homopause and 5 is the upper atmosphere. The inset illustrates the character of daily variations of the temperature profile in the boundary layer (the lowest part of the atmosphere where its motions are influenced by interaction with the surface; the rest of the atmosphere above this layer is treated as the 'free atmosphere'), predicted theoretically.

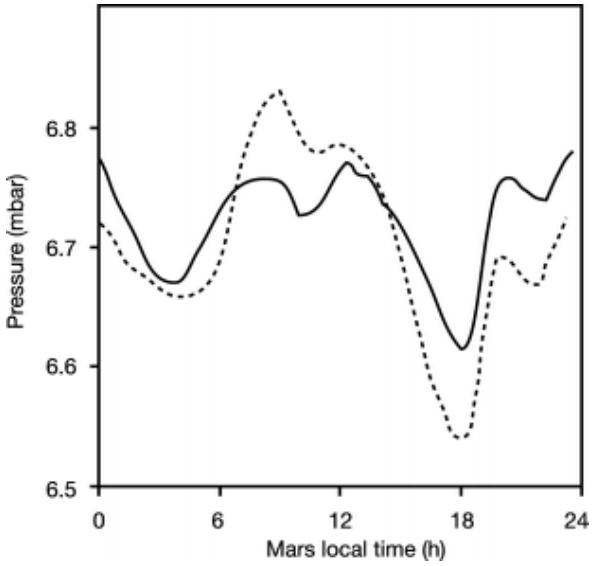


Figure 5. Example of surface pressure measurements by the Pathfinder lander (1997): diurnal pressure cycles for sols 9 and 19, illustrating the observed day-to-day changes in the diurnal pressure cycle and allowing for details of the daily pressure variation. The general shape of the curves is explained by superposition of diurnal and semidiurnal waves occurring due to atmospheric thermal tides.

In Hadley circulation lines of flow go around the whole planet almost along parallels with a much smaller meridional component, at least at low levels. Pure zonally symmetric circulation may transfer only a limited quantity of heat, and is replaced or added to by wave circulation if this limit is reached. Lines of flow in a wave regime oscillate a few times between high and low

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

latitudes going around the planet ('transient baroclinic waves'), wind speeds are higher and more variable and a large meridional component may appear. General circulation models (GCM) developed for Mars show that the observed wind pattern during most of the Martian year is more consistent with a zonally symmetric regime than is the terrestrial atmospheric circulation. (General circulation models are computational tools for the prediction of 3D wind patterns on the planet. Inputs are the thermal properties of the atmosphere and surface, incoming fluxes, surface topography, etc. Using hydrodynamic equations the model provides a time-dependent picture of atmospheric motion. Space and time resolution is restricted of course by computational possibilities.) However, travelling atmospheric waves are known to be prominent and regular in northern mid-latitudes during the fall and the winter. Variabilities in suspended dust content produce time variability (interannual, seasonal, local, diurnal) in atmospheric circulation and also in thermal structure.

Aerosols

There are two main types of Martian aerosol: dust and water ice particles. A mixed composition is also possible: a layer of a condensate around a mineral nucleus. Chemical composition of the dust particles should be the same as that of the surface regolith, although the size distribution can be different. At high altitudes and at high latitudes in winter solid CO_2 particles may appear due to local decreases of the atmospheric temperature to the condensation level.

Mineral aerosol (dust) formations can be divided into two main subclasses: (*a*) constant haze, (*b*) dust storm clouds. There is no sharp boundary between classes. Dust storms appear as a local rise of constant haze opacity and disappear leaving constant haze.

The average optical depth (=the natural logarithm of transmission) of the constant haze varies between 0.1 and 0.5 with a characteristic time scale of a few years. There is also a seasonal trend with the appearance of much denser cloud cover with an optical depth up to 6 during great dust storms (figure 6). The atmosphere is when it contains more dust and colder when the dust dissipates. The number density of the constant haze dust particles near the surface is about n = 1–2 cm^{-3}. Their single scattering ALBEDO (the ratio of omni-directional scattering flux to the incident flux of parallel rays (like the spherical albedo of a planet)) is ~0.86 at wavelengths between 0.55 and 0.6 μm and decreases to 0.63 near 0.32 μm. Corresponding values of the imaginary part of the refractive index m are ~0.005 and 0.015. They are compatible with the presence of a few per cent of materials like Fe_2O_3 in silicate particles. Typical values of mean particle radius in the constant haze are 1.5–2.5 μm. Observations of the vertical distribution of dust in the constant haze were made in experiments on the Phobos mission. They showed that the aerosol scale height is 10–11 km below 15–20 km, almost the same as the gaseous scale height, meaning that dust is well mixed in the lower atmosphere.

A few subclasses of condensate aerosol formations were observed: (*a*) ice haze near the surface, (*b*) high-altitude layers, (*c*) mountain clouds, (*d*) polar hood. Ice haze appears near the surface at night due to a significant decrease in temperature within a narrow (of the order of 1 km) layer near the surface. The typical optical thickness is of the order of 0.1 and depends on the local abundance of water vapor. In the morning the ice particles evaporate and their net effect is clearly visible as an a.m.–p.m. opacity difference. Several seasonal peaks of this difference were observed.

Condensation of water vapour at high altitudes results in the appearance of thin (few km) layers with an optical depth between 0.005 and 0.05. Such layers were observed at heights between 25 and 75 km in the visible, ultraviolet and thermal IR wavelengths. There are also a lot of TV observations of 'mountain' clouds, mainly around huge Tharsis volcanoes, but very few determinations of their optical depth.

Great dust storms

During the northern winter two different dynamic regimes have been observed in various years: in the first case great dust storms (GDS) cover nearly all the planet; in the second regime no great storm occurs, but strong winds produce dust lift in confined regions. Maximal wind speeds are observed during periods of GDS. A lot of dust fills the Martian atmosphere at low and moderate latitudes when a GDS happens. GDS arise when the planet is near perihelion (summer in the southern hemisphere) and consist of one, two or even three peaks covering up to 150 days. Sometimes local dust storms are observed on Mars. There is no terrestrial analogue of Martian GDS, but local dust storms occur on Earth also (in deserts). So, mechanisms of dust storm generation on both planets may have common details. Their study on Mars can be useful for understanding such events on Earth.

Mechanisms of development and decay of great dust storms are not well understood. However, on the qualitative level, the sequence of events may be described as follows. There is a threshold wind velocity that is necessary to lift up small particles from the surface. This threshold is reached only within some narrow limits of seasonal and local conditions. Laboratory experiments show that it is not the smallest but the relatively big (of the order of 100 μm) particles that are lifted first. However, they themselves fall very soon and disrupt the surface of much smaller particles. Soon the general number of lifted particles rises to a level sufficient to cause substantial obscuration, and to cool the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

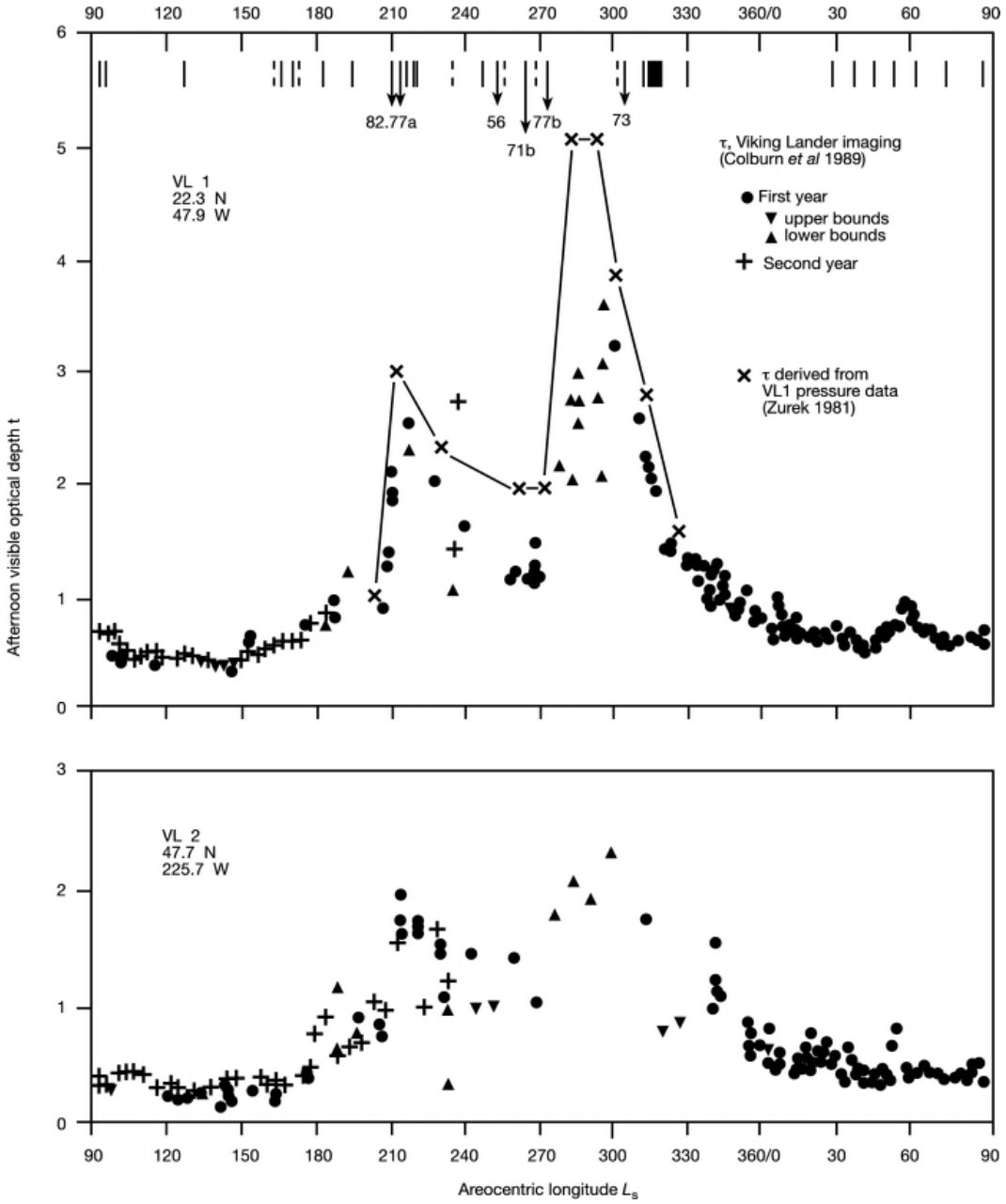


Figure 6. Seasonal trend of the optical depth of dust on Mars according to Viking 1 and 2 data. Two wide spikes correspond to great dust storms. The initial appearence or detection of observed regional dust clouds, hazes or obscurations (vertical lines) and of planet-encircling dust sorms (arrows) are indicated at the top of the upper panel.

atmosphere near the surface and heat it at some kilometers. Local wind speeds rise on the boundaries of this 'storm nucleus' and they expand rapidly. In a few days this originally small nucleus transforms to planetary encircling dust cloud cover. Then the temperature differences disappear and nothing supports higher wind speeds. The decay stage comes: particles lifted to a height of some tens of kilometers set down gradually, separating by sizes. Particles of 10 μm reach the surface in a few days and 1 μm particles in a few months.

Upper atmosphere and ionosphere

The temperature of the uppermost levels of the Martian atmosphere ('thermosphere', above 120 km) depends on the solar activity; its average value is about 210 K, minimum 135 K and maximum 310 K. The reason for this variability is that the thermosphere is heated by extreme solar UV radiation which is strong during solar maximum and much weaker near the minimum of the solar activity cycle. The chemical composition of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

upper atmosphere of Mars above a height of about 125 km (homopause) varies with height due to separation by molecular weight of its different components. However, CO_2 dominates up to a height of about 200 km at the average thermospheric temperature. Atomic oxygen (O) dominates between 200 and 400 km and atomic hydrogen (H) above 400 km. As a result the outer atmosphere of Mars consists of hydrogen, and sometimes even the name 'hydrogen corona' is used. It was observed in the Lα line up to 20 000 km from the planet. Earth and Venus have the same sort of H corona but less expanded due to the higher gravitational field (see VENUS: ATMOSPHERE).

The height of the exobase is about 200 km. Hydrogen atoms continuously leave the Martian atmosphere above these levels; a flux of about 3×10^{11} atoms/m^2 s^{-1} has been estimated. This corresponds to the loss of 1 m of water layer during 4 billion years of Martian geological history.

Ions appear in the upper atmosphere due to absorption of solar extreme UV radiation. Primary CO_2 ions are replaced by O_2 ions due to a set of reactions, and O_2 ions dominate in the Martian ionosphere. The daily peak electronic number density n_e is about 10^5 cm^{-3} at heights from 110 to 145 km, depending on the Sun zenith angle. At night time a peak $n_e = 5 \times 10^3$ cm^{-3} at a height of 110 km was found. The intrinsic magnetic field of Mars is very weak and does not protect the ionosphere from the impact of the solar wind. It influences the shape of the ionosphere profile and creates some induced magnetic field which overlaps the intrinsic one, and substantially masking it.

The long-term atmosphere/surface interaction; origin and evolution of the atmosphere

Mars is more depleted in volatiles than is Earth. However, the terrestrial atmosphere contains only a small part of the total inventory of volatiles. Almost all the water is in the liquid phase and carbon dioxide is hidden in carbonates. So, we may suppose that the full amount of volatiles appearing on Mars during its geological history is much more than in the recent gaseous atmosphere. There are no open reservoirs of liquid water on recent Mars, so POLAR CAPS and permafrost play the same role as the terrestrial oceans. Only rough estimates are possible. They lead to values of an equivalent water layer of the order of 100 m. Seasonal polar caps contain up to about one-third of the atmospheric carbon dioxide. However, much more may be buried in the residual permanent southern cap. Also Martian soil (working as a cryogenic pump) may adsorb a substantial amount of water and even carbon dioxide. Chemical weathering may also bound part some of the water in minerals. If open water reservoirs existed in the distant geological past, then the process of transformation of silicates to carbonates could be effective. So, the search of carbonates on Mars may give an important key to the history of the Martian volatiles.

The recent surface pressure on Mars is very close to the triple point of water. If carbon dioxide slowly enters the atmosphere due to volcanic outgassing, then this puzzling coincidence may be explained by the tempory appearance of liquid water, when the surface pressure exceeds the triple point value. It switches on the process of silicate to carbonate transformation, rapidly removing the excess of carbon dioxide.

In general, there are three types of processes that were (or can be) involved in the formation and the evolution of the Martian atmosphere: (A) supplying volatile (sources), (B) removing them (sinks), (C) reversible processes that work in both directions depending on conditions. Type A processes are: (1) trapping of gases bound in planetesimals from which Mars was composed during accretion; (2) fast release of these gases by impacts immediately during accretion (a catastrophic outgassing); (3) slow outgassing of planetary interiors over billions of years; (4) trapping of gases from the proto-planetary nebula directly; (5) cometary impacts; (6) solar wind implantation. Process (1) probably worked but the destiny of trapped gases is under question. They could be released by processes (2) or (3) predominantly. Some years ago experts preferred process (3), later process (2). Direct trapping from the nebula was never recognized as really important. But processes (1)–(4) all suppose that Martian volatiles came originally from the primary nebula. However, a very different scenario was also possible: the whole primordial atmosphere of Mars was removed by a huge single impact and then process (5) created a new atmosphere. Owen and Bar-Nun (1995) proposed some arguments for this based on the composition of SNC meteorites. The importance of process (6) is not clear, maybe it could be checked by measurements of Ne isotopic ratios.

Type B processes are: (1) impactory blow-off; (2) thermal escape (H and He) (Escape is process of the loss of lighter and faster atoms from the upper atmosphere. Such atoms may go out if their velocity is sufficient to enable them to leave the gravitational field of the planet and they are at a height where the probability of collision with other atoms is small (above the exobase, the height where free path length equals to the scale height); (3) non-thermal escape (oxygen); (4) solar wind blow-off; (5) chemical weathering. It is difficult to observe a flux of escaping materials directly. However, such a flux was measured for oxygen ions during the Phobos mission. An important example of chemical weathering is the binding of water in Martian minerals. Another may be silicate to carbonate transformation and formation of sedimentary rocks when and if open water existed on Mars.

Type C processes are: (1) polar cap formation; (2) permafrost formation; (3) atmosphere/soil exchange.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Water is involved in all three, carbon dioxide in (1). These processes are influenced by different sorts of insolation changes including seasonal and long-term (as solar constant trend) changes, effects of oscillations of equatorial inclination and orbital evolution. Atmosphere/soil exchange acts even within a daily cycle.

The Martian atmosphere is breathing even now, seasonally and from year to year. Much greater changes in the distant past may be anticipated. There are two sorts of evidence for this: (a) clear traces of liquid water activity on the surface and (b) isotopic ratios (N and Ar) showing that a significant part of the atmosphere was lost during evolution of the planet. At least two geological epochs of fluvial activity are supposed: earlier with the formation of valley systems and later with the formation of outflow channels. The first epoch was definitely warmer than now, the second maybe not. Carr (1996) estimates that the average thickness of the water layer may have been 500–1000 m on ancient Mars during the hypothetical 'Mars paradise' period.

Bibliography

Carr M H 1996 *Water on Mars* (New York: Oxford University Press)

Jakosky B M and Farmer C B 1982 The seasonal and global behavior of water vapor in the Mars atmosphere: complete global results of the Viking atmospheric water detector experiment *J. Geophys. Res.* **87** 2999–3019

Kieffer H H, Jakosky, B M, Snyder C W, and Matthews M S (ed) 1992 *Mars* (Tucson, AZ: The University of Arizona Press)

Owen T and Bar-Nun A 1995 Comets, impacts and atmospheres *Icarus* 215–26

Schofield J T, Barnes J R, Crisp D, Haberle R M, Larsen S, Magalhaes J A, Murphy J R, Seiff A and Wilson G 1997 The Mars Pathfinder atmospheric structure investigation/meteorology (ASI/MET) experiment *Science* **278** 1752–8

Vassili Moroz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mars: Climate History

Unravelling the climate history of Mars is one of the main challenges in understanding the planet's evolution. There is substantial evidence for change in the climate and the inventory of volatiles through time. This evidence points to an early environment in which water was either more stable or more abundant at the surface than it is today and to processes of supply of volatiles to the atmosphere (see MARS: ATMOSPHERE) and their removal from it that are able to explain this shift in climate. Determining the forces behind the changing climate, particularly the relationship between the atmosphere, water at the surface and water in the crust, is important for understanding the planet as a whole. In addition, the potential for life to have existed in the past, or even today, is strongly governed by the occurrence of liquid water and its history over time.

The climate system of Mars is inherently complex, involving physical and chemical processes within the deep interior, the CRUST, the surface, the atmosphere and the upper atmosphere and interactions with the solar wind (figure 1). Our goal is to integrate observations and measurements pertinent to each of these areas into a coherent view of Martian climate and its evolution. We will focus on recent results from the MARS GLOBAL SURVEYOR mission and from ongoing analyses of the Martian meteorites and will emphasize observations that provide the strongest constraints rather than describing what is merely plausible.

The present-day atmosphere and climate

The Martian atmosphere consists predominantly of CO_2, with a total atmospheric pressure averaging about 6 mbar (6 hPa), about 0.6% of the Earth's atmospheric pressure. The atmosphere also contains N_2 and Ar at a level of 1–3%, lesser amounts of H_2O, CO and O_2 and additional gases in trace amounts (Owen 1992).

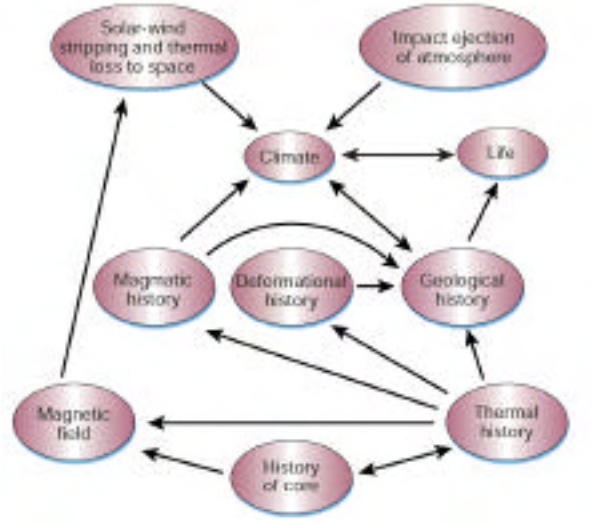


Figure 1. Schematic diagram showing the interconnected nature of the Martian climate system and the relationship between climate and various processes from the deep interior to the upper atmosphere. Arrows indicate how one region affects another, with the arrow pointing towards the processes or area being affected; two-headed arrows indicate mutual interactions.

Water vapour is present in the atmosphere, with a typical partial pressure of 10^{-3} mbar. This corresponds to about 10^{-3} g H_2O residing above each square centimetre, equivalent to a condensed layer about 10 m thick if it all precipitated onto the surface; the water content of the Earth's atmosphere, in contrast, is about 10^4 times greater. The atmospheric water content varies in a manner that is consistent with seasonal supply and removal from the polar caps, exchange with water adsorbed in the near-surface regolith and global transport by means of winds (Jakosky and Haberle 1992).

Temperatures average about 220 K globally, well below the melting temperature of ice and lower than the eutectic freezing temperature of most salt-rich brines. Even though surface temperatures rise above 273 K over large regions near the equator and near noon, liquid water still would not be stable. It could exist as a transient phase, but would quickly evaporate into the relatively dry atmosphere and eventually freeze out at the colder high latitudes (Jakosky and Haberle 1992).

Mars' north polar region is covered with CO_2 ice during winter, which sublimes away and leaves a residual summertime deposit of water ice. The south cap is covered year round by CO_2 frost but almost certainly has water ice mixed in or beneath it (Jakosky and Haberle 1992). The axial obliquity, the tilt of the polar axis with respect to the normal to the orbital plane (currently 25.2°), varies on timescales of 10^5–10^6 yr. On longer timescales, it is chaotic and may have been as low as 0° or as high as 60° in the past few million years. At high obliquity, polar summertime temperatures may increase markedly and substantial amounts of water ice may sublimate into the atmosphere. Exchange of this water between the north and south polar caps, modulated by the changing eccentricity and season of perihelion, is probably responsible for the formation and evolution of layered deposits in the polar regions (Jakosky and Haberle 1992).

Nature of the earliest atmosphere and climate

The observable surface record of Mars' geological history spans 4 billion years (Gyr). The oldest, most heavily cratered surfaces are thought to be about 4.0 Gyr old and the youngest are possibly less than 100 million years (Myr) old (Hartmann and Neukum in press; Hartmann *et al* 1981). Evidence pertaining to the ancient climate is inferred from processes that shaped the surface during the Noachian epoch, which ended when the cratering rate declined dramatically between 3.8 and 3.5 Gyr ago (Hartmann and Neukum in press). (Figure 2 summarizes the history and timing of processes involved in Martian

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

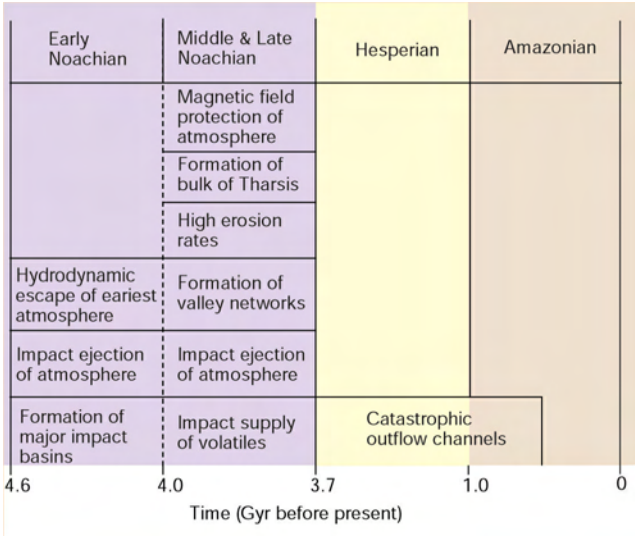


Figure 2. Schematic diagram showing the time history of the Martian volatile system. The 4.6 Gyr history of Mars is divided up into the major geological epochs that have been identified and processes that occur within each epoch are indicated. Note that the timeline is not a linear scale. Some climate-related processes are not shown (including recent polar-cap behaviour and the occurrence of possible crater lakes). Absolute ages marking the boundary between epochs are estimated based on the recent revision of cratering chronology (Hartmann and Neukum in press).

climate evolution.) Although atmospheric gas undoubtedly was present before 4.0 Gyr and was removed in part from the atmosphere by various processes, we focus here on processes that postdated the onset of the visible geological record.

The ancient Martian surfaces contain geological features that indicate that the early climate was different. Dendritic networks of valleys seem similar to those formed by run-off of surface water on the Earth, although the areal density of tributaries is typically lower. There is debate about the relative roles of surface run-off, sapping by release of subsurface water and discharge of water from hydrothermal systems in forming the valleys. However, there is general agreement that water must have flowed at the surface in order to form these features and that their dendritic character and typically V-shaped cross-section requires a gradual rather than catastrophic formation process (Carr 1996).

Erosion rates in general were substantially greater during the Noachian. This can be seen easily on ancient impact craters. The largest craters and basins are severely degraded; ejecta deposits, crater rims and central peaks have all been removed and a paucity of craters smaller than about 15 km diameter suggests that they have been removed in their entirety. Some partially degraded craters have fluting and scalloping along the interior rim that suggests erosion by flowing surface water. Estimated erosion rates were more than 1000 times larger than in subsequent epochs and approach values appropriate for drier regions on Earth (Carr 1996).

Additional evidence for a shift to a colder, drier climate at the end of the Noachian is provided by the initiation of U-shaped valley forms at the downstream ends of V-shaped valleys, suggesting an evolution of valley-forming erosional mechanisms from water related to ice related. This period of U-shaped valley formation and valley network evolution in general, ended rapidly at the end of the Noachian or early in the Hesperian.

Together, these features strongly suggest that liquid water was present during the Noachian and flowed over the surface and that the climate must have been such that water was either more stable or more abundant at the surface than it is at present. Such a climate is thought to have required warmer surface temperatures produced by substantial greenhouse warming (see GREENHOUSE EFFECT). Although the climate must have been 'warmer' and 'wetter', there is no consensus as to what temperatures would have been required, what the greenhouse gas would have been or how much water would have to have been at the surface or in the atmosphere (Squyres and Kasting 1994). If CO_2 were the greenhouse gas, up to several bars pressure would have been required to produce the necessary warming. A warmer climate is indicated even if valley network formation involved subsurface hydrothermal processes, as flow at the surface and gradual erosion require a warmer climate independent of the original source of the liquid water.

Connections between ancient climate and early geology

Other events also were taking place during the Noachian period that would have affected climate. The planet was being bombarded by impacting ASTEROIDS or PLANETESIMALS, it was creating its main division into a southern highlands and a northern lowlands (which is reflected today as the ancient and younger terrains respectively) and volcanism was forming the major Tharsis province on which sit many of the large volcanoes.

The formation of Tharsis was recognized only recently as occurring largely in the Noachian (Phillips *et al* 2001). The heavy loading of the lithosphere, due to the dominantly magmatic formation of the Tharsis rise, resulted in a global warping of the surface (Phillips *et al* 2001). Many of the ancient valley networks are seen to preferentially follow the slopes that resulted from the formation of Tharsis or, in at least one location (Margaritifer Sinus), are of the same age as those that do. This indicates that emplacement of Tharsis must have been nearly complete while the largely Noachian valley networks were still forming. Although much of the surface in Tharsis is sparsely cratered and therefore relatively young, indicating that resurfacing of Tharsis continued throughout Martian history, the bulk of the volume of Tharsis must be old (Phillips *et al* 2001).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The volcanic magma from which Tharsis formed probably contained substantial quantities of both water and CO_2 that would have been released, providing input of gases to the atmosphere and possibly contributing to an early, thicker atmosphere. Geochemical analysis of the Martian METEORITES, for example, suggests a water content of as much as 1.8% by weight. Thus, the timing of valley network formation may be more than coincidental. Tharsis volcanism may have supplied gases that helped maintain a climate conducive to weathering and erosion and the cessation of volcanism may have allowed other processes to begin removal of much of the atmosphere (see below) (Phillips *et al* 2001).

There is another compelling connection between Tharsis and the history of water. The lithospheric deformation due to the weight of Tharsis created a depression or trough encircling it at a radius of about 5000 km (figure 3(*a*)). Most of the large-scale geological features related to surface water are concentrated in this trough (Phillips *et al* 2001). This includes the large-scale drainages that flow northward from the Argyre basin (at 50° S latitude) and the catastrophic flood channels that emanate from the eastern end of Valles Marineris and flow north into the Chryse basin and into the northern lowlands. Argyre, for example, seems to have been filled with water, which subsequently overflowed the rim and flowed northward. Although the catastrophic flooding occurred episodically throughout Martian history, the source and flow regions concentrate in the Tharsis trough. This relationship suggests a long-lasting connection between the geological and geophysical history of Mars and the release, availability and geological effects of liquid water. However, some anomalies exist, such as the paucity of similar features in the western branch of the circum-Tharsis trough and the ill-defined nature of processes, other than simple structural control of water pathways, that would provide a causal connection between the trough and the release of crustal water.

Finally, we note an apparent correlation between the regions on which crustal remanent magnetic fields are imprinted and areas on which valley networks are identified (figure 4). The magnetic anomalies presumably retain an imprint of a global-scale intrinsic magnetic field from an early time in Martian history (Acuña *et al* 1999). Exceptions to the correlation include the large-scale magnetic anomalies occurring in discrete bands in the high southern latitudes, where valley networks may have never formed because of climate constraints and Margaritifer Sinus, where stripping of the surface may have removed the valleys without removing enough material to eliminate the magnetic anomalies. If genuine, this correlation may reflect a coincidence in timing of the formation of both the valleys and the magnetic anomalies or in the geographic location of subsequent

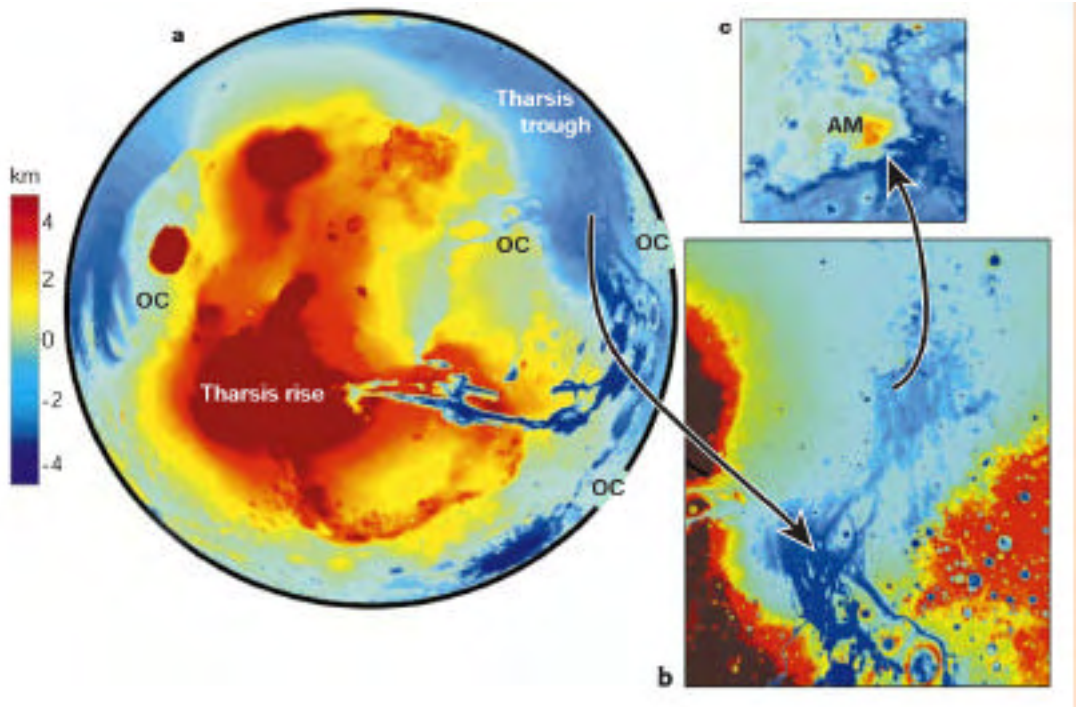


Figure 3. Topography (Smith *et al* 1999) of Tharsis rise and Tharsis trough (Phillips *et al* 2001). (*a*) Topography, which has been saturated at 5 km, with pole-to-pole slope (Smith *et al* 1999) removed. Image is centred on 260° E longitude and view is from 10° north of equator. Outflow channel locations are marked by 'OC'. (*b*) Channel detail in Chryse and Acidalia Planitiae (0°–60° N, 300°–0° E, Mercator projection; elevation range is –3.9 to 0 km). (*c*) The inlier Acidalia Mensa ('AM'), whose southern and eastern boundaries are channels (Mercator projection; elevation range is –3.0 to –1.5 km). Eastern channel extends northwards of 50° N.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

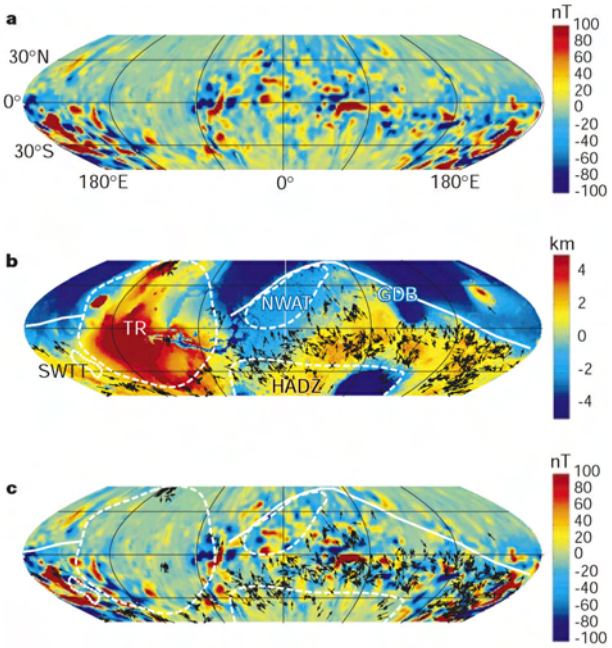


Figure 4. Comparison of Martian magnetic field, topography and valley network locations. (*a*) Radial component of the Martian magnetic field at an altitude of 200 km (after Purucker *et al* 2000). Values saturated at 100 nT. Sinusoidal projection to 47.5° latitude. (*b*) Topographic map of Mars (after Smith *et al* 1999), saturated at 5 km elevation and locations of valley network systems. Database is limited to latitudes less than 47.5°, which is the approximate limit of valley network distribution. Features indicated are the Tharsis rise (TR), global dichotomy boundary (GDB), northwestern Arabia Terra (NWAT), southwestern Tharsis trough (SWTT) and Hellas–Argyre disruption zone (HADZ). TR, NWAT, SWTT and the area north of GDB are regions of valley network exclusion by erosion and/or deposition and cannot be judged for the level of correlation of valley networks with anomalies in the remanent magnetic field. Likewise, TR, HADZ and the area north of GDB are regions of magnetic field exclusion by the presence of post-main-field crustal ages or crustal disruption (Acuña *et al* 1999). (*c*) Valley networks and features in (*b*) superposed on magnetic anomaly map.

alteration processes. Alternatively, it may involve a cause-and-effect relationship, in which valleys could represent the surface manifestation of hydrothermal systems driven by subsurface volcanism that thermally induced the creation of the localized remanent fields. However, as neither the mechanism of formation of the magnetic anomalies nor that of their removal (if they ever were distributed globally) is understood, such connections are speculative at best.

Evolution of the Noachian atmosphere

What processes might have caused the inferred changes in Noachian climate, what sinks were available for atmospheric gases during this epoch and can we infer the degree to which each process acted? These are especially important issues, as the output of the Sun was some 30% less 4 Gyr ago than it is today, making the maintenance of an early warmer climate more difficult.

Impact of large asteroids through the Noachian epoch would have ejected gas in the atmosphere to space. Although inherently capable of removing large amounts of atmosphere, only those impacts occurring since the time of the onset of the geological record would have resulted in climate change that can be inferred from that record (Brain and Jakosky 1998). We know how many impacts occurred, based on the number of large craters and impact basins seen on the ancient terrain and can readily extrapolate to impacts on areas subsequently buried. Combined, impacts are likely to have ejected about 50–90% of the atmosphere present in the early epoch (Brain and Jakosky 1998). This fraction probably is an upper limit, as some volatiles (such as water) might not have resided exclusively in the atmosphere at this time or might have been outgassed late in the epoch.

Recent measurements of surface elevation have allowed identification of previously unknown impact basins buried below a relatively thin veneer in both the southern highlands and the northern plains. The buried craters in the northern plains have similar abundance to the craters in the southern highlands and justify the extrapolation of southern-hemisphere craters to a global inventory. The buried southern-highlands craters predate the geological evidence pertaining to the early climate, so those impacts would not have contributed to the changes in climate and should not be counted in summing up atmospheric loss.

Impacts of volatile-rich objects can supply new volatiles to Mars as well as remove them from the atmosphere. In particular, the analysis of isotope ratios of heavy noble gases in the atmospheres of Earth and Mars suggests that much of the planets' volatiles could have been supplied by comets. However, the relative roles of supply and removal of volatiles by impacts are unknown.

The impingement of the solar wind can strip gases directly from the upper atmosphere. Both atom-on-atom collisions that eject individual molecules (pick-up-ion sputtering) and hydrodynamic collisions that could strip off large volumes of gas en masse would have been important. However, theoretical models of these processes are exceedingly uncertain even at the present epoch and extrapolating to past epochs when the solar wind and solar ultraviolet radiation were more intense compounds the uncertainties. However, two observations of the present-day atmosphere indicate that these processes were significant. First, *in situ* measurements of the energy spectrum of ionospheric electrons distinguish between ionization of gases in the upper atmosphere and those from the incoming solar wind. The morphology of discrete regions of solar-wind versus planetary electrons requires that large masses of upper atmosphere are being

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 1. Martian isotope ratios and atmosphere loss*

Isotope ratio	Measured value†	Amount lost to space (%)‡
D / H	5	–60–74
$^{38}Ar/^{36}Ar$	1.3	–50–90
$^{13}C/^{12}C$	1.05–1.07	–50–90
$^{15}N/^{14}N$	1.7	–90
$^{18}O/^{16}O$	1.025	–25–50

*Values taken from Leshin 2000, Krasnopolsky 2000, Jakosky and Leshin 2001, Romanek *et al* 1994, Watson *et al* 1997 and Jalosky and Jones 1997 and references therein.

†Value estimated, observed or derived for Martian atmosphere relative to terrestrial.

‡Calculated assuming Rayleigh fractionation. D / H range includes uncertainty in escape processes. Other ranges are based on uncertain timing of outgassing relative to escape.

stripped off en masse. Second, isotopic measurements indicate that such stripping has occurred and has been significant. The lighter isotopes of each atom are enriched relative to the heavier ones at the top of the atmosphere by diffusive separation (for example, ^{12}C relative to ^{13}C, ^{14}N to ^{15}N or ^{36}Ar to ^{38}Ar) and will be preferentially removed to space. Consequently, the gas remaining in the atmosphere becomes enriched in the heavier isotope. This enrichment pattern is observed consistently in the modern atmosphere. For example, the ratio $^{38}Ar/^{36}Ar$ is 30% greater on Mars than elsewhere in the solar system and nitrogen and carbon also show enrichments (table 1). This enrichment indicates loss of between about 50% and 90% of the atmospheric species to space, with the actual loss probably tending towards the higher values (Jakosky *et al* 1994).

In addition, compared with terrestrial values, Martian water shows a five-fold enrichment of D relative to H (see Table 1). The lighter H atom can escape thermally to space more readily than D, so the considerable enrichment of observed D requires the loss of substantial quantities of hydrogen to space. The hydrogen comes from water, with the oxygen from water probably also being lost to space. The inferred water loss depends on the relative supply rates of H and D to the upper atmosphere, their relative escape rates and the initial D / H ratio on Mars. Recent analysis of Martian meteorites suggests that the initial D / H value on Mars was twice the terrestrial value (Leshin 2000); in addition, there has been a spectroscopic detection of upper-atmospheric D and a revision of atmospheric reaction rates. Although the initial interpretation was that about 90% of Martian water had been lost to space, the current understanding suggests loss of about two-thirds of the water.

The observed isotopic enrichments indicate the fraction of each species that has been lost but not the total amount lost. Atmospheric gases exchange with the regolith and polar caps, so the fraction lost actually represents that lost from the combined atmospheric plus non-atmospheric reservoir. In addition, different portions of the polar ice, for example, may exchange with the atmosphere on different timescales and outgassing of juvenile gas at later times will affect the isotope ratios. The absolute loss rates to space are uncertain today and may have varied through time, so although we can be confident that loss of substantial volatiles to space has occurred, we are uncertain as to how much has been lost, where it came from or exactly how much has been left behind. (Although the losses of H and O are related and one would expect the isotopic fractionation of D / H to be related to that of $^{18}O/^{16}O$, oxygen has an additional reservoir with which it can exchange — minerals in the crust — so that the connection is likely to be complex.)

Whereas stripping by the solar wind of upper-atmospheric gases will change the isotope ratios, removal by impact will not. Impacts remove the gas from the entire atmospheric column and thereby remove all isotopes with equal efficiency. Thus, the fractions of gas lost by these two processes add together. For example, if each indicates loss of 90%, these represent a different 90% and they sum up to a net loss of 99% of the atmospheric gas. In addition, the two loss processes probably operated at different times, with impact erosion probably being most important early in the Noachian when impact rates were highest and solar-wind stripping occurring late in the Noachian or into the Hesperian.

CO_2 can be removed from the atmosphere in the presence of liquid water to form carbonate minerals on the surface or in the subsurface. This process occurs on Earth, where gaseous CO_2 dissolves in the oceans, combines with calcium ions weathered from the continents and forms deposits of calcite in limestone. If SO_2 was an early greenhouse gas, it might have formed crustal sulphate minerals. Carbonates have been found in the Martian meteorites and have been shown to be indigenous to Mars. However, carbonate and sulphate minerals have not been detected in sufficient quantities that, if released in gaseous form, they could provide enough greenhouse warming to explain the early environment (Christensen *et al* in press). The CO_2 from a several-bar atmosphere would form a global equivalent layer of carbonates perhaps a hundred metres thick; were this distributed throughout the entire volume of the crust, perhaps carried there by circulating groundwater, it might not be detectable at the surface. Thus, the extent to which carbonate minerals could be a sink for gases from an early atmosphere remains uncertain.

However, processes do exist that together can account for loss of an early thicker atmosphere. Theoretical models of the timing of volatile loss suggest that these processes can account quantitatively for the inferred change in climate, but they are not unique (Brain and Jakosky 1998). Two measurements, however, provide supporting evidence.

First, gas trapped in the Martian meteorite ALH84001 may represent a direct sample of the ancient

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

atmosphere. Measurements of nitrogen and argon isotopes indicate that this gas is essentially unfractionated, meaning that most of the isotope-fractionating loss to space through solar-wind stripping had not yet occurred (Mathew and Marti 2001). Isotopic measurements of the gas Xe indicate that it was trapped into the rock 3.9 Gyr ago. If the Xe and lighter gases were incorporated into the rock at the same time and represent a sample of the Martian atmosphere from that epoch, they provide a key time constraint on atmospheric loss (the bulk of the loss to space must have postdated 3.9 Gyr). However, although the gases most plausibly came from the atmosphere and were incorporated into the rock at that time, there are concerns related to their uncertain carrier within the meteorite and to the presumed non-juvenile nature of the gases, such that the interpretation is not unique.

Second, measurements of localized remanent magnetic anomalies detected on Mars today (Acuña *et al* 1999) provide information on the timing of the shut-off of an intrinsic magnetic field. A substantial global magnetic field would cause the solar wind to stand off from the planet, limiting its ability to strip off the atmosphere. In addition, local magnetic anomalies each have the ability to protect the local atmosphere, such that the atmosphere of a planet covered entirely by such anomalies also would be relatively well protected. Thus, the shut-off of an intrinsic magnetic field and the 'erasure' of local remanent magnetic anomalies in the crust, if they ever were distributed globally, would have allowed the turn-on of the stripping of the atmosphere by the solar wind.

The history of the magnetic field and the timing of its shut-off is extremely uncertain (see MARS: CORE AND MAGNETISM). Most of the remanent magnetic anomalies occur in the ancient terrain, indicating that Mars had a substantial intrinsic magnetic field and that it turned off relatively early in Martian history (Acuña *et al* 1999). Suggestion of a late turn-on of the global magnetic field is at odds with magnetization found in carbonate globules in ALH84001, which indicates that a Martian geodynamo was active 4 Gyr ago or earlier. A few younger areas also have remanent fields, although there is a general absence of magnetic anomalies associated with Hesperian and Amazonian volcanics, adding to the confusion. It is clear, however, that an intrinsic magnetic field was most strongly connected to the Noachian epoch and very much less so to later epochs. It may not be possible to resolve the timing issues more clearly, as it is uncertain what the carrier of the remanent magnetic fields is, whether remanent anomalies were originally distributed globally and, if they were, what process would have erased them.

If the turn-off of the magnetic field and the erasure of localized magnetic anomalies allowed the solar wind to begin to strip species out of the atmosphere, then this atmospheric loss would have begun during the Noachian or early Hesperian. The turn-on of atmospheric loss would have been complete only when a substantial fraction of local anomalies had been destroyed. This may not have happened prior to the substantial volcanism associated with Tharsis and the subsequent Hesperian era.

Together, the isotopic and magnetic-field evidence indicates, but does not prove, that the bulk of solar-wind stripping to space occurred subsequent to 3.9 Gyr and that the turn-on of loss occurred sometime between the Middle Noachian and the Early Hesperian. Loss by solar-wind stripping seems to have been roughly contemporaneous with the change in climate inferred from geology, with the main changes in climate coinciding with the end of the Noachian.

Evidence for liquid water in later epochs

There is substantial evidence that liquid water has been present within the Martian crust up to the present. Evidence comes from geological features seen on the surface as well as from geochemical analyses.

Catastrophic outflow channels provide compelling evidence that liquid water was released from within the crust, flowed over the surface and drained into the northern lowlands (see Carr 1996). As these floods involved large quantities of water released catastrophically, with water able to flow substantial distances before freezing, they could occur even in the present cold climate (Carr 1996). They do indicate, however, that liquid water must have been present within the crust. Other eroding agents have been suggested, including liquid CO_2, SO_2, volcanic lava, ice, debris flows and the wind. In particular, the similarity of some of the Martian channels to volcanic channels and the ability of the wind to weather and erode substantial amounts of material, have recently been highlighted, along with the difficulty of reaching unique conclusions about the eroding fluid from morphology alone. However, water or water-filled debris flows remain the most plausible candidates and require the fewest extrapolations. Although the observed floods seem to have drained only about 10% of the surface, connected largely to the Tharsis trough, there is no reason to think that crustal water was not distributed globally.

The Martian meteorites contain weathering products, produced when liquid water was present, filling cracks and voids in the rock. These include carbonate deposits at levels of several per cent in ALH84001, as well as trace amounts of carbonates and the mineral iddingsite (which forms from weathering of basalt in the presence of liquid water) in several of the other meteorites. Detailed microstratigraphy shows that the deposits were present before the rocks were ejected from the Martian subsurface, providing direct evidence that liquid water circulated through the Martian crust.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Indirect geochemical evidence for liquid water in the crust comes from measurements of various isotope ratios within these weathering products. They contain enhanced D / H, $^{13}C/^{12}C$, $^{15}N/^{14}N$ and $^{38}Ar/^{36}Ar$ ratios. Enhancement of each of these ratios is best explained as having resulted from atmospheric processes involving preferential escape of the lighter isotopes to space (as discussed earlier). This requires that these gases once resided in the atmosphere, were left behind as a residuum of loss to space and were subsequently incorporated into the crust. Circulation of groundwater between the surface and the crust provides the best means for exchange of these gases, again suggesting the presence of liquid water (Jakosky and Jones 1997; Watson *et al* 1994).

While the carbonates in ALH84001 were probably precipitated from liquid water about 3.9 Gyr ago, water also must have been present much more recently. The other Martian meteorites are all much younger rocks, having crystallized between about 170 Myr and 1.3 Gyr and also contain weathering deposits (McSween 1994). Thus, crustal water must have been present within the past billion years. Radiogenic dates for the weathering products in the Lafayette meteorite indicate deposition around 650 Myr ago or even more recently.

Several locations on the surface show spectroscopic evidence for the presence of coarse-grained haematite, probably formed in association with liquid water (Christensen *et al* 2000). The haematite may have been deposited from water released from volcanic intrusions or driven by subsurface hydrothermal or aqueous systems (Christensen *et al* 2000), although some could have formed by volcanic processes not requiring water. One of the occurrences is in Aram Chaos, a feature attributed to the release of subsurface water and not connected to volcanism. A second, larger location is in Western Terra Meridiani, located at about 0° latitude and 0° longitude thought to have formed at depth and been exposed subsequently by erosion (Christensen *et al* 2000).

Liquid water also may have been present within the near-surface crust very recently, based on the presence of pristine gullies on the exposed walls of impact craters and valleys (Malin and Edgett 2000). These are identical in size, shape and appearance to gullies on Earth carved by liquid water seeping from aquifers on exposed scarps. Liquid CO_2 has been suggested as a possible eroding agent, based on the similarity of the depth of seeping to that at which the overburden pressure equals the CO_2 liquefaction pressure. However, the absence of a viable charge or recharge mechanism for liquid CO_2 and the inability of CO_2 to discharge as a liquid under Martian conditions preclude its role as a significant erosive agent. Rather, liquid water is much more plausible geologically (see Malin and Edgett 2000). Recent calculations of the stability of liquid water in the crust and of the ability of cyclical oscillations in temperature to freeze and release water, suggest that this is a viable mechanism (Mellon and Phillips in press). There is no means to uniquely determine the age of the features, but they are unweathered and gully debris overlies features such as sand dunes that are themselves thought to be extremely young. If these interpretations are correct, liquid water has been present within a few hundred metres of the surface within the past few million years and may reside there today (Malin and Edgett 2000).

There also is geological evidence for liquid water at the surface, possibly during the later epochs of Martian history. Most intriguing is the evidence for standing lakes within impact craters and the Valles Marineris canyon system (Carr 1996). Potential crater lakes are identified by the presence of channels flowing into and / or out of the craters, providing a source or sink for water. Some deposits within craters have an appearance similar to deltas, sedimentary terraces and shorelines, as might be formed by flowing water. Although most of the craters themselves are relatively old, the age of the lakebed deposits is uncertain. There are few craters on the lakebed deposits, which indicates either that they are very young or that they were buried for long periods and were exhumed very recently (Malin and Edgett 2001); stratigraphic relationships of layered sedimentary deposits within Valles Marineris suggest an old age despite the absence of craters (Malin and Edgett 2001). Well-defined layering within these deposits supports the idea of standing water, although layered lakebed deposits cannot be distinguished uniquely from windblown sediments (Malin and Edgett 2001). However, even if these sediments were windblown, they would have required either liquid water at least in trace amounts or water ice that might be stable at other epochs in order to cement the grains together to form coherent layers.

The catastrophic outflow channels all drain into the northern lowlands (Carr 1996). Possible shoreline features have been identified, suggesting that the water may have accumulated in the lowlands to form a large, long-lived body of water (that is, an ocean). The crudely constant elevation of the innermost of the two main proposed shorelines and the extreme smoothness of the lowlands are held to be consistent with the presence of an ocean. However, the proposed shoreline features are not visible in recent high-resolution images and thus cannot be ascribed with certainty to wave action at an ocean boundary. Features described originally as 'high stands' or shorelines created by a retreating ocean are now recognized as being wrinkle ridges of tectonic derivation. These ridges underlie the Vastitas Borealis formation, a deposit that may in part be sedimentary in origin.

The smoothness of the lowlands (Smith *et al* 1999) may have resulted from fluvially transported sediments associated with channel emplacement during the Hesperian; such a process allows but does not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

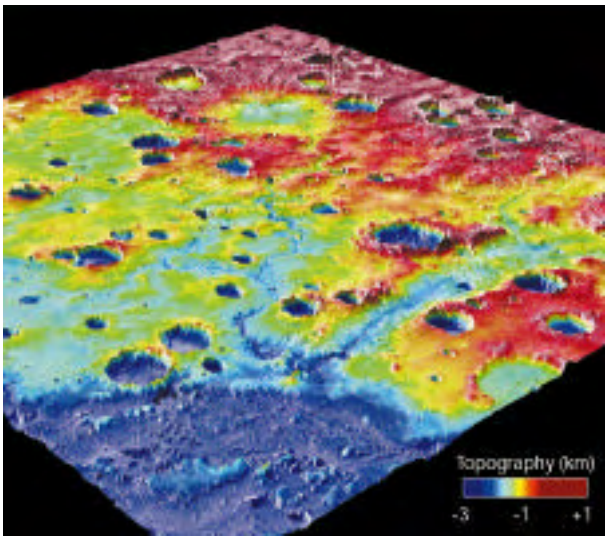


Figure 5. View of topography (from Smith *et al* 1999) looking towards the southeast in Margaritifer Sinus region. Area is 30° x 30° in size and there is 12 vertical exaggeration in topography. Pole-to-pole slope has been removed in topography to emphasize regional variations. The prominent depression in the centre of the image is Parana basin, a possible site of lake sediments; it is drained by a prominent channel (Loire Vallis) that descends northwesterly into the Tharsis trough (foreground). Higher elevations (red shades) are Early- and Middle-Noachian plateau materials. Yellow, green and blue/grey shades of lower elevations represent widespread erosion events in the Late Noachian that removed 1.5 x 10^6 km^3 of Early/Middle-Noachian material from this region. Erosional remnants of older material occur in isolated mesas. (Image courtesy of B Hynek.)

require the presence of an ocean. Erosion of material in the Margaritifer Sinus region and an adjacent area in northwestern Arabia Terra may have provided a significant source of sediments. About a kilometre of material seems to have been stripped away during the Late Noachian and geological evidence points to erosion by liquid water (figure 5). The resulting debris could have filled the northern plains to a depth of about 100 m north of 30° N. The corresponding equivalent depth of water necessary to transport this sediment would have been several times the sediment volume. Whether or not this resulted in a Late Noachian ocean would have depended on the relative rates of erosion and of water removal by ground infiltration, evaporation and so on. Linear gravity anomalies trending northward from Chryse Planitia may be indicative of early channels buried in the northern plains, analogous to the main catastrophic flood channels seen to the east of Valles Marineris (Zuber *et al* 2000). These could be Noachian-era pathways that moved water and sediment from the southern highlands to the northern lowlands.

In addition, the Hesperian outflow flood channels can be traced northward into Acidalia Planitia. There, they converge into a basin (not presently closed) with an area 6 x 10^5 km^2, south of the high-standing Noachian inlier Acidalia Mensa at 45° N latitude (figure 3(*b*)). Acidalia Mensa itself has erosional boundaries on the south and east that are channels (figure 3(c)) and may at one time have formed a natural barrier to water flow. This region of channel convergence coincides closely with one of two main occurrences of kilometre-scale polygonally fractured terrain. The origin of polygonally fractured terrain is controversial, but many of the processes proposed involve water. Thus, it is likely that water (and thus sediments) were carried far into the northern lowlands during the Hesperian, perhaps forming lakes. The region of channel convergence in Acidalia Planitia may mark the largest possible extent of any standing bodies of surface water. If there ever was a Hesperian northern 'ocean', it may have been more of a large lake about the size of the Caspian Sea (4 x 10^5 km^2).

Synthesis of the observations and interpretations

It is clear that there are a number of strong constraints on the history of water on Mars. They tell us that water was present at the surface early in the planet's history and within the crust throughout time. They also allow us to construct a self-consistent scenario of the history of Martian volatiles and climate, although such a scenario is not unique.

The history of Martian volatiles involved the following (see figure 2).

Most of the earliest atmosphere of Mars was lost during the Early Noachian by impact erosion and hydrodynamic escape. A secondary atmosphere was created by water and CO_2 released to the atmosphere as a direct result of Tharsis volcanism and this may have had a strong influence on climate. It is likely that volatiles were also released by non-Tharsis Noachian volcanism presumed to have been responsible for forming the ancient highland crust. Water and CO_2 were lost from the surface and atmosphere system to space, to the polar caps and to carbonate deposits within the crust. There is compelling evidence for the existence of each of these sinks, as described above, although it is not possible at this time to determine uniquely the relative or absolute importance of each. There is a coincidence in the timing of major events in Martian history. The decrease in the impact rate at the end of heavy bombardment, the formation of the bulk of the Tharsis construct by magmatism, the decline in the intensity or existence of an intrinsic magnetic field, the change in climate inferred from the morphological characteristics of the surface and the loss of substantial volatiles to space all occurred at nearly the same time and marked the end of the Noachian epoch about 3.7 Gyr ago. Many of these events are likely to have been causally connected to each other, although some of the similarity in timing may be coincidental. However, the loss of atmos-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

pheric protection by the shutdown of the global magnetic field and the decline in the rate of Tharsis volcanism at the end of the Noachian probably both were instrumental in the major shift in climate as inferred from the geological record. There has been a reservoir of crustal liquid water throughout Martian history, continuing possibly to the present. This reservoir has exchanged, at various times and to various degrees, with water at the surface and in the atmosphere. We have little information as to how much water is contained in this reservoir today, but it is likely to be substantial, globally distributed and readily accessible to the surface in some places.

New measurements from both planned and potential missions will allow us both to test the above hypotheses and to further constrain the processes that have taken place. Although the list of relevant observations is long, some measurements will be key to improving our understanding. These include determining the present-day isotope ratios of the climate-related species (H, C, O and N) and the noble gases, sampling rocks over a broad age spectrum and determining the abundance of weathering products and the isotope ratios of gases contained within them, continuing the search for evidence for trapped carbonate or sulphate minerals within the crust, especially in rocks previously underground and now exposed at the surface, and continuing to explore the morphology of the surface for evidence of climate-related processes.

Discoveries made in the past decade have had a tremendous impact on our understanding of the history of Martian water, volatiles and climate. Analysis of the Martian meteorites and the measurements made from the Mars Global Surveyor spacecraft in particular have been key to progress in this area. Unravelling the history of Martian volatiles and climate will be central to addressing in a meaningful way the potential for Martian life and to interpreting results obtained specifically to look for evidence of present or past life.

Acknowledgements

We thank C. Leovy, S. Stewart, L. Leshin, M. Mellon, H. Frey, P. Withers, B. Hynek, K. Harrison, W. Hartmann and the MOLAscience team for valuable discussions and for providing preprints of their manuscripts. We also thank J. Head, R. Haberle and C. Leovy for detailed reviews of our manuscript. This research was supported by the Mars Global Surveyor Project and the NASA Planetary Geology and Geophysics Program.

This article is a modified version of a previously published article: *Nature* **412** 237–44 (2001). Reproduced by courtesy of *Nature* [http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v412/n6843/abs/412237a0_fs.html]

Bibliography

Acuña M H *et al* 1999 Global distribution of crustal magnetism discovered by the Mars Global Surveyor MAG/ER experiment *Science* **284** 790–3

Brain D A and Jakosky B M 1998 Atmospheric loss since the onset of the martian geologic record: combined role of impact erosion and sputtering *J. Geophys. Res.* **103** 22689–94

Carr M H 1996 *Water on Mars* (New York: Oxford University Press)

Christensen P R *et al* 2000 Detection of crystalline hematite mineralization on Mars by the Thermal Emission Spectrometer: evidence for near-surface water *J. Geophys. Res.* **105** 9623–42

Christensen P R *et al* in press Mars Global Surveyor Thermal Emission Spectrometer experiment: investigation, description and surface science results *J. Geophys. Res.*

Hartmann W K and Neukum G in press Cratering chronology and the evolution of Mars *Space Sci. Rev.*

Hartmann W K *et al* 1981 *Basaltic Volcanism on the Terrestrial Planets* ed Basaltic Volcanism Study Project (New York: Pergamon) pp 1049–127

Jakosky B M and Haberle R M 1992 *Mars* ed H H Kieffer, B M Jakosky, C W Snyder and M S Matthews (Tucson, AZ: University of Arizona Press) pp 969–1016

Jakosky B M and Jones J H 1997 The history of Martian volatiles *Rev. Geophys.* **35** 1–16

Jakosky B M, Pepin R O, Johnson R E and Fox J L 1984 Mars atmospheric loss and isotopic fractionation by solar-wind-induced sputtering and photochemical escape *Icarus* **111** 271–88

Leshin L A 2000 Insights into Martian water reservoirs from analyses of Martian meteorite QUE94201 *Geophys. Res. Lett.* **27** 2017–20 (2000)

Malin M C and Edgett K S 2000 Evidence for recent ground water seepage and surface runoff on Mars *Science* **288** 2330–35

Malin M C and Edgett K S 2001 Sedimentary rocks of early Mars *Science* **290** 1927–37

Mathew K J and Marti K 2001 Early evolution of Martian volatiles: nitrogen and noble gas components in ALH84001 and Chassigny *J. Geophys. Res.* **106** 1401–22

McSween H Y Jr 1994 What we have learned about Mars from SNC meteorites *Meteoritics* **29** 757–79

Mellon M T and Phillips R J in press Recent gullies on Mars and the source of liquid water

Owen T 1992 *Mars* ed H H Kieffer, B M Jakosky, C W Snyder and M S Matthews (Tucson, AZ: University of Arizona Press) pp 818–34

Phillips R J *et al* 2001 Ancient geodynamics and global-scale hydrology on Mars *Science* **291** 2587–91

Purucker M *et al* 2000 An altitude-normalized magnetic map of Mars and its interpretation *Geophys. Res. Lett.* **27** 2449–52

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Smith D E *et al* 1999 The global topography of Mars and implications for surface evolution *Science* **284** 1495–1503

Squyres S W and Kasting J F 1994 Early Mars: how warm and how wet? *Science* **265** 744–9

Watson L L, Hutcheon I D, Epstein S and Stolper E M 1994 Water on Mars: clues from deuterium / hydrogen and water contents of hydrous phases in SNC meteorites *Science* **265** 86–90

Zuber M T *et al* 2000 Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity *Science* **287** 1788–93

Bruce H Jakosky and Roger J Phillips

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mars: Core and Magnetism

The existence of the MARTIAN CORE, which has been accepted for many decades, is interesting for several reasons. First, its size and composition tell us about Mars as a whole — its constituents and provenance. Second, its antiquity tells us about early conditions on Mars; we believe that the core formed early, and this requires that Mars had a hot beginning. Third, this core is the likely source of a magnetic field for some part of Mars' history, probably the earliest part, just as Earth's core is the source of the current geomagnetic field. Fourth, the field may have influenced the early climate through its influence on atmospheric escape. It could also have affected the environment for early life on Mars. Fifth, the heat flow from the core may have fed mantle plumes and influenced volcanic activity, much as hot spots such as Hawaii are thought to be fed by core heat flow on Earth. Sixth, a core, if partly or entirely liquid, influences rotational dynamics, just as (for example) changes in length of day are influenced by Earth's liquid core.

Mars is built from roughly the same ingredients as Earth: silicates and oxides of magnesium and iron, as well as metallic iron (alloyed with various constituents). When we refer to a core for the terrestrial planets Mercury, Venus, Earth and Mars, we mean a central region that is rich in metallic iron. Because this material is about twice as dense as the silicates and oxides making up the CRUST and MANTLE, its presence as a core is revealed through its influence on the mean density of the planet and through its effect on the moment of inertia. Old measurements of gravity and more recent geodetic data from Pathfinder (Folkner *et al* 1997) reveal that the mean moment of inertia for Mars is $0.365MR^2$, where M is the mass of Mars and R is its mean radius. Together with the Martian mean density of 3.93 g cm^{-3}, this suggests a model of Mars that is not too different from a scale model of Earth (that is, similar ingredients, distributed similarly). The Martian core is proportionately a little smaller than Earth's core, and proportionately more iron is found in the mantle (in oxides or silicates). As we do not know the composition of the core we cannot be certain about core size, but a core radius of around 1 300–1500 km (depth to core of 1900–2100 km) is indicated. Figure 1 shows a simple interior structure of Mars. Bertka and Fei (Bertka and Fei 1998) suggest a composition for Mars that is different from partly devolatilized primitive METEORITES.

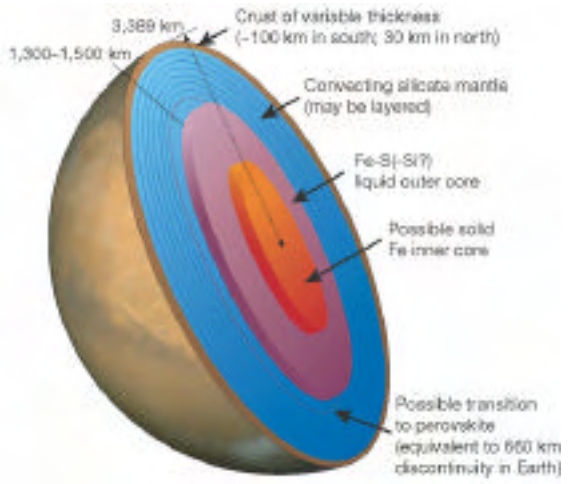


Figure 1. Cutaway view of the Martian interior

Far less is known about the Martian core than Earth's core because we lack seismological evidence or geodetic data of sufficient precision. In particular, we do not know whether the core is entirely liquid, partly liquid (like Earth) or entirely solid, although there are indirect arguments against an entirely solid core. Fortunately, because Mars is at lower pressures than Earth, and so is more accessible to high-pressure experiment, it is possible to assess the likely phase composition for the core.

Core formation

It is widely accepted that terrestrial planetary cores owe their existence to a process of gravitational separation of mostly liquid, immiscible iron from the (partly) solid silicates. The supporting arguments are partly physical (Stevenson 1990), but increasingly geochemical. Although we have no samples of either the core of Earth or the core of Mars, we do have rocks that are probably indicative of mantle composition. For Mars, these are the very limited yet highly important SNC METEORITES (for shergottites, nakhilites and chassignites). As on Earth, these igneous rocks show a striking depletion of 'iron-loving elements' (called siderophiles) whose extraction testifies to the conditions of core formation (Righter, Hervig and Kring 1998). Isotopic data (Chen and Wasserburg 1986, Lee and Halliday 1997) also suggest that this core-forming event was early in Mars' history.

These data, together with physical modelling, suggest a scenario similar to the following. Mars accumulated from smaller bodies over a period of perhaps as long as 100 million yr (Myr), but possibly much shorter, around 4.5–4.6 billion yr (Gyr) ago (Wetherill 1994, Chambers and Wetherill 1998). Isotopic evidence is compatible with a very short accretion time, suggesting that Mars might even have been a runaway, isolated embryo rather than a slowly accumulated body like Earth. In this accretional process, the impacting bodies may have already had iron cores, but the energetics of the impact events would have caused extensive melting and mixing of the immiscible metallic iron and silicate / oxide components, allowing chemical re-equilibration on a small scale (centimetres to metres) (Stevenson 1990). A substantial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

mass of the impacting bodies may have been in the form of giant impacts (bodies of the order of the mass of Earth's Moon), but even much smaller bodies can bury a great deal of heat at depth. The energy of gravitational formation of Mars is roughly $0.6GM/R$ per unit mass, where G is the gravitational constant. If all this were converted into heat it would be sufficient to heat the Mars-forming material to several thousand degrees Celsius above the melting point and, even with the loss of heat by radiation, a magma ocean is likely. This ocean might be transient (surviving for a brief period after each giant impact) or it might be sustained by a dense steam atmosphere, but in either case it will define the conditions of most of the core formation.

Although not as hot or at such high pressure as the likely conditions that formed Earth's core (Righter, Hervig and Kring 1998), the lower gravity on Mars would still permit a thick magma ocean. Metallic iron can settle as droplets in a convecting magma ocean, to accumulate as large blobs ('diapirs') that then ascend by Stokes flow through the possibly more viscous and high-pressure deep mantle. In this scenario, the core might initially be either a few hundred degrees Celsius hotter than the mantle (if the energy of core formation is retained substantially within the iron) or the same temperature as the deep mantle (if efficient thermal equilibration takes place). Much of the Martian crust (particularly that preserved in the south) may have formed in this very earliest epoch.

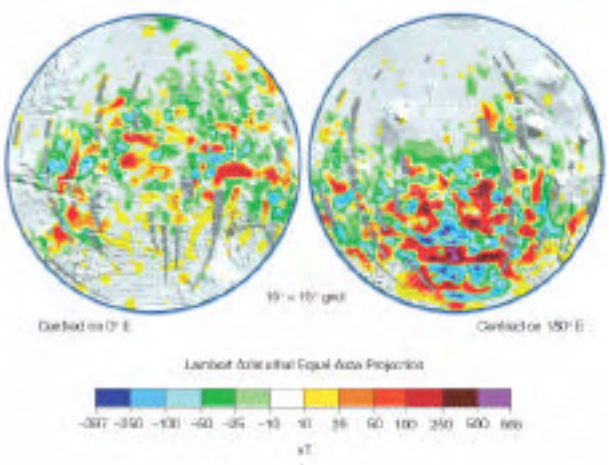


Figure 2. Radial magnetic field at 200 km altitude, based on data collected by the Mars Global Surveyor spacecraft. The southern hemisphere exhibits the largest field anomalies. Notice that the opposite hemisphere has field anomalies that are typically an order of magnitude smaller and exhibit less coherent spatial structure. These maps were prepared by M. Purucker, Goddard Space Flight Center. They are in Lambert azimuthal equidistant projection and were published (in a different projection) in Purucker *et al* 2000.

Martian magnetism

Mars, unlike Earth, has no global dipole magnetic field. The Mars Global Surveyor spacecraft confirmed this but also found strong, spatially variable magnetic fields at altitudes of 200 km down to closest approach of 110 km (Acuña *et al* 1999, Connerny 1999). Figure 2 shows hemispherical maps of the radial field normalized to a constant 200 km altitude. The fields are measured below much of the Martian ionosphere and much of the power in their spatial variability is at length scales comparable with the distance from the surface. 'Inversion' of these data is non-unique, but the source of the field must be confined to the outermost several tens of kilometres of the crust (and possibly confined to an even thinner layer). A deeper layer or source of currents could not provide the observed spatial structure, except with physically implausible assumptions. The inferred crustal magnetizations are up to 10–30 A m^{-1}, an order of magnitude higher than the strongest magnetizations typically encountered in Earth rocks, and even these values are underestimates if one were to require thinner magnetized layers or incoherent magnetization directions. Given the large amounts of magnetized crust required, it seems very likely that the magnetization is thermal remanence acquired during the last time the rocks cooled through the blocking temperature, for example following dike injection. Most of these cooling events took place at a time when a large global field was present. ('Most' rather than 'all' is appropriate here, because the surface magnetizations are so large that it is possible for crustal rock to be substantially magnetized through cooling in the presence of other crustal fields, rather than a global field.) Other origins of the magnetization (for example, due to impact, as suggested for the Moon) are conceivable in principle, but seem insufficient given the magnitude of the requirements.

There is great interest in the meaning of the spatial pattern of magnetization, including possible lineations that suggest an analogy to plate tectonically (see TECTONICS) derived lineations of magnetization on Earth's ocean floor, but the current constant-altitude (that is, constant-resolution) maps do not provide strong support for these speculations. Models of the crustal magnetization suggest that the Martian field may have undergone reversals.

It seems likely that Mars requires at least one (and preferably several) of the following: high abundance of appropriate magnetic materials (for example, magnetite), a particularly favourable magnetic mineralization (for example, single domains), large volumes of crust that are coherently magnetized and/or an unusually large field in which the magnetization was acquired.

The fact that Mars did have a global magnetic field for one or more periods in its early history suggests that it once had an active core dynamo, the process responsible for Earth's current field. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

strongest magnetizations are observed in the ancient southern highlands of Mars, which pre-date 4 Gyr. The antiquity of these regions is inferred by the (imprecise) method of crater counts. However, not all ancient crust on Mars produces large magnetic fields at the spacecraft altitude, and not all younger crust is devoid of magnetization. This prevents firm conclusions being made about the timing of acquisition of magnetization and hence the timing of a postulated Martian core dynamo.

Schubert *et al* (Schubert, Russell and Moore 2000) have suggested a later (post 4.0 Gyr) period of magnetization, which would indicate a later period of dynamo activity. However, some arguments point towards ancient (4.2 Gyr or earlier) acquisition. First, as noted by Acuña *et al* (Acuña, 1999), the ancient impact structure Hellas (believed to be at least 4 Gyr old) lacks any magnetic signature and seems to be surrounded by a region with very little coherent magnetization. This is a plausible outcome were the impact to have occurred when Mars possessed no global magnetic field. It is not a plausible outcome if the southern crust were subsequently reheated and then cooled to acquire magnetization during a later epoch in which a global field was active. Second, it is difficult to imagine any physically plausible scenario in which large provinces of the southern crust were extensively heated later in Mars' history without producing some surficial difference in appearance from those regions that were not so treated. This argument is supported by the recognition that huge volumes of crust are required to explain the observed magnetization, rather than some thin layer of possibly remagnetized material. In particular, revived igneous activity generating new crust would certainly disrupt these terrains because of the large amounts required. Third, evidence from the ancient Martian meteorite ALH84001 suggests that its magnetization was acquired at 4.0 Gyr or even earlier.

A region that lacks large magnetic fields at the spacecraft altitude might still consist of crust that formed in the presence of a global magnetic field. For example, the magnetization may be spatially incoherent, the cooling history may have favoured multidomain magnetite or less favourable mineralization, or the field may have been reversing more rapidly. Moreover, the early rapid pace of planetary evolution means that regions in the south that seem to be of the same age may nonetheless differ in age by 100 Myr and thus cooled in a different magnetic field, even though their surface appearance and crater density seem to be identical. Last but not least, ancient crust may underlay younger crust in some northern localities, thereby allowing preservation of a (relatively weak and patchy) magnetization, even when the surface age postdates any global field.

A Martian dynamo?

The dynamo mechanism is much studied but still imperfectly understood, despite recent advances in numerical simulation. In particular, we do not know the conditions sufficient for the existence of a planetary dynamo. Because we can only speculate about early conditions on Mars, the problem of inferring or predicting the history of a Martian dynamo is indeed formidable. Earth's dynamo is also imperfectly understood, although it is thought that it arises from convection driven largely by inner-core growth (Stevenson, Spohn and Schubert 1983).

If a dynamo exists, then it is likely that the expected field magnitude B inside the region of field generation is given by the Elsasser number of order unity. This implies $B \sim (2\rho\Omega/\sigma)^{1/2}$, where ρ is the fluid density, Ω is the planetary rotation rate and σ is the electrical conductivity. For the Martian core, this yields $\sim 10^{-3}$ tesla, but because this is the same prediction as for present Earth, the predicted palaeofield at the surface of Mars is indistinguishable (at the level of this crude argument) from the present field at Earth's surface. Dynamo theory admits weaker fields as possible solutions, but it does not admit fields substantially larger than $B \sim (2\rho\Omega/\sigma)^{1/2}$. Despite suggestions to the contrary (Stevenson, Spohn and Schubert 1983), there is no theoretical basis at present for the idea that the field scales in some direct way with the energy source, so that it might undergo slow decline over geological time or large changes arising from inner-core nucleation.

One speculative explanation for the origin of magnetization on Mars is that the field was generated in a magma ocean. Plausible numbers are a characteristic fluid velocity v of 10^{-1} m s^{-1} (because of very high heat flows at that time), a characteristic length scale L of 10^6 m and a magnetic diffusivity of 10^4 m^2 s^{-1} (possibly appropriate to high-temperature and high-pressure silicate melts), which together give a magnetic Reynolds number R_m of 10. This is marginal at best, but would be attractive because large fields are predicted (0.01–0.1 T at the Martian surface). The extremely high observed magnetizations might then be explained, although a core dynamo is more plausible.

Thermal or compositional convection?

If one accepts that core convection is needed, then a probable necessary condition for a dynamo is the presence of convection. In terrestrial planets (including Earth), the criterion for core convection is difficult to satisfy. The reason for this is that the natural scale for core heat flows is such that this heat can probably be carried by conduction at a temperature gradient that is stably stratified (that is, it inhibits convection). To obtain core convection, one must appeal to unusually large heat flows or the development of an inner core. In either case, the core must be cooling. To appreciate this argument,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

consider first the simple case of no inner core. Convection will occur provided that the heat flux within the core exceeds that which is carried by conduction along an adiabat:

$$F_{\text{total}} > F_{\text{cond.ad}} \equiv k\alpha Tg / C_{\text{p}} \Leftrightarrow \text{thermal convection}$$

where k is the thermal conductivity, α is the coefficient of thermal expansion, T is the temperature, g is the gravitational acceleration and C_{p} is the specific heat at constant pressure. These parameters are all slowly varying within a core (if T is close to being adiabatically distributed), except for g, which is approximately linear in radius r, the distance from the planet centre. If the core is simply cooling and releasing the stored sensible heat (provided by gravity during planetary accretion), then the total heat flux is also linear in r: $F_{\text{total}}(\text{r}) = -\rho C_{\text{p}} r(\text{d}T_{\text{c}}/\text{d}t)/3$, where T_{c} is the mean core temperature and t is time. It is unlikely that the core contains significant radioactive heat sources (even less likely than Earth, where one can always appeal to unknown, very high pressure effects). Consequently, if thermal convection ceases to operate in the outer part of the core, then it will also cease to operate at about the same time elsewhere in the core.

If the core is cooling and the central temperature drops below the liquidus for the core alloy, then an inner core will nucleate. In Earth, we know from seismic evidence that the core is ~10% less dense than pure iron, and many suggestions have been offered for the identity of the light elements that are mixed with the iron. At the lower pressures relevant to Mars, the dominant light element may be sulphur (Bertka and Fei 1998). For sulphur abundances that are less than cosmic relative to iron, as cosmochemical arguments suggest, the inner core will be nearly pure iron (with some nickel) and the sulphur will be entirely in the outer core. The introduction of this light element into the fluid of the lowermost core will tend to promote convection and cause mixing throughout all or most of the outer core, provided that the cooling is sufficiently fast. Latent heat release at the inner-core–outer-core boundary will also contribute to the likelihood of convection. However, inner-core growth permits outer-core convection even when the heat flow through the core–mantle boundary is less (perhaps much less) than the heat carried by conduction along an adiabat. In this regime (possibly that occupied by Earth), the temperature gradient is very slightly less steep than adiabatic and the compositional convection carries heat downwards. The total heat flux is still outwards, of course, as the heat carried by conduction is large. This state is possible because the buoyancy release associated with the compositional change exceeds the work done against the unfavourable thermal stratification. Unlike thermal convection, compositional convection may not cease everywhere throughout the core at a single epoch. This argument is modified in detail but not in general outline, should the core include a light element that does not exhibit eutectic behaviour (for example, silicon) as well as the (universally expected) complement of sulphur.

Required cooling rates

For plausible choices of parameters, the cooling rate of the core must exceed about 80 K Gyr^{-1} to obtain thermal convection. However, this estimate is uncertain by perhaps as much as a factor of 2. The required cooling rate in the presence of a growing inner core is much smaller, by a factor of several (Stevenson 1983), but has not been studied in detail for Mars. As a consequence, models with an inner core will tend to sustain a dynamo for a long time, perhaps even to the present day, unless there is something unusual about the thermal history (as suggested below). The overlying mantle determines the cooling rate. Indeed, it is the mantle that determines whether a terrestrial planet has core convection and whether it can have a dynamo.

It is also possible that the sufficient condition for a dynamo is not far removed from the necessary condition for the presence of any convection (Stevenson 1983). Simple scaling laws for convection (compatible with the philosophy of Kolmogorov turbulence and known to astrophysicists as mixing length theory) suggest that $v \sim 0.1(F_{\text{conv}}/\rho)^{1/3}$, where ρ is the fluid density and F_{conv} is the convective heat flux (or its compositional equivalent when the convection is driven by compositional density differences). I define

$$\varepsilon = F_{\text{conv}} / F_{\text{cond.ad}} = (F_{\text{total}} - F_{\text{cond.ad}}) / F_{\text{cond.ad}}$$

Substitution above shows immediately that for plausible parameters in the Martian core ($L \sim 10^6$ m, $\lambda \sim 1$ m^2 s^{-1}, $\rho \sim 10^4$ kg m^{-3}), R_{m} may be large even if $\varepsilon \ll 1$. That is, the heat flow has to only slightly exceed that for any convection in order to reach that for convection of sufficient vigour to sustain a dynamo. This claim must be tested by further numerical work. It is conceivable, but difficult energetically, for a dynamo to function for $\varepsilon < 0$ (for example, because of baroclinic instabilities and thermal winds arising from horizontal temperature gradients that are caused by lateral differences in heat flow through the core–mantle boundary). Even in this case, one would expect that a dynamo requires $|\varepsilon| \ll 1$, as the vertical motions would otherwise be strongly suppressed and this inhibits dynamo activity. Alternatives to convective driving (for example precession) still require the core to be close to adiabatic and thus do not escape the constraints discussed above.

In conclusion, if the mantle cools fast enough (or is cool enough to allow inner-core nucleation) then a dynamo occurs, but if the mantle is too hot or fails to cool then there is no dynamo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

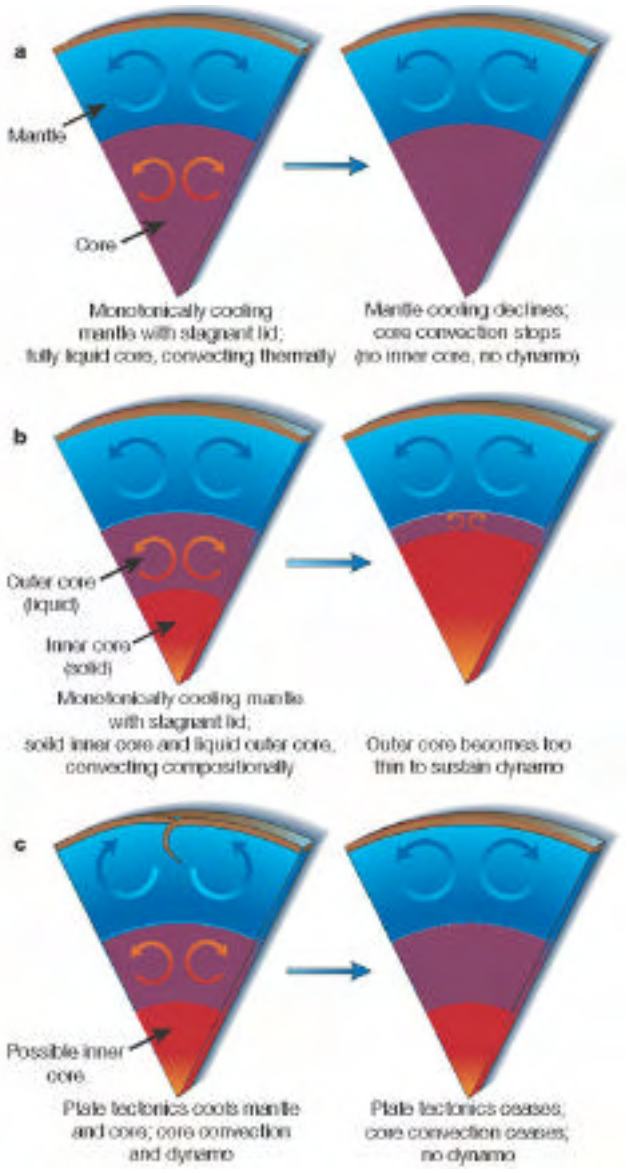


Figure 3. Possible scenarios of Martian core evolution: (a) high-sulphur model; (b) low-sulphur model; (c) plate-tectonic model (works for all sulphur contents).

Possible histories of the Martian core

Three possible scenarios for the history of the Martian dynamo are presented in figure 3. The first is the simplest: the planet starts out very hot and cools quickly at first. The core remains completely liquid throughout. As the cooling rate declines, a point is reached at which the heat flow out of the core can be accommodated by conduction alone. At that epoch, the dynamo turns off (in a very short time geologically, perhaps as little as a few thousand years) and no further field generation is possible, provided that an inner core never develops. This model requires that the core of Mars is sulphur rich, perhaps 10% or more by mass. It also requires tuning of the parameters so that the dynamo turns off as early in Mars' history as some arguments suggest. Most published models are of this kind.

The second scenario is almost the antithesis of the first. It has not been modelled in detail, although it is implicit in the early work of Young and Schubert (Young and Schubert 1974), who considered the possibility of complete core freezing. In this model, the sulphur content for the core is sufficiently low that an inner core develops early and grows rapidly. The liquid outer core becomes progressively more sulphur rich and evolves towards the eutectic composition. The experimental data (Fei, Bertka and Finger 1997) yield a eutectic of around 1400 K at the top of the core. For realistic models of Mars' mantle convection (Nimmo and Stevenson 2000), the expected present-day core–mantle boundary temperature is at least 1850 K, so there is no prospect that the outer core will completely freeze. However, it is conceivable that the outer core will become sufficiently thin that dynamo activity can no longer be sustained. This would seem implausible based on simple scaling arguments, but it might be the state that Mercury currently occupies. It probably requires a lower sulphur content of the Martian core than most would consider plausible, perhaps no more than a few per cent (even less than typical estimates for Earth). Further dynamo simulations are needed to test this hypothesis.

The third scenario invokes a change in mantle convection to trigger the death of the Martian dynamo (Nimmo and Stevenson 2000). It is assumed that early Mars had mobile lid convection in which the lithosphere could be recycled. On Earth, this is accomplished by plate tectonics, and this could also be the case on Mars (Sleep 1994). (It is, however, the recycling of the lithosphere that matters, not the form of the recycling; so there is no need to assume that Mars did exactly what Earth does.) At some time, perhaps after only a few hundred million years, this process ceased and Mars evolved slowly into the stagnant lid regime that it (and all terrestrial bodies except Earth) currently occupies. If this regime follows one of lithospheric recycling, then the mantle must heat up, because the elimination of heat is less efficient. In other words, the coldest time for the Martian mantle was early in Mars' history, despite the inexorable monotonic decline of radioactive heat sources in the mantle and the crust. This scenario has the advantage that it may work for all possible sulphur contents in the core, as the presence of an inner core will not drive a dynamo if the mantle minimum temperature was reached early in Mars' history. An inner core drives a dynamo only while it is growing, and it can grow only if the core is cooling. One problem with this scenario is that it invokes an *ad*

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

hoc timing for the cessation of 'plate tectonics'; it also implies the ability for Mars to be volcanically active throughout geological time.

In all these scenarios, the beginning of dynamo activity may be delayed after Mars' accretion until a thermal boundary layer builds up in the lowermost mantle, depending on the uncertainty in the initial temperature difference between the core and mantle. However, this is unlikely to produce a delay of more than 100 Myr.

There may be other scenarios not yet considered. Unfortunately, none of these scenarios can be tested with great confidence because the parameters that define their chronologies are not known with sufficient accuracy. However, the presence or absence or size of the inner core is clearly a crucial variable and may eventually be determined by a combination of geodesy and seismology. Numerical dynamo modelling will also be important in the coming years.

Consequences of the Martian core and dynamo

Core cooling dictates the presence of a thermal boundary layer at the base of the overlying mantle. Plumes can detach from this layer and may be a cause of hot-spot volcanism. Harder and Christensen (Harder and Christensen 1996) have proposed that Mars may be in a regime where a single plume dominates because of the effect of a major endothermic phase transition near the base of the Martian mantle (the same phase transition that defines the upper-mantle–lower-mantle boundary on Earth). This plume might be stable for a long period of time, perhaps billions of years, and may be responsible for the Tharsis volcanic province. This hypothesis provides the exciting prospect of linking core thermal history with Martian volcanic history. However, it leaves unanswered several questions. If the core heat flow is so low (as required by the absence of a dynamo throughout much of Mars' history), then is it reasonable to suppose that it is responsible for the dominant volcanic activity on Mars? Why would a deep-seated plume happen to produce volcanism at a location just northward of the principal geological feature (the crustal dichotomy)? Why is the plume so stable? Perhaps the answer to Tharsis lies nearer the surface of Mars rather than in the core history.

The history of the atmosphere (Brain and Jakosky 1998) (see MARS: ATMOSPHERE) may also be influenced by the magnetic field history through the effect of the field on atmospheric sputtering. The history of Martian magnetism might even be linked to the history of life on Mars. Perhaps the strongest argument for a biological effect in ALH84001 lies in the single-domain magnetite grains, whose presence in biological organisms is useful only while Mars has a field. This might also push Martian magnetism back to the earliest epoch.

The future

Although Martian core studies can benefit from work in all areas of planetary science (including geochemistry), the greatest contribution is likely to arise from seismological and geodetic efforts. In particular, the Mars Netlander mission and subsequent follow-ups are likely to have the greatest role. It may also be essential to better characterize the surface magnetization, something that no currently funded mission can do. We can also look forward to exciting developments in our understanding of dynamos. Mars' core is at least as interesting as Earth's core for our general understanding of planet evolution.

This article is a modified version of a previously published article: *Nature* **412** 214–19 (2001). Reproduced by courtesy of *Nature* [http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v412/n6843/abs/412214a0_fs.html]

Bibliography

Acuña M H, *et al* 1999 Global distribution of crustal magnetization discovered by the Mars Global Surveyer MAG/ER experiment *Science* **284** 790–3

Bertka C M and Fei Y W 1998 Implications of Mars Pathfinder data for the accretion history of the terrestrial planets *Science* **281** 1838–40

Brain D A and Jakosky B M 1998 Atmospheric loss since the onset of the Martian geologic record: combined role of impact erosion and sputtering *J. Geophys. Res.* **103** 22689–94

Chambers J E and Wetherill G W 1998 Making the terrestrial planets: *N*–body integrations of planetary embryos in three dimensions *Icarus* **136** 304–27

Chen J H and Wasserburg G J 1986 Formation ages and evolution of Shergotty and its parent planet from U–Th–Pb systematics *Geochim. Cosmochim. Acta* **50** 955–68

Connerney J E P *et al* 1999 Magnetic lineations in the ancient crust of Mars *Science* **284** 794–8

Fei Y W, Bertka C W and Finger LW 1997 High–pressure iron–sulfur compound Fe_3S_2 and melting relations in the Fe–FeS system *Science* **275** 1621–3

Folkner W N, Yoder C F, Yuan D N, Standish E M and Preston R A 1997 Interior structure and seasonal mass redistribution of Mars from radio tracking of Mars Pathfinder *Science* **278** 1749–52

Harder H and Christensen U R 1996 A one–plume model of Martian mantle convection *Nature* **380** 507–9

Lee D C and Halliday A N 1997 Core formation on Mars and differentiated asteroids *Nature* **388** 854–7

Nimmo F and Stevenson D J 2000 Influence of early plate tectonics on the thermal evolution and magnetic field of Mars *J. Geophys. Res.* **105** 11969–79

Purucker M *et al* 2000 An altitude–normalized magnetic map of Mars and its interpretation *Geophys. Res. Lett.* **27** 2449–52

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Righter K, Hervig R Land Kring D A 1998 Accretion and core formation on Mars: molybdenum contents of melt inclusion glasses in three SNC meteorites *Geochim. Cosmochim. Acta* **62** 2167–77

Schubert G, Russell C T and Moore W B 2000 Timing of the Martian dynamo *Nature* **408** 666–7

Sleep N H 1994 Martian plate tectonics *J. Geophys. Res.* **99** 5639–55

Stevenson D J 1983 Planetary magnetic fields *Rep. Prog. Phys.* **46** 555–620

Stevenson D J, Spohn T and Schubert G 1983 Magnetism and thermal evolution of the terrestrial planets *Icarus* **54** 466–89

Stevenson D J 1990 *Origin of the Earth* ed H E Newsom and J E Jones (New York: Oxford Univiversity Press) pp 231–50

Wetherill G W 1994 Provenance of the terrestrial planets *Geochim. Cosmochim. Acta* **58** 4513–20

Young R E and Schubert G 1974 Temperatures inside Mars: is the core liquid or solid? *Geophys. Res. Lett.* **1** 157–9

David J Stevenson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mars: Surface

Like all bodies of the solar system, Mars initially experienced a period of heavy meteoritic bombardment that ended about 3.8 Gyr ago. The planet has remained geologically active much longer than Mercury or the Moon, and volcanism and tectonics have continued into a relatively recent past. However, activity was never efficient enough to conceal completely the previous stages of evolution; because of the small amount of energy available to reshape the surface, a large part of the ancient surface is preserved that dates back to the period of heavy bombardment, and the Martian surface presents a sort of catalog of the successive steps in planetary evolution. In the framework of comparative planetology, Mars represents the intermediate situation between the Moon and Mercury on the one hand and Venus and the Earth on the other hand. It is therefore a key element to understand the evolution of telluric bodies. In addition, Mars is probably the most favorable location in the solar system for the development of a prebiotic chemistry, or even for the development of life. These two elements explain why Mars has always been a favored target for space exploration. It is also the most accessible body in the solar system after the Moon, requiring only a 9 month journey with a new launch opportunity every 26 months. As we will see, the study of the surface leads to the inventory of volatile elements, so it is closely related to that of the atmosphere, and gives fascinating hindsight into the past climatic environments.

Historical perspective

Detailed observations of the surface became possible only half a century after GALILEO GALILEI first observed Mars around the year 1610, thanks to the progress of optical instruments. By the end of the 17th century Christiaan HUYGENS and G D CASSINI had produced the first drawings of Mars, determined the rotation period (24 h 40 min) from permanent markings on the surface and identified the polar caps. The most important subsequent observations were performed by Filippo Maraldi (from 1672 to 1719), W HERSCHEL (1777–1783) and J E Schroeter (from 1785 to 1802). By the mid-19th century, the main characteristics of the planet were known; the occurrence of atmospheric phenomena was recognized, together with the presence of dark and bright markings at the surface, and polar caps with large seasonal variations. The first attempt to compile a complete map was made by W BEER and J H MÄDLER in 1840. At this time, the dark areas were generally thought to be seas, following an assumption of W Herschel. Emmanuel Liais made the suggestion, in 1860, that they could be large patches of vegetation, thus explaining the changes in hue and brightness; this idea was widely accepted by the end of the century.

The modern era began with the close opposition of 1877, when G SCHIAPARELLI drew the first map based on precise measurements of surface features. The resulting map was much more detailed than any previous one, with old 'continents' resolved as intricated series of 'islands' and 'bays', so Schiaparelli had to abandon the feature names in use and introduced the foundation of modern nomenclature. This map was also the first one to report numerous 'canals'. Canals, previously mentioned by SECCHI and DAWES, were the first features reported on Mars with no terrestrial equivalent. They remained at the center of discussions about the Martian surface and habitability for several decades, mainly because of their interpretation by Percival LOWELL as artificial constructions. Although this interpretation was always debated, a majority of Mars observers (including C FLAMMARION, O von STRUVE, W Holden, W PICKERING, …) were convinced of the reality of the canals. It was only after E ANTONIADI's observations in 1909 that canals were largely recognized as an interpretation of a very subtle surface structure through the constantly moving terrestrial atmosphere; however, the matter was definitively settled only with the first spacecraft observations in the 1960s.

Planetary astronomy was relatively neglected in the US after the First World War, partly because of the long and sterile controversy about the Martian canals, but it remained vivid in Europe. Along with continuing detailed observations (mainly by Antoniadi, Focas, Maggini, SLIPHER, …), this period saw the first conclusive physical studies of the planet. The classical assumption that Mars was covered with seas did not survive the first atmospheric studies (1920–1950): both pressure and temperature are too low to permit the existence of liquid water. Concerning the surface, polarimetric studies in the 1940s (B LYOT, A Dolffus) suggested that the bright areas are covered by mineral dust with composition dominated by iron oxides; hydrated oxides were actually observed in 1964 by infrared spectroscopy (V Moroz). The assumption that dark areas are covered with vegetation was severely challenged by spectroscopic measurements in the 1950s (Sharonov, Sinton, Vaucouleurs, Kuiper, …) and more so by radar observations in the early 1960s (Goldstein and Gilmore, Kotelnikov) that yielded the first topographic measurements.

The emerging picture of Mars at the eve of the space era was very different from the conceptions of Lowell and Flammarion and resembled the modern view. However, the first close-up images, returned by Mariner 4 in 1965, surprised all Mars observers: the surface appeared much more Moon-like than Earth-like. Craters were never observed from the ground and seldom considered, and they appeared to be the most prominent surface features. Although Mariners 6 and 7 in 1969 confirmed this new view of Mars, they also evidenced more evolved processes in chaotic terrains and uncratered

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

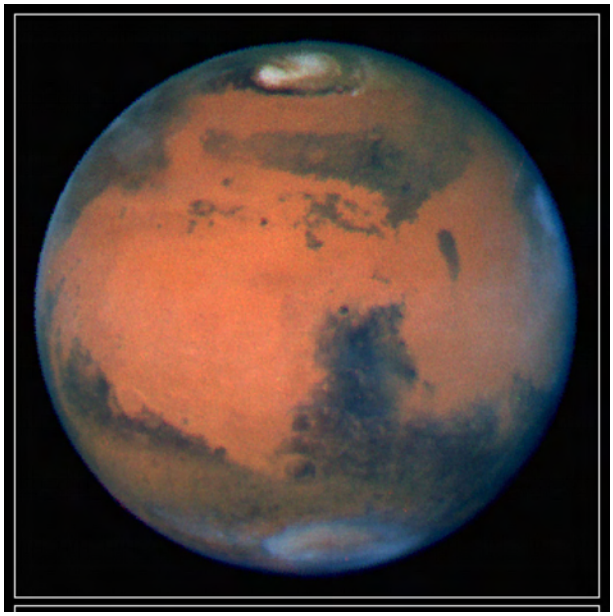


Figure 1. This view of Mars was taken with HST's Wide Field Planetary Camera-2 (WFPC2) on 10 March 1997, just before opposition, when Mars was about 100 million km from the Earth. The resolution is 22 km at the disk center. The image shows the planet during the transition between spring and summer in the northern hemisphere. The annual north polar carbon dioxide ice cap is rapidly sublimating, revealing the much smaller permanent water ice cap, along with a few nearby detached regions of surface frost. Bright water ice clouds can be seen in the giant impact basin, Hellas (near bottom of the image). The largest impact craters can be seen on such images (e.g. Schiaparelli on the left). The dark region at the center is Syrtis Major, a low volcanic shield. (David Crisp and the WFPC2 Science Team (Jet Propulsion Laboratory/California Institute of Technology).)

plains, but missed the more characteristic areas of the planet (see MARINER MISSIONS). The whole complexity of Mars became apparent only in 1971 with the first fully successful orbital spacecraft, Mariner 9. The VIKING MISSION, which consisted of two orbiters and two landers, was then operated for 6 yr (1976–1982). Limited but important additional information was provided by Mars 4 and 5 (1973–1974), PHOBOS-2 (1989) and MARS PATHFINDER (1997).

The main body of our current knowledge of the Martian surface still derives from Viking observations. Most of this information is related to surface morphology however, not directly to composition; even the geological maps are actually based on photointerpretation, while the nature of the soils is still poorly known. In this respect, two other sources of information are particularly important: SNC METEORITES, thought to come from Mars, and telescopic spectral observations in the near-infrared (pioneered in the 1960s by Moroz, Adams, and McCord). SNC meteorites provide direct compositional information but are not as useful as collected samples, since their source areas are unknown. High spatial resolution spectroscopy is less conclusive but provides the opportunity to understand the relationship between morphology and mineralogical composition over large areas, thus hinting at the processes that shaped the surface and at their succession in time.

Major features and main processes

Physiography

The most striking features on Mars, as seen from the ground, are the opposition between bright and dark

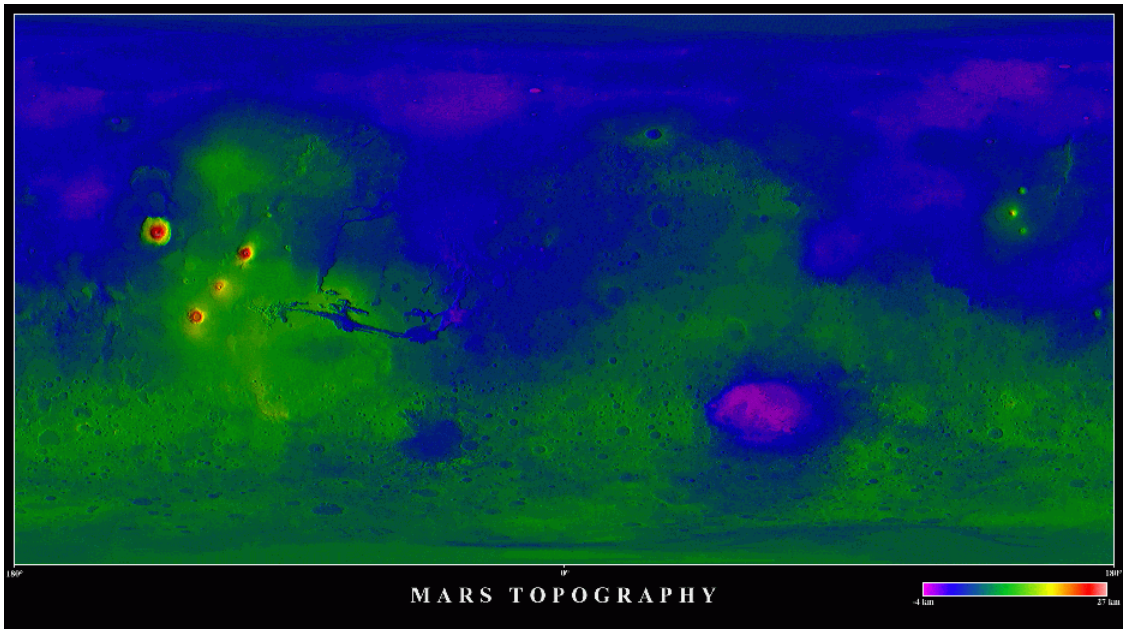


Figure 2. Topographic map of Mars. Derived from USGS, processed by Tayfun Oner.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

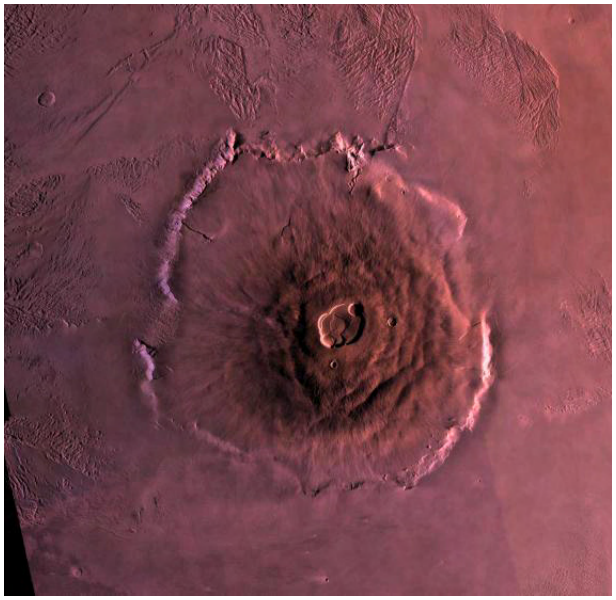


Figure 3. Five giant volcanoes are present on the surface of Mars, including Olympus Mons, the largest known volcano in the solar system. It is 27 km high and 550 km at its base, and the caldera is 90 km across and ~3 km deep. The volcano is surrounded by an outward-facing cliff up to 6 km high, recovered in places by lava flows. The plains surrounding the volcano are covered by the ridged and grooved 'aureole' of the volcano, extending 700 km to the northwest. The origin of the aureole is controversial: it may be related to gravitation-induced sliding of material from the flanks of an ancestral volcano or to a pyroclastic phase in the volcano's history. The small density of craters implies that the last periods of activity are relatively recent. (Mosaic of Viking images, NASA.)

areas, and variations in colors (figure 1). The orbital view is rather different, because local contrast is dominated by shadows, so that relief becomes apparent. The major morphological features consist of the following.

- There is a topographic and physiognomic asymmetry between hemispheres, often referred to as the Martian 'dichotomy' (figure 2). Old, densely cratered highlands are concentrated in the south, and younger low volcanic plains in the north. The difference in altitude is on the order of 3 km in general.
- A broad topographic bulge, Tharsis, is centered at the equator. This region is 8000 km wide and 10 km high and is covered by both old and young terrains. This structure has no equivalent in the solar system. It is surmounted by three huge volcanoes rising another 15 km or so in altitude.
- There are five enormous volcanoes, three of which are aligned on top of Tharsis and another one (Elysium Mons) is on a smaller bulge. The fifth one, Olympus Mons (figure 3), is the largest volcano in the solar system and rests in a low plain. It reaches 27 km at its highest point. The five giant volcanoes are built from fluid basaltic lavas and are similar in type to the Hawaii volcanoes.
- A huge canyon system, Valles Marineris, extends eastward from Tharsis along the equator (figure 4). The complex is 4000 km long, and reaches 600 km wide and 7 km deep in its central part. The canyons appear to be related to faulting and constraints that followed Tharsis formation.
- Channels tens of kilometers wide and hundreds long exist, similar to those produced on Earth by catastrophic flooding (figure 5). Most of them originate in the chaotic terrains east of Valles Marineris, cross the old terrains and merge with the low plains at 45° N.

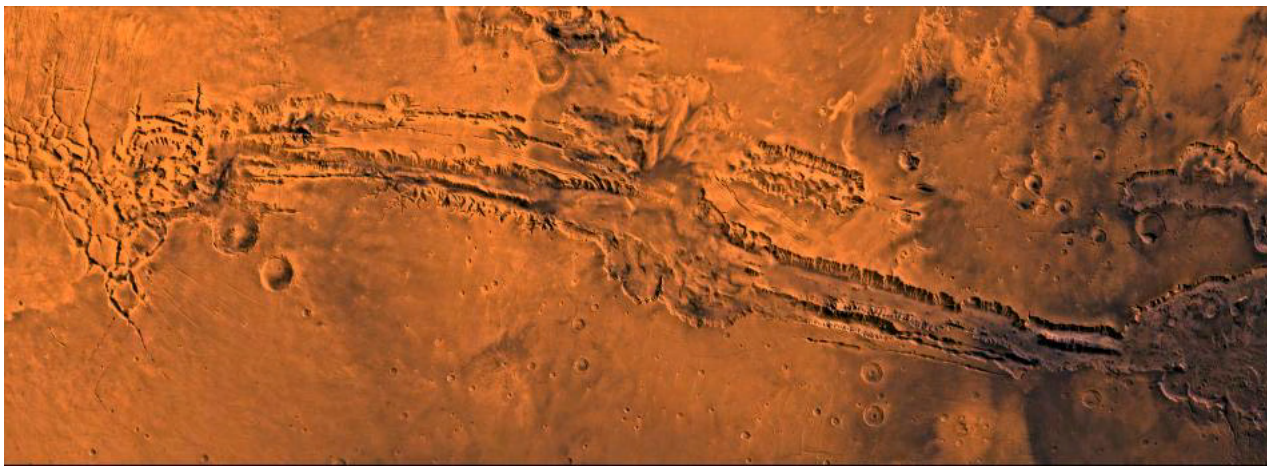


Figure 4. Synoptic view of Valles Marineris, the great canyon of Mars. The scene shows the entire canyon system, over 3000 km long and averaging 8 km deep, extending from Noctis Labyrinthus, the arcuate system of graben on top of the Tharsis bulge, to the chaotic terrains to the east. The connected chasmata of Valles Marineris originated as tectonic structures, and have later been modified by other processes, most notably mass wasting. The central and eastern parts contain layered materials and evidences of erosion by liquid water. Huge channels began from Valles Marineris and from adjacent canyons and ran north. Many of the channels flowed north into Chryse Basin, which contains the Viking Lander 1 and Mars Pathfinder sites. (Mosaic of Viking medium-resolution images in black and white and low-resolution images in color, Mercator projection.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

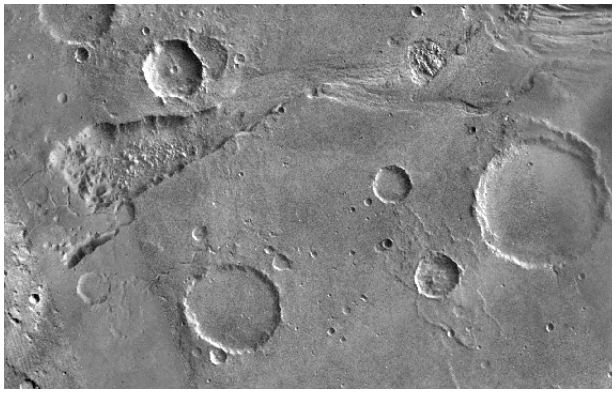


Figure 5. Chaotic terrains and outflow channels are concentrated east of Valles Marineris. This image shows a 300 km long portion of Ravi Vallis. Like many other channels, Ravi Vallis orginates in a closed depression with no apparent tributaries, suggesting that the water was released under great pressure from beneath a confining layer of frozen ground. As the water was released and flowed away the overlying surface collapsed; three such regions of collapsed material (chaotic terrains) are seen in this image, connected by a channel whose floor was scoured by the flowing water. Structures in these channels indicate that they were carved by liquid water moving at high flow rates. Ravi Vallis connects with a system of channels that ultimately flowed northward into the Chryse Basin. A small crater (~10 km diameter) with fluidized ejecta can be seen at the bottom of the image. (Viking mosaic from the Mars Digital Image Map, image processing by Brian Fessler, Lunar and Planetary Institute.)

- Polar deposits are 2–3 km thick with very thin layering (<20 m), cut by spiraling valleys. In the north they are surrounded by a vast, uncratered, dune field.

Although most of these constructs have their equivalent on Earth, they reach enormous dimensions on Mars. Conversely, some of the major terrestrial features do not have their counterpart on Mars, in particular folded mountains and mid-ocean ridges; sedimentary deposits are rare compared with the Earth, while 60% of the present surface is made of volcanic materials. These characteristics derive directly from the largest two differences between Mars and the Earth: their size (Mars' radius is only about half the Earth's, and its mass is one-tenth) and their distance to the Sun (Mars is 1.52 times further).

Main processes

Because of the smaller size the internal temperature of Mars never reached that of the Earth, and cooling was also more effective. In these conditions, the crust thickened and stabilized early in the history of the planet so that plate tectonics, the most effective resurfacing process on Earth, never developed on Mars. Therefore, very old terrains still compose a large part of the present Martian surface: the cratered highlands formed 3.8 Gyr ago, when impacts were the dominant process in the solar system, and are globally well preserved. The enormous size of the Martian volcanoes is another consequence of the absence of plate tectonics: all the material accumulates on the same spot, instead of building series of constructs as the plate shifts above the magma source, like in the Hawaii islands. There is evidence that the volcanoes have accumulated during a long part of Mars history but their surfaces exhibit very few craters, implying that the last eruptive events are relatively recent (~100 Myr?). Other types of volcanoes exist on Mars, in particular small domes and wide low-relief shields (paterae) composed mostly of ash deposits; most plains were also formed by successive deposits of lava flows all along the history of the planet. Although the period of volcanic activity was much longer than on the Moon, volcanism is probably exhausted today. Overall, it was the most widespread and effective process that contributed to the formation of the surface after the period of heavy bombardment.

The small size of Mars has resulted in a short period of geological activity and gas recycling, and the large Sun distance has not permitted a runaway greenhouse effect like that on Venus (see GREENHOUSE EFFECT). The combination of these has resulted in an atmosphere thinner than the Earth's and indirectly in a lower surface temperature and hence in different evolution processes. In particular, liquid water is unstable at the present time, so the most active surface processes on Earth (weathering, aqueous erosion and sedimentation) are only marginal on Mars. For this reason relief created by volcanism, tectonics or impacts remains largely unmodified, and considerable relief exists on Mars despite the small size of the planet: the overall altitude dynamic reaches 37 km, twice the Earth value. The only regulation is provided by the capacity of the crust to sustain the corresponding load, which explains the importance of fracture systems (in particular around Tharsis). On the other hand, the lack of liquid water also has resulted in limited sedimentary processes. Instead, the surface currently evolves mainly under the influence of winds and, to a lesser extent, landslides. Winds related to seasonal sublimation and condensation of atmospheric CO_2 on the polar caps are strong enough to cut the spiraling valleys in the polar deposits and form dunes fields at high latitudes. The polar deposits are successions of layers of ice and dust with proportions related to climatic variations; the absence of craters in these terrains implies that they are constantly modified at the present time. Other wind-related features include erosional and depositional streaks behind craters and erosional forms at lower latitudes (yardangs). Active dust deposition occurs in the northern hemisphere, with net south-to-north transport; this pattern is probably reversed on various time scales because of periodic changes in obliquity and eccentricity (see SOLAR–TERRESTRIAL CONNECTION: SHORT-TERM AND LONG-TERM CLIMATE VARIABILITY). The dust currently seems to be remobilized from southern areas rather than

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

formed from erosion of fresh surfaces, so that the fine materials are generally thought to be homogenized on a global scale and decoupled from the bedrock. Variations of surfaces markings observed from the ground are mostly due to aeolian transport of fines.

Aeolian processes appear to have been significant all along the history of Mars. However, there are numerous indications that other processes were active in a more or less distant past, in particular those related to liquid water and water ice, providing important clues about former climatic conditions and the degree of chemical and perhaps biological evolution.

In 2001, high-resolution pictures taken by the Mars Orbiter Camera on NASA's Mars Global Surveyor spacecraft show distinct, thick layers of rock within craters and other depressions, suggesting that they may have formed in lakes or shallow seas. Such layered rock structures where there were once lakes are common on Earth. The pancake-like layers of sediment have compressed and cemented to form a rock record of the planet's history. The regions of sedimentary layers on Mars are spread out and scattered around the planet. They are most common within impact craters of Western Arabia Terra, the inter-crater plains of northern Terra Meridiani, the chasms of the Valles Marineris, and parts of the northeastern Hellas Basin rim.

Although Martian highlands are very similar to lunar ones, these areas are significantly more evolved on Mars: they are dissected by valley networks and include smooth areas and craters with smoothed shapes. These differences reflect later evolution under the action of erosive processes and highland volcanism in the early stages of the planet's history. Valleys (figure 6) are particularly interesting since they imply the presence of running water at the surface in early times, and consequently a very different climate with higher temperature and pressure. At latitudes higher than 40°, highland craters are more rounded and degraded, and all escarpments are surrounded by lobate debris aprons, suggesting relaxation of soils with a large water ice content (permafrost). This is consistent with present climatic conditions, where underground water ice is potentially stable at high latitudes but not near the equator; accordingly, this process could still be active nowadays. Initial data from NASA's Mars Odyssey gamma ray spectrometer instrument suite has been used to detect hydrogen, which indicates the

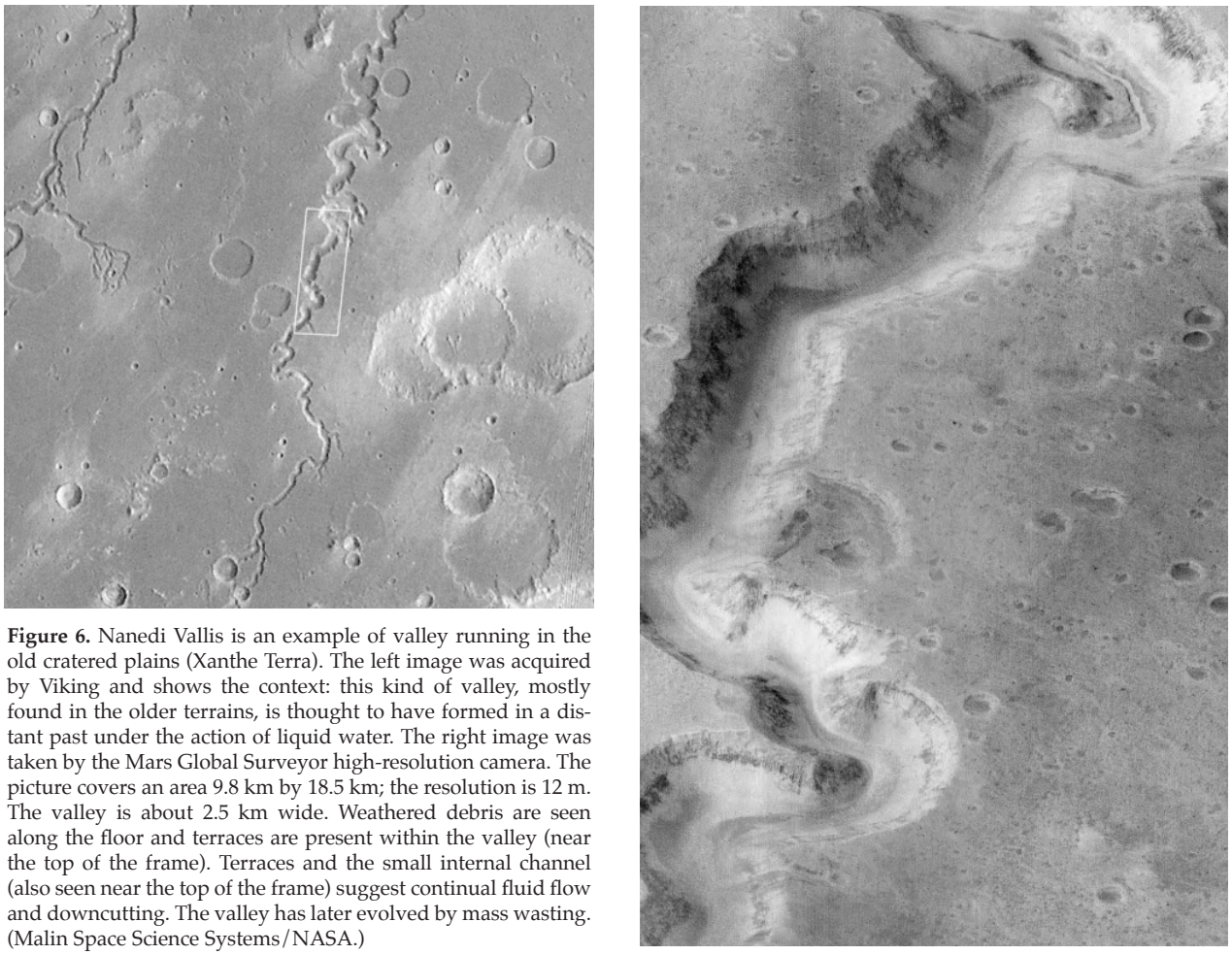


Figure 6. Nanedi Vallis is an example of valley running in the old cratered plains (Xanthe Terra). The left image was acquired by Viking and shows the context: this kind of valley, mostly found in the older terrains, is thought to have formed in a distant past under the action of liquid water. The right image was taken by the Mars Global Surveyor high-resolution camera. The picture covers an area 9.8 km by 18.5 km; the resolution is 12 m. The valley is about 2.5 km wide. Weathered debris are seen along the floor and terraces are present within the valley (near the top of the frame). Terraces and the small internal channel (also seen near the top of the frame) suggest continual fluid flow and downcutting. The valley has later evolved by mass wasting. (Malin Space Science Systems/NASA.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

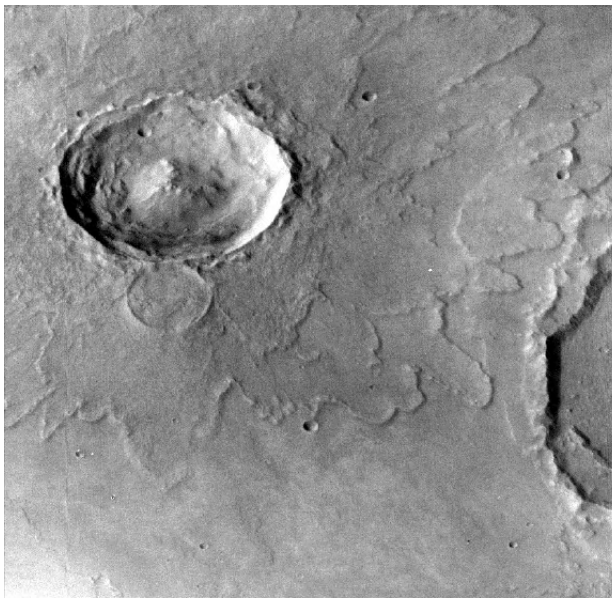


Figure 7. Rampart craters represent a special cratering style found only on Mars: the ejecta deposits consist of one or several sheets with either rounded or lobate contours, sometimes ending with a pedestal escarpment. The 18 km crater Yuty in this figure exhibits overlapping ejecta sheets with complex lobate forms. This type of ejecta morphology is found mostly equatorward of 40°, whereas circular ejecta sheets are frequent poleward of 30°. Rampart craters are believed to form when the impacting object rapidly melts subsurface ice. The presence of liquid water in the ejected material allows it to flow along the surface, giving the ejecta blanket its characteristic, fluidized appearance. The spatial distribution of rampart craters of a given size permits to map the depth of the permafrost. (Viking Orbiter image 3A07.)

presence of water ice in the upper meter (three feet) of soil in a large region surrounding the planet's south pole. The amount of hydrogen detected indicates 20 to 50% ice by mass in the lower layer. Because rock has a greater density than ice, this amount is more than 50% water ice by volume.

Martian craters have another unique particularity: many of them have ejecta blankets lobate in form, clearly due to fluidized materials (figure 7). Although they may result from interactions with the atmosphere during ballistic emplacement, these shapes are better explained by impacts in soils with high water content, with ejecta emplaced as radially directed surface flows. Other glacial forms exist in particular at high latitudes, including thermokarst, patterned grounds, pingos and ridges similar to glacial moraines. These forms suggest that ice has played an important role in the evolution of the surface; they also put constraints on the distribution of ground ice along the planet's history.

Other major landforms related to water are found in the canyons, chaotic terrains and outflow channels. Outflow channels are morphologically different from highland valleys and much younger. Some of them originate in closed collapse zones, which suggests that they are related to sudden discharges of underground water under pressure. Typical forms are present in the channels, in particular streamlined islands and scour marks, that are reminiscent of catastrophic flood features on Earth such as the Channeled Scablands, Washington state, although on a much larger scale. The eastern and central parts of Valles Marineris contain thick layered terrains that were extensively eroded after deposition (figure 8). There is evidence that the layered terrains formed under ice-covered lakes filling the canyons; the water could have been released to the eastern chaotic terrains and channels during catastrophic events (landslides, impacts, …). Several episodes of filling and release of such lakes may have occurred in the canyons. However, various hypotheses have been proposed to explain the evolution of Valles Marineris on the basis of orbital images, and a knowledge of rock composition (volcanic versus sedimentary origin) is required to choose among them. The presence of lakes at low latitudes, even under a frozen surface, could imply an important change in climatic conditions in a relatively recent past. A much more speculative topic concerns the existence of an ocean in the distant past: it has been suggested on the basis of Viking images that the contact between highlands and plains in the northern hemisphere is an ancient shoreline contemporaneous with the layered deposits and outflow

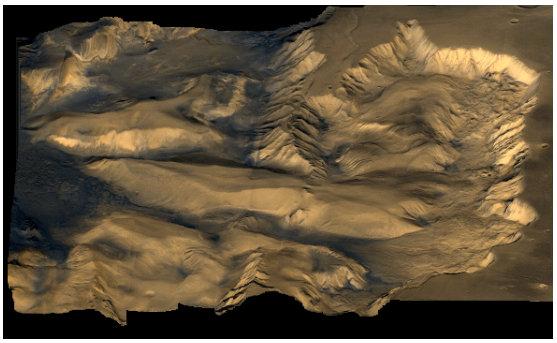


Figure 8. Oblique view of Candor and Ophir Chasmata in central Valles Marineris, from the east. The divide between the two canyons is a remnant of the plateau in which the canyons are cut. This part of the canyon was filled with materials that were later deeply eroded. Two different kinds of layering are visible in the image: in the canyon walls (successive episods of volcanism that formed the plains) and in the mesa-like formations inside the canyon (possible lacustrine deposits). The north–south direction and the gap between the two canyons may indicate drainage of material by catastrophic release of water. A variety of processes have contributed to the present landscape, including tectonics (the canyon itself), mass wasting (canyon walls, top right), wind (the fluted slopes of the layered terrains) and perhaps water and volcanism (layered terrains and canyon floors). (Composite of high-resolution black-and-white images from Viking 1 and color images from Viking 2. Image processing by Jody Swann, Tammy Becker and Alfred McEwen, USGS Flagstaff.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

channels. Radar observations from MARS GLOBAL SURVEYOR show that the possible contact line is close to an equipotential surface (as expected from a sea level) and that the terrains are very smooth below this level, thus supporting the idea of an ancient ocean.

Volatiles inventory and past climates

There is a striking paradox in the extent of water-related landscapes and the present amount of water in the atmosphere, which is equivalent to a layer of ~15 μm precipitated on the surface. In contrast, a layer about 500 m thick seems to be required to form the valley networks observed in the highlands (versus 3000 m present on the Earth). This estimate conflicts with much smaller quantities derived from planetary formation models and noble-gas measurements (10–100 m) but seems more reliable. Hence, most of the water has been stored in surface reservoirs for several billion years or has escaped the planet at the end of the heavy bombardment period. The residual polar caps and the regolith can store only a small amount of water. Conversely, the permafrost could retain up to several hundred meters of water, although no realistic estimates can be made of the actual amount tied up this way. At present, the water stored in the reservoirs can be released in the atmosphere only if the temperature is high enough, which is possible only if a sufficient amount of greenhouse gases, namely CO_2, is present.

Accordingly, periods of large water erosion indicate high temperature and high atmospheric pressure. The CO_2 pressure necessary to maintain liquid water on Mars is consistent with estimates of outgassing and is on the order of 1–5 bar. In such a situation, similar to the present Earth's, atmospheric CO_2 is continually depleted by formation of carbonate minerals in liquid water and refurbished by recycling of buried rocks through volcanic and tectonic processes. This process is ultimately controlled by the planetary heat flow. In the absence

Table 1. Tentative chronology of surface evolution.

Epoch	*Absolute age range (Gyr)*		*Events*	*Main processes*	*Proportion of present surface (%)*
	Hartmann–Tanaka	*Neukum–Wide*			
Late Amazonian	0.25–0.0	0.7–0.0	Aeolian resurfacing Final activity of Tharsis volcanoes Final evolution of Valles Marineris	Aeolian	7
Middle Amazonian	0.7–0.25	2.5–0.7	Polar layered deposits Outflow channels are still active	Volcanic Aeolian Periglacial	8
Early Amazonian	1.8–0.7	3.55–2.5	Late volcanic plains Erosion of layered terrains in Valles Marineris	Volcanic	11
Late Hesperian	3.1–1.8	3.7–3.55	Giant volcanoes Relatively hot climate? Liquid water may be stable or metastable Layered terrains in Valles Marineris Outflow channels and chaotic terrains	Volcanic Fluvial Periglacial	19
Early Hesperian	3.5–3.1	3.8–3.7	Important development of Tharsis and Elysium Earliest central-vent volcanism (highland paterae) First volcanic plains Development of Valles Marineris	Volcanic	16
Late Noachian	3.85–3.5	4.3–3.8	Hot and humid climate? Then dissipation of early atmosphere and cooling Highland valleys, fluvial resurfacing Highland volcanism	Volcanic (flood) Fluvial	12
Middle Noachian	3.92–3.85	4.5–4.3	Early development of Tharsis Highland formation and large basins Formation of megaregolith and sediments	Impact Volcanic (flood)	24
Early Noachian	4.6–3.92	4.6–4.5	Planetary formation First mantle differentiation Outgassing of early atmosphere	Impact Volcanic (flood)	4

Absolute ages are given according to the two time scale models indicated (Hartmann–Tanaka, Neukum-Wise). Adapted from Tanaka *et al* 1993 in Kieffer *et al* (1993).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

of plate tectonics CO_2 recycling was never very effective on Mars, but the early warm and wet period could have lasted hundreds of millions of years, and could have favored rapid chemical reactions, possibly leading to biological evolution. With the secular reduction of the heat flow, this mechanism slowed down: recycling of CO_2 became less efficient, pressure and temperature decreased to the freezing point of water, water was stored mainly as ice under the surface, the remaining CO_2 was trapped in surface carbonates by slower processes, and to a lesser extent adsorbed in the regolith and frozen in the polar caps. The above model is largely theoretical, however; an alternate model states that the dense CO_2 atmosphere was dissipated during large-impact events, in which case carbonates would not have formed in large quantities (see MARS: ATMOSPHERE and PLANETARY ATMOSPHERES). For this reason an estimate of the abundance of carbonates, which have not been positively identified either on the surface or in the airborne dust so far, would give important constraints on the evolution of the primitive atmosphere.

Geological history

Ages of planetary surfaces are estimated through crater counts by comparison with lunar terrains, whose ages are known precisely from the Apollo samples (see CRATERS). Transposing this kind of dating to Mars involves making a hypothesis on the meteoritic flux, so that absolute ages will remain very uncertain until samples are returned from Mars. However, the relative age of the terrains can be estimated from photointerpretation alone. The time scale is divided in three main systems, Noachian, Hesperian and Amazonian, and eight subsystems. Table 1 summarizes the chronology of surface evolution.

Close view

The view from the surface is currently known only by images returned by the three American landers, Viking Landers 1 and 2 and Mars Pathfinder (figure 9). The two Viking landing sites were selected on the basis of minimal surface roughness inferred from the orbital images and are very similar to one another. They mostly consist of rocks standing in a layer of fine debris; VL1 site is less monotonous, with large patches of fines accumulating in dunes from place to place. Rocks range from centimeters to meters in size and are in general angular; they are thought to be ejectas excavated by nearby impacts. Dust particles are less than 10 μm in size but often aggregate in clods looking like rocks. The two Viking landers observed slow movements of the dust layer during their two Martian years of functioning, but little evidence for evolution of the rocks although climatic variations are large. In contrast, the Mars Pathfinder landing site was selected for its geological interest: it is located in an outflow channel east of Valles Marineris, and hence it was hoped to contain materials coming from different source regions. Although it is very similar to the previous two sites the landscape shows some possible evidences of flood but mainly demonstrates again that wind has been the major process at least in these areas for a huge period of time. Most of the rocks are volcanic in origin, but some rocks at the Pathfinder landing site show faint evidence of layering, consistent with a sedimentary origin.

All three landers performed basic chemical analyses of the surface materials. Viking landers analyzed the dust and the duricrust, while Pathfinder's robot, Sojourner, also analyzed rocks. Martian soils are consistent with weathering products of basalts, although they are richer in Fe and Mg, and poorer in K, than most terrestrial basalts; the most probable mineralogical compositions are mixtures of clays and ferric oxides. Duricrust samples have higher S content than the loose fines, suggesting formation by evaporation of salt-rich solutions, with sulfates as the cementing agent. All soils are very similar at the three landing sites, with high degrees of oxidation and high magnetic properties. Pathfinder rocks have composition similar to terrestrial andesites, with relatively high silicon contents. These rocks are obviously not representative of the diversity of Martian rocks, since

Figure 9. Enhanced image from Mars Pathfinder. The image is a mosaic of 300 camera images covering 360° of azimuth and elevations from approximately 4° above the horizon to 45° below it. Dark and bright rocks and bright soils are visible in the image. Dark rocks have elemental composition indicating more silica-rich basaltic andesite. These rocks are typically the small boulders and intermediate-sized cobbles at the Pathfinder site. Bright rocks have elemental spectra indicating basaltic compositions. These rocks are typically larger than 1 m in diameter and many display morphologies suggestive of flood deposition. Bright soils have a strongly reddened spectrum indicative of oxidized ferric minerals. Other materials are also apparent, including pink rocks with morphology indicating a cemented or rock-like structure. This material may be a chemically cemented hardpan that underlies much of the Pathfinder site. (Local solar time increases nearly continuously from about 10:00 at the right-hand edge to about 12:00 at the left.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

they are compositionally different from the dust, and have a much higher Al content.

Mineralogical and chemical compositions are distinct: for a given chemical composition, mineral formation depends strongly on the thermodynamical conditions, and minerals are therefore very useful indicators of past environments and formation processes. No mineralogical analysis has been performed at the surface so far, so the most direct source of information in this regard is provided by the SNC meteorites. These meteorites are believed to be igneous Martian rocks ejected during impacts, based on their young crystallization ages and close match between their gaseous inclusions and the Martian atmosphere. SNCs are not as informative as samples collected *in situ* would be, since there is no precise knowledge of their origin and thus little understanding of how representative they are of Martian crustal materials. These meteorites can be grouped in three large families, a younger one composed of basaltic rocks dominated by two pyroxene mixtures (Shergotty, Zagami, … formed ~180 Myr ago) and an older one dominated by Ca-rich pyroxenes and olivine, with very low concentrations of Al (Chassigny, Nakhla, …formed ~1.3 Gyr ago); ALH84001 is much older (~4.5 Gyr) and has a different origin. These three groups apparently have been ejected during three different impact events, possibly from the Tharsis region. Like materials analyzed *in situ*, all SNCs are richer in Fe than the terrestrial crust, probably reflecting a higher FeO content of the Martian mantle. In contrast with the rocks analyzed by Pathfinder, the SNCs are all poor in Al; hence they derive from an Al-depleted source region and probably formed later, but at least 1.3 Gyr ago. Compositionally, the global dust could derive from a mixture of Mg-poor SNCs (Nakhla, Shergotty, …) and rocks observed at the Pathfinder site. The occurrence of carbonates in SNC meteorites, together with a few ground-based observations at high spectral resolution, suggests that these minerals are indeed the major crustal sink for primitive CO_2. Alternatively, carbonates and other salts could have formed in restricted areas where volcanic activity resulted in a temperature high enough to maintain liquid water at the surface (hydrothermal springs). Globules of carbonates contained in ALH84001 were also presented as a possible byproduct of ancient life on Mars: microscopic shapes included in these globules were tentatively interpreted as fossil bacteria that lived on Mars 1.4–3.6 Gyr ago.

Global surface units

The surface exhibits marked contrast with respect to many observational parameters. Morphological units are commonly interpreted as geological provinces. Surface units defined by other measurements do not necessarily correlate with them, owing to the thickness of the layer scanned by different methods: visible and near-infrared spectroscopy analyzes the first few microns only and gives information on mobile materials, crusts, veneers. Mid-infrared and radar measurements penetrate to a deeper layer, from which rock abundance can be estimated. Multiple observations at different wavelengths help us in understanding the composition and physical state of the materials present on the surface.

As viewed from the ground, the major opposition is between bright and dark regions, with the polar caps constituting a third type. In violet to green light, these soils are equally dark (geometric albedo ~5%) and the surface is almost completely featureless. At longer wavelength, the contrast increases very rapidly and reaches a factor of 4 in the near-infrared (from 0.7 to 2.7 μm). All soils are actually very red in color, the greenish hue reported by 19th century observers being mostly a contrast effect. Seasonal color variations are largely due to scattering by airborne dust and are related to dust storms. At higher spatial resolution, another type of unit with intermediate reflectance and color can be recognized on Viking and HUBBLE SPACE TELESCOPE images. This unit includes bright crater rims and dark patches, so it was first interpreted as a mixture of dark and bright materials; it is currently interpreted as bright materials indurated by local processes. The Viking cameras have also shown a small-scale variability within dark areas, while bright regions are more uniform.

Other important information is provided by mid-infrared radiometry and radar observations, from which thermal inertia and rock abundances can be derived. Rock abundances estimated this way have been confirmed at the Pathfinder site. Four main units have been identified on the basis of all these physical properties at latitudes lower than 60°.

- Bright, fine materials are interpreted as loose dust. The corresponding areas are regions of active dust accumulation several meters thick (e.g. Amazonis, Arabia).
- Dark, coarse materials are interpreted as exposed rocks with constant deposition and removal of a thin dust layer (e.g. Acidalia, SYRTIS MAJOR). In some areas, however, this unit seems to correspond to accumulation of dark sands.
- Intermediate materials with high thermal inertia are interpreted as indurated surfaces (cemented dust).
- Bright materials with high thermal inertia are interpreted as areas of limited dust accumulation. Both Viking landing sites are located in such regions, which are rather uncommon (e.g. Chryse Planitia, Elysium Planitia).

Most units appear relatively young and active, indicating that the upper surface is continually being reworked, in particular by aeolian processes and duricrust formation.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Concerning composition, the most effective method to retrieve information on a global scale is by infrared imaging spectroscopy in the 0.6–20 μm range. Minerals, and therefore rocks, can be identified from their infrared spectra even at modest spectral resolution. Conversely, high spatial resolution is critical to detect materials of limited exposures, e.g. produced by local processes; on Mars, mineralogical interpretation is also complicated by atmospheric absorption and scattering and by mixing of surface features in absorption and emission above 4 μm. Overall, the spectra of every region can be matched by more or less oxidized basalts, so that spectral characteristics are dominated by ferromagnesian ('mafic') volcanic materials and their alteration products. Bright regions are spectrally very homogeneous and are dominated by subtle features of hydrated ferric oxides, while dark areas have spectra typical of ferrous absorptions in mafic minerals such as pyroxenes and exhibit a greater variability. This is consistent with the interpretation that bright regions are covered by a thick layer of weathered dust mixed at the planetary scale, while dark areas have large bedrock exposures and are contaminated at various levels by bright materials.

The freshest dark equatorial regions have spectral characteristics indicating a volcanic mineralogy dominated by two pyroxenes with high and low calcium content (perhaps augite and pigeonite), similar to the basaltic SNC meteorites (e.g. Shergotty, Zagami). The range of ages for these regions indicates that the SNC analyses are relevant to large regions of Mars and to a large lapse of time (3.0–0.13 Gyr). Such materials are not common on the Earth's surface and are probably derived from a source region depleted in Al relative to the hypothesized original mantle composition. An interpretation of these observations is that a primary planetary differentiation occurred very early in the history of Mars (4.5 Gyr) and that the mantle reservoirs have not been remixed since. The depletion in Al could be explained by the presence of a crust enriched in Al, perhaps similar to the lunar highlands; such rocks were actually observed at the Pathfinder site, but their source area is unknown. The occurrence of significant amounts of olivine in dark areas would be inconsistent with the idea of an early differentiation. Although detection of olivine-rich materials was claimed in the past, these observations are not confirmed by recent data. Instead, most of the second-order variations are explained by small amounts of dust and perhaps thin alteration coatings of ferric materials on the rocks themselves.

In bright areas, the only mineral phase that has been positively identified so far is hematite, an iron oxide. Spectral features are subdued and indicate a low degree of crystallization in most areas, so that hematite is thought to occur mainly as extremely fine-grained, or nanophase, particles; furthermore, all bright regions have deep hydration bands, corresponding to an adsorbed- or bound-water abundance of ~1% with local and temporal variations. These materials are clearly alteration products of mafic minerals exposed in the dark areas; the present rate of weathering and dust formation is unknown but is certainly small, and hence most bright materials were probably formed in a distant past. The 'classical' rock analogs for bright regions are palagonites, which are formed by alteration of mafic volcanic glasses under low temperature, often by interaction with water ice. Spectral properties of palagonites are dominated in the near-infrared by nanophase hematite, although it is only a minor component dispersed in a silicate-clay matrix that can include other salts, such as sulfates. Overall, large amounts of palagonites are consistent with spectral observations, *in situ* analyses, thermodynamical considerations and composition of SNC meteorites. Most of this material could have formed during volcanic eruptions through a thick permafrost. The unusual magnetic properties measured by the landers may indicate significant amounts of maghemite, an iron oxide. Airborne dust, which should be compositionally similar to the bright surface materials, has been observed chiefly in the mid-infrared; its spectrum in this domain is more consistent with iron-rich clays.

Subtle local variations are observed in bright regions, indicating that some bright materials are not remobilized and mixed. Most of these variations are ascribed to differences in particle size or degree of crystallization, but some small areas could have high concentrations of other alteration materials: smectite clays, oxyhydroxides, iron sulfates, chlorides, hydrous carbonates, …. These minerals are potentially very helpful indicators of past climates, since they are the major sinks for H_2O, CO_2 and other volatile components such as nitrogen. Some of them have been tentatively detected in the airborne dust, but their spatial distributions and properties must be known to improve our knowledge of the surface evolution. Similarly, 'anomalous' dark materials provide important hints about the major puzzles concerning the surface. In particular, an unusual material observed in the layered deposits of Valles Marineris, with both mafic and hydration characteristics, suggests that hydrovolcanism was an important factor in the development of these areas.

Conclusion

Our present perception of Mars results mostly from the Viking mission, analyses of the SNC meteorites and ground-based spectral observations. Large gaps remain in our current understanding of the planet, however, while Mars is a key factor in the study of the solar system's evolution; in addition, the Martian origin of the SNCs is not definitely demonstrated although a large part of the present ideas concerning Mars derive from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

their study. Some of the most important open questions are related to chronology, composition and volatiles inventory: there is no absolute dating of the terrains; the amount of volatiles outgassed and the current reservoirs are unknown; the present rate of alteration and the main period of dust production are unknown. All these questions would be best addressed by collecting samples in well-characterized locations on the surface and by high-resolution orbital observations of particular areas such as hydrothermal deposits. The canyon walls also provide a natural traverse in the successive volcanic materials accumulated in the plains and a unique opportunity to probe the sequence of event leading to surface formation.

Other questions relevant to the geological evolution include the following. What are the current and past heat flows? How and when did the Tharsis bulge form? What is the history of Valles Marineris (its relations with Tharsis, nature of the layered deposits, extent of volcanic processes, fate of the materials evacuated)? Were there lakes or even oceans on Mars (when, and what is their relation to the channels)? How long did the warm and wet period last? What was the extent of prebiotic evolution, if at all? In this regard, the study of Mars could help understand the first stages of life's evolution on Earth.

NASA's Global Surveyor has revealed images which suggests the Martian climate is changing significantly, even today. The observations were made during a full Martian year, 687 Earth days and show that pits—often referred to as the 'Swiss cheese' terrain—at the southern polar ice cap of Mars have dramatically increased in diameter, indicating the material has evaporated rapidly compared to last year. Images of the south polar cap taken in 1999 were compared with images of the same locations taken in 2001, and it was discovered that pits had enlarged, mesas had shrunk, and small buttes had vanished. In all, the scarps that enclose the pits and bound the mesas and buttes retreated about 3 m in one Martian year. This rapid retreat of polar scarps can only occur if the ice is frozen carbon dioxide (dry ice). Retreat of scarps made of water ice is much slower and would not have been measurable from one Martian year to the next.

Some of these issues will be addressed by the ongoing Mars Surveyor program (NASA), the MARS EXPRESS mission (ESA), and the future Mars sample return program (NASA–CNES (see SOLAR SYSTEM: EXPLORATION)). *In situ* analyses and sample return should be performed in different geological settings, and their results extended to the whole surface by means of orbital remote sensing. Even in the space era, however, ground-based observations remain important. Large orbital observatories like HST and ISO, or the NGST project, have not changed this situation radically, since they are limited by observational schedules and solar pointing constraints. In a threefold approach where spacecraft yield exceptionally detailed but exceptionally rare data, and spaceborne observatories provide high resolution at limited viewing geometry on a year-to-year basis, ground- based observation is still the only method to monitor quick processes on Mars and to clarify the context for future spaceborne observations.

Bibliography

Beatty J and Chaikin A 1990 *The New Solar System* (Cambridge: Cambridge University Press)

Carr M 1996 *Water on Mars* (Oxford: Oxford University Press)

Cattermole P 1992 *Mars* (London: Chapman & Hall)

Kieffer *et al* (eds) 1993 *Mars* (Tucson, AZ: University of Arizona Press) (review papers concerning nearly all aspects of Mars, mostly at professional level; the first chapters present a very good review of historical observations and previous synthesis up to the late 1980s)

Sheehan W 1996 *The Planet Mars: A History of Observation and Discovery* (Tucson, AZ: University of Arizona Press) (accessible to the general reader, very well documented)

Stéphane Erard

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mars: Water and the Martian Landscape

Over the past 30 yr, the water-generated landforms and landscapes of Mars have been revealed in increasing detail by a succession of spacecraft missions. Recent data from the Mars Global Surveyor (MGS) mission confirm the view that brief episodes of water-related activity, including glaciation, punctuated the geological history of Mars. The most recent of these episodes seems to have occurred within the past 10 million years. These new results are anomalous with regard to the prevailing view that the Martian surface has been continuously extremely cold and dry, much as it is today, for the past 3.9 billion years. Interpretations of the new data are controversial, but explaining the anomalies in a consistent manner leads to potentially fruitful hypotheses for understanding the evolution of Mars in relation to Earth.

When the Mariner 9 spacecraft (see MARINER MISSIONS) went into orbit on 14 November 1971, the surface of Mars was shrouded in a global dust storm. Fortunately, by March of 1972, the atmosphere cleared and the true complexity of the Mars landscape was finally revealed to the spacecraft's vidicon cameras. In addition to the immense volcanoes of Tharsis, there was a great equatorial canyon system, Valles Marineris, named for the discovery spacecraft. Most remarkable, however, were sinuous channels and valleys, from whose morphology it was inferred that water had once flowed on the surface of this now dry, unearthly cold planet. The late Harold Masursky, science team leader for the vidicon imaging experiment, wrote in 1973 (Masursky 1973):

> The possible fluviatile channels may record episodes when water was much more abundant in the atmosphere than it is at present. Planet-wide warmer interglacial periods would release not only the water locked in the polar caps but also that frozen in the subsurface as permafrost. Similar warmer and colder periods also are characteristic of terrestrial history.

Although intended primarily as support to landers seeking evidence of Martian life, the Viking orbiters (see VIKING MISSION) of the late 1970s returned 52603 images of Mars, most of them at much higher resolution than the 7329 images returned by Mariner 9. The pictures with the highest resolution have a pixel spacing of 7.5 m, although most frames resolve to several tens of metres. Viking orbiter images (figure 1) provided the basis for an understanding of water and landscape that prevailed until the past few years.

In 1997, MGS spacecraft was inserted into Mars' orbit, but a variety of problems prevented it from achieving a circular mapping orbit until 19 February 1999. Two instruments are particularly relevant to the scientific study of the Mars landscape. The Mars Orbiter Camera (MOC) achieves a resolution of 1.4 m per pixel, but the required high data volume limits scenes to a kilometre or so in width at the highest resolutions. The Mars Orbiter Laser Altimeter (MOLA) maps the topography of Mars with a precision better than 10 m (Smith *et al* 1999). Together these instruments provide new data for studying Mars' landforms (figure 2), although human reasoning about those landforms remains a matter of long-standing scientific experience.

The study of Earth-like planetary surfaces—geomorphology—is not a disjointed collection of observational facts solely with which to test, or against which to constrain, theoretical models. Rather, such scientific inquiry proceeds from the informed colligation of landform observations to the discovery of consistency and coherence and, ultimately, to consilience in the theoretical accounting (explanation) of those observations. The key element of this inquiry is the formulation of one or more working hypotheses, which are most often suggested (but not proved) by analogies of form and context among landscapes of known origin and those under scrutiny. In the retroductive inferences of geomorphology, analogy

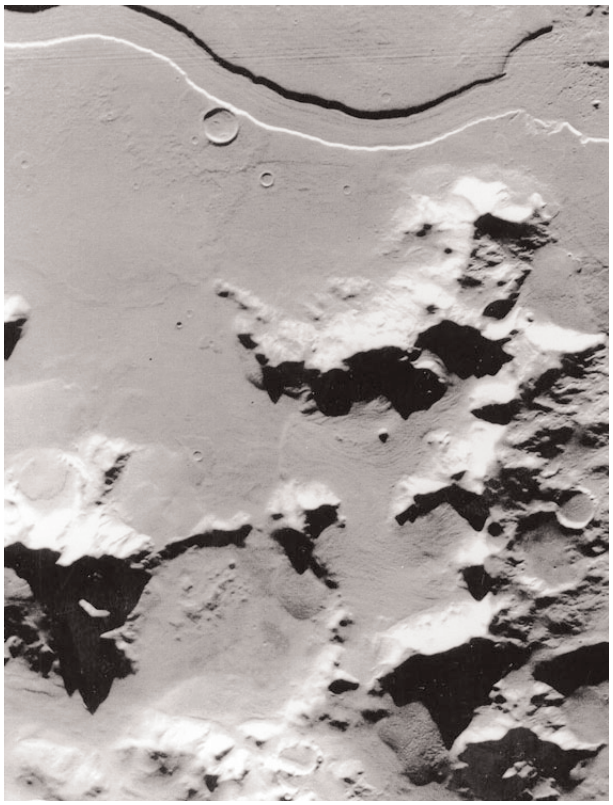


Figure 1. Glaciated terrain east of Hellas Planitia, at latitude 42° S, longitude 252° W.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

serves merely to suggest fruitful working hypotheses, thereby leading to completely new theories that bind together any newly discovered facts. Mars' landscape provides particularly stimulating opportunities to practise geomorphological reasoning, generating hypotheses that may initially strike some researchers as outrageous. Nevertheless, it is the productive pursuit of such hypotheses that leads ultimately to new understanding, not only of Mars, but also of Earth itself.

The surface of Mars is today extremely cold and dry. The atmosphere at the land surface is over 100 times less dense than that of Earth and it holds only minuscule amounts of water vapour. For the present obliquity (tilt of the planet's rotational axis with respect to the orbital plane) of 25°, the residual north polar ice cap sublimates water in the northern spring–summer and the vapour moves to and condenses at south polar areas, a pattern that may reverse in northern autumn–winter (Kieffer, Jakosky, Snyder and Matthews 1992). Given its minor role in comparison with that on Earth, it is not surprising that on present-day Mars, water is replaced by wind as the most continuously active surface-modifying process.

In stark contrast to the current environment, however, numerous landforms provide signs, or indicators, of extensive past activity of water and ice on the Martian surface (see MARS: CLIMATE HISTORY). From the densities of impact craters on the various terrains it is possible to work out a chronology for this activity, such that Mars is divided stratigraphically into ancient heavily cratered uplands that formed prior to about 3.5 billion years ago (the Noachian epoch), intermediate cratered plains (the Hesperian epoch) and more lightly cratered areas (the Amazonian epoch). Water and ice were active on the surface during all these periods. The mode, timing and long-term cycling of water in surficial processes are the phenomena to be considered from the various signs of activity.

Signs of subsurface water and ice

Landforms indicative of ground ice in permafrost (perennially frozen ground) on Mars have been known since the early flyby missions of the 1960s. Images from Viking orbiters provided an overwhelming list of permafrost and ground-ice indicators, including various kinds of patterned ground, thermokarst, hillslope features and mass-movement phenomena. Although originally attributed to permafrost processes, the immense polygons, 3–20 km across, of the northern Mars plains are much larger than the contraction-crack polygons typical of terrestrial permafrost terrains. These features are now explained variously as the tectonic uplift of basin floors, perhaps following removal of load from an overlying standing body of water, or as the result of Rayleigh convection driven by unstable density or temperature gradients in a catastrophic flood deposit positioned over frozen ground (Lane and Christensen 2000). Recently acquired very high-resolution MOC imagery reveals extensive areas of the northern plains and southern highlands of Mars, notably on crater floors (figure 3), where small-scale contraction-crack polygons, tens or hundreds of metres across, cover the landscape (Seibert and Kargel 2001). This polygonal terrain, which closely mimics the ice-wedge polygonal terrains of terrestrial permafrost areas, is essentially uncratered, indicating a surprisingly youthful phase of water-related activity on Mars.

Viking pictures revealed that many Martian craters have a unique morphology, different from that observed elsewhere in the solar system. Ejecta surrounding these craters are layered and each layer has an outer edge terminating in a low ridge or escarpment. Named 'rampart craters', the flow-ejecta morphology most likely represents the incorporation of groundwater and ground ice, although atmospheric effects on ejecta emplacement may also be important. Thus, layered ejecta morphologies can

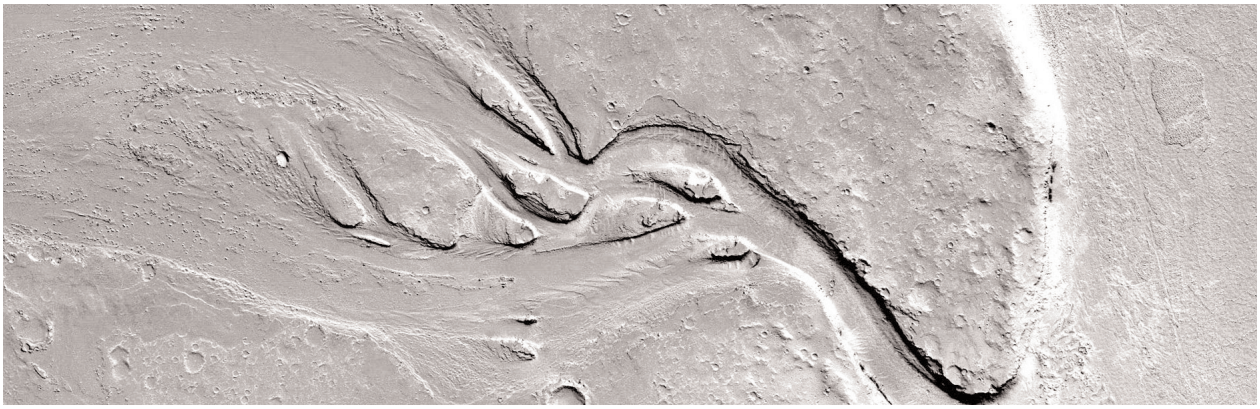


Figure 2. High-resolution MOC image of a fluvial channel system at latitude 7.9° N, longitude 205.8° W, south of Cerberus Rupes (MOC image M21-01914).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

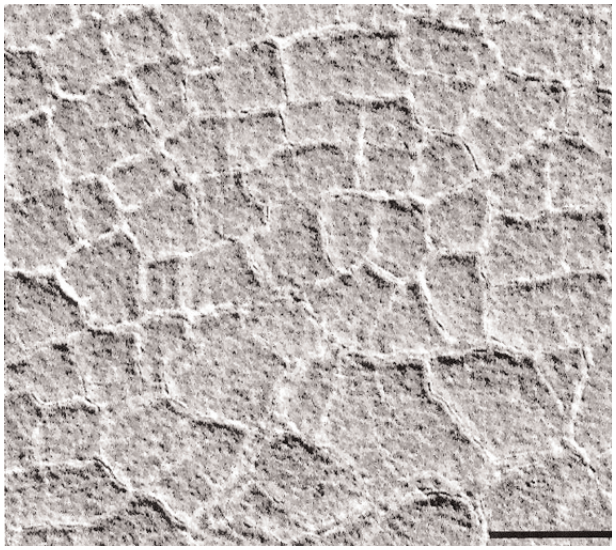


Figure 3. Large contraction-crack polygons developed on the floor of a northern-plains crater at approximately 67.5° N, 312.5° W (MOC image M01-00294).

be used to characterize the past presence of water in various Martian terrains.

First revealed by Mariner 9, the long and complex volcanic history of Mars contains a wealth of examples of interactions among volcanism, ice and water. Much as in the Pleistocene landscapes of Iceland, Martian volcanism has produced features interpreted as table mountains built up on products of subice eruptions, outburst flood channels and extensive pyroclastic landscapes. Some of the youngest volcanic landscapes occur in the Cerberus Rupes and Marte Vallis region (Hartmann and Berman 2000), where cataclysmic flood channels and volcanic lava flows occur in close spatial and temporal association. Even Olympus Mons, one of the largest known volcano constructs, has a morphology interpreted by some to be indicative of water/ice volcanic interactions. Its huge aureole deposits, extending 1000 km, may represent immense submarine landslides, similar to those of Earth's Hawaiian Islands.

Initial data from NASA's Mars Odyssey spacecraft suggests possible identification of significant amounts of frozen water in the upper meter (three feet) of soil in a large region surrounding the planet's south pole. The gamma ray spectrometer instrument suite has been used to detect hydrogen, which indicates the presence of water ice. The amount of hydrogen detected indicates 20 to 50 percent ice by mass in the lower layer. Because rock has a greater density than ice, this amount is more than 50 percent water ice by volume.

Signs of surface water

The heavily cratered Martian highlands are locally dissected by integrated networks of tributaries with widths of about 10 km or less and lengths from <5 km to nearly 1000 km. Drainage densities are generally much lower than for terrestrial valley networks, although an interesting set of valleys on Martian volcanoes is more similar to terrestrial valleys in their general morphology and degree of terrain dissection. Among the other important morphological attributes of Martian valley networks are the following: theatre-like valley heads, prominent structural control, low junction angles, quasi-parallel patterns, hanging valleys, irregular widening and narrowing and indistinct terminal areas. In general, this assemblage of features seems best explained by groundwater sapping processes (Gulick 2001), excellent examples of which occur in the sandstone terrains of the Colorado plateau and in parts of the Hawaiian Islands.

Although most of the networks occur in the heavily cratered terrains of Mars, as many as 25–35% of the valleys may be Hesperian or Amazonian in age. Moreover, MOC images show that some valleys, which formed early in Mars' history, were later reactivated by smaller flows—a discovery that had been expected from Viking data. Another observation, the localized development and incomplete dissection of upland areas by valleys, is reinforced by the lack of fine-scale tributaries to valleys observed at MOC resolution. Explanations for such relationships remain contentious, with lack of precipitation as only one of several alternatives, others of which include the probable high infiltration capacities of the Martian surfaces, the burial of small-scale features beneath mantling deposits and the local hydrothermal and/or snowmelt processes (Gulick 2001).

Many valley networks have orientations consistent with the effects of surface deformation by extensive volcanic loading at Tharsis, inferred to have occurred by late Noachian time (Phillips *et al* 2001). These networks occur in heavily cratered terrains that show evidence of extensive erosion during the Noachian. The average rates for this presumed water erosion are estimated at 10^2–10^4 mm per 10^6 yr, which are comparable with the lower range of terrestrial values. However, post-Noachian rates are estimated at only 10^{-1}–10^{-2} mm per 10^6 years (Golombek and Bridges 2000). This discrepancy led some researchers to postulate an epoch of late Noachian climate conducive to water erosion, followed by post-Noachian conditions that precluded such erosion (Carr 1999). However, if post-Noachian aqueous erosion episodes are limited to very short episodes, then the average post-Noachian erosion will yield the same rate as calculated.

The term 'channel' is properly restricted to one class of Martian trough-like landforms that display at least some evidence for large-scale fluid flow on their floors. The principal landform of interest here is the outflow channel, which shows evidence of flows emanating from zones of regional collapse known as 'chaotic terrain'. The

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

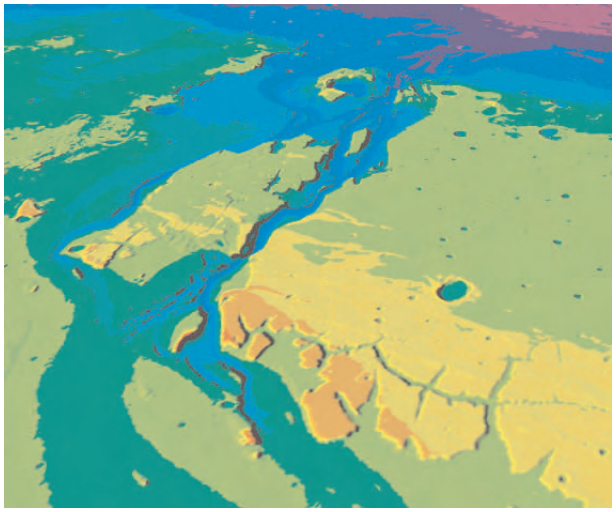


Figure 4. Oblique view of topographic data from the MOLA showing downstream portions of the outflow channel Kasei Vallis.

Martian outflow channels are immense (figure 4), as much as 150 km wide and 2000 km in length. It was recognized shortly after their discovery that they possessed a suite of bedforms and morphological relationships similar to what is exhibited in the 'Channeled Scabland', a terrestrial landscape created by Pleistocene glacially related cataclysmic flooding in the northwestern coterminous United States.

At their largest scale, the outflow channels are broadly anastomosing and split by residual uplands, or 'islands', of pre-flood-modified terrain (figure 2). The channels have low sinuosity and high width–depth ratios. Pronounced flow expansions and constrictions occur, as do prominent divide crossings, hanging valleys and structural control of erosion. At a finer scale, streamlining of the residual uplands is very well developed, as are longitudinal grooves, inner channels, cataract complexes, scabland and bar complexes. Although a variety of other fluid-flow systems have been invoked to explain these features, the whole assemblage of these landforms is best explained by cataclysmic flood processes, with particular analogy to the origin of the Channeled Scabland. Nevertheless, it is also recognized that some important differences derive from the peculiar Martian environment, notably its lower gravitational acceleration, its much lower atmospheric pressure and its prevailing subfreezing temperature in comparison with Earth. These special conditions on Mars would influence such factors as sediment transport mechanics, cavitation and ice formation, debris flow and possible large-scale ice processes in the channels (Lucchitta 2001).

The sizes of Martian outflow channels imply immense discharges of water, exceeding any known flood flows on Earth (figure 5). The calculated peak-flow magnitudes are comparable with those of the high-discharge western boundary currents of Earth's world ocean, such as the Gulf Stream, the Kuroshio and the Agulhas (Baker in press). These currents are integral to Earth's climate system, distributing heat poleward from equatorial latitudes. Major disruption of Earth's circulation pattern provides an endogenetic model for large-scale terrestrial climate change. By analogy, Mars' climate could have been impacted by the megaflood discharges transferring water and heat from the equatorial Tharsis volcanic province to the northern plains of the planet. Outflow activity was probably initiated in the Noachian and outflow events occurred during the early Hesperian, the late Hesperian/early Amazonian (Ivanov and Head 2001) and, surprisingly, during the past 10 million years (Hartmann and Berman 2000). The latter floods occurred in the Marte Vallis–Cerberus Rupes region and the Tharsis area. Thus, principal fluvial events (and possibly associated climate change) occurred right up to most recent Martian history. Alternatively, however, outflow channel events are considered to be random outbursts from a progressively thickening cryo-sphere.

Integral to the controversy over post-Noachian water-related climate change on Mars is the lack of understanding of how such immense discharges can be released from the Martian subsurface. The role of the cryosphere seems important, perhaps for generating pressurized aquifer confinement (Carr 2000). Dike emplacement and penetration of the ice-rich cryosphere (Head and Wilson 2001), pressurized volcanic influences on fracture systems and decompression of gas hydrates deep in the crust (Komatsu *et al* 2000) are other possible factors in the cataclysmic release of outburst floods.

The largest outflow channels delivered their immense discharges to the northern plains of Mars.

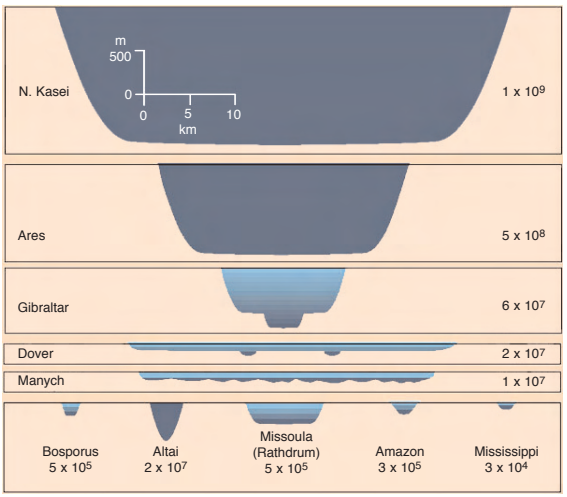


Figure 5. Comparison of simplified channel cross-sections for cataclysmic flood channels on Mars (upper two channels) and Earth (lower channels).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Evidence of temporary inundation of the northern plains was recognized on the basis of Viking data and some aspects of the inundation hypothesis are consistent with new results from the MOLA instrument (Head *et al* 1999). These include the extreme topographic smoothness of the northern plains, the apparent burial of wrinkle ridges, correlation of basin topography with landforms of likely aqueous origin, gradation of outflow-channel floors to ancient inundation levels, consistency of inundation volumes with possible Mars water balances and some correspondence of one of the possible shoreline levels to a deformed ancient equipotential surface (Head *et al* 1999). However, landforms interpreted as diagnostic of shoreline processes are not confirmed by detailed study of MOC imagery. Named 'Oceanus Borealis', the northern-plains inundation remains controversial as to its size, age (or ages), persistence and episodic formation. Other interpretations of the northern-plains inundation include the emplacement of massive debris flows, effects of glaciation and a complex layering of both ice- and water-laid deposits (Clifford and Parker in press).

Evidence for past lakes abounds in numerous impact craters on Mars. Most of the lakes formed in the Upper Hesperian and Lower Amazonian. Lacustrine terraces and Gilbert-type deltas indicate that the lakes must have persisted for 103 to 104 yr (Ori, Marinangeli and Baliva 2000). Standing water in lakes for such timescales requires a major change in climate from present-day conditions, although ice-covered lakes could persist for considerable time at current conditions or for those associated with the range of variation in Mars' orbital parameters. The relationships at the Gale Crater palaeolake are particularly interesting because of extensive work indicating a relatively young age. However, the first interpretation of MOC images claimed that the Gale Crater deposits and all Martian lake deposits were Noachian in age. This ancient age designation was even applied to extensive sedimentary deposits in the Valles Marineris, an interpretation that has been strongly challenged following public release of the images (Weitz 2001). Ironically, the claim of a Noachian age for the deposits requires their extensive exhumation late in Martian history in order to account for the lack of impact craters on the exposed surfaces. Such a presumed erosion event would require even more extreme Martian climate change than would the aqueous emplacement of the deposits that many investigators originally associated with post-Noachian lacustrine activity (Ori, Marinangeli and Baliva 2000). This important scientific controversy continues.

Perhaps the most striking discoveries made from MOC images are the very young, relatively small-scale debris-flow gullies (figure 6), initially explained as the result of groundwater seepage and subsequent surface run-off (Malin and Edgett 2000). The gullies have morphologies essentially identical to those that are common on terrestrial hillslopes in periglacial environments (Costard *et al* 2001), such as coastal Greenland, Iceland, Svalbad and arctic Canada. On Earth, initiation of debris flow occurs when soil water saturation follows surface melting of snow cover or ground ice. Remarkably, the Mars gullies generated debris-flow deposits that are superimposed on aeolian bedforms and on small-scale contraction-crack polygons (Malin and Edgett 2000), all of which are uncratered. Thus, the various age indicators show the gullies and debris flows to be extremely young, probably active within the past several million years of Mars' history.

The profound implications of a recent Earth-like origin for the young gullies led some researchers to seek

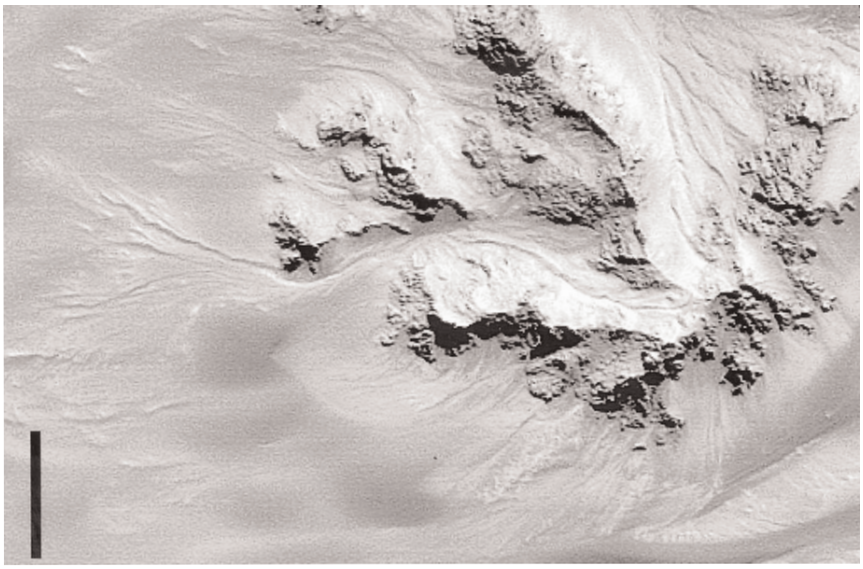


Figure 6. MOC image (M09-04718) of small gullies and other hillslope features in the central peak area of Hale Crater (latitude 36° S, longitude 37° W).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

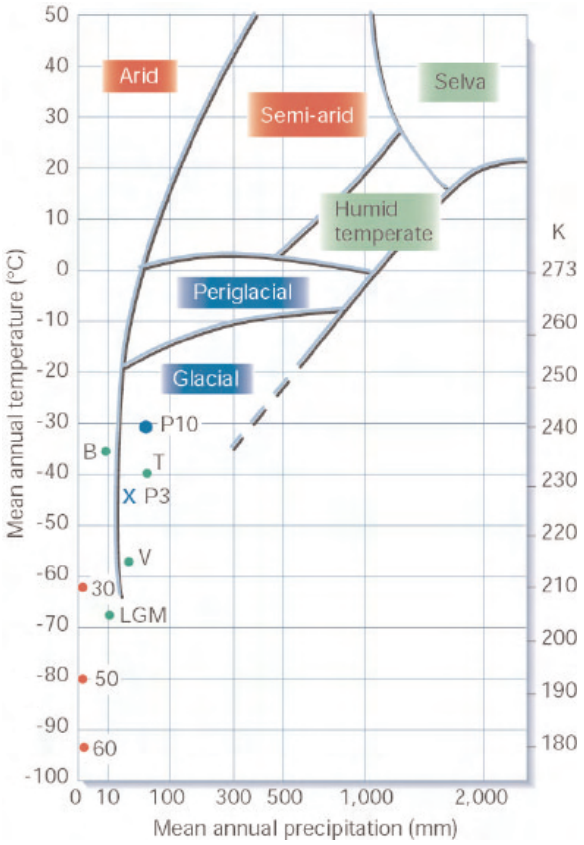


Figure 7. Morphogenetic regions for climate-related landforms on Earth.

alternative mechanisms that might generate debris flows on Mars. Ideally, the modelled processes should be able to occur in the current cold, dry environment. For example, the existence of CO_2-vapour-supported debris flows might, in theory, permit build-up of a liquid-CO_2 aquifer behind a near-surface dry-ice barrier that is subsequently breached by heating. In one extreme scenario, Mars is hypothesized as both currently being and always having been an extreme cryogenic world in which water behaves largely as a solid mineral and nearly all manifestations of fluid-flow phenomena involve CO_2 gas or liquid (Hoffman 2000). This 'white Mars' model postulates a history of mean temperatures continuously lower than those prevailing today.

Signs of glacial ice

Given that periglacial landscapes on Earth involve warmer climatic conditions than do glacial landscapes (figure 7), it is interesting that glacial interpretations of Martian landforms have generally been more controversial than have periglacial interpretations. A strong case was made for the contribution of glacial processes to the origin of landforms associated with certain outflow channels and both glacial and flood outflow processes can operate together or sequentially in large-scale channel morphogenesis (Costard and Baker 2001). Nevertheless, the scale of the presumed climate changes occurring late in Mars' history that would have been necessary to account for glaciation led many researchers to express extreme scepticism with regard to any glacial interpretation.

The most intricate assemblages of glacial landforms are hypothesized to explain landscapes in the mountainous uplands adjacent to the Argyre and Hellas impact basins. Local alpine glacial sculpture in the Charítum Montes, south of Argyre at latitude 57° S, includes cirques, horns, aretes, grooves and V-shaped valley troughs. Many of the cirques and troughs seem to be occupied by extant or relict rock glaciers or debris-covered glaciers (see discussion below). The adjacent floor of Argyre Planitia has landforms that can be interpreted as moraines, drumlins, esker-like ridges, kettles, outwash plains and glaciolacustrine plains.

The esker-like ridges (figure 8), which are 10–200 km long, have proven to be a source of lively controversy at professional meetings. Taken as individual landforms, these sinuous ridges have been variously explained as sand dunes, inverted stream valley fills, igneous or clastic dikes, lake beach ridges, mudflow levees and the wrinkle ridges generated by deep-seated faults. MOC images of the Argyre Planitia sinuous ridges (figure 9) reveal apparent sedimentary strata, boulders and discontinuous ridges, with heights of

Figure 8. Esker-like ridges in southern Argyre Planitia (latitude 56° S, longitude 40° W).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

10–100 m and widths of 200–2000 m. Together with the associated glacial landforms, the esker interpretation seems to provide a consistent hypothesis explaining all relevant observations.

Even more prominent esker-like ridges are associated with the Hesperian-aged Dorsa Argentea Formation at latitudes 75–80° S. Numerous morphological similarities show these ridges to be similar to terrestrial eskers (Head and Hallet 2001). The eskers seem to have been produced by the meltback of ice-rich deposits, perhaps deriving from an extensive south polar ice cap in the middle (Hesperian) portion of Mars' history (Head and Pratt in press). The drainage from this ice cap was carried through prominent valleys to the Argyre impact basin, where it constituted a temporary lake.

MOC images of lineated valley fills in the northern fretted terrain and valleys draining to Hellas, plus images of the lobate debris aprons in uplands near Hellas (figure 1) and Argyre, show features interpreted to represent very recent glacial flow. These include crevasse-like fracture concentrations and medial moraines displaying glacier-like patterns of tributary convergence and down-valley flow. The surfaces of the lineated valley fills and lobate debris aprons are uncratered, indicating likely emplacement within the past several million years. MOLA profiles of lobate debris aprons in Deuteronilus Mensae (latitude 40° N) and Protonilus Mensae (latitude 46° N) reveal shapes best explained by solid-state deformation of ice hundreds of metres thick (Mangold and Allemand 2001). However, at current Mars surface temperatures and very low accumulation rates (<1 cm yr^{-1}), flow rates for these large ice masses would be so slow that they could not be produced on the timescale of 10^6–10^7 yr implied by their uncratered surfaces. Because the hypothesized glaciers are associated with outwash and other water-drainage landforms, higher past temperatures are implied and these could achieve Earth-like strain rates in the deforming ice.

One school of glacial geomorphological thought holds that a continuum exists on Earth from true glaciers to debris-covered glaciers and then to remobilized talus or till. The last phenomena would certainly be classed as rock glaciers, which another geomorphological school holds strictly to be permafrost phenomena. The prevailing hypothesis for the Mars lineated valley fills and lobate debris aprons follows the permafrost theory in its claim that deformation is achieved in ground ice that cements slope-derived debris. In contrast, following the continuum theory, one can hypothesize that recently active, true Martian glaciers carved the V-shaped troughs, cirques, aretes and other landforms of local highland terrains. These ancient glaciers would have had large accumulation areas, requiring atmospheric transport of water to the site and a net surplus of input from snow to offset losses by sublimation and meltwater runoff.

When climate shifted to more adverse conditions, such that input of snow no longer exceeded losses, the glaciers wasted back. However, debris accumulating on the now extensive ablation zone would have greatly reduced sublimation loss. Moreover, if temperatures also fell, meltwater loss would cease. The resulting debris-covered glacier might still have had a very small accumulation zone at its head and this could sustain continued flow of ice at very low rates, as observed in rock glaciers on Earth.

For the extremely cold conditions that currently prevail on Mars, debris-covered ice would have sublimation

Figure 9. Portion of MOC image MOO-01511 showing detail of esker-like ridge in southern Argyre Planitia (figure 8).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Cyclic and episodic change on Mars

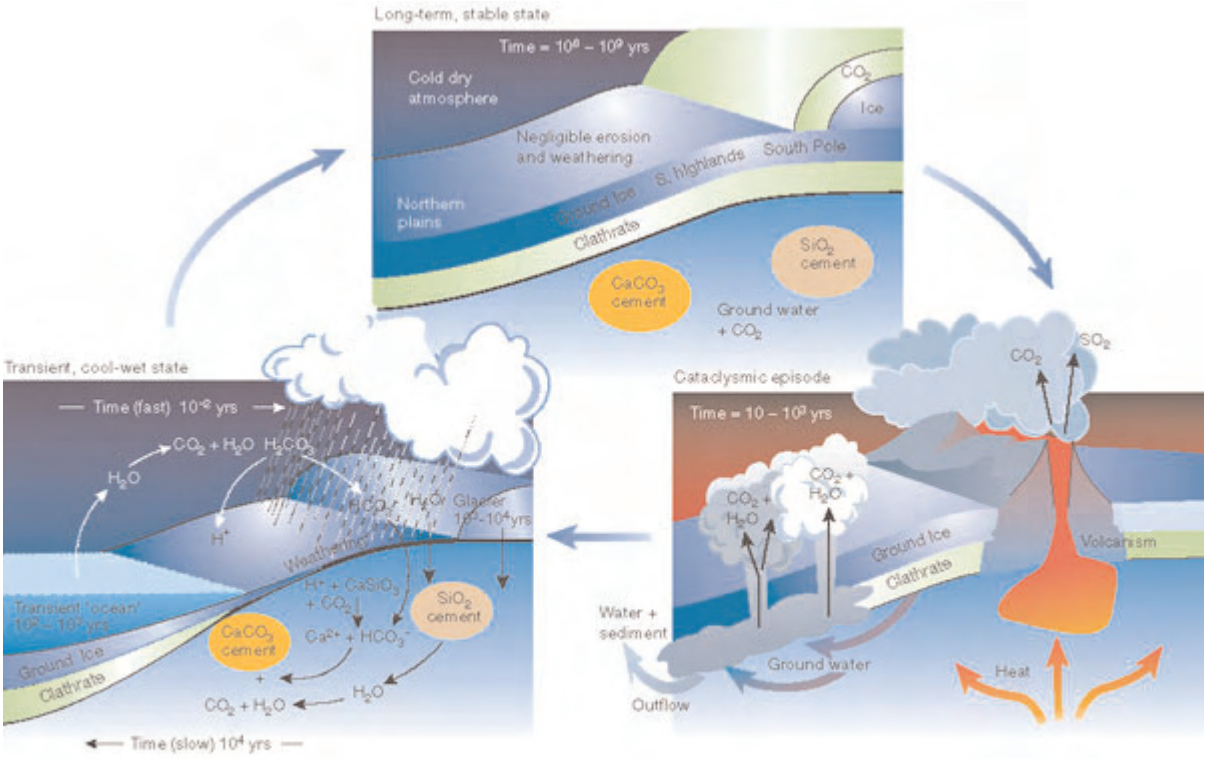


Figure 10. A genetic model, presented in an early form in 1991, predicts both long-term (108–109 yr) and short-term (103–104 yr) aspects of a global Martian hydrological cycle. The episodic inundation of the northern plains of Mars is one component of the short-term portion of the cycle. The hypothesis explains the phenomenally long epochs of post-heavy-bombardment time during which the Mars surface clearly had extremely cold and dry conditions similar to those prevailing today. The alteration of this prevailing condition with quasi-, stable, short-duration warmer (cool) and wetter conditions explains many young landforms that require such conditions, including glaciated terrain, crater lakes, melting of permafrost and other water-related landforms that occur throughout Mars' history. Internal planetary heat could have provided the trigger for the massive outflows that transformed Martian climate during the geologically short epochs of climatic change. Superimposed on the long-term monotonic decline in mantle heat flux for Mars, one can predict short episodes of higher heat flux to the surface. These episodes of higher heat flow seem consistent with the magmatic and tectonic (see TECTONICS) history of Mars (Anderson *et al* 2001). The huge floods would have acted to initiate climate change through release of CO_2, leading to greenhouse warming. During the short-duration thermal episodes of cataclysmic outflow, a temporary cool–wet climate would prevail. Water that evaporated off a transient 'ocean' would be transferred to uplands, including the Tharsis volcanoes and portions of the southern highlands, where precipitation as snow would promote the growth of glaciers. However, such a cool–wet climate is inherently unstable for Mars. Water from the evaporating ocean would be lost to storage in highland glaciers and through infiltration into the porous lithologies of the Martian surface. This would lead to the demise of warm conditions owing to the progressive loss of CO_2 through dissolved gas in infiltrating acidic water and through silicate weathering carrying bicarbonate into the subsurface by infiltration. Subsequent underground carbonate deposition would then release CO_2 to the groundwater, which would become trapped beneath an ice-cemented permafrost zone. The latter develops as the greenhouse effect declines, because of atmospheric loss of water and CO_2 over a timescale of 10^3–10^5 yr. A concurrent decline in planetary heat flow following the triggering peak episode, would lead to a downward-extending permafrost zone that would progressively incorporate the recharging water and groundwater. As the permafrost extended downward into the stability field for CO_2 clathrate, this gas hydrate would accumulate above the gas-charged groundwater. Thus, the underground sequestering of clathrate, gas-charged groundwater and carbonate cements constitute the long-term reservoir for carbon on Mars. Only occasionally, and for relatively short duration, is carbon transferred to the atmosphere, as greenhouse-promoting CO_2, during cataclysmic ocean-forming episodes triggered by pulses of high heat flow. The whole process would be cyclic as illustrated in the diagrams.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

rates as low as 10^{-5} cm yr^{-1} (Carr 1990). Thus, even very small accumulation rates over small areas might sustain a positive net mass balance, although flow rates would be extremely low in a glacier at a mean annual temperature of –75 °C (the current condition on Mars at about 45° latitude). Of course, present water precipitation at this latitude is effectively zero and water condensation occurs today primarily at near-polar latitudes.

The current mean annual temperature at Vostok on the East Antarctic Ice Sheet is –57 °C and the accumulation rate is only 2 cm yr^{-1}. During the Last Glacial Maximum, ice-core data indicate that mean annual temperatures may have been as much as 20 °C colder and accumulation rates may have been less than 1 cm yr^{-1} (Mulvaney *et al* 2000). These conditions are very close to those on Mars today (figure 7), although Earth atmospheric pressures are, of course, much higher. In Beacon Valley, part of the Dry Valley region of Antarctica, precipitation is currently less than 1 cm yr^{-1} water equivalent and the mean annual temperature is -35 °C. The area is icefree, but relict glacial ice occurs beneath a 0.5 m debris cover that is dated at 8.1 million years (Myr) old. The indicated sublimation rate is 10^{-5} cm yr^{-1}, the same rate noted above for debris-covered ice on Mars. With present-day conditions on Mars so close to those in parts of Antarctica (figure 7), it is reasonable to invoke a climate-change scenario to explain the observed glacial phenomena.

Temperatures achieved during the Last Glacial Maximum are representative only of Earth's most recent glacial epoch, the Late Cenozoic. Earlier periods of prolonged, extensive glaciations (megaglaciations) of Earth occurred in the Late Palaeozoic (260–340 Myr ago), Late Devonian/Early Carboniferous (350–360 Myr ago), Ordovician/Silurian (430–450 Myr ago), Late Proterozoic (500–900 Myr ago), Early Proterozoic (about 2200–2400 Myr ago) and Archaean (2910–2990 Myr ago). The Late Proterozoic glaciation is particularly enigmatic, as there is evidence that Earth may have switched into Mars-like ice-house conditions by freezing the surface of the global ocean. The extreme glaciation is explained alternatively as a result of a huge change in planetary obliquity or as a runaway cooling that terminates cataclysmically with a super-greenhouse build-up caused by volcanism (Hoffman and Schrag in press). The rapid melting of ice and weathering of the land surface lead to carbonate precipitation in the ocean, thereby terminating the super-greenhouse.

The glacial epochs of Earth are explained by two classes of theoretical models: external and internal. External causes include the orbital parameters, axial tilt (obliquity), precession and eccentricity, all of which acted as the 'pacemaker' of Pleistocene glaciations. Internal causes include the role of conveyor-belt oceanic circulation and the arrangement of continental relief on the planet. For Mars, it is possible to imagine an internal scenario capable of explaining the diverse observations summarized in this paper (see figure 10), although there have been many objections raised to an earlier version of this model. Alternatively, the complex variability of Mars' obliquity over long timescales may be able to mobilize water in ways that will explain relatively recent manifestations of glacial action, lakes and fluvial gullies and the melting of near-surface ground ice.

Is it possible, as a theoretical matter of pure physics, that the numerous Martian landscapes, temporal associations of landforms and proximal-to-distal relationships—all analogous to water-generated process–form relationships on Earth—could have been produced by a fluid that mimics the behaviour of water? For appropriate relationships with temperature and pressure, CO_2 is a possible candidate for achieving the necessary physical behaviour (Hoffman, 2000). Moreover, the CO_2 hypothesis is parsimonious in the sense of being capable of generating flow phenomena for the current near-surface environment of Mars. If Mars, unlike Earth, were not perturbed over geological time by major epochs of climate change, then a perpetual 'white Mars' might well occur. Nevertheless, it would be remarkable that this white Mars landscape evolved locally to generate surprisingly detailed copies of what on Earth is readily produced by the long-term action of water and ice. It is also a measure of simplicity in a hypothesis that the most natural explanation, in this case an aqueous origin, tends to accord with greater consistency with the observed phenomena.

In choosing to explore the consequences of a water-based explanation for Mars' geomorphology and the associated implications for both ancient and very recent climate change, one is not assured of easily testable hypotheses. The extensive hydrosphere implied by past aqueous activity on Mars may only be extant as ground ice in the thick permafrost zone and as underlying groundwater. Yet, this is the type of environment in which the extremophile progenitors of Earth's biosphere probably evolved. Indeed, early Mars provided an arguably better habitat for the inception and incubation of early life than did early Earth. Episodic, brief episodes of aqueous activity on the Martian surface may have exposed this biosphere to produce possible fossil indicators of its existence. More speculatively, a deep subsurface biosphere containing methanogenic archaea could have produced methane that accumulated beneath a growing Martian cryosphere. The result could destabilize the Martian cryosphere (see figure 10) and perhaps change the climate on short timescales.

Ultimately in geomorphological investigation, one chooses the working hypothesis that is most fruitful with regard to connection to other lines of scientific inquiry, while still providing consistency and coherence in

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

explaining the whole complex of available observations. The water-generated landforms and landscapes of Mars have proven especially difficult to unite under a single working hypothesis. Nevertheless, whatever explanation one chooses as a working hypothesis, its roots in the complex detail of Mars' landscape features ensure that it will provide a highly probable basis for further productive inquiry.

Acknowledgements
NASA provided partial support for the research.

This article is a modified version of a previously published article: *Nature* **412** 228–36 (2001). Reproduced by courtesy of *Nature*.
[http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v412/n6843/abs/412237a0_fs.html</]

Bibliography

Anderson R C *et al* 2001 Primary centers and secondary concentrations of tectonic activity through time in the western hemisphere of Mars *J. Geophys. Res.* **106** 20 563–20 586

Baker V R 2002 *High-energy megafloods: planetary settings and sedimentary dynamics in Flood and Megaflood Deposits: Recent and Ancient Examples* ed I P Martini, V R Baker and M Garzon (International Association for Sedimentol. Spec. Publ. Spec. Publ. No.32 3–15 (Int. Assoc. Sedimentology)

Carr M H 1990 D/H on Mars: effects of floods volcanism impacts and polar processes *Icarus* **87** 21027

Carr M H 1999 Retention of an atmosphere on early Mars *J. Geophys. Res.* **104** 21 897–21 909

Carr M H 2000 Martian oceans valleys and climate *Astron. Geophys.* **41** 3.20–3.26

Clifford S M and Parker T J in press The evolution of the Martian hydrosphere: implications for the fate of a primordial ocean and the current state of the northern plains *Icarus*

Costard F *et al* 2001 Debris flows on Mars: Analogy with terrestrial periglacial environment and climatic implications *Lunar Planet. Sci.* Conf. XXXII Abstr. 1534 http://www.lpi.usra.edu/meetings/lpsc2001

Costard F and Baker V R 2001 Thermokarst landforms and processes in Ares Vallis Mars *Geomorphology* **37** 287–301

Golombek M P and Bridges N T 2000 Erosion rates on Mars and implications for climate change: constraints from the Pathfinder landing site *J. Geophys. Res.* **105** 1841–53

Gulick V C 2001 Origin of the valley networks on Mars: a hydrological perspective *Geomorphology* **37** 241–68

Hartmann W K and Berman D C 2000 Elysium Planitia lava flows: crater count chronology and geological implications *J. Geophys. Res.* **105** 15011–25

Head J W III *et al* 1999 Possible ancient oceans on Mars: evidence from Mars Orbiter Laser Altimeter Data *Science* **286** 2134–37

Head J W III and Wilson L 2001 Mars: geologic setting of magma/H_2O interactions Lunar Planet *Sci. Conf. XXXII* Abstr. 1215 http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1215.pdf

Head J W III and Hallet B 2001 Origin of sinuous ridges in the Dorsa Argentea Formation: new observations and tests of the esker hypothesis *Lunar Planet. Sci.* **32** Abstr. No. 1373

Head J W III and Pratt S in press Extensive Hesperian-aged south polar ice sheet on Mars: evidence for massive melting and retreat and lateral flow and ponding of meltwater *J. Geophys. Res.*

Hoffman N 2000 White Mars: a new model for Mars' surface and atmosphere based on CO_2 *Icarus* **146** 326–42

Hoffman P F and Schrag D P in press The Snowball Earth hypothesis: theory observations and tests *Terra Nova*

Ivanov M A and Head J W 2001 Chryse Planitia Mars: topographic configuration outflow channel continuity and sequence and tests for hypothesized ancient bodies of water using Mars Orbiter Laser Altimeter (MOLA) data *J. Geophys. Res.* **106** 3275–95

Jakosky B M and Haberle R M 1992 *Mars* ed H H Kieffer, B M Jakosky, C W Snyder and M S Matthews 969–1016 (Tucson, AZ: University of Arizona Press)

Komatsu G. *et al* 2000 A chaotic terrain formation hypothesis: explosive outgas and outflow by dissociation of clathrate on Mars Lunar Planet *Sci. Conf. XXXI* Abstr. 1434 http://www.lpi.usra.edu/meet

Lane M D and Christensen P R 2000 Convection in a catastrophic flood deposit as the mechanism for the giant polygons on Mars *J. Geophys. Res.* **105** 17617–27

Lucchitta B K 2001 Antarctic ice streams and outflow channels on Mars *Geophys. Res. Lett.* **28** 403–06

Malin M C and Edgett K S Evidence for recent groundwater seepage and surface runoff on Mars *Science* **288** 2330–35

Mangold N and Allemand P 2001 Topographic analysis of features related to ice on Mars *Geophys. Res. Lett.* **28** 407–10

Masursky H 1973 An overview of geologic results from Mariner 9 *J. Geophys. Res.* **78** 4037–47

Mulvaney R R *et al* 2000 The transition from the last glacial period in inland and near-coastal Antarctica *Geophys. Res. Lett.* **27** 2673–76

Ori G G, Marinangeli L and Baliva A 2000 Terraces and Gilbert-type deltas in crater lakes in Ismenius Lacus and Memnonia (Mars) *J. Geophys. Res.* **105** 17629–41

Phillips R J *et al* 2001 Ancient geodynamics and global scale hydrology on Mars *Science* **291** 2587–91

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Seibert N M and Kargel J S 2001 Small-scale Martian polygons: liquid surface water *Geophys. Res. Lett.* **28** 899–902

Smith D E *et al* 1999 The global topography of Mars and implications for surface evolution *Science* **284** 1495–1502

Weitz C M *et al* 2001 The interior layered deposits of Valles Marineris: layering erosional processes and age relationships *Lunar Planet. Sci.* **32** Abstr. No. 1629

Victor R Baker

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Marshall Space Flight Center

NASA's George C Marshall Space Flight Center is located in Huntsville, Alabama. It is the agency's leading center for space transportation and propulsion development. The Saturn launch vehicles used in the Apollo Moon program were designed and developed here. Today, it provides the solid rocket boosters, main engines and external tank for the Space Shuttle. Apart from upgrades to these systems, Marshall engineers are designing future concepts for reusable, low-cost space transportation.

The Marshall Space Flight Center is also an important center for research into microgravity, space physics and astrophysics. One of its most important recent programs is the Chandra X-ray Observatory (formerly known as AXAF).

For further information see http://www.msfc.nasa.gov/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maskelyne, Nevil (1732–1811)

Astronomer Royal, born in London, England, his interest in astronomy had begun after seeing the eclipse of 1748. After a mathematics education at Cambridge, he was sent by the Royal Society in 1761 to St Helena to observe the transit of Venus. In 1764 he traveled to Barbados, testing JOHN HARRISON's chronometer, and was appointed Astronomer Royal on his return. He initiated publication of the *Nautical Almanac* in 1766. Following BOUGUER and LA CONDAMINE, he proposed an experiment for determining the Earth's density with the use of a plumb line, whose plumb-bob was attracted sideways by the gravitational pull of a mountain. He carried out the experiment in 1774 on Schiehallion, a mountain in Perthshire, Scotland, chosen because it was a regular conical shape and its volume could be calculated accurately. He determined that the Earth's density is approximately 4.5 times that of water.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mason, Charles (1730–87)

English astronomer, an assistant at the Royal Observatory, Greenwich, he went with JEREMIAH DIXON to the Cape of Good Hope to observe the transit of Venus in 1761. Dixon and Mason were engaged to survey the boundary between Maryland and Pennsylvania, and reached a point 224 miles west of the Delaware River, but stopped work due to hostile Indians. The boundary was given the name *Mason–Dixon line*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mass

A measure of the amount of matter in a body or of its inertia (the extent to which it resists acceleration when a force is applied to it).

The mass of a body, when regarded as a measure of the amount of matter contained within it, is one of the factors that determines the strength of its gravitational field. In Newtonian gravitation, the magnitude of the mutual gravitational attraction (F_g) between two bodies, of masses m and M respectively, separated by distance r, is given by $F_g = GmM/r^2$, where G is the gravitational constant. Mass, in this context, is called 'gravitational mass'.

Mass also appears as a constant of proportionality in Newton's third law of motion (the law of inertia), which describes the acceleration (a) experienced by a body when acted on by a force (F). This law is commonly written as $F = ma$, or $a = F/m$, where m, in this context, is called 'inertial mass' (the acceleration produced by a given force is inversely proportional to the mass of the body to which that force is applied). Inertial and gravitational mass are assumed to be identical. Although this assumption has been confirmed experimentally to a very high degree of precision, it remains a theoretical assumption.

The mass of a body when measured in a frame of reference in which that body is stationary is its 'rest-mass'. According to the special theory of relativity, the mass of a moving object is greater than its rest-mass and tends towards an infinite value as the velocity approaches very close to the speed of light. This theory also asserts that mass and energy are equivalent and interchangeable, the relationship between mass (m) and energy (E) being $E = mc^2$, where c denotes the speed of light. The total mass-energy of an isolated system (i.e. the sum of the total mass and total energy in that system) is a constant.

The unit of mass in the SI (International) system is the kilogram (kg). Particle physicists tend to use the term 'mass' to describe the quantity of energy (E) that is equivalent to a subatomic particle of mass m ($E = mc^2$), and quote 'masses' in terms of energy units (electronvolts). Strictly speaking, however, the mass of an elementary particle is its equivalent energy divided by the square of the speed of light (E/c^2). For example, the 'mass' of a proton is often quoted as 938.3 MeV. In conventional units, the mass of a proton (=938.3 MeV/c^2) is 1.6725×10^{-27} kg.

See also: electronvolt, energy, gravitation, inertia, matter, proton, special relativity, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Massey, Harrie Stewart Wilson (1908–83)

Australian-born British space physicist, became professor at University College London.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Matter

Physical substance that occupies finite volumes of space and which has mass. Bulk quantities of matter can exist in any of the following states: solid, liquid, gas or plasma. On the microscopic scale, matter exists as molecules, atoms, ions, subatomic particles or elementary particles. Because of the equivalence of mass and energy, which is a consequence of the special theory of relativity, matter may be considered as a particular form of energy.

All but a tiny fraction of each atom's mass is contained within its central nucleus, which is composed of protons and neutrons (the nucleus is surrounded by a cloud of much lighter electrons). Because protons and neutrons belong to a class of particles called baryons, normal matter is often referred to as 'baryonic matter'. In addition to luminous matter (the stuff of which stars are composed), which is detectable by the radiation that it emits, the universe appears to contain substantial quantities of dark matter that emits no discernible radiation and which is detectable only through its gravitational influence. There are theoretical grounds to suggest that part of the dark matter may be 'non-baryonic' (composed of hypothesized particles that are not baryons).

See also: atom, baryons, elementary particles, hadrons, ion, leptons, mass, molecule, neutron, plasma, proton, special relativity, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maunder, E W (1851–1928) and Maunder, Mrs A S D

Solar astronomers. Maunder became assistant for spectroscopic and solar observations at the Royal Observatory, Greenwich under GEORGE AIRY, aided by his wife. In 1890, while studying the numbers of sunspots over a 300 year time-span he noticed the scarcity of spots in the period 1645–1715. This so-called *Maunder minimum* was confirmed by Jack Eddy (1976) to be a real effect rather than simply a scarcity of observations, by a reduction of carbon-14 in old tree rings at the same time (carbon-14 production in the Earth's atmosphere is modulated by solar activity). In 1904 invented the 'butterfly diagram' of sunspots, showing the solar latitude at which sunspots appear as a function of time. The spots show a progressive drift with the phase of the solar cycle (Spörer law). At the beginning of the solar cycle, spots tend to appear at latitudes 35 degrees N and S. Their number increases and their latitude tends to decrease as the cycle moves on, so that at solar maximum the many spots appear at intermediate latitudes (15 degrees north and south). At the next solar minimum, spots of the old cycle are equatorial, while spots of the new cycle are starting to appear at latitudes 35 degrees again. The plot of this phenomenon has the shape of a butterfly's wings. Maunder also researched into the history of the constellations, for example he calculated the date of origin of the classical constellations at 2700 BC by considerations of the effects of precession on the constellation figures. As an evaluation of the reality of the 'canals of Mars', he carried out experiments on how random unresolved spots tended to be subjectively viewed as linear features.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maupertuis, Pierre Louis Moreau de (1698–1759)

Mathematician, born in Saint Malo, France, published on mathematics, geography, and astronomy, introduced NEWTON's theory of gravitation in France. He developed his theory of dynamics with the principle of least action, a way to determine the orbit of a body by a minimization technique. As a unifying principle, he though that it could be a proof of the existence of God. He went to Lapland in 1736 to determine the flattening of the Earth at the poles into an oblate spheroid, as predicted by Newton.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maurolico [Marul, Marol], Francesco (1494–1575)

Born in Messina, Italy, became a Benedictine and acted in several civil posts in Sicily. He observed the nova of 1572, now known as 'TYCHO's supernova', and wrote new editions of classical Greek mathematics, most published after his death.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Max-Planck-Institut für Astronomie

Founded by the Senate of the Max Planck Society in 1967, it is one of 68 Max-Planck-Institutes in Germany. The Max-Planck-Institut für Astronomie has its headquarters in Heidelberg, where new buildings were erected between 1971 and 1975 on a 5.2 hectare site alongside the State Observatory.

The staff of approximately 150 includes engineers and other technicians as well as about 30 astronomers. They cover a broad spectrum of expertise including laboratory physics, informatics, electronics, instrument-making and photography. It is also an institute of higher education for students of astronomy and physics, and for apprentices in instrument-making and information electronics.

The Institute also operates the CALAR ALTO OBSERVATORY in southern Spain as well as conducting research in astronomy and astrophysics. Studies include star formation, extragalactic objects, astrophysical theory and instrumentation development.

A small station with a 50 cm telescope is also operated on the 2350 m high Gamsberg, a table mountain on the western edge of the Kalahari desert in Namibia. Thousands of photometric measurements on stars of the southern Milky Way have been carried out from this site.

Since 1984, the Institute has also loaned a 2.2 m telescope to the EUROPEAN SOUTHERN OBSERVATORY at La Silla. Under the 25 year agreement, astronomers of the Max Planck Society are permitted sole use of the telescope for 25 per cent of the observing time.

For further information see http://www.mpia-hd.mpg.de/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Max-Planck-Institut für Astrophysik

The Max-Planck-Institut für Astrophysik, now located in the town of Garching north of Munich in Germany, is one of the more than 70 autonomous research institutes of the Max-Planck-Gesellschaft. It was founded in 1958 under the direction of Ludwig Biermann as part of the Max-Planck-Institut für Physik und Astrophysik, directed at that time by Werner Heisenberg. In 1979, when the headquarters of the EUROPEAN SOUTHERN OBSERVATORY (ESO) came to Garching, the Institute also moved to its present location. It is now led by a board of three directors.

Research at the Institute is devoted to a broad range of topics in theoretical astrophysics. Major concentrations of interest lie in the areas of stellar evolution, stellar atmospheres, supernova physics, astrophysical fluid dynamics, high-energy astrophysics, galaxy structure and evolution, the large-scale structure of the universe and cosmology. While most of the Institute's research addresses theoretical issues, the neighboring institutes on the campus of Garching (ESO and the MAX-PLANCK-INSTITUT FÜR EXTRATERRESTRISCHE PHYSIK) provide complementary expertise on observationally oriented astrophysics, and there are many collaborative projects with them.

At any given time the Institute has about 30 scientists working on long-term positions, up to 10 foreign visitors brought in by various programs, and about 20 graduate students, mostly enrolled for degrees in one of the two large universities in Munich. Several of the senior staff members have affiliations with these universities.

For further information see
http://www.mpa-garching.mpa.de.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Max-Planck-Institut für Extraterrestrische Physik

The Max-Planck-Institut für Extraterrestrische Physik (MPE) is one of the world's leading institutes dealing with space physics and astrophysics. The main activities are in infrared, x-ray and γ-ray astronomy and space-plasma physics, and extend to colloidal plasma physics. The institute is primarily involved in developing novel instruments and using them for performing observations and measurements. This is supplemented by theoretical activities. Major projects are Helios, AMPTE, Cluster, ISO, Sofia, First, Rosat, Chandra, XMM, COS-B, CGRO and Integral.

The institute, which is located in Garching near Munich in Germany, was founded in 1963. Today it comprises about 150 scientists and about 250 other staff members. Apart from physical laboratories, the institute has its own workshop, an integration and test facility for space experiments, a test facility (Panter) for calibration of x-ray telescopes and it runs a semiconductor laboratory as a collaborative enterprise with the Max-Planck-Institut für Physik in Munich. The MPE receives institutional support from the Max-Planck Society and project funds from government institutions (BMBF, DLR and DFG). The research activities are often carried out as cooperative projects together with research institutes and universities within and outside Germany, in particular with ESA and NASA, but also with national agencies.

For further information see
http://www.mpe-garching.mpg.de/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Max-Planck-Institut für Radioastronomie

Founded in 1966, the Max-Planck-Institut für Radioastronomie in Bonn is one of the centers for radio astronomical research in Germany. Main fields of research include the interstellar medium of our own and external galaxies, cosmic magnetic fields, pulsars, star formation, late stages of stellar evolution, active galactic nuclei, theoretical high-energy astrophysics, and observational cosmology. Instrumentation for centimeter and (sub)millimeter wavelength observations and for optical and infrared interferometry is developed in-house. The institute operates the EFFELSBERG 100 m RADIO TELESCOPE, which frequently takes part in Very Long Baseline Interferometry (VLBI) observations. The 10 m Heinrich-Hertz Telescope on Mount Graham, Arizona, a facility for work at submillimeter wavelengths, is operated jointly with the University of Arizona's STEWARD OBSERVATORY.

For further information see http://www.mpifr-bonn.mpg.de/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maxwell Montes

The highest mountain range on Venus, situated in the upland region Ishtar Terra, and centered at 65.2 °N, 3.3 °E. It extends for 797 km, and contains the highest point on the planet, nearly 12 km above Venus's average surface level (comparable in this respect to Earth's Mount Everest). It is the only Venusian feature to bear a male name, honoring James Clerk Maxwell. The range showed up prominently in early radar surveys as a 'radar-bright' region. It was given its name in 1970, before the convention of assigning female names was adopted; because the name had become well established it was allowed to survive.

The western slopes are very steep (35°), while to the east the descent is more gentle, toward the region known as Fortuna Tessera; the range itself is a complex of compressive faulting and folding. The steepness and height could point to a relatively recent formation, or it could mean that the underlying crust is thicker or stronger, or even that the rocks in this region are of lower than average density. The radar brightness of Maxwell Montes is thought to indicate that its peaks are covered with a thin deposit of a radar-reflective material similar to the mineral pyrite, which on Venus can exist in solid form only at high altitudes, where the pressure and temperature are low enough. To the east of the range is the 105 km diameter double-ringed impact crater Cleopatra.

See also: Venus: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maxwell, James Clerk (1831–79)

Born in Edinburgh, Scotland, worked at Cambridge on electric and magnetic fields and their interrelation. Entered a prize-winning essay for a competition on the subject 'The Motion of Saturn's Rings' and showed that the rings could be stable only if they consisted of numerous small solid particles in Keplerian motion with the inner edge moving faster than the outer edge. GEORGE AIRY, a competition judge, wrote that it was one of the most remarkable applications of mathematics to physics that he had ever seen. Maxwell worked on the perception of color, inventing the three-color process used, for example by DAVID MALIN in astrophotography. At King's College, London, Maxwell proposed that light is an electromagnetic phenomenon and worked on the kinetic theory of gases, showing that temperatures and heat involved the motion of molecules and that the velocity distribution of molecules was '*Maxwellian*'. In 1871 he became the first Cavendish Professor of Physics, and wrote a *Treatise on Electricity and Magnetism* which unified what appeared to be two separate forces into one set of four partial differential equations, now known as *Maxwell's equations*. The *James Clerk Maxwell Telescope* in Hawaii is named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mayan Astronomy

The Mayas of Mexico, Belize, Guatemala and Honduras developed an urban civilization in the tropical rain forests. In their cities are many monuments, mainly stone, which contain accounts of their rulers' families, conquests, sacrifices, rituals and building programs. These are integrated into an elaborate calendrical system, which gives dates to the day. We also have pottery, painted with elaborate scenes, often with dates and depictions of the cosmic interaction of deities, and four books have survived. It was thought that each day was ruled by many separate deities, gods of numbers, gods of day names, gods of months and years, and others. The fate of each day was determined by the interaction of these gods in numerological and astronomically determined patterns.

The Mayas' knowledge of astronomy allowed them to calculate eclipses and planetary movements over thousands of years. Some inscriptions include calculations over many millions of years. These seem to be combinations of numerology and astronomy. Whereas we would say that the tropical year during the time that the Mayas flourished was 365.242 58 days, the Mayas gave dates indicating that 29 periods of 18 980 days equaled 1508 Maya years of 365 days and that this corresponded to 1507 tropical years. This calculation would require about 70 000 yr to accumulate a full day of error. Much longer calculations imply much greater precision although not necessarily greater accuracy. The greater the length of time specified, the higher the precision which is being claimed. Some such calculations involved periods substantially greater than the present geologically estimated age of the Earth.

The Maya calender

There were three major components of the CALENDAR: a tzolkin (a period of 260 days—20 repetitions of the numbers from 1 to 13 and 13 repetitions of a series of 20 named days); a 'year' of 18 named months of 20 days each plus a set of 5 days referred to by various euphemisms; an account of the number of days from a specified date in the past which served as an era base. The combination of the 365 day year with the 260 day sequence resulted in a repetition of the same combination after 18 980 days ($73 \times 260 = 52 \times 365$). This combination is called a Calendar Round and was found throughout Mesoamerica (the cultural area which shows a large grouping of similar traits—most of the non-Mayan peoples of Mesoamerica were in modern Mexico).

Astronomical observations

We have almost no direct evidence of observational techniques. There are indications that observers posted at fixed points at fixed times of the year, such as EQUINOXES or SOLSTICES, checked solar and planetary observations, often using a fixed alignment defined by two buildings or a building and a mountain peak or low point on a distant horizon. The Mayas had concave mirrors of obsidian (a naturally occurring volcanic glass, often black or green). Such a mirror could have been used directly as a reflecting telescope. This not only would have made fainter objects visible but would have permitted observations substantially closer to the Sun at sunrise and sunset. At Teotihuacan in highland Mexico, light from the Pleiades was projected onto a floor, apparently through a long tubular opening, causing a substantial enlargement (this technique is incorporated into any camera obscura in our own culture). Each dot of light from the 11 brightest stars of the Pleiades was marked on the floor by identical images of a jaguar, drawn by connecting 11 dots forming tiny maps of the Pleiades. These 'maps' or star diagrams are essentially elaborations of the so-called 'ball and link' designs (attested later among the Aztecs) in which solid line 'links' connect circular star markers ('balls'). Although no similar observational technique is attested for the Mayas, the amount of shared astronomy, calendrics and cosmology is such that it seems very probable that they shared the technique, particularly since there is a 'Maya quarter' at Teotihuacan.

The Mayas deliberately built some structures in such a way that the interaction of light and shadow created striking visual effects at specified times of the tropical year (i.e. the year measured from the spring equinox in 0° of the sign of Aries). At Chichen Itza, in Yucatan, there is a four-sided 'pyramid', with 91 steps on each side, which is associated with Kukulcan, the Feathered Snake. The balustrades are carved as gigantic feathered rattle-snakes, head down. At the equinoxes, shadows cast on the balustrades by different levels or stages of the pyramid create the illusion that the giant snakes are descending the stairs. Feathered snakes were primarily identified with Venus at the time when this monument was erected, although I have argued for an earlier identification with Mercury. This may suggest a particular Venus phenomenon which occurred at a particular spring equinox. Despite this equinoctial connnection, the building itself is aligned to the rising point of the Sun at the summer solstice and to its setting point at the winter solstice.

At Palenque, in Chiapas, an image of an old black god with an owl in his headdress is the last monument to be lit by the descending Sun at the winter solstice. I have suggested that he is a personification of Saturn, and that he rules the night, so that we may suppose that this was intended to mark a particular date when Saturn was visible in the east at sunset on a winter solstice. I have postulated that the calendar names of Mesoamerican deities, like those of human beings, were marked from the day of their 'birth', apparently usually identified as conjunction with the Sun for planets and as equinoxes or solstices for the Sun. In the case of the Moon, identification with particular eclipse intervals and series seems to have been usual. I have further argued that all calendar names were calculated from a single calendar round date, the date 12 Lanat on new year's day, the first of Pop. This date was about 3500 yr in the past

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

when the Maya first began erecting monuments, as will be explained. My interpretation of the calendar names of the gods has not been widely accepted, but neither has it been formally criticized and rejected, nor has anyone proposed a coherent alternative. To the extent that it is accepted, the Maya seem to have been interested in all the visible planets both in their movements relative to the sun and relative to the asterisms, or star groups which the Maya recognized. They were deeply interested in ECLIPSES of the Sun and of the Moon, and particularly interested in associating planetary movements with eclipses.

Deities might appear in human form or as animals, from the lowly centipede representing the high god Itzamna, to the deer representing the Sun and the rabbit representing the Moon. An example which clearly integrates these varied concepts is a representation of a heron with the head of a black god incorporated into the body and the glyph for 'star' attached. I think that it may be a representation of Jupiter.

Dates on inscriptions

Returning to the inscriptions, dates are normally given by five numbers counted from 4 Ahao 8 Cumku, previously mentioned, and followed by the resultant calendar round date. Thus the date 9.16.4.10.8 12 Lamat 1 Muan means that 9 baktuns (9×400 tuns of 365 days each, that is, $9 \times 144\,000 = 129\,6000$ days), 16 katuns (16×20 tuns; $16 \times 7200 = 115\,200$), 4 tuns ($4 \times 360 = 1440$), 10 uinals of 20 days each—$10 \times 20 = 200$ and 8 days passed (from the era base 4 Ahau 8 Cumku to 12 Lamat 1 Muan).

In an inscription of Palenque, a god is said to have been born on the date 1.18.5.4.0 1 Ahau 13 Mac. This is probably an implicit reference to a date 9.11.3.2.0 1 Ahau 13 Mac which fell during the lifetime of those who put up the inscription. The interval between 1.18.5.4.0 and 9.11.3.2.0 is 58 calendar rounds. The period of 29 calendar rounds and its multiples ($2 \times 29 = 58$) appears in a number of inscriptions. The astronomical characteristics of this period are very interesting. It consists of 1508 years of 365 days which is equivalent to 1507 tropical years with a minimal error which would amount to a day in about 70 000 yr. The double interval used here is 3016 Maya years, equivalent to 3014 tropical years or 1 100 840 days. We do not know the precise parameters used by the Mayas in any of their calculations except as these may be derived from the calculations themselves. The Maya used a minimal period of 2920 days to relate Venus and the year since $8 \times 365 = 2920$ and 5 approximate Venus periods of 584 days each also equal 2920 days. The Maya day name of Saturn is 13 Cimi which implies 378 days for the synodic period of Saturn. These factors suggest connections between Venus, Saturn and the tropical year, since 2920×377 is 1 100 840 days, and that the day 1.18.5.3.6 13 Cimi 19 Keh, 14 days earlier than 1.18.5.4.0 1 Ahau 13 Mac, is also given here as the birth of a god, the so-called 'Jaguar-Baby', one of twin heroes. Comparative mythology suggests that 'Jaguar-Baby' was the half- brother of a deity with water associations and that he ultimately replaced his Jaguar father as the opponent of youthful heroes. I assume that 'Jaguar-Baby' is Saturn.

One of the most apparent relationships between history and astronomy in the Mayan inscriptions is the not uncommon use of a star glyph with a place-name glyph to indicate that the place has been conquered. It has been suggested that these conquests occurred when Venus was at greatest elongation, but the statistical evidence for the view that Venus was at any specified or specifiable position at all the dates concerned is not strong and other planets may also be involved.

Astronomy in the codices

This brings us to the astronomical evidence found in the Maya books. These are called the *Dresden*, *Madrid*, *Paris* and *Grolier* codices. It has recently been determined that the *Madrid* codex was written after the Spaniards had arrived in the New World, but there is no trace of Spanish culture in the manuscript. The city names of the first three codices have to do with their present locations and nothing to do with their content. Many books were destroyed when the Spaniards conquered the Maya area, but most scholars at the present time recognize four surviving Maya documents which are pre-Spanish in content. The pre-Columbian books and maps of Mesoamerica are usually called codices, although Latin codices were bound books and these are not. The most information comes from the astronomical tables of the *Dresden* codex. The Venus table and the table of lunations and their relationships to eclipses are accepted by virtually all Mayanists. There is also a table of Mars accepted by a substantial minority of scholars and structurally similar tables which I think refer to Mercury, Jupiter and Saturn. In all cases, there is a date in the distant past (somewhat before the Maya era base), then an interval and a date during the Maya Classical Period or somewhat later. This is followed by a multiplication table of factors relevant to astronomy and divination for the particular planet, including apparent correction factors. We have a table of mean motion for the particular planet, sometimes divided in ways which seem unusual to modern Western scholars. Before discussing individual tables, it should be pointed out that the Maya era base is occasionally referred to as 13.0.0.0.0 4 Ahau 8 Cumku counting 13 baktuns or 5200 tuns from a previous era base. This is comparable with our saying that the Christian era base of AD 1 is equivalent to the year 754 of the era base counted from the foundation of Rome. Many early scholars thought that this meant that a new era base would begin every time the count reached 13 baktuns. We know that the Maya used 20 baktuns to make one of the next higher unit (called by Mayanists a pictun) and we know that they also used completely different era bases such as 9 Kan 12 Kayab, which was about 30 000 yr before 13.0.0.0.0 4 Ahau 8 Cumku according to the most likely interpretation of the data. There is, therefore, no reason to think that the end of 13 baktuns from a previous era base would necessarily begin a new world age or era.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In the case of the Venus table, the beginning date was a day 1 Ahau 18 Kayab falling 2200 days before the usual Maya era base, at a date which might be written 12.19.13.16.0 (counted from the previous implicit era base). From this there was a count forward of 9.9.16.0.0 to a date 9.9.9.16.0 1 Ahau 18 Kayab. The interval can be analyzed as 3744 years of 365 days, 2340 conventional Venus rounds of 584 days and 72 calendar rounds. This is 3741 tropical years and 189 days. It is also 2340 true Venus rounds and 189 days. Hence the true relationship of the Sun and Venus would have been the same at both ends of the table but shifted about half a year in the seasons. It is usually supposed that the ending date was approximately in step with reality and the beginning was badly off but it seems strange that they would back-calculate to a base which they knew was grossly in error. The multiplication table starts with 2920 days ($5 \times 584 = 8 \times 365$) going from an implicit 1 Ahau to a specified Ahau. It continues with multiples of 584 by 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60, the last reaching 6 Ahau. It is just possible that the initial 5 Venus cycles are counted from, rather than to, the recorded 9 Ahau, in which case 5 Venus cycles after 6 Ahau would reach 1 Ahau after 2 calendar rounds. We then have a reference to 9100 days reaching 1 Ahau. Read as it stands, this would be 35×260 days, which is not a Venus interval and has usually been regarded as an error. Lounsbury (1990) regarded it as a correction mechanism over 80 calendar rounds. This is certainly more likely than any alternative yet proposed but the specific mechanism suggested involves corrections which seem to me unlikely to have been applied as Lounsbury thought. The table then gives 57×118 and 317 multiples of 584, all of which reach 1 Ahau, before reverting to the 65×584 which restores 1 Ahau, then giving 130×195 and 260×584.

This is followed by a mean (synodic) motion of Venus in 5 sections each totalling 584 days and subdivided into 236, 90, 250 and 8 days. There are three sets of Venus intervals given, creating essentially three tables in one. The base of the top set is 1 Ahau 13 Mac, the base of the second set is 1 Ahau 18 Kayab (the CR date reached by the introduction) and the base of the third set is 1 Ahau 3 Xul. The relationship between these tables is much disputed. Lounsbury's analysis of the relationship between the multiplication table (interpreted as giving various calculations for corrections) and the three-in-one table led him to suggest a date of 11.0.3.1.0 1 Ahau 13 Mac for the top set and a still later date of 11.5.2.0.0 1 Ahau 3 Xul for the bottom set. However, the verbal affix -ah is found in all entries in the 18 Kayab and 3 Xul tables and all but one page (46) of the 13 Mac table. This should mean that the tables are referring to past events, except for part of the 13 Mac table. It seems to be much more likely that the implicit date 9.11.3.2.0 1 Ahau 13 Mac at Palenque in association with apparent Venus calculations is the intended date here as well. That date is 11 960 days after 9.9.9.16.0 1 Ahau 18 Kayab (generally accepted as the intended base for the middle set) which is, in turn, 9360 days later than 9.8.3.16.0 1 Ahau 3 Xul. Lounsbury rejects these dates because of *a priori* assumptions as to how corrections should be applied and the value of common factors. He also assumes all relevant factors of the Venus table are directly related to Venus. I assume, on the other hand, that an important factor of the tables was to relate Venus to other planetary phenomenon, to the tropical year and to eclipses. In this case, an interest in eclipses may be involved, since 11 960 days is the length of the Maya eclipse table ($32 \times \frac{3}{4}$ years) and 9360 seems to be the second most important eclipse interval. This need not imply that any of the Venus table dates were actually eclipse dates, but rather that certain Venus phenomena might serve as markers for eclipses.

This naturally leads to a discussion of the *Dresden* eclipse table. The introductory date falls 8 days after the Maya era base. The usual calculations appear, only partly understood, and then the table. This consists of 69 eclipse syzygies, each with an accompanying text, and with 10 pictures—in all cases except the table end, these pictures follow an interval of 148 days, rather than the normal interval of 177 days (6 lunations) (or occasionally 178 days). The base of the table is the date 9.16.4.10.8 12 Lamat 1 Muan. I have suggested that the calendar names of Mesoamerican deities are the dates reached by a count forward of a relevant astronomical interval from the base 12.8.19.0.8 12 Lamat 1 Pop (of the previous era). In this case, a count forward of 11 960 days reaches 12.10.12.4.8 12 Lamat 1 Muan, which I identify as the calendar name of the eclipse table god. The day 12 Lamat is associated with eclipses in the *Paris* and *Madrid* codices and the highland Mexican *Borgia* codex names an old Moon god as 12 Rabbit, the equivalent of 12 Lamat. If a standardized table of 11 960 days was used as I think, then the *Dresden* table base should be a multiple of 11 960 days from 12.8.19.0.8 12 Lamat 1 Pop and a multiple of calendar rounds from 12.10.12.4.8 12 Lamat 1 Muan. The least common denominator of the two intervals is 48 calendar rounds, which is the equivalent of 73 eclipse table lengths (6.1.5.4.0 or 873 080 days). It turned out that the first common occurrence was at 5.11.17.8.8 12 Lamat 1 Muan and the second at 11.13.2.12.8. 12 Lamat 1 Muan, thus passing over the entire period of the Classic Maya. However, an alternative occurred to me. If one counted 9360 days of the proto-table, one would reach 12.11.18.4.8 12 Lamat 11 Mol. If one then counted forward by multiples of 11 960 days one would reach 12 Lamat 1 Muan after 50 repetitions of the table at 3.14.19.6.8 12 Lamat 1 Muan (32 calendar rounds from 12.10.12.4.8 12 Lamat 1 Muan). If one then counts forward 48 calendar rounds, one reaches 9.16.4.10.8 12 Lamat 1 Muan; the total interval would be 80 calendar rounds plus 11 960 days or 124 eclipse table lengths plus 9360 days (10.12.11.0.0 or 1 530 360 days). The efficacy of that base would then depend in part on the accuracy of the original back calculation and the table itself would serve as a measuring stick against actually observed eclipses (which could theoretically be off by many days or even months). Using modern parameters they would be 41 days off on node passage and 45 days off on lunations, but we do not know what parameters they were using.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It may be noted that the interval from 9.11.3.2.0 1 Ahau 13 Mac to 9.16.4.10.8 12 Lamat 1 Muan is 5.1.8.8 or 36 528 days. This is about 40 days short of an eclipse interval and might suggest that 9.11.3.2.0 was actually an eclipse if the eclipse table base was removed from reality by an equivalent amount.

The first eclipse of the table after 12 Lamat fell on a day 7 Snake, which is the name of a well-known Aztec goddess, Chicomecoatl (Seven Snake). The day 7 Snake is also associated with eclipse syzygy imagery in the *Madrid* codex. The day associated with the first picture in the table is 7 Oc and the associated picture shows a death god on a throne of bones. Again, the *Madrid* codex shows a similar throne of bones with a column of dates including 7 Oc. The so-called 'SAROS' interval of 6585 days from 12 Lamat reaches 7 Ix or 7 Jaguar, said in highland Mexico to be a name for the gourd. The decapitated head of the Sun God was hung in a gourd tree and changed into a gourd. Pottery scenes and sculptures frequently depict the decapitation of the Sun God or Moon Goddess and associated dates suggest eclipses. Perhaps all the gourds in the tree represent previous eclipses. Only a constant back and forth between iconography and calendrical arithmetic will ultimately allow us to understand Mayan astronomy and the role it played in the culture. In the *Madrid* codex there are two cases in which there is a reference to two solar eclipses relative to a single date. In one case, this seems to be 13 tuns after 12 Lamat ($13 \times 360 = 4680$ days). This interval is the ideal short interval to go from a solar eclipse at the node to a lunar eclipse at the node. The latter would be midway between two solar eclipses, one in the far north and one in the far south, neither visible to the Mayas. This suggests at least a minimal concept of the geometry involved and a clear concept that eclipses were occurring even when they were not locally visible.

The table of Mars was satisfactorily identified by Willson in 1924 but was long doubted. More recently, work of the Bricher has again made it acceptable to many scholars, although there are still doubters. I have argued for tables of Mercury, Jupiter and Saturn, still not widely accepted, despite their similarity to the other tables. The stone coffin of Pacal at Palenque has a series of sequential designs which Thomas Barthel identified as symbols of the planets in their Eurasian and African week-day order. This seems to me correct.

Maya constellations

Our knowledge of Mayan CONSTELLATIONS is still very minimal but increasing. The tail of the Rattlesnake is identified with the Pleiades, their Scorpion constellation with ours and Orion's Belt with a turtle. The hearth stones ruled by the Fire God are nearby. I have suggested that the sequence of Mesoamerican day names represents ASTERISMS recognized by the Mayas, but this does not give us very clear definitions. There are problems with other suggested sequences. Some asterisms are still identified by modern Mayan groups.

The correlation problem

Finally, mention should be made of the problem of correlating the Maya chronology with our own. This is usually defined by a correlation constant which matches the era base with a particular Julian Day number. Thus the widely accepted Thompson correlation matches the era base with JD 584 285. I think that this correlation is certainly wrong, disagreeing with colonial and historical evidence, with the archaeology and with the astronomy. Recently, Bryan Wells has proposed alternate correlations at 660 205 or 660 208, the latter being independently suggested by Andreas Fuls. I have found strong historical evidence supporting Wells against my own proposal of 663 310. No yet-proposed correlation agrees well with my interpretation of the *Dresden* eclipse table. The chronological control of a correct correlation is necessary in order to understand Mayan astronomy with any clarity. In my view, we still do not have reasonable certainty about any of the more than 100 proposed correlations, but the Wells correlation is far superior to that proposed by Thompson. Mayan astronomy was complex and sophisticated but it used principles which often differ strikingly from ours, and we still have much work to do.

Bibliography

Justeson J S 1989 Ancient Maya ethnoastronomy: an overview of hieroglyphic sources ed A P Aveni *World Archeoastronomy* (Cambridge: University of Cambridge Press)

Kelley D H 1980 Astronomical identities of the Mesoamerican gods *J. Hist. Astron.* **9** (Archeoastronomy Supplement 2) S1–54

Lounsbury F G 1978 Maya numeration, computation and calendrical astronomy *Dictionary of Scientific Biography* vol 15 (New York: Scribner's) pp 759–818

David Kelley

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mayer, Tobias (1723–62)

Mathematician, born in Marbach, Württemberg, Germany, who, self-taught, discovered the libration (back and forth oscillation on its axis) of the Moon, which won him the chair of mathematics at Göttingen. Made the first accurately measured map of the Moon. Attempted to formulate predictions of the motion of the Sun and Moon as a means to determine longitude at sea, sent them to the British government, which posthumously awarded his widow the sum of £3000. MASKELYNE described the work in the first issue of the *Nautical Almanac*, complimenting their accuracy (they could be used to determine longitude to within a degree, or 60 nautical miles error). Maskelyne noted that 'The Difficulty and Length of the necessary Calculations seem the only Obstacles to hinder them from becoming of general Use'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

McCrea, William Hunter (1904–98)

Theoretical astronomer, relativist and cosmologist, became professor at the University of Sussex. Very influential as a teacher of the principles of cosmology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

McDonald Observatory

McDonald Observatory, located in West Texas near Fort Davis, is the astronomical observatory of the University of Texas at Austin. Discoveries at McDonald Observatory include water vapor on Mars, the abundance of rare-earth chemical elements in stars, the discovery of planets circling around nearby stars and the use of the measurements of rapid oscillations in the brightness of white dwarf stars to deduce their ages.

Telescopes at McDonald Observatory include the 9.2 m (diameter) Hobby–Eberly Telescope (HET) whose partners include (with Texas), Pennsylvania State University, Stanford University, the Ludwig Maximilians University in Munich and the Georg-August University in Göttingen, Germany. The HET came into operation in 1999 and specializes in astronomical spectroscopy. The other telescopes are owned exclusively by Texas and include the 2.7 m Harlan Smith Telescope, the 2.1 m Otto Struve Telescope and the 0.7 m, wide-field survey telescope. In addition, McDonald operates a 0.7 m laser-ranging telescope which measures the distance to reflectors on the Moon.

McDonald Observatory has a staff of 100 people located in Austin and at the Observatory. The operating funds for the Observatory come from the State of Texas, the University of Texas at Austin and from private and corporate benefactors. McDonald Observatory has an extensive set of public information programs including StarDate (English), Universo (Spanish) and Sternzeit (German) radio programs, *StarDate* magazine and a visitor center at the observatory. A new visitor center with extensive exhibits and displays in Spanish and English will be open in late 2001. The new visitor center, at the observatory, is open every day except Christmas Day, Thanksgiving Day, and New Year's Day.

For further information see http://www.as.utexas.edu/mcdonald/mcdonald.html

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MDM Observatory

MDM Observatory was founded by the University of Michigan, Dartmouth College and the Massachusetts Institute of Technology. Current operating partners include Michigan, Dartmouth, MIT, Ohio State University and Columbia University. The observatory is located on the southwest ridge of the KITT PEAK NATIONAL OBSERVATORY near Tucson, Arizona. It operates the 2.4 m Hiltner Telescope and the 1.3 m McGraw-Hill Telescope.

For further information see http://www.astro.lsa.umich.edu/obs/mdm/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mean Solar Time

The time system in use for most civil and many astronomical purposes and based on the motion of a hypothetical object called the mean Sun, the right ascension of which increases from day to day at a uniform rate. The local mean time is defined to be the local hour angle of the mean Sun plus 12 h. Greenwich mean time is taken as the standard for reference; the term Universal Time (UT) is synonymous with Greenwich mean time (GMT).

See also: apparent solar time, equation of time, Universal Time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Measuring Machines

Astronomy is basically an observational science, rather than an experimental one, and the development and advancement of the subject has relied heavily on surveys of the sky at optical wavelengths to expand our knowledge of the observable universe. To date all of the major all-sky optical surveys have been carried out using photographic plates taken on wide-angle SCHMIDT TELESCOPES. Of particular note are the northern hemisphere surveys undertaken by the Palomar Schmidt telescope, and the southern hemisphere surveys using the ESO Schmidt telescope in Chile and the UK Schmidt telescope (UKST) in Australia. These surveys form a basic foundation of observational astronomy playing a similar role to cartographic maps of the Earth's surface.

Formerly, this photographic survey material could only be studied visually with the aid of low-power binocular microscopes. Even with large teams of dedicated observers only a minute fraction of the available wealth of information present on the survey plates could be mined. Objective, quantitative studies were difficult in many cases and impossible in most. However, in the late 1970s and early 1980s the advent of fast optical micro densitometers, or measuring machines, capable of automatically digitizing and routinely analyzing the enormous information content of these survey plates, revolutionized optical astronomy, placing it firmly at the forefront of the computer age. By routinely processing the 4 Gbytes of information on a Schmidt plate in a matter of a few hours, these machines opened up new areas of quantitative survey astronomy. Around 2000 such plates are needed to cover the whole sky in one passband. Add another passband, or two, to provide colour information and possibly different epoch observations to open up proper motion and variability studies, not to mention special plates for specific studies, and the total data volume becomes well over a staggering 50 Tbytes.

This ability to digest huge volumes of information (see also SOFTWARE: DIGITAL SKY SURVEYS) and produce the sort of numbers astronomers are really interested in, such as position, magnitude, colour, star or galaxy and so on, is the key to the success of the measuring machine era. Most of the area on a survey plate belongs to the smoothly varying sky background and hence can be accurately represented using much lower-resolution sampling. Faint images and all stellar images are reducible to simple descriptors such as those above. It is only the handful of bright galaxies where more detailed mapping is required. These straightforward concepts have been exploited by the measuring machine community to significantly compress the original data volume by factors of typically $\times 100$ with little or no loss of pertinent information. Nowadays, it is common practice among astronomers to use the world wide web (WWW) to dial up arbitrary parts of the sky and be presented with catalogs of objects and two-dimensional pictures.

The initial breakthrough came in the 1950s, when Peter Fellgett proposed a machine that could look for and measure images automatically. It was called GALAXY and used a 'flying spot' produced by a cathode ray tube (CRT). When the image of a star on the plate blocked off the light from the CRT the machine registered an object and measured its position using an analogue implementation of a maximum likelihood (i.e. optimal) centering algorithm. GALAXY, completed in 1968 at the Royal Observatory Edinburgh (ROE) by Vincent Reddish, represented a significant advance over previous efforts, but it was still very slow. It could process 1000 stars per hour but, as each survey-grade Schmidt plate contains well over 250 000 stars, measuring a complete plate still took an impracticably long time. Furthermore, as GALAXY was designed to do stellar astrometry only, much of the information content of the plates was still untapped.

GALAXY's successor, COSMOS, also based on a CRT-illuminated flying spot, was commissioned at the ROE in the late 1970s. Unlike GALAXY, COSMOS was designed to digitize first, prior to performing image detection and analysis, and took about a day to process each plate completely. In this way, both stellar and galaxy images could be detected and parameters describing their position, intensity and morphology readily derived. COSMOS has been used to construct digital maps and catalogs of many UKST plates including the complete Southern BJ sky survey and these are available on the WWW. COSMOS was retired in 1993 and has recently been superseded by Super COSMOS, a multichannel system based on a linear CCD and xenon arc lamp source. Telecentric optics helps to control the scattered light. The X–Y table is a modification of a commercial off-the-shelf system and represents the current state-of-the-art in accurate (few tenths of a micron) positional measurement. At the same time, the sampling interval, in this case 10 μm, and the processing speed have been enhanced such that a complete Schmidt plate has a turnround time of some 3 h. Programs for SuperCOMOS include rescanning the UKST BJ survey plates, incorporating the UKST OR and I survey plates and measuring the second epoch Palomar sky survey plates (POSSII) of the northern hemisphere. Both digital maps and image catalogs will be made available over the WWW.

Meanwhile back in Cambridge, also in the 1970s, Ed Kibblewhite, an electronics graduate, started work on a different system, the APM machine. Kibblewhite's breakthrough was to use a laser beam to scan the plate and a series of high-speed special-purpose computers to analyze the data in real time. An acousto-optic deflector generates the 'flying spot', or multichannel parallelism, necessary for fast scanning, by using a sound wave of varying frequency to diffract parts of the incident laser beam. The plate is held on a massive X–Y table which moves under the laser beam and the scanner samples the plate 250 000 times per second with a positional accuracy of a few tenths of a micron. The APM has been used to scan the first epoch Palomar sky survey plates (POSSI) of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

northern hemisphere and the UKST BJ, OR and I survey plates. Catalogs for these plates are available on-line for fields at GALACTIC LATITUDES further than 20° from the plane.

In the US, Willem Luyten had pioneered the use of a manual BLINK COMPARATOR for locating stars of high proper motion. This early work culminated in the publication of measured proper motions of 94 263 stars that form the Bruce Proper Motion Survey, which he finished in 1937. This work continued and Luyten went on using the blink comparator to measure the proper motions of about 100 000 other stars. However, as this work extended to ever fainter stars, measurement by hand became too difficult. As a result, he and the Control Data Corporation designed a new fully automated computer-controlled blink machine based on laser optics, which was constructed in the late 1960s. This machine was specifically designed to do accurate astrometry and played a major role in the production of Luyten's Palomar proper motion catalogs.

After a major electronic upgrade in the 1980s under the direction of Roberta Humphries, this rather specialized machine was renamed the APS and was used on a wide-ranging program to digitize the POSSI red and blue sky survey plates. The APS is based on a laser spot deflected with a rotating prism and is capable of scanning two plates simultaneously in a few hours. Current projects include remeasuring the Luyten proper motion red survey plates and incorporation of catalogs based on POSSII. All the catalogs are available on the WWW.

Similar astrometric measurements using manually operated machines had also been undertaken at Lick Observatory as part of the Lick Proper Motion program. These culminated in the development of the Lick Automatic Measuring Engine (LAME) by Vasilevskis and Popov in 1966. Approximate positions of targets were read from punched cards, and used to drive the plate carriage to within 50–100 μm of the target. Photometric and astrometric measurements were made and recorded on a new card, with an iris being used for photometry and a spinning-sector scanner for centering. Initially LAME required frequent operator intervention, but later modifications to the electronics and computer system in the early 1970s greatly enhanced the efficiency and degree of automation, enabling the goals of the Lick Proper Motion program to be achieved.

Throughout all this, the main difference in focus between UK and US effort in measuring machine development was that GALAXY was the forerunner for a series of general-purpose machines in Britain, while in the US measuring machines were used solely for astrometry until the latter half of the 1980s (see also ASTROMETRY: HISTORY).

One of the main drivers for the latter-day expansion of measuring machine effort in the US was the mission requirement for the Hubble Space Telescope (HST), for a catalog of acquisition and guidance stars covering the whole sky down to roughly 15th magnitude in the visual (see also ASTROMETRY WITH HUBBLE SPACE TELESCOPE). In order to guarantee this objective Barry Lasker at the Space Telescope Science Institute (STSci), in Baltimore, initiated a program using two PDS micro densitometers to scan the Palomar Quick V survey and the UKST BJ survey. This catalog was made widely available to the astronomical community and has become the mainstay for ground-based telescope acquisition and guidance systems too. PDS machines are classical single-channel two-microscope instruments and, although slow compared with the other measuring machines previously mentioned, have the advantage of superior dynamic range. However, by digitizing the survey plates at a more modest resolution, and by being operated around the clock, they were able to digitize the whole sky prior to the launch of HST. In addition to the Guide star catalog, digitized maps of the whole sky were also made available for the first time. The STSci PDS systems were recently upgraded to include laser illumination and an acousto-optic deflector to provide a small number of channels, a compromise hedging between control of scattered light (a problem in all flying spot scanners) and increase of scanning speed. The upgraded system is being used to measure the UKST OR and POSSII survey plates. All the data, including digitized maps and object catalogs, are being made available on the WWW for public access.

The only other major system that is scanning large amounts of sky survey plates is Monet's PMM machine at the US Naval Observatory, Flagstaff. This machine is based on two-dimensional CCD cameras and features extensive software to correct for astrometric field effects in the individual footprint scans from each CCD camera. The PMM is capable of scanning four plates simultaneously in a few hours and has been used to measure all of the POSSI, UKST and POSSII blue and red survey plates, as well as a special set of short exposure plates in the north made on the same grid to improve the astrometric reference frame. Although the main goal of the project was a new ASTROMETRIC SURVEY, an all-sky object catalog with magnitude and color information for over 500 million objects to the limiting depths of the plates, is available on request.

The future of photographic sky survey material is guaranteed by the past. Although the current generation of all-sky survey material is based on photographic plates, the new large-format mosaic CCD cameras are performing the next generation of optical sky surveys to much deeper limits. However, the historic record provided by a century of photographic plates is invaluable and ensures that both archival plates in addition to sky survey plates will have a strong role to play in the foreseeable future.

Bibliography

A good introductory general reference is

Lasker B M 1995 *Publ. Astron. Soc. Pacific* **107** 763

Mike Irwin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Media

Astronomy is one of the sciences that can be experienced by everyone. The skies can be seen directly by all, and it is possible through disciplined observation for everyone to make contributions to current science, e.g. through observations of variable stars, comets, etc. An enhanced view of the sky is available through incredibly beautiful images from the Hubble Space Telescope (see http://oposite.stsci.edu/pubinfo/subject.html) and large ground-based instruments, such as the Anglo-Australian Telescope (see http://www.aao.gov.au/images.html).

Astronomy is mind expanding. The large distances, the effects of the connection between distance and time past ('look-back'), contemplation of the origin of the planets, the Sun, the stars and the Universe, the possibility that there are planets similar to the Earth elsewhere in space: amazing thoughts like these provoke wonder and excitement (see ART AND LITERATURE).

Thus astronomy is one of the most popular sciences, and is represented strongly in the information media, such as newspapers and broadcasting.

Astronomy is also connected with the adventure of space travel. It was possible to follow on TV the landings on the Moon made by the Apollo astronauts, and the later flights of the Space Shuttle. NEIL ARMSTRONG's 'one small step' onto the bleak rocks of the lunar Sea of Tranquillity was watched by many millions of viewers, world wide. Through the internet, very large numbers of people indeed followed the landing of Mars Pathfinder and the explorations by its rover Sojourner. Under its Administrator Dan Goldin, NASA has put an emphasis on a scientific return to the tax-payers of the USA who finance the NASA program, and there has thus been a rise in the opportunities for public participation in NASA space missions. The European Space Agency is increasing its emphasis on the public understanding of science.

Sound broadcasting does not have the immediacy of appeal of television, but is better suited to convey the ideas of astronomy. The superlatives of distance, time, size and energy appeal directly to the imagination but can best be conveyed through words.

A pioneer 'broadcast astronomer' in the UK was Sir JAMES JEANS (1877–1946). Midway in his academic career he was invited to write a popular book and produced *The Universe Around Us* (1929), which was an immediate success. He was soon invited to make a series of radio broadcasts, and these too were so well received that the rest of his career was concerned largely with the popularization of astronomy. Other early broadcasters included Sir ARTHUR EDDINGTON (1882–1944) and Guy Porter, of the Royal Observatory, Greenwich. FRED HOYLE's *The Nature of the Universe* (c 1955) was an influential BBC radio series that developed a reputation for the steady-state theory of the origin of the universe that still survives, although the theory itself is moribund.

The BBC's *The Sky at Night* television series began in 1957, presented by PATRICK MOORE; the programs have been broadcast late at night every four weeks since that time, and are said to be the longest-running TV series in Britain apart from the news. The series has given a large number of astronomers access to an audience of 2 or 3 million in the UK to whom they can directly communicate their work, as well as developing a large number of astronomical media 'experts'. There have also been various other shorter-lived series, on all networks, particularly the BBC channels and Channel 4.

In the United States, CARL SAGAN's *Cosmos* television program attracted a vast following. The Discovery channel in the USA broadcasts many astronomy programs. In France, the annual *Nuit des Etoiles* devotes a whole evening to astronomy and space, with segments of the program originating from across the country. Curiously, no country other than Britain has produced a regular television series like *The Sky at Night*, certainly not one of comparable length.

In presenting a scientific program, there has to be a compromise between the overtechnical and the oversimple approach. Therefore, the facts have to be given clearly without any impression of 'talking down' and to go into more detail means that the script writer and the presenter have to be highly skilled. Confusing programs tend to turn viewers away from science.

Ideas and images combine in the printed media, both in book form and in journals, magazines and newspapers. Newspapers carry predictions of astronomical phenomena, lunar phases, star charts, etc, as well as news reports on astronomical advances. In France, NICOLAS CAMILLE FLAMMARION (1842–1925) was especially noteworthy as a pioneer popularizer of astronomy through these forms, including many lavishly illustrated books (for example *L'Astronomie Populaire*) and a journal *L'Astronomie*, which he founded with the Société Astronomique de France in 1897 and which is still published. In the United States, the magazine *Sky and Telescope*, the largest-circulation astronomical magazine (circulation 125 000) traces its history to a four-page bulletin called the *Amateur Astronomer*, begun in 1929 and absorbed by a monthly magazine called *The Sky* started 1935 by the Hayden Planetarium. *The Sky* went independent in 1941 and merged with *The Telescope*, which had first appeared as a quarterly magazine in March 1931. In Britain, *Astronomy Now* has a circulation between 25 000 and 30 000.

Astronomical and space stories in the press are second only to stories about human biology in attracting public interest. In Britain, the decade of the 1990s has seen in magazines and newspapers an increase in the proportion of articles about astronomy. The rate of increase relative to the volume of published material is some 10% per year. This rise has been associated with a rise in enrolment in university courses in astronomy, doubling over a decade, and an increasing number of universities offering such courses (about half of British universities do so). There is evidence from surveys in Britain that an interest by young people in astronomy and space plays a key role

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in attracting undergraduates into university studies of science and technology in general.

The world geographic distribution of institutes for astronomy, observatories, planetaria, astronomical societies and astronomical publishers is given by André Heck (http://vizier.u-strasbg.fr/~heck/aamaps.htm; see also A Heck *Astronomy Now* August 1998). The Eastern and the Western United States and Japan, the Netherlands, central European countries and the rest of Europe have a strong density of astronomical interest. France has a strong centralization of research in the Paris area (which has the highest concentration in the world of astronomy organizations). By contrast, the distribution of interest in astronomy in the USA and the UK is broadly based and follows the population distribution, particularly the astronomical societies. The London area is a world centre for global astronomical publishing.

The Royal Astronomical Society (http://www.ras.org.uk/ras/press/press.htm) and the American Astronomical Society maintain press offices that regularly issue press releases on scientific stories of astronomical interest. The take-up is by editors who are keen to balance the diet of serious political news, disturbing crimes and tragic accidents, which they feel obliged to cover, with more optimistic stories.

The key to the continued production of successful press releases includes above all the maintenance of scientific integrity, but there are additional factors which help the press take up a science story. There should be some national or local interest in it; it is helpful if one of the scientists is someone with whom the reader can identify. The international nature of science is helpful in getting an astronomical story published in many countries. In general, press interest increases with size of expenditure—the more money, the more important. Journalists are also influenced by the availability of scientists for follow-up interviews—each journalist looks for his/her unique 'angle' to the story via direct contact.

Pictures, especially color pictures, help produce a favorable editorial decision to publish. Such a decision is made by the scientifically uninformed, and an editor can be swayed by a picture (or an animation, in the case of TV). If the story contains a superlative (biggest, furthest, most massive, . . .) or a fancy name or picturesque metaphor, it is also more likely to be published. A widely reported astronomical press release in 1998 in the UK was about a rather technical study of peanut-shaped bipolar instabilities in star distributions within rotating barred spiral galaxies. This is not promising material for a popular scientific press release. Its take-up was due to the inspired label on the press release, using the name of a chocolate-nut confection, and referring to the 'Snickers hypothesis', that there is a peanut in every bar. It was possible to combine the scientific integrity of the work with enjoyment, and this is the essence of the presentation of astronomy in the information media.

Paul Murdin and Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Medieval Cosmology

During the early Middle Ages (ca 500 to ca 1130) scholars with an interest in cosmology had little useful and dependable literature. They relied heavily on a partial Latin translation of PLATO's *Timaeus* by Chalcidius (4th century AD), and on a series of encyclopedic treatises associated with the names of Pliny the Elder (ca AD 23–79), Seneca (4 BC–AD 65), Macrobius (fl 5th century AD), Martianus Capella (ca 365–440), Cassiodorus (ca 480–ca 575), Isidore of Seville (ca 560–636) and Venerable Bede (672–735). Even if we add certain works of the church fathers, especially St Basil (ca 330–379) and St Augustine (354–430), the cosmology embedded in all of these treatises was meager and insubstantial. Beginning in the 12th century, this changed dramatically when a large number of Greek and Arabic scientific treatises were translated into Latin, thus laying the foundation for the development of early modern science in the 17th century.

During the late Middle Ages, cosmology was not an independent discipline, but a branch of natural philosophy, a subject taught at the medieval universities, which had come into existence by 1200. Medieval cosmology was derived primarily from the natural philosophy of ARISTOTLE (ca 384–322 BC), the great Greek philosopher and scientist. Cosmology was also dependent on mathematical astronomy, which was associated with the name of Claudius PTOLEMY (ca 100–ca 170). The works of Aristotle relevant to cosmology, and the works of Ptolemy relevant to astronomy, were unknown in Western Europe until the second half of 12th century, following their translations into Latin from Greek and Arabic.

The large body of cosmological literature produced in the Middle Ages derives largely from questions that medieval natural philosophers drew from Aristotle's books on natural philosophy, especially his treatise *On the Heavens*. Nicole ORESME (ca 1320–1382), a famous natural philosopher and theologian, discussed the following questions in his treatise on Aristotle's *On the Heavens*:

> Whether it is possible that several worlds exist.
>
> Whether the heaven is of another nature than inferior things.
>
> Whether all the stars receive their light from the Sun or [whether] some [stars produce light] in themselves.
>
> Whether the entire Earth always rests.

Responses to these, and hundreds of similar, questions, posed over the centuries, formed the basis of medieval cosmology. There were also straightforward commentaries on Aristotle's works, together with questions and commentaries on a widely used 13th century work by JOHN OF HOLYWOOD, titled *Treatise on the Sphere*.

Aristotle and Christianity

The introduction of Aristotle's works posed a dilemma for medieval natural philosophers. From the Bible, they learned that the world was created by a divine power, who would eventually destroy it. From Aristotle's *On the Heavens* came a powerful contrary message: the world could not have had a beginning and could never come to an end. Aristotle's conviction that the world was without beginning or end conflicted with the basic Christian belief that our unique world was divinely created and of temporal duration. Indeed, by the early 13th century, it was accepted dogma in Latin Christendom that the world was created out of nothing (*ex nihilo*). Reactions to Aristotle's claims for eternity varied. Some, like St Bonaventure (1221–1274), were convinced that the temporal creation of the world was demonstrable. By contrast, Thomas Aquinas (ca 1225–1274) held that logical argument and evidence favored neither judgment. In his view, there was no contradiction in the possibility that God could have willed the existence of the world without a temporal beginning, making it both created and eternal.

The creation account in Genesis compelled Christian cosmologists to explain biblical claims. How, for example, does the heaven (*caelum*), or firmament, created on the first day, differ from the heaven (*caelum*) created on the second day? How does the light created on the first day compare with the light created on the fourth day? How could plants come forth on the third day when the Sun, whose warmth and light are required, was not created until the fourth day? What are the waters above and below the firmament? The various solutions to these questions were proposed in terms of contemporary Aristotelian physics and cosmology.

Can anything exist beyond our finite cosmos?

Most natural philosophers believed the world was a finite sphere. Did anything lie beyond it? Aristotle had argued that it was impossible for bodies, motions, places, time and vacuum to exist outside of our finite cosmos. To Christians, Aristotle's argument implied that not even God could create these things, and therefore He could not create other worlds beyond ours. Aristotle's arguments were viewed as placing constraints on God's absolute power to do as He pleased. In 1277, the Bishop of Paris condemned 219 articles that were deemed theologically offensive and unacceptable. Many of them were directed against Aristotle and his followers. One of the condemned articles claimed that God could not make other worlds. To repudiate Aristotle's claim, scholastic authors showed, by a variety of arguments, that identical and distinct worlds—and this type was most frequently discussed—could exist simultaneously. Many insisted that each of these identical worlds could exist as a self-contained, independent, closed system with its own center and circumference, thus rejecting Aristotle's argument for the existence of a unique center and circumference from which he had inferred a unique spherical world. The possibility of other worlds led to further challenges to Aristotle's authority. What Aristotle had deemed impossible was shown to be possible and intelligible.

Not all extracosmic entities were merely possible. One was real. On theological grounds, Thomas Bradwardine (ca 1290–1349), and a few other theologians,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

assumed the existence of a real, infinite space beyond our world, a space Bradwardine identified with God's infinite, omnipresent immensity. Because God was a being devoid of dimensions, the theologians who identified infinite void space with God's omnipresent immensity followed Bradwardine and also characterized that space as without extension. Bradwardine's ideas about space and God form an integral part of the history of spatial conceptions between the late 16th and 18th centuries, the period of the scientific revolution.

The celestial region of the cosmos

Aristotle divided the world into two radically different parts: (1) a celestial region of stars, planets and orbs that ranged from the Moon to the sphere of the fixed stars, and beyond, and was composed of a special ether, or fifth element; (2) a sublunar realm immediately below, ranging from the concave surface of the lunar orb to the center of the Earth and encompassing four elements—fire, air, water and earth—and the bodies compounded of them.

The celestial region, which is the part that will be described here, consisted of stars, planets and the orbs that carried them around the heavens. They were all made of the same transparent ether, which was assumed indivisible and incorruptible. Although the transparent orbs were invisible, the stars and planets were visible, because they were highly concentrated regions of ether that reflected light to Earthly observers. Most scholars assumed that the stars and planets received their light from the Sun and then transmitted that light to the Earth. Some further suggested that the planets were also weakly self-luminous. Both planets and stars were regarded as spherically shaped.

In what was a virtually self-evident truth, derived ultimately from Aristotle, the celestial bodies were believed to govern the behavior of material things in the terrestrial region below the lunar orb. Not only was this the basis of astrology, but it played an important role in natural philosophy. The influence of the celestial bodies was accomplished by their motions, by light transmitted from their bodies and by general influences that were invoked to explain many otherwise inexplicable phenomena (for example, the formation of metals within the depths of the Earth, magnetism and the tides caused by the Moon).

Natural philosophers were agreed on the existence of seven planets and the fixed stars. Without the detection of parallax, the order of celestial bodies was made on non-astronomical grounds. Since the planets moved against the background of the stars, the latter were assumed farthest from the Earth and equidistant from it. The superior planets—Saturn, Jupiter, and Mars—were ranked by their sidereal periods in that order. Because Venus and Mercury always remained in the vicinity of the Sun, the order of these three planets was not determinable except by non-astronomical arguments. Most followed Ptolemy's arrangement and located the Sun, Venus and Mercury in descending order, followed by the Moon. Because of its life-giving properties, the Sun was regarded as the largest, noblest and most important planet. Counting upward from the Moon or downward from Saturn, the Sun occupied the fourth, or middle, position and was perceived as a 'wise king in the middle of his kingdom' or as the 'heart in the middle of the body'.

The celestial spheres

Although to the naked eye it appeared that planets were self-moved, like fish in water or birds in air, natural philosophers denied appearances and assumed, with Aristotle and Ptolemy, that each planet was embedded in a single transparent, invisible, ethereal sphere, which carried it around. Similarly the stars, which always remained at fixed intervals from each other, were located on a single sphere, which carried them around. Not only was each planet assigned a single sphere, but each motion of each planet—daily motion, sidereal motion, motion in latitude, etc—was also assigned its own orb. Although Ptolemy and Aristotle assumed that each planet was attached to a single sphere, both employed a plurality of spheres to account for the resultant motion of each planet. By assigning one orb to each motion of a planet, both Aristotle and Ptolemy were compelled to assign multiple orbs to each planet. Aristotle assigned as many as 55 for all the planets, while Ptolemy assigned as many as 41.

Aristotle's spheres, however, differed from Ptolemy's. Aristotle's orbs were concentric with respect to the Earth, which lay at the center of the world, whereas Ptolemy's were basically eccentric. This meant that the centers of the planetary orbits were geometric points that did not coincide with the Earth's center, which was assumed identical with the geometric center of the world. It followed that the planets orbited around centers other than the Earth's center, a clear violation of Aristotle's dictum that all planetary orbs move around the Earth's center. It became apparent that Aristotle's concentric orbs could not account for observed variations in the distances of the planets and that Ptolemy's eccentrics and epicycles, which had been devised to cope with that problem, could. Natural philosophers confronted a dilemma: either adopt Aristotle's concentric system with the Earth serving as the center of all planetary orbits, but abandon any attempt to account for variation in planetary distances, or adopt Ptolemy's system of eccentrics and account for the variation of planetary distances, but yield up the central tenet of Aristotle's natural philosophy, namely that the Earth is the sole center of planetary orbits.

Medieval natural philosophers adopted a compromise that retained both concentricity and eccentricity. It was a compromise that had already been made by Ptolemy (in his *Hypotheses of the Planets*) and hinged on a distinction between the concept of a 'total orb' (*orbis totalis*) and a 'partial orb' (*orbis partialis*), to use medieval terminology. The 'total orb' is a concentric orb whose center is the center of the Earth, whereas a 'partial orb' is an eccentric orb whose center is a geometric point lying outside the center of the world. The concentric total orb contains within

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

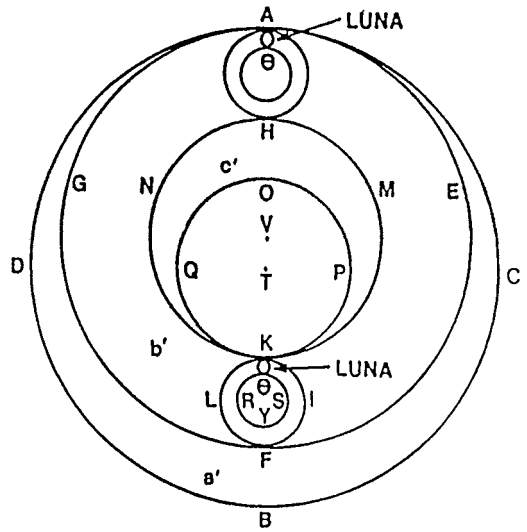


Figure 1. Representation of the Moon's concentric, eccentric and epicyclic orbs as described in Roger Bacon's *Opus tertium* in Duhem P (ed) 1909 *Un fragment inédit de l'Opus Tertium de Roger Bacon précedé d'une étude sur ce fragment* (Florence: Collegium S. Bonaventurae) p 129.

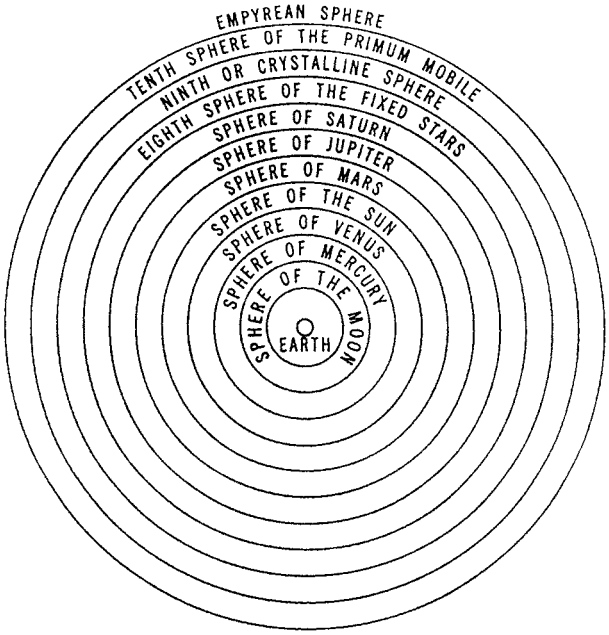


Figure 2. The order of the celestial orbs in the medieval cosmos. Reprinted from Grant E 1977 *Physical Science in the Middle Ages* (Cambridge: Cambridge University Press) p 72.

its concave and convex surfaces at least three partial orbs, one of which, the eccentric deferent, carries an epicycle in which a planet is embedded. (In figure 1, which represents the Moon, the three partial orbs lie in the spaces between ADBC and AGFE, AGFE and HNKM, and HNKM and OQKP.) In this way, Ptolemy's eccentric orbs were encompassed and encircled by the surfaces of Aristotle's concentric orbs.

The number of celestial orbs could thus be determined in two different ways: either as the totality of concentric, or 'total', orbs or as the totality of eccentric, or 'partial', orbs. Aristotelian natural philosophers were interested in the former and rarely in the latter. In fact, the most popular question in medieval cosmology was about the number of spheres. The response required a count of total, or concentric, orbs. The answers to this question varied anywhere from 8 to 11 (see figure 2). Eight concentric orbs represented the seven planets and the sphere of the fixed stars. However, as many as three different motions were assigned to the sphere of the fixed stars (the daily motion, precession and a motion called trepidation that turned out to be non-existent). Applying the principle that a single sphere must be assigned to each distinct celestial motion, two additional spheres, plus an immobile empyrean sphere, were usually added for a total of 11. No celestial bodies were embedded in these three invisible orbs.

The eighth to eleventh orbs had important scriptural associations. The biblical firmament was most frequently identified with the eighth sphere of the fixed stars. Early in the history of Christianity, the waters above the firmament were conceived as crystalline, a term that was sometimes thought of as applying to fluid waters, and sometimes to waters that were congealed and hard, like a crystal. The crystalline orb was usually located above the firmament and was identified with a starless ninth sphere and occasionally with both the ninth and tenth spheres. The outermost orb of the universe was the empyrean sphere, which, in contrast to the perpetually moving orbs it enclosed, was immobile. It was conceived as 'the dwelling place of God and the elect', a place filled with the purest light. If the number of mobile orbs was n, the empyrean orb was always numbered $n + 1$.

Medieval natural philosophers frequently asked: what caused the orbs to move? Most believed that each orb was assigned an immaterial, spiritual intelligence distinct and separate from the orb it moved. Intelligences, often identified with angels, were conceived as voluntary agents divinely endowed with inexhaustible powers to move their respective orbs with uniform circular motion. Not all scholastics invoked intelligences. A small minority held that God created the orbs with an intrinsic power for self-motion. John BURIDAN (ca 1295–ca 1358) even invoked a divinely impressed force, or impetus, that enabled the orbs to move forever through a celestial ether that was assumed to offer no resistance to their motion.

The status of the Earth at the center of the world was much debated. John Buridan held that the Earth's center of gravity was always changing as a consequence of geological alterations of the Earth's surface. These alterations caused the Earth to move continually with small movements as it sought to bring its center of gravity into coincidence with the geometric center of the universe. Buridan and Nicole Oresme contemplated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a more radical motion of the Earth when they asked whether the daily rotation of the heavens might not actually derive from a daily axial rotation of the Earth. For different reasons, they both eventually rejected this alternative, but were agreed that, on the assumption of the immobility of the heavens, the Earth's daily rotation saved the astronomical phenomena just as well as the traditional opinion. NICOLAUS COPERNICUS (1473–1543) found their arguments in favor of an axially rotating Earth worthy of inclusion in defense of his heliocentric system, in which the Earth is assigned both a daily rotation and an annual motion around the Sun.

By modern standards, the medieval cosmos was miniscule. In a popular 13th century astronomical treatise by Campanus of Novara, the distance from the center of the Earth to the concave surface of the fixed stars, which was the outermost orb that contained celestial bodies, was given as approximately 73 million miles. The lunar orb was only 108 000 miles from the Earth, while the Sun was approximately 4 million miles away.

Medieval, Aristotelian cosmology remained functional until the end of the 17th century, by which time the efforts of Copernicus, TYCHO BRAHE, GALILEO GALILEI, KEPLER and NEWTON had produced a new cosmology, which rendered the old obsolete and irrelevant.

Bibliography

Benjamin F S Jr and Toomer G J (ed and transl) 1971 *Campanus of Novara and Medieval Planetary Theory: Theorica Planetarum* (Madison, WI: University of Wisconsin Press)

Grant E 1994 *Planets, Stars and Orbs: The Medieval Cosmos 1200–1687* (Cambridge: Cambridge University Press)

McCluskey S C 1998 *Astronomies and Cultures in Early Medieval Europe* (Cambridge: Cambridge University Press)

Edward Grant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Menelaus of Alexandria (*c*. 70–*c*. 130)

Greek mathematician, born in Alexandria, Egypt (possibly), wrote many books, of which only one, *Sphaerica*, on spherical astronomy, has survived, and including *Menelaus Theorem*, on a property of spherical triangles. PTOLEMY mentions astronomical observations made by Menelaus in Rome in the year 98.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mensa

(the Table; abbrev. Men, gen. Mensae; area 153 sq. deg.) A southern constellation which lies between Hydrus and Volans, and culminates at midnight in mid-December. It was named Mons Mensae (Table Mountain) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky from the Cape of Good Hope, South Africa, in 1751–2, as the presence of the southern portion of the Large Magellanic Cloud in the northern part of the constellation reminded him of cloud enveloping the summit of Table Mountain.

A small, inconspicuous constellation, the brightest star in Mensa is of magnitude 5.1. Interesting objects include NGC 1841, an eleventh-magnitude globular cluster, and IC 2051, an eleventh-magnitude spiral galaxy.

See also: Large Magellanic Cloud.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Menzel, Donald (1901–76)

Solar astrophysicist, born in Florence, CO, became director of Harvard Observatory. Successful popular author of astronomy books.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Merak

The star β Ursae Majoris, apparent magnitude 2.34, one of the leading pair of stars in the Plough known as 'the Pointers'. The Arabic name means 'the Loin' (of the Bear).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mercator, Gerardus [Gerhard Kremer] (1512–94)

Cartographer ('cosmographer'), born in Rupelmonde, Flanders (now Belgium), studied mathematics, its applications to geography and astronomy, and how to make scientific instruments. Constructed terrestrial and celestial globes. Was charged with heresy, a result of suspicion aroused by his travel to collect data for maps, and was imprisoned for several months, after which charges were dropped. Invented the *Mercator map projection* (1569) which has perpendicular longitudes and latitudes, and an expansion towards the poles, such that the bearing from one place to another can be found by a protractor, without the use of spherical trigonometry (a development of an earlier Chinese discovery). He published a comprehensive geographical 'atlas' (he was the first to use the word in this way). His break with PTOLEMY was as important for geography as was Copernicus's for astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mercury

Mercury is the innermost and second smallest planet in the solar system (Pluto is smaller). It has no known satellites. To date, MARINER 10 is the only spacecraft that has explored the planet. It flew past Mercury on 29 March and 21 September 1974, and on 16 March 1975. Mariner 10 imaged only about 45% of the surface at an average resolution of about 1 km, and less than 1% at resolutions between about 100 to 500 meters. This coverage and resolution is comparable to telescopic Earth-based coverage and resolution of the MOON before the advent of space flight. The Mariner 10 data, however, are still our best source of detailed information on the planet. As a consequence, there are still many uncertainties and questions concerning the history and evolution of Mercury. However, recent ground-based observations have provided information on the radar and microwave characteristics of its surface, discovered new constituents in its exosphere, and helped constrain its surface composition. Because of its close proximity to the Sun, missions to Mercury are extremely difficult; more so than to the outer solar system. However, a new mission to Mercury called MESSENGER is scheduled for launch in 2004. It will make one flyby of Earth, two of Venus and two of Mercury before going into Mercury orbit in 2009. This mission will provide vital new information on this poorly known planet. Mercury missions are also being considered by the European and Japanese space agencies.

General characteristics

Mercury's diameter is only 4878 km, but it has a relatively large mass of 3.301×10^{23} kg. Because of its large mass in relation to its volume, Mercury has an exceptionally high mean density of 5.44 g cm^{-3}. The manner in which it reflects light is very similar to the way light is reflected by the Moon. Mercury is covered with a REGOLITH consisting of fragmental material derived from the impact of meteoroids over billions of years. Mercury's surface is heavily cratered with smooth planes filling and surrounding large impact basins. Long lobate scarps traverse the surface for hundreds of kilometres, and large expanses of intercrater plains (the most extensive terrain type) fill regions between clusters of craters in the highlands. A terrain consisting of large blocks and linear troughs occurs antipodal to the CALORIS BASIN. Although Mercury superficially resembles the Moon, there are major differences which set it apart from the Moon, and all other planets and satellites in the solar system.

Motion and temperature

Mercury has the most eccentric (0.205) and inclined (7 degrees) orbit of any planet except Pluto. Although its average distance from the Sun is 0.3871 AU (5.79×10^7 km), its large eccentricity results in a distance varying from 0.3075 AU (4.6×10^7 km) at perihelion to 0.4667 AU (6.98×10^7 km) at aphelion. Consequently, Mercury's orbital velocity averages 47.6 km s^{-1}, but varies from 56.6 km s^{-1} at perihelion to 38.7 km s^{-1} at aphelion.

Mercury's rotation period is 58.646 Earth days and its orbital period is 87.969 Earth days. Therefore, it has a unique 3:2 resonant relationship between its rotational and orbital periods. It turns exactly three times on its axis for every two orbits around the Sun. This resonance was apparently acquired over time as the natural consequence of the dissipative processes of tidal friction and the relative motion between a solid mantle and a liquid core. Because of this resonance, a solar day (sunrise to sunrise) lasts two Mercurian years or 176 Earth days. The obliquity of Mercury is close to 0 $^\circ$ and, therefore, it does not experience seasons. Consequently, the polar regions never receive the direct rays of sunlight and are always frigid compared to very hot sunlit equatorial regions.

Another effect of the 3:2 resonance between the rotational and orbital periods is that the same hemisphere always faces the Sun at alternate perihelion passages. The hemisphere facing the Sun at one perihelion will rotate one and a half times by the next perihelion, placing it directly facing the Sun again. Another consequence of the 3:2 resonance is that an observer on Mercury (depending on location) would witness either a double sunrise, a double sunset, or the Sun would backtrack in the sky at noon during perihelion passage. Near perihelion Mercury's orbital velocity is so great compared to its rotation rate that it controls the Sun's apparent motion in the sky as viewed from Mercury.

Because Mercury is so close to the Sun, has no insulating atmosphere, and has such a long solar day, it experiences the greatest range (day to night) in surface temperatures (635 K) of any planet or satellite in the solar system. Its maximum surface temperature is about 725 K at perihelion on the equator, but drops to about 90 K at night.

Exosphere

Mercury has an extremely tenuous atmosphere with a surface pressure a trillion times less than Earth's. The number density of atoms at the surface is only 10^5 atoms cm^{-3} for the known constituents. It is, therefore, an exosphere where atoms rarely collide; their interaction is primarily with the surface. Mariner 10's ultraviolet spectrometer identified hydrogen, helium and oxygen, and set upper limits on the abundance of argon. The hydrogen and helium are probably derived largely from the solar wind, although a portion of the helium may be of radiogenic origin and some hydrogen could result from the photodissociation of H_2O. The interaction of high-energy particles with surface materials may liberate enough oxygen to be its principal source, but breakdown of water vapor molecules by sunlight could also be a possible source.

In 1985–86, Earth-based telescopic observations detected sodium and potassium in the exosphere. Both sodium and potassium have highly variable abundances (10^4–10^5 Na atoms cm^{-3} and 100–10^4 K atoms cm^{-3})

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

near the surface on time scales of hours to years. Their abundances also vary between day and night by a factor of about 5, the day side being greater. The temperature of the gas is about 500 K, but a hotter more extended Na coma sometimes exists.

Although both sodium and potassium are probably derived from the surface of Mercury, the mechanism by which they are supplied is not well understood. They could be released from sodium- and potassium-bearing minerals by their interaction with solar radiation, or by impact vaporization of micrometeoroid material. Both sodium and potassium show day-to-day changes in their global distribution.

Polar deposits

In 1991, high resolution, full disk radar images of Mercury were obtained from both the Arecibo and the linked Goldstone–Very Large Array radar facilities. The radar signals show very high reflectivities and polarization ratios centered on the poles of Mercury. Mariner 10 images of Mercury's polar regions show cratered surfaces where ice could be concentrated in permanently shadowed portions of the craters, and the radar studies show that the anomalies are indeed concentrated in the permanently shadowed portions of these polar craters. Because the obliquity of Mercury is near 0 degrees, it does not experience seasons, and therefore, temperatures in the polar regions should be <135 K. In permanently shaded polar areas the temperatures should be less than 112 K, and water ice should be stable to evaporation on time scales of billions of years. If the deposits are water ice, then they could originate from comet or water-rich asteroid impacts that released the water to be cold-trapped in the permanently shadowed craters. Because comets and asteroids also impact the Moon, similar deposits would be expected to occur in the permanently shadowed regions of lunar craters. Indeed, the Lunar Prospector mission did detect water ice in these regions of the Moon.

Interior and magnetic field

Mercury's internal structure is unique in the solar system. It also imposes severe constraints on any proposed origin of the planet. Mercury's mean density of 5.44 g cm^{-3} is only slightly less than Earth's (5.52 g cm^{-3}), and larger than Venus' (5.25 g cm^{-3}). Because of Earth's large internal pressures, however, its uncompressed density is only 4.4 g cm^{-3} compared to Mercury's uncompressed density of 5.3 g cm^{-3}. This means that Mercury contains a much larger fraction of iron than any other planet or satellite in the solar system. If this iron is concentrated in a core, then the core must be about 75% of the planet diameter, or some 42% of its volume. Thus, its silicate mantle and crust is only about 600 km thick. For comparison, Earth's iron core is only 54% of its diameter, or just 16% of its volume.

Mercury is the only other terrestrial planet, aside from Earth, with a significant magnetic field. Mariner 10 discovered that Mercury has a global intrinsic dipole magnetic field, but it is not well characterized because the spacecraft only made two passes through the field. The magnitude of the dipole moment is about 330 nT, or over 1000 times smaller than Earth's. Although relatively weak, the field has sufficient strength to hold off the solar wind, creating a bow shock, a magnetosheath, and a magnetosphere (see MERCURY: MAGNETOSPHERE). Because of the weaker field, Mercury occupies a much larger fraction of the volume of its magnetosphere than do other planets, and the solar wind actually reaches the surface at times of highest solar activity. The size of the magnetosphere is about 7.5 times smaller than Earth's. Also, because of the small size of Mercury's magnetosphere, magnetic events happen more rapidly and repeat more often than in Earth's magnetosphere.

Although other models may be possible, the maintenance of terrestrial planet magnetic fields is thought to require an electrically conducting fluid outer core surrounding a solid inner core. Therefore, Mercury's dipole magnetic field suggests that Mercury currently has a fluid outer core of unknown thickness. Thermal history models indicate that Mercury's core should have solidified long ago unless there was some way of maintaining high core temperatures throughout geological history. The preferred way is to lower the melting point of the core material by adding a light alloying element. Sulfur is considered to be the most likely candidate. If sulfur is the alloying element, then Mercury probably contains between 0.2 and 7% sulfur in its core in order to have an outer fluid portion at the present time.

Geology and composition

In general, the surface of Mercury can be divided into four major terrains: (1) heavily cratered regions, (2) intercrater plains, (3) smooth plains, and (4) hilly and lineated terrain. Other relatively minor units have been identified, such as ejecta deposits exterior to the Caloris basin and other large impact CRATERS. The heavily cratered uplands record the period of late heavy meteoroid bombardment that ended about 3.8 billion years ago. In the heavily cratered terrain on Mercury there is an increasing paucity of craters with decreasing crater diameter relative to heavily cratered terrain on the Moon. This paucity of craters is probably due to obliteration of the smaller craters by emplacement of intercrater plains during this period of late heavy bombardment. The crater population superimposed on the smooth plains within and surrounding the Caloris basin shows a same size distribution but at a much lower crater density. This suggests that unlike the lunar maria the Caloris smooth plains formed near the end of late heavy bombardment.

Fresh impact craters on Mercury exhibit similar morphologies as those on the other terrestrial planets. Small craters are bowl-shaped, but with increasing size they develop central peaks, flat floors, and terraces on their inner walls. The transition from simple (bowl-shaped) to complex (central peak and terraces) craters is about 10 km. At diameters between about 130 to 310 km mercurian craters have an interior concentric ring, and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

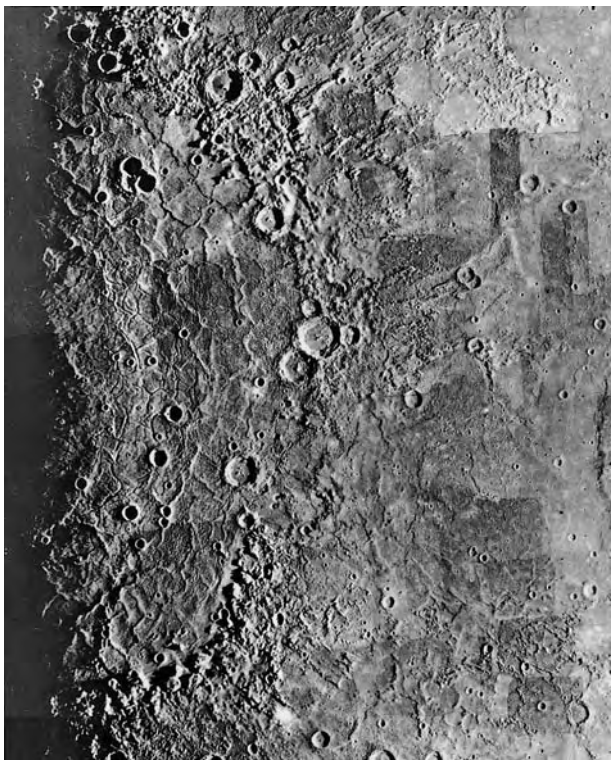


Figure 1. Photomosaic of the 1300 km diameter Caloris impact basin showing the highly ridged and fractured nature of its floor.

at diameters larger than about 300 km they may have multiple inner rings. The freshest craters have extensive ray systems, some of which extend for distances over 1000 km. For a given rim diameter the radial extent of mercurian continuous ejecta is uniformly smaller by a factor of about 0.65 than that for the Moon. Furthermore, the maximum density of secondary impact craters occurs closer to the crater rim than for similar-sized lunar craters. All of these differences are probably due to the larger surface gravity of Mercury (3.70 m s^{-2}) compared to the Moon (1.62 m s^{-2}).

The largest relatively fresh impact feature seen by Mariner 10 is the 1300 km diameter Caloris basin. The floor structure consists of closely spaced ridges and troughs arranged in both a concentric and radial pattern (figure 1). The ridges are probably due to compression, while the troughs are probably tensional graben that post-date the ridges. This pattern may have been caused by subsidence and subsequent uplift of the basin floor.

Directly opposite the Caloris basin (the antipodal point) is the unusual hilly and lineated terrain that disrupts pre-existing landforms, particularly crater rims (figure 2). Stratigraphic relationships suggest that the age of this terrain is the same as the Caloris basin. The hilly and lineated terrain is thought to be the result of seismic waves generated by the Caloris impact and focused at the antipodal region.

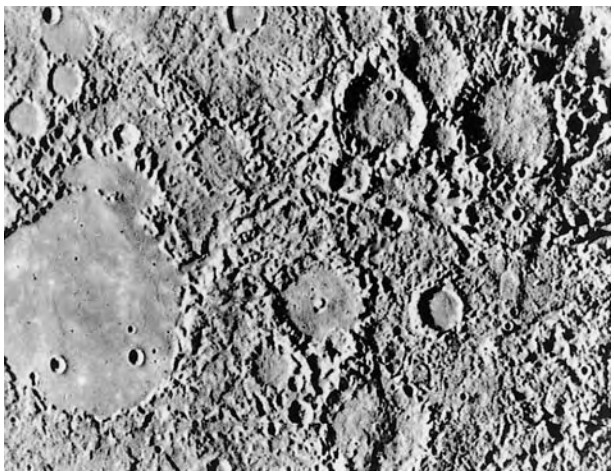


Figure 2. A portion of the hilly and lineated terrain antipodal to the Caloris impact basin. The image is 543 km across.

Mercury's two plains units have been interpreted as lava plains. The older intercrater plains are the most extensive terrain on Mercury (figures 3(a) and 3(b)). They have probably been responsible for obliterating a significant number of craters as evidenced by the paucity of craters less than about 40 km diameter compared to the highlands of the Moon. Intercrater plains were emplaced over a range of ages contemporaneous with the period of late heavy bombardment. They are thought to be volcanic plains erupted through a fractured crust. They are probably about 4.0 to 4.2 billion years old.

The younger smooth plains are primarily associated with large impact basins. They also fill smaller basins and large craters. The largest occurrence of smooth plains fill and surround the Caloris basin, and occupy a large circular area in the north polar region that is probably an old impact basin about 1500 km in diameter. They are similar in morphology and mode of occurrence to the lunar maria, and, therefore, they are interpreted to be lava flows erupted relatively late in mercurian history. They may have an average age of about 3.8 billion years as indicated by crater densities. If so, they are, in general, older than the lunar maria.

Three large radar-bright anomalies have been identified on the unimaged side of Mercury. One of these is similar to the radar signature of a fresh impact crater, and another has a radar signature unlike any other in the solar system. However, the third radar anomaly has a more or less structureless radar-bright halo (500 km diameter) and a radar-dark center (70 km diameter). This is similar to the radar signatures of large shield volcanoes on Venus and Mars. If this radar feature is indeed a shield volcano, then it is as large as Mons Olympus on Mars. Furthermore, the fact that it is bright means it is relatively fresh, otherwise it would not give a radar signature. If this feature is, indeed, a relatively fresh shield volcano, then the history of Mercury is probably radically different than previously believed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

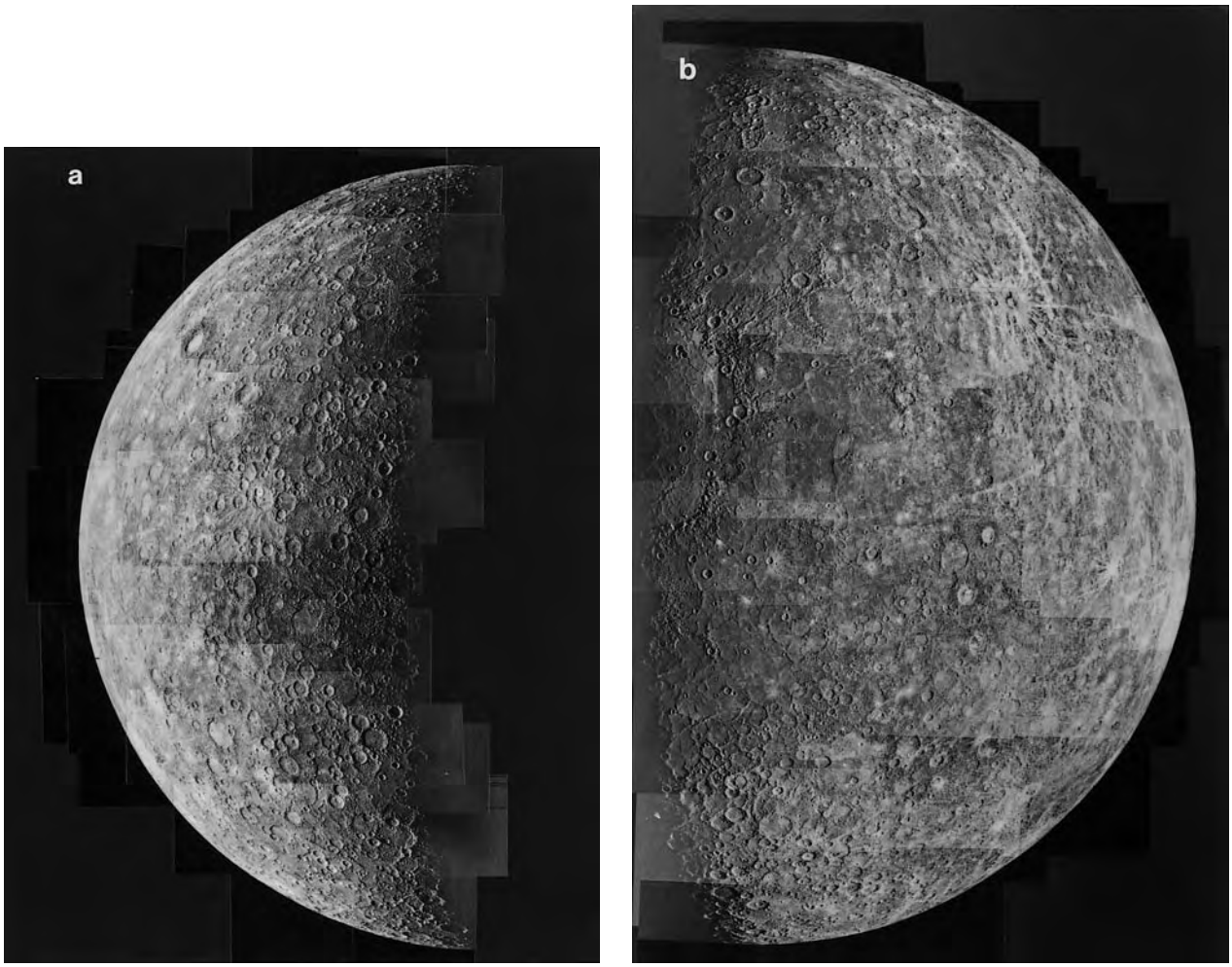


Figure 3. (a) Mercury as viewed by Mariner 10 on its first approach in March, 1974. (b) Mercury's opposite hemisphere viewed by Mariner 10 as it left the planet on the first encounter.

Mercury's tectonic framework consists of a system of compressive thrust faults called lobate scarps. Individual scarps vary in length from about 20 km to over 500 km, and have heights from a few 100 m to about 3 km. They have a random spatial and azimuthal distribution over the imaged half of the planet and presumably occur on a global scale. This suggests that Mercury was subjected to global compressive stresses. The only occurrences of features indicative of tensile stresses are localized fractures associated with the floor of the Caloris basin and at its antipode, both of which are the direct or indirect result of the Caloris impact. Stratigraphic evidence indicates that the system of thrust faults post-dates the formation of intercrater plains and formed relatively late in mercurian history. This tectonic framework was probably caused by crustal shortening resulting from a decrease in the planet radius due to cooling of the planet. The amount of radius decrease is estimated to have been about 2 km.

Very little is known about the surface composition of Mercury. Recalibration and color ratioing of Mariner 10 images have been used to derive the FeO abundance, the opaque mineral content, and the soil maturity over the region viewed by Mariner 10. The probably volcanic smooth plains have a FeO content of <6% by weight, which is similar to the rest of the planet imaged by Mariner 10. The surface of Mercury, therefore, may have a more homogeneous distribution of elements affecting color (e.g. more alkali plagioclase) than does the Moon. At least the smooth plains may be low iron or alkali basalts. There are some dark blue, low albedo, and high opaque mineral regions with diffuse boundaries that may be associated with fractures. These areas could be more mafic volcanic pyroclastic deposits. The bright rayed craters on Mercury have a very low opaque mineral index that may indicate the craters have excavated into an anothositic crust similar to that on the Moon. Earth-based spectroscopic measurements suggest that Mercury's surface is low in titanium and iron compared to the surface of the Moon. This is consistent with Mercury's lower ultraviolet reflectivity and smaller albedo contrast.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This has been interpreted to mean that Mercury's surface is largely devoid of basalt, but it could also mean that the basalts only have a low iron content or are fluid alkali basalts.

Geological history

Mercury's earliest history is very uncertain. The earliest events are the formation of intercrater plains ($\geq$4 billion years ago) during the period of late heavy bombardment. These plains may have been erupted through fractures caused by large impacts in a thin lithosphere. Near the end of late heavy bombardment the Caloris basin was formed by a large impact that caused the hilly and lineated terrain from seismic waves focused at the antipodal region. Further eruption of lava within and surrounding the Caloris and other large basins formed the smooth plains about 3.8 billion years ago. The system of thrust faults formed after the intercrater plains, and resulted in a radius decrease of about 2 km. The volcanic history is very uncertain. If the radar anomaly on the unseen side is truly a relatively fresh shield volcano then the unimaged portion of Mercury may have a totally different volcanic history from the imaged side. We will have to await the results of the MESSENGER mission to fully evaluate the history of Mercury.

Origin

The origin of Mercury and how it acquired such a large fraction of iron compared to the other terrestrial planets is not well determined. Three hypotheses have been put forward to explain the enormous iron core. One involves an enrichment of iron due to mechanical and dynamical accretion processes in the inner-most part of the solar system. Another proposes that intense bombardment by solar electromagnetic and corpuscular radiation in the earliest phases of the Sun's evolution vaporized and drove off much of the silicate fraction of Mercury leaving the core intact. A third proposes a planet-sized object impacted Mercury and essentially blasted away much of the planet's silicate mantle leaving the core largely intact. Discriminating between these hypotheses may be possible from the chemical make-up of the surface because each one predicts a different composition. MESSENGER is designed to measure the composition of Mercury's surface so it may be possible to answer this vital question in the near future.

Robert G Strom

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mercury: Magnetosphere

Among the major discoveries made by the MARINER 10 mission to the inner planets was the existence of an intrinsic magnetic field at Mercury with a dipole moment of $\sim$ 300 nT R_M^3. This magnetic field is sufficient to stand off the solar wind at an altitude of about 1 R_M (i.e. $\sim$2439 km). Hence, Mercury possesses a 'magnetosphere' from which the solar wind plasma is largely excluded and within which the motion of charged particles is controlled by the planetary magnetic field. Despite its small size relative to the magnetospheres of the other planets (see PLANETARY MAGNETOSPHERES), a Mercury orbiter mission is a high priority for the space physics community. The primary reason for this great interest is that Mercury, unlike all the other planets visited thus far, lacks a significant atmosphere; only a vestigial exosphere is present (MERCURY). This results in a unique situation where the magnetosphere interacts directly with the outer layer of the planetary crust (i.e. the regolith). At all of the other planets the topmost regions of their atmospheres become ionized by solar radiation to form ionospheres. These planetary ionospheres then couple to electrodynamically to their magnetospheres or, in the case of the weakly magnetized Venus and Mars, directly to the solar wind. This magnetosphere–ionosphere coupling is mediated largely through field-aligned currents (FACs) flowing along the magnetic field lines linking the magnetosphere and the high-latitude ionosphere. Mercury is unique in that it is expected that FACs will be very short lived due to the low electrical conductivity of the regolith.

Furthermore, at the Earth it has been shown that the outflow of neutral atmospheric species to great altitudes is an important source of magnetospheric plasma (following ionization) whose composition may influence subsequent magnetotail dynamics. However, the dominant source of plasma for most of the terrestrial magnetosphere is the 'leakage' of solar wind across the magnetopause and more direct entry through the northern and southern cusps. Although Mariner 10 did not return plasma composition measurements, the Hermean magnetosphere should be ideal for measuring the manner and rate of solar wind plasma entry due to the lack of strong internal atmospheric sources. Finally, the solar wind conditions experienced by Mercury as it orbits the Sun at 0.31 to 0.47 AU are quite different from those typically encountered by the Earth. This may allow for new understanding of the external factors affecting the transfer of mass, momentum and energy from the solar wind to planetary magnetospheres. This article provides a brief overview of what is now known about Mercury's magnetosphere and why it is a priority target for future planetary missions.

Mariner 10 encounters

Following its launch on 2 November 1973, the Mariner 10 spacecraft executed three close encounters with the planet Mercury on 29 March 1974, 21 September 1974, and 16 March 1975. The first encounter was targeted to pass through the wake of the planet in order to determine its interaction with the solar wind. The closest approach to the surface of the planet occurred over the post-midnight equatorial region at an altitude of about 700 km. Figure 1 displays 1.2 s averages of the magnetic fields observed during the first Mariner 10 encounter in Mercury-centered solar ecliptic coordinates. The intensity of the magnetic field observed by Mariner 10 peaked near closest approach at just under 100 nT. This maximum is due to the increased contribution from the planetary magnetic field which varies as the inverse third power of the distance from the planet. The second encounter passed well upstream of the planet to return dayside images and did not allow observations of the magnetosphere. The third encounter was intended to achieve higher latitudes and lower altitudes in order to confirm the intrinsic nature of the magnetic field. This effort was successful with a closest approach altitude of only 327 km near local midnight at $\sim$68° N latitude and a peak measured magnetic field intensity of $\sim$400 nT.

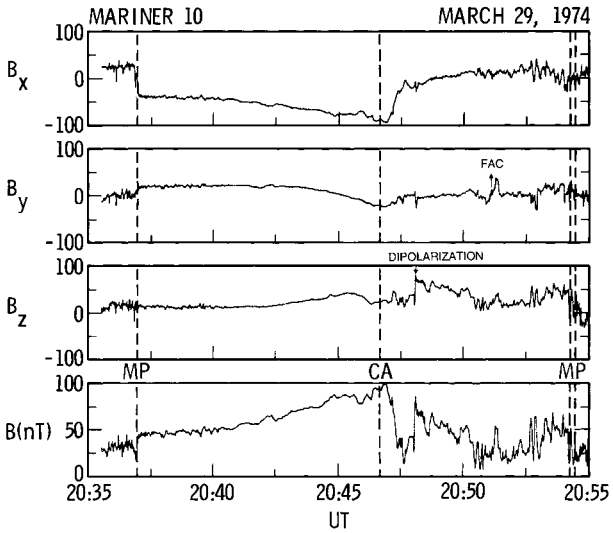


Figure 1. The Mariner 10 magnetic field observations (1.2 s averages) taken during the first encounter are displayed in Mercury centered solar ecliptic coordinates. In this system the X_{ME} axis is directed toward the Sun and the Y_{ME} axis is in the plane of the ecliptic, perpendicular to X_{ME} and directed oppositely from the direction of planetary orbital motion. The Z_{ME} axis completes the right-handed orthogonal system and is positive toward the north. Magnetopause, dipolarization and field-aligned current signatures in the observations are labeled. (From Slavin *et al* (1997).)

Magnetospheric structure

Both of Mariner 10's magnetospheric traversals began and ended with the magnetic field and plasma investigations observing clear magnetopause boundaries (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). The magnetopause is the current layer formed by the interaction between the solar wind plasma and a planetary magnetic field. The magnetopause current layer confines most of the magnetic flux

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

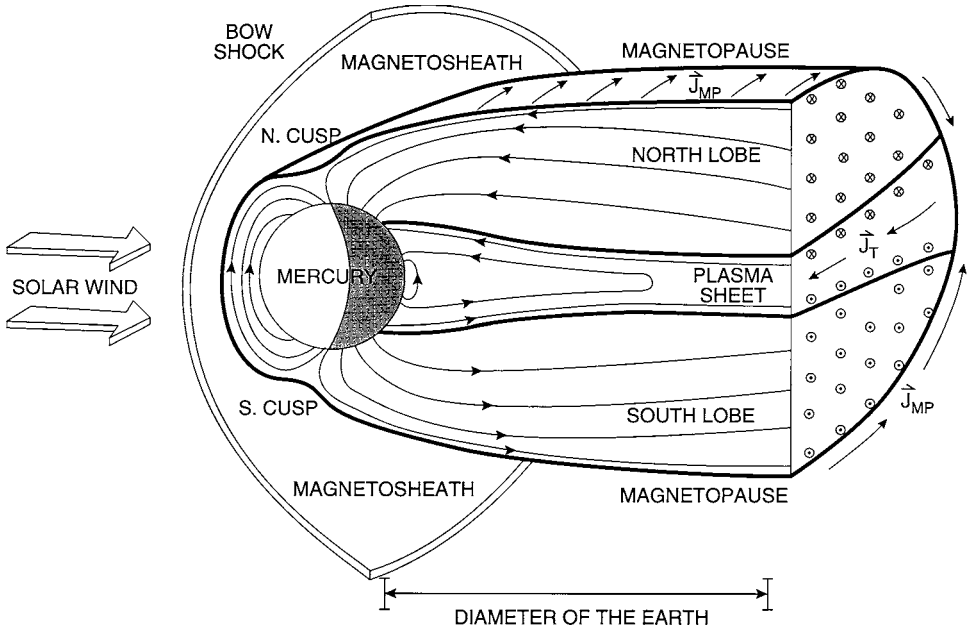


Figure 2. A schematic view of Mercury's magnetosphere as glimpsed by Mariner 10 with its primary regions, some representative magnetic field lines, and the tail current system labelled. Note that the regions typically occupied by radiation belts and the plasmasphere at the Earth are taken up by the planet in this miniature magnetosphere.

tubes rooted in the planet to the magnetospheric cavity. Vertical broken lines in figure 1 mark the inbound and outbound magnetopause crossings as well as the point of closest approach (CA) to the planet. Across the magnetopause current layer the magnetic field rotates in direction to its magnetospheric orientation and increases in magnitude until it is in equilibrium with the external solar wind pressure. The entry of Mariner 10 into the near-tail (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL) of this magnetosphere is apparent in the tail-like nature of the magnetic field with $B_x \gg B_y$, B_z just inside the magnetopause during the first inbound passage.

Analysis of the Hermean magnetopause using the data sets collected by Mariner 10 has revealed some differences from what is usually observed at Earth. For example, the higher solar wind plasma density at 0.3–0.5 AU produces solar wind ram pressure that are about four to ten times greater than at the Earth. This results in higher magnetopause current densities and stronger magnetic fields inside the magnetopause at Mercury than are observed elsewhere in the solar system. The low altitude of the dayside magnetopause at Mercury has provoked much interest as to whether this boundary ever intersects the planetary surface and exposes it to direct impingement by the solar wind. Simple compression of the magnetopause by solar pressure enhancements would probably cause this to happen if it were not for the induction currents which are expected to be driven deep within the planet to oppose such changes in the magnetic field. However, this induction effect will, at times, be countered by the 'erosion' of magnetic flux tubes by reconnection (see MAGNETIC RECONNECTION) at the dayside magnetopause. The net result of dayside reconnection is to remove flux tubes from the dayside magnetosphere, where they participate in standing-off the solar wind, and transport them into the magnetotail. In addition, it has been suggested that the reconnection rate at Mercury may be enhanced by the low Alfvén Mach numbers in the solar wind at 0.3 to 0.5 AU. Ultimately, the balance between these competing effects, induction and erosion, will rest in the actual nature of the electrical conductivity profiles in the planetary interior and the efficiency of the magnetic reconnection at the magnetopause.

During the outbound portion of the first Mariner 10 encounter the magnetic field variations were extremely dynamic. These large-amplitude variations in the field intensity and direction have a number of sources. The large dip in magnetic field strength just after closest approach coincided with the spacecraft becoming immersed in a hot plasma sheet (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). These observations are consistent with the spacecraft abruptly transitioning from the magnetic field pressure-dominated lobes to the plasma pressure-dominated plasma sheet. In addition, there are large-amplitude variations in the B_y field component in figure 1 around 20:51 UT. This magnetic field signature is well known to magnetospheric scientists and can be readily interpreted as being due to the spacecraft traversing several quasi-planar field-aligned current sheets. Given the Mariner 10 trajectory, the main gradient in B_y from negative to positive is indicative of an upward FAC being traversed by the spacecraft. This

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

upward current sheet is largely balanced by two less intense downward current sheets seen just before and after the central upward current sheet. At Earth the occurrence of such multiple current sheets is particularly common over the nightside auroral oval during geomagnetically disturbed intervals. These currents are the primary means by which the magnetosphere transmits energy and momentum to the ionosphere, or in the case of Mercury, the regolith. Given the expectation of a highly electrically resistive regolith at Mercury, field-aligned currents will probably be very short lived and hence, only transient features of the Hermean magnetosphere.

Based on the Mariner 10 measurements, and terrestrial-style extrapolation, figure 2 depicts our present image of Mercury's magnetosphere; included are the primary regions, sample magnetic field lines, and the tail current system. Concerning the physical dimensions of this small magnetosphere, it was noted by the original flight investigators that a scaling in which $1R_M$ at Mercury corresponds to $\sim 6R_E$ at Earth produces good correspondence between the boundaries and regions at these two planets. Hence, the $\sim 2R_M$ distance from the center of Mercury to the subsolar magnetopause maps to about $12R_E$ at the Earth which is, indeed, similar to the average terrestrial magnetopause nose distance of $10–11R_E$. More importantly, the surface of Mercury would correspond roughly to geosynchronous distance within Earth's magnetosphere. This scaling rules out the possibility of a plasmasphere even if the planet possessed the atmospheric plasma source and the rotation rate to create such a region. (Mercury's rotational and orbital periods are 59 and 88 days, respectively.) Trapped radiation belts such as are found in the terrestrial magnetosphere at distances of $2–4R_E$, are also impossible. Finally, the observed near-tail diameter of $4–5R_M$ scales to about $24–30R_E$ which is somewhat less than the typical near-tail diameter at Earth of $30–40R_E$. The reason for this discrepancy in this Earth-based scaling is the higher ratio of static to dynamic pressure in the solar wind at Mercury's orbit. The higher static solar wind pressures reduce the diameter of the tail relative to the subsolar magnetopause radius which is determined largely by the dynamic or 'ram' pressure.

Magnetospheric substorms

A *substorm* is a magnetosphere-wide disturbance which involves the conversion of large amounts of electromagnetic energy, either drawn directly from the solar wind or first stored in magnetic fields of the tail lobes, into plasma sheet heating, high-speed bulk plasma flows, energetic particle acceleration, and field-aligned currents which transfer energy to the planetary ionosphere, or at Mercury, to the regolith (see MAGNETOSPHERE OF EARTH: SUBSTORMS). The energy conversion mechanism is magnetic reconnection between the oppositely directed magnetic fields of the tail lobes in much the same manner as is thought to occur in solar flares. The understanding of the reconnection process and how it is influenced by both local plasma conditions in the tail and electrodynamic coupling to the rest of the magnetosphere and the ionosphere, or regolith at Mercury, is one of the overarching objectives of magnetospheric physics research.

Less than 1 min after Mariner 10 entered the plasma sheet during the first encounter, i.e. beginning around 20:48 UT, it detected a sharp increase in the B_z field component. As displayed in figure 3, the initial sudden B_z increase and subsequent quasi-periodic increases are all coincident with strong enhancements in the flux of >35 keV electrons observed by the Mariner 10 cosmic ray telescopes. This energetic particle signature, and several weaker events observed later in the same pass, have been widely interpreted as evidence for substorm activity. Analyses of the energetic particle and magnetic field signatures for this interval have led to the conclusion that this Mariner 10 event is qualitatively identical to the substorm energetic particle injections and magnetic field dipolarizations observed at altitudes between geosynchronous orbit and the inner edge of the plasma sheet at the Earth. However, there are issues surrounding how a magnetosphere with such small dimensions compared with the relevant charged particle Larmor radii can be such a copious source of energetic particles. More complete energetic particle measurements to be returned by future missions to this planet may yield important clues to how the most energetic charged particles are accelerated in the tails of planetary magnetospheres.

The very short duration of the substorm signatures at Mercury, about 1 min as compared with 1–2 h at the Earth, is also of fundamental interest. Although much remains to be understood concerning substorm processes, it is clear that they correspond to intervals of greatly enhanced magnetospheric convection which leads to the dissipation of large amounts of magnetic energy drawn from the tail lobes. It has been argued that the temporal scale of isolated substorms at the Earth is determined by the time necessary for plasma to convect across the polar cap or, equivalently, from the magnetopause down to the plasma sheet where reconnection can take place. It has been estimated, using typical solar wind and magnetotail parameters, that the time necessary to 'cycle' all of the flux in the tail lobes at the Earth is, indeed, close to the typical duration of a substorm, i.e. ~1–2 h.

For Mercury, corresponding convection time constants of a few minutes have been calculated in agreement with the Mariner 10 measurements. The brevity of the convection time constant at Mercury is due largely to the small tail radius and relatively intense $-\boldsymbol{V} \times \boldsymbol{B}$ electric fields present in the solar wind at 0.3 to 0.5 AU. Furthermore, it must be noted that at Earth the dawn-to-dusk magnetospheric electric field is typically only 10–20 per cent of the total electric potential drop available from the solar wind. However, for Mercury it has been argued that the electric potential drop across the magnetosphere should more closely approach the maximum available potential drop

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

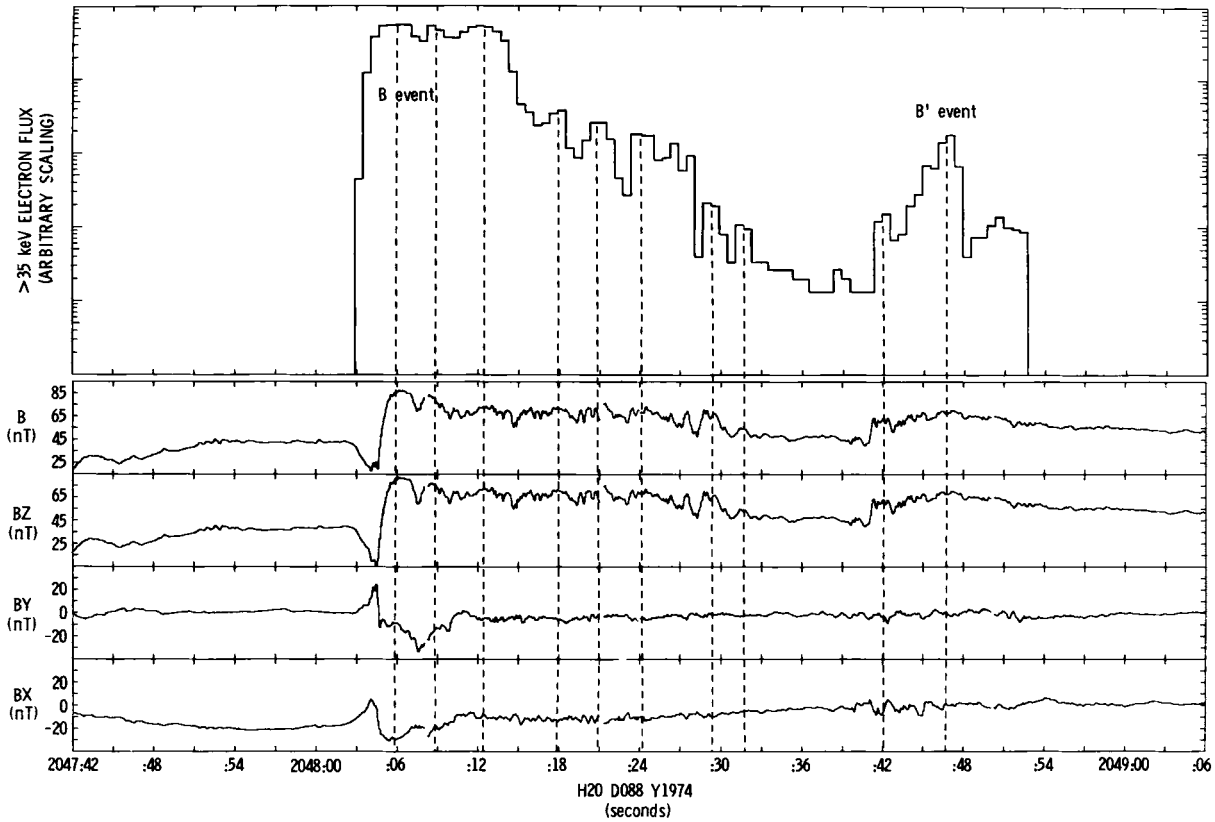


Figure 3. Highest sample rate measurements of the Mariner 10 energetic electron (>35 keV) flux (0.6 s resolution) and magnetic field (0.04 s resolution) are displayed for the substorm dipolarization event observed during the first encounter. (From Christon *et al* (1987).)

than is possible at other planets. The reason is that conducting ionospheres at the other planets act to 'short-out' a portion of this electric field that the solar wind is attempting to impose across the magnetosphere and, therefore, limit the intensity of the internal magnetospheric electric field. Hence, a determination of the factors responsible for the short substorm time scale at Mercury may be very important for our general understanding of electric fields and convection in all planetary magnetospheres.

Summary

From the preceding discussion, it can be seen that Mercury's miniature magnetosphere offers a unique testing ground for theories regarding many important processes. In particular, the very tenuous nature of its atmosphere implies that internal plasma sources are far less significant than at Earth or any of the other known magnetospheres. This fact should make Mercury ideal for quantitative modeling of how and to what degree solar wind plasma enters terrestrial-type magnetospheres. Mercury's relative proximity to the Sun also exposes this magnetosphere to solar wind conditions which are relatively rare at 1 AU. In particular, it has been suggested that the lower Alfvénic Mach numbers and stronger interplanetary magnetic fields at 0.3 to 0.5 AU will result in higher dayside reconnection rates and more intense large-scale magnetospheric electric fields than at the Earth. Hence, measurements at Mercury could also be pivotal to the testing of existing models of reconnection and electrodynamic coupling between the various regions of the magnetosphere. Observations from a future Mercury orbiter mission should also lead to great improvement in our understanding and modeling capability with respect to magnetospheric substorms. With durations at Mercury of only about 1 min, it should be possible to observe huge numbers of complete substorms at a variety of different locations throughout the magnetosphere in a relatively short time. Such a data set would be especially beneficial in the testing of substorm theories with respect to the role of electrodynamic 'feed-back' between the region at the foot of the magnetospheric flux tubes and convection in the magnetotail. Indeed, this small magnetosphere may offer our only opportunity to observe electrodynamic coupling in the asymptotic limit of a magnetosphere rooted in a relatively non-conducting planetary regolith.

Bibliography

Baker D N, Simpson J A and Eraker J H 1986 A model of impulsive acceleration and transport of energetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

particles in Mercury's magnetosphere *J. Geophys. Res.* **91** 742

Christon S P, Feynman J and Slavin J A 1987 Dynamic substorm injections: similar magnetospheric phenomena at Earth and Mercury *Magnetotail Physics* ed A T Y Lui (Baltimore, MD: Johns Hopkins University Press) pp 393–400

Connerney J E P and Ness N F 1988 The magnetosphere of Mercury *Mercury* ed C Chapman (Tucson, AZ: University of Arizona Press) pp 494–513

Russell C T, Baker D N and Slavin J A 1988 The magnetosphere of Mercury *Mercury* ed C Chapman (Tucson, AZ: University of Arizona Press) pp 514–61

Slavin J A, Owen C J, Connerney J E P and Christon S P 1997 Mariner 10 observations of field-aligned currents at Mercury *Planet. Space Sci.* **45** 133

James A Slavin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meridian

A great circle on the Earth's surface, running perpendicular to the equator and passing through the north and south poles. To an observer located at a particular position on the Earth, the meridian represents his north–south line. The Greenwich meridian (i.e. the meridian passing through the Old Greenwich Observatory) is taken as the zero of longitude measurement; thus the longitude of a point on the Earth's surface is the angle (measured parallel to the equator) between the Greenwich meridian and the meridian passing through that point.

For purposes of astronomical position and time measurement, the meridian is taken to be a circle on the celestial sphere passing through the north and south celestial poles and the observer's zenith; i.e. for an observer on the Earth's surface, the celestial meridian is the projection of his terrestrial meridian onto the celestial sphere.

The term 'central meridian' refers to the line joining the north and south poles of a planet that crosses the center of the visible disk of that planet. (The Greenwich meridian is also known as the 'prime meridian'.)

See also: great circle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MERLIN/VLBI National Facility

The Multi-Element Radio Linked Interferometer Network (MERLIN) is a unique synthesis radio telescope which was originally developed by the University of Manchester in the late 1970s and came into operation in 1980. After further development, in 1993 it formally became a National Facility operated by the University of Manchester on behalf of the UK PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL. MERLIN operates in discrete wavelength bands within the range 2 m to 1 cm and has corresponding angular resolutions ranging from 1.4 arcsec to 8 milliarcsec. At its prime operating wavelength of 6 cm, it has a resolution which matches that of the Hubble Space Telescope. In addition to its normal self-contained operations, it is regularly used in conjunction with the European Very Long Baseline Interferometry Network (EVN) to provide angular resolution up to a few hundred microarcsec. MERLIN is used for the study of a wide range of astronomical subjects, including stellar formation, interstellar material, nebulae, active stars, supernovae, external galaxies, and quasars and active galaxies at the extremities of the known universe. It was the first instrument to resolve the outflow of material from novae within the Galaxy and to resolve the shells of supernovae in external galaxies. It has played a key role in the discovery of approximately twenty gravitational lenses and in the establishment of the coordinate frame of the Hubble Deep Field.

For further information see http://www.jb.man.ac.uk/merlin/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Merrill, Paul Willard (1887–1961)

Astronomer, staff-member of the Mount Wilson Observatory, spectroscopist who studied variable stars. He identified the element technetium in an S-type star R Andromedae. The atoms of this element have no stable nuclear isotopes, so the element must have been made within the last million years, although the star was much older. The discovery proved that the element must have been made recently in the star. The process by which it was made was identified by B2FH (see W FOWLER) as the 's-process' of nucleosynthesis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mersenne, Marin (1588–1648)

Born in Oize in Maine, France, joined the order of the Minims. The name signifies that the Minims regard themselves as the least (*minimi*) of all the religious orders, and this encapsulates Mersenne's self-effacing status in science. His own scientific work was modest, but he corresponded with numerous other eminent scientists and acted as a bulletin board among scientists and a scientific commentator at a time when there were no scientific journals. He defended DESCARTES and GALILEO against criticism. He proposed the use of the pendulum as a timing device to HUYGENS, inspiring him to invent the pendulum clock. He published *Les Méchanique de Galilée*, to promulgate Galileo's lectures on mechanics. He translated parts of Galileo's *Dialogo* and his *Discorsi* into French disseminating Galileo's work in Europe outside Italy. After his death letters in his cell were found from nearly a hundred correspondents including Fermat, Huygens, Pell, Galileo and Torricelli.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meshchersky, Ivan Vsevolodovich (1859–1935)

Born in Arkhangelsk, Russia, studied mechanics, in particular the motion of bodies of variable mass, applying this to the doubtful case of comets. But his work formed the basis to calculate the case of rocket flight, developed after 1945. He was the first to study the inverse problem of determining the loss of mass from knowledge of the orbit and the forces.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mesons

Particles that are composed of a quark and an antiquark. Mesons are short-lived particles that decay rapidly into leptons or photons either directly, or by first decaying into other types of meson. Their masses range from about 15% of the mass of a proton, in the case of the pion (pi-meson), to more than ten times the mass of a proton. Their mean lifetimes range from 2.6×10^{-8} s, in the case of the positively charged and negatively charged pions (neutral pions have lifetimes of about 10^{-16} s) to less than 10^{-20} s, in the case of the heaviest mesons.

Although the first meson (the pion) was not discovered experimentally until 1947, the existence of a particle of this kind had been proposed in 1935 by the Japanese physicist, Hideki Yukawa (1907–81), in order to explain how the strong nuclear force could be conveyed between protons and neutrons in an atomic nucleus. At the scale of an atomic nucleus, the pion may be regarded as the 'force-carrying' particle of the strong nuclear force. However, because the pion is composed of quarks, it is not a genuinely elementary (or 'fundamental') particle. At the fundamental level, the strong force is conveyed between quarks by particles called gluons.

See also: antimatter, atom, cosmic rays, elementary particles, fundamental forces, fundamental particles, gauge bosons, hadrons, leptons, photon, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MESSENGER (Mercury Surface, Space Environment, Geochemistry and Ranging Mission)

NASA Discovery mission, scheduled for launch in 2004. Intended to fly past Mercury twice in 2008, then enter orbit around the planet in September 2009. Planned to map the entire planet and study the magnetosphere and composition of its thin atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Messier Catalog

In the eighteenth century the French astronomer, Charles Messier (1730–1817), drew up a catalog of 109 of the brighter nebulae, clusters and galaxies. Objects in this catalog are denoted by the letter M followed by a number, for example, M31 is the Andromeda Galaxy. Messier's prime interest was comets. His purpose was to make comet hunting easier by tabulating permanent deep-sky objects that could be mistaken for comets. He published a preliminary list of 45 such objects in 1771 and compiled the bulk of his catalog 10 years later. The catalog is still a valuable guide to amateur astronomers, although it has been superseded by the New General Catalog (NGC) and modern astro-physically based catalogs.

See also: Fundamental catalogs, Index Catalog, M1 to M110, New General Catalog.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Messier, Charles (1730–1817)

Comet seeker, nebula avoider, born in Badonvillier, Lorraine, France, became interested in astronomy as a child. Went to Paris, employed by JOSEPH-NICOLAS DELISLE to search for Comet Halley on the 1757 return. While doing so, he discovered a comet-like patch in Taurus on the comet's expected path. The patch did not move, and he later identified it as a nebula (now known as the Crab Nebula), earlier found by JOHN BEVIS. Messier decided to make a list of nebulous objects that simulate comets, with Bevis's nebula the first entry, M1. While discovering two comets in 1763–4, he recorded two more nebulae, M2 (previously discovered by Maraldi) and M3 (the unresolved globular cluster, his first original nebular discovery). He undertook systematic searches for comets and nebulae, and searched the literature. This eventually brought his catalog to 110 (some entries are spurious), of which 42 objects were discovered by himself. After 1781, the year of Herschel's discovery of a new planet (which Messier himself observed), his strategy seems to have changed, adding few entries to his catalog although maintaining his comet search. Perhaps he knew of Herschel's survey (see HERSCHEL FAMILY), with superior instrumentation. Messier lost his position in the 'Year of Terror' in 1793–4, but resumed observing afterwards, bringing his total score of comets to 19, 13 being original and 6 being independent co-discoveries. In his old age, Messier accepted the Legion of Honor from Napoleon, but disgraced himself in front of his colleagues by publishing an astrological memoir identifying the great comet of 1769 with Napoleon, born that year. This was, as Admiral Smyth put it, 'The last comet put astrologically before the public by an orthodox astronomer'. The *Messier Catalog* (with the letter 'M' identifying entries in the list) stands as the fitting memorial to a lifetime's astronomical work, its nomenclature used daily by astronomers throughout the world.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteor Crater

A terrestrial impact crater 55 km east of Flagstaff, in the Arizona desert, discovered in 1891. Its several alternative names include Barringer Crater, after Daniel M Barringer, who in the 1930s was the first to investigate it and the first to suggest its extraterrestrial origin. Meteor Crater has an average diameter of 1.2 km, the floor is 180 m below ground level and the rim is 30–45 m high. It is estimated to have been produced 50 000 years ago by the impact of an iron meteorite about 40 m across with a mass in the region of 250 000 tonnes. The energy of the blast produced by the impact, which excavated 175 000 tonnes of limestone, would have been 20 megatonnes or more. Thousands of meteoritic nickel–iron fragments have been found at distances up to 7 km from the crater, suggesting that the incoming body fragmented shortly before impact. Excavations have revealed no masses of nickel–iron below the crater floor, and the main crater-forming mass of the asteroid is assumed to have vaporized.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteor Shower

The appearance in the sky, at a particular time of the year, of higher-than-average numbers of meteors, all appearing to emanate from a specific point on the celestial sphere known as the *radiant*. Most showers are named after the constellation in which the radiant lies, by adding the suffix '-id': the Lyrid shower, or Lyrids, for example, have their radiant in the constellation Lyra. The Quandrantid shower is named after the obsolete constellation Quadrans Muralis, now part of Boötes. Where there is more than one radiant in a constellation the showers are distinguished by adding the Bayer letter of a bright star near the radiant, as for example with the Delta Aquarids and Eta Aquarids. A few showers are named after their *parent comets*—the comets that produced the meteor streams—for example the Giacobinids, after Comet 21P/Giacobini–Zinner.

A shower occurs when the Earth intersects a meteor stream. The activity of a shower, which is calculated as a quantity called the *zenithal hourly rate* (ZHR), lasts for as long as the Earth is passing through the stream, and peaks where the stream is at its most concentrated. Young, tight meteor streams tend to produce showers of short duration which build quickly to a high peak ZHR and then tail off sharply. Older, more evolved streams which have spread either side of the primary orbit are of longer overall duration, building slowly to a more modest ZHR and fading slowly thereafter.

The 10 or so most active showers, with a ZHR of about 10 or more, are known as *major showers*. There are dozens of less active, *minor showers*. table 1 gives details for the most active major showers (where there is more than one radiant, the position given is that of the main one). In the great majority of cases the parent body of a meteor shower is a comet. A rare exception is the Geminid shower, which is associated with the asteroid (3200) Phaethon. The connection between a meteor stream and its parent body is established by correspondences between the orbital elements of each.

Meteors that belong to showers are called *shower meteors*, as distinct from *sporadic meteors*, which are produced by meteoroids that do not come from identifiable meteor streams. *Annual showers* are those whose ZHRs normally vary little from one year to the next, being produced by meteor streams in which the meteoroids are evenly spread around the orbit. Meteor streams in which the meteoroids are concentrated in a swarm in one part of the orbit give rise to *periodic showers*, which are highly active only in years when the Earth intersects the swarm, and show little or no activity in other years. Details of two periodic showers, the Leonids and the Giacobinids, are given in table 1.

See also: meteoroid, meteor storm, meteor stream, radiant, sporadic meteor, zenithal hourly rate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Principal annual and periodic meteor showers.

Shower	Position of radiant at maximum		Duration	Date of maximum	Max. ZHR	Parent body
	RA	Dec.				
Quadrantids	$15^h\ 28^m$	50°	1–6 Jan	3–4 Jan	120	96P/Machholz*
Lyrids	$18^h\ 00^m$	32°	19–25 Apr	21 Apr	15	C/1861 G1 Thatcher
Eta Aquarids	$22^h\ 20^m$	-1°	24 Apr–20 May	5 May	50	1P/Halley
Delta Aquarids	$22^h\ 36^m$	-17°	15 Jul–20 Aug	29 Jul	25	Unknown
Perseids	$03^h\ 04^m$	58°	25 Jul–20 Aug	12–13 Aug	80	109P/Swift–Tuttle
Giacobinids	$17^h\ 23^m$	57°	7–10 Oct	8 Oct	**	21P/Giacobini–Zinner
Orionids	$06^h\ 24^m$	15°	15 Oct–2 Nov	20–22 Oct	30	1P/Halley
Taurids	$03^h\ 44^m$	14°	15 Oct–25 Nov	3 Nov	10	2P/Encke
Leonids	$10^h\ 08^m$	22°	15–20 Nov	17 Nov	***	55P/Tempel–Tuttle
Geminids	$07^h\ 26^m$	32°	7–15 Dec	13 Dec	120	(3200) Phaethon

* The association with this comet is uncertain.
** Outbursts possible every 12 or 13 years. Up to 450 per hour in 1933.
*** Possibility of storm activity every 33 years. May have reached 200 000 per hour in 1833.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteor Storm

An intense burst of meteor activity of short duration (typically about an hour) associated with a periodic meteor shower. A meteor storm occurs when the Earth passes through a meteor swarm or a concentrated ribbon of debris recently shed by a comet. There is no generally agreed quantitative definition of what constitutes storm-level activity, but an hourly rate of 1000 is a reasonable lower limit. Meteor storms are rare events, those produced by the Leonid shower being the best known. The Leonid maximum of 16/17 November 1966 produced the highest documented storm levels of modern times, with an estimated hourly rate of 200 000 over a 40 min period. Some annual meteor showers occasionally show sharp rises above their normal hourly rates; such outbursts are sometimes called *substorms*.

See also: Leonids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteor Stream

A continuous ribbon of meteoroids in a closed orbit around the Sun. (The more correct term *meteoroid stream* is seldom used.) A meteor stream consists of dust particles shed by a comet on its perihelion passages which have gradually spread around the comet's orbit. When the Earth passes through a meteor stream, there is a meteor shower when up to about 100 meteors per hour may be seen. Most of the debris from young comets remains bunched together over a small arc of the orbit in what is known as a *meteor swarm*. When the Earth passes through a swarm the result is a meteor storm, with many thousands of meteors per hour. As a meteor stream ages, and the meteoroids spread round the orbit, the stream assumes the shape of an elliptical toroid. Within the stream there may be filaments containing a greater concentration of meteoroids shed by the parent comet on its most recent circuits of the Sun. The evolution of individual filaments and the stream as a whole may be affected by planetary perturbations, of both the meteoroids and the parent comet, and the structure of a meteor stream can grow increasingly complex. Eventually, under repeated perturbations, a meteor stream will disperse to the point where the concentration of meteoroids is reduced to that of the interplanetary dust.

See also: interplanetary dust, meteor shower, meteor storm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteorites

A meteorite is a natural solid object of extraterrestrial origin, fallen onto the surface of the Earth. Whatever its provenance, it has to traverse the atmosphere of our planet at high speed, which gives rise to a brief luminous phenomenon, called a METEOR. This modifies its shape, size and external aspect, but leaves the interior part intact. Hence the scientific interest of meteorites: they give us the opportunity to study worlds that we have not reached yet. Twenty-six meteorites are widely believed to be samples of the planet MARS. Around twenty-one others come from the MOON, from regions that have not been sampled by the Apollo nor the Luna missions, including probably the far side. The overwhelming majority of the meteorites are fragments of ASTEROIDS, the tiny planets that orbit the Sun between Mars and Jupiter. This has allowed us to study the geology and past evolution of these very diverse small bodies for many decades, while their exploration by space probes is only at its very beginning, and no sample from any of them will be brought back before 2007. But the interest in meteorites is more general: because of their small sizes, many of their parent asteroids did not experience the geological processes which have reprocessed all the materials of the large planets like Earth, Mars, or even the Moon. Although some of them have also gone through significant geological reprocessing, most meteorites are thus samples of 'primitive matter' (the only ones we have at our disposal up to now, if we include the smallest ones, micrometeorites or INTERPLANETARY DUST particles) which keep memories of the sequence of physico-chemical processes which gave birth to the SOLAR SYSTEM, 4.5 billion years ago. Meteorites have allowed us to study the composition and structure of the materials from which the planets were made, and to establish the age of the solar system. In 1987, rare and tiny crystals were isolated from meteorites, which were made in the environment of other stars, before the birth of the Sun, giving us direct access to the nuclear processes which take place in stars and lead to the synthesis of the chemical elements.

Although rare phenomena, meteorite falls have been known since antiquity. Various explanations were advanced, including ejection from volcanoes or transport by violent winds, but by the end of the 17th century, the most widely accepted one was that meteorites were formed in the atmosphere, by the action of lightning on mysterious exhalations from the Earth's interior. Hence the name 'meteoric stones' (stones from the atmosphere) which later gave 'meteorites'. At the beginning of the 18th century, scientists realized that this was unreasonable but, because of the view they had of the universe, they could not imagine that stones could come from beyond the atmosphere. Unable to find a sound explanation of the phenomenon, they simply ignored it. For the whole of the 18th century, the official scientific opinion was that stones which were said to have fallen from the sky were in fact ordinary stones, which had been hit by lightning on the ground. Only at the end of the 18th century, after several widely witnessed falls in various countries, analyses of meteoric stones by the English chemist Howard, and the publication of a visionary book by the German physicist Chladni, did meteorite falls start to gain official recognition again. In 1803, after the historic fall of some 2000 stones near the city of L'Aigle, in northwestern France, and the report on it made by the physicist Biot, no doubt remained. Meteoritics, the science of meteorites, was launched. However, the idea that prevailed then was that meteorites were ejected from lunar volcanoes, and it was not for several decades that the ideas of Chladni, close to the modern ideas on the origin of meteorites, were generally accepted.

Meteorite falls

The fall of a meteorite is a rare but impressive show. As the meteoroid (the initial arriving cosmic body) reaches the top of the atmosphere, friction against the air decelerates it and the conversion of kinetic energy to heat can raise the temperature of the bolide surface and of the surrounding air to several thousand Kelvin. The characteristics of a fall depend on the velocity and mass of the meteoroid, the orientation of its trajectory, and the density and strength of its material. The lower the mass, the more efficient the air drag. Small particles, in the millimeter size range, are stopped in the upper atmosphere and the corresponding meteors do not last more than a fraction of a second: these are the familiar 'shooting stars'. Larger masses penetrate deeper into the atmosphere and make a longer and more spectacular show, in particular in the case of meteorite-producing events. The initial velocity in this case is generally about 15-20 km s^{-1}. A fireball appears at a height of about 100 km. The outermost part of the bolide matter is melted, vaporized and ejected, leaving a trail of vapor, dust and ionized atoms. The mass of the meteoroid thus decreases continuously, a process known as ablation. In addition, because of the very strong tensions to which it is submitted the bolide very often breaks into several, or sometimes thousands of fragments. At a height of about 20 km, the velocity has been so much reduced—it is now about 3 km s^{-1}—that incandescence can no longer be maintained, ablation ceases and the fireball switches off. In general, no more than 10 s have elapsed since it appeared, but in exceptional cases (nearly horizontal trajectories) it can be seen for up to about 40 s. The trajectory now bends over to nearly vertical, and the meteorite finally reaches the ground with its velocity of free fall (one or a few hundred m s^{-1}). If fragmentation occurred in flight, the larger fragments, being less efficiently braked, travel farther, so that the fragments are found distributed on an elliptical

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Figure 1. A few fragments of the L'Aigle meteorite, which fell in 1803. It can be seen on the broken fragment at the bottom left that the heat-affected zone, the black fusion crust, is extremely thin. The largest fragment is approximately 10 cm long. (Photo courtesy of G Segerer, MNHN.)

strewnfield, with the largest fragments at one end and the smallest ones at the other. Meteorite falls are accompanied by sound phenomena. Sonic booms are heard after the disappearance of the fireball, but generally before the landing of the meteorite. Rumbles follow, with those produced in the lowest part of the trajectory arriving first.

Because of this fiery arrival, meteorites have a typical aspect: their surface is black, smooth, with rounded angles (figure 1). However, as the fall is so short, and the heat generated is in large part carried away by the vaporized material and the surrounding air, the interior of the meteorite suffers fortunately very little: the black vitrified pellicle that surrounds it, called the fusion crust, is not usually thicker than a fraction of a mm. A few mm away from the fusion crust, the meteoritic materials are essentially unaffected by the fall.

The largest known meteorite, the Hoba meteorite, found in Namibia, has a mass of about 60 tons. Much bigger meteorites are little decelerated in the atmosphere, which leads to catastrophic issues. Stony meteorites in the size range 10–100 m explode in the atmosphere; the higher the mass and the mechanical strength, the lower the height of the explosion. Such events occur about 8 times per year, without being noticed from the ground. This is probably what happened on 30 June 1908 above the region of the river Tunguska, in Siberia, but in this case, the explosion took place low enough to destroy approximately 2000 km^2 of forest. Still larger stony meteorites, and iron meteorites over ~10 m, reach the ground with almost their original velocity. Most of their enormous kinetic energy transforms into heat and they are entirely vaporized upon impact, together with a much larger quantity of target rock, leading to the formation of an impact CRATER. Some 150 of these craters are known on Earth, and many more on all solid bodies, from Mercury to the satellites of Uranus, making impact cratering a major geological process in the solar system. The largest known crater on Earth, the Chicxulub crater in Mexico with a diameter of ~200 km, resulted from the impact of a meteorite about 10 km in size, 65 Myr ago. There is no doubt that such large impacts cause ecological catastrophes on the scale of the whole planet. 65 Myr ago is the time of the Cretaceous–Tertiary boundary crisis, when a large fraction of the then-living species, including the dinosaurs, disappeared.

Composition and classification

Depending on their structure and composition, meteorites are classified into different classes and groups. This is not just of academic interest, but is aimed at deciphering their various origins. The traditional classification sorted meteorites into stony, iron and stony-iron meteorites. Although this purely descriptive classification can still be convenient, it is now preferred to make the major separation between differentiated and non-differentiated meteorites. Differentiation is the process by which melted rocks separate into different phases having different compositions. The Earth is a differentiated planet: it is composed of a metallic iron core, surrounded by a rocky mantle and a rocky crust. Thus, no single sample is representative of the global composition of the Earth. Moreover, melting not only induced modifications of the chemical composition of the original materials from which the Earth was built, it also erased their structures. This is true for all large planets, including the Moon. Non-differentiated meteorites—also called CHONDRITES—have not been melted in the tiny planets they come from, so that they keep memories of the materials of the solar nebula, the gas and dust cloud from which the solar system was born. They can thus be considered as witnesses of the birth of the solar system. Differentiated meteorites are very diverse, but they have all experienced extensive melting on their parent planets. They have lost to a large extent memories of nebular processes, but are useful to study planetary processes, in particular those which took place on small planets in the early solar system. The various meteorite classes and groups, which are designated by names or symbols, are summarized in table 1. Some groups (CK, CR, CH, R chondrites, Acapulcoites-Lodranites, Brachinites) have been introduced fairly recently, and a few peculiar meteorites which do not fit into existing groups have been omitted.

Non-differentiated meteorites

Chondrites are remarkable in that their chemical composition is very similar to that of the Sun, except for the most volatile elements. Subtle differences have led us to sort them into various groups, as shown in table 1, each of them probably corresponding to at least one parent asteroid. The CI carbonaceous chondrites are believed to

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 1. Meteorite classification.

Non-differentiated meteorites			*Differentiated meteorites*		
FF	Class	Group	Group	Class	FF
4	Carbonaceous chondrites	CI CM CO CV CK CR CH	IAB, IC, IIAB, IIC, IID, IIE, IIF, IIIAB, IIICD, IIIE, IIIF, IVA, IVB, others	Irons	5
			Pallasites		0.4
			Mesosiderites		0.7
<1		R	Eucrites Diogenites Howardites	Basaltic achondrites	6
79	Ordinary chondrites	LL L H	Angrites		<1
			Ureilites		0.5
1.5	Enstatite chondrites	EL EII	Aubrites	Enstatite Achondrites	1
			Anorthosites Basalts	Lunar meteorites	<1
			Shergottites Nakhlites Chassigny Others	Martian meteorites	0.5
		Acapulcoites Lodranites		Primitive achondrites	<1
		Brachinites			

FF: fall frequency (%)

be the most primitive of all chondrites, those whose composition most closely reflects that of the solar nebula. Figure 2 shows a comparison between the chemical abundances measured in CI chondrites and in the solar photosphere. Each point corresponds to an element; a few points have been labeled with the element symbol. The line corresponds to equality between solar and chondritic abundances. The agreement is excellent over 9 orders of magnitude, except for H, He and other rare gases (not plotted), C, N, and O, all volatile elements which have been only partly incorporated into solid bodies, and for Li, which is destroyed by nuclear reactions in the sun. As CI chondrite abundances are measured with a greater precision than those of the solar photosphere and for a larger number of elements, they form the basis for the standard solar system abundances.

Most chondrites are made of silicates, iron-nickel alloys (Fe–Ni) and iron sulfide (troilite, FeS). The most common silicates are olivine ($[Fe,Mg]_2SiO_4$), pyroxene ($[Fe,Mg]SiO_3$), and feldspar (solid solution of $[Na, K]AlSi_3O_8$ and $CaAl_2Si_2O_8$). Many other minerals have been identified in chondrites, but usually only in minor amounts. In the most primitive chondrites, minerals occur as small crystals, arranged in more or less spherical structures, with diameters in the range 50 μm to several mm, called chondrules (see figure 3). The space between chondrules is filled by a matrix made of very fine grains (<1 μm) of similar minerals.

Chemical groups of chondrites differ, in particular, by their oxidation state, which is most conveniently revealed by the chemical form of iron. In table 1, chondrite groups are arranged in roughly decreasing order of degree of oxidation from top to bottom. Carbonaceous chondrites are the most oxidized: many do not contain any Fe–Ni metal, all Fe is in oxidized form, in silicates or other minerals. Enstatite chondrites are the most reduced: they contain large amounts of Fe–Ni metal, but almost no Fe in their silicate which is nearly exclusively enstatite, a purely magnesian pyroxene ($MgSiO_3$). Ordinary chondrites are intermediate, with Fe–Ni alloy content increasing from LL to L to H. These differences probably reflect different local conditions in the solar nebula at formation time.

Although chondrites have a primitive chemical composition, they have nevertheless experienced various amounts of planetary processing. To account for this, they have been sorted into 6 petrologic types. Types 3–6 correspond to increasing effects of thermal metamorphism, i.e. moderate transformations due to temperature increases soon after accretion of their parent asteroids. Type 3s are the least affected: their minerals have variable compositions, in particular the Fe/(Fe+Mg) ratio

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

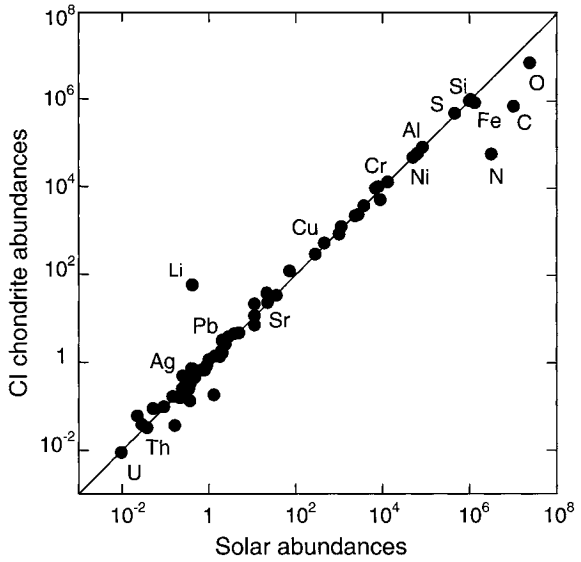


Figure 2. Comparison of element abundances in CI chondrites with those in the solar photosphere, normalized to $Si=10^6$. The line is for perfect correspondence. A few points are labeled with the symbol of the corresponding element. Note the logarithmic scale.

varies among silicates, they are said to be unequilibrated chondrites; chondrules are abundant and sharply defined (figure 3). Type 6s are the most metamorphosed: their minerals have uniform compositions, they are equilibrated chondrites; the chondrule texture has been almost entirely erased by recrystallization, chondrules can hardly be recognized in these chondrites. Types 1 and 2 correspond to negligible planetary thermal processing, but considerable effects of alteration by water on the parent asteroids. Type 1 chondrites, the most primitive from a chemical point of view, are almost entirely composed of very fine-grained hydrous minerals, in particular phyllosilicates analogous to clay minerals found on Earth, which are the products of aqueous alteration of anhydrous silicates. Chondrules are totally absent from these chondrites, but it is not known whether they never existed there, or were destroyed by the action of water. Not all petrologic types are represented in the various chemical groups. For instance, there are no types 1 and 2 among the ordinary and enstatite chondrites, CIs are mostly type 1, CMs are mostly type 2, CVs and COs are type 3.

Chondrule textures imply that they formed as small, independent melted droplets (figure 4). Simulation experiments closely reproduce these textures and suggest formation temperatures in the range 1550–1900°C, and heating times of seconds to hours. However, the physical phenomenon responsible for the (transient) energy inputs necessary to make chondrules is still unknown, although many have been suggested. Most of the hypotheses refer to processes taking place in the solar

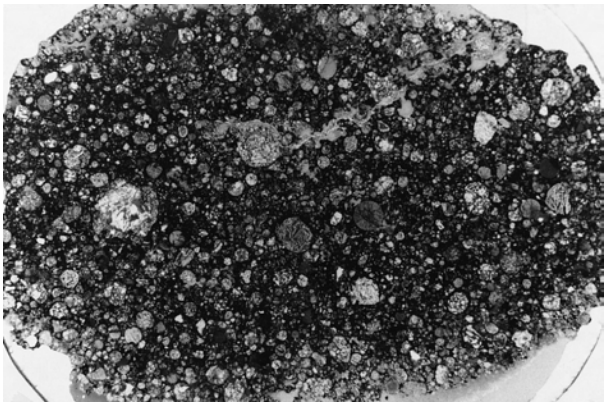


Figure 3. The H3 chondrite Sainte Rose is almost entirely made of chondrules of various textures and sizes. Sample is 2.8 cm long. (Photo courtesy of C Caillet, MNHN.)

nebula, but planetary processes, *a priori* less probable, are periodically revived. This ignorance is all the more irritating in that the high abundance of chondrites (85% of all meteorite falls) and the high proportion of chondrules in chondritic matter (up to 80% in ordinary chondrites) strongly suggest—but do not prove—that chondrule formation was a common process and one of those which most affected the solid matter of the early solar system. Let us mention two of the most recently developed models of chondrule formation: the shock model and the x-wind model. In the former, solid chondrule precursors are melted by friction with the gas accelerated to supersonic speeds by shock waves propagating within the solar nebula. The model accounts for the thermal history of chondrules as inferred from observations and experiments as well as for some other properties of chondrules. The origin of the shocks remains an open question, although the proponents of the model suggest gravitational instabilities in the disk. In the x-wind model,

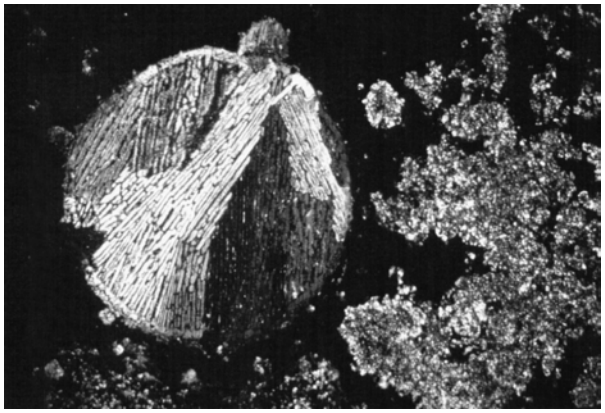


Figure 4. A chondrule (circular, left) and a Ca-Al-rich inclusion (CAI, right) in the Allende CV3 chondrite. These structures once formed independently in the solar nebula. The black background is the fine grained matrix. Width of photo 1 mm. (Photo courtesy of M Bourot-Denise, MNHN.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

chondrules are melted by impulsive protosolar flares at the inner edge of the accretion disk, close to the protosun, and launched to planetary distances by an x-wind outflow. This model is in a large part based on observations of young stellar objects. It offers explanations not only for a number of properties of the chondrules themselves but for more general characteristics of the chondrites (see below). It has not been shown, however, that it can match the chondrule thermal history in detail.

Refractory inclusions are another important type of structure formed at high temperatures (figure 4). They are found in most carbonaceous chondrites, especially in CV3s where they are the largest (few mm to cm), and, rarely, in ordinary and enstatite chondrites. They are assemblages of refractory, Ca- and Al-bearing minerals and are designated for that reason calcium-aluminum-rich inclusions (CAIs). CAIs appear to be the first solids to have formed in the solar system, and they keep memories of events which took place before the birth of the Sun, in the form of ISOTOPIC ANOMALIES for many elements, as will be discussed below. They were made by condensation and/or evaporation in the very first instants of the solar system. In the x-wind model, CAIs are formed by a mechanism similar to chondrules, but closer to the Sun, and they are heated for longer times, including during transport in the x-wind. As a consequence, many elements evaporate, leaving a refractory residue rich in calcium and aluminum oxides and silicates.

Carbon is present in small amounts in all chondrites, and represents up to about 4.5% by mass of the CI and CM chondrites, the most carbon-rich of all. In these chondrites C is mainly in organic matter, the most abundant form of which is a poorly characterized macromolecular component, but many molecules are also present, such as aliphatic and aromatic hydrocarbons, carboxylic and amino acids, etc. Isotopic analyses of hydrogen in this organic matter show that it is strongly enriched in its heavy isotope, deuterium (^{2}H or D), which is difficult to explain by processes that have taken place in the solar system. This suggests that at least part of this organic matter is of interstellar origin, i.e. was synthesized in the interstellar medium before the birth of the Sun, and was thereafter incorporated into the materials of the solar nebula, and survived up to now in some meteorites. It is well known that ion-molecule reactions that occur at low temperatures in interstellar molecular clouds can yield considerable hydrogen isotope fractionation, and indeed D-rich molecules have been detected in the interstellar medium.

The presence in meteorites of complex organic molecules, in particular amino acids that serve to build proteins, have led to some speculation about their possible role in the emergence of life on Earth. It would have been simpler to start making the complex molecules necessary for life from already existing complex molecules coming from space than from the carbon dioxide and nitrogen of the primitive Earth atmosphere. Asymmetric molecules like amino acids possess two forms (two enantiomers), right-handed (R) and left-handed (L), which are mirror images of one another. Biological synthesis only produces the L molecules, while non-biological syntheses lead to a mixture of the two in equal quantities. This latter case applies to most meteoritic organic matter. However, evidences have been found of excesses of the L enantiomer in several meteoric amino acids. These findings need to be scrutinized as terrestrial contamination is an enormous problem for these analyses. Excesses of 34% for L-alanine and 54% for L-glutamic acid in the Murchison (CM2) chondrite, associated with non-terrestrial isotopic ratios of carbon and nitrogen, have been reported by one group, but have not been confirmed by others. Small (1–9.2%) L-enantiomer excesses were found recently in six amino-acids in the Murchison and Murray (CM2) chondrites. These amino-acids are absent or rare in the terrestrial biosphere, thus virtually eliminating the possibility of terrestrial contamination. If confirmed, this could be considered as supporting the hypothesis of extraterrestrial organic matter being at the origin of life.

The discovery, in 2001, of sugar and several related organic compounds in the same two carbonaceous chondrites suggests another fundamental building block of life on Earth may have come from outer space. Sugars and the closely related compounds collectively called polyols, are critical to all known life forms. They act as components of the nucleic acids RNA and DNA, constituents of cell membranes and cellular energy sources. A small sugar called dihydroxyacetone and several sugar acids and sugar alcohols, have been identified in Murchison and Murray. All these are important for life today. One sugar alcohol, glycerol, that is used by all contemporary cells to build cell walls has also been found. In addition, preliminary evidence of other compounds that may contain larger sugars critical in cellular metabolism, such as glucose has been discovered.

Differentiated meteorites

Unlike chondrites, differentiated meteorites have compositions widely different from that of the Sun. They result from total or partial melting of their parent asteroids, followed by a sequence of fractional crystallization. The extreme case is iron meteorites, which are made of only two major elements: iron and nickel. They consist of Fe–Ni alloys, with variable amounts of inclusions of iron sulfide, graphite, silicates and other minerals. On the basis of the concentrations of Ni and minor elements (especially Ga, Ge and Ir), they are sorted into 13 groups (table 1) corresponding to different parent asteroids, but many (~13%) do not fit into any of these groups. Most iron meteorites are believed to be

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

fragments of metallic cores resulting from extensive early melting of asteroids.

Pallasites are made of Fe–Ni and centimeter-sized crystals of olivine in about equal proportions. They probably originate from the interface between the metallic core and the rocky mantle of differentiated asteroids. Eucrites, diogenites and howardites are collectively known as basaltic achondrites, because they resemble a basalt (lava) crystallization sequence. Eucrites are made of pyroxene and feldspar, diogenites of pyroxene, and howardites of agglomerated fragments of eucrites and diogenites. They are magmatic rocks ejected from the crust of a differentiated asteroid. As discussed below, there is evidence that this asteroid is 4 Vesta. Mesosiderites are irregular mixtures of Fe–Ni metal and of silicates similar to those found in eucrites and diogenites. They are thought to result from a collision involving a metallic and a silicate asteroid. Angrites are a very small group of magmatic meteorites which crystallized very early in the solar system history and are most probably from a parent body different from that of the eucrites. Ureilites are distinguished from all other achondrites by their high carbon content (1.5–3%). Their origin is far from clear; various hypotheses include differentiation of a carbonaceous asteroid and/or an asteroid collision. Aubrites consist almost entirely of enstatite. They have a magmatic texture, but are almost surely related to the enstatite chondrites. The rare Acapulcoites–Lodranites and Brachinites are often referred to as primitive achondrites, although they are quite different from each other (the former are Fe–Ni metal-rich, the latter almost entirely made of olivine). In spite of a magmatic texture, they have a nearly chondritic composition. The lunar origin of some 21 achondrites has been established beyond any doubt by comparison with the lunar samples returned by the Apollo and Luna missions. These meteorites comprise both anorthosites from the lunar highlands and basalts from the lunar maria.

Meteorites from Mars

As no sample taken from Mars has yet been brought to Earth, the origin of the Martian meteorites (also called SNCs) cannot be considered as definitely established, but little doubt remains. These meteorites first attracted attention by their very young age (which does not apply to a more recent find) that pointed to an origin from a large planet, as will be discussed in the next section. Gases trapped in vitreous inclusions in two of them exactly match the composition of the Martian atmosphere, as measured by the Viking probes that landed on Mars in 1976. More isotopic and chemical data also support a Martian origin. In 1996, it was claimed that evidence for past life activity on Mars had been discovered in the Martian meteorite ALH84001. The claim was based on several findings including organic matter (polycyclic aromatic hydrocarbons, PAHs), associations of very small grains of iron sulfides and iron oxide, as are produced by bacteria on Earth, and tiny 'worms' about 0.1 μm long, resembling nanobacteria. None of these observations was really compelling by itself, but altogether they could be taken as evidence for primitive life on early Mars. Since then, a number of studies have suggested other interpretations of the observations, including contamination from the Antarctic ice where the meteorite has been collected, and artifacts due to sample preparation. The question is not settled, but it is clear that it has not been proven yet that life was ever present on Mars, and that it will be extremely difficult to do so.

Relation with asteroid classes

Reflectance spectroscopy not only provides information on the mineralogical composition of asteroid surfaces, it also allows a direct comparison between meteorites and asteroids. On the basis of their spectral characteristics in the visible and infrared wavelengths, asteroids have been sorted into some 16 classes. Similarities have been noted for a long time between the spectra of some asteroid classes and laboratory spectra obtained on various powdered meteorites, establishing possible genetic links between asteroid and meteorite classes. For instance M asteroids have been linked either to iron meteorites or enstatite chondrites, and C asteroids to CI and CM chondrites.

A longstanding issue has been that the ordinary chondrites, which make up 80% of all meteorite falls on Earth, have no spectral equivalent in the asteroid belt. Data acquired during the fly-by of S asteroids Gaspra and Ida by the Galileo spacecraft suggest a solution: local spectral variations tend to demonstrate that an ill-defined space weathering process is operating that modifies the spectral properties of asteroid surfaces. Younger terrains have spectra closer to those of ordinary chondrites than older ones. More can be learned on this subject from the data acquired by the NEAR-Shoemaker spacecraft during its orbital rendezvous with the S asteroid Eros from February 2000 to February 2001. Gamma-ray, x-ray, visible and mid infrared spectra recorded by diverse instruments on-board NEAR yield information on the chemistry and mineralogy of Eros surface. The data are far from being as precise and complete as laboratory data, and there are some inconsistencies, which can be due to a number of causes, for instance the fact that, depending on the instrument, the detected photons came from Eros surface layers of different thickness (100 μm for the x-ray spectrometer, several tens of cm for the gamma-ray spectrometer). There is no unambiguous match with any known group of meteorites. Nevertheless, considering the inferred chemistry (relative abundances of O, Mg, Al, Si, K, Ca, Fe) and mineralogy (presence of olivine and pyroxene, the former being more abundant), and taking into account other factors

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

such as the high spectral homogeneity of Eros, the measured values of its density (2.67 g cm^{-3}) and albedo (0.27), the data are overall most consistent with Eros being analogous to ordinary chondrites. A more precise correspondence is not possible, as the x-ray composition favors a link with the H chondrites while the infrared data are more in favor of LL chondrites. Some of the observations (apparently variable Fe/Si abundance ratio, depletion of sulfur with respect to ordinary chondrites) imply that the surface composition of the asteroid has been altered by a range of processes, e.g. impact-related partial melting and migration of metal and sulfide. In any case, it now appears probable that at least part of the S asteroids, the most numerous in the inner asteroid belt, are related to ordinary chondrites.

Most striking is the resemblance between spectra of basaltic achondrites and that of Vesta. The hypothesis that Vesta is the parent body of these meteorites has received strong support from two recent discoveries. Firstly, some 20 small asteroids (<10 km in diameter) with spectra similar to that of Vesta have been found in the main belt. Their orbits are distributed between that of Vesta and the 3:1 RESONANCE, a region where asteroid orbits are unstable and which is believed to be a source of meteorites. Secondly, Hubble Space Telescope images have revealed a huge impact crater (460 km in diameter) on Vesta. This crater may well be the site of origin of both the small Vesta-like asteroids and the basaltic achondrites.

Meteorite ages

Different ages can be defined for meteorites, depending on the method used for the determination.

Crystallization age—age of the solar system

The clock used to determine the age of a rock is generally a couple of isotopes present in the rock, one of which is produced by the radioactive decay of the other, for instance ^{87}Rb–^{87}Sr, ^{40}K–^{40}Ar,^{238}U–^{206}Pb, etc. The transformation of the parent isotope (P) into the daughter isotope (D) follows a law whose time dependence is well established:

$$N_D(t) = N_P(t)(e^{\lambda t} - 1)$$

where $N_D(t)$ is the number of atoms of D produced at time t, $N_P(t)$ is the number of remaining atoms of P, and λ is the decay constant of P. By measuring N_D and N_P, one can thus determine the time t, called the age of the rock. This is just complicated by the fact that one has to take into account the quantity of D initially present at t=0, but there are technical solutions to this problem. Now what does t=0 mean, in other words, when did the clock start? The clock started at the closure of the system, i.e. at the time when no more addition or removal of either P or D occurred in the sample studied, except for what is due to radioactive decay. This happened when the temperature of the sample became low enough to prevent atomic diffusion. The (crystallization) age of a rock is thus the time elapsed since its temperature passed below the closure temperature of the system.

The CAIs mentioned above are present in primitive meteorites, whose temperature always remained low. Their age is thus the time elapsed since the end of the high-temperature event in which they formed. Using the U-Pb clock, an age of 4.566 Gyr has been determined, with an uncertainty of only 2 Myr. No older material has been found, and this age of 4.566 Gyr is taken as the age of the solar system. As all terrestrial rocks have been heated after the formation of the Earth, they are all younger than their planet: the oldest ones have an age of only 3.8 Gyr. Chondrites have ages between 4.56 and 4.50 Gyr, which reflects moderate variations of their thermal histories in their parent asteroids. The ages of the differentiated meteorites vary between 4.56 and 4.45 Gyr, which shows that the thermal activity of their parent bodies, although more intense than that of the chondritic asteroids, occurred at about the same time, very early in the solar system history. The first Martian meteorites studied had very young ages, around 1.3 Gyr. As no asteroid was still hot at such a recent time, the conclusion was inescapable: these meteorites had to come from a large planet that could still have magmatic activity as recently as 1.3 Gyr ago.

Exposure age

When a meteoroid is extracted from its parent body, it starts being exposed to irradiation by cosmic rays, from which it was protected before by overlying rocks. Energetic cosmic ray particles induce nuclear reactions in the meteoroid within the first few meters below the surface, breaking atomic nuclei and creating new ones, called cosmogenic nuclides. This irradiation stops when the meteorite reaches the surface of the Earth, whose atmosphere protects it against cosmic rays. A measurement of the amount of cosmogenic nuclides, for instance ^{21}Ne, yields the duration of the exposure, i.e. the time it took the meteorite to travel from its parent body to the Earth. Typical exposure ages of chondrites are 1–100 Myr, those of irons extend to more than 2 Gyr.

Terrestrial age

As cosmic ray irradiation stops upon arrival on Earth, radioactive cosmogenic nuclides decay within a meteorite without being produced any more. A comparison of their concentration in a collected meteorite with that normally found in a fresh fall gives the time elapsed since arrival on Earth, called the terrestrial age. ^{14}C and ^{36}Cl have convenient half-lives for this kind of measurement (5700 yr and 0.3 Myr, respectively). Terrestrial ages up to 1 Myr have been determined for meteorites collected in Antarctica.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Isotopic anomalies and presolar grains

CAIs not only are the oldest solar system solids, they extend their memory to before the birth of the Sun. In 1976, isotopically anomalous magnesium was found in Al-rich minerals from CAIs of the Allende carbonaceous chondrite: there was an excess of ^{26}Mg of up to a few per cent with respect to the normal composition of Mg. In addition, this excess was correlated with the Al/Mg ratio. The explanation is that, when these minerals formed, they incorporated, together with the stable isotope ^{27}Al, the radioactive isotope ^{26}Al. The decay of ^{26}Al to ^{26}Mg produced the observed excess. ^{26}Al is thought to be synthesized in supernova explosions. As its half-life is only 0.73 Myr, the inferred initial $^{26}Al/^{27}Al$ ratio in the CAIs has been interpreted as implying that a supernova exploded in the vicinity of the molecular cloud that was to become the solar system, no more than a few million years before formation of the CAIs. This interpretation was strengthened by later discoveries of evidence for in-situ decay of other short lived radioactivities (^{53}Mn, ^{41}Ca).

This was the generally accepted picture until very recently, when evidence was found for in-situ decay of ^{10}Be in a CAI from the Allende chondrite. ^{10}Be, as other isotopes of the light elements Li, Be and B, is not produced in stellar nucleosynthesis. It is made by high energy nuclear reactions (spallation reactions) between galactic cosmic ray particles and the interstellar medium. However, the quantity of ^{10}Be inferred to be so produced in an interstellar cloud falls short of explaining that found in the Allende CAI by two orders of magnitude, and ^{10}Be needs to be made in the solar nebula itself. The x-wind model provides a natural framework for that: the protosolar flares that melt the CAIs also irradiate their materials with energetic particles capable of inducing spallation reactions. ^{26}Al and other short lived nuclides can also be made by spallation. With specific, but not unreasonable assumptions, it seems possible to account for all short lived radioactivities at the levels detected in CAIs. In this model, of course, the need for a late supernova disappears. Moreover, the lower amount of ^{26}Al initially present in chondrules compared to CAIs, interpreted in the supernova model as due to a time difference of a few Myr between CAI and chondrule formation, results in the x-wind model from their different formation site: chondrules are made slightly farther away from the Sun than CAIs, at a place where the higher gas pressure partially shields them from the protosolar energetic particles. The question of supernova or not supernova needs further study, but it is clearly among one of the most fundamental questions to answer in meteoritics and related astrophysics at the present time.

Anomalous isotopic compositions had already been found for oxygen in CAIs in 1973. Here, the anomalies could not be explained by the decay of a radioactive isotope. Moreover, it turned out that it was not possible to define a normal solar system isotopic composition of O, because it was different in all analyzed bodies. Indeed, every meteorite group listed in table 1 (hence every parent body) has its own oxygen isotopic composition. Even though this is not well understood, it offers an additional powerful tool for meteorite classification. Oxygen isotope data support an origin from the same parent body for all Martian meteorites. The identical O isotopic composition in enstatite chondrites and enstatite achondrites confirms what their mineralogy suggests, that they are somehow related.

This isotopic variability is generally interpreted in terms of the existence, in the solar nebula, of two oxygen reservoirs, one ^{16}O-rich (probably dust), the other ^{16}O-poor (probably gas). Thus, contrary to what has long been believed, the solar nebula was not thoroughly homogenized. This was confirmed later by the discovery of isotopic anomalies for many other elements (Si, Ca, Ti, Cr, Sr, Ba, Nd, Sm), implying that various nucleosynthetic components issued from different stellar sources escaped homogenization and were preserved with their own isotopic characteristics until they were incorporated into forming CAIs.

However fascinating these isotopic anomalies may be, there is no doubt that CAIs were produced in the solar system, and that individual isotopic components were mixed and diluted in the process. Hence the rather low magnitude of these anomalies: a few per cent at most. In 1987, things changed with the discovery in carbonaceous chondrites of nanometer-sized presolar crystals of diamond, i.e. diamonds made in stellar environments prior to the formation of the solar system, which managed to find their way to the solar nebula and to survive up to now in primitive chondrites. Diamonds were soon followed by larger presolar silicon carbide (SiC, figure 5), graphite, titanium carbide, silicon nitride, and corundum (Al_2O_3). These grains are big enough (up to a few μm) to be analyzed individually with an ion microprobe. The isotopic compositions of their major elements, as well as those of all analyzed minor or trace elements, are very different from the solar system composition. For instance, the $^{12}C/^{13}C$ ratio, which is equal to 89 in the solar system, varies from 2 to 7000 in presolar graphite grains; the $^{16}O/^{18}O$ ratio, 500 in the solar system, varies from 200 to 40000 in presolar corundum grains. Comparison with stellar observation data and with models of stellar nucleosynthesis allows us to identify the stellar sources of the grains. Apparently, many stars provided dust grains to the solar system. Most SiC and corundum grains originate from red giant and asymptotic giant branch stars; diamond, some of the graphite and a small fraction of SiC probably come from supernova(e). The study of presolar grains has quickly expanded to become a field of its own, a new branch of

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

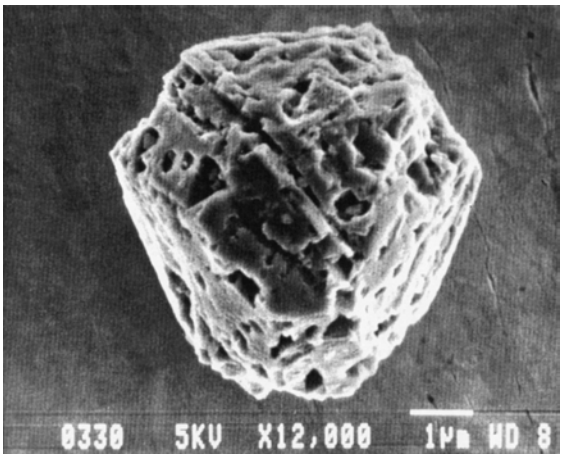


Figure 5. A presolar silicon carbide grain from the Murchison CM2 chondrite, seen in the scanning electron microscope. It probably formed around an AGB star. Scale bar 1 μm. (Photo courtesy of S Amari, Washington University.)

laboratory astrophysics, whose achievements were just unimaginable fifteen years ago. More discoveries are to come. In particular, for experimental reasons, attention has mainly focused up to now on carbonaceous grains. It is conceivable that many presolar oxide grains are still hiding within the oxygen-rich solar system matter, but could be soon unveiled. Not all observations are explained, in particular concerning graphite grains. In addition, large isotopic anomalies have been found in meteorites and interplanetary dust particles for some elements (H, N), suggesting an interstellar origin, but the carrier phases of these anomalies are not known yet. On the other hand, doubts have been raised recently about the presolar nature of the nanodiamonds, as these have been found to be absent in interplanetary dust particles believed to come from comets. Comets are expected to be at least as rich in presolar grains as asteroids. A lower abundance of diamonds in comet samples could mean that at least part of the diamonds formed in the solar system. These are just a few of the questions that remain to be solved.

Bibliography

Bernatowicz T and Zinner E (eds) 1997 *Astrophysical Implications of the Laboratory Study of Presolar Materials* (New York: American Institute of Physics)

Cooper G *et al* 2001 Carbonaceous meteorites as a source of sugar-related organic compounds for the early Earth *Nature* **414** 879–883

Hewins R H, Jones R H and Scott E R D (eds) 1996 *Chondrules and the Protoplanetary Disk* (Cambridge: Cambridge University Press)

Kerridge J F and Matthews M S (eds) 1988 *Meteorites and the Early Solar System* (Tucson: University of Arizona Press)

Pieters C M and McFadden L A 1994 Meteorite and asteroid reflectance spectroscopy: Clues to early Solar System processes *Annu. Rev. Earth Planet. Sci.* **22** 457–97

Zinner E 1998 Stellar nucleosynthesis and the isotopic composition of presolar grains from primitive meteorites *Annu. Rev. Earth Planet. Sci.* **26** 147–88

Claude Perron

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Meteorites from Mars

The SNC METEORITES, named after the initials of the first discovered members (Shergotty, Nakhla and Chassigny), are a group of stony meteorites that are thought to come from Mars, rather than the asteroid belt. They are all igneous rocks (i.e. formed by crystallizing from molten material) and are distinguished from other meteorites by their ages, which are as low as 165 million years old. A young age implies formation on a body that was still active (that is, not totally cooled and solidified) well after the initial aggregation of the solar system about 4560 million years ago. In other words, the SNCs must originate from a planet-sized body, not an asteroid. There are several reasons why the planet from which they originated is almost certainly Mars.

How do we know that the meteorites come from Mars?

On the basis of their young crystallization ages, the SNC meteorites were recognized as different from asteroidal meteorites long before their Martian origin was accepted. It was suggested that the meteorites might actually be terrestrial, or from the Moon, but there are convincing counterarguments for both of these possibilities. Impactors frequently hit the Earth, so it could be argued that the SNC meteorites were broken from the Earth's surface with insufficient energy to escape totally. The counter-argument to this is that all oxygen-bearing rocks from the Earth show a characteristic variation in composition of the three stable oxygen isotopes (see figure 1). Data from the SNCs also fall on a single line indicating that they all come from the same planet, but, because this line is displaced from that for terrestrial samples, the SNCs cannot come from the Earth.

The Moon is heavily cratered and was also considered a possible source object for SNC meteorites. However, the Apollo and Luna samples returned directly from the Moon are so different in age and composition from the SNCs that the latter cannot be from the Moon.

The specific observations that linked SNC meteorites to Mars came with analysis of gases trapped within one of the samples, the Antarctic meteorite named Elephant Moraine (EET) A79001. This meteorite contains inclusions of black glass scattered throughout its mass. The glass was formed by shock melting of mineral grains, presumably during the impact event that lofted the meteorite from the surface of its parent body. Analysis of gas trapped within the glass during the impact shock shows that it is identical in composition to that of the atmosphere on Mars (as measured by the Viking landers in 1976; see figure 2). The only way that this could happen is if EET A79001 came from Mars. Since (on the basis of their oxygen isotopic composition) all the other SNC meteorites come from the same parent as EET A79001, then they too must come from Mars.

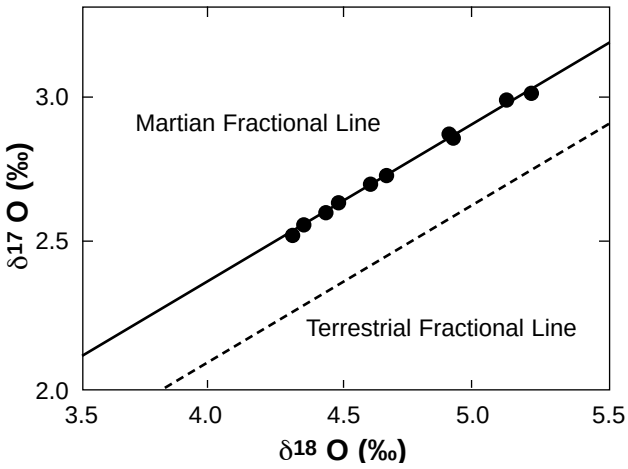


Figure 1. The oxygen isotopic composition of SNC meteorites (adapted from Franchi *et al*, 1999).

What are the SNC meteorites?

The SNCs are named after the type specimens of the three original subgroups. Collection of additional Martian meteorites from Antarctica and the Sahara Desert has extended the number of subgroups to five. The subgroups are rocks that formed in different locations at or below the Martian surface. The groups have different mineralogies and chemistries and cannot all have come from a single impact event. At least three

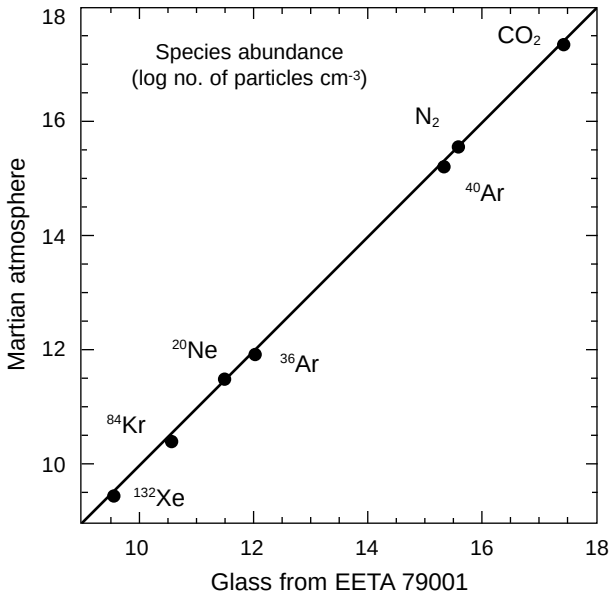


Figure 2. Comparison of the composition of Mars' atmosphere with gas trapped in the shergottite EETA 79001 (after Pepin, 1985).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

craters, with minimum diameters of about 12 km, are required to produce the variety of Martian meteorite types.

The shergottites are silicate rocks that are now further divided into two subgroups, with different formation localities. One group is of fine-grained rocks that originated in a lava flow, while the other is of more coarse-grained rocks that cooled deeper below the Martian surface. The shergottites have crystallization ages of between 165 and 450 Myr. At least two cratering events ejected shergottites from Mars into space; these events happened in much more recent times, 0.5–3 million years ago.

The nakhlites are almost unshocked rocks that formed at or near the Martian surface in a thick flow. They solidified from melts about 1.3 billion years ago, and were ejected from the planet about 10–12 million years ago. The meteorites have been altered by weathering on Mars' surface, leading to the production of secondary minerals (clays, carbonates and sulfates) associated with which are low concentrations of Martian organic material. It has thus been suggested that nakhlites might contain evidence for a Martian biosphere.

Chassigny is the only member of its subgroup. It is almost completely composed of the iron and magnesium silicate mineral olivine, and it crystallized below the Martian surface about 1.3 billion years ago.

Like Chassigny, Allan Hills 84001 (ALH 84001) is the only member of its subgroup. It is the oldest of all the Martian meteorites, having crystallized about 4500 million years ago. It has had a long and complex history of shock and thermal metamorphism and also contains carbonate minerals, indicating that at some stage in its history it has been in contact with Martian water. Since few hydrated minerals (such as are found in clays on Earth) have been identified among the alteration products in ALH 84001, it has been proposed that the carbonates were produced at the surface of Mars in a region of restricted water flow, such as an evaporating pool of brine.

Are there microfossils in Allan Hills 84001?

Throughout ALH 84001 there are patches of bright orange carbonates, up to a few millimeters across. A team in the USA, led by Dr David McKay from NASA in Houston, Texas, studied ALH 84001. They found tiny structures in the carbonates that looked like fossilized bacteria. Associated with the carbonates were organic compounds. Although some of the organics are probably terrestrial contamination, a proportion is Martian. Also present in ALH 84001 are crystals of an iron-rich mineral called magnetite. Magnetite can be produced by bacteria. In 1996, McKay's team reported that the shapes, with their associated organic compounds and magnetite grains, were probably fossilized Martian bacteria. This conclusion has not been widely accepted; however, the report of possible fossils in a meteorite from Mars has inspired much research and re-ignited the debate about life on Mars.

Bibliography

Franchi I A, Wright I P, Sexton A S and Pillinger C T 1999 The oxygen-isotopic composition of Earth and Mars *Meteorit. Planet. Sci.* **34** 657–661

McKay D S, Gibson E K Jr, Thomas-Keprta K L, Vali H, Romanek C S, Clemett S J, Chillier X D F, Maechling C R and Zare R N 1996 Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001 *Science* **273**, 924–930

McSween H Y Jr 1994 What we have learned about Mars from SNC meteorites *Meteoritics* **29**, 757–779

Meyer C *Mars Meteorite Compendium* http://curator.jsc.nasa.gov/antmet/mmc/mmc.html

Nyquist L E, Bogard D D, Shih C-Y, Greshake A, Stöffler D and Eugster O 2001 Ages and geologic histories of Martian meteorites *Space Sci. Rev*. **96** 105–164

Pepin R O 1985 Evidence of Martian origins *Nature* **317**, 473–475

Monica M. Grady

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Meteoroid

A piece of debris from a larger parent body, pursuing its own orbit around the Sun, which has the potential to strike the Earth and produce a meteor or meteorite. There is no agreed dividing line between a meteoroid and a small asteroid, but by 'meteoroid' is usually meant a sub-meter sized particle. A typical meteoroid is roughly 5 mm in size with a density of about 0.2 g cm^{-3}. Such low-density meteoroids are dust particles shed by comets, and are spread around cometary orbits in meteor streams. When these friable particles strike the Earth's atmosphere, they are rapidly ablated and burn up entirely, the friction of their passage producing a brief luminescent trail visible as a meteor. Higher-density meteoroids (about 2 g cm^{-3}) originate from collisions suffered by asteroids. Their higher structural integrity enables them, if of sufficient size, to survive their passage through the Earth's atmosphere and penetrate to the surface as a meteorite.

See also: meteors, meteorites, meteor stream, near-Earth asteroids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Meteors

Meteors are transient luminous phenomena visible in the night sky and popularly called shooting stars. They are caused by small solid cosmic bodies called meteoroids. Meteoroids enter the atmosphere of the Earth with cosmic velocities of several tens of kilometers per second, evaporate and radiate typically at about 90 km above the surface. Most meteoroids are products of disintegration of COMETS and ASTEROIDS but a small percentage may be of interstellar origin. Meteoroids producing typical meteors visible to the naked eye are only few millimeters across.

Meteors brighter than the planet Venus are called FIREBALLS or BOLIDES. Some fireballs may be brighter than the full Moon and in rare occasions a fireball may be visible also during the daytime. Such bright events are caused by meter-sized meteoroids. Fragments of sufficiently large and sufficiently strong meteoroids can survive the atmospheric passage and can reach the ground as METEORITES. However, it is not true that the brighter the fireball the larger is the meteorite that is produced.

Most meteors come from random directions. These are called sporadic meteors. In certain periods of year, however, meteors coming from the same direction on parallel trajectories and with almost identical velocities appear. This phenomenon is called a METEOR SHOWER and the meteors are called shower meteors. All meteors of a particular shower have a common origin in one parent body, in most cases a comet. The meteoroids released from a comet spread along the comet trajectory and form a meteoroid stream. The encounter of the Earth with a meteoroid stream produces the meteor shower. The direction the meteors come from is called the radiant. The apparent paths on the sky of all meteors of a shower intersect in the radiant.

Extremely rich meteor showers are called METEOR STORMS. Meteor storms are rare phenomena occurring a few times per century and lasting only a few hours. Bright fireballs and meteor storms belong to the most impressive astronomical phenomena observable by the naked eye.

Terminology

meteor: luminous phenomenon in the sky, shooting star

meteoroid: cosmic body 0.01 mm–10 m in size; meteoroids produce meteors as they enter the atmosphere

meteorite: any meteoroid remnant which survived the atmospheric flight and reached the ground

Meteor investigation

Meteor astronomy, as a science, started in 1833. On 13 November of that year, a meteor storm was widely observed in Northern America. Although meteor storms had been observed before—there are records in Chinese, Japanese and European chronicles—they were considered as solely atmospheric effects. During the 1833 storm, scientists noted that the apparent point of emergence of the meteors, i.e. the radiant, remained fixed relatively to the stars in the constellation of Leo during the storm. This pointed out the cosmic origin of meteors. The storm was named the Leonids (see figure 1). The radiant was then recognized also for less impressive annual meteor showers such as the PERSEIDS. The relation to comets was established in 1867 when Schiaparelli showed that the Perseid stream has a similar orbit to comet Swift–Tuttle. Subsequently, the parent comets of several other streams were also identified.

The connection of fireballs (see figure 2) and meteorites had also not been recognized for long time. In 1794 CHLADNI argued for a cosmic origin of stones falling from the heavens; nevertheless, his idea was accepted only at the beginning of the 19th century.

During most of the 19th century meteor observations were restricted to visual methods. The first meteor photograph was obtained by Weinek in Prague in 1885 during the Andromedid meteor storm. In 1897 PICKERING in Harvard captured by chance the first photographic meteor spectrum. The first successful program of meteor spectrophotography was started by Blajko in Moscow in 1904. Photography proved to be a very successful method of meteor investigation. Wide-field cameras can be open for hours and wait for a passage of a sufficiently bright meteor. The reduction of photographs from two stations provides the meteor's trajectory. The first meteorite fall was photographed by Ceplecha in 1959—the Příbram meteorite became the first meteorite with a reliably known orbit.

The advancement of electronics during World War II opened the way for radar observation of meteors. The abnormal radio echoes observed by military radars were proved to be due to meteors. Meteors form ionization trails in the atmosphere which reflect radio waves. Meteor radioastronomy was developed in JODRELL BANK OBSERVATORY by LOVELL and his colleagues. The radar technique enables meteors to be observed also during cloudy weather and during daytime.

In the 1970s a television technique started to be used for both direct and spectral meteor observations. The sensitivity of TV systems is higher than photographic sensitivity but the resolution is lower. More recently, atmospheric lidars have been used to detect distinct trails of heavy atoms deposited in the upper atmosphere by meteors.

The description of effects of meteoroids on the atmosphere of the Earth is one motive of meteor studies. Years ago, meteors even served as probes for determining the atmospheric density at large altitudes. This is now better done by rockets. The main goal of meteor astronomy, however, is gathering information on meteoroids themselves: their mechanical properties, chemical composition, abundance and distribution in the solar system, size distribution etc. Meteoroids represent samples of their parent bodies, asteroids and comets and testify to the history and evolution of the solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

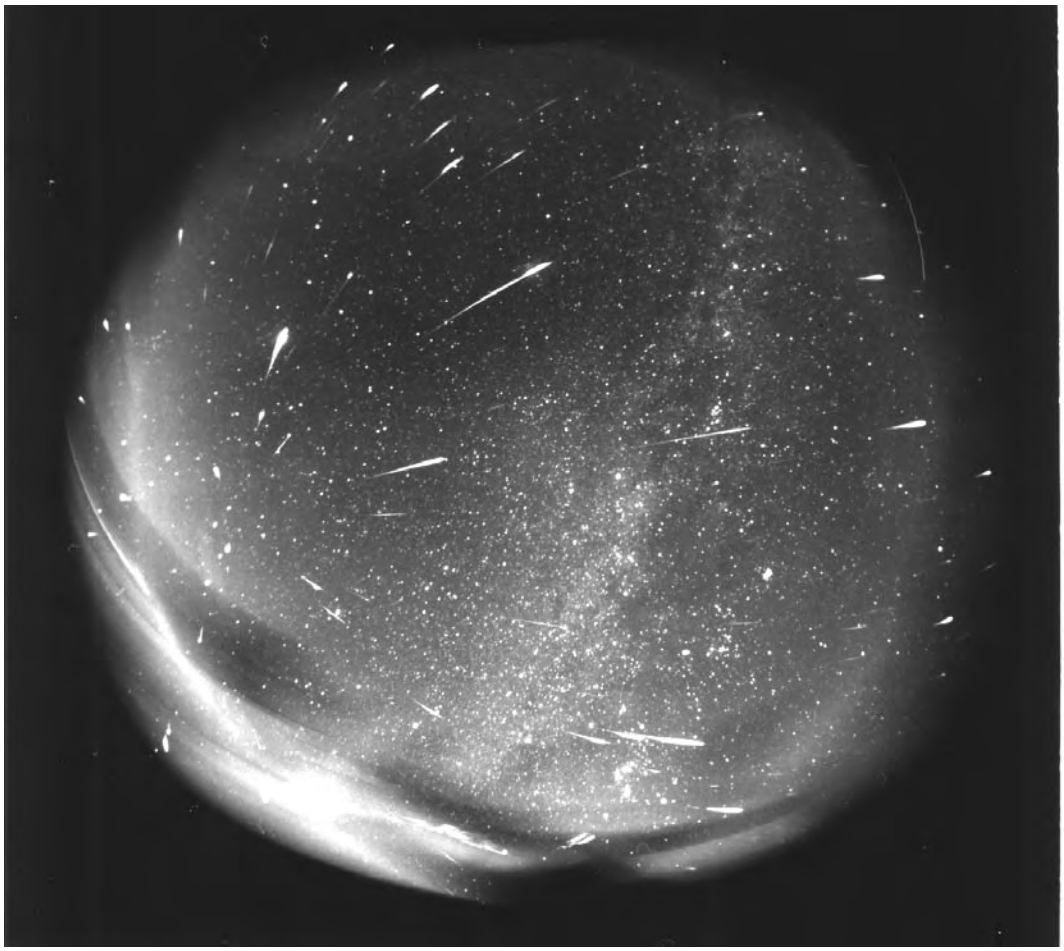


Figure 1. Guided all-sky photograph of the 1998 Leonid meteor shower. About 150 meteors can be seen on the picture. The radiant in the constellation of Leo is clearly recognizable. The exposure lasted for 4 h (23:33–03:37 UT on 16–17 November 1998). Courtesy: Astronomical Observatory Modra, Slovakia.

Meteor atmospheric interaction

The process of meteoroid entry into the atmosphere has several phases. The first phase is preheating, which occurs at the heights about 300–100 km above the Earth's surface. The surface of the meteoroid is heated by the impacts of air molecules. The duration of this phase is from several seconds to tens of seconds. There is no time to heat the body through; the interior remains cool, except for meteoroids smaller than 1 mm which can be heated completely.

After the temperature of the surface reaches 2200–2600 K, ablation, i.e. mass loss, starts. This occurs at heights of 80–130 km, depending on meteoroid velocity, size and material. Ablation can take the form of melting, evaporation and fragmentation of the meteoroid, and all these processes can act simultaneously. A hot gaseous envelope from the ablated material and air is formed. The radiation of the gases produces the meteor phenomenon. The duration is typically 0.1–10 s. As well as losing mass the meteoroid is also decelerated. Deceleration is important for larger meteoroids deep in the atmosphere. Recently, meteor radiation up to heights of 200 km was discovered. The mechanism of this high altitude radiation is not known.

The ablation ceases at a velocity of $\sim$3 km s^{-1}—there is not enough energy to heat the body. If the body is decelerated to this velocity without being completely disintegrated, it continues its flight without ablation and radiation. This phase is called dark flight and takes several minutes. The meteorite is further decelerated and falls by free fall to the surface. The motion is, however, influenced by atmospheric winds. Dark flight starts usually somewhere below a height of 30 km.

Typical meteorites of mass 0.01–10 kg impact the surface with velocities of 10–100 m s^{-1}. The direction at the impact does not reflect the original direction of the meteor; this information is completely lost during the dark flight. Impact creates a pit comparable in size with the meteorite. An explosive impact occurs for very large meteoroids which have not succeeded in being decelerated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

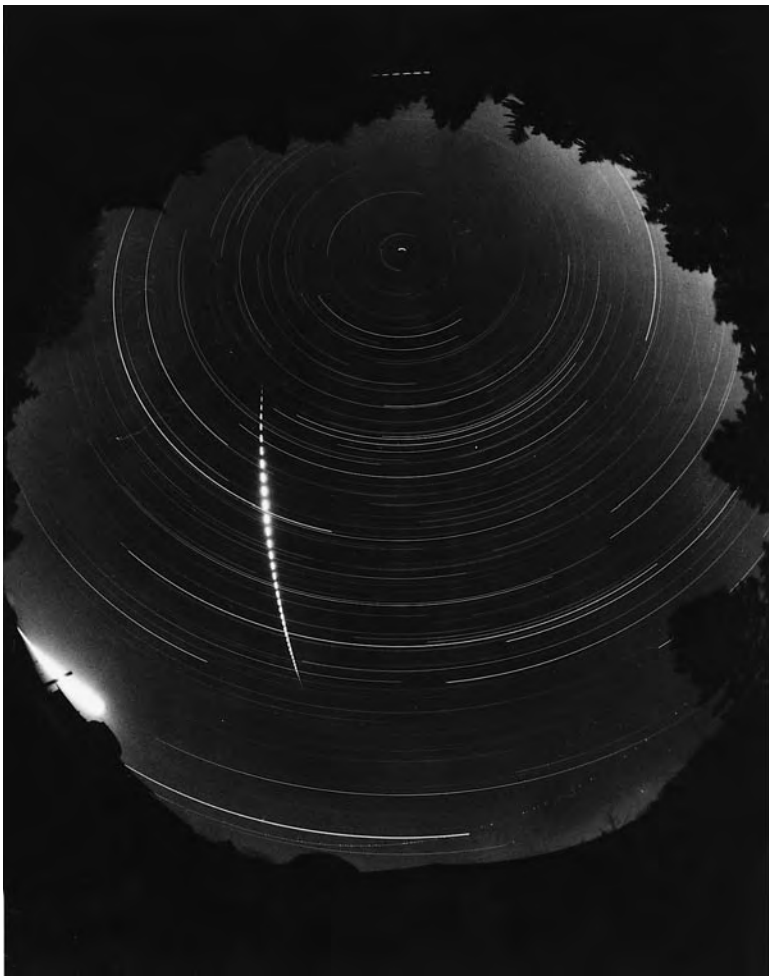


Figure 2. All-sky photograph of a bright fireball EN 220495. The total exposure time was 7 h. The fireball lasted for about 5 s and was brighter than the full Moon. The fireball image has been interrupted by a rotating shutter which enables the fireball velocity to be measured. Photograph: Ondřejov Observatory, Czech Republic.

before reaching surface (no dark flight). An impact crater much larger than the body itself is then formed.

Bodies smaller than about 0.01 mm suffer no ablation and produce no meteors. Such small particles are called dust, not meteoroids. Dust particles are decelerated during the preheating phase and do not reach the temperature necessary for ablation. The heat is radiated out and the particles sediment gradually to the surface. They can be collected in the upper atmosphere as INTERPLANETARY DUST particles (IDPs).

In the upper size range, the terminological boundary between meteoroids and asteroids is rather arbitrary. A 10 m diameter can be used as the dividing value. This nearly corresponds to the largest bodies observed so far during their interaction with the Earth's atmosphere and at the same time to the smallest bodies observed in space in reflected sunlight.

Meteor brightness is usually given in magnitudes as determined by comparison with stars. The absolute magnitude is the magnitude as seen from a distance of 100 km. The meteor light curve is the meteor brightness displayed as a function of time or other quantity (e.g. height). A meteor flare is a sudden increase of brightness. The duration of flares is usually between 0.01 and 0.1 s. A meteor wake is a 'tail' of radiation behind the meteor 'head'. The wake is easily observed in bright fireballs which often have a 'comet-like' appearance in part of their trajectory. The length of the wake may reach a few km. A meteor train is the radiation left in the meteor path after the meteor's disappearance (an afterglow). The visibility of a train may reach 1 h in exceptional cases.

Meteor trajectories and orbits

Meteor trajectories can be considered as straight lines during the few seconds of the ablation phase. Only for long fireballs on nearly horizontal trajectories can the curvature due to gravity become significant. The orientation of the trajectory is described by the equatorial coordinates of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

apparent radiant α_R, δ_R. Another important parameter is the meteor velocity, which, however, changes owing to atmospheric drag. By an extrapolation, the pre-atmospheric velocity, v_∞, can be determined.

The heliocentric orbit of a meteoroid before it encountered the Earth can be computed from α_R, δ_R and v_∞. Since the meteor is observed from the surface of the rotating Earth, the observed velocity must be firstly corrected for this effect which is called diurnal aberration. The second effect is the zenith attraction which is due to the Earth's gravity. Gravity accelerates the meteoroid and bends its trajectory, attracting the radiant to the zenith. The corrected radiant and velocity are called geocentric radiant and velocity. They express the meteoroid velocity vector relative to the Earth's center as it would be without the Earth's gravity. Subtracting the Earth's orbital velocity, the meteoroid heliocentric velocity vector is obtained. From this vector and the position of the Earth at the time of collision, the meteoroid heliocentric orbit is computed.

The minimum meteor initial velocity is 11.2 km s^{-1} which is equal to the escape velocity from the Earth. The meteoroid would be accelerated to this velocity even if the geocentric velocity were zero. The maximum velocity for a meteoroid on a bound heliocentric orbit is 73.5 km s^{-1}. This is given by the maximum meteoroid heliocentric velocity (42.4 km s^{-1}), the maximum Earth orbital velocity (30.2 km s^{-1}) and the gravitational acceleration.

The absolute majority of the meteoroids encountering the Earth move in elliptical heliocentric orbits and are therefore members of our solar system. No interstellar meteoroid has been confirmed among bright photographic meteors. There are indications that about 1% of faint television meteors (produced by meteoroids of the mass of the order of 0.1 mg) are interstellar and that the proportion of interstellar meteoroids may be even higher among much fainter radar meteors.

The distribution of radiants of sporadic meteors on the sky is not isotropic. Five radiant concentrations were identified in radar and photographic surveys. They are called the Helion, Antihelion, Apex, Northern Toroidal and Southern Toroidal sources. The first three lie near the ecliptic, at about 20°, 160° and 90° west from the Sun, respectively. The toroidal sources lie at +60° and −60° from the ecliptic, above and below the apex. The existence of the sources reflects the distribution of meteoroid orbits. The Helion and Antihelion sources are produced by low-inclination prograde meteoroids with significant eccentricity, originating from asteroids and short-period comets. The Apex source is produced by retrograde orbits similar to those of long-period comets. The origin of the toroidal sources is less understood. The corresponding orbits are prograde with high inclination (60°).

Of particular interest are the precise trajectories and orbits of meteorite-dropping fireballs with recovered meteorites. There are only four such cases. The falls of meteorites Příbram, Lost City and Innisfree were photographed and Peekskill was recorded by video cameras. The data are given in table 1. All meteorites are ordinary CHONDRITES and the orbits suggest their origin in the asteroid belt.

Above the sporadic background, meteoroid streams can be recognized. There is no problem in recognizing compact major streams like the Perseids. The existence of low-activity streams with non-distinctive orbits is more difficult to establish. Various criteria which combine the orbital elements and provide a measure of orbital similarity have been proposed. The most widespread is the so-called D-criterion (see p 135 of Ceplecha *et al* 1998). The D-criterion is also used to relate the meteoroid orbits to the orbit of a suspected parent body.

Meteor counts

Meteor counts provide information on distribution of meteoroids in the solar system and on the structure of meteoroid streams.

Meteor showers

The traditional measure of meteor shower activity is the ZENITH HOURLY RATE (ZHR). It is defined as number of meteors observable by one visual observer with the radiant in the zenith under standard conditions. 'Standard conditions' means the limiting magnitude for stars is 6.5. If the radiant lies at an elevation h above the horizon, the ZHR can be computed from the observed hourly rate HR by the formula $\mathrm{ZHR} = \mathrm{HR}/\sin^{\gamma} h$, where the empirical value of the exponent γ is 1.4. If the limiting magnitude is different, the correction depends on the shower magnitude distribution index χ defined as $\chi = n(m+1)/n(m)$, where $n(m)$ is the number of meteors of magnitude m. The correction is $\mathrm{ZHR} = \mathrm{HR}\,\chi^{6.5-L}$, where L is the actual limiting magnitude.

A list of the most pronounced annual showers is given in table 2. The typical maximum ZHR and the duration of noticeable activity (ZHR > 3) are given. The Arietids are a daytime shower observable only by radar. The magnitude distribution index is lower in most showers than in sporadic meteors (where $\chi = 3.4$). The radiant coordinates are given for the time of maximum. The radiant positions drift owing to a systematic change of curvature of heliocentric orbits. Except for GEMINIDS, the known parent bodies of meteor showers are comets. It is not clear whether Phaeton, the parent body of Geminids, is a normal asteroid or an extinct comet.

In addition to the regular annual activity, meteor outbursts occur from time to time. They include exceptional activity of annual showers, observed in LYRIDS, Perseids and especially LEONIDS, as well as the notable appearance of showers which show only negligible or no activity in normal years. The outbursts are often related to the passage of the parent comet near the Earth's orbit. However, outbursts occurred also when the parent comet was far from the inner solar system. Extremely strong outbursts (ZHR > 1000) are called meteor storms. Meteor outbursts and storms are caused by compact filaments of meteoroids and their duration is typically only few hours. The showers which produced meteor storms in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The observed meteorite falls.

	Fireball characteristics						Heliocentric orbit					
Meteorite	v_∞ (km s^{-1})	m_∞ (kg)	z_R (deg)	M (mag)	h_E (km)	m_{Er} (kg)	a (AU)	e	i (deg)	ω (deg)	Ω (deg)	q (AU)
Příbram	20.9	3500	47	−19	(13)	5.8	2.40	0.671	10.5	241.7	17.1	0.789
Lost City	14.2	160	52	−12	20	17.3	1.66	0.417	12.0	161.0	283.0	0.967
Innisfree	14.5	30	22	−12	20	4.6	1.87	0.473	12.3	178.0	316.8	0.986
Peekskill	14.7	1300	86	(−15)	(30)	12.4	1.49	0.41	4.9	308.0	17.0	0.886

v_∞, initial velocity; m_∞, initial mass; z_R, angle to vertical; M, absolute magnitude; h_E, end height; m_{Er}, total mass of recovered meteorites; a, semimajor axis; e, eccentricity; i, inclination; ω, argument of perihelion; Ω, longitude of the ascending node; q, perihelion distance.

Table 2. The most active annual meteor showers.

Name	Maximum date	Duration (days)	ZHR maximum	χ	Parent body	Radiant α	Radiant δ	Velocity (km s^{-1})
Quadrantids	4 January	4	130	2.5	P/Machholz ?	232	+45	43
Lyrids	22 April	6	15	2.7	P/Thatcher	272	+33	49
η-Aquarids	6 May	25	40	2.7	P/Halley	338	−1	66
Arietids	8 June	20	50	2.7	P/Machholz ?	45	+23	38
δ-Aquarids S	29 July	15	10	3.3	P/Machholz ?	339	−17	43
Perseids	13 August	30	80	2.5	P/Swift–Tuttle	44	+58	61
Orionids	22 October	15	25	3.1	P/Halley	95	+16	67
Taurids	3 November	30	10	2.3	P/Encke	48	+18	30
Leonids	17 November	6	20	3.4	P/Tempel–Tuttle	153	+22	71
Geminids	14 December	10	90	2.6	(3200) Phaeton	112	+32	36

last 200 yr were Leonids (1799, 1833, 1866, 1966, 1999), Andromedids (1872, 1885) and Draconids (1933, 1946). The 1966 Leonids reached a ZHR of 144 000 according to some (uncertain) reports.

The year-to-year observation of the activity of meteor showers provides information on the distribution of meteoroids along the cometary orbits and, consequently, on the ejection processes from comets. The storms are probably related to the cometary dust trails observed in infrared light.

Influx of sporadic meteors

Typical hourly rates of sporadic meteors are 5–10. Because of the geometric conditions, more meteors can be seen in morning hours than in evening hours (a diurnal variation). About 700 000 zero magnitude meteors appear on the entire Earth every day.

Sporadic meteors are the major source of cosmic material on the Earth. After relating the meteor brightness and other observable quantities to the meteoroid mass, the influx of meteoroids of various masses can be computed. By adding the information on interplanetary dust and NEAR-EARTH ASTEROIDS, this could be done for the mass range 10^{-21}–10^{15} kg. On average, 1.5×10^8 kg of cosmic material hits the Earth per year. However, in a typical year only half of this mass arrives. The second half comes from the bodies with masses of the order of 10^{14}–10^{15} kg (asteroids several km in diameter) which impact only once in 10^7 years. The second most important bodies for the mass influx are those in the mass range 10^5–10^8 kg (mostly cometary bodies ≈10 m in diameter) with several impacts per year.

The distribution of meteoroids with mass is unequal and no single population index (or magnitude distribution index) can be applied to the whole interval (figure 3). This suggests that several sources of sporadic meteoroids are present. The third most important source is the dust. Meteors in the visual range represent only a minor constituent in terms of mass influx. This may change during short periods of meteor showers and storms.

Evolution of meteoroid streams

Although the parent body of the Geminid meteoroid stream, Phaeton, does not show any cometary activity, typical meteoroid streams are formed by comets. An asteroidal stream can be produced by a collision among asteroids but a powerful collision is rare event and only minor streams are usually formed by asteroidal collisions. A cometary stream can be formed by a complete disintegration of a comet (like COMET BIELA formed the Andromedids) but most streams have been formed less dramatically during the normal cometary activity.

Cometary nuclei are composed by ices and dust grains. Solar radiation causes the ices to sublime, especially near the comet perihelion. The released gas flow also carries solid grains. The grains achieve a speed up to tens of meters per second relative to the nucleus. They then move independently on their own heliocentric orbits. However, since the relative speed is much lower than the comet orbital velocity, the meteoroid orbits remain similar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

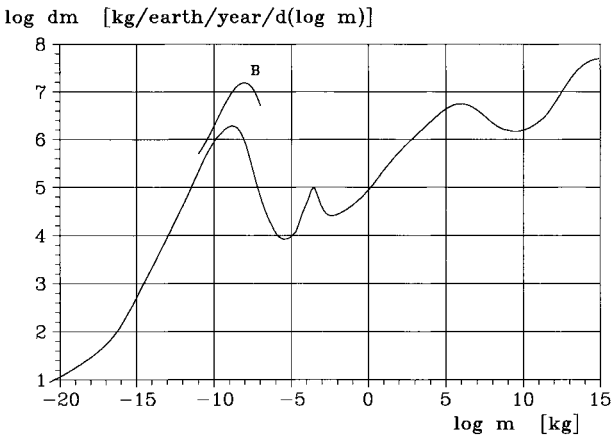


Figure 3. Influx of sporadic meteoroids given in terms of mass per order of mass interval per entire Earth's surface per year as a function of the mass. The four maxima correspond (from the left to the right) to dust, visual meteors, 10 m cometary bodies and km sized asteroids. A newer but unconfirmed calibration for dust is given as curve B. Courtesy: Z Ceplecha.

to the orbit of the parent comet and a well-defined stream is formed. Because of small differences in semimajor axes and thus in the orbital periods, the meteoroids gradually spread along the cometary orbit.

The sizes of particles in a stream formed this way are limited. Small particles are subject to solar radiation pressure. Grains of the size of the order of 0.1 μm are expelled from the solar system. The orbits of somewhat larger dust particles are modified by the RADIATION PRESSURE and by the POYNTING–ROBERTSON EFFECT and a mass segregation within the stream may occur this way. On the other hand, meteoroids larger than several centimeters are too heavy to be carried by the gas outflow against the comet's gravity and are therefore not present in streams formed this way.

In addition to the radiation pressure, both the meteoroids and the parent comet are subject to gravitational perturbations by planets. This leads to gradual dispersion of the stream into sporadic background. The age of the current meteoroid streams is of the order of 10^4 yr only. Their masses are comparable with the masses of the parent comets. Mutual collisions among meteoroids are another effect which leads to the stream dispersion.

The precession of the whole stream can shift the ascending and descending nodes. Any associated meteor showers may then appear or disappear in course of several centuries. This is the case of Geminids which evolve very rapidly. If the orbital evolution of the parent body is complicated, a whole complex of streams can be formed leading to several showers with quite different orbits. This scenario was proposed for comet Machholz 1 with eight possible showers associated, including the Quadrantids.

Classification of meteors

One possible classification is into sporadic and shower meteors. However, meteors differ also in their physical aspects. The most rigorous physical classification is based on the ablation coefficient σ (for definition see the section 'Meteor physics'), fragmentation events or spectra. However, simple quantities such as meteor beginning and end heights can be also used to reveal meteoroid structural classes.

The photographic and television meteors of normal brightness (meteoroids up to about 0.1 kg) have been classified according to the height of the beginning of the luminous trajectory. The height, when normalized to the same initial velocity and incidence angle, shows differences up to 30 km, corresponding to a factor of 50 in the air density. This reflects the differences in the material properties (heat conductivity, heat capacity, density, radiation emissivity). For larger meteoroids (i.e. fireballs), the end height, normalized to the same velocity, angle and mass, proved to be a better discrimination factor. The penetration ability differences are up to a factor of several hundreds when expressed in terms of the air density.

Four main types of fireballs can be recognized. Their typical heights, ablation coefficients and densities are given in table 3. The ablation coefficient can be derived directly for sufficiently well-observed fireballs. The meteoroid bulk density is more difficult to derive and the values given in table 3 are only statistical. The assignment of stones (ordinary chondrites) to the most compact type I was confirmed by the photographed falls of three meteorites. ACHONDRITES may also belong to this type. The connection between types III and comets is confirmed by the fact that the members of typical cometary meteor showers belong to these types. There is no special type for iron meteoroids. Because of their high thermal conductivity, iron meteoroids have high ablation coefficients (between types II and IIIA). This is partly compensated by higher density and the end heights are therefore comparable with those of type II fireballs.

Iron meteoroids are best recognized from meteor spectra. Also achondrites can be distinguished by their spectra. On the other hand, chondrites and cometary meteoroids show similar spectra because they have similar chemical composition despite their very different structure.

Meteor physics

In this section we review the basics of the so-called physical theory of meteors. The theory is mainly used to derive meteoroid properties, such as initial velocity and mass, ablation properties, composition etc, from observations. The observations provide meteor position and brightness as a function of time and, in some cases, the spectrum of the radiation, the intensity and duration of a radar echo or other data. Another approach is numerical modeling of meteors using hydrocodes. Because of the complexity of the problem, these methods provided only partial success and are not discussed here in more detail.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 3. Classification of fireballs.

Type	h_B (km)	h_E (km)	σ (s^2 km^{-2})	ρ_d (g cm^{-3})	Assigned material
I	80	30	0.015	3.7	Stony
II	90	45	0.04	2.0	Carbonaceous
IIIA	100	60	0.10	≈0.8	Cometary
IIIB	110	70	0.25	≈0.3	Soft cometary

h_B, beginning height of meteors ($v_\infty = 25$ km s^{-1}); h_E, end height of fireballs ($v_\infty = 25$ km s^{-1}, $m_\infty = 10$ kg); σ, ablation coefficient; ρ_d, bulk density.

Dynamics and fragmentation

We start with a brief description of the theory of motion of a single ablating body. The body (meteoroid) is characterized at each point of the trajectory by its mass m, head cross-section S, density ρ_d and velocity v. The density of the atmosphere at that point is ρ, t is time. The motion and ablation of the body are governed by two equations, the drag equation,

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = -\Gamma S\rho v^2 \tag{1}$$

and the mass-loss equation

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -\Lambda\frac{S\rho v^3}{2Q}. \tag{2}$$

Here Γ is the drag coefficient, Q is the energy necessary for ablation of a unit mass of the meteoroid ($Q \approx$ (2–8) $\times 10^6$ J kg^{-1}) and Λ is the heat transfer coefficient. In principle Γ can be greater than unity owing to the reactive impulse of reflected and evaporated molecules. Λ is always less than unity, because the energy is consumed also for other purposes, such as heating of the surrounding atmosphere and the evaporated material, radiation, ionization, sonic waves etc. The acceleration due to Earth's gravity is neglected in the above equations.

To solve the equations, new quantities are introduced: the ablation coefficient $\sigma = \Lambda/2\Gamma Q$, the shape factor $A = S/V^{2/3} = Sm^{-2/3}\rho_d^{2/3}$ (A is a dimensionless quantity; for a sphere $A = 1.21$) and the shape-density coefficient $K = \Gamma A\rho_d^{-2/3}$. The drag and mass-loss equations can be then rewritten:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -K\rho m^{-1/3}v^2 \tag{3}$$

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -K\sigma\rho m^{2/3}v^3. \tag{4}$$

The solution is obtained by assuming constant σ and K. One integral is obtained by dividing both equations and integrating:

$$m = m_\infty\,\mathrm{e}^{-(1/2)\sigma(v_\infty^2 - v^2)} \tag{5}$$

where m_∞, v_∞ are the mass and velocity at the beginning of the ablation phase. The ablation coefficient σ is an important parameter describing the meteoroid mass-loss ability.

Equation (5) relates mass to velocity but does not tell us anything about the change of velocity with time. The complete solution can be obtained by means of exponential integral functions and can be found in the specialist literature. The solution is important because observational data (length as a function of time) can be fitted to the resulting functions by optimizing four free parameters. The most important parameters obtained are the ablation coefficient, σ, the initial velocity, v_∞, and the combination $Km_\infty^{-1/3}$. Assuming a typical value for ΓA (≈1.1) and density ρ_d, the initial mass m_∞ can be estimated. Equation (5) can be used to compute mass change during the flight. The mass determined this way is called dynamic mass. Alternatively, if mass is determined from photometry, the density ρ_d can be computed also independently.

The single-body theory is only an approximation. There is much observational evidence that meteoroids are subject to severe fragmentation during their atmospheric penetration. In the case of the Peekskill fireball, 70 fragments are directly visible on one still photograph. More than 1000 meteorites were recovered from a single fireball in Mbale, Uganda, in 1992.

A sudden fragmentation changes the dynamics abruptly because of abrupt change of mass. The single-body theory cannot be used. Nevertheless, one fragmentation point can be found by a sophisticated procedure. It was found that fragmentation occurs in most cases under a dynamic pressure less than 1.2 MPa. In contrast, the mechanical strength of most meteorites is higher than 10 MPa. The structure of meteoroids seems therefore to be strongly affected by the collisions in interplanetary space causing large-scale inhomogeneities.

As well as sudden fragmentation events, continuous fragmentation prevails in some meteoroids, especially of types IIIA, IIIB. Continuous fragmentation can be considered as a mode of ablation and is included in the ablation coefficient. In fact, fragmentation is the most effective ablation process.

Flow regimes

From the aerodynamic point of view three basic flow regimes can be distinguished. They are classified according to the Knudsen number Kn, which is the ratio

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the mean free path of molecules in the vicinity of a meteoroid, L, and the characteristic dimensions of the meteoroid, R: $\mathrm{Kn} = L/R$. In the free molecule regime ($\mathrm{Kn} \gg 1$) the body suffers direct impacts of individual molecules and the collisions among molecules are negligible. In the transition flow regime ($\mathrm{Kn} \approx 1$) the molecular collisions are non-negligible and a shock wave is formed in front of the meteoroid. In the continuum flow regime ($\mathrm{Kn} \ll 1$) air behaves as a continuum, the shock wave is strong and sonic booms are produced.

Radiation and spectra

During the meteoroid interaction with the atmosphere, part of its kinetic energy is transformed into radiation. This statement leads to the conventional luminous equation:

$$I = -\tau \frac{v^2}{2} \frac{\mathrm{d}m}{\mathrm{d}t}. \tag{6}$$

Here I is the meteor luminosity and τ is the luminous efficiency, i.e. the quantity expressing which part of the energy is transformed into radiation. Only the energy loss due to mass loss is considered; the equation neglects the loss of kinetic energy due to $\mathrm{d}v/\mathrm{d}t$, which is smaller in most cases.

The units of I and τ represent some problem. If I is given in absolute units (erg s^{-1} or W), τ is dimensionless. If the meteor brightness is measured in magnitudes M, the intensity is defined as $I = 10^{-0.4M}$ and the commonly used units for τ are 0 mag erg^{-1} s.

The luminous efficiency depends on meteoroid velocity and mass. The actual values are still uncertain. The preferred values for 100 kg bodies are about 4% for $v = 20$ km s^{-1} and about 2% for $v = 10$ km s^{-1}. For smaller bodies the efficiencies are significantly lower. For high velocities the efficiencies are even more uncertain. Usually it is assumed that τ is directly proportional to v. A zero magnitude meteor may be produced by a 0.5 g meteoroid at 20 km s^{-1} or a 0.02 g meteoroid at 60 km s^{-1}.

The integration of meteor light curves using the luminous equation provides the initial meteoroid photometric mass. In some cases the dynamic and photometric masses do not agree, indicating that our knowledge of fragmentation processes and radiation mechanisms is not complete.

More insight into meteor radiation is provided by meteor spectroscopy. Meteor spectra consist primarily of emission lines and molecular bands. Most of them belong to hot vapors of the meteoroid. The strongest lines are usually those of Na, Mg, Ca, Fe, Mn, Cr, Si and atmospheric N and O. The continuous thermal radiation of the meteoroid surface is negligible in most cases. In extremely bright fireballs (mag −20), thermal continuous radiation of the vapors becomes significant.

In the first approximation, meteor spectra can be fitted by a model assuming local thermal equilibrium in the hot gaseous envelope. However, two spectral components are present, one with excitation temperatures of 3500–5000 K and one of about 10 000 K. The high-temperature component is much stronger in fast meteors. The spectra of meteor wakes are far from thermal equilibrium. Little is known about meteor trains.

Meteor spectra can be used to derive the approximate composition of the vapors. The vapor composition does not always directly reflect the composition of the meteoroid. The refractory elements are not efficiently vaporized, especially in slow meteors. This suggests that melting is the main ablation process. Nevertheless, the main compositional classes of meteoroids can be recognized from meteor spectra. See the section 'Classification of meteors'.

Ionization

The collisions among evaporated meteoric atoms and atmospheric atoms and molecules cause ionization of atoms and molecules. The ionized trail persists for some time after the meteor disappearance. The linear electron density α (cm^{-1}) is supposed to be proportional to the meteoroid mass loss rate:

$$\alpha = -\frac{\beta}{\mu v} \frac{\mathrm{d}m}{\mathrm{d}t}. \tag{7}$$

Here μ is the mean mass of meteoric atoms and β is ionization coefficient which is defined as the number of free electrons produced by one evaporated atom. The ionization coefficient depends strongly on velocity; the dependence $\beta \sim v^{3.5}$ has been derived, with $\beta \approx 0.01$ for $v = 20$ km s^{-1} and $\beta \approx 1$ for $v = 70$ km s^{-1}.

Reflections from ionized trails are observed by meteor radars. The reflection lasts as long as the electron density is larger than some critical value (which is dependent on radar wavelength). Electron density gradually decreases owing to electron diffusion. For long-duration trails (>10 s), other effects are also important, in particular the attachment of free electrons to neutral particles. Direct recombination is slow.

The radar reflection occurs not only from the ionized trail but in exceptional cases also directly from a moving source, probably from the vicinity of the meteoroid. This phenomenon is called head echo and is still unexplained.

Bibliography

Bronshten V A 1983 *Physics of Meteoric Phenomena* (Dordrecht: Reidel)

Ceplecha Z, Borovička J, Elford W G, ReVelle D O, Hawkes R L, Porubčan V and Šimek M 1998 Meteor phenomena and bodies *Space Sci. Rev.* **84** 327–471

Jiří Borovička

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Metonic Cycle

An interval of 19 years after which the phases of the Moon recur on the same days of the year, which appears to have been discovered in or around 433 BC by the Greek natural philosopher Meton (*c*. 460 BC–?), and about 50 years later by the Babylonians. This occurs because 235 lunar months is almost exactly equal to 19 tropical years (235 lunar months is equivalent to 6939.689 days, whereas 19 tropical years is equal to 6939.602 days, a discrepancy of about two hours). Meton divided the 19 years into 12 years of 12 months and 7 years of 13 months. The Metonic cycle was used by the Greeks and Babylonians as a reference against which to make adjustments to the calendar and provided the basis of the rules for determining the date of Easter.

See also: calendar, month, year.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Metos of Athens (*c.* 440 BC–?)

Greek astronomer, perhaps born and certainly active in Athens. Made the first recorded attempt to reconcile the lunar and solar calendars. In 432 BC he made observations in Athens of the year and the month. His length for the year was just 1/76 of a day too long. Since the Earth, the Moon and the Sun are independent bodies with quite arbitrary periods, attempts to fit together days (Earth's rotation period), months (the lunar orbital period) and the year (the period of the solar orbit) never quite work, so that for example the year is not a whole number of days. Metos discovered that, over a longer length of time, there almost was a fit between the Moon and the Sun, namely that 235 lunar months (6939.69 days) made up just about 19 tropical years (6939.60 days). This means that after 235 months, the synchronism of the solar and lunar calendars repeats itself and the phases of the Moon recur on the same dates. This became known as the *Metonic cycle* and was adopted from 330 BC as the basis for the Greek calendar until the introduction of the Julian calendar in 46 BC. Since the Metonic cycle is almost equal to 20 eclipse years, 6932.4 days, it is possible for a series (four or five) lunar eclipses to occur on the same dates 19 years apart. CALLIPPUS found an even more exact fit between the lunar and solar calendars after four Metonic cycles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Michell, John (1724–93)

Cleric, geologist and astronomer, born apparently in Nottingham, England. Cambridge-educated and professor of geology there, invented the idea of black holes in a speculative lecture to the Royal Society in 1783 about the gravitational attraction on light in the vicinity of the Sun, or a body more massive than the Sun. Calculated that a body some 500 times the diameter of the Sun (and the same density, so 100 million solar masses and the size of Jupiter's orbit) would attract light such that it could not escape from the Sun's surface without travelling faster than light did (according to RØMER's value). Michell's calculation was based on the now obsolete corpuscular theory of light (that light is made of particles and can be attracted gravitationally), but his speculation contains all the essential ideas behind black holes. Michell also showed by statistical methods (which he initiated) that many of the double stars cataloged for example by Herschel (see HERSCHEL FAMILY) were truly associated and gravitationally bound together. He proposed the experiment with a torsion balance actually carried out by HENRY CAVENDISH to weigh the Earth by determining the gravitational constant G. Demonstrated the inverse square law between magnetic poles, demonstrated light pressure by the deflection of a thin suspended copper sheet, worked on earthquakes and seismology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Michelson, Albert (1852–1931)

German-American physicist, invented the *Michelson-Morley interferometer* to detect, with MORLEY, the motion of the Earth through the ether, but found that irrespective of the orientation of the interferometer, no motion could be detected. This was explained by the Fitzgerald–Lorentz contraction, and was a foundation for Einstein's theory of special relativity. Michelson went on to apply interferometry to astronomy by mounting an interferometer on a steel beam on the 100 in Mount Wilson telescope, measuring the diameters of six stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Michelson–Morley Experiment

A celebrated experiment carried out by A A Michelson (1852–1931) in 1881 and by Michelson together with Edward Morley (1838–1923) in 1887, which attempted and failed to detect the motion of the Earth through the ether. The basis of the experiment was that if a beam of light traveled a known distance in the direction in which the Earth was supposed to be moving through the ether, and another beam traveled the same distance at right angles to this direction then, if light moved at a constant velocity through the ether, the two beams should take different times to cover the distance. The failure to detect any difference is accounted for by the special theory of relativity.

See also: ether, special relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Michigan Radio Astronomy Observatory

Located in Ann Arbor, close to the University of Michigan campus. The primary instrument at the observatory is a 26 m diameter parabolic reflector with an 11 m focal length. This telescope was constructed in 1958, under contract from the Office of Naval Research. Since 1968, the research and operation of the observatory have been supported jointly by the National Science Foundation and the University of Michigan.

The 26 m telescope is used to study total flux density and linear polarization from active extragalactic objects. The discovery that such objects vary in brightness with time-scales of weeks to a few years was made using the Michigan instrument in 1964–5, and the discovery of variability in polarization followed in 1966.

Current research at the observatory is focused on understanding the physical conditions in the relativistic flows in jets and on investigating the relationship between variations in the radio wavelength and the behavior of the radiation in other regions of the electromagnetic spectrum.

For further information see
http://www.astro.lsa.umich.edu/obs/radiotel/radiotel.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Microchannel Plate Detectors

Microchannel plates (MCPs) are imaging electron multipliers of high gain which have been widely used in space astronomy from the visible to the x-ray and in space plasma analysis. MCPs are, in fact, the archetypal detectors for space science: fast, compact, low-mass, low-power devices with a wavelength response extending from the optical to the x-ray, coupled with high sensitivity to ions and electrons.

Technology

The origin of these detectors can be traced to 1930, to P T Farnsworth's US patent for a continuous-dynode electron multiplier—an evacuated hollow tube whose internal surface is coated with a resistive material of high secondary electron yield and along which a large (∼kV) potential difference is applied. When ionizing radiation strikes the inside wall of the tube, low-energy (∼eV) electrons emitted from the surface are accelerated down the channel by the applied electric field. If these electrons gain sufficient energy from the field before next striking the wall, they may each promote more than one secondary electron into the vacuum and an avalanche may develop.

The first practical embodiments of this concept were in the US, UK and USSR in the early 1960s. Single-channel channel electron multipliers (CEMs) with ∼1 mm bores were fabricated from, among other materials, reduced lead (or lead–bismuth) silicate glasses containing significant weight fractions of the alkali metals (Cs, K, Na, Li). The channels of these CEMs were curved in order to give high electron gains (10^7–10^8 for bias voltages of 3 kV) while suppressing unwanted afterpulses from positive ion feedback. The operating pressure for CEMs and MCPs is typically 10^{-6} mbar. With the addition of an input cone to increase collecting area, CEMs were very rapidly adopted for use in space science—as detectors at the foci of paraboloidal x-ray and EUV concentrators on the COPERNICUS (1972) and Apollo–Soyuz (1976) missions and, in particular, as detectors for positive ion and electron analyzers, an instrumental niche they occupy to the present day.

The first MCPs were simply arrays of straight-channel CEMs bonded together using a low-melting point solder glass. By the late 1960s, however, in a development driven by the military need for night vision image intensifiers, the modern 'two-draw' etchable core process, with its similarity to fiber optic production technology, had evolved (variants on this process include 'hollow' and 'triple' drawing). A bar of acid-soluble core glass is inserted into a hollow tube of cladding glass and the composite billet pulled down through a vertical oven heated above the softening point of the glasses. The billet (and hence the final microchannel) is usually circular in cross-section, although square channels are required for MCP optics (below). The ∼1 mm fiber emerging from the bottom of the oven is cut into lengths which are then stacked into a hexagonal or square former to create a multifiber. This multifiber is again drawn down, cut, stacked and pressure fused at high temperature to produce a boule or block from which MCP blanks of the appropriate thickness can be sliced. The blanks are then acid etched to produce the open channel structure. The embryonic channel plate is then chemically treated in order to establish the appropriate channel resistivity and secondary electron yield coefficient and, finally, nickel alloy (Nichrome or Inconel) electrodes are vacuum evaporated onto the front and rear faces, connecting all $\sim 10^7$ channels—each one capable of acting as an independent photomultiplier tube—in parallel. The finished channel walls have a glass composition which varies with depth: in particular, a silica-like surface layer some 100–500 Å thick overlays a thicker semiconducting lead-rich layer. The former layer controls the secondary electron yield and, hence, the quantum efficiency and gain; the latter layer, in which conduction takes place by a hopping mechanism, controls the multiplier current supply and, hence, the gain-versus-count rate characteristic. The very small volume of the active layer also makes MCPs intrinsically radiation hard in a space environment.

Since the linear gain in a continuous-dynode multiplier depends only on the length-to-diameter ratio of the channels (typically 40:1 in night vision MCPs), not explicitly on channel diameter, MCPs can be fabricated with very small pores—leading to imaging detectors of very high spatial resolution. Since the pulse transit time in a 0.5 mm long channel, biased to 1000 V, is of order 100 ps, MCP detectors are also intrinsically fast, a property used to great advantage in time-of-flight space plasma analyzers. The pore diameter in state-of-the-art MCPs is 6 μm, while the open area fraction is typically around 65%. The imaging area of an MCP may now approach 100 cm^2—as in the CHANDRA High Resolution Camera (see table 1). Formats include circular, square, rectangular and annular (for readout of hemispherical charged particle analysers). MCPs with either convex or concave surface profiles have been produced to match the focal surfaces of EUV and UV telescopes and spectrometers. Radioisotope-free 'low-noise' MCP glasses have been formulated, which lack the internal beta-induced background of standard compositions containing K and/or Rb. By varying the production process, the resistance of a 25 mm diameter MCP can be as high as 1 GΩ or as low as 100 kΩ. Low-resistance (high-current) MCPs are of great interest for high count rate applications but, because MCP glass has the negative temperature coefficient (around $-1\%\,^{\circ}C^{-1}$) characteristic of a semiconductor, low-resistance plates are susceptible to 'thermal runaway' which may lead to the local melting of the channel matrix. Very-low-resistance MCPs must be heat-sunk to a conducting anode if they are to be operated stably. Despite its flexibility, however, the two-draw process is costly, complex and close to the end of its technical evolution: for example, the production of large-area glass MCPs with channels much smaller in diameter than 6 μm will be extremely

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

challenging. The long-term future of MCP fabrication may well lie in micromachining, using reactive ion etching or other techniques to produce precisely dimensioned arrays of small channels in Si wafers. Micromachining would provide, in principle, better control of resistivity, better high-vacuum compatibility (new glass MCPs copiously outgas water vapor) and the ability to provide 'one-off' formats at relatively low cost. The first attempt to fabricate such 'alternative technology' Si MCPs was, however, as long ago as 1977; alternative CEM technologies have been successfully based on both ceramics and conducting polymers.

Photocathodes

The intrinsic quantum efficiency of bare glass MCPs in the 'open window' soft x-ray and EUV bands is rather low (∼5–15%) and varies strongly with angle of photon incidence to the channel axis. Alkali halides such as CsI or KBr are therefore thermally evaporated onto the input surface of the MCP to produce substantial enhancements in efficiency. At 120 Å, the detection efficiency may reach 70% for a CsI-coated plate preceded by an open mesh biased slightly more negative than the input face itself. This 'repeller grid' allows events from the interchannel web to be counted by returning low-energy secondary electrons to their plane of emission, where they may enter a channel and initiate an avalanche. Alkali halide coatings may also be used to increase the electron or positive ion detection efficiency of MCP detectors for use in plasma analyzers.

In the optical and far-ultraviolet (FUV) bands, MCP-based detectors are sealed vacuum tubes. The semitransparent photocathode is deposited on a transparent substrate (borosilicate glass or fiber optic faceplate for the optical band; quartz or sapphire for the FUV). Photoelectrons emitted from the cathode are accelerated towards the MCP, which acts as an electron detector, not as a direct photon detector. Photocathodes for the optical are generally antimonized multialkalis such as S20 (Na_2KCsSb) with a peak efficiency of 20% and a response extending out to a wavelength of 0.9 μm. CsTe is also employed in sealed intensifiers for the blue end of the visible spectrum. Negative electron affinity materials such as GaInAs or GaAs, with an extended infrared response but very severe tube vacuum requirements, have found limited use in ground-based astronomy.

Readout

In space astronomy, MCP detectors are usually based on the ion-feedback-free chevron (two MCPs with offset channel axes and an interplate gap between them—an arrangement first developed c 1970 (see figure 1)) or Z-stack (three plates in contact) configurations. The output pulse height distribution (PHD) of a single MCP is generally quasi-exponential, placing a severe requirement on the dynamic range of subsequent processing electronics. The PHD of a multi-stage detector composed of plates with individual length-to-diameter ratio of 80:1 or 120:1, however, is peaked with a modal gain of $\sim 10^7$ (0.625 pC/pulse). A large number of signal location techniques have been developed to encode the position of the MCP charge cloud at the output of the final plate in the stack. Readout methods may be either electrically or optically coupled. In the latter category, commonly used in detectors for visible wavelengths, the output charge cloud is accelerated through a large potential difference and allowed to strike a phosphor screen. The light spot from the phosphor is then channelled via a fiber optic faceplate onto a silicon charge coupled device (CCD) or charge injection device (CID). By using a fiber optic with a suitable taper ratio, a small CCD can be used to encode a much larger MCP input surface. Suitable interpolation algorithms allow the centroid of the light flash (and hence the photon arrival position) to be found to much better than one CCD pixel width, allowing images to be formed which clearly show the microchannel pore structure.

Electrical readout elements are generally based on (a) charge division in a resistive sheet, as in the EXOSAT CMA and ROSAT Wide Field Camera detectors, (b) charge partition among a pattern of conducting elements, for example the wedge-and-strip anodes used in the MCP

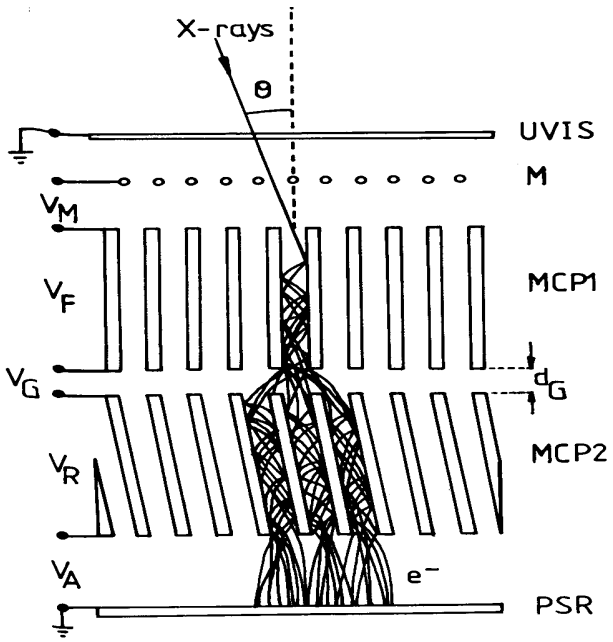


Figure 1. Schematic view of avalanche propagation in a two-stage 'chevron' MCP x-ray detector. The channels of the rear plate (MCP2) are biased at an angle (6°–15°) to those of the front plate (MCP1) in order to stop positive ions generated at the output end propagating back to the entrance surface and initiating afterpulses. PSR, position-sensitive readout; UVIS, UV–ion shield composed of metalized plastic; M, mesh. Typically, the applied potential differences V_F and V_R are ∼1500 V. For electron detection, the front plate input face is held at ground and the PSR at high positive potential.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. MCP detectors.

Mission (launch date)	Instrument	Comments	References
Einstein Observatory (1978)	High Resolution Imager (HRI)	Provided first high-resolution images in x-ray astronomy	Kellogg E *et al* 1976 *Rev. Sci. Instrum.* **47** 282
EXOSAT (1983)	Channel Multiplier Array (CMA)	Focal plane detector for small soft x-ray telescopes	Taylor B *et al* 1981 *Space Sci. Rev.* **30** 479
Rosat (1990)	High Resolution Imager (HRI)	Upgraded copy of Einstein HRI detector, used in pointed mode observations with Rosat x-ray observatory	Zombeck M V *et al* 1995 *Proc. SPIE* **2518** 304
	Wide Field Camera (WFC)	EUV (60–140 Å) detector, input surface spherically curved to match telescope focal surface	Barstow M A and Sansom A 1990 *Proc. SPIE* **1344** 244
Extreme Ultraviolet Explorer (EUVE) (1992)	–	Seven individual survey and spectrometer detectors operating in the EUV	Vallerga J V *et al* 1994 *Proc. SPIE* **2280** 57
ALEXIS (1992)	–	Spherically curved detectors at foci of normal incidence multilayer optics optimized for 133 174 and 188 Å	Priedhorsky W C *et al* 1993 *IEEE Trans. Nucl. Sci.* **40** 863
SOHO (1996)	UVCS	Delay line detector observing solar corona out to 10 solar radii in Ly α and O VI bands	Siegmund O H W *et al* 1995 *Proc. SPIE* **2518** 344
	SUMER	Delay line detector viewing solar atmosphere in 500–1600 Å band	As UVCS
Hubble Space Telescope	STIS	MAMA detector for Space Telescope Imaging Spectrometer	Joseph C *et al* 1994 *Proc. SPIE* **2282** 116
	COS	Cosmic Origins Spectrograph—delay line detector for 2002 HST refurbishment mission	http://www.ozma.ssl.berkeley.edu
Chandra (1999)	High Resolution Camera (HRC-I and HRC-S)	Large area imager 93 × 93 mm^2 area, spectrometer readout detector 300 × 20 mm^2; Radioisotope-free glass	http://hea-www.harvard.edu/HRC/HomePage.html
Far Ultraviolet Spectroscopic Explorer (FUSE) (1999)	–	170 × 12 mm^2 delay line detector with 18 μm spatial resolution, working in the 900–1200 Å band	As COS

detectors on the EUVE satellite, or (c) timing of the charge pulses at the ends of a delay line, as in the detectors on the FUSE mission. Because, crudely, it is simpler to measure a time interval to high precision than it is to digitize a charge to the same precision, delay line readout generally delivers higher spatial resolution (<10 μm) at higher count rates (>100 kHz). More complex position encoding methods use the coincidence counting of pulses on orthogonal strips, as in the Multi Anode Microchannel Array (MAMA) detector in the STIS instrument on the HUBBLE SPACE TELESCOPE. Here, $2M$ amplifier–discriminator chains encode M^2 pixel addresses. The crossed grid charge collector used on the Chandra High Resolution Camera detectors uses orthogonal grids of fine wires to intercept the charge cloud emanating from the rear MCP of a chevron pair. The wires of each grid are terminated by a resistive line subdivided by many (64 per axis) preamplifier inputs. The x or y location of the x-ray interaction is found by taking appropriate ratios of the charges measured by neighboring preamplifier 'taps'. M-fold partitioning of the encoder axis means that the position signal executes its full dynamic range over a distance L/M, where L is the side length of the image field, allowing high resolution (<25 μm) to be obtained over large (L = 10 cm) image areas for modest levels of signal digitization.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Flight instruments

Table 1 describes a number of MCP detectors flown on major space astronomy missions for the UV, EUV and x-ray wavebands.

Microchannel plate x-ray optics

Uniquely, MCP structures can function not only as detectors but also as focusing optics for the UV, EUV and x-ray bands. The operation of MCP x-ray optics was first described by researchers at CSIRO and the University of Melbourne in 1989 and the first images with square-pore MCPs were reported by the University of Leicester in 1993. MCP optics operate by grazing incidence reflection from the internal channel walls. Currently, the best angular resolution for the central component of the cruciform focus is about 5 arcmin. MCPs potentially provide important advantages over conventional x-ray optics because of their very low mass-to-effective area ratio and very large (in principle, 4π sr) fields of view. Applications now being studied include an imaging all-sky monitor (the so-called *lobster eye* telescope) for the 0.1–3 keV band, a hard x-ray telescope for the 2–60 keV band, imaging x-ray fluorescence spectrometers for the surface geochemistry of airless inner solar system bodies and a wide field-of-view aurora imager for the 1300–1600 Å band.

Bibliography

Detailed technical reviews of MCP detector operation are given by

Wiza J L 1979 *Nucl. Instrum. Methods* **162** 587

and by

Fraser G W 1984 *Nucl. Instrum. Methods* **221** 115

Up-to-date information on novel MCP detectors and optics can be found on the website maintained by the Experimental Astrophysics Group, University of California, Berkeley (see Table 1 entry for COS) and the University of Leicester Space Research Centre (http://www.src.le.ac.uk).

G W Fraser

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Microlensing

Microlensing refers to the special case of GRAVITATIONAL LENSING where the multiple images produced are too close together on the sky to be observed as separate images. However, the lensing can still be detected because these multiple images appear as a single object of increased apparent brightness. Although this is not detectable in a one-off observation (since we do not know the 'normal' brightness of the source), with the passage of time the lens moves across the Earth–source line and the amount of brightening changes. Typically the source will appear to brighten, reach a maximum and then fade symmetrically back to normal over the course of a few weeks or months; this is called a 'microlensing event'.

The major application of microlensing, suggested by Paczynski in 1986, is in the search for the DARK MATTER which is strongly believed to exist from rotation curves of spiral galaxies etc. Since the lensing effect depends only on lens mass, it can be used to search for very faint or invisible objects such as brown dwarfs, neutron stars, old white dwarfs or black holes, which might make up the dark matter. These are collectively known as massive compact halo objects or MACHOs, in contrast to the hypothetical weakly interacting massive particles or WIMPs.

To understand the basics of microlensing, consider a small massive object (the lens) situated exactly on the line of sight from Earth to a background star and consider a number of light rays radiating from the star passing the lens at different distances and being bent towards the lens. Since the bending angle for a light ray increases with decreasing distance from the lens, it is clear that there is a unique 'miss distance' such that the ray will be deflected just enough to hit the Earth; this distance is called the Einstein radius. By rotational symmetry about the Earth–star axis, an observer on Earth with perfect resolution would see the star lensed into an annulus centered on its 'true' position, called an Einstein ring. As the lens is moved slightly off the line of sight (e.g. by 0.1 Einstein ring radii), the Einstein ring splits into two banana-shaped arcs, one on the same side of the lens as the source, one on the opposite side. As the lens moves further off (more than 1 Einstein radius), the arcs become more circular, the 'opposite-side' arc fades very rapidly and the 'same-side' arc turns into a slightly deflected and nearly circular image of the star. Figure 1 illustrates a sequence of such images for a typical microlensing event.

Although the perfect alignment giving the Einstein ring will rarely occur in practice, it is still a very important concept because the size of the hypothetical Einstein ring sets the length scale over which substantial brightening will occur. As we will see, for a typical lens in our Galaxy the radius of the Einstein ring $r_{\rm E}$ is roughly $8(M/M_\odot)^{1/2}$ AU (astronomical units), where M is the lens mass. Knowing this scale allows us to understand most of the general characteristics of microlensing: it is extremely small compared with the typical distance to a lens, so the angular separation of the two images will be too small to resolve, hence the 'micro'lensing. However, it is considerably larger than either the size of a star or the size of a MACHO, so we can usually approximate the lens and source as pointlike, which leads to a simple prediction for the lightcurve shape. Also, $r_{\rm E}$ is very small compared with the typical separation of objects in the Galaxy, which implies that microlensing will be a very rare phenomenon. Another notable feature is that $r_{\rm E}$ is proportional to the square root of the lens mass. This means that the area of sky 'covered' by a lens (at fixed distance) is proportional to its mass, so the total fraction of sky covered depends only on the total mass density in lenses, not the individual lens masses. This fraction is called the 'optical depth' τ, and is $\sim 10^{-6}$ for Galactic microlensing. The duration for a microlensing event is given by the time for the lens to move by $2r_{\rm E}$ relative to the Earth–star line; for typical Galactic speeds of 200 km s^{-1}, this is $\sim$130 days $\times$ $(M/M_\odot)^{1/2}$.

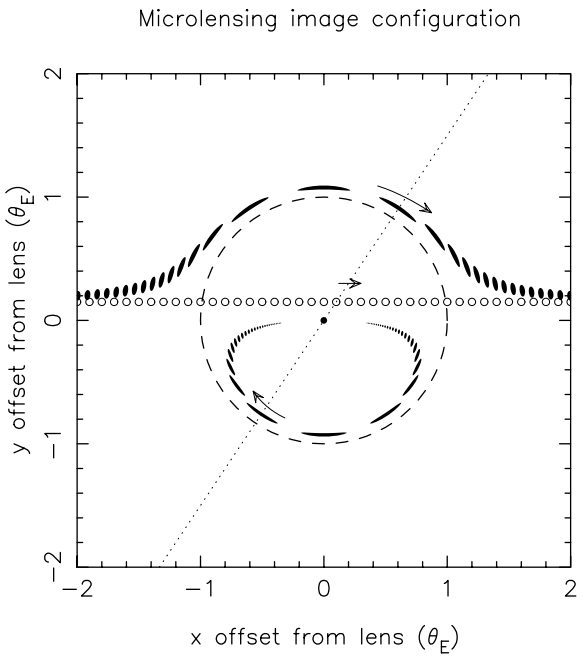


Figure 1. A microlensing event seen at 'perfect' resolution. The axes show angular offsets on the sky from the lens (central dot) in units of the Einstein angle; the dashed circle is the Einstein ring. The series of small open circles shows the 'true' source position at successive timesteps. For each source position, there are two images (solid blobs) collinear with the lens and source, as indicated by the dotted line; the arrows illustrate their motion.

For perfect alignment, simple geometry gives the (small) deflection angle of the light ray meeting Earth as $\alpha = r_{\rm E}/D_{\rm ol} + r_{\rm E}/D_{\rm ls}$, where $D_{\rm ol}$ is the observer–lens distance, $D_{\rm ls}$ is the lens–source distance etc. Requiring this to equal the general relativity deflection, $\alpha = 4GM/c^2 r_{\rm E}$, we obtain

$$r_{\rm E} = \left(\frac{4GM}{c^2} \frac{D_{\rm ol} D_{\rm ls}}{D_{\rm os}} \right)^{0.5}.$$

The angular Einstein radius is just $\theta_{\rm E} \equiv r_{\rm E}/D_{\rm ol}$. If we now introduce a small offset of the lens by a distance b from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

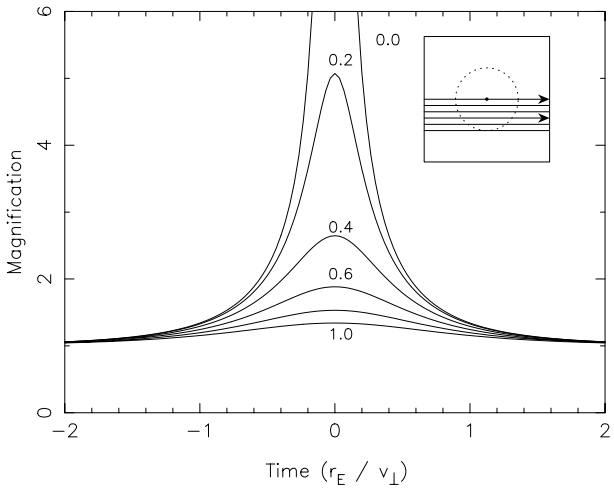


Figure 2. Microlensing event lightcurves (magnification versus time) for six values of the impact parameter $u_{\min} = 0.0, 0.2, \ldots, 1.0$ as labelled. Time is in units of the Einstein radius crossing time $r_E/v_\perp$. The inset illustrates the Einstein ring (dotted circle) and the source paths relative to the lens (dot) for the six curves.

the Earth–source line, i.e. an angle $\beta \equiv b/D_{ol}$, a simple generalization gives the two image angular positions (relative to the lens) as

$$\theta_\pm = 0.5[\beta \pm (\beta^2 + 4\theta_E^2)^{1/2}].$$

Since lensing preserves surface brightness, the magnification A_i of each image is given by the ratio of image to source areas, which for a 'small' source and any axisymmetric lens is just

$$A_i = \left| \frac{\theta_i}{\beta} \frac{\mathrm{d}\theta_i}{\mathrm{d}\beta} \right|.$$

For a point lens, this leads to a total observed magnification as the sum of the two image magnifications,

$$A = A_+ + A_- = \frac{u^2 + 2}{u(u^2+4)^{1/2}} \qquad (1)$$

where $u \equiv \beta/\theta_E = b/r_E$ is the misalignment in units of the Einstein radius. This behaves as $A \approx u^{-1}$ for $u \lesssim 0.5$, so the magnification may be large, but as $A \approx 1 + 2u^{-4}$ for $u \gtrsim 2$, so the magnification rapidly becomes negligible at large u. For uniform motions, we will have $u(t) = \{u_{\min}^2 + [v_\perp(t-t_0)/r_E]^2\}^{1/2}$ where $v_\perp$ is the transverse velocity of the lens relative to the line of sight and $u_{\min}$ is the value of u at closest approach, which occurs at time t_0. Substituting this into equation (1) gives the lightcurve $A(t)$; some examples are shown in figure 2.

Observations of microlensing

To search for microlensing by MACHOs, since it is so rare we need millions of source stars and they should be far enough away to give a good path length for lensing, but not so far away that the individual stars are faint. The obvious targets for dark-matter lenses are the LARGE MAGELLANIC CLOUD (LMC) and SMALL MAGELLANIC CLOUD (SMC), the largest of our Milky Way's many satellite galaxies, since the sight line to them passes mainly through the Galactic halo. The Galactic bulge is also an interesting target, although here because of the high stellar density the microlensing should be mainly due to intervening faint stars rather than dark matter. Around 1991, several teams of astronomers called MACHO, EROS and OGLE began large projects using wide-field 1 m class telescopes to search for microlensing events. The novel feature of these experiments is their massive data volume, requiring large amounts of disk and tape storage and special high-speed photometry programs. The expected number of microlensing events is much smaller than the fraction of intrinsic variable stars, but fortunately microlensing has many strong 'signatures' that are distinct from all previously known types of variable star: the most important is the 'uniqueness', i.e. any given star should be microlensed at most once in a human lifetime, while almost all variable stars are periodic or quasi-periodic. Also, most events should have a symmetrical shape as in figure 2, and they should be achromatic (the source's color should not change during the event), since lensing is independent of wavelength.

The MACHO, EROS and OGLE teams announced their first candidate microlensing events near simultaneously in late 1993 (figure 3). Since then, they have accumulated much more extensive datasets, and completed detailed calculations of their microlensing detection efficiency as a function of event duration; thus a number of general results have become clear.

(i) Hundreds of events have been observed towards the Galactic bulge. They are distributed 'randomly' across the color–magnitude diagram, and their distribution of peak magnifications matches the theoretical prediction, which provides a very convincing check that microlensing works as predicted and that the experiments can detect it. The event rate towards the bulge is greater than initially predicted by Galactic models, favoring a barred structure for the Galaxy. The distribution of timescales is consistent with most of the lenses being low-mass stars as expected and can provide constraints on the stellar mass function at low masses.

(ii) Towards the Magellanic Clouds, no 'short' events (timescales from a few hours up to 20 days) have been seen by any group. This places strong limits on 'Jupiters' in the dark halo: specifically, compact objects in the mass range 10^{-6}–0.05 solar masses contribute less than 10% of the dark matter around our Galaxy. This is a very important result, as these objects were previously thought to be the most plausible form of baryonic dark matter, and (for masses below 0.01 solar masses) they would have been virtually impossible to detect directly.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

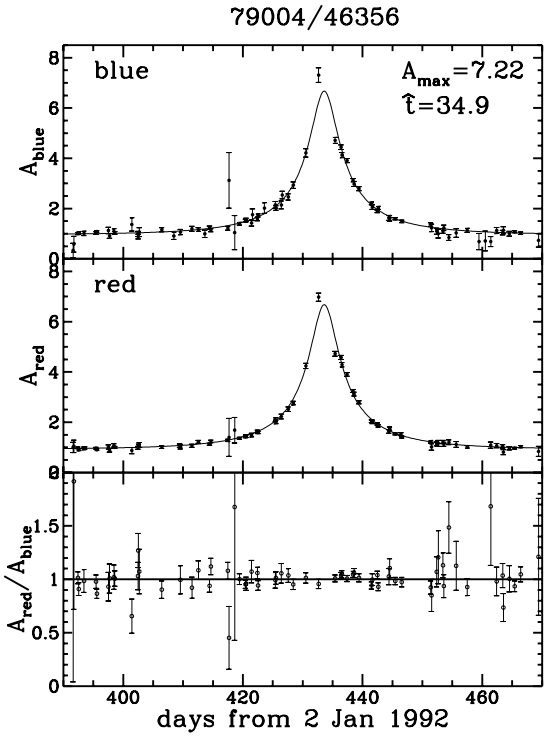


Figure 3. The first LMC microlensing candidate from the MACHO project. (Expanded view: 6 yr of constant data are outside the plot). Upper and middle panels show brightness versus time in blue and red passbands respectively, in units of the baseline value. Points with error bars are observations, the curve is the best microlensing fit. The lower panel shows the ratio of red/blue flux, illustrating the lack of color change.

(iii) Approximately 18 events with durations 30–200 days have been observed towards the Magellanic Clouds. The implied optical depth is well above expectation from 'known' stars, and is roughly 1/3 of the value expected ($\tau \approx 5 \times 10^{-7}$) for an all-MACHO dark halo. The relatively long durations imply lens masses of roughly 0.3–0.7 solar mass, which is a puzzle as discussed below.

(iv) Many events are now detected in real time and announced on the world-wide web. This enables more frequent monitoring of ongoing events, giving more precise tests of the shape. This has attracted considerable interest since, where the lens is a low-mass star (probably the case for most of the events towards the Galactic bulge), a planet around the lens star may give rise to a short-duration 'blip' superimposed on the smooth microlensing lightcurve: this method is potentially sensitive to low-mass planets down to a few Earth masses, below the range accessible to radial-velocity planet searches. Two groups called PLANET and MPS are now searching for this effect.

Current questions in microlensing

The nature of the objects giving rise to the lensing towards the Magellanic Clouds is a significant puzzle at present; the inferred lens masses are well above $0.1M_{\odot}$, which means that they cannot be ordinary hydrogen-burning stars in a spherical halo, as these would be visible in large numbers in deep images, e.g. the Hubble Deep Field. There are two classes of solution: either the lenses are in the halo but are much fainter, e.g. old white dwarfs or possibly primordial black holes, or they are low-mass stars but in a non-halo population, i.e. preferentially concentrated in the foreground of the LMC. The white-dwarf solution has significant problems associated with early metal production but has received some support recently from a possible detection of proper motions of faint objects in the HUBBLE DEEP FIELD. Various classes of non-halo population have been proposed, e.g. a 'thick' LMC disk, a small dwarf galaxy projected in front of the LMC, a warped Milky Way disk etc, but several searches for such populations have been negative.

There are several ways of testing these possibilities: one is to measure distances to individual lenses. In a 'standard' microlensing event as in figure 2 it is impossible to measure the lens distance since the one physical parameter (event duration) depends on three unknowns, the lens mass, distance and transverse velocity. However, as proposed by Gould and others, a small fraction of observed events should deviate from the standard shape for one of several reasons: the non-uniform motion of the Earth, the finite size of the source, a binary lens etc. In these cases we can obtain an additional observable which constrains the location of the lens. At present there are only one or two such cases for Magellanic Cloud lensing. Future satellite observations may help to measure lens distances, either by measuring the 'parallax' effect or the small centroid shift during the event.

Another active area is the search for microlensing towards the ANDROMEDA GALAXY, M31. This is considerably more challenging since the greater distance means that there are multiple unresolved stars per resolution element, and sophisticated image differencing techniques are necessary. However, it has the advantage that microlensing from M31's own halo will produce substantially more events towards the far side of M31's disk than towards the near side, owing to the nearly edge-on inclination. Two groups called AGAPE and MEGA have done pilot studies and are currently undertaking large-scale searches towards M31.

Cosmological microlensing

Although microlensing within the LOCAL GROUP is the main application at present, microlensing of 'point' sources at cosmological distances (e.g. quasars, supernovae or gamma-ray bursts) may also be observable. In the case of a quasar lensed into multiple images by a foreground galaxy, the individual stars in the galaxy may act as microlenses. Here the microlensing appears as brightness changes unique to one image of the multiplet (after correcting for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the time delay); in this case the optical depth is typically close to unity (because the surface density must be of order the 'critical' surface density to produce multiple imaging), so the interpretation of the lightcurve is complex. Other searches are in progress for microlensing of quasars behind the VIRGO CLUSTER by dark objects in the cluster, and the Next Generation Space Telescope should be easily able to detect microlensing of stars in the Virgo galaxy M87.

Bibliography

The first candidate microlensing events from MACHO and EROS appeared back to back in

Alcock *et al* 1993 *Nature* **365** 621
Aubourg *et al* 1993 *Nature* **365** 623

A large review of Local Group microlensing is given by

Paczynski B 1996 *Ann. Rev. Astron. Astrophys.* **34** 419

A shorter review concentrating on recent results from the MACHO team is by

Sutherland W 1999 *Rev. Mod. Phys.* **71** 421

An example of the current state of the art, measuring a lens distance with very detailed lightcurve data is in

Alcock C *et al* 1999 *Astrophys. J.* **518** 44

Searches on NASA ADS or xxx.lanl.gov for the various team acronyms should locate recent papers.

Will Sutherland

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Micrometeorite

An interplanetary dust particle or a particle ablated from a meteoroid during its passage through the Earth's atmosphere that has not itself burnt up in the atmosphere as a meteor. Particles below a mass of around 1 μg are slowed by collisions with molecules in the upper atmosphere to a speed at which further collisions impart insufficient energy for ablation to occur. Such particles drift down and eventually settle on the surface. Micrometeorites have been recovered from cores of sediments drilled from the ocean floor and from the Antarctic ice, and from rainwater. They can also be collected before they reach the ground by high-flying airplanes. Interplanetary dust particles recovered from the atmosphere are sometimes called *Brownlee particles* after Donald E Brownlee, who was the first to investigate them systematically.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Microscopium

(the Microscope; abbrev. Mic, gen. Microscopii; area 210 sq. deg.) A southern constellation which lies between Sagittarius and Piscis Austrinus/Grus, and culminates at midnight in early August. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2.

A small, inconspicuous constellation, the brightest stars in Microscopium are γ and ε Microscopii, both of magnitude 4.7. Interesting objects include α Microscopii, magnitude 4.9, which has a fainter companion, magnitude 9.8, separation 20.5″, and NGC 6925, an eleventh-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mid-Infrared Instrumentation

These astronomical instruments are designed to sense radiation at mid-infrared (also known as thermal) wavelengths (5–30 μm). Mid-infrared instruments operating from ground-based telescopes are limited in sensitivity by the large thermal background from the sky and also the telescope itself but enjoy the spatial resolution benefits provided by large ground-based telescopes. Space-based mid-infrared instruments are not limited by the atmospheric background but can suffer from lack of spatial resolution owing to the smaller telescope apertures available to these instruments.

The mid-infrared spectral region

The infrared (IR) region of the radiation spectrum covers the wavelengths longward of the visible region ($\sim$0.7 μm) and short of the microwave region. The infrared region is further subdivided into three regions: the near-infrared (0.7–5 μm) the mid-infrared (5–30 μm) and the far infrared (30–few hundred μm). In some applications (particularly military infrared devices) the infrared region is divided into the short-wave infrared (SWIR, 1–2 μm), mid-wave infrared (MWIR, 3–5 μm) and long-wave infrared (LWIR, 8–12 μm).

The mid-infrared region is largely synonymous with the thermal infrared, meaning the spectral region where thermal radiation from room temperature type bodies (around 300 K) dominates. The thermal radiation is the radiation emitted by all bodies, because of their own heat, and is characterized by the Planck function. For objects of temperatures around 300 K, the peak of the black body curve (see BLACK-BODY RADIATION) lies at 10 μm in wavelength.

Atmospheric bands and transmission

The transparency of the atmosphere defines a number of bands. This is because atmospheric gases, especially H_2O and CO_2, are efficient absorbers of infrared radiation at certain wavelengths. These absorption bands define 'windows' of relative transparency through the atmosphere. In the mid-infrared region these bands lie at 3–5 μm, 8–13 μm and 16–22 μm, although the 8–13 μm band is by far the 'cleanest', i.e. most free from extra absorption bands. From a high, dry mountain observatory the best part of the 8–13 μm window, near 10.5 μm is about 95% transparent. At the edges of the window it is closer to 50–70%, before falling to 0% transmission outside of the windows.

Thermal backgrounds

At mid-infrared wavelengths, all terrestrial bodies are radiating copiously. This is true of the sky and also the telescope and instrument itself. The fact that the sky is 50–95% transmissive in the atmospheric windows also means that the sky is between 50% and 5% emissive. Similarly, the aluminized mirrors of the telescope emit between 1% and 5% of a black body at the telescope temperature. The dirtier the mirror, the larger the emission. The spider vanes and any part of the telescope structure 'visible' to the instrument radiate as near-perfect black bodies. All these sources combine to provide a large background flux of infrared photons on the infrared detector. The photon shot noise arising from this background radiation sets the background limit to the detector sensitivity. Except at very high resolutions the background 'noise' dominates all other sources of noise such as detector read noise and dark current.

To reduce the background emission from the telescope itself, INFRARED TELESCOPES typically have no optical baffling (which radiates at thermal wavelengths), they maintain a low emissivity mirror surface by regular cleaning and aluminizing, and employ an undersized secondary mirror so that no part of the telescope structure is visible to the infrared detector.

For a typical broad-band imaging situation, using a 4 m class telescope, clear and cold sky, and a pixel field of view of 1 arcsec, then the background from sky and telescope is around 10^9–10^{10} photoelectrons s^{-1}. The astronomical source, however, will be many orders of magnitude (around 10^{-5}) fainter than the background. Extraction of the signal from the background and the background photon shot noise is the key problem for ground-based mid-INFRARED ASTRONOMY.

Observatories

To minimize the sky background, the best mid-infrared observing sites are located atop high, dry mountains (see also OBSERVATORIES AND THEIR LOCATION). Probably the best mid-infrared site in the world is the Mauna Kea observatory in Hawaii, at 4200 m elevation. Here the telescope is above a significant fraction of the Earth's atmosphere and IR-absorbing gases, and in particular it is above 90% of the atmospheric water vapor. The reduced atmospheric temperature and background and improved transmission mean that system sensitivities are significantly improved. Comparable sites are also found in the Atacama desert in Chile, although they are not quite as high.

Chopping

With the advent of large-format, staring detectors at near-infrared wavelengths the practice of chopping has largely disappeared from such instruments, but at mid-infrared wavelengths the background is sufficiently large and the background due to the sky is sufficiently variable that 'sky chopping' is still generally necessary. Chopping provides a form of background subtraction that is very effective and particularly removes the effect of high-frequency fluctuations in the sky background. These fluctuations generally take the form of turbulence which presents regions of slightly different temperature and/or transparency to the detector on very short timescales (i.e. subhertz rates).

Chopping involves wobbling the secondary mirror of the telescope to present alternate views to the detector,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

firstly with source plus background, and then with background alone. The difference of these two images leaves source only.

As the background flux compared with the source is reduced, the requirement for chopping diminishes. So for observations involving very high spectral resolutions, or where the sky is very clear and temperatures low (e.g. South Pole or space), background chopping is usually not required.

Diffraction limit and seeing

At mid-infrared wavelengths the spatial resolution we can obtain with current-generation telescopes (i.e. up to 10 m diameter) is limited not so much by atmospheric SEEING as in the visible and near-infrared regions but rather by the diffraction limit of the telescope itself. The diffraction limit at 10 μm for a 4 m telescope is on the order of half an arcsecond. For most instruments the diffraction limit determines the pixel scale so as to just sample the point spread function (PSF). This means that the point source sensitivity increases as the square of the aperture (D^2) in the critically sampled diffraction limited case, but in the seeing limited case increases as the diameter (D) only.

At 10 μm the seeing is about one-half of that at visible wavelengths. At the best observing sites the median seeing is typically half an arcsecond in the visible, or one-quarter arcsecond in the mid-infrared. So, except for the largest telescopes or the lesser observing sites, the atmospheric seeing is usually dominated by the diffraction limit.

Detector sensitivity

It is very difficult to quote system sensitivities, because the actual sensitivity achieved depends so much on the operating configurations (spectral resolution, pixel size, detector type, telescope size and emissivity, atmospheric background and temperature, weather etc). If we adopt a fairly standard imaging format, however, it is possible to provide some indicative numbers, from which different situations can be scaled. Using a 4 m IR optimized telescope at a good observatory site, a good detector with quantum efficiency around 25%, a pixel size of 0.25 arcsec^2 and a broad-band filter 8–13 μm then a sensitivity figure of around 1–2 mJy arcsec^{-2} $(1\sigma)^{-1}$ h^{-1} should be achievable (the flux density unit Jansky is defined by 1 Jy $= 10^{-26}$ W m^{-2} Hz^{-1}). Sensitivity limits for various mid-infrared instruments available at large observatories are usually available at the observatory web site.

Detectors for use in the mid-infrared

Si photon detectors

The majority of detectors used for mid-infrared astronomy fall into the class of doped-silicon photon detectors. In these devices an incident photon causes an electron to shift from the valence band to the conduction band of the semiconductor, leaving a mobile electron in the conduction band and hole in the valence band. The electron–hole pair can then be sensed using a suitable bias circuit. The efficiency with which the detector layer converts incident photons to photoelectrons is called the detector quantum efficiency (η). The quantum efficiency is dependent on the detector material, its doping and also thickness, and is typically in the range 10–25%.

The drawback with these detectors is that the band-gap energy is low enough that thermally generated electrons (dark current) in the detector itself are excited into the conduction band. At room temperature the dark current would swamp photoelectrons, and so the detector has to be cooled. The longer the wavelength to which the detector needs to operate the colder it needs to be, and temperatures of 4–12 K are typical.

As in the visible and near-infrared wavelength regions, mid-infrared astronomy is being revolutionized by the advent of large format staring focal plane arrays (FPAs). The background from the sky and telescope is large enough that detectors, even with well depths of 10^6–10^7 electrons, still need to be read out at frame rates between 200 and 400 frames s^{-1} for broad-band imaging. These high readout rates and the low required operating temperatures have served to limit detector sizes to just a few hundred pixels square. The most common detectors today have 128 × 128 pixels, generally with between two and four readout channels. Currently under development, or shipping in limited quantities for testing, are 256 × 256 element and 320 × 240 element arrays.

As of today there are only two major manufacturers of astronomy-grade large-format mid-INFRARED ARRAYS, Boeing (until recently made by Rockwell International) and the Santa Barbara Research Center (SBRC, formerly owned by Hughes, but recently acquired by Raytheon). A small number of detectors are also made by LETI of France, Amber Engineering, USA, Aerojet, USA, and some others, but these detectors are rarely competitive for astronomical applications compared with the SBRC and Boeing offerings.

Bolometers

Bolometers are devices that respond to changes in heat, rather than directly detecting the photoelectron. To do this the bolometer has a load that absorbs the incident radiation and a sensing circuit that detects the change in a physical property of the load (e.g. conductivity) as the temperature changes. These detectors have the advantage of operating over a very wide wavelength range but the response times of these detectors tends to be slow. In principle these devices do not require cooling to operate (cf photon detectors) although in practice some cooling is almost always employed in astronomical applications to improve the sensitivity. In spite of some difficulties these detectors can be extremely sensitive and were the mainstay of mid-infrared astronomy before the development of photon detectors and focal plane arrays.

Instruments

While instruments for use at mid-infrared wavelengths are varied they all have some common aspects. Most mid-infrared instruments cool the detector to temperatures

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

between 4 and 20 K. In addition, however, most instruments also cool their complete optical assembly to reduce background emission from the optics themselves. This is taken to the extreme for space- and airborne instruments where the telescope itself is also cooled. Traditionally these instruments were cooled with liquid nitrogen (77 K) and liquid helium (4.2 K). Lately, however, two-stage closed-cycle coolers have become more readily available and for many applications have become the cooling system of choice.

Mid-infrared instruments also use reflective optics in preference to refractive optics wherever possible. This is because reflective optics are achromatic and have high reflectivities compared with the transmission of refractive optics. This is not to say that refractive optics are never used, but reflective are preferable.

In the past mid-infrared instruments have usually been single-purpose instruments for imaging (see INFRARED IMAGERS), low-resolution spectroscopy, high-resolution spectroscopy or polarimetry. As the large observatories have begun to provide facility mid-infrared instruments, however, very sophisticated multifunction instruments have been developed. The most ambitious of these programs are probably the VISIR instrument for the European Southern Observatory, Very large Telescope (ESO VLT 8 m) and Michelle for the United Kingdom Infrared Telescope (UKIRT 4 m) which will later be used on the Gemini 8 m North telescope. Each of these instruments combines broad- and narrow-band imaging with low-, medium- and high-resolution echelle spectroscopy. In each case these instruments also provide polarimetry capabilities.

Space-based and airborne mid-infrared instruments

Ground-based infrared astronomy is severely limited by the background emission produced by the atmosphere. These problems can be completely circumvented by placing the telescope and instrument in Earth orbit. If the telescope itself is cooled, then the sensitivity of the system is limited only by the size of the telescope that can be launched into orbit and the natural scene background. For space-based instruments the sensitivity limit is generally provided by the emission from small interplanetary dust grains heated by the Sun in our own solar system. To achieve the ultimate in sensitivity one needs to place a mid-infrared instrument beyond the orbit of Jupiter where the emission from interplanetary dust grains becomes negligible. Without the limitation of the atmosphere, mid-infrared instruments in space are not restricted to the atmospheric bands as are instruments on the ground, and whole new regions of the spectrum become available for study.

Major contributions to the progress of mid-infrared astronomy were made by the INFRARED ASTRONOMY SATELLITE (IRAS) launched in 1983 and the Infrared Space Observatory (ISO) launched in 1995. IRAS performed an all-sky survey at mid–far-infrared wavelengths while ISO acted as a space observatory facility providing imaging and spectroscopic capability from around 2.5 μm to 240 μm. The next major space facility will be the SPACE INFRARED TELESCOPE FACILITY (SIRTF), currently scheduled for launch in 2001. The Next Generation Space Telescope (NGST) scheduled for launch in 2007 may also carry some mid-infrared instruments (see SPACE INSTRUMENTATION: NEXT GENERATION SPACE TELESCOPE).

One problem for space-based mid-infrared astronomy is that the need to cool instruments (using liquid cryogens) means that mission lifetimes tend to be limited to between 1 and 2 yr depending on the amount of cryogen carried. The limited launch weight, size and cooling requirements limit the aperture size of the infrared telescope (IRAS and ISO both used 60 cm class telescopes). The smaller telescope limits light-gathering power, but in addition the diffraction limit of such telescopes is large (3–4 arcsec at 10 μm and 35 arcsec at 100 μm). This means that the spatial resolution that can be achieved with such instruments is comparatively low. So space-based and ground-based instruments will always be complementary: space-based observations provide access to the full spectrum, without the troublesome effects of the atmosphere, whereas ground-based instruments provide spatial resolution and sometimes sensitivity owing to their large apertures, but in their limited operating regions.

As a compromise between ground- and space-based mid-infrared astronomy, there is the option to place the infrared telescope at high altitude. This can be done by balloon (up to 100 000 ft) or aircraft which reach up to 40 000 ft. At these altitudes the environment is much like that for space instruments, but at a lower cost and with more versatility. Although balloon-borne instruments were common in the early days of IR astronomy, the cost, short flight duration and high risk of such ventures has seen a decline in their use in favour of aircraft. The NASA Kuiper Airborne Observatory (KAO) was a 90 cm cryogenically cooled telescope mounted in a Lockheed C141 Starlifter. It operated very successfully with a variety of mid- and far-infrared instruments until 1995 when the aircraft was decommissioned. The KAO is soon to be replaced by the Stratospheric Observatory for Infrared Astronomy (SOFIA), which will carry a 2.5 m telescope aboard a Boeing 747 jet.

Bibliography

Crawford D and Craine R 1994 Instrumentation in astronomy VIII *SPIE* **2198**

Fowler A 1993 Infrared detectors and instrumentation *SPIE* **1946**

Fowler A 1998 Infrared astronomical instrumentation *SPIE* **3354**

Craig H Smith

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Milankovitch, [Milankovič] Milutin (1879–1958)

Yugoslav mathematician and astronomer who, starting in 1912, showed how the radiation received by the Earth varies cyclically as the Earth's orbit varies. There are three significant *Milankovitch cycles*: the eccentricity of the Earth's orbit varies from more elliptical to more circular with a period of the order of 100 000 years; the obliquity of the Earth's axis oscillates between two extremes with a period of 41 000 years; and the spin axis precesses with a 23 000 year cycle. Palaeoclimatic oscillations with these periods have been discovered in data from deep sea cores dating back 6 million years, which measured global ice volume (indicated by oxygen isotopes in fossil organisms) and fluctuating productivity in surface waters (indicated by calcium carbonate sedimentation), and which thus correlate with the Ice Ages. Although Milankovitch cycles are named after Milankovitch, his ideas were pre-dated by Joseph Adhemar (1842) and James Croll (1875).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Milky Way Galaxy

The galaxy (i.e. assemblage of stars, gas and dust) of which the Sun is a member. Also known as the Galaxy, it is a disk-shaped system, with a diameter of between 80 000 and 100 000 light-years and a thickness of about 2000 light-years, that contains more than a hundred billion (10^{11}) stars. The core, or nucleus, of the Galaxy is surrounded by an ellipsoidal central bulge, containing relatively closely spaced stars, that measures some 15 000 light-years in diameter and about 6000 light-years in the direction perpendicular to the plane of the disk. Surrounding the bulge and extending in a near spherical distribution above and below the galactic plane (the plane of the disk) is the galactic halo. The halo contains about 200 globular clusters (near-spherical clusters that contain between 10 000 and 1 000 000 stars) and an extremely thinly scattered population of individual stars.

The Sun is located just over half way out from the center to the edge of the disk, estimates of its distance from the center ranging from about 22 000 to about 29 000 light-years, with 25 000 light-years being a reasonable working value. In common with other stars, the Sun revolves round the galactic center. The orbital velocity of the Sun is about 220 km s^{-1} and its orbital period is about 225 million years. Overall, the Galaxy exhibits differential rotation, stars and gas clouds closer to the center having shorter orbital periods than those that are located further out.

Stars in the Galaxy are divided into two principal categories, or populations, Population II stars being old stars that formed early in the history of the Galaxy and population I stars being second, or later, generation stars that formed from gas clouds that had been seeded with heavier elements generated in, and expelled from, earlier generations of stars. The halo is composed of Population II objects and the disk, where clouds of gas and dust are concentrated and where ongoing star formation is taking place, is dominated by Population I objects. The central bulge contains both populations but its light output is dominated by radiation from old red giants.

Clouds of dust in the galactic plane absorb starlight and prevent optical astronomers from seeing the galactic center, but infrared and radio wavelengths are largely unaffected. Near-infrared observations have shown that the stars in the central bulge are arranged in an elongated bar, about twice as long as it is wide, that is seen nearly end on from the present location of the solar system. The exact center, or nucleus, of the Galaxy coincides with a strong source of radio emission, called Sagittarius A*, that is less than 15 astronomical units in diameter (i.e. is smaller than the diameter of the orbit of Saturn). Observations of the speeds at which clouds of ionized gas are revolving round the galactic center imply that several million solar masses of material are concentrated within a region of about one light-year in radius. Since only about half of this mass can be accounted for by stars, it seems likely that the balance (about 2.5 million solar masses) is contained in a central black hole and that accretion onto this black hole is the underlying source of the energy radiated by Sagittarius A*.

In 2001, NASA's Chandra X-ray Observatory, detected a sudden x-ray flare while observing Sgr A*. In a few minutes, the source brightened dramatically, eventually reaching a level 45 times brighter than before the flare. After about three hours, the x-ray intensity rapidly declined to the pre-flare level. The rapid rise and fall of the x-rays from this outburst are compelling evidence that the x-ray emission is coming from matter falling into a supermassive black hole. Given the extremely accurate position, it is highly unlikely that the flare is due to an unrelated contaminating source such as an x-ray binary system.

Like many other galaxies, the Milky Way Galaxy has a spiral structure. Within its disk, bright young stars, nebulae, and clouds of gas and dust are concentrated into curved 'arms' that appear to radiate from the central bulge in a spiral pattern. The Galaxy's spiral pattern consists of several major arms and a number of shorter segments, one of which—the Orion arm, or 'spur'—contains the Sun and the Orion star-forming region.

The Milky Way Galaxy derives its name from the Milky Way, a misty band of starlight that stretches across the sky and which may be seen with the unaided eye on a clear dark night. The Milky Way represents the combined light of millions upon millions of stars concentrated close to the galactic plane.

See also: barred spiral galaxy, galaxy, galaxy: structure, halo, galactic, Hubble classification, galactic nucleus, Orion arm, spiral arms, spiral galaxy.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Millikan, Robert Andrews (1868–1953)

Physicist, born in Morrison, IL, Nobel prizewinner (1923) for his determination of the charge on the electron by *Millikan's oil-drop experiment*, and confirmation of Einstein's quantum photoelectric theory. Coined the term 'cosmic rays'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Millimeter and Submillimeter Astronomy

The millimeter wavelength band extends from $\lambda = 4$ mm ($\nu = 75$ GHz) to 1 mm (300 GHz) while the submillimeter band extends from 1 mm to 100 μm (3 THz). The sub-mm band is rather unexplored because of the need to construct large efficient telescopes and sensitive detectors. It is equally important to select excellent sites to reduce the pervasive absorption by the Earth's atmosphere. To a lesser degree, these challenges also affect mm astronomy. Usually the emission from astronomical objects in the mm and sub-mm wavelength range is small compared with that in the meter and centimeter wavelength ranges where intense non-thermal synchrotron emission is common or in the mid- and near-infrared or optical wavelength ranges, where thermal radiation from hot sources is found. Studies in the mm and sub-mm ranges cover a wide range of topics: (1) the physics and chemistry of the interstellar medium (ISM), especially the structure and content of INTERSTELLAR MOLECULAR CLOUDS, (2) star formation, (3) the interaction of stars with molecular clouds, especially during the very early and very late phases of stellar evolution, (4) the distribution of molecular clouds (and neutral carbon) in external galaxies and high-redshift objects and (5) studies of solar system objects. Investigations of the 2.73 K background are also important but specialized.

Spectral line radiation from molecular hydrogen is rare since H_2 has no permanent dipole moment. Thus, extraordinary excitation conditions are needed. The properties of molecular clouds are traced using abundant polar molecules such as carbon monoxide, CO. In the mm and sub-mm wavelength ranges, small linear molecules have rotational transitions, while atomic and ionized carbon have fine-structure lines. In addition to gas, clouds contain dust grains which contain much of the silicon and carbon. Dust grains emit broadband thermal emission which increases at shorter wavelengths where these grains radiate more efficiently. The surfaces of cold grains are covered with condensed atoms and molecules.

There is a sudden transition from neutral atomic hydrogen, H I to H_2 from competition between destruction by the interstellar radiation field and production on grain surfaces. The envelopes of molecular clouds consist of neutral hydrogen, H I, and ionized or neutral carbon, with molecules in cloud cores.

A short history

Observations started with broadband continuum measurements of emission from the Sun, planets and a few quasi-stellar objects. Molecular spectroscopy provided a great impetus for mm and sub-mm astronomy. Radio spectroscopy of molecules began with the detection of the OH radical at a wavelength of $\lambda = 18$ cm in 1963. These molecules were thought to exist in H I REGIONS. In 1968 the detection of ammonia, NH_3, and water vapor, both at 1.3 cm, was made with the Berkeley 6 m telescope. In 1969, formaldehyde, H_2CO, was discovered with the National Radio Astronomy Observatory's (NRAO's) 140 ft telescope at 6 cm. The structures of H_2CO and NH_3 are rather complex, and the abundances are rather small. The most widespread polar molecule, CO, was detected at 2.6 mm in 1970 with the NRAO's 36 ft telescope. After the discovery of CO, there was a rapid increase in the number of interstellar molecules, because radiation processes are more efficient at shorter wavelengths and because there are a larger number of spectral lines in the millimeter and submillimeter wavelength ranges. As of 1999, 120 molecules have been found in the ISM; the heaviest molecule is $HC_{11}N$.

Technical considerations

From diffraction theory, the angular resolution, θ, is

$$\theta = 1.2\lambda/D$$

where D is the telescope size and λ the wavelength. For a 10 m diameter telescope, at $\lambda = 3$ mm, $\theta = 80''$. Intensities are expressed in terms of temperatures, in the Rayleigh–Jeans limit:

$$T = \frac{\lambda^2 I_\nu}{2k} \tag{1}$$

where k is Boltzmann's constant. I_ν is the intensity at frequency ν in units of W Hz^{-1} m^{-2} sr^{-1} and T is the 'brightness temperature'. For thermal emission from planets, T is the surface temperature diluted by the ratio of source to beam size broadened by the source. In most cases, T is not related to a physical source temperature. In terms of I, the flux density is

$$S_\nu = \int I_\nu \, \mathrm{d}\Omega. \tag{2}$$

The solid-angle integration is usually over the beam; if the source size exceeds the telescope beam size, the integration is over the source. Combining equations (1) and (2), we have

$$S_\nu = \frac{2k}{c^2} T \nu^2 \, \Delta\Omega. \tag{3}$$

The power received is

$$P_\nu = \tfrac{1}{2} S_\nu A_{\mathrm{eff}} B \tag{4}$$

where A_{eff} is the effective area and B is the bandwidth.

Receivers

For $\lambda > 0.3$ mm, the most efficient receivers for spectral lines are heterodyne systems. In such coherent receivers, the signal is converted from the sky wavelength to a longer wavelength at which the signal is amplified. Such systems have a wavelength resolution, $\lambda/\Delta\lambda \gg 10^6$. In the 1960s, receiver noise temperatures, T_{rx}, were $\sim$5000 K at 3 mm; now, typical values are <100 K. The ultimate limit for noise in a coherent receiver is $h\nu/k$; practical limits are a few times larger. The first conversion is usually performed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in a mixer. The first mixers were uncooled Schottky devices. Since the 1980s, mixers have junctions which are quantum mechanical superconducting–insulating–superconducting (SIS) sandwiches. SIS mixers are much more sensitive and require less local oscillator power than classical Schottky devices; SIS mixers must be cooled to 4 K. All such receivers are tunable over a large wavelength range. The development of SIS RECEIVERS in the millimeter and sub-mm range has been led by the Caltech Sub-mm Observatory (CSO) and in the millimeter range by NRAO. Other active SIS development groups are at Harvard-Smithsonian Center for Astrophysics (CfA), IRAM and Köln University. From the physics of junction materials, there is a short-wavelength limit to the operation of SIS devices. For spectral line measurements at wavelengths at $\lambda < 0.3$ mm, a new device, the hot-electron bolometer, which has no such limit, has been developed. The first successful astronomical measurements with HEBs were carried out by the CfA group on the Heinrich Hertz 10 m Sub-mm Telescope (HHT) in 1998. At $\lambda > 3$ mm high electron mobility transistors are becoming standard, while, at $\lambda > 2.6$ mm, monolithic indium phosphide microwave transistor amplifiers are now being used on the 16-beam receiver array (SEQUOIA) installed on the University of Massachusetts 14 m telescope. Only SEQUOIA and the eight-beam SIS array installed on the NRAO 12 m telescope have thus far produced published results; all other millimeter heterodyne systems with competitive noise temperatures are single-beam devices.

Bolometers are incoherent receivers; these are preferred for single-dish measurements of broadband continuum radiation at $\lambda < 2$ mm. As with heterodyne receivers, the sensitivity of bolometers has improved greatly since the 1960s. Bolometer sensitivity is given as noise equivalent power (NEP) in W $Hz^{-1/2}$. For practical bolometer systems, background noise, including noise from warm telescope surfaces, the ground and atmosphere, is the limiting factor. A good value of NEP is 3×10^{-17} W $Hz^{-1/2}$. For $\lambda < 0.3$ mm, a combination of bolometer detectors and Fabry–Pérot filters can be effectively used for spectroscopy. In contrast to heterodyne systems, a number of bolometer array receivers are installed on telescopes. The largest is SCUBA, an 81-beam system on the James Clerk Maxwell 15 m Telescope (JCMT) in Hawaii.

The RMS receiver noise in kelvins is

$$\Delta T_{\rm RMS} = \frac{A T_{\rm sys}}{(Bt)^{1/2}} = \frac{A\,{\rm NEP}}{2kBt^{1/2}}. \qquad (5)$$

The constant A is a function of system and data-taking parameters (usually $A \sim 2$), B is the bandwidth in Hz and t is integration time in s. The value of B is limited to one-third of the linewidth for spectral line measurements. For continuum measurements, B can be much larger. The value used for $T_{\rm sys}$ is corrected for absorption of the signal in the Earth's atmosphere and the emission from the Earth's atmosphere and telescope losses. Depending on weather, the value of $T_{\rm sys}$ can exceed the receiver noise temperature, $T_{\rm rx}$, by large factors.

Telescopes

For a RADIO TELESCOPE to be effective, the surface accuracy must be <0.1 of the shortest wavelength measured. Larger radio telescopes have surfaces consisting of individual panels which must be adjusted. The limit to the effective operation of all single telescopes is set by surface deformations and pointing errors. Some of the mm telescopes, with diameters of 14 m and one with 20 m, were built in radomes. The two largest millimeter telescopes, the 30 m, operated by IRAM, and the 45 m in Nobeyama, Japan, are in the open, as will be the 50 m 'Large Millimeter Telescope' (LMT) being constructed near Mexico City. The LMT is designed to be the most advanced single-dish design, useful up to 345 GHz. The construction of open-air telescopes is in part economic; enclosures could cost as much as the telescope itself. The most important improvement in the design of large single dishes has been homology, a scheme in which a paraboloid freely deforms but the shape is maintained. The remaining systematic effects are the influence of wind and temperature gradients in the structure which give rise to differential thermal expansion in the structure. The first radio large telescopes built with high accuracy surfaces of low-expansion material, carbon-fiber-reinforced plastic, were the IRAM 15 m interferometer antennas on Plateau de Bure, France and the 15 m Swedish-ESO-Sub-mm-Telescope (SEST) telescope, at La Silla, Chile, and the HHT. The surface accuracy of the HHT shows that RMS surface accuracies of 12 μm are possible. The 10.4 m diameter telescope of the CSO and the 15 m JCMT are sub-mm telescopes on Mauna Kea, Hawaii, an excellent site. The CSO and JCMT were put into operation in the late 1980s and have produced a great deal of science. Both are constructed of aluminum and are housed in astronomical dome-like shelters.

Sub-mm measurements are also made from aircraft. The Kuiper Airborne Observatory provided such an opportunity with a 91 cm diameter paraboloid. The follow-up facility, SOFIA, will have a 2.5 m diameter mirror. At normal flight altitude, 14 km, the water vapor content is extremely low. Thus, SOFIA can be used for measurements at wavelengths at which the atmosphere is opaque from ground-based sites. The Sub-mm Wavelength Astronomy Satellite has detected emission from the ground-state transition of H_2O, at 0.55 mm, but has not detected emission from O_2. In the future are the Far-Infra-Red-Satellite-Telescope and Satellite-Infra-Red-Telescope-Facility.

Single telescopes have a fixed relation of $A_{\rm eff}$ and θ whereas interferometer systems do not (see equations (1)–(4)). Present interferometer systems consist of $n < 11$ elements, each of a moderate diameter. INTERFEROMETERS consist of a number of parabolic antennas, each equipped with one or more coherent receivers. The receiver outputs are combined to produce a response corresponding to the maximum angular resolution determined by the ratio of the wavelength to the largest separation of the elements, projected in the source direction. Emission

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

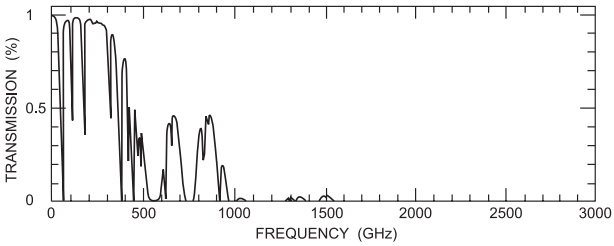


Figure 1. A plot of the transmission through the Earth's atmosphere for the mm and sub-mm wavelength ranges, for a column density of atmospheric water vapor of 1 mm. For an atmospheric optical depth τ, the transmission is $e^{-\tau}$. (Adapted from Phillips and Keene (1992).)

from structures with angular sizes larger than the angle corresponding to the closest separation of the antennas is not recorded. This is the 'smooth structure' in a source. Such information can be recovered by combining the interferometer output with data taken using a large single telescope. Interferometers provide higher angular resolutions and positional accuracies than single-dish telescopes. For a fixed collecting area bandwidth, and ν, the temperature (but not the flux density) uncertainty of an interferometer is worse than that for a single telescope with the same angular resolution. The first mm-wave interferometer, at the Berkeley Radio Astronomy Laboratory, consisted of two 6 m antennas. In the mm range, there are four systems: The Berkeley–Illinois–Maryland association (BIMA) array near Hat Creek, CA, the Caltech Owens Valley Interferometer (OVRO) near Bishop, CA, the Nobeyama Array in Japan and the IRAM array on Plateau de Bure, France. A number of high dynamic range, high-sensitivity images have been produced by the BIMA, OVRO and IRAM arrays. In addition, the 10 m CSO and 15 m JCMT telescopes are occasionally used as a sub-mm interferometer. To obtain even higher angular resolutions, experiments have been carried out using the technique of very long baseline interferometry at 3 mm and 1.3 mm. Finally, a dedicated sub-mm array consisting of eight 6 m dishes has been under construction by the Smithsonian Astrophysical Observatory and the Academica Sinica Institute for Astronomy and Astrophysics, Taiwan, on Mauna Kea, Hawaii. The sensitivity and imaging quality of interferometer systems can be improved by an array with many large antennas on an excellent site. This is the rationale of the Atacama Large Millimeter Array (ALMA) project, to be built in Chile at 5 km elevation, by NRAO and ESO. ALMA is planned to have $n > 60$ elements and a total $A_{\rm eff} > 10^4$ m^2. The results to be obtained with ALMA may revolutionize all of astronomy.

Atmospheric effects

Millimeter and especially sub-mm observing sites are selected to minimize the effect of weather on the observations. Signals are absorbed by O_2 and H_2O vapor. The scale height of H_2O vapor is $\sim$2 km, while that of O_2 is $\sim$8 km, so the influence of H_2O should be reduced at high altitudes. See figure 1 for a plot of transmissions. At 860 GHz ($\lambda = 0.35$ mm), even the very best sites on Earth are useful only a small percentage of the time.

Observing techniques

Because the signal levels are small compared with atmospheric or receiver noise, all measurements must be made by taking the difference of two signals. For spectral line measurements, one uses compared data taken at two different positions or frequencies; for continuum measurements there is a rapid switching by movements of the telescope subreflector or by wobbling a flat mirror. The time intervals are determined by instabilities in the receiver and atmosphere. A technique of slow scanning of the telescope is the normal data-taking mode for continuum measurements. For spectral line measurements this has been called 'on-the-fly' mapping. The data are taken rapidly, before weather or telescope characteristics can change. This greatly increases the reliability of spectral line images and efficiency of data taking.

'Chopper wheel' calibrations are the norm for spectral line measurements. In the simplest version of this method, measurements are made of a room-temperature absorber and the atmosphere; with models, these allow an estimate of the atmospheric emission and hence the absorption. A more complete characterization of the atmosphere requires a 'tipping curve', which involves measurements at a number of elevations. The data are corrected for the absorption and usually for telescope efficiencies to give intensities (usually in the form of temperatures) for a perfect telescope outside the atmosphere. Additional calibration steps involve measurements of sources of known intensity, then scaling the data taken. For interferometry, the measurements of a source and known source or 'calibrator' are interspersed in time. It is assumed that the system performance varies smoothly between calibrations and the corrections obtained are applied to the source measurements.

Astronomical considerations

Spectral lines

In figure 2 we show the energy level diagrams for CO, neutral carbon, C I, and ionized carbon, C II; the energy, E, above the ground state is in kelvins, from $E = kT$. Allowed dipole transitions occur with a unit change in the angular momentum quantum number $\Delta J = \pm 1$. For the CO molecule low-J transitions are frequently thermalized, so the kinetic temperature, $T_{\rm k}$, characterizes the populations. Measurements of CO allow good estimates of the column density of H_2, since CO can be excited at low H_2 densities, since the excitation is close to local thermodynamic equilibrium (LTE) and since there is a large abundance relative to H_2.

If the radiation arises from an isothermal layer much larger than the telescope beam, one can obtain a fairly realistic estimate of kinetic temperatures and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

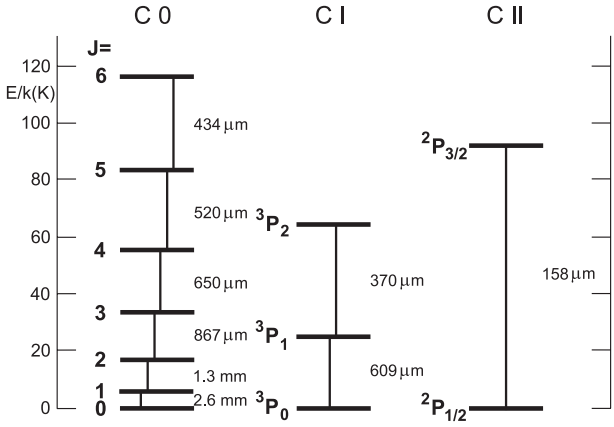


Figure 2. The first few rotational energy levels of selected species. In the ISM, carbon monoxide, CO, is the most abundant molecule with a permanent dipole moment. Also shown are fine-structure transitions of neutral carbon, C I, and ionized carbon, C II.

CO column densities using an LTE analysis. Rotational lines of $^{12}C^{16}O$ are usually optically thick; then peak line temperature T equals T_k, the kinetic temperature. To obtain column densities without models of molecular clouds, one must use rarer, stable isotopes of CO. In LTE, the total population, N, can be obtained from the (optically thin) $J = 1$–0 line of $^{13}C^{16}O$:

$$N(\text{total})^{^{13}\text{CO}} = 2.6 \times 10^{14} \frac{T \int \tau^{13}(v)\, \mathrm{d}v}{1 - \exp(-5.3/T)} \qquad (6)$$

where the column density $N(\text{total})^{^{13}\text{CO}}$ is in cm^{-2}, the linewidth v in km s^{-1} and T in K. If the line optical depth, $\tau^{13} \ll 1$, $T\tau^{13} = T_L^{13}$. If $\tau^{13} \gg 1$, we must use the line of $C^{18}O$; the expression is very nearly the same. Such estimates are usually realistic for cooler clouds ($T_k < 20$ K). To obtain the column density of H_2, one assumes an isotopic ratio and that the CO-to-H_2 ratio is 10^{-4}; the latter value is usually reasonable.

In addition to estimates of column densities, one can estimate local H_2 densities, $n(H_2)$, from the ratio of column density to cloud size to obtain the local density. This leads to a lower limit for $n(H_2)$ since clouds have small-scale structure or 'clumping'. If energy levels are populated by collisions with H_2 molecules, the density needed to balance decay out of a level is n^*. Setting the collision rate equal to the decay rate, we have $n^*\sigma\bar{v} = A$. If the optical depth of the line is $\tau \gg 1$, the analysis becomes more complex. In the Sobolev or large velocity gradient (LVG) approximation, one assumes that the molecular cloud has a velocity gradient $V \sim r$. In the LVG model, the value of the A coefficient is replaced by A/τ. For a multiline analysis of a species, including the fall-off above and below the maximum intensity, one can solve for both T_k and H_2 density (see figure 3).

From LVG calculations applied to optically thin transitions of carbon monoxide, there is a simple relation between total column density and the $J = 2$–1 line, which is valid for a wide range of kinetic temperatures and H_2 densities. Large-scale surveys of ^{13}CO and CO $J = 1 \rightarrow 0$ lines in the disk of our Galaxy show that line shapes and intensity ratios along lines of sight are remarkably similar. Since ^{13}CO lines are optically thin, the CO integrated line intensities must measure mass, even though $\tau^{12} \gg 1$. This can be explained if clouds are in virial equilibrium and consist of a large number of small subthermally excited clumps. An exact relation between the integrated intensity of the $J = 1 \rightarrow 0$ line and column density of H_2 has been determined empirically for the disk. This approach has been often used for molecular mass determinations in external galaxies. However, there are exceptions to this relation: the conversion factor for the galactic center is 1/7 of the factor used for the disk, and the conversion factor appears to differ from galaxy to galaxy.

Dust emission

Another method to determine H_2 column densities is to use broadband dust emission. The dust intensities rise rapidly with frequency, and measurements made with large bandwidths lead to very small values of ΔT_{RMS} (see equation (5)). One drawback to continuum measurements is line-of-sight confusion, since there is no velocity information. The quantitative relation between τ_{dust} and $N(H_2)$ is calibrated by the measurement of emission and extinction toward NGC 7023. The optical extinction has been converted into a column density of hydrogen using the standard gas-to-dust relation (see Mathis 1990). If $h\nu \ll kT$,

$$N(\text{H}) = 1.93 \times 10^{15} \frac{S_\nu}{\theta^2} \frac{\lambda^4}{(Z/Z_\odot) b T_{\text{dust}}}. \qquad (7)$$

S_ν is in Jy, source FWHP size, θ, in arcseconds and λ in μm. $N(\text{H})$ is the column density of hydrogen in all forms, in cm^{-2}. The value of metallicity is Z (for the Sun, $Z_\odot$). The factor b accounts for changes in grain sizes. For moderate densities, $b = 1.9$; for very dense gas, $b = 3.4$, but a wider range is likely. If the uncertainties in equation (7) can be reduced, dust measurements could rival molecular line determinations of mass estimates. Images of dust emission show more structure than images of spectral lines; the explanation for this is not clear, thus far.

A summary of scientific results

It is estimated that there are ~5000 giant molecular clouds in our galaxy. These each have masses of $\sim 10^6 M_\odot$; these are probably cloud complexes with, on average, $n(H_2) \approx 10^2$ cm^{-3} and $T_k \approx 20$ K. Dark dust clouds have larger densities and $T_k = 10$ K. Clouds are assumed to be gravitationally stable and may have a fractal structure. For the Galactic disk inside the solar circle, the masses in H I and H_2 are about equal. Outside the solar circle, the mass of H I is ~6 times that of H_2.

In most clouds, $\Delta v_{1/2}$, the FWHP linewidths, are found to be supersonic. There are some exceptions,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

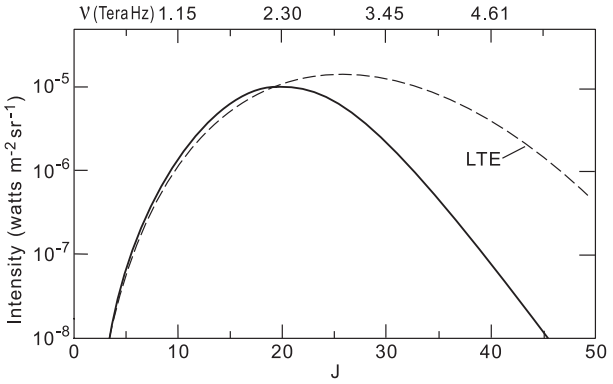


Figure 3. A plot of CO line intensities for the Orion KL source. The dashed curve is a plot of populations for LTE, while the solid curve is an LVG calculation for $T_k = 750$ K, $N(H_2) = 2.8 \times 10^{21}$ cm^{-2}, $n(H_2) = 2.7 \times 10^6$ cm^{-3}. The difference in the curves is caused by the faster decays from higher J levels. With measurements at frequencies above and below the maximum, one can determine both $n(H_2)$ and T_k (Watson *et al* 1985 *Astrophys. J.* **298** 316).

such as the denser cores of dark clouds. Large-scale organized motions, such as rotation, or large-scale collapse are ruled out from observations. It is unlikely that magnetohydrodynamic waves cause most of the line broadening. The linewidths could be caused by outflows from embedded objects, such as young stars, or by the conversion of large-scale motions such as galactic rotation into smaller-scale turbulence. These motions cause shock waves which give rise to broader lines. Theoretical studies show that there are two types of shocks: 'J' or jump shocks and 'C' or continuous shocks. J-shocks locally heat the gas and this leads to the dissociation of molecules. This will change the chemistry of the cloud greatly. C-shocks are cushioned by larger magnetic fields and cause less dissociation. Sub-mm line ratios provide a diagnostic which allows one to differentiate between these types of shocks. For J-shocks, there is a large amount of emission from ions. For C-shocks, the emission is primarily molecular and neutral carbon, but little ionic emission. For a complete understanding, studies of CO, C I and C II are needed (energy level diagrams in figure 2).

The most well-studied Galactic sources are as follows: (1) IRC+10216, a circumstellar envelope of the red giant star CW Leo, ∼150 pc from the Sun; (2) the Orion KL region, at 500 pc, 45″ NW of the Trapezium cluster of stars which ionize the Orion H II region, with FWHP 10″, $T_{kin} = 165$ K and $n(H_2) = 10^7$ cm^{-3}; (3) the dark cloud center TMC-1 in Taurus, with $T_{kin} = 10$ K and $n(H_2) \sim 10^4$ cm^{-3} at 160 pc; (4) Sgr B2, a complex of dense H_2 clouds and H II regions at 8.2 kpc, near the center of our Galaxy; (5) Sgr A, the galactic center itself. In addition, there are a number of well-studied external galaxies. The group is inhomogeneous for a number of observational reasons; the objects are Arp 220, IC342, M82, M51 and NGC 253. In the mm range, all of these were first mapped to a resolution of ∼12″ in a number of lines with single dishes. Later the brighter lines were imaged with interferometers down to resolutions of ∼2″. References to data are given in the bibliography. See especially references in the *Protostars and Planets* series.

INTERSTELLAR CHEMISTRY involves non-equilibrium, non-terrestrial processes. At typical densities, neutral–neutral gas-phase reactions are too slow to form complex molecules in $\sim 10^6$ yr. Rather, the formation involves ion–molecule chemistry. Hydrogen and helium ions are produced by cosmic rays. The H^+ forms H_3^+ or charge exchanges with C or O atoms. The reaction rates for ion–molecule reactions are $>10^3$ times faster than neutral reactions. This explains the abundance of molecules such as CO or OH in dark clouds. Ion–molecule reactions also play a large role in the formation of molecules in the envelopes of low-mass stars in the red giant phase. As a rule of thumb, hydrogen-poor molecules with unsaturated bonds are gas-phase, ion–molecule products, while molecules with saturated bonds are thought to be produced on grains and released into the ISM.

Important goals of mm and sub-mm astronomy are to determine collapse criteria for molecular clouds and to search for individual collapse objects. Stars must form from molecular clouds, because of the close spatial and physical associations. Models for collapse begin with sizes of 10^{16} cm, continuing to pre-planetary disk sizes 10^{14} cm around the resulting star. At present, mm and sub-mm instrumentation limits us to nearby objects which could be imaged on scales $<10^{16}$ cm. After many searches, a number of candidate objects have been found. These are mostly in dark dust clouds where low-mass stars are expected to form. The collapse signature is based on a model where the cloud center has higher temperature and density and where the collapse motions exceed the thermal motions. The measurements of optically thick lines, such as carbon monosulfide, $^{12}C^{32}S$, will show a narrow absorption feature (from gas in the outer part of region). This locates the center of the emission profile. If the cloud is collapsing, the blueshifted portion of this profile has a larger intensity than the redshifted portion, since the radiation from hotter blueshifted gas passes through the redshifted gas without being absorbed. The redshifted radiation from equally hot gas is absorbed by cooler gas at nearly the same velocity. Thus the blue wing arises closer to the warmer gas. If the cloud is expanding, the redshifted portion is more intense. A measurement of the rarer isotope $^{13}C^{32}S$ must show no narrow absorption feature or asymmetry.

An important aspect of molecular cloud physics involves cooling. There are theoretical models, but these are complex even in idealized cases. Surveys of mm and sub-mm spectral line data allow direct tests of the cooling processes by mapping the intensities of the lines of CO, which dominates cooling at lower densities, and H_2O, which should dominate at higher densities. At the highest densities, one expects that hydrides will provide most of the cooling. Measurements of the CO lines from $J = 1$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to 20, H_2O transitions and lines from the more abundant hydrides are needed to experimentally determine the heat flow from clouds.

Fundamental to the STAR-FORMATION process is the outflow phase. This occurs shortly after the core of the star forms. Excess material is expelled in the form of a bipolar high-velocity stream of material which is mainly molecular. Such outflowing material has been traced in mm wavelength lines of CO and other molecules. The central object in this phase is a YOUNG STELLAR OBJECT (YSO). Surrounding the YSO is usually a dense disk of material, perpendicular to the direction of the bipolar outflow; outflows may be driven by magnetic forces.

A completely different type of outflow occurs at the end of the main sequence life of stars of roughly a few solar masses. These asymptotic giant branch stars or red giants are referred to as mass loss stars. Many have been traced in CO lines. The outflow velocities are $\sim$30 km s^{-1}. The mass loss mechanism is thought to be radiation pressure on dust grains, which drags the gas. The moderate velocity flows are sometimes accompanied by flows with velocities larger than 100 km s^{-1}. This material should be hotter and may be atomic rather than molecular. If so, atomic lines of carbon or oxygen might supplement the information based on CO or shock tracers such as SiO to trace these flows.

In PHOTODISSOCIATION REGIONS, the molecules are close to a source of ultraviolet radiation which alters the chemistry and heat balance. The most obvious change is that C II and C I become much more abundant. The layer containing C I may extend to a depth which is a significant fraction of the width of the PDR. The most well known PDR is the 'Orion bar', to the SE of the Orion H II region.

Sub-mm spectroscopy of galaxies is likely to reveal warmer regions of exceptional star-forming or shock activity. Since most galaxies are compact, high angular resolution is needed. Some have moderately intense lines, so can be imaged with interferometers. For most, the emission is usually weak, so lower-resolution data are taken with single telescopes, so far. The most studied galaxy in the mm and sub-mm is Messier 82. It has long since been established that M82 has warm molecular gas; there are HHT maps over a few arcminutes in the $J = 7$–6 line of CO which arises from an energy level 156 K above ground. Thus, there must be very widespread emission from warm molecular gas; details of the heating processes are not fully understood. Another interesting object is the galaxy Centaurus A (see NGC 5128/CENTAURUS A). The nucleus is an intense continuum emitter even in the sub-mm wavelength range, so that species such as hydrides might be detected in absorption. Another fruitful case study is the face-on spiral IC342, thought to be an analog of our Milky Way. Studies will allow a comparison of global properties of these sources. At redshifts >4, CO lines have been detected in the mm wavelength range. In the future, the detection of redshifted lines of molecules, C II and O I would extend chemistry studies to the early universe. See the results summarized by Combes and Wiklind (1998).

Bibliography

Combes and Wiklind 1998 *Highly Redshifted Radio Lines (Astronomical Society of the Pacific Conf. Ser. 156)* ed C L Carili *et al*

Mathis 1990 *Ann. Rev. Astron. Astrophys.* **28** 41

A good reference for sub-mm studies is

Genzel R 1991 *Galactic Interstellar Medium* ed W B Burton *et al* (Heidelberg: Springer)

For the fundamentals of radio astronomy and mm line studies of molecular clouds, see

Rohlfs and Wilson 1996 *Tools of Radio Astronomy* (Heidelberg: Springer)

A recent review of mm and sub-mm interferometric results and techniques is

Sargent A I and Welch W J 1993 *Ann. Rev. Astron. Astrophys.* **31** 297

Sub-mm work is reviewed by

Phillips T G and Keene J 1992 *Proc. IEEE* **80** 1662

Scientific results are summarized in

Levy E H and Lunine J I (ed) 1993 *Protostars and Planets III* (Tucson, AZ: University of Arizona Press)

Hartquist T W and Williams D A (ed) 1998 *The Molecular Astrophysics of Stars and Galaxies* (Oxford: Oxford University Press)

Thomas L Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mills, Bernard Yarnton (1920–)

Australian engineer and astronomer, born near Sydney, and worked with the Australian CSIR Division of Radiophysics. He measured the position of Cygnus A and identified a faint nebulosity as its origin, confirmed by a more accurate position by F. G. Smith and 200 in photographs by W BAADE and R MINKOWSKI showing Cygnus A as a peculiar galaxy. Mills built a cross-type radio interferometer at the Fleurs Field Station: the two arrays in the N–S and E–W arms produced intersecting fan beams which were switched in and out of phase and made a narrow pencil beam at their intersection which could position radio sources accurately. He constructed the *Mills cross* (1954) which operated at 85 MHz and had 1500 foot arms, with which he mapped the Magellanic clouds and conducted source surveys. Mills led the group that constructed the Molonglo Cross, an instrument with one mile arms operating at 408 MHz, built in 1967 and converted in 1981 to a synthesis telescope. Mills discovered the Vela pulsar at the center of the Vela supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Milne, Edward Arthur (1896–1950)

Astrophysicist, born in Hull, Yorkshire, England, became Rouse Ball Professor at Oxford. Milne's theoretical work on stellar atmospheres led to the determination of the temperatures of stars of the different spectral classes and the calculation of limb darkening (the decrease of brightness across the face of a star from center to edge). Milne also showed how molecules escape from stellar and planetary atmospheres. He developed a new form of relativity called kinematic relativity, based on attractive general philosphical ideas. He also suggested that the 'constants' of physics (such as G) changed over the life of the universe. Neither of these propositions is accepted.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mimas

The innermost of Saturn's mid-sized icy satellites, discovered in 1789 by William Herschel. It orbits the planet at a distance of 186 000 km, within the tenuous E ring and near its inner boundary. Mimas is a heavily cratered world, with many craters overlapping. With one exception they are less than 30 km across. The exception is Herschel (named to honor the satellite's discoverer), whose diameter of 130 km is nearly a third of the satellite's diameter, 400 km. Like other mid-sized satellites of Saturn, Mimas may have been destroyed and reassembled itself in the era of heavy bombardment 4 billion years ago. Its density of 1200 kg m^{-3} indicates that its composition is largely of water ice. Mimas's proximity to Saturn's ring system gives it a controlling influence: it keeps the Cassini Division clear, for example, and is responsible for 'corrugations' in the A ring.

See also: Saturn: satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mineur, Henri (1899–1954)

Born in Lille, France, worked at the Paris Observatory, discovered the retrograde rotation of the system of globular clusters around the Galaxy, due to the fact that they are stationary while the Sun moves prograde. Discovered the error in the period–luminosity law for Cepheid variable stars, on which the cosmological distance scale is based. Mineur founded and was the first director of the Institute d'Astrophysique in Paris.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Minkowski, Hermann (1864–1909)

Born in Alexotas, Russian Empire (now Kaunas, Lithuania), became professor at the University of Göttingen, worked on the mathematics of the theory of relativity. Using tensor calculus he cast the work of LORENTZ and EINSTEIN into the form of a four-dimensional non-Euclidean space, now known as *Minkowski space*. This 'space-time continuum' provided the framework on which Einstein developed the general theory of relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Minkowski, Rudolph Leo Bernhard (1895–1976)

Born in Strassburg, Germany, worked on atomic spectroscopy at Hamburg and had to flee the Nazi persecution, joined WALTER BAADE on the Mount Wilson Observatory staff, where he began to apply spectroscopy to astronomy. He investigated nebulae, including supernova remnants, especially the Crab nebula. He classified supernovae into Types I and II, leading to their identification as two similar implosions in different kinds of stars. He headed the National Geographic Society-Palomar Observatory Sky Survey (POSS) which photographed the entire northern sky with the Palomar 48 in Schmidt Telescope. With Walter Baade he optically identified many of the early radio sources, including Cygnus A. In 1960, on his very last observing night with the 200 in Palomar Telescope, he found a galaxy (3C 295) with the then highest redshift, a record which stood for 15 years. After he made the discovery, the night was declared overcast and the occasion toasted in whisky. He discovered Comet 1950b and a small, faint, provocatively shaped nebula is named *Minkowski's footprint*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Minnaert, Marcel Gilles Jozef (1893–1970)

Belgian physicist, became professor of physics in Utrecht in the Netherlands. Worked on the solar spectrum, published a definitive analysis as the *Utrecht Atlas*. Developed the 'curve of growth method' to analyse stellar spectra. His most enduring book is a wonderful explanation of beautiful and interesting natural phenomena called *Light and Color in the Open Air*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mira

The star o Ceti, a red giant, is believed to have been the first variable star ever discovered. It is the type-star of the most common class of long-period pulsating variables, which are usually described as Mira-type stars. Its variability was first recorded by the German theologian David Fabricius (1564–1617) in 1596, and was first studied in detail about 40 years later. It is visible to the unaided eye for only about four months around each maximum, and is a telescopic object for the next seven months—hence Fabricius initially believing it to be a nova. This was of course before the invention of the telescope, so Mira was a star which was then invisible for some two-thirds of the time. It was for many years the only known case of a star varying in brightness. This behavior gave rise to the name Mira ('Wonderful'), which is attributed to Johann Hevelius (1611–87). Despite its long absences it was recorded by Johann Bayer (1572–1625), who showed it as a fourth-magnitude star in his *Uranometria* star atlas of 1603.

Mira's period is variable, but averages 331.96 days (about 11 months). The rise from minimum to maximum is significantly faster than the fall, usually taking about 110 days. There is some evidence from speckle interferometry that Mira is subject to sudden large variations in diameter at particular wavelengths. The spectral type varies during the light-cycle, between M5e and M9e. Its apparent magnitude at maximum is typically 3.0, but can be anywhere in the range 2.0–5.0; minima usually lie between 8.5 and 9.5, but can be as low as 10.1. The star has a parallax of 0.008″, and is at a distance of about 420 light-years. Its absolute magnitude is difficult to calculate with any precision, but is probably about 1.0.

Mira has a close binary companion, the white dwarf VZ Ceti, which is believed to accrete material from Mira's stellar wind. The existence of this faint (thirteenth-magnitude) companion was predicted in 1920 by J H Joy (1882–1973) of the Mount Wilson Observatory, from periodic variations in Mira's spectrum. VZ Ceti is visible only at times when Mira itself is very faint; it was first detected optically with the 36 inch refractor at the Lick Observatory in 1923 by the great double-star observer R G Aitken (1864–1951); the separation was then about 0.9″.

Mira is probably the most thoroughly observed of all variable stars; every maximum since 1638 has been recorded. More than 4000 Mira-type variables are now known.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mira Variables

Mira variables are cool, giant, pulsating stars with periods in the range ∼100 to ∼500 days. The range extends to ∼2000 days if we include OH/IR variables which are generally too faint (due to their low temperatures and to dust shell obscuration) to be readily detected at optical wavelengths and have been mainly found either from their OH-maser emission at radio wavelengths or in the infrared where they are bright. One reason for the importance of Miras in modern astrophysics is that they are low-mass stars in the last, and most luminous, phase of their evolution, before they eject their outer layers as planetary nebula shells (see also STELLAR EVOLUTION). Miras are also important as standard candles for estimating distances within our Galaxy (e.g. the distance from the Sun to center of the Galaxy) and also the distances to nearby galaxies (e.g. the Magellanic Clouds). Matter is being lost from the surfaces of Miras into space in a way that is not yet fully understood. Nor do we precisely understand how the solid grains present in this outflow are formed.

o Ceti, known as 'Mira' (the wonderful), was the first star to have been recorded as variable other than novae or supernovae. It was first noted by David FABRICIUS in 1596, though the fact that it varies in light with a period of about 332 days was not known until the mid-seventeenth century. Its period and light range are typical of this type of star. In the mean it varies from a visual magnitude of 3.4 at maximum, to 9.3 at minimum but the light curve does not repeat exactly from cycle to cycle and it can at times rise to a magnitude of 2.0 and decline as faint as 10.1 mag.

Red (i.e. cool) giant and supergiant variables constitute a large and rather heterogeneous group of objects. The Mira class is restricted to cool stars (surface temperatures lower than ∼3500 to 3000 K) with well defined periods and with visual light ranges of 2.5 mag or more (light ranges of 6 mag or more, as in o Cet itself, are common). Whilst the minimum light amplitude for membership of this class is rather arbitrary it does in fact result in the selection of a group of objects with relatively well defined physical properties. The large light range in the visual region is partly due to the low surface temperatures of the stars. For most Miras the maximum in the energy distribution is in the infrared near 1.6 μm. At such wavelengths the light range is usually much less (typically 1.0 mag or less). Since the visual region is in the short-wavelength tail of the energy distribution, a small change in the surface temperature makes a large difference to the visual brightness. In addition the molecular absorption bands, which dominate the visual region of Mira spectra, strengthen as the surface temperature periodically decreases during the pulsation cycle and their blocking effect further increases the visual light range. Because most of the energy of normal Miras is radiated in the near infrared between 1.0 and 3.0 μm, measurements in this region of the spectrum can be used to estimate the bolometric (i.e. the total) luminosity of the star. There are nearly 6000 Miras or probable Miras currently listed in the *General Catalogue of Variable Stars* (4th edition 1985–1990). But in most parts of the sky only the relatively bright Miras have been so far discovered.

Mira atmospheres

Miras can be grouped into three broad classes according to the chemical composition of their atmospheres. In the largest class (M-type Miras), the spectrum in the optical region is dominated by absorption bands of the TiO molecule. In the S-type Miras absorption bands of ZrO dominate the optical spectrum, though often with strong TiO bands as well. These two classes are grouped together as O-rich Miras (or O-Miras). In the third class (C-type or C-Miras) the spectra are dominated by molecules containing carbon (C_2 etc). Most of the carbon and oxygen in the atmospheres of all these stars is locked up in the stable molecule CO. If there is an excess of oxygen in the atmosphere, the Mira is O-type, if there is an excess of carbon it is of C-type. Both this difference and the difference between the two classes of O-Miras are due to differing amounts of 'dredge-up' that the stars have undergone. 'Dredge-up' is the convective mixing into the upper atmosphere of the star of material that has undergone nuclear processing in its interior. Atomic absorption lines due to the unstable element technetium, which does not occur naturally on the Earth and was first produced in the laboratory in 1937, were found in some S-type Miras in 1951. The half-life of its most stable isotope is 4×10^6 years and the isotope produced by the nuclear s-process (^{99}Tc), the predicted production mechanism in these stars, has a half-life of only 2×10^5 years. This was an important early indication that nuclear processing and the production of complex nuclei take place in some stars, with the processed material being dredged up into the atmosphere. Technetium lines have been subsequently found in many O- and C-Miras and semiregular variables.

Lines of the absorption bands of the CO molecule can be observed in infrared spectra of Miras. The variation in the radial velocity of CO lines of relatively high excitation potential which are formed deep in the atmosphere of the star demonstrates that once per pulsation cycle a shock wave is generated which moves outwards through the stellar atmosphere. Doubling of these absorption lines at certain pulsation phases shows that as one such shock wave is dying out in the higher atmosphere another is being generated lower down. These observations strongly indicate that the variation of Mira stars is due to a pulsation mechanism of some kind. For self-sustaining pulsation one requires there to be a layer in the star which will act as an 'Eddington' valve, absorbing energy when the star contracts and releasing it when the star expands. In the case of Miras this is probably a region in which hydrogen is ionized on compression, recombination and release of energy taking place on expansion.

A characteristic of Mira variables as distinct from most lower-amplitude red variables is that they have emission lines of the hydrogen Balmer series in their spectra (see HYDROGEN SPECTRUM). The intensity of these

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

lines varies systematically in the pulsation period of the star. Most workers believe that the lines are excited by the shock waves just discussed. When first observed, as the brightness of the star increases towards maximum light, the lines have very unusual relative intensities (Hδ much stronger than Hγ for instance). This is due to the emission being excited by the shock deep in the stellar atmosphere. The line strengths are thus greatly modified by absorption in the outer atmosphere (e.g. by molecules of TiO). As the pulsation cycle proceeds, the shock wave moves outwards through the atmosphere and the emission lines which are generated suffer less from the effects of overlying absorption. Many of the well studied Miras also show, at certain phases in their pulsation cycle, emission lines of neutral and ionized metals (e.g. lines of Fe I, Fe II, Mg I, Si I, In I) and even forbidden lines (i.e. lines with very low transition probabilities) of ionized iron ([Fe II] lines). The relative intensities of the metallic emission lines in Mira spectra are quite unlike that seen in other astronomical or laboratory sources. Certain lines of Fe I, Ti I and Zr I, for instance, are abnormally strong compared with other lines of the same element. These lines are excited by fluorescence. This depends on the chance coincidence of a resonance line (that is one whose lower energy level is the ground state) of the element concerned with a strong emission line of another element. Most of the relevant coincidences are in the ultraviolet. Thus long before it was possible to observe Miras from above the Earth's atmosphere it was clear from observations of the fluorescent lines of other elements, seen in the optical region, that the lines of ionized magnesium (Mg II) at 2800 Å must be strongly in emission in some Miras.

The atmospheres of Miras are very extended. A clear indication of this came when the angular diameters of Miras measured in the visual and in the infrared by lunar occultation techniques were compared. For a given O-Mira such as R Leo the diameter at $\sim$2 μm is about half that in the visual. Later it became possible to measure the variation of the angular diameters of Miras with wavelength using interferometric techniques. In the visible region, the angular diameter of o Cet was then found to vary by large amounts (up to a factor of two) depending on whether the observation was made at the wavelength of a strong absorption band of the TiO molecule where the opacity of the atmosphere is very high or just outside the band where the opacity is much lower and where, therefore, we are seeing much deeper into the atmosphere. Thus whilst for most stars the depth of the atmosphere is only a small fraction of the stellar diameter, in the case of Miras the atmosphere occupies most of the star. Evidently the whole concept of 'a diameter' becomes rather vague in these circumstances and different values may well be appropriate for different purposes. Diameters are needed to compare Miras with the theoretical models of these stars and for a discussion of Mira pulsations. In both these cases the diameter is taken to be that measured in the wavelength range in which most of the energy of the star is being radiated (in the infrared at about 1.6 μm). At this wavelength, models suggest that one is measuring the diameter of the stellar photosphere. From Miras with measured angular diameters and parallaxes one can estimate that a Mira of period 300 days has a linear diameter of $\sim 400 R_{\odot}$ (about the distance of Mars from the Sun). The diameters are even larger at longer periods.

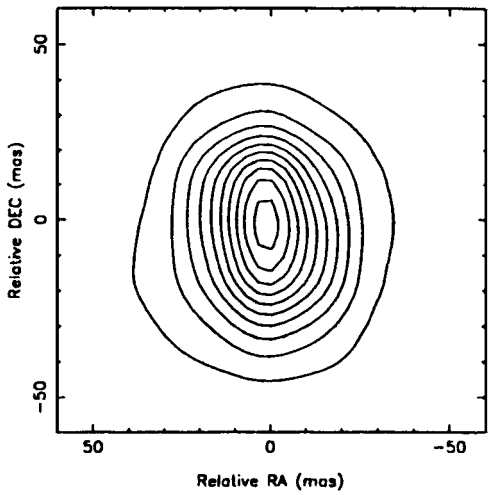


Figure 1. An aperture synthesis image of R Cas showing its non-circular profile (from Haniff 1995).

In recent years it has become possible to reconstruct the images of Miras from interferometric (aperture synthesis) measurements. In several cases (e.g. figure 1) the reconstructed image at optical wavelengths is non-circular. This result applies to the outer region of the atmosphere which, as just discussed, is the region seen at optical wavelengths. It is not yet clear whether it also applies to deeper layers such as seen in the infrared. Three possibilities have been suggested to explain this result: (a) flattening of the outer atmosphere due to rotation; (b) non-radial pulsation of the star; (c) large star spots on the stellar surface which are not fully resolved with current instrumentation and lead to an apparent non-circular profile for the star. The last of these possibilities has generally been favored, mainly because the atmospheres of cool red giants, in general, are expected to contain large convection cells and in these circumstances large over-hot or over-cool regions on the surface might be anticipated. The limited amount of work that has so far been carried out also indicates that the diameter of the outer regions (as seen in the visible) can vary in a manner unconnected with pulsation. In attempting to explain these results it has to be borne in mind that the outer regions are very tenuous and possibly consist of rising and falling clouds with little or no material between them.

The circumstellar region, mass loss, grain formation

One reason why the structure of the outer atmosphere of Miras is of importance is that it is relevant to understanding the process by which Miras lose mass.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mass loss is clearly a key factor in attempting to understand this stage of stellar evolution, and, when properly understood it will lead to a full understanding of how solid grains are formed in the circumstellar environment. At the present time we have only a rudimentary idea of how this process takes place. Remarkably, laboratory studies of small inclusions in some meteorites show them to have the same relative abundances of certain isotopes as those predicted for Mira atmospheres, suggesting that some grains in these meteorites come from Mira outflows.

Miras are losing mass at rates which are typically in the range from $\sim 10^{-7}$ to $\sim 3 \times 10^{-5} M_{\odot}$ yr^{-1}. For extreme OH/IR Miras the rate can be as high as $\sim 10^{-4} M_{\odot}$ yr^{-1}. It has not yet been possible to predict mass-loss rates theoretically. However, observations indicate that the rate is related to the star's pulsation since it is correlated with both the period and the pulsation amplitude. It is generally believed that pulsation raises (gaseous) matter sufficiently above the star that condensation into grains can take place. The concentration of matter into clouds in the atmosphere as mentioned above will clearly aid the process of grain formation. Once formed, the grains are then driven from the star by radiation pressure, dragging gas with them.

The dust round O-Miras is probably composed mainly of silicates whilst that around C-Miras is, probably, generally composed of carbon or carbon-rich compounds. Some C-Miras show large decreases in visual and near-infrared brightnesses on time scales of some thousands of days (i.e. in a time long compared with the pulsation period). These declines in brightness may be similar to dust obscuration events of the RCB type. Thus the dust formation process may be episodic or non-uniform over the stellar surface, possibly connected with clouds in the stellar atmosphere. It seems likely that similar, non-uniform dust formation takes place in the circumstellar envelopes of O-Miras. In this case, though, it is likely to be less obvious in optical observations since, mass for mass, silicate grains of a given size have a much lower opacity in this region than carbon grains. However, temporal changes in the dust shell of o Cet itself are known from interferometric studies in the infrared (at 11 μm) and strongly suggest that the circumstellar dust is concentrated into clouds.

The full complexity of the circumstellar environment of Miras is best seen in the OH/IR stars. These are long-period O-rich Miras many of which were first detected in surveys for objects showing the 18 cm emission lines of the OH molecule. Such lines are emitted by maser action. These objects are usually very faint at optical wavelengths due to heavy circumstellar dust obscuration and probably also to the low temperature of the central star. They tend to be bright at far-infrared wavelengths (at 10 μm and longer wavelengths). This infrared radiation comes from the circumstellar dust heated by the underlying star whose periodic light variations it therefore tends to follow. Besides thermal emission from dust, these objects also show thermal emission from some molecules (e.g. mm and sub-mm lines of CO and SiO) and maser emission from others (H_2O, OH, SiO).

Interferometric studies of Mira itself in the infrared (11 μm) show that dust formation takes place within about two stellar radii of the surface of the star (with the photospheric (infrared) diameter of the star defined as above). Observations of other Miras show that SiO masers are also excited at about this distance from the star's surface—that is, in a shell whose radius is only about 50 per cent greater than that of the outer radius of the star's atmosphere as observed at visual wavelengths. Since a typical Mira of period 350 days has a photospheric radius of $\sim$450 solar radii ($\sim 3 \times 10^{13}$ cm), the inner radius of the dust shell and the radius of the ring of SiO masers is of order 10^{14} cm. H_2O masers are typically of order 10^{15} cm from the center of the star and the OH masers another factor of about ten further away. CO is formed deep in the atmosphere of the star, and provided one has sufficient sensitivity, thermal emission from this molecule can be detected in the outflowing gas and dust mixture up to the point where all the CO has been photodissociated by interstellar ultraviolet radiation. This extent will evidently depend on the mass loss rate and on the speed of the outflow. Observed outer radii of the thermal CO emission around Miras are typically 10^{16} to 10^{17} cm and thus have large angular extent for the nearer Miras, e.g. an angular diameter of $\sim$14 arcsec for o Cet.

OH masers have been detected from a large number of Miras and OH/IR Miras. The lines are generally double peaked. The two peaks come from the front and back of a (presumed) spherical shell expanding at $\sim$10 to 20 km s^{-1}. The OH emission is stimulated by infrared radiation from the heated dust and so varies in the period of the star. There is a time delay (typically 10 to 100 days) between the intensity variations of the two peaks in the spectrum line due to the light travel time across the shell. Since in some cases this shell can also be resolved as a ring in the plane of the sky, one can use the time delay together with the angular diameter of the ring to determine the distance of the source, provided one assumes spherical symmetry.

Circumstellar masers of most extreme OH/IR Miras project on the sky as circular rings and thus do suggest overall spherical symmetry for these objects. For less extreme Miras there is considerable evidence that the outflow from the star can be aspherical and also that the outflowing matter is concentrated into clouds. For instance the fluorescent emission of the KI resonance line 7699 Å is highly asymmetrical around Miras, as is the CO thermal emission. Optical polarization measurements indicate that there must be asymmetrical scattering by the grains in the dust shell. In O-Miras the H_2O maser distribution on the plane of the sky can be asymmetric and in an O-Mira such as R Cas the individual OH masers show that the mass loss is clumped on a scale small compared with the OH shell radius and is intrinsically chaotic with the clumps retaining their identity as they move outwards. These results, together with the apparent non-circular

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

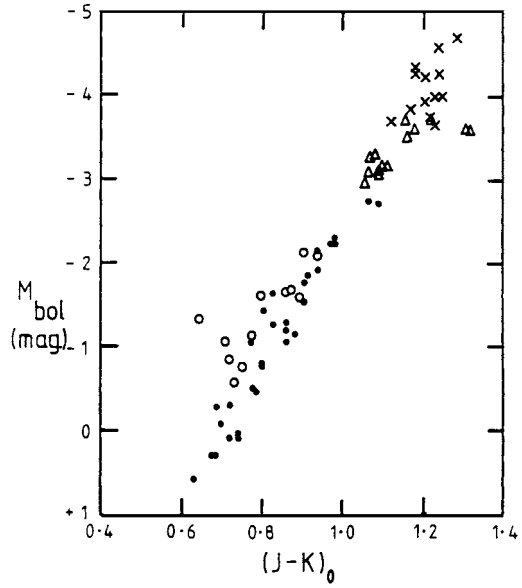


Figure 2. The giant branch of 47 Tuc together with Miras from various clusters; crosses, Miras; triangles, 47 Tuc non-Mira variables; open circles, 47 Tuc AGB stars; filled circles, 47 Tuc RGB stars (from Feast and Whitelock 1987).

profiles of the outer atmosphere of O-Miras, the RCB-like declines of some C-Miras and other evidence, which was mentioned above, suggest that overdense, or unusually cool, regions (clumps, clouds) in the upper atmosphere of the star may be important in the mass-loss and grain formation process.

The circumstellar environment of C-Miras has been somewhat less well studied than that of O-Miras. However, CW Leo, also known as IRC + 10216, a C-Mira with a pulsation period of 630 days, has a shell in which infrared, mm and sub-mm studies have revealed thermal emission lines from more than 40 different chemical compounds. Some of these are quite complex (e.g. HC_9N) and the object is important for our understanding of circumstellar chemistry.

Evolution, pulsation

The evolutionary state of O-Miras is perhaps best shown by those short-period Miras which are members of relatively metal-rich GLOBULAR CLUSTERS. Low-mass stars such as those in globular clusters evolve off the main sequence and climb the red giant branch (RGB). From the top of the RGB they drop to the horizontal branch and then climb through the red-giant region again on the asymptotic giant branch (AGB). Figure 2 shows a color–magnitude diagram with the known cluster Miras together with the RGB/AGB of 47 Tuc, a globular cluster which contains three Miras. Miras are evidently the brightest and coolest of the stars in the clusters in which they occur. Their luminosities are above the predicted maximum for RGB stars (at least for periods greater than about 160 days). They are thus at the tip of the AGB. After this stage they are expected to eject their atmospheres and become planetary nebulas. Near the top of the AGB, stars begin the process of thermal pulsing during which they burn helium and hydrogen alternately in a nuclear active zone. This zone surrounds an inert core which is essentially a white dwarf of about the same size as the Earth. The Miras belong to this thermal pulsing stage, and it seems likely that stars only manifest Mira characteristics during the brighter phases of each thermal cycle. A star may thus become a Mira several times, although this is not yet certain. Between these times it would probably be a low-amplitude semiregular variable such as can be seen marked in figure 2. A thermal pulse is initiated once every $\sim 10^5$ years in a solar mass star by a helium flash (a thermonuclear runaway) which lasts ~500 years. During this relatively short period the structure of the star changes significantly and one might anticipate changes in the pulsation period due to this. Whilst any long-term changes in the periods of most Miras are very small, a few, e.g. R Hya, R Aql, W Dra and BH Cru, show good evidence for significant changes. Thus the period of R Hya has changed from ~500 days in 1700 to 388 days at the present time. It seems likely that these rare objects are currently undergoing helium flashes.

The occurrence of short-period Miras in globular clusters (e.g. three Miras with periods near 200 days in 47 Tuc) shows that Miras at these periods must be old objects with initial masses near solar. The data from globular clusters are limited but they indicate that when there is more than one Mira in a cluster their periods are close to one another. There is also a relation between the periods of Miras in clusters and the cluster metallicity, the period being longer in higher-metallicity clusters. Unfortunately it has not been possible to test this period–metallicity relation for Miras in the general field. The reason for this is that although it is possible to classify Miras spectroscopically into the three broad chemical groups (M, S and C) noted above, their atmospheres are so complex that so far no satisfactory detailed analyses for chemical abundance have been made. However, the galactic kinematics of Miras in the general field are correlated with their periods allowing a classification according to POPULATION. The short period ~200 day stars belong to an old, thick disk or intermediate-halo population as expected since some Miras of this period occur in relatively metal-rich globular clusters. Longer period Miras belong to populations which are more metal-rich and/or younger. Thus the sequence of Mira periods is a sequence of the end points of AGB branches of different ages and/or chemical abundances. It seems likely that the 1000–2000 day OH/IR Miras simply continue this sequence to larger mass stars (perhaps to stars of initial mass $\sim 4 M_\odot$). This is, however, still uncertain, partly because the central stars of these objects are obscured from view by their thick circumstellar shells and so cannot be adequately studied.

The above discussion refers to O-Miras. The place of C-Miras in stellar evolution is less certain. Some at least

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

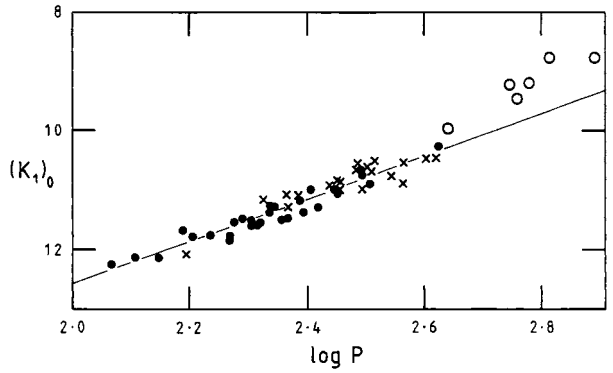


Figure 3. The period–luminosity relation for Miras in the LMC. Ordinate, apparent magnitude at K (2.2 μm); abscissa, log of the period in days; filled circles, O-Miras; crosses, C-Miras: open circles, long-period Miras of uncertain status (from Feast *et al* 1989).

probably evolve from O-rich Miras as a result of dredge-up processes without a significant change of period.

Miras in the LARGE MAGELLANIC CLOUD (LMC), where they are all at closely the same distance from us, show well defined linear relations between bolometric or infrared magnitude and log P for periods between ∼100 days and ∼400 days. The relations for O- and C-Miras are almost the same at 2.2 μm (K-band) (figure 3) and only slightly different in bolometric luminosity. The scatter about the relations is very small (only 0.13 mag about the O-Mira relation at K). This small scatter is a reason for suggesting that Miras may occur only at the brighter phases of a thermal pulse. The absolute magnitudes of the Miras (i.e. the zero point of the Mira period–luminosity (PL) relation) can be obtained in the following ways: by adopting a distance to the LMC (e.g. as derived from Cepheid variables); by using the distance of globular clusters which contain Miras; by using the absolute magnitudes of main-sequence companions to Miras in visual binary systems; or by using the trigonometrical parallaxes of Miras (the Hipparcos astrometry satellite has provided a set of such measures). Infrared color differences between O-Miras of the same period in the LMC and the Galaxy suggest that the period–metallicity relation may not be the same everywhere and such a difference may imply slight differences in luminosity at the same period. The bolometric luminosity of a Mira with a period of 200 days is about 4000 times that of the Sun and that of a 400 day Mira about 9000 times the Sun's.

Once the absolute magnitude, and hence the distance, of a Mira is known this can be combined with an angular diameter to give a linear diameter. The presence of short-period O-Miras in globular clusters as well as the galactic kinematics of O-Miras show that at least those with periods of about a year or less have masses near that of the Sun. One can then use this fact together with the linear diameter and the period to deduce the mode in which Miras pulsate. Good agreement is found in this way with the theoretical predictions for pulsation in the first overtone. This, however, remains a controversial result, mainly because theoretical predictions of shock waves strong enough to excite the observed emission lines discussed above seem to require a smaller, hotter star and this would then be pulsating in the fundamental mode. That the topic has remained controversial is partly due to uncertainties in the definition of the diameter, as discussed earlier, and also to uncertainty in the theoretical predictions of both pulsation and shock wave production. There is evidence for multimode behavior in a few Miras. In such cases the light curve shows evidence of the superposition of two or more periodicities. In principle the period ratios of the excited modes allow mode identification but the theoretical predictions depend on the model of the atmosphere which is adopted. The resolution of these problems will no doubt lead to a deeper understanding of the complex atmospheres of Miras.

The progenitors of Miras are clear in the case of the shorter period Miras that occur in globular clusters, at least if we adopt the standard picture of stars evolving along the AGB to higher luminosities and lower surface temperatures. As figure 2 shows, variability sets in as one ascends the AGB, first as irregular or low-amplitude semiregular variations and then as Mira-type pulsations. In a period–absolute magnitude diagram the low-amplitude semiregular variables evolve along the 'Whitelock' relation of figure 4 with increasing luminosity, period and pulsation amplitude, finally reaching the top of the AGB as Miras on the Mira PL relation. It is presumed that longer-period Miras which evolve from more massive or more metal-rich stars do so in an analogous way. Neither the mass loss of Miras nor their further evolution is fully understood, though the stars are widely believed to eject their outer layers as planetary nebulae.

Mira binaries, red supergiant variables

A number of Miras are known in interacting binary systems. SYMBIOTIC STARS generally show a rich emission spectrum of relatively high excitation, believed to be energized by the infall of gas from the Mira onto a disk round a compact companion. The nearest such system contains the Mira R Aqr (period = 387 days) and is resolved as a bipolar nebula on the plane of the sky. In some symbiotics (e.g. RR Tel, period = 387 days) the presence of a Mira in the system is not obvious from optical observations but the variable dominates the light of the system in the infrared. o Cet itself is transferring mass to a white dwarf companion and this variable should perhaps be regarded as a marginal member of the symbiotic class. QX Pup (also denoted by OH0739 − 14 or OH231.8 + 4.2, depending on whether one specifies the star by equatorial or galactic coordinates) is a 700 day Mira in a complex bipolar nebula showing high-velocity outflows (at ∼200 km s^{-1}). It does not show the characteristic symbiotic emission line spectrum, but this Mira too seems likely to be part of an interacting binary system. The QX Pup system is sometimes known as the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

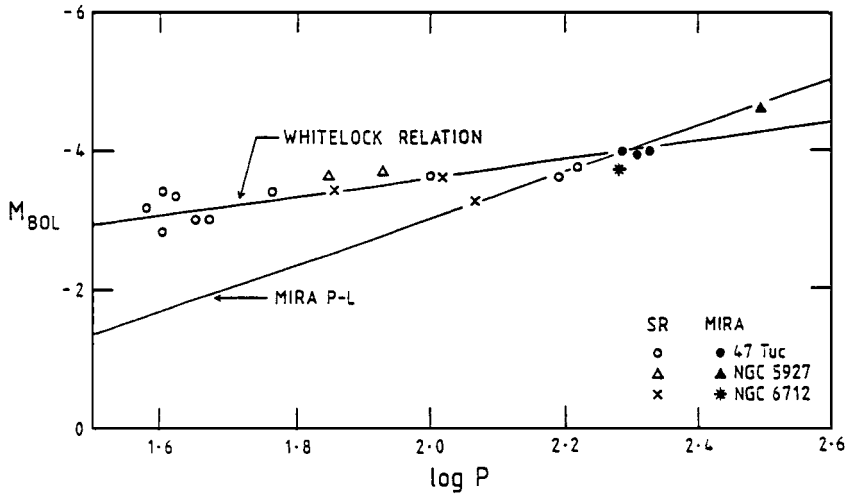


Figure 4. The absolute bolometric magnitude plotted against log period in days for semiregular (SR) and Mira variables in metal-rich globular clusters. The Mira PL relation is shown. The SRs define a line of shallower slope (the Whitelock relation) (from Feast 1989).

Calabash Nebula (from its shape) or the Rotten Egg Nebula (from the great strength of emission lines of H_2S and other sulfur compounds in its mm spectrum).

There is a rare class of supergiant red variables which like the Miras show a rather regular periodicity in their light variations and these are sometimes confused with the Miras. They tend to have periods greater than about 400 days and lower light amplitudes than the Miras. They are also about two magnitudes brighter than Miras but have been much less well investigated. They are known in our Galaxy, the Magellanic Clouds and some have been found in the spiral galaxy M33. Evidence for a period–luminosity relation has been found. These objects are believed to be massive (progenitor mass $\geq 9M_{\odot}$), core helium-burning stars. It is conventional to refer to such objects as supergiants whilst Miras and other stars on the AGB are termed giants. This convention holds even though the spectra of some Miras place them in luminosity classes II or I (i.e. they are spectroscopically classed as supergiants.).

Current research

Infrared, radio and interferometric techniques are currently leading to an understanding of the complex atmospheres of Miras, their circumstellar environment, and the way these are related to stellar pulsation. One aim is to determine the mechanisms of mass loss and grain formation, neither of which are properly understood at present. The final ejection of the outer layers of a Mira, probably as a planetary nebula, is a crucial phase of stellar evolution about which even less is known.

The importance of Miras as distance indicators within our own Galaxy and in extragalactic systems is being increasingly realized. For instance Miras are present in large numbers in the central bulge of our Galaxy and studies of these stars at optical, infrared and radio wavelengths have already made important contributions to our understanding of the composition, kinematics and structure of the bulge, including its bar-like shape. They have also enabled an estimate to be made of the distance to the galactic center. Since Mira periods are related to the stellar population to which they belong, they are valuable indicators of galactic evolution.

Miras are very luminous in the near infrared and it is anticipated that large numbers of them will be found in infrared sky surveys such as that of the European consortium DENIS and of the American consortium 2MASS, both of which began in the late 1990s. The colors of Miras measured in the infrared are distinctive and allow such stars to be recognized even before variability studies are made. Follow-up studies of Miras discovered in the surveys should lead to major improvements in our understanding of both stellar and galactic evolution.

Bibliography

Feast M 1989 *The Use of Pulsating Stars in Fundamental Problems of Astronomy* ed E G Schmidt (Cambridge: Cambridge University Press) p 205

Feast M and Whitelock P A 1987 *Late Stages of Stellar Evolution* eds S Kwok and S R Pottasch (Dordrecht: Reidel) p 33

Feast M W, Glass I S, Whitelock P A and Catchpole R M 1989 *Mon. Not. Roy. Ast. Soc.* **241** 375

Haniff C 1995 *Astrophysical Applications of Stellar Pulsation, Astron. Soc. Pacific Conf. Ser.* vol 83, eds R S Stobie and P A Whitelock (San Francisco: Astronomical Society of the Pacific) p 277

Michael Feast

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Miranda

Miranda is the smallest of URANUS' five major satellites and one of the most unusual objects in the SOLAR SYSTEM. Miranda is such a small world, only about 470 kilometers (300 miles) in diameter, that it was not discovered until 1948 by the American astronomer GERARD P KUIPER. Very little more could be learned about Miranda until the VOYAGER 2 spacecraft flew through the Uranian system in 1986. Because of navigation requirements to give Voyager 2 the gravitational assist necessary to send it on to Neptune, the spacecraft passed closer to Miranda than any other large object in the Uranian system. The targeting of Voyager 2 near Miranda was viewed by most planetary scientists as an unexciting necessity, since all evidence indicated that Miranda would be only an uninteresting, heavily cratered object.

As it turned out, Miranda became the great surprise of the encounter. It is indeed a small object, but its surface is incredibly diverse. The expected heavily cratered terrain is present, but the surface is also criss-crossed by a network of rift-like canyons more characteristic of a much larger object. And, most intriguingly, set amidst the craters and canyons are three large geologically modified patches called 'coronae', unlike any other feature seen elsewhere in the solar system (figure 1).

All three of the coronae have sharply defined boundaries, with surface topographies and textures quite different from the cratered terrain. All three have very few superposed craters, indicating that they formed later than, and on top of, the heavily cratered terrain. Yet each corona is unique in detail. The largest of the coronae, named Arden, consists of a set of concentric light and dark bands and looks something like a great oval racetrack. The outermost 'lane' in the racetrack is one of the huge canyons. The corona over the South Pole, named Inverness, also has light and dark areas, but the pattern is more rectangular than Arden's. The surface of Inverness is also covered by sets of low, parallel ridges arranged in a roughly rectangular pattern. The third corona, named Elsinore, appears uniformly gray in the images. Its surface is covered by parallel sets of low ridges and troughs like Inverness', but the pattern is oval, like Arden's.

Canyons on Miranda occur over a wide range of sizes, but the most prominent are typically 50–60 km wide, 6–8 km deep and hundreds of km long. A few of the canyons appear smooth and 'softened' by age or overlying material, but most appear rugged and fresh. The walls of many of the fresh canyons consist of a series of steep, smooth slopes arranged like tilted steps from canyon floor to rim. However, some canyon walls, like Verona Rupes off the northeast corner of Inverness, consist of a single magnificent step from floor to rim.

Even the population of CRATERS on Miranda is somewhat unusual. On most objects in the solar system, craters range in appearance from fresh and crisp to old and worn down, with all stages of preservation in between. On Miranda some craters appear fresh and some appear worn, but there are almost none in between. Most of the large craters on the cratered terrain appear worn down while the smaller craters on both the cratered terrain and the coronae are sharp and fresh.

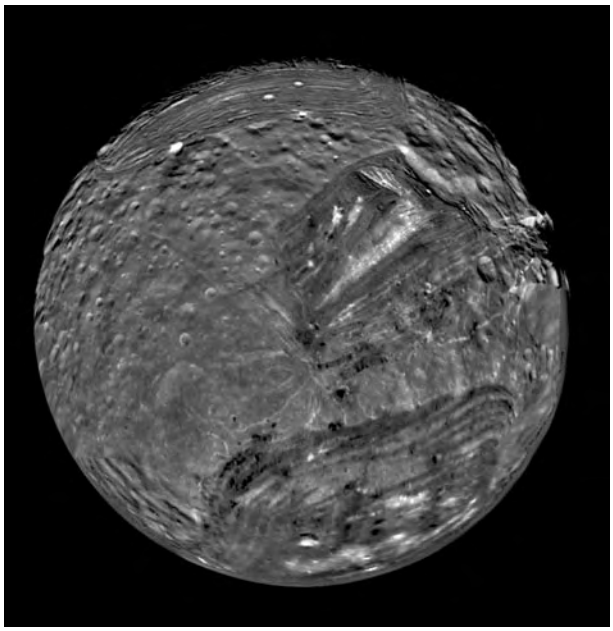


Figure 1. This picture of Miranda is a mosaic of the best images obtained by Voyager 2 in 1986. The projection is orthographic, with the South Pole at the center. The surface shows an incredible variety of features: the heavily cratered terrain at left; large canyons cutting across the surface, including Verona Rupes at right; and the three coronae, Arden at the bottom, Inverness at center, and Elsinore at the top of the image. (NASA photograph.)

The unusual and unexpected collection of features on Miranda has prompted a number of competing explanations for their origin. One early suggestion, based on the theoretical idea that small objects near large planets like Uranus should have been blasted apart many times by incoming PLANETESIMALS, was that Miranda had been catastrophically shattered into many large pieces after it had partially melted into dark rock-rich and light ice-rich layers. The shattered pieces then fell back together in random order, with darker fragments formerly in the deep interior forming the coronae now visible on the surface. An extension on this idea was that the canyons had been formed as rock-rich chunks and had slowly sunk back to Miranda's center, stressing and breaking the surface.

Subsequent detailed analysis of the canyon networks led to their interpretation as cracks or rifts formed by stretching of the surface, which in turn was caused by upwelling of material from below. The appearance of the coronal materials—particularly the different brightness of distinct features (especially in Inverness, see figure 2), the morphology of the surface ridges, and the lobed and flow-like edges—led to their interpretation as huge deposits of lava-like material that flowed out onto the cratered surface, filling low spots and spilling out on top of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

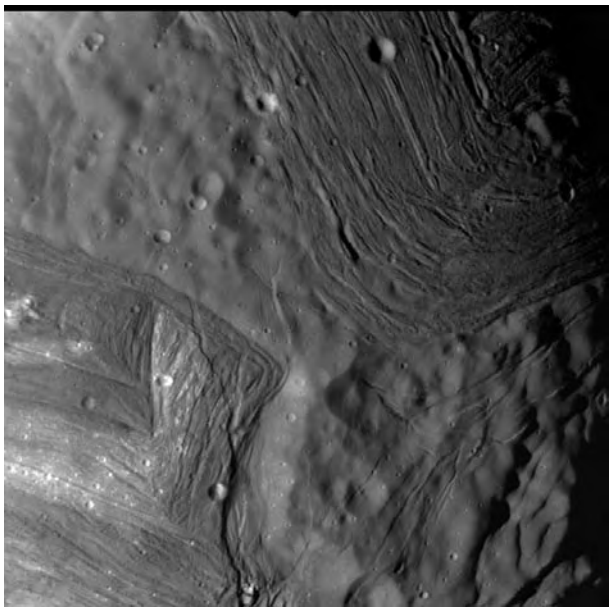


Figure 2. This picture taken by Voyager 2 in 1986 shows portions of Elsinore Corona at upper right, and Inverness Coronae at lower left. A strip of ancient cratered terrain separates them. Note how the coronal materials appear to have flowed over the older cratered surface. (NASA photograph.)

cratered terrain. Both the inferred surface stretching and the volcanism can be accounted for by heating, thermal expansion and partial melting deep in Miranda's interior. The problem is in understanding how sufficiently high temperatures to cause such extensive geologic activity can occur in as small a body as Miranda. Ideas accounting for the apparently excessive heating include episodes of tidal heating, or the presence of an outer layer of material with a very low thermal conductivity, leading to enhanced insulation of the normal internal radioactive heat source. This interpretation does not exclude the idea that Miranda had been blasted apart early in its history, but the present surface features do not appear to result from such an event.

Miranda is an icy body. Spectral analysis of its reflected light shows the presence of water ice on its surface. Miranda's density of 1.15 g cm^{-3} indicates that its interior is also made mostly of water ice with a small fraction of rocky material. Miranda's density is somewhat smaller than the densities of its sister satellites (see URANUS), implying that Miranda might have a smaller fraction of rocky material in its interior than its neighbors. Alternatively, Miranda's outer layers could be more porous than its larger neighbors'.

If the coronal materials are indeed formed by eruptions of lava-like materials, they indicate the presence of at least small amounts of ices other than water ice. 'Lavas' on Miranda would not be made of liquid rock as on Earth, but of liquid water. However, the putative lava deposits on Miranda are up to a kilometer thick and have surface ridges and curved edges indicative of a lava with high viscosity. Pure water has a low viscosity, but mixtures of water with such compounds as ammonia and methanol do have sufficiently high viscosities to form the coronal features. Such icy mixtures also have very low melting points, making melting on small bodies like Miranda easier. Ices of compounds such as methanol have been detected on other icy bodies in the outer solar system, so they could be present on Miranda (see VOLCANISM IN THE SOLAR SYSTEM). Such icy lava would erupt onto Miranda's surface with temperatures near −120 °C and freeze down to Miranda's average surface temperature near −190 °C. At this temperature, water ice is as stiff as rock on Earth and will form permanent features.

Bibliography

General geology:

Croft S K and Soderblom L A 1991 Geology of the Uranian Satellites *Uranus* ed J T Bergstrahl and M S Matthews (Tuscon, AZ: University of Arizona Press) pp 561–628

Competing hypotheses:

Greenberg R, Croft S K, Janes D M, Kargel J S, Lebofsky L A, Lunine J I, Maricalis R L, Melosh H J, Ojakangas G W and Strom R G 1991 Miranda *Uranus* ed J T Bergstrahl and M S Matthews (Tuscon, AZ: University of Arizona Press) pp 693–735

Latest geologic interpretation:

Pappalardo R T, Reynolds S J and Greenly R 1997 Extensional tilt blocks on Miranda: evidence for an upwelling origin of Arden Corona *J. Geophys. Res.* **102** 13 369–79

Steven K Croft

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mirror

An optical component from which light is reflected. Because light is reflected from the front surface of an astronomical mirror, it does not have to travel through the material of which the mirror is made. An astronomical mirror usually consists of a suitably shaped substrate (the body of the mirror) on which a thin layer of highly reflective metal, such as silver or aluminum, is deposited. The substrate is normally a material with a very low coefficient of expansion which, therefore, expands and contracts as little as possible when its temperature changes, typical materials being Pyrex (a low-expansion glass), quartz or ceramics.

Telescopes other than optical telescopes (e.g. for x-rays or radio) use mirrors (or reflectors) in analogous ways; however, the materials of the substrate and reflecting surface are adapted to the properties of the radiation.

In a reflecting telescope, light is collected and brought to a focus by a concave 'primary' mirror, the size of the image formed at its focal plane being proportional to the focal length of the mirror (the distance between the surface of the mirror and the focus). In most reflecting systems, the converging cone of light from the primary is intercepted by a smaller secondary mirror. If the secondary mirror is flat, it reflects the converging cone of light to a different position without changing its angle of convergence and without altering the focal length of the instrument. For example, in the Newtonian design, a flat secondary reflects light to a focal position at the side of the telescope tube. In other designs, the front face of the secondary is curved and changes the angle at which light rays are converging, thereby altering the effective focal length of the telescope. In the Gregorian and Cassegrain designs, the secondary reflects light back down the telescope tube, through a hole in the center of the primary mirror, to a new focal position. The Gregorian uses a concave secondary, the surface of which has an elliptical cross-section, whereas the Cassegrain (and its derivatives) uses a convex secondary with a hyperbolic curve.

If the surface of a mirror has a spherical curve, it will be unable to focus parallel rays striking different parts of its surface to the same focal point (this defect is called 'spherical aberration'). Most reflecting telescopes use paraboloidal primary mirrors (mirrors in which the cross section of the surface is part of a parabola) to overcome this problem. However, paraboloidal mirrors produce crisp, sharp images only in the central parts of their fields of view, images towards the edges of the field suffering from another defect called coma. The Ritchey–Chrétien design (a development of the Cassegrain), which produces good-quality images across most of its field of view, has a hyperboloidal primary and a hyperboloidal secondary.

See also: Cassegrain telescope, Gregorian telescope, focal length, focus, Newtonian telescope, reflector, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mitchell, Maria (1818–89)

First woman astronomer in America, born in Nantucket, MA, became librarian of its Atheneum Library, helped her father in his amateur observatory. In 1847, independently discovered a comet named *Miss Mitchell's Comet*, and became a 'computer' at the US Nautical Almanac Office, calculating the motion of Venus. After a European tour, she was presented with a telescope bought with money collected by American women. Became professor of astronomy and director of the observatory at the newly opened Vassar College in Poughkeepsie, New York, photographing Jupiter, Saturn and the stars, and furthering women's education.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mizar and Alcor

Mizar (Arabic, 'the Girdle') is the name given to ζ Ursae Majoris, the center star in the handle of the Plough, by J J Scaliger (1540–1609). The star was formerly known as Merak or Mirak, from the Arabic Al Marakk—'the Loin' (of the bear). This duplicated the name given to another star, β Ursae Majoris, which may account for Scaliger's introducing a new name. Mizar is an interesting multiple star system, but is perhaps best known for its pairing with Alcor ('the Rider'). Fourth-magnitude Alcor is separated from second-magnitude Mizar by 11.82′ at position angle 72°, and is thus an easy naked-eye double. In ancient Arabia, where it was often referred to as Suha, 'the Forgotten [or Lost or Neglected] One', it was regarded as a test of acute vision. This suggests that it may have brightened over the past few centuries.

Mizar is a main-sequence star of apparent magnitude 2.23 and spectral type A2V; with a parallax of 0.042″ it is 78 light-years distant. Its absolute magnitude is 0.3. It is believed to be the first binary system ever discovered, in 1650 by Giovanni Riccioli (1598–1671). In 1889 its brighter component also became the first spectroscopic binary to be announced, following the discovery by Anna C Maury (1866–1952) of the Harvard College Observatory that the absorption lines in its spectrum alternately appeared double and single, with a period of 20.5 days. The spectroscopic binary, separation about 0.01″, was first resolved telescopically in 1925 by F G Pease (1881–1938), using a 20 foot beam interferometer attached to the 100 inch reflector at the Mount Wilson Observatory—an epoch-making instrument devised by Albert Michelson (1852–1931), with which the first stellar diameters were also measured. The components of the spectroscopic binary are now designated ζ UMa A and B. The secondary star of the main binary system, designated ζ UMa C, is a similar main-sequence star of spectral type A7Vm and apparent magnitude 3.85. It is separated from ζ UMa AB by 14.4″ at position angle 152°. This component is itself also a spectroscopic binary, with a period of 175.5 days.

Alcor (80 UMa) is a white main-sequence dwarf, spectral type A5V, apparent magnitude 3.99. It is a little more distant than Mizar at 81 light-years (parallax 0.040″), but may be a distant member of the same gravitational system. Its absolute magnitude is 2.0. All these stars are part of the Ursa Major moving cluster, having similar and approximately parallel proper motions.

In 1722 an eighth-magnitude star, between and a little to the south of Mizar and Alcor, was discovered by J G Liebknecht (*fl*. 1720), a little-known German professor of mathematics and theology. He was observing with a simple refracting telescope of poor quality, and his inadequate measurements of the star's position led him to conclude erroneously that the object was a planet, which he named 'Sidus Ludoviciana', after the Landgrave Ludwig of Hessen-Darmstadt. The name was copied in many subsequent atlases, a rare distinction for such a faint and unimportant star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Molecule

The smallest particle of a chemical compound that can exist by itself and which has all the chemical properties of that compound. A molecule consists of two or more atoms, of the same elements or of different elements, linked together by chemical bonds. It may be described by its molecular formula. For example, O_2 denotes a molecule consisting of two oxygen atoms (this is the form in which most of the oxygen in the Earth's atmosphere exists), and O_3 denotes a molecule consisting of three oxygen atoms (this particular form of oxygen is called ozone). A molecule of water (a compound), which consists of two hydrogen atoms and an oxygen atom bonded together, is denoted by H_2O, a molecule of carbon dioxide (one carbon atom joined to two oxygen atoms) by CO_2, and so on.

The molecular mass (often referred to as 'molecular weight') of a molecule is its mass expressed in atomic mass units and is equivalent to the sum of the atomic masses of all its constituent atoms. In round figures, the molecular mass of a diatomic (two atom) oxygen molecule is 16 (the atomic mass of oxygen being 8), the molecular mass of water is 10 (the atomic mass of hydrogen being 1) and the molecular mass of carbon dioxide is 28 (the atomic mass of carbon being 12). In practice, atomic and molecular masses differ slightly from precise integral values. Molecules range in size and mass from simple diatomic molecules such as H_2 to macromolecules with molecular masses of more than 10 000.

A molecule has a physical size and shape that depends on the separation and relative orientation of its constituent atomic nuclei. Consequently, it exhibits three types of discrete (quantized) energy levels and transitions (changes between levels): electronic transitions, which involve changes between energy states in the shared cloud of electrons that surrounds its constituent atomic nuclei; vibrational transitions between its various permitted states of vibration (these involve the separations between nuclei); and rotational transitions, which involve discrete changes in the way in which the molecule rotates around various axes. Each of these kinds of transition is accompanied by the emission or absorption of radiation of a specific wavelength, the three classes of molecular transition giving rise to large numbers of lines, many of which overlap to produce broad bands in the resulting spectra. Through studies of their emission and absorption spectra, astronomers have identified more than a hundred species of molecule in the interstellar medium.

See also: absorption spectrum, atom, emission spectrum, interstellar molecular clouds, ion, quantum mechanics, radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Molonglo Observatory Synthesis Telescope

Australian radio telescope, located near Canberra and operated by the University of Sydney. A large Mills Cross radio interferometer was built there in 1966. One arm of the cross has been converted into the Molonglo Observatory Synthesis Telescope (MOST).

MOST consists of two cylindrical 778 m $\times$ 12 m paraboloids, 15 m apart and aligned east–west. The telescope is steered by mechanical rotation of the cylindrical paraboloids about their long axis, and by phasing the feed elements along the arms. The resulting system can follow a sky field for ± 6 h if it lies south of declination -30°.

For further information see
http://www.physics.usyd.edu.au/astrop/most/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Molyneux, William (1656–98) and Molyneux, Samuel (1689–1728)

William was an optical physicist, born in Dublin, independently wealthy and an English government official in Ireland, studied optics and published *Dioptrica Nova* (1692). He obtained a telescope from FLAMSTEED and made astronomical observations, but while his private correspondence with Flamsteed and HALLEY was extensive he never published in astronomy. His second son, Samuel, a prosperous English amateur astronomer, attempted with JAMES BRADLEY to measure the parallax of Gamma Draconis with a 'zenith sector', i.e. a telescope (in this case mounted on the chimney stack of Molyneux's house) looking at stars directly overhead. They found an impossibly large angular shift (up to 40 arcseconds), with the wrong phase characteristics. This turned out to be not parallax but 'the aberration of light' due to the combination of the Earth's orbital velocity (changing throughout the year) with the incoming velocity of light from the star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Molyneux, William (1656–98) and Molyneux, Samuel (1689–1728)

William was an optical physicist, born in Dublin, independently wealthy and an English government official in Ireland, studied optics and published *Dioptrica Nova* (1692). He obtained a telescope from FLAMSTEED and made astronomical observations, but while his private correspondence with Flamsteed and HALLEY was extensive he never published in astronomy. His second son, Samuel, a prosperous English amateur astronomer, attempted with JAMES BRADLEY to measure the parallax of Gamma Draconis with a 'zenith sector', i.e. a telescope (in this case mounted on the chimney stack of Molyneux's house) looking at stars directly overhead. They found an impossibly large angular shift (up to 40 arcseconds), with the wrong phase characteristics. This turned out to be not parallax but 'the aberration of light' due to the combination of the Earth's orbital velocity (changing throughout the year) with the incoming velocity of light from the star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Monoceros

(the Unicorn; abbrev. Mon, gen. Monocerotis; area 482 sq. deg.) An equatorial constellation which lies between Orion and Hydra, and culminates at midnight in early January. It is usually considered to have been introduced by the Dutch theologian and geographer Petrus Plancius, who included it on a celestial globe in 1613, though some authorities regard it as older than this.

An inconspicuous constellation, the brightest star in Monoceros, β Monocerotis (combined magnitude 3.8), is actually a multiple system consisting of three bluish-white (B3) components, magnitudes 4.5, 5.2 and 5.6, separation 7.2″ and 10.1″ from the primary, which also has an unseen companion. Other bright stars are α Monocerotis, magnitude 3.9, and γ Monocerotis, magnitude 4.0. Other interesting stars include Plaskett's Star, a sixth-magnitude spectroscopic binary whose components are among the most massive stars known, the variables U Monocerotis (range 6.1–8.8), an RV Tauri star, and R Monocerotis (range 11.0–13.8), a T Tauri star which is enveloped in and illuminates NGC 2261 (Hubble's Variable Nebula, a Herbig–Haro object), and A0620–00, a possible three-solar-mass black hole.

Monoceros is crossed by the Milky Way and contains several bright open star clusters, including: NGC 2264, a large triangular-shaped cluster of over a 100 stars, including a couple of dozen of eighth or ninth magnitude and also 15 Monocerotis (also known as S Mon., magnitude 4.7v), at the southern end (apex) of which lies the dark Cone Nebula; NGC 2232, a group of a couple of dozen stars, many of sixth to eighth magnitude, surrounding 10 Monocerotis (magnitude 5.1); NGC 2244, another large cluster, including a dozen or so sixth to ninth magnitude stars, embedded in the Rosette Nebula (NGC 2237-9, 2246), a diffuse emission nebula; and M50 (NGC 2323), a rich group of several dozen stars fainter than eighth magnitude. The few galaxies known are relatively small and faint.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Montanari, Geminiano (1633–87)

Italian scientist, born in Modena, practised law in Vienna. Returned to Italy to Modena as court philosopher and mathematician, where he was taught by Count Cornelio Malvasia, with whom he went to Bologna and became professor of mathematics at the university. Montanari made a map of the Moon for Malvasia's ephemeredes of 1662, developing a telescope reticule to make the observations; he published ephemeredes himself in 1665. He studied variable stars (discovering the variability of Algol), observed comets, waged war on astrology, made excellent lenses and developed a machine he called a 'sferologio' that showed all of the movements of the heavenly bodies like an orrery.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Month

A unit of time based on the motion of the Moon around the Earth. The synodic (or lunar) month is the mean time interval between two successive new moons (i.e. the time taken for the Moon to pass through its cycle of phases) and is equal to 29.53059 mean solar days. Because the synodic month (the basis of the lunar calendar) is not equal to a whole number of days, the calendar month is rounded to a whole number of days, as follows: 30 days in April, June, September and November, 28 days in February (29 in a leap year), and 31 days in all other months.

Other definitions of the month include the following. The sidereal month (duration, 27.32166 days) is the time taken for the Moon to move through an angle of 360° relative to the background ('fixed') stars. The anomalistic month is the time interval (27.55455 days) between successive passages of the Moon through perigee. The tropical month (27.32158 days) is the time interval between two successive conjunctions of the Moon with the vernal equinox. The nodical (or draconic) month (27.21222 days) is the time interval between successive passages of the Moon through the ascending node of its orbit.

See also: calendar, equinox, node, orbital elements, sidereal period, synodic period.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moon

The Moon, our closest neighbor in the universe, has a specific relationship with humanity. It is by far the brightest object of the night sky. The encounter of the Sun and the Moon during solar ECLIPSES strikingly demonstrates the similarity in the angular diameter of these two celestial bodies. As queen of the night, the Moon, together with the Sun, has been integrated into the system of symbols and religions of all early civilizations. Because of the obvious link with biological rhythms, the Moon was nearly everywhere associated with a goddess, the lunar 'yin' providing the balance with the solar 'yang'. It was given a particularly important role in Manichaeism (the 'religion of light', from the 3rd to the 14th century), departing souls migrating from the Earth to the Moon during the waxing phase before being transferred to the Sun during the waning phase. Countless pages have been written on the role of the Moon in art, literature, music and culture in general. Most ancient civilizations had CALENDARS which were either lunar (such as the Jewish or Muslim calendars) or lunisolar (such as the Roman calendar which evolved into the Gregorian calendar now widely in use). Because of the importance of calendars for agricultural societies and that of eclipses as portents of dire events, observations of the Moon and predictions of its motion were central to the early development of astronomy, possibly back to prehistoric times.

The Earth–Moon system is unique in the solar system with its ratio of less than 4 between the sizes of the planet and its satellite, and, as seen from afar by departing space probes, it can indeed be considered as a double planet (figure 1). Pluto with its satellite Charon (a factor of 2 in size) are now considered as the two largest members of the Kuiper belt of small outer solar system bodies, even though the International Astronomical Union has maintained the formal status of Pluto as the ninth planet. Because of the large mass of the Moon and its proximity to the Earth, lunar tides are nearly three times larger than solar tides. The orbit of the Moon is quite different from that of other large satellites (with the exception of Triton, around Neptune), which are nearly circular and lie in the equatorial plane of their primaries. One of the major scientific questions associated with the Moon is therefore that of the origin of the Earth–Moon system, which has only recently made significant progress.

In the first section, the history of lunar astronomy will be presented, concluding with the status of lunar science in the middle of the 20th century, before the beginning of the space era.

Figure 1. The double planetary system of the Earth and the Moon observed by the Galileo spacecraft departing to Jupiter. When compared with the Earth, this image demonstrates that the Moon, which looks so bright in the night sky, is actually quite dark.

The second section will focus on the role of the Moon as the major goal of the first 15 yr of space programs, with the pivotal date of 21 July 1969 when Neil Armstrong was the first human being to set a footprint on another planetary body during the historic flight of Apollo 11. After the last Soviet lunar mission, in 1976, the Moon was left alone for nearly 20 yr. The major breakthroughs in lunar science which resulted from this period of intense activity will be reviewed. The return to Earth of lunar samples has revealed the chronology of the geological history of the Moon, to the extent that, during its first billion years, it is better known than that of the Earth itself.

In the third section, the giant impact model for the origin of the Earth–Moon system will be presented. Its formulation in the early 1980s resulted from combining the wealth of information provided by the space missions with advances in dynamical modelling.

These new ideas contributed to the renewed interest in lunar exploration initiated by the American missions Clementine (1994) and Lunar Prospector (1998–99) and presented in the last section on the recent past and future prospects of lunar science. The future of lunar exploration is yet uncertain, with a series of increasingly ambitious missions which could lead to a permanent manned outpost on the Moon before 2030.

The Moon in the history of astronomy

The Moon was undeniably, together with the Sun, the prime target of NAKED-EYE ASTRONOMY, possibly as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

early as the stone age (astronomical relationships have been claimed for most major megalithic monuments, in particular Stonehenge) and definitely from the beginning of historical times, in Chaldea, Egypt, China or Mesoamerica. It was immediately evident that the Moon always turns the same side towards the Earth, which means that both its rotation period and its orbital period are tidally locked at 27.3 days. The variations of the orbital velocity along the slightly elliptical lunar orbit lead to an apparent back and forth rotation of the Moon as seen from the Earth which is called the libration. When taking into account the inclination of the rotation axis, slightly less than 60% of the lunar surface can be observed from the Earth. The familiar cycle of PHASES OF THE MOON (new Moon, first quarter, full Moon, last quarter) is linked to the angle between the directions of the Moon and the Sun, which repeats every 29.53 days (the lunar month). With the naked eye, one can readily distinguish the bright zones (the highlands) from the dark zones (the mare), the latter covering less than 35% of the visible regions. In central and eastern regions, the dark zones appear nearly circular, with diameters of up to 1100 km (Mare Imbrium). On the western side, they merge into the vast Oceanus Procellarum.

The Moon was the first celestial body observed by Galileo with a telescope in 1609, an event which marked the beginning of modern ground-based astronomy. He immediately identified countless round features, the lunar CRATERS. Forty years later, the major features of the near side had been revealed with the names which are still in use today (figure 2). Ray craters such as Tycho had been observed, with bright lines extending radially over thousands of kilometers. The nearly instantaneous disappearance of occulted stars demonstrated that the Moon has no atmosphere. Spatial resolution for lunar observations is essentially limited by atmospheric turbulence to a few km at ground level, down to 1 km at high altitudes. The Moon, a bright extended source, did not benefit from the race to larger and larger mirrors which was essential in the development of galactic and extragalactic astronomy. The major advances of the 19th century were the identification of tectonic features and lava flows on MARE.

Figure 2. Map of the visible face drawn by Grimaldi and Riccioli in 1651, a bare 40 yr after the first telescope observations by Galileo. The two circles mark the limits of the regions which can be observed owing to the lunar librations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Orbital and physical characteristics of the Moon.

Semimajor axis	384 400 km
Eccentricity	0.0549
Inclination on the ecliptic	5.15º
Radius	1740 km
Density	3.34
Mass ratio with respect to the Earth	1/81

The dynamical characteristics of the Moon were in the forefront of the development of celestial dynamics in the 18th and 19th centuries, with applications to astronomical navigation. The characteristics of the orbit of the Moon (table 1) were determined with an accuracy which made possible the development of the theory of perturbations. The influence of the Sun slowly modifies the position of the nodes (intersections of the lunar orbit plane with the orbit plane of the Earth) as well as that of the pericenter. The angular size of the Moon is 10% smaller at apogee than at perigee, where it can completely mask the Sun (total eclipse) if they are on the same line (hence with the Moon at one of the nodes). There is a remarkable triple coincidence between 223 lunar months, 239 returns to the pericenter and 242 returns to the ascending node, all within half an hour of 6585.33 days. This explains the repetition of eclipses with similar characteristics over an 18 yr cycle, with a shift in longitude of 120º (one-third of a day). The eclipse cycle is called the SAROS, a Chaldean word, even though it is now considered unlikely that it was recognized at such an early stage. Tidal dissipation within the Earth's mantle and oceans leads to a long-term evolution of the orbit of the Moon, which slowly drifts away from the Earth. The present and past rate of this long-term motion was a critical missing parameter when attempts were made to retrace this evolution backwards, so as to address the issue of the origin of the Earth–Moon system.

The major questions of lunar science were still open in the middle of the 20th century. Even the origin of ubiquitous features such as the lunar craters was a matter of controversy between proponents of volcanic and impact processes. Theories on the origin of the Moon covered all possibilities (or so it seemed at that time): ejection of a liquid 'droplet' from a rapidly rotating young Earth (fission theory), capture of the Moon by the Earth or accretion of the Moon from PLANETESIMALS orbiting the Earth. All three had critical flaws.

- Hydrodynamic instability of the Earth before the fission requires a total angular momentum of the Earth–Moon system twice as large as that which is observed and does not lead to extremely asymmetric partitions (1:81) as demonstrated by the nuclear fission 'liquid drop' analog (uranium splits into daughter nuclei with mass ratios of at most 2 to 1). Furthermore, the initial orbit would have been equatorial, with no plausible evolution path to the present orbit.
- No plausible dissipation process was proposed to shed the excess energy of a proto-Moon approaching the Earth, which constitutes a major difficulty for the capture hypothesis.
- Accretion in orbit, a smaller-scale version of the Kant–Laplace condensation theory, results in the so-called 'regular' satellite systems (such as that of Jupiter and Saturn), in the equatorial plane of their primaries and with very small eccentricities, which is not the case of the Moon. The low density of the Moon when compared with the Earth represented a major problem if these two bodies formed from the same reservoir of planetesimals.

There was also no information on the ages of lunar features. No transient events were observed, and, because of its small size and density, it was widely considered that the Moon cooled more rapidly than the Earth and was now a dead world. This hypothesis was to be proven correct by lunar sample geochronology.

The Moon and the early phases of exploration of the solar system

The Moon, with its proximity and public appeal, was the obvious first target for spaceflight beyond low Earth orbit. Indeed, only 2 yr elapsed from the first Sputnik to the Soviet lunar probe which crashed on the lunar surface. A major breakthrough was obtained with the first lunar fly-by by Lunar 3, in 1959, which revealed a far side almost entirely covered by highlands (figure 3). The pace of events became quite frantic after the decision by J F Kennedy in 1961 to land a man on the Moon in less than 10 yr. This initiated the 'race to the Moon' as a way to demonstrate the American technological leadership. Only 3 yr separate the first soft lander (Luna 9, a Soviet mission, in 1966) from the historic landing of Apollo 11, in July 1969 (table 2). In 1972, the last of the six Apollo flights left the Moon, and this intensive phase of lunar exploration came to a halt in 1976, when the last automated Soviet sample return mission left Mare Crisium. The science results from this gigantic effort have provided a quantum jump in our understanding of the Moon and its history.

A global photographic coverage was obtained with polar orbiters, with high-resolution images of the potential landing sites. This demonstrated the impact origin of most, if not all, lunar craters, from the similarity to explosion craters and the circular 'ripples' around major basins such as ORIENTALE BASIN. The observation of craters at all scales from basins down to a few micrometers explained the formation of a ubiquitous layer of debris 5–10 m thick, the lunar REGOLITH (figure 4), mainly constituted of very small grains (a few 10 μm)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

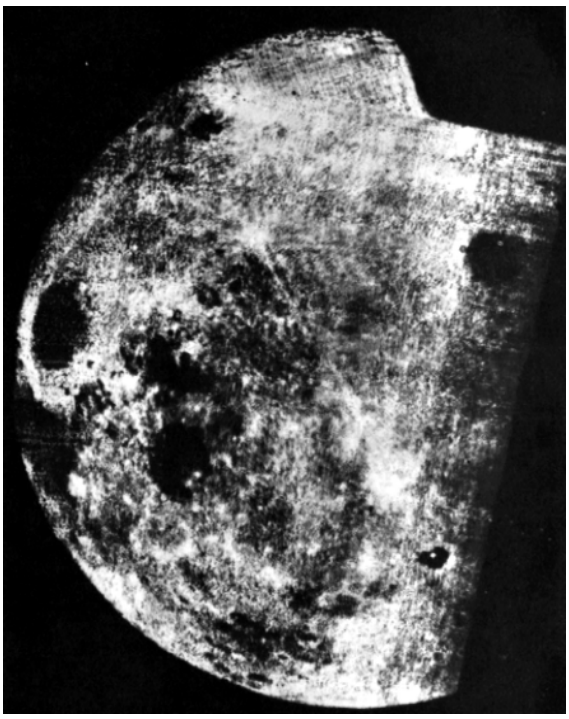


Figure 3. This historic image obtained by the Luna 3 probe in 1959 revealed a far side dominated by highlands (bright regions).

Table 2. Main lunar exploration missions from 1959 to 1976.

Luna 3 (USSR)	October 1959	First images of the far side
Luna 9 (USSR)	January 1966	First soft lander
Lunar Orbiter 4 (USA)	May 1967	Photocartography from a polar orbit
Surveyor 1, 5, 6 (USA)	1966–68	Soft landers
Lunokhod 1, 2 (USSR)	1970, 1973	Remotely controlled rovers (10–40 km)
Luna 16 (USSR)	September 1970	Automated sample return (Mare Fecunditatis)
Luna 20 (USSR)	February 1972	Automated sample return (Apollonius highlands)
Luna 24 (USSR)	August 1976	Automated sample return (Mare Crisium); 2 m deep drill core
The Apollo program (USA)		
Apollo 11	July 1969	First manned landing (Mare Tranquilitatis)
Apollo 12	November 1969	Oceanus Procellarum
Apollo 14	January 1971	Fra Mauro formation (SE of Procellarum)
Apollo 15	July 1971	Hadley (Mare–highland contact zone); First deep drill core, first manned rover (20 km)
Apollo 16	April 1972	Descartes (highlands); deep drill core, rover
Apollo 17	December 1972	Taurus–Littrow (Mare–highland contact zone); field geology (H Schmitt), 35 km explored, deepest drill core (3.2 m)

but also including rocks of all sizes. The accurate determination of the gravity field and altimetry over the low-latitude regions combined with the seismic active and passive experiments provided information on the internal structure of the Moon: either no core or a very small core (at most a few hundred km in radius), a thick mantle and then a crust which is much thinner on the near side (60 km) than on the far side (100 km). The minor contribution of the core to the total mass means that the Moon has about half as much iron as the Earth, a major constraint on formation models. Positive gravity anomalies below basins, the mascons, confirmed the impact origin hypothesis. Magnetometers revealed that the Moon had no dipolar field but observed local crustal fields with widely varying orientations. The origin of these fossil fields is still controversial (frozen ancient dipolar field, induction) and this question is linked to the existence and actual size of a possible lunar core.

The major justification for the large Apollo budgets (up to several per cent of the US GIP in the peak years) was definitely political with military undertones. The goal of the first Apollo flights was quite literally to display the flag, not to do science, and the first astronauts had little or no scientific training. As a result, many samples from Apollo 11 and 12 were collected without adequate documentation, such as taking a photograph before picking up a rock, to know which face was exposed to space. After the political goals had been

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

reached, and after the scientists pointed out the consequences of mishandling the sampling process, a science team was put together for the Apollo 13 flight to instruct and support the astronauts. Of course, this mission is more widely known for its dramatic rescue effort after an explosion in the service module half-way to the Moon than as a breakthrough for the role of science in the program. The last four Apollo flights were science oriented, with much larger and better-documented sets of samples. Apart from sampling, the astronauts performed a wide range of scientific investigations, including the first European space experiment, a solar wind collector. The set of instruments left on the surface remained operational years after the departure of the astronauts. One can quote the important results from seismometers on the internal structure or the heat flux experiment, which provided the global content of long-lived radionuclides (potassium, uranium and thorium) in the lunar interior after nearly 4 yr of integration to screen out the diurnal and annual cycles.

Overall, nearly 400 kg of samples were obtained from nine different sites: the six Apollo sites, but also three automated Soviet sample return missions, which brought back a few hundred grams of samples, a technological feat which is only now being planned for Mars. Lunar sample analyses played a key role in a revolution in analytical techniques which not only benefited planetary sciences but had important technological spin-offs: textural, elemental and isotopic analyses which required grams were performed on micrograms with the development of the electron microscopes, ion probes, mass spectrometers and x-ray analysers. When applied to METEORITES, these new techniques revealed the isotopic anomalies which provide the basis for the most recent theories of the formation of the solar system.

The lunar geologic history

The determination of an absolute age scale for the lunar geologic history was the most important result of sample analyses. Chronometers were developed on the basis of the radioactive decay of uranium to lead, rubidium to strontium and potassium to argon. Argon is a gas, so that the latter age fixes the last shock event in the history of the sample, which can be different from the age of formation of the rock. These analyses demonstrated that the formation of the Moon was nearly simultaneous with that of the Earth, 4.55 billion years ago. After at most 200 million years, the differentiation between a possible small core, a mantle and a crust was complete. Over nearly 500 million years, the Moon (and hence also the Earth) was submitted to an intense bombardment as the newly formed planetary bodies swept out the left-over planetesimals in the inner solar system. This period resulted in the formation of giant impact basins, hundreds of km in diameter and tens of km in depth, with fractures in the underlying crust, while the surface was thoroughly pockmarked by overlapping craters. All the 'mountains' on the Moon (up to 8 km high) result from the accumulation of debris or the rebound of the lunar crust associated with the formation of basins. The far side, still dominated by highlands, gives a good idea of what the whole lunar surface looked like 4 billion years ago (figure 5).

In the meantime, heat from radioactive decay was accumulating slowly in the lunar interior. From 4 to 3 billion years ago, melted mantle material was able to escape through the fractures of the crust as gigantic lava flows to fill out the basins and form the lunar maria. This was much easier on the near side than on the far side, where the crust was twice as thick, which explains the near absence of maria on the far side. As on the Earth, the lava flows from the mantle are much darker than crustal material, which explains why mare are darker than highlands (which are already rather dark, with a maximum albedo of 20%). The cooling of lava flows led to the formation of specific features such as faults, domes and valleys (figure 6).

Another important result of high-precision isotopic analyses was the discovery that the Earth and Moon were formed from a single isotopic reservoir, different from that of the parent bodies of meteorites. The minerals constituting the Moon also show similarities to those of corresponding units on the Earth. However, the content of volatile elements is much lower: hydrogen is nearly absent (the Moon is most likely the driest place in the solar system) and potassium, abundant in terrestrial granites, is replaced in the lunar crust by calcium, a much more refractory element.

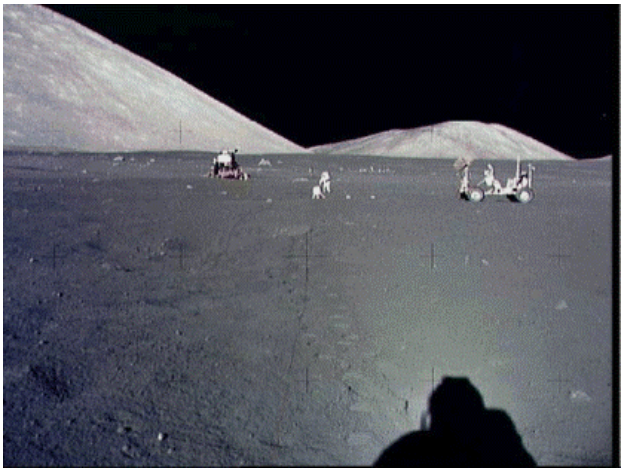


Figure 4. The Apollo 17 landing site. The fine-grained lunar regolith has a very smooth relief and contains only a small fraction of rocks. The footprints of the astronaut (forefront) will remain visible for millions of years. The tiny figure of the astronaut in this profoundly alien landscape is a striking illustration of the magnitude of the Apollo endeavor as well as the true scale of humanity in the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

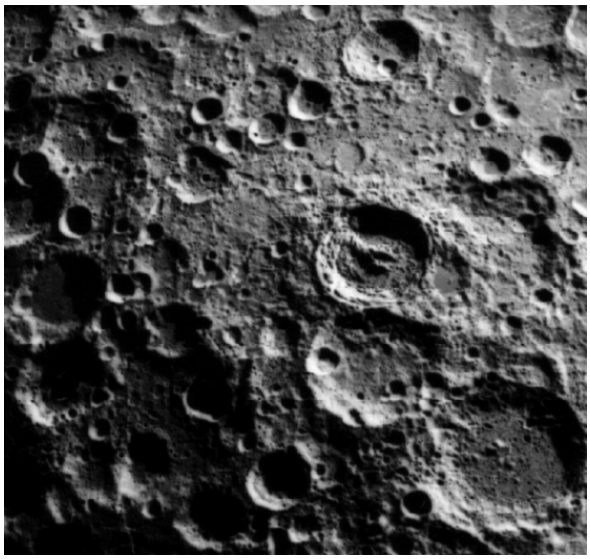


Figure 5. The highlands (here on the far side) have a chaotic landscape resulting from superimposed craters of all sizes. Complex structures such as terraces and central peaks can be observed in the youngest of these craters, such as crater King (right of center, with the lobster-shaped central peak).

Figure 6. The Apollo 15 landing site lies at the eastern edge of the Imbrium basin. Lavas from the center of Imbrium (upper left) filled out the basin up to the first concentric ridges of the rim of the basin (lower right). The sinuous Hadley rille (valley) most likely resulted from the collapse of an oversize lava tube. Tectonic features such as faults were formed by the cooling and relaxation of the lava fills.

Over the last 3 billion years, the role of internal processes was very small. At that stage, there were no planetesimals left in the inner solar system. Later impacts resulted from a small steady-state influx of impactors from the ASTEROID BELT and the KUIPER BELT due to gravitational perturbations. The surface of lunar maria has therefore remained very flat (the mean slope is only a few degrees). The main outcome of this secondary bombardment has been the formation of sparse fresh craters (the most recent large one is TYCHO, a few hundred million years old) and that of the lunar regolith from the slow grinding effect of small micrometeoroids. The relative age of surface units can be determined from crater counts, calibrating on the age of the nine sampled regions. The resulting evolution is extremely slow: the footprints left by the Apollo astronauts may well outlast humanity with a lifetime of a few million years. The absence of an atmosphere means that the main alteration effects of the surface result from the implantation of solar particles and micrometeoroid impacts. This slow evolution is called 'maturity', with an increase in glass content (up to 50%) and a decrease of the albedo. Rays are formed by the bright fresh material thrown out by recent craters and they slowly fade out as maturity increases.

The mare regolith constitutes an invaluable source of information on the history of dust and particle fluxes in the inner solar system. As there is no atmosphere, the surface is directly exposed to the interplanetary medium. The solidifying lavas provided a clean plate more than 3 billion years ago, which registered the effects of particle fluxes and micrometeoroids in a dust layer only 5–10 m thick. The 3.2 m deep drill core obtained during the Apollo 17 mission gives access to 1.7 billion years of history of the interplanetary medium (solar activity, asteroidal and cometary dust, galactic cosmic rays). For comparison, a 4 km deep ice core in the Antarctic corresponds to a few 100 000 yr of history. This unique lunar record should be fully exploited in the future with dedicated new cores extending to the base of the regolith.

The origin of the Earth–Moon system

The tremendous leap in knowledge of the Apollo period did not immediately lead to a widely accepted model for the formation of the Moon. The similarity between the Earth and Moon in terms of trace element and isotopic compositions was considered by geochemists as overwhelming evidence in favor of the fission hypothesis. However, laser echoes using corner reflectors on surface modules allowed the recession rate of the Moon to be measured with precision (4 cm yr^{-1}). Isotopic analysis techniques were also used to follow the growth pattern of corals over time, which constrains the number of days and lunar months per year in the distant past. Integrating this evolution backwards in time led to a highly elliptic and inclined orbit, totally incompatible with fission.

The major breakthrough occurred in the early 1980s:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- the development of chaotic dynamics showed that the orbits of planetary bodies can be stable for millions of years and then evolve rapidly as a result of perturbations by Jupiter and Saturn, hence making possible giant collisions in the early solar system;
- 3D hydrodynamic simulations which became possible with supercomputers demonstrated that a likely outcome of the off-center impact of a Mars-sized body with the Earth was the ejection of part of the mantles of the impactor and the proto-Earth into orbit, while the iron-rich cores fused.

This 'giant impact' model combines elements of all three of the original models: It involves a body formed somewhere else in the solar system (capture), then ejection of part of the outer layers of the transiently merged impactor and proto-Earth (fission), then accretion in orbit of the ejected material. When compared with the fission model, the giant impact model leads naturally to a highly inclined and eccentric early orbit, providing for the first time a dynamical solution consistent with the geochemical evidence. A few years later, numerical models for the condensation of the solar nebula resulted in 20 or more large planetary bodies, so that giant impacts are actually needed to trim down the number of planets. The giant impact model is now widely accepted, even though numerous details have to be sorted out (e.g. possible intermediate stages with smaller objects orbiting the Earth). This is a true Sherlock Holmes outcome, the far-fetched becoming plausible after all the obvious candidates had been eliminated.

Recent simulations (Canup and Asphaug 2001) of potential moon-forming impacts have identified a class of impacts that yield an iron-poor Moon, as well as the current masses and angular momentum of the Earth–Moon system. This class of impacts involves a smaller—and thus more likely—object than previously considered viable, and suggests that the Moon formed near the very end of Earth's accumulation, in contrast to previous simulations that required the impact to occur when the Earth was only about half formed.

Chaotic dynamical behaviors have another important implication for the Earth–Moon system: the inclination of the rotation axis of the Earth could increase to 60º or more if the Moon was not there to stabilize its evolution, with devastating consequences for the weather, each pole alternately facing the Sun. It can therefore be argued that the apparition of life on our planet required the presence of a large satellite such as the Moon.

The renewed interest in the Moon and the future of the lunar exploration program

During nearly 20 yr, the Moon was left aside by PLANETARY EXPLORATION. This led to a paradox: the Moon is indeed the only planetary body other than the Earth for which the complete sequence leading from reconnaissance to complete mapping, surface stations and rovers and eventually sample returns had been achieved. However, the very short timescale and the focus on a manned landing resulted in major gaps in the remote sensing and sampling programs, with a major bias in favor of the maria in the central regions of the near side, the most logical landing sites. Only two of the nine sampling sites (Apollo 16 and Luna 20) are in highlands, which cover nearly 80% of the lunar surface. For obvious communication reasons, all the Apollo sites are close to the center of the near side, and the highest latitude which was reached is 25º (Apollo 15). Only a few per cent of the surface were imaged at high resolution or were observed with chemical and mineralogical remote sensing analysis techniques. Furthermore, these techniques made major advances from 1970 to 1990, with improved detectors in the infrared, x-ray and gamma-ray regions of the electromagnetic spectrum.

The developments on the origin of the Moon confirmed the need for a better understanding of the global composition, a critical parameter for fine-tuning the giant impact model. This requires determining the contribution of a possible core to the total iron content and the chemical and mineralogical composition over the whole surface. Constraints can then be obtained by running petrological models which test how a given initial composition can lead to the observed abundances of minerals.

This renewed scientific interest benefited from two factors. By the early 1990s, missions to other planetary bodies had reduced the gap in knowledge, so that the Moon no longer appeared as the privileged child of the solar system family. More importantly, there was a major shift of policy within NASA, from highly complex and costly programs such as Galileo (Jupiter) and Cassini–Huygens (Saturn and Titan) to much smaller and more frequent missions, which could benefit from the major advances in miniaturization resulting from the 'Star Wars' program. The Moon was an ideal target for this so-called 'faster, better, cheaper' approach, with its small launcher requirements, short travel time and easy communications with the Earth. Similar strategies have been initiated by the European Space Agency (ESA) and Japan (unfortunately, the collapse of the Soviet Union led to a stand-by of the Russian planetary exploration program).

The lunar exploration program restarted on this new basis in 1994, with the CLEMENTINE mission (USA), which was to a large extent a spin-off of a militarily oriented technological effort. This small satellite (100 kg) carried three cameras with 11 channels from the visible to the near infrared and a laser altimeter. Its scientific return is quite significant. Clementine discovered the largest lunar basin, Aitken, covering most of the southern far side (figure 7). It provided the first multicolor database

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

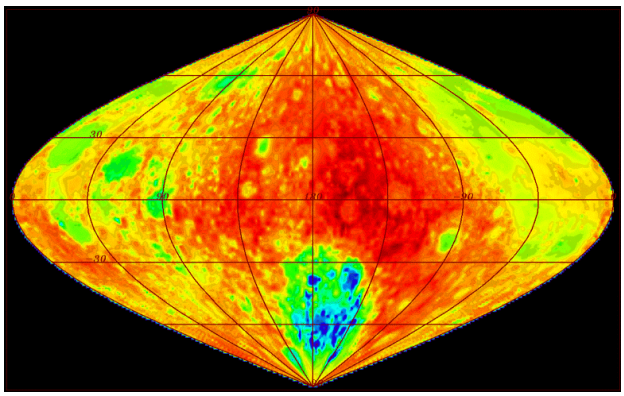


Figure 7. Topographic map of the Moon obtained by the laser altimeter of Clementine. Altitudes range from −12 km (dark blue) to +10 km (red). This map is centered on the far side (longitude 180º), which has a higher elevation and thicker crust than the near side. These observations revealed the largest lunar basin, Aitken, 1800 km in diameter, covering a large fraction of the southern far side. The thick far side crust prevented lavas from filling up this enormous basin, contrarily to similar features on the near side.

for the whole lunar surface, showing that the 'dry' basins of the far side were slightly contaminated by basaltic minerals: lava flows have indeed occurred, but were not quite able to reach the surface, expanding within and below the crust. Clementine also observed polar regions which are always in shadow, providing possible cold traps for volatile species such as water.

A second small polar satellite, LUNAR PROSPECTOR, was launched in early 1998. It provided the first complete coverage with a gamma-ray and neutron spectrometer as well as a state-of-the-art magnetometer. The observation of hydrogen trapped near the lunar poles may be linked to deposits of ice crystals in the permanently shadowed polar craters, although this interpretation is still controversial. If this is confirmed (e.g. by radar sounding) it opens important scientific prospects: such water deposits would then be likely to be provided by cometary impacts on the Moon, so that a core in the polar regolith could give access to the isotopic composition of volatiles from thousands of comets. In the final stage of the Lunar Prospector mission, the orbit was lowered to an altitude of 10 km, which gave unprecendented information on the gravity field. At this altitude, the orbit is not stable, and the inevitable impact was targeted near the south pole. This failed to give direct evidence for water deposits, but the expected signal was extremely marginal.

These discoveries showed that the Moon has quite a few surprises still in store. The next approved mission is LUNAR A, a Japanese mission with two penetrators equipped with seismometers dedicated to the question of the lunar core. This mission was originally planned for 1996, but development problems have delayed it until 2001 or 2002. The Japanese are also planning a very ambitious mission, SELENE, which was scheduled for 2003, but the problems with Lunar A may lead to a shift to a later date. A small ESA mission, SMART-1, will be launched in 2002, mainly as a technological test for a new ion propulsion system. It will provide new high-resolution images, the first comprehensive coverage with an x-ray spectrometer, a chemical analysis instrument which has a better spatial resolution than the gamma-ray spectrometer of Lunar Prospector, and the first medium-resolution IR spectra from lunar orbit, a key constraint on the mineralogical composition.

No discussion of the future of lunar exploration would be complete without mentioning the lunar base issue. The Moon is the obvious stepping stone for human exploration of the solar system, and countless studies have been made on establishing and maintaining a permanent habitat on the lunar surface. Such an endeavor is within reach of present technology, but it requires financial commitments in the scale of those of the Apollo program or the International Space Station. Similarly to these two programs, a lunar base can only be decided on a political basis, with science as a fringe benefit. Many scientific objectives can be reached much more cheaply by automated missions, with capabilities enhanced by the advances in on-board autonomy. The main potential benefit of a lunar base is a new lunar sampling program from medium to high latitudes, highlands or the far side, with deeper drill cores (10 m) giving access to 4 billion years of history of solar activity and the interplanetary medium. Samples from permanently shadowed polar regions would provide the final answer to the question of polar ices, so as to address the diversity and evolution over time of cometary impacts in the inner solar system. The timescale for a lunar base program is still highly uncertain. However, it would be surprising if it was not initiated before 2030, if space exploration worldwide is to retain any momentum.

Bibliography

Canup R M and Asphaug E 2001 *Nature* **412** 708–12

Murray C and Cox C B 1989 *Apollo: the Race to the Moon* (New York: Simon and Shuster)

Spudis P 1996 *The Once and Future Moon* (Smithsonian Institution Press)

Wilhelms D E 1993 *To a Rocky Moon: a Geologist's History of Lunar Exploration* (Tucson, AZ: University of Arizona Press)

Yves Langevin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moon Maps

The MOON is our closest natural neighbor in space. Its mean distance from us is a mere 239 000 miles, and it has to all intents and purposes no atmosphere, so that there is nothing to veil its surface. Even with the naked eye a great deal can be seen, and it is easy to make out the broad dark plains, still known— misleadingly—as seas (maria). So far as is known, the first map of the Moon was drawn early in the 17th century by WILLIAM GILBERT, physician to Queen Elizabeth I, who is best remembered today as a pioneer investigator of magnetic phenomena. Of course he had no telescope, but the main maria can be identified easily enough.

The next map—and the first to be produced with telescopic aid—was drawn in 1609 by THOMAS HARRIOT, in England. Harriot, one-time tutor to Sir Walter Raleigh, was one of the first to use a telescope for astronomical work, and certainly he preceded Galileo, even though he never followed it up (he died in 1621). Harriot showed not only the principal maria, but also some of the craters and mountain ranges. One of his contemporaries was Sir William Lower, a Welsh baronet who seems to have obtained one of the early telescopes and looked at the Moon. His description of the lunar world was certainly graphic: he compared it with a tart which his cook had made—'there some bright stuffe, there some dark, and so confused lie all over'.

GALILEO began his telescopic work in the winter of 1609–1610, and used his primitive optick tube to make a whole series of dramatic discoveries. The Moon, he wrote, was 'full of inequalitie, uneven, full of hollows and protuberances, just like the surface of the Earth itself, which is varied everywhere by lofty mountains and deep valleys'. Frankly, his lunar map was not as good as Harriot's, but he did make a scientific effort to measure the heights of the lunar mountains, and his results were of the right order.

The next major advance came in 1647, and was due to HEVELIUS, a city councillor of Danzig (the modern Gdańsk) in Poland. Hevelius built an observatory on the roof of his house, and equipped it with the best telescopes available. He compiled a lunar map, showing considerable detail, and made the first attempt to name the lunar features. For example, the crater which we now call Copernicus was his 'Etna', while the dark-floored crater Plato was 'the Greater Black Lake'. The system never became popular, and was soon superseded by that of Giovanni RICCIOLI, who drew a new lunar map in 1651—based largely on the observations of his pupil, Grimaldi—and introduced many of the names which we still use today. Riccioli's method was to name the craters after famous personalities, usually—although not always—astronomers; Hevelius' geographical names were largely replaced, and only a few survive (such as the Alps and the Apennines). Riccioli gave the dark areas romantic titles, such as the Mare Imbrium (Sea of Rains) and the Mare Tranquillitatis (Sea of Tranquillity).

Obviously, the system had its drawbacks. Riccioli had no faith in the Copernican system and was convinced that the Earth is the center of the universe; therefore Copernicus was 'flung into the Ocean of Storms', although admittedly his crater is one of the most magnificent on the Moon and is the center of a system of bright rays. Predictably, Riccioli named two very large, dark-floored plains after Grimaldi and himself. Later scientists came off second-best; Newton, for example, was allotted a crater in the far south which is both large and deep, but so foreshortened that from Earth it is none too easy to identify at all.

The next important lunar map was that of TOBIAS MAYER, published posthumously in 1775. It was much smaller than that of Riccioli, a mere 8 in across, but it was reasonably accurate, and Mayer was the first to introduce a system of coordinates, the equivalent of latitude and longitude on the Moon. Unlike Hevelius, Riccioli and most of his predecessors he drew his map with south at the top, which is the way in which most telescopic observers see it—provided, of course, that they live in the northern hemisphere of the Earth.

Mayer was followed by JOHANN HIERONYMUS SCHRÖTER, who was never a professional astronomer—he was chief magistrate of the little town of Lilienthal, near Bremen in Germany—but may be said to be the first of the really great lunar observers. He made a large number of drawings and was responsible for many discoveries; for example, he was the first to show the lunar rills or clefts clearly, and his measurements of the heights of lunar mountains were better than Galileo's. Unfortunately, all his unpublished observations were destroyed in 1814, when the French invaded Germany, sacked Lilienthal and plundered Schröter's brass-tubed telescopes, which they believed to be gold.

The mantle of Schröter fell upon three of his countrymen: WILHELM LOHRMANN, WILHELM BEER and JOHANN VON MÄDLER. Wilhelm Lohrmann, a Dresden land surveyor, set out to compile a really large, accurate lunar map. The first sections were published in 1824, and were amazingly good. Unfortunately his eyesight failed him, and he had to give up.

Next came the superb work by Beer and Mädler. Using the modest $3\frac{3}{4}$ in Fraunhofer REFRACTOR at Beer's private observatory, near Berlin, they studied the Moon intensively for over 10 years, finally producing a map which was the basis for all later studies right up to the start of the space age. They followed it up with a book, *Der Mond*, which contains a full description of the entire visible surface and extended Riccioli's nomenclature. It appeared in 1838, and made a tremendous impact—and, surprisingly, actually held selenography up to some extent. Their map was regarded as definitive; the Moon is virtually changeless, so why observe it further?

Luckily there was one astronomer, Julius Schmidt, who did not agree. He was German by birth, but became Director of the Athens Observatory, and spent much of his life in Greece. He not only revised and completed Lohrmann's map, but also produced one of his own. It

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

was 74 in across, and stands up well when compared with modern charts.

During his observational work, Schmidt made the startling announcement that a small formation, Linné in the Mare Serenitatis (Sea of Serenity) had changed in form since the time of Beer and Mädler; they had described it as a small, deep crater, but Schmidt drew it as a white patch. The 'Linné affair' caused a great deal of discussion. In fact it seems quite definite that no change occurred there, but attention was drawn back to the Moon, and other charts were drawn up. The first important British chart was by Edmund Nevill, who wrote under the name of Neison. His book *The Moon*, published in 1876, contained a map based on Mädler's, with a description of every named formation. Neison was only 25 years old when his book appeared; curiously, he did very little more lunar work after 1882, when he went to South Africa to become director of the new observatory at Natal. Neison died in 1938, in retirement at Eastbourne in Sussex.

Systematic lunar observation was carried out around this time by the short-lived Selenigraphical Society, and, from 1890, by the Lunar Section of the British Astronomical Association. Meanwhile, photography was coming to the fore. Good photographs were obtained by pioneers such as J A Whipple and Warren de la Rue, and in the last decade of the 19th century Loewy and Puiseux, in Paris, produced an excellent series of plates covering the whole of the visible surface; subsequently, all lunar mapping was photographically based. In 1904 W H PICKERING, one of the few professional astronomers of the time to take a real interest in the Moon, issued a lunar atlas in which each region of the Moon was shown under five different conditions of illumination.

Amateur work flourished during the first half of the 20th century, with good maps by T G Elger and Walter Goodacre (successive Directors of the Lunar Section of the British Astronomical Association) and then by H P Wilkins, whose map was 300 in in diameter. In 1935 an 'official' map was published by the International Astronomical Union, drawn by W H Wesley and M A Blagg, but it was not a real success, and was defective near the Moon's limb, where the 'libration regions' are highly foreshortened. The obvious next step was to produce a complete photographic chart, similar to Pickering's but on a larger scale. This was attempted by G P Kuiper and his colleagues in the United States; it appeared in 1959. It was based on the best photographs available, but even so it did not meet all the requirements, and over-enlargement led to blurring of many features, particularly those near the lunar limb.

Then, after 1957, space research methods began to take over, and this caused a minor upheaval in procedure. Previously lunar maps had been printed with south at the top and west to the left (so that the well-formed Mare Crisium was near the western limb). The rocket men wanted to put north at the top. Finally, there was a debate in the International Astronomical Union, and the 'north-toppers' won the day; at the same time east and west were reversed, so that Mare Crisium is now near the eastern limb[1].

The first images of the Moon's far side, which is always turned away from the Earth, were obtained by the Russian spacecraft Lunik (or Luna) 3 in October 1959. Blurred and lacking in detail though they were, judged by modern standards, they represented a tremendous technical triumph. Craters, mountains and ray systems were detected, although one feature believed to be a lofty range, and even given an appropriate name (the Soviet Mountains) turned out to be nothing more than a ray, so that the name was tactfully deleted from later charts.

Once the Americans had announced plans for a manned landing before 1970, energetic efforts at mapping the surface were made. The Ranger probes crash landed, sending back pictures immediately before impact; the Surveyors made controlled landings, and these results, together with those of the Russian soft-landers, finally disproved the theory that the lunar surface was covered with a deep layer of dust. Then, between August 1966 and early 1968, the five United States Orbiters were put into closed paths around the Moon and sent back detailed photographs of both the Earth-turned and the averted hemispheres. By the end of the Orbiter program it was fair to say that the entire Moon had been surveyed in considerable detail.

Since then there have been several photographic probes, notably Clementine (launched January 1994) and Prospector (January 1998). These covered the polar regions of the Moon, which had not been mapped as accurately as the rest of the surface, and also sent back vast amounts of miscellaneous data. Of course, there have also been the observations made by the Apollo astronauts—and only the crews of Apollo 8 and 10–17 have actually had direct views of those regions which are always turned away from the Earth.

The one negative factor is that for various reasons—some of them financial—not all the Orbiter and Apollo photographs have yet been fully analysed. Much work remains to be done. There is, of course, no secrecy; all the images are made freely available to serious researchers—it is purely a matter of time and money.

The story of lunar mapping goes back a long way. It may be said to have started with the naked-eye chart by Gilbert, almost 300 years ago; then came the pioneer telescopic work, and the more detailed maps of the 18th and 19th centuries, many of them due to the work of amateurs; there followed the photographic atlases and, finally, the definitive maps drawn up from the spacecraft images.

It is interesting to look back at a comment made in 1840 by FRANÇOIS ARAGO, the leading French astronomer of the time. Arago was, very understandably, enthusiastic about photography, and said 'In this way we will be able to

[1] Piquantly, the Mare Orientale—the Eastern Sea—is now on the western limb. I have a connection here, because I was the first to show it—in 1939—and also proposed its name—PM.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

accomplish one of the most difficult tasks in astronomy—mapping the Moon—in a few minutes'. Arago's timescale was wrong; we had to wait until the last part of our own century, but by now charting the Moon is to all intents and purposes complete.

Bibliography

Whittaker E A 1999 *Naming and Mapping the Moon* (Cambridge: Cambridge University Press)

Wilhems E 1993 *To a Rocky Moon* (Tuscon, AZ: University of Arizona Press)

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moore, Francis (1657–*c*. 1715)

Astrologer, born in Bridgnorth, Shropshire. In 1700 he began to publish *Vox Stellarum (Voice of the Stars)*, which under the name *Old Moore's Almanac* still attempts to predict annually the events of the following year.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moore, Jonas (1627–79)

Born in Whitelee in Pendle Forest, Lancashire, became mathematics tutor to the Duke of York, brother of the future King Charles II, but lost this job through court intrigue; lived in London as a teacher of mathematics, and as a successful surveyor. He published text-books, and became Surveyor-General of the Ordnance, when he was knighted. His own career was thus more worthy than brilliant. Where he enters the history of astronomy as a first magnitude star is in his strong support of mathematics and astronomy. It was his support that got FLAMSTEED not only astronomical equipment and a degree from Cambridge University, but also an invitation to report as an expert witness to a commission, and to the King himself, on recently proposed methods to determine longitude. Moore then led the delegation that persuaded the King to set up the Royal Observatory at Greenwich in 1674 and to give Flamsteed the job as its first director. Moore provided Flamsteed with two clocks and a 7-foot sextant, which he used until 1688.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moore, Patrick Alfred Caldwell (1923–)

Colorful British amateur astronomer, prolific writer, ubiquitous broadcaster and enthusiastic musician. Since 1957 he has presented monthly the BBC television program *The Sky at Night*, the longest-running such program with the exception of the news. He has published approximately one book per year of his life. His astronomical work has been in observations of the Moon and planets; his greatest impact has been inspiring a generation of British children to astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moreton Waves

'Moreton waves', named for the observer who popularized them, are a solar phenomenon also known in scientific literature as 'Moreton–Ramsey waves', 'flare waves', 'flare-associated waves', 'MHD blast waves', 'chromospheric shock fronts' and various other combinations of terms which connote violently propagating impulsive disturbances. It is unclear whether all of the observations to which these terms have been applied pertain to a single physical phenomenon: there has perhaps been some overlap between the observations and the assumed physical properties of the observed occurrence.

Moreton waves (figure 1) are ideally observed in the wings of H alpha and appear as semicircular fronts propagating at speeds ranging from several hundred to over a thousand km s^{-1}. They form an arc, or 'brow shape', which can span up to 180°. Extrapolating the speed and locations of the arc indicates that the phenomenon's origin intersects well with the impulsive phase of the associated H alpha flare (if the flare exhibits an impulsive phase) (see SOLAR FLARES: IMPULSIVE PHASE). However, the arc may not form or may not be observable until it is tens of megameters from the flaring region, and subsequently can propagate to distances exceeding 100 Mm. The high speeds and distances of propagation, plus the associated radio and energetic particle observations, provided strong evidence of a coronal, rather than a chromospheric, origin. The H alpha manifestation of the wave is assumed to be the 'ground track' or 'skirt' of a three-dimensional disturbance.

The red- and blueshifts of the front propagating on the solar disk into the wings of H alpha appear to indicate evidence of chromospheric depression (and subsequent relaxation) as the wave propagates. The thin arc-shaped structure of the wave, the assumed force which would be required to cause depression in chromospheric features and their association with type II radio bursts (see SOLAR FLARES: RADIO BURSTS), led observers to conclude that these were shock waves or 'blast waves' initiated by the flare.

The 'ideal' observations of Moreton waves, however, may not represent all of the actual events observed. Although the 'depression' interpretation fitted some observations well, some other 'wave' observations did not exhibit strong Doppler shifts in H alpha, and some of these 'waves' were best observed in emission at the H alpha line center. Furthermore, some observations consisted of additional emission superposed on pre-existing structures, instead of the temporary enhancement of a feature due to a compressional disturbance propagating through it. It was concluded that these additional observations were best explained as ejecta moving above the CHROMOSPHERE. The ejecta, which could be SOLAR PROMINENCES or other moving material, could display speeds and structure comparable with those of the prototypical Moreton waves. Had this not been the case, perhaps these observations would have been considered as a phenomenon separate from Moreton waves. However, the wave observations consisting of ejecta and those consisting of depression fronts could also be two aspects of the same phenomenon, and some observations indicate this. The shock wave front could be observed in emission while its impact on the chromosphere could be seen in the depression of chromospheric features.

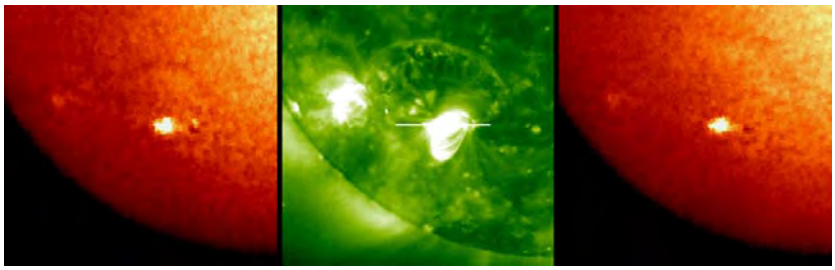
Several other observations have also been tentatively classed as Moreton waves. Before direct observations of wave disturbances were obtained in 1960, observers determined that there was evidence that a flare could initiate another flare or the abrupt activation or oscillation of a filament. Radio observations and the eventual observation of Moreton wave transients in the CORONA suggested a means of communication between the causal event and the affected region. Therefore, sudden filament activity, additional flaring and transient chromospheric brightenings have been identified as wave-associated effects. The inferred speeds of a wave transient (frequently lacking the actual observation of a wave), propagating between a flaring region and an initiated observable phenomenon, could exceed 4000 km s^{-1}. In the absence of supporting wave or radio burst observations, some of these measurements could fall under suspicion of being coincidental and not associative.

Recent observations by the Extreme Ultraviolet Imaging Telescope (EIT) on the SOHO spacecraft have provided unambiguous evidence (figure 2) for large-scale coronal impulses initiated during the early stage of a flare and coronal mass ejection. These waves are typically of lower amplitude than the Moreton wave shown in figure 1, can extend 360° surrounding the initiating region and usually propagate passively through the corona, leaving no observable signs of alteration. It is unclear whether these impulses observed by EIT correspond to the low-amplitude (and hence unobservable in H alpha) limit of Moreton wave phenomena or whether these waves have a different nature entirely. Other evidence of impulsive wave phenomena was recorded in white light coronagraph data as early as Skylab, and observations from the Mauna Loa Solar Observatory have indicated fast-moving disturbances moving in excess of several hundred km s^{-1} in the 10 830 Å He I data.

A clearer understanding of the various phenomena which have been classified as Moreton waves relies on more thorough observations and an increased understanding of their initiation mechanism. Solar flares and coronal mass ejections are magnetic reconfigurations utilized by the solar corona to liberate excess magnetic energy (see SOLAR FLARES: RELATION TO CORONAL MASS EJECTIONS). In expelling flux and mass via a coronal mass ejection, however, the corona frequently undergoes an additional reconfiguration in the form of a solar flare. For this reason, it is extremely difficult to extricate the actual driver or cause of Moreton waves.

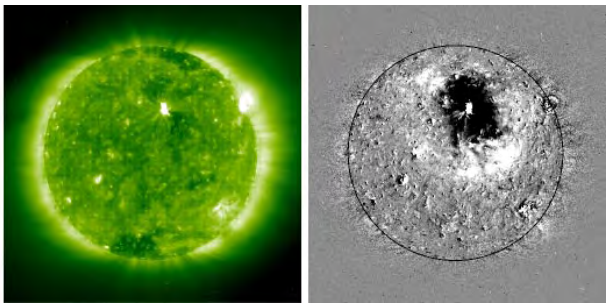
A coronal mass ejection will deliver an impulse to the surrounding corona as the ejection propagates and expands. Some of the SOHO observations have given

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(a) (b) (c)

Figure 1. (*a*) H alpha chromospheric image of a Moreton wave at 02:47 UT on 24 September 1997, (*b*) *SOHO*/EIT 195 Å (dominated by an Fe XII emission line) coronal image of the wave at 02:49 UT and (*c*) H alpha image of the wave at 02:50 UT. The speed of the wave varied spatially and temporally, ranging from 250 to over 600 km s^{-1}. The flaring active region is visible in all three images. H alpha images courtesy of Barry Reynolds, Perth, Australia.



(a) (b)

Figure 2. (*a*) *SOHO*/EIT 195 Å image of the Sun on 12 May 1997 at 05:24 UT. (*b*) *SOHO*/EIT image at 05:07 UT with the pre-event image at 04:34 UT digitally subtracted from it. White (black) regions denote increase (decrease) in emission. The dark regions near the active region visible in both images are darkened regions associated with a coronal mass ejection. The bright circular ring of emission outside the darkened regions in the right-hand image corresponds to an increase in emission propagating at 250 km s^{-1}.

B J Thompson

evidence that some of the wave transients are driven by this effect. However, as mentioned previously, the timing and origin of H alpha Moreton wave observations coincide well with H alpha observations of flares. Because our current understanding of solar flares has advanced beyond the interpretation of 'massive bomb-like explosions', the wave initiation mechanism is unclear. Possible candidates are radiation pressure, rapid thermal expansion and accelerated particle drivers.

Bibliography

Athay R G and Moreton G E 1961 *Astrophys. J.* **133** 935
Meyer F 1968 *Structure and Development of Solar Active Regions (IAU Symp. vol 35)* ed K O Kiepenheuer, pp 458–89
Moreton G E 1961 *Sky Telesc.* **21** 145
Smith S F and Harvey K L 1971 *Physics of the Solar Corona* ed C J Macris, pp 156–67
Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Morgan, William Wilson (1906–94)

American astronomer, staff member of the Yerkes Observatory, with Philip Keenan introduced stellar luminosity classes and the *Morgan–Keenan (MK)* classification scheme of stars based on the appearance of their spectra. With his students Donald E Osterbrock and Stewart Sharpless, he discovered the two nearest spiral arms in the Galaxy by mapping the distribution of bright O- and B-type stars whose distances were derived from their spectral classification. Morgan presented the paper in Cleveland in 1951 to tumultuous applause, unusual in a scientific conference.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Morin, Jean-Baptiste (1583–1656)

Doctor, alchemist and astrologer, born in Villefranche, Beaujolais, France, became professor of mathematics at the Collège Royal, published a defense of ARISTOTLE, attacked GALILEO, being convinced that the Earth was fixed in space. Morin proposed an impractical method to solve the longitude problem, based on measuring absolute time by the position of the Moon. He put the method in for a prize, and disputed for years the adjudication committee's negative verdict. Attacked DESCARTES, and his philosophy. A disputatious reactionary, he unsurprisingly was alienated from the scientific community.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Morley, Edward (1838–1923)

American chemist who collaborated with ALBERT MICHELSON in the *Michelson–Morley experiment*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Morning Star

A name given to Venus when it is visible in the east in the pre-dawn sky. Ancient astronomers believed that morning and evening apparitions of Venus were of two different planets; the morning planet was given the name Phosphor, or Phosphorus. As the morning star, Venus is moving from inferior conduction (when it lies between the Earth and the Sun) to superior conjunction (behind the Sun), and is visible for longest when it reaches the position known as greatest elongation west, when its angular separation from the Sun is greatest. The name 'morning star' is sometimes given to morning apparitions of Mercury.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moulton, Forest Ray (1872–1952)

Mathematical astronomer, born in Le Roy, MI, became professor at the University of Chicago. Proposed (with THOMAS CHAMBERLIN) the planetesimal hypothesis, explaining the origins of the solar system. Provided the crucial theoretical underpinning of BARRINGER's theory of the origin of the Arizona meteor crater.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mount Graham International Observatory

The Mount Graham International Observatory at an elevation of 3200 m, is located near Safford, Arizona. The observatory specializes in advanced technology telescopic facilities that can benefit from the low water vapor and sharp images of the site, and its easy access. The observatory is operated by the University of Arizona. The 10 m diameter Heinrich Hertz Sub-millimeter Telescope (SMT) is a joint project of Arizona and the MAX-PLANCK-INSTITUT FÜR RADIOASTRONOMIE, Germany. The adjustment of the carbon fiber telescope surface has surpassed its goal of 15 microns rms. The Vatican Observatory/Arizona Lennon telescope has a 1.8 m $f/1$ primary figured to 17 nm rms. The third and largest telescope is the 2×8.4 ($f/1.14$) LARGE BINOCULAR TELESCOPE (LBT). It is a partnership between Arizona, Ohio, Italy (Arcetri), Germany and the Research Corporation. The LBT is now under construction and is expected to begin operation in 2002.

For further information see
http://medusa.as.arizona.edu/graham/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mount John University Observatory, New Zealand

The observatory specializes in the observation of variable stars and binary stars by photometry and spectroscopy, especially using a high-dispersion échelle spectrograph. It is located in the center of New Zealand's South Island, at latitude 44 °S. Mount John is owned and operated by the Department of Physics and Astronomy of the University of Canterbury. It was founded in 1965 and currently there are a 1 m and two 61 cm reflecting telescopes. There are four resident staff at the observatory and three academic astronomers at the university site in Christchurch. A major collaboration with Nagoya and Auckland universities, the MOA project, is to study gravitational microlensing by CCD photometry.

For further information see http://www.mjuo.canterbury.ac.nz/mjuo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mount Stromlo and Siding Springs Observatories

Mount Stromlo and Siding Spring Observatories (MSSSO) are the astronomical observatories of the Australian National University. The facilities at Mount Stromlo Observatory were originally established as the Commonwealth Solar Observatory which commenced operation in the Australian Capital Territory in 1924. Expansion after the Second World War saw Mount Stromlo Observatory move into stellar astrophysics, and in 1957 it joined the ANU's Research School of Physical Sciences. Development of Canberra led to the establishment of a dark sky observatory at Siding Spring (SSO) west of Coonabarabran in New South Wales in 1964. Ten years later SSO became the host of the ANGLO-AUSTRALIAN OBSERVATORY with its 4 m and Schmidt telescopes.

The Observatories' major facility is the ANU 2.3 m telescope, opened in 1984. The telescope is a Cassegrain and Nasmyth alt-azimuth under full software control, capable of optical and infrared imaging and spectroscopy. A 1 m telescope is used for imaging, and there are smaller facilities too. At Mount Stromlo Observatory the ANU has a 1.9 m telescope with an excellent Coudé spectrograph and the 1.3 m Great Melbourne Telescope, which is fully automated and used for survey projects, such as the MACHO project. The MACHO project in 1993 recorded the first detection of gravitational microlensing, and over the following six years sought the density of massive compact halo objects by observing the Large Magellanic Cloud and the Galactic nuclear bulge. The project was a collaboration with two US universities/laboratories.

The academic staff of the ANU's Research School of Astronomy and Astrophysics is based at Mount Stromlo Observatory. The School supports a graduate program, offering a PhD in Astronomy and Astrophysics. Mount Stromlo is also the site of a satellite laser ranging facility, operated for the Australian government, and the Stromlo Exploratory, an astronomy visitor center and interactive science center. The ANU's engineering facilities for astronomical instrumentation are also concentrated at Mount Stromlo. The engineering group has built all of the current instrumentation for MSSSO and does external work for, e.g., the Anglo-Australian Observatory and the International Gemini Project.

Siding Spring Observatory also hosts the University of New South Wales patrol telescope. The Uppsala telescope is currently carrying out a Near Earth Object survey for the University of Arizona and NASA. At SSO the ANU operates a lodge for visiting astronomers and workshops for telescope maintenance and site upkeep.

Web update (30 April 2003)
In January 2003, the Mount Stromlo Observatory was largely destroyed by bushfires which ravaged Canberra. The fires destroyed four telescopes, the equipment workshop, eight houses which had been occupied by staff and an administration building. Preliminary estimates have valued the losses at more than $20 million. From a history of astronomy perspective this is a major catastrophe, as the following historically-significant telescopes were lost: the 74-in (Grubb-Parsons) Reflector, the 50-in Great Melbourne Telescope, the 30-in Reynolds Reflector, the 9-in Oddie Refractor, the Solar Telescope and the 26-in Yale-Columbia Refractor. In addition to its telescopes, the Observatory's library was destroyed, along with the Melbourne Observatory archives.

For further information see
http://msowww.anu.edu.au/

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Mount Wilson Observatory

Mount Wilson Observatory, located in the San Gabriel Mountains near Pasadena, California, was founded in 1904 by George Ellery Hale with financial support from Andrew Carnegie. In the 1920s and 1930s, working at the 2.5 m Hooker telescope, Edwin Hubble made two of the most important discoveries in the history of astronomy: first, that 'nebulae' are actually island universes—galaxies—each with billions of stars; second, that these galaxies are moving away from us in all directions, resulting in an expanding universe. This second discovery became the basis for the 'Big Bang' theory of the origin of the universe.

In 1992, under a formal agreement between the Mount Wilson Institute and the Carnegie Institution of Washington, Mount Wilson Institute assumed complete responsibility for the funding and operation of the observatory.

The mission of the Mount Wilson Institute (MWI) is to maintain, manage and operate Mount Wilson Observatory for research, education and other projects in the field of astronomy. Among the institutions conducting research at the observatory in 1999 are the Harvard-Smithsonian Center for Astrophysics, UCLA, UC Berkeley, USC, Jet Propulsion Laboratory, Georgia State University and the University of Illinois. These guest research programs are funded and managed by the respective institutions, which pay a fee to MWI to defray the cost of site maintenance.

Mount Wilson Observatory facilities include the 2.5 m Hooker telescope; the 1.5 m telescope; the Georgia State University CHARA interferometer array, consisting of six 1 m telescopes arranged in a 350 m diameter configuration; the UC Berkeley Infrared Spatial Interferometer; the 45 m and 18 m solar tower telescopes; the Snow horizontal solar telescope; and the 60 cm and 35 cm remotely controlled telescopes, used by the Telescopes in Education (TIE) program for education and student research. A residence facility is maintained for the use of visiting astronomers.

The recent completion of an adaptive optics system on the 2.5 m Hooker Telescope has enabled this instrument to produce the sharpest images in the visible-light portion of the spectrum ever obtained by any telescope in the history of ground-based astronomy, achieving a resolution of 0.06 arcseconds. This is comparable to what would be obtained if the telescope were placed in space.

Mount Wilson's calm atmosphere result in the best 'seeing', i.e. natural sharpness and quality of its telescope images, in North America. This superb seeing is one of the reasons for the world-class performance of the 2.5 m adaptive optics system. It also led to selection of Mount Wilson as the site for Georgia State University's interferometer—the most powerful visible-light interferometer in the world, capable of a resolution of 0.0002 arcseconds.

For further information see
http://www.mtwilson.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mouton, Gabriel (1618–94)

Cleric, born in Lyon, France, observed the apparent diameter of the Sun and Moon as a measurement of their changing distance and an aid to understanding their orbits, published *Observationes Diametrorum Solis et Lunae Apparentium* (1670), studied interpolation, produced tables of logarithmic sines and cosines, and proposed the metric system, with a standard linear measurement based on the circumference of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moving Groups

A moving group is a set of stars which have similar velocities and have a common origin, or, as EDDINGTON stated,

> Since some thousands of proper motions are available, it must be possible, if we take almost any star, to select a number of others the motions of which agree with its motion approximately. ... Clearly the agreement of the motions is no proof of association, unless there is some further condition which indicates that the coincidence is in some way remarkable.

The stars in a moving group are coeval, have the same chemical composition as the parental cloud and share the same space motion to within the internal velocity dispersion of the group. The space configuration of these stars is not restricted to any form. Some examples of moving groups are: GALACTIC OPEN CLUSTERS, OB and T Tauri associations and GLOBULAR CLUSTERS. The common space motion of moving-group members allows a clean identification of these stars. As a result moving groups have always been important in studies regarding STAR FORMATION, STELLAR EVOLUTION, and the UNIVERSAL DISTANCE SCALE.

Selection methods

Traditionally, stars in moving groups were selected using measurements of PROPER MOTIONS: the angular displacements of stars on the sky per unit of time. Because moving-group stars have similar space velocities they also have similar proper motions. Figure 1 illustrates the power of this approach. The top left panel shows the distribution of proper motions in the field of the globular cluster NGC 6397. There are two separate concentrations of stars, one being the cluster members (for practical purposes centered in the frame), the other being the field star population. Figure 1 shows three color–magnitude diagrams (CMDs): (left panel) CMD of all stars in the field, (middle panel) CMD of the moving-group stars and (right panel) CMD of the field stars. The simple selection criterion used in figure 1 already results in a clean CMD which can be studied to test theoretical models for stellar structure and evolution and to use them to derive the age and chemical composition of the moving group. Some more sophisticated methods to select moving groups will be discussed below.

Vector point diagram method

A diagram such as the top left panel of figure 1, in which the two proper-motion components of stars in a field are plotted versus each other, is usually referred to as a vector point diagram. A moving group in the field will reveal itself as an overdensity in this diagram. The vector point diagram selection method tries to fit the distribution of proper motions of the field star population and that of the cluster stars assuming that both can be represented by a smooth function, generally taken to be a two-dimensional Gaussian. For each star the fitted Gaussian of the field and the moving group can be evaluated at its measured proper motion; the ratio of 'moving-group Gaussian to field-plus-moving-group Gaussian' then determines the membership probability of a star (see figure 2).

The vector point diagram method works best if the spread in proper motions of the moving-group stars is small, so that they stand out clearly against the broader field star distribution. This spread in proper motions in the vector point diagram depends on (i) the internal velocity dispersion of the moving-group stars, (ii) the observational errors on the proper motions and (iii) the perspective effect, which causes the proper motions of stars with the same space velocity but different positions in the field to be different (since they are viewed from different directions). As a result, this method is optimally suited for moving groups which cover only a small fraction of the sky, typically less than $\sim$10 square degrees (depending on proper motion errors and internal velocity dispersion). Many such studies have been carried out in the past century, by comparing the stellar positions on photographic plates taken at different epochs. For moving groups which have larger angular sizes the perspective effect will wash out the concentration in the vector point diagram, and other selection methods are needed.

Convergent point method

The convergent point method exploits the perspective effect discussed above. An important characteristic of the perspective effect, although it creates a different proper motion for each star, is that the proper motions of moving-group stars will all be directed towards one point on the sky. This point is called the convergent point (see figure 2). This convergence of the proper motions is analogous to the fact that the two tracks of a rail-line appear to converge in the distance. Stars can easily be selected by determining whether their proper motions point towards the convergent point or not (see figure 2; also note the two stars that are non-members). Although selecting the moving-group stars is relatively easy, finding the convergent point is more difficult. Various versions of the convergent point method use different algorithms for finding the convergent point. They all require some numerical minimization process to search for groups of stars whose proper motions converge towards one point on the sky.

Three-dimensional velocity methods

The previous two methods select moving-group stars based on proper motions and positions. However, the preferred way to select moving-group stars is to search in velocity space directly. To calculate a three-dimensional velocity of a star, its position, proper motion, distance (or PARALLAX) and RADIAL VELOCITY are needed. The distance of a star is required to convert the proper

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

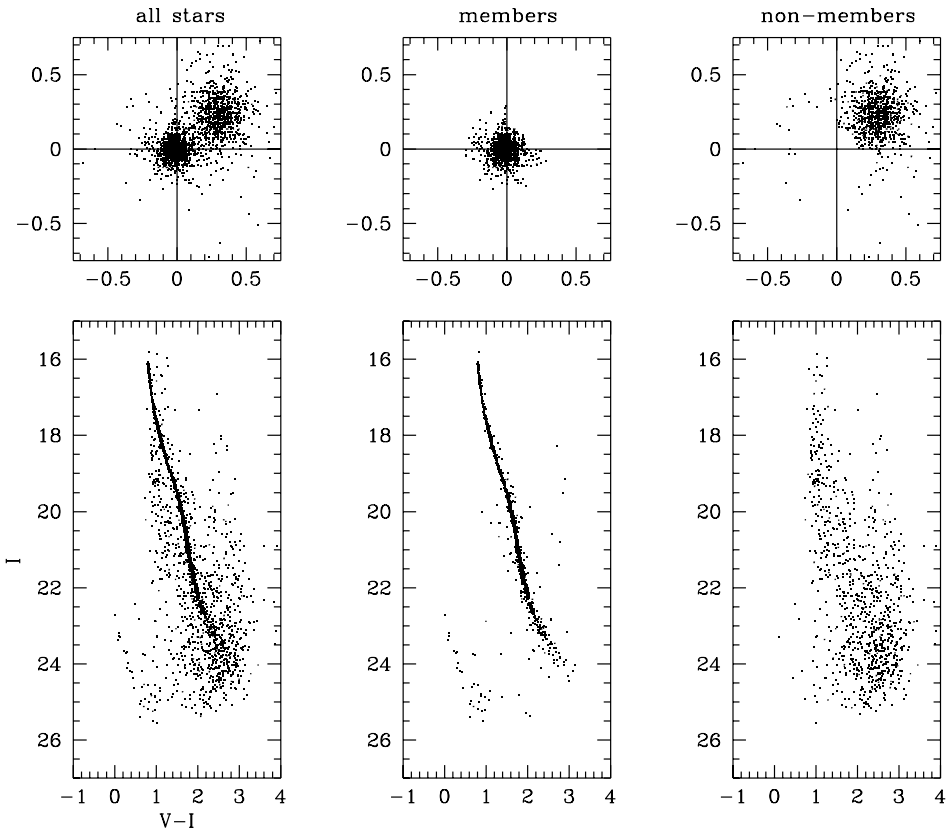


Figure 1. Vector point diagrams (top) and color–magnitude diagrams (bottom) for the globular cluster NGC 6397. All measurements were taken with the HUBBLE SPACE TELESCOPE over a period of 32 months. The left panels show all stars in the field of view of the cluster. The middle panels show the cluster stars; they have been selected based on their proper motions. The right panels show the remaining field stars after selecting the cluster. Note that after the selection of the cluster stars the white dwarf sequence (bottom left of middle panel) becomes apparent as a separate feature in the globular cluster color–magnitude diagram. The proper-motion zero point is chosen as the mean motion of the cluster stars. The proper-motion scale is the displacement in WFC (Wide Field Camera of the Hubble Space Telescope) pixels (1 pixel is 37.5 mas yr^{-1}). Adapted from King *et al* (1998 *Astrophys. J.* **492** L37–L40).

motion into the tangential velocity. The combination of the two components of the tangential velocity and the radial velocity then provides the three-dimensional (space) velocity of a star.

Many schemes have been developed to select moving-group stars using their three-dimensional velocities. The common denominator of all these methods is that they search for stars whose velocities are, within the measurement errors and internal velocity dispersion, equal to each other. The differences between the various three-dimensional velocity methods stem mostly from the differences in available data (e.g. catalog and accuracy of the data) used in the member selection. Most of these methods rely on computationally intensive numerical algorithms.

Moving groups and their astrophysical importance

Several types of moving groups will be discussed below: open clusters, OB associations, globular clusters and halo streamers.

Open clusters

Open clusters are gravitationally bound groups of stars, found in the Galactic disk. The small tidal radius (∼10 pc), combined with the small internal velocity dispersion (a few tenths of km s^{-1}), allows an easy selection of the members based on radial velocities or proper motions, e.g. with the vector point diagram method. The number of stars in an open cluster varies between a few dozen to several thousands.

The nearest open cluster, the Hyades, is located at ∼45 pc from the Sun; several other clusters, including the Pleiades, have distances on the order of 100 pc. The proximity of these clusters allows the acquisition of high-quality data. Besides easy membership selection, this results in well-defined main sequences in the HERTZSPRUNG–RUSSELL DIAGRAM, free of contamination by field stars. These main sequences and other measurements are used to refine and extend the theory of the internal constitution of the stars. The effects of stellar evolution can also be studied because the ages of observed open clusters vary between a few million years to a few billion years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

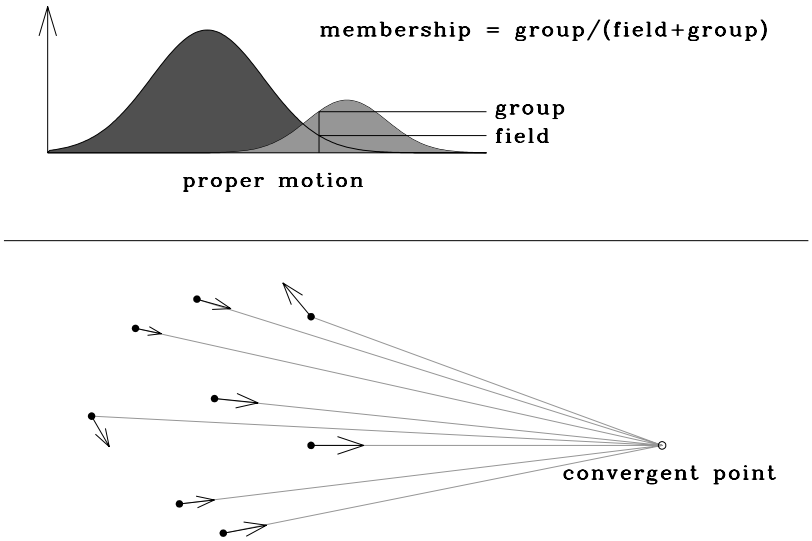


Figure 2. Schematic representation of the vector point diagram method (top): the Gaussians (here represented as one-dimensional) fitted to the proper motion distribution of the field (dark gray) and the cluster stars (light gray) are evaluated at a specific proper motion, i.e. the measured proper motion of a star. The ratio of the value of the moving-group Gaussian and the field-plus-moving-group Gaussian determines the membership probability for each star. Schematic representation of the convergent point method (bottom): stars whose proper motion is directed towards the convergent point are considered members.

Under the assumption that the unevolved parts of the main sequences of open clusters are identical, the distances of the clusters can be determined once the distance to one of them is known. The displacement between the main sequences of two clusters, corrected for interstellar extinction and for differences in metal abundance, is determined by their difference in distance (or distance modulus). For this reason open clusters, the Hyades open cluster in particular, have long been used as the main calibrator of the distance scale in the universe.

OB associations

OB associations are gravitationally unbound groups of typically a few thousand stars in the Galactic disk, including a number of early spectral type (O and B stars). The name OB association was introduced by the Russian astronomer AMBARTSUMIAN in 1947. He noted that these groups were unbound, based on their average density, and concluded that they therefore must be young; any unbound structure would have been disrupted under influence of the Galactic tidal field on a time scale of ∼50 million years. This conclusion was later supported by the theoretical insight, based on stellar evolution–structure theory, that early type (or massive) stars must be young. Further evidence for the youth of these groups was furnished by the discovery of star-forming INTERSTELLAR MOLECULAR CLOUDS nearby OB associations.

The main interest in the study of OB associations is to understand star formation in general and high-mass star formation in particular. It is thought that the majority of stars in the Galactic disk formed in OB associations. Therefore, studying these groups and their properties, such as the initial mass function or binary population, is of great importance in the understanding of stars and their evolution in general. Furthermore, the study of OB associations reveals properties of the Milky Way disk: the bright O and B stars trace the density distibution of the GALACTIC THIN DISK and in particular outline the spiral structure (OB associations generally form in or near spiral arms).

Similar groups containing only low-mass young stars have also been discovered. These are usually referred to as T associations, because of the presence of T TAURI STARS. A well-known example is the Taurus–Auriga association.

Identifying OB and T association members is more difficult than selecting open cluster members. Because associations are unbound they expand and can form large (30–100 pc) structures of which the closest ones (Sco OB2 at ∼130 pc) cover almost 1000 square degrees on the sky. Association members can thus only be identified using a convergent point method or a three-dimensional velocity method. Using photographic techniques for measuring proper motions of sufficient accuracy over such large areas proved to be nearly impossible. However, the globally accurate proper motions and parallaxes provided recently by the HIPPARCOS satellite allowed a reliable member selection for over a dozen nearby associations.

Globular clusters

Globular clusters are dense conglomerations of 10^5–10^6 stars. They have a typical tidal radius of 50–100 pc. The Milky Way contains nearly 150 of these clusters, distributed throughout the Galactic halo (see HALO, GALACTIC).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Globular clusters played a crucial role in our view of the UNIVERSE. Based on the distribution of the visible globular clusters SHAPLEY concluded, in the 1920s, that the universe was 100 kpc in size and that the Sun was located 15 kpc from the center of the system of globular clusters. This enlarged the universe of KAPTEYN, the accepted model of the universe at the time, by a factor of 10 and put the Sun in a non-central position.

Globular clusters still play a key role in the understanding of our Galaxy, although not concerning its size, but rather its formation. They are the oldest objects (∼12 billion years) in the Galaxy and were presumably the first objects to form. Studying them gives information on the processes governing the formation of galaxies. They also provide valuable information on stellar evolution and structure theories and are ideal laboratories to test theories of STELLAR DYNAMICS in high-density environments.

Although the stellar density in globular clusters is very high, so that the cluster stands out against the background stars, member selection based on proper motions is still necessary to remove field stars from the sample (as illustrated in figure 1).

Halo streamers

The term 'halo streamers' is used for the tidal debris of small galaxies, e.g. galaxies the size of the SMALL and LARGE MAGELLANIC CLOUDS, which have fallen into the halo of the Galaxy. When this happens stars are stripped off the satellite galaxy owing to the tidal forces exerted by the Milky Way. These stripped stars then follow their own orbit in the galactic halo. This tidal debris generally forms long arcs, sometimes more than 360° on the sky (cf the MAGELLANIC STREAM and the SAGITTARIUS DWARF GALAXY).

Halo streamers are not moving groups in the sense of the definition given at the beginning of this article (they do not have the same space velocity), but stars belonging to these streams can be selected based on velocity information. The energy and angular momentum (integrals of motion) of these stars are nearly the same as for the infalling satellite galaxy. Thus instead of looking for stars with similar three-dimensional velocities, stars with similar integrals of motion can be identified as belonging to the same group. This approach can, in principle, also be applied to 'normal' moving groups.

The main reason for studying halo streamers and infalling satellite galaxies is to study the process of GALAXY FORMATION. How structure formation in the universe took place is still an open question. Did the small structures form first and did they subsequently form bigger structures (hierarchical structure formation) or were the big structures the first to form and did the smaller structures form later? By studying the Galactic halo and looking for halo streamers the questions about the formation and evolution of the halo of our Galaxy (and the Galaxy itself) may be answered (see also UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION.

Conclusions

The most important characteristic of a moving group is that the common space motion of the stars in such a group allows one to identify a set of stars which have similar ages and a similar formation history. Future astrometric satellites such as FAME and in particular GAIA will provide globally accurate proper motions (and radial velocities) with vastly improved accuracies, which will allow identification of moving groups and streamers not only in the solar neighbourhood but throughout the Galactic disk and halo, and will also allow measurement of the internal motions within these groups and structures. The physical characteristics of these moving groups then serve as tests for theories of star formation, stellar structure, stellar evolution, stellar dynamics and galaxy formation.

Bibliography

Starting points into the literature can be found in

Binney J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)

de Zeeuw P T *et al* 1999 A Hipparcos census of the nearby OB associations *Astron. J.* **117** 354–99

P T de Zeeuw, R Hoogerwerf and J H J de Bruijne

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mullard Radio Astronomy Observatory

The Mullard Radio Astronomy Observatory (MRAO) is operated by the Astrophysics Group of the Department of Physics at the Cavendish Laboratory. The observatory is located 8 km south-west of Cambridge at Lord's Bridge and was established in 1957 through the generosity of Mullard Limited with support from the Science Research Council. Current work is supported by the PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL. The observatory is famed for the pioneering sky surveys of Sir Martin Ryle, who invented the technique of aperture synthesis, and for the discovery of pulsars by Antony Hewish and his student Jocelyn Bell in 1967. These contributions were recognized by the award of the 1974 Nobel prize for physics to Ryle and Hewish.

Instruments currently in operation include the Ryle Telescope, an array of eight parabolic dishes on a 5 km baseline, originally built in 1971 for high-resolution imaging of radio galaxies and quasars. It has been upgraded for mapping faint structures in the cosmic microwave background caused by the Sunyaev–Zel'dovich effect in nearby galaxy clusters and proto-clusters at high redshift. Mapping the microwave background to detect primordial density fluctuations of cosmological significance has been achieved with a prototype Cosmic Anisotropy Telescope (CAT) and this has led to the design of a more advanced instrument, the Very Small Array (VSA) soon to be commisioned on Tenerife; this is a joint project with JODRELL BANK and INSTITUTO DE ASTROFISICA DE CANARIAS. Milliarc second resolution at optical wavelengths for imaging stellar disks has been achieved with the Cambridge Optical Aperture Synthesis Telescope (COAST). Pioneered by John Baldwin, the COAST array of five mirrors on baselines up to 100 m will hopefully be succeeded by a Large Optical Array (LOA) now being designed for a site with clearer skies. At the other end of the spectrum, sky surveys at the low frequency of 151 MHz have been made with the Cambridge Low Frequency Synthesis Telescope (CLFST). This array of 60 tracking Yagi antennas provides 776 baselines up to 4.5 km simultaneously.

In addition to operations at MRAO the Astrophysics Group is developing instrumentation for the JAMES CLERK MAXWELL TELESCOPE (JCMT) on Mauna Kea, Hawaii. MRAO was closely involved with the design and construction of the JCMT, built jointly by the UK, Canada and the Netherlands, and is currently developing a submillimeter wavelength 'camera', a focal plane array of detectors for imaging large areas of sky.

The Astrophysics Group has a total staff (scientific, technical and administrative) of 68, plus 38 research students, and is engaged in a wide variety of observational, technical and theoretical projects.

For further information see
http://www.mrao.cam.ac.uk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mullard Space Science Laboratory

The Mullard Space Science Laboratory (MSSL) is located in the Surrey Hills about 65 km to the southwest of London. It forms the major element of the Department of Space and Climate Physics, University College London.

The department undertakes research into climate physics, high-energy astrophysics, solar physics, space plasma physics and photon and particle detection systems. As the UK's largest university space physics institute, it includes professionally staffed electronic, mechanical and software engineering groups and has designed and built instruments for more than 30 orbiting spacecraft and 200 sounding rockets. Missions in which MSSL has participated include Giotto, Cluster, Cassini and XMM.

For further information see http://www.mssl.ucl.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Multiple Mirror Telescope Observatory

At the time of its dedication in 1979, the 4.5 m Multiple Mirror Telescope (MMT) was the third largest optical telescope in the world. It featured so many ambitious design innovations that its completion heralded the beginning of the current generation of telescope design. The MMT was decommissioned in March 1998, and was replaced in March 1999 with a 6.5 m single-primary telescope.

The 4.5 m MMT was the first large optical telescope to incorporate multiple, lightweight primary mirrors in a common mount, a rotating building instead of a dome, active secondary mirror control during all observations, a high-performance altitude–azimuth mount and detailed finite element analysis of the telescope's structure during the design process. The telescope was a very successful research tool, fueling nearly 1000 scientific publications.

Studies carried out in Arizona in the mid-1980s showed that a 6.5 m primary mirror with a focal ratio of *f*/1.4 or less could be installed in the MMT building. When it was shown that such a telescope could have a substantial field of view, the interest of Smithsonian astronomers was aroused. A letter of intent to carry out such a project sealed the fate of the 4.5 m telescope in 1988. The successful casting (1992) and polishing (1997) of the first 6.5 m, *f* /1.25 primary in the STEWARD OBSERVATORY Mirror Laboratory triggered the decommissioning of the old telescope in 1998.

The new telescope was dedicated in May 2000 and scientific observations began the following month. The successful conversion of the MMT is ongoing. The first of three secondary mirrors has been installed. The new telescope has already improved the observatory's capabilities significantly. Notably, the telescope now has more than double the light-gathering power of its predecessor and two reflections have been removed from the optical path. Two new secondaries will be commissioned in 2001. The first, a 1.7 m mirror, will increase the angular field of view by a factor of 15 (to 1º) and the areal field by a factor of 225. The second, a 0.7 m deformable mirror, will provide a two-reflection, thermally clean adaptive focus.

Located at the 2600 m summit of Mount Hopkins, approximately 65 km south of Tucson, Arizona, the MMT Observatory is a facility of the University of Arizona and the Smithsonian Institution. It is staffed by approximately 20 full-time personnel and is funded jointly by its two parent institutions. Observing time is used by scientists at the University of Arizona, Arizona State University, Northern Arizona University and the Smithsonian Astrophysical Observatory. A significant amount of time is also made available to members of the international astronomical community through a public access programme.

For further information see http://sculptor.as.arizona.edu/foltz/www/mmt.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Multiple Star

Any star system that comprises three or more components that are physically linked, i.e. that have a common gravitational field. It has been estimated that about 30% of all binary systems are actually triple; systems with more component stars are much rarer. A very few systems are known that have as many as six known components; Castor and the Trapezium at the heart of the Orion Nebula are the best-known examples.

Many triple star systems comprise a close binary pair and a much more distant (and often less massive) companion. The α Centauri system is of this type.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Multiple Stellar Systems: Types and Stability

A multiple stellar system is a gravitationally bound group of two or more stars. This rather broad definition includes binary, triple, quadruple and larger stellar systems, all the way up to star clusters and galaxies. It is useful to subdivide this diverse aggregation into three categories: binary stars, hierarchical systems, and non-hierarchical systems. Binary stars are explained at great length elsewhere in this encyclopedia, and so will not be discussed here (see BINARY STARS: OVERVIEW). Hierarchical stellar systems can be broken down into physically separate subgroups of stars that orbit each other in approximately Keplerian fashion. These systems can be completely described by a combination of individual stellar parameters, such as STELLAR MASSES, and relationships between pairs and other subgroups of stars, such as mass ratios, semimajor axes and eccentricities. This is in contrast with non-hierarchical systems, such as STAR CLUSTERS, which are best described by statistical quantities, such as central density, core radius, velocity dispersion, mass function, binary frequency and total mass.

Observed systems

Multiple stellar systems are ubiquitous. For example, among the stars within 17 light-years of the Sun in Alan Batten's 1999 catalog of the nearest stars, there are 38 single stars (including the Sun), 12 binaries and 4 systems of three or more stars. The frequency of triple and larger systems is

$$f_{\geq \mathrm{triple}} = \frac{N_{\geq \mathrm{triple}}}{N_{\mathrm{single}} + N_{\mathrm{binary}} + N_{\geq \mathrm{triple}}} \qquad (1)$$

and exceeds 7% for the 54 systems within 17 light-years. This is consistent with estimates of 5–15% in the literature for larger samples. Extrapolating this to the $\sim 10^{11}$ stars in the entire Galaxy yields a total of $\sim 10^{10}$ systems of three or more stars.

Famous multiple stellar systems include Castor (α Gem), which has six components, α and Proxima Cen, which is at least a triple system, and PSR B1620-26, a pulsar in a triple system in the globular cluster M4. All three display a hierarchical structure.

Essentially all of the systems of three to seven stars in the catalog of 612 multiples compiled by Tokovinin are hierarchical, and much of the research on multiple systems has focused on the hierarchical variety. Apart from some special situations, non-hierarchical stellar systems are dynamically unstable. Such systems will reassemble into stable hierarchical configurations on timescales that are relatively short compared with the stellar evolutionary timescale, and stars will often be ejected in the process (see STELLAR EVOLUTION).

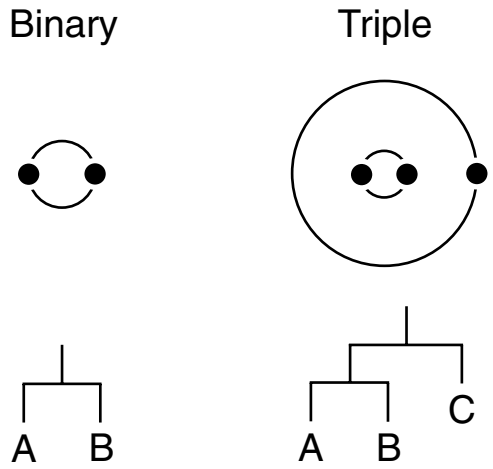


Figure 1. Schematic and 'mobile' diagrams for binary and hierarchical triple systems. A hierarchical triple system consists of two Keplerian orbits.

Some examples of hierarchies

It is useful to have a picture in mind of the structure of hierarchical stellar systems. Figure 1 shows a binary system and a hierarchical triple system. The latter can be broken down into two Keplerian orbits. The close pair of stars is called the inner pair and their orbit is called the inner orbit. The orbit of the third star about the inner pair is called the outer orbit. Figure 2 shows two hierarchical quadruple systems, and both can be broken down into three Keplerian orbits. Hierarchical systems of higher N can be broken down into $N - 1$ Keplerian orbits. Unlike the idealized systems shown in the figures, real systems are made up of eccentric and non-coplanar orbits, and stars of unequal mass, and interactions between the stars cause departures from Keplerian orbits.

Instabilities

Multiple stellar systems are subject to several kinds of instabilities. For example, non-hierarchical systems are dynamically unstable; they will rearrange themselves into stable hierarchical configurations and will often eject stars in the process. Most stability studies have focused on this, but there are additional instabilities. Close encounters in non-hierarchical stellar systems can result in stellar tidal interactions, physical collisions and mergers. The merger process may be assisted by the emission of gravitational radiation, especially when compact objects are involved. Also, stellar evolution of the individual components can result in significant mass loss from those components and mass transfer to other components. This can affect the stability of the system. Finally, external influences, such as intrusive encounters by stars, can also be the cause of dynamical instability in multiple stellar systems.

Stability of isolated systems of point masses

The literature on the stability of isolated systems of three point masses is quite extensive. A great deal of effort has

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

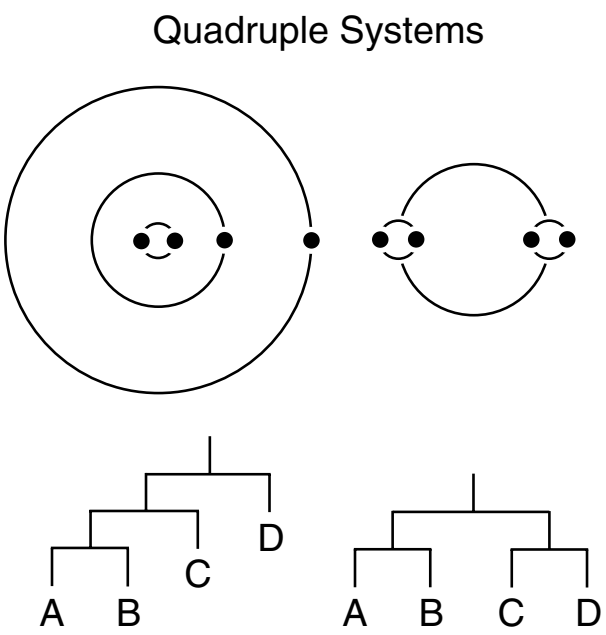


Figure 2. Schematic and 'mobile' diagrams for hierarchical quadruple systems. A hierarchical quadruple system consists of three Keplerian orbits.

gone into studying THREE-BODY PROBLEMS via a technique that was introduced in the 1870s by George Hill, and is described nicely by Szebehely and Zare. The method is based on c^2H in the two-body approximation, where c is the angular momentum of the system and H is the total orbital energy. If $c^2H < (c^2H)_{\rm crit}$, then the inner pair is permanently isolated from the outer body. This guarantees stability, since otherwise interactions between the third body and the inner pair can result in exchanges and ejections. The critical value, $(c^2H)_{\rm crit}$, is calculated for a configuration that is borderline stable, such as when all three bodies remain on a line as they orbit each other. The stability criterion in the case of coplanar circular prograde orbits and equal-mass stars reduces to $a_{\rm outer}/a_{\rm inner} \gtrsim 3$. Application of the c^2H criterion to the real triple systems in the sky indicates that the observed systems are all quite stable.

At least as much effort has gone into investigations of the stability of three-body systems through numerical integrations, which is perhaps the most effective way to study the problem. In N-body simulations, the initial conditions decide the outcome, and therefore the task is reduced to identifying those initial conditions which lead to stable systems and those which do not.

The numerical integrations reveal that many close stellar encounters occur in non-hierarchical systems, and this ultimately results in the ejection of one of the three bodies, usually the body with the smallest mass. Ejection often takes place following a close 'triple' encounter, which is when the eventual escaper approaches the center of mass of the three-body system as the other two bodies move away from it. The result is that the deceleration felt by the escaping body as it moves away from the center of mass is less than the acceleration the body felt as it approached the center of mass, leaving the body with sufficient velocity to escape from the other two. Additional results are that systems with high angular momentum are more stable than those with low angular momentum, retrograde inner and outer orbits are more stable than prograde orbits and systems with circular outer orbits are more stable than those with eccentric outer orbits.

There are many empirical formulae in the literature for the stability of triple systems, and those of Eggleton and Kiseleva are given here as an example. These are based on a large number of numerical experiments of triple systems with a variety of orbital eccentricities and inclinations, and component masses. They correspond to n-stability for $n \gtrsim 2$, which is the persistence of the hierarchical configuration for 10^n times the period of the outer orbit. Systems with an initial period ratio, X_0, exceeding $X_0^{\rm min}$ are stable, where

$$(X_0^{\rm min})^{2/3} = \left(\frac{q_{\rm out}}{1+q_{\rm out}}\right)^{1/3} \frac{1+e_{\rm in}}{1-e_{\rm out}} Y_0^{\rm min} \qquad (2)$$

and

$$(Y_0^{\rm min}) \approx 1 + \frac{3.7}{q_{\rm out}^{1/3}} + \frac{2.2}{1+q_{\rm out}^{1/3}} + \frac{1.4}{q_{\rm in}^{1/3}} \frac{q_{\rm out}^{1/3}-1}{q_{\rm out}^{1/3}+1} \qquad (3)$$

where $q_{\rm in} = m_1/m_2 \geq 1$ and $q_{\rm out} = (m_1+m_2)/m_3$ are the inner and outer mass ratios.

It has been mathematically proven that a three-body system cannot remain bound for all time if it was previously unbound. That is, a temporary triple system created by a collision between a binary and a single star cannot last forever. Ultimately one component will be ejected, probably the least massive one, although it could be any of the three. A binary–binary collision can, on the other hand, create a stable three-body system via the ejection of one star that carries away with it enough binding energy to allow the other three to remain bound indefinitely in a hierarchical configuration.

In comparison with three-body systems, far less has been written on the stability of systems consisting of four or more stars. Non-hierarchical four-body systems can decay into (1) a binary plus two single stars, (2) two binaries and (3) a stable triple plus a single star. There are currently no empirical formulae for the stability of four-body or larger systems. There are, however, many statistical results from numerical integrations regarding the properties of the break-up products. High-velocity escapers from the decay of small-N stellar systems can account for some of the OB runaway stars observed in our Galaxy (see STELLAR N-BODY SYSTEM: DISSIPATION). Kiseleva and collaborators found that 1% of all escapers from such systems have speeds greater than 30 km s^{-1}. Sterzik and Durisen successfully fit the speed distribution of the escapers from $N = 3$, 4 and 5 stellar systems to analytic distributions derived for zero-angular-momentum initial conditions. They also found that the binding energies, E_B, of the remnant binaries follow an $|E_B|^{-7/2}$ distribution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finite stellar sizes

Close encounters in non-hierarchical stellar systems can result in stellar tidal interactions, physical collisions and mergers. For example, if two main-sequence stars pass within a distance of twice the sum of their radii, then tidal effects will cause the pair to merge into a larger main-sequence star, which may be observable as a BLUE STRAGGLER. This mechanism can quickly decrease the number of stars in a non-hierarchical stellar system by one or more. Blue stragglers can be produced by the binary–single and binary–binary interactions that occur in the cores of star clusters. Finite stellar sizes also limit the maximum possible speeds of runaway stars ejected from small-N stellar systems, since it is the interactions with the closest encounters that avoid mergers which result in the highest ejection speeds.

Tidal effects in the inner pair of a hierarchical triple system can affect its stability. Circularization and synchronization are difficult to achieve in the inner pair, as tidal friction tends to circularize the inner orbit, while the presence of the third star does the opposite. Slower than synchronized rotations in the inner pair of three-body systems increase stability by shrinking the inner orbit, while faster than synchronized rotation decreases stability by expanding the orbit.

Angular momentum in three-body systems can be transported from the inner orbit to the outer orbit, and this may be a way of producing short-period binaries. There is observational evidence for an excess of distant companions to contact binaries compared with other types of binaries, which provides some support for this mechanism. If the contact binary eventually merges into a blue straggler, then the hierarchical triple system is reduced to a binary.

Stellar evolution

Mass loss from stars and mass transfer between stars in a multiple stellar system due to stellar evolution can significantly affect the stability of the system (i.e. the criterion given by equations (2) and (3) can change if the masses change). For example, slow mass loss from the outer component can improve the stability by gradually increasing the distance of the outer component from the inner pair of the system. Of course, too much mass loss, especially the essentially instantaneous mass loss that results from a supernova explosion, will unbind the system, so this represents a form of instability. Mass transfer between the inner components will also have an effect, albeit small, on the system stability, according to equations (2) and (3).

Systems in star clusters

The connection between multiple stellar systems and star clusters is intimate, as a star cluster is essentially a very large non-hierarchical multiple stellar system, and the endpoint of the dynamical evolution of a cluster is probably several small-N multiple stellar systems. Multiple stellar systems are created within star clusters via gravitational encounters between binaries and other stars. Subsequent exchange interactions tend to replace the component stars of multiple stellar systems with more massive ones and increase the system binding energy. Strong interactions can even kick the systems out of the cluster. The hierarchical triple system containing the pulsar PSR B1620-26 in the globular cluster M4 was probably formed by a binary–binary interaction. Continual encounters with stars in M4 will alter this system, and limit its stability.

Black hole triple systems

Multiple stellar systems consisting of SUPERMASSIVE BLACK HOLES are possible. If most galaxies contain central black holes, then galaxy mergers will inevitably produce binary black holes, since dynamical friction will guarantee that the holes will sink to the center of the merged galaxy and find each other. A subsequent merger with another galaxy has the potential to produce a triple (or larger) black hole system. Whether the third black hole can get close to the inner pair of black holes before the latter merge owing to dynamical friction and the emission of gravitational radiation depends on v_c, the root-mean-square velocity of stars at the center of the merged galaxy. Makino's calculations indicate that the binary black hole probably will not have merged before the third hole reaches the center if $v_c < 200$ km s^{-1} and that the binary probably will have merged by the time the third hole arrives if $v_c > 600$ km s^{-1}. The observed v_c in real galaxies span this range, so probably both outcomes happen. If the third black hole arrives at the center of the merged galaxy to find a binary black hole already there, then the resulting binary–single interaction may eject one of the black holes from the galaxy.

Bibliography

Observed systems

Batten A H 1998 The nearest stars *Observer's Handbook 1999 of the Royal Astronomical Society of Canada* ed R Bishop (Toronto: University of Toronto Press) pp 234–8

Tokovinin A A 1997 MSC—a catalogue of physical multiple stars *Astron. Astrophys. Suppl. Ser.* **124** 75–84

Stability of triple systems:

Anosova J P and Orlov V V 1992 The types of motion in hierarchical and non-hierarchical triple systems: numerical experiments *Astron. Astrophys.* **260** 473–84

Eggleton P and Kiseleva L 1995 An empirical condition for stability of hierarchical triple systems *Astrophys. J.* **455** 640–5

Ge Y-C and Leng X 1992 An alternative deduction of the Hill-type surfaces of the spatial 3-body problem *Celest. Mech. Dyn. Astron.* **53** 233–54

Szebehely V and Zare K 1977 Stability of classical triplets and of their hierarchy *Astron. Astrophys.* **58** 145–52

Stability of larger-N systems

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kiseleva L G, Colin J, Dauphole B and Eggleton P 1998 High-velocity stars from decay of small stellar systems *Mon. Not. R. Astron. Soc.* **301** 759–66

Sterzik M F and Durisen R H 1998 The dynamic decay of young few-body stellar systems. I. The effect of a mass spectrum for $N = 3$, 4 and 5 *Astron. Astrophys.* **339** 95–112

Finite stellar sizes

Kiseleva L G, Eggleton P P and Mikkola S 1998 Tidal friction in triple stars *Mon. Not. R. Astron. Soc.* **300** 292–302

Leonard P J T 1995 Merged dynamically ejected OB runaway stars *Mon. Not. R. Astron. Soc.* **227** 1080–6

Leonard P J T 1996 The implications of the binary properties of the M67 blue stragglers *Astrophys. J.* **470** 521–7

Stellar evolution

Orlov V V, Petrova A V and Ivanova N S 1996 The effect of stellar wind on the stability of triple systems *Mon. Not. R. Astron. Soc.* **281** 1326–32

Systems in star clusters

McMillan S, Hut P and Makino J 1991 Star cluster evolution with primordial binaries. II. Detailed analysis *Astrophys. J.* **372** 111–24

Sigurdsson S 1995 Assessing the environmental impact of PSR B1620-26 in M4 *Astrophys. J.* **452** 323–31

Black hole triple systems

Makino J 1997 Merging of galaxies with central black holes. II. Evolution of the black hole binary and the structure of the core *Astrophys. J.* **478** 58–65

Valtonen M J, Mikkola S and Pietilä H 1995 Burrau's three-body problem in the post-Newtonian approximation *Mon. Not. R. Astron. Soc.* **273** 751–4

Peter J T Leonard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Musca

(the Fly; abbrev. Mus, gen. Muscae; area 138 sq. deg.) A southern constellation which lies between Carina and Circinus, and culminates at midnight in late March. It was first shown on Petrus Plancius' celestial globe of *c*. 1598 as Apis Indica (the Indian Bee), though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7. It was named Musca Australis (the Southern Fly) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2 (at which time there was also a now defunct constellation Musca Borealis, the Northern Fly). The shortened form was adopted by the IAU in 1922.

A small but not inconspicuous constellation, the brightest stars in Musca are α Muscae, magnitude 2.7, and β Muscae, a close binary with bluish-white (B2) components, magnitudes 3.5 and 4.0, separation 1.2″. There are three other stars brighter than fourth magnitude. Other interesting objects include NGC 5189, a tenth-magnitude planetary nebula of irregular structure, and two seventh-magnitude globular clusters, NGC 4372 and 4833, the former of which is partially obscured by dark nebulosity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Muses-C (Mu-launched Space Engineering Satellite)

Japanese–US sample return mission to asteroid Nereus (1989ML). Scheduled for launch in early 2002. Intended as a technology demonstration mission. A US-built nanorover will image the surface and collect three samples for return to Earth. Other technologies to be tested include use of solar powered electrical propulsion system and autonomous navigation and guidance. The mission will last four years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nadir

The point on the celestial sphere diametrically opposite to the zenith and directly below the observer.

See also: celestial sphere, zenith.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Naked-eye Astronomy

It is often assumed that to take a real interest in astronomy means acquiring a large and expensive telescope. This is emphatically not the case. Binoculars are relatively cheap, and are of great value astronomically (see BINOCULAR ASTRONOMY). But even with the naked eye there is much to see, and much to be learned. After all, telescopes were not invented until less than 400 years ago, and all the knowledge acquired in previous years depended upon the naked eye alone.

Probably the very first step after acquiring a preliminary knowledge of the main facts is to learn how to identify the constellations (figure 1). The stars become so much more interesting when you know which is which, and there is no real difficulty here, because the stars always remain in the same relative positions (their individual or proper motions are too slight to be noticed except over periods of many lifetimes). Moreover, it is seldom or never possible to see more than 2500 naked-eye stars at any one time, and the main constellations stand out clearly. A few nights' observation is enough to grain a working knowledge of the brightest patterns—that is to say, the stars of the first, second and third magnitudes. (The magnitude scale works in the manner of a golfer's handicap, with the brightest stars having the lowest magnitudes. The faintest stars normally visible without an optical aid are of magnitude 6.)

The best procedure is to select a few constellations which are immediately obvious, and use these are 'guides' to the rest. In the northern hemisphere, the two best guides are URSA MAJOR (the Great Bear) and ORION. Ursa Major is so far north in the sky that as seen from latitudes similar to those of Britain it never sets, and remains well above the horizon even when at its lowest; Orion, which is crossed by the celestial equator, is out of view during the middle of the year, when it is above the horizon only during daylight. From countries well south of the Earth's equator, such as Australia and New Zealand, Ursa Major is more or less inaccessible, though Orion can be seen from every inhabited country. However, there are other conspicuous 'guides' in the far south, and most people living in the southern hemisphere can readily identify the Southern Cross.

Consider first Ursa Major, with the seven fairly bright stars which make up the pattern known popularly as the Plough or, in America, as the Big Zipper. Two of these stars show the way to POLARIS, the Pole Star in Ursa Minor (the Little Bear), which lies within one degree of the north celestial pole, and which therefore seems to remain almost motionless in the sky, with the entire celestial sphere turning round it once in 24 hours. The 'handle' of the Plough shows the way to Arcturus, the brilliant orange star in the constellation of Boötes, the Hersdman, and so on. Orion is an even better guide when it is visible, because its stars are much brighter than those of Ursa Major; thus the three stars of the 'Belt' point southward to Sirius, the most brilliant star in the sky, and northward to Aldebaran, the red star in the constellation of Taurus (the Bull).

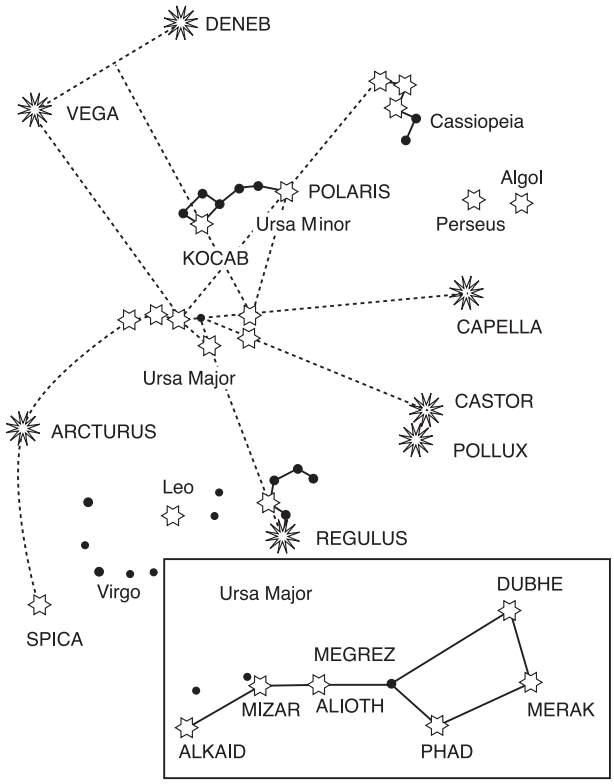


Figure 1.

Only the planets seem to wander around the sky. They keep to the Zodiac, and most are easy to recognize. Venus and Jupiter are always much brighter than any star; Mars is characterized by its strong red colour, and Mercury is always so near the Sun in th sky that it will probably not be seen unless deliberately sought. Only Saturn looks deceptively like a fairly bright star. The positions of the planets for any time can be looked up in astronomical ALMANACS or periodicals; and as soon as they have been identified they will easily be found again.

The MOON is much the closest natural body in the sky, and it has virtually no atmosphere, so that its surface features are always sharp and clear-cut. With the naked eye it is easy to see the outlines of the main dark areas, still miscalled seas (Latin, *maria*) even though there has never been any water in them (figure 2). Really keen-sighted persons can also make out indications of some of the mountain ranges, notably the Lunar Apennines, which form part of the boundary of the huge Mare Imbrium (Sea of Showers).

The Moon has no light of its own, and depends upon reflecting the rays of the Sun. At some full moons—they is to say, when the Sun, Earth and Moon are lined up—the Moon passes into the cone of shadow cast by the Earth; its direct supply of sunlight is cut off, and the Moon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 2. Naked-eye view of the full moon: 1, Mare Frigoris (Sea of Cold); 2, Mare Imbrium (Sea of Showers)' 3, Mare Serenitatis (Sea of Serenity); 4, Mare Crisium (Sea of Crises); 5, Mare Tranquillitatis (Sea of Tranquillity); 6, Oceanus Procellarum (Ocean of Storms); 7, Mare Fœcunditatis (Sea of Fertility); 8, Mare Nectaris (Sea of Nectar); 9, Mare Humorum (Sea of Humours); 10, Mare Nubium (Sea of Clouds).

turns a dim, often coppery colour until it passes out of the shadow again. Lunar ECLIPSES do not happen at every full moon, because the Moon's orbit is appreciably tilted, and most full moons avoid the cone of shadow; eclipses may be either total or partial. Not all eclipses are equally 'dark'; everything depends upon conditions in the Earth's upper atmosphere, through which all sunlight reaching the eclipsed Moon has to pass. On some occasions the eclipsed Moon becomes so faint that it is difficult to see at all.

Another interesting phenomenon can be seen when the Moon is a slender crescent. The 'dark' side can usually be seen shining faintly. This is due to light reflected on to the Moon from the Earth, and is therefore known as the EARTHSHINE.

Naked-eye observations of the Sun can be dangerous; it is unwise to stare directly, even when the Sun is low over the horizon and looks deceptively mild and harmless. Now and then sunspots become large enough to be seen with the naked eye, but a dark filter should always be used. (Never use binoculars or a telescope for direct observation of the Sun, even with a filter; see SUN: SAFE OBSERVATION.) The Sun has a fairly regular cycle of activity; maxima occur about every 11 years. The next maximum is due around 2000–2001, so that naked-eye spots may be expected now and then.

Near maximum activity, electrified particles from the Sun cascade down into the Earth's upper atmosphere, producing the lovely glows known as AURORAS or polar lights—*auroras borealis* in the northern hemisphere, *aurora australis* in the southern. Because they are electrically charged, the particles tend to spiral down in the regions of the magnetic poles. Thus a dark night in, say, North Norway will very probably be lit by an aurora, but from Southern England displays are much less common, and from regions near the Earth's equator they are very rare indeed (though it is on record that an aurora was once seen from Singapore!).

Auroras may take many forms. Some are mere glows; others may produce vivid colours, with rays, curtains and 'coronas', shifting and changing in brightness constantly. A mahjor auroral display is a magnificent sight, and is of course ideally suited to photography. Amateur observations of auroras are of real value, and are collected and correlated by local and national astronomical societies.

When the Sun is setting, a brilliant green flash may occasionally be seen as the last segment of the solar disk passes below the horizon. This 'green flash' is a purely atmospheric phenomenon, and is very uncommon; a very clear horizon is essential, preferably over the sea. Much more common are solar and lunar haloes, luminous rings round the Sun or the full Moon. These are due to very high ice-crystal cloud at well over 20 000 feet above ground level. They can be quite spectacular.

A solar eclipse occurs when he new moon passes directly between the Sun and the Earth, blotting out the brilliant solar dsisk for a brief period. Eclipses of the Sun may be either total, partial or annular. During a total eclipse the entire brilliant disk of the Sun is hidden, and the solar atmosphere flashes into view; for a brief period—never as long as eight minutes, usually much less; the naked-eye observer can see the solar CHROMOSPHERE, the red glowing masses of hydrogen once called red flames and now known as prominences, and the glorious pearly CORONA, which extends outward from the eclipsed Sun. A total eclipse is probably the most magnificent sight in all Nature (figure 3), but the Moon's shadow is only just long enough to touch the Earth, so that as seen from any particular locality total eclipses are rare; from England, the total eclipse of 11 August 1999 was the last until 2090. During totality it is quite safe to look directly at the Sun, without any dark filter, but even a small segment of the brilliant disk remaining visible means that the dangers return. It is only too easy to damage one's eyesight. Partial eclipses are much less interesting, since the Sun's surroundings cannot be seen. If exactly alignment occurs when the Moon is in the far part of its orbit, it cannot cover the Sun completely, and a ring of sunlight is left showing round the dark mass of the Moon; this is an annular eclipse (Latin *annulus*, a ring).

Observations of METEORS are best made with the naked eye, and here the amateur observer can carry out work of real importance (see also AMATEUR–PROFESSIONAL COLLABORATION IN ASTRONOMY). Meteors are cometary débris. As a comet moves along in its orbit, it leave a dusty 'trail' behind it; when the Earth passes through such a tail it collects many particles, and the result is a shower of shooting stars. Many meteor showers occur each year; the most reliable shower is that of the Perseids, seen from late July through to mid August. The average meteor is smaller than a grain of sand, so that what we see is not the particle itself, but the effects which it produces in the atmosphere as it burns away.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3.

Because the meteors in a shower are moving through space in parallel paths, they seem to issue from one particular region of the sky; thus the August meteors come from the direction of the constellation Perseus, hence their name. Other showers are much less predictable; the Leonids, of 17 November, are usually sparse, but may at times produce true 'meteor storms', as in 1966, when it was said that meteors rained down from the sky like snow flakes' for a few hours. There may also be sporadic meteors, not connected with any showers, which may appear from any direction at any moment.

The naked-eye meteor observer can make useful observations simply by counting the numbers of meteors seen over definite periods. To plot their paths, the best method is to hold a rule up again the star background, noting the points where the meteor is first and last seen, and then drawing the track on a chart; the radiant position will soon become evident. For this work, a sound knowledge of the constellations is of course needed.

Naked-eye COMETS are not common, though they do appear sometimes: comets Hyakutake (1996) and Hale-Bopp (1997) have been two recent examples. The periodical comets, which return regularly, are in general too faint to be seen without optical aid. Even *Halley's Comet*, which returns every 76 years and can become very bright indeed, was inconspicuous at its last return in 1986; unfortunately it will be no better at its next return, that of 2061.

Finally, what about naked-eye observations of the stars? They are of different colours; some are white, some bluish and some orange or red. Thus in Orion the two leaders, Betelgeuse and Rigel, are quite different. The warm orange of Betelgeuse contrasts sharply with the glittering whiteness of Rigel. Stars twinkle only because their light is disturbed as it comes to us by way of the Earth's atmosphere; a star such as Sirius will twinkle strongly, and will seem to flash various colours of the rainbow, even though Sirius is in fact a pure white star. A star near the zenith or overhead point will twinkle much less than a star which is lower down in the sky, because its light is coming to us by way of a thinner shell of atmosphere.

Most stars shine steadily for year after year, century after century, but there are some which do not; these are the VARIABLE STARS. One such star is Betelgeuse in Orion, which is well advanced in its evolution and has become unstable; it swells and shrinks, changing its output as it does so. At times it may be almost as bright as Rigel, while at others it is more comparable with Aldebaran. Another variable is Mira in Cetus (the Whale), which can rise to the second magnitude—as bright as the Pole Star—but which remains below naked-eye visibility for much of the year; it reaches maximum about every 331 days (the period is not constant). Algol, in Perseus, fades by more than a magnitude every 2.3 days, remaining faint for a mere 20 minutes before recovering, not because it is intrinsically variable but because it is regularly eclipsed by a darker companion passed in front of it. Naked-eye observers can follow the changes in these variable stars, by comparing them with nearby stars which are constant in brightness.

Occasionally a star will flare up where no naked-eye star has been seen before. This is termed a NOVA. Nova means 'new', but a nova is not in fact a new star at all; what happens is that a formerly dim star suffers a tremendous outburst, and flares up to many times its normal brilliancy before fading back to obscurity. Novae cannot be predicted, and several have been discovered by naked-eye observers. One nova, seen in 1918 in the constellation of Aquila, briefly outshone every star in the sky apart from Sirius, but has now become a very faint telescopic object.

It is clear, then, that there is much to be seen in the skies, even without any optical equipment. the naked-eye observer will always find plenty to occupy their attention.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nançay Radio Observatory

The Nançay Radio Observatory has the largest radio telescopes in France. Founded in 1952, it is located 200 km south of Paris and is operated by the Paris Observatory, in association with the CNRS, and co-financed by the 'Région Centre'. A technical staff of 50 operates the site. Its main instruments are: the decimetric telescope, with the equivalent surface of a 94 m diameter dish used for 21 cm H I and 18 cm OH spectral line observation of comets, evolved stars and galaxies as well as pulsar observations; the radioheliograph, a 43 element interferometer dedicated to solar corona observations at metric wavelength; the decametric array of 144 elements operating at 3–30 m wavelength for studies of Jupiter and the Sun.

For further information see http://www.obs-nancay.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Napier [Napeir, Nepair, Nepeir, Neper, Napare, Naper, Naipper], John [Jhone] (1550–1617)

Born in Merchiston Castle, Edinburgh, Scotland, educated in protestant theology at St Andrews University, and on the continent of Europe (Paris, Italy and the Netherlands). Invented logarithms, formulae used in solving spherical triangles (*Napier's analogies*) and *Napier's bones*, an 'analog computer' used for multiplying dividing and taking square and cube roots. Logarithms were a breakthrough for the astronomical calculation of planetary and cometary orbits to the large number of decimal places that high precision required, and Napier himself wrote of his motivation to invent them: 'Seeing there is nothing (right well-beloved Students of the Mathematics) that is so troublesome to mathematical practice, nor that doth more molest and hinder calculators, than the multiplications, divisions, square and cubical extractions of great numbers, which besides the tedious expense of time are for the most part subject to many slippery errors, I began therefore to consider in my mind by what certain and ready art I might remove those hindrances'. It was through the use of logarithms that KEPLER was able to reduce BRAHE's observations, determine his laws and pave the way for Newton's theory of gravitation. LAPLACE wrote that logarithms, 'by shortening labor, doubled the life of the astronomer'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Narlikar, Jayant Vishnu (1938–)

Theoretical astronomer, born in Kolhapur, Maharashtra, India, became professor at the Tata Institute of Fundamental Research. Worked with FRED HOYLE on the steady state theory of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NASA

The guiding principles for US exploration of air and space have remained remarkably consistent for more than 80 years. In 1915, Congress created an organization that would 'supervise and direct the scientific study of the problems of flight, with a view to their practical solutions'. That organization, the National Advisory Committee for Aeronautics (NACA), evolved into NASA four decades later.

In 1958, NASA was established and the journey begun in 1915 has been continued by aviators, astronauts and robotic spacecraft from the dunes of Kitty Hawk to the edge of the atmosphere and to the surface of the Moon and Mars. American spacecraft have explored more than 60 worlds in our solar system, while methodically peering back in space and time to reveal many of the secrets of the universe.

NASA employs 18 500 civil servants and generates thousands of high-tech jobs in the private sector. The Agency operates nine Field Centers, the contractor-operated Jet Propulsion Laboratory, and the Wallops Flight Facility. The Fiscal Year 2000 NASA budget is $13.6 billion.

NASA provides a variety of on-line services via the Internet. The URL's listed below provide key information about each of the topics. In addition, photographs, scientific and technical information, news releases and organizational data are accessible through a number of World Wide Web sites. Other key starting points for exploring NASA on the 'net' are listed below.

- NASA's main web site is:
  http://www.nasa.gov
- The NASA public affairs web site is:
  http://www.nasa.gov/hqpao/hqpao_home.html
- The Administrator's speeches web site is:
  http://www.nasa.gov/bios/goldin_speeches.html
- The NASA astronaut biographies web site is:
  http://www.jsc.nasa.gov/Bios/
- The International Space Station web site is:
  http://station.nasa.gov/
- The Hubble Space Telescope web site is:
  http://oposite.stsci.edu/1998/19
- The Tropical Rainfall Measuring Mission (TRMM), used to shed light on La Niña, web site is:
  http://www.eorc.nasda.go.jp/TRMM

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NASA Infrared Telescope Facility

The NASA Infrared Telescope Facility (IRTF) is the principal observatory in the world dedicated to solar system astronomy; 50% of the observing time is specifically reserved for such observations. It is a 3 m telescope located at 4145 m on the summit of Mauna Kea, on the island of Hawaii. It has been operated for NASA by the University of Hawaii's Institute for Astronomy since it opened in 1979. It is optimized for infrared observations and has imaging and spectroscopic instrumentation covering the region; active optics also provide very high angular resolution.

For further information see
http://irtf.ifa.hawaii.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nasir al-Din al-Tusi (1201–74)

Astrologer, astronomer and mathematician, born in Tus, Khorasan (now Iran), astrologer to the Isma'ili governor Nasir ad-Din 'Abd ar-Rahim. Captured and imprisoned in Alamut, in a fortress which was the headquarters of the terrorist Assassins, he betrayed the defenses to the invading Mongols. He was rewarded by Ilkhanid dynast Hulagu with an observatory, built with Chinese assistance at Meragha in 1262, and fitted out with instruments some of which were designed by al-Tusi himself. He used them in a 12-year program to observe planetary movements, and published the *Ilkhanic Tables* of planetary and star positions. Al-Tusi wrote about astronomical instruments, translated and commentated on classical works and researched into trigonometry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nasmyth Telescope

An altazimuth reflecting telescope with relatively stable platforms for mounting heavy, large, delicate or developmental equipment which cannot be, or has not been, engineered to cope with attitude changes during the tracking of a star. The optical configuration is the Cassegrain type, with a primary and secondary mirror, and an additional third flat mirror mounted at the intersection of the altitude and azimuth axes. The third, plane mirror, known as the 'Nasmyth flat', reflects the light beam from the secondary mirror through a right angle, along the altitude axis, and through an altitude bearing to a focus at the side of the telescope. Analysis equipment is supported at the Nasmyth focus on a platform mounted on the pillar that supports the altitude bearing. The Nasmyth platform rotates on the azimuth bearing, typically relatively slowly at a speed of one revolution per day (more during the transit of a target star across the meridian), but remains horizontal. The Nasmyth flat can be turned about the optical axis to switch the light beam of the telescope from one bearing to the other, so a Nasmyth telescope usually has two Nasmyth platforms. The William Herschel Telescope on La Palma is a Nasmyth design. The Gemini telescope project considered and rejected the Nasmyth design in order to avoid the thermal effects of the third mirror and its support mechanism in the light beam.

James Nasmyth was an English mechanical engineer and gentleman astronomer who invented this telescope configuration in his less agile later years as a means to provide a convenient viewing arrangement, in which he sat on one of the platforms and viewed the sky comfortably through an eyepiece at the altitude bearing.

See also: Altazimuth mounting; Cassegrain telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nasmyth, James (1808–90)

Engineer, born in Edinburgh, invented the steam hammer and many other devices in mechanical engineering especially for canal working. A childhood interest in astronomy meant that he maintained an interest in astronomical instruments in his professional engineering career, making two conventional small telescopes for himself, and visiting STRUVE during a visit to St Petersburg, for example. He made speculum mirrors and telescopes for WILLIAM LASSELL, WARREN DE LA RUE, and himself. The largest telescope which he made in 1845 had a 20 in mirror, and had a novel mechanical design, which 'brought the stars down to him'. It was an altazimuthally mounted Cassegrain-type, but a third mirror in the optical path reflected the light out of the telescope tube along the altitude axis. This arrangement is called a *Nasmyth telescope mounting*, and is used on all large modern telescopes. Nasmyth sat comfortably and stably with his eye at the center of the altitude bearing, operating handles to move himself and the telescope in azimuth, and to move the telescope in altitude, making his observations of the Moon and the Sun. On the Sun he observed, during occasional moments of superb seeing, what he termed a 'willow leaf' pattern; this appears to have been the discovery of solar granulation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Astronomical Observatory, Japan

The National Astronomical Observatory of Japan (NAOJ) was founded in 1988, integrating the Tokyo Astronomical Observatory of the University of Tokyo, the International Latitude Observatory of Mizusawa, and the solar radio group of the Research Institute of Atmospherics, Nagoya University. It is an inter-university research institute for astronomy under the jurisdiction of the Ministry of Education, Science, Sports and Culture. The headquarters is located in Mitaka near Tokyo. It has six division groups, and the scientific coverage is quite wide, ranging from geodesy to radio astronomy, even including the research and development of gravitational-wave detectors. Some groups are participating in joint space science missions such as the solar x-ray telescope (Yohkoh), radio telescope (Halka).

A total of about 300 permanent staff are working in NAOJ. There are nine ground-based research facilities: Mizusawa Astrogeodynamics Observatory; Norikura Solar Observatory; Okayama Astrophysical Observatory; Dodaira Astronomical Observatory; Nobeyama Solar Radio Observatory; Nobeyama Cosmic Radio Observatory; Astronomical Data Analysis Center; Advanced Technology Center; Public Relations Center; World Solar Data Center; and Subaru Telescope Hawaii Facility, which is the overseas facility having an 8 m optical/infrared telescope 'Subaru' on Mauna Kea, Hawaii.

Since April 1992 NAOJ has been participating in the Department of Astronomical Science, School of Mathematical and Physical Science, the Graduate University for Advanced Studies, and fulfilling the responsibilities of postgraduate education.

For further information see
http://www.nao.ac.jp/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Centre for Radio Astrophysics

India's National Centre for Radio Astrophysics (NCRA), located on the Pune University Campus, is part of the TATA INSTITUTE OF FUNDAMENTAL RESEARCH. At Khodad, 80 km from Pune, NCRA has set up the Giant Metrewave Radio Telescope (GMRT), the world's largest telescope operating at meter wavelengths. GMRT consists of 30 fully steerable dishes of 45 m diameter, spread over a 25 km area. Another meter-wavelength facility operated by NCRA is the Ooty Radio Telescope (ORT), a unique 530 $\times$ 30 m steerable parabolic cylinder located along a hill slope in southern India. NCRA has about 20 astronomers, 30 engineers and other technical and support staff. NCRA's research activities are mainly in the areas of: cosmology and extragalactic sources like radio galaxies, quasars etc; galactic sources including pulsars, supernovae remnants, etc; Sun and solar wind; and the interstellar medium.

For further information see http://ncra.tifr.res.in/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Radio Astronomy Observatory (NRAO), Tucson

One of the radioastronomy facilities funded by the US National Science Foundation, and operated under cooperative agreement by Associated Universities, Incorporated. Located in Tucson on the campus of the University of Arizona.

The NRAO operates and maintains a 12 m telescope on Kitt Peak, 80 km southwest of Tucson, for astronomical observations at millimeter wavelengths. It is the only millimeter-wavelength telescope in the US operated full-time as a national facility.

Originally known as the 36 Foot Telescope, the instrument was responsible for the birth of millimeter-wavelength molecular astronomy. Dozens of molecular species in the interstellar medium were first detected at Tucson.

In 1984, the telescope's reflecting surface and surface support structure were replaced and the 36 Foot was renamed the 12 Meter. Current research covers studies of molecular clouds and galactic star formation, evolved stars, astrochemistry and external galaxies.

Engineers at NRAO Tucson are also involved in development and support for the ATACAMA LARGE MILLIMETER ARRAY (ALMA).

For further information see
http://www.tuc.nrao.edu/Tucson.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Research Council

The National Research Council of Canada (NRC) has the Parliamentary mandate 'to operate and administer any astronomical observatories established by the Government of Canada'. NRC exercises this mandate through its HERZBERG INSTITUTE OF ASTROPHYSICS, which provides astronomical facilities, and related infrastructure and services to university scientists and their students. NRC also provides a time service through its Institute of National Measurement Standards.

For further information see
http://www.nrc.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Solar Observatory

The National Solar Observatory (NSO) operates several solar telescopes on Sacramento Peak in Sunspot, New Mexico, on Kitt Peak near Tucson, Arizona, and in a global network with six sites around the world. These telescopes provide unique capabilities and have an unsurpassed wealth of focal plane instrumentation available, including the first solar adaptive optics system. The NSO mission is to provide state-of-the-art telescope facilities to the solar community and to conduct pioneering research in solar physics. Highlights include the discovery of solar filigree (small magnetic flux tubes), elucidation of the global nature of the solar five-minute oscillations, and mapping of the interior structure of the Sun. The world's first and best-instrumented vacuum tower telescope, the 76 cm Richard B Dunn Solar Telescope, the Evans Coronal Facility (40 cm coronagraph), and Hilltop flare patrol are located at Sacramento Peak. The McMath–Pierce, located on Kitt Peak, is the world's largest (1.5 m) solar telescope and it is the premier facility for infrared studies of the Sun. The Kitt Peak Vacuum Tower Telescope makes daily magnetic maps of the entire visible solar disk. The six GLOBAL OSCILLATION NETWORK GROUP telescopes continuously measure solar oscillations. Research at the NSO concentrates on the nature of solar convection and magnetism, the origins of solar activity and variability, the structure and heating of the solar corona, and the interior structure of the Sun. The NSO was founded in 1983 by merging the solar programs at Sacramento Peak Observatory (1949) with those at Kitt Peak (1950). Most of the funding for the NSO comes from the National Science Foundation, with contributions from the USAF and NASA.

For further information see
http://www.nso.noao.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

National Space Development Agency, Japan

The National Space Development Agency (NASDA) was established by the Japanese government in 1969 to act as the nucleus for the development of space exploration and to promote the peaceful use of space and its practical applications.

The agency's headquarters are in Tokyo, but it also operates launch facilities at Tanegashima Space Center in Kagoshima and tracking stations at Katsuura, Masuda (Kagoshima) and Okinawa. Its main test laboratories and operations buildings are located at Tsukuba Space Center.

NASDA is responsible for the development of communications, meteorological and Earth observation satellites, such as the Engineering Test Satellite series. A number of NASDA experiments and astronauts have flown on the Space Shuttle. Future plans include the addition of the Japanese Experiment Module to the International Space Station. It also oversees the development of launch vehicles, such as the H-II rocket and the Hope-X experimental spaceplane.

NASDA carries out studies into future development and utilization of the Moon and planets, including research into development of materials for use on the Moon and into planetary orbiters and robotic surface explorers. The first practical example of this will be the Selene lunar orbiter–lander mission, a joint project with the INSTITUTE OF SPACE AND ASTRONAUTICAL SCIENCE.

For further information see http://www.nasda.go.jp/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Navigation

Navigation is the knowledge required to sail a ship between two known points by the shortest appropriate way, and in the least possible time. *Coastal navigation* is practised within sight of the coast, and orientation is provided by local data such as the coastline, knowledge of conspicuous landmarks, and the water depth. *Oceanic navigation*, by contrast, is possible only by observation of celestial bodies. Here, a distinction should be made between the passive use of astronomical knowledge based purely on what was visible in the sky, and the active use of scientific instruments and astronomical EPHEMERIDES.

Navigation by observation of the Sun and stars

The level of astronomical knowledge of the Babylonians, Phoenicians, Greeks and Arabs is well documented, but although it is probable that such knowledge was used to improve the navigation of seafarers, little is known about the actual transfer of information between the two disciplines in Antiquity and the early Middle Ages. Seafarers of these peoples were mainly coastal navigators, but they covered large distances and sometimes ventured beyond the sight of land, especially in the enclosed and therefore relatively safe Mediterranean. The first seaman whom we are told steered by stars is Odysseus. At one stage of his imaginary voyage, he had to keep the stars of the Great Bear on his left hand, night after night: the method of steering by horizon stars. A star was observed shortly after rising (or before setting) and used as a compass direction. After a while, the next rising (or setting) star that had similar amplitude was used to steer by in the same way, and thus the ship was held on the same course. A simple way of finding latitude at night which was used in the Mediterranean, was by observation of zenith stars. Seafarers memorized which star was overhead in a particular port. By observing the star's position during their voyage they could see if they were to the north or the south of their destination, and adjust their course accordingly.

During daytime the position of the Sun was observed and used for orientation. Culmination indicated south (or north), while in low latitudes sunrise and sunset pointed to two other major compass directions, east and west. In northern latitudes the Pole Star, although some degrees off the celestial North Pole, also indicated north when observed in combination with the Great Bear. Even south of the equator experienced seafarers could, at least for a while, continue to use the Great Bear to locate north. The Pole Star was also used to determine latitude by means of a primitive finger system. Around AD 650, Arabs, who were experienced in finding their way in the desert by observing stars, used a compass rose of which the points were named after the rising and setting of fixed stars.

The Vikings, when crossing large stretches of ocean, used the Pole Star for orientation at night, and the Sun during daytime. They probably also used these celestial bodies for setting out their courses.

Because of the great distances from the continents and between the islands, the inhabitants of the Pacific had to sail large stretches of water beyond the sight of land. Early documentary evidence about their history is extremely sparse, but information that European voyagers collected from the eighteenth century onwards suggests that Pacific islanders used the Sun for their orientation and steering. Like Mediterranean seamen they steered by horizon stars and used zenith stars for reaching the latitude of their destination.

The dawn of astronomical navigation

When, in the fifteenth century, the Portuguese embarked on voyages along the west coast of Africa their courses were generally north–south. They found their latitude by observing the Pole Star at night, and during daytime the Sun's altitude at noon. Beyond the equator, which was reached by 1474, it was no longer possible to observe the Pole Star. Although the Southern Cross had been seen by Portuguese seamen as early as 1454, it was found that there was no southern Pole Star. This led them to attempt to improve the method of finding latitude from the Sun at noon. Accordingly, the astronomer Abraham Zacuto of Salamanca and his pupil José Vizinho compiled the first day-to-day declination tables of the sun, the *Regimento do Astrolabio e do Quadrante*. When Columbus sailed towards America in 1492, he was supplied with such tables in manuscript. It has been suggested that the first printed edition appeared in 1495, but the only known printed copy dates from about 1509.

The ongoing maritime expansion of both Portugal and Spain called for cooperation between astronomers and navigators. Scientific instruments that had traditionally been astronomers' tools were adapted for navigational purposes. The mariner's ASTROLABE was developed out of the astronomical astrolabe, probably around 1445, and was employed to measure angular altitudes. The astronomer's cross-staff (used to measure the angular separation between celestial bodies) underwent the same transformation, perhaps by the German-born astronomer Martin Behaim, and was used as a navigating instrument after about 1515. Astronomers were involved in teaching and examining navigators, and they were also responsible for providing navigational instruments and correcting charts. Stimulated by commercial successes, Portugal and Spain created institutions that were centers of astronomy, cartography and navigation. The *Casa de Guinea e India* was established in Lisbon, the *Casa de Contratación* in Seville.

After the Atlantic Ocean was crossed in 1492, and the Indian Ocean in 1497, east–west courses were sailed more frequently, and there was an increasing need to measure longitude with accuracy. The difference in longitude between two places is of course equivalent to the difference in local time. In 1494 Columbus tried to find his longitude in the New World from observing a lunar eclipse, and comparing the local time of its occurrence with the time predicted for its occurrence in Spain. The effort failed because of the inaccuracy of the prediction, a consequence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of imperfect knowledge of the Moon's motion. Martin de Magellan's astrologer Andrés de San Martín in 1519 unsuccessfully tried to find the longitude in South America by conjunction of the Moon and Jupiter; failure was similarly the result of errors in the predicted astronomical ephemeris.

The Nuremberg astronomer Johann WERNER, in 1514, was probably the first to suggest that longitude could be determined by measuring lunar distances, the angle between the Moon and a fixed star. The Moon moves relatively rapidly across the sky, and at a given time the angle between the Moon and a given star is virtually the same for every observer who can see them both. The angle between the Moon and a number of stars can be calculated ahead for certain dates and times, for a standard meridian. An observer measuring such an angle would compare the result, and his local time, with the predicted angle and time. When the angles were the same, the difference between the two times would provide the difference in longitude between the observer and the standard meridian. Werner's method is feasible, but until the second half of the eighteenth century it was technically impossible to develop the required instruments and lunar tables.

Astronomy in navigation

In 1598, as the attempts to solve the question of longitude had so far failed, the Spanish king offered a considerable prize for anyone who devised a practical solution. This example was soon followed by the Dutch States-General, the governments of France and Venice, and in 1714 by the British Parliament. Many claims resulted, one of them from GALILEO GALILEI, who proposed that the eclipses and occultations of Jupiter's satellites could provide a solution, as they are visible at the same instant for every observer. For his proposal Galileo received a gold chain from the Dutch States-General in 1636; nevertheless, it was hardly possible to observe Jupiter's satellites from the deck of a moving vessel. Lunar eclipse observations to determine longitude were carried out successfully by Edmond HALLEY at sea near the Cape of Good Hope in 1719, and on land by Captain James COOK in Canada in 1766; but lunar eclipses are rare events.

The method of lunar distances called for an accurate and manageable observing instrument, an accurate mathematical theory of the lunar motion, and accurate tables of the positions of the fixed stars. Important steps towards the third of these goals were taken with the founding of the OBSERVATOIRE DE PARIS in 1667 and the ROYAL OBSERVATORY, GREENWICH in 1675. Both were commissioned to rectify the existing tables of stars to help solve the problem of longitude at sea. The Paris Observatory was the first institution to publish tables of astronomical data for navigation, such as the daily celestial positions of the Sun, Moon and planets, and the Right Ascension of stars. These appeared in the *Connaisance de Temps*, for the first time in 1678. From 1761 onwards the *Connaisance* also contained lunar tables, accurate lunar tables having been calculated for the first time a few years previously by the German mathematician Tobias MAYER. The British Astronomer Royal, Nevil MASKELYNE, became the driving force behind the *Nautical Almanac and Astronomical Ephemeris*, which first appeared in 1765, for the years 1767–9, and also contained lunar tables. Around the time Mayer calculated his tables, the requirement for accurate observing instruments was also fulfilled, with the development of the sextant, and later the reflecting circle.

The lunar method was soon firmly established as a reliable way of determining longitude at sea. Cook was among the first to apply it, in New Zealand in 1769, using the first edition of the *Nautical Almanac*. For many years this method, which was strongly promoted by Maskelyne, competed with that of finding longitude by use of a chronometer. The Astronomer Royal strongly believed that the astronomical solution was the only accurate answer to the longitude question, and opposed the 'mechanical' one. Eventually, however, the chronometer won, once the required horological technology had been developed, for it was immeasurably simpler to use than the method of lunar distances. Nevertheless, lunar distances remained in use throughout the nineteenth century, and until 1907 the necessary tables were published annually in the *Nautical Almanac*. The method was used at sea by those who could not afford an expensive chronometer, and also for checking the rate of chronometers before radio time signals were introduced around 1904.

Once the solution for the longitude problem had been found, astronomers became less involved with navigation. Astronomical ephemerides, however, remained indispensable to navigation, and the nautical almanacs that were annually published in various seafaring countries were of course compiled by astronomers.

Meanwhile there had been parallel improvements in methods of determining latitude. In the eighteenth century the Dutch teacher of navigation, Cornelis Douwes, developed a method of double altitudes by which latitude could be found without restriction to the traditional noon observation. This method was improved in the nineteenth century, when position line navigation gradually also came into use. This is the name for various ways by which, through astronomical observation and subsequent calculation, a ship's position could be found at almost any time of the day. Various methods were developed by which a single observation of the sun or a star would result in a position line being drawn on the chart. The intersection of several such lines, taken from different stars, provided an accurate position. Most of these methods were developed by merchant or naval seamen; but occasionally astronomers were involved, among them the Frenchman Yvon Villarceau who, in the 1860s, helped develop a method which was named 'La nouvelle navigation astronomique'. By the end of the century celestial navigation had reached the level at which it would remain until long after the Second World War. It was adapted for aviation, and seamen and aviators found their position by taking observations of the Sun, selected

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars, the planets and (occasionally) of the Moon, and the use of a nautical almanac. Matters changed when satellite navigation was introduced, in the late 1970s. A position could then be obtained instantly, almost as often as one wished, and it was far more accurate than celestial navigation. In the 1980s and 1990s global positioning systems became cheaper and easier to obtain, and the use of satellites started to replace sextants and nautical almanacs. The first sign that astronomy and navigation will soon part company altogether came from the United States Naval Institute in Annapolis. There, in 1998, it was decided to abandon altogether the teaching of celestial navigation. Future American naval officers, when crossing an ocean, will have to rely entirely on satellites and computers.

Bibliography

Cotter C H 1968 *A History of Nautical Astronomy* (London: Hollis and Carter)

Hourani G F 1951 *Arab Seafaring in the Indian Ocean in Ancient and Early Medieval Times* (Princeton, NJ: Princeton University Press) (2nd edn 1995)

Lewis D 1972 *We, the Navigators: The Ancient Art of Landfinding in the Pacific* (Canberra: Australian National University Press)

Schnall U 1975 *Navigation der Wikinger: Nautische Probleme der Wikingerzeit im Spiegel der schriftlichen Quellen* (Oldenburg and Hamburg: Gerhard Stalling)

Taylor E G R 1956 *The Haven-Finding Art: A History of Navigation from Odysseus to Captain Cook* (London: Hollis and Carter)

Williams J E D 1992 *From Sails to Satellites. The Origin and Development of Navigational Science* (Oxford: Oxford University Press)

Willem F J Mörzer Bruyns

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

n-Body Problem

Calculations involving the gravitational interaction of an arbitrary number (n, greater than two) of masses. The motion of two bodies is easily analyzed but there are no general solutions for three or more bodies interacting. Solutions do exist for particular cases but these take vast amounts of computation. Examples are the restricted three-body problem to determine the effect of Jupiter's perturbations of the asteroids and the use of successive approximations of series solutions to sequentially add the effects of smaller and smaller perturbations for the motion of the Moon. Examples requiring the solution of the n-body problem include the motion of a space probe in the solar system and the orbits of stars within a cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NEAR Shoemaker

The Near Earth Asteroid Rendezvous (NEAR Shoemaker) mission is the first mission to orbit one of the ASTEROIDS. NEAR, a spacecraft of NASA's new Discovery program was launched successfully on 17 February 1996 on a looping 3 yr trajectory to EROS, one of the NEAR-EARTH ASTEROIDS (NEAs). Starting on 14 February 2000, NEAR undertook a 1 yr investigation in orbit around Eros. On 12 February 2001, the mission was terminated with a controlled descent to the surface in order to provide extremely high-resolution images. On its way to Eros the NEAR spacecraft passed within 1225 km of asteroid 253 Mathilde on 27 June 1997. The spacecraft carried five science instruments: a multispectral imager (MSI), a near-infrared spectrometer (NIS), an x-ray and γ-ray spectrometer (XGRS), a magnetometer (MAG), and a laser range finder (NLR). A sixth investigation consisted of analyzing the radio signals of the spacecraft telemetry to measure the effects of the gravitational attraction of the two asteroid targets in order to determine their masses. On its year-long orbital study of Eros, NEAR gave unprecedented views of the asteroid surface with the highest-resolution images obtained so far for a small body.

Introduction

The NEAR Shoemaker mission is the first space mission to an NEA. It was named in honor of the late Eugene M. Shoemaker, a pioneer in the study of asteroid and comet impacts on Earth and other planets. NEAs are a dynamically young population of objects having orbits close to the Earth's. The general idea on the NEAs' origin is that these objects are efficiently removed from other regions of the solar system by collisions and subsequent gravitational interactions with the planets on a time-scale of 10^6–10^8 yr. Since the NEAs have unstable orbits, a continuous resupply of new objects is needed. In fact, some of them can come from the main belt and some can be derived from dead or dormant COMETS. It is believed that most METEORITES are collisional fragments of NEAs, but the meteorites may be a biased and incomplete sample of the materials actually found in NEAs. About 2000 near-Earth objects (NEOs) of all sizes are currently known, but the entire population contains perhaps 1000 objects with diameters larger than 1 km. These objects, thought to be similar in many ways to the ancient planetesimal swarms that accreted to form the planets, are interesting and highly accessible targets for scientific research. They carry records of the solar system's birth and the geological evolution of small bodies in the interplanetary region. Their

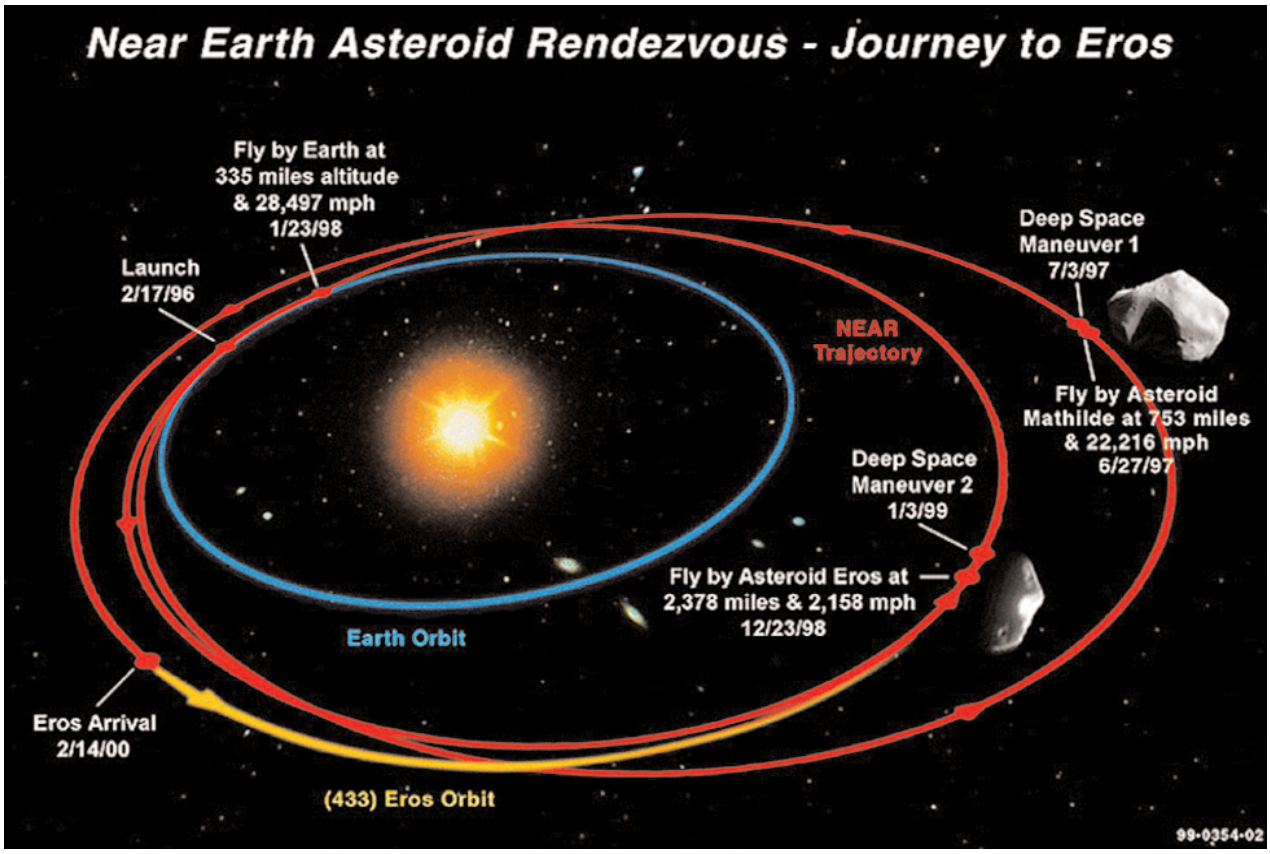


Figure 1. NEAR mission trajectory. Credit: NASA.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

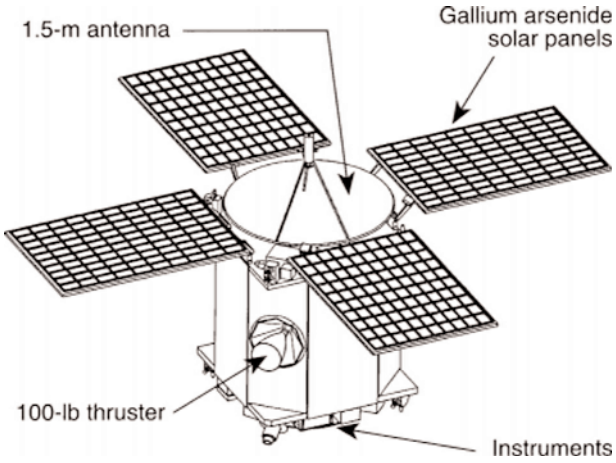


Figure 2. Spacecraft configuration. The solar panels and the fixed high-gain antenna are mounted on the forward deck. The MAG is on the high-gain antenna feed. The other instruments are mounted around the spacecraft adapter. From left to right, they are MSI, NIS, NLR and XGRS. Credit: NASA.

relations with comets, main belt asteroids and meteorites will clarify the links among these small bodies. Because collisions of NEOs with Earth pose a finite hazard to life, the exploration of these objects is particularly urgent. They may also serve as resources for use by future human exploration missions.

Mission design

NEAR was launched on a Delta II-7925 rocket into a 2 yr Earth-gravity assist trajectory (figure 1). The spacecraft passed, during its 23 month orbital loop around the Sun, close to the asteroid 253 Mathilde. NEAR carried out a close flyby (at 460 km) with the Earth on 22 January 1998 to obtain a gravity assist to change the orbital inclination. NEAR was originally planned to rendezvous with Eros in January 1999, but it did not complete its rendezvous burn on schedule, and the ground controllers lost all contact with the spacecraft for 27 h. NEAR was recovered and performed a first flyby of Eros on 23 December 1998. NEAR executed another flyby of Eros on 14 February 2000, and began its year-long orbital mission the same day.

Spacecraft and instruments

The NEAR spacecraft (figure 2) is a solar-powered, three-axis-stabilized spacecraft with a 1.7 m^2 base. The mass is 468 kg (805 kg including propellant). NEAR uses X-band telemetry to the NASA Deep Space Network. The spacecraft structure is composed of aluminum honeycomb decks connected to eight aluminum honeycomb side panels. On the outside of the forward deck there are a fixed X-band high gain antenna of 1.5 m diameter, four fixed solar panels and the solar x-ray monitor system. The NIS has a scan mirror that allows it to look more than 90° away from the common boresight. The spacecraft carries a complement of six science instruments: MSI, NIS, XGRS, MAG and NLR, in addition to a radio science investigation. The MAG is mounted on top of the high-gain antenna feed where it is exposed to the minimum level of spacecraft-generated magnetic fields. The remaining instruments (MSI, NIS, NLR, XGRS) are all mounted on the outside of the aft deck. They are on fixed mounts and are co-aligned to view a common boresight direction.

MSI

MSI is a 537 pixel × 244 pixel charge-coupled device camera with five-element refractive telescope, radiation hard refractive optics. The telescope is $f/3.4$ with a 168 mm focal length. The imager provides a field of view of 2.25° × 2.9° with a pixel resolution of 95 nrad × 161 nrad. The spectral range covers from 400 to 1100 nm. A filter wheel has seven spectral filters (450, 550, 760, 900, 950, 1000, 1050 nm). The main goal of the MSI was to determine the mineralogical and morphological characteristics of the surface at high spatial resolution.

NIS

NIS is a scanning spectrometer which covers the near-infrared wavelength region 800–2600 nm with 64 spectral channels. The NIS slit has two fields of view: 0.38° × 0.76° and 0.76° × 0.76°. A gold scan mirror that rotates over an angle of 140° controls the direction of viewing. Spectral images can be built up by a combination of spacecraft motion and the scan mirror. NIS has a gold calibration target that can be placed to provide a quantitative, in-flight spectral calibration. The InGaAs (indium–gallium arsenide) focal plane array does not require cooling to liquid nitrogen temperature to achieve a high signal-to-noise ratio. NIS was very important for studying the surface composition and, in particular, in determining the distribution and abundance of the asteroid surface minerals, such as olivine and pyroxene.

XGRS

XGRS consists of two sensors: an x-ray fluorescence spectrometer and a γ-ray spectrometer. Redundant solar monitors are used to determine the incident solar x-ray spectrum. The x-ray fluorescence experiment uses three gas-filled proportional counters to cover x-rays in the energy range from 1 to 10 keV. The x-ray spectrometer detects the x-ray line emissions excited by solar x-rays from major elements of the surface. The γ-ray spectrometer uses a body-mounted NaI scintillator and a BGO (bismuth germanate) anticoincidence shield that define a 45° field of view. It detects characteristic γ-rays in the 0.3–10 MeV range emitted from specific elements in the surface. The XGRS is the primary experiment for determining surface/near-surface element composition of the asteroid 433 Eros. In fact x-rays are emitted from the surface (less than a millimeter deep) while γ-rays tend to probe deeper material, up to about 10 cm in depth.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

MAG
MAG is a fluxgate magnetometer that uses ring core sensors made of highly magnetically permeable material. The sensor is mounted approximately 1 m from the main spacecraft bus, on the high-gain antenna feed support. MAG conducted the first close-up search for coherent body-wide magnetism at an asteroid. This search was motivated by the ambiguous evidence of magnetic signatures detected during the Galileo flybys of Gaspra and Ida.

NLR
NLR is an incoherent, direct-detection laser altimeter. NLR uses a chromium–neodymium–yttrium–aluminum garnet (Cr–Nd–YAG) solid-state laser and a compact reflecting telescope. The transmitter generates a brief laser impulse, and the instrument measures the time required for the light to reach the asteroid and return. It measured the distance from the spacecraft to the asteroid surface.

Radio science
There are no dedicated radio science (RS) instruments. The radio transponder allows the spacecraft's position and velocity to be determined in inertial space. This X-band transponder operates at 8438 MHz and is capable of determining the spacecraft's radial velocity to better than 0.1 mm s^{-1} when the data are properly calibrated. The coherent X-band transponder on NEAR was used to conduct an RS investigation by measuring the Doppler shift due to the spacecraft radial velocity component relative to the Earth. Accurate measurements of the Doppler shift allowed us to determine the gravity field and therefore the mass of the asteroid. Despite the NEAR mission being part of the Discovery program, which is synonymous with its being fast and cheap, the instruments were the results of many technical innovations. The x-ray instrument, the γ-ray detector system, the NLR and the InGaAs focal plane array for NIS all flew in space for the first time.

Scientific objectives

The primary objectives of the NEAR Shoemaker mission were the rendezvous with an NEA, and to achieve orbit around it to conduct the first scientific exploration of an object of this population. NEAR flew past Eros on 14 February 2000 at a closest approach distance of 300 km, continued the mission through several altitude orbits and spent 6 months in a 35 km orbit. The spacecraft passed successfully within 1200 km of the main belt 253 Mathilde on 27 June 1997. The measurement objectives of the mission were to determine (1) the bulk properties (size, shape, mass, density, gravity field and spin orientation), (2) the surface properties (elemental and mineralogical composition, geology, morphology, and texture), (3) the internal properties (search for heterogeneity and magnetic field) and (4) to search for possible satellites. All this information was important to characterize the physical and chemical properties of the target asteroids to study their mineralogical composition, to better understand the relationship between comets, asteroids and meteorites, and to improve the understanding of the boundary conditions of the processes that governed the formation and early evolution of the planets.

Figure 3. This mosaic image of Mathilde is constructed from four images acquired when NEAR was at a distance of 2400 km. Details as small as 380 m can be observed. The surface exhibits many large craters, including the deeply shadowed one at the center. Credit: NASA.

Mathilde flyby

On its way to Eros the NEAR spacecraft passed within 1225 km of the asteroid 253 Mathilde on 27 June 1997 at a velocity of 9.9 km s^{-1}. During the encounter with this C-type carbon-rich asteroid, NEAR was at 1.99 AU from the Sun, and because the solar panels were built to provide adequate power for the Eros rendezvous at closer solar distances, only the MSI was turned on. NEAR does not have a movable instrument platform and the whole spacecraft had to be turned to point the camera. NEAR took spectacular images and determined Mathilde's mass to an accuracy of about 5%. The spacecraft approached Mathilde looking close to the direction of the Sun (at an approaching phase of about 140°). The imaging sequence started 5 min before closest approach, obtaining a resolution of about 500 m. The highest resolution (160 m) was obtained at a distance of 1225 km (closest approach), when the phase angle was close to 90°. Then the imaging sequence continued for another 20 min, obtaining images at a phase angle of about 40°. During this time multicolor images with the seven color filters were obtained. Mathilde was on about 330 of the 534 MSI frames obtained. Because of the very slow rotational period of Mathilde (17.4 days), only roughly half of its surface was seen (figure 3) during the 25 min of the imaging coverage. The determined size dimensions supposing a triaxial ellipsoid shape are 66 km × 48 km × 46 km with a volume of $78\,000\ ^{-11\,000}_{+12\,000}$ km^3. On the 50% of the surface imaged by the spacecraft there are five craters with diameters between 19 and 33 km. A total of 91

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

craters have been identified and 70 have a diameter larger than 0.6 km. Analysis of the Mathilde craters shows that those craters larger than 5 km exceed the empirical equilibrium densities seen elsewhere in the solar system and that the largest craters are near the geometric saturation limit (see CRATERS). The similarity of the densities of small craters on Mathilde and IDA suggests similar ages for the surfaces; in the case of Ida, the age has been estimated at 1–2 billion yr. The crater ages depend on the estimation of the collisional strengths of asteroids, but in the case of Mathilde the preservation of giant craters and the low density of the interiors do not allow us to estimate the asteroid's collision age in a reliable way. The crater morphology and actual evidence of downslope movement on crater walls indicate that the surface of Mathilde is covered with regolith, indicating some retention of ejecta produced during crater events. An ALBEDO between 0.035 and 0.050 has been determined. The NEAR colors (between 400 and 1100 nm) show similar results to those previously obtained by ground-based telescopic observations, matching the CM meteorite Murchison, and showing no variations of albedo. On the basis of the volume determined by images ($78\,000\ ^{-11\,000}_{+12000}$ km^3) and the mass determined by the RS instrument ($(1.024\pm 0.058)\times10^{17}$ kg), a mean density of 1.1–1.5 g cm^{-3} has been determined. The density is less than half of that measured for CM meteorites, indicating a porous or underdense structure for the interior of Mathilde. Several hypotheses have been produced to explain this low density, such as water ice in the interior, a rubble pile structure, etc. It is also possible that C-type objects consist of unprocessed primitive material and the low density, in some sense, may be a primordial signature. About 200 images were taken during the closest approach to cover the sky around Mathilde at a distance between 300 and 600 km (10–20 radii) to search for possible satellites. The existence of a satellite could explain the low spin. No objects have been found down to a limit of 40 m across. The NEAR's photometric observations at Mathilde confirm that the ground-based methods used to determine the albedos and global photometric properties of distant asteroids are reliable.

Main results on Eros

Images from the MSI as well as near-infrared, x-ray, laser and radar measurements have been returned from NEAR during its orbital mission around Eros. MSI images were obtained at much higher resolution (~5 m/pixel) than available for any other asteroids and at even higher resolutions (couple of cm/pixel) during landing descent. In addition, γ-ray data were obtained at the landing site.

The asteroid 433 Eros, discovered in 1898, was classified as a typical S-type (silicate-rich) asteroid with a rotational period of 5.27 h but few details on its composition were available from ground-based observations. Eros has shown an irregular, curved shape with dimensions of 34.4 km × 11.2 km × 11.2 km, slightly smaller than previously estimated from the ground (figure 4). The images show chains of craters, numerous blocks, grooves and long ridges, e.g. the Rahe dorsum that extends for several kilometers across the surface. Almost all the features on Eros were named after famous historical and fictional lovers from a sampling of different cultures. The name of the largest feature, Himeros, comes from Greek mythology. Himeros, an attendant of the god Eros, personifies the longing of love.

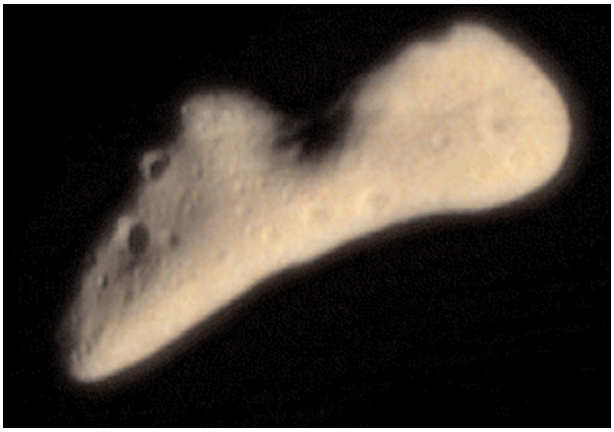


Figure 4. This image of 433 Eros was acquired on 12 February 2000 by NEAR's MSI. The largest depression, Himeros (about 9 km across), is visible in the center top of Eros. Credit: NASA.

These features, and other data, suggest that Eros is a consolidated body rather than an agglomerate of loosely bound, much smaller components. Huge tectonic features are found, which cannot occur in the absence of material strength. Steep slopes are found, where unconsolidated material would be expected to slide off. Many craters have shapes that indicate the influence of material strength in their formation, such as the square craters shown in figure 5. The rubble pile structure—e.g. an agglomerate of smaller component bodies bound mostly by gravity—has been inferred for many asteroids, but it does not apply to Eros. Its irregular, peanut shape houses a homogeneous internal structure. Although Eros is consolidated, the ubiquitous fabric of ridges and grooves suggests an extensively fractured interior.

Composition

The infrared spectrometer showed variations in the mineral composition such as pyroxene, olivine and iron-bearing minerals. The first x-ray results revealed the presence of magnesium, iron, silicon and also aluminum and calcium. Both x-ray data and near-infrared spectra are most consistent with a primitive, largely undifferentiated composition for Eros. Eros, as an undifferentiated body, is made of material that has been scarcely altered since the beginning of the solar system, and probably since before the formation of the Earth. The relative proportions of rock-forming elements in Eros are similar to those found in the Sun itself. Eros was never subjected to

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

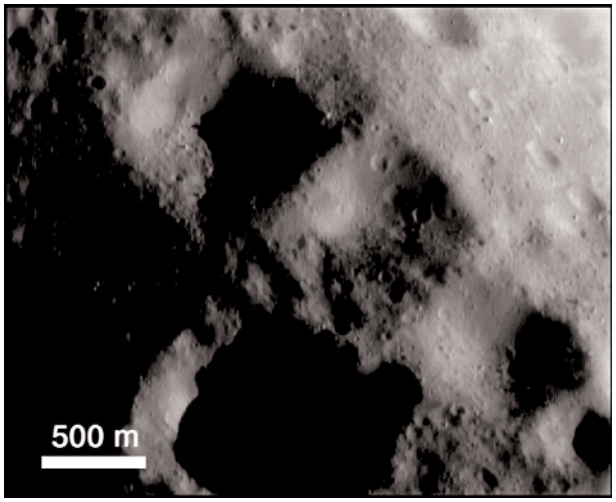


Figure 5. Squared craters on Eros: an example of a structurally controlled feature consistent with Eros as a consolidated body. This image was taken on 26 April 2000, when NEAR Shoemaker was 50 km above the surface. Credit: NASA.

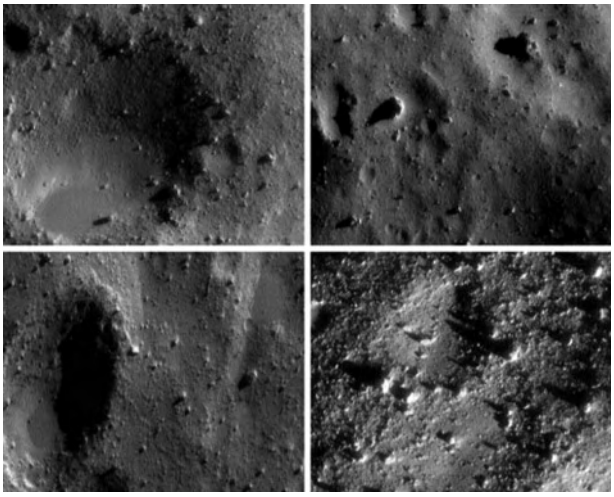


Figure 6. Details of Eros' surface. From upper left to lower right, the images show Eros' featured surface at increasing resolution. Blocks are ubiquitous on Eros. The image at upper left was taken on 27 January of a point 13.5 km away and shows a ponded deposit inside the crater; the one at the upper right was taken on 26 January from 11.1 km away. Signs of blocks partly buried by regolith are apparent. Each upper scene is about 550 m across. The image at bottom left was taken on 26 January from 4.9 km away, and the bottom right image was taken on 28 January from a similar distance. Each lower scene is about 230 m across. Credit: NASA.

the melting and the separation into compositionally distinct layers that Earth, Mercury, Venus and Mars have all experienced (see DIFFERENTIATION). X-ray measurements probing the uppermost 100 μm of surface material showed that abundances of Ca, Al, Mg, Fe and Si are consistent with compositions of either ordinary chondrites or certain primitive achondrites. However, there is a marked depletion of S compared with expectations for these meteorite types. This could be due either to removal of a small fraction of partial melt (which would imply that Eros is primitive achondritic) or to devolatilization of a thin surface layer by impacts.

Using ground-based Doppler and range measurements—and by tracking surface landmarks—scientists have determined that the asteroid's mass is $(6.687 \pm 0.003) \times 10^{15}$ kg and density is 2.67 ± 0.03 g cm^{-3}, which is about the average density of Earth's crust but is less than the average bulk density of ordinary chondritic meteorites, indicating that bulk Eros is significantly porous and/or highly fractured, but not to the same extent as Mathilde. The density is relatively uniform throughout the asteroid. There is a small center-of-mass offset with respect to the center of the asteroid's image that may indicate an underdense regolith layer of up to 100 m depth.

Craters and rocks

As with other asteroids, craters, almost certainly of impact origin, characterize the surface morphology of Eros. The three largest depressions are Himeros (about 9 km across), Shoemaker (7.6 km) and Psyche (5.3 km). The presence of a large number of craters indicates a rather old object. Many large craters have bright albedo features on their interior walls. The morphological and spectral properties of these features, associated with steep slopes, are consistent with a darkened, space-weathered material that slides down crater walls exposing brighter, less weathered underlying material. The space weathering process alters the asteroid regolith which, exposed to the interplanetary environment (micrometeorite impact, solar wind, cosmic rays,…), progressively changes color with time. The asteroid surface appears darker and its reflectance spectrum is redder (with much weaker absorption bands) than the spectrum of its constituent rocks. It has been known that such processes are responsible for altering the lunar regolith and data analyses from Galileo's Ida encounter have provided convincing evidence that the space-weathering process is operating on Ida. However, thorough investigations of spectral properties of these processed materials suggest that space weathering affects the regoliths on Eros and the Moon differently, perhaps because of differences in gravity or location in the solar system.

Flyover images have revealed a profusion of blocks up to 150 m across (see figure 6). These rocks are believed to have been produced as ejecta blocks from large craters. Most of these rocks were inside the Shoemaker crater, positioned near one end of the potato-shaped asteroid. The positions and orientations of the blocks indicate that most of them originated from the relatively young Shoemaker crater. Blocks have shapes ranging from angular to rounded and clodlike. Aprons of debris at the base of some of the larger blocks indicate that varying

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

degrees of erosion exist at the surface of the asteroid, whose origin is not clear.

Ponded deposits

Another one of the big surprises scientists found from NEAR's highest-resolution mapping is that Eros' surface appears to have a global cover of regolith, i.e. surface dust and debris. It is estimated to be tens of meters in depth. The surface appears to be blanketed with a fine material, some of which has created flat deposits, particularly in depressions, such as craters. These fine deposits appear to have been sorted from the upper portion of the asteroid's regolith. These so-called 'ponded' deposits are intriguing because of their extremely smooth and flat surfaces. Their slightly bluish color may be attributed to the optical properties of submicrometer particles. The probe landed in one of these ponds, revealing its contents to be like dry cement (see figure 6). Ponds have not been seen on any other solar system objects and their origin is not clear. Some researchers hypothesized that the ponds formed after seismic shaking or by electrostatic levitation. For the latter effect, particles can build up photoelectric charges with long exposure to the Sun, and this charge might separate out finer particles.

Eros has a stable rotation and an escape velocity that ranges from 3.1 to 17.2 m s^{-1}, which would allow a baseball thrown from its surface to leave forever. The acceleration of gravity on the surface of Eros varies from 2.3 to 5.5 mm s^{-2}, thousands of times smaller than on Earth. A person who weighs 70 kg on Earth would weigh from 16 to 39 g on Eros.

No satellites down to a limiting size of 50 m have been observed. Another negative result was the lack of magnetization of Eros. Even the close look-up at the landing site showed no evidence for an intrinsic magnetic field, which is a surprising result. Indeed, most meteorites, including chondrites, are much more magnetized than Eros.

Conclusion

NEAR provided the first ever close-up look at an asteroid and has increased drastically our knowledge of primitive bodies in the solar system. The spectacular images and surprising results that the spacecraft returned have started to answer the many questions we had about Eros and asteroids but they also revealed new mysteries that scientists will explore for years to come. A lot of work remains to be done with the wealth of data returned by NEAR.

Bibliography

Bell J and Mitton J 2002 *Asteroid Rendezvous: NEAR Shoemaker's Adventures at Eros* (Cambridge: Cambridge University Press)

Cheng A F, *et al* 2001 Laser altimetry of small-scale features on 433 Eros from NEAR-Shoemaker *Science* **292** 488

Cheng A F 2003 Near Earth Asteroid Rendezvous: Mission Summary. Asteroid III book (pp. 351–366)

Robinson M S, Thomas P C, Veverka J, Murchie S and Carcich B 2001 The nature of ponded deposits on Eros *Nature* **413** 396

Thomas P C, Veverka J, Robinson M S and Murchie S 2001 Shoemaker crater as the source of most ejecta blocks on the asteroid 433 Eros *Nature* **413** 394

Veverka J, *et al* 1997 NEAR's flyby of Mathilde: the first look at a C-type asteroid *Science* **278** 2109

Veverka J, *et al* 2001 Imaging of small-scale features on 433 Eros from NEAR: evidence for a complex regolith *Science* **292** 484

Veverka J, *et al* 2001 The landing of the NEAR-Shoemaker spacecraft on asteroid 433 Eros *Nature* **413** 390

Alain Doressoundiram

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Nearby Groups of Galaxies

Groups of galaxies are gravitationally bound aggregates of tens of galaxies, like the LOCAL GROUP (LG), the home of our MILKY WAY GALAXY. What we call a 'nearby' group has to be sufficiently close such that the group can be studied, in terms of structure and contents, with similar (but of course always inferior) detail as the LG. This is roughly met by groups lying in the Local Supercluster—a flattened metagalactic structure of galaxies approximately 120 million light-years (40 Mpc) across and centered on the VIRGO CLUSTER. The LG is in the outskirts of this system. There may be more than 100 groups in the local supercluster, but this is impossible to pin down exactly, as the identification, or definition, of a group gets increasingly difficult with increasing distance: faint group members are harder to detect at greater distances and there is increasing confusion due to projection effects (several groups can be projected on top of each other along the line of sight).

Some of the more conspicuous nearby groups were undoubtedly known already to the early workers in extragalactic astronomy, like E Hubble and H Shapley. But it really was Gerard de Vaucouleurs, in the 1960s, who made the first successful attempt to disentangle the essential supergalactic systems (groups, clouds and clusters) from the sparse data on the distribution of nearby galaxies available at the time. His work is still a major reference in the field, and his nomenclature was largely adopted.

Here we confine ourselves to the six groups of galaxies that are found within a distance of 10 Mpc (or 30 million light-years), excluding the LG. These most nearby groups are of course also the best-studied ones. They are described here one by one. Some of the groups, like the M81 Group and the Cen A Group, are fairly rich (several times the LG) and dense, with clear signs of galaxy interaction. The M101 Group, on the other hand, is very poor: it is as a single giant galaxy with a swarm of dwarf satellites. Others still (CVn I and Sculptor) are even denied the title of a group, they are regions of slightly enhanced galaxy density without appreciable internal motion, called 'clouds'. Finally, there is a large group (IC 342/Maffei) behind the dust absorption band of the Milky Way. Most of its fainter members may still be hidden from our view. But more surprising than the structural variety of the groups is perhaps their external interconnectedness. It seems that the groups are embedded in a coherent supergalactic structure. For example, the LG cannot strictly be separated from the Sculptor Cloud; and there may be other 'cloudy' bridges between the groups.

The neighborhood of the Local Group

The distribution of galaxies in the universe is highly non-uniform. On a scale of about 100 Mpc one finds a sponge-like network of 'filaments' and 'sheets' of galaxies, with rich CLUSTERS OF GALAXIES typically lying at the vertices of these structures. The sheets and filaments are dynamically cold (non-virialized) systems whose internal structure is probably fractal, with (still non-virialized) subsystems on any scale being called 'clouds'. Embedded in the cold clouds of predominantly late-type galaxies (SPIRAL and IRREGULAR GALAXIES) are—as 'hot spots'—dynamically more virialized groups and loose clusters of galaxies. These, depending on density, may contain an appreciable fraction of early-type galaxies (ELLIPTICAL GALAXIES, S0s).

Our home sheet is the Local Supercluster, whose primary hot spot is the loose Virgo cluster. Our home cloud is the 'Coma-Sculptor Cloud', and our home group is of course the Local Group (LG). The Coma-Sculptor Cloud is a cigar-shaped feature of length 15 Mpc, stretching from Sculptor in the anti-Virgo direction all the way to the Coma I Cluster, which is roughly in the direction of Virgo at $D \approx 15$ Mpc. Several known nearby groups belong, like the LG, to the greater environment of the Coma-Sculptor Cloud. Others seem to be connected with neighboring clouds. There are apparently no isolated groups.

Due to this embedding of the groups and also their internal substructure, a clear-cut definition of even a nearby group (in fact, even the LG!) is very difficult to attain. A primary factor is the absolute magnitude down to which galaxies are recorded. At very faint levels every giant galaxy will have a swarm of bound dwarf companions and could therefore be regarded as a group of its own (like the Milky Way Galaxy with its MAGELLANIC CLOUDS and DWARF SPHEROIDAL satellites). So the definition of a group is, to a large extent, a matter of convention.

Groups within 10 Mpc distance

Our most complete knowledge of the LG neighborhood is confined to objects less distant than 10 Mpc from the Milky Way Galaxy. Around 300 (mostly dwarf) galaxies are at present known to lie in this volume. Their sky distribution is shown in figure 1. The strong concentration of the galaxies towards the supergalactic plane, the symmetry plane of the local supercluster, is striking. Roughly 70% of the galaxies appear to be members of a small number of groups; but the remaining 'field galaxies' are unlikely to be isolated but will be members of the embedding cloud(s). Six groups, excluding the LG, are conventionally adopted. Two of these, Sculptor and CVn I, have from the beginning (i.e. from de Vaucouleurs' pioneering study) been known to be rather clouds. As well-distinct nearby supergalactic structures they are treated here along with the other, real groups.

Figure 2 shows the spatial distribution of the galaxies projected onto the supergalactic plane. The cloud nature of Scl and CVn I—by their large extension—is obvious; both appear as mere substructures of the Coma-Sculptor Cloud. A clear separation between the LG and the Sculptor Cloud is almost impossible and perhaps also not physically meaningful.

In the following the six major nearby groups and clouds are briefly described in some detail. The order is by supergalactic longitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

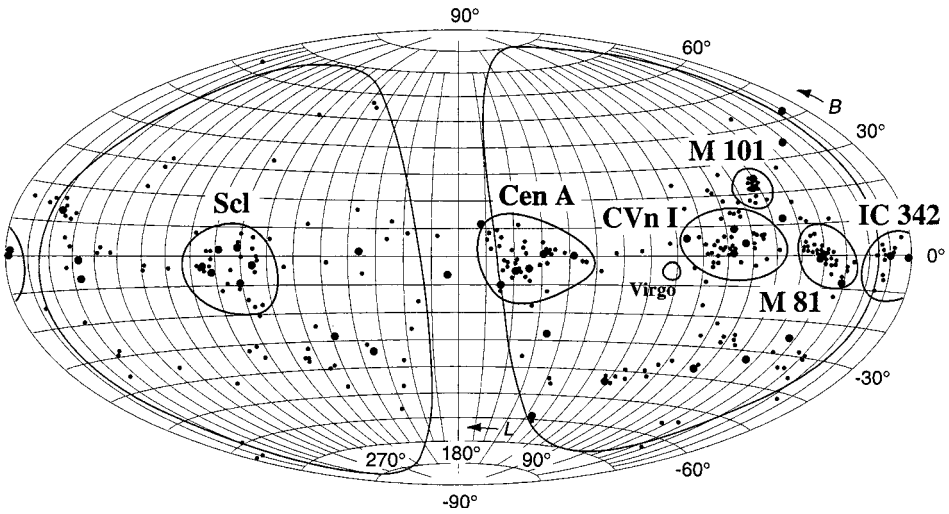


Figure 1. Sky distribution, in supergalactic coordinates, of all galaxies known within a distance of 10 Mpc. Approximate boundaries of the main groups in this volume, as well as the position of the Virgo Cluster, are indicated. The galactic zone of avoidance ($|b| < 15^\circ$) is delineated with a heavy line. Large dots, giant galaxies ($M_B > -18$); small dots, dwarf galaxies ($M_B < -18$).

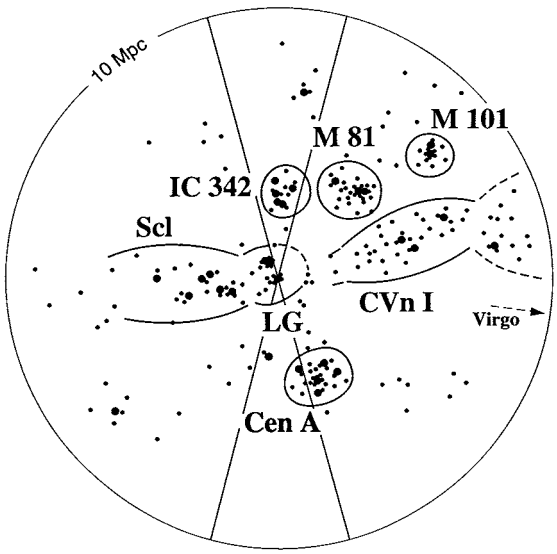


Figure 2. Spatial distribution, projected onto the supergalactic plane, of all galaxies lying closer than 10 Mpc. The Milky Way Galaxy is at the center, with the zone of avoidance as indicated by the double sector ($|b| < 15^\circ$). The orientation is also given by the direction to the Virgo Cluster. Approximate boundaries of the main groups and clouds, including the Local Group, are indicated. Large and small dots as in figure 1. This plot should not be taken as a true map, but rather as a schematic, as most distances used here are rough estimates only.

IC 342/Maffei Group

This galaxy group is named after the Italian astronomer Paolo Maffei who by chance, in the late 1960s, discovered two of its brightest members on a near-infrared-sensitive Schmidt plate. As it later turned out, Maffei discovered the tip of an iceberg whose extension and depth remain largely unexplored even today. What we now call the IC 342/Maffei Group is a fairly massive galaxy group at a distance of ≈3.5 Mpc which, unfortunately, happens to lie entirely in the zone of avoidance of our Galaxy (see figure 1). This means that all of its member galaxies are very heavily obscured by dust absorption, with a dimming of up to 10 (!) magnitudes in the visual. Thus the galaxies are essentially invisible in the optical, but as dust extinction is less severe at longer wavelengths, they may be detected in the near-infrared and radio wave bands. One of the main constituents of the group, Dwingeloo 1, was discovered as late as 1994. Without galactic absorption, Dw 1 would indeed be one of the 10 largest spiral galaxies in the sky.

The four brightest members of the group are IC 342 (of Hubble type Scd), Maffei I (elliptical), Maffei II (SBb), and Dw 1 (SBb). In luminosity they are comparable to our Galaxy and M31. Nine further, fainter galaxies, mostly of the dwarf irregular type, add up to a total of 13 group members identified so far. Clearly, there must be many more dwarf members of the group still awaiting discovery. The mean distance of the group has been controversial until recently, which is not surprising given the enormous amount of dust absorption. Early estimates placed the group at ≈2 Mpc distance, i.e. just barely beyond the classical confines of the LG. If true, this would have had strong consequences for the dynamics of the LG. The mean distance favored at present, 3 to 3.5 Mpc, is more comforting in this respect. However, the zone of avoidance of our Galaxy may harbor more surprises!

M81 Group

The M81 Group is the 'classical' nearby group in the Northern hemisphere, lying at high latitudes in Ursa Major

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(the Big Dipper). It is about two or three times as rich as the LG: around 40 members and possible members have been found so far down to a $M_B \approx -11$ (note that the faintest LG members have about -7). The mean distance of the group is ≈4 Mpc. There are five giant members in the core region: M81, the well known Sb spiral galaxy; M82, equally well known for its peculiar, amorphous appearance which is caused by its STARBURST activity; NGC 3077, again an amorphous; and the late spirals NGC 2976 (Sd) and IC 2574 (Sm). These galaxies (with the exception of IC 2574) have been observed to be embedded in a huge, dynamically complex cloud of neutral hydrogen. This—and probably the starburst activity of M82 as well—must be taken as a sign of ongoing interaction. The core region of the group is indeed rather dense, which, by way of the MORPHOLOGY–DENSITY RELATION GALAXY, is also suggested by the presence of a fair number of bright DWARF ELLIPTICALS. Notwithstanding, the majority of group members are faint dwarf irregulars which tend to lie more in the outskirts of the group. Very faint dwarf spheroidals of the type known in our vicinity have not yet been (but may soon be) detected around M81.

Studies of the M81 Group are traditionally hampered by a strong mixing of Galactic and extragalactic features. Here the problem is not dust obscuration but the presence of Galactic 'cirrus' (REFLECTION NEBULAE) and HIGH-VELOCITY CLOUDS in the region. Deep optical exposures reveil a disturbing criss-cross of cirrus on top of the galaxies. There are low-surface brightness objects whose nature (Galactic versus extragalactic, i.e. cirrus versus dwarf M81 Group members) is still unclear.

M101 Group

The M101 Group is probably the poorest group of all (including our LG). The group is completely dominated by M101 itself, a beautiful giant Sc spiral at a distance of ≈7 Mpc (which can be regarded as group distance). The second-ranked group member is already three magnitudes fainter than M101. Only about 12 members and possible members of the group are known to date, half of which lie very close to M101 and can therefore be considered as M101 satellites. With one exception they are all of the late spiral (Sd, Sm) and irregular (Im) type. One peculiar feature of the group is its LUMINOSITY FUNCTION OF GALAXIES: the population of very faint and diffuse dwarf galaxies (elliptical or irregular), which is so frequent elsewhere, is apparently simply missing here (the faintest member known is as bright as $M_B \approx -14$). Deep Schmidt searches and a recent blind radio survey have not resulted in a single new member of the group. However, it has to be acknowledged that the M101 group is the most distant one considered here. Future surveys of increased sensitivity will likely uncover the sought for exponential rise of the luminosity function.

Canes Venatici (CVn) I Cloud

What is called the Canes Venatici (CVn) I Cloud covers a huge area of sky. There is no obvious concentration of bright galaxies, which would be characteristic of a group, yet the roughly 30 or 40 spirals and irregulars in the area clearly belong together because they have similar velocities and (by implication) distances. G De Vaucouleurs rightly called the feature a cloud. The brightest (possible) cloud members are M64 in the South and M94 in the North, both Sab spirals of roughly Milky Way Galaxy luminosity. But these giants are well separated from each other and almost isolated. The most conspicuous (gravitationally bound?) clumping of galaxies in the cloud, near NGC 4214, involves about ten late-type spirals (Sd, Sm) and irregulars (Im) whose brightest members are not much brighter than the LARGE MAGELLANIC CLOUD. The lack of early-type galaxies is a clear sign of low galaxy density and is typical for clouds.

From the velocities of CVn I members one has inferred a prolonged structure of the cloud, roughly aligned with our line of sight. Direct distance measurements based on the brightest resolved blue stars in CVn I irregulars give 2.3 to 8 Mpc as distance range, i.e. depth of CVn I.

The whole CVn region is very complex. There are several clouds and groups/clusters projected on top of each other. Slightly to the North (in the mean) there is the CVn II cloud which is more distant but may simply be the extension of CVn I backwards to greater distances, merging with the UMa cloud in the North and the Coma I cluster in the South.

Centaurus A Group

The Centaurus A (Cen A) Group is the Southern analog of the M81 Group. It is equally rich and dense, with 40 members known at present, it is equally distant, with $D \approx 4$ Mpc, and there are also a number of giant galaxies, some of which show signs of strong interaction. The 'padrone' of the group is of course the famous peculiar galaxy and strong radio source Cen A (NGC 5128): it is the nearest giant elliptical, the nearest merger, and the nearest ACTIVE GALAXY (AGN) at the same time.

The other giant group members ($M_B > -18$) are M83, a magnificent Sc spiral galaxy, NGC 4945 (also Sc), and NGC 5068 (SBc). Other bright members of the group include NGC 5102 (S0) and NGC 5253 (Im). About 20 faint dwarf irregular and 10 dwarf elliptical members are known in addition. The group shows clear substructure, essentially a double structure (like the LG)—defined by the main clump around Cen A in the South, and a smaller subclump around M83 in the North. Both giant galaxies possess extended swarms of dwarf galaxy satellites. The dwarf ellipticals tend to be clustered around Cen A, as expected from the morphology–density relation of galaxies. The two hot spots of the group, the Cen A and M83 subclumps, are embedded in an oddly triangle-shaped cloud of dwarf irregulars.

Sculptor (Scl) Cloud

The Sculptor, or 'South Polar' Cloud (as it is sometimes called for its unique position in the southern sky), is the most nearby supergalactic structure beyond the LG. It is a prolate feature of length 3 Mpc and thickness 1 Mpc,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which we see pole-on because we are situated near one of its ends. Owing to the low mean galaxy density, there are no real giant, Milky Way-type galaxies, but one finds a handful of bright late-type spirals (Sc, Sd) plus about 15 fainter dwarf irregulars and 5 dwarf spheroidals.

The 28 Scl members known at present seem to fall into three or four (gravitationally bound?) subclumps, centered on NGC 300 (Sc) and NGC 55 (Sc) at the near side ($D \leq 2$ Mpc); NGC 247 (Sc) and NGC 253 (Sc, the most nearby starburst galaxy) at $D \approx 2.7$ Mpc; NGC 45 (Scd) and NGC 59 (dwarf S0) at $D \geq 4$ Mpc; and possibly NGC 7793 (Sd) at $D \approx 3$ Mpc. But not all Scl members are companions to these bright main members; the dwarf irregulars are dispersed in the whole cloud.

It has always been noted that the distribution of LG members is rather asymmetric, with an extended population of Southern dwarf members out to $D \approx 1.5$ Mpc, in the direction of Sculptor (like WLM, Tucana, Pegasus, IC 5152, *et al*). It now appears that these galaxies form a perfect bridge to the Scl Cloud, which means that the LG can be regarded as the major subclump (hot spot) in an extended 'Sculptor-LG Cloud' of dimension 6×2 Mpc (see figure 2), which itself is part of the very large 'Coma-Sculptor Cloud'. The consequences of this new picture for the dynamics of the Local Group have yet to be explored.

Bibliography

De Vaucouleurs G 1975 Nearby groups of galaxies *Stars and Stellar Systems. XI. Galaxies and the Universe* ed A Sandage, M Sandage and J Kristian (Chicago: Chicago University Press) p 552

Karachentsev I 1996 The Local Group in comparison with other nearby groups of galaxies *Astron. Astrophys.* **305** 33

Schmidt K-H and Boller T 1992 Nearby galaxies *Astron. Nachr.* **313** 189 (part I), 329 (part II)

Tully R B and Fisher J R 1987 *Nearby Galaxies Atlas* (Cambridge: Cambridge University Press)

Bruno Binggeli

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Near-Earth Asteroids (NEAs)

Asteroids whose orbits bring them into the inner solar system, by definition those with perihelion distances of less than 1.3 AU. They are also known as *Earth-approaching asteroids*; the term *near-Earth object* (NEO) is sometimes used in recognition that some of these objects are cometary in origin. There are three classes of NEA: Amor asteroids, Apollo asteroids and Aten asteroids. As of April 2000 the numbers known in each class were 450, 464 and 75, respectively. *Amor asteroids* are defined as those with a perihelion distance between 1.017 AU (the Earth's aphelion distance) and 1.3 AU, and are named after the 1 km diameter asteroid (1221) Amor. The largest are (1036) Ganymed, at 40 km, and (433) Eros, at 35×13 km; Eros was the first to be found and is the most studied, having been orbited by the NEAR Shoemaker probe. *Apollo asteroids* are defined as those with a perihelion distance less than 1.017 AU and semi-major axis greater than 1 AU. They are named after the 1.5 km diameter asteroid (1862) Apollo; the largest is the 8 km diameter (1866) Sisyphus. These two groups are sometimes classified together as *Apollo–Amor objects*. *Aten asteroids* are defined as those whose semi-major axis is less than 1 AU, and are named after the 1 km diameter asteroid 2062 Aten. The terms *Mars-crossing*, *Earth-crossing* and *Venus-crossing* are used for asteroids which intersect the orbit of a particular planet.

Three-quarters of NEAs are thought to be asteroids perturbed inwards from the main asteroid belt. The mechanism is believed to be based on chaotic dynamics near the Kirkwood gap corresponding to a 3:1 resonance with Jupiter. The others are thought to be extinct comet nuclei. The lifetime of NEAs is probably not much more than 10 million years, their likely fate collision with one of the terrestrial planets or gravitational ejection from the inner solar system via a close encounter. That objects continue to collide with terrestrial planets is evidenced by, for example, the geologically very recent Meteor Crater on Earth. In 1980 the Spacewatch Program was instituted to track down what have become known as *potentially hazardous asteroids* (PHAs): the closest-approaching NEAs that are considered to be large enough that an impact of one of them with Earth would endanger human civilization. Attempts to quantify the likelihood of future Earth impacts of PHAs suggest that a civilization-threatening event might occur once in several hundred thousand years—in other words, the odds of such an occurrence in a given century are several thousand to one. Estimates of the population of NEAs suggest that there may be 2000 bodies of 1 km or more, and 100 000 larger than 100 m. Objects of 100 m or less would cause local damage, as in the Tunguska event of 1908, while objects of around 10 m would burn up during their passage through the atmosphere. The closest recorded approach was 104 700 km by the 9 m diameter 1994 XM_1; for a 100 m plus object, the record is held by 1996 JA_1 (220 m) at 448 800 km.

Two asteroids hit the news in 2002 because of their predicted close approach to the Earth. On 9 July 2002, MIT astronomers discovered 2002 NT7, a 2 km-wide space rock in a curious orbit. Unlike most asteroids, which circle the Sun in the plane of the planets, 2002 NT7 follows a path that is tilted at 42°. It spends most of its time far above or below the rest of the solar system. Every 2.29 years, the asteroid plunges through the inner solar system not far from Earth's orbit. Researchers calculated there was a chance that 2002 NT7 might hit our planet on 1 February 2019. This made headline news, worrying some members of the public. What was not reported was that the odds of impact are 1 in 250 000.

On 14 July, astronomers discovered 2002 NY40. This asteroid measures about 0.800 km across, and follows an orbit that ranges from the asteroid belt to the inner solar system. Its closest approach was on 18 August 2002, passing only 1.3 times farther away than the Moon, and being visible through binoculars.

See also: KIRKWOOD GAPS, METEOR CRATER, NEAR SHOEMAKER, TUNGUSKA EVENT.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Nebula

A diffuse aggregation of interstellar matter, mostly gas but often also containing dust grains. The name (Latin for 'cloud') is applied to a variety of objects, visible either because they are bright and seen against the dark background sky or because they are dark and seen in silhouette against a brighter background object. The term is also used to describe objects that have a similar appearance when observed in wavebands outside the visible region, in the infrared for example. Until the later 1800s, a 'nebula' was any fuzzy patch that telescopes were incapable of separating into component stars. Spectroscopy demonstrated that some were indeed gaseous, and not unresolvable star clusters; others, that were found to lie well outside the Milky Way, and subsequently found to be other galaxies, were originally referred to as 'extragalactic nebulae'.

Bright nebulae may be either emission nebulae or reflection nebulae. Emission nebulae are self-luminous. Most shine because the gas from which they are composed is ionized by ultraviolet radiation from a nearby star; examples are H II regions and planetary nebulae. Other mechanisms include synchrotron radiation, responsible for the emission of light from supernova remnants such as the Crab Nebula. Reflection nebulae are cool clouds of dust and gas; the dust scatters the light of nearby stars. An example is the nebulosity surrounding the stars of the Pleiades and other star clusters, and many reflection nebulae are the remnants of the interstellar matter from which a cluster of stars was born. Dark nebulae are also known as absorption nebulae: they absorb light from nearby sources. The light is re-emitted as infrared radiation or scattered, making the nebulae appear dark. A famous example is the Horsehead Nebula.

Other uses of the term include protoplanetary nebula (also called a protoplanetary disk), for the cloud of matter around a newly formed star from which planets form (and also applied to an early stage in the development of a planetary nebula), and solar nebula, for the cloud of interstellar matter from which the Sun and the solar system formed.

See: dark nebulae, gaseous nebulae, H II regions, interstellar matter, interstellar molecular clouds, planetary nebulae, reflection nebulae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Needham, Joseph (1900–95)

Historian of science, he was educated at Gonville and Caius College, Cambridge, where he eventually became Master. A friendship with three young Chinese researchers in Cambridge inspired him to learn Chinese, and during the Second World War he was posted to Chongqing, where he became acquainted with Chinese achievements in science and technology, and was astonished. This inspired him to devote the rest of his life to research his massive work on *Science and Civilization in China*, including its astronomical history. Amongst the high points were the first recorded observation of sunspots, invention of quantitative cartography and the Mercator projection, pre-discovery of the solar wind through observation of comet tails, invention of equatorially-mounted astronomical instruments (see KUO SHOU-CHING), and observation of novae, supernovae, comets, auroras, meteors and other phenomena.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neptune

Orbiting the Sun in 164.79 years, at a mean distance of 30.06 AU (4.50×10^9 km), Neptune is the outermost of the four GIANT PLANETS. Its orbit is nearly circular (eccentricity 0.009) and inclined to the Earth's orbit plane by only 1.77°. Neptune's outer atmosphere, like those of the other giant planets, is composed predominantly of molecular hydrogen and helium with traces of methane (CH_4). Yet Neptune, together with URANUS, forms a distinct subclass of ice-rich Jovian planets. Most of the planet's mass is in the form of 'ices', primarily water, with possibly a 'rock' component. Despite its remoteness, Neptune's atmosphere is one of the most dynamically active in the outer solar system. Small bright clouds evolve with a timescale shorter than a day while large dark spots, slowly moving and oscillating in shape, persist for months. In the upper atmosphere, photolysis of methane initiates a suite of photochemical reactions producing heavier hydrocarbons and condensation hazes. Neptune possesses a large and complex magnetic field, much like that on Uranus, with a magnetic dipole offset from the planet's center and tilted by 47° to the spin axis.

Basic data

Table 1 presents basic physical data about Neptune, with some comparison with the Earth (⊕). Uranus and Neptune are similar in mass (14.53 and 17.14 Earth masses respectively), radius (25560 and 24760 km respectively), and mean density (1.27 and 1.64 g cm^{-3} respectively). They are smaller, less massive, and denser than the 'inner' pair of giant planets, Jupiter and Saturn.

The interior of the planet rotates in 16.11 h, slightly more rapidly than Uranus. This rotation is responsible for the slight flattening (or oblateness) of the planet of 1.7%, the polar radius at 1 bar being 420 km smaller than the equatorial radius.

The equatorial plane of Neptune is tilted by 29.6° to the orbit, causing significant seasonal variations of insolation. Neptune emits 2.6 times more energy than it absorbs from the Sun. Reflecting 29% of the incident sunlight, its equilibrium temperature, without any additional heat source, would be 46.7 K. The effective temperature, related to the total energy emitted by the planet, is significantly larger (59.3 K). The excess emission is due to the presence of a substantial source of internal heat, as is the case in Jupiter and Saturn but not in Uranus. This internal heat flow is thought to derive from cooling of Neptune's hot interior.

Table 1. *Neptune's physical parameters*

Mass	1.024×10^{26} kg	(17.14 ⊕)
Equatorial radius at 1 bar	24 760 km	(3.88 ⊕)
Equatorial gravity at 1 bar	10.9 m s^{-2}	(1.11 ⊕)
Oblateness	0.017	
Rotation period	16.11 h	(0.673 ⊕)
Obliquity of rotation axis	29.6°	
Effective temperature	59.3 K	
Bolometric Bond albedo	0.29	

Discovery and exploration

Neptune was first observed and identified as a planet in September 1846. The quest for Neptune actually began in 1781 when WILLIAM HERSCHEL accidentally discovered Uranus. Soon after it was recognized as a planet, it became clear that Uranus' orbit was 'irregular' and perturbed by some unidentified influence. Around 1834 came the first suggestions that a trans-Uranian planet might be causing the deviation from Uranus' predicted orbit. Independently, two astronomers, JOHN COUCH ADAMS from Cambridge University Observatory and Urbain Jean JOSEPH LE VERRIER from Paris Observatory, tackled the problem of the irregularities in the motion of Uranus. By 1845, both ended with very consistent predictions for the position of the disturbing planet. J C Adams wrote to JAMES CHALLIS, professor of astronomy, who in turn contacted George Airy, director of the ROYAL OBSERVATORY, GREENWICH, to conduct a search for the expected planet. His request, however, did not receive immediate attention. It was only in June 1846, on receiving a memoir from Le Verrier, that Airy decided to launch into the search and asked Challis to urgently undertake the observations. With no good map of the relevant sky area, the survey went on very slowly and Challis soon lost faith. On 18 September 1846, Le Verrier decided to contact Johann Gottfried Galle from Berlin Observatory, asking him to search at the position he had predicted. On 23 September, the same day he received Le Verrier's letter, JOHANN GALLE, assisted by a young student HEINRICH D'ARREST, turned the telescope to the expected position. On checking the stars in the field with the recent Berlin Academy's Star Atlas, Galle and D'Arrest very soon noticed an object not in the catalog. The hunt was over: Neptune had been discovered within 1° of its calculated position. This brilliant discovery, based purely on calculations of orbital perturbations, marked the triumph of celestial mechanics.

With a visual magnitude of 7.8, Neptune is too faint to be seen by the naked eye. Through a telescope, it appears as a bluish disk with an angular diameter of 2.3 arcsec at mean opposition. Photographic spectroscopy at the beginning of the century allowed the detection of absorption bands of methane (CH_4). The presence of hydrogen (H_2), the main atmospheric constituent, was only demonstrated in 1952 for all the giant planets. In the 1970s, spectrophotometric observations of the visible and near-infrared spectrum of Neptune yielded a first determination of the CH_4 abundance and of the pressure level of the upper haze layer. Emission bands from ethane (C_2H_6) and acetylene (C_2H_2) were also detected. The development of infrared astronomy allowed observers to measure the thermal radiation emitted by Neptune and to infer the existence of a large internal heat source. Numerous photometric observa-

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

tions in the near-infrared region established that the atmosphere of Neptune shows large temporal variations on timescales from hours to years. Direct imaging of Neptune using sensitive CCD cameras revealed the presence of discrete clouds rotating in the prograde direction, i.e. in the same sense as the orbital revolution. In the early 1980s, the presence of rings around Neptune was sought through stellar OCCULTATION techniques. After a perplexing series of positive and negative detections, a partial occultation was simultaneously detected in 1984 by independent telescopes on one side of the planet only. As a result, the presence of 'ring arcs' around Neptune was proposed.

But most of what we know about Neptune and its environment comes from the VOYAGER 2 flyby in August 1989. Thousands of images were recorded and observations were conducted in the visible, infrared, ultraviolet, and at centimeter wavelengths. Voyager images revealed the presence of prominent cloud systems including an Earth-sized 'Great Dark Spot'. Wind speeds were measured over most of the planet from tracking of cloud motions. The radio science experiment and ultraviolet solar occultations yielded vertical profiles of temperature at selected locations on the planet. The temperature field was mapped in the upper troposphere and lower stratosphere. The mixing ratio of helium (He) was determined as well as the abundance profiles of the main hydrocarbons in the stratosphere. The magnetic field experiment discovered a strong and complex magnetic field and an associated magnetosphere. The rotation period of the interior was determined from the periodicity of radio bursts emitted by Neptune. Broad and narrow rings circling the planet were observed, with the outermost having optically thicker arc segments (see NEPTUNE: RINGS). Six new moons were discovered with diameters between 50 and 400 km (see NEPTUNE: MINOR SATELLITES). Voyager's observations of TRITON, Neptune's largest satellite, revealed a geologically young surface, a covering of seasonal ice over the polar regions, a thin nitrogen (N_2) atmosphere with traces of methane (CH_4), and active geyser-like plumes.

Our knowledge of Neptune's atmosphere continued to grow beyond the wealth of data harvested by the Voyager spacecraft. Hubble Space Telescope (HST) imaging of Neptune in 1994–96 revealed changes in atmospheric features (dark spots and bright clouds) over that period. In addition, the pattern of zonal bands was different from that at the time of the Voyager encounter. Observations at large radiotelescopes yielded detections of two new compounds (hydrogen cyanide HCN, and carbon monoxide CO) in the stratosphere of Neptune. In 1997, the Infrared Space Observatory (ISO) orbiting the Earth recorded the solar reflected and thermal emission spectra of Neptune. These high spectral resolution observations led to the first detections of water vapor (H_2O) and carbon dioxide (CO_2) in the upper atmosphere; methyl radicals (CH_3), direct products of methane photolysis, were also observed.

Atmosphere

Current theoretical models suggest that the atmosphere of Neptune extends down to about 3500 km below the 1 bar level, up to a pressure of ~100 kbar, and overlies a denser fluid comprising mainly water, methane and ammonia. The mass of this outer layer, between 0.5 and 1 Earth mass, is a small fraction of the total mass of Neptune, in contrast to the gas-rich JUPITER and SATURN. Only a tiny fraction of Neptune's atmosphere, the upper 100 bar, is accessible to observations by remote sensors.

Thermal structure

Figure 1 shows a mean vertical profile of temperature in Neptune's atmosphere. In the lower atmosphere—the troposphere—the temperature increases with depth and with increasing pressure. The temperature minimum, located near the 0.1 bar level, defines the tropopause. This is the base of the stratosphere, where temperature increases with altitude. Above the stratosphere, at pressure levels less than ~0.1 microbar, lies the thermosphere.

The temperature profile in the troposphere and stratosphere mostly results from a balance between heating by the internal heat source and the sunlight absorbed and cooling by infrared thermal emission at wavelengths longer than ~7 μm. Deeper than ~0.5 bar, the dominant energy transport mechanism is convection and the lapse rate follows an 'adiabat', i.e.,

$$\frac{\mathrm{d}\ln T}{\mathrm{d}\ln p} = \frac{R}{C_p}$$

where T is temperature, p is pressure, C_p is the molar specific heat of the atmosphere and R the gas constant. Above this level, the atmosphere is more transparent, radiation dominates over convection, and the lapse rate is smaller than the adiabatic one. The 'temperature inversion' above the 0.1 bar level is due to the presence of various heat sources in the stratosphere, including absorption of solar energy by methane gas and cloud/haze particules.

Tropospheric temperatures on Neptune and Uranus are very similar despite Neptune's greater remoteness. The reason for this is the presence of a large internal heat

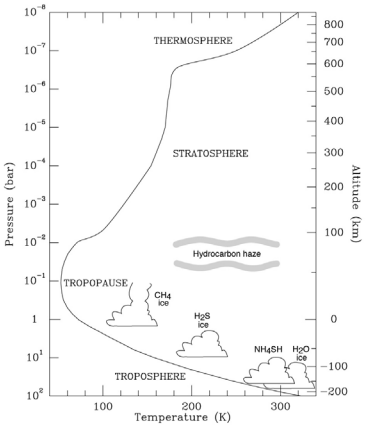


Figure 1. Vertical profile of temperature on Neptune as a function of atmospheric pressure and altitude above the 1 bar level. The location of clouds or hazes derived from thermochemical calculations and assumed abundances of constituents is indicated.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

flow on Neptune that nearly compensates for the reduction in solar heating from Uranus to Neptune. The slope of the temperature profile shows a small change near 1.9 bar. This corresponds to the region where methane condenses and forms haze particles. The change in the lapse rate results from the release of latent heat that accompanies methane condensation.

A striking feature of Neptune's thermal structure is the existence of a relatively warm stratosphere. The temperature reaches ~120 K at the 1 mbar level, increasing up to 170–180 K in the range 10–0.3 μbar. This is higher than observed in Uranus' stratosphere even though Neptune receives 60% less sunlight. This feature may be partly explained by the enhanced solar heating on Neptune due to the larger methane abundance and possibly to stratospheric aerosols that are optically thicker. Radiative heating, however, seems to be insufficient to support the warm temperatures observed and an additional non-radiative heating mechanism may be required. A possibility is the dissipation of GRAVITATIONAL WAVES propagating upward into the stratosphere.

At the base of the thermosphere, the temperature rises steeply to reach ~750 K near 10^{-5}μbar. This rise is caused by the downward transport of heat through molecular conduction. The source(s) of this high-altitude heating is not yet clearly identified. Some proposed mechanisms are dissipation of gravity waves, electron precipitation from the magnetosphere, Joule heating generated by the atmospheric motion in the magnetic field, or collision of supersonic jets generated by the auroral energy deposition (see AURORA).

Composition

Neptune's atmosphere is primarily molecular hydrogen (H_2) and helium (He). Hydrogen is present in both the symmetric (ortho) and antisymmetric (para) states—relative to the spin of the two protons—with an ortho/para abundance ratio consistent with thermodynamic equilibrium. The He mole fraction lies between 16 and 22%. The lower part of this range is compatible with the estimated abundance of He in the presolar nebula from which Neptune formed 4.6 billion years ago. Deuterium (D), the heavier isotope of hydrogen (H), is enriched in Neptune's atmosphere. The D/H ratio is inferred to be 0.5–1×10^{-4}, higher than the value in the primordial solar gas. This enhancement is due to incorporation of deuterium-rich ices during the formation of the planet (see below).

Beyond these two light elements, Neptune's observable atmosphere contains methane. The CH_4 mole fraction is about 2% in the 3 bar region. This corresponds to a ~30-fold carbon enhancement over the solar carbon-to-hydrogen ratio, thought to be representative of the presolar nebula (see discussion below). Absorption of sunlight at red wavelengths by methane is at least partly responsible for the bluish appearance of the planet. Above the 1.9 bar level (T~93 K), the methane abundance is severely depleted by condensation processes. The CH_4 abundance is then limited by the saturation law in the upper troposphere and decreases with altitude as temperature decreases. At the temperature minimum of ~53 K reached near 0.1 bar, the saturated CH_4 mole fraction is ~1×10^{-4}. As methane is carried upward through this 'cold trap', one would expect the stratospheric mole fraction not to exceed 10^{-4}. This is not the case and a mixing ratio ten times larger (~10^{-3}) is measured in the lower stratosphere. This situation, contrasting with Uranus' atmosphere, is discussed later.

A small amount of carbon is also in the form of carbon monoxide (CO). The CO mixing ratio of about 1×10^{-6} is orders of magnitude larger than expected from thermochemical equilibrium at Neptune's cold observable levels. Its presence is very likely due to rapid convective transport from the deep hot interior where it is stable.

Microwave measurements indicate that, in contrast to Jupiter and Saturn, ammonia (NH_3) is severely depleted in Neptune (and Uranus) down to at least the 30 bar pressure level. This is at first glance surprising, considering the large enrichment observed in carbon and the fact that both C- and N-bearing ices were probably present in the planetesimals that accreted to form Neptune. The most likely explanation is that NH_3 is removed by chemical reaction with H_2S (hydrogen sulfide) to form NH_4SH crystals. This, however, requires that H_2S is more abundant than NH_3, and consequently that the S/N ratio is at least five times larger than in the Sun. If H_2S is, like CH_4, enriched by a factor of 30 relative to the solar abundance ratio while NH_3 has a solar abundance ratio, the NH_4SH cloud should form near the 40 bar level and the corresponding gas profiles then provide a satisfactory interpretation of Neptune's microwave spectrum. An alternate possibility is that most of the nitrogen is in the form of N_2 rather than NH_3.

Due to the low atmospheric temperatures, water (H_2O) condenses out even deeper and cannot be probed by remote-sensing observations.

Photochemistry

In the upper stratosphere, methane undergoes photolysis due to ultraviolet radiation from both the Sun and the local interstellar medium. CH_4 photodissociation produces CH, CH_2 and CH_3 radicals. Subsequent reaction schemes of various complexity yield a suite of hydrocarbons, ethane (C_2H_6) and acetylene (C_2H_2) being the most abundant. These compounds slowly diffuse down to the lower (colder) stratosphere, where they are lost by condensation. Hydrocarbon ice particles eventually fall through the tropopause and the warmer troposphere, where they sublimate and are recycled into methane in the deep hot layers. The C_2H_6 and C_2H_2 mole fractions near the 0.1 mbar level are about 2×10^{-6} and 1×10^{-7} respectively. Neptune's carbon cycle is illustrated in figure 2.

A tiny fraction of hydrogen cyanide (HCN) is present in the stratosphere of Neptune. HCN production is initiated by the reaction between CH_3 radicals and N

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

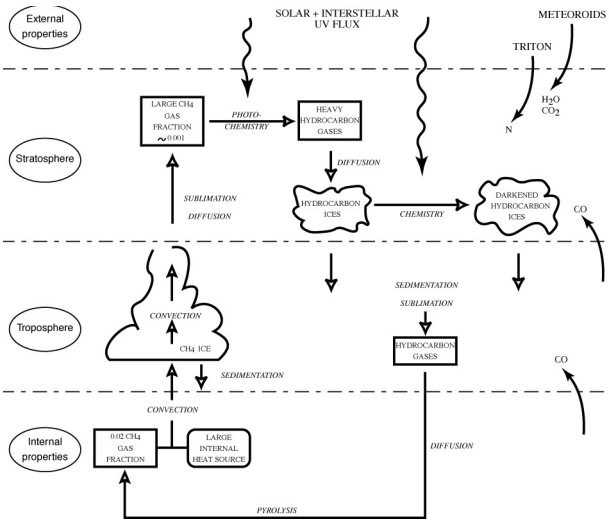


Figure 2. Chemical and physical processes governing the carbon cycle on Neptune. Methane is delivered to the stratosphere by convective plumes that form near the condensation region in the troposphere. Methane photodissociation in the upper atmosphere produces heavier hydrocarbons that diffuse downward to the lower stratosphere, where they condense out. Hydrocarbon ice particles then fall into the warmer troposphere, where they sublimate. Deeper in the atmosphere, hydrocarbon gases are pyrolized back to methane.

atoms. Two possible sources of N atoms can be envisaged: escape from Triton's atmosphere, or galactic cosmic ray impact on N_2 supplied from the deep interior via rapid vertical mixing. The HCN mole fraction is ~1×10^{-9} in the mbar region; condensation near the 2–3 mbar level drastically reduces its abundance at greater depths.

Neptune's upper atmosphere also contains small amounts of oxygen-bearing compounds having an external origin: water vapor (H_2O) and carbon dioxide (CO_2) with column densities of ~3 and 8×10^{14} molecules/cm2 respectively. Water probably originates from the continuous infall of interplanetary dust and meteoroids rich in water ice and silicates. The equivalent of a few liters of water falling every second on the planet is enough to explain the observed quantities. Carbon dioxide may have the same origin, as the infalling meteoroids probably contain CO_2 ice. Another possible source is the reaction of CO, present in Neptune's atmosphere, with OH radicals formed from the photolysis of H_2O.

Clouds and hazes

Formation of condensation clouds in the troposphere of Neptune is expected, considering the atmospheric temperature profile and the presence of various volatile gases. From thermochemical modeling, a solution cloud of liquid water with dissolved H_2S and NH_3 is predicted to form at pressure levels of one or a few hundred bars depending on the H_2O abundance. Moving upward, a water ice cloud is expected around 50 bar (T~270 K). Near 40 bar (T~250 K), reaction of NH_3 and H_2S produces a cloud of NH_4SH solid particles. The amount of H_2S in excess of NH_3, thus not incorporated in the NH_4SH cloud, is expected to condense out at ~5 bar (T~150 K). Methane, more volatile, forms the uppermost cloud layer in the troposphere with a condensation level at 1.5 bar (T~80 K). It should be noted that the precise location, thickness and vertical extent of the cloud layers in Neptune are in fact rather uncertain due to unknowns in the atmospheric composition and to the important role of cloud microphysics and dynamical motions (see CLOUDS IN PLANETARY ATMOSPHERES).

How does this cloud model correlate with actual observations of Neptune's atmosphere? Near-infrared spectroscopy and imaging indicate that the bottom of the visible atmosphere is at a pressure of about 3–4 bar and thus probaby corresponds to the top of the optically thick H_2S cloud. Above this background, the CH_4 cloud appears to be very thin on a planet average, with an optical depth as low as ~0.05 in the red. Neptune's atmosphere is not, however, homogeneous and shows a variety of light and dark features (see below). Some prominent bright cloud features are likely made of CH_4 particles near the 1 bar level. Dark ovals (figure 3) can be interpreted as holes in the H_2S cloud cover.

In the stratosphere, gases formed from CH_4 photochemistry in the microbar region are expected to condense out to form a haze as they diffuse downward. Ethane is probably the main constituent of the 'photochemical' haze (75% by mass), with the remainder provided by heavier hydrocarbons. The haze top, associated with the condensation level, is located at 10–15 mbar. On a disk-average basis, the mean stratospheric haze burden is 1–2 μg cm^{-2}. It is substantially lower in the equatorial region. This haze is composed of sub-micron particles which absorb the ultraviolet radiation from the Sun very efficiently. Linear cloud streaks observed by the Voyager spacecraft in the north have been associated with enhanced wisps of hydrocarbon ice particles.

Dynamic meteorology

Neptune has some of the fastest winds measured on any planet, despite the small amount of energy received from the Sun. The circulation pattern is characterized by large retrograde (westward) winds at the equator transitioning to prograde (eastward) winds poleward of ~50°. Wind speeds measured at the cloud tops vary from 400 m s^{-1} at the equator to 250 m s^{-1}at –70° latitude. The zonal wind structure is similar to that found on Uranus and much simpler than on Jupiter or Saturn. It corresponds to a nearly uniform value of the angular momentum for air parcels aloft over the planet. The observed zonal winds are probably forced by a system of symmetric zonal jets deeply rooted in the atmosphere (yet not in the interior). The amplitude of the winds decays with altitude because of eddy viscosity.

The temperature field in the upper troposphere shows a maximum at the equator, minima at mid-latitudes (~50°) and a rise towards the poles. The latitudinal

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

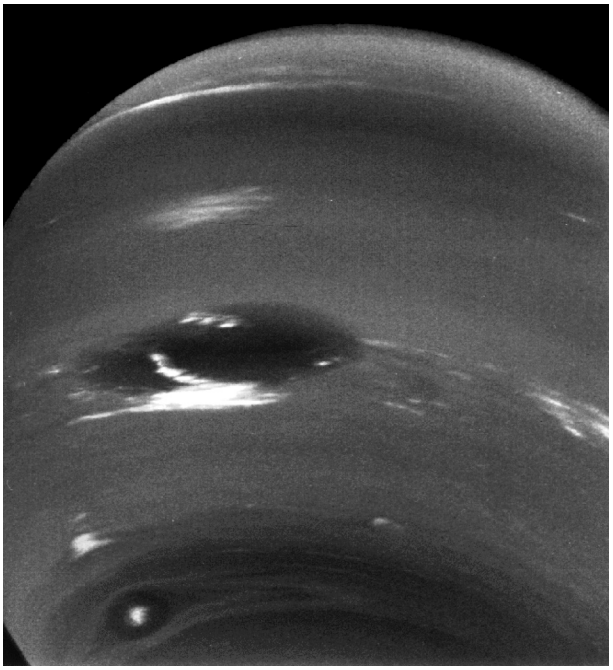


Figure 3. Image of Neptune taken by the Voyager 2 spacecraft in August 1989. The Great Dark Spot (GDS) accompanied by cirrus-like clouds is prominent. The 'Scooter', a small bright cloud feature, is visible to the lower left. Below the Scooter lies another dark spot (DS2) with a bright core (NASA/JPL).

contrast amounts to 3 K near 0.5 bar, 5 K at the tropopause (0.1 bar), and seems to persist in the stratosphere (~10 K around 0.2 mbar). The mid-latitude cool regions are thought to result from upwelling causing adiabatic expansion and cooling. Conversely, subsidence in the equatorial region and at the poles produces atmospheric heating and higher temperatures. The meridional velocities associated with this flow are small, of order 1 cm s^{-1}. The latitudinal temperature gradients and meridional circulation are in fact associated with the frictional damping of the zonal winds, as required by geostrophic balance.

Neptune's atmosphere is visually very active. The most remarkable feature revealed by the Voyager images is the Great Dark Spot (GDS), a dark oval spot in the southern hemisphere (Fig. 3). Between January and August 1989, the GDS moved steadily equatorward from –27 to –17° latitude. Like most of the Jovian vortices (see JUPITER: ATMOSPHERE), it is an anticyclone—a high-pressure center. The GDS was oscillating with an eight-day period, its latitudinal size varying between 12 000 and 18 000 km and its longitudinal size between 5200 and 7400 km. The source that was maintaining the GDS, given that some dissipation must occur, is unknown. Observations with the Hubble Space Telescope (HST) in 1994 showed that the GDS had disappeared or faded away. A second large dark spot (DS2) was present during the Voyager flyby at about –50° latitude. It was 5200 km wide in longitude (~20°) and 2600 km high in latitude (~6°). This spot had a bright central core which was showing strong convective activity. The latitude and longitude of DS2 oscillated with a 36-day period. In 1994 a new spot (GDS-94), similar in size to the GDS, was reported in the northern hemisphere at 31° latitude. It was still present at the same latitude eleven months later.

The dark spots GDS and GDS-94 are accompanied by bright cloud features which are probably composed of condensed methane. The GDS companion, located at its southern border, changed markedly over several hours (figure 4). In 1995 the bright companion of GDS-94 was located northward of the spot whereas, in 1994, a large complex of cloud features extended from slightly north of the spot almost to the equator. Near –41° latitude, a bright long-lived feature, named 'Scooter', was observed by Voyager. Relatively small (3° in latitude and 7.5° in longitude), it consists of many bright east–west streaks which change on timescales of days; its shape itself varies from rectangular to triangular. The Scooter is located beneath the methane cloud top, deeper in the atmosphere than the bright companions of the GDSs. In addition, numerous transient bright features were observed by both Voyager and the HST. In particular, the South Polar Feature (SPF), between the latitudes –68° and –74°, was a very active region at the time of the Voyager encounter, showing many short-lived linear cloud features.

A low-contrast zonal (banded) pattern is also visible on Neptune. The zonal structure present from 1994 to

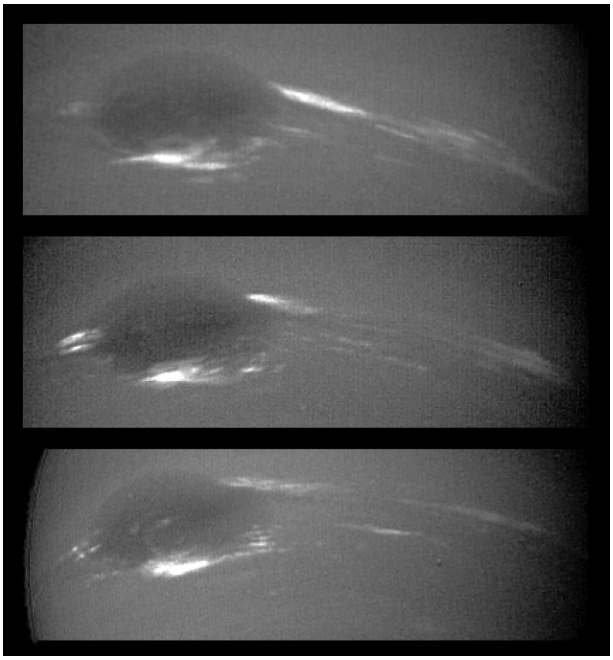


Figure 4. Cloud evolution in the region of the Great Dark Spot (GDS). Panels are separated by 18 hours. The bright cirrus-like clouds of Neptune change rapidly, forming and dissipating over periods of several to tens of hours (NASA/JPL).

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

1996 differed from that in 1989. The prominent dark band seen by Voyager between –40° and –68° was weaker and narrower (–55° to –70°) in 1994–1996.

Other evidence of activity is the presence of a significant abundance of methane in the stratosphere, an order of magnitude larger than allowed from diffusion through the 'cold trap' (see above). It is thought that convective penetration conveys methane in the form of ice crystals from the condensation region in the troposphere (~1.5 bar) upward to the lower stratosphere (~0.05 bar) where they can sublimate and contribute to enhancing the methane mixing ratio (figure 2). In ascending plumes (several km in diameter) uplifted above the cloud base or supersaturated in methane, buoyant moist convection could transport a significant fraction of the condensed methane at least up to the tropopause. Eddy diffusion may further raise the small CH_4 particles into the sublimation region. The delivery of methane to the stratosphere must be of the order of 1000 kg s^{-1} to balance photochemical losses.

Neptune is a surprisingly active planet considering the weakness of the solar energy available to power the wind system and meteorology. It is possible that the low level of insolation actually reduces the source of small-scale turbulence and thus of dissipation. The low frictional losses then allow strong zonal winds to build up and the large dark vortices to persist.

Interior and formation

At pressures higher than 100 kbar, beneath the hydrogen-rich envelope, lies the denser interior of Neptune. Most plausibly, it is mainly composed of the abundant 'ices' H_2O, CH_4 and possibly NH_3. Although these molecules are *not* in the solid state in Neptune's interior, they are referred to as 'ices' because they were in the form of ice in the outer solar nebula, unlike hydrogen and helium. The interior probably includes a significant fraction of 'rocks' (~25% by mass), mostly silicates and iron, and also some hydrogen and helium. The temperature increases inwards from 2000 K at 0.1 Mbar to 8000 K at the center, where the pressure is ~8 Mbar. There is no evidence for a differentiated rocky core at the center although a small one comprising at most 1 Earth mass is possible.

Unlike Jupiter and Saturn, most of the mass in Neptune (80–90%) is in the form of 'ice' and 'rock' rather than hydrogen and helium. The reason lies in the way the giant planets formed. In the outskirts of the solar nebula, dynamical times were longer and the surface density of grains was smaller than closer to the Sun. Neptune thus took longer to accrete a massive core than did the other giant planets, and did not have the time to trap gravitationally a large envelope of gaseous hydrogen and helium before the nebula was dispersed (see SOLAR SYSTEM: FORMATION). It might even have been difficult to pull out the observed amount of hydrogen and helium directly from the nebula. These gases might then have been initially trapped in large ice-rich planetesimals and incorporated into Neptune in the final stages of planetary growth, after dispersal of the nebula.

The enrichment of Neptune's outer atmosphere in deuterium (D) and heavy elements (C, S…) is a signature of the pre-eminence of ices in the planet. It may originate from icy D-rich planetesimals accreted during the last stages of planetary formation, slowed down and dissolved in the outer envelope before reaching the interior. It is also possible that the enrichment results from subsequent mixing between the ice-rich interior and the outer hydrogen–helium atmosphere.

Neptune's intrinsic heat flow amounts to 0.4 W m^{-2}, that is 1.6 times the solar flux absorbed. Transport of heat from the hot interior by convection and radiation is responsible for this energy source and for cooling of the planet since its formation. This behavior contrasts with Uranus, for which only an upper limit of 0.09 W m^{-2} was derived, indicating that the deep-seated heat is no longer transported outwards. Part of the difference in heat flows is due to the fact that Uranus is closer to the Sun so that escape of internal heat is hampered. Compositional stratification, inhibiting convective transport, might also involve a larger fraction of the interior in Uranus than in Neptune.

The magnetic field of Neptune (see below) is probably due to convective motions deep in the interior. At the temperatures and pressures relevant to Neptune's interior, a fluid of ice composition is partly ionized and dissociated and presents some electrical conductivity. Considering the velocities expected for convective motions, a dynamo can be sustained and magnetic field generated. It is possible that the dynamo-generating region is limited to a relatively thin shell (20% of the radius) where the conductivity is high enough, whilst convection is not inhibited by stable stratification.

Magnetic field and plasma

Neptune has a substantial magnetic field with a magnetic dipole tilted from the axis of rotation by 47° and offset 0.55 Neptune radii (R_N) from the planet's center (figure 5). These characteristics are rather similar to those at Uranus. The dipole representation of the field is valid typically at distances of 3 to 20 R_N. Higher-order multipole moments are very significant and affect the magnetic field at shorter distances. The field magnitude at 1 R_N varies on the planet between 0.1 and 0.8 G; it is smaller in the northern hemisphere than in the southern.

A consequence of the large tilt of the magnetic dipole moment is that the internal configuration of the magnetosphere changes rapidly as the planet rotates in the solar wind. The interaction of the solar wind with Neptune's magnetosphere produces a bow shock 35 R_N ahead of the planet. The solar wind plasma is slowed to subsonic speeds and traverses the turbulent magnetosheath as it flows around the magnetopause. The latter, located at 26 R_N sunward from the planet, marks the outer boundary for the pre-eminence of Neptune's own magnetic field. This transition region is a rotational discontinuity across

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

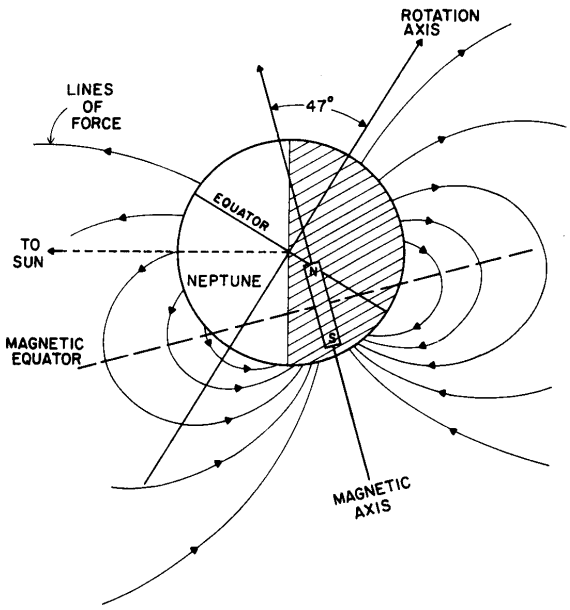


Figure 5. Configuration of Neptune's magnetic field lines. The magnetic field can be represented to first order by a dipole tilted by 47° from the axis of rotation and offset 0.55 Neptune radii. (This figure is adapted from Ness N F, Acuña M H, Burlaga L F, Connerney J E P, Lepping R P and Neubauer F M 1989 Magnetic fields at Neptune *Science* **246** 1473–8.)

which the field direction changes by ~45°, allowing some plasma to flow from the magnetosheath to the magnetosphere. The magnetotail extends to great distances in the anti-sunward direction.

Triton is the main source of magnetospheric plasma. It is created as neutral atoms escape Triton's thin atmosphere and are ionized in the magnetosphere. Neptune's atmosphere may also be a source of neutral H feeding the magnetospheric plasma. The dominant ions are N^+ and H^+. A maximum density of 2 ions/cm^3 was found by the Voyager spacecraft near closest approach to Neptune. This is the lowest plasma number density of all four giant planets. The N^+ ion temperature is about 100 eV (10^6 K) outside 5 R_N and 700 eV (8×10^6 K) near Neptune. The protons H^+ have about the same thermal speed as the N^+ ions and, being less massive, much smaller temperatures.

Plasma is removed by transport mechanisms at the same rate it is produced from the Triton source. This transport is rapid and may occur through an organized convection system and/or by diffusion resulting from stochastic changes in the electric field. Such changes are induced by changes in the atmospheric winds which occur in the planet's ionosphere. Plasma is eventually lost near the planet. Possible loss mechanisms are charge exchange within an extended neutral cloud around Neptune, ion precipitation into Neptune's atmosphere, and absorption by rings as the charged particles move planetward.

Weak auroral emissions have been detected by Voyager on the night side of Neptune: a diffuse emission spanning a wide range of latitudes and a more localized emitting region near the south magnetic pole. It has been suggested that the diffuse emission originates from photoelectrons produced on the day side by the solar flux and transported to the night side along the magnetic field lines. Another possibility is the direct precipitation of trapped particles drifting around the planet into a region of anomalously low magnetic field. This region is very extended due to the particular geometry of Neptune's magnetic field. The second emission is more consistent with usual auroral morphology and is probably associated with precipitation of trapped particles near the south magnetic pole. An aurora is also expected to occur near the north pole. The total power radiated by these auroral processes is $\approx5\times10^7$ W. The much larger magnetospheric input power required to produce the auroras may derive from Neptune's rotational energy or from the solar wind through relatively complex mechanisms.

Weak radio emissions have been detected on Neptune, strikingly similar to those seen on Uranus. They appear strongly controlled by the rotation of the magnetic field and hardly influenced by solar wind variations. Their analysis provides the most accurate measurement of Neptune's interior rotation period (16 h 06 m 30 s). The mechanisms generating these emissions are not well understood. They may originate from regions of low plasma density ('cavities'), or from continuous or impulsive injection of energetic electrons in the auroral regions.

Web update (30 April 2003)

Three previously unknown moons of Neptune have been discovered, boosting the number of known satellites of the gas giant to eleven. These moons are the first to be discovered orbiting Neptune since the Voyager II flyby in 1989, and the first discovered from a ground-based telescope since 1949. The new satellites were a challenge to detect because they are only about 30-40 km (18-24 mi) in size. Their small size and distance from the Sun prevent the satellites from shining any brighter than 25th magnitude, about 100 million times fainter than can be seen with the unaided eye. To locate these new moons, an innovative technique was used with the 4.0-meter Blanco telescope at the Cerro Tololo Inter-American Observatory, Chile, and the 3.6-meter Canada-France-Hawaii Telescope, Hawaii. Multiple exposures of the sky surrounding the planet Neptune were taken. After digitally tracking the motion of the planet as it moved across the sky, many frames were added together to boost the signal of any faint objects. Since the planet's motion was tracked, stars showed up in the final combined image as streaks of light, while the moons accompanying the planet appeared as points of light. It now appears that each giant planet's irregular satellite population is the result of an ancient collision between a former moon and a passing comet or asteroid. These collisional encounters

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

result in the ejection of parts of the original parent moon and the production of families of satellites.

Bibliography

Cruikshank D P (ed) and 78 collaborating authors 1995 *Neptune and Triton* (Tucson, AZ: University of Arizona Press)

Lunine J I 1993 The atmospheres of Uranus and Neptune *Annu. Rev. Astron. Astrophys.* **31** 217–63

Bruno Bézard

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Neptune: Minor Satellites

All but one of Neptune's minor satellites orbit within or just outside its ring system; the exception is the distant object Nereid. Some of them are better described as 'mid-sized' rather than 'minor', but are included under this heading as little is known of them. The inner four, with approximate diameters, are Naiad (60 km), Thalassa (80 km), Despina (150 km) and Galatea (160 km). The first three lie within the Le Verrier ring, to which Despina appears to act as an inner shepherd. All four were discovered during the Voyager 2 flyby in 1989. Diameters of the other satellites are rather more certain. The next, Larissa (195 km), was discovered in 1981 by Harold Reitsema when it occulted a star; it acts as an inner shepherd to the Adams ring. Proteus, orbiting outside the ring system at 118 000 km, is Neptune's second-largest satellite, after Triton. It is a rather blocky object, measuring about 435 by 400 km, and is a dark, heavily cratered body; the largest crater, named Pharos, is some 250 km across. It is another Voyager discovery. The outermost satellite, Nereid, orbiting at an average distance of 5.5 million km, has the most eccentric orbit of any known planetary satellite, at 0.753. It was discovered in 1948 by Gerard Kuiper, and has a diameter of 340 km. It may be a captured object, or it may have been perturbed into this orbit when Triton was captured (if indeed it was).

Web update (30 April 2003)
Three previously unknown moons of Neptune have been discovered, boosting the number of known satellites of the gas giant to eleven. These moons are the first to be discovered orbiting Neptune since the Voyager II flyby in 1989, and the first discovered from a ground-based telescope since 1949. The new satellites were a challenge to detect because they are only about 30–40 km (18–24 mi) in size. Their small size and distance from the Sun prevent the satellites from shining any brighter than 25th magnitude, about 100 million times fainter than can be seen with the unaided eye. To locate these new moons, an innovative technique was used with the 4.0 m Blanco telescope at the Cerro Tololo Inter-American Observatory, Chile, and the 3.6 m Canada–France–Hawaii Telescope, Hawaii. Multiple exposures of the sky surrounding the planet Neptune were taken. After digitally tracking the motion of the planet as it moved across the sky, many frames were added together to boost the signal of any faint objects. Since the planet's motion was tracked, stars showed up in the final combined image as streaks of light, while the moons accompanying the planet appeared as points of light. It now appears that each giant planet's irregular satellite population is the result of an ancient collision between a former moon and a passing comet or asteroid. These collisional encounters result in the ejection of parts of the original parent moon and the production of families of satellites.

See also: NEPTUNE: RINGS, TRITON.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Neptune: Rings

NEPTUNE, the outermost of the giant planets, has held out against ring observers the longest. This planet, 25 000 km in radius, orbits at about four and a half billion km from the Sun. Its apparent diameter on the celestial sphere is a little larger than two arcseconds, which makes it difficult to detect by ground-based telescopes. Unknown until 25 years ago, the rings of Neptune present key dynamical problems not currently understood. Indeed, the presence of puzzling arcs among the rings of Neptune place this ring system in a class by itself. In this article, we review the observations of Neptune's rings attempting, as much as possible, to provide a description sufficiently broad that it may serve to elucidate the possible links between rings and satellites.

Discovery

Stellar OCCULTATIONS provide a powerful tool with which to detect and study the material surrounding a planet. An occultation observation consists of the measurement of the signal of a star as it approaches and then disappears behind the planet during its apparent motion on the celestial sphere. The variations in the strength of the signal as the star nears occultation provide information on the presence or absence of material around the planet. In July 1984, observations at three telescopes in the European Southern Observatory and the Cerro Tololo Inter-American Observatory recorded the near-simultaneous occultation of a bright star by a semi-transparent object 15 km in width and at least 100 km long. The dip was observed on one side indicating that Neptune had a ring, but also that it was incomplete! This discovery led NASA to reprogram observations of Neptune's environment by the VOYAGER spacecraft, both to learn more about its ring system and also to avoid collisions with its circumplanetary material. Voyager 2 encountered Neptune in 1989 and its instruments succesfully collected data on Neptune's rings and satellites. As the initial detection suggested, Neptune's rings are in a class by themselves. The presence of ring-arcs has been particularly puzzling. Such arcs would not remain azimuthally limited for very long, unless actively confined, but would spread into a full ring. Many models have been proposed to account for the occurrence of ring arcs. Before reviewing these models, let us describe the surprising and varied morphology of Neptune's rings.

Morphology

Neptune's ring system comprises four dusty narrow rings, as well as low-optical-depth dust extended over broad regions. Because of the large fractional abundance of fine, micron-sized 'dust', the rings are best imaged at moderate to high phase angles (the angle formed by the line from the Sun to the ring and the line from the ring to the observer) such that the presence of fine material is brought out by DIFFRACTION.

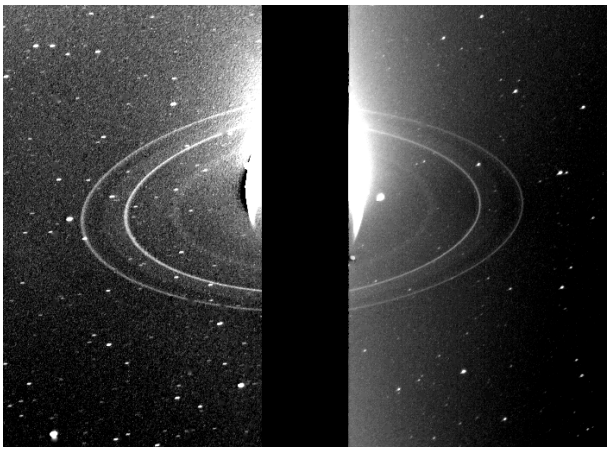


Figure 1. Rings of Neptune. Mosaic of two images of Neptune's rings taken by Voyager spacecraft in 1989. The two bright rings (the inner le Verrier and the outer Adams) are clearly visible. During this period, the bright ring arcs in the Adams ring are not visible. Other discrete structures are also visible like the innermost ring Galle and the broad ring extending from the le Verrier ring towards Adams.

The Adams ring, which houses the ring arcs, is the outermost ring at a radius of $62\,932.57 \pm 0.02$ km from Neptune. Also clearly visible in figure 1, is the narrow le Verrier ring at $53\,200 \pm 20$ km. In between the two outer bright rings lies Lassell, a broad ring, $\sim$ 4000 km, extending from the le Verrier ring and out to the narrow ring Arago. Further in, at an orbital radius of 42 000 km, lies the broad ring ($\sim$ 2000 km) Galle. Finally, a very faint and narrow feature, not visible in figure 1, appears to orbit at the same distance as the satellite Galatea, $\sim$ 980 km interior to the Adam's ring. It is at present unknown whether or not a sheet of low optical depth extends from the innermost ring to the planet.

The arcs in the Adams ring are the most striking example of longitudinal structure in Neptune's rings, and perhaps of any PLANETARY RING (see also JUPITER: RINGS, SATURN: RINGS, URANUS: RINGS). In figure 2 the three arcs named Liberté, Egalité and Fraternité can be seen. They span a region of $\sim$ 40 degrees in longitude. There are, in fact, at least two more arcs, and Egalité itself is composed of two distinct arcs. These features are believed to be clumps of material denser than the average ring particle density found in the Adams ring as a whole. The radially integrated brightness of the Adams ring is observed to vary by about a factor of three, from a maximum near and between the arcs to a minimum far from the arc region. The arcs themselves exhibit clumping and structure on scales ≤ 10 degrees, perhaps resulting from the accumulation of dust-sized particles around bigger particles which might provide the source of the dust through collisions.

Voyager's passage through the ring plane was used to place an upper limit of ≤ 30 km on the vertical extent of the le Verrier and Adams rings. The limit of the vertical extent for one of the arc regions is $\leq$ 110 km.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

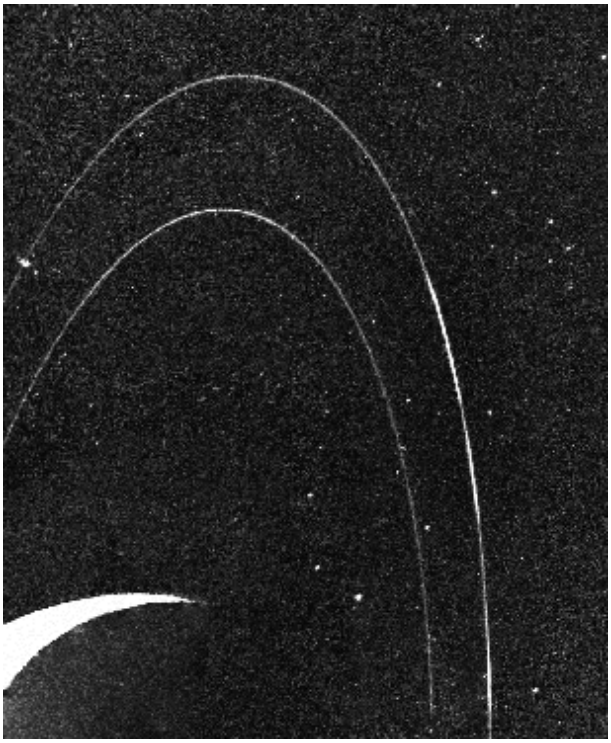


Figure 2. Neptune's arcs from Voyager. The inner ring, le Verrier, is uniform, but the outer ring, Adams, contains three main arcs: from the bottom to the top, Liberté, Egalité and Fraternité.

Further information about the vertical extent from Voyager measurements shows that dust is present throughout the Neptune system, even over Neptune's north pole.

From Voyager photometry the optical depth for the arcs (which measures the degree of extinction of light passing through them) is of the order of 0.1, consistent with occultation observations. For the narrow continuous rings the optical depths depend on the ring widths used to calculate them. The optical depth for the Adams ring is ~ 0.004, and for the le Verrier ring it is ~ 0.002. Both the Lassell and Galle rings are consistent with an optical depth of the order of ~ 0.0001. Uncertainties in the particle size distribution make it difficult to calculate dust fractions for any ring. The problem is exacerbated in the case of the Lassell and Galle rings by the large uncertainties in the low-phase-angle measurements of their scattering properties. Adopting large particle albedos and particle size distribution, the fraction of dust in the arcs, as well as the Adams and le Verrier rings, appears to be comparable with that observed for Saturn's F ring.

The physical and chemical nature of the ring particles is still unknown, mostly because the Voyager spacecraft did not carry a spectrometer, in particular one that operated in the infrared, which would have allowed detection of the absorption band of ices and silicates. However, the study of light diffused by the rings points to a composition of 'dirty ice', perhaps with silicates or some carbon bearing material.

Arc confinement

As soon as Neptune's arcs were discovered, efforts were made to try to understand how they could be azimuthally confined. It was soon realized that unless actively 'herded' the arcs would spread into a complete ring in a matter of years. This is due to the differential rate of rotation around Neptune between those particles located in the inner part of the arc, and those located in the outer part; the latter, being more distant from Neptune, have slightly lower orbital velocities. While one may think that particle self-gravity could counteract this spread, the particles are too small for their gravity to play a significant role.

One possible solution to this conundrum involves satellite confinement in the neighborhood of specific equilibrium locations named Lagrange points. A close analog to such a configuration exists elsewhere in the solar system: a family of asteroids, called the TROJAN ASTEROIDS, are kept in orbit around Jupiter's Lagrange points; however, there is the important distinction that rings are a collisional system. Since Lagrange points correspond to local energy maxima, energy dissipation should lead to the eventual decay away from these locations. As a result, for this model to work it would require the presence of yet another (undiscovered) satellite to counteract the dissipation of energy due to inelastic collisions between ring particles.

An alternative model uses two distinct resonances by one single satellite simultaneously at work to produce ring arcs. For a resonance to take place there needs to be a matching of frequencies; in this context, one possibility is for the orbital period of the ring particles and the satellite to be related by a ratio of integers (i.e. a given number of revolutions by the ring particles corresponds to an integer number of revolutions by the satellite). But for objects orbiting an oblate (slightly non-spherical) planet there are other possibilities; in such a case, the frequency with which they move in and out of the orbital plane (if inclined), and in and out of their orbital radius (if eccentric), is distinct from the frequency with which they revolve around the planet. This gives rise to a panoply of possible resonance conditions which can result in complicated dynamical phenomena. A model that uses two different kinds of resonances to explain ring arcs received some support when analysis of Voyager data found that a 42:43 eccentricity resonance with the satellite Galatea could be controlling the ring's radial excursions. At the same time, its 42:43 inclination resonance (Galatea orbits at a slight inclination inside the Adams ring) could prevent the arc from spreading azimuthally. This model places very stringent conditions on the relative radial locations of the ring and the satellite resonance. While it is clear that Galatea plays a role in shaping the Adams ring, it remains uncertain whether its resonances are directly responsible for arc confinement. In the first place, the number and location of the arcs does not appear to correspond to what one would expect on the basis of the resonance model.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Also, the first ground-based images of the arcs taken in 1999 have improved the orbital parameters for this ring system, and have shown that the mean semi-major axis of the arcs is located outside the region defined by the corotation inclination resonance with Galatea. If true, this calls into question the role of Galatea in confining the arcs.

It is possible that, taken together, the observations indicate that the presence of larger librating particles, which provide the source of (at most) partially confined dust, controls the locations of the ring arcs. Then dust escaping from the arcs would populate first the Adams ring (which might explain the variable opacity for this ring), and possibly other inner rings as well. This sourcing mechanism also appears to be a good candidate to explain the presence of dust at the same orbital radius occupied by Galatea. But this, as well as other models, will have to be tested by further observation and theory if we are to understand this challenging ring system.

Conclusion

Ever since rings were first discovered, and later identified as disks of material nowhere touching the planet, astronomers have grown accustomed to celestial objects which come in two basic shapes: spheres and disks. The prospect of coming to grips with a system distinct from these two groups, in that it breaks the azimuthal symmetry we have come to expect, is too tantalizing for astronomers to resist. Whether these structures have more to tell us with regard to the formation of satellites, the longevity of ring systems in general, or the nature and composition of solar system objects, it is clear that the sheer challenge of Neptune's ring system will keep astronomers coming back for more for a long time to come.

Bibliography

Porco C *et al* 1998 Neptune's ring system *Neptune and Triton* ed D Cruikshank (University of Arizona Press) pp 703–804

Ignacio Mosqueira and François Poulet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neutral Point

A point in space at which an object (strictly, a point mass—a theoretical particle with mass but no volume) would experience no net gravitational force. The simplest circumstance in which such a point could exist would be between two motionless masses, and it would lie on a line joining their centers of mass, closer to the larger mass. In the universe all masses are in motion relative to one another; in these circumstances neutral points do exist, but calculating their positions for the combined gravitational fields of a number of bodies is less straightforward. For the simplest case of two massive bodies in circular orbits around their center of mass, there are five neutral points, known as the Lagrangian points. These points remain fixed with respect to the two bodies.

See: Lagrangian points.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neutrino Astronomy

The Sun generates energy deep in its core by nuclear fusion processes. A SUPERNOVA (of type II) explodes and blows off the entire star by gravitational energy liberated when the central core collapses. These interesting phenomena of energy generation can never be observed with visible light or any other electromagnetic waves. The places where these phenomena take place are buried deep in dense material and the created heat is quickly thermalized before it comes out of the stellar surface. Relevant information is lost during thermal equilibration.

Nevertheless, one may want to observationally probe the deep interior of the Sun or the gravitational collapse in the supernova core. One then has to find electrically neutral particles that have guaranteed straight travel paths in space. They obviously have to be produced in the regions and should not participate in thermal equilibration in order to bring the desired information. NEUTRINOS are exactly the particles that possess these features. Neutrinos are produced in nuclear processes, but interact with matter so weakly that as soon as they are created they leave the region with very nearly the speed of light as their mass is very small.

The universe about 1 s after the big bang had a temperature as high as the core of a supernova, and light particles such as electrons and positrons were in thermal equilibrium. Cosmic expansion in the meantime cooled down the universe and lowered the energy density. Electrons and positrons then annihilated, disappeared and created photons and neutrinos. These relic neutrinos should exist in the present universe and have important information on the early universe.

COSMIC RAYS are accelerated somewhere in our Galaxy or, if they are very energetic, perhaps in distant active galaxies, too. The extremely energetic cosmic rays (above 10^{20} eV) might even be produced by the decay of exotic objects created in the early universe and floating around even at the present, whose existence is conjectured in grand unified theories of elementary particles. Cosmic rays bend in galactic or intergalactic magnetic fields and cannot be traced back to their birthplaces. Even to this day, we do not know for sure where cosmic rays are actually born. Cosmic rays may interact with the medium surrounding the acceleration region and produce mesons, which decay in a short time and produce neutrinos. Neutrinos then escape freely to open space and may reach Earth with fruitful information.

Neutrino astronomy is used to study the central engines of the Sun and supernova, to identify the origin of cosmic rays and to study the very early universe. Astronomical neutrinos can also be used as beams useful for investigating neutrino properties, such as mass and mixing. Indeed, only after having completed the study of neutrino properties can one begin real neutrino astronomy.

Neutrinos from the Sun were first observed by R Davis Jr in the Homestake gold mine in the late 1960s. The experiment is still going on 30 yr after the initial operation. Since then four other experiments have joined this field and two more are coming into operation in the near future. All five experiments did observe solar neutrinos, thus confirming that the Sun is actually generating energy by nuclear fusion processes (see SOLAR INTERIOR: NEUTRINOS). The reason why this field is still so active even now is that there are serious discrepancies between observations and expectations. What the experiments observed is only 30–60% of what the standard solar model predicts. It would be shocking if the Sun were not working as theory tells us, because all life on Earth depends on stable solar heat generation. The so-called solar neutrino problem has not been completely solved. The most popular idea is not that the solar engine works with an efficiency of only 30–60% but that the 'deficit' of solar neutrinos is caused by neutrino oscillations, namely electron neutrinos that are produced in the Sun are converted to a different kind of neutrino, probably muon neutrinos, either in the solar interior or on the way to Earth. Muon neutrinos are impossible or difficult to observe, hence leading to an apparent deficit of solar neutrinos. This idea will be tested by new-generation experiments in the near future.

23 February 1987 will be remembered as the day when a real instance of the gravitational collapse and the subsequent birth of a NEUTRON STAR were first observed with neutrinos. Two large water Cherenkov detectors, Kamiokande and IMB, observed 11 and eight neutrino events, respectively. Although the observed neutrinos were few, these observations beautifully demonstrated that the underlying theory of the explosion of a type II supernova is correct, namely that the explosion was initiated by the gravitational collapse and the total energy released was about 3×10^{53} erg, which roughly corresponds to 500×solar luminosity×4.6 billion years.

These are the only successful observations of astronomical neutrinos. Large under-ice and under-water detectors are being built and tested for high-energy cosmic ray neutrinos with energies above 10^{12} eV. It will still take several more years before these neutrinos are first observed.

There is absolutely no idea how to detect relic neutrinos born in the early universe. They are expected to have an average energy of only 0.000 17 eV which is too low to have a meaningful interaction probability (cross section) with matter.

What neutrinos are

Particle physics tells us that the ELEMENTARY PARTICLES (or simply called particles) are classified in four types: up-quark (u), down-quark (d), electron (e) and electron neutrino (ν_e). These particles are said to form a family. Three families have in fact been found in nature and hence there are 12 elementary particles altogether. They are listed in table 1. Three neutrino species are called the electron neutrino (ν_e), muon neutrino (ν_μ), and tau neutrino (ν_τ), corresponding to charged counterparts, electron (e), muon (μ) and tau lepton (τ), respectively.

The unique characteristics of neutrinos are as follows.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Table of elementary particles. Rows in the second column correspond to u-quark, d-quark, electron and electron neutrino. They are classified by electric charge. The columns correspond to families. Families are distinguished only by the masses of the particles.

Charge			
2/3	u	c	t
−1/3	d	s	b
−1	e	μ	τ
0	ν_e	ν_μ	ν_τ

- They are electrically neutral and fly straight even in the presence of magnetic fields.
- Only weak forces act on neutrinos and hence they have a great ability to penetrate matter.
- Neutrinos are produced in high-temperature and high-density environments such as the center of stars and easily escape out of the star and reach Earth, bringing information on nuclear processes, energy generation etc at the production site.
- Great penetrability in turn makes neutrino detection quite difficult and the target size of a detector should be as large as possible.

Solar neutrinos

According to the standard solar model the central temperature and density of the Sun are about 15 million kelvins ($kT \simeq 14$ keV) and 150 g cm^{-3}, respectively. These conditions allow protons to overcome the repelling force from their positive charges (Coulomb barrier) and to initiate nuclear reactions.

The following nuclear fusion processes are taking place in the Sun:

$$\mathrm{p} + \mathrm{p} \longrightarrow \mathrm{d} + \mathrm{e}^+ + \nu_\mathrm{e} \quad (1)$$

$$\mathrm{d} + \mathrm{p} \longrightarrow {}^3\mathrm{He} + \gamma \quad (2)$$

$${}^3\mathrm{He} + {}^3\mathrm{He} \longrightarrow {}^4\mathrm{He} + 2\mathrm{p} \quad (3)$$

or

$${}^3\mathrm{He} + {}^4\mathrm{He} \longrightarrow {}^7\mathrm{Be} + \gamma \quad (4)$$

$${}^7\mathrm{Be} + \mathrm{e}^- \longrightarrow {}^7\mathrm{Li} + \nu_\mathrm{e} \quad (5)$$

$${}^7\mathrm{Be} + \mathrm{p} \longrightarrow {}^8\mathrm{B} + \gamma \quad (6)$$

$${}^8\mathrm{B} \longrightarrow {}^8\mathrm{Be}^* + \mathrm{e}^+ + \nu_\mathrm{e}. \quad (7)$$

The product ^{7}Li further reacts with a proton and produces two ^{4}He. The product ^{8}Be* immediately decays into two ^{4}He also. The reaction chain is now complete. One can see that electron neutrinos are produced by weak interactions (equation (1) and equation (5)) and a weak decay (equation (7)). These neutrinos are called pp neutrinos, ^{7}Be neutrinos and ^{8}B neutrinos, respectively. pp and ^{8}B neutrinos have continuous energy spectra approximately given as $E_\nu^2 E_\mathrm{e} p_\mathrm{e}$, where E_ν, E_e and p_e are neutrino energy, positron energy and positron momentum, respectively. E_ν is given by $Q - E_\mathrm{e}$, where Q is the available energy for the neutrino+positron pair (Q value); the maximum E_ν are 0.42 MeV and 15 MeV for pp and ^{8}B neutrinos, respectively. ^{7}Be neutrinos have a line spectrum as equation (5) is a two-body reaction. Their energy is either 0.86 MeV (90%) or 0.38 MeV (10%).

The standard solar model calculates the neutrino flux on Earth from each source. The fluxes obviously depend on the temperature, density of each chemical element, thermal conduction of the surrounding medium, etc. The calculated fluxes are enormous: $5.9 \pm 0.06 \times 10^{10}$ cm^{-2} s^{-1}, $0.48 \pm 0.04 \times 10^{10}$ cm^{-2} s^{-1} and $5.2^{+1.0}_{-0.7} \times 10^6$ cm^{-2} s^{-1} for pp, ^{7}Be and ^{8}B neutrinos, respectively.

Neutrinos have to be detected on Earth. In general particle detectors utilize particles' electric charge to produce observable signals. Neutral particles such as neutrinos must first interact with target particles and produce charged particles that produce signals. For solar neutrinos the following reactions have been used:

$$\nu_\mathrm{e} + \mathrm{e}^- \longrightarrow \mathrm{e}^- + \nu_\mathrm{e} \quad (8)$$

$$\nu_\mathrm{e} + {}^{37}\mathrm{Cl} \longrightarrow \mathrm{e}^- + {}^{37}\mathrm{Ar} \quad (9)$$

$$\nu_\mathrm{e} + {}^{71}\mathrm{Ga} \longrightarrow \mathrm{e}^- + {}^{71}\mathrm{Ge}. \quad (10)$$

In water Cherenkov experiments, Kamiokande and Super-Kamiokande (figure 1), water is the target material and electrons emitted from reaction (8) are detected. Only electrons with $E_\mathrm{e} \geq 5.5$ MeV are detected owing to severe background problems. Thus the experiment is sensitive only to ^{8}B neutrinos. For the latter two reactions, the end products ^{37}Ar and ^{71}Ge are extracted by the radiochemical technique and their numbers counted. The pioneering experiment Homestake used reaction (9) with a threshold energy of 0.81 MeV, and was thus sensitive to ^{7}Be and ^{8}B neutrinos, and subsequent SAGE and GALLEX experiments adopted reaction (10) with a threshold energy of 0.23 MeV, sensitive to pp, ^{7}Be and ^{8}B neutrinos.

The observational results are surprising. All five experiments observed much fewer neutrinos than what the standard solar model predicts. If one takes the ratio of observed to predicted numbers, the results were 0.27 ± 0.02, 0.42 ± 0.06, 0.47 ± 0.015, 0.56 ± 0.06, 0.50 ± 0.09 for Homestake, Kamiokande, Super-Kamiokande, GALLEX and SAGE, respectively.

These serious discrepancies are called the solar neutrino problem and its solution has not been found yet, although neutrino oscillation is the most likely clue. Super-Kamiokande and coming experiments, SNO and BOREXINO, will challenge the problem further.

Supernova neutrinos

The theory of the type II supernova explosion can be summarized as follows.

A star with mass larger than $8M_\odot$ (solar mass) rapidly burns out its nuclear fuels. Nuclear ashes form an onionskin-like structure: the iron core at the center, then successively silicon, oxygen, neon, carbon, helium layers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

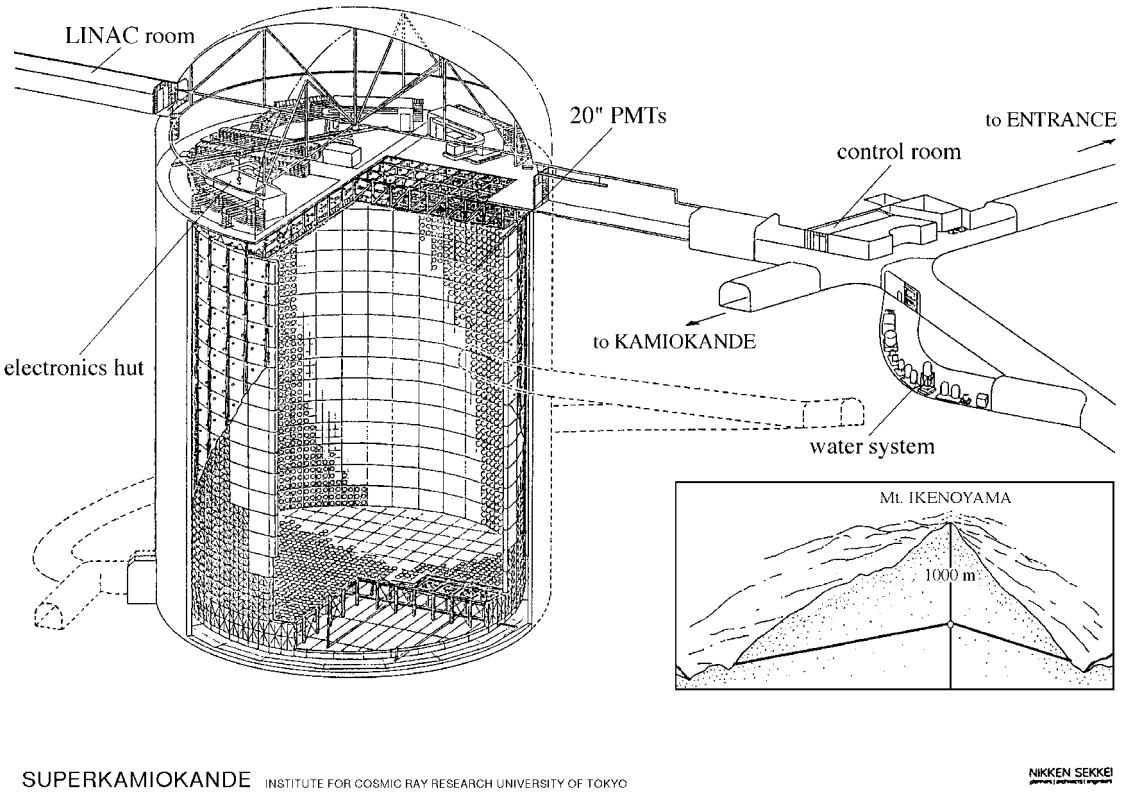


Figure 1. The Super-Kamiokande detector. It consists of inner and outer parts, each of which contains 32 000 tons and 18 000 tons of pure water, surrounded by 11 200 50 cm diameter photomultiplier tubes (PMTs) and 1800 20 cm diameter PMTs, respectively. A charged particle moving faster than the speed of light in water emits blue light named after the discoverer P Cherenkov with an opening angle of about 42°. The energy, direction and type of the particle are determined by measuring the arrival time, intensity and ring pattern of the Cherenkov light. The Super-Kamiokande experiment discovered the finite mass of neutrinos from detailed observation of atmospheric neutrinos.

and the hydrogen layer at the outermost layer. The central iron core becomes unstable when its mass exceeds the Chandrasekhar mass ($\sim 1.4 M_\odot$) and eventually collapses gravitationally. The central core is pressurized by falling matter, stiffened and bounces back. How large the bounce is depends on the elasticity or equation of state of the matter. The temperature rises above 100 billion kelvins ($kT \sim 10$ MeV) and the inverse beta decay process sets in:

$$\mathrm{p} + \mathrm{e}^- \longrightarrow \mathrm{n} + \nu_\mathrm{e}. \tag{11}$$

An electron neutrino burst of about 10 ms duration thus follows. It is called a neutronization burst and signals the real onset of the supernova explosion. A neutron-rich high-density ball forms which is called a proto-neutron star as it will later become a neutron star. The matter density is now above 10^{11} g cm^{-3}, which is high enough that even neutrinos frequently interact with matter and are trapped inside. Neutrinos are in approximate thermal equilibrium with electrons, protons and neutrons. Neutrinos of all kinds now evaporate from the surface of the neutron-rich, high-density sphere called a neutrinosphere, the radius of which was initially several tens of km. The underlying process for neutrino emission is the thermal one:

$$\mathrm{e}^+ + \mathrm{e}^- \longrightarrow \nu_i + \bar{\nu}_i \tag{12}$$

where ν_i denotes any of the three neutrino species and $\bar{\nu}_i$ its antineutrino. The hot proto-neutron star quickly cools down with a time scale of about 5 s, where neutrino and antineutrino pairs carry off thermal energy. Note that the proto-neutron star is still surrounded by thick falling matter. Hence heat cannot be transported by electromagnetic waves. The proto-neutron star shrinks to a radius of about 10 km and the density reaches $\sim 10^{14}$ g cm^{-3}, at which neutrons are packed in contact with each other. It has become a neutron star. The reverse shock produced at the core bounce, and probably the reheat caused by the outgoing neutrinos, later blow off the whole star and the explosion becomes optically visible, thus the birth of a supernova.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The total energy carried off by neutrinos, the luminosity, is equal to the gravitational potential of a neutron star with radius $R \sim 10$ km and mass $M \sim 1.4 M_\odot$:

$$L_\nu = \frac{3}{5}\frac{GM^2}{R} \simeq 3 \times 10^{53} \text{ erg}. \quad (13)$$

The neutrino energy distribution approximately follows the Fermi–Dirac distribution,

$$\frac{\mathrm{d}N_\nu}{\mathrm{d}E} = 4\pi R^2 \frac{g\pi c}{(2\pi\hbar c)^3} \frac{E^2}{\exp[(E-\mu)/kT]+1} \frac{1}{4\pi D^2} \quad (14)$$

where E, T and R are the neutrino energy, temperature and radius of the neutrinosphere, respectively. $g(=1)$ is the helicity weight of the neutrino, μ the chemical potential and D the distance to the supernova. The average neutrino energy is, for zero chemical potential, $\langle E \rangle \simeq 3.15kT$.

The cross sections satisfy the following inequalities: $\sigma_{\nu_\mathrm{e}} > \sigma_{\bar{\nu}_\mathrm{e}} > \sigma_{\nu_\mu} = \sigma_{\nu_\tau} > \sigma_{\bar{\nu}_\mu} = \sigma_{\bar{\nu}_\tau}$. Neutrinos with a smaller cross section can leave the proto-neutron star at the denser and deeper region where the temperature is higher. The average energy of those neutrinos is thus higher. According to calculations, $\langle E_{\nu_\mathrm{e}} \rangle \sim 10$ MeV, $\langle E_{\bar{\nu}_\mathrm{e}} \rangle \sim 15$ MeV and $\langle E_{\nu_\mu} \rangle \sim 20$ MeV. It was also found that the luminosity is equipartitioned among neutrino species, $L_{\nu_\mathrm{e}} \simeq L_{\bar{\nu}_\mathrm{e}} \simeq L_{\nu_\mu} \simeq L_{\bar{\nu}_\mu} \simeq L_{\nu_\tau} \simeq L_{\bar{\nu}_\tau}$.

Neutrinos from SUPERNOVA 1987A (SN1987A) were observed using the following reactions which took place in the detector water:

$$\bar{\nu}_\mathrm{e} + \mathrm{p} \longrightarrow \mathrm{e}^+ + \mathrm{n} \quad (15)$$

$$\nu_i + \mathrm{e}^- \longrightarrow \nu_i + \mathrm{e}^- \quad (16)$$

where ν_i is any of the neutrino species, and emitted e^+ and e^- were detected. However, the cross section of equation (15) is almost 100 times larger than that of the elastic scattering (equation (16)), and hence almost all of the observed events were due to $\bar{\nu}_e$. There was no clear evidence that elastic-scattering events (equation (16)) had been detected.

Assume that the radius of the neutrinosphere R is constant in time while its temperature T behaves like $T = T_0 \exp(-t/4\tau)$. The factor 4 was introduced in order for the luminosity L_ν to take the form $\exp(-t/\tau)$, since the luminosity is proportional to T^4.

The combined Kamiokande and IMB data have been analyzed with an assumption that $\mu = 0$ in equation (14). The results are

$$kT_0 = 4.5^{+1.1}_{-0.9} \text{ MeV}$$

$$\tau = 4.2^{+2.5}_{-1.5} \text{ s}$$

$$R = 23^{+22}_{-10} \text{ km}$$

$$L_{\bar{\nu}_\mathrm{e}} = 5^{+8}_{-3} \times 10^{52} \text{ erg}.$$

$L_{\bar{\nu}_\mathrm{e}}$ was calculated from

$$L_{\bar{\nu}_\mathrm{e}} = \int E \frac{\mathrm{d}N_\nu}{\mathrm{d}E\,\mathrm{d}t} 4\pi D^2 \,\mathrm{d}E_\nu \,\mathrm{d}t$$

where $D = 50$ kpc was assumed and $\mathrm{d}N_\nu/\mathrm{d}E\,\mathrm{d}t$ is given by equation (14).

Note that the total neutrino luminosity L_tot is given as $L_\mathrm{tot} = 6L_{\bar{\nu}_\mathrm{e}}$, which leads to

$$L_\mathrm{tot} = 3^{+5}_{-2} \times 10^{53} \text{ erg}.$$

This is in excellent agreement with the naive expectation of equation (13). The cooling time τ and the temperature T_0 are also quite reasonable. Thus the observations of SN1987A beautifully confirmed the basic part of the supernova theory.

More data are obviously needed for further studies of the gravitational collapse. The new detector, SUPER-KAMIOKANDE, which is 15 times larger than Kamiokande, has been in operation since 1996. If the next supernova occurs at a distance of 10 kpc, i.e. close to the Galactic center, Super-Kamiokande will observe 4000 $\bar{\nu}_e$ events (by equation (15)) and about 250 elastic events (by equation (16)). Super-Kamiokande may be able to see about 10 neutronization-burst events, too. The only question is now when the next supernova will happen.

High-energy cosmic ray neutrinos

The CRAB NEBULA is known to emit gamma rays with energies larger than 10^{12} eV (TeV gamma rays). The energy spectrum was measured and found to obey a power law with a spectral index close to 2. Detailed studies together with other wavelength data revealed that TeV gamma rays are most likely produced by the inverse Compton process, namely collision of high-energy electrons ($\leq 10^{14}$ eV) with synchrotron photons or microwave background photons. Several other PULSARS have also been found as TeV gamma ray sources. Their production mechanism is considered to be the same as that for Crab Nebula. Nonthermal x-ray emission was found at the expanding shell of a supernova remnant SN1006. These x-rays are presumably synchrotron photons radiated by high-energy electrons ($\leq 10^{13}$ eV). TeV gamma rays were indeed found coming exactly from the same shell, indicating again that the inverse Compton process is responsible. TeV gamma rays are also coming from extragalactic objects, BL Lac sources Mkn421 (redshift 0.031) and Mkn501 (redshift 0.032). Their production mechanism is not clear but is presumably the inverse Compton process, too.

Hence acceleration sites of high-energy electrons have been identified. However, where are protons and other nuclei then accelerated? We know that they are dominant components of cosmic rays and moreover we know that their energies extend beyond 10^{20} eV. The acceleration site may be surrounded by intense photons or to a lesser extent a thick gas, i.e. optically thick, so that TeV gamma rays may have been absorbed before escaping the acceleration region. If this is the case, one expects the site to be a strong source of neutrinos, whose energies may extend as high as 10^{17} eV. Production processes of neutrinos are

$$\mathrm{p} + \gamma \longrightarrow \pi^+ + \mathrm{n} \quad (17)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\pi^+ \longrightarrow \mu^+ + \nu_\mu \tag{18}$$

$$\mu^+ \longrightarrow \bar{\nu}_\mu + e^+ + \nu_e \tag{19}$$

where γ is an ambient photon. Muon neutrinos and electron neutrinos are produced with a ratio of 2 to 1.

The flux of these neutrinos is estimated by assuming that observed TeV gamma rays are produced not by the inverse Compton process but by the nuclear processes, namely, corresponding to equation (17),

$$p + \gamma \longrightarrow \pi^0 + p \tag{20}$$

$$\pi^0 \longrightarrow \gamma + \gamma. \tag{21}$$

By estimating somehow the optical thickness, one can relate the gamma ray flux to that of neutrinos. These estimates are, however, quite uncertain owing to the poorly known optical thickness. The energy spectrum is expected to obey a power law with a spectral index of about 2, the same as for gamma rays. The estimated fluxes of $\nu_\mu + \bar{\nu}_\mu$, F_ν, are, for $E_\nu \geq 1$ TeV, $F_\nu \sim 10^{-12}$ cm^{-2} s^{-1} TeV^{-1} for AGN 3C279 and $F_\nu \sim 10^{-11}$ cm^{-2} s^{-1} TeV^{-1} for Mkn421 and 3C273. These numbers may be quite optimistic and could be smaller by several orders of magnitude.

Muon neutrinos from astronomical objects interact with rock in Earth and produce muons:

$$\nu_\mu + N \longrightarrow \mu^- + \cdots. \tag{22}$$

Detection of astronomical muon neutrinos adopts this reaction. A large-area muon detector is deployed deep underground in order to reduce the atmospheric muon background. Rock beneath the detector is the target. A high-energy muon produced in the rock travels a long distance upward, approximately $E_\mu/(2\text{ MeV})$ g cm^{-2}, reaches the detector and is detected. Downward-going muons are completely swamped by atmospheric muons and cannot be detected. In order to overcome the huge background of atmospheric neutrinos, which are produced by cosmic rays in the atmosphere associated with atmospheric muons, a high-energy threshold must be set, say 1 TeV–100 TeV, depending on what objects are to be observed. A rough estimate of the observed rate is

$$N_\mu \sim \frac{0.5\text{ events}(E_\mu \geq 1\text{ TeV})}{10^5\text{ m}^2\text{ yr}} \frac{F_\nu(\geq 1\text{ TeV})}{10^{-12}\text{ cm}^{-2}\text{ s}^{-1}}. \tag{23}$$

The largest detector under test is the AMANDA-B (figure 2) detector which is located 1500 m under Antarctic ice. The detector has cylindrical shape with 120 m radius and 400 m height. Cherenkov light emitted in the ice is detected with a number of photomultipliers. AMANDA-B may still be too small. There is a plan to enlarge it to a size of 1 km^3, and underwater projects of similar size are being seriously considered.

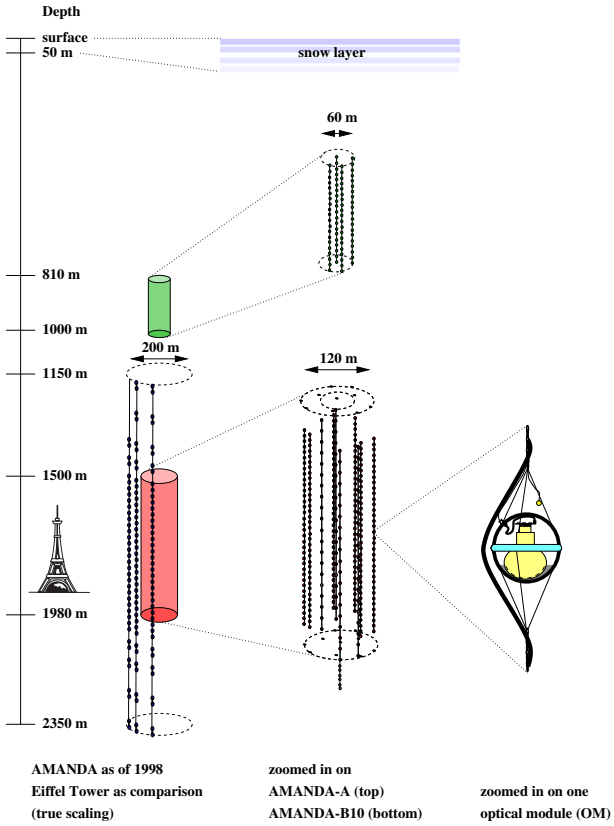


Figure 2. The AMANDA-B detector. A number of optical modules (20 cm diameter photomultipliers) are deployed within a volume 120 m in diameter and 400 m high. A high-energy upward-going muon produced by a cosmic ray muon neutrino in the rock passes through the region and emits Cherenkov light which is detected by optical modules. The direction and energy are obtained by measuring arrival time and intensity of the Cherenkov light. It could be extended to a km^3 detector.

Relic neutrinos

In the hot and dense universe up to about 1 s after the big bang, electrons and neutrinos were in thermal equilibrium. They were interacting each other,

$$e^+ + e^- \longleftrightarrow \nu_i + \bar{\nu}_i \tag{24}$$

where ν_i is a neutrino of any kind.

Surprisingly, the environment at that time was much more modest than that in the core of the supernova. The temperature was 10 billion kelvins ($kT \sim 1$ MeV), which is about 5 times lower, and the energy (or mass) density was almost 9 orders of magnitude less than those of a hot neutron star.

The universe expanded rapidly and hence cooled down (see UNIVERSE: THERMAL HISTORY). The thermal equilibrium could not hold and eventually the reaction from right to left of equation (24) ceased. Neutrinos were left behind in the universe and their energies became lower and lower until the present day. Electrons and positrons

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rapidly annihilated into two photons whose energies were also redshifted to a wavelength of a few mm. These photons are called the COSMIC MICROWAVE BACKGROUND (CMB) and detailed studies are being carried out. Relic neutrinos and photons obey Fermi–Dirac and Planck distributions, respectively. The number densities are

$$n_\nu = 1.5\frac{\zeta(3)}{\pi^2(\hbar c)^3}(kT_\nu)^3 \tag{25}$$

$$n_\gamma = 2\frac{\zeta(3)}{\pi^2(\hbar c)^3}(kT_\gamma)^3 \tag{26}$$

where $\zeta(3)/\pi^2 = 0.121\ldots$. The neutrino number density corresponds to one species only. There is a simple relation between the two temperatures,

$$\frac{T_\nu}{T_\gamma} = \left(\frac{4}{11}\right)^{1/3} = 0.714. \tag{27}$$

The CMB temperature was measured to be 2.74 K. Hence $n_\gamma = 420\ \mathrm{cm}^{-3}$ which is in good agreement with measurement. The neutrino temperature and number density per species should therefore be 1.96 K (0.000 17 eV) and $n_\nu = 115\ \mathrm{cm}^{-3}$, respectively.

These relic neutrinos must exist in the present universe. Recently the Super-Kamiokande experiment has discovered the finite neutrino mass from a detailed study of atmospheric neutrinos. However, the experiment is sensitive only to the mass-squared difference, not to each mass value. The measured value is Δm^2 ($\equiv m^2_{\nu_\tau} - m^2_{\nu_\mu}$) $= (1.5\text{–}6) \times 10^{-3}\ \mathrm{eV}^2$. From this the tau neutrino mass is constrained as $m_{\nu_\tau} \geq 0.04$ eV, which is much larger than the temperature 0.000 17 eV. Thus the relic tau neutrinos should be nonrelativistic. Their velocity is about 300 km s^{-1} which is determined from other astronomical factors such as the Earth's motion in the Galaxy.

The neutrino mass could be as large as a few eV. If this is the case, the relic neutrinos are the hot dark matter component and could significantly contribute to the energy density of the universe.

There is at present no idea how to detect such low-energy neutrinos.

Bibliography

General review:

Totsuka Y 1992 *Rep. Prog. Phys.* **55** 377

High-energy cosmic ray neutrinos:

Gaisser T K, Halzen F and Stanev T 1995 *Phys. Rep.* **258** 173

Solar neutrinos and SN1987A:

Bahcall J N 1989 *Neutrino Astrophysics* (Cambridge: Cambridge University Press)

Neutrino physics and neutrino astronomy:

Winter K (ed) 1991 *Neutrino Physics* (Cambridge: Cambridge University Press)

Recent results on neutrino physics and astrophysics:

Suzuki Y (ed) 1999 *Proc. 18th Int. Conf. on Neutrino Physics and Astrophysics* (Amsterdam: Elsevier)

Yoji Totsuka

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neutrinos

Neutrinos are electrically neutral ELEMENTARY PARTICLES which experience only the weak nuclear force and gravity. Their existence was introduced as a hypothesis by Wolfgang Pauli in 1930 to explain the apparent violation of energy conservation in radioactive beta decay. Chadwick had discovered in 1914 that the energy spectrum of electrons emitted in beta decay was not monoenergetic but continuous, from zero to the maximum energy allowed by the energy balance. Pauli's hypothesis assumed that a neutral particle was emitted together with the electron. This neutral particle had never been observed; a calculation of its interaction probability with matter by Bethe and Peierls gave a mean free path for absorption of thousands of light years in lead. In the opinion of physicists in the 1930s it was therefore undetectable, a ghost.

ENRICO FERMI cast these ideas of Pauli into a theory of beta decay: he introduced a revolutionary new concept in his theory by treating the electron and the neutrino, both spin 1/2 particles, as fields which are created together, in analogy to the creation of a photon in nuclear transitions. Later this was called the Fermi interaction. The agreement between calculated and measured beta spectra and decay rates was excellent. The mass of the neutrino was apparently smaller than the electron mass.

Does the neutrino exist as a particle? It took nearly 30 yr to answer this question. The problem of the small absorption cross-section required a high-intensity neutrino source and a very massive detector which would identify the absorption of the antineutrino by a proton, i.e. $\bar{\nu}_e p \rightarrow e^+ n$, producing a positron and a neutron. After World War II high-power nuclear reactors were available which produce a flux of 10^{13} $\bar{\nu}_e$ cm^{-2} s^{-1} at a distance of 10 m from the core. Reines, Cowan and collaborators built a detector of 200 ton target mass which detected the annihilation of the positron into two prompt 0.5 MeV photons and the capture of the neutron in cadmium dissolved in water, producing several photons after a characteristic delay of about 5 μs. This signature discriminated against background induced by NEUTRONS, PHOTONS and COSMIC RAYS. The measured capture rate agreed with the theoretical cross-section within the rather large uncertainty of the calculated antineutrino flux. In 1956 they obtained evidence that the neutrino is indeed a particle. Five years later, in 1962, Ledermann, Schwartz and Steinberger discovered a second neutrino, the muon-neutrino. They invented a new research tool, a neutrino beam, by intercepting the 20 GeV proton beam of the Brookhaven National Laboratory Synchrotron with a target plate. Positively charged pions are selected by the fringing field of the accelerator; in flight a fraction of a few per cent decay into a muon and a neutrino before they are absorbed by a 5.5 m thick steel wall. The muons were brought to rest by ionization loss in the steel wall. Only neutrinos pass through the wall. They interact with matter in the detector. What kind of neutrino appears in the decay of a pion into a muon? The detector was designed to discriminate between electron-neutrinos which produce electrons and muon-neutrinos which produce muons in the absorption process by a neutron. All 30 single-track events observed had a muon. It was concluded that neutrinos exist in families. Today we know that a third neutrino flavor, the tau-neutrino, exists.

A very sensitive measurement of the number of neutrino flavors was performed at the e^+e^- collider LEP at CERN. From a measurement of the resonance shape of the Z, a mediator of the neutral current weak interaction, discovered at CERN in 1983, the number of neutrino flavors with mass lighter than one-half the Z mass (90 GeV) was determined to be

$$N_\nu = 2.998 \pm 0.029.$$

There are three light neutrinos in nature and three QUARKS.

The method of the neutrino beam created a new area of research called neutrino physics which provided a rich harvest. Because of the point-like nature of the neutrino interaction they are ideal probes. In a bubble chamber exposed to a neutrino beam at CERN (Geneva, Switzerland) it was observed that the interaction probability of neutrinos increased with energy. This observation was attributed to a point-like substructure of matter, due to the quarks which compose the nucleons. The number of quarks per nucleon was found to be three by comparing the interaction probability of neutrinos and electrons, in agreement with the quark model of matter proposed by Gell-Mann, the proton being composed of two u (up) quarks and one d (down) quark, the neutron of two d quarks and one u quark. About 50% of the neutrino interactions in matter do not take place on quarks. These interactions with matter take place on another constituent, the gluon which binds the quarks in matter. It was found that gluons produce a virtual sea of quark–antiquark pairs. Measurements with neutrino and antineutrino beams led to a quantitative analysis of the quark and antiquark structure of matter. A special class of events induced by high-energy neutrinos was found to contain two muons in the final state, a leading muon of negative charge in muon-neutrino interactions and a positively charged muon, in general of lower energy, produced at the hadron vertex. The second muon is a decay product of a heavy quark called the charmed or c quark. It is produced by the fusion of a virtual W particle, the mediator of the charged current weak interaction, and either the s quark or the d quark. The transition of a d quark to a c quark is reduced by the Cabibbo angle, compensating nearly for the smaller presence of s quarks in the sea and producing almost equal contributions. In antineutrino interactions an antiquark fuses with a virtual W^+ particle to produce a $\bar{c}$ quark with decays into μ^- and a K^0 and an anti-neutrino.

A detailed study of these processes led to the experimental determination of the quark mixing matrix which has 3×3 elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In neutrino events observed in the Gargamelle bubblechamber in 1973 a new phenomenon was discovered, a new weak interaction of neutral currents. The observation of muon-neutrino-induced events without a muon in the final state was the experimental evidence. Initially events of two kinds were observed, the elastic scattering of an anti-muon-neutrino on an electron and events in which all charged tracks were identified to be hadrons either by visible interactions or by decays. Soon afterwards the process was identified in general, inclusive reactions. The interaction is mediated by a neutral heavy boson, the Z, which was discovered at CERN, together with the charged boson W, in 1984 by C Rubbia and S van der Meer. A new theory of the weak and the electromagnetic interactions formulated by S Weinberg, S Glashow and A Salam unifies the description of these forces (see GRAND UNIFIED THEORY). The neutral current weak interaction is due to a mixing of the charged current weak interaction and the electromagnetic interaction. The mixing is described by an angle, the electroweak mixing angle, which was determined experimentally from the ratio of neutral and charged current weak interaction cross-sections. The value of this mixing parameter is related to the ratio of the W and the Z masses. The early measurements of the mixing parameter in neutrino experiments led to a prediction of these masses ($M_{\mathrm{W}} \sim 82$ GeV, $M_{\mathrm{Z}} \sim 90$ GeV) which was later confirmed by the direct measurements of the masses.

This agreement was an important confirmation of the unified theory of the electroweak interaction. Even more precise measurements were obtained at the e^+e^- colliders operating at the Z mass. This new theory made real progress compared with the Fermi theory. The unification is achieved by a renormalizable gauge theory which gives finite results for higher-order calculations, whereas in the Fermi theory these calculations lead to divergent results. Hence, the progress is tremendous.

The transformation properties of the fermions, leptons and quarks are given by their doublet structure. There are three lepton doublets and three quark doublets each with an 'up' state of weak isospin +1 and a 'down' state with weak isospin −1. Local gauge symmetry requires this structure.

Experimentally they can be verified by neutrino–electron scattering and by e^+e^- elastic scattering. $\nu_\mu e$ and $\nu_e e$ scattering and $\bar{\nu}_\mu e$ and $\bar{\nu}_e e$ scattering determine two coupling solutions; e^+e^- elastic scattering determines eight cones of coupling values. One solution is found to be common; it is the one which determines that the electron is the down state with respect to the neutrino, as predicted by local gauge symmetry. The measurements of $\nu_\mu e$ and $\bar{\nu}_\mu e$ scattering have been performed in the high-flux neutrino beam at CERN; those of e^+e^- elastic scattering have been performed at LEP.

Astrophysics

The reaction which fuses four hydrogen nuclei to helium by the weak interaction is the main energy source of the Sun and of similar stars. Two neutrinos are emitted in this reaction. On Earth 10^{10} neutrinos cm^{-2} s^{-1} arrive on average. Other reactions involve beryllium and boron. The flux of neutrinos from the Sun has been studied using three different neutrino detection reactions and techniques (see SOLAR INTERIOR: NEUTRINOS):

(1) $^{37}\mathrm{Cl} + \nu_e \rightarrow e^- + {}^{37}\mathrm{Ar}$, threshold $E_\nu > 814$ keV;

(2) $\mathrm{Ga} + \nu_e \rightarrow e^- \mathrm{Ge}$, threshold $E_\nu > 233$ keV;

(3) $\nu_e e$ scattering, detection of the electron by Čerenkov radiation in water; $E_{\mathrm{threshold}}$ of detection >7.5 MeV.

Method (1) was used to record solar neutrinos over ∼20 yr. It is the first experimental confirmation of neutrinos from the Sun. The measured neutrino flux is a factor of 3 smaller than the calculated flux. Only ^{8}B neutrinos are detected. Method (2) is detecting neutrinos from hydrogen fusion and from ^{7}Be and from ^{8}B. The measured flux is a factor of 2 smaller than the calculated flux.

Method (3) can reconstruct the direction of the incident neutrino and is adding the important confirmation that the detected neutrinos indeed come from the Sun. Is the missing flux due to a deficiency of the calculation? The ^{8}B neutrino flux varies with the 18th power of the temperature in the center of the Sun. Alternatively, is it due to a new property of neutrinos which changes their identity over the distance to the Earth? If they 'oscillate' into muon-, tau- or so-called sterile neutrinos they cannot be detected by any of these methods. Combining the results of the three methods and adding the constraint of the total solar luminosity, one can determine the different fluxes separately. One solution gives a strong suppression of ^{7}Be neutrinos, assuming the standard solar model is correct. Such a suppression is predicted by a resonance in the transition of $\nu_e \rightarrow \nu_\mu$; owing to the interaction of ν_e with electrons in the Sun the effective mass of the ν_e crosses the mass of ν_μ. This level crossing leads to a resonant transition $\nu_e \rightarrow \nu_\mu$. This matter-enhanced oscillation is, however, not supported by the present observation of the day/night ratio in method (3). If matter-enhanced oscillations were at work in the Sun it would enhance the ν_e flux when the Earth is between the detector and the Sun. Small changes in the energy spectrum of solar neutrinos resulting from matter-enhanced oscillation have not been detected either. The question of why solar neutrinos are missing cannot be answered at present.

The gravitational collapse of massive stars with a core of more than 1.5 solar masses at the end of the burning of light elements up to Fe is called a supernova. At the high density reached in the collapse electrons are squeezed into protons, producing a neutron core and neutrinos. The potential energy liberated by the collapse of 10^{53} erg is radiated mainly by neutrinos. The radiation of neutrinos was first detected for a supernova observed in 1987 (see SUPERNOVA 1987A).

Twenty neutrino events were detected by Čerenkov radiation of the recoiling electron from ν_e scattering in water. This rate of neutrino events is in good agreement

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with a theory of supernova collapse elaborated by Bethe. The burst of neutrinos lasted for 10–15 s. The two detectors which confirmed each other, one in a mine in the USA and the other in a mine in Japan (see NEUTRINO ASTRONOMY), were originally designed to search for proton decay.

Neutrino oscillation

No other phenomenon has been searched for more intensively. This hypothetical process due to the mixing of different neutrino states was invented by Bruno Pontecorvo and by Z Maki in 1958 and in 1967 to explain the lack of solar neutrinos. Neutrinos propagate as mass eigenstates. If the propagation states have different masses, quantum mechanical interference leads to the disappearance of one neutrino flavor and the appearance of another. The oscillation length depends on the difference of the squares of the masses of the eigenstates and the fraction of neutrino flavor disappearing or appearing on the strength of mixing. A 3×3 mixing matrix describes all possible phenomena. As a theory of flavor mixing does not exist no predictions of neutrino masses and mixing can be made.

The first convincing observation, a disappearance of muon-neutrinos of atmospheric origin, was reported in 1998 by a Japanese group using a large water Čerenkov counter (40 000 ton) in an underground mine in Japan (Kamioka). Atmospheric neutrinos are born in the decay of pions produced by primary cosmic rays in the Earth's atmosphere in the ratio of two muon-neutrinos to one electron-neutrino. Because of the isotropic flux of primary cosmic rays, up- and down-going neutrinos are expected to have equal rates to within 1–2%. The Japanese experiment observed a strong up–down asymmetry; nearly 50% of the up-going neutrinos were missing. The mean disappearance length was estimated to be ∼4000 km at 1 GeV, corresponding to a difference of mass squares of 3×10^{-3} eV^2 or a mass difference of ∼0.05 eV. The mixing appears to be maximal.

The appearance of the corresponding number of neutrinos of another flavor, supposedly tau-neutrinos, has not yet been observed. The water Čerenkov detector cannot identify them. A new experiment has to be built to make this observation. It will use a muon-neutrino beam from the CERN SPS and will be located in the Gran Sasso underground laboratory near Rome at a distance of 730 km. It is planned to start in the year 2005.

Open questions

The phenomenon of a finite neutrino mass cannot be described within the standard model of electroweak interaction. The theory has to be extended to a larger symmetry scheme. The most puzzling feature observed is the nearly maximal mixing of flavors in the case of $\nu_\mu \rightarrow \nu_\tau$ oscillation. This does not seem to be the case for $\nu_e \rightarrow \nu_\tau$ oscillation. A mass difference has not yet been observed for these flavors.

Since neutrinos have a finite mass they can also have a magnetic dipole moment which remains to be detected. The cosmological consequences of the finite neutrino mass have still to be elaborated.

Bibliography

Winter K (ed) 2000 *Neutrino Physics* 2nd edn (Cambridge: Cambridge University Press)

Klaus Winter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neutron

An uncharged subatomic particle that is composed of three quarks. With a mass of 1.6748×10^{-27} kg (which is equivalent to a rest-mass energy of 939.6 MeV), it is slightly heavier than the proton. It has an intrinsic angular momentum, or spin, of 1/2 (in units of the Planck constant, h, divided by 2π).

The neutron is a nucleon, one of the two basic constituents of all atomic nuclei apart from the normal form of hydrogen (which consists of a single proton). The total number of protons and neutrons in an atomic nucleus defines its mass number. An isolated neutron decays into a proton, an electron and an antineutrino on average after about 15 minutes, this process being called 'beta decay'. This process is responsible for the most common form of radioactivity (beta radioactivity), in which the emission of an electron and antineutrino accompanies the conversion of a bound neutron into a proton. This changes the atomic number of the nucleus (and therefore changes, or 'transmutes', it into a nucleus of a different element) but leaves the mass number unchanged. In many other nuclei, a bound neutron is a stable, long-lived, particle.

See also: antimatter, atom, baryons, electron, electron-volt, hadrons, isotope, neutrinos, quark, radioactivity, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neutron Stars

Neutron stars are small compact stars with densities comparable to that inside nuclei, $\rho_0 = 2.8 \times 10^{14}$ g cm^{-3}, and radii $R \simeq$ 10–15 km. They consist predominantly of NEUTRONS and a few percent of PROTONS and ELECTRONS. These huge neutron-rich 'nuclei' are bound by gravitation and require a minimum neutron star mass of $\sim 0.1 M_\odot$. Above a maximum (Chandrasekhar) mass of order $2–3 M_\odot$ neutron stars are unstable towards gravitational collapse to BLACK HOLES.

Neutron stars are formed in type II or Ib SUPERNOVA explosions when massive stars ($M \geq 10 M_\odot$) run out of nuclear fuel after burning for millions of years. When the iron core in the center of the aging stars exceeds its Chandrasekhar mass, $\sim 1.5 M_\odot$, the star undergoes gravitational collapse in just seconds and suffers violent death. Gravitational and kinetic energy of the order $\sim 10^{53}$ ergs is released mainly by neutrino emission that blows off the outer layers. Only ~1% of the energy is actually seen in a brilliant burst, the supernova. Measurements of isotopic abundances of various elements (*r*- and *s*-processes) give some insight into the densities and temperatures during certain stages of the explosion in which all heavier elements in the universe were made. Supernova explosions occur at a rate of one to three per century in our Galaxy. Historically, the Crab supernova was first recorded by the Chinese in AD 1054, and could be seen in daylight. The discovery in 1967 of a rotating neutron star (a radio pulsar) in the CRAB NEBULA supernova remnant confirmed the 1933 hypothesis of Baade and Zwicky, formulated a year after Chadwick discovered the neutron, that neutron stars are formed in supernova explosions. In our Galaxy astrophysicists expect a large abundance, $\sim 10^8$, of neutron stars but they only manifest themselves if powered by rotation, accretion, nuclear power or residual heat.

Radio PULSARS were first discovered by Bell and Hewish in 1967 and extensive searches have led to ~1200 known pulsars at present. Pulsars are rotation powered neutron stars where the magnetic B and rotational $\boldsymbol{\Omega}$ axes are misaligned by an angle θ. Consequently they emit dipole radiation in the form of radio waves that appear to pulse on and off like a lighthouse beacon as the pulsar beam sweeps across the Earth with the spin frequency of the neutron star. The rotational energy loss for dipole radiation is

$$\dot{E} = I\Omega\dot{\Omega} = -\frac{B^2 R^6 \Omega^4 \sin^2\theta}{6c^3} \quad (1)$$

where the moment of inertia for a typical neutron star is $I \sim 10^{45}$ g cm^2. The radio pulses are extremely regular with periods in the range 1.5 ms $\leq P \leq$ 8.5 s and the pulsars are spinning down slowly with period derivatives of the order $\dot{P} \sim 10^{-12}$–10^{-21}. The pulsar age is from equation (1) given by $t \simeq P/2\dot{P}$ and the magnetic field $B \sim 10^{12} G(\dot{P} \times 10^{15} P \text{ s}^{-1})^{1/2}$. The resulting magnetic fields of the order of $B \simeq 10^{12}$ G are independently confirmed by cyclotron absorption lines found in x-ray spectra. Most pulsars are old (~10 million years) and slowly rotating with relatively small period derivatives. In a (P, $\dot{P}$) diagram they cluster around $P \simeq 1$ s and $\dot{P} \simeq 10^{-15}$. A few young pulsars, e.g. those such as the CRAB and VELA PULSARS, have short periods and large period derivatives.

Table 1. Properties of a typical neutron star.

Mass	M	$\sim 1.4 M_\odot$
Radius	R	10–15 km
Central densities	ρ_c	$\sim 10^{15}$ g cm^{-3}
Moment of inertia	I	$\sim 10^{45}$ g cm^{-3}
Rotation period	P	0.0015–8.5 s
Spin down	$\dot{P}$	$\sim 10^{-15}$
Magnetic fields	B	$\sim 10^{12}$ G
Temperatures	T	$\lesssim 5 \times 10^6$ K

The nearest known neutron star, RXJ185635-3754, located in the southern constellation Corona Australis, was imaged by the Hubble Space Telescope in 2001. Its distance is estimated by parallax at only 200 light years. From its black body spectrum, a surface temperature of half a million Kelvin is found, and from its flux a radius of order 10 km is inferred. Circumstantial evidence indicates that it was born in a supernova explosion a million years ago. Such an age and surface temperature is compatible with standard cooling predictions.

Binary pulsars were discovered by Hulse and Taylor in 1973. They form a distinct subclass of mainly millisecond radio pulsars with periods 1.56 ms $\leq P \leq$ 100 ms. The period derivatives are very small corresponding to very small magnetic fields $B \sim 10^8$–10^{10} G. They are believed to be old recycled pulsars, which have been spun up by mass ACCRETION. In this process the magnetic fields of the pulsars have decayed. After the recycling the spun-up millisecond pulsars are expected to emit radio waves for billions of years. About 30—more than half of the known millisecond pulsars—are found in BINARY STARS where the companion is either a WHITE DWARF or a neutron star. Six double neutron stars are known so far. By measuring the inward spiralling or orbital decay one can determine many parameters in the binary systems such as the neutron star and companion masses, orbital periods and period derivatives, orbital distances and inclination. Parameters are overdetermined and provide a test of general

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

relativity to an unprecedented accuracy. The PERIASTRON advance in PSR 1913 + 16 is 4.2º per year as compared with 43 arcseconds per century for Mercury, which was originally used by Einstein to test his theory of general relativity. Binary neutron stars all have masses in the narrow range $1.35\pm0.07M_\odot$. This number is suspiciously close to the Chandrasekhar mass of the iron core, $M_{ch} \simeq 1.5M_\odot$, when subtracting the gravitational energy loss, $\sim GM^2/R \sim 0.1M$, that dissappears by neutrino emission in the collapse.

X-ray pulsars and X-RAY BURSTERS are accretion powered neutron stars (see figure 1). Almost 200 have been discovered within the last three decades by satellite borne x-ray detectors. The orbital period has been determined for about 60 of these X-RAY BINARIES. The x-ray pulsars and bursters are believed to be neutron stars accreting from high-mass ($M \gtrsim 10M_\odot$) and low mass ($M \lesssim 1.2M_\odot$) companions respectively. The former systems are fed by wind accretion whereas the latter are the result of ROCHE-LOBE overflow from a giant companion star. The x-ray pulses are most probably due to strong accretion on the magnetic poles emitting x-rays (as northern lights) with orbital frequency. The x-ray bursts are due to slow accretion spreading all over the neutron star surface before igniting in a thermonuclear flash. The resulting burst time-scales depend on accretion rates rather than orbital periods. One bursting pulsar has recently been discovered. The radiation from x-ray bursters is not blackbody and therefore only upper limits on temperatures can be extracted from observed luminosities in most cases. Masses are less accurately measured than for binary pulsars but a few, for example, the x-ray pulsars Vela X-1 ($M \simeq (1.9\pm0.2)M_\odot$) and the x-ray burster Cygnus X-2, seem to be heavier than the binary pulsars which could be due to accretion.

Recently, a subclass of half a dozen anomalous x-ray pulsars have been discovered. They are slowly rotating, $P \sim 10$ s, but rapidly slowing down. This requires huge magnetic fields of $B \sim 10^{14}$ G and they have appropriately

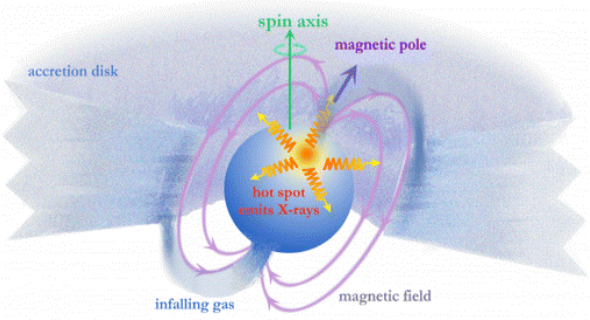


Figure 1. An accreting neutron star with misaligned spin and magnetic axes.

been named 'magnetars'. Four gamma-ray repeaters discovered so far are also slowly rotating. The anomalous x-ray pulsars, and probably also the gamma-ray repeaters, appear to reside inside SUPERNOVA REMNANTS. Together they constitute just as many galactic neutron star candidates inside supernova remnants as the radio pulsars—in spite of the extensive searches for the latter. Of the ~1200 discovered radio pulsars only about ~10 are associated with the 220 known supernova remnants.

Quasi-periodic oscillations (QPO) have very recently been discovered in low-mass x-ray binaries. QPOs provide a promising new method for determining masses and radii of neutron stars. The QPO are brightness oscillations in the x-ray luminosity. In many systems two kHz QPO peaks appear with frequencies that increase with accretion but with almost constant frequency difference. The pair is most likely the orbital frequency of accreting matter in Keplerian orbits around the neutron star and its beat frequency with the neutron star spin. In the case of 4U 1820-30 the accretion may have been tracked to its innermost stable orbit, since the QPO frequency displays a distinct saturation with accretion rate. From the corresponding maximum frequency $\nu_{QPO} = 1060$ Hz, one deduces a neutron star mass of $M \simeq 2.25M_\odot$. The existence of such heavy neutron stars would restrict the equation of state for neutron star matter severely.

X-ray bursts are thermonuclear explosions of accreted matter on the surface of neutron stars. After accumulating hydrogen on the surface for hours, the pressure and temperature become sufficient to trigger a runaway thermonuclear explosion seen as an x-ray burst. The energy involved is of the order of typical nuclear binding energies, ~7 MeV per hydrogen atom. Slow accretion from a small mass companion generates a continuous background of x-rays. Each nucleon radiates its gravitational energy of $\sim m_n GM/R \simeq 200$ MeV. The burst energy is therefore only about a per cent of the time integrated background but is released in only a few seconds. In the case of 4U 1728-34 x-ray modulation due to the neutron star spin frequency of 364 Hz has been found. It decreases, however, to 362 Hz during the first 1–2 s of the burst. A simple explanation is that the thermonuclear explosion elevates the surface of the neutron star. Conserving angular momentum, $L \propto MR^2\nu$, leads to a decrease in rotation by $\Delta\nu/\nu \simeq -2\Delta R/R$. With a frequency change of $\Delta\nu \sim -2$ Hz and typical neutron star radius of order $R \sim 10$ km, an elevation of order $\Delta R \sim 20$ m is found, which is roughly in agreement with expectations but much less than on Earth due to the much stronger gravitational fields on neutron stars.

The structure of a typical $1.4M_\odot$ mass neutron star is to first approximation a ~1 km thick crust and a nuclear liquid in the interior. At terrestrial densities the ground state of matter consists of ^{56}Fe nuclei in a solid body-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

centered cubic lattice and electrons. With increasing density, the atomic cores begin to overlap and matter becomes pressure ionized. At a density $\sim 10^6$ times terrestrial density, or of order 10^6 g cm^{-3}, electrons become relativistic. At a density of about 10^7 g cm^{-3} it becomes energetically favorable to reduce the number of electrons by electron capture, converting protons to neutrons and form ^{62}Ni. At higher densities, the equilibrium nuclei become even more neutron rich, and eventually, at a density of about 4×10^{11} g cm^{-3}, all bound neutron levels in nuclei are occupied, and neutrons 'drip'. The inner crust consists of neutron-rich nuclei embedded in a gas of neutrons and relativistic electrons. As densities approach the density of nuclear matter the nuclei may merge into intricate rod-, plate- and bubble-like structures. The sizes of the structures are determined by a competition between Coulomb and surface energies. The inner crust terminates at a density close to nuclear matter density at $\rho \simeq 2 \times 10^{14}$ g cm^{-3}, where the transition to a spatially uniform state occurs which is believed to be a superfluid (see figure 2).

In the core of neutron stars many other problems arise. The basic difficulty in predicting neutron star properties is that matter at supernuclear densities is not well understood. There are a number of possible states for matter. The simplest is a uniform liquid of neutrons and protons in β-equilibrium, with a neutralizing background of negative charge from electrons, and, when the electron chemical potential exceeds the muon rest mass, also muons. There could be other particles present, such as hyperons. Another class of possibilities is Bose condensation of mesons, either pions or kaons. Yet another is quark matter, in which the nucleons are broken down into quarks. To make progress it is necessary to develop a better understanding of many aspects of physics (see e.g. PARTICLE ASTROPHYSICS). One is the basic interactions between hadrons. While two-particle interactions between nucleons are well characterized from scattering data, the corresponding interactions for strange particles are less secure. Knowledge of many-body interactions is much more rudimentary.

The equation of state (EoS) of dense matter, e.g. pressure versus energy density $P(\epsilon)$, can therefore only be calculated reliably up to a few times normal nuclear saturation densities by using measured nucleon–nucleon scattering data and binding energies and saturation density of nuclei. At higher densities causality provides some constraint on the EoS. Given the EoS, the structure of non-rotating neutron stars can be calculated from the Oppenheimer–Volkoff equation

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -G\frac{(P+\epsilon)(m+4\pi r^3 P)}{r^2(1-2Gm/r)} \qquad (2)$$

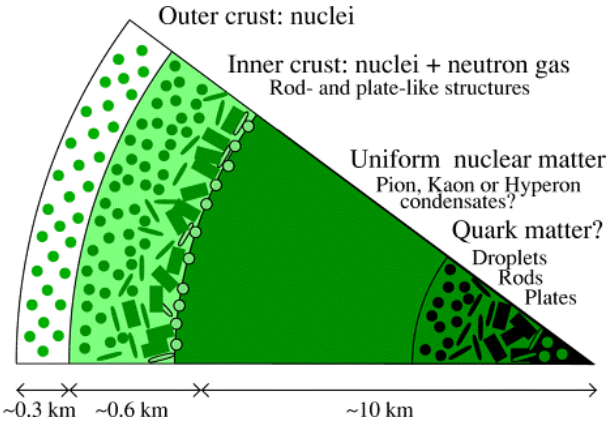


Figure 2. Schematic cross section of a neutron star. Possible components are listed, and estimates of radial dimensions are drawn (not to scale).

where $m(r) = 4\pi \int_0^r \epsilon(r')r'^2\,\mathrm{d}r'$. Rotation was included by Hartle. The maximum (Chandrasekhar) mass is expected to be below $\sim 2.5 M_\odot$ depending on the stiffness of the nuclear EoS, but it could be lower if phase transitions take place. Observations of large neutron star masses of order $\sim 2.3 M_\odot$ would therefore restrict the equation of state (EoS) severely for dense matter.

Glitches are sudden speed-ups in the rotation rate observed in a number of younger pulsars, e.g. the Crab and Vela. Glitches occur due to quakes in solid structures such as the crust, superfluid vortices or possibly a quark matter lattice in the core. As the rotating neutron star gradually slows down and becomes less deformed, the rigid component is strained and eventually cracks/quakes and changes its structure towards being more spherical. The moment of inertia of the rigid component, I_c, decreases abruptly and its rotation and pulsar frequency increase due to angular momentum conservation resulting in a glitch, $\Delta\Omega/\Omega \sim 10^{-6}$–$10^{-8}$, which slowly relaxes on a time scale of days as the rigid and fluid components approach a common rotational frequency.

Surface temperatures may provide information about the interiors and thermal evolution of neutron stars. In recent years much effort has been devoted to measuring neutron star temperatures, especially with the Einstein Observatory and ROSAT. Neutron stars are born with interior temperatures of the order of 10^{12} K, but cool rapidly via neutrino emission to temperatures of the order of 10^{10} K within minutes and $\lesssim 10^6$ K in 10^5 yr. The only information on neutron star temperatures stems from surface temperatures observed in x-ray or UV bands. Most observations are for pulsars, and it is unclear how much of the observed radiation is due to pulsar phenomena, to a synchrotron-emitting nebula or to the neutron star itself. Surface temperatures of neutron stars have been measured in a few cases or upper limits have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

been set. These seem to be compatible with predictions from standard modified URCA cooling processes. Non-rotating and non-accreting neutron stars are virtually undetectable but the Hubble Space Telescope has observed one single thermally radiating neutron star RX J185635-3754. Its surface temperature is $T \simeq 6 \times 10^5$ K $\simeq 50$ eV and its distance is less than 120 pc from Earth. Circumstantial evidence indicates a distance of ~80 pc which leads to a neutron star radius of 12–13 km in good agreement with modern EoS.

Gamma ray bursters (GRB) have recently been associated with afterglows from which very high REDSHIFTS ($z \geq 1$) could be determined. An enormous energy output of $E \sim 10^{53}$ ergs can be inferred, if isotropically emitted, but there is evidence for beaming or jets and the GRB therefore 'only' requires a total energy of $E \sim 10^{51}$ ergs in gamma rays. Candidates for such violent events include neutron star mergers or, as indicated by recent observations of coinciding supernovae, a special class of type Ic supernova (hypernovae) where cores collapse to black holes.

NEUTRINO and GRAVITATIONAL WAVE detectors on Earth will look into the formation and cores of neutron stars. LIGO will be able to detect gravitational waves from core collapse in supernova and coalescence of binary neutron stars. Neutrinos carry away ~99% of the energy in supernova explosions and their abundance, timing and spectra reveal detailed information about the explosion and the proto-neutron star. Nineteen neutrinos were detected from SN-1987A in the Large Magellanic Cloud. Modern neutrino detectors should catch several thousand neutrinos from future supernovae in our galaxy or the neighboring Magellanic Clouds.

The future of neutron stars looks bright as new windows are about to open. A new fleet of x- and gamma-ray satellites will be launched. With upgraded ground-based observatories and detectors new classes of neutron stars will be discovered and our understanding of neutron star properties will be improved.

Bibliography

Shapiro S L and Teukolsky S A 1983 *Black Holes, White Dwarfs and Neutron Stars* (New York: Wiley)

Henning Heiselberg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

New General Catalog of Nebulae and Clusters of Stars (NGC)

A basic reference list of star clusters, nebulas and galaxies compiled in 1888 by Danish astronomer Johan Ludvig Emil Dreyer. Dreyer worked at the Armagh Observatory in Northern Ireland. Here he based his work on lists made earlier by the Herschel family of British astronomers and John Herschel's General Catalog of Nebulae, published in 1864. Dreyer included 7840 celestial objects. He later extended the list by 5386 objects with his first and second Index Catalogs (IC), published in 1895 and 1908 respectively. With these supplements the NGC covers the entire sky, although many objects visible with modern instruments are not listed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Newcomb, Simon (1835–1909)

Mathematical astronomer, born in Wallace, Nova Scotia, Canada, moved with his family to Maryland and self-studied mathematics. Worked at the US Nautical Almanac Office (then in Cambridge, MA), studied at Harvard, and was appointed to the US Naval Observatory at Washington, DC, becoming director. In his own words, because of the 'confusion which pervaded the whole system of exact astronomy, arising from the diversity of the fundamental data made use of by the astronomers of foreign countries and various institutions in their work' he started 'a systematic determination of the constants of astronomy from the best existing data, a reinvestigation of the theories of the celestial motions, and the preparation of tables, formulae, and precepts for the construction of ephemeredes, and for other applications of the same results'. He thus used careful measurements of stellar and planetary positions to compute formulae for their future motions, with an analysis of the constants in the formulae (such as the masses of the Sun, Moon and planets), and the speed of light. This set in train a century of astronomical activity focused in the work of the International Astronomical Union to define the astronomical constants.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Newton

Isaac Newton (1642–1727) is known pre-eminently for discoveries in mathematics (binomial theorem and fundamental theorem of the calculus), optics (the heterogeneity of white light) and mechanics (laws of motion and universal gravitation). Not undisputed are some questions of priority and how in detail to characterize these achievements. Beyond question, however, is the foundational character of Newton's chief works for all later physical science: his *Principia* (editions in 1687, 1713 and 1726) and *Opticks* (English editions in 1704 and 1717, Latin editions in 1706 and 1719). Various understandings of Newton's methodology played an important role in Enlightenment thought.

Newton's paternal forebears were illiterate Lincolnshire yeomen. His grandfather, prospering, purchased Woolsthorpe Manor. His father on inheriting the estate married Hannah Ayscough, then died six months later before his son Isaac was born. When Isaac was three, his mother remarried and went to live with her new husband, the Rev. Barnabas Smith, leaving Isaac till age ten (when Smith also died) in the care of his maternal grandmother at Woolsthorpe. Newton's boyhood was lonely and punctuated by fits of anger against his mother and step-father. Neurotic isolation remained characteristic into his middle years.

Between ages 12 and 17 Newton attended grammar school in Grantham, becoming fluent in Latin. Put then to managing the Woolsthorpe estate, he was a disaster, interested only in bookish learning. His mother at length relented, sending him back to Grantham to prepare for university entrance. He matriculated at Trinity College, Cambridge, in July 1661.

The Cambridge curriculum was a moribund Aristotelianism, but Newton's solitary reading by 1664 had led him beyond the prescribed course to GALILEO's *Dialogue*, DESCARTES' *Principia*, Walter Charleton's epitome of Gassendi's writings, and other proponents of new philosophy. He discovered the new mathematics in Schooten's second Latin edition of Descartes' *Geometry* and in John Wallis's *Arithmetic of Infinites*.

Despite the aberrancy of his studies, Newton in 1664 was elected scholar at Trinity, and in 1667 Minor Fellow, then Major Fellow in 1668. In 1669 Isaac Barrow saw to it that Newton became his successor as Lucasian Professor of Mathematics. To continue as a Fellow required taking holy orders within seven years, but study of Scripture and the Church Fathers convinced Newton that the Trinity was a hoax. By royal dispensation, obtained through Isaac Barrow's intercession, he was freed from the requirement. Newton would remain in Cambridge till 1696 when he moved to London to become Warden of the Mint, then Master of the Mint in late 1699; he resigned his Trinity College fellowship and professorial chair in 1701. His scientifically most fruitful years were over.

Newton's discoveries in mathematics

Newton's most intense study of the new mathematics began sometime in 1664, and lasted until November 1665, by which time he had made his principal discoveries. Quickly he assimilated the central insight of DESCARTES' *Geometry*: curves were to be understood through their equations. He reduced to formula Descartes' method of finding a tangent to a curve (it involved constructing the normal). He then considered normals at two nearby points on the curve with abscissas x and $x + o$. The limiting position of the intersection of the two normals as o went to zero gave the center of curvature, and thence the radius of curvature—a result beyond anything in the texts he was studying.

Wallis's *Arithmetic of Infinites* introduced him to the definite integral, an area under a curve between fixed ordinates, computed by dividing the area into infinitesimal rectangles and summing them as an infinite series. Newton made the upper limits variable, and so arrived at indefinite integrals, for instance (in our symbolism):

$$\int_0^x y\,\mathrm{d}x = \int_0^x x^3\,\mathrm{d}x = x^4/4.$$

Wallis often solved problems by a kind of induction, and Newton used just such a process in solving one of the problems Wallis had posed: to find the area under a circle, that is, to integrate $\int(1 - x^2)^n\,\mathrm{d}x$ where n is $1/2$. If n were a positive integer the integral would be obtainable by expanding the binomial and integrating the resulting polynomial term by term. The successive terms of the integral in this case, Newton observed, are given by multiplying x, $-x^3/3$, $x^5/5$, $-x^7/7$, etc, respectively by the successive binomial coefficients. Discerning the rule by which the binomial coefficients are produced, and extending it to the cases in which $n = 1/2, 3/2, 5/2$, etc, Newton obtained the integrals $\int(1 - x^2)^n\,\mathrm{d}x$ for these fractional values of n as infinite series; thus for $n = 1/2$ the integral is

$$1x + (1/2)(-x^3/3) + (-1/8)(x^5/5) + \cdots.$$

In this manner all radicals could be integrated. Here and elsewhere Newton intuited general rules and illustrated them in specific examples, without strictly demonstrating them or expressing them in a general symbolism.

Dissatisfied with the concept of infinitesimals, Newton shifted to a kinematic view: areas were swept out by a moving ordinate, its length giving the momentary rate of areal increase. Likewise, a curve could be viewed as produced by a moving point, the tangent at a point being determined by the limiting ratio of the component speeds in the x and y directions. To these 'speeds' in the production of an area or curve Newton gave the name 'fluxions'. The determination of the tangent now appeared as the inverse of the finding of an area under a curve. This inverse relation is the fundamental theorem of the calculus. Here again, Newton illustrated the theorem in examples without expressing it in a general symbolism.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

After mid-November 1665, Newton's interest in mathematics lapsed. In autumn 1666 he produced a summary of his discoveries, and in 1669, at Barrow's urging, he set himself to producing a full-fledged treatise on his methods of series and fluxions. Fatefully, he abandoned the unfinished treatise in 1671, apparently overcome by fear of criticisms that publication could bring. The incomplete treatise was not published until after his death.

Newton's optical discoveries

Newton's work in optics began in early 1666 when he obtained a prism 'to try the celebrated *Phaenomena of Colours*'. Descartes, Boyle and Hooke had taken color to be a modification of white light—the traditional view. Newton's understanding of color grew gradually, in tandem with experiments. A circular beam of white light, on passing through the prism, gave an elongated spectrum, implying different refractions of its different parts. Other experiments showed the separate rays to have unalterable degrees of refrangibility. These rays on being recombined gave white light. White light was thus a mixture composed of individual rays of constant refrangibility.

Newton devoted his first series of professorial lectures (in 1670) to the theory of colors. The phenomena implied, he believed, the corpuscularity of light: he saw no way in which a wave theory could agree with distinctive natures in the rays. HOOKE in his *Micrographia* (1665) had investigated the colors of thin films, and Newton now carried Hooke's investigation further, determining the varying thickness of the film between a spherical lens and a plate glass to hundredths of an inch, and thereby establishing a periodicity associated with the successive rings of color and darkness that presented themselves ('Newton's rings'). He attempted to explain the periodicity as due to vibrations set up in the aether by the impact of the light corpuscles; the aether at the second interface between glass and air was thus put into 'alternate fits of easy reflection and refraction'.

By his own labor Newton designed and constructed a reflecting telescope. Completed in early 1669, it was six inches long, magnified 40 times, and was free of the chromatic aberration affecting refractors. Newton showed it off in Cambridge; word of it spread to London. Members of the Royal Society asked to see it, and in late 1671 Barrow presented it to the Society. It caused a sensation. Flattered, Newton agreed to provide an account of his theory and experiments. His 'New Theory about Light and Colors' appeared in the *Philosophical Transactions* in February 1672.

With this, his first publication, Newton's worst fears came true. The force of his experiments was not understood. Hooke took his main point to be the corpuscular nature of light, and was unpersuaded by Newton's attempt to separate the issues of corpuscularity and heterogeneity. When HUYGENS referred to the theory as 'an hypothesis', Newton bridled. Most wearying of all were the objections and experimental incompetence of three English Jesuits in Liège. By 1678 Newton had had enough. Now immersed in alchemical experiments, he wanted no part in further controversy. Only after his death, he said, would his further writings be published.

Newton on the laws of motion and gravitation

Newton's study of the laws of impact began in the mid-1660s. From Descartes he accepted the law of inertia: bodies persist in their state of rest or motion unless acted upon by an exterior cause. Descartes' rules of impact, derived by *a priori* thinking, were contradicted by experiments; Newton by thought and experiment derived correct rules. The center of mass of two bodies must move uniformly whether they impact or not. Implicit here was the equality and oppositeness of action and reaction, and the constancy of the ratio of the inertias or masses determining how the two bodies exchanged or shared speeds. Unlike Descartes, Newton took account of the vectorial character of the motions. He verified his rules in both inelastic and elastic impacts.

In the same period, Newton quantified Descartes' concept of the 'endeavor to recede' of a body whirled in circular motion. The acceleration, which Newton like Descartes thought of as radially outward, was $v^2/r = 4\pi^2 r/T^2$, where v is the body's linear speed, r its distance from the center and T its period of circulation. In Streete's *Astronomia Carolina* (1661), Newton had met with Kepler's harmonic law for the planets, $T^2 \propto r^3$, and this substituted into the formula showed the 'endeavor to recede' of the planets to vary inversely as the squares of their mean solar distances.

Like most of his contemporaries, Newton during the 1660s and 1670s conceived the planets to be borne about in a Sun-centered vortex of aether, and the Moon to be borne in a smaller, Earth-centered vortex. To balance the endeavor to recede developed in these circulations, he imagined a subtler variety of aether rushing into the central body. This aether might serve, in the solar case, as fuel to conserve the Sun's shining, and in the terrestrial case, after chemical transformation within the Earth's body, as nutriment for living things. All things, he proposed, may be composed of different kinds of aether that transform into one another: 'Nature is a great circulatory worker'. In a paper sent to the Royal Society (though not for publication) in late 1675, Newton cited numerous phenomena in support of this aethereal philosophy.

In November 1679 Hooke as a secretary of the Royal Society sought to engage Newton in correspondence, and asked him, in particular, what he thought of Hooke's idea of 'compounding the celestial motions of the planets of a direct motion by the tangent & an attractive motion towards the central body'. This idea had emerged for CHRISTOPHER WREN and Hooke out of their reflections on GILBERT's magnetic philosophy. There is no clear evidence that Newton had seriously entertained it before.

In perhaps August 1684 HALLEY visited Newton in Cambridge to ask what path a planet would follow if

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

subject to an inverse-square force toward the Sun. By now the inverse-square variation of the solar force was accepted by Wren, Hooke and others as deducible from Huygens' formula for centrifugal force (published in 1673 and equivalent to Newton's unpublished formula) and the Keplerian rule $T^2 \propto r^3$. Newton reportedly answered 'an ellipse!', and promised to produce a demonstration. His response, sent to the Royal Society in November, was a small treatise *On the motion of bodies in gyres*, giving quantitative expression to Hooke's conception of orbital motion. A centripetal force, Newton showed, implies equable description of areas. Given the proportionality of time to area, he derived a formula for the planet's instantaneous radial acceleration. Applying this to an elliptical orbit with center of force in a focus, he showed that the force would be inverse square—the converse of the theorem he had promised. 'Therefore the major planets revolve in ellipses having a focus in the centre of the Sun…'. The logic was less than clear.

In the next two and a half years the small treatise expanded into the 510 pages of the *Principia*. But a critical threshold of substantiation had first to be crossed: 'Now I am upon this subject I would gladly know the bottom of it before I publish my papers' (Newton to FLAMSTEED, 12 January 1685). Flamsteed's data for the satellites of Jupiter helped: their periods and distances (measured by micrometer) were as exactly in accord with the $T^2 \propto r^3$ rule 'as it is possible for our senses to determine'. Taking the inverse-square law to be universally exact, Newton could compare the absolute attractive powers of the Sun and the planets having satellites (Jupiter, Saturn, Earth), by comparing the accelerations they would produce in satellites at a fixed distance. Identifying the attractive powers with the masses, Newton found that the center of mass or gravity of the system fell within the Sun or remained close to it, thus demonstrating the Copernican system *a priori*.

In another basic confirmation of his 'mathematical way', Newton carried out the earliest version of what is now called the 'Eötvös experiment'. In contrast to the changes of motion produced in impacts, gravitational accelerations appeared to be independent of the masses of the bodies accelerated; implying a proportionality of mass to gravitational force. Using pendulums with bobs of nine different materials (gold, silver, lead, glass, etc), Newton tested this proportionality, and found it to hold to within one part in a thousand. Hence aethereal mechanisms for gravity could be dismissed.

Newton's method demanded universalization of any rule found to hold in all known cases; results thus arrived at, however, were subject to correction or refutation from further experiments or observations. The law of universal gravitation had therefore to be tested unendingly. In an analysis of the three-body problem (*Principia* I.66) Newton derived qualitatively all the known inequalities of the Moon, and claimed to derive the tides and the precession of the equinoxes (the latter two derivations were flawed). Newton's procedure for fitting orbits to comets enabled Halley to predict the return of Halley's comet in 1759. In the late nineteenth century the anomalous precession of Mercury's perihelion was at length recognized as requiring modification of Newton's law; Einstein in 1915 accounted for the anomaly by his Theory of General Relativity.

Newton failed to generate a following of applied mathematicians able competently to pursue the testing of his theory. His mathematical methods, partly geometrical and partly algebraic, were too various and special to lend themselves to successful imitation. Wrongheadedly and fiercely Newton and Leibniz wasted energy charging each other with plagiarism in the invention of the calculus. The systematic attack on the THREE-BODY PROBLEM would come only after Newton's death, when EULER combined Leibnizian symbolic mathematics with the calculus of trigonometric functions that Newton had known. The resulting science, known today by LAPLACE's name for it, CELESTIAL MECHANICS, led to new mathematics and eventually to new understandings of the scope and limits of rational mechanics; but its agenda had been set by Newton's *Principia*.

Bibliography

Brackenbridge J B 1995 *The Key to Newton's Dynamics* (Berkeley, CA: University of California Press)

De Gandt F 1995 *Force and Geometry in Newton's Principia* transl. C Wilson (Princeton, NJ: Princeton University Press)

Densmore D 1995 *Newton's Principia: The Central Argument* (Santa Fe, NM: Green Lion)

Westfall R S 1980 *Never at Rest* (Cambridge: Cambridge University Press)

Wilson C 1989 The Newtonian achievement in astronomy *Planetary Astronomy from the Renaissance to the Rise of Astrophysics, Part A: Tycho Brahe to Newton* (Cambridge: Cambridge University Press) pp 233–74

Curtis Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Newton Mission

The European Space Agency's long-term plan for its science program is called 'Horizons 2000' and is designed to provide a balanced combination of satellite missions in astronomy, space science and fundamental physics. One of the cornerstones of this program is the XMM–Newton x-ray observatory, launched on 10 December 1999 from the European spaceport in French Guyana on the first commercial flight of the Ariane 5 rocket. In order to observe x-rays from astronomical objects it is necessary to be above the Earth's atmosphere since this absorbs incident x-rays. XMM–Newton was launched into a highly eccentric orbit, which takes the satellite once around the Earth every 48 h. The reason for this highly eccentric, long-duration, orbit is that it allows for long and uninterrupted observations (which are often required in x-ray astronomy) free of the radiation belts around the Earth. XMM–Newton cannot operate within the radiation belts since they disturb the measurements and can even damage some of the sensitive instrumentation (if left unprotected).

Originally, XMM–Newton was called the 'High-Throughput Imaging X-ray Spectroscopy Mission', which accurately describes its capabilities. 'High throughput' is provided by the three largest x-ray telescopes ever flown in space. Each telescope consists of 58 tightly nested grazing-incidence x-ray mirrors with a focal length of 7.5 m. Grazing-incidence optics are necessary because most materials simply absorb x-rays, rather than reflecting them, and only at very small incidence angles will ultrasmooth gold reflect x-rays. The mirrors focus the x-rays with a precision of about 15 arcsec which accounts for the 'imaging'. The 'spectroscopy' is provided by the instruments on board XMM–Newton:

EPIC — European Photon Imaging Camera (principal investigator, Dr M. Turner, Leicester University, UK);
RGS — Reflection Grating Spectrometer (principal investigator, Dr A. Brinkman, SRON, Utrecht, The Netherlands);
OM — Optical Monitor (principal investigator, Professor K. Mason, MSSL, UCL, UK).

The EPIC instrument provides the medium spectral resolution and simultaneous imaging capability through three cameras using two different technologies. The EPIC–MOS is geared towards lower energies, while EPIC–PN excels at unparalleled throughput and timing capabilities, especially at higher (5–10 keV) x-ray energies. The RGS is a dispersive instrument (much like a prism is for visible light), which provides medium- to high-resolution dispersive spectroscopy only. The concept of grazing incidence has once more been applied here in two arrays, each consisting of 182 grating plates, mounted behind two of the three x-ray telescopes. The co-aligned Optical Monitor provides the capability to observe the x-ray sources at visual and ultraviolet wavelengths, a concept that has proven its worth in the past. It is one of the great strengths of the XMM–Newton observatory that all of the above instruments work in parallel.

The first results from the XMM–Newton Observatory have seriously changed our understanding of many astrophysical sources of x-rays: from the extraction of energy from a black hole, through the distribution and state of the hot gas pervading clusters of galaxies to the accretion disks of material swirling around collapsed stars. It is clear that XMM–Newton is paving the way in x-ray imaging spectroscopy, so much so that astronomers are already seriously thinking about the potential successor which is called XEUS. This will be even larger than XMM–Newton, able to observe the first massive black holes as they were formed when the universe was only a small fraction of its current age.

Web update (31 July 2002)

New results from the XMM–Newton suggest a supernova was responsible for the gamma-ray burst that occurred on 11 December 2001. Observing the afterglow of the gamma-ray burst GRB 011211, XMM–Newton has produced the first ever evidence of the presence of chemical elements which were the unmistakable remnants of a supernova explosion which had occurred just a few days before. The observations revealed two important facts: first, the material in the source was moving quickly towards Earth, at a tenth percent of the speed of light; and second, chemical analysis of this material showed that it had to be the remnant of a supernova explosion.

Gamma-ray bursts are the most powerful explosions ever detected in the universe. So far there has been no conclusive evidence to prove what causes them. Until now, there have been two prime suspects for what makes gamma-ray bursts: the collision of neutron stars—dead, ultra-dense stars—or the death of very massive stars in supernova explosions.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Newton's Laws of Motion

Laws governing the motion of all bodies which were set out by Isaac Newton (1642–1727) in 1687. They form the basis of Newtonian mechanics. The three laws are as follows:

First law. Every body continues in a state of rest or uniform motion in a straight line unless acted upon by a force.

Second law. If a body is acted upon by an external force, it accelerates, the acceleration being directly proportional to the force and inversely proportional to the mass of the body; the acceleration takes place in the direction of the force. This may be expressed as the rate of change of *momentum* of a body is proportional to the applied force. This can be written as:

$$F = ma$$

where F is the applied force, m is the mass of the body and a the acceleration.

Third law. To every action there is an equal and opposite reaction; in other words, if a force acts on one body an equal and opposite force must act on another body. For example, a person standing on the Earth is pressing down on its surface with a force equal to his weight; the Earth's surface in resisting this force (so preventing the person from falling towards the center of the Earth) exerts a reaction, an upward force equal to that person's weight.

The essence of some of these laws was appreciated by other workers prior to and contemporary with Newton (e.g. Galilieo), but he was the first fully to appreciate and to formalize them. Newton's laws of motion express a completely different point of view from the earlier belief that force is necessary to maintain motion; in Newtonian theory force is only required to *change* the state of motion of a body.

See also: gravitation, weight.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Newtonian Telescope

A reflecting telescope in which the converging cone of light from a concave parabolic primary mirror is reflected to the side of the telescope tube by a small flat mirror set at an angle of 45° to the optical axis. The light is brought to a focus at the side of the tube, at the opposite end of the tube from the primary mirror, and this is where the eyepiece is located. Originally devised by Isaac Newton (1642–1727), who built the first working reflecting telescope to this design in 1668, this optical system is still widely used today in small reflecting telescopes, since the observer's head does not block incoming light. The design is less practical when the telescope focal length is large, say larger than a person's height.

Newton's first reflector had an aperture of 2.5 cm (1 inch) and its mirrors were made from speculum metal (an alloy of various metals, predominantly copper and tin) which, even when freshly polished, reflected no more than 60% of the incident light. With two such mirrors, a tiny aperture, and a magnification of about 25 times, it produced very faint images. Furthermore, because the curve of its primary mirror was spherical rather than parabolic, it suffered from an optical defect called spherical aberration.

See also: aperture, focal length, mirror, reflector, spherical aberration, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Next Generation Space Telescope

The Next Generation Space Telescope (NGST) will be an 8 m class deployable, radiatively cooled telescope, optimized for the 1–5 μm band, with zodiacal background limited sensitivity from 0.6 to 10 μm or longer, operating for 10 yr near the Earth–Sun second LAGRANGIAN POINT (L2). It will be a general-purpose observatory, operated by the SPACE TELESCOPE SCIENCE INSTITUTE (STScI) for competitively selected observers from the international astronomy community. NASA, the European Space Agency (ESA) and the Canadian Space Agency (CSA) will build NGST, with construction to start in 2004 and a planned launch of 2009.

NGST will be able to observe the first generations of stars and galaxies, including individual starburst regions, protogalactic fragments and supernovae out to REDSHIFTS of z = 5–20. NGST will resolve individual stars in nearby galaxies, penetrate dust clouds around local star-forming regions and discover thousands of isolated substellar and KUIPER BELT objects.

NGST science goals

The primary science goals for NGST have been generated by the astronomical community through reports to NASA headquarters and science working group studies. Although NGST will be a general-purpose observatory capable of a variety of investigations, several types of observational problems are being examined in detail to help design the mission. The following sections highlight these programs.

The origin and evolution of galaxies

Motivated by the spectacular success of the HUBBLE SPACE TELESCOPE (HST) in pioneering the exploration of high-redshift (z > 1) galaxies and clusters, the HST & Beyond Committee foresaw the enormous potential of a scientific successor to HST, optimized for the near infrared (NIR, 1–5 μm). Such an instrument would be especially powerful in studying the origin and evolution of galaxies. In 2008, NGST will be poised to complete our understanding of the formation and early evolution of galaxies such as the Milky Way by doing the following.

- The earliest phases of star and galaxy formation—the end of the 'dark ages'—will be detected (figure 1). This requires superb NIR sensitivity (> 1 nJy (1.0 × 10^{-35} W m^{-2} Hz^{-1}), 1–5 μm) in deep-broadband imaging (in a ~10^5 s exposure).
- The first galactic substructures larger than individual star clusters (whose diameters are typically ~300 pc (~9 × 10^{16} km) for 0.5 < z < 5.0) will be resolved. This requires HST-like resolution in the NIR (~0.060 arcsec at 2 μm).
- The fundamental properties of individual galaxies will be quantitatively measured. This will be enabled by emission-line and absorption-line spectroscopy, with broad spectral coverage and low-to-moderate spectral resolution ($R = \lambda/\delta\lambda$): R ~ 300 (0.6–5.0 μm) for redshift confirmation, cluster membership and ages of stellar populations; R ~ 1000 (0.6–5.0 μm or longer) for star formation rates, metallicity and reddening; R ~ 3000 (1.0–10 μm) for dynamics (mass).
- Dust-enshrouded regions hiding massive star formation or active galactic nuclei during the epoch of greatest star formation to a minimum of z~2 will be detected and diagnosed.

The structure and chemical enrichment of the universe

The geometry and structure of the universe, as well as its history of element formation, are intimately related to the formation of galaxies. Future microwave telescope missions will measure the power spectrum of the COSMIC MICROWAVE BACKGROUND (CMB) at z ~ 1300 and, using standard models, will provide or constrain important cosmological constants. NGST will play a powerful complementary role in determining the distribution of mass and light on small scales. Large MICROLENSING imaging surveys will use the wide field, superb angular resolution and excellent 0.6–5.0 μm sensitivity of NGST to measure the mass structure of the universe at z = 1–5 on scales smaller than those probed by CMB measurements from space or possible from the ground or HST.

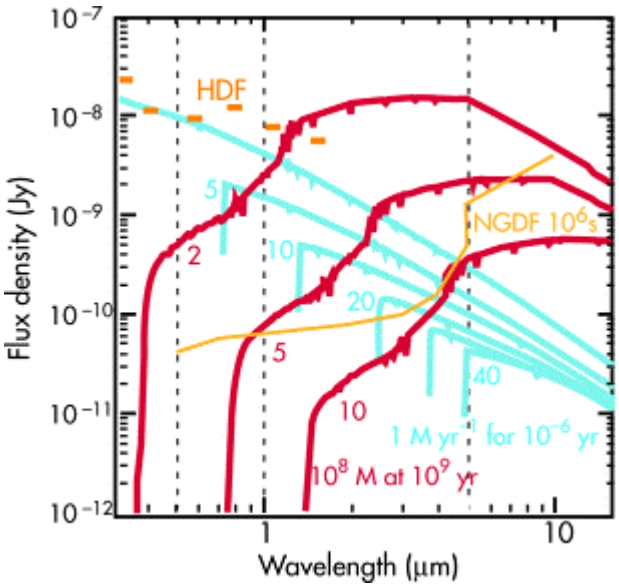


Figure 1. The sensitivity of an NGST deep field (10^6 s in 30% bandwidths, 10σ detection). Also indicated are the spectra of starburst regions (10^6 solar masses (2 × 10^{36} kg) in 10^6 yr) and established populations (10^8 solar masses at 1 Gyr age) at various redshifts (Ω_m = 0.2). Comparable sensitivities also are shown for the Hubble Deep Field (HDF) using NICMOS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Expected science programs include the following.

- The dark matter halos of galaxies to redshifts of z~5 will be weighed statistically by deep imaging of selected fields.
- The growth of galaxy clusters to redshifts of $z \sim 1$–3 will be measured using multicolor deep imaging of selected high-redshift clusters and protoclusters discovered in other surveys.
- The statistical properties of the distribution of matter on scales of 1–10 Mpc (~$3 \times 10^{19–20}$ km) can be found from wide-area, high-resolution NGST imaging surveys (>1 deg^2).

These imaging programs are comparable in depth and required field of view with those used for the study of GALAXY EVOLUTION. Such surveys also provide an excellent method for discovering type Ia and type II SUPERNOVAE at redshifts in the range $1 < z < 5$. Measuring the rates and galactic associations of type Ia and type II supernovae will provide an independent assessment of the history of element production. We expect that NGST will be crucial in extending the observations of type Ia supernovae beyond z~0.9 to z~5. Only at the higher redshifts is it possible to distinguish between the behavior of type Ia supernovae with cosmologies involving only H_0, Ω_m and Γ and models with significant supernova evolution or smoothly distributed gray obscuration. Such data will provide measurements of the cosmological parameters, which are independent of and complementary to those derived from the CMB missions.

The processes of star and planet formation

The potential studies in this arena are essentially limitless and depend crucially on the available spectral resolutions and mid-infrared (MIR; defined here as ~5 μm < λ < ~30 μm) wavelength coverage. Examples include the following.

- The infall and outflow processes through which stars are built and their final masses determined will be characterized. MIR spectroscopy will diagnose the accretion shocks in protostellar systems, while NIR imaging will reveal outflow shocks and jets near their source, with a resolution of ~2 astronomical units (~300 million kilometers).
- The structure and evolution of circumstellar material will be traced, from the massive envelopes of class 0 protostars to the protoplanetary disks of pre-main-sequence stars, and finally to the dissipation of these disks into mature debris disks of main sequence stars. NIR and MIR spectroscopy of gas and dust features, their excitation and their radial variation within the circumstellar region will permit study of the growth of dust grains toward planetesimals, the chemical processing of disk gas and the disk dissipation mechanisms that define the time available for planet formation.
- Substellar objects will be detected and characterized. Ground-based sky surveys and adaptive optics programs are now beginning to discover significant numbers of isolated and companion brown dwarf stars. However, these observations will be limited to the bright (high mass–low age) end of the substellar luminosity function and to wide binary companions. NGST will have the needed combination of high angular resolution, high sensitivity and a stable PSF for high-contrast imaging of faint substellar companions in planetary orbits. Observations at 5 μm with a graded-mask coronagraph would be able to directly detect planets with Jupiter's mass, age and orbital semimajor axis in more than 90% of the single stars within 8 pc (2.4×10^{14} km) of the Sun (>50 systems). By detecting planetary photons directly, NGST could provide the first opportunity to spectrally characterize exoplanet atmospheres.

NGST mission concept

The science goals for NGST require a telescope with high sensitivity covering the wavelength range from 0.6 to 10 μm with capability out to 28 μm, and with NIR angular resolution comparable with that of HST at visible wavelengths. Industry and NASA have studied several mission architectures. While they differ in details the concepts share an overall similar design. For simplicity, the NASA version, referred to as the Reference Architecture, is discussed here. Figure 2 shows the observatory and its main components: the Optical Telescope Assembly (OTA), the Integrated Science Instruments Module (ISIM) and the Spacecraft Support Module (SSM). Details for the Reference Architecture may be found in the book *Visiting a Time When Galaxies Were Young*.

The optical telescope assembly

The Reference Architecture optical configuration is a three-mirror anastigmat that provides a real, accessible pupil and permits a relatively fast primary mirror to minimize telescope length. This design provides excellent imaging over a field of more than 20 arcmin with achievable alignment tolerances. A real pupil permits the use of a deformable mirror (DM) for wavefront correction and a fast-steering mirror for fine pointing using image compensation. The primary mirror is a compact 8 m diameter segmented aperture. It is composed of a central mirror segment, with a diameter of 3.3 m, surrounded by eight petals.

The Reference Architecture mirror is made of beryllium, thermally controlled with very-low-power heaters so that its figure remains insensitive to rapid or

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

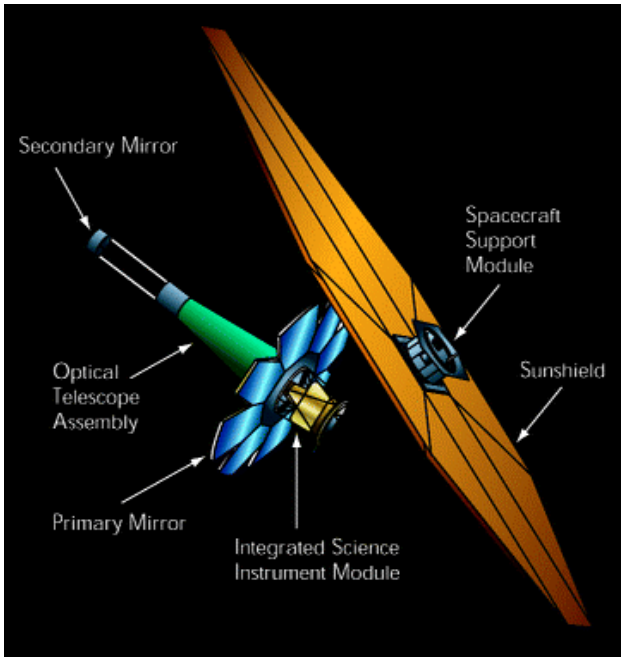


Figure 2. The elements of the Reference Architecture concept.

large positioning slews. The areal density of the primary mirror assembly (mirror, actuators and backup structure) is 13 kg m^{-2}. This represents over an order of magnitude improvement over the HST mirror areal density. The DM provides a design margin for figure errors in the primary mirror, including those due to gravity release, thermal gradients or edge effects.

Unlike telescopes such as HST that are launched fully assembled, NGST must be able to compensate for errors in deployment position, long-term dimensional changes and on-orbit thermal variations. Optics are aligned and phased by observing the image of a star and deriving mirror position corrections. Wavefront errors are determined by obtaining defocused star images and analyzing the image with a 'phase retrieval' computer algorithm.

The Reference Architecture Integrated Science Instrument Module

The ISIM consists of a cryogenic instrument module integrated with the OTA, and processors, software and other electronics located in the Spacecraft Support Module. The ISIM provides the structure, environment and data handling for several science instruments.

This ISIM design is illustrative and is not intended to define NGST's final complement of instruments. The Reference Architecture instrument suite includes the following.

- An NIR camera covers 0.6–5 μm, critically sampled at 2 μm. Efficient surveying capability, as well as guiding requirements, set the field at about 4 arcmin × 4 arcmin, apportioned over four subcameras each 2 arcmin on a side. The NIR detectors (InSb or thinned HgCdTe) are radiatively cooled to 30 K.
- An NIR multi-object spectrometer, with spectral resolutions of 300 and 3000 and a spatial resolution of 100 marcsec, covers a field of 3 arcmin × 3 arcmin. Multi-object capability is enabled by an array of 2048^2 micromirrors used to form a reflective slit mask, directing light into or away from the spectrometer.
- An MIR camera–spectrometer covers a field of 2 arcmin × 2 arcmin with a spectral range of 5–28 μm using a 1 K × 1 K Si:As array as detector and a long-slit cross-dispersed grism. Its spectral resolution is ~10^3. The camera employs a selection of slits and a no-slit option to enable direct imaging with filters. The MIR detector is cooled to 6 K by a miniaturized reverse turbo-Brayton cooler; open-cycle solid hydrogen cooling has been identified as a viable alternative.

Passive cooling and thermal control

NGST designs solve the problem of cooling to the cryogenic temperatures required for NIR and MIR operation passively by (a) protecting the observatory from the Sun with a multilayer shield, (b) using a heliocentric orbit to decrease the Earth's thermal input and (c) configuring the telescope to have a large area exposed to space to improve radiative cooling.

To make the thermal emissivity negligible compared with the zodiacal light, the back of the sunshield must be below 100 K. This is accomplished by adding five low-emissivity layers behind the surface of the shield facing the Sun. The main optics then reach very low temperatures (<40 K) and do not contribute significantly to the overall emissivity of the observatory.

Sky coverage

The observatory can be pitched ±25º off the sunline and rolled 360º about the sunline. The portion of the sky accessible is a 50º wide spherical band centered 20º away from the perpendicular to the sunline. This represents ~40% of the entire sky. Full sky coverage is achieved in slightly less than 6 months. At any time, any target in the accessible zone can be tracked for a minimum of 7 weeks.

Launch and orbit

The overall mass of the Reference Observatory NGST is approximately 3300 kg, within the capability of an Atlas IIAS or the next generation of medium launchers (EELV Medium). Deployment of the OTA occurs soon after launch, before the sunshield is deployed, while all the mechanisms are still relatively warm. Optics alignment can then begin, followed by science calibration as the telescope cools. The halo orbit at L2 is reached about 3 months later.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Science and mission operations

NGST's Science and Mission Operations will be simple and efficient because of the telescope's location at L2. Operations for NGST will be based on an optimized long-range observing plan that consists of sequenced science programs ranging from large imaging surveys, often with pre-planned spectroscopic follow-ups, to intensive studies of individual objects.

Discovery potential

NGST will be an observatory with enormous discovery potential both at 0.6–10 μm and at longer wavelengths. This is illustrated in figure 3, which shows the time to achieve a broadband, high-resolution wide-field image with a variety of facilities: NGST, HST, GEMINI OBSERVATORY (representing IR-optimized ground-based 8 m telescopes) and SPACE INFRARED TELESCOPE FACILITY (SIRTF). NGST enjoys a significant background advantage over the ground at all wavelengths, a larger field of view over which high-resolution images can be obtained (4 arcmin × 4 arcmin assumed) and a significant aperture advantage over SIRTF. The shorter times required to achieve a given threshold sensitivity can translate into larger fields observed (more targets) or greater sensitivities.

At wavelengths between 2.2 and 10 μm, NGST is many orders of magnitude faster than any other planned facility. In practice, it will have sensitivities several orders of magnitude better than even SIRTF because of source confusion in very deep observations. In this regard, SIRTF is a superb instrument for surveys of 10–100 square degrees. NGST is best used for deeper observations of smaller pre-selected fields. At wavelengths longer than 10 μm, NGST will surpass SIRTF by an order of magnitude in sensitivity despite a much higher MIR background. For NIR and MIR spectroscopy, it will be unsurpassed between 0.9 and 28 μm (28 μm is the cutoff for Si:As detectors) because NGST's thermal background is much reduced at high spectral resolution.

The 'uniqueness space' for NGST is defined relative to the many large ground-based telescopes that will be operating by 2009. The criterion is that NGST be at least 100 times faster than an optimally operated and equipped 8 m telescope with laser guide stars and adaptive optics. In the NIR, the NGST imaging uniqueness space is very large. For moderate spectral resolution 2D spectroscopy the effects of detector noise significantly narrow the uniqueness space. At this resolution and detector dark current, ~0.02 e^- pixel s^{-1}, NGST is no longer background limited in the visible or NIR (< 4 μm). For spectroscopy of single faint targets, for instance, NGST would be used primarily in the NIR (>0.9 μm). For high-resolution spectroscopy in the visible and J(1.25 μm) and K(2.2 μm) bands (R > 5000) or for visible imaging of large fields with 0.4 arcsec resolution, large ground-based telescopes are competitive with NGST. This is where the large telescopes planned for the next decade will make major contributions.

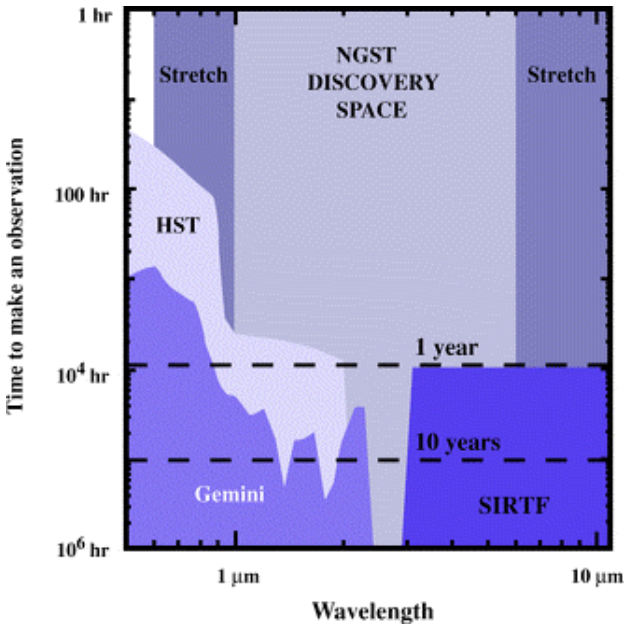


Figure 3. NGST Discovery Space: the relative speed of broadband NGST high-resolution, wide-field imaging compared with other observatories (HST, ACS, NICMOS, Gemini and SIRTF).

It is instructive to consider the relative power of NGST compared with that of HST. HST is currently unique in the windowed ultraviolet (0.1–0.3 μm, a factor of 3 in wavelength). It has comparable sensitivity but superior resolution over ground-based telescopes in the visible and NIR (0.3–1.8 μm, a factor of 6). NGST will have 10–100 times more imaging sensitivity than HST and superior resolution at 2.5 μm (a factor of 4 or comparable with HST at 0.6 μm). It will be unique in imaging and spectroscopy from 2.5 to 28 μm (a factor of 10 in wavelength). It is clear from this simple analysis that NGST will have at least as great an impact on astronomy as HST.

Bibliography

Dressler A (ed) 1996 HST and Beyond *Exploration and the Search for Origins: a Vision for Ultraviolet–Optical–Infrared Space Astronomy* (Association of Universities for Research in Astronomy)

Stockman H S (ed) 1997 Next Generation Space Telescope *Visiting a Time When Galaxies Were Young* (Association of Universities for Research in Astronomy)

Eric Smith

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NGC 1068: Prototypical Seyfert 2 Galaxy

NGC 1068 is a nearby spiral galaxy which harbors the prototypical example of a Seyfert type 2 nucleus. However, all evidence points to the conclusion that its active nucleus is completely hidden to our view by dense obscuring material. If we could see NGC 1068 along an unobscured line of sight it would appear radically different. This unveils the importance of orientation of the observable properties of ACTIVE GALACTIC NUCLEI. The extended emission line region in NGC 1068 bears a strong morphological connection with the radio emitting plasma. This close association suggests that nuclear outflows play a fundamental role in the origin of narrow line regions in SEYFERT GALAXIES.

Basic data

NGC 1068 is an early type, BARRED SPIRAL GALAXY located in the Cetus–Aries group. It lies very close to the celestial equator and this makes it visible from both hemispheres (RA $02^h42^m\ 40.711^s$ DEC-$00^d00^m47.81^s$, J2000). Its recessional velocity is 1148 ± 5 km s^{-1} which corresponds, after correction for the Virgocentric infall and adopting $H_0 = 75$ km s^{-1} Mpc^{-1}, to a distance of 14.4 Mpc. At this distance an angular separation of $1''$ translates to 72 pc. NGC 1068 is very bright over the whole electromagnetic spectrum. For this reason, and for its many peculiarities, it belongs to several catalogs of galaxies. Table 1 lists the most common aliases for NGC 1068.

Table 1. NGC 1068 aliases

Messier 77	UGC 2188
ARP 37	IRAS 02401-0013
3C 71	PKS 0240-00

The optical appearance of NGC 1068 is dominated by a 13 kpc oval bulge elongated on a NS orientation from which originate two spiral arms (see figure 1). The most vigorous star formation, however, occurs in a ring-like region with a diameter of ∼2 kpc (see figures 2(*a*) and 2(*b*)). On a smaller scale HUBBLE SPACE TELESCOPE (HST) observations shown a bright compact knot which stands out from the diffuse emission of the central regions of the stellar bulge (see figure 2(*c*)). Active star formation is often associated with nuclear activity in Seyfert galaxies (see the article on STARBURST GALAXIES). Several absorption features form a filamentary spiral-like structure around the nucleus and a dust lane (extending over $\sim 180 \times 35$ pc) is projected just south of the brightest knot.

Infrared images reveal a quite different morphology. In particular they are dominated by the presence of a conspicuous stellar bar which extends over ∼3 kpc along position angle $\sim 50^\circ$. Furthermore, a bright central source is also seen at these longer wavelengths but, as will be discussed later, it is not coincident with the optical knot.

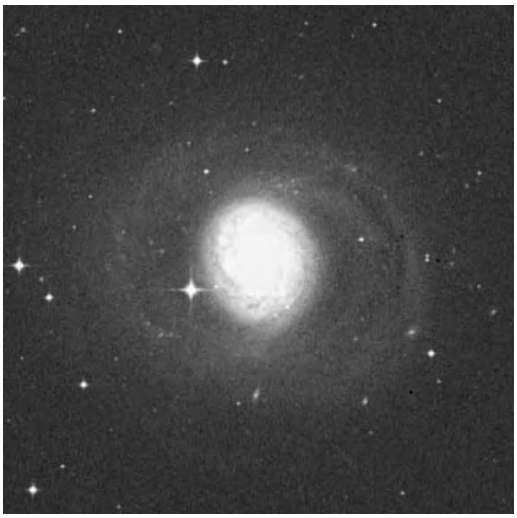


Figure 1. Digital Sky Survey optical image of NGC 1068. The field of view is 10×10 arcmin.

NGC 1068 hosts one of the most studied active galactic nuclei (AGN) in the sky. It is the first object listed in the original discovery paper by SEYFERT containing galaxies with strong ultraviolet excess and in this sense it represents the prototypical Seyfert galaxy. Its optical spectrum is characterized by prominent emission lines. Since permitted and forbidden lines share a similar width (∼1000 km s^{-1}) NGC 1068 is classified as a Seyfert type 2 galaxy (as opposed to Seyfert type I in which permitted lines are broader than the forbidden ones). Ground-based studies have shown that the region where the narrow emission lines are produced is extended and covers several arcsec north of the nucleus. Deep spectroscopy and imaging also revealed a much fainter and more extended (50 arcsec) emission line region with a well defined conical morphology.

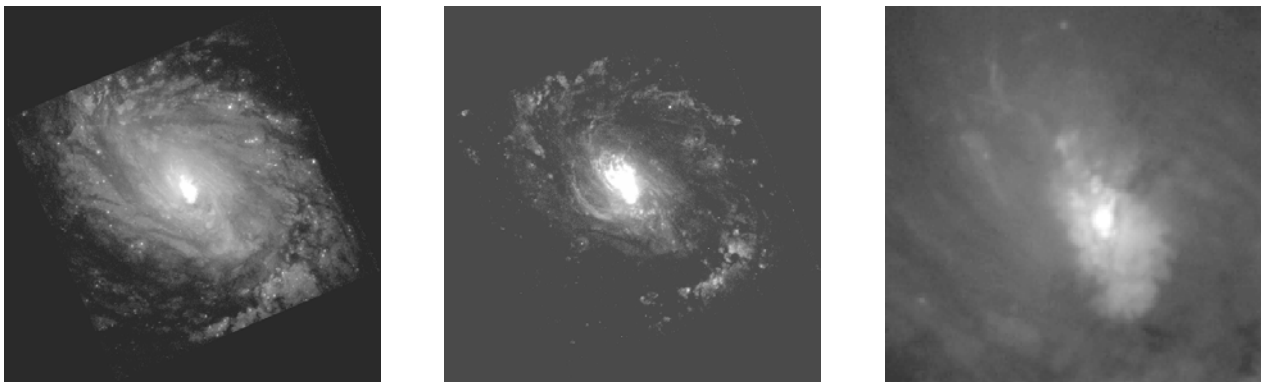
A bright radio source (16 Jy at 178 MHz) is also associated with NGC 1068. The radio emission is dominated by a central component which, in high resolution images, reveals a jet-like morphology originating from the nucleus and extending towards the north. Two radio lobes form a bipolar structure which extends over $\sim 15''$.

NGC 1068 harbors a hidden active nucleus

Since the discovery of significant optical POLARIZATION in Seyfert nuclei the study of their polarized flux has proved to be a powerful tool for probing the properties and the geometry of their inner regions.

Three astrophysical processes are known to produce polarized emission: synchrotron radiation, transmission of light through aligned dust particles, and scattering. However, in Seyfert galaxies, synchrotron radiation at optical wavelength is negligible and dichroic transmission has a relatively low efficiency and thus cannot account for the high degree of polarization observed (as high as 60%) in NGC 1068. Conversely, when a photon is scattered

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*) (*b*) (*c*)

Figure 2. NGC 1068 as seen by HST: (*a*) Visual continuum, (*b*) Hα emission line image showing the 2 kpc ring of star formation, (*c*) innermost nuclear regions of NGC 1068. Fields of view are 50 arcsec for (*a*) and (*b*) and 15 arcsec for (*c*).

it acquires a net linear polarization which, for a right angle scattering, can be as high as 100%. This indicates that scattering is the dominant polarization mechanism in Seyfert galaxies. The degree of polarization in NGC 1068, after correction for the starlight contribution, is essentially constant ($\sim$16%) over the UV and optical wavelength range. This strongly argues that electrons, whose scattering cross section is independent of wavelength, represent the dominant scattering medium, at least in the innermost nuclear regions.

Polarization observations represent a unique tool to distinguish scattered light from light produced *in situ*. In particular, optical spectropolarimetry revealed that the spectrum of the *polarized* light of NGC 1068 is radically different from that observed in traditional spectroscopic observations. The polarized spectrum of NGC 1068 closely reproduces the spectrum of a type I Seyfert galaxies with the presence of broad permitted hydrogen and iron lines. The simplest interpretation is that NGC 1068 harbors a Seyfert I nucleus which is hidden to our view by obscuring material, probably arranged in a torus or on a highly warped disk. The minimum vertical height of the torus is set by the requirement that the whole region in which the broad lines are produced must be covered, $\sim$0.1–1 pc. Light produced from the nucleus can reach us only after it has been scattered into our line of sight and is therefore polarized.

Imaging polarimetry has been particularly useful to confirm this scenario (see POLARIMETRY: IMAGING POLARIMETRY). From two-dimensional mapping of polarized light it is possible to derive the location of the illuminating sources and of the scattering medium. In fact, the plane of polarization is perpendicular to the direction of the last free path before the scattering and this enable us to trace this path back to the source of illumination. The polarization vectors are thus expected to be tangential to circles drawn around the nucleus and to form a centro-symmetric pattern. This behavior has been indeed clearly observed in NGC 1068 which, in this sense, can be described as a giant REFLECTION NEBULA. By modeling such a pattern it is possible to accurately locate the hidden nucleus as the center of the polarization circles. In agreement with the proposed unified model, there appears to be no visible counterpart at this location since the nucleus is not seen directly. Conversely, the hidden nucleus lies at the center of the nuclear dust lane which probably mark the location of the obscuring torus.

The distribution of polarized light also provides evidence that the nuclear radiation field is anisotropic, as expected if the torus intercepts the nuclear light over a substantial solid angle. Indeed, the polarized emission is concentrated within a region which cover an angle of $\sim 70^\circ$ as seen from the nucleus. This is reminiscent of the conical distribution of the large-scale ionized gas. The torus is 'shading' a large portion of the galaxy from the illumination of the nuclear radiation field. While in the optical the scattering nebula and the emission line region is strongly asymmetric and essentially only the northeastern side is visible (as well as the extended line emission), at IR wavelengths they are clearly seen also on the southwestern quadrant.

The strongest direct evidence in favor of the presence of dense obscuring nuclear material comes from high resolution interferometric radio observations. When looked at with a resolution of $0.002''$($\sim$1.5 pc) the radio component closest to the hidden nucleus shows an elongated morphology, approximately 20 mas in size, oriented *perpendicularly* to the radio jet. Its radio spectrum is well accounted for as thermal emission from hot ionized gas at a temperature of 10^6 degrees. However, water maser emission has been detected associated to the same physical region and this requires the presence of warm ($\sim$400 K) high density ($\sim 10^8$–10^{10}) gas. These two results, combined, point to the idea that we are indeed seeing the inner portion of an edge-on dense disk or torus.

Is the NGC 1068 active nucleus hidden at all wavelengths?
As already discussed the nucleus is completely obscured from ultraviolet to optical wavelengths. The region of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

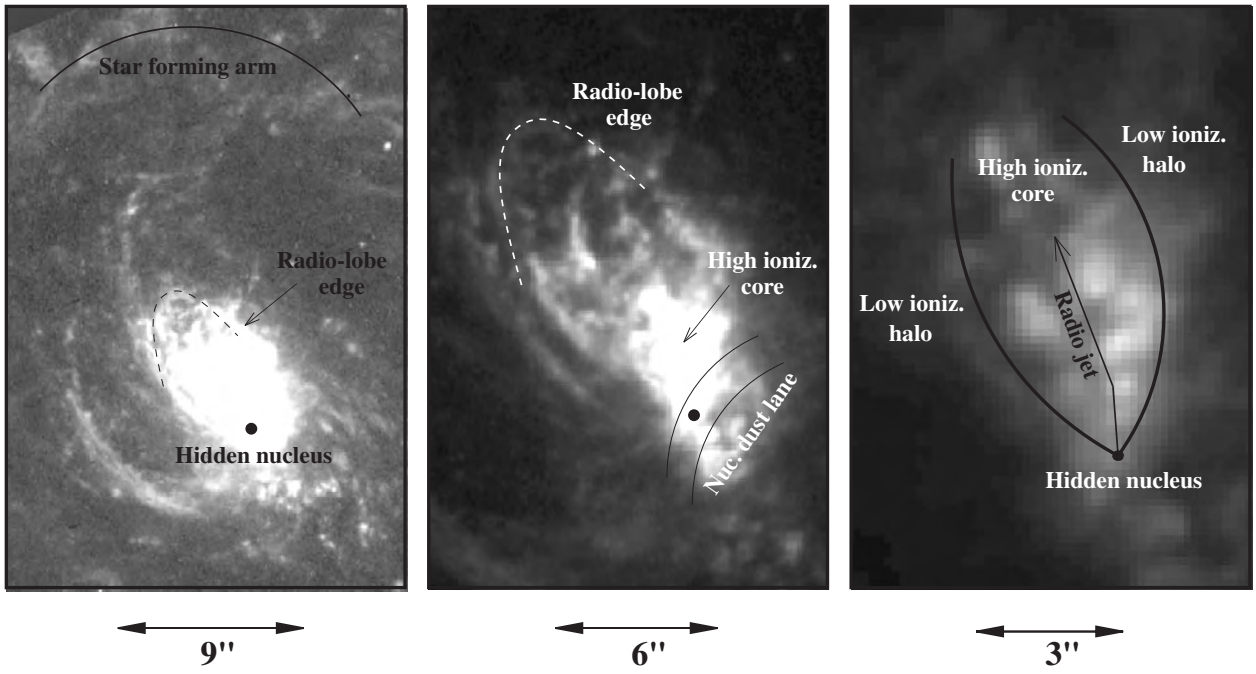


Figure 3. Schematic representation of the various components of the NLR and other features of interest in NGC 1068 overlaid on Hα+[N II] HST images. The black circle indicates the location of the hidden nucleus.

brightest optical emission is located ∼20 pc north of the hidden nucleus as determined by imaging polarimetry and radio observations. In the near-infrared, where dust extinction is less severe, there is a compact central source (smaller than 2 pc) coincident to the location of the nucleus. However, this emission is produced by the hot dust located in the innermost regions of the obscuring torus rather than by the nucleus itself. This is confirmed by the lack of detection of infrared permitted broad lines which, if the nucleus was seen directly at these wavelengths, should be associated with the nuclear emission as in the case of Seyfert I galaxies. Similarly, radio observations of the nuclear region do not show the presence of radio continuum emission near the center of the torus.

X-ray observations seem to represent our best chance to see the nucleus directly. In fact, the amount of material needed to absorb the x-ray emission is much larger that in the optical band, particularly in the hard x-ray band (∼2–10 keV) where a gas column density of the order of $N_H \sim 10^{25}$ atom cm^{-2} is needed. As a comparison, in the optical band a column density of $\sim 5 \times 10^{21}$ atom cm^{-2} produces an absorption depth of 1 and typical galactic column densities are $\sim 10^{21}$ atom cm^{-2}. Indeed, NGC 1068 is a strong x-ray emitter and its hard x-ray spectrum is similar to those of Seyfert I with one important exception: no variability is seen over a timescale of 10 yr. Another puzzling result is that the x-ray spectrum does not show the effects of significant absorption; if we were seeing the x-ray nucleus, these high-energy photons should cross the same material which is responsible for the obscuration in the optical and infrared bands, and thus produce easily detectable absorption features. These results can be understood if the x-ray nuclear emission is not seen directly but is scattered, similar to what is seen in the optical. In this picture the scattering region is not affected by absorption from the nuclear torus and it is extended so that the intrinsic variability of the incident nuclear continuum is smoothed out. In support of this interpretation the hard x-ray spectrum of NGC 1068, and in particular the prominent iron emission lines, can be successfully modeled by scattering over a combination of neutral and ionized reflectors. Recent observations moved the limit in photon energy at which only the scattered component can be seen, to ∼100 keV, consequently increasing the minimum column density of the torus to $\sim 10^{26}$ atom cm^{-2}.

We can conclude that we have no direct view of the nucleus of NGC 1068 but only through a small fraction of nuclear light which is scattered or reprocessed into our line of sight. Interestingly, the detection of scattered light implies that light *can* escape the nuclear regions, although not in our direction. If we could see NGC 1068 from these unobscured lines of sight it will probably look like an ordinary Seyfert I galaxy.

The key relevance of these obserations of NGC 1068 is that they clearly show that the properties of an active galaxy depend critically on its orientation. This prompted the development of a unified model for Seyfert galaxies in which the two Seyfert classes are intrinsically identical and differ only for their orientation with respect to the observer and this scheme is now proving to be widely applicable

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

also to other classes of AGN.

The origin of the emission line region in NGC 1068

The large-scale distribution of line-emitting gas in NGC 1068 shows a sharp conical morphology, clearly indicative of illumination from an anisotropic source of ionizing photons. Recent HST imaging shown that such a cone-like morphology is also seen on a smaller scale, ∼4 arcsec, but the axis of symmetry of this inner 'cone' is rotated some 30° clockwise from the outer one. The changing orientation with distance suggests that the gas does not fill uniformly the region illuminated by the nucleus.

In this respect, particularly intriguing is the relationship between the emission-line region and the radio emission (see figure 3). The inner radio jet lies in a region of relatively low optical emission and is surrounded by most prominent line-emitting clouds. Similarly, on a larger scale, a network of line-emission filaments completely envelop the northeastern radio-lobe. These results can be understood as due to the interaction between the outflowing radio plasma and the surrounding medium: in this scenario the outflowing plasma is sweeping, compressing and heating the interstellar gas. As a result of this interaction and compression the gas takes the form of dense shells and the line emission is highly enhanced along the edges of the radio-jet and lobes.

Strong support to this scenario comes from the velocity field of the narrow line region (NLR). Within 0.5 arcsec of the radio-jet the emission lines are kinematically disturbed and split into two components whose velocity separation is 1500 km s^{-1}, while the filaments associated with the radio lobe are redshifted by 300 km s^{-1}. This is clearly a consequence of the acceleration of the ionized gas driven by the expansion of the radio plasma.

The radio outflow might also be important for the ionization properties of the gas. In fact, the material along the radio-jet is in a much higher ionization state than the remaining of the NLR. Because the electron density is substantially larger in this region compared to the surrounding gas, these results can only be explained if there is a more intense ionizing continuum associated with the jet. This can be accomplished in a variety of ways which include an intrinsically anisotropic nuclear radiation field but also the presence of a local ionization source such as thermal emission from hot gas heated by the interaction with the radio-jet.

It therefore appears that the morphology of line-emitting region in NGC 1068 is dominated by the presence of a radio outflow, which sets the distribution of ionized gas, rather than to the anisotropy of the nuclear illuminating source and it may also play a crucial role in determining its ionization properties.

Bibliography

In 1996, a four-day workshop was dedicated to the discussion of the latest results on NGC 1068. The proceedings of this meeting, published in 1997 in the 248th volume of *Astrophysics and Space Science*, provide a valuable source of additional information and further reading on this galaxy.

A Capetti

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NGC 4151

NGC 4151 is one of the nearest and brightest active galaxies, being among the original six 'Seyfert' galaxies singled out by Carl Seyfert in 1947 for the unusually strong, broad emission lines in their spectra (see SEYFERT GALAXIES). NGC 4151's rather anonymous designation as the 4151st entry in Dreyer's 1888 *New General Catalogue* belies its prominence in modern astronomical studies. Its proximity and brightness make possible detailed observations with good spatial resolution using a wide variety of techniques across the entire electromagnetic spectrum from the radio through the gamma ray. These observations have provided valuable insights into the general workings of all active galaxies (see ACTIVE GALAXIES: OVERVIEW, QUASISTELLAR OBJECTS: OVERVIEW and ACTIVE GALAXIES: OBSERVATIONS).

NGC 4151 is located in the constellation Canes Venatici (the Hunting Dogs), and it is visible through a modest-size telescope in the northern spring sky. As shown in figure 1, it is a spiral galaxy with a moderately bright bulge (Sab on the Hubble classification sequence). The most prominent feature of its visual appearance, however, is the bright, point-like nucleus that is the center of activity. The whole galaxy has an integrated visual magnitude of 10.8, and 35% of this light is emitted by the nucleus itself.

Modern surveys have catalogued thousands of Seyfert galaxies, and astronomers have subdivided these unusual galaxies into two major classes. The nuclei of Seyfert 1 galaxies display unusually broad emission lines as well as some narrower lines. (These 'narrow' lines are typically much broader than the emission lines seen in normal galaxies or Galactic nebulae.) In Seyfert 2 galaxies, all the emission lines have similar widths comparable to the narrow lines in Seyfert 1s. The nuclear spectrum of NGC 4151 displays both broad lines and unusually strong narrow lines, and occasionally the broad lines are so faint that they are difficult to see. It is therefore often classified in an intermediate category, 'Seyfert 1.5'.

Owing to the expansion of the universe, NGC 4151 is receding from us at a velocity (relative to the Sun) of $cz = 995$ km s^{-1}, as measured by the neutral hydrogen gas that fills its galactic disk. For a Hubble constant of 65 km s^{-1} Mpc^{-1}, this places it at a distance of 15 Mpc, or 50 million light years. At this distance, features 0.1 arc seconds across, capable of being resolved by the Hubble Space Telescope (HST), span a physical distance of 7.4 pc, or 24 light years.

Visual appearance

NGC 4151 has a visual size of 4.5 × 6.3 arc minutes, bounded by a set of outer spiral arms that merge into a faint ring. The plane of the spiral disk is inclined to our line of sight by 21°, and the major axis of the disk is oriented at a position angle of 26°. As is usual for active galaxies, the nuclear light is bluer than that of the surrounding galaxy. The galaxy has an integrated color of $B - V = 0.73$, while the nucleus has $B - V = 0.58$. Spectra of the nucleus reveal broad emission lines and a non-stellar, power law continuum, but, depending on the size of the aperture used, galactic starlight is seen in varying degrees as well.

Figure 1. Visible light image of NGC 4151 from the Digital Sky Survey. The image covers an area of 160 × 160 square arc seconds on the sky.

Figure 2. Emission-line image of NGC 4151 in the light of [O III] obtained with WFPC2 on HST. The image size is 15 × 12 square arc seconds.

HST images of NGC 4151 resolve the narrow-line emission in the nuclear region on the sub-arc second scale. As shown in figure 2, its morphology is biconical, similar to that seen in Seyfert 2 galaxies (like the prototype NGC 1068). The cones have an opening angle of 75°, and they are oriented at a position angle of 60° on the sky. Long-slit spectroscopy of this emission using the Space Telescope Imaging Spectrograph (STIS) on HST shows that the gas is outflowing from the nuclear region. Images with ground-based telescopes trace the narrow-line emission in filaments extending up to 30 arc seconds from the nucleus. This is often called the 'extended narrow-line region.'

Ultraviolet properties

At shorter ultraviolet wavelengths, the starlight of the galaxy proper in NGC 4151 nearly disappears, and its

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

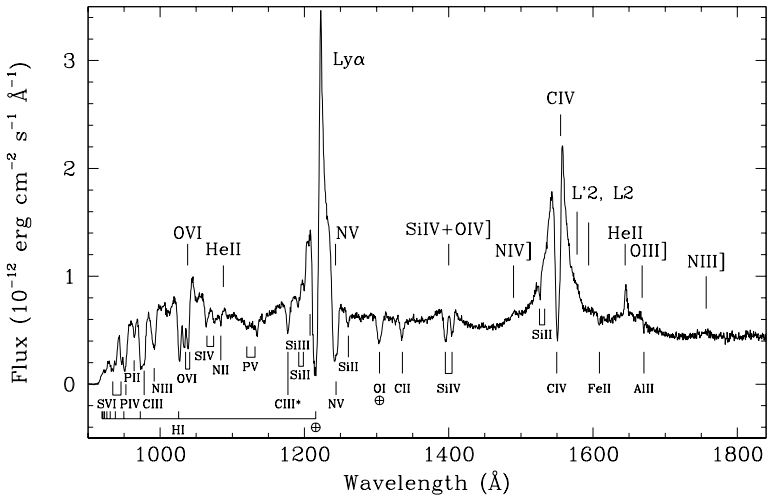


Figure 3. Far-ultraviolet spectrum of NGC 4151 obtained with the Hopkins Ultraviolet Telescope during the Astro-2 mission.

appearance is completely dominated by the bright, point-like nucleus. As in visible light, the nuclear spectrum shows a non-stellar continuum and prominent broad absorption lines. This spectrum, shown in figure 3, is similar to that seen in other Seyfert 1 galaxies (and in quasars) except for one unusual characteristic, which is the strong, complex assortment of absorption lines. While weak UV absorption lines are seen to be common in HST spectra of Seyfert galaxies, these are usually restricted to the transitions of highly ionized atoms and the Lyman α transition of hydrogen. In NGC 4151 absorption lines are associated with nearly all ionization levels, and they are particularly strong, broad (200–300 km s^{-1}), and blue shifted, indicating that the absorbing gas is outflowing from the nucleus.

X-ray characteristics

At x-ray energies, NGC 4151 is one of the brightest sources in the sky outside our own Galaxy. The 2–10 keV x-ray flux is variable on time scales as short as days, and it roughly averages 2.5×10^{-10} ergs cm^{-2} s^{-1}. This corresponds to a luminosity of 7×10^{42} ergs s^{-1}, making NGC 4151 one of the intrinsically least luminous Seyfert galaxies. The x-ray emission is dominated by the point-like nucleus, but there is some surrounding faint extended emission.

The x-ray spectrum of NGC 4151 is complex, and it displays a variety of features that have both illuminated and puzzled our understanding of active galaxies. The 2–10 keV spectrum is dominated by a power law continuum showing strong, variable absorption at energies below 4 keV. The power law continues to higher x-ray energies before showing an exponential cutoff beyond 100 keV. A strong 6.4 keV emission line attributed to Fe Kα emission is also present.

The most unusual aspect of the low-energy absorption is that x-ray emission is still visible at low energies at levels above what would be predicted by simple absorption by cold gas at the required column densities of $\sim 10^{23}$ cm^{-2}. When this was first noticed in observations obtained with the Solid State Spectrometer on the Einstein Observatory, it was interpreted as 'partial covering' of the x-ray source by the absorbing material, as might happen when a myriad of individual small clouds each obscured only a small portion of the x-ray emitting region, thereby allowing some of the x-ray flux to leak through holes between the clouds. More recent and detailed x-ray observations with the Japanese ASCA satellite show that the excess emission is more complex. Matching the observations requires the absorbing material to be partially ionized, thereby making it more transparent at x-ray energies, having a small percentage of the original x-rays scattered into our line of sight, and adding a soft thermal component at energies below 0.7 keV. In this model the ionized absorber could also serve as the scattering medium as well as the source of the soft thermal emission.

The shape of the hard x-ray spectrum and the presence of the iron emission line are signals of Compton x-ray reflection. This type of emission is generally seen in Seyfert galaxies, and it arises when the primary x-ray source illuminates surrounding colder material covering a large solid angle, such as an accretion disk (see ACCRETION DISKS). The Fe Kα emission shows both a narrow component as well as one that is relativistically broadened with the shape expected for material in an accretion disk orbiting the central black hole. Once the effects of the Compton reflection are taken into account, the intrinsic x-ray spectrum can be modeled as the emission produced when a hot (60 keV), optically thick plasma cloud Compton scatters low-energy seed photons. A natural source for these low-energy photons is the ultraviolet and optical radiation from the outer portions of the accretion disk. Thus Compton scattering is operating in two regimes and in two separate locations here—first, at high temperatures in the production of the primary x-rays, and second, in the reflection of those x-rays from cold surrounding material.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio morphology

Among active galaxies, NGC 4151 belongs to the class known as 'radio quiet.' Radio quiet doesn't mean radio invisible, however. At radio wavelengths, NGC 4151 also displays unusual activity compared to a normal galaxy. The nuclear region has a total radio flux of 145 mJy at 6 cm (5 GHz). This is resolved on arc-second scales into a two-sided radio jet oriented at a position angle of 77° and extending about two arc seconds from the nucleus. The radio core coincident with the nucleus is itself extended, and it is resolved with very long baseline interferometry (VLBI) into an elongated structure that is misaligned with the larger scale radio jet, but at an orientation (position angle 57°) that is nearly aligned with the biconical narrow-line region imaged with HST.

Unified models

Unified models that link the Seyfert 1 and Seyfert 2 classes of active galaxies rely on a combination of geometrical factors to explain the disparate appearance of what are intrinsically identical objects (see ACTIVE GALAXIES: UNIFIED MODEL). The basic Seyfert galaxy of either type has a 'central engine' in which material accreting onto a central black hole generates the radiation that we see and which illuminates the surrounding material in the host galaxy. Fast-moving gas clouds near the central engine emit the characteristic broad lines seen in Seyfert 1 galaxies; this is called the 'broad-line region'. Slower moving clouds at distances of 10–100 pc in the 'narrow-line region' produce the narrower lines typical of Seyfert 2 galaxies. An opaque ring, or torus, surrounds the central engine and the broad-line region. This torus blocks radiation from escaping in its plane. Thus, the narrow-line region is illuminated by radiation that has been collimated into two opposing cones. In Seyfert 2 galaxies, our line of sight lies in the plane of the torus. Our view of the central engine and broad-line region is blocked, and what we see is dominated by the extended, conically shaped narrow-line region. In contrast, our view of Seyfert 1 galaxies is from above the plane of the torus, where we have a clear line of sight to the broad-line region and the central engine. Since our line of sight lies within the cone of collimated radiation, any extended narrow-line structure has a symmetrical distribution about the point-like nucleus rather then the twin cones often seen in Seyfert 2s. In addition, the brilliance of the radiation from the nuclear region often makes the much fainter narrow-line clouds difficult to see.

The appearance of NGC 4151 presents a challenge to the simple unified model outlined above. The biconical narrow-line region imaged with HST and shown in figure 2 is typical of what one would expect in a Seyfert 2 galaxy, and it implies that our line of sight lies in the shadow of whatever structure is collimating the radiation. Yet, we clearly have a direct view of the central engine and broad-line region since the broad lines characteristic of a Seyfert 1 are prominent in its optical and UV spectra. Thus the collimation cannot be provided by an optically thick torus. Two possible explanations have been discussed in the literature.

The narrow-line region can still take on the appearance of a cone if the illuminated material lies in a thin plane parallel to the disk of the galaxy rather than uniformly filling the conical region of collimated nuclear radiation. Since the conical structure has sharp edges that can be traced down to distances less than 0.5 arc second from the nucleus, the plane must be no thicker than 35 pc and have sharp upper and lower boundaries. The virtue of this explanation is that the difference between the observed axis of the ionization cone and the radio jet is simply due to a projection effect. There are two difficulties, however. First, the ionization cone must have an intrinsic opening angle of 130°, much higher than the $\sim 90°$ inferred for other Seyferts. Second, the vertical distribution of the gas is unusually thin compared to our own Galaxy, and it must be sharply bounded rather than have a smooth vertical gradient.

The other alternative notes that while our line of sight to the central engine in NGC 4151 is fairly clear, it is not completely unobstructed. The x-ray spectrum shows heavy (but partial) low-energy absorption, and the UV and optical spectrum is rich in absorption features, including an optically thick Lyman limit. These absorption features are unusual for Seyfert 1 galaxies, and they are only found in those Seyfert 1s in which a biconical narrow-line region is visible. (NGC 3516 is the only other known case.) A slight modification of the unified model in which the obscuring torus is surrounded with a thin atmosphere can explain the absorption features and the biconical appearance of the narrow-line region. The gas in this atmosphere will be photoionized by the central source, and radiation pressure may drive it in an outflowing wind as seen in the high luminosity broad absorption-line quasars. This can produce the blue-shifted, ionized UV absorption features. For lines of sight through the atmosphere in which the total column density is $\sim 10^{21}$ cm^{-2}, the view is clear at all wavelengths longer than the Lyman limit of neutral hydrogen (912 Å), but it is optically thick throughout the ionizing UV and into the soft x-ray region of the spectrum. This atmosphere can then effectively collimate the ionizing radiation and still provide a clear line of sight to the central regions. One can then explain the unusual characteristics of NGC 4151 within the context of unified models by simply requiring a special line of sight for this unique object.

Bibliography

HST imaging of NGC 4151 and opposing viewpoints of its implications for unified models of Seyfert galaxies are presented by

Evans I N *et al* 1993 *Astrophys. J.* **417** 82
Boksenberg A *et al* 1995 *Astrophys. J.* **440** 151

Recent radio imaging of NGC 4151 is tied into this debate by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pedlar A *et al* 1993 *Mon. Not. R. Astron. Soc.* **263** 471

The ultraviolet spectrum of NGC 4151 and the implications of the absorbing gas for collimating the ionizing radiation are discussed by

Kriss G A *et al* 1995 *Astrophys. J.* **454** L7

The original paper on the soft x-ray spectrum of NGC 4151 and the 'leaky' absorber can be found in

Holt S S *et al* 1980 *Astrophys. J.* **241** L13

More recent x-ray observations are discussed by

Weaver K A *et al* 1994 *Astrophys. J.* **436** L27

The relativistically broadened iron line in NGC 4151 is shown by

Yaqoob T *et al* 1995 *Astrophys. J.* **453** L81

The x-ray though γ-ray spectrum of NGC 4151 and models for its formation are discussed by

Zdziarski A A, Johnson W N and Magdziarz P 1996 *Mon. Not. R. Astron. Soc.* **283** 193

Gerard A Kriss

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

NGC 5128/Centaurus A: 150 Years of Wonder

Just over 150 years ago, in 1847, Sir John Herschel saw in the southern sky 'a most wonderful object ... cut asunder ... by a broad obscure band' (see HERSCHEL FAMILY). At the time, there was no clue as to the nature of this celestial object. It took well over half a century for observational astronomy to advance to the point where it could hypothesize that it was extra-galactic, and only in the last 50 years have advances in radio, infra-red, x-ray, γ-ray and optical astronomy permitted us to think we understand this unusual and visually spectacular object.

We now know it as the galaxy NGC 5128. We know that its dominant visual feature is a dust lane, a large, edge-on disk of gas and dust around the galaxy, with areas of recent star formation, thought to have resulted from a past merger with another galaxy—updating the very early model for the morphology of a collision between two ELLIPTICAL GALAXIES. We know that it is a strong radio source, and hosts an ACTIVE GALACTIC NUCLEUS (AGN) which is a powerful source of emission at many wavelengths—AGN are thought to be SUPER-MASSIVE BLACK HOLES, accreting interstellar material from the host galaxy. In fact, we more commonly use its radio designation CENTAURUS A. Finally, we also know that Centaurus A is not unique. There are untold numbers of galaxies with significant dust lanes and features; dust is thought to play a major role in the evolution of galaxies and in star formation. We know that mergers play a major role in galaxy evolution and star formation. And we know that there is a broad variety of nuclear activity in galaxies, that massive black holes may be present at the centers of most, if not all, galaxies, that mergers may also be related to the presence of nuclear activity and/or massive black holes, and that nuclear activity may in turn affect star formation.

The phenomena we see in Centaurus A are thus, in fact, quite common, and Centaurus A is likely representative of many galaxies. Much of its uniqueness, and its spectacular appearance, arises from its proximity to us. At a distance of 3.5 Mpc, it is one of the nearest galaxies outside the LOCAL GROUP—the nearest giant elliptical galaxy, the nearest radio galaxy, the nearest active galactic nucleus, and the nearest example of a recent merger. The unprecedented detail potentially visible in Centaurus A due to its proximity allows us to learn more about all these phenomena: it should be an excellent laboratory for investigating the nature of active galactic nuclei, of radio and x-ray jets, of mergers, and of the relation among these phenomena. But in one of nature's perversities, the very dust lane which gives the galaxy its spectacular appearance also blocks visible investigations of the nuclear region. We must rely on all of modern astronomy's observational tools to understand this object.

The galaxy as a whole—its optical appearance

Optically, NGC 5128 appears as an elliptical galaxy with a dark lane cutting through the central region (figure 1). The obvious visual image stretches out some 5–10 arcmin, although intensity contours can be traced several times further out. Kinematic studies and dynamic modeling suggest a triaxial geometry. Estimates for the distance to NGC 5128 have varied though the years, ranging up to 8 Mpc, but it is now widely accepted to be between 3 and 3.5 Mpc away; we use the value of 3.5 Mpc—11 or 12 million light years away. At this distance, one arcsec corresponds to 17 pc; an arcmin is just about a kpc.

If we ignore the dust lane and inner region, NGC 5128 actually appears visually as a fairly normal giant elliptical galaxy. Most of the light comes from an older stellar population, and the light distribution follows a de Vaucouleurs law. The integrated visual magnitude of about 7 makes it one of the brightest galaxies in the sky, even without compensating for the obscuring dust lane (the intrinsic magnitude is ~ -22). Outside the dust lane, the light has an R-I color $\simeq$0.7, average for an elliptical. Inside the dust lane, the color varies generally from 0.9 to 1.5, with occasional higher peaks. Correcting for the estimated extinction reveals a relatively smooth brightness distribution, confirmed by near-infrared observations which are less affected by the dust. Within the dust lane, polarized red and infrared emission is seen, with spatially variable polarized intensity proportional to extinction, but with a relatively constant position angle of $\sim$110°. This suggests that the POLARIZATION is caused by dichroic transmission through aligned dust grains. The highest polarization region at the center has a different angle and coincides with a flatter spectrum compact knot, interpreted as a scattering knot very close to the active nucleus.

In the outer regions, NGC 5128 has a faint blue shell structure. These shells are interpreted as being caused by a galaxy merger—a small, gas and dust-rich disk galaxy colliding with the existing giant elliptical galaxy some 10^8–10^9 yr ago. The outer shells are debris of the intruder galaxy—stars and gas stripped from the small galaxy during the interaction. The blue color suggests the shells include new stars formed as a result of the collision. Neutral hydrogen is also seen near the shells.

The dust lane itself comprises the remaining gas and dust of the tidally disrupted incoming galaxy, having formed a disk around the center of the original giant elliptical galaxy. Molecular (e.g. CO, HCO^+, HCN), Hα, and H I kinematic mapping of the dust lane in emission and absorption are well modeled by a warped, twisted gas disk around the nucleus, viewed nearly edge-on. This disk, with a projected width of approximately an arcmin, effectively obscures the nucleus and the inner half-kiloparsec of the galaxy in the optical. The outer parts of the warped disk reflect the orbital plane of the incoming galaxy, the inner portion having settled into a principal plane of the elliptical galaxy. Sub-millimeter and recent ISO mid-infrared observations (see INFRARED SPACE MISSIONS) have shown a 5 kpc structure along and within the dust lane, symmetrical around the nucleus. This structure resembles a barred spiral associated with the thin disk

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

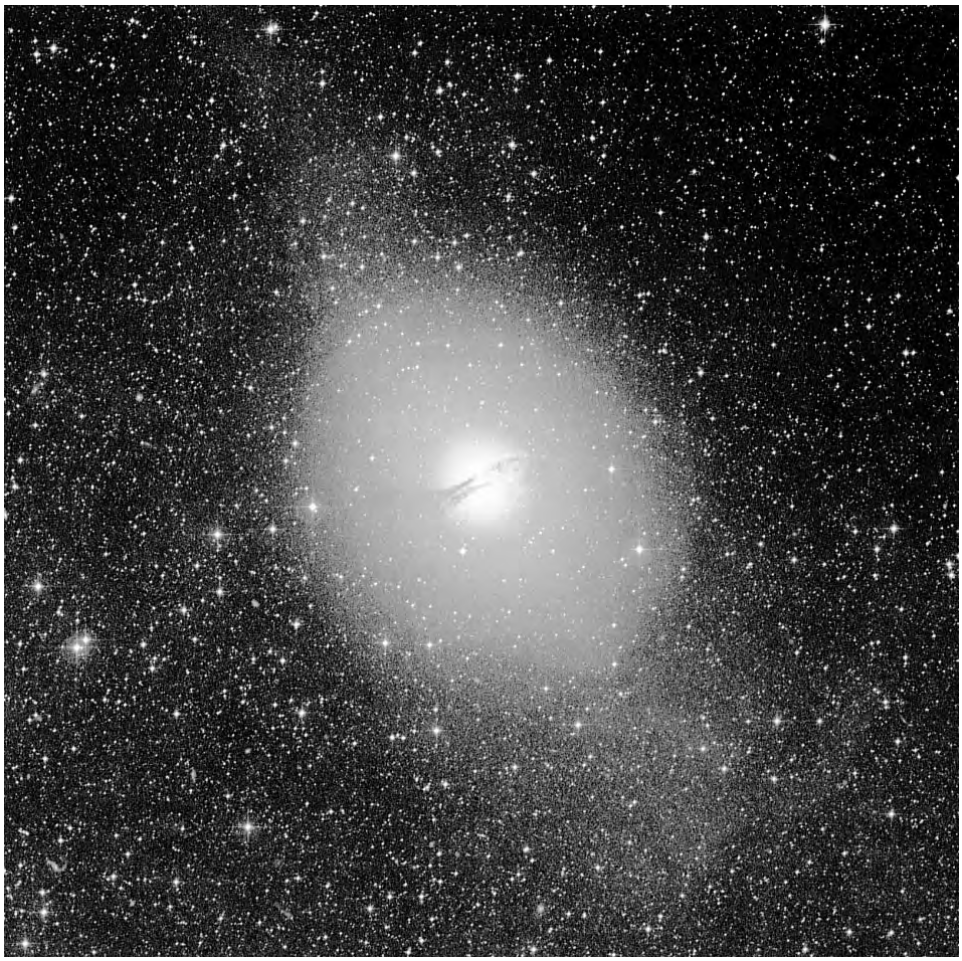


Figure 1. Deep visible light image of NGC 5128, showing the large (50′ ∼ 50 kpc) low-luminosity extent of the elliptical galaxy, the bright bulge, and the warped dustlane.

formed by tidal debris; it is, perhaps, related to the transfer of material into the central region around the AGN.

Associated with the edges of the dust lane are young blue stars, stellar associations, globular clusters, and H II regions. Early HUBBLE SPACE TELESCOPE (WFPC1) observations revealed some 74 sources above detection limits (23.0, 22.6 in R, I) along the northern edge of the dust lane, much larger than the density in neighboring parts of the galaxy. The luminosities of many of these objects exceed that expected for single stars, and many were resolved on the several parsec scale. A large color dispersion, with many blue objects dominated by spectral type A or earlier stars, is consistent with recent star formation along the dust lane. The blue stars, stellar associations and globular clusters, also studied with HST/WFPC2, indicate that new stars were likely formed along the dust lane as a result of the merger.

The active galaxy—Centaurus A in radio, x-rays and infrared

Centaurus A was one of the earliest powerful discrete radio sources discovered, and is, of course, the nearest of the giant RADIO GALAXIES (see figure 2). It was the identification of NGC 5128 with Centaurus A in 1949 that sparked interest in the galaxy, suggesting that it was indeed an unusual object. One of the well-known early optical images of the galaxy was labeled 'NGC5128—peculiar galaxy, source of radio noise'. Baade and Minkowski confirmed its extragalactic nature by radial velocity measurements in 1954. Early radio observations revealed giant outer lobes on a scale of several degrees and intermediate and smaller scale structure down to the scale of several arcmin. It was only in 1980, however, that high resolution observations with the VLA revealed the detailed morphology of the inner lobes, nucleus and jet (see below).

The nucleus itself, although invisible in the optical, has by now been studied for years via radio, x-ray, gamma ray, infrared, and mm observations. It was one of the first

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

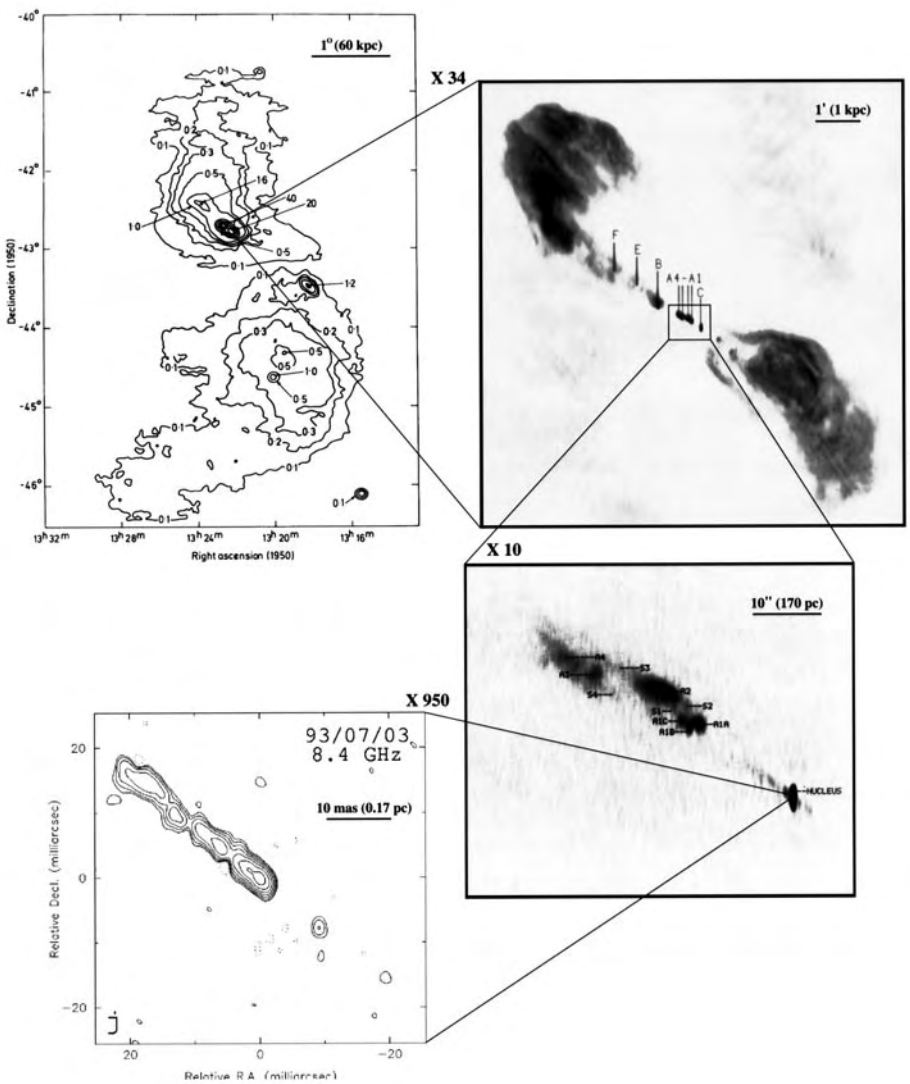


Figure 2. Radio structure of Centaurus A on scales of 6° down to milliarcsec, a dynamic range of several million.

active galactic nuclei identified, even before we recognized AGN as a major category of astrophysical phenomena. A strong x-ray source was found in the vicinity of Centaurus A in 1970. The x-ray source was confirmed to be associated with the nucleus of the galaxy by UHURU and ANS. Its luminosity is about 10^{42} erg s^{-1}, not particularly unusual for AGN. It shows a simple power law spectrum in the few keV to tens of keV range, similar to that of Seyferts seen in the same energy range. Strong absorption is seen at low x-ray energies, historically thought to derive from the optically obscuring dust lane, but now better interpreted via the so-called standard model of AGN, in which we are looking through a dense accretion torus around a black hole. OSSE gamma-ray data show a slightly flatter power law out to 120 keV, and other gamma-ray observations have shown some steepening again at higher energies. Fe K-line emission has been detected by ASCA, consistent with reprocessing by the 10^{23} H cm^{-2} column, and some other Hα lines are seen at lower x-ray energies, suggesting scattering and fluorescence in a thinner, cooler medium outside the nucleus. An infrared hotspot was reported as the nucleus in 1971, but modern observations reveal a lot of structure in the nuclear region (see below).

When x-ray imaging with the Einstein Observatory became possible in 1979, a jet and other diffuse features were discovered in addition to the nucleus. The spectacular jet was the first such discovered via its x-ray emission, and only subsequently studied in radio with the VLA. Extended x-ray emission on the scale of the galaxy as a whole, at the level of about 10^{39}–10^{40} erg s^{-1}, originally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

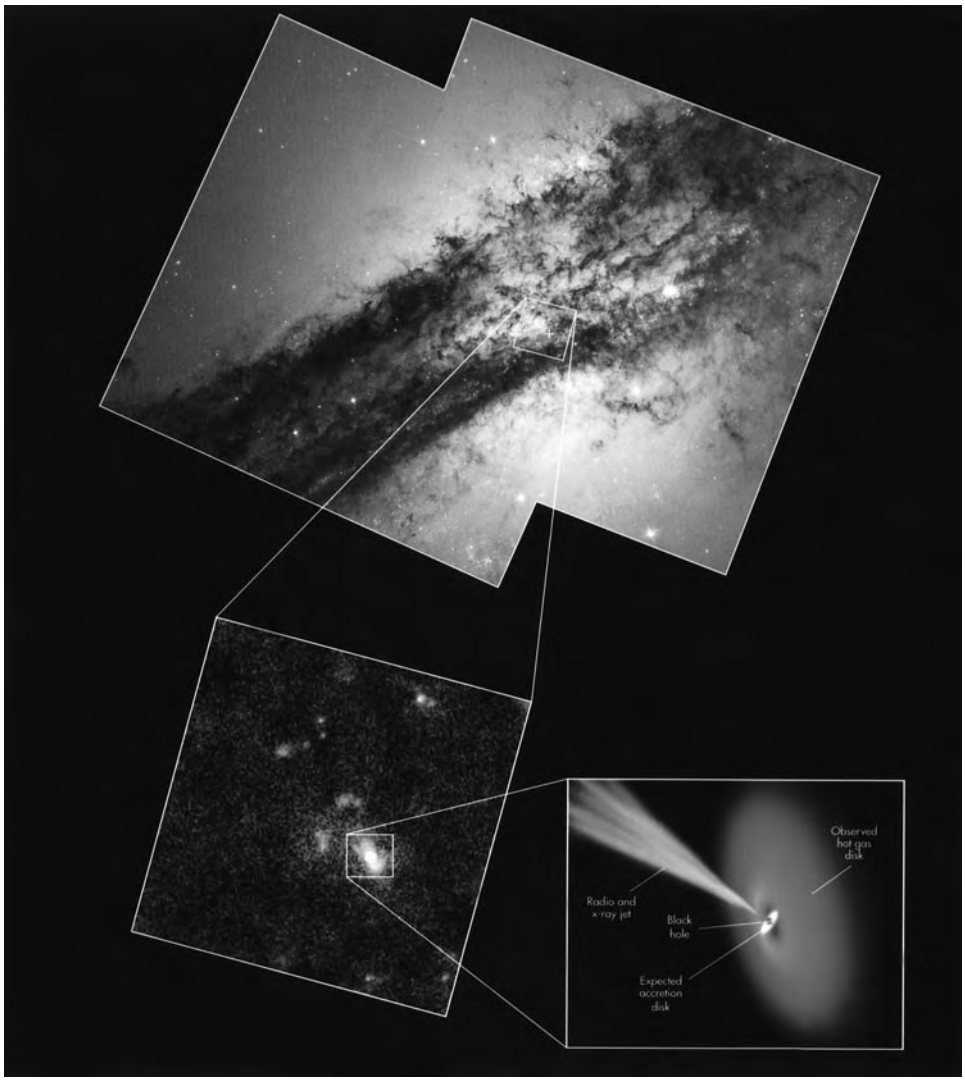


Figure 3. Upper panel: HST WFPC2 image of the core of NGC 5128. Note the bright young stars along the edge of the dustlane. Lower left: HST NICMOS image of the ionized gas at the nucleus of the galaxy. Lower right: Schematic model of the hot gas disk around the active nucleus and the radio/x-ray jet.

hypothesized as inverse Compton emission from the radio lobes, was later found to be common in other galaxies, and is likely due to emission from hot gas produced by stellar mass loss. Ridges of x-ray emission parallel to but outside the dust lane have been interpreted as being associated with the young population in the disk of the galaxy, possibly the integrated emission of massive x-ray binaries.

The radio source in general, and the jet in particular, is not unusual. The large-scale structure is typical of the large class of medium power, double-lobed radio sources, although the presence of both inner and outer sets of lobes suggests at least two different epochs of radio activity. Possibly related to this, the symmetry axis changes progressively from the outer to inner lobes, and in to the observed jet, suggesting changes in the jet direction over time, caused, perhaps, by precession. The jet morphology is rather similar to that of M87, with constant opening angle and knots, having an inner core and an outer halo. The nucleus has an inverted spectrum ($\alpha = -0.55$), and jet shows a synchrotron spectrum, steepening slightly with distance ($\alpha \simeq 0.48 \rightarrow 0.72 \rightarrow 0.86$ in the lobe) and with rather good correlation between radio and x-ray morphology in the inner knots, suggesting that the x-ray emission might also be synchrotron. Unfortunately, x-ray spatial and spectral resolution has not yet been adequate to unambiguously distinguish among synchrotron, thermal, and inverse Compton emission mechanisms. A synchrotron model would require continuous re-acceleration of electrons to replenish their energy, since the x-ray emitting electron lifetimes would only be $\sim$50 yr or less, far less than the light

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

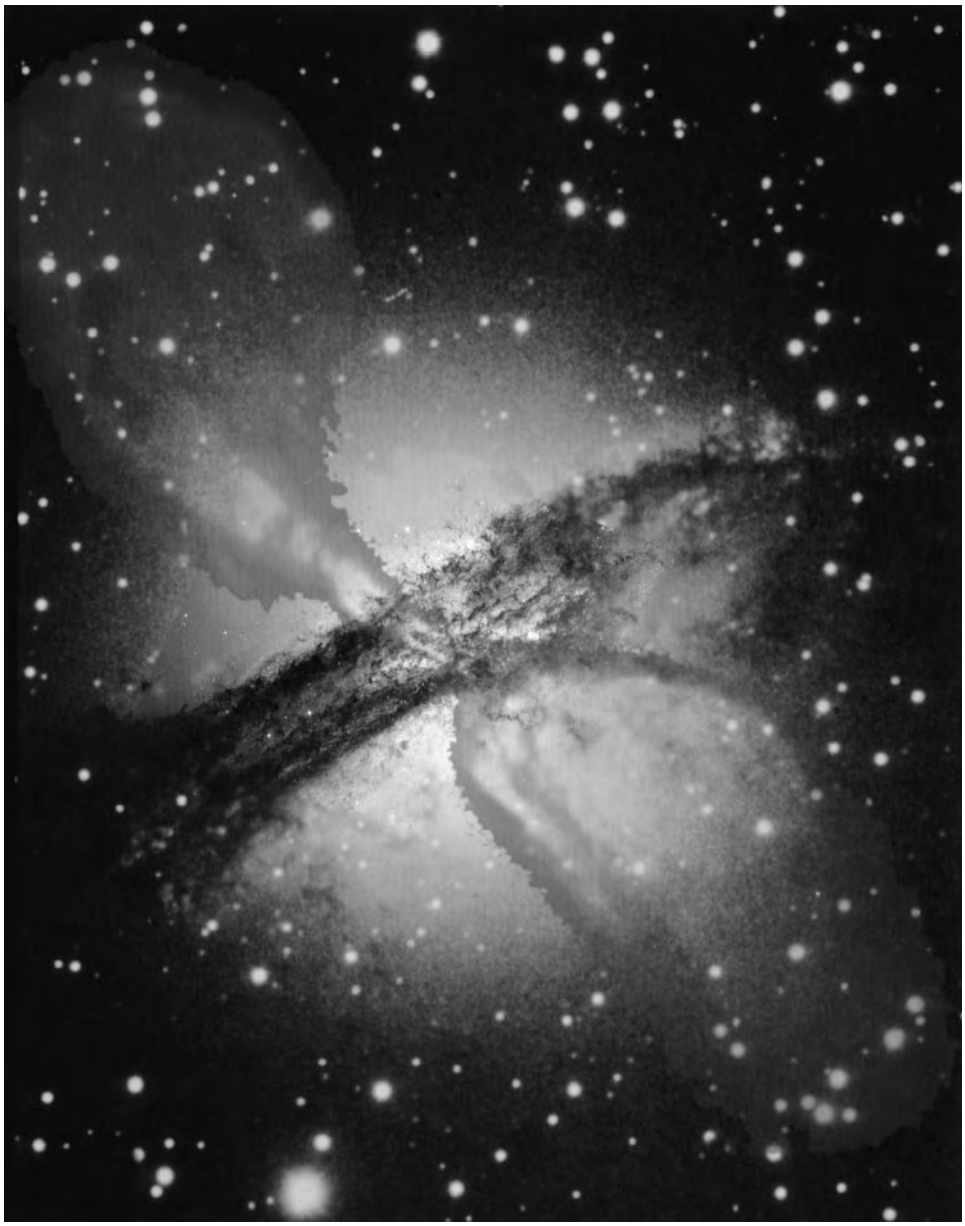


Figure 4. NGC 5128 = Centaurus A. The inner radio lobes and jet (blue outline, stretching from lower right to upper left) and the high resolution HST true-color image (central regions), superimposed on the ground-based, lower resolution visible light image. **This figure is reproduced as Color Plate 37.**

travel time along the jet. Then again, in a perverse 'catch 22' of nature, if there is indeed re-acceleration, we might also expect to see shocks, which could in turn produce thermal x-rays.

In addition to the jet components seen in x-rays, there are also several radio components within 1–5″ of the nucleus, a VLBI jet on the scale of 100–150 mas, showing sub-luminal motion, and a VLBI counterjet. The VLBI observations, on the one parsec scale, are consistent with Centaurus A being an FRII source; the jet is not relativistically beamed toward us—it is close to the plane of the sky. VLBA observations limit the size of the nucleus to $\sim$0.5 mas $\simeq$0.01 pc $\simeq$10 light days, the smallest known extragalactic radio source. X-ray variability suggests an even smaller size, since the source is known to vary on scales ranging down to less than a day.

Optical emission from the jet has been variously reported through the years at different places in the galaxy. There are indeed faint blue filaments of line-emitting gas and stars about 8′ from the nucleus along the line of the jet, but outside the radio lobe. These suggest jet interaction with the interstellar medium, but are not

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

directly correlated to specific x-ray/radio features. There is also a faint linear optical feature along the position angle of the jet near the nucleus seen by HST, but there have been no confirmed detections of optical counterparts to any radio or x-ray knots.

The active galactic nucleus—a black hole

The proximity of this nearest active galaxy, in which 1″ corresponds to only 17 pc, presents the expectation that it would be a good test bed to explore the standard model for AGNs at the parsec scale. Nonetheless, the obscuring dust lane, with its high extinction, has hindered high resolution studies of the innermost regions around the presumed BLACK HOLE. X-ray spectroscopy has suggested the presence of thermal components near the nucleus in addition to the highly cut-off, non-thermal emission from the nucleus itself, and AXAF is expected to further improve our understanding of the high energy emissions from the AGN. High resolution imaging and spectroscopy in the near-infrared, made possible by HST and by the new generation of ground-based 8 m telescopes, are rapidly remedying our lack of insight into the immediate environment of the massive black hole, as we write. HST has produced diffraction-limited images in U, V, I, H and K bands, and HST and VLT are likely to soon produce spectra with spatial resolutions on the 0.1″ scale in R through K.

HST/WFPC2 visible light images have not only revealed new detail of the morphology within the dust lane, but have finally revealed the unresolved nucleus in V and I (see figure 3). HST/NICMOS near-infrared images also show the nucleus and extended galaxy emission, but in addition reveal an elongated ionized structure, interpreted as a gas disk emitting in Paschen α and [Fe II]λ1.643 μm. This inclined disk is approximately 20 pc in radius and roughly perpendicular to the larger dust lane, perhaps on another principal plane of the galaxy. Notably, the disk is not perpendicular to the jet, despite the fact that the standard model suggests that the central torus around a black hole would be perpendicular to a jet. The ionized feature is thought to be an extended ACCRETION DISK, aligned with the gravitational potential of the galaxy, perhaps transferring material from the dust lane and the barred structure seen in the mid-infrared into the dense torus around the black hole at the center. Ground-based near-infrared spectroscopy has revealed broadened lines at the nucleus in the 1–2.5 μm range, likely dominated by the integrated emission from this disk, leading to hopes that we will indeed be able to use HST or the 8 m telescopes to map the velocity structure in the disk, measure the mass of the black hole, and perhaps better understand the details of the accretion process itself (see figure 4).

Summary

NGC 5128—Centaurus A—is indeed a wondrous object, as first noted more than 150 years ago. We now know that its wonder goes far beyond its visual appearance. It gives us a close-up look at a number of astrophysical phenomena which are fundamental to our current world view of the universe and its evolution: the nature of AGNs and the presence of super-massive black holes at the centers of galaxies; the role of mergers in star formation, in the building of elliptical galaxies, and in the presence of activity in galaxies; and the presence of dust and its relation to all these processes. The spectacular appearance is due merely to its proximity and orientation—it is otherwise representative of a large class of giant elliptical galaxies with nuclear activity. It illustrates the fact that the nearest objects of a class often do not initially fit into canonical models—their proximity allows us to see far more detail and more strictly test standard models and interpretations of phenomena derived from observing more distant objects, seen with fuzzier vision. Our growing understanding of our close neighbor Centaurus A will allow us to extrapolate our understanding to galaxies at much greater distances and ages, gaining new insight into the evolution of galaxies and of the universe.

The author gratefully thanks Anton Koekemoer for help in preparing this article.

Bibliography

Israel F 1998 Centaurus A—NGC5128 *Astron. Astrophys. Rev.* **8** 237

Marconi A *et al* 1999 Unveiling the active nucleus of Centaurus A *Astrophys. J.* at press

E J Schreier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nicholas of Cusa [Nicolaus Cryffts, Chrypffs or Krebs, Nicolaus de Cusa, Cusanus] (1401–64)

Cardinal and philosopher, born in Cues or Cusa, Germany. He studied at Heidleberg and Padua, and became a cardinal in 1448. He marks the end of the philosophy of the Middle Ages. In his writing on astronomy, he denied the finitude of the world and its enclosure by crystal walls, suggesting instead that the universe was 'interminate'—'without boundaries'—and that the Earth therefore cannot be at the center of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nicholson, Seth Barnes (1891–1963)

Astronomer, born in Springfield, IL, joined the staff of the Mount Wilson Observatory. While photographing the recently discovered eighth moon of Jupiter with the 36 in Crossley reflector, he discovered the ninth, whose orbit he computed. He discovered three more of Jupiter's satellites, as well as a Trojan asteroid. He observed the Sun, including its sunspot activity and magnetism. With Edison Pettit he used a thermocouple to measure the temperature of the Moon (illuminated and eclipsed) and the planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nicolaus Copernicus

Nicolaus Copernicus (1473–1543) was a Polish astronomer whose *De Revolutionibus Orbium Coelestium* (1543) was the final achievement under the agenda of the ancient Greek astronomers and—by its claim that the Earth orbits the Sun—provided a major contribution to the later development of a dynamics of inertial motion.

Mainstream Greek astronomy conceived the universe as being bounded by the spherical heavens, in the midst of which was the spherical (and motionless) Earth. Astronomers attempted to 'save the appearances' (to reproduce the observed movements) of the planets by means of calculations based on geometrical models centered on the Earth; these models comprised circles on whose circumferences the center of another circle, or the planet in question, moved (PLANETARY THEORY FROM EUDOXUS TO COPERNICUS). Since Venus (for example) orbits the Sun which appears to us to orbit the Earth, and since the orbits are in each case not very different from circles and the speeds of the two bodies not very far from uniform, a model of the motion of Venus in which the planet moves on a circle whose center moves uniformly in a circular though somewhat eccentric orbit about the Earth can give an encouraging first approximation to the observed behavior of Venus. On the other hand, the desire for accuracy (on the part of astronomers and astrologers alike) guaranteed that no such model, nor any elaboration of it, would ever prove completely satisfactory: there would always be an incentive to develop and refine further.

In the second century AD, PTOLEMY in his *Almagest* and elsewhere had developed models that came near to saving the appearances. However, in these models Ptolemy had found it necessary to introduce non-uniform motion, namely motion that appeared uniform when viewed from an off-center 'equant' point that we recognize as analogous to the empty focus in a Keplerian ellipse. This violated a fundamental truth about the cosmos, and so brought about a rift between natural philosophers interested in understanding the universe and astronomers glad to have models that, whatever their defects, gave good predictions.

Another limitation of the Ptolemaic models was their *ad hoc* character. In the *Almagest*, which was to dominate astronomy for 14 centuries, Ptolemy used devices that had no rationale other than that they worked. Thus, faced with the observed fact that Venus and Mercury are never seen far from the Sun, he simply required the centers of the models for Venus and Mercury to be aligned with the mean Sun, without offering any explanation of why this should be so. Furthermore, in the *Almagest* Ptolemy made no attempt to combine the individual models into a unified system with a common scale. In the Renaissance, when both Aristotle's concentric nested spheres model was recovered, and the competing detailed mechanisms of Ptolemy understood, an integrated and intellectually satisfying picture of the cosmos became once more an important desideratum, and this came to be seen as a serious defect.

By the fifteenth century, Latin scholars were in the later stages of assimilating the achievement of Ptolemy, as mediated and elaborated by Islamic astronomers (see ISLAMIC ASTRONOMY). They recognized that the Western planetary tables based on the *Almagest* and named for the thirteenth-century Alfonso X of Castile, were of good, though by no means perfect, accuracy. On the other hand, Ptolemy's use of the equant violated a basic truth of natural philosophy, and his failure in the *Almagest* to offer an intellectually satisfying view of the cosmos as a whole was becoming increasingly troubling.

The man who developed geometrical models that more than matched Ptolemy's in predictive accuracy and achieved this without resorting to equants, and who also provided an elegant and convincing picture of the cosmos, was born in Toruń on the Vistula on 19 February 1473. Nicolaus Copernicus was the son of a prosperous merchant who died a decade or so after his son's birth. Copernicus then came under the protection of his uncle, who was to become bishop of Warmia in 1489. From 1491 to 1495 Copernicus studied at Cracow, where his purchases included a volume of the Alfonsine Tables. In 1495 his uncle obtained for him a canonry of the chapter of Warmia (a position that did not necessitate his being ordained priest), but he did not formally take up the post until 1497. In 1496 he enrolled at Bologna, the leading university for law, where he studied both canon and civil law, though he resided with the professor of astronomy. In 1497 and (twice) in 1500 he made astronomical observations that he was later to use in his major publication, and about 1500 he is reported to have lectured in Rome on astronomy. In 1501, after Copernicus had returned to Poland, the chapter agreed to his studying medicine at Padua for two years. This done, he again returned to Poland, where he was to remain for the rest of his life. At first he became personal physician and administrative assistant to his bishop uncle, but in 1510 he moved to Frauenberg, the headquarters of the chapter.

Meanwhile Copernicus's interests in astronomy had developed, and somewhere he had encountered the planetary theory of the Islamic astronomers of Maragha, elements of which were to appear in his own mature planetary models, to the wonderment of modern historians. The catalog of the library of a Cracow professor dated 1 May 1514 contains an entry, 'Next a quire of six leaves of a *Theorica* maintaining that the Earth moves while the Sun is at rest', and this must refer to a copy of the little treatise that Copernicus was circulating anonymously and in manuscript and is known as the *Commentariolus*. Its modern editor thinks it 'may have been written in a momentary burst of enthusiasm, perhaps immediately upon devising the heliocentric theory'.

Already in this sketch we encounter the problem that would bedevil the later acceptance of his developed theory: he was attempting two different tasks, the promulgation of a new cosmology in which the Earth was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in orbit about the Sun, and the provision for mathematical astronomers of a new set of planetary models purged of the equant, that philosophically suspect device employed by Ptolemy.

Of his great *De Revolutionibus Orbium Coelestium Libri VI (Six Books on the Revolutions of the Heavenly Spheres)*, which eventually appeared in 1543, only the first 10 chapters of Book I, or about 6% of the whole, were dedicated to cosmology. The remaining 94% was densely mathematical and inpenetrable to all but a handful of specialists. These specialists would for the most part focus on what interested them and ignore the preliminary cosmology, while the many who might have debated the cosmology would pass by so forbidding a volume—especially since (as we shall see) an anonymous preface explained that the motion of the Earth was proposed merely as a mathematical device to facilitate computational accuracy.

The problematic nature of this assertion, that the book was intended merely to save the appearances, would have been clear to anyone willing to master its contents; but, as Copernicus was well aware, it was indeed formally impossible, simply by means of applied geometry, to *demonstrate* the motion of the Earth. Only when Kepler taught astronomers to think dynamically would it become clear that the Sun must be the focus of the planetary system. All Copernicus could do was to argue that the harmony and simplicity of the heliocentric vision was convincing, and not only to him but to anyone who shared Plato's sense of cosmic beauty.

The planets had worried Greek astronomers not only by moving against the background of the 'fixed' stars—which was bad enough—but by coming to a halt from time to time and temporarily reversing their direction of movement. But on the heliocentric approach such retrogressions were to be expected; indeed, it would be worrying if they did *not* occur. For if the Earth is one of several planets orbiting the Sun, then we will expect the Earth from time to time to overtake each of the outer planets, and the outer planet will then appear to retrogress, not because it is in fact moving backwards, but because the Earth is moving forwards faster. We will expect this to occur when the outer planet is in opposition to the Sun—and in fact this is exactly what we observe. Similarly, when the Earth is overtaken by an inner planet, the planet will appear to reverse its normal direction; and this will happen when it is in inferior conjunction with the Sun.

A problem with Ptolemaic astronomy had been that while Saturn was slower than Jupiter in circuiting the sky, and therefore was plausibly further away, while Mars was quicker and therefore nearer, the Sun, Venus and Mercury all took one year to go round against the background of the stars; the order of their respective distances from Earth was therefore moot. In other words, in Ptolemaic astronomy even the very order of the planets was open to debate. But on the heliocentric view, not only the order of the planets but the actual radii of their orbits could be determined. If we observe the angle Sun–Earth–Venus (and Sun–Earth–Mercury) when the planets are at maximum elongation from the Sun, then elementary trigonometry will give us the radii of their orbits as fractions of an astronomical unit. Observations of the outer planets can likewise yield the radii of their orbits, and in this way we can envision a true system of the planets, which we may depict with the (basic) orbits all drawn to scale.

Furthermore, the annual period of the Sun, Venus and Mercury was revealed as no more than a reflection of the annual period of the Earth: Venus and Mercury had quite different periods which could now be determined. It then emerged that the order of the planets by distance exactly matched the order of the planets by period—not a formal proof of the heliocentric hypothesis, but thoroughly persuasive to those of a certain cast of mind. Indeed, given this geometry it was easy to predict just when a planet would appear to stop its forward movement, and when it would subsequently resume this forward movement once more: the phenomena that had given the 'wandering stars' their name had been shown to be thoroughly lawlike and amenable to prediction, no more than a natural consequence of the Earth's status as one planet among several.

The merits of a heliocentric cosmology are sketched in the *Commentariolus*. The author also offers geometrical models for the motions of the various planets, in longitude and in latitude. The models for longitude are modified from the Maragha models, but referred to the center of the Earth's orbit, while those for latitude derive more directly from Ptolemy. These models are no more than outline, and would require elaboration before becoming the basis for computations of sufficient accuracy. The parameters involved in this elaboration would depend upon observations to be made when the planets concerned were in particular configurations, and the opportunities for such observations would be spread over many years.

Copernicus's interest in astronomy had to take second place to his duties as a canon. So for three years, and later for a part of a fourth, he was Administrator of Benefices, required to travel throughout the chapter's estates collecting revenues and administering justice. During these three years he drew up a set of recommendations for the reform of the coinage, whose debasement he ranked as a disaster alongside civil strife, disease and famine. For much of 1523 he was Administrator General of the diocese between the death of one bishop and the appointment of his successor.

Furthermore, he lived in troubled times; and even in the rare intervals of peace his medical skills were much in demand. But by about 1530 the necessary astronomical observations for the determination of parameters had been made, and soon his treatise was taking shape. By the middle 1530s he was able to compute an almanac (now lost), as a demonstration that the heliocentric approach was more than empty rhetoric. By 1539 the work was essentially complete, and in May of that year he allowed sight of it to a young visitor, Georg Joachim RHETICUS

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(1514–74). Rheticus was a professor of mathematics at the University of Wittenberg, and he brought with him as gifts a number of mathematical and astronomical books that had been published in Nuremberg by Johann Petreius. By this time, despite Copernicus's isolation, word of his astronomical work had spread in the German-speaking world, and even as far as Rome, and Rheticus's curiosity had been aroused. Copernicus allowed Rheticus to summarize his treatise in a *Narratio Prima*, or *First Account*, which appeared in 1540. By then the elderly canon was revising his work for publication, while Petreius had indicated a strong interest in being the publisher.

In October 1541 Rheticus returned to Wittenberg to resume his teaching, and he either took Copernicus's manuscript with him or had it sent to him soon thereafter. In May 1542 Rheticus was able to deliver the fair copy to Petreius, and in the coming months he corrected the proofs, until he was appointed to a post at Leipzig. The task of seeing the book through the press then passed to a Lutheran clergyman, ANDREAS OSIANDER, who took it upon himself to forestall criticism of Copernicus by adding an unsigned preface to say that the motion of the Earth was not being proposed as a truth of nature but merely for purposes of calculation. Rheticus was involved in an attempt to persuade the Nuremberg town council to compel Petreius to issue a corrected edition in which Copernicus's true position on the question was clearly stated, but this failed, and the misleading preface was taken to be the author's until KEPLER revealed Osiander's role after the turn of the century.

Late in 1542, Copernicus suffered a cerebral hemorrhage and paralysis of his right side, but he lingered on until 24 May of the following year. We are told he received the final pages of *De Revolutionibus* on the day of his death; but if so, he would hardly have been aware of it.

For the remainder of the century Copernicus's treatise was mined by mathematical astronomers for the ingenuity of its planetary models. Not until 1596 did Kepler publish an unabashedly heliocentric cosmology, and not until 1610 did GALILEO's telescopic observations lead him to embark on a reform of physics that would make it possible for Earth-dwellers to accept the notion that they inhabited a spinning, orbiting planet without feeling any sensation of movement.

Bibliography

Duncan A M (trans) 1976 *Copernicus: On the Revolutions of the Heavenly Spheres* (Newton Abbot: David & Charles)

Gingerich O 1983 Ptolemy, Copernicus, and Kepler *The Great Ideas Today: 1983* (Chicago: Encyclopaedia Britannica) pp 137–80

Rosen E 1959 *Three Copernican Treatises* revised edn (New York: Dover)

Swerdlow N 1973 The derivation and first draft of Copernicus's Planetary Theory: a translation of the *Commentariolus* with commentary *Proc. Am. Phil. Soc.* **117** (6)

Swerdlow N M and Neugebauer O 1984 *Mathematical Astronomy in Copernicus's De Revolutionibus* (New York: Springer)

Owen Gingerich and Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nicolaus Copernicus Astronomical Center

Nicolaus Copernicus Astronomical Center is the largest astronomical institution in Poland, located in Warsaw and founded in 1956. At present it is a government-funded research institute supervised by the Polish Academy of Sciences and licensed by the government of Poland to award PhD and doctor habilitatus degrees in astronomy and astrophysics. In September 1999 staff included 21 senior scientists with permanent or tenure-track positions; 13 post-doctorates; 16 PhD students. The main research fields are theory and observations of binary stars; stellar structure and pulsations; stellar atmospheres and circumstellar matter; cosmology and large-scale structure of the universe.

A PhD program in astronomy and astrophysics is taught and there is international cooperation with various observational projects: Integral (a gamma-ray telescope to be launched in 2001); SALT (South-African Large Telescope) and WET (Whole Earth Telescope); French–Polish program Jumelage (collaboration in theoretical astrophysics). Main achievements are the structure of and mechanisms operating in cataclysmic binaries; evolution of single and binary stars; theory of thick accretion disks; helioseismological models of the Sun's interior; limits on the equation of state of neutron stars; Compton reflection in AGNs.

For further information see http://www.camk.edu.pl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Niels Bohr Institute for Astronomy, Physics and Geophysics

NBIfAFG is located at Blegdamsvej 17, DK-2100 Copenhagen, Denmark. NBIfAFG has four physics departments including COPENHAGEN UNIVERSITY ASTRONOMICAL OBSERVATORY (CUAO).

For further information see
http://www.nbi.dk/NBIfAFG

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nitrogen (N)

The element that is the principal constituent of the Earth's atmosphere (making up, at ground level, 78.08% by volume). By number of atoms, it is the fifth most abundant element in the universe (by mass, it is seventh). The chemical symbol for nitrogen is N, and its normal molecular form, N_2 (i.e. a molecule made up of two nitrogen atoms). Its atomic mass is most commonly 14 and its atomic number is 7. Although nitrogen makes up the major part of the Earth's atmosphere, it comprises only 3.5% of the atmosphere of Venus and 2.7% of the atmosphere of Mars.

See also: cosmic abundance of elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nobel Prize Winners in Astronomy

One of the greatest honours which can be bestowed upon any human being is the Nobel Prize. It is named after Alfred Nobel, a Swedish physicist who was particularly noted for his work on explosives. Nobel was born in 1833, and had a long and successful career, so that, when he died, on 10 December 1896, he left a large sum of money. This was to be put into a trust to establish an annual prize to those who, in the preceding year, shall have conferred the greatest benefit on mankind; there were to be prizes for physics, chemistry, physiology or medicine, literature and peace. Note that there was to be no specific Nobel prize for astronomy, but it is impossible to separate astronomy from any other branch of science, and several Nobel laureates have earned their prizes for work which is mainly astronomical.

The decisions are made by four institutions, three Swedish and one Norwegian; the prize involves a gold medal, a diploma and a sum of money.

The first prizes were awarded in 1901, five years after Alfred Nobel's death. An early laureate was SVANTE AUGUST ARRHENIUS, a Swedish chemist (1859–1927). The citation read that he received the prize 'in recognition of the extraordinary services he has rendered to the advancement of chemistry by his electrolytic theory of dissociation', but he was also very much concerned with astronomy; he studied cosmic physics, and was particularly concerned with radiation pressure on comet tails. He also put forward the 'panspermia' theory, according to which life on Earth did not originate here, but was brought to our world by a meteorite. The theory never gained wide acceptance, and neither has a much more modern theory, due to Sir Fred Hoyle and Chandra Wickramasinghe, that life was brought here via comets. However, it caused a great deal of discussion, and in his own field Arrhenius certainly deserved the Nobel Prize.

In 1907 the Physics Prize went to ALBERT A MICHELSON (1852–1931), German by birth but who spent most of his career in America. He was awarded the prize for his work on designing optical instruments and carrying out experiments with them. He is best remembered for his part in the Michelson–Morley experiment, in which it was demonstrated that the velocity of light is constant and is independent of the Earth's motion in space—a result which finally disproved the existence of the hypothetical 'aether', a medium which was supposed to permeate all space. He also designed an interferometer, which was used with the Mount Wilson 100 in reflector to make the first direct measurements of the angular diameters of some nearby giant stars such as Betelgeuse.

Obviously we cannot forget ALBERT EINSTEIN (1879–1955), widely and justifiably regarded as the world's greatest mathematician since Newton. In 1905, while still working at the Patent Office at Berne in Switzerland, he published three theoretical papers—and in retrospect any one of these was good enough to earn a Nobel Prize. In 1915 he produced his general theory of relativity, on which all subsequent work has been based. He was finally awarded the Nobel Prize for Physics in 1921.

It seems strange that Einstein had to wait several years for this recognition, but an even longer delay was experienced by the two American physicists ARNO PENZIAS (1933–) and ROBERT WOODROW WILSON (1936–), who shared the 1978 Physics Prize with the Russian, Peter Kapitsa. The story here goes back for more than a dozen years.

It began with some theoretical work in the United States by R H DICKE, who was investigating the very early stages of the universe. According to the 'big bang' theory, the universe as we know it came into existence around 12 000 million years ago; space, time and matter were created simultaneously. (One cannot ask what happened before the big bang, because if time started at this moment there was no 'before'.)[1] The temperature at this moment was of the order of 10^{32} K, but cooling was rapid, and 3 min after the big bang the temperature was down to a few thousand million kelvins. By now, according to Dicke's calculations, the overall temperature of the universe should have fallen to 3 K—that is to say, 3 degrees above absolute zero, the coldest temperature that there can possibly be ($-273\,^{\circ}$C). In this case, a background radiation should be detectable at radio wavelengths.

Meanwhile, Arno Penzias and Robert Wilson had been carrying out work with a specially built radio telescope, not to search for any background radiation but for quite different reasons. They found a steady, persistent signal coming from all directions and could not explain it; for a time they even put it down to pigeon droppings in the antenna of their radio telescope. When Dicke learned about this, it took very little time to establish that what Penzias and Wilson had found was in fact the expected radiation—the last remnant of the big bang. There can be little doubt that this was one of the most important cosmological discoveries of the 20th century, and it finally disposed of the steady-state or continuous creation theory, according to which the universe has always existed and will exist for ever.

The Austrian physicist VICTOR FRANZ HESS (1883–1964) shared the 1936 Physics Prize with CARL D ANDERSON. Hess' main contribution to science was his discovery of cosmic rays, in 1911–1912. He was then at the university of Vienna, specializing in studies of the upper atmosphere, and he made ten balloon ascents, establishing for the first time that cosmic rays are bombarding the Earth from all directions all the time (in fact the name is misleading; cosmic rays are not rays at all, but atomic particles). In modern times many cosmic ray satellites have been launched, and cosmic ray astronomy has become of immense importance, but it all dates back to the balloon ascents made by Hess.

Another early investigator of cosmic rays was Arthur Holly Compton (1892–1962), professor of physics at the University of Chicago. He discovered what is now known as the Compton effect—the interaction between

[1] The term 'big bang' was scornfully introduced much later by Sir Fred Hoyle, who does not believe in the theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a photon and a charged particle, so that the photon is scattered with lower energy and the particle is given extra energy. Compton shared the 1927 Nobel Physics Prize with Charles Wilson. The important Gamma-Ray Observatory satellite (GRO), launched on 5 April 1991, was named the Compton Observatory in his honor.

Several physics laureates have carried out work which is of vital importance in astronomy as well as in other branches of science, for example Professor Patrick Blackett (1948), Sir John Cockcroft and Ernest Walton (1951) and Pavel Čerenkov (1958). In the field of what may be called pure astronomy, we must note HANS BETHE (1906–). For many years the source of stellar energy was not known; initially it was thought that a star shone because it was slowly contracting, and then came a theory that atomic particles were simply annihilating each other with the release of energy. The key to the whole problem was found by Bethe in 1939—it is said that he actually worked it out during a train journey between Washington and Cornell University. The essential 'fuel' of a normal star is hydrogen. Deep inside the star, where the temperatures are very high and the pressures are immense, nuclei of hydrogen are combining to form nuclei of helium; it takes four hydrogen nuclei to make one nucleus of helium, and in the process a little energy is set free and a little mass is lost. (In the case of the Sun, the mass loss amounts to 4 000 000 tons s^{-1}; fortunately there is plenty of material left.) The process is not straightforward, and Bethe worked out a whole series of reactions on which future investigations have been based. His Nobel Prize for Physics did not come until 1967, a lapse of over a quarter of a century, so that one can hardly accuse the Nobel Committees of being over-hasty. (It is only fair to add that similar work was carried out at around the same time by GEORGE GAMOW.)

Two more pioneers in studies of stellar evolution shared the Nobel Physics Prize in 1983; the Indian astrophysicist SUBRAHMANYAN CHANDRASEKHAR (1910–1995) and WILLIAM A FOWLER (1911–1995). At one stage Chandrasekhar came into conflict of ideas with SIR ARTHUR EDDINGTON, the great British astrophysicist. A star such as the Sun will end its main career by collapsing into the white dwarf condition—that is to say it will become very small and very dense, as the companion of Sirius is at present. Chandrasekhar calculated that if the mass of the star is more than 1.4 times that of the Sun, a white dwarf cannot be formed, and the whole evolutionary sequence will be different. Eddington disagreed, but it is now known that Chandrasekhar was right, and the value of 1.4 times the mass of the Sun is still known as the Chandrasekhar limit.

In 1967 a team of radio astronomers at Cambridge University (England), led by Professor A HEWISH, was carrying out a sky survey with a special form of radio telescope which, it was said, looked rather like a collection of barbers' poles. One member of the team was a graduate student, JOCELYN BELL (now Professor Jocelyn Bell Burnell). Examining the records, she found a weak radio source which was fluctuating so quickly and so regularly that it seemed to be 'ticking'. For a while it disappeared, but then returned. At first it was regarded with scepticism and was dismissed either as instrumental fault or terrestrial interference, but eventually it was found to be real, and it had to be explained. It was even suggested that it might be an artificial transmission from an alien civilization far out in the Galaxy—and this was treated so seriously that no official announcement of the discovery was made until the LGM or little green men theory had been discounted. It is now known that the object was a rapidly rotating neutron star. As it spins, it sends out pulses of radio energy and is therefore known as a pulsar. Pulsars are now known to be very common, and a few have been optically identified; for example the Crab Nebula, known to be the remnant of the supernova seen in 1054, contains a pulsar which is seen optically as a very faint, flashing object.

In 1974 Hewish, leader of the team, shared the Nobel Prize for Physics with another outstanding radio astronomer, SIR MARTIN RYLE (who was also Astronomer Royal). That Hewish and Ryle deserved the prize is not in doubt, although it was also suggested that recognition should also have been given to Jocelyn Bell as the actual discoverer of the first pulsar.

Pulsars were again to the fore in 1993, when the Nobel Prize for Physics went jointly to RUSSELL HULSE and JOSEPH TAYLOR, of Princeton University, for their discovery of binary pulsars. In a system of this sort there are two pulsars, separated by a distance of only a few million miles. Each component is around 6 miles in diameter, but is as massive as the Sun, and they are so close that their movements can give valuable clues as to the nature of gravitation, attributed by Einstein to the curvature of space–time close to very massive bodies.

These are only a few of the Nobel laureates who have made outstanding contributions to astronomy. Yet there are also some notable omissions, and of these two are particularly curious. EDWIN HUBBLE, the American astronomer who was the first to prove that the so-called 'spiral nebulae' are outer galaxies, and who was also the first to show that the universe is expanding, never received a Nobel Prize; the Hubble Space Telescope is, of course, named after him. Neither did Sir BERNARD LOVELL, who masterminded the 250 foot radio telescope at Jodrell Bank in Cheshire (England). One could also advance the claims of Sir Arthur Eddington, HARLOW SHAPLEY (who made the first good measurements of the size of the Galaxy), GEORGE ELLERY HALE (who was responsible for the world's first really large reflecting telescopes) and the first radio astronomer, KARL JANSKY. Like the 20th century Nobel laureates, they will not be forgotten.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nobeyama Radio Observatory

Nobeyama Radio Observatory has telescopes at millimeter and submillimeter wavelengths. It was established in 1982 as an observatory of Tokyo Astronomical Observatory (NATIONAL ASTRONOMICAL OBSERVATORY, JAPAN since 1987), and operates the 45 m telescope, Nobeyama Millimeter Array, and Radioheliograph. High-resolution images of star forming regions and molecular clouds have revealed many aspects of first stages of stellar evolution, chemical evolution of molecular clouds, and galaxy evolution. Activities of very long baseline interferometry (VLBI) made it possible to realize the first space VLBI mission VSOP, and discover one of the most reliable black hole candidates in NGC 4258.

For further information see
http://www.nro.nao.ac.jp/index-e.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Node

Either of the points at which an orbit intersects a reference plane. On the celestial sphere, a node is either of the points at which the great circle representing the orbital plane intersects the great circle corresponding to the reference plane (usually the ecliptic or the celestial equator). In the context of planetary motion, the reference plane is the ecliptic. The point at which the orbiting body crosses the reference plane from south to north is called the ascending node, and the point at which it crosses from north to south, the descending node. The line joining the two nodes, which is the line of intersection between the orbital plane and the reference plane, is called the line of nodes.

In the case of the Moon, the gravitational influence of the Sun causes the line of nodes to rotate slowly around the Earth in a westerly direction, this motion being called the regression of the line of nodes. The line of the nodes of the lunar orbit makes one complete circuit of the Earth in 18.61 years.

See also: celestial equator, celestial sphere, ecliptic, great circle, orbit, orbital elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Non-gravitational Force

A force which is not gravitational in origin that acts on a celestial body to alter its orbit. The term is used in connection with comets, in which the outgassing of volatile material through vents leading to the surface of the nucleus can give rise to jets. Strong jets can behave as rockets and alter comets' orbits, in much the same way as the gravitational perturbations to which they are also susceptible. Such non-gravitational forces are believed to have changed the orbit of Comet 109P/Swift–Tuttle, which was calculated on its observation in 1862 to have a period of 120 years, but did not return until 1992. The Poynting–Robertson effect may also be regarded as a non-gravitational force.

See also: Poynting–Robertson effect, Comet Swift–Tuttle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Non-radial Stellar Pulsations

Non-radial STELLAR PULSATIONS are complex, non-radially (non-spherically) symmetric pulsations of a star in which part of the stellar surface moves outward (expands) while other parts of the stellar surface move inward (contract). Therefore, differential changes (displacements) in state variables (e.g., pressure, temperature) would follow the expansion and compression of different regions and would have positive values in some areas and negative values in others. There exists a wave equation for non-radial oscillations, just as in radial oscillations. Moreover, the angular variation of solutions is represented in terms of spherical harmonics as is the case with other important physical equations, e.g., the Schrödinger equation, Laplace's equation and others. The non-radial patterns are characterized by two integers, l and m, and correspond to the real parts of the spherical harmonic functions, $Y_l^m(\theta, \phi)$, where l is a non-negative integer, m is equal to any of the $2l + 1$ integers between $-l$ and l, and θ and ϕ are the polar and azimuthal angles, respectively. The spherical harmonics are defined as

$$Y_l^m(\theta, \phi) = (-1)^m \left(\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!} \right)^{1/2} P_l^m(\cos\theta)\, \mathrm{e}^{\mathrm{i}m\phi}. \quad (1)$$

Thus, the patterns can be described by l nodal circles, where the radial displacement $(\delta r) = 0$, with $|m|$ of these circles passing through the poles of the star and the remainder parallel to the star's equator. For $l = m = 0$, the pulsation is purely radial.

Since physical information is propagated with the velocity of small amplitude (sound) waves, non-radial pulsations correspond to sound waves travelling horizontally as well as radially, producing waves that travel around the star. Since pressure provides the restoring force for sound waves, then pulsations of this type are known as *p-modes*. The p-modes, as with radial modes of pulsation, undergo most of their motion near the stellar surface. The non-radial p-modes have their largest relative displacements, $\delta r/R$, in the outer envelope of a star.

The other major class of non-radial stellar pulsations are the *g-modes* or gravity modes. Here, gravity is the source of the restoring force for these modes, which are produced by internal gravity waves. These oscillations are intimately related to the buoyancy of the stellar gas since they involve small blobs of material exhibiting small pressure and density variations and are largely transverse in their physical dsplacement. Accordingly, g-modes can be understood by considering the displacement of a gas bubble from its equilibrium position upward by an amount $\mathrm{d}r$. If the motion occurs sufficiently slowly that its pressure remains the same as the ambient pressure and if its motion is adiabatic (no heat exchange with the surrounding gas), then the net restoring force per unit volume is

$$f_{\mathrm{vol}} = \left(\frac{1}{\rho} \frac{\mathrm{d}\rho}{\mathrm{d}r} - \frac{1}{\gamma P} \right) \rho g \, \mathrm{d}r \quad (2)$$

where P is the pressure, ρ is the gas density, γ is the ratio of specific heats, g is the gravitational acceleration (GM_r/r^2) and $\mathrm{d}r$ is the displacement.

If the quantity in the parentheses in (2) is positive, then the gas is unstable toward convection. The bubble will continue to rise. If the quantity in parentheses is negative, then the bubble will sink back toward its equilibrium position. The restoring force is proportional to the displacement and the bubble will execute oscillations about its equilibrium position in the form of simple harmonic motion. The force per unit mass or acceleration is given by

$$\frac{f_{\mathrm{vol}}}{\rho} = -N^2 \mathrm{d}r = \left(\frac{1}{\rho} \frac{\mathrm{d}\rho}{\mathrm{d}r} - \frac{1}{\gamma P} \right) g \, \mathrm{d}r \quad (3)$$

where N is the angular frequency of the bubble about its equilibrium position and is called the Brunt–Väisälä frequency:

$$N = \sqrt{\left(\frac{1}{\gamma P} - \frac{1}{\rho} \frac{\mathrm{d}\rho}{\mathrm{d}r} \right) g} \quad (4)$$

The frequency of a g-mode is determined by the average value of N in the regions of the star where the oscillations are most energetic.

The g-modes involve significant movement of the gas in the deep interior of the star whereas the p-modes have their greatest motion near the stellar surface. The use of these modes as probes of the STELLAR INTERIOR and surface layers underlies stellar seismology.

Bibliography

One of the classic technical works on non-radial pulsation theory is:

Ledoux P and Walraven Th 1958 *Variable Stars* Handbuch der Physik vol 51 (Berlin: Springer)

Outstanding introductions to the subject at the advanced undergraduate or graduate level are found in:

Collins G W 1989 *The Fundamentals of Stellar Astrophysics* (New York: W H Freeman)

Ostlie D A and Carroll B W 1996 *An Introduction to Modern Astrophysics* (New York: Addison-Wesley)

Other excellent technical reviews are found in:

Cox J P 1976 Non-radial oscillations of stars: theories and observations *Ann. Rev. Astron. Astrophys.* **14** 247

Pesnell W D 1990 Non-radial, non-adiabatic stellar pulsations *Astrophys. J.* **363** 227

Dziembowski W 1984 Theory and application of stellar seismology *Adv. Space Res.* **4** 143

Edward M Sion

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nordic Optical Telescope

The Nordic Optical Telescope (NOT) is a 2.56 m, (Super) Ritchey–Chrétien, altitude-azimuth telescope with an actively controlled primary mirror sited at 2400 m on the Roque de Los Muchachos, La Palma, Canary Islands, Spain. It was built by Denmark, Finland, Norway and Sweden, achieving first light in 1988. Iceland joined the association in 1997. Its twelve on-site astronomers, students and technical staff use and maintain the telescope, develop new instrumentation, and provide support to visiting astronomers. Core instruments on the Cassegrain-only telescope are a high-resolution imager (HiRAC) and a faint-object spectrographic camera (ALFOSC), a photopolarimeter (TurPol) and a high-resolution spectrograph (SOFIN). An infrared spectrographic camera (NOTCam) will be commissioned in 2000.

For further information see
http://www.not.iac.es/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Norma

(the Level; abbrev. Nor, gen. Normae; area 165 sq. deg.) A southern constellation which lies in the Milky Way between Lupus and Ara, and culminates at midnight in mid-May. It was named Norma et Regula (the Level and Square) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2. The shortened form was adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest star in Norma is γ^2 Normae, magnitude 4.0, which forms a naked-eye double with γ^1 Normae, magnitude 5.0. Other interesting objects include ε Normae, a wide binary with components of magnitudes 4.8 and 7.5, separation 23″, the Mira-type stars T Normae (range 6.2–13.6, period about 241 days) and R Normae (range 6.5–13.9, period about 507 days), NGC 6067, an open cluster of more than 100 stars fainter than eighth magnitude, NGC 6087, another open cluster, which contains about 40 stars between sixth and tenth magnitude (including the Cepheid variable S Normae, range 6.12–6.77, period 9.75 days), and NGC 5946, a tenth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

North America Nebula (NGC 7000)

An emission nebula in the constellation Cygnus, position RA $20^{h}\ 58.8^{m}$, dec. $+44^{\circ}\ 20'$. It is large, measuring $120'$ by $100'$. In outline it bears a strong resemblance to the North American continent, delineated by surrounding dark nebulae, including one known appropriately as the Gulf of Mexico. NGC 7000's main source of illumination is believed to be the hot blue star HR 8023. The adjacent Pelican Nebula is part of the same nebulosity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

North American Indian Astronomy

As one distinguished astronomer observed more than 20 years ago, North American Indian astronomy seems to hold a particular fascination for the North American audience 'who now occupies the land once held by a race of people so long regarded as distant noble savages' (Aveni 1980, p294). This interest has been both sustained and expanded on in the last decade or so, particularly in the case of those indigenous people who, in historic times (*c*.1600–1900), occupied the very heartland of North America — the Plains Indians.

This is a land of sun, wind and vast grasslands, stretching from north to south more than 3000 km from the Saskatchewan River in Canada almost to the Rio Grande in Mexico. The western and eastern boundaries are approximately those of the foothills of the Rocky Mountains and the Mississippi–Missouri valleys. In all, the region encompasses an area of some 2.5 million km^2.

During the 18th and 19th centuries, when contact was first made with the white man, the Great Plains were largely dominated by Siouan and Algonquian linguistic groups although others, such as Uto-Aztecan and Athapaskan, were also represented. In the northern part of the region (now Alberta, Saskatchewan and Montana) lived such tribes as the Blackfeet, Cree, Plains Ojibwa and Gros Ventre; to their south (present-day Wyoming, southern Montana and the Dakotas) were the Sioux, Cheyenne, Crow, Mandan, Hidatsa and Arikara; the Southern Plains (present-day Texas, Nebraska and Oklahoma) were occupied by the Comanche, Kiowa, Wichita and Pawnee. Most of these tribes had migrated into the region after the introduction of the horse by the Spaniards in the southwest and of the gun by the English and French in the northeast — a gradual process which started in about 1650. It brought together peoples of great diversity of background and history — not least the way their ancestors incorporated astronomical knowledge into their culture.

There is considerable evidence which suggests that many tribal groups who occupied the various cultural areas of North America (it is convenient to consider nine cultural areas in North America, north of Mexico; see Sturtevant and Taylor 1991) and led largely sedentary lifestyles did accurate timing relating to planting and harvesting by observation of the regular celestial events. Hunters and gatherers, however — and this would include the Plains Indians — put more emphasis on use of the changing sky largely to order ritual and religious activities.

Other phenomena, such as comets, solar and lunar eclipses, meteor showers as well as terrestrial events — the thunder and lightning in particular — were not infrequently incorporated in religious and ceremonial activities. Symbols of such phenomena also adorned artefacts of various kinds.

The North American Indians, however, viewed the canopy of complex stellar patterns which nightly unfolded above them (in an ever changing but largely cyclic pattern) somewhat differently to the professional astronomer. Sophisticated reflective, refractive and radio instrumentation tends to sample but relatively small regions of the sky. In contrast, naked eye astronomy, as practiced by American Indians, led to the idea of constructing accurate calendars or patterns which could give order to planting and harvesting, hunting, ceremonial, ritual and religion.

A common form of religious activities throughout the Americas — and indeed a theme common to all humanity — was annual renewal ceremonials. Many were complex, lasting for several days, even weeks. There is considerable evidence that astronomical phenomena played an important role in many of them.

For example, scholars who have documented the lifeways of the Mahicans (these are probably the Mohicans of James Fenimore Cooper; they were an Algonquian-speaking group who, in ancient times, occupied the Hudson River Valley of present-day Vermont) recorded that they traditionally enacted a complex world renewal ceremonial which commenced at signals from the sky. The changing sky patterns associated with the constellation Ursa Major was viewed as an annual celestial bear hunt. Four of the stars defined the body of the bear while three stars that form the handle of the Dipper represented waiting hunters, the star Alcor being their dog. This 'cosmic bear' was observed to revolve around the North Star and then, in spring, seen to leave his den — the Corona Borealis. It was now followed by seven stars representing hunters, the original three being joined by stars from Boötes and Arcturus. Throughout the summer, this bear was trailed and then finally slain in the autumn. The seasonal reddening of the forest foliage was attributed to this sky hunt, the red tinting due to that of the slain bear, and the mantle of early winter snow was some of the bear's grease, tried out by the star hunters.

The ceremonials were held at night within a special structure — the Big House, *Xwate'k'an*, which itself was replete in sky symbolism. Thus, the constellation Ursa Major was represented on the floor of the house. The furnishings and positions of officials corresponded to the positions of the stars in the constellation, while the movements and acts of the performers paralleled those of the stars. In turn, the *Xwate'k'an* was said to represent the universe. This renewal ceremonial commenced again the following spring when bear once more emerged from his celestial den.

Further west, in what is now present-day Nebraska, the Pawnee — particularly one branch of the tribe known as the *Skidi* — were much governed by the panorama of the changing sky. When the villages of the various Pawnee bands assembled for annual ceremonials, they arranged themselves according to the place of their stars in the sky and nothing on Earth could move without evoking the power of the heavens. The whole 'Spring

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Awakening' could not be undertaken without reference to the position of the stars to guide the time relating to the commencement of the appropriate ceremonials. The earth lodge itself acted as a type of astronomical observatory, the priests viewing the sky through both the smoke hole and the oriented entrance way.

Astronomical and cyclic patterns were documented and understood by the tribal intellectuals and woven into mythology, religion and ceremonial. Thus, Morning Star, *Opirikus* — almost certainly Mars — was viewed as the leader of men who travelled with his brother, the Sun, to the land of the western stars in an attempt to overcome the power of the Moon. Credited with helping the creation of the universe and the fathering of the human race, Morning Star utilizes the power of sacred bundles (housed within the ceremonial lodges) to overcome the obstacles which Moon puts in his way and is thus able to enter the woman-star village. This aspect of Pawnee mythology was clearly based on accurate and detailed observations of the sky.

Such mythology and religious concepts, however, led to one sinister ritual. Because of his favours to mankind, and particularly to the Pawnee themselves, *Opirikus* demanded human sacrifice — the Morning Star Ceremonial. Tribal priests stated that the sacrifice not only ensured continued abundant crops and game but also, perhaps more than anything, was an acknowledgement of a celestial heritage and essential to the well-being of the tribe — although not all agreed.

While the record relating to the sky watchers' craft is far less complete for tribes north of Mexico, much significant evidence is scattered within the historical and anthropological literature. Here, there are references made by North American Indians to the nature of gravitational forces and the Milky Way, to meteor showers, eclipses of the Sun and Moon and the images of the lunar landscape. Star groups, such as the Great Bear and Pleiades, the diurnal motion of stars across the sky and the heliacal rising of such stars as Aldebaren and Sirius figured as calendrical markers for ceremonial.

Some tribes, such as the Pawnee and Lakota, did star maps as well as recognizing certain geographical features as a mirror of the sky. Others used stone alignments as markers to identify the directions of set or rise of certain celestial bodies; all this, and almost certainly much more.

As such data are drawn together and analysed, they will surely demonstrate that astronomical phenomena were, not surprisingly, an important component of the indigenous Americans' culture. (This is partly based on Taylor (2000 and 2001, in press). I am indebted to both Paul Murdin and Patrick Moore for their encouragement for me to give consideration of the astronomical knowledge of the North American Indian, in the course of my researches in American ethnology.)

Bibliography

Aveni A F 1980 *Skywatchers of Ancient Mexico* (Austin, TX: University of Texas Press)

Sturtevant W C and Taylor C 1991 *The Native Americans* (London: Salamander)

Taylor C 2000 Astronomy of the Plains Indians *2001 Yearbook of Astronomy* ed P Moore (London: Macmillan) pp147–61

Taylor C in press Guided by the stars? The Big Horn medicine wheel in northern Wyoming *2002 Yearbook of Astronomy* ed P Moore (London: Macmillan)

Colin Taylor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Nostradamus (1503–66)

Doctor, astrologer, born in St Rémy, France. Took on the role of a prophet and wrote *Centuries*, a collection of predictions in rhyme (1555–8). The predictions are expressed in obscure and enigmatic terms, which are both difficult to interpret and open to many interpretations, and so can be interpreted as successful prophesies, including what Catherine of Medici interpreted as the manner of death in a jousting tournament of her husband King Henry II. At the royal court in France, this brought their author a fame that continues today as notoriety.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nova V1974 Cygni

A NOVA outburst is the second most violent explosion that can occur in a galaxy. It is exceeded in violence only by a SUPERNOVA explosion, but a nova occurs much more frequently so that astronomers expect a bright naked eye nova to appear in our skies every few years. In a day or two a nova can go from an intrinsic brightness much fainter than that of the Sun to more than 10 000 times that of the Sun. At the same time it is ejecting an amount of gas that exceeds the mass of the Earth at speeds of up to 5000 km s^{-1}. The gas ejected into space by the nova explosion will then mix with other gas in the interstellar medium (ISM) and eventually become part of a new solar system. There is evidence that our own solar system contains material that was once ejected in a nova explosion.

Evolution at ultraviolet wavelengths

V1974 Cygni was first seen as a naked eye star on the morning of 19 February 1992 and rapidly climbed to a peak brightness of about 4th magnitude. Because of its brightness, it was studied by more telescopes at more wavelengths than any nova before or since. The first spectra were obtained in the ultraviolet with the INTERNATIONAL ULTRAVIOLET EXPLORER satellite (IUE). They covered the wavelength range from 1150–3300 Å. They were unlike any previous IUE nova spectra obtained early in the outburst. While energy was being emitted throughout the ultraviolet, there was an increasing amount of energy at longer wavelengths. Analysis of these spectra showed that the nova had been caught in the 'fireball' phase when the hot dense material ejected by the explosion is in its first cooling phase. Because the gas was cooling, the amount of energy emitted at ultraviolet wavelengths decreased and by the second day of observations, it had dropped by a factor of 10. The rapid decline in ultraviolet light is caused by the temperature of the hot, highly ionized, expanding material dropping below 30 000 K. The ultraviolet data show that the nova had reached this temperature at the time of the first IUE observations. Below this temperature, the electrons start to recombine with the ions and form both neutral and singly ionized atoms of the iron group elements (among others) which are capable of blocking the ultraviolet light through tens of millions of individual absorption lines.

It is the gas becoming opaque in the ultraviolet which forces the radiation from the nova to emerge at optical and infrared wavelengths and then become bright in the optical. This is the time that a nova, typically, is first discovered. Luckily, V1974 Cyg was caught before maximum brightness in the optical. By the time V1974 Cyg reached maximum brightness in the optical, its temperature had declined to below 15 000 K and it was emitting most of its energy in the optical and infrared, not the ultraviolet. This phase did not last long and within a few days the optical brightness started to decline. This decline is caused by the continuing drop in density as the ejected gases expand and the outermost layers gradually become transparent. The increasing transparency allows us to see deeper into the gas, where the material is hotter, and thus the decline in optical light goes hand-in-hand with an increase in ultraviolet light. Over the next two months of IUE observations, the ultraviolet brightness of V1974 Cyg increased to the levels seen on the first spectra. This result demonstrated that the energy radiated by the underlying star had remained virtually constant during this time. This period of the nova outburst is designated as the 'constant bolometric luminosity phase' and its existence was an early prediction of the theoretical simulations of the nova outburst.

Evolution at x-ray wavelengths

Further evidence that the heat from the underlying layers was getting through the ejected material came from observations in April 1992 with ROSAT, a joint German–British–US x-ray satellite launched in 1990. The first observations showed that the nova was faint in x-rays but in a series of observations done over the next 12 months it steadily brightened. It was later realized that ROSAT had detected the underlying hot object which was gradually appearing through the expanding shell of gas ejected by the explosion. By July 1993, V1974 Cyg had become one of the brightest objects in the low-energy x-ray sky. It was emitting nearly as much energy in x-rays as it had in the optical and ultraviolet a year earlier in the outburst. The ROSAT observations thus provided important new data on the constant bolometric luminosity phase of the outburst. It then began to fade rapidly and by December 1993 it had become too faint to observe with ROSAT.

Observations with the Hubble Space Telescope

V1974 Cyg was also observed a number of times with various instruments on the HUBBLE SPACE TELESCOPE (HST). Analyses of the first HST observations showed that the nova had not ejected material in a smooth uniform sphere but rather that the expanding gas had formed a large number of dense knots inside a lower density, more rapidly expanding gas. The ejected gas was distributed in a non-spherical hollow shell with a thickness of order 30–50% of the outer radius. In a second set of observations, it was found that the same knots were still visible and were moving at approximately the same speeds as when first seen seven months earlier. However, the fast moving, low density, material had largely vanished. This means that the HST observations were now seeing completely *through* the ejected gas. As the gaseous debris from the explosion expanded, it thinned and continued to clear so that it became possible to see completely through it to the underlying object on which the explosion took place. It then became possible to study the structure of this object and determine the cause of the explosion.

Cause of the nova outburst

Studies of V1974 Cyg after the explosion showed that it has the same structure as all other novae and the entire class of stars designated 'cataclysmic variables.' More than 40

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

years ago it was demonstrated that novae are CLOSE BINARY STAR systems with one component a white dwarf and the other component a larger, cooler star. The two stars have orbital periods of a few hours, about 2 h for V1974 Cyg, so that tidal forces of the WHITE DWARF are sufficient to cause material to be stripped off the larger star. Because the binary is rotating rapidly, the gas forced off the cooler star forms a torus of gas around the white dwarf called an ACCRETION DISK. The gas slowly spirals through the accretion disk and ultimately falls onto the surface of the white dwarf. Over many thousands of years the gas flowing from the companion, through the accretion disk, lands on the white dwarf star where it is compressed and heated by the continuous accretion of more and more material. It is the gradually thickening layer of accreted gas on the surface of the white dwarf that provides the fuel for the explosion. It was also realized later that the accreted material must be mixed with the core material of the white dwarf since such gas is observed in the ejected gas.

The detailed calculations of the consequences of this accretion plus mixing process show that, if sufficient material falls onto the white dwarf, then compressional heating alone can cause the temperatures in the accreted gas to reach to 10^7 K or higher. At this temperature, hydrogen nuclear fusion reactions begin in this material and, because fusion is occurring on a white dwarf, the reactions quickly increase to high temperatures, just as in a terrestrial hydrogen bomb. This type of explosion is called a *thermonuclear runaway* (TNR). The white dwarf must accrete about 100 times the mass of the Earth before the temperature in the newly added material can grow to the temperatures required for initiating the TNR. By this time, the density in this layer will exceed 10 000 g cm^{-1}. Under these conditions the hydrogen nuclei fuse into helium and release energy via the carbon, nitrogen and oxygen (CNO) cycle.

Since carbon and oxygen are catalysts in the CNO cycle, the higher the percentage of carbon or oxygen in the nuclear fusion region, the greater the energy production and the greater the ultimate violence of the explosion. Since the accreted layers mix with the core material of the white dwarf, which is mostly carbon and oxygen, a violent explosion is guaranteed. One concern was that the heat from the nuclear fusion should cause the outer layers to expand and cool and prevent an explosion from occurring. That this does not happen was explained by numerical calculations which showed that the gas at the bottom of the accreted layers had become degenerate.

A degenerate gas is so dense that the pressure comes from the inability of electrons to interpenetrate each other and, unlike an ordinary gas, the temperature does not play much of a role in determining the pressure. The gas can still become hot, but it cannot expand as one would expect if the gas were normal. It is also known that the speed at which the CNO nuclear reactions occurs depends on the temperature to a very high power (between 16 and 18). Thus, even a small increase in temperature can cause a tremendous increase in the speed of the reactions between the hydrogen nuclei and the CNO nuclei. This increase in speed of the reactions acts to release heat even more rapidly into the gas. As long as the region where nuclear fusion is occurring does not expand and cool, the rate of energy release must rise. This is a vicious cycle; as the temperature increases, the rate of energy release into the layers also increases, which in turn causes a rise in temperature and more energy to be produced.

Something new happens when the temperature in the nuclear fusion region grows to values above 30×10^6 K. Nuclear energy is being produced so rapidly that the rate of energy transport by radiation is insufficient to carry the heat away from the region of peak nuclear fusion deep in the accreted layers. Even though the material is degenerate it can still become buoyant and it starts to mix with the zones above it. This is called convection. As the temperature grows from the continuous release of nuclear energy deep within the accreted layers, the newly formed convective region grows toward the surface carrying both heat and the products of the nuclear fusion reactions. Since the accreted material mixes with core material at some timeduring the explosion, convection then mixes core material to the surface. Thus, one of the predictions was that, when these layers were blown into space by the explosion, they would carry both white dwarf core gas and the nuclear reaction products along with the accreted layers. This prediction could be tested since once the ejected material became sufficiently dispersed, it would be possible to analyse it and determine its chemical composition. If this gas actually showed that core material were present, then the analysis would tell astronomers not only what happened inside the white dwarf star during the explosion but also shed light on the prior evolution of the white dwarf. This prediction has been verified by a number of recent studies.

Production of radioactive isotopes

Other calculations predict that two important nuclear ISOTOPES were produced during the explosion on V1974 Cyg. The first is an isotope of sodium, ^{22}Na, and the second is an isotope of aluminum, ^{26}Al. Each of these is radioactive with a distinct γ-ray signature and a relatively short half-life. The importance of ^{26}Al is twofold. First, its decay product, ^{26}Mg, is found in some classes of meteorites and suggests that material has been found that has existed virtually unchanged since before the formation of the solar system. It is also thought possible that the heat produced by the decay of the aluminum isotope could be responsible for heating the asteroids and other small bodies early in the evolution of the solar system. Second, the γ-rays produced by the decay of ^{26}Al have been seen throughout the Galaxy by both the SOLAR MAXIMUM MISSION and by the COMPTON GAMMA RAY OBSERVATORY. The source of the interstellar ^{26}Al is not known, but novae are one candidate. The existence of novae that eject large amounts of aluminum has now been verified by observational studies of V1974 Cygni. Unfortunately, one can only determine the total amount of aluminum, not the isotopic abundance in the ejected

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

material. ^{22}Na, the other radioactive isotope produced in the explosion, should have been detected by the Compton Gamma Ray Observatory. It pointed at it in September 1993, however, but no γ-rays were detected. Since other studies have shown that the nova is at a distance of 2000–4000 ps, it may have been too far to be detected in γ-rays since theoretical studies imply that a reasonable limit to detection is about 1000 ps.

Finally, this nova has provided the most important and complete datasets ever obtained for a nova in outburst and it is still sufficiently bright that it is continuing to be observed by astronomers using both observatories in space and on the Earth.

Bibliography

Gehrz R D, Truran J W, Williams R E and Starrfield S 1998 Theoretical and observational studies of the contributions of galactic classical novae to the interstellar medium *Publ. Astron. Soc. Pacific* **110** 3

Shore S N, Starrfield S, Ake T B and Hauschildt P H 1997 Spatially resolved spectra of Nova V1974 Cygni 1992 with the Goddard High Resolution Spectrograph *Astrophys. J.* **490** 393

Starrfield S, Gehrz R D and Truran J W 1997 Dust formation and nucleosynthesis in novae *Astrophysical Implications of the Laboratory Study of Presolar Grains AIP Conf. Proc.* **402** ed T Bernatowicz and E Zinner (American Institute of Physics) pp 203–36

Starrfield S and Shore S N 1994 Nova Cygni 1992: nova of the century *Sky and Telescope* **87** 2–25

Starrfield S and Shore S N 1995 The birth and death of Nova V1974 Cygni *Sci. Am.* Jan. 76–81

Starrfield S, Truran J W, Wiescher M and Sparks W M 1998 Evolutionary sequences for V1974 Cyg using new nuclear reaction rates and opacities *Mon. Not. R. Astron. Soc.* **296** 502–21

Sumner Starrfield

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Novae

A nova, short for *nova stella* (new star), is a star that brightens suddenly several hundred- to a million-fold, remains bright for a few days to several months and then returns to its former, low luminosity. The spectrum, too, changes considerably along with the changing luminosity. The eruption is recurrent, although for most novae the time elapsed between outbursts is thousands to tens of thousands of years, and hence only one outburst is recorded. These are often referred to as classical novae, to be distinguished from recurrent novae that erupt at intervals of tens of years, so that a number of such outbursts have been recorded for each. Nova outbursts are accompanied by mass ejection and the formation of nova shells, which slowly disperse into the interstellar medium. Although the term 'nova' refers to the VARIABLE STAR that undergoes temporary explosive eruptions, it is sometimes used to designate the outburst itself (as in the case of SUPERNOVAE). Thus terms such as prenova, post-nova, or nova progenitor are quite common.

Historical record

In ancient times, novae were classed with the guest stars, which also included supernovae, as well as COMETS, all transient objects. Early observations of such objects were made mostly in the Far East—China, Japan and Korea—where professional astronomers (astrologers, in fact) were employed by rulers to constantly watch the sky for signs of impending dangers. Meanwhile, ancient and even medieval Europe showed little interest in these temporary stars, which were in marked conflict with the dominant Aristotelian doctrine of a perfect, immutable celestial sphere. Detailed records from China go back to about 200 BC; in Korea and Japan, regular observations began around AD 800. While comets were quite early suspected and then recognized as being a separate class of objects, the distinction of novae from supernovae was made only in the 1930s, when it was realized that the two differed in maximal brightness by about 6 orders of magnitude. As novae spend most of their time in the quiescent, low-luminosity state, old novae are difficult to detect; the oldest nova that has been recovered was discovered in 1670 in the constellation Vulpecula.

By the early 1960s ample observational evidence had accumulated, mainly through the work of Robert Kraft, indicating that novae were invariably members of close binary systems (see CATACLYSMIC BINARIES: CLASSICAL AND RECURRENT NOVAE). The nova companion was found to be a low-mass main sequence star. Observations of novae after eruption, and in a few cases prior to eruption, showed them to be hot compact stars. Mass estimates, albeit scarce and uncertain, suggested that the erupting stars were WHITE DWARFS. This led to the hypothesis that the RED DWARF companion is extended enough (i.e. fills its ROCHE LOBE) to allow mass transfer to the hotter star (through the inner Lagrangian point of the binary system's gravitational field). Indeed, in some cases, a rapidly rotating region,

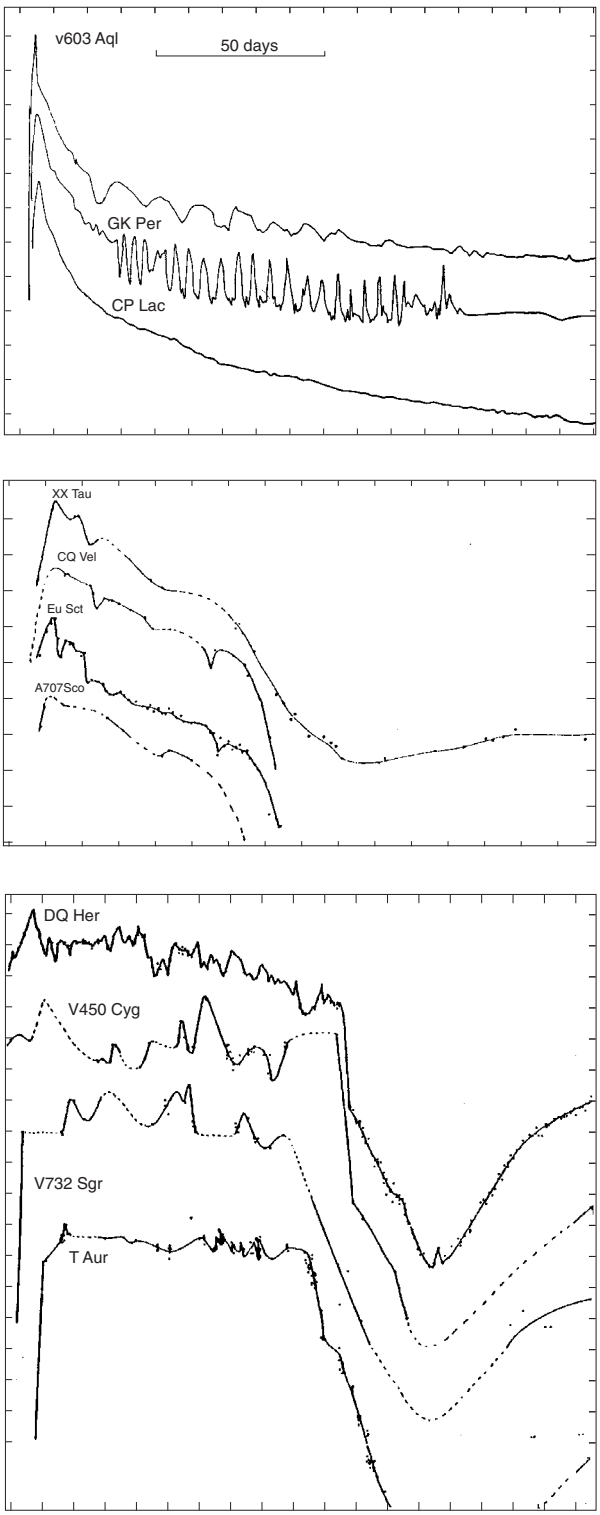


Figure 1. Light curves of (top) very fast, (middle) moderate and (bottom) slow novae. Abscissae are marked at intervals of 10 days and ordinates at intervals of 1 magnitude. (Adapted from Payne-Gaposhkin (1957).)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

identified as an ACCRETION DISK, was detected around the hot star.

Thus, novae appear to be hot white dwarf members of close binary systems, which accrete matter from a cool red dwarf companion. This sets the scene for the theory that explains the outburst mechanism, its many distinctive features and its consequences.

Observations

Rate and distribution

Around the turn of the 20th century, the number of nova discoveries rose considerably, but since then it has remained fairly constant, at an average of ∼4 Galactic novae yr^{-1}. There appear to be two populations of novae: disk novae, observed in the solar neighborhood, which are brighter and evenly distributed, and bulge novae, concentrated towards the Galactic center with a larger scale height, which are fainter and have a generally slower outburst development. One should note that the bulge novae are subject to severe observational selection (interstellar extinction, etc). The true Galactic nova rate is estimated at $\sim 40 \pm 20$ yr^{-1}. For comparison, the nova population of the Large Magellanic Cloud (LMC) consists almost exclusively of disk novae, while in spiral galaxy M31 (Andromeda), mainly the novae in the pronounced bulge are observed. The established nova rate in M31 is about 30 yr^{-1}. Much lower rates are derived for M33, about 5 yr^{-1}, and for the LMC, about 2 yr^{-1}. Considering the galactic rate of nova outbursts on the one hand and the restrictive requirements for a system to undergo a nova outburst on the other, one arrives at the inevitable conclusion that nova outbursts must recur in the same system a great many times, as was realized already in the late 1930s. Most of the time, however, is spent in quiescence, while the white dwarf accretes mass from its companion.

Light curves and spectra

Early studies of novae already revealed that they varied considerably in the rate of evolution after maximum. Following Dean McLaughlin, novae are classified according to the time of decline from maximum light, usually by 3 magnitudes, denoted by t_3. Thus three—and sometimes five—speed classes are identified: very fast novae, with $t_3 \lesssim 10$ days, slow novae with $t_3 \gtrsim 100$ days and one to three intermediate classes in between. However, these are not physically distinct classes of objects, but rather convenient groupings; the variation of t_3 is continuous.

Light curves of novae belonging to different classes are shown in figure 1 and a characteristic spectral evolution is illustrated in figure 2. Several stages, distinguished by optical brightness variations and typical spectra, may be identified in the evolution of a nova outburst. Their characterization, dating back to the work of McLaughlin in the early 1940s, has remained practically unchanged to this day.

(a) The prenova stage is characterized by weak brightness variations and the spectrum of a hot star. There appears to be a small increase in (visual) luminosity 1–15 yr prior to eruption.

(b) The premaximum stage consists of a steep rise to maximum visual brightness and has a characteristic premaximum spectrum: a strong blue continuum, blueshifted absorption lines and weak or absent emission lines. The line shifts indicate lower velocities than those observed after maximum. The evolution of the blueshifted absorption line spectrum indicates an expanding and cooling photosphere. During final expansion towards maximum, the photospheric temperature declines rapidly, reaching values down to 9000–4000 K for different novae. Spectral types are B to A and sometimes F. Three successive phases are generally distinguished: the initial rise that usually lasts for a day or two, the brief premaximum halt, which in a few cases may extend to months, and the quick final rise.

(c) The maximum stage includes the early decline with the appearance of the principal spectrum, soon followed by the diffuse enhanced spectrum. The optical maximum is defined by the peak in the optical light curve. Near maximum and during early decline, nova colors are similar to those of bright supergiants. The luminosity in most novae remains constant for some time after maximum light. This is confirmed by observations in the UV, where the radiation maximum of the hot photosphere is located, and, in the presence of dust, by IR measurements of re-radiated light from dust grains that form at the outer edge of the expanding nova shell. The principal spectrum is characterized by strong P Cygni lines superimposed on a relatively cool and very strong continuum, which soon starts fading, shifting into the blue. Emission lines of hydrogen and metals in low ionization stages are strong. Their complex profiles suggest that material has been ejected in blobs, cloudlets or rings. Particle velocities deduced from the blueshifts lie in the range of a few to several hundred km s^{-1} for slow novae and around a thousand km s^{-1} for fast novae. The diffuse enhanced spectrum appears 1 or 2 days after maximum in fast novae and 1 or 2 weeks after maximum in slow ones. It is characterized by diffuse emission and absorption lines of abundant elements, such as C, N, O, Si, S and Ne. The line strengths of some elements suggest high abundances. The derived velocities are high and the large line widths indicate large velocity dispersions.

(d) The transition stage, during which the decline continues, is characterized by the Orion spectrum, line flaring and coronal lines. Magnitude measurements during this period reveal a smooth decline, or oscillations with typical timescales of the order of days or a deep decline with a possible recovery after a few weeks. The decline is attributed to the formation of dust and the light loss is due to absorption. The color

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

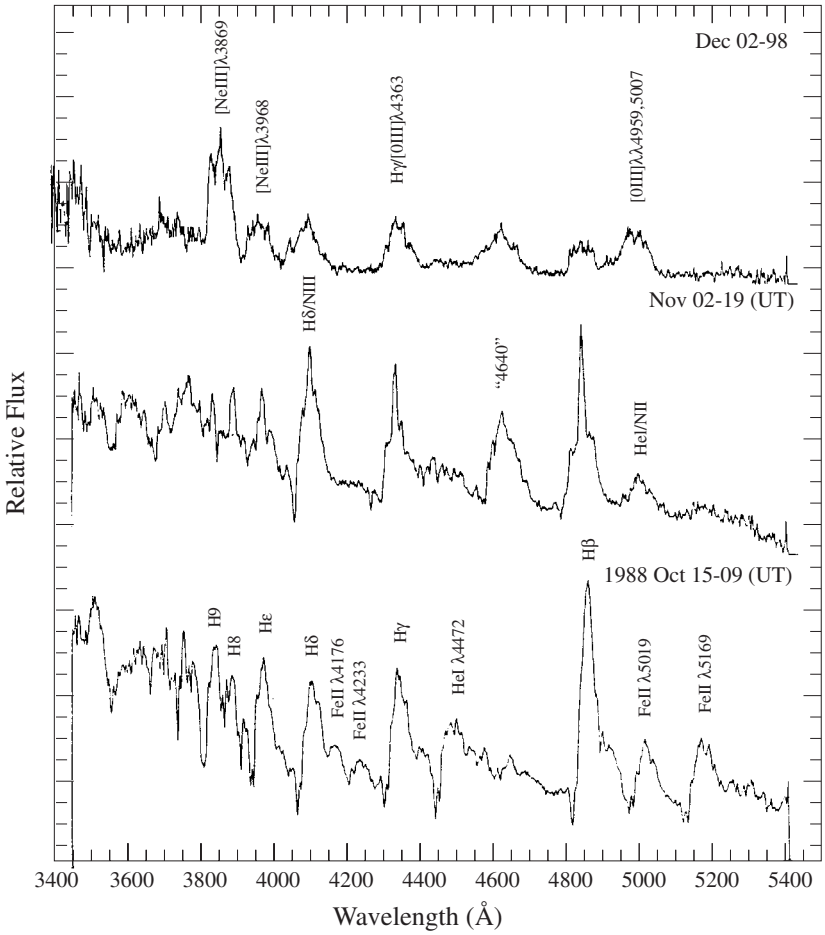


Figure 2. Illustration of the evolution of a nova spectrum (Nova LMC 1988 No 2, a very fast nova with $t_2 \approx 5$ days): (bottom) 1.34 days after maximum, showing the principal and diffuse-enhanced spectrum; (middle) 19.44 days after maximum—a typical Orion spectrum; (top) 50.23 days after maximum, showing an early nebular spectrum. The ordinate is an arbitrarily normalized flux. (From Sekiguchi K, Kilkenny D, Winkler H and Doyle J G 1989 *Mon. Not. R. Astron. Soc.* **241** 827.)

effects of dust absorption are complex and require different particle sizes at different times. As the expanding shell grows optically thin, the photosphere recedes towards deeper and hotter regions. Velocities derived from the broad spectral lines are even higher than before. An even stronger sign for a thinning atmosphere is the advent of forbidden lines. Infrared spectra of some novae at this stage are dominated by strong coronal lines.

(e) The nebular stage that follows is characterized by forbidden lines. The nova has now become fainter by several magnitudes. The photosphere is small and radiates mainly in the UV and x spectral ranges. The nebula that has formed around the nova is clumpy; its mass is estimated to be of order $(10^{-5}$–$10^{-4})M_{\odot}$.

(f) The post-nova or quiescence stage is reached when the nova returns to its pre-outburst appearance. Photometric and spectroscopic observations indicate that the accretion disk has been regenerated. The ejecta now appear as a resolved shell made of numerous blobs, as shown in figure 3 for two novae. It also appears that different parts of the ejecta have noticeable abundance differences. Observations in the UV range by the IUE satellite and, for a few novae, in the x-ray range, first by the EXOSAT satellite and later by ROSAT, indicate that the hot star remains luminous for several years after outburst. Furthermore, a correlation emerges between the decline time of the UV luminosity and the time span of the much earlier decline of the visual luminosity (t_3). On observational grounds, little can be said concerning timescales of more than a hundred years, with the possible exceptions of two old novae that have been recovered, CK Vul of 1670 and WY Sge of 1783, both very faint.

Despite the similarity in spectral and brightness evolution, novae exhibit an overwhelming variety of properties, which may differ by orders of magnitude

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

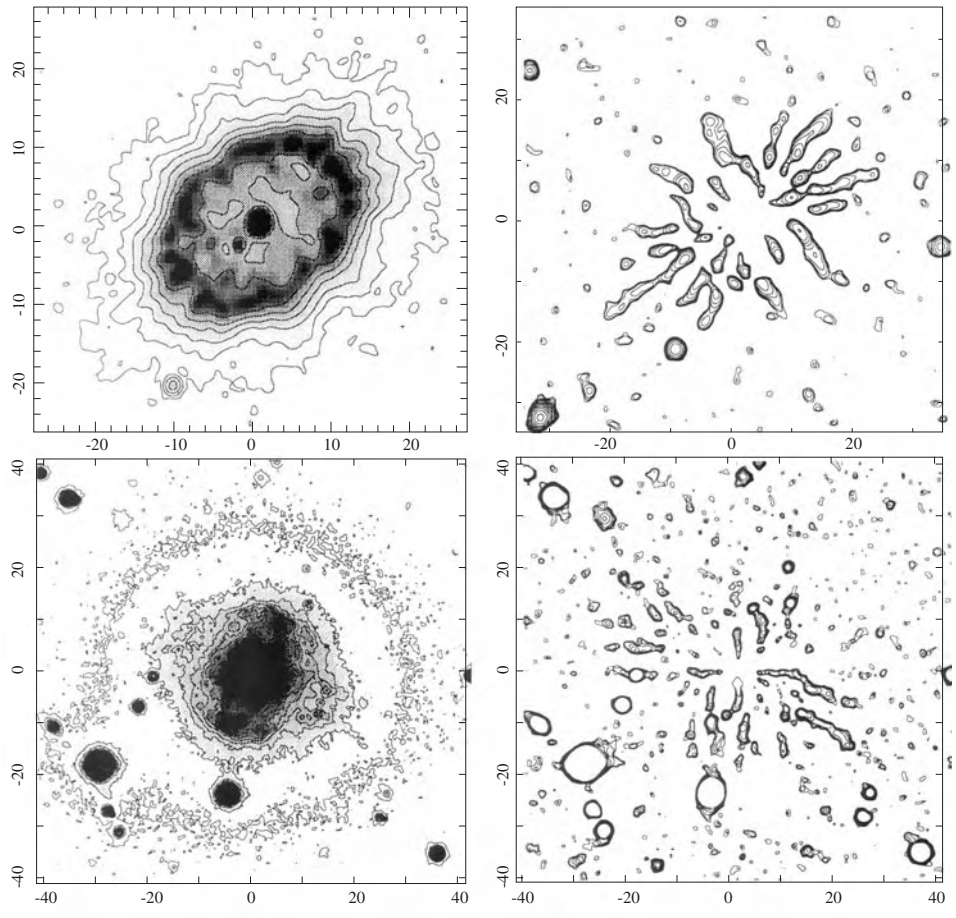


Figure 3. Hα images of two nova shells: (top) DQ Her and (bottom) RR Pic, exhibiting the typical, highly clumpy, structure. The right panels are contour plots obtained by removing the elliptically symmetric component. Axes are marked in arcsec. (From O'Brien T J and Slavin A J 1996 *Cataclysmic Variables and Related Objects* ed A Evans and J H Wood (Dordrecht: Kluwer).)

between one nova class and another, and even between individual novae within the same class. These must be accounted for by the nova theory.

Theory

The outburst mechanism

Theoretical studies of the nova outburst started with the work of Evry Schatzman in the late 1940s and 1950s, in which explosive hydrogen burning was first proposed as the triggering mechanism. Hydrogen accretion onto a white dwarf and the ensuing unstable burning under conditions of electron degeneracy were first studied by Leon Mestel in 1952. It is noteworthy that all this preceded the discovery that novae were members of binary systems. More than 20 yr later, it became commonly accepted that classical nova outbursts result from explosive hydrogen burning—thermonuclear runaways—on the surfaces of white dwarfs that are accreting hydrogen-rich material from red dwarf (late-type) companions in close binary systems. The process is cyclic, with the greatest part of the cycle spent in quiescence. A quite detailed picture of the evolution of such cyclic outbursts emerges from the numerous theoretical studies—numerical hydrodynamical evolutionary calculations—that have been carried out since the late 1960s.

The material gradually accumulating on the white dwarf's surface becomes compressed and the electrons at the bottom of the accreted envelope become partially degenerate. At the same time the temperature at the bottom of the hydrogen-rich envelope rises. When it reaches 2×10^7 K, hydrogen is ignited in a thin shell by the CNO nuclear reaction cycle. The energy released raises the temperature further, but since the degeneracy pressure is insensitive to temperature, no expansion and cooling results and the temperature keeps rising exponentially, boosting the nuclear reaction rates in a runaway process. However, above about 10^8 K, the CNO cycle rate is limited by the decay rates of β^+-unstable nuclei, which are temperature independent. In addition, the temperature becomes sufficiently high for the degeneracy to be lifted, turning the gas into an ideal one. The runaway is thus quenched: the shell expands, while the burning

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

temperature starts dropping after having reached a few 10^8 K. Following ignition, a convective region forms just above the thin shell source and extends towards the surface. Convection mixes the β^+-unstable nuclei throughout the envelope and brings fresh CNO nuclei into the burning shell, until the unstable isotopes prevail over the extent of the envelope. Energy generation continues, supplied by the decay of the β^+-unstable nuclei. Heat absorption now results in rapid expansion and cooling of the accreted envelope.

During the runaway, the white dwarf's luminosity rises until it attains—or even surpasses briefly—the Eddington critical luminosity, $L_{\rm Edd} = 4\pi cGM_{\rm wd}/\kappa$, where $M_{\rm wd}$ is the star's mass and κ is the photospheric opacity. When the star reaches bolometric maximum, its radius is still relatively small and its surface still hot; hence it radiates mostly in the UV or even EUV. A relatively long phase of constant bolometric luminosity, a luminosity plateau, ensues. The rapid expansion of the envelope proceeds now at constant, close to critical, luminosity. Hydrostatic equilibrium cannot be achieved; instead, mass is driven out by radiation pressure in an optically thick wind. The importance of radiation pressure for the ejection of matter by novae was first recognized by William McCrea, as early as 1937. The idea was later pursued in a vast number of analytical and numerical studies of optically thick steady winds. These calculations yield velocities of the expanding material, mass loss rates and the total ejected mass. Expansion velocities vary between a few hundred km s^{-1} for slow novae to a few thousand km s^{-1}, for very fast ones.

The optical maximum is obtained when the maximal photospheric radius is reached, of order $100R_\odot$, corresponding to effective temperatures of several thousand kelvins. Thereafter, when the envelope becomes highly diluted, and the opacity drops, the photosphere recedes through the expanding mass; its radius decreases, while the effective temperature rises. Thus the evolution traces a double horizontal path in the H–R diagram, from high to low effective temperatures and back, as shown in figure 4.

When most of the envelope has been ejected, mass loss comes to an end. The small remnant shell on the white dwarf's surface contracts and then starts cooling slowly. Hydrogen burning continues, in thermal equilibrium, with the nuclear and bolometric luminosities both slowly declining, now at constant radius. When almost all of the hydrogen in the remnant shell has turned into helium, nuclear burning ceases and the white dwarf returns to its pre-outburst state. The decline takes roughly one to several years and the white dwarf remains almost unaffected by the outburst that has taken place. Accretion resumes towards the next outburst and a new nova cycle begins.

The quiescent state

The mass transferred from the red dwarf carries substantial ANGULAR MOMENTUM. Therefore, it does not settle immediately on the surface of the white dwarf, but forms a ring around it, which is spread by viscosity into a disk. Accreting material spirals in through the disk, releasing gravitational energy, which is processed and radiated by the disk. The disk luminosity—essentially, accretion luminosity of order $GM_{\rm wd}\dot{M}/R_{\rm wd}$, where $R_{\rm wd}$ is the radius of the accreting white dwarf and $\dot{M}$ is the accretion rate—dominates the luminosities of both binary components between outbursts. Thus the visible spectrum is the featureless blue continuum of the accretion disk, typical of cataclysmic variables in general.

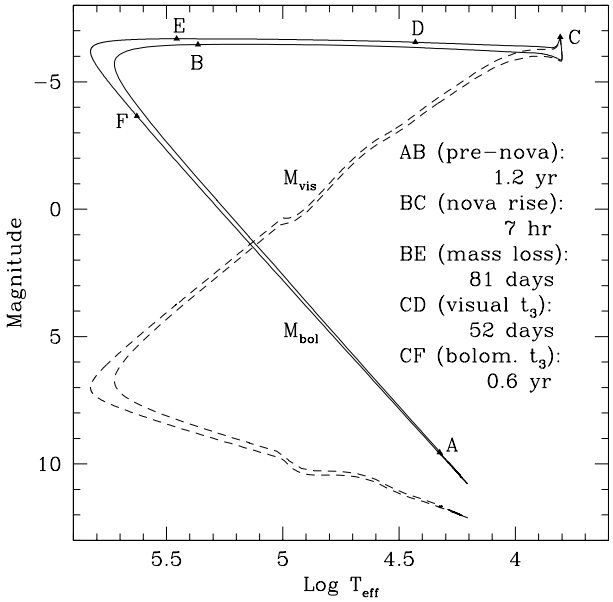


Figure 4. Illustration of the nova evolution pattern in the H–R diagram, as obtained from model calculations. Bolometric and visual magnitudes are indicated. Significant evolutionary landmarks are labeled. (Adapted from Kovetz (1998).)

From the space density of cataclysmic variables obtained in galactic surveys, of order 10^{-6} pc^{-3}, combined with the galactic rate of nova outbursts, an upper limit of about 10^3 yr is derived for the recurrence time. This demands, in turn, high mass transfer rates, which are incompatible with those inferred from nova theory. A solution to this apparent conflict is a variable mass transfer rate between outbursts: a high rate for tens to hundreds of years after outburst, which declines thereafter to lower values maintained for most of the quiescent state, and increases again centuries to decades prior to a new outburst. During the low state novae would escape luminosity-limited surveys. Thus nova systems are more numerous than they appear to be, but they erupt at much longer intervals. The very low luminosity of the oldest known novae, the relatively high luminosity of post- and prenovae, and the declining luminosity of decades-old novae support this theory.

The mass transfer rate is determined by the extent of Roche lobe overflow by the mass-donating star and may change when either the Roche lobe radius or the star's

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

radius change. Such changes are believed to be caused by several factors. Because of the mass lost by the binary system, the binary components are driven apart slightly during an outburst. The increased Roche lobe radius should result in a diminished mass transfer rate, but this effect is delayed so long as the companion star, irradiated by the hot nova, swells to some extent in order to get rid of the surplus luminosity. Eventually, as the nova fades away and the companion star contracts back to its equilibrium radius, the mass transfer rate does decline and remains low for a long time, between a few 10^3 yr and $\sim 10^5$ yr. During this long 'hibernation' phase the binary components are driven closer again owing to angular momentum losses caused by gravitational radiation or magnetic braking by a stellar wind. Thus the mass transfer rate increases again, leading to a new eruption.

General estimates

The large variation in observed nova features is attributed by the theory to differences in the values of three basic parameters that characterize the binary system: the mass of the accreting white dwarf, its internal temperature, which is a measure of the system's age, and the mass transfer rate, which is determined by the mass of the companion star and by the binary separation (orbital period). Remarkably, varying these parameters within acceptable limits reproduces roughly the entire range of observed properties. Furthermore, correlations are obtained between t_3 and the peak luminosity, the expansion velocity and the decline time of the bolometric luminosity, compatible with those derived from observations. Nevertheless, the stage where all the properties of a particular observed nova can be reproduced by one model has not yet been reached.

Despite the complexity of nova outbursts, some simple relations between the basic nova parameters and quantities that characterize the development of such outbursts may be obtained from analytical considerations. It has been shown, for example, that a thermonuclear runaway occurs when the pressure at the base of the accreted envelope—essentially, the degeneracy pressure of electrons—exceeds a critical value $P_{\rm crit}$ of order a few times 10^{18} N m^{-2}. Since the envelope, of mass $m_{\rm env}$, has negligible thickness,

$$P_{\rm crit} \approx \frac{G M_{\rm wd} m_{\rm env}}{4\pi R_{\rm wd}^4}. \tag{1}$$

For a white dwarf, the radius and mass are correlated; a simple relation, $R_{\rm wd} \propto M_{\rm wd}^{-1/3}$, holds for $M_{\rm wd} \lesssim 0.5 M_\odot$, leading, with equation (1), to

$$m_{\rm env} \propto M_{\rm wd}^{-7/3}. \tag{2}$$

For larger white dwarf masses, $R_{\rm wd}$ decreases much more strongly with increasing $M_{\rm wd}$. Thus massive white dwarfs require considerably smaller accreted envelopes in order to erupt. This conclusion has further consequences. The outburst recurrence time is given by $\mathcal{P} = m_{\rm env}/\dot{M}$ and, as the mass accretion rate $\dot{M}$ is independent of $M_{\rm wd}$, outbursts are more frequent on massive white dwarfs than on low-mass ones. In addition, since the companion's mass is another independent parameter, massive white dwarfs undergo, on average, a larger number of outbursts than low-mass ones. This results in a strong observational selection effect: the probability of discovering massive novae is higher, although they are not necessarily more numerous.

The luminosity at outburst is close to $L_{\rm Edd}$, which is proportional to $M_{\rm wd}$. Since the $M_{\rm wd}$ range is rather restricted, and since outbursts on massive white dwarfs are statistically favored, one should expect a narrow range of variation for the luminosities of novae in eruption. This should turn novae into reliable distance indicators (see below). During quiescence, the luminosity is that of accretion. Hence the outburst's bolometric amplitude A (in magnitudes) may be estimated by

$$A \approx 2.5 \log \left(\frac{\kappa \dot{M}}{4\pi c R_{\rm wd}(M_{\rm wd})} \right) \tag{3}$$

which yields, roughly, the observed range of ~ 7 to ~ 16.

The ejected mass $m_{\rm ej}$ may be estimated by assuming that it is supported against gravity at the star's surface—at the extended radius R—by the radiation pressure $P_{\rm rad} = \frac{1}{3} a T^4$, where a is the radiation constant. Thus, $m_{\rm ej} g = 4\pi R^2 P_{\rm rad}$, where g is the gravitational acceleration at the surface. Substituting the effective temperature for T and the Eddington luminosity for L,

$$4\pi R^2 \sigma T_{\rm eff}^4 = L = \frac{4\pi c G M_{\rm wd}}{\kappa} \tag{4}$$

leads to

$$m_{\rm ej} \approx \frac{16\pi R^2}{3\kappa}. \tag{5}$$

The range obtained—between $\sim 10^{-5} M_\odot$ and a few $10^{-4} M_\odot$—overlaps with that resulting from the independent estimate (1) and agrees with masses determined both observationally and by evolutionary calculations.

The energy required to power a nova outburst may be estimated as follows: the radiated energy is roughly the Eddington luminosity multiplied by the duration of an outburst, typically several weeks, which yields $\sim 5 \times 10^{37}$ J, and the kinetic energy of the expanding shell is $\sim 2.5 \times 10^{37}$ J, assuming a shell mass of $10^{-4} M_\odot$ and an average velocity of 500 km s^{-1}. However, both are negligible compared with the energy required to remove the shell from the deep gravitational potential well of the white dwarf, which is of order 10^{39} J. This energy is supplied by nuclear burning of a fraction α of the accreted hydrogen-rich envelope mass, which may be estimated by

$$\frac{G M_{\rm wd}}{R_{\rm wd}} \approx \alpha X Q \tag{6}$$

where X is the hydrogen mass fraction in the accreted matter, typically 0.7, and Q is the energy released per gram

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Nova properties and abundances.

Nova	Year	t_3 (days)	v (km s^{-1})	X	Y	Z	C	N	O	Ne, etc
V1370 Aql	1982	13	2800	0.053	0.088	0.859	0.035	0.14	0.051	0.633
T Aur	1891	100	655	0.47	0.40	0.13		0.079	0.051	
V842 Cen	1986	48	1200	0.41	0.23	0.36	0.12	0.21	0.030	0.0047
V693 CrA	1981	12	2210	0.29	0.32	0.39	0.0046	0.080	0.12	0.186
V1500 Cyg	1975	3.6	1180	0.57	0.27	0.16	0.058	0.041	0.050	0.010
V1668 Cyg	1978	23	1300	0.45	0.23	0.32	0.047	0.14	0.13	0.0068
HR Del	1967	230	550	0.44	0.48	0.077		0.027	0.047	0.0030
DQ Her	1934	94	350	0.31	0.31	0.38	0.056	0.13	0.20	0.007
V827 Her	1987	55	1000	0.36	0.29	0.35	0.087	0.24	0.016	0.0028
GQ Mus	1983	40	535	0.27	0.32	0.41	0.016	0.19	0.19	0.010
V2214 Oph	1988	73	500	0.34	0.26	0.40		0.31	0.060	0.032
RR Pic	1925	150	475	0.53	0.43	0.043	0.0039	0.022	0.0058	0.011
PW Vul	1984	147	285	0.54	0.28	0.18	0.032	0.11	0.038	0.0008
QU Vul	1984	40	?	0.44	0.46	0.10	0.0014	0.015	0.038	0.0447
QV Vul	1987	60	700	0.68	0.27	0.053		0.010	0.041	0.0020
Sun				0.71	0.27	0.0019	0.0031	0.001	0.0097	0.0052

of burnt hydrogen, 6×10^{14} J kg^{-1}. This yields $\alpha \approx 5\%$, which means that only a very small fraction of the accreted matter must be burnt in order to power the outburst.

The composition of nova ejecta—observation and theory

Characteristic abundances

The fact that nova ejecta exhibit unusual abundance features has long been recognized. Recent determinations of heavy-element abundances in the ejecta of classical novae have been based either on emission line analyses during decline or on analyses of resolved shells of older novae. The main characteristics of the nova ejecta composition derived from observations, as given in table 1, are as follows. (a) All novae show some degree of He–CNO enrichment. (b) There is simultaneous enrichment of He and CNO and/or ONeMg. (c) No apparent correlation is found between the helium mass fraction Y and the heavy-element mass fraction Z. (d) According to their neon content, novae may be roughly divided into two categories: those with a modest neon abundance, in spite of a considerable enrichment in CNO nuclei, and those which appear to have a very high abundance of neon and heavier elements. Although the theoretical prediction of significantly enriched ejecta preceded the actual observation, the enrichment mechanism has eluded astronomers for quite a long time.

Sources of enrichment

In principle, there are three possible sources for the unusual composition: (a) the accreted matter itself is rich in He, CNO and heavier isotopes; (b) heavy elements are synthesized during outburst; (c) core material is dredged up into the accreted hydrogen-rich matter. The first possibility is highly unlikely, the donor being a K or M main sequence star. In fact, no stars are known with outer layers of the nova ejecta composition. Regarding the second, the energy required to power a nova outburst in all its aspects is readily supplied, as shown above, by burning a tiny fraction of the accreted mass, and temperatures are not high enough to produce elements heavier than helium. Hence there remains the possibility of mixing of white dwarf material with the accreted mass. Two mechanisms have been suggested: (1) diffusion of hydrogen into the core, which causes ignition to take place below the core boundary and results in convective mixing of core material lying above the ignition radius into the envelope; (2) turbulent mixing due to shear-induced instabilities at the boundary between the white dwarf and the accreted material. Both may be supplemented by convective overshoot during the initial phases of the runaway.

At the onset of an outburst, the nova progenitor consists of two parts—core and envelope. The burning front separating them is located at the depth reached by the downward-diffusing hydrogen during the accretion phase in the diffusion scenario, or at the bottom of the unstable, turbulent layer, in the shear-mixing scenario. The explosive ignition of hydrogen leads to the formation of a convective zone that sweeps the entire envelope. Thus, regardless of the element distribution prior to outburst, the convective envelope's composition becomes well mixed before any matter is ejected. Hence observed abundances do not bear direct witness to the mixing mechanism, but they reflect the composition of the white dwarf. Accordingly, two types of white dwarf appear to be involved in nova outbursts: carbon–oxygen white dwarfs, in which carbon and oxygen are in roughly equal mass fractions, and oxygen–neon–magnesium white dwarfs, for which relative abundances are less certain. STELLAR EVOLUTION theory predicts that the latter must have larger masses than the former.

The source of helium excess may be either a helium-rich outer layer of the white dwarf progenitor—a relic of its early evolutionary stages—or the hydrogen-exhausted envelope remnant of a previous nova outburst. Since

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

helium and heavier-element enrichments are concomitant, the helium layer overlying the white dwarf core cannot be too thick. Thus, even if the white dwarf was originally formed with a relatively thick helium mantle, this mantle was eroded during the first nova outbursts. Hence the common helium source for the ejecta is the helium-rich remnant that forms during the decline phase of each nova outburst, to be destroyed and reformed in the next. The cyclic evolution and composition changes of a nova outburst are sketched in figure 5. The fact that helium and heavy-element enrichments have different sources explains the lack of correlation between them.

Nuclear processing

At outburst, temperatures exceed 10^8 K and densities are of order 10^7 kg m^{-3}. Under such conditions, the hydrogen burning reaction chain is

$$^{12}\mathrm{C}(\mathrm{p},\gamma)^{13}\mathrm{N}(\mathrm{p},\gamma)^{14}\mathrm{O}(\beta^+,\nu)^{14}\mathrm{N}(\mathrm{p},\gamma)^{15}\mathrm{O}(\beta^+,\nu)^{15}\mathrm{N}$$

ramifying into

$$^{15}\mathrm{N}(\mathrm{p},\alpha)^{12}\mathrm{C}$$

or

$$^{15}\mathrm{N}(\mathrm{p},\gamma)^{16}\mathrm{O}(\mathrm{p},\gamma)^{17}\mathrm{F}(\beta^+,\nu)^{17}\mathrm{O}(\mathrm{p},\alpha)^{14}\mathrm{N}.$$

Its characteristic timescale is limited by the lifetimes of ^{13}N (863 s), ^{14}O (102 s), ^{15}O (176 s) and ^{17}F (93 s), which become the most abundant heavy isotopes in the nova envelope. Their decay products, ^{13}C, ^{14}N, ^{15}N and ^{17}O, are thus greatly enhanced compared with the equilibrium CNO cycle abundances characteristic of solar composition.

The rate at which novae return mass to the interstellar medium may be estimated from the typical nova rate of about 20 yr^{-1} per galaxy multiplied by the average ejected mass per outburst, $\sim 10^{-4} M_\odot$, yielding $\sim 0.002 M_\odot$ yr^{-1}. This is about 50 times less than the rate corresponding to supernovae and over 200 times less than that corresponding to red giants and supergiants. Thus novae are significant contributors to interstellar abundances only for those isotopes which are enhanced by NUCLEAR PROCESSING more than 100 times compared with solar values. This is, indeed, the case for the rarer C, N and O isotopes. Consequently, novae have been repeatedly invoked as sources of the rarer isotopes, although their contribution to the dominant isotopes appears to be negligible (with the possible exception of ^{14}N). Isotopic ratios obtained from model calculations show that ^{13}C/^{12}C and ^{17}O/^{16}O are higher than solar by large factors for all parameter combinations, while ^{15}N/^{14}N is higher for some and lower for others. Consequently, novae are also considered potential ^{14}N sources in the interstellar medium, since this does not necessarily imply ^{15}N production in too large amounts.

Despite the high temperatures attained at outburst, the only abundant species in the ejecta of carbon–oxygen nova models are the isotopes of C, N and O. Neon production is entirely insignificant. Hence, it appears that the only way to obtain any neon in the ejecta would be to have it present in the white dwarf's composition. Indeed, carbon–oxygen white dwarfs formed in close binary systems do contain a few per cent of neon, in agreement with the amount indicated by observations.

In the case of novae occurring on oxygen–neon–magnesium white dwarfs, hydrogen burning proceeds through the NaMg reaction chain,

$$^{20}\mathrm{Ne}(\mathrm{p},\gamma)^{21}\mathrm{Na}(\mathrm{p},\gamma)^{22}\mathrm{Mg}(\beta^+,\nu)^{22}\mathrm{Na}(\mathrm{p},\gamma)^{23}\mathrm{Mg}(\beta^+,\nu)$$
$$^{23}\mathrm{Na}(\mathrm{p},\alpha)^{20}\mathrm{Ne}$$

and the MgAl reaction chain,

$$^{24}\mathrm{Mg}(\mathrm{p},\gamma)^{25}\mathrm{Al}(\beta^+,\nu)^{25}\mathrm{Mg}(\mathrm{p},\gamma)^{26}\mathrm{Al}(\mathrm{p},\gamma)^{27}\mathrm{Si}(\beta^+,\nu)$$
$$^{27}\mathrm{Al}(\mathrm{p},\alpha)^{24}\mathrm{Mg}$$

which yield substantial abundances of the radioactive species ^{22}Na and ^{26}Al. These isotopes have much longer lifetimes than the CNO radioactive isotopes, and hence they are live in the nova ejecta and in the dust grains that form in the shell. They are responsible, at least in part, for the enrichments in ^{22}Ne and ^{26}Mg—their decay products—that have been identified in meteoritic inclusions, believed to be representative of the primordial solar nebula composition. Moreover, ^{26}Al has been detected in the Galaxy through the 1.809 MeV γ-ray line and novae are considered potential sources of this isotope, competing with supernovae and red giants.

Novae as distance indicators

Novae are extremely luminous at optical wavelengths, brighter than Cepheids and surpassed only by supernovae, and they are about a hundred times more frequent than supernovae. They are therefore easy to detect in external galaxies, certainly within the Local Group, and potentially beyond. In particular, the constant luminosity maintained during part of the outburst can be used as a standard candle; if moderate to slow novae are considered, the magnitude variation among different novae does not exceed 1 magnitude.

The notorious property of nova outbursts is the apparent correlation between the maximum magnitude attained at outburst and the rate of decline, or more precisely, the time of decline by 3 magnitudes, t_3 (or, sometimes, by 2 magnitudes, t_2), a relationship that was already pointed out by Fritz Zwicky in 1936 and was first calibrated by McLaughlin in 1945. Since then, great effort has been devoted to the absolute calibration of this relation, which has became known as the maximum magnitude versus rate of decline (MMRD) relationship. The MMRD is considered a reliable distance indicator, in spite of the relatively large scatter of points, which results from observations and is expected from the theoretical point of view.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

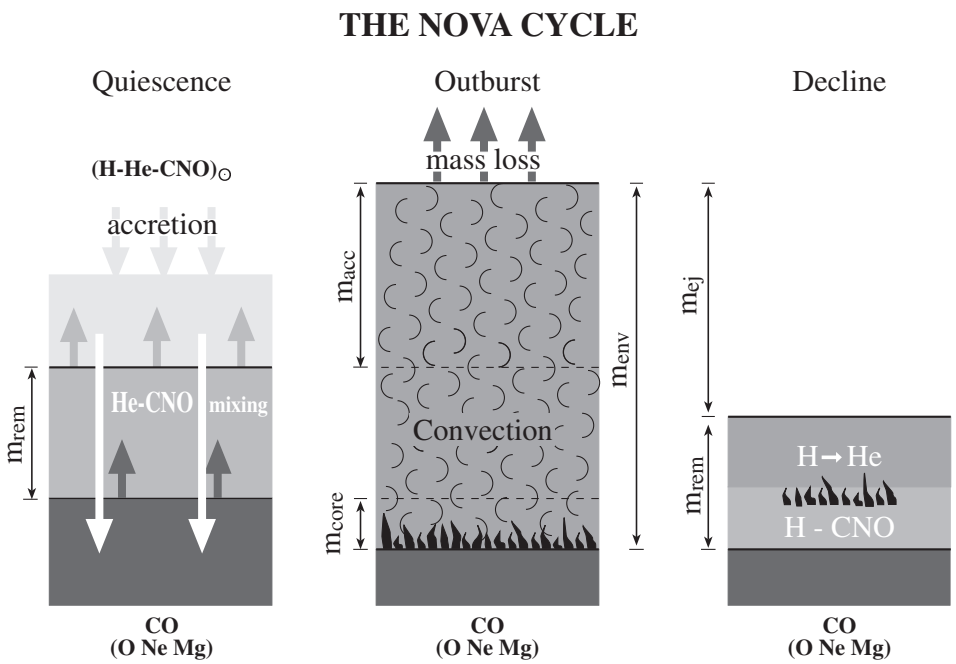


Figure 5. Schematic illustration of a nova cycle, focusing on composition changes caused by mixing, convection, nuclear burning. The white dwarf's mass mixed into the nova envelope is denoted $m_{\rm core}$; the remnant envelope mass after completion of mass ejection is denoted $m_{\rm rem}$.

Bibliography

An extensive account of Galactic novae, including a detailed catalogue, is given in

Payne-Gaposhkin C 1957 *The Galactic Novae* (Amsterdam: North-Holland)

and more recently in

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

The most comprehensive catalogue of Galactic classical novae (and similar objects) to date, including finding charts, has been compiled by

Duerbeck H W 1987 *A Reference Catalogue and Atlas of Galactic Novae* (Dordrecht: Reidel)

Collections of articles and reviews on novae may be found in

Bode M F and Evans A (ed) 1989 *Classical Novae* (New York: Wiley)
Cassatella A and Viotti R (ed) 1991 *Physics of Classical Novae* (Berlin: Springer)
Nussbaumer H and Orr A (ed) 1994 *Interacting Binaries* (Berlin: Springer)

as well as in numerous reviews and proceedings of conferences devoted to the broader class of cataclysmic variables.

The instability leading to nova outbursts was studied analytically by

Fujimoto Y F 1982 *Astrophys. J.* **257** 752, 767
Paczynski B 1983 *Astrophys. J.* **264** 282

The first numerical simulation of hydrogen burning on the surface of white dwarf was performed by

Giannone P and Weigert A 1967 *Z. Astrophys.* **67** 41

followed by detailed numerical evolutionary calculations of nova outbursts carried out by

Starrfield S, Sparks W M and Truran J W 1974 *Astrophys. J.* **192** 647
——1974 *Astrophys. J. Suppl.* **28** 247

who were also the first to investigate outbursts on ONeMg white dwarfs:

Starrfield S, Sparks W M and Truran J W 1986 *Astrophys. J. Lett.* **303** L5

The 'hibernation' scenario for nova systems was proposed by

Shara M M, Livio M, Moffat A M and Orio M 1986 *Astrophys. J.* **311** 163

following the extensive survey of novae—within cataclysmic variables—carried out by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Patterson J 1984 *Astrophys. J. Suppl.* **54** 443

The diffusion–convection mechanism for heavy-element enrichment in nova ejecta was proposed by

Prialnik D and Kovetz A 1984 *Astrophys. J.* **281** 367

who were also the first to explore the entire nova parameter space by multicycle evolutionary calculations:

Prialnik D and Kovetz A 1995 *Astrophys. J.* **445** 789
Kovetz A and Prialnik D 1997 *Astrophys. J.* **477** 356

Shear mixing resulting from accretion onto a rotating white dwarf was first proposed by

Kippenhahn R and Thomas H-C 1978 *Astron. Astrophys.* **63** 265

Nova winds, first studied by

Bath G and Shaviv G 1976 *Mon. Not. R. Astron. Soc.* **175** 305

have been widely investigated, particularly by

Kato M and Hachisu I 1994 *Astrophys. J.* **437** 802

The outburst and wind were first calculated self-consistently by

Kovetz A 1998 *Astrophys. J.* **495** 401

The first three-dimensional calculation of a nova outburst was performed by

Kercek A, Hillebrandt W and Truran J W 1999 *Astron. Astrophys.* **345** 831

Dina Prialnik

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nozomi (Planet-B)

Japanese mission to orbit Mars, launched in July 1998. Initially placed into a gravity assist orbit around the Earth and Moon. Carries instruments provided by Canada, Sweden, Germany and the United States. Intended to map the planet's surface, study the Martian atmosphere and the interaction of Mars with the solar wind. Arrival at Mars delayed until at least December 2003 after an engine malfunction necessitated an unscheduled course correction and heavy fuel consumption on leaving Earth orbit. Nozomi means 'hope'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuñez Salaciense, Pedro [Nunes, Pedro; Nonius, Petrus] (1502–78)

Doctor, philosopher and navigator, born in Alcácer do Sal, Portugal, holding professorial appointments at Lisbon and Coimbra, where he taught navigation during the heyday of the Portuguese world exploration, appointed Royal Cosmographer. Invented a vernier scale for measuring fractional parts of a degree on a quadrant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuclear Beta Decay

For over a century it has been known that atomic nuclei can undergo radioactive decay. These decays can be induced by the strong, weak or electromagnetic interaction in the case of excited nuclei. Those nuclear decays which proceed via the weak interaction are the beta decays. 'Beta' originally referred to electrons, but these decays can involve electrons or positrons, and either electron neutrinos or antineutrinos. The term beta decay covers all nuclear decays in which the charge number Z of the nucleus X for a constant mass number A changes by one unit. This includes the following decay processes:

$$\beta^-\text{decay}: \quad {}^{A}_{Z+1}\mathrm{X} \longrightarrow {}^{A}_{Z}\mathrm{Y} + \mathrm{e}^- + \overline{\nu}_\mathrm{e} \tag{1}$$

$$\beta^+\text{decay}: \quad {}^{A}_{Z-1}\mathrm{X} \longrightarrow {}^{A}_{Z}\mathrm{Y} + \mathrm{e}^+ + \nu_\mathrm{e} \tag{2}$$

$$\text{electron capture}: \quad \mathrm{e}^- + {}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A}_{Z-1}\mathrm{Y} + \nu_\mathrm{e}. \tag{3}$$

The most elementary beta decay process is the free neutron decay, in which a neutron decays into a proton, an electron and an antineutrino

$$\mathrm{n} \longrightarrow \mathrm{p} + \mathrm{e} + \overline{\nu}_\mathrm{e}, \tag{4}$$

or, on the quark level, a d quark decays into a u quark, an electron and an antineutrino

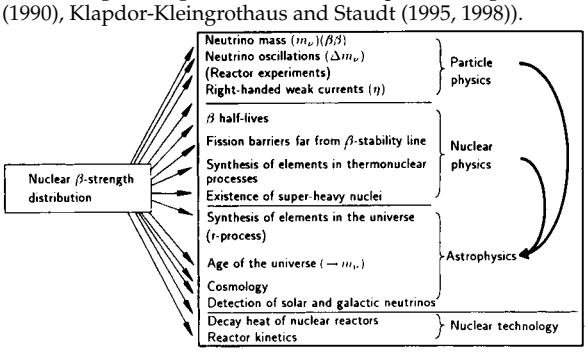
$$\mathrm{d} \longrightarrow \mathrm{u} + \mathrm{e} + \overline{\nu}_\mathrm{e}. \tag{5}$$

From a perturbation theory point of view, beta decay is a first-order effect of the classical theory. In the Glashow–Weinberg–Salam (GWS) theory, in which the pointlike current–current interaction is replaced by a boson exchange interaction, beta decay is a second-order effect.

In nuclear beta decay it is first and foremost possible to investigate the mass spectrum of the neutrino state which couples to the electron. The electron energy spectrum of an allowed transition is modified by a finite (Dirac or Majorana) neutrino mass. At present the investigations of tritium decay give the sharpest model-independent limits on the electron neutrino mass, of 2.8 eV (at 95% c.l.) (Weinheimer 2000, Lobashev 2000).

Even though the mechanism of β decay at the nucleon (respectively quark) level is in principle understood, only quite recently did it become possible to produce a quantitatively satisfactory description of the β decay processes of atomic nuclei (Klapdor 1983, 1986, Klapdor-Kleingrothaus 1991, Staudt *et al* 1990, Staudt and Klapdor-Kleingrothaus 1992, Hirsch *et al* 1992, 1993, Homma *et al* 1996). This is because the nucleus is a complex many-particle system, in which residual interactions of various types may lead to collective excitations, which can have a massive effect on the β decay properties. The distribution of the β strength in the daughter nucleus determines not only the β half-lives, and the rates for β-delayed particle emission (neutron or proton and fission) but also the shape of the emitted electron (positron) and (anti)neutrino spectra. The ability to calculate these quantities is of central importance to numerous applications in nuclear physics and neighboring disciplines such as astrophysics, particle physics and also nuclear engineering (see table 1). Some examples are (for details we refer to Klapdor (1983, 1986), Klapdor-Kleingrothaus (1991, 1998, 1999), Grotz and Klapdor-Kleingrothaus (1990), Klapdor-Kleingrothaus and Staudt (1995, 1998)):

Table 1. Nuclear beta decay and its links with some fundamental problems in nuclear, particle and astrophysics, and nuclear engineering (from Grotz and Klapdor-Kleingrothaus (1990), Klapdor-Kleingrothaus and Staudt (1995, 1998)).



Nuclear physics. Investigation of the nuclear structure of nuclei far from stability; determination of the fission barriers of nuclei from β-delayed fission; understanding the production of transuranic elements by thermonuclear devices; calculation of the electron and antineutrino spectra in nuclear reactors, originating from β decay of the around 1000 fission products (important for neutrino oscillation experiments).

Particle Physics. Calculation of $\beta\beta$-decay matrix elements as a prerequisite for determination of the neutrino mass and other particle theory parameters from measured decay rates (see below and also PARTICLE ASTROPHYSICS).

Astrophysics. Understanding the development of stars, particularly the gravitational collapse of heavy stars and the synthesis of elements in the *r* process; determining the age of the galaxy from cosmochronometers; response function of gallium detectors (and others) in solar neutrino detection. Stars shine because of the weak interaction involving either the pp reaction to form deuterium, or in the case of CNO burning, because of nuclear beta (actually positron) decays.

A precise knowledge of the beta decay of neutron-rich nuclei is crucial for the understanding of the r-process (Klapdor 1983, 1986, Klapdor-Kleingrothaus 1991). Both the element distribution on the r-path, and the resulting element distribution of stable elements—and of the cosmochronometers—are sensitive to the beta decay properties (half-lives, beta-delayed rates) of the neutron-rich nuclei involved in the process (about 6000 nuclei between the beta stability line and the neutron drip line (see figure 1)). The rates for beta-delayed fission determine where

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the r-process breaks off at high mass number A, and also influence the production rate of heavy nuclei. On the basis of microscopic calculations of the beta decay properties (since the early 1980s) of all of these nuclei for the first time reliable ages of the Galaxy and the universe were obtained with the method of cosmochronology (Klapdor 1983, 1986, Klapdor-Kleingrothaus 1991), which hinted more than ten years ago at a non-vanishing COSMOLOGICAL CONSTANT Λ (Klapdor and Grotz 1986). These hints were confirmed in 1998 by observation of high-redshift SUPERNOVAE of type Ia in the Supernova Cosmology Project.

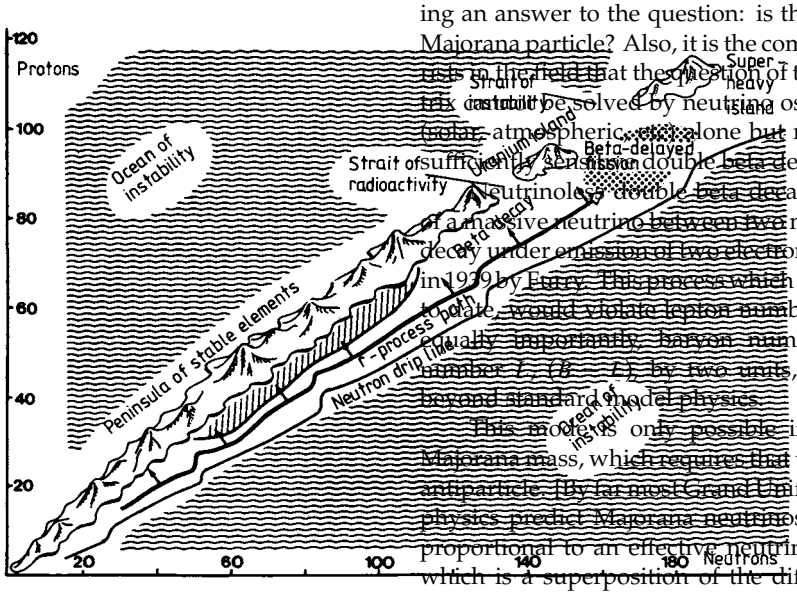


Figure 1. Schematic diagram of element synthesis by rapid neutron capture (r-process) (from Grotz and Klapdor-Kleingrothaus (1990), Klapdor-Kleingrothaus and Staudt (1995, 1998)).

Reactor physics. Calculation of the residual heat released in the β decay of fission products on switching off reactors.

It is a second order effect of the classical theory (fourth order in the GWS theory), namely double beta decay, which yields—besides proton decay—perhaps the most promising possibilities to probe physics beyond the standard model of particle physics (figure 2).

The investigation of double beta decay is indispensable to solve one of the fundamental questions of all unified theories of modern particle physics, that of neutrino mass and nature. Double beta decay is the only process allowing an answer to the question: is the neutrino a Dirac or Majorana particle? Also, it is the common opinion of theorists in the field that the question of the neutrino mass matrix can only be solved by neutrino oscillation experiments (solar, atmospheric, etc.) alone but requires in addition a sufficiently sensitive double beta decay experiment.

Neutrinoless double beta decay, in which exchange of a massive neutrino between two nucleons triggers their decay under emission of two electrons, was discussed first in 1939 by Furry. This process which has not been observed to date, would violate lepton number conservation, and, equally importantly, baryon number B minus lepton number L, $(B - L)$, by two units, which would imply beyond standard model physics.

This mode is only possible if the neutrino has a Majorana mass, which requires that the neutrino is its own antiparticle. [By far most Grand Unified Models of particle physics predict Majorana neutrinos.] The decay rate is proportional to an effective neutrino mass $\langle m \rangle$ squared, which is a superposition of the different neutrino mass eigenstates m_i

$$\langle m_\nu \rangle = \sum_i m_i U_{ei}^2. \tag{6}$$

respecting the possibility of the electron neutrino to be a mixed state (mass matrix not diagonal in the flavor space)

$$|\nu_e\rangle = \sum_i U_{ei} |\nu_i\rangle \tag{7}$$

U is the (unitary) neutrino mixing matrix.

Consequently double beta decay yields important information on the parameters of the neutrino mass and mixing matrix. The information from double beta decay is particularly important to fix the absolute neutrino mass scale, since neutrino oscillation experiments measure only differences between mass eigenstates.

There are worldwide several double beta experiments running looking for this type of decay for various nuclei (figure 3). The Heidelberg–Moscow experiment is by far the most sensitive one at present, and also for the next few years (figure 3 and see Klapdor-Kleingrothaus (1998, 1999)).

The team from the Max Planck Institute for Nuclear Physics in Heidelberg and the Kurchatov Institute in Moscow constructed a set-up of five high-purity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

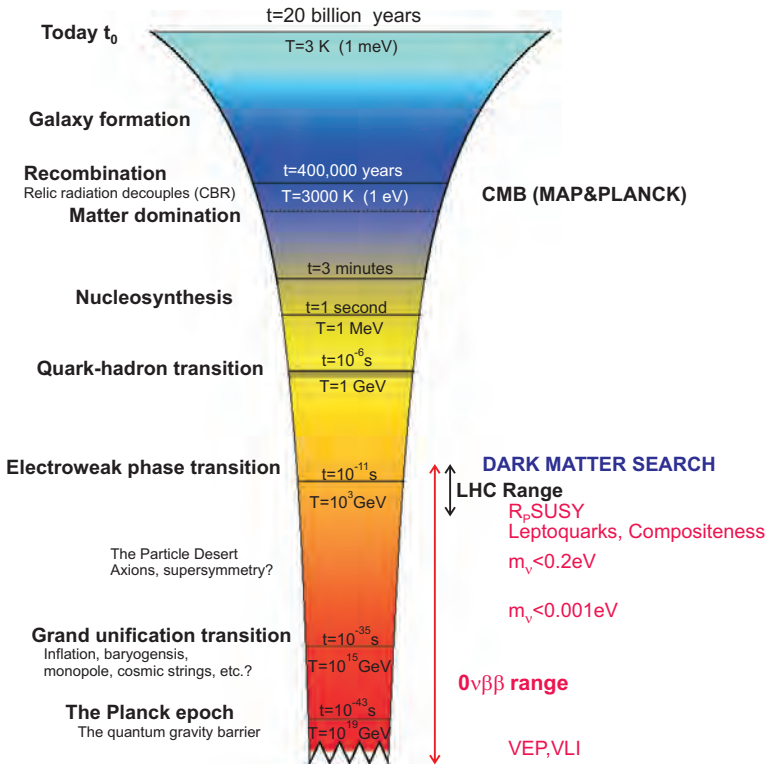


Figure 2. Energy (time) scales in the early universe where double beta decay can contribute to the related particle physics research. For the (light) neutrino masses the scale is indicated for the right-handed heavy counterparts according to the see-saw mechanism.

germanium detectors enriched in the isotope germanium-76 to 86% (natural abundance 7.8%) of total mass of 11.5 kg. This results in the largest source strength ever used in a double beta decay experiment. The set-up is operated in a heavy shielding under 1500 meters of rock (corresponding to 3500 m of water shielding) in the Gran Sasso Underground Laboratory in Italy. The background level reached by the experiment is the lowest worldwide for this kind of experiment—0.06 events per year and kilogram of detector mass in the energy range of the expected double beta decay signal, which would be a peak in the spectrum produced by the decay electrons, at 2038 keV.

In almost 10 years of measurement, the Heidelberg–Moscow team has obtained a lower limit for the half-life of neutrinoless double beta decay of germanium-76 of several 10^{25} years. This is the world record under all running investigations. With the deduced limit for the Majorana neutrino mass of 0.27 eV, the Heidelberg–Moscow experiment is not only the first one exploring the sub-eV range of the neutrino mass, but also enters in a range of values, which has stringent consequences for neutrino mass models and for cosmological models assuming neutrinos as hot DARK MATTER in the universe (see figure 4 and Klapdor-Kleingrothaus *et al* 2000a).

For example, *in degenerate neutrino mass models*, in which the neutrino mass eigenstates have very close values, the above value (and even a more conservative value of 0.35 eV) excludes models with cold and hot dark matter (CHDM models), and also those including a non-vanishing cosmological constant Λ (ΛCHDM models) for the case of the small mixing angle solution of the solar neutrino problem. (Minakata 1997, Adhikari *et al* 2000, Ellis and Lola 1999, Georgi and Glashow 2000, Yasuda 1999). This result is consistent with the recent disfavoring of the small mixing angle solution for the solar neutrino problem as reported by SUPER-KAMIOKANDE (Suzuki 2000, see also Gonzales-Garcia 2000). Assuming the best fits of the large mixing angle solution or the vacuum oscillation solutions, still CHDM models are excluded, while ΛCHDM models are still marginally possible (requiring $\langle m \rangle = 0.15$–0.30 eV). In the case of inverse hierarchical neutrino mass scenarios, with only two neutrinos contributing to hot dark matter, only ΛCDM models with a small Hubble constant $h = 0.5$ are marginally not yet excluded.

Already now, still assuming the small angle solution of the solar neutrino problem to be valid, the experiment would rule out almost the whole range of sensitivity of the future satellite experiments MAP and PLANCK for cosmological models of the above types involving neutrinos as hot dark matter (see figure 4). Figure 4 also demonstrates that, combined with the neutrino oscillation results obtained in the recent solar and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

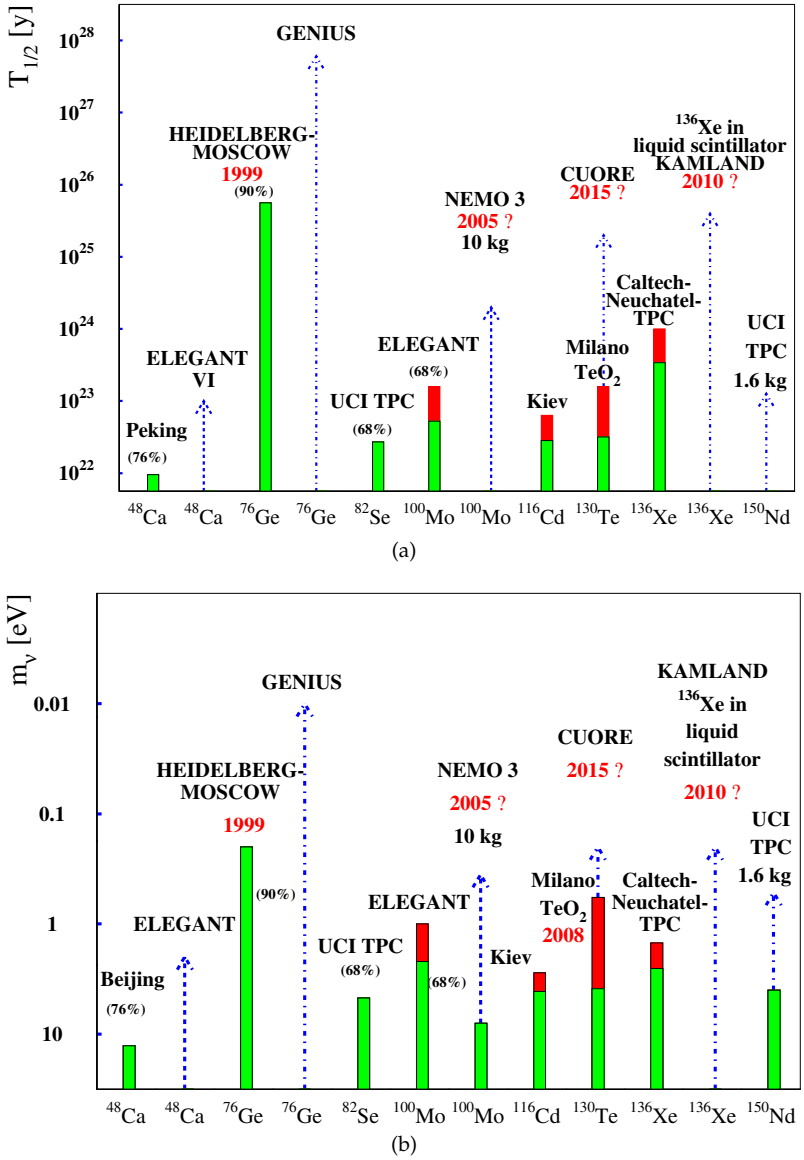


Figure 3. Present situation, 2000, and expectation for the near future and beyond, of the most promising $\beta\beta$-experiments concerning accessible half-life (a) and neutrino mass limits (b). The light-shaded parts of the bars correspond to the present status, the dark parts of the bars to expectations for running experiments, dashed lines to experiments under construction and dash-dotted lines to proposed experiments.

atmospheric neutrino oscillation experiments (see also SOLAR INTERIOR: NEUTRINOS) and with precise determinations of cosmological parameters, double beta decay is obviously the only way to obtain precise informations about the neutrino mixing and the absolute mass scale in partially degenerate and degenerate neutrino mass scenarios.

Looking into four-neutrino scenarios, including in addition to the three known neutrino flavours a fourth 'sterile' neutrino, there are as shown recently by Giunti *et al*, Bilenky *et al*, and others, only two schemes that can accommodate the results of *all* neutrino oscillation experiments (including the Los Alamos experiment LSND). The first of these two scenarios is ruled out by the Heidelberg–Moscow experiment.

An application of the extremely low background of the germanium double beta detectors is to search for cold dark matter in the form of WIMPs (weakly interacting massive particles; see WIMPS AND MACHOS) which are expected to populate the halo of our Galaxy. They could be looked for by elastic scattering off the Ge nuclei and the following ionization by the recoiling nucleus in the double beta decay detectors. The deposited energy for neutralinos with masses between 10 GeV and 1 TeV

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

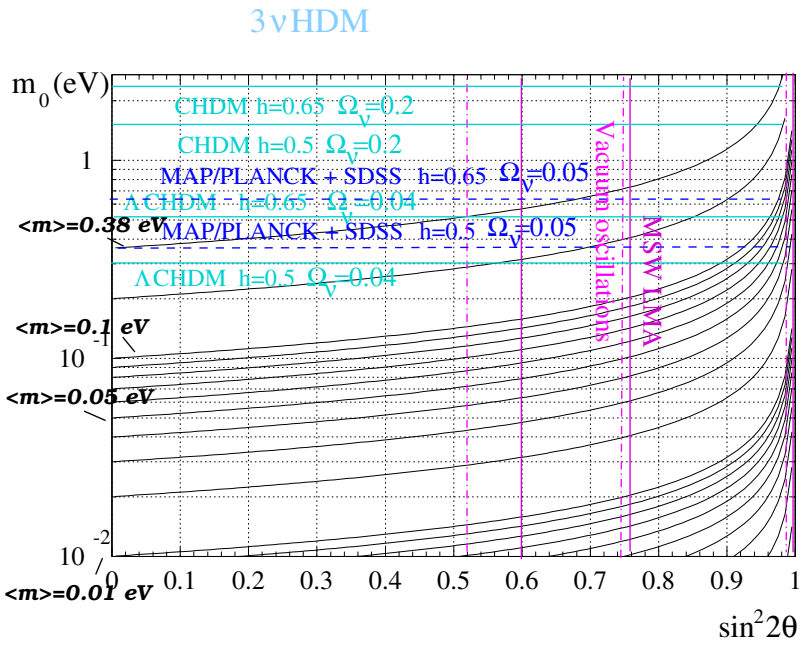


Figure 4. Isomass lines for the effective neutrino mass $\langle m \rangle$ measured in double beta decay in the m_0–$\sin^2(2\theta)$ plane *for the case of a degenerate three-neutrino scenario* with mass m_0. Also shown are the best fits for CHDM and ΛCHDM for different values of the Hubble constant, the sensitivity of MAP/PLANCK combined with SDSS, and the regions of the MSW LMA and of the vacuum oscillation solutions of the solar neutrino problem. The Heidelberg–Moscow experiment excludes already most of the cosmological models and also, assuming the MSW SMA solution, almost the whole range of sensitivity of the future satellite experiments MAP and PLANCK for cosmological models involving neutrinos as hot dark matter (from Klapdor-Kleingrothaus *et al* 2000a).

is below 100 keV. The best current limits on WIMP nucleon cross sections come from the DAMA experiment, from the CDMS experiment and from the Heidelberg–Moscow experiment, the latter experiment yielding the most stringent limits for using raw data without pulse shape analysis. All of these experiments at present just marginally touch only the upper part of the parameter space predicted for neutralinos as cold dark matter. A huge increase in sensitivity could be expected from the future double beta (and dark matter) detector GENIUS (see Klapdor-Kleingrothaus (1997, 1998, 1999), Klapdor-Kleingrothaus and Hirsch 1997, Klapdor-Kleingrothaus *et al* (1999a, 2000a)).

In addition to the information on the neutrino mass, double beta decay experiments, and in particular the Heidelberg–Moscow experiment yield already now information on beyond standard model physics on the TeV scale, where new physics could be expected. This is possible since the $\Delta L = 2$ process of neutrinoless double beta decay could occur in *any* lepton-number violating theory, so, e.g. by exchange of supersymmetric particles like neutralinos, gluinos, sleptons, etc, and thus the process would yield information on the underlying theories and on properties of the involved particles. This has been realized in the last 10–15 years only (see Klapdor-Kleingrothaus (1998, 1999), Klapdor-Kleingrothaus and Päs (1998) and Klapdor-Kleingrothaus and Krivosheina (2000)). It is important to note, and has been proved theoretically already in the early 1980s by Schechter and Valle, that independent of the mechanism underlying the double beta decay, its occurrence *always* would imply a nonvanishing Majorana neutrino mass.

The half-life limit measured in the Heidelberg–Moscow experiment yields a *lower* limit on the mass of a superheavy left-handed neutrino of $M > 8 \times 10^5$ GeV, a limit which could be reached only by a far-future 2 TeV linear electron–electron collider. It yields further an upper limit on the Yukawa coupling Λ'_{111} in the R-parity breaking part of the superpotential of the minimal supersymmetric standard model (MSSM), which is more stringent than present limits from the TEVATRON and HERA colliders, and which immediately excluded the possibility of squarks of first generation being produced in the recently discussed high-Q squared events at HERA.

The Heidelberg–Moscow result also restricts stringently products of higher generation Yukawa couplings, it yields the sharpest limits for a Majorana-type mass of the sneutrino, the supersymmetric partner of the neutrino, sets limits to compositeness (assuming a substructure of quarks and leptons), which are, as shown recently by Panella *et al* (2000), more stringent than those from the LEP II accelerator, yields bounds on violation of Lorentz invariance (VLI) and equivalence principle (VEP) in the neutrino sector, in the range of small mixing, which cannot be constrained by other experiments. This implies that double beta decay probes even physics at the Planck scale,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

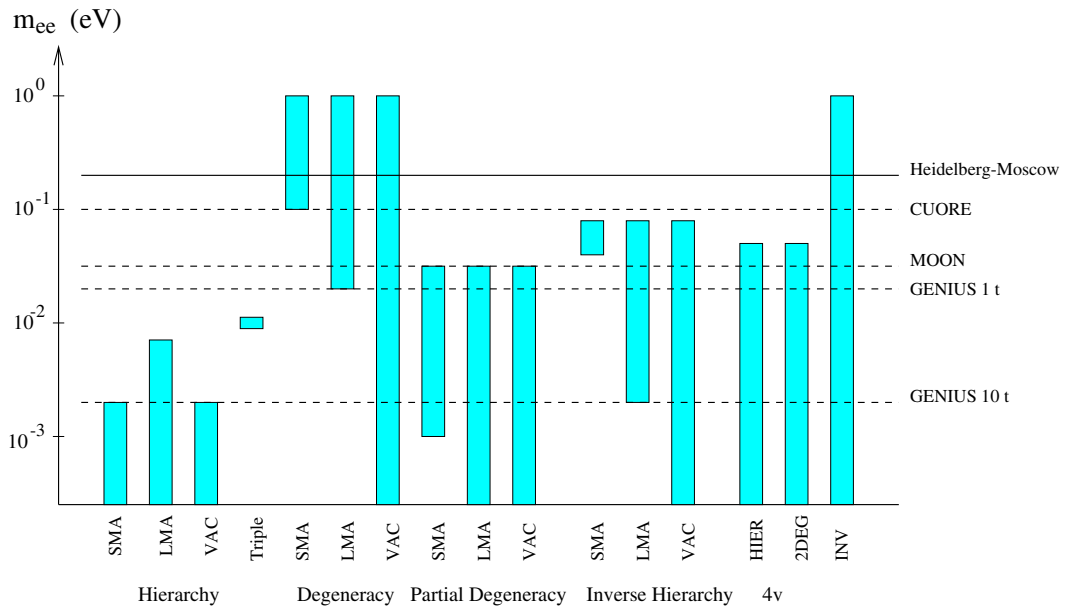


Figure 5. Summary of expected values for the *effective neutrino mass* $m_{ee} \equiv \langle m_\nu \rangle$ in different neutrino mass schemes and the potential of future double beta experiments to probe them (from Klapdor-Kleingrothaus *et al* 2000a).

since violation of Lorentz invariance and of equivalence principle (as well as CPT violation) could occur in superstring theories (see STRING THEORY/SUPERSTRING THEORY). Violation of the latter could result from contributions of dilatons (excitations in compactified spacetime dimensions) to gravitation by flavour-characteristic couplings. This could lead to neutrino oscillations. Violation of Lorentz invariance can originate from the fact that in QUANTUM GRAVITY theories the vacuum has a non-trivial SPACETIME structure, and thus particles propagating through the quantum foam show energy-dependent recoil and vacuum polarization effects (for details see Klapdor-Kleingrothaus (1998, 1999), Klapdor-Kleingrothaus *et al* (1999b, 2000a, b, c)).

In summary the most sensitive present double beta experiments yield important complementary information to accelerator experiments, which partly can be covered only by future colliders.

The method of double beta decay could be pushed in another type of set-up to a limit of 0.1 eV or ultimately 0.001 eV, which probably would be the ultimate value obtainable in double beta decay experiments. For this purpose the Heidelberg group has proposed the new project GENIUS (Klapdor-Kleingrothaus 1997, Klapdor-Kleingrothaus and Hirsch 1997, Klapdor-Kleingrothaus *et al* 1999a, Klapdor-Kleingrothaus and Zuber 1997). Its research potential for neutrino physics is demonstrated in figure 5. It is complementary to that of solar neutrino experiments (see, e.g., Bahcall *et al* 2000). More generally, this would have implications for a broad area of physics encompassing particle physics in the multi TeV range, and astrophysics, and which would serve as an important bridge between the physics that will be gleaned from high energy accelerators such as LHC and NLC on the one hand and satellite experiments such as MAP and PLANCK on the other.

For further information please look at Klapdor-Kleingrothaus H V (2000), and at the home page:

http://www.mpi-hd.mpg.de/non_acc.

Bibliography

Adhikari R and Rajasekaran G 2000 *Phys. Rev.* D **61** 031301-1

Bahcall J N, Krastev P I and Smirnov A Yu 2000 *Preprint hep-ph/0002293*

Ellis J and Lola S 1999 *Phys. Lett.* B **458** 310

Georgi H and Glashow S L 2000 *Phys. Rev.* D **61** 097301-1

Gonzales-Garcia M 2000 *Proc. Int. Conf. 'Neutrino Physics and Astrophysics', NEUTRINO 2000*, Sudbury, Canada, 16–21 June 2000, ed A McDonald *Nucl. Phys. Proc. Suppl.* B

Grotz K and Klapdor-Kleingrothaus H V 1990 *The Weak Interaction in Nuclear, Particle and Astrophysics* (Bristol: Adam Hilger)

Hirsch M, Staudt A and Klapdor-Kleingrothaus H V 1992 *At. Data Nucl. Data Tables* **51** 243

Hirsch M, Staudt A, Muto K and Klapdor-Kleingrothaus H V 1993 *At. Data Nucl. Data Tables* **53** 165

Homma H, Bender E, Hirsch M, Muto K, Klapdor-Kleingrothaus H V and Oda T 1996 *Phys. Rev.* C **54** 2972

Klapdor H V 1983 *Prog. Part. Nucl. Phys.* **10** 131

Klapdor H V 1986 *Prog. Part. Nucl. Phys.* **17** 419

Klapdor H V and Grotz K 1986 *Astrophys. J.* **304** L39

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Klapdor-Kleingrothaus H V 1991 *Nuclei in the Cosmos* ed H Oberhummer (Heidelberg: Springer) p 199

Klapdor-Kleingrothaus H V 1997 *Proc. Beyond the Desert 1997—Accelerator and Non-Accelerator Approaches, BEYOND '97*, eds H V Klapdor-Kleingrothaus and H Päs (Bristol: Institute of Physics) p 485

Klapdor-Kleingrothaus H V 1998 *Int. J. Mod. Phys.* A **13** 3953

Klapdor-Kleingrothaus H V 1999 *Lepton and Baryon Number Violation in Particle Physics, Astrophysics and Cosmology* ed H V Klapdor-Kleingrothaus and I V Krivosheina (Bristol: Institute of Physics) p 251

Klapdor-Kleingrothaus H V 2000 *Sixty Years of Double Beta Decay—From Nuclear Physics to Particle Physics Beyond the Standard Model*, (World Scientific: Singapore)

Klapdor-Kleingrothaus H V, Baudis L, Heusser G, Majorovits B and Päs H 1999a *MPI-Report* MPI-H-V26-1999 *Preprint hep-ph/9910205*

Klapdor-Kleingrothaus H V and Hirsch M 1997 *Z. Phys.* A **359** 361

Klapdor-Kleingrothaus H V and Krivosheina I V (ed) 2000 *Beyond the Desert 1999—Accelerator, Non-Accelerator and Space Approaches* (Bristol: Institute of Physics)

Klapdor-Kleingrothaus H V and Päs H (ed) 1998 *Beyond the Desert 1997—Accelerator and Non-Accelerator Approaches* (Bristol: Institute of Physics)

Klapdor-Kleingrothaus H V and Päs H 2000b *Proc. COSMO'99 Trieste, Italy, 27 September–3 October 1999* and *Preprint hep–ph/0002109*

Klapdor-Kleingrothaus H V, Päs H and Sarkar U 1999b *Eur. Phys. J.* A **5** 3

Klapdor-Kleingrothaus H V, Päs H and Sarkar U 2000c *Preprint hep–ph/0002215*

Klapdor-Kleingrothaus H V, Päs H and Smirnov A 2000a submitted for publication, *Preprint hep-ph/0003219*

Klapdor-Kleingrothaus H V and Staudt A 1995 *Non-Accelerator Particle Physics* (Bristol: Institute of Physics)

Klapdor-Kleingrothaus H V and Staudt A 1998 *Non-Accelerator Particle Physics* (Bristol: Institute of Physics), second edition

Klapdor-Kleingrothaus H V and Zuber K 1997 *Particle Astrophysics* (Bristol: Institute of Physics), second edition 2000

Lobashev V 2000 *Proc. Int. Conf. 'Neutrino Physics and Astrophysics', NEUTRINO 2000*, Sudbury, Canada, 16–21 June 2000, ed A McDonald *Nucl. Phys. Proc. Suppl.* B

Minakata H and Yasuda O 1998 *Phys. Rev.* D **56** 1692

Panella O, Carimalo C and Srivastava Y N 2000 *Phys. Rev.* D **62** 015013-1

Staudt A, Bender E, Muto K and Klapdor-Kleingrothaus H V 1990 *At. Data Nucl. Data Tables* **44** 79

Staudt A and Klapdor-Kleingrothaus H V 1992 *Nucl. Phys.* A **549** 254

Suzuki Y (for the SUPERKAMIOKANDE collaboration) 2000 *Proc. Int. Conf. 'Neutrino Physics and Astrophysics', NEUTRINO 2000*, Sudbury, Canada, 16–21 June 2000, ed A McDonald *Nucl. Phys. Proc. Suppl.* B

Weinheimer C 2000 in *Proc. Int. Conf. 'Neutrino Physics and Astrophysics', NEUTRINO 2000*, Sudbury, Canada, 16–21 June 2000, ed A McDonald *Nucl. Phys. Proc. Suppl.* B

Yasuda O 1999 *Proc. 'Beyond the Desert '99': Accelerator, Non-accelerator and Space Approaches, 'Conference on Physics Beyond the Standard Model'* eds H V Klapdor-Kleingrothaus and I V Krivosheina, Tegernsee, Germany, 6–12 June 1999 (Bristol: Institute of Physics)

Hans V Klapdor-Kleingrothaus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuclear Combustion

Nuclear combustion is the rapid consumption of stellar material by a thermonuclear fusion front, propagating subsonically (deflagration) or supersonically (detonation), releasing large amounts of energy that are observable as astrophysical flashes.

Overview

In some astrophysical events, stellar material becomes sufficiently strongly compressed and heated that nuclear fusion reactions occur much more rapidly than during the slow, gradual nuclear burning that takes place for most of every star's lifetime. If this happens, a thin sheet of reacting material, a so-called nuclear combustion front, forms and travels from the point of its ignition outward. The huge amount of energy released by the combustion front within a very short time is believed to be the source for the bursts of light that are observed as astrophysical flashes. There are mainly three types of astrophysical explosions whose theoretical explanations invoke fast nuclear combustion: classical NOVAE, X-RAY BURSTERS and—most violently—type Ia SUPERNOVAE. In the former two, the hydrogen gas accreted onto a WHITE DWARF or NEUTRON STAR in a BINARY SYSTEM becomes more and more compressed by the strong gravitational field of the accreting star and eventually explodes, leaving the underlying star intact. In the case of a type Ia supernova, the accreting white dwarf ignites nuclear combustion deep in its interior, leading to the explosion of the entire star.

In order to understand these astrophysical phenomena, one must construct mathematical models that describe the structure of nuclear combustion fronts and their modes of propagation. If one neglects the influence of turbulent fluid motions, all combustion fronts can be classified as either deflagrations or detonations. Deflagrations (or 'flames') are thin layers of hot nuclear burning material that diffuse heat into the cold, unburned material ahead of the front, igniting it and thus moving the front forward. They propagate subsonically and resemble, for instance, flames in car engines. Detonations, on the other hand, are driven supersonically by a shock wave that is formed by the energy released in the reacting layer.

In astrophysical explosions, however, the burning material is often not at rest but undergoes strong, turbulent motions—often stirred by the combustion front itself—making the modeling of the event much more difficult. Turbulent deflagrations can propagate much faster than laminar ones, and may sometimes even turn into detonations. Computers are needed to solve the mathematical models in all but the simplest cases, but even they can only provide us with a rough glimpse of the complexity of turbulent nuclear combustion fronts.

Basic physics of nuclear combustion fronts

Just like the hydrostatic nuclear burning that takes place in the core of our Sun, astrophysical thermonuclear flashes obtain their energy from nuclear FUSION reactions that combine light atomic nuclei to heavier ones. They differ mainly in the speed at which these reactions take place, described by the energy generation rate $\dot{w}$. Since the reaction rate depends on the probability of two nuclei colliding and therefore on their speeds and mean separation, $\dot{w}$ is a very steeply rising function of the density ρ and, even more crucially, the temperature T. How do thermonuclear flashes reach sufficiently high temperatures to burn their nuclear fuel explosively?

Stars control their central temperature, and thus their energy generation rate, by expansion or contraction that causes cooling or heating respectively. This feedback control becomes impossible if the pressure is almost independent of the temperature, as in the case of electron degeneracy. All three cases of known astrophysical thermonuclear flashes involve degenerate nuclear fuel (we shall use the terms 'fuel' and 'ashes' in a general sense for the nuclear reactants and reaction products respectively): in novae and x-ray bursts, the hydrogen layer on top of a white dwarf or a neutron star is compressed by the strong gravitational field and becomes degenerate, whereas in type Ia supernovae the nuclear fuel is the degenerate carbon and oxygen core of a Chandrasekhar mass white dwarf itself. In all three cases, accretion from a binary companion heats the material and triggers nuclear reactions. In the absence of an efficient cooling mechanism, T, and therefore $\dot{w}$, rises at an accelerating pace that is only stopped when all of the material is burned to nuclear ashes. A thermonuclear runaway begins. It is localized in a small region where the initial temperature fluctuation was highest, the so-called ignition region.

When a critical temperature T_c is exceeded, all of the fuel inside the ignition region is consumed almost instantaneously, and a thin reaction zone forms at the interface of burned and unburned material. It propagates into the surrounding fuel by one of two mechanisms: if the overpressure created by the heat of the burning products is sufficiently high, a hydrodynamical shock wave forms that ignites the fuel by compressional heating. A self-sustaining combustion front that propagates by shock-heating is called a detonation (cf figure 1). Detonations generally move supersonically, and therefore do not allow the unburned medium to expand before it is burned (this is important in the context of type Ia supernovae). Their speed depends only on the total amount of energy released per unit mass, ϵ. On the other hand, if the initial overpressure is too weak, the temperature gradient at the fuel–ashes interface steepens until an equilibrium between heat diffusion (carried out predominantly by electron–ion collisions) and energy generation is reached. The resulting combustion front consists of a diffusion zone that heats the fuel up to T_c, followed by a thin reaction layer where the fuel is consumed and energy is generated. It is called a deflagration or simply a flame and moves subsonically with respect to the unburned material (cf figure 2). Flames, unlike detonations, may therefore be strongly affected by turbulent velocity fluctuations of the fuel. Only if the unburned material is at rest, can a unique

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

laminar flame speed S_l be found which depends on the detailed interaction of burning and diffusion within the flame region. It can be estimated by assuming that in order for burning and diffusion to be in equilibrium, the respective time scales, $\tau_b \sim \epsilon/\dot{w}$ and $\tau_d \sim \delta^2/\kappa$, where δ is the flame thickness and κ is the thermal diffusivity, must be similar: $\tau_b \sim \tau_d$. Defining $S_l = \delta/\tau_b$, one finds $S_l \sim (\kappa\dot{w}/\epsilon)^{1/2}$, where $\dot{w}$ should be evaluated at $T \approx T_c$. This is only a crude estimate due to the strong T dependence of $\dot{w}$. Numerical solutions of the full equations of hydrodynamics including nuclear energy generation and heat diffusion are needed to obtain more accurate values for S_l as a function of ρ and fuel composition.

Flame instabilities and turbulent nuclear combustion

Even if the laminar, one-dimensional structure of flames and detonations is understood, one has to ask whether small perturbations of this structure grow in three-dimensional reality, i.e. whether the solution is hydrodynamically unstable (there are also non-hydrodynamical, one-dimensional flame instabilities which will not be discussed here). If it is, the perturbations may either grow to a finite size and stabilize, or grow until the fluid becomes fully turbulent. In either case, the exact mathematical description of the combustion front becomes very complicated or even impossible. Instead, phenomenological or statistical methods must be employed to model important global quantities such as the total extent, propagation speed and energy generation rate of the front. The following discussion will be limited to flames; although detonations are known to develop cellular instabilities, their relevance in astrophysical explosions has not been as clearly established as flame instabilities.

Let us first consider very thin flames, such that their internal structure can be neglected when observed on scales of the perturbation wavelength. These flames are unstable to the Landau–Darrieus (LD) instability which arises as a consequence of the density jump produced by thermal expansion within the flame, and is present whenever a density discontinuity moves subsonically in the direction of the higher density. The instability grows until a web of cellular structures forms and stabilizes the front against further growth of the perturbations. The LD instability therefore does not, in general, lead to the production of turbulence.

If the flame moves upward in a gravitational field, it is also subject to the Rayleigh–Taylor (RT) instability resulting from the buoyancy of the hot, burned fluid with respect to the dense, unburned material. Here, small perturbations grow until they form bubbles that begin to float upward. After a short time, bubbles of fuel and ashes with various sizes create a foamy RT mixing layer that continually grows in size. Secondary instabilities related to the velocity shear along the bubble surfaces may then lead to the production of turbulence.

The problem of turbulent combustion has been researched extensively in the engineering sciences in the context of car engines, turbines and chemical explosions. Owing to the nonlinearity of the underlying equations, the Navier–Stokes equations of FLUID DYNAMICS including reactions and diffusion, exact solutions are rarely obtainable. Even numerical simulations on supercomputers can only yield limited insight into the complexity of turbulent combustion. The main problem is the tremendous range of length scales that need to be observed (or solved for), ranging from the size of the largest turbulent eddy L (reaching some fraction of the stellar radius in the astrophysical case) down to the smallest one, which can be smaller than a millimeter. Since no computer is capable of resolving this range of scales, one has to resort to statistical or scaling approximations of those length scales that are not properly resolved. The most prominent scaling relation in the science of turbulence is Kolmogorov's law for the cascade of velocity fluctuations: it states that under the assumption of isotropy and statistical stationarity, the mean velocity v of turbulent eddies with size l scales as $v \sim l^{1/3}$. Given the velocity of large eddies, e.g. from computer simulations, one can use this relation to extrapolate the eddy velocity distribution down to smaller scales under the assumption of isotropic, fully developed turbulence.

Knowledge of the eddy velocity as a function of length scale is important to classify the burning regime of the turbulent combustion front. The ratio of the laminar flame speed and the turbulent velocity on the scale of the flame thickness, $K = S_l/v(\delta)$, plays an important role: if $K \gg 1$, the laminar flame structure is nearly unaffected by turbulent fluctuations. Turbulence does, however, wrinkle and deform the flame on scales l where $S_l \ll v(l)$, i.e. above the Gibson scale l_g defined by $S_l = v(l_g)$. These wrinkles increase the flame surface area and therefore the total energy generation rate of the turbulent front. One can define the turbulent flame speed, S_t, as the mean overall propagation velocity of the turbulent flame front (obviously, $S_t > S_l$). If the turbulence is sufficiently strong ($v(L) \gg S_l$) the turbulent flame speed becomes independent of the laminar speed (and therefore of the microphysics of burning and diffusion) and scales only with the velocity of the largest eddy: $S_t \sim v(L)$. Because of the unperturbed laminar flame properties on very small scales, and the wrinkling of the flame on large scales, the burning regime where $K \gg 1$ is called the 'corrugated flamelet regime'.

The opposite limit, $K \ll 1$, is the 'well-stirred reactor' burning regime. In this regime, a laminar flame structure ceases to exist because the fluid is stirred by turbulence on scales of δ and below. Nuclear reactions are no longer confined to thin layers but occur everywhere at once at varying intensities, depending on the local temperature and fuel concentration. The flame instabilities mentioned above are replaced by reactive convection as a source for the turbulence. Unlike the flamelet regime, the well-stirred reactor regime does not give rise to a simple estimate for S_t, as the details of burning and diffusion continue to matter. It is, however, accessible to simplified

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

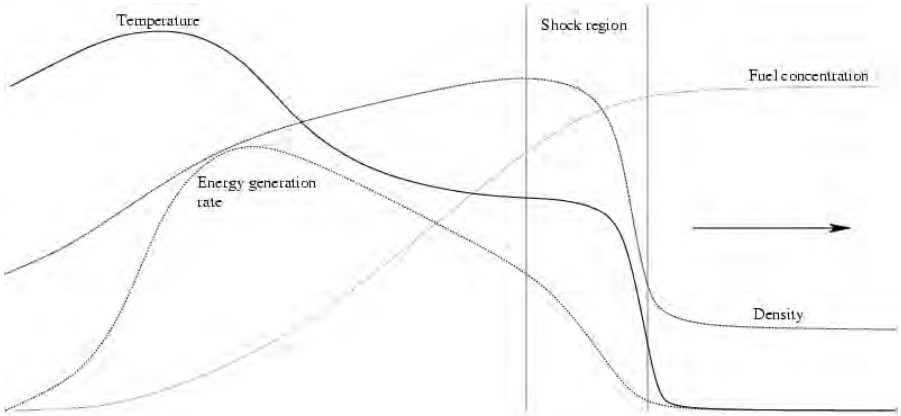


Figure 1. Sketch of the variation of flow properties across a detonation front.

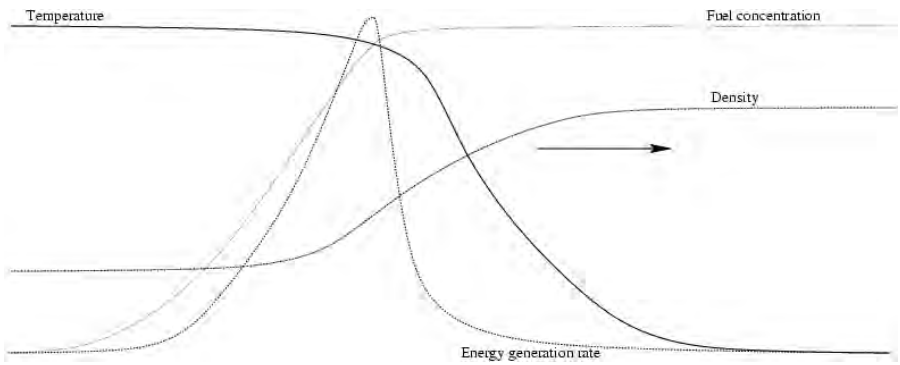


Figure 2. Sketch of the variation of flow properties across a deflagration front.

treatments by approximating the effect of turbulent mixing with the help of statistical methods.

The intermediate case of $K \approx 1$ is called the 'distributed burning regime' and is the least understood of all three. It recently received some attention in the framework of type Ia supernova models suggesting that the turbulent deflagration front, after beginning its life in the flamelet regime, may turn into a detonation (deflagration–detonation–transition, DDT) when it passes through the distributed burning regime. It is very difficult to study theories that involve distributed burning, since neither nuclear burning and diffusion nor turbulent mixing can be properly described by simplified models in this regime. Phenomenologically, the laminar flame structure is believed to be disrupted by turbulence and to form a distribution of reaction zones with various lengths and thicknesses.

Burning regimes in astrophysical flashes

This section gives a brief summary of the burning regimes of nuclear combustion fronts that are currently believed to occur in the three most studied astrophysical flashes: novae, x-ray bursts and type Ia supernovae.

In novae, hydrogen accreted onto a white dwarf star burns explosively. Even at the high temperatures $T \approx 3 \times 10^8$ K that may be reached in nova outbursts, τ_b for hydrogen burning in the CNO cycle is larger than the turnover time scale of turbulent eddies driven by the convective instability of the hot burning region at the base of the cold hydrogen envelope. Consequently, $K \ll 1$, and the burning proceeds in the well-stirred reactor regime. Additional carbon enrichment, for instance caused by convective dredge-up of white dwarf material, may locally enhance the burning rate significantly by acting as a catalyst for the nuclear reactions.

X-ray burst models invoke the nuclear runaway of accreted hydrogen, helium or a mixture of both on the surface of a neutron star. Very little is known about the propagation mode and burning regime of the combustion front. In the case of pure helium flashes, the front will most likely propagate as a convectively driven turbulent deflagration or even a detonation. An additional complication is the possible presence of dynamically important strong magnetic fields.

The currently most successful model for type Ia supernovae involves the thermonuclear explosion of a Chandrasekhar mass carbon–oxygen white dwarf. Here, it is known from spectroscopical observations of the explosion products that the star needs to expand before being consumed by the combustion front, ruling out the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

possibility of a single, centrally ignited detonation. If the explosion begins as a deflagration, the flame front immediately becomes turbulent due to the RT instability. Until the density has declined by roughly a factor of 100, the front is in the flamelet regime and therefore depends only weakly on the microphysics of burning and diffusion. The final fate of the turbulent flame front is still the subject of active research: it may (a) release a sufficient amount of energy to power the explosion exclusively by turbulent combustion, (b) transition into a detonation (delayed detonation) upon entering the distributed burning regime, or (c) fail to release enough energy to unbind the star, whereupon it re-collapses and possibly ignites a detonation (pulsational detonation).

Bibliography

Canal R and Ruiz-Lapuente P 1996 *Thermonuclear Supernovae* (Dordrecht: Kluwer)

Williams F A 1985 *Combustion Theory* (Menlo Park, CA: Benjamin/Cummings)

Zeldovich Ya B, Barenblatt G I, Librovich V B and Makhviladze G M 1985 *The Mathematical Theory of Combustion and Explosions* (New York: Consultants Bureau)

Jens Niemeyer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuclear Processing

Solar abundances, s-, r- and p-nuclei, reaction networks, thermonuclear reactions

Nuclear processing is the of the heavy elements (better, ISOTOPES) with mass numbers $A > 60$ in stars by thermonuclear non-explosive as well as explosive breeding processes. The fuel and seed material are gravitationally contained in the hot interior of the star. Efficient breeding of heavy nuclei is possible with a fuel of neutral nuclear particles, i.e. neutrons and photons. The heavy isotopic abundances are generated by processing iron seed with neutrons (bulk of material) and by reprocessing so-formed heavy abundances under shock heating. The isotopic abundances created in the stellar cauldrons are ejected into the interstellar medium.

The solar system is the object of the universe which provides the most complete set of high-quality cosmic abundance data. It is understood that solar system material is made of the ashes of many nucleosynthetic events that have been mixed in the interstellar medium before the formation of the solar system. The SOLAR ABUNDANCES are chiefly extracted from the spectrum of the Sun and from primitive METEORITES of unfractionized material (Anders and Grevesse 1989; Palme and Beer 1993).

The origin of the heavy isotopes can be understood with only three different nuclear formation processes, the s-, r- and p-processes (figure 1)(see, e.g., the reviews of Burbidge *et al* 1957; Meyer 1994). In the solar abundance distribution the nuclei of the s-process population show up in sharp maxima at mass numbers 88, 138 and 208 which correspond to isotopes of the magic neutron numbers 50, 82 and 126. The s-process nuclei are made by a neutron capture process (successive neutron captures) on a slow time scale. The synthesis path of the isotopes ${}^{A}Z$ with A the mass and Z the element numbers lies in the valley of β stability of the $Z,A–Z$ plane (figure 2). The nuclei of the r-process population show up in broader maxima near the mass numbers 80, 130 and 195. The r-process nuclei including ^{232}Th and ^{235,238}U are formed by a neutron capture process on a rapid time scale. The synthesis path is located far off the β stability valley at the neutron-rich side of the $Z,A–Z$ plane (figure 2). The peak structures are associated with the magic neutron shells at 50, 82 and 126 as well. The shift to smaller mass numbers compared with the s-process peaks and the broader structure of the peaks indicate that the abundances are the decay products of the originally synthesized unstable neutron-rich nuclei. Additionally there is an ensemble of neutron-deficient isotopes with considerably smaller abundances than r- and s-isotopes. These rare isotopes form a distribution with peaks at $A = 92$ and 144 which are also associated with the magic neutron shells 50 and 82. The p-abundances are suggested to be made from s- and r-process seeds shifted by photodisintegration reactions to unstable neutron-deficient isotopes. Post-p-process β decay then leads eventually to the abundances of the p-nuclei (figure 2).

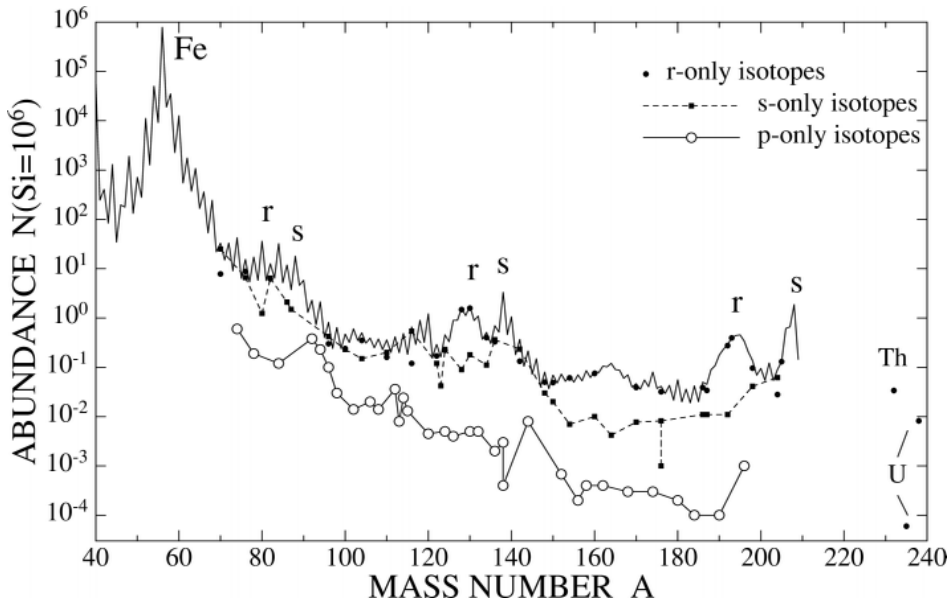


Figure 1. Solar system abundances by mass number of the heavy isotopes. The double peaks at A =80,88, 130,138 and 195,208 are from r- and s-process nucleosynthesis. The abundances of isobars which are added in the global curve are also shown separately as full circles (r-only isotopes) and full squares (s-only isotopes connected by a dashed line). The p-process abundance population (open circles) is significantly lower than the s- and r-process. (From Palme and Beer 1993.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

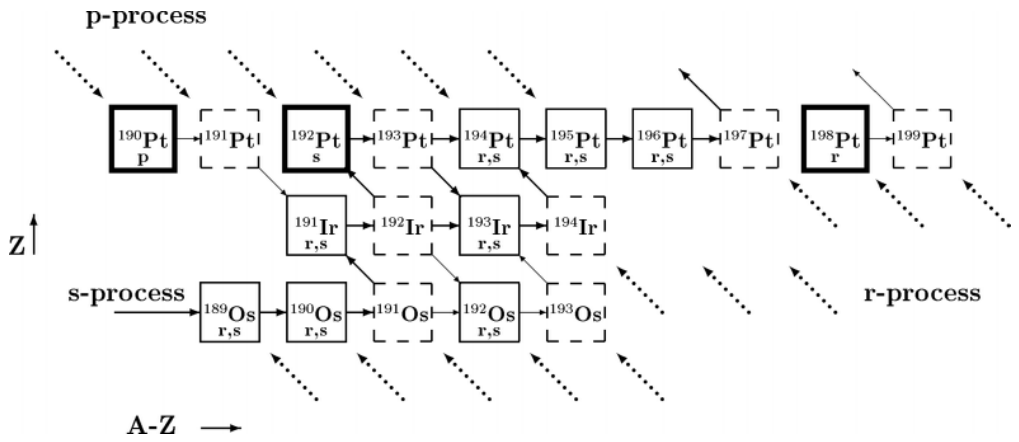


Figure 2. A section of the chart of nuclides in the stability valley and the s-process path through the stable and radioactive isotopes in this mass region (solid and dashed boxes). The p-only ^{190}Pt and r-only ^{198}Pt nuclei not on the path are destroyed in the s-process. At ^{191}Os, ^{192}Ir and ¹⁹³Pt the s-process flow is branched because of competition between β decay and neutron capture. The p- and r-process occurs off the stability valley. Abundance contributions to the stable isotopes are obtained via isobaric β decay chains. This flow of post r- and p-process is indicated by the inclined dotted arrows. The s-only ^{192}Pt isotope is shielded from r-abundance contributions.

The network equations can be written in the form

$$\frac{dN}{dt} = -\ (\text{sum of destruction terms}) + (\text{sum of creation terms}). \quad (1)$$

In s- and r-process nucleosynthesis these terms are essentially proportional to neutron capture reaction rates, λ^{n}, and beta decay rates λ^{β} of the types β^-, β^+ and electron capture decay; in the p-process to photodisintegration rates of neutrons, protons and α particles, $\lambda^{\gamma n}$, $\lambda^{\gamma p}$ and $\lambda^{\gamma\alpha}$, respectively. The neutron capture rate is given by $\lambda^{n} = n_n\langle\sigma_{n\gamma}v\rangle$ with the in-general time-dependent neutron density n_n and the radiative capture cross section $\sigma_{n\gamma}$ folded with the Maxwell–Boltzmann velocity distribution of the particles at the specified thermal energy kT, k being the Boltzmann constant and T the temperature. The radiative proton and alpha capture rates are defined equivalently. As for the majority of heavy isotopes $\sigma_{n\gamma}$ follows a v^{-1} energy dependence λ^{n} is approximately independent of temperature. It is also practice to form a Maxwellian average capture (MAC) cross section $\sigma = \langle\sigma_{n\gamma}v\rangle/v_T$ by dividing $\langle\sigma_{n\gamma}v\rangle$ by the thermal velocity $v_T = (2kT/\mu)^{1/2}$ with the reduced mass $\mu = M_A M_n/(M_A + M_n)$. M_n and M_A are the neutron mass and mass of the isotope, respectively (Beer *et al* 1992). Then λ^{n} can be expressed as $\lambda^{n} = n_n v_T \sigma$.

The photodisintegration rates $\lambda^{\gamma n}$, $\lambda^{\gamma p}$ and λ^{α} are related to the inverse radiative capture reactions by the requirements of statistical physics and thermodynamics. For example the (γ, n) rate is related to the (n, γ) reaction by

$$\lambda^{\gamma n}_{A,Z} = \frac{n_n N_{A-1,Z}}{N_{A,Z}}\langle\sigma_{n\gamma}v\rangle_{A-1,Z}. \quad (2)$$

The equilibrium concentrations of the neutrons n_n and the particles $N_{A,Z}$, $N_{A-1,Z}$ which form a non-relativistic, non-degenerate Boltzmann gas at temperature T are described by a nuclear Saha equation

$$\frac{n_n N_{A-1,Z}}{N_{A,Z}} = \frac{(2\pi\mu kT)^{3/2}}{h^3}\frac{g_n g_{A-1,Z}}{g_{A,Z}} \times \exp(-B_{n,A,Z}/kT) \quad (3)$$

with the binding energy $B_{n,A,Z}$ of the neutron in the nucleus ^{A}Z, h Planck's constant, μ the reduced mass of neutron and isotope ^{A-1}Z, $g_n = 2$ the statistical spin factor of the neutron, and with $g_{A,Z}$ the partition function for isotope ^{A}Z

$$g_{A,Z} = \sum_i (2j_{i,A,Z} + 1)\exp(-E_{i,A,Z}/kT).$$

The summation is over all excited states with spin $j_{i,A,Z}$ and energy $E_{i,A,Z}$.

s-Processing

The canonical s-process

According to the concept of equation (1) the buildup of nuclei in the s-process is formulated by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\frac{dN_{s,A,Z}}{dt} = \lambda^{n}_{A-1,Z} N_{s,A-1,Z} - \lambda^{n}_{A,Z} N_{s,A,Z} + \lambda^{\beta}_{A-1,Z} N_{s,A-1,Z} - \lambda^{\beta}_{A,Z} N_{s,A,Z} + \text{termination terms due to alpha decay at } A > 209 \tag{4}$$

where λ^{β} is zero if ${}^{A-1}Z$ and/or ${}^{A}Z$ is stable. In the case of an unstable isotope ${}^{A}Z$ and an s-process neutron density n_n between 10^7 and 10^9 cm^{-3} corresponding to capture lifetimes between 35 yr and 128 days for an MAC cross section of 300 mbarn the condition $\lambda^{n}_{A,Z} < \lambda^{\beta}_{A,Z}$ holds in general; then the term with $\lambda^{n}_{A,Z}$ is neglected and the term $\lambda^{\beta}_{A,Z} N_{s,A,Z}$ quickly transforms into $\lambda^{n}_{A,Z+1} N_{s,A,Z+1}$, the s-process synthesis path has reached a higher Z and remains in the valley of β stability of the Z,A–Z plane. However, as β decay rates near the stability valley can vary over many orders of magnitude (see NUCLEAR BETA DECAY), there are a number of cases where λ^{β} is comparable with λ^{n} and then the synthesis path is branched (figure 2).

Starting in equation (4) from ^{56}Fe as the major seed each isotope is formed and destroyed by neutron capture and β decay. In terms of a convenient notation $\psi_A = \sigma_A N_{s,A}/N_{\text{tot},0}$ with σ_A the MAC cross section for neutron capture of isotope A, N_A the abundance of isotope A on the synthesis path from ^{56}Fe to ^{209}Bi and N<inf>tot,0</inf> the total seed abundance, we formulate a set of differential equations dependent on the exposure $\tau = \int_0^t n_n v_T \, dt$:

$$\begin{aligned}
&\frac{1}{\sigma_{56}}\frac{d\Psi_{56}}{d\tau} = -\Psi_{56} \\
&\vdots \\
&\frac{1}{\sigma_A}\frac{d\Psi_A}{d\tau} = \Psi_{A-1} - \Psi_A \\
&\vdots \\
&\frac{1}{\sigma_{A'}}\frac{d\Psi_{A'}}{d\tau} = \Psi_{A'-1} - \left[1 + \frac{\lambda^{\beta}_{A'}}{\lambda^{n}_{A'}}\right]\Psi_{A'} \\
&\frac{1}{\sigma_{A''}}\frac{d\Psi_{A''}}{d\tau} = \frac{\lambda^{\beta}_{A'}}{\lambda^{n}_{A'}}\Psi_{A'} - \Psi_{A''} \\
&\vdots \\
&\frac{1}{\sigma_{A'''}}\frac{d\Psi_{A'''}}{d\tau} = -\Psi_{A'''} \\
&\vdots
\end{aligned} \tag{5}$$

The isotope A is a typical stable nucleus on the unique synthesis path, whereas A' represents a radionuclide with a β decay rate $\lambda^{\beta}_{A'}$ in competition with the neutron capture rate $\lambda^{n}_{A'} = \sigma_{A'} n_n v_T$. Note that the occurrence of radionuclide A' with coefficients containing explicitly the neutron density n_n requires a constant neutron density. Otherwise, the time dependence of n_n must also be considered.

The stable isotope A'' is, therefore, formed by the s-process in proportion to this competition. The quantities ψ_{56} and $\psi_{A''}$ refer to the seed-only isotopes ^{56}Fe and A''', respectively. The isotopes which feel the neutron density explicitly are the few located within a branch (see figure 2). Equations (5) can be written in matrix form $d\psi/d\tau = \mathbf{M}\psi$ with $\mathbf{M}$ the coefficient matrix. Equations (5) can be solved by special numerical techniques or analytically. The matrix $\mathbf{M}$ is diagonalized: $\mathbf{M} = \mathbf{ABA}^{-1}$. With $\mathbf{B}$ the diagonal matrix the equation is solved with $\exp(\mathbf{M}\tau) = \mathbf{A}\exp(\mathbf{B}\tau)\mathbf{A}^{-1}$. The matrix $\mathbf{A}$ is the matrix of eigenvectors with respect to the eigenvalues contained in $\mathbf{B}$. The solution is given by

$$\mathbf{N}(\tau) = (\boldsymbol{\sigma})^{-1}\mathbf{A}\exp(\mathbf{B}\tau)\mathbf{A}^{-1}\boldsymbol{\sigma}\mathbf{N}_0 \tag{6}$$

with $\boldsymbol{\sigma}$ the diagonal matrix of the MAC cross sections and $\mathbf{N}_0$ the vector of seed abundances. Special solutions to be derived from this ansatz are the Bateman solution which does, however, not include branchings. The solution $\mathbf{N}(\tau)$ for single exposures is relevant to special s-process scenarios, s-processing of nuclei in the mass range 56–90 (weak s-process, figure 3) and from sulfur to calcium.

The s-process formation of the bulk of heavy isotopes ($85 < A < 209$) (main s-process, figure 3) is not the result of one exposure τ but a superposition of exposures $\rho(\tau)$. The differential equations describing the synthesis have to be integrated over that distribution. For a stable isotope A on the unique synthesis path we obtain

$$\begin{aligned}
&\frac{1}{\sigma_A}\int_0^\infty \rho(\tau)\frac{d\Psi_A(\tau)}{d\tau}\,d\tau \\
&\quad = \int_0^\infty \rho(\tau)\Psi_{A-1}(\tau)\,d\tau - \int_0^\infty \rho(\tau)\Psi_A(\tau)\,d\tau.
\end{aligned} \tag{7}$$

Partial integration of the left-hand side of equation (7) yields

$$\begin{aligned}
&-\frac{1}{\sigma_A}\Psi_A(0)\rho(0) - \frac{1}{\sigma_A}\int_0^\infty \Psi_A(\tau)\frac{d\rho(\tau)}{d\tau}\,d\tau \\
&\quad = \sigma_{A-1}N_{s,A-1} - \sigma_A N_{s,A}.
\end{aligned} \tag{8}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

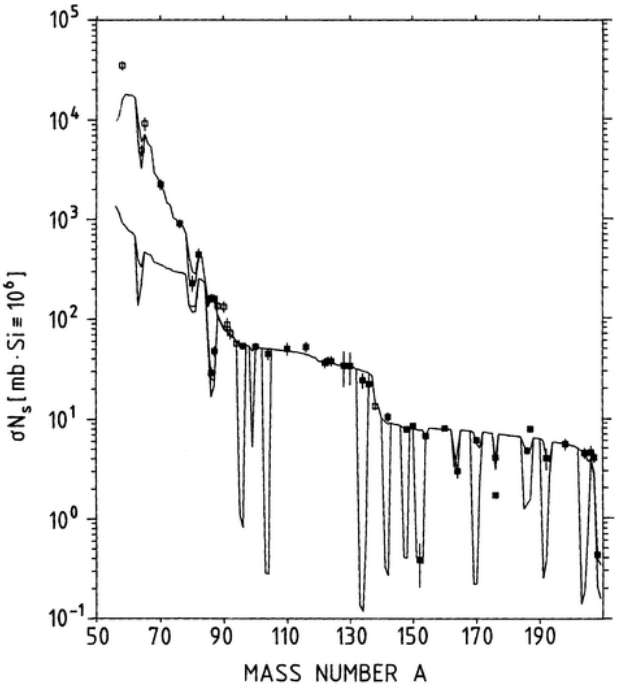


Figure 3. σ*N* curve versus mass number, characterized by a ledge precipice structure at nuclei with magic neutron shells (*A* = 88, 138 and 208) and flat portions between with approximately equilibrium abundance flow. The negative spikes indicate branchings. A weak s-process component is superimposed responsible for the steep increase for $A < 90$. Not all empirical data points shown are used to adjust the curves. Full symbols are s-only nuclei. Open symbols are data at the magic neutron numbers 50 and 82 (^{88}Sr, the Zr isotopes and ^{138}Ba) with dominant s-process abundance as well as the data points ^{58}Fe, ^{64}Ni and ^{65}Cu. Other empirical data too low (^{176}Lu) or too high (^{187}Os, ^{206,207,208}Pb) are affected by long-lived radioactive decay and in the case of ^{206,207,208}Pb contain extra r-process contributions from transbismuth nuclei and from an additional strong s-process exposure. (From Beer 1991.)

As the empirical σN_s values of s-only isotopes are decreasing with increasing mass number a monotonically decreasing function $\rho(\tau)$ is required. The form $d\rho(\tau)/d\tau = -\tau_0^{-1}\rho(\tau)$ with the mean exposure τ_0 yields analytical solutions. From equation (8) we obtain the recursion formula

$$\left(1+\frac{1}{\tau_0\sigma_A}\right)\sigma_A N_{s,A} = \sigma_{A-1}N_{s,A-1} + \tau_0^{-1}N_{A,0} \tag{9}$$

and in the same way from the two differential equations describing a branching at the nucleus *A*' (see equation (5)) the recursion formulae

$$\left(1+\frac{\lambda^{\beta}_{A'}}{\lambda^{n}_{A'}}+\frac{1}{\tau_0\sigma_{A'}}\right)\sigma_{A'}N_{s,A'} = \sigma_{A'-1}N_{s,A'-1}$$

$$\left(1+\frac{1}{\tau_0\sigma_{A''}}\right)\sigma_{A''}N_{s,A''} = \frac{\lambda^{\beta}_{A'}}{\lambda^{n}_{A'}}\sigma_{A'}N_{s,A'} + \tau_0^{-1}N_{A'',0}. \tag{10}$$

Finally for an isotope *A*''' not lying on the synthesis path we obtain

$$\left(1+\frac{1}{\tau_0\sigma_{A'''}}\right)\sigma_{A'''}N_{s,A'''} = \tau_0^{-1}N_{A''',0}. \tag{11}$$

The quantity $N_{A,0}$ represents the seed abundance of the nuclide *A*. The exposure distribution $\rho(\tau)$ is given by $\rho(\tau) = G\exp(-\tau/\tau_0)$ with $G = N_{\mathrm{tot},0}/\tau_0$, $N_{\mathrm{tot},0}$ the total seed abundance. This form was suggested to arise from the effect of galactic reprocessing, i.e. the total exposure experienced by some fraction of material would relate to the number of times that material had been processed through stars (Seeger *et al* 1965).

We write the final result for all isotopes involved in the s-process in the convenient matrix form (Ward *et al* 1976; Beer *et al* 1997)

$$(\mathbf{M}-\mathbf{U}\tau_0^{-1})\sigma\mathbf{N} = (-\tau_0)^{-1}\sigma\mathbf{N}_0 \tag{12}$$

where **U** is the unity matrix. The vector $\mathbf{N}(\tau_0)$ represents the wanted abundances. Normally a solar seed abundance distribution $\mathbf{N}_\odot$ is used and a seed fraction F_{seed} defined by $\mathbf{N}_0 = F_{\mathrm{seed}}\mathbf{N}_\odot$. If only the dominant ^{56}Fe is taken as seed material, the vector of solar seeds $\mathbf{N}_0$ reduces to a scalar as only the first element is non-zero $N_0(1) = F_{\mathrm{seed}}N_\odot$ (^{56}Fe) (Ward *et al* 1976). Without branchings and ^{56}Fe as the only seed material, $\sigma_A N_{s,A} = \xi_A\xi_{A-1}\ldots\xi_{56}\tau_0 - {}^1\mathrm{N}_0(1)$ is simply the product of propagators $\xi;_A = (1 + (\sigma_A\tau_0)^{-1})^{-1}$. An s-process analysis using the analytical solutions of Ward *et al* (1976) is shown in figure 3 (Beer 1991). To treat reprocessing the s-process solutions of equations (6) and (12) can be used as seed distribution input for s-processing by a second exposure (Beer *et al* 1997).

In the solutions equations (9), (10) and (11) of the main component s-process the efficiency of the s-process exposure of the isotopes on and beyond the synthesis path depends on the size of their cross sections (Beer *et al* 1997). For $\sigma_A \gg \tau_0^{-1}$ stable isotopes on the path are characterized by the well-known correlation, $N_A \sim \sigma^{-1}{}_A$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(equation (9)), and stable isotopes beyond the path are destroyed, $N_{A''} \approx 0$ (equation (11)). For $\sigma_A \approx \tau_0^{-1}$ stable isotopes on and beyond the path are sensitive to the average exposure τ_0 (bottle-neck or waiting-point isotopes) and may contain information on the burning temperature of the neutron source. The synthesis is controlled by these isotopes. For $\sigma_A \ll \tau_0^{-1}$ the abundances of stable isotopes on the path become independent of the MAC cross section and are determined by τ_0 which is a function of temperature. The seed abundances of isotopes beside the path are practically unaffected. The MAC cross sections of the isotopes on the synthesis path have been compiled and tabulated as a function of temperature (Bao *et al* 2000). For unstable isotopes with $\lambda_{A'}^{\beta} \approx \lambda_{A'}^{n} = \sigma_{A'} n_n v_T$ a branching in the synthesis path occurs which constrains n_n. In addition the β decay rate can be a strong function of temperature and even electron density which in turn depends on the mass density of the stellar environment; then also temperature T and mass density are constrained (figure 2). Calculated values of the β decay rate $\lambda_{A'}^{\beta}$ as a function of the temperature and electron density for the significant branch point isotopes have been published (Takahashi and Yokoi 1987).

Experimental and observational constraints, stellar models

The s-processing of the heavy nuclei with the canonical s-process using a best set of chiefly experimentally determined input parameters (figure 3) requires two components, the weak and main components. Each component is globally characterized by the fraction of seed material and the neutron exposure. The bottle-neck isotopes Sr, Y, Zr, Ba, La, Ce and ^{208}Pb with magic neutron shells at 50, 82 and 126 control the main s-process synthesis. The processing can go up to Pb and Bi if ^{138}Ba is bypassed; ^{138}Ba is, next to ^{208}Pb, the isotope with the smallest MAC cross section on the unique synthesis path ($\sigma(^{138}\text{Ba}) = 4.0\pm0.2$ mbarn, $\sigma(^{208}\text{Pb}) = 0.36\pm0.03$ mbarn at $kT = 30$ keV). The required mean exposure must therefore be comparable with $\tau_0(kT = 30 \text{ keV}) \approx \sigma(^{138}\text{Ba})^{-1} = 0.25$ mbarn^{-1} (value found in the calculations $\tau_0(kT = 30 \text{ keV}) \approx 0.30$ mbarn^{-1}) (Beer *et al* 1997). Only the MAC cross section of ^{208}Pb is still smaller and requires, therefore, an extra-strong exposure. At the nuclei with a magic neutron shell, abundance is accumulating because of the small destruction rate. For the majority of nuclei lying in flat portions of the σN curve with practically steady flow conditions the $N_A \approx \sigma_A^{-1}$ correlation is approximately fulfilled. From the local abundance structures of branchings, neutron and mass density and temperature are deduced ($n_n = (2\text{–}4) \times 10^8$ cm^{-3}, $\rho = (2.6\text{–}13) \times 10^3$ g cm^{-3} $kT = 25\text{–}30$ keV).

The basic questions of s-processing to be clarified are the possible stellar site or sites and the neutron source or sources which produced the solar s-abundances. Weak s-processing can be located in the STELLAR EVOLUTION of massive stars ($M \geq 15M_\odot$) from the core He burning to the subsequent shell C and Ne burning phases. Neutrons are provided by the ^{22}Ne reaction. The ^{22}Ne being produced from the ashes of the pre-existing nuclei of the CNO cycle via the reaction sequence $^{14}\text{N}(\alpha,\gamma)^{18}\text{F}(\beta^+)^{18}\text{O}(\alpha,\gamma)^{22}\text{Ne}$ makes the s-process yields increase with metallicity. In the stellar model calculations the accuracy of the ^{22}Ne(α, n) and competing ^{22}Ne(α,γ) rates is crucial for the final results.

For the main s-process component the temperature estimated with the canonical s-process would suggest also the ^{22}Ne(α, n) neutron source and shell He burning in the stellar interior of thermally pulsing intermediate-mass stars (IMSs; with $4 < M/M_\odot < 8$) as the s-process site. The exponential exposure distribution was shown to derive from the neutron release during the convective instabilities in the He shell (Ulrich 1973). It is obtained if the total exposure of the seed material occurs in a series of individual neutron exposures, where a decreasing fraction of seed material $r, r^2, r^3, \ldots, r^l$ ($r < 1$) obtains an increasing number of exposures $\Delta\tau$, 2 $\Delta\tau$, 3 $\Delta\tau$,…,$l\Delta\tau$. The exponential exposure distribution $\rho(\tau = l\Delta\tau)$ is given as follows: $\rho(\tau = l\Delta\tau) = Gr^l = G[\exp(-\Delta\tau/\tau_0)]^l = G\exp(-l\Delta\tau/\tau_0) = G\exp(-\tau/\tau_0)$ (Ulrich 1973). This would relate the mean exposure τ_0 to r and $\Delta\tau$ (note the equivalent arguments in the case of reprocessing in stars during the galactic s-process enrichment). However, IMSs which could provide this scenario are too rare and the predicted temperature and neutron densities were far too high to reproduce the abundance patterns of the s-process branchings. In addition the increase of the s-process yields with metallicity associated with the ^{22}Ne(α, n) neutron source and an enhancement of ^{25}Mg in these stars are not observed.

Since the discovery of s-process Tc in the stellar atmospheres of RED GIANTS (the short 2.1×10^5 yr half-life of ^{99}Tc suggests Tc nucleosynthesis in the star itself) elemental s-process abundances from Rb to Nd were found and studied in many low-mass asymptotic giant branch (AGB) stars. The s-process enrichment in these cool MS and S type stars (surface temperatures between 2800 and 3800 K) is considered intrinsic in contrast to s-enhancements in the hotter Ba stars interpreted to originate from mass overflow in a binary system. The observation of Rb and ^{96}Zr in the MS and S stars gives constraints for a low neutron density in accordance with the ignition of the ^{13}C(α, n) neutron source at a temperature of only $T_8 = 1$. Additional evidence for a dominantly primary ^{13}C s-process neutron source comes from studies of the observed galactic enrichment of s-process yields which are independent of metallicity. From very low metallicity stars very strong neutron exposures can then be expected which are necessary to build up the abundance of ^{208}Pb.

The idea that the solar s-process abundances are a mixture of s-process yields from different stellar sources

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is also supported by cosmochemistry. The primitive unfractionized meteorites which represent well in their total composition the solar abundance mixture also carry presolar SiC grains containing elements with preserved isotopic abundance patterns of the s-process. These grains are supposed to be formed in the cicumstellar environments of low-mass AGB stars. Evidence for this s-process origin is the correlation between the isotopic ratios and the inverse of the corresponding MAC cross sections (e.g. s-Xe, s-Nd). The isotopic ratios of s-Kr, s-Sr and s-Ba which are sensitive to details of the s-process neutron exposure show additional variations for different grains. The scatter of the data with regard to the isotopes ^{86}Kr, ^{88}Sr and ^{138}Ba is interpreted to indicate an origin from a variety of stars with different masses and neutron exposures and suggests that the solar system s-process abundances are indeed an average over many sources.

Low-mass thermally pulsing AGB stars ($1 < M/M_{\odot} < 3$), where the ^{13}C(α,n) reaction is most likely the dominant neutron source, are therefore the most favored site for the main s-process component. The ^{13}C is supposed to be produced in a thin layer of the He shell by the reaction chain ^{12}C(p,γ)^{13}N(β^{+})^{13}C from primary ^{12}C with protons which have penetrated the envelope by diffusion or semiconvection. The formation and size of the ^{13}C pocket which are crucial for the quantitative description are still treated in a parametrized way (Gallino *et al* 1998). It should be noted that in these stellar model calculations the solar s-process abundances are reproduced for the supposed model star with an exposure distribution that is much more complex than the exponential distribution of neutron exposure. In order to meet the constraints on s-process temperature imposed by the branching analysis of the canonical s-process the ^{22}Ne neutron source must also be invoked marginally towards the end of the thermal pulse, providing a small additional exposure to reset the abundances of the branch point thermometer isotopes to the higher temperature of the ^{22}Ne(α, n) ignition. As the average MAC cross section in the mass region of the s-process thermometers is of the order of 300 mbarn this reprocessing is possible for a mean exposure of $\tau_0(kT = 30\ \text{keV}) \approx 1/300 = 0.0033$ mbarn^{-1} without affecting the global structure of the s-process determined by the very small cross sections of the bottle-neck isotopes (see also figure 4). Evidence for the ^{13}C(α, n) source is also obtained from a significant temperature dependence of the neutron capture rate λ^{n} of some bottle-neck isotopes (e.g. ^{142}Nd, ^{208}Pb; see also the general temperature dependence of λ^{n}). The stellar model is put to the test by laboratory neutron capture measurements (see, e.g., Guber *et al* 1997, Koehler *et al* 1998, Arlandini *et al* 1999).

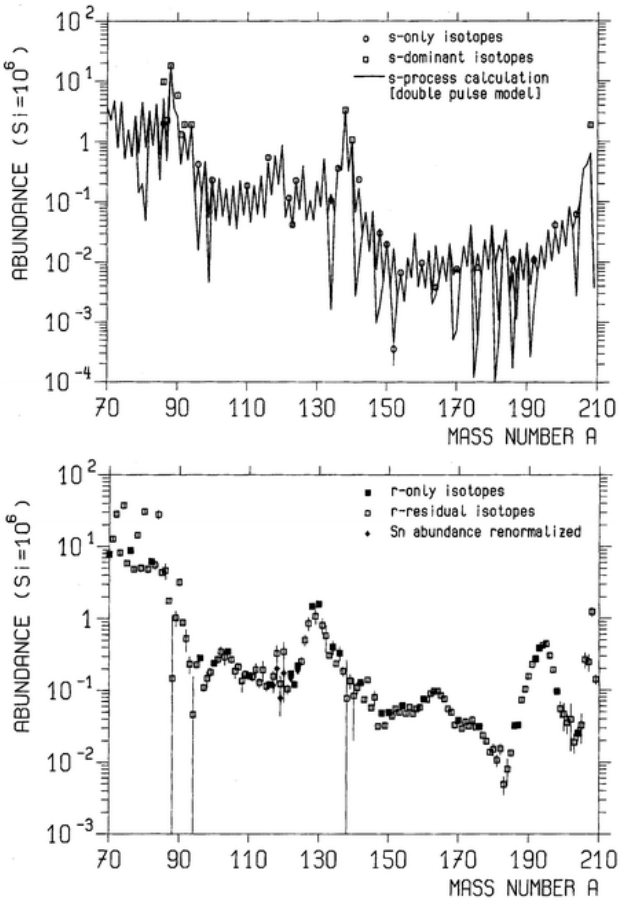


Figure 4. Decomposition of the solar abundances. (Top) Parametrized s-process abundance calculation with combined neutron exposure at $kT = 8$ and 28 keV associated with the ^{13}C(α, n) and ^{22}Ne(α, n) neutron sources, respectively. The calculation is a fit to the empirical s-only isotopes shown as full symbols. The typical signature of the s-process, an odd–even abundance staggering, is visible. (Bottom) The empirical r-process distribution $N_r = N_{\odot} - N_s$ exhibits a smooth behavior. The r-only isotopes are shown as full symbols. Note that for the similarities of s- and r-isotopes produced in about equal amounts no obvious genetic relation is known. (From Beer *et al* 1997.)

r-Processing

Empirical r-process distribution

The empirical r-process distribution is obtained by subtraction of the calculated s-contributions from the solar abundances. Possible p-process contributions are small and were, therefore, neglected:

$$N_{r,A,Z} = N_{\odot,A,Z} - N_{s,A,Z}. \tag{13}$$

The so-constructed r-process distribution (figure 4) with abundance maxima at A = 80, 130 and 195 (the waiting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

points) related to the magic neutron shells at 50, 82 and 126 is fundamental. It forms the basis for comparisons with r-process model calculations. Tables of the calculated empirical r-process distribution have been reported elsewhere (Käppeler *et al* 1989; Palme and Beer 1993; Beer *et al* 1997; Arlandini *et al* 1999).

The canonical r-process

The general equation for the r-process is

$$\begin{aligned}\frac{\mathrm{d}N_{\mathrm{r},A,Z}}{\mathrm{d}t} &= \lambda^{\mathrm{n}}_{A-1,Z}N_{\mathrm{r},A-1,Z} - \lambda^{\mathrm{n}}_{A,Z}N_{\mathrm{r},A,Z} \\ &+ \lambda^{\beta}_{A-1,Z}N_{\mathrm{r},A-1,Z} - \lambda^{\beta}_{A,Z}N_{\mathrm{r},A,Z} \\ &+ \lambda^{\gamma\mathrm{n}}_{A-1,Z}N_{\mathrm{r},A-1,Z} - \lambda^{\gamma\mathrm{n}}_{A,Z}N_{\mathrm{r},A,Z} \\ &+ \text{termination terms due to fission for } A > 260.\end{aligned} \tag{14}$$

Successive neutron captures in a thermonuclear environment ($n_{\mathrm{n}} > 10^{20}$ cm^{-3}, $T > 10^{9}$ K) will continuously build up for each isotopic chain more and more neutron-rich isotopes on a rapid time scale (10^{-4} s) compared with the much longer β decay half-lives (10^{-1}–10^{-3} s) (λ^{n}>λ^{β}). This goes on until an equilibrium is reached via the reactions

$$^{A-1}Z + \mathrm{n} \rightleftharpoons {}^{A}Z + \gamma. \tag{15}$$

This is because continuous neutron capture leads to nuclei with decreasing neutron binding energy. This process will go on until a neutron binding energy of B_{n} = 2–4 MeV is reached because of the incidence of (γ, n) reactions at the temperature of the thermonuclear environment. If this point is reached with $\lambda^{\mathrm{n}}_{A,Z} \approx \lambda^{\gamma\mathrm{n}}_{A-1,Z}$ and

$$\frac{\mathrm{d}N_{\mathrm{r},A,Z}}{\mathrm{d}t} = \lambda^{\beta}_{A,Z-1}N_{\mathrm{r},A,Z-1} - \lambda^{\beta}_{A,Z}N_{\mathrm{r},A,Z} \tag{16}$$

then the isotopes have to wait for β decay to isotopes with $Z + 1$ where quickly a new different (n, γ) ⇌ (γ, n) equilibrium is established. At the magic neutron shells 50, 82 and 126 β decays are long and therefore large abundances are accumulated. In (n, γ) ⇌ (γ, n) equilibrium the relative abundances for isotopes of a given element are determined by the properties of a Boltzmann gas at temperature T. The nuclear abundances and the neutron density are related by a nuclear Saha equation (3). Via equation (3) the distribution of nuclei of an isotopic chain Z is determined. We can describe the abundance flow from one isotopic chain to the next as (Seeger *et al* 1965)

$$\begin{aligned}\frac{\mathrm{d}N_{\mathrm{r},Z}}{\mathrm{d}t} &= \sum_{A}\lambda^{\beta}_{A,Z-1}N_{\mathrm{r},A,Z-1} - \sum_{A}\lambda^{\beta}_{A,Z}N_{\mathrm{r},A,Z} \\ &= \lambda^{\beta}_{Z-1}N_{\mathrm{r},Z-1} - \lambda^{\beta}_{Z}N_{\mathrm{r},Z}\end{aligned} \tag{17}$$

where we have defined

$$N_{\mathrm{r},Z} = \sum_{A}N_{\mathrm{r},A,Z} \qquad \text{and} \qquad \lambda^{\beta}_{Z} = \frac{1}{N_{\mathrm{r},Z}}\sum_{A}\lambda^{\beta}_{A,Z}N_{\mathrm{r},A,Z}.$$

Starting from ^{56}Fe as seed material the canonical r-process characterized by an ^{A-1}Z + n ⇌ ^{A}Z + γ equilibrium proceeds for constant neutron density and temperature along a contour line of constant neutron binding energy B_{n} towards the heavy nuclei according to the nuclear Saha equation. This binding energy B_{n} is estimated in the abundance peak maxima where the nuclear flow of a chain is concentrated mainly at one isotope pair with $N_{A-1,Z}/N_{A,Z} \approx 1$. Via the equations (17) for all element numbers Z on the r-process path a set of differential equations has to be solved by special numerical techniques as a function of duration time t_{dur}. With the solutions $N_Z(t_{\mathrm{dur}})$ and the relative isotopic abundances from the Saha equation (3) isotopic abundances are determined. The final r-process abundances of the stable isotopes are obtained in the post-r-process phase after radioactive decay. Therefore, besides the nuclear properties, i.e. binding energies and essentially β decay rates, the astrophysical quantities n_{n}, T and t_{dur} are sufficient to predict the whole set of r-abundances.

Experimental and observational constraints, stellar models

The unstable isotopes of the r-process path, 15–35 mass units off the valley of β stability in the Z,A–Z plane, are in their majority not studied directly in laboratory experiment. Theoretical nuclear physics has to provide a reliable data basis: mass formulas to determine r-process paths (equation (3)), β decay rates to solve equation (17) eventually with effective β decay rates, excited states properties of nuclei on the path for the partition functions (equation (4)), fission barriers and yields to assess the termination of the r-process, β delayed neutron emission and fission and α decay half-lives for the post r-process phase to obtain the r-process yield at the stable isotopes. To improve the applied nuclear mass formulas quenching of shell effects was suggested. This led to considerably better reproductions of the r-abundances around A = 120 and 140. An important clue to an understanding of the r-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

process was given by the measurement of the half-lives of waiting point nuclei at the magic neutron shells 50 and 82. These data give evidence that the r-process achieves really ${}^{A-1}Z + n \rightleftharpoons {}^{A}Z + \gamma$ equilibrium and steady flow locally in the abundance peaks (Kratz *et al* 1993). Eventually the r-process calculations show that after each r-process abundance maximum the ${}^{A-1}Z + n \rightleftharpoons {}^{A}Z + \gamma$ equilibrium breaks down so that the reproduction of the empirical r-process distribution requires for the specified temperature ($T_9 = 1.35$) at least three different canonical r-process paths superimposed with three sets of neutron densities (10^{20}–10^{24} cm^{-3}). The duration time of the synthesis lies between 1.2 and 2.3 s (figure 5) (Kratz *et al* 1993). An important question about the canonical r-process is a check of the validity of the waiting point approximation by a boundary condition which marks the transition from equilibrium to conditions that require accounting for the competition of neutron capture and β decay. To what extent this limit, which is dependent on the local nuclear parameters, affects the r-abundances at freeze-out is a matter of debate. Detailed investigations stress the importance of the transition phase at r-process freeze-out for the final abundances. The suggested mechanisms of freeze-out can prevent abundance troughs in the r-process calculations ascribed to deficiencies of the nuclear input physics and explain the observed smoothness of the empirical r-process abundance distribution (figure 4).

The yield of r-process elements in the galactic chemical evolution appears to require a neutron source which is not dependent on initial stellar metallicity. This is one of the arguments against a secondary r-process in an explosive environment fuelled by neutrons from the ^{22}Ne(α, n) source interacting with preexisting heavy element seeds (e.g. nova outbursts, shock-induced explosive He or C burning).

The heavy element abundance pattern of old very low metallicity stars matches rather well the corresponding solar system r-process (derived from the empirical r-process, figure 4, by summation over isotopic abundances). This would suggest that the r-process is primary and contributed to the galactic element enrichment before the s-process. A secondary r-process would have to rely on an already abundant s-process enrichment. From the resemblance of the elemental stellar r-process and solar r-process distributions for elements heavier than Ba it is concluded that a unique r-process mechanism exists that always produces the abundances in similar proportions. Also, in these low-metallicity old stars, indications of the existence of a second r-process source for elements in the nuclear mass range A ≈ 80–120 were detected (Sneden *et al* 2000). The definitive identification of these heavy elements in the low-metallicity stars as r-process products would, however, require the isotopic abundance determination.

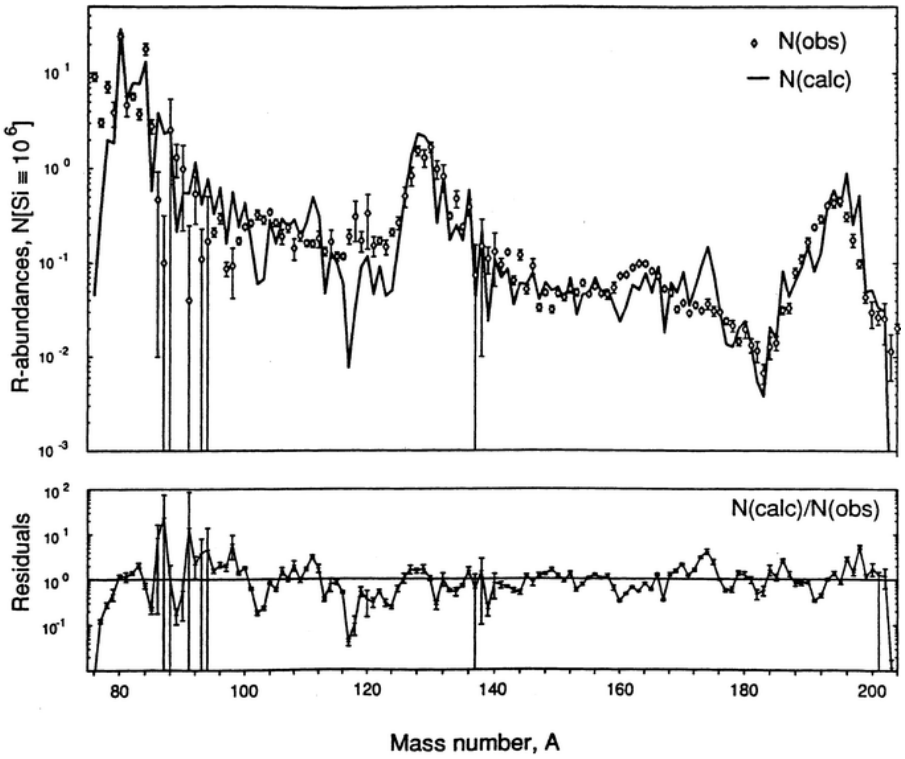


Figure 5. Global r-process abundance curve obtained from a superposition of three time-dependent calculations with the best fit n_n–T_9 values for the $A = 80$ peak and the $90 \leq A \leq 130$ and $135 \leq A \leq 195$ mass ranges. The weights of the individual components are 10:2.6:1. In the lower part the ratio of calculated and observed r-process abundances is shown. (From Kratz *et al* 1993.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The question of the site of the r-process is not settled. Several scenarios, ranging from the neutrino-driven winds in type II SUPERNOVA explosions (Woosley *et al* 1994) to merging neutron stars, have been suggested. Supernova explosions of type II, understood to be initiated by the iron core collapse of a massive star with the formation of a NEUTRON STAR, are assumed to be a prominent site for an r-process at the mass cut of matter exploded and locked up in the neutron star. However, it was also argued that the total mass fraction of r-process material of $\sim 10^{-7}$ (figure 4) which amounts to only $\sim 10^{-7} \times 10^{11} M_\odot \approx \sim 10^4 M_\odot$, where $\sim 10^{11} M_\odot$ is the total mass of the Galaxy, would require a contribution per supernova explosion over the history of the Galaxy (estimated age of the Galaxy at the birth of the solar system $\sim 10^{10}$ yr) of only $\sim 10^4 M_\odot / (10^{-2} \times 10^{10}) \approx \sim 10^{-4} M_\odot$. This estimate, where a supernova rate of $\sim 10^{-2}$ yr^{-1} is assumed, would suggest that either the r-process occurs for only a small subset of supernova explosions or the efficiency of r-production is as small as $\sim 10^{-4} M_\odot$. The latter possibility has been studied with some success in the framework of a primary r-process model site located at the formation of a hot neutron star predominantly cooling via neutrino emission (Woosley *et al* 1994). The compelling new scenario, a neutrino-driven explosion, appears to meet the time scales, temperature and neutron density conditions for a canonical r-process with β-flow equilibrium in the r-abundance peaks beyond nuclear masses A=110. In addition, as most neutron stars possibly pass through a similar early cooling development, the expectation that the same r-process relics will always be observed in old low-metallicity stars seems to be fulfilled.

p-Processing

Photodisintegration p-process

In the p-process it is assumed that preexisting r- and s-process seed nuclei are exposed to a high temperature ($2.1 \leq T_9 \leq 3.2$) for a short time (≈1 s). The first reactions to occur are the (γ, n) reactions which produce quite proton-rich nuclei. With each neutron ejected the neutron binding energy B_n becomes on the average larger and the photoneutron reaction rate $\lambda^{\gamma n}$ slower (see equations (2) and (3)). At the same time the binding energy for a proton and α particle is decreasing and the rate for photodisintegration accompanied by charged-particle emission $\lambda^{\gamma p}$, $\lambda^{\gamma \alpha}$, is increasing. Once the nuclei become sufficiently proton rich about 5 mass units away from the stability valley, they begin a (γ, p) and (γ,α) cascade. The reaction network consists, therefore, of equations of the form

$$\frac{dN_{p,A,Z}}{dt} = -(\lambda^{\gamma n}_{A,Z} + \lambda^{\gamma p}_{A,Z} + \lambda^{\gamma \alpha}_{A,Z}) N_{p,A,Z} + \lambda^{\gamma n}_{A+1,Z} N_{p,A+1,Z} + \lambda^{\gamma p}_{A+1,Z+1} N_{p,A+1,Z+1} + \lambda^{\gamma \alpha}_{A+4,Z+2} N_{p,A+4,Z+2} \quad (18)$$

ordered in terms of decreasing element and mass numbers. In this way nuclei are shifted towards iron. In the already large network of equations (18) to be solved the possible reactions of the n, p and α background particles as well as the weak interactions ((e^-,ν), (e^+,ν)) are neglected owing to assumed particle reaction and weak decay time scales much longer than the p-process duration. This makes the results independent of the composition and density of the p-process site. The results are only dependent on the abundance distribution of the input seed nuclei and the photon temperature. During the synthesis abundance is accumulating especially at the nuclei, where the photodisintegration flow is deflected to lower-Z elements by the (γ, p) or (γ,α) reactions because of a long disintegration time (waiting points). When the disintegration flow crosses the closed neutron shells at 50 and 82, the disintegration time scales become especially large because of the particularly strong binding energies. The abundances are then built up directly at the closed shell nuclei ^{92}Mo and ^{144}Sm. The abundances of the primary unstable p-nuclei eventually β decay after freeze-out to stable p-nuclei. The concept of the photodisintegration p-process was elaborated by Woosley and Howard (1978).

No single temperature exposure can produce all p-nuclei. However, with a distribution of exposures an approximate representation of the solar abundances is obtained. This is true also for the extended network including the n, p and α background reactions (figure 6) (Rayet *et al* 1990). Only the (n, γ) reactions have some influence on the yield of individual p-nuclei. In particular, it appears to be impossible to reproduce the solar abundances of the isotopes ^{92,94}Mo and ^{96,98}Ru by the photodisintegration p-process because they are too abundant and there is insufficient s- and r-process seed at $A > 96$. In the search for possible solutions of this problem, a significantly higher s-process seed abundance is suggested (Costa *et al* 2000), which would result from a higher rate of the ^{22}Ne(α,n)^{25}Mg neutron-producing reaction varied within the quoted uncertainty. In another suggestion, a different synthesis process, the rp-process on accreting neutron stars, is considered (Schatz *et al* 2001).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

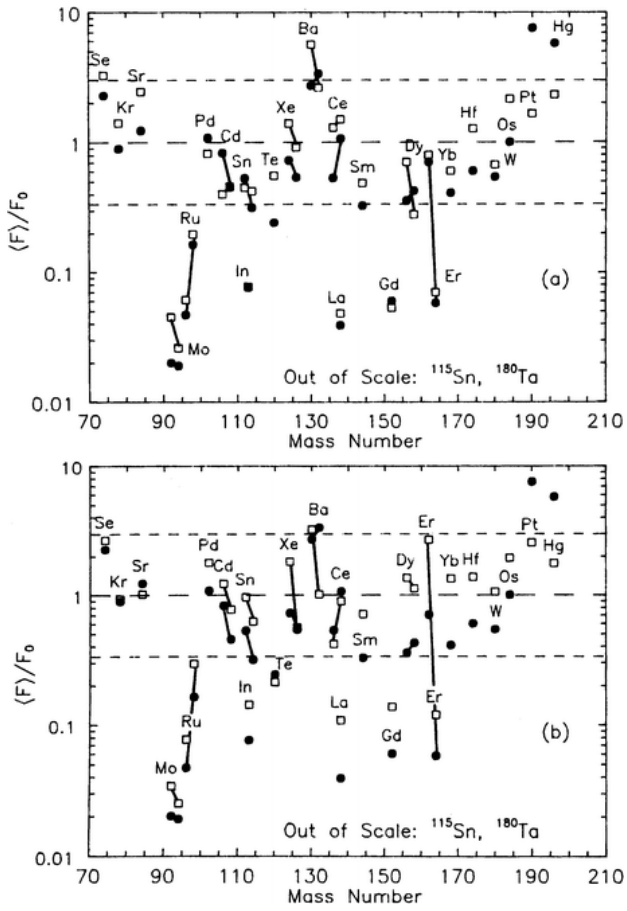


Figure 6. Relative overproduction factors for p-process nuclei. Lines connect isotopes of the same element. (*a*) Predictions for a uniform weighting of the initial temperatures: filled and open symbols correspond to different time scales and mean overproduction factors (t_{HYD}(s), F_0) = (0.446,200) and (1, 145), respectively. (*b*) Comparison of predictions for t_{HYD} = 0.446 s for uniform weighting of the initial temperatures (filled symbols) and for a best fit mix of temperatures (open symbols and F_0 = 114). (From Rayet *et al* 1990.)

Experimental and observational constraints, stellar models

The correlation found between neutron binding energy (hence photoneutron disintegration rate) and solar abundance of the p-process nuclei represents the first experimental evidence pointing to nuclear p-processing in terms of a photodisintegration process (Macklin 1970). This correlation is verified more properly between the total photodisintegration rate ($\lambda^{\gamma n} + \lambda^{\gamma p} + \lambda^{\gamma \alpha}$) of the synthesized p-nuclei and solar p-abundances at that mass number.

An important constraint on the p-process is the presence of live ^{146}Sm in the early solar system detected in meteorites and the inferred production ratio of ^{146}Sm/^{144}Sm. It appears to be a challenge to reproduce this ratio with the photodisintegration p-process.

From the success of the parametrized photodisintegration models it was astrophysically plausible to place p-processing in the oxygen/neon shell of a highly evolved massive star, exploding as a type II supernova. Such calculations were performed for the SUPERNOVA 1987A (Prantzos *et al* 1990). The supernova shock that heated up this shell provided temperature and density profiles to initiate the photodisintegration reactions on the preexisting heavy isotope s-process seed from core He burning. The results are essentially similar to the parametrized model calculations. A detailed review of the origin of the p-process nuclei was given by Lambert (1992).

Epilogue

As the understanding of nuclear processing in stars relies mainly on the analysis of the solar system heavy element composition it has to be stressed that the stellar evolution and composition of stars including stellar mixing processes and the chemical evolution of the galaxy have to play a large role in the calculations. For example the solar r-process abundances are assumed to result from contributions of ~10^8 supernova explosions.

Bibliography

Concepts of stellar nuclear s-, r- and p-processing were first reviewed in the article

Burbidge E M, Burbidge G R, Fowler W A and Hoyle F 1957 Synthesis of the elements in stars *Rev. Mod. Phys.* **29** 547–650

The concept of the canonical s-process with an exponential exposure distribution and detailed canonical s- and r-process calculations date back to

Seeger P A, Fowler W A and Clayton D D 1965 Nucleosynthesis of heavy elements by neutron capture *Astrophys. J. Suppl.* **11** 121–66

A recent complilation of solar abundances was published by

Palme H and Beer H 1993 Abundances of the elements in the solar system *Landolt–Börnstein, New Series, Group VI, Astronomy and Astrophysics* subvol 3a (Berlin: Springer) pp 196–221

An understanding of r-, s- and p-processing from general astrophysical principles with many references to the experimental and observational work which constrained the concepts of nuclear processing was given by

Meyer B S 1994 The r-, s-, and p-processes in nucleosynthesis *Annu. Rev. Astron. Astrophys.* **32** 153–90

The canonical s-process with detailed analytical solutions and discussion of branchings was given by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ward R A, Newman M J and Clayton D D 1976 s-Process studies: branching and time scale *Astrophys. J. Suppl.* **31** 33–59

The formulation of a pulsed s-process was suggested by

Ulrich R K 1973 The s-process in stars *Proc. Conf. on Explosive Nucleosynthesis* ed D N Schramm and W D Arnett (Austin, TX: University of Texas Press) pp 139–67

An s-process analysis using the analytical solutions of Ward *et al* (1976) was performed by

Beer H 1991 Capture cross section measurements of krypton and xenon isotopes and the fundamental parameters of the s-process *Astrophys. J.* **375** 823–32

The measurement of the crucial s-process bottle-neck isotopes, an update of the empirical r-process distribution and the representation of the canonical s-process to treat reprocessing were given by

Beer H, Corvi F and Mutti P 1997 Neutron capture of the bottle-neck isotopes ^{138}Ba and ^{208}Pb, s-process studies, and the r-process abundance distribution *Astrophys. J.* **474** 843–61

The most recent stellar s-process model for low-mass AGB stars is found in

Gallino R, Arlandini C, Busso M, Lugaro M, Travaglio C, Straniero O, Chieffi A and Limongi M 1998 Evolution and nucleosynthesis in low-mass asymptotic giant branch stars II. Neutron capture and the s-process *Astrophys. J.* **497** 388–403

The present understanding of the canonical r-process based on measured waiting point half-lives requiring a minimum of three components is found in

Kratz K L, Bitouzet J P, Thielemann F K, Möller P and Pfeiffer B 1993 Isotopic r-process abundances and nuclear structure far from stability: implications for the r-process mechanism *Astrophys. J.* **403** 216–38

The most recent stellar r-process model is given in

Woosley S E, Wilson J R, Mathews G J, Hoffman R D and Meyer B S 1994 The r-process and neutrino-heated supernova ejecta *Astrophys. J.* **433** 229–46

The concept of photodisintegration p-process was elaborated by

Woosley S E and Howard W M 1978 The p-process in supernovae *Astrophys. J. Suppl.* **36** 285–304

and applied with an enlarged network, with the effect of the background particles, especially neutrons, also being taken into account, by

Rayet M, Prantzos N and Arnould M 1990 The p-process revisited *Astron. Astrophys.* **227** 271–81

These parametrized extended network calculations were applied to a realistic stellar model by

Prantzos N, Hashimoto M, Rayet M and Arnould M 1990 The p-process in SN 1987A *Astron. Astrophys.* **238** 455–61

The most recent detailed review of the origin of the p-process nuclei was given by

Lambert D L 1992 The p-nuclei: abundances and origins *Astron. Astrophys. Rev.* **3** 201–56

Other references cited in the text:

Anders E and Grevesse N 1989 Abundances of the elements: Meteoritic and solar *Geochim. Cosmochim. Acta* **53** 197–214

Arlandini C, Käppeler F, Wisshak K, Gallino R, Lugaro M, Busso M, and Straniero O 1999 Neutron capture in low mass asymptotic giant branch stars: Cross sections and abundance signatures *Astrophys. J.* **525** 886–900

Bao Z Y, Beer H, Käppeler F, Voss F, Wisshak K, and Rauscher T 2000 Neutron cross sections for nucleosynthesis studies *Atomic Data and Nuclear Data Tables* **76** 70–154

Beer H, Voss F, and Winters R R 1992 On the calculation of Maxwellian-averaged capture cross sections *Astrophys. J. Suppl.* **80** 403–424

Costa V, Rayet M, Zappalà R A and Arnould M 2000 The synthesis of the light Mo and Ru isotopes: How now, no need for an exotic solution? *Astron. Astrophys.* **358** L67–70

Guber K H, Spencer R R, and Koehler P E 1997 New ^{142,144}Nd (n,γ) cross sections and the s-process origin of the Nd anomalies in presolar meteoritc silicon carbide grains *Phys. Rev. Lett.* **78** 2704–7

Käppeler F, Beer H and Wisshak K 1989 s-process nucleosynthesis—nuclear physics and the classical model *Rep. Prog. Phys.* **52** 945–1013

Koehler P E, Spencer R R, Guber K H, Winters R R, Raman S, Harvey J A, Hill N W, Blackmon J C, Bardayan D W, Larson D C, Lewis T A, Pierce D E and Smith M S 1998 *Phys. Rev.* **57** R1558–61

Macklin R L 1970 Were the lightest stable isotopes produced by photodissociation? *Astrophys. J.* **162** 353–5

Schatz H, Aprahamian A, Barnard V, Bildsten L, Cumming A, Ouellette M, Rauscher T, Thielemann F-K and Wiescher M 2001 End point of the rp process on accreting neutron stars *Phys. Rev. Lett.* **86** 3471–4

Sneden C, Cowan J J, Ivans I I, Fuller G M, Burles S, Beers T C and Lawler J E 2000 Evidence of multiple r-process sites in the early galaxy: New observations of CS 22892-052 *Astrophys. J. Lett.* **533** L139–42

Takahashi K and Yokoi K 1987 Beta-decay rates of highly ionized heavy atoms in stellar interiors *Atomic Data and Nuclear Data Tables* **36** 375–409

Hermann Beer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuclear Properties

Nuclei consist of NEUTRONS n and PROTONS p, which belong to the family of BARYONS (a combination of three quarks (q_1, q_2, q_3)). These two combinations with the lowest masses contain only up and down QUARKS u and d. We have n = (udd) and p = (uud) with electric charge 0 and +1 and masses of 939.570 MeV/c^2 and 938.265 MeV/c^2. They are fermions, i.e. particles with intrinsic half-integer spin 1/2. The binding energy of a nucleus of mass $M(Z, N)$ with Z protons and N neutrons is given by the corresponding mass difference: $B(Z, N) = Zm_{\rm p} + Nm_{\rm n} - M(Z, N)$. The mass excess of a nucleus with a total number of nucleons $A = Z + N$ is defined with the aid of the atomic mass unit $m_{\rm u} = M(^{12}C)/12$, i.e. the average mass of nucleons inside the carbon 12 nucleus, via $M_{\rm ex}(Z, N) = M(Z, N) - Am_{\rm u}$. Here we use the notation $^A Z$ with mass number A, where Z is usually replaced by the element symbol representing that charge or proton number.

The binding energy per nucleon B/A describes how tightly bound a nucleus is. Experimental binding energies range from 0 for neutrons and protons over about 7 MeV per nucleon for ^{4}He, reaching a maximum of 8.7 MeV per nucleon for Fe and then declining to about 7.5 MeV per nucleon towards the heaviest nuclei such as Th and U. These variations of nuclear binding are responsible for the energy generation in stars during a sequence of burning stages from hydrogen burning to silicon burning, transforming in total H to Fe-group nuclei. Variations in nuclear binding are also reflected in the abundances of nuclei produced in astrophysical environments.

The, nevertheless, relatively constant binding per nucleon is related to an approximately constant density ρ_0 in the central part of the nucleus with a typically small surface diffuseness a of 0.55 fm = 5.5×10^{-14} cm. This permits one to express the nuclear mass M ($\approx Am_{\rm u}$) in terms of the mass of a constant-density sphere and explains the A dependence of the nuclear radius $R = (3/4\pi)^{1/3}(m_{\rm u}/\rho_0)^{1/3}A^{1/3} = r_0 A^{1/3}$, where $r_0 = 1.14$ fm is a constant for all nuclei if ρ_0 ($\approx$0.16 nucleons fm^{-3} = 2.7×10^{14} g cm^{-3}) does not depend on A. Such a constant density can be explained by a nucleon–nucleon interaction which drops off quickly with increasing distance and has a repulsive term for short distances. The strong repulsion ensures that the density cannot increase to infinite values, whilst the fast drop-off for large distances causes vanishing interaction with distant neighbors. Thus, only immediate neighbors are important for the interaction.

Interactions are generally explained by the exchange of (virtual) particles. The strong interaction between nucleons is due to the exchange of MESONS, particles consisting of a quark–antiquark pair ($q_1\bar{q}_2$). The lowest mass mesons—consisting of a combinations of up (u) and down (d) quarks—are π-mesons with an approximate mass of 140 MeV. If we assume that these mesons move with the speed of light c, they can at most move a distance $r_{\rm ex} = c\,\Delta t \approx 1.4$ fm when Δt is the time for which the violation of energy conservation (140 MeV) is permitted to create a particle–antiparticle pair according to the UNCERTAINTY PRINCIPLE $\Delta E\ \Delta t = \hbar$. Thus, as this concerns the lightest mesons with the largest $r_{\rm ex}$, the nuclear force has to drop off very quickly beyond that distance.

The above description is related to the dominant bulk feature of nuclear matter (either infinite in neutron stars or finite in nuclei), but it needs a substantial refinement responsible for nuclear shapes and shell structure, and the resulting effects on decay probabilities and nuclear reactions. This will be the focus of the following sections.

Mass models

The bulk part of nuclear binding is equivalent to replacing the (microscopic) many-body Schrödinger equation by independent particle models in an average potential and assuming for this average potential a simple square well potential of depth $v_0 \approx 40$ MeV. This leads to Fermi energies of degenerate nucleons (the highest populated state at temperature $T = 0$ K) of

$$E_{\rm F} = \frac{h^2}{2m}\left(\frac{3}{8\pi}\frac{N}{V}\right)^{2/3}.$$

As N/V, the nucleon density, is roughly equal for all nuclei, $E_{\rm F}$ is also a constant for all nuclei with $E_{\rm F} \approx 29$ MeV. The average energy $\bar{E}$ of a nucleon in a nucleus is $\bar{E}=\frac{3}{5}E_{\rm F}$. In order to unbind an ensemble of nucleons with average energy $\bar{E} = \frac{3}{5}E_{\rm F}$, a total energy of $N(v_0-\bar{E}) = N(v_0-\frac{3}{5}E_{\rm F})$ is needed. With v_0 and $E_{\rm F}$ being roughly constant for all nuclei, the total binding energy is just proportional to the number of nucleons. Thus, for a nucleus consisting out of A nucleons, the total binding is $B = A(v_0 - \bar{E})$. This is the dominant (volume) term of the binding energy $B_{\rm V} = a_{\rm v}A$.

This leading term has to be corrected for the errors introduced by the simplifying approximations. Nucleons which reside at the surface have fewer neighbors and are therefore less bound. This reduction in binding is proportional to the surface area of the nucleus $4\pi R^2 \propto A^{2/3}$, i.e. $B_{\rm S} = -a_{\rm s}A^{2/3}$. The electrostatic repulsion of protons is measured by the Coulomb self-energy (of a uniform charge distribution) $E_{\rm Coul} = \frac{3}{5}\,Q^2/R$ with $Q = Ze$ and $R = r_0 A^{1/3}$, leading to $B_{\rm C} = -a_{\rm c}Z^2A^{-1/3}$. The derivation of the bulk term $B_V = a_{\rm v}A$ relied on an equal density of neutrons N/V and protons Z/V. Correcting for $N \neq Z$ causes the symmetry energy correction $B_{\rm sym} = -a_{\rm sym}(N - Z)^2/A = -a_{\rm sym}(A - 2Z)^2/A$.

Neutrons and protons are both fermions with intrinsic spin 1/2. Pairs can couple to a total spin of 0 or 1. The 0 state corresponds to a lower energy level and contributes to the binding energy. The binding energies of nuclei fluctuate by approximately $B_{\rm pair} = \pm 12/A^{1/2}$ MeV around the mean defined by nuclei with an odd number of nucleons, with Z, N even-even nuclei being more and odd-odd nuclei being less bound. The total binding energy is the sum of all these terms with typical values of $a_{\rm V} = 16$ MeV, $a_{\rm S} = 18.5$ MeV, $a_{\rm C} = 0.72$ MeV and $a_{\rm sym} = 23.4$ MeV and we can express the nuclear mass via

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

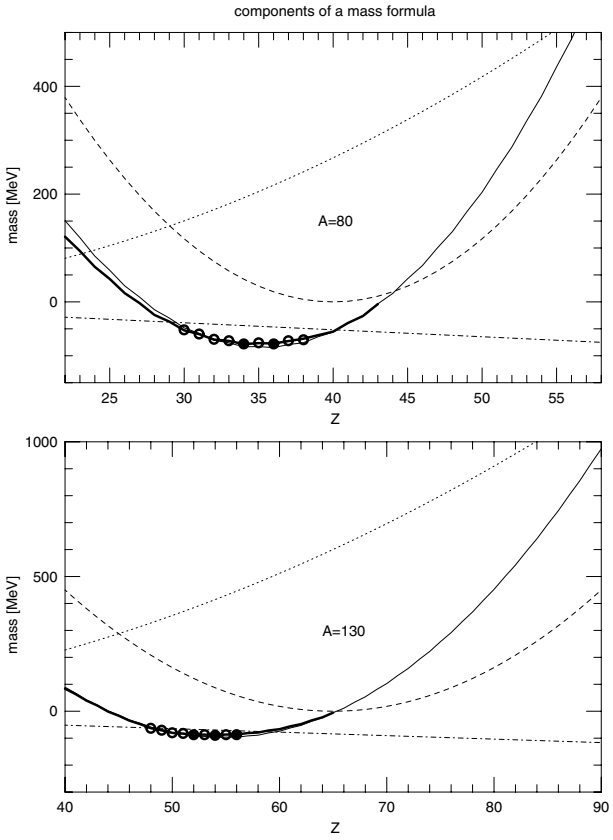


Figure 1. Several components in a liquid drop or Bethe–Weizsäcker mass formula: the dotted curves indicate the Coulomb term, the dashed curves the asymmetry term and the dash-dotted curves the Z-dependent term $Z(m_p - m_n)$. A is taken as a constant in either plot. The solid curves give the atomic mass excess. The thick solid curves result from a more elaborate modern mass model, the circles indicate experimental mass measurements, the solid circles stand for stable nuclei. The larger Coulomb term for heavier nuclei leads to a larger shift of the valley of stability from the location $Z = A/2$.

$M(Z, A) = Zm_{\rm p} + (A-Z)m_{\rm n} - B(Z, A)$, obtaining the well-known Bethe–Weizsäcker (or liquid drop) mass formula with pairing corrections (see figure 1).

In reality the single-particle levels of a nuclear potential are not as evenly distributed as for a Fermi gas in a cubic box, used for this derivation. For a realistic, spherically symmetric nuclear potential which also makes use of a 'spin–orbit' force proportional to $\vec{l}\vec{s}$, where l stands for the orbital angular momentum and s for the intrinsic spin of a nucleon, the levels are unevenly bunched in excitation energy and large spaces can be interpreted as shell closures. Mass formulas which include such shell structure add an empirical shell correction term $S(Z, A)$

$$M(Z, A) = Zm_{\rm p} + (A - Z)m_{\rm n} - a_V A + a_{\rm s}A^{2/3} + a_{\rm c}Z^2A^{-1/3} + a_{\rm sym}\frac{(A-2Z)^2}{A} + S(Z, A) - B_{\rm pair}(Z, A).$$

Shell closures (magic numbers at Z, $N = 2$, 8, 20, 28, 50, 82, 114 or 126) are very important for understanding nuclear properties. It is well known that 'magic' nuclei which are tightly bound are spherical in general, while nuclei in between closed shells try to maximize their binding via deformation, which e.g. rearranges the charge distribution and leads to smaller Coulomb repulsion. Nuclear mass models of the 'microscopic–macroscopic' type which describe the macroscopic part in terms of such a nuclear droplet and the shell corrections in a microscopic way, including also deformation effects, can quite accurately predict nuclear masses within less than 1 MeV and are at present mostly used in astrophysical applications dealing with unstable nuclei where masses are not known experimentally.

An accurate prediction of the shell corrections is essential. Figure 2 shows predicted shell corrections from the finite range droplet model (FRDM) for stable and unstable nuclei. Only unstable nuclei which do not decay by proton or neutron emission are considered. A fully microscopic many-body treatment is also possible by now within the 'shell model' for nuclei up to around Fe. Microscopic 'mean-field' theories allow the heaviest nuclei to be treated as well. The present uncertainties of the latter, due to uncertainties of nuclear forces, are still of the order of a few MeV. One of the major remaining open questions is whether shell closures (magic numbers) and the size of shell gaps change with the distance from the valley of stability.

Decay modes

A necessary condition for the decay of a nucleus is that the transition is energetically possible. This is the case when the total mass of the (final) end products of a decay or reaction is smaller than the sum of the original (initial) masses. In other words, the reaction Q value has to be positive:

$$Q = \sum_{\rm initial} M_i - \sum_{\rm final} M_i > 0.$$

The well known nuclear decay modes are α decay, $\beta^{\pm}$ decay, electron capture and fission.

(i) α decay	$(Z, A) \rightarrow (Z-2, A-4) + {}^4{\rm He}$
(ii) fission	$(Z, A) \rightarrow 2(\approx Z/2, \approx A/2)$
(iii) β^- decay	$(Z, A) \rightarrow (Z+1, A) + {\rm e}^- + \bar{\nu}_{\rm e}$
(iv) β^+ decay	$(Z, A) \rightarrow (Z-1, A) + {\rm e}^+ + \nu_{\rm e}$
(v) electron capture	$(Z, A) + {\rm e}^- \rightarrow (Z-1, A) + \nu_{\rm e}$.

Ground state fission is not fully symmetric in its end products, as indicated above. Larger energies are released if at least one of the fission products is a nucleus near a closed shell. At higher excitation energies the shell effects vanish and symmetric fission occurs. Figure 3 displays in grey scale predicted β^- decay half-lives, based on the FRDM mass model and its microscopic features of excited states in the daughter nucleus. Electron captures occur in atoms by capturing one of the electrons of the low-energy orbitals. For an ionized atom, i.e. a nucleus, electron capture can only occur as a result of continuum capture

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

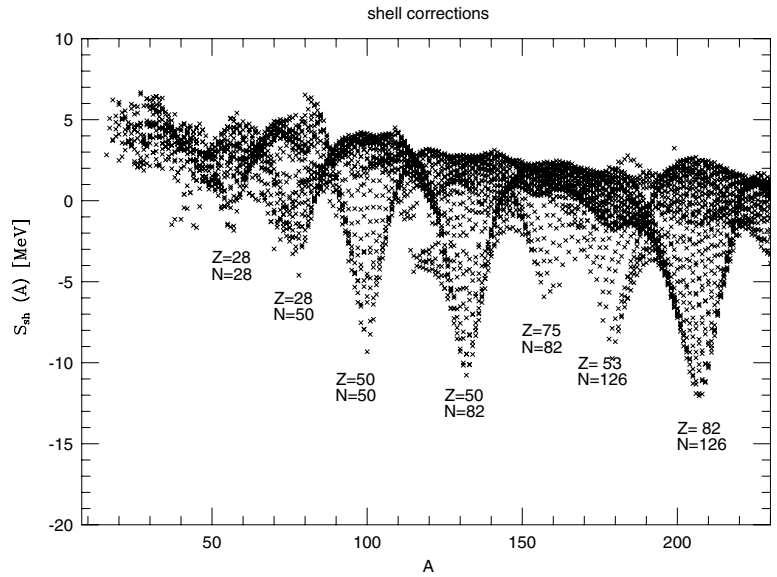


Figure 2. Shell corrections of a modern mass model. They are strongest at closed shells and can be larger than 10 MeV. The two subminima between $A = 150$ and 190 are the onset of stronger effects expected at the closed proton shells $Z = 82$ and 50, which can, however, not be realized between the proton and neutron drip-lines (see next section).

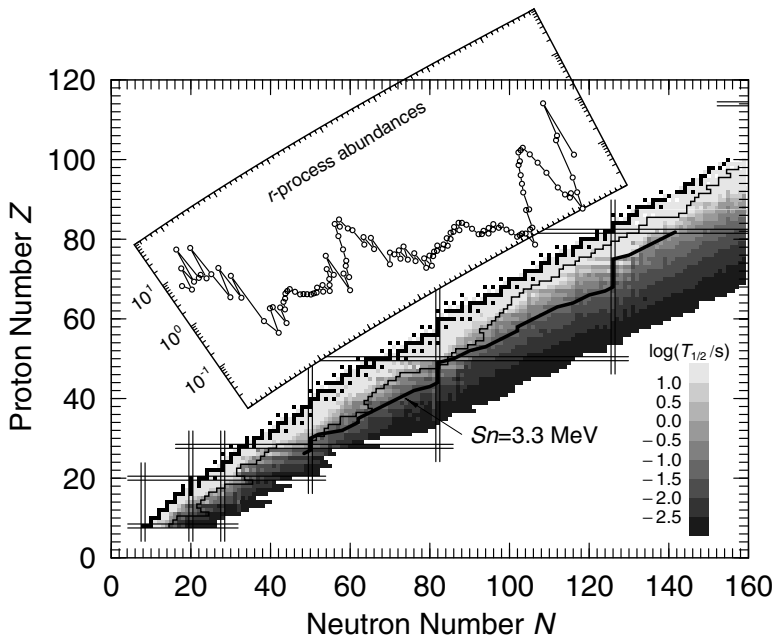


Figure 3. Some features of nuclei in the (N, Z) chart of isotopes. Stable nuclei are indicated by black filled squares. The thin solid curve represents the present limit of experimentally known nuclear masses. The magic numbers are shown as double bars. The thick solid curve is the contour line of constant $S_n = 3.3$ MeV. It relates nuclear properties to astrophysical abundances of the so-called r-process due to rapid neutron capture. It can be recognized that the abundances are proportional to the β^- decay half-lives [indicated by grey shades in $\log_{10}(\tau_{1/2})$] along S_n contour lines.

from a background electron gas. In case of a negative Q value the missing energy has to be provided by the kinetic energy of the ELECTRON.

These decay modes are only of interest between two boundaries in the nuclear chart related to the onset of (much faster) particle emission, i.e. decay of nuclei via neutron and proton emission $(Z, A) \rightarrow (Z, A-1) + \mathrm{n}$ or $(Z, A) \rightarrow (Z-1, A-1) + \mathrm{p}$. The Q values of these decays

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are related to the neutron and proton separation energies via $Q_n = -S_n$ and $Q_p = -S_p$:

$$S_n(Z, A) = M(Z, A-1) + m_n - M(Z, A)$$

$$S_p(Z, A) = M(Z-1, A-1) + m_p - M(Z, A).$$

Nuclei with positive S_n or S_p are stable against neutron or proton emission. The connecting lines of S_p or $S_n = 0$ in the (N, Z) chart are called proton and neutron drip-lines. Figure 3 shows the predicted contour line for $S_n = 3.3$ MeV. The kinks are due to magic numbers. One of the currently open questions is whether a quenching of shell effects decreases these features towards the neutron drip-line.

Nuclear reactions

In a nuclear reaction a projectile will penetrate a target nucleus in order to form a new compound nucleus. This penetration can be described by solutions to the Schrödinger equation for a given potential via transmission coefficients T into the nucleus. Bombarding a nucleus, described by a simple box potential, i.e. $V = 0$ outside (1) and $V = -V_0$ inside (2) the nucleus, with a free neutron of energy E, momentum $\hbar k$, $k = [2\mu(E - V)]^{1/2}/\hbar$ and a reduced mass μ leads to

$$T = \frac{4k_1k_2}{|k_1 + k_2|^2}.$$

As $k_2 > k_1$, because of the large negative potential $-V_0$ inside the nucleus, $T \approx 4k_1/k_2$. Protons experience a Coulomb repulsion potential outside the nucleus, which can be larger than the projectile energy E. Expressed in the quantum mechanical WKB approximation for the Coulomb potential of a nucleus with charge Z_i and a projectile of charge Z_j, $V(x) = Z_iZ_je^2/x$, this results in

$$T = \mathrm{e}^{-2\pi\eta} \qquad \eta = \left(\frac{\mu}{2E}\right)^{1/2} \frac{Z_iZ_je^2}{\hbar}$$

where η is called the Sommerfeld parameter. This transmission coefficient decreases for larger charges and increases with increasing energy.

A quantitative measure for the probability of reactions between a target nucleus and incoming projectiles is the reaction cross section, defined as

$$\sigma = \frac{\text{number of reactions per target s}^{-1}}{\text{flux of incoming projectiles}}.$$

Because the flux of incoming particles is measured in cm^{-2} s^{-1}, σ has units cm^2. All particles passing through a circular area of that size perpendicular to the projectile trajectory will undergo a reaction. Introductory QUANTUM MECHANICS relates the cross section to the sum of transmission coefficients over all possible angular momenta l

$$\sigma = \frac{\pi}{k^2} \sum_{l=0}^{\infty} (2l + 1)T_l$$

if we neglect the fact that projectile and target have an intrinsic spin. Here l denotes the angular momentum and T_l the related transmission coefficient for incoming projectiles of momentum $\hbar k$.

This is the cross section for a successful production of the compound nucleus; however not in its ground state. With a reaction Q value, the projectile with center of mass energy E creates a nucleus at the excitation energy $Q + E$. This excited state can again (because it is not stable) decay into other states: electromagnetic transitions into lower states of the compound nucleus, i.e. γ decay, or decay via particle emission into another nucleus i.e. $i + j \rightarrow \gamma + m$ $(i(j, \gamma)m)$ or $i + j \rightarrow o + m$ $(i(j, o)m)$. In most cases of low-energy projectiles γ transitions to the ground state will dominate. In addition, s-waves ($l = 0$) dominate at low energies and the capture cross section is approximated by

$$\sigma \approx \frac{\pi}{k^2} T_{l=0}.$$

The $l = 0$ transmission coefficients are exactly the ones of the previously described one-dimensional approach, i.e. central collisions where no angular momentum is involved. From the approximate behavior of neutron and charged-particle transmission coefficients, the energy dependence (at low energies E or velocities v) for neutron and charged-particle capture reactions follows.

(i) For neutrons with $T_{n,0} \approx 4k_1/k_2 \propto E^{1/2}$,

$$\sigma = \frac{\pi}{k^2} E^{1/2} \propto \frac{1}{E^{1/2}} \propto \frac{1}{v}.$$

(ii) For charged-particle captures with $T_{c,0} = \mathrm{e}^{-2\pi\eta}$,

$$\sigma = \frac{\pi}{k^2} \mathrm{e}^{-2\pi\eta} = \frac{\hbar^2\pi}{2\mu E} \mathrm{e}^{-2\pi\eta} \propto \frac{1}{E} \mathrm{e}^{-2\pi\eta}.$$

These are key features of neutron-induced and charged-particle-induced reaction cross sections which are essential for astrophysical applications. They are, in addition, mass dependent and modulated by shell effects due to the density of available excited states.

In general, target and projectile are not spinless particles. If a reaction of the type $i(j, o)m$ takes place, the compound nucleus formation cross section has also to be multiplied by the probability of that specific outgoing channel o, $P_o = T_o / \sum_n T_n$, where n sums over all possible decay channels. In a generalized treatment the Hauser–Feshbach or statistical model formula results. With more realistic potentials than the ones discussed before, a correct inclusion of isospin conservation and mixing, an adequate treatment of γ transitions in nuclei, a level density prescription required for transitions to higher energies where individual levels are hard to resolve and with width fluctuation corrections which include correlations between incoming and outgoing channels, a quite reasonable agreement can be found with experimental cross section determinations, provided that a high density of excited states (resonances) is found in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the compound nucleus at the appropriate bombarding energy. Reactions for light nuclei, which are dominated by individual resonances, have been discussed elsewhere (see NUCLEAR REACTION RATES).

Bibliography

Aboussir Y, Pearson J M, Dutta A K and Tondeur F 1995 Nuclear mass formula via an approximation to the Hatree-Fock method *At. Data Nucl. Data Tables* **61** 127

Martinez-Pinedo G, Langanke K and Dean D J 2000 Electron capture rates and β^- decay of pf-shell nuclei *Astrophys. J. Suppl.* **126** 493

Möller P, Nix J R, Myers W D and Swiatecki W J 1995 *At. Data Nucl. Data Tables* **59** 185

Möller P, Nix J R and Kratz K-L 1997 Nuclear properties for astrophysical and radioactive-ion-beam applications *At. Data Nucl. Data Tables* **66** 131

Oda T, Hino M, Muto K, Takahara M and Sato K 1994 Rate tables for the weak processes of sd-shell nuclei in stellar matter *At. Data Nucl. Data Tables* **56** 231

Rauscher T and Thielemann F-K 2000 Astrophysical reaction rates from statistical model calculations *At. Data Nucl. Data Tables* **75** at press

F K Thielemann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuclear Reactions Rates

Investigations during the last 60 yr have shown that we are connected with distance space and time not only by our imagination but also through a common cosmic heritage: the chemical ELEMENTS that make up our bodies. These elements were created by nuclear reactions in the hot interiors of remote and long-vanished stars over many billions of years. Their nuclear fuels finally spent, these giant stars met death in cataclysmic explosions, scattering far and wide the atoms of heavy elements synthesized deep within their cores. Eventually this material, as well as material lost by smaller stars during red-giant stages, collected into clouds of gas in interstellar space; these, in turn, slowly collapsed giving birth to new generations of stars, thus leading to a cyclic evolution that is still going on. In this scenario, the Sun and its complement of planets were formed nearly 5 billion years ago. Drawing on the material gathered from the debris of its stellar ancestors, the planet Earth provided the conditions that eventually made life possible. Like every object in the solar system, each living creature on Earth incorporates atoms from distant corners of our galaxy. Every molecule in our bodies contains matter that once was subjected to the tremendous temperatures and pressures at the center of a star. This is where the iron in our blood cells originated, the oxygen we breathe, the carbon and nitrogen in our tissues and the calcium in our bones. All were formed predominantly in nuclear fusion reactions of smaller atoms in the interior of stars. The present picture is that all elements from carbon to uranium have been produced entirely within stars during their fiery lifetimes and explosive deaths. A few of the lightest elements were formed before the stars even existed, during the birth of the universe itself. In addition, a few of the most reactive light elements appear to have been synthesized in intergalactic space by COSMIC RAYS. Thus, theories of nucleosynthesis have identified the most important sites of element formation and also the diverse nuclear processes involved in their production. The detailed understanding of our cosmic heritage combines astrophysics and nuclear physics and forms what is called nuclear astrophysics. There are excellent theories of NUCLEOSYNTHESIS, but theories, in all of science, are only guides to understanding. They must be constantly confronted with experimental and observational data to ensure that they evolve in a meaningful direction. In turn, nuclear reactions are at the heart of nuclear astrophysics: they influence sensitively the nucleosynthesis of the elements in the earliest stages of the universe and in all the objects formed thereafter and control the associated energy generation, neutrino luminosity and evolution of stars. A good knowledge of the rates of these reactions is essential for an understanding of this broad picture.

Nuclear astrophysical rates

In a stellar plasma the constituent nuclei have kinetic energy (see below) and move in random directions. Occasionally they collide with other nuclei, whereby they experience one of two types of processes. In one type they are scattered on each other (like billiard balls): this elastic scattering process may be written symbolically as $1 + 2 \rightarrow 1+2$ or $x+A \rightarrow x+A$ (the latter is often written as A(x,x)A). The symbol x (or 1) represents the projectile and A (or 2) the target nucleus which together constitute the entrance channel. In the elastic scattering process the same types of nuclei emerge from the collision and constitute the exit channel. In the other type of process two different nuclei emerge from the collision, $1 + 2 \rightarrow 3 + 4$ or $x + A \rightarrow y + B$ (or A(x,y)B), where y (or 3) and B (or 4) represent the emerging nuclei which together constitute here the exit channel. This type of process is called a nuclear reaction, transmuting nuclei of the entrance channel into different nuclei of the exit channel and representing thus the key to the understanding of elemental nucleosynthesis. If the nuclei in the entrance channel have nuclear masses m_1 and m_2 and those in the exit channel have nuclear masses m_3 and m_4, energy conservation together with EINSTEIN's relation $E = mc^2$ leads to the nuclear reaction Q value, defined as $Q = (m_1 + m_2 - m_3 - m_4)c^2$. If the Q value is positive, there is a net production of energy in the reaction for each event: mass is converted to kinetic energy. Clearly, such reactions are most important for the energy production in stars. The Q value can be obtained from nuclear mass tables, which are the results of decades of painstaking work. Of equal importance is the intrinsic nuclear probability that a given reaction will take place. This probability, expressed as an energy-dependent cross section $\sigma(E)$ (see below), determines how many reactions occur per unit of time and unit volume: hence, together with the Q value, $\sigma(E)$ provides important information on nuclear energy production in stars.

We know that when one is shooting at a target the probability of hitting it increases as its area increases. By analogy, one may associate with each nucleus a geometrical area, which is directly related to the probability of a projectile interacting with the nucleus. This area is referred to as the cross section for the reaction. Classically, this cross section σ is equal to the combined geometrical area of the projectile and the target nucleus; if the projectile and the target nucleus have radii R_p and R_t, respectively, the cross section may be written as $\sigma = \pi(R_p + R_t)^2$ and does not depend on energy. In reality, the nuclear reactions are governed by the laws of quantum mechanics, whereby the geometrical cross section must be replaced by the energy-dependent quantity $\sigma = \pi(\lambda/2\pi)^2 \propto E^{-1}$ (E is the energy of the nuclei), where λ represents the de Broglie wavelength reflecting the wave aspect of quantum mechanical processes. In some cases the Coulomb and centrifugal barriers related to nuclear charge and angular momentum strongly inhibit the penetration of one nucleus into another. This barrier penetration leads to a steep energy dependence of the cross section $\sigma(E)$. Other energy-dependent effects, such as resonances and their mutual interference effects, also play important roles and the energy dependence of $\sigma(E)$ can be quite complex. In addition to the effects just discussed, the cross section for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a particular nuclear reaction depends on the nature of the force involved (i.e. strong, electromagnetic or weak force), with strengths that vary by many orders of magnitude. It is the challenge to the experimentalist to make precise $\sigma(E)$ measurements over a wide range of energies, as our fragmented knowledge of nuclear physics prevents us from predicting $\sigma(E)$ on purely theoretical grounds.

In hot stellar matter the energies of the moving nuclei can be described by a Maxwell–Boltzmann distribution, $\Phi(E) \propto E \exp(-E/kT)$, where T is the local temperature and k the Boltzmann constant. At low energies, $E \ll kT$, the function $\Phi(E)$ increases linearly with energy E. $\Phi(E)$ reaches a maximum value at $E = kT$, and at higher energies, $E \gg kT$, $\Phi(E)$ decreases exponentially (figure 1). Folding the cross section with this energy (or velocity) distribution leads to the nuclear reaction rate per pair of nuclei:

$$\langle \sigma v \rangle = (8/\pi\mu)^{1/2}(kT)^{-3/2} \int_0^\infty \sigma(E) \exp(-E/kT)\, \mathrm{d}E \quad (1)$$

where v is the relative velocity of the pair of nuclei, E is the center-of-mass energy and $\mu = m_1 m_2/(m_1 + m_2)$ is the reduced mass of the entrance channel. Multiplication of $\langle \sigma v \rangle$ by the number densities of the interacting species in the entrance channel leads to the total reaction rate per unit of time and volume for a star of given chemical composition. In order to cover the different evolution phases of stars, i.e. from main-sequence stars ($T \approx 10^7$ K) to supernovae ($T \approx 10^9$ K), one must know the rates over a wide range of temperatures, which in turn requires the availability of $\sigma(E)$ data over a wide range of energies.

For the important class of charged-particle-induced reactions, there is a repulsive Coulomb barrier in the entrance channel of height $E_c = Z_1 Z_2 e^2/r$, where Z_1 and Z_2 are the integral nuclear charges of the interacting particles, e is the unit of electric charge and r is the nuclear interaction radius. Because of the tunneling effect through this barrier, $\sigma(E)$ drops nearly exponentially with decreasing energy (figure 2):

$$\sigma(E) = S(E) E^{-1} \exp(-2\pi\eta) \quad (2)$$

where $\eta = 2\pi Z_1 Z_2 e^2/hv$ is the Sommerfield parameter (h is the Planck constant). The function $S(E)$, defined by this equation, contains all the strictly nuclear effects, and is usually referred to as the nuclear or astrophysical S factor. If equation (2) is inserted in equation (1), one obtains

$$\langle \sigma v \rangle = (8/\pi\mu)^{1/2}(kT)^{-3/2} \times \int_0^\infty S(E) \exp(-E/kT - b/E^{1/2})\, \mathrm{d}E \quad (3)$$

with $b = E_G^{1/2} = 2(2\mu)^{1/2}\pi^2 e^2 Z_1 Z_2/h$. Since for nonresonant reactions $S(E)$ varies slowly with energy, the steep energy dependence of the integrand in equation (3) is governed primarily by the exponential term. The penetration through the Coulomb barrier gives rise to the term $\exp(-b/E^{1/2})$, which becomes very small at low energies. The other exponential term, $\exp(-E/kT)$, which vanishes at high energy, is a measure of the number of particles available in the high-energy tail of the Maxwell–Boltzmann distribution. The product of the two terms leads to a peak of the integrand near the energy E_0, which is usually much larger than kT. The peak is frequently referred to as the Gamow peak (figure 1). For a constant $S(E)$ value over the energy region of the peak, one finds $E_0 = (bkT/2)^{2/3}$; this is the effective mean energy for a given reaction at a given temperature. Approximating the peak by a Gaussian function, one finds an effective width $\Delta = 4(E_0 kT)^{1/2}/3^{1/2}$. Thus, nuclear burning takes place predominantly over the energy window $E_0 \pm \Delta/2$, the stellar thermal energy range for which information on $\sigma(E)$ must be obtained.

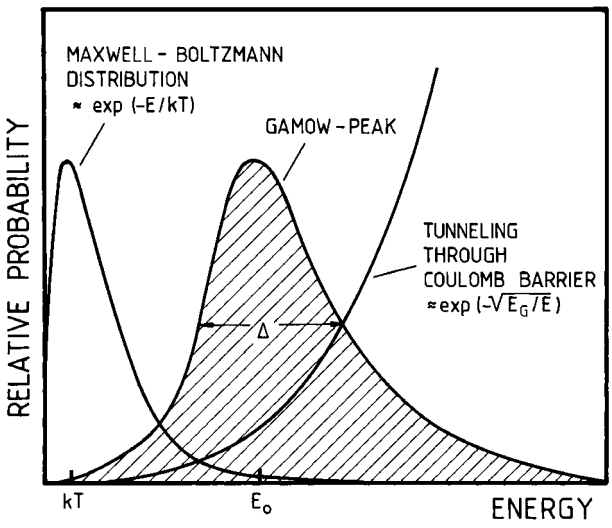


Figure 1. The dominant energy-dependent functions are shown for nuclear reactions between charged particles. While both the energy distribution function (Maxwell–Boltzmann) and the quantum mechanical tunneling function through the Coulomb barrier are small for the overlap region, the convolution of the two functions results in a peak (the Gamow peak) near the energy E_0, giving a sufficiently high probability to allow a significant number of reactions to occur. The energy of the Gamow peak is generally much larger than kT.

Because of the steep drop of $\sigma(E)$ at subcoulomb energies ($E < E_c$), it becomes increasingly difficult to measure $\sigma(E)$ as E is lowered. Although experimental techniques have improved significantly over the years, extending $\sigma(E)$ measurements to lower energies (with a low-energy limit at E_L, corresponding to a reaction yield in a detector of roughly 1 event per hour), it has not been possible to measure $\sigma(E)$ at stellar thermal energies, as $E_0/E_c \approx 1/100$ for hydrogen-burning reactions (i.e. pp chain and CNO cycles) in main-sequence stars such as our Sun. Instead, the measured energy dependence of $\sigma(E)$ at higher energies ($E > E_L$) must be extrapolated to stellar energies, using the $S(E)$ factor defined in equation (2). Such an 'extrapolation into the unknown' can lead

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

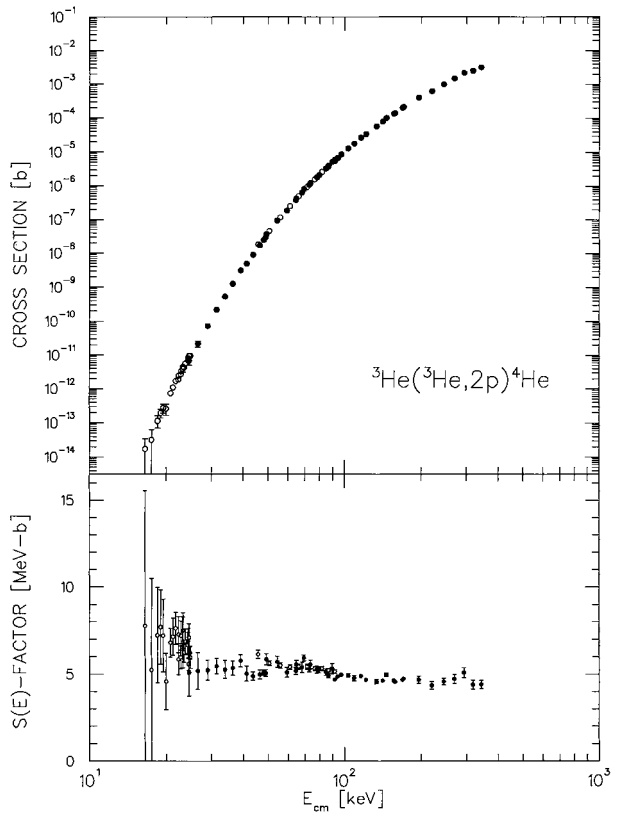


Figure 2. Energy dependence of the cross section and the astrophysical $S(E)$ factor for the ^{3}He(^{3}He, 2p)^{4}He reaction. The Gamow peak is at $E_0 = 21$ keV for solar conditions (central temperature $T = 15 \times 10^6$ K).

to considerable uncertainty. At energies lower than E_L there might be a change of reaction mechanism, or of the centrifugal barrier, or there might be a contribution of narrow or subthreshold resonances to $\sigma(E)$ at stellar energies. The danger of such extrapolations was strikingly demonstrated in the case of ^{2}H(d, γ)^{4}He, for example, where new low-energy data increased the extrapolated values by a factor 1000, mainly as a result of change in the centrifugal barrier (d- to s-waves). For another example, the observed solar neutrino fluxes in the existing neutrino detectors are not consistent with the current 'standard' picture of the microscopic processes in the Sun. A possible solution for this solar neutrino problem may be found in one or more of the areas of neutrino physics (e.g. the recent experimental evidence for neutrino oscillations at Super-Kamiokande), solar physics (models) or nuclear physics. In view of the important and fundamental conclusions regarding nonstandard physics, which may be deduced from the results of present and future solar neutrino observations, it is of utmost importance to place the predictions on a solid basis. The nuclear physics component of this problem involves the cross sections of the hydrogen-burning reactions extrapolated to the relevant solar energies. Yet another example is the reactions ^{12}C(α, γ)^{16}O, which takes places in the helium-burning phase of red-giant stars, at $E_0 \approx 0.30$ MeV. It has been found theoretically that this reaction influences sensitively not only the nucleosynthesis of the elements between C and Fe but also the evolution of stars, the dynamics of supernovae and the kind of remnant (neutron star or black hole) of a supernova explosion (see SUPERNOVA REMNANTS). At present, useful but still inadequate $\sigma(E)$ data are available only above $E_L = 1.3$ MeV. New experimental approaches are needed to reduce the uncertainties of the extrapolations.

The LUNA project

Low-energy studies to thermonuclear reactions in a laboratory at the Earth's surface are hampered predominantly by background effects of cosmic rays in the detectors, leading typically to more than 10 background events per hour in common detectors. Conventional passive or active shielding around the detectors can only partially reduce the problem of cosmic ray background. The best solution is to install an accelerator facility in a laboratory deep underground. As a pilot project, a 50 kV accelerator facility has been installed in the Laboratori Nazionali del Gran Sasso (LNGS), where the flux of cosmic-ray muons is reduced by a factor of 10^6 compared with the flux at the surface. This unique project, called LUNA (Laboratory for Underground Nuclear Astrophysics), was designed primarily for a renewed study of ^{3}He(^{3}He,2p)^{4}He at low energies, aiming to reach the solar Gamow peak at $E_0 \pm \Delta/2 = 21 \pm 5$ keV. This goal has been reached with a detected reaction yield of about 1 event per month at the lowest energy, $E = 16$ keV, with $\sigma \approx 20$ fb of 2×10^{-38} cm^2. Thus, the cross section of an important reaction of the pp chain has been directly measured for the first time at solar thermal energies (figure 2); in principle, extrapolation is no longer needed in this reaction. The work demonstrated the research potential of LUNA and that all of the experimental requirements in such low-rate, time-consuming experiments can be fulfilled. Installing larger facilities at LNGS in the near future will open the possibility of improving our knowledge of other key reactions by shifting their low-energy limit E_L significantly closer to E_0 or even to within the Gamow peak.

In the extrapolation of $\sigma(E)$ using equation (2), it is assumed that the Coulomb potential of the target nucleus and projectile is that resulting from bare nuclei. However, for nuclear reactions studied in the laboratory, the target nuclei and the projectiles are usually in the form of neutral atoms or molecules and ions, respectively. The electron clouds surrounding the interacting nuclides act as a screening potential: the projectile effectively sees a reduced Coulomb barrier. This in turn leads to a higher cross section, $\sigma_s(E)$, than would be the case for bare nuclei, $\sigma_b(E)$. There is an enhancement factor, $f_{lab}(E) = \sigma_s(E)/\sigma_b(E) \approx \exp(\pi\eta U_e/E)$, where U_e is the electron-screening potential energy (e.g. $U_e \approx Z_1 Z_2 e^2/R_a$, with R_a an atomic radius). Note that $f_{lab}(E)$ increases exponentially with decreasing energy. For ratios $E/U_e >$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1000, shielding effects are negligible, and laboratory experiments can be regarded as essentially measuring $\sigma_b(E)$. However, for $E/U_e < 100$, shielding effects become important for understanding and extrapolating low-energy data. Relatively small enhancements from electron screening at $E/U_e \approx 100$ can cause significant errors in the extrapolation of cross sections to lower energies, if the curve of the cross section is forced to follow the trend of the enhanced cross sections, without correction for the screening. Notice that, for a stellar plasma, the value $\sigma_b(E)$ must be known because the screening in the plasma will be quite different from that in the laboratory nuclear-reaction studies, i.e. $\sigma_{plasma}(E) = f_{plasma}(E)\sigma_b(E)$, and $f_{plasma}(E)$ must be explicitly included for each situation. A good understanding of electron-screening effects in the laboratory is needed to arrive at reliable $\sigma_b(E)$ data at low energies. Experimental studies of reactions involving light nuclides have shown the expected exponential enhancement of the cross section at low energies (figure 2). However, the observed enhancements were in all cases significantly larger than could be accounted for from available atomic-physics models. This situation is disturbing because, if the effects of electron screening are not understood under laboratory conditions, they are most likely to be not fully understood in a stellar plasma. A solution to the laboratory puzzle might be found in one (or all) of the following areas: the assumed energy-loss predictions from stopping-power codes at low energies, the assumed nuclear-reaction models at energies far below the Coulomb barrier, and the assumed atomic-physics models. All of these areas require additional experimental and theoretical efforts. An improved understanding of laboratory electron screening may also help eventually to improve the corresponding understanding of electron screening in stellar plasmas. It is in the nature of astrophysics that many of the processes and most of the objects one tries to understand are physically inaccessible. Thus, it is important that those aspects that can be studied in the laboratory be rather well understood. The electron-screening project addresses one such aspect.

Other rapidly developing experimental techniques

In explosive scenarios with temperatures of the order of 10^9 K (e.g. early universe, novae and supernovae), where E_0 approaches E_c, nuclear burning times can be greatly reduced, even down to seconds. If the half-life of a radioactive nuclide is longer than or of the same order as the burning time, that nuclide will be involved in the nuclear burning and thus will contribute to the composition of the final nuclear ashes. Since H and He are the most abundant species in these scenarios, the burning consists predominantly of (p,γ) and (α, γ) capture reactions involving these radioactive nuclides. If the half-life of a nuclide is longer than a day or so, it may be made into a radioactive target: e.g. ^{7}Be(p,γ)^{8}B with $T_{1/2}(^7\text{Be}) =$ 53 days. However, in a great majority of interesting cases, the half-lives are too short to make targets: e.g. ^{13}N(p,γ)^{14}O with $T_{1/2}(^{13}\text{N}) = 10$ min. In this case, a possible method is to create the radioactive nuclides in an accelerator, separate them, accelerate them in a second accelerator and allow the radioactive ion beam (RIB) to interact with an H or He target (inverted kinematics). All of this must be achieved in a time shorter than the decay lifetime of the radioactive nuclei. A growing number of laboratories have already produced RIBs of astrophysical interest or are in the stage of technical development of such beams. For the study of these capture reactions, the detection of the capture γ-rays or of the residual recoil nuclei (via their radioactive decay signals) leads in general to efficiencies far below 100%. In addition, the radioactivity of the RIB seriously hampers the detection methods. Only p(^{13}N,γ)^{14}O has been successfully studied so far via γ-ray spectroscopy, because of an unusually large cross section arising from a strong, broad resonance. For all other capture reactions studied so far, only upper limits for $\sigma(E)$ have been derived at the relevant energies $E_0 \pm \Delta/2$ owing to more typical (i.e. small) cross sections. Thus, future experiments require a significantly improved detection efficiency, with a detector that is less sensitive to background. Since the recoil nuclei are kinematically confined to a forward cone of a few degrees, a recoil mass separator (RMS) can be used to filter the recoils from the RIB and to focus all recoils onto a suitable detector for particle identification. Of course, in most RMSs one must make a choice of the charge state of the recoils; however, there is usually a charge state representing about 50% of the recoils produced and thus the capture reactions can be studied with an approximately 50% detection efficiency. Such an RMS is in preparation at Bochum, called ERNA (European Recoil Separator for Nuclear Astrophysics), for the study of ^{4}He(^{12}C,γ)^{16}O and other key reactions. Although this reaction does not involve RIBs, a similar RMS should be appropriate for work with RIBs.

It should be pointed out that the capture reactions ^{7}Be(p,γ)^{8}B and ^{13}N(p,γ)^{14}O have also been studied using the alternative method of Coulomb dissociation (Coulomb excitation into the continuum). However, there remain uncertainties about the interpretation and achievable precision of the method. It is to be hoped that these uncertainties can be resolved and that this technique will eventually prove to be useful for the indirect studies of capture reactions involving both stable and radioactive nuclides. Unique applications for this technique would be the study of capture reactions involving three particles in the entrance channel (e.g. ^{4}He(αn,γ)^{9}Be) of two radioactive nuclides in the entrance channel (e.g. a neutron plus another radioactive nucleus).

Neutron-induced reactions, mainly (n,γ), play an important role in the early universe, as well as in stars where they are involved in the nucleosynthesis of the transiron elements via the s- and r-processes. Because of the short lifetime of the neutrons ($T_{1/2} = 10$ min), they must be produced by nuclear reactions within the star (e.g. via ^{13}C(α,n)^{16}O) and are quickly thermalized through elastic scattering. Because of the absence of a Coulomb barrier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the entrance channel, the most probable energy is near $E_0 = kT \approx 30$ keV for $T = 3 \times 10^8$ K. Usually $\sigma(E)$ is very large and increases with decreasing E. Thus, $\sigma(E)$ measurements of the s-process can in principle be carried out directly in the relevant stellar energy region. Since the r-process involves radioactive nuclei near the neutron-drip line, the relevant $\sigma(E)$ measurements require RIB techniques (see above).

As pointed out earlier, a few of the most reactive light elements, i.e. Li, Be and B, appear to have been synthesized in intergalactic space through spallation processes induced by high-energy cosmic rays. This possibility is currently under quantitative study using space vehicles (measurement of cosmic-ray composition and spectra) and in laboratories (measurement of high-energy spallation cross sections).

Summary

Impressive progress has been achieved in the knowledge of nuclear reaction rates. However, there remains much critical work to be done in the future to arrive at reliable data for many key reactions and processes. New techniques continue to be developed for this purpose, for nuclear reaction rates involving both stable and radioactive nuclei, and research in nuclear astrophysics will remain an exciting pursuit for many years to come.

Bibliography

Burbridge E M, Burbridge G R, Fowler W A and Hoyle F 1957 *Rev. Mod. Phys.* **29** 547

Fiorentini G, Kavanagh R W and Rolfs C 1995 *Z. Phys.* A **350** 289

Fowler W A 1984 *Rev. Mod. Phys.* **56** 149

Rolfs C and Rodney W S 1988 *Cauldrons in the Cosmos* (Chicago, IL: University of Chicago Press)

U Greife and C Rolfs

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nucleosynthesis

Nucleosynthesis is the science related to all astrophysical processes which are responsible for the abundances of the elements and their isotopes in the universe. The astrophysical sites are the Big Bang and stellar objects, either during their stable (hydrostatic) evolution and wind ejection or during explosions like novae and supernovae, or possibly other events, where binary stellar systems are involved. The understanding of each of these environments requires in general hydro(fluid/gas)dynamics, thermodynamics and energy transport. Nuclear abundances and energy generation are determined by thermonuclear reactions.

We give a brief review of hydrostatic burning phases in stellar evolution and the specific burning features of explosive stellar events. This is followed by a discussion of the basic equations for reaction networks which govern composition changes and the required nuclear physics input for thermonuclear reactions.

The working of nucleosynthesis is presented in a survey of events which act as nucleosynthesis sources and an analysis of their role in the puzzle to explain the evolution of the elemental and isotopic composition in galaxies and especially the mixture found in the solar system (see figure 1). Different timescales due to the mass dependence of the endpoints of stellar evolution (type II supernova explosions—SNe II versus planetary nebulae) or single versus binary stellar systems (the latter being responsible for novae, type Ia supernovae—SNe Ia, or x-ray bursts) are the key to understanding galactic evolution. We also speculate on the role of (i) very massive stars not undergoing the SN II explosions but rather causing 'hypernovae' after the formation of a central black hole via core collapse, and (ii) binary neutron star or neutron star–black hole pairs.

Nuclear burning

Nuclear burning can in general be classified into two categories: (1) hydrostatic burning stages on timescales dictated by stellar energy loss and (2) explosive burning due to hydrodynamics of the specific event. Hydrostatic burning stages are characterized by temperature thresholds, permitting thermal Maxwell–Boltzmann distributions of (charged) particles (nuclei) to penetrate increasingly larger Coulomb barriers of electrostatic repulsion. We can describe the burning stages in terms of an abbreviated nomenclature for reactions with target i, projectile j, emitted particle k, and main reaction product l [$i + j \rightarrow k + l$ or $i(j,k)l$]. Projectiles are either neutrons n, protons p, α-particles (^{4}He) or nuclei, emitted particles include those listed above, photons γ or in $\beta^{\pm}$-decays positrons/electrons $e^{\pm}$ and neutrinos/antineutrinos $\upsilon/\overline{\upsilon}$. H-burning converts ^{1}H into ^{4}He via pp-chains or the CNO-cycles. Here we use the notation $^A Z$ where the mass

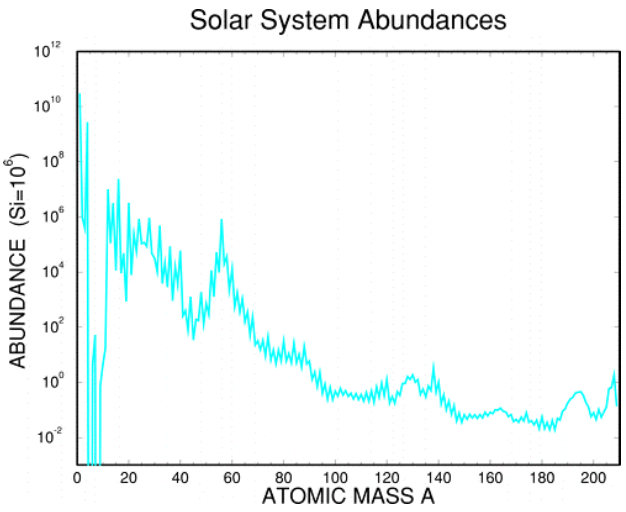


Figure 1. Abundances by number as found in the solar system from meteorite samples and solar spectra. Abundances of isotopes of different elements with the same mass number A are added. The units are arbitrarily scaled to an abundance of Si of the order 10^6.

number A is the sum of neutrons and protons $N + Z$ and Z is usually replaced by the element symbol representing that charge or proton number (see NUCLEAR PROPERTIES). The simplest pp-chain is initiated by ^{1}H(p,$e^+\upsilon$)^{2}H(p,γ)^{3}He and completed by ^{3}He(^{3}He,2p)^{4}He. The dominant CNO-cycle chain ^{12}C(p,γ)^{13}N($e^+\upsilon$)^{13}C(p,γ)^{14}N(p,γ)^{15}O($e^+\upsilon$) ^{15}N(p,α)^{12}C is controlled by the slowest reaction ^{14}N(p,γ)^{15}O. Further burning stages are characterized by their major reactions, which are in He-burning ^{4}He(2α,γ)^{12}C (triple-alpha) and ^{12}C(α,γ)^{16}O, in C-burning ^{12}C(^{12}C, α)^{20}Ne and in O-burning ^{16}O(^{16}O, α)^{28}Si. The alternative to fusion is photodisintegration which starts to play a role at sufficiently high temperatures T when $30kT \approx Q$ (the Q-value or energy release of the inverse capture reaction). This ensures the existence of photons with energies $>Q$ in the Planck distribution and leads to Ne-burning (^{20}Ne(γ,α)^{16}O, ^{20}Ne(α,γ)^{24}Mg) at $T > 1.5 \times 10^9$ K (preceding O-burning) due to a small Q-value of ≈4 MeV and Si-burning at temperatures in excess of 3×10^9 K (initiated like Ne-burning by photodisintegrations). The latter ends with nuclear reactions in a complete chemical equilibrium (nuclear statistical equilibrium, NSE) and an abundance distribution around Fe. These temperatures permit photodisintegrations with typical Q-values of 8–10 MeV as well as the penetration of Coulomb barriers in capture reaction.

In such an NSE the abundance of each nucleus Y_i, with mass number A_i, charge Z_i, and neutron number $N_i = A_I - Z_i$ is only dependent on temperature T, density ρ, and its nuclear binding energy B_i, while fulfilling mass conservation $\sum_i A_i Y_i = 1$ and charge conservation $\sum_i Z_i Y_i = Y_e$ (equal total abundances of protons and electrons). Y_e is changed by weak interactions (beta-decays and electron captures) on longer timescales. Quasi-equilibrium can

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

occur when different nuclear mass regions are locally in equilibrium with the background of free neutrons, protons and alphas but offset from their NSE values.

During core and shell He-burning specific α-induced reactions can liberate neutrons which are responsible for the slow neutron capture process (s-process). It leads to the build-up of elements up to Pb and Bi through a series of neutron captures and β^--decays, starting on existing heavy nuclei around Fe.

Many of the hydrostatic burning processes occur also under explosive conditions at higher temperatures and on shorter timescales (see figure 2), when often the β-decay half-lives are longer than the explosive timescales. This requires in general the additional knowledge of nuclear reactions for unstable nuclei. The fuels for explosive nucleosynthesis consist mainly of $N = Z$ nuclei like ^{12}C, ^{16}O, ^{20}Ne, ^{24}Mg, or ^{28}Si (the ashes of prior hydrostatic burning), resulting in heavier nuclei, again with $N \approx Z$. At high densities also substantial electron captures on nuclei $e^- + {}^AZ \rightarrow {}^A(Z-1) + \upsilon$ can occur due to energetic, degenerate electrons with high Fermi energies (equivalent to a β^+-decay which converts a proton into a neutron inside a nucleus).

Explosive Si-burning differs strongly from its hydrostatic counterpart and can be divided into three different regimes: (i) incomplete Si-burning and complete Si-burning with either (ii) a normal (high density, low entropy) or (iii) an alpha-rich (low density, high entropy) freeze-out of charged-particle reactions during cooling from NSE. At high temperatures or during a 'normal' freeze-out, the abundances are in a full NSE. An alpha-rich freeze-out is caused by the inability of the triple-alpha reaction ^{4}He(2α,γ)^{12}C, and the ^{4}He(αn,γ)^{9}Be reaction to keep light nuclei like n, p and ^{4}He, and nuclei beyond $A = 12$ in an NSE during declining temperatures, when the densities are small. This causes a large alpha abundance after freeze-out. This effect, most pronounced for SNe II, is a function of entropy.

r-process nucleosynthesis (rapid neutron capture) relates to subsets of explosive Si-burning, either with low or high entropies, experiencing a normal or alpha-rich freeze-out. The requirement of a neutron/seed ratio of 10 to 150 after charged particle freeze-out translates for a normal freeze-out into $Y_e = 0.12$–0.3. For a moderate $Y_e > 0.40$ an extremely alpha-rich freeze-out is needed.

Nuclear reaction rates and networks

In this section we want to outline the essential features of thermonuclear reaction rates and nuclear reaction networks (see also NUCLEAR REACTION RATES). The nuclear cross section for a reaction between target j and projectile k is defined by

$$\sigma = \frac{\text{number of reactions target}^{-1}\,\text{s}^{-1}}{\text{flux of incoming projectiles}} = \frac{r/n_j}{n_k v}.$$

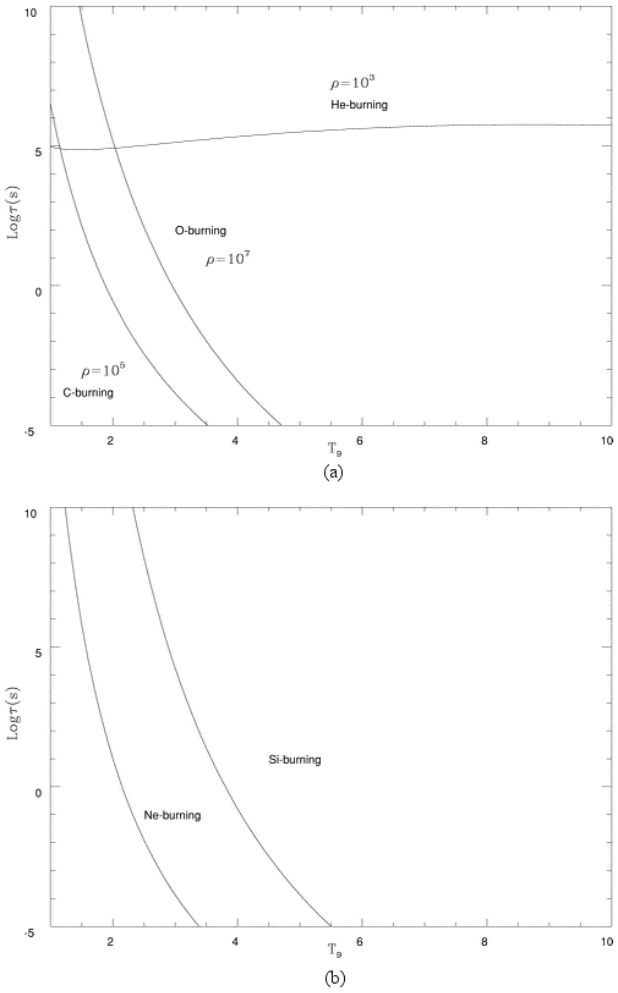


Figure 2. Burning timescales in seconds for fuel destruction of He-, C-, and O-burning (top) and Ne- and Si-burning (bottom) as a function of temperature. Density-dependent timescales are labeled with a chosen typical density (in g cm^{-3}). They scale with $1/\rho$ for C- and O-burning and $1/\rho^2$ for He-burning. Ne- and Si-burning, initiated by photodisintegrations, are not density-dependent. The almost constant He-burning timescale beyond $T_9 = T/10^9$ K = 1 permits efficient destruction on explosive timescales only for high densities.

The second equality holds for the case that the relative velocity between targets with the number density n_j and projectiles with number density n_k is constant and has the value v. Then r, the number of reactions per cm^3 and s, can be expressed as $r = \sigma v n_j n_k$. More generally, when targets and projectiles follow specific thermal momentum distributions, r is given by

$$r_{j,k} = \int \sigma(|\boldsymbol{v}_j - \boldsymbol{v}_k|)|\boldsymbol{v}_j - \boldsymbol{v}_k| \mathrm{d}n_j\, \mathrm{d}n_k.$$

The evaluation of this integral depends on the types of particles (fermions, bosons) and distributions which are involved. For nuclei j and k in an astrophysical plasma,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

obeying a Maxwell–Boltzmann distribution, this simplifies to $r_{j,k} = <\sigma v> n_j n_k$. The thermonuclear reaction rates have the form

$$\langle \sigma v \rangle_{j,k} = \left(\frac{8}{\mu \pi}\right)^{1/2} (kT)^{-3/2} \int_0^\infty E \sigma(E) e^{-E/kT} dE.$$

Here μ and E denote the reduced mass and center of mass energy of the target–projectile system. When particle k is a photon, the relative velocity is always c and quantities in the integral are not dependent on dn_j. This simplifies to $r_j = \lambda_{j,\gamma} n_j$. $\lambda_{j,\gamma}(T)$ results from an integration over a Planck distribution for photons of temperature T.

A similar procedure is used for electron captures by nuclei. Because the electron is about 2000 times less massive than a nucleon, the velocity of the nucleus j is negligible in the center of mass system in comparison to the electron velocity ($|\boldsymbol{v}_j - \boldsymbol{v}_e| \approx |\boldsymbol{v}_e|$). The electron capture cross section has to be integrated over a Fermi distribution of electrons. The electron capture rates are a function of T and $n_e = Y_e \rho N_A$, the electron number density, where ρ denotes the matter density and N_A Avogadro's number. In a neutral, completely ionized plasma, the electron abundance is equal to the total proton abundance $Y_e = \sum_i Z_i Y_i$ and $r_j = \lambda_{j,e}(T, \rho Y_e) n_j$.

This treatment can be generalized for the capture of positrons, which are in a thermal equilibrium with photons, electrons, and nuclei. At high densities ($\rho > 10^{12}$ g cm^{-3}) the size of the neutrino scattering cross section on nuclei and electrons ensures that enough scattering events occur to thermalize a neutrino distribution. Then also the inverse process to electron capture (neutrino capture) can occur and be expressed in a similar way to photon and electron captures, integrating now over the neutrino distribution. Finally, for normal decays with half-life τ<inf>1/2</inf>, we obtain an expression with the decay rate $\lambda_j = \ln 2/\tau<inf>1/2</inf>$ independent of temperature or density.

The time derivative of the number densities of each of the species in an astrophysical plasma is governed by the different expressions for r, as discussed above for the different reaction mechanisms

$$\left(\frac{\partial n_i}{\partial t}\right)_{\rho=\text{const}} = \sum_j N_j^i r_j + \sum_{j,k} N_{j,k}^i r_{j,k} + \sum_{j,k,l} N_{j,k,l}^i r_{j,k,l}.$$

The reactions listed on the right-hand side of the equation belong to the three categories of reactions: (1) decays, photodisintegrations, electron and positron captures and neutrino induced reactions ($r_j = \lambda_j n_j$), (2) two-particle reactions ($r_{j,k} = <j,k> n_j n_k$), and (3) three-particle reactions ($r_{j,k,l} = <j,k,l> n_j n_k n_l$) like the triple-alpha process, which can be interpreted as successive captures with an intermediate unstable target. The capital Ns indicate how many nuclei of type i are destroyed (-) or produced (+) in the corresponding reaction. In order to exclude changes in the number densities $\dot{n}_i$ which are only due to expansion or contraction of the gas, the nuclear abundances $Y_i = n_i/(\rho N_A)$ were introduced. With A_i being the atomic weight, $A_i Y_i$ represents the mass fraction of this nucleus. In terms of abundances Y_i the previous equation translates into a set of differential equations

$$\dot{Y}_i = \sum_j N_j^i \lambda_j Y_j + \sum_{j,k} N_{j,k}^i \rho N_A \langle j,k \rangle Y_j Y_k + \sum_{j,k,l} N_{j,k,l}^i \rho^2 N_A^2 \langle j,k,l \rangle Y_j Y_k Y_l.$$

This reaction network has in general to be solved numerically. A simplified situation is given in the case of a nuclear statistical equilibrium (NSE) at high temperatures. Instead of solving reaction networks, which would result in the same, but much more time consuming, solution the abundances can be expressed by a chemical equilibrium for the build-up of all nuclei from neutrons and protons via the chemical potentials $\mu(Z,A) = Z\mu_p + (A - Z)\mu_n$. For the typical Boltzmann distributions this leads to

$$Y_i = (\rho N_A)^{A_i - 1} \frac{G_i}{2^{A_i}} A_i^{3/2} \left(\frac{2\pi\hbar^2}{m_u kT}\right)^{3(A_i-1)/2} e^{B_i/kT} Y_p^{Z_i} Y_n^{N_i}$$

for the abundance of nucleus Y_i, with nuclear binding energy B_i and partition function $G_i(T)$, and m_u denoting the atomic mass unit. The required solution for neutron and proton abundances Y_n and Y_p is found from mass conservation $\sum_i A_i Y_i = 1$ and charge conservation $\sum_i Z_i Y_i = Y_e$. In that case only the slow weak interaction processes, which change Y_e as a function of time, have to be followed explicitly.

The total energy generation rate per gram, due to nuclear reactions with abundance changes $\dot{Y}_i$, is expressed in terms of the mass M_i of the participating nuclei

$$\dot{\epsilon} = -\sum_i \dot{Y}_i N_A M_i c^2$$

where c is the speed of light.

As noted above, the important ingredients to nucleosynthesis calculations are decay half-lives, electron and positron capture rates, photodisintegrations, neutrino induced reaction rates, and strong interaction cross sections. Photodisintegration rates can be calculated via detailed balance from the reverse capture rates.

The early universe

The dominant contributions to the cosmic abundances of the light nuclides ^{1}H, ^{2}H, ^{3}He, ^{4}He and ^{7}Li are due to BIG BANG nucleosynthesis and thus relics of a hot and dense state of the early universe which underwent adiabatic expansion. When assuming that all microphysics is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

understood, the final abundances of the light elements are determined by only one parameter, the entropy, expressable in terms of the baryon to photon ratio $\eta = n_B/n_\gamma$, because this initial state possesses also a weak interaction equilibrium, determining Y_e. The Big Bang entropies correspond to an extremely α-rich freeze-out (high entropies = a low density for a given temperature), where essentially no element with $A > 7$ is produced. The build-up of light elements through non-equilibrium reactions is only determined after the reaction $^1H(n,\gamma)^2H$ falls out of equilibrium close to $T_9 \approx 1$, because of a very small Q-value of 2 MeV. Uncertainties, affecting especially the primordial ^{4}He abundance, were related to the number of neutrino species and the neutron lifetime. LEP experiments at CERN proved the existence of three neutrino species in the late 1980s, and the neutron lifetime is determined by experiments with trapped ultracold neutrons. The cross sections of the relevant thermonuclear reactions linking nuclei from ^{1}H to ^{7}Li have been measured at the appropriate energies to an accuracy better than 10%. The major uncertainties are related to the determination of the primordial abundance of ^{2}H, ^{3,4}He and ^{7}Li with astronomical means. They constrain η to $2–5 \times 10^{-10}$.

Low mass stars and the s-process

STELLAR EVOLUTION is an advanced subject with many successes. Major modeling uncertainties seem to be related to the treatment of convection, the consistent inclusion of effects due to stellar rotation, and the treatment of wind losses. Reactions involving stable nuclei are in quite reasonable shape, but with a number of open questions left until now. A few are (a) pp-chain reactions in H-burning, e.g. $^3He(\alpha,\gamma)^7Be$, $^7Be(p,\gamma)^8B$, and $^7Li(p,\alpha)^4He$, which affect sub-solar and solar type stars and are directly connected to the solar neutrino emission, CNO-cycle reactions like $^{14}N(p,\gamma)^{15}O$ and $^{17,18}O(p,\alpha)^{14,15}N$, which directly relate to the surface abundances in intermediate mass stars, NeNa-cycle reactions in intermediate and massive stars, (b) the $^{12}C(\alpha,\gamma)^{16}O$ reaction in He-burning, and (c) fusion reactions of late burning stages, e.g. $^{12}C + {}^{12}C$, which are not well known at low energies.

Low and intermediate mass stars which enrich the interstellar medium via stellar winds and/or planetary nebula ejection before forming central C + O WHITE DWARFS, are contributing important amounts of He and C as well as other H- and He-burning products. Among the He-burning products one also finds about half of the heavy elements and isotopes beyond Fe and up to Pb and Bi, resulting from s-process nucleosynthesis which is due to neutron capture by pre-existing intermediate and heavy nuclei. New data (Van Eck *et al*) support this view, which hitherto had little observational support. In core and shell He-burning neutrons are provided by a side branch of He-burning reactions acting on ^{14}N, the dominant CNO nucleus after H-burning, $^{14}N(\alpha,\gamma)^{18}F(e^+\upsilon)^{18}O$ $(\alpha,\gamma)^{22}Ne(\alpha,n)^{25}Mg$. An alternative stronger neutron source during the period of He-shell flashes in low and intermediate mass stars is the reaction $^{13}C(\alpha,n)^{16}O$, which requires admixture of hydrogen and the production of ^{13}C via $^{12}C(p,\gamma)^{13}N(e^+\upsilon)^{13}C$. The s-process is characterized by neutron captures which are slow in comparison to beta-decays.

Figure 3 shows a section of the chart of nuclei and the s-process path indicated by horizontal arrows for neutron captures with reaction rates $\langle\sigma v\rangle_{n,\gamma}(A) \equiv \langle\sigma v\rangle_A$, connected by β^--decays whenever an unstable nucleus is encountered. The path can be simplified by abundances related only to mass numbers A and the abundances approach a steady flow equilibrium $\dot{Y}_A = 0$ of creation and distruction

$$\dot{Y}_A = \rho N_A \langle\sigma v\rangle_{A-1} Y_n Y_{A-1} - \rho N_A \langle\sigma v\rangle_A Y_n Y_A$$
$$= n_n \langle\sigma v\rangle_{A-1} Y_{A-1} - n_n \langle\sigma v\rangle_A Y_A .$$

This leads to $\langle\sigma v\rangle_A Y_A \approx$ const for neighboring As, with exceptions at closed neutron shells where the cross sections and reaction rates are small, i.e. the neutron capture timescales are too long to attain a steady flow equilibrium. For a large fraction of nuclei the neutron cross sections scale with $E^{-1/2}$ ($l = 0$ s-wave capture), which causes a temperature independent $\langle\sigma v\rangle$. Then the resulting abundance pattern is only dependent on the neutron exposure $\tau = \int n_n \, dt$, and in this classical analysis a very good fit to all s-only isotopes (see figure 3) can be found with a continuous (exponential) superposition of

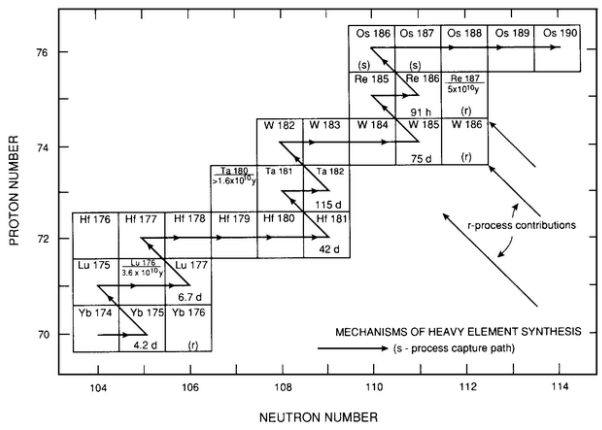


Figure 3. A typical portion of the nuclear chart for heavy elements beyond the 'Fe-group'. The s-process path, which is constrained to the 'valley' of stability, is indicated. Neutron-rich r-only isotopes formed by the r-process are identified as well as the shielded s-only isotopes, which cannot have an r-process contribution. Stable isotopes, more proton-rich than the s-process path, have typically an abundance smaller by two orders of magnitude than s- and r-isotopes. Their origin is due to the p- or γ-process (discussed later).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

exposures τ. This analysis can also provide the r-process abundances (to be discussed later) by subtracting the s-process component from the total solar abundances for those nuclei which have s- and r-contributions (see figure 4). The s-process contribution has abundance peaks at stable nuclei with closed neutron shells (N = 50, 82, 126 for A = 88, 138, and 208) where small cross sections are encountered.

There exist two complications to the simple picture given above: (i) nuclei with very long β-decay half-lives (comparable to the s-process duration time) can cause branchings into two flows, and (ii) cross sections which are not s-wave dominated lead to a temperature dependent $<\sigma v>$. In that case the results are not just a function of τ but can depend on $n_n(t)$ and $T(t)$. Among the most important and difficult examples is the branching occurring at ^{85}Kr, which involves Kr, the Rb and Sr isotopes. The abundance of ^{142}Nd can only be reproduced with realistic stellar conditions due to a temperature dependent $<\sigma v>$. The s-process is recognized to occur in at least two different astrophysical sites, the weak component accounting for most of the s-nuclei below the Kr–Rb–Sr abundance peak, and the main component accounting for heavier nuclei up to Pb. The main s-process component is due to the thermally-pulsing phase of shell He- and H-burning in low mass stars ($M \leq 3M_\odot$) and primarily driven by ^{13}C(α,n)^{16}O rather than ^{22}Ne(α,n)^{25}Mg. Recent research indicates that ^{13}C(α,n)^{16}O burns in the interpulse phases at low temperatures. Core He-burning with the ^{22}Ne source contributes to the weak s-component which is a complex combination of core He-, and shell C-, and Ne-burning in massive stars.

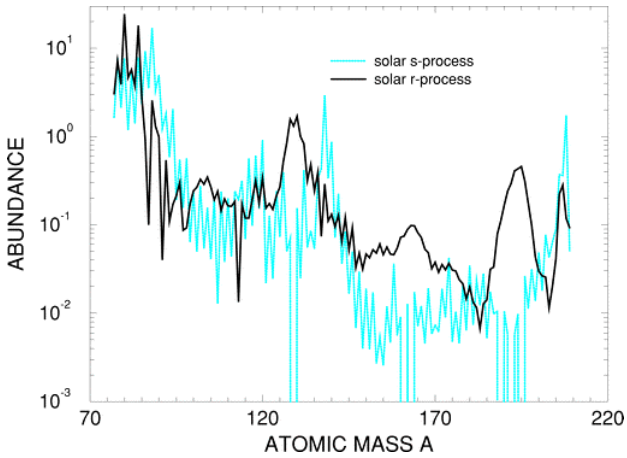


Figure 4. Decompositions of s-process and r-process abundances. Notice the opposite behavior of the two processes with respect to odd–even staggering.

Massive stars and type II supernovae

Stars with main sequence masses $M > 8M_\odot$ follow the hydrostatic burning stages beyond He-burning, develop an onion-like composition structure, and produce a collapsing core after the end of their hydrostatic evolution, which proceeds to nuclear densities (see SUPERNOVAE). The latter is due to the fact that Fe-group nuclei possess the highest binding energy per nucleon and nuclear energy generation ceases to be available. The total energy released in the collapse, 2–3 × 10^{53} erg, equals the gravitational binding energy of a neutron star. Because neutrinos are the particles with the longest mean free path, they are able to carry away that energy in the fastest fashion as seen for SN1987A in the Kamiokande, IMB and Baksan experiments. The apparently most promising mechanism for such type II supernova explosions (SNeII) is based on neutrino heating beyond the hot proto-neutron star via the dominant processes $\upsilon_e + n \rightarrow p + e^-$ and $\overline{\upsilon}_e + p \rightarrow n + e^+$ with a (hopefully) about 1% efficiency in energy deposition. The neutrino heating efficiency depends on the neutrino luminosity, which in turn is affected by neutrino opacities. The explosion via neutrino heating is delayed after core collapse for a timescale of seconds or less. The exact delay time t_{de} and other aspects of the explosion mechanism are still uncertain and depend on Fe-cores from stellar evolution and the supranuclear equation of state and maximum neutron star mass. The observational fact that many core collapse supernovae show polarized light emission is an indication of a nonspherical explosion mechanism.

Nucleosynthesis calculations

The still existing uncertainties of self-consistent models suggest making use of the fact that typical kinetic energies of 10^{51} erg are observed and light curve as well as explosive nucleosynthesis calculations can be performed by introducing a shock of appropriate energy in the pre-collapse stellar model, either by depositing thermal energy or invoking a piston. Induced calculations (lacking self-consistency) utilize the constraint of requiring ejected ^{56}Ni-masses from the innermost explosive Si-burning layers in agreement with supernova light curves, being powered by the decay chain ^{56}Ni–^{56}Co–^{56}Fe (see figure 5). However, it should be clear that, even if ^{56}Ni ejecta from a variety of progenitor masses are known, this is not a simple one parameter problem. Explosion energies can vary as a function of progenitor mass (or implicit parameters like rotation) and the delay time between core collapse and explosion can vary as well.

Figure 5 shows the radial composition of explosively processed mass zones caused by the shock wave which initiates a supernova explosion. The outer ejected layers ($M(r) > 2M_\odot$) are unprocessed by the explosion and contain results of prior H-, He-, C-, and Ne-burning in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stellar evolution. The interior parts of SNe II contain products of explosive Si, O, and Ne burning. In the inner ejecta, which experience incomplete and complete Si-burning, we see an alpha-rich freeze-out (^{4}He). The Y_e originates from the pre-explosive hydrostatic fuel in these layers and can change from 0.4989 to 0.494. This brings huge changes in the Fe-group composition for mass zones below $M(r) = 1.63M_\odot$, making the abundances of ^{58}Ni and ^{56}Ni comparable. All neutron-rich isotopes increase (^{57}Ni, ^{58}Ni, ^{59}Cu, ^{61}Zn, and ^{62}Zn); the even-mass isotopes (^{58}Ni and ^{62}Zn) show the strongest effect. One can also recognize the increase of ^{40}Ca, ^{44}Ti, ^{48}Cr, and ^{52}Fe with an increasing alpha-rich freeze-out, but a reduction of these $N = Z$ nuclei in the inner more neutron-rich layers.

Observational constraints

There is limited direct observational information from individual supernovae with known progenitors, leading to direct O, Ti or Fe(Ni) observations. When including also supernova remnants, there exist observations for element abundances, like C, O, Si, Cl, Ar, Co, and Ni, and isotopic abundances, like ^{56}Ni, ^{57}Ni, ^{44}Ti, and ^{26}Al, from optical, UV, and x-ray spectra, supernova lightcurves, as well as γ-ray lines following nuclear decay. They constrain the products of hydrostatic burning stages like C, O, Ne and Mg as well as the composition close to the mass cut between the central neutron star and the supernova ejecta (^{56}Co, ^{57}Co, ^{44}Ti). The latter are a measure of temperature, entropy, and Y_e and give a consistent picture for neutrino heating delay times of 0.3–0.5 s, explosion energies of the order 10^{51} erg, and a mass cut close to or outside the O-burning shell with minute permitted admixtures of deeper layers so that only matter with $Y_e \geq 0.497$–0.498 is ejected.

One has to investigate the question how the features discussed here might depend on the supernova progenitor mass and possibly other parameters. Figure 5 shows also that ^{55}Co (decaying to ^{55}Mn) and ^{52}Fe (decaying to ^{52}Cr) are products of incomplete Si-burning, while ^{59}Cu (decaying to ^{59}Co) is resulting from alpha-rich freeze-out. As these nuclei are the only (or dominant) stable isotopes of the respective element, this fact could be related to the Mn, Cr, and Co/Fe ratios observed in galactic evolution and give clues to changing mass cut positions as a function of progenitor mass.

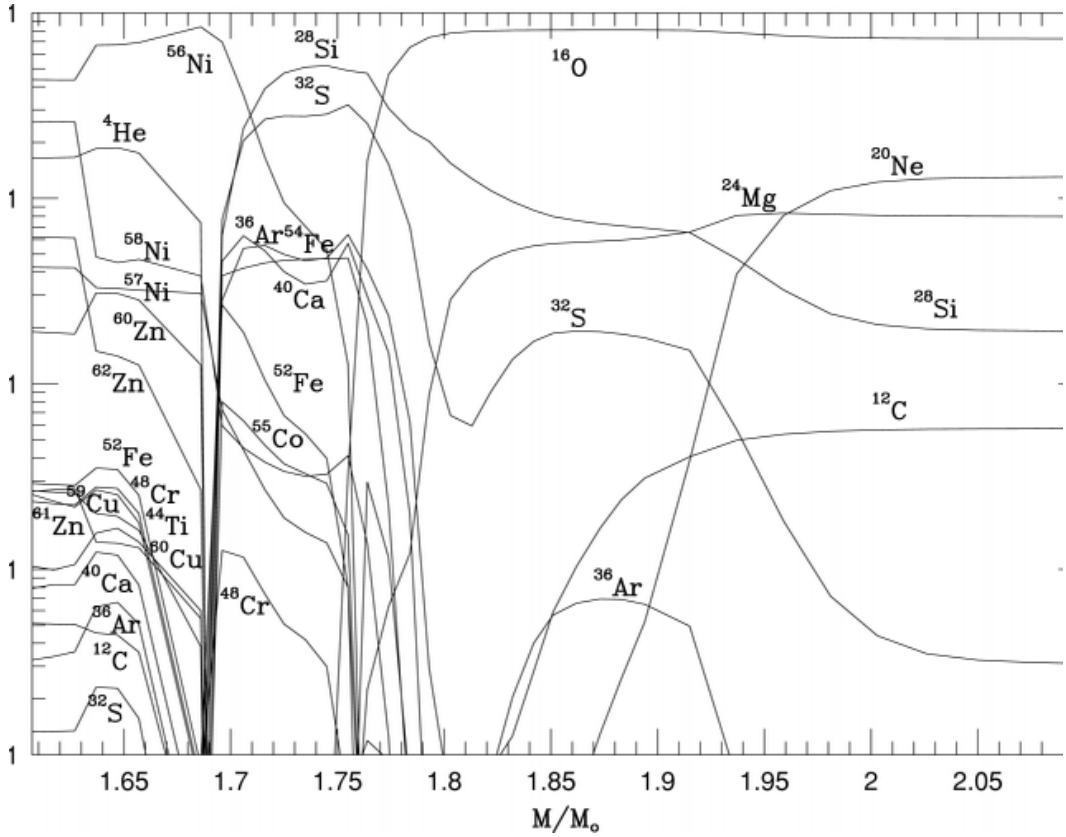


Figure 5. Isotopic composition for a core collapse supernova from a $20M_\odot$ progenitor star with a $6M_\odot$ He-core and a net explosion energy of 10^{51} erg, remaining in kinetic energy of the ejecta. The exact mass cut in $M(r)$ between neutron star and ejecta depends on the details of the delayed explosion mechanism.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finally, the question has to be answered for which progenitor masses a large Fe-core at the end of stellar evolution leads the creation of a black hole instead of a neutron star and a supernova explosion. With SN 1998bw and SN 1997ef, which appear in optical spectra like core collapse supernovae, the transition to 'hypernovae' (failed supernovae which produce black holes and are probably powered by black hole accretion and jet ejection) is encountered. Their nucleosynthesis products have not yet been seriously analysed nor included in chemical evolution studies.

Nucleosynthesis as a witness of uncertainties

The synthesized elements form three different classes which are sensitive to different aspects of the stellar models and supernovae explosion mechanism: (1) stellar evolution treatment, (2) stellar evolution plus the explosion energy, and (3) details of the explosion mechanism which includes aspects of stellar evolution determining the size of the collapsing Fe-core.

(1) The abundances of C, O, Ne, and Mg originate from the unaltered (essentially only hydrostatically processed) C-core and from explosive Ne/C-burning. They are mainly dependent on the structure and zone sizes of the pre-explosion models and thus serve as a test for the treatment of stellar evolution. These zones and therefore the amount of ejected mass varies strongly over the progenitor mass range.

(2) The amount of mass for the elements S, Ar, and Ca, originating from explosive O- and Si-burning, is similar for all massive stars. They do not show the strong progenitor mass dependence of C, O, Ne, and Mg. Si has some contribution from hydrostatic burning and varies by a factor of 2–3. This second set (Si, S, Ar, Ca) tests the progenitor models and the explosion energy, because the amount of explosive burning depends on the structure of the model plus the energy of the shock wave which passes through it.

(3) The amount of Fe-group nuclei ejected (which includes also one of the so-called alpha elements, i.e. Ti) and their relative composition depends directly on the explosion mechanism, connected also to the size of the collapsing Fe-core. Observational checks of individual supernovae are presently still required to test the detailed working of a supernova. Three types of uncertainties are inherent in the Fe-group ejecta, related to (i) the total amount of Fe(group) nuclei ejected and the mass cut between neutron star and ejecta, mostly measured by ^{56}Ni decaying to ^{56}Fe, (ii) the total explosion energy which influences the entropy of the ejecta and with it the degree of alpha-rich freeze-out from explosive Si-burning and the abundances of radioactive ^{44}Ti as well as ^{48}Cr, the latter decaying later to ^{48}Ti and being responsible for elemental Ti, and (iii) finally the neutron richness or Y_e of the ejecta, dependent on stellar structure and the delay time between collapse and explosion. Y_e influences strongly the ratios of isotopes 57/56 in Ni(Co,Fe) and the overall elemental Ni/Fe ratio being dominated by ^{58}Ni and ^{56}Fe.

(4) r-process yields. SNe II have long been expected to be the source of r-process elements. Some recent calculations seemed to be able to reproduce the solar r-process abundances well in the high entropy neutrino driven wind, emitted from the hot proto-neutron star after the SN II explosion. However, we should keep in mind that present-day supernova models have difficulties to reproduce the entropies required for such abundance calculations. In addition, incorrect abundance features of lighter r-process nuclei are encountered. The inclusion of non-standard neutrino properties may perhaps achieve low enough Y_es, for intermediate entropies to correct for such unwanted features. However, recent observations shed also some doubts on the supernova origin. On average SNe II produce Fe to intermediate mass elements in ratios within a factor of 3 of solar. If they were also responsible for the r-process, the same limits should apply. But the observed bulk r-process/Fe ratios vary widely. In CS 22892-052 the r/Fe ratio is 30 times solar!

Type Ia supernovae

There are strong observational and theoretical indications that SNe Ia are thermonuclear explosions of accreting white dwarfs (see BINARY STARS: OVERVIEW). Theoretically, both (i) Chandrasekhar mass white dwarf models and (ii) sub-Chandrasekhar mass models have been considered. These correspond to a single white dwarf scenario, involving accretion of H or He via mass transfer from a binary companion onto the white dwarf. High rates of H-accretion cause high temperatures at the base of the accreted matter and lead to quasi-stable H-burning and subsequent He-burning in shells surrounding the white dwarf, probably related to supersoft x-ray sources. This increases the mass of the white dwarf consisting of C and O which approaches the maximum stable Chandrasekhar mass. In the case of He accretion at low rates, He detonates (supersonic burning front) at the base of the accreted layer before the system reaches the Chandrasekhar mass. A double white dwarf scenario is related to the merging of two C + O binary white dwarfs with a combined mass exceeding the Chandrasekhar limit.

Observational features of SNe Ia in early phases clearly indicate that Chandrasekhar models give a more consistent picture than the sub-Chandrasekhar models of helium detonations.

Ignition and burning front propagation

For Chandrasekhar mass models contraction and carbon ignition in the central region lead to a thermonuclear runway and a complete explosive disruption of the white dwarf. High accretion rates cause a higher central temperature and pressure, favoring lower ignition

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

densities. A flame front then propagates at a subsonic speed as a *deflagration wave* due to heat transport across the front. Here the most uncertain quantity is the flame speed which depends on the development of instabilities of various scales at the flame front. Multi-dimensional hydro simulations of the flame propagation have suggested that a carbon deflagration wave might propagate at a speed v_{def} as slow as a few per cent of the sound speed v_s in the central region of the white dwarf. The nucleosynthesis consequences of such slow flame speeds witness the actual burning front velocities and can thus serve as a constraint. Electron capture on free protons, intermediate mass and Fe-group nuclei affects the central electron fraction Y_e, which determines the composition of the ejecta from such explosions. The amount of electron capture depends on both v_{def} (influencing the time duration of matter at high temperatures, and with it the availability of free protons) and the central density of the white dwarf ρ_{ign} (increasing the electron chemical potential i.e. their Fermi energy).

After an initial deflagration in the central layers, the deflagration can turn into a detonation (supersonic burning front) at lower densities. The transition from a deflagration to a detonation (delayed detonation model) leads to a change in the ratios of incomplete Si-burning, complete Si-burning with normal freeze-out and with alpha-rich freeze-out, the latter being less alpha-rich than in SNe II due to the smaller entropies or higher densities. This also leaves an imprint on the Fe-group composition. Nucleosynthesis constraints can help to find the 'average' SN Ia conditions responsible for their contribution to galactic evolution, i.e. especially the Fe-group composition. SNe Ia contribute essentially no elements lighter than Al, about $\frac{1}{3}$ of the elements from Si to Ca, and the dominant (50–70%) amount of Fe group nuclei (Ti to Ni). In addition, the average Fe-group yields of SNe II differ from those of SNe Ia (e.g. in the elemental Cr/Fe, Mn/Fe ratios). The latter fact, combined with O–Ca/Fe ratios, gives the strongest constraints on the transition density ρ_{tr} from deflagrations to detonations.

Nucleosynthesis consequences

Figure 6 shows the influence of central ignition densities ρ_{ign} 1.37 (C) and 2.12×10^9 g cm^{-3} (W) at the onset of the thermonuclear runaway and slow (S) deflagration speeds of v_{def}/v_s = 0.015 (WS15, CS15), 0.03 (WS30,CS30) or 0.05 (CS50) on the resulting $Y_e(r)$ due to different amounts of electron capture. Y_e values of 0.47–0.485 lead to dominant abundances of ^{54}Fe and ^{58}Ni, values between 0.46 and 0.47 produce dominantly ^{56}Fe, values in the range of 0.45 and below are responsible for ^{58}Fe, ^{54}Cr, ^{50}Ti, ^{64}Ni, and values below 0.43–0.42 are responsible for ^{48}Ca. The intermediate Y_e-values 0.47–0.485 exist in all cases, but the masses encountered which experience these conditions depend on the Y_e-gradient determined by v_{def}. Whether the lower values with $Y_e < 0.45$ are attained depends on the central ignition density ρ_{ign}. Therefore, ^{54}Fe and ^{58}Ni are indicators of v_{def} while ^{58}Fe, ^{54}Cr, ^{50}Ti, ^{64}Ni, and ^{48}Ca are a measure of ρ_{ign}. Conclusions drawn from these results are that (i) v_{def} in the range 1.5–3% of the sound speed is preferred, and (ii) ignition densities $\rho_{ign} < 2 \times 10^9$ g cm^{-3} provide apparently a better agreement with the solar abundances of very neutron-rich species (case C rather than W). The latter interpretation is changed somewhat with improved electron capture rates on pf-shell (Fe-group) nuclei. In either case the ignition densities are in agreement with identifying supersoft x-ray sources as SN Ia progenitor systems. Their high accretion rates (close to $\dot{M} = 10^{-7} M_\odot$) would be in agreement with such ignition densities.

If a deflagration turns into a detonation, the transition density ρ_{tr} affects the total amount of ^{56}Ni, the intermediate mass elements Si–Ca, and the ratios of different explosive Si-burning regimes like alpha-rich freeze-out to incomplete Si-burning as seen in figure 7 (indicated by ^{58}Ni versus ^{54}Fe plateaux). Light curve modeling and spectra show that the typical ^{56}Ni mass should be in the range 0.5–0.7$M_\odot$, which compensates for the well known overproduction of Si–Ca/Fe in SNe II during galactic evolution with the observed Ia/II ratios in order to obtain a solar mix.

Small transition densities favor larger amounts of matter which experience incomplete Si-burning. Observations require an overproduction of Mn (and Cr) in SNe Ia. These elements are mostly made as ^{55}Co and ^{52}Fe (decaying to Mn and Cr), which are favorably

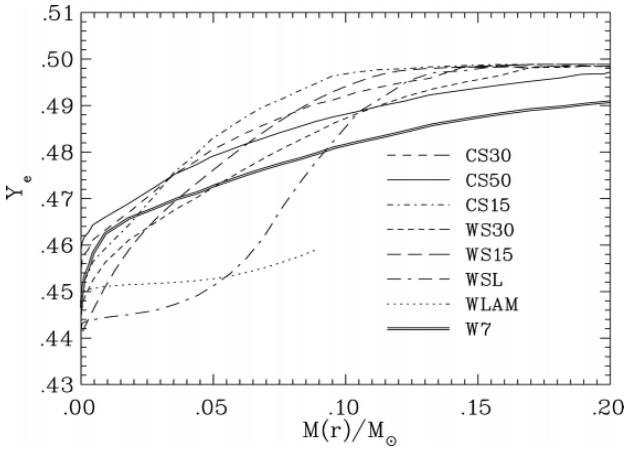


Figure 6. Y_e as a function of radial mass coordinate $M(r)$ after freeze-out of nuclear reactions measures the electron captures on free protons and nuclei. Small burning front velocities lead to steep Y_e-gradients which flatten with increasing velocities (see the series of models CS15, CS30, and CS50 or WS15, WS30, and W7). Lower central ignition densities shift the curves up (C versus W), but the gradient is the same for the same propagation speed. Only when the Y_e from electron captures is smaller than for stable Fe-group nuclei, will subsequent β^--decays reverse this effect (WSL and WLAM).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

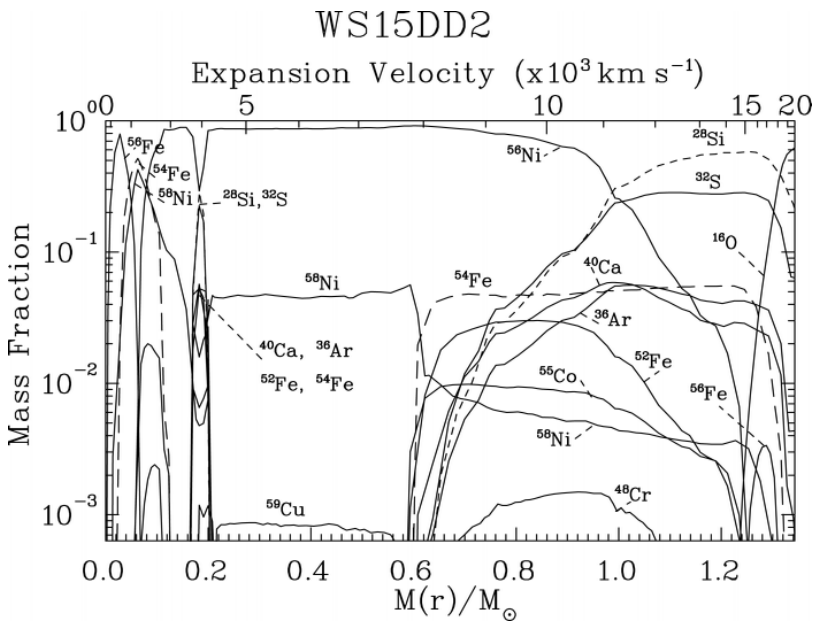


Figure 7. Composition of model WS15DD2 as a function of the expansion velocity and $M(r)$ for a delayed detonation model with a deflagration/detonation transition density of 2.2×10^7 g cm^{-3}. We see intermediate mass elements (Si–Ca) in the outer layers (at higher expansion velocities after the explosion), ^{56}Ni from explosive Si-burning dominating from about 0.2 to 1$M_\odot$, with features of incomplete Si-burning (^{54}Fe-plateau) and an alpha-rich freeze-out (^{58}Ni-plateau). The very central regions experience substantial electron capture and show more neutron rich nuclei (see figure 6).

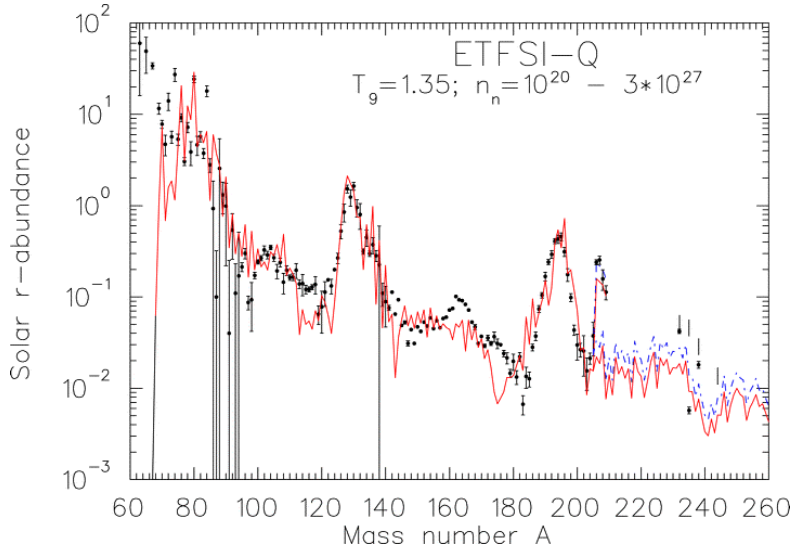


Figure 8. Comparison of theoretical abundances prior to and after β- and α-decay with solar r-process abundances (small filled circles). For $A > 206$ two superpositions are shown, obtained from fitting either the three r-process peaks or the global mass region $A = 125$–209 with r-process calculations obtained from 17 equidistant $S_n(n_n,T)$ components between 1 and 4 MeV. Both calculations make use of the ETFSI-Q mass model which assumes a quenching of shell effects far from stability.

produced in incomplete Si-burning. Combining all requirements on transition densities from total ^{56}Ni-yields, Si/Fe and Ia/II ratios, as well as specific elements favored in incomplete Si-burning, argues for a transition density close to 2×10^7 g cm^{-3}.

The r-process

The site-independent classical analysis based on neutron number densities n_n, temperatures T, and duration times τ assumed constant values of n_n and T and a fast drop of these quantities at $t = \tau$. The combination of n_n and T is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

related to an r-process path in the nuclear chart along nuclei with a neutron separation energy $S_n(n_n,T)$ or neutron capture Q-value (see figure 3 in the article on NUCLEAR PROPERTIES). A continuous superposition of components with neutron separation energies in the range 4–1 MeV on timescales of 1–2.5 s, provides a good overall fit (see figure 8). For the heavier elements beyond A = 130 this reduces to about S_n = 3–1 MeV. The beta-decay properties along contour lines of constant S_n towards heavy nuclei are responsible for the resulting abundance pattern. These are predominantly very unstable nuclei not accessible in laboratory experiments to date. A recent detailed analysis of the A = 206–209 abundance contributions to Pb and Bi isotopes from α-decay chains of heavier nuclei also allows the prediction of abundances of nuclei as heavy as Th with reasonable accuracy. The results of this study are shown in figure 8.

A different question is related to the actual astrophysical realization of such conditions. Stellar spectra of low metallicity stars, stemming from the very early phases of galactic evolution, are all consistent with a solar r-abundance pattern for elements heavier than Ba, and the relative abundances among heavy elements do not apparently show any time evolution. This suggests that all contributing events produce the same relative r-process abundances for the heavy masses, although a single astrophysical site will still have varying conditions in different ejected mass zones, leading to a superposition of individual components. The physical conditions must vary smoothly, as expected from a single astrophysical site. (This is opposite to the s-process, where the solar abundances are only an average of many sites with different features.)

Some meteoritic abundances and very recent observations of low metallicity stars indicate that at least two r-process sources have to contribute to the solar r-process abundances. The observed non-solar r-process pattern for e.g. Ag, I, and Pd demonstrates the need for a second r-process component in the nuclear mass range $A \approx 80$–120 in addition to the main process which provides a solar r-process pattern for $A > 130$. It is not clear, yet, which of the two processes is related to SNe II and which one is related to possible other sources.

To identify possible stellar sites, the basic requirement is a surplus of 10 to 150 neutrons per seed nucleus (in the Fe-peak or somewhat beyond) which have to be available to form all heavier r-process nuclei by neutron capture. For a composition of Fe-group nuclei and free neutrons that translates into a Y_e = 0.12–0.3. Such a high neutron excess is only possible for high densities in neutron stars under beta equilibrium ($e^- + p \rightarrow n + \upsilon$, $n \rightarrow p + e^- + \overline{\upsilon}$), based on high electron Fermi energies which are comparable to the neutron–proton mass difference. Neutron star mergers from binary neutron star systems eject such matter and are a possible (low entropy) site. Recent calculations show that on average about $10^{-2}M_\odot$ of neutron-rich matter can be ejected and preliminary calculations with assumptions on Y_e predict a solar-type r-process pattern for nuclei beyond A = 130. The smaller masses are depleted due to a long duration r-process with a large neutron supply in such neutron-rich matter, which also leads to fission cycling. This seems (accidentally?) to be in accordance with the main observed r-process component. Given the frequency ($\approx 10^{-5}$ y^{-1} per galaxy) and amount of ejected matter, this component alone could be responsible for the heavy solar r-process pattern and also explain the large scatter of r/Fe elements found in low metallicity stars. Neutron star–black hole mergers have not been analysed yet with the same accuracy, but bear similar options.

Another option is an extremely alpha-rich (i.e. high entropy) freeze-out in complete Si-burning with moderate $Y_e > 0.40$, related to SNe II. Given the galactic occurrence frequency, SNe II would need to eject $\sim 10^{-5}M_\odot$ per event. Explaining the r-process by ejecta of SNe II faces two difficulties: (i) whether the required entropies can really be attained in supernova explosions has still to be verified, (ii) the mass region 80–110 cannot be reproduced adequately. It has to be seen whether the inclusion of non-standard neutrino properties can cure both difficulties. It is also possible that none of this very high entropy matter escapes the neutron star surface and that outer mass zones in explosive He-burning introduce a 'fast' s-process, also dubbed an n-process, which just supplies the lower mass region ($A < 120$), missing or underproduced in some of the low metallicity stars, with the strongest r-process pollution. We should keep in mind that the question whether we understand fully all astrophysical sites leading to an r-process is not a settled one.

Conclusions

After an introduction into hydrostatic and explosive burning, the working of nucleosynthesis in general and its mathematical description, this overview concentrated on astrophysical sites or nucleosynthesis processes in the big bang, low and intermediate mass stars, type II and type Ia supernovae and analysed the options and sites of r-process nucleosynthesis. These are the major contributions to galactic chemical evolution.

Nucleosynthesis calculations have a right on their own to predict abundance patterns for many stellar events, but they can also serve as a tool to test the correctness of model descriptions, either in comparison to direct observations or indirect information from galactic evolution. We tried to show especially for SNe Ia and II how specific isotopic abundances can test ignition densities, burning front speeds, deflagration/detonation transitions or explosion energies (entropies), temperatures and delay times between collapse and explosion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

There exist other nucleosynthesis sites not mentioned here, e.g. (i) the p or γ process (explosive neon burning in possibly both types of supernovae, producing proton-rich stable nuclei via photodisintegration from pre-existing heavy nuclei), or events in binary systems like (ii) novae (small hydrogen accretion rates onto a white dwarf and explosive ejection of the accreted layer), and (iii) x-ray bursts (hydrogen accretion onto a neutron star and explosive processing of the accreted layer). With the possible exception of ^{15}N, ^{26}Al and other proton-rich nuclei lighter than S (in novae) or possibly some of the hard-to-make p-process nuclei in the range of Sr and Mo (in x-ray bursts), neither novae nor x-ray bursts are primary nucleosynthesis sites. This is due to the small or negligible amounts of ejected mass. The energy in x-ray bursts might not be at all sufficient to overcome the gravitational potential. But in all of these cases there exists very interesting nuclear physics on the proton-rich side of stability (rp-process), which is important to understand these explosive astrophysical events.

Bibliography

Arnett W D 1996 *Nucleosynthesis and Supernovae* (Princeton, NJ: Princeton University Press)

Grevesse N and Sauval A J 1998 Standard solar composition *Space Sci. Rev.* **85** 161–74

Iwamoto K *et al* 1999 Nucleosynthesis in Chandrasekhar mass models for type Ia supernovae *Astrophys. J. Suppl.* **125** 439–62

Käppeler F, Thielemann F-K and Wiescher M 1998 Current quests in nuclear astrophysics and experimental approaches *Annu. Rev. Nucl. Part. S.* **48** 175–251

Kratz K-L, Pfeiffer B and Thielemann F-K 1998 Nuclear-structure input to r-process calculations *Nucl. Phys. A* **630** 352–67

Schatz H *et al* 1998 rp-process nucleosynthesis at extreme temperature and density conditions *Phys. Rep.* **294** 167–264

Van Eck S, Goriely S, Jorissen A and Plez B 2001 Discovery of three lead-rich stars *Nature* **412** 793–795

Wallerstein G *et al* 1997 Synthesis of the elements in stars: forty years of progress *Rev. Mod. Phys.* **69** 995–1084

Woosley S E and Weaver T A 1995 The evolution and explosion of massive stars *Astrophys. J. Suppl.* **101** 181–235

F-K Thielemann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nuffield Radio Astronomy Laboratories

The Nuffield Radio Astronomy Laboratories are a part of the University of Manchester's Department of Physics and Astronomy. They are located at Jodrell Bank in Cheshire, UK.

In 1945, while it still functioned as the university's botanical station, the site was used by Bernard Lovell to observe cosmic rays. The radio echoes he detected were soon found to come from meteors burning up in the atmosphere.

Under the initiative of Lovell and engineer Charles Husband, the world's largest fully steerable telescope was completed in 1957. One of its first tasks was to use radar to track Sputnik 1, the world's first artificial satellite, and its rocket's upper stage. It was then in great demand for sending and receiving telemetry from US and Soviet satellites, particularly lunar and planetary missions.

Through a donation by Lord Nuffield and the Nuffield Foundation, the debts outstanding on the telescope's construction were cleared. The Jodrell Bank Experimental Station then became Nuffield Radio Astronomy Laboratories. On the telescope's thirtieth anniversary in 1987, the 76 m Mark I telescope was renamed the Lovell Radio Telescope.

Small portable telescopes were later linked to the Mark I telescope by microwave radio links to create a radio interferometer. These instruments played a major role in the discovery of quasars in the early 1960s.

The parabolic 38 × 25 m MKII telescope, completed in 1964, was built to a similar design to the Goonhilly 1 telescope which received the first TV signals relayed by the Telstar satellite across the Atlantic. It was the first telescope of any type in the world to be controlled by a digital computer, the Ferranti Argus 100.

In 1976, work began on the building of a 134 km array of telescopes linked by microwave radio links to Jodrell Bank. Three new 25 m telescopes were built to add to the MKII, MKIII and Defford telescopes already operated by Jodrell Bank. These were located at Knockin, near Oswestry, Pickmere, north of Northwich, and Darnhall, south of Winsford.

In the early 1990s the array, now called MERLIN, was extended to include a new 32 m telescope at Cambridge. It can be linked to other observatories around the world for VLBI studies.

Current research at Jodrell Bank includes studies of stars, pulsars, galaxies, quasars, gravitational lenses, cosmic microwave background and SETI.

For further information see http://www.jb.man.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Numerical Simulation of Stellar Convection: Comparison with Mixing-length Theory

The energy released by nuclear fusion in the STELLAR INTERIOR is carried to the surface by two different transport mechanisms, *radiation* and *convection*. In the presence of a temperature gradient, there is always a net radiative energy flux since more photons travel from hot to cool regions than in the other direction. If the temperature gradient exceeds a certain threshold, convection sets in. The resulting large-scale circulation of hot rising and cool sinking gas constitutes a very efficient mode of energy transport, usually dominating over radiation.

While an elaborate theory of RADIATIVE TRANSFER is well established, a rigorous theory of stellar convection is lacking due to the complexity of the underlying hydrodynamical problem. So far, stellar structure models still rely on a phenomenological approximation, the so called *mixing-length theory* (MLT). However, radiation hydrodynamics simulations based on first principles can now provide physically consistent *ab initio* models of stellar convection which can serve to check the validity of MLT.

Stellar convection

Convection is a universal feature: essentially all types of stars have either a convective core, a convective envelope, or both. Low-mass stars are fully convective, giants may accommodate several distinct convective shells. In the case of the Sun, the energy transport in the inner parts is entirely due to radiation (radiative layers), while in the outer 28.7% (in radius) it is primarily due to large-scale convective currents. At the surface, the solar granulation is the visible imprint of gas flows in the outermost layers of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE).

The role of stellar convection is far-reaching: convective energy transport determines the internal temperature structure of a star and its radius (which decreases with increasing convective efficiency), and hence controls the star's global properties. Convective regions are chemically completely mixed, and overshooting convective flows lead to partial mixing of the adjacent radiative layers. 'Overshoot' and similar mixing processes which are not confined to the convectively unstable layers are thought to be responsible for the existence of carbon stars, carbon-rich white dwarfs, and for the destruction of lithium in solar-type stars. Convective motions and concomitant temperature inhomogeneities exert a direct influence on stellar spectra, causing small but practically relevant changes in wavelength position, shape and strength of spectral lines. Stochastic convective motions can excite stellar oscillations (like the five-minute oscillations in the Sun) and are a source of acoustic energy, contributing to the heating of stellar chromospheres. In combination with DIFFERENTIAL ROTATION, convection is a necessary precondition for the operation of the magnetic dynamo mechanism, and hence for STELLAR ACTIVITY.

Schwarzschild criterion

A region in a star is convectively unstable if a gas element which is displaced from its equilibrium position is further accelerated due to buoyancy forces. In 1906, K Schwarzschild discovered that this situation occurs whenever the temperature stratification is superadiabatic, i.e. $\mathrm{d}\ln T/\mathrm{d}\ln P > (\mathrm{d}\ln T/\mathrm{d}\ln P)_{\mathrm{ad}}$, where T and P denote temperature and gas pressure, respectively (*Schwarzschild criterion*). For stellar applications (inviscid fluids), this is the necessary and sufficient condition for convective instability of a chemically homogeneous layer. In terms of the specific entropy, s, the criterion can be rewritten as $\mathrm{d}s/\mathrm{d}z < 0$, implying convective instability wherever entropy of the gas decreases with height, z.

In the envelope of stars, convective instability is favored by large radiative opacities (steep radiative temperature gradient) and large specific heats (small adiabatic temperature gradient). This situation is found particularly in regions where an abundant element (hydrogen or helium) is partially ionized. Convective instability also occurs in the core of main-sequence stars more massive than $\approx$1.3 times the Sun, where the strong temperature dependence of nuclear fusion rates leads to very steep temperature gradients.

For a given stellar structure, the Schwarzschild criterion can be used to determine the location of convective regions. However, it does not give any information about the properties of the corresponding convection zone, like flow velocities and convective energy flux. Moreover, since convection changes the thermal structure, the boundaries of the unstable region depend on the convective efficiency. In addition to the Schwarzschild criterion, a theory of convection is required to compute the structure of stellar convection zones in a self-consistent way.

Mixing-length theory

As a 'first approximation', the mixing-length theory was developed mainly by Biermann (1932), Cowling (1935), Vitense (1953), and Böhm-Vitense (1958). The underlying picture is that of turbulent elements ('bubbles') moving relative to a mean background stratification. In an unstable region, elements accelerated upwards gain an excess of thermal energy, while downward moving elements cool relative to their surroundings. After traveling some characteristic distance, ℓ, the 'bubbles' are assumed to dissolve completely. The mixing length, ℓ, determines the mean flow velocity, the typical amplitude of horizontal temperature fluctuations and the resulting convective energy flux.

This simplistic formalism is unsatisfactory in several respects. First of all, the mixing length, ℓ, which is usually expressed as $\ell = \alpha H_{\mathrm{p}}$ (where the local pressure scale height, H_{p}, measures the distance over which the gas pressure changes by a factor e^{-1}), is a *free parameter*. The *mixing-length parameter* α is *a priori* unknown and must be determined empirically; its value may depend on stellar type. Moreover, MLT is a local theory: convection

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is restricted to the unstable region as defined by the Schwarzschild criterion, and the transport properties of convection are assumed to depend only on local thermodynamic variables and their local gradient. But actually convection is non-local: the velocity of a gas element at some point, for example, is determined by the acceleration it has experienced before along its path, not by local conditions. Obviously, MLT is not based on sound physical principles, and some of the basic assumptions are not even valid in real stars. Still, almost all current stellar evolution models rely on this phenomenological 'theory'. The same holds for modeling convective STELLAR ATMOSPHERES.

Numerical simulations

The differential equations governing the physics of stellar convection are well known but difficult to solve: these are the conservation equations of hydrodynamics, coupled with the equations of radiative transfer. While their highly nonlinear and non-local character precludes an adequate analytical treatment, the numerical integration of the system of partial differential equations has now become feasible. This approach constitutes an increasingly powerful method to study in detail the time-dependent properties of a radiating, partially ionized fluid under stellar conditions.

Using modern supercomputers, it is possible today to perform three-dimensional numerical simulations of stellar convection with realistic background physics (equation of state, radiative opacities). Although such simulations are usually confined to a small section near the surface, even two-dimensional convection models require substantial amounts of computer time, particularly for solving the multidimensional, non-local radiative transfer problem in several wavelength bands, taking into account the influence of spectral lines. In order to yield meaningful information, the simulated time sequence must be long enough to ensure relaxation towards a statistical steady state which is independent of initial conditions. Also, the horizontal size of the computational domain must be large enough to accommodate at least two 'convection cells'.

Just like 'classical' stellar atmospheres, the hydrodynamical models are characterized by the average total energy flux per unit area and time F (related to the *effective temperature* of the stellar atmosphere $T_{\rm eff}$ via $F = \sigma T_{\rm eff}^4$, σ being the Stefan–Boltzmann constant), *surface gravity* g, and *chemical composition*. But in contrast to the mixing-length models, there is no longer any free parameter to adjust the efficiency of the convective energy transport. Similarly, spatial and temporal scales are not imposed on the simulations, but rather develop naturally as a consequence of the underlying equations.

MLT versus hydrodynamics

Time-dependent numerical simulations of stellar convection can be used for a qualitative comparison with the basic picture and assumptions underlying the mixing-length theory, and for a quantitative comparison of the average transport properties.

While the granules ('bubbles') seen at the solar surface have a typical size (dominant contribution to area coverage) of ≈1 000 km, the 'bubbles' considered in mixing-length models have a size of the order of $\alpha H_{\rm p}$, which is about five times smaller. Clearly, the MLT concept of single-sized 'bubbles' must not be taken too literally. On the other hand, the numerical convection models are capable of reproducing the size distribution of solar granules and other statistical properties rather well. As these models teach us, the typical granule size is determined by a complex interplay of nonlinear dynamics and radiative energy exchange.

According to the simulations, the dynamics of convection is dominated by *fast, cool, narrow downdrafts* (sometimes also called 'plumes' or 'jets'), embedded in much more extended, slowly ascending, warm material (granules). This pronounced up/down asymmetry is a general feature of the numerical models (2D as well as 3D), and is seen for all types of stars investigated so far. Obviously, this result is in stark contrast to the picture of MLT: instead of 'bubbles' traveling for about one pressure scale height before dissolving, we actually have deep-reaching 'jets' surviving over many $H_{\rm p}$, possibly extending all the way to the base of even deep convective

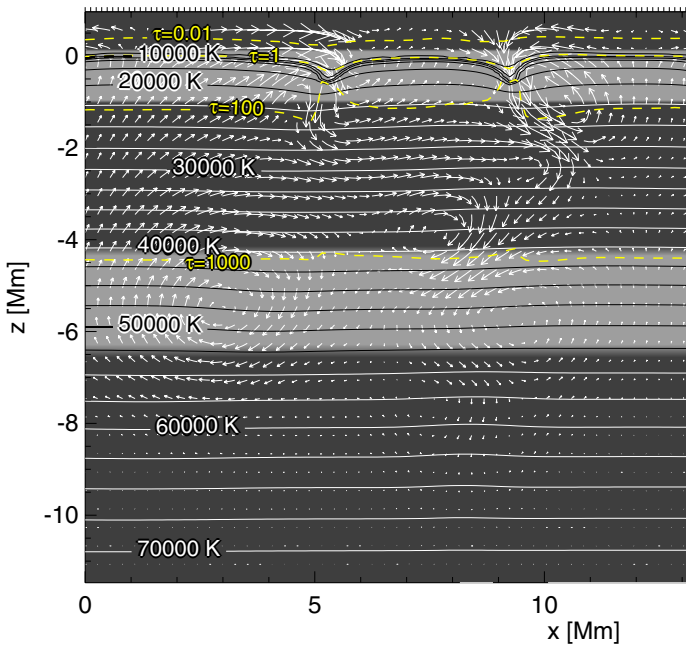


Figure 1. Snapshot from a 2D simulation of convection in an A-type star ($T_{\rm eff} = 80$ two separate convection zones. The unstable regions ($d\bar{s}/dz < 0$, $\bar{s}$ being the horiz gray. The velocity field is represented by pseudo streamlines, indicating the displa velocity 15.4 km s^{-1} at this moment); the thermal structure is outlined by tempera levels of constant optical depth ($\tau = 0.01, 1, 100, 1000$). Geometrical height $z = 0$ c non-local, multidimensional radiative energy exchange and partial ionization. The zone; lateral boundaries are periodic. The 'dents' seen at the 'surface' ($z = 0$) are t intergranular lanes. (From Freytag *et al* (1996)).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

envelopes. The downward 'jets' originate and are accelerated in the stellar surface layers, which are highly unstable (steep temperature gradient, density inversion) as a consequence of efficient radiative cooling. Since their temperature deficit relative to the surroundings is determined by the efficiency of surface cooling, convection is an extremely non-local process. The properties of the surface determine the dynamics and structure of the whole convection zone. Certainly, the inherently 'linear' approach enforcing symmetry between up and down, as well as the assumption of locality, are the major problems with MLT.

A variety of observations (e.g. the fact that the granulation is still visible at the solar limb, corresponding to optical depths $\tau < 0.3$), and numerical simulations indicate that convective flows are not limited to the Schwarzschild-unstable regions, but rather must penetrate into the adjacent stable layers by some amount. Though not inherent in MLT, the problem of 'overshoot' was first recognized, considered and discussed in the original work by Vitense (1953). There is no simple rule to predict the amount of 'overshoot', since it depends on the nature of the particular convective boundary under consideration.

The role of 'overshoot' is illustrated in figure 1. In this A-type main sequence star there are two distinct convection zones: the one at the surface is driven by the combined first ionization of hydrogen and helium, the deeper one is related to the second ionization of helium. Although the Schwarzschild-unstable regions are separated by a stable radiative layer extending over more than two pressure scale heights, they are effectively connected by vigorous convective currents; the stable buffer layer is completely mixed, in contrast to MLT predictions.

It is a physical property of overshoot that the pressure field that serves to drive the motions in the unstable regions extends far into the stable layers and accelerates the fluid there, even though the thermal stratification is convectively stable. The hydrodynamical simulations demonstrate that the velocity field due to overshooting declines exponentially with distance from the Schwarzschild boundary. In the example of figure 1, the velocity scale height below the He II convection zone is $H_v \approx 0.3 H_p$. 'Overshoot' above the surface convection zone is the explanation for 'microturbulence' and the asymmetry of spectral lines; it cools the upper photosphere slightly below radiative equilibrium temperatures.

In order to achieve agreement between observed and computed stars (isochrone fitting), stellar evolution models have to include some amount of 'overshoot', which is usually implemented by introducing another parameter, d_{over}, specifying the distance (in terms of H_p) by which convection penetrates into the stable layers. This empirical approach ignores the hydrodynamical results indicating that 'overshoot' is characterized by an exponential velocity field causing (partial) diffusive mixing.

A quantitative comparison of mixing-length models and averaged results from numerical simulations has been performed for various stellar types, including the Sun, A-type main sequence stars and DA WHITE DWARFS. The result is similar in all cases: no single choice of α can produce a detailed match of the structure of the hydrodynamical models in the framework of MLT. Adjusting α can only serve to fit particular properties, like maximum convective flux or maximum convective velocity, but it is impossible to reproduce the depth dependence of these quantities correctly. This point is illustrated in figure 2: while $\alpha = 1.5$ gives the correct maximum enthalpy flux, $F_{h,max}$, MLT fails to match the flux profile $F_h(z)$ even roughly. A much larger α would be required to reproduce the correct depth of the convection zone ($F_h > 0$). This result resolves an apparent inconsistency which arises when modeling ZZ CETI STARS (a class of non-radially pulsating white dwarfs) with MLT. The observed spectrum (atmospheric structure) is best described adopting a moderate value of $\alpha \approx 1.5$, but at the same time a large convective efficiency ($\alpha > 2.5$) is required for driving the observed pulsations of these stars. Note that the simulations show a distinct region of negative (downward) F_h—due to 'penetration' into the stable layers—and a significant negative kinetic energy flux. Both these features are missing in MLT.

While it seems hopeless to successfully model the structure of the superadiabatic surface layers within the framework of MLT, it is possible to calibrate MLT through

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

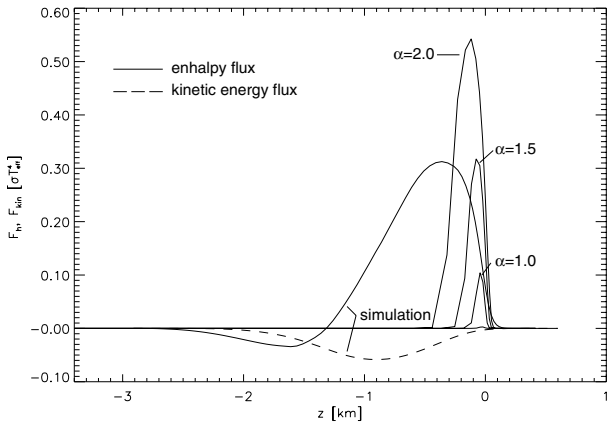


Figure 2. Comparison of the enthalpy flux, $F_{\rm h}$, predicted by MLT for different α and a 2D hydrodynamical simulation for the outer layers of a DA white dwarf ($T_{\rm eff} = 12\,200$ K, $\log g = 8.0$ [cgs units]). The flux of kinetic energy, which is ignored in MLT, is negative (downward) according to the simulation. (Adapted from Freytag *et al* (1996)).

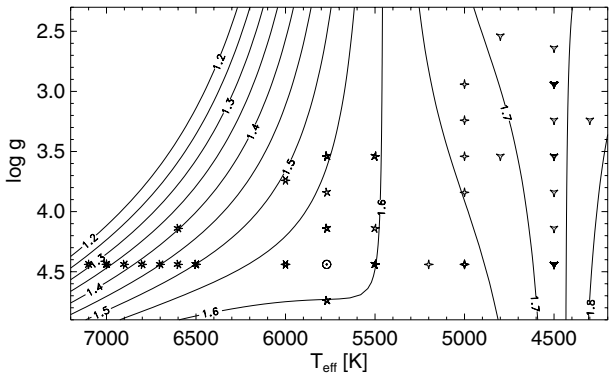


Figure 3. Calibration of the mixing-length theory for stellar evolution calculations based on 2D hydrodynamical simulations of surface convection. Symbols indicate that at least one simulation has been performed for the corresponding stellar parameters. For each simulation, an effective mixing-length parameter α has been determined from a differential comparison with MLT models. Contour lines, labeled with the respective value of α, visualize the derived variation of α as a function of $T_{\rm eff}$ and $\log g$ (From Ludwig *et al* (1999)).

hydrodynamical simulations for application to STELLAR EVOLUTION. The key quantity here is the entropy jump, Δs^*, from the surface to the interior, which can be 'measured' from the simulations. MLT can then be calibrated by adjusting α to produce the given Δs^*. This procedure gives the correct adiabat and depth of the convective envelope, as verified for the Sun by comparison with helioseismic measurements.

For a sample of 'solar-type' stars, the result of this calibration is shown in figure 3. It is the first theoretical prediction of how the mixing-length parameter α depends on stellar type. According to this calibration (which needs to be verified by 3D hydrodynamics), the variation of α over the investigated region of the HERTZSPRUNG–RUSSELL DIAGRAM is only moderate (from 1.8 to 1.2 for spectral type K to F). Certainly, this is one of the reasons for the surprising success of MLT over the past 40 years.

There have been many attempts to improve MLT. A recent example is the CGM model (Canuto, Goldman and Mazzitelli 1996), which accounts for the full spectrum of eddy sizes in a turbulent flow and adopts a mixing-length of the form $\ell = z + \zeta H_{\rm p}^{\rm top}$, where z is the distance to the (nearest) Schwarzschild boundary and ζ is a free parameter related to 'overshoot'. In comparison with MLT, CGM makes efficient convection more efficient and inefficient convection less efficient, implying that the variation of ζ across the Hertzsprung–Russell diagram is even smaller than that of α. Further claims that CGM is superior to MLT need to be verified. It is highly doubtful, however, that convection can be properly described in terms of classical turbulence, since the basic assumptions typically made in these models are found to be violated in the hydrodynamical simulations: in a stratified medium, large-scale motion is strongly anisotropic while small-scale turbulence is confined to the downdrafts and contributes little to the energy transport, which is dominated by scales exceeding the local pressure scale height by an order of magnitude.

Further details can be found in the article on SOLAR INTERIOR: CONVECTION THEORY.

Bibliography

Biermann L 1932 Untersuchungen über den inneren Aufbau der Sterne. IV. Konvektionszonen im Inneren der Sterne *Z. Astrophys.* **5** 117

Böhm-Vitense E 1958 Über die Wasserstoffkonvektionszone in Sternen verschiedener Effektivtemperaturen und Leuchtkräfte *Z. Astrophys.* **46** 108

Canuto V M, Goldman I and Mazzitelli I 1996 Stellar turbulent convection: a self-consistent model *Astrophys. J.* **473** 550

Cowling T G 1935 The stability of gaseous stars *Mon. Not. R. Astron. Soc.* **96** 42

Freytag B, Ludwig H-G and Steffen M 1996 Hydrodynamical models of stellar convection *Astron. Astrophys.* **313** 497

Gimenez A, Guinan E F and Montesinos B (ed) 1999 *Theory and Tests of Convection in Stellar Structure (Astron. Soc. Pacific Conf. Ser. 173)*

Ludwig H-G, Freytag B and Steffen M 1999 A calibration of the mixing-length for solar-type stars based on hydrodynamical simulations *Astron. Astrophys.* **346** 111

Ludwig H-G, Jordan S and Steffen M 1994 Numerical simulations of convection at the surface of a ZZ Ceti white dwarf *Astron. Astrophys.* **284** 105

Schwarzschild K 1906 Über das Gleichgewicht der Sonnenatmosphäre *Göttinger Nachr.* **1** 41

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spruit H C, Nordlund Å and Title A M 1990 Solar convection *Ann. Rev. Astron. Astrophys.* **28** 263
Stein R F and Nordlund Å 1998 Simulations of solar granulation. I. General properties *Astrophys. J.* **499** 914
Vitense E 1953 Die Wasserstoffkonvektionszone der Sonne *Z. Astrophys.* **32** 135

Matthias Steffen and Stefan Jordan

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nutation

Short-period oscillations in the long-term motion of the pole of rotation of a rotating body that is being acted on by external perturbing forces. In the case of the Earth, the gravitational attractions of the Sun and Moon on its equatorial bulge causes the direction of its axis, and hence the positions of the celestial poles, to revolve slowly around the pole of the ecliptic (the points on the celestial sphere 90° away from the ecliptic), this phenomenon being called precession. Because the orbital plane of the Moon is inclined to the plane of the ecliptic by an angle of about 5°, and revolves round the Earth in a period of 18.6 years, an additional small oscillation in the position of the celestial poles is superimposed on the general precessional motion. The effect of this additional motion is to cause the Earth's poles (and hence the celestial poles) to tilt, or 'nod', periodically toward and away from the poles of the ecliptic, the term 'nutation' deriving from this 'nodding' motion. The magnitude of this nutational motion is about 9 arcsec, and its period of oscillation is 18.6 years. Nutation also gives rise to a corresponding periodic variation in the obliquity of the ecliptic.

See also: celestial poles, ecliptic, node, obliquity of the ecliptic, precession.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

OB Associations

The term OB association is used to describe a large collection of gas and young stars (O-type and B-type) located primarily in the disk of SPIRAL GALAXIES. There is a great variance of size and star number among these associations, from ten stars or fewer at the lower end up to tens of thousands of stars. The latter case is sometimes referred to as an OB superassociation (NGC 206 in the ANDROMEDA GALAXY, for example).

OB associations are among the least dense of the STAR CLUSTERS. Most OB associations lack the structure and central compactness present in the more easily discernible GLOBULAR CLUSTERS. In some cases, the overpopulation of O-type and B-type stars is the only factor that allows the association to be distinguished from the remaining stars in its vicinity. Because of this lack of structure, OB associations fall within the category of open clusters.

There is a considerable difficulty in studying a complete sample of OB associations in our own galaxy. This difficulty arises from the fact that OB associations, as well as the other types of open clusters, are highly concentrated toward the GALACTIC PLANE. This means that the more distant OB associations in our galaxy will most likely be obscured by dust. As a result, studying the distribution of OB associations in other nearby galaxies allows us to learn more about the probable distribution of such associations in the Milky Way.

One of the most important roles for OB associations in the field of astronomy is to act as tracers for both current and recent massive STAR FORMATION. In other words, regions where OB associations are found are interpreted as areas where massive stars are currently being formed or have just recently been formed. If knowledge can be gained regarding the environments in which OB associations tend to be found, then that knowledge should directly lead toward what environments trigger massive star formation.

Determining the most accurate description for the dimension and content of any OB association requires knowledge about the spectral type of each potential member star. For our own galaxy, as well as other nearby galaxies, this is at least within the realm of possibility. It should be noted that nearby galaxies with small inclination angles are preferred, since this minimizes complications due to projection effects. For galaxies that are more distant, only signatures from the associations as a whole can be detected. This can be done from the actual stars themselves (ultraviolet light) or from the gas surrounding the stars as it becomes excited (Hα emission). It should be noted that while all OB associations will have an overdensity of O-type and B-type stars by definition, some may be of sufficient age that they will not have a detectable amount of gas remaining around them.

Current research is being done to locate OB associations both in the Milky Way and in other nearby galaxies. Information from both catalogs, such as HIPPARCOS AND TYCHO CATALOGUES, and observations, including some from the HUBBLE SPACE TELESCOPE, is being used in these projects. This research is very important to the understanding of the distribution of OB associations and the environments in which massive stars are most likely to form.

Robert Mohr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oberon

A mid-sized satellite of Uranus, discovered by William Herschel in 1787. Its diameter is 1520 km and it orbits at a distance of 191 000 km. The images of Oberon obtained by Voyager 2 were not clear, but were enough to show an extensively cratered surface, some craters surrounded by bright ejecta or rays, and some with dark floors, possibly a result of the eruption of water ice mixed with organic material. The largest crater is the 206 km diameter Hamlet. An 11 km mountain peak, possibly the central peak of an impact basin, was visible on the limb. The preponderance of craters suggest that this is an ancient surface, largely undisturbed by geological activity, save for the 500 km muted fault feature Mommur Chasma. One hemisphere is rather redder than the other, perhaps as a result of dust accumulated from the two small outer retrograde satellites of Uranus. Like most of Uranus's satellites, Oberon has captured rotation, so the dust would be swept up by the leading hemisphere.

See also: Uranus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oberth, Hermann (1894–1989)

German space travel visionary, born in Sibiu, Hungary. His book *Die Rakete zu den Planetenrämen* (*The Rocket into Interplanetary Space*), established his reputation in 1923, and he became president of the German Society for Space Travel. In the Second World War he worked on rockets at Peenemünde, and went with WERNHER VON BRAUN to the US Army Ballistic Missile Agency in Huntsville, Alabama to develop rockets in the American space program.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Objective Lens

The principal lens of a refracting telescope, which collects light from a distant object and forms an image of the object at its focus, or focal plane. The objective lens, or objective, is also known as the object glass (OG) because, in a telescope, it is the lens nearest to the object that is being imaged.

The simplest form of objective is a single biconvex lens in which the curvature of each face corresponds to part of a sphere, the radius of which is called the radius of curvature. A ray of light passing through the center of the lens, perpendicular to its surface, passes through undeviated. Rays entering other parts of the lens meet its surface at an angle to the perpendicular and are refracted (deflected) so as to converge together and form an image. A simple objective lens of this kind suffers from a number of optical defects, or aberrations, notably spherical aberration (an inability to bring rays that enter different parts of the lens to the same focal point) and chromatic aberration (an inability to bring all wavelengths, or colors, to a focus at the same point). Chromatic aberration arises because different wavelengths of light are refracted by differing amounts, shorter wavelengths (e.g. blue) being refracted more than longer wavelengths (e.g. red). Consequently, blue light is brought to a focus closer to the lens than red light and the focal points for the various different colors lie at different points along the optical axis (the line that passes perpendicularly through the center of the lens). The range of focal points for light of different wavelengths is called the longitudinal chromatic aberration, or secondary spectrum.

Both of these aberrations are considerably reduced by the achromatic lens, a compound lens consisting of two components, each made of different kinds of glass with different optical properties. The combination of two lenses is called an achromatic doublet. An achromatic doublet normally consists of a converging (biconvex) lens, usually made of crown glass, and a diverging lens (usually plano-concave: concave on the front surface and flat on the back), usually made of flint glass. The radii of curvature are chosen so that the positive spherical aberration introduced by one component is to some extent cancelled by the negative spherical aberration induced by the other. The different dispersive powers (ability to refract different wavelengths by differing amounts) of the two types of glass are used to minimize the spread of focal positions. An achromatic doublet is normally figured (shaped) to bring two particular wavelengths, for example 400 nm (indigo) and 600 nm (orange), to the same focal point. The spread in focal positions for the other wavelengths are usually reduced by a factor of 10–20 compared to the spread that would be produced by a single lens. The addition of a third lens further reduces the spread of focal positions. A three-element objective lens is called an apochromat.

As the residual chromatic aberration decreases with increasing focal ratio, achromatic doublets usually have focal lengths that are at least ten times their apertures (focal ratio f:10 or more). Objective lenses with large focal ratios have small fields of view. Where a wider field of view and shorter focal ratio is required, an apochromat is normally used.

See also: aperture, chromatic aberration, field of view, focal length, focal ratio, refractor, spherical aberration, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oblateness

The degree of flattening of an oblate spheroid, the solid body obtained by rotating an ellipse about its minor axis. The shape of the Earth is approximately an oblate spheroid, for the equatorial diameter is slightly greater than the polar diameter. Jupiter and Saturn are considerably more oblate than the Earth. If a hypothetical planet had a polar radius of 10 000 km and an equatorial radius of 20 000 km, the oblateness (or flattening) would be:

$$\frac{20\,000 - 10\,000}{20\,000} = 0.5.$$

By measuring the degree of flattening at the poles of a planet as compared with the speed of rotation, it is possible to infer the density distribution inside the planet. If two planets had the same mass and bulk density, the planet with most of its mass concentrated close to the center would be more flattened by rotation. For example, from the relatively small oblateness of Uranus combined with the planet's relatively rapid rotation, it would appear that its constituents, ice and gas, are well mixed and a rocky core is small or non-existent.

See also: axis, rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Obliquity of the Ecliptic

The angle between the planes of the ecliptic and the equator. On the celestial sphere, the angle at which the ecliptic intersects the celestial equator. The current (year 2000) value of the obliquity of ecliptic, which is denoted by the symbol ε, is $23^\circ\ 26'\ 21''$. Its value varies by $\pm 9''$ over a period of 18.6 years as a consequence of a phenomenon called nutation. Over a much longer period (about 40 000 years) the perturbing influence of the planets causes its value to vary between $21^\circ\ 55'$ and $28^\circ\ 18'$; currently ε is decreasing at an average rate of about $0.5''$ per year.

See also: celestial equator, ecliptic, nutation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Besançon

Founded in 1878 to provide an astronomically derived time reference service for the region's watch-making industry, the Observatory today continues its research and service activities in the field of time and frequency metrology and has also developed active research groups in the fields of stellar and galactic evolution and the chemical physics of astrophysical media. A galactic stellar population simulation model and a database of double and multiple stars may be consulted via the Observatory's web site. Occupying an attractive park and buildings, the Observatory's instruments include a meridian telescope (2.37 m focus, 19 cm diameter) dating from the 1880s, an astrograph (two 3.5 m and one 3 m focal length instruments) commissioned in 1956 and a modern time and frequency laboratory equipped with three cesium atomic clocks. The staff is made up of 16 university lecturers and other researchers, 11 engineers and technical personnel and 6 support staff, plus a variable number of students up to PhD level. The Observatory is a part of the Université de Franche-Comté and is recognized by the Centre National de Recherche Scientifique (CNRS/INSU) and the Bureau National de Metrologie (BNM).

For further information see www.obs-besancon.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de la Côte d'Azur

This French public institute was founded in 1988, with the mission to collect and interpret astronomical data. In 1881 Raphael Bischoffsheim founded and equipped its oldest component, Nice Observatory. The dome of its great refractor, 76 cm, was built by Eiffel. The second component, the CERGA, founded in 1974 for positional astronomy, is located near Grasse.

For further information see http://www.obs-nice.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Lyon

The Observatoire de Lyon is a laboratory of Université Claude Bernard Lyon1, under the authority of the INSTITUT NATIONAL DES SCIENCES DE L'UNIVERS. It constitutes the Centre de Recherche astrophysique de Lyon together with the astrophysics group of the École Normale Supérieure de Lyon.

It operates a 1 m Ritchey–Chrétien telescope, which is now mainly used by students and the general public. Formerly installed on the summit of Gornergrat in Switzerland, the instrument was returned to Lyon in 1983 for trials of various instruments—including photometers, an infrared camera and the OASIS spectrograph—that were developed at the observatory.

For further information see
http://www-obs.univ-lyon1.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Marseille

The Marseille Observatory was founded by the Jesuits in 1702 and taken over by the State in 1763. A century later it was transferred to its present site on the Longchamp Plateau in Marseille, southern France.

The observatory, which is part of the University of Provence, belongs to the Provence Astronomical Federation. It owns an 80 cm telescope, made in 1867 by Léon Foucault, which was the first large telescope with a silvered mirror ever built. It was used for observations until 1960 and designated an historical monument in 1993.

The observatory is funded by a number of sources including the Ministry of Higher Education and Research, CNRS, the Gassendi Institute and regional authorities. It also carries out research and contracts for industry, and participates in international projects such as the CANADA–FRANCE–HAWAII TELESCOPE and the EUROPEAN SOUTHERN OBSERVATORY.

Scientific research covers all areas of astronomy, from optical to radio and infrared observations.

For further information see http://www-obs.cnrs-mrs.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Paris

The Observatoire de Paris is the largest French institution dedicated to astronomy. Created in 1667 by Louis XIV to define the 'Méridien de Paris' and then devoted to metrology, celestial mechanics and positional astrometry, it further developed many branches of astrophysicis and is now established on three sites: the initial Paris Observatory, the Meudon Observatory (south-west suburb of Paris) appended in 1926, and the radioastronomy facility of Nançay (central France), appended in 1953. Research activities cover many fields from theoretical to observational astronomy, experimental physics and chemistry, instrument development for international ground-based observatories and space experiments for astronomy and planetology missions. Services related to its research activities are provided to the astronomical community (time and frequency metrology, solar monitoring, International Earth Rotation Service, etc). The observatory also has an educational mission and provides an astrophysical research degree and PhD course, courses in astronomy for high school teachers, visits and exhibitions for the public.

The Observatoire de Paris is a governmental institution. Its sources of support are, for the most part, contracts with the Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT), the Centre National de la Recherche Scientifique (CNRS), the Centre National d'Etudes Spatiales (CNES), the European Community, or with European scientific agencies such as the EUROPEAN SOUTHERN OBSERVATORY (ESO) or the EUROPEAN SPACE AGENCY (ESA).

The Observatoire de Paris comprises ten research groups. Its staff is of about 800 people, of whom 640 are on permanent jobs (230 scientists) and 160 are students and research fellows on long-term contracts.

For further information see
http://www.obspm.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Section de Meudon, Paris

The Meudon Observatory in France was founded on an old royal estate in 1876. It was built with public funds and placed at disposal of astronomer Jules Janssen, in order to allow him to develop his research on solar spectra far from urban pollution.

Several instruments were installed by 1893: an 83 cm refractor placed under the dome over the castle; a 1 m reflector, which was restored in 1969; and various spectroheliographs and siderostats. Later additions included an equatorial table, a 60 cm telescope, and a solar tower, which was used for spectroscopic studies of the Sun.

It merged with the Paris Observatory in 1926. The section is now devoted to theoretical astrophysics, with departments devoted to studies of stars, the interstellar medium, galaxies and cosmology.

For further information see
http://www.obspm.fr/index.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire de Strasbourg

The Strasbourg Astronomical Observatory, located in Strasbourg, France, is a research unit of Université Louis Pasteur and CNRS. It hosts teaching and research activities, the services of the Strasbourg Astronomical Data Center (CDS), and public outreach activities, with its Planetarium.

Founded in 1881, the Strasbourg Observatory currently carries out research in the following fields: stellar populations and galactic evolution, galactic dynamics, high-energy astrophysics, data and information management. It is active in the development and exploitation of space missions such as Hipparcos or XMM.

The Data Center is used by professional astronomers worldwide for its databases and services related to astronomical objects outside the solar system.

For further information see
http://astro.u-strasbg.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire des Sciences de l'Univers de Bordeaux

Ground-based observations (meridian circle, radiotelescope, magnetic resonance spectrometer, radiometer, specrophotometer) are one of the main features of 'Observatoire des Sciences de l'Univers de Bordeaux, laboratoire d'astrodynamique, d'astrophysique et d'aéronomie', which is located in its original place of foundation, in 1878, at Floirac, in the suburbs of Bordeaux.

Dedicated to research, teaching and science for the general public, the Observatory is run by the Centre National de la Recherche Scientifique (CNRS), and the Teaching and Research Ministry. There are about thirty scientists and students, and six research groups: astrodynamics: (astrometry/Hipparcos, galactic dynamics, dynamic of the solar system, nucleus of galaxies); radio-astronomy (interstellar medium, galaxies and quasars, star formation, space missions (First, Rosetta)); radio-aeronomy (ground based microwave radiometry, space radiometry (UARS, Odin), ozone international network); atmosphere of planets (photochemistry of giant planets, primitive atmospheres, atmosphere of Mars (Mars-Express probe), and Saturn (Huygens probe)); Exobiology (prebiotic atmosphere of the Earth, extra-terrestrial molecules and matter); and exoplanets.

For further information see
http://www.observ.u-bordeaux.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire des Sciences de l'Univers de Grenoble

Young stars and their environment, their formation and evolution, plasmas in active galactic nuclei, together with adaptive optics and optical interferometry research and development, are the main activities of the Laboratoire d'Astrophysique. Created in 1979 on the campus, later forming the Observatoire with geophysical laboratories, it has a permanent staff of around 50. It is a joint facility of CNRS and the Joseph Fourier University.

For further information see http://www-laog.obs.ujf-grenoble.fr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatoire Royal de Belgique

This observatory was founded in 1826 and installed in the present site in 1891. It is organized into four departments, concerned with public services and scientific activities in: time; Earth rotation; space geodesy; seismology; Earth tides; gravimetry; astrometry and dynamics of objects in and outside the solar system, astrophysics of Galactic and extra-Galactic objects; stellar atmospheres; and solar physics (structure and dynamics of the solar atmosphere, solar activity). The library contains about 120 000 books. A large Planetarium, receiving about 30 000 visitors each year, is managed by the Royal Observatory. The budget is supported by the Ministry on Science Policy (60%) and by special projects initiated by the Royal Observatory (40%).

The staff is composed of 32 permanent researchers and 50 technicians, who collaborate in public services and scientific researches, administration and maintenance. Also there are, on average, 10 to 15 researchers on contract (PhD students, post-doctoral positions. . .).

For further information see
http://www.OMA.BE/KSB-ORB/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatories and their Location

As astronomy and astronomical instruments have evolved over the centuries, so too have observatories.

Up to the seventeenth century, observatories were usually privately owned. Rulers, wealthy land-owners and members of the aristocracy sponsored both the instruments and the astronomers who operated them. The first observatories were built where the sponsors had land that was available. For example, Castle of the Heavens. TYCHO BRAHE was sponsored by King Frederick II of Denmark, who built him a castle called Uraniborg (*Star-city*) on the island of Hveen (now Ven).

In their second phase of development, observatories were a mixture of private facilities, perhaps owned and operated by a wealthy amateur, and public organizations, owned and funded by universities (too many to mention), cities and regional governments (for example, Florence, Palermo, ...), churches or other religious foundations (for example, the ARMAGH OBSERVATORY, founded by the Archbishop of Ireland), and scientific societies (for example, the Portuguese Royal Academy of Sciences).

In recent times, the costs of big telescopes being so large, major observatories have been funded by government agencies (the Russian Academy of Sciences, the French Institut National des Sciences de l'Univers, the National Research Council of Canada, ...), or by associations of public bodies (such as AURA, the ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY, operator of the KITT PEAK NATIONAL OBSERVATORY and the CERRO TOLOLO INTERAMERICAN OBSERVATORY for American university astronomers). Private finance is still a factor, however, as in the W M KECK OBSERVATORY (capital finance from the Keck private foundation and running costs through the California Institute of Technology and the grants that it obtains) and the SETI INSTITUTE (privately financed).

The major observatories of the current generation and the future are and will be financed and operated by international consortia. Sometimes this is obvious in their names (like the ANGLO-AUSTRALIAN OBSERVATORY, the CANADA–FRANCE–HAWAII TELESCOPE, and the EUROPEAN SOUTHERN OBSERVATORY), sometimes not (like the ISAAC NEWTON GROUP OF TELESCOPES on La Palma, the GEMINI OBSERVATORY, or the ATACAMA LARGE MILLIMETER ARRAY, ALMA).

The transition to state-operated observatories began in the seventeenth century. In Britain, King Charles II approved the foundation of the ROYAL OBSERVATORY AT GREENWICH in 1675 in order to produce a new star catalog that would assist British seafarers. Since the Royal Navy paid for it, the observatory was located on a hill overlooking the busy wharves along the River Thames and within an easy carriage drive of the royal palace and Admiralty buildings in London. Because of its location in a royal park, the observatory's architect, SIR CHRISTOPHER WREN, acknowledged that he designed it 'a little for pompe'.

Although the importance of dark skies was generally recognized, astronomers had to accept poor weather as a price for accessibility in the days when transportation was not as easy as now. Thus, two optical telescopes that held the title of 'world's largest' prior to the twentieth century were built in locations that today would be considered completely inappropriate for world-class observatories. In 1789, funded personally by King George III, William Herschel (see HERSCHEL FAMILY) placed his 40 ft (focal length), 48 in (aperture) reflector in Slough, a low-lying rural area near Windsor next to the River Thames. Lord ROSSE himself financed his 72 in 'Leviathan of Parsonstown', which was built in 1845 in the grounds of his home, Birr Castle in Ireland, a site surrounded by numerous peat bogs.

As time went by, urban expansion, increasing air pollution from smoke and light pollution meant that observatories located near cities were increasingly at a disadvantage. Thus, the VATICAN OBSERVATORY, which was formally founded by Pope Leo XIII in 1891, was originally located in the Vatican gardens behind St. Peter's Basilica. As light pollution increased, it was moved out of Rome to the Papal summer residence at Castel Gandolfo. By the 1970s, light pollution from the nearby city had become so bad that it was decided to relocate the observatory to Arizona.

A similar process occurred in the case of the Royal Observatory at Greenwich. In the 1950s, the observatory's main telescopes were removed from the growing, smog-ridden metropolis of London to the Sussex countryside at Herstmonceux. This site, whether or not it was among the best possible sites in the UK, suffered from the damp English climate with its frequent cloud cover and poor seeing conditions, and the population growth of the south coast area increased sky background. As a result, the main instrument, the 2.5 m (98 in) Isaac Newton Telescope, was relocated to La Palma in the Canary Islands in 1990.

Much the same happened with the Paris Observatory (OBSERVATOIRE DE PARIS) that was commissioned by King Louis XIV in 1667. Although accessible from the King's main place of residence it was originally built so far from the center of Paris that it never grew, as planned, into the cultural center of the then Royal Academy of Sciences. It is now surrounded by the city. The observatory established a new campus south of the city, at Meudon, for observing facilities, but French astronomers' attention has shifted since to more remote telescopes in ESO and on Hawaii. As one further example, the national operation of the Helsinki Observatory in Finland was also moved from its hill-top site, in land originally outside the city, first to Turku and then to Tuorla in the countryside. Finnish astronomers' attention is now focused on the NORDIC OPTICAL TELESCOPE on La Palma and towards ESO.

Radio observatories have likewise been established, first in fields near to universities and research institutes, and then in radio-quiet zones, like GREEN BANK, West Virginia.

Nevertheless, the vast majority of observatories remain located with accessibility in mind, near the sponsoring body; if a telescope is intended for use by students or by the public, then accessibility is the over-riding concern. This means that most observatories for general use are in or near cities.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mountain sites and city lights

In a famous passage in *Opticks* (1730), ISAAC NEWTON wrote that telescopes 'cannot be so formed as to take away that confusion of rays which arises from the tremors of the atmosphere. The only remedy is a most serene and quiet air, such as may perhaps be found on the tops of the highest mountains above the grosser clouds'. Modern technology has, in recent times, proved Newton wrong in the first statement but his second statement was proved correct in 1856. The Astronomer Royal for Scotland, CHARLES PIAZZI SMYTH, mounted an expedition to the mountain of El Teide on the island of Tenerife in the Canary Islands. Re-erecting the 18 cm telescope with which he had made calibrating observations in Edinburgh, approximately at sea level, he found that he could see stars that were four magnitudes fainter through the more transparent mountain air at altitudes of 2700–3200 meters. He could also distinguish closer double stars, because the seeing was better.

The first large observatory built on a mountain summit according to Newton's conjecture and Smyth's experiments was LICK OBSERVATORY, opened on Mount Hamilton in northern California in 1888 at an altitude of 1280 meters. There one of the world's largest refracting telescopes was built (the 36 in refractor), funded by a wealthy private donor, James Lick.

Over the following decades, many leading mountain observatories sprang up in the dry, clear air of the American west. In California, the world's largest (at the time) reflecting telescopes were installed. The 100 in Hooker Reflector was built in 1917 at MOUNT WILSON OBSERVATORY near Pasadena and the 200 in Hale Reflector in 1948 at Mount PALOMAR OBSERVATORY near San Diego. An almost permanent temperature inversion below the mountain tops more or less guaranteed low atmospheric turbulence and a cloudless sky. Even with the increase in urban sky glow as a result of the development of California since those times, these sites remain competitive in terms of 'SEEING' and image quality (table 1). Mount Wilson, for example, is the host site of CHARA, an optical interferometer, to which, since it works with bright stars, artificial sky glow matters little.

Table 1. Loss of effective aperture and value due to urban sky glow. In the table that follows, *X* is the sky glow level. A value of 1.0 designates the natural sky background, without any human contribution; a value of 1.2 means a 20% increase above the natural sky background due to a man-made skyglow; 2.0 is double the natural background, and so on.

X	Equivalent aperture in meters	Equivalent aperture in inches	Per cent of original value
For a 4 m aperture telescope			
1.00	4.00	157	100
1.10	3.81	150	88
1.20	3.65	144	78
1.25	3.58	141	74
1.50	3.27	129	58
2.00	2.83	111	39
3.00	2.31	91	23
5.00	1.79	70	11
For a 4 m telescope on Kitt Peak			
1.06	3.89	152	93
For a 5 m (200 in) telescope on Mount Palomar			
2.00	3.58	141	39
For a 3 m (120 in) telescope on Lick Observatory			
3.00	1.75	69	22
For the 2.5 m (100 in) at Mt. Wilson Observatory			
5.00	1.14	45	11

(Copyright the International Dark Sky Association, December 1996.)

Desirable physical characteristics for an observatory

All observatories need road access for the installation, maintenance and operation of the telescopes and instrumentation, and a level-enough site to build on, with appropriate foundation material. Power from transmission lines is desirable, but can be supplied by local generators, which in any case are needed as back-up if power fails over the sometimes long and vulnerable lines. Telecommunication facilities are necessary, and in these days of high-speed data links optical fiber connectivity or telecommunication satellite bandwidth is becoming the norm. In the installation of such facilities, care has to be taken not to degrade the electromagnetic environment of the observatory—small currents can be induced through pickup in the detector electronics and amplified, with interference to the observations (see also TELESCOPE ENGINEERING).

For an optical-infrared observatory, chosen for lack of cloud, water run-off from buildings is not likely to suffice for the operation of the facility, or for use by the staff, and must be pumped from artesian supplies or brought in. Remoteness from population sources minimizes artificial sky background, industrial smoke or dust generation, and background radio noise from electrical machinery, so it is not likely there will be abundant water, since water would probably have already attracted settlements.

Prime requirements for an optical observatory are good atmospheric transparency and seeing characteristics. For an infrared or millimeter wave observatory, transparency is linked to (lack of) water vapor content in the air, measured by precipitable water vapor, and carbon dioxide absorption.

There will undoubtedly be a winter–summer cycle in all these weather-related characteristics, but the less seasonal the variation, the more all-sky access there will be. If winters are severe, ice or snow may also affect building costs and access. To establish the weather characteristics of candidate sites for a major telescope, a period of site testing is necessary, and has to be at least a year long to establish the annual cycle. Moreover, since there are year to year differences across a region or even a hemisphere, similar testing should be carried out at a reference site whose long-term characteristics are known, to establish the differences between the new site and the established one.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

As Newton conjectured, high altitude sites have the advantage of being above the inversion layer, which restricts aerosol content to the Earth's surface layer. The inversion layer can be between 2 km and 10 km in altitude, depending on ground or sea heating.

Local effects on seeing are minimized if the mountain presents a streamlined contour to the prevailing wind, so that air heated by the ground layer does not mix with colder, higher air. The thickness of the skin of air above the land surface in which there is local heating and turbulence drives the decision about the height at which the telescope is placed above the ground on a pillar, and is a cost-driver through the building costs. The prize of locating the telescope above the ground layer is that the seeing created by the air above this skin could be as small as 0.3 arcsec.

Merle Walker (1983) identified a group of good observatory sites on mountain tops to the east (this is usually the down-wind direction) of cold oceanic currents, which stabilize air flow, reduce the scale height of convection and therefore keep the inversion layer low (at say 2 km). California, the Canary Islands, and Chile all fit this pattern; Walker also identified Namibia in southern Africa as a relatively under-exploited astronomically desirable country.

Today, the vast majority of the world's largest, most powerful, optical-infrared telescopes are to be found in a handful of high-altitude sites. There are concentrations of such sites in the western half of the continental United States, in the fiftieth state, Hawaii, in the arid Andean mountains of Chile, and in the Canary Islands. The local environmental factors outlined above are the driving force behind these concentrations. As remote observation and telescope control become increasingly common, while the jet plane makes travel quick and easy, accessibility for the astronomers who use these instruments and the engineering staff who build and maintain them drops down the list of priorities.

Large regions of the world are downgraded as possible observational sites through their geographical location. Much of eastern North America, continental Europe, southeast Asia and the tropical regions of Africa and South America suffer not only from proximity to man-made pollution but from higher humidity, a higher percentage of cloudy days and poorer atmospheric transparency caused by water droplets, dust, smoke or other aerosols. Nevertheless, for the reasons already stated, most observatories, except the very largest, are located in these regions, near to the population.

Of course, there are penalties associated with mountain top locations. Their geographical remoteness inevitably means that road and air links are probably limited or non-existent and these may have to be provided before construction can begin. Part of the costs of operating the observatory will consist of access maintenance. Access to the PIC DU MIDI OBSERVATORY in the French Pyrenees is by cable car. The IRAM millimeter-wave observatory at Plateau du Bure near Grenoble is also accessed by cable car, whose collapse in 1999 caused loss of life and suspension of operations.

Once the site has been selected, its rocky, sloping surface then has to be flattened and prepared for the foundations for the observatory building and its subsidiary facilities. Apart from the difficulty of building in remote conditions, staff and visiting astronomers have to survive low temperatures, strong winds, occasional blizzards and altitude sickness caused by the thin air. Acclimatization is important for visitors unfamiliar with the oxygen deprivation. On Mauna Kea, where the telescopes are at 4200 m, a 'halfway house' has been set up at 3800 m. Here a visitor may acclimatize before heading for the summit, an astronomer may sleep during the daytime after his or her night-time observations, and engineering staff may think about problems encountered at the summit but unsolved in an oxygen-poor, headache-inducing atmosphere.

Earth tremors pose a threat to any facilities located nearby. One notable example was the earthquake which struck the BIG BEAR SOLAR OBSERVATORY in June 1992. The tremor, which measured 6.6 on the Richter scale, occurred only 9 km from the observatory. The causeway which links the observatory with the lake shore suffered from subsidence and fractures, while the support pedestal for the telescopes and the tracking machinery were also damaged. The observatory was out of action for more than four months while repairs were undertaken. Telescopes on the volcanoes of Mauna Kea and La Palma are protected against earthquakes of a size that is thought possible there, with spring-loaded pivots pulling gears apart in the case of tremors that could damage the main bearings.

Observatories in Australia and La Palma have been uncomfortably close to bush fires, which threaten buildings, and drift ash onto delicate optical surfaces.

Astronomical observatories are also affected by air pollution, vibrations and thermal pollution. For example, studies by the US NAVAL OBSERVATORY in Washington, DC have shown that thermal turbulence and refractive index asymmetries over the site created by large heated buildings nearby are compromising the astrometric value of the site. Even aircraft lights and contrails can cause problems for observatories located beneath flight paths. Civilian air traffic has been re-routed away from La Palma and use of the take-off and approach path to Nice airport, which passes over the Observatoire de Calern, is not permitted at night for this reason.

Environmental and ecological considerations may also play a part in influencing location. Many observatories are located in remote regions that are noted for their scenic beauty and rare fauna and flora. Some, such as Australia's Siding Spring Observatory, are built close to a national park.

Conflicts may sometimes arise with local residents or environmental groups. When the University of Arizona and its international partners selected Mount Graham as the site for a series of world class instruments, including the LARGE BINOCULAR TELESCOPE, the 3200 m mountain seemed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

an ideal location, not only for its excellent 'seeing' but also because it already had a two-lane access road maintained for summer residences and rural industries such as logging. After a lengthy environmental impact assessment, permission to go ahead with the first stage of construction was duly granted by the US Fish and Wildlife Service. Unfortunately, the astronomers then became embroiled in a series of costly legal arguments centered around the fact that Mount Graham is a 'sky island', an isolated pocket of rare fauna and flora. Environmentalists argued that the unique ecosystem, which included an endangered subspecies of the red squirrel, would be irrevocably damaged by the new observatory, even though the initial area of construction was limited to 3.5 hectares (8.6 acres). Even after two bills of Congress were passed in order to allow the Large Binocular Telescope to go ahead, further appeals and protest meetings caused the project to be badly delayed and its costs to escalate dramatically.

Environmental issues and the religious concerns of the Hawaiian native people to whom the mountains are sacred areas are factors that limit the development of Mauna Kea and Haleakala in Hawaii.

Socio-economic conditions

Economic development and political stability also play a significant part in influencing decisions about where to locate modern astronomical observatories. The governments of developed countries, sometimes assisted by private funds, seek to maintain astronomical facilities at the cutting edge of cosmological research by placing them in the most advantageous locations. These can be anywhere in the world where the environmental requirements are met.

The underdeveloped infrastructure that is likely at an observatory location means that development of an observatory has to be undertaken from scratch, entailing a much longer, more expensive process of preparation and construction. The intangible asset of confidence is engendered by an economically solvent, politically stable and astronomically supportive regime in the host country, providing a basis for investment of large sums by the external partners. The establishment of a major astronomical center can help to improve economic conditions in the host country, by providing jobs, by improving levels of education and technical training, and by providing an aspirational example for future development. As a 'site fee' the host nation usually receives a proportion of the observatory's observing time, providing guaranteed access for its scientists to world class observational facilities at relatively little cost.

The southern hemisphere

The latitude of an observatory is important, with low latitudes generally favored. The natural sky glow decreases away from the polar regions (but this is a disadvantage if geomagnetic phenomena are the subject of investigation), the nights are of more uniform length near the equator, and a larger fraction of the sky passes above the horizon. The southern hemisphere sky contains unique astronomical phenomena, such as the galactic center and the Magellanic Clouds. Such objects are available for longer periods when they pass overhead, and for this reason, for multiwavelength studies driven by satellite telescopes that orbit the earth and see both hemispheres equally, and to provide a complete picture of the Galaxy and the universe, astronomers have required southern observatories. The need to improve coverage of the southern hemisphere skies was one of the driving forces behind new observatories in the twentieth century. In the early years of the twentieth century, the main observational centers south of the equator were the Radcliffe Observatory at Pretoria and Mount Stromlo near Canberra. The 1.9 m (74 in) reflectors that were built in Pretoria and in Canberra remained the largest telescopes in the southern hemisphere for 20 years.

As light pollution from the nearby conurbation of Australia's capital city increased, the Australian National University (ANU), which became responsible for Mount Stromlo Observatory, established a new facility at Siding Spring (MOUNT STROMLO AND SIDING SPRING OBSERVATORIES) near the town of Coonabarabran in New South Wales. In addition to its good local infrastructure and easy access, its observing conditions were among the best that could be found in Australia. The UK government joined Australia to add to the ANU facilities at Siding Spring, including the UK Schmidt telescope and the 3.9 m Anglo-Australian Telescope, in the 1970s.

Today, the Siding Spring site is at a disadvantage when compared with other newer southern locations. First, it is only 1000 m (3200 ft) above sea level. As a result, atmospheric turbulence is significant, while observations on cloud-free nights are only possible for about 60% of the year. Light pollution is also beginning to encroach from conurbations far to the south. At a time when funding for astronomical facilities is increasingly tight, these factors may militate against further development or even continued support for such middle-range facilities in the future.

This experience compares and contrasts with the fortunes of South Africa. During the early 1970s, the British Science Research Council joined with its South African counterpart, the Council for Scientific and Industrial Research, to set up a modern observatory in the semi-desert region of the Karoo, near Sutherland, northeast of Cape Town. The Radcliffe Telescope was relocated from Pretoria and several smaller telescopes were added—new and relocated from the Cape Observatory. Despite its relatively low altitude (1760 m), this site has now been selected by an international consortium for a new 9 m class instrument (SALT, the SOUTHERN AFRICAN LARGE TELESCOPE). South Africa looks to enjoy an astronomical revival in the next decade, once more operating the largest telescope in the southern hemisphere, in competition with the VLT, whose individual telescopes will be slightly smaller but which together lay joint or greater claim to that title.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Island sites

Because mountain tops are favored for large astronomical telescopes, and building on the development work carried out by the INSTITUTO DE ASTROFISICA DE CANARIAS, the UK prospected La Palma in the Canary Islands for its 'northern hemisphere observatory'. The La Palma site became available in 1979 after Britain and Spain, to which the Canary Islands belong, signed an agreement to establish the OBSERVATORIO DEL ROQUE DE LOS MUCHACHOS on the rocky volcanic island. Site protection was guaranteed in Spanish law. British astronomers selected the site for the relocation of the Isaac Newton Telescope, then western Europe's largest optical telescope, after examining other promising locations in Tenerife, Madeira and Italy. The 4.2 m William Herschel Telescope followed in 1989, funded by the UK in partnership with The Netherlands.

The summit of La Palma is 2500 m above sea level and is home to half a dozen major astronomical instruments that are owned and operated by various European countries. La Palma is the site for the Gran Telescopio de Canarias, virtually a clone of the Keck telescope and the nucleus of the 'European northern observatory'.

However, despite its advantages, La Palma has been overtaken by Hawaii as the prime choice for today's 8–10 m class telescopes in the northern hemisphere. Indeed, the huge extinct volcano of Mauna Kea in Hawaii lays claim to be the best observatory site in the world. At an altitude of 4200 m it is also by far the highest of the major observatory sites.

This extreme altitude is important because it opens the way to observations at optical, infrared and sub-millimeter wavelengths. The observatory lies above some 90% of the atmospheric water vapor, and almost all of the remainder is removed from the air as frost or snow, while the trade winds which blow across the Pacific are forced around the mountain, leaving the summit in a pocket of exceptionally stable air. Since the island is far from any large land mass, the air is also free of dust. It is unaffected by light pollution, and there are strong zoning laws controlling lights in the coastal towns.

On the other hand, as its name (Mauna Kea means 'White Mountain') suggests, weather conditions on the summit can sometimes be severe, with heavy snowfall and 150 km h^{-1} winds which can last for days. Oxygen levels are 60% those at sea level, and oxygen deprivation affects visitors' ability to breathe, move around and make rapid decisions, as well as the cooling capacity of fans, the stability of computer disk drives and the power of diesel engines. Remote control facilities and remote diagnostic capability are emphasized for the Mauna Kea telescopes. The 10 m Keck telescopes on Mauna Kea are operated from a base at the foot of the volcano at Waimea, the James Clerk Maxwell Telescope likewise.

Sub-millimeter and radioastronomy

In contrast with optical observations, radioastronomy has a relatively short history. However, it follows a similar pattern. From the earliest experiments, radio interference from terrestrial sources has been an obstacle to studies of the universe.

One of the earliest examples of this dilemma came in December 1945, when BERNARD LOVELL borrowed three trailers of radio equipment from the British army for cosmic ray studies. In order to avoid electrical interference on the main campus, Lovell moved his equipment to the rural setting of the University of Manchester's botany department at Jodrell Bank. In 1957 this became the site of the world-famous 76 m fully steerable antenna, later known as the Lovell Telescope, and is now the home of the NUFFIELD RADIO ASTRONOMY LABORATORIES, and the center of the MERLIN interferometer complex distributed across the UK.

Unfortunately, facilities such as Jodrell Bank are now immersed in an electromagnetic jungle composed of signals from innumerable sources, including everything from TV transmitters to electric motors and CD players. Inevitably, modern radioastronomy observatories, such as the Goldstone and OWENS VALLEY RADIO OBSERVATORIES in California, and the Australia Telescope are built in radio-quiet areas, far from large population centers and often shielded by hills.

Sometimes, the sites are able to benefit from local legislation that prevents interference. For example, the new Green Bank Telescope, the largest fully steerable dish in the world, lies in a 150 km wide National Radio Quiet Zone in West Virginia. This means that the observatory director is consulted over the installation of any new radio transmitters within this zone or any significant modification to existing transmitters. The radio telescopes of the MULLARD RADIO ASTRONOMY OBSERVATORY in Cambridge were protected from radio noise from car-ignition systems on the nearby motorway by sinking the road and erecting a metallic fence that acts as an electromagnetic shield.

Today, interferometry is a vital technique for radioastronomers, and a number of multi-telescope systems have been created. These occupy large areas, so remote desert sites where flat land is cheap and plentiful are ideal. The largest of these observatories is the VERY LARGE ARRAY in New Mexico, where 27 separate dishes are spread in a Y-shape along three arms, each of which is 21 km long.

While sheltered valleys are popular sites, local topography can sometimes be utilized in unusual ways. The most famous example of this is the ARECIBO OBSERVATORY, the largest radio telescope in the world, which lies within a natural rock basin in the limestone hills of Puerto Rico. Since the antenna is not intended to be steerable, the giant depression provided an ideal site for the construction of the 305 m diameter dish. China has plans to construct a similar dish, FAST, the Five-hundred meter Aperture Spherical Telescope, in a region of southern China with many hemispherical karst depressions in which the telescope could be mounted. This project might be helped to fruition if a radio-quiet zone could be implemented in the area.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Physical factors are relatively insignificant for long-wavelength radio observations. At sea level, observations at wavelengths of 2 cm or more are possible without serious interference, but at shorter wavelengths, it is necessary to observe from high-altitude sites where absorption by atmospheric water vapor is minimized. In particular, sub-millimeter observations benefit greatly from such high-altitude locations. Not surprisingly, the world's largest sub-millimeter telescope, the JAMES CLERK MAXWELL TELESCOPE, is located on the summit of Mauna Kea.

The VLT at Paranal—a case study

The European Southern Observatory's Very Large Telescope (VLT) site at Cerro Paranal in Chile, located 2650 m above sea level, exemplifies many of the advantages and problems involved in the construction and operation of a modern mountain-top observatory. Its physical advantages are fairly obvious. Its altitude and its low rainfall (a storm every decade) result in up to 350 clear nights a year, unusually stable air which allows a spatial resolution of 0.45 arcsec for 15% of the viewing time and low water vapor concentrations which increase atmospheric transparency at infrared wavelengths.

On the other hand, these advantages were bought by major investments of time and money. Seven years of extensive site testing were undertaken before preparation work at Paranal began in 1991. Once a road was built from a nearby highway to the summit, some 350 000 cubic meters of rock and soil were removed from the site, flattening the mountain to create the 20 000 square meter platform on which the four 8.2 m telescopes and the interferometric complex would be built.

Before construction began, political factors had to be resolved between the European Southern Observatory (ESO) and the government of Chile. Following the signature of a treaty between ESO and the Chilean government in 1963, 15 world-class telescopes had been located at Cerro La Silla in the arid mountains of the Atacama Desert. However, when ESO requested permission to build the VLT at Cerro Paranal, in 1991, the government of Chile asked for modification of the original Convention.

A protracted period of negotiations ensued, concerning ownership of the Paranal site, labor relations, guaranteed observing time on the ESO telescopes for Chilean astronomers, and the rules governing the installation of additional telescopes in the future. These difficult discussions, spread over several years, inevitably delayed completion of the Paranal site, increased costs, and prompted ESO management to pursue studies of other first-class astronomical sites elsewhere in the world. The difficult issues were eventually resolved, the VLT is now in operation at Paranal and it is likely that the Atacama Large Millimeter Array will be built nearby.

Prior to construction of the main accommodation and office complex at Paranal, which began in 1999, ESO staff and visitors were housed in containers at a 'Base Camp', while the workers and engineers of the contracting firms were housed in similar quarters along the mountain road.

Since the nearest town, Antofagasta, is 120 km away, and Paranal lies in the middle of the Atacama Desert, everything from food and water to telescope spare parts, is brought in by road. Several diesel generators are used to maintain a stable electricity supply for the sensitive astronomical instruments and to ensure a minimal level of comfort for staff and visitors. A hotel and observatory maintenance facilities, including the advanced aluminizing plant for the giant VLT mirrors, are situated at the foot of the mountain.

In order to squeeze as much performance as possible out of such substantial investments, every effort is made to reduce the impact of the physical environment on the observational efficiency. At Paranal, for example, the laboratories, control building and most of the office/accommodation block are buried underground, in order to minimize disturbances to the wind flow around the mountain summit.

Future sites

Three major threats to future research have been recognized by the astronomical community:

- Growing levels of radio pollution, interference at radio frequencies from telecommunications satellites and their ever-increasing demand for new wavebands cloud the future of radio astronomy.
- Space debris is a growing threat to scientific satellites and also interferes with ground-based observations as satellites or space junk leave streaks across sky images.
- Observatories around the world are also in danger from the proliferation of outdoor lighting, increased urbanization, and projects to launch highly luminous objects into space for various purposes such as Earth illumination or artistic, celebratory and advertising goals (see table 1).

The only undeveloped site on Earth known to rival Mauna Kea for infrared and sub-millimeter observations, and for freedom from radio noise and urban sky glow is Antarctica. Stratospheric air is pulled down onto the South Pole by convection upwards at latitudes closer to the equator and the cold, dry and stable conditions on the high-level Antarctic Plateau allow small telescopes at the South Pole to outperform much larger telescopes at temperate sites. Difficult access and the extreme climate have discouraged construction and observations, but there are astronomical observational facilities at the American CENTER FOR ASTROPHYSICAL RESEARCH IN ANTARCTICA (CARA), so far all quite modest in size and capability.

As even near-Earth space becomes polluted with space debris, some astronomers have begun to look towards future observatories on the far side of the Moon. Problems with cloud cover, atmospheric turbulence and absorption of incoming radiation by water vapor are non-existent on the airless Moon. The absence of an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

atmosphere means that every type of radiation can stream down to the surface without being filtered by a blanket of air. Optical and infrared interferometry would also be much easier in the lunar environment. Whereas terrestrial optical and infrared telescopes can only operate at night, lunar telescopes would be able to operate around the clock. With the Moon's slow rotation, it would be possible to observe faint objects non-stop during the two-week-long nights. And in a one-sixth gravity environment, much larger, lighter mirrors or reflector dishes could be installed without being deformed.

The difficulties in establishing a telescope on the Moon are not only the obvious ones of doing anything in space, but also the dusty environment (bad news for mechanical and optical devices). A study by the EUROPEAN SPACE AGENCY concluded that the theoretical advantages of an optical-infrared or millemeter wave observatory on the Moon were outweighed by the disadvantages. Free-flying observatories like the Hubble Space Telescope (HST), the Next Generation Space Telescope (NGST) and the Far Infra Red Space Telescope (FIRST) were the way to go.

The lunar far side would, however, be particularly favorable for radio astronomy. Shielded by the body of the Moon from terrestrial signals, the telescopes would have a clear view of the radio universe. Lunar craters would make ideal sites for giant Arecibo-style radio antennae. There would even be the opportunity to explore an entirely new region of the electromagnetic spectrum. Very-low-frequency radio waves (below about 15 MHz), which are absorbed or scattered by the Earth's atmosphere, would be readily accessible to telescopes on the Moon. Moreover, since the body of the Moon screens radio telescopes from artificial signals engendered on the Earth, this might be the place from which artificial signals from other civilizations can be identified most readily. To protect the operational electromagnetic environment of the Moon for these purposes, astronomers have already established norms for radio communication between lunar colonies.

Peter Bond and Paul Murdin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatories: History

An astronomical OBSERVATORY is a building, installation or institution dedicated to the systematic and regular observation of celestial objects for the purpose of understanding their physical nature, or for purposes of time reckoning and keeping the calendar. At a *bona fide* observatory such work constitutes a main activity, not just an incidental one. While the ancient Egyptians, Babylonians, Chinese, Greeks and Mayans made many observations of scientific and historical interest, their efforts were not carried out at observatories *per se*. In this article we shall concentrate on installations that were intended to be *bona fide* research centers.

Observatories in Islam

There was a strong motivation in medieval Islam to preserve the knowledge of the Greeks and if possible to add to it. In a large city the presence of a hospital, a university and a public library was considered sensible and desirable. Within this context, in which knowledge and piety were highly encouraged and intertwined, the tradition of observatory building began during the reign of al-Ma'mūn, the seventh Abbasid caliph (AD 813–833). He organized a scientific academy in Baghdad called the House of Wisdom, which included a library and observatory. This was the most significant scientific endeavor since the establishment of the Alexandrian Museum in the 3rd century BC. A second observatory was built on the plains of Tadmor in Syria. al-Ma'mūn's astronomers measured the obliquity of the ecliptic (obtaining a value of $23^\circ 33'$), concluded that the precession of the equinoxes amounted to 54 arcsec yr^{-1}, determined that the Earth was 6500 miles in diameter and produced tables of motion of the planets based on PTOLEMY's work, but updated on the basis of new observations. Astronomical work continued in Baghdad under the ninth Abassid caliph, but then ceased (see also ISLAMIC ASTRONOMY).

A second scientific academy, the Hall of Wisdom, was established in Cairo in 1005. It lasted until 1171. An observatory was begun there in 1120 and nearly completed, but was ordered destroyed by the caliph in 1125 after the death of the vizier in charge of the observatory and its instruments. Apparently, serious conflicts had arisen between this vizier and the caliph, and it did not help that the superstitious populace presumed that trying to understand the motions of the planets was akin to delving into magic.

The Istanbul Observatory, completed in 1577, met a similar fate. Political back-stabbing and superstitions regarding astronomy led to that observatory's destruction by 1580.

The two most successful and extensive Islamic observatories were those at Maragha (in Azerbaijan, in modern-day Iran) and at Samarkand, in modern-day Uzbekistan. The Maragha Observatory was built under the direction of NASIR AL-DIN AL-TUSI (1201–1274). It was situated on the flattened top of a hill, covering 400 m by 150 m. Completed in 1264, it was active at least until 1304, perhaps until 1316. It was the first observatory whose activity did not greatly diminish after the death of its founder. While the motivation for the construction of the observatory was *astrological*, the prediction of future celestial events needed to be carried out on the basis of exact physical measurements. A significant number of instruments were built, among them equinoctial and solsticial armillary spheres, and a mural quadrant of radius 4.3 m. The observatory's library at one time contained 400 000 volumes. The principal accomplishment at Maragha was the compilation of the Ilkhanic Tables, which were tables of motion of the Moon, Sun and planets. These identified some of the major shortcomings of Ptolemaic astronomy.

The most important astronomer of the 15th century was ULUGH BEG (1394–1449). For most of Ulugh Beg's life his father was the ruler of Transoxiana, a region situated between the River Oxus (Amur Darya) and the River Jaxartes (Syr Darya). This provided the son with the opportunity to indulge his passion for astronomy. As a young man Ulugh Beg visited the remains of the Maragha Observatory, and he endeavored to build an even greater institution in Samarkand. As many as 70 astronomers were active there between 1408 and 1437. The largest instrument constructed was a sextant of radius 40 m, a fixed instrument mounted on the north–south meridian which could achieve a resolution of several arc seconds, a value not to be exceeded until the invention of the telescope and the micrometer in the 17th century. This sextant was primarily used for observations of the Sun.

On the basis of new observations, Ulugh Beg and his fellow astronomers compiled a catalog of the brightest 1000 stars visible at the latitude of Samarkand. No one since Ptolemy (ca 150 AD) or perhaps as far back as Hipparchus had made such extensive stellar observations. Ulugh Beg's catalog was most likely based on observations with a zodiacal armillary sphere graduated to 15 arcmin, with interpolation to two-tenths of a mark, or 3 arcmin. The typical positional accuracy in the catalog is ± 16 arcmin.

China and India

At the very end of the 17th century one Louis Lecomte published his *Memoirs and Observations of a Journey through China*. He noted in some detail the vigilant activity carried out by astronomers at the Imperial Observatory in Beijing: 'Five mathematicians spend every night on the tower in watching what passes overhead; one is gazing towards the zenith, another to the east, a third to the west, the fourth turns his eyes southwards, and a fifth northwards, that nothing of what happens in the four corners of the world may escape their diligent observation'. Such activity had been going on for three millennia. As a result, the Chinese had accumulated records pertaining to many hundreds of lunar and solar eclipses, observations of novae, supernovae, comets, meteor showers, aurorae,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

naked-eye sunspots and even a possible observation of Jupiter's moon Ganymede in 364 BC, nearly two millennia before the invention of the telescope (see also CHINESE ASTRONOMY IN HISTORY). The first clock drive was built by the Chinese in AD 132; it was powered by a constant pressure-head of water in a clepsydra, or water clock. The Chinese astronomer Yü Xi independently discovered the precession of the equinoxes in AD 320, finding a value of about 72 arcsec yr^{-1}, twice the value found by Hipparchus. Sunspot records in the Chinese annals demonstrate the 11 yr solar cycle.

Just as spices were transported along the caravan routes, so too was astronomical knowledge. Astronomers from Persia arrived in China in AD 719. The astronomical ties between the mid-East and far East became quite extensive during the Yüan Dynasty (1271–1368). In 1267, only 8 yr after the founding of the Maragha observatory, blueprints for seven instruments were sent to China. Actual models may have followed.

In 1270 the astronomer Guo Shoujing (1231–1316) built the first equatorially mounted instrument. After the Beijing Observatory was re-equipped in 1276–1279 under his direction, it was equal in stature to the Maragha Observatory. The Chinese also operated an observatory at Nanjing.

The Chinese did not appreciate or understand the Greek geometrical models of planetary motions or the Arabic use of geometry, particularly spherical trigonometry and stereographic projection. However, 300 yr before Tycho Brahe became convinced that right ascension and declination were the 'coordinates of the future', the Chinese were fully committed to their use. Just as the influence of the Emperor radiated in all directions, so the hour circles radiated from the pole, 'like the spars of an umbrella'. The Chinese laid out a system of 28 lunar mansions (*xiu*), which were defined by the points at which these hour circles intersected the celestial equator.

One of the most important events in the history of Chinese science was the arrival of the Jesuits in 1600, towards the end of the Ming Dynasty. They had a very specific motivation. By demonstrating the superiority of Western science, they hoped to convince the Chinese of the superiority of Western religion. Western science clearly won out over the traditional Chinese and Muslim methods of predicting the solar eclipses of 15 December 1610 and 21 June 1629. In November 1629 a new calendar bureau was established under the direction of the Chinese Christian convert Xu Guangqi, who supervised the work of 50 astronomers, many of whom converted to Christianity. However, in 1664, at the start of the Qing Dynasty, politics led to the dismissal of the Jesuits and most of the Chinese converts from the calendar bureau. Some of these astronomers were later beheaded.

The Jesuits regained favor in Chinese astronomy in 1669. From 1673 to 1676 the Beijing Observatory was re-equipped with a new set of Tychonic instruments, which are still in place today. However, the astronomical activities in China were hindered by several factors, notably the limited interest of the Chinese in questions other than calendrical revision, the Jesuits' observance of the Church's prohibition of the discussion of heliocentrism, which was in effect until 1757, and their attraction for the hybrid geo-heliocentric model of Tycho Brahe.

In India the most significant observatory construction was carried out under the direction of Jai Singh (1686-1743), a Hindu prince in the court of a Muslim Mogul emperor. Large instruments of masonry were constructed at Delhi, Jaipur, Ujjain, Benares, and Mathura. The largest instrument was a sundial 27 meters tall. Jai Singh wished to follow in the footsteps of Ulugh Beg, minimizing observational errors by using the largest instruments possible. His star catalog was an update of Ulugh Beg's star catalog, but apparently with no new observations: $4^\circ 8'$ were simply added to Ulugh Beg's ecliptic longitudes to account for precession over 288 years (see also INDIAN ASTRONOMY).

Early European observatories

Astronomical observations were carried out in Moslem Spain at Cordova in the 10th century, at Toledo in the 11th, and at Castille in the 13th century under the patronage of the Christian King ALFONSO X. The Toledan astronomical tables were translated into Latin in the 12th century. The Alfonsine Tables, the first state-sponsored astronomical ephemerides published for general use, were reprinted as late as 1641.

The first European observatory worthy of the name was built by Bernard Walther (1430–1504), a wealthy private citizen of Nuremberg, who was at the same time the pupil and patron of REGIOMONTANUS (1436–1476). Together they found that the positions of the planets differed to a significant degree from the predictions of the Alfonsine Tables. They printed astronomical treatises on a printing press set up in Walther's house. One such treatise by Regiomontanus laid out the 'method of lunar distances' for determining longitude at sea. A significant innovation at Walther's observatory was determination of the times of observations by mechanical clocks instead of by astrolabes or armillary spheres. Observations with Walther's instruments exhibited an improved accuracy, to ± 10 arcmin.

Wilhelm IV (1532–1592), the Landgrave of Hesse-Cassel, began systematic observations in 1561. In 1567 his father, the Landgrave of Hesse, died, and the landgraviate was divided up amongst the four sons. Wilhelm was compelled to take over the administration of his province and had to lay aside his astronomical endeavors. In 1575, however, he was visited by Tycho Brahe and was inspired to resume observing.

Wilhelm constructed a number of metal instruments, among them an azimuthal quadrant of radius 0.4 m, a torquetrum, a sextant of radius 1.3 m, a quadrant of radius 1.5 m, and armillary spheres. He demonstrated the superiority of metal instruments over wooden ones. Wilhelm's most significant innovation was the rotating dome; he observed from one built in a tower of his castle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

His star catalog of about 400 stars, although far short of the intended number of stars he had hoped to measure, was one of the first assembled in Europe.

The most significant observatory prior to the invention of the telescope was that of TYCHO BRAHE (1546–1601). It was situated on the 2000 acre island of Hven, in the Danish sound between Copenhagen and Elsinore. There Tycho built Uraniborg, the 'castle of the heavens'. It sat in the middle of a square 300 feet on a side and enclosed by a wall 22 feet high. The castle itself had two main levels and was taller than a seven-story building. In addition to the observing rooms and verandahs on the upper level, it contained a dozen or so bedrooms, a dining room, library, chemical laboratory and even a jail in the basement. Begun in 1576, it was completed in 1580. Over the north portal was an inscription, carved in stone: *nec fasces nec opes sola artis sceptra perennant* ('neither wealth nor power, but knowledge alone endures'). Uraniborg was the first scientific research institute in Renaissance Europe, and Tycho was the first full-time scientist.

At Uraniborg Tycho and his assistants used a dozen instruments. The one that produced the most accurate stellar positions was a mural quadrant of radius 2 m, which was located in the southwest room on the ground floor. A mural quadrant is attached to a wall, fixed on the celestial meridian. Its fine construction and the use of transversals allowed the determination of declinations with a resolution of 10 arcsec.

An auxiliary observatory, Stjerneborg, was built in 1584 a short distance to the south of the main building. There were five new instruments, the largest of which was an equatorial armillary made of iron, which was 2.9 m in diameter and could be used to measure declinations with a resolution of 15 arcsec. Each of these five instruments was situated in a subterranean crypt. Three were sheltered by folding roofs, while two had revolving domes.

From observations at Hven Tycho discovered four inequalities in the Moon's motion, two in longitude (the 'variation' and the annual equation with a period of one solar year) and two in latitude. His observations of Mars were used later by Kepler to discover the elliptical nature of planetary orbits. From observations of 1588–1591, Tycho produced a catalog of 777 stars with an improvement in positional accuracy of an order of magnitude. More than one of Tycho's instruments produced positional accuracies for his nine principal reference stars of ± 0.6 arcmin. The average uncertainty of the (other) brighter stars in Tycho's catalog is 1.9 arcmin in ecliptic longitude and 1.2 arcmin in ecliptic latitude. For the fainter stars the uncertainties amount to 2.8 and 2.6 arcmin, respectively.

Tycho was able to afford such a magnificent observatory because his annual income from various fiefdoms was the equivalent of 1% of the Danish government's income, and this continued for a period of nearly 30 yr. It was not just money, of course, that led to the improvement in instrumentation. More than one of Tycho's metal instruments required the workmanship of five or six people over a period of 3 yr.

Tycho's fortunes, both literally and figuratively, began to decline in 1588 after the death of Tycho's patron, the Danish king Frederick II. Tycho left Hven in 1597 and his observatory fell into disuse.

JOHANNES HEVELIUS (1611–1687), the son of a prosperous brewer of Danzig (Gdansk), built what was for a short time the world's leading observatory. He copied a number of Tycho's instruments, building many quadrants and sextants of wood or copper. For observing the Moon and planets he constructed refractors up to 150 feet in focal length, in which the objective was mounted on a tall mast. However, for stellar positions he preferred the unaided eye, as he explicitly stated in his *Machina coelestis* (1673). This precipitated a controversy with Robert Hooke, who was a strong advocate of telescopic sights. EDMOND HALLEY was compelled to visit Danzig in 1679 to try to resolve the controversy. For 2 months Halley observed with a telescope fitted with sights while Hevelius used naked-eye instruments. Their observations were of equal accuracy, slightly less than 1 arcmin. The *Selenographia* (1647) of Hevelius had been a milestone in lunar mapping, while his posthumous star catalog (1690) contains the positions of 1564 stars.

The rise of national observatories

First era: 1576–1725

We may distinguish three eras in the construction of national observatories. Tycho's island observatory was, in effect, a national observatory, but the Paris Observatory (established 1667) was the prototype of the national observatories that followed. Astronomers in Paris, along with those at the Greenwich Observatory (1675), Berlin (1701), and St Petersburg (1725), were dedicated to practical matters of national importance: improving navigation (especially the determination of longitude at sea), geodesy, calendar reform, producing accurate stellar coordinates and the determination of ephemerides of the Sun, Moon and planets.

The Paris and Greenwich observatories stand out as institutional models on which many subsequent observatories were based. The Paris Observatory reflected the splendor of the court of Louis XIV, and many significant astronomers were associated with it. Jean Dominique Cassini (1625–1712), his son, grandson and great-grandson, all of whom were unofficial directors at Paris, formed a most notable astronomical dynasty (see CASSINI DYNASTY). The first Cassini discovered four moons of Saturn and the division of the ring system named after him. Paris astronomers used extremely long focal length refractors (up to 136 feet), mounted on tall masts, to observe the planets and Moon. Ole Römer demonstrated the finite nature of the speed of light from observations of Jupiter's Galilean satellites. Paris astronomers carried out the first geodetic surveys, covering the full arc of meridian of France by 1700, and were later involved in expeditions to Peru and Lapland.

From the outset both Paris and Greenwich were mandated to aid navigation, and the production of nautical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. Pulkovo Observatory, originally completed in 1839, was completely destroyed during the Second World War. It was rebuilt by 1954. This is the main building as it appeared in 1989. (Photo by Kevin Krisciunas.)

almanacs was a way of doing this. The French *Connoissance de Temps* first appeared in 1679, and the *Nautical Almanac* in 1767. In 1795 production of the *Connoissance* became the responsibility of the newly created Bureau des Longitudes, which supervised the work of the Paris Observatory. In Britain the Board of Longitude supervised the work of the Greenwich Observatory from 1714 to 1828. By then, with the existence of accurate star positions, ephemerides and the invention of the marine chronometer, the problem of determining one's geographical position at sea was essentially solved.

The buildings, staff size and funding level of the Greenwich Observatory were, until the 20th century, modest compared with those of Paris. Several of the Astronomers Royal, such as Flamsteed, Halley, Bradley, Maskelyne and especially GEORGE BIDDELL AIRY (1801–1892), were among the most important astronomers of their day. Airy organized the observatory as an astronomical factory with himself as factory director and sought to carry out the observatory's mission (data acquisition, reduction and publication of results) as accurately and efficiently as possible.

Second era: 1820–1918

The second era of national observatory building was characterized by offshoots from previous national observatories (such as Royal Observatory Cape, South Africa, 1820), by newer observatories of younger nations (United States Naval Observatory, 1839) and later the rise of astrophysical observatories (Potsdam, Prussia, 1874). Other national observatories of the second era include Pulkovo (Russia, 1839) (figure 1), the Chilean National Observatory (1852), the Argentine National Observatory (1870), the Smithsonian Astrophysical Observatory (USA, 1891) and Canada's Dominion Observatory (1903) and Dominion Astrophysical Observatory (1918).

Of 19th century observatories, Pulkovo deserves special mention. Tycho had been given *carte blanche* by his sovereign to build the finest observatory in the world, and WILHELM STRUVE (1793–1864) was amply funded likewise by Tsar Nicholas I. Struve ordered an Ertel transit instrument (for the determination of absolute right ascensions), an Ertel vertical circle (for the determination of absolute declinations), a Repsold meridian circle (for the determination of differential stellar coordinates) and prime vertical transit (for the determination of aberration and nutation), a Merz & Mahler heliometer (for the determination of angular sizes or angular separations) and the 15 in Merz & Mahler refractor (for the discovery and measurement of double stars). This refractor was the largest telescope in the world at that time; a twin was built for Harvard College Observatory in 1847.

The instruments listed above testify to the maturation of positional astronomy as a field of endeavor. Not only did Pulkovo astronomers measure some of the first stellar parallaxes and discover many double stars, they also produced the most accurate values for the constants of precession, aberration and nutation—values not superseded until 1964.

It is important to note that progress in astronomy depended very greatly on progress in instrument making and optics. Thomas Tompion, George Graham, John Bird, Jesse Ramsden, Joseph Fraunhofer, Edward Troughton, Carl Zeiss, Alvan Clark, Alvan Graham Clark, Howard Grubb and George Ritchey are some of the men who made significant contributions to instrument making during this era.

Third era: 1956–present

The present era of national observatory building is characterized by national or international consortia, large budgets and the investigation of celestial objects at all electromagnetic wavelengths. From ground-based observatories we may investigate optical, infrared, radio, and submillimeter waves. Because of absorption by water vapor in the Earth's atmosphere of certain infrared and submillimeter waves, we must make some observations from balloons and aircraft. Ultraviolet, x-ray and gamma-ray astronomy *must* be carried out from satellites.

Significant national observatories of the third era include the National Radio Astronomy Observatory (USA, 1956), Kitt Peak National Observatory (USA, 1957), National Radio Astronomy Observatory (Australia, 1959), Cerro-Tololo Inter-American Observatory (Chile, 1963) (figure 2), European Southern Observatory (Chile, 1964), Anglo-Australian Observatory (Australia, 1967), the Kuiper Airborne Observatory (USA, 1975–1995) and the Space Telescope Science Institute (1981).

Private observatories

Not all significant astronomical work is carried out at observatories funded by federal governments, universities and research foundations. In the late 18th century and throughout the 19th religious orders (notably the Jesuits) and various (usually wealthy) individuals established their own private observatories.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The most accomplished private astronomer of all time was the German–English astronomer William Herschel (1738–1822) (see HERSCHEL FAMILY). He pioneered the production of larger and larger *reflecting* telescopes (up to 48 in in diameter), whose light-gathering power enabled him to establish the field of galactic astronomy. His discovery of the planet Uranus in 1781 won him a royal pension that enabled him to give up his profession of music, and he moved to the neighborhood of Windsor Castle where his only obligation was to demonstrate the heavens to the royal family when requested to do so. He carried out deeper and deeper surveys of the northern sky, discovered hundreds of double stars and many hundreds of nebulae, and laid the groundwork for studies of the evolution of stars and stellar systems.

Herschel's son John followed in his father's footsteps and took an 18 in, 20 foot focal length reflector to South Africa where from 1834 to 1838 he conducted a southern sky survey that completed his father's work. The nebulae discovered by the Herschels were to form the basis of Dreyer's *New General Catalogue* (1888).

Following Herschel's discovery of Uranus, the German astronomer JOHANN HIERONYMUS SCHRÖTER (1745–1816) quit his post in Hanover and moved to the small town of Lilienthal to devote himself to observational astronomy. He built what was then the largest observatory on the European continent, and was assisted by Karl Ludwig Harding (who discovered the third asteroid Juno in 1804) and Friedrich Wilhelm Bessel. Schröter made detailed drawings of features on the lunar surface and believed he detected changes in the structure of some of those features. He inspired many observers to dedicate themselves to planetary studies.

William Parsons, the THIRD EARL OF ROSSE (1800–1867), in 1839 constructed at Parsonstown in central Ireland a reflector with a mirror 3 feet in diameter, and he followed this in 1845 with a monster with a mirror 6 feet in diameter. The 'Leviathan of Parsonstown' faced in a southerly direction, slung between masonry walls. Within weeks of completion it revealed the spiral nature of certain nebulae.

In 1856 WILLIAM HUGGINS (1824–1910) established a private observatory at Tulse Hill, near London, where he observed the spectra of stars and other celestial objects. Huggins was the first to prove that some nebulae are composed entirely of glowing gas. After his marriage in 1875 to Margaret Lindsay Murray, husband and wife formed a productive scientific partnership.

In 1868, simultaneously with the Frenchman Jules Janssen, the English amateur NORMAN LOCKYER (1836–1920) used a spectroscope attached to a refractor to observe solar prominences outside of a total solar eclipse. He was subsequently appointed head of the Solar Physics Observatory at South Kensington. When he retired in 1911 he established his own observatory at Salcombe Regis.

The American HENRY DRAPER (1837–1882) was a pioneer of astrophotography. His widow funded significant spectroscopic research carried out by astronomers at the Harvard College Observatory.

Figure 2. Cerro Tololo Inter-American Observatory illuminated by moonlight. The largest dome houses the 4 m reflector. (Photo by Kevin Krisciunas.)

Figure 3. Lick Observatory, ca 1936. The large white dome houses the 36 in refractor, still the second largest in the world.

GEORGE ELLERY HALE (1868–1938), the son of a wealthy Chicago industrialist, invented the spectroheliograph while an undergraduate at the Massachusetts Institute of Technology. He established a private solar observatory in his parents' backyard in Kenwood, then a suburb of Chicago; this later became part of the University of Chicago. Hale became the greatest observatory entrepreneur of all time, setting up the Yerkes Observatory (1897) and the Mt Wilson Observatory (1904). He was the driving force behind the establishment of Palomar Observatory (1948).

Today, as a result of advances in electronics and imaging devices, many amateurs obtain significant hard data for professional research projects. Photoelectric and CCD photometry by amateurs has been very important for variable star research and has even led to the discovery of a new class of pulsating stars. Astrometry from CCD imagery routinely provides data on recently discovered asteroids, which allows the calculation of their orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Until very recently the most successful discoverer of supernovae in other galaxies was the Australian amateur Robert Evans.

Mountaintop observatories and the modern era

As a result of his optical researches, Isaac Newton realized that the Earth's atmosphere acts as a lens of sorts and that one could achieve much better astronomical image clarity by siting one's telescope on 'some lofty mountain' (see OBSERVATORIES AND THEIR LOCATION).

In 1741 the French astronomer François de Plantade carried out the first high-altitude astronomical observations near Pic du Midi in the French Pyrenees. Unfortunately, he died at the foot of the Pic, although the cause of his death is not known.

The next small step leading to improved observing was taken with the establishment in 1783 of Dunsink Observatory, an institute of Trinity College, Dublin. The observatory was intentionally sited some 8 km outside the city. It also had the first fully functional rotating hemispherical dome, and the primary telescope was mounted on a pier which was structurally (and therefore vibrationally) isolated from the walls of the building.

In 1856 the Scottish astronomer Charles Piazzi Smyth spent nearly 4 months at Tenerife in the Canary Islands and demonstrated that mountaintop observing was both desirable (for image clarity) and viable (for humans).

The first mountaintop observatory intended to be permanent was Lick Observatory (figure 3), which was completed in 1888. Situated at the 1280 m summit of Mt Hamilton, near San Jose, California, it boasted a 36 in refractor by Alvan Clark, then the world's largest. After the subsequent spectacular success of the Mt Wilson Observatory and the Palomar Observatory, it became clear that the best seeing was obtained at mountaintop sites within 50 miles of the ocean. Because of the proximity of the ocean, one gets a smooth, laminar flow of air over the coastal mountains when the prevailing wind conditions are in effect. This has been borne out by the establishment of many other observatories during the second half of the 20th century. The highest major observatory in the world (at 4205 m) is at Mauna Kea on the island of Hawaii. From a very modest start in 1964, Mauna Kea's astronomical installations expanded greatly and now include the University of Hawaii 2.2 m telescope, the Canada–France–Hawaii Telescope (3.6 m), the NASA Infrared Telescope Facility (3.0 m), the United Kingdom Infrared Telescope (3.8 m), the James Clerk Maxwell submillimeter telescope (15 m), the Caltech Submillimeter Observatory (10.4 m), one of the nine elements of the Very Long Baseline Array (25 m radio telescope), the two Keck telescopes (each 10 m), the Japanese optical and infrared telescope Subaru (8.3 m), the northern Gemini telescope (8.1 m) and the Smithsonian Astrophysical Observatory Submillimeter Array.

At present, the highest established telescopes are the twin 0.7 m reflectors of the Meyer–Womble Observatory at Mt Evans, Colorado (elevation 4313 m). A still higher

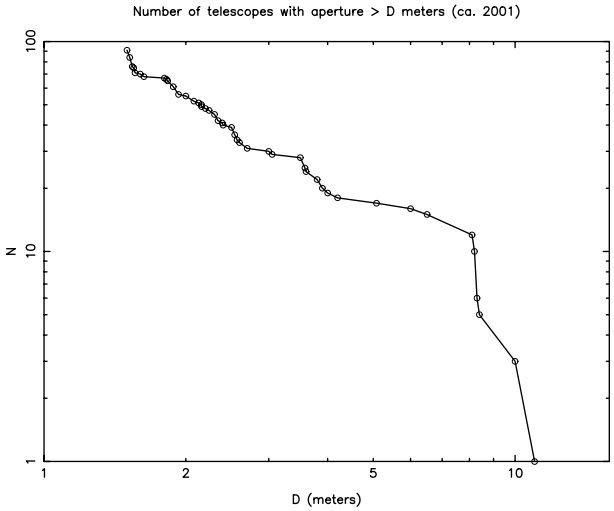


Figure 4. The number of optical and infrared telescopes with primary objectives larger than a given size, ca 2001. We include all telescopes with diameter D greater than or equal to 1.5 m. We count the two 8.4 m diameter elements of the Large Binocular Telescope, the four 8.2 m diameter elements of the Very Large Telescope and the two 6.5 m telescopes of the Magellan Project as separate instruments. Each of the 10 m Keck Telescopes contains 36 hexagonal mirror segments, while the 11 m Hobby–Eberly Telescope has 91 hexagonal segments. (Diagram by Kevin Krisciunas.)

observatory is being built at 5000 m about 40 km east of San Pedro de Atacama in Chile. This is the centimeter-wavelength Cosmic Background Imager, which will investigate the small ripples in the 3 K background radiation left over from the Big Bang. This observatory will have an oxygen-enhanced environment for the operators.

At the time of the publication of this article there will be more than 90 optical or infrared telescopes in operation around the world with diameters of 1.5 m or larger (see figure 4). This includes the now single-mirror 6.5 m MMT at Mt Hopkins, Arizona, the twin 6.5 m reflectors of the Magellan Project (situated at Las Campanas, Chile), the twin 8.1 m Gemini telescopes (one at Mauna Kea, Hawaii, the other at Cerro Pachon, Chile), the four 8.2 m elements of the European Southern Observatory's Very Large Telescope (at Cerro Paranal, Chile), Japan's 8.3 m Subaru Telescope (at Mauna Kea), the two 8.4 m elements of the Large Binocular Telescope (at Mt Graham, Arizona), the two 10 m Keck telescopes at Mauna Kea and the 11 m Hobby–Eberly spectroscopic telescope at Mt Locke, Texas. In fact, by the beginning of the 21st century more than half the *area* of telescopes larger than 1.0 m will be in telescopes larger than 8 m.

Given the expense of such facilities (measured in units of hundreds of millions of dollars), one might naturally ask why so many large telescopes are being built. It is a reflection of the blossoming of extragalactic observational astronomy, the desire to give hard data to cosmologists for testing their models and the wish to discover and study

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

intrinsically faint objects such as asteroids in the outer solar system, brown dwarf stars and extrasolar planets.

The design and construction of these state-of-the-art telescopes has required the development of computer-controlled mirror mounts, better-ventilated domes (to eliminate as much as possible the degradation of the visibility due to air in the dome), thin mirrors (as in the case of most of the telescopes mentioned above) or multimirror systems (36 hexagonal segments in the case of each of the Keck Telescopes, 91 segments in the case of the Hobby–Eberly Telescope). Modern instrumentation often uses techniques borrowed from spy satellites to achieve diffraction-limited performance. Whereas photographic plates a century ago had a quantum efficiency (QE) of less than 1% and photomultiplier tubes typically had peak QEs of a few per cent, modern instrumentation can achieve QEs of 70%. Such instruments, fed by the large objectives of the present generation of telescopes, are producing many spectacular results. Furthermore, modern computing capabilities are allowing digital celestial surveys which would have been technically impossible only a few years ago.

Bibliography

Dick S J 1990 Pulkovo Observatory and the national observatory movement: an historical overview *Inertial Coordinate System on the Sky* ed J H Lieske and V K Abalakin (Dordrecht: Kluwer) pp 29–38

Howse D 1986 The Greenwich list of observatories: a world list of astronomical observatories, instruments and clocks, 1670–1850 *J. Hist. Astron.* **17**(4) (1994 Amendment list no. 1 *J. Hist. Astron.* **25**(3) 207–18

Krisciunas K 1988 *Astronomical Centers of the World* (Cambridge: Cambridge University Press)

Müller P 1992 *Sternwarten in Bildern: Architektur und Geschichte der Sternwarten von den Anfängen bis ca. 1950* (Berlin: Springer)

Needham J and Ling W 1959 *Mathematics and the Sciences of the Heavens and the Earth (Science and Civilisation in China vol 3)* (Cambridge: Cambridge University Press)

Sayili A 1981 *The Observatory in Islam* (New York: Arno)

Thoren V E 1990 *The Lord of Uraniborg: a Biography of Tycho Brahe* (Cambridge: Cambridge University Press)

Kevin Krisciunas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatorio Astronómico Nacional, Colombia

The Observatorio Astronómico Nacional (OAN) was founded by José Celestino Mutis in the framework of the 'Expedicion Botanica' on 20 August 1803. His first astronomer was Francisco Jose de Caldas. By the turn of the century, Julio Garavito worked in celestial mechanics—a crater on the far side of the Moon bears his name.

Nowadays, the OAN leads a postgraduate program, 'Especializacion en Astronomia', and the staff work in areas of celestial mechanics, statistical astronomy, cosmology, stellar structure, instrumentation, national ephemerides and parallel computing. The OAN is situated in Bogotà, Colombia (South America), and at present belongs to the 'Universidad Nacional de Colombia' at the Faculty of Sciences.

For further information see
http://www.observatorio.unal.edu.co/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatorio Astronómico Nacional, Mexico (OAN)

Located in the mountains of the Sierra San Pedro Martir in Baja California. The observatory operates three telescopes—a 2.1 m, 1.5 m and 0.84 m. The observatory offices and workshops are located in Ensenada, beside the Pacific Ocean. OAN is a part of the Instituto de Astronomía of the Universidad Nacional Autónoma de México.

For further information see http://www.astrosen.unam.mx/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatorio del Roque de Los Muchachos

The Observatorio del Roque de los Muchachos, at a height of 2400 m above sea level, is situated on the island of La Palma (Canary Islands) and belongs to the INSTITUTO DE ASTROFISICA DE CANARIAS (IAC). It contains one of the world's most extensive collections of telescopes, which are operated by several countries, the largest being the William Herschel Telescope (4.2 m). Apart from several nocturnal telescopes, the Observatory houses two solar telescopes, an automatic meridian circle and an array of cosmic-ray detectors. In 2003, the Gran Telescopio Canarias (GTC), an optico-infrared telescope of 10 m diameter, is planned to come into operation.

For further information see
http://www.iac.es/folleto/orml.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatorio del Teide

The Observatorio del Teide, at a height of 2400 m above sea level on the island of Tenerife (Canary Islands), forms part of the INSTITUTO DE ASTROFISICA DE CANARIAS (IAC). It contains several solar telescopes (the largest of these being greater than 90 cm in diameter) operated by different countries, as well as helioseismological instrumentation, radiotelescopes to study the cosmic microwave background, various optical telescopes, a 1.55 m infrared telescope and an optical ground station for communication with satellites and the cataloging of space debris.

For further information see http://www.iac.es/folleto/teide.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observatory

A structure from which astronomical observations may be made. In the case of ground-based optical and infrared astronomy, an observatory is a structure within which a telescope is housed and which allows the selected area of sky to be observed while at the same time providing the instrument and observer with protection from the wind and, to some extent, extraneous light. In the context of radioastronomy, where, in most cases, the radiotelescopes are out in the open and the observers in a building, the term 'observatory' relates to the complex of instruments and buildings on the observing site.

The classic design of observatory has an opening slot in a hemispherical dome that rotates on a circular wall. In some cases, the 'dome' is cylindrical (like a cake tin) or polygonal. In others, the entire building rotates or, in smaller sizes, the entire roof slides off to one side. In order to minimize the effects of light pollution, absorption of light in the atmosphere and scattering of light by dust particles suspended in the atmosphere, and to attain the steadiest possible atmospheric conditions, most of the major new and recently constructed telescopes are housed in observatories that are clustered together on one or other of a small number of high, isolated, mountain sites. Among the principal sites are Mauna Kea, Hawaii (at an altitude of 4200 m, the world's highest major concentration of observatories), Cerro Tololo, La Silla and Cerro Paranal, all in the Chilean Andes, La Palma (Canary Isles) and Kitt Peak, Arizona.

See also: environmental threat to astronomy, infrared astronomy, optical astronomy, radioastronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observing Asteroids

The asteroids were so named in the early 1800s by William Herschel as they appeared star like to an observer (figure 1). They appear just like stars when observed with amateur equipment, but stars that move relative to the background star field.

Initially amateurs observed asteroids purely for the thrill of discovery, but in the 1940s, they became known as vermin because they 'contaminated' the sky survey plates taken at the large observatories for mapping the stars and galaxies. Interest returned in the 1960s when the prospect of sending space probes through the main asteroid belt to the outer planets became a reality. The discovery of asteroids in the outer solar system in 1976 and beyond Neptune in 1992 prompted further studies as these were examples of bodies that had formed in the early solar system. There is now much observational activity directed to the discovery and follow-up of NEAR-EARTH ASTEROIDS as realization of the risk and effects of impact have slowly dawned on both scientists and the public at large.

Many amateurs begin asteroid observing by learning to track them down and keeping a log of those viewed. Beyond this 'trophy count' there is quite a variety of observation that can be done visually and much more with the use of a CHARGE-COUPLED DEVICE (CCD) camera. It is possible to image the brighter asteroids with a CCD camera and a 135 mm focal length telephoto lens mounted on an undriven camera tripod (figure 2). With no instrument at all, there is little to be done in the field of asteroid observing—at a favorable opposition 4 Vesta can reach naked eye visibility but even then a clear dark sky is necessary to allow it to be seen. However, with even small binoculars several examples can be seen at favorable opposition and this number increases significantly as the size of instrument used is increased. A permanent record can be made with photographic film or a CCD imager for study later.

Figure 1. Asteroid 15 Eunomia (arrowed) in Pegasus on 11 September 2001—135mm lens and CCD.

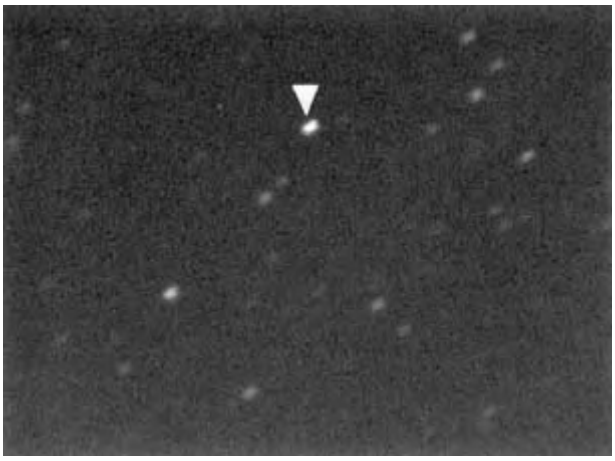


Figure 2. An undriven 4 second exposure of 4 Vesta in Taurus —31 December 2001.

The first task is to find the asteroid. As a minimum it is necessary to plot the position on a star chart to take to the telescope so that a visual identification can be made. On confusing fields where there are a number of possible

Figure 3. 135mm focal length lens and Starlight Express MX-516 CCD on a Celestron NexStar GoTo tripod mounting.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

stars it is necessary to wait for perhaps an hour to identify which object has moved and is hence the asteroid. I currently use a telephoto lens–CCD on a driven GoTo mount (figure 3) so I only need to know the position to send the imager and can find the asteroid on the resulting image. Modern planetarium-type computer programs, such as GUIDE8 from Project Pluto, can plot a starfield with the asteroid shown and can in some cases also drive the telescope to the field (a far cry from what could be done in the 1950s).

Once the asteroid is found there are a number of things that can be done. The brightness can be deduced either visually or preferably from an image. Suitable comparisons can be taken from GUIDE8 by choosing HIPPARCOS stars of similar color to that of the asteroid. For a CCD image a V-filter should be used to ensure that the light passed is similar to eyeball observation. The Sony chip has a pseudo-visual window and can be used carefully with no filter as long as care is taken in the choice of comparison star. From this an indication of how the asteroid's brightness changes as it comes to and recedes from opposition (figure 4). Taking images of one asteroid over a night allows the rotation period to be determined (figure 5). Typically these are between 6 and 12 h although there are many examples known with periods shorter or longer than this.

Modern computer software allows CCD images to be used for positional measures. This is assuming that the start and end of the exposure is accurately timed and a suitable positional catalog is used (or downloaded from the internet for the field imaged).

Frequently asteroids occult (see OCCULTATION) field stars as they follow their orbital motion. As asteroids are relatively small (most less than 300 km in diameter) the path across the Earth is small. In recent years the positions of the stars and the orbits of the asteroids have been so refined that most predictions now have an uncertainty of typically 100 km—before HIPPARCOS and the orbital work, the uncertainty was 10 times greater or even more. Timing the events (or recording a miss) are

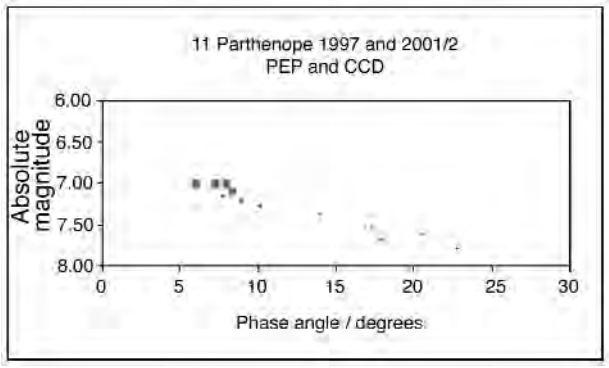


Figure 4. Phase curve for asteroid 11 Parthenope from CCD and photoelectric photometry.

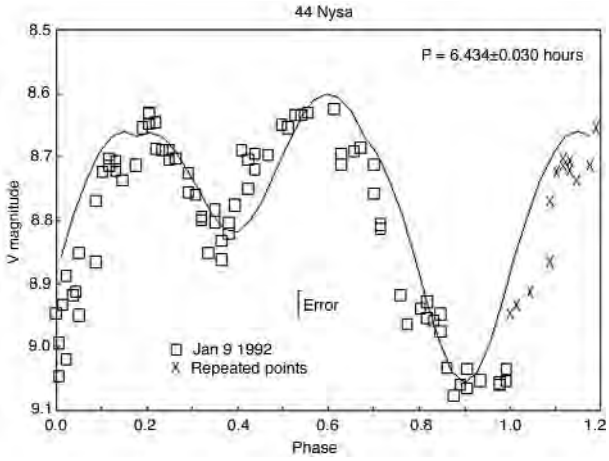


Figure 5. Rotation light curve for 44 Nysa from photoelectric photometry.

used to prepare a shape and size profile of the asteroids. Most of the information we have is from Earth-based observation as there are only a limited number of space missions that can be flown and a larger number of potential targets.

There are many observers who cooperate with professional programs to study near-Earth objects (comets can also be in this group). Very few observers are now discovering new objects (pre 2000 there were many unknown at 16th magnitude) as most discoveries are now 19th magnitude or fainter. Important aspects in this field are follow-up observations to define the orbits or photometry to check the rotation periods. Notices are given for those which will be studied by radar from Arecibo and rotations are studied using CCDs which help with the radar work.

From being a Cinderella field, the study of asteroids is now of high importance. All levels of study can be undertaken using simple equipment or the most modern of technology. As such there is much to interest all levels of observer.

Andrew Hollis

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Obsevatoire d'Haute-Provence

The Obsevatoire d'Haute-Provence is situated in south-eastern France at Saint-Michel-l'Observatoire, at an altitude of 650 m. Founded in 1937, it was originally an observational national facility for French astronomers. Now it has become an astronomical institute of the French CNRS with several fields of activity.

A staff of astronomers and post-graduate students are active in fields such as quasars, x-ray binaries, interstellar matter etc. Visiting astronomers use the 1.93 m and 1.52 m telescopes and their high-accuracy spectrographs with which the first extra-solar planets have been discovered (1995). Smaller telescopes are used for educational purposes or specific experiments.

All instrumentation including CCD controllers is observatory designed and built. Partnerships are developed with other institutes for the design of space experiments and the construction of the focal instrumentation of ESO VLT.

Research and development of new technologies related to large arrays of telescopes and interferometry are carried out.

A Department of Geophysics is active in research on the Earth's atmosphere as a member of the 'Network for Detection of Stratospheric Changes'.

For further information see
http://www.obs-hp.fr/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Occultation

The motion of celestial bodies sometimes causes their paths to cross on the heavens. One says that the background object (usually a star) is occulted by the front object. Usually, a star is occulted by a solar system body: a planet, a satellite or an asteroid. The occultation of the Sun by the Moon is called a 'solar eclipse'. In fact, an eclipse is the disappearance of a body in the shadow of another one and an occultation (from the Latin *occultare*, to hide) happens when a body hides another body: in a lunar eclipse the Moon is in the Earth's shadow but a solar eclipse is in fact the occultation of the Sun by the Moon.

The occultation of stars by the Moon has been observed for a long time to study the stars. Later, stellar occultations by any solar system objects have been observed to study the planets, their atmospheres and their ring systems and small bodies such as satellites or asteroids. Occultations have now become a tool to explore the environment of other stars.

Lunar occultations

The disappearance of a bright star (Regulus, Antares, Aldebaran) behind the Moon is not a rare phenomenon, even if it is very difficult to observe without resorting to instrumental aid. However, some disappearances are reported in astronomical treatises. Copernicus observed a lunar occultation of Aldebaran in 1497. These observations allow one to deduce the relative proximity of the Moon and the very small apparent size of the stars. Sometimes, the star disappears in two steps, and this is interpreted as being because the light source is a binary. The abruptness of the occultation leads also to the conclusion that the Moon has no or at most a very tenuous atmosphere.

When photographic emulsions became sensitive enough, it was possible to record the rapid fluctuations of the stellar light before its disappearance and to evaluate the star's diameter. By 1980, a hundred star diameters had been measured with this method. The accuracy of the record was such that one had to take into account the fact that the occultation of a star by the Moon is a diffraction phenomenon: this means that the propagation of light when the star approaches the Moon's edge creates interference. The light fluctuates before extinction and reaches a maximum intensity which is 1.37 times the nominal star intensity. This flickering is extremely rapid, two successive peaks being separated by 0.02 s. That is why records on rapidly moving film were necessary to detect the diffraction fringes. We had to await the discovery of photometric detectors sensitive and rapid enough to exploit this phenomenon and obtain accurate measures of stellar diameters.

A tool to explore the solar system

Because they have a very small apparent size, occultations of bright stars by objects of the solar system are rare. Modern instruments, such as rapid aperture-photometers and charge-coupled devices, have allowed one since 1952 (photoelectric record of the occultation of σ Arietis by Jupiter) to observe occultations by planets, satellites, comets and asteroids. These observations are very fruitful tools to explore these objects.

- The exact position of the occulting object can be computed. The mutual occultations (and eclipses) of the satellites of JUPITER and SATURN occur each time the Earth passes through their satellite orbital planes, every 6 yr in the case of Jupiter. The principal value of the observation of these 'mutual phenomena' is in the improvement of our knowledge of the orbits of the satellites but it also allows us to detect small features on the satellites such as volcanoes on Io.
- When an occultation is observed from several observatories, the corresponding sections on the occulting object are separated by the same distance as between the observatories and the comparison of the lightcurves allows us to reconstruct the shape of the occulting object (figure 1). When their mass is known, the extremely precise size of some asteroids obtained with this method has made possible the determination of their density which provides clues to both composition and origin. An unexpected offshoot of minor planet occultations was the detection of possible satellites to some minor planets (Herculina, Melpomene, 146 Lucina), each satellite detected as a brief secondary dip in the occultation lightcurve. The observation by GALILEO of the double asteroid Ida and Dactyl has confirmed the reality of asteroid satellites.

Figure 1. Limb profile of the asteroid 2 Pallas deduced from observations of chords during the 29 May 1983 occultation of 1 Vulpeculae. (From Dunham *et al* 1983.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- The extinction of a star's light behind a planet's atmosphere reveals the latter's temperature profile and the possible presence of haze but also the presence of gravity waves, some structures which seem to be a constant in all the dense planetary atmosphere from the Earth to Neptune. When the star disappears behind the atmosphere, it strongly sparkles before disappearing. In fact, the atmosphere is not isothermal and local temperature variations betray variation of the refractive index. Just as the warm air above a heater distorts images, so the star twinkles when it passes behind the planetary atmosphere (figure 2 and figure 3). The measurement of these fluctuations provides a way to discover atmospheres. Pluto's atmosphere has been discovered by this method.
- A few lucky occultations have provided information about the shape of the planet's atmosphere. The star's light is focused by a circular atmospheric layer and this creates a bright spot in the middle of the planet's shadow. The oblateness of the atmosphere gives to this spot the shape of a cardioid. If the observatory is at less than few tens of kilometers from the center of the planet's shadow, a scan of this 'central flash' is observed, which gives the global shape of the atmosphere. This phenomenon has been observed for Mars, Neptune, Titan (figure 2) and Triton.
- Observations of occultations led to the discovery of the very narrow (few kilometers) and/or thin rings system of URANUS (1977) and NEPTUNE (1984), which are almost invisible with other methods of observation. The Uranian rings reveal their presence by nine symmetrical brief drops in the stellar flux before the immersion behind the planet and after the emersion. The incomplete rings of Neptune, called arcs, left astronomers puzzled as they observed only one drop. They had not imagined that incomplete rings could exist: in fact they cannot exist without the help of a satellite. This shepherd satellite was predicted by theoreticians after the discovery of the arcs and observed much later by Voyager 2. Thanks to successive occultations, the positions of the Uranian rings are now known with a precision of a few hundred meters.

For these results it is necessary to record the star's light with a very high frequency (100 points s^{-1}) and to know precisely the timing of the occultation: this is now possible with the GPS satellite network. This method of observation is very favorable for Uranus and Neptune, for which some tens of occultations have been observed. Taking advantage of the presence of methane in the atmosphere of these planets, the occultations are observed at the infrared wavelengths of 2.2 and 0.9 μm: at these wavelengths, which correspond to absorption

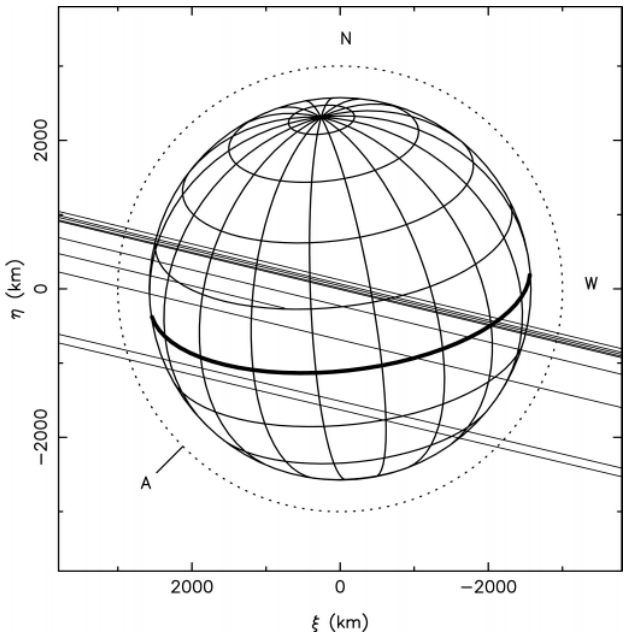


Figure 2. Geometry of the occultation of 28 Sagittarii by Titan, observed on 3 July 1989. W and N denote celestial west and north. The lines are the tracks of various stations in Titan's shadow. Observers move from the right to the left in this diagram. The upper tracks, packed together, correspond to stations in Northern Europe (including Hanover); they pass near the center of the shadow and observe the central flash. The middle and lower tracks correspond respectively to stations in Italy and Israel. The dotted circle is at 425 km altitude, where the atmospheric pressure is 7 μbar. (From Sicardy *et al* 1999.)

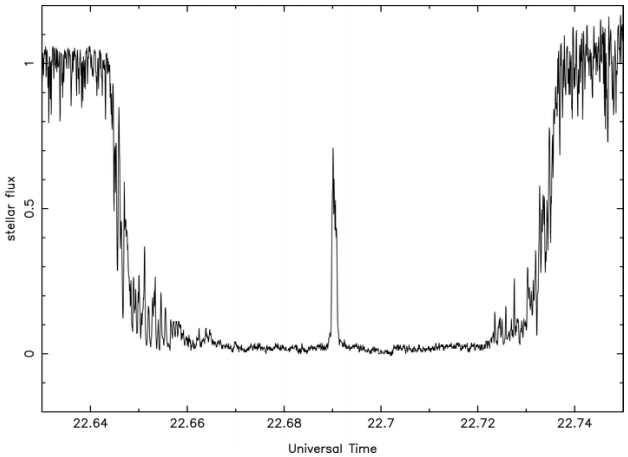


Figure 3. Light curve of the occultation of 28 Sagittarii by Titan, observed from Hanover, Germany, by H J Bode and C Bittner with a 0.4 m telescope at 0.65 μm wavelength. The normalized stellar flux is plotted in term of decimal Universal Time. The spikes visible during the immersion and the emersion are due to local temperature variations of the atmosphere. The central flash is created by the focusing of stellar light by a deep atmospheric layer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

bands of methane, the planet is darker than in the visible wavelengths, and the signal-to-noise ratio of the light curves is increased. Jupiter and Saturn are not such good candidates for this method of observation because they are too bright. However, in 1989, the very bright star 28 Sagittarii (its visible magnitude is 5.5) was occulted by Saturn and its ring system (see SATURN: RINGS) and, 1 day later, by its moon TITAN: this exceptional event has been a mine of information about the atmosphere of this satellite (see figures 2 and 3).

Predictions and observations of occultations

Satellites, or asteroids, much smaller than the planets, have a smaller chance to occult a bright star; moreover, the necessary task of prediction of occultation is then much more difficult. It is necessary to know as exactly as possible when and where the shadow of the occulting object will sweep the Earth to prepare for the observation. For this, astronomers compare the EPHEMERIDES of the planet or satellite with a sky map. The uncertainty in the shadow trajectory make the observation chancy from a given place. Some scientific teams have developed portable kits with a telescope and an acquisition system. They wait to determine precisely the shadow's trajectory, a few days before the event, and they place the systems at the right location. The short-term predictions, the brevity of the events (less than a second for a ring and roughly 1 h for a giant planet) and their uniqueness make these observations very difficult. However, they are an unequalled tool for the exploration of the solar system, thanks to their high accuracy (up to 1 km) for the determination of the dimensions of solar system bodies.

Prospects for the future

Stellar occultations are the only way to detect small objects in the outer solar system which are too faint to be detected directly: the future will make stellar occultations a tool for exploration of the asteroid family and the small satellites such as Charon or Triton but also of the KUIPER BELT, the residual protoplanetary disk which lies beyond Neptune.

Sometimes, a star of a binary system is occulted by its dimmer companion. These so-called eclipsing binary are observed to determine their relative mass.

An extraordinary object is the star β Pictoris: spectral signatures of objects passing in front of the star are observed and these phenomena are interpreted as comets occulting the star when they fall into it.

The environment of stars can also be explored by searching for the extinction of the primary during the transit of giant planets.

The search for transits of planets in front of the central star is a promising method to detect extrasolar planets: a planet passing in front of the star creates a very faint extinction. Thus, this method needs to detect an extremely small decrease of the stellar flux. The main limit to this method is the scintillation due to the Earth's atmosphere. Observation from space will substantially improve the power of this method.

Bibliography

Dunham D W, Van Flandern T C, Millis R L , Chapman C R, Maley P D and Povenmire H 1983 The size and shape of (2) Pallas from its occultation of 1 Vulpeculae on 1983 May 29. *Bull. Am. Astron. Soc.* **15** 822

Elliot J L and Olkin C B 1996 Probing planetary atmospheres with stellar occultations *Annu. Rev. Earth Pl. Sci.* **24** 89–123

Warner B 1988 *High Speed Astronomical Photometry* (Cambridge: Cambridge University Press)

Sicardy B, Ferri F, Roques F, Lecacheux J, Pau S, Brosh N, Nevo Y, Hubbard W, Reitsema H, Blanco C, Carreira E, Beisker W, Bittner C, Bode H, Bruns M, Denzau H, Nezel M, Riedel E, Struckmann H, Appleby G, Forrest R, Nicolson T, Hollis A and Miles R 1999 The structure of Titan's stratosphere from the 28 Sgr occultation. *Icarus* **142** 357–390

Françoise Roques

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oceans

Earth is a planet covered with oceans. This article describes these great bodies of water and the forces that shape their circulations. Also discussed are the latest observational techniques in oceanography with emphasis on remote sensing. Turning to the possibility of oceans on other worlds, the indirect evidence for ice-covered seas on Jupiter's largest moons is reviewed.

D H Lawrence once wrote: 'They say the sea is cold, but the sea contains the hottest blood of all.' These evocative words refer to whales in the sea, and they remind us that the topic of Earth's oceans is larger than just the physics of waves and currents, but also encompasses the biology and ecology of much of the planet's living space.

Humankind has always held a deep-seated fascination for the marine world, and more than just as an abundant source of food and an avenue for transportation. Interestingly, when spoken in reverse the poet's words still make sense. Consider that seawater is about ten times saltier than blood, but that, by mass, ocean salt consists of 55% chlorine (Cl^-), 30% sodium (Na^+), 8% sulfate (SO_4^-), 4% magnesium (Mg^+), 1% potassium (K^+), and 1% calcium (Ca^{2+}). The corresponding percentages in human blood plasma are 51%, 44%, 0.6%, 0.3%, 2% and 0.7%. Except for sulfate and magnesium, which are much reduced in blood, the salt ions appear in roughly similar proportions in blood and seawater. In this sense 'our blood contains the sea'.

Seven seas

Like a map of the constellations, a world map of Earth shows the patterns and locations of the ocean basins as they appear to us in the present epoch. But, just as the last frame of a movie provides only a hint of the plot, so too a world map hides most of the history of the endless movements of the continents. The solar system is 4.6 billion years old, based on radiogenic dating of primitive meteorites, and one need travel back in time only 2% of that age to see the Atlantic Ocean disappear. Continental drift proceeds at a typical rate of about 2 cm yr^{-1}, which adds up to more than two trips around the circumference of Earth over the age of the solar system. The Seven Seas are going places!

The Seven Seas are the Mediterranean Sea, the Indian Ocean, the Red Sea, the Persian Gulf, the Black Sea, the Adriatic Sea and the Caspian Sea. Today, we have a global view of Earth and speak in terms of four oceans, the Pacific, the Atlantic, the Indian and the Arctic.

Arctic Ocean

Earth's south polar region is covered by the continent of Antarctica, but its north polar region has no continental land mass or islands, instead, it is the location of the planet's smallest ocean, the Arctic Ocean. This remote body of water is nearly landlocked, with half of its sea floor made up by continental shelf. There is a central basin that contains three submarine ridges; the deepest point is 4.7 km below sea level.

The majority of the Arctic Ocean is north of the Arctic Circle and hence stays in continuous darkness during the winter and continuous sunlight during the summer. The ocean is mostly covered by sea ice that is 3 to 4 m thick, but where the ice opens up there is abundant life. The salinity of the ocean decreases from wintertime to summertime as the ice cap melts and shrinks in area (sea-ice melt has about one-tenth the salinity of average sea water).

The shortest route between North America and Russia is via the Arctic Ocean, and there are active collaborations between Russian, Canadian and US scientists focused on establishing the ocean's circulation patterns and interactions with the Atlantic Ocean via the Greenland Sea and Pacific Ocean via the Bering Strait. There have even been some recent attempts to find analogies between the morphology of pressure ridges in Arctic sea ice and the linear features seen in spacecraft images of Jupiter's icy satellite Europa.

Indian Ocean

The only ocean included in the original Seven Seas is the Indian Ocean, which is nearest to the world's oldest civilizations. It covers about one-fifth of the Earth's surface and is somewhat smaller than the Atlantic Ocean. It is bordered by Africa and the Arabian Peninsula to the west and Asia to the north, and because of this prime location has had a long, and sometimes contentious, maritime history.

Weather in the Indian Ocean region is dominated by monsoons (which comes from the Arabic word for season). These torrential wind and rain storms start June, with the heaviest rains hitting India in October, November and December. The winds are driven by differences in heating between the land and sea, and because they arise suddenly and cause much damage there is an effort to better monitor the storm breeding grounds in the Indian Ocean using satellite remote sensing.

Near the surface, the Indian Ocean's water is relatively poor in nutrients (it is warm and buoyant and hence resists mixing with nutrient-rich deep water). For this reason, phytoplankton (microscopic plants) do not thrive, and the ocean is not heavily populated by fish. The circulation in the southern basin is counterclockwise and the winds there are relatively mild. In the winter, the monsoon winds reverse the ocean currents around the northern basin from clockwise to counterclockwise.

Effects of rotation

Large-scale circulations in the oceans are affected by the EARTH'S ROTATION in the same manner that winds are affected in the atmosphere. Instead of rushing down-gradient to fill in low-pressure regions, which is the 'nature abhors a vacuum' direction that a fluid takes in a non-rotating system, Coriolis accelerations cause ocean currents to circulate around highs and lows in the water.

The balance between the horizontal pressure-gradient force and the Coriolis acceleration that occurs in all rapidly rotating atmospheres and oceans is called geostrophic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

balance. The Earth-orbiting TOPEX/Poseidon and ERS-2 (European Remote Sensing) satellites both generate daily global maps of sea-surface height that can be converted to ocean surface circulations via geostrophy. The Atlantic's Gulf Stream, for example, stands out in surface altimetry data as a sharp ∼1 m step (low on the western side, high on the eastern side). It is a step because sea-surface height is a surrogate for pressure, and it is the pressure gradient that is proportional to the current. It has been demonstrated that the abyssal circulation is controlled by the pressure boundary condition above, meaning that a useful estimate of the circulation throughout the entire ocean can be made by combining satellite altimetry data with numerical models. In addition to correlating undersea circulations with sea-surface height, satellites are also capable of determining ocean wave heights, as was first demonstrated by Seasat in 1978.

There are two major differences between circulations in oceans versus atmospheres. The first is that ocean water is confined to move inside closed basins, which can be viewed as the most confining type of boundary condition encountered in geophysical FLUID DYNAMICS (GFD). In contrast, atmospheres are not closed off, although mountain ranges strongly influence wind patterns on Venus, Earth and Mars. Interestingly, the winds on the 'bottomless' gas-giant planets Jupiter, Saturn, Uranus and Neptune can be viewed as having rather benign boundary conditions since they suffer no mountain ranges or ocean basins, and they all sport east–west circulations.

Given the much greater mass of fluid in Earth's oceans than its atmosphere, it may come as a surprise that the second major difference is that the large-scale ocean circulation is driven by wind stress. The tail wags the dog, so to speak. Take away the atmosphere and its trade winds, and great currents like the North Equatorial Current, which travels the Pacific from east to west at latitude 15°N, would run down.

The winds do not merely push the water, except at the equator. If they did, the result would be confined to a shallow current that would head off perpendicular to the wind-stress direction. Away from the equator where rotational effects are important, two boundary layers called Ekman layers form in the wind-driven circulation, one in the atmosphere that is about 1 km thick and one in the ocean that is about 0.01–0.1 km thick and corresponds roughly to the ocean's well-mixed layer (the upper layer below which there is little response to passing storms). Significantly, the net (vertically averaged) atmospheric transport and net oceanic transport in these boundary layers are in opposite directions, and both are perpendicular to the surface wind-stress direction. This intricate three-dimensional picture emerges because the Coriolis acceleration balances the applied force, causing the bulk mass motion to be perpendicular to the forcing direction.

A horizontal variation in the wind stress leads to convergence or divergence in the oceanic Ekman layer, which leads to water being expelled from or drawn into the boundary layer, respectively. This process is called Ekman pumping. The amount of Ekman pumping is proportional to the curl of the wind stress. This vertical motion, a result of conservation of mass, deforms the constant-density surfaces underneath the boundary layer with the result that an ocean current is set up to achieve geostrophic balance. It is by this indirect mechanism that the wind drives the major ocean currents.

An important practical result of Ekman pumping is that surface water can be drawn away from a coastline by a persistent breeze that blows along the coast in the appropriate direction (depending on the hemisphere). To conserve mass, cold, nutrient-rich deep water (enriched by the settling of dead organisms) then upwells near the coast, which causes the phytoplankton level to rise dramatically, which in turn attracts great quantities of fish. Half the world's fishing relies on this sensitive mechanism. Given that Earth's coupled atmosphere–ocean system can exhibit more than one type of circulation with similar boundary conditions, there is strong economic motivation to discover and understand the feedbacks. For example, in normal years one-fifth of the world's catch is taken off the coast of Peru alone, but during El Niño years the equatorward wind is disrupted and Peruvian fishing declines dramatically.

Western boundary currents

Each ocean basin contains a strong current that runs along its western boundary, the Atlantic's Gulf Stream being a well-known example. In the Pacific, the North Equatorial Current turns into a western boundary current called the Kuroshio as it curves northward past Japan. In three dimensions, these fast moving rivers-in-the-sea have the appearance of writhing snakes.

Why not eastern boundary currents too, or alternatively, no lateral intensification at all? The answer depends on the shape of the planet itself: if Earth was flat instead of round, then there would be no Gulf Stream.

Because planets are round, the angle between their rotation axis and the local horizontal plane of motion changes with latitude. This means that the component of the Coriolis acceleration that is vertical (parallel with the local direction of gravity), which is the component that couples to horizontal motions, vanishes at the equator and is strongest, with opposite signs, at the two poles. The gradient of the Coriolis acceleration with respect to latitude is important because it appears in the equations of motion as a vorticity gradient. It is traditionally denoted β, hence the phrase 'beta effect'. The planetary vorticity gradient would be zero on a turntable Earth. On a sphere, β is zero only at the poles and is positive everywhere else (in both the northern and southern hemispheres); it is largest at the equator.

The spatial anisotropy that generates western intensification in ocean basins comes from the dispersion properties of a particular kind of wave, not a sound wave or a buoyancy wave, but a larger, more slowly moving wave called a Rossby wave (or planetary wave). Rossby waves

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

travel along gradients in the background vorticity (to be precise, gradients in the background potential vorticity, the ratio of the inertial-frame vorticity to any appropriate measure of vertical column stretching). On Jupiter and Saturn, the large amount of local vorticity in the alternating jet streams tends to dominate the beta effect, and consequently the local shape of the wind profile is important to the Rossby wave dynamics. But on Earth, the winds and ocean currents are much weaker and the beta effect governs the behavior of Rossby waves.

Rossby waves are strongly dispersive. A wave packet that reflects off the western boundary of a beta-dominated ocean basin and travels eastward is composed of small wavelengths only, because only small wavelengths have eastward group velocities. Conversely, a wave packet that reflects off the eastern boundary and heads westward is composed of only large wavelengths. Thus, the western boundary is a source of small scales and the eastern boundary is a source of large scales. All models of the western intensification include some type of damping that prevents the small scales generated at the western boundary from leaking into the ocean's interior, but the details do not affect the existence of the boundary current itself, which is a certain feature of ocean basins on a rotating, spherical planet.

Western boundary currents play a role in marine biology and climate. Meanderings of the Atlantic's Gulf Stream regularly pinch off to form drifting vortices called Gulf Stream rings that can transport phytoplankton over large distances. Regarding climate, the warmth of the Gulf Stream moderates the weather of the British Isles and western Europe. In fact, the Scottish islands to the north have a milder climate than the rest of Britain because of their close proximity to the Gulf Stream.

Atlantic Ocean

The Atlantic Ocean is separated from the Indian Ocean by a dividing line at the 20°E meridian between the southern tip of Africa and Antarctica. An association with Atlas in Greek mythology gives the Atlantic its name.

Many gulfs and seas ring the north and south Atlantic basins, including the Mediterranean Sea to the east, the Gulf of Mexico and Caribbean Sea to the west, the North Sea and several seas around Greenland and Canada to the north, and the Weddell Sea off the coast of Antarctica to the south. There is ample fishing in the waters over the continental shelves, which constitute about one-tenth of the ocean's floor.

The northern and southern halves of the Atlantic combine to form a shape like the letter S, and suggestively, the sea floor is dominated by a submerged mountain range with the same shape called the Mid-Atlantic Ridge. The main evidence that first won widespread acceptance of the theory of plate TECTONICS is the fact that the orientation of the magnetic field in magnetic rocks on the ocean floor, which record the direction of Earth's field when they cool below their Curie temperature, alternates in stripes that are mirror symmetric about the Mid-Atlantic Ridge. This demonstrates that the ridge is a sea-floor spreading center. In fact, the Atlantic is only 100 million years old, which makes it the youngest of the four oceans.

In the North Atlantic between the West Indies and the Azores is a stable region away from strong currents that is notable for containing large patches of floating seaweed. This is the Sargasso Sea, named after the Portuguese word for grape, sargaco, which describes the appearance of its most common seaweed, sargassum. The total area of the Sargasso Sea covers about 6% of the Atlantic (not counting adjacent seas). The free-floating seaweed does not hinder ships, although fishing is generally poor in the region because the water is warm and the phytoplankton concentration is reduced by a factor of three from average.

The sargassum plays host to many small marine animals, but there is one fish in particular whose association with this sea-within-a-sea is remarkable, the fresh-water eel. The manner in which eels in eastern North America, Europe and the Mediterranean reproduced had long been a mystery, until it was discovered in the 1920s that they *all* swim to the Sargasso Sea to spawn (details, such as how long the eels live after spawning, are not yet known). One cannot help wondering what sort of sagacious greeting a 30-year-old American eel gives to a 3-year-old Atlantic salmon as they swim past each other on these incredible early autumn reproductive migrations. How and why all this two-way traffic got started is a mystery (a similar puzzle is the migration of sea turtles in the South Atlantic, discussed below). The Pacific is home to many varieties of salmon, but eels spend their long lives inland, and in this respect it is notable that the Atlantic receives river outflow from a drainage area four times bigger than that received by either the Pacific or the Indian Oceans.

Meddies

A simple application of comparative PLANETOLOGY to the solar system's atmospheres and oceans generates the following question: why do Jupiter, Saturn and Neptune have long-lived storm systems, numbering hundreds, a dozen or so, and a couple, respectively, with the largest and most famous being Jupiter's GREAT RED SPOT, whereas Earth has only short-lived storm systems? Where are Earth's great spots? Hurricanes might fit the bill, except they turn out to be a uniquely terrestrial phenomenon that requires surface friction to operate. To be specific, one is looking for free, coherent high-pressure storm systems, called anticyclones, not cyclones like hurricanes.

In fact, EARTH'S ATMOSPHERE does have great high-pressure storm systems. One just needs to know where to look to find them. For example, a stubborn anticyclone called a blocking high often settles over North America or Russia for months at a time. Unfortunately, in satellite images these blocking highs are cloud-free and hence invisible, instead of cloud-filled and easy to visualize like spots on Jupiter. Blocking highs tend to divert rain storms from their normal paths and can cause droughts. One of several reasons they do not last long is simply that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Earth's atmosphere is so small in horizontal area that its atmospheric eddies jostle against each other incessantly, and so do not remain coherent for long. In constrast, anticyclones on the gas giants have one hundred times more surface area to move around in, which results in fewer eddy–eddy interactions and more predictable weather.

In contrast to its atmosphere, Earth's oceans turn out to be as roomy as the gas-giant atmospheres. The typical ocean eddy is two orders of magnitude smaller than the typical atmospheric eddy because oceans having much slower buoyancy waves, a consequence of much reduced density contrasts. One could go so far as to refer to Earth's oceans as 'basin-giants', to distinguish them from the 'gas-giant' atmospheres of Jupiter and Saturn.

If indeed the Atlantic Ocean was populated with Jupiter-like anticyclones spinning around underneath the surface, how would we know it? Flow visualization for Jupiter requires simply running a movie of the swirling cloud tops that have been photographed from afar by a spacecraft or a space telescope like Hubble, but not so for Earth's oceans. One discovers anticyclones in an ocean by hunting them down painstakingly with well-equipped ships. As it turns out, the hunting is good in the eastern Atlantic.

The reason is that extra-salty water from the Mediterranean Sea regularly breaks away into the eastern Atlantic and forms undersea lens-shaped anticyclones called Mediterranean water eddies, or meddies for short. Large numbers of meddies have been tracked over the last two decades. Individuals persist for several years and travel great distances, and collectively they are responsible for 20–40% of the salt flux coming into the Atlantic from the Mediterranean.

Subsurface SOFAR floats (SOund Fixing And Ranging floats) are used to determine the circulation structure inside meddies. They have core radii of about 20 km and turnaround times of about 4 days, with longer looping periods above and below the core. This is similar to the turnaround times of anticyclones on Jupiter. Because both meddies and Gulf Stream rings are likely to be mathematically similar to gas-giant anticyclones, there is a great deal of interest in both oceanography and planetary science to understand how environmental conditions like shear affect their drift speeds and directions, and how fluid nonlinearities allow them to remain coherent.

Labrador Sea

The general ocean circulation is called the thermohaline circulation because gradients in both temperature and salinity affect the density of seawater. This makes the dynamics of ocean currents particularly interesting. For example, a parcel of seawater can sink because it is cold and salty, or because it is warm and very salty, or because it is fresh and very cold. In addition, the diffusion rates of temperature and salinity are different, and this leads to a set of convective phenomena called double-diffusive convection, which includes unexpected structures like salt fingers and layered convection.

Seawater that becomes dense enough will sink to the bottom of the ocean and then spread out until it is constrained by the effects of planetary rotation. This is especially true in the Arctic Ocean and the poleward borders of the Atlantic Ocean where the formation of sea ice increases the salinity and hence the density of the already-cold underlying water. One accurate method for tracking the deep circulation is to monitor the progress of radioactive isotopes introduced into the oceans by early nuclear bomb tests.

An ongoing campaign that combines observations and numerical modeling from many research groups has the goal of understanding the formation of bottom water and the process of deep, penetrative convection. The Labrador Sea, which is the water between Greenland and Canada, is the unenviable location of the field laboratory for these mid-winter experiments. Often, 'chimneys' of downward convection are found there that have a significantly smaller horizontal scale than the atmospheric forcing above. They may occur preferentially in regions where the ocean stratification is weakened by the presence of pre-existing eddies (with cold centers), or be triggered by topography, or by some other small-scale, destabilizing mechanism.

Understanding the rate of formation of deep water is important for climate studies because it affects the rate in which carbon dioxide is exchanged with the atmosphere, and because deep water circulation contributes to the redistribution of heat from the equator to the poles. The process of Coriolis-influenced deep convection occurs not only in the coldest seas on Earth but also in stars and planets. For example, it may play a role in the formation of hot spots (large holes in the clouds) on Jupiter, like the hot spot that the GALILEO atmospheric probe sampled in December, 1995.

Pacific Ocean

Earth's oldest and largest ocean is the Pacific Ocean. It contains the deepest point on the planet's surface, the Mariana Trench (11.7 km below sea level), and contains the continent of Australia, the island arcs that make up Polynesia bordered by Hawaii, New Zealand and Easter Island, and about 25 000 islands all told. This single body of water covers one-third of Earth's surface. Away from island arcs, their trenches, seamounts and mid-ocean spreading centers like the East Pacific Rise, the bottom of the Pacific basin is basically flat with an average water depth of 4.3 km.

Ocean acoustics

The high opacity of salt water prohibits direct observations of the three-dimensional structure of Earth's oceans. But unlike light waves, sound waves travel readily in the ocean, and for this reason sound is used extensively by marine mammals for echolocation (high-frequency clicks) and communication (low-frequency songs). Physical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

oceanographers use acoustic tools as well, like inverted echo sounders (IESs), to exploit the linear relationship that exists between acoustic travel time and interesting properties such as temperature and the depth of the ocean's well-mixed layer. Acoustic tomography uses coded transmissions sent between arrays of half a dozen or so elements that last for around 30 s at a time.

It is known that dolphins coordinate with each other when hunting in groups. Killer whale pods appear to have dialects or accents. During the mating season, humpback whales sing repeating songs that can be heard more than 100 km away. The repertoire of marine mammals is impressive, spanning a range of clicks, rasps, buzzes, whistles, moans and haunting trumpets. Low-frequency man-made sounds ($\sim$120 dB), as from underwater drilling, have caused whales to change their migratory paths, so there are studies underway to determine the impact of military, commercial and scientific noise pollution on marine animals.

Sea monsters

In the process of gaining complete knowledge of the oceans and their inhabitants, we are losing something precious, namely the mystery of the sea and its ability, since antiquity, to fire our imaginations. No one expressed this mystery better than the early map-makers who filled uncharted expanses of ocean with illustrations of sea monsters. Besides the many species of whales in the oceans, there are at least two giants of the deep that have, over the centuries, greatly influenced the cultural view of what a sea monster looks like.

First, consider the world's largest invertebrate, and undoubtedly the source of the Norwegian kraken legend, the giant squid (*Architeuthis dux*). This mysterious behemoth has not yet been photographed alive, and virtually nothing is known about its social organization, but it is thought to live in total darkness a kilometer below the surface of the sea. Only a couple hundred giant squid have washed ashore around the world in the last century. The longest were 20 m long (the length of two school buses). The evidence suggests their typical lifespan is only about 5 years. To grow so large they must be terrific hunters that require an abundant food supply, as can be found in the Kaikoura Canyon off of New Zealand. Although no sunlight reaches the depths favored by the giant squid, most deep-sea fish are bioluminescent, and the giant squid has eyes the size of a human head, the biggest of any animal.

Giant squid have been found in the stomachs of sperm whales, which is the only known predator of the adult squid, and in turn, dead sperm whales with giant sucker marks have been found washed ashore. Even though the sperm whale's favorite food appears to be giant squid and it weighs 30 times more than its prey, the whale does not always hold the advantage in what is present-day Earth's most titantic one-on-one rivalry. Legends of kraken attacks are most likely based on the actions of giant squids that mistook small ships for whales.

Second, consider the world's largest and longest bony fish, and undoubtedly the source of the sea-serpent legend, the crested oarfish (*Regalecus glesne*). This incredible ribbonfish grows to over 10 m in length, and there are old reports of specimens twice that length, which rivals the length of the giant squid. The fish's name comes from a pair of elongated pelvic fins that rotate like oars as it swims. Its dorsal fin runs the length of its back like a red-gold mane, and the first ray at the head extends upwards in the manner of a long feather. When seen swimming on the surface, the oarfish has the appearance of an impossibly long, red, undulating sea serpent. Oarfish are known to inhabit the Mediterranean, the Atlantic and the Pacific, and like the giant squid, they can survive to depths of 1 km. They have large deep-sea eyes, a small toothless mouth, and a silvery appearance that is thought to attract prey. Studies of specimens washed ashore reveal that the fish's stomach runs almost half the length of its body, but amazingly, the oarfish can apparently survive losing the second half of its body to attack by predators, and then regenerate it.

Sightings of the oarfish occur only about once a decade, but unlike the giant squid, an oarfish has recently been photographed underwater. In May 1996, a pair of divers working near New Providence in the Bahamas encountered and photographed an oarfish rising from the depths. It was positioned vertically and was swimming by undulating its dorsal fin without body movement.

Throughout the ages, there have been many detailed eyewitness accounts of sea monsters that can be attributed to giant squid or oarfish (examples of the former include an old report by sailors that describes a whale being pulled beneath the waves by an enormous 'snake', and another that tells how a giant squid crawled up the bow of a ship before being shot off), but there are other stories, from equally credible sources, that appear to require alternate explanations. For now, the fact that the most common undersea vessel, the military submarine, carefully listens to but does not look at its surroundings, means that the reader is free to imagine (and if talent permits to illustrate) the size and shape of unknown deep-sea creatures that occasionally glide by.

Remote sensing

Providing a complementary perspective of the inner space of the sea from the relative safety of outer space, oceanography's newest observing platform is the orbiting satellite. The remote sensing of Earth's oceans is, in many respects, similar to any other astronomical observation, except for the 180° turn of the space-borne telescope and the inherent advantages of proximity to the subject. Space-based detectors have generated so much high-quality data in so short a time that modelers are struggling to keep pace. They are developing optimized data assimilation techniques like those used in meteorology to gain the greatest accuracy from the blending of observations with numerical models.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sea-surface temperature

In addition to sea-surface height and spectroscopic information, satellites produce maps of sea-surface temperature (SST). Surface temperatures range in approximately zonal bands from freezing at the poles to 20–30 °C in the tropics.

Changes in SST reflect changes in the transfer rate of energy between the air and sea. For example, hurricanes are known to develop in regions where the surface temperature is greater than 26 °C. Since most tropical depressions do not grow into hurricanes, an energetic threshold must be overcome. Strong winds produce large evaporation rates that increase the moisture content of the atmosphere, which enhances the convection and has a positive feedback on the winds. Modelers are working on developing a reliable means of predicting which disturbances will become hurricanes and what specific field observations are needed to do so.

Since December 1998, the Geostationary Operational Environmental Satellites (GOES) have been producing hourly SST estimates, with the ability to sample a position 48 times per day with a 4 km field of view. Sea-surface temperatures cannot be determined over cloudy areas. This disadvantage is turned into an advantage by observers who exploit the fact that SST maps show precisely the location of cloudy areas, which is useful for other studies.

Sea turtles

Satellites are not only having a positive effect on the study of physical oceanography, but also on the study of marine biology, as the following example illustrates.

In the middle of the Atlantic Ocean, just 80 km east of the Mid-Atlantic Ridge, at coordinates 7.95°S, 14.37°W, there is a solitary island less than 90 km^2 in surface area that World War II pilots were fond of quipping about, dryly, saying 'If you miss Ascension, your wife gets a pension'. Across half the ocean, about 2300 km west of Ascension Island, is the coast of Brazil in South America. For the past 100 million years since the birth of the Atlantic, plate tectonics has increased the distance between the Mid-Atlantic Ridge and Brazil by a few centimeters per year.

Based on tagging studies, a remarkable group of green sea turtles (*Chelonia mydas*) are known to use Ascension Island as their only nesting site. Females return every 2 to 4 years to their favorite beach on the island to lay eggs. No one can say for sure how long these turtles have been nesting exclusively on Ascension Island, but it is known that they have unique mitochondrial DNA that implies females do not trade places with other females. The species itself is many millions of years old.

What makes Ascension turtles remarkable is that their feeding grounds are along the coast of Brazil. The pioneer of sea turtle studies, Archie Carr, first suggested the idea that plate tectonics has slowly but incessantly stretched the turtles' foraging path from a reasonable foray millions of years ago into today's mighty migration.

The study of marine animal migrations has become a precise science with the advent of satellite tracking, which has been applied not only to green sea turtles but also to loggerhead turtles (*Caretta caretta*) and to elephant seals. As an example, six Ascension green turtles were outfitted with Telonics ST-14 satellite transmitters between April and July 1997 and then tracked as they migrated to Brazil. Of course, the 'if you miss Ascension' return trip would be even more interesting to track, but it is hard to predict when a turtle is going to leave Brazil, and current transmitters have relatively short lifetimes. The signal from one of the six turtles was lost after only 19 days, during which time she did not head off to Brazil but went on an eastward exploratory path that eventually looped back to the island. The other five turtles headed west towards Brazil in a no-nonsense fashion, with three of them separately following indistinguishable paths during the first 300 km. The data show that the trip to the coast takes the turtles about 6 weeks (in this particular study all of the transmitters failed just shy of Brazil).

The tracking results establish that turtles are able to swim in a straight line for weeks at a time with no islands or coastlines to guide them across the deep sea. They also like to punctuate their journey with occasional exploratory meanders that are then followed by the appropriate course corrections. The ability of sea turtles to achieve such precise navigation is not well understood. The satellite data show that most turtles tend to slow down somewhat at night, but the position of the Moon does not affect their navigation, and since they are known to be severely nearsighted when their heads are above water, they cannot focus on stars, which is one trick used by migrating birds. The leading hypothesis is that sea turtles are sensitive to both the intensity and inclination of Earth's magnetic field and can navigate it like a grid. This ability has been established for hatchling by controlled experiments. Earth's magnetic field is known to have changed some of its details in just the last 100 years, so the complete story has yet to be uncovered; however, the turtles will undoubtedly be patient with us as we try to come up to speed with them using the latest satellite technology.

Icy satellites: the case for an ocean

Of the nine planets, Earth's orbit alone occupies the narrow corridor where surface liquid water is stable by virtue of sunlight, which happens to occur at about one hundred solar diameters from the Sun. However, sunlight is not the only heat source in the solar system, there is also geothermal energy, and with it, the possibility of underground oceans.

There is no lack of potential sites for such an ocean in the solar system. Jupiter is the closest gas giant to the Sun and has four large moons, CALLISTO, GANYMEDE, EUROPA and IO, which are referred to collectively as the Galilean satellites in honor of their discovery in 1610 by Galileo, and which are easily seen with binoculars in the night sky. These are planets in their own right — in fact, to within 2% Callisto has the same radius as Mercury, while

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ganymede, the largest moon in the solar system (out of 62 and counting), is 8% bigger. Saturn's moon Titan is also 6% bigger than Mercury (much more will be learned about Titan's surface and interior during the Cassini mission).

But, what Mercury has in over abundance these moons lack, namely proximity to the Sun and its warming rays. Surface temperatures on Europa do not exceed 130 K and average 103 K, which is $-170\,^{\circ}$C. Given this, why are icy satellites viable candidates for bodies that might harbor subsurface seas?

Firstly, one needs a ready supply of H_2O in some phase, and the icy satellites are aptly named in this regard. Their $\sim$2–3 g cm^{-3} bulk densities imply a substantial icy component to their make up, moment-of-inertia gravity data place the ice predominantly in a 200–300 km outer shell (depending on satellite), and the spectroscopic signature of ice has long been detected on the surfaces of Callisto, Ganymede and Europa. No water has been detected on Io's surface, which is covered by sulfurous volcanic material, mostly SO_2 frost. Excluding Io, much of the effort in icy-satellite spectroscopy is to peer through the strong H_2O absorption lines and determine the composition of the small amount of dirt that is mixed in and to determine whether there are salts like $MgSO_4$ on the surface.

Secondly, one needs to melt the ice. At a pressure of 1 bar $= 10^5$ Pa, or approximately sea-level pressure on Earth, the melting temperature is 273 K. The lowest temperature possible for *pure* liquid water is 251 K at a pressure of 2080 bar, which is the pressure found a few hundred kilometers below the surface of an icy satellite. This depth is likely to be past the icy shell and into the rocky mantle on Europa. However, it is not hard to make antifreeze out of water by adding small amounts of ammonia or salt, which lowers the freezing point by many tens of degrees. It is also beneficial that when ice forms it leaves a large portion of these contaminants behind and this increases the brininess of the remaining liquid. If enough ammonia is present, the target minimum temperature for liquid water will be less than 200 K. The trick, then, is to raise the temperature from the surface value of about 100 K to the melting temperature as near as possible to the surface.

Sources of heating

To form liquid inside an icy satellite requires a source of heating. One source is the decay of naturally occurring radioactive ISOTOPES, called radiogenic heating. The amount is dependent on how much silicate-rock is mixed in with the ice, since it is this component that bears trace amounts of radioactive uranium, rubidium, samarium, potassium and thorium. It is reasonable to assume that these elements occur in proportions similar to what is measured in meteorites, so an estimate of the ice-to-rock ratio yields a heating rate.

Tidal dissipation has the potential to produce a great deal more heating inside an icy satellite than radioactivity, which is the case for Io and Europa. By TIDES and tidal dissipation is meant the flexing of the body, both its solid and liquid components, in response to changes in the external gravity field. Flexing a solid object repeatedly produces an impressive rise in temperature, as can be demonstrated by bending a paperclip back and forth several times. Io, the innermost Galilean satellite, experiences so much of this heating that it is more volcanically active than Earth, in spite of the fact that Io has nearly the same size and density as Earth's Moon.

The reason that Io and Europa have high tidal heating starts with the fact that their orbits are continuously forced to be slightly elliptical. This is the price they pay for cooperating with Ganymede to form a 4:2:1 orbital-period resonance called the Laplace Resonance. Io is the closest to Jupiter and hence the fastest orbiter, with a period of 42 h. The result of the three-body resonance is a forced eccentricity for Europa of about 0.01 and for Io of about 0.004. These seemingly small values turn out to be not so small when one carries out tidal stress calculations, as was first done by Peale, Cassen and Reynolds in 1979, who correctly predicted widespread volcanism on Io just a week before VOYAGER discovered the same.

Orbital eccentricity is converted into tidal flexing, and ultimately into heating that could be maintaining a subsurface ocean today, by two means. A small amount of flexing is generated by the change in the strength of the external gravity field as a satellite moves closer to and then farther away from its planet during each orbit. However, for synchronous rotators like Io and Europa, that is, satellites that keep the same face pointed towards their planet in the same manner that the Moon points towards Earth, most of the flexing arises because each satellite rotates around its axis at a constant rate like a clock but revolves at a variable rate described by KEPLER'S LAWS, moving along its orbit faster when closer to the planet and slower when farther away. This produces a constant nodding back and forth across the satellite–planet direction that is called 'optical libration'.

Estimates of Europa's present-day heat flux, based on theoretical evolution models, run as high as 5 K km^{-1}, which implies melted ice only 20 or 30 km below the surface. There are many details left to be worked out, for example solid-state convection of ice may replace conduction as the primary heat transfer mechanism for ice thicker than about 10 km. In any case, the strong tidal heating on Europa and Io cause these bodies to have two of the most intriguing and beautiful surfaces seen anywhere.

Magnetometer data

Is Europa's beauty only skip deep? There have been some tantalizing signals obtained by the Galileo Orbiter's magnetometer that suggest a subsurface ocean may be a reality. Jupiter's strong magnetic field is tilted about 10° from its rotation axis and rotates with the planet on a 10 h cycle, which means the background magnetic field in which the Galilean satellites are immersed changes its direction every few hours. If an icy satellite has a global conductive component, for example a briny interior

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sea, then an alternating magnetic field will be induced that could be distinguished from a permanent dipole by sampling the satellite's magnetic field at different phases of Jupiter's rotation.

Surprisingly, the outermost and most heavily cratered Galilean satellite, Callisto, has yielded a positive initial result to this experiment. The first two passes of Callisto by Galileo (passes C3 and C9; the notation signifies that the third and ninth orbits involved close encounters with Callisto) revealed that the satellite has a magnetic field with a surface strength of about 15 nT (4000 times weaker than Earth's surface magnetic field). Significantly, there was a flip in direction of the satellite's field that matched the large change in the rotational phase of Jupiter between the two flybys. The conductivity of a hypothetical ionosphere on Callisto is insufficient to provide the needed current path, but an ocean as salty as on Earth that is about 10 km thick, and at most only 100 km below the surface, can account for the data. This caught the Galileo team off guard, because Callisto has a much older surface than Europa and it is not part of the 4:2:1 orbital resonance.

At Europa, an induced-field model like the one that works for Callisto can fit the Galileo magnetometer data obtained from orbits E4, E11, E12 and E14 (magnetometer data were not acquired on pass E6). However, for these orbits the picture is complicated by Europa's interaction with the plasma caught in Jupiter's magnetosphere. In particular, large magnetic fluctuations in the data can be attributed to ions that are sputtered off of the satellite itself. Also, for these passes the rotational phase of Jupiter was not as distinct as it was for Callisto. To settle the question, the E26 orbit on 3 January 2000 was designed to make it clear whether or not the north pole on Europa moves with Jupiter's magnetic field. The experiment netted a positive result. Europa's north pole is moving and thus it is likely that there is a salty ocean somewhere in the outer 100 km of Europa. At the time of this writing the jury is still out for an induced field at Ganymede, but there are more encounters to come as part of the extended Galileo Millennium mission, which will also feature simultaneous observations of Jupiter by Cassini in December 2000 as it slingshots towards Saturn.

Europa's surface versus terrestrial sea ice

The reason that the spotlight has been on Europa is that Voyager discovered it has very few craters and no large mountains, implying that its surface has unusually ductile underpinnings. Additionally, Europa's darkest regions are all lighter in color than the lightest regions on Callisto and Ganymede—it is truly an ice world. But is it a sea-ice world like the Arctic?

Voyager images reveal many long, thin dark–light–dark markings called triple bands that criss-cross Europa's surface and some dark, wedge-shaped bands that imply more than 20 km of lateral spreading. On Earth, various processes modify the appearance of sea ice on horizontal scales ranging from kilometers down to the size of ice crystals, and it is instructive to make comparisons with Europa where possible. Caution is needed, however, because in some cases the formation mechanisms for ridges and jumbled terrain in sea ice are not completely understood, and so, even after accounting for the much lower surface temperature and gravity on Europa, one must always be conservative when attempting analogies. (Recall that before the 1960s, cloud-covered Venus was thought to harbor tropical rainforests by analogy with Earth's tropics; when in truth, the Venusian surface is hotter than the inside of a self-cleaning oven.)

Europa's triple bands often exceed 1000 km in length and this is much longer than any open cracks (called leads) or pressure ridges observed in Earth's sea ice. An important deformation mechanism on Europa is global-scale tidal flexing, whereas on Earth the shape of the ice is governed by local changes in temperature, sea currents, and the wind. Also, mature terrestrial sea ice reaches an average thickness of only 3 to 4 m, which is three orders of magnitude thinner than models suggest for Europa. The large difference in thickness, plus the large difference in temperature, prevent simple comparisons of ice-plate deformations on the two worlds.

The best geological hints of water inside Europa are revealed in high-resolution imagery taken by the Galileo orbiter; reports of all the Galileo imaging results are collected together in a special issue of *Icarus*, (vol 135, 1998). The spacecraft passed close enough to photograph Europa's surface at 20 m/pixel, which is 100 times better resolution than obtained by Voyager. A type of terrain dubbed 'chaos' was discovered that consists of $\sim$ 1 km and larger broken blocks that have been obviously rotated, translated and in some cases tilted, and appear to be frozen into an ice field. There is an active debate as to whether a global ocean is needed to explain the chaotic terrain. Alternative ideas include having only a local 'slush field', or having no liquid at all, just an underpinning of soft ice that, given enough time, could yield the same jumbled appearance.

Evidence for local cryovulcanism is found in some moderate-resolution Galileo images that show flowlike masses covering portions of ridges, and in some high-resolution images that show small, smooth areas that appear to be frozen flood plains. Efforts to discover current cryovolcanic activity, for example by looking for scattered light from erupting particles while Europa is in Jupiter's shadow, have yielded null results to date, but the Galileo observations are by no means exhaustive and the search continues.

Life on Europa?

Assume for the sake of argument that Europa contains an ice-covered salty ocean. Given what we know about the robustness of life on Earth, which routinely surpasses the expectations of even the experts, what are the chances that there is life on Europa today? Slim to none, is the present estimation, because of a lack of fresh oxidants.

In addition to water and nutrients, which we are assuming Europa has, life requires both a source of energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to grow and reproduce and a mechanism for harnessing that energy. On Earth, the primary source of energy is sunlight. The most that can be expected of sunlight at Europa is a weak illumination of the frigid surface and a penetration to depths of a few tens of meters, which is at best one-thousandth of the 20 km distance needed to reach an interior ocean.

Even so, energy is not the main problem for life on Europa, for two reasons. First, life on Earth is thought to have started before photosynthesis evolved. Simpler chemical reactions involving reduction–oxidation (redox) pairs of chemicals, kept in disequilibrium by deep-sea hydrothermal vents, were apparently an early means by which Earth's organisms harnessed energy. Second, present-day Earth has impressive biota associated with these hydrothermal vents (called 'black smokers' because of their appearance) that are disconnected from other biological systems and see no sunlight, and yet are thriving. By analogy, we can say that as long as Europa has something like hydrothermal vents, its inhabitants would not have to suffer from an energy crisis.

It is the chemistry needed to harness that energy that is the bigger problem, in particular, the oxidant budget. Present-day terrestrial vent organisms rely on redox reactions involving SO_4^{2-}, O_2 and CO_2, which are oxidants that are delivered to the deep sea from the ocean surface. Other energy-producing reactions exist that involve elemental sulfur or the production of methane (CH_4), but these are low-energy alternatives that do not meet the needs of terrestrial organisms. In other words, Earth's current vent biota is thriving because it is chemically connected to the surface.

By occasionally turning over some of the ice on Europa (after all its surface is relatively young) some oxidants could be transported to the ocean below. However, calculations suggest that, at best, the resulting energy flux is a thousand times weaker than what is needed to support ecosystems on Earth. If the low-energy alternatives are to be used, then without terrestrial-style plate tectonics to keep Europa's rocky mantle from becoming a reducing agent, mature hydrothermal vents will yield CH_4 instead of CO_2, in which case biological methane-producing reactions would be pre-empted and therefore shut down.

The study of terrestrial lifeforms living in extreme environments, and their implications for life on icy-satellites, is a young research field that continues to develop. Experience with hearty strains of bacteria provides an optimistic note for life on Europa, namely, that we should be guarded against being too guarded when the topic is the tenacity of life. For example, there exist unusual bacteria that gain energy by using hydrogen to reduce minerals, and these are perhaps not too different from what a Europan might be like.

With regard to Earth-like planets orbiting other worlds, NASA engineers are busy designing the Terrestrial Planet Finder, a space-based telescope system consisting of four 4 m wide mirrors. The goal is to be able to detect (or rule out) planets as small as Earth orbiting stars that are within 50 light-years, starting around 2010.

Lights in the dark—an anecdote

It has been pointed out, rightly, that astronomy is the study of outer space and oceanography is the study of inner space, and that we have only just begun to uncover the mysteries of both. There is much beauty in each that transcends scientific description. If the reader has ever been lucky enough to be at the sea on a still, clear night, then he or she has hovered between the blackness of both worlds, and at such a time, who does not sense a bit of mystery in the lights flickering above and below?

For the writer, such a moment came with a group of fellow geophysical fluid dynamics students late one August night as we were wading and whispering in calm seawater near Woods Hole, Cape Cod. We were enchanted by the discovery that any movement of our arms and legs through the water elicited green, glowing bioluminescence that trailed like pixie dust, and was punctuated by flashing disks of blue light from tiny, harmless jellies, while above the occasional trail of a falling star could be seen.

Apparently, such a transcendent moment can only be a fleeting one. The following night our attempt at a sequel was brought to an end by the appearance of a darkened patrol car called in by annoyed residents to contain the most recent outbreak of after-hours hotbloodedness on their beach.

Bibliography

Gaidos E J, Nealson K H and Kirschvink J L 1999 Life in ice-covered oceans *Science* **284** 1631–3

Gill A 1982 *Atmosphere–Ocean Dynamics* (New York: Academic)

Kivelson M G, Khurana K K, Stevenson D J, Bennett L, Joy S, Russell C T, Walker R J, Zimmer C and Polanskey C 1999 Europa and Callisto: induced or intrinsic fields in a periodically varying plasma environment *J. Geophys. Res.* **104** 4609–25

Luschi P, Hays G C, Del Seppia C, Marsh R and Papi F 1998 The navigational feats of green sea turtles migrating from Ascension Island investigated by satellite telemetry *Proc. R. Soc. Ser. B Biol.* **265** 2279–84

Pedlosky J 1987 *Geophysical Fluid Dynamics* 2nd edn (Berlin: Springer)

Turner J S 1973 *Buoyancy Effects in Fluids* (Cambridge: Cambridge University Press)

Timothy E Dowling

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Octans

(the Octant; abbrev. Oct, gen. Octantis; area 291 sq. deg.) A southern circumpolar constellation which lies between Indus/Pavo and Chamaeleon/Mensa, and includes the south celestial pole. It was named after the reflecting quadrant invented by John Hadley in 1731 by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2.

A small, inconspicuous constellation, the brightest star in Octans is ν Octantis, magnitude 3.7. There are no other stars brighter than fourth magnitude. σ Octantis, magnitude 5.5, is the closest naked-eye star to the pole, being situated just over 1° away. Interesting objects include λ Octantis, a close binary with yellow and orange (G8 and K0) components, magnitudes 5.6 and 7.3, separation $3.2''$.

There are no bright star clusters, nebulae or galaxies in Octans.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Odierna [Hodierna], Gioanbatista [Giovan or Giovanni Battista] (1597–1660)

Polymath, born in Ragusa, Sicily, became a priest, and was supported as chaplain and parish priest of the town of Palma di Montechiaro in Sicily by the Dukes of Palma. They made him court mathematician and gave him an apartment on the top floor of their palace for his astronomical observations. Observed the three comets of 1618–19, and later published *De Systemate Orbis Cometici* (1654). His studies on the satellites of Jupiter were intended to help the determination of longitude at sea, by timing their eclipses (Odierna timed the first such eclipse recorded, in 1652). His observations were published in *Medicaeorum Ephemeredes* (1656). He wrote on Saturn and astrology, and compiled a catalog (*De Admirandi Coeli Caracteribus* 1654) of some 40 nebulous stars, including 19 real nebulae found with a simple Galilean refractor of magnification 20. Believing that all nebulous objects were actually clusters of stars he classified his objects *Luminosae* (stars visible to the naked eye), *Nebulosae* (nebulous for the naked eye, but resolved in a telescope) and *Occultae* (nebulosity unresolved even in a telescope). His observations include an independent rediscovery of the Andromeda nebula (M31), and at least nine (probably 10) own true discoveries. Odierna made one of the earliest surviving drawings of the Orion nebula, M42. He was also interested in natural phenomena in meteorology, the rainbow and the spectrum, and he invented a microscope or camera obscura with which he studied the eyes of insects and the poison glands of vipers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oenopides of Chios (*c*. 500–*c*. 430 BC)

Greek natural philosopher credited by Eudemus with the identification of the ecliptic, within the zodiacal band discovered by Egyptian priests (Aëtius credits PYTHAGORAS), and determined its inclination to be 24 degrees (actually 23.5 degrees).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ohio State University Radio Observatory

The Ohio State University Radio Observatory (OSURO), Columbus, Ohio, started in 1951. Its Director, John Kraus, designed, built and operated radio telescopes including a 96-helix array and the 'Big Ear' which ceased operation in 1998.

Achievements include: the Ohio Sky Survey, which measured over 19 000 sources, over half previously undetected; the Ohio Specials, sources with unusual radio spectra, which led to the discovery of the two most distant objects in the universe; overlays to the Palomar Sky Survey prints; master lists of radio sources and non-stellar optical objects; and the 'Wow!' signal, the strongest narrowband signal detected (its origin is still unknown).

For further information see
http://www.bigear.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Okayama Astrophysical Observatory

The Okayama Astrophysical Observatory (OAO) is a branch Observatory of the NATIONAL ASTRONOMICAL OBSERVATORY, JAPAN. Its main facilities are 188 cm and 91 cm telescopes, equipped with newly built instruments with CCD/IR cameras (e.g. OASIS). OAO accepts nearly 300 astronomers a year, according to the observation program scheduled by the committee.

OAO conducts observations of various astronomical objects and phenomena, including the Sun, solar system objects, stars, galaxies and quasars. Its information and data archives are open to outside researchers and institutes through the network. OAO accepts around 30 000 public visitors a year.

For further information see
http://home.oao.nao.ac.jp/index-e.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Olbers, Heinrich Wilhelm (1758–1840)

Doctor, astronomer, born in Arbergen, Germany, enthusiast for astronomy. He discovered several comets. In 1800 he joined the team of 24 'celestial police', organized by FRANZ VON ZACH, who were to patrol a share of the zodiac looking for the planet missing (according to BODE'S law) between Mars and Jupiter. On New Year's Day 1801 PIAZZI discovered Ceres, and, in March 1802, Olbers discovered Pallas. Both planets were small, leading to the name 'minor planet', or asteroid, and Olbers suggested, to preserve Bode's law, that they were fragments of a full-sized planet. This in turn suggested that there might be other fragments and in 1807 Olbers discovered Vesta. In contemplation of more distant problems, he formulated *Olber's paradox*, asking the question 'why is the night sky not as uniformly bright as the surface of the Sun?' If the universe has infinitely many stars, any line of sight should end up on the surface of a star, and the sky should be as bright as the Sun. Olbers' paradox can be traced as far back to KEPLER (1610), HALLEY and DE CHESEAUX. The explanation is that the universe is expanding, so distant stars are red-shifted into obscurity, and young, so distant light has not reached us yet.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Olbers' Paradox

A paradox relating to the darkness of the night sky that was discussed by Heinrich Olbers (1758–1840) in 1826, but which had been noted previously by other scientists. Essentially the problem is that if the universe were infinite, static and uniformly populated with stars, the whole of the night sky should be as bright as the surface of a star (no matter in what direction one looked, eventually one would be looking at the surface of a star). The fact that the sky is dark at night clearly demonstrates that at least one of these assumptions is wrong.

It had been suggested that the paradox could be resolved if it were assumed that space is filled with matter that absorbs starlight (clouds of interstellar dust do, indeed, exist). However, the absorbed light would cause the dust to heat up and glow. We now know that stars are clumped into galaxies and are not themselves uniformly distributed through space. Even so, in an infinite static universe, a similar paradox could be posed for galaxies. The observed expansion of the universe resolves the paradox. If galaxies are receding with speeds proportional to their distances, the more distant they are, the greater the redshift of their radiation (wavelengths emitted by these galaxies are stretched to longer wavelengths). Beyond a certain range, the intensity of their radiation would be so greatly reduced by the redshift that they could not be detected (even if the universe were infinite). Likewise, if the universe has a finite age (as the Big Bang theory implies), there has not been sufficient time for light to reach us from galaxies beyond a certain range. The paradox is also resolved if the universe is finite in extent.

See also: Big Bang theory, cosmology, Hubble law, redshift.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oljato

An Apollo asteroid discovered by Henry Giclas in 1947, designated (2201) Oljato. About 2 km in diameter, it follows an orbit that shifts chaotically as a result of frequent close approaches to Earth and Venus. Its present orbit takes it from within the orbit of Venus at perihelion (0.67 AU) to the outer reaches of the main asteroid belt at aphelion (3.72 AU); its mean distance from the Sun is 2.18 AU (325 million km), period 3.20 years, inclination 3°, eccentricity 0.71. The rotation period is roughly 24 h. Oljato was lost after its discovery and not recovered until 1979. It is very high on the list of potentially hazardous asteroids, with a potential closest approach of less than 75 000 km. Its reflection spectrum bears no resemblance to that of any other asteroid, meteorite or comet so far obtained. It may well be an extinct comet nucleus.

See also: near-Earth asteroids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Olympus Mons

The largest volcano on Mars, and the largest in the solar system, centered at 18.4 °N, 133.1 °W. It coincides with the albedo feature visible from Earth known as Nix Olympica, the 'Snows of Olympus', after the mountain in Greece which in legend was the home of the gods. Olympus Mons rises to a height of 27 km above Mars's mean surface level, and in its greatest horizontal dimension measures 624 km. The central caldera has a diameter of about 90 km; from there, undulating flanks slope away at an average gradient of 4° toward the peripheral cliff, whose walls tower up to 6 km above the surrounding plateau. This huge volcanic edifice contains a hundred times the mass of the Earth's largest volcano, Hawaii's Mauna Loa, which is 120 km long and rises to just over 9 km above the Pacific Ocean floor. On the Earth, plate tectonics operates to carry a volcano away from the underlying source of magma. With no such movement on Mars, volcanoes could continue to grow for as long as magma was available.

See also: Mars: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Omega Nebula (M17, NGC 6618)

An emission nebula in the constellation Sagittarius, position RA 18^{h} 20.8^{m}, dec. -16° $11'$, also known as the Horseshoe Nebula and the Swan Nebula. All three names are suggestive of its shape. It measures $47'$ by $37'$, and was discovered by Philippe de Chéseaux in 1746.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ondrejov Observatory

The Ondrejov Observatory is located 20 miles from Prague in the village of Ondrejov. It was established in 1898 as a private observatory and donated to the state of Czechoslovakia in 1928. Since 1953 it has been part of the Astronomical Institute, Academy of Sciences of the Czech Republic; there are 40 astronomers.

The observatory's instruments include: 2 m stellar telescope, photometric telescope, multichannel solar flare spectrograph, solar magnetograph, solar telescope, solar radiospectrographs, meteor radar and photographic zenith tube. The observatory is part of the photographic European fireball network.

For further information see
http://www.asu.cas.cz.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Onsala Space Observatory

The Onsala Space Observatory (OSO), the Swedish National Facility for Radio Astronomy, operates two radio telescopes: the 20 m diameter millimeter and the 25 m decimeter wave telescopes at Onsala and the Swedish–ESO Submillimeter Telescope, SEST, in Chile.

OSO was the first European observatory to participate in high-resolution observations of radio sources using the very long baseline interferometry (VLBI) technique. This technique is used in astronomy and geodesy when transatlantic distances are measured with high precision. OSO is a pioneer of VLBI at millimeter wavelength, providing images of AGN with 50 microarcsec resolution. Important VLBI observations of hydrogen absorption reveal tori around AGN; VLBI shows methanol masers delineate rotating protostellar disks.

Research includes millimeter wave astronomy used for spectroscopy of interstellar molecules in the Galaxy, and external galaxies; molecular distribution in Magellanic Clouds measured with SEST. Several molecules have been discovered by Onsala scientists—molecules detected in distant galaxies in absorption against even more distant quasars leading to chemistry of the early universe ($z > 1$).

For further information see
http://www.oso.chalmers.se/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oort Cloud

The Oort cloud is a huge, spherical swarm of some 10^{12}–10^{13} comets surrounding the solar system and extending roughly halfway to the nearest stars. The cloud is named for the Dutch astronomer JAN OORT who first suggested its existence in 1950 in order to explain the orbits of the observed LONG-PERIOD (LP) COMETS.

The orbits of COMETS in the Oort cloud are so distant—typical APHELION distances are on the order of 30 000 to 60 000 AU—that they are perturbed by random passing stars, giant molecular clouds and the tidal field of the Galactic disk and core. These external perturbers randomize the orbits and occasionally send comets back into the planetary system where they can be observed. However, the comets in the Oort cloud likely did not form at these great distances from the Sun where the density of the solar nebula must have been exceedingly low. Rather, the comets likely formed as icy PLANETESIMALS in the region of the giant planets and were ejected to distant orbits by the growing proto-planets early in the solar system's history. Icy planetesimals that formed beyond the orbits of the giant planets remained in a low-inclination, low-eccentricity belt of comets, which is known as the KUIPER BELT. The Oort cloud and the Kuiper belt serve as long-lived storage reservoirs for comets in the outer solar system.

LP comets are defined as comets with orbital periods exceeding 200 yr. The definition is somewhat arbitrary and is based on the fact that we only have good orbital records over the past 200–300 yr. LP comets passing through the planetary region have highly eccentric, near-parabolic orbits that take them out to very large distances from the Sun. Indeed, many of the orbits appear to be hyperbolic, that is they appear to come from beyond the solar system, from interstellar space. This is shown in figure 1(*a*) where the orbits of the observed LP comets are plotted as a function of the inverse of their semimajor axes, $1/a$, which is proportional to their orbital energies. About one-third of the comets have negative values of $1/a$ which correspond to positive energies. These are the apparently hyperbolic orbits. The remainder of the LP comets have positive values of $1/a$, corresponding to negative or bound energies. These comets are in elliptical orbits around the Sun, although ones of exceedingly long period. These apparent or instantaneous orbits are known as 'osculating' orbits.

However, if one integrates the orbits backwards in time, to a point before the comets entered the planetary region, and then references the orbits to the barycenter (center of mass) of the solar system, rather than to the Sun, the distribution of orbital energies changes radically. These are known as the 'original' orbits. The distribution of inverse 'original' semimajor axes, $1/a_o$, is shown in figure 1(*b*). Almost all of the apparently hyperbolic orbits are shifted to bound, elliptical orbits. Also, the distribution is characterized by a large spike of comets at near-zero but positive (bound) values of $1/a_o$ and a low continuous distribution of comets in more tightly bound, less eccentric

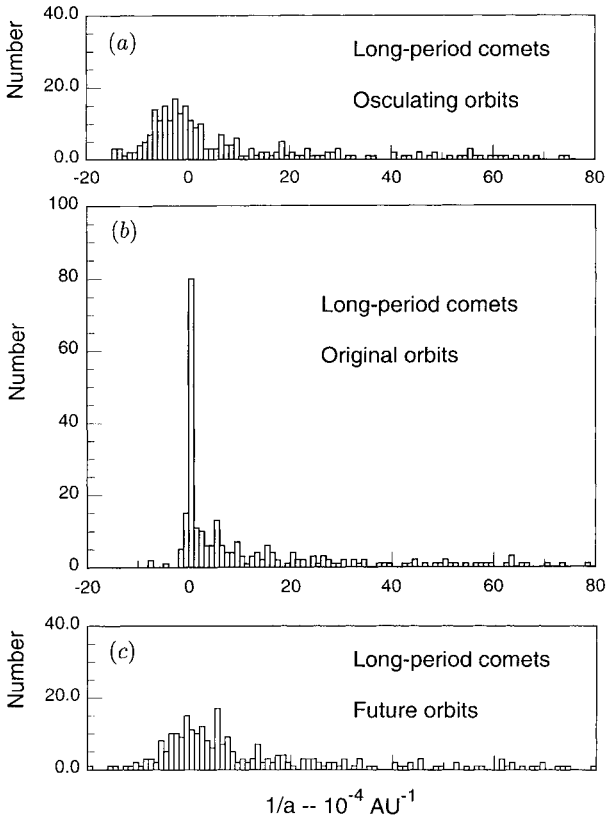


Figure 1. Distribution of inverse semimajor axes, $1/a$, for the observed LP comets: (*a*) for the instantaneous or 'osculating' orbit as the comet passes through the planetary system, (*b*) for the original orbit, integrated backwards in time to before the comet entered the planetary system, and (*c*) for the future orbit, integrated forward in time to after the comet has left the planetary system. The sharp 'spike' of comets at near-zero but bound energies in (*b*) is the 'dynamically new' comets from the Oort cloud. The few hyperbolic comets in (*b*) are likely the result of errors in their orbit determinations.

orbits. Comets in the spike have orbital semimajor axes between 10^4 AU and infinity and orbital periods greater than one million years. A small number of comets still appear to have hyperbolic orbits, but these have been attributed to small errors in their orbit determination and to jetting forces from the icy cometary nuclei, which also act to change the energy of the orbits. In the 1940s researchers showed that Jupiter would perturb the orbits of comets passing through the planetary system so as to spread them randomly in orbital energy, giving rise to the low, continuous distribution. However, what then was the explanation for the spike of comets?

Oort recognized that the spike had to be the source of the LP comets, a vast, spherical cloud of comets at distances greater than 10^4 AU from the Sun, but still gravitationally bound to it (the spherical nature of the Oort cloud is revealed by the distributions of orbital elements for the LP comets, which show that they appear from all directions on the sky). Oort showed that comets in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cloud are so far from the Sun that distant perturbations from random passing stars can change their orbits and occasionally send some comets back into the planetary system. On their first pass through the planetary system, random planetary perturbations, primarily by Jupiter, eject roughly half of the 'new' comets to interstellar space, while capturing the other half to more tightly bound, less eccentric orbits. Only about 5% of the new comets are returned to Oort cloud distances. On subsequent returns the comets continue to random walk in orbital energy until they are ejected, captured to a short-period orbit, collide with a planet or the Sun, or are destroyed by one of several poorly understood physical mechanisms. Analysis of the observed cometary orbits shows that the average 'dynamically new' LP comet entering the planetary system from the Oort cloud comes from an aphelion distance of $\sim 4.3 \times 10^4$ AU.

Interestingly, speculations by EDMOND HALLEY in his classic work on comets in 1705 could be interpreted as implying a distant comet cloud. Halley was only able to fit parabolic elements to the 24 comet orbits he derived, but he argued that the orbits would prove to be elliptic, writing,

> For so their Number will be determinate and, perhaps, not so very great. Besides, the Space between the Sun and the fix'd Stars is so immense that there is Room enough for a Comet to revolve, tho' the Period of its Revolution be vastly long.

It is possible to simulate the dynamical evolution of comets in the Oort cloud using computer-based Monte Carlo simulation models. These models include the effects of dynamical perturbations by the planets, random passing stars and nongravitational forces and of physical loss due to collisions, random disruption (splitting) and loss of volatiles (sublimation of ices). By 'tuning' such a model to fit the observed data in figure 1, some insight into the possible physical and dynamical loss mechanisms can be obtained.

The dynamical simulations suggest that ~65% of the LP comets are dynamically ejected from the solar system on hyperbolic orbits, 27% are randomly disrupted—10% on the first PERIHELION passage—and the remainder are lost by a variety of processes such as loss of all volatiles and collision with the Sun and planets. The average hyperbolic ejection velocity is only 0.6 km s^{-1}. Some comets may become unobservable owing to the formation of nonvolatile lag deposits, or 'crusts', on their surfaces, which would cut off further cometary activity. The average LP comet with perihelion distance less than 4 AU makes only five passages through the planetary region before arriving at one of the physical or dynamical end-states, with a mean lifetime of 6×10^5 yr between the first and last passages.

Another way to illustrate the dynamical evolution of LP comets from the Oort cloud is shown by the scatter diagram in figure 2, which plots each observed LP comet as a function of its perihelion distance and inverse original semimajor axis (omitting those comets for which only parabolic orbital solutions are available). The dynamically new, Oort cloud comets appear as a horizontal band of objects at near-zero $1/a_o$. As comets are perturbed by the giant planets, they random walk vertically in the figure (planetary perturbations of the perihelion distance, q, are typically small unless a very close approach occurs). However, if a comet random walks to a negative value of $1/a_o$, it escapes the solar system and does not return.

Cometary fading

Some researchers have proposed that the height of the low, continuous distribution of returning comets seen in figure 1(*b*) is too low relative to the height of the Oort cloud spike. This problem was also noted by Oort. These researchers suggest that comets fade sharply after their first appearance. However, the problem is a complex one involving both observational selection effects and poorly understood physical processes in comets. Oort cloud comets on their first perihelion passage are often anomalously bright at large heliocentric distances, and thus their probability of discovery is considerably enhanced. Indeed, as seen in figure 2, most observed LP comets with perihelion distances greater than 3 AU are Oort cloud comets. Suggested explanations for this anomalous brightness include a veneer of volatiles accreted from the interstellar medium and lost on the first perihelion passage near the Sun, blow-off of a primordial cosmic-ray-processed nucleus crust or the amorphous-to-crystalline water ice phase transformation that occurs at about 5 AU inbound on the first perihelion passage. When these Oort cloud comets return, they are not observed unless they come within 3 AU of the Sun, where water ice can begin to sublimate at a sufficient rate to produce an easily visible coma (see COMA, COMETARY). The failure to observe returning LP comets with perihelia greater than 3 AU must be an observational selection effect, as there is no recognized physical and/or dynamical mechanism for preferentially removing them.

On the other hand, physical 'fading' of comets is not well demonstrated. Suggestions of fading in some SHORT-PERIOD COMETS have largely been explained by changes in instrumentation and observing techniques. Observations of three well-known intermediate-period comets (those with periods between 20 and 200 yr), 1P/Halley, 109P/Swift–Tuttle and 55P/Tempel–Tuttle, have been traced back up to two millennia in human records. The observations suggest that these comets have remained remarkably constant in maximum brightness during that time and have not faded appreciably (if the comets had been brighter in the past then they should have been detected on additional past apparitions; this is not seen).

A physical mechanism that might explain fading is the formation of nonvolatile lag deposits on the surfaces of COMETARY NUCLEI. As cometary ices sublimate, the evolving

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

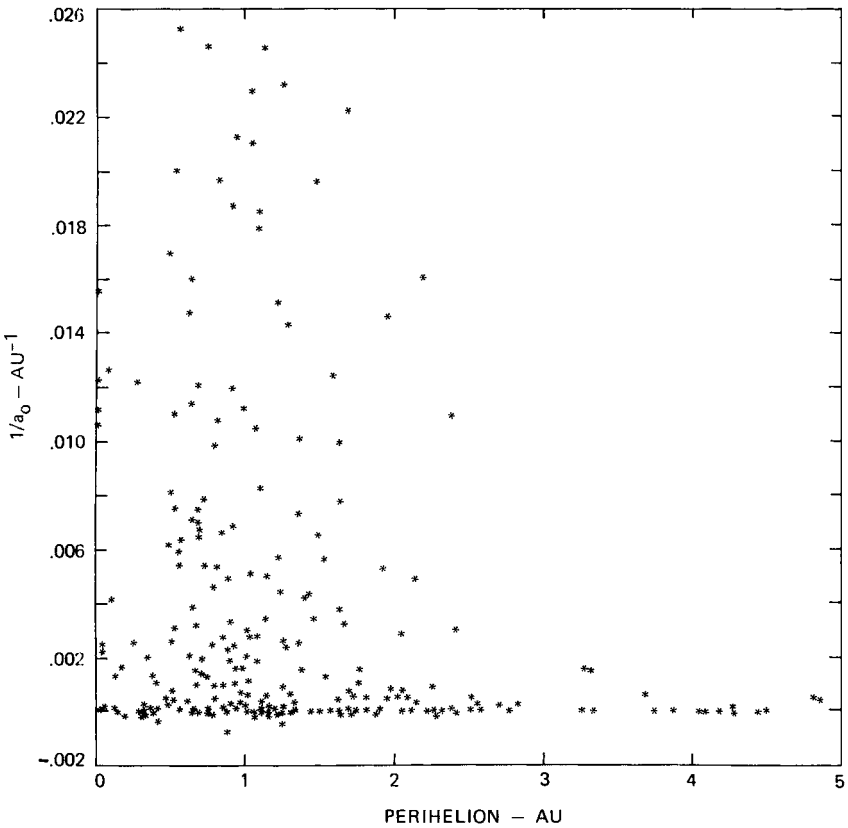


Figure 2. Scatter diagram in perihelion distance and inverse original semimajor axis for the observed LP comets. The horizontal band of comets at near-zero $1/a_o$ are comets making their first perihelion passage from the Oort cloud. Comets diffuse up and down in the diagram as a result of planetary perturbations, primarily by Jupiter (from Weissman 1996).

gases carry grains of dust and ice into the cometary coma. However, large grains may be too heavy to be lifted off the nucleus surface by the evolving gases, and thus would form a lag deposit. The effect of this lag deposit is to insulate the ices below from solar heating, thus decreasing the sublimation rate and the resulting observable cometary activity. Thermal models have suggested that a layer of nonvolatile grains only a few centimeters thick could completely insulate the underlying icy-conglomerate material from solar insolation. Thus comets may slowly 'turn off' as these lag deposits build up on their surfaces. Observed outbursts in comets have been attributed to the sudden blow-off of surface crusts as the pressure of evolved gases builds up beneath them.

It should also be recognized that random disruption, one of the physical loss mechanisms mentioned above, can mimic the apparent effect of cometary fading, removing a fraction of the LP comets on each perihelion passage. Disruption events which do not totally destroy the cometary nucleus may still result in smaller nuclei less capable of producing visible coma, i.e. faded comets.

Oort cloud perturbers

Since first proposed in 1950, Oort's vision of a cometary cloud gently stirred by perturbations from distant passing stars has evolved considerably. Additional perturbers have been recognized: giant molecular clouds (GMCs) in the Galaxy, which were unknown before 1970, and the Galactic gravitational field itself, in particular the tidal field of the Galactic disk. GMC encounters are rare, occurring with a mean interval of $\sim 3 \times 10^8$ yr, but can result in major perturbations on orbits in the Oort cloud.

The Galactic field sets the limits on the outer dimensions of the Oort cloud. The cloud is a prolate spheroid with the long axis oriented toward the Galactic nucleus. Maximum semimajor axes are about 10^5 AU for direct orbits (relative to Galactic rotation) oriented along the Galactic radius vector, decreasing to about 8×10^4 AU for orbits perpendicular to the radius vector, and increasing to 1.2×10^5 AU for retrograde orbits (opposite to Galactic rotation).

In addition, random stars will occasionally pass directly through the Oort cloud, ejecting a substantial number of the comets and severely perturbing the orbits of others. A star passage drills a narrow tunnel through the Oort cloud, ejecting all comets within a radius of $\sim$200 AU, for a 1 solar mass star passing 10^4 AU from the Sun at a velocity of 40 km s^{-1}. Over the history of the solar system, $\sim 10^4$ stars have passed within 10^5 AU of the Sun,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ejecting about 10% of the Oort cloud population. Although the ejected comets will all be positioned close to the path of the perturbing star, comets which are thrown into the planetary system in a 'cometary shower' can come from any location in the spherical Oort cloud.

Galactic tidal perturbations result from the different distances of the Sun and a comet in the Oort cloud above the Galactic plane and (to a lesser extent) their different distances from the Galactic nucleus. Galactic tidal perturbations peak for orbits with their line of apsides at Galactic latitudes of $\pm 45^\circ$ and go to zero at the Galactic equator and poles. The distribution of Galactic latitudes of the aphelion directions of the observed LP comets mimics that dependence. Although a lack of comet discoveries near the Galactic equator could be the result of observational selection effects (e.g. confusion with Galactic nebulae), the lack of comets near the poles appears to confirm the importance of the Galactic tidal field on the Oort cloud.

The Galactic tide acts to step the cometary perihelia in a regular fashion, as compared with the random walk nature of stellar perturbations. Thus the tide results in comets being brought into the observable region more efficiently, making it easier to overcome the dynamical barrier that Jupiter and Saturn present to cometary diffusion into the inner planets region.

The Oort cloud perturbers discussed above are all evaluated assuming the Sun's current position in the Galaxy, its current velocity relative to the local standard of rest (LSR) and the current perturbing environment of stars, GMCs and the Galactic gravitational field. However, the Sun's motion has likely varied considerably over the history of the solar system, precisely because of random encounters with GMCs. Also, there is some evidence that the Sun and solar system may have formed closer to the Galactic center, at about 6 kpc, as compared with its current estimated distance of 8.5 kpc.

The Sun's past motion can only be reconstructed statistically, and thus one can only give probabilities as to the perturbation history of the Oort cloud from stars, GMCs, and the Galactic tide. An interesting feature of the Sun's current motion is that the Sun moves anomalously slowly relative to neighboring stars, as compared with stars with similar ages. The Sun's current velocity is only $\sim$16–20 km s^{-1} relative to the LSR, as compared with an rms velocity dispersion of $\sim$42 km s^{-1} for G-type stars. The effect of this lower velocity is to decrease the rate of encounters with stars and GMCs but to increase the perturbing effect of those encounters that do occur. Also, the low velocity restricts the Sun to moving in a low-eccentricity orbit around the Galactic center, and an orbit that does not carry it very far out of the Galactic plane. The Sun's current Galactic orbit has it moving 'inward' towards Galactic periapsis and 'upwards', having recently passed through the Galactic plane.

The inner Oort cloud

As a result of this better understanding of the roles of Oort cloud perturbers, it is now estimated that the mean dynamical lifetime of comets in the cloud is only about 60% of the age of the solar system; some researchers have estimated even shorter lifetimes. The Oort cloud must somehow be replenished, either by capture of comets from interstellar space or from a more populous inner Oort cloud reservoir, comets in orbits closer to the Sun that are pumped up to replace the lost comets.

Cometary capture is a highly unlikely process because a three-body gravitational interaction is required to dissipate the excess hyperbolic energy. The probability of capture is proportional to V_∞^{-7}, where V_∞ is the hyperbolic excess velocity. Capture is possible at encounter velocities $\leq$1 km s^{-1}, but is highly unlikely at the Sun's velocity of $\sim$16–20 km s^{-1} relative to the LSR.

On the other hand, Monte Carlo simulations have shown that the dynamical evolution of icy planetesimals formed in the giant planets' zone would naturally lead to an inner Oort cloud. The icy planetesimals would be perturbed by the forming proto-planets into eccentric long-period orbits with semimajor axes of several times 10^3 AU or more. Stellar and Galactic perturbations at aphelion would raise the perihelia of such orbits to heliocentric distances greater than Neptune's semimajor axis, detaching the comets from planetary perturbations. In this manner, an inner Oort cloud reservoir, with a population about 2–5 times that of the outer cloud, could be formed.

Note that the difference between the inner and outer Oort clouds is one of definition only. The two are really one continuous cloud of comets, formed coevally and evolving under the same combination of external perturbers. The difference is defined by the different dynamical regimes. The outer Oort cloud is 'dynamically active'; perturbations on the orbits of comets in the outer cloud are sufficient to throw the comets back into the planetary region where they can be observed. The comets in the inner cloud are dynamically less active, unless the cloud undergoes a major perturbation. The transition between the inner and outer Oort clouds is generally taken to be at semimajor axes of $\sim 10^4$ AU, or at heliocentric distances of $\sim 2 \times 10^4$ AU. As comets are stripped away from the outer Oort cloud by close stellar and GMC encounters, the same perturbations will pump up the orbits of comets in the inner Oort cloud to replace them.

An example of a Monte Carlo simulation of Oort cloud formation is shown in figure 3 for four different times in the solar system's history. Initially the comets are confined to the ecliptic plane (which is inclined approximately 62° to the Galactic plane). As perturbations drive the orbits outward, their inclinations relative to the ecliptic plane slowly increase. At the end of 4.5×10^9 years, the cloud is essentially randomized in both inclination and eccentricity beyond about 6×10^3 AU from the Sun. However, inward of that distance the cloud remains flattened toward the ecliptic plane and the orbits remain highly eccentric.

Approximately 40–60% of the original comets ejected to the Oort cloud (both inner and outer) have been lost over the history of the solar system. Comets are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

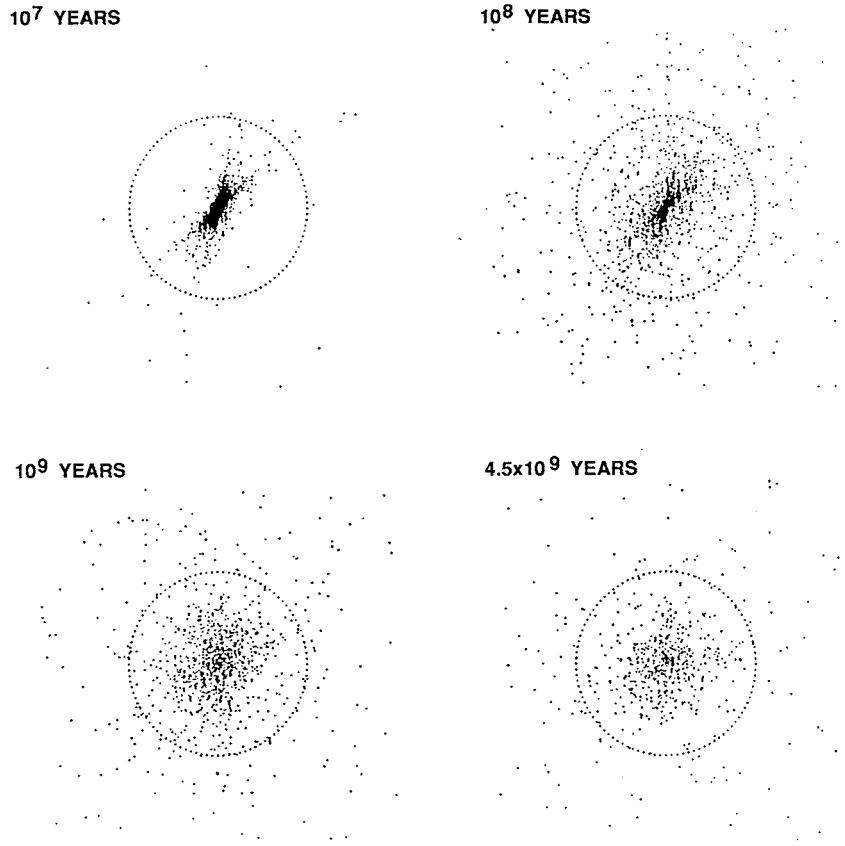


Figure 3. Dynamical evolution of a hypothetical cloud of comets ejected out of the giant planets' zone, at several times during the history of the solar system, under a combination of Galactic, stellar and planetary perturbations (projected onto a plane perpendicular to the Galactic plane). The dotted circle is at a radius of 2×10^4 AU, the approximate boundary between the inner and outer Oort clouds (from Duncan *et al* 1987).

lost by direct ejection due to close stellar passages, by diffusion back into the planetary region where Jupiter and Saturn perturbations will eject them hyperbolically or where they will be physically destroyed and by the slow growth of some orbits to distances greater than the Sun's gravitational sphere of influence, $\sim 2 \times 10^5$ AU (1 pc).

The population, mass and angular momentum of the Oort cloud

Estimates of the population of the Oort cloud can be made by comparing the results of dynamical simulations with the observed flux of comets in the planetary region. To account for the observed flux, Oort estimated that the population of the cometary cloud was 1.9×10^{11} objects. More recent dynamical models have produced somewhat higher estimates, by up to an order of magnitude. These result in part from higher estimates of the flux of LP comets through the planetary system and in part from a recognition of the role of the giant planets in blocking the diffusion of cometary orbits back into the planetary region. Comets perturbed inward to perihelia near the orbits of Jupiter and Saturn will likely be hyperbolically ejected before they can diffuse to smaller perihelia and be observed. Thus, the terrestrial planets' region is under supplied in LP comets as compared with the outer planets' region.

About ten new LP comets are discovered each year, with about two or three of those having perihelia inside the Earth's orbit. However, researchers have shown that many comets are missed because of observational selection effects. This is particularly true for comets that reach their greatest brightness on the opposite side of the Sun from the Earth. Estimates of the incompleteness range from 50% to 95%. Using these numbers, the best current guess for the population of the outer, dynamically active Oort cloud is about 1×10^{12} comets. The population of the inner Oort cloud is estimated at 2–5 times that of the outer cloud. These population estimates refer to comets brighter than absolute magnitude $H_{10} = 11$. Absolute magnitude 11 corresponds to a nucleus mass of $\sim 4 \times 10^{15}$ g, or a nucleus radius of 1.2 km, assuming a density of 0.6 g cm^{-3} (although there is not good agreement on the cometary mass–brightness relationship).

The population estimates depend on the assumption

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

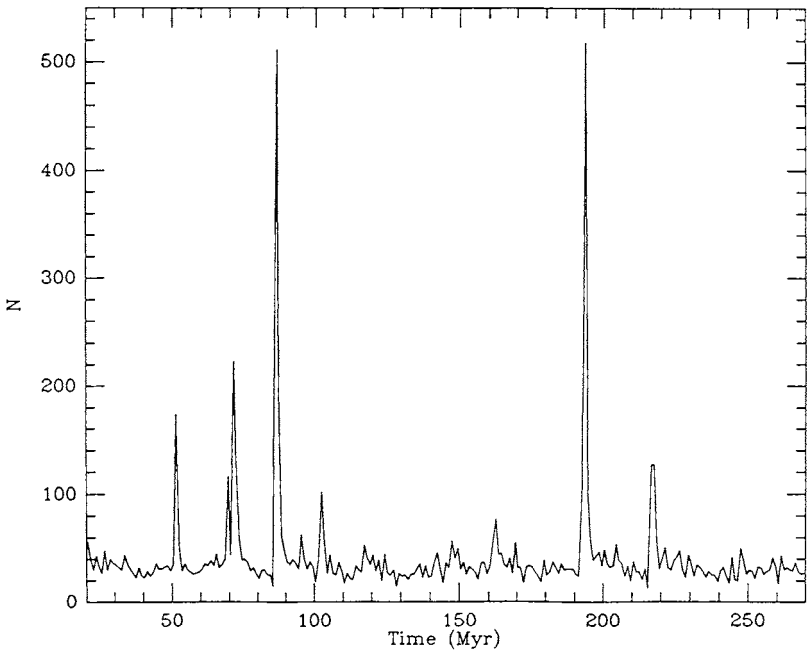


Figure 4. Number of new LP comets from the Oort cloud entering the terrestrial planets' region, $q < 2$ AU, versus time, based on a Monte Carlo simulation that included random passing stars and Galactic tidal perturbations. The large spikes are comet showers due to random stars penetrating the Oort cloud (from Heisler 1990).

that the currently observed LP comet flux through the inner planets region is equivalent to the long-term average flux. If the current flux is enhanced owing to a recent perturbation on the Oort cloud, then the population estimate for the cloud is too high, and vice versa. However, it has been pointed out that the Galactic latitude dependence of the observed LP comet aphelion directions would not be present if most comets seen were the result of a recent major perturbation on the cloud by a close stellar passage or by a GMC. Also, researchers have shown that the $1/a_o$ distribution does not exhibit any indication of a recent major perturbation on the comet cloud. Thus, the Oort cloud population estimates above may be lower limits.

The mass of the Oort cloud is highly uncertain because both the nucleus size distribution and the bulk density of cometary nuclei are not well known. Estimates of the density of the nucleus of COMET HALLEY range from 0.2 to 1.2 g cm^{-3}, with an even larger range of uncertainties. Measurement of the bulk density of recovered interplanetary dust particles which are believed to be of cometary origin finds values of ~0.7–1.9 g cm^{-3}. Researchers found a bulk density of 0.6–1.1 g cm^{-3} for the progenitor nucleus of comet Shoemaker–Levy 9, based on computer simulations of the tidal disruption of that comet.

The cometary mass distribution is even more uncertain. Several attempts to derive the mass distribution have been based on the observed distribution of cometary magnitudes (including coma), H_{10}, after correction for observational selection effects. Another method is to estimate the size of cometary nuclei from their observed brightness at large distances from the Sun, when they presumably are inactive: this method also requires one to assume an albedo for the nucleus surface; actual cometary albedos have only been measured in a few cases. A third method has been to construct a size or mass distribution from estimates of the population of the Kuiper belt in various size ranges. Using these various methods, a plausible value for the average nucleus mass is $\sim 4 \times 10^{16}$ g. Assuming a bulk density of 0.6 g cm^{-3}, the average nucleus would have a mean radius of 2.5 km. Using this value, a current best guess for the total mass of the Oort cloud is $\sim$(1.2–2.4) $\times$ 10^{29} g or ~20–40 Earth masses, assuming an outer Oort cloud population of 10^{12} comets and an inner Oort cloud population of (2–5) $\times$ 10^{12} comets.

Assuming that population and mass, and distributions of cometary orbits from dynamical simulations, the total angular momentum of the Oort cloud at the present time is (3–6) $\times$ 10^{50} g cm^2 s^{-1}. The original angular momentum of the comet cloud at the time of formation was approximately a factor of 5 less than the present value; angular momentum has been pumped into the cloud by stellar and GMC perturbations. The original angular momentum is ~3 times that of Uranus and Neptune combined, which is consistent with estimates of the total mass of icy planetesimals originally in that zone.

Temporal variations in the cometary flux

The cometary flux into the planetary region is expected to vary as a result of major perturbations on the Oort

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cloud and as a result of the variation in the Galactic tidal perturbation. The latter results from the solar system's harmonic motion above and below the Galactic plane, which currently carries the planetary system ∼75 pc out of the plane. The period of the oscillation is ∼(62–68) × 10^6 yr. This is expected to cause the cometary flux to vary sinusoidally by a factor of 4 over that period, with the maximum flux occurring just after passage through the Galactic plane. However, the role of stellar perturbations during the oscillation has not been evaluated and they may act either to enhance the variations in the flux or to actually mitigate the variations. The solar system has just passed through the Galactic plane in the last few million years, so the current steady-state flux is likely at a local maximum.

Random close approaches by stars can cause more extreme, although short-lived, changes in the cometary flux. A computer-based simulation of the flux of dynamically new comets perturbed to perihelion distances <2 AU is shown in figure 4. The large spikes are the result of close stellar passages which actually penetrate the Oort cloud to <10^4 AU.

These extreme increases in the cometary flux caused by penetrating stellar passages are of particular interest. A $1M_\odot$ star passage at 20 km s^{-1} at 3 × 10^3 AU from the Sun would perturb a shower of ∼5 × 10^8 comets into Earth-crossing orbits, raising the peak flux by a factor of 300 or more, and lasting (2–3) × 10^6 yr. Contributing to the very high flux rate is the fact that the shower comets from the inner Oort cloud start from shorter-period orbits than outer Oort cloud comets, and returning comets tend to be perturbed to even shorter-period orbits. They thus make many returns in a relatively brief period of time. The temporal profile and fraction of surviving comets for a major cometary shower are shown in figure 5.

Fortunately, major cometary showers as a result of deep, ∼3 × 10^3 AU, penetrating stellar encounters are rare, occurring perhaps once every (3–5) × 10^8 yr. Cometary showers should also occur with a similar frequency owing to random encounters with GMCs, but with perhaps an order of magnitude less total flux into the planetary region. Lesser showers from more distant, but still penetrating stellar passages at heliocentric distances ∼10^4 AU occur more frequently, on the order of every (3–5) × 10^7 yr. Cometary showers have also been suggested as a possible cause of biological extinction events on the Earth. LP comet impacts, including comet showers, are expected to account for approximately 10–25% of the craters larger than 10 km diameter on the Earth and Moon.

The suggestion that biological extinction events on the Earth repeat with an approximately 26 Myr period led to several hypotheses in the 1980s that each invoked cometary showers as the cause of the extinctions. These hypotheses involved (1) a dwarf companion star to the Sun in a distant, eccentric, 26 Myr period orbit with its perihelion deep in the Oort cloud, (2) a 10th planet circulating in a highly inclined, precessing orbit at about 150 AU from the Sun, at the inner edge of the inner Oort cloud, (3) the solar system's epicyclic motion above and below the Galactic plane with a half-period of 32–34 Myrs, with GMC encounters at Galactic plane crossings. There is an apparent coincidence between Galactic plane crossings by the solar system and some terrestrial extinction boundaries.

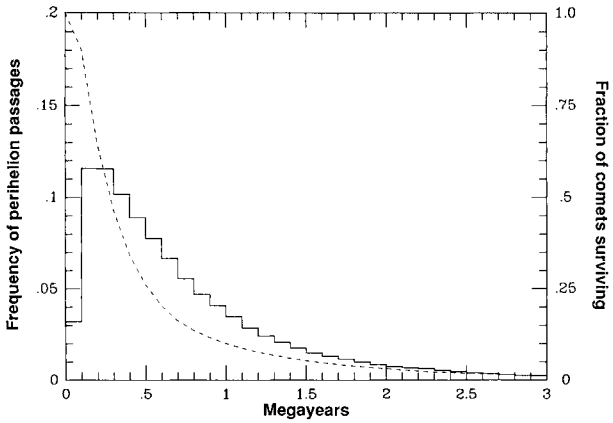


Figure 5. Dynamical evolution of a shower of comets from the inner Oort cloud due to a close, penetrating stellar passage at 20 km s^{-1} at 3 × 10^3 AU from the Sun. The solid histogram is the relative number of comets crossing the Earth's orbit versus time; the dashed curve is the fraction of the original shower comets still evolving in the system. On the order of 5 × 10^8 comets (brighter than $H_{10} = 11$) are expected to be thrown into Earth-crossing orbits by the $1M_\odot$ star's passage (from Hut *et al* 1987).

However, a variety of dynamical problems have been identified with each of these hypotheses, and no evidence in support of any of them has been found. As a result, periodic comet shower hypotheses have not gained wide acceptance. Questions have also been raised about the reality of the periodicity in the fossil extinction record.

Physical processing of comets in the Oort cloud

It had generally been thought that the Oort cloud is a fairly benign storage location for the LP comets. The typical spacing between comets in the outer cloud is ∼15 AU, and in the inner cloud it is ∼1 AU. The typical temperature is that of interstellar space, <10 K. However, it is now recognized that a variety of processes combine to act on the cometary nuclei, in particular their surface layers, over the history of the solar system. These include (1) irradiation, sputtering, and polymerization by Galactic cosmic rays, (2) heating by passing stars and nearby supernovae, (3) gardening by debris impacts and (4) the accretion of interstellar dust and gas and the accompanying erosion by hypervelocity dust impacts. In addition, the interiors of the nuclei may have been warmed by long-lived radionuclides, and even melted if short-lived radionuclides such as ^{26}Al are there in similar abundances to those found for some primitive carbonaceous meteorites.

The combined effects of these competing processes are not well understood. They may lead to the development

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a permanent nonvolatile crust over the nucleus surface before it ever enters the planetary system, or they may simply result in a heterogeneous collection of unbound, highly processed materials in a well-gardened nucleus regolith. Surface heating may cause both loss and inward diffusion of the more volatile cometary species, while internal heating will also mobilize volatile species and could conceivably result in the nucleus core transitioning from amorphous to crystalline ice, or even perhaps being melted and refrozen.

Extra-solar Oort clouds and interstellar comets

Presumably, the same processes that led to the formation of planetesimals and their ejection to form the Oort cloud in our planetary system can also occur around other forming stars. However, researchers have suggested that stars younger than about 10^7 yr ($M_* > 7M_\odot$) likely do not have sufficient time to form Oort clouds during their main sequence lifetimes. Also, the formation of Oort clouds is likely dependent on the presence of giant planets at distances greater than 3 AU from the central star (where ices could form) in order to eject enough comets to Oort cloud distances.

Because the typical dimensions of Oort clouds are so large, ~0.1–0.5 pc, it may be possible to actually detect and resolve such comet clouds. Astronomers have searched for excess thermal emission from cometary dust around nearby stars in data from the Infrared Astronomical Satellite. Their results were negative. However, this was expected, since dust at Oort cloud distances will be very cold ($T \approx 10$ K at 10^3 AU from the Sun), and radiation pressure and collisions with interstellar dust and gas should rapidly sweep fine dust from the comet clouds.

Astronomers have also proposed searching for Oort clouds around RED GIANT STARS. These stars are sufficiently luminous that the comets in their Kuiper belts, if they had them, would be actively sublimating at rates comparable with the gas production rates of comets at 1 AU in our own solar system. Observed OH-IR stars may be an example of this phenomenon.

It is interesting to speculate on the fate of the many comets ejected to interstellar space in forming the Oort cloud and over its history. Dynamical ejection is the most common loss mechanism for comets in the cloud, either due to close stellar and GMC perturbations or as a result of Jupiter perturbations during passage through the planetary system.

No comet on a clearly interstellar trajectory has been observed passing through our planetary system. This fact sets an upper limit on the space density of interstellar comets of $\sim 4 \times 10^{12}$ comets pc^{-3}. For comparison, this is ~100–300 times less than the density of material in the solar neighborhood, $\sim(0.07\text{–}0.18)M_\odot$ pc^{-3}, so interstellar comets cannot contribute significantly to the 'missing mass' in the galaxy. Note also that this value is about half the mean density of comets in the outer Oort cloud, assuming a population of 10^{12} comets in a sphere of radius 10^5 AU centered on the Sun. Thus, the limit is not very strict.

It is also possible to compare the limit above with the estimated space density of interstellar comets, if one assumes that all stars produce cometary clouds. Dynamical models estimate that between 2 and 20 times as many comets are ejected by the proto-planets as are placed in the Oort cloud. Another factor of 2–3 comes from the comets lost from the Oort cloud over the history of the solar system. Thus, taking a nominal current Oort cloud population of 3×10^{12} comets, the solar system has ejected $\sim 10^{13}$ to 2×10^{14} comets to interstellar space. Taking a mean volume per star in the solar neighborhood of ~12 pc^3 and assuming that all stars produce comet clouds, the predicted space density is 10^{12}–2×10^{13} comets pc^{-3}. This is 0.25–5 times the upper limit noted above.

The lower limit in the above calculation is obviously not a problem. Since half of all stars form in multiple systems, and that process may prevent the formation of a PROTOPLANETARY DISK leading to cometesimals (although that conjecture has not actually been demonstrated), factors of a few times the interstellar comet density limit are also likely not a cause for great concern. However, a factor of 5 excess is not consistent with the Oort cloud models and population estimates presented here. This may suggest that Oort clouds are efficient at capturing a significant fraction of the planetesimals ejected from giant planet zones, rather than allowing them to escape to interstellar space. Alternatively, it may be that only a small fraction of stars actually produce Oort clouds.

Bibliography

Fernández J A 1994 Dynamics of comets: recent developments and new challenges *Asteroids, Comets, Meteors 1993* ed A Milani, M Di Martino and A Cellino (Dordrecht: Kluwer) pp 223–40

Weissman P R 1996 The Oort cloud *Completing the Inventory of the Solar System (ASP Conference Series 107)* ed T W Rettig and J M Hahn pp 265–88

Paul R Weissman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oort, Jan Hendrik (1900–92)

Astronomer, born in Franeker, Netherlands, became director of the Leiden Observatory, confirmed BERTIL LINDBLAD's hypothesis of galactic rotation by analysing motions of distant stars, with our Sun orbiting around the mass of the Galaxy which lies inside its path. The *Oort constants* are measurements of the Galaxy's rotation, and from them Oort was able to calculate the distance of the Sun from the center of the Galaxy and the period of its orbit. During the Second World War, Oort inspired HENDRIK VAN DE HULST to calculate the 21 cm radio spectral line from hydrogen, and led the Dutch group which co-discovered and used the line to map hydrogen gas in the Galaxy, discovering spiral structure and the galactic center. He suggested the existence of a sphere of primordial cometary material surrounding the Solar System. He proposed that comets detached themselves from this *Oort cloud*, and went into orbits about the Sun. Oort showed that light from the Crab nebula is polarized, confirming IOSIF SHKLOVSKII'S suggestion that it was synchrotron radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oosterhoff, P Th (1904–78)

Dutch astronomer, shared with JAN OORT the administration of the Leiden Observatory. Was in charge of the Leiden southern station in South Africa. His main interest was in variable stars and photometric systems. He discovered the *Oosterhoff dichotomy*, that the type-ab RR Lyrae variable stars in globular clusters are divided into two groups with different average periods, connected with the abundance of metals in the atmospheres of the stars in the globular clusters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ophiuchus

(the Serpent-bearer; abbrev. Oph, gen. Ophiuchi; area 948 sq. deg.) An equatorial constellation which lies between Hercules and Scorpius, and culminates at midnight in mid-June. The ecliptic cuts across the southern part of Ophiuchus, but the constellation is not included among the constellations of the zodiac. Ophiuchus is usually said to represent Asclepius, the Greek god of medicine, and is shown on early celestial charts as a man holding a huge snake (represented by the neighboring constellation of Serpens). The brightest stars of Ophiuchus were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large constellation, the brightest stars in Ophiuchus are α Ophiuchi (Rasalhague), magnitude 2.1, η Ophiuchi (Sabik), magnitude 2.4, and ζ Ophiuchi, magnitude 2.5. There are 11 other stars brighter than fourth magnitude. Interesting stars include ρ Ophiuchi, a multiple system consisting of three bluish-white (B2, B9 and B3) components, magnitudes 5.1, 7.1v and 7.0, separation 152″ and 156″, embedded in a reflection nebula (IC 4604), that have the same proper motion, the former and last of which are close binaries with a bluish-white (B2) component, magnitude 5.7, separation 3.1″, and a magnitude 8.7 component, separation 0.4″, respectively. Other interesting binaries include 70 Ophiuchi, which has yellow (K0) and orange (K4) components, magnitudes 4.3 and 6.3, separation 1.6″, period 88 years, and 36 Ophiuchi, which has two orange (K2) components, both of magnitude 5.2, separation 4.7″, period about 550 years, and a third orange (K5) component, magnitude 6.3v, separation 12′ north-east, which has the same proper motion and parallax. The latter system is about 20′ north of the 'stem' of the Pipe Nebula, a large area of dark nebulosity which extends east-west for several degrees and is visible to the unaided eye.

Other interesting stars include Barnard's Star, magnitude 9.5, which, at a distance of 5.9 light-years, is the second closest star to the Sun and has the highest proper motion of any star, RS Ophiuchi, a recurrent nova (range 4.3–12.5), which is normally between tenth and eleventh magnitude but flared up to naked-eye visibility in 1898, 1933, 1958, 1967 and 1985, and Kepler's Star (V843 Ophiuchi), the most recent example of a supernova in our Galaxy, observed by Kepler in 1604.

Other interesting objects in Ophiuchus include NGC 6572, an eighth-magnitude planetary nebula, and IC 4665 and NGC 6633, two open clusters, both of a few dozen stars of seventh magnitude and fainter, but the constellation is renowned for its globular clusters, of which over two dozen are known. Of the globular clusters, the finest are M12 (NGC 6218), M10 (NGC 6254) and M62 (NGC 6266), all sixth magnitude, and M14 (NGC 6402) and M19 (NGC 6273), which are seventh magnitude. The Milky Way crosses the southern part of Ophiuchus and many of the globular clusters in this region appear fainter than they would otherwise, owing to interstellar absorption.

See also: Barnard's Star, Kepler's Star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oppenheimer, Julius Robert (1904–67)

Nuclear physicist, born in New York City, joined the Manhattan Project and directed the Los Alamos Laboratory, where he became known as the 'father of the atomic bomb'. He directed the Institute for Advanced Studies, Princeton, NJ. He worked on cosmic rays and calculated theoretically the structure of neutron stars, remarking that it seemed unlikely that they existed. Neutron stars were discovered as pulsars in 1967 by JOCELYN BELL.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oppolzer, Theodor Egon Ritter von (1841–86)

A gifted mathematician, he knew 14 000 logarithms by heart. He calculated the *Canon der Finsternisse* (*Canon of Eclipses*) listing the time of occurrence and the visibility track of 8000 solar and 5200 lunar eclipses between 1208 BC and AD 2161, used for historical research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Opposition

The position of a planet in its orbit when it and the Earth are aligned such that it is directly opposite the Sun in the sky, so that it has an elongation of 180°. The term is also used for the time at which this alignment occurs. Only superior planets (those outside the Earth's orbit) can come to opposition. As this brings them closer to the Earth than at any other position in their orbit, opposition is the most favorable time for observing them. This applies in particular to Mars: because Mars's orbit is markedly eccentric, its distance from the Earth at opposition varies considerably. At *aphelic opposition*, when the planet is farthest from the Sun, the distance is 101 million km, while at *perihelic opposition*, when it is closest to the Sun, it is only 56 million km from the Earth.

See: elongation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Optical Astronomy

Optical astronomy is the subset of astronomy that is studied through use of light (and near infrared) from the universe. It comprises the most common observational techniques used by astronomers and accesses a wide range of celestial objects.

It is increasingly common for astronomers to concentrate on their subject matter, rather than the observational techniques which they use, and thus to use a variety of multiwavelength techniques to study their specialty subjects. Nevertheless, the different techniques of astronomy are individually complex and astronomers also, in practice, specialize in observational techniques too. They form collaborations with others when they venture outside their specialty technique to make the necessary observations of their specialty subject matter.

Spectral range

The spectral range of optical astronomy is set, not by the physical properties of its objects, but by its technology, which stems from the spectral response of the human eye. Two astronomical environmental considerations have, however, helped to define the optical spectral range. (i) The peak of the energy distribution of the Sun lies near to 550 nm, at which wavelength there is the maximum illumination of the Earth. (ii) The Earth's atmosphere has evolved to contain oxygen produced by living creatures, and therefore ozone and water vapor, which set respectively a sharp lower and fuzzy upper boundary to a transparency window between about 330 nm and, say, 1.4 μm.

Of course, there were strong biological factors in the development of the spectral response of the eye. In the evolution of primates, biological considerations favored light as a means to inspect vegetation, rather than, as in snakes, favoring infrared with which to detect thermal emission from warm-blooded prey. Together with these important biological considerations, astronomical considerations have presumably been at least minor factors to pull the evolution of the human eye to respond to light.

The precise spectral response of the eye is also a matter of the photochemical reactions of the retina and absorption in the material of the eye lensing system. The optical range fits within the atmospheric window, from 400 nm to 700 nm. Shortwards of 400 nm lies the ultraviolet, longwards of 700 nm lies the infrared. This spectral range is overlapped by the spectral responses of the photographic emulsion and CCD detectors, developed industrially to mimic human vision through the use of cameras fed by transparent lenses, and adopted as imaging and recording technology by astronomers (see DETECTORS: PHOTOGRAPHY, CHARGE-COUPLED DEVICE).

These rather ill-connected but consistent considerations define optical astronomy as in practice the study of the constituents of the universe that emit in the spectral range from 330 nm to say 1 μm. Beyond about 1 μm, optical detectors must be replaced by suitable infrared detectors, but optical telescopes give adequate performance up to about 2.4 μm. Above this wavelength, the increasing thermal emission from telescope and atmosphere becomes a major limitation. Ground-based telescopes can be made which mitigate the problems. Ground-based INFRARED TELESCOPES look more or less like optical telescopes, but are used with special techniques. They are usable up to a wavelength of, say, 35 μm at most. The spectral band from 330 nm to, say, 35 μm is termed 'optical–infrared' or OptIR.

By contrast with astronomy using much longer wavelengths (radio and millimeter), telescopes for OptIR astronomy are of manageable size. By contrast with astronomy using longer wavelengths (long-wavelength infrared to submillimeter), telescopes and detectors for OptIR astronomy operate at ordinary temperatures under the atmosphere (although with limitations of atmospheric seeing, and perhaps in cold mountain locations). By contrast with astronomy using shorter wavelengths (ultraviolet and x-ray), telescopes for OptIR astronomy have high throughput and do not have to be operated in space. Because of the commercial importance of optical detectors, they have been developed into panoramic detectors for astronomy with many detector elements (pixels) and produce rich amounts of data.

These considerations account for the historical primacy of optical astronomy. It was the sole area of astronomy up to the Second World War period. A large proportion of astronomers and observatories still work in OptIR astronomy. It remains the broad basis for astronomy. Often in the last 50 yr, of course, it has been fed by exciting and innovative discoveries made in other wavebands. Non-optical emission from a celestial object is a flag which waves 'look at me, I am interesting' and distinguishes its optical counterpart from the numerous similar-looking optical stars or galaxies. Even so, as the technology of OptIR telescopes develops, for example to larger-sized telescopes with greater collecting area, or through access by OptIR telescopes to space (HUBBLE SPACE TELESCOPE, Next Generation Space Telescope), its own capability for discovery is undiminished.

Subject matter of optical astronomy

The cosmic bodies that emit OptIR constitute a rich proportion of the material contents of the Universe. Not coincidentally (because of the connection between the evolution of the human eye in a location near to the Sun), many stars are of a surface temperature such that they emit significant radiation in this waveband. Not coincidentally (because the human eye is made of a molecular fluid and photochemical in the nature of its detector, and it operates under an atmosphere), many of the weaker electronic transitions in atoms and some in molecules involve photons in the optical range.

Optical astronomy therefore consists mainly of the study of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- thermal emission from stars of temperature above, say, 1000 K,
- thermal emission from collections of stars such as clusters and galaxies and
- spectral lines from atoms and ions in stellar atmospheres and nebulae.

Because SYNCHROTRON RADIATION has a broad band nature, synchrotron sources are also seen through optical astronomy, even if relatively weak.

The OptIR band brings into the subject matter for study

- thermal emission from cooler stars and the dust material that surrounds them, at temperatures down to about 150 K, and
- spectral lines of a number of molecules.

Techniques of optical astronomy

The broad range of techniques that lie in the science of optics have been applied to OptIR astronomy. Glass is transparent to the OptIR spectral range. Ionic impurities can limit the spectral range of transmission, a limitation to the wider spectral range of glass or an exploitable asset as color FILTERS.

Mirrors reflect light and infrared. At first mirrors were of polished metal (speculum) and later of glass or ceramic material coated with aluminum, silver or gold (depending on the spectral region within or near the OptIR range that is most important). Lenses and mirrors can be used to focus and filter light and form optical images for detection and analysis.

The DISPERSION properties of transparent materials produce a means to form spectra for analysis. The different dispersion properties of different transparent materials form the basis to compensate for defects of optical systems caused by it. The wavelength of light is longer than the separation of atoms and substructure in glass masses, so parallel scratches can create DIFFRACTION GRATINGS. The wavelength is not so small that the scratch separation or the controlled separation of parallel planes of glass is outside technological reach. The wavelength is small enough that a large density of scratches is feasible, making high spectral resolution possible in the dispersed light. These considerations facilitate spectroscopy, spectral line analysis and measurement of radial velocities at the levels common in the motions of stars and nebulae.

The mechanical rigidity of glass materials and their relatively low thermal expansion coefficient make large telescope lenses and mirrors possible. This has enabled optical telescopes to reach low flux limits and drawn into consideration a substantial fraction of the universe of stars and galaxies. The limits to mechanical and thermal stability also ultimately set limits on the size of the ground-based telescopes that can be made, even with the active control systems now possible. Because of the flux gradients from celestial objects such as the Sun and Moon, thermal stability remains a limit to the performance of OptIR telescopes even in space, such as the Next Generation Space Telescope.

Paul Murdin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Optical Double

This term is used for an observed double star that consists of two quite unrelated stars which just happen to lie in almost the same line of sight, and therefore appear to be close together in the sky. They may in fact be at very different distances from us and each other. If one of the stars is sufficiently close to have a measurable parallax, the true nature of the system can be quickly established. For most pairs this is not the case, and only measurements of their separation and position angle repeated over many decades will confirm that they have no orbital motion and do not constitute a true binary system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Optical Interferometers

Optical INTERFEROMETERS are astronomical instruments designed to provide higher angular resolving power than is possible with a conventional optical telescope or, in other words, the capacity to measure smaller angles and to discern finer details in an image.

The ability of a telescope to resolve fine detail in an astronomical object is fundamentally limited by the wavelength (color) of the light used for the observations and the diameter of the mirror or lens that collects light from the object. In practice the Earth's atmosphere imposes a limit on the resolution achievable unless special techniques are used, but even above the atmosphere the only way to see finer detail with a telescope would be to observe at a shorter wavelength (bluer light) or to build a larger telescope.

The measurement of the sizes of stars was the challenge that led to the development of optical interferometry. The size of a star cannot be measured directly but, if its angular size $\theta_\star$ and its distance $D_\star$ can be determined, its diameter $d_\star$ can be calculated from $d_\star = D_\star \times \theta_\star$, as illustrated in figure 1. From above the Earth's atmosphere the images of all but a very few stars appear as points of light blurred only by the wave nature of light. From the ground, turbulence in the Earth's atmosphere corrupts the light from a star and blurs its image so badly that no star, other than our own Sun, can have its size determined from its image. Astronomers call this blurring effect 'SEEING'.

FIZEAU (1868) was the first to propose an interferometric method to overcome the effects of seeing. He noticed that when a star was viewed through a small aperture its image was blurred but was otherwise regular in shape—unlike that from a large telescope where the image is blurred but also continually changing. This led him to the idea of putting a mask with two small holes in front of a telescope and observing the superimposed images of the star. The combined image would be crossed by interference fringes—alternate bright and dark bands. Fizeau realized that the contrast of the fringes, the difference in brightness of the bright and dark bands, would diminish as the separation of the apertures was increased, and that the fringes would vanish at a separation directly related to the angular size of the star. Stephan (1873) tested Fizeau's idea but was unable to observe the disappearance of the fringes since the largest telescope available had an aperture diameter of only 0.8 m, which was too small. Nevertheless, Fizeau's idea is the basis of modern optical interferometers which, by measuring the fringe contrast as a function of the separation of the apertures, are able to determine the angular sizes of stars and the angular separations of close binary (double) stars as illustrated in figure 1.

The first measurement of the angular diameter of a star was by MICHELSON and PEASE in 1920 when they measured the angular diameter of Betelgeuse, the bright red supergiant star in the constellation of Orion. Because even the 100 inch (2.5 m) Hooker telescope on

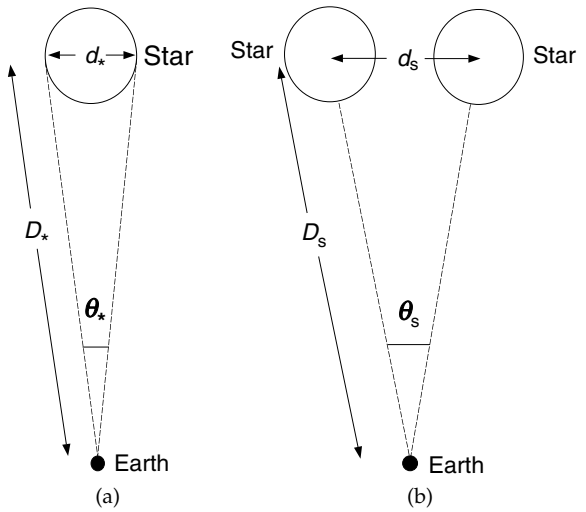


Figure 1. The angular diameter of a star $\theta_\star$ and the angular separation θ_s of the components of a binary, or double star.

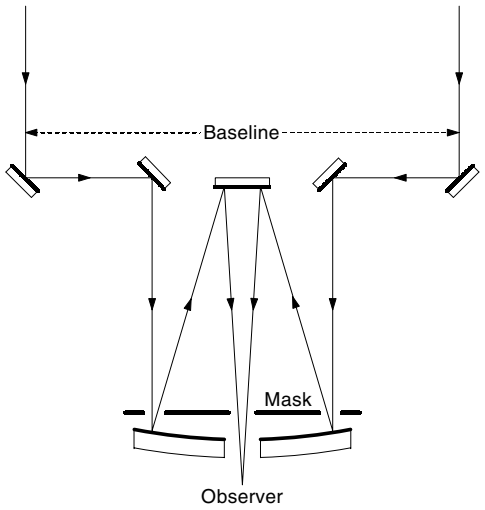


Figure 2. Cross section of the Michelson stellar interferometer mounted on the 100 inch Hooker telescope on Mount Wilson in California.

Mount Wilson in California was too small to observe the disappearance of the fringes, Michelson had the brilliant idea of attaching a long beam across the front of the telescope to carry mirrors to relay starlight from separations greater than the telescope aperture into the telescope. A diagram of Michelson's stellar interferometer is shown in figure 2.

In the absence of a turbulent atmosphere even the largest optical telescopes, with aperture diameters of 8–10 m, would only reveal about a dozen stars as images whose sizes could be measured. While interferometric techniques, like speckle interferometry and developments of the masking interferometry originally proposed by Fizeau, have enabled telescopes to measure stars to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

limits set by their aperture diameters and the wavelength of light, interferometry has moved on to combine the light from separate apertures in order to increase the angular resolution. This development required new technology in the form of laser metrology, adaptive optics, fast electronics and computer control and, since the driving force was primarily the measurement of stars and binary stars, most instruments have been called stellar interferometers. The increasing sensitivity and capabilities of interferometers offer other possibilities that are exciting astronomers and astrophysicists as the potential of interferometry is recognized. For example, these include the determination of accurate stellar positions (ASTROMETRY), the search for EXOPLANETS (planets other than those in the solar system) and the measurement of the sizes of cores of ACTIVE GALAXIES.

Basic theory

The minimum angle $\theta_{\min}$ resolvable by a circular aperture of diameter d at a wavelength λ is given by

$$\theta_{\min} = \frac{1.2\lambda}{d} \text{ radians.}$$

For a 10 m telescope $\theta_{\min}$ is 0.012 arcsec at a wavelength of 500 nm. The angular diameters of all but a few stars are less than a very few milliarcsec (mas) and interferometers with aperture separations (baselines) in excess of 10 m are required to resolve them.

The combination of the light from two separate apertures is similar to a Young's double-slit experiment with interference fringes being formed in the combined images of the source. A measure of the contrast of the fringes is Michelson's fringe visibility V defined as

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$

where $I_{\max}$ and $I_{\min}$ represent the maximum and minimum intensities in the fringe pattern.

A more sophisticated approach involves consideration of the coherence of the light at the two apertures. The complex degree of coherence $\gamma_{12}(\tau)$ can be expressed in terms of a modulus $|\gamma_{12}(\tau)|$ and a phase $\phi_{12}(\tau)$, where τ is the difference in arrival time of the two beams at the point of combination and the subscripts 1 and 2 represent the two apertures:

$$\gamma_{12}(\tau) = |\gamma_{12}(\tau)| \exp[\mathrm{i}\phi_{12}(\tau)].$$

When the intensities from the two apertures are equal, the fringe visibility $V_{12}(0) = |\gamma_{12}(0)|$. The complex degree of spatial coherence $\gamma_{12}(0)$ (i.e. with $\tau = 0$) and the angular distribution of intensity across the source I_α are a Fourier transform pair—the van Cittert–Zernike theorem.

$$\gamma_{12}(0) = \int_{-\infty}^{\infty} I_\alpha \exp\left(\frac{-2\pi \mathrm{i}\alpha}{\lambda}\right) \mathrm{d}\alpha.$$

If $|\gamma_{12}(0)|$ and $\phi_{12}(0)$ are measured over an appropriate range of baseline lengths and orientations, the angular intensity distribution (I_α) of the source can be found. The fringe phase $\phi_{12}(0)$, as a function of the aperture separation, provides information on asymmetries in the angular intensity distribution—for a symmetrical source the fringe phase will be zero or 180° for all aperture separations.

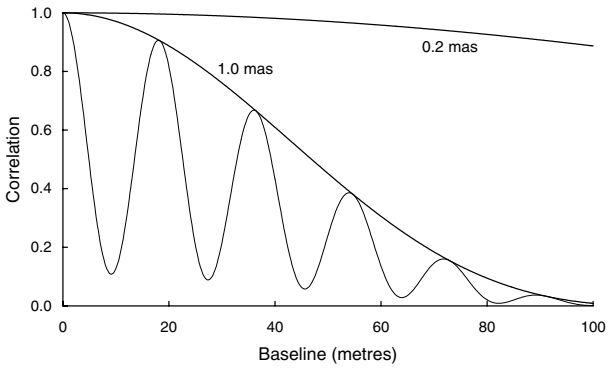


Figure 3. The variation of correlation with baseline for two single stars whose angular diameters are shown in the figure. The oscillating curve is the response for a binary star. The primary star of the binary has an angular diameter of 1 mas and is twice as bright as the secondary. The angular separation of the two stars is 10 mas.

In practice, the Earth's turbulent atmosphere disrupts the phase and, for observations made with a single baseline at a time, only the degree of coherence $|\gamma_{12}(0)|$ is measurable. In the absence of the phase, departures from symmetry in the intensity distribution cannot be detected. Thus, measurements with a single baseline yield only the visibility, and these are interpreted as the angular size of an equivalent strip source or the angular diameter of a radially symmetrical disk source. To obtain images it is necessary to measure both the modulus and phase for at least three baselines simultaneously.

In practice, many interferometers measure the square of the fringe visibility, which is usually called the correlation $C_{12}(\tau)$. Thus

$$C_{12}(\tau) = [V_{12}(\tau)]^2 = |\gamma_{12}(\tau)|^2.$$

As an example of the variation of correlation with baseline figure 3 shows the relationships for two stars with different angular diameters, both assumed to be uniformly bright across their circular disks. Also shown is the response for a BINARY STAR.

If there is a path difference between the beams in an interferometer at the point where they are combined there will be a loss in fringe visibility and in correlation. The magnitude of the loss depends on the spectral bandwidth used for the measurements. The wider the bandwidth, the greater the loss. The reduction in correlation can be expressed in terms of the coherence length of the light Δl, which is given by $\Delta l = \bar{\lambda}^2/\Delta\lambda$, where $\Delta\lambda$ is the spectral bandwidth and $\bar{\lambda}$ is the mean wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The Fourier transform of the intensity distribution across the spectral bandpass is the complex degree of coherence and, as an example, the correlation for a flat-topped spectral bandpass is given by

$$C_{12}(\Delta\mathrm{OPL}) = [V_{12}(\Delta\mathrm{OPL})]^2 = \left|\frac{\sin x}{x}\right|^2$$

where $x = \pi\,\Delta\mathrm{OPL}/\Delta l$ and ΔOPL is the optical path length difference between the beams (optical path length is the product of the geometrical path length in a medium and the refractive index of the medium).

In order to obtain accurate measurements of fringe visibility or correlation, the optical paths from the source to the point of combination of the beams must be equal to within a small fraction of the coherence length ($\Delta\mathrm{OPL} \ll \Delta l$).

Overcoming the effects of atmospheric turbulence

The effect of atmospheric turbulence on incoming light is to distort the shapes of wavefronts and introduce randomly varying phase shifts, wavefront tilts, and wavefront curvatures. The significance of these effects for interferometry depends on the observing wavelength. The spatial scale of the wavefront distortions is measured by the parameter r_0, which is the diameter of a circle on a wavefront over which the phase fluctuations have an rms value of one radian. The temporal scale of the phase fluctuations is represented by t_0, the time interval at the end of which the rms value of the phase fluctuations is one radian.

Both r_0 and t_0 vary as $\lambda^{6/5}$ so both improve for longer wavelengths. Observed values depend on the site, and both vary from night to night and during the course of each night. Typical values for the visual region of the spectrum are $r_0 \sim 10$ cm and $t_0 \sim 5$ ms.

In order to overcome the effects of atmospheric turbulence most optical and infrared interferometers use apertures of diameter less than r_0 and signal sampling times less than t_0. In addition, tilts in the sections of the wavefront accepted by the apertures of an interferometer are individually measured and corrected in real time by tip-tilt mirrors (first-order ADAPTIVE OPTICS). This ensures that the beams from the apertures are combined optimally for the measurement of the interference fringe contrast.

In principle, instrumental loses and residual seeing effects are calibrated by observing unresolved stars between observations of the star or object to be measured.

Two-aperture interferometers

Modern optical/infrared interferometers operate with baselines ranging from a few metres up to a few hundred metres (see INTERFEROMETRY: GROUND). This necessitates mounting the input apertures on separate structures anchored to the ground. Because of the short wavelengths involved, extremely good stability and freedom from vibration is essential. With the input apertures at fixed locations the arrival time of the light from a star will differ at the input apertures and an internal variable optical delay must be incorporated in the instrument to equalize the optical path lengths at the point of beam combination. One method of achieving this is shown in figure 4.

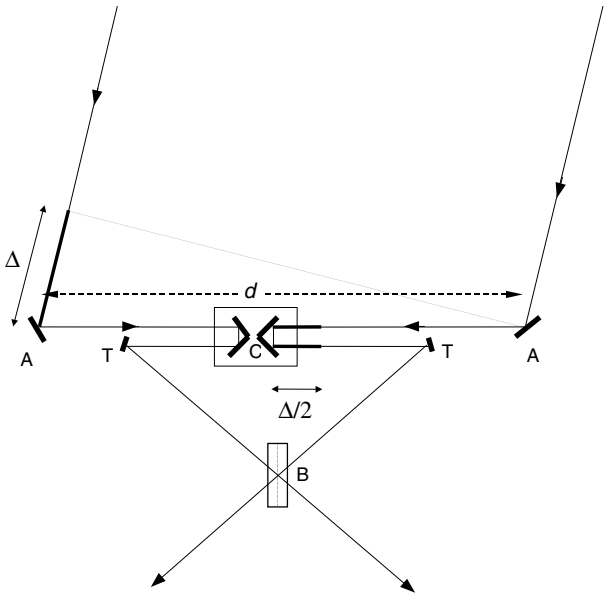


Figure 4. A simplified layout for a long-baseline optical interferometer illustrating the main features. A represents the input apertures and T the wavefront tip-tilt correcting mirrors, Δ is the difference in path length for the two beams of starlight at the input apertures, C is the path-equalizing carriage, B is the beam-combiner and d is the baseline of the interferometer. Further details are discussed in the text.

For interferometers in which the light from only two apertures at a time can be combined, only the fringe visibility or the correlation is measurable since the phase is disrupted by the atmosphere. For such measurements, made with only a single baseline at a time, the Fourier transform yields the intensity distribution for the equivalent strip source (i.e. the source intensity distribution reduced to the equivalent strip source parallel to the baseline direction). No information on asymmetries in the intensity distribution across the source can be derived and true images cannot be constructed. For angular diameter determinations and studies of binary systems this is generally of no concern—the relative brightnesses of the components of a binary system can be determined but with a 180° ambiguity in the orientation of the system. The combination of stellar angular diameters with complementary measurements made with conventional telescopes enables fundamental stellar properties such as surface fluxes, effective temperatures, radii and luminosities to be determined. For many binary stars radial velocities can be determined spectroscopically. For such systems stellar masses and distances can also be found.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Interferometers and interferometric arrays.

Instrument (acronym)	Location (country)	Aperture diameter (m)	Longest baseline (m)	Wavelength range (μm)	Status†
Stellar Intensity Interferometer (SII)	Australia	2×6.8	188	0.44	C
Sydney University Stellar Interferometer (SUSI) Prototype	Australia	2×0.10	11	0.4–0.5	C
Mark III Interferometer	USA	2×0.05	32	0.45–0.8	C
Interféromètre à 2 Télescopes (I2T)	France	2×0.26	144	visible	C
Grand Interféromètre à 2 Télescopes (GI2T)	France	2×1.5	65	visible	W
Infrared spatial interferometer (ISI)	USA	2×1.65	70	10	W
Sydney University Stellar Interferometer (SUSI)	Australia	2×0.14	640	0.4–0.9	W
Infrared–Optical Telescope Array (IOTA)	USA	3×0.45	38	visible/IR	W
Cambridge Optical Aperture Synthesis Telescope (COAST)	UK	5×0.4	100	red/near IR	W
Palomar Test Interferometer (PTI)	USA	2×0.4	100	2.2	W
Navy Prototype Optical Interferometer (NPOI) (Astrometry)	USA	4×0.125	38	0.45–0.9	W
(Imaging)		6×0.35	437	0.45–0.9	UC
Center for High Angular Resolution Astronomy Array (CHARA Array)	USA	5×1.0	354	0.55–0.9 2.1–2.5	UC
European Southern Observatory (ESO)	Chile	4×8	200	0.45–20	UC
Very Large Telescope Interferometer (VLTI)		3×1.8			UC
Keck Interferometer	USA	2×10	165	2.2–10	UC
		$4 \times \sim 2$			UC

† Key: C = program completed and instrument shutdown; W = instrument working; UC = undergoing commissioning or under construction.

Multiple-aperture interferometers

In order to form images with an interferometer it is necessary to use more than two apertures simultaneously. For each pair of apertures in an imaging array the same constraints must be met as for a single-baseline instrument. This includes matching the optical paths to within a small fraction of the coherence length, dealing with the effects of wavefront distortion, and using rapid signal sampling to limit phase smearing during each observational sample period.

Even though the phase measurement for a single baseline is corrupted by the atmosphere, it is possible to determine a 'closure phase', free of corruption, for each independent triangle of baselines formed by three apertures of an array. A closure phase is solely a function of the fringe phases due to the source since, in the calculation of the closure phase, the atmospherically induced phase differences cancel out. Closure phases are a property of the angular intensity distribution across the source and, together with the fringe visibilities measured for the baselines, enable an image of the source to be constructed. This technique, first developed for radio wavelengths, has been adapted to optical and infrared wavelengths.

Multiple-aperture interferometric arrays are aimed at imaging a range of astronomical phenomena including the birth of stars and planetary systems, close binary systems and the cores of active galaxies and quasars.

Modern optical interferometers

Table 1 contains a list of optical/infrared interferometers and interferometric arrays with the acronyms by which they are generally known. The instruments shown as having two apertures combine the light from only two apertures at a time. In most of these cases there is either an array of apertures, from which the two to be used can be selected, or there is the possibility of moving input optics between stations to provide a range of baselines. Instruments listed with three or more apertures are generally intended to combine the light from several, or all, of the apertures simultaneously for the construction of images.

Interferometers and interferometric arrays intended for astrometric measurements are similar in principle to other interferometers and arrays, but differ in having extensive monitoring of the instrument stability by laser metrology and an optical delay compensator operating in vacuum. Astrometric interferometers are of two broad types—global and narrow-angle (differential). Global astrometric interferometers, such as the NPOI Astrometric Array in Arizona, measure the relative positions of stars over the whole sky with the aim of maintaining the Hipparcos frame and improving proper motion determinations by extending the timebase of measurements. Narrow-angle interferometers are effectively double interferometers with each interferometer linked to the other by precision laser metrology. One interferometer is phase locked to a reference source while the second is locked to the target source to search for relative motion produced by a faint unseen companion such as a brown dwarf or planet.

An alternative interferometric technique for the detection of exoplanets is the nulling interferometer. By working in the infrared, and suppressing the stellar flux by combining the star's light out of phase, it should be possible to detect planets down to Earth size but with an instrument operating in space.

Figure 5 shows an aerial view of SUSI, which is located in northern New South Wales in Australia. SUSI has the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

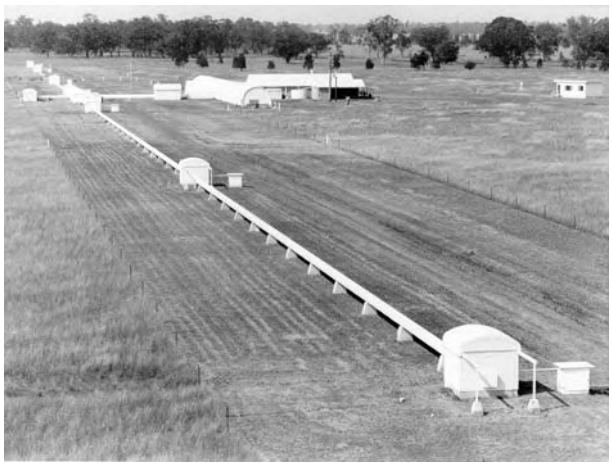


Figure 5. The Sydney University Stellar Interferometer (SUSI) seen from the northern end of its 640 m long baseline array.

distinction of having the longest baseline array of any optical/infrared interferometer. Since it can only observe with two apertures at a time it does not form images. Instruments such as NPOI, the CHARA Array on Mount Wilson in California, and COAST in England all have arrays of input apertures arranged along three arms like the letter Y in order to form images by combining the light from three or more apertures simultaneously. The Keck interferometer and ESO's VLTI will combine the light from separate 8–10 m class telescopes for imaging of fainter objects.

Optical and infrared interferometry and interferometric imaging is in its infancy but promises to make significant contributions to astrophysical research in the 21st century.

Bibliography

An overview of optical interferometry and its potential for astrophysics, as presented at a NATO Advanced Study Institute written by authors expert in various aspects of the field is contained in *High Angular Resolution in Astrophysics* ed A-M Lagrange, D Mourard and P Léna (Dordrecht: Kluwer, 1997).

A large proportion of reports on the developments in optical and infrared interferometry is contained in the proceedings of international symposia and workshops. However, the key technical papers in the development of optical interferometry have been collected in *Selected Papers on Long Baseline Stellar Interferometry* SPIE Milestone Series, Volume MS 139, ed P R Lawson (Bellingham, WA: SPIE, 1997).

The theory of interference with partially coherent light is treated comprehensively by Born M and Wolf E 1975 *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light* 5th edn (Elmsford, NY: Pergamon).

A review of the scientific potential, particularly for interferometric arrays involving large aperture telescopes, is presented in *Science with the VLT Interferometer* ed F Paresce (Berlin: Springer, 1997).

John Davis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orbit

The path pursued by a body moving in a field of force. In most astronomical contexts we consider motion in a gravitational field, but the term applies equally well to motion under the action of other types of forces, for example to charged particles moving in a magnetic field. Bodies moving freely in the gravitational field of a massive body (e.g. the planets moving around the Sun or an artificial satellite moving around the Earth) follow conic orbits, which may be elliptical, circular, parabolic or hyperbolic. The precise orbit that such a body will pursue depends upon the strength of the gravitational field and the velocity of the body. A planetary orbit is completely described by six geometric properties called its elements. From these the future positions of a planet can be calculated. In the case of a binary system, where the masses of the stars are unknown, seven elements are required.

See also: conic sections, ellipse, hyperbola, orbital elements, parabola, revolution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orbital Elements

The parameters required to define the shape and orientation in space of an orbit and to fix the position in that orbit at any time of the planet or satellite moving in that orbit. For a planet moving around the Sun, six elements are required and these would be as follows:

- a the semi-major axis of the ellipse
- e the eccentricity of the ellipse
- i the inclination of the orbit to the plane of the ecliptic
- Ω the longitude of the ascending node, i.e. the angle measured anti-clockwise in the ecliptic from the vernal equinox to the point of intersection of the orbit with the ecliptic at which the planet passes from south to north of this plane
- ω the longitude of perihelion (sometimes called the 'argument of perihelion'), i.e. the angle, measured anti-clockwise in the orbit plane, from the ascending node to perihelion
- T the time at which the planet passed perihelion.

The Sun occupies one of the two foci of the ellipse of a planet's orbit. A line drawn through the point of the planet's perihelion and aphelion passes through the Sun and is called the line of apsides or major axis of the orbit. One-half of this line's length is the semi-major axis, equivalent to the planet's mean distance from the Sun. The eccentricity of an elliptical orbit is a measure of the amount by which it deviates from a circle; it is found by dividing the distance between the focal points of the ellipse by the length of the major axis. The position of a planet at any time can be predicted if it is known when it passed through any definite position; for example, its time of perihelion passage.

For a satellite in orbit around the Earth, the ascending node and inclination would be referred to the plane of the celestial equator.

In the case of a binary system of unknown mass, a seventh element is necessary: the orbital period P, or the mean motion, $n(=2\pi/P)$. In the case of the solar system this additional element, although sometimes listed, is related to the value of the semi-major axis (a) by Kepler's third law. T, in the case of a binary system, relates to the time of periastron passage (the instant at which the two stars make their closest approach to each other).

See also: conic sections, ellipse, hyperbola, orbit, parabola.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orbiting Astronomical Observatory (OAO)

Series of four NASA satellites. No data were returned from OAO-1 or OAO-B. The second satellite (OAO-2) operated for four years. It carried two experiments which surveyed the sky at ultraviolet wavelengths. Most successful was COPERNICUS (OAO-3), which was launched in August 1972 and operated for nine years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orbiting Solar Observatory

Series of eight NASA satellites launched 1962–75. Designed to study the Sun at ultraviolet and x-ray wavelengths. OSO-1 and 3 also carried gamma ray instruments for extrasolar studies. Most successful was OSO-8 which operated for more than three years and studied the diffuse cosmic x-ray background.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oresme, Nicole (1320–82)

Born in Allemagne, France, became dean of Rouen and chaplain to King Charles V. He invented coordinate geometry before DESCARTES, finding the logical equivalence between a function and a graph. He taught the motion of the Earth, 200 years before COPERNICUS. He argued that there were objections to the 'proofs' that the Earth was stationary. These were that (i) the Sun and the stars are observed to rotate around the Earth, (ii) a great wind from the east would be expected if the Earth rotated, and (iii) a stone would be expected to fall in a curved path if the Earth rotated. But in the end he lost confidence in his own objections and ended back with a stationary Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orientale Basin

A large multiringed structure on the Moon, about 930 km across in its largest dimension. It is partly on the nearside and partly on the farside, centered at 19.4 °S, 92.8 °W, and at best is only poorly visible from the Earth, and then only when libration brings it partly into view. It has been only partly filled by basaltic lava flooding, to form the Mare Orientale, 327 km across at its widest. (This name means 'Eastern Sea'; its longitude demonstrates that the convention for directions on the Moon has not always been the same.) The surrounding ring structure is the best preserved of any of the Moon's large impact features. The mountain range forming the outer ring is named Montes Cordillera, and the inner ring is Montes Rook. Orientale is the youngest of the Moon's large impact features, having been excavated around 3.8 billion years ago, which explains why the structure is largely intact. Ejecta from the impact traveled up to 1000 km, and the basin is surrounded by chains of secondary craters and other radial features, including a number of lunar valleys.

See also: Libration, Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Origin of The Solar System

The origin of the SOLAR SYSTEM, and of the Earth in particular, is one of the most fundamental problems in astronomy. Apart from the intrinsic fascination of understanding how and when our planet was formed, a plausible solution to that problem could suggest an answer to a more general question: is the formation of a planetary system a normal feature of STAR FORMATION, or a rare accident? Are planetary systems likely to be so abundant throughout the universe that the chance of communicating with other civilizations is high, or is life on Earth the lonely exception in a dead galaxy?

Before the 18th century, European astronomers believed that God created the Solar System in its present state only a few thousand years ago. In the 1700s, speculations proliferated about physical processes that might have formed the Sun and planets over a longer period of time. The French naturalist BUFFON imagined a close encounter of the Sun with a comet, which drew from it hot gases that later condensed to form the planets. He estimated that the Earth had cooled down over a period of about 75 000 years.

The nebular hypothesis

During the 19th century most astronomers accepted the 'nebular hypothesis' proposed in 1796 by the French theorist Pierre Simon de LAPLACE and the German–British astronomer William HERSCHEL. Laplace stressed the uniformity of motions in the Solar System: all planets go in the same direction around the Sun in nearly the same plane and rotate in the same direction around their own axes, and all satellites go in the same direction around their primaries. Such uniformity would be extremely improbable if these bodies had been formed independently. For NEWTON that was a proof that the system had been designed by God; for Laplace it showed that the system had developed by a unitary physical process. (He was not worried by exceptions such as the satellites of Uranus.)

Laplace suggested that the atmosphere of the Sun originally extended throughout the entire space now occupied by planetary orbits; it was a hot, luminous rotating cloud of gas, gradually losing its original heat to the cold surrounding space. As the cloud cooled it contracted, rotated faster, flattened into a disk, and broke up into rings, which then condensed to planets. Satellites could be formed by a similar process from clouds surrounding the planets. The remaining central portion became the Sun (see figure 1). (There is one obvious defect of the theory: the Sun should rotate faster than the speed of revolution of Mercury around the Sun, but in fact it rotates much more slowly.)

Herschel suggested that stars are formed by contraction from nebulae like those he observed through his telescope. Thus the Laplace–Herschel nebular hypothesis, a 'monistic' process requiring no outside stimulus to start

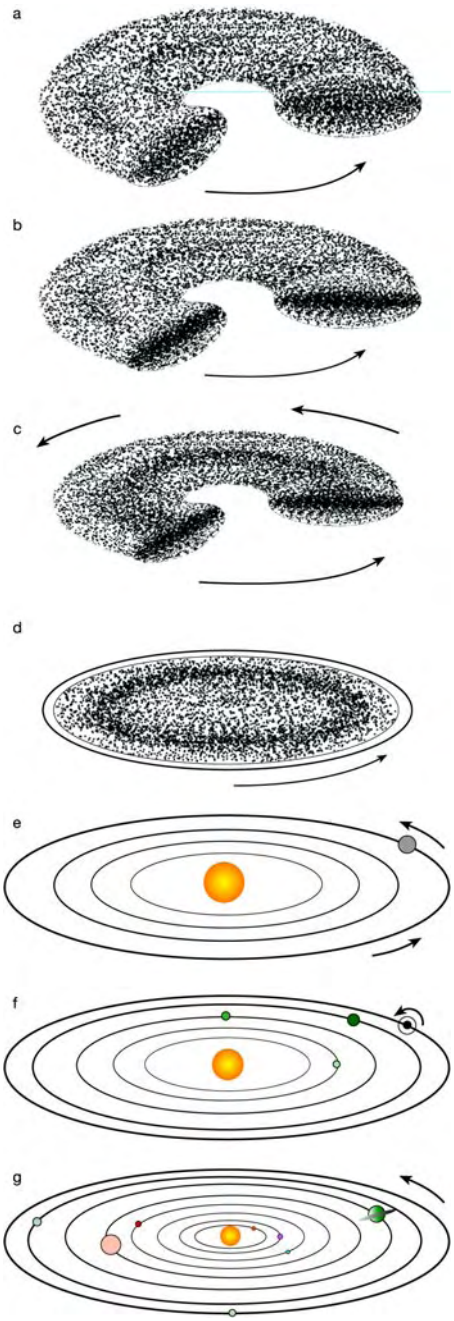


Figure 1. Nebular hypothesis. (a) Rotating spherical cloud has started to flatten. (b) Most of the material has fallen to equatorial plane, forming a flat disk. (c) Disk cools and contracts, spins faster (conservation of angular momentum). (d) When the outside of the disk spins so fast its centrifugal force equals gravitational attraction of material inside the disk, it no longer contracts but is left as a ring. (e) This process is repeated, forming several rings. Material in the outer ring condenses to a protoplanet. (f) Each ring forms a protoplanet; outer protoplanets spin off rings which form satellites. (g) All rings have formed planets, most planets have satellites and one (Saturn) also has a flat ring that did not condense into a satellite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the development of the system, implied that planetary formation is a universal accompaniment of the birth of stars, and that with sufficiently powerful telescopes we should be able to observe the later stages of the process occurring in our Galaxy right now. Also, it would seem likely that life has arisen or will arise on some of these many planets. A 'dualistic' process like Buffon's, on the other hand, would be much rarer, especially when it is realized that the foreign body would have to be another star rather than a comet.

The German scientist Hermann von HELMHOLTZ, in 1854, modified the Laplace–Herschel theory by assuming that the primordial nebula was cold rather than hot; gravitational contraction would convert mechanical energy to heat (in accordance with the Law of Conservation of Energy), and thus account for the present high temperature of the Sun.

The nebular hypothesis was closely connected with 19th-century geological theories, which generally presumed that the Earth had been formed as a hot fluid ball and then cooled down, solidifying on the outside first.

The British physicist Lord Kelvin (William THOMSON) used the cooling model to estimate the age of the Earth and found it to be only 20–100 million years, compared with the several hundred million years assumed by geologists to have been available for slow processes like erosion to produce their observed effects. Adopting Helmholtz's hypothesis of heating by gravitational contraction, Kelvin also estimated that the Sun could not be much older than 20 million years. Kelvin's estimates caused difficulties for Charles Darwin's theory of biological evolution, since Darwin had assumed, following the geologists, that long periods of time would be available for present-day species (including humans) to emerge by the slow process of natural selection.

At the beginning of the 20th century the credibility of the Laplace–Herschel–Kelvin scenario for the development and present state of the Earth was undermined by the discovery of radioactivity. The Earth's crust contained enough radium, the highly radioactive element isolated by Marie and Pierre Curie in Paris, to generate the heat needed to replace that lost from the Earth's crust by conduction into space; the Earth might actually be warming up rather than cooling down. Analysis of the amounts of uranium, radium, lead and helium in rocks quickly led to estimates that these rocks are several hundred million years old. In 1953 the American geochemist Clair Cameron Patterson used very accurate determinations of lead isotopes in rocks and meteorites to arrive at the presently accepted value for the age of the Earth: about 4500 million years.

Similarly it was proposed that radioactivity (or, later, nuclear fusion reactions) in the Sun could provide enough energy to keep it shining for much longer than 20 million years; thus the evolution of the Solar System also took place over thousands of millions of years.

Planetesimal and tidal theories

Even before radioactivity had cast doubts on the 19th-century cooling Earth model, the American geologist Thomas Chrowder CHAMBERLIN had proposed to replace that model by a completely different hypothesis about the origin of the Earth and other planets. His researches on past Ice Ages led him to doubt the doctrine that the Earth had cooled down from an initial temperature of several thousand degrees. With the help of the American astronomer Forest Ray MOULTON, he showed that the nebular hypothesis could not satisfactorily explain the physical properties of the Solar System (for example, the fact that the giant planets have most of the angular momentum rather than the Sun).

Chamberlin proposed that the planets and satellites had been formed by the aggregation of small cold particles, which he called 'PLANETESIMALS' (meaning 'infinitesimal planets'). The formation process would generate enough heat to bring them up to their present temperatures.

Having rejected Laplace's theory of the formation of the Solar System, Chamberlin still seemed to be attached to Herschel's old idea that the nebulae one sees in the sky are associated with the early stages of stellar evolution. Looking at photographs of spiral nebulae taken by the American astronomer James KEELER, Chamberlin speculated that the two prominent arms belonged to two previously distinct celestial objects. From this thought, and contemplation of solar prominences, he was led to the idea that a planetary system could be generated when another star passed close to the Sun. Just as the Moon raises tides on the near and far sides of the Earth, the gravitational force of the intruder would cancel the gravitational force holding in the solar gases on the near and far sides of the Sun, allowing two filaments of material to flow out; the filaments would then be curved by the continued action of the intruder as it receded. Chamberlin assumed that the filaments would eventually condense into small solid particles, which would be captured into orbits around the Sun (see figure 2).

When it became clear in the 1920s that spiral nebulae are galaxies rather than objects that could be as small as planetary systems, Chamberlin dropped this part of his theory but retained the assumption that two stars interacted in order to release into space the material from which planets formed. As it happened, the British geophysicist Harold JEFFREYS and the British physicist–astronomer James Hopwood JEANS independently proposed a similar assumption in 1916 and 1917, respectively. It became known as the 'tidal' theory of the origin of the Solar System. But Jeans and Jeffreys rejected Chamberlin's planetesimal hypothesis, retaining instead the older assumption that the planets formed from fluid balls (see figure 3).

Astronomers realized that any theory that required the close encounter of two stars in order to form planets would entail an extremely small number of planetary systems in the universe (and an even smaller number of planets that could support the evolution of life). This was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

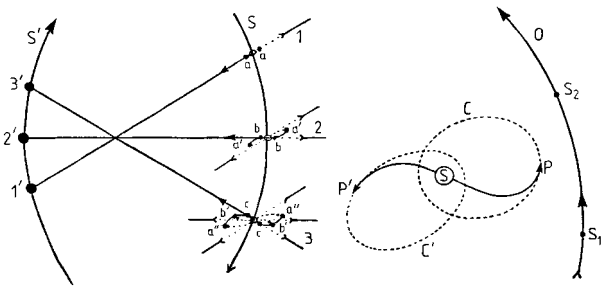


Figure 2. Creation of the Solar System by tidal action of a passing star on the Sun (Chamberlin's version). Chamberlin, inspired by Keeler's photographs of spiral nebulae, imagined that such a nebula could be produced when an intruder star S' (in successive positions 1', 2', 3') exerts a tidal force on the Sun S (in successive positions 1, 2, 3). As a result, material is drawn out of the Sun on near and far sides (a). In the second stage, more material (b) has emerged from the Sun, but it will experience a force in a different direction because of the subsequent motion of S' from 1' to 2'; similarly for material (c) emerging in the third stage. The resulting curvature of the filaments produces the spiral-arm appearance. The gaseous filaments condense first to small solid particles (planetesimals) which then aggregate into planets.

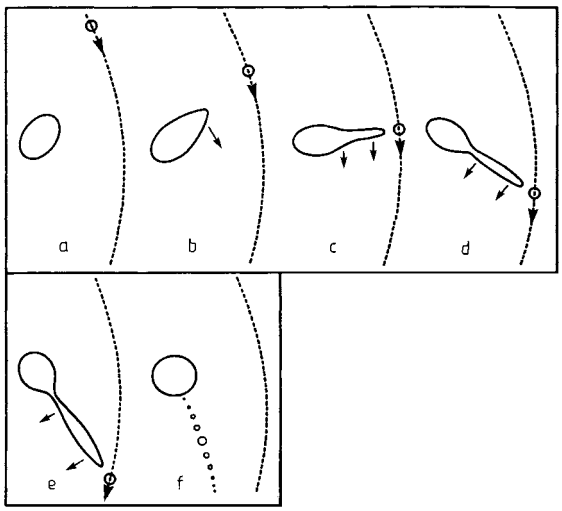


Figure 3. Creation of the Solar System by tidal action of a passing star on the Sun (Jeans' version). Jeans imagined that the Sun would first be deformed by the tidal force of the passing star, then a single coherent filament would be detached and subsequently break up into several fluid blobs which would condense directly into planets. According to Jeans 'This sequence of six diagrams (a, b, c, d, e, f) is intended to suggest the course of events in the Sun, or other star, when its planets were coming into existence. For simplicity the second star is represented throughout by a small circle, although actually this would, of course, also experience deformation and possible break-up'.

consistent with the failure to find convincing evidence for planetary systems surrounding other stars than our own. Jeans seemed to take perverse pleasure in the idea that we are the result of a chance event that has happened only once in the history of the universe and (because the stars are decaying and thinning out by expansion) will probably never happen again.

The tidal theory, whether the Chamberlin–Moulton or Jeans–Jeffreys version, was generally accepted by astronomers until 1935, even though it was never worked out in sufficient detail to provide a convincing explanation of the quantitative properties of the Solar System. Its supporters believed that the tidal theory could overcome the major defect of the nebular hypothesis by showing at least qualitatively how most of the original angular momentum could have been given to the major planets rather than to the Sun.

But the tidal theory turned out to have serious defects as well. The American astronomer Henry Norris RUSSELL found two major objections. First, theories of stellar structure developed by A S EDDINGTON and others in the 1920s indicated that the gases drawn from the interior of the Sun would be at such a high temperature—on the order of a million degrees—that they would dissipate into space before they could condense into planets. Second, a simple dynamical calculation showed that it would be impossible for the tidal encounter to leave enough material in orbits at distances from the Sun corresponding to the giant planets. (In other words, like the nebular hypothesis it failed to explain the distribution of angular momentum in the Solar System.)

There was now *no* satisfactory theory of the origin of the Solar System, since the earlier objections to the nebular hypothesis still seemed valid.

Revival of monistic theories after 1940

New ideas and observations provided a plausible basis for reviving a monistic theory after 1940. For example, the Swedish plasma physicist Hannes ALFVÉN proposed that the early Sun had a strong magnetic field and was surrounded by an ionized gas. Lines of magnetic force, rotating with the Sun, would be trapped in the ionized gas and transfer angular momentum to it. In this way one might explain why most of the angular momentum of the present Solar System is carried by the giant planets, rather than by the Sun. Alfvén's 'magnetic braking' mechanism to slow the Sun's rotation was adopted by several theorists in the 1950s and 1960s even though they rejected other aspects of Alfvén's theory of planetary formation.

Another problem with the nebular hypothesis was the implausibility of the assumption that the nebula would start to condense by itself. As Jeans and the French mathematician Henri POINCARÉ had pointed out early in the 20th century, if the total mass of the planets and Sun were uniformly distributed throughout the entire volume of the Solar System, the density would be so small that (unless the temperature is very low) the atoms would simply dissipate into space rather than condensing. This difficulty could be avoided if the nebula had originally been much more massive, then (after the condensation process started) lost most of its mass. The clue as to how

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

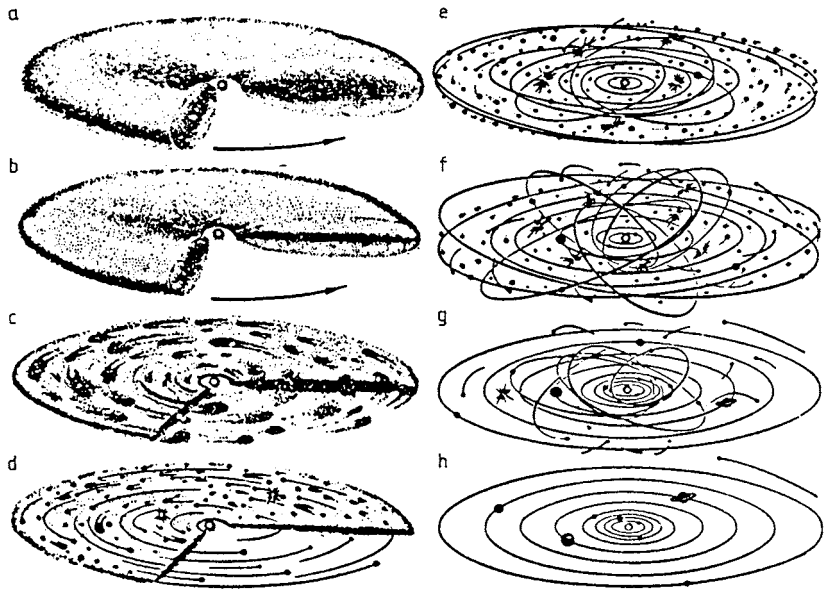


Figure 4. Origin of the Solar System according to Schmidt, Safronov and colleagues. The primordial gas–dust cloud begins its evolution somewhat as described by Laplace but does not form rings and large protoplanets at an early stage. Instead it condenses to small solid particles (planetesimals) which then aggregate to the planets.

this could have happened was furnished by Cecilia PAYNE-GAPOSCHKIN's discovery, in 1925, that the Sun is mostly hydrogen and helium. Yet both planets and Sun are supposed to be formed from the same nebula. Since the planets are composed mostly of heavier elements, perhaps the nebula started with a large amount of hydrogen and helium (the same proportion as in the present-day Sun) but lost it after the formation of the planets. That idea provided the basis for monistic theories proposed in the 1940s and 1950s by C von Weizsäcker, G P KUIPER, and H C UREY.

The revival of monistic theories began with a 1944 paper by the German astrophysicist C F von Weizsäcker. At the same time Alfvén and the Russian astronomer Otto Schmidt developed theories that were dualistic, insofar as they postulated previously formed clouds of material captured by the Sun, but their theories concentrated on the subsequent development of the Solar System and eventually became monistic theories.

Schmidt's theory was developed by Victor Safronov and other Soviet scientists throughout the 1960s and 1970s. It became primarily a model for the accumulation of small solid particles (Chamberlin's planetesimals) from the protoplanetary cloud into planets. Safronov's model was adopted, with some modifications, by the American geophysicist George W Wetherill, who explored its consequences with the help of computer calculations. The Safronov–Wetherill model is now considered the most plausible one for the formation of the terrestrial planets, though it does not yet account quantitatively for their properties (see figure 4).

After the 1950s, a major role in ideas about the origin of the solar system was played by 'ISOTOPIC ANOMALIES'. These are discrepancies between the average abundances of isotopes estimated for the Solar System and the abundances found in particular meteorites. Although these anomalies have little bearing on most of the traditional problems of planet formation, they were believed to offer important clues to the initial stages of formation and contraction of the solar nebula, relating them to nuclear processes in the Sun and other stars. The best-known example was the 'supernova trigger' hypothesis, based in part on the excess ^{26}Mg found in the ALLENDE METEORITE which fell in 1969. The anomaly was established by Typhoon Lee, D A Papanastassiou and G J Wasserburg at Caltech in 1976, and was ascribed to the earlier presence in the Solar System of ^{26}Al, which has a half- life of only 700 000 years. It was generally supposed that the ^{26}Al must have been synthesized in a supernova less than a few million years before the formation of the Solar System. Since a supernova explosion also produces a shock wave that might compress rarefied material to densities high enough for them to become unstable against gravitational collapse, the isotopic anomalies might indicate that the supernova *caused* the Solar System to form.

The most influential advocate of the supernova trigger hypothesis was the Canadian–American astrophysicist Alistair Graham Walter Cameron. But Cameron abandoned the hypothesis in 1984 when astrophysical observations indicated that other modes of synthesis were probably responsible for producing ^{26}Al. (He revived it later for other reasons.) In the meantime Cameron had developed a quantitative theory of the collapse of the solar nebula and the formation of protoplanets. He argued that turbulent viscosity, rather than magnetic braking, is primarily

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

responsible for the transfer of angular momentum from Sun to planets.

In the 1990s, theorists were challenged to explain the proliferating observations of extrasolar planets (see EXOPLANETS). These discoveries, suggesting that many stars are surrounded by planets, seem to favor the monistic theory, although they do not exclude the possibility that supernovae could have triggered the formation of several systems at once. But the assumption, based on our own Solar System, that every system would have a few dense 'terrestrial' planets in orbits close to the central star, with gaseous 'giant' planets located further out, has been thrown into doubt.

It now seems likely that the next few decades will see a revolution in theories of the origin of planetary systems, with important consequences for our understanding of our own system: which of its features are the result of general laws, and which are accidental?

Bibliography

Ahrens T J 1994 The origin of the Earth *Phys. Today* **47** (8) 38–45

Brush S G 1996 *A History of Modern Planetary Physics* (Cambridge: Cambridge University Press)

Fisher D E 1987 *The Birth of the Earth: A Wanderlied through Space, Time, and the Human Imagination* (Princeton, NJ: Columbia University Press)

Levin A E and Brush S G 1995 *The Origin of the Solar System: Soviet Research 1925–1991* (New Yrok: AIP Press)

Levy E H and Lunine J L 1993 *Protostars and Planets. III* (Tucson, AZ: University of Arizona Press)

Stephen G Brush

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orion

(the Hunter; abbrev. Ori, gen. Orionis; area 594 sq. deg.) An equatorial constellation which lies between Taurus and Monoceros, and culminates at midnight in mid-December. Its origin dates back to Sumerian times, when it was identified with the hero Gilgamesh and his fight against the Bull of Heaven (represented by Taurus), but today it is associated with the son of Poseidon, in Greek mythology, a mighty hunter who is accompanied in the sky by his two dogs (the constellations Canis Major and Canis Minor) in pursuit of a hare (the constellation Lepus). The brightest stars of Orion were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A conspicuous constellation and one of the few that even remotely resembles the object after which it was named, Orion is easily recognized by a quadrilateral of first and second magnitude stars, α, γ, β and κ Orionis, marking the hunter's right and left shoulders, left foot and right knee, respectively, within which lies a trio of second-magnitude stars, δ, ε and ζ Orionis, in a straight line, marking his belt, and below the belt another line of fainter stars, 42, θ and ι Orionis, marking his sword.

The brightest stars in Orion are β Orionis (Rigel), a triple system consisting of two bluish-white (B8 and B5) components, magnitudes 0.2 and 6.8, separation 9.5″, the primary of which is the seventh brightest star in the sky, and a third, unseen component which revolves around the latter in a period of 9.86 days, α Orionis (Betelgeuse), a red giant semiregular variable (range 0.0–1.3, average period about 2335 days), which is usually considered to be the tenth brightest star in the sky, γ Orionis (Bellatrix), magnitude 1.6, ε Orionis (Alnilam), magnitude 1.7, ζ Orionis (Alnitak), a close binary with bluish-white (O9.5 and B0) components, magnitudes 1.8 and 4.0, separation 2.4″, κ Orionis (Saiph), magnitude 2.1, and δ Orionis (Mintaka), a multiple system consisting of a bluish-white (O9.5) eclipsing binary (range 2.14–2.26, period 5.73 days) with a third, very close companion, magnitude 3.8, separation 0.27″, and a fourth, bluish-white (B2) component, magnitude 6.3, separation 53″. There are nine other stars brighter than fourth magnitude.

Another interesting multiple system is σ Orionis, which consists of a very close binary with bluish-white (O9.5 and B0.5) components, magnitudes 3.8 and 5.2, separation 0.2″, period about 170 years, and three other bluish-white (A2, B2 and B2) components, magnitudes 8.8, 6.9 and 6.6, separations 11″, 13″ and 43″ respectively.

Other interesting objects include the middle star in the sword of Orion, θ Orionis, which is a wide double with components θ^1 and θ^2 135″ apart. Each component is itself multiple, θ^2 consisting of a bluish-white (O9.5) primary, magnitude 5.0, and a fainter, magnitude 8.3, secondary, separation 0.4″, while θ^1 has four main components: A (also known as V1016 Orionis), a bluish-white (O7) eclipsing binary (range 6.7–7.7, period 65.43 days), B (also known as BM Orionis), another bluish-white (B0) eclipsing binary (range 7.9–8.7, period 6.47 days), separation 8.8″, C, a bluish-white (O6) star, magnitude 5.1, separation 12.8″, and D, another bluish-white (B0.5) star, magnitude 6.7, separation 21.5″. The latter four stars, known as the Trapezium, are the brightest members of an open cluster at the heart of the Orion Nebula (M42, NGC 1976), one of the brightest emission nebulae in the sky, just visible to the unaided eye as a faint misty patch.

The Orion Nebula is an area of ionized hydrogen (HII) associated with a region of mostly molecular hydrogen (H_2) that forms part of the Orion Molecular Clouds, which extend across the constellation. These are believed to be sites of star formation, and infrared observations have detected the presence of several young stars and star clusters, such as the Becklin–Neugebauer (BN) object and the Kleinmann–Low (KL) cluster, embedded in the gas.

Another well-known nebula is the Horsehead Nebula (Barnard 33), an protuberance of dark nebulosity silhouetted against a swathe of emission nebulosity (IC 434) that extends south from ζ Orionis. There are no bright galaxies in Orion.

The Orionid meteor shower appears to radiate from the constellation.

See also: Becklin–Neugebauer object, Betelgeuse, Horsehead Nebula, Orion Nebula, Orionids, Rigel.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orion Arm

The local spiral arm of the Milky Way Galaxy within which the Sun and the star-forming region of Orion are embedded. The stars and clouds of gas and dust that comprise the Orion arm curve outwards (relative to the galactic center) from the general direction of the constellation of Cygnus to the constellation of Vela. The Orion region, which includes the hot young stars of the Orion O, B association, the Orion Nebula and the Orion molecular cloud, is the most prominent star-forming complex in the local arm. Sometimes referred to as the 'Orion spur', the Orion arm is believed to be an extended arm segment rather than a major arm that originates directly in the central bulge of the Galaxy.

See also: Milky Way Galaxy, Orion, Orion Nebula, spiral arms, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orion Nebula (M42, NGC 1976)

An emission nebula in the constellation Orion, position RA $05^{\mathrm{h}}\ 35.4^{\mathrm{m}}$, dec. $-05^{\circ}\,27'$. At fourth magnitude and measuring $66'$ by $60'$, it is the most prominent emission nebula in the sky and is visible to the naked eye. It surrounds the quadruple star θ^1 Orionis, known as the Trapezium, ultraviolet light from which energizes the nebula. The northwestern part of the nebula is designated M43 (NGC 1982). A dark lane of dust, named the Fish Mouth from its shape, separates M43 from M42. The Orion Nebula abuts onto OMC-1, the largest of many molecular clouds which occupy much of the constellation Orion. This is a region of active star formation, the Trapezium stars being some of many young, hot stars associated with the nebulosity. The Becklin–Neugebauer Object and the Kleinmann–Low Nebula lie behind the Orion Nebula.

See also: Becklin–Neugebauer object, Kleinmann–Low Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orionids

A meteor shower that takes place in late October. The radiant lies in the constellation Orion, near the border with Gemini. The Orionids occur when the Earth intersects the ascending node of the meteor stream from Halley's Comet; the Eta Aquarids in May are produced by the Earth's passage through the descending node. Because Halley's orbit is retrograde, Orionid meteoroids impact the Earth at a high relative velocity and produce very fast meteors. This is a less active shower than the Eta Aquarids, because in October the Earth encounters the meteor stream off-center.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Orrery

A mechanical model that illustrates the relative positions and motions of bodies in the solar system. The name derives from the fourth Earl of Orrery, Charles Boyle, for whom such a device was constructed, probably by George Graham (d. 1751) about 1725.

See also: planetarium, planisphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oscillating Universe

A universe that expands and contracts in a cyclic fashion. The universe is expanding at present but the rate of expansion is generally believed to be slowing dow. If the expansion comes to a halt at some time in the future, the universe will then begin to contract, slowly at first, then ever more rapidly until it collapses back into a hot, highly compressed state (the 'big crunch'). The oscillating universe theory suggests that the collapse of the universe is followed by a new 'Big Bang' and a new cycle of expansion and contraction. If this theory is valid, the current cycle of expansion and contraction may have been preceded by an indefinite, possibly infinite, number of cycles and will be followed by an indefinite number of cycles.

There is no known physical process that could cause a collapsing universe to 'rebound' in this way, although some have suggested that a hitherto unknown force or forces may come into play to prevent the universe from collapsing to a genuine singularity (a point of infinite compression) and to trigger a new expansion phase. It has also been argued that if the universe is indeed oscillating, then each successive cycle will expand further, and last longer, than its predecessor. This argument is based on the contention that because the evolution of stars converts matter to radiation as the universe ages, and the radiation content of the universe is blue-shifted to higher energies during the collapse phase, the energy content of the 'big crunch' and the subsequent expansion cycle will be greater than that of the previous 'Big Bang'.

See also: Big Bang theory, cosmological model, expanding universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Osiander, Andreas (1498–1552)

A theologian, born Gunzenhausen, Bavaria, Germany. A minister, Osiander had mathematics as his hobby and came into contact with COPERNICUS. In his old age, Copernicus had been persuaded to publish his theory of the heliocentric universe. He entrusted the publication of his work *De Revolutionibus* to GEORGE JOACHIM RHETICUS. The young Rheticus was forced to leave Nuremberg by a particularly attractive offer which he could not refuse of a job with a high salary, and, even knowing Osiander's prejudices against the motion of the Earth, he left Osiander in charge of the publication of the book. It was Osiander who inserted an anonymous preface to Copernicus' book, denying that the book intended to propose more than a mathematical hypothesis, useful for calculations—'not necessarily true or even probable'. This was at first damaging to Copernicus' credibility. In Rheticus' own copy of the book he has angrily crossed out Osiander's preface with red crayon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Osservatorio Astronomico di Bologna

Astronomy in Bologna has a long history. Official records of systematic teaching of astronomy at the Bologna 'Studio' date back to 1297. Records of systematic observing work can be traced back to the beginning of the seventeenth century, when G D Cassini built a huge sundial in the major church of the city.

The present institute has a staff of about 70 people, half of whom are scientists. It manages a 1.5 m telescope in Loiano. Research activity is mainly addressed to stellar population, extragalactic astrophysics, cosmology and high-energy astrophysics. The Observatory is affiliated to the National Institute of Astrophysics (INAF), the main financial support coming from the National Ministry for Scientific and Technological Research.

For further information see
http://www.bo.astro.it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Osservatorio Astronomico di Brera

The Observatory (OAB) is a world leader in x-ray mirror technology. (The mirrors of the Beppo-SAX, JET-X, and XMM satellites have been made using the electro-forming technology developed in Milan.) Important progress is now being made also in x-ray mirror multilayer technology.

In ground-based instrumentation, OAB is part of both national and international teams to build various spectrographs: Dolores for the Italian National Telescope, NIMOS, VIMOS and AVES for the ESO-VLT. The research issues mainly address observational cosmology, high-energy astrophysics and stellar astronomy. The ESP survey has allowed a robust estimate of the local luminosity function for galaxies, and the Reflex survey is giving fundamental data about the large scale structure of the universe. A group of Brera scientists was the first to detect, using Beppo-SAX alert and the ESO-VLT, polarization from gamma ray bursts (GRBs). The Observatory is now involved, in collaboration with the USA and the UK, in a satellite (Swift) designed for detection and study of GRBs.

The Astronomical Observatory of Brera, located in Milano Via Brera 28, dates back to the year 1764 when it was conceived by Ruggiero Boscovich, who became the first Director in 1770. Later, in 1923, the construction started of the observing station in Merate and, in 1926, the first telescope, a Zeiss reflector of 102 cm aperture, started operations.

At present the Observatory consists of two Institutes, one in Milan and one in Merate. The administration and accounting are kept in Milan while all the laboratories—clean rooms, advanced metrology and design—are situated in Merate. The present staff consists of 30 astronomers; 38 staff technicians, research engineers, and administrative personnel; 5 post-docs and about 15 students. A very tight collaboration exists with the universities of Milan, Como and Pavia.

Funding originates from the Ministry of Research and Technology (MURST), from the Italian Space Agency (ASI), from the Italian Consortium of Astronomy and Astrophysics (CNAA) and from the European Community and European Southern Observatory in the case of finalized projects.

For further information see
http://www.merate.mi.astro.it or
http://www.brera.mi.astro.it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Osservatorio Astronomico di Roma

The Osservatorio Astronomico di Roma is a non-profit institute dedicated to astrophysical research. It was founded in the early nineteenth century, and moved to its present residence, the ancient Villa Mellini, not far from St Peter's Cathedral, in 1935.

It has also acquired a second site at Monteporzio Catone, in the scientific area of Frascati, close to Rome. Completed in 1965, the Frascati buildings were originally intended to serve as headquarters of the national observational facilities, but the rapid expansion of the Rome urban area forced to abandon this project. The Rome Observatory was then entrusted with its management.

The observatory also runs a mountain station at Campo Imperatore. Its 60–90 cm Schmidt telescope is used to search for supernovae.

Research interests include globular clusters, the distribution of galaxies on cosmological scales and the distribution and properties of Lyman-α clouds in the early universe. Theoretical work covers solar activity, binary stars, active galactic nuclei, the properties of galaxy clusters in the optical and x-ray bands and open stellar clusters.

Instrument work concerns the development of CCDs and the development of infrared spectral instruments.

The Astronomical Museum of the Observatory was originally assembled in 1873 as a Copernican Museum on the occasion of the fourth centenary of Nicholas Copernicus' birth.

For further information see
http://oar.rm.astro.it/home.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Outer Planets

A collective term for the major planets Jupiter, Saturn, Uranus, Neptune and Pluto: those whose orbits lie farthest from the Sun, beyond the main asteroid belt, as opposed to the inner planets, orbiting inside the main asteroid belt.

Compare: superior planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Owens Valley Radio Observatory

The Owens Valley Radio Observatory (OVRO) is located in a deep dry valley 400 km north of Pasadena, California, where the California Institute of Technology—which operates the Observatory—is located. It was founded in the mid 1950s and the first major instrument was an interferometer composed of two 26 m antennas. These dishes moved on east–west and north–south tracks and so the pair provided excellent resolution in two dimensions. At the time it was the largest such system in the world, and made many high-resolution studies especially of extragalactic radio sources. In the late 1960s a 40 m paraboloid was constructed. For many years it was the major west-coast station of the Very Long Baseline Interferometer, which linked radio telescopes around the world to get angular resolution as fine as one-thousandth of an arcsec. More recently it has been used to study pulsars and the microwave background radiation.

At present the major instrument at OVRO is an interferometer comprising six 10 m telescopes, and operating at millimeter wavelengths. These dishes are on tracks and the system can synthesize a 0.5 arcsec beam at 1 mm wavelength. Extensive studies are being undertaken in star formation, molecular studies in the Milky Way and other galaxies, and in studies of galaxies at high redshift.

For further information see
http://www.ovro.caltech.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Owl Nebula (M97, NGC 3587)

A planetary nebula in the constellation Ursa Major, position RA 11^{h} 14.8^{m}, dec. $+55^{\circ}$ $01'$. The Owl is $3'$ across and gets its name from two adjacent dark patches that have the appearance of large eyes. The nebula is eleventh magnitude, and the central star is a faint magnitude 16.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Oxygen (O)

A gas that makes up 20.95% by volume of the Earth's atmosphere at ground level, 89% by weight in seawater and 46.6% in the Earth's crust. It appears to be the third most abundant element in the universe (after hydrogen and helium), but its abundance—in terms of number of atoms—is only about 1/1500 that of hydrogen. The chemical symbol is O, and it normally occurs in the atmosphere in molecular form (two atoms linked), O_2. Triatomic oxygen, O_3, called ozone, and monatomic oxygen, O, are more predominant in the upper atmosphere, where ozone shields the Earth from the Sun's ultraviolet radiation. Pure oxygen is 1.1 times heavier than air. Its atomic mass is 15.99 and its atomic number 8. Oxygen molecules are essential to the sustenance of human and animal life on Earth. Oxygen combines with most other elements and, when this reaction occurs rapidly with the release of heat and light, we describe the reaction as combustion, or burning. Oxygen liquefies under normal pressure at about $-183\,^{\circ}$C and it becomes solid at about $-218\,^{\circ}$C. Liquid oxygen (LOX) is commonly used as the oxidant in the propellant of chemical rockets.

Free molecular oxygen is almost entirely absent from the atmospheres of Venus and Mars.

See also: cosmic abundance of elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ozone Hole

The ozone hole is a recent seasonal phenomenon which corresponds to a significant reduction in the total amount of ozone in the EARTH'S ATMOSPHERE during springtime over Antarctica. Evidence that this phenomenon is caused by human activities through the release into the atmosphere of man-made chlorine and bromine compounds has been building up over the last 15 years. The depletion of the ozone layer is not restricted to Antarctica, but unlike the sudden and near total loss of ozone over this region at certain altitudes, ozone depletion at mid-latitudes is much less and much slower (only a few per cent per year). It is nevertheless a worrying trend which is the subject of intense scientific research at present. In view of this global phenomenon, regulations to restrict the release of man-made chlorine and bromine compounds were promptly decided with the Montreal Protocol in 1987 and successive amendments in 1990 and 1992. Full compliance with the protocol should lead to a recovery of the ozone layer by the middle of the twenty-first century. However, other environmental changes such as CLIMATE changes might impact the time of the recovery.

The ozone layer

Ozone (O_3: three OXYGEN atoms) occurs naturally in the atmosphere. It is mainly found in two regions of the Earth's atmosphere (see figure 1). Most ozone (about 90%) resides in a layer located in the 10–50 km altitude range, a region called the *stratosphere*. This stratospheric ozone is commonly known as the 'ozone layer'. The remaining ozone is located in the *troposphere*, the region which extends from the Earth's surface up to about 10 km. Even in the stratosphere, ozone represents a very small fraction of the atmospheric gas with a maximum of about six ozone molecules for every million air molecules. Despite this small amount, ozone plays a vital role in the atmosphere: stratospheric ozone absorbs most of the biologically damaging ultraviolet sunlight (called UV-B), allowing only a small amount to reach the Earth's surface. The absorption of UV radiation by ozone creates a source of heat, which actually forms the stratosphere itself. Ozone thus plays a key role in the temperature structure of the Earth's atmosphere. On the other hand, ozone at the Earth's surface can be harmful to living systems: it reacts strongly with other molecules and high levels of ozone are toxic to living organisms and biological life.

The amount of ozone in the Earth's atmosphere is governed by the balance between permanent ozone production through molecular oxygen photolysis and ozone destruction through chemical processes which involve numerous constituents whose abundance can be very low (less than one molecule for every billion air molecules). Besides, although ozone is mainly formed in the tropics where the solar radiation is strongest, the ozone layer is thinnest in the tropics (around the equator) and denser towards the poles due to the transport processes which redistribute the ozone over the globe. The amount of ozone above a point on the Earth's surface is measured in DOBSON UNITS (DU), typically 300 DU in global average, with large seasonal fluctuations. This quantity corresponds to the integrated ozone column density above one point, generally called 'total ozone'. As a convention, it is reduced to the standard pressure and temperature conditions of the Earth's surface (respectively 1 atmosphere and 0 °C). In these conditions, the thickness of a pure ozone layer above the Earth is about 3 mm. As a comparison, the total thickness of the atmosphere defined in the same conditions amounts to 8 km. The Dobson unit is defined as 10^{-5} m of ozone at standard pressure and temperature conditions.

The first theory on ozone equilibrium was introduced by Sydney Chapman around 1930. According to this theory, ozone equilibrium is governed by a cycle referred to now as the Chapman cycle. According to this cycle, production occurs when ultraviolet light breaks down oxygen molecules, while destruction involves chemical reactions among oxygen compounds:

$$O_2 + \text{radiation} \rightarrow O + O \quad (1)$$

$$O + O_2 + M \rightarrow O_3 + M \quad (2)$$

$$O_3 + \text{radiation} \rightarrow O_2 + O \quad (3)$$

$$O_3 + O \rightarrow O_2 + O_2 \quad (4)$$

where M is any air molecule.

Later ozone measurements showed that this cycle produces amounts of ozone which are too high. This led to the introduction of other atmospheric compounds to balance ozone production by reactions (1) and (2). The concept of a catalytic cycle was then introduced, where a constituent X participates in ozone destruction through the following reactions:

$$X + O_3 \rightarrow XO + O_2 \quad (5)$$

$$XO + O \rightarrow X + O_2. \quad (6)$$

The net budget of reactions (5) and (6) corresponds to reaction (4) which destroys ozone and X is re-formed at the end of the cycle. The main components acting on ozone equilibrium involve hydrogen, oxygen, nitrogen, chlorine and bromine compounds. These constituents are produced chemically from source gases such as methane, water vapor, nitrous oxide (N_2O), chlorine and bromine compounds, including the man-made chlorofluorocarbons or CFCs and the halons which contain bromine. Chlorofluorocarbons were mainly used in refrigerators, air conditioners and as industrial solvents. After their emission into the atmosphere through physical, chemical or biological processes, these gases, chemically inert in the lower atmosphere, diffuse towards the stratosphere due to the permanent mixing of air in the atmosphere. After 7 to 10 years, they reach the stratosphere where the molecules are broken down by the more and more intense solar radiation as a function

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

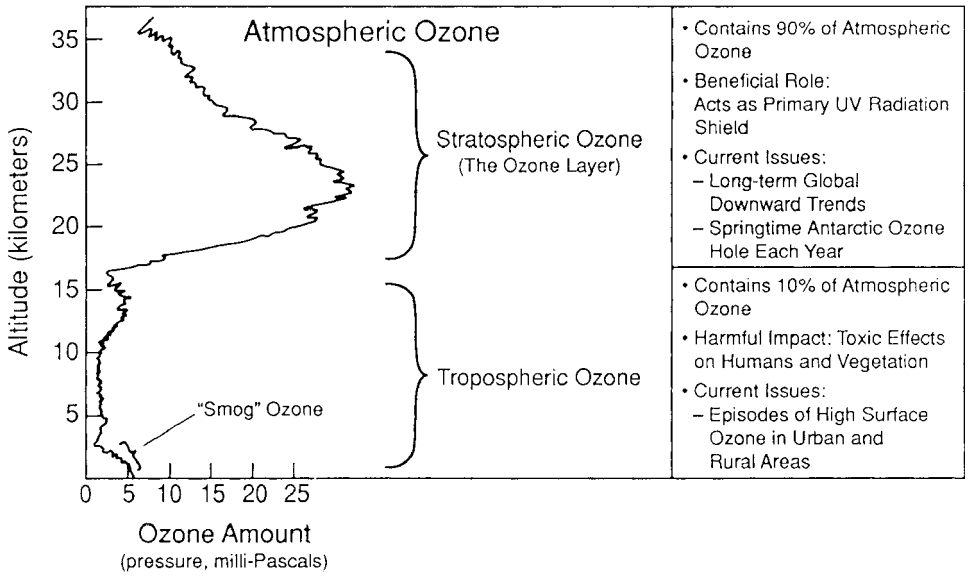


Figure 1. The ozone layer.

of altitude. Chemically active compounds are produced, which destroy ozone through chain reactions at the end of which they are re-formed. This means that one molecule of these active compounds can destroy 1000 to 10 000 ozone molecules, before being eliminated by recombination processes.

The ozone hole

Worries about the fate of the ozone layer first arose in the 1970s when nitrogen and chlorine compounds involving man-made products were introduced into the ozone catalytic cycles. The role of nitrogen compounds was suggested by Paul Crutzen in 1970 while Molina and Rowland pointed out the danger of increasing emission of chlorofluorocarbons in 1974. A ban on aerosol sprays containing CFCs was made by the USA in the late 1970s. These investigations, as well as their subsequent studies on the ozone hole, led to the award to these authors of the Nobel Prize for Chemistry in 1995.

Up to the mid-1980s, all the analyses of the impact of increasing chlorine levels on the ozone layer led to rather small decreases, so the first report on the Antarctic ozone hole was mostly unexpected. It was first reported in 1984–5 that ozone losses reached 50% as compared with the measurements performed in October during the previous decades from 1957 to 1975 (see figure 2). Such measurements were obtained in the Syowa station and Halley Bay Antarctic station. Subsequent satellite measurements showed that the ozone hole was spread over the whole Antarctic continent. In parallel, measurements of the ozone vertical profile revealed a destruction of more than 95% in the 16–22 km altitude range (figure 3). The intensive experimental and theoretical investigations which followed the discovery of the ozone hole allowed the link between the man-made chlorine and bromine compounds and the rapid ozone depletion during the Antarctic springtime to be precisely established: they showed that the ozone hole results from the conjunction of two conditions—high levels of stratospheric chlorine and bromine and the specific meteorological conditions which prevail in the polar stratosphere during winter.

The first condition is ensured by the continuous horizontal and vertical mixing of the Earth's atmosphere over the globe. Thus, while human emission of CFCs and halons occurred mainly in the northern hemisphere, these constituents are present in the Antarctic stratosphere, where they have been measured. After emission at the Earth's surface, they are rapidly mixed in the troposphere. This well-mixed air rises into the stratosphere essentially at tropical latitudes. Winds then move this air poleward, so that air throughout the stratosphere contains nearly the same amount of chlorine and bromine.

The second condition is due to the particular location of the polar stratosphere: during the winter polar night, sunlight does not reach the south pole, leading to the development of strong circumpolar winds in the middle to lower stratosphere. These strong winds are known as the 'polar vortex' which isolates the air over the polar region from the surrounding mid-latitude regions. Since there is no sunlight, the air within the polar vortex can get very cold, so cold that special clouds can form once the air temperature gets below about $-80\,^{\circ}$C. These clouds are called polar stratospheric clouds (PSCs). These PSCs are formed of liquid and solid particles whose composition includes nitric acid, sulfuric acid and water. Their exact composition and formation temperature is still the subject of intense scientific debate. As the temperature gets colder, below $-85\,^{\circ}$C, larger droplets of water-ice with nitric acid

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

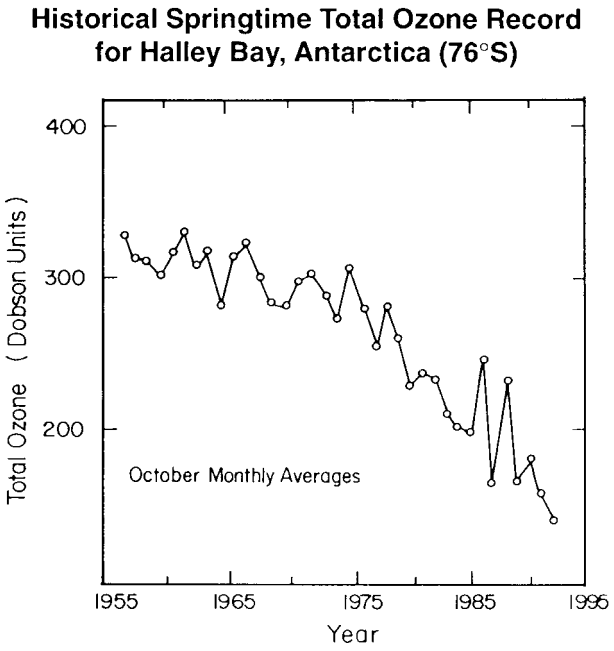


Figure 2. The decrease of the total amount of ozone in October at Halley Bay.

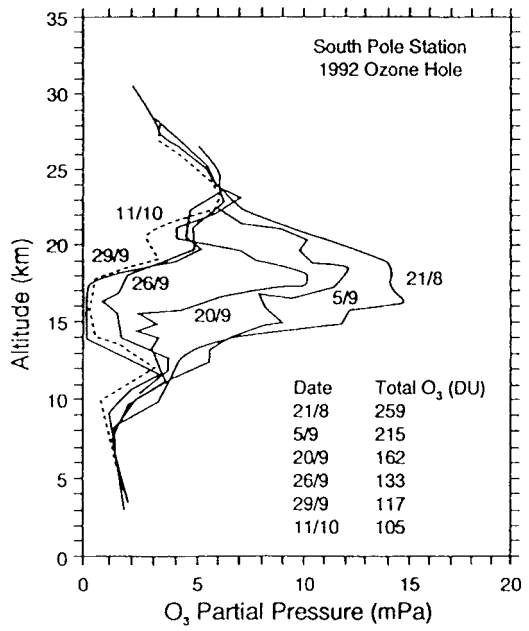


Figure 3. The evolution of the vertical distribution of ozone during the ozone hole period in Antarctica.

dissolved in them can form and freeze. These various PSCs are crucial for ozone loss to occur, since the presence of these particles allows the onset of a particular chemistry called 'heterogeneous chemistry' which transforms most of the chlorine compounds into potentially active chlorine, likely to destroy ozone. The most important reactions in the preconditioning of the polar stratosphere during winter are

$$HCl + ClONO_2 \rightarrow HNO_3 + Cl_2 \quad (7)$$

$$ClONO_2 + H_2O \rightarrow HNO_3 + HOCl \quad (8)$$

$$HCl + HOCl \rightarrow H_2O + Cl_2 \quad (9)$$

$$N_2O_5 + H_2O \rightarrow 2HNO_3. \quad (10)$$

These reactions are very fast and can only take place on the surface of polar stratospheric clouds. This is why the ozone hole was not predicted. Heterogeneous reactions were neglected in atmospheric chemistry (at least in the stratosphere) before the ozone hole was discovered. The nitric acid (HNO_3) formed in these reactions remains in the PSC particles, so that the gas phase concentrations of nitrogen oxides are reduced. This reduction, called 'denoxification', is important as it slows down the removal rate of ClO that would otherwise occur by the reaction

$$ClO + NO_2 + M \rightarrow ClONO_2 + M \quad (11)$$

where M is any air molecule.

Without reaction (6), high levels of active chlorine (multiplied by a factor of about 1000) are maintained. The destruction of ozone requires the return of sunlight over the pole. First, HOCl and Cl_2 are photodissociated:

$$Cl_2 + \text{radiation} \rightarrow Cl + Cl \quad (12)$$

$$HOCl + \text{radiation} \rightarrow Cl + OH. \quad (13)$$

Then, ozone destruction takes place through catalytic cycles which allow the re-formation of the destroying compounds. Thus, active chlorine constituents which are about 1000 time less abundant than ozone can destroy nearly all the ozone molecules in the lower stratosphere. Two main cycles are considered to be responsible for the ozone destruction:

(I)
$$ClO + ClO + M \rightarrow Cl_2O_2 + M$$
$$Cl_2O_2 + \text{radiation} \rightarrow Cl + ClO_2$$
$$ClO_2 + M \rightarrow Cl + O_2 + M$$
$$2(Cl + O_3) \rightarrow 2(ClO + O_2)$$

net: $2O_3 \rightarrow 3O_2$.

(II)
$$ClO + BrO \rightarrow Br + Cl + O_2$$
$$Cl + O_3 \rightarrow ClO + O_2$$
$$Br + O_3 \rightarrow BrO + O_2$$

net: $2O_3 \rightarrow 3O_2$.

The dimer (Cl_2O_2) of the chlorine monoxide radical involved in cycle I is thermally unstable, and the cycle is most effective at low temperatures. It is thought to be responsible for most (70%) of the ozone loss in Antarctica. In the warmer Arctic a large proportion of the loss may be driven by cycle II. These cycles lead to a very rapid ozone decrease which reach more than 1% per day at some altitude levels. This destruction is interrupted at the end of the spring when the general stratospheric warming breaks the polar vortex and the isolation of the polar airmasses. The ozone hole is then diluted in the southern hemisphere leading to a net global ozone loss.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

What is happening in the Arctic?

Significant reduction in ozone content in the stratosphere above the Arctic has been observed during the late winter and early spring in five of the last eight years (1990, 1993, 1995, 1996 and 1997). However, these reductions, which can reach 25%, did not last as long and occurred over a much smaller area than the Antarctic ozone hole. The difference between the two polar regions is due to the different weather patterns: Antarctica is a very large land mass surrounded by oceans. These symmetrical conditions produce very low stratospheric temperature within the meteorologically isolated polar vortex region. The conditions that maintain elevated levels of chemically active chlorine and bromine persist into September and October when sunlight returns over the region to initiate ozone depletion. In the Arctic, large-scale weather systems disturb this main flow, making it less stable than over Antarctica. These disturbances prevent the temperature from being as cold as in the Antarctic stratosphere and fewer polar stratospheric clouds are formed. Nevertheless, chemically active chlorine and bromine compounds are also formed from reaction on the surface of the clouds. But the cold conditions rarely persist into March when sufficient sunlight is available to trigger substantial ozone depletion. Several American and European scientific campaigns have been organized since the beginning of the 1990s in order to evaluate ozone depletion in the Arctic and evaluate its possible impact on the surrounding mid-latitude regions. Such campaigns generally involve measurements from numerous ground-based stations, large stratospheric balloons, aircraft, dedicated satellite platforms and numerical simulations of the atmosphere in order to interpret the experimental results. The last major campaign, called THESEO, was organized in winter 1998–9. It was the third major European campaign to study the effects of CFCs and other ozone-depleting substances on the stratosphere.

Evolution of the ozone layer at mid-latitude

The last evaluation of the stratospheric ozone trends in mid-latitude regions show that between 25 and 60°N, ozone abundance over the last years averaged about 4% below 1979 values. The winter/spring and summer/fall ozone losses are of the order of 5.5% and 2.8% respectively. As in the polar regions, the ozone depletion occurs mainly in the lower stratosphere. It is due to the dilution after the break up of the polar vortex at the end of spring, of the polar airmasses in which ozone was destroyed and to heterogeneous reactions which take place on the surface of the stratospheric aerosol layer which exists at all latitudes around 18 km. In the latter case, the processes are similar to the ones which occur at high latitudes, but the global surface of the stratospheric particles is much less than the PSC one.

Regulations for protecting the ozone layer

In 1987, the recognition of the potential for chlorine and bromine compounds to destroy stratospheric ozone led to the 'Montreal Protocol on Substances that Deplete the Ozone Layer' which aimed at reducing the global production of ozone-depleting substances. Subsequently, global observations of significant ozone depletion have prompted amendments to strengthen the treaty. The 1990 London Amendments call for a ban of the most damaging compounds by 2000. The 1992 Copenhagen Amendments changed the date of the ban to 1996. Only some exceptions for the production of halocarbons were tolerated for essential uses, such as for medical sprays, before safe substitutes were found. Further restrictions on ozone depleting substances such as methyl bromide and strengthening measures to ensure compliance with the Montreal Protocol have been agreed upon in Vienna (1995) and Montreal (1997). Full compliance should lead to a recovery of the ozone layer by the middle of the twenty-first century. However, other environmental changes such as climate changes might impact the exact time of the recovery.

Bibliography

As a requirement of the Montreal Protocol, extensive reports on the latest findings concerning the ozone hole layer are produced regularly by the scientific community. The latest reports are the following:

World Meteorological Organization 1998 Global ozone research and monitoring project, assessment of trends in the vertical distribution of ozone *Report* 43 (Geneva: WMO)

World Meteorological Organization 1999 Global ozone research and monitoring project, scientific assessment of ozone depletion (Geneva: WMO) at press

Sophie Godin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Padova/Asiago Observatories

The Asiago Highlands, about 90 km north-west of Padova, at an elevation of about 1000 m, host, in two sites, the main complex of optical telescopes operating in Italy. The telescopes belong to two independent institutions, based in Padova: the Astronomical Observatory (OAP) and the Astronomy Department of the local University.

The OAP, a research institute, was founded in 1767 by the Venetian Republic and located in a tower built by the tyrant Ezzelino da Romano in 1242, and restructured by the architect Domenico Cerato.

Research in the eighteenth and nineteenth centuries dealt primarily with meteorology, geodesy, compilation of stellar catalogs, and determination of orbits of solar system bodies. The construction by the Padova University of the 122 cm reflector, then the largest telescope in Europe, completed in 1942, fostered the development of modern astrophysical research in Padova/Asiago. Three more telescopes were subsequently built by OAP: a Schmidt 50/40 (in 1958), a Schmidt 92/67 (in 1965) and the 182 cm 'Copernicus' reflector (in 1973). At present they are all located at the top of Cima Ekar (elevation 1350 m). A large collection of plates and spectra was secured over several decades of observations: about 10 000 direct photographs and more than 18 000 spectra taken with the 122 cm reflector; 14 500 and 18 000 wide field plates and films with the two Schmidt telescopes.

Research groups (both observational and theoretical) cover a broad range of subjects: solar system, stellar astrophysics, external galaxies, active galactic nuclei, cosmology, and the history of astronomy. The Observatory played, and is still playing, a primary role in major technological projects: construction and commissioning of the 'Galileo' telescope at La Palma (Canary Islands) and of its instruments (optical imager, high-resolution spectrograph, adaptive optics), realization of the camera for the VST and of the LBT prime focus imager, ESO VLT–LGS program; development of innovative methods for handling very large databases. It also has a significant involvment in several space projects (Rosetta, Planck Surveyor, Gaia). Also very important are the educational activities exploiting the Web—in 1998 OAP was awarded the 'New Media Prize' for the best educational site for children.

Padova is one of two Italian universities having a degree (Laurea) in astronomy (since 1968). Since 1983 it has also run PhD courses in astronomy.

The Observatory staff is: 38 astronomers (5 senior, 14 associates, 19 researchers), 38 technicians, and 28 employees for administration and general services.

The Department staff is: 16 professors (6 full, 10 associate), 8 researchers, 25 PhD students and post-doctorates; 21 employees.

For further information see
http://www.pd.astro.it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Palimpsest

A bright circular feature on an icy surface with very little vertical relief. The term is used to describe features on the outer three Galilean satellites of Jupiter; it originally meant a piece of parchment prepared for reuse by erasing what had been written on it. Palimpsests are interpreted as impact features whose original morphology has been smoothed out. The impacts are supposed to have occurred when the crust was thin, and subsurface 'slush' rapidly extruded out to fill the crater; alternatively, the raised rim and depressed floor of the crater were able to 'relax' towards the mean surface layer since the lack of rigidity of the underlying material meant that it could not support surface relief. Palimpsests are typically 100 km in diameter. They are found on Callisto at the centers of the large impact features Asgrad and Valhalla; and on Europa (e.g., the 148 km diameter Tyre Macula), where they have almost no discernible surface relief, and on Ganymede (e.g., the 348 km diameter Memphis Facula).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Palitzsch, Johann Georg (1723–88)

German farmer-astronomer in Dresden who recovered Halley's Comet at its predicted reappearance in 1758. During the 1761 transit of Venus across the Sun, he observed a black band linking Venus to the edge of the Sun at the beginning and end of the transit (the 'tear-drop effect') and correctly interpreted this as an effect of Venus' atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pallas

The second asteroid to be discovered, by Wilhelm Olbers in 1802, and so designated (2) Pallas. It is also the second largest asteroid, irregular in shape with an average diameter of 525 km. It orbits the Sun in the main asteroid belt at a mean distance of 2.77 AU (415 million km) in a period of 4.61 years; the orbital inclination is 35°, high for a main-belt asteroid, and the eccentricity 0.23. Pallas rotates in 7.81 h. It is a B-type asteroid, with a reflectance spectrum similar to that of carbonaceous chondrites, but rather brighter, with an albedo of 0.16.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Palomar Observatory

The observatory is located on Mount Palomar, between Los Angeles and San Diego. It is owned and operated by the California Institute of Technology, a privately endowed educational and research institution located in Pasadena, and is used to support the scientific research programs of Caltech's faculty and students.

The principal instrument at Palomar is the 5 m Hale Telescope, which from 1948 to 1976 was the largest optical telescope in the world. Palomar also houses the 1.2 m Oschin Telescope, the 46 cm Schmidt telescope, and a 1.5 m reflecting telescope (operated jointly by Caltech and the Carnegie Institute of Washington). By agreement, Caltech shares use of the Hale Telescope with astronomers from Cornell University and the Jet Propulsion Laboratory.

Through the efforts of George Ellery Hale, the International Education Board (one of the Rockefeller Foundations) awarded a grant to Caltech for the construction of a 5 m telescope in 1928. Mount Palomar was selected as the optimum site for the new instrument in 1934.

The 5 m Pyrex glass disk was cast on 2 December 1934, and after a cooling period of eight months, the 20 ton disk was shipped by rail to Pasadena for the long process of grinding and polishing that would transform it to the precise shape and size required for astronomical observations.

Construction of the building (including the 1000 ton rotating dome) and the telescope structure (the moving parts weigh about 530 tons) began in the mid 1930s and was nearly complete by 1941 when the United States entered the Second World War. But the war delayed polishing of the mirror, and it was not until 18 November 1947 that the finished mirror, now weighing only 14.5 tons, began its two-day trip to Mount Palomar where it was installed in the telescope for a period of testing and adjustment.

Scientific research at Palomar Observatory has been undertaken since 1948. The Hale Telescope has been used on virtually every clear night to study objects ranging from nearby asteroids and comets, to the stars that comprise the Milky Way, distant galaxies and quasars.

The 5 m telescope has recently been upgraded. The telescope is fitted with CCD cameras which are 100 times more sensitive than the photographic plates used when the telescope first went into service. A new adaptive optics system corrects for the effects of air turbulence by using a mirror placed between the telescope and camera and which is adjusted up to 500 times a second. An infrared camera, the Palomar High Angular Resolution Observer (PHARO), has also been installed.

Although electronic detectors have replaced photography at the 5 m telescope, the 1.2 m Oschin Telescope, which is designed for wide-field viewing, still uses glass photographic plates to record images. Its Schmidt camera carried out the famous National Geographic–Palomar Sky Survey in the early 1950s, producing an atlas of the entire sky north of declination -33°. Each 6° square segment of the sky was photographed on both blue- and red-sensitive plates, a total of 879 pairs of plates.

A second sky survey was carried out in the early 1980s, and a third multi-year-long program to obtain complete photographic coverage of the northern sky is now under way. Unfortunately, light pollution is an increasing problem because of the rapid urbanization of southern California.

For further information see
http://astro.caltech.edu/observatories/palomar/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parabola

A conic section obtained by cutting a right circular cone by a plane parallel to the side of the cone. Such a curve is open (i.e. it does not form a closed shape such as an ellipse or circle), with eccentricity = 1, and stretches to an infinite distance with the two arms tending to become parallel. The parabola forms the boundary between closed ellipses and open hyperbolas.

See also: conic sections, ellipse , hyperbola, orbit, orbital elements, revolution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parallaxes

The trigonometric parallax (π) of a star is the angle subtended by 1 astronomical unit (AU) at the star's distance (d) from the Sun (see figure 1). As the Earth annually orbits the Sun, the observer's changing position causes an annual shift in the star's measured position, tracing a small ellipse on the sky that reflects the size and orientation of the Earth's orbit as it might be viewed from the star. The size of this 'parallactic ellipse' is inversely proportional to the distance. For convenience, the astronomer's unit of distance, the parsec (pc), is defined as the distance at which a star's parallax would be one second of arc (arcsec), the unit in which parallaxes are traditionally measured. This gives the simple relation $d = \pi^{-1}$.

The measurement of trigonometric parallaxes is fundamental to our knowledge of stellar distances in our Galaxy and is the first step in establishing the cosmic distance scale. However, most stars are so distant that their parallaxes are very small and difficult to measure accurately. The nearest known star, Proxima Centauri (α Cen C), has a parallax of 0.77 arcsec. A typical naked-eye star's parallax is about 0.01 arcsec. Most of the 118 000 parallaxes in the Hipparcos Catalogue are only a few milliarcseconds (mas) in size, barely larger than the errors of measurement (see HIPPARCOS AND TYCHO CATALOGUES). Indeed, the history of parallax observations has been as much a quest for increasing accuracy as for increasing numbers of stars measured. Moreover, the use of parallax data is fraught with difficulties caused by the small size of most stellar parallaxes.

Parallaxes are also vital to the measurement and calibration of stellar luminosities, because a star's distance needs to be known to convert its apparent magnitude into an absolute magnitude. Likewise, parallaxes are useful in the study of stellar kinematics, because the distance is needed to convert a star's PROPER MOTION into its velocity in the plane of the sky (transverse velocity), and in the determination of stellar masses, where the distance converts a visual binary's apparent orbit into the linear scale needed to apply Kepler's third law.

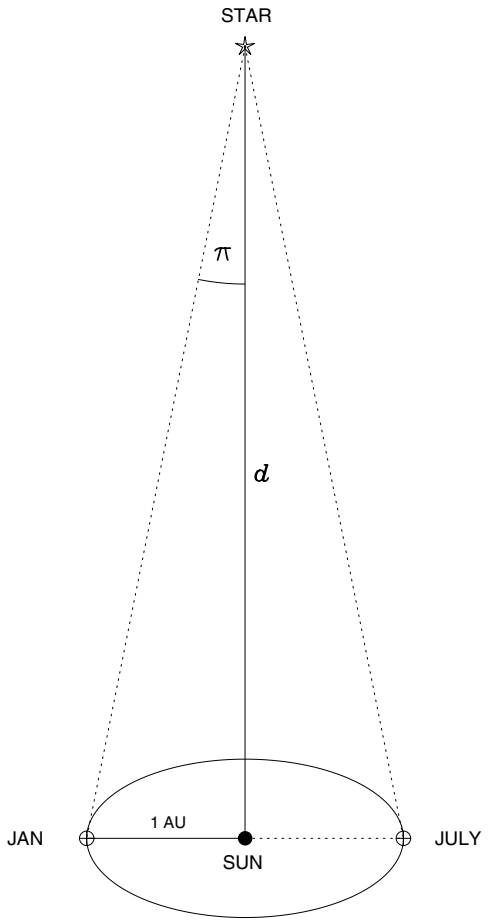


Figure 1. Parallax (π) of a star at distance (d). In this example, the star is observed at two opposite points (January and July) on the Earth's orbit, which is assumed circular with radius 1 AU, but appears as an ellipse when viewed at an angle. The distance d is not drawn to scale. Even the nearest stars are several hundred thousand AU from the Sun.

Historical background

Copernicus' 16th-century heliocentric theory of the solar system held two important implications for astrometry. First, if the stars were at finite distances, they should show annual parallactic motions. Detection of stellar parallaxes would directly confirm the COPERNICAN SYSTEM. Second, the distances to the stars themselves could be determined by measuring their parallaxes. However, more than two centuries of observational efforts, with gradually increasing accuracy, were frustrated because parallaxes remained too small to detect. Larger effects, such as proper motions (up to 10 arcsec yr^{-1}), annual aberration (20.5 arcsec), and nutation (9 arcsec) were discovered in the first half of the 18th century, but it was not until 1838 that Bessel finally measured the parallax (0.3 arcsec) of the 5th-magnitude high-proper-motion star 61 Cygni with the Königsberg heliometer. Contemporaneously, parallaxes of two bright stars, Vega and α Centauri, were measured by Struve and Henderson, with different instruments. The rest of the 19th century saw much work but little progress (see ASTROMETRY: HISTORY). The errors of visual observations were too large to measure more than a few parallaxes with any real accuracy. Parallaxes for the same stars measured by different observers and instruments often disagreed badly. The introduction of photography failed to improve matters significantly. By the end of the century, parallaxes were known for fewer than 60 stars, often with a precision not much better than 0.1 arcsec.

In the first two decades of the 20th century, SCHLESINGER revolutionized astrometry and opened the era of modern parallax determinations. Starting with the Yerkes 1 m refractor in 1903, Schlesinger perfected methods for accurate photographic observation with long-focus REFRACTING TELESCOPES. He systematized the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observing, measuring and reduction methods to eliminate or minimize errors. By 1920, Schlesinger's techniques improved parallax precision to the 0.01 arcsec level, making it possible to measure distances for thousands of stars. This led to a world-wide program, with five major observatories (Allegheny, Cape, Greenwich, McCormick and Yale) working for nearly a half-century on an all-sky survey, complete to visual magnitude 5.5, supplemented by high-proper-motion stars as faint as 14th magnitude. Starting in 1924, Schlesinger and his successors at Yale compiled these observations into a series of catalogues. By 1960 the classic program had nearly ended; the 1963 edition of the Yale General Catalogue of Trigonometric Stellar Parallaxes (GCTSP) lists 6390 stars.

Modern observations

By the 1960s it was widely recognized that future progress in parallax observations would require another leap in accuracy, which modern technology could now provide. Vasilevskis at Lick Observatory pioneered sophisticated observational techniques, automatic plate measurements and computerized data reductions. Strand obtained a 1.5 m astrometric reflector for the US Naval Observatory. By the 1970s, parallaxes for faint, low-luminosity stars could be determined with a precision of several mas. The 1980s saw the replacement of photographic plates by electronic detectors (CCDs at USNO, the Multichannel Astrometric Photometer (MAP) at Allegheny), further reducing errors to the 1 mas level. The 1995 edition of the Yale GCTSP now lists 8112 stars. While this is only a small increase in number since 1963, the added stars represent nearly an order of magnitude increase in accuracy.

In the 1990s, astrometry was revolutionized again by the European Space Agency's HIPPARCOS satellite, which measured parallaxes and proper motions, as accurate as the best ground-based data, for 118 000 stars. Scanning the sky for just over 3 yr, Hipparcos observed 15 times as many stars as ground-based parallax observatories had accumulated over the past century. Working from a predetermined Input Catalogue, Hipparcos observed a two-part program: (1) 52 000 bright stars ($V < 7$–9, depending on galactic latitude), constituting a virtually complete magnitude-limited survey; (2) 66 000 fainter stars ($V \lesssim 12.4$) selected for astrophysical or astrometric interest. Additionally, the Tycho experiment used the satellite's star mapper to measure parallaxes of modest precision (7 mas for $V < 9$; 25 mas for $V < 11.5$) for 1.06 million stars, complete to $V \simeq 10.5$.

So, by the close of the 20th century, space astrometry has almost entirely replaced ground-based observations as the source of parallax data for bright stars. However, it is worth noting that ground-based telescopes remain the only source of parallaxes for faint stars.

Applications and critical analysis

The utility of trigonometric parallaxes to calculate stellar distances, kinematics, masses or luminosities depends critically on the size of the parallax in relation to the observational error σ_π. Parallax errors propagate into stellar distances and transverse velocities in proportion to the ratio σ_π/π, i.e.

$$\sigma_d/d = \sigma_v/v = \sigma_\pi/\pi$$

and into visual binary star masses as

$$\sigma_{\mathcal{M}}/\mathcal{M} = 3\sigma_\pi/\pi.$$

When the absolute magnitude is calculated from the apparent magnitude and parallax (in arcsec)

$$M = m + 5 + 5\log_{10}\pi$$

the parallax error propagates as

$$\sigma_M = 5\sigma(\log_{10}\pi) \simeq 2.17\sigma_\pi/\pi.$$

Because of this, only 'large' parallaxes ($\pi \gg \sigma_\pi$) are individually useful on a star-by-star basis. However, most parallaxes are 'small', i.e. $\pi \simeq \sigma_\pi$, and can only be used statistically. In either case, care must be taken to ensure that (1) the parallaxes are on an accurate absolute zero point, (2) the observational errors are accurately estimated and (3) statistical issues are properly addressed.

Absolute zero-point calibration

Ground-based parallax observations are made by repeatedly pointing the telescope at the particular field (generally $<1^\circ$ in size) containing the 'target' star(s) and a number (usually 4–10) of background 'reference' stars. The parallactic ellipse of the target star is measured relative to the reference stars, whose own motions are in phase with the target star, only (presumably) smaller. The result is a 'relative' parallax, which must be corrected to 'absolute' by adding the mean parallax of the reference stars. This correction must be determined independently. This can be done statistically, by using Galactic models to estimate the reference frame's parallax as a function of the stars' apparent magnitudes and the Galactic latitude and longitude of the field. Reference star parallaxes can also be estimated individually from photometry and spectroscopy. For the 1995 Yale GCTSP, van Altena re-calculated the corrections to absolute for all parallaxes published to that date. The size of the corrections is generally 1–5 mas, with an uncertainty of ~10%, i.e. a few tenths of 1 mas.

The Hipparcos satellite simultaneously imaged two $0.9^\circ \times 0.9^\circ$ fields of view, separated by 58° on the sky. Because the parallactic motions in the two widely separated fields were out of phase, absolute parallaxes could be determined directly from the Hipparcos observations themselves. Global tests confirm that the overall absolute zero point of the Hipparcos Catalogue is accurate to 0.1 mas. However, this does not represent the zero-point accuracy in any individual field, where small-angular-scale zero-point errors may reach the order of 1 mas. (See the discussion of cluster parallaxes below.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Observational errors

Any parallax determination will have a finite accidental error $\pi_{\rm obs} - \pi_{\rm true}$, the size of which, σ_π, is estimated statistically from the dispersion of the positional measurements. It is usually assumed that the observing process $C(\pi_{\rm obs} \mid \pi_{\rm true})$ is represented by a normal distribution with mean $\pi_{\rm true}$ and variance σ_π^2. Great effort has been taken by the compilers of the Yale GCTSP and the Hipparcos Catalogue to detect and eliminate systematic errors and to assess the real precision of the parallaxes. This allows the user to rely on the π, σ_π values listed in these catalogues to be the best statistical estimates of the true quantities. For the 1995 GCTSP, van Altena comprehensively assessed the 'external' errors of each observatory, as a function of the published 'internal' errors. Before the 1997 publication of the Hipparcos Catalogue, the Hipparcos team verified that the errors were not underestimated by more than 10%.

Parallax errors can be correlated in two ways that may be important. First, the unknowns in a star's parallax solution (position, proper motion, parallax) will be correlated unless the observations are well distributed in time. This is simple to achieve with ground-based observations. However, for certain areas of the sky, Hipparcos was not able to provide equal numbers of observations on both sides of the parallactic ellipse, causing large (up to ± 0.6) correlations between parallax and ecliptic longitude. The Hipparcos Catalogue lists for each star the correlations between all the observational variables; the data should be used with caution when these values are not small ($\lesssim 0.2$).

Second, the parallaxes of two or more stars in one telescope's field of view will be correlated (i.e. not statistically independent) because they share the same reference frame. This is rarely a problem for ground-based observations, which generally have one 'target' star per field. Hipparcos presents a very different case. The satellite scanned great circles on the sky, timing each star's passage through the 0.9° field of view as a one-dimensional 'abscissa' relative to the positional zero point adopted for each 'reference great circle'. All stars on any one circle share the same zero-point error. As the scans repeated with different orientations, stars separated by $\gg 1^\circ$ were measured on different sets of scans, and their parallaxes are statistically independent. However, stars close together on the sky ($\lesssim 1^\circ$) share many scans in common, regardless of their orientation. As the same abscissa errors propagate into the parallax errors, the individual parallaxes are no longer statistically independent. The corresponding correlations, approaching $\rho = +0.6$ for small separations, must be carefully taken into account whenever Hipparcos data for small areas of the sky, such as in a cluster, are used.

'Large' parallaxes: Lutz–Kelker corrections

Perhaps the most important use of trigonometric parallaxes is the calibration of luminosities of different types of nearby stars. For 'large' parallaxes ($\pi \gg \sigma_\pi$), individual absolute magnitudes are calculated from

$$M = m + 5 + 5\log_{10}\pi.$$

However, this requires selecting stars by the size of their observed parallaxes, which introduces a statistical bias caused by the distribution of stars in space. This is because the conditional distribution of true parallax given the observed value, $C(\pi_{\rm true}|\pi_{\rm obs})$ is not the same as the (unbiased) conditional $C(\pi_{\rm obs}|\pi_{\rm true})$ representing the parallax measuring process. Rather,

$$C(\pi_{\rm true}|\pi_{\rm obs}) \sim C(\pi_{\rm obs}|\pi_{\rm true}) \times N_d(\pi_{\rm true})$$

where $N_d(\pi_{\rm true})$ is the spatial distribution of the stars in question. For nearby stars distributed uniformly in space, $N_d(\pi_{\rm true}) \sim \pi^{-4}$. The sense of the bias is that $\pi_{\rm true}$ is most likely less than $\pi_{\rm obs}$, because of the larger volume of space corresponding to smaller parallax.

The impact of this volume bias on luminosity calibrations was not fully understood until 1973, when Lutz and Kelker showed how to calculate absolute magnitude corrections

$$\Delta M = \langle M_{\rm true} - M_{\rm obs}\rangle = \langle 5\log_{10}(\pi_{\rm true}/\pi_{\rm obs})\rangle$$

with the averages taken over the conditional distribution $C(\pi_{\rm true}|\pi_{\rm obs})$. These 'Lutz–Kelker corrections' reach $\Delta M = -0.4$ mag for the limiting value $\sigma_\pi/\pi = 0.175$.

A serious objection to Lutz and Kelker's simple formulation is that, in practice, stars selected for parallax observation are not uniformly distributed in space. Common observational criteria (apparent magnitude, proper motion) purposely select nearby stars in preference to more distant stars. So, the space distribution of parallax stars is generally much less steep than π^{-4}, and the appropriate Lutz–Kelker corrections are correspondingly smaller. In 1979, Hanson showed how to use the power-law distribution of the proper motions of the sample of stars in question to estimate the exponent n in the generalized distribution $N_{\rm d}(\pi) \sim \pi^{-n}$ and how to calculate absolute magnitude corrections as a function of n and σ_π/π. For a typical case, $n = 2$ and $\sigma_\pi/\pi = 0.15$, $\Delta M = -0.1$ mag.

Even with the large number of mas-precision parallaxes provided by Hipparcos, many stellar luminosity calibration problems necessitate working with data of modest relative precision ($0.1 \lesssim \sigma_\pi/\pi \lesssim 0.2$) in order to obtain a large enough sample size. For this reason, Lutz–Kelker corrections remain a vital step in the careful statistical treatment of the data.

An outstanding example is the use of Hipparcos parallaxes of nearby field subdwarfs to calibrate the metal-poor main sequence which is fitted to Galactic GLOBULAR CLUSTERS to derive their distances and ages. Most studies have restricted $\sigma_\pi/\pi < 0.15$ to keep the Lutz–Kelker corrections small, but this leaves too few (~10–30) subdwarfs to calibrate M_V versus $B - V$ empirically as a function of the metallicity [Fe/H], and almost

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

no subdwarfs comparable with the most metal-poor ([Fe/H] < -2) clusters. Hipparcos actually measured parallaxes for several hundred metal-poor subdwarfs. The calibration sample could be increased substantially by extending the precision limit to $\sigma_\pi/\pi \simeq 0.2$ or greater, and applying appropriate Lutz–Kelker corrections.

'Small' parallaxes: reduced parallaxes

Most classes of high-luminosity distance indicators (e.g. RR Lyraes, Cepheids) are sufficiently rare in space that even the nearest examples are too distant for their parallaxes to be individually useful. For example, RR Lyrae itself is the only variable of that type in the Hipparcos Catalogue with $\sigma_\pi/\pi < 0.15$. Over 200 Cepheid variables were observed by Hipparcos; most are so distant that $\sigma_\pi/\pi \simeq 1$. Many of the parallaxes are negative. Although useless individually, 'small' parallaxes for large numbers of stars do contain valuable distance information, which can be extracted statistically, as long as the absolute zero point is reliable and the parallaxes are independent.

The proper tool for this work is the classic method of reduced parallaxes, first introduced by Charlier as early as 1916, and improved by Malmquist in the 1930s and again by Roman in 1952. The 'reduced parallax' P is defined by

$$P = \pi \times 10^{0.2(m+5)} = 10^{0.2M}$$

which is just a rearrangement of the usual formula relating parallax and absolute magnitude. The procedure is simple. For a group of stars presumed to have the same absolute magnitude M, the parallaxes are 'reduced' to a common apparent magnitude, thus scaling each parallax to a common distance. The reduced parallaxes P can then be averaged (with appropriate weights), and then

$$M = 5\log_{10}\langle P\rangle.$$

With the advent of the Hipparcos parallaxes, the classic method of reduced parallaxes has been re-discovered and applied with considerable success. Most notable is the work of Feast's group on the CEPHEID PERIOD–LUMINOSITY RELATION. The work of Arenou and Luri shows how reduced parallaxes can be used individually to make an 'astrometric H–R diagram' and clearly outlines the statistical advantages of the method. Respecting the long history of reduced parallaxes, however, one does regret their unfortunate coinage 'astrometry-based luminosity' in place of the classic terminology.

Cluster parallaxes: small-scale errors

For a set of stars at virtually the same distance, such as in a GALACTIC OPEN CLUSTER, the parallaxes can be simply averaged to obtain the cluster's parallax, which will be accurate if the absolute zero point is reliable and the individual parallaxes are uncorrelated. With ground-based parallaxes, only a few nearby clusters, most notably the Hyades, have enough data to provide useful results. (The parallaxes of several Hyades members have also been measured with the Hubble Space Telescope.) Without parallaxes, main-sequence (MS) fitting has been the only way to estimate distances for most open clusters.

Hipparcos observed stars in more than 200 open clusters, opening the possibility to determine their distances directly, without assumptions about stellar structure or chemical composition. Several teams have analyzed Hipparcos' cluster parallaxes. The Hyades distance was determined by Perryman *et al* to better than 1% precision ($d = 46.3 \pm 0.3$ pc) using proper motions as well as parallaxes. Robichon *et al* studied 18 other clusters with $d < 500$ pc and eight or more members observed by Hipparcos. van Leeuwen independently analyzed the nearest nine of these. Because the angular size of the clusters ($\sim 1^\circ$, except for the Hyades) is roughly the size of Hipparcos' field of view, and the star density in these clusters (9–17 stars deg^{-1}) is so high, one must account carefully for the correlations between measurements on the same great circles. Both Robichon and van Leeuwen did this using van Leeuwen's 'combined abscissae method', which uses the Hipparcos intermediate data to solve for a single parallax and proper motion for each cluster. These cluster parallaxes were claimed to be accurate to $\pm\sim$0.2–0.4 mas.

Controversy arose in 1997–9 when the Hipparcos parallax analyses gave anomalous distances (compared with MS fitting) for several clusters (e.g. Pleiades, Coma Berenices). Moreover, there were unexplained discrepancies between the two groups' results. Robichon concluded that only the Pleiades is anomalous, and the other clusters are consistent with MS fitting when their metallicities are taken into account. To the contrary, van Leeuwen claimed that the Pleiades and five other young clusters are subluminous, without regard to their metallicities, in stark contrast to modern understanding of stellar structure. Furthermore, the Hipparcos Pleiades parallax derived by both groups (8.5 ± 0.2 mas) was nearly 1 mas larger than the values determined from radial-velocity gradients by Narayanan and Gould, and from Allegheny MAP parallaxes by Gatewood.

The solution lay in recognizing the size of the small-angular-scale zero-point errors in Hipparcos' parallaxes. Regardless of what method of analysis is used, if the local zero point at the position of the cluster differs from absolute zero, the cluster parallax will be wrong by that amount. Hipparcos' zero point had been tested globally before the Catalogue was published, but no tests on small ($\sim 1^\circ$) angular scales were performed—such tests would require the cluster parallaxes themselves. van Leeuwen asserted that the local errors were only $\sim$0.1 mas, as small as the global errors, but the Pleiades parallaxes cast doubt on this.

It is important to note that the Pleiades has the highest central density (17 stars deg^{-1}) of any cluster observed by Hipparcos, and thus its parallax may be most vulnerable to small-scale systematic errors. Indeed, Pinsonneault *et al* discovered that the stars in the central 1° of the Pleiades have $\pi \simeq 9$ mas, but the outer stars have $\langle\pi\rangle = 7.5$ mas. Most of the central Pleiads have high correlations ($\rho >$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

+0.34) between parallax and right ascension, identifying these stars as sharing many scans in common, and having little statistical independence. The central parallax is dominated by the local error, which does not 'average out' as it could over a larger area, such as the outer Pleiades or the Hyades.

Narayanan and Gould found correlated parallax errors with amplitude up to 2 mas on scales up to 2° for both the Pleiades and the Hyades regions. The RMS size of the zero-point fluctuations is 0.5–0.6 mas. Pinsonneault *et al* suggested that the RMS size of Hipparcos' small-scale zero-point errors over the whole sky could be up to 0.4 mas. Perryman pointed out that the parallax differences in the separate Hipparcos data reductions by the FAST and NDAC consortia reach 2 mas or more averaged over $2^\circ \times 2^\circ$ fields near the ecliptic. Merging the FAST–NDAC results for the final Catalogue reduces the local systematics by a factor of 2, to $\sim$1 mas. The Pleiades and Hyades results are fully consistent with the expected uncertainties of the Hipparcos Catalogue.

When the small-angular-scale errors in the Hipparcos parallaxes are taken into account, the cluster parallaxes are quite consistent with MS-fitting distances. In hindsight, it is now apparent that Hipparcos, for all its precision, was not the ideal instrument to measure cluster distances. The accuracy of a cluster's parallax is effectively limited to the level of the small-scale zero-point errors, no matter how many stars per cluster are observed. Future parallax surveys from space, based on Hipparcos' scanning principle, may also be affected by small-scale errors, and might well use cluster parallaxes to test them.

Future prospects

The future of parallaxes lies largely in space. Ground-based observations will continue to have importance for low-luminosity stars, binaries and clusters. Most late M dwarfs, white dwarfs and brown dwarf candidates are too faint to be observed even by the ambitious satellite programs mentioned below. Binaries generally need to be followed for longer periods than the limited lifespans of satellites. Hipparcos' experience points out the need for ground-based cluster parallaxes. At present, several observatories (Allegheny and US Naval in the north, Cerro Tololo and Mt Stromlo in the south) have active parallax programs.

The success of Hipparcos in producing large numbers of precise parallaxes for bright stars has proved the potential of space astrometry and spawned several projects to reach well beyond the magnitude and precision limits of Hipparcos.

The US National Aeronautics and Space Administration (NASA) has selected the US Naval Observatory's Full-sky Astrometric Mapping Explorer (FAME) for a 2.5–5 yr mission, scheduled for launch in 2004. With an operating principle much like Hipparcos (input catalogue, two fields of view, great circle scanning) but with an advanced 24-CCD-array detector, FAME will measure astrometry (positions, proper motions, parallaxes) and four-color photometry for nearly all stars to $V = 15$, a total of 4×10^7 stars. Projected astrometric accuracy ranges from 0.05 mas at $V = 9$ to 0.50 mas at $V = 15$. Photometric accuracy will be at the milli-magnitude level.

Several other space astrometry projects are in various stages of planning. DIVA (Double Interferometer for Visual Astrometry), proposed for launch in 2003, is a small Hipparcos-like scanning satellite which will provide astrometry and broad-band photometry to $V = 15$, and spectrophotometry to $V = 13.5$. Parallax accuracy at $V = 9$ will be 0.20 mas.

GAIA is a satellite observatory proposed for the European Space Agency's Horizon 2000+ plan, with the goal to build a catalogue of positions, distances, and three-dimensional space velocities for 10^9 stars to $V \simeq 20$. Parallax accuracy will be 0.004 mas for $V = 10$, 0.14 mas for $V = 15$ and 0.30 mas for $V = 20$. Photometry and radial velocities will be measured by separate instruments on the satellite.

NASA's SPACE INTERFEROMETRY MISSION (SIM), proposed for launch in 2005, is a 10 m baseline optical Michelson interferometer, making pointed observations from an Earth-trailing solar orbit. Astrometric precision will be 0.001 mas in relative measurements over small angles ($\leq 1^\circ$). Absolute parallax accuracy will be 0.004 mas.

Microarcsecond (μas) astrometry from space will revolutionize distance measurements once again. To give just one example, the parallax of the Galactic center is 120 μas. At that distance, an RR Lyrae star of $M_V = +0.5$ could be measured to 10% precision in parallax. Provided that systematic errors can also be reduced to the μas level, the prospects for 21st-century Galactic astronomy seem virtually unlimited.

Bibliography

Binney J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)

ESA 1997 *The Hipparcos and Tycho Catalogues* ESA SP-1200 (Noordwijk: European Space Agency)

Kovalevsky J 1998 First results from Hipparcos *Ann. Rev. Astron. Astrophys.* **36** 99–129

Reid I N 1999 The HR diagram and the galactic distance scale after Hipparcos *Ann. Rev. Astron. Astrophys.* **37** 191–237

van Altena W F, Lee, J T and Hoffleit D 1995 *The General Catalogue of Trigonometric Stellar Parallaxes* 4th edn (Schenectady, NY: Davis)

Vasilevskis S 1966 The accuracy of trigonometric parallaxes of stars *Ann. Rev. Astron. Astrophys.* **45** 57–76

Robert B Hanson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parker, Eugene ['Gene'] Newman (1927–)

American solar-terrestrial physicist, Parker worked at the University of Chicago, studying the solar wind and the effects of magnetic fields in the heliosphere. At first this work was through the study of the effect of the Sun on comet tails, but more recently through satellite measurements *in situ* he has developed understanding of the solar corona, the solar wind, the magnetic fields of both Earth and Sun, and their interactions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parkes Observatory

Arguably one of the world's great and most productive research instruments, the 64 m fully steerable paraboloid, Parkes Radio Telescope, is located in rural New South Wales, 400 km west of Sydney, Australia.

Commissioned in 1961 to further the pioneering research of the CSIRO Division of Radiophysics, continual upgrades to its surface, receivers and correlators have kept it at the forefront of radio astronomy.

Operated as part of the AUSTRALIA TELESCOPE NATIONAL FACILITY, much of its time is allocated to international pulsar, galaxy and Very Long Baseline Interferometry (VLBI) programs, with occasional support for JPL planetary missions. The H I Parkes All-Sky Survey for galaxies is a current major project.

For further information see
http://www.parkes.atnf.csiro.au.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parsec (pc)

A unit of distance commonly used by astronomers. It is the distance at which a star would have an annual parallax of exactly one second of arc; from this distance, the Earth's orbit would have an apparent radius of one second of arc. One parsec is equivalent to 3.26 light-years (3.09×10^{13} km) or 206 265 astronomical units.

The distance of a star expressed in parsecs = 1/annual parallax (in seconds of arc); thus a star whose annual parallax is 0.1 seconds of arc lies at a distance of 1/0.1 = 10 parsecs. For example, the nearest triple-star system, Alpha Centauri, has a parallax of 0.753 seconds of arc; hence, its distance from the Sun and the Earth is 1.33 parsec.

Multiples of this unit are kiloparsecs (kpc) = 1000 parsecs and megaparsec (Mpc) = 1 000 000 parsecs. In our Galaxy distances to remote stars are measured in terms of kiloparsecs. The Sun is at a distance of 8.5 kiloparsecs from the center of the Galaxy. When dealing with other galaxies or clusters of galaxies, the convenient unit is the megaparsec. The distance to the Andromeda Galaxy (M31) is about 0.7 megaparsecs. Remoter galaxies and quasars have distances in the order of 3000 megaparsecs, or 9 000 000 000 to 10 000 000 000 light-years.

See also: parallaxes, second of arc.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Parsons, Charles (1854–1931)

Youngest son of the third Earl of Rosse, brother to the fourth. Inventor of the steam turbine. Interested in telescope building through his father's activities with the Birr telescope, took over the machine engineering business of THOMAS and HOWARD GRUBB in 1925, and formed Sir Howard Grubb-Parsons Ltd in Newcastle, which made large telescopes starting with the 74 in reflector for the David Dunlap Observatory in Toronto. The firm took over Cooke, TROUGHTON and Simms Ltd. The firm was wound up after completing the 4.2 m William Herschel Telescope on La Palma in 1985.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Particle Acceleration in Magnetic Fields

Tenuous PLASMAS in astrophysical settings often contain particle populations which are not in thermodynamic equilibrium. Such non-thermal populations frequently possess power-law distributions. They are either observed directly or inferred from radiative properties. In low-density astrophysical plasmas the particles do not scatter against other particles, but scatter collisionlessly against magnetic field turbulence (see also PARTICLE ASTROPHYSICS; ASTROPHYSICAL BOUNDS ON PARTICLE PROPERTIES). Thus the particles are forced to equilibrate with the collective body of particles held together by the nearly frozen-in magnetic field. When the scattering centers, i.e. magnetic field turbulence, are moving toward each other, charged particles trapped in between are accelerated like tennis balls bouncing off two approaching walls. This is the basic concept of magnetic field acceleration first proposed by FERMI in 1949. Two very distinct cases can be separated: in the first case scattering is due to the magnetic field turbulence upstream and downstream of fast magnetosonic shocks, i.e. shocks which have an upstream plasma velocity larger than the fast magnetosonic velocity. The process of acceleration of particles by scattering back and forth across a shock in the presence of magnetic turbulence in the background plasma is called first-order Fermi acceleration or the diffusive acceleration mechanism. Since the phase velocity of the waves is small in comparison with the upstream flow speed the scattering centers move essentially with the flow, and a particle scattered back and forth across the shock senses the difference in the upstream and downstream flow speeds. In the second case the scattering centers are low-frequency plasma waves traveling in the plasma frame in opposite directions. In the plasma frame, a particle makes either a head-on or an overtaking collision with a scattering center, leading to either energy gain or energy loss respectively. Due to the large relative velocity between particle and wave the frequency of head-on collisions is greater than the frequency of trailing collisions, which leads to a net energy gain. This is the second-order Fermi acceleration process. Other acceleration mechanisms relay on the possibility of strong magnetic field-aligned electrostatic potential drops.

Diffusive shock acceleration

The transport of energetic particles in fluctuating magnetic fields is effectively described by the equation

$$\frac{\partial f}{\partial t} = \nabla\cdot(\bar{\kappa}\cdot\nabla f) - (\boldsymbol{V}+\boldsymbol{V}_D)\cdot\nabla f + \nabla\cdot\boldsymbol{V}\frac{p}{3}\frac{\partial f}{\partial p} + \frac{1}{p^2}\frac{\partial}{\partial p}\left(p^2\langle D_{pp}\rangle\frac{\partial f}{\partial p}\right) + Q \qquad (1)$$

where $f(\boldsymbol{r}, p, t)$ is the particle omnidirectional distribution function, $p = m\gamma v$ is momentum magnitude, $\boldsymbol{V}(\boldsymbol{r}, t)$ is the bulk velocity of the background plasma and $Q(\boldsymbol{r}, p, t)$ is a source term (to describe for example ion injection at a shock). The control of the magnetic field on the spatial transport and transport in momentum space is described by the symmetric part of the spatial diffusion tensor, $\bar{\kappa}$, the drift velocity, $\boldsymbol{V}_D$, and the diffusion coefficient in momentum space, $\langle D_{pp}\rangle$. Equation (1) is based on the assumption that $v \gg V$ and the spatial scale length, L, is sufficiently larger than the scattering mean-free path. The first term on the right-hand side describes spatial diffusion of the particles. The second term describes the effect of spatial convection and energetic particle drift, and the third term is usually called the adiabatic deceleration term. This term is a momentum convection term and describes adiabatic deceleration as well as first-order Fermi acceleration in the case of converging bulk flow as, for instance, at a shock where $\nabla\cdot\boldsymbol{V} > 0$. The drift term $\boldsymbol{V}_D\cdot\nabla f$ describes transport due to curvature, gradient and magnetization drifts in the 'average' (usually viewed as an ensemble average) magnetic field. The diffusion tensor is often assumed to be axisymmetric about the unit vector $\boldsymbol{b} = \boldsymbol{B}/B$ and therefore has the form $\bar{\kappa} = \kappa_\perp\delta_{ij} + (\kappa_\parallel - \kappa_\perp)b_i b_j$ where δ_{ij} is the Kronecker symbol. $\kappa_\parallel$ ($\kappa_\perp$) describes spatial diffusion parallel (perpendicular) to $\boldsymbol{b}$ due to magnetic field fluctuations. The parallel diffusion coefficient $\kappa_\parallel$ and the diffusion coefficient in momentum space $\langle D_{pp}\rangle$ are determined by pitch angle averages

$$\kappa_\parallel = \frac{v^2}{8}\int_{-1}^{1}\mathrm{d}\mu\frac{(1-\mu^2)}{D_{\mu\mu}} \qquad (2)$$

and

$$\langle D_{pp}\rangle = \frac{1}{2}\int_{-1}^{1}\mathrm{d}\mu\left(D_{pp} - \frac{D_{\mu p}^2}{D_{\mu\mu}}\right) \qquad (3)$$

where $D_{\mu\mu}$, D_{pp}, and $D_{\mu p}$ are the three Fokker–Planck coefficients, and μ is the cosine of the pitch angle α, $\mu = \cos\alpha = v_\parallel/v$ ($v_\parallel$ = velocity parallel to the magnetic field).

Associated with the solution to the distribution function f is a streaming flux $\boldsymbol{S}$

$$\boldsymbol{S} = -\kappa_\parallel\nabla f - \frac{1}{3}\boldsymbol{V}p\frac{\partial f}{\partial p} \qquad (4)$$

and an associated anisotropy magnitude $\delta = 3|\boldsymbol{S}|/(vf)$. In the case of a planar shock with the shock at $x = 0$ the steady-state solution to equation (1) can be obtained by solving it in the upstream and downstream regions and relating the solutions across the shock by the jump condition that the streaming flux $\boldsymbol{S}$ is continuous. With x in the direction of the shock normal the jump condition reads

$$\left[\kappa_{xx}\frac{\partial f}{\partial x} + \frac{v_x}{3}\frac{\partial f}{\partial \ln p}\right]_1^2 = \tilde{Q}. \qquad (5)$$

The sub- and superscripts 1 and 2 denote upstream and downstream states, $\tilde{Q}$ is that part of the source which is concentrated at the shock and κ_{xx}, the diffusion coefficient normal to the shock surface, is given by the angle Θ

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

between the magnetic field and the x direction as $\kappa_{xx} = \kappa_\parallel \cos^2\Theta + \kappa_\perp \sin^2\Theta$. The spectrum produced at a planar shock is a power-law spectrum in momentum, determined only by the shock compression ratio, $R = V_1/V_2 = \rho_2/\rho_1$, where ρ is the density:

$$f(p) \propto p^{-3R/(R-1)}. \tag{6}$$

This spectrum falls off upstream with e-folding distance $L = \kappa_{xx}/V_1$. The important quantity that determines the time scale for acceleration is the diffusion coefficient in the normal direction. In the case of quasi-parallel shocks this is the diffusion coefficient parallel to the magnetic field, whereas in the case of quasi-perpendicular shocks cross-field diffusion eventually determines the normal diffusion coefficient. From a time-dependent solution to equation (1) it can be shown that the steady state is reached in a characteristic time $\tau \approx 4\kappa_{xx}/V_1^2$. Other linear modifications arise due to the fact that shocks are not necessarily planar, i.e. the radius of curvature of the shock front is no longer small compared with the diffusion length scale $L = \kappa_{xx}/V_1$. This is the case for blast waves from a point explosion or for shocks terminating stellar winds.

The convection–diffusion equation (1) can be generalized to include, in addition to the term that describes energy changes due to the divergence of the flow, a term that describes energy changes due to a shear in the flow. In the case when the plasma velocity $\boldsymbol{V}$ varies over spatial scales that are large compared with the scattering mean free path, this term has the form

$$\frac{\Gamma}{p^2}\frac{\partial}{\partial p}\left(\tau p^4 \frac{\partial f}{\partial p}\right) \tag{7}$$

with Γ the viscous momentum transfer and τ the scattering time. Γ is determined by the velocity shear $\partial V_i/\partial V_j$. For a simple shear flow $\boldsymbol{V} = V_y(x)\boldsymbol{e}_y$ the coefficient of momentum transfer reduces to $\Gamma = (\partial V_y/\partial x)^2/15$. The transfer of energy to the energetic particles leads in turn to a viscosity of the ambient plasma.

The spatial diffusion coefficients

The diffusion–convection equation (1) is derived by pitch-angle averaging of a Fokker–Planck equation. The Fokker–Planck equation is obtained by applying the quasilinear approximation to the behavior of energetic particles in a uniform magnetic field with superposed small-amplitude plasma turbulence. The effect of the turbulence is determined by calculating first-order corrections to the particle's orbit in the uniform magnetic field and ensemble averaging over the statistical properties of the plasma turbulence. The ensemble-averaged first-order particle correction to the equation of motion in momentum and pitch angle space determines the three Fokker–Planck coefficients D_{pp}, $D_{\mu p}$ and $D_{\mu\mu}$. The particles interact resonantly with the magnetic field turbulence, i.e. a particle is scattered only if the Doppler shifted wave frequency equals the particle's gyrofrequency Ω times an integer n:

$$\omega - k_\parallel v_\parallel = n\Omega \tag{8}$$

where $v_\parallel = v\mu$ and $k_\parallel$ are the parallel velocity component and the parallel wavenumber respectively. The gyrofrequency is given by $\Omega = qB/mc\gamma$, where q = charge, m = mass and γ is the relativistic factor of the particle. The resulting Fokker–Planck coefficients depend strongly on the type of turbulence present. Most analytical investigations have been done for the so-called slab model, where the plasma turbulence is assumed to consist of waves propagating parallel to the magnetic field. Furthermore, these waves are usually assumed to be Alfvén waves. The dispersion relation then reads $\omega = k_\parallel V_A$, with V_A the Alfvén speed. In this case quasi-linear theory yields the pitch-angle diffusion coefficient

$$D_{\mu\mu} = \frac{\pi q^2}{2m^2c^2|\mu|v} P\left(\frac{\Omega}{\mu v}\right). \tag{9}$$

$P(k)$ is the wave power at wavenumber k:

$$P(k) = \frac{1}{2\pi}\int \mathrm{d}s \langle \delta\boldsymbol{B}(s_0)\cdot\delta\boldsymbol{B}(s_0+s)\rangle \,\mathrm{e}^{-iks}. \tag{10}$$

If $P \propto k^{-\delta}$ this gives for the spatial diffusion coefficient in the non-relativistic regime an energy dependence $\kappa_\parallel \propto E^{(3-\delta)/2}$. If fast magnetosonic waves are present equation (9) has to be changed accordingly. Very low frequency fast magnetosonic waves have a dispersion relation given by $\omega = kV_A$. For oblique propagation the wave magnetic field has both a compressive (along $\boldsymbol{b}$) and a linearly polarized transverse component. Unlike the case of slab turbulence, the compressive component of oblique fast-mode waves allows energetic particles to resonate with the waves also through the $n = 0$ (Landau) resonance. However, this resonance is unimportant for the diffusion coefficient but, as will be shown below, is the dominant factor for the momentum diffusion coefficient. Summarizing, in order to determine the parallel spatial diffusion coefficient the following parameters have to be known: the total intensities in the Alfvén and fast-mode waves, the power-law shape, the distribution of the fast mode wave power in $\boldsymbol{k}$ space and the minimum cut-off frequency of the wave power. Depending on these parameters, the energy and mass/charge dependence of the parallel spatial diffusion coefficient may vary considerably.

Resonant scattering which causes pitch angle diffusion causes particles to shift the field line about which they gyrate and thus contributes to $\kappa_\perp$. In addition, a field line randomly walks about the average field $\boldsymbol{B}$, which causes perpendicular transport of a particle that follows the field line. There is no generally accepted theory for $\kappa_\perp$. For large-amplitude fluctuations it is often assumed that $\kappa_\perp \approx \kappa_\parallel$. Another model assumes that scattering of particles by the magnetic field turbulence occurs as in hard sphere scattering. With a parallel scattering time $\tau = 3\kappa_\parallel/v^2$, the cross-field diffusion coefficient is given in this case by

$$\kappa_\perp = \kappa_\parallel \frac{1}{1+\Omega^2\tau^2}. \tag{11}$$

The contribution to $\kappa_\perp$ from field line random walk is proportional to the power $P(0)$ of the magnetic field at zero wavenumber.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nonlinear modifications to shock acceleration

In general, the reaction of the shock-accelerated particles on the background plasma cannot be ignored. Two such kinds of reaction have to be considered. The streaming of the shock accelerated particles through the upstream plasma will excite waves through an ion/ion beam instability. This then leads to enhanced scattering with the particles themselves providing the scattering centers needed for diffusive shock acceleration. A more serious modification is concerned with the fact that for a power-law downstream distribution of energetic particles the pressure of these particles diverges for strong shocks and the reaction on the flow must be taken into account. Wave generation by the streaming of energetic particles can be dealt with in the framework of quasilinear theory. The diffusion–convection equation is supplemented by an equation for the wave energy density E_W. The time dependence $\partial E_W/\partial t$ is given by the convection of the waves with the flow and the wave growth rate γ_W, which is proportional to particle streaming. The spatial diffusion coefficient in the diffusion–convection equation according to quasilinear theory is proportional to E_W^{-1}. These two equations simultaneously determine f and E_W.

The problem of the reaction of the energetic particle pressure on the flow is essentially one in fluid dynamics. The background plasma is usually described by the Euler equations (conservation of mass flux, momentum flux and energy flux), and the reaction of the energetic particles is incorporated by an additional pressure term in the momentum equation due to energetic particles. An equation for the energetic particle pressure $P_c = (4\pi/3)\int \mathrm{d}pp^3vf(p)$ can be obtained by taking a suitable moment of the diffusion–convection equation with respect to momentum p. In practice, however, two supplemental assumptions are introduced instead. These are that the pressure and the energy flux of energetic particles are both continuous across the shock transition:

$$[P_c]_2^1 = [\gamma_c V P_c - \tilde{\kappa}\partial P_c/\partial x]_2^1 = 0 \qquad (12)$$

where γ_c is the ratio of specific heats for the energetic particles (which is 4/3 for relativistic, and 5/3 for non-relativistic particles). $\tilde{\kappa}$ is the suitable momentum averaged spatial diffusion coefficient. For low Mach number shocks the relativistic particle pressure has the effect of decreasing the effective specific heat ratio of the gas so that the shock compression ratio is not limited to four, but can increase up to seven. For stronger shocks the entire transition itself can be mediated by the presence of the energetic particles. Besides this two-fluid approach the mediation of shocks by energetic particles has been investigated through Monte Carlo simulations.

The momentum diffusion coefficient and second-order Fermi acceleration

The momentum diffusion coefficient can be computed from the Fokker–Planck coefficients according to equation (3). In the case of equal intensity of forward and backward traveling waves (vanishing cross helicity) equation (3) reduces to the much simpler equation

$$\langle D_{pp}\rangle = \int_0^1 D_{pp}\,\mathrm{d}\mu. \qquad (13)$$

In the case of slab-Alfvén waves D_{pp} has only a contribution D_{pp}^A from gyroresonance: the particles gain or lose speed when they are resonantly pitch-angle scattered. The gain or loss in speed is equal to the Alfvén speed. The Fokker–Planck coefficient D_{pp} is additive; if magnetosonic turbulence is present, additional terms D_{pp}^T and D_{pp}^G have to be added to D_{pp}^A. Here D_{pp}^G is the Fokker–Planck coefficient due to gyroresonant interaction (integer n in equation (8)), and D_{pp}^T is the Fokker–Planck coefficient for the $n = 0$ resonance, i.e. $\omega - k_\parallel v_\parallel = 0$. The latter condition means that interaction between the wave and the particle occurs when the particle is moving at about the parallel wave phase speed. This interaction is called transit-time damping and is due to the magnetic mirror force in the wave compressions. Introducing the parallel wavelength $\lambda_\parallel = 2\pi/k_\parallel$ and the wave period $T = 2\pi/\omega$, the resonance condition reads $\lambda_\parallel/v_\parallel = T$. This implies that a particle will interact with the wave when the particle transit time across the wave compression is about equal to the wave period. Transit-time damping can only occur for particles with a velocity exceeding the Alfvén velocity. In the case of a mixture of isotropically distributed fast magnetosonic turbulence and slab Alfvén waves of similar power, the contribution from the gyroresonances due to both Alfvén waves and magnetosonic waves to the Fokker–Planck coefficient D_{pp} is less than the contribution from the $n = 0$ resonance of the magnetosonic waves. If fast magnetosonic turbulence is present, energetic particle acceleration is mainly due to transit-time damping.

Other acceleration mechanisms

In addition to the above mechanisms of accelerating particles by magnetic waves and shocks, collisionless plasmas (see COLLISIONLESS PROCESSES IN ASTROPHYSICAL PLASMAS) are capable of dumping energy into small groups of particles by means of generating strong field-aligned electric potential drops, so-called electrostatic double layers. Large-scale potential drops can be produced by the action of the magnetic mirror force $\boldsymbol{F}_{m\parallel} = -M\nabla_\parallel B(s)$ on the particle distributions in a magnetic trapping configuration (e.g. the dipole magnetic field of a neutron star). Here $M = P_{c\perp}/NB$ is the magnetic moment of the particles of mass m, perpendicular pressure $P_{c\perp}$ and density N, and s is the coordinate along $\boldsymbol{b}$. Since ions and electrons mirror on the field line at different locations s, a field-aligned electric potential is produced that may accelerate particles.

Other important ways of producing strong parallel quasistatic electric fields require the presence of large-amplitude microturbulence which backscatters the electrons causing anomalous collisions, friction and ohmic electric fields. The energy for excitation of microturbulence is taken either from electric currents flowing along

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or perpendicular to the magnetic field, density or temperature gradients, or from electron beams. Once the current density $|\boldsymbol{j}| \geq j_c = eN|\boldsymbol{v}_D|$ exceeds a critical threshold j_c (here e is electron charge, N density and v_D the current-drift speed), high-frequency electric fluctuations near some typical plasma resonance frequencies are driven unstable until they grow to large amplitudes. For very strong currents $v_D > \sqrt{k_B T_e/2m_e}$ exceeding the electron thermal velocity, this frequency is the Buneman frequency $\omega_B = 0.03\omega_{pe}$, with ω_{pe} the plasma frequency. At weaker currents $v_D > c_{ia}$, with c_{ia} the ion sound speed, the ion acoustic wave $\omega = kc_{ia} < \omega_{pe}/43$ is excited. Electron beams generate Langmuir waves near $\omega = \omega_{pe}$ and Bernstein waves near $\omega = n\Omega_e$ ($n = 1, 2, \ldots$ and $\Omega_e = eB/m_e$ the electron cyclotron frequency). They also drive electron acoustic waves at frequencies between the ion and electron plasma frequencies. Gradients and transverse currents excite waves near the lower-hybrid frequency $\omega = \sqrt{\Omega_i \Omega_e} = 43\Omega_e$. In the region of such turbulence (extension L along $\boldsymbol{b}$) the total resulting potential drop can be estimated from

$$U = j_\parallel L \eta_{\rm an} \approx \frac{eN\nu_{\rm an}}{\omega_{pe}^2} L v_D. \tag{14}$$

Here $\nu_{\rm an}$ is the anomalous collision frequency. The maximum available energy to which an electron may be lifted is just $K_{\rm max} = eU$. For a monochromatic electron flux $F = N_{\rm acc} v_0$ passing the potential drop U at speed v_0 the number of electrons that can be accelerated is $\mathcal{N} = Fv_0/L$. This yields for the energy that can actually be reached the expression

$$K_{\rm acc} \approx \frac{eN\nu_{\rm an}}{\omega_{pe}^2} \frac{v_D L^2}{F v_0}. \tag{15}$$

When the amplitude of the microturbulence reaches large values, typically of the order of $E_W/Nk_BT \approx 10^{-4}$, the quasilinear effects causing anomalous resistivities are superceded by nonlinear effects. Particles become trapped in the wave potential and are accelerated and released in energetic bunches. These are effects well known from nonlinear processes encountered in chaos theory. In addition the interaction between the wave and the various groups of particles leads to the emergence of small-scale structures of the order of the Debye length, electron skin depth and particle gyroradii which trap the waves locally. Interaction of particles with these so-called solitons, solitary waves, cavitons, phase space holes or BGK modes causes microscopic stationary potential drops to develop across the solitary structures. Because in well developed turbulence myriads of such microscopic drops exist along and transverse to a field line their total potential drop may add up to very large values. Any small groups of particles falling across these potentials will then experience transit time acceleration in the interaction with the potential and the turbulence and will damp the turbulence. Some of them may thereby reach very high energies. Theory of this kind of particle acceleration is still under development and relies heavily on the availability of realistic numerical simulation models in many dimensions.

Bibliography

Excellent reviews on diffusive shock acceleration are:

Drury L O'C 1983 An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas *Rep. Prog. Phys.* **46** 973–1027

Forman M A and Webb G M 1985 Acceleration of energetic particles *Collisionless Shocks in the Heliosphere: A Tutorial Review* ed R G Stone and B T Tsurutani (Washington, DC: American Geophysical Union) pp 91–114

Jones F C and Ellison D C 1991 The plasma physics of shock acceleration *Space Sci. Rev.* **58** 259–346

The original work on the diffusion-convection equation and on quasilinear theory is found in:

Parker E N 1965 The passage of energetic particles through interplanetary space *Planet. Space Sci.* **13** 9–49

Jokipii J R 1966 Cosmic ray propagation—1. Charged particles in a random magnetic field *Astrophys. J.* **146** 480–7

Acceleration due to velocity shear has been investigated in:

Earl J A, Jokipii J R and Morfill G 1988 Cosmic ray viscosity *Astrophys. J.* **331** L91–L94

A recent article which deals with momentum diffusion in magnetosonic turbulence is:

Schlickeiser R and Miller J A 1998 Quasi-linear theory of cosmic ray transport and acceleration: the role of oblique magnetohydrodynamic waves and transit-time damping *Astrophys. J.* **492** 352–78

Manfred Scholer and Rudolf A Treumann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Particle Astrophysics

Particle astrophysics (or, equivalently, astro-particle physics), as the name implies, is an area of research which lies at the cross-roads of high-energy particle physics and astrophysics. Because most of the interest in the field is focused on high-energy phenomena, COSMOLOGY, particularly the early universe, is at the core of research activity.

The related area of nuclear astrophysics was developed as a means to understand the mechanism of solar energy production and stellar evolution. For example, nuclear processes are fundamental to our understanding of element production in stars. Particle astrophysics was developed in the 1970s and one could perhaps identify three problems that sparked the field. The first was the recognition that NEUTRINOS, if massive, could have a profound effect on the overall mass density of the universe. In the STANDARD MODEL OF PARTICLE PHYSICS, neutrinos are massless. The theory was constructed this way to account for the lack of any experimental evidence for a neutrino mass. As such, the theory was quite economical, as it predicted the minimum number of new undetected particles. By expanding the theory, neutrino masses can be accommodated, and if present, even if very slight in comparison with any known massive particle, would come to dominate the dynamics of the universe.

Another major motivating force in the development of particle astrophysics was the realization that theories which unify three of the fundamental forces in nature (electromagnetism, and the weak and strong nuclear forces), commonly called GRAND UNIFIED THEORIES or GUTs, possessed the ingredients necessary to solve a long-standing problem in cosmology relating to the asymmetry of MATTER versus ANTIMATTER. As such, these two problems made it clear that any complete description of our evolving universe would hinge on our understanding of the fundamental processes in nature, namely particle physics.

There are three known types or flavors of neutrinos. In the early universe, even massless neutrinos played a significant role in the dynamics of the rapidly expanding SPACE-TIME. If more than the three known neutrinos were present, the increased energy density would lead to an increase in the expansion rate and would alter the big bang NUCLEOSYNTHESIS (BBN) predictions of the light element abundances (of D, ^{3}He, ^{4}He and ^{7}Li). The concordance of these predictions with the observational determination of the abundances places a strong constraint on the expansion rate at the time of BBN (about 2 minutes after the big bang) and hence on the particle content of the universe at that time. The establishment of a limit on the number of neutrino flavors also played a key role in the development of the particle-astrophysics connection. The limit on the number of neutrino flavors has been generalized to constraints on a multitude of particle properties including limits on particle masses and lifetimes, and even limits on the variation of the fundamental constants of nature.

As the field developed, particle astrophysics has grown to include a wide variety of topics beyond neutrino masses and the matter asymmetry. The cosmological significance of stable particle masses provided an alternative means of solving the long-standing problem of DARK MATTER. Indeed, the theoretical constraints on theories with dark matter candidates and the experimental effort towards the detection of these candidates is one of the major thrusts in particle astrophysics today.

The study of grand unification in a cosmological context has also spread and diversified the domain of particle astrophysics. In addition to its potential for resolving the origin of the matter-antimatter asymmetry, grand unified theories predict the existence of very massive magnetic monopoles. Like electrons which have a unit of electric charge, these particles would carry a single unit of magnetic charge. However, the predicted number of these cosmological relics is clearly ruled out by experiment. Attempts at solving the monopole problem led to one of the major changes in what is referred to as the standard COSMOLOGICAL MODEL by introducing the notion of INFLATION, a period of extremely rapid expansion in the very early universe.

As a symmetry of nature, grand unification is not realized at the energies available to laboratory experiments. The breakdown of the symmetry and its effects (including the production of magnetic monopoles) has been an active area of research. Indeed, inflation was first thought of as a direct result of this symmetry breakdown. Although the early models of inflation were not successful, the study of other exotic phenomena, such as the production of topological defects which include cosmic strings, domain walls and textures, thrived. As theories of unification were pushed to include gravity, even the number of spatial dimensions was no longer held sacred. In STRING THEORY, the universe is a ten-dimensional space-time and in extensions of string theory, called M-theory, the universe is eleven-dimensional.

Although the description of all phenomena considered to fall within the domain of particle astrophysics is clearly beyond the scope of this article, probing a few of the topics described above will serve to define the nature of the field.

As noted above, although the standard model of electroweak interactions precludes the presence of neutrino masses, simple extensions of the theory can easily account for such masses. In the standard model, neutrinos and electrons form a weak left-handed doublet of the $SU(2)_L$ gauge group. This means that there are gauge interactions, mediated by a spin-1 particle, which are capable of interconverting (left-handed) electrons and neutrinos. In addition, the theory includes an $SU(2)_L$ singlet - a right-handed electron. Through its interaction with the Higgs boson, the left- and right-handed electrons mix to form a single, massive state. By adding a right-handed neutrino to the model, a neutrino mass can also be generated. However, we know experimentally that neutrino masses must be very small. To obtain a small

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mass, it is common to introduce a right-handed neutrino with a very large so-called Majorana (right-right) mass, M, in addition to the usual (left-right) Dirac mass, $m_{\rm D}$. Because of the left-right mixing, the physical neutrinos have masses which are different from the input masses, M and $m_{\rm D}$. To determine the physical masses one can write down a simple 'mass matrix' for neutrinos, which in a $(\nu_{\rm L}, \nu_{\rm R})$ basis can be of the form

$$\begin{pmatrix} 0 & m_{\rm D} \\ m_{\rm D} & M \end{pmatrix}. \quad (1)$$

Diagonalization of this matrix determines the physical neutrino masses, giving one massive state $\nu_1' \simeq \nu_{\rm R}$ with mass $m_{\nu_2} \simeq M$ and a lighter state, $\nu_1 \simeq \nu_L + {\rm O}(m_{\rm D}/M)\nu_{\rm R}$ with mass $m_{\nu_1} \simeq m_{\rm D}^2/M$. For $M \gg m_{\rm D}$, ν_1 can be made acceptably light. If all three flavors of neutrinos received mass in this way, the light states could mix (quantum mechanically) or oscillate.

Neutrino oscillations occur when the physical masses and input masses are not identical as described above. For example, nuclear processes in the Sun may produce an electron neutrino which if massive may not be a propagating state. Rather than a $\nu_{\rm e}$, it is one of the three light states $\nu_{\rm i}$ which travels to Earth from the Sun. But $\nu_{\rm i}$ contains some mixture of the three flavors of neutrinos $\nu_{\rm e}$, ν_μ and ν_τ. Depending on the degree of mixing, the distance between production and detection, and the energy of the neutrino, a ν_μ or ν_τ may be detected instead of $\nu_{\rm e}$. Neutrino experiments designed to detect $\nu_{\rm e}$s from the Sun have for decades recorded a deficit in the flux of electron neutrinos from the Sun. A similar problem is seen when detecting neutrinos produced in cosmic-ray showers in the atmosphere (see COSMIC RAYS: EXTENSIVE AIR SHOWERS). In a cosmic-ray collision with a nucleus in the atmosphere, many pions are produced. Pions decay primarily into muons which subsequently decay to electrons. Simple counting predicts that the ratio of muon to electron neutrinos is 2:1. However, observations indicate that this ratio is only about 60% of what is expected. Again, the explanation could be neutrino oscillations. In fact, it is widely believed that these experiments provide the first positive indication of a neutrino mass. Indeed, neutrino astrophysics is one of the strongest and most experimentally based components of particle astrophysics.

A light neutrino (with mostly left-handed interactions) would have been produced cosmologically and remained in equilibrium down to a temperature of about 3 MeV (the temperature scale at which the rate of weak interactions falls below the expansion rate of the universe). The relic number density of neutrinos is related to the neutrino temperature, $n_\nu \sim T_\nu^3$, and the mass density of relic neutrinos is simply $\rho_\nu = m_\nu n_\nu$. The neutrino temperature is predicted to be somewhat below the photon or microwave background temperature due to the annihilation of electron-positron pairs shortly after neutrino decoupling, $T_\nu = (4/11)^{1/3} T_\gamma$. Therefore the total mass density in neutrinos can be compared with the critical density necessary to close the universe ($\rho_{\rm c} = 1.05 \times 10^4 h^2$ eV cm^{-3} where h is the scaled Hubble parameter, $h = H/(100$ km Mpc^{-1} s$^{-1})$) and depends on the mass of the neutrino. The fraction of critical density in neutrinos is

$$\Omega_\nu = \rho_\nu/\rho_{\rm c} = 0.011 m_\nu({\rm eV}) h^{-2}. \quad (2)$$

The cosmological limit of $\Omega h^2 < 0.3$ (assuming that the age of the universe is greater than 12 Gyr) then places a constraint on the mass of the neutrino. In fact, it places a limit on the sum of all light neutrino masses

$$\sum_\nu m_\nu < 28 \text{ eV}. \quad (3)$$

The tiny mass scale should be compared with the electron mass $m_{\rm e} \simeq 0.511$ MeV, and the proton $m_{\rm p} \simeq 938$ MeV.

The connection between particle astrophysics and dark matter extends far beyond the case of light neutrinos. Very massive neutrinos (in the GeV mass range) were also considered as dark matter candidates. Neutrino interactions with matter (e.g. electrons) drop out of equilibrium when the average energy in the radiation background is about 3 MeV. At lower energies, the number densities and cross sections of neutrinos are too small to ensure scatterings with electrons in an expanding and cooling universe. Therefore, the relic density of much heavier neutrinos is determined by annihilations of ν and $\bar{\nu}$ pairs. At temperature scales below the neutrino mass ($kT \sim m_\nu c^2$ or m_ν in units where $c = 1$), neutrino annihilation is energetically preferred to neutrino production, and the density of neutrinos decreases. When the density becomes very low, the annihilations also stop (or freeze-out) and the relic density (relative to photons) becomes fixed. Because the efficiency of the neutrino annihilation scales with the neutrino mass, the cosmological limit $\Omega h^2 < 0.3$, places a *lower* bound on the mass of a heavy neutrino of about 4-7 GeV.

Experimental results based on the production of Z^0 gauge bosons at CERN and SLAC have shown that there are only three neutrinos ($\nu_{\rm e}$, ν_μ and ν_τ) with a mass less than m_Z. Therefore, the mass of a possible fourth heavy neutrino must exceed $m_Z/2 \simeq 45$ GeV. However, for such a heavy neutrino, the relic density is far too small for it to constitute a significant amount of dark matter, thereby excluding a heavy ($m_\nu > 1$ MeV) neutrino as dark matter.

The study of heavy neutrinos was nonetheless very fruitful, as many of the same relic density calculations also apply to other heavy dark matter candidates. For example, the well studied extension of the standard model of electroweak and strong interactions, known as SUPERSYMMETRY, predicts the existence of a new stable massive particle, the neutralino, which could account for the dark matter of the universe. Supersymmetry is a symmetry which relates particles of different spin. For example, all particles of spin-1, like the photon, must have a fermionic partner of spin-1/2 (for the photon, the partner is called the photino). For fermions, like the electron, a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spin-0 partner must exist (in this case, called the selectron). Supersymmetry predicts new interactions in addition to new particle types. In fact, some of these interactions could be responsible for the extremely rapid decay of the proton, which is clearly not observed. Therefore a new quantum number called R-parity is assigned to all particles. R-parity is related to baryon number, lepton number (see below for a description of these) and spin. As such, all observed particles are assigned an R-parity of +1, and all of the supersymmetric partners have R-parity -1. The lightest supersymmetric particle (commonly referred to as the LSP) is therefore stable as there are no lighter $R = -1$ particles to decay into. Supersymmetry is a theory which is currently under intensive experimental and theoretical investigation in both particle physics and particle astrophysics. Major efforts are under way to specifically search for the dark matter candidate that is predicted by supersymmetric theories and as such is a main focus in particle astrophysics today.

Indeed, the methods for experimental searches for dark matter can be placed into three categories: (a) accelerator searches, (b) indirect and (c) direct searches. Accelerator searches for dark matter are in principle the most direct, since one can hope to produce and detect the dark matter particle (or an associated particle) and establish its identity. However, these searches are typically designed to search for a very specific candidate with theoretically predicted interactions. The so-called direct and indirect dark matter searches are more general and do not depend on the specific properties of the candidate. Indirect dark matter searches typically key on some other particle rather than the dark matter candidate itself. Weakly interacting dark matter in our Galaxy (and solar system) continuously flows through the Sun and the Earth. For example, as a dark matter particle passes through the Sun, it will occasionally scatter and lose enough energy to become trapped in the Sun. As the number of these particles builds up, they begin to annihilate in the Sun and provide a source for high-energy neutrinos. Neutrino detectors (the same ones as used to measure neutrino oscillations) search for a flux of high energy neutrinos from the core of the Sun as well as the center of the Earth. Direct searches are based on laboratory detectors which are designed to detect the passage of dark matter particles. These are typically ultra-cold detectors which attempt to measure slight changes in the temperature of the detector induced by the interaction of the dark matter candidate and the material in the detector.

As can be inferred from the discussion above, the number of (light) neutrino flavors has been limited to three by experiment. However, particle astrophysics had provided a limit to this quantity ($N_\nu < 4$) years prior to the accelerator experimental results. Next to the prediction and discovery of the the 3 K microwave background, the prediction of the light element abundances of D, ^{3}He, ^{4}He and ^{7}Li is the most important test of big bang cosmology. The production of the light elements, and in particular ^{4}He, depends on a competition between the weak interaction rates which interconvert neutrons and protons and the rate of expansion of the universe. Because the standard models of cosmology, nucleosynthesis and particle physics (with three neutrino flavors) are so successful, most deviations from any of the standard models spoil the agreement between the theoretical predictions of the light element abundances and their observational determinations. Hence, BBN can limit the number of neutrino flavors. Other departures from the standard electroweak theory can also be constrained by BBN.

Nuclear processes in the early universe are very sensitive to the rate at which the universe expands and cools. The expansion rate of the universe (quantified by the Hubble parameter) is determined by the energy density of the universe. At early times (when curvature and the cosmological constant can be neglected) the Hubble expansion rate is given by

$$H^2 = \frac{8\pi G_{\rm N}}{3}\rho \qquad (4)$$

where the energy density in radiation is given in terms of the black body temperature and the total number of relativistic particle species present, $\rho \sim NT^4$. In the standard model, at the time of nucleosynthesis ($kT \sim 1$ MeV), $N = 2 + \frac{7}{2} + \frac{7}{4}N_\nu$, corresponding to the contributions of photons, electrons and neutrinos respectively. For $N_\nu = 3$, big bang nucleosynthesis predicts the correct abundances of the light elements. Due to the uncertainties in the observational abundances of the light elements, the limit from nucleosynthesis is $\sim$4, that is, the agreement between theory and observation becomes unacceptable when N_ν exceeds 4.

The limit to N_ν has proven to be one of the most valuable in the particle astrophysics connection. The limit in fact can be extended to any new particle degree of freedom present in the early universe at the time of big bang nucleosynthesis. Therefore, many new extensions of the standard model of particle physics can be, and have been, constrained by cosmology.

A key factor in the growth of particle astrophysics was the development of a wide class of constraints on theories of particle physics from cosmology and astrophysics. In addition to the BBN limits described above, our understanding of stellar processes has also provided valuable constraints. For example, certain theories predict new, massless or nearly massless particles. If these particles could be produced in the Sun, they would affect the rate of energy production and upset the agreement between observations and standard stellar models. In some cases, much stronger constraints are obtained by energetic arguments concerning RED GIANT STARS, which are relatively short lived. A new source of energy production could grossly affect the lifetime of these stars and make them in effect unobservable. The observation of neutrinos from the SUPERNOVA 1987A also provided a great opportunity for particle astrophysics to place constraints on many

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

phenomena beyond the standard model. In addition, this event, along with the new age of neutrino detectors (detecting neutrinos from the Sun and neutrinos produced in cosmic-ray collisions in the atmosphere), has opened a new door (NEUTRINO ASTRONOMY) which has grown to become a major focus of particle astrophysics.

In the late 1970s, the success of gauge theories in explaining the standard model of electroweak interactions led to ambitious efforts towards the unification of the strong and weak nuclear forces with electromagnetism, or Grand Unification. In the standard model, proton stability is guaranteed by the conservation of a quantum number called baryon number. If conserved, the net baryon number, like electric charge, must not change in any physical process. To date, all experimental results are consistent with the exact conservation of baryon number. Particles which are built out of three constituent quarks, like neutrons and protons, are ascribed a baryon number of 1 (quarks have baryon number 1/3). Electrons and neutrinos in contrast, being elementary and not comprised of quarks, have no baryon number. Although pions and other similar particles called mesons are built from quarks, because they are described by a quark-antiquark pair they also have no baryon number. The neutron, being heavier than the proton, can decay (into a proton, electron and anti-neutrino). However, the proton is the lightest state with baryon number $B = 1$, and is stable if baryon number is conserved.

One of the remarkable predictions of grand unified theories (GUTs) is the non-conservation of baryon number. The standard model interactions are mediated by particles called gauge bosons. There are 12 standard model gauge bosons: the photon, mediating electromagnetic interactions; the $W^{\pm}$ and Z^0, mediating the weak nuclear force, and eight gluons, mediating the strong nuclear force. GUTs not only place these 12 particles in a common multiplet, but predict the existence of additional gauge bosons and hence new interactions. The gauge sector of the smallest grand unified theory has 24 gauge bosons. The 12 new particles (three 'colors' of Xs and their antiparticles $\bar{X}$, and three 'colors' of Ys and their antiparticles $\bar{Y}$) mediate interactions between quarks and the class of particle which include the electron and neutrino called leptons. For example, an X gauge boson can take two quarks and turn them into an anti-quark and a positron (the antiparticle of the electron). In doing so, the X takes an initial state of baryon number +2/3, and converts it to a final state with baryon number $-1/3$, thus violating baryon number by one unit. In fact this process would allow the proton to decay into a positron and a pion.

Life, the universe and everything as we know it depend on the relative stability of the proton. Stars, planets and people could not exist if the proton lifetime were too short. The neutron lifetime or half-life, determined by the weak nuclear force and the mass of the $W^{\pm}$ gauge boson which is about 80 GeV (or about 80 times the mass of the proton), is just over 10 minutes. That means that a collection of free neutrons would be reduced by half after about 10 minutes. After 20 minutes, only 1/4 would remain, and so on. The lifetime of the decaying particle scales as $1/M^4$, where M is the mass of the gauge boson mediating the interaction.

The mass of the new GUT gauge bosons is determined by the unification energy scale, which in turn is determined by the point where the three standard model coupling constants meet after an energy-dependent renormalization. These constants determine the strength of the interactions. For example, the electromagnetic coupling constant, e, fixes the quantity known as the fine-structure constant $\alpha = e^2/4\pi$. The value of these constants must be adjusted at high energies according to the predictions of quantum field theory and as verified with high precision by experiment. At a very high energy scale, the constants take the same (unified) value. Though the exact energy scale is not known, it is typically about 10^{15-16} GeV. The masses of the X and Y gauge bosons are similarly large. The lifetime of the proton then scales as M_X^{-4} and is at least 10^{32} years. This extremely long lifetime should be compared with the current age of the universe of about 1.4×10^{10} years. Thus, although protons are allowed to decay, they do so very rarely. The search for proton decay has been under investigation since the late 1970s. To date, no decays have been observed.

Grand unified theories have had a profound impact on cosmology and were in part responsible for the development of particle astrophysics. GUTs provided the missing ingredients to the long-standing problem of a matter-antimatter asymmetry in the universe.

All particles exist as particle-antiparticle pairs (though a few particles like the photon and Z^0 are their own antiparticles). Antiparticles are produced either in high-energy accelerator experiments or naturally, when high-energy cosmic rays collide with nuclei in the Earth's atmosphere. However, apart from the secondary production of antimatter, there is no evidence of antimatter elsewhere in the universe. We know that the Earth and the solar system are composed solely of matter. If cosmic rays are of Galactic origin (as is widely accepted), then the lack of evidence of primary (not produced in a collision) antimatter in cosmic rays indicates that the Galaxy as a whole is also composed of matter. If nearby galaxies were composed of antimatter, the radiation produced by the annihilations of gaseous matter and antimatter along the boundary of these regions would be easily detected.

The question posed for cosmology and particle astrophysics is two-fold: (a) Why is there an apparent asymmetry between matter and antimatter in the universe? (b) Why is the asymmetry, commonly referred to as the baryon asymmetry (due to the asymmetry of neutrons and protons, though it is expected that there is a similar asymmetry in leptons - or electrons and neutrinos), small compared with the number density of photons in the universe? In fact, it is common to define a parameter measuring the baryon asymmetry

$$\eta = \frac{n_B - n_{\bar{B}}}{n_\gamma} \approx \frac{n_B}{n_\gamma} \approx (2-5) \times 10^{-10}. \qquad (5)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This quantity is estimated by comparing the observed density of matter, as determined for example by the visual and x-ray-output of galaxies and clusters of galaxies, with the density of photons determined by the measured temperature of the microwave background. This ratio can also be fixed by the observed abundances of the light elements, as η is the one free parameter of BBN theory.

In 1967, Sakharov derived three essential ingredients to solve at least the first problem. Namely the necessary conditions to generate an asymmetry of matter over antimatter. They are: (a) baryon number violating interactions; (b) a violation of two other symmetries of nature called C and CP - for charge symmetry and a combination of charge and parity; (c) a departure from thermal equilibrium. The necessity for the first condition is rather obvious. Without baryon number violating interactions, it would not be possible to change the total baryon number of the universe. As described above, grand unified theories provide baryon number violating interactions automatically. The second condition is more subtle. C, P and T (for time reversal invariance) are three fundamental symmetries of nature. In fact, all of particle theory is built on the exact conservation of the full combination of CPT (that is when all three operations are performed (charge conjugation, parity or $x \rightarrow -x$ and time reversal or $t \rightarrow -t$) simultaneously). However, individually they are known to be violated in certain interactions. Parity violation is explicitly demonstrated by the weak interactions, and was first observed in the decay of ^{60}Co in 1957. CP violation was observed in the decay properties of K^0 and $\bar{K}^0$ mesons in 1964. GUTs also contain CP violation, although the simplest GUT does not contain enough to produce the observed asymmetry of matter. Thus cosmology again places strong constraints on the building of particle theories. The violation of C and CP is necessary in order for the reactions to preferentially produce baryons rather than anti-baryons. Without this violation, any baryon number produced by a baryon number violating interaction would be exactly canceled by the corresponding reaction (under the charge symmetry), producing a net negative baryon number. Finally, the departure from thermal equilibrium is necessary so that reverse processes to those which produce a net baryon number are not operative. If they were, they would simply erase any baryon number produced. The departure from thermal equilibrium is a common feature of a rapidly expanding universe. If the expansion rate of the universe is fast compared with a specific interaction rate, that process will not be in equilibrium. In the late 1970s and early 1980s, it was shown that GUTs and standard big bang cosmology have all that it takes to explain the existence of the matter-antimatter asymmetry. The precise value of η has become a (GUT) model-dependent question. Although there are now many specific mechanisms proposed to produce the asymmetry, the exact mechanism is still an open question in particle astrophysics.

Along with the cosmological success of grand unification, a serious problem was discovered. Since a grand unified symmetry is not realized in nature today (or else we would find massless X and Y gauge bosons), it must be broken. In the process of breaking such a (grand) symmetry, it was discovered that new very massive particles called magnetic monopoles must exist. These particles, as their name implies, carry a single unit of magnetic charge much like the electron carrying a unit of electric charge. This would, in effect, appear like a magnet with only a north or south pole, but not both. Needless to say, such a particle has never been found.

In a cosmological context, grand unified theories predict the abundance of magnetic monopoles and their masses. Therefore, it is possible to compute their contribution to the energy density today. The masses of the monopoles are superheavy, of order M_X. Because the existence of monopoles is associated with the breakdown of the grand unified symmetry, their abundance is determined by the correlation length of the symmetry during the phase transition during which the symmetry is lost. The longer the correlation length, the fewer the monopoles produced. The maximum correlation length possible is determined by cosmology, and is related to the size of the observable universe at the time of the phase transition. This size is in turn determined by the age of the universe and is set by the maximum distance light can travel in that time. This distance is called the cosmological (or particle) horizon, and corresponds to the maximum distance over which information can be communicated. Therefore, this distance $d_{\rm H}$ is proportional to $t_{\rm U}$, the age of the universe. Since it is expected that the grand unified symmetry is broken at a temperature scale related to M_X, and since in the early universe $t_{\rm U} \propto T^{-2}$, the size of the horizon is very small and many massive monopoles are produced. The monopole problem, as it is referred to, is the expectation that the universe is dominated by monopoles today. In fact, the naive calculation predicts that the ratio of monopoles to baryons is about $\mathcal{O}(1)$. That is, as many monopoles as protons! This is clearly not the case.

There are several other problems which occur in the standard big bang model. Although the cosmological density parameter, Ω, is uncertain, it is known that its value is close to 1. As the universe evolves, this parameter also evolves, from an initial value extremely close to 1 (in the limit that $t_{\rm U} \rightarrow 0$, $\Omega \rightarrow 1$) to values much larger than 1 if the universe is closed, and to values much smaller than 1 if the universe is open. Since the evolution of the universe is basically a gravitational phenomenon, one might expect that Ω evolves on a typical gravitational time-scale which is set by Newton's constant $G_{\rm N}$. The natural time-scale, called the Planck time, is

$$t_{\rm P} = (G_{\rm N}\hbar/c^5)^{1/2} \approx 10^{-44}\ {\rm s}.$$

However, Ω has remained near 1 for over 10^{10} years or 10^{17} s, thus outlasting its welcome by over 61 orders of magnitude.

Another problem with the standard cosmological model is related to the near perfect isotropy of the COSMIC

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MICROWAVE BACKGROUND. The temperature of the microwave background was extremely accurately measured to be 2.728 K by the COBE satellite. Furthermore, it is measured to have the same temperature in all directions to an accuracy of better than 100 parts per million. In the standard cosmological model, there is no explanation for the isotropy of the background radiation. The detected background photons were last scattered when the background temperature was a few thousand degrees and the age of the universe was about 10^5 years. The horizon size at that time was considerably smaller than it is today. (Our horizon today comprises about 10^5 horizons corresponding to the time of last scattering.) Since there is no information communicated from one horizon to another, there is no intrinsic reason why different horizons would have the same temperatures at the time of last scattering. If horizons at last scattering did have different temperatures, it would be reflected as a large anisotropy in the microwave background today, as it is measured in different directions.

A solution to this problem is known as the inflationary universe. Originally formulated as a period of rapid (exponential) expansion in the very early universe, inflation was thought to have been produced during the grand unified phase transition. That is, while the large symmetry of GUTs was being broken, the universe was dominated by the (false) vacuum energy density of the the GUT, rather than by radiation. The result of the expansion and the subsequent rethermalization of the universe was a resolution of the cosmological problems discussed above. If monopoles are not reproduced at the end of inflation (if the GUT is not restored), the density of monopoles is diluted to the extent that we only expect about one magnetic monopole in the entire visible universe! The dramatic expansion of the universe has flattened the universe to such an extent that we would predict that Ω is in fact indistinguishable from 1 today. Then, so long as the universe can live long enough to enter an inflationary stage, it is perfectly natural that it would last far beyond the gravitational lifetime discussed above. Finally, since the size of our horizon was dramatically inflated, our entire visible universe and more originated from a small and causal region of the universe. In this case, one expects the isotropy of the background radiation (up to perturbations caused by quantum fluctuations during inflation). Although it is now recognized that standard GUTs cannot be responsible for inflation, there are many alternative models. The development of these models has been another main focus of particle astrophysics since the early 1980s.

Bibliography

Borner G 1993 *The Early Universe* (Heidelberg: Springer)
Kolb E W and Turner M S 1990 *The Early Universe* (Redwood City: Addison-Wesley)

Keith Olive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Particle Physics and Astronomy Research Council

PPARC is the agency through which UK taxpayers' money flows to fund astronomy (as well as particle physics). PPARC supports the EUROPEAN SPACE AGENCY as well as the telescopes with UK involvement such as the 8 m Gemini Telescopes, UK Infra Red Telescope (UKIRT), James Clerk Maxwell Telescope (JCMT), William Herschel Telescope (WHT) and the Anglo-Australian Telescope (AAT) along with Jodrell Bank and the Joint Institute for VBLI in Europe (JIVE). The Astronomy Technology Centre in Edinburgh is a PPARC establishment and provides the technology backbone for instruments for the telescopes.

PPARC, based in Swindon, runs the peer review and allocation system for the financing of astronomy research particularly by making grants to UK universities and argues the case for funding to the Office of Science and Technology and the Science Minister. Its budget for astronomy is approaching £100M p.a.

For further information see
http://www.pparc.ac.uk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pauli Exclusion Principle

A principle of quantum theory, devised in 1925 by Wolfgang Pauli (1900–58), which states that no two fermions may exist in the same quantum state. The quantum state of a particle is defined by a set of numbers that describe quantities such as energy, angular momentum and spin. Fermions are particles such as quarks, protons, neutrons and electrons, that have spin $= \frac{1}{2}$ (in units of $h/2\pi$, where h is the Planck constant). The principle implies, for example, that no two electrons in a given atom can have identical quantum values of energy, angular momentum and spin. It thereby restricts the number of electrons that can exist in the various electron shells that surround the nucleus of an atom and hence determines the electronic structure and chemical properties of atoms of different chemical elements. For example, two electrons can exist in the ground state (lowest energy level) of a helium atom only if they have opposite values of spin. The permitted permutations of quantum numbers allow a maximum of two electrons to exist in the innermost shell, eight in the second, and so on.

The principle also applies to the protons and neutrons in an atomic nucleus and to the quarks that make up protons, neutrons and other kinds of hadrons. When applied to free electrons, the principle implies essentially that no two electrons can have the same position and momentum. This limits the extent to which electrons can be squeezed together and gives rise to a pressure, called electron-degeneracy pressure, which prevents further contraction of compact stars such as white dwarfs. A similar argument applies to the close-packed neutrons inside neutron stars.

See also: angular momentum, atom, degenerate matter, electron, energy, hadrons, quantum mechanics, quantum theory, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pavo

(the Peacock; abbrev. Pav, gen. Pavonis; area 378 sq. deg.) A southern constellation which lies between Telescopium and Octans, and culminates at midnight in mid-July. It was first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, rather inconspicuous constellation, the brightest stars in Pavo are α Pavonis (Peacock), magnitude 1.9, and β Pavonis, magnitude 3.4. There are four other stars of magnitude 4.0 or brighter. Interesting objects include κ Pavonis, a Cepheid variable (range 3.9–4.8, period 9.09 days), NGC 6752, a fifth-magnitude globular cluster, and NGC 6744, an eighth-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pawsey, Joseph Lade (1908–62)

Radiophysicist and astronomer, born Ararat, Victoria, Australia, pioneered the use of a Lloyd's mirror arrangement for radio interferometry at Dover Heights in Australia, and located the source of solar radio noise within the disc of the Sun. As John Hey had suggested, the radio noise came from sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Payne-Gaposchkin, Cecilia Helena [née Payne] (1900–79)

Astronomer, born in England, married Sergei Gaposchkin, first woman to become a full professor at Harvard. Worked on stellar atmospheres, and in her 1925 dissertation suggested correctly that the great range in strength, from star to star, of absorption lines in stellar spectra was due to differing amounts of ionization (differing temperatures), not differing chemical composition. She suggested that hydrogen was the most abundant element in stars. This conclusion was at first resisted but accepted by 1929, in particular by HENRY RUSSELL.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pease, Francis G (1881–1938)

Astronomer and designer of optical instruments, born in Cambridge, MA, designed the 100 in telescope. With John Anderson used an interferometer to observe Betelgeuse in 1920 and obtained a diameter for the star of 0.047 arcseconds, very close to the value predicted by EDDINGTON.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Peculiar Galaxies

A peculiar galaxy is an object that cannot be easily classified as a SPIRAL, ELLIPTICAL, or IRREGULAR GALAXY based on its optical morphology. They constitute between 5% and 10% of the known galaxy population, although most 'normal' galaxies will show peculiar features if examined carefully. Peculiar galaxies show a great diversity of form. The vast majority can be attributed to strong gravitational tides generated in the close passage of two galaxies, to the extent that the terms 'peculiar galaxy' and 'interacting galaxy' are now virtually synonymous. These objects provided the first clear evidence that a galaxy's environment can profoundly affect its evolution (see GALAXY EVOLUTION).

Many peculiar galaxies are experiencing episodes of enhanced star formation, called starbursts. They also show a greater tendency to host ACTIVE GALACTIC NUCLEI (AGN) compared with the normal galaxy population. Conversely, when one examines the most luminous objects in the universe—quasars, radio galaxies and infrared galaxies—one often finds peculiar morphologies, suggesting a direct link between interactions and extreme levels of emission. Tidally induced mergers of separate galaxies almost certainly account for a substantial fraction of the local elliptical population. This process has probably played a key role in GALAXY FORMATION.

Peculiar galaxies can be used to deduce the structure of normal galaxies and in a few cases probe large-scale mass distributions. Observational and numerical studies of these objects offer important insights into phenomena that have shaped the formation and evolution of galaxies.

Galaxies that will not fit

EDWIN HUBBLE introduced his 'tuning fork' classification scheme in 1929, in which galaxies were grouped into spiral, elliptical and irregular families based on their appearance in photographic plates. However, even at this early stage a small number of galaxies defied these broad categories. Deep exposures of NGC 4038/39 (aka 'the Antennae') and NGC 5216/18 revealed long and faint filaments of light (see figure 1). Equally perplexing was the elliptical galaxy Messier 87 (M87, figure 1), which showed a remarkable blue linear 'jet' emanating from its core. As intriguing as these objects were, they seemed to be rare exceptions and attracted little attention. Their apparent scarcity was due to difficulties in recording extended and low-surface-brightness emission with the telescopes and photographic plates then in use. This situation changed dramatically with the advent of wide-field Schmidt cameras. On completion of the National Geographic–Palomar Sky Survey in 1956, the number of peculiar galaxies grew from a mere handful to many thousands—roughly 10% of the galaxies visible in the Sky Survey plates. This new population displayed a bewildering array of morphologies: luminous bridges and tails, often extending many galaxy diameters, ellipticals with faint ripples or dust lanes, ring-shaped galaxies, even systems that appeared to be dissolving. Figure 1 shows several examples taken from the Digitized Sky Survey and gives some idea of their varied (and beautiful) forms. Many of these were included in the photographic atlases of peculiar galaxies by Vorontsov–Vel'yaminov (1956) and Arp (1963), which are still widely used reference works.

To the early investigators, these striking objects suggested gigantic explosions, galaxies in birth, galaxies in fragmentation, material ejections from nuclei, galaxies compelled by strong magnetic forces—or entirely new forces, even the creation of matter. However, the earliest explanation has proven to be the most successful, namely *peculiar galaxies are the result of strong gravitational tides generated by near collisions of ordinary galaxies.* This was originally suggested by the fact that peculiar galaxies are usually found in close pairs, and often with symmetrically placed features. This collisional interpretation had been rejected initially because galaxy encounters were thought to be exceedingly rare and because tides were considered incapable of creating long and narrow plumes like those seen in 'the Antennae'. However, it was eventually realized that galaxies exist in groups, sometimes separated by only a few disk diameters. Galaxies should therefore experience several close passages over their lifetimes (see GALAXIES: INTERACTIONS AND MERGERS). It was Alar and Juri Toomre's landmark 1972 paper that conclusively showed how gravitational tides could produce systems like the 'the Antennae' through a near collision of two spirals. A wide range of peculiar morphologies could in fact result depending on the relative masses, inclinations, orbital speed and collision geometry of the two galaxies. Gravitational tides vary with distance like R^{-3}, so they are most significant over a brief interval near closest approach. It is here that large-scale peculiarities such as plumes and bridges are abruptly 'launched' into their trajectories. The Toomres also extended the work of Holmberg by suggesting that sufficient orbital angular momentum could be lost in an encounter for two galaxies to coalesce, producing something resembling an elliptical galaxy. Simple statistical considerations implied that most ellipticals in the NGC catalog could have arisen this way. The idea that galaxies were isolated island universes had been dramatically overthrown.

Peculiar galaxies have been studied extensively over the past 30 yr, both observationally and theoretically. During this period the tidal interpretation has proven to be so successful that the terms 'peculiar galaxy' and 'interacting galaxy' are often used interchangeably.

Peculiar galaxy morphologies

There is a great diversity of form among peculiar galaxies, as even a quick examination of the Vorontsov–Vel'yaminov or Arp atlases will show. Even so, a number of clear categories can be recognized:

Interacting spirals with tails and/or bridges

'Antennae'-like systems result from close passages of similarly massive spirals. Bridges (e.g. NGC 2535/36)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

require lower-mass companions to prevent quick capture and are thus less common. Both may extend 100–200 kpc from the host galaxies. Bridges and plumes originate in the outer disks, and are therefore often blue in color and gas rich. Some contain dwarf-galaxy-sized condensations of stars and gas that appear to be gravitationally bound.

Interacting spiral–elliptical pairs
Because their internal motions are not dominated by rotation, interacting ellipticals produce diffuse sprays of stars rather than tails and bridges. In spiral–elliptical encounters, therefore, a single tail or bridge is seen, and the spiral typically displays the most visible damage (e.g. NGC 274/5).

Shell galaxies
Roughly 40% of ellipticals show faint interleaved shells or ripples in their outer regions (e.g. Arp 230). These may be the result of weak tidal interactions or the cannibalization of another galaxy. The peculiar elliptical Centaurus A, for example, has both shells and a dust lane, suggesting the consumption of a gas-rich spiral.

Galaxies with optical jets
Remarkably linear filaments are sometimes found near peculiar spirals. Most are tidal tails and bridges that only appear linear owing to edge-brightening and projection effects (e.g. NGC 4676). The much rarer optical synchrotron jets are manifestations of AGN. M87 and 3C 273 are the best known examples.

cD and 'dumb-bell' galaxies
The centers of rich galaxy clusters are often occupied by giant ellipticals called cD galaxies. These frequently possess multiple nuclei, most likely from the cannibalization of another cluster galaxy. An extreme form of this class are 'dumb-bell' galaxies, which are binary ellipticals embedded in a common stellar envelope (e.g. NGC 750). These are expected to coalesce after a few orbits (i.e. $\sim 5 \times 10^8$ yr).

Merger remnants
Both the Vorontsov–Vel'yaminov and Arp atlases include objects now believed to represent the late stages in the merger of two galaxies (e.g. Arp 220). At this point they are roughly elliptical in appearance, although their vestigial tidal tails, large gas masses and double nuclei—generally visible only at radio and infrared wavelengths—betray their true origin.

Polar-ring galaxies
Roughly 0.5% of known S0 galaxies possess rings orbiting a plane perpendicular to the disk. NGC 4650A (figure 2) is the prototype polar-ring galaxy. The rings are often rich in gas and young stars and are believed to represent material captured during an interaction that has settled in semistable polar orbits. Spiral–spiral mergers, in which one is transformed into an S0, have also been proposed as a formation mechanism.

Ring galaxies
Ring galaxies such as the Cartwheel (figure 2) are created by the passage of a companion through a spiral's disk near the nucleus. This interaction crowds the disk's orbits into a large ring ($D_{ring} \sim$ 10–35 kpc) that propagates outward at speeds of ~50–100 km s^{-1}. Observations show that ~90% of the spiral's gas supply is concentrated in the ring, along with essentially all star formation activity.

The expected lifetimes of these peculiarities vary greatly. The multiple nuclei of cD and dumb-bell galaxies evolve in a deep and rapidly changing gravitational potential, and so are believed to be very transient, perhaps needing less than $\sim 10^8$ yr to merge. Tidal tails, bridges and sprays experience a much shallower potential, and should persist much longer, perhaps several billion years.

Most spirals show slightly distorted arms or minor asymmetries in their disks. These can be attributed to weak interactions with companions. Many otherwise 'normal' galaxies show peculiar morphologies at low light levels. Particularly dramatic examples have been discovered by David Malin using special techniques (see DETECTORS: PHOTOGRAPHY) to explore emission as faint as 0.5% of the night-sky glow (see figure 3). These structures may represent long-lived remnants of ancient interactions. There is in fact a wide range of morphological oddities among the general galaxy population. What we have been calling peculiar galaxies throughout this article actually represent the extreme end of a continuum.

Star formation in peculiar galaxies

It had been long noted that peculiar galaxies tended to be very blue in color, and often with unusually luminous star-forming complexes. Analysis of the UV excess Byurakan Survey galaxies, for example, showed that 40% possessed double nuclei or obvious tidal features. By the 1970s sufficient observational data had accumulated to examine possible links between unusual morphologies and enhanced STAR FORMATION activity. The most influential of these early investigations was that of Larson and Tinsley (1978), who compared the optical UBV colors of a normal galaxy sample taken from the Hubble atlas with peculiar galaxies from the Arp atlas. Figure 4, taken from that paper, shows that the Arp atlas galaxies are on average significantly bluer than the normal galaxy sample (i.e. smaller U-B for a given B-V) with a much wider spread of values. This could be simply understood if the peculiar galaxy sample had recently experienced brief episodes (~10^7–10^8 yr duration) of elevated star formation activity called a starburst. The large color dispersion primarily reflects differences in the starburst's age and strength, and/or dust obscuration. Research over the last two decades has established a clear connection between peculiar morphology and the occurrence of starburst activity. Theoretical studies show that tidal interactions can be very effective in triggering starbursts by driving strong spiral or ring density waves or by transporting gas to the inner nucleus. Observations show that star

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

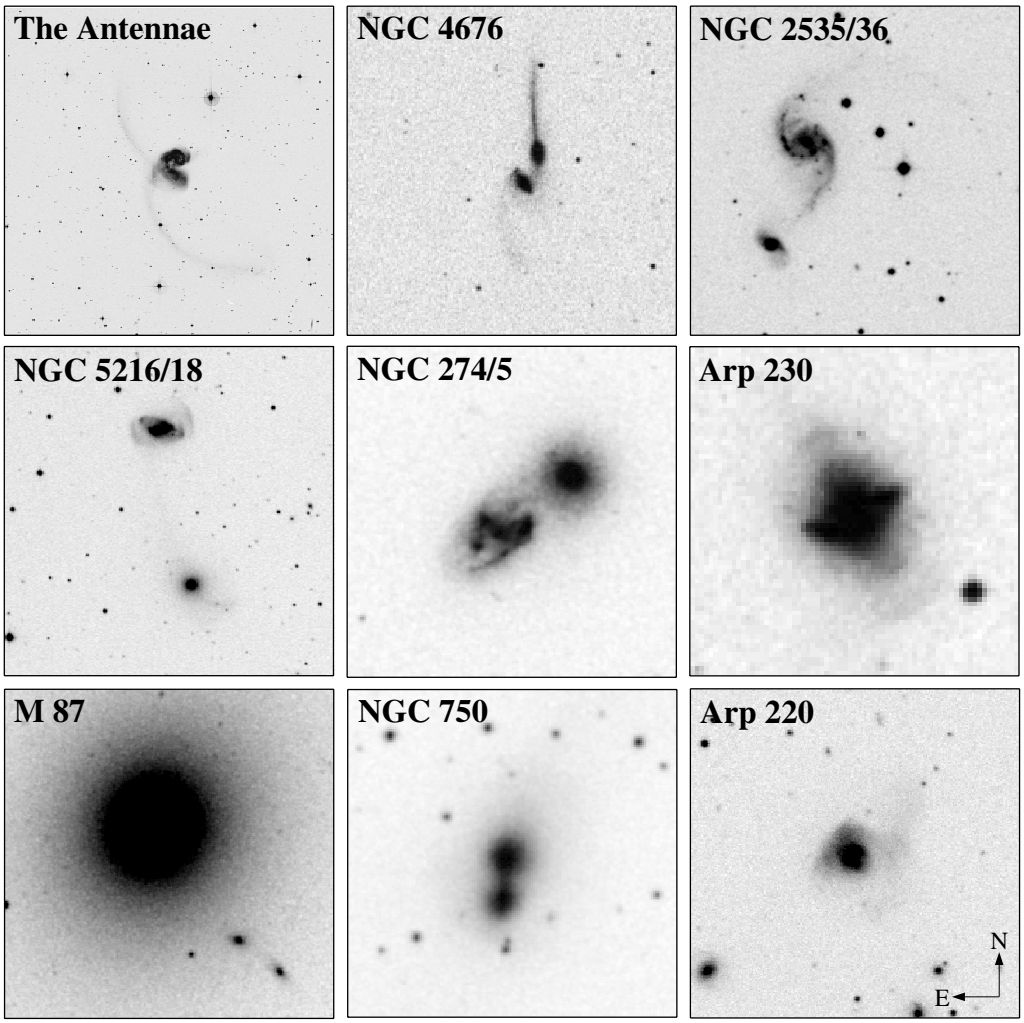


Figure 1. A montage of peculiar galaxies reproduced from the Digitized Sky Survey. They are presented in negative to make faint emission easier to see.

formation is increased only by modest factors of 2–3 over normal galaxies when averaged over the entire peculiar galaxy population. However, there is a wide range in star formation enhancement among individual peculiar galaxies, from essentially zero to several orders of magnitude. The induced star formation may be global in extent (e.g. the Cartwheel's ring). However, the most intense starbursts occur within the galaxy's inner kiloparsec. These nuclear starbursts tend to be found in the most strongly interacting systems. Excellent examples are the luminous infrared galaxies (LIRGs, e.g. Arp 220). LIRGs emit most of their total luminosity at infrared wavelengths ($\lambda > 5\,\mu$m), in at least half of the cases owing to dust heated by young massive stars. Astonishing star formation rates (SFR$\sim(10^2$–$10^3)M_\odot$ yr^{-1}) and gas masses ($M_{gas} \sim (10^{10}$–$10^{11})M_\odot$, i.e. a substantial fraction of the *total* gas mass) characterize their inner nuclear regions. LIRGs show a strong tendency to possess unusual optical morphologies. In the most luminous examples, essentially all are peculiar.

It should be emphasized that not all STARBURST GALAXIES appear highly peculiar. Also, not all peculiar galaxies are experiencing starbursts. This may reflect time delays between interaction and peak star formation, weak and quickly damped tidal perturbations, details of the collision or a lack of sufficient gas to fuel a starburst in the first place. Nevertheless, the frequency of starburst activity is much higher in peculiar galaxies. And, as the intensity of the starburst increases, so does the likelihood that the galaxy will possess a highly disturbed optical morphology.

Peculiar galaxies and active galactic nuclei

The first securely identified optical counterparts of powerful extragalactic radio sources (Cygnus A and Centaurus A) were strikingly peculiar. This fact suggested a direct connection between galaxy collisions and radio emission. However, subsequent observations found luminous radio emission to be far more likely to be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

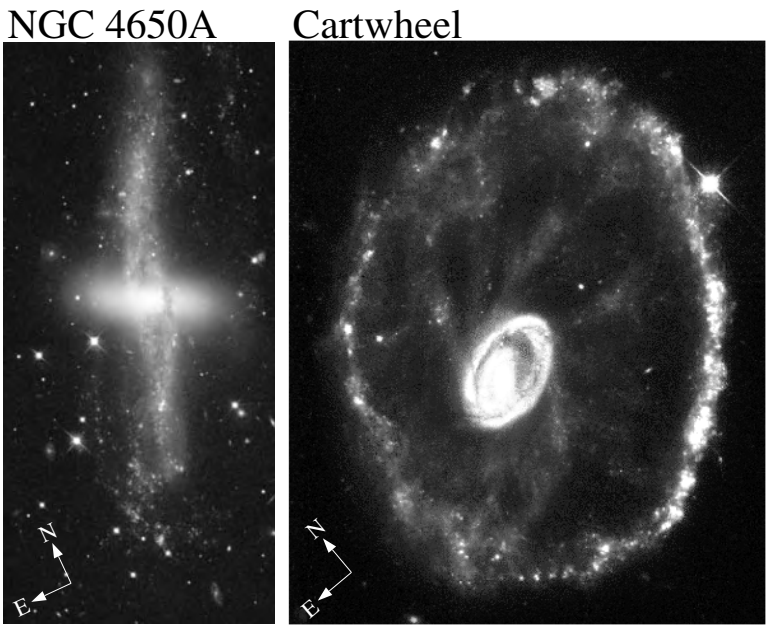


Figure 2. Two examples of peculiar rings: NGC 4650A (polar-ring galaxy) and the Cartwheel (ring galaxy). Both images were taken with the Hubble Space Telescope.

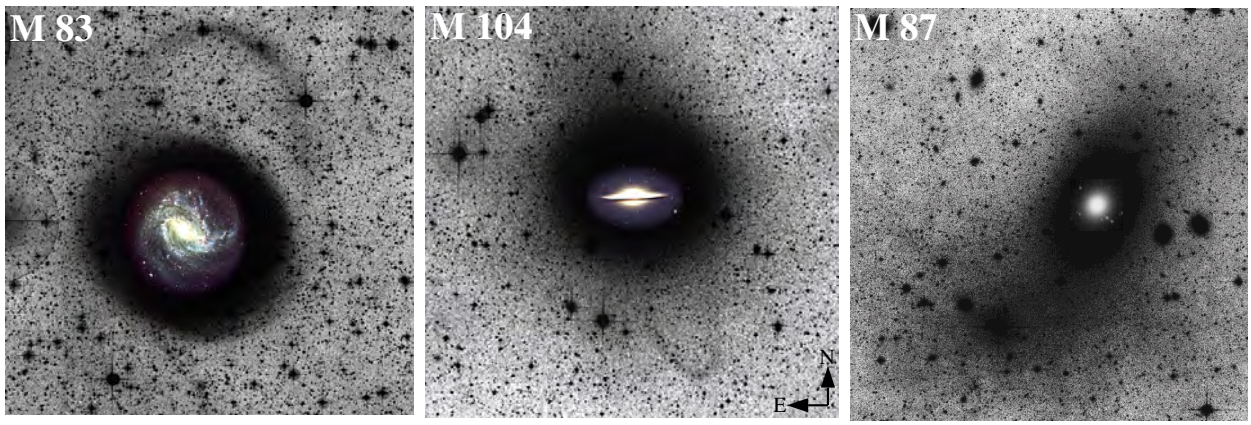


Figure 3. Enhanced photographs of nearby 'normal' galaxies by David Malin. The spiral galaxies M83 and M104 both show faint loops of starlight, while the elliptical galaxy M87 is embedded in an asymmetric spray of stars. More typical exposures have been inserted into the over-exposed centers to give an idea of their relative scales.

associated with the nuclei of rather ordinary looking ellipticals. The issue was reopened in the 1970s with the discovery that nuclear starbursts were common in peculiar galaxies. Theoretical results also suggested that interactions could efficiently transport gas to small radii and fuel an AGN's central engine. Recent surveys show that while the majority of galaxies with AGN in the local universe do not appear obviously peculiar, there is an increased tendency for peculiar galaxies to possess active nuclei. For example, Seyfert nuclei—the most common local type of AGN—occur more frequently in peculiar spirals than in normal spirals. The significance of this association increases with the degree of the morphological distortion. Further, one-third to one-half of the most powerful LIRGs appear to harbor heavily obscured AGN. These objects are invariably peculiar at optical wavelengths. In addition, ground-based imaging of low-redshift QUASARS shows that a significant fraction possess large-scale distortions, multiple nuclei or at least an excess of nearby peculiar companions. HST observations confirm that many quasars are morphologically peculiar. For example, of the 20 $z < 0.3$ quasars studied by Bahcall and collaborators (figure 5), more than a third are clearly interacting tidally, with half of the remainder showing faint extended structures resembling those of local merger remnants.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

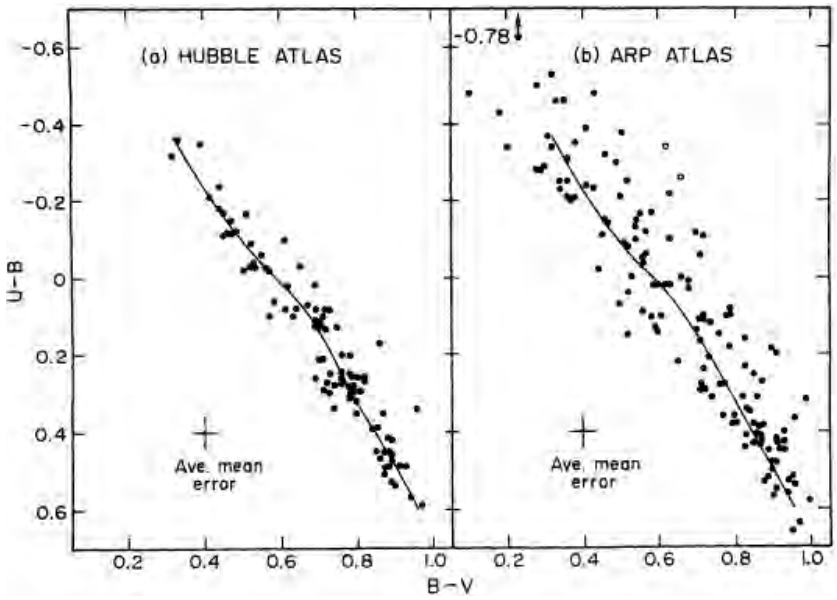


Figure 4. The UBV two-color distributions of normal (Hubble atlas) and peculiar (Arp atlas) galaxies reproduced from Larson and Tinsley (1978). The solid curve represents the mean through the Hubble atlas sample. Typical uncertainties in the colors are shown in each panel.

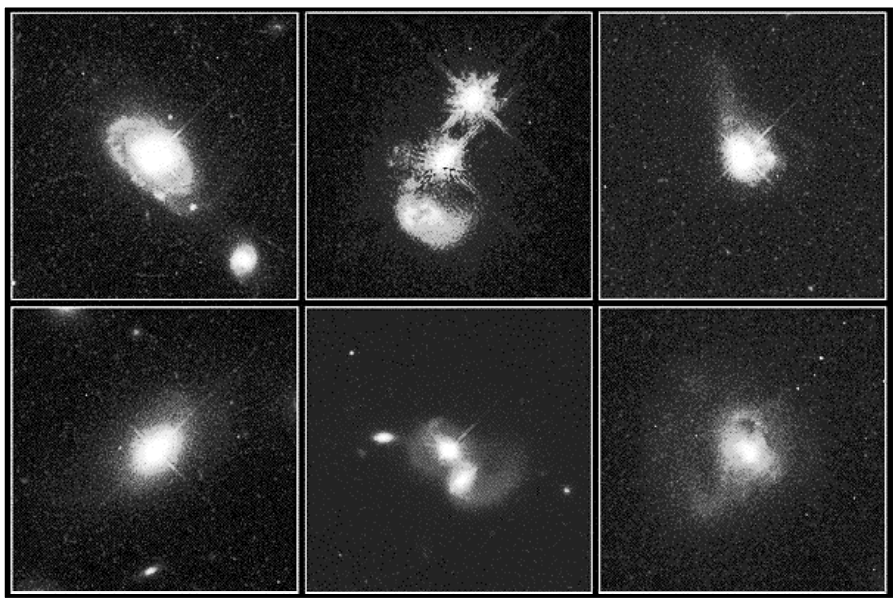


Figure 5. Optical images of low-redshift ($z < 0.3$) quasars taken with HST (Bahcall and collaborators, 1997). Note the close companions and/or obvious tidal features.

Again, most galaxies with AGN are not noticeably peculiar, and most peculiar galaxies do not possess AGN. Nevertheless, there is at least indirect evidence for a connection between galaxy interactions and at least some forms of powerful AGN.

Concluding remarks

The vast majority of peculiar galaxies represent highly perturbed ordinary galaxies. Observational and numerical studies of these systems can therefore provide insights into galactic structure. The kinematics of polar rings and tidal tails, for example, can be used to probe dark halos on scales much larger than ordinary optical or gas disks. Detailed computer simulations of interacting systems not only help

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

refine the input galaxy models, they also allow one to better understand how galaxies might evolve in more complex environments such as compact groups or rich clusters. Further, several lines of evidence suggest that galaxies are assembled through the coalescense of dwarf galaxy sized components at high redshift, and there is evidence that at least some nearby ellipticals have arisen through mergers. The ongoing mergers and merger remnants in the peculiar galaxy population can thus serve as templates for the study of galaxy formation and evolution. Finally, the energy released by starbursts and AGN can have a major impact on the energy balance in a galaxy's interstellar medium. Likewise, studies of peculiar galaxies such as Arp 220 show that starbursts can also supply substantial amounts of heavy elements. Both processes appear to have been much more common at higher redshifts. These few examples show that peculiar galaxies are important for what they can tell us about the structure and evolution of galaxies and for providing local examples of processes that have shaped the chemical and luminosity evolution of the universe.

Bibliography

The peculiar galaxy literature is vast. Fortunately, there are a number of excellent reviews with extensive references, e.g.

Kennicutt R, Schweizer F and Barnes J 1996 *Galaxies: Interactions and Induced Star Formation (Saas-Fee Advanced Course 26, Lecture Notes 1996)* (Swiss Society for Astrophysics and Astronomy)

One would also benefit by reading such classic papers as

Holmberg E 1941 *Astrophys. J.* **94** 385
Larson R and Tinsley B 1978 *Astrophys. J.* **219** 46

and of course

Toomre A and Toomre J 1972 *Astrophys. J.* **178** 623

Perhaps the best introduction to this subject is simply to browse through the Vorontsov–Vel'yaminov and Arp atlases. Electronic versions exist on the web (both atlases can be found at nedwww.ipac.caltech.edu/level5/pi-galaxies.html), although the quality of Arp's atlas is such that one should try to locate a large format copy in a university science library.

James L Higdon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Peebles, Phillip James E (1935–)

Canadian cosmologist, born in Winnipeg, professor at Princeton University, predicted the existence of the cosmic background radiation just before it was found by A PENZIAS and R WILSON. Investigated the clustering and superclustering of galaxies, providing evidence of the existence of large quantities of dark matter in the haloes of galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pegasus

(the Winged Horse; abbrev. Peg, gen. Pegasi; area 1121 sq. deg.) A northern constellation that extends from Cygnus, Lacerta and Andromeda almost to the celestial equator, and culminates at midnight in early September. It is named after the winged horse in Greek mythology that sprang from the body of Medusa, the Gorgon, when she was beheaded by Perseus, and later was tamed by the hero Bellerophon. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A large constellation, Pegasus is easily recognized by the asterism of the Square of Pegasus, formed by the stars β Pegasi (Scheat), α Pegasi (Markab), γ Pegasi (Algenib) and α Andromedae (Alpheratz). The last, which marks the north-east corner of the Square, was originally designated δ Pegasi. The brightest stars in Pegasus are ε Pegasi (Enif), magnitude 2.4, α Pegasi, magnitude 2.5, β Pegasi, a red giant irregular variable (range 2.3-2.7), γ Pegasi, a Beta Cephei variable (range 2.8–2.9, period 0.15 day), and η Pegasi (Matar), magnitude 2.9. There are five other stars of magnitude 5.0 or brighter. Other interesting objects include the W Ursae Majoris type variable star U Pegasi (range 9.2–10.1, period 0.37 day) and M15 (NGC 7078), a sixth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Peiresc, Nicolas-Claude Fabri de (1580–1637)

Polymath and astronomer, born Belgentier, Var, France, widely educated in Europe. In 1610 his patron, du Vair, acquired a telescope with which Peiresc and Joseph Gaultier were the first in France to see the satellites of Jupiter. With assistants in various locations, he recorded the positions of the satellites of Jupiter and used these observations to calculate terrestrial longitudes. Organized an international campaign to observe the lunar eclipse of 28 August 1635, and was able to correct the considerably over-estimated length of the Mediterranean. Discovered the Orion nebula, M42, the first deep sky discovery with a telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pelican Nebula (IC 5070)

An emission nebula in the constellation Cygnus, position RA $20^{h}\,50.8^{m}$, dec. $+44^{\circ}\,21'$. It measures $80'$ by $70'$, but has a low surface brightness. Its eastern border, the 'pelican' profile, is delineated by dark nebulosity which separates it from the North America Nebula (NGC 7000).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Penston, Michael Victor (1943–90)

Astronomer, born in London, a 'lover of active galaxies' (a LAG), as he described himself, he witnessed the change of NGC 4151 from one type of active Seyfert galaxy to another and determined the mass of the black hole at its center. This went towards a proof of the unified theory of active galaxies by which it is hypothesized that the various types are fundamentally the same phenomena seen at different angles and in different circumstances, and set the way for the Hubble Space Telescope's more accurate determination of the mass of black holes in different active galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Penzias, Arno Allan (1933–)

Radioscientist, born in Munich in Germany, Nobel prizewinner (1978) 'for the discovery of cosmic microwave background radiation', a refugee from Germany at the age of 6, found his way to America and experience in microwave physics. Joined Bell Laboratories, Holmdel, New Jersey, searched for and investigated line emission from the interstellar OH molecule. Was able to gain the use of a large radio telescope (the Holmdel horn) and a new ultra low-noise 7 cm travelling wave maser available for radioastronomy. With ROBERT WILSON began a series of radioastronomical observations intended to make the best use of the careful calibration and extreme sensitivity of the system, including a measurement of the radiation intensity from the Galaxy at high latitudes. Discovered the cosmic microwave background radiation. Made a millimeter-wave receiver and discovered a number of interstellar molecular species, including deuterated hydrogen cyanide.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Periapsis

The point in an elliptical orbit at which the orbiting body is farthest from the body it is orbiting. The prefix 'peri-' may be attached to various words or roots depending on the body being orbited: for example, periastron for an orbit round a star; perijove for an orbit around Jupiter.

See also: perigee, perihelion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Periastron

The closest point in the relative orbit of one component of a binary system to the other component. The term is also used for the closest position of a planet orbiting a star other than the Sun.

See also: binary system, perihelion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perigee

The point in its orbit around the Earth at which the Moon or an orbiting spacecraft is closest to the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perihelion

The point in the elliptical orbit of a planet or other object around the Sun at which it is closest to the Sun. The Earth, for example, reaches perihelion in January, when it is about 147 million km from the Sun. This is the Earth's *perihelion distance*. The heliographic longitude of a body's perihelion point steadily moves around the Sun in the direction of the object's orbital motion as the major axis of the ellipse precesses. This *advance of perihelion* is largely a result of gravitational perturbations by the planets on one another, but there is a small component caused by the curvature of spacetime in the vicinity of the Sun. The latter factor is most pronounced for Mercury, and provided an early test for the general theory of relativity.

Compare: aphelion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perrine, Charles Dillon (1867–1951)

American astronomer, born in Teubenville, Ohio, became director of Cordoba Observatory in Argentina, constructed a 30 in telescope and observed galaxies. Was retired prematurely when, in political unrest, he became the target of nationalist sentiment and an assassination attempt. Discovered Jupiter's sixth and seventh satellites with the 36 in Crossley reflector at Lick Observatory, numerous comets and the expanding light echo around Nova Persei as the delayed nova flash reflected off nearby dust clouds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perseids

The best-known meteor shower and one of the most prolific, occurring in late July and August. In medieval times the shower was known as the Tears of Saint Lawrence, who was martyred on 10 August 258, close to the time of maximum Perseid activity. The radiant begins in the constellation Cassiopeia and moves through northern Perseus. The high inclination (113°) of the parent comet, 109P/Swift–Tuttle, keeps the Perseid meteor stream largely free from planetary perturbations, and its activity has remained fairly constant for many centuries. Zenithal hourly rates rose during the 1980s and 1990s, either side of the parent comet's return in 1992.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perseus

(abbrev. Per, gen. Persei; area 615 sq. deg.) A northern constellation which lies between Andromeda and Auriga, and culminates at midnight in early November. It is named after the hero in Greek mythology who beheaded the Gorgon Medusa and rescued Andromeda from being sacrificed to the sea monster Cetus. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A conspicuous constellation, the brightest stars in Perseus are α Persei (Mirfak or Algenib), magnitude 1.8, β Persei (Algol, the 'Demon Star'), the prototype eclipsing binary (range 2.1–3.4, period 2.86 days), ζ Persei, magnitude 2.8, and γ and ε Persei, both magnitude 2.9. There are 11 other stars of magnitude 4.0 or brighter.

The Milky Way passes through Perseus and the constellation contains a number of open star clusters, including the Double Cluster (NGC 869 and 884, or h and χ Persei), the two components of which are each about 45′ across, contain a few hundred stars, and are just visible to the unaided eye as a faint misty patch midway between γ Persei and δ Cassiopeiae, M34 (NGC 1039), which contains about 80 stars fainter than eighth magnitude, and Melotte 20, a loose assemblage surrounding α Persei. Other interesting objects include M76 (NGC 650-51), the Little Dumbbell Nebula, a tenth-magnitude planetary nebula, NGC 1023, a tenth-magnitude spiral galaxy, and NGC 1275 (Perseus A), a twelfth-magnitude elliptical galaxy that is the brightest member of the Perseus cluster of galaxies and a strong radio source.

The Perseid meteor shower appears to radiate from a point about 4° north of γ Persei.

See also: Algol, Perseids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Peurbach [Peuerbach, Purbach], Georg von (1423–61)

Born in Peuerbach, Austria, became court astronomer of King Ladislaus of Hungary and was served as professor of astronomy at the University of Vienna. Observed Halley's and other comets in and, with his student REGIOMONTANUS, recorded the lunar eclipse of 3 September 1457 checking his own tables of eclipse calculations (*Tabulae Ecclipsium*) based on Ptolemy's epicyclic theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

PG 1159 Stars

The PG 1159 stars form a small group of about 30 objects. They are named after the prototype PG 1159-035 (=GW Vir), which was discovered in the Palomar-Green survey. They are hot post-asymptotic giant branch (post-AGB) stars and their location in the HR diagram coincides with the hottest central stars of PLANETARY NEBULAE (CSPN) and the hottest WHITE DWARFS (WD). The most peculiar characteristic is their chemical surface composition. They are hydrogen deficient and the most abundant elements are He, C and O with typical abundances of 33%, 50% and 17% respectively. This suggests that their immediate progenitors are CSPN with Wolf–Rayet type spectra, whose analyses yield similar abundance patterns. It is thought that most of the PG 1159 stars evolve into non-DA WDs, i.e. helium-rich objects from which the heavy elements were removed from the surface layer by gravitational settling. The origin of the PG 1159 surface chemistry is under debate. The most popular picture describes these stars as the result of a late He-shell flash, i.e. the re-ignition of helium fusion in a CSPN or even in a cooling, contracting WD. This event causes the star to return onto the AGB and to retrace its post-AGB evolution for a second time (born-again AGB star scenario). As a consequence, envelope convection causes ingestion and burning of surface hydrogen and, possibly in conjunction with mass loss, carbon and oxygen as helium burning products are mixed into surface layers.

The mere existence of PG 1159 stars represents a challenge to post-AGB STELLAR EVOLUTION theory. Details of the born-again scenario are being worked out, in order to prove or disprove it. An alternative scenario proposes that close binary evolution causes stripping of the envelope during the RED GIANT phase. Since some PG 1159 stars are associated with a planetary nebula, it is hoped that analyses of the planetary nebulae will give further hints as to the stellar history. Some PG 1159 stars are variable and these multi-periodic g-mode pulsators are defining the GW Vir (or DOV) instability strip in the HR diagram. The pulsational properties are used to probe the interior structure of these stars. Hence spectroscopic analyses with model atmospheres in combination with asteroseismology using pulsating stellar models help to constrain theoretical evolutionary calculations for PG 1159 stars.

The hydrogen-deficient post-AGB sequence [WC] → PG 1159

Hot post-AGB stars ($T_{\rm eff} > 20\,000$ K) can be grouped into either H-deficient or H-'normal' objects. The evolution of the latter group is in principle understood, so research focuses on the problems with H-deficient objects. Spectroscopic analyses have firmly established an evolutionary sequence formed by the Wolf–Rayet CSPN and the PG 1159 stars. This is suggested by their location in the g–$T_{\rm eff}$ plane (figure 1) as well as the similarity of their chemical abundance patterns. The Wolf–Rayet CSPN, dominated by He, C and O emission lines, are subdivided into late and early spectral types, [WCL] and [WCE] (brackets are used to discriminate these CSPN from the massive WOLF–RAYET stars). The link between the [WCE] and PG 1159 stars is marked by a few [WC]–PG 1159 transition objects which show a mixed emission/absorption line spectrum. The PG 1159 stars display essentially absorption line spectra. Their spectroscopic hallmark is a broad absorption trough formed by He II and C IV lines (figure 2). While [WC] stars have been known for a long time, the PG 1159 stars have been established as a spectral class more recently (Wesemael *et al* 1985). We know 31 PG 1159 stars and about every other one is a CSPN. Note that the so-called 'OVI' classification of CSPN is no longer used because objects having very different spectra were subsumed under it. All former 'OVI' stars are now [WCE] stars of different subclasses according to an appropriate classification scheme. An exception is the well known CSPN NGC 246, also a former 'OVI' object, which is now classified as PG 1159.

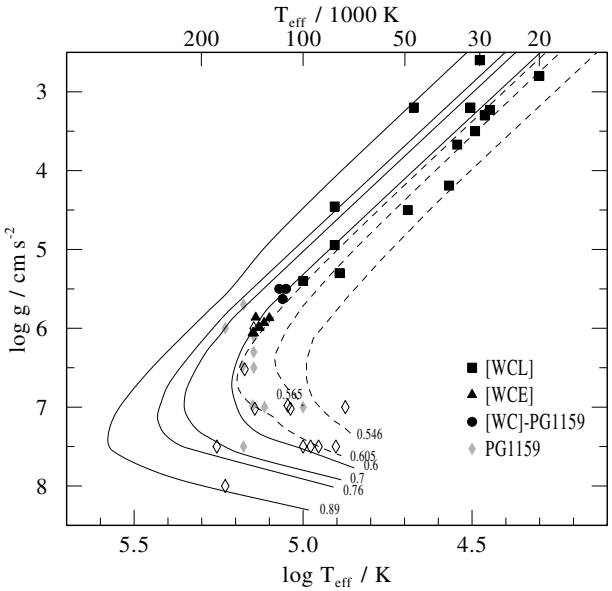


Figure 1. Hydrogen-deficient post-AGB stars in the effective temperature versus surface gravity plane. We identify Wolf–Rayet central stars of late and early spectral subtype ([WCE] and [WCL]), PG 1159 stars as well as [WC]–PG 1159 transition objects. Open symbols denote stars not associated with a planetary nebula. Evolutionary tracks (labels: mass in $M_{\odot}$) are only shown for a rough comparison, because they predict normal, H-rich surface abundances.

An overview of knowledge of PG 1159 stars with detailed references is found in Werner *et al* (1999). For detailed spectroscopic results on the [WC] stars we refer to a review by Hamann (1997).

Origin of H-deficient post-AGB stars

The typical chemical composition of PG 1159 atmospheres is displayed by the prototype PG 1159-035 itself. It was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

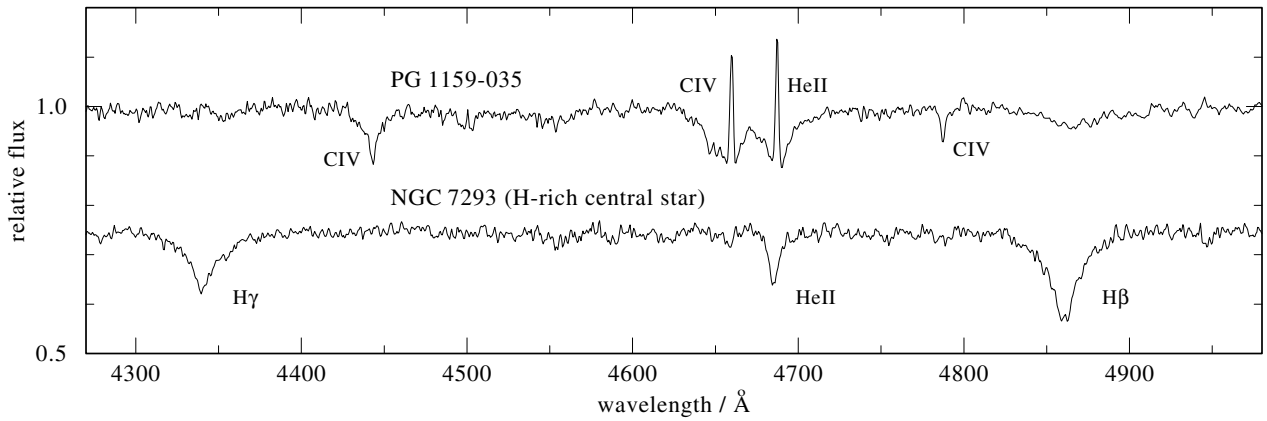


Figure 2. Optical spectrum of PG 1159-035 as compared to a central star with a usual hydrogen-rich composition. The characteristic PG 1159 spectrum lacks hydrogen lines and shows a prominent He II/C IV absorption line trough near 4670 Å, accompanied by central emission reversals in the line cores in the case of the prototype.

found to be He:C:O=33:50:17 (mass fractions) while hydrogen is not detectable. A comparison of this abundance pattern with the interior structure of post-AGB stellar models suggests that the PG 1159 stars must have lost their entire H-rich envelope ($\approx 10^{-4} M_{\odot}$) and also most of their He-rich intershell layer ($\approx 10^{-2} M_{\odot}$) because the high C and O abundances are found only at its bottom. With the first reliable non-LTE model atmosphere analyses of PG 1159 stars available it was suggested that the H deficiency is a consequence of a late He-shell flash. Originally, Iben (1984) introduced this scenario in order to explain the discovery of H-deficient zones in the inner parts of the PNe Abell 30 and Abell 78 (whose nuclei are today classified as [WC]–PG 1159 transition objects). In this scenario He-shell burning re-ignites when the post-AGB star is already on the WD cooling track. As a consequence, the star moves back to the AGB ('born-again' AGB star). The star suffers another superwind phase and during its second descent from the AGB the star loses its remaining hydrogen envelope by a fast wind. FG SAGITTAE and SAKURAI'S OBJECT (V 4334 Sgr) are thought to be such stars which are on their way back to the AGB (Asplund 1999).

There are several problems with this scenario, arising from several observational facts, which suggest that the H deficiency is not only formed during the second post-AGB evolution, but already during the second AGB phase. In spite of the rapid evolutionary rate, the existence of cool (i.e. young) [WC] stars requires the complete ejection of the H-rich envelope on the AGB. This is also corroborated by the fact that the ejection of H-deficient material in the planetary nebulae Abell 30 and Abell 78 occurs with low expansion velocities (<100 km s^{-1}), which means that the rather extended AGB precursor of the more compact CSPN must already have been H deficient.

The born-again scenario is faced with another problem, which is related to the expansion ages of the nebulae of [WC] stars. Their planetary nebulae should be systematically older than those of the CSPN with normal surface chemistry; however, this is not the case. This finding suggests that the H deficiency possibly occurs at the first departure from the AGB, i.e. without invoking the born-again scenario at all. What then is the origin of the H deficiency? One possibility is close binary evolution during which the stellar envelope could be stripped off by interaction with the companion star. However, for [WC] stars binarity is extremely difficult to detect directly from radial velocity or photometry. For PG 1159 stars the binary frequency does not appear to be unusually high. Bipolarity of planetary nebulae has frequently been invoked as a tracer of binarity. Here we face a puzzling fact: while the frequency of bipolar planetary nebulae around [WC] stars appears to be 'normal' (20%), the fraction of PG 1159 stars with a bipolar planetary nebula appears to be high (75%). The latter result, however, is limited by small-number statistics and inhomogeneous observational material.

A further problem represents the high photospheric oxygen abundance found in many PG 1159 and [WC] stars, which approaches almost 20% by mass. As already mentioned above, this means that not only must the hydrogen-rich envelope be removed from the star, but also the 100 times more massive helium envelope. This conflicts even more severely with the existence of the young [WCL] stars: removal of the He-rich intershell during the post-AGB evolution can be enforced in theoretical calculations by high mass-loss rates, but envelope removal is not completed until the star is very hot ($T_{\rm eff} > 100\,000$ K), because the main effect of mass loss is to increase the evolutionary rate. However, new evolution calculations show that this problem is probably much less severe. These calculations, which include a hydrodynamically based overshoot prescription, provide a very high oxygen abundance throughout the He-rich intershell, and not only at its bottom as suggested by earlier computations. Though the problem of envelope removal seems less serious now, it is still unsolved. In any case, we

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

have to face the exceptional case of H 1504+65, a PG 1159 star which is completely devoid of hydrogen and helium, i.e. its atmosphere is mainly composed of C and O by equal parts, with an admixture of neon (Werner and Wolff 1999). Obviously this naked C/O stellar core must have lost also its massive He-rich intershell.

Further evidence that the H deficiency is formed on the AGB or at least very early on the post-AGB track comes from the analysis of ISO spectroscopy of two [WCL] CSPN (BD+30°3639 and He 2-113.). C-rich dust (PAHs) in the inner part of the planetary nebula and crystalline (O-rich) silicates in the outer part suggest that the chemical composition of both stars changed to a C-rich mixture very recently, i.e. less than 1000 years ago. The H-rich nature of the C-rich dust suggests that some hydrogen was still left on the star, which is in accordance with the identification of hydrogen in the atmosphere of one of these CSPN (BD+30°3639).

The detection of residual hydrogen in some [WC] and PG 1159 stars as well as the detection of nitrogen in a few cases suggests that mixing and incomplete burning of the hydrogen envelope takes (or took) place in these stars. According to evolutionary calculations this can only occur after a very late He-shell flash (i.e. on the WD cooling track, where H-shell burning is extinct) and not on or soon after the AGB phase (where H-shell burning is still active). This is in clear contradiction with the planetary nebula observations of H-deficient gas and dust around the stars (see above).

To conclude, the true origin for H deficiency in post-AGB stars is not known. Further progress can be expected from evolutionary calculations which must include both overshoot as described above and time-dependent mixing.

Another H-deficient post-AGB sequence

A few hot H-deficient post-AGB stars cannot be classified as PG 1159 stars, because their absorption line spectra exhibit almost exclusively He lines and, if at all, only weak carbon or other metal lines. This small group of stars comprises four objects, two of which have an associated planetary nebula. These stars are classified as 'O(He)' and the prototype is the CSPN K1-27. Spectral analyses confirm that only trace amounts of metals are present in the helium-dominated atmospheres but, interestingly, two of them have a considerable fraction of hydrogen left (10% and 30%, by number). They are very hot ($T_{\rm eff}$ = 100 000–140 000 K) and have relatively low surface gravities ($\log g$ =5.5–6.5); hence, their location in the g–$T_{\rm eff}$ plane is among the PG 1159 and [WCE] stars. It is possible that the O(He) stars are the successors of those He-rich subdwarf O stars, which are on post-AGB evolutionary tracks. These, in turn, might be connected to the R CrB stars, which are He-rich giants very similar to Sakurai's object in its present state.

The variety of abundance patterns among PG 1159 stars and [WC] stars is rather diverse and the existence of O(He) stars makes an explanation for the origin of H-deficient stars even more difficult. At present it can only be speculated that, among other details, the phase in the thermal pulse cycle at the moment of AGB departure plays an important role.

Mass loss from H-deficient post-AGB stars

From the above discussions it is clear that knowledge about mass loss from [WC] and PG 1159 stars is essential for our understanding of the H-deficient post-AGB stars. Mass-loss rates from [WC] stars can be inferred from analyses of the optical emission line spectra. Good spectral resolution is required in order to eliminate the contribution from nebular emission lines to the photospheric features. As a result, high mass-loss rates are found, ranging between 10^{-5} and $10^{-6.5} M_{\odot}$ yr^{-1}, with rates decreasing along with proceeding post-AGB evolution. The mass-loss rate of two [WC]–PG 1159 transition objects amounts to $10^{-7.5} M_{\odot}$ yr^{-1}. For the PG 1159 stars one expects even smaller rates. As a consequence, the optical spectrum is formed in the hydrostatic layers of the photosphere and only the strong UV resonance lines of C IV and O VI can be used to study the properties of the weak winds. IUE, HST and ORFEUS spectra were analysed of four low-gravity PG 1159 stars, i.e. immediate successors of [WCE] stars, and one found $\dot{M} = 10^{-7}$–$10^{-8} M_{\odot}$ yr^{-1}. Comparison with theoretical studies suggests that the winds are radiation driven. For the [WC] stars the highly stratified ionization structure of metals in photospheric models lends further support to the hypothesis that the strong [WC] winds are radiation driven, too. This might explain why the supposed progenitors of the O(He) stars, the He-rich and metal poor SDO STARS, do have 'normal' mass-loss rates (only observable in the UV spectra) in contrast to the C- and O-rich [WC] stars which are found in the same part of the g–$T_{\rm eff}$ plane.

Occurrence of super-hot winds at the end of pre-white dwarf evolution

The bifurcation of post-AGB evolution into H-deficient and H-normal channels seems to continue into the WD domain, dividing the cooling sequence into non-DA and DA white dwarfs. It is natural to assume that the hottest He-rich WDs (spectral type DO) are descendants of the PG 1159 stars, however, some H-deficient post-AGB stars will become DA WDs, namely those which have retained at least some hydrogen in their envelope. The transition of post-AGB stars into the hot WD phase is encompassed by a spectacular phenomenon, which has been discovered only recently. It was found that a large fraction (50%) of the hot DO WDs shows signatures of a super-hot wind in the optical spectrum. High-ionization absorption lines of the CNO elements were identified (e.g. C VI, N VII, O VIII, and even Ne X). The high excitation potentials involved require temperatures approaching almost 10^6 K and the triangular shaped line profiles suggest their formation in a rapidly accelerating wind from the WD, reaching a terminal speed of the order of 10 000 km s^{-1}. Preliminary model calculations suggest mass-loss rates between 10^{-7} and $10^{-10} M_{\odot}$ yr^{-1}. However,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

at present these radiative equilibrium models are crude at best, as they simply assume a very high $T_{\rm eff}$ in order to produce the high-ionization stages. Since the post-AGB age is of the order 10^5–10^6 yr, consequences of the stellar wind on evolutionary times scales and chemical surface composition are possible, if the mass-loss rates are in fact as high as estimated right now.

It is emphasized that this wind phenomenon is not restricted to DO WDs. It was also detected with one WD of spectral type DAO (H rich with traces of He) with similar high-ionization features. This is the only case discovered among DA/DAO WDs so far, and the question arises of why there is such a strong preference among DOs to show this phenomenon.

Bibliograpy

Asplund M 1999 Sakurai's object—stellar evolution in real time *IAU Symp. 191: AGB Stars* ed T Le Bertre, A Lebre and C Waelkens (San Francisco, CA: Astronomical Society of the Pacific) pp 481–6

Hamann W-R 1997 Spectra of Wolf–Rayet type central stars and their analysis *IAU Symp. 180: Planetary Nebulae* ed H J Habing and H J G L M Lamers (Dordrecht: Kluwer) pp 91–6

Iben I Jr 1984 On the frequency of planetary nebula nuclei powered by helium burning and on the frequency of white dwarfs with hydrogen-deficient atmospheres *Astrophys. J.* **277** 333–54

Werner K, Dreizler S, Rauch T, Koesterke L and Heber U 1999 Born-again AGB stars: starting point of the H-deficient post-AGB evolutionary sequence? *IAU Symp. 191: AGB Stars* ed T Le Bertre, A Lebre and C Waelkens (San Francisco, CA: Astronomical Society of the Pacific) pp 493–8

Werner K and Wolff B 1999 The EUV spectrum of the unique bare stellar core H1504+65 *Astron. Astrophys.* **347** L9–L13

Wesemael F, Green R F and Liebert J 1985 Spectrophotometric and model-atmosphere analyses of the hot DO and DAO white dwarfs from the Palomar-Green survey *Astrophys. J. Suppl.* **58** 379–411

Klaus Werner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Phaethon

An Apollo asteroid discovered by Simon Green and John Davies in 1983 from data returned by the IRAS satellite; it is designated (3200) Phaethon. It follows a highly elliptical orbit and at perihelion is just 0.14 AU from the Sun, well within the orbit of Mercury. It was named after the son of the Sun-god Helios, who took the reins of the Sun-chariot and, losing control, nearly incinerated the Earth. By April 2000, only two other asteroids were known to have a smaller perihelion distance. Phaethon's aphelion distance is 2.40 AU, and its mean distance from the Sun is 1.27 AU (190 million miles); orbital period 1.43 years, inclination 22°, eccentricity 0.89 and rotation period about 4 h. The discovery of Phaethon solved the mystery of the Geminid meteor shower's 'missing' parent comet, for the orbital elements of the asteroid match those of the Geminid meteor stream.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

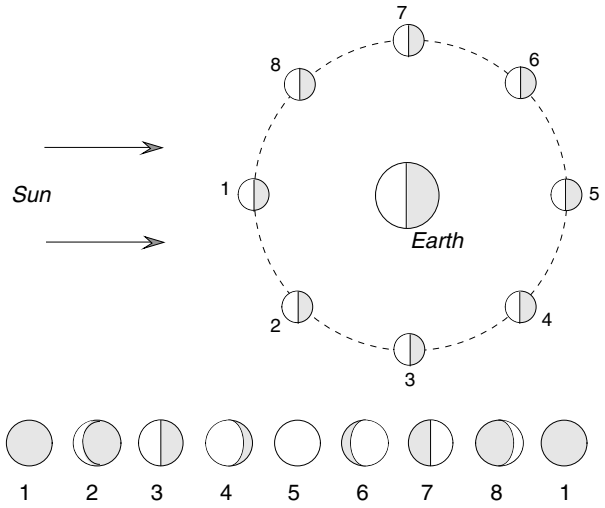


Figure 2. The Earth–Moon system as seen from space.

Phase

In the different phases of its motion, a celestial body of the solar system, as seen from the Earth, shows different apparent shapes. This comes from the relative positions of the Earth (the location of the observer), of the body (that is observed) and of the Sun (which is illuminating the body). These different successive aspects of the illuminated body that is being observed are named phases (for example: the PHASES OF THE MOON). From these first definitions, the phase angle is defined as the angle measured from the center of the illuminated observed body between the source of light (the Sun) and the observer (the Earth). It is the Sun–object–observer angle. Knowledge of the value of this angle will provide the aspect of the body, the apparent disk of which is smaller than the true apparent disk (cf figure 1).

The Moon phases

In the case of the Moon, it is easy to understand its changing appearance day after day (figure 2).

Let L_{m} and L_{s} be the longitudes of the Moon and of the Sun and φ the phase angle. The relationship between these are given in table 1 and the illumination from the Moon as a function of the phase angle is given in table 2.

Phases of the planets

Similarly to the Moon, all the bodies of the solar system, and especially the planets, are affected by the phase effect. This effect is quite different if the planet is located between the Earth and the Sun (as are Mercury and Venus) or if the planet is located outside the Earth's orbit (Mars, Jupiter, Saturn, Uranus and Neptune).

In the case of Mercury and Venus, the phases are observed just as they are for the Moon and all the aspects of the apparent disk of these planets are as for the Moon.

In the case of the other planets, the phase angle is always less than 90° (cf table 3) since these planets are never between the Earth and the Sun.

Before the use of telescopes, it was impossible to see the phases of the planets with the naked eye. This was cited as a proof to astronomers that the Earth was the center of the universe. Let us see why. If the Sun is the center of the solar system, then the distance between the Earth and Venus should have very large variations, which is not the case if the Earth is the center of the solar system. However, observation shows that VENUS keeps the same magnitude all the time, so that Venus is rotating around the Earth. Such an argument does not take into account the existence of the phase effect. GALILEO, when observing Venus through his telescope, understood easily why the magnitude of Venus was not changing much during its revolution around the Sun. The phases of Venus were compensating for the variation of distance to the Earth. Venus showed a full disk when furthest away and a thin crescent when nearest.

The phase effect

The phase effect is a tool for astronomers when observing the planets: first, the variation of the angle of incidence of the illumination of the planets by the Sun will provide information on the nature of the planets themselves. The 'flat' aspect of the Moon, independent of the phase, shows that the Moon has no atmosphere. In contrast, the relief aspect of the planet Jupiter shows that this planet has a thick atmosphere. Whether the line marking the terminator between the illuminated side and the dark side of a planet is sharp or not is an indicator of the existence of twilight due to the atmosphere on the observed planet. This shows the importance of the phase effect for astronomers. For that, astronomers have theoretical models providing the light flux received from solar system objects depending on the phase angle. Different laws of distribution of the light on the observed bodies exist including parameters depending on the characteristics of the surface of the bodies. For example, Lambert's law describes very well a planetary surface with a thick atmosphere. More recently, Hapke's law describes more precisely the granulometry of the dust and the porosity of the surface of the objects.

For example, the phase angle helps us to understand the nature of the surfaces of solar system objects: for a smooth surface, the reflectivity depends directly on the phase angle; for an irregular surface, the reflectivity changes when the phase angle becomes zero. The reflected light will increase suddenly owing to the disappearance of any shadow made by the particles of dust. This is called the opposition effect. Modeling the reflected light is very useful for the determination of the center of mass of the bodies of the solar system (in order to build dynamical models of their motion). Because of the phase effect, it is necessary to reconstruct the geometrical aspect of the planet from the apparent observed disk. Uncorrected, this can induce systematic errors in the observed astrometric positions of the body.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

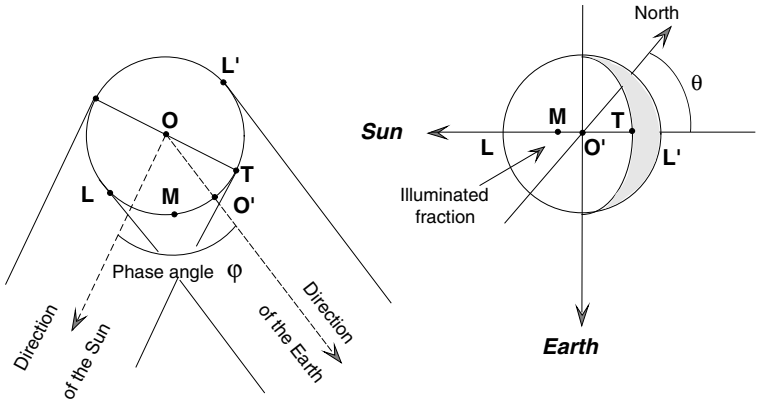


Figure 1. Phase angle and illuminated fraction of the disk. Quantities related to the phase: φ, phase angle; M, mid-point of the phase; L, L′, limbs of the body as seen from the Earth; T, terminator (limit of the illuminated part of the body on the apparent disk); OO′, direction of the Earth; θ, position angle of the phase; $K = (1 + \cos\varphi)/2$, illuminated fraction of the apparent disk (ratio of the illuminated area to the total area).

Table 1. Relationships between longitudes of Moon and Sun and phase angles.

Moon 1	New Moon or conjunction	$L_m = L_s$	$\varphi = 180^\circ$	$k = 0$
Moon 2	First crescent	$L_m - L_s < 90^\circ$	$180^\circ > \varphi > 90^\circ$	
Moon 3	First quarter or quadrature	$L_m = L_s + 90^\circ$	$\varphi = 90^\circ$	$k = 1/2$
Moon 4	Waxing gibbous Moon	$90^\circ < L_m - L_s < 180^\circ$	$90^\circ > \varphi > 0^\circ$	
Moon 5	Full Moon or opposition	$L_m = L_s + 180^\circ$	$\varphi = 0^\circ$	$k = 1$
Moon 6	Waning gibbous Moon	$180^\circ < L_m - L_s < 270^\circ$	$180^\circ > \varphi > 90^\circ$	
Moon 7	Last quarter or quadrature	$L_m = L_s + 270^\circ$	$\varphi = 90^\circ$	$k = 1/2$
Moon 8	Last crescent	$L_m - L_s > 270^\circ$	$\varphi < 90^\circ$	

Table 2. Illumination from the Moon as a function of phase angle (according to the Annuaire du Bureau des Longitudes).

Phase angle (deg)	Illumination (%) before the new Moon	after the new Moon
0 (Full Moon)	100	100
10	78.8	76.0
20	60.3	58.7
30	46.6	45.3
40	35.7	35.0
50	27.6	27.3
60	21.1	21.1
70	16.2	15.6
80	12.0	11.1
90 (quarter)	8.2	7.8
100	5.6	5.8
110	3.8	4.1
120	2.5	2.6
130	1.5	1.6
140		0.9
150		0.4

Table 3. Maximum phase angles of the outer planets.

Planet	Mars	Jupiter	Saturn	Uranus	Neptune
Maximum value of the phase angle	20°	6°	3°	2°	1°

Phases of eclipses

In the case of phenomena such as ECLIPSES, the phases correspond to the successive configurations of the phenomenon such as partial eclipse and total eclipse. The analogy between these events led to the use of the same word.

J-E Arlot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Phases of the Moon

The various aspects that the Moon presents to observers on the Earth as the proportion of its sunlit side which is visible changes in the course of its orbit around the Earth. There are four principle phases: new moon, first quarter, full moon and last quarter. One complete cycle of phases is termed a *lunation*, and is completed in just over $29\frac{1}{2}$ days, the Moon's synodic period.

At *new moon* the Moon is between the Earth and the Sun, and is invisible as it presents its night side to the observer. If the alignment is exact a total solar eclipse will occur, but because the Moon's orbital plane is inclined to the Earth's, the solar and lunar disks do not usually coincide at new moon. The *age* of the Moon is reckoned in days from one new moon to the next. After a day or two, a sliver of the sunlit side becomes visible as a thin *crescent* (under exceptional conditions a crescent may be visible when the Moon is only about 12 hours old). From new moon until full moon the Moon is said to be *waxing*, as more of the sunlit side comes into view. A week after new moon comes *first quarter*, when the Moon is a quarter of the way around its orbit. The phase at first quarter is also termed a 'half moon'—the terminator, the line marking the boundary between night and day, bisects the Moon's disk. As the Moon continues to wax, the phase becomes *gibbous*, when most of the sunlit side is visible, a crescent-shaped portion of the disk remaining in darkness. *Full moon* occurs at just under 15 days. As at new moon, Earth, Sun and Moon are again aligned, but this time the Moon is opposite the Sun in the sky; if the alignment is close enough, a total lunar eclipse occurs. After full moon the phases then repeat in reverse, the terminator continuing its passage across the disk, through the gibbous phase to *last quarter*, when the Moon enters the last quarter of its orbit at nearly 23 days old. In this half of its orbit the Moon is said to be *waning*. After passing through a crescent phase again, and the lunation ends with another new moon.

See also: phase.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Philolaus of Tarentum or Croton (*c*. 480 BC–?)

Pythagorean philosopher, wrote explanations on Pythagorean views. Thought the Earth and planets revolved around a central fire.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Phobos

Two large Soviet spacecraft launched in 1988. Intended to explore Mars and drop instruments onto the surface of its moon Phobos. Contact was lost with Phobos-1 en route to Mars. Phobos-2 failed during manoeuvres in Martian orbit. Some images and other data on the planet and Phobos were returned.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Phobos and Deimos

Asaph Hall discovered the two moons of Mars, Phobos and Deimos, in August 1877. Although other observers claimed to have discovered several other moons during this opposition, a consensus was rapidly reached about the presence of two satellites only. This discovery allowed Hall to make the first determination of the mass of MARS from the observation of the satellites' orbits with a very high accuracy, differing by less than 0.1% from the currently accepted value of 0.1074 times the mass of the Earth.

Both Phobos (figure 1) and Deimos (figure 2) have a direct circular motion in the equatorial plane of Mars, at very short distances from the planet (9400 and 23 500 km). They are very dark and small (~27 and ~15 km diameter) and are difficult to observe from the ground—particularly Phobos, which rotates rapidly and is always close to the much brighter Mars. Their shape and surface morphology are known through spacecraft observations only: both objects are very irregular with surfaces saturated by impact craters, very similar to the few ASTEROIDS observed so far, although Deimos' surface is significantly smoother. These characteristics suggest that the two satellites are asteroids formed well beyond the orbit of Mars, and captured after the accretion of the planet. The capture mechanism is unknown, however. Although their low density (~1.9) is difficult to reconcile with the most favored meteoritic analogs, it is comparable to the only C-type asteroid observed so far. The surface composition has been observed only on Phobos; its spatial variations imply that, whatever its origin, Phobos is not a primitive body but instead experienced a long and complicated history. The orbital position around Mars probably favors very specific evolution processes, in particular concerning impact cratering.

Figure 1. Phobos seen by Mars Global Surveyor on 19 August 1998. The resolution is 4 meters per pixel. Individual boulders on the rim of the crater are ejecta blocks from the impact that formed Stickney, the large crater on top. The grooves are the elongated depressions near the rim of Stickney. High-resolution MOC image PIA01333 (Malin Space Science Systems, NASA).

Figure 2. Viking Orbiter image of Deimos acquired on 5 May 1977, from 1400 km. Deimos appears smooth. Higher-resolution images taken during closer approaches show that the surface is covered with craters, but many of these have been partially buried or subdued by regolith. Deimos is about 14 km from top to bottom in this image. (Viking Orbiter 413B83, NASA.)

Discovery and observations

Several astronomers, including William Herschel, had conducted unsuccessful searches for Martian satellites, and it was generally accepted that Mars had no moon when Asaph Hall systematically looked at very close distances from the planet. He eventually discovered the two satellites in the bright halo of Mars with the 66 cm refractor at the US Naval Observatory during the very close opposition of 1877—the same year Schiaparelli drew the first detailed map of Mars. Phobos and Deimos already had a literary history, however, since two satellites of Mars are mentioned by Jonathan Swift in <it>*Gulliver's Travels* (1726) and by Voltaire in *Micromégas* (1750). Swift was probably influenced by Kepler's view that, Mercury having no satellite, Earth only one, and Jupiter four (only the Galilean satellites were known at the time), Mars should have two satellites for symmetry reasons.

The first studies of the physical properties of Phobos and Deimos were conducted by E Pickering at Harvard between 1877 and 1882; however, their detailed characteristics remained unknown for nearly a century. Detailed observations of the orbital motions permitted

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

accurate determination of Mars' oblateness and rotation axis even before the space era. Phobos was first imaged by Mariner 7 in 1969, and appeared as a small and very irregular rock about 25 km long. Such small objects cannot be resolved by ground-based telescopes or the Hubble Space telescope at the distance of Mars, and therefore most of our current knowledge derives from spacecraft observations by Mariner 9 (1971), Viking (1977) and Phobos-2 (1989) and from a few ground-based spectroscopic observations. The Soviet Phobos mission was dedicated to the largest satellite of Mars, and represented the first attempt to land on a small body of the solar system. Unfortunately the spacecraft was lost during its approach to Phobos, after two months in Martian orbit; therefore no *in situ* observations were performed although high-quality remote sensing data were acquired, mostly from the camera and the two spectrometers.

The current Martian program is focused on sample return from Mars, and does not include specific observations of the satellites. However, several projects dedicated to Phobos have been proposed. The reason for this interest is that Phobos is the most accessible asteroid-like object and would give information about both the primitive matter from which terrestrial planets were built, and the evolution processes on small bodies of the solar system.

Orbits, orbital histories and possible origins

Deimos lies at 23 500 km from Mars's center, Phobos at only 9400 km. Phobos is so close that it is located well inside the synchronous orbit of Mars, so it rises in the west and sets in the east twice a day as seen from Mars. This also means that Phobos is slowed down by the tidal forces raised inside the volume of Mars. Phobos is therefore spiraling towards Mars, and will eventually crash to the surface; the predicted collapse time is of the order of 40 million years. Conversely, this implies that the original orbit of Phobos was very different. Both orbits currently have low eccentricity and inclination, and this puts strong constraints on the satellites' origin.

Two types of hypothesis have been proposed concerning the origin of the Martian satellites: the first type states that they formed in Martian orbit, and are remnants of the planetary formation process as are the large satellites of giant planets; in the second type of hypotheses the satellites are asteroids formed further from the Sun and captured by aerodynamic drag on either the planetary nebula or a dense primitive atmosphere. As detailed below, compositional arguments favor the second hypothesis, but the capture mechanism is difficult to explain. Dynamicists generally argue that the current orbits cannot be produced following capture of outer solar system objects; in particular the orbit of Deimos is stable on very long time scales and cannot result from a more eccentric orbit under the single influence of tidal forces. Several hints have been proposed that could accommodate the capture theory, including chaotic tumbling, collisions, gas drag on a dissipating nebula, eccentricity jumps consecutive to gravitational resonances etc. It has also been proposed that the two satellites are the two parts of a single object in synchronous orbit, split apart at the time of heavy bombardment. In the framework of the capture hypothesis, Phobos and Deimos would be the last of many objects captured by the proto-Mars, which eventually crashed on the planet and contributed to its formation (see also PLANETESIMALS).

Morphology and bulk properties

Both satellites are very irregular, as expected for such small bodies. To a first approximation they are elongated ellipsoids with dimensions given in table 1. Like most satellites in the solar system they are in synchronous rotation, with their longest axis turned toward the planet and the shortest one perpendicular to the orbital plane. Large departures from this ellipsoidal form exist however, and precise estimates of the density rely on accurate topographic models of the satellites. The currently accepted value (~1.9) is surprisingly low; it can be due to the material itself, to large porosity in a thick regolith, or to the existence of large voids inside the volume.

Phobos' shape is dominated by three large craters: Stickney (10 km diameter), Hall and Roche (5 km each). No secondary craters are identified and no central peak crater is observed, which is consistent with the small size of the satellite. On Deimos the largest identified crater is only 3 km across, and the density of large craters (>200 m) is at most half that of Phobos. However, a large saddle-like depression 10 km across, i.e. relatively much larger than Stickney on Phobos, could also result from an impact. Both surfaces are saturated with small craters and are thought to be at least 2 billion years old. All measurements demonstrate the existence of a regolith up to 200 m thick on Phobos, a consequence of the impacts. On Deimos the regolith is apparently five to ten times thinner, but it fills most of the craters.

A peculiarity of Phobos, which was long thought unique in the solar system, is the presence of several systems of parallel grooves with directions related to the main axes of the satellite. The grooves are typically 30 m

Table 1. *Satellite data.*

	Phobos	*Deimos*
Semi-major axis (km)	9378.5 (2.76 R_M)	23 459 (6.92 R_M)
Period	7 h 39 min	30 h 18 min
Eccentricity	0.015 15±0.000 04	0.000 196±0.000 034
Inclination with respect to Laplace plane	1.02°	1.82°
Orbital velocity (km s^{-1})	2.14	1.36
Mass (kg)	$(1.08 \pm 0.01) \times 10^{16}$	1.8×10^{15}
Axes (radii, km)	13.4×11.1×9.3 ± 0.3	7.5×6.2× 5.4±0.5
Volume (km^3)	5680±250	1000±300
Density (g cm^{-3})	1.90±0.10	1.70±0.4
Escape velocity (m s^{-1})	12	6
Geometric albedo	0.071±0.012	0.068±0.007

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

deep, 100–200 m wide and up to 20 km long. In many areas, grooves are alignments of coalescent pits that could result from outgassing mechanisms or more probably from regolith drainage into large cracks in the volume of Phobos. Most of them are centered on Stickney and are best developed in its vicinity. The main impact suffered by the satellite, which is close to the maximum that a body of this size can tolerate without disruption, is sufficient to cause these cracks to form, but tidal effects could have later played a role in the evolution of the grooves. Deimos has a much more rounded shape with large regular facets separated by bright ridges; other bright patches are visible on crater rims. The whole surface has a distinctly smoother appearance that is due to downslope movement of surface materials. This difference in morphology is unexpected, given the similarity in bulk properties of the two satellites.

The first high-resolution observations of asteroids by the Galileo and NEAR spacecraft have shown that Phobos is actually the most 'normal' body, while Deimos is markedly different. Gaspra and Ida, which are S-type asteroids with a much larger density, are the only other known solar system bodies with grooves, although theirs are less developed than those on Phobos. Mathilde, currently the only C-type asteroid observed, is very much like Phobos: its overall morphology, its density (1.5–2.0 g cm^{-3}) and its spectral properties are similar; conversely, it exhibits at least five impact craters with radius half its size, and no grooves.

Evolution mechanisms

Internal evolution of small bodies stops very early in their history because their volume-to-surface ratio is too small to allow significant heating. In these conditions the most efficient evolution process is meteoritic cratering, which is responsible for the shape and relief but also forms a regolith by accumulation of debris and can even modify the surface composition. Other significant processes include space weathering (formation of dark glasses and modification of surface chemistry by interaction with the solar wind and cosmic particles) and possibly tidal effects.

Crater formation (see CRATERS) is in general in the strength-scaling regime, where ejecta leave the immediate vicinity of the crater and well-defined rims do not form. On Phobos, only craters more than 5 km large have a significant ejecta blanket; this suggests that the bulk material has a compressive strength similar to anhydrous carbonaceous chondrites, while the regolith has properties similar to that of the Moon. In these conditions, materials can be excavated from one or two kilometers deep; due to the small gravity, ejecta are also transported very far from the impact. However, observations from the Phobos-2 spacecraft showed significant spectral variability on a kilometer scale, related in particular to ejecta blankets and deep craters. Therefore, lateral and vertical mixing are not efficient enough to homogenize the whole surface.

Cratering processes are deeply affected by the orbital situation around Mars. Because of the strong asymmetry of the effective gravitational field, ejecta suffer a different fate depending on the longitude of impact and on the speed and direction of ejection; this mechanism is clearly seen on Stickney's ejecta, which are more confined in the sub-Mars direction. In addition, the average escape velocities are much smaller than the orbital velocities (~8 m s^{-1} versus ~2 km s^{-1}), therefore most ejecta escaping the satellites remain on the same orbit around Mars and are reaccreted by low-velocity impacts in a very short time (~10^4 yr), which strongly increases the meteorite flux and regolith thickness. Dust belts may have formed along the orbits of Phobos and Deimos when impacts were very frequent; these dust belts were never observed, however, and the lifetime of such objects seems rather short. Over long periods of times, surface materials can be ejected and reaccreted over and over

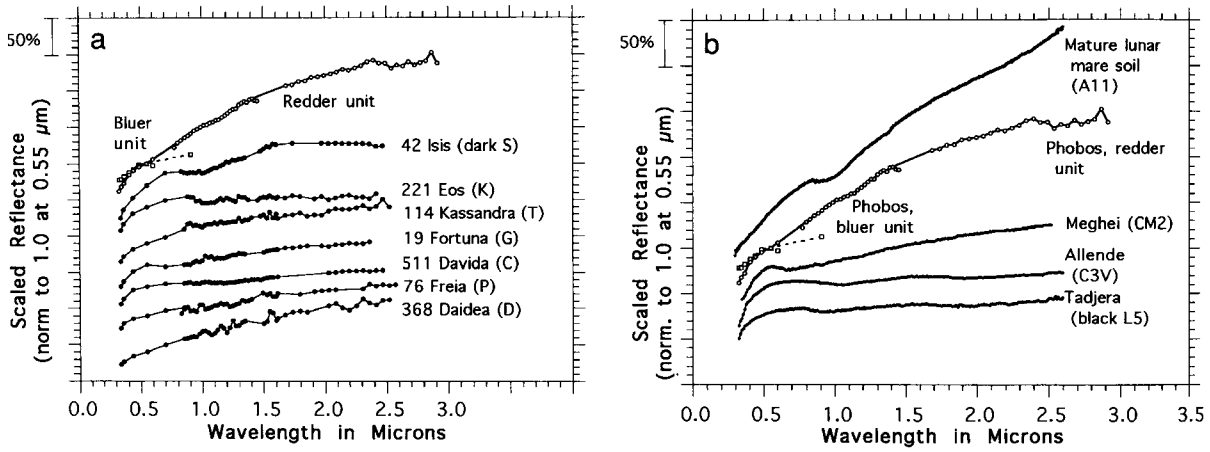


Figure 3. Comparison of the spectra of Phobos and various dark asteroids and meteorites. (From Murchie and Erard 1996 *Icarus* **123** 63–86.)

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

again. Shock heating would then dehydrate the surface materials and could profoundly modify the composition and spectral properties of the regolith. The regolith itself can insulate the interior, which could retain large quantities of volatiles.

The morphological peculiarities of Deimos are currently best explained by the effect of a large impact corresponding to the saddle-like scar mentioned above; seismic shaking and reaccumulation of ~200 m of debris would have concealed previous craters and possible grooves formed during the impact, thus explaining the lower density of large craters, the thinner regolith in the craters, and global downslope movements. The location in the Martian orbit permits isotropic reaccumulation of large quantities of ejecta, which is not possible for an asteroid in solar orbit, and could contribute to the particularly smooth appearance of Deimos.

Composition and internal structure

Composition is inferred from spectral observations, with a strong constraint from the density. The observed properties (figure 3) strongly suggest that the satellites are captured asteroids but do not strictly preclude formation in the inner solar system, although independent from Mars' formation. The two satellites are among the darkest bodies in the solar system, with geometric albedo half that of the Moon. Variations in brightness and photometric functions are observed, but are usually ascribed to variations in particle size rather than composition.

The Martian satellites have often been presented as homogeneous mixtures of materials similar to hydrated carbonaceous chondrites, inferred to be the primary component of C-type asteroids. This interpretation was based on the low albedo, on limited spectral coverage by Mariner 9 and Viking, and on the low density. More recent observations contradict this view, however. The major spectral properties observed by Phobos-2 are a very red continuum slope, no hydration signature, a weak and variable pyroxene/olivine absorption, and a strong fall-off in reflectance below 0.5 μm; the Hubble Space Telescope has provided consistent results for both satellites in the visible range. These properties do not fit any known meteorite spectrum; the best match is given by some dark ordinary chondrites (L and LL types), but is not incompatible with dry carbonaceous chondrites (CO and CV types). However, only the mature lunar mare soils exhibit similarly steep spectral slopes, which demonstrates that exposure to the space environment has modified the regolith extensively. Consistently, Stickney's ejecta have flatter spectra, and either are less altered or consist of a different material excavated by the impact. As for asteroids, only the driest C-type have low hydration bands like Phobos and Deimos, but the best spectral matches are given by P, D and T types. T-type asteroids are poorly understood, but there appears to be a gradual evolution in spectral properties from C to P and D. Most P and D objects are thought to be primitive, rich in organic materials and possibly in anhydrous silicates, and are located in the outer asteroid belt (2.8 to 3.8 AU); interestingly, these dry asteroids are often suspected to contain large amounts of water ice preserved since the formation of the solar system.

The low density of the Martian satellites is in fact inconsistent with that of any low-albedo METEORITES. The density of the best spectral analogs is so much higher (3.4–3.8 for black ordinary chondrites, 3.3–3.6 for dry carbonaceous chondrites) that regolith porosity alone cannot explain the discrepancy. The presence of large quantities of ice in the interior would resolve this paradox. On Phobos, an ice core with radius 7–9.5 km is required, depending on the density of the upper layer; the latter would mostly consist of space-weathered primitive material, but could also have accreted Martian material ejected during the formation of impact basins.

Another possibility often suggested to explain the low density and surface heterogeneity of Phobos is that of the 'rubble pile' of fragments. Such objects are thought to result from disruption of larger parent bodies, and to have large internal porosity and therefore low density. However, impact modeling of Stickney suggests that Phobos' grooves could have formed only in a coherent body; in addition, the satellite was apparently not fragmented in the process and the resulting cracks are not large enough to explain the unusual low density. Furthermore, a rubble pile is expected to exhibit only a small departure from sphericity, which is hardly compatible with Phobos' shape. The rheological properties of an ice core at low temperature would probably permit the formation of grooves during a large impact. Considering the absence of grooves and the more spherical shape, the rubble pile hypothesis cannot be rejected for Deimos in the framework of the large impact theory mentioned above; it is also favored for Mathilde, in spite of its close similarities with Phobos.

Conclusion

The last point above is representative of the uncertainty that persists concerning the origin and evolution of the Martian satellites: although it is more and more widely accepted that they are captured bodies formed in the outer asteroid belt, their structure and evolution are still poorly understood. The more attractive assumption is perhaps that they are D or T-type asteroids made of relatively primitive material accreted on a water ice core, with surfaces extensively modified by exposure to the space environment; specific processes related to the orbital position and the unusual proximity of the Sun have probably resulted in surface characteristics that are uncommon among asteroids of similar origin.

However, numerous intriguing questions remain, including: the origin of the bodies (do they represent primitive objects still containing large amounts of ice formed in the solar nebula; from which part of the asteroid belt do they come?), the regolith composition (are they made of a material similar to carbonaceous chondrites or a more evolved one; is there a significant

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

amount of Martian material accreted on the surfaces?), the internal structure, the nature and efficiency of specific evolution processes, the origin of the difference between Phobos and Deimos, and so on. The possibility of capturing distant objects around Mars is actually still problematic.

A better understanding will require further observations of the Martian satellites and more asteroids. Planetary missions now systematically consider asteroid flybys during the cruise phase (e.g. Cassini/Huygens, Rosetta), so our knowledge of these objects is rapidly improving. In the long term, sample return from Phobos may prove the only way to address open issues concerning its origin. In the meantime, global imaging spectroscopy coverage in the near-infrared would provide important information about its composition; such observations could be performed from a spacecraft dedicated to the exploration of Mars.

Bibliography

Sheehan W 1996 *The Planet Mars: A History of Observation and Discovery* (Tucson, AZ: University of Arizona Press) [discovery of the satellites]

Burns J 1993 Contradictory clues as to the origin of the Martian moons (in *Mars* ed Kieffer *et al* (Tucson, AZ: University of Arizona Press) [orbital and dynamic review]

Thomas P, Veverka J, Bell J, Lunine J and Cruikshank D 1993 Satellites of Mars: geologic history (in *Mars* ed Kieffer *et al* (Tucson, AZ: University of Arizona Press) [composition and geology]

Stéphane Erard

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Phoenix

(the Phoenix; abbrev. Phe, gen. Phoenicis; area 469 sq. deg.) A southern constellation which lies between Sculptor and Tucana, and culminates at midnight in early October. It was probably first shown on Petrus Plancius' celestial globe of *c*. 1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A rather inconspicuous constellation, the brightest stars in Phoenix are α Phoenicis (Ankaa or Nair al Zaurak), magnitude 2.4, β Phoenicis (combined magnitude 3.3), a very close binary with a yellow (G8) primary, magnitude 4.0, and a slightly fainter secondary, magnitude 4.2, separation 0.7″, and γ Phoenicis, magnitude 3.4. There are four other stars brighter than fourth magnitude, including ζ Phoenicis, a multiple system with bluish-white/white (B6 and A7) components, magnitudes 3.9v and 8.0, separation 6.5″, the former of which is an eclipsing binary (range 3.9–4.4, period 1.67 days). Other interesting objects include NGC 625, an eleventh-magnitude galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Photodissociation Regions

Photodissociation regions, or photon-dominated regions (PDRs), are dense ($\sim 10^3$–10^7 cm^{-3}) regions in the interstellar medium (ISM) where the hydrogen gas is neutral, and is either atomic or molecular (see H I and H II regions), and where externally incident far-ultraviolet (FUV: 6–13.6 eV) radiation controls the chemical structure and dominates the thermal properties of the gas. This definition excludes the diffuse and translucent clouds which are less dense and generally more transparent to FUV radiation, although some authors refer to the latter objects as 'low-density' PDRs. PDRs are ubiquitous in the ISM and are present in the surface layers of dense molecular clouds, reflection nebulae, planetary nebulae, neutral stellar winds and circumstellar disks.

PDRs are widespread sources of optical, infrared and millimeter atomic and molecular line emission, as well as thermal infrared dust continuum emission. PDR emission is particularly bright in gas exposed to enhanced FUV radiation fields, such as exist in massive-star-forming regions in the Milky Way. PDRs are prominent in starburst galaxy nuclei and active galactic nuclei.

Famous and well-studied Galactic PDRs are located in the A and B clouds and Bright-Bar region in the Orion molecular cloud complex, the star-forming regions M17 and S140, molecular clouds in the Galactic center, the reflection nebulae NGC 2023, NGC 7023 and IC 63, and the planetary nebulae NGC 7027 and BD+30°3639. PDR emission has been observed in numerous galaxies including the well-studied starburst galaxies M82 and NGC 253.

Chemical structure

The chemical processes which occur in PDRs are similar to those which occur in the diffuse and translucent clouds (see INTERSTELLAR MOLECULAR CLOUDS). Molecules are destroyed by photodissociation (XY+$h\nu$ $\rightarrow$ X + Y), and atoms with ionization potentials less than that of atomic hydrogen (13.6 eV) are photoionized (X+$h\nu$ $\rightarrow$ X + e) by the FUV radiation. Molecules form via sequences of two-body gas-phase processes including radiative association (X+Y $\rightarrow$ XY + $h\nu$), ion–molecule exchange (X^+ + YZ $\rightarrow$ XY^+ + Z), charge-transfer (X^+ + YZ $\rightarrow$ X + YZ^+) and neutral-neutral (X+YZ $\rightarrow$ XY + Z) chemical reactions. Molecular ions are removed by dissociative recombination (XY^+ + e $\rightarrow$ X + Y). Endothermic neutral–neutral reactions play a more important role in PDRs compared with diffuse or translucent clouds owing to the generally higher gas temperatures that prevail in PDRs.

Formation of molecules on grain surfaces and their ejection into the gas phase may also play an important role, particularly in the production of hydrogenated species. Molecular hydrogen (H_2) is not formed efficiently by two-body reactions, and catalytic formation on grain surfaces dominates its production in PDRs (and elsewhere in the ISM).

The typical structure of a uniform-density (model) PDR is illustrated schematically in figure 1. This particular model is for a cloud with a total hydrogen density equal to 10^4 cm^{-3} exposed to an incident FUV field which is 2×10^3 times more intense than the mean interstellar FUV field. The relative abundances of H, H_2, C^+, C, CO, S^+, S, SO and SO_2 are shown as functions of increasing visual extinction, A_V, into the cloud.

A thin ionization front (IF) lies at the outer boundary of the PDR. Lyman continuum photons are absorbed in the IF, and longer-wavelength FUV radiation penetrates into the PDR. At the outer edge of the PDR molecules (in particular H_2 and CO) are rapidly photodissociated, and atomic carbon is photoionized. The incident FUV radiation is attenuated by a combination of dust grain continuum absorption and molecular line absorptions. The gas becomes progressively more molecular as the destructive effects of the FUV field diminish with increasing cloud depth.

Figure 1 illustrates some important chemical layers which occur in PDRs. The first is the H–H_2 'dissociation front' across which the hydrogen gas becomes molecular. H_2 is photodissociated in a two-step process. FUV photons are first absorbed in electronic excitations in lines of the Lyman and Werner band systems. This is then followed by spontaneous radiative decays to the ground-state continuum, which lead to dissociation. The dominant H_2 absorption lines generally become very optically thick in PDRs, and once this occurs the hydrogen molecules are said to 'self-shield' against the destructive FUV radiation.

A second important transition is the C^+–C–CO layer across which the gas-phase carbon is incorporated into CO molecules. Atomic carbon becomes abundant in this transition layer. At large cloud depths the excess oxygen either remains atomic or is incoporated into O_2. A third transition illustrated in figure 1 is the S^+–S–SO,SO_2 layer across which sulfur is incorporated into molecules.

The locations of the various transition layers depend primarily on the hydrogen particle density and the intensity of the incident FUV field. Other cloud parameters include the elemental abundances, the grain absorption and scattering properties, and the cloud geometry.

The observations are generally consistent with the structure illustrated in figure 1. However, in inhomogeneous clumpy clouds the behavior can become more complicated, owing to enhanced penetration and internal scattering of the FUV radiation and the presence of unresolved dense clumps within the observational beams. For example, photodissociation of CO molecules by scattered radiation in clumpy media probably accounts for the extended atomic carbon emission observed in star-forming molecular clouds.

The ionization fraction is an important parameter of interstellar clouds. Near the outer edges of PDRs the positive charge is carried mainly by C^+ ions, and beyond the C^+–C–CO transition photoionization of silicon and sulfur provides the free electrons. In clouds which are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

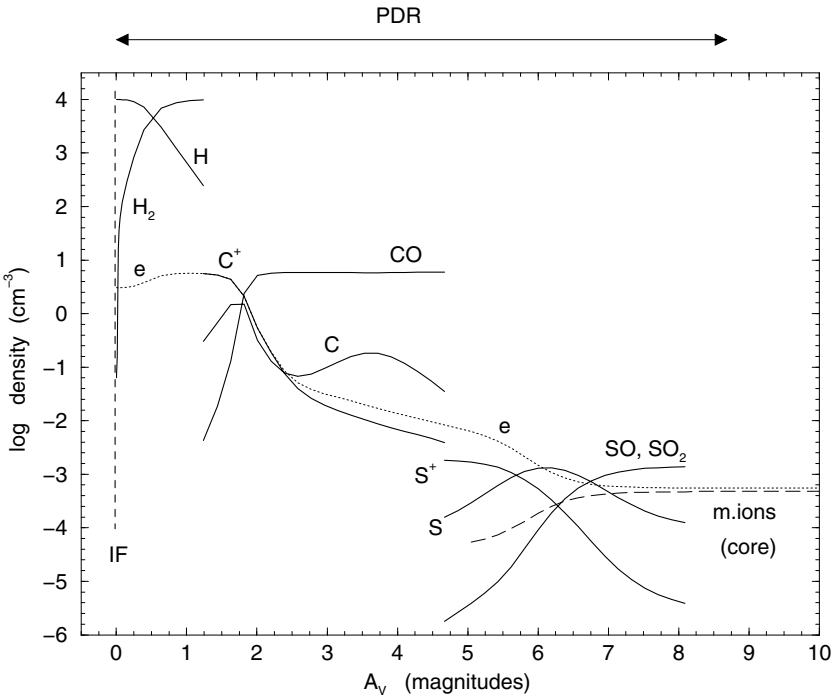


Figure 1. Chemical structure of a PDR, illustrating the H–H_2 dissociation front and the C^+–C–CO and S^+–S–SO,SO_2 transition layers. The ionization front (IF) marks the outer boundary of the PDR.

sufficiently thick to absorb all of the incident FUV radiation an inner boundary to the PDR exists beyond which other processes (e.g. cosmic-ray ionization) dominate the thermal and chemical properties of the gas. In the interior shielded regions the positive charge is carried mainly by abundant molecular ions (e.g. H_3^+ and HCO^+) and metal ions.

PDR emission lines

The gas temerature in PDRs is fixed by the balance between FUV heating and atomic and molecular emission line cooling. Observations and theoretical computations show that the gas temperatures in the outer parts of PDRs range from a few 100 K in lower density clouds to more than 1000 K in very dense clouds. The gas temperature decreases with increasing cloud depth. Generally the gas temperature is significantly higher than the dust temperature in PDRs.

The FUV radiation energy is effectively transferred to the gas via photoelectric emission from small dust grains and by the photoionization of 'large molecules' or polycyclic aromatic hydrocarbons (PAHs). Evidence for the presence of PAHs in PDRs is provided by observations of broad emission features at 3.3, 6.2, 7.7, 8.6 and 11.3 μm, although the exact identifications of these features remain uncertain. Dense PDRs are also heated by collisional de-excitation of H_2 molecules which have been vibrationally excited by the FUV-pumping process (see below).

PDRs are cooled by atomic and ionic fine-structure emission lines and molecular rotational transitions. Important fine-structure emission lines include the [C II] 158 μm, [O I] 63 and 146 μm, [Si II] 35 μm, and [C I] 370 and 609 μm transitions. Their relative intensities serve as diagnostics of the conditions in PDRs. The [C II] and [O I] emission lines are particularly useful diagnostics, and can emit up to $\sim$ 1% of the absorbed FUV luminosity (most of the FUV energy is reradiated as thermal far-infrared dust emission).

Figure 2 shows model computations of the [C II] 158 μm and [O I] 63 μm emission line strengths as functions of the incident FUV intensities for a range of assumed cloud densities. A selection of observed line strengths from several astronomical sources is also displayed in figure 2. The emission line intensities saturate at large values of the FUV intensity owing to the effects of dust absorption. The critical densities of the [C II] 158 μm and [O I] 63 μm transitions (for collisions with neutral hydrogen) are 3×10^3 and 3×10^5 cm^{-3} respectively, so that the [O I]/[C II] emission line ratio becomes large in high-density PDRs. The [C II] line generally remains optically thin and can be used as a tracer of the cloud mass. The [O I] line becomes optically thick in PDRs.

PDRs are prominent sources of H_2 and CO emission lines. Because H_2 is a light molecule with a small moment of inertia its rotational energy levels are widely spaced, and elevated temperatures are required for efficient excitation. This occurs in PDRs, and mid-infrared H_2 emission lines, such as the fundamental $J = 2$–0 28.2 μm line (and other transitions including the 17.0, 12.3, 9.7 and 8.0 μm lines) are readily excited and have have been observed (from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

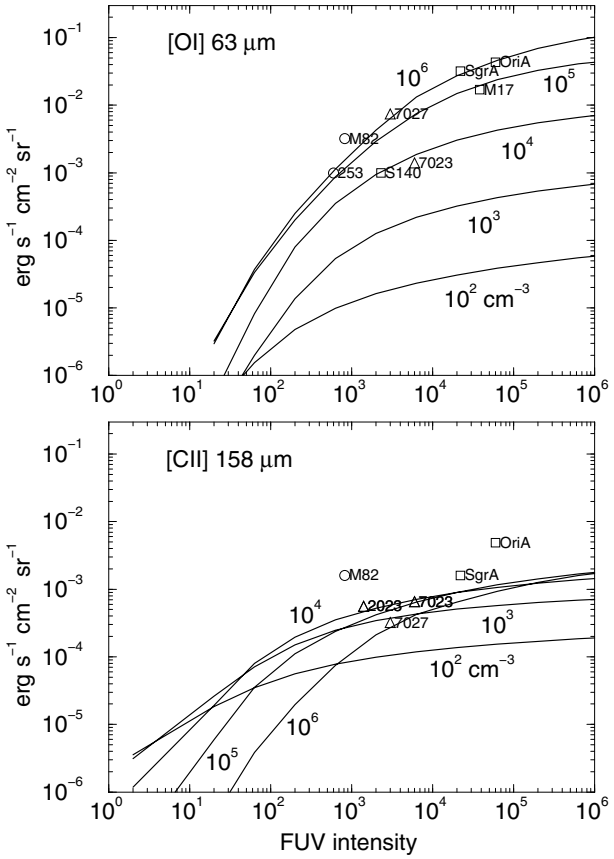


Figure 2. [C II] 158 μm and [O I] 63 μm fine-structure emission line intensities as functions of the incident FUV intensity for hydrogen particle densities in the range 10^2–10^7 cm^{-3}. Square symbols indicate PDRs adjacent to H II regions, triangles are reflection and planetary nebulae and circles are starburst galaxy nuclei.

space) in several Galactic PDRs. Mid-IR H_2 emissions have also been observed in several STARBURST GALAXIES, and may arise in PDRs. Near-infrared vibrational H_2 emissions, including the commonly observed $v = 1$–0 S(1) 2.12 μm line can also be collisionally excited in hot ($T > 1000$ K) and dense ($>10^4$ cm^{-3}) PDRs. The relative intensities of the collisionally excited H_2 emission lines are useful probes of the gas temperatures in the outer atomic hydrogen gas layer and the H–H_2 dissociation front.

H_2 is also excited directly by the FUV photons in PDRs via an important process called FUV pumping in which the Lyman and Werner band FUV line absorptions are followed by decays to discrete excited vibrational levels of the ground electronic state. The transitions to the excited vibrational levels initiate a cascade of slow radiative transitions at optical and near-infrared wavelengths. Figure 3 displays a model computation of the FUV-pumped fluorescent vibrational emission line spectrum produced in a PDR with a hydrogen density equal to 10^3 cm^{-3} exposed to an FUV field with an intensity of 10^3. Fluorescent H_2 emissions have been observed in

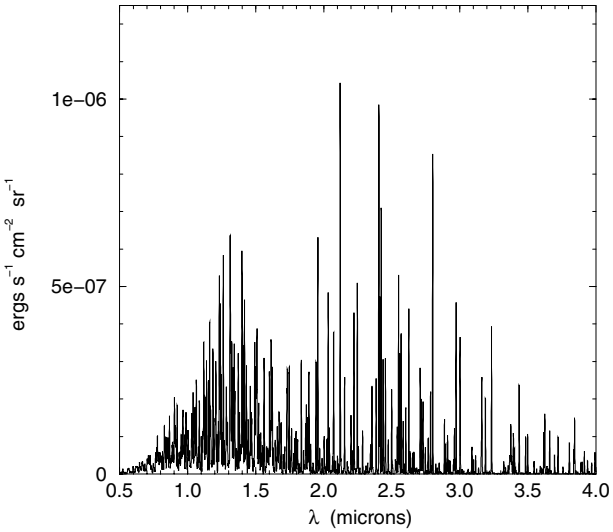


Figure 3. The FUV-pumped fluorescent H_2 emission line spectrum in the range 0.5–4 μm.

numerous PDRs (e.g. the REFLECTION NEBULA NGC 2023 and the PLANETARY NEBULA NGC 7027). The fluorescent emission line intensities are useful diagnostics of the hydrogen gas density, the FUV field intensity, the grain-surface H_2 formation efficiency, the dust grain scattering properties and the PDR lifetimes.

The intensity ratio of the 2.12 μm 1–0 S(1) and the 2.25 μm 2–1 S(1) lines may be used to discriminate between collisional and fluorescent H_2 excitation in PDRs. The 1–0/2–1 line intensity ratio is small ($\sim$2) when FUV pumping dominates the excitation and becomes large ($>$10) when collisions dominate the excitation and the level distributions approach local thermodynamic equilibrium. For example, observations of H_2 line emissions from the planetary nebula BD+30°3639 show that the 1–0/2–1 line ratio is large in the hot outer part of the PDR which is closest to the central star. Collisional excitation becomes less effective with increasing distance from the star, and the FUV pumping by the stellar photons then dominates the molecular excitation, and the 1–0/2–1 intensity ratio becomes small.

PDRs are prominent sources of CO rotational emission lines, including the fundamental millimeter wave CO $J = 1$–0 rotational transition at 115 GHz which is the standard tracer of interstellar molecular clouds. This important emission line generally forms in the inner cooler parts of the PDRs. PDRs are also sources of submillimeter and far-infrared mid-J (e.g. CO $J = 7$–6) and high-J (e.g. CO $J = 14$–13) emission lines. The high-J CO emissions trace warm and high-density gas, which is usually confined to isolated clumps within molecular clouds. These high-pressure clumps may be self-gravitating and collapsing objects.

Despite the destructive effects of the FUV radiation fields some molecules are formed preferentially in PDRs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Important examples are the radicals OH and CN and the molecular ion CO^+. In warm PDRs OH and CO^+ are produced by the endothermic hydrogen abstraction reaction $O+H_2 \rightarrow OH + H$, followed by the ion–molecule exchange $OH + C^+ \rightarrow CO^+ + H$. This simple sequence may explain the large abundances of CO^+ ions that have been observed in the planetary nebula NGC 7027 and molecular cloud interfaces in M17 and the Orion Bar. Another interesting molecular diagnostic of PDRs is the CN/HCN abundance ratio which, because of the photodissociation of HCN, is expected to be larger in PDRs compared with shielded regions. Indeed, observations of HCN and CN molecules in the reflection nebulae NGC 2023 and NGC 7023 show that the CN/HCN density ratios are enhanced near the sources of FUV radiation and decrease by large factors with increasing distance from the illuminating stars.

FUV regulated star formation

The ionization fraction in PDRs (see above) may control the rate of low-mass star-formation in the Galaxy. In this scenario the STAR-FORMATION rate is fixed by the rate at which magnetic fields are removed from the clouds via ambipolar diffusion. The diffusion rates are inversely proportional to the ionization fractions which are fixed by the visual extinctions through the clouds. A feedback loop becomes possible in which energy injection by newly formed stars leads to cloud expansion, greater penetration of the background interstellar radiation and therefore larger ionization fractions, so that the star-formation rate decreases. As the energy injection rates decrease the clouds collapse, become better shielded against the FUV radiation and become less ionized, thereby leading to an increased star-formation rate. Theoretical computations suggest that an equilibrium occurs for clouds with total visual extinctions $A_V \sim 7.5$, generally consistent with the typical observed opacities of molecular clouds.

Bibliography

Hollenbach D J and Tielens A G G M 1999 Photodissociation regions (PDRs) in the interstellar medium of galaxies *Rev. Mod. Phys.* **71** 173

Hollenbach D J and Tielens A G G M 1997 Dense photodissociation regions (PDRs) *Ann. Rev. Astron. Astrophys.* **35** 179–215

Sternberg A 1998 Photon-dominated regions *The Molecular Astrophysics of Stars and Galaxies* ed T W Hartquist and D A Williams (Oxford: Clarendon) pp 201–19

Amiel Sternberg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Photon

An elementary particle that is equivalent to a discrete packet, or 'quantum', of electromagnetic energy.

In accordance with the wave–particle duality, light, or any other form of electromagnetic radiation, behaves in some respect like a wave motion and in others as a stream of photons. The energy (E) of a photon is directly proportional to the frequency (f), and inversely proportional to the wavelength (λ), of the radiation. It is given by $E = hf = hc/\lambda$, where c denotes the speed of light and h is the Planck constant. Gamma-ray photons, which correspond to the shortest wavelengths of electromagnetic radiation, have the highest energies, and radio photons, which correspond to the longest wavelengths, have the lowest.

The photon is the force-carrying particle (gauge boson) that conveys the electromagnetic force between charged particles. It has zero mass, zero charge, and spin = 1 (in units of $h/2\pi$).

See also: electromagnetic radiation, electromagnetic spectrum, elementary particles, fundamental forces, gauge bosons, light, quantum mechanics, wave–particle duality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Physics of Molecules

Many varieties of molecule have been detected in the Milky Way and in other galaxies. The processes by which these molecules are formed and destroyed are now broadly understood (see INTERSTELLAR CHEMISTRY). These molecules are important components of galaxies in two ways. Firstly, radiation emitted by molecules enables us to trace the presence of diffuse gas, to infer its physical properties and to estimate the propensity for further STAR FORMATION in that galaxy. Secondly, the emitted radiation removes energy from the emitting regions, allowing them to cool. This cooling is an essential part of the process of the gravitational collapse of gas clouds in the initial stages of star formation.

Molecules are found in many environments in space, but most readily in the denser, cooler parts of the Galaxy and other galaxies. Molecules were discovered first in interstellar clouds and are also abundant in cool circumstellar envelopes around evolved stars. They have also been detected in planetary nebulae, in the ejecta of novae and supernovae, and in the atmospheres of relatively cool stars. Molecules have even been detected in the Sun: SUNSPOTS contain water molecules (admittedly, very short-lived). In general, molecules are likely to be present in astronomical situations where the gas temperatures are less than about 4000 K and the gas number densities are above about 10^6 H atoms/m^3 (assuming that other elements are also present with the usual cosmic abundances).

Molecules were even important in the early universe where hydrogen molecules were an important cooling agent for pre-galactic gas clouds, allowing them to collapse under gravity to form PROTOGALAXIES.

Physical processes

Physical processes are here defined to be those in which the molecules interact with their environment but remain fundamentally unchanged. In a chemical process, by contrast, an exchange of atoms occurs, and new species are created.

The main physical processes to which astronomical molecules are subjected are the following:

- absorption and emission of electromagnetic radiation;
- collisions with fast particles (COSMIC RAYS) and with other atoms and molecules on the gas;
- sticking to and desorption from the surfaces of dust grains.

Since these processes generally involve the excitation of energy levels in molecules, we shall first give a brief summary of molecular energy level structure, emphasizing in particular how it differs from that of atoms.

Energy level structure in molecules

Molecules have degrees of freedom unavailable to atoms: molecules may vibrate and rotate. Of course, molecules may also be excited electronically. Molecular electronic excitation energies are different from those of the individual atoms in the molecule. However, in the limit of large separations between the atoms (when the interaction energy between them becomes small) the molecular electronic energy levels approach those of the atoms. Hence, molecular electronic energy levels are normally of the same order of magnitude as atomic energy levels, and are separated typically by energies of a few electron volts. Energies associated with vibrational transitions are generally one to two orders of magnitude smaller than this, and transitions between rotational levels have energies that are generally again much smaller than those of vibrational transitions.

The essential features of molecular vibration and rotation are evident in the smallest molecules, i.e. diatomics, so the discussion here is introduced in terms of diatomic molecules. In the simplest approximation, the vibrations of diatomic molecules can be treated as simple harmonic oscillations. Classically, a harmonic oscillator may have any energy, but for molecules quantum mechanics restricts the permitted energies, E_v, to the discrete values

$$E_v = \hbar\omega(v + \tfrac{1}{2})$$

where $\hbar$ is the reduced Planck's constant, ω is an angular frequency measured in radians per second, related to the force constant in the oscillator, and v is the vibrational quantum number that may take integer values or zero. The energies permitted by quantum mechanics for the harmonic oscillator are, therefore, $\frac{1}{2}\hbar\omega, \frac{3}{2}\hbar\omega, \frac{5}{2}\hbar\omega, \ldots$. There is no restriction on transitions between these permitted energy levels, in this approximation.

The rotation of a molecule can be considered in a simple approximation to be that of a rigid rotator. Quantum mechanics confines the permitted energy levels of a molecular rigid rotator, E_r to the set of discrete values

$$E_r = B_r J(J+1)$$

where B_r is related to the moment of inertia, I, of the molecule

$$B_r = \hbar^2/2I$$

and J is the rotational quantum number and is confined to integer values or zero. Thus, the permitted energy levels are $0, 2B_r, 6B_r, 12B_r, \ldots$ in this approximation and transitions between the levels are restricted by selection rules. Hence, massive molecules with large moments of inertia have relatively closely spaced energy levels, and the associated wavelengths of absorption or emission are relatively long, and vice versa.

The total energy of a diatomic molecule is in the simplest approximation the sum of electronic, vibrational and rotational energies:

$$E_{\rm tot}(n, v, J) = E_e(n) + E_v(v) + E_r(J)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where E_e is the electronic energy which depends on one or several quantum numbers, n. In general, transitions may occur between states of different n, v, and J, subject, of course, to any selection rules. For example, interstellar molecular hydrogen was first detected by absorption in its electronic transition from the ground electronic state, $v = 0$ vibrational level, and $J = 0$ and 1 rotational levels to the first accessible excited electronic state in a variety of vibrational and rotational levels.

Vibration–rotation transitions may also be important. For example, molecular hydrogen is readily detected in warm astronomical regions by its emission at wavelengths near 2 μm arising from transitions in level $v = 1$ to level $v = 0$ in the ground electronic state. In both v levels, several J levels are involved. Finally, pure rotation transitions may occur, and are the means by which most interstellar molecules have been detected. For example, carbon monoxide, CO, is an important tracer of cold gas through its emission from $J = 1$ to $J = 0$, occurring at a wavelength of 2.6 mm. The energy difference corresponding to this is equivalent to about 5 K$\times k$ (k is Boltzmann's constant), and so the $J = 1$ level is readily excited in collisions in gas at a temperature as low as about 10 K (which is often the case for interstellar CO).

Of course, many cosmic molecules are much more complex than diatomics. For polyatomics, the vibrational spectra can be regarded as a superposition of harmonic oscillations for each of the normal modes of the system, so the diatomic description can be very easily extended. However, polyatomics have a considerably more complex rotational energy level structure than diatomics. When there is an axis of symmetry in the molecule (a so-called symmetric top) then one more rotational quantum number, K, in addition to J, is required to define the energy state and a simple extension can be given:

$$E_r = B_r J(J+1) + (A_r - B_r)K^2$$

where A_r is related to the moment of inertia about an axis which is not the axis of symmetry. Figure 1 illustrates the rotational energy level structure that arises for a molecule of this type. However, when the molecule does not have an axis of symmetry (an asymmetric rotator) then no simple formula can be given.

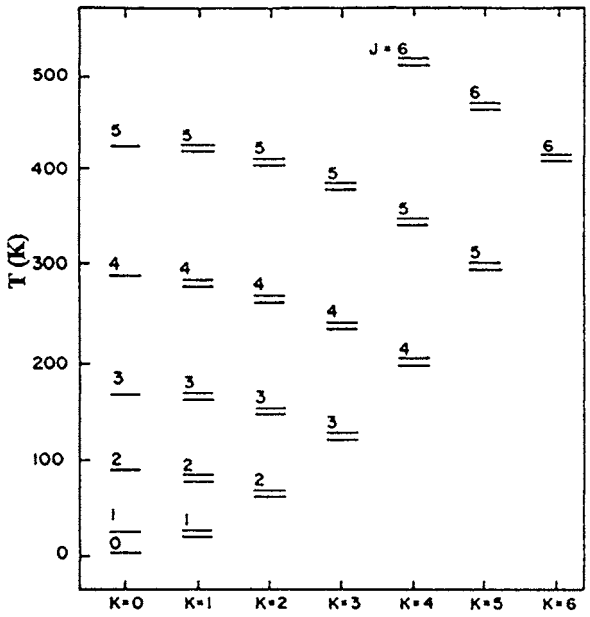


Figure 1. A partial energy level diagram for the rotational energy levels of a molecule with an axis of symmetry (a symmetric top); the levels illustrated are for the ammonia molecule NH_3. The quantum number J is indicated for each level; the quantum number K is indicated on the abscissa. In the case of ammonia, each level with $K > 0$ is doubled because there are two equivalent positions for the N atom, one of each side of the plane defined by the three H atoms. Perturbations in the molecule (not discussed here) ensure that these two levels have slightly differing energy.

Excitation of molecules

Radiative excitation

Stellar radiation may raise molecules to high states of excitation from which relaxation will occur, though not necessarily back to the initial state. In the mainly neutral regions of space where molecules are found, stellar radiation is truncated at wavelengths of 91.2 nm corresponding to the ionization potential of hydrogen (13.6 eV). All radiation shortward of this value is confined within the Strömgren sphere of ionized hydrogen (H II REGION) around the star. Within the Strömgren sphere, any molecules that form are very rapidly destroyed through processes of direct photodissociation and direct photoionization. For example, in the case of molecular hydrogen in its ground vibrational state, the following processes occur:

$$\begin{aligned} H_2 + h\nu &\rightarrow 2H \quad (h\nu > 14.7\ \text{eV}) \\ H_2 + h\nu &\rightarrow H_2^+ + e \quad (h\nu > 15.4\ \text{eV}) \\ H_2^+ + h\nu &\rightarrow H + H^+ \quad (h\nu > 2.65\ \text{eV}) \end{aligned}$$

for photons ($h\nu$) of appropriate energy. Here, 'direct' processes signify transitions to states which are either repulsive (so that the molecule falls apart), or states in which the excess energy in the excited molecule is sufficient to dissociate or ionize the molecule (see also EXCITATION AND IONIZATION).

In molecular clouds which are predominantly neutral and where the stellar radiation is restricted to wavelengths greater than 91.2 nm (photon energies less than 13.6 eV), many of these direct processes are inaccessible. However, electronic excitation of H_2, and many other molecular species, is still possible, and this enables the detection of H_2 and provides a means of both populating vibrational energy levels and of destroying molecules. For H_2, absorption of starlight at wavelengths around 100 nm raises this molecule from the ground electronic state, X, and ground vibrational state $v'' = 0$ to an excited electronic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

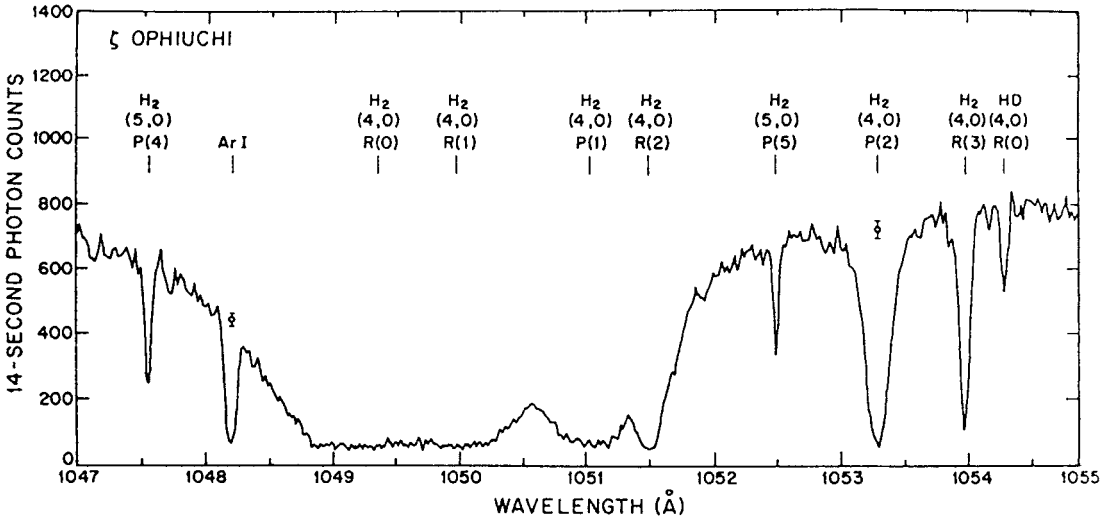


Figure 3. The ultraviolet absorption spectrum of H_2 (the wavelength scale is in angströms, Å; 1 Å is 0.1 nm). The radiation from a bright star ζ Ophiuchi is absorbed by foreground H_2 in a diffuse interstellar cloud. The lines of H_2 arise in the ground electronic state and ground vibrational level. For example, the absorption at a wavelength of 1054 Å is from X, $v = 0$, to the upper electronic state, B, in the $v = 4$ vibrational level. The transitions also involve a change of rotational levels; in the notation used in the figure, $R(0)$ refers to an increase in J from 0 to 1, while $P(1)$ implies a reduction in J from 1 to 0. Some of the lines are saturated. (From Spitzer and Jenkins 1975.)

state, B or C, in a variety of vibrational and rotational states

$$H_2(X;\, v'' = 0;\, J'') \rightarrow H_2(B \text{ or } C;\, v';\, J')$$

where the v' and J' quantum numbers refer to the upper electronic state, B or C (see figure 2). The absorption of ultraviolet radiation from a background star enables the detection of H_2 in foreground diffuse interstellar clouds (figure 3). This excitation is very rapidly (on a timescale of about 10^{-8} s) followed by radiative relaxation into the ground electronic state, X, and into a range of vibrational and rotational levels of X:

$$H_2(B \text{ or } C;\, v';\, J') \rightarrow H_2(X;\, v'' > 0;\, J'') + h\nu.$$

The distribution over the v'' levels can be calculated, and for excitation by radiation around wavelengths of 100 nm peaks around $v'' = 6$. Then, further relaxation occurs through vibration–rotation transitions within the ground electronic state of the molecule; for example, the vibrational quantum number may change by one

$$H_2(X;\, v'';\, J'') \rightarrow H_2(X;\, v'' - 1,\, J'' \pm 2).$$

Therefore, electronic excitation of a molecule can lead to the population of a range of vibration–rotation states in the ground electronic state, and radiative relaxation from those levels then occurs through infrared and radio wavelength transitions.

In the important case of H_2, the relaxation from the upper electronic state may end in the vibrational continuum of the ground electronic state, X, implying that the molecule falls apart. This vibrational continuum arises in the following way: the harmonic oscillator approximation breaks down for high vibrational energies, and the discrete levels become closer together (see figure 2) and eventually merge into a continuum for states above $v'' = 14$. For excitation of H_2 by radiation near 100 nm, the probability that the molecule relaxes into the vibrational continuum is about 10%. This mechanism constitutes the main destruction route for H_2 in diffuse interstellar clouds. However, one should note that the excitation occurs through absorption in spectral lines. Therefore, if H_2 formation is also occurring, then H_2 may shield itself from the dissociating radiation by absorbing all the photons within the narrow spectral lines. In diffuse clouds, a transition from mostly H atomic to mostly H_2 molecular gas occurs at depths into a cloud corresponding to an optical depth (caused by interstellar dust) at visual wavelengths of a few tenths. For molecules other than H_2, however, this indirect destruction mechanism is generally unimportant. Its special role for H_2 arises because the minima of the potential energy curves of groundstate X and of excited states B and C differ considerably; this tends to promote a relaxation into high v'' states of X.

Other simple molecules may respond radiatively in the direct or indirect modes described here. For example, H_2O molecules absorb strongly in the wavelength range 150 nm to 200 nm, and are directly dissociated. However, larger molecules have more complex responses to radiation, including internal conversion of absorbed energy into vibrational energy of many modes of the molecule. The energy is radiated away in the infrared, on relatively long time scales (on the order of seconds), in many photons.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

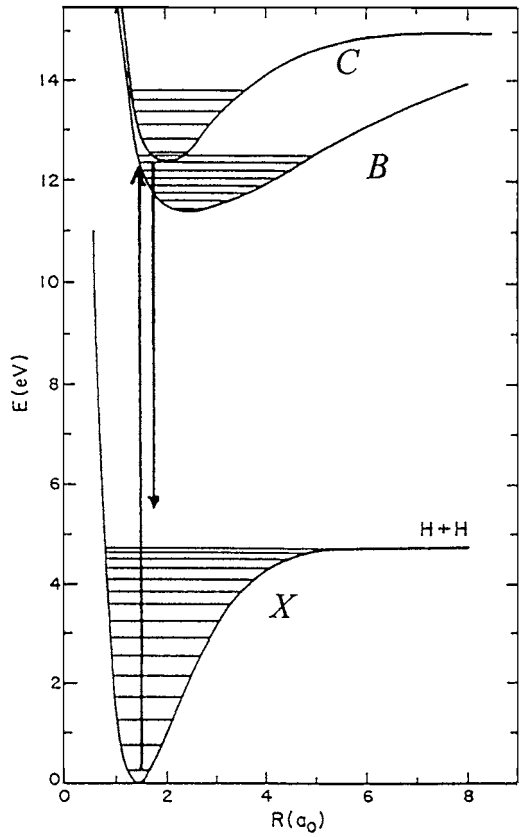


Figure 2. The potential energy curves of H_2, in several electronic states. The curves show how two H atoms interact as the distance between them, R (measured in H-atom radii), is allowed to vary. The curves are attractive at long range, and repulsive at short range. The vibrational energy levels associated with each electronic state are shown. The transition from the lowest vibration level of the ground electronic state, X, to the upper state, B, is shown; this excitation is rapidly followed by a relaxation into discrete or continuum vibrational levels of the ground electronic state.

Collisional excitation

In the collision between a molecule, M, and a collision partner, P

$$M + P \rightarrow M^* + P$$

energy may be transferred to or from the molecule, M, changing its energy state. The final state of the molecule M^*, may differ from the initial state in its electronic, vibrational or rotational energies, or in a combination of these energies. If the process is one of excitation, then the internal energy of M^* is increased at the expense of kinetic energy of the colliding pair. Conversely, in de-excitation processes, internal energy is converted into kinetic energy.

Collisional excitation of molecules into electronic states via thermal processes is generally rare in astronomical situations, because the energies required (a few eV) correspond to temperatures that are too high for molecules to survive for very long. The final state of the molecule, M^*, may then correspond to a dissociating state. Non-thermal excitation by fast particles can, however, be very important. Molecular hydrogen in collision with cosmic rays (mainly MeV protons and electrons) is directly excited electronically into the ionization continuum of the molecule

$$H_2 + \text{c.r.} \rightarrow H_2^+ + e^* + \text{c.r.}$$

and the free electron carries away some 30 eV of energy. This energy is partly lost in collision with other molecules and with atoms, e.g.

$$e^* + H_2 \rightarrow e^* + H_2^*$$

followed by

$$H_2^* \rightarrow H_2 + h\nu$$

so that at least some of the energy released in the ionization is radiated away. However, some of the electron's energy is shared in elastic collisions with atoms, e.g.

$$e^* + H \rightarrow e^* + H^*$$

and this represents a heating process for the gas which is important in situations such as molecular clouds from which starlight (the alternative energy source) has been excluded by interstellar dust. The molecular ion, H_2^+, created in the ionization of H_2 by cosmic rays, is responsible for initiating an extensive ion–molecule interstellar chemistry.

Thermal excitation and de-excitation of vibrational and rotational levels occurs in warm (around 2000 K) and cold (around 10 K) interstellar clouds (see figure 4). These processes compete with radiative relaxation, and together determine which molecular transitions are important in tracing and in cooling the gas. Emission is maximized when the collision rate is high, implying a high gas density, but not so high that radiative relaxation is quenched by collisional de-excitation. Hence, each transition of each molecule has associated with it a range of physical conditions in which it radiates most effectively, and this range is determined by equating the radiative rate (the Einstein A coefficient for spontaneous emission, $A_{\rm em}$) with the collisional excitation rate

$$n_{\rm coll} v_{\rm coll} \sigma_{\rm coll} \simeq A_{\rm em}$$

where $n_{\rm coll}$, $v_{\rm coll}$ and $\sigma_{\rm coll}$ are the number density of the colliding partner (usually H_2), and the velocity and cross section for the collisions.

For a given temperature, therefore, this equation gives the approximate number density range of collision partners for which particular transitions of particular molecules are significant radiators, the so-called critical density. For example, in molecular clouds at temperatures of about 10 K, then CO(J = 1–0) is an effective tracer of gas in which the number density of H_2 is about 10^9 H_2 molecules/m^3, whereas the 1-0 rotational emission of CS traces the gas most effectively when the density is about 100 times greater. The large differences between these

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

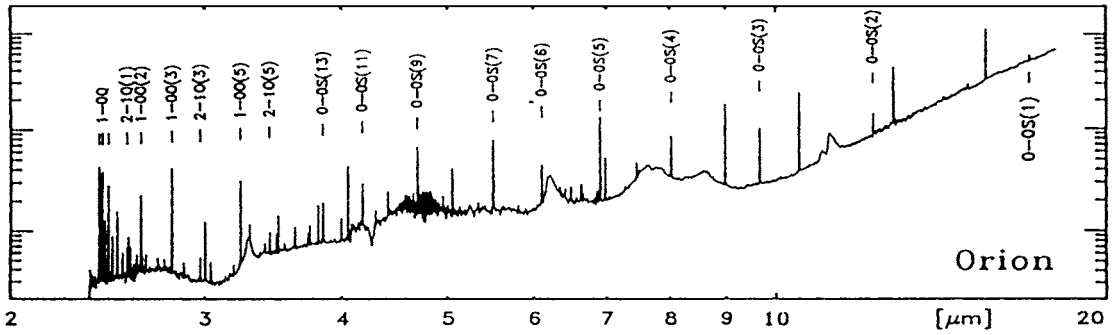


Figure 4. Vibrational emission from collisionally excited H_2 in warm interstellar gas in the Orion Molecular Cloud, detected by the Short Wavelength Spectrometer on the Infrared Space Observatory (ISO). This gas has been heated by the passage of an interstellar shock. The emission in the spectrum arises because the collisions between atoms and molecules in the gas are sufficiently energetic to populate the $v = 1$ and $v = 2$ levels and high rotational states. Relaxation to lower energy states then provides the radiation that is detected. (From Rosenthal *et al* 1999.)

values are caused mainly by the differences in the Einstein A values for the various transitions of each molecular species. In practice, the wealth of the chemistry and the richness available in molecular spectra ensure that there is a molecule and transition that is an appropriate tracer for almost any physical conditions in which molecules can exist in astronomy.

Surface physics

In interstellar molecular clouds, interstellar dust particles are generally well mixed with the gas. The particles have a range of sizes from about a nanometer to about a micrometer, and they probably have a porous open structure. They are probably composed mainly of silicates and carbons, and the total mass of interstellar dust is about 1% of the mass of interstellar gas. Both the gas and dust in molecular clouds are normally cold, with gas phase and solid temperatures typically around 10 K. At such low temperatures, there is a high probability that molecules stick to dust surfaces when they collide with them. For cold interstellar clouds with a normal gas:dust ratio then the time scale for loss of molecules from the gas phase by sticking to dust grains is rather shorter than $10^6(10^{10}\ \mathrm{m}^{-3}/n)(1/S_X)$ years, where n is the total number density of H atoms in all forms in the gas, and S_X is the sticking probability per collision of species X. Therefore, in a molecular cloud with number density $n = 10^{10}$ H atoms/m^3 and with a sticking probability of $S_X = 1$, the time scale t_0 for loss of molecules X from the gas is about one million years, i.e. the number of molecules of X would, in the absence of other processes, decay exponentially as e^{-t/t_0}. Therefore, significant depletion of gas phase molecules may occur within the typical age of a cloud (probably a few million years).

The interaction between a perfect surface of the dust grain and a closed shell molecule (i.e. one with all valencies satisfied) is a physical process involving long-range van der Waals forces. In fact, the surfaces of dust grains are likely to be defective, and stronger, shorter-range forces may also be involved. In any case, the interaction must involve a transfer of energy from the incident molecule into the grain if sticking is to occur.

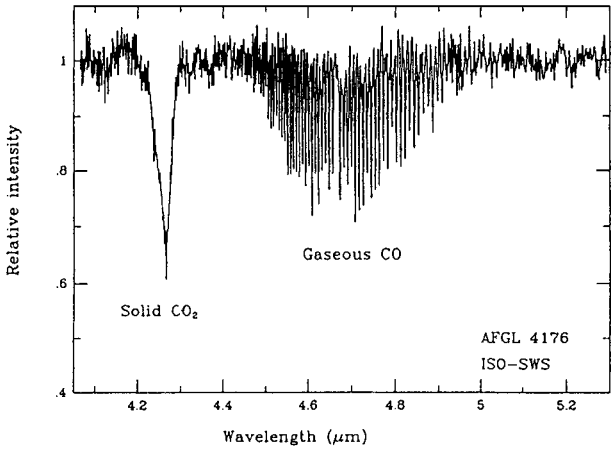


Figure 5. The normalized spectrum obtained with the Short Wavelength Spectrometer on the Infrared Space Observatory (ISO) towards massive young stellar object AFGL 4176 embedded in a dense molecular cloud. The strong broad absorption at 4.27 μm is due to solid CO_2, while the characteristic rotational structure near 4.4–4.9 μm indicates the line of sight. (From van Dishoeck *et al* 1996.)

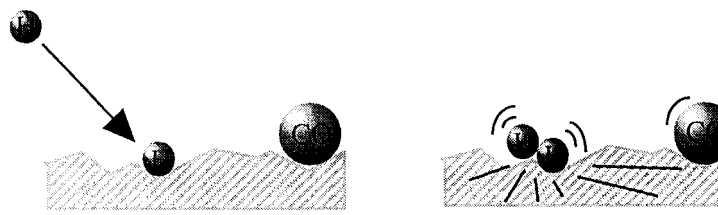


Figure 6. Chemically driven desorption of molecules. In this schematic diagram t
H_2 deposits some energy in the grain, causing a local heating of the grain. A CO m
desorbed by this transient rise in temperature. (From Takahashi and Williams 199

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Experimental and theoretical studies suggest that for most interstellar molecules and likely surfaces this interaction leads to energy transfer (possibly in repeated encounters) and that the sticking probability is on the order of unity. The detection of molecular ices on dust grains in molecular clouds confirms that sticking must be efficient. These ices are found to contain H_2O, CO, CO_2 and some other simple molecules, and molecules on the grains have a quite distinct spectrum from those in the gas (see figure 5). The ices are all detected by means of absorption in the infrared, corresponding to stretching vibrations; e.g. the CO molecule has a vibrational transition at a wavelength near 4.7 μm, and all rotational structure is suppressed for CO in the ice.

Observations show, however, that the fraction of any particular molecule locked up in ice on dust grains is often rather small. This suggests either that all such clouds are young, compared to t_0, or that some process is returning molecules from the solid to the gas phase. Since the former possibility seems implausible, several desorption mechanisms have been invoked. Thermal evaporation at the low temperature of interstellar dust is too slow, and the desorption mechanisms depend on non-thermal processes to create a local 'hot-spot' from which some evaporation may occur. Sources of energy for the 'hot spot' include deposition of energy by cosmic rays in the grains, photodesorption by a cosmic ray generated radiation field internal to the cloud, or chemical energy from reactions occurring at the surface of grains (see figure 6).

Bibliography

Elitzur M 1992 Interstellar masers *Ann. Rev. Astron. Astrophys.* **30** 543–74

Hartquist T W and Viti S 1998 The basics of the structures and spectra of simple molecules *The Molecular Astrophysics of Stars and Galaxies* ed T W Hartquist and D A Williams (Oxford: Oxford Science Publication) pp 1–10

Kirby K P 1990 Molecular photoabsorption processes *Molecular Astrophysics* ed T W Hartquist (Cambridge: Cambridge University Press) pp 159–80

Rosenthal D, Bertoldi F, Drapatz S and Timmerman R 1999 *Proc. Conf. ¡The Universe as seen by 150' (Paris, France, 20–23 October 1998) (ESA Special Publications SP-427)* ed M Kessler

Roueff E 1990 Collisional excitation processes *Molecular Astrophysics* ed T W Hartquist (Cambridge: Cambridge University Press) pp 232–58

Shull J M and Beckwith S 1982 Interstellar molecular hydrogen *Ann. Rev. Astron. Astrophys.* **20** 163–90

Spitzer L Jr and Jenkins E B 1975 Ultraviolet studies of the interstellar gas *Ann. Rev. Astron. Astrophys.* **13** 133

Takahashi J and Williams D A 1999 Chemically-driven desorption of CO from icy grains in dark clouds *Mon. Not. R. Astron. Soc.* at press

van Dishoeck E F *et al* 1996 A search for interstellar gas-phase CO_2; gas:solid state abundance ratios *Astron. Astrophys.* **315** L349

Williams D A 1993 Physical and chemical processes on dust *Dust and Chemistry in Astronomy* ed T J Millar and D A Williams (Bristol: IOP Publishing) pp 143–70

David Williams

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Piazzi, Giuseppe (1746–1826)

Born in Ponte di Valtellina, Italy, became a monk, professor of theology and professor of mathematics in Palermo, where he set up an observatory in 1789. Measured the positions of 7646 stars, discovering that the star 61 Cygni had a large proper motion. Because this suggests that the star is nearby, BESSEL chose it to investigate its parallax, making the first successful such measurement. He discovered the first minor planet, Ceres, on 1 January 1801, but was able to make only three observations before it was lost behind the Sun. Fortunately, GAUSS had developed mathematical techniques with which he could calculate the orbit and Ceres was recovered. The thousandth asteroid discovered was named Piazzia in his honour, as was the British astronomer CHARLES PIAZZI SMYTH.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Picard, Jean (1620–82)

Astronomer, born in La Flèche, France, became professor of astronomy at the Collège de France in Paris and a member of the Académie Royale des Sciences. He made instruments to make more accurate observations, including (with Auzout) the movable-wire micrometer to measure small angles, such as the diameters of the planets, and a sighting telescope for a quadrant and sector to determine star positions. He also measured the length of a degree on the meridian at Paris; the data were used by NEWTON in his calculations of the shape of the Earth. Picard measured the location of TYCHO BRAHE's observatory at Hven so that TYCHO's observations could be re-interpreted.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pic-du-Midi Observatory (Observatoire Midi-Pyrenees) (OMP)

OMP is under the administrative supervision of both the Institute des Sciences de l'Univers (INSU) of the French National Center for Scientific Research (CNRS) and the Ministry of Research, Technology and Education. It has laboratories located at the Université Paul Sabatier in Toulouse, Bagnères, Lannemezan and at the summit of Pic du Midi de Bigorre.

The Pic du Midi Observatory is located in the Pyrenees, 2876 m above sea level—the highest observatory in France. Construction began on the site of a meteorological station in 1878. It was affiliated to the Toulouse University Observatory in 1903.

The main instrument is the 2 m Bernard Lyot Memorial Telescope, which began operation in 1980. It is named after the scientist who carried out the first experiments with a coronagraph in the 1930s. Other instruments include a 1.06 m lunar and planetary telescope, a solar telescope, four coronographs and a 60 cm telescope used by amateur astronomers.

The observatory is perhaps best known for solar and planetary observations and for the detailed lunar map produced for NASA's Apollo program in 1968. It is being developed as a cultural center for the Bigorre region in the Hautes-Pyrénées.

For further information see http://www.obs-mip.fr/omp/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pickering, Edward Charles (1846–1919) and Pickering, William Henry (1858–1938)

Astronomers—two brothers. Born in Boston, MA, Edward became director of the Harvard College Observatory and observed the brightnesses of 45 000 stars. He hired a number of women, including WILLIAMINA P FLEMING, ANNIE J CANNON, Antonia Maury and HENRIETTA LEAVITT, and produced the *Henry Draper Catalog*, classifying the spectra of hundreds of thousands of stars. He and HERMANN CARL VOGEL independently discovered the first spectroscopic binary stars. Edward's brother William, born in Boston, worked in Harvard's astronomy department with the new astronomical detector, dry-plate photography, becoming the first person to discover a satellite (Phoebe, Saturn's ninth moon) by photography in 1899. William established an outpost observatory for Harvard at Arequipa, Peru, and published sensationalist accounts of Martian canals, using the Arequipa telescopes, not for the stellar spectroscopy for which they had been established. For this and his inability to stay within budget his brother dismissed him from his post at Arequipa. William took up with PERCIVAL LOWELL and founded the Lowell Observatory to observe Mars. William's and Percival Lowell's careers overlapped later when William, independently of Lowell, predicted the position of a trans-Neptunian planet using the discrepancies in both the orbits of Uranus and Neptune as data, without it being found.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pictor

(the Painter's Easel; abbrev. Pic, gen. Pictoris; area 247 sq. deg.) A southern constellation which lies between Columba and Dorado, and culminates at midnight in mid-December. It was introduced as Equuleus Pictoris (the Painter's Easel) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2. The shortened form (which strictly translated means 'the Painter') was introduced by the American astronomer Benjamin A Gould (1824–96) in the 1870s and adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest stars in Pictor are α Pictoris, magnitude 3.2, and β Pictoris, magnitude 3.9, the first star found, in 1983, to be surrounded by a dusty disk, out of which planets may be condensing. There are no other stars brighter than fourth magnitude. Other interesting objects include ι Pictoris, a double star with two yellow (F9) components, magnitudes 5.6 and 6.4, separation 12.3″, that have the same proper motion, and Kapteyn's Star, a nearby star which has the second largest proper motion of any star.

There are no bright star clusters, nebulae or galaxies in Pictor.

See also: Beta Pictoris, Kapteyn's Star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pioneer Missions

Development of the first 'Pioneer' space missions began with the advent of the space age in 1957. After NASA was established by Congress in 1958, a further series of Pioneer missions were developed by NASA's Ames Research Center. The series included humanity's first probes of space outside Earth's influence, first passage through the Asteroid Belt, first flyby of Jupiter, first flyby of Saturn and first to escape the solar system. The final Pioneer mission was the United States' first intense study of Venus. Pioneers became consistently successful with numerous discoveries and long productive lives.

Pioneers 1–5 were part of the US Air Force's urgent program to get America competitive in the space age. These were the United States' first probes into space beyond altitudes of the early satellites to measure radiation and detect micrometeoroids. Pioneer 1 was launched in 1958, the year NASA was created. The program was completed in 1960.

Pioneers 6–9, launched in the years 1965 through 1969, were NASA's first spacecraft to explore space beyond any of Earth's influence. These four orbited the Sun independently from Earth, not far inside and outside Earth's orbital path at 1 astronomical unit (AU), 93 million miles, from the Sun. They measured the million-miles-per-hour wind of plasma (electrons and ions) emanating from the Sun. Pioneers 6–9 identified storms on the Sun's surface with magnetic storms from solar wind affecting communications and power grids on Earth. Pioneers 6–9 proved interplanetary space flight safe from destructive dust particles.

Pioneers 10 and 11 were launched in 1972 and 1973. Greatly more complex than earlier Pioneers, they were the first spacecraft to explore beyond Mars (1.6 AU),

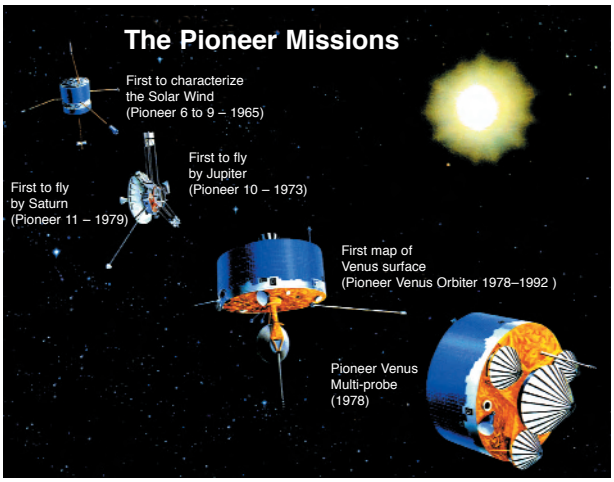


Figure 1. The Pioneer missions. From left to right: Pioneers 6 through 9, Pioneers 10 and 11, Pioneer 12 and Pioneer 13. (NASA / Ames Research Center.)

through the Asteroid Belt (2.5–4 AU), near Jupiter (5 AU) and beyond. Pioneer 11 later made a close reconnaissance with Saturn (10 AU). As the first man-made craft to escape the solar system, Pioneer 10 carried a plaque to identify where in our galaxy it originated and to introduce the creatures who made it. (Pioneer 11 carried a duplicate.) Data collection from Pioneer 10 continued through 2000.

The Pioneer Venus Orbiter and Multiprobe spacecraft, launched in 1978, were USA's first dedicated investigations of VENUS. The Orbiter mapped Venus's cloud-shrouded surface by radar, examined Venus's upper atmosphere, and explored interaction of the solar wind with Venus's atmosphere over a 14 yr period. (The Orbiter also measured rate of water loss from Halley's

Table 1. Pioneer 10 and 11 science payloads.

Experiment	*Principal investigator*	*Institution*
Helium Vector Magnetometer	Edward J Smith	Jet Propulsion Laboratory
Flux-gate Magnetometer (Poineer 11)	Mario H Acuna	Goddard Space Flight Center
Plasma Analyzer	John H Wolfe	Ames Research Center
Charged Particle Composition	John A Simpson	University of Chicago
Cosmic Ray Energy Spectra	Frank B McDonald	Goddard Space Flight Center
Jovian Charged Particles	James A Van Allen	University of Iowa
Jovian Trapped Radiation	R Walker Fillius	University of California at San Diego
Asteroid–Meteoroid Astronomy	Robert K Soberman	General Electric Co., Drexel University
Meteoroid Detection	William H Kinard	Langley Research Center
Celestial Mechanics[a]	John D Anderson	Jet Propulsion Laboratory
Ultraviolet Photometry	Darrell L Judge	University of Southern California
Imaging Photopolarimetry	Tom Gehrels	University of Arizona
Jovian Infrared Thermal Structure	Guido Munch	California Institute of Technology
S-Band Radio Occultation[a]	Arvydas J Kliore	Jet Propulsion Laboratory

[a] No on-board experiment. Used tracking data only.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

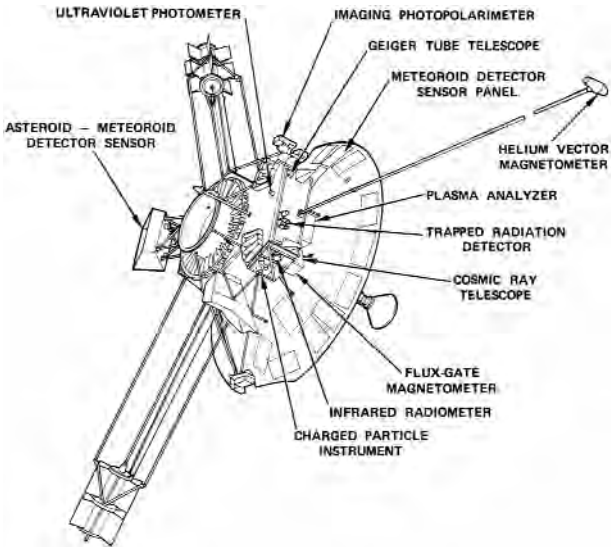


Figure 2. Pioneer 10 and 11 configuration showing locations of scientific instruments. (NASA / Ames Research Center.)

comet as it passed near the Sun.) The Multiprobe deployed four probes into Venus's atmosphere describing the atmosphere down to the surface where lead and zinc would melt.

PIONEER VENUS, the last launched of NASA's 'Pioneer' series of programs, is summarized in a separate article. Below, from latest to earliest, are summary descriptions of the missions of Pioneers 10 and 11, Pioneers 6–9 and Pioneers 1–5. Forms of NASA's Pioneer spacecraft are illustrated in figure 1.

Pioneers 10 and 11

Dominating scientific objectives for Pioneers 10 and 11 were important practical questions about outer space: would exploration beyond the orbit of Mars with more sophisticated spacecraft be feasible and in what environments? Of special concern were possibilities of an impenetrable population of micrometeoroids among asteroids and of destructively intense radiation belts (charged particles trapped in a magnetic field) then theorized to be source of JUPITER's radio noise. Scientific objectives included study of the solar wind en route to Jupiter and beyond, and the maximum practical definition of Jupiter and its environment from a small spacecraft. Penetration of the boundary of the heliosphere with the interstellar medium (heliopause) was expected to occur between 5 and 10 AU. The interstellar medium was to be characterized.

The scientific instruments carried on Pioneers 10 and 11 are listed in table 1 and are indicated in figure 2.

Designed for 21 month or longer missions beginning in 1972 and 1973, Pioneer 10 was still transmitting data after 28 yr, and Pioneer 11 produced data for 21 yr. Launched in heliocentric orbits with energies to reach about 6 AU, their final planetary encounters in 1973

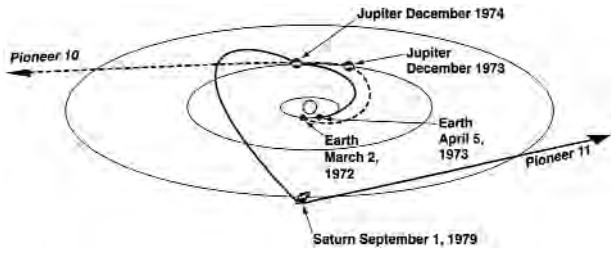


Figure 3. Pioneer 10 and 11 trajectories. (NASA / Ames Research Center.)

(Pioneer 10, Jupiter) and 1979 (Pioneer 11, Saturn) gave them sufficient velocities to escape the solar system. They recede at about 2.5 AU per year forward (Pioneer 11) and backward (Pioneer 10) along the Sun's motion among nearest stars (see figure 3).

Numerical density of dust particles was detected with an array of pressurized cells and was found to follow a diminishing gradient with solar distance unaffected by the ASTEROID BELT. (However, density was surprisingly greater in Jupiter's equatorial plane where VOYAGER later showed a faint ring.) Size distribution of particles was found as expected by the fewer particles larger than 10^{-9} g that could puncture thicker-skinned cells on Pioneer 11. An optical detector for tiny asteroids did not produce convincing observations.

At Jupiter, radiation belts were measured. Giant inflation and compression of Jupiter's shock wave in the solar wind were discovered dependent on the solar wind's energy. Intensity of trapped radiation was found to be several thousand times greater than Earth's, and the performance of some of Pioneer 10's subsystems and instruments was temporarily impaired by radiation. Jupiter's magnetic field, 19 000 times Earth's magnetic moment, was precisely modeled. Axis and rate of rotation of the nongaseous core of the planet were established for the first time. Cloud-top temperatures and an estimate of the planet's radiation heat balance were determined with the infrared scanner. Hydrogen and helium proportions in Jupiter's upper atmosphere and in space were measured. Abundance of gas and dust above the clouds were estimated with data from the Imaging Photo-Polarimeter (IPP). The first close-up 'pictures' of Jupiter, and of SATURN and its rings, came from the two-color optical imaging mode of the IPP. ('Images' were constructed from recordings of light intensity each half-milliradian of rotation of the spacecraft as it passed by Jupiter and Saturn, and the instrument's small telescope was stepped such that the planet would appear to move correspondingly in cone angle during each rotation. Clarity and scientific value of these 'images' greatly exceeded expectations.)

Exploration of Saturn by Pioneer 11 was first contemplated several months after launch of Pioneer 10. Feasibility of a Jupiter gravity assist to Saturn was

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

identified then, and the longevity of Pioneers' newly redesigned plutonium power sources was showing promise. After Pioneer 10's successful equatorial flyby of Jupiter, Pioneer 11 was targeted for a much closer steeply spiraled 'reverse' (clockwise looking south) flyby of Jupiter. The new trajectory greatly enhanced definition of Jupiter's characteristics, such as magnetic field, charged particle distribution with latitude, polar optical depth etc, admitted a much smaller radiation dose to the spacecraft (from higher average Jovian latitudes and higher speed) and induced a trajectory inclined 16° to the Sun's equatorial plane that subsequently encountered Saturn (see figure 3). A proposal for high-risk passage between Saturn and its visible rings was abandoned in 1977 for a test of Voyager's intended penetration of the ring plane just outside the A ring. Approach to Saturn was at low inclination to its ring plane, so *in situ* measurements of the environment could be made about 3/4 inward from the rings' visible outer edge. A magnetic field and radiation belt was discovered and defined at Saturn. Interaction of radiation particles with ring material was found to produce a uniquely radiation-free space under the rings. Corresponding to results obtained at Jupiter, particulars of Saturn's atmosphere, external heat balance, internal structure, mass and rotation were gleaned from Pioneer 11's encounter.

A confusing picture of solar influence in interplanetary space was initially presented by Pioneer data. Energetic interactions between magnetic field lines and charged particles led to a number of theories to explain the data from Pioneers, later combined with data from other spacecraft. Reversal in polarity of the Sun's magnetic field every 22 yr, i.e. every other 11 yr solar cycle, was discovered through Pioneer 11's data. Pioneer 10 and 11's data, together with data from Voyagers and others, have been analyzed to show a dynamic pattern of plasma embedded in outward-moving magnetic field lines beyond 70 AU. The heliopause lies very much farther out from the Sun. Low-energy helium, oxygen and nitrogen ions, increasingly encountered by both spacecraft toward extremes of their productive travels, suggest incursions from the still unexamined interstellar medium.

Densities of cosmic rays of energetic electrons and ions from interstellar space were found to increase as Pioneers 10 and 11 ranged beyond 10 AU from the Sun, but not as rapidly as expected. This implication of strength of the Sun's influence is further evidenced by modulations in cosmic ray spectra. High-speed streams of solar wind correlate with decreased local cosmic rays in lower energy ranges. These observations add to suggestions that the radius of the heliosphere varies and is perhaps much greater than 100 AU.

Precise Doppler tracking of Pioneers 10 and 11 has provided intervals of data with which to test hypotheses of long (several AU) gravity waves emanating from celestial cataclysms. None has been detected. Mysteriously, these tracking data persist in showing 10^{-10} g excess in gravitational attraction of the solar system, apparently due to unidentified mass.

On March 1, 2002, engineers and technicians at the NASA Jet Propulsion Laboratory's (JPL) Deep Space Network in Goldstone, Calif., sent a signal to Pioneer 10 which is still hurtling toward the fringes of the solar system. Twenty-two hours later, at 1:47 p.m. EST, engineers and technicians at the network's facility in Madrid, Spain, monitoring a 70 meter dish antenna, heard Pioneer's response. The signal was received on the 30th anniversary of the spacecraft's launch (2 March, 1972). NASA previously lost contact with Pioneer 10 in August 2000, but made contact again in April 2001 by switching the spacecraft to a different communications mode. Before 1 March, NASA had most recently made contact with the spacecraft on 9 July, 2001. There was another successful contact with telemetry received on 27 April 2002 with more possible in the future depending upon the overall communication system link margin, spacecraft power, and the durability of the transmitter. Scientific data were received during these contacts from the single remaining operating instrument onboard and passed on to the NASA Ames Research Center mission operations center and then relayed to the Principal Investigator, James Van Allen, at the University of Iowa.

Spacecraft

The form of the Pioneer 10 and 11 twins is an 8 ft diameter parabolic dish reflector with equipment and instruments in (and on) a box mounted behind (figure 2). Two pairs of radioactive thermoelectric generators (RTGs) and a magnetometer protrude from the equipment box on three booms, and an antenna feed is mounted on a tripod in front of the reflector. Mass was about 285 kg. The spacecraft were built by TRW Inc.

Three especially difficult technical problems were overcome in development of Pioneers 10 and 11: (1) electrical power was needed to at least 6 AU from the Sun; (2) reliable communications were required through 7 AU of the unexplored interplanetary medium; (3) magnetic cleanliness was required to permit measurement of the weak interplanetary field levels. Plutonium RTGs had been flown but required basic improvements to reach past Jupiter reliably and to gain more from the isotope's 93 yr half-life. The generators were redesigned and rigorously qualified against impacts and temperatures of accident to ensure against leakage into Earth's environment. The telemetry system was designed with convolutional encoding so that data bit errors could be corrected for a then-theoretical improvement of communications at extremely low signal power margins. The Deep Space Network, developed and operated by JPL, improved

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

designs of receivers on Earth so much during development and flights of the Pioneers that data were ultimately recorded from more than 10 times the original 7 AU system design capability. The spacecraft's material content and arrangement of its circuits were planned and tested to minimize magnetic fields. Partly because of the spacecraft's simplicity (e.g. no camera or gyroscopes), a uniquely low background was achieved with Pioneer allowing unexcelled measurement of the interplanetary and planetary magnetic fields.

Pioneers 6–9

Pioneers 6, 7, 8 and 9 were the first Pioneers developed under management of NASA's Ames Research Center. Each was of the same general design, built by TRW Inc. They carried scientific instruments to detect and measure magnetic field, electric field, plasma, cosmic rays and their anisotropy, and cosmic dust. Their orbits, at ranges from 0.8 to 1.1 AU from the Sun, provided descriptions of SOLAR WIND from varying longitudes around the Sun in Earth's orbital plane.

Pioneer 6–9's data showed strong correlations of solar wind velocities and densities with eruptions on the Sun's surface. 'Magnetic storms' which upset communications and power grids on Earth became predictable when Pioneer 6–9 spacecraft found their sources on the rotating Sun before Earth was exposed. Solar wind was found to travel past the spacecraft at about 400 km sec^{-1}. Embedded magnetic field lines appeared to move outward with solar wind. Severe solar storms could propel plasma to Earth in as little as 20 min. The Earth and its magnetic field were observed to create a huge 'bow wave' in the solar wind with a 'tail' of turbulence outward from the Sun. Variations in speed and density of solar wind result in large changes in the volume behind its shock wave shielding Earth's magnetic field.

Precise tracking of Pioneers 6–9 over their extended lives enabled NASA to substantially improve fundamental metrics of the solar system needed for guiding interplanetary trajectories. The Sun/Earth mass ratio uncertainty was reduced to 2.5×10^{-6}. Orbits of the Earth and other planets were much better defined than before. Radio transmission from Pioneer 6 began to be occulted by the Sun's corona in 1968. Effects of the corona on the radio carrier signal provided direct data on changing magnetic and electrical characteristics of the corona.

Spacecraft

Structures of Pioneer 6–9 spacecraft were simple rotating cylinders with axes aligned such that their flat 'pancake' antenna patterns always illuminated Earth (figure 1). Long-term stability of these aircraft spinning at 60 rev min^{-1} was the first such application of the principle for pointing scientific sensors and for controlling antenna pattern. Data telemetered at low rates were originally communicated from tracking stations via standard teletype format in blocks of alphabetic characters before breakdown into engineering and scientific intelligence by computers. Delta E and Delta L launch vehicles guided Pioneers 6–9 into the desired orbits around the Sun; compressed gas was pulsed to reorient spin axes normal to their orbital plane.

Designed for at least 6 months of service, Pioneers 6, 7, 8 and 9 produced much more than hoped and gained wide notice, with operating lives of 32, 19, 19 and 25 yr, respectively.

Pioneers 1–5

Pioneer 1 was launched in 1958, the year NASA was created by Congress. Intended to determine the extent of radiation belts in Earth's magnetic field above altitudes of the earliest satellites, Pioneer 1 fell short of plan at about one-third the Moon's altitude. Pioneer 2's launch failed completely. (A number for Pioneer 2 would not have been assigned by subsequent NASA policy.) Pioneers 3 and 4 discovered the VAN ALLEN (radiation) BELTS to consist of two distinct bands of charged particles trapped in Earth's magnetic field. Pioneer 4 passed close to the MOON (37 000 miles) to measure radiation in its vicinity. Although second by 1 week to pass the Moon, Pioneer 4 became humanity's first spacecraft to escape Earth's gravity. Pioneer 5 made the first measurement of the interplanetary magnetic field outside the Earth's influence and operated for the then record of 106 days. These five earliest Pioneers were launched on Thor Able (Pioneer's 1, 2 and 5) and Juno 2 (Pioneer's 3 and 4) launch vehicles. Instruments they carried were magnetometers, radiation detectors, Geiger counters and micrometeoroid detectors.

Bibliography

More about the Pioneer 10 and 11 mission, and both encounters with Jupiter, is in
1975 *Science* **188** 445–77

Early results of Pioneer 11 encounter with Saturn are in
1980 *Science* **207** 400–53

Examples of scientific analyses of the Solar system's interplanetary medium are in
1979 *J. Geophys. Res.* **84**

More about Pioneer 10 and 11 in relatively popular form is in
1980 *Pioneer: First to Jupiter, Saturn, and Beyond* NASA SP-446

Larry Lasher and John Dyer

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Pioneer Venus Mission

Rather contrary to its name, Pioneer was not the first, although certainly the most comprehensive mission yet, to explore VENUS, and most of our current knowledge, of the atmosphere in particular, derives from the measurements it made around 20 years ago. It also made the first detailed radar maps of Venus' surface, although these are now largely superseded by the higher-resolution data from the more recent MAGELLAN mission.

Pioneer Venus was in fact two missions, launched separately within a few months of each other. The first launch placed in orbit the first artificial satellite of Venus (the 'orbiter' mission), while the second delivered four probes, one large and three small, directly into Venus' atmosphere (the 'multiprobe' mission). The orbiter mapped the atmosphere above the clouds at infrared wavelengths, the clouds themselves in the visible and ultraviolet, and the surface using radar. Operating from a highly elliptical orbit, it also investigated the thermosphere of Venus, utilizing measurements of the drag on the spacecraft as it dipped down into the atmosphere, to altitudes as low as 150 km, and occasionally lower. Aided by the angular coverage which it obtained as a result of its spin-stabilized configuration, the orbiter also obtained the first comprehensive maps of the Venus magnetosphere.

History and scientific objectives

The Pioneer Venus program was born of a series of studies carried out in the USA in the late 1960s and early 1970s, aimed at defining the best approach to addressing the major mysteries posed by Venus. The mission in its final form began in January 1972, when NASA formed a Pioneer Venus Science Steering Group (SSG) and assigned the project to its Ames Research Center in Mountain View, California. The SSG laid out a list of the most important scientific questions to be addressed and worked with the project to devise the mission design which would provide the best answers within the imposed cost ceiling of $200 million. From this the concept of the two-launch strategy, within a single launch opportunity (originally 1976, but eventually 1978), evolved.

The multiprobe was to address questions of the nature of the clouds on Venus, their layering, microstructure, and composition, solar heating of the atmosphere as a function of depth, the atmospheric circulation and its driving forces, the bulk composition of the atmosphere, the loss of water and stability of carbon dioxide, the deep atmosphere vertical temperature structure, ionospheric turbulence, ion chemistry and exospheric temperature. The orbiter, to be in position before, and operating during, the probe descent, was to address cloud-top morphology, the cause and extent of the 4 day 'super-rotation' at the cloud tops, vertical temperature structure, gravitational moments, exospheric temperature, composition and loss of water and ionospheric temperatures and motions.

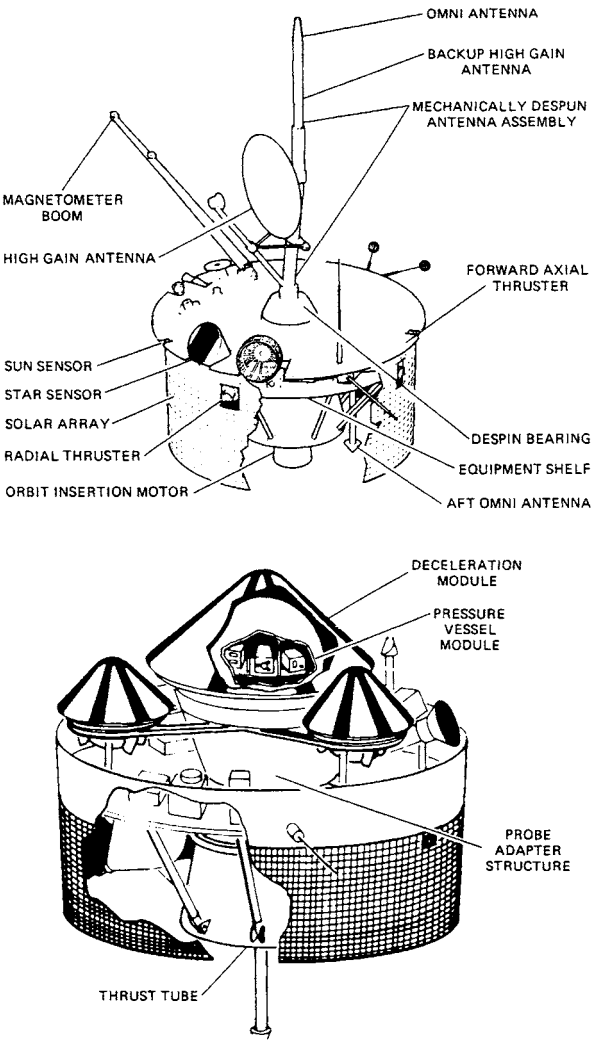


Figure 1. The Pioneer Venus orbiter (top) and multiprobe (bottom) spacecraft, showing the accommodation for the scientific instruments on the upper platform below the antenna on the orbiter, and the positioning of the cone-shaped probes on the corresponding platform of the multiprobe bus (NASA).

Mission description

Both spacecraft used the same basic body, a flat cylinder 2.5 m in diameter and 1.2 m high, built by the Hughes Aircraft Corporation (figure 1). The orbiter was spin stabilized, using a 1.09 m diameter despun dish antenna to communicate with the Earth. It also carried a solid-fuel rocket motor to decelerate the spacecraft on arrival and allow it to achieve orbit insertion. It weighed 553 kg when launched, including 55 kg of scientific instruments and 179 kg of propellant. The multiprobe weighed 875 kg, including the probes themselves at 585 kg.

The mission was launched from Cape Canaveral using two Atlas–Centaur vehicles. The orbiter was launched first, on 20 May 1978, followed by the multiprobe on 6 August. The orbiter arrived at Venus on 4 December

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

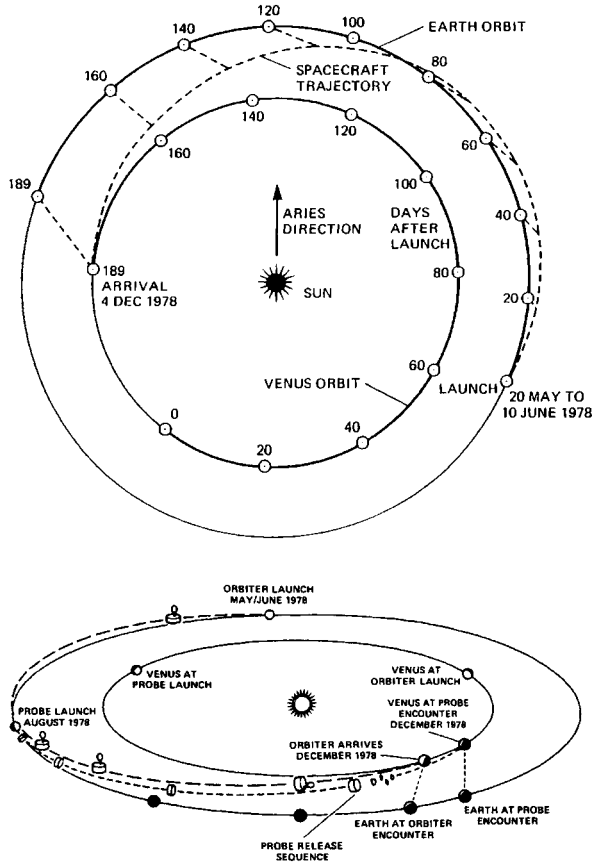


Figure 2. The trajectories followed by the Pioneer Venus orbiter (top) and multiprobe (bottom) spacecraft, en route to Venus.

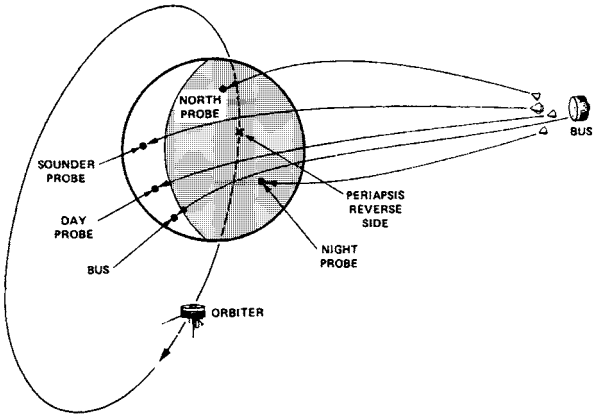


Figure 3. Geometry of the Pioneer Venus probe entries on 9 December 1978, showing the orbiter already on station (NASA).

of the same year, followed by the probes only 4 days later; the latter could follow a faster, more direct trajectory since their velocity relative to Venus on arrival was less critical. Figure 2 shows the trajectories they followed and figure 3 the geometry of their deployment at Venus. The large, or 'Sounder' probe landed near the equator, at 4.0° latitude and 304.0° longitude, the three small probes, now named 'North', 'Day' and 'Night' at 60.2°, 3.2°, −31.3°, 317.7°, and −27.4°, 56.6° respectively. In addition, the bus which had carried the probes prior to separation entered at −37.9°, 290.9°, carrying two instruments (a neutral mass spectrometer and an ion mass spectrometer) and operated down to 100 km altitude. The probe instruments functioned down to a height of about 12 km above the surface, when most of them stopped working simultaneously. An enquiry showed the cause of this to have been the failure of insulating material on one of the external sensors, which caused a short circuit. One of the probes survived its impact with the surface and continued to transmit data for a further hour.

The orbiter functioned throughout its nominal mission of 1 Venus year with the loss of only one instrument, the infrared radiometer, after 72 days of excellent performance. It went on to obtain valuable data for nearly 14 Earth years, finally entering the atmosphere and burning up during mid-1992.

Scientific payload

The total complement of scientific instruments on all six spacecraft is summarized in table 1. The instruments are grouped according to their principal scientific objective, and the code names begin with O, L, S or B to signify mounting on orbiter, large probe, each of the three small probes, or probe carrier bus, respectively. The principal investigator and institution responsible for each experiment are also shown.

Scientific achievements

The scientific results from the mission are very extensive and in some cases still being accumulated. Compilations can be found in the references given at the end of this article. The following is a brief summary.

Pioneer Venus obtained radar altimetry for about 90% of the surface of the planet and made a number of radar images (see VENUS: SURFACE). With these, volcanic and tectonic features such as rift valleys, mountains, continents and volcanoes were discovered. It was found that there is a unimodal distribution of topography on Venus, whereas Earth has a more bimodal distribution, resulting from a dearth of elevated regions of continental size. The detailed mapping confirmed the existence of great troughs or 'rift valleys', the presence of which had been expected from Earth-based radar data. No evidence was found for a Venusian analogue to the continuous ridge systems which are characteristic of the terrestrial plate tectonics system.

Over the course of its 14 yr lifetime, the Pioneer Venus orbiter obtained measurements of the gravity field that, when combined with the radar altimetry results, showed that the interior behavior of Venus is more Earthlike than Mars or the Moon. However, there is a great difference between Venus and Earth in that on Venus there is a strong positive correlation of gravity with topography at all wavelengths.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Pioneer Venus science payload.

Code name	Experiment	Principal investigator, affiliation
Composition and Structure		
LNMS	Large probe mass spectrometer	J Hoffman, University of Texas at Dallas
LGC	Large probe gas chromatograph	V Oyama,Ames Research Center
BNMS	Bus neutral mass spectrometer	U von Zahn, University of Bonn
ONMS	Orbiter neutral mass spectrometer	H Niemann, Goddard Space Flight Center
OUVS	Orbiter ultraviolet spectrometer	I Stewart, University of Colorado
LAS–SAS	Large–small probe atmosphere structure	A Seiff, Ames Research Center
OGPE	Atmospheric propagation experiment	T Croft, SRI International
OAD	Orbiter atmospheric drag experiment	G Keating, Langley Research Center
Clouds		
LN–SN	Large–small probe nephelometer	B Ragent, Ames Research Center
LCPS	Large probe cloud particle size spectrometer	R Knollenberg, Particle Measuring Systems, Inc
OCPP	Orbiter cloud photopolarimeter	L Travis, Goddard Institute for Space Studies
Thermal balance		
LSFR	Large probe solar flux radiometer	M Tomasko, University of Arizona
LIR	Large probe infrared radiometer	R Boese, Ames Research Center
SNFR	Small probe net flux radiometer	V Suomi, University of Wisconsin
OIR	Orbiter infrared radiometer	F Taylor, Jet Propulsion Laboratory and Oxford University
Dynamics		
DLBI	Differential long-baseline interferometry	C Counselman, Massachusetts Institute of Technology
MWIN	Doppler tracking of probes	A Kliore, Jet Propulsion Laboratory
MTUR–OTUR	Atmospheric turbulence experiment	R Woo, Jet Propulsion Laboratory
Solar wind–ionosphere		
BIMS	Bus ion mass spectrometer	H Taylor, Goddard Space Flight Center
OIMS	Orbiter ion mass spectrometer	H Taylor, Goddard Space Flight Center
OETP	Orbiter electron temperature probe	L Brace, Goddard Space Flight Center
ORPA	Orbiter retarding potential analyzer	W Knudsen, Lockheed Palo Alto Research Laboratory
OMAG	Orbiter magnetometer	C Russell, University of California at Los Angeles
OPA	Orbiter plasma analyzer	J Wolfe, Ames Research Center
OEFD	Orbiter electric field detector	F Scarf, TRW
ORO	Orbiter dual-frequency occultation experiment	A Kliore, Jet Propulsion Laboratory
Surface and interior		
ORAD	Orbiter radar mapper	G Pettengill, Massachusetts Institute of Technology
OIDD	Orbiter internal density distribution experiment	R Phillips, Jet Propulsion Laboratory
OCM	Orbiter celestial mechanics experiment	I Shapiro, Massachusetts Institute of Technology
High-energy astronomy		
OGBD	Orbiter gamma burst detector	W Evans, Los Alamos Scientific Laboratory
Interdisciplinary scientists		
	Atmosphere, geology and geophysics	G Schubert, University of California
	Fields and particles	A Nagy, University of Michigan
	Geology and geophysics	G McGill, University of Massachusetts
	Surface features	H Masursky, US Geological Survey
	Atmosphere	T Donahue, University of Michigan
	Atmosphere	J Pollack, NASA Ames Research Center

Probe and orbiter made refined measurements of the abundances of major, minor and noble gas species in the lower, mixed atmosphere and in the upper, diffusively separated atmosphere (see VENUS: ATMOSPHERE). The orbiter discovered that sulfur dioxide is an important absorber of ultraviolet radiation a wavelengths below 3200 Å, accounting in part for the well-known ultraviolet markings in the clouds. However, another absorber, still unidentified but probably a solid or liquid component of the clouds, must be present to account for absorption at longer wavelengths.

The structure of the clouds was investigated globally and vertically, producing (when taken with results from earlier Soviet probes) the first good data on cloud layering, the distribution of particles of different sizes, composition and optical properties. These quantities are now known to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

vary enormously across the planet, so the Pioneer Venus large probe nephelometer (cloud sensor) and particle size spectrometer data, once taken to represent the mean global structure in a relatively horizontally homogeneous cloud, are now regarded as a snapshot of a very dynamic cloud system (see also CLOUDS IN PLANETARY ATMOSPHERES). Net flux measurements determined the sinks for solar radiation and the sources and sinks of infrared radiation in the lower atmosphere and clouds at four locations characterizing daytime, night-time, low-latitude and high-latitude conditions. Taken altogether, these provided an improved understanding of the Venusian version of the 'greenhouse effect' which, coupled with global dynamics, explains the high surface temperature. The orbiter instruments detected radio signals that are thought to originate from lightning discharges in the clouds of Venus, thereby confirming earlier observations made by VENERA probes.

One of the biggest mysteries concerning Venus is the dynamics of its atmosphere. Measurements of vertical profiles of wind velocities at four probe locations, and global wind measurements at the cloud tops and above by remote sensing, helped to clarify the vertical profile of the four-day winds observed in UV markings in the upper cloud layers, while orbiter infrared imaging discovered remarkable structure in the polar regions of the atmosphere. This included the 'polar dipole', a kind of double vortex several thousands of km across rotating around the pole. This and the polar collar, a cold wave-like disturbance surrounding the dipole, are aspects of the circulation pattern in those regions which still remain to be understood.

Direct entry-probe measurements and remote sensing of temperature profiles with the infrared radiometer produced much new data concerning the atmospheric state properties (temperature, pressure and density) globally and vertically from the surface through the clouds and into the upper atmosphere. Several remarkable features were revealed for the first time, including a reverse temperature gradient from the equator to the pole (making the atmosphere above the pole warmer than that above the equator) and global-scale wave motions. Prominent among the latter are the solar tides, i.e. the diurnal cycle of heating of the atmosphere by the Sun, which on Venus have two maxima per day in contrast to the one on Earth. Pioneer Venus also discovered that the high atmosphere well above the cloud tops is much colder at night than in the daytime, with a surprisingly abrupt boundary between them.

Pioneer Venus mapped the AIRGLOW on the dark side of Venus, determined the global characteristics of the ionosphere—its ion composition, temperature, flows, electron concentration and temperature, modification of ionospheric characteristics by input from the solar wind and the production and maintenance of a night-side ionosphere. It determined the nature of the solar-wind interaction with the planet, including temporal and spatial studies of the location of the bow shock and ionopause, and particle and energy input to the atmosphere (see VENUS: INTERACTION WITH SOLAR WIND). Finally, it confirmed that Venus has little if any intrinsic magnetic field, and set a very low upper limit on the magnetic moment of the planet.

Bibliography

A description of the Pioneer Venus mission and its instruments may be found in a special edition of the following journal:

1980 *IEEE Trans. Geosci. Remote Sens.* **18**

The results and their interpretation are discussed in another special issue:

1980 *J. Geophys. Res.* **85**

See also the following books:

Hunten D M, Colin L, Donahue T and Moroz V (eds) 1982 *Venus* (Tucson, AZ: University of Arizona Press)
Bougher S W, Hunten D M and Phillips R J (eds) 1997 *Venus 2* (Tucson, AZ: University of Arizona Press)

F W Taylor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pipe Nebula (B59, B65–7, B79)

A large dark nebula in the constellation Scorpius, resembling a tobacco pipe. E E Barnard assigned five separate catalog numbers to sections of the nebula. B78 is the 'bowl', about $3\frac{1}{2}^\circ$ by $2\frac{1}{2}^\circ$, with approximate position RA $17^{\rm h}\,33^{\rm m}$, dec. -26°, while B59 and B65–7 comprise the 'stem', 5° by 1°, with approximate position RA $17^{\rm h}\,21^{\rm m}$, dec. -27°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pisces

(the Fishes; abbrev. Psc, gen. Piscium; area 889 sq. deg.) A northern zodiacal constellation which lies between Pegasus and Cetus, and culminates at midnight in late September. Its origin dates back to Babylonian times and it is said to represent Aphrodite, the goddess of love in Greek mythology, and her son Eros, who jumped into the Euphrates to escape from the multi-headed Typhon and were turned into fish. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A large but rather inconspicuous constellation, the brightest stars in Pisces are η Piscium, magnitude 3.6, γ Piscium, magnitude 3.7, α Piscium (Alrescha), a close binary with bluish-white (A0 and A3) components, magnitudes 4.2 and 5.3, separation 1.9″, and ω Piscium, magnitude 4.0. Other interesting objects include ζ Piscium, a double star consisting of white (A7) and pale yellow (F7) components, magnitudes 5.2 and 6.4, separation 23″, that share the same proper motion, each of which has a very close (separation <0.05″) companion, 19 Piscium (also called TX Piscium), one of the few so-called 'carbon stars' visible to the unaided eye (magnitude range 4.8–5.2), and M74 (NGC 628), a ninth-magnitude face-on spiral galaxy mistakenly cataloged by John Herschel as a globular cluster.

Pisces also contains the First Point of Aries, or vernal equinox (see First Point of Aries), which when named lay in the neighboring constellation of Aries but has since moved into Pisces as a result of precession.

See also: Fomalhaut.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Piscis Austrinus

(sometimes called Piscis Australis; the Southern Fish; abbrev. PsA, gen. Piscis Austrini; area 245 sq. deg.) A southern constellation which lies between Aquarius and Grus, and culminates at midnight in late August. Its origin dates back to Babylonian times, and it is usually shown on early celestial charts unnaturally drinking water from an urn carried by Aquarius. Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*, though he included α Piscis Austrini (Fomalhaut) in the neighboring constellation of Aquarius.

A small but not inconspicuous constellation, Piscis Austrinus is easily recognized by virtue of α Piscis Austrini, magnitude 1.3. There are no other stars brighter than fourth magnitude. Interesting objects include β Piscis Austrini, a wide binary with bluish-white (A0) and yellow (G2?) components, magnitudes 4.3 and 7.9, separation 30″, γ Piscis Austrini, a close binary with bluish-white (A0) and pale yellow (F5) components, magnitudes 4.5 and 8.2, separation 4.1″, and NGC 7314, an eleventh-magnitude spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plössl, Georg Simon (1794–1868)

Born in Wieden near Vienna, optical-instrument maker. The *Plössl eyepiece* is a four-element design for a telescope eyepiece, well known to amateur astronomers.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plage

A region of intensified emission in the solar chromosphere. Plages are revealed in monochromatic images obtained in the light of various emission lines, for example, the hydrogen-α line and the 'H' and 'K' lines of ionized calcium. The name derives from the French word for 'beach', an allusion to the fact that these features stand out like bright sandy beaches against the fainter background of the chromosphere as a whole.

Physically, plages are regions of enhanced temperature and density that float in the chromosphere. They are heated by magnetic fields, with strengths of upwards of 0.01 T, which are concentrated within solar active regions. Plages usually cover most of the area occupied by an active region, forming before sunspots appear in the region and persisting for some time after the spots have disappeared.

They are similar in nature to the faculae that are observed in the photosphere.

See also: active region, chromosphere, emission spectrum, faculae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planck Surveyor

European Space Agency mission to measure temperature fluctuations in the cosmic microwave background with a precision of about 2 parts in a million and an angular resolution of about 10 arcseconds. Named after Nobel Prize winning physicist Max Planck. Will carry a 1.5 m telescope and two instruments to cover frequencies between 30 and 857 GHz. Planned for launch together with FIRST in 2007.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planck, Max (1858–1947)

Born in Kiel, Schleswig-Holstein, Germany, Nobel prizewinner for physics (1918). Calculated the radiation of a black-body by means of the introduction of the quantum theory of light. The Planck satellite launched to study the black-body spectrum of the cosmic microwave background was named for him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planet

A body, in an independent orbit around the Sun or another star, that is not self-luminous. The major planets of the solar system show a variety of sizes and physical states. The inner planets are of rock and metal, while the outer planets are much larger and predominantly gaseous (and liquid, under the high pressures in their interiors), with relatively small rock/metal cores. Asteroids in the main belt are mostly rocky; Pluto is icy with a rocky core, and the Kuiper Belt objects (which have been given minor planet designations) are icy. The icy comets, though, are excluded from the definition. The term *planetary body* is sometimes used to include sizeable planetary satellites and asteroids.

There is no agreed lower size limit which determines what constitutes a planet. The largest bodies orbiting the Sun are sometimes referred to as major planets to distinguish them from the minor planets, and the category 'minor planets' includes near-Earth objects only meters in size. Meteoroidal particles and dust do not count as 'planetary'. The question of an upper size limit has been highlighted with the discovery in the 1990s of *exoplanets*—planets orbiting other stars. All but a handful of them are several times the mass of Jupiter; the most massive are better described as *substellar objects*. The more massive a substellar object, the more closely its characteristics approach those of a brown dwarf. A common rule of thumb is that 10 Jupiter masses marks the division between planets/substellar objects and brown dwarfs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planet X

A name given to a hypothetical tenth major planet once believed to exist in the outer solar system, beyond the orbit of Neptune. The 'X', which stood for 'unknown', was also appropriate as the roman numeral for 'ten'. The label 'Planet X' was originated by Percival Lowell. From the late nineteenth century, he and others, including William H Pickering, worked out orbits for a large tenth planet which they believed responsible for gravitational perturbations of Uranus, which, even since the discovery of Neptune in 1846, did not appear to be following the orbit calculated for it. In 1930 Clyde Tombaugh discovered Pluto. Though Pluto was only 5° from Lowell's predicted position, and in a similar orbit, it soon became clear that the new planet was too small to be Lowell's Planet X. Tombaugh continued searching, and by 1943 had surveyed the entire sky, but without finding a Planet X.

From the 1940s, further predictions were made for a tenth planet whose perturbations would account for the deflection of comets from an outer reservoir into orbits that would bring them into the inner solar system. Charles Kowal's search from 1977 to 1984 turned up the 180 km Chiron, orbiting between Saturn and Uranus, briefly hailed in the media as the tenth planet. Others, including Robert Harrington and Thomas Van Flandern, returned to the problem of discrepancies in the orbits of Uranus and Neptune, and issued a variety of predictions for a tenth planet; but searches revealed nothing.

In the 1990s the need to invoke a Planet X disappeared. Voyager 2 in 1989 had established a new mass for Neptune, and Myles Standish and others demonstrated that the discrepancies in the orbital motions of Uranus and Neptune could be explained away by this and by observational errors. The discovery from 1992 onwards of the first members of what appears to be a large population of asteroidal objects beyond the orbit of Neptune seems finally to have brought the Planet X chapter to a close.

See also: Kuiper Belt, Pluto and Charon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetarium

Usually a domed building housing a special projector that is used to simulate the night sky. The institution associated with a planetarium is devoted to popular education in astronomy. The term was originally used to describe teaching devices, known as orreries, designed to portray the orbital motions of the planets and their satellites. The first planetarium was opened at the Deutsches Museum in Munich in 1923.

Planetarium installations vary greatly. The larger planetariums have an extensive exhibition area, museum collections and a seating capacity in excess of 600. The most common projector, since the 1940s, is manufactured by Zeiss in Oberkochen, Germany. Modern Zeiss instruments are large, technically advanced combinations of lenses, lights, gears and motors. They are designed to place the planets, Sun and Moon in their correct locations among the stars for thousands of years past and future. Manufacturers in Japan and the United States compete with Zeiss, producing an array of instruments appropriate for a variety of dome sizes and for different purposes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Atmospheres

Planetary atmospheres are the external gaseous envelopes which surround the planets. In the case of the telluric planets, they represent only a negligible fraction of their mass, but they play an essential role in the energy balance between the surfaces and the Sun. In the case of the GIANT PLANETS, which are mostly gaseous, they account for a large fraction of their total mass and constitute their major component. Planetary atmospheres show extremely diverse physical conditions, with surface pressures ranging from nanobars to hundreds of bars and surface temperatures ranging from 40 to over 700 K; they also show large differences in their chemical compositions, from the lightest, hydrogen-dominated, to the densest, dominated by carbon dioxide.

In spite of these differences, there are striking similarities in the physical and chemical processes involved: photochemistry and photodissociation take place as a consequence of solar radiation and, in some cases, magnetospheric effects; condensation takes place on almost all planets, generating cloud structures, sometimes subject to seasonal effects; volcanism is present, or has been present, on several planets and/or satellites, and sometimes drives the surface/atmosphere interaction.

From Mercury to Pluto

Among the nine planets of the SOLAR SYSTEM, all except one have an atmosphere (table 1). MERCURY, the planet closest to the Sun, is too small and too hot to retain a permanent atmosphere. The other planets are all surrounded by a gaseous envelope, but there is an incredible variety in their characteristics. Even within the group of the TERRESTRIAL PLANETS, the atmospheres of Venus, the Earth and Mars show extreme differences in their physical conditions: on Venus, the surface pressure reaches almost 100 bar and its surface pressure is 730 K, while at the surface of Mars, the mean pressure is only 6 mbar and the mean temperature ranges around 210–230 K (see VENUS: ATMOSPHERE, MARS: ATMOSPHERE). Between these two extremes, the EARTH'S ATMOSPHERE shows intermediate physical conditions. Concerning their visual aspect, there is little in common between the cratered surface of Mercury, the thick white-yellow cloud layer of Venus, presumably composed of sulfuric acid, the blue planet Earth surrounded by white clouds, and the red rocky surface of Mars. The terrestrial planets, however, exhibit more similarities in their chemical atmospheric composition, dominated by carbon dioxide and nitrogen. In the case of the Earth, carbon dioxide is indeed present too, trapped under the oceans in the form of carbonates. The main difference, for our planet, comes from the presence of large amounts of oxygen, due to the presence of life.

At further distances from the Sun, the four giant planets—JUPITER, SATURN, URANUS and NEPTUNE—also show striking differences in their visual appearance. Although they are all composed mostly of hydrogen and helium, their clouds, composed of various minor constituents, are very different. Giant planets have no 'surface'; they are mostly gaseous and their central core is not directly observable. Finally, PLUTO, the ninth planet, is a single case by itself. This small icy body, surrounded by a very tenuous atmosphere of nitrogen, shows more similarities with the satellites of the outer solar system than with any other planet.

Some satellites are also surrounded with a stable atmosphere: TITAN, Saturn's largest satellite, has a nitrogen-dominated atmosphere with a surface pressure comparable to that of the Earth. TRITON, Neptune's largest satellite, also has a nitrogen atmosphere with a surface pressure of a few nanobars, comparable to that of Pluto when the planet is close to perihelion. Finally IO, the Galilean satellite closest to Jupiter, has a very tenuous atmosphere of sulfur dioxide generated by active volcanism.

What is an atmosphere?

All planetary atmosphere are governed by the hydrostatic law, which expresses the balance between pressure and gravity. Its equation is

$$\mathrm{d}P/\mathrm{d}z = -\rho g \tag{1}$$

where P is the pressure, z the altitude, ρ is the density and g is the acceleration due to gravity. Combining this equation with the ideal gas law, one can derive the pressure distribution as a function of altitude. In an isothermal atmosphere, its equation is

$$P = P_0 \exp(-z/H) \tag{2}$$

where H is the scale height:

$$H = RT/\mu g \tag{3}$$

R is the universal gas constant and μ is the mean molecular weight. H is the height at which the pressure is reduced by a factor e; the value of H, at the surface of terrestrial planets, is about 8 km for the Earth, 14 km for Venus and 10 km for Mars; it reaches 20 and 40 km in the observable regions (at a pressure of about 0.5 bar) of Jupiter and Saturn respectively. For a given planet, the scale height is proportional to the temperature and thus slightly varies with altitude following the thermal profile. Atmospheric constituents are mixed up to high altitudes; the separation level is the homopause, at a pressure of a few microbars in the case of the Earth. Below the homopause is the homosphere, where mixing is achieved by convection in the lower part and by eddy diffusion above.

Thermal structure

The thermal structure of an atmosphere is a direct result of its interaction with the incoming solar flux. In the case of the telluric atmospheres, a small fraction of the solar radiation (especially in the UV and the near IR) is absorbed by the atmosphere itself (gas and/or

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Physical characteristics of the planets.

Name	Equatorial diameter relative the Earth	Equatorial diameter (km)	Flattening	Mass relative to the Earth†	Mean density	Surface gravity ($m\ s^{-2}$)	Escape velocity ($km\ s^{-1}$)	Sideral rotation	Inclination of the equator to the orbital plane	Principal atmospheric components
Mercury	0.382	4878	0	0.055	5.44	3.78	4.25	58 646 d	0°	H, He, Ne (solar wind)
Venus	0.949	12 104	0	0.815	5.25	8.60	10.36	243 d (r)‡	2°07′	CO_2 (97%)
Earth	1	12 756	0.003 353	1	5.52	9.78	11.18	23h 56 min 04 s	23°26′	N_2 (78%) O_2 (21%)
Mars	0.533	6794	0.005	0.107	3.94	3.72	5.02	24 h 37 min 23 s	23°59′	CO_2 (95%)
Jupiter	11.19	142 800	0.062	317.80	1.24	24.8	59.64	9 h 50 min to 9 h 56 min	3°04′	H, He, CH_4, NH_3
Saturn	9.41	120 000	0.0912	95.1	0.63	10.5	35.41	10 h 14 min to 10 h 39 min	26°44′	H, He, CH_4, NH_3
Uranus	3.98	50 800	0.06	14.6	1.21	8.5	21.41	17 h 06 min§	98°	H, He, CH_4, NH_3
Neptune	3.81	48 600	0.02	17.2	1.67	10.8	23.52	15 h 48 min	29°	H, He, CH_4, NH_3
Pluto	∼0.2	2300	?	0.002	∼2		?	6 d 9 h 18 min	?	—

† $m_E = 5.976 \times 10^{24}$ kg.

‡ (r) indicates that the rotation is retrograde.

§ Rotation period of the magnetic field.

aerosols) while a larger fraction (in the visible range) is absorbed by the surface; a third part is scattered and reflected outside. Near the surface, the troposphere is heated by the infrared radiation of the surface, itself heated by the solar flux. This temperature gradient induces convection which transports the energy toward higher levels; this part of the atmosphere is called the troposphere. The tropopause, at the top of the convective cells, corresponds to a temperature minimum. Above it, in the stratosphere, the energy is transferred by radiation. In several cases, the temperature increases again with altitude, due to the absorption of the solar UV flux by atmospheric gases or aerosols. In the case of the Earth, ozone absorbs in this region while, in the giant planets, the solar flux is absorbed by methane and stratospheric aerosols. The stratosphere may be more or less complex, depending on the spectroscopic properties of the atmospheric constituents. Other heating mechanisms contribute at higher altitudes, including dissociation of radicals and propagation of gravity waves. At even higher altitudes, the solar flux is energetic enough to ionize all atoms; this region, called the ionosphere, strongly interacts with the magnetic field, when present, to produce, in particular, auroral phenomena (see AURORA).

The thermal structure described above basically applies for all planets (figures 1 and 2), except that in the case of the giant planets there is no surface to absorb the visible solar flux, which is absorbed in the troposphere. In addition, with the exception of Uranus, all giant planets exhibit an internal source of energy, comparable to the energy received from the Sun. This internal source is probably the result of specific evolutionary processes.

In the case of the telluric planets, the GREENHOUSE EFFECT can strongly affect the temperature distribution of the lower troposphere. Its mechanism is as follows: the surface, heated by the visible solar flux, mostly emits in the infrared range. This radiation can be absorbed by atmospheric gases if they present spectroscopic transitions in this spectral range: this is the case for CO_2, H_2O and CH_4 in particular. The lower atmosphere is then heated, and radiates back to the surface to heat it even more, and the process accelerates. The extreme case is Venus, which now has a surface temperature of 730 K while its equilibrium temperature, in the absence of greenhouse effect, would be 229 K! Both the Earth and Mars also present a significant greenhouse effect, which causes an increase of 33 K for the Earth and 4 K for Mars.

General circulation

Because the solar flux falls preferentially near the equator of all planets (except Uranus), latitudinal temperature and pressure gradients are created, and the heat is redistributed by thermal winds. The effect of the planet's rotation then determines the general circulation. Due to Coriolis forces, pole-ward winds are apparently deflected in the direction of the planet's rotation. A typical example is, on the Earth, the cyclonic motions (in the northern hemisphere) around low-pressure centers at mid-latitudes; similar examples exist in the southern hemisphere of Jupiter. If nonlinear terms, like centrifugal force, dominate in the momentum balance, the balance with the pressure gradient is called cyclostrophic balance. The Rossby number gives the ratio of nonlinear terms to Coriolis terms in the momentum balance:

$$R_0 = U/fL \tag{4}$$

where U is the mean zonal wind and L is a typical horizontal length scale of the weather system. f is given by

$$f = 2\omega \sin\theta \tag{5}$$

where ω is the rotation angular velocity and θ is the latitude. Geostrophy or cyclostrophy take place respectively for small or large values of R_0. The thermal wind equation is derived from a combination of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

geostrophic balance, the hydrostatic equilibrium and the ideal gas law

$$H\frac{\mathrm{d}U}{\mathrm{d}T} = -\frac{R}{fa}\frac{\mathrm{d}T}{\mathrm{d}\theta} \qquad (6)$$

where a is the planetary radius. If the temperature decreases toward the poles (as is usually the case), the zonal winds increase with height; this is observed on terrestrial planets up to the tropopause. In the case of Jupiter the situation is different, as strong winds have been measured in the deep troposphere by the Galileo probe; the circulation in this case is also influenced by the existence of the internal heat source.

The rapidly rotating planets Mars and Earth, as well as the giant planets, are characterized by geostrophic dynamics, with low values of R_0 (0.1–0.5). In contrast, for Venus and Titan, which have very slow rotation, R_0 is in the range 10-100. Venus exhibits a strong superrotation, and it seems also to be the case for Titan.

The exploration of planetary atmospheres

How can we study planetary atmospheres? There are two distinct classes of methods. The first one, used for centuries, is ground-based telescopic observation; the second, which started in the 1970s, is *in situ* exploration from space probes.

Observations using ground-based telescopes consist first of images, which allow one to study the atmospheric cloud structures and to monitor their evolution with time, an essential tool for climatology studies. The main technique used for remote sensing is spectroscopy: spectra recorded at different wavelengths, from the ultraviolet to the millimeter range, allow us to probe different altitude levels of the planetary atmospheres (depending upon the atmospheric composition). Infrared spectra of planetary atmospheres, in particular, are essential for determining their chemical composition and their thermal profile. Another powerful method is the photometric observation of stellar occultation events: the stellar flux decreases at the time of the OCCULTATION and increases as the star emerges from the planet's atmosphere. The study of the lightcurve provides valuable information about the refractive index of the atmosphere, which allows us to determine the temperature if the mean molecular weight is known; this technique has been successfully used in the case of Uranus, Neptune and Titan.

With the space era, we now have access to direct measurement of the physical parameters and the chemical composition of planetary atmospheres. Sensors are used for direct measurements of the pressure and the temperature, and mass spectrometers provide the chemical analysis of the atmosphere. The first space probes devoted to PLANETARY EXPLORATION were flyby or orbiter spacecraft, equipped with remote sensing instruments (cameras and spectrometers): then landers were sent for *in situ* analysis of the atmosphere and the surface. Some of these missions have been especially fruitful: VIKING on Mars (launch in 1974, operation in 1976–7), VOYAGER toward the giant planets (launch in 1979, operation in 1979–89), and, more recently, GALILEO toward Jupiter (launch in 1989, operation in 1995–2000). In parallel, after the launch of Mars Pathfinder and Mars Global Surveyor, the exploration of Mars is the object of an extensive on-going space program including, in forthcoming years, the planned launch of several planetary probes.

For the terrestrial planets, most of our knowledge has

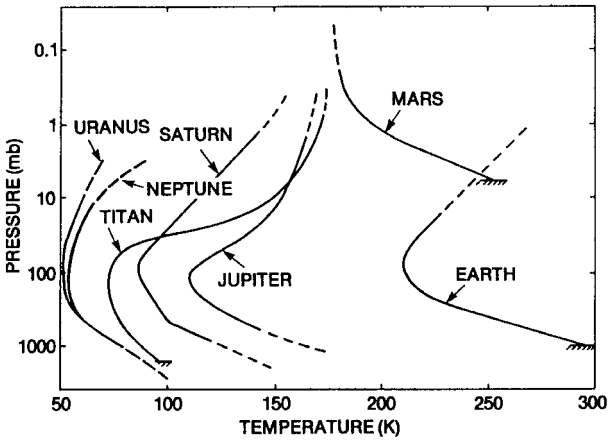


Figure 1. Vertical temperature profiles as a function of pressure for the planets (except Venus, out of scale) and Titan. The figure is taken from Samuelson (1997).

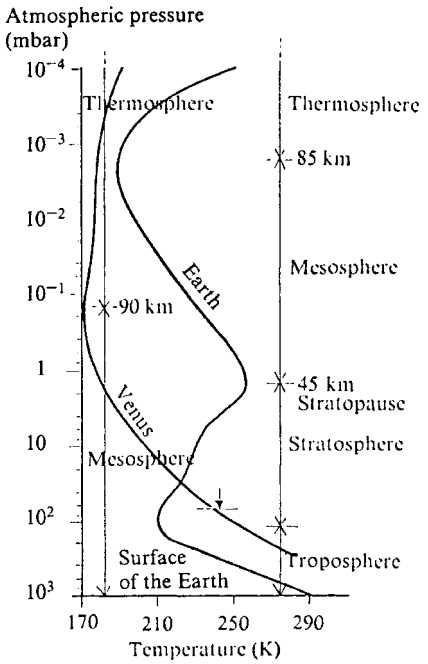


Figure 2. Temperature profiles in the atmospheres of the Earth and Venus. The pressure scale is common to both planets. The altitude scales refer to the Earth (left) and Venus (right). On Venus, the surface temperature is 730 K and the surface pressure is 94 bar. The figure is taken from Encrenaz *et al* (1995).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

come from space exploration. The atmosphere of Venus was extensively explored in the 1970s and 1980s by the Venera and Pioneer Venus missions with, in addition, the Galileo flyby (1990); its surface has been mapped by the radar of the Magellan mission (1991–3). Our knowledge of Mars mostly relies on the results of the Viking mission which, with its two orbiters and landers, studied the Martian atmosphere and surface over a full Martian seasonal cycle (2 years). For both Mars and Venus, the space missions provided most of our knowledge about the atmospheric composition, the thermal profile and the properties of the clouds, as well as their spatio-temporal variation.

In the case of the giant planets, our knowledge about the chemical atmospheric composition and the cloud and thermal structure first came from ground-based observations, especially in the infrared range, until the major achievement of the Voyager mission, which followed the precursor approach of the Pioneer 10 and 11 spacecraft in the 1970s. Voyager was especially successful in revealing the thermal structure and the chemical composition of Titan's atmosphere, largely unknown before. Finally, the Galileo mission, with the first launch of a probe in the atmosphere of a giant planet, has opened a new era in the space exploration of the outer solar system, with an extended monitoring of the planetary system. The next step of this program will be the CASSINI MISSION, which includes the descent of the Huygens probe into Titan's atmosphere and the 4-year monitoring of the whole Saturnian system by the Cassini orbiter.

The origin of planetary atmospheres

In order to understand the formation of planetary atmospheres and their subsequent evolution, we need to know how the planets themselves formed. The history of the solar system is now understood in its basic concept, and generally accepted among the scientific community. Like other stars and stellar systems, the Sun and the solar system generated from a mass of interstellar gas, called the primordial nebula. The chemical composition of this gas must have been close to the composition generally found in the universe, i.e. mostly hydrogen (about 75%) and helium (25%), with all other elements (C, O, N,...) with their cosmic abundances, i.e. below 1%. Due to some instability (for example, the explosion of a nearby star), this primordial nebula starts rotating and contracting. At some point, the rotation becomes fast enough for the gas to collapse into a rotating disk. The central part will become the Sun, and the planets will form from the outside disk. We now observe some examples of such PROTOPLANETARY DISKS outside the solar system, around young nearby stars, so this formation process seems to be common in the universe.

As the temperature cools down, the matter of the disk, mostly gaseous at the time of the collapse, combines from atoms to molecules and starts to condense (minerals and silicates first, then the H_2O ice, then the other ices like CO_2, NH_3, CH_4, etc). Some molecules like CO and H_2 stay in gaseous form because of their low condensation temperatures. These are the bricks which will form the future planets, through the process of repeated collisions, leading to the formation of PLANETESIMALS. These small particles will continue to collide with each other and to stick together, giving rise to a larger and larger body. The protoplanet grows through the repeated accretion of infalling bodies, finally acquiring, for diameters larger than about 100 km, a spherical shape as a result of self-gravity.

The major differences that we see today in the composition of the various planets and their atmospheres can largely be explained, at least to first order, by an effect of the condensation sequence which takes place as the temperature decreases from the Sun to the outside edge of the protoplanetary disk. This condensation sequence may have been the following: in the vicinity of the Sun, at heliocentric distances smaller than about 2 astronomical units (AU), only the heavy elements (minerals, silicates, etc) could be found in solid form. These elements were used in the planetesimals which formed the terrestrial planets. This explains why these planets are relatively dense and small. In contrast, at large heliocentric distances (at 5 AU and beyond) most of the molecules are in solid form, especially the molecules made with the most abundant elements H, O, C and N: H_2O, CO_2, CH_4, NH_3. Molecular hydrogen H_2, however, is still in gaseous form.

As a result, the cores formed by the ACCRETION process were much larger than in the case of the terrestrial planets, since, following the cosmic abundances, the mass contained in the ices is by far larger than the mass contained in the heavier elements. Theoretical models estimate that the cores of the protoplanets formed at 5 AU and beyond may have been about 15 times the mass of the Earth. At this stage, a qualitative difference appears in the evolution of the two classes of planets. Indeed, cores of 15 terrestrial masses have a gravity which is sufficient to accrete the surrounding nebula, mostly composed of gaseous hydrogen and helium; in the case of the terrestrial planets, the gravity field is not sufficient for this accretion to take place. This scenario explains why the giant planets are so massive (Jupiter's mass is over 300 times the Earth's one), mostly gaseous, and mostly formed of hydrogen and helium.

At the heliocentric distances of Uranus and Neptune, the matter available for the core accretion must have been less abundant than close to the Sun; as a result, the time needed to accrete a core of 10–15 terrestrial masses must have been significantly longer. It is possible that in the meantime a large fraction of the light elements of the disk were blown away during the T-Tauri phase of the young Sun. In this case, Uranus and Neptune had little gas available in the surrounding nebula to be accreted around their cores. This would explain why Uranus and Neptune are smaller than Jupiter and Saturn, and why the relative mass fraction of their cores (about 60 to 80%) is much higher than for Jupiter and Saturn (3 to 15%).

In the case of the terrestrial planets, the process of formation must have been very different. It is generally

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

believed that these atmospheres have been at least partly outgassed from the interior of the planets by active volcanism. It has been suggested, in particular by Owen, that part of their atmosphere might have been brought from the exterior through the infalling flux of meteorites (see Atreya *et al* 1989); a significant part of these meteorites must have been of cometary origin, i.e. rich in volatiles (H_2O, CO_2, etc). As will be discussed below, the recent measurement of the deuterium abundance in comets provides a constraint to this hypothesis. However, because the terrestrial planets are much less massive than the giant planets, they could not retain in their atmospheres the lightest elements like H_2 and He. Only heavier molecules like CO, N_2, O_2 and CO_2 could stay within the lifetime of the solar system.

Chemical composition of planetary atmospheres

In view of their chemical composition, planetary atmospheres also fall into two distinct classes. The terrestrial planetary atmospheres are mostly composed of CO_2 and N_2, with variable amounts of H_2O, O_2 and traces of CO and O_3. In contrast, the giant planets' composition includes, apart from hydrogen and helium, CH_4 and NH_3, with traces of many other minor constituents.

In which form were carbon and nitrogen in the primordial nebula and the protoplanetary disk? Calculations of thermochemical equilibrium indicate that C is in the form of CH_4 or CO, and N is in the form of NH_3 or N_2, according to the following reversible chemical reactions

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2$$

$$2\,NH_3 \rightleftharpoons N_2 + 3\,H_2$$

which evolve toward the formation of CH_4 and NH_3 at low temperature, and the formation of CO and N_2 at high temperature. To first order, this favors the formation of CH_4 and NH_3 in the cold and dense environment of the giant planets where thermochemical equilibrium can effectively occur (this may not be the case in dilute media, like the interstellar medium or the medium from which comets formed). In the outer solar system, this scenario, however, does not explain the origin of the atmospheres surrounding the small bodies, Titan, Triton and Pluto, all dominated by nitrogen with a small fraction of methane. As for the terrestrial planets, these atmospheres are 'secondary', i.e. probably outgassed from the cores, with some possible contribution from meteoritic impacts. The presence of both N_2 and CH_4 on Titan, Triton and Pluto is not fully understood at present. In the case of Titan, formed within the Saturnian sub-nebula, the atmosphere may have originally contained CH_4 and NH_3, ammonia being subsequently converted into N_2 by photolysis, and H_2 escaping in the process. The case of Pluto and Triton is different, as they were not formed in a dense environment, and they may have accreted like comets in a dilute medium far from equilibrium.

In the vicinity of the Sun, the reactions evolved toward CO and N_2. CO in turn reacts with H_2O according to the chemical reaction

$$CO + H_2O \rightarrow CO_2 + H_2.$$

This favors the formation of CO, CO_2, N_2 and H_2 in the terrestrial atmospheres. Hydrogen escapes, as it is too light to be trapped by the planetary gravity field, and the planets are left with an atmosphere of N_2, CO_2, CO and H_2O in various amounts. As mentioned above, the presence of a large fraction of O_2 on Earth is due to the presence of life.

The relative evolution of the telluric planets

By which processes did Venus, the Earth and Mars evolve toward their extreme present conditions, while they started from more or less similar initial conditions? The three planets indeed are in the same range of heliocentric distances, diameters and densities, yet exhibit extreme differences in their present surface temperatures and surface pressures. In addition, water has almost completely disappeared from the atmosphere of Venus and Mars, while large amounts are present in the terrestrial oceans. How was such an evolution possible? Several authors, including Pollack, Hunten and Donahue (see Atreya *et al* 1989), have tried to rebuild the history of the three planets on the basis of the data available now: physical parameters, chemical composition, abundances of rare gases, isotopic ratios, etc. Although the question is still not completely solved, it is generally believed that two factors were especially important: the difference in heliocentric distance and the variation in the planetary mass.

Although not more than a factor 2, the change in heliocentric distance implies a significant variation of the equilibrium temperature at the corresponding orbit. This temperature change is such that a key atmospheric element, water, tends to take different states on the three planets: mostly gaseous in the Venusian atmosphere, liquid at the Earth's surface and solid on Mars. We know that water was very abundant in the early stages of Venus' history, through the measurement of the deuterium-to-hydrogen ratio, derived from the HDO/H_2O ratio. The abundances of HDO and H_2O have been measured from their infrared signatures in the spectra of Venus and Mars. As previously suggested from the Pioneer Venus measurements in 1982, the D/H ratio in Venus today, as derived from HDO/H_2O, is 120 times higher than the terrestrial value. In the case of Mars, it is 6 times higher than the Earth value. The explanation for HDO enrichment is the following: deuterated water—or heavy water, HDO—escapes slightly more slowly than H_2O; over the age of the solar system, this process leads to an accumulation of deuterium and an enhancement of the D/H ratio. The very high D/H ratio in Venus today is the signature of a very strong water outgassing over the history of Venus; the smaller, but still significant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

enrichment of deuterium in Mars today indicates a moderate outgassing over the history of Mars. In the case of the Earth, no outgassing occurred since the water is trapped in the oceans. It is worth noticing that the comparison of the D/H ratio in the Earth's oceans and in comets provides an important constraint upon the origin of the atmosphere of the Earth. As recently measured on COMET HYAKUTAKE and COMET HALE–BOPP, the D/H ratio derived from H_2O and HDO in comets is twice as high as its value in the terrestrial oceans. There is no mechanism which could explain a deuterium depletion within the oceans over the history of the Earth. As a consequence, the water on Earth cannot come entirely from comets; about at least half of it must have been outgassed from its interior.

In the case of Venus, the presence of gaseous water in massive amounts must have been important in increasing its surface temperature, through the greenhouse effect: H_2O, like CO_2, is a gas which is spectroscopically very active in the infrared. This process is an amplifying one and is believed to be responsible for the very high surface temperature of Venus observed today (see VENUS: SURFACE). How did water escape from Venus? Several processes have been proposed, including dissociation by solar UV radiation, or reaction with the surface, or reaction with the atmospheric CO, but the question is not fully answered yet, and will require the development of more sophisticated photochemical models.

In the case of Mars, the situation is completely different. Due to the lower temperature, the abundance of gaseous water is severely constrained by saturation and the greenhouse effect has to be much less significant. There was some time in Mars' history, however, where the atmosphere must have been significantly more dense than today, as demonstrated by the signs of intense flooding shown on the surface. How did Mars lose its atmosphere, and how did water disappear? First, because of the smaller mass of the Martian globe as compared with Venus and the Earth, volcanism (whose energy is generated by the radioactive elements of its core) must have stopped earlier, thus interrupting a major contribution to the origin of the atmosphere. In addition, CO_2 and N_2 may have been trapped in the surface in the forms of carbonates and nitrates respectively, but these constituents have not been firmly detected yet. Nitrogen might also have partly escaped gravitationally. Water might be present in the form of permafrost under the Martian surface, as also observed in some places on the Earth. Discovering such water reservoirs under the Martian surface would be of major importance for understanding the whole history of the Martian climate, and this research will be a main objective of the future space exploration of Mars (see MARS: SURFACE).

Finally, the evolution of the terrestrial atmosphere follows a third different scenario: the primitive atmosphere of the Earth probably contained large amounts of CO_2, which would have been able to generate a greenhouse effect comparable to the one on Venus. However, the presence of liquid oceans of water prevented this fatal evolution. Carbon dioxide was trapped in the oceans and reacted with the rocks to form carbonates ($CaCO_3$). The amount of gaseous CO_2 consequently decreased and the greenhouse effect remained limited, so the surface temperature was never high enough for the oceans to evaporate.

Dynamics and circulation of planetary atmospheres

All planetary atmospheres show evidence for temporal variations, on very different time scales, diurnal and seasonal. In the case of the terrestrial planets, the climate is, as in the case of the Earth, dominated by the incoming solar flux and its variation as a function of latitude. The dynamics of the giant planets are more complex, as they are driven both by the incoming solar flux and the internal heat source.

The case of Mars is of special interest as it shows striking analogies with the case of the Earth. The inclination of Mars' rotation axis on the ecliptic is 25°, which generates strong seasonal effects. Both CO_2 and H_2O can condense on Mars. These constituents condense on the northern and southern caps during winter, and partially sublimate (especially CO_2) to recondense on the opposite pole. After the pressure measurements achieved by the two Viking landers over a full seasonal cycle (1976–7), it was realized that this process involves as much as 25% of the mass of the total atmosphere. This enormous variation in the atmospheric mass induces strong meridional transport of heat and momentum; it is also responsible for the transport of water from one pole to another, the permanent CO_2 southern polar cap acting as a cold trap for water vapor. These transport phenomena are associated with violent winds which can generate huge dust storms, lifting particles as high as 50 km. The dust in turn influences the Martian meteorology by increasing the opacity of the atmosphere, heating the atmosphere and/or acting as nuclei for CO_2 or H_2O condensation. The Martian climate is thus the complex product of various seasonal cycles—carbon dioxide, water and dust—which are closely interconnected.

It should be mentioned that another solar system body exhibits climatic changes even stronger than those of Mars, but still with some analogies: Triton, Neptune's largest satellite, has a subsolar point which varies from +52 to −52° in latitude. Its very tiny N_2 atmosphere, constrained by saturation, is subject to seasonal variations in abundance which are probably much stronger than in the case of Mars. On Triton there is probably a transfer from pole to pole comparable to the one on Mars, with N_2 and CH_4 on Triton playing the roles of CO_2 and H_2O respectively on Mars.

Jupiter and Saturn show a cloud structure along parallel belts and zones, which have been monitored by astronomers over the last three centuries. These striking features result from the fast rotation of the giant planets (about 10 h), associated with convective atmospheric motions. After the Voyager investigations—and more recently after the exploration of Jupiter by the Galileo probe—it is now generally believed that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

zones are cloudy regions of ascending motion, while the belts, more free (if not completely free) of clouds, are regions of subsidence. This type of circulation is similar to the Hadley circulation on Earth, which moves the humid air upward at the equator and brings back dry air downward at the tropics. The space results on the Jovian circulation are also supported by ground-based observations, especially in the infrared range, which show that the belts are associated with 'hot spots': the infrared radiation is stronger in these regions, because it comes from deep atmospheric levels where the temperature is higher. The latest Galileo observations do confirm this interpretation, but also show that the reality is much more complex that this simple scheme. The dimension scale of the hot spots is apparently very small, and the circulation is more complex by far than the simple zone/belt scheme. In addition to the latitudinal variation of the incoming solar flux which governs the climate in the terrestrial atmospheres, one has to take into account the effect of the internal energy which, for all giant planets except Uranus, is comparable to the absorbed solar energy.

Another striking feature of the dynamics of the giant planets is the presence of cyclonic or anticyclonic features which consist of columns of ascending motion rising above their immediate surroundings and which are colder than their environment. The most famous of these features is the GREAT RED SPOT (GRS), in the southern hemisphere of Jupiter, which has been observed over more than three centuries, since the early days of planetary telescopic observations. The stability of this feature is still an enigma. Many structures of this kind exist elsewhere, on the giant planets but also on Earth (they are equivalent to our giant hurricanes), but these structures have short time scales. In particular, a similar feature, the dark spot, was discovered on Neptune at the time of the Voyager 2 flyby, in 1989; however, recent observations by the HUBBLE SPACE TELESCOPE, over the past 2 years, have shown that this feature has now disappeared. Most of the similar features on Jupiter—the so-called 'white ovals'—are also variable with time.

What is the nature of the red material responsible for the Great Red Spot, and what is the constituent which makes the Jovian belts brown? We still do not know. The red color could be due to phosphorus or sulfur compounds, but there is at present no identification of them. More spectroscopic observations of these regions in the infrared and the ultraviolet range, from the Galileo orbiter or with the Hubble Space Telescope, might help to determine their chemical nature. We better understand the white ovals and plumes, most likely associated with ammonia clouds. These NH_3 clouds are also likely to be responsible for the yellow color of Saturn. The greenish color of Uranus might be due to the presence of methane. In large abundances, gaseous methane is responsible for the intense blue color of Neptune, while the white spots are methane cirrus.

The small bodies of the outer solar system

One planet—Pluto—and three satellites—Titan, Triton and Io—are surrounded with a stable atmosphere. Apart from Io, they are all composed of nitrogen and methane; apart from Titan, they are all tenuous, with a surface pressure of a few microbars (Triton and Pluto) or even a few nanobars (Io).

Io, Jupiter's closest Galilean satellite, is another unique object in the solar system: it is the only body showing evidence for active volcanism. The strong tidal forces generated by Jupiter's gravity field, coupled with the eccentricity of Io's orbit and the perturbations due to EUROPA, induce a permanent remodeling of Io's surface. The active volcanoes of Io were first visually identified by Voyager, which also detected a plume of sulfur dioxide above one of them. The evidence for a stable atmosphere came from millimeter observations which detected transitions of SO_2 (its major constituent) and SO. The atmosphere of Io was found to be permanent but variable with time and localized, either produced by SO_2 frost sublimation or by direct volcanic outgassing.

Titan's atmosphere was largely unknown before its exploration by the Voyager spacecraft. Not only did this mission reveal the nature of its atmosphere—primarily nitrogen—and its thermal structure, but the Voyager infrared spectrometer also detected a large number of unexpected complex molecules, mostly hydrocarbons and nitriles. These species are the products of methane photodissociation—the second atmospheric constituent in term of abundance—and the dissociation of nitrogen by energetic particles associated with the magnetic field of Saturn. The fascinating aspect of this discovery is that the complex molecules which were found are exactly those which appear in the laboratory in simulation experiments aiming to form HCN polymers and amino acids from the irradiation of CH_4, H_2O and NH_3 ices. Could a prebiotic life have developed on Titan? Although the presence of a surface, and perhaps even lakes, may favor the development of heavier molecules, the low temperature is definitely a handicap that slows down all chemical reactions considerably. Nevertheless, the study of Titan's atmosphere could possibly shed some light on our understanding of the early stages of the history of our own atmosphere, and this is among the major motivations for the Cassini mission.

Pluto and Triton show striking similarities in their physical properties (diameter and density) but also in the composition of their atmosphere. Pluto's orbit is highly elliptical with a period of 248 years. For a period of 20 years, its heliocentric distance is very close to 30 AU, which makes its incoming solar flux similar that one of Triton. This probably explains why they both at present have a surface pressure of a few microbars, mostly composed of nitrogen, with traces of methane; the surface pressure is apparently constrained by N_2 saturation.

Bibliography

Atreya S K, Pollack J B and Matthews M S (eds) 1989 *Origin and Evolution of Planetary and Satellite Atmospheres* (Tucson, AZ: University of Arizona Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Del Genio A D 1997 Atmospheres *Encyclopedia of Planetary Sciences* (London: Chapman and Hall)
Encrenaz T, Bibring J-P and Blanc M 1995 *The Solar System* (Berlin: Springer)
Hanel R A, Conrath B J, Jennings D E and Samuelson R E 1992 *Exploration of the Solar System by Infrared Remote Sensing* (Cambridge: Cambridge University Press)
Samuelson R E 1997 Atmospheric thermal structure *Encyclopedia of Planetary Sciences* (London: Chapman and Hall)

Thérèse Encrenaz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Exploration

Nine known planets, a large number of natural satellites, planetary rings, comets and asteroids are the constituents of the solar system. Knowing when and how they all formed, understanding the physical and chemical processes that shaped the planets and their moons as well as their relations with the Sun and the wider cosmos are the reasons why we want to explore the solar system. However, the fulfillment of the above scientific goals requires efforts that have embraced, and still do, large sectors of our society. The development of new technologies, more and more sophisticated and miniaturized, is essential for the progress of the exploratory programs, but such development requires investments which can only be supported at the highest political level. That is why the history of the exploration of the solar system is, in a way, the history of the post-world-war-II society.

The first steps to the Moon

It all started with the launch by the USSR of the Luna 1 spacecraft to the MOON, on 2 January 1959. It accomplished its mission successfully and will be remembered as the first man-made craft to leave the Earth's gravity field. The Soviet Union launched, on 12 September 1959, Luna 2, which impacted the lunar surface as planned, and, on 4 October of the same year, Luna 3, which circumnavigated the Moon and sent back the first images of the far side. The effort carried out by the USSR to develop reliable launchers had paid off and put the Soviet planetary program in a leading position for about 5 yr. Indeed, it was only on July 1964 that the US Ranger 7 reached the Moon and sent back thousands of images of that part of the surface named Mare Cognitum. Such a late success of a NASA spacecraft gives to the landing of US astronauts on the surface of Moon, only 5 yr later, a miraculous touch. In fact, the first six Ranger missions had failed for a wealth of technical reasons. The Soviet Union, meanwhile, had launched from 1959 throughout the 1960s a total number of 17 fully successful missions to the Moon.

It is common understanding to consider the landing of Apollo 11 and its crew on the surface of the Moon as the apex of lunar exploration. This landing is also considered as the final victory of the technology of the US over that of USSR. However, we tend to forget that already in 1970 the USSR had been able to return samples from the Moon's surface with a fully automated mission (Luna 16) and that with a periodicity of a mission every other year, until 1976, lunar sample return missions were carried out successfully. NASA landed its first automated rover on Mars in 1996 and we will have to wait until 2003 to see a European lander on another planet.

In the US, between 1967 and 1969, an incredible number of missions to the Moon were successfully launched in preparation of the human landing. The data acquired during those years constituted the main database, together with the Russian data, to be used by the scientists to understand the geology, the geophysics and the geochemistry of the Moon. Five other landings followed the first human landing in 1969, the last of them being the mission of Apollo 17 in 1972. The US had demonstrated, in the course of their successful human landings, their technological superiority and, in particular, their capacity to define a goal that involved politically, scientifically and technologically the entire nation. On the contrary, the hyperprogrammed USSR had failed in coagulating around the conquest of the Moon the different interests of the nation. Contrasting industrial, political and technological interests did not favor the accomplishment of the ultimate goal: landing a Soviet cosmonaut on the surface of the Moon.

New data were made available to the scientific community almost 20 yr later by the CLEMENTINE Mission in 1994 and the LUNAR PROSPECTOR Mission in 1998. The Clementine spacecraft was developed by the US Department of Defense in order to test both new microtechnologies and management procedure for the faster construction of smaller spacecraft. It provided us with extensive maps of the lunar surface, including some multispectral maps. Lunar Prospector, launched on 6 January 1998, was to determine mainly surface chemistry and the lunar gravity field. The most noticeable result acquired by this last lunar mission is the determination of large quantities of hydrogen in the cold lunar polar regions. Some scientists speculated from this result that water ice might be present near the poles of the Moon; however, such a presence has yet to be proven, even after the impact of Lunar Prospector on the surface on 31 July 1998.

The major scientific result of the exploration of the Moon is that we now have a well-measured time scale of its evolution, thus providing us with a comparative basis to interpret the evolution of other planetary bodies, including the Earth. The larger CRATERS on the Moon record a period of intense bombardment which ended about 3.7 billion years ago. Presumably, this is a process that affected all the inner planets more or less at the same time. By counting and measuring the lunar crater density, and by determining accurately the age of the lunar samples with radioisotope dating techniques, we have been able to construct a chronological reference which allow us to interpret the geological history of Mars, Mercury and Venus.

The analysis of the returned samples has revealed that two major provinces constitute the lunar crust: the young, sparsely crated and low-albedo mare terrains, and the old, heavily cratered, high-albedo highlands. Isotopic dating has shown that the youngest mare basaltic lava flows have an age of about 2.5 billion years; the oldest-dated lunar rocks have been found in the lunar highlands and have shown a radiometric age of about 4.5 billion years. Geophysics of the Moon, i.e. its internal constitution and state, is still the subject of intense debate within the scientific community. We know that the Moon had a strong magnetic field during the early phases of its history and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that this field had disappeared since. We still do not have clear data and understanding of the internal structure and the presence of a small molten core still needs to be proven definitively. Moreover, all the available geochemical and geophysical data refer to the near side of the Moon and there is still a need to obtain data from the far side in order to complete our geological and geophysical picture of our natural satellite.

Exploring the void: the interplanetary missions

The space within which the solid constituents of the solar system revolve around the Sun is far from being void and knowledge about it is essential to understand the 'environment' of that part of the universe where we live. Indeed, the particles emitted by the Sun, as well as the electric and magnetic fields associated with such emissions, have a strong influence on the planetary bodies and on the spacecraft supposed to explore them. It goes without saying that manned flights outside the Van Allen belt required detailed information on the possible natural hazards the astronauts were supposed to face.

At the beginning of the space age a good number of spacecraft which missed their primary scientific goals and became lost in the interplanetary space provided interesting and novel information on the solar system 'environment'. The PIONEERS 1, 2, 3 and 4 launched between 1958 and 1959, belong to this category. They were indeed supposed to flyby the Moon and reach a final orbit around the Sun, but never accomplished their primary goals. Nevertheless, these missions were the first to show that human beings could have reached the Moon without major problems due to the interplanetary radiation. The following series of Pioneers (6–9), launched between 1965 and 1969, were successfully put in an orbit around the Sun similar to that of our planet and have provided us for many years with vital information on the interplanetary fields and particles and on the Sun itself.

In 1990 a US shuttle launched the ESA-made Ulysses spacecraft. No other name would have been more appropriate for this spacecraft which is still travelling along an out-of-the-ecliptic orbit which takes it periodically above the poles of the Sun.

Ulysses is one of the many spacecraft dedicated to the study of the interplanetary space in the framework of the International Solar Terrestrial Physics Program (ISTP). This program includes a large number of spacecraft mostly provided by the Russian Space Agency (RKA, recently renamed the Russian Aviation and Space Agency, Rosaviakosmos), the Japanese Institute for Space Science (ISAS), NASA, NOAA and ESA (ACE, AKEBONO, ASTRID, CLUSTER, DMSP, EQUATOR-S, ETS-VI, FAST, FREJA, GEOTAIL, GMS-4, GOES, GOMS, IMP-8, INTERBALL-TAIL, INTERBALL-AURORA, LANL S/C, MSX, NOAA/TIROS, OERSTED, POLAR, PRACTICE-4, ROCSAT, SAMPEX, SOHO UARS, WIND).

The exploration of the inner region

Taking for granted that exploration of our planet can be neglected in this context, and having described previously the exploration of the Moon, we are left with two planets: MERCURY and VENUS.

All the planetary bodies orbiting the Sun in a belt between 0.4 and 1.5 astronomical units from the Sun, i.e. Mercury, Venus, the Earth–Moon system and Mars, are mainly composed of rock and metal and are characterized by a relatively small amount of volatiles. A trend of decreasing density from Mercury to Mars is also a noticeable characteristic of the inner planets. This trend could be determined by the thermodynamical behavior of the PROTOPLANETARY NEBULA in which pressure and temperature decrease with increasing distance from Sun. However, it could also be due to local anomalies in the course of the fractionation and re-aggregation of these bodies during the early phases of their collisional accretion. The comparative study of the formation and evolution of the inner planets provides us with a better understanding of the Earth as well of those parameters which determine planetary evolution.

Mercury

What we know of Mercury is due to the NASA mission, MARINER 10, which, launched in 1973, flewby the planet three times. This spacecraft will be remembered as the first one to make use of the 'gravity assist' technique thanks to the brilliant calculations of the Italian scientist Giuseppe Colombo. Indeed, the gravitational attraction of Venus was used to slow Mariner 10's solar orbital velocity down in order to let the spacecraft 'fall' towards the Sun and reach the orbit of Mercury. In the course of its three flybys, Mariner 10 was able to return about 12 000 images, corresponding to a coverage of about 40% of the surface of the planet. The Mariner 10 images revealed a densely cratered surface, dominated by the Caloris Basin, a huge multi-ringed impact crater. Caloris appears to have been filled by lava flows similarly to the Lunar Maria.

The little information we have on Mercury is pushing the major world space agencies to plan new missions to this planet. In particular ESA has identified Mercury as the target of one of its 'Cornerstone' missions, with a launch date in the 2008–10 time frame. NASA and ISAS have also manifested interest in exploring Mercury. At the time of writing an effort of international coordination, led by ESA, is ongoing.

Venus

As for the exploration of the Moon, the Soviet Union had taken the lead in the exploration of Venus. In 1967 the Soviets, after three initial failures, succeeded in letting the VENERA 4 probe descend through the atmosphere of Venus. Shortly afterwards the Venera 5 and 6 probes repeated the atmospheric measurements providing us with the first extended database of '*in-situ*' atmospheric measurements. During the 1970s the Soviets increased the ambition of their program of exploration of Venus by carrying out eight

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

landings during which the only images of the surface still available today were obtained. In the early 1960s NASA was facing a number of technical problems and failures with its first missions to the Moon and did not have many resources to invest in the exploration of other planets. Nevertheless, they had succeeded in sending Mariner 2 to Venus in 1962 (the first probe ever to visit another planet) and in 1967 Mariner 5, which succeeded in measuring the crushing atmospheric pressure at the surface.

The Mariner 10 and PIONEER VENUS (PV) 1 and 2 missions, launched in 1975 and in 1978 respectively, were real successes. The first mission returned 3500 images of the complex atmospheric structure of Venus characterized by a Hadley cell type of circulation as proposed already in 1735 by the eponymous scientist. The missions of the two Pioneers completed our understanding of the Venusian upper atmospheric processes (PV 1) and of the deep structure (PV 2) thanks to the four probes deployed inside the atmosphere.

The complex geology of the planet was revealed by the two Soviet missions Venera 15 and 16 and by the NASA MAGELLAN mission. In fact, these spacecraft made use of imaging radars capable of penetrating the thick atmosphere of Venus and returning images of the surface. Venus's surface is characterized by a large variety of structures: volcanoes, valleys, depressions, faults, plateaus, mountain ridges. Volcanic activity appears to have played a major role in the evolution of the surface, but, strangely enough, Venus's crust seems to be much more static than that of the Earth. Magellan is a typical example of a space saga. Originally called VOIR (Venus Orbital Imaging Radar), this mission was to be launched at the beginning of the 1980s. For financial reasons it was first cancelled and then resurrected under the name of Magellan. However, also the new launch date of 1984 could not be maintained owing to the delays of the shuttle program. Eventually, Magellan was launched in 1989, the first planetary spacecraft launched on a shuttle.

Exploring the outer regions

It is difficult to understand why the Soviet Union was completely absent from the exploration of the outer regions of the solar system: was it lack of interest, technological problems in operating the spacecraft at large distances from the Earth or shortage of funds? Perhaps a combination of all these reasons, plus the large investments made in the manned orbital infrastructure, explains such a strange absence. What is certain is that the missions to the giant planets are made only in the US, with the exception of the CASSINI MISSION that includes major participation by ESA and ASI, the Italian space agency.

Pioneers 10 and 11 were the first missions to be launched towards the external regions of the solar system. Very small and simple spacecraft, they have shown an incredible reliability which has allowed them to survive the hostile environment of JUPITER and SATURN. The most outstanding result of the Pioneer 10 mission was the discovery of an intense magnetic field tilted 11° from the axis of rotation of Jupiter which created wobbling radiation belts around the planet. Pioneer 11 was targeted at Saturn where it discovered very intense magnetic field radiation belts interacting with the dusty rings and a very structured magnetosphere.

Between 1977 and 1989 the NASA VOYAGERS 1 and 2 transformed our understanding of the outer solar system. These planetary probes were designed and launched to exploit the rare opportunity (available every 176 yr) to explore the outer regions of our planetary system: the alignment of the outer planets. In the course of a mission lasting about 13 yr the two Voyagers explored all the gaseous giants, Jupiter, Saturn, Uranus and Neptune. These spacecraft have been able to transmit telemetry and science data from distances of hundreds of millions of kilometers with an emitting power of a few tens of watts. When Voyager 2 reached Neptune the radio signals emitted by the on board antenna took more than 4 h to reach us.

Jupiter was the first planet to be visited by the two spacecraft. They reached this planet following different trajectories characterized by different cruise times. The spacecraft orbits were such that Voyager 1 could obtain the closest approach to the Galilean satellites of Jupiter. Voyager 2 was kept further away from Jupiter in order to minimize the interaction with its radiation belts and preserve it in the best possible conditions in view of the long voyage to Neptune. The acquisition of science data was extended to all phases of exploration of the Jovian system. This has allowed the spacecraft to measure the temporal variability of Jupiter's atmosphere and magnetosphere, of the plasma torus associated with IO's orbit and of the large-scale morphology of the satellites. Among the many results obtained during the long cruise phases it is worth remembering the discovery, at a distance of about 130 000 km from the center of Jupiter, of a 30 km thick dust ring , the measurement of a 5 million volt electric current that flows along the magnetic flux lines between Jupiter and Io and the generation of radio signals probably due to the oscillations of the plasma cloud associated with Io.

After one more year of navigation the Voyagers reached Saturn, which is at an average distance from the Earth of about 1.6 billion kilometers. At that distance, the solar radiation is only 1% of the amount that reaches our planet. The prominent ring system of this planet was eventually studied in great detail as well as TITAN, the only satellite in the solar system characterized by a thick atmosphere that resembles that of the Earth in its early phases. A completely unexpected feature of the ring system was the presence of transient radial spokes due to levitation of dust grains above the plane of the rings. The images of the icy satellites have shown that they all have been subject to an intense meteoritic bombardment. However, noticeable differences in the crater density were detected among the satellites, suggesting the possibility of a non-uniform distribution of the impactors and/or

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

geological reprocessing of the surface, effective just after the most intense phase of the meteoritical bombardment.

The exploration of the Saturnian system concluded the 'nominal' part of the Voyagers' mission in 1981. Immediately a political debate arose on the opportunity to turn the spacecraft off and terminate the funding of the mission. Fortunately, NASA succeeded in obtaining approval for an extended mission, which consisted in the targeting of Voyager 2 towards Uranus and Neptune. In January 1986 Voyager 2 reached URANUS, the only planet of the solar system with a spin axis parallel to the plane of the ecliptic and pointing towards the Sun. Uranus's magnetic field axis is tilted by about 60°. This suggests the hypothesis that the planet's magnetic field might be generated at relatively shallow depths. The images of Uranus returned by Voyager appear particularly sharp thanks to the reprogramming of both the attitude control of the spacecraft and the pointing of the imaging system which now was capable of locking on the target and compensating the motion of the spacecraft with appropriate rotations. It is thanks to this technique that the discovery of Uranus's satellite 1985U1 became possible.

After three more years of interplanetary cruising, Voyager 2 reached NEPTUNE. The scientific payload was still working nominally more than 12 yr after launch. In order to receive in the best possible way the images transmitted by the spacecraft from the borders of the solar system, the antennae of the Deep Space Network were all strung together in order to create a 'virtual' antenna of a much larger size. The images of Neptune show a blue planet, with cloud patterns similar to those of Jupiter, surrounded by rings segmented in arcs and only thinly connected. The discovery of active geysers and plumes on the surface of TRITON, a retrograde satellite of Neptune, was a great surprise.

The two Voyager spacecraft, which are now flying in the depths of interstellar space, have accomplished flawlessly a fantastic mission of exploration to be remembered as one of the brightest successes of the US space technology.

In 1989 the exploration of Jupiter started again with the launch of the GALILEO MISSION. The launch should have concluded a 20 yr long saga caused by changes in the US and NASA launching policy. Unfortunately, the problems started again after launch, when the atypical Galileo deployable high-gain antenna became stuck. It took 2 yr for the mission controllers to find a back-up solution to the antenna malfunction and to put the spacecraft back into an operational mode, which permitted 70% of the original scientific goals to be rescued. On 7 December 1995, Galileo reached Jupiter and, after a close flyby of Io, entered its operational orbit. Extended imaging of the Galilean satellites and in particular of EUROPA, which could be the target of a future NASA mission, was carried out in the course of 4 yr of operations in Jupiter orbit.

Cassini, as mentioned above, is the only mission to the outer planets that includes European participation. The mission was launched from Cape Canaveral on 15 October 1997, on a Titan 4/Centaur expendable launcher. A complex trajectory that includes gravity assists from Venus and the Earth will take Cassini to Saturn in about 6.5 yr. The highlight of the mission will be the deployment of the ESA-built Huygens probe into the dense Titanian atmosphere. Huygens will probe Titan's atmosphere looking for traces of organic and prebiotic compounds.

The leftovers of the cosmic construction

According to the theories of formation and evolution of the solar system, the ASTEROIDS and the COMETS represent the leftovers of the cosmic working area which give rise to our solar system. The study of these numerous but still little-known bodies will allow us to jump back in time and analyze the most ancient and/or uncontaminated materials in the solar system.

The exploration of Halley's comet (COMET HALLEY) marked the beginning of a new era in the understanding of the comets. ESA, the Soviet Union and Japan decided to go chasing the comet that the British scientist Edmund Halley predicted would return in 1758, using Newton's newly formulated laws of motion, after the appearances of 1531, 1607 and 1682. The comet did indeed return as predicted and has since continued to return every 76 yr. The space agencies involved in the exploration of Halley could not miss the launch opportunity when at the beginning of the 1980s they decided to go exploring this comet. An 'armada' of five spacecraft was indeed launched in time: the two Soviet Vegas went on 10 and 14 June 1985; the ESA's Giotto was launched from Kourou on 2 July 1985; the Japanese Sakigake and Suisei probes were launched on 8 January and 18 August 1985 respectively. The closest approach to Halley's nucleus was carried out by the Giotto probe that flewby at a distance of less than 600 km on 14 March 1986. The on-board Multicolor Camera obtained a sequence of images of the nucleus that appeared to be elongated, very dark and characterized by very irregular morphology.

Apart from the interesting scientific results the exploration of Halley's comet will be remembered as the first worldwide effort of coordination of scientific space missions. NASA, which could not provide a spacecraft, joined the other participants by making its Deep Space Network available. Since then the four major space agencies in the world have tried to coordinate their space science programs through the Inter-Agency Consultative Group, an open forum which meets periodically at least once per year.

The GIOTTO spacecraft surprisingly survived its close encounter with Halley. It was decided to put it on an orbit which would have brought the spacecraft back in the vicinity of the Earth every 5 yr, and then hibernate it. In February 1990, 1419 days later, Giotto was reawakened by the ESA mission controllers. Once again Giotto surprised everyone by showing no major sign of deterioration due to the long exposure to the harsh deep space environment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The spacecraft was then retargeted towards the Grigg–Skjellerup comet and a very close (100–200 km) flyby of its nucleus took place on 7 July 1992.

The only images of asteroids obtained at the time of writing were obtained in the course of a couple of flybys by the Galileo spacecraft, en route to Jupiter, and by the Near Earth Asteroid Rendezvous (NEAR) mission. NEAR was launched on 17 February 1996 and it is the first of a new class of NASA missions called Discovery. The main characteristic of the Discovery missions is that they are a kind of Principal Investigator mission which have to be implemented with a NASA budget but completely outside the NASA infrastructure. In principle this procedure allows a faster and cheaper implementation of the missions but it also reduces dramatically the technical reliability of the missions and makes international coordination and collaboration very difficult.

The NASA DEEP SPACE 1 mission, mainly devoted to testing electric propulsion as the primary propulsion, was launched in 1998 with the goal to fly by an asteroid and a comet. Some technical problems in the propulsion system have obliged NASA to change the orbital strategy in flight. Asteroid Braille flyby took place successfully on 28 July 1999 at a distance of 15 km.

The Japanese scientific agency ISAS has announced the launch of a technology mission devoted to returning samples from the Nereus asteroid in 2002. The name of the mission is MUSES-C.

In 2003 the ESA ROSETTA mission will be launched with an Ariane 5 from Kourou. This is a major mission of the ESA science program, supposed to reach comet Wirtanen after 8 yr of deep space cruise and orbit its nucleus for at least 2 yr. En route to Wirtanen, Rosetta will fly by two main-belt asteroids, Otawara in 2006 and Siwa in 2008. A major feature of this mission is the deployment of the Roland Lander on the surface of the Wirtanen nucleus. This lander will carry out high-resolution geochemical, mineralogical and geophysical measurements.

The invasion of the red planet

MARS has always had a special place in people's minds since the time that, rather erroneously, some astronomers thought they had identified artificial structures on the surface of the planet. However, it is true that Mars, both climatically and geologically, is the planet that resembles more closely the Earth. Since the 1970s a large number of missions have been sent to Mars; recently, after a break of about 20 yr, the exploration of the red planet has started again with the involvement of a large number of countries.

In the 1960s NASA was the only agency that dared sending spacecraft to Mars. In 1964 Mariner 4, an adaptation of the lunar Ranger spacecraft, reached the red planet, returning 22 images, mostly from the southern hemisphere. NASA sent two more spacecraft, Mariner 6 and 7, 5 yr later; these were technologically more advanced especially in the field of telecommunications. Despite some problems with the on-board battery pack, the spacecraft reached Mars successfully and carried out imaging campaigns as well as measurements of the composition and density of the Martian atmosphere. Unfortunately, because of celestial mechanics constraints on the orbital scenario of the two missions, Mariner 6 and 7 also observed the southern hemisphere, confirming the wrong impression that Mars was a Moon-like body, heavily cratered and lacking signs of geological evolution. When a new favorable launch opportunity presented itself 2 yr later, the Soviet Union decided to launch MARS 2, a simple orbiter, and Mars 3, consisting of an orbiter and a lander. NASA launched Mariner 9. An agreement was stipulated between the two countries to exchange data and assist each other's missions. In particular, Mariner 9, due to reach Mars before the Mars probes, was supposed to provide assistance for the choice of the landing site. When Mariner 9 entered Mars orbit, the on-board cameras found out that the entire surface was obscured by a planet-wide dust storm. The high winds were presumably the cause of the failure of the Soviet lander. When eventually the dust storm subsided, Mariner 9 initiated an observational campaign which revealed the true 'face' of the planet. Hemispherical asymmetry due to very intense and large-scale geological processes, gigantic shield volcanoes and flood-like morphologies demonstrated that Mars had undergone a very intense geological evolution and that perhaps, in a phase of its evolution, liquid water could have been flowing on its surface. After a new failure of the Soviet Mars 4 and 5 missions, NASA 4 yr later launched the VIKING 1 and 2 missions, both consisting of an orbiter and a lander. The Viking missions were extremely successful, but they also caused a terrible disappointment when the biology experiments showed that that there was no active life on the surface of Mars. This outcome, together with a new emphasis at NASA on the manned flights, brought to a halt the exploration of Mars for almost 20 yr. The Vikings will be remembered, however, for having obtained the first high-resolution and panoramic images of two different (landing) sites on Mars, for having determined atmospheric profiles down to the surface, for having studied for extended periods of time the local meteorology and for having provided a medium–high-resolution coverage of the entire surface from orbit.

In 1988 the Soviets tried again an adventure on Mars with the launch of two very complex spacecraft PHOBOS 1 and 2 consisting of orbiters and landers to be deployed on the surface of Phobos, one of the two natural satellites of Mars. The Mars malediction struck again, however. Phobos 1 was lost shortly after launch, while Phobos 2 failed just prior to its rendezvous with the moon Phobos.

The US went back to Mars on 4 July 1997. The MARS PATHFINDER mission, carrying a rover called Sojourner, landed on the surface of the red planet making use of an innovative passive landing technique, the air bag. Pathfinder had an incredible public relations success, rather than a scientific one, and the great merit of spreading 'Mars fever' everywhere in the world. Since then, a robust program of exploration, internationally coordinated by the International Mars Exploration Working Group, has

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

started with the participation of most of the space-faring nations and organizations.

After the failure of the Mars Observer, which exploded during cruise, NASA decided to decrease the complexity of each mission but to increase the launch frequency. The science objectives of Mars Observer were spread over a number of Mars Surveyor Orbiters, the Mars Global Surveyor, launched in 1996, being the first of the new series.

In the US for the first time ever a budget spreading over 10 yr was approved in order to allow the launch of at least two spacecraft at each launch window: Mars Climate Orbiter and the Mars polar Lander in 1998–9; a Mars Surveyor Orbiter and a Lander in 2001 and in 2003; a Mars Sample Return in 2005 and 2007.

Europe reacted as well. After many years of studies and setbacks, ESA succeeded eventually in starting the construction of MARS EXPRESS. This mission is breaking new ground in Europe in terms of rapidity of approval, simplified management procedures and reduction of the cost. This mission will be launched in 2003 also thanks to the strong collaboration of the ASI and the UK which is providing the Beagle 2 lander. ASI could also be involved in other NASA-led missions to Mars. A sample return mission might see a major participation of the French space agency CNES, which is also supposed to coordinate the construction of the Netlander probes meant to constitute the first geophysical network on the surface of Mars.

Last but not least, the Japanese ISAS launched the NOZOMI (formerly Planet-B) spacecraft in 1998. Because of a technical problem this spacecraft will have a delayed arrival at Mars, but it will be able to carry out a 'tandem' mission with Mars Express supposed to arrive on Mars almost simultaneously.

Bibliography

Lewis J S 1997 *Physics and Chemistry of the Solar System* revised ed (New York: Academic)

Shirley J H and Fairbridge R H (ed) 1997 *Encyclopedia of Planetary Sciences* (London: Chapman & Hall)

Smith A E 1989 *Mars, the Next Step* (Bristol: Adam Hilger)

The ESA web page is at
http://www.esrin.esa.it/

Marcello Coradini

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Interiors

The part of a planet that we can see or directly explore is tiny by comparison to its interior. Inaccessibility does not imply unimportance, however. There are four reasons why planetary interiors are studied. (1) The interpretation of the interior is part of any unifying story for the planet structure and evolution. You cannot hope to understand the formation and history of a planet if you do not know its major constituents and how they are distributed internally. (2) The explanation for many observable features lies inside a planet. For example, you cannot understand why some planets have volcanoes, a magnetic field or a large heat flow without knowing what goes on inside. (3) Even when a phenomenon or planetary attribute is not directly determined by what is inside, the interior still matters. For example, the atmosphere of a planet is often affected by recycling i.e. transport of material into or out of the interior. There is at least as much water inside Earth as in Earth's oceans. (4) Large planets have materials under extreme pressure and high temperature; this is of interest to those who study condensed matter physics (the fundamental science of materials).

What is a planet?

Since the study of planetary interiors involves the study of materials and conditions very different from those at the surface, and since these conditions are determined primarily by the mass of the body and not by whether it orbits the Sun, it is natural to include large satellites in the designation 'planet'. Indeed, Pluto is much smaller than many satellites and even Mercury has a smaller radius than Ganymede or Titan. Small satellites and asteroids are not included in the discussion here, because their internal conditions differ rather little from their surface conditions. At the other extreme, some of the bodies recently discovered in orbit around other stars are ten Jupiter masses or more and could perhaps also be called planets. This designation is uncertain because we know so little about them. They might in some instances have formed in the same way stars formed and thus could better be called brown dwarf companions. Any object exceeding about eighty Jupiter masses can burn light hydrogen and reach the main sequence of conventional stellar evolution and thus should certainly not be called a planet.

Pressures and temperatures

Pressures inside planets are high because of the weight of overlying material. From this physical principle and our knowledge of gravity, we can estimate a pressure deep within a planet of around (1400 bars)$\langle\rho\rangle^2(R/1000\ \mathrm{km})^2$. Here, $\langle\rho\rangle$ is the mean density (mass divided by volume) of the planet in grams per cm^3 (water at one bar pressure has a density of unity in these units), and R is the planet's radius. One bar (=10^5 pascals or 10^6 dynes cm^{-2}) is close to the pressure of EARTH'S ATMOSPHERE at sea level. For Earth's MOON, this equation yields an estimate of around 50 thousand bars (50 kilobars). For Earth, the equation yields an estimate of around 2 million bars or megabars (the 'exact' value for Earth's center being around 3.6 megabars). For JUPITER, the largest planet in our solar system, the estimated 'typical' pressure is around ten megabars. These numbers are of interest because if you ask how much work is done in compressing material to a megabar pressure, it is of the order of an electron volt per atom, and this is the natural energy scale of atoms. In other words, the pressures within planets are sufficient to change the structure of materials and modify greatly their behavior. In a very real sense, the inside of a planet is different from the outside (even when it is made of the same atoms!). Moreover, planets have difficulty eliminating heat, and as a consequence have high internal temperatures, irrespective of whether they are hot or cold externally. The central temperature of Earth is around 6000 K, and temperatures deep within Jupiter are 10^4 K or more. Even the Moon may have a central temperature of around 1600 K. Planets are 'cold' in the sense that thermal energy is small compared to the gravitational energy (indeed, this is much of what distinguishes planets from stars) but they are nonetheless hot enough to be partly molten, or, in the case of giant planets, mostly in the fluid state. In planets larger than about the size of MARS, the high internal temperatures are assured just by the large gravitational energy of formation, much of which is unavoidably converted into heat. These high temperatures can allow massive gravitational *differentiation* (e.g., the separation of a dense iron core from a less dense silicate mantle), particularly during the planet's early history.

Composition and planetary classes

The behavior of a planet depends on the materials from which it is constructed and this depends in turn on the cosmic environment in which the planet formed. Massive planets are made of the most common materials (especially hydrogen), and relatively low mass bodies such as Earth are made of relatively uncommon materials which we call rock. The abundance of elements in the universe is similar to the abundances in the Sun, in most respects, and is dictated by NUCLEOSYNTHESIS, which favors light elements. Hydrogen dominates (it makes up around 70% of the mass of the material from which the Sun formed), followed by helium (27% by mass) and much smaller amounts of oxygen, carbon, neon and nitrogen. Rock forming elements (magnesium, silicon, iron) have still lower abundance.

It is convenient, both on the basis of cosmic abundance and because of how these elements behave, to define three classes of materials: 'gases', 'ices' and 'rock'. These labels refer to composition and are listed in order of decreasing volatility (which refers to their tendency to evaporate easily or resist condensation). Quotation marks are needed to remind us that these labels do not refer to the *state* of the material (e.g., 'gas' can be a metal, as explained below). 'Gases' refer primarily to hydrogen and helium, which do not condense (i.e., form a solid or a liquid)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

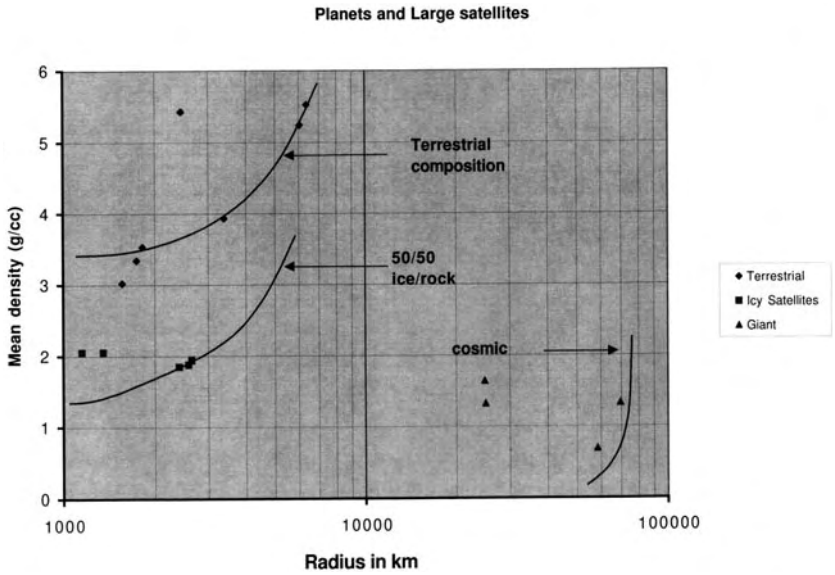


Figure 1. The mean densities and radii for all planets and large satellites in the solar system.

under conditions encountered during the formation of planets. Jupiter and SATURN are the only planets in our solar system that are predominantly 'gas'. 'Ices' refer to the dominant compounds involving oxygen, carbon and nitrogen, and especially to H_2O. The ices partially condense under the cold conditions encountered beyond the asteroid belt and are thus important in all objects from the Jupiter-forming region outwards. 'Rock' refers to the least volatile materials: the silicates, oxides and metallic iron that dominate all bodies interior to Jupiter's orbit. The behavior of these materials can be very different from our everyday experience, primarily because of the high pressures, but also because of high temperature. Thus, 'gas' becomes a liquid metal or semi-metal at the extreme pressures inside Jupiter and Saturn.

Water may become a very good ionic conductor, methane might decompose into hydrogen and diamond, and the minerals that make up rock adopt very different lattice architectures (packing of the atoms). We know about these phenomena through laboratory experiments on Earth, and understand them through the application of quantum mechanics to condensed matter; we then apply this knowledge to our understanding of the planets. As a result, we can predict the expected mean density of a planet for a particular assumption about its composition. The actual mean density is easily determined from the observed mass and radius.

Figure 1 shows the mean densities and radii for all planets and large satellites in the solar system. In the giant planet group, there are two clusters, each containing two points: the gas giants, Jupiter and Saturn, are to be found in the lower right corner, with the ice rich bodies, URANUS and NEPTUNE, at somewhat higher densities and lower radii.

At somewhat similar densities but much smaller radii, there are bodies that have comparable amounts of ice and rock (GANYMEDE, CALLISTO, TITAN, PLUTO and TRITON, although the latter two are somewhat less icy than the first three). If one made bodies of much larger mass (and hence radius) than this composition, they would have much higher density. The bodies that are primarily rocky, MERCURY, VENUS, Earth, Moon, Mars, IO and perhaps EUROPA (though the latter has a significant component of H_2O), form the remaining higher density group. Mercury, Venus, Earth and Mars are called the *terrestrial planets* for obvious reasons.

How is internal structure determined?

Although we can estimate mean composition from mean density, the layering, the thermodynamic state and the dynamics of a planet's interior are much more difficult to determine. *Seismology* is by far the best technique for this. On solid planets, seismology usually means the study of travel times for body or surface waves created by the brittle failure of the rock, but the study of normal modes (oscillatory deformations of the entire body, analogous to the ringing of a bell) are also very important. Normal modes are readily detected for the Sun and you might hope to achieve something similar on giant planets. But the motions exciting the modes are far weaker, so no convincing detection has been made yet. Seismology has proved to be of paramount importance in our study of Earth, where it has enabled to determine crustal thickness, to confirm the existence of major mineral phase transitions in the mantle, and to establish the fluidity of the outer core and the solidity of the inner core. Seismic tomography (the 'imaging' of lateral velocity anomalies) also provides clues on the dynamic state of the mantle (warm upwellings and cold downwellings of large scale convection, carrying heat out of EARTH'S INTERIOR and driving plate tectonics). However, seismology is difficult to employ elsewhere. We

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

have only limited seismic data for our Moon, little yet of value for Mars, and no data for other planets.

Gravity

Gravity has assumed a central role in determining planet structure because of the technical or fundamental limitations of planetary seismology. Planets are not spherically symmetric and, as a consequence, the external gravity field may have a rich harmonic spectrum which influences the trajectory of a passing or orbiting spacecraft detectably. The planet's gravity field can usually be thought of as three rather well separated components: the part that comes from the total mass (and which yields the familiar contribution to the total gravitational force that varies as the inverse square of distance from the planet's center of mass); the part that comes from the rotational (and sometimes tidal) distortion of the planet; and the part that comes from internal dynamics or finite strength of the materials. The rotational or tidal responses are particularly important because (unlike the internal dynamical effects) the driving force for the planet's distortion is precisely known. Consider, for simplicity, the case of a rapidly rotating planet with relatively small tidal effects (e.g. Jupiter or even Earth). The next most important contribution to the gravitational force beyond inverse square is a term that varies as the inverse fourth power of radial distance (the gravitational quadrupole). The dimensionless number that is a measure of its strength is a 'gravitational moment' called J_2. Theory relates this J_2 to the difference between the polar and equatorial moments of inertia of the planet. These two moments of inertia differ by only a small amount but we would really like to know the mean moment of inertia, since that can tell us whether the planet is differentiated or well mixed. However, theory also tells us that provided the body behaves like a fluid on long time scales, J_2 is proportional to the ratio of centrifugal to direct gravitational forces at the equator (a known quantity) with a constant of proportionality that depends on the mean moment of inertia. This *theory of figures*, developed a century ago and refined subsequently, is our main way of estimating the internal structure of all giant planets and also the Galilean satellites (where the permanent tidal bulges also provide similar information). It is an *approximate* theory and, in the case of giant planets, it imposes a much more complicated constraint than merely determining the mean moment of inertia. It does not work for Mercury or Venus because they rotate so slowly that J_2 is very small, thus poorly determined and not dominated by rotational effects. Major uncertainties persist for the interiors of these planets for this reason. It might seem remarkable that this theory should work for Earth, a primarily solid body, but in fact most of Earth's solid material flows in response to stresses as though it were a fluid, provided we allow sufficiently long times to elapse. A timescale less than a million years is sufficient, far less than the ages of the planets (all about 4.5 billion years). This approach has been superseded for Earth, Moon and Mars where we possess additional geodetic information (precession or nutation of the polar axis) and can directly determine the principal moments of inertia, confirming but also refining the approach based on fluidity of the body.

Our everyday concept of a liquid or fluid is an implicit estimate of the viscosity. A low viscosity material flows readily under stress but a very high viscosity material such as water ice or a terrestrial planetary mantle (silicate rocks) might seem to us to be solid on a human timescale, despite flowing. Planets which are fluid in the everyday sense of the word (Jupiter, Saturn, Uranus and Neptune) are made predominantly of materials that flow so easily that there are no detectable long-lived non-hydrostatic contributions to gravity. Terrestrial planets by contrast can often create and support non-hydrostatic loads corresponding to stresses of the order of ten to even thousands of bars. Though these may be tiny compared to the mean stress state deep down (the pressure) they yield detectable gravity and shape anomalies. These are generally interpreted to tell us: (a) the depth above which the planetary material has finite strength on long timescales rather than behaving like a fluid; and (b) something about the scale and magnitude of the dynamic (convective) density anomalies deep within the planet's solid mantle. Indirectly, these tell us something about the temperature and viscosity structure, since the ability of a solid to flow depends very strongly on its temperature.

Heat flow

Heat flow also provides information on temperature structure. There are two major sources of heat in planets, radioactivity and gravitational energy release. Gravitational energy can be converted into heat either primordially (as the planet forms and perhaps differentiates to form a core) or as an ongoing process (the gradual settling of heavier material). Tidal flexing is an additional source that is mainly of interest for large satellites around giant planets; it leads to Io being the most volcanically active body in the solar system.

In terrestrial planets (including the Moon), the dominant source of heating is long-lived radioactive elements (primarily ^{238}U, ^{232}Th and ^{40}K). For a volumetric heating source like this, the heat flux (energy flow per unit area) scales as mass to surface area, or approximately one power of radius. Large planets thus have higher heat fluxes. Planets are also expected to cool on billion year timescales, if only because the radioactive sources slowly decay, but also because they began hot due to the partial conversion of their gravitational energy of formation to internal heat. Gradual cooling might contribute as much as one half of the total heat output, but this is still poorly understood. Heat flows of terrestrial planets are difficult to measure and it is well known only for Earth. However, there is little doubt that the heat flow is so large as to require transport to be predominantly convective.

In Jupiter, Saturn, Uranus and Neptune, gravitational energy dominates, partly because these bodies are so massive but also because only a small fraction of the mass is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the rocky component that carries radioactive elements. A simple model assumes that the heat flow arises from steady cooling from an initial hot state. For a fully convective, adiabatic planet, the heat flow is then limited by the opacity of the atmosphere. This model works quite well for Jupiter but underpredicts the Saturnian heat flow and overpredicts the heat flows of Uranus and Neptune. The shortfall for Saturn is most likely explained by the rainout of helium, which has limited solubility in metallic hydrogen.

Some satellites, notably Io and Europa, are in eccentric, close orbits and undergo tidal heating in addition to radioactivity. This arises because all real solid or partial solid materials are imperfectly elastic and dissipate heat as they are flexed. Io's heat flux is roughly thirty times greater than Earth, and as a consequence, Io is the most volcanically active body in the solar system.

Magnetism

Magnetism can arise as a microscopic (quantum) property of the constituent materials, as in the permanent magnetism of iron or magnetite, but it can also arise from macroscopic currents in an electrical conductor. Since permanent magnetism can only occur below the Curie temperature of the material, it is thought to be restricted to a thin outer layer of solid planets, all of which have hot interiors. The fractional abundance of permanently magnetized materials is also frequently small compared to the major rock-forming minerals, and the spatial distribution of magnetization directions is often complex because of tectonic processes or magnetic field reversals. On Earth, the net contribution of permanent magnetization to the measured field is only one part in a thousand or less at most locations. On this basis, it is thought likely that when coherent, large dipole fields (e.g. surface fields of the order of one Gauss) are observed, they must be caused by macroscopic electrical currents deep within the planet. These currents can be sustained by a *dynamo*, which depends on electromagnetic induction within an electrically conducting, circulating fluid. Temporal variation of the field confirms that it is a dynamic property, but this temporal variation has only been seen for Earth. Planetary magnetic fields (PLANETARY MAGNETOSPHERES) can therefore tell us something of the internal dynamical state and material properties. Earth is the only terrestrial planet that currently possesses a large field, but Mars exhibits evidence of an ancient large field, expressed in unusually high permanent magnetism of ancient rocks. Mercury has a global field that is over two orders of magnitude smaller than Earth's and not understood. Venus has no detectable global field. The Moon may have had an early large field, but the evidence could also possibly be explained by some local process (e.g., impact magnetization). The inference is that only Earth among the terrestrial planets still has the convective core motions needed for a dynamo. Although the other terrestrial bodies are likely to have at least partially liquid cores, they may not be convecting, because the heat flow is sufficiently low that it can be carried by conduction alone. Surprisingly, Ganymede appears to have a dynamo field, probably generated in a liquid iron core. Despite the presence of a water ice mantle, Ganymede thus has some Earth-like attributes.

All the giant planets possess substantial fields, indicating that their deep interiors are convecting and electrically conducting. In Jupiter and Saturn, the conductivity arises from the partial metallization of hydrogen at extreme pressure (a semimetal is sufficient). In Uranus and Neptune, ionic conduction (transport by mobile protons) is the likely explanation.

Volcanism and petrology

Terrestrial planetary mantles consist of a mixture of minerals and as a consequence do not have a single melting point, but a range of melting temperatures. Despite the ability of solids to flow and convect heat at high temperature, some melting occurs and this is expressed at the surface as volcanism and the formation of crust. This melt mainly forms basalt on freezing, which is also a mixture of minerals (those that melt most readily). Basaltic volcanism has dominated the volcanic history of all but the most earliest periods for all terrestrial bodies. The composition and distribution of these basalts, which is the science of petrology, tells us much about what takes place within the planet. The continents on Earth are *not* basalt and are not thought to arise in a simple way; there are no known close analogs to continents on any other planet (though there are places where the topography is high). Volcanism is also a major, sometimes dominant, factor in determining the atmospheric composition of a planet.

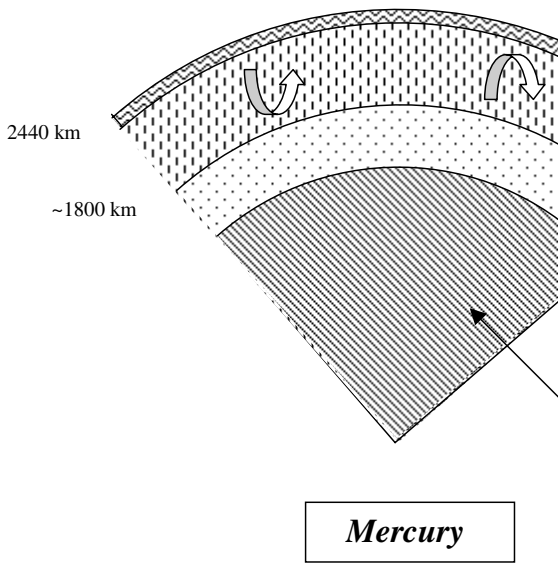


Figure 2. The interior of Mercur

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Giant planets are fluid throughout, except maybe for the very central region, and cannot have volcanism in the usual sense. However, their atmospheric compositions are affected by mixing and differentiation processes within. The most striking example is the lower than cosmic abundance of helium and neon in the atmosphere of Jupiter, and probably Saturn, arising from the formation of helium raindrops deep within the planet.

Icy bodies may also have volcanism if the ice mixture includes appropriate constituents other than water ice. Ammonia and methane are good candidates for driving volcanic activity. (See also VOLCANISM IN THE SOLAR SYSTEM.)

Electromagnetic response

When a planet with a moderately conducting interior is subjected to time-variable external magnetic fields, internal currents are induced and can be detected. On Earth, these are called magnetotelluric currents and they flow both in the ocean water and in the (moist) continental crust. On the Moon, the detection of these induced currents enabled estimates of the conductivity of the deep interior, implying that the rocks are hot (around 1500 K or more). In Europa and Callisto, induced currents and their associated magnetic fields appear to require a conducting layer that is near (i.e. within about 100 km of) the surface. Were this layer to be of salty water (like Earth's oceans) then it must have a thickness exceeding about ten kilometers, strong evidence for at least a thin water ocean in these bodies. Ganymede may also have a water ocean, but the effect is presumably masked by the much larger dynamo field that this body possesses.

Models of planetary interiors

It is easier to send a spacecraft to Jupiter than to send a probe to Earth's core. Not surprisingly, our understanding of planetary interiors is incomplete. 'Models', as described below, provide a summary of our current best efforts to organize our understanding and expectations.

Mercury is unusually dense for its size, and thus must be iron rich. Most of this iron is thought to be in an iron core, much of which may now have frozen. Possibly, Mercury started out with a conventional (Earth-like) composition and then lost part of its silicate mantle in a massive collision with another body similar in mass.

Venus is remarkably similar to Earth in mass and mean composition and may have a similar internal structure. Volcanic structures, presumably built of basalt, provide topographic relief that is of similar magnitude to Earth. The surface is devoid of ancient (heavily cratered terrains) implying resurfacing on a half billion to billion year timescale. However, Venus does not appear to have plate tectonics, at least at present, and may even be heating up internally. Either this or the absence of an inner core may explain why Venus does not currently have a magnetic field. The high surface temperatures may preclude retention of significant permanent magnetism in the rocks.

Earth is in some ways the best understood of the planets, because of the large amount of seismic information, yet many aspects remain mysterious. The greatest puzzle centers around plate tectonics, which is a very efficient way of eliminating Earth's internal heat, yet appears to be unique, at least at present, among the terrestrial planets. Plate tectonics also provides for a large amount of volcanism, most of which takes place on the sea floor. Although the ocean basins would exist even if there were no water to fill them (provided plate tectonics persisted), it is possible that water is indirectly responsible for plate tectonics through its influence on the ability of solid rocks to flow easily beneath the plates. The solid inner core may provide a crucial part of the buoyancy needed for convection in Earth's core and thus sustain the dynamo and the geomagnetic field.

Moon is iron poor and thus has little or no core. The lunar highlands are made of rocks that would occur by freezing from a fully molten outer layer of early Moon; this and other evidence attests to an early traumatic and very hot epoch of Moon's history. The currently favored explanation is that the Moon was born from a giant impact of a Mars-sized or larger body on early Earth. The splash-out from this impact could have placed a short-lived disk of material in Earth's orbit, from which the Moon quickly formed, perhaps in 100 years! All of this took place around 4.5 billion years ago. The mare on the Moon (the darkest regions) are basaltic and came up to a billion years later. The Moon is 'geologically dead' in the sense that there are no surface features that result from internal processes that are more recent than a few billion years ago.

Mars is only one tenth the mass of Earth and appears to have similarity to Earth in bulk composition. It evidently had a large magnetic field for at least the first few hundred million years of its history, during a period of massive crustal production. The northern hemisphere, at least, may have had plate tectonics or some form of recycling, thus explaining why it now has a much thinner crust than in the south. The volcanic province, just northward of the equator, is more recent geologically, and it is likely that volcanic activity on Mars has not entirely ceased. The iron core of Mars is likely to be at least partially liquid and may be entirely liquid if it possesses enough sulfur.

Jupiter is easily the most massive planet and the most Sun-like in its composition. Even so, it does

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

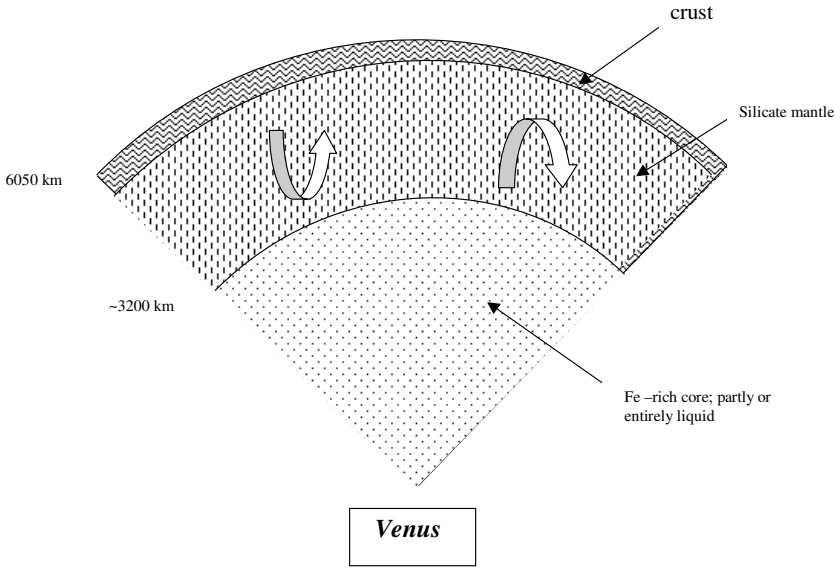


Figure 3. The interior of Venus.

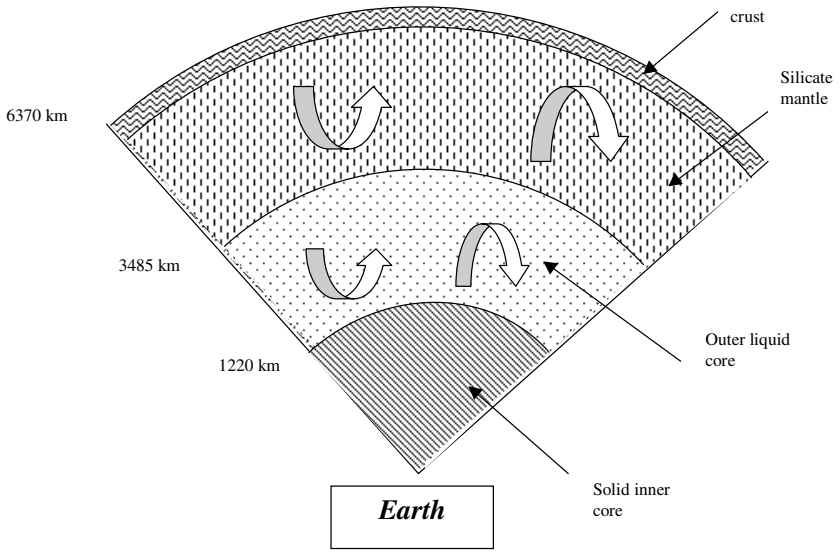


Figure 4. The interior of Earth.

not have the same composition as the primordial Sun, and the differences in composition, though small, are very important. It has long been suspected that Jupiter possesses a core of rock and/or ice, perhaps about ten Earth masses. However, this would be only ∼0.3% or less of the entire mass of the planet and thus very hard to detect! The very precise gravity data obtained by spacecraft do not tell us with certainty whether this core exists, partly because there are still uncertainties in the high pressure behavior of the dominant constituents (hydrogen and helium), and partly because gravity data always admit multiple interpretations. Unlike terrestrial planets, where the core is thought to form by a differentiation event (because of the inability of iron and silicates to remain well-mixed when partially or completely molten), the presence of a core in Jupiter and other giant planets is expected for the popular formation scenario in which a rock–ice core forms first and then gas is added on top. The atmosphere is enriched in a number of elements, notably carbon (in the form of methane) and heavier noble gases; this is probably due to the infall of icy bodies, perhaps including bodies that formed at far greater distance from the Sun where temperatures were very low.

The Galilean satellites are as rich and complex as an entire planetary system. *Io* is roughly Earth-like in composition, though possibly with more volatiles (e.g. a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

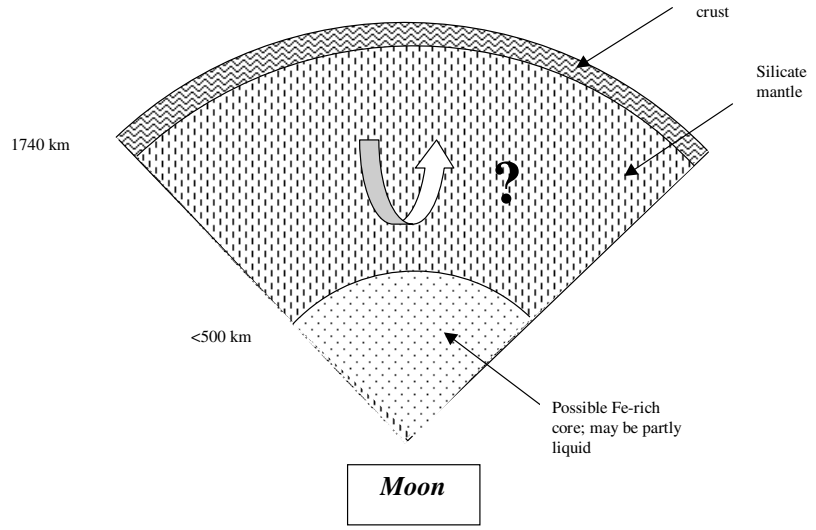


Figure 5. The interior of the Moon.

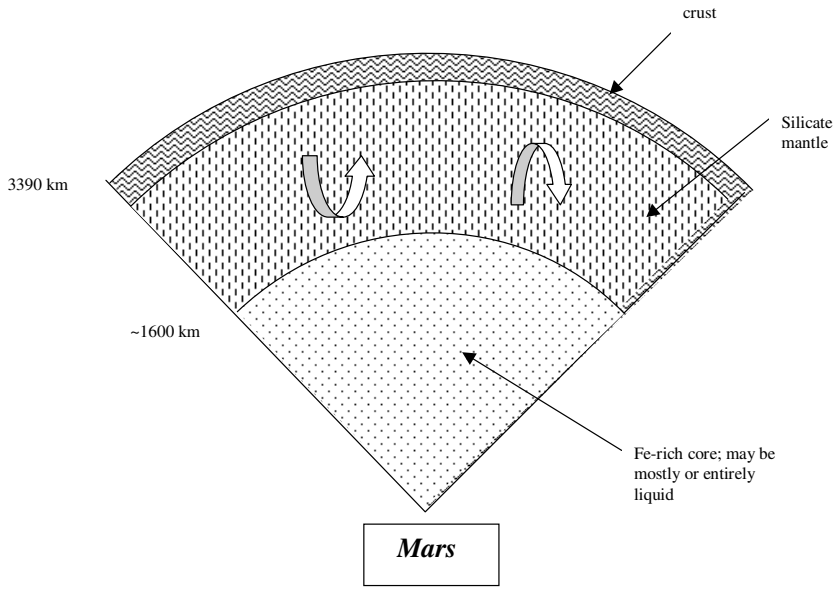


Figure 6. The interior of Mars.

more nearly cosmic sulfur to iron ratio). Gravity data tell us that Io possesses a core, presumably a liquid core of iron and sulfur. Tidal heating has made it the most volcanically active body in the solar system and it has been suggested that a large fraction of the interior is continuously molten. *Europa* is also mostly Earth-like in composition but has a 100–200 km layer of H_2O (as ice or water) on the outside. Tidal heating may cause a large part of this layer to be in the form of an ocean, and the proximity of this layer to the surface has led to speculation about Europa as a possible habitat for past or present life. *Ganymede* is the largest satellite in the solar system and is roughly half (water-dominated) ice and half rock by mass. Gravity data tell us that the body is differentiated, with the central region being Io-like in structure. The evidence of past ice tectonics on the surface suggest that Ganymede has been through an epoch of severe internal heating, possibly tidal in origin. *Callisto* is very similar to Ganymede in size and bulk composition yet remarkably different in other ways: it is only partially differentiated (i.e., the rock and ice are partly mixed), it does not have a large magnetic field and it exhibits none of the tectonic terrains (so-called grooved terrains) that are so common on Ganymede. The differences between Ganymede and Callisto are not well understood.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

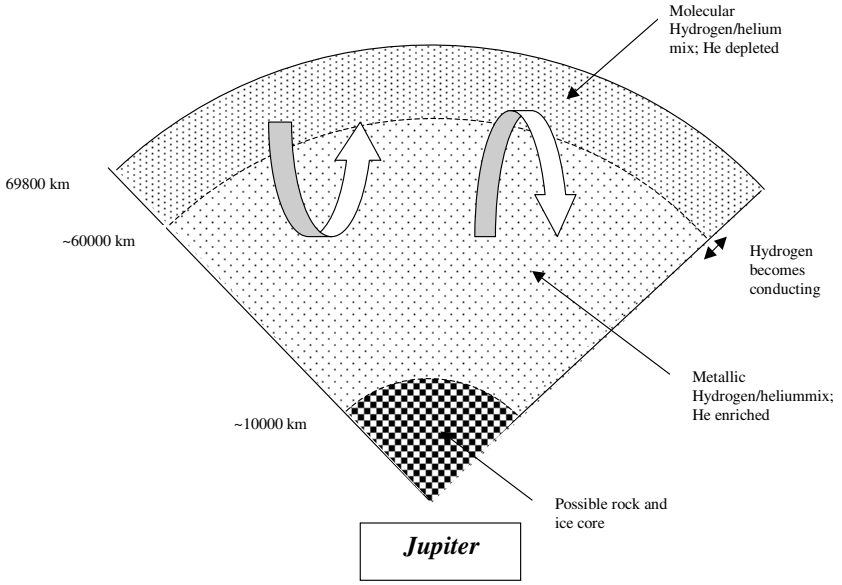


Figure 7. The interior of Jupiter.

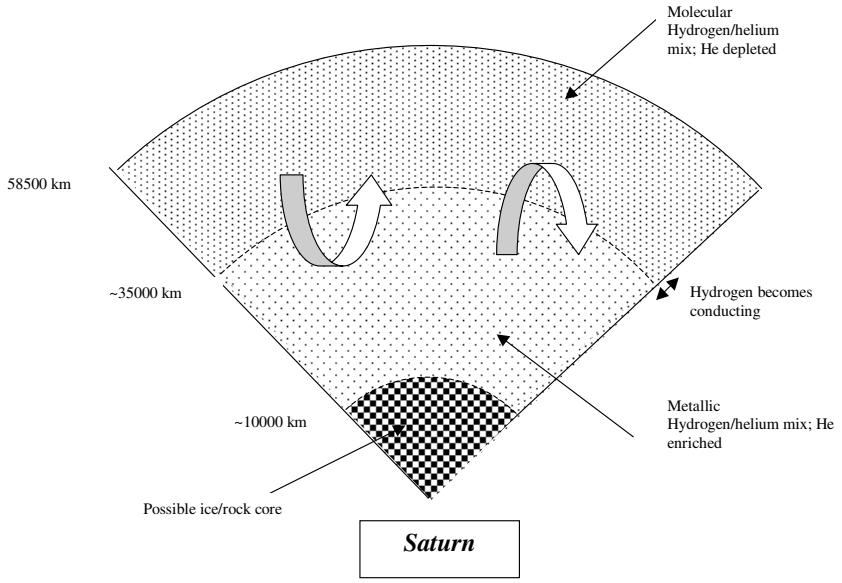


Figure 8. The interior of Saturn.

Saturn is in many ways similar to Jupiter: similar bulk composition, large magnetic field and heat flow. The lower mass means that hydrogen becomes an electrical conductor at greater depths, the magnetic field is generated deeper down, and there is more reliance on helium rain as an energy source. There is somewhat stronger evidence for a central ice/rock core in Saturn than in Jupiter, mainly because it is plausibly around ten Earth masses and thus a bigger fraction of the total mass and easier to detect. Saturn's atmosphere is more enriched in elements such as carbon than is Jupiter's atmosphere, consistent with a large influx of comet-like planetesimals after planet formation (see PLANETARY ATMOSPHERES).

Titan has similarities to Ganymede and Callisto, except that it is richer in some volatile elements and accordingly has a dense, opaque atmosphere consisting mainly of nitrogen and methane. The continuous destruction of methane by solar UV leads to the requirement that Titan has the equivalent of a kilometer of liquid methane or more (some of which has now been converted to liquid ethane and other hydrocarbons). This hydrocarbon reservoir may be partly on the surface as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

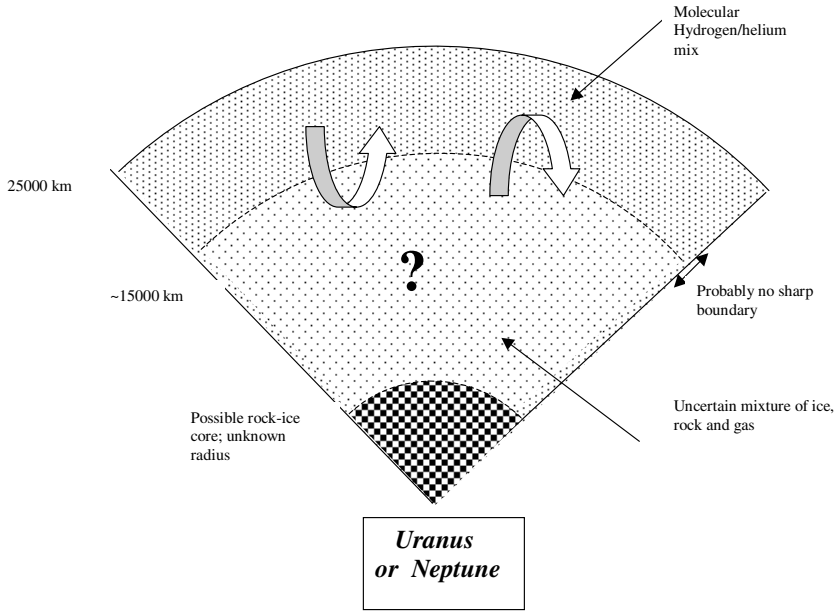


Figure 9. The interior of Uranus or Neptune.

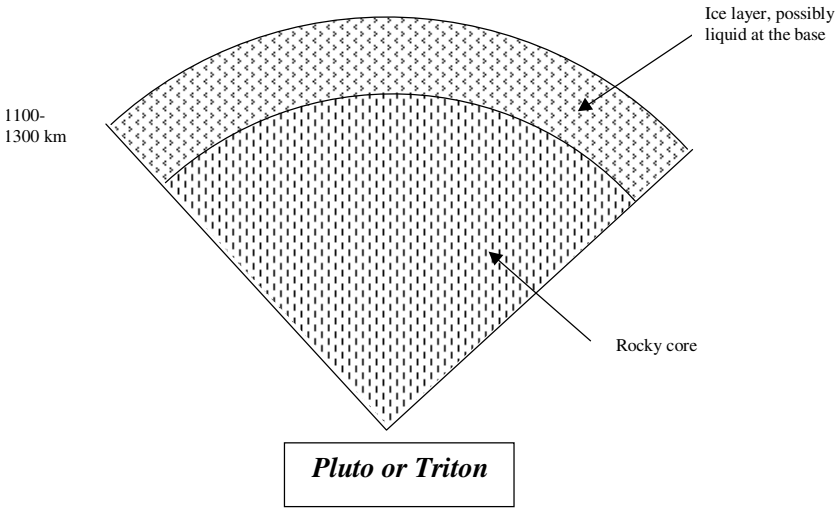


Figure 10. The interior of Pluto or Triton.

oceans or lakes and partly subsurface (analogous to aquifer storage of water in Earth's crust).

Uranus and *Neptune* are very similar but form a distinct class of planets in which no one or two of the three categories of materials ('gas', 'ice' and 'rock') dominates over the other(s). Certainly they possess several Earth masses of solar composition (hydrogen-dominated) material primarily in the outer regions which is responsible for the large radius and low mean density. However, most of the mass is in ice and rock, distributed in an unknown way deep down. Neptune has a higher heat flow than Uranus, but both have magnetic fields, suggesting that both have convective shells at pressures and temperatures where water can conduct electricity.

Triton, *Pluto* and *Charon* have many similarities. Triton is a moon of Neptune, though probably captured, and thus formed in a way that may be indistinguishable from the formation of Pluto. Charon is the moon of Pluto. All of these bodies have mean densities of around 2 g cm^{-3} and must have a lower ice mass relative to rock than we find in the large icy satellites (Ganymede, Callisto, Titan), though with uncertain amounts of other volatiles (methane, molecular nitrogen) that are mostly frozen out on the surface or contained internally.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The future

Progress in understanding the interiors of planets may be slow, partly because it relies heavily on seismology (solid bodies) or deep atmospheric probes (giant planets)—and these are difficult missions—and partly because appropriate laboratory experiments are difficult. Mars is probably the body where the greatest advances, including the installation of a seismic network, can be expected in the coming decade. Detection and characterization of extra-solar system planets (EXOPLANETS) provides much hope for a fuller display of planetary menageries.

Bibliography

A general reference on planetary properties:

Lodders K and Fegley B 1998 *The Planetary Scientist's Companion* (Oxford: Oxford University Press).

See also

Lewis J S 1995 *Physics and Chemistry of the Solar System* (San Diego: Academic)

There is no recent general monograph on the subject of this article. Excepting recent developments, a good general text is

Hubbard W B 1984 *Planetary Interiors* (Princeton, NJ: Van Nostrand-Reinhold)

A general reference on Earth and some aspects of earth-like planets:

Anderson D L 1989 *Theory of the Earth* (Boston: Blackwell)

Excellent information on specific planets can be found in the books published by University of Arizona Press, especially

Mercury, *Venus II*, *Mars*, *Saturn*, *Uranus*, *Neptune* and *Triton*.

For a good review of the giant planets:

Guillot T 1999 *Science* **286** 72

For a good review of the Galilean satellites:

Showman A P and Malhotra R 1999 *Science* **286** 77

David J Stevenson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Ionospheres

Most planets and many satellites in our solar system are surrounded by envelopes of gravitationally bound gases. The interaction of solar radiation and charged particles of solar wind and planetary magnetospheric origin with these gases produces weak IONIZATION that creates planetary ionospheres embedded within the more dense PLANETARY ATMOSPHERES. Additional sources of ionization which are relatively more important in the outer solar system include resonance scattering of solar radiation by interplanetary atomic hydrogen, cosmic rays and UV starlight from the southern Milky Way and the Orion region. The free electrons and ions that result from this interaction are dominated by a thermal population with energies generally comparable to, but not much higher than, the neutral atmosphere and bound to the planet or satellite by gravitational and/or magnetic forces. This thermal population of gravitationally bound electrons and ions is one possible way to define a planetary ionosphere. The interaction of plasma in the inner Jovian magnetosphere with the atmospheres of the Galilean satellites produces ionospheres that strain this definition (see MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS). In this case frequent collisions between the ions and the neutral atmosphere can be a sufficient constraining force on the ionospheric plasma with respect to $\boldsymbol{v} \times \boldsymbol{B}$ convection away from the satellite.

Planetary ionospheres can be classified according to the major categories of planetary atmospheres: (a) the H_2–He atmospheres of the GIANT PLANETS, (b) the CO_2 atmospheres of MARS and VENUS and (c) the N_2 atmospheres of EARTH, PLUTO and satellites TITAN and TRITON. In addition to these categories there is the volcanic generated SO_2 atmosphere of IO, tenuous sputter generated O_2 atmospheres of the other Galilean satellites, and exospheres on MERCURY and the MOON.

Nomenclature

When an electrically conducting layer was discovered (or perhaps better, verified experimentally) in the Earth's upper atmosphere by reflected radio waves, E V APPLETON, to whom the Nobel prize for physics was awarded in 1947, labelled it the 'E-layer' for the reflected wave's electric field. When he inferred additional layers experimentally, he named them D and F layers, and since then this 'letter' terminology has been firmly established in the nomenclature of Earth's ionosphere and by analogy to ionospheres elsewhere in the solar system. Figure 1 presents a schematic of these layers/regions, which are defined below.

D region

The 60–95 km region in the Earth's ionosphere is characterized by a total neutral atmosphere column density of $\sim 10^{20}$ cm^{-2} between the region and the top of the atmosphere. Absorption of solar Lyman-α by NO and solar x-rays below 10 Å and cosmic rays by air are the principal ionization mechanisms. Positive ions are molecular: NO^+ and hydrated ion clusters with three-body reactions being important in the formation of these clusters. A Chapman layer description is generally valid for this region (see below). Near the bottom of this region, negative ions are present.

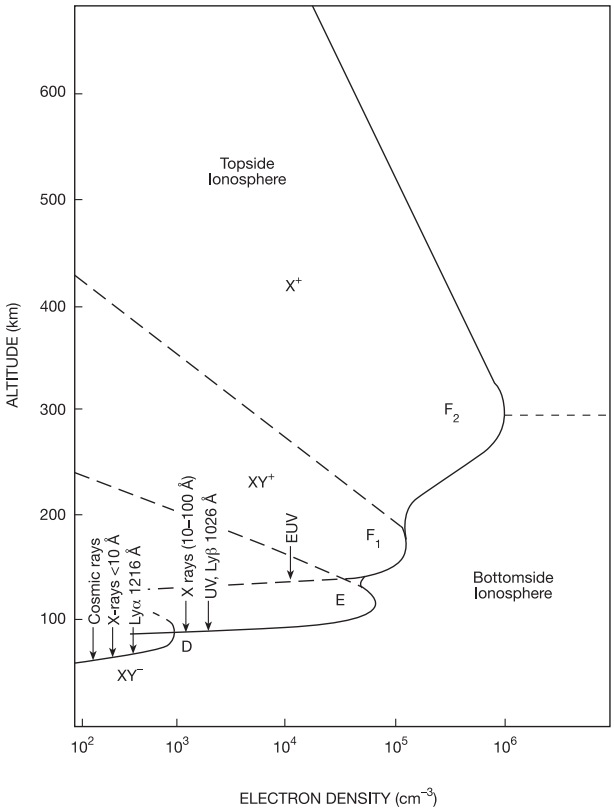


Figure 1. Schematic diagram of the Earth's principal ionospheric regions. X^+, XY^+, and XY^- designate positive atomic ion, positive molecular ion, and negative molecular ion respectively. (After Bauer 1973.)

E region

The 95–135 km region in the Earth's ionosphere is characterized by a total neutral atmosphere column density of $\sim 10^{19}$ cm^{-2} between the region and the top of the atmosphere. Atmospheric absorption of soft solar x-rays and EUV radiation produces the ionization and the terminal ions are predominantly O_2^+ and NO^+. A Chapman layer model with photochemical equilibrium and chemical loss proportional to the electron density squared are a good description. Plasma transport is generally negligible.

F1 region

The 135–200 km region where the peak production of atomic ions occurs and ion loss is controlled by dissociative recombination (see below) is known as the F1 region. In many instances this region is not a distinct layer, but merely a ledge in electron density. The total neutral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

atmosphere column density between the region and the top of the atmosphere is $\sim 10^{17}$ cm^{-2}, the inverse of a typical absorption/ionization cross section of an atom and molecule for FUV and near EUV radiation.

F2 region
This is the region above 200 km, where production of ionization is in the optically thin limit and ion loss is limited by ion–atom and ion–molecule reaction rates with a linear dependence on electron density. Plasma transport due to diffusion and neutral wind induced drifts is extremely important. The electron density in the Earth's ionosphere reaches an absolute maximum where the chemical ion loss time constant is equal to the time constant for plasma transport.

Ionization

The binding energy of an outer shell electron known as the ionization potential (IP) is fairly large, with a typical range of 9–25 eV, as given in table 1. The notable exceptions are the alkali metals with IPs as low as 5 eV. Thus only solar radiation shortward of ~1300 Å and mostly below 1000 Å is important in producing ionospheric plasma. For comparison purposes the energy required to break apart a simple molecule known as the bond dissociation energy is typically ~5 eV, with CO and N_2 being the most notable exceptions. Charged particles must have a high velocity to have sufficient energy to remove an electron from a neutral atom or molecule and to attain a large cross section or probability for ionization of neutrals. Consequently, faster moving electrons are far more important than ions in generating ionospheric plasma.

Absorption of solar radiation and the classical Chapman production rate expression

The attenuation of monochromatic solar radiation at wavelength, λ, through a purely absorbing atmosphere with height, z, measured from the surface, is given by the expression

$$f(\lambda, z) = f(\lambda, \infty) \exp[-\tau(\lambda, z) \sec \chi] \qquad (1)$$

where

$$\tau(\lambda, z) = \sigma_a \int_z^\infty n \, \mathrm{d}z$$

σ_a is the total absorption cross section and χ is the solar zenith angle. If the absorbing gas is in hydrostatic equilibrium, then $\int_z^\infty n \, \mathrm{d}z = nH$, where H is the scale height of the absorbing gas. If more than one constituent absorbs solar radiation at the same wavelength, the optical depth, τ, must be summed over all absorbing species. When absorption of solar radiation leads to ionization, the photoionization rate of the ith constituent, q_i, at a given height is proportional to the attenuated solar flux, the cross section for ionization σ_i and the atmospheric concentration of the ith constituent. Integrating over wavelength, the photoionization rate is

$$q_i(z) = n_i(z) \int_{\lambda(\mathrm{IP})}^0 \sigma_i(\lambda) f(\lambda, z) \, \mathrm{d}\lambda. \qquad (2)$$

Table 1. Ionization potentials (IP) for atoms and molecules of interest in planetary ionospheres. Dissociation energies for molecules.

Species	IP (eV)	$\lambda(IP)$ (Å)	Dissociation energy (eV)
K	4.34	2857	—
Na	5.14	2412	—
Mg	7.64	1621	—
NO	9.26	1340	6.5
CH_3	9.83	1261	4.69
NH_3	10.15	1222	4.40
SO	10.29	1205	5.36
S	10.36	1197	—
C	11.256	1102	—
C_2H_2	11.4	1088	5.38
O_2	12.1	1025	5.115
H_2O	12.6	984	5.118
SO_2	13.1	947	5.65
CH_4	13.1	947	4.5
H	13.595	912	—
O	13.614	911	—
CO_2	13.79	899	5.453
CO	14.0	886	11.09
N	14.545	853	—
H_2	15.41	805	4.48
N_2	15.58	796	9.76
Ar	15.75	787	—
Ne	21.56	575	—
He	24.58	504	—

For simplicity assume $\sigma_i = \sigma_a = \sigma$, a single atmospheric constituent in hydrostatic equilibrium, and a monochromatic beam of sunlight, then

$$q(z) = \sigma n(z) f(\lambda, \infty) \exp[-\sigma n(z) H \sec \chi] \qquad (3)$$

provided $\chi \leq 80^\circ$. The peak production, $q_{\max}$, occurs where $\mathrm{d}q/\mathrm{d}z = 0$. This occurs where the slant optical depth is unity

$$q = q_{\max} = \sigma n f(\lambda, \infty) \, \mathrm{e}^{-1} = \frac{f(\lambda, \infty) \, \mathrm{e}^{-1}}{H \sec \chi}$$

where

$$\tau_{\mathrm{slant}} = \sigma H n \sec \chi = 1. \qquad (4)$$

Note that $q_{\max} \propto \cos \chi$, and $q_{\max}$ is largest when $\chi = 0^\circ$, i.e. overhead sun. Also the height of the peak production z_{peak} is at a minimum altitude when $\chi = 0^\circ$, since at $\tau = 1$, $\sigma H n \propto 1/\sec \chi$ reaches its maximum value with $\chi = 0^\circ$.

Let $h = z/H$, and $n = n_0 \exp(-h)$, then (3) can be transformed to the classical Chapman production rate expression

$$q(h) = \frac{f(\lambda, \infty)}{H} \exp(-h - \mathrm{e}^{-h} \sec \chi) \qquad (5)$$

where the reference levels $h = 0$ and $z = 0$ are the location of $q_{\max}$ when $\chi = 0^\circ$, and $q(h = 0) = q_{\max}(\chi = 0)$. The peak, when $\chi \neq 0$, is located at $z_{\mathrm{peak}} = H \ln \sec \chi$ and $h_{\mathrm{peak}} = z_{\mathrm{peak}}/H$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

For the major constituents in planetary atmospheres, the ionization rate

$$I_\infty = \int_{\lambda(\mathrm{IP})}^{0} \sigma_i(\lambda)\, f(\lambda, \infty)\, \mathrm{d}\lambda \tag{6}$$

at 1 AU (the Earth's orbital distance from the Sun) is in the range of $(0.5\text{–}10) \times 10^{-7}\ \mathrm{s}^{-1}$ with the low values appropriate for H, H_2, He and the large values for species with low ionization potentials such as NO. For O, O_2, N_2, the rates are in range of $(2\text{–}5) \times 10^{-7}\ \mathrm{s}^{-1}$. Of course these rates must be adjusted by the inverse of the planetary distance from the Sun squared ($1/r^2$ factor) for other planets and there is also a factor of 2–3 variability due to solar cycle variation of the Sun's EUV and FUV flux output.

A description of the penetration and ionization rate profile for charged particles is beyond the scope of this article. Generally the required description involves the solution of either a Boltzmann equation or a Fokker–Planck equation. Based on experimental data for ionization and penetration of electrons and some ions, range formulae have been determined to estimate penetration depth in an atmosphere in units of g cm^{-2}. For electron energies above 100 eV, the electron interaction with most gases generates an ion–electron pair for every 35 eV of energy expended. Thus a 3.5 keV electron would produce a total ionization rate of 100 electron–ion pairs.

Ionospheric chemistry

The primary production of ions in a planetary atmosphere is due to ionization of the dominant constituents. For the EARTH'S ATMOSPHERE of N_2, O_2 and O in the thermosphere, the major ionization sources are

$$\begin{aligned} &\mathrm{O} + h\nu \rightarrow \mathrm{O}^+ + \mathrm{e} \\ &\mathrm{O}_2 + h\nu \rightarrow \mathrm{O}_2^+ + \mathrm{e} \text{ or } \mathrm{O} + \mathrm{O}^+ + \mathrm{e} \\ &\mathrm{N}_2 + h\nu \rightarrow \mathrm{N}_2^+ + \mathrm{e} \text{ or } \mathrm{N} + \mathrm{N}^+ + \mathrm{e} \end{aligned} \tag{7}$$

where the electron is ejected with most of the excess translation energy and may have sufficient energy to ionize other atoms or molecules and collisionally excite atoms and molecules. For the H_2–He atmospheres of the giant planets, one can substitute H for O in the first two equations above. Atomic helium ions are also created. In CO_2 atmospheres, CO_2^+ is the major ion initially produced. Once the primary ions are formed they may undergo a variety of chemical reactions. Symmetric charge exchange reactions ($X^+ + X \rightarrow X + X^+$) have no net chemical effect, but provide a mechanism for converting hot ions into hot neutrals which can enhance the escape rate of atmospheric species near the exobase. Charge exchange reactions involving unlike atoms are generally very slow at normal ionospheric temperatures, unless the reaction is accidently resonant (i.e. the atoms have almost the same ionization potential) with the most notable example being $H^+ + O \leftrightarrow H + O^+$. Ions may subsequently react in ion–atom and ion–molecule interchange reactions generically written as

$$\mathrm{A}^+ + \mathrm{BC} \rightarrow \mathrm{AB}^+ + \mathrm{C}$$

$$\mathrm{AB}^+ + \mathrm{C} \rightarrow \mathrm{A}^+ + \mathrm{BC}. \tag{8}$$

At normal ionospheric temperatures these reactions occur by an induced dipole polarization interaction for which the theoretical rate constant known as the Langevin rate constant is given by

$$k_L = 2.6 \times 10^{-9} \left(\frac{\alpha}{\mu}\right)^{1/2}\ \mathrm{cm}^3\ \mathrm{s}^{-1} \tag{9}$$

where α is the polarizability of the neutral in units of $10^{-24}\ \mathrm{cm}^3$ and μ is the reduced mass of the system in atomic mass units (amu). Most ion–atom and ion–molecule reactions proceed at or near the Langevin rate. There are, however, some notable exceptions which are extremely important reactions in the Earth's upper ionosphere, which allow the F region to attain electron densities as high $10^6\ \mathrm{cm}^{-3}$:

$$\begin{aligned} &\mathrm{O}^+ + \mathrm{N}_2 \rightarrow \mathrm{NO}^+ + \mathrm{N} \qquad k \sim 10^{-12}\ \mathrm{cm}^3\ \mathrm{s}^{-1} \\ &\mathrm{O}^+ + \mathrm{O}_2 \rightarrow \mathrm{O}_2^+ + \mathrm{O} \qquad k \sim 10^{-11}\ \mathrm{cm}^3\ \mathrm{s}^{-1}. \end{aligned} \tag{10}$$

These reaction are two and three orders of magnitude slower than the Langevin rate, but

$$\mathrm{N}_2^+ + \mathrm{O} \rightarrow \mathrm{NO}^+ + \mathrm{N} \qquad k \sim 10^{-10}\ \mathrm{cm}^3\ \mathrm{s}^{-1} \tag{11}$$

is faster.

The direction of ion–atom and ion–molecule reactions is always to yield an ion whose neutral progenitor has a lower ionization potential, as a consequence of the necessity for the reaction to proceed via an exothermic channel. This leads to the concept of the terminal ion. In the absence of loss processes which remove ionization (e.g. recombination or transport) ion–atom and ion–molecule reactions will convert all ions to the one whose neutral progenitor has the lowest ionization potential, hence the term terminal ion. For example, in pure air containing only atoms and molecules of nitrogen and oxygen, NO has the lowest ionization potential (cf table 1) and thus NO^+ will be the terminal ion. Introduce an alkali metal such as Na and it will now have the lowest ionization potential and the terminal ion will be Na^+.

Recombination

Ionospheric plasma can recombine by radiative recombination also known as radiative association

$$\mathrm{A}^+ + \mathrm{e} \rightarrow \mathrm{A} + h\nu \qquad \alpha_{\mathrm{rad}} \sim (1\text{–}10) \times 10^{-12}\ \mathrm{cm}^3\ \mathrm{s}^{-1} \tag{12}$$

which is extremely slow because the excess energy (=the ionization potential) must be radiated away. This is generally an inefficient energy loss process due to the mismatch in radiative lifetimes and the time interval when the electron and ion are in close vicinity to each other. Thus terminal atomic ions can potentially attain high densities in the absence of other loss processes. The preferred path for recombination of ionospheric plasma is the conversion of atomic ions to molecular ions which then proceeds via a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

dissociative path involving the formation and dissociation of intermediate state AB*

$$AB^+ + e \rightarrow AB^* \rightarrow A + B \qquad \alpha_d \sim (0.1\text{–}10) \times 10^{-7}\ \mathrm{cm^3\ s^{-1}}. \tag{13}$$

The ionization potential energy is partially consumed by the dissociation of the molecule AB and in many instances either fragment A or B or both are electronically excited. With two fragments the rest of the energy can easily be carried away as translational energy.

The relative velocity for recombination is the electron's velocity and dissociative and radiative recombination rates are faster at lower electron temperatures due to the longer Coulombic interaction time interval at low velocities. Dissociative and radiative recombination rates vary with electron temperature typically as $T_e^{-(0.75\pm0.25)}$.

For the Earth's ionosphere the three major recombination reactions are

$$\begin{aligned} &O_2^+ + e \rightarrow O + O \\ &NO^+ + e \rightarrow N + O \qquad \alpha_d \sim (2\text{–}4) \times 10^{-7}\ \mathrm{cm^3\ s^{-1}} \\ &N_2^+ + e \rightarrow N + N \end{aligned} \tag{14}$$

whereas the following reactions are negligible loss paths

$$\begin{aligned} &O^+ + e \rightarrow O + h\nu \qquad \alpha_r \sim 3 \times 10^{-12}\ \mathrm{cm^3\ s^{-1}} \\ &O_2^+ + e \rightarrow O_2 + h\nu \qquad \alpha_r \sim 10^{-10}\ \mathrm{cm^3\ s^{-1}} \end{aligned} \tag{15}$$

due to their slow rates.

But $O^+ + O^- \rightarrow O + O$ is fast, $\alpha = 2.7 \times 10^{-7}(T/300)^{-0.45}\ \mathrm{cm^3\ s^{-1}}$. However, negative ion densities are negligible above the mesopause (pressures less than 0.01 mbar). Negative ions are primarily produced by three-body reactions

$$e + AB + M \rightarrow AB^- + M \tag{16}$$

which require high pressures (>0.01 mbar) to form negative ions quickly, and are destroyed by the photodetachment process

$$AB^- + h\nu \rightarrow AB + e. \tag{17}$$

Generally the electron affinities of neutral atmospheric species are small (<1 eV, e.g. 0.44 eV for O_2) so visible sunlight can readily detach the electron. A notable exception is NO_3^-, with an electron affinity of 3.9 eV. At night in the absence of an ionization source the ionosphere recombines rapidly and negative ions are not important even though photodetachment is inoperative.

To understand some of the general ionospheric features let us consider the ionospheric plasma to be composed of atomic ions A^+ and molecular ions M^+ in a neutral gas M with electron density [e]. Charge neutrality is an excellent approximation throughout the ionosphere and requires that $[e] = [A^+] + [M^+]$.

The relevant reactions are

$$\begin{aligned} &A + h\nu \rightarrow A^+ + e \qquad q = I_\infty[A](\mathrm{cm^{-3}\ s^{-1}}) \\ &A^+ + M \rightarrow M^+ + A \qquad \gamma(\mathrm{cm^3\ s^{-1}}) \\ &A^+ + e \rightarrow A + h\nu \qquad \alpha_r(\mathrm{cm^3\ s^{-1}}) \\ &M^+ + e \rightarrow A + A \qquad \alpha_d(\mathrm{cm^3\ s^{-1}}) \end{aligned} \tag{18}$$

where $\gamma = \epsilon k_L$ and ϵ is ratio of the actual rate to the Langevin rate (9). The relevant continuity equations are

$$\frac{\partial[e]}{\partial t} = q - \alpha_d[M^+][e] - \alpha_r[A^+][e] \tag{19}$$

$$\frac{\partial[A^+]}{\partial t} = q - \gamma[M][A^+] - \alpha_r[A^+][e] \tag{20}$$

$$\frac{\partial[M^+]}{\partial t} = \gamma[M][A^+] - \alpha_d[M^+][e] \tag{21}$$

where [] denotes number density. In photochemical (stationary or steady) state equilibrium the last continuity equation is

$$\frac{[(A)^+]}{[M^+]} = \frac{\alpha_d[e]}{\gamma[M]} \tag{22}$$

and then charge neutrality can be written as

$$[e] = \left(\frac{\alpha_d[e]}{\gamma[M]} + 1\right)[M^+] \tag{23}$$

and in photochemical equilibrium, the electron continuity equation becomes

$$q = \left\{\left(\frac{\alpha_d[e]}{\gamma[M]} + 1\right)^{-1} + \frac{\alpha_r[M^+]}{\gamma[M]}\right\}\alpha_d[e]^2. \tag{24}$$

The second term on the right-hand side is due to radiative recombination and will always be small in comparison to the first term ($\alpha_d\alpha_r[M^+][e] \ll \gamma^2[M]^2$) for weakly ionized planetary atmospheres. Hence this equation can be rewritten as

$$\frac{1}{q} = \frac{1}{\gamma[M][e]} + \frac{1}{\alpha_d[e]^2}. \tag{25}$$

When $\gamma[M] \gg \alpha_d[e]$, appropriate for the lower ionosphere, then $[M^+] \gg [A^+]$, i.e. mostly molecular ions, and $q = \alpha_d[e]^2$. This is known as the E-region solution which is most appropriate for the 95–135 km region of the Earth's ionosphere and is also called the square-law loss formula due to quadratic dependence on electron density. When $\gamma[M] \ll \alpha_d[e]$, which is valid for low pressures and high in the ionosphere, then $[M^+] \ll [A^+]$ and $q = \gamma[M][e]$ which is known as the F2-region solution, appropriate for above 200 km and is also called the linear loss formula. In this region the ionosphere is composed of mostly atomic ions. For the Earth's ionosphere one may adopt typical values of $q \sim 10^3\ \mathrm{cm^{-3}\ s^{-1}}$, $\alpha_d \sim 10^{-7}\ \mathrm{cm^3\ s^{-1}}$, $\gamma \sim 10^{-12}\ \mathrm{cm^3\ s^{-1}}$, $[M] \sim 10^9\ \mathrm{cm^{-3}}$ to estimate $[e] = (q/\alpha_d)^{1/2} \sim 10^5\ \mathrm{cm^{-3}}$ in the square loss region and $[e] = q/\gamma[M] \sim 10^6\ \mathrm{cm^{-3}}$ in the linear loss region as representative electron concentrations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The continuity equation with plasma transport

At and above the F_2 peak (in electron density) plasma transport becomes important. Plasma can be transported by ambipolar diffusion and by neutral winds dragging ions and electrons up and down field lines. Ionospheric plasma satisfies the plasma approximation which means that the inverse of the product of the plasma density times the Debye length cubed is a small number. Hence the ionospheric plasma behaves as if it were an ideal gas, in spite of the Coulombic interaction of electrons and ions, because the mean kinetic energy is much greater than the interparticle potential energy. Ionospheric temperatures generally are equal to or exceed the neutral temperatures in planetary ionospheres. While it is not necessary for a planet or satellite to have a magnetic field, one test of the relative importance of a planetary magnetic field is the plasma beta parameter, which is the ratio of total plasma pressure to magnetic pressure, or equivalently, the ratio of plasma thermal energy to magnetic energy

$$\beta = \frac{n_i k T_i + n_e k T_e}{B^2/8\pi} \tag{26}$$

where k is Boltzmann's constant. In a magnetic field in the absence of collisions a charged particle executes helical motion perpendicular to the magnetic field with a cyclotron or gyrofrequency given by $\omega = eB/mc$ and corresponding gyroradius or Larmor radius $a_L = v_{\rm perp}/\omega$. The ratio of the gyrofrequency to the collision frequency with neutrals determines whether the motion of electrons and ions is constrained by magnetic fields (large ratio) or unaffected by magnetic fields (small ratio).

In an ionized gas electrons diffuse much more rapidly than ions, but for length scales in excess of the Debye length, the electrons and ions must diffuse with the same velocity by virtue of a strong electrostatic polarization field coupling the electrons and the ions. This process is known as ambipolar diffusion. In planetary ionospheres plasma diffusion and wind induced transport is most important at low pressures and high altitudes, where the ratio of the electron gyrofrequency to the electron collision frequency with neutrals is extremely large (effectively infinite) and the ratio of the ion gyrofrequency to the ion collision frequency with neutrals is large ($\gg$1).

To illustrate the importance of plasma transport which can strongly influence the vertical density distribution of ionospheric plasma let us consider the continuity equation for the number density of the major ion species i

$$\frac{\partial n_i}{\partial t} + \nabla \cdot (n_i \boldsymbol{v}_i) = P_i - L_i n_i \tag{27}$$

where P_i is the production rate per unit volume per unit time, L_i is the loss rate per unit time and v_i is the ion velocity of species i. On a planet or satellite with a magnetic field which may be internal and intrinsic, induced, or imposed by the parent planet or solar wind, the vertical ion flux, ϕ_i, in the limit of a large ratio of the ion gyrofrequency to the ion collision frequency with neutrals, is given by

$$\begin{aligned}\phi_i = n_i w_i &= -n_i \sin^2 I [D_a \mathcal{D} + v_n \cos I \sin I] \\ &= -\sin^2 I D_a \left(\frac{\partial n_i}{\partial z} + \frac{n_i}{H_{eff}} \right)\end{aligned} \tag{28}$$

where w_i is the vertical ion velocity

$$\nu_{\rm in} = k_L n_n$$

is the ion–neutral collision frequency

$$D_a = \frac{k(T_e + T_i)}{\nu_{\rm in} m_i}$$

the ambipolar diffusion coefficient

$$H_i = \frac{k(T_e + T_i)}{m_i g}$$

is the plasma scale height

$$H_{\rm eff} = H_i \left(1 + \frac{H_i}{T_e + T_i} \frac{\partial (T_e + T_i)}{\partial z} + \frac{H_i v_n \cot I}{D_a} \right)^{-1}$$

is the effective plasma scale height and

$$\mathcal{D} = \frac{1}{n_i (T_e + T_i)} \frac{\partial}{\partial z} [n_i (T_e + T_i)] + \frac{1}{H_i}$$

where the latter expression represents the net imbalance between the upward plasma pressure gradient and the downward force of gravity. The subscript n denotes neutral gas and I is the magnetic dip angle measured from the horizontal to the magnetic field vector, which is positive in the downward direction. Thus a neutral horizontal wind component along the magnetic field line ($v_n \cos I > 0$), induces a downward ion flux when the magnetic dip angle is positive.

From the continuity equation terms, time constants for the three important processes can be defined as

(a) diffusion time constant

$$\tau_{\rm diff} = \frac{H_i^2}{\mathcal{D}_a \sin^2 I} \tag{29}$$

(b) wind time constant

$$\tau_{\rm wind} = \frac{H_i}{|v_n| \cos I \sin I} \tag{30}$$

(c) transport time constant

$$\tau_{\rm tran} = \{\tau_{\rm diff}^{-1} + \tau_{\rm wind}^{-1}\}^{-1} = \frac{H_i}{(\mathcal{D}_a \sin^2 I / H_i) + v_n \cos I \sin I} \tag{31}$$

(d) chemical time constant

$$\tau_{\rm chem} = L_i^{-1}. \tag{32}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

With transport included, the peak electron density will occur approximately where the time constants for chemistry and transport are equal, i.e. at a neutral density given by

$$n_n = \frac{b_a \sin^2 I}{H_i\{H_i L_i^{-1} - v_n \cos I \sin I\}} \tag{33}$$

where $b_a = [k(T_e + T_i)]/(k_L m_i) = n\mathcal{D}_a$ is the ambipolar collision coefficient. When winds drag ions up the field lines ($v_n < 0$), then the ionospheric peak is raised in altitude. But there must be net flow downward due to the presence of a chemical sink at lower altitudes and therefore the diffusion velocity should exceed the vertical drift velocity induced by the neutral winds. This statement implies that the peak must occur at a density where

$$n_n = \left| \frac{b_a \tan I}{v_n H_i} \right|. \tag{34}$$

When $v_n = 0$, then ambipolar diffusion and chemistry determine the location of the electron density peak. When winds drag ions down the field lines ($v_n > 0$), then the ionospheric peak is lowered and ambipolar diffusion and winds act in concert.

Note that winds also alter the shape of the electron density profile, especially in the vicinity of the maximum density as evident from the ratio of the effective to actual plasma scale height $H_{\rm eff}/H_i$. Because D_a decreases exponentially with altitude, $H_{\rm eff}$ rapidly approaches H_i within one to two scale heights above the density peak.

At extremely high altitudes in a planetary ionosphere, ambipolar diffusion overpowers wind induced transport and the electron density profile depends on the upper boundary conditions. The diffusion differential equation possesses two basic solutions: (a) a no-flow, static equilibrium with the plasma density distribution given by the plasma scale height H_i, and (b) a non-zero diffusion velocity with the plasma density distribution given by the neutral gas scale height which in thermal equilibrium is one-half the plasma scale height. The latter solution represents the maximum upward flux under steady state conditions and is applicable when the magnetic field lines are open or non-existent and plasma is sufficiently energetic to escape the gravitational field of the planet or satellite. Along closed magnetic field lines the ionospheric plasma is confined and the no-flow solution is applicable.

Unlike the neutral gas species, which above the homopause are in gravitational diffusive equilibrium independent of each other, a mixture of multiple ion species in gravitational diffusive equilibrium are *not* independent of each other because the polarization electric field acts on all ions according to their masses, charge states and temperatures.

Ionospheric currents

In the Earth's E region, especially the ratio of the electron gyrofrequency to the electron collision frequency with neutrals, $\omega_e/\nu_{\rm en}$ is extremely large whereas the ratio of the ion gyrofrequency to the ion collision frequency with neutrals $\omega_i/\nu_{\rm in}$ is $\sim$1. As a consequence the ion motion is strongly coupled to neutral winds through frequent collisions, but the electrons are frozen to move along magnetic field lines. The net effect is a substantial difference in the electron and ion velocities, which leads to current flow in the ionosphere. The presence of a magnetic field generates a highly anisotropic conductivity in the ionosphere, which can be represented mathematically by a tensor. The current flowing parallel to the magnetic field lines (or in the absence of a magnetic field) is given by

$$j_\parallel = \sigma_0 E_\parallel$$

where

$$\sigma_0 = e^2 \sum_{e,i} \frac{n_j}{m_j \nu_{jn}} \tag{35}$$

where the sum is performed over electrons and ions and σ_0 is the longitudinal or direct conductivity and the current perpendicular to the magnetic field lines is given by

$$\boldsymbol{j}_{\rm perp} = \sigma_1 \boldsymbol{E}_{\rm perp} + \sigma_2 c \frac{\boldsymbol{B} \times \boldsymbol{E}_{\rm perp}}{B^2} \tag{36}$$

where

$$\sigma_1 = e^2 \sum_{e,i} \frac{n_j}{m_j \nu_{jn}} \left(\frac{\nu_{jn}^2}{\nu_{jn}^2 + \omega_j^2} \right)$$

and

$$\sigma_2 = e^2 \sum_{e,i} \frac{n_j \omega_j}{m_j (\nu_{jn}^2 + \omega_j^2)} \tag{37}$$

with σ_1 known as the Pedersen conductivity which is parallel to the electric field and perpendicular to the magnetic field and σ_2, the Hall conductivity, which is perpendicular to both the electric and magnetic fields. Note that ω_j is positive for ions and negative for electrons in (37). The electric field can be an imposed field from the magnetosphere, a polarization field, or an induction field $\boldsymbol{v}_n \times \boldsymbol{B}$ due to neutral winds.

Generally in planetary ionospheres, vertical length scales are much smaller than horizontal length scales and to lowest order the vertical electric current is zero. Thus the electric currents in the ionosphere are essentially horizontal. With a right-handed horizontal coordinate system (x, y) and y parallel to the vertical plane containing $\boldsymbol{B}$, Ohm's law may be written as

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \begin{pmatrix} \sigma_{xx} \sigma_{xy} \\ -\sigma_{xy} \sigma_{yy} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \tag{38}$$

where for magnetic dip angles $I \geq 3^\circ$,

$$\sigma_{xx} = \sigma_1, \sigma_{xy} = \frac{\sigma_2}{\sin I}, \sigma_{yy} = \frac{\sigma_1}{\sin^2 I} \tag{39}$$

and in the vicinity of the magnetic equator, $I \leq 3^\circ$,

$$\sigma_{xx} = \sigma_1 + \frac{\sigma_2^2}{\sigma_1} = \sigma_3 = \text{Cowling conductivity},$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\sigma_{xy} = 0, \sigma_{yy} = \sigma_0. \tag{40}$$

The Cowling conductivity is exceedingly large, of the order of the direct conductivity. A highly conducting strip is formed along the magnetic equator which carries a large current known as the 'equatorial electrojet' and the region is known as the dynamo region, which is confined to a few degrees in width, where $\sigma_0 \sin^2 I \ll \sigma_1 \cos^2 I$. The very high conductivity along the magnetic field lines ensures that these lines are approximately electric equipotentials. Because $\sigma_2 \gg \sigma_1$ in the dynamo region, the major electric field is the polarization field rather than the $\boldsymbol{v} \times \boldsymbol{B}$ field, and the major current is the Hall current.

Plasma temperatures in the ionosphere

In the ionization process by either photons or energetic particles the majority of the excess energy is converted into translation energy of the ejected 'photoelectrons' with very little energy going to the ion. At low altitudes in the ionosphere, collision frequencies are sufficiently high to ensure thermal equilibrium among electrons, ions and neutrals. With increasing altitude and exponentially decreasing collision frequencies of electrons and ions with neutrals, the time constant for photoelectrons to lose their energy to neutrals is no longer extremely short in comparison to plasma recombination times and the average electron random kinetic energy can exceed both ion and neutral temperatures. The photoelectrons can cool by collisions with thermal electrons, ions and neutrals. Their collisions with thermal electrons elevate the electron temperature, where it is assumed that the electron velocity distribution has a Maxwellian shape at low velocities and the photoelectrons constitute the high-energy tail of the distribution. If the thermal electrons cool preferentially by collisions with the ions and there is an approximate balance between heat input into the electron gas, Q_e, and heat transfer to the ions, then a first order representation of this equilibrium is given by

$$\frac{Q_e}{n_e n_i} \sim 8 \times 10^{-6} \frac{T_e - T_i}{m_i T_e^{3/2}} \text{ eV cm}^3 \text{ s}^{-1} \tag{41}$$

where the right-hand side has a maximum cooling rate for fixed T_i at $T_e = 3T_i$. Beyond this limit the electrons would have to cool by collisions with neutrals and/or by thermal heat conduction downward to denser regions where cooling rates are faster. Application of similar arguments to the ion gas when an equilibrium between heat input from the electrons is balanced by cooling through collisions with the neutrals leads to the first-order expression

$$T_i \sim \frac{T_n + \psi T_e^{-1/2}}{1 + \psi T_e^{-3/2}} \qquad \psi \sim 6 \times 10^6 \frac{n_e}{n_n}. \tag{42}$$

At low altitudes where the ionosphere is a very weakly ionized gas, $T_i \to T_n$ and thermal equilibrium holds, whereas at high altitudes ψ is large and $T_i \to T_e$. While these simple expressions are only approximate they do illustrate the essential aspects of disequilibrium among the electron, ion and neutral gases in planetary ionospheres. For accurate quantitative vertical profiles of the temperatures of these gases, detailed calculations of thermal heat conduction, non-local effects of photoelectron transport when their mean free path exceeds a scale height and comprehensive treatment of all important energy transfer processes among the gases are required.

The ionospheres of the terrestrial CO_2 planets

In the CO_2 atmospheres of Mars and Venus, photoionization of CO_2 is the dominant source of ions

$$CO_2 + h\nu \to CO_2^+ + e. \tag{43}$$

The CO_2^+ reacts rapidly with atomic oxygen, which is abundant in the upper atmosphere, to form oxygen ions

$$CO_2^+ + O \to O^+ + CO_2$$

$$CO_2^+ + O \to O_2^+ + CO. \tag{44}$$

The formation of an ionospheric F2 region of predominantly O^+ is suppressed by the fast reaction of

$$O^+ + CO_2 \to O_2^+ + CO \tag{45}$$

ensuring that O_2^+ is the dominant, terminal ion in the ionospheres of Mars and Venus. In terms of Earth-based nomenclature, the peak ionization region is an F1 region.

The ionospheres of the giant planets

The giant planets with an upper atmosphere comprising of H_2, He and H have as the major ionization sources

$$\begin{aligned} &H + h\nu \to H^+ + e \\ &H_2 + h\nu \to H_2^+ + e \text{ or } H + H^+ + e \\ &He + h\nu \to He^+ + e \end{aligned} \tag{46}$$

and an equivalent set due to electron impact on these species. Atomic hydrogen has the lowest ionization potential (cf table 1) and the expectation would be that H^+ would be the terminal ion as

$$\begin{aligned} &H_2^+ + H_2 \to H_3^+ + H \\ &He^+ + H_2 \to H^+ + H + He \end{aligned} \tag{47}$$

and H_3^+ would recombine dissociatively with a fast rate leaving behind H^+ which could only recombine radiatively and hence very slowly. In view of its long chemical time constant, protons are subject to plasma transport. With the exception of Saturn with its nearly dipole magnetic field aligned with the rotational axis, the other giant planets have highly distorted magnetic fields with quadrupole and octupole components comparable to the dipole components at ionospheric heights. Thus the calculation of plasma transport is considerably more complicated than in the Earth's ionosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

There is one potential loss path for protons at high altitudes and low H_2 densities involving vibrationally excited H_2

$$H^+ + H_2(v \geq 4) \rightarrow H_2^+ + H. \quad (48)$$

This reaction is exothermic when $v \geq 4$, or equivalently H_2 has a lower ionization potential when it is excited into the $v \geq 4$ vibrational level. At higher pressures, but still sub-microbar, the three-body association reaction becomes important

$$H^+ + H_2 + H_2 \rightarrow H_3^+ + H \quad (49)$$

and at still higher pressures, in the vicinity of the homopause, protons react with CH_4 to produce initially CH_3^+ and CH_4^+. Subsequent reactions with more complex C_2H_x hydrocarbons lead to more complex ions.

The ionospheres of planets and satellites with nitrogen atmospheres

The Earth's ionosphere has been discussed in some detail and is unique because of the presence of life and the associated highly oxidizing atmosphere. The other nitrogen atmospheres of Pluto, Titan and Triton are mildly reducing with CO as the most important oxygen-bearing molecule with a maximum volume mixing ratio of, at most, 0.001 in any of these atmospheres. The primary ions produced are N_2^+ and N^+, which in the atmospheres of Pluto and Titan react with CH_4

$$\begin{aligned} N_2^+ + CH_4 &\rightarrow CH_3^+ + N_2 + H \\ N^+ + CH_4 &\rightarrow CH_3^+ + N + H \\ &\rightarrow CH_4^+ + N \\ &\rightarrow H_2CN^+ + H + H \\ &\rightarrow HCN^+ + N + H_2. \end{aligned} \quad (50)$$

Given the large suite of organic molecules in Titan's atmosphere and possibly in Pluto's atmosphere, the terminal ions are certainly more complex than indicated above and definitely molecular in composition. This ensures fast recombination and low electron densities with negligible plasma transport. On Triton with less CH_4, a volume mixing ratio $\sim 2 \times 10^{-4}$ at the surface and subject to significant photochemical depletion in the lower atmosphere, CO plays a fundamental role in the ionosphere as the dominant source of C atoms and C^+ ions, while suppressing the N^+ concentrations

$$\begin{aligned} N_2^+ + CO &\rightarrow CO^+ + N_2 \\ N^+ + CO &\rightarrow CO^+ + N \\ N^+ + CO &\rightarrow NO^+ + C \\ CO^+ + e &\rightarrow C + O \\ C + h\nu &\rightarrow C^+ + e. \end{aligned}$$

The net effect is to make C^+ the dominant, terminal atomic ion in the lower and middle ionosphere and N^+ the dominant ion in the upper ionosphere.

More specific information on ionospheres of specific planets and satellites may be found in the separate articles on each planet and the major solar system satellites, as indicated in the cross-references above. Additionally, the references listed below may be consulted.

Bibliography

For additional information on ionospheric physics and chemistry, there are a number of books which thoroughly discuss the subject in depth. Highly recommended are:

Banks P M and Kockarts G 1973 *Aeronomy, Parts A and B* (New York: Academic)

Bauer S J 1973 *Physics of Planetary Ionospheres* (Berlin: Springer)

Kelley M C 1989 *The Earth's Ionosphere: Plasma Physics and Electrodynamics* (San Diego: Academic)

Rishbeth H and Garriott O K 1969 *Introduction to Ionospheric Physics* (New York: Academic)

For discussions of individual ionospheres, an excellent set of chapters can be found in the University of Arizona Press's space science series of books: *Jupiter, Venus, Saturn, Mercury, Uranus, Mars, Neptune and Triton* and *Pluto and Charon*, listed here in chronological order of publication.

Darrell F Strobel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Magnetospheres

A definition of a planetary magnetosphere is the region surrounding a planet within which its own magnetic field dominates the behavior of electrically charged particles. A magnetosphere can also be thought of as a magnetic cavity around which the SOLAR WIND is deflected. These definitions are not precise and the term magnetosphere does not imply a spherical shape but is used in a looser sense, as in the phrase 'sphere of influence'. The solar wind is not completely excluded from a magnetosphere and important phenomena result from the coupling between the solar wind and the PLASMAS and magnetic fields that are inside the magnetosphere. Furthermore, modern usage is more inclusive and is generally used to also encompass the solar wind interaction with non-magnetized objects as well as the interactions of satellites that are themselves embedded within magnetospheric plasmas. If a moon has sufficiently strong magnetization, it may carve out a true magnetosphere completely contained within the magnetosphere of the parent planet.

A schematic illustration of a 'generic' magnetosphere is given in figure 1 which shows how the solar wind is diverted around a surface called the magnetopause that encloses the planet. Upstream of the magnetosphere in the solar wind there is a bow shock, a standing shock wave, because the solar wind flow is 'supersonic' (the flow speed is faster than the speed of waves that propagate in the solar wind plasma). After passing through the bow shock, the solar wind flows around the sides of the magnetic barrier formed by planet's magnetic field. The magnetopause is the interface between the solar wind and magnetospheric plasmas. Most of the solar wind is deflected around the magnetosphere with some solar wind plasma leaking into the magnetospheric cavity. A characteristic dimension of a magnetosphere is the distance, on the sunward side, of the magnetopause from the center of the planet (R_M). To first order, this distance is controlled by the balance between the internal pressure of the planet's magnetic field and the external ram pressure of the solar wind. The magnetic field accompanying the solar-wind plasma merges with that of a planet and stretches it out to produce a long, turbulent magnetotail, or wake, on the downwind side of the planet. Spacecraft have determined that a magnetotail can extend up to 100 times farther down stream than the subsolar magnetopause distance (specifically in the cases of Earth and Jupiter; see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE, JUPITER: MAGNETOSPHERE). Thus, magnetospheres are very far from spherical, having an aspect ratio (1:100) that is similar to that of a comet (or a 30 m long pencil).

The nature of the interaction between the solar wind and non-magnetized objects (figure 2) depends on the electrical conductivity of the body. If conducting paths exist across the planet's interior or ionosphere, then electrical currents flow through the body and into the solar wind where they create forces that slow and divert the incident flow. The net result is a region very similar to a magnetosphere around which the solar

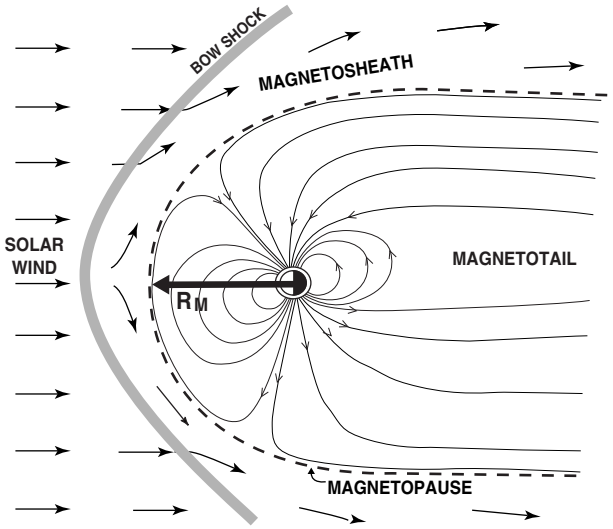


Figure 1. Magnetospheric structures formed by the solar wind interaction with a magnetized planet. The subsolar distance of the magnetopause, R_M, is measured from the center of the planet.

wind is diverted. The interplanetary magnetic field is draped around the planet and stretched out downstream to form a magnetotail (figure 2(*a*)). Mars and Venus have ionospheres which provide the required conducting paths to generate such an 'induced magnetosphere'. The barrier that separates the planetary plasma from the solar wind plasma is referred to as an ionopause. A bow shock forms upstream of the conducting obstacle to the solar wind. The Earth's Moon, with no ionosphere and a surface of low conductivity, does not carry sufficient current to deflect the solar wind. Instead, the solar wind runs directly into the surface of the Moon where it is absorbed (figure 2(*b*)). The absorption leaves the region immediately downstream of the Moon, the wake, devoid of plasma. Farther downstream, the void fills in as the solar wind plasma flows towards the center of the wake.

The interactions of magnetospheric plasma with moons have similarities with the solar wind interaction with planets. An important aspect of the nature of the interaction is that the plasma flows into the moons at subsonic speeds so that the magnetospheric flow is deflected and no bow shock is formed. Since Titan (Saturn), Io (Jupiter) and Triton (Neptune) each have an atmosphere and ionosphere the interaction of these satellites with the magnetospheric plasma can be considered variations of the Mars–Venus case. Europa and Callisto are closer to the Moon's case except the variable magnetic field of Jupiter induces substantial currents (probably in a liquid layer) within these satellites.

Planetary magnetic fields

Because the characteristic time scale for thermal diffusion is greater than the age of the solar system, the planets tend to have retained their heat of formation. At the same

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

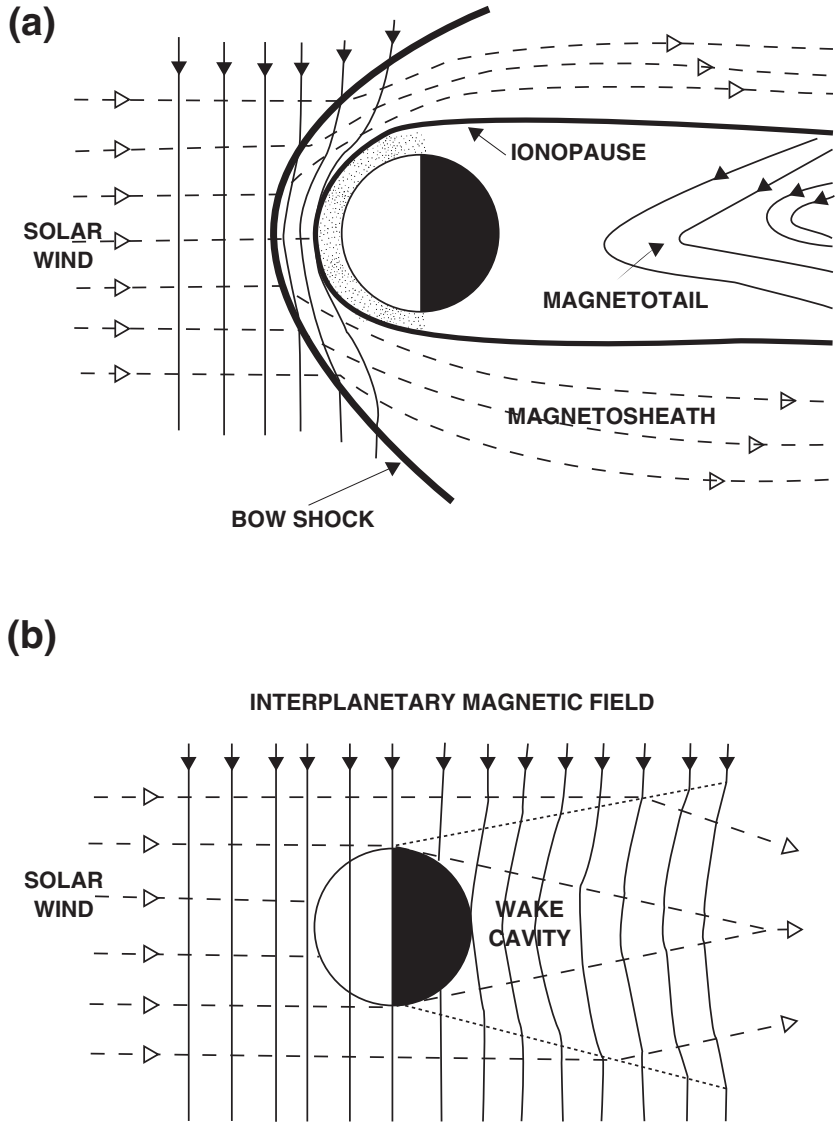


Figure 2. The interaction regions surrounding non-magnetized objects: (*a*) a planet such as Mars or Venus which is sufficiently conducting that currents close through the planet or its Ionosphere (shown by dots on the dayside); (*b*) a body such as the Moon which has no ionosphere and low surface conductivity.

time, the characteristic time scale for diffusive decay of a magnetic field in a PLANETARY INTERIOR is much less than the age of the planets. Consequently, primordial fields and permanent magnetism are small and the only means of providing a substantial planetary magnetic field is an internal dynamo (see DYNAMO THEORY). For a planet to have a magnetic dynamo, it must have a large region that is fluid, electrically conducting and undergoing sufficiently vigorous convective motion. The deep interiors of the planets and many larger satellites are expected to contain electrically conducting fluids: terrestrial planets and the larger satellites have differentiated cores of liquid iron alloys; at the high pressures in the interiors of the giant planets JUPITER and SATURN, hydrogen behaves like a liquid metal; for Uranus and Neptune a water–ammonia–methane mixture forms a deep conducting 'ocean'. The fact that some planets and satellites do not have dynamos tells us that their interiors are stably stratified and do not convect. Models of the thermal evolution of terrestrial planets show that, as the object cools, the liquid core ceases to convect and further heat is lost by conduction alone. In some cases, such as the Earth, convection continues because the nearly pure iron solidifies out of the alloy in the outer core, producing an inner solid core and releasing gravitational energy that can drive continued convection in the liquid outer core. The more gradual cooling of the giant planets allows convective motions to persist.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Planetary magnetospheres.

	Rotation period (days)	Dipole moment (Earth = 1)[a]	Field at equator (G)	Field ratio[b] maximum/ minimum	Tilt of Tilt of dipole[c]	Typical magnetopause distance (R_{plane})	Typical magnetopause distance (km)	Plasma sources[d]
Mercury	59	0.0007	0.003	2	+14°	1.5	0.04×10^5	W
Venus	243 (R)[e]	<0.0004	<0.0003	?	–	–	–	A, W
Earth	1.00	1	0.305	2.8	+10.8°	11	0.7×10^5	W, A
Mars	1.03	$<2.5 \times 10^{-5}$[f]	$<5 \times 10^{-5}$[f]	?	–	–	–	A, W
Jupiter	0.41	20 000	4.2	4.5	−9.6°	80	60×10^5	S, A, W
Saturn	0.44	600	0.20	4.6	<−1°	20	12×10^5	S, A, W
Uranus	0.72 (R)[e]	50	0.23	12	−59°	20	5×10^5	A, W
Neptune	0.74	25	0.14	9	−47°	25	6×10^5	S, A, W

[a] Earth's dipole moment = 7.906×10^{25} G cm^3 = 7.906×10^{15} T m^3.

[b] Ratio of maximum surface field to minimum equals 2 for centered dipole field.

[c] Angle between the magnetic and rotation axes. Minus/plus sign indicates dipole moment antiparallel/parallel to spin vector.

[d] W, solar wind; A, atmosphere; S, satellite(s) or rings.

[e] (R), retrograde.

[f] These values are upper limits on a global magnetic field. The Mars Global Surveyor has shown Mars to have 2000 km × 200 km regions of strong, local magnetization, presumed to be remnant magnetization produced by an ancient (now inactive) dynamo.

The characteristics of the six known planetary fields are listed in table 1. Assuming each planet's magnetic field has the simplest structure, a dipole, we are able to compare the equatorial field strength (B_0) and the tilt of the dipole axis with respect to the planet's spin axis. Note that while the net magnetic moment of the magnetized planets varies by nearly eight orders of magnitude, the surface fields are on the order of a gauss = 10^{-4} T (except for MERCURY) indicating that the strength of the dynamo scales with planetary size. The degree to which this dipole model is an oversimplification of a more complex structure is indicated by the ratio of maximum to minimum values of the surface field. This ratio has a value of 2 for a dipole. The larger values, particularly for URANUS and NEPTUNE, are indications of strong non-dipolar contributions to the planets' magnetic fields. Similarly, the fact that the magnetic axes of these two planets are strongly tilted (see figure 3) also suggests that the dynamos in the icy giant planets may be significantly different to those of the aligned, dipolar planetary magnetic fields.

Exploration of VENUS has provided an upper limit to the degree of magnetization comparable with the crustal magnetization of the Earth suggesting that its core is stably stratified and that it does not have an active dynamo. In the case of MARS, until the Mars Global Surveyor, the only spacecraft that were instrumented with magnetometers did not approach very close to the surface. There has been vigorous debate about whether the weak fields that were detected are due to a dynamo or are due to currents in the ionosphere. The upper limit to the Martian dipole moment is 2×10^{-5} that of the Earth, corresponding to a surface field of 5×10^{-5} G (5 nT). In September 1997 the Mars Global Surveyor spacecraft began its orbital tour of Mars, the first few orbits bringing the spacecraft within ~120 km of the surface where the magnetometer measured regions of surprisingly strong magnetic fields (~400 nT on the surface). The patchy nature of the magnetic field suggests that it is crustal rather than generated by an active internal dynamo. From the preliminary data the magnetization appears to be frozen into the rock when regions of molten lava solidified and is a remnant of a dynamo that was active in Mars' early history. PLUTO has yet to be explored. While models of Pluto's interior suggest it is probably differentiated, its small size makes one doubt that its core is convecting and any magnetization is likely to be remanent. GANYMEDE is the only moon where a substantial magnetic field has been measured. Earth's MOON has a negligibly small planet-scale magnetic field although localized regions of the surface are highly magnetized. The first Galileo flyby of Io has provided suggestions that Io might have an internally generated magnetic field but close-up measurements are needed to separate a dynamo field from the effects of the plasma interaction with Io's atmosphere (see IO: PLASMA TORUS).

For the six planets known to generate magnetic fields in their interiors the size of a planet's magnetosphere (R_P) depends on the ambient solar wind density and the planet's radius and magnetic field. Thus we expect a planet with a strong magnetic field to have a large magnetosphere, and even the weak fields of Uranus and Neptune produce moderately sized magnetospheres in the tenuous solar wind of the outer solar system (see URANUS AND NEPTUNE: ATMOSPHERES, IONOSPHERES, AND MAGNETOSPHERES). The measured sizes of planetary magnetospheres generally agree quite well with the theoretical R_M values. Jupiter is the only notable exception, where the plasma pressure inside the magnetosphere is sufficient to further 'inflate' the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

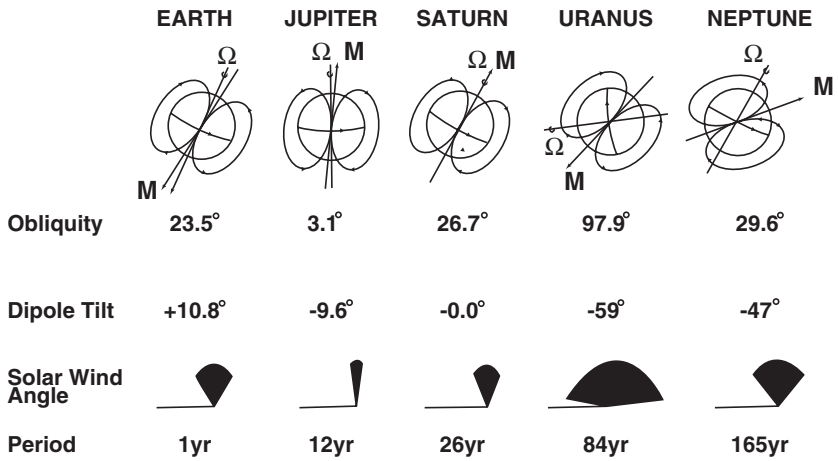


Figure 3. Magnetic field orientations at the Earth and giant planets. The obliquity is the angle between the rotation axis and the normal to the ecliptic. The dipole tilt is the angle between the planet's magnetic and rotation axes. The above two angles contribute to the seasonal variation in the angle between the solar wind flow direction (approximately radial from the Sun) and the planet's magnetic axis (range shaded) over the planet's rotation and orbital periods.

magnetosphere. This makes the magnetosphere of Jupiter a huge object—about 1000 times the volume of the Sun, with a tail that extends at least 6 AU in the antisunward direction, beyond the orbit of Saturn. If the Jovian magnetosphere were visible from Earth, its angular size would be twice that of the Sun, even though it is at least 4 times farther away. The magnetospheres of the other giant planets are smaller (although large compared with the MAGNETOSPHERE OF EARTH), having a similar scale of about 20 times the planetary radius, comparable with the size of the Sun. With only a weak magnetic field and being embedded in the denser solar wind close to the Sun, Mercury has a very small magnetosphere (see MERCURY: MAGNETOSPHERE).

While the size of a planetary magnetosphere depends on the strength of a planet's magnetic field, the configuration and internal dynamics depend on the field orientation (illustrated in figure 3). The orientation of a planet's magnetic field is described by two angles: the tilt of the magnetic dipole field with respect to the planet's spin axis and the angle between the planet's spin axis and the solar wind direction, which is generally within a few degrees of being radially outward from the Sun. Since the direction of the spin axis with respect to the solar wind direction varies only over a planetary year (many Earth years for the outer planets), and the planet's magnetic field is assumed to vary only on geological time scales, these two angles are constant for the purposes of describing the magnetospheric configuration at a particular epoch. Earth, Jupiter and Saturn have both small dipole tilts and small obliquities. This means that the orientation of the magnetic field with respect to the solar wind does not vary appreciably over a planetary rotation period and that seasonal effects are small. Thus Mercury, Earth, Jupiter and Saturn have symmetric and quasi-stationary magnetospheres, each exhibiting only a small wobble at the planetary rotation period owing to their $\sim 10^\circ$ dipole tilts (or a barely detectable wobble in the case of Saturn). In contrast, the large dipole tilt angles of Uranus and Neptune mean that the orientation of their magnetic fields with respect to the interplanetary flow direction varies considerably over a planetary rotation period, resulting in highly asymmetric and time-variable magnetospheres. Furthermore, Uranus' large obliquity means that the configuration of Uranus' magnetosphere will have strong seasonal changes over Uranus' 84 yr orbit.

Magnetospheric plasmas

Magnetospheres contain considerable amounts of plasma, electrically charged particles in equal proportions of positive charge on ions and negative charge on electrons, from various sources (see table 1). The main source of plasma in the solar system is clearly the Sun. The solar CORONA, the upper atmosphere of the Sun (which has been heated by some, as yet undetermined, process to temperatures of 1–2 million K), streams away from the Sun at a more or less steady rate of 10^9 kg s^{-1} in equal numbers (8×10^{35} s^{-1}) of electrons and ions. The boundary between the solar wind and a planet's magnetosphere, the magnetopause, is not entirely plasma tight. Wherever the interplanetary magnetic field has a component antiparallel to the planetary magnetic field, magnetic reconnection is likely to occur and solar wind plasma will enter the magnetosphere across the magnetopause. Solar wind material is identified in the magnetosphere by its energy and characteristic composition of protons (H^+) with $\sim$4% alpha particles (He^{2+}) and trace heavy ions, many of which are highly ionized.

Secondly, although ionospheric plasma is generally cold and gravitationally bound to the planet, a small fraction has sufficient energy to escape up magnetic field lines and into the magnetosphere. In some cases, field-aligned potential drops accelerate ionospheric ions and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

increase the escape rate. Ionospheric plasma has a composition that reflects the composition of the planet's atmosphere (e.g. abundant O^+ for the Earth and H^+ for the outer planets).

Thirdly, the interaction of magnetospheric plasma with any natural satellites or ring particles that orbit within the magnetosphere can generate significant quantities of plasma. Magnetospheric plasma flowing by a satellite can ionize the outermost layers of its atmosphere thus providing a major source of plasma (as much as 1 ton^{-1} of sulfur and oxygen ions in the case of Jupiter's satellite Io). Energetic particle sputtering of the satellite surface or atmosphere can create an extensive cloud of neutral atoms that are eventually ionized, possibly far from the satellite. The distributed sources of water-product ions (totaling $\sim$2 kg s^{-1}) in the magnetosphere of Saturn suggest that energetic particle sputtering of the rings and icy satellites is an important process (see SATURN: MAGNETOSPHERE). Although the sputtering process, which removes at most a few microns of surface ice per thousand years, is probably insignificant in geological terms, sputtering has important consequences for the optical properties of the satellite or ring surfaces.

The composition of the ionic species indicates the primary sources of magnetospheric plasma: satellites in the cases of Jupiter, Saturn and Neptune; the planet's ionosphere in the case of Uranus. In the magnetospheres where plasma motions are driven by the solar wind, solar wind plasma enters the magnetosphere, becoming the primary source of plasma in the case of Mercury's small magnetosphere and a secondary plasma source at Uranus and Neptune. At Earth, both the ionosphere and the solar wind are important sources. Earth's moon remains well beyond the region in which sputtering or other plasma effects are important. In the magnetospheres where plasma flows are dominated by the planet's rotation (Jupiter, Saturn and within the Earth's plasmasphere), the plasma is confined by the planet's strong magnetic field for many days so that substantial densities are accumulated.

Giant rotating magnetospheres

The magnetospheres of Jupiter and Saturn are gigantic, dwarfing all the other planetary magnetospheres. The strong planetary magnetic fields allow the magnetospheric plasma to tap the planet's rotational energy and accelerate particles to high energies. The interactions of magnetospheric plasma with the moons (and ring matter, in the case of Saturn) produce prodigious amounts of new plasma (see MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS). The magnetospheric plasma generates powerful radio emissions and stimulates strong auroral emissions in the planets' polar atmospheres.

The study of Jupiter's magnetosphere has an interesting history. When Burke and Franklin discovered in 1955 that Jupiter is a source of radio emission, it was soon realized that this radio emission must come from energetic charged particles in a strong magnetic field (see MAGNETOSPHERES: JUPITER, RADIO EMISSIONS). This remarkable discovery came before Van Allen's detection of the Earth's radiation belts and the *in situ* verification of the solar wind. A more puzzling discovery came in 1964 when it was revealed that the component of the Jovian radio emission at decametric ($\sim$tens of meters) wavelengths was influenced by Io, the innermost of the four Galilean satellites. The component of the radio emission at decimetric ($\sim$tens of centimeters) wavelengths was assumed to be synchrotron radiation from electrons with energies of >10 MeV that gyrate around dipolar magnetic field lines at a distance of a few Jovian radii. This basic picture of a strong magnetic field trapping a large, energetic particle population was confirmed by PIONEERS 10 and 11 which reached Jupiter in 1973 and 1974, respectively. The Pioneers also revealed that farther from the planet the magnetic field is considerably stretched out so that the Jovian magnetosphere is shaped more like a disk than a sphere (see figure 4). Although the large size and radial distension implied the presence of a substantial amount of plasma at lower energies, the Pioneer plasma detector provided little information on the thermal population. Nevertheless, the theorists had already come out strongly in favor of a magnetosphere dominated by the planet's rotation. A few months before the Pioneer 10 encounter optical emission was detected using a ground-based telescope, coming from neutral sodium atoms in the vicinity of Io. The first direct evidence of the presence of high densities of ionized material at low energies near Jupiter came with the discovery of optical emission from S^+ ions. Borrowing techniques from studies of more remote astronomical gaseous nebulae, planetary astronomers concluded that the S^+ emission comes from a dense (500–3000 cm^{-3}) ring of cold (2.3×10^5 K) plasma that is corotating with Jupiter inside the orbit of Io. In 1979 the VOYAGER 1 spacecraft confirmed that Io is the major source of plasma in the Jovian magnetosphere; a fact that seemed less surprising when the Voyager cameras revealed the satellite's active volcanoes. Bright ultraviolet emission and local plasma measurements revealed an extensive torus of hotter ($\sim10^6$ K) plasma outside the inner ring of cold plasma.

In the two years between September 1979 and August 1981 the magnetosphere of Saturn was explored by three spacecraft, Pioneer 11 and Voyagers 1 and 2. Overall, Saturn's magnetosphere (figure 5) was found to be similar to the Jovian magnetosphere: satellites are the major source of magnetospheric plasma and the plasma dynamics are dominated by the planet's rotation. Nevertheless, the magnetosphere of Saturn is considerably smaller and the multiple sources of plasma are much weaker and less distinct than at Jupiter. Moreover, Saturn's magnetosphere is less compressible. This behavior indicates that the planetary field stands off the solar wind with little contribution from the internal plasma pressure. The magnetosphere of Saturn is separated by a boundary at about 15 Saturn radii into two regions: an inner region where the plasma density and temperature vary smoothly with distance and an outer region where

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

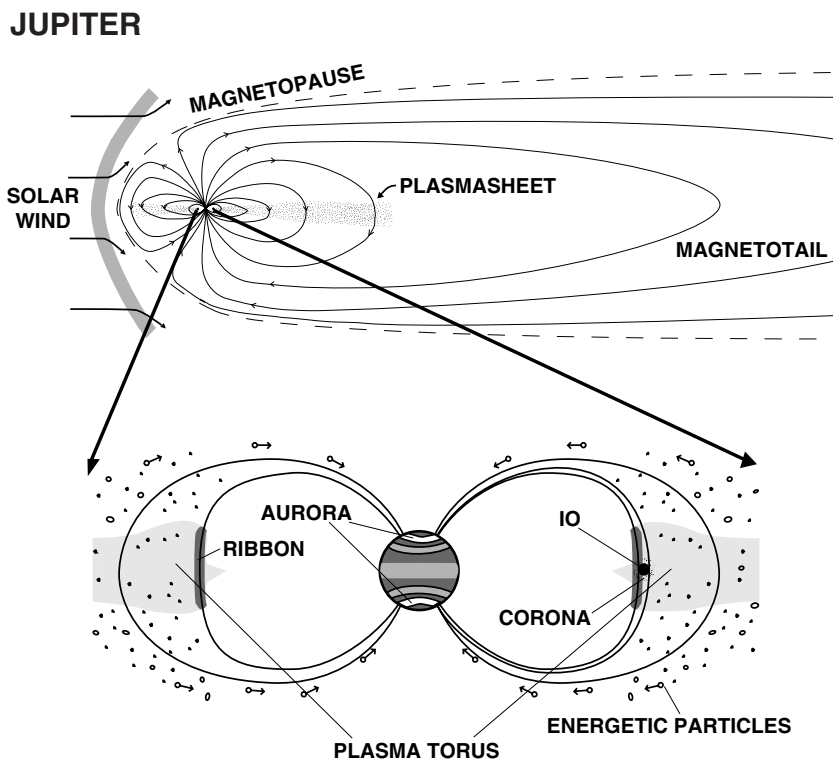


Figure 4. Magnetosphere of Jupiter. Deeply embedded within the magnetosphere of Jupiter is the satellite Io. The ionization of Io's atmosphere and extended neutral corona deposits 1 ton s^{-1} of plasma into the torus. In the process of plasma spreading out from the torus in to the extended plasmasheet, the charged particles are accelerated. Some of the energetic particles are scattered and stream along magnetic field lines into the atmosphere of Jupiter where they stimulate intense auroral emissions.

the plasma densities and temperatures vary erratically. While there is wide agreement that the source of plasma in the inner region of the magnetosphere is the icy satellites Dione and Tethys (with lesser contributions from Rhea, the small inner satellites and the rings) there is strong debate over the source and loss processes. In the outer region, debate also remains about the role of Titan (see SATURN: MAGNETOSPHERE INTERACTION WITH TITAN).

Asymmetric magnetospheres

Early magnetospheric studies were based on experience of Earth, Jupiter and Saturn and only considered symmetric magnetospheres. When the planet's magnetic field can be closely approximated by a dipole at the center of the planet with its axis roughly parallel to the planet's spin axis then the magnetospheric configuration is symmetric and remains constant as the planet spins (convective motions of plasma in the magnetosphere are quasi-steady in the inertial reference frame). Once the magnetic tilt angle becomes appreciable this is no longer true.

Voyager 2 made the first-ever encounter with Uranus on 24 January 1986. It was realized ahead of time that the rotational axis of Uranus would, in early 1986, be within 8° of the planet–Sun line, whereas those of all other planets (except Pluto) are approximately perpendicular to their orbital planes. If Uranus' magnetic and rotational axes were nearly parallel, as is the case for other magnetized planets, one magnetic pole would be pointed almost directly at the Sun and a very unusual magnetospheric topology would be expected. If they were inclined markedly to one another, each magnetic pole would move in a sweeping, conical path as the planet spun—creating an even more exotic magnetosphere.

Voyager 2 ended our many years of conjecture. It found that Uranus has a substantial magnetic field with an orientation and asymmetry very different from expectations. Uranus' magnetic axis is tilted a huge 59° from the rotational axis and is highly asymmetric (see table 1 and figure 3). Its large magnetosphere contains plasma and a large population of energetic particles. As at Saturn, the absorption of energetic particles by satellites Miranda, Ariel and Umbriel, and by particulate matter in rings, controls the development of the inner magnetosphere. However, unlike the situation at Saturn, the diurnal wobble of the tilted magnetic equator creates a very complex relationship between the orbiting satellites and the magnetic field's topology.

At Uranus the rotation axis is currently nearly parallel to the solar wind direction (figure 3) so that the solar wind driven convection and the direction of planetary rotation are orthogonal and hence de-coupled. In the reference frame rotating with Uranus the solar wind driven

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

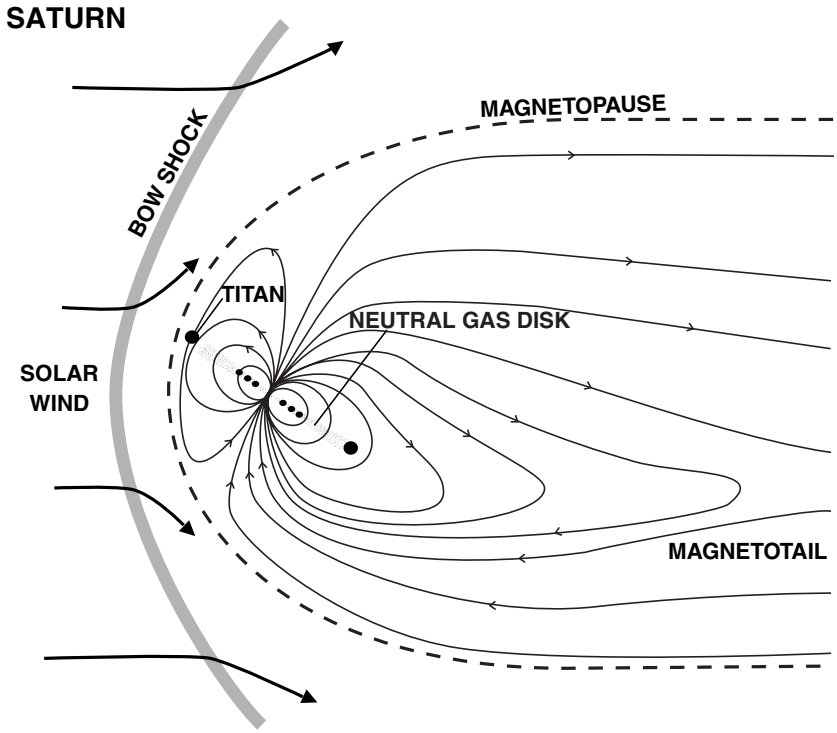


Figure 5. Magnetosphere of Saturn. The magnetosphere of Saturn is populated by the ionization of an extensive disk of neutral gas which was generated by the sputtering of the rings and icy satellites by energetic radiation belt particles.

convection is quasi-steady and permeates throughout the magnetosphere. Thus the plasma corotates with the planet once every 17 h but on a longer time scale (days) the plasma is circulated through the magnetosphere by the solar wind driven convection. Elements of plasma exhibit helical trajectories, spiraling sunward at the magnetic equator and anti-sunward at high latitudes. Figure 6 (top) shows two configurations of Uranus' magnetosphere separated by half a planetary rotation. To a first approximation, the magnetosphere of Uranus resembles that of the Earth but revolving every 17 h around the planet–Sun line.

When Voyager 2 flew past Neptune in August 1989 it discovered an extensive magnetosphere with a large population of trapped charged particles. The magnetometers detected a magnetic field whose surface intensity is similar to the Earth's, but whose distribution was completely different. Neptune appeared to have a magnetic field with a higher degree of asymmetry than any other planet. To first order, Neptune's magnetic field can be described as a dipole that is tilted 47° from the planet's spin axis. As the planet spins over Neptune's 16 h spin period, the orientation of the planet's magnetic field drastically changes with respect to the solar wind. When the magnetic dipole is perpendicular to the solar wind, the configuration is, momentarily, symmetrical like that of Earth, Jupiter and Saturn. When the angle between the magnetic dipole and the solar wind direction is small, we have a unique configuration with the magnetic axis pointed 'pole-on' into the solar wind. These configurations lead to very different magnetic field topologies (see figure 6 (bottom)). Complete reconfiguration of Neptune's magnetosphere must occur every 16 h spin period.

Neptune's large satellite TRITON is comparable in size with Ganymede which raises the question of whether it too could have an internal magnetic field. The large moon orbits at $14.6R_N$ ($1R_N$ is a Neptune radius), well inside the magnetosphere. Voyager 2 detected protons and nitrogen ions in Neptune's magnetosphere, consistent with a source of $\sim$200 g s^{-1} from Triton and strong inward diffusion with a time scale of a few days. It appears that the rapid transport does not allow large plasma densities to build up. The nitrogen ions are thought to be produced directly in the interaction of the magnetospheric plasma with the satellite's atmosphere while the protons come from the ionization of a large hydrogen cloud that extends from Triton's orbit in to about $8R_N$. Closer to the planet there appear to be significant losses due to waves which scatter the ions and electrons along the magnetic field into the atmosphere of Neptune. The charged particles transfer their energy to atmospheric molecules which then glow as dramatic auroras. Neptune's tilted and asymmetric magnetic field concentrates the bombarding charged particles into an irregular region at mid-latitudes rather than the simple auroral ovals at the poles of other planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

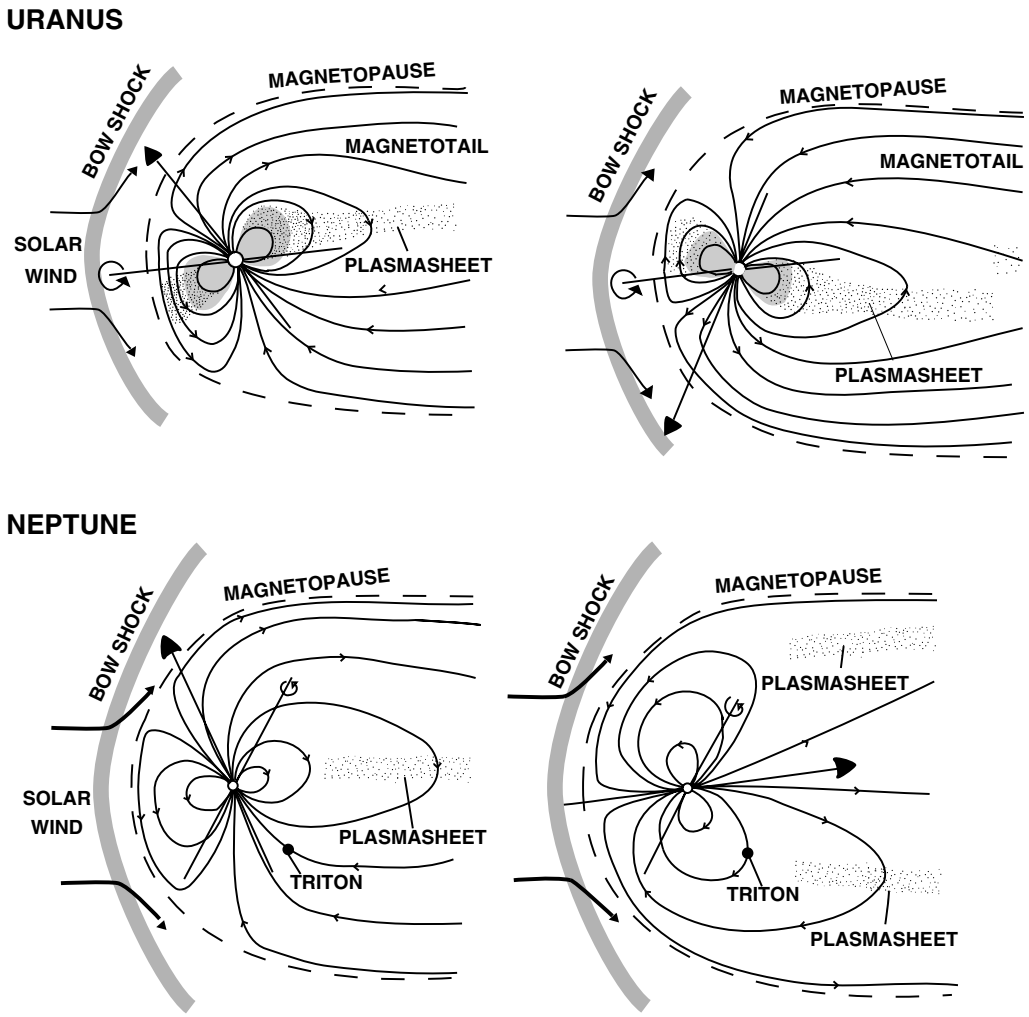


Figure 6. The magnetospheres of Uranus and Neptune. The large angle between these planets' magnetic and rotation axes leads to substantial variations over a planet's rotation period, particularly for Neptune.

Small magnetospheres

Before 1974 it was thought that only large planets would have magnetic fields and that convective motions in smaller objects would have halted as the interiors cooled. The discovery of weak but significant magnetic fields of smaller bodies, first Mercury and recently Ganymede, has challenged our view of both the dynamo process and the nature of small-scale magnetospheres. Mercury and Ganymede have magnetospheres that are comparable in size with the diameter of the Earth (see figure 7).

The magnetosphere of Mercury (figure 7 (top)) contains field lines that connect to the planet at both ends but the Mercurian magnetic field is too weak to maintain a belt of trapped particles. Low-energy plasma was detected by the MARINER spacecraft's traversals through the nightside magnetosphere but it is not clear whether this plasma originated in the solar wind or from ionization of Mercury's tenuous atmosphere. More notable, however, are the intense bursts of energetic particles that were measured in the magnetotail. These bursts have been compared with magnetic storms on Earth and suggest Mercury's magnetosphere is dynamic. The time scales of $\sim$1 min for the Mercurian storms (compared with hours at Earth) are consistent with the much smaller scale of the magnetosphere.

One of the greatest surprises of the Galileo mission to Jupiter was the discovery of Ganymede's magnetic field. This was the first magnetosphere within a magnetosphere (figure 7 (bottom)). The Jovian plasma, in which Ganymede is embedded, flows at speeds much slower than the solar wind and slower than the local speeds of waves that can propagate in the magnetospheric plasma. Hence, there is no bow shock upstream of Ganymede. Models of the interaction of plasma in Jupiter's magnetosphere with Ganymede's magnetic field suggest that there is a small region where magnetic field lines are closed. However, the magnetosphere of Ganymede is too small to have stably trapped radiation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

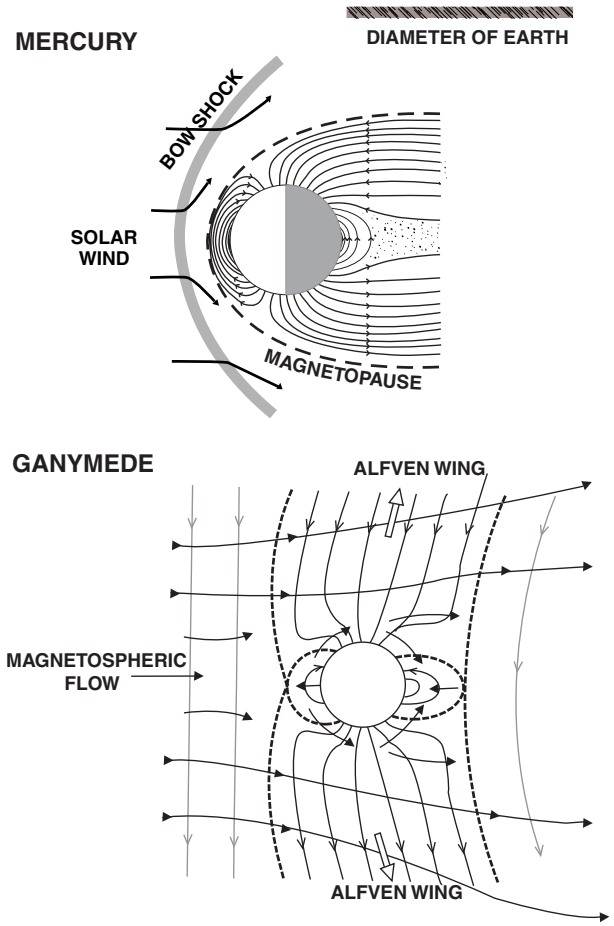


Figure 7. The small magnetospheres of Mercury and Ganymede are comparable in size with the Earth's diameter.

belts or a magnetotail or contain substantial amounts of plasma. There is thought to be a pattern of convection, similar to solar wind driven convection at Earth, but in this case driven by the interaction of Ganymede's magnetic field with the surrounding magnetospheric plasma. Orbiting Jupiter at a distance of 19 Jovian radii, Ganymede is located in the middle magnetosphere region (see figure 4). Ganymede experiences a periodic change in plasma conditions and background magnetic field orientation as the plasma sheet, tilted by the 10° of Jupiter's magnetic field with respect to the equator, passes over the satellite twice per 10 h rotation.

Magnetospheric scaling laws

Magnetopause distance.

In a true magnetosphere, the scale size is set by the distance, R_{MP}, along the planet–Sun line at which the sum of the pressure of the planetary magnetic field and the pressure exerted by plasma confined within that field balances the dynamic pressure of the solar wind. (The dynamic pressure is ρu^2 where ρ is the mass density and u is its flow velocity in the rest frame of the planet. The thermal and magnetic pressures of the solar wind are small compared with its dynamic pressure.) Assuming that the planetary magnetic field is dominated by its dipole moment and that the plasma pressure within the magnetosphere is small, one can estimate R_{MP} as $R_{MP} \approx R_P(B_o r/2\mu_o \rho u^2)^{1/6}$. Here B_o is the surface equatorial field of the planet and R_P is its radius. Table 1 shows the vast range of scale sizes in terms both of the planetary radii and of absolute distance. The large scale sizes of the magnetospheres of the outer planets arise both because R_P is large and because the solar wind density is so small at the orbits of the outer planets that the solar wind pressure is not effective in balancing a strong magnetic field.

Dynamics

Magnetospheric configuration is generally well described by MAGNETOHYDRODYNAMICS (MHD) in which the magnetic field can be considered to be frozen into the plasma flow. Thus we need to consider the processes controlling magnetospheric flows. The two largest sources of momentum in planetary magnetospheres are the planet's rotation and the solar wind (see figure 8). The nature of any large-scale circulation of material in the magnetosphere depends on which momentum source is tapped. For planetary magnetospheres, corotation of plasma with the planet is a useful first approximation with any departures from strict corotation occurring when certain conditions break down. It may be helpful to think of plasma in the magnetosphere as mass that is coupled by means of magnetic field lines to a giant flywheel (the planet) with the ionosphere acting as the clutch.

For magnetospheric plasma to rotate with the planet, the first two conditions are that the upper region of the neutral atmosphere must corotate with the planet and must be closely coupled to the ionosphere by collisions. The (Pedersen) electrical conductivity of the ionosphere, σ^i, is large so that in a corotating ionosphere (with velocity $\boldsymbol{V}^i$) horizontal currents (perpendicular to the local magnetic field) are given by Ohm's law, $\boldsymbol{J}^i_\perp = \sigma^i_\perp(\boldsymbol{E}^i + \boldsymbol{V}^i \times \boldsymbol{B})$. Just above the ionosphere the plasma particles in the (collision-free) magnetosphere are tightly bound to the magnetic field and do not easily conduct currents across the magnetic field so that $\boldsymbol{E}^m = -\boldsymbol{V}^m \times \boldsymbol{B}$. Because the plasma particles are far more mobile in the direction of the local magnetic field, the parallel conductivity in the magnetosphere $\sigma^m_\parallel$ is large and the field lines can be considered to be equipotentials ($\boldsymbol{E} \cdot \boldsymbol{B} = 0$). Thus the electric field in the magnetosphere can be mapped into the ionosphere. Because the ionosphere is relatively thin, the electric field $\boldsymbol{E}^m$ just above the ionosphere is the same as $\boldsymbol{E}^i$ so that we can write $\boldsymbol{J}^i_\perp = \sigma^i_\perp(\boldsymbol{V}^i - \boldsymbol{V}^m) \times \boldsymbol{B}$. The condition for corotation of the magnetospheric plasma is that the ratio J^i/σ^i is sufficiently small that $\boldsymbol{V}^m = \boldsymbol{V}^i = \Omega \times \boldsymbol{r}$. For a dipolar magnetic field that is aligned with the rotation axis, the corotational electric field (in the equatorial plane) is therefore radial with magnitude $E_{co} = \Omega B_o/r^2$.

It is clear that large ionospheric conductivities facilitate corotation. The large $\sigma^m_\parallel$ also means that any

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

currents in the magnetosphere that result from mechanical stresses on the plasma are directly coupled by field-aligned currents to the ionosphere. Thus corotation breaks down when mechanical stresses on the magnetospheric plasma drive ionospheric currents which are sufficiently large that the ratio $J^{\mathrm{i}}/\sigma^{\mathrm{i}}$ becomes significant. Such conditions might occur in regions of the magnetosphere where there are large increases in mass density due to local ionization of neutral material, where there are strong radial motions of the plasma or where there are sharp gradients in plasma pressure. When the magnetosphere imposes too large a load, the ionospheric clutch begins to slip.

Next let us consider how the momentum of the solar wind may be harnessed by processes occurring near the magnetopause where the external solar magnetic field interconnects with the planetary magnetic field. At the poles the planetary magnetic field lines are open to the solar wind. The solar wind drives a plasma flow across the polar caps and the field lines from the polar region move in the direction of the solar wind flow, being pulled by the solar wind over the poles and back into the extended magnetotail. Conservation of flux requires that field lines are further cut and reconnected in the tail.

The MHD condition of the field being frozen to the flow can be written as $\boldsymbol{E} + \boldsymbol{V} \times \boldsymbol{B} = 0$ which allows the convection electric field to be written $\boldsymbol{E}_{\mathrm{cv}} = -\eta \boldsymbol{V}_{\mathrm{sw}} \times B_{\mathrm{o}}/R_{\mathrm{M}}^3$ (where η is the efficiency of the reconnection process in harnessing the solar wind momentum, $\sim$0.1 for the Earth). In simple magnetospheric models E_{cv} is assumed constant and the corresponding circulation is given by the $\boldsymbol{E} \times \boldsymbol{B}$ drift, $V_{\mathrm{cv}} = \eta V_{\mathrm{sw}}(r/R_{\mathrm{M}})^3$ where R_{M} is the magnetopause distance. After being carried tailward at high latitudes, the plasma then drifts towards the equatorial plane and eventually returns in a sunward flow to the dayside magnetopause.

Comparison of the corresponding electric fields indicates whether the magnetospheric circulation is driven primarily by the solar wind or the planet's rotation. Since E_{co} is proportional to R^{-2} and E_{cv} proportional to R^3 it seems reasonable to expect that corotation dominates close to the planet while solar wind driven convection dominates outside a critical distance R_c. Thus, the magnetosphere corotates out to a distance of $R_{\mathrm{c}}/R_{\mathrm{M}} = (R_M\Omega/\eta V_{\mathrm{sw}})^{1/2}$. This formula simply says that magnetospheres of rapidly rotating planets with strong magnetic fields are dominated by rotation while the solar wind controls the plasma flow in smaller magnetospheres of slowly rotating planets.

Orientation and seasonal variations

If a planetary rotation axis is not approximately aligned with either the magnetic dipole axis or the solar wind flow direction then there exists no reference frame in which the plasma flow is steady. The degree of variability over a planetary rotation period is governed by the solar wind angle (see figure 3), the angle between the solar wind and the dipole axis. When the solar wind angle is near 90° the configuration is symmetrical (as approximated by Earth, Jupiter and Saturn). When the solar wind angle is small the configuration has the magnetic axis pointed 'pole-on' into the solar wind, a configuration that was expected for Uranus before Voyager 2 found a large dipole tilt. These configurations lead to very different magnetic field topologies. Figure 3 shows how the solar wind angle varies for the major magnetospheres. At the time of the Voyager 2 encounter in 1989, Neptune's northern hemisphere was close to mid-winter with the rotation axis tipped 23° degrees from the Sun, i.e. 67° from the radial, solar wind direction. The 47° tilt of the magnetic dipole means that the angle between the solar wind and the dipole axis changes between 20° and 114° over the 16.1 h planetary rotation. Complete re-configuring of the magnetosphere must occur every planetary rotation (figure 6(*b*)).

A planet's obliquity is responsible for seasonal variations over the planet's orbital period. For a planet such as Jupiter with a very small obliquity there are very small changes in the solar wind angle and, hence, magnetospheric configuration over Jupiter's 12 yr orbital period. On the other hand, the high obliquity of Uranus leads to dramatic variations in the solar wind angle (covering all orientations) over the planet's 84 yr orbital period. Earth, Saturn and Neptune, with moderate obliquities (23.5°–30°), are intermediate cases with moderate seasonal variations.

Energetic particle populations

All magnetospheres have significant populations of particles with energies (keV–MeV) well above the thermal population. These particles are largely trapped by the strong planetary magnetic field in long-lived radiation belts. Where do these energetic particles come from? Since the interplanetary medium includes energetic particles of solar and galactic origins an obvious possibility is that these energetic particles were 'captured' from the external medium. In the cases of the giant planets, the observed high fluxes are hard to explain without additional internal sources. Compositional evidence also implies that some fraction of the thermal plasma is accelerated to high energies, either by tapping the rotational energy of the planet, in the cases of Jupiter and Saturn, or by processes in the distorted magnetic field in the tail, in the cases of Earth, Uranus and Neptune. Particle drifts in a non-uniform magnetic field lead to the ions and electrons drifting in opposite directions around the planet, producing an azimuthal electric current: the *ring current*. If the energy density of the energetic particle populations is comparable with the magnetic field energy then the ring current produces a magnetic field that significantly perturbs the planetary magnetic field. This is the case for Jupiter and Saturn, where the high particle pressures inflate and stretch out the magnetic field and generate a strong ring current in the magnetodisk. While Uranus and Neptune have significant radiation belts, the energy density remains small compared with the magnetic field

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

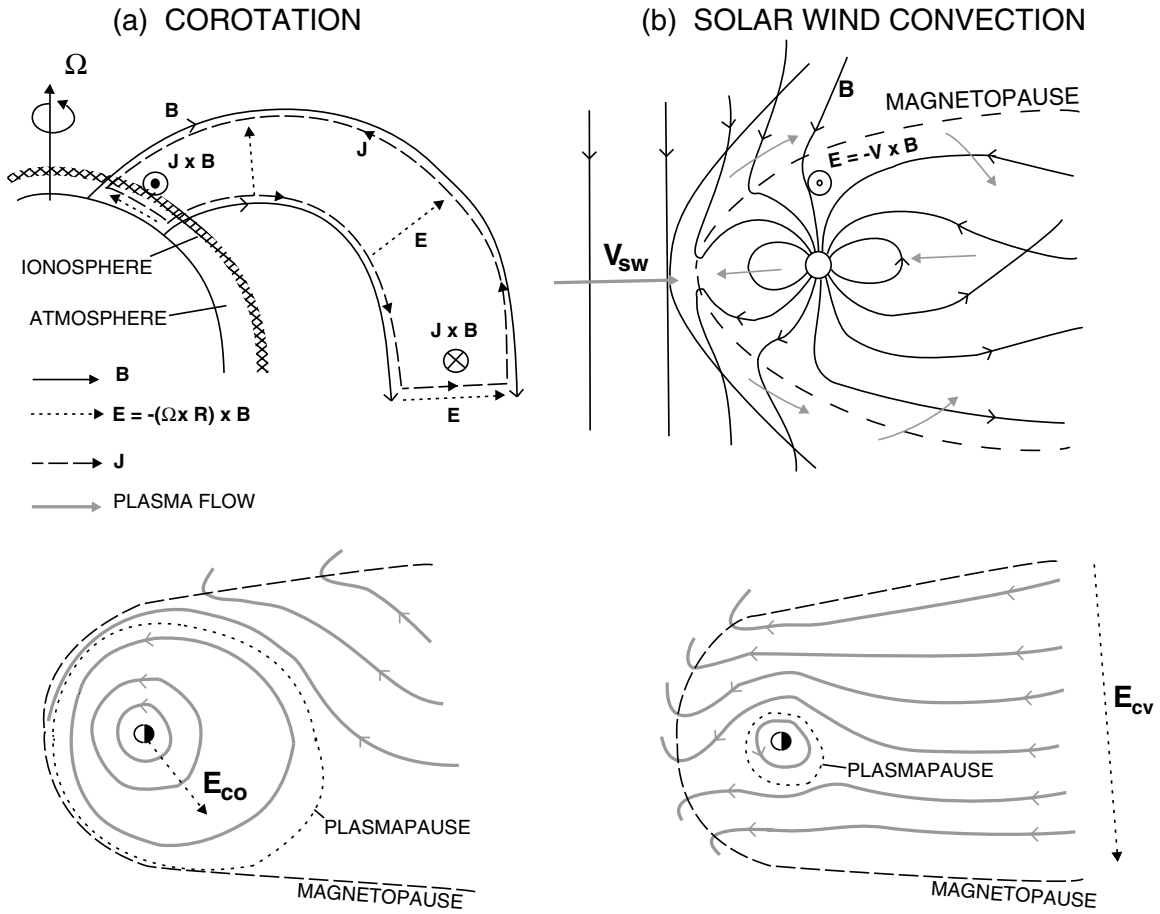


Figure 8. Large scale magnetospheric circulation driven by (*a*) corotation and (*b*) the solar wind. In each case, the upper (lower) diagram shows the meridional view (the view in the equatorial plane).

and the ring current is very weak. The Dessler–Parker–Sckopke relation compares the magnetic field produced by the ring current with the kinetic energy of the trapped particle population, scaled to the dipole magnetic energy external to the planet. Applying the Dessler–Parker–Sckopke relation to planetary magnetospheres we find that while the total energy content of magnetospheres varies by many orders of magnitude and the sources are very different, it appears that the particle energy builds up to only 1/1000 of the magnetic field energy in each magnetosphere. Earth, Jupiter and Saturn all have energetic particle populations close to this limit. The radiation belts of Uranus and Neptune are much less than this limit, perhaps because it is harder to trap particles in non-dipolar magnetic fields.

Where do these energetic particles go? The majority appear to diffuse inwards towards the planet. Loss processes for energetic particles in the inner magnetospheres are satellite absorption, charge exchange with neutral clouds and scattering by waves so that the particles stream into the upper atmospheres of the planets where they can excite auroral emission and deposit large amounts of energy.

Bibliography

Bagenal F 1992 Giant planet magnetospheres *Ann. Rev. Earth Planet. Sci.* **20** 289

Cravens T E 1997 *Physics of Solar System Plasmas* (Cambridge: Cambridge University Press)

Gombosi T I 1998 *Physics of the Space Environment* (Cambridge: Cambridge University Press)

Kivelson M G and C T Russell (ed) 1995 *Introduction to Space Physics* (Cambridge: Cambridge University Press)

Van Allen J A and F Bagenal 1999 Planetary magnetospheres and the interplanetary medium *The New Solar System* ed J K Beatty, C C Petersen and A Chaikin (Cambridge: Cambridge University Press) pp 39–58

F Bagenal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Nebula Central Star Mass Loss/Winds

The ultraviolet spectra of the central stars of PLANETARY NEBULAE frequently show broad emission–absorption features which are indicative of stellar winds with typical velocities of several thousand km s^{-1}. The collision of these winds with the surrounding slower-moving nebular gas is important in shaping the nebula. In some cases, mass loss in the form of narrow jets is observed, which may indicate the presence of a companion to the central star.

Mass loss and the formation of planetary nebulae

Planetary nebulae are isolated clouds of gas, ionized by a hot central star. It is generally accepted that planetary nebulae form during the penultimate stage in the evolution of low- and intermediate-mass ($<8M_{\odot}$) stars, after they have become RED GIANTS for the second time. The red giant progenitor eventually expels its envelope at a relatively low velocity ($\sim$15 km s^{-1}), and this ejected gas is subsequently ionized by the ultraviolet radiation of the red giant core. The stellar core—now the central star—will eventually cool to its final configuration, a WHITE DWARF. After the primary mass-loss event which expelled the nebula, mass loss from the central star does not immediately cease. A low-density, high-velocity (typically, several thousand km s^{-1}) stellar wind develops, driven by the radiation pressure of the central star (see RADIATIVELY DRIVEN STELLAR WINDS FROM HOT STARS). This wind will have important dynamical effects on the morphology and dynamics of the nebula, sweeping the previously ejected envelope into a well-defined shell.

The details of the mechanism by which the red giant (technically, an *asymptotic giant branch star*) expels its envelope remain obscure. We know that, as the star evolves, it experiences a series of thermonuclear flashes in a shell near the degenerate core where helium is being converted to carbon. With each shell flash, the star becomes more luminous and its radius increases. A slow ($\sim$10 km s^{-1}) stellar wind ensues, which carries off an increasing amount of mass; the mass loss eventually reaches values of $\sim 10^{-4} M_{\odot}$ yr^{-1}, the so-called 'superwind phase'. Such a wind could by itself remove the stellar envelope in a few thousand years. On the other hand, the mass loss may terminate with a more dramatic event. One suggestion is that mixing of carbon into the base of the envelope increases the opacity suddenly, increasing the radiation pressure and lifting off the entire envelope. In the case of stars with close binary companions, the companion may spiral in to the giant star, spinning up and ejecting its envelope: we know such interactions must occur, because some planetary nebula central stars have binary companions whose distance from the central star is less than the size of their asymptotic giant precursors (see CLOSE BINARY STARS).

Following the removal of the hydrogen-rich envelope, there will be a period when the central star is not hot enough to produce much ultraviolet radiation ($T_{\rm eff} <$ 30 000 K) and the nebula will not be ionized. During this phase, the central star is heavily obscured by a dense cloud of neutral gas. This is the PROTOPLANETARY NEBULA phase. Many of these nebulae are strongly asymmetric and have bipolar jets. The jet velocity may be a few hundred km s^{-1}, intermediate between the speed of the nebula and the speed of the later stellar winds. It is suspected that many of the morphological features seen in mature planetary nebulae are determined in the protoplanetary stage.

Observational evidence of winds

The presence of stellar winds is usually signaled by the appearance of 'P-Cygni profiles' in the star's spectrum. These are broad lines with a deep absorption component on the short-wavelength ('blue') side and an emission component on the long-wavelength ('red') side. The absorption component is formed in the gas between the star and the observer, which is blueshifted by the DOPPLER EFFECT as it streams outwards, while the emission component is due to gas beside and beyond the star. P-Cygni profiles are only seen in strong lines of those ionization stages which are dominant in the wind. Because of the high temperatures of planetary nebula central stars, these are lines at ultraviolet wavelengths, such as C IV 155 nm and N V 124 nm. As a result, it was only with access to the ultraviolet by satellites such as the International Ultraviolet Explorer (IUE) that it was discovered that up to half of all central stars have winds.

The maximum blueshift of the absorption component of the P-Cygni profile provides v_{∞}, the terminal velocity of the stellar wind. It is believed that the wind is driven by the pressure of radiation scattered by strong absorption lines, and v_{∞} is always a few times the escape velocity from the stellar surface. Since the surface gravity (and hence the escape velocity) increases greatly as the central star evolves toward the white dwarf stage, we find a considerable range of v_{∞}, with values generally between 500 and 4000 km s^{-1}.

Another wind parameter is the density, which multiplied by v_{∞} gives us the rate of mass loss, $\dot{M}$. The mass-loss rate is more difficult to determine than v_{∞}, but results are available for a number of stars. $\dot{M}$ is found to vary greatly and is not a function of a single parameter such as stellar temperature or v_{∞}. In table 1 we give values for a few central stars, chosen to illustrate the range of properties encountered. The second column gives the surface temperature of the star, $T_{\rm eff}$, in thousands of kelvins, but it should be noted that the definition of the 'surface' is somewhat arbitrary for a star with an extended atmosphere. The last column is the mass-loss rate in units of $10^{-8} M_{\odot}$ yr^{-1}.

Part of the variation in $\dot{M}$ is associated with the chemical composition of the central star's atmosphere. One group of central stars retain a remnant (less than $\sim 10^{-4} M_{\odot}$) of the red giant envelope, which gives them a hydrogen-rich stellar atmosphere—they are stars of spectral type 'O' or 'Of'. The other group has hydrogen-poor atmospheres: a typical chemical composition would

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Parameters for selected central stars.

Name	$T_{\rm eff}$ (K)	Type	v_∞ (km s^{-1})	$\dot{M}$ (10^{-8} $M_\odot$ yr^{-1})
BD+30 3639	30	WR	700	600
IC 418	37	Of	700	26
IC 4593	40	Of	900	10
NGC 2392	45	Of	400	$\leq$3
NGC 6543	48	Of–WR	1600	16
IC 4637	55	Of	1500	$\leq$2
NGC 3242	75	Of	2300	$\leq$2
NGC 40	78	WR	1000	250
Abell 30	110	O VI	4000	5
Abell 78	115	O VI	3700	2.5
NGC 6369	150	WR	1200	70

be 60% helium, 30% carbon and 10% oxygen. These stars may show Wolf–Rayet spectra. Wolf–Rayet spectra are dominated by broad emission lines of carbon and helium; such stars have very extended atmospheres and strong stellar winds. (Planetary nebula central stars with Wolf–Rayet spectra should not be confused with the more massive and luminous 'classical' WOLF–RAYET STARS.) Not all hydrogen-poor central stars are Wolf–Rayet stars, however. Those which are hotter and which have higher gravities may be PG 1159 STARS or O VI stars. The PG 1159s are hot, hydrogen-deficient stars with at most feeble stellar winds. Some of them pulsate, and only about half of the known members of this class are associated with planetary nebulae. The O VI stars (now usually called (WC)-PG 1159 stars) seem to be intermediate between Wolf–Rayets and PG 1159s, very hot but still having a substantial fast wind. In fact, one PG 1159 star, Longmore 4, became an O VI star for a few weeks. During this outburst, the carbon and helium absorption lines went into emission and a stellar wind developed; the star then returned to its previous state.

We might summarize by saying that, as planetary nebulae evolve, the temperatures and surface gravities of their central stars increase. As a result, the velocities of the central star winds increase, while the wind mass-loss rate drops. However, there is a second factor at work: those stars which have lost their hydrogen will have atmospheres of higher opacity, increasing the effective radiation pressure and driving stronger winds than in the case of stars with a hydrogen atmosphere.

Dynamical effects of winds on nebulae

After it was discovered that strong central star winds were common, it was realized that these winds could have a dynamical effect on the nebula itself. Pikel'ner considered such effects in 1968 and by the early 1980s the 'two-wind model' developed by S Kwok and others was widely accepted. In this model, the fast stellar wind overtakes the earlier slow red giant wind and sweeps it into the hollow shell we see as a planetary nebula. Because the nebulae are rarely spherical, but often have an axis of symmetry, it was necessary to modify the two-wind model by postulating

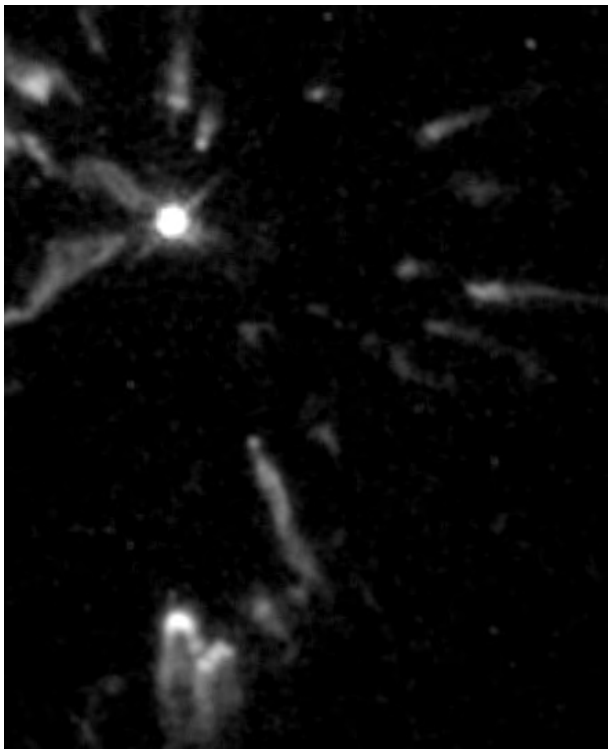


Figure 1. The wind from the central star of Abell 30 (upper left) sweeps past nearby nebular material.

that the red giant wind was not spherical but had a denser equatorial plane. Computer simulations based on this modified two-wind model have been successful in reproducing many of the basic observed morphologies. Such models, however, leave unanswered the question of why the red giant wind should be strongly aspherical, since it is not expected that the red giant would have a significant rotational velocity. This has led to speculation about the presence of companions—which might include substellar brown dwarfs or even massive planets—which could spin up the progenitor star and impart the needed asymmetry to the slow wind.

Another outstanding question concerns the physical processes affecting the fast stellar wind after its collision with the nebula. A wind moving at 1000 km s^{-1} will be heated to over 10^7 K by such a collision. If this gas is able to cool within a few hundred years, only its momentum will accelerate the nebula. If it does not cool, however, the heated wind will fill the interior of the nebula, and the pressure of this hot, thin gas will be more effective in accelerating the planetary nebula outwards than the momentum alone. In the case of one nebula, BD+30 3639, x-rays have been observed from a 3×10^6 K plasma, providing direct evidence for the still-hot bubble of shocked stellar wind.

The high spatial resolution provided by the HUBBLE SPACE TELESCOPE (HST) has revealed a wealth of unexpected structure in planetary nebulae. Many nebulae have been

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

found to have neutral globules embedded in the ionized gas, which are in the process of evaporation. The comet-like tails streaming outwards from these globules give dramatic evidence of the interaction of the stellar wind with the nebula (figure 1).

Jets in planetary nebulae

In a few nebulae, such as NGC 6543 and IC 4593, HST images have revealed narrow jets. Many more nebulae have been found to possess 'point symmetry', where features are located symmetrically with respect to the star, even though there is no single *axis* of symmetry. Such structures immediately suggest the action of a bipolar jet, with a shifting axis. The discovery of such structures came as a surprise, because previously known stellar jets, such as those from protostars, arise from ACCRETION DISKS, which are not expected around isolated planetary nebula central stars—stars from which material has been expelled not accreted. Nevertheless, the existence of jets must be explained, and many astronomers now believe they are strong evidence for the existence of some sort of companion to the central star. The presence of such a companion could allow mass transfer, resulting in at least a transient accretion disk and collimated jets. As mentioned above, there are planetary nebulae where the central star has a close binary companion and which must have evolved through a mass-transfer phase. Companions have also been invoked to explain the dense equatorial wind of the modified two-wind model. These considerations point to a whole range of complicated mass-loss phenomena that have yet to be explored.

Bibliography

There is some discussion of the central stars of planetary nebulae in this recent text:

Lamers H J G L M and Cassinelli J P 1999 *Introduction to Stellar Winds* (Cambridge: Cambridge University Press)

A good entry point into the literature is:

IAU Symposium No 180: Planetary Nebulae ed H J Habing and H J G L M Lamers (Dordrecht: Kluwer, 1997)

Patrick Harrington

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Nebulae

Planetary nebulae, shells of illuminated gas that surround old and dying stars, were discovered by William Herschel, who in 1785 referred to the 'planetary', or 'disk-like', appearance of the object we now call NGC 7009 (figure 1). He later discovered that one of his 'planetaries' (NGC 6543) had a central condensation or star. Some 2000 planetary nebulae are known today, the best known being the Ring Nebula in Lyra (see figure 2). All have central stars. The central stars were once the nuclear burning zones of giant stars that have lost their outer envelopes and are on their way to becoming WHITE DWARFS, the fleeing envelopes making the ephemeral planetaries.

Spectra and parameters

Although Herschel thought the nebulae were made of 'shining fluid', their actual natures were not proven until William Huggins examined NGC 6543 (figure 3) in 1864 with a visual spectroscope and found three emission lines, demonstrating that the nebulae are made of low-density gas. One in the blue (Hβ at 4861 Å) was quickly associated with hydrogen, but the green lines at 4959 Å and 5007 Å defied identification, becoming known as the 'nebulium' lines. Huggins and others subsequently discovered additional emission lines, including more from hydrogen, others later found to be helium, and yet others that added to the mysterious 'nebulium' list.

The origin of the hydrogen emissions was worked out in 1927 by Hermann Zanstra. The blue central stars are very hot, above 25 000 K, and emit copiously in the ultraviolet beyond the Lyman ionization limit of hydrogen at 912 Å. The nebula is primarily a plasma filled with protons, free electrons and neutral and ionized atoms that have their electrons in the lowest orbits or energy levels (the ground states). Each Lyman photon is capable of ionizing one hydrogen atom, the electron freed with an energy equal to the difference between the ionizing photon's energy and hydrogen's ionization energy. The electron is eventually recaptured by another proton and can land on any energy level. The best chance is for the

Figure 1. The discovery object, NGC 7009, is about 3500 ly away and about 1 ly across. Ionized by a hot star with a temperature of 90 000 K, the nebula consists of a complex, highly elliptical shell punched through by 'FLIERS' that make the ansae that gave the nebula the name 'Saturn Nebula'. B Balick and J Alexander (University of Washington), A Hajian (US Naval Observatory), Y Terzian (Cornell University), M Perinotto (University of Florence) and P Patriarchi (Arcetri Observatory), Space Telescope Science Institute and NASA.

Figure 2. The Ring Nebula in Lyra (M57, NGC 6720), the best known of all planetary nebulae, is viewed in close to true color by the Hubble Space Telescope. Blue indicates ionized helium, green doubly ionized oxygen, and red hydrogen and ionized nitrogen. The 15th magnitude central star has passed its maximum temperature and (although still at about 150 000 K) is cooling and dimming. H Bond (Space Telescope Science Institute) and the Hubble Heritage Team (AURA, Space Telescope Science Institute and NASA). **This figure is reproduced as Color Plate 39**.

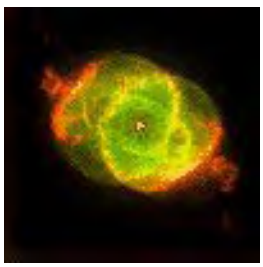


Figure 3. NGC 6543, an enormously complex planetary nebula imaged by Hubble, was the first nebula to be observed spectroscopically. The young central star, with a temperature of 47 000 K, is still heating as the nebula expands. J P Harrington (University of Maryland) and K Borkowski (North Carolina State University), Space Telescope Science Institute and NASA.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

electron to land back in the ground state. If it does, the UV photon is re-created and will ionize another atom.

Absorption and re-creation continue until the free electron lands in an upper level. If it goes to the second, it produces a photon of the Balmer continuum, which flies out of the nebula, as there are few atoms able to absorb it. From level 2, the electron drops to level 1, creating a Lyman α photon that bounces around in the nebula until it escapes or is destroyed by hitting a dust grain, by ionizing another element or in the creation of a pair of lower-energy photons that escape.

If the electron lands above level 2, it has several choices. If it first goes to, say, level 4 (as in figure 4), it can jump directly to level 1, creating a Lyman γ photon that is immediately reabsorbed by a neutral atom, the electron returning to level 4. Eventually the electron will jump to level 2 or 3. If it goes to 2, a Balmer photon is created that leaves the nebula. If it goes to 3, the same process operates until the electron jumps to level 2. The result is that each Lyman photon from the star produces one Balmer photon from the nebula. The ionization probabilities are high, while the recombination probabilities are low. As a result, the optically visible nebular gas is well over 99% ionized, there being essentially one free electron for each proton. In a real nebula, a variety of atoms act in a similar way to produce 'recombination lines' of helium, oxygen, nitrogen, neon and others.

The nebulium problem was solved in 1928 by I S Bowen. While exploring the energy-level structure of doubly ionized oxygen, he found levels just above the ground state with energy differences appropriate to Huggins' 4959 Å and 5007 Å (and eventually other) lines. Downward transitions from these levels to the ground state are 'forbidden' under simplified quantum rules. In truth, they are allowed, just improbable, the upper levels being 'metastable' with long lifetimes. Electrons freed from hydrogen with sufficient energy (a few electron volts for the optical lines) collide with O^{2+} ions and kick the bound electrons into the metastable states. Upward and downward collisions establish an equilibrium and relatively high populations in the metastable states. Given a stable population, there will be a steady stream of downward radiative transitions that produce the 'forbidden lines', denoted by square brackets around the spectral signatures (the forbidden lines of doubly ionized oxygen called '[O III]', etc). Forbidden lines are produced by various ionization stages of oxygen, nitrogen, neon, sulfur, iron and other elements from the ultraviolet through the infrared. Although the forbidden lines are intrinsically weak, the recombination lines are weaker (the odds of a recombination quite low as a result of low density). The strongest forbidden lines therefore dominate the nebulae.

Temperatures, densities and chemical compositions are easily derived. The free electrons thermalize quickly by rapid collision and establish gas kinetic (electron) temperatures (that reflect velocities) between about 8000 K and 20 000 K that depend on the balance between heating and cooling. The gas is heated by ultraviolet radiation from the central star, so the hotter the star (temperatures range to over 200 000 K), the hotter the nebula. The electrons are cooled when they lose energy in exciting the forbidden lines, principally [O III], so the more heavy atoms, the cooler the nebula. Electron temperatures are measured by comparing forbidden lines from a given ion that arise from levels that lie at different energies above the ion's ground state, for example from the ratio of the 4363 Å and 5007 Å intensities of [O III] or the 5754 Å and 6584 Å intensities of [N II]. Electron densities (in the range from 10 to 10^5 electrons cm^{-3}) can be measured from pairs of forbidden lines that arise from closely spaced levels that have different excitation and radiative probabilities, notably from the ratios of the 3729 Å and 3726 Å lines of [O II] and the 6731 Å and 6717 Å lines of [S II].

Once the physical conditions are known, we find the ionic abundances of the nebulae from the strengths of the recombination and forbidden lines. To find the elemental compositions we add the individual ionic compositions and use predictive radiative models to calculate the abundances of unobserved ions, in effect fine-tuning the model's abundances and parameters until the predicted spectra match the observed spectra. Planetary nebulae in the Galactic disk have basic compositions similar to those of the Sun and other POPULATION I STARS, while those in the Galactic halo are notably deficient in heavy elements. Superimposed on these basic compositions can be powerful enrichments of helium, nitrogen and carbon, these elements having been created in the giant stars and dredged to the stellar surfaces before the creation of the nebulae. Helium abundances can be double normal, while those of nitrogen and carbon can be up by over a factor of 10. Nitrogen- and helium-enriched nebulae (figure 5) are generally referred to as 'Type I', the other chemical types (whose terms are rarely used) having lower enrichments or none at all.

Distances, dimensions and masses

Physical diameters of the planetaries follow from angular diameters (which range from sub second of arc to a quarter of a degree) and distances. Masses equal the volumes multiplied by the mean proton (electron) densities multiplied by the proton mass. Distances, however, are poorly known. All but a handful of nebulae are too far away for PARALLAXES, and these have high errors. Distances can be found for some objects through association with clusters or companions, from their degree of interstellar reddening by dust or by comparing their angular expansions with their measured expansion velocities. We can also examine nebulae in the MAGELLANIC CLOUDS. From all these we find ionized masses typically a few tenths that of the Sun.

Since the total energy radiated in the standard Hβ line, $E(\mathrm{H}\beta)$, depends on the product of the electron and proton densities (multiplied by a recombination coefficient), $E(\mathrm{H}\beta)$ can be related to mass. The flux in Hβ at the Earth, $F(\mathrm{H}\beta)$, depends on $E(\mathrm{H}\beta)$ and distance. Adopting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

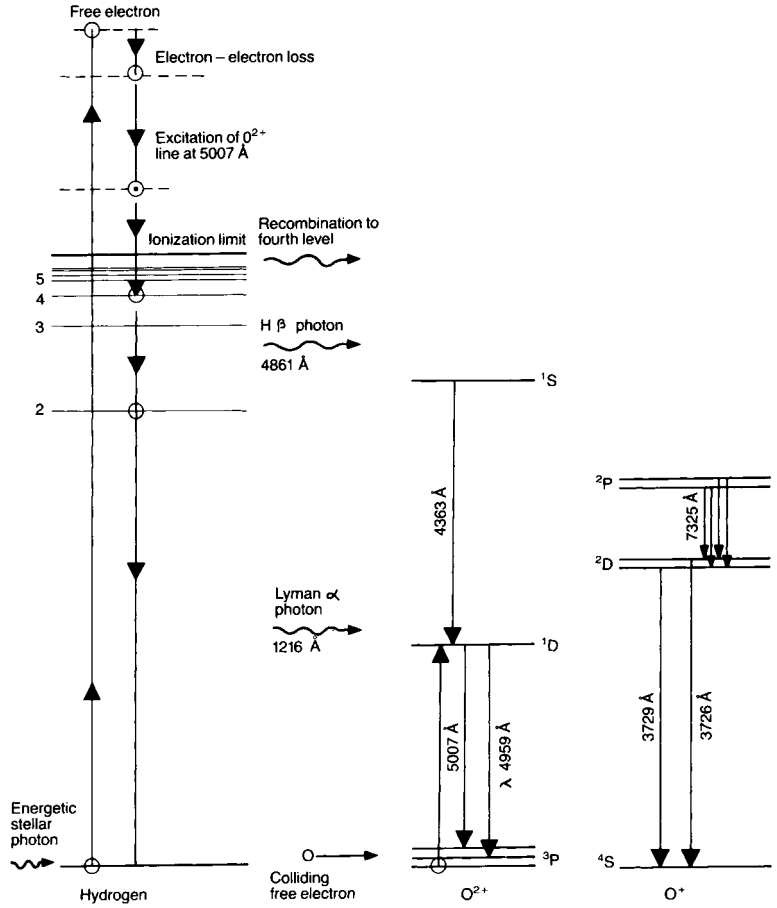


Figure 4. Nebular emission lines are produced by a combination of ionization, recombination and collisional excitation. On the left, a hydrogen atom is ionized by a stellar photon. The freed electron undergoes various energy losses and recombines with another proton (here to the fourth energy level) to create the hydrogen lines. While free, the electron can collisionally excite the electrons of heavier atoms into long-lived metastable states, descent from which makes the forbidden lines. In the center we see the excitation of doubly ionized oxygen, at right of singly ionized oxygen (these two diagrams doubled in size for clarity). The ratio of the intensities of the lines at 4363 Å and 5007 Å (or 7325 Å with 3726 Å + 3729 Å of [O II]) gives electron temperature, whereas the 3726:3729 intensity ratio gives density. From Kaler J B 1989 *Stars and their Spectra* (Cambridge: Cambridge University Press).

a standard mass allows the computation of 'statistical distances'. These are used to find most nebular radii, which range from about 0.01 pc to over 1 pc. Distances combined with radial velocities allow determination of nebular kinematics. Nebulae in the disk have the motions of other population I stars, those in the halo having high velocities typical of population II.

The ionized mass is only a portion of a nebula. Planetaries fall into two broad categories, those in which the central star ionizes all the surrounding matter and some of the stellar radiation escapes ('optically thin') and those in which the radiation is all absorbed, leading to a bubble (a Strömgren sphere) of ionization that is nested inside a neutral shell ('optically thick'). In this simplified view, only optically thin nebulae are subject to the above statistical distance method, as the optically thick planetaries may have very low ionized masses, leading to distance overestimates, although corrective estimates can be made with enough calibrating objects.

Reality is more complex. A nebula may be optically thin in one direction, thick in another. Planetaries also exhibit a great range in total mass, which leads to an even greater range in ionized mass, from $0.03M_{\rm Sun}$ to $0.5M_{\rm Sun}$ or so (making statistical distances even more uncertain). Nebular structures are complex as well, commonly consisting of nested shells and disks. A dense inner shell may be optically thin, radiation leaking to a more tenuous outer shell that may be optically thick. This shell may also be optically thin, radiation leaking to a larger outer halo that may also be thick or thin. If the inner shells are thick, the outer will not be visible, or might be seen as fragmented if radiation leaks in one direction and not in another.

High-dispersion spectra reveal internal velocities. The emission lines are broadened or split from the expansion of a shell. Expansion velocities increase

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

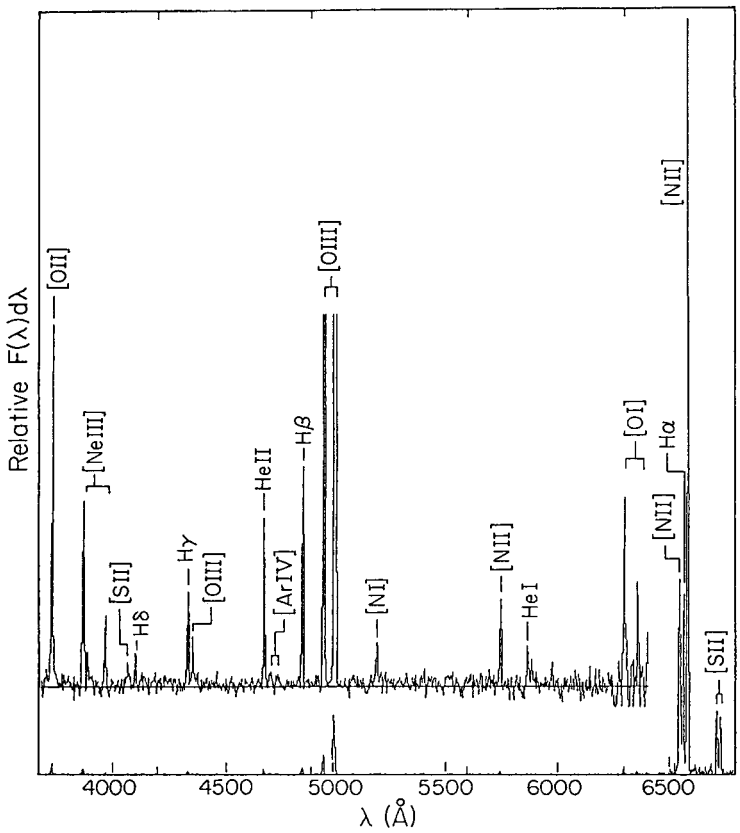


Figure 5. The spectrum of a type I nebula (BV-1), one highly enriched in nitrogen and helium, displays powerful singly ionized nitrogen lines as well as a host of others. Two scales are used to show the full range of intensities. NOAO, from Kaler J B, Chu Y-H and Jacoby G H 1988 *Astron. J.* **96** 1407.

outward from the center, and are typically a few tens of kilometers per second. From dimensions and expansion velocities we find ages that range from a few hundred to tens of thousands of years. Some 50 000 yr after the nebulae are formed by ejection from their parent stars, the nebulae merge with the interstellar medium and are too faint to be seen, leaving behind white dwarfs.

Infrared observations of molecular emissions show the traditional nebulae to be surrounded by molecule-rich shells in which we find H_2, CO and even polycyclic aromatic hydrocarbons. These shells may or may not also be visible in the optical. Radiation longward of 912 Å leaks into the molecular gas, exciting it and producing a 'partially dissociated region' that will someday become ionized as the nebula expands and high-energy radiation attacks it. In other cases, molecular gas is rather thoroughly mixed in with the ionized gas, at least in projection, and in some cases physically. Intimately involved is dust, which radiates an infrared continuum and emits spectral signatures of carbon and silicates. Dust can both be buried in the nebula and surround it. When all these structures are taken into account, nebular masses—ionized, neutral and molecular—can range upward to well beyond solar.

The central stars

Excepting neutron stars, planetary nebula central stars are the hottest known and are exceedingly luminous. They are faint to the eye only because of high bolometric corrections for invisible ultraviolet radiation. From Zanstra's mechanism, the total Balmer photon luminosity equals the stellar ultraviolet photon luminosity. Recombination theory relates the total Balmer photon luminosity to the Hβ energy luminosity. Assuming a blackbody, the stellar ultraviolet photon luminosity can be related to its ultraviolet energy luminosity, which can therefore be found from the Hβ energy luminosity (and distance). Comparison of the stellar ultraviolet luminosity with the star's visual luminosity yields the effective (surface) temperature (the 'Zanstra temperature'), distance factoring out. If a nebula is optically thin and ionizing radiation escapes, we do not count all the emerging ultraviolet photons and therefore derive only a lower limit to temperature.

Because the strengths of the ionized helium lines depend on the amount of ionizing radiation shortward of the helium Lyman limit at 228 Å, a similar temperature can be found by comparing the flux in the He II 4686 Å line (Paschen α) with the visual magnitude. A nebula

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

may be optically thin for hydrogen, but thick for helium. As a result, the He II Zanstra temperature is almost always greater than or equal to the hydrogen Zanstra temperature. If the two are equal, the nebula is probably thick. Reliable Zanstra temperatures range from around 25 000 K to nearly 300 000 K. Temperatures combined with visual magnitudes and distances yield bolometric luminosities that range from a few solar luminosities to tens of thousands. Other temperature methods use the slope of the stellar flux in the observed ultraviolet and the determination of the energy balance in the nebula, equating heating (from the recombination lines) with cooling (from the forbidden lines).

Temperatures and luminosities, as well as surface gravities and stellar compositions, can also be found by modelling the stellar absorption line spectra. Surface gravities and assumed masses combined with temperatures yield luminosities and thus another way to find distances. Modelled temperatures are in the range of the Zanstra temperatures. Specific disagreements with Zanstra temperatures suggest that some stars may not be blackbodies.

Central star spectra are varied and complex. Some stars exhibit simple O-type absorption spectra. Lower-luminosity stars can be hydrogen rich or hydrogen poor, appearing as preludes to different kinds of white dwarfs. Higher-luminosity stars commonly show emissions from winds, some mimicking the spectra of the carbon WOLF–RAYET (WC) STARS. Spectra of lower-temperature stars can be so filled with emission lines that the continuum cannot even be seen. At the high-temperature end, the stars may radiate strong O VI lines. Many of these 'O VI stars' vary with periods of minutes, much as do the PG 1159 white dwarfs. We commonly observe P Cygni lines in the ultraviolet that reveal terminal wind velocities that range from a few hundred km s^{-1} to 5000 km s^{-1} and give mass loss rates of 10^{-7}–10^{-9} solar masses per year. The winds can be variable, leading to variations in line strengths. The winds also add considerable complications to modelling and may result in poorly determined temperatures (see PLANETARY NEBULA CENTRAL STAR MASS LOSS/WINDS).

Wind modelling, however, shows enrichments in helium in the Of type of spectra, with He/H ratios of 0.2 or so. The WC spectra exhibit no hydrogen lines. These helium stars have C/He ratios of unity and oxygen mass fractions of 4–10%. White dwarf DA and DB characteristics begin early, during the planetary nebula stage. The compositions of the stars are consistent with those of the dust grains found in the nebulae, carbon-rich stars having carbon grains, the others having silicate grains.

Temperatures and luminosities allow the placement of planetary nuclei on the HR diagram (the log L–log T plane) for the establishment of empirical evolutionary tracks and for comparison with theoretical tracks (figure 6). Distance errors, however, greatly smear the luminosities. The best recourse is to use planetary nebulae in the Magellanic Clouds, although the low metallicity of the Clouds may make these nebulae different.

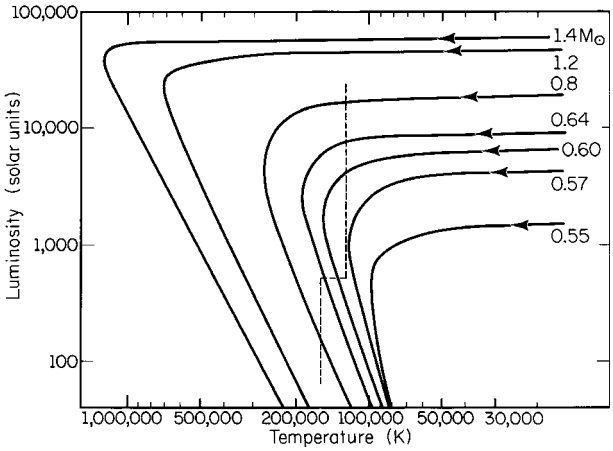


Figure 6. Planetary nebula central stars evolve from the right (from AGB stars) along horizontal evolutionary tracks, reach maximum temperatures, and then cool and descend to become white dwarfs, the tracks depending critically on final (core) masses. Most observed planetary nebula stars fall to the right of the dashed line. From R A Shaw, evolutionary tracks by I Iben Jr.

The direction of evolution on an evolutionary track is revealed by nebular expansion. The stars enter the planetary regime of the HR diagram at high luminosity and low temperature, then move to the left, increasing T at constant L, where they exhibit the wind phenomena that make for highly divergent spectra. At a high temperature that typically falls between 100 000 and 200 000 K, the empirical track reverses, and the stars descend in luminosity while cooling to temperatures of 70 000 K or so, at which point the nebulae effectively dissipate. The empirical track leads directly to the tracks followed by white dwarfs. The birth rate of planetary nebulae found from the space density (pc^{-3}) and the mean expansion velocity is of the order of 10^{-12} pc^{-3} yr^{-1}, within the large errors comparable with the birth rate of white dwarfs.

Since asymptotic giant branch (AGB) stars (those that have completed core fusion and have contracting carbon–oxygen cores surrounded by fusing shells of hydrogen or helium) are losing their outer envelopes at fierce rates, and their luminosities are similar to those of planetary nuclei on the heating 'horizontal track', it is clear that they are the planetaries' predecessors. The planetary nebula nuclei are therefore essentially the spent carbon–oxygen cores that are still surrounded by hydrogen-burning shells and topped by thin hydrogen envelopes. The planetary nebulae therefore link the AGB stars directly to the white dwarfs.

Theory fits observation. The luminosity of a remnant core on a horizontal heating track increases with its mass, which in turn depends on initial mass. The cores run from about 0.55 solar masses for an initial mass of 0.8 solar to the Chandrasekhar limit for a 10 solar mass star. As the remnant outer envelope is thinned from the top by a fast hot wind and from the bottom by nuclear fusion, the radius

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the core goes down and the effective temperature goes up, the star traversing the HR diagram to the left. The traversal time depends on a high power of the core mass. As a result, masses on the horizontal tracks are strongly concentrated toward an upper limit around 0.6 solar masses. Planetary nebulae thus become useful distance indicators for other galaxies. They are easy to find and the upper limit of their brightness distribution represents a standard candle for distances to tens of megaparsecs.

The turnaround point on an evolutionary track at specific mass is reached when the nuclear burning begins to shut down, the stars then cooling and dimming. The high-temperature maximum is proportional to the core mass, low-mass stars reaching only 80 000 K or so, high ones exceeding 200 000 K. The pace of evolution for different masses then changes, allowing the higher-mass cores to be observed. The ages of the nebulae can be found by comparing the positions of their stars with the calculated rates of evolution. However, this picture is complicated by the style of nuclear fusion, a large fraction (perhaps up to 40%) of the AGB stars becoming planetary nebulae while their inner shells are fusing helium instead of hydrogen, which slows the evolutionary pace. Evolutionary ages tend strongly to exceed the kinematic ages found from expansion velocities.

The chemical enrichments seen in the planetary nebulae are interpreted in terms of internal processes that take place in the predecessor giants before nebular ejection. While on the red giant branch (the first ascent), the star's convective envelope penetrates inward to the hydrogen-burning shell to dredge up the by-products of the carbon cycle, the process at least doubling the surface nitrogen content and altering carbon isotope ratios. A second dredge-up takes place in higher-mass stars beyond about 3–5 solar masses after the cessation of core helium burning, at the beginning of the AGB phase, when convection brings up helium and nitrogen created in what had been the hydrogen-burning shell. Third dredge-up occurs after the start of helium shell flashes, when convection enters the zone between the hydrogen- and helium-burning shells and lofts upward more helium and now carbon. Finally, in the 'hot bottom burning' stage, the carbon cycle operates at the base of the convection zone, causing yet more nitrogen to rise to the surface. The observations are qualitatively in accord with theory, although the nebulae can be more enriched than predictions allow. Only cores above about 0.7 solar masses, which come from higher initial masses, have highly enriched nitrogen, in rough accord with prediction. The carbon-rich nebulae are presumably the descendants of CARBON STARS.

Along with helium, nitrogen, carbon, planetaries must also eject vast amounts of elements created in the helium-burning shells by slow neutron capture (the s-process). Most of the galactic carbon and a great deal of nitrogen, as well as those elements made by the s-process, find their way into the interstellar medium—and into new stars—via the planetaries.

Structure, formation and evolution

Planetary nebulae exhibit a complex, sometimes bewildering, variety of structures that have elicited numerous classification schemes. The most general classification simply divides the nebulae into those with two lobes ('binebulous', now more commonly 'bipolar') and those without, the non-bipolar ranging from almost circular to quite elliptical. The more extreme bipolar nebulae have a tight 'waistband' that encircles the middle between two extended poles that fall along a prominent major axis. The majority of nebulae, bipolar or not, display double or even multiple shells that sometimes include vast extended outer haloes. A significant number are also 'point symmetric', in which each feature on one side of the nebula is reflected through the center almost exactly onto the other side.

The bipolar nebulae have distinctive common features. They tend to reside close to the plane of the Milky Way, suggesting that they are the progeny of higher-mass stars, which apparently eject their envelopes in a manner different from stars of lower mass. The bipolars also tend to have high enrichments of helium and nitrogen, in keeping with AGB dredge-up theory (although there are distinct exceptions). That is, highly enriched type I nebulae are population I and stick to the galactic plane. Moreover, bipolars are richer in molecular gas, the molecules strongly confined to the waistband. Waistband H_2 seems to be excited more by shocks rather than by radiation.

The nebulae are also filled with small 'microstructures' that are not well explained. Some planetaries have rod-like features that extend along the major axis. They can extend past the main nebula to form jug-handle 'ansae' and are often broken into pieces with complex substructures. Moving rapidly, they are called 'FLIERS' (from 'fast low ionization emission regions') and have very low excitations indicative of shock heating. Buried within the closest planetary nebula, the Helix (NGC 7293), are thousands of comet-like knots that point away from the central star, the knots filled with neutral gas, molecules and dust, their surfaces heated by starlight.

These structural features all somehow emerge from giant (AGB) stars. However, the process long remained mysterious because of limited resolution and imaging capability and because the AGB stars surround themselves in dusty cocoons, which rendered the birth process invisible until the advent of infrared capabilities. Understanding requires that we see the transition objects that cover the gap in the $\log L$–$\log T$ plane between the AGB stars and the planetaries. A number of such 'PROTOPLANETARY NEBULAE', nebulae in the developmental state, are now recognized. The Egg Nebula in Cygnus (figure 7), the paradigm, is a bipolar reflection nebula whose bright lobes have the spectrum of a class F supergiant apparently buried within a thick, dark, molecule-filled waistband. Shining through the poles are twin beams that illuminate dozens of rings of bright reflective dust. The rings show that the mass loss from the progenitor AGB star is episodic, with sudden

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

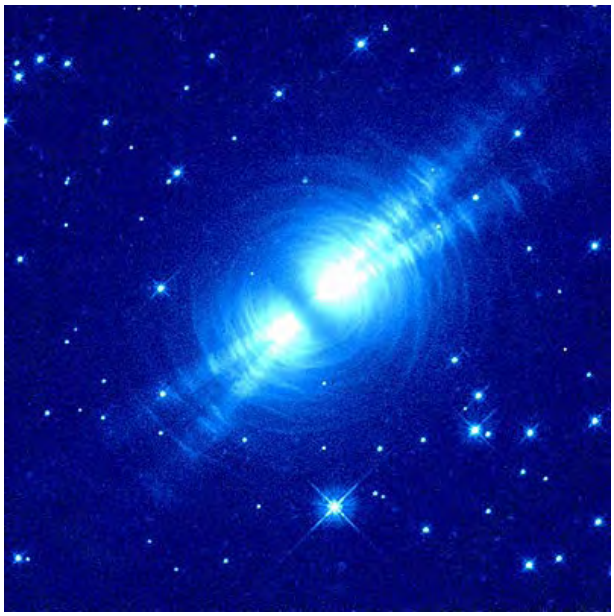


Figure 7. The Egg Nebula is the best-known protoplanetary nebula: a planetary nebula in the making. Outflowing columns of dusty gas, arranged in symmetrical rings, reflect the light of a class F supergiant that was probably once a Mira variable. The star is embedded in the dark thick central disk. H Bond (Space Telescope Science Institute) and the Hubble Heritage Team (AURA, Space Telescope Science Institute, and NASA). R Sahai and J Trauger (Jet Propulsion Laboratory), Space Telescope Science Institute and NASA.

increases every couple of hundred years, for reasons not understood. In other transition objects the star is hotter and the ionization of the interior just beginning. Very young nebulae clearly show oppositely directed flows from the buried stars that appear collimated by thick central disks. As the AGB star loses its envelope, the 'stellar' body that remains heats through the various spectral classes, from class M through F then to A to B as it loses ever more mass and the core is revealed. Finally, it is able to ionize the surrounding thick nebula.

The guiding principle to the formation of the visible nebular structures involves interacting winds (figure 8). As the hot core is revealed, the old, cool, slow wind of the AGB star is transformed into a fast hot wind that runs into the more slowly fleeing dusty gas and shovels it into a ring. When the star becomes hot enough, we then see a 'snowplowed' thick ring nested within a fainter outer structure that has not yet been shovelled up. The different structures are formed by the distribution of the mass that is lost by the AGB star and then magnified by the hot wind's interaction. If AGB mass is lost with great uniformity, the resulting planetary will be round. If the AGB wind is slightly axial symmetric, thicker at the equator than at the poles, the hot wind will blow more vigorously through the poles, and the nebula will take on an elliptical shape. If the slow wind is yet more concentrated into a waistband, the resulting nebula takes on a distinct bipolar structure as the hot wind creates huge bubbles that finally break through to interstellar space.

MIRA VARIABLES (AGB stars that will soon produce planetary nebulae) have been shown not to be round. How such asymmetries factor into shaping planetary nebulae, however, is unknown. Interactions with binary companions could produce latitude-dependent winds that give rise to elliptical and bipolar nebulae. As the giant expands, it approaches the tidal boundary between it and its companion, and may lose mass to it through the common LAGRANGIAN POINT. If the wind is strong enough, the giant can surround its companion in a common envelope that enhances mass loss in the orbital plane. When the envelope has been lost, revealing the white dwarf to be, and the fast wind begins, the result is a highly bipolar nebula, bipolar flows also being created in the interaction. Yet though some central stars are observed as binaries, most that inhabit bipolar planetaries appear to be single. The central stars, however, are so luminous that small companions could be lost. That binaries are indeed important is perhaps revealed by the point-symmetric nebulae, as one explanation is precession caused by orbiting companions.

At the extreme, the AGB wind loss pattern might be produced by interactions with planets still in orbit about the AGB star. The incorporation of planets into the growing giant may act to spin up the star, dust from their destruction adding to the dust formed in the giants' winds. Moreover, it has been suggested, on the basis that planets form from dusty disks and even from the remains of pulsar-ablated stars, that new planets might form within the dusty waistbands of protoplanetary nebulae.

As the star passes through the protoplanetary nebula stage, it finally reaches the critical 25 000 K mark and begins to ionize the inner edge of the snowplowed nebula, shocks helping in the heating of the gas. As the nebula continues to expand and the density decreases, radiation can penetrate ever deeper into the cloud, dissociating the surrounding molecules, and ionizing more of the gas until radiation leaks through to illuminate the unplowed portions of the structure and then the outer haloes, which include matter lost while the star was still in its earlier giant phases. Jets presumably formed by the star penetrate through to act somehow in supplying the FLIERS. As the nebula continues to expand, the surrounding molecular shells disappear. Nevertheless, knots of dense matter in the wind can survive, leaving packets of molecules and dust behind perhaps to make a variety of globules that can survive for nearly the life of the nebula. The nebulae eventually become so large and tenuous that they and their motions through space are affected by the interstellar medium with which they are gradually merging. Continuing to move, the parent stars leave their nebulae behind, the star of one old object actually leaving a wake. All is now over except for the dim, cooling white dwarfs that will orbit the Galaxy forever.

Yet a surprise may await. As the nascent white dwarf cools along its descending evolutionary track, the old

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

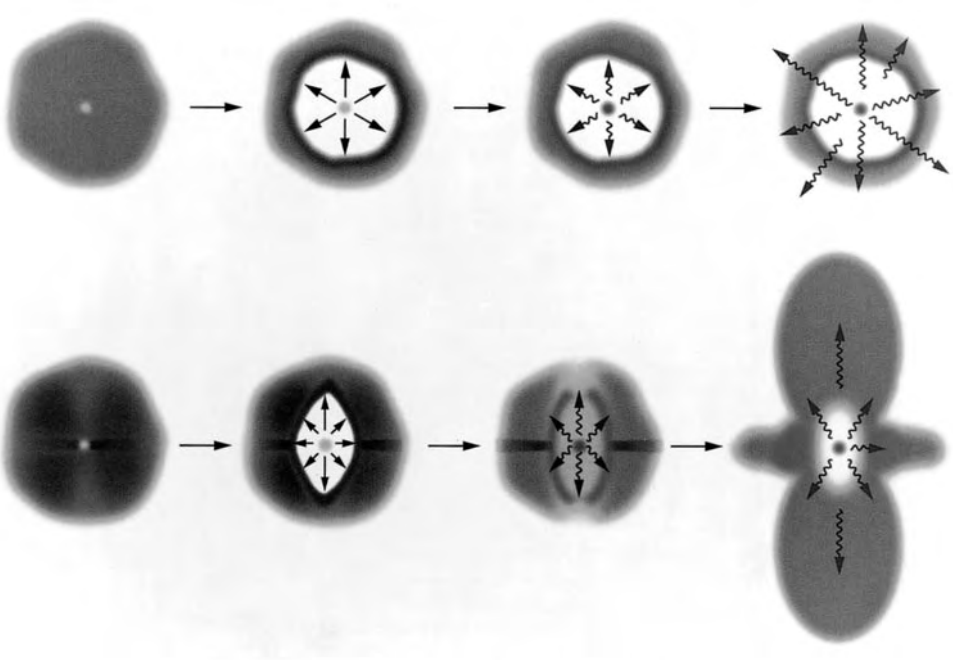


Figure 8. Planetary nebulae evolve with time. Each strip shows (1) a warming star in a protoplanetary nebula surrounded by a cloud of lost mass, (2) the development of a fast wind that shovels the matter into a shell, (3) the heating star radiating in the ultraviolet and illuminating the shell and (4) an advanced state in which the radiation penetrates the shell. The top strip shows the result of spherical mass loss that makes a spherical planetary; the bottom strip shows of mass preferentially lost in an equatorial waistband, the result a bipolar nebula. Art from Kaler J B 1997 *Cosmic Clouds (Scientific American Library)* (New York: Freeman).

helium-burning shell can reactivate itself in a final thermal pulse. The new energy source re-expands the star to giant proportions. New sets of thermonuclear reactions coupled with convection can dramatically change the chemical composition of the surface, enriching it with s-process elements. FG SAGITTAE, which has brightened and cooled over the past several decades, and whose spectrum has become increasingly bizarre, seems to be one of these.

The low surface gravity of such a 'born-again' giant allows more matter to be ejected to produce a planetary within a planetary, the star then approximately retracing its old path on the log L–log T plane. However, the star had already lost most of its hydrogen envelope before it produced its first planetary, so the new ejecta are ultra-rich in helium and carbon, a composition totally different from that of the old nebula. Fine examples are Abell 30 and Abell 78, which have knots of hydrogen-poor matter surrounding their nuclei, their stars now shrunken back to the usual dimensions. Such events may be responsible for a variety of other objects, including the rare R CORONAE BOREALIS STARS, giants with carbon-rich helium atmospheres.

Current research

Numerous problems involving both star and nebula remain. There are still no general methods that can unambiguously find distances to planetary nebulae. The best hope for now is to use nebulae in other galaxies, particularly in M31, where metallicity is no problem. Except in the Magellanic Clouds, however, extragalactic distances are too great to determine much of anything about morphologies. Chemical compositions of galactic nebulae also remain a problem. Abundances of carbon and oxygen found from recombination lines do not fit with those found from forbidden lines, perhaps because of temperature fluctuations.

Wind theory is still in a primitive state, and we cannot easily relate the initial main sequence mass of the star to that of the final product, rendering it difficult to trace the full path of evolution. Moreover, kinematical ages do not fit stellar evolutionary ages, the predicted sizes of the nebulae for a given point on the HR diagram being too large. Either evolutionary theory is wrong or we are not properly taking into account the accelerations and decelerations of the nebular gas.

The cause and the time of origins of the bipolar structures that so many of the nebulae assume are not known; neither are those of the bipolar flows. We are only beginning to penetrate into the mystery of the protoplanetaries. The degree to which binaries (including planets) are important is unknown. We do know, however, that Herschel's discovery eventually led to a fuller understanding of STELLAR EVOLUTION, of how stars meet their ends and of the processes of mass and chemical recycling in the Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Aller L H 1984 *Physics of Thermal Gaseous Nebulae* (Dordrecht: Reidel)
Frank A 1997 Blowing cosmic bubbles *Sky Telesc.* (February) p 36
Gurzadyan G A 1997 *The Physics and Dynamics of Planetary Nebulae* (New York: Springer)
Habing H J and Lamers H J G L M (ed) 1997 *Planetary Nebulae: International Astronomical Union Symp. 180* (Dordrecht: Kluwer)
Iben I 1995 Planetary nebulae and their central stars—origin and evolution *Phys. Rep.* **250** 1
Kaler J B 1985 Planetary nebulae and their central stars *Ann. Rev. Astron. Astrophys.* **23** 89
Kwok S 1996 A modern view of planetary nebulae *Sky Telesc.* (July) 38
——1998 Metamorphosis *Sky Telesc.* (October) 30
Peimbert M 1990 Planetary nebulae *Rep. Prog. Phys.* **53** 1559
Pottasch S R 1984 *Planetary Nebulae* (Dordrecht: Reidel)

James B Kaler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Rings

Planetary rings are common to all the giant planets of our solar system. They form a disk of particles following a nearly circular orbit in the equatorial plane of the central planet. Ring disks have various dimensions and radial structures. They have been discovered from both ground-based and spacecraft observations. SATURN'S RINGS were discovered by GALILEO GALILEI in 1610 but clearly identified as rings by CHRISTIAAN HUYGENS in 1659. The development of ring science has been tightly linked with the history of their observations during three centuries. The discovery of a new ring system around Uranus in 1977, during the ground-based observation of a stellar OCCULTATION by the planet (see URANUS: RINGS), boosted research in this area and a quest for new rings started. Suspected after the charge particle measurements of PIONEER MISSION 11 in 1974, the very faint JUPITER'S RINGS were finally detected by the VOYAGER spacecraft in 1979. The story of NEPTUNE'S RINGS started when ring 'arcs' were detected in 1984 with the same method of stellar occultation. The complete ring system was revealed by Voyager 2 cameras in 1989. The view of ring systems has been completely renewed in this 20 yr long 'golden age' period. Many theories have been developed to explain the huge amount of unexpected observations sent back by Voyager probes. Great progress has been made in our understanding of planetary rings but many questions remain on their origin and evolution. In the following sections the possible scenarios on their origin are discussed and the main questions raised after these discoveries are developed. Finally open questions and the main science objectives of the next CASSINI/HUYGENS MISSION to Saturn's rings are presented.

Origin

Rings are a common feature in the close vicinity of giant planets. They are located within the ROCHE LIMIT, an approximate distance below which any rock is expected to break up under the tidal effects exerted by the planet. Ring disks in the Roche limit are vertically very thin, a few hundred meters thick for a few ten or hundred thousand kilometers wide. This is the natural evolution of a cloud of particles which is losing energy in inelastic interparticle collisions while conserving angular momentum in the gravitational field of a central planet. An out-of-plane particle, when crossing the mean plane of the other particles, has its out-of-plane velocity component reduced in each collision and finally reaches the common plane. Meanwhile particle orbits are precessing in the equatorial plane if the planet is oblate.

Ring particles may have condensed from circumplanetary material when the planet formed and have collapsed into a thin disk. The creation of rings after the disruption of a parent body is another plausible scenario. The progenitor, a satellite or a captured comet or asteroid, would have drifted into the Roche limit and been disrupted by tidal effects or by a large impact. Observations tend to show that the lifetime of particles and the timescales of evolution of planetary rings are much shorter than the age of the solar system. Then primordial rings should have disappeared already or should be continuously replenished and destroyed over ages. The equivalent mass of a 200 km sized satellite should be reinjected into the Saturn's rings every 500 Myr. The less massive main rings of Neptune, Uranus and Jupiter might have found enough material in their nearby small satellites to be fed with. An alternative scenario of a recent and widespread ring formation would imply a hardly reconcilable anthropocentric viewpoint.

The origin of diffuse rings may be less uncertain. Their particles are typically micron sized and should be consequently removed very rapidly by drag forces. They are most probably replenished with the ejecta produced by the meteoroid bombardment of the surfaces of close satellites. The E ring of Saturn, which extends from 3 to 8 Saturn radii, may be the most illustrative example. Its peak density is correlated with the orbit of the icy satellite Enceladus. The ejected dust would be dispersed around its orbit and further dispersed at large scale under the action of electromagnetic forces. Recent Galileo observations of Jupiter's rings have shown direct evidence of very faint dusty rings associated with small satellites Amalthea and Thebe outside of the main ring system (see GALILEO MISSION TO JUPITER). Spectroscopic measurements tend to confirm the compositional link between Jupiter's rings and nearby satellites (see JUPITER: SATELLITES).

The chemical composition and bulk structure (density, porosity, ...) of ring particles may provide hints of possible scenarios of formation but relevant spectroscopic data are often lacking. Water ice has been identified as the main constituent on the surface of Saturn's ring particles. Impurities are also present but not yet identified. The compositions of the other ring systems are still unknown. Water ice has not been detected. Reaching conclusions about the origins of rings from their current composition remains, however, a tricky task as ring particle surfaces are eroded or polluted on relatively short timescales by meteoroid bombardment, high-energy particles sputtering or ejecta of nearby satellites. Clues to origin may be deeper under the surface of particles and therefore still hard to probe with current data.

Evolution

Interparticle collisions in the ring plane disperse particles radially even more. The final disk should be uniform with smooth edges. The most remarkable discovery in the past 30 yr has been the unexpected and extreme diversity of structures in planetary rings, such as divisions, sharp edges, spiral waves, narrow ringlets or azimuthal asymmetries such as arcs and clumps in narrow rings or spokes in Saturn's rings. Most of them are common to all ring systems. Any narrow ring should be dispersed radially because of collisions and differential motion (the orbital velocity decreases with increasing distance to the planet). Any longitudinal asymmetry should also be slowly stretched along the ring and erased on timescales

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a few years by this differential motion. The observed structures are therefore either very young or radially and azimuthally confined.

Among the processes proposed to explain the observed structures, the resonant gravitational interaction of ring particles with nearby satellites is certainly one of the most important. Theories of the radial confinement of a ring, or a 'shepherding' mechanism, proposed by Goldreich and Tremaine in the 1970s, estimate the torque exchange between colliding ring particles and a satellite. A transfer of angular momentum takes place at resonant locations where the orbital frequencies of the satellite and ring particle are commensurable, with mean motions in a ratio $m:m+1$ for example. The resonant perturbation of an outer satellite removes angular momentum from ring particles which drift inwards while an inner satellite adds angular momentum to outer ring particles which drift outwards. Both interacting bodies tend to move away from the resonant orbit. This interaction results in various structures depending on satellite mass and distance. A close or massive outer satellite creates a wake in the ring and repels particles along a sharp edge. The edge of Jupiter's main ring is maintained by the close 10 km sized satellite Adrastea. The outer edge of the B ring is located at the 2:1 resonance with 200 km sized Mimas and corresponds to the Huygens gap in the Cassini division. The massive Mimas is then able to remotely open a gap. The outer edge of Saturn's A ring is located at a 7:6 resonance with the coorbital 100 km sized satellite Janus (see SATURN: SATELLITES). Gaps with sharp edges can also be opened in rings by embedded moonlets. The Encke gap is carved out of Saturn's A ring by the 10 km sized moonlet Pan. The satellite was suspected because of wakes detected on gap edges and later discovered in Voyager images. Density waves or bending waves are driven by smaller and distant satellites at high-order (m) resonance locations. They provide useful information on the collective behavior of particles (or 'viscosity') in the densest parts of the rings. Waves a few kilometers wide driven by Prometheus, Pandora, Janus or Mimas are visible in Saturn's rings. Because of these interactions, the orbits of particles drift a few cm to 1 m per year. The outer A ring is expected to fall onto the Cassini division in few hundred million years.

Two shepherding moons on each side of a ring repel particles the same way and can confine the ring radially. The discovery of Saturn's F ring shepherds Prometheus and Pandora and of Uranus's ε ring shepherds Cordelia and Ophelia were the most spectacular success of the Goldreich and Tremaine theory. The singularity of URANUS'S RINGS at the time of their discovery was their extreme narrowness of a few kilometers. Dermott and Gold had proposed in 1977 another confinement theory involving a single small satellite. In a frame corotating with the moon, a ring particle of nearly the same semimajor axis describes a 'horse shoe' trajectory forming a narrow ring. However, this kind of orbit is not stable against collisions. No shepherding moons were discovered around or within the other eight narrow rings of Uranus. However, resonant relationships were found between the sharp edges of Uranus's δ, λ and γ rings and the satellites Cordelia and Ophelia. The radial confinement of the 6, 5, 4, α, β and η rings remains unexplained.

The discovery of Neptune's arcs set the new puzzling question of azimuthal confinement. From 1984 to 1989, a campaign of stellar occultation observations led to the conclusion that Neptune's 'rings' were in fact a family of longitudinally discontinuous rings, called 'arcs'. Lissauer in 1985 formulated a model where an arc could be confined at the Lagrangian L_4 or L_5 points of a 200 km sized satellite. An inner satellite was required to prevent the radial spreading due to collisions. Goldreich, Tremaine and Borderies in 1986 proposed that a single nearby inclined satellite would confine azimuthally the whole family of arcs in very specific sites. Voyager 2 images in 1989 showed that arcs were in fact the densest parts of one narrow ring, the Adams ring. No coorbital massive satellite was found but clumps, maybe embedded km-sized objects, were discovered in the arcs themselves. After ground-based observations and Voyager 2 data were compared, the arcs appeared to be stable over 5 yr despite differential motion. Porco detected in Voyager 2 images a wavy radial distortion in the Adams ring that she attributed to the excitation of the ring by a 42:43 resonance with the nearby satellite Galatea. Arcs would be confined in 4.2° long azimuthal sectors along the ring according to the theory of Goldreich and colleagues mentioned above. However, some arcs were not completely confined to these sites. Thanks to the recent first direct imaging of Neptune's arcs from Earth with the HUBBLE SPACE TELESCOPE, the mean motion of the arcs has been measured more accurately. The orbit is unambiguously outside the 42:43 resonance with Galatea. The role of this satellite and the azimuthal confinement of arcs therefore have to be studied again. Other arcs and clumps have been detected in the F ring, in the Encke gap ringlets around Saturn and in Uranus's λ ring. Their temporal evolution is still poorly known but recent ground-based and HST images of the F ring arcs obtained during Sun and Earth ring plane crossings in 1995 suggest that the structure of the F ring is evolving on a timescale of weeks. This result appears to be compatible with former Voyager observations. Observing heterogeneous rings is certainly a key towards the understanding of the close relationship between rings and satellites, the radial and azimuthal confinements of ring material and then the origin and evolution of rings.

Other processes at play in planetary rings can also severely reduce their lifetime. Erosion mechanisms due to interparticle collisions, meteoroid bombardment, sputtering by cosmic rays, etc supply new molecules to a gaseous ring atmosphere or new small particles in the ring system while removing others. The meteoroid bombardment appears to be the most efficient at eroding particle surfaces at a rate of about 10^{-3} cm yr^{-1}. These outstanding rates even for cm- to m-sized ring particles

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

still depend on poor data. The size distribution and number of particles are also changed by interparticle collisions but the surface erosion, fragmentation or accretion rates of particles are very uncertain as the restitution factor and bulk properties of particles are unknown and difficult to constrain. Ring particles immersed in the surrounding magnetospheric plasma are charged. Within the planetary magnetic field, sub-micron-sized grains can experience significant Lorentz force compared with gravity and be driven out of the ring plane. This process is certainly involved in the creation of Saturn's B ring spokes and of Jupiter's ring halo and in the spatial distribution of Saturn's E ring. Small particles are also sensitive to drag forces such as Poynting–Robertson or plasma drags. The POYNTING–ROBERTSON EFFECT due to solar radiation induces an orbital decay and reduces the lifetime of dusty particles in the system to less than 1 Myr. The plasma drag is due to collisions with electrons and ions or long-range charge interactions. Charged dust particles inside or outside the synchronous orbit (where particles and magnetic field rotation speeds are equal) will spiral inwards or outwards respectively. The related lifetime may be as short as few hundred years in Jupiter's and Saturn's systems.

Future

Planetary rings are complex systems. They apparently evolve on timescales much shorter than the age of the solar system but are also a common component around a planet. The different processes at play transfer angular momentum between rings and satellites, confine ring material at some places or push other material onto the planet, provide or remove particles. Rings are therefore marvellous laboratories for 'live' tests of theories of dynamical processes at work in disks, the same processes which have been certainly at work, at a larger scale, sometimes during the formation of planetary systems. Many important ring structures are still not understood. Timescales of evolution and lifetimes of particles have to be confirmed. The Cassini spacecraft will be placed into orbit around Saturn in 2004. Observations in a uniquely large spectral range will give decisive constraints on Saturn's ring particles properties and on the compositional relationship between ring and satellites. The temporal evolution of structures will be carefully followed during 4 yr to better understand confinement mechanisms or collective behavior of particles in dense disks. This mission will then allow major advances in answering the questions of the origin and evolution of Saturn's rings and of planetary rings in general. Meanwhile technical progress will provide always-richer ground-based data on the other ring systems to constrain their physical properties and any short-evolution timescales, until a new generation of spacecraft is sent to Uranus and Neptune.

Bibliography

Greenberg R and Brahic A (ed) 1984 *Planetary Rings* (Tucson, AZ: University of Arizona Press)

Esposito L, Brahic A, Burns J A and Marouf E A 1991 Particle properties and processes in Uranus' rings *Uranus* ed J T Bergstrahl, E D Miner and M S Matthews (Tucson, AZ: The University of Arizona Press) pp 410–65

Esposito L, Cuzzi J N, Holdberg J B, Marouf E A, Tyler G L and Porco C C 1984 Saturn's rings: structure, dynamics and particle properties *Saturn* ed T Gehrels and M S Matthews (Tucson, AZ: The University of Arizona Press) pp 463–545

French R, Nicholson P D, Porco C C and Marouf E A 1991 Dynamics and structure of the Uranian rings *Uranus* ed J T Bergstrahl, E D Miner and M S Matthews (Tucson, AZ: The University of Arizona Press) pp 327–409

Porco C C, Nicholson P D, Lissauer J J, Cuzzi J N and Esposito L W 1995 Neptune's rings *Neptune and Triton* ed D Cruikshank and M S Matthews (Tucson, AZ: The University of Arizona Press) pp 703–806

Cécile Ferrari

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Society

Carl Sagan, Bruce Murray and Louis Friedman founded the non-profit Planetary Society in 1979 to advance the exploration of the solar system and to continue the search for extraterrestrial life. The Society has its headquarters in Pasadena, California, but is international in scope, with 100 000 members worldwide, making it the largest space interest group in the world. The Society funds a variety of projects and programs, including Red Rover Goes to Mars (for students), the Mars Microphone on the Mars Polar Lander mission, SETI@home, the Gene Shoemaker Near Earth Object Grants, etc.

For further information see http://planetary.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Surfaces

For most of the planets and satellites, the knowledge we have of their global properties comes from the observation of their surfaces. Indeed, the information used to characterize solar system objects derives essentially from the analysis of the light we receive from them. Only rarely does one have access to additional sources of data, such as *in situ* seismic or magnetic probing, or soil chemical analysis. When the objects are surrounded by a thick—and therefore opaque—atmosphere, as for VENUS or JUPITER, we cannot see the surface. In all other cases, including the inner and minor planets, as well as planetary satellites, the surface characteristics are used to decipher clues about the past and present properties of the objects as a whole.

Surface features do in general result from the complex interaction between internal processes, which initiated tectonic and volcanic activity, atmospheric phenomena and exogenic processes, due to the contact with the interplanetary medium, such as irradiation with solar particles and bombardment by meteorites. These processes control the evolution of planetary surfaces over a wide variety of time constants. Therefore, by studying their surfaces, one has the potential to trace back the history of planetary bodies, from seasonal variations to geological time scales.

Heavily cratered terrains

Some of the most widely observed surface features are impact CRATERS, as on the MOON and MERCURY. These craters are remnants of the very intense bombardment that affected all solar system bodies after their accretion. Actually, in the inner solar system, a huge number of bodies were formed, on trajectories leading to violent collisions. It took hundreds of millions years to clear up the solar cavity, largely by mutual destruction of the impacting bodies. These impacts constituted the dominant process in the early solar system. If pictures of the inner planets had been taken four billion years ago, they would essentially all look the same.

On the few remaining objects, records of this era remained in the form of impact craters of all sizes, covering their entire surface. Bodies that have not evolved much since still exhibit their impacted surface. ASTEROIDS, planetary satellites and planet areas are labeled 'old' when their surfaces appear heavily cratered, which only means that they have not been further remodeled by global activity.

As a general rule, the smaller the objects are, the less they have evolved after this early bombardment. Thus, by studying the Moon and Mercury, we have access to the first billion years of solar system evolution. MARS has a greater mass and has transformed about half its surface by more recent activity, while preserving heavily cratered highlands on its southern hemisphere. On the largest of the inner planets, the Earth, geology has erased the record of that early history. Attempts to identify impact craters from that era on the terrestrial crust continue.

Tectonic and volcanic features

Surface TECTONIC features have been produced in solid bodies with internal energies large enough to have sustained a mantle activity through the crust. For most planets and satellites, for which the energy source is dominated by the radioactivity of long-lived (U, Th and K) elements, the rise in temperature, and thus the level of internal activity, increases with the size. On the Moon, the only surface features associated with this internal activity come from the filling with lava flows of the largest—and thus the deeper—impact-produced basins, giving birth to the so-called 'MARE'. Predominantly made of Mg/Fe silicates (pyroxenes and olivines), they appear darker than the surrounding 'continents', rich in bright Al-rich feldspar (anorthosite). On larger bodies, internal activity induced the lifting of large areas, like THARSIS MONTES, on Mars, or ISHTAR TERRA on Venus, with MAXWELL MONTES dominating the surrounding plains from an altitude of 11 km. The rise of these features is sometimes associated with great faults or canyons, such as VALLES MARINERIS on Mars, 5000 km long, 6 km deep on average and up to tens of km wide.

Magmatic activity, restricted to mare production on the Moon, produced five impressive volcanoes on Mars, the largest in the solar system, with OLYMPUS MONS the highest, 27 km in altitude. Each of these volcanoes results from a process similar to the one responsible for volcanic chains on the Earth, the difference being the lack of plate tectonics on Mars: a 'hot spot' under the lithosphere generates distinct 'bubbles' of magma which, on Mars, contribute to the same volcano while, on the Earth, the moving lithosphere is 'pierced' through aligned blisters, building distinct volcanoes, as one can see for example with the Hawaiian islands.

In a very few cases, tidal effects initiate volcanic activity at a very high rate. The most striking example is Io, closest of Jupiter's four Galilean satellites, whose entire surface is remodeled by volcanic activity on time scales of only thousands of years (see IO: VOLCANISM AND GEOPHYSICS). The differences in surface properties of these four objects reflect the high degree of diversity of evolutionary patterns in the solar system. Formed almost at the same time from the same material, with about equal masses, they exhibit today totally different surfaces: the more distant one, CALLISTO, is very heavily cratered, and thus described as the 'oldest' one. On GANYMEDE, bright 'young' terrains contrast with darker cratered units. EUROPA is covered with a fractured but smooth and uncratered ice sheet, possibly floating on an underground water ocean; Io is the only planetary object, apart from the Earth, where active volcanoes have been observed (see also VOLCANISM IN THE SOLAR SYSTEM).

Impact craters and volcanoes constitute surface structures with elevations that can reach tens of kilometers. However, mountains similar to those on Earth are extremely rare. They result from continental drifts directly linked to plate tectonics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Features induced by interaction with the atmosphere

Surface structures can also be driven by interactions with PLANETARY ATMOSPHERES. Polar caps and frosts are currently observed on Mars; liquid, icy and eolian erosion features are also spread over large areas of its surface. Some may have been formed billions of years ago, such as valley networks, while dunes are formed by dust transportation processes occurring on a seasonal basis, at present, when CO_2 condensation at the poles induces drastic changes in atmospheric pressure, giving rise to giant dust storms (see MARS: SURFACE).

Past and present: the surface records

Altogether, the variety of information one can derive from an in-depth study of surface features, in the framework of comparative PLANETOLOGY, has the potential to trace back the entire history of the solar system. This requires an extended coverage of the surface mapping, together with a high spatial resolution, a wide spectral range and a large spectral resolution. Indeed, one needs not only to distinguish extended geological areas, but also to identify structures at a much smaller scale, such as outflows, crater rims and stratified canyon slopes. Moreover, in order to determine the composition of the observed terrains, the observation in the visible has to be extended towards the infrared, where most materials have specific transition features enabling unambiguous characterization. When the atmosphere is opaque in the visible, as for Venus and, to a lesser extent, TITAN, one needs to image the surface at longer wavelengths. A new generation of instruments, imaging spectrometers operating in a variety of wavelength ranges, will provide a major source of information for the future space exploration of the solar system.

Jean-Pierre Bibring

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetary Theory from Eudoxus to Copernicus

By the term planetary theory, in its widest sense, is meant a method of representing, either qualitatively or quantitatively, the apparent geocentric motions of the Sun, Moon and planets. Prior to the development of modern celestial mechanics by Newton, planetary theory could be purely arithmetic, geometrical with or without computation, or physical, the last being an extension of geometrical planetary theory to mechanisms presumed to exist in the heavens and to produce the apparent motions of the Sun, Moon and planets. The principal arithmetic planetary theory is that of the Babylonians, considered in the article on BABYLONIAN ASTRONOMY. Here we shall consider primarily geometrical and secondarily physical planetary theory in Greek, Arabic and European astronomy up to the time of Copernicus.

The earliest geometrical and physical planetary theory is that of EUDOXUS OF CNIDUS (*c*. 350 BC), known only from rather imprecise descriptions by ARISTOTLE (384–322 BC) and Simplicius (sixth century AD), which has been reconstructed by SCHIAPARELLI and more elaborately by Yavetz. Eudoxus's theory, and its slightly later modification by CALLIPPUS (*c*. 330 BC), is composed exclusively of uniformly rotating spheres concentric to the Earth, perhaps on the analogy of the sphere of the fixed stars, which appears to rotate daily about a central Earth. The Sun and Moon each receive three spheres, the outermost, with poles at the celestial poles and equator in the plane of the celestial equator, produces the diurnal rotation from east to west, the second, with poles at the poles of the ecliptic and equator in the plane of the ecliptic, slowly shifts the inclined plane of the path of each body—it was assumed that the Sun had some small motion in latitude—from west to east for the Sun, from east to west for the Moon, and the third, with poles inclined slightly to the poles of the second, carries the Sun in its annual motion and the Moon in its monthly motion through the zodiac from west to east. The planets are more interesting. The outer sphere produces the daily rotation just as for the Sun and Moon, the second sphere, with poles at the poles of the ecliptic and equator in the plane of the ecliptic, produces the motion through the zodiac from west to east in the zodiacal period specific to each planet; then a third sphere with poles in the plane of the ecliptic and a fourth with poles slightly inclined to the third complete rotations in opposite directions, in principle in the planet's synodic period, the period between conjunctions with the Sun.

In Schiaparelli's reconstruction, the planet is at the equator of the fourth sphere and the combined motion of the third and fourth spheres produces a figure called by Simplicius a *hippopede* (horse-fetter), an elongated figure '8' lying on its side and bisected by the ecliptic, with the inclination of the two spheres determining both its length along the ecliptic and its width above and below the ecliptic. When carried about the ecliptic by the second sphere, this produces periodic and invariable retrograde arcs and deviations from the ecliptic in latitude for Saturn, Jupiter and Mercury, but fails badly for Mars and Venus, producing either no retrogradation, because their synodic periods are longer than their zodiacal periods, or enormous deviations in latitude if the inclinations are made large enough to produce a retrogradation. In Yavetz's reconstruction, the poles of the fourth sphere are not only inclined to those of the third, but are offset from the plane of the ecliptic, in which the poles of the third are located, and the planet itself is offset from the equator of the fourth sphere. This gives three degrees of freedom and allows a far better approximation to the apparent path of a planet. It is not incompatible with Simplicius's description, and can even produce a figure like the *hippopede*, but whether Eudoxus himself actually did something so complex, and so ingenious, is uncertain. The models were modified by Callippus by the addition of two spheres each to the Sun and Moon, to account for inequalities in their motions, and of one sphere each to Mars, Venus and Mercury, the reasons for which are not known. Both Schiaparelli and Yavetz have investigated such models, and Yavetz has shown that retrogradations of different shape can be produced, remarkably close to those of Mars, although they still remain invariable throughout the zodiac, contrary to the behavior of the planets.

Whatever the models of Eudoxus and Callippus, which were intended as physical mechanisms in the heavens, may have achieved as geometrical representations of the apparent motion of a planet, they could not be used for calculation of any kind; in this sense they were purely qualitative. There was another geometrical method of representing the motions of planets, by means of concentric circles, or spheres, surrounding the Earth and carrying epicycles, or of eccentric circles, or spheres, about the Earth with movable centers, which are equivalent. Since, as shown in figures 7 and 8 below, these are also geometrically equivalent to the heliocentric theory for, respectively, superior and inferior planets, proposed by ARISTARCHUS OF SAMOS (early third century BC), they must have been known by that time, as it is unthinkable that Aristarchus could have worked out the heliocentric theory without an understanding of these transformations and equivalencies; in the most interesting equivalence of such models, APOLLONIUS OF PERGA (late third century BC) showed how to find the stationary points between which the motion of the planet is retrograde. By the second century BC, there were computational planetary theories in Greek astronomy, but these were arithmetic rather than geometrical and based upon Babylonian methods, probably with no alteration except for adaptation to the Egyptian calendar and, late in the first century, to the Alexandrian calendar. HIPPARCHUS (mid second century BC) developed epicyclic and eccentric models for the Sun and Moon, and even derived parameters for them from observation. Nevertheless, it appears that for computation he too used Babylonian methods rather than geometrical models. Nearly 300 years elapse between Hipparchus and PTOLEMY (mid second century AD), and although it is probable that computation from geometrical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

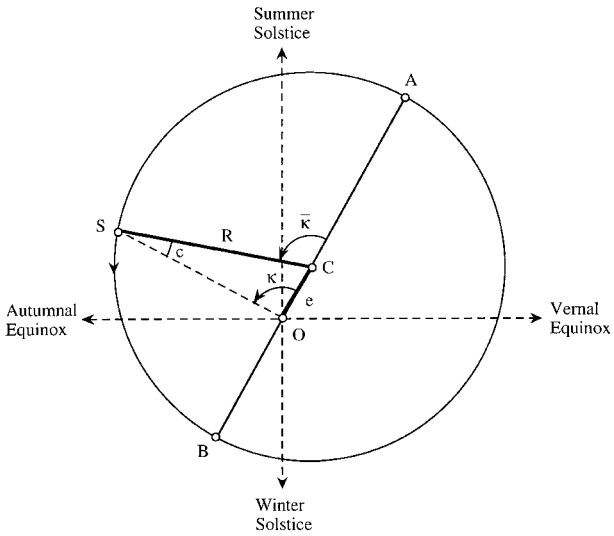


Figure 1.

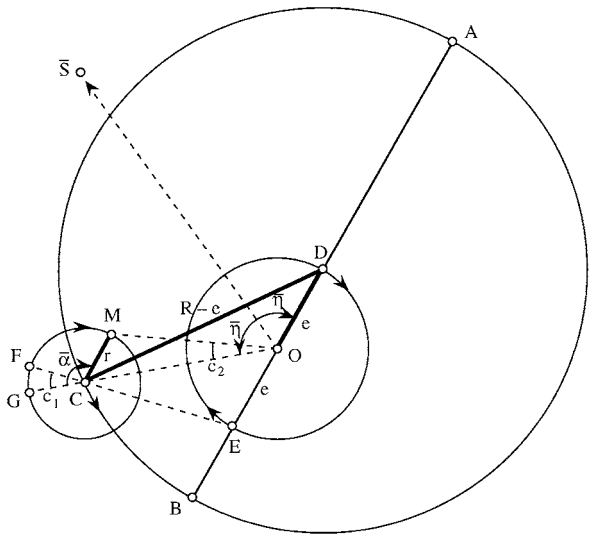


Figure 2.

models was developed during this period, as shown by survivals in Indian planetary theory, the earliest examples in Greek are in the works of Ptolemy, in particular, the *Almagest* and the *Handy Tables*. These remain central to mathematical astronomy for the next 1400 years, as does Ptolemy's *Planetary Hypotheses* to physical astronomy.

In what follows, we shall briefly go through the elements of Ptolemy's theories of the Sun, Moon and planets. Ptolemy's method may be characterized simply: it is rigorously empirical and rigorously mathematical. Every model is either derived or confirmed by observation, although not all the observations are presented in the text, and every numerical parameter is derived directly and uniquely from observation by strict mathematical procedures. There is, however, a large range of precision in his observations, from positions and times measured to within a few minutes for the derivation of parameters—although their accuracy is more variable and there are systematic errors—to rough, even qualitative, observations for demonstrating the properties of models.

In the case of the Sun there is a single inequality most evident in observations showing that the intervals of time between the equinoxes and solstices, although separated exactly by quadrants, are unequal, meaning that the apparent motion of the Sun is non-uniform. Ptolemy accounts for this inequality by letting the motion of the Sun take place on a circle eccentric to the Earth, a model earlier used by Hipparchus, and an equivalent epicyclic model may also be used. In figure 1, the equinoxes and solstices seen from the Earth at O are separated by quadrants, but the arcs of the Sun's motion on an eccentric circle about C are greater or less than quadrants. The motion of the Sun S may thus be considered as a uniform motion $\bar{\kappa}$ about C and a variable correction c, resulting in a non-uniform apparent motion $\kappa = \bar{\kappa} \pm c$ about O. The mean anomaly $\bar{\kappa}$, the eccentricity e as a fraction of the radius R of the eccentric, and the direction of the apogee A and perigee B may be derived from the very observations of intervals of time between the equinoxes and solstices that showed the Sun's inequality. Although later astronomers found improved values of these parameters, the solar model itself received no improvement prior to KEPLER.

The theory of the Moon is considerably more complex as Ptolemy found that it displayed two inequalities, the first with a period of the anomalistic month, in which the Moon returns to the same point in its (movable) orbit, already known to Hipparchus, and the second, which modifies the first, with a period of half a mean synodic month, in which the Moon returns to conjunction with or opposition to the 'mean Sun', that is, the Sun with only its mean motion ($\bar{\kappa}$ in figure 1). He accounted for the first inequality by placing the Moon on an epicycle with the period of the anomalistic month and for the second by placing the epicycle on an eccentric such that it completes a full revolution on the eccentric in half a synodic month. A related inequality, also with a period of half a mean synodic month, affects the mean anomaly, the motion of the moon on the epicycle. In figure 2, the Earth is at O, the mean Sun is at $\bar{S}$, from which the center of the epicycle C moves through the mean elongation $\bar{\eta}$, and the center of the eccentric D moves in the opposite direction through the same $\bar{\eta}$. The result is that C completes one revolution with respect to $\bar{S}$ in a mean synodic month, but two revolutions with respect to the apogee A and perigee B of the eccentric, the effect being to draw the epicycle closer to the Earth at quadrature than at conjunction and opposition, increasing the inequality due to the motion of the Moon on the epicycle. The Moon moves on the epicycle through the mean anomaly $\bar{\alpha}$, completing one revolution in an anomalistic month, measured from a direction FCE that has an 'inclination' to a point E that lies on the apsidal line AB exactly opposite D such that the distances OD

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and OE are equal, the effect being to modify $\bar{\alpha}$ to the true anomaly $\bar{\alpha} \pm c_1$ measured from the direction GCO to the Earth. The true elongation of the Moon from the mean Sun is then $\bar{\eta} \pm c_2$, where c_2 is a function of $\bar{\alpha} \pm c_1$ and the variable distance OC. The parameters of the model are found from observation: the radius of the epicycle r, the mean anomaly $\bar{\alpha}$ and mean elongation $\bar{\eta}$ from lunar eclipses, the eccentricity e from observations of the Moon at quadrature and octants from the Sun. A peculiarity, really a defect, of Ptolemy's lunar model is that it produces a variation of the distance of the Moon from the Earth so large that it may be about half as distant at quadrature as at opposition and conjunction, which is directly contradicted by the small change in the Moon's apparent diameter and parallax.

The most impressive achievement in Ptolemy's astronomy is the theory of the planets, which provided the foundation of all later planetary theory, including that of Copernicus, and was not improved upon until Kepler. Ptolemy's criterion for ordering the planets in the *Almagest* is zodiacal period: in descending order from the fixed stars, Saturn, Jupiter, Mars, Sun, Venus, Mercury, Moon. The Sun, Venus and Mercury have the same period, one year, and Ptolemy places the Sun between the superior planets, Saturn, Jupiter and Mars, which reach opposition to the Sun, and the inferior planets, Venus and Mercury, which reach only a limited elongation. The Moon of course reaches opposition, but is placed lowest since it has the shortest period, may occult all the higher bodies, and has a large, measurable parallax showing that it is close to the Earth. In the *Planetary Hypotheses*, Ptolemy sets out a method of computing the distances of the planets which confirms that the Moon, Mercury and Venus are below the Sun, since there is adequate space only for them, and the remaining planets above.

The planets have two inequalities, the first a function of longitude, of location in the zodiac, the second a function of elongation from the Sun. It is the second inequality that produces the most obvious feature of planetary motion, a periodic retrogradation bounded by two stations, which occurs near opposition for the superior planets and near inferior conjunction for the inferior planets. The second inequality is produced by the motion of the planet on an epicycle and the first inequality by the motion of the center of the epicycle. For the superior planets, the motion of the center of the epicycle corresponds to the planet's own motion about the Sun in a heliocentric model, and the motion of the planet on the epicycle to the Earth's motion about the Sun; for the inferior planets the motion of the center of the epicycle corresponds to the Earth's motion about the Sun and the motion of the planet on the epicycle to the planet's own motion about the Sun. Ptolemy's model for both inequalities is shown in figure 3. The earth is at O and at an eccentricity e let M be the center of an eccentric circle of radius R with apogee A and perigee B. Symmetrically to M at the same eccentricity e let E be the center about which the epicycle moves uniformly,

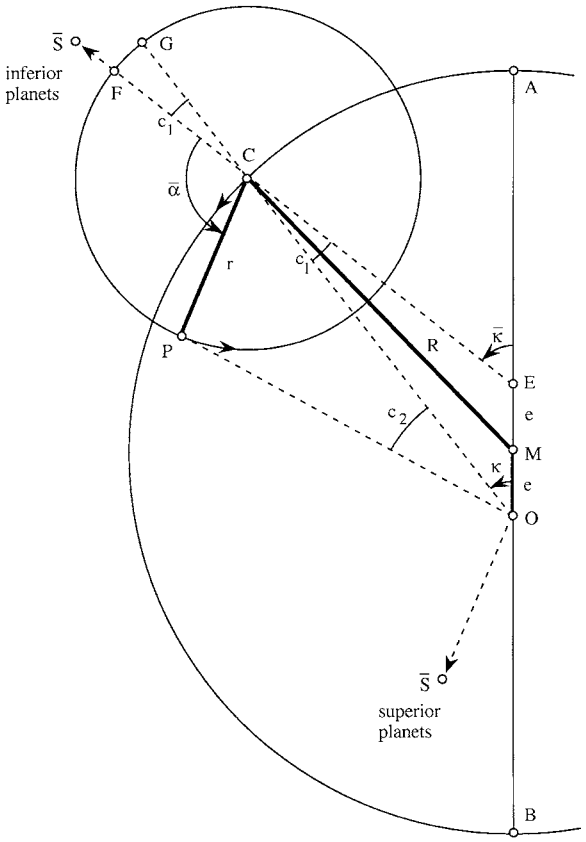


Figure 3.

often called the 'equant' point, and let the center of the epicycle C move on the circle about M uniformly about E through mean eccentric anomaly $\bar{\kappa}$. The effect of the separation of the center of constant distance M from the center of uniform motion E closely approximates motion according to Kepler's ellipse and area laws since the circle about M corresponds to the circle constructed on the major axis of the ellipse, to which for small eccentricity the ellipse very nearly coincides, and E corresponds to the empty focus of the ellipse, about which the angular motion of the planet is very nearly uniform. In principle, the models of Ptolemy and Kepler differ by only a few minutes of arc, although inaccuracies in Ptolemy's parameters and other approximations introduce greater errors.

The planet P moves on the epicycle of radius r though the mean anomaly $\bar{\alpha}$ measured from the direction EC extended to the mean apogee of the epicycle F such that for a superior planet CP is always parallel to the direction $O\bar{S}$ from the Earth to the mean Sun. For an inferior planet EC always lies in the direction of the mean Sun $\bar{S}$. The model produces variable corrections to two inequalities. The first, c_1, called the 'equation of center', corrects $\bar{\kappa}$ to the true eccentric anomaly $\kappa = \bar{\kappa} \pm c_1$ seen from O and $\bar{\alpha}$ to the true anomaly $\alpha = \bar{\alpha} \pm c_1$ measured from the direction OC extended to the true apogee of the epicycle G. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

equation of the anomaly c_2, subtended by the radius of the epicycle r, then corrects the direction of the center of the epicycle OC to the direction of the planet OP. Hence, where the longitude of the apogee A is λ_A, the true longitude of the planet is given by $\lambda_A + \bar{\kappa} \pm c_1 \pm c_2$. The parameters of the model are derived rigorously from observation. For the superior planets e, λ_A and $\bar{\kappa}$ are found from an iterative computation of considerable complexity using three oppositions of the planet to the mean Sun, and r and $\bar{\alpha}$ from two observations not at opposition. For the inferior planets e, λ_A and r are found through observations of greatest elongations from the mean Sun, and $\bar{\alpha}$ from an observation not at greatest elongation.

This is but a brief summary of Ptolemy's planetary theory in the *Almagest*, upon which the *Handy Tables* are based; there are additional complications for Mercury and rather complicated models for motion in latitude, due to errors in observations of extreme latitudes, in which the inclinations of the planes of the epicycles oscillate. The *Planetary Hypotheses* is notable both for the physical, spherical models that Ptolemy developed from the mathematical, circular models of the *Almagest*, and for a method of computing the distances and sizes of the planets (see DISTANCES (FROM ANTIQUITY TO 1900)). The two are related since the spherical models are contiguous, with no empty spaces, and the distances depend upon this contiguity. Figure 4 shows a cross section of the spherical model based upon the circles of figure 3. The epicycle, a complete sphere with the planet just inside it at P is carried by the eccentric sphere, the inner and outer surfaces of which are concentric to M. Within the eccentric sphere is an inner sphere with its outer surface concentric to M and its inner surface concentric to O. Outside the eccentric sphere is a corresponding outer sphere with its inner surface concentric to M and its outer surface concentric to O. The spheres of a lower planet lie inside the inner sphere and the spheres of a higher planet lie outside the outer sphere in the same way. The center of the eccentric sphere is M, but the sphere rotates such that the center of the epicycle C moves uniformly about the equant point E, and the epicycle rotates such that the planet moves uniformly with respect to the direction EC extended to the far side of the epicycle. But how can rigid, spherical bodies do this? For the spheres must be rigid in order to carry the epicycle and the planet, that is, the bodies are not moving through a fluid medium or empty space. This is the celebrated violation of uniform, circular motion in Ptolemy's models that was noted by Arabic astronomers and Copernicus, and both developed nearly identical alternative models to maintain uniform motion about the equant point strictly through the uniform rotation of spheres (see ISLAMIC ASTRONOMY, NICOLAUS COPERNICUS).

The original modifications of Ptolemy's models were the work of a number of astronomers associated with the observatory of Marāgha in north-western Persia in the second half of the thirteenth century, Naṣīr ad-Dīn aṭ-Ṭūsī, Mū'ayyad ad-Dīn al-'Urḍī, Quṭb ad-Dīn ash-Shīrazī, and in the following century in Damascus Ibn ash-Shāṭir. Their

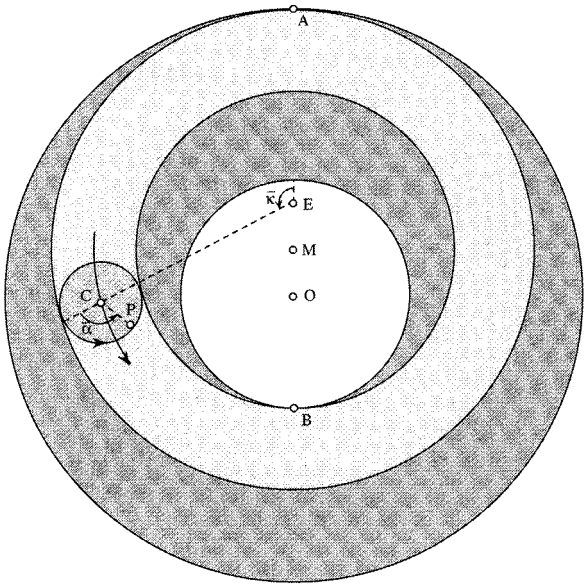


Figure 4.

models work on the principle of breaking up the equant motion in Ptolemy's model into two or more components of uniform circular motion that may be produced by the rotation of spheres. The models for the first inequality of the superior planets of al-'Urḍī and ash-Shīrazī, which are identical, and the model of Ibn ash-Shāṭir are shown superimposed on Ptolemy's model in figure 5. The center of the Earth is O, the center of the eccentric in Ptolemy's model is M, the center of the equant is E, and the two eccentricities $OM = ME = e$ are equal; the center C of the epicycle of the second inequality of figure 3, which is not shown here, moves in a circle of radius R about M such that its angular motion, the mean eccentric anomaly $\bar{\kappa}$, is uniform about E. In Ibn ash-Shāṭir's model the center C_1 of an epicycle of radius r_1 moves through $\bar{\kappa}$ in a circle of radius R about O, and the center C_2 of an epicycle of radius r_2 moves in the opposite direction through $\bar{\kappa}$ so that r_1 is always parallel to the apsidal line OA. The center C' of the epicycle of the second inequality moves in the same direction as C_1 through $2\bar{\kappa}$ such that r_2 coincides with r_1 when C_1 is in the apsidal line. Provided that $r_1 = 3/2e$ and $r_2 = 1/2e$, C' will lie on the line EC, thus moving uniformly about E, and the apparent directions OC and OC' will nearly coincide. In 'Urḍī's and Shīrazī's model, we let $OM' = r_1 = 3/2e$ and let C_2 move about M' through $\bar{\kappa}$ on a circle of radius R, so that the motion of C' measured from $M'C_2$ will be $\bar{\kappa}$, and the model is identical in effect to Ibn ash-Shāṭir's. Hence, both models preserve uniform motion about the equant and are observationally indistinguishable from Ptolemy's. We shall see that the same is true of Copernicus's model, to which the points $\bar{S}$ and P' in parentheses refer.

The Marāgha astronomers also devised notable modifications of Ptolemy's lunar model, both preserving

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

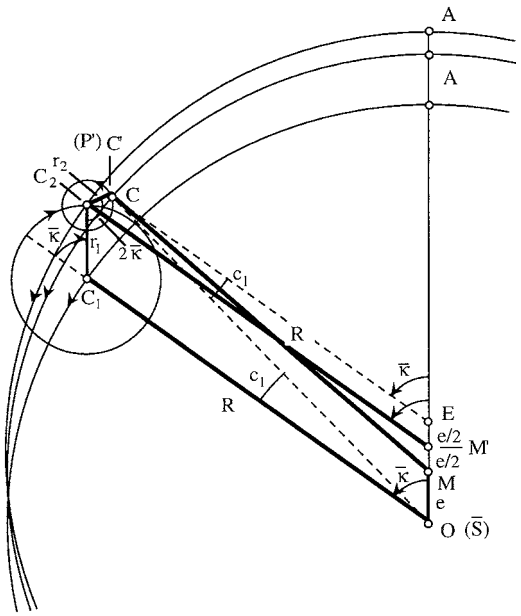


Figure 5.

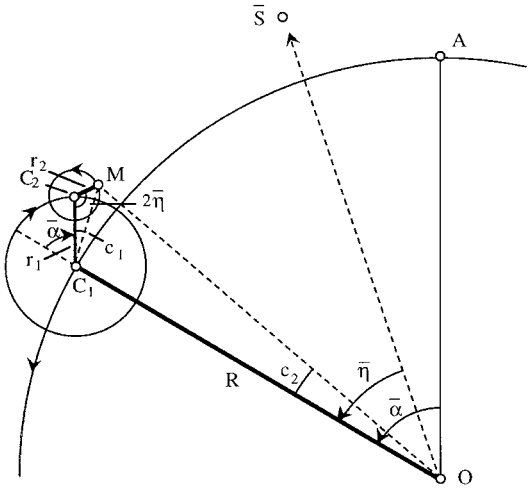


Figure 6.

uniform rotation of spheres and greatly reducing the variation of lunar distance. The model of Ibn ash-Shāṭir, to which Copernicus's model is identical, is shown here in figure 6. The Earth is at O, the apogee, which moves about $7'$ per day or 40° per year, is A, from which the center C_1 of a larger epicycle of radius r_1 moves through the mean anomaly $\bar{\alpha}$ and through the mean elongation $\bar{\eta}$ with respect to the mean Sun $\bar{S}$. The radius r_1 moves in the opposite direction through $\bar{\alpha}$, thus remaining parallel to OA, and the radius r_2 of a smaller epicycle, with center C_2, carries the Moon M in the same direction as C_1 through $2\bar{\eta}$, twice the elongation from the mean Sun. The effect of r_2 is to introduce a correction c_1 to the mean anomaly, and to vary the equation of the anomaly c_2 from a minimum at conjunction and opposition when $2\bar{\eta} = 0^\circ$ to a maximum at quadrature when $2\bar{\eta} = 180^\circ$. Both effects are also found in Ptolemy's model, but at the cost of a variation in lunar distance reaching about 1.9/1, which is here reduced to about 1.3/1, still more than twice as great as it should be; the correct variation was found by TYCHO BRAHE, whose lunar theory was a great improvement over all predecessors.

These innovations in planetary and lunar theory first appear in Europe in the sixteenth century in the work of Copernicus, although just how they were transmitted, perhaps by way of Italy, is uncertain. Ptolemaic astronomy itself only became known in the late twelfth century with the translation into Latin of the *Almagest* and of derivatives of the *Handy Tables*, first the *Toledan Tables* and then in the early fourteenth century the *Alfonsine Tables*. The most important exposition of Ptolemy's astronomy was the *Epitome of the Almagest* by GEORG PEURBACH (1423–61) and Johannes REGIOMONTANUS (1436–76), which first made Ptolemy's difficult book intelligible and was used along with, or even in place of, the *Almagest* until the beginning of the seventeenth century. In addition to Ptolemaic astronomy, the early sixteenth century saw a renewed interest in concentric-sphere models among Aristotelians, particularly at Padua, but these were little more successful than the models of Eudoxus and were never used for computation. The first serious innovation in planetary theory was that of Copernicus (1473–1543), built upon Ptolemy's and upon the models of the Marāgha astronomers. Copernicus's principal innovation is in accounting for the second inequality of the superior planets and the first inequality of the inferior planets by the annual motion of the Earth about the Sun. He arrived at his theory by way of transformations of Ptolemy's models explained by Regiomontanus in the *Epitome of the Almagest*, a work he used even in preference to the *Almagest* as his guide to Ptolemy's astronomy. He first explained his new theory in a short work known as the *Commentariolus*, written at some time between 1510 and 1514, in which the parameters of the models were merely extracted from the *Alfonsine Tables*. The original theory was modified and expanded greatly in *De Revolutionibus Orbium Coelestium* (On the Revolutions of the Heavenly Spheres), published in 1543, essentially a new *Almagest*, in which he derived new parameters from observation following Ptolemy's methods, again as explained by Regiomontanus.

The relation of Ptolemy's and Copernicus's models for the superior planets is shown in figure 7, in which we assume only circular motions. In Ptolemy's model, the Earth O is at rest, about it is described a circle of radius R', on which moves an epicycle with center C and radius r'. The planet P moves on the epicycle through the mean anomaly $\bar{\alpha}$ measured from the apogee of the epicycle A' such that the radius CP is parallel to the direction $O\bar{S}'$ from the Earth to the mean Sun, the Sun with only its uniform motion. The distance $O\bar{S}'$ is arbitrary in Ptolemy's model, but if a parallelogram $OCP\bar{S}$ is constructed and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

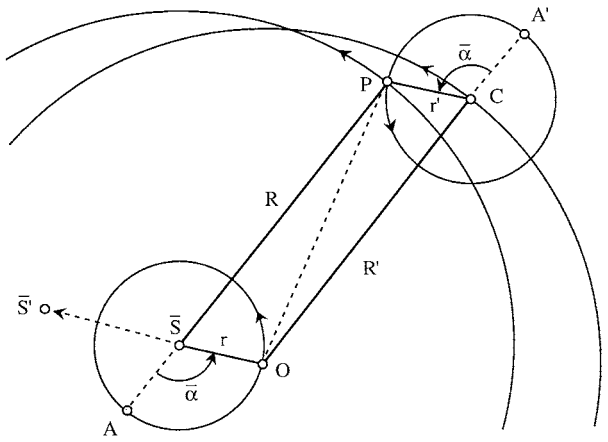


Figure 7.

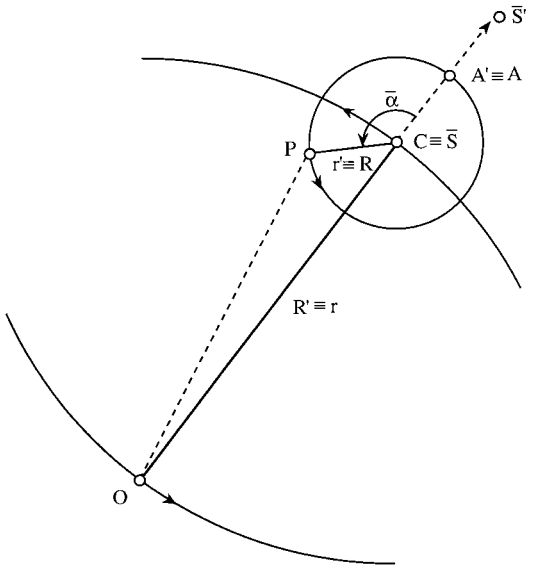


Figure 8.

it is assumed that the mean Sun is at $\bar{S}$, then both the Earth O and the planet P may be taken to move about $\bar{S}$ in circles of radii r and R where $r/R = r'/R'$, and the direction OP from the Earth to the planet, the diagonal of the parallelogram, is identical in both models. The relation of the models for the inferior planets is shown in figure 8, again assuming only circular motions. As before in Ptolemy's model, about the Earth O at rest is described a circle of radius R', on which is an epicycle with center C and radius r'. But now OC lies in the direction $O\bar{S}'$ from the Earth to the mean Sun, where the distance $O\bar{S}'$ is arbitrary, and the planet P again moves on the epicycle through the mean anomaly $\bar{\alpha}$ measured from the apogee of the epicycle A'. If it is assumed that the mean sun $\bar{S}$ coincides with C, is identical to $(\equiv)C$, then again both the Earth O and the planet P may be taken to move about $\bar{S}$ in circles of radii r and R, where $r/R = r'/R'$, and the direction OP from the Earth to the planet is identical in both models.

The essential advantage of the Copernican theory is concerned with the order and distances of the planets, the unification of the planetary models into a single system in which everything is determined and nothing arbitrary. In Ptolemy's models, the radii r' and R' are arbitrary for each planet, only the ratio r'/R' is determined by observation; thus there is no common measure of distances within the planetary system, and even the order of the planets is arbitrary without further assumptions. However, in Copernicus's models, the radii R of the orbits of all the planets are measured in terms of the unit r, the radius of the Earth's orbit, which is common to the model for every planet, so neither the distances nor the order of the planets are arbitrary, for they are fixed by the ratio r/R, which, as in Ptolemy's models, is determined by observation. There are also certain arbitrary rules governing Ptolemy's models that are explained when it is seen that geocentric planetary models are actually transformations of underlying heliocentric models. The most obvious in the figures are why the

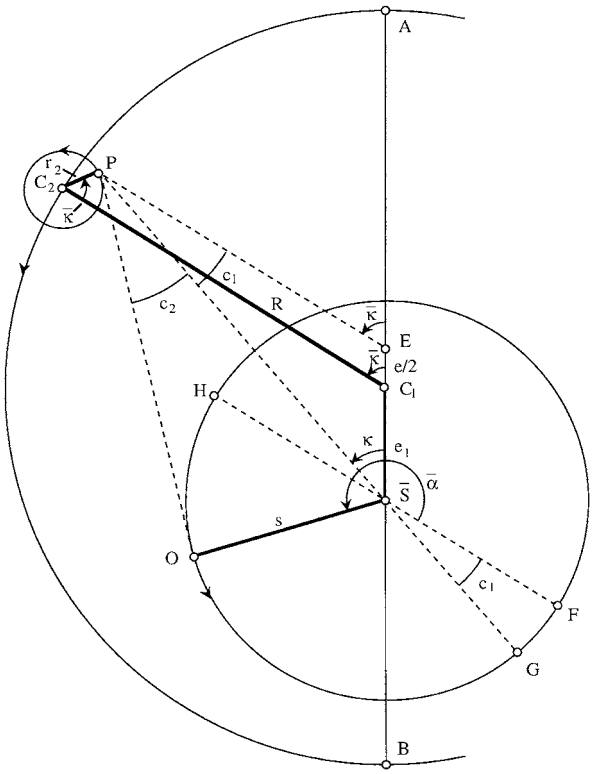


Figure 9.

radii of the epicycles of the superior planets are parallel to the direction from the Earth to the mean Sun, and why the centers of the epicycles of the inferior planets lie in the direction of the mean Sun, but there are many more properties of geocentric models, even as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

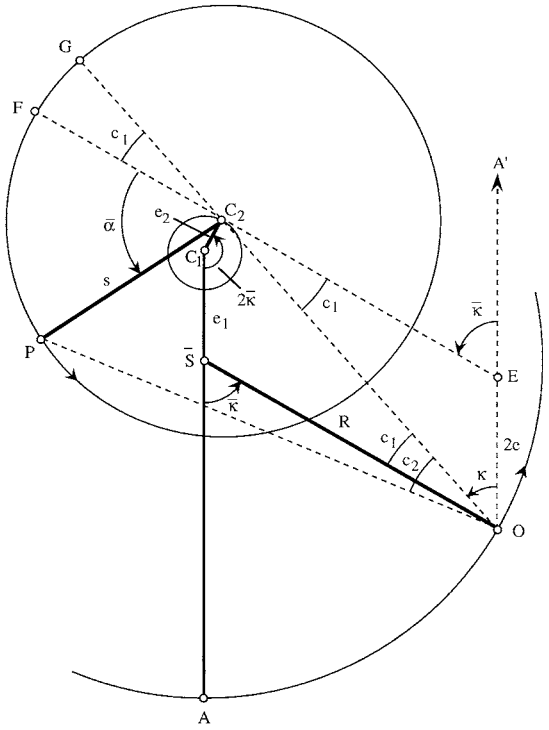


Figure 10.

basic as why the superior planets move in the same directions on their epicycles, that are only explained as transformations of heliocentric models. The most well-known consequence of the heliocentric theory is the explanation of retrogradation (see RETROGRADE MOTION), which takes place in Ptolemy's models near the perigee of the epicycle when the apparent backward motion of the planet on the epicycle exceeds the forward motion of the center of the epicycle, which happens to be near opposition for a superior planet and near inferior conjunction for an inferior. In Copernicus's models retrogradation occurs when the Earth passes a slower superior planet near opposition or is passed by a faster inferior planet near inferior conjunction, so the location of the retrogradation is also explained.

Copernicus's solar theory is the same as Ptolemy's, a simple eccentric, now with the Earth moving about the Sun, but, following medieval sources, he found a smaller solar eccentricity than did Ptolemy and he developed a model to produce a variation in the eccentricity, along with an inequality in the motion of the solar apsidal line, such that the eccentricity is near maximum at Ptolemy's time and near minimum at his own. Following the Marāgha astronomers, Copernicus objected to Ptolemy's lunar model on the grounds that it violates uniform, circular motion, since the center of the epicycle does not move uniformly about the center of the eccentric on which it is carried, and also that it produces a large variation in lunar distance contradicted by the small variation in the apparent size and parallax of the Moon. His own lunar model is identical to that of Ibn ash-Shāṭir shown in figure 6.

Copernicus's model for the first inequality of the planets in the *Commentariolus* is also that of Ibn ash-Shāṭir using two epicycles, shown above in figure 5, but with the mean Sun $\bar{S}$ in place of the Earth O and the planet itself P' in place of the center C' of the epicycle of the second inequality, as shown in parentheses in the figure. It thus preserves uniform motion about the equant E just as in Ptolemy's model. The model for the superior planets is altered in *De Revolutionibus* to that of al-'Urḍī and ash-Shīrazī, and is shown in figure 9, which shows both inequalities. As in figure 5, the larger epicycle is replaced by an equal eccentricity $e_1 = r_1 = 3/2e$, which has the same effect since r_1 is always parallel to e_1; r_2 is unchanged, although its motion, because it is measured from C_1C_2, is now $\bar{\kappa}$. Thus P still moves uniformly about E and, as in Ptolemy's model, the true eccentric anomaly $\kappa = \bar{\kappa} \pm c_1$. The second inequality is accounted for by the motion of the Earth O in a circle of radius s about the mean Sun $\bar{S}$ through the mean anomaly $\bar{\alpha}$, uniformly with respect to a line FH which is drawn parallel to EP; the true anomaly, from line $G\bar{S}$, is $\bar{\alpha} \pm c_1$. The true position of the planet as seen from the Earth is found through two corrections, c_1, the equation of center, the difference in direction between the uniform motion of the planet about the equant and the non-uniform motion about the mean Sun, and c_2, the equation of the anomaly, found from the true anomaly, which reduces the heliocentric direction of the planet $\bar{S}P$ to the geocentric direction OP. Given the longitude of the apogee λ_A, the longitude of the planet is $\lambda_A + \bar{\kappa} \pm c_1 \pm c_2$. The corrections are nearly identical to those used by Ptolemy, and Copernicus's tabulations of the corrections are also nearly identical to Ptolemy's. Likewise, Copernicus finds the parameters of the model for each planet by observation through exactly the same methods as used by Ptolemy, and except for the directions of the apsidal lines, his results are the same or quite close.

The model for the inferior planets, shown for Venus in figure 10, is somewhat different in that the equation of center depends, not upon the distance of the planet from its own apogee, but, remarkably, upon the distance of the Earth from the planet's apogee, due to a direct adaptation of Ptolemy's model. The Earth O moves through $\bar{\kappa}$ with respect to the planet's apsidal line while at the end of the fixed eccentricity $e_1 = 3/2e$ a movable eccentricity $e_2 = 1/2e$ rotates through $2\bar{\kappa}$. The result is that the equation of center c_1 is measured at the Earth, just as in Ptolemy's model, and the direction EC_2 from the equant to the center of the planet's orbit lies in the direction of $O\bar{S}$ from the Earth to the mean Sun, just as though C_2 moved uniformly through $\bar{\kappa}$ about E at a distance $2e$ from O in the direction of the geocentric apogee A', again as in Ptolemy's model. The equation of center c_1 and the equation of the anomaly c_2 are computed and tabulated as for the superior planets, and again the longitude of the planet is $\lambda_A + \bar{\kappa} \pm c_1 \pm c_2$. There are further complications for Mercury, which are identical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to those of Ibn ash-Shāṭir's model, most notably that the radius of the planet's orbit is itself variable as a function of $2\bar{\kappa}$, twice the Earth's distance $\bar{\kappa}$ from the apsidal line, the result being an excellent approximation of Ptolemy's model.

Copernicus's models for the latitudes of the planets are also direct adaptations of Ptolemy's, which require oscillations of orbital planes that Kepler found implausible and prompted his famous remark, 'Copernicus, ignorant of his own riches, took it upon himself for the most part to represent Ptolemy, not nature, to which he had nevertheless come the closest of all.' Kepler could not be more correct. Copernicus's astronomy was built upon Ptolemy's descriptions of the apparent motions of the heavens, upon Ptolemy's observations, upon transformations of Ptolemy's models, and upon Ptolemy's methods of deriving parameters from observation. Thus, he did, as Kepler said, represent Ptolemy rather than nature.

Bibliography

On planetary theory in general:

Dreyer J L E 1906 *History of the Planetary Systems from Thales to Kepler* (Cambridge: Cambridge University Press) (reprinted 1953 as *A History of Astronomy from Thales to Kepler* (New York: Dover)

Herz N 1887–94 *Geschichte der Bahnbestimmung von Planeten und Kometen* 2 vols (Leipzig: Teubner)

Neugebauer O 1975 *A History of Ancient Mathematical Astronomy* 3 parts (New York: Springer)

On Eudoxus:

Heath T 1913 *Aristarchus of Samos, the Ancient Copernicus* (Oxford: Oxford University Press) pp 190–224

Schiaparelli G V 1875 Le sfere omocentriche di Eudosso, di Callippo e di Aristotele *Pubblicazioni del Reale Osservatorio di Brera in Milano* **9** (also in Schiaparelli G V 1925–7 *Scriti sulla Storia della Astronomia Antica* vol 2 (Bologna: Zanichelli) pp 3–112

Yavetz I 1998 On the homocentric spheres of Eudoxus *Arch. Hist. Exact Sci.* **52** 221–78

On Ptolemy:

Pedersen O 1974 *A Survey of the Almagest* (Odense: Odense University Press)

Toomer G J 1984 *Ptolemy's Almagest* (New York: Springer)

On Marāgha planetary theory:

Kennedy E S 1983 *Studies in the Islamic Exact Science* (Beirut: American University of Beirut) pp 50–97

Ragep F J 1993 *Naṣī al-Dīn al-Ṭūsī's Memoir on Astronomy (al-Tadhkira fī 'ilm al hay'a)* 2 vols (New York: Springer)

Saliba G 1993 *A History of Arabic Astronomy, Planetary Theories During the Golden Age of Islam* (New York: New York University Press)

On Copernicus:

Armitage A 1957 *Copernicus, the Founder of Modern Astronomy* (New York: Thomas Yoseloff)

Swerdlow N M 1973 The derivation and first draft of Copernicus's planetary theory. A translation of the *Commentariolus* with commentary *Proc. Am. Phil. Soc.* **117** 423–512

Swerdlow N M and Neugebauer O 1984 *Mathematical Astronomy in Copernicus's De Revolutionibus* 2 parts (New York: Springer)

N M Swerdlow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetesimal

In cosmogony, a planetesimal is one of numerous solid bodies of sub-planetary size orbiting the SUN (or another star), from which planets may accumulate. The term was coined near the beginning of the twentieth century by T C Chamberlin and F Moulton, who developed the 'planetesimal hypothesis' for the origin of the SOLAR SYSTEM. In their original concept, solid particles condensed from matter ejected from the Sun by tidal forces during a close encounter with another star. Collisions between objects in intersecting orbits caused them to accumulate into progressively larger bodies, forming the planets. Many features of the Chamberlin–Moulton theory have been rejected or superseded, but the concept of growth of planets from planetesimals has been adapted to the context of modern theories.

Formation of planetesimals

Both theory and observation indicate that a star forms by collapse of an interstellar cloud, composed mostly of hydrogen gas, when internal pressure cannot support it against its own gravity. If the cloud has some initial rotation, conservation of angular momentum ensures that some of its mass is left orbiting the star in a flattened, disk-shaped gaseous NEBULA. The nebula contains microscopic solid grains, a mixture of surviving interstellar grains and condensates from the cooling gas. Near the star, these consist of metals and silicate minerals; at larger distances organic compounds and ices are included. The small grains, with large area/mass ratios, are constrained by drag forces to move with the gas. They accumulate into planetesimals, which are sufficiently large to be in Keplerian motion about the star, controlled by gravity rather than gas drag. This is the case for objects of about kilometer size or larger.

The mechanism by which kilometer-sized bodies form from micrometer-sized grains has been a topic of much effort and some controversy. The idea of gravitational instability was suggested by K Edgeworth in 1949. In this concept, solid grains settle to form a thin layer in the central plane of the nebula. If this layer's density exceeds a critical value it becomes unstable, and localized regions collapse under their own gravitational attraction. This model was developed in quantitative form by V S Safronov in the Soviet Union, and independently by P Goldreich and W Ward in the United States. They showed that this process would tend to produce solid bodies a few kilometers in size in the inner solar system, and larger objects at greater distances from the Sun.

The simplicity and elegance of gravitational instability caused it to be widely accepted. The process was believed to be independent of the poorly defined properties of the grains themselves, requiring no sticking mechanism or force other than gravity. However, later investigations have cast doubt upon this concept. Microscopic grains would settle so slowly that the slightest degree of turbulence in the nebula would inhibit formation of the particle layer. Some non-gravitational coagulation mechanism is needed to produce larger aggregates of grains to allow their descent to the central plane of the nebula. Many possible sticking processes have been suggested, including electrostatic charging of grains, magnetic attraction of metallic grains, van der Waals surface forces, sticky coatings of organic material, physical interlocking of irregularly shaped particles, and condensation of ices upon grain surfaces. Some or all of these mechanisms may have dominated in different density–temperature regimes in various regions of the nebula, and are current topics of theoretical and experimental study. Because the rate of settling increases with particle size, larger particles can grow by overtaking and sweeping up smaller ones as they settle. If coagulation occurs, centimeter-sized aggregates of grains can grow in this manner on time scales of the order of a thousand orbital periods at any given distance from the star.

Coagulation during settling allows particles to become concentrated into a layer in the central plane of the nebula, but gravitational instability is inhibited by interactions of particles with the gas. The pressure in the nebula decreases with distance from the Sun. This radial gradient acts to support the gas against the central gravity, and causes the nebula to rotate at slightly less than the Keplerian orbital velocity. This velocity difference is less than one per cent for any plausible nebular structure, but has a typical magnitude of tens of meters per second. A solid body in the nebula lacks this pressure support, and as a consequence cannot be at rest with respect to the gas. Its motion occurs within one of two dynamical regimes that depend on its size. A small particle, coupled to the gas by a large area/mass ratio, moves with the nebula's angular velocity and drifts inward due to the residual radial component of gravity. The drift rate is proportional to the particle's size and density, with magnitude of order 100 times its diameter per second for a body of density one gram per cubic centimeter. A large body with low area/mass is less affected by drag, and pursues a Keplerian orbit. As it moves faster than the gas, it experiences a 'headwind', that causes its orbit to decay. In this regime, the orbital decay rate decreases with size and density, and has a magnitude of order $(100/D)$ m s^{-1}, where D is the diameter in meters. The radial velocity reaches a peak, equal to the difference between the Kepler velocity and gas velocity, at the transition between these regimes, at sizes of about a meter. Thus, solid bodies of all sizes move radially inward, while large bodies also have significant transverse velocities (tens of meters per second) relative to the gas.

If the particle layer becomes denser than the gas, as would be necessary for gravitational instability, collective effects become important. High mass loading allows the particles to become dominant, diminishing the pressure support of the gas within the layer, and causing its motion to be more nearly Keplerian, i.e. faster than the pressure-supported gas on either side. It can be demonstrated analytically that such a shear flow is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

unstable and would become turbulent; this behavior has been confirmed with numerical fluid-dynamic simulations (see FLUID DYNAMICS). This self-generated turbulence is sufficient to prevent further settling of centimeter-sized particles. The layer cannot attain the critical density unless collisional coagulation produces meter-sized or larger bodies, which are unaffected by the turbulence. However, the critical density is necessary, but not sufficient, to allow gravitational instability. The particles must also have sufficiently low relative velocities to allow density perturbations in the layer to grow. The meter-sized bodies have radial and transverse motions induced by gas drag. As these velocities are size dependent, any system of non-identical bodies would have significant velocity dispersion, sufficient to prevent gravitational collapse.

The failure of gravitational instability seems to require that planetesimals grew gradually by collisional coagulation. The details of this evolution depend on their mechanical properties, such as impact strength and energy dissipation, which are unknown. However, collisions driven by nebular gas drag are relatively favorable for particle growth. The size dependence of radial and transverse motions induced by drag implies that for bodies of comparable size impact velocities would be low, so coagulation may be more likely than disruption. Objects of very different sizes would have more frequent collisions at higher speeds, in which the smaller projectile could be embedded in the larger target. Drag-induced velocities decrease with size, so collisions become more gentle as the planetesimals grow. When they reach sizes of the order of a kilometer, gravitational attraction increases their impact velocities. Thus, planetesimals built up in this manner may have a lumpy structure on scales from meters to hundreds of meters. Numerical simulations of collisional growth driven by drag produce kilometer-sized or larger bodies on time scales of a few thousand orbital periods at any given heliocentric distance. In these simulations, coagulation and settling is only moderately efficient, so a significant population of small grains remains after planetesimals have formed. This result is consistent with observations of DUSTY CIRCUMSTELLAR DISKS; the presence of dust does not mean that planetesimals have not yet formed. Some disks appear to have dust-free inner regions; such clearing probably is not a direct result of planetesimal formation, but is due to other processes operating on longer time scales.

Evolution of planetesimals to planets

After planetesimals reach sizes of a few kilometers, gravity becomes more important than gas drag, at least for short-term behavior, and their motions can be described as Keplerian orbits. As the planetesimals become larger, their mutual gravitational perturbations stir up random velocities, i.e. orbital eccentricities and inclinations. Analyses of this stirring process by kinetic theory show that the mean random velocity tends to be comparable to the escape velocity of the bodies that dominate the mass distribution in the planetesimal swarm. As the collisions are inelastic, gravitational binding ensures that the planetesimals experience net growth during this stage, regardless of their impact strength or mechanical properties.

The continuing evolution of the swarm of planetesimals is controlled by the interplay of stirring and collisions. V S Safronov developed an analytic solution to the coupled distributions of sizes and velocities, in which the mass distribution is assumed to be a power law. Its slope is such that it is dominated by the largest bodies. At any stage of the swarm's evolution, random velocities are comparable to the escape velocity of these bodies, increasing as they grow. This process, sometimes referred to as 'orderly growth', continues with collisions at larger scales and higher velocities until a few large bodies (PLANETS) survive in widely separated stable orbits that do not intersect.

Numerical simulations of planetesimal ACCRETION have revealed another possible path for the growth of planets. At low relative velocities, gravitational attraction deflects trajectories, allowing collisions between orbits that would otherwise not intersect. This enhancement of the collisional cross section increases as the square of the ratio of the mutual escape velocity of two planetesimals to their relative velocity. In a swarm containing bodies with some range of sizes, this factor tends to favor growth of the larger bodies due to their higher escape velocities. In the orderly growth model relative velocities are high enough to minimize this advantage. However, more recent models of velocity stirring include dynamical friction, a tendency for gravitational interactions to produce equipartition of the energy in random motions. More massive bodies tend to have lower velocities, giving them a greater advantage. In numerical simulations of accretion in such a system, a small fraction of planetesimals, the first to become larger than the average size, experience 'runaway growth'. The size distribution becomes bimodal, with a few large bodies (protoplanetary embryos) accreting while the mass distribution is still dominated by small planetesimals. This runaway growth is self-limiting, being halted by depletion of the small bodies in orbits near those of the embryos, when the latter have attained masses of the order of one per cent that of the Earth. Many such bodies must collide in order to produce a small number of widely spaced planets as found in our own solar system. Mutual perturbations among the embryos may cause their orbits to cross after the smaller planetesimals have been depleted. The final stage of planetary accretion probably involved very large impacts, like that which has been invoked to explain the formation of the Moon.

This scenario appears to account for the formation of the TERRESTRIAL PLANETS. The estimated time scale for their accretion, of the order of tens of millions of years, does not violate geophysical constraints on the age of the Earth–Moon system. However, it is less successful when applied to the outer planets. One widely accepted model for the origin of gas GIANT PLANETS, Jupiter and Saturn, is accumulation of planetesimals of mixed rock/ice

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

composition into cores with masses of the order of ten times that of Earth, which then accrete gas from the solar nebula. Despite the higher surface density of solids due to condensation of ice, the longer orbital periods at greater heliocentric distances yield slower growth of the cores. Current models do not produce cores of the requisite mass within the estimated age of the nebula. Models for accretion of Uranus and Neptune from icy planetesimals fare even worse, failing to produce those planets within the age of the solar system. Probably unmodeled phenomena, such as tidal interactions between massive planetary embryos and the nebula, were involved.

ASTEROIDS and COMETS represent populations of planetesimals that survive in the present solar system. The former have been greatly altered since their formation by collisional evolution; the smaller asteroids are not primordial, but are fragments of once larger bodies that have been disrupted. Comets in the Oort cloud and Kuiper belt have probably undergone little change since their formation. *In situ* examination by spacecraft may reveal structures elucidating their accretion in the solar nebula.

Bibliography

Brush S 1996 *Fruitful Encounters: The Origin of the Solar System from Chamberlin to Apollo* (Cambridge: Cambridge University Press)

Lissauer J J 1993 Planet formation *Ann. Rev. Astron. Astrophys.* **31** 129–74

Safronov V S 1969 *Evolution of the Protoplanetary Cloud and Formation of the Earth and Planets* (Jerusalem: Keter) (NASA Technical Translation TTF-677)

Weidenschilling S J 1997 The origin of comets in the solar nebula: a unified model *Icarus* **127** 290–306

Wetherill G W 1990 Formation of the Earth *Ann. Rev. Earth Planet. Sci.* **18** 205–56

S J Weidenschilling

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planetology

Although created from the same solar nebula, the bodies residing in our solar system are all very different, each with its own distinct characteristics. The comparison of features and processes between different bodies can tell us about the evolution of those planets, or give us an idea of how things may have evolved if certain physical parameters were just slightly different. Comparing a particular process on two similar, but subtly different worlds allows us to isolate the contribution that just one or two physical parameters makes to the final form we see. Only in recent years, with the advent of spacecraft visiting other planets and taking detailed measurements of those systems, have we truly been able to compare one planet or satellite with another. But now, even superficial analysis reveals that the planets and satellites in our solar system are very different, each telling its own story and each able to tell us a little about our own past, present and perhaps even the future.

Comparison of surface processes

The surfaces of solid bodies can tell us much about its history. The different processes that can occur on the surface of a planet or satellite indicate properties of both its internal and external evolution. There are many processes that can act upon planets, satellites and asteroids, some affecting all, and others operating on only one or two. Here we look at the main processes that have acted to shape the PLANETARY SURFACES we see, revealing their differences and what those differences can tell us about the bodies themselves.

Impact cratering

Impact cratering is the most important process acting upon the surface of solid bodies in the solar system. Without exception, every planet, satellite and small body is subject to the effects of impacts, both large and small, past and present. The impact cratering process can tell us much about the history of a body as well as its present state; it can highlight the differences found between worlds and tell us something about the characteristics of the solar system itself. Without impact CRATERS, our knowledge of the solar system and the planets within it would be much poorer.

While every planet in the solar system has impact craters on its surface (figure 1), some have more than others or one region of a planet may have a higher density of craters than another area on the same planet. The MOON, for example, is completely saturated with impact craters on its highland surface—in other words, every new impact crater obliterates an older one, so that the crater density observed is constant in the highlands. On the dark maria (solidified ancient lava flows) however, the number of craters per unit area is much lower. The reason for this is quite logical—the maria have been exposed to space for less time than the highlands, and hence have had less time to accumulate impact craters. Therefore, the simple observation that the maria have less craters immediately tells us that they are younger than the highlands. This is true for any solid planetary body in the solar system although it starts to get more complicated when trying to compare the age of one planetary surface with another purely by counting craters. Firstly, the cratering flux through time is not constant. We can tell this by looking at the Moon. The highlands are known to be 4 Gyr or older and the maria are between 3.1 and 3.8 Gyr old. Yet the difference in the number of impact craters gives the appearance that the maria are very much younger. This implies that in the very earliest times, before the maria were emplaced, the impact cratering flux was much higher, and this has now become known as the period of heavy bombardment. The cratering flux dropped off dramatically after that time, and is now believed to be quite low. As the Earth is in close proximity to the Moon, it can be assumed that our planet underwent the same bombardment as the Moon. In fact, we may have had a slightly higher flux due to our greater gravitational attraction compared to the Moon. Therefore, the study of impact craters on the Moon can tell us something about the conditions on Earth about 4 Gyr ago—a period where terrestrial records have long been wiped out. While it is logical to believe that the whole solar system underwent a similar period of bombardment, the intensity of the flux may have varied quite considerably through space. It is possible that the immense gravitational pull of Jupiter may have caused more impacts to occur on the surfaces of its satellites than on the inner planets, for example. Or perhaps large impact events in the inner solar system threw off more debris destined to impact on the terrestrial planets. There may be many factors affecting the cratering flux through time and space, and hence to use the crater density on a planetary surface as an age indicator can give a great error on the value obtained. However, at the present time, counting craters is the only way we can estimate the age of a surface for which we have no sample, and this means every surface except the few areas for which we had rocks brought back from the Moon. Nevertheless, we are able to obtain crude estimates for relative ages, and in some cases this can tell us much about a planet's history. For example, on VENUS, the crater density is very low. Although Venus is relatively large (approximately the same size as the Earth) it has less than 1000 craters across its entire surface, distributed in a random fashion. This number is far less than would be expected if the surface had been exposed for the length of time that the lunar highlands have. Therefore the surface of Venus must be relatively young, with some scientists believing that Venus underwent a global resurfacing event only 500 million years ago, effectively resetting its age to zero.

Similarly, MARS also must have undergone some kind of resurfacing event, this time across only one half of the planet. Mars can be split into two hemispheres; the south, which is higher and has a high crater density; and the north, which is lower and has a low crater density. Lava flows can be seen across the northern plains, and it is believed that widespread volcanism is one of the reasons

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

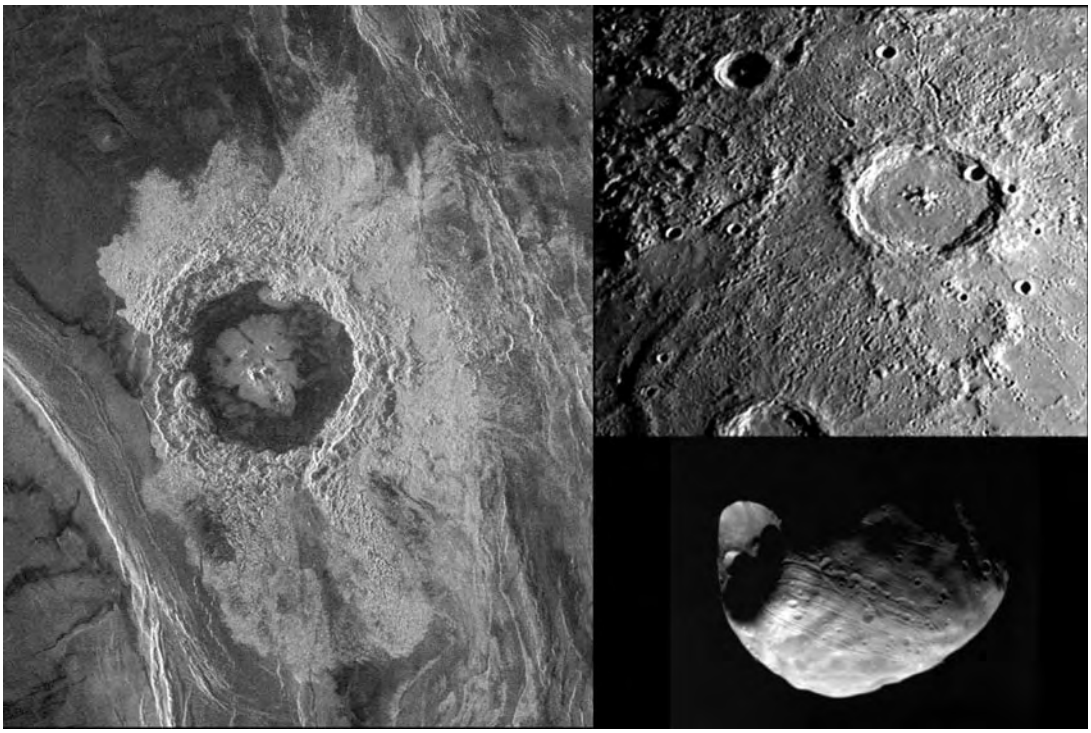


Figure 1. Impact craters on selected bodies. Left: a Magellan radar image of the 43 km diameter crater Dickinson, showing a complex structure with a central peak, terraced walls and flat floor. (Courtesy of NASA JPL/Caltech.) Top right: a large crater on Mercury, 140 km in diameter again showing a complex shape. Note the many smaller impact craters also in the image. (Courtesy of NASA JPL/Caltech.) Bottom right: Mars' moon Phobos has a large impact crater almost as large as the moon itself. Phobos is only 28 km by 20 km but still suffers from the effects of impact cratering. (Courtesy of NASA.)

for the dearth of impact craters across this hemisphere of the planet. MERCURY, the closest planet to the Sun, is very Moon-like. It is a small planet, rocky with no atmosphere and therefore no weathering process to destroy the record of impacts across its surface. We would therefore expect to see a similar cratering record on Mercury as we do in the highlands of the Moon, but we do not. Although Mercury's surface is heavily cratered, most of the surface is not saturated, and once again, it is believed that some kind of resurfacing event occurred early in Mercury's past. The satellites of the outer planets do not escape this kind of analysis either, and we find some very interesting results. Many of the larger satellites of, for example, JUPITER, show very few impact craters, indicating that their surfaces are young. IO, one of Jupiter's moons, is the only solid body that does not have impact craters on its surface (at least not at the resolution of our images). This does not mean that Io has never experienced an impact, it instead shows that the surface is extremely young in geological terms, some say as young as 10 000 years. This is not difficult to imagine, as we know that the surface of Io is constantly being renewed by volcanism (see later). Another satellite of Jupiter, EUROPA, has a low crater density. Its surface is known to be covered in a layer of thick water ice, and it is possible that water from beneath the icy crust is periodically brought to the surface covering the previous record of craters upon freezing. To obtain absolute ages it is necessary to get samples from areas of different crater densities and date them in the laboratory. Therefore, it is our Moon that provides the only key to this technique at the moment, because we have those samples to work with. We can infer much about the conditions in the solar system during its very earliest times as the Moon's surface does not have any significant weathering processes to degrade the impacts on its surface. The impact craters that we see there today (along with the footprints and tyre tracks left by the Apollo astronauts) will still be there in a billion years' time.

The Earth has a very young surface, as evidenced by the relatively few impact craters we have found here. The volcanic and tectonic activity that is still going on today are constantly renewing the Earth's surface, and the weathering caused by the action of wind and rain quickly degrade those craters we do see. Weathering also occurs on planets such as Mars and Venus, as these worlds both have atmospheres. In fact some craters may appear to be older than they truly are where weathering is particularly efficient. Other factors may also influence the shape of the impact craters, such as the angle at which the projectile hits the surface, the atmospheric density and, perhaps most importantly, the volatile content of the surface at the time of impact. For example, we can use impact craters to tell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

us about the distribution of water across the surface of Mars through time by looking for distinct crater patterns. When an impact crater forms, material is ejected from the crater and thrown out onto the surrounding surface. When volatiles (i.e. water) are present, the ejecta takes on a mud-like appearance, as though it flowed as it hit the surface. The presence of a crater with this feature indicates that some water must have been present in the area (probably as a permafrost layer) at the time of the impact. Mapping the distribution of these craters across Mars gives an idea of where water has existed in the past, and hence may give us an idea of where life may have existed on the planet, if it did at all. These fluid patterns are unique to Mars and show how a single factor (in this case volatiles) can have a significant effect on the appearance of a crater. Similarly, by comparing the Moon and Mercury (both small worlds with no atmosphere and no current day activity), we can directly study the effects of gravity in the impact cratering process (Mercury has a higher gravitational pull than the Moon).

Of course, it is not only solid bodies that attract projectiles. The giant planets, Jupiter, Saturn, Uranus and Neptune, also suffer from the impact of large and small bodies, but as they have no solid surface, they hold no record of impact events of the past. Scientists were recently privileged to witness such an impact. The comet Shoemaker-Levy 9 was captured by the gravitational influence of Jupiter, the largest planet in the solar system, and was subsequently torn apart before being dragged into the atmosphere of the planet. Scientists watched each explosion as the comet pieces were consumed, and found that traces of each impact were visible for some time before finally disappearing. This was the first large impact event observed by modern-day scientists and their instruments. If nothing else, it brought home the fact that these impacting bodies are still out there and can still collide with any body in its way, even the Earth. By studying these impacts we can learn about how they may have affected the history of our own planet. It was the impact of an asteroid in the Yucatan Peninsula 65 million years ago that is believed to have caused the extinction of the dinosaurs. Such an impact today would cause world-wide devastation, and while we would hope that technological advances would allow some people to survive, the majority of Earth's human population and that of the large animals would certainly perish. It is a sobering thought that less than 100 years ago, a comet exploded in the Earth's atmosphere above Siberia flattening trees for many kilometers. Even a cometary explosion like this would kill thousands if located over a densely populated city. It is therefore very important to study the impact process and to identify precisely how many impacts we can expect over the next few hundred years—our lives may depend on it.

Volcanism

Volcanism, although not obvious on all planets, does occur, or at least has occurred, on many of the bodies we can see. Volcanism clearly still plays a role in the evolution of the Earth today and there are many instances of volcanic eruptions causing disruption to populations over recorded history. Volcanism was once far more extensive in the distant past however, with vast areas of the planet being covered with lava. These flood lavas, as they are known, are not active on Earth today, but it is the most common form of volcanism that has occurred on other planets. The dark maria on the Moon are composed of basaltic lava of low viscosity, erupted from large fissures (cracks) in its crust. The volcanism on the Moon occurred over a relatively short period of time (3.1–3.8 Gyr ago) and has remained dormant since that time, apart from the possibility of minor outbreaks of volcanic activity up to 2.5 Gyr ago. Mercury, on the other hand, is somewhat enigmatic when it comes to working out whether volcanism has occurred on this planet. We have photographs of only 45% of its surface and it is not conclusive from those alone whether volcanic activity has shaped the planet in any way. It seems likely that some of the plains regions are volcanic, and recent re-analysis of Mariner 10 data, taken through ultraviolet and orange filters, suggests that this is the case. One thing that does remain clear is that further study of Mercury is essential if we are to understand more of the role volcanism has played in its past.

From a world where volcanism has played an unknown role, we move to a world where there is little doubt that volcanism has significantly shaped the evolution of its surface. Across the entire planet of Venus, volcanic features can be seen. There are estimated to be over a million small volcanoes (<20 km diameter) across the planet, with a significant number of larger volcanoes to be found also. The volcanoes take on different shapes, which may be attributable to differences in the viscosity and perhaps composition of the lavas extruded. A similar situation occurs on Earth, with volcanoes at different locations erupting lavas of a different composition. Lava channels on Venus may extend for several hundred kilometers, which indicates that the lava was molten for a significant length of time, far longer than would be the case on Earth. It is highly likely that the elevated surface temperatures on Venus contributed to the ability of the lava to remain molten long enough to create these features. As we saw in the section on impact cratering above, the cratering record implies that Venus underwent a global resurfacing event a relatively short time ago. This event would have been volcanic, with the release of huge quantities of lava onto the surface covering the previous record of Venus' past. There is still a debate as to whether this event happened simultaneously (in geological time) across the whole surface or as a continuous set of discrete eruptions occurring over a longer period of time. Whichever theory is correct, the planet Venus obviously underwent a massive change, which effectively prevents us from learning about the very early stages of Venus' evolution.

Mars is very much easier to study than Venus. For one, it does not have a thick optically opaque atmosphere

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

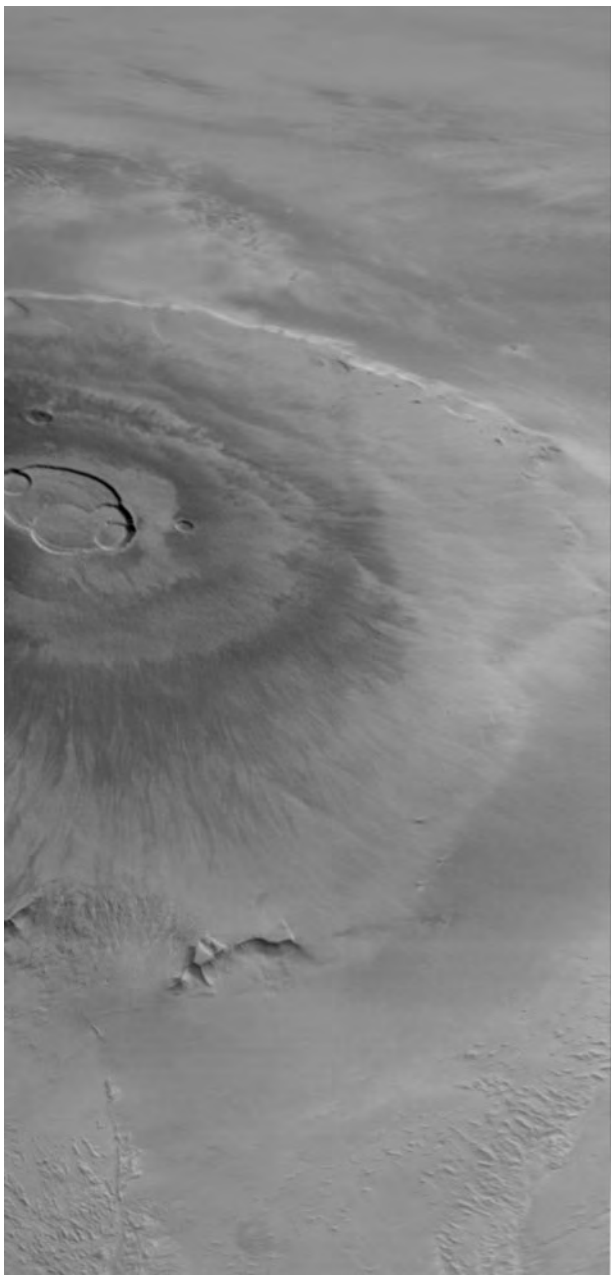


Figure 2. A Mars Global Surveyor/MOC image of Olympus Mons, the largest known volcano in the solar system. With a basal diameter of 600 km and a height of 24 km it is much larger than any terrestrial equivalent. (Courtesy of NASA/Malin Space Science Systems.)

like Venus. This means we can use standard photographic and multispectral remote sensing data to observe the surface and deduce its volcanic past. Mars has had an interesting history in volcanic terms. Although half the diameter of the Earth, it is home to the largest volcano in the solar system, Olympus Mons (figure 2). With a basal diameter of 600 km, and rising more than 24 km above its surroundings, it is many times larger than the largest volcano on the Earth, Mauna Loa, Hawaii (120 km diameter, 9 km above the sea-floor). The type of volcanism forming Olympus Mons and Mauna Loa appears to be the same. They are both shield volcanoes with Olympus Mons likely to be formed by the non-explosive eruption of basalts over a long period of time, as in the case of Mauna Loa. The main difference between them is the size. This is most likely due to the fact that Mars is a one-plate planet and does not experience motion of plates as the Earth does. The size of the Hawaiian volcanoes is limited by the fact that the plate on which they stand constantly moves. Once the volcano is no longer connected to the magma source, activity stops and a new volcano forms in the chain. On Mars this cannot happen and the volcano remains over the magma source and hence is active for as long as magma is available. The outermost flows on Olympus Mons appear to be very young, judging by the lack of large impact craters on its surface. However, while the youngest flows may have erupted only a few hundred million years ago, the volcano as a whole is likely to be much older. Olympus Mons is located just off the edge of a region known as the Tharsis bulge, a huge area uplifted relative to the surrounding terrain and which contains three very large shield volcanoes. Another volcanic area of uplift on Mars is the Elysium region, again containing large shield volcanoes. These regions and the long lava flows seen across the northern hemisphere indicate that volcanism has played an important role in the evolution of Mars.

But do any of these planets have active volcanism today? We cannot be sure. We are reasonably confident that no volcanism occurs on the Moon today. There is no image-based evidence for volcanism occurring on Venus or Mars, but Venus does seem to have an excess of SO_2 in its atmosphere. Chemical reactions with the surface should quickly deplete the atmosphere of sulfur, so there must be a continual replenishment of this element into the atmosphere. The most likely explanation is volcanism, but we cannot be certain until we get direct evidence of eruptions taking place there. The only other body where we are certain that present day volcanism is occurring is on Io, the third largest satellite of Jupiter. The tidal forces from Jupiter and the surrounding satellites maintain a high temperature within Io, which in turn drives the volcanism we see today (figure 3). Evidence for active volcanism on Io was first found by one of the Voyager craft as it flew past the Jovian system. A large plume of material was seen on the limb of the satellite being erupted from a volcano. Subsequent analysis of Voyager data showed Io to have many volcanoes across its surface and it is the most volcanically active body in the solar system today. The Galileo mission to Jupiter has returned stunning pictures of eruptions in action, revealing changes visible over only a few days. But, how does the volcanism on Io compare with that of the Earth? Firstly, Io's eruptions are likely to be more similar to terrestrial geysers. In addition, the type of volcanism occurring has become

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

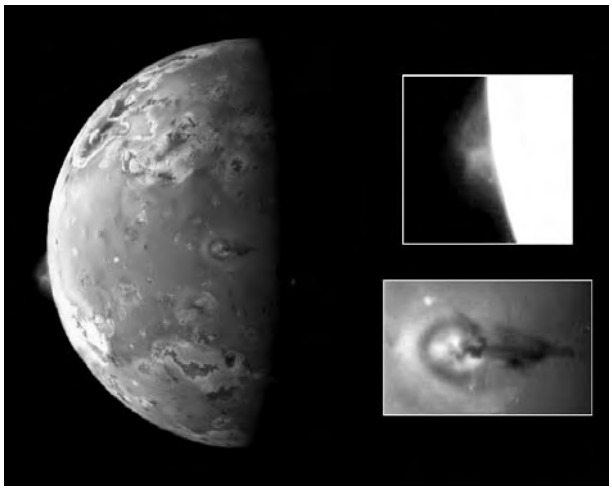


Figure 3. Images from the Galileo Orbiter showing volcanic activity on Io. The plume seen on the limb of Io in the main image (and enhanced in the top right insert) is 140 km high. The bottom right insert shows another area of activity with a 75 km high plume erupting from the surface. The shadow of this plume is clearly seen to the right of the main center of activity. (Courtesy of NASA JPL/Caltech.)

a matter of debate. On Earth, most of the volcanism is silicate based with some sulfur gases being released when the magma rises and lava erupts. For some time it was suggested that Io had sulfur based eruptions, due to the orange and brown colors of Io. These are similar to allotropes of sulfur with the different colors representing sulfur at various temperatures. In addition, spectroscopy of the surface and tenuous Ionian atmosphere revealed the presence of substantial amounts of sulfur being present. However, the density of Io ($\sim$3.5 g cm^{-3}) is more consistent with silicate rocks, and the crust has to be made of a strong material such as silicate in order to support the mountains of Io, some of which reach heights of 5–10 km. Finally, some of the 'hotspots' on Io reach temperatures of 1000 K or more, far too hot for sulfur. So it seems that Io, like the Earth, does have silicate volcanism, but also has an added contribution from sulfur (see also VOLCANISM IN THE SOLAR SYSTEM).

Tectonism

The Earth is today a tectonically active world, and has likely always been that way. The other planets and satellites in the solar system also show evidence of TECTONIC activity, but not like that of the Earth (figure 4). Without plate tectonics, the Earth would have been a very different world. Neither Mercury nor Venus nor Mars show any evidence for having multiple plates, and hence plate tectonics. On Mars, this means that huge volcanoes such as Olympus Mons could form. Mars' huge canyon, Valles Marineris, is likely to be at least partly formed by tectonic activity and forms a scar approximately 4000 km long. On Venus, global tectonism could have provided the vents through which the magma rose to completely resurface the planet. If true, evidence for this is now buried, but Venus still shows complex tectonic forms covering much of the planet. Tectonism is not limited just to the larger bodies. Smaller worlds like the Moon, Mercury and some satellites also show evidence for tectonism. Mercury in particular is another world where global events may have shaped the surface. Long lobate scarps are to be found all over the photographed hemisphere. These are contractional features, but have no associated extensional tectonism to compensate. This implies that at some point Mercury's crust must have undergone a global contraction by a few kilometers to produce the features seen. Looking just at the four terrestrial planets, we get very different views of how tectonism has shaped their evolution and provides a convincing argument that the Earth would indeed have been very different without the action of plate tectonics.

Comparison of planetary atmospheres

The PLANETARY ATMOSPHERES in our solar system should be able to tell us much about the evolution of the planets themselves. After all, each atmosphere came from the same nebula that made up the planets and all their atmospheres, so we would think that to a first approximation all the atmospheres would be of a similar composition. It is now obvious that they are not, and this is likely to be the result of the differences in the formation and evolution through time of the host planet. There is an obvious dichotomy between atmospheres in different parts of the solar system. Close in to the Sun we have the terrestrial planets with their sometimes tenuous, sometimes thick atmospheres enveloping the solid rocky body beneath. Once we get past the orbits of Mars and the main asteroid belt, we come to the gas giants, which are composed entirely of gases and have no solid surface beneath. The difference between these two kinds of world are quite dramatic, with the gas giants showing spectacular 'weather' systems and storms which, in the case of Jupiter, are larger than the planet Earth itself. However, we can learn most about our own planet by looking at the atmospheres of the terrestrial planets. The gas giants have undergone different evolutionary processes due to their nature, whereas the terrestrial planets share the same region of space and will have had initial starting conditions closer to that of the Earth.

Venus is often referred to as 'Earth's twin' due to its similar size and density. We would perhaps expect its atmosphere also to be similar to Earth, but we now know that Venus' atmosphere is very different. The atmosphere of Venus is composed mainly of CO_2, whereas the Earth has an atmosphere dominated by N_2 and O_2. Also, the atmosphere of Venus is very much denser than the Earth's. It has a surface pressure 90 times greater than that at our sea-level, and contains 100 times the amount of gas as our atmosphere. If Venus' atmosphere was once similar to our own, something quite dramatic happened to change it in this way. Venus is known to have a runaway GREENHOUSE EFFECT that has raised its temperatures to far in excess of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

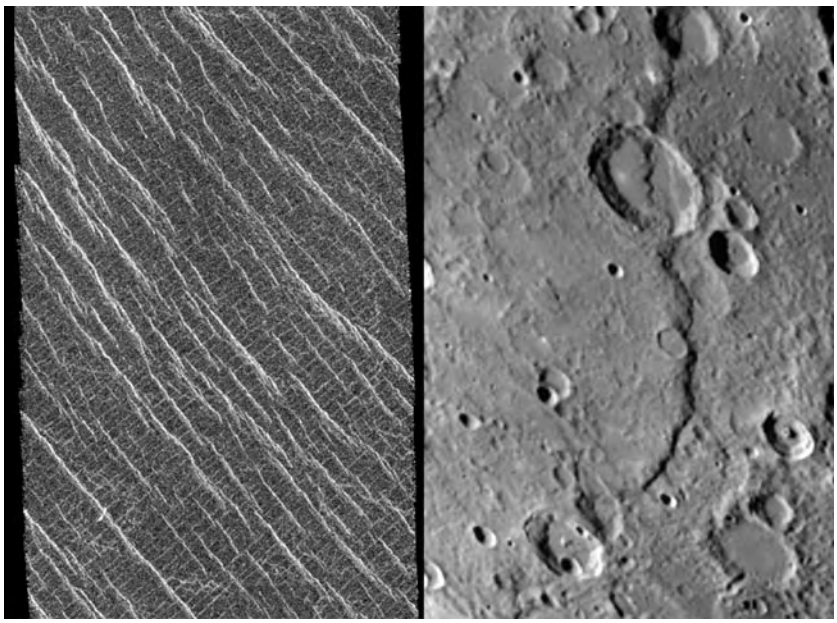


Figure 4. Two examples of tectonism on different planets. Left: lineated plains in Lakshmi Region on Venus. It is not known whether the two sets of intersecting lineations are faults or fractures, but this kind of terrain is unique to Venus. The image is 37 km wide, and 80 km long. (Courtesy of NASA JPL/Caltech.) Right: Mariner 10 image of part of the surface of Mercury. The long fault running through the center of the image is called Discovery scarp and is 350 km long. The large crater it intersects is 55 km in diameter and the maximum height of the scarp just below this crater is 3 km. (Courtesy of NASA JPL/Caltech.)

those it would have if it had an Earth-like atmosphere. In fact, this effect raises the surface temperature on Venus to approximately 750 K—hotter even than the surface of Mercury, which is closer to the Sun. The abundant CO_2 in Venus' atmosphere acts as an excellent greenhouse gas. One reason why the Earth's atmosphere has so little CO_2 compared to Venus is that CO_2 is easily dissolved into the Earth's oceans where it ends up being converted into carbonates, such as limestone. If you look at the total amount of CO_2 Earth and Venus have, including that locked into carbonates, we find that the amount on each planet is rather similar, but distributed in a very different way.

The water content in the atmospheres of Earth, Venus and Mars are all very low. Overall, the total volatile content as a percentage of the mass of the planet is not very high either. However, water plays an important role where Earth is concerned—it supports the life on the planet. We also know that water played a role in Mars' past, due to the numerous features on the surface that must have been cut by the action of running water or glacial action. For this to be the case though, the atmosphere of Mars must have been very different to what it is today. Under today's atmospheric conditions, water is not stable at the surface of Mars. There is a small amount to be found in the north polar region as water ice, and it is likely that some water ice exists beneath the surface as a permafrost layer. As mentioned earlier, impact craters can be used to trace the past presence of ice beneath the surface, but it is difficult to say whether that ice layer still exists in those regions today. It is clear, however, that Mars lost a lot of its atmosphere some time in the past, and its surface water along with it.

Why did the atmospheres of Mars and Venus change so dramatically over time? While this is still a matter of debate, Mars may have lost its atmosphere for a variety of reasons, including stripping of the atmosphere due to the solar wind, or erosion caused by impacts ejecting atmospheric gas into space. Venus, on the other hand, may have had a more hospitable atmosphere in the early solar system while the Sun was less luminous. As the Sun output more energy, Venus heated up, starting its runaway greenhouse effect and leaving the planet as it is today. Could this happen to the Earth? Conceivably, yes, but it would require a substantial heating of the Earth's atmosphere and the appropriate levels of greenhouse gases to cause it. Venus does, however, give us an insight into what Earth may have been like if we were formed as close to the Sun as our twin.

The rubble of the solar system: asteroids, comets and other small bodies

While most of our comparative planetology concentrates on the planets and satellites of the solar system, we must not forget the smallest bodies, as they can tell us much about the characteristics of the young solar system. METEORITES are pieces of ASTEROIDS and other bodies, broken apart from their parent body, probably by an impact and sent on a course that sees it captured by the Earth's gravity and fall to the surface. These give us the chance to study other bodies without the expense of sending a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spacecraft to it, although with meteorites, there is the problem of coping with terrestrial contamination in the results. Some meteorites have been found to contain material that dates back to the start of the solar system, and hence we can get an idea of the initial makeup of the solar nebula. Meteorites give us clues as to the nature of asteroids too. Asteroids are larger bodies, most of which reside in a belt between the orbit of Mars and Jupiter. The compositions of the asteroids vary, as evidenced by meteoritic compositions. In fact, there are a rare few with properties which indicate that the parent body underwent melting. The sizes of the asteroids today are too small for this to have happened, and so it is likely that the asteroids we see were once part of a larger body or bodies. The gravitational perturbations of Jupiter probably prevented this body from reaching planetary proportions.

Further out in the solar system, beyond the orbit of Pluto are the KUIPER BELT objects. These again are large lumps of rock, left over from the formation of the solar system. Although we do not have any samples of these objects (or at least we do not know if we have a sample in a meteorite on Earth) scientists are discovering more and more Kuiper belt members. Being able to get a sample of one of these, or at least to be able to determine its composition remotely will tell us much about what we were made of. This region of space may also be the home of some of the comets we see on a regular basis. COMETS are made mostly of ice and rock. They can make spectacular displays when close to the Earth, but more importantly it may have been the comets that were responsible for the delivery of a substantial amount of the volatiles to be found on Earth today. By studying the complete composition of these bodies, we may get a better understanding of what the comets delivered to us.

There is much to learn by comparing one world with another. Some may be able to tell us more about our own planet and what will or may have happened to it. In other cases, we can learn more about an individual process that occurs on many bodies. Comparative planetology was not possible a few decades ago, and we are now in a phase where our knowledge of other worlds is expanding at a fantastic rate. The more we learn about the differences between planets and why those differences occur, the more we will understand about the evolution of the entire solar system and perhaps even that of the planetary systems we are discovering around other stars (EXOPLANETS).

Bibliography

Beatty J K, Peterson C C and Chaikin A (ed) 1999 *The New Solar System* 4th edn (Cambridge: Cambridge University Press)

Lewis J S 1997 *Physics & Chemistry of the Solar System* revised edn (California: Academic)

McSween H Y Jr 1999 *Meteorites and their Parent Planets* 2nd edn (Cambridge: Cambridge University Press)

Spudis P D 1996 *The Once and Future Moon* (Washington: Smithsonian Institution Press)

Sarah K Dunkin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Planisphere

A simple device that displays the stars visible from a particular place at a particular time and date. It consists of a disk-shaped chart of all of that part of the celestial sphere that can potentially be seen from a particular latitude, overlaid by a rotatable disk-shaped mask in which an aperture has been cut. Using graduated scales around the perimeters of the two disks, the mask may be rotated so that the aperture reveals the stars that are visible on any selected date or time. The principle of the planisphere is broadly similar to that of the astrolabe, an ancient astronomical and navigational device.

See also: astrolabe, celestial sphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plaskett, John Stanley (1865–1941)

Astronomer and spectroscopist, born in Woodstock, Ontario, worked as a mechanic. He was a member of the original staff of the Dominion Observatory, Ottawa, where he became director and designed and constructed the 72 in telescope there. He worked on stellar radial velocities and spectroscopic binaries, and studied spectra of O- and B-type stars, and the interstellar calcium that showed on their spectra, mapping galactic structure. *Plaskett's star* HD47129 is the most massive known star, in fact a pair of O-type stars in a binary system, each 55 solar masses, and close to the maximum permissible for a star, beyond which radiation pressure disrupts the star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plaskett's Star

A spectroscopic binary in Monoceros, discovered in 1922 by the Canadian astronomer J S Plaskett (1865–1941) using a spectrograph he had designed specifically for the measurement of stellar radial velocities, attached to the 72 inch reflector of the Dominion Astrophysical Observatory in Victoria, British Columbia. The telescope was also largely of Plaskett's design, and is now known as the 'Plaskett telescope'.

Plaskett's Star, also designated V640 Mon, is in the region between *Betelgeuse* and *Procyon*, about 10° south of the star γ Geminorum. It is a member of a major cluster of hot, young stars associated with the Rosette Nebula, NGC 2244. Its apparent magnitude varies between 6.04 and 6.08 in a period of 14.396 days. The system, believed to be one of the most massive pairs of stars in the Galaxy, comprises two blue supergiants of spectral class O8e which may be separated by as little as 80 million kilometers. There must therefore be a turbulent mass of gas surrounding, and flowing between, the components. The latest determinations indicate that the mass of each of the components is of the order of 50 solar masses.

Sir James Jeans (1877–1946) described Plaskett's Star as 'the most massive and absolutely the brightest star whose elements are known with fair certainty. Temperature 28 000 °C. Every square centimeter of surface emits sufficient energy to run a locomotive at full speed for millions of years.'

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plasma

An almost completely ionized gas that contains equal numbers of electrons and positive ions moving freely and independently of each other. Although a plasma is electrically neutral, because it contains equal numbers of positively and negatively charged particles, it is highly conductive. Plasmas, which normally have very high temperatures, are found in a wide variety of astrophysical contexts, for example deep inside stars, where matter is very highly ionized, or in the high-temperature outer atmospheres of stars (coronas). The solar corona and the solar wind are both examples of plasmas that consist predominantly of electrons and protons (hydrogen nuclei).

Magnetic fields and plasmas interact in interesting and important ways. Because a charged particle moving in a magnetic field is subject to an accelerating force acting at right angles to the direction in which it is traveling, it is compelled to trace out a circle around the field line. If it also has a component of motion along the direction of the field, the combination of the two motions will cause it to follow a helical path along the direction of the magnetic field and prevent it traveling for any distance at right angles to the field. Therefore, provided the density is not so high that frequent collisions between particles break the ties between particles and field lines and thereby enable charged particles to diffuse across the direction of the magnetic field, plasma can flow along the direction of a magnetic field but not at right angles to it. In these circumstances, magnetic field and plasma are closely tied together, and the field is said to be 'frozen in' to the plasma. Where the pressure of the plasma exceeds the pressure of the magnetic field, moving plasma pushes the field lines around; conversely, where the magnetic pressure exceeds the plasma pressure, moving magnetic field structures push the plasma around. In the case of the Sun, closed magnetic structures (field lines that curve out from, and back into, its surface) confine the hot coronal plasma and prevent it escaping into space, whereas open field lines (which lead out into interplanetary space) allow plasma to flow out into the solar wind. The interplay between magnetic fields and plasmas underlies practically all aspects of solar activity.

See also: ion, ionization, magnetosphere of Earth, solar wind, Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plate-Measuring Machine

A device that enables the positions of images on a photographic plate to be measured to a high degree of accuracy. The plate is mounted on a flat table that may be moved in two perpendicular directions ('x' and 'y'). In traditional machines the plate is moved manually in the x- and y-directions by means of two fine screws until the star image of interest is centered on crosswires in a microscope eyepiece. The amount of movement is registered on a finely graduated drum or vernier, typically to an accuracy of about a micron (μm). The plate may then be reversed (and, if necessary, inverted) and re-measured to minimize positional errors. This is a tedious and time-consuming process. Automatic plate-measuring machines have speeded up the process by many orders of magnitude. Usually, in a machine of this kind, the plate is mounted on a table that moves slowly, or in very small steps, in one direction while a spot of light (produced, for example by a cathode ray tube or a laser) scans rapidly across the plate in the other, images being identified by measuring the amount of light transmitted through the plate. In addition to measuring position, some machines can also measure brightness and distinguish between different kinds of objects (for example stars and galaxies). The machines are usually computer controlled and the computer carries out much of the data analysis. The SuperCOSMOS machine at the Royal Observatory, Edinburgh, which uses a linear CCD array to scan successive strips of a plate, can scan a complete Schmidt plate (which covers an area of sky 6° square and typically contains about a million star images) in two hours.

See also: celestial coordinates, Schmidt telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plato (428–347 BC)

Philosopher, real name said to be Aristocles, with 'Plato' a nickname ('broad') derived either from the breadth of his knowledge, shoulders, or forehead, born in Athens, Greece, taught by Socrates, and travelled in Egypt, Sicily and Italy. Back in Athens, he founded, in 387 BC, a school of learning called the Academy, devoted to research and instruction in philosophy and the sciences, particularly to the development of mathematical models of natural phenomena. The Platonic solids are the cube, tetrahedron, octahedron and icosahedron, envisaged as the shapes of the atoms of the four elements, earth, fire, air and water, with the fifth Platonic solid, the dodecahedron, a model for the whole universe (see KEPLER). Plato envisaged the universe as a series of shells of crystalline spheres on which the stars, planets, Sun and Moon move round the Earth. He believed that the Moon shines by reflected sunlight and that astronomers were the wisest people. Plato's Academy was deemed a pagan establishment and closed in AD 529 by the Christian Emperor Justinian, but at 900 years of age it was the longest surviving university.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plough

A well-known asterism, being the most readily recognizable part of the rather disjointed constellation of Ursa Major; it comprises the seven stars α, β, γ, δ, ε, ζ and η UMa. The arrangement resembles the outline of the old-style, hand-steered plough drawn by a horse or ox. The stars α and β UMa are also termed 'the Pointers', as they indicate of the direction of Polaris, the pole star. The star ζ Uma (Mizar), is a well-known binary system which has a fourth-magnitude naked-eye companion, Alcor (80 UMa).

Five of the stars (β, γ, δ, ε and ζ UMa) are members of the Ursa Major open cluster and have very similar proper motions. As the proper motions of the other two stars (α and η UMa) are in an opposing direction, the present familiar appearance of the Plough is temporary; thousands of years ago its shape was entirely different, and will be different again thousands of years hence.

See also: Mizar and Alcor.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Plutino

Any of the Kuiper Belt objects in or near the 3:2 mean motion orbital resonance with Neptune. This means that they complete two orbits around the Sun in the time it takes Neptune to complete three orbits. The 3:2 resonance acts to stabilize the Plutinos against gravitational perturbations by Neptune.The same resonance is also occupied by Pluto. 'Plutino' means 'little Pluto'. Approximately 35% of the known trans-Neptunian objects are Plutinos, and there are approximately 25 000 larger than 100 km diameter. Pluto is considered by some astronomers as by far the largest Plutino.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pluto and Charon

Pluto is an unusual planet in many respects. It is the smallest and the most distant from the Sun of all nine planets. Its orbit is unusually eccentric and inclined to the ecliptic. It is accompanied in this orbit by a satellite, Charon, whose radius is half of Pluto's. Its surface is among the coldest and, except for IAPETUS, the most contrasted of all PLANETARY SURFACES. Although Pluto is the only planet never visited by a space probe (but see PLUTO–KUIPER EXPRESS), a series of breakthrough discoveries made since about 1975 have now provided us with a relatively detailed picture of the Pluto–Charon system, which is reviewed here.

Pluto's heliocentric orbit

Although Pluto was discovered in 1930, prediscovery photographs establish its position back to 1914. Pluto has thus been observed for about 85 years, i.e. slightly over one third of its orbital period of 248 years. Pluto's mean distance from the Sun is about 40 Astronomical Units (semi-major axis of the orbit, a = 39.54 AU), i.e. about 6 billion kilometers. However, due to its strongly elliptical orbit (eccentricity, $e = 0.249$), its distance from the Sun varies by 20 AU between PERIHELION (q = 29.7 AU) and APHELION (Q = 49.85 AU). Pluto's orbit is therefore 'Neptune-crossing' in the sense that during approximately two decades within its orbital period, it can get closer to the Sun than NEPTUNE, whose perihelion distance is 29.8 AU. However, due to its orbit high-inclination ($i = 17.1^\circ$), Pluto, unlike any other planet, makes large excursions of 8 AU above and 13 AU below the plane of the ecliptic and, despite the fact that the orbits of Pluto and Neptune appear to cross when projected onto the ecliptic plane, Pluto never gets very close to Neptune (never closer than 17 AU). Pluto in fact makes closer (12 AU) and more frequency approaches to URANUS than to Neptune. The reason for this interesting situation lies in the so-called 3:2 mean motion resonance between Pluto and Neptune. Over a period of 495 years, Pluto orbits the Sun twice while Neptune orbits three times (figure 1). As a consequence, the relative orbital phases of Pluto and Neptune recur periodically, so that the perturbative forces of Neptune on Pluto act in nearly the same phase at each repetition. Numerical integrations of the motion of the five outer planets over time scales of more than a million years indicate that Pluto's conjunctions with Neptune (i.e. the configurations when the two planets have the same heliocentric longitude) occur when Pluto is near aphelion. More specifically, calculations show that the 'resonance argument', defined as $\Phi = 3\lambda_P - 2\lambda_N - \varpi$, where λ_P and λ_N are the longitudes of Pluto and Neptune, and ϖ is the longitude of Pluto's perihelion, varies ('librates') about a mean value of 180° with an amplitude of 82° over a libration period of 19 912 years. Therefore, the 3:2 resonance has a strong stabilizing influence on Pluto's orbit. In addition, Pluto's perihelion and aphelion cannot occur at any position on the orbit and, in particular, they never cross Neptune's orbital plane, further helping to keep the two planets apart. Finally, it is important to note that most modern observations of Pluto occurred near the last perihelion (1989). Thus, the vision of the planet (in particular the surface and atmospheric properties) described below essentially refers to perihelion and would be drastically different at aphelion.

Charon's orbit and Pluto–Charon self-rotation

Charon, discovered in 1978, orbits about Pluto at a mean distance of 19 640 km (corresponding to about 17 Pluto radii, or an apparent separation on the sky of 0.9″ at 30 AU). (Because the masses of the two objects are not very different (see below), it is more correct to state that Pluto and Charon orbit about their common center-of-mass.) Charon is the only satellite of Pluto known to date, with a visible magnitude of $m_v = 17.25$ at opposition. Searches for additional satellites, motivated by the fact that the volume of space around Pluto that is stable against solar perturbation (i.e. in which other satellites could exist) extends to about 100 Charon orbit radii, have concluded to the absence of any object brighter than $m_v = 20.6$ in the region from 6 to 10 arcsec and $m_v = 22.6$ beyond that. The orbital period of the Pluto–Charon system is 6.387 223±0.000 017 days. The rotational period of Pluto, as determined from its photometric lightcurve (see below), is 6.387 260 ± 0.000 070 days, The two numbers are consistent within error bars, implying that Pluto's spin is tidally locked to Charon's orbital motion. Charon's rotational period is not well constrained but consistent with SYNCHRONOUS ROTATION. In addition, theoretical arguments indicate that Charon's rotational motion must also be locked to the orbital motion, because the timescale over which the satellite's self-rotation is slowed by frictional effects (called the tidal damping time) is much shorter than for Pluto. Therefore, just as the Moon always turns the same face to the Earth, each body (not only the satellite) keeps the same hemisphere turned towards the other. In addition, as seen from the surface of one of the two bodies, the other body is either invisible or always at the same position in the sky. This situation of mutual synchronism is the stable end of tidal evolution and Pluto–Charon is the only known example of such a case.

The orientation of Charon's orbit with respect to Pluto's orbital motion gives rise to so-called 'mutual events'. Every 124 years (i.e. every half-orbit of Pluto), and for a period of approximately 5 years, Charon's orbit is seen edge-on from the Earth (and the Sun), allowing observers to view Pluto and Charon passing directly in front of and behind one another. By unusual chance, this 'mutual events season' started in 1985, i.e. only seven years after the discovery of Charon. The scientific return of the mutual events was enormous, not only because they allowed an improved knowledge of Charon's orbit (thereby providing the total mass of the Pluto–Charon system by virtue of Kepler's third law) but more importantly because they provided the first resolved view of the two bodies, with estimates for their individual

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

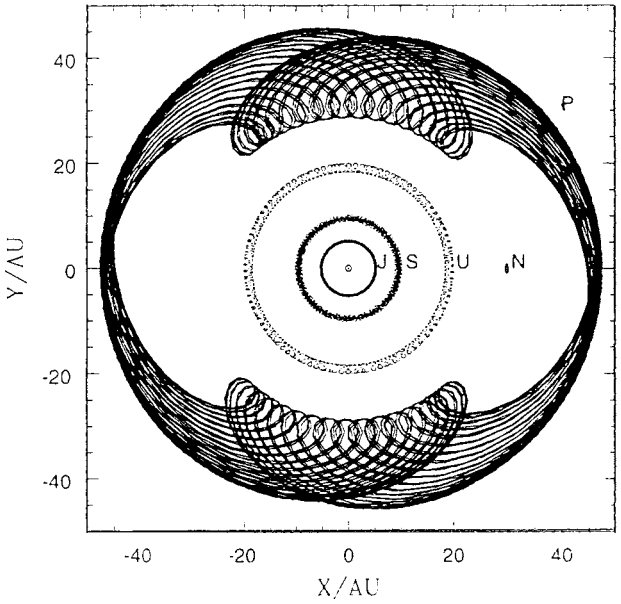


Figure 1. The orbits of the outer planets for 40 000 years in a reference frame corotating with the mean motion of Neptune. In this reference frame, two revolutions of Pluto trace a complete loop in 495 years. It can be seen that Pluto's perihelion librates 90° away from Neptune (with a period of about 19 900 years). From Malhotra and Williams (1997).

diameters (thus the average density of the system), their surface composition and their albedo distribution (see below). The best determination of Charon's orbit now relies on direct HST imaging. These observations yield a semi-major axis of 19 636 ± 8 km and a small but significantly non-zero eccentricity of 0.0076 for Charon's orbit. This result is surprising given that tidal effects can damp out such an orbital eccentricity on a time scale of 1-10 million years, and may suggest that the system was disturbed relatively recently. The orientation of Charon's orbit is such that its orbital angular momentum vector is inclined by 119.6° with respect to Pluto's orbital plane. Again for tidal effect reasons, Pluto's and Charon's polar axes must be parallel to Charon's angular momentum. Thus, with an obliquity of ∼120°, the latitude of the subsolar point varies between −60° and +60° over a Pluto year. This excursion, which dictates the annual variation of the insolation for a given point on Pluto's surface, is the largest experienced by any planet except Uranus. Due to the torques exerted by the Sun over the Pluto–Charon system, Pluto's equatorial plane precesses about the plane of its heliocentric orbit, so that Pluto's obliquity varies between 103° and 128° with a period of about 3 million years.

Bulk properties: radius, mass, density

Determinations of Pluto and Charon radii are based primarily on mutual event photometry, stellar occultations, and recently, from direct HST imaging. Radius values from mutual events and direct imaging tend to converge towards radii of about 1160 km for Pluto and 600 km for Charon, but the absolute accuracy, limited by the uncertain effects of limb darkening, is probably not better than 10-20 km. Stellar OCCULTATIONS by planets are a powerful means for measuring diameters (and possibly oblatenesses) but in Pluto's case, the small size and slow motion of the planet make these events rare and difficult to predict. To date, only occultation (June 1988) has provided good-quality data. Unfortunately, the uncertain structure of Pluto's atmosphere (see below) makes the radius determination ambiguous. Nominal values are 1180-1195 km which, at face value, are inconsistent with the above numbers; however, whether the occultation has probed Pluto's atmosphere down to the surface is actually not certain (see below).

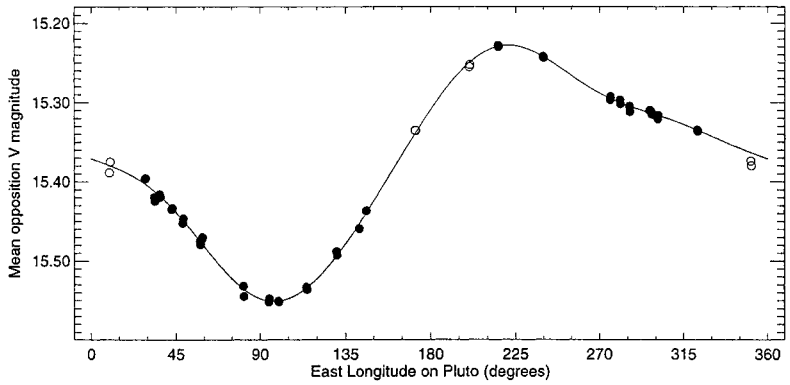
The total mass of the Pluto–Charon system is straightforwardly determined from the orbit parameters of Charon to be equal to 0.002 433 terrestrial masses. (Incidentally, this value is 1/4000 of the mass deduced by Lowell in 1915 to account for the then believed perturbations in the orbit of Uranus. Therefore although Pluto was indeed discovered somewhere near the position predicted by Lowell, this appears to be a mere coincidence.) The individual masses of Pluto and Charon are much harder to determine, as this requires examination of the motion of Pluto with respect to the barycenter of the system, which is about an order of magnitude smaller than Charon's. Nonetheless, and in spite of remaining divergences, the current 'consensus' value for the Charon/Pluto mass ratio is about 0.119. This gives individual masses of 1.30×10^{25}g for Pluto and 0.15×10^{25}g for Charon. Adopting radii of 1151-1178 km for Pluto and 591-628 km for Charon, based on the various modern measurements, the average mass density of the Pluto–Charon system is 1.87-2.03 g cm^{-3}; individual densities are 1.92-2.06 g cm^{-3} for Pluto and 1.51-1.81 g cm^{-3} for Charon.

Surface properties

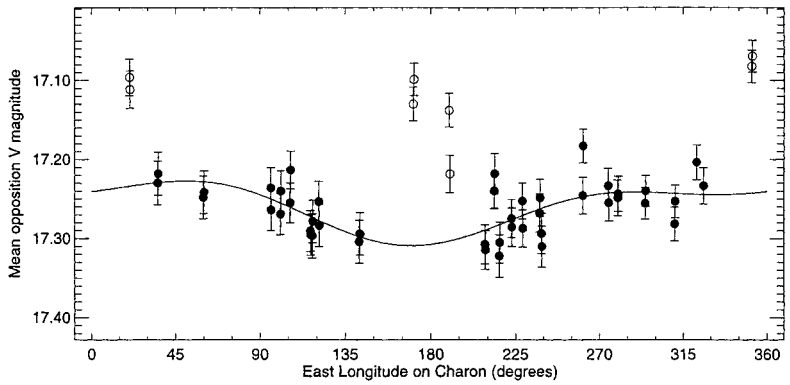
Visual appearance

One of the earliest findings (1955) about Pluto is that its brightness varies by about ±0.15 magnitudes (about ±15%) around a mean magnitude of 15.9 with a period of 6.3872 days. At that time, this result, called the 'Pluto visible lightcurve', actually pertained to the 'Pluto+undiscovered Charon' system. HST observations, which separate Pluto from Charon, prove that the variation is essentially entirely due to Pluto, with Charon having a very weakly marked lightcurve of ±0.04 magnitude (figure 2). The existence of the lightcurve indicates that Pluto's surface is not uniformly bright, but has bright and dark regions. Observations of the mutual events and now direct imaging, in combination with lightcurve information, indicate the gross structure of Pluto's surface. It includes a large bright southern polar cap, a broad dark equatorial band, and a large region of intermediate brightness in the Northern hemisphere (figure 3). The equatorial and mid-latitude regions appear

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(a)



(b)

Figure 2. Pluto and Charon separate lightcurves. These curves give the visual magnitude (for a mean opposition) as a function of longitude on the planet and satellite. From Buie *et al* (1997).

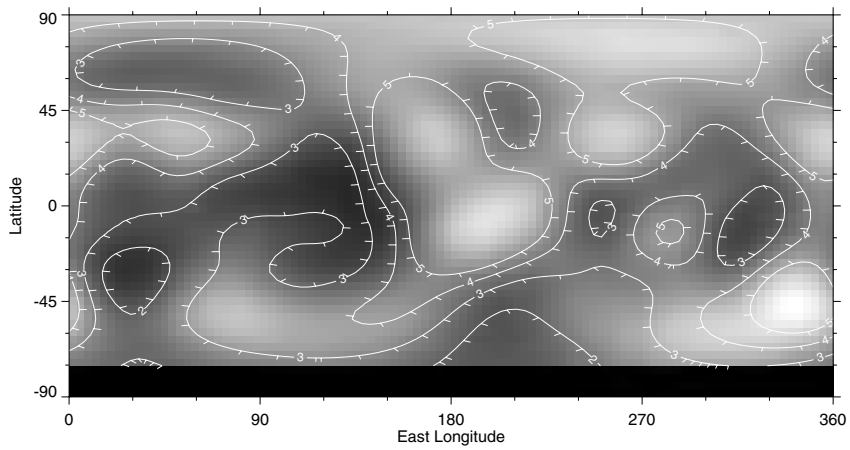


Figure 3. Surface appearance of Pluto, from HST imaging. Albedo contrasts reach values of 6:1. From Stern *et al* (1997).

very variegated. Brightness contrast ratios over the planet can reach values of 6:1, nearly as high as on Iapetus. Brightness variations probably result from the effect of volatile condensation/sublimation cycles and transport over very many seasons (see below). Charon's surface is not so well known but does not seem to show any strong spatial variations. Its average geometric albedo is 0.36-0.39, substantially less than the average Pluto value (0.58), suggesting that its surface is not as frost-rich.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

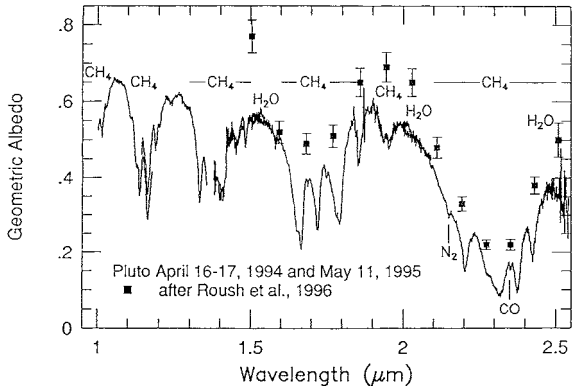


Figure 4. The near-infrared spectrum of Pluto at 1-2.5 μm, showing absorption bands due to CH_4, N_2, CO and tentatively H_2O. From Cruikshank *et al* (1997).

Composition

Spectroscopic studies at 0.7-2.5 μm indicate that Pluto's surface is dominantly covered by volatile ices. Ices of N_2, CH_4, CO and tentatively H_2O have been detected (figure 4). N_2 appears to be the dominant species, with CO and CH_4 ices constituting less than 2% by mass of the surface materials. In the case of CH_4, there appear to be at least two components: one corresponds to CH_4 ice intimately mixed at small concentrations with N_2 ice (e.g. N_2:CH_4 in ratios 100:1), and the other corresponds to a fraction of the surface covered by a thick layer of pure CH_4. Some models also suggest that the N_2:CH_4 layer is actually overlain by a thin CH_4-enriched layer. The methane bands appear to vary in intensity with the phase, with generally weaker bands near the minimum of the visible lightcurve (i.e. when the largest fraction of dark regions are in view). The interpretation of this variation is, however, not clear. Variations in methane depth can be due, for example, to variations in the spatial coverage of the CH_4 ice deposits, in the abundance of CH_4 in the N_2-CH_4-CO mixture, or in the size of the icy grains. It also appears that the overall spectrum of Pluto at visible wavelengths is somewhat 'red', i.e. that the reflectance of the surface increases with the wavelength between 0.4 and 1 μm. This may be related to a compositional effect, but the reddening agent remains to be identified. On Charon, H_2O ice has been clearly identified. It dominates the infrared spectrum at all longitudes. The visible spectrum of Charon, unlike Pluto's, is spectrally neutral.

Temperature

Pluto's and Charon's surface temperatures remain uncertain. At 30 AU from the Sun, the maximum temperature of any point of their surfaces is 72 K (the value corresponding to instantaneous equilibrium with the solar insolation). Observationally, the surface temperature of a planet with no atmosphere or a tenuous atmosphere is mostly determined by the thermal flux emitted by the object. In the Pluto–Charon system case, measurements have been performed at far-infrared (60-200 μm) and millimeter/submillimeter (0.8-1.3 mm) wavelengths. Note that these measurements cannot separate Pluto from Charon, so they primarily sample Pluto, which contributes ~80% of the total flux. The two techniques give apparently discrepant results, with an average surface temperature on the visible disk (i.e. on the dayside) of 47-58 K from the far-IR data and 30-44 K from the mm/submm. The far-IR values are thought to be the most representative of the true surface temperature, because the flux emitted at the longest wavelengths can be significantly lowered by non-unity surface emissivities, leading to an apparently lower temperature. However, studies of the shape of the N_2 absorption band at 2.15 μm on Pluto suggest that the N_2 ice temperature is about 40 ± 2 K. This suggests that Pluto's surface is not isothermal. This situation is also expected given the existence of Pluto's visible lightcurve (dark regions reflect less energy than bright regions and therefore must be warmer) and given the atmospheric composition (see below). A direct proof that Pluto's surface is not isothermal is provided by recent IR observations which indicate that the thermal flux emitted by the Pluto–Charon system varies with orbital longitude, with, as expected, more flux (i.e. a higher average temperature) when Pluto is darker. In summary, the current paradigm is that Pluto possesses cold (40 K) regions dominated by N_2 ice and warmer regions at temperatures well above 50 K. No information is available on Charon.

Pluto's atmosphere and interaction with the surface

The existence of an atmosphere around Pluto was demonstrated by the observation of a stellar occultation in June 1988 (figure 5). It is not surprising in the presence of volatile species on its surface, which can sustain an atmosphere by sublimation equilibrium. The observations indicates a pressure of a few μbar, but the full interpretation of the occultation data requires a knowledge of the main atmospheric chemical composition (which in fact was obtained later).

Chemical composition

Methane is the only gas to have been spectroscopically detected in Pluto's atmosphere. However, N_2 must dominate the composition of Pluto's atmosphere, because it is the most volatile of the various ices detected on Pluto's surface, and because its abundance at the surface is also the largest. Based on the relative abundance of ices, estimates of the surface temperature and direct abundance measurement of CH_4, Pluto's atmosphere can be estimated to consist primarily of ~99% of N_2, ~1% of CH_4 and ~0.1% of CO (the latter two numbers being uncertain by factors of a few).

Atmospheric structure and surface pressure

The occultation lightcurve (figure 5) is very unusual in that it exhibits a break in slope. This break occurs when the stellar rays cross Pluto's atmosphere at 1215 km from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

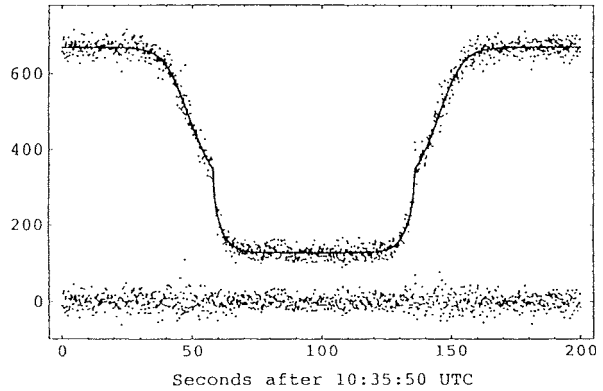


Figure 5. The occultation lightcurve for the stellar occultation by Pluto of 9 June 1988. The dots represent the stellar signal as the star is progressively extinguished by Pluto's atmosphere. Note the 'breaks' in the lightcurve. They occurred when the image of the star was 1215 km away from Pluto's center. Above this level, the lightcurve can be fitted by an isothermal atmosphere. Below this level, the data can be interpreted as due to a haze layer or to a strong temperature gradient. From Elliot and Young (1992).

Pluto's center, at a pressure level of 2.3 μbar. Above this level, Pluto's atmosphere is vertically isothermal with a temperature of about 102 K. Below this level, the shape of the occultation curve can be explained by either (i) a very strong temperature gradient (i.e. the temperature decreases towards the surface at a rate of 5-10 K km^{-1}) or (ii) an extinction layer due to a relatively thick haze, with an opacity of about 0.15. A temperature of ∼100 K for the upper atmosphere while the surface temperature is only 40-60 K is quantitatively explained by the fact that gaseous methane is an efficient heating agent, from absorption of solar insolation in its near-infrared bands. Models indicate that an atmospheric abundance of ∼1% for methane may also be capable of sustaining a strong temperature gradient near the surface. On the other hand, the 'haze' interpretation of the shape of the occultation curve is more difficult to justify physically. In addition, both interpretations imply a surface radius in the range 1180-1195 km, which is inconsistent with the values inferred from the mutual events and direct imaging. A way to resolve this discrepancy is to postulate that the occultation has not probed Pluto down to the surface itself but that the atmosphere extends 20-40 km below the strong thermal gradient layer. Specifically it has been proposed that Pluto possesses a 'hidden' 40 km deep troposphere. The uncertainty in the structure of the lower atmosphere reflects on the surface pressure. The occultation probes the atmosphere down to pressures of 3-5 μbar but the surface pressure could be significantly higher: for example with a 40 km troposphere at ∼40 K the surface pressure would be about 30 μbar.

Interaction with the surface

Pluto's atmosphere is primarily produced by the sublimation of the surface ices. At a temperature of 38-42 K, the surface pressure of nitrogen ice is 19- 159 μbar. To first order, this is in agreement with (although possibly somewhat higher than) the observed surface pressure. The equilibrium between frost and gas implies that the frost must be isothermal. Indeed, on a planet covered by frost, the frosts in the two hemispheres are at a uniform temperature if about half of the solar energy received by the summer hemisphere is transported to the winter hemisphere in the form of latent heat across the equator. This can be expressed by the following equation:

$$\tfrac{1}{2}S(1-A)\pi R^2 = 2\pi RL\Phi$$

where S is the solar constant, A is the surface albedo, R is the planetary radius, L is the latent heat per unit of mass and Φ is the mass flux across the equator (g cm^{-1}s^{-1}). Φ is equal to the product of the mass density (ρ) by the meridional flow velocity (υ) and the vertical height (d) of the flow. One can solve this equation for the flow velocity. Taking d = 1 km, one finds $\upsilon \sim 3$ m s^{-1}. Thus, modest meridional winds are able to equalize the N_2 frost temperature and surface pressure over the globe. The situation is similar to MARS but with the important difference that on Mars the temperature of the CO_2 polar caps is determined by the annually averaged insolation. On Pluto, the characteristic time of hemisphere transport is of the order of a few planetary days only. Thus the frosts are in instantaneous equilibrium with the diurnally averaged insolation. As a consequence, the seasonal variations in surface pressure and frost temperature are expected to be much larger than on Mars. Over the 21st century, when Pluto recedes from perihelion, the atmospheric pressure is expected to decrease by about two orders of magnitude. Seasonal variations are accompanied by a redistribution of volatiles, thereby modifying the visual appearance of the planet. Note that the above discussion applies as well to TRITON, which exhibits a 14 μbar N_2 atmosphere in equilibrium with N_2 frost at 38 K.

This simple picture does not work for methane, whose atmospheric abundance (∼1%) is orders of magnitude higher than expected from thermodynamical equilibrium. This methane enrichment is explained by postulating the existence either of a thin CH_4-enriched layer at the surface, or of thick and warmer isolated patches of pure CH_4 ice. Both scenarios could result from the known fact that in a CH_4-N_2 solid mixture, N_2 preferentially sublimes because of this higher volatility. Remarkably, both of them are supported by the most recent analyses of the near-IR spectra (see above). The presence of a thin CH_4-enriched layer is also expected to somewhat decrease the equilibrium atmospheric pressure of N_2.

Atmospheric escape

Reaching 100 K within a few kilometers from the surface, Pluto's atmosphere is much warmer than Triton's, where the 100 K upper atmosphere temperature is not attained before several hundreds of kilometers. As a consequence,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pluto's atmosphere is very extended, with an exobase (i.e. the level where the mean free path of the molecules is comparable to the scale height) near 4000 km altitude (750 km for Triton). Given the low gravity field, Pluto's atmosphere escapes very significantly, with an estimated escape flux of a few 10^{27} N_2 molecules/s (and about 10 times less of CH_4). In fact, Pluto's atmosphere is so weakly bound that escape occurs hydrodynamically (i.e. vertically 'blowing off') rather than thermally (i.e. from the fact that a fraction of the molecules has a velocity exceeding the escape velocity). Note, however, that Pluto's escape flux is expected to drastically decrease towards aphelion, not only because of the expected collapse in surface pressure, but, more importantly, because the atmospheric temperature and extent, controlled by the atmospheric abundance of methane, will strongly decrease with increasing heliocentric distance. Pluto's present averaged escape flux, if maintained over the age of the Solar System, represents a loss of a volatile (N_2, CH_4 . . .) ice layer at least several kilometers thick.

Possible internal structure

The internal structure of any planet is necessarily uncertain to some degree, as it escapes direct observation and is primarily based on models. The philosophy of structural modeling can be summarized as follows: a set of components (e.g. water ice, other volatile ices, rock) for the interior of the planet is selected based on observational (surface data) and theoretical (cosmogonic) considerations, an internal arrangement of these components is guessed (for example, is the planet differentiated or undifferentiated, i.e. are the ice and rock vertically statigraphed or homogeneously mixed?), and using the equation of state of the components and hydrostatic equilibrium, the equation of state is computed throughout the body. Parameters are adjusted until the model predicts correct values of the observables (mass, radius, quadrupolar moments of the gravity field, etc). In the case of Pluto and Charon, the internal models are particularly uncertain because only the individual masses and radii are known as constraints to the models, and still with a considerably uncertainty (see above). Nonetheless, internal models are useful because they allow the estimation of parameters such as the rock-to-ice ratio and the bulk chemistry, which are of key importance for understanding the origin of the bodies.

Constraints to the models

As discussed above, the most recently determined masses and radii provide values for the mass densities of 1.92-2.06 g cm^{-3} for Pluto and 1.51- 1.81 g cm^{-3} for Charon. This clearly indicates that Pluto and Charon are some mixtures of ice and rock. The dominant ice must be water ice because oxygen is the most abundant of the ice-forming elements (C, N, O). Given the temperature and pressure range in Pluto's and Charon interiors, H_2O ices I, II, III, V and IV may be present. Like Pluto's surface, Pluto's interior may contain other volatile ices. However, they are probably concentrated near Pluto's surface so that the present mantle structure and rheology must be dominated by water ice. The exact composition of the rock component is unknown but it is almost certainly at least partially hydrated. Finally, refractory organic compounds may represent an important fraction of Pluto and Charon interiors. This consideration is based on chemical evidence from COMET HALLEY in which CHON (carbon-hydrogen-oxygen-nitrogen) particles contained in a diversity of refractory organic materials represent as much as 10% of the mass.

Pluto's interior is probably differentiated. This stems from arguments based on Pluto's thermal history and the current presence of volatiles. Radiogenic heating release is more than sufficient to allow solid state flow of water ice. In addition, the accretional heating of Pluto may be important, especially considering the fact that the collision that probably formed Charon could have triggered water ice melting, followed by DIFFERENTIATION. Moreover, surface volatiles, as the source of Pluto's atmosphere, are subject to escape. Therefore, their continued existence implies either an initially substantial reservoir at the surface, which would suggest differentiation, or continual resupply. On the other hand, Charon's differentiation state remains ambiguous, due primarily to the large remaining uncertainty on its density and surface frost coverage.

Interior models

Based on the previous arguments, the most likely internal structure for Pluto includes a ∼300 km thick water ice layer (either pure ice I or a combination of ices I and II) and a ∼850 km thick layer of partially hydrogenated rock. The deeper part of the ice layer probably undergoes solid-state convection. The boundary occurs at a typical pressure of 0.2 gigapascals (GPa) and a temperature of 250 K. Conditions at the center are a pressure of 1.30 GPa and a temperature in excess of 1000 K. This model corresponds to a rock fraction of 0.65. If Pluto is organic-rich (say organic/rock mass ratio= 0.2), it may contain a separate ∼100 km thick layer of organics. For Charon, both differentiated structures, similar to Pluto, and undifferentiated (in this case the interior is a mixture of water ice I and II and hydrated rock) are possible, and the pressure at the center is about 0.2 GPa.

Origin of Pluto and Charon

Our understanding of the origin of the Pluto–Charon system has enormously improved in recent years, especially as the true nature of the dynamical link between Pluto and Neptune became apparent and as new objects were discovered in the Kuiper-Edgeworth disk. The most popular early hypothesis, based on the fact that Pluto's orbit is 'Neptune-crossing', was that Pluto was formerly a satellite of Neptune, which had been ejected to its present orbit by a collision with Triton (reversing, by the same mechanism, Triton's orbit). In fact, Pluto's 3:2 orbit resonance with Neptune, by preventing close approaches between Neptune and Pluto, implies that no Neptune

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ejection event could have placed Pluto on its present orbit. In addition, the high rock-to-ice ratio forbids formation in a planetary subnebula. Other bulk and chemical characteristics (density, presence of volatile ices) indicate instead a formation in the outer solar nebula, beyond at least 10 AU, although the constraints on the exact location are loose. The existence of the Pluto–Charon pair, with a satellite mass ratio of about 1/9, and the obliquity of the system suggests that one or more substantial physical collisions occurred in the past. It must be stressed, however, that the astronomically low probability of such collisions seems to imply that Pluto and its impactor (the Charon progenitor) were members of a large population of small planets. Charon's orbit attributes suggest that it has undergone quick tidal evolution. Finally, the discovery in recent years of numerous objects with diameters of possibly tens to several hundreds of kilometers in the Kuiper- Edgeworth disk between 32 and 48 AU has placed Pluto in a clear dynamical and population context.

With these various constrains in hand, a baseline scenario can now be elaborated. In this picture, Pluto (and independently the Charon progenitor) formed in a near-circular, low-inclination heliocentric orbit, probably beyond Neptune. At that time, a much higher density of population probably existed in this region, although it is not known whether this population included numerous objects similar in size to Pluto or Charon. Once formed, due to the perturbations by Neptune, these bodies underwent orbital migration and were captured in a stable suite of resonances and ultimately in the 3:2 mean motion resonance, amplifying their eccentricity and inclination. At some point, probably when the two bodies already had significant eccentricities, Pluto and the Charon progenitor underwent a mutual collision which formed the binary system. Charon's orbit then quickly (in $< 10^8$ yr) evolved to its present synchronous state. Neptune perturbations on the 35-50 AU zone cleared the Kuiper-Edgeworth disk of most of its mass and prevented further growth of Pluto into a larger body.

Bibliography

Buie M W *et al* 1997 *Icarus* **25** 233-44
Cruikshank D P *et al* 1997 *Pluto and Charon* (Tucson, AZ: University of Arizona) pp 221-67
Elliot J L and Young L A 1992 *Astron. J.* **103** 991-1015
Malhotra R and Williams J G 1997 *Pluto and Charon* (Tucson, AZ: University of Arizona) pp 127-57
Stern S A *et al* 1997 *Astron. J.* **113** 827-85

Emmanuel Lellouch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pluto–Kuiper Express

Proposed NASA mission to fly past Pluto and its moon Charon, and study Kuiper Belt objects. Possible launch in 2004, with arrival 2012–18. Named after distinguished planetary scientist Gerard Kuiper.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Poincaré, Jules Henri (1854–1912)

Born in Nancy, Lorraine, France, became professor of mathematics at the Sorbonne in Paris, contributed to all areas of mathematics, including the three-body problem of celestial mechanics and studies of the stability of the solar system. He was first to consider the possibility of chaos in planetary orbits, foreshadowing the modern study that restarted in 1963. His work began as the answer to a question in a contest to show rigorously that the solar system, as a multi-body modelled by Newton's equations, is dynamically stable. Poincaré realized that the outcome prediction was sensitive to the initial conditions: 'If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. but even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation approximately. . . it may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible. . .' Poincaré also determined the shape of a rotating fluid subject only to gravity. He wrote *Les Méthods Nouvelle de la Méchanique Celeste* (1892–99) and *Leçons de Mecanique Celeste* (1905). He wrote on the philosophy of science, worked on equations of mathematical physics, and, with ALBERT EINSTEIN and HENDRICK LORENTZ, developed the special theory of relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pointers

A name sometimes used for the two leading stars of the Plough, Dubhe (α UMa) and Merak (β UMa), arising from their frequent citation as locating stars for Polaris, and hence for the north celestial pole.

See also: Dubhe, Merak.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polar

Part of NASA's Global Geospace Science programme and the International Solar Terrestrial Physics (ISTP) programme. Launched in February 1996 to study the interaction between Earth's magnetosphere and the solar wind at high latitudes. Imaging instruments measure visible, ultraviolet and x-ray spectra of the polar regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polar Caps

Bright, icy surface deposits in the polar regions of a planetary body. In the solar system, the Earth, Mars and Neptune's major satellite Triton are known to possess polar caps. The Earth's polar caps are permanent and consist of water ice, with a maximum thickness of about 4 km. The northern cap lies over Greenland and the Arctic Ocean; the southern covers the continent of Antarctica. Both show a little seasonal variation, with different limits of sea ice in winter and summer. Mars's polar caps show a marked seasonal variation. At their minimum extent (about 600 km across for the northern cap, 400 km for the southern) they have a distinctive swirled pattern, the ice filling valleys that curve outward from the poles. In the winter the southern cap extends to around 60 °S (50 °S in the Argyre Basin), the northern to about 65 °N, as carbon dioxide sublimes out of the atmosphere to form extensive deposits of frost. While the permanent southern polar cap consists of carbon dioxide ice, the northern one is of water ice. Triton has a pink southern polar cap, probably of nitrogen ice and frost. This is expected to migrate to the north polar region and back as the angle of solar illumination changes over the course of Neptune's 165-year orbit round the Sun.

See also: Mars: surface, Triton.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polar Motion and Length of Day

Polar motion

Polar motion refers to the movement of the axis about which the Earth rotates. The length of day is the difference between the day measured using astronomical observations and the day as measured by an atomic clock. Both quantities are needed to specify the orientation of the Earth in space and both vary with time.

The angles which characterize the direction of the rotational pole with respect to the Earth's surface are called the polar coordinates, x and y. Variation of these coordinates is called polar motion. The polar coordinates specify the direction of the Earth's pole of rotation in a reference frame which is defined by the adopted locations of terrestrial observatories. The coordinate x is measured along the 0º (Greenwich) MERIDIAN while the coordinate y is measured along the 90ºW meridian. These two coordinates determine the directions on a plane onto which the polar motion is projected (figure 1).

Polar motion consists largely of two motions. The largest is a mostly circular path which repeats after approximately 435 days and is called the 'Chandler' motion after Seth Chandler, who discovered it in 1890. It is caused by the misalignment of the EARTH'S ROTATION axis with respect to its axis of symmetry. The second motion is one which follows an elliptical path and repeats annually. This motion is caused by the redistribution of atmospheric mass during the year. These two motions account for the spiral motion of the pole as seen from the Earth's surface (figure 2).

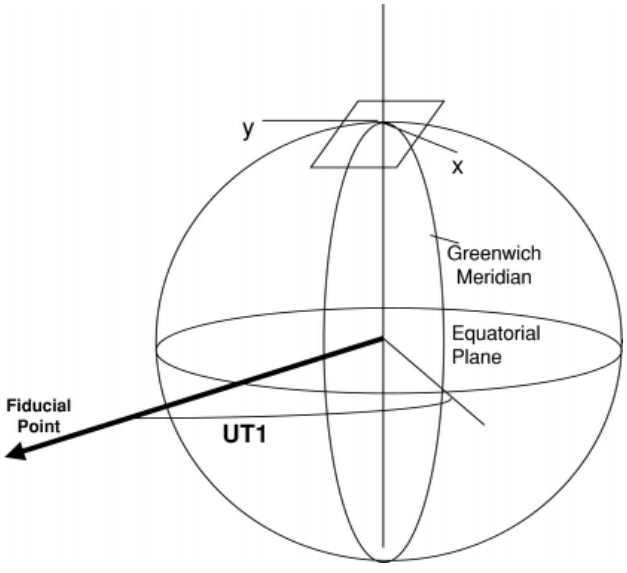


Figure 1. UT1 and polar coordinates x and y.

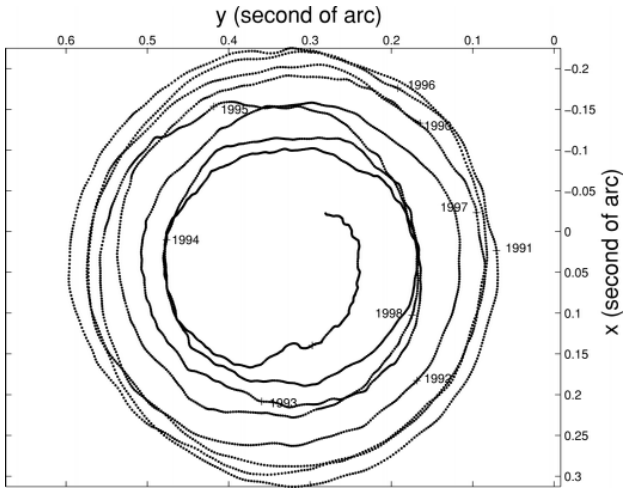


Figure 2. The path of the rotational pole of the Earth. 0.1 arcsecond corresponds to 3 m.

Earth rotation

The Earth rotation coordinate is the angle through which the Earth has turned in a given period of time. This angle expressed as the difference between a time scale measured by the rotation of the Earth, UT1, and a uniform time scale, UTC, refers to the angular separation between the direction of the 0º meridian on the Earth and the direction to a point defined in space astronomically (figure 1).

Historically some form of time based on the rotation of the Earth has always been the basis for civil time, the definition and the measurement procedures depending on available technology and precision requirements (see TIME SCALES: TERRESTRIAL). In modern practice, UT1 is determined using a fiducial direction defined mathematically in the celestial reference system. This direction is referred to as the Mean Sun.

UTC, standing for Coordinated Universal Time, designates the atomic time scale which approximates the rotational time of the Earth (see UNIVERSAL TIME, TIME SCALES: ATOMIC). From the time of its inception, its rate and/or epoch have been adjusted to keep it near UT1. The current practice is to adjust UTC in epoch by integral seconds (leap seconds) to keep the difference between UT1and UTC less than 0.9 s. UTC as defined by the International Radio Consultative Committee (CCIR) Recommendation 460-4, differs from TAI (Temps Atomique International INTERNATIONAL ATOMIC TIME) by an integral number of seconds. TAI is an atomic time scale determined by the Bureau International des Poids et Mesures (BIPM). Its unit is exactly one Systeme Internationale (SI) second at mean sea level.

The difference between the length of the day measured astronomically and 86 400 s (24 h × 60 min × 60 s) is defined as the excess length of day. It is often referred to as the 'length of day' and abbreviated 'LOD'.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This quantity is related to the rotational speed of the Earth, such that a slower rotational speed will be seen as an increase in the LOD and vice versa.

Until recent times the rotation of the Earth has served as the basis for timekeeping. The assumption was made that the rotational speed of the Earth was essentially constant and repeatable, and that the length of the day which resulted from this constant rotational speed was naturally useful as a measure of the passage of time. Comparisons of astronomical observations of different kinds and of terrestrial clocks have shown, however, that the speed with which the Earth is rotating is not constant with time. Variations in rotational speed may be classified into three types: secular, irregular and periodic. The secular variation of the rotational speed refers to the approximately linear increase in the length of the day due chiefly to tidal friction. This effect causes a slowing of the Earth's rotational speed resulting in a lengthening of the day by about 0.0005 to 0.0035 s per century.

The irregular changes in speed appear to be the result of random accelerations, but may be correlated with physical processes occurring on or within the Earth. These have caused the length of the day to vary by as much as 0.01 s over the past 200 years. Irregular changes consist of so-called decade fluctuations with characteristic periods of 5–10 years as well as variations that occur at shorter time scales. The decade fluctuations are apparently related to processes occurring within the Earth. The higher-frequency variations are now known to be related to the changes in the total angular momentum of the atmosphere.

Periodic variations are associated with periodically repeatable physical processes affecting the Earth. Tides raised in the solid Earth by the Moon and the Sun produce periodic variations in the length of the day as large as 0.0005 s with periods of 1 year, 1/2 year, 27.55 days and 13.66 days, among others. Seasonal changes in global weather patterns occurring with approximately annual and semi-annual periods also cause variations in the length of the day of this order.

Knowledge of the rotational speed of the Earth is required for observers on the Earth who find it necessary to know the orientation of the Earth in an inertial reference frame. This includes navigators, astronomers and geodesists. The rotational speed of the Earth remains essentially unpredictable in nature due to the incompletely understood irregular variations. Because of this, astronomical observations continue to be made regularly with increasing accuracy, and the resulting data are the subject of continuing research in the field (see EARTH'S ROTATION: THEORY).

Analyses of astronomical observations reveal different types of variations in the speed of rotation. The ancient observational data form the basis for estimates of the secular deceleration in the speed of rotation. The more recent information, having been obtained with higher accuracy and more regularity, has shown the changes in the acceleration causing irregular variations in the length of the day. These data have also been used to detect the periodic variations in the length of the day. Figure 3 shows the excess length of day in milliseconds or the difference between the rotationally determined length of day and 24 h of UTC time.

Causes of variations in the Earth's orientation

The conservation of ANGULAR MOMENTUM of the Earth requires that changes in the Earth's moments of inertia produce changes in the Earth's orientation. The moments of inertia are dependent on the distribution of mass on and within the Earth. This includes the mass contained in the atmosphere and in the oceans. As mass is redistributed, the moments of inertia change, producing the subsequent changes in the Earth's orientation. An example of such an effect is that of the spinning skater who is able to change rotational speed by redistributing mass through the changing of the positions of the arms.

The crust and the interior of the Earth are not strictly rigid. Tides generated by the Sun and the Moon change the shape of the Earth. The Earth is thought to have a fluid core with material possibly in motion. Coupling between such a core and the mantle would change the speed of rotation of the mantle from which the observations are made. The Earth may undergo changes in the moments of inertia as landmasses move or rise and fall. All of these motions cause changes in the Earth's moments of inertia and may cause changes in the rotational speed. Seasonal variations are caused by an exchange of momentum between winds and the crust of the Earth.

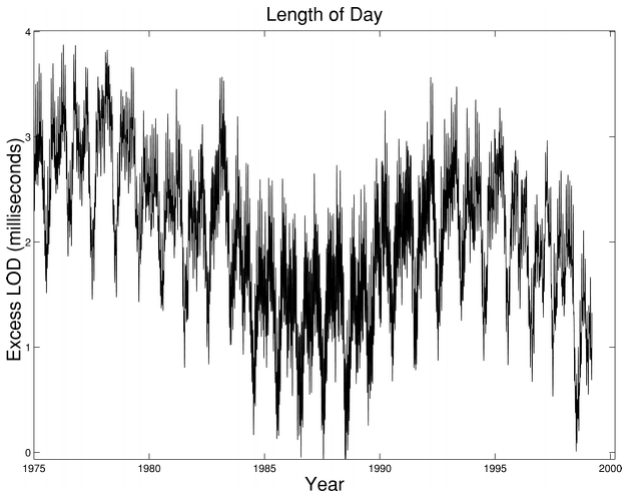


Figure 3. Excess length of day.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Changes in the total angular momentum of the atmosphere have also been shown to be correlated with changes in the length of the day. Precise calculations of the total angular momentum using global observations of the wind speed and direction have been used to demonstrate this effect. The variations in the rotational speed are apparently caused by the interaction of the winds with the Earth's topography, and may be the leading cause for all of the irregular variations in the rotation of the Earth.

Another known cause of variations in the rotational speed is tidal friction. The Moon raises tides in the ocean. Friction carries the maximum tide ahead of the line joining the center of the Earth and Moon. The resulting couple diminishes the speed of rotation of the Earth, and this reacts on the moon to increase its orbital momentum. The sum of the angular momentum of the Earth and the orbital momentum of the Moon remains constant. This produces an increase in the size of the orbit of the Moon and a reduction of its angular speed about the Earth. Tidal friction should be distinguished from actual changes in the moments of inertia of the Earth brought about by the tides.

Observations of variations in the Earth's rotation

In order to detect variations in the Earth's orientation, observations must be made from the Earth of objects located in space. These include stars, artificial satellites, the Moon and distant radio sources called quasars. These bodies provide useful reference directions with which to measure the Earth's orientation.

Stars were observed photographically for decades to determine the motion of the pole and the rotation of the Earth. Recently, more accurate methods have been devised including the use of lasers, the Global Positioning System (GPS) and radio telescopes. Laser bursts can be bounced off of artificial satellites or the Moon. The length of time to receive the reflected pulse provides information on exactly where the object is at a particular time which, in turn, can be used to determine the Earth's orientation in space.

Observations of the GPS satellites at stations participating in the International GPS Service provide estimates of the Earth's orientation with respect to the constellation of the satellites. Measurements of the length of time taken for a radio signal to travel to the observer on the Earth from the different satellites are used to locate the observer's site on the surface of the Earth as well as determine the orientation of the entire globe in the celestial reference system.

Radio telescopes can also be utilized in a technique called Very Long Baseline Interferometry (VLBI). By having several radio telescopes looking at the same quasar at the same time and recording the information that is seen at each telescope, the Earth's orientation can be determined. The recorded information then needs to be processed further before the final results can be determined. The reduction procedure involves the use of a highly specialized computer called a correlator. Once the data from a VLBI experiment have been correlated, they can be processed further to produce Earth orientation information.

Astronomical observations are made routinely by a number of observatories located around the world for this purpose. The International Earth Rotation Service (http://www.iers.org/) is the international organization responsible for the coordination of observations of polar motion and nutation as well as astronomical time. The Central Bureau is located at the Bundesamt für Kartographie und Geodäsie (BKG) Frankfurt am Main, Germany. The Sub-bureau devoted to rapid service and predictions is located at the US Naval Observatory. Observations are contributed to the IERS by observatories and laboratories around the world and treated at analysis centers. The information is recomputed twice per week and disseminated by E-mail, anonymous ftp and the world-wide web.

Bibliography

Lambeck K 1980 *The Earth's Variable Rotation: Geophysical Causes and Consequences* (London: Cambridge University Press)

D D McCarthy

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polar Plumes

Polar plumes (figure 1) are thin, bright, linear structures that are seen to emanate from the poles of the Sun during the 'minimum' phase of the solar cycle, when there are well-developed polar CORONAL HOLES. They may be seen with the naked eye during a solar eclipse or at any time with coronagraph instruments sensitive to visible or ultraviolet light. When seen with visible light, they are often called 'coronal rays'.

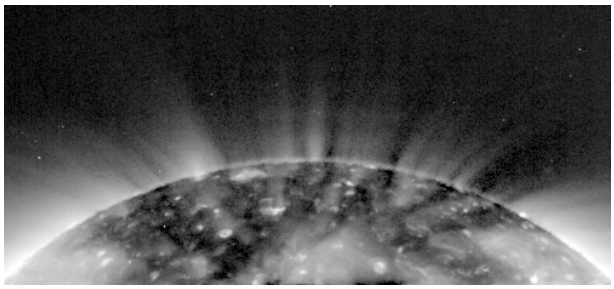


Figure 1. Polar plumes emanating from the pole of the Sun. The photograph was captured with the EUV Imaging Telescope aboard the SOHO spacecraft at Earth's L-1 point, in a cluster of spectral lines given off by Fe IX and X (Fe^{8+} and Fe^{9+}) ions. These ion species are present at approximately 0.9–1.2 million K.

Polar plumes are denser and cooler than the surrounding coronal hole material. They arise from approximately unipolar concentrations of magnetic flux under the polar coronal holes, expanding and spreading with altitude as do the holes themselves. Material is constantly accelerated along the plumes, forming part of the fast SOLAR WIND that streams from the polar coronal holes to fill most of the solar system. The plumes trace out the predominantly open magnetic field lines that connect their magnetized footpoints to the interplanetary medium. Near the limb of the Sun, they subtend about 2° relative to the center of the solar disk, but they expand superradially with altitude, subtending up to 30° at 15 solar radii above the disk in the plane of the sky. Above 25 solar radii, they fade from view, merging with the surrounding fast solar wind from other parts of the coronal hole.

Polar plumes are important to the study of the solar corona because they present the simplest, most quiescent system within the corona and because they may contribute significantly to the high-speed solar wind that fills virtually the entire solar system. Because the flux tubes that make up the plumes are magnetic open structures, no magnetic stresses can build up along their length: displacements of individual field lines are free to propagate out the top of the plume at the Alfvén speed. Without concentration via magnetic stress of the energy available from the turbulent motions of the SOLAR PHOTOSPHERE, the physical mechanisms available to heat plumes are a small, strict subset of those available elsewhere in the corona. Likewise, the unipolar nature of the overall polar fields near solar minimum ensures that

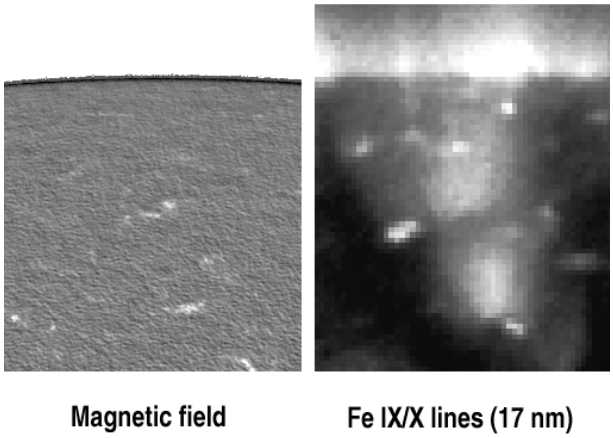


Figure 2. A close-up of a pair of polar plume footpoints from figure 1. The left panel shows the line-of-sight component of the magnetic field at the visible surface of the photosphere. The right panel shows corresponding plume structures in the corona, using spectral line emission at 17 nm. The plumes arise from unipolar magnetic structures in the photosphere; their shape is defined by the lines of magnetic force that are 'pinned' to the surface at the magnetic footpoints.

the plumes are only weakly coupled to other features in the corona, so they are well isolated from the more complex activity elsewhere. These two aspects make polar plumes a simple but important test case for our understanding of the more general problem of the coronal heat.

Overall shape and magnetic structure

Plumes take the form of tall, thin structures that follow a particular bundle of magnetic field lines from the surface out to interplanetary space. They arise from unipolar magnetic flux concentrations (i.e. magnetic dipoles with one pole buried inside the Sun and the other poking out through the photosphere). The flux concentrations are approximately 3000–10 000 km in size with typical total magnetic fluxes of 10^{19}–10^{20} Mx and peak field strengths of up to several hundred gauss. The field lines diverge rapidly at first from their photospheric roots, expanding sideways with altitude until they begin to interact with other field lines of similar polarity from elsewhere in the coronal hole (see figure 2). This initial turning point takes place between 30 000 and 50 000 km above the surface of the Sun, giving them the appearance of blunt pencils near the surface of the Sun.

At higher altitudes, the plumes expand and diverge slightly faster than radial (vertical) lines would. This is due to the fact that the entire coronal hole expands with altitude to fill the void left by closed magnetic structures (such as streamers and loops) at lower latitudes. By 4–5 solar radii (3 million km) over the surface, each polar coronal hole fills almost a complete hemisphere around the Sun, and its expansion turns radial. The plumes stop diverging, but continue expanding with altitude until ∼10 solar radii over the surface (figure 3). They have been detected as far as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

25 solar radii from the Sun using long-duration exposures with space-based coronagraphs.

Plume-like magnetic flux concentrations account for approximately half of the field lines that escape from the coronal hole, although not all flux concentrations support visible plumes. Of the several hundred flux concentrations that are present at any given moment in the solar minimum coronal hole, 10–20 support bright, readily distinguishable plumes. The flux concentrations that harbor plumes are typically associated with weaker opposing flux regions nearby, suggesting that reconnection of the opposing field may be important to plume formation.

At low altitudes, the relatively strong magnetic field dominates the plasma pressure (the plasma β parameter is low), so there is little difference in shape between field lines that harbor plumes and those that do not. Because the coronal hole expands superradially with altitude to occupy approximately an entire hemisphere, and because β remains low to altitudes of several solar radii, these plume-like magnetic structures are connected to a large fraction of the volume of the solar system.

Plume density and temperature

One may determine the temperature of solar coronal features in three principal ways: through spectrography, or by measuring the Doppler broadening of spectral lines by Brownian motion of atoms in the feature ('dynamic temperature'); or by measuring the relative brightness of several spectral lines, indicating the relative abundance of different ionization species ('ionization temperature'); by measuring the rate of falloff of density with altitude, which indicates the equilibrium between gravity and the gas pressure of the plasma ('scale height temperature'). All three methods show that polar plumes are cooler than their surrounding environment at altitudes of 1.05–1.3 solar radii from the center of the Sun (50 000–400 000 km from the surface of the Sun). Spectral line width maps of the visible coronal hole, which measure dynamic temperature versus position, show that polar plumes have somewhat narrower spectral lines than do interplume regions. Ionization temperature measurements indicate that plumes have temperatures close to or under 1×10^6 K, while the interplume coronal hole may be as hot as 1.5×10^6 K. At low altitudes (<1.5 solar radii), where the plasma motion is negligible compared with the local speed of sound, plumes' scale heights are shorter than those of the surrounding interplume regions, resulting in a decreasing contrast with altitude. Above about 1.5 solar radii, the plumes are not in hydrostatic equilibrium, because the outward and upward gas flow is not negligible compared with the local speed of sound, and scale height is not a meaningful measure.

Plumes appear bright in visible light instruments such as SOHO/LASCO and the HAO Mk III White Light Coronameter located on Mauna Loa. Those instruments image visible white (continuum) sunlight that is scattered by free electrons in the corona; the amount of scatter depends only on the number of electrons and not on their temperature. The plumes must be denser than the surrounding media to be seen. Several measurements of plume density have been made using both integrated emission measure in visible and EUV images and ratios of particular density-sensitive emission lines. Near their bases, plumes are seen to be anywhere from 2 to 10 times denser than the surrounding media, with the brightest plumes perhaps 10 times denser than the surrounding coronal hole. A typical low-altitude coronal hole density is a few 10^{13} m^{-3}; the very brightest plumes may be as dense as 10^{15} m^{-3} near their bases.

The fact that plumes are cooler and denser than their surroundings indicates that they are heated principally at their bases. Heat deposited more than a few 10^5 km above the surface of the Sun would mainly increase the local temperature, yielding hot, rarefied plumes. However, heat deposited near the surface is readily conducted downward into the dense, cool CHROMOSPHERE below, causing cool material to evaporate into the corona and yielding cooler, dense plumes. In the latter case, the additional density causes the plumes to be cooler than the surrounding coronal hole, because the denser plasma radiates energy more efficiently than a more rarefied one would. Coronal plasma is cooled mainly by radiation of collisionally excited spectral lines. These lines' brightness varies as the square of the plasma density, so that the denser plumes have a proportionally higher rate of energy loss and hence a lower equilibrium temperature than does the less dense interplume plasma.

Temporal evolution

On human timescales of a few minutes to a few hours, plumes are quiescent structures: their brightness and overall shape remain almost constant over several hours. However, each plume comprises many smaller filamentary structures less than 5 arcsec (3000 km) across at the base of the plume. These smaller structures fluctuate in brightness very rapidly, lighting up and dimming in just a few minutes. A typical plume may support three to five bright filamentary structures at a time. The bright substructures are thought to be the result of numerous small magnetic reconnection events inside the base of each plume. Each entire plume stays bright for about 20–30 h at a time, which roughly corresponds to the time it takes for its plasma to cool by radiation alone. However, after a plume has vanished new plumes are seen to recur in roughly the same location on the solar surface, brightening repeatedly in the same location over periods of up to several weeks. This recurrence suggests that the field lines over a particular flux concentration on the surface are re-filled with plasma from time to time, perhaps because of reconnection with new opposing flux regions that are carried to them by the motion of supergranules in the photosphere (see SOLAR PHOTOSPHERE: SUPERGRANULATION).

Plumes support low-frequency magneto-acoustic waves that propagate upward from the photosphere into the corona. These wave fronts are quasi-periodic, coming in trains of only a few wavefronts at a time, with periods

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

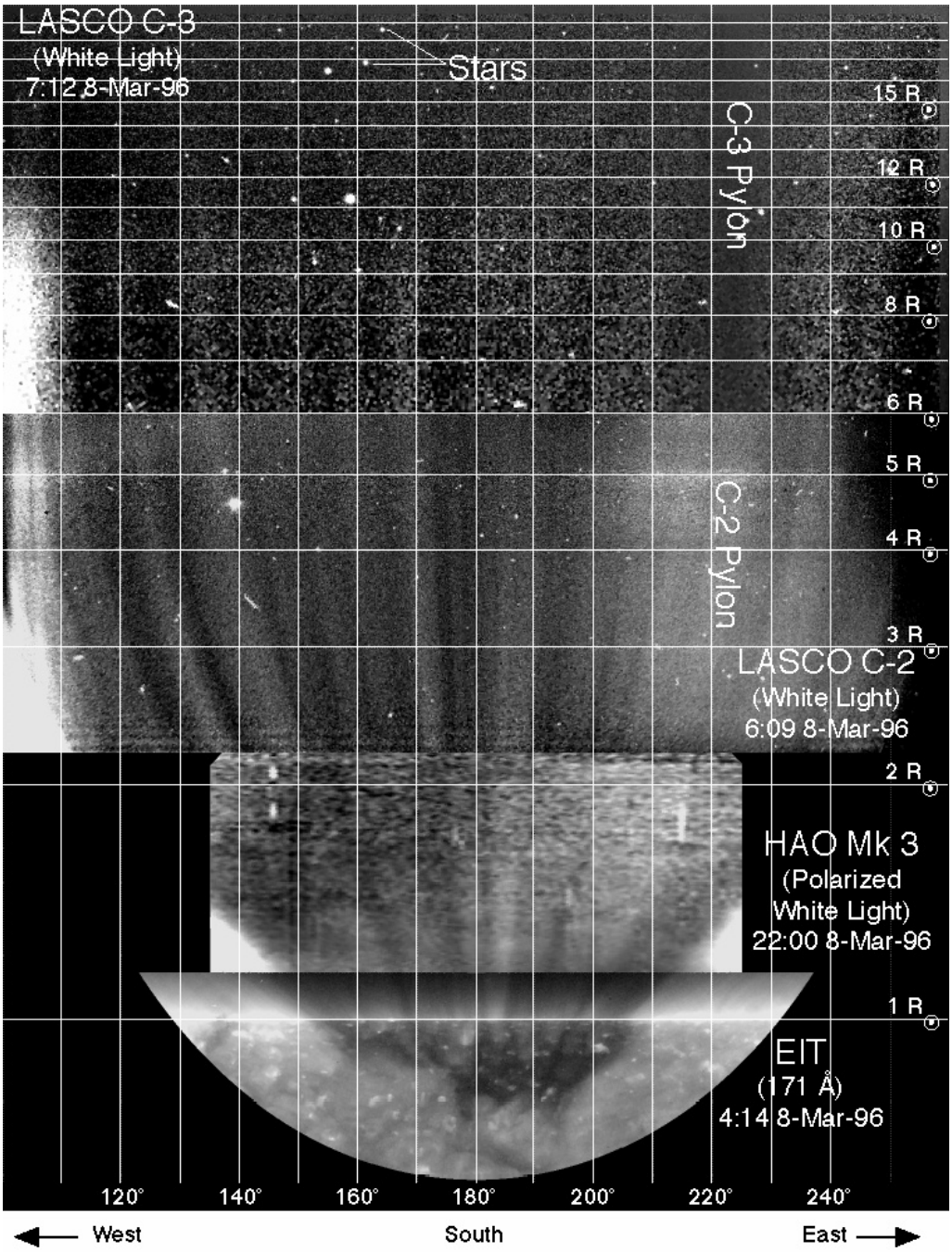


Figure 3. The plumes in figure 1 in the context of the whole corona. Data from four instruments have been combined to make this montage image, in which the horizontal direction measures azimuth around the Sun and the vertical direction measures distance from the center of the solar disk. The vertical scale is logarithmic, to show both the small roots of the plumes and their vast extensions into interplanetary space.

of 10–20 min (1–2 mHz). They have density amplitudes of up to 10%, resulting in traveling fluctuations in brightness in the extreme ultraviolet seen by spaceborne telescopes such as TRACE and EIT. The waves carry energy fluxes of about 50 W m^{-2}, or a few per cent of the energy flux (about 1 kW m^{-2}) required throughout the coronal hole to heat the corona and drive the high-speed solar wind. The origin of the waves is not known.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Connection to the high-speed solar wind

At any given time, there are only a few bright plumes in existence; these brightest plumes cannot account for more than a small proportion of material streaming out in the solar wind. However, the vast majority of plume-like magnetic structures are only slightly brighter–denser than the interplume regions between them; these fainter plumes do subtend a significant part of the solid angle in the coronal hole, and in turn must make up a large fraction of the solar wind. Hence, understanding solar wind flow in polar plumes is important to the broader problem of understanding the high-speed solar wind.

With the advent of both stable coronagraphs and accurate coronal UV spectrometers aboard the SOHO spacecraft, it has been possible to place limits on the flow speeds inside polar plumes. Direct Doppler measurements by the SUMER instrument of spectral lines emitted by the magnesium IX and oxygen V ion species in the lower corona show that plumes' outward flow is very subsonic in the lower reaches of the coronal hole: 150 Mm above the surface, SUMER measurements constrain the radial outflow speed to less than 30 km s^{-1}. Higher up, about 1 solar radius over the limb of the Sun, Doppler dimming measurements with UVCS show outflow speeds of 100–200 km s^{-1}, indicating that the sonic point occurs between 1 and 2 solar radii over the limb (2 and 3 solar radii from disk center). This rapid acceleration cannot be explained by hydrostatic forces alone; some other form of momentum transfer, such as waves launched from the photosphere or chromosphere, must be at work.

Because plumes appear to extend well into the outer corona, one might expect traces of them to remain in the high-speed solar wind at interplanetary distances. The Ulysses spacecraft has flown over the poles of the Sun, sampling the solar wind throughout the coronal hole at latitudes from 0° to over 80° at distances greater than 1 AU. Fluctuations in the density of the solar wind are seen on timescales of about a day, suggesting that the traces of plumes may survive as coherent structures to distances of more than 1 AU from the Sun itself. There is no direct evidence that the density fluctuations are actually polar plumes, in part because there was no spaceborne coronagraph in operation to compare with the *in situ* data during Ulysses' first overflight of the poles. The frequency spectrum of the fluctuations shows no major features at particular timescales and does not vary strongly with latitude within the coronal hole. This pattern could result either from randomly sized plume-like structures in the heliosphere or from time-dependent fluctuations in the solar wind itself. Whether or not the plumes survive to high heliographic altitudes, and how much they affect the solar wind, are important unanswered questions.

Bibliography

DeForest C E *et al* 1997 Polar plume anatomy: results of a coordinated observation *Sol. Phys.* **175** 393

Hassler D, Wilhelm K, Lemaire P and Schühle U 1997 Observations of polar plumes with the Sumer instrument on SOHO *Sol. Phys.* **175** 375

Newkirk G Jr and Harvey J 1968 Coronal polar plumes *Sol. Phys.* **3** 321

Poletto G *et al* 1996 Searching for coronal plumes in Ulysses observations of the far solar wind *Astron. Astrophys.* **316** 374

Wang Y M 1994 Polar plumes and the solar wind *Astrophys. J. Lett.* **435** L153

C E DeForest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polar Ring Galaxies

Polar ring galaxies are peculiar objects in the sky, composed of a primary galaxy disk and a ring of stars orbiting in a plane almost perpendicular to the primary disk (i.e. almost polar). The prototype of these peculiar galaxies is NGC 4650A, in figure 1.

Polar rings are most often found around early-type galaxies, essentially classified lenticular (or S0) in the Hubble sequence (see HUBBLE CLASSIFICATION). In very rare cases, these rings have been found around flattened ELLIPTICAL GALAXIES and, even less frequently, around late-type SPIRAL GALAXIES, which are gas rich themselves. In those cases, they are not well relaxed and are still composed of many tilted rings, in the process of settling to some symmetry plane in the potential of the primary galaxies (e.g. the Spindle galaxy NGC 2685, or NGC 660).

Spectroscopy has shown that the two disks are rotating and that their axes of rotation are almost perpendicular. In the most frequent cases, the polar rings have radii larger than the main galaxy (they are then more detectable); when there is gas in the two systems, they are never at the same radii, since gas cloud collisions will scatter them apart.

Note that a certain class of rare ringed galaxy could also be related to polar ring galaxies, those possessing only one broad ring around a central unbarred disk. The prototype is Hoag's object (figure 2), also called the 'bull eye'. We would then observe the central early-type galaxy end on, and the polar ring face on. The much more abundant barred galaxies possessing one or more resonant rings can be explained easily by the action of the bar. Also, head-on collisions between two galaxies can develop in the disk of each of them a density wave propagating radially as a ring. The prototype is the well-known CARTWHEEL GALAXY. These are recognizable from polar rings, since the galaxy nucleus is off-centered with respect to the ring.

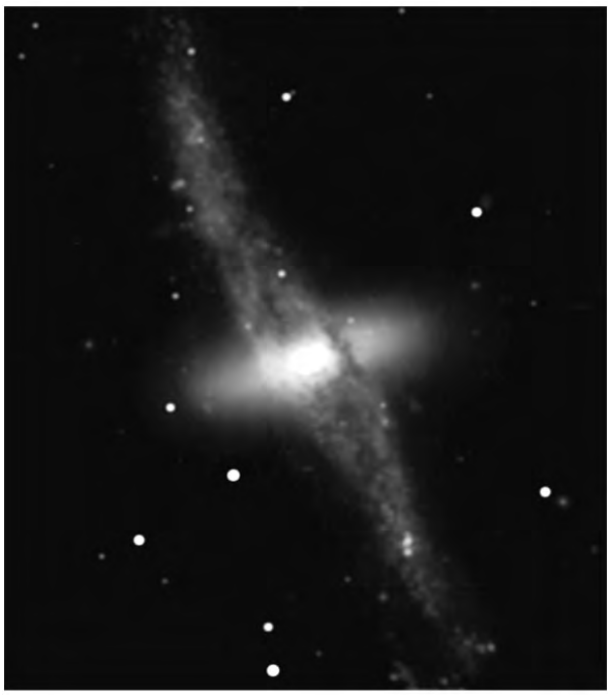


Figure 1. Prototype of the polar ring galaxies, NGC 4650A, a southern sky object in the constellation of Centaurus. Two components can be seen: a primary disk galaxy, oriented almost horizontal in this picture, which is of lenticular type, and an extended ring-like distribution of stars, almost perpendicular to it, which is called the polar ring itself. The latter looks younger, with an irregular and clumpy shape, containing gas and dust, while the primary disk is old, smooth, and mostly stellar. (Photograph from VLT ESO-Press Release 19/98.)

Origin of these galaxies

The existence of two systems with almost perpendicular directions of rotation is not possible without a secondary event in the formation of the system, such as a galaxy interaction or merger. When a galaxy forms, it is composed essentially of gas, and, through dissipative processes, this gas settles in a disk, perpendicular to the average angular momentum of the system (see GALAXY FORMATION, GAS IN GALAXIES). Clouds of gas with other directions of rotation lose their misaligned angular momentum in collisions that mutually re-orient their orbits. The equilibrium state is a disk of gas in rotation, in which stars form progressively, consuming the gas. All the matter is rotating in the same sense.

However, galaxies form in a hierarchical way: they are formed in groups, that belong themselves to clusters, then superclusters, etc.... Galaxies are not distributed homogeneously in the sky, and they have frequent gravitational interactions that lead to mass exchange and mergers. A merger between two galaxy disks, with different angular momenta, is the way to build polar ring systems. When a small galaxy companion is tidally destroyed and swallowed by the primary galaxy, the stellar component is heated and spread all around, since stars cannot collide and dissipate energy. The gas on the contrary can dissipate its energy and fall inside the potential well of the primary galaxy. It settles into a disk, perpendicular to the average angular momentum, as in the case of first formation, and, in general, it becomes aligned in the equatorial plane of the primary galaxy. In rare cases, however, there exists a stable polar orbit, where the gas can be trapped. If the gas disk then formed is massive enough, it changes the gravitational potential in its vicinity, which stabilizes the new disk. Stars can then form in this perpendicular disk, and a polar ring system is born. In most of the polar rings known, stars in the polar disk are a few Gyr old, demonstrating the age and stability of the system.

The formation of a polar ring is rare, since most of the time mass is accreted in the symmetry plane of the primary galaxy, sometimes in a sense opposite to the primary sense, forming a galaxy with two counter-rotating streams of stars. Also, polar rings are difficult to detect,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

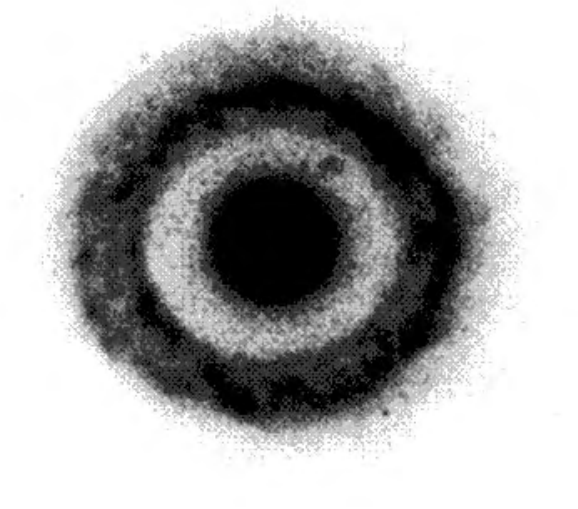


Figure 2. Photograph of Hoag's object, the 'bull eye', showing the central core and surrounding broad ring. The size of the ring is about 15 kpc. Objects of this type have been suggested to be polar rings seen face on. (From Schweizer *et al* 1987 *Astrophys. J.* **320** 454.)

since they are conspicuous only for special viewing angles (see figure 3), when the polar ring appears nearly edge on. The percentage of polar ring systems observed at the present time is about 0.5%, and taking into account the viewing selection bias, together with other factors (such as dimming or short lifetime of the rings), it is estimated at 5% in reality. This percentage should have been even higher in the past, since the rate of galaxy merging is predicted to peak at about one-third of the universe's age, in modern cosmological formation scenarios (see GALAXIES: INTERACTIONS AND MERGERS).

Stability

If the ring is not too massive, and the orbiting matter can be considered as test particles in the potential of the main galaxy, it must evolve quite rapidly, through differential precession. The rings are not exactly polar (tilted by about 20° from the pole), and therefore do not lie in a symmetry plane of the inner galaxy. The gravity field is not spherical, since the primary galaxy is flattened, but has a quadrupole moment, which produces a torque on particles in the ring, causing their orbit to precess. The precession rate depends on the orbital inclination and radius of each particle and is differential, since all particles do not precess at the same rate (the ring is broad and spans a significant range in radius). Then, it can be computed that the polar ring should look warped and twisted in much less than a Hubble time, since the particles in the ring acquire orbits with different orientations according to their radius.

This is a dynamical puzzle, since most present-day polar rings are flat, but also because the gas clouds will collide and dissipate when the disk is twisted and warped,

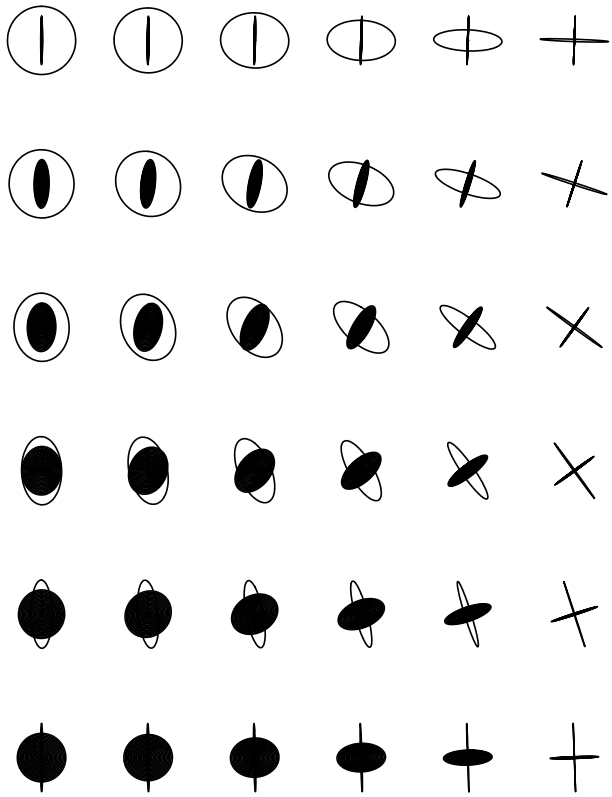


Figure 3. Schematic representations of a polar ring galaxy, as seen from different viewing orientations. The full disk represents the primary (lenticular) galaxy, and the thin ellipse, the polar ring itself. Note that, in the majority of cases, it is difficult to identify the projected system as a polar ring galaxy, the most favorable cases being the two systems nearly edge on.

and rapid inflow and settling to the equatorial plane will follow. The gas will not have time to form a disk of stars, as is observed.

A solution to this problem is that the polar disk is warped in such a manner that the tilt with respect to the pole compensates the decrease of the torque with radius: the polar disk is slightly tilted towards the pole at large radius. Also, the self-gravity of the polar disk produces compensating torques that stabilize the polar ring. In any case, if the ring is too broad, it can break up in a number of subrings that will precess independently. The inner rings will precess more rapidly and settle to the equator, allowing the remaining outer ring to become more polar and more stable.

The 3D shape of dark matter halos

Polar ring systems are very useful, since they give us the possibility not only to test galaxy formation scenarios but also to probe the dark matter distribution around galaxies (see DARK MATTER IN GALAXIES). Indeed, we know from the rotation curves of spiral galaxies, that they are embedded in dark matter halos. In particular, the interstellar gas that can extend in radius much farther than the stars in a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

galaxy disk is observed to rotate around the center at a high velocity. For this gas to be in rotation equilibrium requires a lot more mass in the disk than is visible by radiation. In other words, there must exist some dark mass that is detected only through its gravitational action.

Since the gravitational action of the dark matter is usually known only in the galaxy disk, through measurements of velocities here, we cannot know the 3D shape of the dark matter distribution: is it a spherical halo or a dark matter disk or has it some flattening in between? The answer to this question will considerably help to determine the nature of dark matter. If it is a non-baryonic exotic particle, interacting only gravitationally with ordinary matter, than it is not dissipative, and cannot be flattened as a disk. On the contrary, a distribution of dark baryons could be more flattened.

In polar ring systems, matter is rotating in two perpendicular disks, allowing us to probe the rotational velocity outside the main disk, in the dark matter halo. This is the occasion to deduce the 3D shape of dark halos. The problem is, however, more tricky to solve than expected. First, there is a selection bias to detect polar ring systems where the polar disk is bright enough: in that case, it is also massive enough to perturb the gravity field and mask the true dark halo potential. In the case of NGC 4650A, for instance, when all the mass is added (gas and stars), the polar accreted disk is almost as massive as the primary visible disk. The mass ratio is so extreme that one has proposed the opposite scenario, where the primary disk is the 'polar ring' and the accreted object is the lenticular old galaxy in the middle.

Second, in general, the polar system is outside and not overlapping with the inner disk, and the velocities in perpendicular planes are not known at the same radii. Extrapolations are therefore required, which are model dependent. This is true in particular when there is gas in both systems, although gas velocities are more precise.

Third, because of the flattened nature of the potential (at least due to the visible mass), there is no matter in circular orbit anywhere in the system (neither in the primary disk, which is no longer axisymmetric, owing to the mass of the polar ring, nor in the polar ring itself). Orbits are all elongated and explore a large range of radii. They must be modelled through shooting techniques and depend on the mass distribution in a complicated way.

Finally, it is also possible that the dark matter, if not spherical, is flattened around the polar disk and not around the primary disk. This has been shown to be a viable model for NGC 4650A in particular. In summary, polar ring galaxies carry a real hope that they will allow the 3D shape of the dark matter halos (and clues to their nature) to be determined, but precise spectroscopy and photometry data are needed to constrain the models, together with further exploration of the many possible models.

Bibliography

Book:

Casertano S, Sackett P D and Briggs F H (ed) 1991 *Warped Disks and Inclined Rings around Galaxies Proc.* (Cambridge: Cambridge University Press)

Journals articles:

Combes F and Arnaboldi M 1996 The dark halo of polar ring galaxy NGC 4650A: flattened towards the polar ring? *Astron. Astrophys.* **305** 763

Sackett P D, Rix H-W, Jarvis B J and Freeman K C 1994 The flattened dark halo of polar ring galaxy NGC 4650A: a conspiracy of shapes? *Astrophys. J.* **436** 629

Whitmore B C, Lucas R A, McElroy D B, Steiman-Cameron T Y, Sackett P D and Olling R P 1990 New observations and a photometric atlas of polar-ring galaxies *Astron. J.* **100** 1489–522

Françoise Combes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polarimetry: Imaging Polarimetry

Knowledge of the POLARIZATION state of radiation provides far more astrophysical information than intensity alone. Although the degree of polarization can be small, its measurement is relatively straightforward and it can reveal the underlying nature of sources such as young stars and the nuclei of active galaxies that are often hidden from our direct view.

Outline of the article

Our knowledge of the universe comes almost entirely through observing the ELECTROMAGNETIC RADIATION emitted from the great variety of objects contained within it. Yet measuring just the intensity of that radiation, whether it be its spatial variation or spectral composition, gives only a fraction of the information that is potentially available. This article gives a more complete description of light, by including its polarization properties, and then outlines the main physical processes responsible for producing polarized radiation in astronomy and the way it can be used to understand astrophysical processes. The basic components of an astronomical polarimeter are introduced followed by some recent results from astronomical research based on imaging polarimetry.

Basic physics

All electromagnetic radiation, from radio through to γ-rays, consists of an oscillating electric and magnetic field, with both oscillations transverse to each other and to the direction of propagation of the radiation. The polarization state of the radiation is, by convention, defined by the behaviour of the electric field. In the case of unpolarized radiation, the electric field orientation behaves in a random manner. For polarized radiation, the tip of the vector that represents the instantaneous electric field traces out an ellipse, a circle or a straight line, and the radiation is said to be elliptically, circularly or linearly polarized, respectively, with circular and linear polarization being particular cases of the more general elliptical form.

Degrees of polarization p can take any value from 0 to 1, and thus a mathematical description is required that can accommodate zero, complete and partial polarization. The most convenient and most widely used parametrization is due to Sir GEORGE STOKES, who used a 4-vector representation of radiation with components I, Q, U and V.

The link between the Stokes parameters and the geometry of a polarization ellipse is given by

$$I = a^2 \qquad Q = a^2 \cos 2\beta \cos 2\theta$$

$$U = a^2 \cos 2\beta \sin 2\theta \qquad V = a^2 \sin 2\beta$$

where a represents the ellipse size, $\tan\beta$ the ratio of the minor to major axes ($\beta = 0$ for linear and $\pi/2$ for circular polarization) and θ is the orientation of the ellipse. For linear polarization the ellipse collapses to a straight line and θ is known as the position angle of polarization, essentially giving the preferred angle of the radiation's electric field, as projected onto the plane of the sky. By convention this is measured from north, with increasing values rotating eastwards.

The degree of polarization p, linear polarization $p_{\rm lin}$, circular polarization $p_{\rm circ}$, polarized intensity $I_{\rm p}$ and θ are given by

$$p = \frac{(Q^2 + U^2 + V^2)^{1/2}}{I} \qquad p_{\rm lin} = \frac{(Q^2 + U^2)^{1/2}}{I}$$

$$p_{\rm circ} = \frac{V}{I} \qquad I_{\rm p} = Ip \qquad \theta = 0.5 \tan^{-1}\left(\frac{U}{Q}\right)$$

and the sign of V gives the handedness of the circular polarization.

Producing polarized radiation

Some physical processes directly produce polarized radiation. The main examples of these intrinsic processes are cyclotron radiation, which is circularly polarised and produced by the motion of non-relativistic electrons in a magnetic field, and SYNCHROTRON RADIATION, which is linearly polarized and produced by relativistic electrons in the presence of a magnetic field, with the polarization perpendicular to the magnetic field.

More commonly, polarized radiation is produced in secondary processes through the interaction of radiation with matter. For example, scattering off electrons (Thomson scattering) or dust grains whose size is small compared with the wavelength of the radiation (Rayleigh scattering) produces linear polarization, with the degree of polarization $p = (1 - \cos^2\theta)/(1 + \cos^2\theta)$, where θ is the angle through which the radiation is scattered. In both cases the direction of the electric vector is perpendicular to the scattering plane, the plane containing the incident and scattered light, and the degrees of polarization are independent of wavelength. For larger grain sizes (the so-called Mie regime for spherical particles), the scattering becomes more complicated and the degrees of polarization are generally smaller.

Polarized radiation as a diagnostic tool

The detection of scattered (polarized) light is a very important diagnostic tool in astrophysics. It allows regions to be observed that are normally obscured from direct view. Examples are young stars and ACTIVE GALACTIC NUCLEI (AGN), with both surrounded by optically and geometrically thick DUSTY CIRCUMSTELLAR DISKS (see figure 1). These disks obscure the central source unless our view is within typically $\sim 30^\circ$ of the normal to the disk plane (the disk polar axis). Optical and near-infrared radiation is unable to penetrate the dust but can escape along the poles of the disk and is then scattered to us by particles located along the polar axis of the disk. These particles are mostly dust grains in the case of young stars and electrons in the case of AGN. The scattering of light, above and below the plane of the disk, produces a REFLECTION NEBULA

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

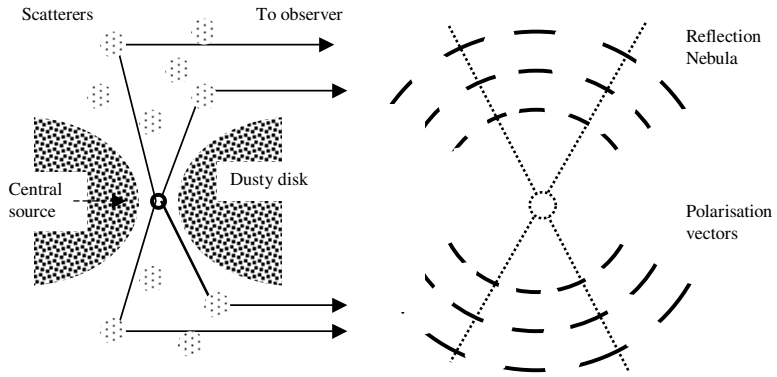


Figure 1. Light scattered from a central source is scattered to the observer by particles located along the poles of the obscuring disk. The reflection nebula is observed as a centrosymmetric polarization pattern with the length of each vector representing the degree of polarization and the direction of the line the position angle of polarization. The intersection of the normals to the polarization vectors (shown by dotted lines) gives the location of the illuminating source.

with observed polarizations ranging typically from a few per cent to 50%.

The scattered (polarized) flux in the reflection nebula gives the spectrum of the obscured source, although this will be modified if the scattering is off dust grains (in the Rayleigh domain, the scattered flux $\propto 1/\lambda^4$), but is unchanged by electron scattering. The degree and position angle of polarization give the geometrical relation between the source and scatterers relative to the observer, e.g. the inclination and position angle of the polar axis respectively, as well as providing the location of the illuminating source, given by the intersection of the normals to the polarization vectors. Importantly, polarimetry is able to provide geometrical information on scales that could never be spatially resolved directly at optical or infrared wavelengths, even with the largest telescopes.

In many astrophysical situations light is scattered more than once, and the scattering of linearly polarized radiation off non-Rayleigh particles produces circular in addition to linear polarization. Whereas linear polarization is dominated by the last scatter, circular polarization depends on the initial polarization state of the last-scattered photon and hence provides information on the previous history of that photon. Typical degrees of circular polarization produced by dust scattering are $\sim$2%, or less, presenting a significant technical challenge.

Circular polarization is also produced by the scattering of radiation (of any state) off aligned non-spherical grains. Dust grains can be aligned in many situations, and provided that the grains are paramagnetic they will always align themselves with their short axis parallel to the local magnetic field.

Grains in the interstellar medium (ISM) are aligned, and radiation from any astronomical source, in passing through the ISM, is polarized by dichroic absorption, i.e. the differential extinction of the light, with polarized components parallel to the long axis of the grain being preferentially absorbed. This produces a polarization with the electric field vector parallel to the magnetic field and can be used to measure the direction of the fields. Dichroic absorption is the same process that filters polarized light in sheet Polaroid. Warm dust grains will also emit radiation (at wavelengths longer than a few microns), and if they are aligned the radiation will be polarized perpendicular to the magnetic field.

Measuring polarization

Most imagers and spectrometers can be converted to polarimeters by the addition of two standard optical components. The first is some form of modulator, which is usually a waveplate (also known as a retarder), and the second an analyzer.

Both components are made from a birefringent crystal such as quartz or magnesium fluoride. Birefringent crystals have the property that the speed of light in the crystal depends on the direction of propagation and the polarization of the light. However, they do have one axis, known as the optical axis, along which all light—irrespective of the polarization state—travels with the same speed.

A birefringent crystal has two refractive indices (n_1 and n_2, with $n_1 - n_2$ known as the crystal birefringence), and light incident on a crystal surface, that is neither parallel to nor perpendicular to the optic axis, is doubly refracted. This is best understood by noting that, for polarized radiation, the electric field consists of two orthogonally polarized components which are refracted through different angles, producing two separate beams of light with orthogonal polarizations. This is the basis of the polarization analyzer.

If light is incident normally on a crystal, with its optical axis parallel to the crystal surface, the two light components coincide but there is a phase difference or retardance between them. If the difference in crystal optical thickness $d(n_1-n_2)$, where d is the crystal thickness, is exactly one-half or one-quarter of a wavelength then the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

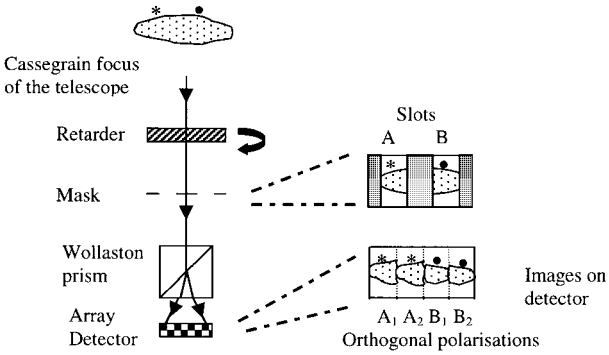


Figure 2. Typical polarimeter system used in optical and near-infrared imagers.

crystal is known as a half-wave and quarter-wave retarder respectively.

For measuring linear polarization a half-wave retarder is used. When rotated in the telescope beam it rotates the radiation's plane of polarization through an angle twice the mechanical rotation. Following the modulator is the analyzer (here a Wollaston prism, see figure 2). The intensity of the emergent, orthogonally polarized, beams depends on the direction of the electric field of the light incident on the analyzer.

As the half-wave retarder is rotated, the electric field of the emergent light rotates. On passing through the analyzer two light beams are produced with their relative intensity depending on the angle of the retarder. The two light beams are then recorded on an array detector. For light that is unpolarized no change in light intensity will occur as the waveplate is rotated, whereas if the light is 100% polarized the two beams will each have zero intensity for positions of the retarder separated by 45°.

The most common polarizers are calcite plates, which produce a lateral separation of the two beams equal to one-tenth of the calcite thickness, and Wollaston prisms that produce an angular divergence given by $2(n_1 - n_2)\tan\Omega$, where $n_1 - n_2$ is the birefringence of the prism material and Ω is the prism angle.

Exposures are normally taken with the retarder at 0°, 22.5°, 45° and 67.5°, where 0° is along the equatorial north–south direction. The Stokes parameter Q is calculated from the relative difference between the intensities of the two beams with the retarder at 0° and 45°, and U is calculated from the relative difference with the retarder at 22.5° and 67.5°. In principle, only the 0° and 22.5° positions are required but the 45° and 67.5° exposures provide some element of redundancy and also allow corrections for any difference in sensitivity between the two beams.

Essentially the same procedures are used for measuring circular polarization except that a quarter-wave, rather than a half-wave, retarder is used. The action of this retarder is to convert the circular polarization to linear, which is then analyzed.

In an imaging polarimeter, the two beams form two images on the detector, with both images representing the same area of the sky. A focal plane mask has to be employed to prevent overlap of these images and thus loss of all the required information (see figure 2).

Results from imaging polarimetry

(a) Studying reflection nebulae around young stars, embedded in dusty molecular clouds, has been one of the major uses of imaging polarimetry at optical and particularly near-infrared wavelengths. Up to the mid-1990s all measurements were of linear polarization. The degrees of polarization give the inclination of the system to our line of sight, the symmetry axis of the polarization pattern gives the position angle of the normal to the circumstellar disk and the wavelength dependence of polarization gives the size of dust grains responsible for scattering the light to us.

More recently, circular imaging polarimetry has been introduced. The degrees of circular polarization, produced by multiple scattering, are usually small (<2%), but nonetheless give important information on scatters close to the source, including in the circumstellar disk. Combined with the linear polarization, this technique provides far more information on grain composition and sizes.

For some high-mass star-forming regions very high degrees of circular polarization have been reported (∼20%). This is most likely produced by scattering off dust grains that have been aligned by magnetic fields in the region. Such high values, in a star-forming region, add weight to the suggestion that homochirality, the (left)-handedness of the amino acids of all living organisms on Earth, has an extraterrestrial origin, with the selective destruction of the other (right)-handed molecules by circularly polarized light.

(b) Imaging polarimetry of ACTIVE GALAXIES, in the optical and near-infrared, has been used to study those cases where the nucleus, presumed to be a massive black hole fuelled via an accretion disk, is obscured by direct view by an optically and geometrically thick torus (type II AGN). The spectrum of scattered light reveals the nature of the obscured source and its spatial distribution gives the geometrical distribution of the scatterers.

Light scattered from material along the torus polar axis (pointing away from us) is obscured by the torus. Hence the torus itself can be viewed in silhouette, back-illuminated by the scattered flux, and its size and inclination determined.

(c) Imaging polarimetry of galactic disks is used to map the galactic magnetic field structure, with the polarization presumed to arise from dichroic absorption at optical and near-infrared wavelengths, although contributions from scattered radiation can cause confusion. A more powerful technique is imaging polarimetry at radio wavelengths where the polarization is produced by synchrotron emission from charged particles, moving in the galactic magnetic field, with the electric field perpendicular to the direction of the magnetic field. At radio wavelengths scattering is unimportant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(d) Polarization images in the mid- and far-infrared of the central regions of the Galaxy, and some star-forming regions, have been obtained. Here the polarization is due to thermal emission from dust grains aligned by the ambient magnetic field and the position angle of polarization is at right-angles to the projected field direction. The mechanisms involved in STAR FORMATION are a major unsolved problem in astrophysics and this technique is at present the only one capable of studying the magnetic fields in the dense, neutral, molecular clouds where stars form.

Bibliography

A recent text covering polarimetry in astronomy:

Tinbergen J 1996 *Astronomical Polarimetry* (Cambridge: Cambridge University Press)

A text on the physics and applications of polarised light:

Collett E 1993 *Polarized Light: Fundamentals and Applications* (New York: Dekker)

Reviews of polarimetry of galaxies and of AGN are given in

Hough J H 1996 Polarimetry of external galaxies *Astron. Soc. Pacific Conf Series* vol 97, p 569
Hough J H 1997 Polarization properties of active galactic nuclei *Astron. Soc. Pacific Conf Series* vol 121, p 600
Hough J H 1997 IR polarimetry and the torus *Astrophys. Space Sci.* **248** 269

A paper presenting linear and circular polarization of a star-forming region, and discussing the origins of homochirality, is

Bailey J A, Chrysostomou A, Hough J H, Gledhill T M, McCall A, Clark S, Menard F and Tamura M 1998 Circular polarization in star formation regions: implications for the origin of biomolecular homochirality *Science* **281** 672

A paper on the polarization of the Galactic center is

Aitken D K, Smith C H, Moore T J T and Roche P F 1998 *Mon. Not. R. Astron. Soc.* **299** 743

James Hough

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polarimetry: Spectropolarimetry

Spectropolarimetry is an observational procedure in which the POLARIZATION characteristics of the light from an object are measured as a function of WAVELENGTH. Modern spectropolarimeters are capable of measuring not only the wavelength dependence of linear and/or circular polarization but also the total intensity of light at the same spectral resolution. In fact many spectropolarimeters are constructed by retrofitting a module containing the polarization optics to an existing spectrograph. The technique is most useful in objects that show relatively rapid changes in polarization with wavelength; otherwise imaging polarimetry (see POLARIMETRY: IMAGING POLARIMETRY) may suffice.

Polarization is caused by anisotropy. This can be the simple geometry of scattering from electrons, dust or molecules or it can be the direction of a magnetic field. Spatial resolution in astronomy is always at a premium, and complex geometries often go unrecognized because they cannot be seen directly. Polarimetry in general, and especially spectropolarimetry, can give information on spatially unresolved geometries in objects which would be nearly impossible to study with conventional techniques.

Spectropolarimetry has been used to discover hidden regions in the nuclei of active galaxies, study the asymmetric winds of stars, find parent stars of Herbig–Haro objects, separate reflected emission from emission produced *in situ* in the lobes of nebulae and characterize the magnetic fields of white dwarfs, among other things.

Polarization analyzers

Different wavelengths require different solutions to the measurement of polarization. For example in the x-rays Bragg scattering may be used. This article concentrates on UV–optical–IR applications. The simplest form of polarization analyzer is a material that preferentially absorbs one sense of polarization and which can be rotated in front of the detector. This is analogous to rotating a camera's polarizing filter in front of the lens to change the contrast of the sky in a daylight scene or reduce the reflection on a lake's surface.

Almost all spectropolarimeters, however, use some sort of optic (the *analyzer*) that can split light into two beams of opposite linear polarizations[1], plus some mechanism (the *modulator*) for changing the relationship between the analyzer and the incoming light. In *dual-beam* instruments the analyzer outputs, and the detector records, two orthogonally polarized beams. The calibrated intensities of the two beams are compared, and the difference implies the polarization level. In *single-beam* instruments one of the polarization senses is discarded, either by absorption within the analyzer or by instrument design, and two separate measurements must be made to determine one sense of polarization. In this case the modulator is necessary to present the polarization sense differently to the analyzer. In practice some number of inherently redundant measurements may be made, in either technique, in order to remove other effects within the system, such as drifting gains, changing backgrounds, etc.

Many analyzers use the intrinsic birefringence of some crystals, most notably calcite, magnesium fluoride and quartz. Sapphire and lithium niobate are common choices in the infrared. In birefringent crystals the different linear polarizations have different refractive indices. Light will naturally separate into two side-by-side beams in such crystals, although the separation of the beams is sometimes inadequate for use in real instruments. In analyzers such as the Glan–Taylor (or Glan–air) and Glan–Thompson prisms, one polarization undergoes total internal reflection at an interface, effectively separating it from the other polarization, which passes through the interface. In single-beam instruments the separated beam is discarded; in dual-beam instruments both beams are used. These analyzers can have relatively large beam separations and are achromatic over wide wavelength ranges.

In instruments that use smaller beam deviations, a 'comb dekker' is often used. This consists of alternating clear and opaque strips, each roughly one beam separation wide. This allows the second, offset beam from one clear strip to fall on a part of the detector that normally would be illuminated by the first beam of the adjacent opaque strip. A comb dekker allows relatively thin analyzers to cover large fields of view, but effects from the edges of the strips often further reduce the field available in a single image.

Polarization modulators

The modulator is also required to measure the other sense of polarization using the same analyzer. Sometimes the entire instrument may be rotated, but more common is the use of a *wave plate*. These are typically two or more birefringent materials combined to provide either a half wave or quarter wave of phase difference between the two polarization senses. A half-wave plate will rotate the plane of polarization of linearly polarized light; hence a rotating half-wave plate is a common component of a linear spectropolarimeter. A quarter-wave plate can be used to convert linearly polarized light to circularly polarized light and vice versa.

Wave plates are often designed to work over a large range in wavelength. Materials with different birefringent properties can be combined to produce achromatic or even 'superachromatic' wave plates, similar to the way in which different lens materials are combined to produce achromats or apochromats in lens design. The 'Pancharatnam' design, which uses six pieces of two different materials, is the most popular form of 'superachromatic' wave plate design. It is used in instruments which span a factor of 4 in wavelength coverage, such as optical spectrographs working from 3000 Å to 1.2 μm.

[1] Outside of the realm of radio astronomy, there are few analyzers that can efficiently split *circularly* polarized light.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

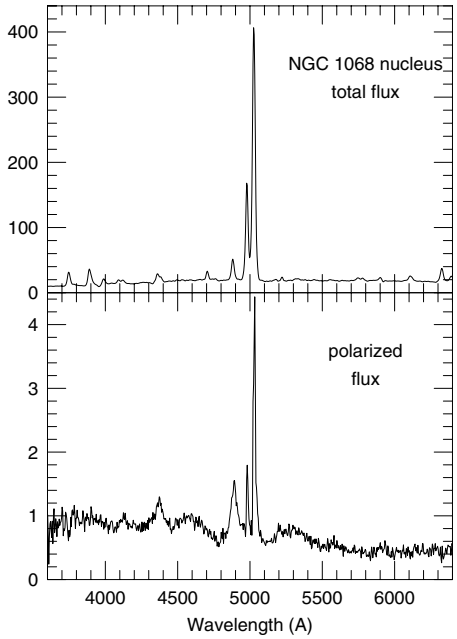


Figure 1. Spectropolarimetry of NGC 1068. In total flux (top panel) the spectrum shows relatively narrow emission lines. In polarized flux (bottom panel), however, the spectrum looks quite different, with broad permitted lines and strong blends of Fe II emission, both characteristic of high-density regions. In total flux NGC 1068 was classified as a Seyfert 2 galaxy, while in polarized flux it would be a Seyfert 1.

Applications of spectropolarimetry

$P(\lambda)$ is particularly powerful when emission or absorption features are expected to show some polarization signature. It allows understanding of the geometry and kinematics of regions separated by angular distances far smaller than the diffraction limit of the telescopes.

A dramatic example comes from the field of research into ACTIVE GALACTIC NUCLEI (AGN). In many AGN our view into the center of the nucleus is blocked. Yet the nucleus is bright enough that its light reflects off surrounding material—dust and electrons—that we can see. Since reflection generally polarizes light, the light from the blocked nucleus is polarized. Light generated in the outer, unblocked regions seen directly is unpolarized; hence spectropolarimetry provides a way of separating the two. Normal spectroscopy may be overwhelmed by light from the regions in direct view and show little evidence of the obscured region.

Such 'mirrors' into hidden regions will also impart their own signature on the polarized spectrum. Electrons, for example, have a scattering cross-section independent of wavelength and will not change the overall shape of the scattered spectrum. Hot electrons will also broaden spectral features. Dust grains tend to 'bluen' the spectrum, having a scattering cross-section that rises towards the blue. The velocity of the mirror can be revealed as a wavelength shift between polarized and unpolarized spectral lines.

These effects can be seen in figure 1, spectropolarimetry of the Seyfert 2 galaxy NGC 1068. A normal spectrum, shown in the top panel, displays relatively narrow emission lines of hydrogen, oxygen, nitrogen, sulfur, etc. The polarized flux spectrum at bottom shows a large broad component to the Balmer lines and strong emission of Fe II, something not obvious in the top spectrum.

If the scatterers are flowing toward or away from the light source, we have a 'moving mirror', in which the emission features in the source are Doppler shifted by two effects: the mirror moving relative to the source itself and the mirror moving relative to the observer. For an outflow velocity, $V(\text{out})$, of the scatterers, the observed Doppler shift is given by

$$V(\text{obs}) = V(\text{out}) - V(\text{out})\cos\theta(\text{obs})$$

where $\theta(\text{obs})$ is the observation angle, 0° if the mirror is moving directly towards us, 180° if it moves away from us. The first term is simply the mirror-source Doppler shift, while the second is the mirror–observer shift. Note that when the mirror is moving towards us the Doppler shift is small, and when the mirror moves away from us the shift is largest. Realistic situations, in which the outflow covers a range of angles, require integration of the polarized flux as a function of velocity, favoring velocities closer to $V(\text{out})$ because forward and back scattering generally show low polarization compared with scattering at 90°.

Some types of stars show features similar to active galaxies, albeit on a smaller scale. An interesting case is GL 1403, a cool, highly polarized carbon star. Figure 2 shows its unusual polarization spectrum. The polarization is high towards both the blue and the red, but drops to

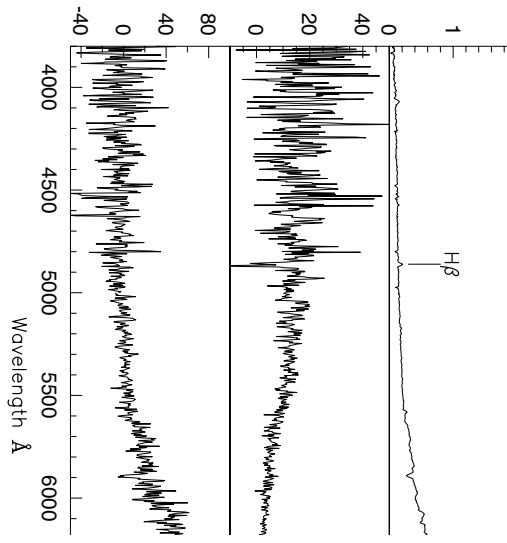


Figure 2. The very red carbon star GL 1403 shows strong polarization at both the b
between. Coincident with this drop is a dramatic swing in the angle of polarizatio
borne out by HST imaging. In the red one sees light reflected by the equatorial tor
dust in the torus becomes opaque, and instead one sees light dominated by scatter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

nearly zero in between. At this minimum, the polarization position angle swings through a nearly 90° change.

The interpretation of this spectrum is that in the blue we see scattering of light from the star, hidden by an equatorial torus, much the same as in some active galaxies. Towards the red, however, the dust in the torus becomes progressively less opaque, until we finally can see in to the stronger scattering occurring within the torus itself. The polar scattering seen in the blue occurs at a 90° angle to the scattering in the torus seen in the red, hence the position angle swing. High spatial resolution images using HST have confirmed that the blue image of the nebula is extended more or less orthogonally to the red image.

Spectropolarimetry has also been used to determine the parent stars of HERBIG–HARO OBJECTS AND EXCITING STARS (HH objects), small gas blobs ejected from the vicinity of young stars. Quite often the young stars are obscured by their parent molecular cloud. They also tend to lie in regions with many other stars, so that infrared surveys that might penetrate the surrounding dust in fact show a number of stars, any one of which might be the HH object's parent. When the spectrum of the HH object contains enough light reflected from its parent, the angle of polarization indicates the direction in which the parent lies (with a 180° ambiguity). Further, the polarized flux will give a spectrum of the hidden parent, providing some information on its spectral type, activity, etc. Note that measurements of the PROPER MOTION of the HH object supply similar directional clues, without the 180° ambiguity inherent in the polarization measurements. Of course, the proper motion data do not give any indication of the type of star.

Another common scenario is light passing through nonspherical dust grains, which are magnetically aligned by the local magnetic field (see INTERSTELLAR GRAINS). The asymmetry leads to scattering of the light that prefers one polarization sense over another. The transmitted light is hence polarized, at a level dependent on the shape, size and composition of the dust grains and the efficiency of the alignment mechanism. This 'dust transmission' polarization is accompanied by reddening, and in most cases is small. In our Galaxy an empirical upper limit to the polarization is given by $P(\max) \times 9\%\ E(B-V)$. As a function of wavelength the polarization shows a convex curvature, with the peak polarization typically occurring at visual wavelengths near 5500 Å. The peak can be shifted towards the red for dust grains larger than typical and to the blue for smaller grains. Except for interstellar absorption lines, the shape of $P(\lambda)$ is smooth and can be studied using multicolor imaging polarimetry as well as spectropolarimetry. However, there have been notable cases in which broadband polarimetry has misled researchers into thinking they were seeing polarization by interstellar dust, when in fact a far more complicated situation was revealed by spectropolarimetry.

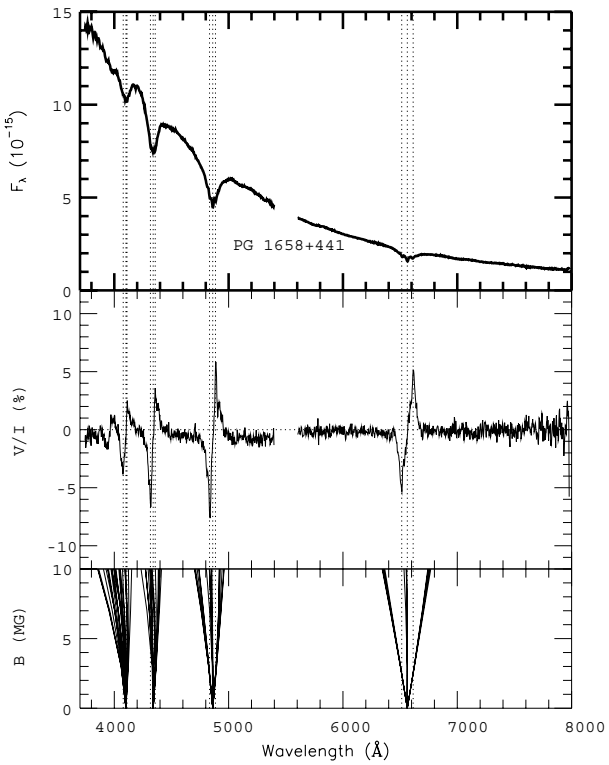


Figure 3. The white dwarf PG 1658 + 441 shows only broad Balmer absorption lines in total flux (top panel). Circular spectropolarimetry, however, shows that some of the broadening is actually due to the splitting of the lines in a strong magnetic field (middle panel). The amount of splitting can be estimated from theoretical models and can be used to measure the mean field strength (bottom panel).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A third source of polarization involves magnetic fields. As electrons spiral around magnetic field lines, their emission is polarized. If the electrons are relatively slow moving, cyclotron emission, both circularly and linearly polarized, is produced, while if the electrons are relativistic linearly polarized synchrotron emission results. Cyclotron emission is seen in white dwarfs with magnetic fields and synchrotron emission is often seen in the powerful nuclei of radio galaxies.

The magnetic fields in the atmospheres of some stars will also cause the splitting of absorption features in the stellar spectra (see STELLAR ATMOSPHERES). The presence of the magnetic field splits previously degenerate energy levels in the atoms, causing a single transition to break up into multiple transitions. Consequently the spectrum will show multiple components of normally single absorption lines. Depending on the geometry and details of the energy levels, some of these components will be circularly polarized in one sense, some in the other, and some will be linearly polarized. In the weak-field cases the absorption features simply look broadened, and spectropolarimetry can be used to cleanly separate the components of different polarization and determine the field strength (see figure 3). In high-field white dwarfs the lines can be shifted so far from their original wavelengths that polarization information is necessary even to identify them.

Polarization measurements are usually photon starved, meaning that the accuracy obtained is limited by the number of photons one can collect. The new generation of 10 m class telescopes is opening new realms of research using the powerful techniques of imaging polarimetry and spectropolarimetry, spurring renewed interest in a number of fields spanning much of astronomy and astrophysics.

Bibliography

Tinbergen J 1996 *Astronomical Polarimetry* (Cambridge: Cambridge University Press)

Robert W Goodrich

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polaris

The star α Ursae Minoris, the (north) Pole Star, currently situated 44.16′ from the north celestial pole. Owing to the effect of precession, Polaris is currently approaching the pole; its closest approach will be 27.52 in AD 2102. A pale yellow supergiant of spectral type F7:Ib–IIv, apparent magnitude 1.97, Polaris is at a distance of 430 light-years (parallax 0.008″, absolute magnitude 3.6). It is a Cepheid variable with a period of 3.97 days, magnitude range 1.9–2.1. Polaris is also a binary system, having a ninth-magnitude companion at a separation of 18.4″, a useful test object for small telescopes. Two fainter companions have also been detected.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polarization

The extent to which the vibrations of an electromagnetic wave display a non-random orientation. Light, and other forms of electromagnetic radiation, consists of a periodically varying electric and magnetic disturbance that travels through space as a transverse wave motion (a wave that vibrates perpendicular to the direction in which it is propagating, like a wave on water). A beam of light is said to be unpolarized if it contains waves vibrating with equal amplitudes (heights) in all directions perpendicular to the direction in which the beam is propagating. By contrast, a plane-polarized (or linearly polarized) beam contains waves vibrating in one plane only. If, as a wave advances, the direction of vibration rotates (clockwise or anticlockwise, with the same frequency as the wave itself) but the amplitude remains the same, it is circularly polarized. The most general case is elliptical polarization, in which both the amplitude and the orientation vary as the frequency of the radiation.

Where a beam contains a mixture of polarized and unpolarized radiation, it is said to be partially polarized, the degree of polarization being given by the ratio of the strength of the polarized component to the total strength of the beam. The state of polarization of a beam of radiation provides information on the mechanism responsible for producing the radiation at its source and/or on what has happened to that radiation between leaving the source and arriving at the observer. For example, thermal radiation and blackbody radiation are unpolarized, whereas synchrotron radiation (radiation emitted by charged particles moving at high speeds in magnetic fields) is highly polarized (elliptically or linearly). Light emitted from an ordinary star is usually unpolarized, but becomes partially polarized when it passes through clouds of interstellar dust grains. Radiation from pulsars and supernova remnants, which is produced largely by the synchrotron process, is strongly polarized.

See also: black-body radiation, electromagnetic radiation, frequency, pulsars, radiation, supernova remnants, synchrotron radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Polarization in Active Galaxies

In the absence of a strong magnetic field, the emission of photons is a random process and light is unpolarized. The mere presence of POLARIZATION therefore signifies asymmetry in the emission mechanism or light path through a source—structural details which in principle may be inferred from observations of sufficient quality. Since radiation sources in the nucleus of an ACTIVE GALAXY generally overwhelm all other components, these techniques can be extended to very high REDSHIFT where the entire object is unresolved. POLARIMETRY of active galactic nuclei (AGN) therefore offers the potential of not only discerning structural characteristics of individual objects, but also identifying relationships between various classes of objects and testing for evolution in the dimensions and activity level of the central engines.

Taxonomy of active galaxies

The classification of AGN reflects an historical background in which categories were defined more by observational than physical characteristics. The myriad of classes can be broken down according to two primary criteria: optical 'spectral type' and radio 'loudness' (see ACTIVE GALAXIES: OVERVIEW). Objects dominated by a featureless power-law continuum and broad emission lines ($\gtrsim$3000 km s^{-1} full width at half-maximum) are termed spectral Type 1, while those exhibiting a prominent starlight component and narrow lines are Type 2. The division by radio power commonly considers the ratio $R = F(5\,\mathrm{GHz})/F(4400\,\text{Å})$, with $\log R \gtrsim 2$ indicating a radio-loud object and $\log R \lesssim 0.5$ radio-quiet. As indicated in table 1, QSOs, Seyfert 1 and broad-line radio galaxies (BLRGs) are Type 1, while Seyfert 2 and narrow-line radio galaxies (NLRGs) are Type 2 examples.

Table 1. AGN classification*.

	Radio-quiet	Radio-loud
Type 1		
Strong featureless cont.	Seyfert 1	BLRG
Broad permitted lines	**QSO**	**Quasar**
Narrow forbidden lines		
Type 2		
Weak featureless cont.	Seyfert 2	NLRG
Galaxy starlight	**IRAS QSO**	**Hi-z RG**
Narrow forbidden lines		
Blazar		
Variable polarized cont.	Don't exist?	**Blazar**
BALQSO		
Type 1 +	**BALQSO**	Don't
Broad absorption lines		exist?

* Boldface indicates objects with $L \gtrsim 5 \times 10^{12} L_{\odot}$.

The unified scheme

Beginning in the late 1970s, evidence began to point to relationships between certain classes of AGN. For example, it was found that SUPERLUMINAL MOTION is confined to objects which exhibit strong, highly variable, polarization—the BLAZARS. SEYFERT 1 GALAXIES and QSOs (QUASISTELLAR OBJECTS) seemed to form something of a luminosity continuum, as did RADIO GALAXIES and QUASARS. However, the breakthrough in understanding came with Miller and Antonucci's detection in 1985 of polarized broad emission lines in the prototype Seyfert 2 galaxy NGC 1068. This remarkable discovery established two crucial points: (a) at least some objects displaying Type 2 spectra contain broad emission-line regions, and (b) the fact that these regions can be detected in light polarized by scattering implies that other vantage points (i.e. those in the direction of the scatterers) would observe an object with prominent broad lines (Type 1). The observations quickly led to models which 'unified' broad- and narrow-line AGN into a single type of object whose observed properties depend on viewing aspect (see ACTIVE GALAXIES: UNIFIED MODEL).

The key element of the unified scheme (figure 1) is a dense dusty/molecular torus which obscures the central engine and compact broad emission-line region for low-inclination sightlines. Small portions of the intense featureless continuum and broad emission-line spectrum of the internal QSO that escape through the polar hole of the torus is scattered by small particles (electrons and/or dust grains); this reflected component is linearly polarized. Evidence for a torus as opposed to other geometries comes from 'ionization cones' revealed by emission-line imaging and which are thought to arise from gas within the toroidal opening illuminated by the nuclear UV radiation field. Imaging polarimetry of favorable examples also depicts highly polarized centrosymmetric scattering regions (figure 2) very similar to the bipolar reflection nebulae observed in our own Galaxy. Finally, direct images of some nearby AGN have uncovered toroidal structures several hundred LY (light-years) across—these may be large-scale extensions of the structures which absorb much of the nuclear radiation and define the opening angle of the escaping radiation.

The unified scheme represents an enormous improvement over previous pictures; indeed the polarization properties of essentially all AGN are explicable in this context. However, it is important that the picture should not be over-interpreted. The scheme is perhaps better regarded as a valuable tool rather than a physical model—a starting point from which many mysteries of the nature of AGN must still be explained.

Polarization of AGN

The critical parameter of the unified scheme—in radio-quiet and radio-loud objects alike—is the inclination of the line of sight to the toroidal axis.

High-inclination objects (Type 2)

Type 2 spectra are measured for near edge-on examples. In this case the optical spectrum is dominated by starlight from the host galaxy and by the extended narrow-line emission region. The direct view of the nucleus is highly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

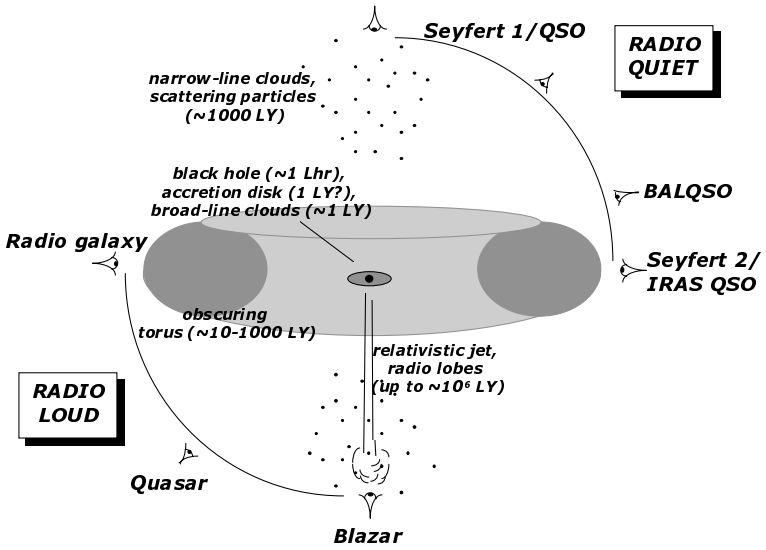


Figure 1. The unified scheme for AGN, indicating object class as a function of viewpoint. For objects with Type 2 spectra (e.g. Seyfert 2s and IRAS QSOs) where the central continuum and broad emission-line source are hidden from direct view, nuclear light is visible only as a component polarized by scattering off polar clouds of small particles. The degree of polarization declines with decreasing inclination and as the nucleus becomes progressively more exposed for the BALQSOs and objects displaying Type 1 spectra. In nearly pole-on radio-loud examples (Blazars), an additional highly polarized, variable component appears due to beamed synchrotron emission from a relativistic jet directed nearly along the line of sight. Drawing is not to scale.

extinguished at wavelengths shorter than 1 μm, and re-radiation by the dusty torus in the mid-infrared ($10 - 30$ μm) is prominent relative to the optical light. Narrow-line radio and Seyfert 2 galaxies are viewed at these angles, as are the luminous IRAS QSOs selected by the Infrared Astronomical Satellite (IRAS).

Light escaping through the polar hole of the torus is scattered at large angles, resulting in high linear polarization. With the dusty torus blocking the nucleus that would otherwise dilute the scattered light, the net polarization of the object can exceed 15%. Because the scattering particles are mixed with the extended narrow-line emitting gas, the broad emission lines and continuum share the same polarization, but the narrow lines are unpolarized. Finally, lack of dilution from the central source results in narrow emission lines which are overemphasized in total flux spectra (figure 3).

Low-inclination objects (Type 1)

At a lower-inclination sightline, our view is unhindered by the torus, and a Type 1 spectrum is measured. Thus, optical surveys which discriminate according to near-UV excess or strong, broad emission lines catalog primarily classical broad-line QSOs and Seyfert 1 galaxies. The overall polarization is found to be weak, $P \lesssim 2\%$, and spectropolarimetry of the emission lines reveals a range of properties: polarization which is weaker, but at the same position angle θ as the surrounding continuum; a position angle which differs from the continuum value; even a rotation in θ through the lines. The strongly polarized but faint scattered component is presumably

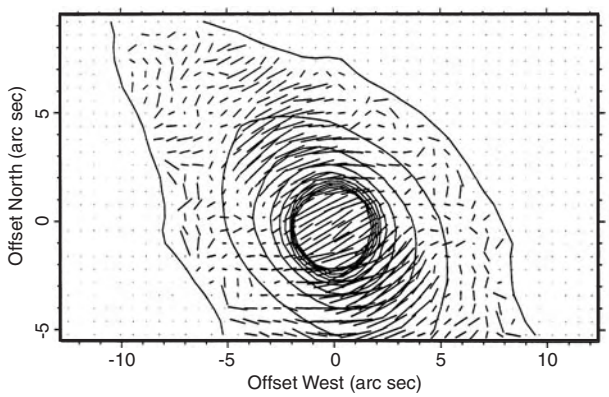


Figure 2. Imaging polarimetry of the nucleus of NGC 1068 at 2.2 μm. A scattering bicone is evident in the centrosymmetric polarization pattern represented by the (pseudo)-vectors. Contours represent the total intensity. (After figure 3 of Packam C *et al* 1997 Near-infrared imaging polarimetry of NGC 1068 *Mon. Not. R. Astron. Soc.* **288** 375.)

still present, but it is diluted to obscurity by the brilliant featureless continuum and broad emission-line spectrum. Instead, the observed polarization of Type 1 objects is thought to originate in the heart of the nucleus: through scattering off electrons in the broad-line clouds, scattering off the inner torus or the accretion disk which feeds the supermassive black hole, and in the radio-loud objects, possibly residual synchrotron emission from an inclined jet (see below). Much like the observations that led to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

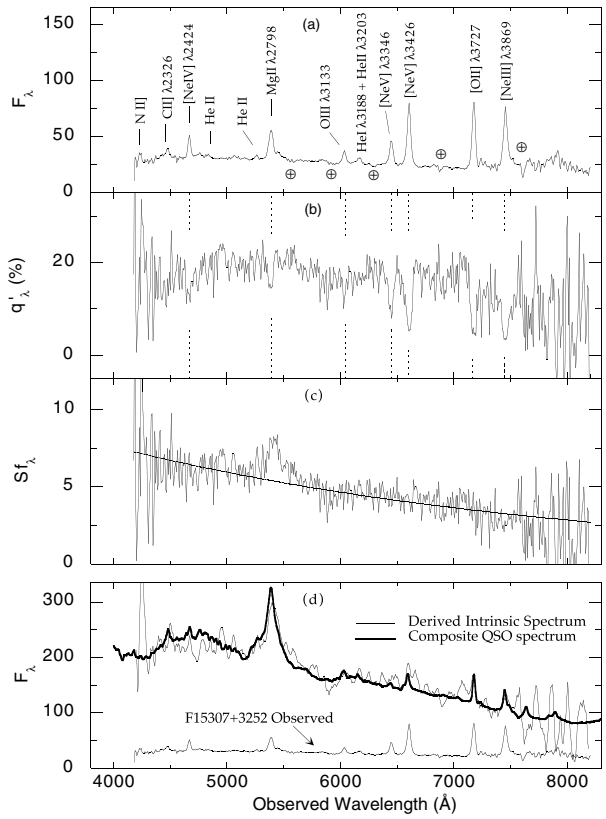


Figure 3. Spectropolarimetry of the IRAS QSO F15307+3252. (*a*) The observed (Type 2) total flux spectrum. (*b*) The degree of polarization. (*c*) The polarized flux with same units as for (*a*). (*d*) Comparison of the 'intrinsic' nuclear spectrum of F15307+3252 with that of a composite quasar (heavy line). The intrinsic spectrum was constructed by summing the polarized flux and observed narrow emission-line spectra, assuming isotropic narrow-line emission with equivalent widths typical of a luminous QSO. The smooth curve in (*c*) is a power-law of slope typical of unobscured Type 1 spectra.

the unified scheme itself, detailed polarimetric studies of these objects may dramatically revise our thinking about the innermost regions of the central engine.

Transition objects

Evidence has accumulated over several years suggesting that QSOs which show broad absorption lines displaced to the blue of the resonance emission lines represent objects seen from a vantage point that just skims the surface of the obscuring torus. Emission-line spectra of these so-called broad absorption-line QSOs (BALQSOs) closely resemble unobscured QSOs, but extinction by intervening dust is indicated by the significant reddening exhibited by certain examples (see also QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES). In some models, the absorbing clouds represent material ablated off the dense torus and radiatively accelerated to outflow velocities of at least 20 000 km s^{-1}.

Polarimetric evidence supports the interpretation of BALQSOs as 'transition' objects between Type 1 and Type 2. Not only do they exhibit a degree of polarization which is intermediate between the types, but spectropolarimetry indicates that we are presented with multiple light paths from the nucleus—a direct route which provides a large portion of the continuum but is heavily extinguished in the absorption features, a polarized continuum component which is presumably the same scattered light that is seen in the edge-on examples, and the largely unpolarized broad emission lines whose larger emitting area is not completely occulted by the absorbing clouds.

Pole-on radio-loud sources

The event which originally focused the attention of polarimetrists on AGN was the discovery in the early 1970s of strong linear polarization in BL Lac (BL LACERTAE) and several other compact radio sources. The property was quickly attributed to non-thermal (synchrotron) emission of relativistic electrons in a magnetic field, and a common opinion held that the objects, which collectively became known as Blazars, represented 'naked' QSOs where the central engine is viewed directly. This static picture led to many difficulties in explaining the short variability timescales, high apparent luminosities and superluminal motion—difficulties which Blandford and Rees proposed would be largely alleviated if the emission occurred in a relativistic plasma jet whose velocity v is directed roughly along our line of sight. Doppler boosting enhances the brightness within a cone of opening angle $\theta \sim \sqrt{2/\Gamma}$, where $\Gamma = 1/\sqrt{1 - v^2/c^2}$, leading to more than a hundred-fold increase in brightness on-axis for modest Doppler factors ($\Gamma \sim 5$). The polarization of optically thin synchrotron emission is almost entirely linear, with

$$P = \frac{\alpha + 1}{\alpha + 5/3} = 75\%$$

for a uniform magnetic field and power-law emission spectrum characterized by $\alpha = 1$ ($F_\nu \propto \nu^{-\alpha}$). The position angle is perpendicular to the projection of the magnetic field axis on the plane of the sky. Because of the many physical attributes of the beamed picture and its natural place in the unified scheme (figure 1), it is nearly universally adopted today.

Unfortunately, it is proving difficult to extract physical details of the central engine from the enormous observed polarizations (up to 40%) and dramatic variability of Blazars. The beamed picture means that modest changes in the injection direction into a jet or of the magnetic field arrangement within a plasma blob can translate into wild variations in P and θ at the observer (see figure 4); in this sense the problem is one of 'the tail wagging the dog'. Time dilation further complicates the situation. Certain sources exhibit 'preferred' position angles of polarization which are often thought to indicate the general projection of the jet axis on the sky. BL Lac (figure 4) commonly exhibits a position angle which hovers around $\theta \sim 30°$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

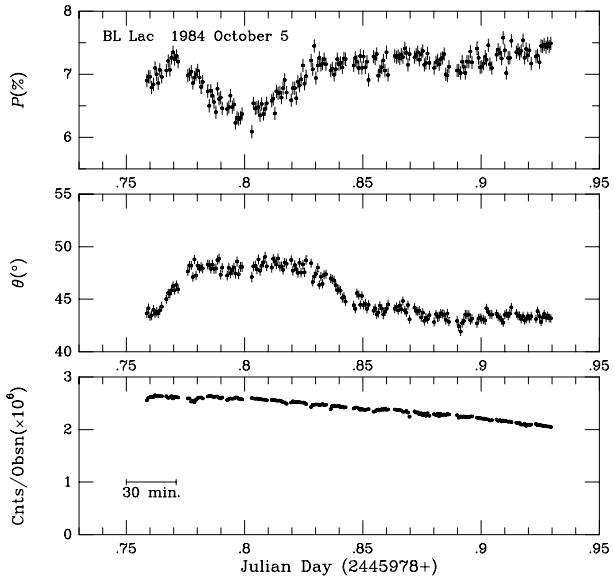


Figure 4. Optical polarization of the synchrotron source BL Lac recorded at 1 minute time resolution. The 'event' centered near fractional JD = 0.8 can be represented by an emission component which rotates by 90°—and back!—in an interval of 2 h. The slow decline in brightness is due to increasing airmass as the object sets in the sky.

Remaining mysteries

The unified scheme has proven qualitatively correct for both radio-quiet and radio-loud AGN over at least eight orders of magnitude in luminosity. It is a tool which can be used to disentangle pure orientation from fundamental properties of the nucleus and eventually lead to 3D models. From the combination of multiwavelength imaging and spectroscopic observations, it is now possible to infer the sizes and structures of obscuring tori, and to measure scattering cone opening angles, accretion rates and black hole masses. Polarimetric observations formed the foundation of the unified scheme, and polarimetry plays a key role in these studies as well, particularly at distances where the nuclei cannot be resolved directly.

We have still not solved the radio-quiet/radio-loud dichotomy. Progress in identifying this essential ingredient of the central engine will likely require much improved theories of the fate of magnetic fields in the innermost regions of the accretion disk and their role in launching relativistic jets. Interferometric and polarimetric observations at the shortest possible wavelengths will probe deep into the base of the jets and lend clues to their formation. For some as yet unknown reason, a correlation between radio loudness and galactic morphology persists. That puzzle touches upon the overriding question of the feeding and evolution of AGN, and of how they form in the first place. The historical record of this field has shown that answers do not come easily. When they do, as in the polarimetric catalysing of the unified scheme, they can appear from the most unlikely corners.

Bibliography

Comprehensive reviews of the unified scheme for AGN are given in:

Antonucci R 1993 Unified models for active galactic nuclei and quasars *Ann. Rev. Astron. Astrophys.* **31** 473

A similar review for Blazars can be found in:

Urry C M and Padovani P 1995 Unified schemes for radio-loud active galactic nuclei *Publ. Astron. Soc. Pacific* **107** 803

An example of the power of polarimetry for understanding the nature of extremely luminous infrared galaxies and the identification of the class of Type 2 QSOs is given in:

Hines D C *et al* 1995 The buried QSO in IRAS F15307+3252 and the nature of hyperluminous infrared galaxies *Astrophys. J.* **450** L1

Gary Schmidt and Dean Hines

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pollux

The star β Geminorum, an orange giant of spectral class K0IIIvar. At apparent magnitude 1.16 it is the brightest star in Gemini, suggesting that it may have brightened (or Castor faded) since the Bayer letters were allocated in 1603. Pollux is rather closer than Castor, at 34 light-years (parallax 0.097″); its absolute magnitude is 1.1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pond, John (1767–1836)

As an amateur astronomer, Pond had used the mural quadrant at Greenwich and analysed its inaccuracies. He participated in the committee which supervised its replacement by a 6 foot mural circle by EDWARD TROUGHTON. When Maskelyne died in office, Pond was appointed to succeed him and, as sixth Astronomer Royal, concentrated on upgrading the instrumentation at the Royal Observatory. Its measurements of star positions and time were raised to an accuracy never before achieved. During his administration the observatory was taken over by the Royal Navy, for which the observatory regulated chronometers, and he instituted in 1833 the public dissemination of time through the dropping of a time ball at 1 p.m. each day (this still happens). What gave under the pressure of all this new work was the *Nautical Almanac*, and it fell into disrepute, being transferred for recovery to the responsibility of a separate office, called the Nautical Almanac Office.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pons, Jean-Louis (1761–1831)

Born in Peyre, France, joined the Observatory at Marseilles as concierge, becoming an astronomer there. Like Australian supernova-finder Robert Evans, he remembered the star fields that he observed and recognized changes; this enabled him to discover comets at the rate of about one per year from 1801 until 1827. He suggested to J F ENCKE that a comet found by him was one already discovered by Encke in 1805. Calculating a period of 3.3 years, Encke successfully predicted its return for 1822. This comet is known as 'Encke's Comet', but Encke himself referred to it as 'Pons' Comet'. Pons became director of the observatory at Marlia (Italy), and director of the Florence Observatory. 26 comets bear his name, but he discovered or co-discovered up to 37.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Population I Stars

The relatively young, metal-rich stars that are found in the disks and spiral arms of spiral galaxies. Originally differentiated from Population II stars by the German-born American astronomer Walter Baade (1893–1960) in 1944, on the basis of their metallicity and spatial distribution, they are now seen as representing stars that have formed in the latest stages of galaxies' morphological and chemical evolution. Their orbits about the galactic center have a low eccentricity and a low inclination to the galactic plane.

The oldest Population I stars are fairly strongly concentrated towards the galactic plane, where their spatial distribution overlaps with the youngest of Population II. Intermediate Population I stars have approximately circular orbits, are strongly concentrated along the galactic plane, and have an irregular distribution. The Sun belongs to this population, together with the stars in older open clusters, many A stars, giants, and emission-line (Me) dwarfs. Extreme Population I stars have an extremely uneven distribution within galaxies' spiral arms and have the greatest abundance of heavy elements. They include young stars in the process of formation, T-Tauri stars, OB associations, stars of spectral classes O and B, supergiants, classical Cepheids, and young open clusters.

See also: Population II stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Population II Stars

Old red stars that occur throughout elliptical and lenticular galaxies, and in the haloes, globular clusters and central bulges of spiral galaxies. They differ from Population I stars in having a very low metal content, and those heavier elements that they do contain are thought to have been produced by massive, short-lived earlier stars which form a hypothetical Population III. Population II stars are now seen as forming a continuum with the younger Population I stars, although, for convenience, various subclasses are recognized. Their respective distribution may be related to the gradual concentration of material towards the galactic plane that occurs in galactic evolution, together with the chemical evolution of stars themselves, and hence of the interstellar medium.

The oldest stars with the lowest metal content have an essentially spherical distribution, and are known as Halo Population II stars. This group includes the stars in globular clusters, long-period RR Lyrae stars (period >0.4 days), the W Virginis stars (also known as Type II Cepheids), and certain metal-poor subdwarfs. The stars of Intermediate Population II, which include most high-velocity stars and many Mira-type variables, are also metal-poor, but are more strongly concentrated towards the galactic plane. Both subclasses show an increasing spatial density towards the galactic center.

The youngest Population II stars are those of the galactic bulge, where they intermingle with the very oldest Population-I stars, and form part of the Disk Population. They include short-period RR Lyrae stars (period <0.4 days). The orbital eccentricity and inclination to the galactic plane show a decrease from Halo Population II stars to the Disk Population.

See also: high-velocity star, Mira variables, Population I stars, Population III stars, RR Lyrae stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Population III Stars

A hypothetical population of stars that may have existed in the very early universe. As originally proposed, these extremely massive stars would have formed before the earliest galaxies, and have been responsible for the 'primordial' helium content of the universe and also for a large part of the cosmic microwave background radiation. This hypothesis does not agree with the currently generally accepted theory of the Big Bang, nor with 'top-down' theories of galaxy and star formation. Subsequently, the term has been used for an early population of massive stars, long extinct, that would have exploded as supernovae, and thus account for the metal content of the Population II stars in globular clusters. The remnants of this early population would be neutron stars or black holes, and could thus account for some of the dark matter postulated to explain the missing mass in galactic haloes.

See also: cosmic microwave background, Population II stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Porter, Russell (1871–1949)

Arctic explorer, astronomer, telescope-maker; born in Springfield, VT, helped to design the 200 in Mount Palomar telescope which he illustrated in the book *Giant of Palomar*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Poseidonius of Apamea (*c*. 135–*c*. 50 BC)

Greek Stoic philosopher born in Apamea, Syria, settled at Rhodes, and from the rather imprecise observation that Canopus just rose above the horizon at Rhodes, but was 7 degrees 30 min above the horizon at Alexandria, he determined a too-low value for the circumference of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Position Angle

A measure of the direction of one object relative to another on the celestial sphere, for example the direction of one component (usually the fainter) of a double star from its companion. Together with the separation, this defines the observed form of the system, and such measurements repeated over many years enable any binary motion to be determined.

Position angle, usually abbreviated 'PA', has been defined in the literature in various ways, often confusing and sometimes mutually contradictory. The best definition of the system, now accepted as standard, is 'position angle is measured anticlockwise as seen on the sky, from the $+\delta$ direction' (the '$+\delta$ direction' being the direction of increasing declination, i.e. northwards).

Position angle may also be used in recording the location of an observed feature on the disk of a planet or the Sun, or to define the direction of an axis of rotation. In such cases, PA is measured eastwards from a line joining the center of the disk to the direction of the north point.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Positron

The antiparticle of the electron. It has the same mass as the electron but carries unit positive charge (1.602×10^{-19} coulomb). Like the electron, it has a spin of 1/2 (in units of the Planck constant, h, divided by 2π) but when an electron–positron pair is created, the positron will have the opposite spin to the electron. Electron–positron pairs can be produced from high-energy gamma-ray photons. Conversely, when electrons and positrons collide, they liberate energy on the form of gamma-ray photons. Where the electrons and positrons have been accelerated to very high speeds and kinetic energies, for example in particle accelerators, the energy released when they collide and mutually annihilate may be high enough to create hadrons and quarks.

See also: antimatter, electron, gamma rays, hadrons, photon, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Poynting–Robertson Effect

A net force acting on small particles which causes their orbits to decay, also known as *Poynting–Robertson drag*. When photons of light strike a dust particle in orbit around the Sun, they impinge not directly 'side-on' but slightly on its leading side (the 'forward' side in terms of its orbital motion). Their energy is absorbed, and subsequently re-radiated, but it is re-radiated isotropically—in all directions. The transfer of energy from photon to particle can be thought of as the application of a force to the particle. There is a component of this force that acts in the opposite direction to the tangential component of the particle's orbital motion which is not compensated for when the absorbed energy is re-radiated. As a result the particle loses kinetic energy, which reduces its velocity, so it spirals inward, toward the Sun. This simple picture is complicated by, among other things, the fact that particles have different shapes and spins; a ring of dust particles of the same mean radius but with a distribution of shapes would gradually broaden out. The decay time is shorter for smaller and low-density particles. The effect can be counteracted by radiation pressure, more so with decreasing particle size below about 1 μm. The Poynting–Robertson effect also acts on particles in planetary ring systems. John Poynting was the first to describe the effect, in 1903; Howard Robertson derived it from relativity theory in 1937.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Poznań Astronomical Observatory

This Poznań Astronomical Observatory is a unit of the Adam Mickiewicz University, located in Poznań, Poland. From its foundation in 1919, it has specialized in astrometry and celestial mechanics (reference frames, dynamics of satellites and small solar system bodies). Recently, research activities have also included planetary and stellar astrophysics (asteroid photometry, catalysmic binaries). The scientific staff consists of twelve researchers and six PhD students. In 1949, a new minor planet was discovered at the Observatory, receiving the name 1572 Posnania. Nowadays, due to light pollution, observations are carried out abroad as well as at the 0.4 m photometric telescope located outside the city.

For further information see http://www.astro.amu.edu.pl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Precession

The slow periodic change in the orientation of the Earth's axis of rotation due to gravitational attractions of the Sun and Moon on the non-spherical globe (geoid) of the Earth. The effect is often referred to as luni-solar precession. The term is also applied to the periodic change in the rotation axis of any spinning body, for example, an artificial satellite. The axis of the Earth is inclined to the perpendicular to the ecliptic by an angle of 23.44° and precession causes the celestial pole to trace out a circle of this radius around the pole of the ecliptic in a period of some 25 800 years. Thus the north celestial pole is at present located close to the fairly conspicuous star Polaris, known as the Pole Star. It lay close to the star Thuban in the constellation Draco 4500 years ago. In 12 000 years time, it will be near the bright star Vega.

As the celestial pole moves, so the celestial equator moves relative to the ecliptic. Thus the points of intersection of these two circles, in other words the equinoxes, move along the ecliptic in a clockwise direction at a mean rate of about 50 seconds of arc per annum (i.e. 360× in 25 800 years). Since the celestial longitude of a star is measured from the vernal equinox, it follows that this must increase at a rate of about 50 arcseconds per annum due to precession.

The gravitational attractions of the planets cause a very slow charge in the orientation of the ecliptic (the plane of the Earth's orbit) which causes the right ascensions of stars to decrease by about 0.13 arcsec per annum and which is known as planetary precession. The combined effects of luni-solar and planetary precession are called general precession.

Because of precession, star catalogs and charts must be drawn up for a particular equinox (e.g. 00 h on 1 January for the years 1900, 1950, 2000 etc). If a precise position for a star is required at some other time, the effects of precession must be calculated to determine by how much the stars coordinates will have changed from its tabulated position.

Hipparchus first noted the effects of precession in the second century BC.

See also: rotation, equinox.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Prehistoric Astronomy

For many decades arguments have raged about the nature and level of sophistication of astronomical achievements among pre-literate cultures in the Old World and among pre-Columbian peoples in the New. The often-heated debates have tended to revolve around some claim that astronomical knowledge reached a remarkable degree of sophistication among a particular group of prehistoric people—for example, the people who built and used conspicuous stone monuments during the Neolithic and earlier part of the Bronze Age in Britain and Ireland—and sceptical reactions by archaeologists and anthropologists.

A fundamental problem that is still not always recognized is that to talk about levels of sophistication is already to measure prehistoric 'achievements' against our own and by implication to see scientific astronomy in the modern Western world as a pinnacle of achievement with all other developments in understanding either as steps along the road towards it or as dead ends along the way. Many archaeologists and anthropologists still feel that to speak of prehistoric 'astronomy' at all is to fall into this very trap. It is crucial to appreciate at the outset that many other 'astronomies' are simply different, and if we are to attempt to understand and evaluate them in their own terms—that is, if we are to study what celestial bodies and events meant to different groups of prehistoric people and how astronomical 'knowledge' was used—then we must attempt to discover something of the wider frameworks of understanding within which the sky was observed and perceived by those people.

Such world-views or 'cosmologies' are likely to have been very different from the Western perspective, as is evident from the many examples still existing in the world today among a variety of indigenous peoples. Things that from a Western point of view seem widely separated and to bear no relationship to one another may in other world-views be understood as inextricably linked. Indigenous cosmologies generally draw no clear distinction between terrestrial and celestial realms, between sacred and mundane activities, between animate and inanimate objects or between (what to us is) empirically real and what is imagined: thus, to the Barasana of the Colombian Amazon, 'Old Star', chief protector of the inhabitants of the sacred *He* world, is at once a human warrior, the fierce thunder jaguar, a constellation corresponding to Orion and a short trumpet.

Space and time in the Western view are abstractions: backdrops upon which objects and events can be placed and plotted. To indigenous communities, on the other hand, particular places may acquire specific meanings and significance for any number of reasons, and spatial perceptions of the world are developed, in very particular ways, in relation to such places. Similarly, the timing of various human actions tends to be defined in terms of natural events and cycles, celestial ones included. In some measure this is for very practical reasons to do with subsistence, but it is also the case that where and when certain activities take place tend to be defined by the perceived need to keep human action in harmony with the COSMOS as it was perceived and understood. It is this propensity that enables archaeologists to extrapolate from spatial patterning in the material record to make deductions about prehistoric cosmologies.

It is in this context that astronomy is of particular importance to the modern archaeologist, because we can reliably reconstruct recurrent elements of prehistoric skies. Thus celestial correlations in the archaeological record (such as the alignment of buildings on solstitial sunrise or sunset) are easily recognizable as such, whereas many other referents that might have been crucial in determining the location or orientation of a house, temple or tomb are no longer detectable. On the other hand, an astronomical alignment could easily have arisen fortuitously, as a result of factors quite unrelated to astronomy, and where there is no independent evidence to support the astronomical interpretation it often takes the recurrence of a similar phenomenon among a group of related structures (e.g. a consistent pattern of astronomical alignment among a group of similar monuments constructed in a local area at a particular stage in prehistory) to provide convincing evidence that astronomical considerations were indeed important.

It is impossible to make sweeping generalities about 'prehistoric astronomy'. For a start, the pre- (and 'proto'-) historic societies about which one might speak differ hugely from one another, spanning a time period from the Upper Palaeolithic ca 30 000 BC to times immediately prior to European contact in the Americas and elsewhere. They also vary from small groups of hunter-gatherers to complex city-states such as those of the Maya, Aztec and Inca which, unlike the urban states of the Near and Middle East, were in no way precursors of the Western tradition. Furthermore, there can be huge disparities between the nature of world-views even of communities of a similar nature in similar locations, as is evident from certain indigenous groups living in relative proximity today.

The earliest claim for astronomical observations relates to the Upper Palaeolithic, the period when fine depictions of animals first appeared in western European caves. A large number of engraved portable artefacts such as pieces of bone and antler are also known from this period. One, part of an eagle's wing from a cave at Abri Blanchard in the Dordogne valley, France, contains a series of notched marks in a serpentine pattern. In the early 1970s Alexander Marshack concluded from microscopic studies of these marks that they had been accumulated over a period of time. Furthermore, the numbers of marks in different parts of the pattern suggested that they were tallies of days in the lunar phase cycle. Others have questioned these conclusions, and they remain controversial.

Marshack's claims aside, there is little direct evidence of any interest in the skies among hunter-gatherer communities during the Upper Palaeolithic and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Mesolithic. However, the evidence from modern hunter-gatherer groups such as Australian aboriginals suggests this should not be taken to indicate a lack of interest in the sky or that world-views of considerable complexity were not developed by such communities; it is merely a reflection of the paucity of the material evidence relating to such groups.

It is in the subsequent Neolithic period, when early farmers around the Atlantic fringes of Europe started to construct the large communal tombs that still remain in considerable numbers as conspicuous monuments in the modern landscape, that substantive evidence relating to prehistoric astronomy first appears in many areas. It is clear that orientation was important as early as the fifth and fourth millennia BC: regional trends in tomb orientations are sufficiently strong and distinctive in many places that simple factors such as local topography or prevailing wind may be ruled out as a satisfactory explanation. Regional orientation trends are also evident among groups of communal long-houses built in the earlier part of the Neolithic across northern Europe. Each spread of orientations has a distinctive 'signature', many of which seem to be related in a general way to the motions of the Sun: for example spreads between north-east and south are quite common, and may reflect a need for tomb entrances to face the rising Sun or the Sun climbing in the sky. A similar remark holds for tombs and other forms of stone monument such as the megalithic rings and rows built during the Neolithic and well into the Bronze Age, perhaps as late as 1000 BC in remote parts of Scotland and Ireland. The orientation signatures of a few of the later groups such as the Scottish 'recumbent stone circles' and short stone rows in south-west Ireland are strongly suggestive that it was the Moon rather than the Sun that was of primary importance for their builders and perhaps regulated the timing of rituals or ceremonials held there.

This picture is very different from that put forward in the 1960s and 1970s by ALEXANDER THOM, who propounded the idea that many standing stone monuments incorporated intentional high-precision alignments on the Sun and Moon, reflecting their use as high-precision 'observatories'. These ideas did not stand up to close scrutiny, the main problem being the fair selection of data. Nonetheless, certain Neolithic chambered tombs do seem to have incorporated more specific solar alignments, a famous example being Newgrange in Ireland, where a specially constructed roof-box above the entrance permitted sunlight to enter the tomb for a few minutes after dawn on days around the winter solstice, shining down the entire length of the 19 m long passage to light up the bones of the dead in the central chambers. The monument was not an observatory in any modern sense of the word; the alignment is probably best seen as a symbolic expression of a relationship between ancestors and the Sun that was of great significance to the builders because of the way they understood their own place, and that of their ancestors, in the cosmos. Similarly, Stonehenge—which for some 500 years had existed as a unexceptional ditch-and-bank enclosure in which various timber structures had been constructed—was reoriented on the solstitial Sun at around the time, ca 2500 BC, when it was transformed into the stone edifice whose remains are so famous today. It could well be this realignment helped to strengthen its symbolic power by legitimizing its place at the center of the perceived cosmos.

In the European Iron Age, the evidence shifts to domestic structures such as British roundhouses, which not only manifest distinctive orientation trends but where there is evidence to suggest that certain activities may only have been appropriate in particular quarters of the house, paralleling what is known from various indigenous groups in north America. It seems too that certain places were preferred when disposing of the dead, of animals or of other objects, also reflecting symbolic or cosmological principles. Celtic cult sites also show distinctive orientation trends.

In other parts of the world astronomies developed in distinctive ways. Polynesian navigators, for example, developed a practical astronomy based on 36 linear star-to-star 'constellations' that formed a three-dimensional 'star-compass' enabling navigators to locate distant islands. Central to Inca cosmology was the system of *ceques*, conceptually straight lines radiating outwards from the Coricancha temple in the Incaic capital Cuzco, which at once reflected Inca knowledge of the sky, dictated the layout of the city and organized social groups within it, and functioned as a calendar. Life in the Mayan world revolved around a multitude of ritual acts which were dictated by the calendar, and in particular by a repeating cycle of 260 days used widely in pre-Columbian Mesoamerica, each day of which had particular symbolic associations and omens. However, the Mayans, who had developed a base 20 system of numeration together with a form of hieroglyphic writing, also developed a quantitative, predictive astronomy capable of forecasting eclipses, the appearances of Venus and Mars, and the passage of the Moon through the zodiac. They were prepossessed with the passage of time, always trying to equate integer multiples of different cycles: thus 73 cycles of the 260 day count equals 52 years of 365 days, and 405 lunations (11 960 days) equals 46 cycles of the 260 day count equals (actually to within 3 h). This prepossession developed not in a spirit of abstract enquiry but in the context of a world-view in which the ruling class communicated with, and were legitimized by, the gods; there was a cosmic 'right to rule' that needed to be constantly reinforced by keeping their activities in tune with various celestial cycles (see MAYAN ASTRONOMY).

Many investigations of prehistoric astronomies now take place within the framework of archaeoastronomy, an interdiscipline which arose in the 1970s as a result of the need to resolve the cross-disciplinary disputes raised by the work of Alexander Thom on the British megaliths. In north-west Europe, archaeoastronomy was mainly concerned with statistical appraisals of putative astronomical alignments amongst prehistoric

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stone monuments, but in the Americas it quickly began to assess such alignments in the context of a wider variety of cultural evidence bearing on the nature and meaning of astronomies in pre-conquest times, such as historic accounts, ethnographic accounts of modern descendants of pre-Columbian peoples and the written record in the form of codices, inscriptions and iconography.

Archaeoastronomers have moved on from the days when they simply looked for astronomical alignments in prehistoric architecture—a sort of celestial 'butterfly collecting' which fails to consider the social context in which prehistoric astronomies might have operated and to address the cognitive issues that are vital for their interpretation. This has been replaced by a much broader approach which tends to highlight particular themes such as concepts of space and time, calendar development and the place of celestial referents in different schemes of 'sacred geography' which are the concrete reflection in the physical landscape of the cosmologies which structured human action within the world as prehistoric communities perceived it.

The sky is a symbolic resource of immense significance, not only because many of its cycles are convenient and reliable for regulating human activity, but also because its immutability confers power on those individuals who can convince others that they wield control over celestial cycles and events. Most archaeologists expect that the meanings attached to what was seen in the sky were highly specific, varying greatly between different cultural contexts. This specificity works against the need to seek repeated instances of patterning in the material record in order to be better assured that they were meaningful to the people who created them. As we have said, a solstitial alignment could easily have arisen fortuitously, as a result of factors quite unrelated to astronomy. On the other hand there is evidence to suggest that traditions of public ritual tend to change slowly, even in the face of social upheaval, and it is this—reinforced by clear trends in the archaeological record such as the strong patterning in sets of orientations of prehistoric tombs in many parts of Europe from Neolithic times onwards—that most clearly suggests the existence of a great deal of evidence yet to be explored that will give us new insights into prehistoric astronomies, and into the broader frameworks of cosmology in which they operated.

Bibliography

Aveni A F (ed) 1989 *World Archaeoastronomy* (Cambridge: Cambridge University Press)

Ruggles C L N (ed) 1993 *Archaeoastronomy in the 1990s* (Loughborough: Group D Publications)

Ruggles C L N 1999 *Astronomy in Prehistoric Britain and Ireland* (New Haven: Yale University Press)

Ruggles C L N and Saunders N J (ed) 1993 *Astronomies and Cultures* (Niwot: University Press of Colorado)

Clive Ruggles

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pre-main-sequence Binary Stars

The typical product of the STAR FORMATION process is a multiple star system, most commonly a BINARY STAR. A pre-main-sequence (PMS) binary star is a pair in which, typically, both stars have not yet fully contracted to the main sequence and are not yet fusing hydrogen at their cores. Such binary stars are recently formed, with usual ages of less than 10 million years. Since stellar contraction times decrease with increasing stellar mass, most known PMS binary stars include stars with masses from a few solar masses to masses of less than one-tenth of a solar mass, at the threshold of brown dwarfs. Although formally the definition of a PMS binary star includes only two stars, they are sometimes found in higher-order systems (e.g. triples and quadruples) with a hierarchical ordering of their separations.

Given their youth, PMS binary stars are often found still associated with the parent molecular clouds out of which they formed. They are also often surrounded by protostellar (and likely, PROTOPLANETARY) disks. Observations show that a PMS binary star can dynamically clear regions in such disks on size scales comparable to the binary separation, as predicted by theory. The impact of such cleared regions on the evolution of binary systems and the formation of planets is at the forefront of the field of star and planet formation.

Pre-main-sequence binary star populations

It has long been recognized that most stars in the galactic field are binary, and thus it was expected that most PMS STARS would also be binary. Even so, it came as a surprise to find that in some star-forming regions PMS binary star frequencies are higher than the binary star frequencies found among main-sequence stars. For example, near the Taurus dark clouds the binary star frequency is a factor of two higher than the main-sequence binary star frequency, a remarkable result given that the main-sequence binary star frequency is more than 50% for solar-type stars. (This apparent mathematical contradiction is the result of triple systems being counted as two binary stars and so on when deriving binary star frequencies.) Such high binary star frequencies among PMS stars leads to the profound result that most stars form in the immediate vicinity of another star.

Presumably the main-sequence stars in the solar neighborhood derive from earlier generations of stars which formed in molecular clouds elsewhere in the galaxy. Hence the higher binary star frequencies among PMS stars also leads to the conundrum of how the main-sequence binary star population derives from the apparently more abundant PMS binary star population. While the answer to this question is not secure, present studies suggest that this is not primarily an evolutionary process—that is, PMS binary stars do not break up to become single field stars. More likely, the PMS binary star frequency may vary with the environment of star formation. For example, PMS binary star frequencies may be higher among stars formed in cold dark clouds, and lower in young clusters and OB associations formed in dense, hot environments. If, as has been suggested, young clusters and OB associations are the primary contributors to the galactic STELLAR POPULATION, then the conundrum is resolved.

Orbital properties

PMS binary stars have been found with orbital periods ranging from 2 days to longer than 100 000 years (the latter inferred from projected separations). The logarithmic distribution of periods is roughly Gaussian, peaking at periods of about 50 years and decreasing to shorter and longer periods. For binary stars made up of two solar-mass stars this peak in the period distribution corresponds to an orbital semi-major axis of 17 AU, similar to the orbital radius of Uranus. So for more than half of the binary stars both stars would reside within the dimensions of our planetary system.

At present our knowledge of the distribution of orbital eccentricity of PMS binary stars rests entirely upon spectroscopically determined orbital elements. Orbital eccentricities from $e = 0$ to $e = 0.8$ have been found. Most of the shortest-period binary stars ($P < 6$ days) have circular orbits; on the other hand, no circular orbits are found at longer periods. Together, these results suggest that circular orbits are not favored by the binary star formation process, and that the circular orbits at short periods are the result of rapid tidal circularization between these large convective stars.

The PMS binary star period and orbital eccentricity distributions are similar to those found for main-sequence solar-type binary stars. Evidently the orbital properties are largely established by the PMS phase of stellar evolution. The secondary mass distribution of PMS binary stars is not yet well determined.

Disks around young binary stars

Circumstellar ACCRETION DISKS are an inevitable consequence of conservation of angular momentum in a contracting cloud, and recent studies have found them to be an essential component of single star formation. The same is almost certainly true for binary star formation. Nonetheless, the conceptual framework of circumstellar disks must be transferred to binary stars with care. Perhaps most importantly, a binary star introduces its semi-major axis as a scale length in the process. We cannot expect disks in binary stars with separations of thousands of astronomical units (AU) to mimic disks in binary stars with separations of a few AU. Furthermore, the concept of a single disk must be discarded. A binary star embedded within a coplanar disk is expected to dynamically clear a region on the scale-length of its semi-major axis. Consequently three distinct disks may be identified (figure 1). Each star can have an associated disk within its ROCHE LOBE, referred to generally as circumstellar disks and specifically as circumprimary and circumsecondary disks (where the primary star is taken to be the most massive star). In addition, a circumbinary disk may be present external to the binary star orbit. One

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

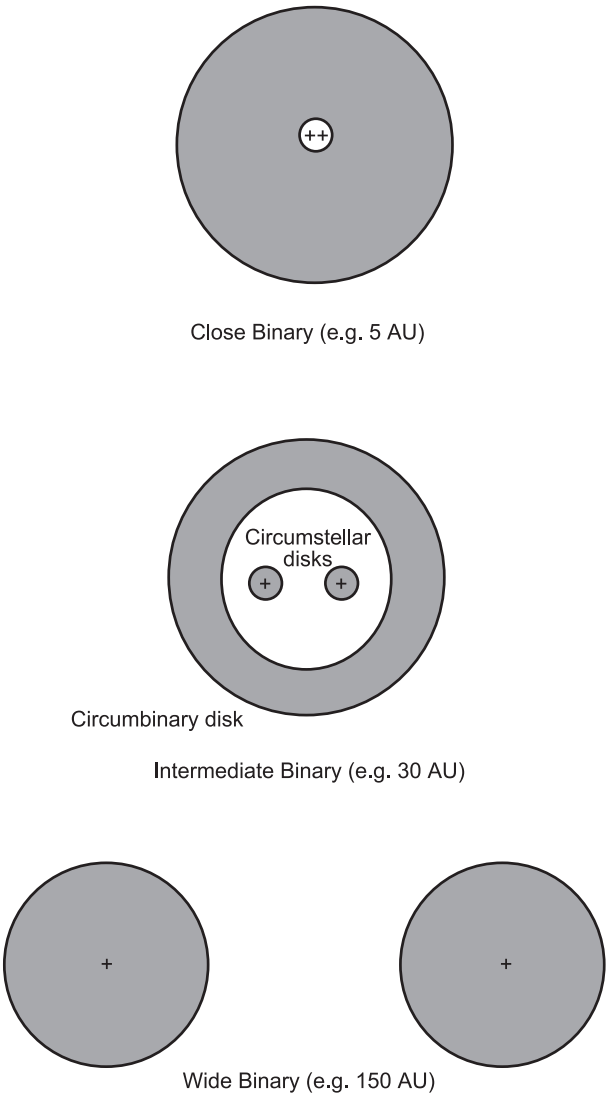


Figure 1. Schematic illustration of circumstellar and circumbinary disks among pre-main-sequence binaries of different separations. The outer radii of the upper two circumbinary disks are not to scale; actual disks have larger outer radii by factors of a few.

or all of these disks may be absent in any given PMS binary star system. Finally, the coplanarity of these disks has not been established by observation, and eventually may be a critical property for distinguishing among binary star formation hypotheses.

Disk–star interactions

Observations of disks around binary stars are best understood in the context of the dynamical theory for the gravitational interaction of stellar companions with an associated disk. In the presence of the periodically varying gravitational potential of a binary star, gravitational resonances are established in disks (similarly to resonances established in planetary rings by orbiting moons, or Lindblad resonances). Torques at these resonances act to transfer energy and angular momentum to or from the binary star orbit, resulting in evolution of both the orbit and the disk structure.

A more detailed discussion is not appropriate here, but the essential elements of the evolution of the binary-disk system can be presented. In this spirit we develop a simple scheme for the evolution of an isolated, coplanar, binary-disk system. We consider initially the case of a binary star with a circular orbit and a small secondary/primary mass ratio. We ignore the details of binary star formation and *ad hoc* begin with the binary star embedded in a continuous coplanar disk having an outer radius much greater than the binary star separation and a mass less than that of the secondary star. Given the posited binary star mass ratio the primary star lies near the center of this disk. On a dynamical (orbital) time-scale the gravity of the companion will initially clear an annular gap in the disk which includes its orbital path. Two circumstellar disks are created, with the circumprimary disk being the larger of the two. A circumbinary disk also remains, with its inner radius set by a balance between the angular momentum transport in the disk (e.g. from viscosity) and the torque provided by the binary star. Calculations find the inner radius of the circumbinary disk to be approximately twice the semi-major axis of the circular orbit.

Once this annular gap is established, disk evolution continues on slower (e.g. viscous) time-scales. Accretion of material from the circumstellar disks onto the stellar surfaces may proceed. However, because of the torques provided by the binary star at the inner edge of the circumbinary disk, material is not expected to flow from the circumbinary disk to the circumstellar disks (but see below). Unless replenished from otherwise infalling material, the circumstellar disks will exhaust themselves. Depending on the time-scale of this exhaustion with respect to the lifetime of the circumbinary disk, the binary star may then reside in a hole within the circumbinary disk. Outside the binary star orbit the circumbinary disk will adjust its surface density distribution so as to match the angular momentum drawn from the binary star with the outward angular momentum flow of the disk. Once the appropriate surface density distribution has been established, a prolonged phase ensues during which the circumbinary disk is expelled.

Such disk evolution requires a response from the binary star. From energy arguments the semi-major axis of the binary star must decrease. The direction in which the orbital eccentricity will evolve depends on the relative rates of extraction of energy and angular momentum from the binary star orbit by the disk. Binary star orbital elements can be expected to change significantly in thousands of orbital periods. For binary stars with semi-major axes less than 100 AU this time-scale is shorter than PMS disk lifetimes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This binary-disk system is the simplest case. Among other things, the detailed distribution of disk material will depend on both the mass ratio of the binary star and the orbital eccentricity. For example, stars in a circular binary star with unit mass ratio are equidistant from the center of mass of the system. Consequently the morphology of the dynamically cleared region changes qualitatively, but the essential nature of the evolutionary history of the disk will not change. The interaction of a disk with an eccentric binary star is a more difficult problem. In terms of resonant clearing, the primary modification is that for given semi-major axis the inner radius of the circumbinary disk increases with eccentricity and the outer radii of the circumstellar disks decrease.

Observations of disks around pre-main-sequence binary stars

The observational case for disks in young binary stars is well established. Observations of PMS binary stars reveal many of the classic signatures of disk material and accretion, including excess emission at near-infrared through millimeter wavelengths, spectral veiling, Balmer and forbidden emission lines and polarization. Every indication is that the frequency of disk material in the environments of PMS binary stars is similar to that around PMS single stars.

Given the scenario above, most attention has focused upon the structure of these disks. Flux measurements at submillimeter wavelengths provide insight into total disk masses. Interestingly, while submillimeter flux is easily detected from many PMS binary stars with large (>100 AU) and small (<1 AU) separations, most binary stars with intermediate separations have gone undetected, indicating reduced disk masses. These observations suggest the rather intuitive picture that binary stars much wider or much closer than typical disk radii do not significantly alter disk structure, while those binary stars whose separations are comparable to disk radii substantially modify the associated disks as in the theoretical scenario above (figure 1). The threshold of roughly 100 AU is similar to the sizes of circumstellar disks around PMS single stars.

The submillimeter observations strongly suggest that companions do alter disk structure, but images are required to understand in detail the actual structure of the resulting disks. The PMS binary star GG Tau has proven particularly fruitful as a case study (figure 2). Millimeter wavelength interferometric observations reveal a circumbinary disk around the binary star. The total circumbinary mass is $0.12 M_0$, about 10% of the stellar mass. The images clearly show the circumbinary disk to be separated from circumstellar disks by regions of very low surface density, indicative of disk clearing. In addition, the circumbinary disk has an inner radius consistent with theory under reasonable assumptions about the orientation of the binary star. These observations are supported by beautiful near-infrared ADAPTIVE OPTICS images which also reveal a ring with the same dimensions. These images are seeing stellar light scattered off the inner edge of the circumbinary disk, confirming the presence of a physical structure.

While observational evidence seems to support the theoretical picture for the structure of disks in binary environments, the frequent observation of ongoing mass accretion at stellar surfaces in PMS binary stars of all separations and ages has been a surprise. Although theory suggests that circumstellar disks in binary stars cannot be replenished with material from circumbinary disks, there is no evidence that circumstellar disks in binaries are in fact being depleted at a faster rate than circumstellar disks around single stars. This is all the more remarkable when evidence for mass accretion is found in the very closest binaries, with separations as small as tenths of an AU (see CLOSE BINARY STARS). These observations have led theorists to re-examine the torque barrier at the inner edge of circumbinary disks, and recent calculations have found that given appropriate conditions accretion streams with substantial flow rates may be able to flow inward from a circumbinary disk. The implication of such streams for the evolution of both the disks and the binary orbit are an important theoretical problem.

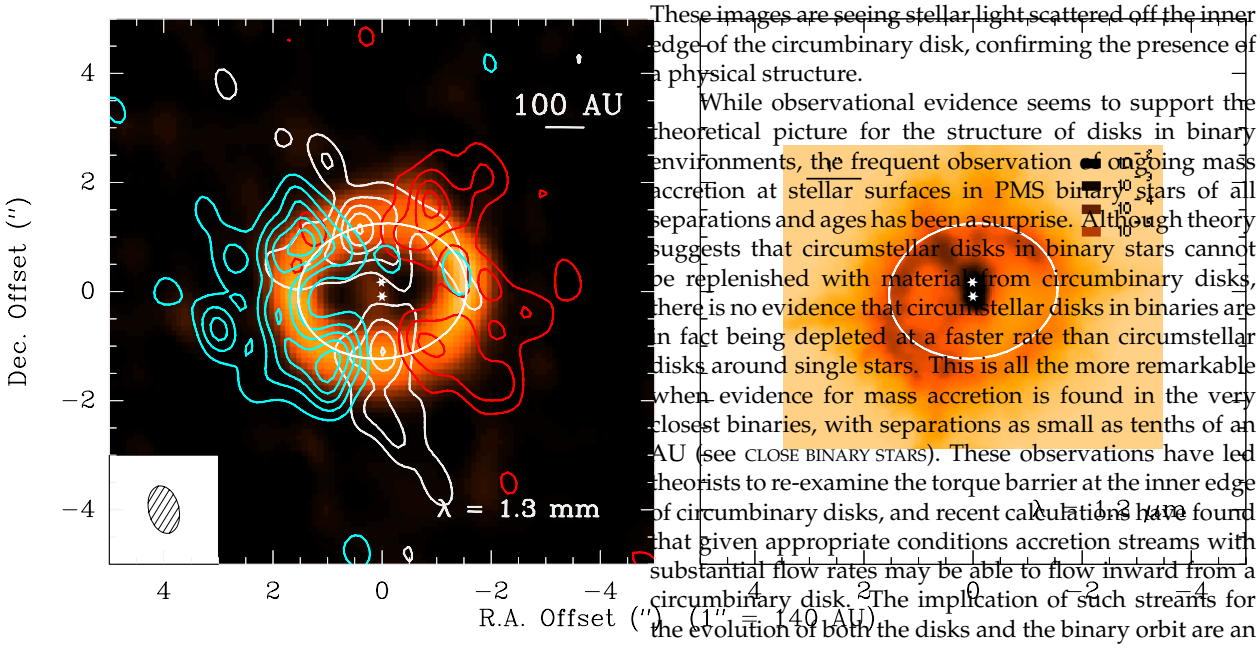


Figure 2. Left: the stellar components of the GG Tau binary are shown as stars. The circumbinary disk is evident as the orange shaded region which shows millimeter wave (1.3 mm) continuum emission. The contours display three velocity channels of ^{13}CO (2-1) line emission, 5.55 km s^{-1}, 6.30 km s^{-1} and 7.05 km s^{-1}, from left to right in the figure. The change in velocity from blueshifted on one side of the binary to redshifted on the other is consistent with Keplerian orbital motion around the binary. Right: the near-infrared (J-band) adaptive optics image. The white ellipse is the same physical size in both images. Evidently the light in the near-infrared image is associated with the circumbinary ring seen in millimeter light, and is interpreted as scattered light off the inner edge of the ring. **This figure is reproduced as Color Plate 38.**

Protobinary stars

Evidently binary star formation occurs prior to the PMS phase. Evolutionary models of the more massive PMS binary stars suggest that the properties of main-sequence binary stars are already at least approximately present in the PMS phase, and likely

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

were established earlier. With the PMS binary stars now opened to observational study, the forefront of discovery is moving to yet younger, embedded protobinary stars. Several very wide pairs have been found embedded within dark clouds. The case of L1551 IRS5 best shows the prospects for the study of protobinary stars.

L1551 IRS5 is the canonical protostar still undergoing infall from an envelope. Recent millimeter wavelength observations demonstrate that the system is a protobinary star with a projected separation of 45 AU (in projection on the sky). Observations reveal a system with three components: a large surrounding envelope of dust and gas, a circumbinary disk or extended structure with a size scale of order 150 AU, and the inner binary star system with two circumstellar disks. Very Large Array millimeter wavelength observations indicate circumstellar disk radii of 10 AU, with disk masses of $0.06M_0$ and $0.03M_0$. These disk masses are comparable with the minimum mass required to form a planetary system like our own. It is also notable that such large circumstellar masses are found in a system with a projected separation of 45 AU, since, as noted above, PMS binary stars typically do not have detectable millimeter flux at this separation. Perhaps these substantial circumstellar disks are maintained by rapid accretion from the envelope, or by an accretion stream from a circumbinary disk.

The presence of such massive circumstellar disks in L1551 IRS5 and near-infrared and submillimeter observations of PMS binary stars suggest that while the radii of disks in binary systems may be constrained by the presence of a companion, the circumstellar disk material may be similar in temperature and surface density to that in disks around single stars. The implication is that in a significant fraction of young binary systems planet formation may be able to proceed relatively undisturbed. Clearly this is true for the widest systems, in which extrasolar planets have already been discovered (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS). The possibilities for planet formation among closer systems remains to be explored. Given that most stars are in binary star systems, the answer will have an important impact on the frequency of planetary systems.

Bibliography

Two more comprehensive and detailed reviews of the field are:

Mathieu R D 1994 Pre-main-sequence binary stars *Ann. Rev. Astron. Astrophys.* **32** 465–530

Mathieu R D, Ghez A M, Jensen E L and Simon M 2000 Young binary stars and associated disks *Protostars and Planets IV* ed V Mannings, A Boss and S Russell (Tucson, AZ : University of Arizona Press) p 703

An insightful (but technical) observational paper is:

Guilloteau S, Dutrey A and Simon M 1999 GG Tau: the ring world *Astron. Astrophys.* **355** 165

Robert Mathieu

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pre-main-sequence Stars

Prior to their ignition of hydrogen, stars undergo a phase of bulk contraction. This process drives both the central density and the temperature upward, until nuclear fusion can begin. Somewhat paradoxically, the star's luminosity, which stems from the released gravitational binding, is relatively high at this early stage. The luminosity generally decreases as a result of contraction, reaches a minimum just at the point of hydrogen ignition, and then rises again during post-main-sequence evolution. Thus, observed pre-main-sequence stars are relatively bright, a fact which aids in their detection. Several thousand low-mass examples have now been cataloged, despite the fact that the relevant lifetime is only about 0.1% that of the main-sequence phase.

A snapshot of the star taken during the contraction epoch would show it to be very nearly in hydrostatic equilibrium. That is, the compressive force of self-gravity closely balances the internal pressure gradient. Evolution is driven by radiative emission from the photosphere, which gradually drains thermal energy. Note that the star's interior temperature rises even as its heat content falls. Thus, the entire object is characterized by negative heat capacity. Note also that the inward velocity of any mass element is always well below the local sound speed. The contraction is therefore termed quasi-static, to distinguish it from more rapid, dynamical motion.

This basic physical picture was already well established by the 1950s, when L Henyey constructed the first detailed, numerical models of young stars. At the time, the role of thermal convection in STELLAR INTERIORS was still not fully appreciated, even in mature objects such as the Sun. Henyey assumed that pre-main-sequence stars would be radiatively stable and found that they contracted at nearly constant luminosity. However, the opacity in the sub-photospheric layers is enhanced by the presence of H^- ions. It was C Hayashi who first realized, in the early 1960s, that the H^- opacity forces the star to radiate at nearly constant surface temperature. The luminosity L_* is therefore high when the radius and surface area are both large. In particular, L_* is more than can be transported by radiation alone, and the interior is convectively unstable.

Within a few years after Hayashi's discovery, the extensive numerical models of Iben, Cameron and others added more quantitative details. The theory now assumed the form it was to retain for many years. Each stellar mass has its own pre-main-sequence track in the HERTZSPRUNG–RUSSELL (HR) DIAGRAM, which it follows down to the zero-age main sequence (ZAMS). Early on, the star indeed contracts almost vertically, as found by Hayashi. The track then veers sharply to the left, and proceeds to the ZAMS along the more horizontal path obtained by Henyey. This change reflects a transition to radiative stability with falling luminosity.

Figure 1 shows the traditional set of tracks, as obtained by Ezer and Cameron. Each track is labeled by the corresponding mass is solar units. Also indicated at a number of points are the times required to evolve along the track to the position in question. It is immediately apparent that more massive stars evolve faster than their low-mass counterparts. The great utility of pre-main-sequence theory is its ability to supply both a mass and an age for any observed object with known bolometric luminosity and effective temperature. These data help us not only to analyze individual stars, but to discern the star formation history of entire clusters.

Despite the systematic drop in contraction time with mass, all the tracks in figure 1 are qualitatively the same. Each begins by descending vertically from a very high luminosity, before turning onto its more horizontal, radiatively stable portion. In contrast, figure 2 displays results from a more recent calculation. Although the individual tracks are still recognizable, the landscape has markedly changed. The upper portion of the diagram, corresponding to the highest luminosities, is now inaccessible to pre-main-sequence stars. So too is the area to the left containing objects of the highest effective temperature. In the new picture, the most massive stars have no pre-main-sequence contraction phase at all, but appear directly on the ZAMS.

What has brought about this change? How, in turn, have these revisions impacted our understanding of STAR FORMATION? My purpose in this review is to answer these two questions. It should be recognized at the outset that the theoretical developments, initiated over 20 yr ago, still continue, although the key ideas are probably secure. During the same period, our empirical knowledge of pre-main-sequence stars has increased tremendously. Here, I can only sketch the major advances and indicate the ongoing effort to accommodate these observations with a more refined theoretical treatment.

Protostars and the birthline

Most of the alterations to pre-main-sequence theory reflect the impact of new initial conditions. Compared with Hayashi and his contemporaries, we now have a better, although still imperfect, understanding of how each star begins its contraction phase. This knowledge in turn translates into the stellar birthline, a rather well-defined upper boundary for the set of evolutionary tracks in the HR diagram. Referring again to figure 2, it is the presence of the birthline that now restricts the luminosities to relatively modest values. In addition, the joining of this curve to the ZAMS sets the upper mass limit for all pre-main-sequence objects.

Earlier calculations of quasi-static contraction needed to adopt specific initial conditions, whatever their physical plausibility. In the absence of relevant observational data, researchers set the stellar radii at very large values. The distended young stars had correspondingly high luminosities, as is evident in figure 1. Hayashi and others attempted to justify this choice on theoretical grounds, arguing that the youngest stars gain their mass too quickly to radiate away appreciable energy. In reality, the hope

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

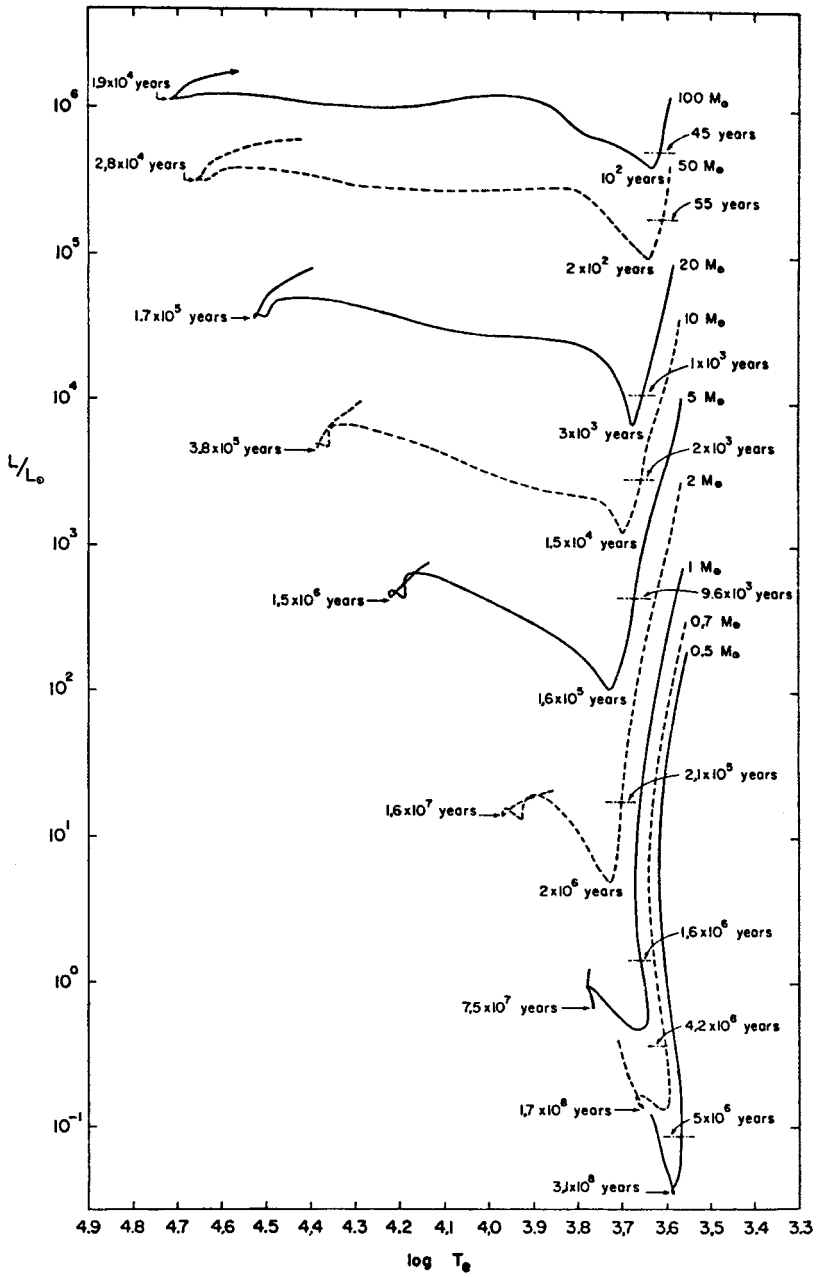


Figure 1. Pre-main-sequence tracks in the traditional theory. Each track is labeled by the appropriate mass. Dashed lines and corresponding times denote the point at which the stars cease to be fully convective. The times to reach the ZAMS are also indicated.

was that all memory of the earliest epoch would quickly be lost during subsequent evolution.

This idea does receive support from the temporal behavior of contracting stars. As we have just noted, any self-gravitating object must lose a large portion of its total energy content in order to reduce its size substantially. Now the virial theorem tells us that this energy is comparable with that from gravitational binding alone. Thus a star of mass M_*, radius R_* and luminosity L_* contracts over the Kelvin–Helmholtz time scale, given by

$$t_{\rm KH} = \frac{GM_*^2}{R_* L_*} = 3 \times 10^7 \text{ yr} \times \left(\frac{M_*}{1M_\odot}\right)^2 \left(\frac{R_*}{1R_\odot}\right)^{-1} \left(\frac{L_*}{1L_\odot}\right)^{-1}. \quad (1)$$

The luminosity, in turn, is that due to a radiating blackbody:

$$L_* = 4\pi R_*^2 \sigma_{\rm B} T_{\rm eff}^4 \quad (2)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

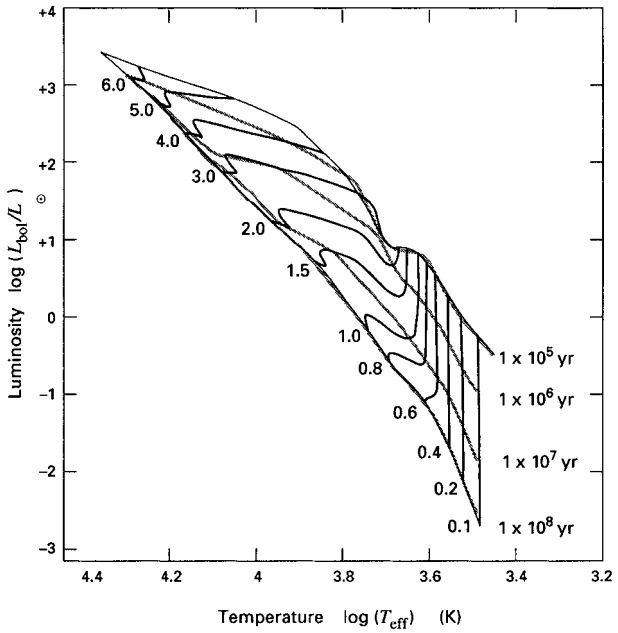


Figure 2. Recent calculation of pre-main-sequence tracks. The upper curve is the stellar birthline, and the lower one the ZAMS. Each track is labeled by the mass in solar units. Also shown are isochrones for selected times.

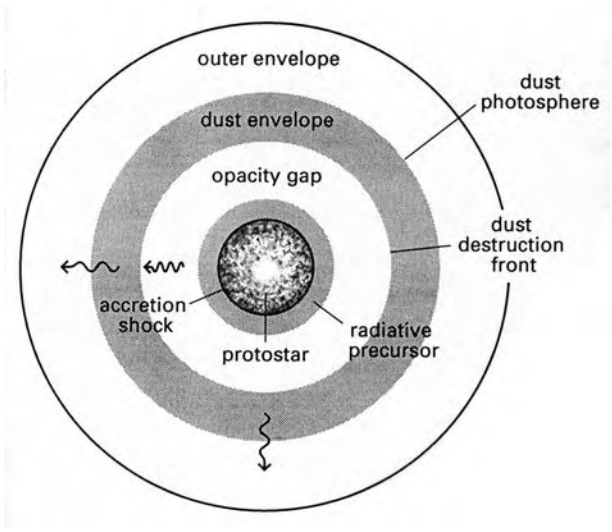


Figure 3. Structure of a protostar and its environment. The range of distances is greatly compressed in this schematic rendering.

where $T_{\rm eff}$ is the star's effective temperature. As the star contracts with nearly constant $T_{\rm eff}$, we see that $t_{\rm KH}$ increases as R_*^{-3}. Thus, the contraction slows down markedly, and any early, distended state indeed occupies a small fraction of the total time to reach the ZAMS.

From an observational perspective, their relatively brief evolutionary lifetimes would make the brightest pre-main-sequence objects relatively uncommon. The difficulty is that stars with the highest L_* values depicted in figure 1 are never seen. Given the large sample sizes, one must conclude that the adopted initial conditions were mistaken. A star actually begins contracting at a much smaller radius and luminosity than was originally envisioned. Although the contraction indeed slows down, the time spent in the earliest state is not negligible. In other words, initial conditions do matter.

The three decades since the original pre-main-sequence calculations have witnessed a substantial growth in our knowledge of stellar formation. Astronomers of the 1960s had no knowledge of molecular clouds, now established as the breeding grounds for stars of all mass. Nor did they understand well the mechanism of gravitational collapse, i.e. the transformation from diffuse cloud fragments to compact stars. Needless to say, these central developments have impacted the theory of pre-main-sequence evolution.

Radio astronomers have by now supplied detailed maps of molecular clouds over a vast range of size scales. Largest are the giant complexes, elongated structures some 50 pc in size, and containing perhaps $10^5 M_\odot$ of gas, mostly in the form of molecular hydrogen. Complexes are self-gravitating swarms of clumps, each one containing about $10^3 M_\odot$. These clumps, in turn, are dotted with numerous dense cores. Each core measures some 0.1 pc in diameter, has a mass of $3M_\odot$ or so and shares the same low temperature (from 10 to 20 K) of its parent cloud. Although the origin of these smallest units remains mysterious, we do know that they are responsible for individual stars.

The first phase of STELLAR EVOLUTION, then, is not the pre-main-sequence, but that earlier epoch when a dense core is undergoing gravitational collapse. The hydrostatic object forming at the center of this converging flow is known as a protostar. Theoretical work on protostars began almost 30 yr ago, with the first direct simulation of a collapsing, spherical cloud by R Larson. It is the subsequent development of this theory that has led to our present, altered conception of pre-main-sequence stars.

Figure 3 depicts schematically a protostar and its surrounding cloud environment. Material in the collapsing dense core rains down onto the surface of the star, forming a strongly radiating accretion shock front. The ultraviolet and x-ray photons produced here work their way back upstream through the infalling gas. As they do so, they elevate the temperature of the envelope to the point where all dust grains thermally sublimate. Since dust is the cloud's chief source of opacity, the protostar is surrounded by a relatively transparent region known as the opacity gap. Beyond this region, photons diffusing through the dusty envelope become degraded to infrared wavelengths. The full accretion luminosity then leaves the cloud at the dust photosphere, typically located about 10 AU from the central star.

A more complete theoretical description endows the parent dense core with a finite degree of rotation. Some portion of the infalling matter then misses the star itself,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and instead forms a circumstellar disk. This material then spirals onto the stellar surface. The structure and growth of the protostar are largely determined by the total mass accretion rate, conventionally designated $\dot{M}$. The geometrical pattern of infall, i.e. the proportion coming through a disk or by direct infall, is less significant than the overall magnitude and temporal behavior of this rate. The former is set by the basic physics of gravitational collapse. If the parent cloud was largely supported by thermal pressure, its subsequent collapse yields a characteristic accretion rate toward the center:

$$\dot{M} \approx \frac{a_T^3}{G}. \tag{3}$$

Here, a_T is the isothermal sound speed in the dense core, a quantity depending only on the ambient temperature T and molecular weight μ. In a molecular cloud with T between 10 and 20 K, equation (3) yields $\dot{M}$ values ranging from $10^{-6} M_\odot$ to $10^{-5} M_\odot$ yr^{-1}.

The most detailed protostar calculations focus on the star itself and represent the infalling cloud envelope solely through the mass accretion rate. Collapse simulations designed to cover the large-scale dynamics of the cloud find that $\dot{M}$ has a rather slow variation with time, at least beyond an initial, transient phase. Thus, it is worthwhile examining the idealized case of a protostar accumulating mass at a strictly constant rate.

Figure 4 shows the results of such a theoretical calculation. Here we see the protostellar radius R_*, displayed as a function of the mass M_* for $\dot{M} = 1 \times 10^{-5} M_\odot$ yr^{-1}. The first thing to notice is that the radii are not much larger than the corresponding main-sequence values. This result contradicts the original assumption of Hayashi and his contemporaries. Protostars have modest sizes because they lose significant amounts of energy while accumulating their mass. Radiation is generated at the accretion shock front bounding the stellar and disk surfaces. As we have seen, this energy readily escapes the cloud by virtue of the opacity gap.

The relatively small sizes of protostars create elevated internal temperatures. Specifically, the stars become hot enough for deuterium to ignite, in the reaction

$$^2\mathrm{H} + \mathrm{p} \rightarrow {}^3\mathrm{He} + \gamma. \tag{4}$$

Although the number fraction of this isotope is only of order 10^{-5}, heat from the fusion swells the star's radius. This swelling begins at subsolar mass and is evident as the first knee in the curve of figure 4. The second, even more dramatic, rise occurs when deuterium subsequently ignites in a subsurface shell. At still higher mass, $M_* \gtrsim 4M_\odot$, the protostar largely exhausts its deuterium supply and starts to contract under the influence of self-gravity. Its temperature rise concurrently accelerates, and ordinary hydrogen ignites by the time M_* reaches about $8M_\odot$. Even more massive protostars are essentially main-sequence objects still gathering material from their surrounding clouds.

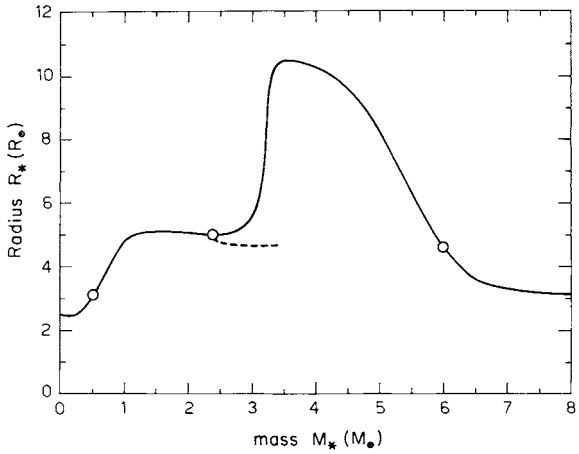


Figure 4. Protostar radius as a function of mass. Open circles mark the onset of deuterium-induced central convection, appearance of a radiatively stable interior region, and the onset of central convection due to hydrogen burning.

Deuterium-generated energy not only swells the protostar but acts as an effective thermostat, locking the central temperature to a value not far from 10^6 K. One important consequence is that the mass–radius relationship is rather insensitive to the adopted $\dot{M}$ value, at least for changes of about a factor of 10. On the other hand, the protostar's history would be very different indeed if no deuterium at all were present. The short dashed curve in figure 4 was obtained by artificially ignoring the fusion reaction in the subsurface shell. In this hypothetical situation, the pronounced swelling would not occur at all.

Protostar calculations like the one illustrated assume a continuous flow of mass onto the central object. Naturally, something must eventually truncate this flow to produce stars of finite mass. One possible factor is rotation of the parent cloud. We have remarked, however, that rotation merely shifts the infall trajectories from the stellar surface to a circumstellar disk. It is likely that internal torques within the disk then promote transfer of this material onto the star. Another possibility is that the star itself blows back its infalling envelope through the action of a strong wind. Finally, infall may cease because the remaining dense core is mechanically prevented from collapsing.

We shall later revisit these last two possibilities. Even without a definitive answer to the question of infall termination, however, we can still address our original concern of pre-main-sequence initial conditions. As long as infall ends relatively quickly, the curve in figure 4 yields the radius of any star of mass M_* at the beginning of its quasi-static contraction phase. The stipulation here is that $\dot{M}$ must decrease substantially over a time that is brief compared with t_{KH} for the mass in question. Provided that this condition is met, we may construct a sequence of initial pre-main-sequence models with a range of masses and the appropriate radii. These objects, unlike their protostellar antecedents, are optically visible, and so may be placed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

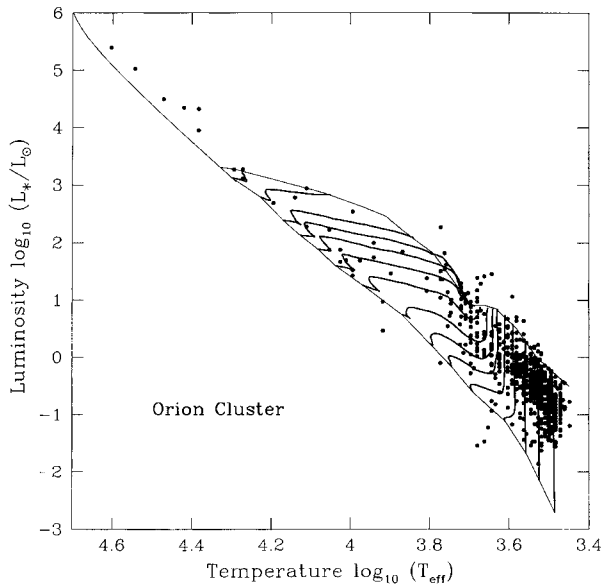


Figure 5. HR diagram of the Orion Nebula Cluster. The theoretical birthline, tracks and ZAMS are identical to those of figure 2, but the ZAMS has been extended to include massive stars.

in the HR diagram. The resulting curve is the birthline depicted in figure 2.

The existence of the birthline is thus motivated by basic theoretical considerations. From an observational perspective, the curve should represent the upper envelope to the positions of young stars in the HR diagram. That is, no stars should be found above the birthline, whether they are observed in isolation or in heavily populated clusters. So far, the voluminous photometric observations of pre-main-sequence stars largely conform to this expectation. For example, Hillenbrand has recently placed over 900 stars of the Orion Nebula Cluster in the HR diagram, after obtaining their bolometric luminosities from I-band observations. The crowded diagram (see figure 5) shows a reasonably sharp upper boundary, which coincides well with the theoretical birthline.

Quasi-static contraction

Low-mass objects

The presence of a birthline in the HR diagram means that pre-main-sequence tracks now fall into two broad categories. Those of lower mass depart from the birthline vertically downward. Observationally, these correspond to T TAURI STARS. In the second category are tracks that begin more horizontally. Here, the theoretical models have radiatively stable interiors, and represent the observed class of Herbig Ae and Be stars. Figure 2 shows that the dividing point is about $2M_\odot$.

We may understand this difference, i.e. predict the fate of any newly formed pre-main-sequence star, by comparing two luminosities. One is the surface value L_*, as given by the blackbody formula of equation (2). The second is $L_{\rm rad}$, the maximum value that can be transported through the interior by radiation alone. This critical luminosity actually varies spatially within a star and also depends on the specific stellar model. Dimensional arguments show, however, that it must be approximately

$$L_{\rm rad} = L_1 \left(\frac{R_*}{R_\odot}\right)^{-1/2} \left(\frac{M_*}{M_\odot}\right)^{11/2} \tag{5}$$

where $L_1 \approx 1L_\odot$. Note the steep mass dependence, reflecting the temperature sensitivity of the interior opacity.

Consider now a sequence of very young pre-main-sequence stars. For sufficiently low mass, $L_{\rm rad}$ is much less than L_*. These objects are radiating more energy than can be transported by radiation. Their interiors are thus fully convective. On the other hand, stars of greater mass have $L_{\rm rad} > L_*$, and are radiatively stable. Because $L_{\rm rad}$ climbs so rapidly with M_*, the transition between the two cases occurs at a rather well-determined mass value, even considering plausible uncertainties in the initial stellar radii.

Schwarzschild's criterion dictates that stars become convective once their specific entropy s decreases outward, i.e. when $s(M_{\rm r})$ is a declining function of the internal mass coordinate $M_{\rm r}$. In practice, however, convection is so efficient at transporting energy that $s(M_{\rm r})$ is nearly uniform, apart from a thin super-adiabatic layer near the surface. Now the heat equation governing evolution of the entropy reads

$$\frac{\partial L_{\rm int}}{\partial M_{\rm r}} = \epsilon - T\frac{\partial s}{\partial t}. \tag{6}$$

Here, $L_{\rm int}(M_{\rm r})$ is the interior luminosity crossing any mass shell and ϵ the local nuclear energy generation rate. Prior to deuterium ignition, we may ignore the latter term. The star loses heat during contraction, so that $\partial s/\partial t$ is negative. It follows that $L_{\rm int}$ is a monotonically increasing function of $M_{\rm r}$, rising from zero at the center to the surface value L_*.

The kind of quasi-static contraction we have just described may be termed homologous. Here, the stellar interior at any one time is simply a rescaled version of its earlier state. This simplification arises from the spatial uniformity of the specific entropy. In mathematical terms, the star may be accurately represented as an $n = 3/2$ polytrope, i.e. the pressure P and density ρ are related at every location through

$$P = K\rho^{5/3}. \tag{7}$$

The coefficient K is a surrogate for the specific entropy and also decreases with time.

Even if a pre-main-sequence object begins its life as a convective object, figure 2 indicates that it eventually becomes radiatively stable, at least for $M_* \gtrsim 0.5M_\odot$. We may again understand this transition in terms of our two

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

luminosities. Contraction reduces R_* and therefore, by equation (2), L_* as well. On the other hand, equation (5) shows that $L_{\rm rad}$ weakly increases at fixed M_*. Thus, the original inequality $L_* > L_{\rm rad}$ may eventually be reversed and the star becomes stable against convection. In stars of the lowest mass, L_* is so much greater than $L_{\rm rad}$ initially that the two cannot reach equality before hydrogen ignites and halts further contraction. That is, the star settles onto the ZAMS while still fully convective.

The transition of a star to radiative stability occurs first at the center, where the temperature is highest and the interior opacity a minimum. The luminosity in the spreading radiative region obeys equation (5), at least in order of magnitude. Thus, the average $L_{\rm rad}$ rises slowly as R_* continues to drop. The surface luminosity must keep pace with $L_{\rm rad}$, so L_* also increases. This rise is evident in the tracks of figure 2. From equation (2), $T_{\rm eff}$ can no longer be fixed but is forced to increase. The tracks now veer to the left in the HR diagram.

This picture of low-mass young stars accords well with a variety of observational data. We note, however, that the evidence for convection itself is entirely circumstantial. The phenomena actually observed are consistent with surface activity, an enhanced version of that present in the Sun and nearby dwarf stars. Thus, many T Tauri stars exhibit periodic fluctuations in light that indicate large, rotating starspots. Prominent optical emission lines, such as Ca H and K, are thought to stem from chromospheric heating. So too is the x-ray emission that has been especially useful in locating pre-main-sequence objects in dense cloud environments.

It is conventional to divide T Tauri stars into two types, classical and weak lined. Objects in the first category have been studied for over 50 yr, ever since the pioneering work of A Joy. They display, besides the emission lines and other signs of surface activity, excess broadband flux in both the infrared and the ultraviolet regimes. Some of their emission lines exhibit P Cygni profiles, signaling the presence of an outflowing wind. Weak-lined stars derive their name from the suppression or absence of Hα emission, another diagnostic of heated, possibly outflowing gas that appears in the classical objects. Originally dubbed 'naked' stars by Walter, they have normal broadband energy distributions for their spectral types.

In the HR diagrams of young clusters and associations, the two types are thoroughly intermingled close to the birthline. The weak-lined variety predominates, however, as one approaches the ZAMS. The generally accepted interpretation of this pattern is that classical stars are surrounded by disks of dust and gas. Stellar luminosity heats the dust, causing it to reradiate at infrared and longer wavelengths. Apparently, the disks vanish as stars contract toward the ZAMS. Note finally that many classical stars also display inverse P Cygni profiles. The suggestion is that matter from the disk may be spilling onto the stellar surface, perhaps along magnetic field lines. In any case, the associated mass transfer rate of roughly $10^{-8} M_\odot\ {\rm yr}^{-1}$ is too low to cause substantial deviation from the constant-mass theoretical tracks of figure 2.

The variety of T Tauri broadband spectra implies that some pre-main-sequence stars first appear with substantial disks, while others seem to lack them entirely. As we have noted, disks represent material from the parent dense core that had too much angular momentum to impact the star directly during collapse. The production of objects both with and without disks may therefore point to a range in cloud rotation rates. However, it will be difficult to quantify this idea before we have a better understanding of how the infall process actually terminates.

There is a smooth morphological progression from T Tauri stars with disks to other stars exhibiting far greater levels of infrared emission. The so-called 'flat-spectrum' and even more extreme 'class I' sources have far too much flux at long wavelengths to be attributable to disk reradiation. Indeed, the typical class I spectral energy distribution shows a broad maximum near 100 μm, far beyond the near-infrared peaks of classical T Tauri stars. What is the nature of these more embedded objects?

One critical observation is that class I sources are invariably located near the centers of dense cores of molecular gas. Thus, their infrared and millimeter radiation plausibly stems from heated dust within more massive circumstellar envelopes. A number of researchers have suggested that these envelopes are collapsing onto their central stars, i.e. that the objects are true protostars. One should then observe, in addition to the star's photospheric luminosity, the additional energy release from infall. Indeed, current theory predicts that this infall component should actually dominate the total. Observationally, however, the bolometric luminosities of class I sources are quite similar to those of optically visible T Tauri stars, whether classical or weak lined. It thus appears that, while some pre-main-sequence stars appear promptly as revealed objects on the birthline, others remain embedded for some time in remnant cloud material. How this gas is mechanically suspended is not well understood, but the ambient magnetic field almost certainly plays a role. In this regard, it will be interesting to see whether high-resolution mapping at millimeter wavelengths eventually reveals systematic differences between dense cores with and without internal stars.

Stars of intermediate mass

Consider next the evolution of objects above the transition near $2M_\odot$. Here $L_{\rm rad}$, the maximum luminosity carried by radiation, exceeds the surface value L_*. The interior opacity is now low enough that the star can have a fully radiative interior. As we have noted, $L_{\rm rad}$ is not just an upper limit, but also a good estimate for the actual luminosity under the condition of radiative stability. The inequality $L_{\rm rad} > L_*$ then implies that the surface cannot emit all the energy generated by the deep interior through gravitational contraction. Very young stars of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

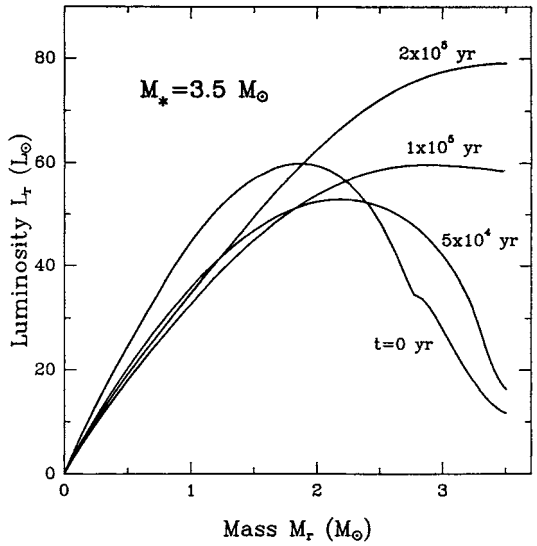


Figure 6. Evolution of the internal luminosity in a $3.5M_\odot$ pre-main-sequence star. The luminosity crossing any shell is plotted against the mass interior to that shell, for the indicated times.

intermediate mass are out of thermal balance and must readjust accordingly.

Figure 6 displays a temporal sequence of luminosity profiles from detailed modeling of a $3.5M_\odot$ star. Plotted is $L_{\rm int}(M_{\rm r})$, starting when the object first appears at the birthline. At this earliest time, we see that $L_{\rm int}$ reaches a maximum value of about $60L_\odot$, while the surface luminosity L_* is only about $10L_\odot$. What becomes of the excess energy? From the heat equation (6), $\partial s/\partial t$ is now positive for all mass shells between the luminosity peak and the surface. The energy being released by contraction thus heats these layers, at least temporarily.

Equation (6) also implies that the entropy is still falling inside of the luminosity maximum. This region therefore contracts as in lower-mass objects. Most of the energy expelled is transferred to the outer layers, which actually expand. The structural evolution of the star is now nonhomologous, and the earlier description in terms of polytropes breaks down. In physical terms, the star has accumulated its mass at a rate which outpaced its ability to thermally relax.

Heating of the outer layers persists until the surface temperature $T_{\rm eff}$ climbs substantially. The luminosity L_* then rises as well, eventually matching the interior peak value. This development may be followed through the alteration of the profiles in figure 6. In effect, the star has released a thermal pulse to the surface. From a more observational perspective, the object may be said to undergo a luminosity jump, a feature evident in the evolutionary tracks of figure 2. The duration of this phenomenon is the thermal relaxation time, or roughly 10^5 yr at the relevant masses. Note that this period is shorter than $t_{\rm KH}$ in equation (1). In the present case, the energy content of the star refers only to its outer layers and is therefore less than the full GM_*^2/R_*.

The observed objects corresponding to these theoretical models are the Herbig Ae and Be stars, first identified nearly 40 yr ago. While these stars are rare compared with their low-mass counterparts, they have come under increasing scrutiny in recent years. There is still no empirical evidence for the thermal transition that should occur at these masses, including the luminosity jump. On the other hand, the very existence of this class is a testament to the essential correctness of the theoretical approach. It should be apparent by now that the location of the birthline in the HR diagram depends solely on the mass–radius relation for accreting protostars. We have seen how those of intermediate mass swell because of deuterium fusion (recall figure 4). Were it not for this effect, the birthline would be shifted downward toward the ZAMS, excluding the region where Herbig Ae and Be stars are actually found. Physically, smaller-radius protostars would have higher central temperatures and would ignite ordinary hydrogen earlier, thus joining the ZAMS at a lower mass.

Pre-main-sequence stars of intermediate mass also display, to varying degrees, excess broadband emission at infrared wavelengths. Subtracting off the likely photospheric contribution to the spectral energy distribution, one often finds that the long-wavelength flux constitutes a large fraction of the total observed luminosity. It was first argued that this emission must stem from accretion through a circumstellar disk, since a more isotropic distribution of dust would have obscured the visible radiation that is actually seen (see DUSTY CIRCUMSTELLAR DISKS). However, the required accretion rates can be so high that the disk would still need continual replenishment from an external envelope of dust and gas. In addition, observers have now spatially resolved the infrared emission in a number of cases, a clear demonstration that the source is much larger than a disk. We are led once more to a picture in which remnant cloud material hovers about the star for a protracted time.

Even more puzzling than the infrared excesses are the signs of surface activity in a large fraction of Herbig Ae and Be stars. Emission in Hα and other lines is present in these cases and frequently displays the P Cygni profiles indicative of winds; other objects have substantial x-ray luminosities. Now the theoretical models do have transient surface convection, powered by the diminishing supply of deuterium. However, convection always vanishes by the time $T_{\rm eff}$ increases to the range appropriate for A and B stars. Could the x-rays and winds instead originate in lower-mass companions that have not been spatially resolved or do they signal residual infall, as is believed for T Tauri stars? Further observations are needed to sort out this issue.

The thermal relaxation characterizing pre-main-sequence stars also occurs in protostars, provided that they accumulate sufficient mass. Thus, the steep increase of R_* seen in figure 4 is caused both by deuterium shell burning and by the shift of internal entropy accompanying the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

thermal pulse. At these masses, the protostar's luminosity stems more from internal energy release than from external infall. The time scale $t_{\rm KH}$ is now significantly shorter than the evolutionary time $M_*/\dot{M}$, and the object rapidly contracts even as it accretes. The mass at which ordinary hydrogen ignites, about $8M_\odot$ for $\dot{M} = 1 \times 10^{-5} M_\odot$ yr^{-1}, also marks the point in the HR diagram where the birthline joins onto the ZAMS. Observations of Herbig Ae and Be stars indeed find an upper mass limit close to this value.

Stars of even higher mass never undergo pre-main-sequence contraction. They exist as obscured protostars, as objects on the ZAMS or as post-main-sequence stars rapidly depleting their nuclear reserves. It is doubtful, moreover, that their protostar phase resembles that at lower masses. Powerful winds from these objects drive back infalling matter, a tendency that is even present to a lesser extent at lower masses. The extraordinary luminosities of massive stars also disrupt the envelope, through radiation pressure acting on the grains. A number of researchers have pointed out that anisotropic infall might lessen the radiative effect. For example, gas in a rotating envelope would not fall directly onto the star, but onto a disk, where it could be shielded by the high optical thickness. However, the disk itself must remain intact for an extended period, despite the star's harsh ionizing flux. Observations of disks in Orion being ablated by the Trapezium stars show that this is no easy matter.

How then do massive stars form? It is significant that the youngest such objects appear close to the density peaks of crowded stellar clusters. The more commonly observed OB ASSOCIATIONS are the expanding descendants of such groups, while systems such as the ORION NEBULA Cluster represent them at an earlier, more compact stage. In the latter example, numerical calculations have demonstrated that the massive stars have not had time to drift to their present, central locations but must have formed *in situ*. One possibility is that these objects originate through the coalescence of previously existing stars. The interaction cross sections of the stars themselves, or even binaries, are probably too small to be relevant, even in the densest observed clusters. However, we have noted that pre-main-sequence stars appear to remain for some time within their parent cloud cores. The merger of these gaseous entities and the subsequent joining of their internal stars is an attractive picture that remains to be developed.

Unanswered questions

Throughout our discussion, the initiation of the pre-main-sequence phase has been a frequently recurring topic. Specifically, we still need to understand what physical event marks the termination of dynamical, protostellar collapse. The derivation of the birthline sidestepped this question by simply assuming that the critical event is suitably brief. We saw, however, that the relevant interval may not be small for many of the stars still surrounded by residual gas and dust. More fundamentally, the issue of infall termination is central in our quest to understand the empirical distribution of STELLAR MASSES.

One possibility we have mentioned is that stars themselves are the responsible agents, through the momentum of their outflowing winds. This idea is motivated by the nearly universal association of optical jets and molecular outflows with embedded stellar objects. Outflows are thought to be cloud gas entrained by the high-velocity wind, where the latter is channeled into a narrow jet. On the other hand, we have no clear example of a cloud being shredded by its internal jet. Such well-observed objects as the Bok globule B335 show both high column densities in molecular gas and vigorous, well-collimated flows. The available evidence is that winds and dense cores coexist for some period. The proposition that the former eventually destroys the latter in all cases remains to be demonstrated.

An alternative viewpoint is that stars accrete all the nearby material available to them. Now the typical masses of cloud cores, as obtained from molecular line observations, are larger by nearly an order of magnitude than the average mass of a newborn star. Assuming that the densest, inner portion of the cloud first collapses, the implication is that the outermost envelope is somehow prevented from joining onto the central star.

The most likely means of preventing, or at least forestalling, collapse is through the embedded magnetic field. Measurements of Zeeman splitting have established a well-ordered field of typical magnitude 10 μG in the lower-density gas surrounding cores. Unfortunately, this observational technique fails within the core itself, but we may assume that the field penetrates this inner region. The narrow linewidths of NH_3 and other tracers show that the core gas is relatively quiescent. Thus, the dominant expansive force aiding thermal pressure is tension from a smooth field, rather than the pressure associated with MAGNETOHYDRODYNAMIC (MHD) WAVES . By the same token, the broad lines observed outside the cores indicate a strong wave influence throughout the larger cloud volume.

Although bending of field lines can effectively resist gravity, it cannot do so indefinitely. Neutral matter slowly drifts across the field through the process known as ambipolar diffusion. Thus we may picture, at least crudely, the collapsing inner core tugging on the ambient field until tension halts rapid infall onto the central star. This event marks the end of the protostar phase. Further mass addition and depletion of the core would occur much more slowly, over the diffusive time scale. The stars receiving this matter would be young, pre-main-sequence objects with strong emission at long wavelengths, perhaps the observed class I sources.

Other vexing problems remain in the study of pre-main-sequence evolution. Chief among them is the issue of ANGULAR MOMENTUM. We remarked before that disks represent cloud material of excessive angular momentum that misses the star during collapse. It would seem to follow that the stars themselves must rotate close to breakup. In fact, the observed rotation speeds of T Tauri stars are very slow, typically under 10% of the breakup value. Furthermore, there is no clear indication of spinup

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

due to contraction along the evolutionary tracks, except very close to the ZAMS.

It has long been accepted that the torquing action of magnetic winds provides an effective brake on stellar rotation, in all low-mass stars including the Sun. Given a field strength and geometry, theory can predict with confidence the rate of angular momentum loss from the stellar surface. Less well understood is the interior response of the star to this braking action. Nor can theory yet specify what field is generated by a given stellar model. Until these gaps are closed, we will be unable to provide a convincing history of pre-main-sequence rotation.

These problems are fundamental, involving both dynamo theory and the stability of rotating fluids. According to current thinking, regions that are stable against convection may still undergo a kind of anisotropic mixing that serves to transport angular momentum outward. This same process would also affect the internal CHEMICAL COMPOSITIONS OF STARS. Thus, the angular momentum problem relates to yet another outstanding issue, the depletion of lithium and other light elements. In the future, the pre-main-sequence theory outlined here will continue to provide the basic framework for these and other investigations.

Bibliography

For a brief history of the theory of young stars, see

Stahler S W 1988 *Publ. Astr. Soc. Pac.* **100** 1474

An influential calculation of pre-main-sequence evolutionary tracks prior to the development of protostar theory was

Ezer D and Cameron A G W 1967 *Can. J. Phys.* **45** 3429

while a modern result is

Palla F and Stahler S W 1999 *Astrophys. J.* **525** 772

Calculations of protostellar collapse have been both numerical and analytic. Examples of each include

Larson R B 1969 *Mon. Not. R. Astron. Soc.* **145** 271
Shu F H 1977 *Astrophys. J.* **214** 488

For the growth and structure of the protostar itself, see

Palla F and Stahler S W *Astrophys. J.* **375** 288
Stahler S W, Shu F H and Taam R E 1980 *Astrophys. J.* **241** 637

Note that the first article focuses on intermediate-mass stars.

The class of T Tauri stars was first recognized by

Joy A H 1945 *Astrophys. J.* **102** 68

while the weak-lined subclass was distinguished by

Walter F W 1986 *Astrophys. J.* **306** 573

A comparison of pre-main-sequence theory with observations, concentrating on low-mass objects, is in

Stahler S W and Walter F W 1993 *Protostars and Planets III* ed E H Levy and J I Lunine (Tucson, AZ: University of Arizona Press)

The categorization of Herbig Ae and Be stars is originally due to

Herbig G H 1960 *Astrophys. J. Suppl.* **4** 337

A conference proceedings devoted to this latter topic is

Thé P S, Pérez M R and Van den Heuvel E P J (eds) 1994 *The Nature and Evolutionary Status of Herbig Ae/Be Stars* (ASP)

Steven W Stahler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Primeval Fireball

A term coined by P J E Peebles, to describe the early phase of the Big Bang universe during which space was filled with an opaque, high-temperature, mixture of radiation (photons) and particles (protons, neutrons, electrons and other elementary particles). Also known as the 'primordial fireball'.

See also: Big Bang theory, cosmic microwave background.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Procyon

The star α Canis Minoris, a cream subgiant of spectral type F5IV–V, apparent magnitude 0.40. Though long known as the 'Little Dog Star' or 'Lesser Dog Star', from the constellation name, the name Procyon is derived from the ancient Greek for 'before the dog', no doubt referring to the fact that it rises before its companion dog star Sirius. The two are the only bright stars in the region to the east of Orion, and hence are especially prominent as they rise in the eastern sky on winter evenings. Procyon is the twelfth star in distance from the Sun at 11.4 light-years, having a parallax of 0.286″. Its absolute magnitude is 2.7.

Procyon is a spectroscopic binary, with a period of 40.4 years. There is also a white dwarf optical companion (magnitude 11.7) at a separation of 122″.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Profession

There was a time, not very many years ago, when professional astronomers were widely regarded as being divorced from the real world. It was thought that they spent their nights in lonely observatories watching the stars, and that many were decidedly advanced in years. This picture was never true, and today it is as wide of the mark as it is possible to be. Your average professional astronomer is young or at most middle-aged, works in a team and very seldom actually looks through a telescope.

In bygone days, of course, there were very few full-time astronomers, and these were supported by the patronage of dukes, princesses, kings and queens, or perhaps through membership of religious orders. Professional astronomers supported by patronage were almost always at least part-time astrologers.

With the rise in Europe, and later in America and the countries of the new world, of the universities, astronomers were employed in larger numbers both to teach astronomy and to research into it, in particular into what we would now call dynamical astronomy or celestial mechanics. With the development of the concept of state responsibility, other astronomers were employed in civic-funded observatories and institutes, financed on a city, regional or national level. Such observatories and institutes were established both for the development of navigation and time-keeping (the Royal Observatories at Paris and Greenwich, and the US Naval Observatory were founded for these reasons), and for making accurate positional observations and calculations about the motions of stars, double stars etc. (astrometry).

After amateurs led the way through observation of the intrinsic properties of the Moon, the planets, the stars and the nebulae (see AMATEUR–PROFESSIONAL COLLABORATION IN ASTRONOMY), by contrast to their positions, more and more astronomers turned to astrophysics, the application of scientific principles to elucidate the nature of celestial bodies—the majority of astronomers at the present time are astrophysicists. The definition of a 'celestial body' is usually taken to mean everything outside the Earth, i.e. in practice everything more than 100 km above the Earth's surface, and the Earth itself in its planetary properties. This includes a very wide variety of natural phenomena! Astronomy is thus an interdisciplinary subject, linked inextricably with all other branches of science: space research (of course), physics, mathematics, chemistry, biology—there are no hard and fast boundaries. The astronomer was never really 'a man apart', but he is even less so now.

The AMERICAN ASTRONOMICAL SOCIETY estimates that there are 6,000 professional astronomers in the USA. In Europe there is about the same number (an estimated 1000 in the UK at post-doctoral level and up). Perhaps the world total is some 15,000–20,000. Established professional astronomers (also, rarely, distinguished amateurs) may become members of the INTERNATIONAL ASTRONOMICAL UNION, through nomination by the Union's national adhering organisations. Global membership is 6,000. The international language of astronomy is English. The number of astronomers is small enough to form a genuinely international community. This tendency is encouraged by the outward-looking nature of the subject, and the scale of investigations of large numbers of stars and galaxies, perhaps from both hemispheres of the Earth, which means that collaborations of astronomers are necessary to tackle the big astronomical questions.

Although astronomy is a universal subject and has a wide appeal, professional astronomers are by no means a representative sample of the human population. They are most numerous (relative to the population) in Europe, North America, Russia and Australia. There are significant numbers of professional astronomers in South America and Asia. There are few in Africa. The proportion of female professional astronomers varies widely from country to country, and is never 50%. Vatican City's Jesuit astronomers are 100% male, for understandable cultural reasons. The northern European nations have the smallest proportion of females. The United States forms an intermediate case between these and the southern European and South American countries. In all countries, the proportion of females decreases with seniority. (In what follows we will have to accept the limitations of the English language and often use male pronouns to denote astronomers of either gender.)

It is difficult to give a simple picture of the prospective development of astronomy as a profession, since the situation varies from country to country. In the USA and the UK, public spending is limited by national economic policy and there is increasing pressure on astronomical budgets in observatories and public institutes. This means that the number of astronomers employed in such places is static, or decreasing. On the other hand, the attraction of astronomy to students has meant that an increasing number of universities offer astronomy components to degree courses, and the number of teachers has been increasing. Telescopes are provided centrally, not necessarily in a big laboratory on campus, so astronomical research is relatively cheap to the universities, and teachers have the opportunity for independent research in bite-sized morsels which have high impact. In general university authorities thus support astronomy research by their teachers.

In continental Europe there has been a history of a larger proportion of astronomers being employed through national, regional or city observatories, and the same economic pressure developing in France and Germany is causing reductions in numbers of astronomers, which may be countered by university growth. Russia and China are special cases, in which astronomers have had high status as respected intellectuals and government employees under the communist governments, a status which has eroded in a negative reaction to everything governmental and disconnected with day to day survival. The difficult economic situation and greater competitivity in both countries has led to a reduction of the numbers

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of astronomers, although this seems to have been accomplished mostly at the expense of the less productive ones. In China this process has been called 'Going down to the sea,' meaning making your own way and surviving by a struggle.

At the same time that national facilities are under pressure, established international ones are under pressure too, such as ESA and ESO. However, new international facilities are being set up all the time, such as Gemini and ALMA. The net trend in science is positive as the old gives way to the modern; the net trend in the population of astronomers is hard to discern. In the established astronomical countries, there were periods of great expansion in the population of astronomers after the 1957 launch of Sputnik, and the broadening of tertiary education. These periods of great expansion in traditionally astronomical countries are over, and astronomy in them is now in steady state. The future expansion may be in the astronomically developing countries, like Greece, Spain and South America, with its large mass of astronomical assets especially in Chile.

Astronomy is based on observation of the natural experiments in the Universe. Interpretation of the observations is by theory. Theoretical astronomy suggests further observations. Most astronomers combine theory and observation in their studies to some degree, but most are either principally observational astronomers or principally theorists. In principle, astronomers focus on the natural experiments, and deploy what is required to understand the phenomena, whether theory and calculation, or use of the range of available technology (radio, x-ray or optical telescopes, interferometers, spectroscopy, photometry, space astronomy,...). In practice, a given astronomer will develop skills in a subset of techniques which he or she deploys on the phenomena to which they are most appropriate. An investigation might therefore require the cooperation of a team of astronomers, especially when the investigation is large-scale.

In past times the astronomer had to make his observations at night, often on his own in the cold, up a ladder. The astronomer's presence was required to make and document the observations, initially with the eye and later by photography, and at the same time to operate the telescope, pushing it to new positions and relocating the ladders, while monitoring the performance of the telescope. Travel to an observatory in a remote place may have been arduous, and the sleeping arrangements may have been primitive. This all demanded a certain amount of physical stamina. EDWIN HUBBLE, the great astronomer who first proved that the so-called spiral nebulae are in fact external galaxies, used to relate how he spent night after night in the dome of the Mount Wilson 100 in reflector, checking to make sure that the target object was still perfectly positioned. The degree of domestic comfort may be judged by the name of the hostel on Mt Wilson—'The Monastery'. Today, with computer control of the telescope and data recording, the astronomer need not be in the telescope dome, or even in the same country. Indeed, if the telescope is in space he cannot be in contact with it. This has meant that modern astronomers can specialise in astronomy, rather than technology, and more astronomers than before are 'wavelength independent', happy to use any telescope in pursuit of the astrophysics. Some astronomers lament this, in part from nostalgia, more importantly because there is a generation of astronomers growing up who do not understand the limitations of the observation process and instrumental performance.

It is possible for an individual astronomer to own and have exclusive access to a small telescope, using it whenever he or she wants and conditions are suitable. Moderate sized facilities may be within the scope of a small group of astronomers, say a university department. Large telescopes are however expensive and often provided from national funds and therefore probably with an expectation that access will be 'open'. Competition for access to the facilities may therefore be high—'telescope time' is at a premium because there are simply not sufficient numbers of large telescopes to satisfy everybody in the profession. The same may be true of access to computing power.

All these facts are highly relevant from the viewpoint of the young enthusiast who is considering a career in professional astronomy. As a desirable profession, and a competitive one, modern astronomy is populated by men and women of exceptional ability and drive. Astronomy is an essentially mathematical and physical science; anyone who does not have a natural aptitude for the mathematical sciences may find employment in the role of an assistant to astronomers, for example as engineering or computing support staff, or (as exemplified by one of us) in the astronomical MEDIA, but is unlikely to succeed as a professional astronomer.

In the progression towards being an astronomer it is wise to consider the competitiveness of the profession and to defer specialising wholly in astronomy until the later stages. If you are trained in maths or physics, you can take up astronomy and also a number of other professions; if you are trained exclusively in astronomy you are not so versatile or so desirable to an employer, if your plans to be an astronomer come to nothing. In school it would be wise to take up the maximum quota of science and mathematics subjects, and certainly to aim for a university degree. The first degree (BSc or equivalent) need not be in pure astronomy, indeed, according to the principle of least specialisation perhaps is better not to be, but many general science, physics or mathematics courses in universities offer astronomy and astrophysics options. There is, in fact, evidence in Britain that such options enhance the attractiveness of such courses in universities to students, even those who do not intend to take up astronomy as a career (see EDUCATION). Once a basic science, physics or math degree has been obtained, a further degree in astronomy can be added at Masters level, or in the US must be added as part of the Qualifying Examination for a PhD course. During the astronomy study, the student

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

will be able to build on the necessary mathematical and physical background acquired in the first three or four years. In France the parallel sequence is the DEUG in mathematical or physical sciences, followed by a "maîtrise" in mathematical or physical sciences with an Astronomy or Astrophysics option.

At Bachelor's or Master's degree level, or its equivalents, career options remain limited to the role of an assistant helping better qualified astronomers, i.e. those with doctorate (PhD or equivalent). The doctorate course (in France Diplôme d'Etudes Approfondies followed by Thèse de doctorat) is a training, not in the acquisition of knowledge discovered by someone else, but in research to acquire new knowledge. The training is supervised by an established astronomer, who will offer guidance on the choice of topic and execution of the research programme. In most cases the supervisor will work with the doctoral student on the research, including any observational work with telescopes, data reduction, and theoretical calculations as appropriate to the topic and the student's inclinations and abilities. Often, a long lasting professional relationship builds up between a supervisor and his students, especially the talented ones. The research is written up by the student into publishable or published scientific papers, and assembled into a thesis.

Success at the PhD stage is the entry qualification into astronomy as a profession. It is likely that the astronomer will be further informally tested for a few years in fixed-term or temporary positions, such as fellowships. It is not uncommon for such positions to be in other countries—indeed in the Netherlands it is regarded as a requirement to become an established Dutch astronomer that at the post-doctoral stage the astronomer should have been overseas, often in North America. During this period he or she has an opportunity to make a name for him- or herself by making important astronomical advances. The result could be to secure that highly desirable, tenured and long-term position as an astronomer in a university, an observatory or an institute.

In past times, the student astronomer worked with a more senior astronomer in a kind of apprenticeship. Just as Johannes Kepler took over the role of his teacher Tycho Brahe, the student would succeed his master. A modern PhD course lasts for three or four years, and a professional astronomer's career lasts perhaps 30 or 40. An active astronomer might supervise a large number of PhD students; even if an astronomer has just one student at a time he might supervise more than ten altogether. It is clear that not all these students can succeed their supervisor, given that that employment opportunities for astronomers are in steady state. As in any population, steady state is not compatible with reproducing oneself several times over. This is the numerical basis for the competition to become a professional astronomer.

In the UK approximately 20% of PhD astronomy students remain in astronomy as a profession. On the other hand there are none who are unemployed. Student astronomers at this level are intelligent, numerically literate, used to interpreting incomplete data (just as you have to in real life), practised in working in teams, often with international colleagues—and they have something to say at their interview. They are very employable, if not all in astronomy. This security has acted as a back stop for astronomy students in the UK who continue in increasing numbers to seek out opportunities to contribute to astronomy and to compete to become professional astronomers. We too were attracted into astronomy at an early age, and we know how powerful is the attraction. If we had to summarise our advice to an aspiring astronomer, it would be 'Recognise the realities and make fall-back plans; but it's not impossible, and it's very rewarding to succeed, so give it your best try.'

Bibliography

Careers advice on joining the community of professional astronomers in the UK is available on the Web from the Cardiff Astronomical Society (http://www.astro.cf.ac.uk/cas/sheet3.pdf). The Royal Astronomical Society publishes a leaflet: Becoming an Astronomer. An Australian perspective is available from the Astronomical Society of Australia (http://www.atnf.csiro.au/educate/become.html). Three comprehensive American discussions on astronomy as a profession are available from NOAO (http://www.noao.edu/education/astfaq.html), from the American Astronomical Society (http://www.aas.org/education/career.html) and from the US Bureau of Labor Statistics (http://stats.bls.gov/oco/ocos052.htm). All these references include realistic statements about employment prospects. The French route to an astronomical career is described by the Société Française d'Astronomie et d'Astrophysique (SF2A) at http://www.iap.fr/sf2a/metier.html.

Paul Murdin and Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Prognoz

Series of 10 Soviet spacecraft launched 1971–85 to investigate the solar wind and its interaction with Earth's magnetosphere. Prognoz-9 (launched July 1983) carried experiments to study the Sun, x-ray and gamma ray bursts as well as the Relict-1 radio astronomy experiment. Prognoz 10 (launched April 1985) carried instruments for studying the plasma, the magnetic field, electric and magnetic components of waves, solar flare x-ray bursts and kilometre radiation. (See also INTERBALL and RELICT.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Prominence

A plume of luminous gas in the solar atmosphere visible beyond the solar limb (the edge of the visible disk). Although their densities are about a hundred times greater than that of the surrounding coronal material, they are too low to enable them to emit a continuous white-light spectrum. Their visible light consists mainly of emission lines of hydrogen and ionized calcium and, except during total eclipses of the Sun (when they may be seen directly by eye), they are normally studied in monochromatic light (wavelengths corresponding to particular emission lines, such as the hydrogen-α line in the red part of the spectrum).

There are two principal types of prominence—quiescent and active. Quiescent prominences hang like clouds in the corona for weeks or months with relatively little overall change, whereas active prominences undergo rapid changes. Quiescent prominences have lengths ranging from several tens of thousands of kilometers to several hundred thousand kilometers and heights of around 30 000 km. They are suspended above the chromosphere by magnetic fields. Most quiescent prominences look like curtains of gas that connect with the chromosphere at various points and contain many near-vertical strands down which gas flows. From time to time, a quiescent prominence will suddenly erupt outwards and dissipate, often to be replaced, in the same location, by a new one. Active prominences, which are associated with active regions on the Sun, and are often associated with, or triggered by, solar flares, are short-lived phenomena, most of which survive no more than a few hours. They take a variety of forms. Among the most active are loop prominences, structures that connect regions of opposite magnetic polarity and within which coronal material condenses and flows down into the chromosphere. Sprays scatter material over a wide range of angles at speeds ranging from 200 km s^{-1} to as much as 2000 km s^{-1}, whereas surges consist of more closely collimated ejections of material. Giant eruptive prominences (often loop-shaped) surge upwards to heights of 500 000 km or more.

When solar rotation carries prominences across the face of the Sun, they then appear as dark features against the bright solar disk. Quiescent prominences correspond to the extensive, long-lived, dark filaments that feature on monochromatic images of the solar disk.

See also: active region, chromosphere, corona, emission spectrum, filament, flare.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proper Motion

Proper motion is the change of position of an object on the celestial sphere, usually measured in seconds of arc (″) per year (figure 1). 'Proper' means intrinsic, as opposed to a motion due to change of coordinates (such as precession). The term is most often applied to stars but is also used for changes in position of clusters of stars, galaxies outside our Milky Way and features within these objects. In 1718 Edmund Halley was the first to recognize proper motions, when he realized the positions relative to nearby stars of Arcturus, Aldebaran and Sirius had changed since antiquity. Proper motion is an angular measurement and depends directly on an object's transverse velocity (the component of velocity across the line of sight) and inversely on its distance. For a given proper motion, the more distant an object is, the higher its transverse velocity must be (figure 2). Barnard's star has the highest proper motion known, 10.″3 yr^{-1}. The term also refers to changes in position of features observed at infrared and radio wavelengths. Proper motions are determined by measuring positions with respect to other objects, typically stars or galaxies. If the comparison objects cannot be assumed to be stationary on the sky, a correction for their motion must be made to determine the absolute proper motion. Together with the RADIAL VELOCITY, the proper motion defines the motion of an object relative to the SUN. If the distance to the object is known, the proper motion and radial velocity define the three-dimensional space velocity of the object, again relative to the Sun.

Proper motions are used to study stellar motions and the correlation, if any, between motions and age, metal abundances and spectral types of stars. Proper motions are also used to calculate the rate of revolution around the center of our Galaxy. Recently proper motions have been applied to the investigation of features associated with young stars, the possible existence of a massive black hole at the center of our Galaxy and the motion of objects in other galaxies. The latter investigations show promise for the direct determination of the distances to other galaxies. Space missions currently under discussion will further extend investigations of objects in our own and other galaxies through the use of proper motions.

There are several components to the motion of stars in our Galaxy. Stars in the galactic disk, like the Sun, partake of a general revolution in a common direction around the center of the Galaxy (see GALACTIC THICK DISK, GALACTIC THIN DISK). Stars in the halo revolve around the galactic center in orbits with random eccentricities and orientations. In addition to the motion around the galactic center, disk stars each have their own peculiar, or random, motions, typically 10–20 km s^{-1}. Peculiar motions are defined with respect to the local standard of rest, a reference frame in which the averaged random velocities of stars near the Sun is zero. The local standard of rest moves around the center of the Galaxy at the velocity corresponding to galactic rotation at the Sun's distance from the galactic center. The apparent motion of stars relative to the Sun is a combination of their motion and the Sun's motion. The position of objects on the sky changes owing to their relative motion perpendicular to the line of sight with respect to the Sun.

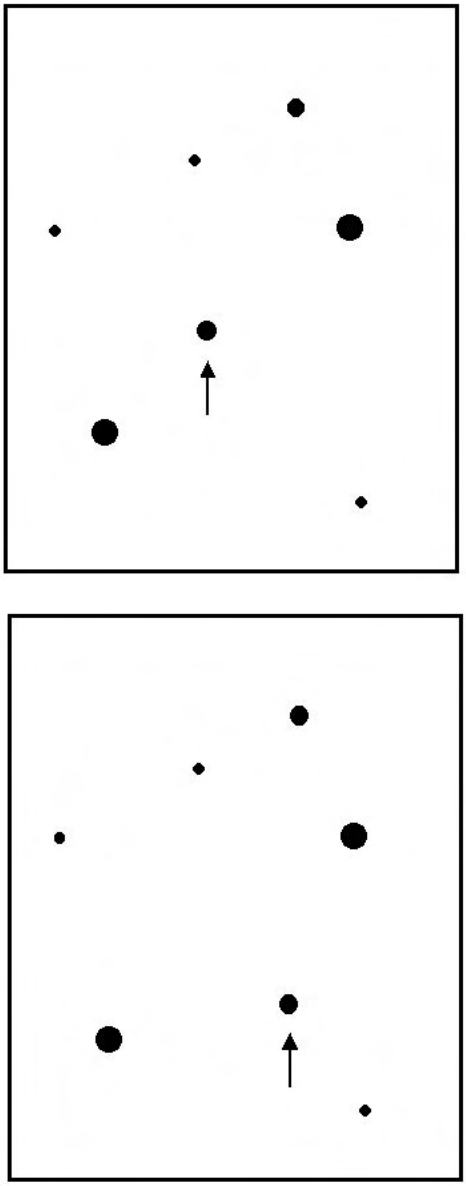


Figure 1. An example of the change of position on the sky of a star, indicated by arrows, relative to other stars.

Because more distant stars tend to have small proper motions, in the past they were used to define a coordinate system in which to measure the motions of nearby stars. As measurement techniques improve, even the proper motions of distant stars can be detected. Modern measurements use the position of faint galaxies to define the reference frame. Motion along the line of sight directly towards or away from the Sun (the radial velocity) does

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

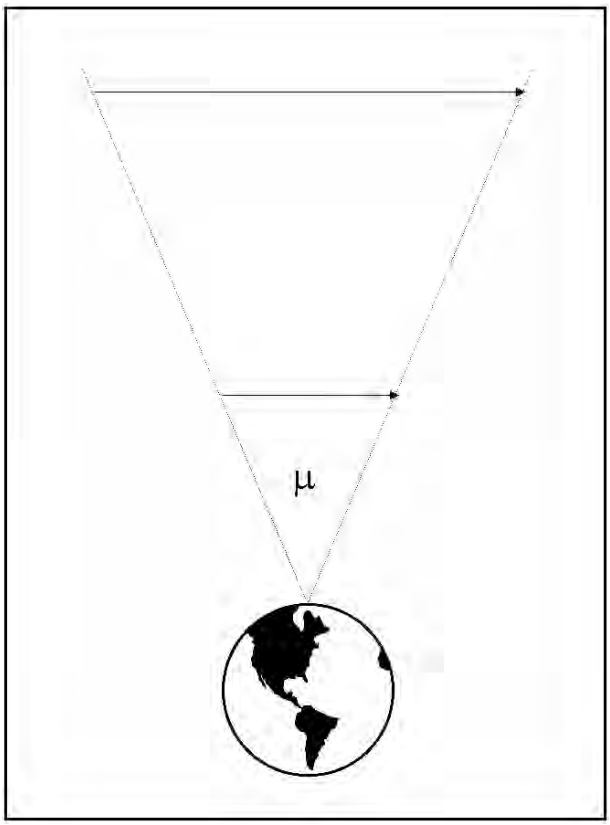


Figure 2. Proper motion, indicated by the Greek letter μ, is the angular change of position of an object on the sky, usually reported in seconds of arc per year. The angle shown is highly exaggerated; the maximum proper motion measured is $10\farcs3\ \mathrm{yr}^{-1}$.

not contribute to the proper motion. Motion relative to the Sun is usually reported, even though measurements made from the Earth contain a component due to the Earth's motion around the Sun, which must be removed.

Measurement

While the concept of proper motion is straightforward, proper motions are difficult and time consuming to measure. In principle, two images of the object of interest separated by 1 yr, to remove the apparent change of position on the sky due to the Earth's revolution around the Sun (see PARALLAXES), are sufficient. The position of the object relative to other objects in the field would then be measured. In practice, because the annual changes in position are small, successive images are often separated by several years or decades. Since one does not know in advance which objects will have detectable proper motions, proper motion surveys often use large photographic plates to observe as many objects as possible simultaneously. Large-area detectors also ensure that there are many stars and galaxies in the field to define the reference frame against which motions are measured. Positions of objects on the images must then be compared. Historically, this was done manually with a blink comparator, a device which allowed the astronomer to look first at one image, then at the other, to see which objects appeared to move relative to the others. Now photographic plates are automatically scanned with densitometers, the output of which is fed into computers which find and catalog stellar positions. Computerized comparison of positions on plates taken at two different times identifies objects whose positions have changed significantly and the proper motion is calculated. Because the changes in position are very small and errors can easily be introduced into the results, it is advisable to take all data on the same telescope and, if possible, use the same instrument. If stars are used to define the reference frame for the measurements, the final results must be corrected for the mean parallax and proper motion of the reference stars.

Results of proper motion surveys are generally produced as catalogs containing positions, proper motions and often photometric information for thousands or tens of thousands of stars. Modern surveys use faint galaxies to define the reference frame, which can then be linked to the International Celestial Reference Frame, based on extragalactic RADIO SOURCES. Accuracies of 3–8 marcsec yr^{-1} are typical for ground-based observations, depending on the brightness of the star.

Techniques to improve this accuracy and thus obtain more precise proper motions are constantly being investigated. One source of uncertainty is the size of the stellar images on the photographic plates. Motion in the EARTH'S ATMOSPHERE blurs astronomical images as the light passes through. This blurring, referred to as seeing, produces stellar images larger and more diffuse than they would be outside the Earth's atmosphere and makes it difficult to determine the center of the image.

Problems associated with seeing can be minimized by obtaining a series of images, each exposed for a fraction of a second, and then combining those data to produce the final image. In these exposures a star image is broken up into a series of blobs, or speckles, produced as the light passes through cells in the atmosphere. The speckle pattern changes rapidly, producing the blurred image normally observed. By aligning the speckle patterns of all the images, the spatial resolution of the telescope is significantly enhanced. Because the exposures must be taken rapidly in succession, an electronically read-out detector must be used, so the area of sky observable is limited by the available detector arrays.

Space astronomy has also contributed to improved measurement of proper motions. Taken above the Earth's atmosphere, these images do not suffer from the atmospheric seeing. Space observatories also avoid the problem of deflection of the telescope, optics and instruments due to the gravitational attraction of the Earth. They can, however, be subject to large temperature fluctuations, which can also cause deflection problems unless controlled or counterbalanced.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In 1989 the EUROPEAN SPACE AGENCY launched the HIPPARCOS (High Precision Parallax Collecting Satellite) astrometric satellite, which measured the positions, parallaxes and proper motions of 118 000 stars to an average accuracy of 1 marcsec. Astrometric data were obtained for a second group of more than one million stars at lower precision. The satellite was designed to spin slowly, scanning the entire sky several times during its mission. While only early results of the Hipparcos mission have been reported, they are already having a large impact on our understanding of stellar motions and other fundamental questions in astronomy.

Scientific results

Nearby and high-velocity stars

One of the first applications of proper motions was to identify nearby and/or high-velocity stars, because stars with large proper motions are either nearby or have high transverse velocities relative to the Sun. Proper motions are an efficient way to identify intrinsically faint, nearby stars, which are otherwise difficult to distinguish from brighter stars that are more distant. Once nearby stars are identified, parallaxes and other data can be obtained. If several measurements of the stellar position are made and the star appears to follow a wavy path across the sky, the star has a fainter companion. In such systems, the center of mass follows a straight line, while the combination of star's motion around the system center of mass and the system's motion through the Galaxy produces the apparent wobble. Such systems can be used to derive stellar mass data. Astronomers are particularly interested in two classes of intrinsically faint stars: RED DWARFS and WHITE DWARFS. Red dwarfs are the faintest, yet most numerous, stars on the main sequence (stars which generate energy by converting hydrogen to helium in their cores). Red dwarfs are studied to determine the physical characteristics and space densities of the least massive stars. Because they are very numerous, red dwarfs are a significant component of the stellar mass in the Milky Way. A white dwarf is a stellar remnant, the remains of a star after it has consumed all of its fuel. Such stars are studied to test theories of the end state of STELLAR EVOLUTION and also to determine their contribution to the stellar density of the Galaxy.

Proper motions are also an efficient way to identify stars moving rapidly with respect to the Sun. Such stars probably originated outside the Galaxy's disk, and their high velocities may be clues to the star's location at its time of formation, its age, chemical composition and even the evolutionary state of the Galaxy when the star was formed. According to some theories of GALAXY FORMATION, the Milky Way formed from a large rotating cloud of gas which collapsed into a disk containing most of the stars. In this scenario, some stars formed during the collapse. On the other hand, recent work suggests that some of the stars far from the Galaxy disk originated in stellar satellites which were destroyed by gravitational interaction with our Galaxy. Whatever their origin, stars far from the disk in the galactic halo have large components of motion perpendicular to the galactic plane and relatively large (100–200 km s^{-1}) velocities compared with stars in the disk. Therefore stars with high proper motions may belong to a population that formed while the gas in the Galaxy collapsed or to a population that formed in star clusters or satellite galaxies accreted by our Galaxy. Investigations of these stars can reveal the physical conditions in their place of origin and thus how the galactic halo formed.

In principle, astronomers can identify high-velocity stars from radial velocities as well as proper motions. However, proper motion surveys have the advantage of screening large numbers of stars simultaneously. Until the advent of SPECTROGRAPHS with two-dimensional detectors, stellar radial velocities were measured individually. Even with modern electronic detectors, spectra of at most a few hundred objects can be obtained simultaneously. Measurements of positions and therefore proper motions can be obtained for thousands of stars simultaneously from imaging surveys.

'Runaway' stars are a second type of high-velocity star that can be identified from proper motions. These are stars that are too young to belong to the halo population whatever its origin. Of particular interest are runaway O and B stars, which have lifetimes of a few million to a few hundred million years. Such stars are found at greater distances from their birthplaces in the galactic disk than they could travel in their lifetimes at the normal peculiar velocities of such stars.

Two hypotheses have been suggested to explain the location of runaway O and B stars: they were ejected from their birthplace by a dynamical interaction among stars in a young, compact STAR CLUSTER or they were originally members of a BINARY SYSTEM containing two massive stars in which one star became a SUPERNOVA. Stellar evolution theory predicts that massive stars become supernovae when they have exhausted all of their nuclear fuel. Models indicate that, in binary systems with two massive stars, the system may be disrupted if the first supernova is asymmetric. Such a system could produce a high-velocity O star and a high-velocity PULSAR, the remnant of the exploded star. Even if the first supernova does not disrupt the system, the second supernova can, producing two high-velocity single pulsars. The two hypotheses make different predictions concerning the kinematic ages of the stars, the times since they left their birthplace. Kinematic ages can be estimated from the space velocity of the stars and the distance back to their point of origin or, if that is not known, to the plane of the galaxy. Recent investigations of O and B stars by Moffat and collaborators tend to favor the cluster ejection scenario, although the binary disruption mechanism cannot be ruled out.

Kinematics of stellar populations

One of the fundamental goals of galactic astronomy is to understand the spatial distribution and kinematic properties of the stars in the Milky Way. These data not only reveal the current state of these objects but also contain information about the gravitational potential of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the Galaxy and the history of STAR FORMATION. The proper motions, distances and radial velocities of a group of stars are needed to completely define their space velocities. The data are typically analyzed by comparing the observations with a model of the Galaxy and deriving the desired parameters. In some cases the model is expressed only in terms of transverse velocities, so that radial velocities may not be required. Quantities that can be estimated for the stellar group include the mean velocity and velocity dispersion perpendicular to the galactic plane, in the direction of galactic rotation in the plane and in the radial direction in the galactic plane. The velocity of rotation of the local standard of rest, the distance of the Sun from the galactic center, the peculiar velocity of the Sun and the density gradient parallel to the galactic plane in the solar neighborhood also can be determined.

Because different types of stars have different kinematic properties, it is important to use a homogeneous stellar sample to derive the velocities. Then one may study the average velocity perpendicular to the galactic plane as a function of spectral type or other parameter. One result of such studies has been an anti-correlation between the age of a stellar population and the degree of concentration of its space distribution to the galactic plane. Investigations like these led to the classification of stars into multiple STELLAR POPULATIONS. Recent work by Wielen investigated the mean velocities and velocity dispersions of RR LYRAE stars as a function of metallicity. (Astronomers refer to all elements other than hydrogen and helium as metals. It is generally agreed that on average older stars have lower abundances of metals than younger stars.) RR Lyrae stars are old pulsating stars often found in GLOBULAR CLUSTERS. Such investigations relate the increase of metals in the Galaxy, which is directly dependent on the rate of massive star formation, to the rate of collapse of the gas in the galactic halo. The deviation of the stellar velocity from the circular rotation velocity also may be investigated. It has been found that, at the Sun's distance from the galactic center, the rotational velocity derived from the youngest stars is closest to the circular velocity.

Statistical analysis of proper motions is used to determine the mean parallax, and thus distance, to a sample of stars. If the stars are selected to have similar spectral type and luminosity class, and thus absolute magnitude, the mean distance can be used to determine a value for the mean absolute magnitude. There are two methods to determine distances, secular and statistical parallaxes. For secular parallaxes, the averaged velocity of a sample of nearby stars is assumed to be zero. The mean proper motion of the sample is therefore just due to the average of the Sun's motion, 20 km s^{-1}, projected onto the plane of the sky at the positions of the stars. The average distance for the sample is directly calculated from the ratio of the distance traveled by the Sun during a year projected on the sky and the mean proper motion. Statistical parallaxes assume that the three-dimensional space velocities of the stars are isotropically distributed. If the radial velocities of the stars are known, by the isotropic assumption, the mean of the radial velocities must be equal to the mean of the transverse velocities. Since the proper motion is the distance traveled by a star in a year at its transverse velocity divided by its distance, the mean distance to the stars can be determined. More complex analyses for stars far from the Sun assume a model for motion of stars as a function of location in the Galaxy. The resulting statistical parallax is of course dependent on the model, just as the results described previously depend on the assumption that the average velocity of the sample stars is zero or the velocities are isotropic.

Statistical parallaxes are useful for stars too distant to measure individual stellar parallaxes. Statistical parallaxes have been used to estimate the absolute magnitude of RR Lyrae stars. Because these stars vary in a recognizable way, they can easily be identified in globular clusters, in the galactic halo or in nearby galaxies and are useful as a distance indicator. If their absolute magnitude is known, the distance to the cluster or galaxy can be determined. Distances to globular clusters are particularly important because the ages of the clusters are determined by the absolute magnitude and spectral type of stars beginning to evolve off the main sequence. Globular cluster ages are a lower limit to the age of the universe. RR Lyrae variables are also observed in nearby galaxies and can be used to determine their distances. Unfortunately they are relatively faint and can only be detected in the closest galaxies.

Star clusters and associations, nearby galaxies

Detailed studies of the proper motions of stars in clusters (groups of physically associated stars thought to have a common origin) and associations (looser groupings of young stars which are probably not gravitationally bound) can reveal a great deal about the membership and dynamics of these groups. Since clusters are first identified as a concentration of stars in a small area of the sky, proper motions enable astronomers to confirm membership of individual stars and even to establish that a cluster really exists and is not just a random superposition of stars along the line of sight. Stars which are not obviously part of the group, for example F stars in an O,B association, can be identified. The proper motion of the cluster can be combined with its radial velocity to calculate the space velocity. For nearby clusters such as the Hyades, these data can be used to improve the precision of the cluster's parallax and thus distance. For globular clusters, which are several billion years old, the space velocity provides information about the kinematics of the gas out of which the clusters formed. Proper motions can be combined with radial velocities to determine the distance to the cluster using the method of statistical parallax. Because even the closest clusters are somewhat distant, a long time baseline is required to measure accurately the proper motions of stars in clusters.

If the relative proper motions are sufficiently accurate, they can be used to investigate the motions within the star cluster. Because most clusters are distant and the relative

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

motions small and of limited accuracy, the dispersion of the proper motions is often used. If the dispersion can be calculated as a function of magnitude, which relates directly to stellar mass, and distance from the center of the cluster, the mass of the cluster can be estimated. The azimuthal and radial proper motion dispersions relative to the cluster center measure the amount of anisotropy in the relative motions, which tests theories of cluster dynamics.

Proper motions can also be used to study the kinematics of the MAGELLANIC CLOUDS, satellite galaxies gravitationally bound to the Milky Way. Proper motions combined with radial velocities determine the space motions of the galaxies and constrain the gravitational potential of the Milky Way.

Black hole at the center of the Galaxy

It has been suggested that the Milky Way, like many other galaxies, harbors a black hole at its center. In the 1990s the development of infrared array detectors and speckle interferometry has enabled astronomers to test this theory by measuring the proper motions of stars at the center of our Galaxy. Stars at the center of our Galaxy cannot be observed in visible light, owing to dust in the plane of the Galaxy along the line of sight. Infrared wavelengths are not absorbed to the same degree as visible light and can be used to detect stars obscured by dust. Eckart and Genzel (1996) and Ghez and collaborators (1998) have applied the speckle technique to near-infrared (2.2 μm) observations of the stars at the center of our Galaxy. The distribution of the proper motions of the stars is centered at the position of Sgr A*, a radio source previously suspected of being at the center of the Milky Way (figure 3). In general the stars closest to Sgr A* are moving fastest. The dispersion of the velocities obtained from the proper motions also peaks at the position of Sgr A*. If the stars are gravitationally bound, the central mass implied is 2.6×10^6 solar masses. While exotic alternatives such as clusters of elementary particles and very-low-mass black holes cannot be completely eliminated, the mass is most likely due to a supermassive black hole.

Radio proper motions

While proper motions were first observed and used at optical wavelengths, they can be measured at any wavelength for which repeated observations indicate motion on the plane of the sky. Since the late 1960s radio astronomers have used interferometry to obtain high-resolution measurements of the positions of radio sources (see RADIO INTERFEROMETER). Spatial resolutions of 1 marcsec are readily attained with modern arrays of radio telescopes. While there are relatively few strong stellar radio sources, radio astronomers have observed proper motions of features in the continuum radio emission from nuclei of QUASISTELLAR OBJECTS and galaxies. In many cases there appears to be transverse motion at several times the speed of light, probably because of orientation effects in a relativistic beam pointed towards the observer (within a few degrees of the line of sight) (figure 4). These

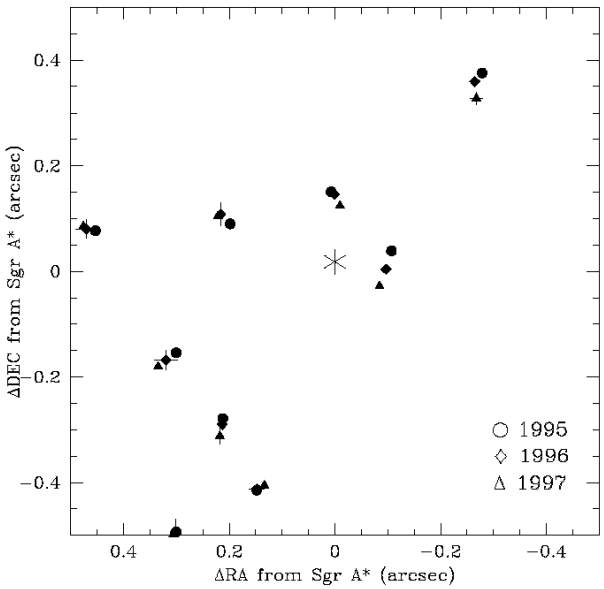


Figure 3. Measured positions of stars in 1 arcsec^2 centered on Sgr A*, indicated by starred point. Significant velocities are readily detected in this region. Each year's measurement is represented by a different symbol: 1995 by triangles, 1996 by squares and 1997 by circles. (Figure courtesy of A Ghez and American Astronomical Society.)

observations help constrain models for the evolution of radio sources.

Statistical methods have also been applied to the proper motions of radio sources to determine their distances and three-dimensional velocity field. Observations of star-forming regions in our Galaxy reveal that some of them contain numerous microwave radio sources called masers. These are small regions which emit enhanced molecular spectral lines, usually from OH or H_2O. The enhancement is thought to be associated with star formation. Starting in 1977, Genzel and collaborators measured radial velocities and changes in the positions of water masers in star-forming regions. Distances may be determined by two methods. If the motions are random so that all velocity components have the same dispersion, the distance can be directly determined from the ratio of the velocity dispersion in the radial direction to the dispersion of the proper motions. If the kinematics in the source is not random, the proper motions, radial velocities and positions on the sky are used to model the source in three dimensions. The distance is then calculated by scaling the observed proper motions to the transverse velocities predicted by the model. In this method not only is the distance to the source estimated but a model of the source is also determined. If the distance to the star-forming region is known *a priori*, the proper motions and radial velocities can be used to investigate the dynamics within the source. Attempts have also been made to apply these techniques to clouds of molecular masers in nearby galaxies, thus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

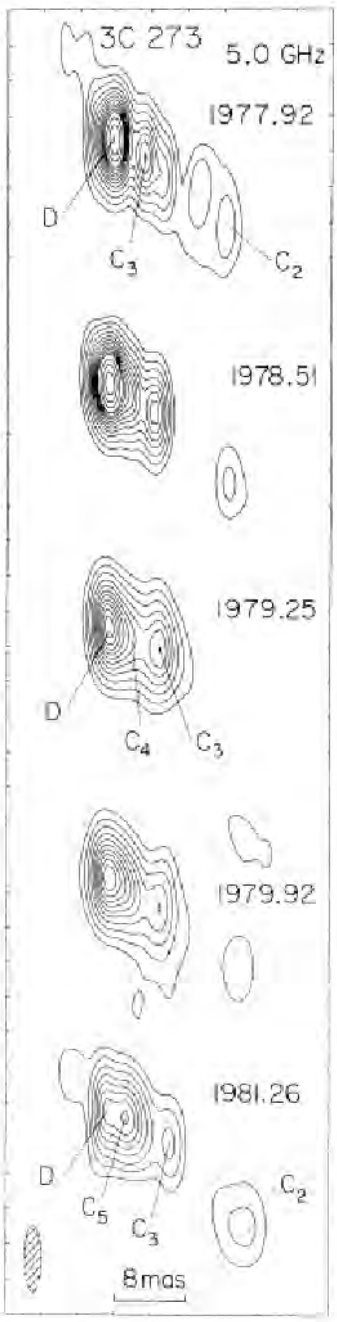


Figure 4. Map of the radio structure in the nucleus of the quasar 3C 273 measured at frequency 10.7 GHz. The horizontal scale indicated is 4 marcsec. Changes in positions of components C_3, C_4 and C_5 are readily visible. (Figure courtesy of S Unwin and American Astronomical Society.)

directly measuring their distances without recourse to assumed luminosity standards.

Proper motions of water masers associated with the centers of galaxies also show promise of enabling astronomers to directly determine distances to galaxies. This technique was used to determine the distance to the center of our Galaxy by Reid and collaborators in 1988. In 1995 Greenhill and collaborators detected water masers which apparently trace a thin, nearly edge-on disk rotating around a central mass in the galaxy NGC 4258. The rotation of the disk was predicted to cause the radial velocities and positions of the masers in the disk to change, with the change in position, i.e. proper motion, dependent on the distance to the galaxy. A series of radio observations indicated that the positions of the masers had indeed changed. Assuming a model for the rotation including the radial velocities, a distance to NGC 4258 was calculated based only on geometry. If such masers, appropriately positioned with respect to the plane of the sky, can be found in other galaxies, this will be a powerful new technique for directly determining distances to galaxies. Not only will the distances to the galaxies be calculated but the results will make possible the calibration of variable stars and other luminosity standards using objects in other galaxies, not just the Milky Way.

Future initiatives

In the future increasingly accurate proper motions can be expected from observations on the ground and in space. Efforts are underway to combine the results of the Hipparcos mission with older ground-based photographic plates to extend the time baseline and accuracy of proper motions.

With the success of the Hipparcos mission, several other space astrometry missions have been suggested. The two that seem most likely to fly at this time are SIM, the SPACE INTERFEROMETRY MISSION, proposed by NASA, and GAIA, Global Astrometric Interferometer for Astrophysics, under consideration by the European Space Agency. GAIA's goal is to assemble a catalog of more than one thousand million stars with three-dimensional positions, space velocities and related information. Astrometric measurements of position, parallax, and proper motions will be obtained for all stars to apparent visual magnitude 20, about one-millionth the brightness of the faintest star visible to the naked eye. Radial velocities for stars down to apparent visual magnitude 17 will also be measured. This mission is expected to measure distances to the center of the galaxy to 10% accuracy and catalog about 1% of the stars in the galaxy. Assuming a 5 yr mission, the predicted precision for proper motions of stars brighter than apparent visual magnitude 15 is better than 10 μarcsec yr^{-1}.

The SIM mission will measure proper motions of individual, bright stars in several nearby intermediate- and late-type spiral galaxies. Assuming the averaged proper motions and ground-based radial velocities are both due to the nearly circular rotation in the disks of the galaxies, rotational parallaxes, distances independent of luminosity-based distance indicators, can be determined. Directly measured luminosities of bright stars, including cepheid and RR Lyrae variables, will then be available for a variety of galaxies, strengthening the calibration of the distance scale. As proposed, the mission is expected

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to achieve an astrometric accuracy of 5 μarcsec or better, depending on the target. Other goals for SIM include determining positions and distances for stars throughout the Milky Way and probing nearby stars for Earth-like planets.

Bibliography

Genzel R, Reid M J, Moran J M and Downes D 1981 Proper motions and distances of H_2O maser sources. I. The outflow in Orion-KL *Astrophys. J.* **244** 884–902

Ghez A M, Klein B L, Morris M and Becklin E E 1998 *Astrophys. J.* **509** 678–86

Kovalevsky J 1998 First results from Hipparcos *Ann. Rev. Astron. Astrophys.* **36** 99–129

Monet D G 1988 Recent advances in optical astrometry *Ann. Rev. Astron. Astrophys.* **26** 413–40

Strand K Aa 1963 *Basic Astronomical Data* (Chicago, IL: University of Chicago Press)

Van Altena W F 1983 Astrometry *Ann. Rev. Astron. Astrophys.* **21** 131–64

Zensus J A 1997 Parsec-scale jets in extragalactic radio sources *Ann. Rev. Astron. Astrophys.* **35** 607–36

Donna Weistrop

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proper Motion: Optical/Infrared

The proper motion of a star, or other object, is its angular change in direction on the sky over a period of time, often expressed in arcseconds per year. Its definition is specific with regard to location of the observer, usually at the Sun, and the background frame of reference against which the motion is compared.

Most of the stars in our Galaxy occupy a flattened, disk-shaped region and central bulge, while the remainder are sparsely distributed in a spherical halo. Some 400 billion stars in our Galaxy, the Sun included, are revolving mostly in the same sense about a distant point, the galactic center, with different speeds and directions. The GALACTIC OPEN CLUSTERS and GLOBULAR CLUSTERS are also in orbit about the galactic center. Within each cluster each member star moves in an orbit about its center of mass. Even the non-stellar galactic nebulae and the dark absorption clouds are also in motion, as well as changing in size and shape. There is more. As we look to our neighboring galaxies, we are aware that they, too, are moving at different speeds and directions. Space motion is a universal characteristic, one component of which is the proper motion. Therefore, we need to measure proper motions in order to see the pattern of galactic motions for stars and clusters, as well as for stellar motions within clusters. The measurement of proper motions for nearby galaxies is becoming a reality that enables tracing their dynamical evolution and tidal interactions, including possible mergers with our Galaxy. These are forefront problems of astronomy.

Consider a star viewed from the Sun in a certain direction on the sky (figure 1). Over a time interval the space motion (S) of the star, pictured here in two dimensions, is described by a vector, like an arrow of certain length and direction. The component of this space velocity projected on the sky is called the *tangential velocity* (T), while the component along the line of sight is the *radial velocity* (V), all expressed in kilometers per second. Proper motion μ is simply the change in direction after some time interval, say one year, for a star moving at the tangential velocity. It may be expressed in units of arcseconds, milliarcseconds (mas), or microarcseconds (μas) per year, where 1 mas = 0.001 arcsec = 1000 μas. Simple formulae relate these motions with distance of the object from the Sun d in parsecs (1 pc = 3.26 light years).

$$S = (V^2 + T^2)^{1/2} \qquad \text{and} \qquad T = 4.74\,\mu\,d.$$

If we measure the distance, proper motion, and radial velocity of a star, we can compute its space velocity with respect to the Sun as origin. The distance and PARALLAX (π in arcsec) are inversely related by $d = \pi^{-1}$.

The observations are made from the surface of a rotating Earth (topocentric), which is revolving about the Sun, so that certain corrections are needed to specify a space motion with respect to the Sun (heliocentric), or to the center of mass of the solar system (barycentric). Observed positions represent the combined effects of parallax, significant for nearer stars, and proper motions. Observed radial velocities need corrections for the Earth's rotation and orbital motion and, possibly, motion of the Sun about the center of mass of the solar system. Other small corrections are relevant. Other definitions of the space motion of stars are used for studies of stellar kinematics, including origins located at the average motion of stars in the solar neighborhood, the Local Standard of Rest (LSR), or at the galactic center (galactocentric motion). Transformation formulae enable conversions of space motions amongst these origins and axes.

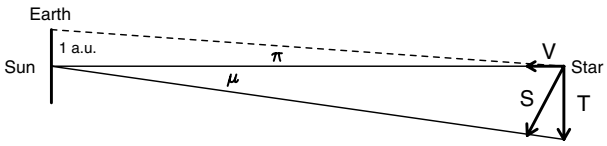


Figure 1. Space velocity (S) with components radial velocity (V) and tangential velocity (T). Proper motion displacement (μ) corresponds to the tangential velocity and parallax (π) is the angular change in direction of a star as seen from ends of a baseline of length one astronomical unit.

The proper motions for galactic stars reflect the wide range of their space velocities. The open clusters and the bulk of the stars near the galactic plane have heliocentric motions of low tens of kilometers per second, with some well over 100 km s^{-1}. Large motions prevail for globular clusters and many halo stars, which reach several hundred kilometers per second. Internal motions in globular clusters are typically around 5 km s^{-1}, while in open clusters it is only one-tenth of this.

BARNARD'S STAR, the second nearest after the α Centauri system, has the largest known proper motion of 10.36 arcsec per year and lies at a distance of 1.82 pc. It traverses an angle corresponding to the diameter of the Moon in only 175 yr. Interestingly, among the nearby stars, there is the RED DWARF Gliese 710 with an annual proper motion of only 0.003 arcsec (or 3 mas). Now at a distance of 19.3 pc, it is approaching the solar system with a radial velocity of about -18.6 km sec^{-1} and should pass at a miss distance of only 0.3 pc in about 1 million years. When closest to the Sun, its radial velocity becomes zero, while its tangential velocity equals its space velocity and its proper motion reaches a maximum of 13 arcsec per year. This change in size of proper motion illustrates the effect of *perspective acceleration*, now detectable for very few stars. Future astrometry from space, with sub-milliarcsecond levels of accuracy, will be able to measure this effect for many stars.

Historical background

In 1718 the English astronomer Edmond HALLEY discovered that stars exhibit proper motions, when he compared recent positions observed by TYCHO BRAHE and others in the 1580s and 1590s with those in the Almagest, the early Middle Age Arabic version of the Greek PTOLEMY's star

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

catalog (AD 138). The positions listed by Ptolemy may actually be much older and refer to those observed by the Greek astronomer HIPPARCOS (ca. 130 BCE), corrected for PRECESSION to AD 138. Halley found that some stars showed large residual shifts in excess of expected errors, two of these being the bright stars Aldebaran and Betelgeuse. The discovery of the proper motions had no immediate impact. It was only in 1738 that Halley's proper motion of Arcturus was confirmed by JACQUES CASSINI. The first catalog containing proper motions for many stars came a few years later with the work of Tobias Mayer, who compared RØMER's observations of 1706 with those made by himself and LACAILLE in the 1750s. Positional astronomy grew over the following 250 years with ever-improving instrumentation and observing techniques.

The legacy from this long period is the many star catalogs containing the equatorial coordinates right ascension and declination, based on meridian circles and related devices for observing positions for the brighter stars to magnitude 8–9. Unfortunately, this limit was too bright to include the remote galaxies, which would have provided fixed anchor points against which to measure positions of moving stars. Instead, fundamental observations used the Sun and some relatively nearby stars to locate the positions of the moving celestial equator and equinox point, as a practical way of defining the equatorial coordinate system and its progressive change with time. The secular effect of precession and the periodic effect of NUTATION measure this change. The practice of correcting for systematic differences among catalogs was begun by Auwers after 1865. It led to a series of catalogs, each defining a new version of the Fundamental System of reference with improved precision and accuracy, culminating with the final *Fundamental Katalog FK5* (1988), containing nearly 5000 stars around the sky. A new level of accuracy and precision was achieved in 1995 by the Hipparcos catalog, which is linked to the extragalactic frame and adjusted to the framework of the FK5 at epoch J2000.

Determination of proper motion

The tools of astrometry take various forms. The photographic plate has dominated direct imaging for over a century. The final decade of the 20th century saw electronic imaging sensors, such as the charged-coupled device (CCD), infrared detectors, and others, singly or in mosaic form, becoming the detectors of preference. Their advantages include high quantum efficiency, direct digital recording, and rapid real-time data processing by computers. Imaging detectors sensitive to the infrared domain offer great opportunities as astrometric probes in optically opaque regions of the Galaxy. Infrared observations take advantage of 'windows' in the electromagnetic spectrum, where the Earth's atmosphere is relatively transparent, such as those at near-infrared wavelengths 1.25, 1.65, and 2.20 microns (μm) for the J, H, and K bands, respectively. The HUBBLE SPACE TELESCOPE (HST) provides capabilities in optical and near-infrared regions for the measurement of proper motions and parallaxes with precisions at least comparable to the *Hipparcos Catalog*, but reaching to much fainter magnitudes. Optical INTERFEROMETRY shows the feasibility of differential astrometry to very high precision (tens of μas) in very small fields, with further advantages expected by observing in the near-infrared region using large telescopes.

The determination of proper motions is based on the comparison images of a star field obtained at suitably spaced epochs of observation. The images may be photographic plates or frames from electronic detectors. Precision coordinate measuring machines provide rectangular coordinates from photographic plates, while software associated with electronic detectors provides coordinates directly. The term plate is used below for images.

Basically the proper motion is resolved into components, nominally aligned along right ascension and declination, expressed by coordinate differences divided by epoch difference Δt, as $\mu_x = \Delta x/\Delta t$ in x and $\mu_y = \Delta y/\Delta t$ in y. In practice, the case is more complex. Each plate is exposed and measured with small differences in centering and alignment of axes. The telescope contributes effects from guiding errors, optical distortion, flexure, plate tilt and others, which are not precisely the same for each exposure. Variations in atmospheric refraction and chromatic effects, depending on the color of the star, contribute to the imaging on each plate. The plates may even be taken with different telescopes. These effects must be accounted for in astrometric reductions.

In order to map precisely each plate of a series of two or more plates into one selected as a standard frame, we use a model that accounts for the above-cited effects. For a pair of plates ($\Delta t = t_2 - t_1$), we have in each coordinate the form

$$x_2 - x_1 = \mu_x \, \Delta t + F(x, y, m, c; a_i) + \epsilon \qquad i = 1, n$$

where the model F is some function of position (x, y), stellar magnitude (m), and color (c). The n unknown parameters a_i are called plate constants, while ϵ is a residual error.

The simplest case is a linear model, often adequate for small fields, where we have for the x coordinate (similarly in y)

$$\Delta x = \Delta t \, \mu_x + a_1 \, x + a_2 \, y + a_3 + \epsilon_x$$

with unknown plate constants (a_1, a_2, a_3). Two basic approaches exist for finding the plate constants. If a subset of the measured stars is also contained in a star catalog of positions and proper motions, they may be used as reference stars to evaluate the plate constants and, hence, permit the calculation of proper motions, close to the astrometric system of the parent catalog, for the other stars in the field.

If no external astrometric data exists, then we fall back to a solution for *relative* proper motions, which has an unknown zero-point offset. In practical terms, we use all

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(or part) measured stars as reference stars and, assuming their average proper motion is zero, solve for *relative* plate constants and then compute relative proper motions for all stars. However, if the measurements include faint, compact galaxies, we may compute their mean 'fictitious proper motion', taken with opposite sign, and apply it as a zero-point correction to convert from relative to absolute proper motions for the stars. An alternative statistical method (*mean secular parallaxes*) may be used to estimate the zero-point correction—the parallactic motion component of the reflex effect of solar motion—evaluated in the direction of the field and at the mean apparent magnitude and mean distance of the reference stars.

Techniques for evaluating the plate constants include a simple least squares adjustment for a pair of plates. The more complex method of plate overlap, pioneered by Eichhorn (1960) and now used extensively, solves simultaneously by least squares for both plate constants and star constants (proper motion and parallax) for a series of plates.

Some applications

Cluster membership

Proper motions provide a means for isolating members of open and globular clusters from foreground and background field stars. Goals include the construction of color–magnitude diagrams or studies of individual members. The internal motions in clusters are generally very small. This translates into very nearly equal and parallel proper motions, in contrast to the wide range of motions displayed by the surrounding field stars. The proper-motion vector point diagram, a plot of the two components of (usually) relative proper motion, forms the basis for estimating stellar membership in star clusters. The diagram displays two distributions of points with centroids representing the cluster (usually smaller dispersion) and the field stars. Estimates of membership have been made for some 100 open clusters and about 15 globular clusters. Recent observations with large ground-based telescopes and with the Hubble Space Telescope (HST), focusing on small regions, particularly the cores of globular clusters, are raising the astrometry to new levels of precision and magnitude limits unattainable by traditional photography (see CLUSTER MEMBERSHIPS).

Space motions of clusters

The trajectories of clusters around the galactic center may be computed from their space motions. These orbits provide clues to the origin and evolution of the Galaxy. For globular clusters the issue concerns correlations between cluster metallicity, as age indicator, and cluster kinematics—whether the motion supports membership in the halo or disk population. Motions have been traced back over the age of the Galaxy, to expose the changing character of the orbits with time. For the generally much younger open clusters the trajectories provide information about their places of origin in the framework of spiral structure as it was in the past and correlations of their atmospheric metal contents with galactocentric distance.

The number of globular clusters with measured absolute proper motions of quality ranging from useful to submarginal is about 35 and growing. Many are measured on photographs using galaxies or stars from the *Hipparcos Catalog* as reference frame, while others are based on a statistical reduction for conversion to absolute motion. Absolute proper motions have been measured on photographs for more than 200 open clusters, using photographic star catalogs for the reference frame, as most clusters lie in the Milky Way. The *Hipparcos Catalog* provides astrometric and photometric data for 241 clusters, ranging from one star per cluster to over 100 stars (Hyades).

Astrometric distances of clusters

The history of estimating distances to star clusters, using various astrophysical methods, is a long one. An alternative astrometric method for estimating distance involves the comparison of dispersions in proper motion and radial velocity, using the simple expression

$$d = \frac{\sigma_v}{4.74\,\sigma_\mu}$$

where σ_v and σ_μ are the error-corrected dispersions in radial velocity (km s^{-1}) and proper motion (arcsec yr^{-1}), respectively. The astrometric method provides an independent way for finding distance that is insensitive to the interstellar medium. A more realistic approach that accounts for omissions in the simple method uses dynamical models of the cluster. Few clusters have been calibrated by these methods so far.

Space motions of nearby galaxies

Our Galaxy belongs to a clustering of some 30 known members called the LOCAL GROUP. First steps towards measuring the trajectories for some nearer members have been made. These reveal the complex nature of interactions of our Galaxy with the MAGELLANIC CLOUDS and other neighbors. The absolute proper motion measured for the LMC in 1994 reveals it is gravitationally bound to the Galaxy and now lies at perigalacticon on a strongly eccentric orbit. Results using the *Hipparcos Catalog* for both LMC and SMC provide independent measurements of the motions. These indicate that more precise absolute proper motions are still needed to firmly establish the nature of the motions of the Clouds.

Absolute proper motions measured for other satellite galaxies (the nearer dwarf spheroidal systems Ursa Minor, Draco, and Sculptor) show them to be generally moving along one of two great circles on the sky, either the MAGELLANIC STREAM or a second stream with the Fornax dwarf spheroidal galaxy. The motion for the SAGITTARIUS DWARF SPHEROIDAL galaxy, discovered in 1994, shows that it is a tidally distorted system now near perigalacticum, only 16 kpc from the galactic center. The motions of galactic satellite systems provide evidence for the mass and mass distribution in the Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

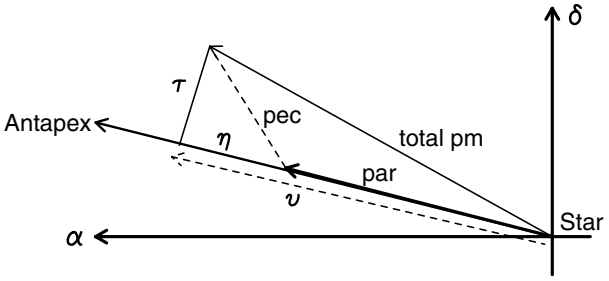


Figure 2. Total proper motion resolved along (υ) and normal (τ) to great circle from star to antapex of Solar Motion. Peculiar motion (pec) is resolved into τ and η components. The parallactic motion (par) is the reflex solar motion.

Solar motion and galactic rotation

The effect of galactic rotation is that *on average* stars closer to the galactic center than the Sun move faster around the center, while stars farther from the center than the Sun orbit more slowly. This differential galactic rotation effect, measured with respect to the average motions of stars in the Sun's neighborhood, called the Local Standard of Rest (LSR), may be expressed by mathematical models as a function of absolute proper motions and radial velocities. The simplest is the linear Oort model, valid for nearer stars with circular orbits in the galactocentric plane. It provides the Oort constants A and B, which enable computing the angular velocity and circular velocity in the solar neighborhood.

Statistical luminosity calibration

The observed total proper motion is resolvable onto orthogonal axes, one aligned along (υ) and the other at right angles (τ) to the great circle passing through a star and extending from apex to antapex defined by the solar motion (figure 2). The two proper motion components form the basis of two variations of statistical methods for finding mean parallaxes and distances for specific classes of stars, usually applied to those with narrow range in luminosity, and, therefrom, the calibration of their mean absolute magnitudes. Certain classes of stars have too few members near enough for the accurate measurement of individual parallaxes. Consequently, statistical methods based on proper motions provide alternative approaches.

The mean *parallactic motion* for a stellar sample is found from averaging the υ components. For the special case of a small region at angular distance λ from the Solar apex, we have the mean $\langle\upsilon\rangle = \langle\mathrm{par}\rangle + \langle\eta\rangle = \langle\mathrm{par}\rangle$, because, by definition of solar motion, the average peculiar motion of the stars (here η component) is zero with respect to their LSR. The mean parallax for the region, where $V_\odot$ is the solar motion, is

$$\langle\pi\rangle = \frac{4.74\,\langle\mathrm{par}\rangle}{V_\odot \sin\lambda}.$$

The *statistical parallax* is based on τ components, which represent pure peculiar motions of stars. It is measured by the ratio of mean $\langle|\tau|\rangle$ and mean peculiar radial velocity $\langle|V|\rangle$ components. The practical determination of mean parallax by either method is more complex, in that the distribution of peculiar space velocities is generally ellipsoidal. This means that forming averages of proper motions and radial velocities for stars located over the sky must be done carefully using the tools of statistical astronomy.

Nebular astrometry

Normally we regard proper motions as properties of stars and stellar systems. Proper motions can also refer to the angular movement of material in the stellar environment. Nebular astrometry has its roots in photographic imaging extending back to the early years of the 20th century, with observations of expanding shells and light echos of NOVAE. Observed proper motion expansions prevail in several classes of objects, including novae, SUPERNOVAE, PLANETARY NEBULAE, YOUNG STELLAR OBJECTS (Herbig–Haro), and various classes of luminous stars undergoing mass loss. These motions are observable in the optical and other wavebands.

Measurements of proper motion expansion provide timescales and evidence for underlying stellar processes. Proper motions combined with radial velocities of shells, or other features, provide a valuable geometrical method for measuring distances and luminosities of central stars that is independent of corrections for intervening interstellar absorption. Moreover, observations of supernovae provide an independent measure of distances to galaxies. The advent of exceptional resolution provided by HST and ground-based CCD imaging of non-stellar features is leading to a new level of precision unattainable by the older photographic technique.

Infrared astrometry and catalogs

Historically color-sensitized photographic plates have recorded beyond the visual wavelength limit near 0.7 micron up to a near-infrared limit of about 1.1 micron. The record shows scant precision photographic astrometry in this wavelength domain. However, infrared astrometry is emerging as cameras are being developed with electronic detectors possessing sensitivity, imaging characteristics, and field size suitable for precision measurements. The ability to penetrate deeply into a variety of optically highly opaque regions of the Galaxy is their prime advantage. These include star-forming regions, enabling studies of the kinematics of stars in the process of formation. Fields are now being exploited through proper motions of both stars and associated non-stellar matter, such as jets, streams, shells, and other parts. An outstanding success of near-infrared K-band astrometry is the measurement of relative proper motions of stars near the optically highly opaque galactic center using the Keck 10 m telescope. The motions support a massive BLACK HOLE candidate. Systematic large-scale infrared sky surveys are unveiling now-hidden populations throughout the galactic disk and bulge and to the far side of the Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Infrared surveys, such as *IRAS*, provide large numbers of sources, including many galactic stellar objects. The large positional uncertainty at a single epoch precludes their use for measuring accurate proper motions. However, cross-identifications with large optical catalogs, containing proper motion and other data, enables these catalogs to be used for kinematical studies. An example is the *Infrared Astrometric Catalog (IAC)*, based on space observations made with the Mid-Course Space Experiment telescope Spirit III at a wavelength of 8 microns. It forms a new infrared database of 177 860 stars, containing improved positions as well as proper motions and other data compiled from optical catalogs. Other surveys in progress, giving positions at the 0.5–1.0 arcsecond level for over 100 million objects, include the Sloan Digital Sky Survey (SDSS), Deep Near Infrared Survey (DENIS), 2 Micron All Sky Survey (2MASS), and Infrared Space Observatory (ISO). Eventually observations for high-precision proper motions will enable greater exploitation of these databases.

Astrometric catalogs

The precision of a catalog is characterized by the standard errors of its (equatorial) coordinates (ϵ_c) at its mean epoch t_c and its proper motions (ϵ_μ). The inexorable temporal degradation in accuracy of positions computed at some epoch t away from the catalog epoch ($\Delta t = t - t_c$) is given by the law of error propagation

$$\epsilon_t = [\epsilon_c^2 + (\epsilon_\mu \, \Delta t)^2]^{1/2}.$$

For the *Hipparcos Catalog* with errors of 0.8 mas in positions and proper motions at 1991.25, the error in computed positions grows to 8 mas after 10 years. This tells us that catalogs have finite useful lifetimes, as a source of positions projected to distant dates.

During the preparation of the *Hipparcos Catalog* it became apparent that binary and multiple stars could affect both the derived proper motions and the estimates of the total errors. This comes from the fact that the observations for the Hipparcos program spanned only ~3.5 years, while many binary stars have Keplerian orbital periods ranging from much shorter to much greater values. The measured instantaneous positions on the sky are the sum of the proper motion of the center of mass of the system and the location of each stellar component on a sky-projected orbit. The picture is more complex. The Hipparchos detector measures the equivalent of a photocenter, which at one extreme senses only the brighter component of a system where the fainter one lies below the detection limit. At the other extreme the measured photocenter senses the partially merged light of two components of comparable brightness. Moreover, one or more components may be variable in brightness, thus complicating the interpretation of the displacement of the photocenter with time. If observations of photocenters are made in different wavebands, for stars of different colors, a color-induced effect may also be significant. Therefore, the total positional error in star positions projected to other epochs may be dominated by an orbital-motion-induced cosmic error, in excess of the measuring error. Conversely, a comparison of Hipparcos proper motions with those based on traditional decades-to-century-long baselines provides an independent way of discovering close binary stars. New techniques of statistical astrometry were developed to complement high-precision space astrometry made over short time baselines.

The following sections describe major astrometric programs and star catalogs currently available or in progress. Table 1 summarizes these catalogs and provides references. Data centers or observatories providing these catalogs usually include detailed information for the user.

Surveys for large proper motions

Surveys to faint limits have uncovered many stars of large proper motion, mainly nearby red and white dwarfs and subdwarfs. These are useful collectively for stellar statistics of the solar neighborhood and individually for astrophysical observations. Stars of larger proper motion, typically over about 0.1–0.2 arcsec per year, are discovered by comparing photographs taken a decade or more apart. Major recent surveys include those of Luyten (Minnesota), Giclas group (Lowell Observatory), Wroblewski group (Cerro Calan Observatory), and others. About 70 mainly bright stars have proper motions over 2.0 arcsec per year. Some 4000 stars have annual motions over 0.5 arcsec (over half found by Luyten), while 60 000 stars exceed 0.2 arcsec (70% found by Luyten). An estimated 500 000 stars have annual motions over 0.10 arcsec. Digitized Schmidt and other sky surveys will provide a full count of proper motions above their error thresholds.

Astrograph surveys

The Lick Northern Proper Motion (NPM) and Yale/San Juan Southern Proper Motion (SPM) programs provide absolute proper motions for large numbers of faint stars, based on photographs exposed 2–4 decades apart. These use faint galaxies to approximate an inertial frame for the non-Milky Way sky and the Hipparcos system in the Milky Way zone. The NPM program covers the sky northward of declination -23°, while the SPM program extends southward from -17°, with a some fields north of this. Both programs, still in progress, have produced large astrometric catalogs: *NPM1* with 149 000 stars for the non-Milky Way sky and *SPM 1.0* with 59 000 stars in a 720 deg^2 region around the south galactic pole. Both programs include large numbers of special stars selected for their astrophysical and kinematical importance.

Similar surveys include the Faint Star Catalog (*KSZ*) plan (Pulkovo), involving photography at several observatories for 205 fields north of declination -25°. The separate parts, when completed at Kiev, will form the *General Compiled Catalog of Absolute Proper Motions (GPM)*. The Main Meridional Section of the Galaxy (MEGA) program at Kiev, including contributions from Potsdam, contains fields along an arc linking the galactic poles

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Large astrometric catalogs.

Catalog	Source	Date	Mag Color	No. objects	Content†
HIP	ESA	1995	12.4V	118 218	p/pm
TYC	ESA	1995	11.5V	1058 332	p/pm
TYC2	Høg+			2.5 million	p/pm
AC2000	USNO	1998	12.5B	4621 186	p
ACT	USNO	1997	11.5V	988 758	p/pm
TRC	Høg+	1998	11.5V	990 182	p/pm
TAC 1.0	USNO	1996	12.5V	705 679	p/pm
TAC 2.0	USNO				p/pm
Starnet	ARI	1996	12.5B	4.3 million	p/pm
GSC II	STScI			2 billion	p/pm
DPOSS	CalTech			2 billion	p/pm
APS	Minn	1997	21B,20R	40 million	p
APS-Luyten	Minn		20R		p/pm
USNO A1.0	USNO	1997	20-22BVR	488 million	p
USNO A2.0	USNO	1998	20-22BVG	526 280 881	p
USNO B	USNO				p/pm
UCAC-S	USNO		16VR	40 million	p
NPM1	Lick	1993	19B,18V	148 996	p/pm
NPM2	Lick		19B,18V	300 000	p/pm
SPM 1.0	Yale/S.Juan	1998	19B,18V	58 800	p/pm
SPM 2.0	Yale/S.Juan		19B,18V	320 000	p/pm
GPM	Kiev		15.5B	60 000	p/pm
IAC	MSX	1996	K-band	61 242	p

† Positions (p) and proper motions (pm).
References:
HIP & TYC: *The Hipparcos and Tycho Catalogu* 1997, ESA SP-1200
TYC2: Høg 1997 *ESA Symp. Hipparcos-Venice '97*, ESA SP-402, 25
AC2000: Urban 1997 *Astron. J.* **115** 1212
ACT: Urban 1998 *Astron. J.* **115** 2161
TRC: Kuzmin 1999 *Astron. Astrophys. Suppl.* **136** 491
TAC 1.0 & TAC 2.0: Zacharias 1996 *Astron. J.* **112** 2336
Starnet: Röser 1996 *IAU Symp. No.* 172, 481
GSC II: Morrison 1997 *ESA Symp. Hipparcos-Venice '97*, ESA SP-402, 73.
DPOSS: Djorgovski 1998 *Bul. Am. Ast. Soc.* **30** 1270
APS: Cornuelle 1997 *ASP Conf.* **127** 55
APS-Luyten: Humphreys 1997 *ASP Conf.* **127** 41
USNO A1.0 & USNO B: Monet 1997 *ASP Conf.* **127** 31
USNO A2.0: Monet 1999 *Bul. Am. Astron. Soc.* **30** 1427
UCAC-S: Zacharias 1997 ESA Symp. *Hipparcos-Venice '97* ESA SP-402, 177
NPM1: Klemola 1987 *Astron. J.* **94** 501
NPM2: Hanson 1997 *ASP Conf.* **127** 23
SPM 1.0: Platais 1998 *Astron. J.* **116** 2556
SPM 2.0: van Altena 1990 *IAU Symp. No.* 141, 419.
GPM: Rybka 1997 *Astron. Astrophys. Suppl.* **121** 243
IAC: Egan 1996 *Astron. J.* **112** 2862

for studies of stellar distributions and kinematics. The Potsdam program uses the Tautenburg Schmidt telescope for deep photography in numerous regions over the northern sky.

Hipparcos/Tycho catalogs

The Hipparcos and Tycho Catalogs provide positions, proper motions, photometry, and other data in the Hipparcos system, which is regarded as a representation of the ICRS system in optical wavelengths. This is a large catalog of unprecedented astrometric precision in optical wavelengths. For the brighter Hipparcos stars ($H_p \leq 9$) the median precisions are 0.8 mas for position, 0.8 mas yr^{-1} for proper motion, and 1.0 mas for parallax. For the much larger *Tycho Catalog* the median precisions for brighter stars ($V_t \leq 9$) is 7 mas, while for all stars it is 25 mas. A new version *Tycho2 Catalog* will have 2.5 million stars, including proper motions derived from re-reduced *AC* positions.

The adopted astrometric reference system is now the International Celestial Reference System (ICRS). The ICRS is that astrometric system that closely approximates the conventional equatorial coordinate system for mean equator and equinox of J2000 and the fundamental FK5 system. It is represented by the *Hipparcos Catalog* for stars and by the nearly inertial extragalactic frame defined at radio wavelengths by a special set of more than 200 (later 600) faint quasars and compact galaxies uniformly distributed around the sky. The ICRS system is used

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hereafter for both optical and radio observations to ensure compatibility of their reference systems.

Extensions of Hipparcos system

The *Astrographic Catalog (AC)*, also known as *Carte du Ciel (CdC)*, was a massive program started in 1887 by some 22 observatories to photograph and measure positions for stars to magnitude about $B = 12.5$. The resulting observatory catalogs contain rectangular coordinates of varying precision for over 4.5 million stars, spanning a very wide range of epochs around the average near 1905. These positions, now in computer-readable form, have been reduced to equatorial coordinates nominally in the Hipparcos (ICRS, J2000) system. One of several versions, completed or in progress, is the *AC2000*.

The *ACT Reference Catalog (ACT)* and *Tycho Reference Catalog (TRC)* are parallel catalogs based on the same material, the *Tycho* and *AC*, but using different strategies for reducing the *AC* into the Hipparcos system for deriving improved proper motions for the Tycho and other stars. The *ACT* uses *AC2000*, while *TRC* uses its own version.

The *Twin Astrographic Catalog (TAC) Version 1.0* is based on photography at the US Naval Observatory (USNO) for the sky north of declination -18° with epochs 1977–1986. It is an independent catalog with higher density than *Tycho*, where proper motions are found by comparison with the *AC* reduced to the (FK5, J2000) system. A new version *TAC 2.0* reduced to the Hipparcos system is in progress. The *US Naval Observatory CCD Astrometry Catalog (UCAC-S)* represents a CCD survey in progress to produce a very high density catalog in the Hipparcos system for the southern sky to magnitude 16.

The high-density *Starnet Catalog* and *Four-million Star Catalog of Positions and Proper Motions (4M Cat)* are parallel in using the *HST Guide Star Catalog 1.0* and reduced *AC*, referred to the *Positions and Proper Motions Catalogue (PPM)*.

Schmidt telescope surveys

The digitization of northern and southern photographic surveys obtained at several epochs with Schmidt telescopes forms massive databases of positions and photometry, and (in progress) proper motions, to faint magnitudes.

The Minnesota *APS Catalog* is based on copies of the *Palomar Observatory Sky Survey I (POSS-I)* plates for galactic latitude $|b| > 20^\circ$. The *APS-Luyten Proper Motion Catalog*, in progress, includes red Palomar plates taken by Luyten in the 1960s to derive absolute proper motions with respect to galaxies.

The *USNO A2.0* catalog provides positions in the ICRS system from scans of POSS-I, UK(SRC), and ESO Schmidt plates made at the USNO (Flagstaff). The Space Telescope Science Institute (STScI) digitized the POSS-II and UK Schmidt plates intended for the second generation *Hubble Guide Star Catalog (GST-II)*, containing positions, proper motions, and photometry. It is organized on the sky (32 768 equal areas) for efficient cross-matching with other optical surveys and data from other wavebands. The *DPOSS* (Caltech) is the digitized POSS-II for two billion stars and 50 million galaxies.

Future prospects: space astrometry

The success of the Hipparcos Mission in reducing errors to the milliarcsecond level has emphasized the decisive advantages of astrometry from space. Discussions for future missions now center on the scientific and technical aspects for achieving astrometric accuracy in the microarcsecond domain, some 1–3 orders of magnitude smaller than from the Hipparcos Mission. More optimistic scenarios envision measuring proper motions and parallaxes for objects throughout the Galaxy and deep into extragalactic space. Such a success would impact profoundly on problems that today are only imperfectly understood. Beneficiaries would extend from the solar system to stellar kinematics and galactic rotation, distance scale and luminosity calibrations, galaxy trajectories and their implications on the formation and evolution of the Galaxy, and much more. The needed metrology to achieve this goal is technologically challenging. Discussions envision a Space Interferometry Mision (SIM) and a German interferometer for a multichannel photometry and astrometry (DIVA) mission sometime after 2000. Some concepts include GAIA (ESA), POINTS (Center for Astronomy: Cambridge), OSI (Jet Propulsion Laboratory (JPL)), NEWCOMB and FAME (JPL, USNO, SAO, NRL). Direct imaging with HST and future space telescopes in optical and other spectral regions offer alternative options for space astrometry.

Bibliography

Eichhorn H 1974 *Astronomy of Star Positions* (New York: Fredrick Ungar)

Høg E and Seidelmann P K (eds) 1995 *Astronomical and Astrophysical Objectives of Sub-microarcsecond Astrometry* (IAU Symp. 166) (Dordrecht: Kluwer)

Humphreys R M (ed) 1997 *Proper Motions and Galactic Astronomy (ASP Conf. Ser. 127)* (San Francisco, CA: ASP)

MacGillivray H T, Thomson E B, Lasker B M, Reid I N, Malin D F, West R M and Lorenz H (eds) 1994 *Astronomy from Wide-Field Imaging* (IAU Symp. 161) (Dordrecht: Kluwer)

Morrison L V and Gilmore G F (eds) 1994 *Galactic and Solar System Optical Astrometry* (Cambridge Workshop) (Cambridge: Cambridge University Press)

van Altena W F 1983 Astrometry *Ann. Rev. Astron. Astrophys* **21** 131–164

Arnold Klemola

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Protogalaxies

The term protogalaxies or sometimes primeval galaxies (hereafter PGs) has been used in the literature with a range of different meanings, but a general definition may be something like this: progenitors of the present-day (normal) galaxies, in the early stages of formation; also, many, but not all, authors would also add to that the words 'at high redshifts'. The key phrase is 'the early stages of formation', which is fundamentally not well defined.

The ambiguities are mainly due to the fact that our understanding of what GALAXY FORMATION means has been evolving over the past few decades. Moreover, even at any given time, experts of different technical orientations (e.g. observers versus n-body simulators) who would otherwise agree about the general picture of galaxy formation would have different and equally legitimate definitions in mind. For example, an observer may mean 'the first major burst of star formation in a progenitor of a present-day elliptical galaxy', whereas a theorist may mean 'the peak merging epoch of dark halos of the fragments which assemble to produce an average galaxy today'; others would define a PG as a still gaseous body before any STAR FORMATION has taken place or as an overdense region of DARK MATTER in the very early universe, destined to become gravitationally bound and to collapse. In some sense all of them are right.

It may be useful here to offer a brief account of our present understanding of galaxy formation and then describe some of the relevant observations, thus defining PGs in an implicit way. While the subject is still evolving rapidly as of this writing (late 1999), most cosmologists would agree that the basic picture we now have is likely to be essentially correct. Instead of seeking a hypothetical magic epoch when PGs appear and then evolve into the familiar types of galaxies today, the focus is now on understanding of the complete history of galaxy and large-scale structure assembly, as well as the history of star formation in all galaxies and the chemical enrichment of gas in the universe at large. It is then almost a matter of taste which slice of formative history would be called the epoch of galaxy formation and which objects or which evolutionary stages of normal galaxies would be called PGs.

In what follows, we outline a broad-brush picture and give general arguments and order-of-magnitude estimates, which are likely to be at least roughly correct, however much our understanding of galaxy formation evolves in the forthcoming years. Ever more detailed models will be producing more precise predictions, which can then be tested by future observations.

From density fluctuations to protogalaxies

The basic paradigm of structure formation is that it takes place through the gravitational instability of bound, overdense regions in the early universe. The distribution of such regions in mass (or, equivalently, density) is quantified by the initial density perturbation spectrum, which is produced by some type of quantum fluctuation processes in the early universe. Sufficiently overdense regions separate from the universal expansion and collapse under their own self-gravity, roughly on the free-fall time scale:

$$t_{\rm ff} = \left(\frac{\pi^2}{8G}\right)^{1/2} R_{\rm init}^{3/2}\, M^{-1/2}$$

$$\approx 5\times 10^8\ {\rm yr} \times \left(\frac{R_{\rm init}}{100\ {\rm kpc}}\right)^{3/2} \left(\frac{M}{10^{12} M_\odot}\right)^{-1/2}$$

$$\approx\ 1.6\times 10^9 {\rm yr} \times \left(\frac{R_{\rm init}}{10\ {\rm Mpc}}\right)^{3/2} \left(\frac{M}{10^{15} M_\odot}\right)^{-1/2}$$

where $R_{\rm init}$ is the initial radius of the bound region and M is the enclosed mass. It can be seen that for a galactic or subgalactic mass fragment the corresponding cosmic epochs ($\sim 10^8$ yr) imply high REDSHIFTS of collapse, whereas CLUSTERS OF GALAXIES or larger structures may be still collapsing today. The free-fall time scale is a lower limit to the formation time scale, its duration and the cosmic epoch: in practice, it may take several free-fall times for a given mass overdensity to be assembled through a process of hierarchical merging and virialization. To a first order, this simple argument implies that the peak epoch of galaxy formation is likely to be found at the cosmic epochs of a few $\times 10^8$ to a few $\times 10^9$ yr, or roughly in the redshift range $z \sim$ 2–20 or so, whereas the epoch of cluster formation lasts many billions of years and is still going on now.

Collapse

Since the non-baryonic dark matter appears to dominate the total mass, the overdense regions, i.e. proto-dark-halos, can start collapsing even while the universe is still ionized, e.g. at $z \sim 10^4$, with the baryonic matter following. The smallest mass fluctuations collapse the fastest; however, they are also most readily erased by a variety of damping processes operating in the early universe, such as the streaming of matter and photons, sound waves, etc. It is now believed that the smallest structures which survive at the recombination epoch are similar in mass to GLOBULAR CLUSTERS or DWARF GALAXIES, i.e. $(10^5$–$10^8)M_\odot$. They may be the basic building blocks of galaxies.

Images of the COSMIC MICROWAVE BACKGROUND (CMB) photosphere show a snapshot of overdense regions at the recombination epoch at $z \sim 1100$. At the present time, the resolution of such observations corresponds to physical scales of large clusters and superclusters of galaxies rather than galaxies themselves. Nevertheless, these observations support the basic picture of structure formation via gravitational instability.

Any energy dissipation in the baryonic component of the mass (also called the cooling of PGs) would of course accelerate the collapse and lead to the formation of denser objects, since systems which dissipate energy at a fixed mass become more tightly bound. The distinction between galaxies and large-scale structures such as galaxy groups or clusters (or larger) is that the galaxies (or

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

their protogalactic fragments) can cool faster than the free-fall time. Physical mechanisms which enable this energy dissipation in collapsing PGs include the inverse-Compton cooling of hot gas on the CMB radiation (CMBR) photons and shocks in the infalling and colliding gas clouds.

Thus galaxies (or their building blocks) become distinct concentrations within the overall large-scale structure, whose evolution is dominated by the gravity of the dark matter. Simple arguments based on the mean density of galaxies, the spin-up of galaxy disks, etc all suggest that they collapsed by about a factor of 10 or 20 in radius, whereas pure dissipationless collapse can produce only a factor of 2 (this is a direct consequence of the virial theorem for systems bound by gravity). Given the ratio of the comoving rms separation of galaxies today and their typical radii, this must have happened at $z < 20$ or so. It is also hard to have an effective cooling of PGs at redshifts much higher than that.

There are two fundamental aspects of galaxy formation: assembly of the mass (mostly dark), and conversion of the primordial gas (see GAS IN GALAXIES) into stars, and the subsequent chemical enrichment through SUPERNOVA explosions. The former is easier to model numerically; the latter is what is actually observable. The two may be connected, as bursts of star formation are likely to result from mergers of gas-rich PG fragments.

Assembly of protogalaxies

At every spatial (or mass) scale, there is a gradual, hierarchical merging of fragments into ever larger structures. This process operates first at the smaller scales, and moves to ever larger ones as the age of the universe increases and as the smaller fragments are consumed in mergers. While mergers of galaxies are relatively rare (yet readily observed) in the present-day universe, such processes must have been much more common at higher redshifts, when galaxies and their initial building blocks were closer together, with less developed peculiar velocities, and when there were more fragments to merge with (see GALAXIES: INTERACTIONS AND MERGERS). However, this process of hierarchical assembly never really ceases, and today it is the principal mechanism of galaxy transformation.

In 1960s and early 1970s, a simple picture of galaxy formation as a monolithic collapse of a single large PG cloud was promoted by Eggen, Lynden-Bell and Sandage, with a majority of the POPULATION II (bulge and halo) STARS formed more or less during a single free-fall time. This picture has now been completely replaced by the hierarchical formation scenario. The remaining issue is when and where did most of the stars form? A merger of a large number of PG fragments on a time scale comparable with the free-fall time of the entire system, and accompanied by a rapid star formation activity, is in practice indistinguishable from the old monolithic collapse picture. It is quite possible that massive ELLIPTICAL GALAXIES in clusters formed in such a manner. In another extreme scenario, mergers of large fragments occur over a more protracted time scale, comparable with the Hubble time, with most of the stars already formed in the merging units and some formed in merger-induced starbursts. It is likely that there was a full spectrum of galaxy assembly scenarios or formation histories at work.

The interplay of mass assembly and star formation fundamentally determines the galaxy morphology and the origin of disks in SPIRAL GALAXIES. In general, random merging of fragments leads to the formation of spheroidal (or, more accurately, triaxial) systems such as the bulges, stellar halos and dark halos, where kinetic energy needed to support the system against its own gravity is in random motions. In contrast to that, if protogalactic gas settles into a dark halo potential well gradually, it radiates away its kinetic energy, but retains most of the angular momentum (which can be only mechanically transferred away). PGs acquire their angular momenta through tidal interactions in the early universe, as passing-by mass concentrations exert tidal torques on each other. The gas then assumes the lowest-energy configuration possible for a given amount of angular momentum, which is a centrifugally supported, rotating thin disk (similar arguments apply to the origin of any disks in astronomy, including PROTOPLANETARY DISKS, ACCRETION DISKS, etc, not just DISK GALAXIES). Any stars formed within the gaseous disk then inherit the same kinematics.

Disks are dynamically fragile: major mergers would disrupt their orderly rotation and convert them to random motions. Thus, any disk galaxy is unlikely to have had a major merger since its disk was formed, but accretion of smaller satellites is still possible. Conversely, one often hears assertions that (some) ellipticals are made by merging spirals, which is indeed observed in the nearby universe. Obviously this is not the only possible path: a young elliptical may be made by merging a thousand small fragments early on. While some ellipticals may be formed in late, major mergers, many studies of elliptical galaxy properties in clusters at low and moderate redshifts suggest that most of their star formation was completed at high redshifts, $z > 3$ or higher, although some fraction of their stars may have formed since then.

At large redshifts, the merging was much more rapid and frequent, and the formation of durable, massive galactic disks was not likely. As the merging subsided and galaxies were carried apart by the universal expansion, the necessary gradual assembly of disks increased (incidentally, the same argument is used to explain the morphology–density relation in clusters and groups of galaxies: spirals are found mainly in low-density regions, and vice versa for ellipticals). Thus, we expect that the epoch of halo and bulge formation in galaxies in general preceded the epoch of disk formation, which is consistent with their star formation histories: bulges and ellipticals show little or no star formation today, and have spectra and colors characteristic of old stellar populations, where the bulk of the stars formed ∼6–12 billion years ago, whereas the disks are actively forming stars now,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and apparently had similar star formation rates over their entire history.

To summarize, the expected history of galaxy formation may be something like this. The first PG fragments (gravitationally bound clumps in the early universe) may have had masses comparable with those of dwarf galaxies, $M\sim(10^6$–$10^8)M_\odot$ at the epoch of recombination, at $z \sim 1100$. Over the subsequent few billion years they merged hierarchically into ever more massive structures, stimulating star formation in the process. The first stars may have formed within a few hundred million years, at redshifts $z \sim 20 \pm 10$. The process of hierarchical assembly of PGs reached a peak maybe 1–3 billion years later, at redshifts $z \sim 2.5 \pm 1$. At subsequent epochs, this activity subsided as galaxies were carried apart by the universal expansion, and the supply of fragments and gas to merge with was diminished. Galaxy formation became GALAXY EVOLUTION. Most of the early star formation was in proto-ellipticals and bulges, with the disks forming on longer time scales.

Assembly of the mass into the present-day galaxies is one fundamental aspect of galaxy formation, but since most of the mass is dark this process is not readily observable. The other fundamental aspect is the conversion of primordial gas into stars and the early evolution of stellar populations, which is what we can observe. Let us then examine the sources of luminosity in PGs.

Luminosity of protogalaxies

The first is the release of excess binding energy, since PGs have collapsed. This is true regardless of the details of mass assembly, be it hierarchical merging or a monolithic collapse. A variety of arguments suggest that the PGs collapsed by about a factor of 10. First, the mean densities of galaxies exceed the extrapolated densities of the large-scale structure down to the typical radial scales of galaxies (a few kpc) by about a factor of $\sim 10^3$, suggesting a collapse by about a factor of 10. Second, the spinup of galactic disks requires a collapse by a similar radial factor. We can estimate the amount of the binding energy released by assuming that the present-day galaxies are in a virial equilibrium and estimating their total intrinsic kinetic energy. According to the virial theorem, this is equal to the amount of the excess binding energy lost in the process of virialization: a PG starts as a marginally bound system, with the kinetic energy equal to the potential one, but ends up as a virialized system where the kinetic energy is one half of that, with the other half presumably radiated away or carried out by galactic winds. An estimate of the kinetic energy of a normal, average galaxy today can be written as

$$|E_{\rm k}| \simeq \frac{1}{2} M_{\rm cool} \langle V_{\rm 3d}^2 \rangle$$

$$\simeq 6.2 \times 10^{58}\ {\rm erg} \times \frac{M_{\rm cool}}{10^{11} M_\odot} \left(\frac{V_{\rm 3d}}{250 {\rm km\ s^{-1}}} \right)^2$$

where $M_{\rm cool}$ is the total mass which can cool radiatively and $V_{\rm 3d}$ is the typical rms velocity. This estimate represents a minimum amount of energy released by a PG, even with no star formation. A comparable amount of binding energy is released in the collapse of protostars within a PG.

There are several possible mechanisms for release of this energy. The gas may be cooling via the inverse Compton effect on the photons of the CMBR. It may be radiated away in the shocks from colliding, still gaseous PG clouds and fragments. If the collapse happens on a free-fall time scale, $t_{\rm ff} \sim 10^8$ yr, the corresponding mean luminosity is $L_{\rm coll} \sim 2 \times 10^{43}$ erg s^{-1} $\sim 5 \times 10^9\ L_\odot$.

However, a much more energetically important process is nuclear burning in stars. It has been estimated that about a half of the STELLAR MASS in galaxies today is in the old populations, bulges and stellar halos, and thus about a half of all star formation in the universe's history can be associated with PGs (this is admittedly dependent on how one wishes to define PGs). The key to estimating this energy is through the resulting chemical enrichment: the same stars which produce the energy also produce the heavier elements. Most of the energy comes from hydrogen burning into helium, ~7 MeV per nucleon, i.e. ~0.7% or the initial rest mass of the processed hydrogen. The rest of the nucleosynthetic chain up to iron adds <0.1% to that.

Unfortunately, it is practically impossible to measure directly the accurate abundance of helium in the old STELLAR POPULATIONS today. Instead, we can use the observed abundances by mass of heavier elements and a conversion from chemical evolution models: $\Delta Y/\Delta Z = f$, where ΔY is the increase in the fractional mass abundance of helium, ΔZ is the increase in the fractional abundance of heavier elements (i.e. the mean metallicity), and f is typically found to be ~2–5. Since the initial abundance of heavier elements is essentially zero, $\Delta Z = Z$, which is observable. Estimates of the mean metallicity of old stellar populations are of the order of solar, i.e. $\langle Z \rangle \approx 0.01$–$0.02$. Thus the fraction of the rest mass of the primordial gas converted into energy during this initial chemical enrichment is

$$\Delta X \approx\ 0.07 f \langle Z \rangle \approx (4 \pm 2) \times\ 10^{-4}.$$

We can then estimate the total energy release from nuclear burning in stars responsible for this initial chemical enrichment:

$$E_{\rm nuc} \simeq M_\star\, c^2 \Delta X$$

$$\simeq 7.2 \times 10^{61} {\rm erg} \times \frac{M_\star}{10^{11} M_\odot} \frac{\Delta X}{4 \times\ 10^{-4}}$$

where $M_\star$ is the total mass burned in stars during the protogalactic starburst. This is much larger than the release of the binding energy and would thus be the dominant signal in detection of PGs. It may be a lower limit, since some of the produced metals must have been ejected in the resulting galactic winds and are not counted in our estimate of the mean metallicity of stellar populations today. Direct evidence for these ejected metals is in the intracluster gas seen in x-rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

If this energy is released over a free-fall time scale, $t_{\rm ff} \sim 10^8$ yr, the corresponding bolometric luminosity would be $L_{\rm nuc} \sim 2 \times 10^{46}$ erg s^{-1} $\sim 5 \times 10^{12}\, L_\odot$, comparable with the luminosities of bright QUASARS and utraluminous IRAS STARBURST GALAXIES today. Some authors have indeed proposed that such PG starbursts are the source of luminosity of quasars, but a majority belief is that quasars are powered by accretion onto massive BLACK HOLES. The mean star formation rate would be several hundred $M_\odot$ yr^{-1}. Even if this process proceeded on a more extended, hierarchical merging time scale $t_{hm} \sim 10^9$ yr, the resulting luminosities would still be an order of magnitude higher than the mean galaxy luminosity today (L_*) and comparable with that of the brightest cluster ellipticals today.

Finally, if formation of AGN accompanied some or all of these PG starbursts in young ellipticals and bulges, a comparable amount of energy may have been contributed by these early AGN. There is now growing evidence that massive black holes exist in the centers of most ellipticals and bulges, with the mean mass proportional to that of the stars, $M_{\rm bh} \sim 0.002 M_\star$. If about 10% of the rest mass of the black hole is radiated away in dissipative processes during its formation and growth (this is a typical estimate), then a comparable fraction of the bulge rest mass is radiated in the production of these ubiquitous black holes as in the nuclear burning in stars. It would be hard to separate observationally the effects of primordial starbursts in PGs housing quasars at large redshifts in terms of the energetics alone.

Thus, a 'generic' PG would have a typical bolometric luminosity $L_{\rm PG} \sim 10^{11\pm1} L_\odot$. To make an order-of-magnitude estimate of the observable fluxes, we note that in the relevant range of redshifts, $z \sim$ 3–10, and for a reasonable range of cosmological models, the luminosity distances are $D_{\rm L} \sim$ a few $\times 10^{29}$ cm, so that the expected bolometric fluxes are of the order of $\sim 10^{-15\pm1}$ erg cm^{-2} s^{-1}. Assuming for simplicity $L \sim \nu P_\nu$, the corresponding observed R band magnitudes would be $R \sim 25$ mag, or $F_\nu \sim 0.1$ mJy at $\lambda_{\rm obs} \approx 0.3$ mm, give or take an order of magnitude in flux. Much of this flux range is detectable with present (or soon forthcoming) instruments.

Detection of protogalaxies

The key question about detectability of PGs is that of obscuration. Most of the energy is released in the restframe UV, either from the photospheres of massive stars or from accretion disks in AGN, and would be now redshifted in the visible regime. On the other hand, if these PG starbursts were obscured by dust, the captured energy would be reradiated in the far-IR and would now be observable in the sub-mm regime (the AGN may also be detectable in x-rays). Approximately a third of all star formation in the present-day universe is obscured by dust, and perhaps a similar fraction applied to PGs (note, however, that some tens of millions of years of stellar evolution and chemical enrichment are needed before any dust can be produced, and the early PG phases must have been unobscured). This dichotomy of energy release channels leads to a dichotomy of detection strategies. We will discuss the two cases in turn.

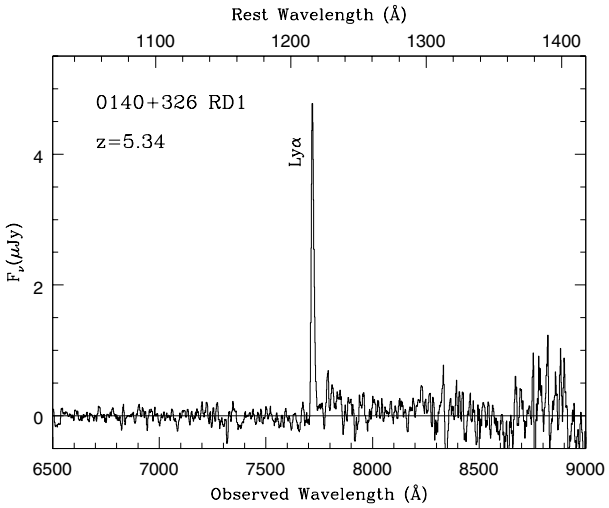


Figure 1. An example of a spectrum of a galaxy at $z = 5.34$, discovered serendipitously in a long-slit spectroscopic search. The Lyα emission line is very prominent. Objects like this can also be found in a narrow-band imaging search, provided that the line falls within the narrow filter bandpass. Reprinted with permission from Dey A *et al* 1998 *Astrophys. J. Lett.* **498** L93. © The American Astronomical Society.

Most of the early, and some of the modern searches for PGs were based on the detection of recombination emission lines, and most notably the Lyα λ 1215.7 Å line of the ionized hydrogen. For an unobscured, young stellar population with a normal IMF, several per cent of the total bolometric luminosity should emerge in this line. A useful rule of thumb is that about two Lyα photons are produced for every three ionizing (i.e. $\lambda < 912$ Å) photons. The line is readily detected (and is very luminous) in quasars, but its detection from galaxies powered by star formation was only achieved in 1990s. An approximate conversion is that unobscured star formation rate of $1 M_\odot$ yr^{-1} would produce the Lyα line luminosity of about 10^{41} erg s^{-1} in the restframe, but this number is highly sensitive to the form of the stellar initial mass function, since most of the ionizing photons are produced at the high-mass end but most of the mass is in the low-mass stars.

There are two observational approaches to detection of high-redshift, emission line objects, including PGs: narrow-band imaging and slit spectroscopy. Most of the emission line PG searches in the visible light have targeted the Lyα line, but exactly the same methodology was also used in the near-infrared, where searches targeted the Balmer lines of ionized hydrogen (mainly Hα λ 6563 and Hβ λ 4861) and the nebular lines of ionized oxygen, [O II] λ 3727 or [O III] λ 5007.

In the former, deep images are obtained in a narrow-band filter (usually an interference filter, or a Fabry–Perot tunable filter) and compared with images in a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

broad bandpass with a comparable central wavelength. Objects with a strong emission line falling in the narrow bandpass would appear much brighter than in the broad band, where this line flux would be averaged out. This method can cover a relatively large area, but only a very narrow slice of redshift, corresponding to the redshifted line traversing the width of the filter, i.e. $\Delta z = \Delta\lambda/\lambda_c$, where $\Delta\lambda$ is the filter width and λ_c its central wavelength. Typically, $\Delta z \approx 0.01$–0.02 is used in PG searches.

The second approach is to obtain deep long-slit spectra of 'blank' patches of the sky, hoping that an emission line object would be covered by the slit. This method covers only a small area but a large range in redshift. One benefit of this approach is that almost any deep spectra taken for some other purpose can be used for a serendipitous search for PGs (figure 1).

The first detections of Lyα emission from GALAXIES AT HIGH REDSHIFT were from powerful RADIO GALAXIES, which are ostensibly ionized mainly by an active nucleus, and from quasar companion galaxies. However, no clear population of 'ordinary' PGs, i.e. powered by star formation, was found despite considerable efforts until the late 1990s. The early limits and the more recent detections correspond to modest star formation rates, up to a few $M_\odot$ yr^{-1}.

This conspicuous absence of a large population of Lyα luminous PGs was readily explainable by invoking extinction: both Lyα photons and the ionizing UV photons which produce them are absorbed by interstellar dust, which is likely to be present at some level in all but the most primordial star-forming regions. To complicate things further, Lyα photons are resonantly scattered by the interstellar hydrogen, which greatly increases their probability of being absorbed before escaping the PG.

The real breakthrough in the detection of high-redshift, star-forming galaxies came with the use of a different technique, the Lyman-break method (figure 2). Actively star-forming PGs would have relatively featureless, flat continua, except for one very significant feature: a strong drop in flux shortward of the Lyman limit, at 912 Å in the restframe, as the more energetic photons are readily absorbed by interstellar hydrogen. Thus, by imaging in at least three filters, one of which would cover the part of the spectrum shortward of the redshifted Lyman break, and two of which would sample the relatively flat continuum longward of the break, one can select candidate high-redshift galaxies through the resulting color signature. Actual spectra are then obtain to verify the candidates and measure their redshifts. This method is very efficient, and to date many hundreds of high-redshift galaxies (i.e. at $z > 2$ or so) have been identified in this manner.

At somewhat higher redshifts, i.e. $z > 4$ or so, a variant of this technique is used to detect both galaxies and quasars. There the net absorption by the intergalactic Lyα forest clouds along the line of sight decreases the flux shortward of the Lyα line in the object itself and sufficiently so as to produce a similar color signature. For example, quasars and PGs at $z \sim 4$–4.5 or so are very red in the $B-R$ or a similar color owing to this absorption effect, yet blue in the $R-I$ color, which is intrinsic and is not affected by the Lyα forest. This is actually the principal discovery technique for $z > 4$ quasars, but it works equally well for galaxies at comparable redshifts. The effectiveness of this technique increases with redshift, as the optical depth of the Lyα forest increases.

What if the star-forming regions are effectively completely hidden by dust? The absorbed energy is then reradiated in the restframe far-infrared and would now be observable in the sub-mm and mm region of the spectrum. Adequate detection technology in these wavelength regimes is just maturing, and first detections of high-redshift sub-mm sources (intrinsically very luminous ones) are just starting as of this writing (late 1999). This is clearly a very promising area of research, and it is likely that future observations in this regime will uncover a more substantial, optically obscured component of PG population.

We note that the integrated emission of unobscured and obscured galaxies, both forming and evolving, is observable in the form of diffuse optical and far-infrared backgrounds. Discussion of this fascinating topic is beyond the scope of the present article, but suffice it to say that the data so far indicate that the two components (obscured and unobscured) are roughly comparable in terms of the energy content.

A complementary approach to a direct detection of PGs in emission is to detect them in absorption, in the spectra of unrelated background quasars. Metallic line absorbers and high column density hydrogen absorbers are now believed to originate in galaxies or their halos, which at high redshifts may still be in the process of assembly. This approach eliminates all selection biases inherent in direct photometric detection, but it is limited in the information it can provide, primarily the gas column density, metallicity (an indicator of the chemical evolution) and perhaps rudimentary kinematics. Direct detections of a small number of high-redshift galaxies responsible for absorption systems have been made, and the data suggest that they are very similar to the general field population detected with the Lyman-break technique at comparable redshifts.

Formation of galaxies and that of the large-scale structure are closely related at early epochs. It is plausible that the first galaxies form in the densest spots of the primordial density field. These 'density peaks' are expected to be intrinsically highly clustered, as they would be mainly located within the larger overdensity regions. The first PGs may thus be found in the cores of future rich clusters, and then galaxy formation would gradually spread to ever lower density environments. This concept of so-called biased galaxy formation is generic to almost all models of structure formation. There is now some evidence that galaxies and quasars at $z > 3$ or 4 may indeed be strongly clustered. It should be noted that this primordial large-scale structure derives from initial conditions, with the first PGs and quasars lighting up

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

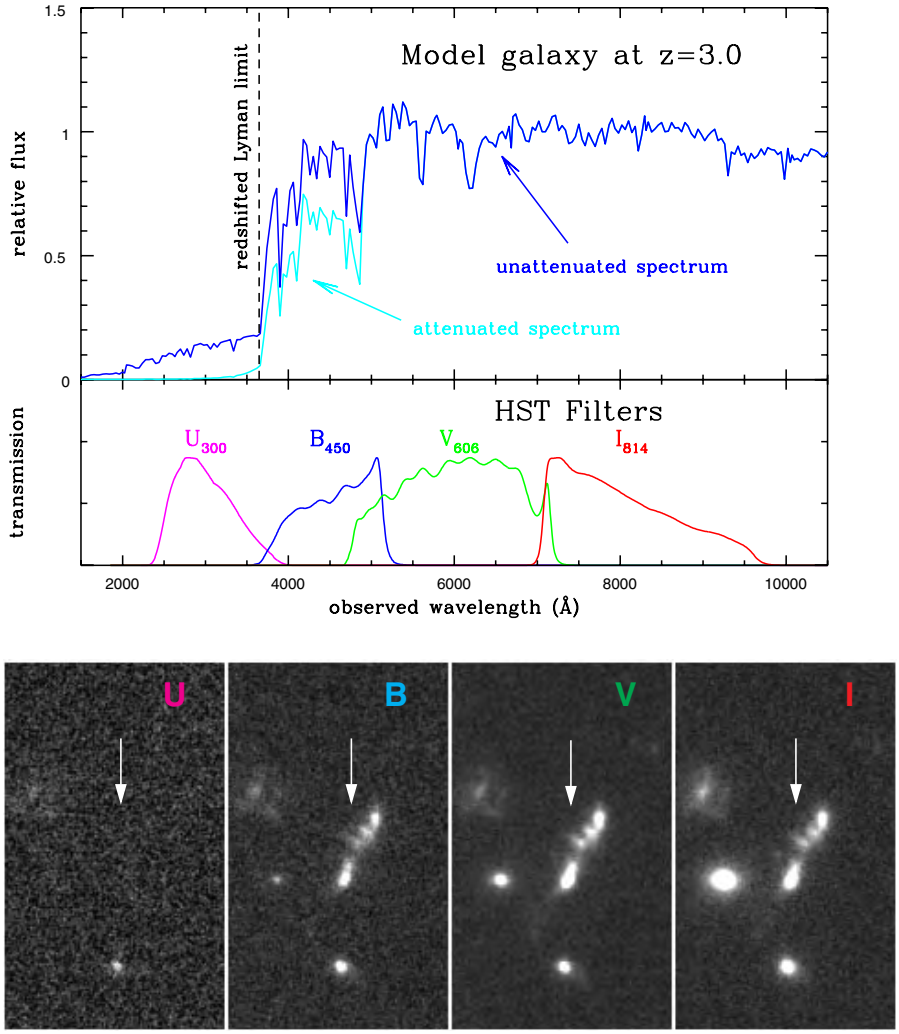


Figure 2. An illustration of the Lyman-break technique at work. The top panel shows a population synthesis model spectrum of an actively star-forming galaxy redshifted to $z = 3$, with and without attenuation due to the Lyα forest. Note the sharp drop in the continuum shortward of the Lyman break, at the observed wavelength of 912 Å $\times(1 + z)$. A set of transmission curves of filter used in the imaging of the Hubble Deep Field is shown below the spectrum. One would expect that such a galaxy would appear about equally bright in the B, V and R filters, but be practically undetectable in the U filter. This is indeed observed in the actual images of a galaxy at $z = 2.8$ found in the Hubble Deep Field images, shown in the bottom row. Reprinted with permission from Dickinson M 1998 *The Hubble Deep Field* ed M Livio *et al* (Cambridge: Cambridge University Press) p 219. © (The Cambridge University. Press).

the highest peaks of the primordial density field. This is distinct from the large-scale structure seen in the local universe, which grows gradually over the Hubble time through the gravitational instability.

The net result of the studies to date is at least a broad-brush understanding of the history of galaxy formation and evolution since $z \sim 5$ or so. The average star-forming activity has a broad maximum around $z \sim 2$ or 3, with a rapid decline since $z \sim 1$ or 2 towards the present day, $z \sim 0$. However, it is not clear whether we have yet seen the onset of galaxy formation. From absorption line studies we know that the universe was already fully ionized at $z \sim 5$, presumably by some combination of PGs and early AGN. While most of the stars may have formed since $z \sim 5$, clearly the first stars and thus PGs must have formed earlier.

We know that the universe entered 'dark ages' after the recombination at $z \sim 1100$, a few hundred thousand years after the big bang. Current theory suggests that the first protogalactic fragments started forming at $z \sim 10$–20 or so, a few hundred million years after the big bang. This 'cosmic renaissance' ended the dark ages and marked the onset of galaxy formation. The challenge for the future is to map this critical era, from $z \sim 20$ to $z \sim 5$, where the first PGs were forming.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

This is a rapidly developing field, and perhaps the best way to learn about the up-to-date results is by browsing recent conference proceedings on cosmology, high-redshift universe, etc, of which there are many. However, the following reviews give more details on the subject of this article.

A good review of PG searches, problems, and techniques up to the mid-1990s is

Pritchet C 1994 *Publ. Astron. Soc. Pac.* **106** 1052–67

An updated review of searches for distant galaxies up to 1999 is

Stern D and Spinrad H 1999 *Publ. Astron. Soc. Pac.* **111** 1475–502

A good overview of the Lyman-break technique and some of the results is

Steidel C 2000 *Proc. Pritzker Symp. on the Status of Inflationary Cosmology* (Chicago, IL: University of Chicago Press)

S G Djorgovski

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proton

Series of four large Soviet magnetospheric research satellites launched 1965–68. Carried experiments to investigate ultra-high-energy cosmic particles such as cosmic rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proton

A positively charged elementary particle that is composed of three quarks. It has a charge of +1, equal in magnitude to the charge on the electron (1.602×10^{-19} coulomb), and a mass of 1.6725×10^{-27} kg, 1836 times greater than the mass of the electron. In terms of energy, its mass is equivalent to 938.3 MeV. It has an intrinsic angular momentum, or spin, of 1/2 (expressed in units of the Planck constant, h, divided by 2π).

The proton is a nucleon, one of the two basic constituents of an atomic nucleus (the other being the marginally heavier neutron). The number of protons in an atomic nucleus (its atomic number) determines its net charge and, therefore, the chemical element of which that atom is an example. For example, a hydrogen nucleus contains one proton, a helium nucleus, two, and a uranium nucleus, 92. Although it has a lifetime of at least 1032 years, and may be completely stable, some versions of grand unified theory predict that protons eventually decay into leptons and photons.

See also: atom, baryons, electron, electronvolt, elementary particles, hadrons, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Protoplanet

An object representing an intermediate stage in the process of planet formation. A protoplanet is formed by the accretional collision of planetesimals, the primary products of accretion, themselves formed initially by the coalescence of dust grains and then the gradual aggregation of further material. Protoplanets collide with one another to form larger objects; the end result of this stage is a system of planets—a small number of still larger objects which have swept up most of the debris from collisions and occupy orbits which have long-term stability.

See: accretion, planetesimal, solar system: formation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Protoplanetary Disks

Protoplanetary disks are disks of dust and/or gas orbiting very young stars. Based upon what we know about the formation of our own solar system, these objects are deemed to be very likely to evolve into solar systems like our own.

This is a very fast-paced field of study and many of the scientific advances have occurred within the past 15 yr. In 1983, a satellite was launched to study the infrared emissions from the sky. This satellite, called the Infrared Astronomy Satellite or *IRAS*, revealed that about one-fifth of the main-sequence A stars near the Sun exhibited an excess infrared emission that could only be explained by warm dust orbiting these stars. This suggested that a significant fraction of all A stars, and presumably other young stars as well, have dust orbiting them. Since A stars have lifetimes of 10^8–10^9 yr, their implied ages overlap with the era of planet formation and heavy meteoritic and cometary bombardment of planets in our early solar system.

Shortly after this discovery, attempts were made to image the dust surrounding these stars. This was very tricky, since one must see the dust against the glare of the bright star. A small disk was placed inside the camera to block out the light coming from the star. In this way, it was hoped that the light scattered by the dust surrounding the star could be seen. One star, β Pictoris, an A star at a distance of 62.5 light-years, revealed a very flat disk of dust extending over 500 astronomical units (AU) from the star. (An astronomical unit is the distance defined by the distance between the Earth and the Sun.) For comparison, Pluto has an average distance of 39 AU from the Sun. This was the first evidence for truly solar system sized disks of dust orbiting nearby stars. The size of this dust disk is comparable to that of the KUIPER BELT of our own solar system. The Kuiper Belt is thought to extend from just inside the orbit of Neptune to beyond 200 AU from the Sun. The Kuiper Belt is the home of an estimated 100 million to 1 billion COMETS. A small fraction of these comets have orbits that penetrate into the inner solar system near the Earth and the Sun.

Spectroscopic studies of β Pictoris using the International Ultraviolet Explorer (IUE) satellite and the Hubble Space Telescope (HST), as well as ground-based telescopes, revealed an even more complicated picture. These spectra showed red-shifted spectral absorption lines of low-ionization atoms, varying with time, indicating gas possibly from comets falling onto the the stellar surface. The age estimates for β Pictoris are typically less than 100 million years. β Pictoris and other stars similar to it show spectral signatures of ice. These data, collectively, suggest that large numbers of comets are present around many young A stars. Hopefully, by studying these objects we can gain some understanding of how our early solar system evolved.

Thanks to a variety of observations, using both ground-based and space-borne instruments, we are beginning to get a consistent picture on how young protostars and disks representing PROTOPLANETARY NEBULAE evolve. The process starts with the gas on an interstellar cloud collapsing to form stars. The gas and dust forming an individual star fall inward and produce a star and an envelope which contains a large fraction of the angular momentum. An outward flow of gas, or 'wind', develops that clears the gas and dust of the inner envelope away from the young stellar object along the rotational axis of the star. The polar winds, or jets, can extend to very large distances. These extended bipolar flows can have distinct 'knots' of emitting gas along the flow. These flows can interact with the dense nebulosity typically found around these objects and produce shock-heated boundaries. These objects, if enough obscuring dust is removed so they can be seen at optical wavelengths, are sometimes called HERBIG–HARO OBJECTS.

The youngest obscured systems

Recent HST images have been acquired of YOUNG STELLAR OBJECTS (YSOs) in the constellation Taurus with the Near Infrared Camera and Multi-Object Spectrometer (NICMOS). These young systems are typically only observable at infrared (IR) wavelengths because the thick veil of dust absorbs all the visible light. However, the dust re-radiates this light energy at infrared wavelengths. The NICMOS infrared images at wavelengths from about 1 to 2 μm indicate YSOs with masses similar to our Sun and ages $\leq$1 Myr. These objects show dark obscuring dust lanes with diameters of 500–900 AU perpendicular to the bipolar flows emanating from the YSOs. The HST observations indicate these dust lanes are somewhat-flattened disks with equatorial extents of 500–900 AU and vertical thicknesses of 50–350 AU. The central protostar and the surrounding disk are completely obscured at optical wavelengths.

Other presumably young objects are the 'proplyds'. These objects are very compact knots of ionized gas that are seen near the core of the ORION NEBULA. The Orion Nebula is also a well-known site of active star formation. The proplyds appear to be very small with radii on the order of 50 AU. All observations suggest that they are a distinct class of very young objects. Proplyds appear to be flattened circumstellar clouds of gas and dust that surround stars collapsing toward the main sequence. Just how these objects fit into the observational scheme of the evolution of protoplanetary disks is not at all clear.

Ages from 1 to 10 million years: the young stars are unveiled

At ages from 1 Myr to 10 Myr, the envelope disperses, the disk continues to flatten and the central pre-main-sequence star becomes visible. Although much of the stellar radiation is absorbed by dust and re-emitted at IR wavelengths, the star itself shows very strong emission lines of ionized gas. The emission is produced by the gaseous infall onto the stellar surface and by the strong magnetic activity in these young objects. These young

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

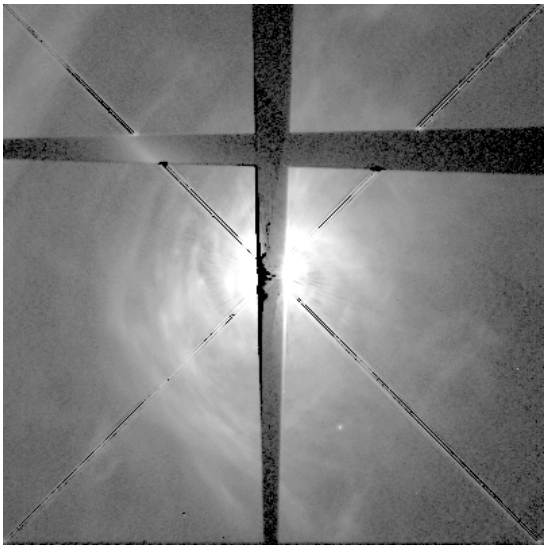


Figure 1. HST/STIS image of the circumstellar disk of the pre-main-sequence star, AB Aurigae (after Grady *et al* 1999). This image obtained with the STIS shows stellar light reflected from dust of the circumstellar disk. Notice the crossbar structure that blocks the direct light of the central star. The circumstellar disk appears to have a clumpy band or spiral structure centered on the star.

emission-line stars are called T TAURI STARS and are named after their prototypical example.

HST images of protoplanetary disks have been acquired with both the NICMOS camera in the near-IR and the Space Telescope Imaging Spectrograph (STIS) at visual wavelengths using techniques to dramatically reduce the contributions of the central stars. These techniques enhance the visibility of the surrounding circumstellar dust disk. These images have provided important glimpses of what happens at ages from 4 to about 10 Myr. These observations are very important because they can be used to limit the time scale of the protoplanetary disk phase. At later times, if a massive, Jupiter-sized planet is present, its gravitational perturbations can act to clear away dust from the region interior to the planet's orbit.

Recent STIS images of the star AB Aurigae, at a distance of about 470 light-years, show a dust disk with a radius of about 1300 AU (figure 1). Age estimates are generally not highly accurate, but determinations for AB Aurigae yield an age of about 2–4 Myr. This star also displays a gas disk that has been seen in line emission from molecular CO. The images do indicate a nebulosity that is circularly symmetric about the star with some spiral or band-like structure superposed. In general, there is a uniform drop-off in surface density with distance from the star. No definite evidence of clearing due to a massive planet is seen.

Near-IR images obtained by the NICMOS camera aboard the HST show annular dust rings instead of uniform dust disks about two young stars, HR 4796A and HD 141569. Both stars are hotter and much more massive than our Sun. The spectral types are A0 Ve and B9.5 Ve for HR 4796A and HD 141569, respectively. The 'e' in both spectral types indicates that these stars have detectable hydrogen Balmer lines in emission.

The age for HD 141569 is ill defined. All that can be said is that it has an age between 1 Myr and 10 Myr. The NICMOS images reveal a disk with a gap at a distance of approximately 250 AU from the star. In addition the brightness of the dust disk peaks at about 185 AU from the star and falls off in brightness at both smaller and larger distances. If recent modeling efforts are correct, the gravitational perturbations of a Jupiter-sized planet at a distance of 250 AU could produce the observed gap. This planet would be far too faint to be detected with the available ground-based and space borne instruments.

The NICMOS imagery of the star HR 4796A is even more revealing (figure 2). The near-IR images at 1.1 and 1.6 μm of this star clearly show a sharply defined annular ring. The inclination on the plane-of-the-sky implies that the ring brightness peaks at a distance of about 70 AU from the star. The resolved width of the ring further shows that it is sharply delineated with a width of 17 AU. The sharply defined morphology of the annular disk implies one, possibly two, unseen planets near 70 AU from the star, HR 4796A. The age estimate for HR 4796A suggests that it is 8$\pm$ 3 Myr old.

The HST observations for HD 141569 and HR 4796A impose strong constraints on the theory of planet formation. Based upon theories attempting to explain how planets form in our own solar system, it would be extremely surprising to see massive, Jupiter-sized planets at such great distances (70–250 AU) from the central star, a region where the density in the protoplanetary disk should be far lower than at distances corresponding to the realm of the giant planets in the solar system. Also, these observations suggest the formation time needed to form giant planets is quite short ($\leq$8–10 Myr) and poses strong constraints on theories for planetary formation.

The end of the protoplanetary phase

At times greater than 10 Myr, if planets are, indeed, forming at times shorter than 10 Myr, we may not be seeing protoplanetary disks, but debris disks, representing the residues of planetary formation. Although meteoritic and cometary impacts will continue the planetary growth process, the protoplanetary part of the evolution of the circumstellar disk has come to an end. There are several physical processes that act to destroy and disperse circumstellar dust. Observations of stars older than 10 Myr can place important constraints on these processes.

The most notable stars which appear to have circumstellar dust disks and that have ages $\geq$10 Myr are the so-called 'big three', β Pictoris, α Piscis Austrini (Fomalhaut) and α Lyrae (Vega). Based upon their deduced positions in the H–R diagram and theoretical evolutionary tracks, the ages of these three stars are thought to be near 100, 200 and 400 Myr, respectively. However, there is some indirect evidence which suggests

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

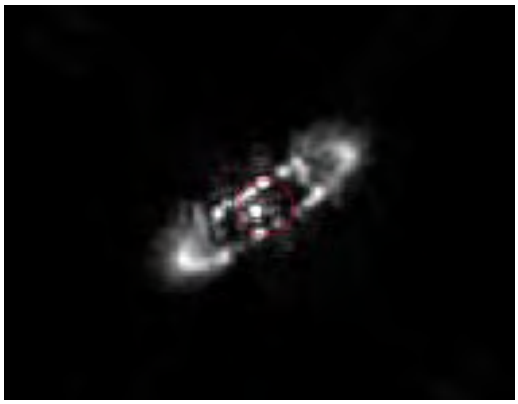


Figure 2. HST/NICMOS image of the circumstellar dust ring of HR 4796A (after Schneider *et al* 1999). This is the processed image of the disk as seen in the infrared at 1.1 μm. The central star has been placed in a 0.3″ radius hole in an attempt to minimize the light seen from the star. Note the sharply defined dust ring. Its infrared brightness peaks at ∼70 AU from the star. The sharply defined nature of the ring may indicate the presence of one or more unseen Jupiter-sized planets.

that β Pic might be as young as 20 Myr. All of these stars have significant IRAS infrared excesses as has been known since the mid-1980s.

The processes that lead to the destruction of the dust disk are (a) expulsion by pressure of starlight, (b) Poynting–Robertson drag and (c) sublimation of icy grains. Small grains have a high cross-sectional area relative to their mass and can be preferentially blown away when the force due to radiation exceeds that of gravity. The smaller grains, with sizes $\leq$2–10 μm, are quickly expelled even at early ages. The larger remaining grains, because they orbit the star, have a component of the radiation force, Poynting–Robertson drag, that opposes the orbital velocity of the grains. This causes the orbiting grains to eventually spiral into the star. This preferentially clears the inner disk, but the outer regions at distances of 1000 AU can be cleared on time scales comparable with the evolutionary time of the A stars.

However, it is still unresolved how such extensive dust disks remain in these older systems. Perhaps this dust and the presence of massive planets inferred at larger distances suggest that the protoplanetary disks are quite massive and dense at very large distances from the protostar. Recent models even suggest that planets can form at much greater distances and then, through interactions with the disk, migrate inward toward the star. This is a controversy that hopefully will be solved with future observations of other young forming planetary systems and extrasolar planets (EXOPLANETS).

Bibliography

Comprehensive reviews of the research on protoplanetary disks are regularly presented in scientific conferences. The best source of these compilations can be found in the series published by The University of Arizona Press called *Protostars and Planets I–IV*. These published books detail the proceedings of regular meetings held every few years in this field.

Much of what has been presented here comes from very recent observational results obtained with instruments aboard the Hubble Space Telescope. More exciting results are sure to come. Some of the recent observations and theoretical models can be found in the following references.

Bryden G, Chen X, Lin D, Nelson R and Papaloizou J 1999 Tidally induced gap formation in protostellar disks: gap clearing and suppression of protoplanetary growth *Astrophys. J.* **514** 344

Grady C, Devine D, Woodgate B, Bruhweiler F, Boggess A, Plaitt P, Clampin M and Kalas P 1999 STIS coronographic imaging of the Herbig Ae star: AB Aurigae *Astrophys. J. Lett.* **523** L151

O'Dell C R 1998 Observational properties of the Orion proplyds *Astron. J.* **115** 263

Padgett D, Brander W, Stapelfeldt, Strom S, Terebey S and Koerner D 1999 HST/NICMOS imaging of disks and envelopes around very young stars *Astron. J.* **117** 1490

Schneider G *et al* 1999 NICMOS imaging of the HR 4796A circumstellar disk *Astrophys. J. Lett.* **513** L127

Fred Bruhweiler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Protoplanetary Nebulae

Protoplanetary nebulae (or pre-planetary nebulae, PPNs) are defined as objects that are in transition between the asymptotic giant branch (AGB) and planetary nebula phases of STELLAR EVOLUTION. Stars on the AGB lose mass at a high rate ($(10^{-7}$–$10^{-4})M_{\odot}$ yr^{-1}) in the form of a stellar wind. Such mass loss eventually depletes the hydrogen envelope of the star and exposes the electron-degenerate carbon–oxygen core. This leads to an increase of the effective temperature of the star and the termination of mass loss. As the small amount of remaining hydrogen above the core continues to be consumed by hydrogen shell burning, the effective temperature of the star will gradually increase. When the stellar temperature reaches approximately 30 000 K, the amount of stellar photons with energies above the Lyman limit (13.6 eV) will be large enough to photoionize the surrounding material. The emergence of strong recombination lines of hydrogen (e.g. Hα) and helium and collisionally excited lines of metals (e.g. the [O III] line at 5007 Å) signifies the beginning of the planetary nebula phase. The evolutionary phase between the end of AGB mass loss and the beginning of photoionization is referred to as the PPN phase.

Observational properties

PPNs are primarily identified by their large infrared excesses. Such excesses originate from dust in the remnant of the stellar wind ejected during the AGB. As the wind disperses, the dust cools and the thermal radiation shifts to longer wavelengths. Most of the PPNs are discovered by finding IRAS sources that have dust temperatures intermediate between those of AGB stars and planetary nebulae. A typical PPN shows an infrared excess peaking at $\sim$20 μm, corresponding to a dust temperature of $\sim$150 K. Some examples of PPNs are given in table 1.

Table 1. Examples of protoplanetary nebulae.

Object	Spectral type	Chemistry	[s/Fe]
IRAS 04296 + 3429	G0 Ia	C	+1.4
IRAS 07134 + 1005	F5 I	C	+2.0
IRAS 18095 + 2704	F3 Ib	O	−0.1
IRAS 19500 − 1709	A2–3 I	C	+1.2
IRAS 22272 + 5435	G5 Ia	C	+2.5

[s/Fe] denotes the abundance of s-process elements.

Molecular emission is also a characteristics of PPNs. Molecules (e.g. CO, HCN) in the remnant AGB envelope emit line radiation as the result of rotational transitions. These molecular lines have typical linewidths of 20–40 km s^{-1}, implying expansion velocities of 10–20 km s^{-1} for the envelope. From these velocities, we can estimate the dynamical ages of PPNs to be $\sim 10^2$–10^3 yr. The total lifetime of the PPN phase is $\sim 10^3$ yr, depending on the mass of the star.

The central stars of most PPNs are of spectral types between G and B (corresponding to effective temperatures of 5000–15 000 K) and luminosity class I. The high luminosity class of PPNs is a manifestation of the low surface gravity ($\log g \sim 1$) in these objects. Because of their spectral class, PPNs are sometimes referred to as yellow supergiants.

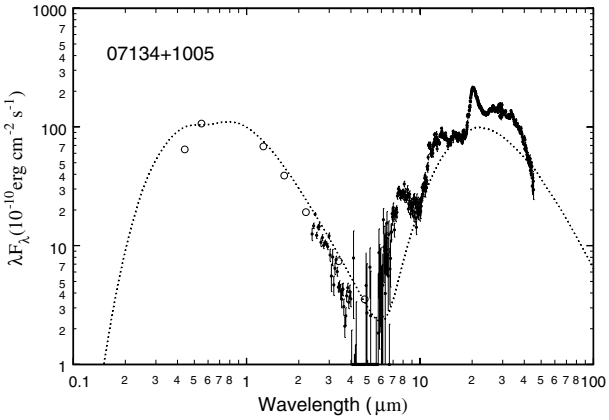


Figure 1. The spectral energy distribution of the PPN IRAS 07134 + 1005. The dust component (right) is clearly separated from the photospheric component (left). The fact that the dust shell is detached from the photosphere is confirmed by a radiation transfer model (shown as a dotted curve).

Evidence for the end of AGB mass loss and the detachment of the AGB envelope from the photosphere can be found in the spectral energy distribution. As the envelope expands, the decrease in circumstellar extinction allows the central star to become more prominent. This results in a double-peaked spectral energy distribution made up of a cool dust component and a photospheric component reddened by circumstellar extinction (see figure 1).

The morphology

Optical and infrared imaging has found that many PPNs have asymmetric envelopes. For objects that are viewed edge on, two bipolar reflection lobes can be seen. These lobes are the result of light from the central star scattered off the dust in the circumstellar envelope. An example of such bipolar structure is shown in figure 2. The dark region between the bipolar lobes is believed to be due to extinction from an equatorial disk. With the use of mid-infrared cameras on new large ground based telescopes, such disks can be directly resolved and imaged (see INFRARED IMAGERS, INFRARED TELESCOPES).

PLANETARY NEBULAE are found to have diverse morphologies including bipolar (or butterfly) shapes. When and how the spherical AGB envelopes are shaped into such bipolar forms are topics of high current interest. The detection of bipolar nebulosity in PPNs suggests that the shaping of planetary nebulae occurs early, possibly only a few hundred years after the end of the AGB.

Some PPNs show two pairs of searchlight beams and a series of concentric arcs (figures 2 and 3). The beams can be explained as light reflected off a cavity in the polar directions, but the origin of the arcs is not known. The cavity is probably created by a collimated fast outflow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

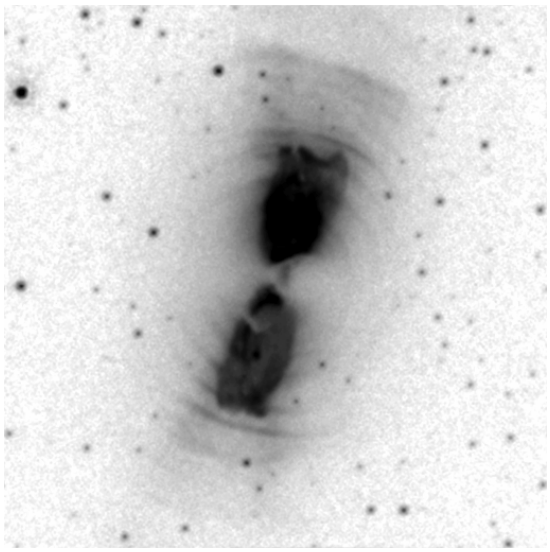


Figure 2. The *HST* V-band image of the PPN IRAS 17150 − 3224 (the Cotton Candy Nebula) showing two bipolar lobes and eight concentric pairs of arcs. The nebulosity is seen from scattered light, not emission lines as in planetary nebulae. **This figure is reproduced as Color Plate 40.**

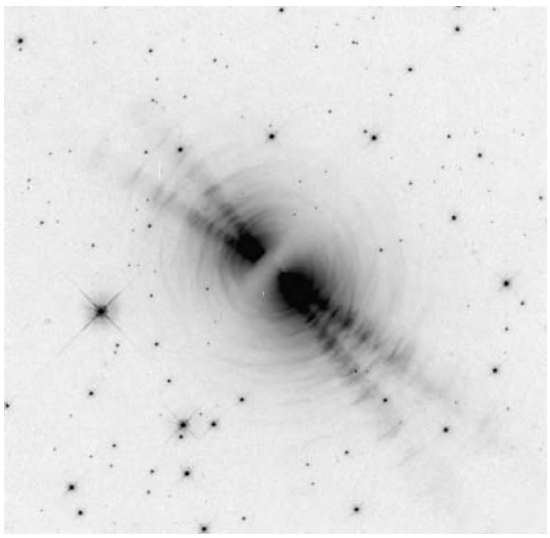


Figure 3. *HST* V-band image of the PPN AFGL 2688 (the Egg Nebula), showing both concentric arcs and searchlight beams. **This figure is reproduced as Color Plate 41.**

that developed during the PPN phase. Support for the existence of a collimated outflow can be found in PPNs such as Hen 401 (figure 4) which have an elongated shape.

Masses and luminosities

The central star of a PPN represents the core of its AGB progenitor. Since the envelope mass is very small ($<10^{-3}M_{\odot}$), the total mass of the central star is approximately equal to the core mass (M_c). The typical

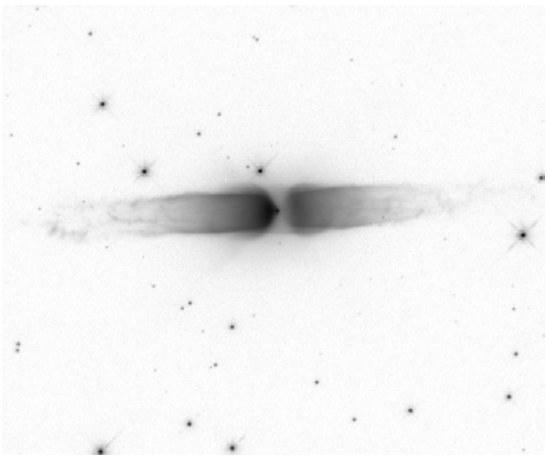


Figure 4. *HST* V-band image of the PPN Hen 401. **This figure is reproduced as Color Plate 42.**

mass of the central star is $0.6M_{\odot}$, but theoretically it can be as high as $1.4M_{\odot}$, the maximum mass limit for a WHITE DWARF (the Chandrasekhar limit). The luminosity of a hydrogen-burning PPN is directly proportional to the mass of the core:

$$L_* \sim 60\,000(M_c/M_{\odot} - 0.5)L_{\odot}.$$

For example, PPNs with $M_c = 0.60M_{\odot}$ and $0.64M_{\odot}$ will have luminosities of $\sim 6000L_{\odot}$ and $8300L_{\odot}$ respectively. PPNs are therefore among the most luminous stars in the Galaxy (see also HIGH-LUMINOSITY STARS).

Which stars will become PPNs?

The discovery of white dwarfs in young open clusters suggests that stars with initial (main-sequence) mass as

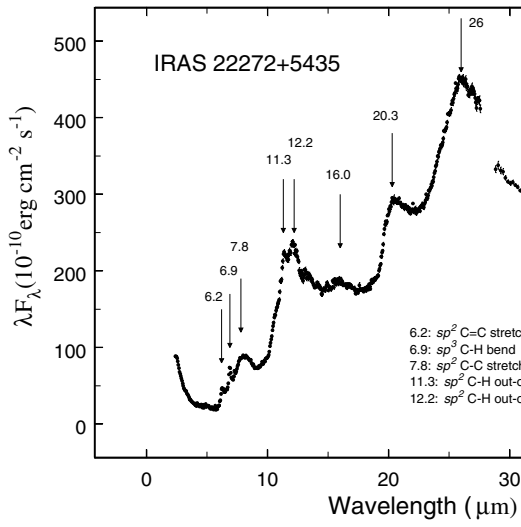


Figure 5. *ISO* spectrum of the PPN IRAS 22272 + 5435 showing various stretching

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

high as (8–10)$M_\odot$ will become white dwarfs. This implies that approximately 95% of stars will end their lives as white dwarfs. However, not all white dwarfs necessarily passed through the planetary nebula stage. Specifically, post-AGB stars with low core masses ($M_c < 0.57 M_\odot$) are believed to be evolving too slowly to cross the HERTZSPRUNG–RUSSELL (H–R) DIAGRAM in time to ionize the circumstellar envelope before it disperses. These post-AGB stars without surrounding nebulae will be observed as blue hot stars radiating primarily in the ultraviolet. Such stars have been suggested to be responsible for the ultraviolet excesses observed in ELLIPTICAL GALAXIES.

Variability

As PPNs evolve across the H–R diagram at constant luminosity, the radii of the central stars will decrease. Consequently, their pulsation periods are also expected to decrease from the several hundred days observed for AGB stars. Pulsations of PPNs have been observed through both spectroscopic (radial velocity) and photometric variations, with periods ranging from 25 to 150 days.

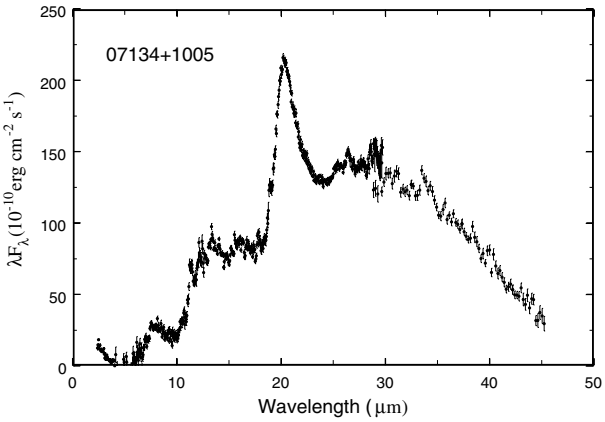


Figure 6. *ISO* spectrum of the PPN IRAS 07134 + 1005 showing the unidentified 21 μm emission feature.

Chemical abundances

PPNs can be classified as oxygen or carbon rich depending on whether the O/C abundance ratio is larger or less than unity. Since the less abundant element is tied up almost exclusively in CO, it is not available to form other molecules. Oxygen-rich PPNs can be identified by the presence of the 9.7 μm circumstellar silicate feature whereas carbon-rich PPNs show molecular absorption bands of C_2, C_3 and CN.

Stars on the AGB undergo hydrogen and helium shell burning as well as neutron capture processes. Some PPNs show a large overabundance of s-process elements (Y, Zr, Ba, La, Ce, Pr, Nd, Sm and Eu), confirming that these objects have indeed evolved through the AGB phase (see table 1).

Chemistry of the circumstellar envelope

Infrared spectroscopy of PPNs has revealed the presence of organic compounds with aromatic and aliphatic structures. A series of infrared emission features at 3.3, 6.2, 7.7 and 11.3 μm due to aromatic C–H and C–C stretching and bending modes are commonly seen in carbon-rich PPNs. In addition, features at 3.4 and 6.9 μm due to stretching and bending modes of aliphatic subgroups are also detected (see figure 5). These features, as well as the underlying broad emission plateaux around 8 and 12 μm, can be found in the laboratory infrared spectra of polycyclic aromatic hydrocarbons, quenched carbonaceous composites, hydrogenated amorphous carbons and coal.

In addition, strong unidentified emission features at 21 and 26 μm are also seen in carbon-rich PPNs (see figures 5 and 6). In particular, the 21 μm feature is unique to PPNs and has not been seen elsewhere in stars in other phases of evolution or in the interstellar medium.

The detection of these carbonaceous compounds suggests that complex organic molecules can be formed over a short ($\sim 10^3$ yr) timescale. This process is believed to first begin during the late AGB phase, where simple organic molecules such as acetylene (C_2H_2) are formed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the AGB wind. Chemical reactions in the envelope can lead to the formation of benzene (C_6H_6) and possibly other aromatic hydrocarbons. PPNs are therefore useful laboratories for the study of ASTROCHEMISTRY.

Bibliography

Reviews on the discovery and observations of PPNs can be found in

Hrivnak B J 1997 *IAU Symp. 180* p 303
Kwok S 1993 *Ann. Rev. Astron. Astrophys* **31** 63

A comprehensive discussion of PPNs in the context of planetary nebulae evolution is given in the book

Kwok S 2000 *The Origin and Evolution of Planetary Nebulae* (Cambridge: Cambridge University Press)

Sun Kwok

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proxima Centauri

The star α Centauri C, a red dwarf member of the Alpha Centauri gravitational system, spectral type M5Ve, apparent magnitude 11.01. The separation from the principal stars of the system, the binary pair (Cen A and B, is so great (2.2°) that although its similar distance from the Sun confirms that it is a member of the system no orbital motion can be detected. It is a star of very low luminosity indeed, absolute magnitude 15.4, and is also a flare star which frequently brightens briefly by about one magnitude. It is a good example of a main-sequence star nearing the end of its life. With a parallax of 0.772″, equivalent to a distance of 4.22 light-years, it is 0.17 light-years closer than the central pair and is thus the closest known star to the solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ptolemaic System

The geocentric view of the universe as expounded by Ptolemy (Claudius Ptolemaus, *c*. AD 100–170, librarian at Alexandria) in the second century AD. According to this view, the Earth lay at the center of the universe and around it, in order of distance, moved the Moon, the planets Mercury and Venus, the Sun, and the planets Mars, Jupiter and Saturn. Beyond the outermost planet lay the sphere of the stars. The Sun and Moon were envisaged as moving around the Earth in circular paths while, in order to account for their observed motions, each planet was considered to move around a circle or epicycle, the center of which moved around a larger circle, the deferent, which was itself centered on the Earth.

The system required additional complex devices in order to account fully for planetary motions and was never entirely satisfactory (although surprisingly successful predictions of planetary positions were achieved). At that time it was felt that the circle was the perfect geometrical form, and that only perfect motion was possible in the heavens. It was not until the work of Kepler (1571–1630) that it was appreciated that the planets move around the Sun in elliptical orbits.

See also: Copernican system, Kepler's laws, Tychonic system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ptolemy

The last important astronomer of classical antiquity, Ptolemy is also the only one whose works have come down to us substantially intact. His *Almagest*, a comprehensive empirical and mathematical deduction of models to explain the motions and phenomena of the heavenly bodies, laid the methodological foundation for the Islamic and European mathematical astronomy of the Middle Ages and Renaissance.

Biography and writings

Claudius Ptolemaeus (Ptolemy) worked at or near Alexandria in Egypt during the middle decades of the 2nd century AD. He began making astronomical observations at least as early as AD 127. By 147, he had completed a preliminary version of his system of planetary models, which he described in outline in a public inscription that was later preserved in manuscript copies. The full presentation of the system, incorporating some significant revisions, was given in a large treatise entitled *Mathematical Composition*, which much later came to be universally called the *Almagest* (from an Arabic title meaning 'the greatest'). Ptolemy wrote several other works on physical science, most of them after the *Almagest*. Those that are directly concerned with astronomy include the *Phases of the Fixed Stars*, the *Handy Tables*, the *Planetary Hypotheses* and brief works on stereographic projection and sundial theory. The original Greek texts of the *Almagest* and *Handy Tables* have been preserved in medieval copies, whereas the other works are incomplete, or extant only in Arabic versions.

The *Almagest*

The purpose of the *Almagest* is to establish quantitative geometrical models based on uniform circular motion for the Sun, the Moon, the five known planets and the fixed stars (see also PLANETARY THEORY FROM EUDOXUS TO COPERNICUS). Ptolemy attempted to make the work thoroughly empirical, deductive and demonstrative, following a course that he succinctly delineates in the introduction to his planetary theory: 'anyone who was to convince himself and his future audience must demonstrate the size and the period of each of the two anomalies [in a planet's motion] by means of well-attested phenomena which everyone agrees on, must then combine both anomalies, and discover the position and order of the circles by which they are brought about, and the type of their motion; and finally must make practically all the phenomena fit the particular character of the arrangement of circles in his hypothesis'. This is an axiomatic approach, resembling Greek geometry, rather than a heuristic one. The starting points are (a) hypothetical model types for the heavenly bodies, (b) approximate values for the periodicities, taken over from earlier astronomers, and (c) records of observations and measurements.

Ptolemy's plan also resembles a Greek geometrical treatise in being cumulative, building demonstrations on prior demonstrations. He begins with arguments, chiefly from qualitative observation, that the Earth and heavens are spheres with a common center, and that the Earth is stationary and negligible in size relative to the celestial sphere. This is followed by a table of the lengths of chords in a standard circle (used for trigonometrical calculations), a measurement of the obliquity of the ecliptic using a meridian ring and a group of geometrical theorems useful for solving problems in spherical astronomy, which Ptolemy immediately applies to deducing tables for declinations and right and oblique ascensions.

The natural frame of reference for planetary positions in ancient astronomy was the backdrop of fixed stars. Because he knew from Hipparchus' work (see HIPPARCHUS) that the stars have a precessional motion relative to the equinoctial points, and considered it necessary to demonstrate the nature of that motion before using the stars as reference points in observations, Ptolemy instead chose to build his chain of deductions from a solar theory demonstrated from observations of the dates of solstices and equinoxes. Drawing heavily on Hipparchus' writings on solar theory, Ptolemy argues that the tropical year, i.e. the interval measured between observed equinoxes or solstices of the same kind, is constant, and approximately 365;14,48 days (where the numerals after the semicolon represent sixtieths, sixtieths of sixtieths, and so forth). Hence it also follows that the lengths of the astronomical seasons do not change, and Ptolemy claims to have confirmed by 'very precise observations' that Hipparchus' values ($94\frac{1}{2}$ days for astronomical spring and $92\frac{1}{2}$ days for astronomical summer) were valid in his own time. On this basis Ptolemy took over (with acknowledgment) Hipparchus' solar model, according to which the Sun travels with a period of 365;14,48 days on a circle whose center is displaced from the Earth's center by 1/24 of the orbit's radius in a fixed direction $65\frac{1}{2}^\circ$ east of the vernal equinoctial point.

Ptolemy's handling of the solar theory calls for several remarks. First, without a constant tropical year his rationale for giving priority in the *Almagest* to the solar theory would have been undermined. Secondly, while there is no obvious reason why Ptolemy should not have been able to observe equinoxes and solstices generally within the quarter-day tolerance that Hipparchus correctly claimed for his own observations, consistent results unambiguously confirming the constancy of the seasons would have been much harder to achieve. Thirdly, Ptolemy's observations of the equinoxes and summer solstice in AD 139 and 140, on which his demonstration rests, are not genuine, since they are consistently about a day too late. Why Ptolemy introduced falsified observations at this point, whether he made real ones and whether his proceedings were justifiable are topics of controversy. Since Ptolemy reckons longitudes from the vernal equinoctial point rather than relative to the stars, the error of a day introduces a systematic and gradually increasing error into all longitudes derived from his models, and the false observation reports continued

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to bedevil later astronomers' solar theories for centuries afterwards.

The lunar theory in the *Almagest* begins along the same lines as Hipparchus', with a calculation of the radius of the Moon's epicycle (or, equivalently, the eccentricity of its orbit) from sets of three observed lunar eclipses on the hypothesis of a simple epicyclic model. Unlike Hipparchus, Ptolemy obtains a consistent result, namely an epicycle radius of $5\frac{1}{4}$ units where the deferent circle's radius is 60 units; the consistency is enhanced, however, by small 'fudgings' in Ptolemy's arithmetic. Ptolemy revises Hipparchus' lunar periodicities after finding the dimensions and configuration of the model, so that he is able to use for this purpose very widely spaced eclipse observations subject to fewer restrictions than Hipparchus required.

From a study of observations recorded by Hipparchus, Ptolemy discovered that the lunar anomaly predicted by this model was too small in situations other than syzygy, with the maximum discrepancy when the Moon was elongated 90° from the Sun. He therefore modified the model by replacing the Moon's deferent circle with a rapidly revolving eccentric circle so as to draw the epicycle closer to the Earth at quadrature than at SYZYGY. The revised model, which is quite successful in predicting lunar longitudes, has the defect of falsely implying that the lunar disk ought to appear much larger at half Moon than at full Moon. Ptolemy does not mention the problem, although he can hardly have been unaware of it.

Ptolemy was now in a position to calculate lunar positions accurately and to compare these with observed positions to measure the Moon's parallax and hence its distance from the Earth. The calculation is marred by errors, yet arrives at a fairly accurate result (mean distance at syzygy of 59 Earth radii), a clear illustration of Ptolemy's contriving to make a derivation arrive at an answer that he already knew by other means. In this instance, an inaccurate parallax would have shown up in the prediction of solar eclipses, and Ptolemy also could have been influenced by previous estimates by Hipparchus and others. Ptolemy uses Hipparchus' eclipse diagram (see HIPPARCHUS, figure 1) in reverse to find a solar distance from his lunar distance, disregarding the extreme sensitivity of this calculation to small variations in the data.

After the Moon Ptolemy turns to the fixed stars, showing from observations spanning more than four centuries that the configurations of stars do not vary (that is, they have no proper motion), and that the entire 'sphere of fixed stars' revolves about the poles of the ecliptic. Ptolemy finds too small a rate of precession (1° in 100 yr instead of in about 72 yr), but this almost exactly cancels out the error in his tropical year, so that the mean sidereal year implicit in his theory is unexpectedly good. A catalog of 1022 fixed stars, with longitudes (normed for AD 137), latitudes and numerical magnitudes ranging from 1 to 6, follows at this point. Ptolemy describes this as the product of his own observations; historians are in disagreement about the truth of this assertion, and whether the catalogue actually incorporates an older catalogue by Hipparchus adjusted by Ptolemy for precession.

Ptolemy's planetary models impose an epicycle (to exhibit the synodic cycle of direct and retrograde motion) on an eccentric orbit (to exhibit the variations in the synodic cycle). For the INFERIOR PLANETS the epicycle's center revolves about the Earth with the same period as the Sun, whereas the SUPERIOR PLANETS revolve about their epicycles with a period such that they are approximately at the perigee of the epicycle when in opposition to the mean Sun. Ptolemy further modified the eccentric model by introducing the so-called equant point, a point twice as distant from the Earth as the center of the eccentric deferent along the same line; the revolution of the epicycle around the deferent is postulated as uniform as seen from the equant point rather than from the deferent's own center. The effect of this model is roughly to make the planet's deferent appear to have twice as great an eccentricity at opposition and conjunction as at situations half-way between.

The procedure for measuring the components of the models is different for the inferior and superior planets. In the case of Venus and Mercury, Ptolemy is able, so to speak, to observe the edges of the epicycle directly when the planet is at its greatest elongation from the Sun; hence he can straightforwardly establish the apogee of the deferent and the size of the epicycle. In the case of the remaining planets Ptolemy selects sets of three observations of oppositions with the mean Sun, when the planet is theoretically seen along the same line with the epicycle's center, and applies the same Hipparchian method of determining an eccentricity and apogee as in the solar and lunar models. The procedure is made much more difficult here by the equant motion, to account for which Ptolemy carries out a sequence of iterative corrections to his first calculation. The epicycle is then measured using an observation of the planet at intermediate elongation from the Sun. The verification of the periodicities is carried out after the determination of the radii and apogee, as was the case with the Moon. In contrast to the closely reasoned treatment of the planetary models so far as they concern motion along the ecliptic, Ptolemy's rather complicated system of wobbles to produce their latitudinal motion is presented as a *fait accompli*, unsupported by specific observations.

Other writings

Ptolemy concluded the *Almagest* with a theory of the limits of planetary visibility, which served by its agreement with the accepted planetary phenomena as a confirmation of the validity of Ptolemy's models. The essential concept here was the *arcus visionis*, a critical arc by which the Sun must be below the horizon along an altitude circle if the planet is to be visible. In his *Phases of the Fixed Stars* (the theoretical part of which has not survived) Ptolemy extended this visibility model to the stars, taking into account the effects of the observer's latitude and the latitudes of the stars. The extant part of the book applies this information to produce

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a calendar correlating computed dates of appearance and disappearance of stars with regular patterns of weather in different localities.

Interspersed in various parts of the *Almagest* are many tables for computing positions and phenomena of the heavenly bodies, derived from the models. These have an internal application in the deductions of the *Almagest*, but also were potentially useful for other purposes, especially ASTROLOGY. After completing the *Almagest* Ptolemy revised the tables, and published them separately as the *Handy Tables*. In addition to many improvements in the format of the tables, Ptolemy tacitly simplified the underlying models for planetary latitude and reworked the tables for planetary visibility. The *Handy Tables* was widely disseminated in antiquity.

The *Planetary Hypotheses* represents Ptolemy's last word on the celestial motions. In this work he treats the models of the *Almagest* in a more physical way, as a system of rotating spheres. Most of the *Planetary Hypotheses* exists only in an Arabic translation (see also MEDIEVAL COSMOLOGY) that is often difficult to interpret, and there is some ambiguity about the extent to which Ptolemy meant to delineate the actual structure of the heavens, since he also evidently has in mind the construction of a sort of ORRERY. A feature of this book that was profoundly influential in medieval astronomy was Ptolemy's presentation of the models as a series of nested spheres, with the outermost sphere of the Moon directly continguous with the innermost sphere of Mercury, and so on outwards to the sphere of the fixed stars (see ISLAMIC ASTRONOMY). By a remarkable coincidence, as Ptolemy discovered, such a packing is almost perfectly in agreement with the solar and lunar distances that he determined in the *Almagest*.

Bibliography

Neugebauer O 1975 *A History of Ancient Mathematical Astronomy* (Berlin: Springer)

Pedersen O 1974 *A Survey of the Almagest* (Odense: Odense University Press)

Toomer G J 1984 *Ptolemy's Almagest* (London: Duckworth)

Alexander Jones

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pulsar Backends

Present-day pulsar astronomers require equipment which will perform two main tasks. The first of these is to search for new PULSARS and the second is to make observations of known ones, primarily to make precision timing measurements, but also to study the polarization and morphology of the pulses. These require rather different capabilities from the equipment, but the two things they have in common are the necessity for high time resolution and high sensitivity. The signal-to-noise ratio, which is the main measure of the quality of a pulsar observation, is approximately

$$\mathrm{SNR} = K\left[NBT\left(\frac{P}{W}-1\right)\right]^{1/2} \qquad (1)$$

where K is a constant which depends on the size and performance of the radiotelescope, N is the number of polarization channels (usually 2), B is the total receiver bandwidth, T is the duration of the observation and P and W are the period and width of the pulses. This equation clearly indicates that the use of a receiver with a large bandwidth is important. Moreover, we see that the sensitivity is greatest for small values of the width of the observed pulses W. Thus any instrumental process which broadens the pulses reduces the sensitivity.

Dispersion

What makes the tasks non-trivial is the dispersion of the signals caused by the weak frequency dependence of the velocity of the pulses during their passage through the interstellar medium. The gas in the interstellar space is ionized to a greater or lesser extent by ultraviolet and x-ray emission from stars. The pulsar signals travel through the resultant electron plasma at a radiofrequency ν with the group velocity of

$$v_\mathrm{g} = c\left(1-\frac{n_\mathrm{e}e^2}{2\pi m\nu^2}\right) \qquad (2)$$

where n_e is the free-electron plasma density, c is the free-space velocity of light and e and m are the charge and mass of the electron. The effect of this can be seen most clearly in figure 1 which shows the dispersion of the radio pulses from the pulsar B1356 − 61, the pulses at low frequencies being delayed relative to those at higher ones. Clearly, if one uses a receiver with bandwidth exceeding a few of these frequency channels, the observed pulse will be much wider than the true pulse, so reducing the sensitivity.

The amount of frequency drift is determined by the column density of free electrons along the line of sight to the pulsar, a quantity known as the dispersion measure.

Dedispersion

Of course one could combat the broadening effects of this dispersion simply by using a very narrowband receiver but this would be at the penalty of much reduced sensitivity as can be seen from equation (1). However, there are several ways in which the effects of dispersion may be removed from the data, while retaining full sensitivity, using processes which come under the generic name of dedispersion.

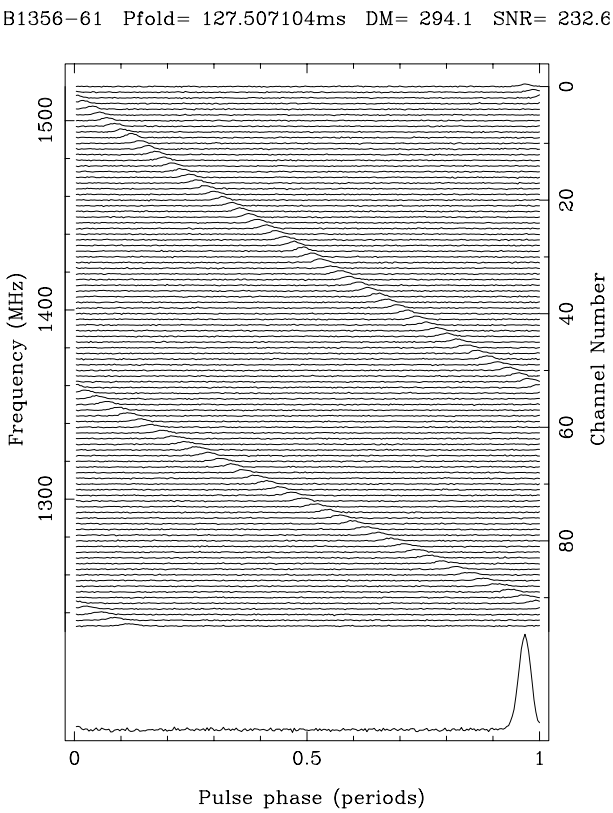


Figure 1. Frequency dispersion in the pulses from PSR B1356 − 61. Each horizontal trace shows the pulse profile recorded in one of 96 adjacent frequency channels, each 3 MHz wide, centered on 1380 MHz. The pulse profile at the bottom is the result of correcting the individual profiles for the drift in frequency and then adding them together.

The most widely used technique is to segment the receiver band into a number of much smaller bands in which the dispersion broadening is small and to record the power from each (see figure 1). If the dispersion measure is known, the outputs from the individual channels may be delayed by appropriate amounts before adding them all together to give a 'dedispersed' time sequence with a resolution determined by the broadening arising in the single channels.

There are two main practical ways of achieving this—by the use of FILTERBANKS or the use of autocorrelation spectrometers. The former essentially requires a bank of separate contiguous, narrowband receivers, with bandwidths equal to their frequency separation. These can often be constructed simply and cheaply using simple LC-tuned circuits. These have the disadvantage of inflexibility and more expensive digital filters have been employed in which the bandwidth and frequency are programmable.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The alternative digital autocorrelation spectrometers (CORRELATORS) form the basis of most spectral analysis in modern RADIO ASTRONOMY and are also used in pulsar astronomy. These first involve the computation of the autocorrelation function (ACF) of the data. Fortunately, the ACF is easily calculated digitally and accumulated over a period of time. Its Fourier transform may be calculated in a computer to provide the power spectrum (from a relationship known as the Wiener–Khinchin theorem). This is essentially the output of a filterbank and, as before, after the application of appropriate delays and additions, a dedispersed time sequence may be obtained.

Unfortunately there is a fundamental limit to the achievable time resolution of such post-detection, or incoherent, dedispersion techniques. As the bandwidth of the individual channels is reduced, the dispersion broadening is reduced. However, the use of a filter of bandwidth b limits the response time of the output to roughly $1/b$ and hence there is a critical bandwidth below which the time broadening starts increasing again.

There is only one technique which can overcome this limitation; pre-detection, or coherent, dedispersion. This requires sampling the voltages of the full receiver bandwidth digitally and then applying exactly the inverse of the filter which has been applied by the interstellar medium, which essentially rotates the phase of the signals as a function of frequency according to the dispersion law. This is usually achieved by performing a Fourier transform of the data, rotating the phases of every element in the transform and then performing the inverse Fourier transform. The result is a time sequence which is essentially the output of a receiver which was very close to the pulsar. In this case, the time resolution is as little as $1/B$, the inverse of the full receiver bandwidth. For example, with a bandwidth of 100 MHz, time resolution of only 10 ns is possible.

This of course is the ideal process but requires very large computational resources and, while it is used for a number of studies of millisecond pulsars, it is not worthwhile for most slower pulsars.

Searching for pulsars

Most modern searches for pulsars are based on the identification of signals which are dispersed and periodic in nature. These searches are usually conducted by recording the output of a filterbank or spectrometer, using a sampling interval which is small compared with the intrinsic width of the pulses of the target pulsars. If this is not so, the effective pulse width will be increased and the sensitivity of the search will be reduced according to equation (1). The recorded data are then dedispersed for a number of trial dispersion measures, each trial producing a time sequence which is then searched for pulse trains of unknown periodicity and pulse width. This search is usually carried out by performing a fast Fourier transform of the data and seeking large spikes in the resultant spectrum. The search sensitivity can be improved by summing up to $P/2W$ harmonics of the fundamental period P, where W is the pulse width; the sensitivity is then improved by a factor $(P/2W)^{1/2}$ and is consistent with that given by equation (1).

While these computations can in principle be carried out in real time, the required computing resource is considerable and there is no possibility of reprocessing the data, for instance if the data from some filterbank channels are spoilt by interference. For these reasons, data are usually recorded directly onto magnetic tape and processed later on off-line computers in a more leisurely fashion. The volume of recorded data in modern searches is often large. It is not unusual for 512 filterbank channels to be sampled every 100 μs, requiring the storage of 5 million data elements per second. Recording say 8 bits for each element requires a sustained data rate of 5 Mbytes s^{-1}. However the number, and hence cost, of tapes can be reduced by representing each element by a smaller number of bits, resulting in a somewhat reduced sensitivity. In fact, in the extreme case of the small signals expected from undiscovered pulsars, single-bit sampling provides a sensitivity which is reduced by a factor of only $(2/\pi)^{1/2} = 0.80$ for an eight-fold reduction in data volume. Such digitization simply requires the recording of a '0' or a '1' depending on whether the signal level is less than or greater than the long-term mean level. The simplicity and economy of such sampling make it widely used (see also SOFTWARE: RADIO PULSAR SOFTWARE).

Timing measurements

Much of the most interesting astrophysics in pulsar astronomy arises from the ability to monitor the rotation of pulsars through very high-precision measurements of their arrival times. This is usually achieved by dedispersing and folding the data at the known period of the pulsar over a sufficiently long interval of time, typically 10 min, to give a high signal-to-noise ratio profile. Such folding may be carried out within special-purpose hardware or in general-purpose computers.

In order to perform this operation, the pulsar dispersion measure and the instantaneous period must be known. The latter requires an accurate model of the rotation of the pulsar, as it would be observed at the BARYCENTER of the solar system, together with an ephemeris for the rather complicated motion of the telescope around the barycenter. This then allows calculation of the phase of the data samples within the pulse period so that they may be added into one of a number of hardware or software registers, each containing data for a single pulse 'phase'.

The input data for this process is usually the output of a filterbank or spectrometer or a coherent dedispersion process. The data may be dedispersed completely, partially or not at all prior to folding, in which case the dedispersion is completed afterwards. The reason for this is that the dispersion measure may not be well determined at the time of observation, and incorrect dedispersion may lead to distortion, or smearing, of the pulse. Likewise, an incorrectly known period may lead to smearing of the pulse. To combat this, an observation is usually

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

divided into a number of subintegrations within which the distortion is small. These may be combined subsequently when the period is better determined.

In order to use such data for timing measurements, the local observatory clock must be synchronized with UTC or some other atomic time standard. This is usually carried out using the Global Positioning Satellites (GPS) with an accuracy of better than 0.1 μs. The profile must then be appropriately time tagged and, from this, a time of arrival is usually obtained by convolution with a template, which is a standard, noise-free profile. The time of arrival is taken to be the peak of the resulting function.

Single pulses

Most pulsars show a considerable amount of detail and variation from pulse to pulse, revealing much about the details of the emission process. The acquisition of such data uses most of the principles discussed in the previous sections and usually involves the recording of dedispersed time sequences. Sometimes, the amount of data recorded can be reduced by storing only the data lying within narrow 'windows' centered on the pulse, rather than the whole period.

Polarization measurements

The measurement of the polarization of the radio pulses is important in establishing the configuration of the pulsar rotation axis and the NEUTRON STAR magnetosphere. Most modern telescopes have receivers providing two channels, A and B, which are sensitive to either left-hand and right-hand circular polarizations or to orthogonal linear polarizations. These pairs of signals can be combined to provide the four Stokes parameters, I, V, Q and U, which completely specify the polarization state of the radiation. For incoherent dedispersion systems, these are obtained electronically for each element of the filterbank and each of the four Stokes parameters is then dedispersed separately. For coherent dedispersion systems, the phase information is still available in the dedispersed time sequences so that only A and B need be recorded and dedispersed, the full Stokes parameters being calculated digitally from the resulting dedispersed time sequences.

Bibliography

General introductions to the astrophysics of pulsars are given in

Manchester R N and Taylor J H 1975 *Pulsars* (San Francisco, CA: Freeman)

and

Lyne A G and Graham-Smith F 1998 *Pulsar Astronomy* (Cambridge: Cambridge University Press)

An excellent discourse on the principles of pulsar data acquisition and signal processing is given by

Hankins T H and Rickett B J 1975 Pulsar signal processing *Methods Comput. Phys.* **14** 55–129

Andrew Lyne

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pulsars

Pulsars are the lighthouses of the Galaxy—rapidly spinning NEUTRON STARS whose strong magnetic fields produce conal beams of electromagnetic radiation that sweep past the Earth with each rotation of the star, producing the eponymous pulses that are observed primarily at radio wavelengths.

Like other neutron stars, radio pulsars are born in the SUPERNOVA explosions that accompany the collapse of massive stars. The nascent pulsars are born rotating at up to about 100 times per second. It is this stored rotational kinetic energy that powers the pulsar, so like a spinning top the pulsar gradually slows, reaching spin periods of about a second within a few million years. Eventually, within about 100 million years, the pulsar is spinning too slowly to maintain its radio emission, and it fades from view. Some old pulsars that have binary companions can be 'recycled', or spun back up to fast rotation periods by mass transfer from their companions. Because the resulting 'millisecond pulsars' have relatively low magnetic field strengths and hence low energy loss rates, they can continue to spin rapidly for times that are long compared with the age of the galaxy.

Just over 1000 pulsars are now known, all in our own galaxy except for a few pulsars detected in the MAGELLANIC CLOUDS. They are studied because neutron stars are intrinsically interesting astronomical objects, but also because the study of pulsars is deeply intertwined with many different branches of both astronomy and physics. Pulsars are, for example, very useful astrophysical probes. For example, a sharp radio pulse emitted by a pulsar is delayed and broadened during its propagation through the dispersive, turbulent interstellar medium, in a way that depends on the frequency and polarization of the signal as well as the properties of the medium. Multifrequency studies of pulsar signals have been used to map the distribution and turbulence structure of ionized material in the Galaxy, as well as the average Galactic magnetic field.

Pulsars are also, by virtue of their very regular, clocklike pulses, useful probes of the gravitational environments in which they are found. The Doppler shifts of the signals from pulsars in binaries can be used to study the binary properties, just as spectral lines are used with normal stars. Some very close pulsar binaries have orbits that are substantially deformed from Keplerian ellipses by general relativistic effects; in these systems, very precise tests of 'post-Keplerian' gravity theory have been possible.

Discovery and early history

Pulsars were discovered in 1967 by a Cambridge University graduate student, JOCELYN BELL (now Jocelyn Bell Burnell), working under the supervision of her thesis advisor, ANTHONY HEWISH. Hewish, Bell and their collaborators had designed and built a new telescope near Cambridge that was designed to study the rapid SCINTILLATION, or twinkling, of radio signals from QUASARS as they propagated through the SOLAR WIND. Studying the thousands of feet of chart recorder paper produced their receivers, Bell noticed some unusual 'scruff', from a direction well away from the Sun, that recurred on several nights at a fixed sidereal time. Follow-up observations showed that the signal had the form of a series of short radio pulses, repeated every 1.337 s. The extreme regularity of the pulsations was puzzling. After ruling out terrestrial interference sources, the team briefly considered the possibility of an artificial extraterrestrial origin—leading Bell to jokingly call the source LGM, for 'little green men'—but her quick discovery of three more pulsating radio sources demonstrated that they must arise from some natural source. A distance estimate of 65 pc was obtained for the first object, from measurement of the DISPERSION of the signal in the ionized interstellar medium, placing the source far beyond the solar system.

The rapid variability of the signal required a source that was much smaller than a light-second across: smaller, that is, than an ordinary star. The name 'pulsar'—pulsating star—enshrines the original model put forth by Hewish and his collaborators in which the pulses were produced by radial oscillations of a WHITE DWARF or neutron star. Models involving rotation rather than pulsation were, however, strongly favored by observations that the pulsation periods were gradually lengthening. (A top, as it loses energy, slows down; a bell rings with a constant frequency but a decreasing amplitude.) The identification of pulsars with neutron stars rather than white dwarfs came with the discovery in 1968 of pulsars in the Vela and Crab supernova remnants. Not only were neutron stars thought to be born in supernovae, but the Crab pulsar had a period of only 33 ms, much faster than allowed by any white dwarf model. (Remarkably, Franco Pacini had predicted that rotating, highly magnetized neutron stars were a potential source of electromagnetic radiation just a few months before pulsars were discovered.)

Hewish was awarded the 1974 Nobel Prize in Physics for his 'decisive role in the discovery of pulsars' and his pioneering work in radio astronomy. Bell's key role in the discovery has been widely recognized: among other awards, she has received the Michelson Medal of the Franklin Institute in Philadelphia (jointly with Hewish), the Tinsley Prize of the American Astronomical Society and the Herschel Medal from the Royal Astronomical Society.

The pulsar population

There are just over 1000 known radio pulsars (March 1999), with pulse periods ranging from 1.557 ms to over 8 s. This known sample represents fewer than 1% of the active pulsars in the Galaxy, with the others either too faint or too distant to detect, or with radio beams whose sweeps do not intersect the Earth. Most known pulsars lie within a few kiloparsecs, although some very luminous pulsars are observed in the distant reaches of our Galaxy and a few in the Magellanic clouds. No known pulsar is bright

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

enough to be detectable if it were placed in our neighbor galaxy M31.

A new pulsar is born in the Galaxy every 50–300 yr. This rate is somewhat lower than the estimated rate of core collapse supernova explosions (type II and type Ib/c) in the Galaxy, thought to be one every 20–50 yr. Both rate estimates are subject to many uncertainties, so the significance of the discrepancy is unknown. It is possible, however, that many core collapse supernovae do not produce active pulsars.

Pulsars in supernova remnants

Only a relatively small number of known SUPERNOVA REMNANTS harbor known radio pulsars. There are perhaps eight secure associations, and another dozen proposed associations where the possibility of a chance line-of-sight coincidence between the pulsar and remnant cannot be ruled out. Because pulsars remain active for over 10 million years, and their remnants fade within a hundred thousand years, the absence of remnants around most pulsars is easily understood. Remnants without pulsars may result from type Ia supernovae, may have pulsars that are too faint to be seen or are beaming in a direction away from Earth, or may contain neutron stars that are not active as pulsars for some reason (e.g. the soft gamma repeaters).

The association between the CRAB NEBULA, a plerionic supernova remnant, and the pulsar PSR B0531+21 has been particularly important. Chinese records of the supernova, which was first detected by Sung astronomers on 4 July 1054 and which remained visible in daylight for 3 weeks and at night for nearly 2 yr, have given modern astronomers a precise age for the pulsar. Tracing the motion of the filaments in the nebula backwards in time, they converge nearly a century after the supernova took place, implying an acceleration of the expansion after the original explosion. The pulsar provides the long-lived energy source for this acceleration.

Pulsar ages and magnetic fields

Observations of pulsars over extended periods of time allow the measurement of both the pulsar period P and the rate at which the period is lengthening $\dot{P}$ (or equivalently the frequency Ω and $\dot{\Omega}$). Often, it is assumed that the braking torque on the pulsar is proportional to some power of the pulsar spin frequency, so $\dot{\Omega} \propto \Omega^n$ where n is the 'braking index' ($n = 3$ for braking by magnetic spin-down radiation). If the original spin frequency was Ω_0, then the current age of the pulsar is

$$t = -\frac{\Omega}{(n-1)\,\dot{\Omega}}\left[1 - \left(\frac{\Omega}{\Omega_0}\right)^{n-1}\right]. \qquad (1)$$

A 'characteristic age' is often defined by assuming that $\Omega \ll \Omega_0$ and $n = 3$, so

$$\tau = -\frac{\Omega}{2\dot{\Omega}} = \frac{P}{2\dot{P}}. \qquad (2)$$

In a few cases, the braking index can be directly determined from measurements of the second derivative of the pulsar frequency, since $n = \Omega\ddot{\Omega}/\dot{\Omega}^2$. Measured values are between 1.4 and 3, slightly lower than prediction of the magnetic dipole model.

With this caveat in mind, the assumption that magnetic dipole radiation dominates the pulsar energy loss rate leads to an estimate of the surface magnetic field strength:

$$B \approx \left(\frac{3Ic^3P\dot{P}}{8\pi^2R^6}\right)^{1/2} \qquad (3)$$

where I is the moment of inertia of the neutron star and R its radius. For typical values from neutron star models, and P in seconds, we have

$$B \approx 3.2 \times 10^{19}(P\dot{P})^{1/2}\ G. \qquad (4)$$

In figure 1, we show the known pulsars with measured P and $\dot{P}$, including lines of constant characteristic age (equation (2)) and surface magnetic field (equation (4)). The pulsars in supernova remnants are shown by stars; as expected, they are among the youngest known pulsars. The pulsars in the lower left region of the figure are the 'recycled' pulsars; they will be discussed further below. The great majority of known pulsars are seen to have inferred magnetic field strengths of 10^{11}–10^{13} G and ages of 10^3–10^8 yr.

After pulsars are formed, pulsar spin periods lengthen and individual pulsars move from left to right in the figure. The precise shape of their tracks depends on the evolution of their magnetic fields, and hence the spin-down torques. For many years it was assumed that fields would decay owing to ohmic losses on timescales of 10^6–10^7 yr, but the current understanding is that the high conductivity interior to the star prevents field decay at least on 10^8 yr timescales. Pulsars would then move along lines of constant magnetic field. Eventually, pulsars reach the region in the lower right of the figure and are lost from the sample, either fading below detectable luminosity or crossing a 'death line' below which the radio emission process does not function.

The pulsar velocity distribution

Pulsars have high space velocities. This is observed directly, with interferometric measurements of PROPER MOTIONS corresponding to typical transverse speeds of around 300 km s^{-1}. Although most pulsars are born in the explosions of massive stars that are found near the Galactic plane, their high velocities carry them to a large observed scale height above (and below) the plane, $z \sim 500$ pc.

The origin of the pulsar velocities is still uncertain. Many pulsars are born in binaries, and binary disruption during a supernova explosion can leave the individual stars with large velocities. Binary disruption cannot account for the highest observed velocities, however. A more promising source of a large 'kick' is anisotropy in the supernova birth event. Even a small asymmetry in the neutrino emission from a supernova explosion is sufficient to leave the neutron star moving at 1000 km s^{-1} or faster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

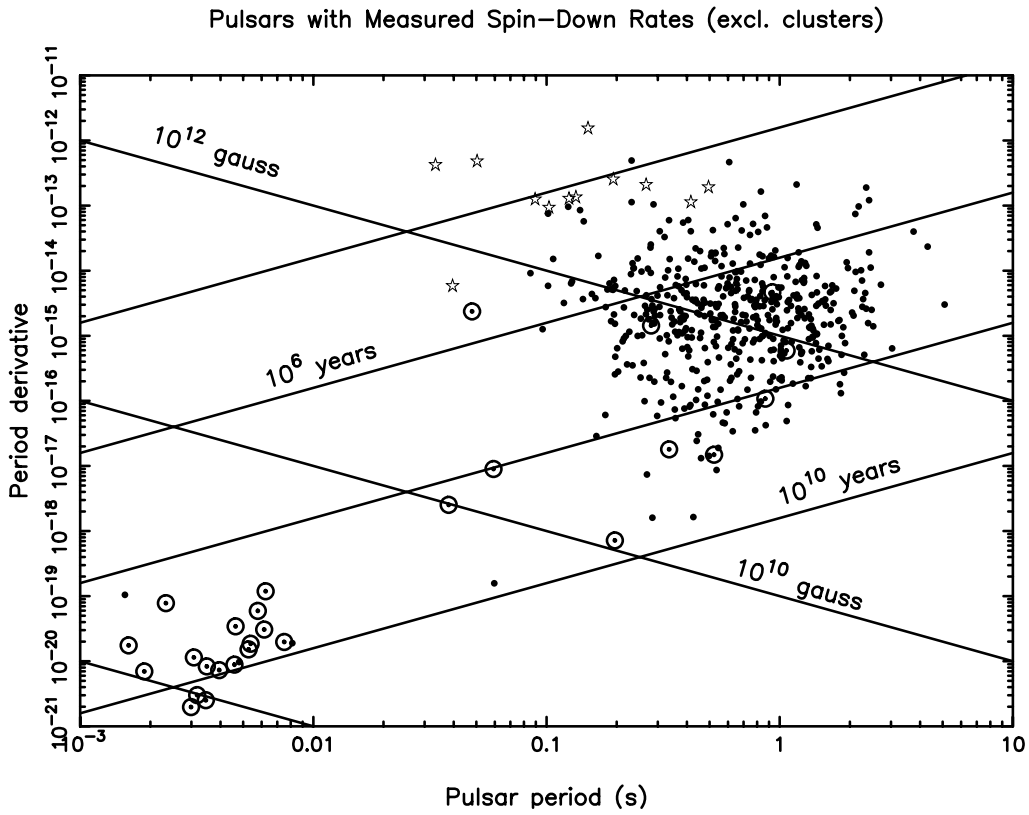


Figure 1. Pulsars with measured spin-down rates (excluding pulsars in globular clusters). Binary pulsars are marked with open circles, pulsars in supernova remnants with stars.

Radio and high-energy emission from pulsars

Although a basic physical model of pulsars was developed within a few years of their discovery, the pulsar radiation source itself remains rather mysterious. Astronomer Jon Arons has observed that we know why pulsars pulse, but not how they shine.

Some aspects of the emission process are straightforward. The neutron star is very strongly magnetized, with a field that is at least roughly dipolar and whose axis is misaligned with the rotation axis of the star. The rotating magnetic field produces a strong induced electric field which accelerates charged particles off the surface of the star into an ionized 'magnetosphere'. The conductivity is high along the magnetic field lines but very low perpendicular to the lines, so the charged particles are forced into corotation with the star. At a radial distance $r_c = c/\Omega$ a corotating particle would travel at the speed of light, so corotation must break down within this 'light cylinder'. Magnetic field lines that close back to the star within this radius can support a static charge distribution, but the 'open' field lines that cross the light cylinder define a conal region above the pulsar's magnetic polar cap in which particles can be accelerated to high energies. It is this polar cap region where pulsar radio emission is probably generated. Particle acceleration may also occur in a depleted region that develops just inside the light cylinder; this 'outer gap' is frequently invoked as the source of high energy (optical, x-ray and γ-ray) emission observed from some pulsars.

This simple picture remains incomplete, with such questions as 'current closure' (global charge conservation) remaining unsolved. It is remarkable that the seemingly straightforward electrodynamical problem of a non-aligned, magnetized rotator has so far defied analysis.

The radio emission mechanism

Observed pulsar radio luminosities, together with the small source size, imply extraordinarily high brightness temperatures—as high as 10^{31} K. To avoid implausibly high particle energies, coherent RADIATION processes are invoked. A maser-like mechanism, involving particles bunched in momentum space, is attractive, if only because maser action has been observed elsewhere in astrophysics, but models with coherent emission from bunches of particles have also been widely discussed, with a bunch of N particles localized in physical space radiating power proportional to N^2. Coherence by bunching is seen in terrestrial lightning flashes. In detail, severe problems remain in understanding pulsar emission by either the maser or the bunching models, and no consensus has emerged.

The emission mechanism itself also remains uncertain. Charged particles gyrating around magnetic field

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

lines produce SYNCHROTRON RADIATION. In the strong magnetic fields of the pulsar magnetosphere, a particle will quickly radiate aways its components of momentum that are perpendicular to the field lines, so will be confined to the lowest Landau level. Roughly, the charged particle can be pictured as a bead on a wire, along which the bead is free to move. As a particle moves out along a curved field line, it will produce synchrotron-like radiation that is conventionally called 'curvature radiation'. Coherent curvature radiation is currently the most widely accepted model for pulsar radio emission, but many other possible models have been discussed, including models based on relativistic plasma instabilities that are varients of the mechanisms proposed for type III solar radio bursts (see SOLAR FLARES: RADIO BURSTS).

Polarization

The radio emission from pulsars is frequently highly polarized, with linear polarization fractions reaching nearly 100% in some cases. In most emission models, the electric vector in the radiation can be associated with a preferred direction: either in the plane of the curving magnetic field lines or perpendicular to that plane. In many cases, there is evidence for radiation in both of two orthogonal modes of linear polarization. This can be observed, for example, by monitoring the pulse-to-pulse variations in the polarization vector. The position angle of the linear polarization often swings smoothly across the pulse period, as expected from the changing projection on the plane of the sky of the magnetic field direction. With the assumption that the field geometry is roughly dipolar, this position angle swing can be modeled to produce estimates of the angles between the spin axis and magnetic axis, and between the spin axis and the observer's line of sight.

Significant circular polarization ($\gtrsim$10%) is also often observed. Often, the circular polarization changes handedness near the center of the pulse profile. It is not known whether the circular polarization is intrinsic to the pulsar radiation processes or arises as a linearly polarized signal propagates, for example, through a birefringent region in the pulsar magnetosphere.

Radio phenomenology

The observed radio beam of a typical pulsar is quite narrow, with the pulsar 'on' only a few per cent of the time. When averaged over hundreds to thousands of pulse periods, the radio emission pattern is usually very stable. Some pulsars confine their radiation to a single Gaussian beam; others have complex pulse shapes, or show 'interpulses' that most likely arise from the second magnetic pole. A few pulsars have two or more stable emission patterns, and switch, apparently at random, between the two emission 'modes'.

From one pulse to the next, pulsars display a bewildering array of complex behaviors, both random and highly structured. A histogram of pulse strengths is generally roughly exponential in form, with the strongest pulses about 10 times the mean. Some pulsars undergo extended stretches, called 'nulls', with no detectable radio emission at all. Two pulsars, the Crab pulsar and the millisecond pulsar B1937+21, exhibit a power-law tail of 'giant pulses' that extends to hundreds of times the mean pulse intensity.

Individual pulses are often composed of one or more 'subpulses', which are usually significantly narrower than the length of the average pulse window. These subpulses occasionally show highly organized behavior from pulse to pulse, drifting across the pulse window. On much shorter timescales, 'micropulses' with widths as short as a microsecond are seen.

Rotation-powered X-ray and γ-ray pulsars

The coherent radio emission from pulsars is a nearly insignificant fraction of the total luminosity inferred from their spin-down rates. A handful of radio pulsars, primarily the youngest, also have pulsed emission at optical, x-ray and γ-ray energies. The Crab pulsar, for example, has a spectrum that has been measured all the way from 10 MHz radio waves to at least 5 GeV γ rays; the emitted power peaks in the hard x-ray band. The pulsar in the Vela supernova remnant has a spectrum that peaks in the γ rays, where it is the brightest object in the sky. (These rotation-powered high-energy pulsars should not be confused with X-RAY BINARY STARS that are accreting material from a companion.) One pulsar is known, GEMINGA, that appears quite ordinary except that it is seen only at high energies and not in the radio. It is believed that Geminga is a very nearby radio pulsar whose radio beam does not intersect the Earth. As noted above, some models place generation of the high-energy emission from pulsars in an outer gap, rather than the polar cap region where the radio emission probably originates.

Binary and millisecond pulsars

There are about 50 pulsars now known in binary systems. Most of these pulsars are clustered in the lower left of figure 1, with shorter pulse periods, larger characteristic ages, and lower magnetic fields than the general pulsar population. Only a handful of ordinary pulsars have binary companions, certainly a far smaller number than might be expected from studies of BINARY STARS. Although most pulsar progenitors were originally in binaries, most of these binaries were gravitationally unbound by the mass loss that accompanied supernova explosions, leaving the pulsars isolated.

Recycling

Under normal circumstances, the more massive component of a binary star system will evolve most rapidly, resulting, if the star is above about $8M_\odot$, in a supernova explosion. If the binary remains bound, the system may appear as a young pulsar with a main sequence companion. Eventually, the second star will finish core hydrogen burning and move off the main sequence, gradually

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

swelling. If the binary is sufficiently wide, the companion will continue its evolution as if it were an isolated star. However, if the binary is sufficiently compact, the surface of the companion will eventually come into contact with its Roche surface and begin transferring mass (and angular momentum) to the neutron star, spinning it up to shorter rotation periods. During this stage, the system will be visible as a luminous x-ray binary.

The fate of the system depends on the mass of the companion. A massive companion will evolve quickly, exploding as a supernova after a relatively short mass transfer period (as a high-mass x-ray binary). The result will be a pulsar with spin period around 30–50 ms, with a neutron star companion or no companion at all if the system becomes unbound. A light companion will evolve slowly, with a long mass transfer period (as a low-mass x-ray binary), leaving a pulsar with period 1–10 msec and a light, helium white dwarf companion. An intermediate-mass companion will leave a pulsar with intermediate spin period and a massive, carbon–oxygen white dwarf companion.

It is observed that the magnetic fields of these 'recycled' pulsars are significantly smaller than those of young pulsars. (The reason is not well understood, but is probably related to the mass accretion.) The resulting torques are very small, so recycled pulsars can spin down much more slowly than ordinary pulsars and remain active for billions or even tens of billions of years.

Selected binary pulsars

A few binary pulsars of particular interest are described here.

The Hulse–Taylor binary. The first binary pulsar was discovered by RUSSELL HULSE and JOSEPH TAYLOR in 1974, during a survey for new pulsars done at the ARECIBO OBSERVATORY as part of Hulse's PhD thesis work. Follow-up observations showed that the pulsar is in a high-eccentricity ($e = 0.6$), short-period ($P_b = 8$ h) orbit with another star, which is almost certainly a second neutron star. It was immediately realized that the high velocities and strong gravitational fields in this binary make it an extraordinary laboratory for studying fundamental physics, as described below. In 1993, Hulse and Taylor were awarded the Nobel Prize in Physics 'for the discovery of a new type of pulsar, a discovery that has opened up new possibilities for the study of gravitation'.

Two other close, double neutron star binaries are now known. Loss of energy to GRAVITATIONAL RADIATION causes the neutron stars in these binaries to slowly spiral together; on timescales of 10^8–10^9 yr, the neutron stars will collide and coalesce, probably forming a BLACK HOLE and possibly powering a GAMMA-RAY burst.

The eclipsing binaries. In 1988, a pulsar (B1957+20) was discovered with a $0.025M_\odot$ companion. The pulsar was eclipsed by its companion for about 10% of each orbit. The eclipse region is significantly bigger than the companion's ROCHE LOBE, leading to the conclusion that the eclipse is caused by a wind of material streaming off the companion. Although the mass loss rate inferred for this system is fairly small, the excitation of a wind by pulsar radiation has been suggested as a means of ablation or evaporation of a companion. The handful of known millisecond pulsars without binary companions may have emerged from such 'black widow binaries', of which several examples are now known. (See also ECLIPSING BINARY STARS.)

Pulsars with main sequence companions. Two pulsars are known with massive main sequence companion stars. Both are young pulsars that have not yet been recycled during companion evolution. The pulsar B1259−63 is in a very wide (3.4 yr), high-eccentricity orbit around a $10M_\odot$ Be star; J0045−7319 orbits a $\sim 9M_\odot$ B1 V star whose radial velocity curve has been measured with optical observations, making it the only known 'double-line' spectroscopic pulsar binary. Quadrupolar distortions of the companions in these systems lead to classical spin–orbit coupling and hence to measurable perturbations of the pulsar orbits, providing a test of stellar structure models of these massive stars.

A pulsar with planets. PSR B1257+12, a pulsar with a 6 ms spin period, has two planetary companions of roughly terrestrial mass, $3.4M_\oplus$ and $2.8M_\oplus$, orbiting with periods of 66.5 and 98.2 days. A third body, of roughly lunar mass, orbits in a 1 month orbit. This was the first planetary system detected outside the solar system, and these remain the lightest known extrasolar planets. The planets are in nearly circular orbits, suggesting that they were formed after the supernova that created the pulsar. It is possible that they formed from a disk left by a companion star that was destroyed by the ablation processes seen in the eclipsing binary pulsars.

A triple pulsar system. The pulsar PSR B1620−26, in the globular cluster M4, has a $\sim 0.3M_\odot$, white dwarf companion in a low-eccentricity, half-year orbit. An observed acceleration of this binary has been interpreted as the reflex motion of the binary around its center of mass with a second, more distant companion. The outer body is light, probably either a planet or a brown dwarf, and the orbital period is most likely a few centuries.

Isolated millisecond pulsars

There are a few pulsars known that have short spin periods and low magnetic fields that place them with the binary pulsars in figure 1 but which have no observed binary companion. These include the first millisecond pulsar discovered, PSR B1937+21, which is still the fastest known pulsar, with spin period $P = 1.557$ ms. It is generally thought that B1937+21 and similar pulsars were recycled in binary systems but later lost their companions, perhaps to a process like the evaporation that appears to be occurring in the eclipsing binary systems.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pulsars in globular clusters

Over 40 pulsars are now known in the Galactic GLOBULAR CLUSTERS, 20 in 47 Tucanae alone. Most have low magnetic fields and short spin periods and are probably products of recycling in a binary system. About half are now in binary systems. Since globular clusters contain very old stellar populations and no active star formation, no core collapse supernovae are currently producing young pulsars. The apparent presence of a few relatively young pulsars in clusters has led to the suggestion that some neutron stars could be created in 'accretion-induced collapse' of white dwarfs that are pushed over the Chandrasekhar mass limit by accretion from a binary companion.

The stellar density in the core of a globular cluster is sufficiently high that stellar interactions, including star–binary interactions and binary–binary interactions, are an important dynamical process. Thus clusters contain a relatively high fraction of isolated millisecond pulsars, whose companions have been unbound in three or four body interactions. (See also BINARY STARS IN GLOBULAR CLUSTERS.)

Pulsar timing

Because of their great rotational inertia, the rotation of pulsars is extremely stable. The individual pulses from pulsars are the ticks of celestial clocks that are, in the best cases, as precise as the best terrestrial atomic time standards (long-term frequency stability of about a part in 10^{14}). Measuring the differential light travel time of the pulsar signal as the Earth moves around its orbit allows, for example, the position of the pulsar in ECLIPTIC coordinates to be precisely determined. Similarly, the pulsar signal will be periodically advanced or delayed as the pulsar travels in a binary orbit. Pulsar timing experiments—clock comparison experiments that compare the pulse arrival time with a reference atomic time standard—can thus be used to make very precise estimates of the orbital parameters of binary pulsars. (For further details, see also RADIO PULSAR BACKENDS and SOFTWARE: PULSAR SOFTWARE.)

Tests of general relativity

The most famous application of pulsar timing techniques has been to tests of experimental gravitation. In most cases, binary orbits are well approximated as Keplerian ellipses. The high velocities ($\sim 10^{-3}c$) and strong gravitational fields in some binary pulsar systems cause relativistic deviations from Keplerian motion to be significant. Five relativistic corrections have been measured: the advance of the angle of periastron of the elliptical orbit (as is seen in the orbit of Mercury); the combined effect of the transverse Doppler shift and the changing gravitational redshift as the eccentric orbit carries the pulsars closer and further from its companion; two parameters describing the Shapiro time delay of the pulsar signal as it propagates through the gravitational potential well of the companion; the decay of the binary orbit due to gravitational radiation back reaction. The measurement of any two of these effects allows the amplitude of the other three to be predicted, making possible very precise tests of general relativity and alternative gravity theories. This was the subject of the 1993 Nobel Prize in Physics, discussed above. (See also GENERAL RELATIVITY TESTS: BINARY PULSARS.)

Pulsars and nuclear physics

The average density of a neutron star is comparable with the density in heavy atomic nuclei, and the density reaches several times this value at the core. Neutron stars are thus useful laboratories for studying nuclear materials at densities beyond what can be reached in the laboratory. One topic of considerable interest is the maximum neutron star mass, above which the star becomes unstable to collapse to a black hole. Pulsar timing measurements allow, in some cases, very accurate stellar mass estimates. The most massive known neutron star in a radio pulsar binary is $1.44M_{\odot}$, considerably below the $\sim 2M_{\odot}$ limit of a typical nuclear equation of state.

A more direct probe of neutron star interiors comes from the study of timing 'glitches', that are observed in some young pulsars. During a glitch event, the pulsar period decreases in a step function, with amplitude from a few parts in 10^9 to a part in 10^6. At the same time, the spin-down rate increases, and the pulsar recovers over a time period of weeks to months to a period close to the pre-glitch value. Smaller glitches may be caused by a sudden cracking of the crust of the neutron star and its relaxation to a more spherical shape; larger glitches are probably due to variations in the coupling between the superfluid interior of the star and its independently rotating crust.

Bibliography

The most recent introductory textbook is

Lyne A G and Smith F G 1998 *Pulsar Astronomy* 2nd edn (Cambridge: Cambridge University Press)

A more advanced, classic textbook, now dated, is

Manchester R N and Taylor J H 1977 *Pulsars* (San Francisco, CA: Freeman)

A collection of more advanced review articles on a variety of topics have been published in book form as

Blandford R D, Hewish A, Lyne A G and Mestel L (eds) 1993 *Pulsars as Physics Laboratories* (Oxford: Oxford University Press)

A good introduction to the problem of the pulsar emission mechanism is

Michel F C 1991 *Theory of Neutron Star Magnetospheres* (Chicago: University of Chicago Press)

An annotated guide to the pulsar literature has been compiled:

Weisberg J M 1993 *Am. J. Phys.* **61** 13

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This bibliography and more than two dozen papers on a broad range of topics of historical or current interest have been collected as

Weisberg J M (ed) 1994 *Radio Pulsars: Selected Reprints* (College Park: American Association of Physics Teachers)

Stephen E Thorsett

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pulsating and Chemically Peculiar Upper Main Sequence Stars

There is a natural astrophysical division of the main sequence which occurs for early F stars. Stars cooler than this have deep surface convection zones, radiative cores, generate their energy by the proton–proton cycle, rotate relatively slowly and any global magnetic field present is weak, although strong fields may exist in tangled flux tubes, as in sunspots in the Sun. Stars hotter than early F have thin or no surface convection zones, convective cores, generate their energy by the CNO cycle, rotate relatively rapidly and, if they have magnetic fields, those fields are strong, global and approximately dipolar. Given this natural division this article concerns the stars of the Upper Main Sequence, from early F to the O stars. A guide through the names of most of the types and classes for these stars is provided.

Mostly, the peculiar stars do not pulsate, and the pulsating stars (see STELLAR PULSATION) are not peculiar. However, there is a small overlap; the relationship between pulsation and peculiarity is discussed, and the most widely accepted model, the diffusion hypothesis, is introduced and explained, with examples given of its wide applicability to many fields of stellar astrophysics.

Nomenclature for the peculiar stars

On and near the main sequence for $T_{\rm eff} > 6600$ K there is a plethora of spectrally peculiar stars and photometric VARIABLE STARS with a bewildering confusion of names. Table 1 shows some of the subgroups of chemically peculiar (CP) stars as a function of temperature. These stars all show spectral peculiarities. Members of the magnetic group are known to have global magnetic fields which are roughly dipolar with strengths of hundreds to tens of thousands of G. (The global magnetic field of the Sun is about 0.5 G; the magnetic field in sunspots is about 1500 G.) The magnetic stars are mostly known as Ap stars (for A peculiar), but many of the 'Ap' stars are B stars, so sometimes those are called Bp stars. Other A stars which have peculiar spectra, but not the same peculiarities as the Ap stars, and are non-magnetic, are called Am stars (for A metallic-lined).

Table 1. Magnetic and non-magnetic peculiar stars of the upper main sequence.

$T_{\rm eff}$ (K)	Magnetic stars	Non-magnetic stars
7000–10 000	Ap SrCrEu	Am, λ Boo
	A3 – F0	A0 – F1
10 000–14 000	Ap Si	Ap HgMn
	B8–A2	B6–B9
13 000–18 000	He-weak Si, SrTi	He-weak PGa
	B3–B7	B4–B5
18 000–22 000	He-strong	
	B1–B2	

Am stars are given three spectral classifications: one based on the Balmer lines which gives a good measure of the effective temperature; one based on the Ca II K-line which, because of its weakness relative to normal stars, gives an earlier spectral type; and one based on the metal lines which, because of the enhanced strength of the metal lines, gives a later spectral type. 'Classical Am' stars have K-line and metal line spectral types that differ by five or more spectral subclasses; this difference is less than five subtypes in 'marginal Am' stars (designated 'Am:'). Classical Am stars have H-line types that lie between A3 and F1. For hotter stars the Am phenomenon is harder to recognize, but continues to A0 (Sirius is an Am star), so Am stars with H-line spectral types between A0 and A3 are known as 'hot Am' stars. Evolved Am stars with luminosity classes IV and III are classified as δ Del stars, or as ρ Pup stars (see also CLASSIFICATION OF STELLAR SPECTRA).

To rationalize this situation another system of nomenclature was devised in which there were four subtypes of chemically peculiar (CP) stars: CP1 ≡ Am stars, CP2 ≡ magnetic Ap, Bp stars, CP3 ≡ HgMn stars, and CP4 ≡ He-weak B stars. This system is widely used—principally, but not exclusively, in Europe—whereas North Americans seem to prefer the Ap–Bp terminology.

There are also the λ Bootis stars which have H-line types between A0 and F0, a Ca II K-line type of A0, or slightly later, and *weak* metallic lines, particularly Mg II 4481 Å. Abundance analyses show marked *under*-abundances of the Fe-peak elements, with lighter elements essentially normal.

The oblique rotator model

The magnetic field strengths, spectra and luminosities of the magnetic CP stars all vary synchronously on timescales of about a day to over a century. These variations are well-understood in terms of the oblique rotator model. Figure 1 shows the geometry of this model. The line-of-sight to the Earth is to the right. The inclination of the rotation pole is i, the obliquity of the magnetic axis is β, and the variable angle between the magnetic pole and the line-of-sight is α. As the star rotates, the magnetic field is seen from varying aspect, so its effective strength is observed to be variable. Because the stars have anomalous abundance patches at their magnetic poles, the spectrum also varies with rotation as does the luminosity—both of these in phase with the magnetic variations. These rotational variables are known as α^2 CVn stars.

For a centered dipolar magnetic field the measured effective magnetic field strength is:

$$H_{\rm eff} = \frac{1}{20}\frac{15+\mu}{3-\mu} H_{\rm p}(\cos i \cos\beta + \sin i \sin\beta \cos \Omega t)$$

where $H_{\rm p}$ is the polar magnetic field strength, i and β are the rotational inclination and magnetic obliquity, μ is the limb-darkening coefficient and Ω is the rotation frequency. It can be seen from this equation that dipolar magnetic fields give rise to magnetic variations which are sinusoidal, with amplitudes proportional to $\sin i \sin\beta$ and zero points proportional to $\cos i \cos\beta$; this immediately constrains the values of i and β.

© Astronomical Society of the Pacific 2000

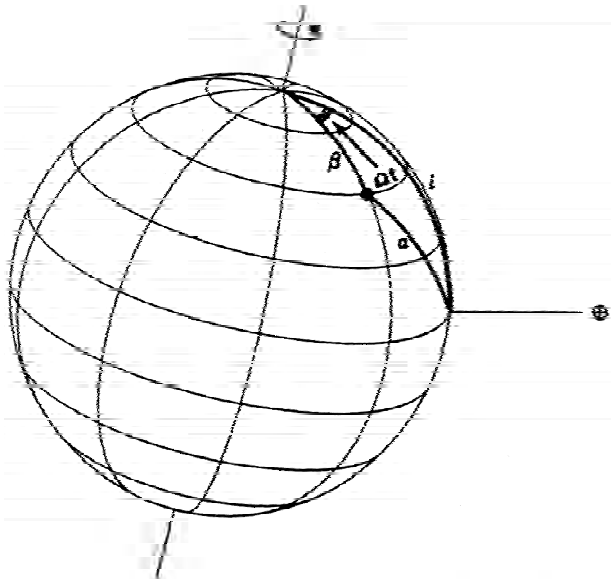


Figure 1. The geometry of the oblique rotator model. Earth is in the direction marked.

Nomenclature for the variable stars

There are many kinds of pulsating stars near the upper main sequence:

- The δ Sct stars lie in the instability strip, which ranges from A2 to F0 on the main sequence and from A3 to F5 at luminosity class III. They pulsate primarily in low-overtone radial and non-radial p modes with periods between about 30 min and 6 h, and amplitudes ranging from the detection limit of a few mmag up to 0.8 mag. Some δ Sct stars may also pulsate in g modes.
- The high-amplitude δ Scuti stars are known as HADS and as dwarf Cepheids if they are POPULATION I STARS. Low metallicity, POPULATION II subdwarfs with δ Scuti pulsation are called SX Phe stars.
- The rapidly oscillating Ap (roAp) stars are mid-A to early-F main sequence Ap SrCrEu (CP2) stars which pulsate in high-overtone p modes with periods in the range 5–16 min and amplitudes $\leq$0.016 mag—generally much less. The pulsation modes are aligned with the oblique magnetic axis, so the geometry of the oblique rotator model in figure 1 applies to the pulsation of these stars in the similar oblique pulsator model. They mostly lie within the δ Sct instability strip, but a few of them are cooler. The driving and mode selection mechanisms are not known. Asteroseismic study of the pulsation frequencies of these stars provides information about their pulsation modes, rotational inclination, the inclination of the magnetic field to the rotation axis, atmospheric structure, luminosity, internal structure and internal magnetic field strength and configuration. Asteroseismically, they are the most studied main sequence stars other than the Sun.
- The Herbig Ae and Be stars, known collectively as HAeBe stars, are PRE-MAIN SEQUENCE A and B STARS which show emission lines from strong stellar winds and from cocoons of remnant gas from which they collapsed. These are massive counterparts of the T TAURI STARS. Two Herbig Ae stars are now known to be δ Sct stars.
- The γ Dor stars are multi-periodic, non-radial g-mode main sequence pulsators with periods in the range 0.3–3 days, and amplitudes of a few hundredths of a magnitude. They are confined to an instability strip which partially overlaps the δ Sct stars in the HERTZSPRUNG–RUSSELL (HR) DIAGRAM, ranging in temperature from 7200–7700 K on the main sequence, and 6900–7500 for $\log g \approx 4$.
- The slowly pulsating B (SPB) stars (also known as 53 Per stars) are multi-periodic, non-radial g-mode main sequence pulsators with periods in the range 0.6–3 days and amplitudes typically less than a few hundredths of a magnitude. They lie in a narrow instability box in the HR diagram which ranges from B2 to B9, thus they do not overlap with the δ Sct stars.
- The β Cephei stars are giant and sub-giant, p-mode pulsators with periods in the range 2–7 h. Some are singly periodic, some are multi-periodic; they pulsate in both radial and non-radial modes of low degree, $\ell \leq 3$, and low overtone, n. Their photometric amplitudes are $\leq$0.1 mag, except for the star BW Vul.
- Even hotter than the β Cep stars there are (possibly) pulsating O stars known as α Cygni variables. Their periods are typically of the order of 1 to 2 months, and their amplitudes are less than a few tenths of a magnitude. They are thought to be pulsating in g modes, or in so-called 'strange modes'—modes associated with a sound-speed inversion, caused by a density inversion, caused by an opacity bump, most likely from Fe, H and/or He.
- Amongst the B stars there are emission-line Be stars which have periodic variations with periods between 0.3 and 3 days and amplitudes between 0.01 and 0.3 mag. These are known as λ Eri stars. It is contentious as to whether their variability is caused by g-mode pulsation or by some rotational effect.

Non-pulsating variables near the upper main sequence include:

- Eruptive LUMINOUS BLUE VARIABLES (LBVs) called S Doradus stars with variations of about 0.2 mag on time-scales of weeks to months. If there is thought to be periodicity in the variations, then these are called α Cyg stars as discussed above. LBVs also have eruptions of 0.5–2 mag on a time-scale of years to decades, and giant eruptions on the time-scale of a millennium, η Car being the best known, most spectacular example of this.

© Astronomical Society of the Pacific 2000

- WOLF–RAYET STARS which show non-periodic variations with amplitudes of the order of 0.02 mag caused by variable winds and possibly by rotation.
- Non-periodic, long-time-scale light variable Oe and Be stars known as γ Cas stars which are in addition to the periodic Be stars (λ Eri stars) mentioned previously. They range from spectral type O6 to B9 and through luminosity classes V to III. Their light variations are caused by variable winds.
- α^2 CVn stars which are magnetic stars with spots associated with their magnetic poles. These spots are seen from varying aspect as the stars rotate, causing photometric variability with the rotation period. The amplitudes are typically a few hundredths up to (rarely) a few tenths of a magnitude.

In general, there is an exclusion between the CP stars and the pulsating δ Sct stars. But there are some CP stars that do pulsate! The diffusion hypothesis is the favored explanation for this. It also has important implications in many other fields of stellar astrophysics.

The diffusion hypothesis

There is no universally accepted model that explains all of the observed characteristics of the CP stars. Models involving anomalous atmospheres, nucleosynthesis and dredging, surface spallation, magnetic accretion, planetesimal impacts, binary mass transfer and element separation (diffusion) have all been tried with varying success. The most successful model, the *diffusion hypothesis*, is the working model of choice for the majority of investigators.

The idea is simple in principle: if there are layers in a star which are stable against turbulent mixing, then elements heavier than H will tend to sink gravitationally, unless they have many absorption lines near the local flux maximum, in which case the asymmetry in the flux arising from the temperature gradient will mean that ions absorb more radiation from below than above, hence are radiatively driven towards the surface. These two competing effects can cause some elements to sink and other to rise, thus producing peculiar atmospheric abundances.

In the CP stars the diffusion hypothesis accounts for:

- Overabundances of the Fe peak, rare earth and lanthanide elements: these elements, which are principally in their neutral and first-ionized states, have many absorption lines near the flux maximum, hence they are driven upwards to the surface layer.
- Underabundances of Ca, Sc, C, He in Am stars: the principal atmospheric ionization states of Ca and Sc are Ca III and Sc IV which leave each of these elements, numbers 20 and 21, with 18 electrons so that they are in closed electronic shell configurations similar to Ar. With tightly bound electrons the absorption lines have high excitation potentials, hence lie in the UV where there is insufficient flux to levitate these ions. Thus they sink and are deficient.
- Isotopic ratios of Hg in HgMn stars: there are seven observable isotopes of Hg in the HgMn stars—196, 198, 199, 200, 201, 202 and 204—which show drastically different abundances from the terrestrial mixture, and which show large variation from star to star. Radiative levitation preferentially lifts the less abundant isotopes to the surface. Because their lines are less saturated, they absorb more flux per ion than the more abundant isotopes, and thus the isotopic ratios are altered. This same effect is seen in the He^3/He^4 ratio embedded in Moon rocks from the solar wind. He^3 is preferentially driven off the Sun in the solar wind because its lines are less saturated than the more abundant He^4.
- Ages of CP stars: calculations indicate that diffusion can produce observable anomalies in 10^6 years, the age of the youngest Am stars.
- Disappearance of anomalies in RED GIANTS with CP precursors: all models of CP stars can account for this, because it is known that the abundance anomalies must be confined to a thin atmospheric layer. One strong demonstration of this is the overabundance of some rare earths and lanthanides in Ap stars by a factor of 10^4–10^5. More than one in a hundred thousand stars is a member of this group. Hence, if the abundance anomalies were global, the Ap stars would contain almost all the rare earths and lanthanides in the universe, a clear absurdity. In all models, the growing depth of the surface convection layer with age, as stars move to the right in the HR diagram towards the zone of completely convective stars, leads to mixing away of the surface anomalies.
- Slow rotation in Am and Ap stars: A-type stars may show equatorial rotational velocities up to $v_{\rm eq}\sin i \approx$ 250–300 km s^{-1}, whereas Am stars have $v_{\rm eq}\sin i \leq$ 125 km s^{-1} and Ap stars have $v_{\rm eq}\sin i \leq 100$ km s^{-1}, although for the latter $v_{\rm eq}\sin i \leq 10$ km s^{-1} is not unusual. Rotation generates meridional circulation currents which flow faster with higher $v_{\rm eq}\sin i$. At some value of $v_{\rm eq}\sin i$ they become turbulent; that value is calculated to be about 50–100 km s^{-1}, in good agreement with the observations. The resultant turbulent mixing inhibits diffusion.
- Binary nature of Am stars: nearly 100% of Am stars are BINARY STARS with $1 \leq P_{\rm orbital} \leq 10$ days. Tidal synchronism locks the rotational and orbital periods and slows the rotation. A 2-$R_\odot$ star rotating with a period of 1 day has $v_{\rm eq}\sin i \approx 100$ km s^{-1}. Binary stars with $P_{\rm orbital} < 1$ day rotate too quickly; those with $P_{\rm orbital} > 10$ days do not become rotationally synchronized, thus tending to rotate too quickly.
- Magnetic braking of Ap stars: The very low rotational velocities of the Ap stars are considered to be the result of magnetic braking.
- Concentration of elements in spots near the magnetic poles in Ap stars: Ap stars have 'spots' or 'rings' of overabundant elements near their magnetic poles. Diffusion is thought to concentrate certain elements

© Astronomical Society of the Pacific 2000

in these regions, perhaps because quadrupole magnetic field components may produce magnetic loops arranged in rings which halt radiative levitation and trap rising elements.

- Near-exclusion of CP and δ Sct stars: about 30% of the stars in the lower instability strip are δ Sct stars. Most of the non-pulsating stars in this part of the instability strip are Am and Ap stars. Only a few Am and Ap stars pulsate, and it is a strength of diffusion theory that it can explain this: in stable A- and early F-star atmospheres helium settles gravitationally. This shuts off the κ-mechanism operating in the He II ionization zone which drives δ Sct pulsation.

There are problems for the diffusion hypothesis in CP stars:

- Diffusion is at best semi-quantitative. Calculated diffusion velocities are very low—only 10^{-4} to 1 cm s^{-1}. Hence stars must be stable to turbulence at that level in the diffusive layers.
- Pulsation in peculiar stars which do pulsate must be laminar; i.e. it must not mix away the anomalies. In a few cases this demands no turbulence with pulsation velocities of the order of 10 km s^{-1}.
- The odd–even effect in abundances is apparent in Ap and Am stars. Opponents of the diffusion hypothesis say this argues for a partial role for nucleosynthesis. Proponents say diffusion does not erase completely the initial abundance patterns.

Thus, while the diffusion hypothesis is the most popular working model for the CP stars, it must be remembered that it has some weaknesses which are yet to be dealt with. That diffusion acts in CP stars is reasonably accepted; whether it is sufficient to explain all the peculiarities, or whether other mechanisms are working as well is still to be determined.

In other stars diffusion can also be important:

- WHITE DWARF STARS (see also DB PULSATING STARS) are layered by gravitational settling.
- The gravitational settling of He is an important part of the 'Standard Solar Model' changing the internal solar temperature gradient enough to decrease the calculated neutrino production rate by 7%. Helium and some metal settling have also been included in calculations of Li and Be mixing in the Sun.
- Gravitational settling of He in older stars converts gravitational potential energy to thermal energy, increasing core temperatures and nuclear reaction rates, hence shortening stellar lifetimes. This could possibly reduce GLOBULAR CLUSTER ages by 1–4 Gyr, a significant amount in the on-going cosmological conflict of low Hubble time, H_0^{-1}, and high stellar age from globular cluster evolutionary tracks.
- Gravitational settling also partially accounts for the Li-gap in F stars.

Thus the CP stars of the upper main sequence are the best laboratory in which to test diffusion theory which has many applications is stellar astronomy. The pulsating stars are being mined for rich asteroseismic information about stellar structure, and the small group of stars that are both peculiar and pulsate are providing particularly stringent tests of theory, plus a unique look at the interaction of stellar pulsation and strong magnetic fields.

Bibliography

For an introduction to the CP stars see

Wolff 1983 The A-type stars: problems and perspectives (Monograph series on nonthermal phenomena in stellar atmospheres) *NASA* 463 Special Publication

For more recent discussions see

Dworetsky M M, Castelli F and Faraggiania R (eds) 1993 *Peculiar Versus Normal Phenomena in A-type and Related Stars (ASP Conf. Ser. 44)* (San Fransico: Astronomical Society of the Pacific)

Asteroseismology is an exciting, rapidly developing field: see

Deubner F, Christensen-Dalsgaard J and Kurtz D W (eds) 1998 *New Eyes to See Inside the Sun and Stars: Pushing the Limits of Helio- and Asteroseismology with New Observations from Earth and Space (Proc. IAU Symp. 185)* (Dordrecht: Kluwer)

For a very useful introduction to the many kinds of variable stars see

Sterken C and Jaschek C 1996 *The Light Curves of Variable Stars* (Cambridge: Cambridge University Press)

D W Kurtz

© Astronomical Society of the Pacific 2000

Puppis

(the Poop or Stern; abbrev. Pup, gen. Puppis; area 673 sq. deg.) A southern constellation which lies between Canis Major and Vela, and culminates at midnight in early January. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, from stars that formed part of the ancient constellation of Argo Navis (the Ship), which had been included by Ptolemy (*c*. AD 100–175) in the *Almagest*.

The brightest stars in Puppis are ζ Puppis (Naos), magnitude 2.2, π Puppis, magnitude 2.7, and τ Puppis, magnitude 2.9. There are seven other stars brighter than fourth magnitude, including σ Puppis, a double star with orange (K5) and yellow (G5) components, magnitudes 3.3 and 8.6, separation 22.3″, the former of which has an unseen companion which revolves around it in about 258 days. An interesting binary star is k Puppis (HR 2948–2949), which has two bluish-white (B5) components, magnitudes 4.5 and 4.7, separation 9.9″. Other interesting stars include L^2 Puppis, a semiregular variable (range 2.6–6.2, period about 140 days), and V Puppis, an eclipsing binary (range 4.4–4.9, period 1.45 days).

The Milky Way passes through Puppis and the constellation contains a number of open clusters, the brightest of which are M46 (NGC 2437), which has more than 150 stars fainter than eighth magnitude, M47 (NGC 2422), which has a few dozen stars, the brightest of magnitude 5.7, M93 (NGC 2447), which consists of about 80 stars fainter than eighth magnitude, and NGC 2477, which has about 300 stars fainter than ninth magnitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Purcell, Edward Mills (1912–)

Physicist, born in Taylorville, IL, Nobel prizewinner (with Bloch 1952) for his studies of the nuclear magnetic moment of helium. With H I Ewen in 1951, he formed one of the three groups that independently detected cosmic 21 cm radiation from the spin transition of the electron in neutral hydrogen, predicted by H VAN DE HULST. The detection was soon confirmed in the Netherlands by C A Muller and J OORT and in Australia by W CHRISTIANSEN and J V Hindman.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Purple Mountain Observatory

The Purple Mountain Observatory was built in 1934. It is situated on the Purple mountain in the eastern suburb of Nanjing, China.

At present it is engaged in the research of astrophysics, celestial mechanics, radio astronomy, space astronomy and practical astronomy. The programs consist of solar physics, stellar physics, planetary science, high-energy astrophysics, orbital theory of artificial satellites, etc.

The observatory has a staff of about 300 people, and has three observing stations outside Nanjing. A 13.7 m millimeter wave telescope was set up at the Qinghai station.

For further information see
http://www.pmo.ac.cn/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pushchino Radio Astronomy Observatory

The Pushchino Radio Astronomy Observatory is 120 km south of Moscow. The staff consists of 45 astronomers and 42 engineers, supported by the Russian Academy of Sciences, Ministry of Sciences and Russian Foundation for Basic Research.

The most sensitive meter-wavelength radiotelescopes BSA and DKR-1000, as well as the 22 m dish of the Observatory are used for studies of galaxies, pulsars, interstellar and interplanetary plasma.

Discoveries and studies of Geminga's and Magnetar's (SGR1900+14) pulse radio emission are among the achivements of the last two years.

For further information see
http://www.prao.psn.ru.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pythagoras of Samos (*c*. 580–*c*. 500 BC)

Mathematician, born in Samos, Ionia, taught by Thales and Anaximander on Miletus, founded a philosophical and religious school in Croton (in southern Italy) whose members, as well as having various beliefs that we would recognize as religious, believed that at its deepest level, reality is mathematical in nature. Pythagoras observed that vibrating strings produce harmonious tones when the ratios of the lengths of the strings are whole numbers, and proved *Pythagoras's theorem*, and other geometrical theorems. In astronomy Pythagoras taught that the Earth was a sphere at the centre of the universe, and is credited as the first to realize that Venus as an evening star (Hesperus) was the same planet as Venus as a morning star (Phosphorus).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pyxis

(the Compass; abbrev. Pyx, gen. Pyxidis; area 221 sq. deg.) A southern constellation which lies between Puppis and Antlia, and culminates at midnight in early February. It was introduced as Pyxis Nautica (the Mariner's Compass) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, from stars that formed the mast of the ancient constellation of Argo Navis (the Ship), but its name fell into disuse until resurrected in its shortened form by the American astronomer Benjamin A Gould (1824–96) in 1879. The shortened form was adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest stars in Pyxis are α Pyxidis, magnitude 3.7, β Pyxidis, magnitude 4.0, and γ Pyxidis, also magnitude 4.0. Interesting objects include T Pyxidis, a recurrent nova (range 7.0–15.8), which is normally about fourteenth magnitude but flared up to seventh to eighth magnitude in 1890, 1902, 1920, 1944 and 1966–7, and NGC 2818, an open cluster of about 40 stars fainter than eleventh magnitude which contains a twelfth-magnitude planetary nebula (PK261 + 08.1).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Qadi Zada, al'Rumi (1364–1436)

Born in Bursa, Turkey, professor of mathematics at Samarkand and taught ULUGH BEG. Became director of the Samarkand Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quadrantids

A meteor shower that takes place in early January. It is one of the most prolific, with a peak zenithal hourly rate of 120 in a short, sharp maximum of about 12 hours. The brightest meteors show a blue or yellow-green tinge. The radiant lies in the constellation Boötes, near its border with Hercules. The shower is named after the obsolete constellation Quadrans Muralis (the Mural Quadrant) which used to occupy this region. Perturbations of the meteor stream by Jupiter cause it to oscillate up and down through the ecliptic; the current phase of Quadrantid activity began in the 1700s and will last until the 2200s.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quadrature

The position of a planet or the Moon when the angle it makes with the Sun (the difference between the two bodies' celestial longitudes) is 90°.

See: elongation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quantum Gravity

GENERAL RELATIVITY and QUANTUM MECHANICS are among the greatest intellectual achievements of the twentieth century. Quantum gravity is the theory that is being sought to unify them.

Each of the two theories has profoundly altered our understanding of the physical world. Furthermore, each has been successful in describing the physical phenomena in its own domain to an astonishing degree of accuracy. And yet they offer strikingly different pictures of physical reality. The world of general relativity is geometric, marked by precision and certainty. The world of quantum mechanics is fuzzy, marked by chance and uncertainty. Physicists can happily maintain a schizophrenic attitude and switch between the two because the natural domains of the two theories are distinct: while general relativity governs astrophysical and cosmological phenomena, quantum mechanics governs atomic and subatomic ones. But there do exist situations which straddle the two, in particular near the BIG BANG and the final stages of gravitational collapse. More importantly, everything in our past experience with physics tells us that such strikingly different descriptions of nature cannot both be fundamental. Since neither encompasses the other, they must both be special cases of a single, grander theory.

The search for this grander theory has therefore become a holy grail of theoretical physics. Indeed, some of the most distinguished personalities of this century have devoted many years of work to it. These include Peter Bergmann, Paul Dirac, Richard Feynman, Stephen Hawking, Roger Penrose, John Wheeler and Edward Witten, to mention just a few. The search for the theory began in the 1930s in the writings of Rosenfeld. However, the area remained dormant till the late 1950s. From then on, there has been a steady increase in activity. Over the last 10 years, this activity has intensified many fold. In particular, striking advances have occurred within two approaches. While the final theory still eludes us, many intriguing and suggestive results have been obtained which have provided glimpses of the desired theory. Therefore, there is now a heightened sense of optimism, enthusiasm and rapid progress in the field.

Why has the problem been so difficult? The most important obstacle has been the complete lack of experimental data with direct bearing on quantum gravity. The situation is thus very different from the evolution of quantum mechanics where the theory could be guided at critical junctures by masses of accumulated data from spectroscopy, radioactivity and solid state physics. But there also are, in addition, deep conceptual difficulties. Einstein taught us that gravity is a manifestation of curvature of SPACE-TIME. Unlike other forces of nature, gravity is geometry. A quantum theory of gravity is thus simultaneously a quantum theory of space-time structure. Therefore, in quantum gravity, geometry itself has to be subject to uncertainties, probabilities and fuzziness. Suddenly, the familiar arena of physics is no more and a host of fundamental questions arise. What happens to causality? To geodesics along which particles and lights rays travel? To differential equations obeyed by the electromagnetic waves? In short, how does one do physics without the familiar space and time?

Key issues

Although these difficulties seem enormous, we also have a number of hints. The first two come from the intrinsic limitations of the two theories themselves.

In the mid-1960s, Penrose, Hawking and others showed that Einstein's equations of general relativity themselves signal the limitation of the theory. More precisely, even if one begins with physically reasonable initial conditions, the resulting solutions to Einstein's equations often lead to singularities, regions in space-time where the gravitational field and other physical quantities become infinite. The theory predicts 'non-sense'. This is typically what happens when a theory is pushed beyond its domain of applicability. In the case under consideration, as the curvature of space-time increases, quantum effects which are ignored in the equations of general relativity are expected to become more and more important and make physical quantities finite. The challenge is to show that this does indeed happen.

Quantum field theory—the theory that results by unifying quantum mechanics with principles of SPECIAL RELATIVITY—also has its own limitations. Calculations of physical quantities often lead to infinities. These arise because one has to allow virtual processes of arbitrarily high energies, or equivalently, because the calculations of physical quantities involve integration over arbitrarily small distances. Thus, their origin can be traced back to the assumption that the continuum picture holds at arbitrarily small scales. From the very early days, it was conjectured by Planck, Heisenberg and others that the continuum hypothesis probably fails in the real world due to quantum gravity effects. This is because with the three fundamental constants of nature that quantum gravity would have—Newton's gravitational constant G, speed of light c and the Planck constant $\hbar$—one can construct a unique combination of length, $\ell_{\rm P} = \sqrt{G\hbar/c^3}$ whose numerical value is approximately 10^{-33} cm. This was noted by Planck in his seminal paper on quantum mechanics and $\ell_{\rm P}$ is now called the Planck length. The existence of a natural scale in a theory signals a transition; physics below that scale is generally quite different from that above the scale. It is widely believed that the familiar picture of space-time breaks down below Planck length. The quantum geometry effects become important. If they are ignored and the classical continuum approximation is pushed beyond the domain of its validity, one obtains infinities, i.e. non-sense. A satisfactory quantum gravity theory of gravity should be free of such infinities.

The clues provided by these limitations of the two theories have been valuable because they provide goals and general directions in the search of quantum

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

gravity. More concrete hints came from rather unexpected quarters in the early 1970s, namely the physics of BLACK HOLES. Through ingenious thought experiments, Jacob Bekenstein first pointed out that black holes can be assigned thermodynamic properties. This idea became more concrete in the work of James Bardeen, Brandon Carter and Stephen Hawking. They introduced a notion of 'surface gravity' κ of black holes, which is analogous to g on the surface of the Earth, and showed that within classical general relativity, black holes in equilibrium are subject to laws which are completely parallel to the standard laws of thermodynamics; a multiple of κ plays the role of temperature and a related multiple of the area A of the black hole horizon plays the role of entropy (see table).

Black hole mechanics	Thermodynamics
κ constant	T is constant
$\delta M = \dfrac{\kappa \delta A}{8\pi G}$	$\delta E = T \delta S$
$\delta A \geq 0$	$\delta S \geq 0$

However, within classical general relativity this could only be a mathematical analogy because, since black holes can only absorb energy, they must have zero physical temperature. This view changed dramatically with Hawking's discovery that, if quantum effects are included, black holes do in fact radiate due to the tunneling phenomenon and, furthermore, they do so exactly as if they were black bodies at temperature $T_{\rm bh} = \hbar\kappa/2\pi c$. The form of the first law of black hole mechanics then suggested that black holes carry an entropy $S_{\rm bh} = Ac^3/4G\hbar = A/\ell_{\rm p}^2$. Note that, as $\hbar$ goes to zero, $T_{\rm bh}$ tends to zero and $S_{\rm bh}$ tends to infinity as expected within classical general relativity. More importantly, subsequent and more detailed thought experiments by William Unruh, Robert Wald and others confirmed the precise expressions of $T_{\rm bh}$ and $S_{\rm bh}$.

These considerations led to concrete questions for potential quantum theories of gravity because Hawking's calculations included quantum matter but treated gravity classically. Indeed, the considerations outlined above contain an *ad hoc* mixture of classical and semiclassical ideas, reminiscent of the Bohr model of the hydrogen atom. However, because the final picture seems compelling and naturally brings together the three pillars of fundamental physics—general relativity, quantum mechanics and statistical mechanics—the general consensus is that the results must continue to hold for large black holes also in full quantum gravity. The challenge is to account for black hole thermodynamics from first principles.

Recent developments

The problem of quantum gravity has been approached from several directions. For examples, methods from quantum field theory which have been highly successful in the description of other forces of nature were adapted to general relativity by Feynman, Bryce De Witt and others. Unfortunately, the resulting theory has uncontrollable infinities. To cure these, theories with SUPERSYMMETRY—a new type of symmetry between bosons and fermions—were extended to include gravity. While some of the infinities were tamed by this process unfortunately, others continued to persist. Another avenue, developed by Hawking, James Hartle, Garry Gibbons and others, extends methods of quantum field theory in which the basic Green functions are first constructed in a four dimensional Euclidean space and the results are then used to calculate physical quantities of interest in the real, Lorentzian world. This approach was used especially in the cosmological context and also provided a unified description of black hole thermodynamics in the semiclassical regime. A completely new line of attack was adopted by Penrose and collaborators who developed twistor theory. Here, a four-dimensional complex space is regarded as fundamental and space-time emerges as a secondary, derived concept.

While these and other approaches have provided a wealth of insights, they are yet to address the key questions discussed above in a concrete fashion. However, over the past decade, two approaches have been successful in addressing some of these questions and have provided concrete directions to address the remaining ones. I will conclude with a brief summary of this work.

The first approach is STRING THEORY. Here, the basic objects are extended, rather than point-like. Historically, point particles were first replaced by strings. However, recent work has indicated that not only does one have to include higher dimensional objects—membranes—but, in a certain well-defined sense, all these objects are equally important. The fundamental theory is formulated in nine space dimensions. The three-dimensional space of our everyday experience arises because six of the dimensions are tiny and manifest themselves only at extremely small distances or in processes involving very high energies.

At present, string theory is still formulated on a background space-time geometry and the methods used are natural extensions of those developed in quantum field theory. However, because of the use of extended objects, there are fundamental differences from local quantum field theories such as supergravity. In particular, the so-called 'ultraviolet' infinities which arise in local field theories because of the integration over arbitrarily high energies are all absent. Even though the theory is not entirely free of divergences, taming of these ultraviolet infinities is a major triumph. Furthermore, there are situations in which the theory has been able to demonstrate how singularities of general relativity could be cured. More precisely, there exist space-times which are singular in general relativity but non-singular when embedded in string theory. In these situations, none of the physical quantities diverge in the 'fundamental theory' and infinities of general relativity can be seen as artifacts of the approximations in which one ignores the deeper structure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perhaps the most exciting contribution of string theory to the basic issues of quantum gravity lies in the domain of black hole thermodynamics. Specifically, for large black holes with electric charge near the maximum allowed value, there is a procedure to count the number of string microstates that can constitute the black hole. The logarithm of this number turns out to give precisely the Hawking–Bekenstein entropy of the black hole. Furthermore, when interactions between string states are included, one obtains the Hawking spectrum of black hole radiation.

The second approach that is addressing the key questions discussed above is based on a QUANTUM THEORY of geometry. Here, one takes the physical role of geometry as fundamental and begins by probing its quantum nature. In contrast to string theory, one works with three-dimensional space and the approach is manifestly background independent. Thus, to begin with, there is no classical space-time whatsoever; everything including geometry is treated quantum mechanically. A systematic mathematical and conceptual framework is introduced to substitute the standard one used in quantum theories of other forces.

Detailed analysis shows that the fundamental excitations of geometry are one-dimensional, rather like a polymer. The fabric of your shirt, for example, is woven by one-dimensional threads. However, for all practical purposes, the shirt appears to be a two-dimensional continuum. Similarly, the three-dimensional continuum of our everyday experience is 'woven' by the 'quantum threads' of geometry. More generally, smooth classical geometries arise from coarse graining of semiclassical, 'weave states'. In extreme circumstances, like those near the Big Bang and black hole singularities, the continuum picture breaks down and one must describe physics directly in terms of the fundamental one-dimensional excitations. Furthermore, in this theory, geometrical quantities such as lengths, areas and volumes are all quantized, i.e. can take on only discrete values. In the case of area, all of these permissible values are known explicitly. As one would expect, the elementary quanta are of the order of $\ell_{\rm P}^2$. But for large areas, the allowed discrete values crowd exponentially so that the continuum limit is reached very rapidly.

The questions raised by black hole thermodynamics have also been addressed in this approach. The microstates underlying a large black hole are now attributed to the quantum excitations of geometry itself. For non-rotating black holes, it is known that these microstates can account for the Hawking–Bekenstein entropy. (The restriction to the non-rotating case is only for technical simplicity and is not regarded as a fundamental limitation.) The Hawking process now corresponds to physical processes in which the quanta of area are converted to quanta of matter; the black hole shrinks as it emits radiation. Furthermore, somewhat surprisingly, the Hawking evaporation of *large* black holes provides a check on the details of predictions of quantum geometry. Specifically, the crowding of the eigenvalues found in the theory is essential for reproducing from first principles the black body spectrum found by Hawking using semiclassical methods.

Because of these successes, research in string theory and quantum geometry has intensified in recent years. (According to a survey by Carlo Rovelli, the number of quantum gravity publications in string theory is approximately three times that in the approach based on quantum geometry.) The strengths of the two approaches are complementary. In the analysis of black hole thermodynamics, for example, detailed calculations in string theory refer to black holes which are not of direct interest to astrophysics. Furthermore, by and large these calculations are performed in flat space and the relation of the microscopic degrees of freedom to the curved geometry of a black hole space-time is yet to become clear. The quantum geometry treatment encompasses black holes which are of interest to astrophysics and the relation to curved geometry is clear. However, in a certain well-defined sense, the quantum geometry treatment provides only an 'effective description' of a large black hole while the stringy description is 'fundamental' within that theory. In particular, string theory is much closer to addressing the issue of the final stages of black hole evaporation and the related 'information loss paradox'.

More generally, string theory has a natural built-in principle to specify the fundamental matter fields and their couplings. Over the last year, an interesting set of conjectures have emerged, pioneered by Juan Maldecena, which have allowed considerable flexibility in the choice of background space-times on which the theory is defined. But to be a satisfactory theory of quantum gravity, it would have to get rid of background dependence altogether. The approach based on quantum geometry is free of this problem. However, like general relativity, it has no principle to determine matter couplings. In the last two years, developments pioneered by Thomas Thiemann have led to considerable progress in formulating a quantum theory based on matter couplings taken from the standard model of particle physics. However, the issue of satisfactory quantum dynamics is yet to be settled. The two approaches seem to be complementary in many ways. Much progress would occur if in the coming years their strengths can be combined to overcome their present limitations.

Bibliography

Ashtekar A and Stachel J 1991 *Conceptual Problems of Quantum Gravity* (New York: Birkhauser)

Schwarz J Web-page on string theory http://theory.caltech.edu/people/jhs/strings/index.html

Abhay Ashtekar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quantum Mechanics

A development of quantum theory that was initiated in the 1920s by Werner Heisenberg (1901–76) and Erwin Schrödinger (1887–1961). The theory drew on a proposal made in 1925 Prince Louis de Broglie (1892–1987), that particles have wavelike properties (the wave–particle duality) and that an electron, for example, could in some respects be regarded as a wave with a wavelength that depended on its momentum. In this new formulation, the permitted energy levels in the Bohr theory of the hydrogen atom corresponded to electron orbits into which integral numbers of electron wavelengths could fit. For example, the lowest level (the lowest energy level of the Bohr atom) contains exactly one wavelength, the second level two, and so on. Thus, the discrete orbits and energy states of the atom followed as direct consequences of the wave properties of electrons.

Schrödinger developed a formulation of quantum mechanics, known as wave mechanics, in which an orbiting electron is treated as a standing wave, represented by a 'wavefunction', which is described by the Schrödinger wave equation (an equivalent formulation, based on matrices, was developed by Heisenberg). In accordance with the Heisenberg uncertainty principle, which implies that it is not possible to determine, simultaneously, the position and velocity of an electron, the wavefunction gives only the probability that an electron is in a particular state, or at a particular point at a particular time. Therefore, Bohr's precise orbits are replaced by a set of more fuzzy 'orbitals' within which electrons of a particular energy will lie.

The quantum state of an electron (or other subatomic particle) is defined by a set of numbers (quantum numbers) that specify quantities such as energy, angular momentum and spin and which are associated with solutions to the wave equation. Each energy level within an atom corresponds to a unique set of quantum numbers and various selection rules determine which transitions between levels, and hence which spectral lines, are permitted.

See also: atom, emission spectrum, uncertainty principle, Pauli exclusion principle, quantum theory, wave–particle duality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quantum Theory

A theory based on the premise that, on the microscopic scale, physical quantities have discrete, rather than a continuous range of, values. The theory was devised in the early part of the twentieth century to account for certain phenomena that could not be explained by classical physics. In 1900, the German physicist, Max Planck (1858–1947), was able precisely to describe the previously unexplained distribution of energy radiated by a black body by proposing that light, and other forms of electromagnetic radiation, exist in the form of discrete 'packets' of energy called quanta (singular: 'quantum'). The energy (E) of a quantum of electromagnetic radiation is directly proportional to the frequency (f) of the radiation and inversely proportional to its wavelength (λ); thus, $E = hf = hc/\lambda$, where h is a constant of proportionality, called Planck's constant (or the Planck constant), and c denotes the speed of light. The value of the Planck constant is 6.63×10^{-34} J s.

In 1905, Albert Einstein (1879–1955) used the concept of quanta to explain the photoelectric effect, a phenomenon whereby certain substances will emit electrons when illuminated by light of a sufficiently short wavelength. Einstein showed that, because a certain minimum amount of energy has to be imparted to an electron in order for it to be ejected, the emission of electrons would only take place if the incoming light consisted of quanta (or 'photons') with energies above that minimum and, therefore, wavelengths shorter than a particular value.

In 1913, Danish physicist Neils Bohr (1885–1962) extended the quantum hypothesis to matter as well as radiation by proposing that electrons in atoms could have only certain particular values of energy, values that were simple multiples of the Planck constant. While an electron was in one of the permitted energy states it could remain in orbit around the atomic nucleus without radiating energy (this was in marked contrast to classical theory, according to which an orbiting electron would radiate continuously, lose energy and spiral in towards the nucleus). If an electron were to drop down to a lower level, it would radiate a photon with a particular energy (and hence a particular wavelength) equal to the energy difference between the two permitted levels. Bohr's theory gave very good agreement with the observed wavelengths of the various lines in the spectrum of the element hydrogen. Further development of the theory led to the suggestion that the angular momenta of orbiting electrons were restricted to integer multiples of $h/2\pi$ and to the idea that the electron itself had an intrinsic spin (as if it were a tiny sphere rotating around its axis) of magnitude $h/4\pi$. The precise quantum state of an electron (or other subatomic particle) is defined by a set of numbers (quantum numbers) that specify quantities such as energy, angular momentum and spin.

A more comprehensive, and deeper, version of the theory, called quantum mechanics, was developed during the 1920s. Nowadays, quantum theory embraces practically all aspects of microscopic physics.

See also: angular momentum, atom, black-body radiation, electron, electromagnetic radiation, uncertainty principle, Pauli exclusion principle, photon, quantum mechanics, wave–particle duality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quark

A fundamental particle that joins with others to form hadrons (baryons and mesons).

There are six varieties, or 'flavors', of quark, each of which has an equivalent antiparticle. The six flavors, in order of increasing mass, are called up, down, strange, charm, bottom, and top. Quarks possess fractional charge (i.e. they carry an electrical charge that is a simple fraction of the charge on the electron) and have half-integer values of spin. For example, an up quark (symbol 'u') carries a charge of $+\frac{2}{3}$ and an anti-up ($\bar{u}$), $-\frac{2}{3}$; a down quark (d) carries a charge of $-\frac{1}{3}$ and an anti-down ($\bar{d}$), $+\frac{1}{3}$. Quarks join together in clusters of three to form baryons, such as protons and neutrons, and as quark–antiquark pairs to form mesons. For example, a proton consists of two ups and a down (net charge: $+\frac{2}{3} + \frac{2}{3} - \frac{1}{3} = +1$) and an antiproton of two anti-ups and an anti-down (net charge: $-\frac{2}{3} - \frac{2}{3} + \frac{1}{3} = -1$); a neutron consists of two downs and an up (net charge $-\frac{1}{3} - \frac{1}{3} + \frac{2}{3} = 0$), and so on. A neutral pion (pi-meson) consists of an up and an anti-up ($u\bar{u}$) or a down and an anti-down ($d\bar{d}$), a positive pion of an up and an anti-down ($u\bar{d}$), and so on. Of the six flavors of quark, only the up and down quarks are needed to make the nucleons (the building blocks of atomic nuclei) and pions, the other varieties being required to construct some of the many short-lived particles that are found among cosmic rays and created in high-energy particle accelerators.

Six leptons are associated with the six quarks in three families, or 'generations': the electron and electron-neutrino with the up and down quarks (first generation); the muon and muon-neutrino with the charm and strange quarks (second generation); and the tau and tau-neutrino with the top and bottom quarks (third generation).

In addition to electrical charge, quarks possess another property, called 'color', which in some respects is analogous to electrical charge. There are three types of color (called red, blue, and yellow by analogy with the primary colors), each of which can be positive or negative. As with electrical charge, positive and negative colors attract; but, in addition, different colors attract and like colors repel. In a hadron (a particle that consists of quarks), the net color has to be neutral (or 'white'). This can be achieved either by combining three different colored quarks to make a baryon, or by linking a quark with an antiquark (in which case the opposite color charges neutralize each other) to form a meson. The interaction between the color charges, which is the believed to be the underlying source of the strong nuclear force, is described by a theory called quantum chromodynamics. The strong (color) force is conveyed between bound quarks by force-carrying particles called gluons and is so powerful that, at least to the limits of present-day experiments, individual quarks cannot be ejected from within hadrons.

In the present-day universe, quarks exist only inside hadrons, not as free particles. However, during the very high-energy phase of the very early universe (during the first 10^{-6} s), free quarks would have been abundant.

See also: antimatter, atom, baryons, Big Bang theory, elementary particles, fundamental forces, fundamental particles, grand unified theory, hadrons, mesons, neutrinos, neutron, proton.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasar

An extragalactic object (one that lies well beyond the confines of our own Galaxy) that has a very small angular size, but which has a high redshift. Because the first objects of this kind to be discovered, in the early 1960s, were radio sources that coincided in appearance with what looked like stars, they were called 'quasi-stellar radio sources', a name that was then abbreviated to quasar. Although it was shown subsequently that only about 10% of objects of this kind were actually strong radio emitters, and the more general term, 'quasi-stellar object' (QSO) was invoked to include the radio-quiet versions, in practice, the term 'quasar' is widely used to describe both kinds of object.

The spectrum of a quasar typically displays emission lines superimposed on a non-thermal (non-stellar) continuous spectrum (continuum). The redshifts of the lines in quasar spectra range from 0.06 to around 5, these values of redshift corresponding to velocities of recession ranging from about 6% to about 95% of the speed of light. If, as is generally accepted, the redshifts imply that quasars are remote objects that are receding because, like conventional galaxies, they are participating in the general expansion of the universe in accordance with the Hubble law, then they must all lie at very great distances. To appear as bright as they do, despite their very great distances, their luminosities must range from about a hundred to about ten thousand times greater than that of a conventional galaxy like the Milky Way (the implied luminosities are in the range 10^{39}–10^{41} watts). Many quasars vary substantially in brightness (by factors of up to 10, or even more in exceptional cases) over periods of time as short as months or even days. This implies that most of their energy is emitted from regions of space that are very small, in some cases less than a light-day in diameter.

Since the 1980s a number of quasars have been shown to be surrounded by faint fuzzy patches of light. The spectra of these 'fuzzy blobs' contain stellar-type absorption lines, consistent with the idea that quasars are embedded within remote galaxies. In addition, studies of quasars that appear to lie within clusters of galaxies have shown that they have the same redshifts as those clusters, consistent with their being located at very great distances. Furthermore, quasars share similar characteristics to active galactic nuclei in general. They radiate strongly over a wide range of wavelengths, have a non-thermal spectrum, display short-term variability, are compact and highly luminous and, in some cases, appear to be emitting jets and blobs of material. In light of this evidence, astronomers (with a few notable exceptions) are convinced that a quasar is an extremely brilliant active galactic nucleus, so remote and so highly luminous compared with the galaxy within which it is embedded that, in most cases, only the compact nucleus can be seen.

As with other active galactic nuclei, the energy source within a quasar is widely believed to be a supermassive black hole that is accreting matter from its surroundings. The broad emission lines in a quasar's spectrum are believed to originate in heated gas clouds revolving around the central black hole at speeds of up to several thousand kilometers per second.

See also: accretion, active galaxy, active galactic nucleus, black hole, blazar, BL Lacertae objects, Hubble law, redshift, spectrum, synchrotron radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasiperiodic Oscillations in X-ray Binaries

The term quasiperiodic oscillation (QPO) is used in high-energy astrophysics for any type of non-periodic variability that is constrained to a relatively narrow range of variability frequencies. X-RAY BINARIES are systems in which a 'compact object', either a BLACK HOLE or a NEUTRON STAR, orbits a normal star and captures matter from it. The matter spirals down to the compact object and heats up to temperatures of 10^7 K, which leads to the emission of x-rays. Such systems therefore provide ways of obtaining information about black holes and neutron stars that cannot be obtained in any other way, as the basic process taking place in them involves matter in close orbit around such a compact object. QPOs are an important diagnostic of these flows of matter in x-ray binaries. They are usually studied by calculating power spectra of a time series of, most commonly, x-ray flux samples by means of the Fourier transform. In such a power spectrum, the QPO shows up as a peak of finite width centered around the oscillation frequency (see figures 1 and 3). If the peak is too narrow, such as in the case of PULSARS, the signal is called periodic instead of quasiperiodic, and if the peak is too broad (conventionally, if its full width at half-maximum is broader than half its centroid frequency) the feature is usally called a 'noise' feature. There is no sharp physical distinction between QPO and noise features, and they are discussed together here. Two important classes of noise are 'power-law noise', which shows up in the power spectrum as a power law and thus has no measured associated time scale, and 'band-limited noise', which steepens towards higher frequency above some break or cut-off frequency $\nu_{\rm break}$.

A variety of processes can produce a similar power-spectral peak. Often the statistical quality of the data is insufficient to see the QPO directly in the time series. Then, it can only be detected by calculating the average power spectrum of a large amount of data. The exact signal causing a given power-spectral QPO peak then remains unknown. Popular examples of signals in the time series that may explain the QPO peaks observed in the power spectra of x-ray binaries are oscillating-shot signals, essentially a superposition of many short-lived oscillations, and frequency-modulated signals, which are oscillations whose frequency varies on a time scale much longer than the oscillation period.

The presence of short-lived 'clumps' of matter orbiting the central object at some preferred radius in the ACCRETION DISK and producing a QPO with a frequency equal to the Keplerian orbital frequency corresponding to that radius may be the most popular explanation for QPOs in x-ray binaries (see the section on models). However, as several different types of QPO are observed, several different models are required. Most QPO phenomena increase in frequency when the accretion rate increases; in Keplerian models this implies the orbits come closer with increasing mass accretion rate. From KEPLER'S LAWS, the closest possible orbits around neutron stars and stellar-mass black holes could have periods as short as 0.5 ms, and from the beginning astronomers have searched for QPOs with frequencies in the corresponding kilohertz domain. Only very recently have such QPOs been found (see the section on the current era). Their exploitation for the study of strong-field general-relativistic effects and the basic properties of neutron stars and black holes has only just begun.

History: the discoveries with EXOSAT and Ginga

Early reports of QPOs in x-ray binaries included observations with the Japanese Hakucho satellite of ∼2 Hz oscillations during bursts of the Rapid Burster and with the European Space Agency's EXOSAT of 0.7–20 mHz QPOs in Cygnus X-3. Both of these objects are unique, very peculiar x-ray binaries. The discoveries of two different QPO phenomena (the 15–60 Hz 'horizontal branch oscillation' and the 6–20 Hz 'normal–flaring branch oscillation', or HBO and NFBO; figure 1) with EXOSAT in the brightest low-mass x-ray binaries (LMXBs) during 1985–1987 were the first which had sufficiently suggestive properties to lead to an appreciable body of interpretative work. In particular, the relatively high frequencies of these oscillations suggested an origin connected in some way with the inner accretion flow, where the characteristic time scales are short.

By the early 1990s, the magnetospheric beat frequency model, in which the oscillations occur as a result of a quasiperiodic modulation of the accretion rate caused by an interaction between the inner accretion disk and the magnetic field of a rapidly spinning neutron star remained as a prominent, although not universally accepted, model for the HBO. The radiative-feedback model, where a radiation-force-driven instability quasiperiodically modulates the optical depth of an approximately radial, near-Eddington accretion flow and thus causes the oscillations by variations in the radiative transfer through this flow, enjoyed a similar status for the NFBO. Studies of correlated changes of the x-ray spectrum and the power spectrum of accreting low-magnetic-field neutron stars in LMXBs led to a subdivision into two groups, named, after the characteristic shapes of the tracks they trace out in x-ray color–color diagrams, the 'Z sources' and the 'atoll sources' (figure 1). In such diagrams, the x-ray 'colors' or spectral hardnesses in two different photon energy bands are plotted versus each other. The tracks in these diagrams are usually quite repeatable in shape (less so in the position in the diagram) and are traced out over and over again on time scales of hours to days. Sources in both groups were found to show several 'source states' characterized by their x-ray spectral and power-spectral properties; the transitions between these states are gradual and occur with intervals of hours to many days. The HBO and the NFBO were observed in the Z sources only; common to both source types were two broad power-spectral components, a power-law noise component called very-low-frequency noise, dominating the power spectra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

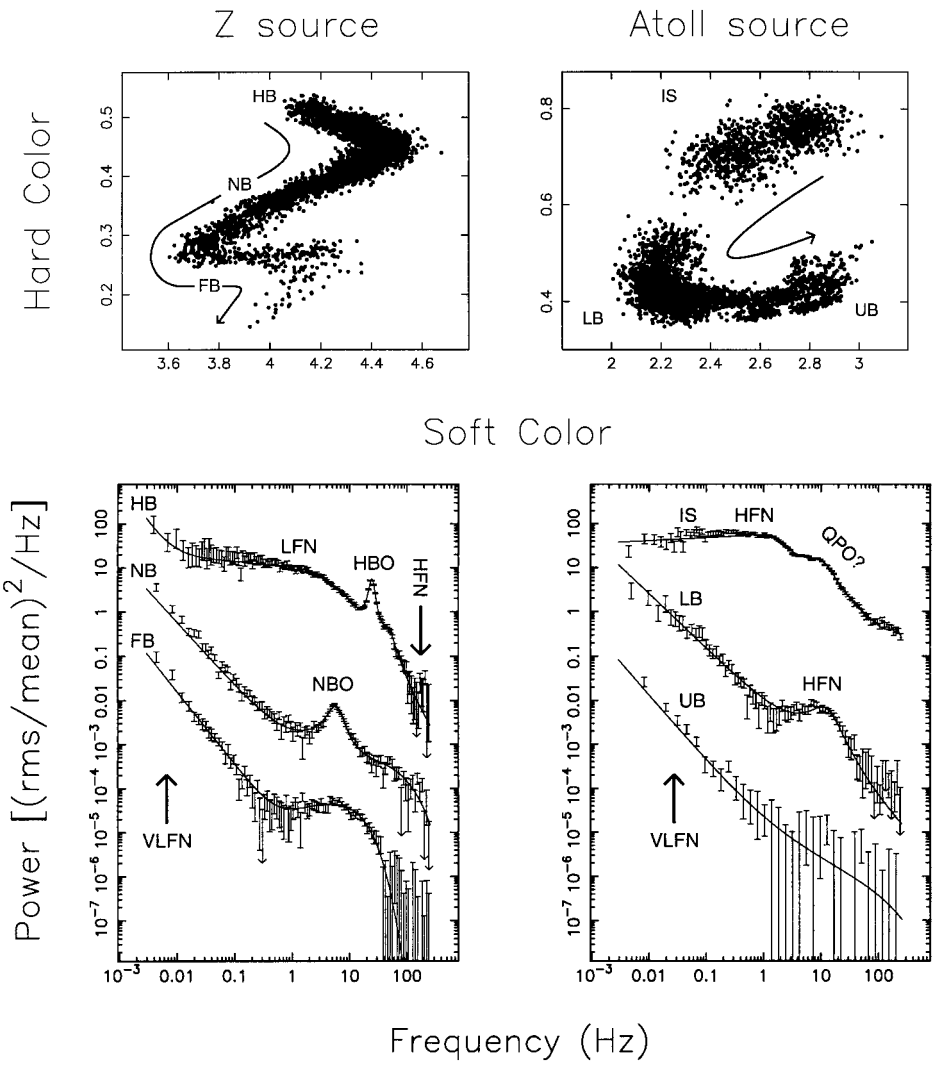


Figure 1. X-ray color–color diagrams (top) and power spectra (bottom) of Z and atoll sources. Top: the soft and hard colors represent the x-ray spectral hardness in the 3–6 and 6–10 keV photon energy bands, respectively. Arrows indicate the sense in which $\dot{M}$ is inferred to increase along each track. The various source states as they appear in the color–color diagram, horizontal branch, normal branch, flaring branch, island, lower banana and upper banana are indicated as HB, NB, FB, IS, LB and UB, respectively. Bottom: typical power spectra are displayed for each of the source states identified in the color–color diagrams. The various power spectral noise (very-low-frequency noise, low-frequency noise and high-frequency noise; VLFN, LFN and HFN) and QPO (horizontal branch oscillation, normal branch oscillation; HBO and NBO) components are indicated. Despite its name, the HFN in the atoll sources is probably similar to the LFN in Z sources.

below ∼1 Hz, and a band-limited noise component called either low- or high- frequency noise, with $\nu_{\rm break}$ typically between 0.1 and several tens of Hz. All these variability phenomena vary in their properties, often monotonically, along the tracks in the color–color diagram. The position of a source in its track was therefore interpreted as a measure of its instantaneous accretion rate, $\dot{M}$; this interpretation implied that x-ray count rates in the usual photon energy bands (e.g. 2–20 keV) are not proportional to $\dot{M}$; it is even possible for the count rate to rise when $\dot{M}$ drops.

Observations with the GINGA satellite in the years 1987–1991 led to a much clearer view of the correlated x-ray spectral and power-spectral properties of the black hole candidates (BHCs), which turned out to be in many ways very reminiscent of those of the (neutron-star) atoll sources. The source states in BHCs were called, in order of increasing inferred $\dot{M}$, the low state, high state and very high state. In the very high state clear 3–10 Hz QPO were seen that are perhaps related to some of the QPOs seen in the Z sources. In the low state the power spectra are dominated by a band-limited noise component similar to that seen in the atoll sources. The break frequency is variable between 0.03 and 1 Hz and correlated with inferred $\dot{M}$. In the high state very little power is observed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

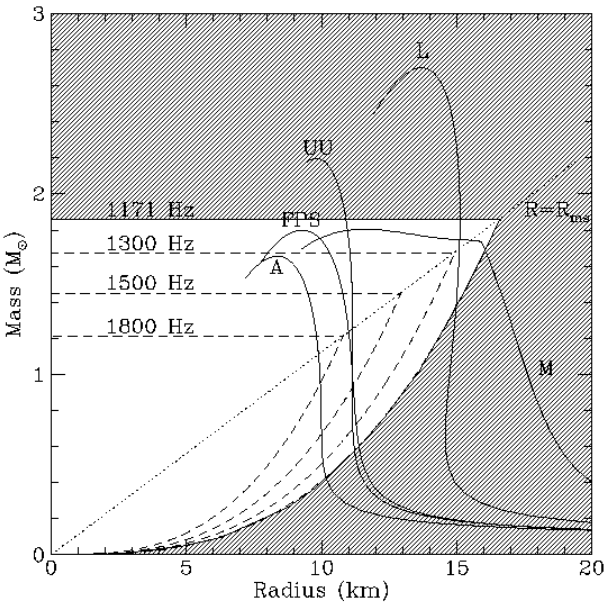


Figure 2. Mass–radius diagram of neutron stars. The curves labeled A, FPS, UU, L and M represent neutron-star mass–radius relations for various different assumed equations of state of high-density matter. The shaded area is the region that can be excluded if orbital motion takes place around the neutron star at a frequency of 1171 Hz, as suggested by observed kHz QPO frequencies. For higher observed frequencies, larger areas can be excluded as indicated. The currently highest measured kHz QPO frequency is 1228 Hz. Adapted from Miller *et al* (1998).

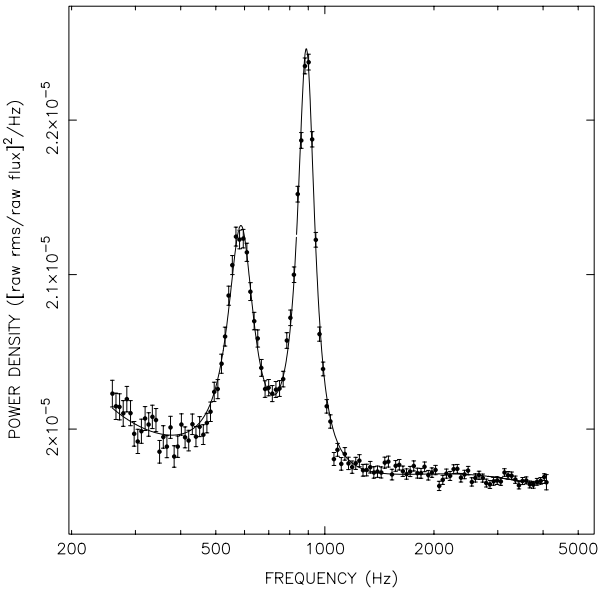


Figure 3. Power spectrum of the Z source Scorpius X-1 showing the twin kHz QPOs.

In the very high state a band-limited noise component reappears with $\nu_{\rm break}$ between 1 and 10 Hz. Later, an intermediate state was identified that fits between the low and high states and has properties similar to those of the very high state.

The current era: RXTE

Launched on 30 December 1995, the ROSSI X-RAY TIMING EXPLORER (RXTE) satellite provided the latest, and largest, upswing in our knowledge of QPO phenomena in x-ray binaries. It seems likely that for the first time we are detecting signals from the bottom of the gravitational potential wells of neutron stars and BHCs related to the orbital motion of matter very close around the compact objects. In these regions, space–time is strongly curved and strong-field general relativity is required to describe the motion of matter. Indeed, interpretations involving strong-field general-relativistic effects are being debated extensively with respect to the newly discovered phenomena. The inferred orbits are so close that they directly constrain the density of the central neutron star, and thus the compressibility (i.e. the equation of state) of the supranuclear-density matter of which they are composed (figure 2).

In both Z and atoll sources, the long-predicted phenomena on millisecond and submillisecond time scales were found at last. Surprisingly, these phenomena are in the form of often quite narrow QPO peaks with frequencies up to about 1200 Hz that come in pairs, the 'twin kHz QPOs' (figure 3). It seems likely that one of these peaks is caused by the orbital motion at the inner edge of the accretion disk. In x-ray bursts of several atoll sources a different type of oscillation was discovered, with a frequency, near constant in each source, between 290 and 590 Hz, that was interpreted as due to the neutron-star spin. RXTE also finally, after a search of nearly 20 yr, found the first accreting millisecond pulsar, which, at a spin frequency of 401 Hz, supports the interpretation of the burst oscillations as being due to the neutron-star spin. More importantly, this discovery provides the first direct confirmation of the evolutionary ideas about the origin of the millisecond radio pulsars.

QPOs at frequencies of a few tens of Hz in atoll sources evident in the RXTE data may be similar to Z source HBOs. QPOs found in two BHCs stand out as perhaps different and possibly related to each other: the 67 Hz QPO in GRS 1915 + 105, which seems to always occur at the same frequency, and the 300 Hz QPO in GRO J1655 − 40, the highest-frequency QPO in a BHC so far. These QPOs may be related to some basic property of accretion onto a black hole of given mass and spin, such as the orbital, Lense–Thirring or apsidal motion frequency near the general-relativistic innermost stable orbit.

Models

Beat-frequency models have played a central part in interpretations of QPO phenomena observed in accreting neutron stars. In such models, there is an interaction between the spin of the neutron star and the orbital

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

motion at some preferred radius in the accretion disk which produces a third phenomenon, the beat, which has a frequency ν_B that is equal to the difference between the neutron-star spin frequency ν_S and the orbital frequency in the disk ν_D according to the relation $\nu_B = \nu_D - \nu_S$. No signal is generated at the sum of these frequencies because the interacting frequencies are those of two angular motions that take place in the same sense (a sum frequency would be observed if the neutron-star spin was in the opposite sense to the orbital motion in the disk, a situation which because of angular-momentum transfer from the disk to the star cannot persist for a long time).

Two mechanisms have been worked out to provide the spin–orbit interaction in a way that would lead to observable QPO phenomena. In the magnetospheric beat-frequency model the accretion of matter from short-lived clumps of plasma in the inner accretion disk is modulated by electromagnetic interactions with the neutron-star magnetic field, which spins with the spin of the neutron star. In this model the preferred disk radius at which we observe the Keplerian orbital frequency is the radius of the magnetosphere. This model has been used to explain the 15–55 Hz HBO in Z sources as due to the beat between the (unobserved) disk orbital and neutron-star spin frequencies, and to explain much slower QPO phenomena that are seen in some accreting x-ray pulsars with periods of $1–10^2$ s, but has also been discussed as a possible explanation of the kHz QPOs. In the sonic-point beat-frequency model the preferred radius in the disk is located at the sonic point, where the disk flow becomes radially supersonic owing to either general-relativistic effects or interaction of the accretion flow with the emerging radiation. This model has been proposed for the twin kHz QPO peaks, where the higher frequency of the twin QPO peaks would be at ν_D and the lower-frequency one at ν_B.

The spin frequency of the neutron stars in some of the twin kHz QPO sources as deduced from the burst oscillations is found to be close to the frequency difference between the two kHz peaks in the same source, in accordance with such a spin–orbit beat-frequency model. However, several discrepancies between the predicted and observed frequencies have emerged since the model was proposed, and it is clear that any spin–orbit beat-frequency model for the twin kHz QPOs cannot be as straightforward as was initially believed. Other models relying on various kinds of hydrodynamic or radiation–hydrodynamic accretion–flow instabilities are also being explored.

Since the discovery of the kHz QPOs, which from their high frequencies probably originate very close to the compact objects, general-relativistic effects have played an increasingly prominent role in the attempts at interpretation of all QPOs. It has been proposed that Z source HBOs and related QPOs in atoll sources are due to Lense–Thirring precession of the inner accretion disk, and evidence has been mounting that the kHz QPO frequencies, which increase in each given source with mass accretion rate, reach a maximum value near ∼1200 Hz. If this is because the orbits corresponding to this frequency are near the general-relativistic innermost stable orbit (at $6GM/c^2$ with small corrections for frame dragging due to the spin of the neutron star), then neutron-star masses near 2 $M_\odot$ are derived for the neutron stars in low-mass x-ray binaries. It so far remains a mystery why 1200 Hz is reached in different sources at luminosities that differ by factors of up to several 100. Lense–Thirring precession of orbits around near-maximally rotating black holes has also been speculated to explain the 67 and 300 Hz QPOs in GRS 1915 + 105 and GRO J1655 − 40. Finally, a model has been proposed where the lower frequency of the twin kHz QPO is due to general-relativistic apsidal motion. In particular, in the case of the neutron stars where HBO, burst QPO and twin kHz QPOs are all seen in a single source, the data can constrain models of this type very well, and a shake-out among these models can be foreseen when data accumulates.

With the discovery of the kHz QPOs, the historical goal of obtaining diagnostic information about the motion of matter at the bottom of the gravitational well of stellar-mass compact objects from high-time-resolution x-ray data appears to have been reached. Future work, requiring much larger-area instruments than even RXTE, will probably concentrate on mapping out space–time in these regions in detail by precisely reconstructing the orbits of matter in these regions.

Bibliography

Bildsten L and Strohmayer T 1999 New views of neutron stars *Phys. Today* February 40–6

Charles P A and Seward F D 1995 Low-mass x-ray binary stars *Exploring the X-ray Universe* (Cambridge: Cambridge University Press) pp 187–217

Miller M C *et al* 1998 *Astrophys. J.* **508** 791

van der Klis M 1995 Rapid aperiodic variability in x-ray binaries *X-ray Binaries* ed W H G Lewin, J van Paradijs and E P van den Heuvel (Cambridge: Cambridge University Press) pp 252–300

van der Klis M 1998 Kilohertz quasi-periodic oscillations in low-mass x-ray binaries *The Many Faces of Neutron Stars* ed R Buccheri, J van Paradijs and M A Alpar (Dordrecht: Kluwer Academic) pp 337–68

M van der Klis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasistellar Objects: Intervening Absorption Lines

Every parcel of gas along the line of sight to a distant QUASAR will selectively absorb certain wavelengths of continuum light of the quasar due to the presence of the various chemical elements in the gas. Through the analysis of these quasar absorption lines we can study the spatial distributions, motions, chemical enrichment and ionization histories of gaseous structures from REDSHIFT five until the present. This includes the GAS IN GALAXIES of all morphological types as well as the diffuse gas in the INTERGALACTIC MEDIUM.

Basics of quasar spectra

Figure 1 illustrates many of the common features of a quasar spectrum. The relatively flat quasar continuum and broad emission features are produced by the quasar itself (near the black hole and its accretion disk). In some cases, gas near the quasar central engine also produces 'intrinsic' absorption lines, most notably Ly α, and relatively high ionization metal transitions such as C IV, N V, and O VI. These intrinsic absorption lines can be broad (thousands or even tens of thousands of km s^{-1} in which case the quasar is called a broad absorption line (BAL) QSO; see QUASISTELLAR OBJECTS: INTRINSIC AGN ABSORPTION LINES), or narrow (tens to hundreds of km s^{-1}). However, the vast majority of absorption lines in a typical quasar spectrum are 'intervening', produced by gas unrelated to the quasar that is located along the line of sight between the quasar and the Earth.

A structure along the line of sight to the quasar can be described by its neutral hydrogen column density, N(H I), the number of atoms per cm^2. N(H I) is given by the product of the density of the material and the pathlength along the line of sight through the gas. Each structure will produce an absorption line in the quasar spectrum at a wavelength of $\lambda_{\rm obs} = \lambda_{\rm rest}(1 + z_{\rm abs})$, where $z_{\rm abs}$ is the redshift of the absorbing gas and $\lambda_{\rm rest} = 1215.67$ Å is the rest wavelength of the Ly α transition. Since $z_{\rm abs} < z_{\rm QSO}$, the redshift of the quasar, these Ly α absorption lines form a 'forest' at wavelengths blueward of the Ly α emission. The region redward of the Ly α emission will be populated only by absorption through other chemical transitions with longer $\lambda_{\rm rest}$. Historically, absorption systems with N(H I) $< 10^{17.2}$ cm^{-2} have been called LYMAN ALPHA FOREST lines, those with $10^{17.2} < N$(H I) $< 10^{20.3}$ cm^{-2} are Lyman limit systems, and those with N(H I) $> 10^{20.3}$ cm^{-2} are damped Ly α systems (see LYMAN ALPHA ABSORPTION: THE DAMPED SYSTEMS).

The number of systems per unit redshift increases dramatically with decreasing column density, as illustrated in the schematic diagram in figure 2. Lyman limit systems are defined by a sharp break in the spectrum due to absorption of photons capable of ionizing H I, i.e. those with energies greater than 13.6 eV. The optical depth, τ, of the break is given by the product N(H I)σ, where the cross section for ionization of hydrogen, $\sigma = 6.3\times10^{-18}(E_\gamma/13.6\,{\rm eV})^{-3}\,{\rm cm}^2$, (and the flux is reduced by the factor $e^{-\tau}$). The energy dependence of σ leads to a recovery of the Lyman limit break at higher energies (shorter wavelengths), unless N(H I) $\gg 10^{17.2}$ cm^{-2} (see figure 1).

The curve of growth describes the relationship between the equivalent width of an absorption line, W (the integral of the normalized profile), and its column density, N. Figure 3 shows that for small N(H I) the number of absorbed photons, and therefore the flux removed, increases in direct proportion to the number of atoms. This is called the linear part of the curve of growth. As N is increased, the line saturates so that photons are only absorbed in the wings of the lines; in this regime the equivalent width is sensitive to the amount of line broadening (characterized by the Doppler parameter b), but does not depend very strongly on N(H I). This is the flat part of the curve of growth. Finally, at N(H I) $> 10^{20.3}$ cm^{-2}, there are enough atoms that the damping wings of the line become populated and the equivalent width increases as the square root of N(H I), and is no longer sensitive to b.

In addition to the Ly α (1s $\rightarrow$ 2p) and higher-order (1s $\rightarrow$ np) Lyman series lines, quasar spectra also show absorption due to different ionization states of the various species of metals. Figure 1 illustrates that the damped Ly α system at $z = 0.86$ that is responsible for the Ly α absorption line at $\lambda_{\rm obs} = 2260$ Å and a Lyman limit break at $\lambda_{\rm obs} = 1700$ Å also produces absorption at $\lambda_{\rm obs} = 2870$ Å due to the presence of C IV in the absorbing gas at that same redshift. Like many of the strongest metal lines seen in quasar spectra, C IV is a resonant doublet transition due to transitions from $^2S_{1/2}$ energy levels to the $^2P_{1/2}$ and to the $^2P_{3/2}$ energy levels. (The left superscript '2' represents the number of orientations of the electron spin, the letter S or P represents the total orbital angular momentum, L, and the right subscript represents the total angular momentum, J.) Doublet transitions are easy to identify. The dichotomy between rest wavelength and redshift is resolved because the observed wavelength separation of the doublet members increases as $1+z$. Table 1 lists some of the metal lines that are commonly detected for intervening absorption systems. Many of these are only strong enough to be observable for quasar lines of sight that pass through the higher N(H I) regions of galaxies.

History, surveys and revolutionary progress in the 1990s

The history of quasar absorption lines began within a couple of years of the identification of the first quasar in 1963. In 1965, Gunn and Peterson considered the detection of flux blueward of the Ly α emission line in the quasar 3C 9, observed by Schmidt, and derived a limit on the amount of neutral hydrogen that could be present in intergalactic space. In that same year, Bahcall and Salpeter predicted that intervening material should produce observable discrete absorption features in quasar spectra. Such features were detected in 1967 in the quasar PKS 0237 $-$ 23 by Greenstein and Schmidt, and in 1968 in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

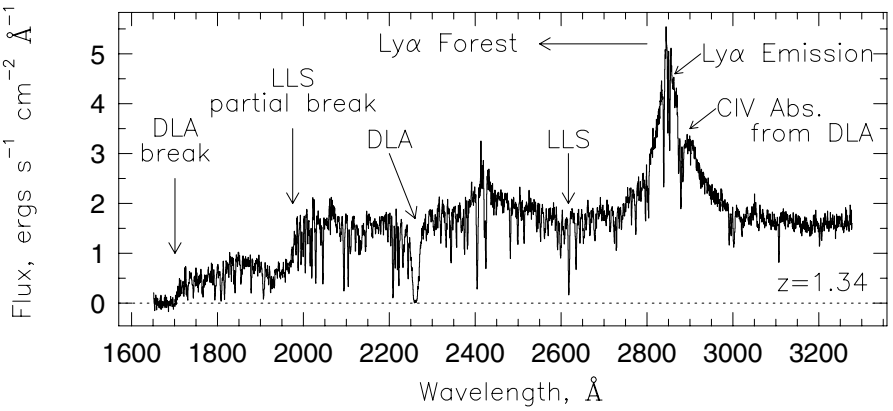


Figure 1. Typical spectrum of a quasar, showing the quasar continuum and emission lines, and the absorption lines produced by galaxies and intergalactic material that lie between the quasar and the observer. This spectrum of the $z = 1.34$ quasar PKS0454 + 039 was obtained with the Faint Object Spectrograph on the Hubble Space Telescope. The emission lines at $\sim$ 2400 Å and $\sim$2850 Å are Ly β and Ly α. The Ly α forest absorption, produced by various intergalactic clouds, is apparent at wavelengths blueward of the Ly α emission line. The two strongest absorbers, due to galaxies, are a damped Ly α absorber at $z = 0.86$ and a Lyman limit system at $z = 1.15$. The former produces a Lyman limit break at $\sim$1700 Å and the latter a partial Lyman limit break at $\sim$1950 Å since the neutral hydrogen column density is not large enough for it to absorb all ionizing photons. Many absorption lines are produced by the damped Ly α system at $z = 0.86$ (C IV $\lambda\lambda$1548, 1550, for example, is redshifted onto the red wing of the quasar's Ly α emission line).

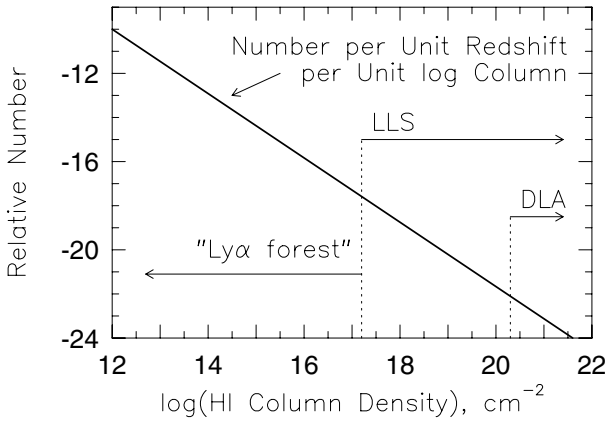


Figure 2. The column density distribution of Ly α clouds, $f(N(\mathrm{H\,I}))$, roughly follows a power law over ten orders of magnitude; there are many more weak lines than strong lines. The column density regions for the three categories of systems are shown: Ly α forest, Lyman limit, and damped Ly α. The term 'Ly α forest' has at times been used to refer to metal-free hydrogen clouds, perhaps those with $N(\mathrm{H\,I}) < 10^{16}$ cm^{-2}, but now metals have been found associated with weaker systems down to the detection limit.

Table 1. Common transitions.

Transition	λ_{rest} (Å)
LL	$\sim$912
Ly γ	972.537
Ly β	1025.722
Ly α	1215.670
Si IV 1393	1393.755
Si IV 1402	1402.770
C IV 1548	1548.195
C IV 1550	1550.770
Fe II 2382	2382.765
Fe II 2600	2600.173
Mg II 2796	2796.352
Mg II 2803	2803.531

PHL 938 by Burbidge, Lynds and Stockton. By 1969 many intervening systems had been discovered, and Bahcall and Spitzer proposed that most with metals were produced by the haloes of normal galaxies. As more data accumulated, the sheer number of Ly α forest lines strongly supported the idea that galactic and intergalactic gas, and not only material intrinsic to the quasar, is the source of most quasar absorption lines.

In the 1980s many more quasar spectra were obtained and many large statistical surveys of the different classes of absorption line systems were published. The emphasis was on characterizing the number of lines per unit redshift, $\mathrm{d}N/\mathrm{d}z$, stronger than some specified equivalent width limit. With 4 m class telescopes (equipped with charge coupled device (CCD) detectors) it was possible to conduct surveys with a spectral resolution of $R \sim 1000$. The spectral resolution is defined as $R = \lambda/\Delta\lambda = c/\Delta v$, so that $R = 1000$ corresponds to 300 km s^{-1} or 5 Å at $\lambda = 5000$ Å. Separate surveys were conducted for Ly α lines, Mg II doublets, C IV doublets, and also for Lyman limit breaks, all as a function of redshift. The Ly α line is observable in the optical part of the spectrum for $z > 2.2$, Mg II for $0.4 < z < 2.2$, C IV for $1.7 < z < 5.0$, and the Lyman limit break for $z > 3$. However, a break is also easily identified in lower-resolution space-based UV spectra, which extended Lyman limit surveys to lower redshift.

In order to consider the cross section of the sky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

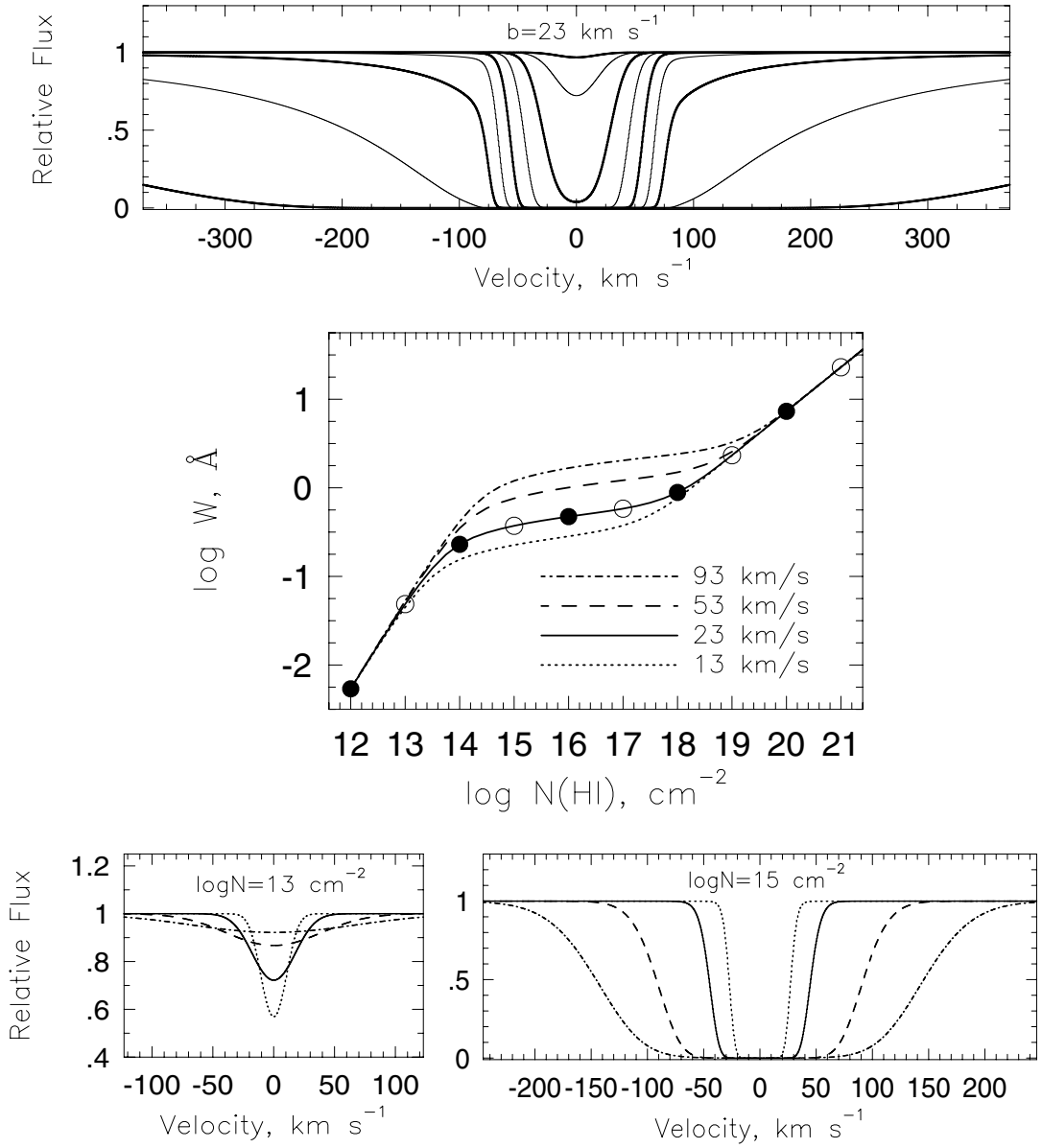


Figure 3. Illustration of the different regimes of the curve of growth. The middle panel shows the curve of growth for the Ly α transition, relating the equivalent width, W, of the absorption profile to the column density, N(H I). The different curves represent four different values of the Doppler parameter: $b = 13$, 23, 53 and 93 km s^{-1}. The upper panel shows absorption profiles with Doppler parameter $b = 23$ km s^{-1} for the series of neutral hydrogen column densities N(H I) = 10^{12}–10^{20} cm^{-2}. The thick (thin) curves correspond to the full (open) points on the $b = 23$ km s^{-1} curve of growth (middle panel), starting at N(H I) = 10^{12} cm^{-2}. For N(H I) < 10^{13} cm^{-2}, known as the linear part of the curve of growth, the equivalent width does not depend on b. The lower left panel shows that, at fixed N(H I), the depth of the profile is smaller for large b, such that the equivalent width remains constant. On the flat part of the curve of growth, profiles are saturated and the equivalent width increases with b for constant N(H I). For N(H I) > 10^{20} cm^{-2}, the profile develops damping wings, which dominate the equivalent width.

covered by the different populations, it can be assumed that absorption will be observed for all lines of sight within some radius of every luminous galaxy (>0.05L_K^*). (L_K^* represents the Schechter luminosity, i.e. the transition between the exponential and the power law forms of the luminosity function, and corresponds to a K-band absolute magnitude of $M_K = -25$.) To explain the observed dN/dz at $z \sim 1.5$, this radius would be 70 kpc for strong C IV (detection sensitivity 0.4 Å), and 40 kpc for strong Mg II (detection sensitivity 0.3 Å) and also for Lyman limit systems, implying that the latter two populations are in fact produced in the same gas. The higher N(H I) damped

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ly α absorbers would be produced within 15 kpc of the center of each galaxy, while the Ly α forest lines would require a considerably larger region, hundreds of kpcs around each galaxy, to produce a cross section consistent with the observed number of weak lines.

Up until the 1990s, the focus of quasar absorption line work was to separately consider the properties of the individual classes of absorbers (e.g. Ly α forest or Mg II absorbers). In the 1990s, however, three different observational advances led to recognition of the direct connections between the different classes of quasar absorption lines, and of direct associations with the population of galaxies:

(a) Deep images of quasar fields could be obtained, and redshifts of the galaxies in the field could be determined from low-resolution spectra. Steidel found that whenever Mg II absorption with $W_r(\text{Mg II}) > 0.3$ Å is observed, a luminous galaxy ($L_K > 0.06L_K^*$) is found within an impact parameter of 38 h^{-1} $(L/L_K^*)^{-0.15}$ kpc with a redshift coincident with that determined from the absorption lines. Also, it is rare to find a galaxy within this impact parameter that does not produce Mg II absorption. There appears to be a one-to-one correspondence between strong Mg II absorption and luminous galaxies. The Mg II absorbing galaxies span a range of morphological types.

(b) The High Resolution Spectrograph on the Keck I 10 m telescope made it possible to obtain quasar spectra at a resolution of $R = 45\,000$, which corresponds to $\sim$6 km s^{-1}. Previous surveys with a resolution of the order of hundreds of km s^{-1} identified absorption due to entire galaxies and their environments. With 6 km s^{-1} resolution it became possible to resolve structure within a galaxy: the clouds in its halo, the interstellar medium of its disk, and the satellites and infalling gas clouds in its environment. Figure 4 is a dramatic illustration of this contrast for the Mg II absorber at $z = 0.93$ toward the quasar PG 1206+459.

(c) The Faint Object Spectrograph (FOS) on the Hubble Space Telescope provided resolution $R \sim 1000$ in the UV, from 1400–3300 Å. Observations of Ly α forest clouds could be extended from $z = 2.2$ down to the present epoch. Furthermore, absorption from a given galaxy could be observed in numerous transitions; if Mg II was observed in the optical, the Lyman series and C IV could be studied in the UV (see figure 4). With information on transitions with a range of ionization states, consideration of the degree of ionization (related to the gas density and the intensity and shape of the ionizing radiation field) and the multiple phase structure of galactic gas became possible.

No longer is analysis of absorption lines in quasar spectra an esoteric subject. It has developed into a powerful tool to be used in the study of GALAXY EVOLUTION (e.g. similar to imaging the stellar components of the galaxies). At least in principle, quasar spectra can be used for an unbiased study of the gaseous environments of galaxies from the present back to the highest redshifts at which quasars are observed. Gas structures smaller than $1M_\odot$ can be detected if they are intercepted by the quasar line of sight, irrespective of whether they emit light. Through the tool of quasar absorption lines, protogalactic structures and low surface brightness galaxies can be studied as well as high-luminosity galaxies.

Developing physical intuition

With high-resolution spectra of quasars, it is possible to consider the physical conditions of the gaseous structures that produce absorption. However, it is challenging to separate the various effects that 'shape' the spectral features in the different chemical transitions. The absorption profiles observed for the different chemical transitions are determined by a combination of the spatial distribution of material along the line of sight, its bulk kinematics, temperature, metallicity and abundance pattern. The ionization structure is influenced by gas densities and by the UV radiation field, which is a combination of the extragalactic background radiation due to the accumulated effect of quasars and stellar photons escaped from galaxies (and corrected for absorption by the intergalactic medium).

The shape of an absorption line can be modeled with a Voigt profile, which is a combination of the natural, quantum mechanical Lorentzian broadening and the Gaussian broadening caused by the thermal and turbulent motions in the gas. Several Voigt profiles can be blended together to form an overall complex absorption feature (see figure 4). The 'width' of a single Voigt profile is characterized by the Doppler parameter, b (expressed in velocity units and related to the Gaussian σ by $b = 2^{1/2}\sigma$). Physically, the Doppler parameter is the sum of thermal and turbulent components, $b_{\text{tot}}^2 = 2kT/m + b_{\text{turb}}^2$, where T is the temperature of the gas, and m is the mass of an atom.

Kinematic models

Two of the simplest types of organized kinematics in galaxies are illustrated in figure 5: clouds distributed in a rotating disk, and radial infall of clouds in a spherical distribution. Here, Mg II absorbers are used as an example, but the same kinematic arguments would apply to other transitions. For radial infall, clouds can be distributed over the range of velocities, with a tendency for a 'double peak' from material that is redshifted and blueshifted but with a considerable amount of variation if there are typically several discrete clouds along the line of sight. A rotating disk with a vertical velocity dispersion characteristic of a spiral galaxy disk (10–20 km s^{-1}) will have clouds superimposed in velocity space, and an overall kinematic spread of tens of km s^{-1}. Strong Mg II absorption has been found to arise along nearly all lines of within $\sim$40 kpc of normal galaxies (i.e. the covering factor is nearly unity within that radius). The large variety of kinematics evident in Mg II absorption profiles is, in fact, consistent with a superposition of disk and radial infall (halo) motions, and not with just one or the other. In addition to these simple, toy models, insights can be gleaned by passing lines of sight through the structures in cosmological N-body/hydrodynamic simulations. In a few studies, metals

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

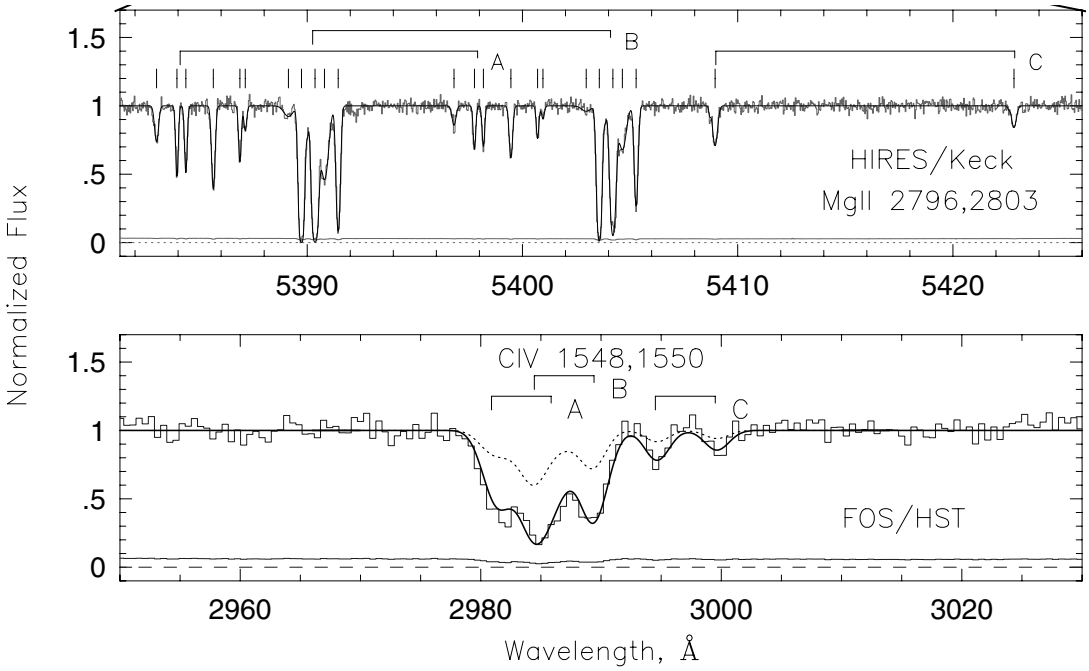


Figure 4. Dramatic demonstration of gains due to high-resolution spectroscopy of the Mg II doublet. The top panel is an $R = 3000$ spectrum of PG1206 + 459. The doublet that is apparent at an observed wavelength of ∼5400 Å is due to Mg II absorption from a system at $z = 0.927$. The middle panel shows the remarkable kinematic structure that is revealed at the resolution ($R = 45\,000$) of the Keck/HIRES spectrograph of the same quasar. The 2796 Å transition is resolved into multiple components (5583–5592 Å), which also appear in the 2803 Å transition (5396–5406 Å). This system can be separated in two 'clusters' of clouds, labeled 'A' and 'B'. Another weaker Mg II doublet is observed at 5409 and 5423 Å, from a system at $z = 0.934$ Å, labeled with a 'C'. The solid line through these complex Mg II profiles is the result of multiple Voigt profile fitting, with a cloud centered on each of the ticks drawn above the spectrum. The lower panel shows the C IV doublets associated with the same three systems, observed with the Faint Object Spectrograph on HST, but at much lower resolution ($R = 1300$). The C IV is in three different concentrations around the three systems 'A', 'B' and 'C'. The C IVλ1550 transition from system A is blended with the C IVλ1548 transition from system B. The C IV equivalent width is too large for this absorption to be produced by the same phase of gas that produces the Mg II cloud absorption. The maximum absorption that can arise in the Mg II phase is given by the dotted line; a plausible model with a kinematically broader C IV phase yields the solid curve.

have been added uniformly throughout the simulation box and photoionization models used to predict the absorption expected from different structures. This is especially important for establishing the kinematics that would be observed from the process of structure formation at high redshifts.

Photoionization models

Consider a cloud of material, modeled by a plane parallel slab with a certain total column density of hydrogen, $N(\mathrm{H}) = N(\mathrm{H\,I}) + N(\mathrm{H\,II})$, and with a constant total number density $n_{\mathrm{H}} = n(\mathrm{H\,I}) + n(\mathrm{H\,II})$ along the line of sight. The cloud is also characterized by its metallicity, Z, which is the ratio of Fe/H expressed relative to the solar value, $Z_{\odot}$, and by an abundance pattern (the abundance ratios of all other elements to Fe). The degree of ionization in the gas depends upon the intensity and shape of the spectrum of ionizing radiation. The intensity is characterized by the ionization parameter, $U = n_{\gamma}/n_{\mathrm{H}}$, which is the ratio of the number density of photons at the Lyman edge to the number density of hydrogen ($n_{\mathrm{H}} = n_e$, where n_e is the total number density of electrons). The larger the value of U, the more ionized the gas. Collisional ionization can also be an important process for some absorption systems with gas at high temperatures (hundreds of thousands of degrees). Photoionization equilibrium models typically yield temperatures of tens of thousands of degrees.

Once the metallicity, abundance pattern, ionization parameter and spectral shape are specified the equations of RADIATIVE TRANSFER can be solved to find the column densities of all the different ionization states of various chemical elements. Figure 6 illustrates, for $N(\mathrm{H\,I}) = 10^{16}$ and 10^{19} cm^{-2}, the dependence of column densities of various transitions on the ionization parameter, U. For optically thin gas ($N(\mathrm{H\,I}) < 10^{17.2}$ cm^{-2}), the column density ratios of the various metal transitions are not dependent on the overall metallicity, i.e. the curves shift vertically in proportion to Z. For optically thick gas, ionization structure develops, with an outer ionized layer around a neutral core, and there is no simple scaling relation with metallicity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

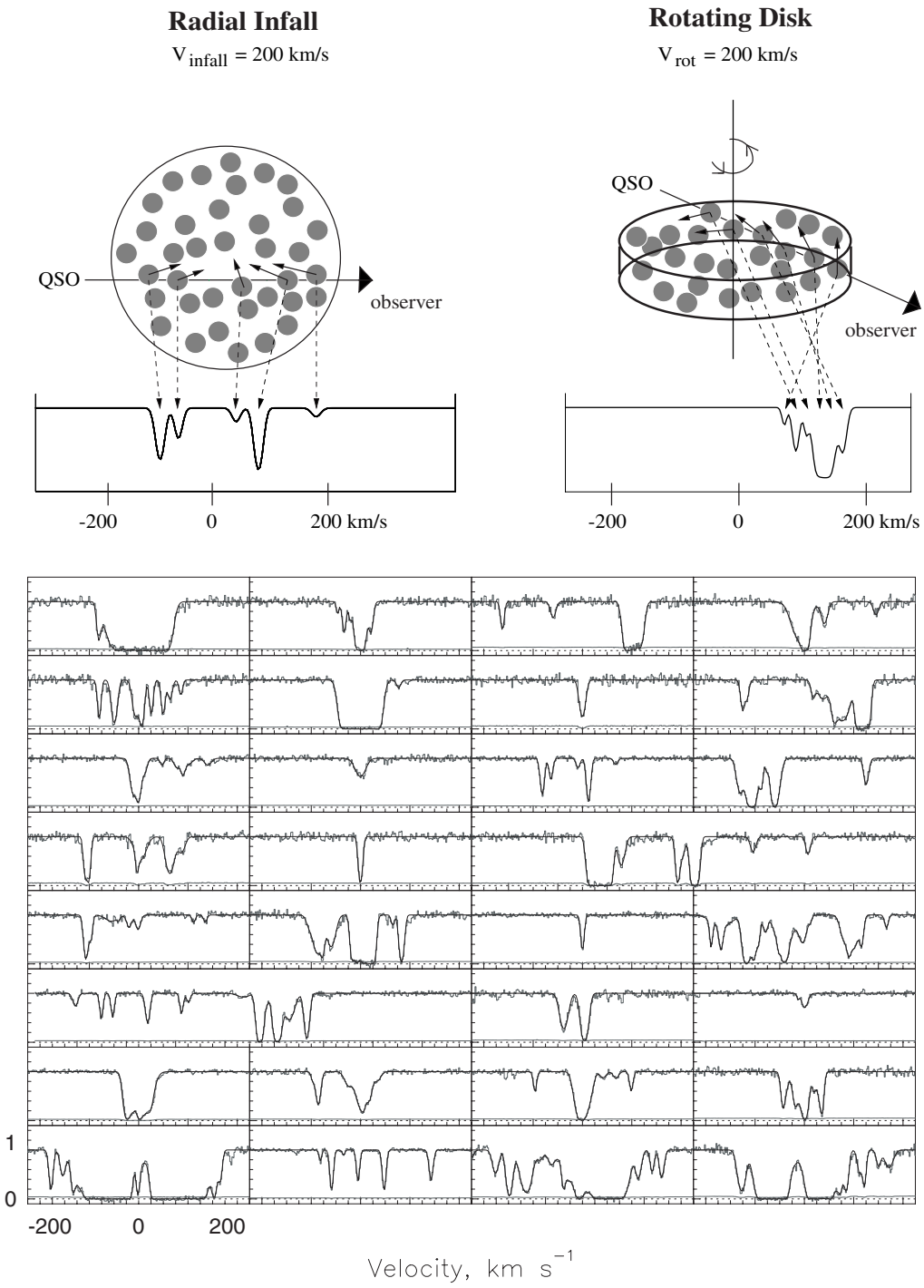


Figure 5. Illustrations of two simple kinematic models are shown in the top panel. To the left, the model is radial infall of clouds to the center of a sphere, with constant velocity. The line of sight passes through five clouds, which leads to five different absorption features (for a single transition) in the quasar spectrum. Two of the features are blueshifted relative to the standard of rest of the absorbing galaxy, and the other three are redshifted. The absorption features from a radial infall model can be spread over a velocity of 100–200 km s^{-1}, typical of the velocity dispersion of a galaxy halo. To the right, a rotating disk model is illustrated. In this case all the 'clouds' along the line of sight have a component of motion that is redshifted, and they tend to be clustered together in velocity space, with a typical spread of 20–60 km s^{-1}. The lower panel shows a sample of $0.4 < z < 1.4$ Mg II absorption profiles observed with the Keck/HIRES spectrograph at $R = 45\,000$, corresponding to a resolution of ∼6 km s^{-1}. The solid lines through these data are Voigt profile fits and the ticks drawn above the spectrum represent the cloud velocities. Some of these profiles are consistent with the kinematics of a rotating disk, and others with radial infall kinematics. However, to explain the full ensemble of profiles a model combining these two basic types of kinematics is needed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In practice, if we assume that a cloud has a simple, single phase structure, the ratios of the column densities can be used to infer the ionization parameter, which relates to the density of the gas. However, the abundance pattern can differ from the solar abundance pattern because of differing degrees of depletion onto dust, or because of different processing histories. Most of the so-called α particle nuclei (such as Mg and Si) are synthesized primarily by Type II SUPERNOVAE during the early history of a galaxy when most massive stars form and quickly evolve to reach their end states. On the other hand, the Fe group elements are primarily produced by Type Ia supernovae, and therefore build up over a longer timescale. In the basic picture of galactic evolution, the halo stars are formed early, have been enriched only by Type II supernova, and therefore are α-element enhanced. Younger disk stars have also incorporated the Type Ia processed material and therefore have relatively larger Fe group abundances. Ideally, several different ionization states of the same chemical element are observed so that there is no ambiguity between the ionization parameter and the abundance pattern, but this has generally not yet been possible because of limited wavelength coverage at high resolution.

Examples of the variation of column density ratios with velocity in two absorption systems are shown in figures 7 and 8. In figure 7, N(Fe II)/N(Mg II) varies by an order of magnitude over the four components in the $z = 1.325$ system toward the quasar Q0117 + 213. This represents a variation of an order of magnitude in the ionization parameter ($10^{-4} < U < 10^{-3}$), or an order of magnitude variation in the abundance pattern. Figure 8 is a very unusual system with two clouds separated by only 20 km s^{-1} in velocity, one of which has a silicon to aluminum ratio similar to the Milky Way ISM, and the other which requires a significant enhancement of aluminum.

Multiphase conditions

The gaseous component of the the Milky Way and nearby galaxies has a phase structure (i.e. spatial locations with different densities and/or temperatures). Examples are the disk/halo interface (galactic coronas) and the cold, warm and hot phases of the interstellar medium. From photoionization models, it is not usually possible to generate absorption that is simultaneously consistent with all observed chemical transitions for a given system. For example, in single cloud Mg II systems, figure 6 (with N(H I) = 10^{16} cm^{-2}) shows that if Fe II is detected at a similar column density to Mg II, the ionization parameter must be small, and W_r(C IV) cannot be large. Many systems have C IV absorption which exceeds this limit and requires a higher ionization (lower density) phase; generally, this phase must have structure over a large velocity range (a large 'effective' Doppler parameter). The $z = 0.93$ system toward the quasar PG 1206+459 is another case that requires multiphase structure. The observed C IV profile in figure 4 is much too strong for this absorption to arise in the same clouds that produce the Mg II, even if their ionization parameters are pushed to the largest values consistent with the data.

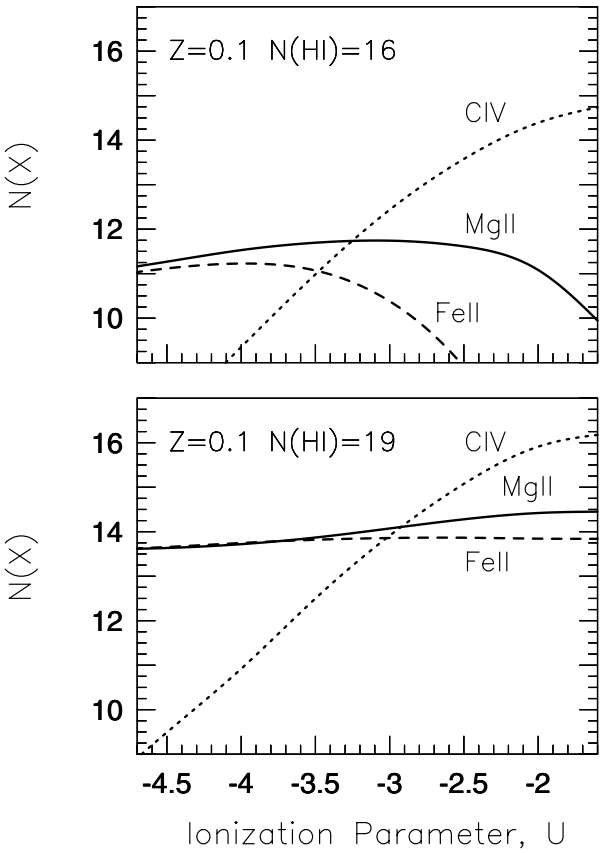


Figure 6. Photoionization model predictions of the column densities of Mg II, Fe II and C IV as a function of the ionization parameter (the ratio of ionizing photons to the electron number density in the gas). The spectrum incident on the cloud, represented by a constant density slab, is the 'Haardt–Madau' spectrum (attenuated spectrum due to the integrated effect of quasars and young galaxies). The predicted column densities are presented in two series of models with N(H I) = 10^{16} cm^{-2} and with N(H I) = 10^{19} cm^{-2}, the optically thin and optically thick cases. For both, the metallicity is fixed at 10% of the solar value. For the optically thin case, the column densities scale with metallicity, i.e. the ratios remain constant, but for the optically thick case the situation is more complex.

Statistics, evolution and interpretation

Future quasar absorption line studies will combine insights gained from detailed analyses of individual systems with conclusions drawn from the large statistical samples assembled over cosmic time. Evolution of the ensemble of absorption profiles generated by the universal collective of intervening structures is a result of the combined effects of numerous processes. These include growth of structure, star formation, morphological evolution of galaxies, galaxy mergers and changes in the extragalactic background radiation. Here, we summarize

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

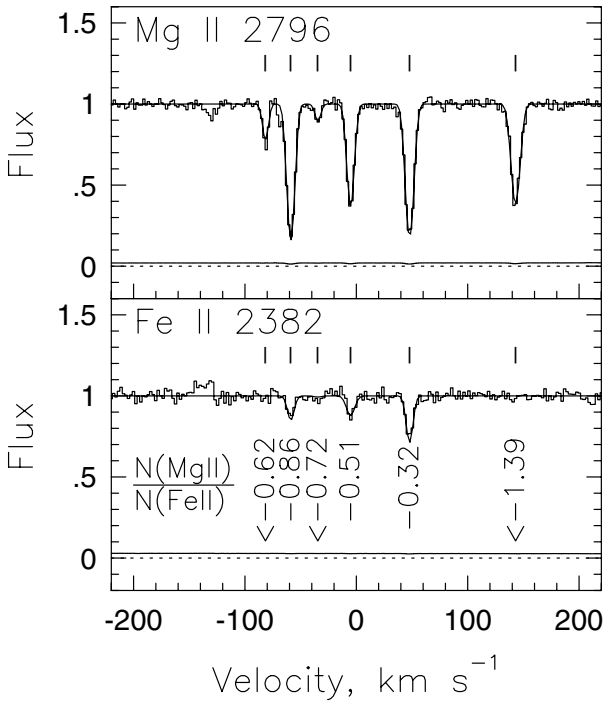


Figure 7. HIRES/Keck Fe II and Mg II absorption profiles for the $z = 1.325$ system in the spectrum of the quasar Q0117 + 213. The six clouds in this system show a range of more than an order of magnitude in N(Fe II)/N(Mg II), given below each cloud in the lower panel. These variations could be due to cloud to cloud variations of ionization parameter (density) or of abundance pattern within the system.

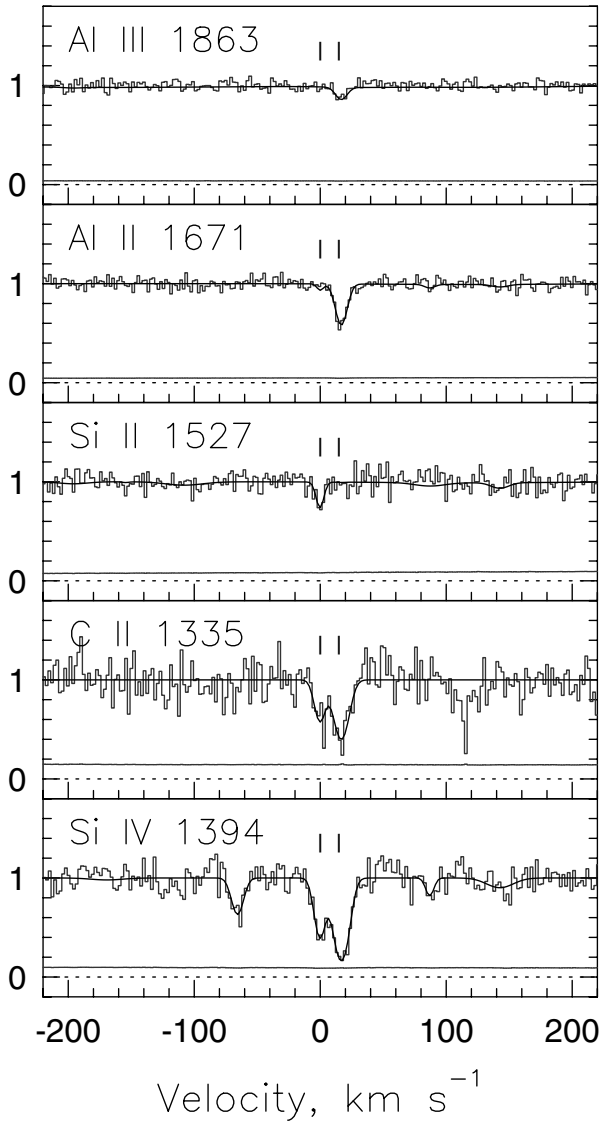


Figure 8. An unusual aluminum-rich cloud is apparent in the $z = 1.93$ system toward the quasar Q1222 + 228, and it is close in velocity space to a normal (relative to galactic clouds) cloud which has detected Si II. Note the different kinematic structure in the higher ionization transitions. The excess of Al II and Al III in the cloud at $v = 9$ km s^{-1} is best explained by an abundance pattern variation, since Si II and Al II are transitions with very similar ionization states.

the best present statistical data and likely interpretations for the different classes of absorbers. The number of lines per unit redshift for various populations of absorbers is represented by a power law $\mathrm{d}N/\mathrm{d}z \propto (1+z)^{\gamma}$. For a universe with only the cosmological evolution due to expansion, $\gamma = 1.0$ for deceleration parameter $q_0 = 0$ and $\gamma = 0.5$ for $q_0 = 0.5$.

Ly α *forest*

The Ly α forest evolves away dramatically from high to low redshift, as is strikingly clear from the spectra of $z \sim 3$ and $z \sim 1$ quasars in figure 9. The evolution of the Ly α lines with W_r(Ly α) > 0.3 Å can be characterized by a double power law with $\gamma \sim 2$ for $1.8 < z < 4.5$ and $\gamma \sim 0.2$ for $z < 1.8$. Help in understanding the physical picture has come from sophisticated N-body/hydrodynamic simulations that incorporate the gas physics and consider cosmological expansion of the simulation box. The dynamical evolution of the H I gas can be described as outflow from the centers of voids to their surrounding shells, and flows along these sheets toward their intersections where the densest structures form. This picture is consistent with observational determinations of the 'sizes' of Ly α structures. It is difficult to obtain direct measurements of sizes except in some special cases to use 'double lines of sight', close quasar pairs, either physical or apparent due to gravitational lensing. If the spectra of the two quasars both have a Ly α absorption line at the same wavelength that implies a 'structure' which covers both lines of sight. From these studies, it is found that 'structures' are at least hundreds of kpc in extent.

At redshifts $z = 5$ to $z = 2$ $\mathrm{d}N/\mathrm{d}z$ for Ly α forest absorption is quite large, but it is declining very rapidly over that range. This dramatic evolution in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

number of forest clouds is mostly due to the expansion of the universe, with a modest contribution from structure growth. At $z < 2$, the extragalactic background radiation field is falling, and Ly α structures are becoming more neutral. Therefore, the more numerous, smaller N(H) structures are observed at a larger N(H I) and this will counteract the effect of expansion, thus slowing the decline of the forest.

The high-redshift Ly α forest was once thought to be primordial material, but in fact it is observed to have a metallicity of 0.1% solar, even at $z = 3$. For $N(\text{H I}) < 10^{14}$ cm^{-2}, the expected N(C IV) would be below the detection thresholds of current observations, so truly pristine material still eludes us. Perhaps it does not exist. To spread metals all through the intergalactic medium may have required a 'pre-galactic' population of stars at $z > 10$ that polluted all of intergalactic space.

Lyman limit and metal line systems

The dN/dz of Lyman limit systems is consistent with that of strong Mg II absorbers (with $W_r(\text{Mg II}) > 0.3$ Å) over the redshift range for which both have been observed, $0.4 < z < 2.2$. For $W_r(\text{Mg II}) > 0.3$ Å, $\gamma = 1.0 \pm 0.1$, consistent with no evolution. For even stronger Mg II systems ($W_r(\text{Mg II}) > 1$ Å), dN/dz increases more dramatically with z, with $\gamma = 2.3 \pm 1.0$.

The number of Mg II systems (equivalent width distribution) continues to increase down to the sensitivity of the best surveys, $W_r(\text{Mg II}) > 0.02$ Å, such that $dN/dz = 2.7 \pm 0.15$ at $z \sim 1$. The 'weak' Mg II absorbers are therefore more common than the strong systems ($W_r(\text{Mg II}) > 0.3$ Å) known to be associated with luminous galaxies. Unlike the strong Mg II absorbers, the weak Mg II absorbers are sub-Lyman limit systems (they do not have Lyman limit breaks), and no galaxies have been identified at the redshift of absorption. Yet, photoionization models indicate that the metallicities of these weak absorbers are at least 10% of the solar value, and in some cases comparable to solar. They are a varied population: some have relatively strong Fe II while others have no Fe II detected, and some have strong C IV that requires a separate phase while others have no C IV detected. Those with strong Fe II are constrained to be smaller than 10 pc (the ionization parameter must be small and n_e large, as can be seen in figure 6). Also, since Fe is produced primarily by Type Ia supernovae they must be enriched by a relatively old STELLAR POPULATION. Those with weaker or undetected Fe II could be larger (kpcs or tens of kpcs) and possibly enriched by Type II supernovae. Candidate environments that could be traced by weak Mg II absorption are: remnants of pre-galactic star clusters formed in mini-haloes at $z > 10$, super star clusters formed in interactions, tidally stripped material, low surface brightness galaxies and ejected or infalling clouds (analogous to the Milky Way high-velocity clouds).

The evolution of dN/dz for C IV absorbers can be studied in the optical for high redshifts. For $W(\text{C IV}) > 0.4$ Å and $z > 1.2$, the number decreases with increasing z, as $\gamma = -2.4 \pm 0.8$. In this same interval, the number of Lyman limit systems is still increasing with redshift, with $\gamma = 1.5 \pm 0.4$. This implies that the dramatic evolution in the number of C IV systems is either due to a change in metallicity or a change in ionization state. The dN/dz for C IV systems peaks at intermediate z and declines, consistent with no evolution, until the present. Combining optical and UV data, C IV and Mg II have been compared at $0.4 < z < 2.2$. The fraction of systems with large $W_r(\text{C IV})/W_r(\text{Mg II})$ decreases rapidly with decreasing redshift; there is a shift toward 'lower ionization systems'.

It is important to consider that the H I, Mg II and C IV absorptions do not always arise in the same phase. It is possible that the C IV in many $z \sim 1$ Mg II absorption systems arises in a phase similar to the galactic coronas. If the origin of this phase is related to star-forming processes in the disk, then it might be expected to diminish below $z = 1.2$ since peak star formation rate is passed.

Another important trend is the fact that the very strongest Mg II absorbers evolve away from $z = 2$ until the present. If we study the kinematic structure of these objects, we find that they commonly have a 'double' structure, with two separate kinematic regions in the Mg II profile. These objects also have strong C IV which also has separate components around the two Mg II regions in the 'double' structure. The C IV does not arise primarily in the individual Mg II clouds, nor is it in a smooth, 'common halo' structure that extends in velocity space around the entire Mg II profile. As more data are collected on the kinematic structure of various transitions in these 'double' systems, it will be interesting to consider the hypothesis that galaxy pairs in the process of merger are responsible. The number of these is thought to have been dramatically larger in the past.

Damped Ly α systems (DLAs)

The $N(\text{H I}) > 10^{20.3}$ cm^{-2} systems are of particular interest because it is possible to observe many different chemical elements (such as Zn, Cr, Fe, Mn and Ni) in these objects back to high redshift. Metallicities and abundance patterns can be studied and compared with those of old stellar populations in the Milky Way. Back to $z = 3$, the metallicity in DLAs, as measured by the undepleted element zinc, is about 10% of the solar value, but it may decline at $z > 3$. The identity of sites responsible for DLAs at high z remains controversial, but they do contain most of the neutral hydrogen in the universe, from which most of its stars form. The kinematic structure of the absorption profiles of neutral and low ionization species is consistent with the rotation of a thick disk, so that it is possible that these are the $z = 3$ progenitors of normal spiral galaxies. However, this signature is not unique. It could also be the consequence of directed infall in an hierarchical structure formation scenario. The higher-ionization species show complex kinematics which vary in relation to those of the lower-ionization gas; in some systems they appear to trace

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

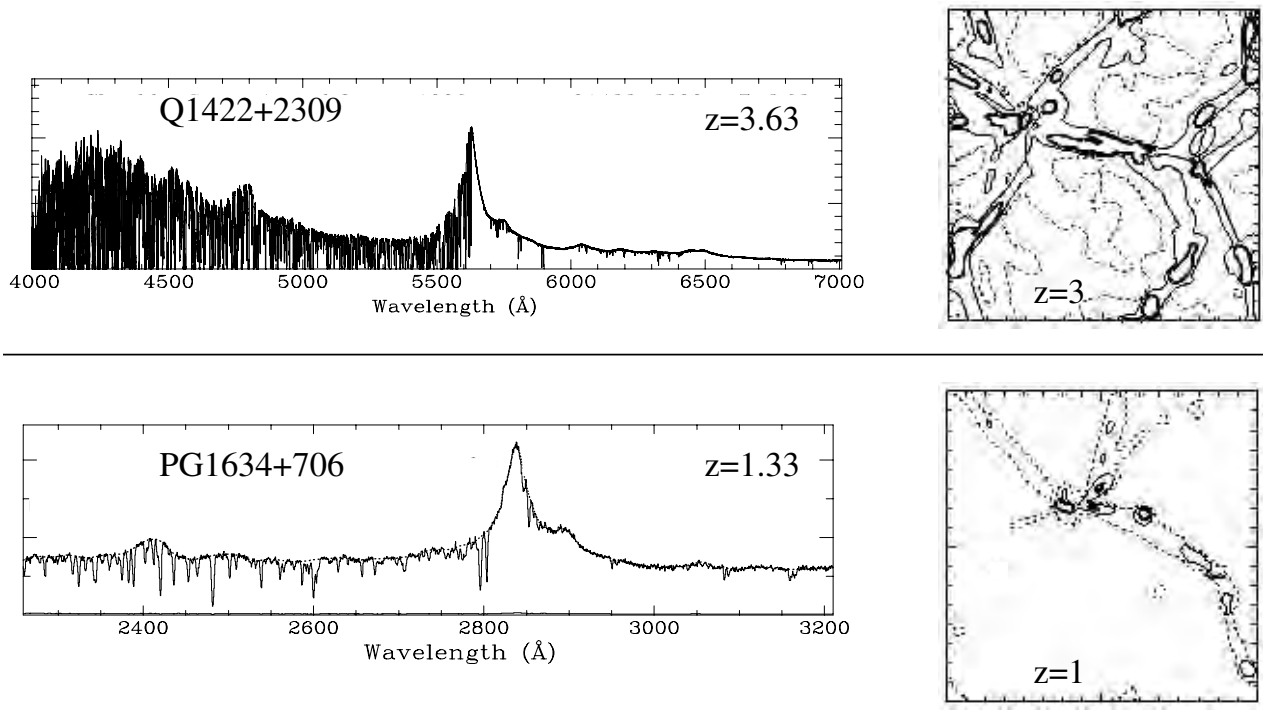


Figure 9. Illustration of structure evolution of intergalactic gas from high to low redshift. The upper spectrum of a $z = 3.6$ quasar is a Keck/HIRES observation, while the lower spectrum is a FOS/HST observation of a $z = 1.3$ quasar. Higher-redshift quasars show a much thicker forest of Ly α lines. Slices through N-body/hydrodynamic simulation results at the two epochs $z = 3$ and $z = 1$ are shown in the right-hand panels. Three contour levels are shown: 10^{11} cm^{-2} (dotted lines), 10^{12} cm^{-2} (solid lines) and 10^{13} cm^{-2} (thick solid lines). Evolution proceeds so that the voids become more empty so that even the low column density material is found in filamentary structures at low redshifts.

relatively similar structure, and in others there are clearly several different phases.

At low redshift, many of the galaxies that are responsible for the DLA absorption can be directly identified. These galaxies are a heterogeneous population. They are not just the most luminous galaxies, but include dwarf and low surface brightness galaxies, and even cases where no galaxy has been identified to sensitive limits. Damped Ly α absorption does not trace the most luminous objects, but rather it traces the largest neutral gas reservoirs. An additional selection effect may be important. The most dust-rich galaxies that have the potential to produce DLA absorption could produce enough extinction so that their background quasars will not be included in quasar surveys. In this way, the population of DLAs that are actually observed could be significantly biased against dusty galaxy hosts.

Future prospects

The next decade will see the synthesis of the various techniques for the study of galactic evolution, through their stars and through their gas. Higher-resolution quasar spectra will be obtained in the ultraviolet (with the Space Telescope Imaging Spectrograph (STIS) and with the Cosmic Origins Spectrograph (COS) on the HST, and later, hopefully, with a larger UV space telescope). It will then be possible to conduct a systematic analysis of the relationships between the different ionization species that trace the different phases of gas in $0.4 < z < 1.5$ galaxies. In this redshift regime, comparisons with the detailed morphological structure and orientations of the absorbing galaxies is possible from HST images.

Invaluable insights into the origin of quasar absorption lines have been gleaned from absorption studies of nearby galaxies, for which it is possible to directly observe the processes that are involved. Making more observations of this type will be possible by discoveries of bright quasars that fall behind nearby galaxies. The discoveries of quasars in large surveys will also include multiple lines of sight behind distant absorption line systems which can be used to produce 3D maps of the structures.

The interstellar medium of the Milky Way shows structure on sub-pc scales, and absorption features can only be resolved with resolution <1 km s^{-1}. Such a resolution will soon be available on 8 m class telescopes. This is important for separating blends and for looking for metallicity, ionization and abundance pattern gradients along the line of sight.

The key low-ionization transitions of Mg II and Fe II are shifted into the near-IR region of the spectrum for $z >$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

2.5. Very soon, near-IR quasar spectra will be obtained at relatively high resolution ($\sim$20 km s^{-1}). Also, IR-imaging, narrow-band techniques and multi-object spectroscopy in the near-IR should provide much more information about absorbing galaxies at higher redshifts. This will extend evolutionary studies back to an epoch at which formation processes may be contributing significantly to evolution.

Bibliography
Articles in review journals:

Churchill C W and Charlton J C 2000 Mg II absorbers: a review *Publ. Astron. Soc. Pacific* in press

Rauch M 1998 The Lyman alpha forest in the spectra of QSOs *Ann. Rev. Astron. Astrophys.* **36** 267

Conference proceedings:

Blades J C, Turnshek D A and Norman C 1988 *QSO Absorption Lines: Probing the Universe, Proceedings of the QSO Absorption Line Meeting (Baltimore 1987)* (Cambridge: Cambridge University Press)

Meylan G 1995 *QSO Absorption Lines: Proceedings of the ESO Workshop (Munich 1994)* (Berlin: Springer)

Petitjean P and Charlot S 1997 *Structure and Evolution of the Intergalactic Medium from QSO Absorption Lines* (Paris: Editions Frontiéres)

Jane Charlton and Chris Churchill

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasistellar Objects: Intrinsic AGN Absorption Lines

Any gaseous material along our line of sight to distant QUASARS or, more generally, ACTIVE GALACTIC NUCLEI (AGNs) will absorb light according to the amounts and range of ions present. Strong absorption lines are common in rest-frame ultraviolet (UV) spectra of AGNs due to a variety of resonant transitions, for example the Lyman series lines of neutral hydrogen (H I, most notably Lyα λ1216; see also LYMAN ALPHA ABSORPTION: THE DAMPED SYSTEMS) and high-ionization doublets like CIV $\lambda\lambda$1549, 1551 (of triply ionized carbon).

The lines are called *intrinsic* if the absorbing gas is physically related to the AGN, e.g. if the absorber resides broadly within the radius of the AGN's surrounding 'host' galaxy. Intrinsic absorption lines are thus valuable probes of the kinematics, physical conditions and elemental abundances in the gas near AGNs.

Studies of intrinsic absorbers have historically emphasized quasar broad absorption lines (BALs), which clearly identify energetic outflows from the central engines. Today we recognize a wider variety of intrinsic lines in a wider range of objects. For example, we now know that SEYFERT 1 GALAXIES (the less luminous cousins of quasars) have intrinsic absorption. We also realize that intrinsic lines can form in a range of AGN environments—from the dynamic inner regions like the BALs, to the more quiescent outer host galaxies $>$ 10 kpc away. One complicating factor is that most AGNs also have absorption lines due to unrelated gas or galaxies far from the AGN (see QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES). Part of the effort, therefore, is to identify the intrinsic lines and locate their absorbing regions relative to the AGNs. Empirical line classifications are a good starting point for this work.

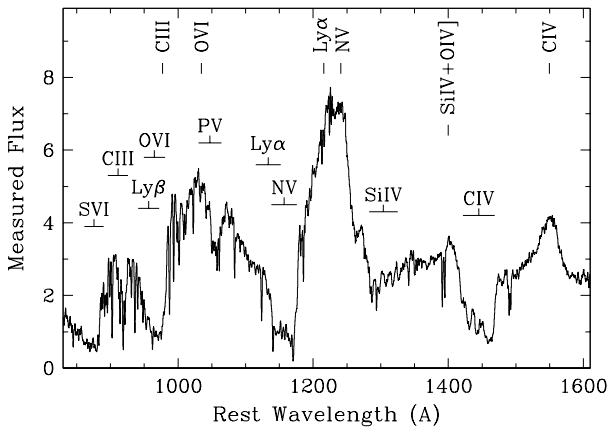


Figure 1. Detached BALs in the quasar PG 1254 + 047. The BALs are labelled just above the spectrum. The wavelengths of prominent broad emission lines are marked across the top. The flux has units 10^{-15} erg s^{-1} cm^{-2} Å^{-1}.

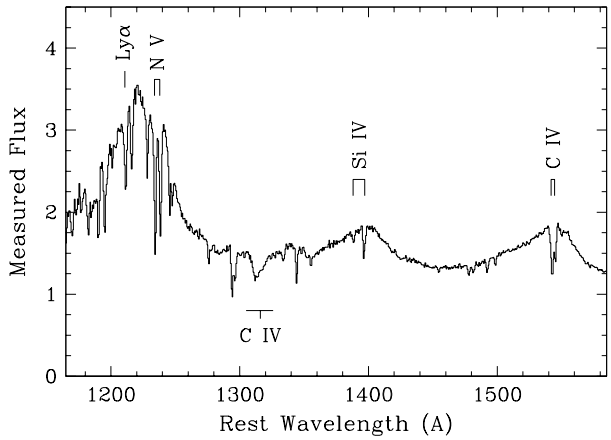


Figure 2. Intrinsic absorption in the quasar PG 0935 + 417. A system of associated NALs is labelled above the spectrum, with open brackets showing the doublet separations. A mini-BAL due to CIV blueshifted by $\sim$ 51 000 km s^{-1} is labelled below. The flux has units 10^{-15} erg s^{-1} cm^{-2} Å^{-1}.

Empirical line types

AGN absorption lines are usually classified according to the widths of their profiles. These classes separate the clearly intrinsic broad lines from the many others of uncertain origin. The main empirical classes are (1) the BALs, (2) the narrow absorption lines (NALs), and (3) the intermediate 'mini-BALs.'

Broad absorption lines (figure 1) are blueshifted relative to the AGN emission lines, implying outflow velocities from near 0 km s^{-1} to as much as $\sim$ 60 000 km s^{-1} ($\sim$ 20% of the speed of light). A representative line width is $\sim$ 10 000 km s^{-1}, although there is considerable diversity among BAL profiles. Velocity widths $\gtrsim$ 3000 km s^{-1} and blueshifted velocity extrema $\gtrsim$ 5000 km s^{-1} are usually considered minimum requirements for classification as a BAL. Some BALs have several distinct absorption troughs, while others are strictly 'detached' from the emission lines—such that the absorption appears only at blueshifts exceeding several thousand km s^{-1} .

Narrow absorption lines (figure 2) have widths less than a few hundred km s^{-1}. NALs with absorption redshifts, z_a, within $\pm$5000 km s^{-1} of the emission redshift, z_e, are called 'associated' (or $z_a \approx z_e$) lines because of their likely physical connection to the AGN. In general, NALs can appear at redshifts from $z_a \approx z_e$ to $z_a \approx 0$. Many of these systems are not related to the AGNs.

Absorption lines with intermediate widths between the BALs and NALs are increasingly called mini-BALs (figure 2). Their strictly blueshifted profiles appear smooth and BAL-like in high resolution spectra, and their centroid velocities span the same range as the BAL troughs (from near 0 km s^{-1} to almost 0.2c). Mini-BALs evidently form in outflows similar to the BALs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

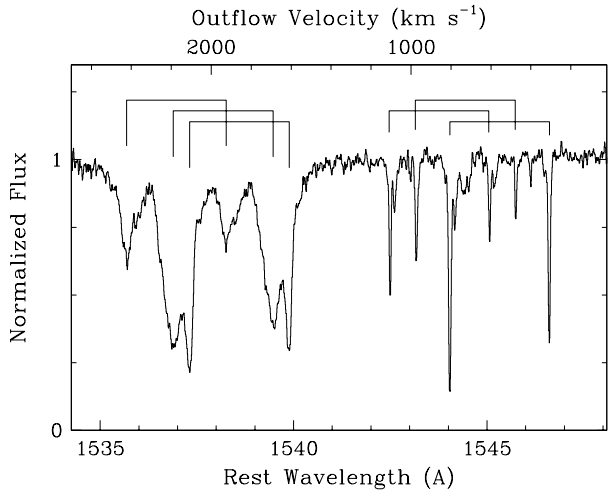


Figure 3. High-resolution blow-up of the CIV NALs in PG 0935 + 417, revealing multiple components (cf figure 2). The open brackets mark the strongest doublet pairs. The outflow velocities are appropriate for the short-wavelength doublet members. The absorption complex at 1535–1540 Å is intrinsic based on the broad line profiles and doublet ratios that imply partial line-of-sight coverage.

Identifying intrinsic NALs

The first evidence that some NALs are intrinsic came from statistical tendencies, namely, (1) quasar NALs appear with greater frequency near the emission-line redshift, and (2) the strengths of these $z_a \approx z_e$ systems correlate with the quasars' radio properties. The first tendency might be explained by external galaxies clustering around quasars, but the second clearly demonstrates a physical relationship to the quasars themselves.

Direct evidence for the intrinsic origin of specific NALs has come from spectroscopic indicators, such as (1) time-variable line strengths, (2) well-resolved line profiles that are smooth and broad compared to the thermal velocity, (3) multiplet ratios that imply partial coverage of the background light source(s), and (4) high space densities ($\gtrsim$100 cm s^{-3}) inferred from the presence of excited-state absorption lines. These properties signal intrinsic absorption because they are most easily understood in terms of the dense and dynamic environments near AGNs. Unrelated intervening absorbers—typically inter-galactic gas clouds or extended galactic halos—should generally be larger, more quiescent, and less ionized for a given gas density. The link between the first three properties and the near-quasar environment is further strengthened by the fact that they are common in BALs and mini-BALs. NALs with these characteristics probably also form in outflows; they have been measured at blueshifted (ejection) velocities from ~ 0 to $\sim$ 24 000 km s^{-1} (e.g. figure 3).

Global covering factors and the ubiquity of intrinsic absorbers

All varieties of quasars and Seyfert 1 galaxies show some type of intrinsic absorption, but different objects favor different types of absorbers. For example, BALs and other high-velocity ($\gtrsim$ 3000 km s^{-1}) intrinsic lines appear only in quasars and never in the Seyfert galaxies. BALs also tend to avoid quasars with strong radio emission, while associated NALs seem to favor them. Preferences like these are probably tied to the unique physical conditions, whereby different AGNs drive different types of outflows.

Intrinsic lines also do not appear in every individual spectrum. For example, BALs are detected in only $\sim$ 12% of quasars. The detection rate of associated (and so probably intrinsic) NALs in quasars is not well known, but it appears to be roughly similar to the BALs. Mini-BALs may be somewhat rarer than both BALs and associated NALs in quasars, but they can appear in either 'radio-loud' or 'radio-quiet' objects. Mini-BALs and intrinsic NALs are both common in Seyfert 1 galaxies, with at least one of these features appearing in 50–70% of the low-luminosity sources.

These detection rates (for any given object class) depend on viewing angle effects and on the global covering factors of the absorbing gas. If the objects are randomly oriented, the absorption-line detection frequencies should approximately equal the average value of the global covering factor, $Q \equiv \Omega/4\pi$, where Ω is the solid angle subtended by the absorber as seen from the central light source. In practice, the situation can be more complicated. For example, attenuation by BAL gas might bias quasar samples against the inclusion of sources with BALs. Thus the true global covering factor of BAL regions could be 20–30%, instead of the $\sim$ 12% implied by the detection rate. Nonetheless, the overall conclusion is that intrinsic absorbers are present in all AGNs, with global covering factors *similar* to the line detection frequencies in each type of object.

Absorber geometry

A consistent picture of the geometry has emerged in which the absorbing gas resides primarily near the equatorial plane of the AGN ACCRETION DISK (figure 4). In particular, spectropolarimetric observations show that the continuum light from BAL quasars is more polarized than non-BAL quasars. Also, the percentage polarization is typically much greater inside BAL troughs compared to the adjacent continuum. These results are understood in terms of light scattering in the disk geometry. Quasars viewed close to edge-on exhibit BALs because our sight-line intersects the the absorbing gas. These objects are also more polarized because the direct (unpolarized) light through the BAL region is slightly attenuated; thus the scattered (polarized) light makes a relatively larger contribution to their measured flux. BAL troughs have higher percentage polarizations because the direct (unpolarized) light in the troughs is more attenuated than the continuum. Quasars without BALs, viewed from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

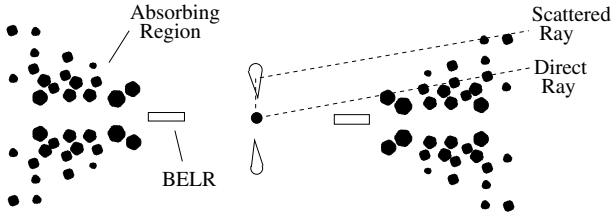


Figure 4. Schematic absorber geometry. The dotted lines are representative light rays from the continuum source (central dot). The open rectangles mark the broad emission line region (BELR) near the accretion disk plane. The teardrop lobes above and below are putative scattering regions near the disk axis. Extended radio jets, when present, would lie along this axis. The radius of the BELR (equation (3) below) sets the scale.

above or below the disk plane, have low polarizations because their direct (unpolarized) light dominates the measured flux.

This picture is also supported by observations showing that quasars with reddened spectra, presumably caused by dust near the disk plane, are more likely to have BALs and high polarizations. Similarly, studies of radio-loud quasars show that associated NALs are stronger and more common in sources with their radio jets aligned perpendicular to our line of sight (such that their unresolved inner disks are viewed nearly edge-on).

Basic physical properties

Kinematics

One surprising property of intrinsic absorbers is that none of them have changed velocity between observations that now span 10–20 years. In one well-studied case, distinct features in a quasar BAL show < 30 km s^{-1} of movement over 5 years in the quasar rest frame, implying an acceleration of < 0.02 cm s^{-2}. The outflow speeds of 10 000–20 000 km s^{-1} are therefore stable to $\lesssim 0.2\%$ on this time scale.

Another general property is the kinematic complexity (e.g. figures 1–3). Many intrinsic absorbers have multiple distinct absorption troughs. Sometimes BALs, mini-BALs and/or NALs appear together in the same spectrum, either blended together or at different (non-overlapping) outflow velocities. The narrower lines, i.e. the NALs and mini-BALs, generally have small velocity dispersions compared to the outflow speeds. Evidently, the outflows producing these lines are not continuously accelerated from rest along our line of sight. These properties all present challenges for the physical models discussed below.

Column densities and partial coverage

Estimates of the column densities in each absorbing ion are complicated by the fact that many (most?) intrinsic absorbers only partially cover the background light source(s) along our line(s) of sight. Partial coverage might occur if the absorbing regions are porous or they are smaller than the emission sources in overall projected area (see figure 4). In any case, partial coverage leads to unabsorbed flux filling in the bottoms of the measured troughs. The observed line intensities thus depend on both the line-of-sight coverage fraction, $C_{\rm f}$, and the line optical depths as

$$I_v \;=\; (1 - C_{\rm f})\, I_o \;+\; C_{\rm f}\, I_o\, {\rm e}^{-\tau_v} \qquad (1)$$

where $0 \leqslant C_{\rm f} \leqslant 1$, I_v and I_o are the observed and emitted (unabsorbed) intensities, and τ_v is the line optical depth, at each line velocity v. The first term on the right-hand side represents the unabsorbed light that fills in the troughs. In the limit $\tau_v \gg 1$ we have simply,

$$C_{\rm f} = 1 - \frac{I_v}{I_o}$$

Outside of that limit, we can compare lines whose true optical depth ratios are fixed by atomic constants, such as the HI Lyman lines or doublets like CIV $\lambda\lambda$1548, 1550, to determine uniquely both the coverage fractions and the true optical depths across the line profiles. For the commonly measured doublets (like CIV), where the true optical depth ratios are ~ 2, the coverage fraction at each velocity follows from equation (1) such that,

$$C_{\rm f} \;=\; \frac{I_1^2 - 2I_1 + 1}{I_2 - 2I_1 + 1} \qquad (2)$$

where I_1 and I_2 are the observed line intensities, normalized by I_o, at the same velocity in the weaker and stronger line troughs, respectively. The column densities, N, follow from the line optical depths via,

$$N = \frac{m_e c}{\pi e^2 f \lambda_o} \int \tau_v {\rm d}v$$

where f and λ_o are the line's oscillator strength and laboratory wavelength. The integral is performed over all or part of the line profile.

This analysis has been applied to well-resolved multiplets in NALs and mini-BALs. The derived coverage fractions range from $\sim 10\%$ to 100%. The column densities in commonly measured ions (e.g. H I, C IV, N V, Si IV, etc) are usually in the range $10^{13} \lesssim N \lesssim 10^{16}$ cm s^{-2}, consistent with the usual absence of Lyman edge absorption at 912 Å. Most BALs are too broad and blended for this analysis. They probably have $C{\rm f} < 1$ in general, but the exact values are rarely known. Crude estimates suggest that the range of BAL column densities is perhaps an order of magnitude larger than the well-measured NALs.

Ionization and total column densities

Intrinsic absorbers are in general highly ionized. Their lowest levels of ionization are usually characterized by ions such as C III, N III, Si IV or C IV. Lower ionization species like Si II, Fe II, Mg II or C II are rarely present. The upper limits of ionization are generally not known because higher ions have their resonance lines at shorter (and more difficult-to-measure) wavelengths. Existing data suggest that ionizations up to at least Ne VIII are common.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The strong radiative flux of the AGN provides a natural ionization source. Theoretical simulations of photoionized plasmas in equilibrium with the AGN radiation field are used to match the measured column densities, and thus derive the overall ionization state(s) and total column densities (in H I + H II). The ionization state is described by an ionization parameter, U, which is the dimensionless ratio of hydrogen-ionizing photon to hydrogen particle densities in the absorber:

$$U \equiv \frac{1}{4\pi R^2 c n_{\rm H}} \int_{\nu_{\rm LL}}^{\infty} \frac{L_\nu}{h\nu} {\rm d}\nu$$

where $n_{\rm H}$ is the absorber's hydrogen density, R is its radial distance from the central quasar, L_ν is the quasar luminosity distribution and $\nu_{\rm LL}$ is the frequency at the H I Lyman limit. For a 'typical' AGN spectral shape (with $L_\nu \propto \nu^{-1.5}$ in the Lyman continuum) we have

$$U \approx 0.3\, n_{10}^{-1}\, R_{0.1}^{-2}\, L_{46} \qquad (3)$$

where n_{10} is the density in units of 10^{10} cm s^{-3}, $R_{0.1}$ is the radial distance in units of 0.1 pc, and L_{46} is the AGN luminosity relative to 10^{46} erg s^{-1}. Typical U values are between ~ 0.01 and ~ 1. Absorbers at the same velocity, or at different velocities in the same spectrum, often have a range of U values implying a range of densities or radii in the overall absorbing region. Derived total column densities are usually $N_{\rm H} \approx 10^{19}$ to 10^{21} cm s^{-2} for NAL regions and $\gtrsim 10^{22}$ cm s^{-2} for the BAL gas.

Correlated x-ray absorption

Recent studies have shown that the presence of intrinsic lines in the UV correlates with continuous (bound-free) absorption in soft x-rays. Evidently, the UV lines are just one aspect of the intrinsic absorber phenomenon. The x-ray results are important because they reveal generally higher ionizations, $U > 1$, and much higher total column densities, by 1–2 orders of magnitude, compared to the UV lines. A critical issue, therefore, is the physical relationship between the UV and x-ray absorbers.

The data clearly show that different absorbers (of either type) can have a wide range of physical conditions. Some well-measured UV absorbers, in particular, have ionizations and total column densities that should produce only minimal x-ray absorption. Moreover, the best ensemble data are often inconsistent with a 1-zone medium producing all of the measured features. Spatially distinct zones with different ionizations and column densities are at least sometimes present. It is therefore likely that the UV and x-ray absorbers are physically related but not, in general, identically the same.

Space densities and radial distances

Most constraints on the radial distance come indirectly from estimates of the space density and ionization parameter (via equation (2)). The densities are derived in several ways. For example, the absence of broad or blueshifted forbidden emission lines, e.g. [O III] λ5007, suggests (for BAL regions only) that this emission is suppressed by collisional de-excitation at high densities. Some AGNs have absorption lines from excited energy states, implying significant densities to support the upper level populations. In other cases, time-variable absorption lines require minimum densities to allow the gas to adjust its ionization structure within the variability time. This last method assumes that the variability time exceeds the recombination time, $t_{\rm recomb} \approx (n_{\rm e}\alpha)^{-1}$, where $n_{\rm e}$ is the electron density and α is the recombination rate coefficient. Even if the line variability is not caused by changes in the ionization, it turns out that dynamical limits on clouds moving across our line of sight lead to similar R constraints.

The overall results imply densities from $\lesssim 7$ cm s^{-3} to $> 10^6$ cm s^{-3} and corresponding distances from $\gtrsim 300$ kpc to $\lesssim 10$ pc in different sources. The largest confirmed distances apply only to narrow and non-variable NALs. Lines with clear dynamic signatures, e.g. the BALs, mini-BALs and some NALs, probably form closer to the AGN than even the smallest of these upper limits. The minimum radial distance, $R_{\rm min}$, is set by the fact that many intrinsic absorbers, e.g. most BALs, suppress both the continuum and broad line emissions. The absorber distances therefore cannot be much less than the broad emission line region radius, $R_{\rm BELR}$, which is known independently to scale with AGN luminosity, such that

$$R_{\rm min} \approx R_{\rm BELR} \approx 0.1\ (L_{46})^{1/2}\ \ {\rm pc} \qquad (4)$$

where $L_{46} \approx 1$ is 'typical' for quasars. (Values of L_{46} can actually range from $\lesssim 0.001$ in Seyfert nuclei to > 100 in the most luminous quasars.) If there is absorbing gas near this minimum radius, the densities could reach $\sim 10^{10}$ cm s^{-3} (see equation (2)).

Physical models

Most physical models of intrinsic absorbers feature a wind-disk geometry similar to figure 4. The outflowing gas stays close to the accretion-disk plane, and we observe it (via blueshifted absorption lines) only if our sight line(s) to the emission source(s) intersect wind material. The disk provides a likely source for the wind material, and its acceleration to high speeds probably occurs via radiation pressure from the central emission source. The outward transfer of angular momentum in these winds might facilitate the *inward* flow of matter in the accretion disk, thus promoting the growth and fueling of the central black hole.

The measured line velocities suggest that the outflows often originate near the minimum radius implied by equation (3). In particular, the radial acceleration of a wind driven by radiation pressure from a central point source with luminosity L and mass M is

$$\frac{v\,{\rm d}v}{{\rm d}R} = \frac{f_L L}{4\pi R^2 c m_{\rm H} N_{\rm H}} - \frac{GM}{R^2}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where f_L is the fraction of the luminosity absorbed or scattered in the wind. Integrating this equation from R to infinity yields the terminal velocity,

$$v_\infty \approx 10\,000\, R_{0.1}^{-1/2} \left(\frac{f_{0.1} L_{46}}{N_{22}} - 0.8 M_9 \right)^{1/2} \quad \mathrm{km\ s^{-1}} \quad (5)$$

where N_{22} is the total column density in 10^{22} cm s^{-2}, M_9 is the central black hole mass relative to 10^9 $M_\odot$, and $f_{0.1}$ is the absorbed fraction compared to 10%. These expressions hold strictly for open geometries, where the photons escaping one location in the wind are not scattered or absorbed in another location. Estimates of the total absorption indicate that $f_{0.1}$ could be as large as a few for BAL flows, and proportionately less for the narrower mini-BALs and NALs. Estimates of N_{22} range from $\lesssim 0.01$ for some intrinsic NALs to $\gtrsim 1$ for typical BALs. Radiative acceleration therefore requires small radii—very roughly similar to the radius of the broad emission line region.

The radial scale is important for defining the mass and kinetic energy of outflowing gas. The total wind mass at any instant is,

$$M_w \approx 1.1\, Q_{0.1}\, N_{22}\, R_{0.1}^2 \quad M_\odot$$

where $Q_{0.1}$ is the global covering factor relative to 10%. The mass loss rate, $\dot{M}_w$, depends further on the characteristic flow time, $v/\Delta R$, such that,

$$\dot{M}_w \approx 0.11\, Q_{0.1}\, N_{22}\, \frac{R_{0.1}^2}{\Delta R_{0.1}}\, v_4 \quad M_\odot\ \mathrm{yr^{-1}}$$

where v_4 is the flow velocity in units of 10^4 km s^{-1} and $\Delta R_{0.1}$ is its radial thickness in units of 0.1 pc. The kinetic energy luminosity is $L_\mathrm{K} \approx \frac{1}{2}\dot{M}_w v^2$, or,

$$L_K \approx 4 \times 10^{42}\, Q_{0.1}\, N_{22}\, \frac{R_{0.1}^2}{\Delta R_{0.1}}\, v_4^3 \quad \mathrm{ergs\ s^{-1}}$$

During a quasar's lifetime, perhaps $\sim 10^8$ years, outflows with these parameter values will eject a total of $\sim 10^7$ $M_\odot$ of gas with kinetic energy $\sim 10^{58}$ ergs. BAL winds might often have total masses, kinetic energies, etc, that are an order of magnitude larger, based on our best estimates of $Q_{0.1} = 1$–3 and $N_{22} \gtrsim 1$ for these outflows.

One unresolved issue is whether the outflowing gas is smoothly distributed or residing in discrete clouds. This seemingly simple question goes to the heart of the wind physics. It was long believed that the flows consist of many discrete clouds because, for example, a flow with $N_{22} \approx 1$ and $\Delta R_{0.1} \approx R_{0.1} \approx 1$ would have a mean density of only $n \approx 3 \times 10^4$ cm s^{-3} . This density would lead to U values several orders of magnitude larger than expected from the data (equation (2)). Moreover, a gas with this high ionization cannot be radiatively accelerated because the ions would be too stripped of electrons to absorb enough incident flux. The flows must therefore be distributed in much denser clouds that fill only part of the wind volume. If these clouds individually have velocity dispersions close to the sound/thermal speed (roughly 15 km s^{-1} for a nominal 15 000 K gas), then a smooth BAL profile of width 10^4 km s^{-1} requires $\gtrsim 1000$ clouds along the line of sight. The main objection to this scenario is that the clouds must be very small, $\lesssim 10^9$ cm across, and they cannot survive as discrete entities without some *ad hoc* external pressure.

An alternative model has emerged in which the flows are, in fact, smoothly distributed with high ionization parameters. The high U values, $\gtrsim 100$, are reconciled with the data (and with radiative acceleration) by invoking a large column density ($N_\mathrm{H} \gtrsim 10^{23}$ cm s^{-2} for BAL winds) of highly ionized gas at the base of the flow. This gas is not radiatively accelerated because it is too ionized. But its bound–free absorption (in soft x-rays and the extreme UV) greatly 'softens' the spectrum seen by the wind material downstream, thereby lowering the wind's ionization level and facilitating its acceleration. The most compelling aspect of this model is that it provides a physical basis for the observed correlation between UV and x-ray absorption: the winds revealed by the UV lines cannot exist without the shielding x-ray absorber. The leading alternative explanation, which simply equates the two absorbing regions, presents a serious problem because the high column densities implied by the x-ray absorption might be impossible to radiatively accelerate (equation (4)).

A serious challenge to all models is the kinematic complexity. It is surprising, for example, that no intrinsic absorbers have shown line-of-sight acceleration. The crossing time for a flow with $v_4 = 2$ and $\Delta R_{0.1} = 1$ is only ~ 5 years, yet repeated observations over this time frame reveal no velocity changes. Just as challenging are the multi-component line troughs and the frequent lack of absorption near zero velocity. These characteristics might be caused by the episodic ejection of discrete 'blobs,' or by well-collimated flows that cross our line of sight and thus reveal only part of their full velocity extent. The latter hypothesis seems more likely for at least the broader lines, because their supersonic velocity dispersions (> 1000 km s^{-1}) should quickly dissipate discrete blobs. Collimated accretion disk winds might cross our line of sight if they are driven at first vertically off the disk, before being bent into fully radial motion by the AGN's radiative force. The collimation and vertical ejection might both be facilitated by magnetic fields running perpendicular through the disk plane. However, multiple line troughs would require multiple collimated outflows. A major problem with this picture is that these intricate flow structures, which are tied to the accretion disk, must remain stable while the disk rotates.

A more general problem is determining how much, and what aspects, of the diversity among intrinsic absorbers results simply from orientation effects. Do the various outflows identified by BALs, mini-BALs and intrinsic NALs coexist generally in AGNs? Orientation is clearly not the only factor. For example, the complete lack of high-velocity lines in Seyfert galaxies suggests a luminosity dependence, probably related to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the requirements for radiative acceleration. Similarly, the relationships to AGN radio emissions suggest that there is some (unknown) physical connection between the disk-wind properties and the formation of radio jets.

Element abundances and host galaxy evolution

Our understanding of the element abundances near AGNs is based on general principles of stellar NUCLEOSYNTHESIS and galactic chemical evolution. All of the heavy elements from carbon on up are synthesized from primordial H and He in stars. The amounts of these elements are thus revealing of both the amount of local STAR FORMATION and the evolutionary status of the galactic environment. The elements near AGNs specifically probe these properties in the centers of big galaxies. For distant quasars, the local abundance evolution might involve some of the first stars forming in collapsed structures after the Big Bang.

Abundance measurements from absorption lines are, in principle, quite simple; one has only to apply appropriate correction factors for the ionization to convert the ionic column densities into relative abundances. For example, the abundance ratio of any two elements a and b can be written simply as

$$\frac{a}{b} = \frac{N(a_i)}{N(b_j)} \frac{F(b_j)}{F(a_i)}$$

where N and F are respectively the column densities and number fractions of element a in ion state i, etc, in the absorbing gas. The F values are generally adopted from photoionization calculations.

The results from intrinsic NALs, and independently from the broad emission lines, indicate that quasar environments have roughly solar or higher heavy-element abundances out to redshifts > 4. (Results based on the BALs are so far unreliable because of the problems with partial line-of-sight coverage mentioned earlier.) The local star formation must be both rapid and extensive to achieve these high abundances at early cosmic times. In particular, much of the gas must have already collapsed into stars by the time the quasars 'turned on' or became observable—just a few billion years after the Big Bang. These findings support prior expectations for the rapid, early-epoch evolution of massive galactic nuclei.

Bibliography

Studies of intrinsic AGN absorption lines began in the 1960s with measurements of BALs and associated NALs in quasars. The status of this evolving field circa 1997 is summarized by numerous articles and reviews in the conference proceedings:

Arav A, Shlosman I and Weymann R J (eds) *Mass Ejection from AGN* (San Francisco: ASP Conference Series)

More recent results on quasar BALs appear in

Hamann F 1998 Broad PV absorption in the QSO PG 1254+047: column densities, ionizations and metal abundances in broad absorption line winds *Astrophys. J.* **500** 798–809

The status of BAL polarization work is summarized by

Ogle P M, Cohen M H, Miller J S, Tran H D, Goodrich R W and Martel A R 1999 Polarization of broad absorption Line QSOs. I. A spectropolarimetric atlas *Astrophys. J. Suppl.* **125** 1–34

The latest results on intrinsic x-ray absorption are discussed by

Brandt W N, Laor A and Wills B J 1999 On the nature of soft x-ray weak quasi-stellar objects *Astrophys. J.* **528** 637–49

A comprehensive study of intrinsic UV absorption in Seyfert galaxies appears in

Crenshaw D M, Kraemer S B, Boggess A, Maran S P, Mushotzky R F and Wu C-C 1999 Intrinsic absorption lines in Seyfert 1 galaxies: I. Ultraviolet spectra from the Hubble space telescope *Astrophys. J.* **516** 750–68

The work on AGN element abundances is reviewed by

Hamann F H and Ferland G J 1999 Elemental abundances in QSOs: star formation and galactic nuclear evolution at high redshifts *Ann. Rev. Astron. Astrophys.* **37** 487–531

Frederick Hamann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasistellar Objects: Overview

Quasistellar objects, or quasars, were defined originally as star-like objects of large redshift. Quasars are believed to be powered by the accretion of matter onto massive black holes at the centers of galaxies, a process that emits more energy than thermonuclear reactions. Today, quasars are considered to be the most luminous members of the general class of objects called active galactic nuclei, or AGNs. Quasars are the most luminous objects in the universe.

This article begins with a brief history of the discovery of quasars. Next it describes their main properties and the concepts that have been developed to explain them. It continues with a description of their nature and theoretical models. Then additional properties and topics are considered: absorption lines, host galaxies, and luminosity functions and evolution.

History

In 1960 Mathews and Sandage identified the radio source 3C 48 (the 48th object in the 3rd Cambridge catalog of radio sources) with a star-like object of 16th magnitude. Its optical spectrum showed unusual emission lines that could not be identified, and the object was considered to be a radio star, more of which were identified in the course of their work. In 1962, Hazard, Mackey and Shimmins established an accurate radio position for another source, the BRIGHT QUASAR 3C 273, by observing the object as it was being occulted by the Moon. Schmidt used this position to identify 3C 273 with a 13th magnitude star-like object (figure 1), whose spectrum also displayed unusual emission lines. Schmidt subsequently realized that the emission lines could be identified with lines from hydrogen and oxygen redshifted by 15.8% of their rest wavelength. A redshift of 0.158 was unprecedented for such a bright object. If the redshift arose from the expansion of the universe, it meant that 3C 273 was about two billion light-years distant, according to the Hubble law, in which the redshift of objects increases in proportion to their distance. Furthermore, the combination of the observed brightness of 3C 273 with the inferred distance meant that it had an intrinsic luminosity 100 times that of an entire galaxy like the Milky Way.

Following Schmidt's work, it was immediately realized that the redshift of 3C 48 was 0.37. By 1965, 3C 9 was found to have a redshift of 2.01, well in excess of the 0.46 value for 3C 295, the most distant galaxy known at the time. It had taken more than 30 yr of work on galaxies with the largest telescopes in the world for astronomers to push the limit of the observable universe to the distance of 3C 295. Only 2 yr after the discovery of quasars, the maximum observed redshift had increased by more than a factor of 4. The light from 3C 9 had taken 80% of the age of the universe, or more than 10 billion years, to reach Earth. The ability to observe objects at redshifts greater than 2 meant that astronomers could probe back in time to within a few billion years of the big bang.

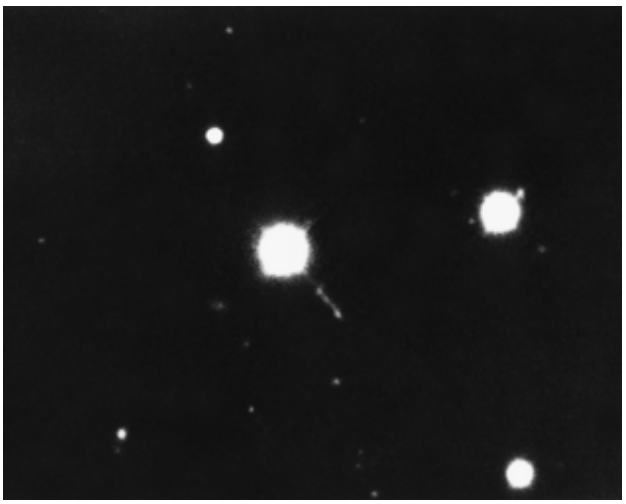


Figure 1. Optical image of 3C273. The object looks stellar in this image except that it is accompanied by a jet of radiation extending to the lower right. (Credit National Optical Astronomy Observatories/National Science Foundation. Copyright Association of Universities for Research in Astronomy Inc (AURA), all rights reserved.)

The unusual nature of quasars was compounded by the work of Smith and Hoffleit, who showed that 3C 273 varied significantly in brightness on timescales of months. This variability indicated that the characteristic size of the emission region was light-months and led immediately to the question, how could such a small object produce more light than an entire galaxy, with a typical size of 100 000 light-years?

To summarize, quasars revolutionized astronomy in the 1960s for two reasons: (1) they pushed back the limits of the observable universe significantly in both distance and lookback time and (2) their compact size and great luminosities could not be explained in terms of the stars and galaxies known at the time. Quasars opened new frontiers in cosmology and stimulated the development of a new subject—relativistic astrophysics.

Main properties and concepts

The main observed properties of quasars are the following.

1. Nuclei that appear starlike in optical images. Extended emission can now often be detected around the nucleus, and jets extending away from the nucleus are occasionally seen.
2. Spectra showing broad emission lines with widths greater than observed in normal galaxies and a range

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of ionization broader than seen in typical nebulae ionized by stars.
3. Structures in quasars with observable radio emission that range in angular extent from m arcsec to tens of arcsec. The most compact structures are often observed to expand on timescales of years.
4. Radiation of similar power at energies ranging from γ-rays to the far-infrared (100 μm) region of the spectrum, with a decrease in energy at radio frequencies.
5. Redshifts, z, of spectral features ranging from 0.1 to 5.
6. Variability in brightness at different wavelengths on timescales as short as days to weeks.
7. Luminosities as high as 10^{14} Suns.
8. Total energy in high-energy electrons in the region of radio emission that can exceed 10^{60} erg.

Additional information about these properties and the concepts they imply about the nature of quasars follows below.

Redshifts

The large redshifts of quasars are one of their distinctive features, and the value of the greatest known redshift, z, continued to increase after 1965, surpassing $z = 3$ in 1973 and $z = 4$ in 1987. The maximum value reached $z = 4.9$ in 1991, and an object at $z = 5$ was discovered in the first stages of the Sloan Digital Sky Survey in 1998 (see QUASISTELLAR OBJECTS: SURVEYS). The great majority of quasars, however, have redshifts less than 2.5. This occurs for two reasons: (1) such quasars are easier to find because they are brighter in the ultraviolet than most stars and (2) they are intrinsically more numerous than quasars with $z > 2.5$. The low-redshift limit of quasars is normally set at 0.1 because objects at lower redshift are usually resolvable as galaxies and are cataloged as ACTIVE GALACTIC NUCLEI.

In addition to providing our first view of the universe at $z > 0.5$, quasars also proved to be valuable cosmological probes because they were luminous enough to yield clues on the properties of matter along the line of sight back to the Earth (see discussion on absorption lines below and in more detail in QUASISTELLAR OBJECTS: INTERVENING ABSORPTION LINES).

Variability

Observations indicate that variability is a common feature of quasars and most quasars are believed to show variations in light at the 10–40% level over timescales of months to years. Furthermore, variations are observed in the strength and shape of emission lines and in continuum emission at x-ray, ultraviolet and radio wavelengths. The variations are irregular in nature, with periods of flaring and relative quiescence being intermixed. Some objects, which are called BLAZARS, show variations in light on timescales as short as 1 day (see BL LACERTAE, BL LACERTAE OBJECTS and AGN: VARIABILITY).

Variability provides powerful tools for determining the inner structure and nature of quasars. When the central source itself varies in brightness, the effects of the variation propagate outward with time and produce changes in, for example, the emission line spectra and continuum radiation from the material surrounding the central source. In the case of blazars, we are seeing almost directly down the axis of the jets of material that are being ejected from the central regions. By following how the radiation from the jets varies at different wavelengths with time, we are able to infer the physical conditions and processes that are occurring in the jets.

Observations of the variability of quasars carried out at different wavelengths in intensive campaigns and also in programs extending for years have provided some of the most direct information on the nature of the inner regions in quasars and AGNs. These programs have shown that the emission line regions are considerably smaller and closer to the central source than was originally thought.

Continuum emission and multi-wavelength observations

Quasars emit radiation strongly in bands ranging from γ-rays and x-rays to the far infrared (100 μm). The amount of energy emitted in each band by most objects is remarkably similar, in contrast to normal thermal radiation from stars, which is much more peaked and restricted in wavelength. For example, most stars are cool and are faint at ultraviolet wavelengths. In contrast, most quasars with redshifts less than 2.5 are bright at ultraviolet wavelengths, a property that is very helpful in distinguishing them from the more numerous stars that are found in sky surveys.

Although quasars were discovered through their radio emission, about 90% of quasars do not emit strongly at radio wavelengths and are classified as radio quiet. Radio-loud quasars are typically 100 times brighter at radio wavelengths than radio-quiet quasars.

The continuum emission in quasars appears to arise from a combination of thermal and non-thermal processes. For the latter, synchrotron emission from relativistic electrons produces the lower-energy radiation. The x-ray and γ-ray emission may come from the inverse Compton scattering of lower-energy photons by high-energy electrons, although thermal mechanisms may also play a part in x-ray emission. In any event, the continuum radiation from quasars demonstrates that some very energetic processes are involved. Furthermore, the continuum radiation at the highest energies tends to show the most dramatic variability and the shortest timescales, which is another indication of the extreme conditions that exist near quasars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Emission lines

The emission line spectra of quasars (figure 2) are characterized by the large widths of the lines and wide range of ionization. The full widths at half-maximum of the emission lines are typically 5000 km s^{-1} and range from about 500 km s^{-1} to greater than 10 000 km s^{-1}. Full widths at the base of the lines can exceed a tenth the speed of light.

The widths are produced by motions of gas in the emitting region, where a mixture of infall, rotation and ejection probably occurs. The widths are consistent with the emission region being at a distance of light-months to a few light-years from a central black hole.

One of the most striking observational properties of the spectra of quasars and AGNs is the anticorrelation of the strength of the C IV emission line at 1550 Å and the continuum luminosity, that is the emission line is observed to be weaker relative to the continuum in more luminous objects. This relation, which is called the Baldwin effect, is perhaps the most prominent observed correlation that has been found for quasar and AGN spectra. It is important for two main reasons: (1) the clues it provides about the nature of the emission line region and central source and (2) the potential it offers for using quasars as cosmological probes at high redshifts.

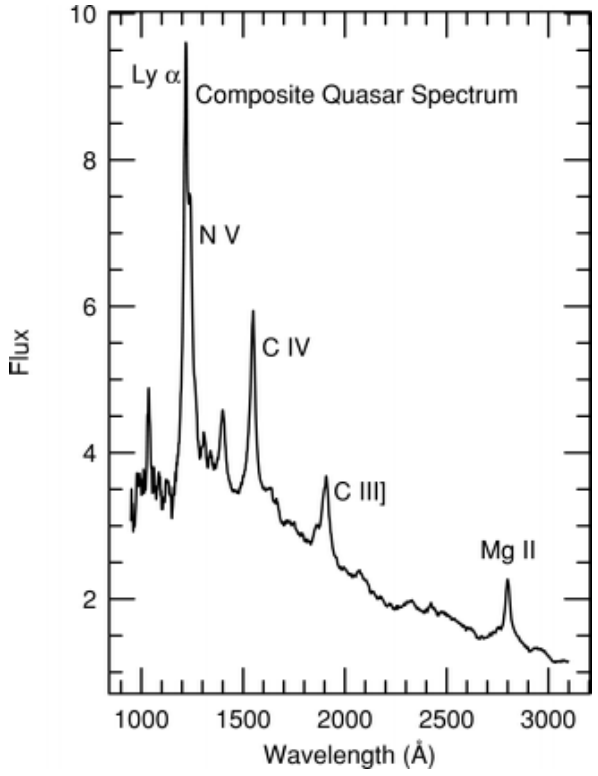


Figure 2. A composite quasar spectrum covering the ultraviolet region of the spectrum. It shows strong emission lines from hydrogen, carbon, nitrogen and magnesium. (Credit Francis P *et al* 1991 *Astrophys. J.* **373** 465.)

The strongest emission lines in quasar spectra come from hydrogen, carbon and magnesium, with lines of nitrogen, oxygen, iron and other elements also being visible. The observed levels of ionization range from neutral for hydrogen and oxygen up to five-times ionized oxygen and even more highly ionized iron. The range of ionization is another indicator of the strength of the non-thermal continuum radiation in quasars and its extent to high energies.

Although it is difficult to make a reliable determination of the chemical abundances in quasar emission line regions, results to date indicate that the abundances are similar to those observed in present-day stars and nebulae, even in quasars at the highest redshifts. This is a striking result when one considers that the most distant quasars correspond to cosmic times of 1 or 2 billion years after the big bang. Except for hydrogen, all the elements mentioned above originate in the thermonuclear reactions that occur in stars. The abundances of these elements increase with cosmic time as successive generations of stars complete their evolution. When stars die, they return part of their newly produced elements to interstellar space, and the next generation of stars forms from the enriched material.

Thus, the apparently normal abundances of the elements in high-redshift quasars implies that they were preceded by substantial stellar activity in the early universe.

Nature

The prodigious luminosities of quasars in combination with their small size led theorists to consider gravity as their energy source immediately on their discovery. It was realized that normal stars and thermonuclear reactions were entirely insufficient to produce the observed properties of quasars. However, the required gravitational potential energy could only be extracted if quasars contained compact objects with masses of a hundred million Suns and sizes about that of the solar system. In addition the gravitational field of such objects was so strong that the effects of general relativity would be dominant in their vicinity. Since that time, gravity has remained as the consensus source of energy for quasars, although the concepts of the nature of the central source have changed considerably. For example, supermassive 'stars', were originally postulated as the central objects in quasars, but enough difficulties were found with such ideas that they were soon abandoned.

Subsequently, it was realized that BLACK HOLES were the most likely central engines of quasars. A black hole was an object with gravity so strong that not even light could escape from its surface. The widths of the emission lines in quasars and estimates of the size of the emission line region suggested that the central mass was about 10^8 Suns. Such objects would not be able to support themselves against the pull of their of their own gravity,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and they would collapse to form a black hole. The collapse could liberate 10% or more of the rest energy, mc^2, of the object, that is, more than 10 times the energy of the thermonuclear reactions that occur in stars. Similarly, any material later captured by the black hole would liberate the same fraction of rest energy. This of course provided an explanation for the energies found in quasars. As a result, quasars stimulated much work on the origin and nature of black holes and the detailed processes of how the observed energy could be obtained from them.

Today, the working hypothesis for quasars and AGNs is that massive (10^6–10^9 solar masses) black holes exist at their centers. The radiated energy we detect from them comes from matter being accreted onto the black hole. According to general relativity, the radius of a black hole is 3 km per solar mass, and the radii of black holes in AGNs and quasars range from a few solar radii to 20 times the distance of the Earth from the Sun.

It is most likely that much of the matter surrounding the black hole is in the shape of a disk. The matter in the disk orbits the black hole and moves inward as it loses angular momentum from some source of viscosity in the disk. The inner edge of the disk may extend to within a few radii of the central black hole. The gas in the inner regions of the disk is expected to be hot enough, about $10^{4–6}$ K, to account for the thermal component of the continuum radiation at ultraviolet wavelengths.

The x-ray observations of quasars provide evidence of the existence of a corona of energetic particles above the accretion disk. The x-rays may be produced by upscattering of the ultraviolet photons from the disk on the energetic particles via the inverse Compton process. Some x-rays may also be produced in shocks located farther from the central source.

Jets of beamed radiation are also an important feature. They are frequently observed in radio-loud quasars and, as in the case of 3C 273, sometimes are seen at optical wavelengths. The jets are thought to be accelerated electromagnetically and ejected along the rotational axis of the black hole. They can be moving at relativistic speeds and contain high-energy particles, such as the relativistic electrons that produce the radio emission through the synchrotron process. The radio emission in the 3C 273 jet is located several light-years distant from the central source.

A striking feature of jets in some radio sources is that repeated observations on the smallest angular scales with radio interferometers show them to be moving outwards at speeds that appear to be several times greater than the speed of light. This of course would violate a basic limit of relativity. The most satisfactory explanation for this paradox is that the jets are moving at speeds near that of light in the direction of the Earth. In this case the apparent time interval in the expansion motion of the jet appears smaller than it really is because of the motion toward the Earth. See the article on SUPERLUMINAL MOTION for more information.

The emission line spectra of quasars indicate that there are also gas clouds surrounding the black hole. The clouds occupy about 10% of the volume of the region where the broad emission lines are produced. The broad emission line region in quasars can extend to several light-years from the black hole.

Understanding the nature of the inner regions of quasars and the central black holes continues to be an important subject of research both observationally and theoretically. Specific topics include the formation of black holes and their relation to the formation of galaxies, the role that galaxy interactions play in providing the fuel supply of gas for accretion onto the black hole and how quasars evolve in time and what is their relation to the inactive black holes now being observed in the centers of nearby galaxies.

Alternative theories

One of the first alternative explanations for the large redshifts of quasars to be investigated was gravitational. If the emission arose near a compact massive object with a sufficiently strong gravitational field, then the emission lines could exhibit large redshifts. However, it was realized almost immediately that such a hypothesis was inconsistent with the widths and other properties of the emission lines, and the hypothesis was ruled out.

Another hypothesis was that quasars were being gravitationally lensed and thus appeared much brighter than they really were. Interestingly, many years after this idea was proposed, some quasars were discovered to be lensed, but they are only a small fraction of the total.

From observations of the apparent associations of quasars with galaxies of much lower redshift, Arp has challenged the concept that quasar redshift arise from the cosmological expansion of the universe. In this picture, the redshifts arise from some other, non-cosmological effect whose nature has yet to be determined.

Subsequent observations, however, have shown that quasars do reside in galaxies. In the cases where it has been possible to measure the redshifts of the host galaxies, they agree with the redshifts of the quasars. Furthermore, low-luminosity quasars and luminous Seyfert galaxies overlap in luminosity and show a general continuity of properties.

Today, it is generally accepted that the large redshifts of quasars do arise from the expansion of the universe and that they are at the distances indicated by the redshifts.

A different approach has been taken by Terlevich and collaborators, who have proposed that starbursts, intense episodes of star formation, can account for quasars. In this case energy from hot, young stars and the supernova explosions that occur at the end of their evolution provides the radiation and spectral properties

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observed in quasars and AGNs. While there are abundant observations showing that starbursts are common in regions around AGNs and may well be related to the processes that fuel the central object, the starburst model has great difficulty in accounting for the most luminous quasars and for the rapid variability in x-rays that is observed in AGNs. Furthermore, the observational evidence for black holes in nearby galaxies has become increasingly strong.

Additional topics

Absorption lines

Absorption lines were detected in quasars within a few years of their discovery and were initially believed to be rare. However, with improved sensitivity and spectral resolution, it became clear that absorption lines were present in all quasars with redshifts 2 and greater. A contributing factor to their presence was that ultraviolet, ground-state lines were being redshifted to wavelengths where they could be observed with ground-based telescopes. Today, the study of quasar absorption lines is a major topic and the article on this subject should be consulted for more details.

Absorption lines may be placed in three categories according to the distance of the absorbing gas from the central emission source of the quasar: (1) intrinsic systems, (2) associated systems and (3) intervening systems.

Intrinsic systems arise close to the quasar itself. The most conspicuous members of this class are the broad absorption line (BAL) quasars, whose spectra show broad troughs of absorption on the short-wavelength side of the Lyman α line of hydrogen and such high-ionization lines as C IV λ1549 and N V λ1240, for example. The troughs indicate outflow velocities that can exceed 30 000 km s^{-1}. BAL features are seen in about 10% of high-redshift quasars, a value that gives an indication of the average covering factor of the outflowing gas.

Associated systems show absorption features from elements such as hydrogen, carbon and magnesium. They have redshifts close to that of the emission lines in quasars but are much narrower. They are believed to arise in gas associated with the host galaxy of the quasar or the environment in which the host galaxy resides, which might be a group or cluster of galaxies.

Intervening systems have redshifts smaller than the emission lines in quasars and arise in clouds of gas unrelated to the quasar that lie along the line of sight to the Earth. In this case, the quasar serves as a background beacon that enables the study of otherwise invisible gas. The sight lines to quasars probe representative samples of intergalactic space, and absorption line spectra provide an unbiased way to determine the nature and distribution of the gas. Studies of absorption lines in quasars have provided crucial information about gas in the high-redshift universe. Intervening systems may themselves be classified in three groups according to the column density of the absorbing gas: (a) damped Lyman α systems, (b) intermediate (metal line and/or Lyman limit) systems and (c) the Lyman α forest.

Damped Lyman α systems represent the highest column densities, typically in excess of 10^{20} neutral hydrogen atoms cm^{-2}, and are so called because absorption is so strong that the damping wings of the line are seen. Such systems are thought to occur when the line of sight passes through a disk galaxy. Some may also arise in dense clouds or filaments of intergalactic gas at high redshift.

Damped Lyman α systems, because of their high column densities, also exhibit absorption lines from heavier elements such as carbon, silicon, magnesium and iron, with additional elements being detected in the most favorable cases. Analyses of these lines indicate that the abundance of metals in the systems is about one-tenth of the solar value. This is consistent with our understanding of the chemical evolution of galaxies, in which the abundance of metals increases with cosmic time owing to the continuing enrichment of the interstellar medium by supernovae and red giant stars as they complete their life cycle. Damped Lyman α systems can only be observed from the ground for redshifts greater than 1.8, which corresponds to a lookback time of approximately 80% of the age the universe.

Intermediate systems, with a characteristic column density of 10^{17} neutral hydrogen atoms cm^{-2}, also show absorption lines from elements such as carbon, magnesium and other metals plus a discontinuity at the Lyman limit of hydrogen. They are commonly attributed to the halos of galaxies. Their metal abundances range down to 1% of the solar value, indicating that they have had less chemical enrichment than the gas in damped Lyman α systems.

The Lyman α forest lines are the most numerous and ubiquitous absorption features and are attributed to clouds of gas in the intergalactic medium. They are seen in a range of column densities beginning at the limit of detectability around $10^{12–13}$ neutral hydrogen atoms cm^{-2} and extending up into the range of the intermediate systems. Their number density increases with redshift; by redshift 4, the integrated effect of absorption from the Lyman α forest produces a significant depression of the flux shortward of the Lyman α emission in quasar spectra. Observations of the Lyman α forest are used to deduce the state of ionization of the intergalactic medium. Much of the ionizing flux at high redshift is produced by quasars. The data are also important for understanding the formation of structure in the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

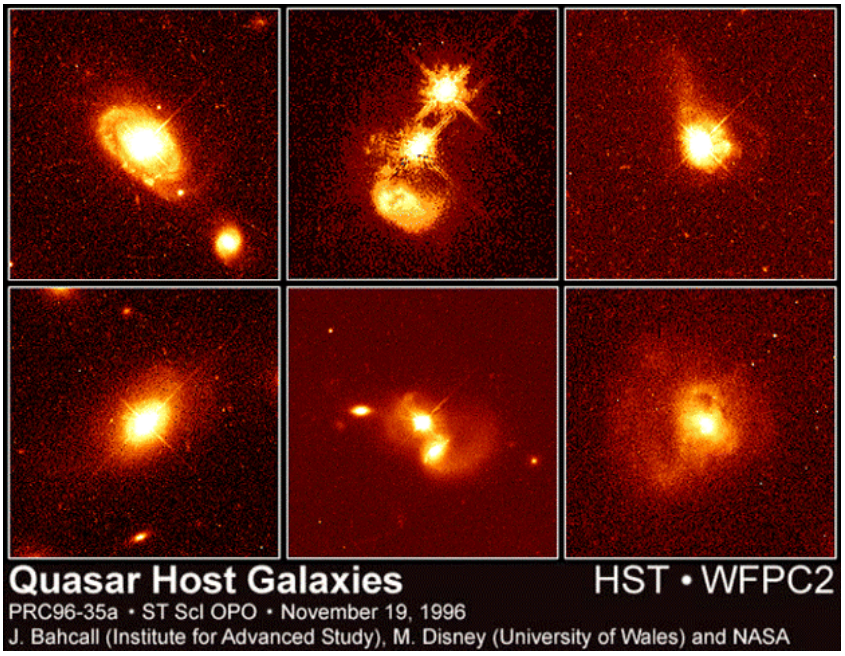


Figure 3. A sample of images showing the host galaxies of quasars. The hosts can be normal-looking spiral and elliptical galaxies (left panels) but are frequently seen to be interacting or merging galaxies (center and right panels). (Credit J Bahcall, M Disney, NASA. This image was created with support of Space Telescope Science Institute, operated by the Association of Universities for Research in Astronomy, Inc from NASA contract NAS5-26555, and is reproduced with permission from AURA/STScI.)

Host galaxies

The regions immediately around quasars are difficult to observe because of the glare from the central source, which is why quasars were originally called star-like objects. However, with improved observations, extended emission began to be detected around quasars. In the meantime the realization was developing that quasars, Seyfert galaxies and other active galactic nuclei were members of the same family. One of the main differences among the different classes of objects was the luminosity of the central source, and the host galaxies with low-luminosity nuclei were much easier to study.

The excellent image quality and spatial resolution of the Hubble Space Telescope have enabled a significant advance of our knowledge of quasar host galaxies (figure 3). Results to date for low-redshift objects show that quasars occur in elliptical, spiral and interacting galaxies, with some of the latter appearing very disturbed by the interaction. On average, the host galaxies are brighter than normal field galaxies. The radio-loud quasars tend to occur in elliptical or interacting galaxies. The radio-quiet quasars can occur in elliptical or spiral galaxies. These results are in contrast to the situation for lower-luminosity AGNs, where there are stronger tendencies for radio-quiet objects to occur in disk galaxies and radio-loud objects in ellipticals.

The study of host galaxies is important to improving our understanding of how both galaxies and quasars form and evolve. The quasar is in the nucleus of the galaxy and represents the endpoint of the process that produced the central concentration of matter in the galaxy. How a massive black hole formed and grew from the central concentration remains an important research question. At the same time, for the black hole to be visible and attain the luminosity observed for bright quasars, it must accrete several solar masses of material per year. What causes this inflow of matter from the outer parts of the galaxy to the nucleus is also an important topic. It is thought that the interaction of two galaxies, that is their collision or near collision, can produce enough disruption to the orbit and angular momentum of stars and gas to produce the inflow, but the details of the process are not yet fully understood.

Luminosity functions and evolution

The luminosity function of quasars is the distribution of their volume density as a function of luminosity and redshift. As with galaxies, the number density decreases steeply at the highest luminosities. At low luminosities the slope of the function is smaller than at high luminosities, so the rate of increase of objects slows with decreasing luminosity.

Luminosity functions are among the most basic parameters in astronomy and provide a basis for determining the amounts of matter in different forms in the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

One of the most striking properties of quasars is the evolution of their luminosity function with redshift. This was first noticed by Schmidt, who realized there were too many quasars at high redshift in early surveys if the space density of objects were constant. Interpreted as density evolution, the space density of luminous quasars is more than a thousand times greater at $z = 2$ than in the local universe. An alternative explanation is that quasars were brighter at high redshift than at present.

At redshifts greater than 3, there is considerable observational evidence that the space density of quasars declines steeply. A logical interpretation of the results (figure 4) is that we are seeing back to the epoch of peak quasar activity. Alternatively, clouds of dust at high redshift could be blocking our view of more distant quasars. For example, such clouds could be associated with star formation regions in the host galaxy of the quasar. However, the best information currently available indicates that dust is an important factor for some types of quasars but does not account by itself for the decline in space density at high redshift. In any event, the most distant objects now known in the universe are galaxies, the first time this has happened since 1965. Although this may be a coincidental result, it is another indication that quasars at redshifts greater than 5 are genuinely very rare objects.

Current observations of quasars and distant galaxies are consistent with bursts of star formation occurring in galaxies before the quasar activity builds up to its peak around redshift 2–3. As was mentioned earlier, the gas in the emission line regions of distant quasars appears to have relatively normal abundances of elements heavier than hydrogen, which implies that the gas has been enriched by massive stars before reaching the quasar. Furthermore, it takes time for a supermassive black hole to attain the masses needed for quasars, time that allows for a few generations of massive stars to complete their evolution. For reference, the age of the universe is only about a billion years at redshift 5.

Subsequently, the overall rate of star formation in the universe reaches a peak as galaxies continue to form and assemble. After that, both quasar activity and star formation rates decline until they reach the values observed today in the local universe. Both declines are probably related to galaxies completing their formation and assembly. In addition, the rate of interactions and mergers of galaxies decreases as the universe continues to expand. As a result less material is available (1) to fall to the central regions of galaxies and fuel the massive black holes and (2) to form stars.

In this picture, the quiescent black holes that today are found to be ubiquitous in nearby galaxies are the products of the spectacular galaxy formation process and quasar activity that occurred in the first third of the history of the universe.

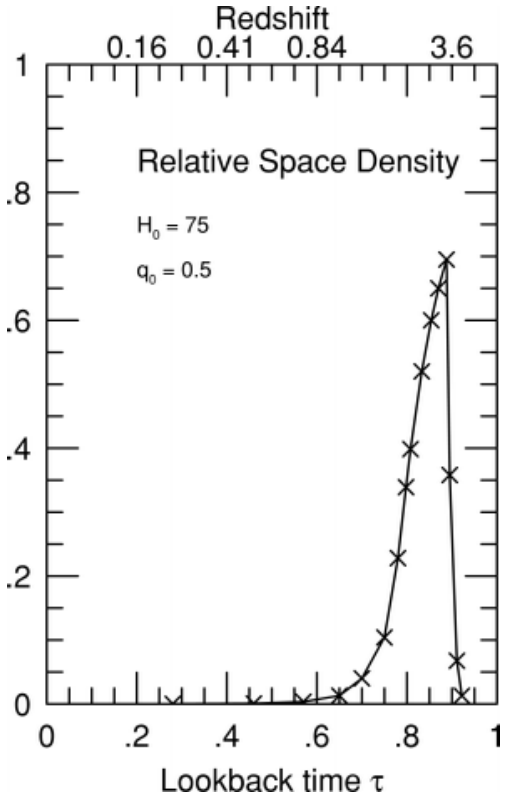


Figure 4. The evolution of the space density of luminous quasars as a function of cosmic lookback time, expressed as a fraction of the age of the universe. The space density shows a strong peak at 0.85, when the universe was 15% of its present age. (Credit Warren S, Hewett P and Osmer P 1994 *Astrophys. J.* **421** 412.)

However, the picture is but a sketch and is in need of both development and confirmation. We may expect significant progress in the near future on several fronts.

- Large new surveys for quasars that are being carried out at optical and radio wavelengths, such as the Sloan Digital Sky Survey, the Australian 2dF Survey and the FIRST radio survey, will increase the number of known quasars by a factor of ten. (In June 2001, two quasars with redshifts of 6.0 and 6.2, the most distant objects ever observed in the universe were identified in the Sloan Digital Sky Survey.)
- The new generation of large telescopes with 8–10 m apertures that has come into operation and space observatories such as Chandra and XMM that will soon be launched provide powerful tools for observing quasars with significantly more sensitivity and resolution at optical, infrared and x-ray wavelengths than has been available previously.
- Advances in theoretical modeling and the continuing increase in computing power will enable significant progress in interpreting the new observations and in modeling the inner regions of quasars and key processes such as accretion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thus, the prospects for improving our understanding of how quasars form and evolve and for determining their relation to galaxies are very bright.

Bibliography

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

Robinson I, Schild A and Schucking E L (ed) 1965 Quasi-stellar Sources and Gravitational Collapse: *Proc. 1st Texas Symp. on Relativistic Astrophysics* (Chicago, IL: University of Chicago Press)

Patrick Osmer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quasistellar Objects: Surveys

Since the discovery of QUASARS in 1963, surveys have been a prerequisite both for investigating models of the central regions of quasars themselves and to provide samples of quasars for use as cosmological probes. The early history of surveys for quasars was determined by our incomplete knowledge of their nature and of their relationship to galaxies. Discovered initially through the campaign to identify the strong RADIO SOURCES in the Third Cambridge Radio Catalogue (3CR), the optical counterparts were found to be stellar, i.e. unresolved point sources, and to possess blue optical colors. Following the initial spectroscopic identification of 3CR273 as an object with the then unprecedented REDSHIFT of $z = 0.16$ (see BRIGHT QUASAR 3C 273), by Maarten Schmidt, a significant number of quasars were soon found by obtaining optical spectra of unresolved blue objects within the positional error boxes of radio sources. A further major advance was made in 1965 when Allan Sandage showed that there existed a much larger population of unresolved blue sources with very similar properties to 'quasars' except that the objects were not detectable as radio sources. Some researchers continue to draw a distinction between quasi-stellar radio sources (quasars) and quasi-stellar objects (QSOs). However, many workers regard the distinction as primarily historical and both 'quasar' and 'QSO' may be used to refer to objects irrespective of whether they possess significant radio emission.

An enormous observational effort at radio, infrared, optical and x-ray wavelengths in the three decades following the discovery of quasars has resulted in the identification of some 12 000 quasars. Most recently, the ability to undertake sensitive surveys over very large areas of sky has marked the beginning of projects that will produce catalogs of hundreds of thousands of quasars, allowing entirely new scientific investigations to be undertaken.

Identifying quasars

Quasars are rare objects that exhibit a very broad range of intrinsic properties. Normal stars and the majority of galaxies, whose spectral energy distributions are dominated by the radiation from their constituent stars, possess blackbody-like spectral energy distributions that are strongly peaked. The characteristic temperature of the stellar photospheres determines the wavelength at which the bulk of the radiation is emitted. By contrast, the extreme physical conditions in the central regions of quasars lead to the emission of radiation over many decades of wavelength and quasars emit significantly in the x-ray, ultraviolet, optical, infrared and radio portions of the spectrum. With only a few exceptions, surveys for quasars rely on identifying objects whose spectral energy distributions differ from the blackbody-like spectra of normal stars and galaxies.

The majority of known quasars have been identified using data from the optical region of the electromagnetic spectrum. At redshifts $z \leq 2.2$ quasars exhibit an excess of near-ultraviolet light compared with common Galactic stars. The 'ultraviolet-excess' technique, involving the selection of objects with unusually blue broad-band colors derived from B and U band magnitudes, has produced several thousand quasars—figure 1. The technique can be made both more efficient, by reducing the number of contaminating hot stars included, and more complete, by increasing the sensitivity to a larger fraction of the quasar population, by extending the wavelength interval through the use of additional broad-band magnitudes (e.g. one or more of V, R, I). The so-called 'multicolor' technique has proved very effective, allowing the identification of quasars over the redshift range $0 < z < 5$.

The second primary methodology employed in the optical regime relies on the identification of the prominent emission lines, present in most quasars, visible in low-resolution spectra. Slitless spectroscopy, a technique that employs a thin prism (objective prism) or a combination of prism and transmission grating (grism) enables spectra of thousands of objects over large fields to be obtained. Prior to the development of the multi-color selection techniques virtually all quasars with redshifts $z > 2.2$ were located through the identification

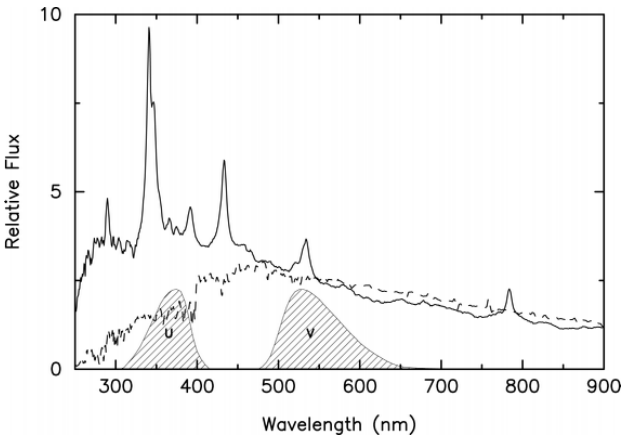


Figure 1. Spectra of a solar-like star (dashed curve) and a typical quasar of redshift $z = 1.8$ (solid curve). The measured fluxes of both objects in an optical passband, such as the V band represented by the hatched region to the right, are very similar. However, the flux of the quasar measured using a passband in the near-ultraviolet, such as the U band represented by the hatched region to the left, is relatively much greater than for the star. Employing this characteristic excess of ultraviolet flux is one of the simplest and most effective techniques for identifying quasars with redshifts $z \leq 2.2$. Another class of discovery techniques focus on the detection of the strong emission lines which are visible in the quasar spectrum but absent in the spectra of normal stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

of the strong hydrogen Lyman-α 121.6 nm and carbon C IV 154.9 nm emission lines which, for redshifts $1.5 < z < 3.5$, appear in the optical portion of the spectrum.

While the basic methodologies of the color, slitless-spectroscopic and other discovery techniques have been established for several decades, there has been a longstanding debate over the relative effectiveness of quasar survey techniques. Advocates of particular methodologies have often claimed to detect a greater fraction of the quasar population compared with other approaches. Originally, the selection of candidate quasars was performed by individual researchers visually inspecting photographic plates through a binocular microscope. In recent years a significant improvement in the agreement between the results of different surveys has come about through the use of digital data and the implementation of automated, computer-based selection procedures. The computer-based selection ensures homogeneity in the candidate selection and, most importantly, allows a precise quantification of the effectiveness of a particular survey (see e.g. SOFTWARE: DIGITAL SKY SURVEYS). A now well-established technique is to simulate the appearance of quasars of given properties—redshift, magnitude and spectral energy distribution—in the data set under examination. The survey selection procedure can then be applied to the simulated objects to establish what fraction of quasars with particular properties can be successfully identified—figure 2. An example of the success of such methodology was the remarkable agreement in the derived space density of HIGH-REDSHIFT QUASARS, $z = 3$–4.5, from two groups, one employing multicolor selection and the other using slitless-spectroscopic techniques.

The extreme luminosities and very extended spectral energy distributions of quasars mean that many of the apparently brightest sources of radiation at x-ray, infrared and radio wavelengths are quasars. Employing the very-large-wavelength baseline between, for example, the optical and x-ray portions of the electromagnetic spectrum should provide a highly efficient technique for identifying quasars; given a sample of quasars and stars of similar optical brightness, the quasars will appear unusually 'x-ray bright'. Relatively small numbers of quasars have been identified through investigation of radio, infrared and x-ray catalogs and these samples have proved particularly important for studies of the physics of the central energy source of quasars. However, the sensitivity of telescopes and their associated detectors at wavelengths away from the optical is, in a relative sense, poor. Thus, only a small fraction of the quasars detectable via optical techniques have been accessible to observations made at the extremes of the electromagnetic spectrum.

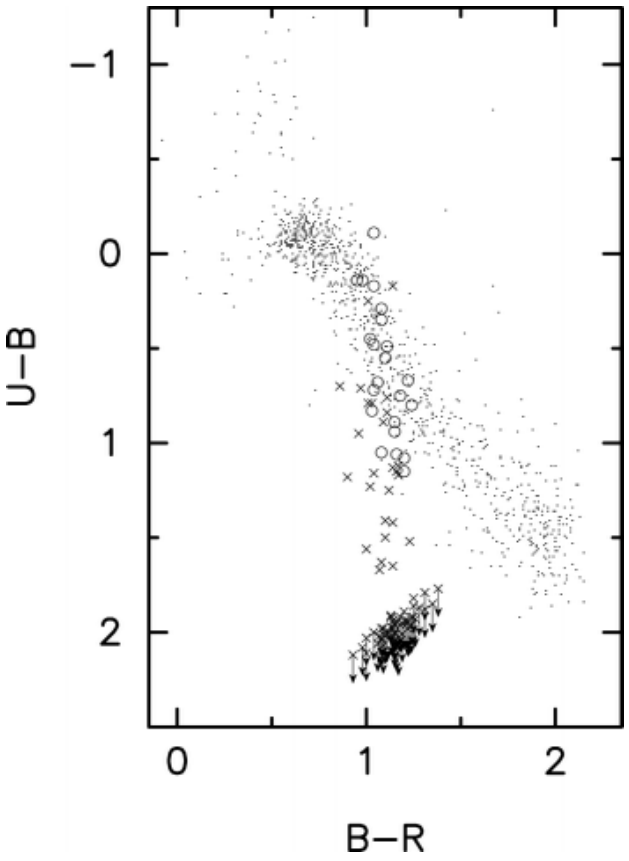


Figure 2. Many modern surveys aim to detect quasars with a variety of spectral energy distributions over a large range in redshift. Simulations of the appearance of a quasar in a particular survey, incorporating a detailed model of the effects of intervening absorption and the properties of the survey itself, allow the probability that a quasar of given brightness, redshift and spectrum will be detected to be calculated. The figure shows a two-color diagram with the colors of 980 of the total of ~30 000 objects in the survey area (•). The larger symbols show the predicted colors of a quasar with redshift $z = 3.33$ derived from such a simulation. ○s indicate occasions when the quasar would not have been identified, while ×s represent occasions when the quasar was successfully picked out. The simulation shows that the probability of detecting a quasar of the specified type is $P = 0.78$. Symbols with downward pointing arrows correspond to occasions when the simulated U magnitude was below the survey detection limit and the quasar lies somewhere below the plotted symbol, with a larger $U - B$ color.

Notwithstanding a general convergence in the results from, and understanding of the differences between, various survey techniques, some researchers continue to believe that a significant fraction of the quasar population remains to be identified. Two alternative techniques, based on lack of proper motion and the presence of photometric variability, have been proposed to address this question. Over timescales of decades essentially every star in our own Galaxy exhibits a small, but detectable, motion on the sky and surveys for quasars employing so-called PROPER MOTIONS focus on objects

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that do not appear to move. Photometric variability, with a wide range of amplitudes, was recognized soon after the discovery of quasars as a distinctive feature of many objects. The cause in the majority of quasars is believed to be variations in the rate at which material is accreted by the central massive object. The statistical properties of photometric variability are not well determined but, if monitored at approximately monthly intervals over a period of one or two decades, virtually all quasars exhibit brightness changes of at least 10%. Variability surveys thus treat any object showing evidence of variability, not explicable as a variable or binary star, as a quasar candidate. Recent results from both techniques agree well with those derived from other surveys.

Present status

Impressive progress has been made over the last 15 years in compiling samples of quasars from x-ray to radio wavelengths. As a result, the general behavior of the quasar luminosity function over the redshift range $0 < z < 5$ is fairly well established. Significant constraints on models of the central regions of quasars have been derived from the spectroscopic and other properties of quasars in existing surveys. However, the popular unified scheme models for quasars and ACTIVE GALACTIC NUCLEI predict that the population should exhibit a very large spread in luminosity and spectral energy distribution. Similarly, quasars viewed along a line of sight that passes through an intervening galaxy containing dust may be dimmed by very large factors. Virtually all current surveys, independent of wavelength, are sensitive to objects with only a narrow range in luminosity at fixed redshift (see also GALAXY REDSHIFT SURVEYS). As a consequence, existing surveys do not in general provide strong tests of a number of different theoretical models. While the total number of quasars known, ~12 000, may seem large, for many investigations, including those that exploit quasars as cosmological probes, very large numbers of objects are required. Progress in areas where the signal sought is weak, such as defining the clustering behavior of quasars as a function of redshift, are still limited by the lack of surveys containing many thousands or tens of thousands of quasars.

Future surveys

Undertaking surveys for many thousands of quasars has in practice been impossible because of the lack of an instrument capable of obtaining spectroscopic observations of hundreds of faint candidate objects over areas of at least a square degree on the sky. Recent developments in instrumentation have overcome this practical restriction and two major surveys now in progress will produce samples of tens of thousands of quasars within the next five years.

In Australia, at the 3.9 m Anglo-Australian Telescope, a multiobject fiber spectrograph with 400 fibers and a 2º diameter field of view, the so-called '2 degree field' (2dF), has been commissioned (see also SPECTROGRAPHS: FIBER-FED SPECTROGRAPHS). Simultaneous spectroscopy of quasar candidates to $B = 21$ over an area of three square degrees is relatively straightforward and the '2dF Quasar Survey', with a target of 25 000 quasars, is a key program. The survey area consists of two long strips on the sky, 75º × 5º in extent, with the selection of the targets based on a three-color (*UBR*) technique that is sensitive to quasars over the redshift range $0.2 < z < 3.0$. The magnitude limit of the survey and the geometry of the survey regions have been chosen to optimize the quasar sample for investigation of spatial clustering properties as a function of redshift. The quasars act as probes of large-scale structure and the extended redshift range over which quasars are detected will provide an order of magnitude improvement in the determination of the development of large-scale clustering with redshift, providing an important new constraint on theories of structure formation.

A still more ambitious program involves the SLOAN DIGITAL SKY SURVEY (SDSS) which has recently begun to acquire imaging in five broad optical passbands (*u*,*g*,*r*,*i*,*z*) with the aim of covering 10 000 square degrees of the sky at high Galactic latitude. The second element of the project is to obtain spectra of the brighter galaxies and quasar candidates identified from the broad-band imaging data. Although the study of some 1 000 000 galaxies is the principal motivation for the SDSS some 150 000 quasar candidates will be observed using the dedicated multifiber spectrograph mounted on a 2.5 m telescope at Apache Point in New Mexico. The broad-band filters employed in the SDSS extend from the near-ultraviolet through to a '*z*'-band at the far-red extreme of the optical atmospheric window. The *u* and *g* colors allow the detection of low-redshift quasars via the ultraviolet-excess technique with additional information from the redder filters, while sensitivity to radiation with wavelengths as red as 1000 nm means the SDSS is sensitive to quasars with redshifts as high as $z = 6$ (see below). The primary, 10 000 square degree, survey will identify relatively bright quasars, $B \leq 20$, and the number of very-high-redshift objects is not expected to be large. However, a smaller area of 100 square degrees will be surveyed to much fainter limits. It will also be possible to select subsets of potentially rare quasar candidates to rather fainter limits that may be followed up using telescopes of larger aperture. The first results from such a strategy are already encouraging, with a number of high-redshift quasar candidates observed in late 1998 producing the first quasar to be found with a redshift $z = 5$. More recently, a group of astronomers in California discovered a quasar with $z = 5.50$, and in April 2000 a group from Chicago reported a quasar with $z = 5.8$. In June 2001, two quasars with redshifts of 6.0 and 6.2, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

most distant objects ever observed in the universe, were identified in the Sloan Digital Sky Survey.

The use of the FIRST (FAR-INFRARED AND SUBMILLIMETRE TELESCOPE) radio source catalog to identify optically bright quasars has proved highly successful, with an initial catalog of more than 650 objects. The origin of the very large range in radio luminosity at fixed optical luminosity among the quasar population has been an outstanding question for several decades. The FIRST sample and future extensions to fainter optical magnitudes will make a significant contribution in this and a number of other areas. Similarly, the quasar samples to be defined from the CHANDRA and XMM x-ray satellites will be well suited to addressing key questions such as the contribution of quasars to the x-ray background and testing the predictions of unified schemes for quasars and active galactic nuclei.

The capabilities of optical telescopes with apertures of 8–10 m will allow fainter quasars at redshifts $z \sim 2$–3 to be investigated, helping to extend the dynamic range of the surveys. Probably of even greater significance will be the ability to conduct surveys for quasars at near-infrared wavelengths to relatively faint magnitudes over many hundreds of square degrees of sky. Extinction of radiation by dust is greatly reduced at wavelengths of ~2000 nm compared with optical wavelengths and quasars at very high redshifts, with very faint optical colors owing to the presence of intervening hydrogen clouds, should also be detectable in the near-infrared. Such infrared surveys will be possible early in the millennium and a resolution to the ongoing debate concerning whether a significant proportion of the quasar population has eluded detection owing to extinction by dust and gas, either associated with the quasars themselves or present in intervening systems, should result.

Bibliography

A general introduction to quasars and active galactic nuclei, including the principles behind surveys is

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

A more specific review of quasar surveys

Hewett P C and Foltz C B 1994 Quasar surveys *Publ. Astron. Soc. Pac.* **106** 113

Results from the FIRST Bright Quasar Survey can be found in

White R L *et al* 2000 *Astrophys. J. Suppl.* at press

The potential of the Sloan Digital Sky Survey for detecting quasars is demonstrated by the description of the discovery of a number of high-redshift quasars by the SDSS Collaboration;

Fan X *et al* 2000 *Astron. J.* **119** 1

Initial results from the Anglo-Australian Observatory 2dF quasar survey are described in

Boyle B J *et al* 2000 *Mon. Not. R. Astron. Soc.* at press

Paul Hewett

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Quetelet, Lambert Adolphe (1796–1874)

Statistician, born in Ghent, Flanders, Belgium, founder (1833) and director of the Brussels Observatory. Studied astronomy at the Paris Observatory under FRANÇOIS ARAGO, and probability under Joseph Fourier and PIERRE LAPLACE. Apart from social statistics (crime, mortality, census taking), he worked on statistical, geophysical and meteorological data, and established statistical methods. Following a suggestion of Arago, he studied the statistics of the visibility of meteors and showed that there was a greater frequency of meteors in the month of August. He found archival records of earlier exceptional meteor displays in August, and predicted the shower (the Perseids) of 10 August 1837 (see EDWARD CLAUDIUS HERRICK). He published the *Catalogue des Principales Apparitions d'Étoiles Filantes* (1839) listing 315 meteor displays recorded until then.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

R Coronae Borealis Stars

The R Coronae Borealis stars (RCB) stars fade dramatically and unpredictably by factors of up to 1000 within a few weeks. Over succeeding months, they gradually recover their original brightness. This spectacular fading is caused by the formation of sooty dust clouds above the surface of the star. The surfaces of RCB stars are unusually poor in hydrogen, and rich in carbon and nitrogen, which implies that they are the remnants of evolved stars. Practically all RCB stars pulsate, which may help to explain the dust formation episodes.

R Coronae Borealis, or RCB, stars are an important class of VARIABLE STAR. Stars vary in many different ways, often by changing the amount and color of the light they emit. They can change regularly or irregularly. RCB stars, remarkably, do both.

RCB stars are best known because of their irregular 'fades'. Unpredictably and every few years or so, their brightness suddenly drops by a huge amount. Then they get slowly brighter until they look the same as before the fade. The fades are caused by enormous clouds of dust which cover a large part of the star's surface. The dust is thought to be made of small particles of carbon—rather like ground-up pencil sharpenings.

Why the dust clouds form is harder to understand. As well as irregular fades, the light output from RCBs goes up and down by a few per cent every 40 days or so. This is due to the star swelling and shrinking—known as pulsation. It seems that these pulsations may cause dust to form which is then thrown out from the star. Thus, indirectly, they may be responsible for the irregular fades.

Another remarkable thing about RCB stars is that their surfaces contain almost no hydrogen, the most common element in space. In fact, they are mostly made of helium and carbon. These elements are made by nuclear reactions deep inside stars. Since they are now seen on the surface, RCB stars must be very old.

This article introduces the history and basic properties of RCBs, before exploring them in more detail. Three brief sections address the origin of RCBs, the occurrence of RCB activity in other stars, and recent progress towards solving some of the outstanding problems.

Historical background

RCB stars are named after a star just visible to the naked eye in the constellation of Corona Borealis. R CrB was discovered by Edward Pigott, an English amateur astronomer who had been observing a star in the 'Northern Crown' for over a decade. In the spring of 1795 he noticed that it had disappeared, but over succeeding months it gradually reappeared and recovered its original brightness. Pigott estimated a period of 10.5 months for this phenomenon, but he was puzzled that the changes did not repeat themselves precisely.

In fact, unlike δ CEPHEI and other variable stars known at the time, R CrB's changes in brightness are irregular and unpredictable. This irregularity has prompted numerous astronomers to observe R CrB so regularly that its light curve has now been monitored continuously for almost 200 years.

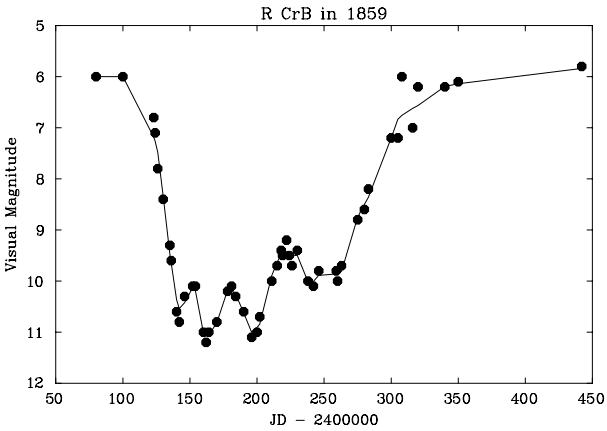


Figure 1. The light curve of R Coronae Borealis during a fading event in 1859.

Since 1795, other RCB stars have been discovered and their properties have been carefully observed. The defining property is their characteristic light curve, marked by the irregular fadings first seen by Pigott. In addition, nearly all RCB stars pulsate with periods between 40 and 100 days. Their surfaces contain very little hydrogen, unlike the vast majority of stars which have about 90% hydrogen. Instead their surfaces are rich in helium and carbon, which must have been produced by nuclear reactions deep within the star.

It is established that the sudden fadings of RCBs are caused by clouds of carbon-rich dust passing in front of the star, but it is unclear what causes carbon gas in the stellar atmosphere to condense into dust. Other questions remain about the precise composition and origin of RCB surfaces.

Remarkable as their properties are and extensive as two centuries of observations have been, RCB stars continue to pose many problems. The irony is that the original question posed by Pigott's observations in 1795 remains essentially unsolved—'what causes the sudden fadings of R CrB ?'

Basic properties

RCB stars were first identified by their characteristic light curves. A part of the light curve of R CrB itself is shown in figure 1.

Spectrum

Many light-curve identifications were made retrospectively from photographic plates originally obtained to patrol the sky for asteroids, supernovas and other variable phenomena. As a consequence the light curves were often incomplete so that symbiotic, cataclysmic and semiregular variables were sometimes identified as RCBs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

RCB stars have a spectrum which resembles that of an F or G supergiant. Balmer lines are weak or absent, there are many lines of neutral atomic carbon and strong bands of molecular carbon (C_2 and CN). In one of the earliest analyses of stellar surface composition, it was shown that R CrB itself was extremely hydrogen deficient. A handful of hot RCB stars have B-type spectra. These characteristics have enabled the identification of 34 RCB stars in the Galaxy to be confirmed.

Galactic distribution and absolute magnitudes

The distribution of galactic RCB stars is controversial. Several authors have argued that they belong to a thick galactic disk, and are thus old Population I stars. However, they may alternatively belong to the even older galactic bulge (Population II). The controversy may simply be a result of the different ways in which surveys for RCB stars and related objects have been carried out. In either case, the populations are sufficiently old that RCB stars must have relatively low masses ($\leq 1 M_{\odot}$).

A growing number of RCBs have been discovered in the LARGE MAGELLANIC CLOUD (LMC), most recently as a result of wide field surveys. These are the only RCB stars for which a distance is known and hence for which an absolute magnitude can be deduced. With $M_V \sim -4$ to -5, the LMC RCBs are 3000–10 000 times more luminous than the Sun.

Numbers and distances

The absolute magnitude of the RCBs in the LMC and the small number of galactic RCBs implies that they are very rare stars. Together with statistics for related stars, the total number of hydrogen-deficient stars in the Galaxy is estimated to be about 1000. Whilst the luminosity of RCBs is similar to that of other low-mass stars as they evolve rapidly from the asymptotic RED GIANT branch to become WHITE DWARFS, their frequency is much lower. They do not, it seems, represent a stage in the evolution of all normal low-mass stars.

From the preceding estimates of mass and luminosity, it is seen that RCBs have very high luminosity-to-mass ratios ($L/M \sim 10\,000$) and lie close to the Eddington limit for radiative stability. If they were any more luminous, their atmospheres would be expelled by radiation pressure.

Physical characteristics

Atmospheres

The fundamental properties of RCBs may be deduced from their spectra at maximum light. A typical spectrum looks like that of an F or G supergiant. RCB effective temperatures fall mostly around 7000 K, although a small number are as cool as 5500 K. The spectra of most RCBs have now been analysed in detail. They show that hydrogen makes up less than one part per thousand down to less than one part per million of the STELLAR ATMOSPHERE. Most of the atmosphere is neutral helium. Although carbon is enriched, it is not possible to establish its abundance reliably. All RCB atmospheres show products of hydrogen- and helium-burning NUCLEOSYNTHESIS in several different episodes, as well as mild enhancements of s-process elements.

Pulsations

Most RCB stars are small-amplitude variables at maximum light. Light variations are typically a few tenths of a magnitude with periods between 40 and 100 days. In well-studied cases, radial-velocity variations have been found with amplitudes proportional to the light variations. The visual light curve of RY Sgr (period ~38 days, amplitude 0.2–0.3 mag) correlates well with the radial velocity curve (amplitude 40 km s^{-1}) and indicates that RCBs are radially pulsating stars.

The light curves of RCB stars are not strictly periodic. Some authors suggest that this is due to interference between a number of radial modes with different periods. It is more likely due to extreme non-adiabacity in the stellar envelope which can lead to chaotic behavior in the pulsation cycle length. Since pulsation periods are directly related to the mean density of a star, evidence for period changes was thought to imply secular changes in RCB radii. It now seems likely that the observed period changes are due to chaotic behavior.

The driving mechanism for pulsations in RCB stars is not the classical κ mechanism found in δ Cepheid variables; RCB stars do not lie in any of the classical instability strips. Their pulsations are due to 'strange-mode' oscillations whereby waves trapped in parts of the acoustic cavity of the stellar envelope are amplified. Such pulsation modes are also seen in other HIGH-LUMINOSITY STARS, such as luminous blue variables and α Cyg variables.

A consequence of the high L/M ratios of RCB stars is that densities in their atmospheres are low. The action of the pulsations on the photosphere resembles that of a piston imparting an outward impulse at regular intervals. In between these impulses the photosphere is virtually in free-fall. During the impulse, highly nonlinear processes can generate shock waves within the photosphere. Evidence for such nonlinearity has been observed in RY Sgr where, at minimum radius, absorption lines are seen to double. A redshifted component is due to infalling material and a blueshifted component is due to lower-lying plasma that has already being accelerated outwards.

Frequency of fading events

The most obvious and spectacular properties of RCB stars are their dramatic fading events. They spend the majority of their time at maximum light. Initial decline is sudden and steep, the star can fade by up to eight magnitudes in a few weeks. The decline may show a series of standstills, partial recoveries and subsequent declines. The final recovery to maximum light may take several months to a year.

The first outline of a mechanism for the fading events was proposed in the 1930s when it was 'shown that the shape of the light curve of R CrB and its spectral variations

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

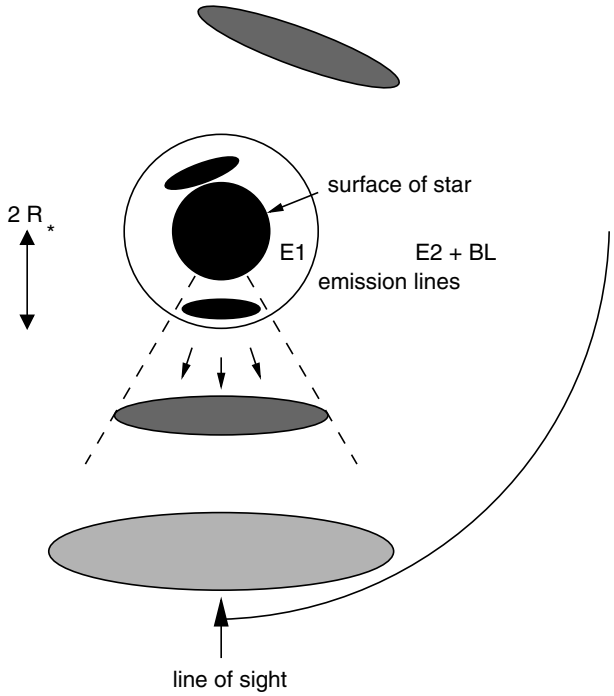


Figure 2. RCB dust-cloud ejection and evolution. (Based on a figure in Clayton (1996) and reproduced by permission.)

at minimum can be accounted for by supposing it to eject matter which condenses at a considerable distance and forms obscuring clouds. The solid matter is believed to be principally carbon.'

From nearly two centuries of observations of R CrB, a characteristic interval of 1100 days between fading events has been deduced. Several searches have consistently failed to find any regular patterns in RCB declines which might help to identify a causal factor such as rotation or binarity. For two cases only, RY Sgr and V854 Cen, it has been shown that the onset of the decline episodes may be closely correlated with a particular phase of the pulsation cycle. It is attractive to relate the onset of a dust formation episode to the pulsation of the stellar envelope, but the evidence for such a connection is not yet compelling.

The ratio between the characteristic interval between fades and the typical pulsation period of RCBs implies that a decline occurs about once every 20 pulsation cycles.

Anatomy of a decline

Infrared observations of RCBs show that, even at maximum light, there are copious amounts of dust surrounding the star. A very interesting result is that during a decline, when the V-band brightness is plummeting, the infrared L-band brightness shows no significant change and continues to show pulsational variations. It seems that the V-band brightness is dominated by the obscured atmosphere, but the L-band brightness is dominated by dust surrounding the star at some distance. Evidently the dust continues to be illuminated by light from the photosphere, including its pulsations, whilst light coming directly from the photosphere to the observer is obstructed. Therefore the obscuring dust can only cover a fraction of the stellar surface.

The picture that emerges (figure 2) is that at a particular phase in the pulsation cycle, conditions in the photosphere can compel the carbon-rich gas to be expelled from a part of the stellar surface to cool, condense and form dust grains. Roughly once every 20 cycles, this 'puff of soot' is directed into our line of sight, and obscures the photosphere from view.

Another important observation associated with RCB declines is that, during the early part of the decline, a rich emission spectrum appears, consisting of many narrow lines of neutral and singly ionized metals. These lines (known as E1 lines) appear to be short-lived so that within two or three weeks, they are replaced by a simpler broad-line spectrum (BL), although a few narrow lines (E2) persist for an extended period. The E1 spectrum is probably due to hot gas close to the stellar surface which is visible only when the photosphere alone is obscured. As the dust cloud moves away from the stellar surface it expands, obscuring the E1 region. Since the E2 and BL regions remain visible, they must arise from cooler gas spread over a much larger volume around the star.

This picture of clouds of dust ejected by the pulsating stellar surface has been successful in explaining many properties of RCBs. There remain a number of difficult questions, the hardest being 'What is the origin of RCB stars?'

The evolutionary status of RCBs

Nuclear processes within stars first convert hydrogen to helium, then helium to carbon, and eventually to heavier elements. RCB surfaces consist primarily of a mixture of helium and carbon-rich layers, with a trace of hydrogen. Single-star evolution does not normally succeed in mixing such different layers of a star, so special models have been proposed to explain their origin.

In one model, a low-mass star which has finished its evolution as a red giant contracts to become a white dwarf, passing through a phase when it illuminates a PLANETARY NEBULA. It may happen that sufficient unprocessed helium remains on the surface of the white dwarf that nuclear reactions can be reignited. The star then expands suddenly to become a helium-burning red giant for a second time. Convection will thoroughly mix the outer layers of this star to give the mixture of helium and carbon seen in RCBs.

In another model, it is supposed that two white dwarfs are in orbit around one another. Over a long time scale ($\sim 10^{10}$ years), GRAVITATIONAL RADIATION will make the orbit decay and the stars will spiral in towards one another. If one star is a helium white dwarf (HeWD), and the other a carbon/oxygen white dwarf (COWD), the HeWD will be cannibalized by the more massive COWD. This helium will be capable of initiating new nuclear reactions and,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

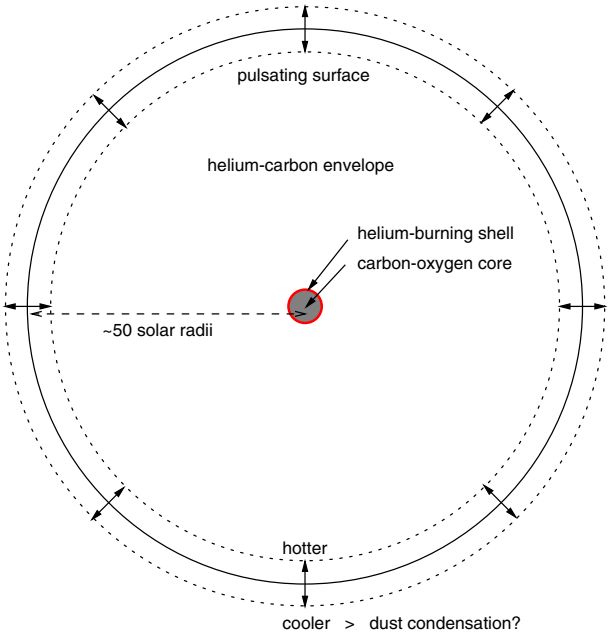


Figure 3. Internal structure of an RCB star.

like the previous model, the star will expand to become a helium-burning giant, with a helium- and carbon-rich surface.

It is difficult to resolve which model is correct, if either, because RCBs show a range of surface compositions, and the crucial carbon abundance is not well known. However, it is agreed that RCBs probably have a degenerate carbon-oxygen core, and their energy comes from a nuclear-burning shell at the bottom of the helium-carbon envelope (figure 3).

Other hydrogen-deficient stars

The Galaxy hosts a variety of other extraordinary stars with little or no surface hydrogen. Most are rare, but nevertheless intriguing. Massive WOLF–RAYET STARS are young (Population I) hydrogen-deficient stars, as are binary systems such as υ Sgr. Low-mass stars like the RCBs belong to an old population. A group of hydrogen-deficient giants with spectra similar to RCBs show no RCB fading events. These have probably failed to become RCB stars because their luminosities are too low.

As well as RCB stars with F- and G-type spectra, several other stars show RCB-type light curves. Two, MV Sgr and DY Cen, are B-type helium-rich giants and are very similar to a group of about 20 'extreme helium stars', A- and B-type stars with weak or no Balmer lines, strong neutral helium and ionized carbon lines. If they are indeed RCB stars, they are not very active. Their surface carbon abundance is ~1% by number, the remainder being helium. A more active B-type helium star, V348 Sgr shows strong emission lines at all times, and has a surface carbon abundance of ~10% by numbers. This star may be related to about 20 hydrogen-deficient central stars of planetary nebulae which have similar carbon abundances and strong emission line spectra. These stars are all overluminous and have strong stellar winds. Carbon-rich knots may form in the winds and give rise to RCB behavior. Of importance to understanding the RCB fading events, the discovery in 1997 of weak RCB-like fading events in CARBON STARS suggests that the process of carbon condensation in stellar atmospheres is not confined to classical RCB stars.

During the past century three stars have been observed as being both hot subdwarfs and RCB stars at different times. V605 Aql is now the hydrogen-deficient central star of a planetary nebula (Abell 58). In 1919 it brightened as a slow nova, and there is a report that its spectrum at one time resembled an RCB star. FG Sge was a faint blue star in 1900. Since then, it has become progressively redder and brighter so that now it has the spectrum of a K-type giant. In 1960 it started to pulsate and in 1992 it started to show RCB-type fading events. Its surface composition contains the products of many nuclear reaction processes, and is thought to be hydrogen poor. FG Sge may be a new-born RCB star. Most recently, in 1996, an unremarkable faint blue star suddenly brightened and within three years had become a very cool and luminous carbon star, now known as SAKURAI'S OBJECT or V4334 Sgr. The spectacular evolution of these three stars and their similarities with RCB stars may eventually help to explain the origin of RCB stars.

Unsolved problems

Since the discovery of R CrB itself, the mechanism that produces fadings has been elusive. Primary data connecting pulsation phase and the trigger for fadings exist for only two RCBs (V854 Cen and RY Sgr), and protracted photometry of several RCBs will be necessary to establish any connection firmly.

A second difficulty is encountered by the physical conditions necessary for dust to condense above the surface of the star. The frequency and duration of fading events implies a geometry in which the dust clouds form within two stellar radii ($2R_\star$). Under normal conditions, the local temperature would be too high for dust to condense at this distance, and a condensation distance of $20R_\star$ would be expected. Recent models treat the chemistry, energy balance and dust nucleation in pulsating star atmospheres in considerable detail. They show that excess cooling can occur during adiabatic expansion after the passage of a shock wave, reducing the local temperature to about 1500 K within 1.5–3$R_\star$. It remains to be shown that pulsations in all RCB stars provide the necessary conditions for dust nucleation to occur at this distance.

The RCB carbon abundance remains an enigma. Whether it is ~1% or ~10% determines if RCBs are related to the C-rich remnants like the extreme helium stars or to C-strong remnants like the H-deficient central stars of planetary nebulae.

The problem is a natural consequence of stellar atmosphere physics. The strength of a stellar absorption

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

line represents the ratio of line opacity to continuous opacity in the atmosphere; both are related to the number density of the absorbing atoms and so provide the number ratio of line absorbers to continuum absorbers. When the line and continuum absorbers are the same, it becomes impossible to measure the abundance of the absorbing species. In hydrogen-rich stars, hydrogen is normally the main continuum absorber and, since it is also assumed to be the most abundant species, the problem does not arise. Since the predominant continuum absorber in RCB atmospheres is neutral carbon, it has not yet proved possible to measure the carbon abundance from carbon absorption lines.

Finally, the emission line spectrum seen during fading events is difficult to explain, especially the sodium D lines which indicate an expansion (or wind) velocity of several hundred km s^{-1}. Superimposed are a number of narrow absorption components, possibly representing cooling gas from previous ejections. The presence of heavily blueshifted absorption from infalling material remains a puzzle.

RCB stars offer the astronomer a wealth of information, from their spectacular fadings, through their pulsations and dust shells, to the nuclear waste on their surfaces that traces previous evolution. Even with all of these data, RCBs retain much of their original mystery.

Bibliography

Most general astronomy texts and books on stellar variability refer to RCB stars, but only briefly. For example:

Hoffmeister C, Richter G and Wenzel W 1985 *Variable Stars* (Berlin: Springer)

In contrast, the academic literature is extensive and this article can only provide a few pointers. A recent and comprehensive review provides many important references:

Clayton G C 1996 The R Coronae Borealis stars *Publ. Astron. Soc. Pacific* **108** 225–41

A conference held in 1995 was devoted to hydrogen-deficient stars in general. Its proceedings contain many excellent reviews covering recent work on RCBs and related objects:

Jeffery C S and Heber U (eds) 1996 *Hydrogen-deficient Stars (Astr. Soc. Pac. Conf. Ser. 96.)*

Simon Jeffery

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rømer [Römer, Roemer], Ole [Olaf] Christensen (1644–1710)

Born in Aarhus, Denmark, studied at the University of Copenhagen under Thomas and Erasmus Bartholin, who gave him TYCHO BRAHE's manuscripts to edit and his own daughter to wed. Rømer accompanied Bartholin and JEAN PICARD to Hven to measure the position of Tycho's observatory, the better to reduce Tycho's observations. He went on to the Paris Observatory where he made and used instruments for the measurement of star positions. While in Paris, he timed the eclipses of Io, the innermost of Jupiter's moons. He discovered that the period between eclipses was changing. The change of period was due to a Doppler shift, caused by the changing velocity of Earth relative to Jupiter, and led Rømer to a value of 11 minutes for the value of the light travel time from the Sun to Earth (the 'light equation'). The modern value is 499 s. Rømer's explanation of the eclipses of Io was criticized by CASSINI, who could not repeat it with the other Galilean satellites, and by an incredulous ROBERT HOOKE. Hooke maintained that light travels instantaneously, and ''tis so exceeding swift that 'tis beyond Imagination; for so far he thinks indubitable, that it moves a Space equal to the Diameter of the Earth, or near 8000 miles, in less than one single Second of the time, which is in as short time as one can well pronounce 1, 2, 3, 4: And if so, why not be as well instantaneous I know no reason. . .'. Rømer was supported by HUYGENS and NEWTON. Rømer became professor of astronomy and of mathematics in Copenhagen, astronomer royal to Kings Christian V and Frederick IV of Denmark, and director of the Copenhagen Observatory. He carried out a succession of state tasks, including development of a reformed measurement system for Denmark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radar Astronomy

The branch of astronomy that is concerned with the investigation of solar system bodies by radar methods. The technique involves transmitting a beam of microwave radiation from an Earth-based radio telescope or an orbiting spacecraft towards a target and analyzing the faint 'echo' that returns from the target's surface. Radar techniques can yield information on the precise distance between the Earth and the target body, the rate at which the target body is rotating, the altitude of a spacecraft above a body's surface, the vertical relief and topography on the surface of a planet or satellite, and the nature and roughness of that surface.

The transmitted beam may be continuous (continuous wave) or pulsed. Typical pulses transmitted by Earth-based astronomical radar systems have lengths (durations) of between 10 μs and 10 ms (10^{-5} to 10^{-2} seconds) and peak powers of tens of megawatts. Pulsed radar is used to measure planetary distances. During the time interval (Δt) between the transmission of a pulse and the arrival back at Earth of the reflected pulse, the signal has traveled, at the speed of light (c) a distance equal to twice the distance between Earth and the target planet. Therefore the distance (d) between Earth and the target planet is given by: $d = c\Delta t/2$. The distance to a planet (Venus) was measured for the first time in 1962 and, from this result, the value of the astronomical unit (the mean distance between Earth and the Sun) was calculated to a much higher precision than had previously been possible. The altitude of a spacecraft above any particular point on a planet's surface (and hence the heights of features on the planet's surface above or below the mean surface level) can be measured in a similar way.

If the target body is rotating, part of the reflected radar pulse will be reflected from the side of the planet that is approaching the observer and part from the side that is receding. Compared with the transmitted signal, the part that has been reflected from the 'approaching' side will be Doppler shifted to a higher frequency and the part from the 'receding' side Doppler-shifted to a lower frequency. The spread of frequencies in the returning signal, therefore, provides information on the rotation rate of the target body. This technique enabled the retrograde rotation period of cloud-covered Venus to be determined in 1962 and the 59-day rotation period of Mercury to be determined in 1965.

Detailed mapping of the surface topography of Venus has been carried out using ground-based radar and radar instrumentation carried on a succession of spacecraft that have been placed in orbit round that planet, including Pioneer Venus (1978), Veneras 15 and 16 (1983) and Magellan (1990–4). The key to achieving high-resolution radar mapping is synthetic aperture radar (SAR), a technique that uses the movement of the spacecraft in the time interval between the transmission of the signal and the return of its 'echo' to simulate the performance of a 'dish' much larger than the onboard radar dish. By this means, the Magellan mission achieved best resolutions of about 120 m on the surface of Venus. Radar techniques have been used to study meteors, planets, satellites, comets and asteroids.

See also: Magellan, Pioneer Venus, radio telescope, resolving power, Venera.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radial Velocities

The radial velocity of a star is the component of its motion along the line of sight to the observer. A typical value for the radial velocity of a star in the solar neighborhood is 10–20 km s^{-1}. The early work with radial velocities, dating back more than 100 yr, is reviewed in the first section of this article. Then the efforts over the past 25 yr to survey many thousands of stars with correlation techniques are described, together with examples of applications to Galactic structure and to studies of the frequency and characteristics of spectroscopic binaries. Recent efforts to achieve velocity precisions better than 10 m s^{-1} are outlined next, along with their successful use for the discovery of extrasolar planets. The final section speculates about the future of radial velocity research, especially in conjunction with space astrometric missions.

Early radial-velocity work

The measurement of stellar radial velocities has a long and rich history. KIRCHHOFF and Bunsen electrified the scientific and intellectual world in 1862 with their wonderful demonstration that spectral analysis could be used to unlock the chemistry of the Sun and stars. Doppler and Fizeau had already proposed, independently in 1842 and 1848, that the motion of a star along the line of sight would introduce a shift in its spectral features, such as the dark Fraunhofer absorption lines, according to the formula[1]

$$\frac{\lambda - \lambda_0}{\lambda_0} = \frac{v}{c} \qquad (1)$$

and Kirchhoff and Bunsen's discoveries inspired several stellar spectroscopists to look for Doppler shifts in the spectra of stars.

For example, at his private observatory in England, the great spectroscopist SIR WILLIAM HUGGINS compared the position of a prominent hydrogen emission line from a Geissler tube with the position of the corresponding hydrogen absorption line in the spectrum of Sirius. Huggins reported the detection of a Doppler shift for Sirius in 1868, but his result was off by more than 50 km s^{-1} and did not even have the correct sign; Sirius is moving towards the solar system, not away from it. Except in the hands of JAMES KEELER with the great 36 in refractor at the Lick Observatory on Mt Hamilton in California, visual spectroscopy was unable to achieve the precision needed to make sense out of the radial velocities of nearby stars.

A breakthrough in this field came with the application of photography to stellar spectroscopy, which led to an order-of-magnitude improvement in the precision of the measurements. HERMANN CARL VOGEL at Potsdam pioneered the use of dry plates for radial-velocity work and was able to achieve a precision of 2 or 3 km s^{-1} for bright stars.

Leadership in this field soon moved to the Lick Observatory, where WILLIAM WALLACE CAMPBELL had taken over the reins from Keeler. With generous long-term financial support from D O Mills, Campbell and his staff were able to undertake an ambitious survey of all the stars brighter than visual magnitude 5.5, first with the 36 in refractor in the north, and subsequently with a 36 in reflector at a southern station near Santiago in Chile. For solar-type stars the precision of their measurements was about 1 km s^{-1}, more than good enough for studies of Galactic structure, and well suited for solving the orbits of SPECTROSCOPIC BINARY STARS. The great catalog reporting the results of this work was published by Campbell and Moore in 1928, with center-of-mass velocities listed for 57 spectroscopic binaries among the 2771 stars in the catalog.

Photographing stellar spectra for the purpose of measuring radial velocities continued to be a major enterprise at several observatories on the North American continent up through the time of the Second World War, especially with the telescopes at the Mt Wilson Observatory, built by GEORGE ELLERY HALE. The fruits of these labors were summarized in 1953 by Ralph E Wilson in a catalog of mean radial velocities for more than 20 000 stars, with the Mt Wilson and Lick Observatories contributing more than half of the total between them.

Radial-velocity measurements using correlation techniques

In that same year, 1953, a revolutionary new approach to measuring stellar radial velocities was proposed by Peter Felgett. One of the main goals of the new scheme was to take advantage of the much higher quantum efficiency of photoemissive cathodes, which offered a gain of at least a factor of 10 compared with photography. The best photographic emulsions were barely able to register 1% of the incident light, even when hypersensitized for the low light levels encountered in ASTRONOMICAL SPECTROSCOPY. Felgett's second idea was to measure many FRAUNHOFER LINES simultaneously, thus gaining a major advantage over the traditional procedure of measuring the positions of only a small number of lines carefully selected to be unblended.

Correlations using a physical mask

Felgett's ideas were implemented by Roger F Griffin with a new instrument at the Cambridge Observatory. A physical mask containing 240 fine slots, chosen to match selected absorption lines in the spectrum of the K2 giant ARCTURUS, was placed at the focus of the spectrometer, and the light passing through the slots was gathered onto the cathode of a photomultiplier tube. By driving the mask back and forth in the dispersion direction with a micrometer, it was possible to determine the shift of the mask that gave the best match to the stellar spectrum being observed. The shift of the mask could then be converted into a Doppler shift using a calibration curve. Griffin used this

[1] Where c is the speed of light, v is the velocity of the star along the line of sight to the observer, λ_0 is the wavelength measured in the laboratory for a chosen spectral feature and λ is the wavelength observed for the same feature in the spectrum of the star. A negative velocity, or blueshift, indicates that the star is approaching the observer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

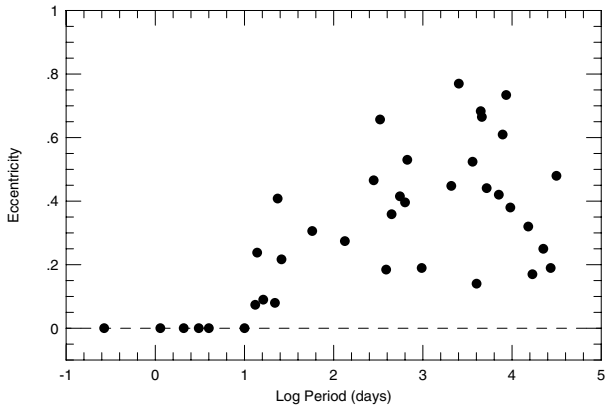


Figure 1. Eccentricity versus log(period) for the stellar companions orbiting nearby solar-type stars.

instrument for nearly 30 yr to monitor the radial velocities of thousands of stars. One of the main achievements of this work was a series of more than 150 papers on the orbits of spectroscopic binaries.

Similar instruments with physical masks were put into operation at several other observatories, including the Dominion Astrophysical Observatory near Victoria, British Columbia, and the 200 in Hale Telescope on Palomar Mountain. Especially noteworthy were the CORAVELs, operated for many years by a team led by Michel Mayor at the Geneva Observatory, both on the 1 m Swiss telescope at the Observatoire de Haute Provence in France and on the 1.5 m Danish telescope at the European Southern Observatory (ESO) on La Silla in Chile. These two, nearly identical, instruments also used physical masks, but were based on a spectrometer design that employed a cross-dispersed echelle grating. This provided significantly larger spectral coverage than Griffin's spectrometer, with the result that the CORAVELs could go fainter, easily reaching to visual magnitude 13. Altogether the two CORAVELs accumulated 286 181 measurements of 60 811 stars, with velocity precisions mostly in the range 0.3–1.0 km s^{-1}.

The CORAVELs were used to study pulsating stars, the memberships and kinematics of stars in clusters, and especially the frequency and orbital characteristics of spectroscopic binaries in a variety of stellar populations, including open clusters and globular clusters. Perhaps the most famous result from the CORAVELs was a comprehensive summary of the stellar companions found in a volume-limited sample of 164 solar-type dwarfs within 22 pc of the Sun. Figure 1 shows the distribution of eccentricity versus period for the binaries in this sample with orbital solutions. Notice that there is a transition at about 10 days. For longer periods the orbital eccentricity shows a wide distribution with a median value of about 0.3, but for periods shorter than 10 days the orbits all have very small eccentricities, presumably because they have been circularized by tidal mechanisms.

In addition to measuring radial velocities, it was possible to calibrate the shape and area of the cross-correlation dips produced by the CORAVELs, in order to derive rotational velocities up to about $v \sin i = 50$ km s^{-1}, and even metallicities.

Correlations with digital template spectra

Physical-mask instruments suffer from the disadvantage that the detector provides just a single channel. Only one wavelength can be registered at a time, as the mask is scanned through the range of Doppler shifts expected for the star. This shortcoming can be overcome by using an imaging detector directly at the focal plane of the spectrograph to record simultaneously all the wavelengths in the spectrum, in just the same way that photography was used to make a spectrogram, except that a modern digital detector can now be used. The determination of the Doppler shift can then be carried out digitally in software by cross-correlating a template spectrum against the observed spectrum.

Perhaps the most dramatic advantage of this approach is that silicon diode arrays, such as large-format CHARGE-COUPLED DEVICES (CCDs), can be used as the detectors, and these can have detective quantum efficiencies that are so good that they detect nearly all the incident light over a wide range of wavelengths from the ultraviolet to the near infrared. Also, with the advent of large array detectors with good sensitivity in the near infrared, this approach can now be extended to considerably longer wavelengths.

A second advantage of digital cross-correlations is that the template spectrum can be chosen after the fact from a large library of templates and thus can be optimized to match the temperature, rotational velocity, metallicity and surface gravity of the target star. If better templates become available, or better software is developed, then the analysis can be carried out all over again from scratch. With a physical mask spectrometer, you are stuck with the pattern of slots that was built into the mask.

The digital speedometers developed at the HARVARD–SMITHSONIAN CENTER FOR ASTROPHYSICS (CfA) are early examples of instruments designed to use digital correlation techniques. They used photon-counting intensified Reticons for the detectors, and normally recorded about 50 Å centered near 5187 Å in a single order of an echelle spectrograph. Operationally the photon-counting Reticons had the advantage that they provided a real-time display of the incoming count rates and of the spectrum as it accumulated. This was useful for optimizing the telescope guiding and for deciding when to end an exposure. However, because of the limited spectral coverage and a peak detective quantum efficiency of not much more than 10%, these detectors fell far short of what is now obtainable with CCDs.

To calibrate the wavelength scale, every observation of a stellar spectrum with the CfA digital speedometers is bracketed by an exposure of a thorium–argon hollow-cathode lamp. This naturally establishes the scale for converting a wavelength shift to a Doppler shift, but it

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

leaves open the possibility of a systematic error in setting the zero point of the velocity system. Exposures of the dawn and dusk sky are obtained every night, and these are used to monitor the velocity zero point (the actual radial velocity of the Sun can be calculated very accurately for any given observatory and time, from the known orbit and rotation of the Earth). As a double-check on the velocity zero point, more than a thousand observations of several dozen minor planets have been made over the years (absolute radial velocities can be calculated very accurately from the known astrometric orbits of the minor planets). The velocity zero point based on minor planet observations agrees with the dawn and dusk sky calibration within about 100 m s^{-1}.

Initially the CfA digital speedometer spectra were analyzed using observed spectra as the templates. Subsequently a large library of more than 60 000 synthetic spectra was calculated by Jon Morse using Kurucz model atmospheres. Not only did this make it possible to choose an optimum template during the reductions, but it also opened up the possibility of determining temperatures, rotational velocities up to about $v \sin i = 140$ km s^{-1}, metallicities and even surface gravities from the same spectra. Another important advantage of having a closely spaced grid of synthetic template spectra came in the analysis of composite spectra, such as the spectra of double-lined spectroscopic binaries, using a two-dimensional correlation technique, TODCOR. The choice of templates for the primary and secondary stars could be optimized simultaneously for the TODCOR analysis.

Three nearly identical CfA digital speedometers were put into operation, on the 1.5 m Wyeth Reflector at the Oak Ridge Observatory in the town of Harvard, Massachusetts, and on the 1.5 m Tillinghast Reflector and MMT, both located at the F L Whipple Observatory atop Mt Hopkins, Arizona. With the MMT it was possible to reach stars as faint as visual magnitude 16. Altogether nearly 200 000 radial velocities have been measured with the CfA digital speedometers, with velocity precisions mostly in the range 0.5–1.0 km s^{-1}.

A major application of the CfA digital speedometers was a study of the structure and early history of the Galaxy using a sample of nearly 1500 nearby stars selected for high PROPER MOTION. An example of how these observations were used to study the relationship between kinematics and chemistry is shown in figure 2, which plots the Galactic V velocity (the component in the direction of Galactic rotation, relative to the local standard of rest) versus metallicity, [m/H] (the logarithm of the metal-to-hydrogen ratio compared with the Sun). The broad distribution of stars centered near $V = -210$ km s^{-1} and [m/H] $= -1.5$ is the halo population, the first generation of stars that formed even as the Galaxy was still collapsing out of a vast intergalactic cloud composed of metal-poor gas and dust. These stars are distributed roughly in a giant sphere centered on the Galaxy, and on average they do not participate in the rotation of the Galaxy. As the Sun passes them by in its circular orbit around the center of the Galaxy, the halo stars appear to be streaming past us into the solar system's wake. The more concentrated distribution of stars near $V = 0$ and [m/H] $= 0$ is the disk population, stars that lie in a highly flattened disk and orbit the center of the Galaxy along with the Sun. These stars formed after the Galaxy had collapsed into its present disk configuration and after the interstellar medium had been enriched by initial generations of massive stars that finished their lives quickly and spewed metals from their interiors, where they were created by nuclear reactions, back into the raw material for the next generation of stars.

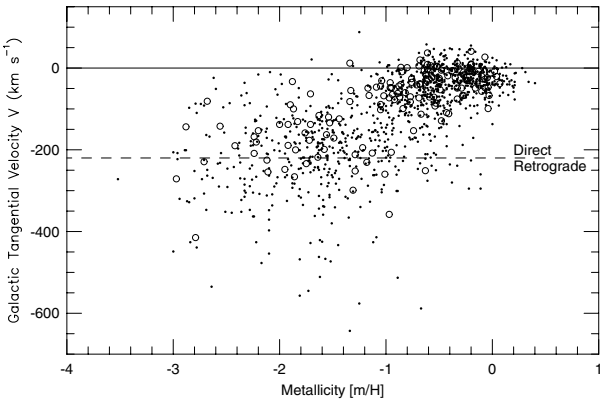


Figure 2. Galactic V velocity versus metallicity for the stars in a large sample selected for high proper motion. The single-lined binaries are plotted as open circles, the stars without detected velocity variations are plotted as small filled circles.

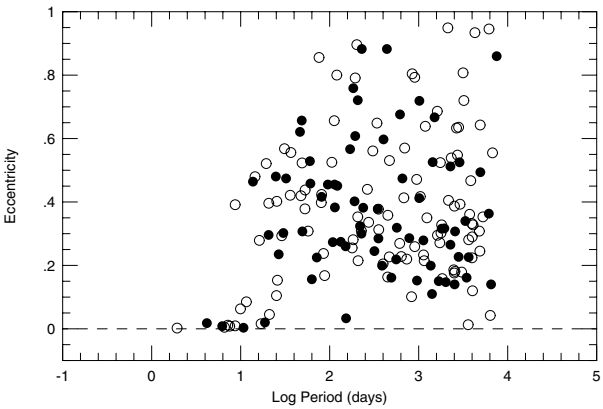


Figure 3. Eccentricity versus log(period) for the spectroscopic binaries in the proper-motion sample. The halo stars are plotted as filled circles, the disk stars as open circles.

Long-term monitoring of the radial velocities of the stars in the proper-motion sample over a span of more than 15 yr revealed that the frequency and orbital characteristics of spectroscopic binaries are pretty much the same for solar-type dwarfs in the halo as in the disk populations of the Galaxy. Figure 3 shows the distribution of eccentricity versus period for all the binaries with orbital solutions in the proper-motion sample. The halo stars are plotted as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

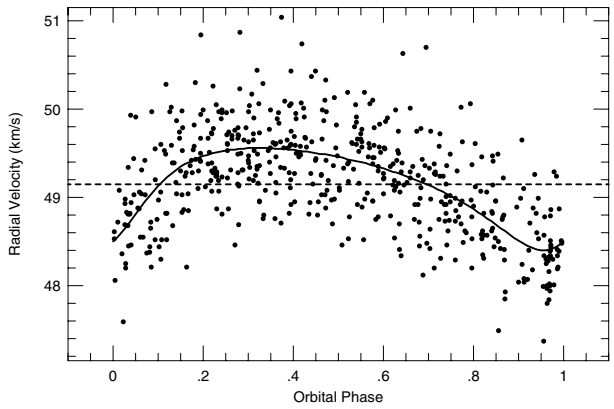


Figure 4. The orbital solution for HD 114762. The theoretical velocity curve is plotted as a solid line, the individual velocity measurements from the CfA digital speedometers are plotted as filled circles.

filled circles and the disk stars as open circles. The two populations show similar distributions, with the exception that the transition from circular to eccentric orbits occurs at a longer period for the halo binaries because they are older and tidal circularization has managed to reach out to binaries with wider separations and longer periods.

Together with the CORAVELs, the CfA digital speedometers have been used to monitor large samples of solar-type stars, from which optimum targets have been selected for extrasolar planet searches and searches for extraterrestrial intelligence (SETI), namely targets without evidence for a companion star that would interfere with the formation and/or survival of planets in the habitable zone.

A prophetic discovery, made with observations jointly from the CfA digital speedometers and the CORAVELS, was the detection in 1988 of an unseen low-mass companion orbiting the solar-type star HD 114762. This target had been especially well observed by both teams because it was one of the INTERNATIONAL ASTRONOMICAL UNION (IAU) radial-velocity standard stars. The mass of the companion could be as small as 10 Jupiter masses (M_J) if the orbit happens to be oriented edge-on to the line of sight (the orbital inclination cannot be determined from radial-velocity observations alone). The orbital characterisitic of this companion that dissuaded most astronomers from thinking it could be a giant planet was the relatively short period, $P = 84$ days. This implied that the companion was orbiting too close to the parent star to have formed the way that giant planets were thought to form, by core accretion in the icy regions of the stellar system, out beyond the snow line. Figure 4 shows the velocity curve for HD 114762's orbital solution, together with the velocities observed with the CfA digital speedometers. The observed velocities show a scatter that is similar to the amplitude of the orbital motion, but the solution is secure and gives reliable parameters because many cycles have been covered with a large number of observations.

Very precise radial velocities

A proper exploration of the realm of extrasolar planets using stellar radial velocities required the development of a new generation of instruments: ultrastable echelle spectrographs coupled to large-format CCDs as the detectors. Two examples of such instruments are the High Resolution Echelle Spectrograph (HIRES) built by Steven S Vogt for the Keck I telescope on Mauna Kea in Hawaii and ELODIE built by Michel Mayor's team for the 1.93 m reflector at the Observatoire de Haute Provence in France. With these instruments it is possible to record enough photons in a single CCD spectrogram of a slowly rotating solar-type dwarf so that the photon-noise limit to the determination of the Doppler shift is only a few m s^{-1}. At this level of precision, where the shift of the system of absorption lines must be determined to less than one-thousandth of the width of a typical line, it is the systematic drifts and image distortions in the spectrograph that set the most difficult limitations.

Gas absorption cells

Two different approaches have been developed for monitoring the spectrograph drifts so that systematic errors in the velocities can be corrected to better than 10 m s^{-1}. The first technique was pioneered by a Canadian team, who used a hydrogen-fluoride gas absorption cell to impress many sharp absorption lines of known wavelength on each stellar spectrum. An improvement of this technique was developed by R Paul Butler and Geoffrey W Marcy. For the absorbing gas they chose to use iodine, which has a rich spectrum of lines in the region 5000–6000 Å.

One approach to the reduction of these spectra is to use a very-high-resolution Fourier transform spectrometer (FTS) scan of the iodine lamp to model the temporal and spatial variations of the instrumental profile of the spectrograph. A high-quality spectrum of each star without the iodine absorption lines is also needed, to serve as a template, and velocities relative to the template are then determined for each stellar spectrum taken with the iodine gas absorption cell. With careful attention to the characteristics of the CCD readout electronics and detailed modeling of the instrumental profile, it has been possible to achieve a precision of a few m s^{-1} with HIRES. A version of the two-dimensional cross- correlation technique TODCOR has also been used to analyze stellar spectra taken with the iodine gas absorption cell and HIRES and has achieved a precision of 10 m s^{-1}.

Optical fiber feeds

A second approach for achieving very high precision is to use optical fibers to scramble the stellar light as it is fed onto the spectrograph slit and to monitor the drifts in the spectrograph during a stellar exposure using a simultaneous exposure of a thorium–argon comparison lamp. This approach has the advantage that it can provide wider spectral coverage than the limited window over which iodine absorption lines are available, at the cost of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

minor light loses in the optical fibers. This approach has been used both with ELODIE and with CORALIE on the 1.2 m Swiss telescope at ESO on La Silla. These instruments can both produce velocities with a precision of better than 10 m s^{-1}.

The speed of the Earth in its annual orbit around the Sun is nearly 30 km s^{-1}, and the daily rotation amounts to about 0.5 km s^{-1} at the Earth's equator. When reporting a radial velocity, it has been customary to remove these motions and to state the velocity relative to the Sun. Now that it is feasible to measure velocities with precisions of a few m s^{-1}, additional motions must also be taken into account, including the monthly orbit of the Earth and Moon and the orbits of the other planets in the solar system. In particular, the gravitational pull of Jupiter causes a reflex motion in the Sun of about 12 m s^{-1} with a period of about 12 yr. Thus, for very precise work it is necessary to refer radial velocities to the center of mass, or BARYCENTER, of the solar system.

Extrasolar planets

Most of the excitement in stellar radial-velocity research during the 1990s was the discovery of dozens of extrasolar planets orbiting nearby solar-type stars (see EXTRASOLAR PLANETARY COMPANIONS AND BROWN DWARFS). Perhaps the biggest surprise was the announcement in 1995 of a Jupiter-sized companion orbiting 51 Pegasi with a period of only 4 days. This placed the companion in a very tight orbit, 20 times smaller than the Earth's orbit around the Sun, and led to the development of new theoretical ideas for how to get giant planets to migrate inward from where they originally formed out beyond the snow line. The problem of how to stop the migration and park the planet in an orbit like that of 51 Pegasi remained unsolved.

A particularly significant discovery was the detection in 1999 of transits for the planet orbiting the solar-type star HD 209458. This marked the opening of a new era in extrasolar planet research, when the physical characteristics of extrasolar planets could be studied in detail. By measuring the amount of dimming the parent star undergoes as the planet transits in front of the star, it is possible to deduce how much of the stellar surface is being shadowed, which allows the radius of the planet to be determined, assuming the radius of the star can be estimated accurately. Moreover, from the shape of the light curve it is possible to deduce the exact inclination of the planet's orbit, thus eliminating any ambiguity in the mass of the planet. In turn this makes it possible to deduce the bulk density of the planet, simply by dividing the mass by the volume.

In the case of the planet orbiting HD 209458, the mass is $0.69 M_J$, the radius is $1.40 R_J$ and the density works out to be 0.31 g cm^{-3}, about one-third the density of water. This implies that the planet must be made mostly of gas, more or less like Jupiter and Saturn, but puffed up by the intense heating due to its close proximity to its parent star. However, despite this heating, the gravity of the planet is strong enough to keep its atmosphere from boiling away. The escape velocity from the planet's surface works out to about 40 km s^{-1}, while the typical thermal velocity for an atom in the planet's atmosphere is only a few km s^{-1}. The prospects for learning more about this planet are fascinating. For example, astronomers expect to be able to study the atmosphere of the planet by observing the absorption lines that are introduced into the spectrum of the parent star during transits, when the planet passes between us and the star and some of the starlight passes through the planet's atmosphere. Another possibility is to look for the light reflected from the surface of close-in giant planets and to measure radial velocities from the Doppler shift of the reflected stellar absorption lines.

Another important development was the discovery in 1999 of two more planets orbiting υ Andromedae, in addition to the close-in giant planet reported in 1996, making this the first planetary *system* to be discovered around a solar-type star. The outer two planets are more massive than Jupiter, and their orbits are more eccentric and smaller than that of Jupiter, so the υ Andromedae system does not look anything at all like our own solar system. Nevertheless, this is path-breaking progress towards understanding the characteristics of systems of extrasolar planets and whether or not our own system is unusual (see also EXOPLANETS).

Future developments for radial-velocity research

So far, we have not detected any true Jupiters, planets the size of Jupiter orbiting with periods of 10 or 20 yr. This does not mean that such systems do not exist. It does not mean that the configuration of our solar system is unique, with the terrestrial planets huddled in close to the parent star and the giant planets patrolling the outskirts, warding off incoming intruders such as asteroids or comets that might be instruments of mass extinction. Such a conclusion is premature, because we have so far barely achieved the Doppler precision that is needed to detect a true Jupiter, and with few exceptions the observations do not yet cover long enough time spans. Both these shortcomings are being addressed. New and better instruments and techniques are being developed. Gradually enough additional data with sufficient time coverage will be accumulated to detect true Jupiters.

It is not yet clear what the ultimate limit to the precision of stellar radial velocities will be. The limit may already have been reached for stars rotating more rapidly than a few km s^{-1} in $v \sin i$, where the velocity residuals are often larger than 10 m s^{-1}, perhaps because of stellar spots that distort the stellar line profiles as the star rotates. For giant stars there is good evidence that some astrophysical phenomena, such as oscillations in the stellar atmospheres, are the source of significant velocity jitter, often exceeding 1 km s^{-1} for the most luminous evolved stars. For well-behaved solar-type dwarfs that are hardly rotating and have no unusual surface activity, the ultimate limit to the velocity precision may turn out to be smaller than 1 m s^{-1}. Instruments that should be able to reach that level of precision are now under development.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The future of radial-velocity research will be intimately connected with the success of space astrometric missions, such as the Full-Sky Astrometric Mapping Explorer (FAME) and the Space Astrometric Mission (SIM) under development by the National Aeronautics and Space Administration (NASA). In the case of the European Space Agency's (ESA's) Global Astrometric Interferometer for Astrophysics (GAIA) there are even plans to include the capability to measure radial velocities of millions of stars, although the velocity precision cannot be much better than a few km s^{-1} because of photon starvation.

Data from the space astrometric missions will usher in a new era in the study of the structure and kinematics of the Galaxy and of stellar subsystems such as star-forming regions and clusters. Radial velocities will be an important part of this research, because they can be combined with proper motions and distances to derive space motions for literally millions of stars.

Another important area of study for the space astrometric missions will be the characterization of extrasolar planets. For most of the planets discovered by radial-velocity observations it should be possible to derive the orbital inclinations using ASTROMETRY, thus removing the ambiguity about the actual masses of the planets. Astrometry can also look for evidence of outer planets with orbital velocities too small to be detected using radial velocities. When applied to systems with known radial-velocity planets in short-period orbits, astrometry should be able to explore what kind of systems those planets come in.

Despite all the excitement about the discovery of extrasolar planets and the outstanding prospects for learning more about the characteristics of these planets and their systems in the coming years, one should not forget that we have a lot more to learn about the frequency and orbital characteristics of stellar companions in a variety of populations, building on the good progress in this field over recent years. Indeed, a full understanding of how and where planets form and evolve will also require a fuller understanding of the formation and evolution of stellar companions. Only then can we hope to understand how stars decide whether they will have planets or stars for companions.

Bibliography

Philip A G D and Latham D W 1985 (ed) *Stellar Radial Velocities (IAU Colloq. 88)* (Schenectady, NY: L Davis)

Hearnshaw J B 1986 *The Analysis of Starlight: One Hundred and Fifty Years of Astronomical Spectroscopy* (New York: Cambridge University Press)

Hearnshaw J B and Scarfe C D 1999 (ed) *Precise Stellar Radial Velocities (IAU Colloq. 170)* (San Francisco: Astronomical Society of the Pacific)

David W Latham

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radiant

The point on the celestial sphere from which the meteors of a particular meteor shower all appear to diverge. Meteoroids in a particular meteor stream impinging on the Earth's atmosphere do so on essentially parallel paths. That they appear to an observer to radiate from a single point is an effect of perspective. Some showers have more than one radiant, representing different strands of meteoroids within the meteor stream. Radiants move eastward at a rate of 4 min of RA per day as a result of the Earth's orbital motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radiation

The emission and propagation through space of electromagnetic energy (photons) or subatomic particles. Electromagnetic radiation, of which visible light is an example, can be regarded as a wave-like disturbance that travels through space at the speed of light, 300 000 km s^{-1}; it may also be regarded as a stream of particles, 'packets' of energy called photons. Examples of particle radiation include alpha and beta 'radiation' resulting from radioactivity, the solar wind, and cosmic rays. High-energy subatomic particles or electromagnetic radiation (gamma rays and x-rays) cause ionization and are harmful to living tissue. Most of the ionizing radiation absorbed by living creatures at the surface of the Earth comes from natural radioactivity and cosmic rays.

Electromagnetic radiation may be emitted or absorbed in a variety of ways. In general, any electrically charged particle will emit or absorb radiation when it is accelerated or oscillated. For example, if an electron moving freely through a gas, makes a close encounter with an ion, and is not captured by that ion, its kinetic energy will be changed by a finite amount, the change in energy corresponding to the emission or absorption of a photon (a 'packet' of electromagnetic energy). Radiation emitted in this way is called free–free radiation (the electron moves freely before and after the encounter). When a large number of electrons and ions are involved in these kinds of encounters (the assemblage of ions and electrons is called a plasma), their individual free–free emissions add together to produce a continuous spectrum, or continuum, of radiation. Because the temperature of the plasma determines the energies of the electrons, radiation produced in this way is often called thermal emission.

If a freely moving electron is captured by an ion, the kinetic energy that it loses is radiated away as a photon. Electromagnetic energy released in this way is called free–bound radiation (the electron, which was moving freely before the encounter, ends up bound to the ion). Again, where a large number of electrons are involved, with a spread of kinetic energies, free–bound emission takes place over a wide range of wavelengths, so producing a continuous spectrum. An electron that is bound to an atom exists in one of a restricted number of possible energy levels. If an electron drops from a higher to a lower energy level, it emits a photon of energy equal to the difference between the two levels; this corresponds to the emission of light of one particular wavelength. The various possible transitions (changes in level) correspond to emissions (or absorptions) at a discrete set of wavelengths and give rise to various series of emission or absorption lines. Electron transitions of this kind are called bound–bound transitions (the electron remains bound to the atom before and after the transition) and the resulting radiation is sometimes referred to as bound–bound radiation.

See also: absorption spectrum, atom, continuous spectrum, cosmic rays, electromagnetic radiation, electron, emission spectrum, ion, ionization, photon, plasma, quantum mechanics, radioactivity, solar wind, subatomic particles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radiative Transfer

Astronomy is an observational science and astronomers rely mostly on the electromagnetic radiation emitted by cosmic sources. Indeed over 95% of the information comes from observed radiation, with the remaining 5% coming from COSMIC RAYS, METEORITES, NEUTRINOS and possibly GRAVITATIONAL WAVES. Before reaching an observer, the radiation has interacted with matter in the source and in the intervening space between the source and observer. Understanding the interaction with and transport of radiation through matter is crucial for astronomers to interpret the observations. Radiative transfer deals with the propagation and interaction of radiation with matter.

When radiation passes through a region of space containing gas particles and dust grains, energy is generally removed from the beam either by absorption or by scattering. In addition, the material in a medium may emit radiation, with the source of energy coming from the internal energy of excitation or indirectly from the kinetic energy of motion. In scattering, the photon energy is only slightly altered (energy transferred to the internal energy of the scatterer) and is not being converted into the kinetic or thermal energy. In general scattering does not couple the radiation field to the thermodynamic properties of the medium. Examples of scattering are: Thomson or Compton scattering by electrons, Raman or Rayleigh scattering by molecules and coherent scattering by dust grains. On the other hand, in absorption the photon energy is being converted into kinetic energy or thermal energy. Radiation is coupled to the thermodynamic properties of the medium. Examples of absorption are photoionization of gas particles, bremsstrahlung of free electrons and absorption of radiation by dust grains (see also PHYSICS OF MOLECULES, SOLAR SPECTROSCOPY: ATOMIC PROCESSES).

The theory of radiation transport has many applications. Some important applications in astrophysics include: (a) formation of spectral lines and continuum radiation in STELLAR ATMOSPHERES, (b) excitation mechanisms of interstellar molecular lines and cosmic masers, (c) interpretation of dust spectral features in the energy spectrum of circumstellar envelopes and molecular clouds, (d) modeling PLANETARY ATMOSPHERES, and (e) ionization and thermal structures of HII REGIONS. Some applications in physics and engineering include: (a) neutron transport in nuclear reactors, (b) radiation hydrodynamics in nuclear explosions, and (c) GREENHOUSE EFFECTS in the Earth's atmosphere.

Basic concepts and definitions

To understand the theory of radiation transport, we first introduce a few basic concepts and definitions.

Description of radiation field

Microscopically a radiation field is simply an ensemble of photons. A photon may be described as a point particle of zero mass which moves with constant speed c in free space. It has spin 1 and obeys Bose–Einstein statistics. At any time t, we may represent a photon as a point in phase space, characterized by six phase-space variables: (a) three spatial coordinates (components of position vector $\boldsymbol{r}$) which specifies the location in some reference frame, (b) two angular variables (direction cosines of direction unit vector $\widehat{\Omega}$) which denotes the direction of propagation, and (c) a frequency ν such that the photon has energy $h\nu$ and momentum $h\nu/c$, where h is the Planck constant. Hence a radiation field is a function of $\boldsymbol{r}$, t, $\widehat{\Omega}$ and ν. We often specify $\widehat{\Omega}$ by a polar angle θ and an azimuthal angle ϕ. The components of $\widehat{\Omega}$ in Cartesian coordinates are known as *direction cosines*: $\widehat{\Omega} = \Omega_x\widehat{\boldsymbol{x}} + \Omega_y\widehat{\boldsymbol{y}} + \Omega_z\widehat{\boldsymbol{z}}$, where $\Omega_x = \sin\theta\cos\phi$, $\Omega_y = \sin\theta\sin\phi$, $\Omega_z = \cos\theta$, and $|\widehat{\Omega}| = \sqrt{\Omega_x^2 + \Omega_y^2 + \Omega_z^2} = 1$. Furthermore, $\widehat{\Omega}$ is related to the solid angle element $\mathrm{d}\Omega$ defined by

$$\mathrm{d}\Omega \equiv \frac{\mathrm{d}A}{r^2} = \frac{(r\sin\theta\,\mathrm{d}\phi)(r\,\mathrm{d}\theta)}{r^2} = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi. \qquad (1)$$

When integrated over all directions, the total solid angle, measured in steradians (sr), is 4π.

Photon distribution function

To describe the radiation field as a statistical ensemble of photons, we define a photon distribution function $\Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ such that

$$\mathrm{d}n = \Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\,\mathrm{d}V\,\mathrm{d}\Omega\,\mathrm{d}\nu \qquad (2)$$

is the number of photons at time t with frequency ν in a frequency interval $\mathrm{d}\nu$ located within a differential volume $\mathrm{d}V$ near a point $\boldsymbol{r}$ and traveling with speed c in a direction $\widehat{\Omega}$ in a solid angle element $\mathrm{d}\Omega$. In other words, Ψ is the number of photons at time t per unit phase-space volume (not spatial volume) centered at position $(\boldsymbol{r}, \widehat{\Omega}, \nu)$ in phase space. In general $\boldsymbol{r}$ and $\widehat{\Omega}$ are not specified by the same coordinate system. A radiation field is said to be homogeneous, stationary, isotropic or gray according to whether Ψ is independent of $\boldsymbol{r}$, t, $\widehat{\Omega}$ or ν. Blackbody radiation is an example of a homogeneous, stationary and isotropic radiation field.

In the KINETIC THEORY OF GASES, the particle distribution function f is usually defined such that $f(\boldsymbol{r}, t; \boldsymbol{p})\,\mathrm{d}\boldsymbol{p}$ is the number of particles per unit spatial volume at $(\boldsymbol{r}, t)$ with momenta in the range $(\boldsymbol{p}, \boldsymbol{p} + \mathrm{d}\boldsymbol{p})$. By noting that $\boldsymbol{p} = (h\nu/c)\widehat{\Omega}$ for photons and that the number of particles per unit spatial volume is the same irrespective of what distribution function we use, one can show that $\Psi = (h^3\nu^2/c^3)f$.

Specific intensity

Consider a beam of photons with frequency ν at a position $\boldsymbol{r}$ traveling in direction $\widehat{\Omega}$ (see figure 1). During a time interval $\mathrm{d}t$, the number of photons $\mathrm{d}n$ crossing an element of surface area $\mathrm{d}A$ which is oriented at an angle θ with $\widehat{\Omega}$ is simply $\mathrm{d}n = \Psi\,\mathrm{d}V\,\mathrm{d}\Omega\,\mathrm{d}\nu$, where $\mathrm{d}V$ is the differential volume swept through by those photons in crossing $\mathrm{d}A$ during $\mathrm{d}t$. Hence,

$$\mathrm{d}V = (c\,\mathrm{d}t)(\widehat{\Omega}\cdot\mathrm{d}\boldsymbol{A}) = (c\,\mathrm{d}t)(\cos\theta\,\mathrm{d}A) \qquad (3)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

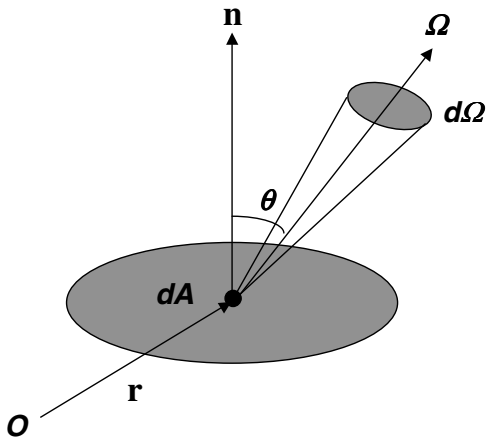


Figure 1.

$$\mathrm{d}n = \Psi(c\,\mathrm{d}t)(\cos\theta\,\mathrm{d}A)\,\mathrm{d}\Omega\,\mathrm{d}\nu. \tag{4}$$

The amount of energy transported across $\mathrm{d}A$ is then

$$\mathrm{d}E = (h\nu)\,\mathrm{d}n = (ch\nu)\Psi\cos\theta\,\mathrm{d}A\,\mathrm{d}\Omega\,\mathrm{d}\nu\,\mathrm{d}t. \tag{5}$$

The specific intensity $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ of radiation at position $\boldsymbol{r}$, traveling in direction $\widehat{\Omega}$, with frequency ν, at time t is defined as the amount of energy, per unit frequency interval, passing through a unit area oriented normal to the beam, into a unit solid angle, in a unit time, i.e.

$$I(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \equiv \lim_{\mathrm{d}A,\,\mathrm{d}\Omega,\,\mathrm{d}\nu,\,\mathrm{d}t\to 0}\left(\frac{\mathrm{d}E}{\cos\theta\,\mathrm{d}A\,\mathrm{d}\Omega\,\mathrm{d}\nu\,\mathrm{d}t}\right). \tag{6}$$

The specific intensity is a scalar quantity with dimensions erg cm^{-2} sr^{-1} Hz^{-1} s^{-1}, and is spatially invariant in free space, i.e. independent of the distance between the source and the observer if there are no sources or sinks of radiation along the line of sight. Observationally, the quantity I can be measured approximately if a source is spatially resolved. Provided POLARIZATION effects can be ignored, the specific intensity completely describes the radiation field. From the above, we see that I and Ψ are related by $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = ch\nu\Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$.

Characterization of radiation by either $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ or $\Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ ignores the wave nature of radiation. Such a description is valid only if the uncertainty principle is satisfied. This condition implies that $\mathrm{d}A\,\mathrm{d}\Omega \geq \lambda^2$ and $\mathrm{d}\nu\,\mathrm{d}t \geq 1$.

Energy density and mean intensity

The energy density $E_R(\boldsymbol{r}, t; \nu)$ of radiation at frequency ν and at time t is defined as the amount of energy at position $\boldsymbol{r}$ per unit frequency interval $\mathrm{d}\nu$ per unit volume $\mathrm{d}V$

$$E_R(\boldsymbol{r}, \nu, t) = \oint \frac{\mathrm{d}E}{\mathrm{d}\nu\,\mathrm{d}V} = h\nu \oint \Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\,\mathrm{d}\Omega. \tag{7}$$

This is physically consistent since Ψ is the number of photons, per unit volume, of energy $h\nu$, propagating in direction $\widehat{\Omega}$ into intervals $\mathrm{d}\nu\,\mathrm{d}\widehat{\Omega}$, so that the energy density is just this number, multiplied by the energy per photon and summed over all solid angles. E_R is a scalar quantity and has dimensions erg cm^{-3} Hz^{-1}.

The mean intensity $J(\boldsymbol{r}, t; \nu)$ is defined as the average of the specific intensity over all solid angles, i.e.

$$J(\boldsymbol{r}, t; \nu) = \frac{1}{4\pi}\oint I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\,\mathrm{d}\Omega. \tag{8}$$

The mean intensity is the *zeroth angular moment* of the specific intensity. It is also a scalar quantity and has the same dimensions as specific intensity, erg cm^{-2} Hz^{-1} s^{-1} sr^{-1}. From the above, we see that

$$E_R(\boldsymbol{r}, t; \nu) = \frac{4\pi}{c}J(\boldsymbol{r}, t; \nu) \tag{9}$$

so that the mean intensity is basically a measure of the energy density.

Radiative flux

The radiative flux $\boldsymbol{F}(\boldsymbol{r}, t; \nu)$ at position $\boldsymbol{r}$, frequency ν, and time t is a vector defined as the net rate of energy flowing across a fixed surface of unit area in all directions per unit frequency interval (flow $\rightarrow$ direction)

$$\boldsymbol{F}(\boldsymbol{r}, t; \nu) = h\nu \oint \Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)c\widehat{\Omega}\,\mathrm{d}\Omega \tag{10}$$

where $(c\widehat{\Omega})$ is just the velocity vector indicating the direction of photon propagation. For an arbitrarily oriented surface $\mathrm{d}\boldsymbol{A} = \widehat{\boldsymbol{n}}\,\mathrm{d}A$, $\boldsymbol{F}\cdot\mathrm{d}\boldsymbol{A}$ gives the net rate of energy flow across the surface $\mathrm{d}\boldsymbol{A}$ per unit frequency interval. The component of $\boldsymbol{F}$ in the direction $\widehat{\boldsymbol{n}}$ is just

$$\begin{aligned} F(\widehat{\boldsymbol{n}}) = \widehat{\boldsymbol{n}}\cdot\boldsymbol{F} &= \oint \frac{1}{\mathrm{d}A\,\mathrm{d}\nu}\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) \\ &= ch\nu \oint \Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\cos\theta\,\mathrm{d}\Omega \end{aligned} \tag{11}$$

where θ is the angle between $\widehat{\Omega}$ and the normal to $\mathrm{d}A$ ($\cos\theta = \widehat{\boldsymbol{n}}\cdot\widehat{\Omega}$). The net flux $F(\widehat{\Omega}')$ in the given direction $\widehat{\Omega}'$ can be written as (see figure 2)

$$F(\widehat{\Omega}') = F^+ - F^- \tag{12}$$

where $\left\{\begin{matrix} F^+ \text{ (outward net flux)} \\ F^- \text{ (inward net flux)} \end{matrix}\right\}$ represents the part of $F(\widehat{\Omega}')$ due to the radiation from the direction $\widehat{\Omega}$ such that $\widehat{\Omega}'\cdot\widehat{\Omega}\left\{\begin{matrix} > 0 \\ < 0 \end{matrix}\right\}$. Note that the amount of radiant energy per unit volume at a given point is characterized by energy density E_R whereas the direction and velocity of its flow at the point are given by the radiative flux $\boldsymbol{F}$. The radiative flux has dimensions erg cm^{-2} Hz^{-1} s^{-1}. It is the *first angular moment* of the radiation field. For an isotropic radiation field $\boldsymbol{F} = 0$. Observationally, if a source is spatially unresolved, we can measure only the flux, and the energy received falls off as the inverse square of the distance between the source and the observer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

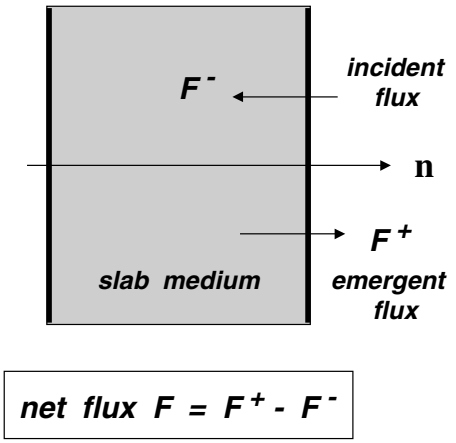


Figure 2.

Radiation pressure tensor

In the kinetic theory of gases, pressure at a given point is defined as the net rate of momentum transfer normal to an arbitrarily oriented surface of unit area at that point. The same definition applies to radiation field which is essentially a photon gas. In general, the radiation pressure tensor is defined as

$$\mathsf{P}(\boldsymbol{r}, t; \nu) = h\nu \oint \Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\widehat{\Omega}\widehat{\Omega}\, \mathrm{d}\Omega. \qquad (13)$$

Physically we may interpret the $i - j$ component of the radiation pressure tensor as the rate of momentum flow in the j-direction from radiation of frequency ν through a surface of unit area oriented perpendicular to the i-direction

$$P_{ij}(\boldsymbol{r}, t; \nu) = \oint (c\Omega_i)(h\nu\Omega_j/c)\Psi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\, \mathrm{d}\Omega \qquad (14)$$

where $(c\Omega_i)$ is the projected velocity in the i-direction and $(h\nu\Omega_j/c)$ is the momentum in the j-direction. $\mathsf{P}(\boldsymbol{r}, t; \nu)$ is a tensor of second rank and has nine components. It is the *second angular moment* of the radiation field. It has the same dimensions as energy density, erg cm^{-3} Hz^{-1}. It is symmetric, i.e. $P_{ij} = P_{ji}$, and obeys the trace relationship $\sum_i P_{ii} = P_{xx} + P_{yy} + P_{zz} = E_R$. For an isotropic radiation field, P is diagonal and isotropic

$$P_{ij} = \left\{ \begin{array}{ll} 0 & i \neq j \text{ (all off-diagonal terms are zero)} \\ \frac{1}{3}E_R & i = j \text{ (all diagonal terms are equal)} \end{array} \right\}. \qquad (15)$$

Note that for an isotropic radiation field $\mathsf{P} \neq 0$ while $\boldsymbol{F} = 0$.

Simplifications in 1D geometries

In most astrophysical problems, the following one-dimensional (1D) geometries are assumed:

(a) Planar or slab geometry: if one employs Cartesian (x, y, z) coordinates to specify $\boldsymbol{r}$ and introduces polar and azimuthal angles (θ, ϕ) to denote $\widehat{\Omega}$, 1D planar or slab geometry implies azimuthal independence and infinite extent in x and y, i.e. $\boldsymbol{r}$ is independent of x and y while $\widehat{\Omega}$ is independent of ϕ, so that $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ becomes $I(z, t; \mu, \nu)$, where $\mu = \cos\theta = \widehat{\Omega} \cdot \widehat{z}$.

(b) Spherical geometry: in spherical coordinates, the position vector $\boldsymbol{r}$ is specified by (r, Θ, Φ) while the direction $\widehat{\Omega}$ is specified in terms of the polar and azimuthal angles (θ, ϕ), now measured with respect to a unit vector $\widehat{r}$ in the radial direction. 1D spherical geometry implies that $\boldsymbol{r}$ depends only on r and is independent of Θ and Φ, while $\widehat{\Omega}$ is independent of ϕ so that $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ becomes $I(r, t; \mu, \nu)$, where $\mu = \cos\theta = \widehat{\Omega} \cdot \widehat{r}$.

In both geometries, $\widehat{\Omega} \rightarrow \mu$ so that $\widehat{\Omega}\widehat{\Omega} \rightarrow \mu^2$ and

$$\oint \ldots \mathrm{d}\Omega = \int_0^{2\pi} \ldots \mathrm{d}\phi \int_0^{\pi} \sin\theta\, \mathrm{d}\theta = 2\pi \int_{-1}^{+1} \ldots \mathrm{d}\mu. \qquad (16)$$

With these simplifications, the angular moments of the specific intensity can be written as

$$M^n(r, t; \nu) = \frac{1}{2} \int_{-1}^{+1} I(r, t; \mu, \nu)\mu^n\, \mathrm{d}\mu \qquad (17)$$

so that we have mean intensity $J = M^0$, *Eddington flux* $H = M^1$, K-integral $K = M^2$.

For 1D geometries, the angular distribution of a radiation field may be characterized by the scalar ratio $f \equiv K/J$, known as the *Eddington factor* or *anisotropy factor*.

$f \equiv K/J$	Radiation field anisotropy
$0 < f < 1/3$	Sideward peaking
$f = 1/3$	Isotropic or linearly anisotropic
$1/3 < f < 1$	Forward peaking

The assumption of $f = 1/3$ is known as the *Eddington approximation* which is valid for stellar interiors. In the outer layers of stellar atmospheres and for centrally heated sources, the radiation tends to stream radially and $f \rightarrow 1$. On the other hand, for externally illuminated sources (interstellar clouds embedded in the ambient interstellar radiation field) in which radiation tends to peak sideways and non-radially, $f < 1/3$.

Observational implications

The energy received from an unresolved astronomical source by a distant observer depends on the geometry of the source. The energy received, per unit area normal to the line of sight, from a differential area $\mathrm{d}A$ on the surface of the source with its normal $\hat{\boldsymbol{n}}$ making an angle θ with the observer's line of sight is

$$\mathrm{d}F = I\, \mathrm{d}\Omega = I(\hat{\boldsymbol{n}} \cdot \mathrm{d}\boldsymbol{A})D^{-2} = I\, \mathrm{d}A \cos\theta D^{-2} \qquad (18)$$

where $\mathrm{d}\Omega$ is the solid angle subtended by the surface element, D is the distance of the source from the observer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and I is the specific intensity emergent at the surface of the source.

Consider two sources of the same surface brightness I_0: one is a sphere of radius R and the other is a disk of radius R and half-thickness Z. In both cases $D \gg R$ or Z so that all rays from the source to the observer may be considered to be parallel. Below we show that if $R = Z$, $F_{\rm disk} \geq F_{\rm sphere}$ for all viewing angles.

For a spherically symmetric source of radius R, the surface element is just a differential annulus on the disk (projected image of sphere), i.e. $\hat{\boldsymbol{n}}\cdot\mathrm{d}\boldsymbol{A} = 2\pi r\,\mathrm{d}r = 2\pi(R\sin\theta)(R\cos\theta\,\mathrm{d}\theta)$ so that the flux received by a distant observer ($D \gg R$) from the visible face of the sphere is

$$F_{\rm sphere} = \int I(\hat{\boldsymbol{n}}\cdot\mathrm{d}\boldsymbol{A})D^{-2} = \left(\frac{R}{D}\right)^2 I\cos\theta\sin\theta\,\mathrm{d}\theta. \quad (19)$$

The upper limit of integration for θ is $\pi/2 = 90^\circ$ because we assume that $R \ll D$ so that all rays from the source to the observer may be considered to be parallel. For a sphere of uniform brightness I_0, this reduces to $F_{\rm sphere} = (\pi R^2/D^2)I_0$ which is independent of θ. Note that $\Delta\Omega \equiv (\pi R^2/D^2)$ is just the solid angle subtended by the source at the observer, so that $F_{\rm sphere} = I_0\Delta\Omega$.

For a disk of radius R, half-thickness Z, and with its symmetry axis making an angle θ with the line of sight, we can view the element area $\mathrm{d}A$ as having two components: $(\hat{\boldsymbol{n}}\cdot\mathrm{d}\boldsymbol{A})_{\rm face} = \cos\theta(r\,\mathrm{d}r\,\mathrm{d}\phi)$ for the planar surface of the disk and $(\hat{\boldsymbol{n}}\cdot\mathrm{d}\boldsymbol{A})_{\rm edge} = \sin\theta(\cos\phi R\,\mathrm{d}\phi\,\mathrm{d}z)$ for the curved surface of the disk edge. Hence the total flux received by a distant observer ($D \gg Z$ or R) is

$$F_{\rm disk} = \left(\cos\theta\int_{\phi=0}^{2\pi}\int_{r=0}^{R} I_{\rm face} r\,\mathrm{d}r\,\mathrm{d}\phi + R\sin\theta\int_{z=-Z}^{+Z}\int_{\phi=-\pi/2}^{\pi/2} I_{\rm edge}\cos\phi\,\mathrm{d}\phi\,\mathrm{d}z\right)D^{-2}. \quad (20)$$

For a disk of uniform surface brightness I_0, this reduces to $F_{\rm disk} = I_0\Delta\Omega$, where again $\Delta\Omega \equiv (\pi R^2\cos\theta + 4RZ\sin\theta)/D^2$ is the solid angle subtended by the total projected area of the disk. Compared to a sphere with the same I_0 and R, and if $R = Z$, the ratio of observed fluxes for the two geometries is just given by the ratio of their projected areas

$$\frac{F_{\rm disk}}{F_{\rm sphere}} = \frac{\pi R^2\cos\theta + 4R^2\sin\theta}{\pi R^2} \Rightarrow 1 \leq \frac{F_{\rm disk}}{F_{\rm sphere}} \leq \frac{4}{\pi} \quad (21)$$

which indicates that $F_{\rm disk} \geq F_{\rm sphere}$ for all viewing angles.

Equation of radiation transport

To derive the equation of radiation transport, we introduce a few terms to quantify the physical processes of absorption, scattering and emission.

Basic definitions

When radiation passes through material, energy is generally removed from the beam either by pure absorption or by scattering. The sum of scattering and absorption is called extinction. In addition, the material in a medium may emit radiation.

Extinction coefficient The *volume extinction coefficient* χ is defined such that an element of material of cross section $\mathrm{d}A$ and length $\mathrm{d}s$, oriented normal to a beam of radiation having specific intensity $I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ propagating along $\widehat{\Omega}$ into solid angle $\mathrm{d}\Omega$ within a frequency interval $\mathrm{d}\nu$, removes from the beam in a time $\mathrm{d}t$ an amount of energy

$$\mathrm{d}E_- = \chi(\boldsymbol{r}, t; \widehat{\Omega}, \nu)I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\,\mathrm{d}A\,\mathrm{d}s\,\mathrm{d}\Omega\,\mathrm{d}\nu\,\mathrm{d}t. \quad (22)$$

χ has the dimension of cm^{-1} and is given by the product of an extinction cross section $\sigma(\mathrm{cm}^2)$ and the number density of interacting particles $n(\mathrm{cm}^{-3})$ summed over all states that can interact with photons of frequency ν, i.e. $\chi = n\sigma$. Some authors use the term mass extinction coefficient κ which is related to χ by $\chi = \kappa\rho$, where ρ is the mass density. It can be shown that $1/\chi$ is the photon mean free path, defined as the average distance traveled by a photon before it is absorbed or scattered.

The extinction coefficient χ is made up of two contributions, pure absorption and scattering, with the corresponding coefficients χ^a and χ^s

$$\chi(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = \chi^a(\boldsymbol{r}, t; \widehat{\Omega}, \nu) + \chi^s(\boldsymbol{r}, t; \widehat{\Omega}, \nu). \quad (23)$$

Emission coefficient The *volume emission coefficient* or simply *emissivity* η is defined such that the amount of energy released from a material element of cross section $\mathrm{d}A$ and length $\mathrm{d}s$, into a solid angle $\mathrm{d}\Omega$ around a direction $\widehat{\Omega}$, within a frequency interval $\mathrm{d}\nu$ in a time $\mathrm{d}t$ is

$$\mathrm{d}E_+ = \eta(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\,\mathrm{d}A\,\mathrm{d}s\,\mathrm{d}\Omega\,\mathrm{d}\nu\,\mathrm{d}t. \quad (24)$$

η has units of erg cm^{-3} s^{-1} sr^{-1} Hz^{-1}, the same as specific intensity I. The emissivity may be isotropic in the rest frame of the material, but is anisotropic in the laboratory frame when the material moves.

The emission coefficient η can be represented as a combination of two different coefficients

$$\eta(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = \eta^t(\boldsymbol{r}, t; \widehat{\Omega}, \nu) + \eta^s(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \quad (25)$$

where η^t is the true or thermal emission coefficient and η^s is the scattering emission coefficient.

For a system in thermodynamic equilibrium, χ and η are related by *Kirchhoff's law* $\eta^*(\nu) = \chi^*(\nu)B(\nu, T)$, where $B(\nu, T)$ is the Planck function.

Source function The *source function* $S(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ is defined as the ratio of η and χ

$$S(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \equiv \eta(\boldsymbol{r}, t; \widehat{\Omega}, \nu)/\chi(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \quad (26)$$

so that from Kirchhoff's law $S(\boldsymbol{r}, t; \nu) = B(\nu, T)$ in thermodynamic equilibrium.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Optical depth The optical depth τ is defined as the volume extinction coefficient integrated over a path length

$$\tau \equiv \int \chi \, \mathrm{d}s. \tag{27}$$

Note that τ is a dimensionless quantity. A medium is said to be *optically thin* if $\tau \ll 1$ and *optically thick* if $\tau \gg 1$. In dimensionless units, a path length of one photon mean free path corresponds to $\tau = 1$.

Scattering and redistribution functions During scattering, there are changes in both direction and frequency of the radiation. These changes are described in terms of a *redistribution function* $R(\widehat{\Omega}', \nu'; \widehat{\Omega}, \nu)$ which is the probability of a photon in the frequency interval $(\nu', \nu' + \mathrm{d}\nu')$ in the direction $\widehat{\Omega}'$ within a solid angle $\mathrm{d}\Omega'$ being scattered in the direction $\widehat{\Omega}$ within solid angle $\mathrm{d}\Omega$ in the frequency interval $(\nu, \nu + \mathrm{d}\nu)$. More simply, it is the joint probability that a photon with direction and frequency $(\widehat{\Omega}', \nu')$ is absorbed and a photon with $(\widehat{\Omega}, \nu)$ is emitted. It satisfies the normalization condition

$$\left(\frac{1}{4\pi}\right)^2 \oint_{\Omega'} \mathrm{d}\Omega' \oint_{\Omega} \mathrm{d}\Omega \int_{\nu'} \mathrm{d}\nu' \int_{\nu} \mathrm{d}\nu R(\widehat{\Omega}', \nu'; \widehat{\Omega}, \nu) = 1. \tag{28}$$

If $\sigma_{\mathrm{sca}}(\boldsymbol{r}, t; \widehat{\Omega}, \nu)$ is the scattering cross section, and n is the number denisty of scatters, then the scattering emission coefficient η^s is given by

$$\eta^s(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = n\sigma_{\mathrm{sca}}(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \oint_{\Omega'} \frac{\mathrm{d}\Omega'}{4\pi} \int_{\nu'} \mathrm{d}\nu' \times [I(\boldsymbol{r}, t; \widehat{\Omega}', \nu') R(\widehat{\Omega}', \nu'; \widehat{\Omega}, \nu)]. \tag{29}$$

Similarly, the scattering coefficient χ^s is given by

$$\chi^s(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = n\sigma_{\mathrm{sca}}(\boldsymbol{r}, t; \widehat{\Omega}, \nu)\varphi(\nu) \tag{30}$$

where we have defined a scattering profile $\varphi(\nu)$ as

$$\varphi(\nu) \equiv \oint_{\Omega'} \frac{\mathrm{d}\Omega'}{4\pi} \oint_{\Omega} \frac{\mathrm{d}\Omega}{4\pi} \int_{\nu'} \mathrm{d}\nu' R(\widehat{\Omega}', \nu'; \widehat{\Omega}, \nu). \tag{31}$$

For coherent scattering, $\nu = \nu'$, the redistribution function can be expressed as

$$R(\widehat{\Omega}', \nu'; \widehat{\Omega}, \nu) = p(\widehat{\Omega}', \widehat{\Omega})\varphi(\nu')\delta(\nu - \nu') \tag{32}$$

where $p(\widehat{\Omega}', \widehat{\Omega})$ is known as the *phase function* which gives the probability of scattering in the direction $\widehat{\Omega}$ of radiation arriving from the direction $\widehat{\Omega}'$ on a volume element at point $\boldsymbol{r}$. It is normalized as

$$\frac{1}{4\pi} \oint p(\widehat{\Omega}, \widehat{\Omega}') \, \mathrm{d}\widehat{\Omega}' = 1. \tag{33}$$

For coherent scattering (e.g. scattering by dust grains), the scattering emission coefficient becomes

$$\eta^s(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = n\sigma_{\mathrm{sca}}(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \oint_{\Omega'} \frac{\mathrm{d}\Omega'}{4\pi} \times [I(\boldsymbol{r}, t; \widehat{\Omega}', \nu) p(\widehat{\Omega}', \widehat{\Omega})]. \tag{34}$$

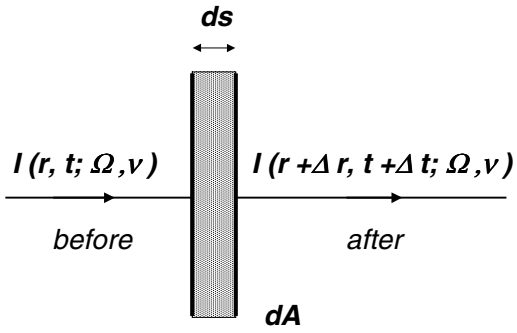


Figure 3.

Derivation of the transport equation
Consider an element of material of length $\mathrm{d}s$ and cross section $\mathrm{d}A$, fixed in the laboratory frame. We calculate the change, in a time $\mathrm{d}t$, in the energy of the radiation field contained in a frequency interval $\mathrm{d}\nu$, traveling into solid angle $\mathrm{d}\Omega$ along a direction $\widehat{\Omega}$ normal to $\mathrm{d}A$, as it passes through the material. The difference between the amount of energy that emerges at position $\boldsymbol{r} + \Delta\boldsymbol{r}$ at time $t + \mathrm{d}t$ and the amount incident at position $\boldsymbol{r}$ at time t must equal the difference between the amount of energy created by emission from the material and the amount removed by extinction (see figure 3). The equation of radiation transport is just a statement of the energy conservation along a ray path $\mathrm{d}s$

$$\underbrace{\Delta E}_{\text{energy change}} = \underbrace{\mathrm{d}E_+}_{\text{energy emitted}} - \underbrace{\mathrm{d}E_-}_{\text{energy removed}}. \tag{35}$$

From the definitions of I, η, and χ, this means

$$[I(\boldsymbol{r} + \Delta\boldsymbol{r}, t + \Delta t; \widehat{\Omega}, \nu) - I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)] \, \mathrm{d}A \, \mathrm{d}\Omega \, \mathrm{d}\nu \, \mathrm{d}t = [\eta(\boldsymbol{r}, t; \widehat{\Omega}, \nu) - \chi(\boldsymbol{r}, t; \widehat{\Omega}, \nu) I(\boldsymbol{r}, t; \widehat{\Omega}, \nu)] \, \mathrm{d}A \, \mathrm{d}s \, \mathrm{d}\Omega \, \mathrm{d}\nu \, \mathrm{d}t \tag{36}$$

where we have $\Delta t = \mathrm{d}s/c$. Applying Taylor's expansion in $(\boldsymbol{r}, t)$ to $I(\boldsymbol{r} + \Delta\boldsymbol{r}, t + \Delta t; \widehat{\Omega}, \nu)$

$$I(\boldsymbol{r} + \Delta\boldsymbol{r}, t + \Delta t; \widehat{\Omega}, \nu) \approx I(\boldsymbol{r}, t; \widehat{\Omega}, \nu) + \left(\frac{1}{c}\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) I(\boldsymbol{r}, t; \widehat{\Omega}, \nu) \, \mathrm{d}s + \text{higher-order terms}.$$

Neglecting the higher-order terms and substituting into the energy conservation statement, we obtain the time-dependent equation of radiation transport

$$\left(\frac{1}{c}\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) I(\boldsymbol{r}, t; \widehat{\Omega}, \nu) = \eta(\boldsymbol{r}, t; \widehat{\Omega}, \nu) - \chi(\boldsymbol{r}, t; \widehat{\Omega}, \nu) I(\boldsymbol{r}, t; \widehat{\Omega}, \nu). \tag{37}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Suppressing the functional dependence of the quantities involved, we may write

$$\underbrace{\frac{1}{c}\frac{\partial I}{\partial t}}_{\text{temporal change}} + \underbrace{\frac{\partial I}{\partial s}}_{\text{spatial change}} = \eta - \chi I. \tag{38}$$

This derivation is based entirely on a classical, macroscopic and phenomenological approach. It neglects the wave nature of radiation such as polarization, dispersion, coherence, interference and other quantum effects. For situations in which

$$\frac{\partial I}{\partial s} \gg \frac{1}{c}\frac{\partial I}{\partial t} \approx 0 \tag{39}$$

we have the time-independent equation of radiation transport

$$\underbrace{\frac{\partial I}{\partial s}}_{\text{geometry}} = \underbrace{\eta - \chi I}_{\text{physics}}. \tag{40}$$

The term on the left depends on the *geometry* of the medium while the term on the right represents the *physics* of how radiation interacts with matter. Since s is a coordinate-independent path length, the above equation is valid in arbitrary coordinate systems, provided we use an appropriate expression to evaluate $\partial/\partial s$.

Expressions for $\partial/\partial s$ in different geometries

The directional derivative $\partial/\partial s$ basically describes how $\widehat{\Omega}$ changes with $\boldsymbol{r}$. In general $\boldsymbol{r}$ and $\widehat{\Omega}$ are not specified by the same coordinate system. If both $\widehat{\Omega}$ and $\boldsymbol{r}$ are specified in Cartesian coordinates ($\boldsymbol{r} = x\widehat{\boldsymbol{x}} + y\widehat{\boldsymbol{y}} + z\widehat{\boldsymbol{z}}$, $\widehat{\Omega} = \Omega_x\widehat{\boldsymbol{x}} + \Omega_y\widehat{\boldsymbol{y}} + \Omega_z\widehat{\boldsymbol{z}}$), then $\partial/\partial s$ can be expressed in vector notation as

$$\begin{aligned}\frac{\partial}{\partial s} &= \left(\frac{\partial x}{\partial s}\right)\left(\frac{\partial}{\partial x}\right) + \left(\frac{\partial y}{\partial s}\right)\left(\frac{\partial}{\partial y}\right) + \left(\frac{\partial z}{\partial s}\right)\left(\frac{\partial}{\partial z}\right) \\ &= \Omega_x\frac{\partial}{\partial x} + \Omega_y\frac{\partial}{\partial y} + \Omega_z\frac{\partial}{\partial z} = \widehat{\Omega}\cdot\nabla.\end{aligned} \tag{41}$$

In curvilinear coordinates, the coordinate basis vectors rotate with respect to the direction determined by $\widehat{\Omega}$ and we must evaluate $\partial/\partial s$ as

$$\partial/\partial s = \widehat{\Omega}\cdot\nabla + (\mathrm{d}\widehat{\Omega}/\mathrm{d}s)\cdot\nabla_\Omega \tag{42}$$

where ∇_Ω denotes differentiation with respect to the direction cosines of $\widehat{\Omega}$. Specifically, for arbitrary curvilinear coordinates (x_1, x_2, x_3) with unit vectors $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$, the directional derivative is

$$\begin{aligned}\frac{\partial}{\partial s} = {} & \frac{1}{h_1}\sin\theta\cos\phi\frac{\partial}{\partial x_1} + \frac{1}{h_2}\sin\theta\sin\phi\frac{\partial}{\partial x_2} + \frac{1}{h_3}\cos\theta\frac{\partial}{\partial x_3} \\ & - \frac{1}{\sin\theta}\left\{\widehat{\Omega}\left[\frac{1}{\sin\theta}\nabla(\widehat{\Omega}\boldsymbol{i}) - \frac{\cos\theta\cos\phi}{\sin^2\theta}\nabla(\widehat{\Omega}\boldsymbol{k})\right]\right\}\frac{\partial}{\partial\phi} \\ & - \frac{1}{\sin\theta}\{\widehat{\Omega}[\nabla(\widehat{\Omega}\mathbf{k})]\}\frac{\partial}{\partial\theta}\end{aligned} \tag{43}$$

where

$$h_i = \left[\left(\frac{\partial x}{\partial x_i}\right)^2 + \left(\frac{\partial y}{\partial x_i}\right)^2 + \left(\frac{\partial z}{\partial x_i}\right)^2\right]^{1/2}, \quad i = 1, 2, 3 \tag{44}$$

are the Lamé coefficients. Below we summarize the expressions for $\partial/\partial s$ in various 3D geometries

Geometry	$\boldsymbol{r}, \widehat{\Omega}$	Directional derivative ($\partial/\partial s = \widehat{\Omega}\cdot\nabla$)
Slab	$(x, y, z), (\theta, \phi)$	$\sin\theta\cos\phi\frac{\partial}{\partial x} + \sin\theta\sin\phi\frac{\partial}{\partial y} + \cos\theta\frac{\partial}{\partial z}$
Sphere	$(r, \Theta, \Phi), (\theta, \phi)$	$\cos\theta\frac{\partial}{\partial r} + \frac{\sin\theta\cos\phi}{r}\frac{\partial}{\partial\Theta} - \frac{\sin\theta\sin\phi}{r\sin\Theta}\frac{\partial}{\partial\Phi} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta} - \frac{\sin\theta\sin\phi}{r\tan\Theta}\frac{\partial}{\partial\phi}$
Cylinder	$(r, z, \Phi), (\theta, \phi)$	$\sin\theta\cos\phi\frac{\partial}{\partial r} + \frac{\sin\theta\sin\phi}{r}\frac{\partial}{\partial\Phi} + \cos\theta\frac{\partial}{\partial z} - \frac{\sin\theta\sin\phi}{r}\frac{\partial}{\partial\phi}$

For 1D geometries, the above expressions simplify to ($\mu \equiv \cos\theta$)

1D geometry	$(\boldsymbol{r};\widehat{\Omega})$	Directional derivative $(\partial/\partial s)$
Slab	$(z;\theta)$	$\cos\theta\frac{\partial}{\partial z} = \mu\frac{\mathrm{d}}{\mathrm{d}z}$
Sphere	$(r;\theta)$	$\cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta} = \mu\frac{\partial}{\partial r} + \frac{1-\mu^2}{r}\frac{\partial}{\partial\mu}$
Cylinder	$(r;\theta,\phi)$	$\sin\theta\left[\cos\phi\frac{\partial}{\partial r} - \frac{\sin\phi}{r}\frac{\partial}{\partial\phi}\right]$

For 2D disk geometry with axial symmetry, we have

$$\frac{\partial}{\partial s} = \cos\theta\frac{\partial}{\partial z} + \sin\theta\left(\cos\phi\frac{\partial}{\partial r} - \frac{\sin\phi}{r}\frac{\partial}{\partial\phi}\right). \tag{45}$$

Moments of the transport equation

For simplicity we suppress the functional dependence of the quantities involved, except for the frequency dependence which we denote by subscript ν. Starting with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the time-dependent transport equation and rewriting $\partial/\partial s$ as $\nabla\cdot\widehat{\Omega}$, we integrate over all directions and frequencies to get

$$\int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega\left(\frac{1}{c}\frac{\partial I_\nu}{\partial t}+(\nabla\cdot\widehat{\Omega})I_\nu\right) = \int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega(\eta_\nu - I_\nu\chi_\nu). \tag{46}$$

Re-arranging the terms yields

$$\frac{\partial E_R}{\partial t} = \int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega(\eta_\nu - I_\nu\chi_\nu) - \nabla\cdot\boldsymbol{F}. \tag{47}$$

This is the *zeroth moment equation* and is a statement of energy conservation for the radiation field. Integrating the equation over a fixed volume element and applying the divergence theorem, we see that the rate of change of radiative energy in the volume equals the emission from the material, minus the absorption by the material, and minus the net flow of radiant energy (flux) through the volume element's boundary surface.

In steady state when the energy density does not depend on time, and if there is no internal energy source so that the radiative energy absorbed in any volume element is balanced locally by the energy emitted, we have the condition of *radiative equilibrium*

$$\int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega(\eta_\nu - I_\nu\chi_\nu) = 0 \tag{48}$$

which leads to the vanishing of flux divergence, $\nabla\cdot\boldsymbol{F} = 0$. Thus radiative equilibrium implies that the integrated net flux is constant with depth in 1D planar geometry while the luminosity is constant with radius in 1D spherical geometry.

The *first moment equation* is obtained by multiplying the transport equation by $\widehat{\Omega}$ and then integrating over all directions and frequencies

$$\int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega\left(\frac{1}{c}\frac{\partial I_\nu}{\partial t}\widehat{\Omega} + (\nabla\cdot\widehat{\Omega})I_\nu\widehat{\Omega}\right) = \int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega \times [(\eta_\nu - I_\nu\chi_\nu)\widehat{\Omega}] \tag{49}$$

which simplifies to

$$\frac{1}{c^2}\frac{\partial \boldsymbol{F}}{\partial t} = \frac{1}{c}\int_0^\infty \mathrm{d}\nu \oint \mathrm{d}\Omega[(\eta_\nu - I_\nu\chi_\nu)\widehat{\Omega}] - \nabla\cdot\mathsf{P}. \tag{50}$$

This may be viewed as the momentum equation of the radiation field. Now the momentum of a photon with energy $h\nu$ is $(h\nu/c)\widehat{\Omega}$ so that the net rate of radiative momentum transport across a surface $\mathrm{d}\boldsymbol{A}$ is $c^{-1}\mathbf{F}\cdot\mathrm{d}\boldsymbol{A}$. Since the transport by particles occurs at a speed of c, the radiative momentum density vector is just $c^{-2}\boldsymbol{F}$. Integrating the equation over a fixed volume element and applying the divergence theorem, we see that the rate of change of radiative momentum density equals the gain from material emission, minus the loss due to material absorption and the transport across the boundary surface of the volume element.

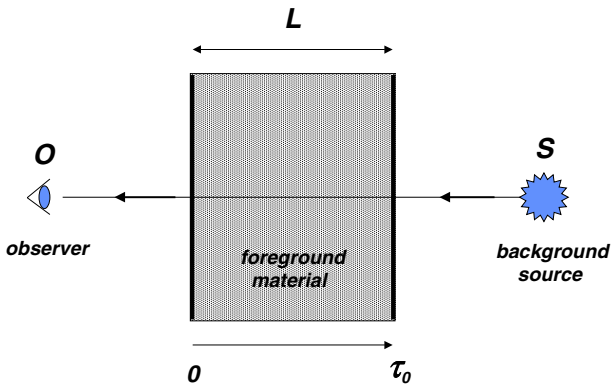


Figure 4.

Formal solution of the transport equation

The solution to the transport equation under the assumption that either χ and η (or equivalently τ and S) of the medium are known is called the *formal solution*. In this case, the transport equation becomes a linear first-order differential equation.

Consider an observer looking through an intervening medium that absorbs and emits radiation. There is a background source behind the medium (see figure 4). In terms of the optical depth τ and source function S, the time-independent transport equation may be written as

$$\frac{\mathrm{d}I}{\mathrm{d}s} = -\chi I + \eta \Rightarrow \frac{\mathrm{d}I}{\mathrm{d}\tau} = I - S \tag{51}$$

where, by convention, we measure τ inward from the observer into the medium

$$\mathrm{d}\tau \equiv -\chi\,\mathrm{d}s \Rightarrow \tau(s) = \int_0^s \chi(s')\,\mathrm{d}s'. \tag{52}$$

Using the integrating factor $\exp(-\tau)$, we obtain the formal solution

$$I^+(o) = I(s)\,\mathrm{e}^{-\tau} + \int_0^\tau S(\tau')\,\mathrm{e}^{-\tau'}\,\mathrm{d}\tau'. \tag{53}$$

The superscript '+' means emergent intensity. Hence the observed intensity is the sum of two contributions: contribution from attenuated background source and contribution from the emitting/absorbing medium of the foreground layer. For a medium with constant properties (both S and τ are uniform), the above result simplifies to

$$I^+(o) = I(s)\,\mathrm{e}^{-\tau} + S(1 - \mathrm{e}^{-\tau}). \tag{54}$$

In terms of χ and η, the formal solution is

$$I^+(o) = I(s)\exp\left(-\int_0^s \chi(s')\,\mathrm{d}s'\right) + \int_0^s \eta(s')\exp\left(-\int_0^{s'} \chi(s'')\,\mathrm{d}s''\right)\mathrm{d}s'. \tag{55}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Consider the following situations:

(a) No background source; only uniform foreground material. The formal solution becomes

$$I^+(o) = S(1 - \mathrm{e}^{-\tau}). \tag{56}$$

If the layer is optically thin, i.e. $\tau \ll 1$, we have

$$I^+(o) \approx S\tau = \left(\frac{\eta}{\chi}\right)(\chi L) = \eta L. \tag{57}$$

This means we can see through the layer and the observed intensity is just the sum of all the emission from different points along the total path length L. On the other hand, if the medium is optically thick, i.e. $\tau \gg 1$, we have

$$I^+(o) \approx S = \eta\left(\frac{1}{\chi}\right). \tag{58}$$

This means we can see up to a distance of $\tau = 1$ in the layer if we regard $\left(\frac{1}{\chi}\right)$ as the distance to $\tau = 1$ or just one *photon mean-free-path*.

(b) Background source only; no emission by foreground material. In this case, the formal solution is

$$I^+(o) = I(s)\,\mathrm{e}^{-\tau} \tag{59}$$

which is just simple attenuation of background intensity by the foreground material.

(c) Background source + uniform foreground material. The formal solution is

$$I^{+'}(o) = I^+(s)\,\mathrm{e}^{-\tau} + S(1 - \mathrm{e}^{-\tau}) \tag{60}$$

Consider the observational technique of '*position switching*' which consists of combining two observations:

on source: $I^+(o) = I(s)\,\mathrm{e}^{-\tau} + S(1 - \mathrm{e}^{-\tau})$
off source: $I^{+'}(o) = I(s)$

so that

$$\Delta I^+(o) = I^+(o) - I^{+'}(o) = [S - I(s)]\,(1 - \mathrm{e}^{-\tau}). \tag{61}$$

This implies that one would see (i) an emission line if $\Delta I^+(o) > 0$ or $S > I(s)$, (ii) an absorption line if $\Delta I^+(o) < 0$ or $S > I(s)$, and (iii) no line or signal if $\Delta I^+(o) = 0$ or $S = I(s)$.

We may view the equation of radiation transport as a description of a relaxation process in which the radiation field asymptotically approaches the source function in the limit of large optical depth, i.e.

$$I^+(o) = I^+(s)\,\mathrm{e}^{-\tau} + S(1 - \mathrm{e}^{-\tau}) = S + \mathrm{e}^{-\tau}[I^+(s) - S] \tag{62}$$

which implies that $\lim_{\tau\to\infty} I^+(o) \to S$.

Input physics

To solve the problem of radiation transport for a medium containing dust grains and gas particles, we need to express the extinction and emission coefficients in terms of microphysical parameters. Below we first consider the interaction of radiation with a gaseous medium.

Gas component

When a gaseous particle (atom, ion or molecule) absorbs radiation, it will undergo a transition from a lower energy level i to a higher energy level j. Subsequently two different recombination transitions may take place: *spontaneous emission* for which the emitted photons are isotropic and *induced or stimulated emission* in which the emitted and incident photons have the same direction. Since the emitted and incident photons for an induced transition have the same direction, the induced emission is often treated as negative absorption. If both levels i and j are bound states, we have bound–bound transitions. If j is a free state (or continuum level), we have bound–free transitions (e.g. ionization). Free–free transitions occur when both i and j are free states. The total volume extinction coefficient (of a gaseous species at a given degree of ionization) due to bound–bound, bound–free and free–free transitions can be written as

$$\begin{aligned}\chi(\nu) = &\sum_{j=2}^{N}\sum_{i=1}^{j-1}\sigma_{ij}(\nu)\left(n_i - \frac{g_i}{g_j}n_j\right)\\ &+ \sum_{i=1}^{N}\sigma_{ik}(\nu)\left[n_i - n_e n_k \phi_i(T)\exp\left(-\frac{h\nu}{kT}\right)\right]\\ &+ n_e n_k \sigma_{kk}(\nu)\left[1 - \exp\left(-\frac{h\nu}{kT}\right)\right]\end{aligned} \tag{63}$$

where the subscript k refers to a continuum or ionized state, n_e is the number density of electrons and $\phi_i(T)$ is the Saha–Boltzmann factor given by

$$\phi_i(T) = \left(\frac{h^2}{2\pi mkT}\right)^{3/2}\frac{U_i}{2U_k}\exp\left(\frac{h}{kT}(\nu_k - \nu_i)\right). \tag{64}$$

The gs and Us are the statistical weights of the energy levels and partition functions of the ionization states. The σs are the relevant cross sections for radiative excitations (either photoabsorption or photoionization). For example, for bound–bound transitions

$$\sigma_{ij}(\nu) = \frac{c^2}{8\pi}\frac{g_j}{g_i}\frac{A_{ji}}{\nu_{ji}^2}\phi(\nu) \tag{65}$$

where A_{ji} is Einstein's coefficient for spontaneous emission and $\phi(\nu)$ is the line profile function. The interaction cross sections for bound–free and free–free are normally determined from laboratory measurements. For example, an empirical formula for bound–free transitions is

$$\sigma_{ik}(\nu) = \sigma_0\left[\beta\left(\frac{\nu_0}{\nu}\right)^{\gamma} + (1 - \beta)\left(\frac{\nu_0}{\nu}\right)^{\gamma+1}\right] \tag{66}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where $\sigma_0, \beta, \gamma, \nu_0$ are fitting parameters. Likewise, the corresponding emission coefficient is

$$\eta(\nu) = \frac{2h\nu^3}{c^2}\left\{\sum_{j=2}^{N}\sum_{i=1}^{j-1}\sigma_{ij}(\nu)\frac{g_i}{g_j}n_j + n_e\sum_{i=1}^{N}\sigma_{ik}(\nu)n_k\phi_i(T)\exp\left(-\frac{h\nu}{kT}\right) + n_e n_k\sigma_{kk}(\nu)\exp\left(-\frac{h\nu}{kT}\right)\right\}. \quad (67)$$

Both $\chi(\nu)$ and $\eta(\nu)$ depend on the level populations, which in turn are determined by both collisional and radiative processes, the latter dependent upon the intensity of the radiation field. On the other hand, the radiation intensity depends on $\chi(\nu)$ and $\eta(\nu)$ of the medium. Hence a self-consistent solution of the radiation transport problem involves solving simultaneously the equation of radiation transport and equations which govern the level populations. These depend on the thermodynamic state of the medium as discussed below.

Local thermodynamic equilibrium A system of gas particles (e.g. atoms, ions, molecules) and photons are said to be in *thermodynamic equilibrium* (TE) when all of the following conditions are met:

(a) The velocities of each species of particles have a *Maxwellian velocity distribution* $f_M(v)$, characterized by the same temperature T for all species

$$f_M(v) = 4\pi\frac{m^3}{(2\pi mkT)^{\frac{3}{2}}}v^2\exp\left(-\frac{mv^2}{2kT}\right) \quad (68)$$

or in terms of kinetic energy $E = \frac{1}{2}mv^2$

$$f_M(E) = \frac{2\pi}{(\pi kT)^{\frac{3}{2}}}\sqrt{E}\exp\left(-\frac{E}{kT}\right). \quad (69)$$

(b) The distribution of atoms or molecules over the energy levels obeys the *Boltzmann law or distribution*, i.e. for levels i and j, the populations n_i^* and n_j^* are given by

$$\frac{n_j^*}{n_i^*} = \frac{g_j}{g_i}\exp\left(-\frac{h\nu_{ji}}{kT}\right) \quad (70)$$

where $h\nu_{ji} = E_j - E_i$ and g_i, g_j are the statistical weights.

(c) The distribution of electrons, ionized and neutral atoms is given by the *Saha equation*

$$n_e\frac{n_+^*}{n_0^*} = \frac{U_+}{U_0}\frac{2(2\pi mkT)^{\frac{3}{2}}}{h^3}\exp\left(-\frac{h\nu_0}{kT}\right) \quad (71)$$

where $h\nu_0 = E_0$ is the ionization potential, U_+ and U_0 are the partition functions of the ionized and neutral atoms.

(d) The radiation field is isotropic, homogeneous and stationary. The frequency dependence of the intensity is a universal function of the system's temperature T. It is given by the *Planck function* and does not depend upon the properties of the matter

$$I(\nu) = B(\nu) = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{kT}\right) - 1\right]^{-1}. \quad (72)$$

(e) The *principle of detailed balancing* applies, which states that in thermodynamic equilibrium (TE) every detailed microscopic process is statistically balanced by its precise inverse (reciprocal) process.

(f) *Kirchhoff's law* applies, which relates the emission and extinction coefficients to the Planck function

$$\eta(\nu) = \chi(\nu)\,B(\nu, T), \text{ or } S(\nu) \equiv \frac{\eta(\nu)}{\chi(\nu)} = B(\nu, T). \quad (73)$$

In other words, if a system of gas particles and photons is in TE, its properties can be characterized by a single parameter, the temperature T. In most astronomical situations, the conditions for TE are rarely met. However, it is often found that the above conditions may be satisfied locally in a small region. If we can characterize the properties of each region by a different T, then it may be possible to describe the properties of the whole system by a temperature distribution. If these conditions are satisfied at each point in a gas, and the temperature T is permitted to change from point to point, the gas is said to be in *local thermodynamic equilibrium (LTE)*. One often finds that when collisional processes dominate radiative processes in the ionization and excitation of atoms, the state of the gas is close to LTE, such as in stellar interiors.

When more than one of the conditions for TE are violated in a system, either locally or globally, the system is said to be in *non-local thermodynamic equilibrium (NLTE)*. Even in systems which are described as NLTE, sometimes one or more species of particles has a Maxwellian velocity distribution and it is possible to introduce the concept of kinetic temperature. In most astrophysical problems, deviation from the Maxwellian velocity distribution is small, although the assumption of TE or LTE cannot be made.

Under LTE, the level populations are governed by the Boltzmann distribution for bound levels and by the Saha equation for ionized states. Hence under LTE conditions both the extinction and emission coefficients can be determined if the gas temperature is known. On the other hand, for NLTE one must determine the level populations which in turn are governed by both collisional and radiative processes, the latter dependent upon the intensity of the radiation field.

Equations of statistical equilibrium Let us now consider the equations by which we determine the level populations of atoms. These are governed by the rates of all processes,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

collisional and radiative, by which an atom leaves a certain state i to some other state j and return rates of processes from all states j back to state i. The collisional and radiative processes that determine the level populations for bound–bound and bound–free transitions are:

Bound–bound	Bound–free
Radiative	
Photoabsorption (R_{ij})	Photoionization (R_{ik})
Spontaneous + stimulated emission (R_{ji})	Spontaneous + stimulated recombination (R_{ki})
Collisional	
Collisional excitation (C_{ij})	Collisional ionization (C_{ik})
Collisional de-excitation (C_{ji})	Three-body recombination (C_{ki})

Let the total rate be:

$$\begin{array}{ll} \text{upward rate:} & P_{ij} = C_{ij} + R_{ij} \\ \text{downward rate:} & P_{ji} = C_{ji} + R_{ji}. \end{array} \tag{74}$$

The *equations of statistical equilibrium*, also known as *rate equations*, for an atom with N levels are given by

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = \sum_{j\neq i}^{N} n_j P_{ji} - n_i \sum_{j\neq i}^{N} P_{ij} = 0, \qquad \text{for level } i. \tag{75}$$

The first term represents the gain in population due to contributions from all other levels while the second term is the loss of population to all other levels. Note that while we can write down one such equation for each of the N levels, only $(N-1)$ of these are truly independent. These $(N-1)$ equations must be solved subject to the constraint of particle conservation

$$\sum_{i=1}^{N} n_i = n_T \tag{76}$$

leading to N equations in N unknowns.

For a moving medium with local fluid velocity $\boldsymbol{v}$ and time-dependent situations, the equations governing level populations take the form

$$\frac{\partial n_i}{\partial t} = -\nabla\cdot(n_i\boldsymbol{v}) + \sum_{j\neq i}^{N} n_j P_{ji} - n_i \sum_{j\neq i}^{N} P_{ij} \tag{77}$$

which simply states that the rate of change of n_i is determined by the net rate of flow of particles into the volume and the net rate of transitions into level i from all other levels j by collisional and radiative processes. Summing over all states, the rate terms cancel one another, and we obtain a continuity equation

$$\frac{\partial n_T}{\partial t} = -\nabla\cdot(n_T\boldsymbol{v}) \tag{78}$$

where we made use of the constraint of particle conservation. The equations of statistical equilibrium are linear equations if the collisional and radiative rates are known.

Expressions for radiative and collisional rates Hence, to describe a system under NLTE conditions we must determine in detail the collisional and radiative processes by solving the rate equations. In particular, for each transition (or pair of energy levels), we need to determine the cross sections: collisional excitation, collisional de-excitation, radiative excitation and radiative de-excitation. We can express the radiative rates for both bound–bound and bound–free transitions in a common form

$$\text{upward rate:} \quad R_{ij} = 4\pi \int \frac{\sigma_{ij}(\nu)J(\nu)}{h\nu}\,\mathrm{d}\nu \tag{79}$$

$$\text{downward rate:} \quad R_{ji} = 4\pi \int \frac{\sigma_{ij}(\nu)J(\nu)}{h\nu} G_{ij}(\nu) \times \left(\frac{2h\nu^3}{c^2} + J(\nu)\right)\mathrm{d}\nu \tag{80}$$

$$G_{ij}(\nu) = \begin{cases} g_i/g_j & \text{(b-b)} \\ n_e\phi_i(T)\exp\left(-\frac{h\nu}{kT}\right) & \text{(b-f)} \end{cases} \tag{81}$$

where the subscript k stands for continuum or ionized state. $\phi_i(T)$ is the Saha–Boltzmann factor and $\sigma_{ij}(\nu)$ is the appropriate cross section for radiative excitations (either photoabsorption or photoionization). $J(\nu)$ is the mean intensity at frequency ν.

Consider the excitation scheme in which an electron with kinetic energy E excites an atom and so loses energy E. In the corresponding de-excitation, the electron follows the reverse path. The total collisional rates are given by

$$\text{excitation:} \quad C_{ij} = n_e \int_{E_0}^{\infty} \sigma_{ij}(E)\sqrt{\frac{2E}{m}}\, f(E)\,\mathrm{d}E \tag{82}$$

$$\text{de-excitation:} \quad C_{ji} = n_e \int_{0}^{\infty} \sigma_{ji}(E-E_0)\sqrt{\frac{2(E-E_0)}{m}} \times f(E-E_0)\,\mathrm{d}(E-E_0) \tag{83}$$

where $f(E)$ is the energy distribution of the colliding particles. It can be shown that the collisional cross sections are related by

$$\frac{\sigma_{ij}(E)}{\sigma_{ji}(E-E_0)} = \frac{g_j}{g_i}\frac{E-E_0}{E}. \tag{84}$$

This relation is valid in general, independent of the energy (or velocity) distribution of the electrons because the ratio σ_{ij}/σ_{ji} depends only on atomic parameters (e.g. g_i, g_j and E_0) and is independent of quantities like n_i, n_j, and $f(E)$ that depend on the thermodynamic state of the system. For a Maxwellian velocity distribution, the ratio of collisional rates is

$$\frac{C_{ij}}{C_{ji}} = \frac{g_j}{g_i}\exp\left(-\frac{E_0}{kT}\right). \tag{85}$$

For non-Maxwellian velocity distribution, one must use the relation σ_{ij}/σ_{ji} to derive a different expression for the ratio C_{ij}/C_{ji}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Consider the process of collisional ionization in which an electron of energy E produces an ion together with a pair of electrons of energies E' and $E - E_0 - E'$, where E_0 is the ionization potential of an atom in state i. The cross sections in collisional ionization (σ_{ik}) and three-body recombination (σ_{ki}) are related by

$$\frac{\sigma_{ik}(E, E')}{\sigma_{ki}(E', E - E_0 - E')} = \frac{16\pi}{h^3}\frac{U_k}{U_i}\frac{mE'(E - E_0 - E')}{E} \quad (86)$$

where U_i and U_k are partition functions of the atoms of states i and k. The collisional ionization rate is given by

$$C_{ik} = n_e \int_{E_0}^{\infty} \mathrm{d}E \int_0^{E-E_0} \sigma_{ik}(E, E')\sqrt{\frac{2E}{m}} f(E)\,\mathrm{d}E' \quad (87)$$

where $f(E)$ is the velocity distribution of electrons. A similar expression can be written for C_{ki}, the three-body recombination rate. For Maxwellian velocity distribution, we have

$$\frac{C_{ik}}{C_{ki}} = \frac{U_k}{U_i}\frac{2(2\pi mkT)^{3/2}}{h^3}\exp\left(-\frac{E_0}{kT}\right). \quad (88)$$

Gas cooling by line radiation Atoms or molecules are excited either by collisions or by radiation. The excitation energy is supplied either from the kinetic energy of the exciting particle, in the case of collisional excitation, or from the absorption of a photon, as in the case of radiative excitation. Similarly, excited molecules may lose their energy either by collisional de-excitation or by spontaneous and stimulated emission. In stimulated emission, a photon stimulates the emission of a similar photon from an excited molecule which is then de-excited. Thus the conversion of radiative to kinetic energy occurs through collisional de-excitation of an excited molecule following the absorption of a photon. On the other hand, kinetic energy is converted into radiative energy in collisional excitation followed by stimulated or spontaneous emission.

To derive an expression for the rate of gas cooling, we consider a two-component system consisting of gas atoms (or molecules) and photons. Since collisional and radiative processes are the only physical processes occurring in the system, the net rate of energy loss by the gas component must be equal to the net rate of radiative energy gain by the photon component, from the principle of conservation of energy, i.e. the energy gain or loss through collisional processes must be equal to that by radiative processes. We may generalize the above results to atoms with N bound levels. In this case, the net cooling rate becomes

$$\Lambda = \sum_{j=2}^{N}\sum_{i=1}^{j-1}(n_i C_{ij} - n_j C_{ji})E_{ji} = \sum_{j=2}^{N}\sum_{i=1}^{j-1}(n_j R_{ji} - n_i R_{ij})E_{ji}. \quad (89)$$

The net cooling rate due to line emission can also be expressed in terms of the emission and extinction coefficients as

$$\Lambda = \int_0^{\infty} \mathrm{d}\nu \oint \mathrm{d}\Omega[\eta(\nu) - \chi(\nu)I(\nu)]. \quad (90)$$

When a system of gas and photons is in TE, the principle of detailed balancing applies, i.e.

$$\left.\begin{array}{l} n_i C_{ij} = n_j C_{ji} \\ n_j R_{ji} = n_i R_{ij} \end{array}\right\} \Rightarrow \Lambda = 0 \quad (91)$$

so that there is no net cooling. While TE clearly implies $\Lambda = 0$, the reverse is not necessarily true. Indeed the condition that $\Lambda = 0$ is equivalent to that of radiative equilibrium, which states that the radiative energy absorbed in any volume element is balanced locally by the energy emitted. It does not require the balancing of individual microscopic processes.

Theoretically the temperature distribution in the gas component is determined by the thermal balance between various heating and cooling processes. The gas cools by the emission of line photons. In general, the gas cooling rate is determined by the difference between the kinetic energy loss through collisional excitations and the kinetic energy gain through collisional de-excitations. For isothermal and homogeneous media, the cooling rate decreases toward the interior region where radiation trapping is most significant. Because cooling results from a binary process (collisions), we expect the cooling rate to depend quadratically on the gas density under many astrophysical environments. In extreme cases of very high densities and optical depth, the cooling rate is roughly independent of density, because most of the cooling lines are saturated and thermalized. The cooling rate increases with temperature because more cooling species are excited collisionally to higher levels. Because photons from higher transitions not only carry more energy but also are created more frequently, the net cooling rate is increased. In general, the line transitions contributing most to the total cooling rate are those in which the energy for excitation is comparable to the thermal energy of the gas.

Dust component

When radiation passes through a region of space containing dust grains, the photons can be either absorbed, with conversion of radiation into heat, or scattered, with a change in the photon's direction of travel. The sum of these two processes is called extinction, since a collimated beam of light can be extinguished by either process. The probabilities of absorbing and scattering photons are given respectively by the absorption and scattering cross sections, the sum of which is called the extinction cross section. The ratio of the scattering cross section to the total extinction cross section is called the ALBEDO which measures the relative importance of scattering as opposed to true absorption. For dust grains in astrophysical environments, both the extinction cross section and the albedo decrease as the wavelength increases, with scattering dominating absorption in the UV and visible and vice versa in the infrared. The albedo approaches unity in the UV and visible, while in the infrared there is very little scattering and the albedo essentially goes to zero.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In general, the scattering and absorption cross sections for dust grains depend on the grain composition, shape, size and wavelength of radiation. The interaction of radiation with a particle of homogeneous solid material and refractive index m is completely described by Maxwell's equations of electromagnetism. If we can solve these equations for the appropriate boundary conditions, we can determine the cross sections. For spherical particles which are homogeneous and optically isotropic, these cross sections can be calculated from the Mie theory of scattering based on optical constants of the grain material. The cross sections are often expressed as the product of a dimensionless efficiency factor Q and the geometric cross section πa^2, where a is the grain radius

$$\text{absorption efficiency factor } Q_{\rm abs}(\lambda) \equiv \frac{\sigma_{\rm abs}(\lambda)}{\pi a^2} \qquad (92)$$

$$\text{scattering efficiency factor } Q_{\rm sca}(\lambda) \equiv \frac{\sigma_{\rm sca}(\lambda)}{\pi a^2}. \qquad (93)$$

In general we find:

(a) $Q_{\rm abs}(\lambda)$ and $Q_{\rm sca}(\lambda)$ depend on the wavelength λ only via the dimensionless quantity $x \equiv 2\pi a/\lambda$, i.e. $Q_\alpha = Q_\alpha(x)$. Furthermore, $Q_\alpha(x)$ may be greater than unity for some values of x.
(b) As $x \to \infty$ ($\lambda \gg a$), we have the geometrical limit so that $Q_{\rm sca} \to 1$ and $Q_{\rm abs} \to 1$.
(c) As $x \to 0$ ($\lambda \ll a$), we have Rayleigh scattering for which $Q_{\rm sca} \propto x^4 \propto \lambda^{-4}$ and $Q_{\rm abs} \propto x \propto \lambda^{-1}$.

Although the interaction cross sections for spherical grains may be calculated from the Mie scattering theory, cosmic dust grains have irregular shapes as they are most likely formed by stochastic growth processes. The cross sections for grains of arbitrary shapes may be calculated using a technique called *discrete dipole approximation* (DDA) in which an irregularly shaped grain is approximated by a collection of dipoles. The DDA is valid only when the dipole lattice spacing (Δ) is less than the wavelength of the incident radiation. For a grain approximated by N dipoles, $\Delta = (4\pi/3N)^{1/3} a_{\rm eff}$, where $a_{\rm eff}$ is the radius of a sphere containing the same number of dipoles. For $a_{\rm eff} = 0.1\ \mu$m and $N = 512$, the DDA introduces typical errors of 5% in spectral regions where the approximation is valid.

Before using the DDA to calculate the interaction cross sections, one must first generate a grain shape by computer simulation. A useful parameter to characterize the topology of irregularly shaped grains is the fractal dimension (D), defined by the relation $N(r) \propto r^D$, where $N(r)$ is the number of monomers within a sphere of radius r. For a spherical grain, $D = 3$, while for a rod shaped grain, $D = 1$. In general $1 < D < 3$ and the smaller the fractal dimension, the more filamentary the structure. The thermal and radiative properties of a dust grain depend on the ratio of volume to surface area, which is determined by D. For a given mass, grains with the same D show no significant difference in their absorption cross sections, temperature and energy spectrum, independent of the overall grain shape. Hence in modeling phenomena involving irregularly shaped grains, we need to introduce just one parameter, the fractal dimension D, to characterize the shape. It should be emphasized that porosity [$\equiv$1$-$(volume of solid)/(total volume)] and fractal dimension are not uniquely related. Porosity measures the effect of volume filling factor in a grain and is ill-defined for a dust grain with an open and filamentary structure. Grains with the same porosity may have very different fractal dimensions, and vice versa.

The absorption and scattering coefficients as a function of frequency ν are respectively given by

$$\chi^a(\nu) = n_d \sigma_{\rm abs}(\nu) = n_d Q_{\rm abs}(\nu) \pi a^2 \qquad (94)$$

$$\chi^s(\nu) = n_d \sigma_{\rm sca}(\nu) = n_d Q_{\rm sca}(\nu) \pi a^2. \qquad (95)$$

In general dust grains emit radiation over a continuum of wavelengths, similar to a blackbody but mediated by the overall $\sigma_{\rm abs}$ profile. For continuum radiation, the assumption of LTE is valid so that the thermal emission coefficient can be derived from Kirchhoff's law

$$\begin{aligned}\eta^t(\nu) &= [\chi^a(\nu) + \chi^s(\nu)] B_\nu(T_d) \\ &= n_d [Q_{\rm abs}(\nu) + Q_{\rm sca}(\nu)] \pi a^2 B_\nu(T_d)\end{aligned} \qquad (96)$$

where $B_\nu(T_d)$ is the Planck function at dust temperature T_d. Since coherent scattering is valid for dust grains, the scattering emission coefficient is given by

$$\eta^s(\nu) = \chi^s(\nu) \left(\oint_{\Omega'} \frac{{\rm d}\Omega'}{4\pi} I(\widehat{\Omega}', \nu) p(\widehat{\Omega}', \widehat{\Omega}) \right) \qquad (97)$$

where $p(\widehat{\Omega}', \widehat{\Omega})$ is the scattering phase function.

Phase function and asymmetry parameter The scattering process involving dust grains is coherent since it involves only a change of direction of radiation but not of its wavelength. Coherent scattering by dust particles is described in terms of a phase function $p(\widehat{\Omega}, \widehat{\Omega}')$ which gives the probability of scattering in the direction $\widehat{\Omega}$ of radiation arriving from the direction $\widehat{\Omega}'$ on a volume element at point $\boldsymbol{r}$. The scattering is isotropic or linearly anisotropic when $p(\theta) = 1$ or $p(\theta) = 1 + x \cos\theta$, where $x = \pm 1$ and $\cos\theta = \widehat{\Omega} \cdot \widehat{\Omega}'$. The phase function may either be given in the form of a numerical table from measurements or computations, or represented by a mathematical expression. A useful form is an expansion in Legendre polynomials written as

$$p(\theta) = \sum_{l=0}^{L} \beta_l P_l(\cos\theta);$$

$$\beta_l = \frac{2l+1}{2} \int_{-1}^{+1} p(\theta) P_l(\cos\theta)\, {\rm d}(\cos\theta) \qquad (98)$$

with $\beta_0 = 1$ from the normalization condition and $\beta_1 = 3g$ where g is the *asymmetry parameter* defined as

$$g = \frac{1}{2} \int_{-1}^{+1} p(\theta) \cos\theta\, {\rm d}(\cos\theta) = \frac{\beta_1}{3}. \qquad (99)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Two commonly used analytic approximations for the phase function are the Henyey–Greenstein phase function

$$p(\theta) = \frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{3/2}} \tag{100}$$

and the two-parameter phase function for highly anisotropic scattering

$$p(\theta) = \frac{4\alpha g(1 - g^2)^{2\alpha}}{[(1 + g)^{2\alpha} - (1 - g)^{2\alpha}](1 + g^2 - 2g\cos\theta)^{\alpha+1}} \tag{101}$$

which reduces to the Henyey–Greenstein form for $\alpha = \frac{1}{2}$. The phase function for Rayleigh scattering is given by

$$p(\theta) = \frac{3}{4}(1 + \cos^2\theta). \tag{102}$$

Once the extinction and emission coefficients quantifying the interaction of radiation with gas and dust particles are expressed in terms of microphysical parameters, the equation of radiation transport can be solved in principle. For physically realistic situations, analytical solutions are seldom available so one must resort to numerical methods. Even for time-independent situations involving 1D geometries, the problem of radiation transport is three-dimensional because of the intrinsic dependence of the radiation field on direction and wavelength.

Due to space limitation, the following topics are not covered in this article: methods of solution (numerical and semi-analytic) of the equation of radiation transport; transfer of polarized radiation; radiation hydrodynamics that deal with transfer phenomena involving velocity (relativistic and non-relativistic) flows and/or time dependence.

Bibliography

Bohren C F and Huffman D R 1998 *Absorption and Scattering of Light by Small Particles* (New York: Wiley)

Chandrasekhar S 1960 *Radiative Transfer* (New York: Dover)

Pomraning G C 1973 *The Equations of Radiation Hydrodynamics* (Oxford: Pergamon)

Rybicki G B and Lightman A P 1979 *Radiative Processes in Astrophysics* (New York: Wiley)

Mihalas D and Mihalas B W 1999 *Foundations of Radiation Hydrodynamics* (New York: Dover)

Sen K K and Wilson S J 1998 *Radiative Transfer in Moving Media: Basic Mathematical Methods for Radiative Transfer in Spherically Symmetrical Moving Media* (Berlin: Springer)

Chun Ming Leung

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radiatively Driven Stellar Winds from Hot Stars

A stellar wind is the continuous, supersonic outflow of matter from a star. Among the most massive stars—which tend also to be the hottest and most luminous—the winds can be very strong, with important consequences both for the star's own evolution as well as for the surrounding interstellar medium. Such hot-star winds are understood to be driven by the pressure of the star's emitted radiation.

Our Sun has a SOLAR WIND, but it is so tenuous and transparent that it would be difficult to detect directly from a more distant star. However, many stars have winds that are dense enough to be opaque at certain wavelengths of the star's radiation, and this makes it possible to study them remotely through careful interpretation of the observed stellar spectra. The winds from massive, hot stars—with surface temperatures above about 10 000 K—form a well-defined class, distinct from the winds from cooler stars, and characterized by the central role of the star's own radiation in driving the mass outflow. The high temperature of such stars means that they have a very high surface brightness. Because light carries momentum as well as energy, this high surface brightness imparts a force to the atoms that scatter the light. At the level of the star's atmosphere where this force exceeds the inward force of the stellar gravity, material is accelerated upward and becomes the stellar wind.

An important aspect of this radiative driving process is that it stems mostly from *line* scattering, wherein an electron is shuffled between two discrete, bound energy levels of an atom. In a static medium such scattering is confined to radiation with a photon energy near the energy difference between the levels, corresponding to a range of wavelengths near a distinct, line-center value. However, in an accelerating stellar wind flow, the Doppler effect shifts the resonance to increasingly longer wavelengths, allowing the line scattering to sweep gradually through a much broader portion of the stellar spectrum. This gives the dynamics of such winds an intricate feedback character, in which the radiative driving force that accelerates the stellar wind depends itself on that acceleration.

Overview and comparison with the solar wind

The general concept that stars can sometimes lose or eject matter is apparent from some spectacular, catastrophic examples, such as nova or supernova explosions. However, the specific notion of a *continuous*, quiescent-phase stellar wind outflow stems largely from discovery, during the 1950s and 1960s, of the *solar* wind. The Sun is a relatively low-mass, cool star with a surface temperature about 6000 K, but curiously its wind arises from pressure expansion of the very hot, million kelvin solar corona, which is somehow superheated by the mechanical energy generated from convection in the Sun's subsurface layers. By contrast, high-mass stars with much higher surface temperatures (10 000–100 000 K) are thought to lack the strong convection zone needed to heat a circumstellar corona. Their stellar winds thus remain at temperatures comparable with the star's surface, and so lack the very high gas pressure needed to drive an outward expansion against the stellar gravity. However, such hot stars have a quite high radiative flux, since by the Stefan–Boltzmann law this scales as the fourth power of the surface temperature. It is the pressure of this *radiation* (not of the gas itself) that drives the wind expansion.

The typical flow speeds of hot-star winds—up to about 3000 km s^{-1}—are a factor of a few faster than the 400–700 km s^{-1} speed of the solar wind. However, the inferred mass loss rates of hot stars greatly exceed—by up to a factor of a billion—that of the Sun. At the Sun's current rate of mass loss, about $10^{-14} M_\odot$ yr^{-1}, its mass would be reduced by only $\sim$0.01% during its entire characteristic lifespan of 10 billion (10^{10}) years. By contrast, even during the much shorter, few million (10^6) year lifetime typical for a massive star, its wind mass loss at a rate of up to $10^{-5} M_\odot$ yr^{-1} can substantially reduce, by a factor of 2 or more, the original stellar mass of a few times $10 M_\odot$. Indeed, massive stars typically end up as 'WOLF–RAYET' STARS, which often appear to have completely lost their original envelope of hydrogen, leaving exposed at their surface the elements such as carbon, nitrogen and oxygen that were synthesized by nuclear processes in the stellar core. (Much of this cumulative mass loss might also occur during relatively brief 'superwinds' or epsisodic ejections as a 'luminous blue variable'.) In addition to directly affecting the star's own evolution, hot-star winds often form 'wind-blown bubbles' in nearby interstellar gas. Overall they represent a substantial contribution to the energy, momentum and chemical enrichment of the interstellar medium in the Milky Way and other galaxies.

Wind detection through asymmetric, P-Cygni line profiles

The high density of hot-star winds means that the expanding material is itself opaque, or *optically thick*, to scattering in many atomic spectral-line transitions. When observed as part of a stellar spectrum, radiation that is line scattered within the wind develops a characteristic line profile, known as a 'P-Cygni profile', after the star P-Cygni in which its significance as a signature of mass outflow was first broadly recognized.

Figure 1 illustrates the formation of a P-Cygni line profile. Wind material approaching an observer within a column in front of the star has its line resonance blueshifted by the Doppler effect. Thus scattering of stellar radiation out of this direction causes an absorption trough, or reduction in the observed flux, on the blue side of the line profile. However, from the lobes on either side of this absorption column, wind material can also scatter radiation *toward* the observer. Since this can occur from either the approaching or the receding hemisphere, this scattered radiation can be either blueshifted or redshifted. The associated extra flux seen by the observer thus occurs as a symmetric emission component on both sides of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

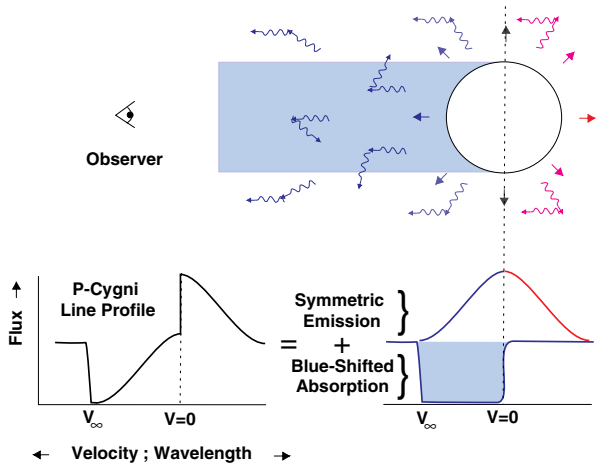


Figure 1. Schematic illustration of the formation of a P-Cygni type line profile in an expanding wind outflow.

the line center. Combined with the reduced blue-side flux, the overall profile has a distinctly asymmetric form, with apparent net blueward absorption and redward emission. Moreover, the wavelength of the blue edge of the absorption provides a quite unambiguous measure of the asymptotic wind speed, v_∞.

The radiative force from electron scattering

A radiative force results from the material interception of radiative momentum. A particularly simple case is scattering by free electrons, which is a 'gray', or frequency-independent, process. Since gray scattering cannot alter the star's total luminosity L, the radiative energy flux at any radius r is simply given by $L/4\pi r^2$. This corresponds to a radiative momentum flux of $L/4\pi r^2 c$, where c is the speed of light. The material interception of this flux is characterized by an interaction cross section per unit mass, κ, which is formally named the 'mass absorption coefficient' but more succinctly is also commonly called the 'opacity'. For electron scattering in an ionized medium, the opacity is simply a constant given by $\kappa_e = \sigma_e/\mu_e$, where σ_e (=0.66 $\times$ 10^{-24} cm^2) is the classical Thompson cross section, and the mean atomic mass per free electron is $\mu_e = 2m_H/(1+X)$, with m_H and X the mass and mass fraction of hydrogen. This works out to a value $\kappa_e = 0.2(1+X) = 0.34$ cm^2 g^{-1}, where the latter result applies for the standard (solar) hydrogen mass fraction $X = 0.72$. The product of this opacity and the radiative momentum flux yields the radiative acceleration (force per unit mass) from free-electron scattering,

$$g_e(r) = \frac{\kappa_e L}{4\pi r^2 c}.$$

It is of interest to compare this with the star's gravitational acceleration, given by GM/r^2, where G is the gravitation constant and the M is the stellar mass. Since both accelerations have the same $1/r^2$ dependence on radius, their ratio is spatially constant, fixed by the ratio of luminosity to mass,

$$\Gamma_e = \frac{\kappa_e L}{4\pi G M c}.$$

This ratio, sometimes called the Eddington parameter, thus has a characteristic value for each star. For the Sun it is very small, of order 2 $\times$ 10^{-5}, but for hot, massive stars it is often within a factor of 2 below unity. As noted by Eddington, electron scattering thus represents a basal radiative acceleration that effectively counteracts the stellar gravity. The limit $\Gamma_e \to 1$ is known as the Eddington limit, for which the star would become gravitationally unbound.

It is certainly significant that hot stars with strong stellar winds have Γ_e only a factor of 2 or so below this limit, since it suggests that only a modest additional opacity could succeed in fully overcoming gravity to drive an outflow. However, it is important to realize that a stellar wind represents the outer envelope outflow from a nearly static, gravitationally bound base, and as such is not consistent with an entire star exceeding the Eddington limit. Rather the key requirement for a wind is that the driving force increase naturally from being smaller to larger than gravity at some radius near the stellar surface. We shall now see that the force from line scattering is ideally suited for just such a spatial modulation.

The Doppler-shifted resonance of line scattering

When an electron is bound into one of the discrete energy levels of an atom, its scattering of radiation is primarily with photons of just the right energy to induce shuffling with another discrete level. The process is called line scattering, because it often results in the appearance of narrowly defined lines in a star's energy spectrum. At first glance, it may seem unlikely that such line scattering could be effective in driving mass loss, simply because the opacity only interacts with a small fraction of the available stellar flux. However, there are two key factors that work to make line scattering in fact the key driving mechanism for hot-star winds.

The first is the resonant nature of line scattering. The binding of an electron into discrete energy levels of an atom represents a kind of resonance cavity that can greatly amplify the interaction cross section with photons of just the right energy to induce transition among the levels. The effect is somewhat analogous to blowing into a whistle versus just into open air. Like the sound of a whistle, the response occurs at a well-tuned frequency and has a greatly enhanced strength. Relative to free-electron scattering, the overall amplification factor for a broad-band, untuned radiation source is set by the quality of the resonance, $Q \approx \nu_0/A$, where ν_0 is the line frequency and A is the decay rate of the excited state. For quantum mechanically allowed atomic transitions, this can be very large, of order 10^7. Thus, even though only a very small fraction ($\sim 10^{-4}$) of electrons in a hot-star atmosphere are bound into atoms, illumination of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

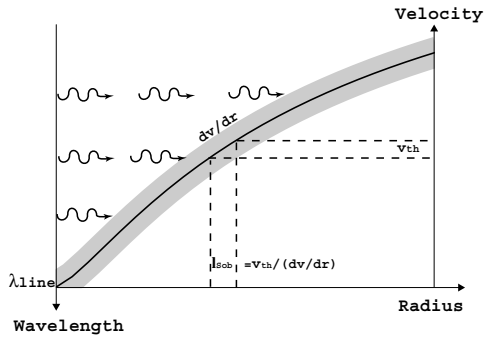


Figure 2. Doppler-shifted line resonance absorption in an accelerating flow. Photons with a wavelength just shortward of a line propagate freely from the stellar surface until redshifted, within a reference frame moving with the wind outflow velocity, into a narrow line resonance, whose width is set by the thermal broadening of the line, as represented here by the shading.

these atoms by an unattenuated (i.e. optically thin), broad-band radiation source would yield a collective line force that exceeds that from free electrons by about a factor $\bar{Q} \approx 10^7 \times 10^{-4} = 1000$. For stars within a factor of 2 of the free-electron Eddington limit, this implies that line scattering is capable, in principle, of driving material outward with an acceleration on the order of a thousand times the inward acceleration of gravity.

In practice, of course, this does not normally occur, since any sufficiently large collection of atoms scattering in this way would quickly block the limited flux available within just the narrow frequency bands tuned to the lines. Indeed, in the static portion of the atmosphere, the flux is greatly reduced at the line frequencies. Such line 'saturation' keeps the overall line force quite small, in fact well below the gravitational force, which thus allows the inner parts of the atmosphere to remain gravitationally bound.

This, however, is where the second key factor, the Doppler effect, comes into play. In the outward-moving portions of the outer atmosphere, the Doppler effect redshifts the local line resonance, effectively desaturating the lines by allowing the atoms to resonate with relatively unattenuated stellar flux that was initially at slightly higher frequencies. By effectively sweeping a broader range of the stellar flux spectrum, this makes it possible for the line force to overcome gravity and accelerate the very outflow it itself requires. As quantified within the CAK wind theory described later, the amount of mass accelerated adjusts such that the self-absorption of the radiation reduces the overall line driving to being just somewhat (not a factor of 1000) above what is needed to overcome gravity.

The Sobolev approximation for line scattering in an accelerating flow

The late Russian astrophysicist V V Sobolev developed an extremely useful approach for treating line scattering in such a rapidly accelerating flow. As illustrated in figure 2, he noted how radiation emitted from the star at a wavelength somewhat blueward of a line propagates freely until it is redshifted by the accelerating flow into a local line resonance. For the usual case that the broadening of the line is set by the ion thermal speed $v_{\rm th}$, the geometric width of this resonance is about a Sobolev length, $l_{\rm Sob} \equiv v_{\rm th}/(\mathrm{d}v/\mathrm{d}r)$. In a supersonic flow, this Sobolev length is of order $v_{\rm th}/v \ll 1$ smaller than a typical flow variation scale, like the density–velocity scale length $H \equiv |\rho/(\mathrm{d}\rho/\mathrm{d}r)| \approx v/(\mathrm{d}v/\mathrm{d}r)$.

The nearly homogeneous conditions within such resonance layers imply that the key parameters of the line scattering can be described in terms of strictly *local* conditions at any radius. In particular, the total optical depth of radiation passing through the line resonance—which normally requires evaluation of a nonlocal spatial integral—can in this case be well approximated simply in terms of the local density and velocity gradient,

$$\tau = \frac{\kappa \rho v_{\rm th}}{\mathrm{d}v/\mathrm{d}r} \tag{1}$$

where κ is the line opacity. This allows a localized solution for how much the flow absorption reduces the illumination of the line resonance by the stellar flux and leads to a simple, general expression for the line acceleration,

$$g_{\rm line} \approx g_{\rm thin} \frac{1 - \mathrm{e}^{-\tau}}{\tau}. \tag{2}$$

In the optically thin limit $\tau \ll 1$, the line acceleration reduces to a form similar to the electron scattering case,

$$g_{\rm thin} \equiv \frac{\kappa v_{\rm th} \nu_0 L_\nu}{4\pi r^2 c^2} \approx \frac{\kappa}{\kappa_{\rm e}} \frac{v_{\rm th}}{c} g_{\rm e} \tag{3}$$

wherein the last approximate equality applies the usual assumption that the line frequency ν_0 is near the peak of the stellar luminosity spectrum L_ν, so that $\nu_0 L_\nu \approx L$. In the opposite limit of an optically thick line with $\tau \gg 1$, there results a quite different form,

$$g_{\rm thick} \approx \frac{g_{\rm thin}}{\tau} = \frac{L}{4\pi r^2 \rho c^2} \frac{\mathrm{d}v}{\mathrm{d}r} = \frac{L}{\dot{M} c^2} v \frac{\mathrm{d}v}{\mathrm{d}r} \tag{4}$$

where the last equality uses the definition of the wind mass loss rate, $\dot{M} \equiv 4\pi \rho v r^2$.

A key result here is that the optically thick line force is independent of the opacity and instead varies in proportion to the velocity gradient $\mathrm{d}v/\mathrm{d}r$. The basis of this result is illustrated by figure 2, which shows that the local rate at which stellar radiation is redshifted into a line resonance depends on the slope of the velocity. By Newton's famous equation of motion, a force is normally understood to *cause* an acceleration. However, here we see that an optically thick line force also *depends* on the wind's advective rate of acceleration, $v\,\mathrm{d}v/\mathrm{d}r$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Derivation of direct absorption force for a single line

Let us derive this key property of line driving in more quantitative terms. Under the simplifying approximation that the stellar radiation flux is purely radial (as from a central point source), the force per unit mass associated with direct absorption by a single line at a radius r is given by

$$g_{\rm line}(r) = g_{\rm thin} \int_{-\infty}^{\infty} {\rm d}x\, \phi[x - v(r)/v_{\rm th}]\, {\rm e}^{-t(x,r)}. \qquad (5)$$

The integration is over a scaled frequency $x \equiv (\nu/\nu_0 - 1)\, c/v_{\rm th}$, defined from a line center frequency, ν_0, in units of the frequency broadening associated with the ion thermal motion, $v_{\rm th}$. The integrand is weighted by the line profile function $\phi(x)$, which for thermal broadening typically has the Gaussian form $\phi(x) \sim {\rm e}^{-x^2}$. The exponential reduction takes account of absorption, as set by the frequency-dependent optical depth to the stellar surface radius R_*,

$$t(x, r) \equiv \int_{R_*}^{r} {\rm d}r'\, \kappa\rho(r')\phi[x - v(r')/v_{\rm th}].$$

A crucial point in evaluating this integral is that, in a supersonic wind, the variation of the integrand is dominated by the velocity variation within the line profile. As noted above, this variation has a scale given by the Sobolev length $l_{\rm Sob} \equiv v_{\rm th}/({\rm d}v/{\rm d}r)$, which is smaller by a factor $v_{\rm th}/v$ than the competing density–velocity scale, $H \equiv |\rho/({\rm d}\rho/{\rm d}r)| \approx v/({\rm d}v/{\rm d}r)$. A key step in the Sobolev approximation is thus to recast this spatial integration as an integration over the comoving frame frequency $x' \equiv x - v(r')/v_{\rm th}$,

$$\begin{aligned} t(x, r) &= -\int_{x'(R_*)}^{x'(r)} \frac{{\rm d}x'\, v_{\rm th}}{{\rm d}v/{\rm d}r'} \kappa\rho(x')\, \phi(x') \\ &\approx \tau(r) \int_{x - v(r)/v_{\rm th}}^{\infty} {\rm d}x'\, \phi(x') \end{aligned} \qquad (6)$$

where the latter approximation uses the assumption that $v(r) \gg v_{\rm th}$ to formally extend the surface frequency $x'(R_*)$ to infinity relative to the local resonance $x'(r)$. The quantity $\tau(r)$, which is just the local Sobolev optical thickness defined in equation (1), arises as a collection of spatial variables that are assumed to be nearly constant over the Sobolev resonance zone and thus can be extracted outside the integral.

Finally, a remarkable, extra bonus from this approximation is that the resulting optical depth (6) now has precisely the form needed to allow analytical evaluation of the line force integral (5), yielding directly the general expression given in equation (2).

The mass loss rate from inertial force balance

We can actually estimate the wind mass loss rate from a simple consideration of how this line force overcomes the wind inertia (for now ignoring gravity). For a single optically thick line near the peak of the stellar flux spectrum, the resulting steadystate flow acceleration is given by

$$v\frac{{\rm d}v}{{\rm d}r} = g_{\rm line} \approx \frac{L}{\dot{M}c^2} v \frac{{\rm d}v}{{\rm d}r}.$$

Since the acceleration cancels from both sides, this simply implies $\dot{M} \approx L/c^2$, or that the total mass loss of the wind is roughly equal to just the mass loss associated with the stellar luminosity. In practice, hot-star winds are driven by not just one but many optically thick lines. As long as these are independent of each other, the total resulting mass loss simply accumulates in proportion to the total number of thick lines,

$$\dot{M} \approx N_{\rm thick} \frac{L}{c^2}. \qquad (7)$$

For a wind with terminal speed v_∞, each line sweeps through about a fraction v_∞/c of the spectrum. This means that there can at most be about $N_{\rm thick} \approx c/v_\infty$ nonoverlapping thick lines spread throughout the spectrum. This implies a single-scattering limit for the mass loss,

$$\dot{M} v_\infty < \frac{L}{c}.$$

The winds from stars of spectral types O and B are generally within this limit, but winds from Wolf–Rayet (WR) stars can exceed it by a factor of 10 or more. As further discussed below, the driving of such WR winds thus requires a more intricate, multiline scattering process.

The CAK theory for line-driven winds

In practice, the number of thick lines depends itself on the mass loss rate and so is not known *a priori*. Self-consistent solution of the wind properties is possible, however, through the formalism developed by Castor, Abbott and Klein (CAK). A key simplification is to approximate the flux-weighted number distribution of lines as a power law in the line opacity κ,

$$\frac{{\rm d}N}{{\rm d}\kappa} = \frac{1}{\kappa_0} \left(\frac{\kappa}{\kappa_0}\right)^{\alpha-2}$$

where the CAK power-law index satisfies $0 < \alpha < 1$, and κ_0 is a normalization constant defined such that $\kappa_0\, {\rm d}N/{\rm d}\kappa \equiv 1$. (The latter is related to the commonly defined CAK constant k by $k = \Gamma(\alpha)(v_{\rm th}/c)(\kappa_0/\kappa_{\rm e})^{1-\alpha}/(1-\alpha)$, where Γ denotes the complete gamma function.) The cumulative force from this line ensemble can be obtained by an integration of the single-line result (2) over this opacity distribution. This yields a kind of 'geometric mean' between the optically thin and thick forms (equations (3) and (4)) for a single line,

$$g_{\rm cak} = \frac{g_{\rm thin,o}}{\tau_{\rm o}^{\alpha}}. \qquad (8)$$

Here $\tau_{\rm o}$ and $g_{\rm thin,o}$ are respectively the Sobolev optical thickness and the optically thin acceleration for a single,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

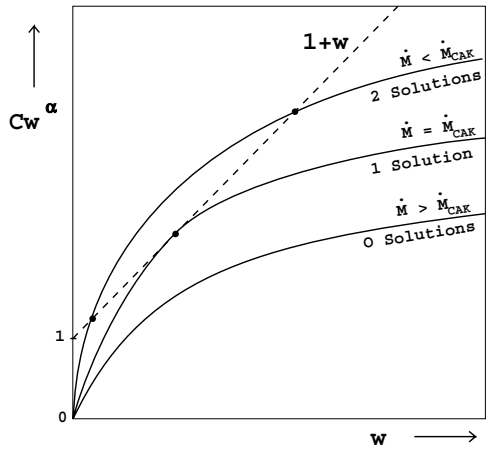


Figure 3. Graphical solutions of the dimensionless equation of motion (9).

spectral-peak line of opacity κ_o. This normalization opacity is given roughly by the line resonance amplification factor discussed previously,

$$\frac{\kappa_o v_{th}}{c} \approx \bar{Q}\kappa_e \approx 10^3 \kappa_e$$

which thus implies that $g_{thin,o} \approx 10^3 g_e$, as noted earlier. In this context, $\bar{Q}$ can be thus thought of an alternative, dimensionless way to specify the opacity normalization.

Ignoring the generally unimportant contribution of gas pressure, the steady-state equation of motion simply requires that the wind acceleration must equal the line acceleration minus the inward acceleration of gravity,

$$v\frac{dv}{dr} = g_{cak} - \frac{GM(1-\Gamma_e)}{r^2}$$

wherein we have also taken into account the effective reduction of gravity by the free-electron scattering factor Γ_e.

Defining the flow acceleration in units of the gravity as

$$w \equiv r^2 v v' / GM(1-\Gamma_e)$$

the equation of motion can be rewritten in the simple, dimensionless form,

$$w = C\, w^\alpha - 1 \qquad (9)$$

where the constant is given by

$$C = \left(\frac{\bar{Q}\Gamma_e}{1-\Gamma_e}\right)^{1-\alpha} \left(\frac{L}{\dot{M}c^2}\right)^\alpha .$$

Figure 3 illustrates the graphical solution of this dimensionless equation of motion for various values of the constant C. For fixed stellar and opacity distribution parameters, this corresponds to assuming various values of the mass loss rate $\dot{M}$. For high $\dot{M}$ there are no solutions, while for low $\dot{M}$ there are two solutions. The two limits are separated by a critical case with one solution—corresponding to the *maximal* mass loss rate—for which the function Cw^α intersects the line $1 + w$ at a tangent. For this critical case, the tangency requirement implies $\alpha C_c w_c^{\alpha-1} = 1$, which together with the original equation (9) yields the critical conditions $w_c = (\alpha/1-\alpha)$ and $C_c = 1/(\alpha^\alpha(1-\alpha)^{1-\alpha})$.

Since equation (9) has no explicit spatial dependence, these conditions hold at all radii. By spatial integration of the critical acceleration w_c from the surface radius R_*, we thus obtain the CAK velocity law,

$$v(r) = v_\infty \left(1 - \frac{R_*}{r}\right)^{1/2} \qquad (10)$$

where the terminal speed is given by $v_\infty = v_{esc}[\alpha/(1-\alpha)]^{1/2}$, where v_{esc} is the effective surface escape speed. Likewise, the associated critical value of C_c defines the maximal CAK mass loss rate

$$\dot{M}_{cak} = \alpha \left(\frac{(1-\alpha)\bar{Q}\Gamma_e}{1-\Gamma_e}\right)^{(1-\alpha)/\alpha} \frac{L}{c^2}. \qquad (11)$$

Comparison with equation (7) shows that the first two factors in equation (11) now provide an explicit expression for the number of thick lines.

These results strictly apply only under the idealized assumption made above that the stellar radiation is radially streaming from a point source. If one takes into account the finite angular extent of the stellar disk, then the mass loss is slightly reduced (typically by about a factor of 2), and the velocity law becomes slightly flatter (approximated by replacing the exponent 1/2 in equation (10) by a value $\beta \approx 0.8$), with a somewhat higher terminal speed ($v_\infty \approx 3v_{esc}$).

Finally, another simplifying assumption of this analysis is that the line opacities are spatially constant, which implies a fixed wind ionization. The effect of a radial change in ionization can be approximately taken into account by correcting the CAK force (8) by a factor of the form $(n_e/W)^\delta$, where n_e is the electron density, $W \equiv 0.5[1-(1-R_*/r)^{1/2}]$ is the radiation 'dilution factor', and the exponent has a typical value $\delta \approx 0.1$. This factor introduces an additional density dependence to that already implied by the optical depth factor $1/\tau_o^\alpha$ given in equation (8). Its general effect on the mass loss can be roughly accounted with the simple substitution $\alpha \to \alpha' \equiv \alpha - \delta$ in the power exponents of the CAK mass loss scaling law (11). The general tendency is to moderately increase $\dot{M}$ and accordingly to somewhat decrease the wind speed.

The wind momentum–luminosity relation

An important success of these CAK scaling laws is the theoretical explanation they provide for an empirically observed 'wind momentum–luminosity' (WML) relation. Combining the CAK mass loss law (11) together with the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scaling of the terminal speed v_∞ with the effective escape speed $v_{\rm esc} = [2GM(1-\Gamma)/R]^{1/2}$, we obtain a WML relation of the form,

$$\dot{M} v_\infty R_*^{1/2} \sim L^{1/\alpha'} Q^{1/\alpha'-1}$$

wherein we have neglected a residual dependence on $M(1-\Gamma)$ that is generally very weak for the usual case that α' is near 2/3. Stellar mass loss rates can be routinely derived by fits to the observed Balmer (Hα) line emission, and wind terminal speeds can be inferred from the blue edge of observed P-Cygni profiles. Combined with spectroscopic estimates of the stellar luminosity and radius (the latter of which in any case enters only weakly as a square root), empirical fits for a large sample of galactic OB supergiants have demonstrated a remarkably tight agreement with a single WML relation. These fits give a luminosity slope that implies $\alpha' \approx 0.57$ and an overall normalization that is quite consistent with the expectation that $\bar{Q} \approx 10^3$.

Less extensive fits have also been derived for OB stars in the Magellanic clouds. These are each separately consistent with a single WML relation, but with a lower normalization, especially for the small cloud. This probably just reflects the weaker radiative driving from the lower metallicity Z, since the line opacity normalization varies in direct proportion to the metallicity, $\bar{Q} \sim Z$. There are currently efforts toward carefully calibrating this WML relation and its dependence on metallicity. The ultimate goal is to apply observationally inferred wind parameters of luminous hypergiant stars in external galaxies as an alternative 'standard candle' for extragalactic distance measurements. In this way hot-star winds may play a role in calibrating the distance scale of the universe.

The Wolf–Rayet wind 'momentum problem'

WR STARS are evolved, massive, hot stars for which the cumulative mass loss has led to depletion of the original hydrogen envelope. They typically show broad wind emission lines of elements such as carbon, nitrogen and/or oxygen that are the products of core nucleosynthesis. Overall, observations indicate that WR winds are especially strong, and even optically thick to continuum scattering by electrons. Notably, inferred WR wind momenta $\dot{M} v_\infty$ are generally substantially higher than for OB stars of comparable luminosity, placing them well above the OB star line in the WML relation. In fact, the inferred ratio of the wind momentum to that of stellar luminosity, $\eta \equiv \dot{M} v_\infty/(L/c)$, is typically much above unity, sometimes as high as $\eta = 10$–50.

This last property has generally been cast as the WR wind 'momentum problem', sometimes with the implication that it implies that WR winds cannot be radiatively driven. In fact, it merely means that, unlike for OB stars, WR winds cannot be treated in the standard, single-scattering formalism, which assumes optically thick lines do not overlap within the wind. However, momentum ratios above unity can, in principle, be achieved by multiple scattering between overlapping

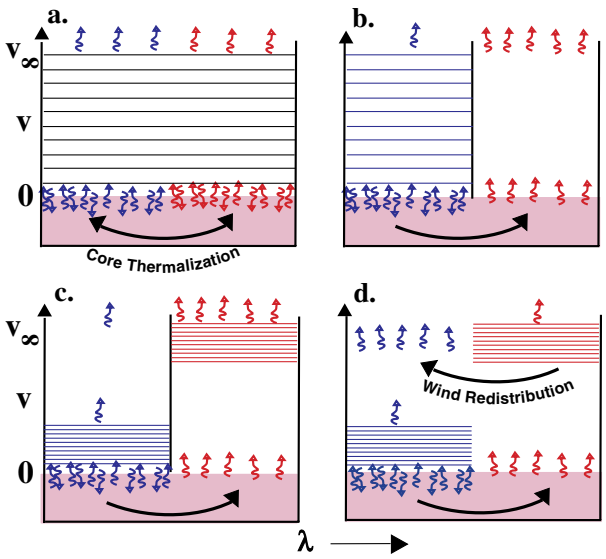


Figure 4. Schematic diagram illustrating the role of line gaps, line bunches and photon thermalization for WR wind momentum deposition. The horizontal lines represent the velocity–frequency spacing of optically thick lines in the wind. The four parts represent (*a*) an effectively gray model, (*b*) a wind with fixed ionization and extensive gaps, (*c*) a wind with ionization stratification that fills gaps and (*d*) the importance of limiting energy redistribution in the wind.

thick lines with a velocity unit frequency separation $\Delta v < v_\infty$. Indeed, if the spectral distribution of lines is 'gray'—i.e. spread randomly through the spectrum without substantial gaps or bunches—then the momentum attained scales simply with the mean thick-line velocity separation as $\eta \approx v_\infty/\Delta v$ (figure 4(*a*)).

However, realistic line lists are not gray, and leakage through gaps in the line spectral distribution tends to limit the effective scattering to $\eta \lesssim 1$ (figure 4(*b*)). However, the optically thick nature of the WR winds gives them a substantial ionization stratification, and this helps to spread line bunches and so fill in gaps, allowing for more effective global trapping of radiation that makes $\eta > 1$ (figure 4(*c*)). Redistribution of photon energy can reduce the local effectiveness of line driving (figure 4(*d*)). In particular, redistribution by photon thermalization near the stellar core makes it especially difficult for radiation alone to initiate the wind. The relative complexity of WR wind initiation may be associated with the extensive turbulent structure inferred from observed variability in WR wind emission lines. Overall, the understanding of WR winds is perhaps best viewed as an 'opacity problem', i.e. identifying the enhanced opacity that can adequately block the radiation flux throughout the wind and thus drive a WR mass loss that is much greater than from OB stars of comparable luminosity.

Wind instability and variability

There is extensive evidence that hot-star winds are not the smooth, steady outflows idealized in the simple

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

CAK theory but rather have extensive structure and variability on a variety of scales. The most direct example is the explicit variability often detected in unsaturated absorption troughs of P-Cygni line profiles. To produce such explicit line profile variability, the associated wind structure must be relatively large scale, covering a substantial fraction of the stellar disk. (By comparison, the emission component of P-Cygni profiles is generally much less variable, since it forms from a more global average of the wind.)

Much of the variability can be characterized as 'discrete absorption components'. These begin as broad absorption enhancements in the inner part of the blue absorption trough of an unsaturated P-Cygni line, which then gradually narrow as they drift, over a period of days, toward the blue edge of the profile. Their occurrence is generally irregular, and their apparent acceleration is quite slow, much slower than expected for an element of the steady wind. They may reflect the slow acceleration of mass ejected by some kind of semirandom disturbance from the underlying stellar surface. The exact nature of this is unknown, but might possibly be linked to either magnetic activity or stellar pulsation.

In a second class of explicit variation are the 'periodic absorption modulations'. These occur at regular, sometimes multiple periods, and include both a modulated increase and decrease in the absorption strength. The periods are generally of order a few days, which is consistent with a low-order multiple of the stellar rotation period. They probably represent rotational modulation of wind structure by an unknown kind of surface irregularity (again perhaps magnetic).

The optical emission lines formed in WR winds also commonly show a low- amplitude variability. Such optical lines have been monitored using ground-based telescopes at much higher signal-to-noise ratio than has been typical for the ultraviolet wind lines monitored from OB stars by orbiting satellites. Analysis suggests a turbulent hierarchy of structure extending down to quite small spatial scales.

There is also indirect evidence that OB winds have an extensive, small-scale, turbulent structure. Saturated P-Cygni lines have extended black troughs thought to be a signature that the wind velocity is highly nonmonotonic. Such stars commonly show soft x-ray emission and sometimes also nonthermal radio emission, both of which are thought to originate from embedded wind shocks.

A promising explanation for this small-scale, turbulent structure is the intrinsically strong instability of line driving to small-scale velocity perturbations. As noted above, there is a strong hidden potential in line scattering to drive wind material with accelerations that greatly exceed the mean outward acceleration. Numerical simulations suggest this can indeed lead naturally to a highly structured flow dominated by multiple shock compressions. However, it is not yet clear whether the simulated flow properties can robustly explain key observational features such as the soft x-ray emission.

Frontiers: colliding winds, disks, etc

The theory for line-driven winds is mature enough that it is now finding application in many more complex circumstances. Let us thus briefly summarize some of these frontiers.

One example is colliding winds of close, massive star binaries. In several cases these involve WR + O stars with separations of only a few O-star radii. Generally the WR wind is so much stronger that a simple hydrodynamic ram balance between the two stars is not possible. This suggests the WR wind should simply overwhelm its companion's outflow and penetrate down to the O-star surface. Orbital phase monitoring of such systems suggests, however, that this does not occur, but that a wind–wind interface is generally kept away from the O-star surface. An attractive explanation for this comes from considering the additional pressure provided by the O-star light in a *radiative braking* of the impacting WR wind. Such a radiative deceleration has many of the same properties of the usual wind acceleration, including a crucial role for the Doppler shift of line resonance in allowing a strong line force. An interesting result is that radiative braking is generally only effective if the interaction of O-star light with WR material is characterized by a stronger line opacity than needed for driving the O-star wind. This represents a new twist to the usual 'opacity problem' of how WR stars drive such strong winds and even suggests that observations of WR + O binaries could help to constrain answers to this fundamental question.

Hot stars generally have a rapid rotation, but among the fastest rotators are the BE STARS, which are characterized by strong Balmer line emission thought to originate in a circumstellar disk. A quite appealing idea is that these form naturally as 'wind-compressed disks' by rotational focusing of the star's radiatively driven stellar wind toward the equatorial plane. Initial dynamical simulations that assumed a simple, radial form for the line force generally supported this idea. Surprisingly, however, subsequent, more realistic models show that *nonradial* components of this line force can actually completely *inhibit* the formation of such a disk by effectively driving material away from the equatorial plane. Further analysis has shown how this, and other peculiar properties of line driving, follows naturally from the characteristic scaling of the line force with the velocity gradient.

Finally, line-driven winds can also originate from luminous accretion disks, such as those found in cataclysmic variables (see also CATACLYSMIC BINARIES: CLASSICAL AND RECURRENT NOVAE). Recent work has thus focused on extending the basic CAK theory to account for key differences from the stellar case, for example the 2D nature of the flow and the peculiar nonmonotonic variation of the effective gravity, which initially increases from the disk plane before eventually falling off with distance. There is even renewed interest in developing such disk wind models to explain the broad-line flows observed from ACTIVE GALACTIC NUCLEI and QUASARS.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The theory of hot-star winds is nearly three decades old, and in this time there has amassed an appropriately large literature.

The basic notion of line driving was first broached in a prescient paper by

Milne E A 1926 *Mon. Not. R. Astron. Soc.* **86** 459

The seminal work laying out the CAK formalism still used today was in

Abbott D 1982 *Astrophys. J.* **263** 723
Castor J I, Abbott D C and Klein R I 1975 *Astrophys. J.* **195** 157

The key Sobolev approximation was developed in

Sobolev V V 1960 *Moving Envelopes of Stars* (Cambridge, MA: Harvard University Press)

The reformulation of the CAK scalings in terms of the Q of line resonances was by

Gayley K G *Astrophys. J.* **454** 410

A key paper testing the validity of the Sobolev approach was also one of the first to relax the CAK point-star approximation:

Pauldrach A, Puls J and Kudritzki R 1986 *Astron. Astrophys.* **164** 86

Wind variability is the subject of several recent conference proceedings:

Kaper L and Fullerton A (eds) 1998 *Cyclical Variability in Stellar Winds* (Dordrecht: Kluwer)
Moffat A, Owocki S, St-Louis N and Fullerton A (eds) 1994 *Instability and Variability in Hot Star Winds* (Dordrecht: Kluwer)
Wolf B, Fullerton A and Stahl O (eds) 1999 *Variable and Non-spherical Stellar Winds in Luminous Hot Stars (Springer Lecture Series in Physics)*

A recent paper on instability, which also addresses the general validity of the Sobolev approach, is

Owocki S P and Puls J 1999 *Astrophys. J.* **510** 355

WR stars have also been the subject of several recent conferences, with proceedings containing several reviews on the dynamics of WR winds both for single stars and in binary systems:

van der Hucht K, Koenigsberger G and Eenens P (eds) 1999 *Wolf–Rayet Phenomena in Massive Stars and Starburst Galaxies (IAU Symposium 143)* (Dordrecht: Kluwer)

Some general reviews on hot-star winds can be found in

Cassinelli J 1979 *Ann. Rev. Astron. Astrophys.* **17** 275
Howarth I (ed) 1997 *Boulder–Munich Workshop II* (San Francisco: Astronomical Society of the Pacific)
Kudritzki R and Hummer D 1990 *Ann. Rev. Astron. Astrophys.* **28** 303
Owocki S 1990 *Rev. Mod. Astron.* **3** 98
Puls J *et al* 1993 *Rev. Mod. Astron.* **6** 271

Finally, an essential general reference is the new introductory book on stellar winds:

Lamers H J G L M and Cassinelli J 1999 *Introduction to Stellar Winds* (Cambridge: Cambridge University Press)

Stanley Owocki

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Emission from Binary Stars

Stellar radio emission is most common in double star systems where each star provides something essential in producing the large amounts of radio radiation needed for it to be detectable by RADIO TELESCOPES. They transfer mass, supply energy or, when one of the stars is a NEUTRON STAR or BLACK HOLE, have the strong gravitational fields needed for the energetic particles and magnetic fields needed for the radio emission process. However, while this is an area of science where we know what has been observed, the underlying causes are not well understood.

Radio emission from stars in general

Radio emission may be detected from either single or binary stars, although the latter are generally the most luminous. Because it is the closest star, the Sun is the brightest source of radio emission that we see produced by natural processes. This is mainly because it is so close. Other stars are 200 thousand times, or more, further away and, because the strength of observed radio emission weakens with the square of the distance, our Sun would appear about 40 billion times weaker if it were at the distance of the nearest star.

Only a small number of nearby single stars, called flare stars (see RED DWARFS/FLARE STARS), emit enough radio emission to be observable with the largest radio telescopes. Detecting the radio emission from stars usually requires physical conditions that produce stronger radio emission than is normal for stars. These extreme conditions occur most often when there are two or more stars in close orbits around each other. The strongest and most interesting radio emission comes when one star in a binary system is in its last stages of its life as a star.

BINARY STARS are common in the universe, but very few of them are detectable as sources of radio emission. This generally occurs under the special conditions where one of the companions either has an unusually high level of surface activity, or is a collapsed star at the end of its life. There are four types of special stars that are important for the strongest radio emission: 'normal' stars with much more surface activity than seen on the Sun; WHITE DWARF STARS, which are special because of their compactness and the sensitivity of their electron degenerate matter; neutron stars, which are special because of the neutron degeneracy of their matter and the stronger gravitational fields due to their compactness; and black holes, which are the most compact and massive 'dead' stars. White dwarfs, neutron stars and black holes are the final stages in the life of stars with masses in the range of 1–1.44 solar masses ($M_{\rm sun}$), (1–3)$M_{\rm sun}$ and $\geq 3\ M_{\rm sun}$, respectively. The extreme conditions of these dying, or dead, stars are important in producing strong radio emission.

Radio emission mechanisms

Radio emission, particularly from radio stars, comes from four basic processes involving some combination of radiating electrons and magnetic fields, with the latter needed for three of the most important processes.

The first process is thermal emission which arises because the temperature of the gas is hot enough so that interactions between electrons and the electrical field of protons in the gas, which must be in a partially or totally ionized plasma state, radiate by so-called braking radiation (bremsstrahlung). This form of radio emission limits the observed flux densities, the total emission that can be collected by a radio telescope, to

$$S_\nu \leq \frac{2kT_{\rm e}}{\lambda^2}\frac{\pi}{4}\left(\frac{D}{d}\right)^2$$

where k is Boltzmann's constant, $T_{\rm e}$ is the temperature of the electrons in the plasma, λ is the wavelength of observation by the telescope, D is the diameter of the source and d is the distance between the source and the observer. The maximum thermal emission occurs when the plasma is optically thick, that is, all radio photons are absorbed and re-emitted one or more times before escaping from the radiating plasma. Because ionized gases have temperatures that are generally either about 10 000 K or in extreme cases a million kelvins, this severely limits their radio emission. The non-thermal processes that have effective radiation temperatures from 10^8 to 10^{12} K are needed for most of the observed radio stars.

The next three processes differ from thermal emission only in that the emission is from electrons moving at a velocity v and interacting with large-scale magnetic field structures rather than the electric fields of protons. These three differ only in the energy of the radiating electrons. The lowest-energy electron–magnetic field radiation process is from electrons with Lorentz factors, $\gamma = 1/\sqrt{1-(v/c)^2}$, which are much less than 1; this is cyclotron or gyroresonance emission. The next highest energy process is gyrosynchrotron radiation where the electrons have $\gamma \sim 1$. The most important, and highest-energy, process, synchrotron emission, requires $\gamma \gg 1$ so the electrons are highly relativistic. Most radio stars are seen only because they have regions which produce highly relativistic electrons, moving very close to the speed of light, which then interact with large-scale magnetic field structures in the binary system or its outflows.

Need for acceleration of relativistic electrons

How nature can so easily produce highly relativistic plasmas is still a mystery that we only partly understand. We know that particle acceleration occurs in strong electric fields and in so-called reconnection events in magnetic field structures. However, the only process we believe can produce the needed highest-energy electrons is acceleration in shocks. Shocks occur whenever plasma or gas from different sources collides or interacts with each other, sometimes simply as internal structures in an extended plasma. The case where the matter ejected from a SUPERNOVA explosion collides with the interstellar medium is one reasonably well-understood process and is believed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

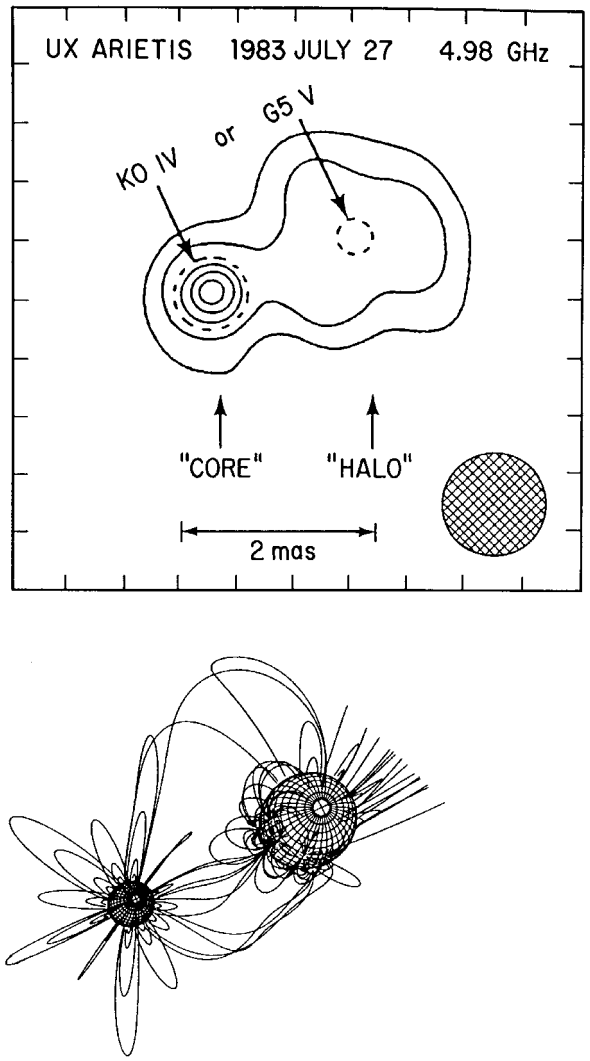


Figure 1. A VLBI image of the radio emission from the active binary UX Ari together with a drawing showing the active regions and magnetic field structures linking the two stars.

to supply most if not all observed COSMIC RAYS. In the cases of stars little is known about the shock interaction regions that produce the needed acceleration to highly relativistic energies. Observing strong radio emission is the main reason for believing that shocks accelerating relativistic electrons do exist in observed radio binaries.

Active binaries and flare stars

Nearby red dwarf flare stars were the first stars reported to have radio emission, although much of the early work is doubted because simultaneous monitoring for man-made radio interferences was not done. It was not until that was done, and interferometers less subject to interference were used in the 1970s, that stellar radio emission was definitely established. The radio emission from dwarf flare stars, called dMe stars because of their small size, M spectral type and characteristic occurrence of strong optical emission lines, is a scaled-up version of solar radio emission. The radiation is mainly from relativistic particles created by magnetic loop reconnection and other non-linear plasma phenomena. Many of the observed flare stars are in double star systems where both systems show radio emission. Active binaries where one star is a relatively cool K star with extreme magnetic surface activity are the greatest source of radio binaries involving normal stars. Figure 1 shows a radio image of the active binary UX Arietis together with a cartoon showing the probable magnetic fields linking the two stars.

It is generally believed, based on statistical evidence of their frequent occurrence, that active binaries are more active than equivalent single stars because one effect of having a companion star is the increase in the rotation rate of the active star. This rotation probably leads to increasing magnetic dynamo effects in convective regions near the outer layers of the stars, and this increases the number and size of large starspots and related magnetic activity leading to radio emission. It is still unclear whether intrastellar plasma phenomena are important in augmenting this activity or whether the regions between the stars are simply regions into which radio-emitting plasma, mainly radiating gyrosynchrotron emission, extends during magnetic activity.

Radio emission from white dwarf binaries

Binaries containing an evolved 'normal star' and a white dwarf produce episodic outbursts of optical, infrared and radio emission when shells of hot, ionized gas are ejected by thermonuclear explosions ocurring on the surface of the white dwarf. This happens when a companion star has expanded to fill its local gravitational (Roche) lobe and this transfers matter to the surface of the white dwarf. This increases the amount of non-degenerate matter on top of the degenerate electron material of the white dwarf. When this layer reaches roughly 1/10 000 of a solar mass, a thermonuclear explosion results and the non-degenerate, built-up layer is ejected at velocities of several hundreds to a few thousands of km s^{-1}. This is the cause of the NOVA, or 'new star' phenomenon, which is first seen by the sudden appearance of a bright, new star usually seen first by amateur astronomers. The nova shell is an ionized gas with temperatures near 10 000 K, and its thermal radio emission becomes observable in a few days or weeks when the size of the shell becomes large enough.

The radio emission from nova shells is first seen as optically thick surface emission, and with time the less dense outer layers become transparent, and more weakly emitting, as the opaque surface slowly moves from the outer part of the shell to the inner boundary of the shell. The resolved structure of the nova shells has been observed with large arrays of radio telescopes such as the Very Large Array in New Mexico and the Merlin array in England, the first being nova QU Vulpecula 1974 and the most recent being nova V1974 Cygni in 1992. These imaging

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observations indicate shell masses higher than expected from theoretical models of the nova explosion process. Bright novae that are easy to study at all wavelengths occur only every several years.

Neutron star binaries and x-ray sources

Among the first stellar systems observed to show radio emission were strong x-ray sources such as SCORPIUS X-1 and CYGNUS X-1, which have these names because in the 1960s they were the strongest x-ray sources in their constellations. Sco X-1 is one of the class of radio-emitting binaries where a neutron star with a strong magnetic field and magnetosphere probably plays a major role in the process; Cyg X-1 was the first stellar system in our galaxy suspected of containing a black hole, and is in the class that we discuss in the next section.

In the early 1970s Sco X-1 would occasionally exhibit radio flares up to levels of 1/4 Jy that lasted a few hours; then it seemed to weaken in its flaring, only to return to 1/4 Jy level flaring in the late 1990s. Sco X-1 belongs to a small class of x-ray sources called Z-sources because of the Z-like shape they exhibit in plots where the ratio of the observed intensities at two high-energy ranges is plotted against the ratio of intensity for two lower-energy ranges. The Z-sources at any one time occupy only one location in their Z-diagram and this identifies the state of the X-ray source. The properties of these sources, including their radio emission, vary in the Z-diagram. Sco X-1 and other Z-sources such as Cygnus X-2, GX17+2 and GX349+2 show the weakest and most steady radio emission when at the lower end of the Z and have stronger and more erratically variable radio sources as the sources move to the other end of the Z.

At the time of this writing we can only speculate that the weak steady emission may be due to jet-like radio emission while the stronger, more erratic emission is related to complex, flare-like phenomena arising from the magnetospheric regions generated by the neutron stars. One could also speculate that some of the unusual radiation processes seen in single neutron stars seen as radio pulsars, may occur in a weaker and non-periodic form in X-RAY BINARIES containing neutron stars.

There are a number of other radio-emitting star systems where a neutron star component is probably involved. One is the periodically flaring system LSI $+61^\circ$ 303, which is an x-ray binary with a period of 29 days, where the compact star is moving inside the strong stellar wind from its companion star with what is probably an elliptical orbit. In this system it is believed that when the compact star is closest to the other star radio flaring is initiated, perhaps because of the strengthening of the bow shock region where the winds from the compact star's ACCRETION DISK interact with the wind of the other star. Circinus X-1 is a similar case where radio flaring is triggered once in each orbit of the compact star around its companion. The neutron star binaries Aquila X-1 and Centaurus X-4 have been shown to exhibit radio flares associated with recurring x-ray outbursts.

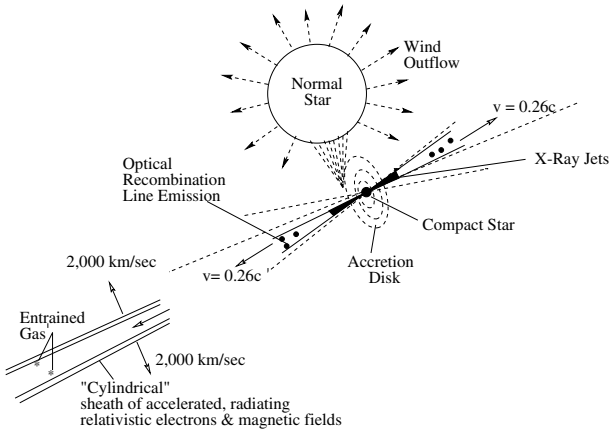


Figure 2. A cartoon showing the x-ray, optical and radio-emitting jets of SS433. They are ejected along the axis of the accretion disk at a velocity of 0.26*c*.

There are two binary systems where the evidence is not clear whether the compact star is a neutron star or a black hole: Cygnus X-3 and SS433. Because black hole binaries seem to show stronger radio events than neutron star binaries, we will discuss Cygnus X-3 and SS433 with the black holes and black hole candidates.

Binary stars with black holes and black hole candidates

In the late 1970s the 433th entry in the Stephenson–Sandeulek catalog of emission line stars, called SS433 (figure 2), was found to show a strong 162.5 day periodicity in the velocities of its optical emission lines which indicated jets of emission with velocities of 26% of the speed of light, or 0.26*c*. That this was due to precession of jets ejected along the rotational axis of the accretion disk was verified when x-ray emission lines and images of radio emission were seen to show changes explainable by a precessing jet. The x-ray and optical lines vary with radial velocity along the line of sight, with a stationary line system, one line system producing emission with redshifts indicating that they are moving away from us and another line system with the blueshifts of material moving toward us. The radio emission is seen in radio images on size scales ranging from a few milliarcseconds up to several arcseconds and exhibits lateral proper motions that provide us with the full three-dimensional picture of the behavior of the 0.26*c* jets. It is still uncertain whether SS433 is a binary with a neutron star or a black hole. Depending on which evidence in emphasized the conclusion can go either way. However, although SS433 seemed to be unique for 15 yr, the phenomenon of highly relativistic radio jets has recently been shown to be associated with some binary systems that are either clearly established or strongly suspected to be binary stars with black hole systems.

The discovery of highly relativistic radio jets in the recurring but transient x-ray binaries GRS 1915 + 105, GRO J1655-40 and XTE J1748-288 in 1994, 1994 and 1998,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

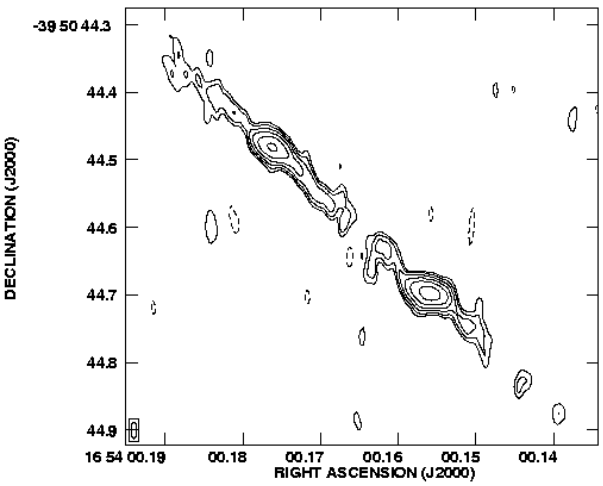


Figure 3. An 18 cm VLBA radio image of GRO J1655-40 made on 18–19 August 1994 showing its jets, which appear superluminal on one side. The brightest region at the lower right is the location of the black hole x-ray binary and the material to the upper left is moving with an apparent lateral velocity of $1.1c$ owing to relativistic effects resulting from its $0.92c$ velocity.

respectively, brought the realization that strong radio flares are often associated with the x-ray events where highly relativistic radio jets are ejected with velocities near the speed of light. These three systems show apparent motions of radio-emitting matter moving towards us that seem to be moving faster than the speed of light. The finiteness of the speed of light causes the radio emission seen at a particular time to be emitted earlier from the far side jet and later from the near side jet such that the apparent velocities are

$$v_{\text{apparent}} = \frac{v}{1 \pm (v/c)\cos(\Theta)}$$

where the + sign is for the receding jet, the − sign is for the aproaching jet and Θ is that angle between the line of jet ejection and the line of sight to an observer. For the three superluminal jet sources, v_{apparent}/c is 1.4, 1.1 and 1.3, respectively. It is because v/c is close to unity in these systems that it is possible for the denominator in this equation to become small, leading to $v_{\text{apparent}} > c$ and producing a so-called superluminal jet source. Superlumiminal jets are commonly present in extragalactic systems such as quasars and active galactic nuclei.

Figure 3 shows the high-resolution radio image of GRO J1655-40 observed with the Very Long Baseline Array (VLBA) on 18–19 August 1994, about a week after this black hole binary ejected a jet with a velocity of $0.92c$. The binary x-ray source is located in the middle of the central bright source in the lower right and the material to the upper left is the approaching jet moving with an apparent velocity of $1.1c$. It has been established that the GRO J1655-40 binary system consists of a 7 solar mass black hole orbiting around a 1.2 solar mass F2 IV star every 2.61 days. Matter from the F2 IV star falls toward the black hole forming an accretion disk whose inner regions are disappearing into the event horizon of the black hole. The x-ray outburst associated with the relativistic jet ejection in early August 1994 was the result of a state change in the black hole's disk of accreting matter when it went from weakly to strongly radiating in x-rays, with the ejection of the observed radio jet along the rotation axis of the accretion disk.

SS433, GRS1915 + 105, GRO J1655-40 and XTE J1748-288 all exhibit another effect due to the relativistic nature of their jets: Doppler boosting, due to the relativistic distortion of the area of the emitting regions as seen by a distant observer, so the apparent ratio of the flux density of the approaching and receding jets is given by

$$\frac{S_{\text{approaching}}}{S_{\text{receding}}} = \left(\frac{(1 + (v/c)\cos(\Theta)}{1 - (v/c)\cos(\Theta)}\right)^{k-\alpha}$$

where α is the spectral index of the radio emission and $k = 2$ or 3 depending on whether the jet is continuous or blob like.

The x-ray binary Cygnus X-3 is the system with the most frequent, strong radio flares, exhibiting an average of two big flares per year since first discovered in 1972. In early 1997 an 11 Jy flare in Cyg X-3 was imaged at very high resolution, showing an entirely one-sided jet with $v \geq 0.8c$ and $\Theta \leq 8^{\circ}$. The Doppler boosting of the Cyg X-3 jets pointing nearly at the Earth is part of the reason why Cyg X-3 flares are so strong, often rivaling the apparent radio brightness of distant QUASARS. These systems with highly relativistic jets are frequently described as microquasars.

Future prospects

The study of radio events in binary stars with one stellar component having magnetic-field-driven surface events, or being a compact white dwarf, neutron star or black hole, is one of the best ways to study relativistic plasma outflows in stars. With X-RAY ASTRONOMY satellites monitoring the sky for x-ray outbursts, astronomers can quickly turn x-ray, radio and infrared or optical telescopes to observe these objects with their accretion disks and outflows that develop synchrotron radio emitting plasmas. Arrays of radio telescopes such as the Very Large Array, the Westerbork Synthesis Radio Telescope, the Australian Telescope Compact Array, the VLBA, the European VLBI Network, the Giant Metre-wave Radio Telescope and other arrays of radio telescopes promise to continue generating images detailing the evolution of these relativistic plasma events. These images, when combined with x-ray, optical and infrared information about the events, will continue to play a major role in the study of the major events in exotic binary star systems.

Bibliography

Hjellming R M 1988 Radio Stars *Galactic and Extragalactic Radio Astronomy* ed G L Verschuur and K I Kellerman (New York: Springer) pp 381–438

Hjellming R M and Han X–H 1995 Radio Properties of X-ray Binaries *X-ray Astronomy* ed W H G Lewin, J

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

van Paradijs and E P J van den Huevel (Cambridge: Cambridge University Press) pp 308–30
Mirabel I F and Rodriguez L F 1999 Sources of Relativistic Jets in Out Galaxies *Ann. Rev. Astron. Astrophys.* **37** 409–43

Robert M Hjellming

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Galaxy

A galaxy that is an intense source of radio-frequency emission. The radio luminosity of a strong radio galaxy (10^{37}–10^{39} watts) can be up to a million times greater than the radio output of a conventional galaxy (typically about 10^{33} watts) and up to a hundred times greater than the optical luminosity of a galaxy such as the Milky Way. The optical counterparts of radio galaxies are usually elliptical galaxies

In a typical radio galaxy, most of the radio emission comes from two 'lobes' (clouds of radio-emitting material), often located beyond and on either side of, the visible galaxy. Typically, the radio-emitting clouds span a region of space five to ten times larger than the optical galaxy. The overall extent of a radio galaxy (measured across its lobes) may be several million light-years; indeed, the lobes of the giant radio galaxy 3C 236 ('3C' denotes the Third Cambridge Catalogue of radio sources) have an overall diameter of nearly 20 million light-years. High-resolution images show that many radio galaxies also possess a compact central radio source from which there emerges a jet, or a pair of oppositely directed jets, of radio-emitting material, pointing outwards towards the distant lobes. The radio-emitting lobes are believed to be clouds of electrically charged particles (predominantly electrons) that have been expelled at some time in the past, from a central 'powerhouse' in the core of the galaxy. The jets appear to be streams of highly energetic electrons that have been accelerated to a substantial fraction of the speed of light. Spots of enhanced intensity are often seen where the outflowing jets of particles plough into the distant lobes.

The jet and lobes radiate synchrotron radiation (a type of radiation that is emitted by charged particles traveling at large fractions of the speed of light in magnetic fields). The spectrum of a radio galaxy is relatively 'flat', displaying the smooth slow decline in intensity from low to high frequencies (from long to short wavelengths) that is typical of a synchrotron source (by contrast, the spectrum of a galaxy like the Milky Way is dominated by the combined light of its constituent stars and, like that of a star, rises steeply at the short-wave end of the spectrum, reaches a peak in the visible and declines in the infrared).

The energy that is needed to produce the lobes and sustain them for tens of millions of years is very substantial; in the region of 10^{53} joules for a strong radio galaxy. The 'energy machine' that is responsible for ejecting the lobes and powering the jets is believed to reside in the core of the galaxy and is widely considered to be a supermassive black hole surrounded by an accretion disk of heated gas.

The first radio galaxy to be identified was Cygnus A, a classic double-lobed strong radio galaxy. The nearest (rather weak) radio galaxy, Centaurus A, lies at a distance of about 15 million light-years.

See also: accretion, active galactic nucleus, active galaxy, astrophysical jets, Centaurus A, Cygnus A, radioastronomy, synchrotron radiation, Virgo A.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Instrumentation

RADIO TELESCOPES generally fall into two categories: single-dish telescopes, where one antenna is used on its own as a telescope, and synthesis arrays or INTERFEROMETERS, where several antennas are combined to synthesize the effect of one very large telescope. Whilst instrumentation on a single-dish radio telescope can often be changed in a similar way to that on an optical telescope, instrumentation on a synthesis array tends to be very tightly integrated into the telescope system. In this article, instrumentation for the two classes of radio telescope is therefore considered separately.

Single dish continuum detectors

The earliest radio telescopes worked in a simple 'single dish' mode, in which the entire signal of the telescope was focused on to a single detector. This detector was typically a low-noise amplifier followed by a detector that simply measured the amount of continuum radio power falling onto it. Much of the early pioneering work of radio-astronomy was performed with this simple arrangement. To make an image of the source, it was necessary to move the entire telescope back and forth in a raster pattern, and thus gradually build up a radio image of the area being surveyed. In practice, this raster pattern was generally composed of a number of points set in a rectangular grid. The telescope would be pointed at each of these grid points in turn, and a total power measured. However, this technique is limited by fluctuations in the receiver power caused either by the atmosphere or by the electronics. A modern variant of this technique is 'on-the-fly' mapping, where the telescope is kept in constant motion. As a result, fluctuations in receiver power due to atmosphere or electronics occur on a longer timescale than the observed changes in power caused by the motion of the telescope beam past a radio source in the sky, and so the two effects can be separated out. Thus, on-the-fly mapping yields more reliable images than the earlier gridded images.

The main limitation of the technique of using a single dish for continuum imaging of the sky is the relatively poor spatial resolution obtainable with a single-dish telescope. However, this simple single-dish technique is still sometimes used for specialist purposes, particularly at millimeter and sub-millimeter wavelengths where receivers tend to be too complex to build in multiple packages (see RADIO TELESCOPE DATA PATH FOR SINGLE DISH AND INTERFEROMETERS).

Modern radio-astronomy RECEIVERS come in two varieties. The most common is a heterodyne receiver, in which the incoming signal is successively amplified and mixed with local oscillators to produce a coherent signal at low frequency which can then be processed in 'back-end' instrumentation. At millimeter and sub-millimeter wavelengths, the receiving element is sometimes a bolometer, which is a sensitive device that measures the instantaneous incoherent radio power falling on the detector.

Spectroscopy

In the early days of radio-astronomy, measurements of the radio spectrum of a source were made by using a set of filters, each tuned to a slightly different frequency. Such an instrument was used in many early pioneering studies of the neutral hydrogen line. However, it suffered both from instability in the electronics and from a relatively limited frequency resolution.

In the 1970s, Fourier transform or autocorrelation spectrometers (or 'CORRELATORS') were developed. In these instruments, a signal was successively delayed electronically in a shift register, and then the delayed samples were then cross-correlated with the original signal. The resulting autocorrelation function was then Fourier transformed to produce a spectrum of the signal. This technique was not possible until the advent of fast digital computers to perform the Fourier transform, but once established it quickly replaced the older filter banks. Its advantages were much higher stability and flexibility, and a scalability which allowed very high frequency resolution. In recent times, correlators with up to one million channels have been produced for specialist projects such as that of the SETI INSTITUTE.

Until recently, these autocorrelation spectrometers still suffered from a relatively low total bandwidth, which was limited by the speed of the available electronic components. This bandwidth was generally too low for millimeter work, and so a different type of spectrometer was born. This was the acousto-optic spectrometer (or AOS) which was widely used at millimeter and submillimeter telescopes. In this type of spectrometer, the radio signal is applied to a Kerr cell, causing a periodic modulation of opacity across the Kerr cell. A laser beam applied to this Kerr cell is diffracted by the resulting grating, and the intensity distribution of the diffracted laser beam is effectively the Fourier transform of the input signal, and thus gives the spectrum of the input signal. Such a spectrometer can cover a very wide bandwidth (up to several GHz), but suffers from a high sensitivity to thermal instability, so that these spectrometers need elaborate calibration schemes to be usable for radio-astronomy. They have recently started to be supplanted by autocorrelation spectrometers, since the advent of high-speed samplers has meant that autocorrelation spectrometers can now operate up to bandwidths of several GHz.

Pulsar timing

Specialist receivers are used to search for and study PULSARS. These receivers differ from conventional ones both in their ability to combat interstellar DISPERSION and in their ability to record data in 'bins' of short time duration (see PULSAR BACKENDS).

In pulsar searches, it is necessary to record data at a very high rate, as the data need to be sampled at millisecond intervals with a wide bandwidth. This typically results in tens of gigabytes of data per day of observation. These data are then analyzed offline using

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

powerful but computationally expensive algorithms that search for periodicities in the data (see SOFTWARE: RADIO PULSAR SOFTWARE).

For studying the characteristics and timing of known pulsars, the data rates can be much lower as the data can be synchronously added into a number of 'bins', each covering a small (millisecond) time interval across the pulsar period. The data may then be integrated over a period of minutes, and at the end of this integration only an amplitude for each bin need be recorded.

In all studies of pulsars, allowance must be made for the dispersion of the pulsar signal as it passes through the interstellar medium. This dispersion occurs because high-frequency radio waves travel faster through an ionized medium than low-frequency waves. So, a sharp pulse emitted by the pulsar will be blurred out over time, and if no correction is made for this effect, the pulse recorded back on Earth using a wide bandwidth will be weakened and broadened. To correct for this, 'de-dispersers' are used, in which the received band is broken into a number of small frequency channels using either a filter bank or a correlator, in the same way as for spectroscopy, as described above. The signals in the different channels are electronically delayed by an appropriate amount such that the pulses in all channels are once more aligned in time. The data from all the channels can then be integrated together. Recently, the increasing power of high-speed electronics and computers has enabled an additional technique to be used, in which the raw signal from the telescope is recorded as a voltage at very high speed (at hundreds of millions of samples per second). Then all these corrections to the data can be made off-line on high-speed computers at the same time as other processing.

Multibeam and array feeds

Traditional single-dish radio-astronomy typically uses just one or two receiving elements at the focus of the telescope. Increasingly, there is pressure to mount an array of receiving elements at the focus. This evolution is similar to the replacement of photomultipliers on optical telescopes by ccd arrays. However, the challenge is harder at radio wavelengths because the physical size and the complexity of the receivers makes it difficult to mount more than one receiver at the focus simultaneously. However, a feed array with n elements increases the power of the telescope by a factor of n, and so the pressure to develop such multibeam receivers is huge. At millimeter wavelengths, the physical size of the receivers does not present such a problem as those operating at longer wavelengths, and so the first experimental multibeam receivers, developed at several observatories, typically had a few elements and operated at millimeter wavelengths.

However, the first multibeam receiver to make a major impact on radio-astronomy was the seven-beam receiver developed by NRAO and successfully deployed first on the GREENBANK 70 m telescope in the USA and then on the PARKES OBSERVATORY 64 m telescope in Australia. It was used to make a very successful continuum survey of the entire sky at a wavelength of 6 cm.

In 1997, a 13-beam receiver operating at 21 cm was mounted on the Parkes 64 m telescope, representing a milestone in radio-astronomy instrumentation. This receiver, which required a major upgrade of the telescope in order to accommodate its enormous size, was used for both spectroscopic surveys of neutral hydrogen and for pulsar searches.

At millimeter wavelengths, development of array feeds is continuing, using both heterodyne receivers and bolometers. A major advance was signalled by the development of SCUBA (Submillimeter Common-User Bolometer Array) which was installed on the James Clerk Maxwell Telescope in 1996. SCUBA consists of two arrays of bolometers, containing 37 and 91 elements respectively. Plans are already being made for similar receivers with an even larger number of elements.

Another type of array feed which is in its early stages of development is the focal plane array in which the focal plane is sampled sufficiently densely to be able to measure the electric field across the focal plane. The signals from the different elements are cross-correlated, and the Fourier transform of the cross-correlation function yields a holographic image of the electric field distribution across the telescope. Corrections can then be made to this image for effects such as telescope deformation, pointing errors and atmospheric effects. In principle, this technique can yield a near-perfect image of the sky, regardless of the quality of the telescope or the atmosphere.

Synthesis arrays and interferometry

All the types of instrumentation fitted to single-dish radio telescopes, with the exception of the bolometer, can also be fitted to radio synthesis arrays or interferometers. In practice, because each instrument generally has to be duplicated many times over, synthesis arrays usually have far less flexibility than single-dish radio telescopes, and so the range of instrumentation tends to be smaller.

Modern synthesis arrays all use correlators to combine the data from their telescopes, and these correlators follow the same principles as those correlators used for spectroscopy in single-dish radio-astronomy. Thus, the two functions are relatively easily combined, so that a synthesis array can generally yield images in each of many frequency channels across the observing band. In practice, the size of the correlator tends to be limited, and so older instruments such as the VERY LARGE ARRAY tend to have only a modest spectroscopic capability, whereas more recent instruments tend to have much higher spectroscopic resolution. Similarly, the correlators on synthesis arrays can be designed to include 'binning' for pulsars, and several instruments offer this capability for pulsar astronomy.

In very long baseline interferometry (VLBI), the same principles apply, and most VLBI correlators offer both a spectroscopic mode and a pulsar binning mode.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The future

Instrumentation on radio telescopes is changing rapidly on several fronts. Increasingly, the distinction between on-line instrumentation and off-line processing is becoming blurred.

For example, the increasing speed and power of general-purpose computers means that some functions that once had to be performed in hardware can now be done in software. One example, mentioned above, is the class of pulsar search and de-dispersion techniques, which is now within the reach of general-purpose computers.

Another aspect of this blurring is the move towards real-time processing. At present, data from a synthesis array are typically collected over a period of hours, then written to tape or CD-ROM, and then taken back to a user's institution for analysis. However, the amount of processing time needed is now typically smaller than the observing time, and so this processing is starting to be performed in real-time on computers attached to the on-line system.

Another change that is rapidly occurring is the increase in radio-interference contamination of radio-astronomy signals. There are two causes of this. One is the rapidly growing rise in interference that occurs within supposedly protected radio-astronomy bands. For example, the designated 18 cm radio-astronomy bands are becoming very badly contaminated by out-of-band emissions from constellations of low-Earth orbiting satellites. The other cause of the problem is the need for radio-astronomers to receive radiation from a greater fraction of the available spectrum than that officially designated as a radio-astronomy protected band. For example, technology now allows us to study the neutral hydrogen emission from GALAXIES AT HIGH REDSHIFT. The redshifted hydrogen lines fall well outside the band protected for radio-astronomers.

To enable radio-astronomy to continue in the face of this interference, it is clear that techniques need to be developed to excise or cancel radio interference from the radio-astronomy signal received by a telescope. Such radio interference mitigation techniques are in their infancy, but will become increasingly important as the next-generation radio telescope, the Square Kilometer Array (SKA), is developed.

R Norris

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Interferometer

A device that utilizes interference effects between radio waves in order to obtain higher angular resolutions than those achievable with individual radio telescopes.

The simplest form of radio interferometer consists of two antennae or dishes that point in the same direction but are separated by a distance (the 'baseline') that is large compared with the wavelength of the radiation that is being observed. Depending on the angle between the direction of the source and the direction of the baseline, wavefronts arriving at the two dishes from a given radio source may be in phase (a wavecrest arriving simultaneously at each dish) or out of phase. If the signals received by the two dishes are combined, they will interfere with each other. In particular, if they are exactly in phase, the two waves will add together to produce a signal of greater amplitude (constructive interference), whereas if they are 180° out of phase (a 'trough' arriving at one dish when a 'crest' arrives at the other), they will cancel to produce a signal of zero amplitude (destructive interference). Constructive interference occurs whenever the angle (θ) between the direction of the source and the normal to the baseline is such that the extra distance (d) traveled by a wavefront between arriving at one dish and the next is equal to a whole number of wavelengths. The condition for constructive interference is given by $\sin\theta = n\lambda/D$ where n is an integer, D is the separation between the dishes and λ is the wavelength of the radiation.

As a point source moves across the sky, and its direction relative to the baseline changes, the signals will shift in and out of phase with each other, so producing an interference pattern. Two neighboring point sources will produce overlapping interference patterns. Analysis of the interference pattern produced by the system enables the position of a point source to be determined and neighboring point sources to be resolved. Since an extended source can be considered to consist of a set of point sources, an interferometer can be used to resolve an extended source in detail. However, this resolution is achieved only along a direction parallel to the baseline. The resolving power of a two-element radio interferometer of this kind is given by $\lambda/2D$ (in radians), or $1.03 \times 105\lambda/D$ (in arcsec). Thus, for example, an interferometer consisting of two dishes separated by 1000 m and operating at a wavelength of 21 cm (0.21 m) would have a resolution of $0.21 \times 1.03 \times 105/103 \approx 20$ arcsec.

In more sophisticated interferometers, the separations between the dishes can be varied and more than two dishes may be included. Very-long baseline interferometry (VLBI) utilizes radio dishes separated by very large distances. The baselines in such systems can be thousands of kilometers long and, in principle, can be almost as great as the diameter of the Earth. The signals received by the two (or more) dishes are stored along with precise timing information and combined at a later date to produce interference patterns and to achieve resolutions that can approach 0.001 arcsec at centimeter wavelengths and 100 arcsec at millimeter wavelengths. VLBI using space-based instruments will improve resolutions even further.

See also: aperture synthesis, radio telescope, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Sources

Extragalactic radio sources can be divided into two groups. One group consists of the star-forming galaxies—most optically bright galaxies, including STARBURST GALAXIES—which produce radio emission at a fairly low rate, about 10^{37} to 10^{40} erg s^{-1}. The radio emission from these is associated with by-products of stellar activity, such as particle acceleration caused by SUPERNOVAE, and is limited in extent to the size of the star-forming regions in the optical host galaxy. The second, much more powerful, group includes RADIO GALAXIES, radio-loud QUASARS, SEYFERT GALAXIES, and other normal galaxies with active nuclei. These galaxies, often collectively referred to as AGNs (ACTIVE GALACTIC NUCLEI), vary by a factor of more than 10^6 in their emitted power, reaching up to 10^{45} erg s^{-1} in the radio. In both star-forming galaxies and AGNs the radio emission is non-thermal and is almost certainly synchrotron emission from relativistic electrons moving in a magnetic field. How these various groups of galaxies are related, and what causes the huge variation in their levels of activity, remain major questions for astronomy.

Observations of H I in the COMA CLUSTER in the early 1980s revealed that for many normal (star-forming) galaxies there is a strong correlation between their radio continuum (RC) emission and far infrared emission (FIR). Since the radio emission is related to high-energy particles and magnetic fields, while the FIR is produced by warm dust, the physical basis for this correlation is not direct, and the tightness of the correlation is surprising: it has not yet been fully explained. Young massive stars appear to play an important role, presumably as a common energy source. The relationship is quite tight, holds for over five orders of magnitude in luminosity, and covers sources as diverse as blue compact DWARF GALAXIES, normal star-forming galaxies, Seyferts and radio-quiet quasars. Since the K-correction is quite different for the FIR emission from the dust (varying as λ^{-3}) and the radio continuum (varying as $\lambda^{0.8}$), this correlation makes it possible to get distances of starburst galaxies in the absence of optical REDSHIFTS.

Radio galaxies and radio-loud quasars have much more powerful radio emmisions, and a FIR-to-RC ratio that is about two orders of magnitude lower than that for normal star-forming galaxies, and with much greater scatter.

From the radio data alone, radio galaxies and quasars are indistinguishable. Both have extremely compact cores (nuclei) less than 1 pc in extent. Quasars differ from radio galaxies only by the extremely bright optical emission from their nuclei. In most cases this 'quasi-stellar' light makes it difficult to detect the fainter emission from the host galaxy, although such detections have now been made for many quasars using the Hubble Space Telescope. In both radio galaxies and quasars, thin jets emerge from the compact core and feed into diffuse radio lobes.

The first radio galaxy to be discovered, CYGNUS A, has just this classic form. This object, the strongest extragalactic radio source, appears in one of the earliest radio surveys, GROTE REBER's 1944 map. It was found again in subsequent surveys in 1948 and the source was double-lobed and more than a minute of arc in size: clearly, unusual astrophysics was at work here. The object was optically identified in 1954. Other radio sources—Centaurus A (NGC 5128), Virgo A (M87) and Fornax A (NGC1316)—had been identified with galaxies using the Dover Heights interferometer, and so it was clear that they were extragalactic. Cygnus A was found to be distant ($z = 0.06$), but still so intense that it would be in the 3C catalogue even if it were 20 times further away, and readily detectable anywhere in the universe. Hence, radio galaxies offered the first possibility of observing well beyond $z = 1$. At this time, the most distant objects known were at $z \ll 0.2$.

Early counts of these radio sources showed that they had to be a strongly evolving population: this effectively ruled out the then popular Steady State cosmology and provided early evidence for an evolving 'history' universe. But ironically, because of this strong evolution, radio galaxies did not turn out to be the powerful new tool for cosmology that astronomers hoped for. Until very recently, all radio source surveys have been dominated by a shell of sources at very great distance, picking up very few sources in the local universe, because the effects of cosmic evolution have totally outweighed those of geometry: more distant, and hence younger, sources have power great enough to completely offset the effects of the inverse-square law. Only the most recent surveys, those with enough sensitivity to detect sources below 1 mJy, have been able to detect fainter radio sources where the population starts to be dominated by the normal star-forming galaxies. These can be observed in large numbers at all distances, and so the study of distant radio sources has returned to the mainstream of extragalactic astronomy.

Fifty years after extragalactic radio sources were discovered we are still seeking to understand their details, but we do have a broad understanding of their energetics and evolution. After radio galaxies were discovered, the first questions that arose were: 'What supplies their energy?', 'Why are they double-lobed?' and 'What was it that made some galaxies active?' Some of these questions have been answered: we now understand that the energy supply is gravitational, and have coarse models to explain the double-lobed structure. But today we still need to understand: which galaxies have 'central engines'; the detailed relationship between the various classes of active galaxies and quasars; issues related to the speed, structure, collimation and symmetry of jets; which came first, AGNs or the black holes they are thought to house?; and how radio galaxies interact with their environment.

Ron Ekers and Helen Sim

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Telescope

An instrument used to collect and measure radio waves from cosmic sources. In essence, a radio telescope consists of a collector of radiation, a detector, an amplifier (to enhance the weak cosmic signals) and a device to store, monitor and display the output from the system. Although the most familiar type of radio telescopes uses a concave dish to collect radiation, radio telescopes may take other forms, for example arrays of aerials or antennae arranged in rows or grids.

The archetypal collector is a parabolic dish that, like the primary mirror of an optical telescope, reflects incoming radiation to a focus. Radio waves arriving at the focal plane set up an electrical disturbance in an aerial, or antenna, that is fed to a receiver, tuned to the frequency of interest, which amplifies the weak incoming signal by factors of up to around 10^{15}. The output from the receiver may then be displayed directly by an output device, such as a chart recorder, or stored and processed in digital form by a computer. An image of an extended radio source may be produced by scanning the dish to and fro to record the intensity of the signal originating at different points on the source, storing the data and then constructing a map in which intensity values are plotted as contour lines or are color-coded to make up a false-color image.

Most dishes are steerable, and can be pointed to any location on the sky, but some are fixed. The largest steerable dish is the 100 m instrument at Effelsberg, near Bonn, Germany. The largest fixed dish is the 300 m (1000 foot) instrument that is built into a natural hollow at Arecibo, Puerto Rico.

As with an optical reflector, the resolving power of a radio dish is given by $1.22\lambda/D$, where λ is the wavelength of the radiation and D the aperture of the instrument. Because the wavelengths studied by radio telescopes are orders of magnitudes longer than optical wavelengths, the resolving power of a radio telescope is orders of magnitude poorer than that of an optical telescope of the same aperture. For example, the resolving power of a radio dish of aperture 100 m, operating at a wavelength of 0.5 m (about one million times longer than visible wavelengths) would be about 20 arcminutes (a third of a degree), much poorer than that of the human eye. The theoretical resolution of a 1 m aperture optical telescope is about 0.1 arcsec. In order to achieve this resolution, a radio telescope operating at a wavelength of 0.5 m would require an aperture of 1000 km; even at the shorter wavelength of 1 cm, an aperture of 20 km would be required. Clearly, this cannot be achieved with a single dish. Much improved resolutions can, however, be achieved by radio interferometers and aperture synthesis systems, instruments that combine the signals received by pairs or sets of dishes.

Because the wavelengths of radio waves are so much longer than those of visible light, radio dishes do not require to be shaped with the same degree of precision as the surfaces of optical mirrors. For example, irregularities in the reflecting surface of a centimeter, correspond to deviations of less than one-twentieth of the wavelength (a good level of precision for a reflecting surface) for an instrument that is studying emissions by clouds of hydrogen at a wavelength of 21 cm. At shorter wavelengths (centimeter or millimeter waves), the tolerances are proportionately finer. At longer radio wavelengths (above, say, 20 cm), a wire mesh surface rather than a solid metal skin can achieve good reflection. This lightens the dish considerably, and reduced the effect of the wind.

See also: aperture synthesis, electromagnetic spectrum, flux density, radar astronomy, radioastronomy, radio interferometer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radio Telescope Data Path for Single Dish and Interferometers

Radioastronomers derive all their observational data from the ELECTROMAGNETIC RADIATION which arrives at the RADIO TELESCOPE from the region of interest. The subset of parameter space of interest to the astronomer is modest: what part of the spectrum; what polarization; which direction; when and where the observations are made. The telescope is the filter which determines the actual subset, and care is required in order to ensure that the filter's characteristics are well understood.

This overview outlines the path from the raw data to the end product in the astronomer's hands.

The data

The radiation to be studied is generated by random natural processes, which leads to an incoherent mix of frequencies, in the form of Gaussian noise with statistical properties which are stationary on time scales which are usefully long—minutes to hours.

The incident wavefront can be described in terms of its *flux density* (power/unit area/unit frequency interval)[1]; Stokes parameters provide a convenient framework for describing the polarization characteristics. We also need to account for the directional properties if the radiation is from an extended region of space.

Single dish

This section examines the data flow in a single radio telescope, whether operating in a stand-alone mode ('single dish') or as an element of an array.

Figure 1 locates the major components of the data path.

The telescope optics

At the top of the chain we impose an angular distribution filter.

1. The reflecting surfaces. We want the telescope to be as sensitive as possible to the radiation from the target. To that end we build a structure with as large a collecting area as possible. In general the design is based on a paraboloid with a prime focus or with a secondary cassegrain or gregorian subreflector.

2. The feed antenna. In every telescope there is the problem of converting the energy in the focal plane to a useful form. Optical telescopes are able to exploit simultaneously a sizeable region of the focal plane, with the silver grains of a photographic emulsion or the active elements of a CCD array. Most radio telescopes are single-pixel devices, with just one active element in the focal plane. The interface device is the *feed antenna*, able to convert the converging ('focused') wavefront to an electrical signal suitable for the downstream electronics.

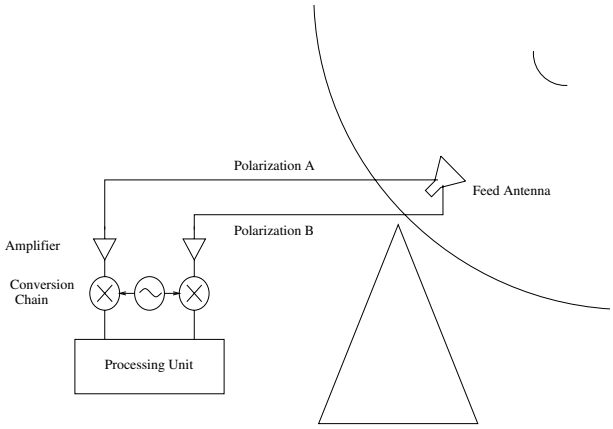


Figure 1. A generic radio telescope, identifying the main components in the data path.

The feed antenna needs to be tailored to the telescope optics to optimize the efficiency. In electrical terms this is a *matching* problem. The feed antenna is optimized for a particular electrical field over its aperture; therefore the telescope focussing must endeavour to produce that optimal field if the combination is to reach its peak efficiency.

3. The radiation pattern. There are several related ways to characterize the antenna's performance: effective aperture, efficiency, radiation pattern.

A wavefront from a distant target can be described in terms of its *flux density S*. An antenna with its optical axis normal to this wavefront will produce a signal of power density P (W Hz^{-1}) at the terminals of the feed antenna. The ratio P/S is called the effective aperture, and is related to the physical aperture by the efficiency:

$$A_{\text{eff}} = \eta A_{\text{physical}}.$$

The telescope does respond at some level to all wavefront directions; this is often expressed in terms of the peak effective aperture and the *radiation pattern*, $G(\theta)$, the effective aperture function normalized to the peak value.

In the general case of a telescope observing an extended source distribution the output at the feed antenna's terminals will be a convolution of the source distribution of wavefronts and the radiation pattern:

$$T(\theta_0) = \int T_{\text{source}}(\theta) G(\theta - \theta_0)\, \mathrm{d}(\theta).$$

$G(\theta)$ is a slow function of frequency—modern antennas, for example, often take the form

$$G(\theta) = 2\frac{J_1(U)}{U}$$

where J_1 is the first-order Bessel function, $U = 2\pi R \sin(\theta)$ and R is the aperture radius in wavelengths.

[1] The radioastronomer's unit of flux density is the Jansky (after the father of RADIO ASTRONOMY), and is : 1 Jy $= 10^{-26}$ W m^{-2} Hz^{-1}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

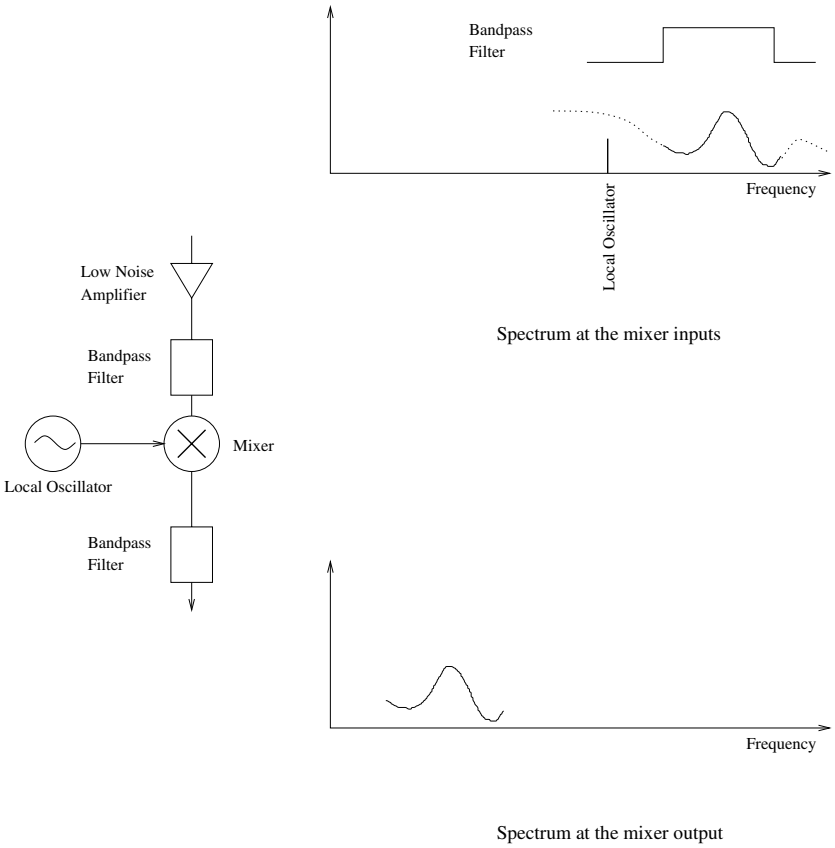


Figure 2. A generic mixing stage.

The polarization splitter

The polarization filter is the next stage—generally a device built into the feed antenna. The incoming wavefront is now transformed to two electrical waveforms, $V_A(f,t)$ and $V_B(f,t)$. The standard alternatives are orthogonal linear or circular polarizations. The critical point to note is that polarization information is encoded in the cross-correlations:

$$V_{AB}(t) = \langle V_A . V_B \rangle = \frac{1}{T} \int_{t-T/2}^{t+T/2} V_A(t).V_B(t)\,\mathrm{d}t.$$

Since the cross-correlations are only computed at the end of the conversion chain, it is important to ensure that the two waveforms are treated in identical ways, and that the relative phasing is carefully controlled. This has implications for the local oscillator machinery.

The discussion in the sections below will describe the treatment of a single electrical signal as it progresses down the conversion chain. A polarization-sensitive system will have two such chains, constructed to be as similar as possible.

The low noise amplifier

The signal at the output of the feed antenna is weak; it would be $P = 5 \times 10^{-24}$ W Hz^{-1} for a 1 Jy source observed by a 100% efficient 25 m diameter antenna; this is equal to the signal from a matched resistor at a temperature of 0.35K.

This poses a serious problem for the signal processing machinery which needs to detect this signal in the face of competing sources of noise.

The first line of defense is to provide amplification immediately after the feed antenna. Modern telescopes use low-noise amplifiers, with cooled electronics (10 K units are now in widespread use). The total noise competing against the signal will be somewhat higher than the raw temperature of the amplifier, as there will be contributions from lossy elements ahead of the amplifier, as well as contributions from the atmosphere, the universal cosmic mirowave background and the ground pick-up from the antenna's environment.

In a well designed system the signal-to-noise characteristics are established at this stage. The astronomical signal can be described in terms of T_A, the antenna temperature; this is the temperature of a hypothetical matched resistor which if placed at the terminals of the receiver would produce the same power (within the specified frequency range) as does the astronomical source.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

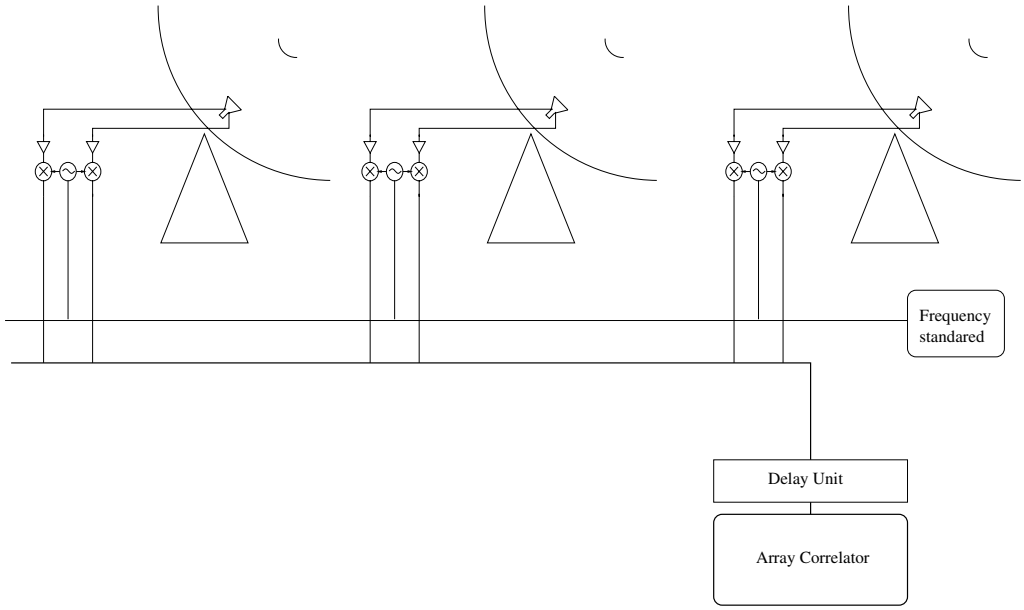


Figure 3. A generic interferometer array.

Competing against T_A is the system temperature ($T_{\rm sys}$) which describes in a similar way the noise generated by the system.

Modern systems operating in the centimeter wavelength band now achieve $T_{\rm sys}$ in the 20–50 K range.

The conversion chain

The feed antenna and the low-noise amplifier operate at the astronomer's chosen sky frequency, while the final stages of the signal processing operate at standard, low frequencies. This means that some form of frequency translation will be required; the conversion chain will isolate a section of the spectrum, a band of width B centered on $f_{\rm sky}$, and translate it to a band B centered on $f_{\rm base}$. In the process it must avoid introducing aliased artifacts. We need also to take care that the spectrum is not distorted.

The process, termed *mixing*, amounts to multiplying the input data stream by a reference signal from a *local oscillator* (LO). A generic mixing stage is shown in figure 2.

The mixing process is best understood in terms of a *Fourier series* description of the data. Given a stretch of data, T seconds long, then the voltage can be rephrased as

$$v(t) = \sum C_n \cos(2\pi f_n t + \phi_n).$$

The coefficient C_n is an estimate of the spectral power density at frequency f_n; the phase terms ϕ_n are uniformly distributed over the interval 0 to 2π.

The mixer performs the operation

$$v_{\rm mix}(t) = v(t)\cos(2\pi f_{\rm lo} t).$$

There will be a sum and a difference product for every frequency component in the input signal. Two filtering operations are therefore required: on the input side we need to ensure that all the frequencies are on just one side of $f_{\rm lo}$ if we wish to avoid a folding of the spectrum; and on the output side we need to separate the 'sum' components from the 'difference' components.

A single LO, split between the two polarization signals will ensure that the polarization properties (the cross-correlations) are not compromised.

The processing unit

A sampler (an analog-to-digital device) is almost inevitable as it is probable that the data will be processed by a computer. The sampler must operate at least at the Nyquist rate, $2B$, where B is the bandwidth of the signal if spectral information is not to be lost.

A CORRELATOR is increasingly the workhorse at this point in the chain:

- in the autocorrelator mode it will produce the spectral distribution.
- the polarization (Stokes) parameters can be derived from the four correlation products of the two polarization signals ($C_{AA}, C_{AB}, C_{BA}, C_{BB}$).

Interferometer array

The INTERFEROMETER array is the radioastronomer's imaging device.

Figure 3 shows the additional components needed to combine a number of radio telescopes into an array.

The subunits of the array are single dish telescopes. The array operation requires the telescopes to operate coherently; the hardware modifications to achieve this are:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(a) The conversion chains must be phase coherent; this can be achieved by driving all the local oscillators from a common frequency standard[2].
(b) All the data streams are returned to a common processing unit, the array correlator.
(c) The imaging algorithms require the data to be transformed to the coordinate frame which has its z-axis oriented towards the center of the field of interest. Let $\hat{\varepsilon}$ be the unit vector towards the field center, and $\vec{r}_{AB}$ be the baseline vector for the antenna pair (A,B); then the data from antenna (A) will need to be delayed by an amount $\tau = (\hat{\varepsilon} \cdot \vec{r}_{AB})/c$ in order that the data streams remain aligned on common wavefronts. Since $(\hat{\varepsilon} \cdot \vec{r}_{AB})/c$ changes with time (as the Earth rotates) τ will need to be adjusted continuously.

An additional (dynamic) phase correction must be made at the local oscillator if, as is usually the case, the delays are applied at baseband (at the end of the conversion chain).

A computer is obviously part of the data chain; this is particularly so when imaging with an interferometer array. Imaging is possible because the array exploits the Van Cittert–Zernicke theorem. It samples the complex coherence function of the extended target. A number of operations are required before the image is delivered to the astronomer:

(a) The data collected in time-order need to be sorted so that the 2D coherence function can be estimated.
(b) The coherence function is Fourier transformed to the image domain.
(c) Calibration and data quality checks will be needed.
(d) The coherence function is rarely fully sampled—the location of the sampling is set by the antenna locations; this means that the raw image contains signatures ('artefacts') reflecting the sampling function. Procedures, some based on *maximum entropy* algorithms, are available to account for these signatures and provide a usable image.

Bibliography

Thompson A R, Moran J M and Swenson G W 1986 *Interferometry and Synthesis in Radio Astronomy* (New York: Wiley)

M Kesteven

[2] Intercontinental interferometry (VLBI) uses independent standards (e.g. masers) of sufficient stability that the local oscillators are phase coherent over adequate observing periods.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radioactivity

The spontaneous decay of certain unstable atomic nuclei through the emission of alpha particles (helium nuclei), beta particles (electrons) or gamma radiation. This process, otherwise known as radioactive decay, changes ('transmutes') the original ('parent') atomic nucleus into a ('daughter') nucleus of another element. If the resulting nucleus is itself unstable, further radioactive decay will ensue until, after a series of stages, the original nucleus has been transmuted into a stable nucleus and radioactivity ceases.

For a nucleus of element X, with atomic number Z (equal to the number of protons in the nucleus) and mass number A (equal to the total number of protons and neutrons in the nucleus), which is denoted by ${}^{A}_{Z}\mathrm{X}$, alpha emission (the emission of a helium nucleus, ${}^{4}_{2}\mathrm{He}$, which contains two protons and two neutrons) reduces A by 4 and Z by 2. Beta emission, the most common form of radioactivity, usually results from the spontaneous decay, within the nucleus, of one of its constituent neutrons into a proton, an electron and an antineutrino. This process (beta decay) increases the atomic number by 1 but leaves the mass number unchanged (the number of protons increases by one and the number of neutrons decreases by one) and is accompanied by the emission of an electron and an antineutrino. With either type of decay, if the resulting nucleus is formed in an excited state it will 'de-excite' (drop to a lower energy level) by emitting gamma radiation.

The time taken for half the nuclei in a sample of radioactive material to decay is known as its half-life, each radioactive nuclide (nucleus of a particular isotope) having its own characteristic half-life. The half-lives of the various radioactive nuclides range from fractions of a microsecond to billions of years. For example, ${}^{238}_{92}\mathrm{U}$, by far the most abundant of the 14 different isotopes of the element uranium, undergoes alpha decay with a half-life of 4.5 billion years. The decay process continues via a series of much faster decays, terminating eventually in the production of the stable isotope of lead, ${}^{206}_{82}\mathrm{Pb}$. Radioactive decay can be used to estimate the time that has elapsed since a sample of material (e.g. rock) has solidified. For example, if a sample of lead, formed by the radioactive decay of uranium, is analyzed, and the relative proportions of the remaining quantity of the parent uranium and daughter lead compared, then, knowing the half-life of uranium, the age of the sample can be calculated. The ages of the Earth, of meteorites and of samples of lunar rock have been calculated by this technique. Younger materials can be dated by means of shorter-lived radioactive isotopes such as carbon-14 (${}^{14}_{6}\mathrm{C}$), which decays through beta emission with a half-life of 5570 years.

See also: alpha particle, atom, electron, isotope, gamma rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radioastron (Spectrum-R)

Russian–US radio astronomy satellite which may launch in 2000–2001. Designed as part of a very long baseline interferometry programme which involves linking the satellite with ground receivers. The satellite carries a 10 metre antenna and will observe from a highly elliptical orbit of $4000 \times 77\,000$ km. Intended to study compact radio sources such as quasars and galactic nuclei at extremely high resolution (3×10^{-5} arcseconds).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radioastronomical Institute, Ukraine

Located in Chervonopraporna near Kharkov. The institute aims its research efforts at: developing the theory and design principles of radio telescopes; ground-based radioastronomical observations at decimeter and millimeter wavelengths; investigations of the Earth and other planets, terrestrial environment and interplanetary space by radioastronomy methods.

For further information see http://www.nas.gov.ua/d/d4/nas_d4i9.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Radioastronomy

Mankind has been fascinated by what one can see in the heavens for thousands of years. The human eye is sensitive to a narrow band of the electromagnetic spectrum from about 4000 to 6600 Å wavelength. Astronomical objects, however, emit electromagnetic radiation across most of the spectrum from meter wavelengths in the radio band to sub-angstrom wavelengths in the x-ray/gamma-ray bands. This article concerns itself with the radio band, which traditionally covered meter and centimeter wavelengths, but more recently has expanded to shorter wavelengths as radio instrumentation has been developed for millimeter and sub-millimeter wavelengths.

If one were to have eyes sensitive only to radio waves, the Sun, the planets, and the Milky Way would be recognizable, yet different. The Sun would still be extraordinarily bright. However, the tenuous and extremely hot plasma comprising the solar corona would cast a diffuse, luminous haze around the Sun. The planets would shine on their own, instead of in reflected light from the Sun. The Milky Way, the spiral galaxy in which we live, would still glow. However, the radio light would not come from stars too numerous to discern, but instead it would come from highly energized electrons.

The radio sky would also have many new and fascinating features. Few stars could be seen, and these would be double stars so tightly bound together that they interact strongly gravitationally and magnetically. Old stars which have undergone massive explosions, called supernovae, would form large shell-like 'lanterns' which decay in strength slowly over human lifetimes. As these lanterns dimmed, one would occasionally find a remnant beacon or pulsar providing regular pulsating radio waves. The radio sky would also be littered with numerous dumbbell-shaped structures associated with distant galaxies, some near the 'edge' of the universe. Finally, the sky would not be totally dark, as the remnant light from the Big Bang explosion that created the universe continues to bathe us in faint radio waves.

Early history

KARL JANSKY made the first radioastronomy observations in the early 1930s. Jansky was working for the Bell Telephone Laboratories, investigating the sources of interference which might affect transatlantic phone calls. He built a rotatable antenna sensitive to emissions at a wavelength of 15 m and found sources of 'static' coming from local thunderstorms, distant thunderstorms in the tropics and an unknown steady source. The peak of the steady source drifted in position at the same rate at which stars drift across the sky from night to night (about 4 min per day), indicating an extraterrestrial origin for the emissions. Jansky had discovered synchrotron emission associated with energetic electrons accelerated in the magnetic field of the Milky Way.

The astronomical community did not fully appreciate the significance of Jansky's discovery for decades. One of the few who followed up on Jansky's results was GROTE REBER. Reber built a 9 m diameter parabolic reflector to investigate the strength of the radio emission from the Milky Way at meter and centimeter wavelengths. His antenna–receiver system was not sensitive enough to detect the Milky Way at centimeter wavelengths, but he did detect and map the Milky Way at 1.9 m wavelength.

Both Jansky and Reber detected broad-band radiation, analogous to the continuous static in radio reception found between man-made radio channels. Dutch astronomers sought a spectral feature, radiation at a particular frequency analogous to the narrow-band emission from a radio station. Spectral features, or lines, in the radio band come from atoms or molecules, similar to those seen optically. With a spectral line, one can measure a frequency (or Doppler) shift indicative of the relative motion between the source and the observer. A transition of atomic hydrogen at a wavelength of 21 cm provided a candidate, and some years later this line was detected at Harvard University. Observations of atomic hydrogen have been used to study the structure of the Milky Way as well as to measure the mass of other galaxies. A major advantage of observations at radio waves, compared with visible light, is that one can see through interstellar dust grains which hide most of the Milky Way from view at optical wavelengths.

Radioastronomy grew rapidly following the Second World War, aided by the significant wartime investment in radio equipment for use in radar systems. The Sun was shown to be a strong radio source, along with some other isolated sources. The strongest source in the constellation of Taurus, called Taurus A, was identified with the Crab Nebula, a remnant of the AD 1054 supernova chronicled by Chinese astronomers. Another isolated radio source, Cygnus A, proved difficult to identify. Ultimately it was associated with a distant galaxy with a peculiar optical morphology, and Cygnus A became the archetypal member of the important class of objects called radio galaxies.

Radio telescopes

Radioastronomy developed rapidly and some very large telescopes were built. A 250 ft (76 m) diameter, fully steerable telescope was built in England in the 1950s, and other large telescopes were built in Australia, the United States and elsewhere. In the early 1970s, the Effelsberg 100 m telescope (see figure 1) in Germany, the world's largest fully steerable radio telescope, came into operation. The largest (partially steerable) telescope is the Arecibo 1000 ft (305 m) telescope in Puerto Rico. This telescope has a fixed spherical primary reflector, built over a large natural depression, and complex movable radio 'optics' suspended near the primary focus.

The diffraction-limited resolution of a telescope is approximately λ/D, where λ is the observing wavelength and D is the telescope diameter. For the ARECIBO OBSERVATORY operating at 6 cm wavelength, this corresponds to a resolution of about 40 arcseconds, comparable to that of the human eye, but far poorer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

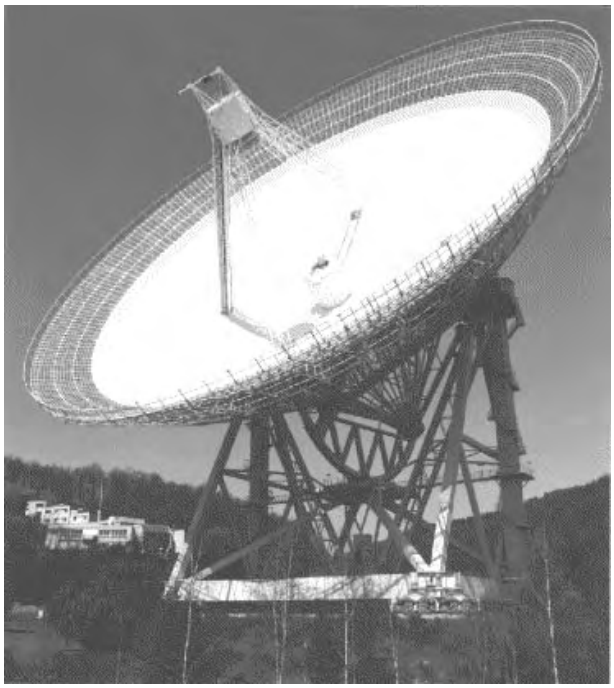


Figure 1. The Bonn 100 m diameter telescope, operated by the Max-Planck-Institut für Radioastronomie in Germany, is the largest, fully steerable, telescope in the world.

than the ≈1 arcsecond (atmospherically limited) resolution achievable by telescopes at optical wavelengths. The construction of a centimeter wavelength radio telescope with ≈1 arcsecond resolution would require a diameter of ≈10 kilometers, which is not technologically feasible. However, arcsecond and better resolution has been achieved though radio interferometry as discussed below.

Imaging in radioastronomy

Single telescopes

Photography produces an image which typically contains millions of picture elements, or pixels. This is achieved with photographic film (or plates) placed in the focal plane of a camera (or telescope). There is no analogous photographic film for radio waves, and radio astronomers have been able to place only a small number of detectors in the focal plane. Thus, maps of radio sources have been generated laboriously by moving the telescope across the source and recording its intensity pixel by pixel. Ultimately, an 'image' of a radio source can be displayed on a computer monitor. While single radio telescopes can image an arbitrarily large field of view, the finest angular resolution, or smallest useful pixel size, is limited by the diameter of the telescope.

Interferometry

The appetite for radio images with angular resolution finer than can be reasonably obtained with single telescopes was satisfied in an elegant manner with RADIO INTERFEROMETERS. Unlike optical interferometry, for which direct combination of light from different telescopes (or segments of one telescope) is employed, a radio interferometer makes use of coherent amplifiers that preserve both the amplitude and phase of the incoming electromagnetic waves. The signals from the individual telescopes can then be replicated without significant loss, distributed and combined with signals from other telescopes.

Combining signals coherently from two telescopes is analogous to what occurs when light from a laser propagates through two pin holes to yield a sinusoidal interference pattern on a screen. One can think of two radio telescopes as pin holes to view the sky and then understand that a radio interferometer is responsive to the sky brightness modulated by a sinusoidal interference pattern. If a mathematical function is multiplied by a sinusoid and then summed (integrated), one obtains a Fourier coefficient of the function. Thus, radio interferometers provide Fourier coefficients of the sky brightness.

One can accumulate many Fourier coefficients of the sky brightness by combining signals from many pairs of telescopes and also by taking advantage of different projections of telescope separations (as seen from the source) provided by the Earth's rotation. These Fourier coefficients can be placed on a grid, called the (u, v)-plane, whose axes are the eastward and northward projection of the antenna separations, and a two-dimensional inverse Fourier transform performed to recover an image of the sky. The result is that each telescope acts as a piece of a giant telescope, equal in size to the maximum separation of individual telescopes. One achieves the angular resolution, but not the light collecting capacity, of the giant telescope. This technique, called aperture or Earth-rotation synthesis, was pioneered by MARTIN RYLE in England who received a Nobel Prize for this work in 1974.

Large-scale interferometers involving many individual telescopes have been built in the Netherlands (Westerbork) and in the United States (the VERY LARGE ARRAY NRAO; see figure 2) and have achieved angular resolutions of better than 0.1 arcseconds. In the mid 1960s some sources of radio emission (radio galaxies and molecular masers) were found to contain structures much smaller than could be resolved with existing interferometers. This motivated radioastronomers to seek even higher resolution by combining signals interferometrically from telescopes across the Earth. This was achieved by tape recording the voltages received at each telescope with a precise timing signal available from atomic clocks. The tapes were then shipped to a 'correlation' facility where the data streams from pairs of telescopes were played back, carefully aligned in time, and cross-multiplied. This technique, called Very Long Baseline Interferometry (VLBI), achieved the astounding angular resolution of ≈0.001 arcseconds and revealed remarkable properties of bright sources. Recently, VLBI techniques have been extended to link a radio telescope in space (the Japanese 'Halca' spacecraft with the 8 m 'VSOP' telescope)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 2. The Very Large Array (VLA) on the plains of St Augustine in New Mexico, operated by the National Radio Astronomy Observatory of the USA, contains 27 25 m diameter telescopes.

with arrays of ground telescopes, yielding interferometer baselines of about three times the Earth's diameter.

Millimeter wavelengths

In recent decades much of the emphasis in radioastronomy has been on observations at shorter wavelengths. The pioneering work of millimeter wave telescopes, such as the Bell Labs 7 m, the Texas 5 m, the NRAO 12 m (originally 36 ft), the FCRAO 14 m and the Onsala 20 m, opened up the field of observational molecular spectroscopy. Carbon monoxide (CO) was detected in 1970 from giant molecular clouds, whose dense cores have now been studied in line emission from many other molecules. Molecular line observations have enabled astronomers to study the star formation process in these clouds in great detail. Following up on these breakthroughs, two large millimeter wave telescopes were constructed in the 1980s: the NOBEYAMA RADIO OBSERVATORY 45 m in Japan and the INSTITUTE DE RADIOASTRONOMIE MILLIMETRIQUE (IRAM) 30 m in Spain. These telescopes, which operate at wavelengths as short as 1 millimeter, have greatly added to our understanding of the distribution of molecular clouds in the Milky Way and the chemical processes which operate in these clouds.

Interferometry techniques have been developed to work at millimeter wavelengths at the Berkeley–Illinois–Maryland (originally HAT CREEK RADIO OBSERVATORY) and the OWENS VALLEY RADIO OBSERVATORY arrays in California, the Nobeyama array in Japan, and the IRAM array in France. The high angular resolution afforded by these interferometers has allowed astronomers to study disks surrounding newly formed stars, elucidate the nature of spiral arms in galaxies, and detect emission from molecular gas and dust at cosmological distances.

Electronics

Radioastronomy has benefited immeasurably from the development of low-noise amplifiers. The added noise can be exceedingly small, equivalent to a black-body radiator whose temperature is only a few degrees above absolute zero. Today the preferred amplifier is a cooled HFET (heterostructure field-effect transistor) device. These devices work best at centimeter wavelengths. For shorter wavelengths it is advantageous first to shift, or mix, the observing band to a longer wavelength where HFET amplifiers can be used. Currently this is accomplished with an electronic device made up of a triple-decker 'sandwich' comprising a superconducting layer, an insulating layer and another superconducting layer (SIS), which provides the mixing with little added noise.

Digital electronics have also had a dramatic effect on radioastronomy. In 1961, Sander Weinreb constructed the first digital spectrometer for radioastronomy. This instrument digitized the amplified voltage from a telescope and formed the autocorrelation function by multiplying replicas of the voltage delayed in time by different amounts. The autocorrelation function was then Fourier transformed to yield the source spectrum. In early CORRELATORS, digitization involved clipping, or 1-bit sampling, a process by which positive voltage is conceptually represented by a +1 and negative voltage by a −1. Although one would suspect that retaining only the sign of the voltage would lose most of the information in the signal, a mathematical relationship (derived by J H Van Vleck during the Second World War) demonstrates that one retains the essential character of the spectrum, losing only a factor of $2/\pi$ in signal-to-noise ratio. Often this loss of signal-to-noise ratio is offset by the exceptional stability of digital correlation, allowing one to integrate noise levels down for long periods of time without encountering systematic effects, such as frequency-dependent gain variations, that plague analog spectrometers.

In addition to spectrometers, digital correlators are used to cross-multiply signals from pairs of telescopes used as an interferometer. These correlators are special purpose digital computers which perform an astounding number of operations per second. For example, the Very Large Array correlator, which was built in the early 1980s, processes data from each telescope coming in at hundreds of megasamples per second. It cross-multiplies these signals for all 351 independent pairings of the 27 individual telescopes, corresponding to an aggregate processing speed of over 100 000 megahertz!

Discoveries in radioastronomy

Radio galaxies

The discovery of radio galaxies is one of the many achievements of radioastronomy. At first, the origin of these sources was uncertain. The radio emission was resolved with interferometers and presented as two giant 'lobes' as shown in figure 3. Once accurate positions for the radio emission from these sources were obtained, astronomers noted that they were often centered on distant, giant galaxies—usually the dominant galaxy at the center of a large cluster of galaxies. These lobes are among the largest structures in the universe—some separated by more than a million light years, or nearly a hundred times the size of our Milky Way. The radiation mechanism is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

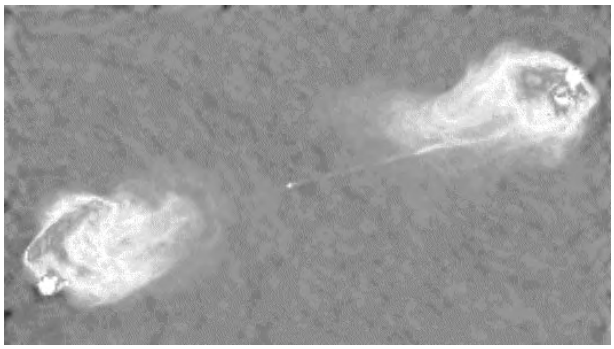


Figure 3. The radio galaxy Cygnus A as imaged by the VLA at a wavelength of 4 cm. Note the large double lobes of radio emission and the thin jets connecting the lobes to the nucleus of the galaxy at the center of the image.

now understood to be incoherent synchrotron emission, which comes from electrons excited to nearly the speed of light and then accelerated in a spiral motion around a magnetic field line. Remarkably the minimum possible energy contained in energetic particles and magnetic fields needed to power radio galaxies can exceed $10^7 M_\odot c^2$—that is the total conversion of the mass of 10 million Suns into energy!

What is the source of this phenomenal amount of energy in radio galaxies? The symmetrical placement of the lobes about a central galaxy clearly suggested a galactic link. Next, the discovery of thin jet-like structures connecting the lobes with the centers of the galaxy indicated a channel for the energy to flow. Finally, images of the nuclei of radio galaxies made with VLBI indicated that these JETS continued down to scales less than 0.001 arcseconds, or typically less than 1 parsec. Thus, a strong circumstantial case was made that the mass of far more than 10 million Suns must have flowed through a tiny region at the center of a radio galaxy, smaller than that from the Sun to its nearest neighboring star, and then was converted into energy! This reasoning lead to the current understanding that SUPERMASSIVE BLACK HOLES power radio galaxies.

Quasars and superluminal motion

Large portions of sky were observed at ≈1 m wavelength, such as by the third Cambridge (or 3C) survey, and astronomers sought optical identifications for the newly discovered sources. While many radio sources could be associated with supernova remnants in the Milky Way or with radio galaxies, one class of sources appeared as bluish, star-like objects called quasistellar radio sources or quasars. In 1963, astronomers at Caltech identified spectral lines in two quasars, the BRIGHT QUASAR 3C 273 and 3C 48, as having extraordinarily high recessional velocities (or 'redshifts'). Distances to these objects obtained from their redshifts using the Hubble law indicated extremely great, cosmological, distances and, correspondingly, extremely high intrinsic luminosities. The cosmological interpretation of high redshifts was controversial since quasars would then be the most luminous objects in the universe. (Recently, however, nearly all astronomers have come to accept this interpretation. Indeed, 'normal' galaxies have been found with red shifts greater than 4, making the 'extreme' redshift of 3C 48 of 0.4 unremarkable today.)

Quasars, along with radio galaxies and many other types of galaxies which exhibit strong nuclear activity, are all thought to be powered by accretion of matter onto supermassive black holes. Early VLBI observations of some quasars revealed an usual morphology: emission dominated by a small number of 'blobs' which appeared to expand at speeds exceeding that of light. This phenomenon, called SUPERLUMINAL MOTION, has been well studied for decades and numerous quasars have been found to exhibit superluminal motion. Apparent superluminal motion is explained quite simply in terms of the constant and finite speed of light (and radio waves). A source moving at nearly the speed of light and nearly along the line of sight toward an observer can almost keep up with its own light, causing the observer to see events in the source greatly speeded up.

The cosmic background

In 1965 ARNO PENZIAS and ROBERT WILSON, working at the Bell Telephone Laboratories, discovered that 'blank sky' is not totally dark at radio wavelengths. Instead, the universe is bathed in a 'cosmic background' (see COSMIC MICROWAVE BACKGROUND) of radiation, which has an intensity equivalent to that of emitting material (a black-body) at a temperature of 2.7 K (equivalent to about −270.4 °C). The cosmic background is the cooling remnant of the Big Bang 'explosion' that marked the origin of the universe. Recently, the Cosmic Background Explorer (COBE) satellite confirmed that the spectral shape of the emission corresponds precisely to what would be expected for a black-body. COBE also detected fluctuations in the background, which could possibly explain the formation of galaxies in the Universe.

Pulsars and supernova remnants

In the mid-1960s, ANTHONY HEWISH in England led the construction of a telescope capable of studying variable radio emissions. Some radio sources were known to be variable, owing to propagation effects (scintillation by free electrons) in the interstellar and interplanetary medium. In 1967, JOCELYN BELL observing with this telescope unexpectedly discovered sources of pulsed radio emission with very regular periods (∼1 s). Pulsars, as they are now called, are formed in explosions of evolving, massive stars. During the explosion the outer envelope of the star is expelled, while the inner portions are compressed so strongly that protons and electrons join to make neutrons. These 'neutron stars' typically contain about 40% more mass than the Sun and yet have a radius of only about 10 kilometers! The formation of a neutron star also leads to astounding magnetic field strengths (∼10^{12} G)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and to rapid rotation (∼1 Hz). We see intense radio emission when the magnetic pole sweeps by at the star's rotation rate, if the rotation axis and the magnetic pole are misaligned.

More recently, RUSSELL HULSE and JOSEPH TAYLOR discovered a double-star (binary) system containing a pulsar. Precise timing of the pulsar signals over many years has lead to the detection of changes in the orbit of this binary pulsar caused by energy loss owing to gravitational radiation as predicted by Einstein's theory of general relativity. Hulse and Taylor received the 1993 Nobel Prize for this discovery, joining Hewish who received this prize in 1974 for his role in discovering pulsars.

The supernova explosion also creates a shock wave which expands into the interstellar environment, plowing up and exciting a hot sphere or shell-like region. The hot electrons in such SUPERNOVA 'REMNANTS' are strong radio sources and can be seen throughout the Milky Way and in other galaxies.

The galactic center

The central region of our Milky Way has proven to be rich in a variety of phenomena, containing numerous supernova remnants, massive star-forming clouds, strong x-ray and gamma-ray sources, and towering magnetic structures which protrude from the galactic plane (see figure 4). In the early 1970s, an intense radio source was discovered in the galactic center. This radio source, called SAGITTARIUS A (Sgr A)*, is very compact, with an intrinsic size less than the distance from the Sun to the Earth (1 AU), and appears to mark the center of our rotating galaxy. Stars in a dense cluster surrounding Sgr A* move at speeds exceeding 1000 km s^{-1} at projected distances of 3000 AU from Sgr A*. These high speeds, if resulting from gravitationally bound orbits, indicate that Sgr A* contains about 3 million solar masses. While stars around Sgr A* move rapidly, Sgr A* itself appears stationary, suggesting that it is far more massive than the stars surrounding it. The evidence is very strong that Sgr A* is a supermassive black hole, a diminutive relative of the supermassive black holes that power quasars and radio galaxies.

Gravitational lensing

One of the predictions of Einstein's theory of general relativity is that gravitational mass bends light and therefore acts as a lens. Gravitational lenses magnify background sources, distort their appearance, and even create multiple images of a single source. In 1979, 0956+571, a source discovered in a large radio survey, was imaged and found to consist of two compact objects separated by only about 6 arcseconds. Subsequent optical observations revealed that the two objects were quasar-like and had nearly identical spectra. While it was conceivable that two quasars could form in the same environment and have similar properties, the possibility that these 'quasars' were two images of the same object was a simpler explanation. Recently, fluctuations in the two 'quasars' have been shown to be nearly identical, except for a time difference of roughly one year as expected if there is one quasar whose light reaches us along two different paths with different path lengths. This provides conclusive evidence of GRAVITATIONAL LENSING. In the past 20 years dozens of other lensed systems have been identified. In these systems, the lens is formed by the gravitational mass of a galaxy or a cluster of galaxies.

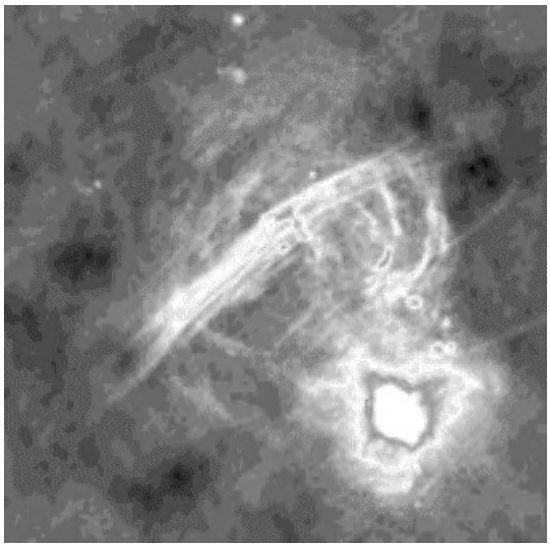


Figure 4. An image of the central region of the Milky Way made with the VLA at a wavelength of 4 cm. The filamentary structure near the center of the image indicates a complex network of magnetic field lines. The actual center of the Milky Way, probably a supermassive black hole containing a few million solar masses, is hidden in the center of the brightest region near the lower right corner of the image.

Molecular clouds

Life as we know it on Earth involves very complex chemical processes which give rise to many organic molecules. It had been thought that such molecular complexity could only evolve on planets. However, in the 1970s radioastronomers started to discover many molecules, including water, ammonia and simple organic molecules, in clouds of interstellar gas. At centimeter wavelengths detectable spectral lines are rare. However, at millimeter, and especially at submillimeter wavelengths, the spectrum is nearly completely covered with detectable lines of molecules, e.g. methyl alcohol (CH_3OH). Over 100 different molecular species have been identified; some species, such as $HC_{11}N$, contain more than a dozen atoms. These molecular clouds, as they are now called, are the sites of the formation of stars and planets and provide the building blocks for life.

Masers

One of the early surprises of molecular spectroscopy in the radio band was the discovery of extremely strong emission lines from the hydroxyl (OH) radical. These lines were not only thousands of times stronger than

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

expected, they also were extremely narrow and highly polarized. Originally dubbed 'mysterium' because of these remarkable properties, the emission was correctly understood as coming from a natural amplification process, called a maser, analogous to the process that results in (optical) lasers. Masers have been detected in many molecular species, such as water vapor and methyl alcohol. Water vapor displays the most spectacularly strong signals of any known maser species. In fact, a water maser in a star-forming region called W49, located most of the way across the Milky Way, puts out as much energy in one narrow spectral line as the Sun does across the entire visible band.

Sun and planets

As early as the 1940s, the Sun was found to be a very strong and complex source of radio emission. Intense, sporadic, low-frequency emissions were detected by radar equipment during the Second World War but these discoveries were kept classified until the end of the war. A variety of solar flares were seen and found to be associated with sun spots and other active regions.

Early radio observations of the planets also revealed diverse and fascinating characteristics. Optical astronomers thought they had established that Mercury's spin and orbital periods were equal (as occurs for our Moon owing to tidal effects). Thus, it was a great surprise when early radar observations revealed that MERCURY instead rotates three times for every two orbits about the Sun. In retrospect, the erroneous conclusion that Mercury's spin and orbit were locked was based on images which were obtained only when the planet was near the maximum elongation from the Sun; with such poor sampling many spin periods could fit the same data.

Early radio measurements of the brightness of VENUS indicated very high temperatures, and this was thought to indicate a hot ionosphere. However, the polarization of the radio emissions confirmed that the high temperature, about 750 K, came from the planet's surface. The high surface temperature of Venus is understood to be due to a strong 'greenhouse' effect, in which visible light from the Sun heats the planet, evaporates its oceans and produces an extremely dense atmosphere. Infrared emission cannot easily escape the atmosphere, preventing cooling, and further elevating the temperature.

JUPITER, surprisingly, was found to have strong emission at about 10 cm wavelength, which was shown to be synchrotron in nature. Even more surprising was the discovery of sporadic emission at about 10 m wavelength, which comes from electrical discharge involving the interaction of Jupiter's innermost satellite, Io, with the magnetic field of the giant planet.

Future directions

The current trend in developing major new telescopes for radioastronomy is toward interferometric arrays. Distributing the collecting area of an instrument among many individual telescopes can greatly reduce costs and at the same time increase angular resolution compared with building single large telescopes. Extremely ambitious projects are currently planned or being developed. For example, a millimeter/sub-millimeter wavelength array called the Atacama Large Millimeter Array (ALMA) will soon be built by an international collaboration involving the United States, Europe and possibly Japan. This array is planned to consist of more than 50 12 m diameter telescopes spread over an area of about 10 km on an exceptionally dry 5000 m elevation site in northern Chile.

Looking toward the future, radioastronomers in many countries are working to develop the technology necessary to build a centimeter wavelength telescope with a collecting area of one square kilometer! This would represent an increase in collecting area of about a factor of 70 over the VLA and a factor of about 15 over Arecibo.

The far future for radioastronomy will probably involve radio telescopes in space. The weightless environment of space should allow large, low-mass telescopes to be constructed, extremely long interferometer baselines and decreased exposure to man-made interference. The latter benefit of space may be the most important, as the explosion of electronic technologies has resulted in a dramatic increase in the pollution of the radio band from fixed and satellite transmitters broadcasting a variety of data from telephone calls to television programs.

Bibliography

Rolufs K and Wilson T L 1996 *Tools of Astronomy* 2nd edn (Berlin: Springer)

Verschuur G L and Kellermann K I (ed) 1988 *Galactic and Extragalactic Radio Astronomy* 2nd edn (New York: Springer)

Mark Reid

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ramsden, Jesse (1735–1800)

Instrument-maker, born in Halifax, Yorkshire, England, the most outstanding maker of precision scientific instruments of the eighteenth century, producing extremely precise scaled circles. In 1775, he made the first dividing machine that could enscribe precise divisions on the borders of small-diameter circles. He made the theodolites with 3 foot diameter circles used in triangulation to connect the geodetic networks of France and England, and the large altazimuth instrument with a 5 foot diameter vertical circle built for PIAZZI. His telescopes used achromatic lenses according to DOLLOND's design (Ramsden was married to his younger sister). His name is commemorated in the design of the *Ramsden eyepiece*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ranger

Series of NASA missions launched 1961–65. Designed to send back detailed images of the Moon prior to crash landing on the lunar surface. Only the last three, Rangers 7, 8 and 9, were successful.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ray

A bright linear feature on a planetary body formed by ejecta from a recent crater-forming impact. Surface material darkens with age through exposure to radiation, so rays, consisting of freshly excavated material, are bright by contrast. In addition, some impacts have sufficient force to vaporize surface material, which can resolidify in a reflective, glassy form. Rays extend from their associated crater, often in a radial pattern known as a *ray system*. The best-known ray systems are lunar, the most prominent being the one surrounding the crater Tycho, which is very prominent at full moon, individual rays extending for a thousand kilometers or more. Other solar system bodies with rays are Mercury and Ganymede.

See also: Tycho (Crater).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rayet, Georges Antoine Pons (1839–1906)

French astronomer, who co-discovered *Wolf–Rayet (WR) stars* (see CHARLES WOLF). Became director of the Floriac Bordeaux Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rayleigh Limit

The theoretical resolving power of a telescope according to a criterion devised by Lord Rayleigh (1842–1919). Because of the phenomenon of diffraction the image of a point source of light (such as a star) produced even by a perfect optical instrument consists of a central bright spot (the Airy disk) surrounded by concentric dark and light rings. If two point sources are very close together, the resulting image will consist of two overlapping diffraction patterns. According to Rayleigh's criterion, two close point sources of light may just be resolved if the angle between them is such that the center of the Airy disk corresponding to the image of one source falls on the first minimum in intensity in the diffraction pattern of the other (i.e. the first dark fringe around the Airy disk of the other source). Under these circumstances, assuming an ideal optical system, there will be a drop in light intensity of 20% mid-way between the centers of the two Airy disks.

The Rayleigh limit is given, in radians, by $\alpha = 1.22\lambda/D$, where λ denotes wavelength and D denotes the aperture of the telescope; and in seconds of arc by $\alpha = 2.52 \times 10^5 \lambda/D$. Thus, for example, the theoretical resolution of a telescope with an aperture of 0.1 m, for light of wavelength 550 nm (5.50×10^{-7} m), is $2.52 \times 10^5 \times (5.50 \times 10^{-7}/10^{-1}) \approx 1.4$ arcsec.

See also: Airy disk, aperture, Dawes' limit, resolving power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reber, Grote (1911–2002)

American radio engineer, born in Chicago, IL. After reading JANSKY's articles on radioastronomy he built (1937) the world's first radio telescope (a 9 m tiltable paraboloid) in his backyard in Wheaton, IL. With it he detected radio emission ('cosmic static') at 1.9 m wavelength from the Milky Way. With an improved receiver in 1941 he detected the Sun and a strong source in Cassiopeia (Cas A). As a self-employed researcher in Hawaii and Tasmania he mapped the background radiation at 1–2 MHz. He produced the first radio maps of the sky.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Receivers: Millimeter and Longer-Wavelength Feeds and Receivers

The frequency range over which ground-based RADIO ASTRONOMY observations can be carried out extends from about 2 MHz ($\lambda = 150$ m) at the low-frequency end to near 1 THz ($\lambda = 350\ \mu$m) at the high end. This discussion treats systems and techniques from the low-frequency limit up to a frequency of 300 GHz ($\lambda = 1$ mm); submillimeter systems and techniques are dealt with in the next article (see also MILLIMETER AND SUBMILLIMETER ASTRONOMY).

Within this very broad range of frequencies, astronomers measure the properties of incoming cosmic radiation in the form of broadband continuum or narrowband spectral-line emission. This incoming radiation is collected by the RADIOTELESCOPE, amplified, detected and recorded by a sensitive radio receiver mounted on the telescope, and then processed in such a way as to display the output as a flux density measurement, or as a spectrum of the frequency band to which the receiver is tuned.

The power levels of the received signals are extremely small. Even after collection by the largest telescopes, they amount to as little as 10^{-16} W for a continuum source or 10^{-21} W for a weak spectral line. For this reason, the receiver must contribute a minimum of self-generated noise to the overall system and be highly gain stable in order to detect very small changes in noise power when the telescope is pointed at a RADIO SOURCE. In addition, it must be highly stable in frequency to detect narrowband spectral-line emissions and, when arrayed with other telescopes to form an INTERFEROMETER, must have a high degree of overall phase stability.

Modern radio astronomy receivers—general features

The modern radio astronomy receiver system can be divided into four main subsystems as depicted in figure 1. The first of these is a feed horn–front end combination situated at one of the focal regions of the telescope system. The front-end electronics will consist of low-noise preamplifiers for centimeter wavelength operation or a heterodyne mixer detection stage for millimeter wavelengths (for which amplifiers are not readily available). The output signal from the mixer will usually be at centimeter wavelengths and similar to the output of longer-wavelength receivers. The feed horns, amplifiers and mixers are integrated in such a way as to minimize transmission line loss between these elements, thus avoiding signal loss and excess noise. Finally, to reduce overall front-end receiver noise the preamplifiers, mixers and millimeter wave feed horns are all cooled to cryogenic temperatures.

The next part of the signal path employs the heterodyne mixing process again, where the amplified centimeter wavelength signals, or the down-converted millimeter mixer outputs, are converted to an intermediate frequency (IF) which can be more readily amplified. It is at this frequency that further signal conditioning such as filtering, gain control and detection can take place.

The principle of heterodyne detection can be understood by reference to figure 1. Here the received sky frequencies ν_{mm} and ν_{cm} are mixed in non-linear elements with local oscillator (LO) frequencies ν_{LOmm} and ν_{LOcm} to produce IFs ν_{cm} and ν_{IF}.

In fact, the mixing process produces many frequency products consisting of the sum, difference and all harmonic combinations of the signal and LO inputs. However, it is the difference frequencies which are selected by a filter and amplifier following the mixer, to produce the IF output band required. In any broadband mixer system, two input frequency bands will contribute to the signal content in the mixer IF output. These bands are given by $\nu_{LO} + \nu_{IF}$ and $\nu_{LO} - \nu_{IF}$, forming what is known as a double-sideband mixing system. This scheme was commonly used for millimeter-wavelength systems in years past; however, recent improvements in sideband rejection techniques allow single-sideband measurements to be made with most modern millimeter receiver systems. At centimeter wavelengths, which employ low-noise preamplifiers before the mixing process, filters can be used between the amplifier and the mixer to select the desired sideband for the mixer input.

The heterodyne mixing process requires an appropriate source of LO signal, and this forms the third subsystem necessary for the receiver. The LO system will provide all the necessary frequency sources at an appropriate power level to be mixed with the incoming sky signals and all the subsequent IF signals present in a multiple conversion receiver system. Frequency stability of these sources is paramount, particularly for spectral-line systems, and if the telescopes form part of an interferometer array, the phase behavior of these sources requires control along with a high degree of spectral purity and low phase noise.

A final signal subsystem deals with the output of the IF section in two ways. The first is to detect the IF power in a square law detector (a diode rectifier whose output dc voltage is proportional to the input IF power), integrate its output for a fixed integration time and then digitally record this output via an A–D converter. This output can then be calibrated as a measure of flux density.

The second method involves the use of a digital spectrometer to obtain a power spectrum of the final IF passband. This is done by first digitally sampling the IF noise waveform with sufficient speed and precision as to have minimal effect on the measurement accuracy. A digital CORRELATOR is then used to generate an autocorrelation function of the time-sampled IF noise, after which a Fourier transform is used to obtain the power spectrum. This spectrum is a direct representation of the spectrum at the incoming sky frequency, transposed by the various local oscillators used in the converter systems.

Receiver sensitivity

The contributions to excess noise in a practical radio astronomy system come from several sources which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

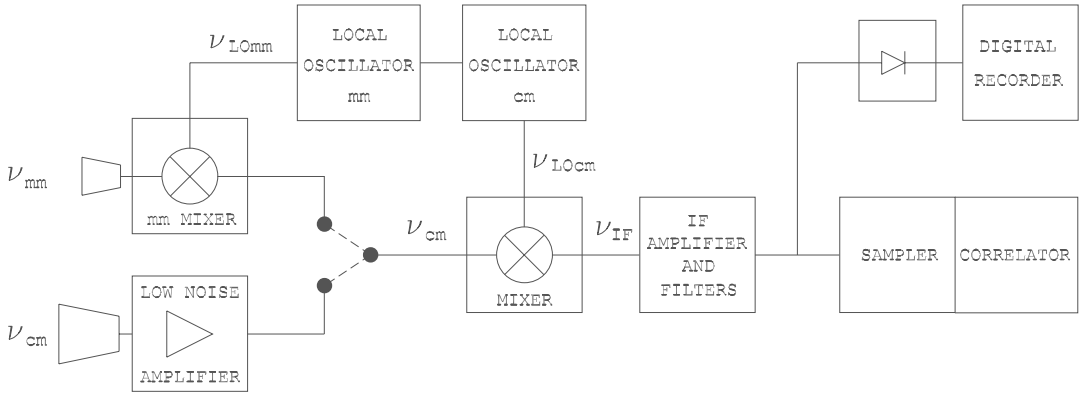


Figure 1. Block diagram of a typical radio astronomy receiver for both millimeter and centimeter wavelengths.

include atmospheric attenuation, antenna spillover and diffraction, feed horn and transmission line losses, as well as the receiver itself. The sensitivity of the receiver system depends on the level of this excess noise power. To quantify this, a figure of merit called the system temperature, T_{sys}, can be assigned to the receiver and telescope system. This is defined as the physical temperature of an ideal resistive termination, placed at the input of an ideal receiver with the same gain and bandwidth and producing the same noise power output as the practical system.

The minimum signal detectable against this noisy background is then defined by the radiometer equation, $\Delta T_{rms} \propto T_{sys}(B\tau)^{-1/2}$, where B is the final bandwidth just prior to detection and τ is the integration time. This expression indicates how important it is to minimize T_{sys} to obtain maximum sensitivity.

For frequencies of 1–300 GHz the values of T_{sys} range from about 12 K to ∼1000 K. Bandwidths can vary from 1 kHz for spectral-line measurements to several GHz for broadband continuum observations, while integration times vary from fractions of a second for strong time-varying phenomena to tens of hours for weak spectral lines.

The feeds

Reflector antennas need a feed horn or some other type of radiating element to couple the receivers to the reflector systems. For effective coupling the feed should be dual polarization, have low cross-polarization coupling, illuminate the reflector surfaces efficiently, have a highly circular beam and have sufficient bandwidth to match the receiver. Much development has occurred in feed-horn design over the years, starting with simple circular waveguide horns, progressing through choke-ringed aperture and early multiple-mode designs, to the modern corrugated hybrid-mode horn in its single- or dual-mode form. These corrugated horns are characterized by radial quarter-wavelength slots (operating frequency) in the walls of the waveguide along its length, and a short mode conversion section at the waveguide input end to excite the required hybrid mode. This mode, on radiation from the horn aperture, displays very low cross-polarization, excellent pattern symmetry and low sidelobe structure thus making it an ideal solution for radio astronomy application. A further development to this type of horn resulted in the internal corrugated surface being profiled to produce a more compact design, and this, when excited by a ring-loaded slot mode converter, produces a feed-horn system with an octave bandwidth. An excellent review of the design parameters of all these types of feed is given by Olver *et al* (1994).

A recent development in hybrid-mode feed design is the dielectric-loaded horn. This is a smooth-walled, profiled horn filled with a dielectric of appropriate permittivity. A small gap between the dielectric and horn wall allows the hybrid mode to propagate and, because there are no quarter wave slots, the potential bandwidth of this design is enormous. A recent design indicates a horn performance covering 16–120 GHz, with pattern symmetry, low cross polarization and high return loss being maintained over this wide frequency range.

For feed arraying applications such as focal-plane imaging arrays, while the corrugated horn is still the most efficient, its size makes it difficult to pack closely. Simpler horns, such as the Potter, stepped and diagonal horns, are better suited for these applications, such as multibeam horn arrays for fast surveying. Figure 2 shows smooth-wall stepped horns in a 13-beam hexagonal array, at 21 cm on a 64 m prime focus fed paraboloid.

Millimeter wave receivers

Most modern millimeter wave receiver systems will have a quasi-optical input beam waveguide, in which polarization splitting and LO power coupling can take place before the incoming radiation finally illuminates a lens-corrected feed horn. The feed horn will be closely integrated with a heterodyne mixer block, both of which will be cooled within a cryogenic vacuum dewar to 4 K. The horn coupling to the beam waveguide will be via a transparent vacuum window.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

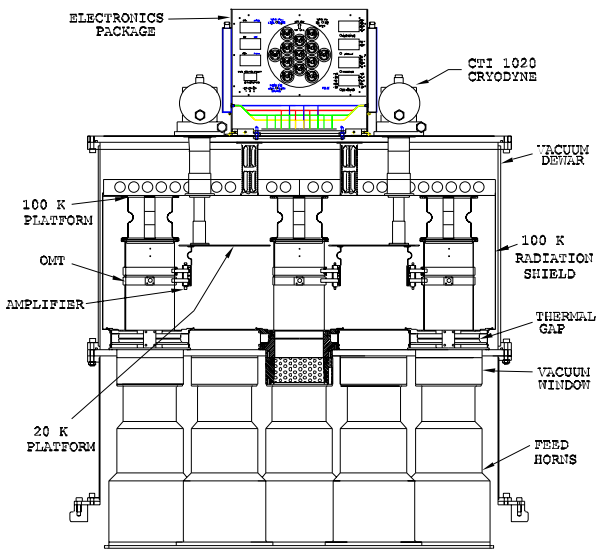


Figure 2. Cross-section view of a 21 cm multibeam front end consisting of a 13 stepped horn hexagonal array, cryogenic receiver dewar and cooled orthomode waveguide-coaxial transitions.

The non-linear element used in the heterodyne mixer is a superconducting tunnel-junction type or superconductor–insulator–superconductor (SIS) junction. Mixers utilizing these devices have very low noise temperatures and because of this have replaced gallium arsenide (GaAs) Schottky diode mixers in almost all of the millimeter wave systems currently operating. In practice the single-sideband (SSB) noise temperature of the best SIS receivers will be less than 0.5 K GHz^{-1} across the whole millimeter spectrum. Early tuning of these mixers was accomplished by adjustable backshorts in the waveguide structure. However, current mixer designs now achieve broadband tuning by integrated tuning circuits using thin film inductances to resonate the SIS junction capacitance.

Because the LO power required by SIS mixers is quite small ($\sim$1 μW), coupling to the junctions can be made either via the beam waveguide system through the feed horn or by $-$20 db directional coupler between the feed horn and mixer block. Sources of LO are now derived directly from phase-locked Gunn oscillators for the longer millimeter wavelength systems in the 1 cm–2 mm region or Gunn oscillator–multiplier combinations for wavelengths in the 1 mm region.

A significant development in recent years has been the rapid development of low-noise high electron mobility transistor (HEMT) amplifiers with good performance in the millimeter-wave region.

The availability of robust indium phosphide (InP) HEMT devices has now resulted in reliable hybrid 'chip and wire' amplifiers being developed, to operate at frequencies up to 110 GHz, with noise temperatures of 50 K when cooled to a physical temperature of 15 K. Monolithic microwave integrated circuit (MMIC) amplifiers also using InP, and designed to operate over the 75–110 GHz band, have also recently been tested at 15 K physical and found to yield amplifier temperatures of 50–60 K. This makes these low-noise amplifier systems competitive with SIS mixer receivers at frequencies up to 115 GHz, with the added advantage of a much simpler cryogenic cooling system.

Centimeter wavelength receivers

Centimeter wavelength receiver systems are characterized by preamplification of the incoming sky signal in a low-noise amplifier (LNA) before heterodyne mixing to a lower IF. The LNA is almost universally cryogenically cooled (except for frequencies below 1 GHz where cooling is not necessary) and utilizes GaAs HEMT amplifiers for frequencies up to $\sim$10 GHz and more recently InP HEMT amplifiers for frequencies up to 45 GHz. Cooled noise temperature performance of these LNAs is excellent, being in the region of 1 K GHz^{-1} for GaAs HEMT amplifiers and $<$0.5 K GHz^{-1} for InP HEMT amplifiers. The feeds used with these receivers are almost exclusively corrugated hybrid mode horns giving maximum efficiency with the telescope system, minimum spillover noise and good pattern symmetry for precision polarization measurements. In addition, wide bandwidths can be achieved by using the profiled form of corrugated feed.

While these centimeter feeds are often much too large to cool cryogenically, the transmission lines between the feed and LNA often are cooled cryogenically, so the feed horns must be integrated with the cryogenic dewar. Figure 3 illustrates this, showing a receiver–feed horn combination spanning 2 cm–3 mm wavelength and utilizing three horns of the compact corrugated profiled form. The linear polarization splitters, and following InP HEMT amplifiers for all three receivers, are incorporated within the cryogenic dewar and cooled to a physical temperature of 15 K.

The heterodyne mixer and IF systems following these centimeter front ends are now simple and straightforward. Wide-bandwidth balanced and double-balanced diode mixers are readily available with operating wavelengths up to 1 cm (30 GHz). The LO power requirements for these mixers is readily obtained from YIG (yttrium iron garnet) tuned transistor oscillators or dielectric resonator oscillators, phase locked to a reference frequency derived from an atomic standard oscillator.

A novel development in recent years has been the use of optical-fiber transmission systems to distribute these LO reference frequencies within the telescope and across arrays of telescopes. Fiber systems have also been used to transport band-limited sky signal noise from remote antenna front ends and to distribute IF signals within a telescope system.

Future developments

The impetus for future radio astronomy receiver development will be driven by all types of arraying applications.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

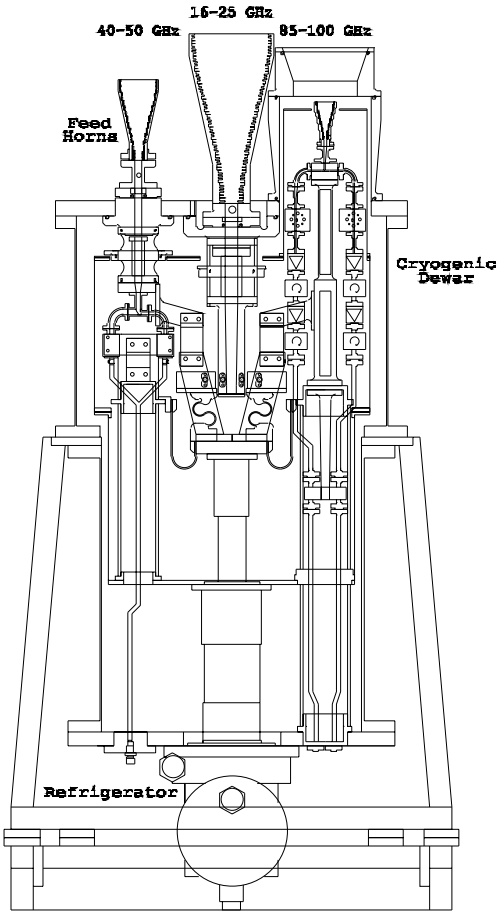


Figure 3. Wideband front end utilizing compact corrugated feed horns. Frequency selection is accomplished by feed horn translation to the antenna Cassegrain focus.

Focal-plane feed arrays will require close-packed feeds backed by compact receivers to image the focal-plane region of single paraboloids, while distributed arrays of large collecting area will require very large numbers of sensitive, wideband, reliable and inexpensive receivers.

These challenges will best be met by having all parts of the basic receiver subsystems in the form of MMICs. Already tentative steps are being taken to develop low-noise amplifiers, heterodyne mixers and oscillators in this form for both centimeter and millimeter bands, leading to the possibility of a complete 'receiver on a chip' in the very near future. A successful integration of these devices, combined with fiber-optic connections between elements, will be a major step forward towards developing large-area arrays.

Bibliography

Bird T S and James G L 19xx Design and practice of reflector antennas and feed systems in the 1990s *URSI Review of Radio Science 1996–1999* ed W Ross Stone (Oxford: Oxford University Press) pp 81–117

Carlstrom J E and Zmuidzinas J 19xx Millimeter and submillimeter techniques *URSI Review of Radio Science 1993–1995* ed W Ross Stone (Oxford: Oxford University Press) pp 839–82

Olver A D, Clarricoats P J B, Kishk A A and Shafai L 1994 *Microwave Horns and Feeds (IEE Electromagnetic Wave Series 39)* (IEEE Press)

Pospieszalski M W and Wollack E J 1998 Characteristics of broadband INP HFET millimeter-wave amplifiers and their applications in radio astronomy receivers *Proc. 2nd ESA Workshop on Millimeter Wave Technology and Applications: Antennas, Circuits and Systems (Espoo, 27–29 May 1998)* pp 221–6

Tiuri E and Räisänen A V 1986 Radio telescope receivers *Radio Astronomy* 2nd edn, ed J D Kraas (Powell, OH: Cygnus-Quasar Books) 7-0

Malcolm Sinclair

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Receivers: Submillimeter Feeds and Receivers

The submillimeter band of the ELECTROMAGNETIC SPECTRUM is generally defined as wavelengths in the range 1 mm–100 μm. This spectral range is at present only partially explored owing to molecular (H_2O and O_2) absorption in the Earth's atmosphere. Also, it is difficult to extend the performance of radio-style telescopes to this band. Finally, it is still early in the development of both continuum detectors, usually bolometers, and heterodyne, or radio, detectors, which are used for high-resolution spectroscopy. It is the latter devices which form the subject of this article.

The submillimeter band contains several important aspects of astrophysics, e.g. the spectral features of the rotational motion of interstellar molecules such as CO and CH_3OH and the fundamental lines of light, hydride molecules such as H_2O, NH_3, HCl and HF. The band also contains the ground state fine-structure lines of atoms and ions, e.g. atomic carbon [C I]. The interstellar medium also contains dust, seen in the continuum, and the combination of spectral and continuum data provides information which enables the study of STAR FORMATION in our Galaxy. The submillimeter band naturally contains emission mostly from objects whose temperatures range from about 10 to 100 K. This is indeed the temperature range of much of the interstellar medium of our Galaxy but is also the temperature of the redshifted, but intrinsically hotter, gas and dust in distant galaxies. This means that much of the information about distant galaxies, now seen to $z \sim 5$, comes to us in the submillimeter band (see MILLIMETER AND SUBMILLIMETER ASTRONOMY.

Astronomy demands the very best possible receivers to detect the limited number of photons arriving from such distant objects. Several new facilities are under construction or in the design and development phase, complementing existing facilities such as the CALTECH SUBMILLIMETER OBSERVATORY and JAMES CLERK MAXWELL TELESCOPE. These include the Smithsonian interferometer on Mauna Kea, ALMA, the international interferometer planned for the Atacama desert in Chile, SOFIA, the US–German airborne observatory, and FIRST, the ESA–NASA space observatory.

Receiver principles

The operation of a heterodyne receiver in the submillimeter is similar to that for a radio receiver. The basic principle is to mix the weak signal ($E_1 \cos \omega_1 t$) from the telescope with a strong local oscillator signal ($E_2 \cos \omega_2 t$) in a non-linear element, so generating the difference term $E_1 E_2 \cos(\omega_2 - \omega_1)t$ of much greater power than E_1^2 and at the frequency $\omega_2 - \omega_1$, which is determined by the frequency range of the preamplifier. The amplified intermediate-frequency (IF) signal, say 1–2 GHz, is finally integrated in small frequency intervals, say 1 MHz, with the aid of some kind of spectrometer, such as a FILTERBANK, CORRELATOR etc.

Traditionally the receiver acquires the radiation from the telescope by means of a feed horn which condenses the free space wave from the telescope down to a waveguide mode. The feed horn is usually of the type known as a 'scalar' feed, which has the property of generating an approximately Gaussian beam of the same width in both dimensions. It does this by means of circular grooves inside the cone which act as a high-impedance boundary. The radiation is then fed from the horn to the mixer element by a short run of rectangular waveguide (see figure 1). The leads providing the current through the mixer will be across the waveguide in the strong electric field region for the TE_{01} mode. Sometimes the mixer element may be matched to the waveguide by mechanical tuners, or better by microcircuit elements, photolithographically fabricated. An alternative scheme for coupling radiation to the mixer element is 'quasioptical', which avoids the difficulties in construction and alignment associated with very high-frequency waveguide. Here a high dielectric constant material plano-convex (hyperhemispheric) lens focuses the radiation to a planar microantenna on a flat substrate fixed to the flat side of the lens. (See figure 2). The planar antenna, which feeds current to the mixer, can take many possible forms, such as bow-ties, spirals, dipoles or slots. The best results have been obtained with twin slots, where the beam profiles are similar to those for scalar feed horns.

The critical element in the receiver, and the one which usually gives the receiver its name, is the mixer. This may be a classical diode, typified by the Schottky barrier elements formed from photolithographically shaped gold dots on gallium arsenide substrates, or a quantum detection device, usually involving a superconductor, such as a Josephson junction or a quasiparticle tunneling junction. These last two devices are based on a superconducting phenomenon, the tunneling of Cooper pairs and single quasiparticles respectively, both of which can occur in the physical structure of a superconductor–insulator–superconductor (SIS) tunnel junction (see DETECTORS: SUPERCONDUCTING TUNNEL JUNCTION DEVICES). For better or worse, the name 'SIS mixer' is now applied to the device in use in nearly all high-frequency RADIO TELESCOPES in the world, which is the single-quasiparticle, photon-assisted tunneling device (an early attempt to introduce the acronym SQPAT failed).

There is a certain similarity between a semiconducting photoconductor device and an SIS device in that, in both cases, an incoming photon causes a transition of an electron (or quasiparticle) from a state below an energy gap to a state above, where conduction occurs. The essential difference is that the semiconductor energy gap is about 0.6 eV, suitable for optical–infrared photon detection, whereas the superconducting gap is $\sim 10^{-3}$ eV, equivalent to a submillimeter photon's energy. This is the reason for choosing SIS devices for frequencies of detection up to the energy gap of the superconductor (about 2.9 meV = 700 GHz for Nb—the most suitable material found to date).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

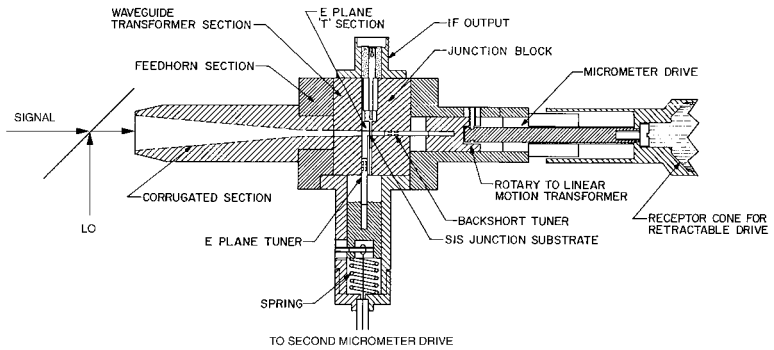


Figure 1. The structure of a typical submillimeter waveguide SIS receiver. The case shown has beamsplitter LO injection, a scalar feed and E-plane and backshort tuners.

Actually, it is not required that the mixer element be a diode-type device; it is only necessary that the mixer element have a non-linear I–V characteristic. The device can even be bulk material and in fact the first successful heterodyne receiver in the submillimeter employed an InSb hot electron bolometer (HEB) element as mixer, in the form of a rod across the waveguide. The problem with such devices is that there is a relatively slow response of the hot electron current due to the weak electron–phonon coupling and, in the InSb receiver case, it is limited to about 1 MHz. Such a narrow-band device would not be considered useful today. Surprisingly, a new style of HEB employing a very small (micron-size) superconducting film is having considerable success and is almost certainly the device of choice for frequencies well above the band gap of niobium.

A measure of the effectiveness of a receiver is the noise temperature ($T_{\rm n}$). That is the temperature of a blackbody (in the Rayleigh–Jeans approximation) which, when coupled fully to the input of the receiver, produces the measured noise at the output. All receivers are governed by the Dicke radiometer equation, where the signal-to-noise ratio is given by:

$$S/N \approx \frac{T_{\rm A}^*}{T_{\rm n}}\sqrt{\Delta\nu\Delta\tau}.$$

$T_{\rm A}^*$ is the Rayleigh–Jeans temperature of an astronomical emission object, $\Delta\nu$ is the channel width which is being integrated and $\Delta\tau$ is the integration time. In general a receiver will have two sidebands [upper $(\omega_2 - \omega_1)$ and lower $(\omega_1 - \omega_2)$]. $T_{\rm n}$ is usually measured by replacing $T_{\rm A}^*$ with known temperature loads and, if both sidebands are equally sensitive, then $T_{\rm n}$ is the double-sideband (DSB) value, relevant to continuum detection. For spectroscopy, where the astronomical line of interest falls in only a single sideband (SSB), the relevant noise temperature to employ in the Dicke formula will be $T_{\rm n}({\rm SSB}) = 2T_{\rm n}({\rm DSB})$.

A non-trivial component of a submillimeter receiver is the source of local oscillator (LO) power, which has to be adequately intense and stable. This becomes an extreme technological challenge when applied to a receiver operating at 1000 GHz (1 THz) or more. The LO power required by a submillimeter receiver varies depending on the type of detector element and the frequency. LO power is coupled to the receiver either by means of a dielectric beamsplitter, which must allow most of the signal to pass and therefore couples only a small fraction of the LO power, or by means of a frequency selective device such as an INTERFEROMETER. A balanced mixer configuration is also a possibility. Roughly speaking, in most cases the LO power needed is the same order of magnitude as the dc power supplied by the bias circuit, because the device is in thermodynamic equilibrium somewhere between the sinks of the dc and LO supplies, to which it is coupled. The typical superconducting device will require a bias voltage of ∼2.5 mV (related to the gap voltage), whereas a semiconducting device needs a bias of ∼0.5 V. Since both devices must have a suitable circuit impedance (∼100 Ω), the superconducting device should require roughly 4 orders of magnitude less LO power, making it the device of choice in this respect. The LO power is usually generated by a series of diode varactor doubler or tripler units multiplying the output of a phase-locked Gunn oscillator.

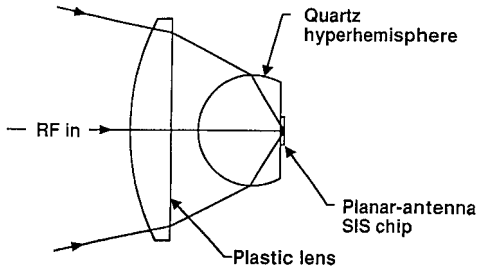


Figure 2. A quasioptical coupling scheme for an SIS receiver. The LO could be injected as in figure 1, but the waveguide components are all replaced with optical systems. The two-lens design shown here can be replaced with a single silicon lens suitably antireflection coated.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

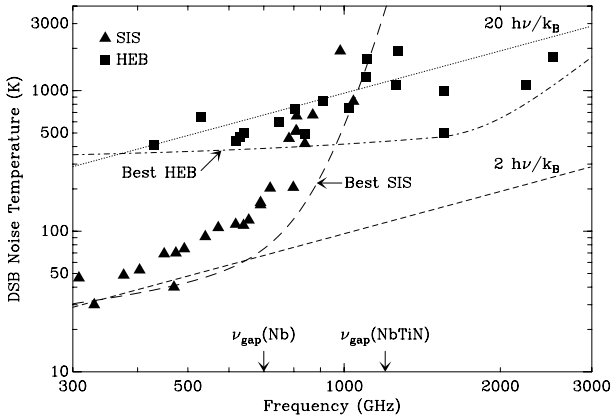


Figure 3. Noise temperatures for SIS receivers (triangles) and HEB receivers (squares) compared with the quantum limit.

The SIS receiver

The SIS receiver is now the workhorse for millimeter and submillimeter observatories. Originally, when the SIS junctions were first introduced they were made from a lead–indium–gold alloy which suffered mechanical and chemical problems and was unreliable. Now the junction elements are usually made from niobium, which is robust and provides a high degree of reliability. It is expected that the order of 1000 such receivers will be constructed for the ALMA project.

Clearly, the requirement for the receiver is that it provides the best S/N in a given observation, so in the Dicke formula $T_{\rm n}$ must be minimized. $T_{\rm n}$, the noise from the receiver, is limited by the quantum noise due to the detection of photons with simultaneous measurement of phase and amplitude. This fundamental noise source provides a minimum noise temperature of $T_{\rm n} \sim h\nu/k$ where h is Planck's constant and k Boltzmann's constant. To this are added excess noises from the inefficiency of the frequency conversion process, noise from the preamplifier and noise due to signal loss in the warm input optics. SIS receivers are able to achieve $T_{\rm n}$ values only a few times the quantum limit throughout the millimeter and submillimeter bands, up to the gap frequency of niobium. Above this frequency a microstrip circuit made from niobium becomes lossy. A plot of the noise temperature of receivers including a set of niobium junction SIS receivers is shown in figure 3.

A second factor in the performance of the receiver is the IF bandwidth which is primarily limited by the achievable bandwidth of low noise cryogenic HEMT preamplifiers. This is currently about 10 GHz but is expected to increase further yet. The third factor is the integration time which can be achieved before the system noise ceases to integrate down with the square root of time. This depends on the care taken in thermal stabilization of components, for example, the integration channel width and the type and rate of switching in use.

In order to extend the low-noise performance achieved below the niobium gap to higher frequencies, there are several possibilities. One is simply to use higher-gap superconducting material for the microstrip circuit. Niobium titanium nitride has been successfully used. However, that solution is limited by the band gap of high-quality superconductors (see figure 3). A completely different superconducting device, which is not limited in frequency by the band gap, is the HEB. In principle this is similar to the older bulk semiconductor device, but has a much shorter response time and therefore should be more useful in a practical device.

The superconducting hot electron bolometer

SIS receivers are the devices of choice for most applications, when they are operating below the gap frequency. For higher frequencies the choices are between SIS receivers with normal metal tuning circuitry, which extends the range for SIS receivers to about 1200 GHz, Schottky diode receivers, which are generally more noisy but operate to several THz if sufficient LO power is available, and superconducting HEB receivers, which also have performance to several THz, but are relatively low noise and require less LO power than do SIS receivers. For astronomy applications such as the ESA–NASA submillimeter space telescope, FIRST, and the US–German airborne telescope project, SOFIA, the choice for frequencies above 1200 GHz may be the HEB.

HEBs use a heating-induced non-linearity in a small superconducting film held near the critical temperature, T_c. The electronic system is heated above the lattice temperature by dc and rf power. The device will act as a mixer as long as the IF is restricted to the regime below the hot electron relaxation rate, otherwise the electron temperature cannot follow the power variation at the IF. There are two types of superconducting HEB, the first depending on the electron–phonon interaction for cooling the hot electron system and the second using the diffusion process to carry the excess thermal energy out of the superconductor and into the normal metal leads. Both techniques appear to work well, with the diffusion type able to sustain IF bandwidths of about 10 GHz for niobium films. The plot of figure 3 includes the HEB results for very high frequencies.

Bibliography

The first astronomical submillimeter heterodyne spectroscopy employed the InSb hot electron bolometer

Phillips T G and Jefferts K B 1973 *Rev. Sci. Instrum.* **44** 1009

However, the development of SIS receivers came partly as a result of the intense effort in the late 1970s to develop superconducting switching devices for computers, using the Josephson effect. Although this was never successful for computing, it provided materials and a fabrication technology for making small-area superconducting tunnel junctions. The first SIS detectors were described in adjoining papers in 1979 by Berkeley and Bell Telephone Labs groups working independently:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1979 *Appl. Phys. Lett.* **34** 345, 347

Interpretation of the performance was helped a great deal by a quantum theory for quasiparticle tunneling devices by

Tucker J R 1979 *IEEE J. Quantum. Electron.* **15** 477

published shortly thereafter. Subsequent milestones have been the initial attempts at planar-antenna, quasioptical structures

Wengler M J, Woody D P, Miller R E and Phillips T G 1985 *Int. J. Infrared Millim. Waves* **6** 697

improved on subsequently by a twin-slot design

Gaidis M C, LeDuc H G, Bin M, Miller D, Stern J A and Zmuidzinas J 1996 *IEEE Trans. Microw. Theory Tech.* **44** 1130

Superconducting HEBs were first considered for mixers in the phonon cooling mode by

Gershenzon E, Gol'tsman G, Gogidze I G, Gusev Y P, Elant'ev A I, Karasik B S and Semenov A 1990 *Superconductivity* **3** 1582

and in the diffusion cooling style by

Prober D E 1993 *Appl. Phys. Lett.* **62** 2119

A detailed description of the scalar, corrugated feed horn is given by

Clarricoats P J B and Olver A D 1984 *Corrugated Horns for Microwave Antennas (IEE Electromagnetic Wave Series 18)* (Peter Peregrinus)

A comprehensive modern review article is that by

Carlstrom J and Zmuidzinas J 1996 Millimeter and submillimeter techniques *Review of Radio Science 1993–1996* ed W Ross Stone (Oxford: Oxford University Press) pp 839–82

Thomas Phillips

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Red Dwarfs/Flare Stars

The red dwarf stars are main-sequence stars characterized by masses in the approximate range from $0.51M_{\odot}$ to $0.08M_{\odot}$. The upper mass limit corresponds to a spectral type of M0 V and the lower limit is the critical mass for the onset of nuclear burning in the central cores of stars with chemical abundances similar to the Sun. Below the range of approximately $(0.074–0.08)M_{\odot}$ is the realm of substellar objects known as *brown dwarfs*. The radii of the red dwarfs span from $0.60R_{\odot}$ to $\approx 0.18R_{\odot}$ while their effective temperatures are in the approximate range 2500 K–4000 K. Thus, the red dwarf stars are cooler, smaller and less massive than the Sun. Correspondingly, their luminosities are low relative to the Sun, ranging from less than a few tenths of a per cent (or less) to about 8% of the solar luminosity. The flare stars are observationally defined as those red dwarfs that exhibit transient brightenings on time scales of seconds to minutes as seen in the visible region of the electromagnetic spectrum. These short-lived brightenings are considered to be qualitatively similar to SOLAR FLARES.

The red dwarfs represent the largest constituent of the STELLAR POPULATION, composing at least 80% of the stars in the Galaxy. Therefore, the typical star in the Galaxy is smaller, less massive and cooler than the Sun. From this perspective, the Sun is not a common star. Rather, the Sun and Sun-like stars are actually atypical compared with the more numerous red dwarf stars. The large mass-to-luminosity ratios of the red dwarfs combined with their high flare activity and relatively large space density suggest that these objects could play an important role in galactic dynamics.

Within K–M spectral types, the red dwarfs are divided into subclasses according to the appearance of certain spectral features. In particular, the dMe stars are those red dwarfs that exhibit Hα line emission. It is nearly always the case that a given dMe star is also a flare star. Those red dwarfs without Hα in emission are non-dMe stars or, simply, dM stars. The spectrum of a dM star exhibits either Hα in absorption or no apparent Hα feature. Flare activity has only rarely been observed in dM stars, and these objects are usually not identified as flare stars.

Interior and atmospheres

The interior structure of red dwarf stars is characterized by a thick outer convection zone relative to the radius of the star. STELLAR INTERIOR models suggest that at a mass of $0.4M_{\odot}$, the depth of the convection zone is nearly 60% of the stellar radius. At masses less than approximately $0.30M_{\odot}$, the interiors of red dwarfs are completely convective. The photospheric spectra of red dwarfs are dominated by molecular features. Molecular bands of TiO and, for the latest types, VO are strongly present together with atomic lines of Ca II, Na I and K I. The progression from K to M dwarfs is characterized by increasing strength of TiO absorption. The coolest M dwarfs (later than M7 V) exhibit prominent absorption by VO. Bands of MgH are visible in K through early M dwarfs ($\sim$ M3.5) after which they become dominated by overlying TiO. Bands of CaOH at 553–557 nm and near 625 nm begin to appear at spectral type M3.5. These bands become especially prominent toward the latest M dwarfs (at least as late as M8) where they remain visible through the thicket of TiO lines. The CaH B band at 638.5 nm is also sensitive to temperature, first appearing at spectral type M0 and remaining identifiable through the latest M spectral subclasses. Atomic lines of Fe I and Ca II that are strong in the K through mid-M dwarfs disappear in the coolest M dwarf stars as molecular absorption by TiO and VO dominates. The appearance of the red spectrum of these cool dwarfs is illustrated in figure 1.

The red dwarfs can exhibit periodic or quasi-periodic behavior in their photometric light curves. This is due to the rotational modulation of large, cool spots analogous to SUNSPOTS, on the surfaces of these stars. As rotation carries a spotted region on the star through the line of sight, a dip in the photometric light curve is observed. Repeated observation of this phenomenon yields the rotation period of the star, independent of the inclination of the rotation axis to the line of sight. The rotation rates of red dwarfs are correlated with their activity in the form of flares, emission lines and x-ray emission from a CORONA. Red dwarfs with rotation rates exceeding 3–5 km s^{-1} are generally more active and tend to be classified as dMe stars. Those objects with slower rotation rates are non-dMe or, simply, dM stars. These red dwarfs have lower levels of magnetic-field-related activity and no Hα emission.

All red dwarfs thus far observed with sufficient sensitivity and spectral resolution exhibit Ca II H and K line emission, indicative of the presence of CHROMOSPHERES. In addition to Ca II resonance line emission, ultraviolet spectra obtained with satellite observatories have revealed Mg II h and k line emission as well. Emission from these metal species together with Hα emission (as seen in the dMe stars) establishes the presence of significant nonradiative heating in the red dwarfs that is presumed to be related to surface magnetic fields. Interestingly, chromospheric models have shown that Hα absorption is also a result of chromospheric heating in the non-dMe (dM) stars. The photospheres in these red dwarfs are so cool that, in the absence of a chromosphere, only weak Hα absorption would be observed in their spectrum. However, as chromospheric heating increases, the population of hydrogen atoms in the $n = 2$ quantum level increases, thus leading to increasing opacity in the Hα line to initially produce an absorption feature. From the perspective of stellar chromospheric activity, this result suggests that the principal difference between dMe and dM stars is, in part, due to differing intrinsic heating rates that are present in their respective chromospheres.

The far-ultraviolet spectrum of active red dwarfs is dominated by emission lines arising from high-temperature ($T > 50\,000$ K) plasma in the so-called transition region of the outer atmosphere. Distinctive spectral features include lines arising from C IV, Si IV

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

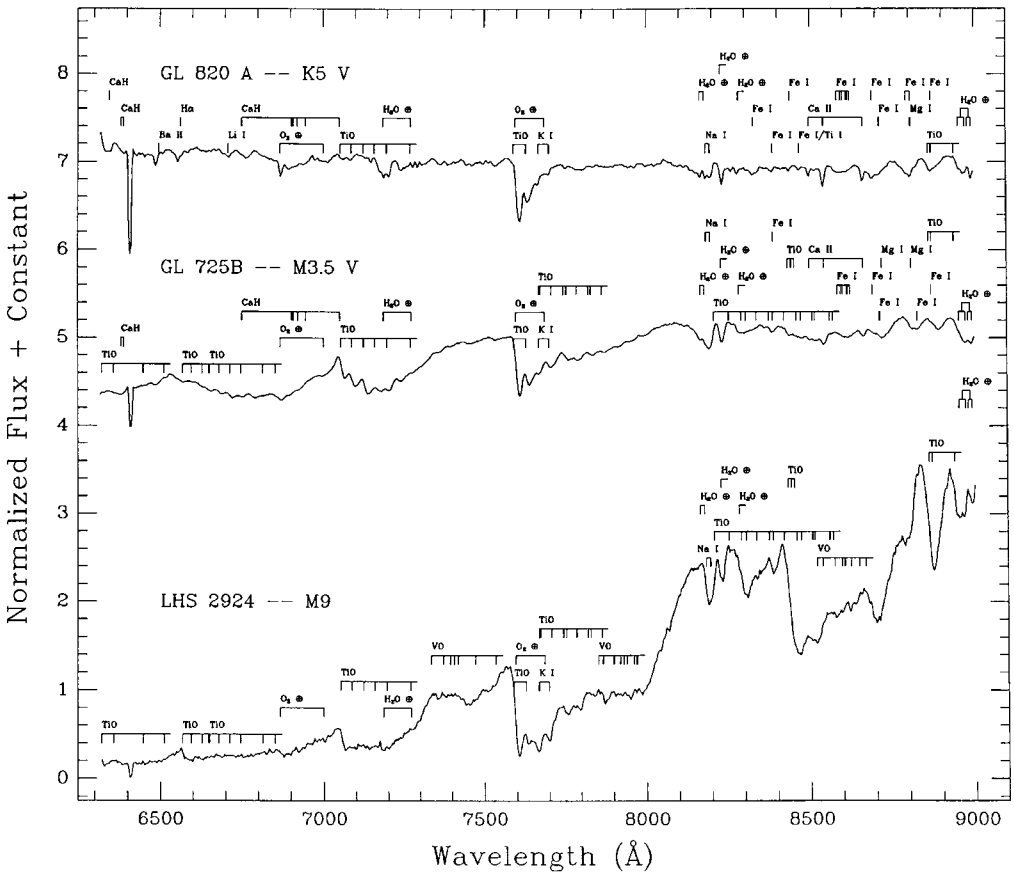


Figure 1. Photospheric spectra of red dwarfs. Illustration of the principal features in the red spectrum of cool dwarf stars extending from spectral types mid-K to M9. Molecular bands of titanium oxide (TiO) begin to dominate the spectrum in the M dwarfs while, in the coolest dwarfs, vanadium oxide (VO) becomes prominent. Other features include CaH and neutral metal lines of iron (Fe), magnesium (Mg) and sodium (Na). Terrestrial lines due mainly to atmospheric water are indicated with a ⊕ symbol. (From Kirkpatrick J D, Henry T J and McCarthy D W Jr 1991 *Astrophys. J. Suppl.* **77** 417.)

and N V. The surface fluxes of these high-temperature lines in the dMe stars are 10–100 times greater than the corresponding line surface fluxes in the quiet Sun. The far-UV spectra of the less active red dwarfs (dM stars) are, by contrast, dominated by lines arising from neutral metal species and other low-temperature features. An illustration of the far-ultraviolet spectra of active and inactive red dwarfs is given in figure 2.

X-ray emission and coronae in red dwarfs

The red dwarfs are x-ray sources with levels of detected x-ray emission that range in luminosity from ∼0.1 to ∼1000 times that of the value of the quiet solar corona (∼10^{27} erg s^{-1}). The absolute x-ray luminosity of red dwarfs declines toward lower masses. However, the x-ray luminosity normalized to the bolometric luminosity of the star, which is often expressed as L_x/L_{bol}, can persist among the active (dMe) stars at the so-called saturation level of approximately $L_x/L_{bol} \simeq 10^{-2.5}$ to spectral types at least as late as M7. Thus, the active red dwarfs are characterized by relatively high coronal heating efficiencies compared with that of the active Sun (i.e. the Sun as viewed in the x-ray at the maximum of the solar activity cycle) where, for the active Sun, we have that $L_x/L_{bol} \sim 10^{-6}$. If there is any sharp decline in coronal heating efficiency, as measured by the parameter L_x/L_{bol}, then it occurs at spectral subtype M8. However, any possible decline in coronal heating efficiency cannot be verified until more red dwarf stars of subtypes M8 and M9 are discovered and observed with an x-ray telescope such as the new Chandra x-ray satellite observatory.

Analysis of the x-ray results shows that those red dwarfs that are more luminous in the x-ray have higher average coronal temperatures. Red dwarfs that exhibit little chromospheric activity (i.e. dM and dK stars) do indeed have cooler and weaker coronae than the more active (dMe) stars. Detailed model results based on x-ray observations reveal that the coronae of the low-mass dwarfs consist of two distinct thermal components: a so-called 'soft' component with temperatures $T \sim$ (2–4) $\times 10^6$ K and a 'hard' component with $T \sim 10^7$ K.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

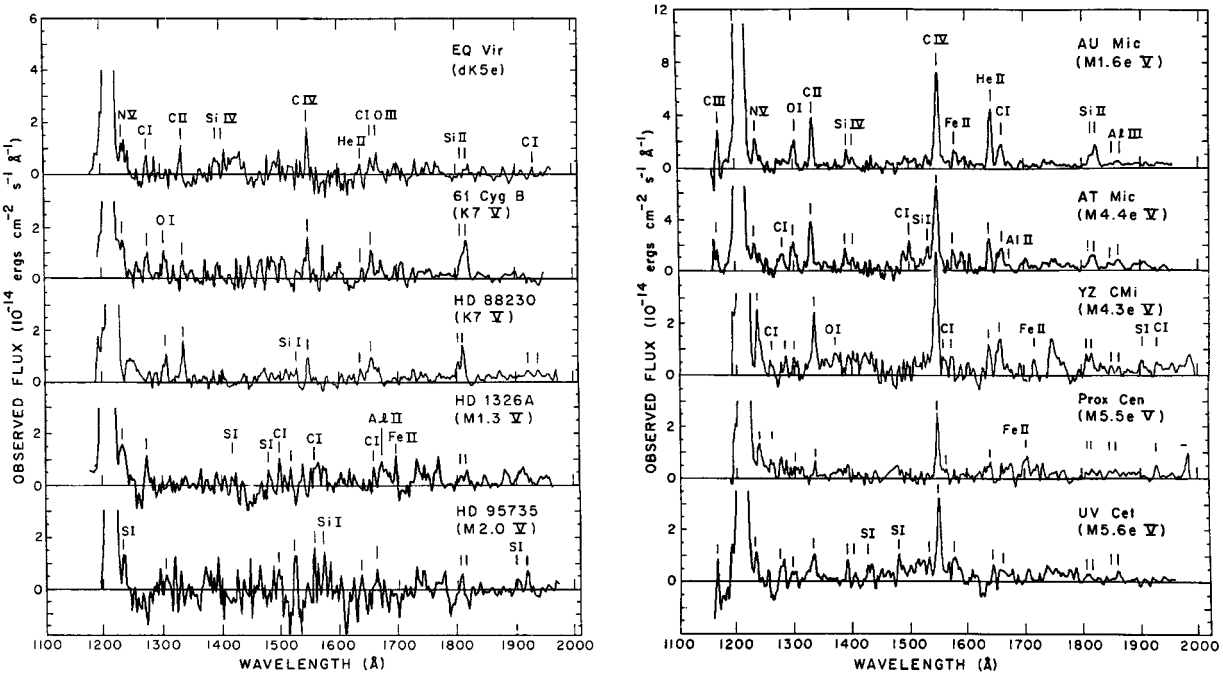


Figure 2. Far-ultraviolet spectra of red dwarfs and flare stars. The far ultraviolet spectra of active and inactive red dwarf stars as obtained with the International Ultraviolet Explorer satellite observatory. The identification of emission lines is indicated. The active red dwarfs (denoted as types dKe or dMe) exhibit emission features from highly ionized species, indicative of hot plasma. By contrast, the far UV spectra of the inactive (dK or dM) red dwarfs are dominated by emission from low-temperature species thus implying that a significantly lower density of hot plasma is present in their outer atmospheres. (From Linsky J L, Bornmann P L, Carpenter K G, Wing R F, Giampapa M S, Worden S P and Hege E K 1982 *Astrophys. J.* **260** 670.)

Based on the observational and model results, the author and his collaborators concluded that coronal emission in active red dwarfs (dMe stars) can be interpreted as arising from quiescent active regions (a quiescent, low-temperature component) and compact flaring structures (variable, high-temperature component).

Active red dwarfs are also radio sources. The quiescent component of radio emission is attributed to gyrosynchrotron radiation from electrons with energies of 20–200 keV. Surveys of both radio and x-ray emission reveal that the x-ray and radio luminosities are directly correlated over a range of 3 orders of magnitude, irrespective of spectral type. This result strongly suggests that there is a physical relation between the particle populations that give rise to the x-ray and to the radio emissions.

Underlying the wealth of activity that is observed in red dwarfs are magnetic fields. Emergent magnetic flux as a result of interior dynamo action is presumably directly related to the origins of spotted regions, flares, chromospheres and coronae, as is seen on the Sun. Direct spectroscopic detections of surface magnetic flux on active red dwarf (dMe) stars reveal field strengths in the broad range of 2–4 kG covering over 50% of the star. These kinds of field strengths are consistent with a model consisting of magnetic flux tubes in equilibrium with the surrounding photospheric gas pressure. Direct measurements of magnetic field properties in the relatively inactive red (dM) dwarfs have not yet been accomplished.

Flare activity in red dwarfs

The vast majority of observations of flare outbursts in red dwarf flare stars have been recorded in broad-band photometry, typically with Johnson UBV filters. During a flare, the U band exhibits a larger relative change than the B band which, in turn, shows a larger relative change than the V band. The blue color of the optical flare energy distribution is due both to enhanced UV emission lines and to a continuum suggestive of an optically thick, radiating plasma which is much hotter than the underlying red dwarf photosphere. The continuum emission in stellar flares is consistent with free–free emission at temperatures of (5–10) $\times$ 10^4 K, bound–free emission at a temperature of 2 $\times$ 10^4 K, or a combination of both processes. A black-body component due to heating of the lower stellar chromosphere and upper photosphere with a characteristic effective temperature of $\sim 10^4$ K may also contribute to the flare continuum emission. Based on the assumption that the flare continuum emission is predominantly composed of black-body radiation, effective areas of the flaring region are derived that correspond to a range from 0.5% to 2% of the stellar disk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

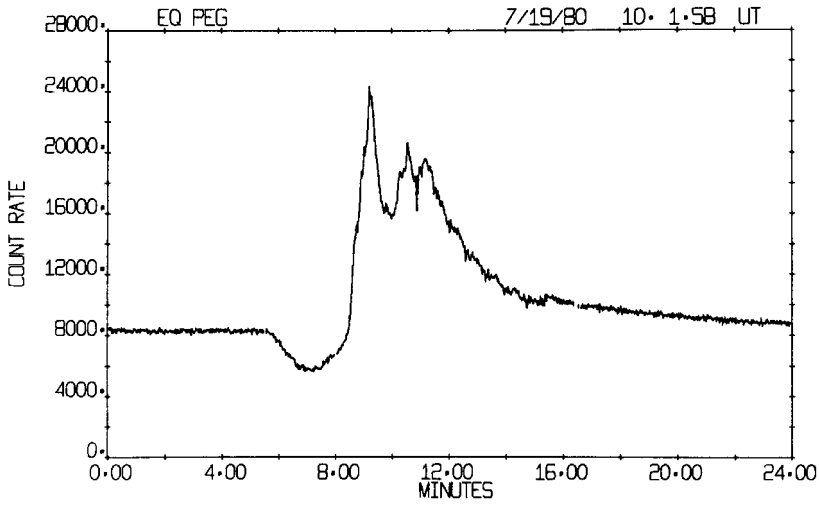


Figure 3. Flare event on a red dwarf flare star. A particularly striking example of the variety of phenomena that can be seen in stellar flares is encapsulated here in a single, extraordinary event observed on the binary flare star EQ Pegasi (dM3.5e + dM4.5e). This event was observed at the Cloudcroft Observatory 1.2 m utilizing high-speed photometry in the Johnson U band filter. The flare is preceded by a 'pre-flare dip', i.e. a decrease in the U band brightness of the star, in this case, of roughly 0.1 mag min^{-1} for 2.7 min. The signal then leveled off at 75% of the quiescent brightness for 1 min before the flare begins. The peak of the flare is 3 times the quiescent level brightness, and is followed by two more smaller peaks separated by 60 and 30 s. The flare event then decays exponentially in brightness, finally returning to the pre-flare level 19 min after the onset of the event. (From Giampapa M S, Africano J L, Klimke A, Parks J, Quigley R J, Robinson R D and Worden S P 1982 *Astrophys. J. Lett.* **252** L39.)

The total energy released during a stellar flare event can be as much as 2–3 orders of magnitude greater than the largest solar flares.

Photometric observations of stellar flares display a wealth of variety in temporal profile and color. Generally, however, the flares are divided into two categories: 'spike flares' typically attain their peak intensity in less than a minute while 'slow flares' take several minutes before reaching maximum. The initial decay toward quiescence is rapid, falling below half-maximum in several minutes. In most flares the rate of decay quite suddenly slows and then can take several hours to relax fully to quiescence. Large flares typically exhibit the superposition of slow and fast events that may be the result of several related flares, or a 'wavelike' flaring in the stellar active regions. A distinctly non-solar-like phenomenon that occasionally occurs in flare stars is a 'pre-flare dip'. A particularly striking example of this sudden diminution in the light from the star, followed by a flare, is shown in figure 3.

Time-resolved, optical spectroscopic observations of red dwarf flares reveal that a strong continuum rises to dominate the violet and blue region during maximum light. The Balmer series is seen out to H14, Ca II H and K emission, and He I emission lines all rise with the continuum, but usually peak a few minutes afterwards. The rapid decline of the continuum is responsible for the initial photometric decay, while the emission lines slowly return to quiescence. The emission lines can remain at enhanced levels for hours following the initial flare outburst and decline. A sequence of spectra obtained during a flare on the prototype flare star, UV Ceti, is displayed in figure 4.

High-resolution line profiles obtained during flares can show evidence of significant mass motions with blueshifted emitting material rising at $\sim$70 km s^{-1} in one case, and perhaps much higher velocities in other cases. In brief summary, the qualitative picture of the flare would include upward-moving material filling a large volume (with a characteristic length scale $\sim$0.1 stellar radius) with hot plasma at chromospheric densities of $n_e \sim 10^{11-12}\,\text{cm}^{-3}$. The flare volume produces line emission that is strongly enhanced in total flux compared with the pre-flare emission but is often otherwise similar in basic line profile shape to the profile in the quiescent chromosphere (apart from components that can be attributed to bulk motions in the line of sight). The flare luminosity can be a few per cent of the quiescent stellar bolometric luminosity which is an order of magnitude greater than the corresponding ratio for either chromospheric or coronal heating. Hence, the conversion of flare energy to radiative flux is significantly more efficient than the processes that control coronal and chromospheric heating in red dwarfs and flare stars.

Observations of stellar flares in the x-ray reveal that the total energy in soft x-rays is often comparable with that in the largest solar flares. Stellar flare x-ray light curves morphologically resemble the x-ray light curves observed for compact and two-ribbon flares that occur on the Sun. Very recent observations by the author and his collaborators using the ROSAT x-ray satellite observatory

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

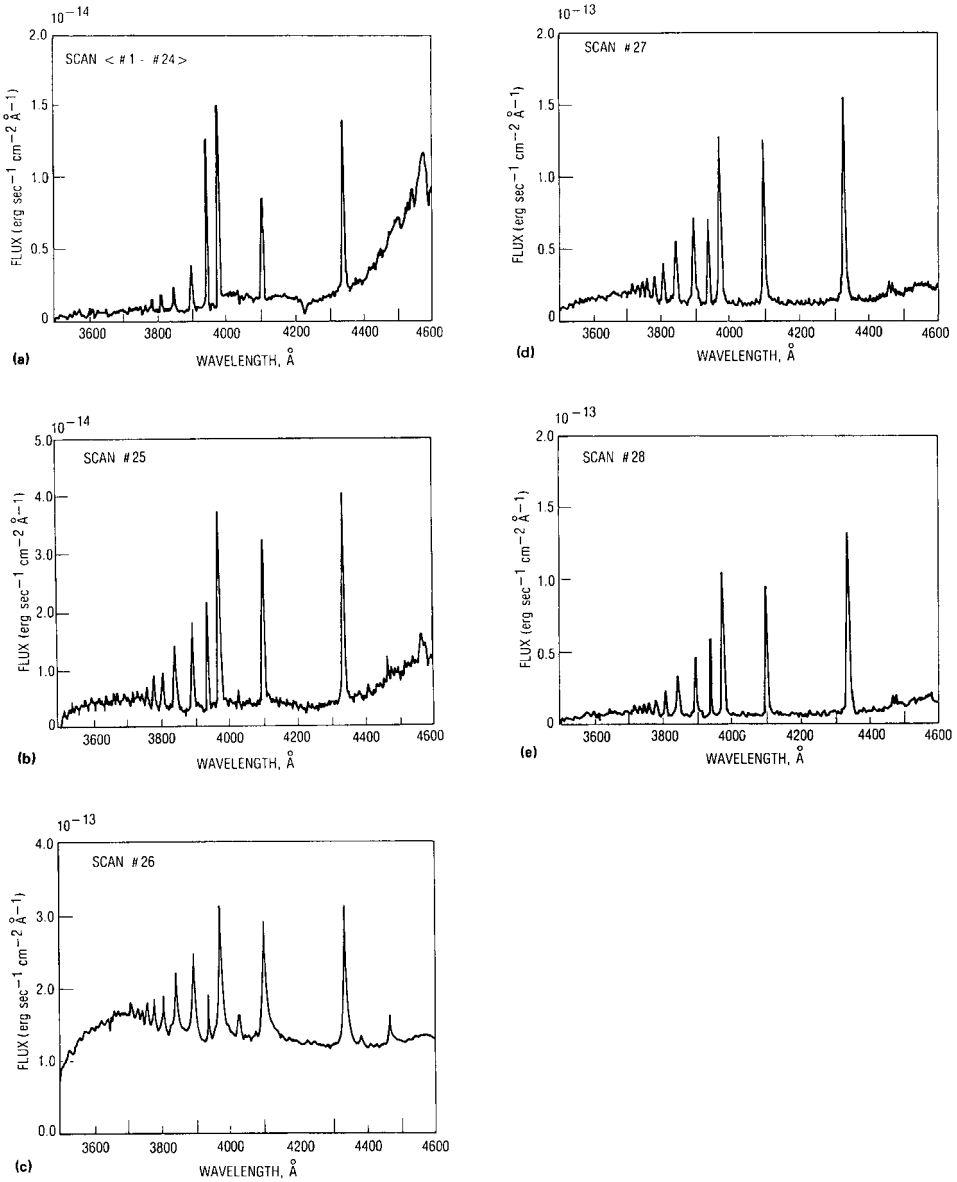


Figure 4. Spectrum of a large flare event. Time sequence of the blue spectrum of the flare star UV Ceti (dM5.6e) during the onset of a strong flare that exhibited a 5 mag increase in the photometric *U* band. The first scan is the average, quiescent spectrum of the star. The Ca II H and K lines are strong, and the Balmer series is evident out to H14, even during quiescence. The following scans document the onset of the flare event. Note the varying vertical axis scales. Emission lines from hydrogen are the most numerous and prominent emission features during the flare itself. Strong continuum emission is seen at the peak of the event in scan 26. Note the red asymmetry that appears in the emission lines at flare maximum, indicating the possible presence of mass motions. These spectra were originally obtained with the 2.3 m telescope and Reticon detector of the Steward Observatory of the University of Arizona. (From Eason E L E, Giampapa M S, Radick R R, Worden S P and Hege E K 1992 *Astron. J.* **104** 1161.)

disclosed the occurrence of an x-ray flare on the very-low-mass star VB 10 (dM8e). VB 10 is the latest-type, lowest-mass main-sequence star known to exhibit coronal activity. The peak luminosity of the flare was $\sim 10^{27}$ ergs s^{-1}. This is at least 2 orders of magnitude greater than the quiescent x-ray luminosity of VB 10, which has yet to be measured.

Concluding remarks

The manifestations of magnetic-field-related activity in the red dwarfs and flare stars are especially vivid because of the sharp contrast between the high-temperature emission features, arising from nonradiatively heated plasma at temperatures extending from $\sim 10^4$ to 10^7 K, against the cool photospheres with effective temperatures in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

~2500 K–4000 K range.

The origin of the surface magnetic field itself in the Sun and cool stars is ascribed to interior dynamo action involving the interaction between rotation, convection and extant magnetic fields. Empirically, coronal x-ray emission and chromospheric line emission strengths are correlated with rotation in red dwarfs. Magnetic field measurements by S Saar suggest an inverse relationship between rotation period and the area coverage, f, of magnetic flux of the form $f \sim P_{\rm rot}^{-1.8}$ for rotation periods greater than 3 days. At more rapid rotation periods, the surface is saturated with magnetic fields. However, the total magnetic flux, i.e. the product of the field strength and fractional area coverage, continues to increase with more rapid rotation rates according to the preliminary, empirical relation $fB \propto P_{\rm rot}^{-1.7}$. The implication is that magnetic field strengths continue to increase with increasing rotation rates, perhaps forming dark starspots.

In the Sun, the magnetic field that eventually appears on the surface in the form of plages, spots and flares is widely considered to have originated in an interface region between the outer convection zone and the interior radiative core. While this dynamo model may indeed apply to cooler dwarfs, it is the regime of the red dwarfs where the transition from a partial to a fully convective interior occurs. Thus, in the absence of an interface layer between the radiative and convective regions, another dynamo mechanism must then operate to continue to give rise to the variety of magnetic activity that persists into the wholly convective red dwarfs and flare stars. The nature of this new dynamo mechanism is a key question in current research efforts involving red dwarfs.

The physical mechanism that leads to the spectacular flare events in these cool stars is also unknown but is believed to involve, as in the Sun, magnetic reconnection. The energetics of stellar flares can often be accounted for within the context of two-ribbon flare models. Nevertheless, neither the actual flare morphology nor the true plasma-heating mechanism is known in the case of stellar flare events. Large stellar flares, like their solar counterparts, probably represent the superposition in space and time of many evolving flare regions in a flaring complex that could encompass a geometrically extended, large volume of emitting plasma.

Other key questions involving red dwarfs and flare stars include the possible contributions by these active objects to the diffuse galactic x-ray background, the evolutionary relationship between the active dMe flare stars and the relatively quiescent dM stars; whether red dwarfs exhibit cycles in their activity analogous to the solar cycle, the relationship of red dwarfs to brown dwarfs and the frequency of extrasolar planet formation among the low-mass stars. As always in astronomy, advances in instrumentation and analysis techniques will yield new insights on these common yet fascinating stars.

Bibliography

Bastian T S 1990 Radio emission from flare stars *Sol. Phys.* **130** 265–94

Delfosse X, Forveille T, Perrier C and Mayor M 1998 Rotation and chromospheric activity in field M dwarfs *Astron. Astrophys.* **331** 581–95

Fleming T A, Schmitt J H M M and Giampapa M S 1995 Correlations of coronal x-ray emission with activity, mass, and age of the nearby K and M dwarfs *Astrophys. J.* **450** 401–10

Giampapa M S 1983 Results from optical and UV stellar flare spectroscopy *Activity in Red-Dwarf Stars* ed P B Byrne and M Rodonò (Dordrecht: Reidel) pp 223–35

Giampapa M S, Rosner R, Kashyap V K, Fleming T A, Schmitt J H M M and Bookbinder J A 1996 The coronae of low-mass dwarf stars *Astrophys. J.* **463** 707–25

Henry T J and McCarthy D W Jr 1993 The mass–luminosity relation for stars of mass 1.0 to 0.08$M_\odot$ *Astron. J.* **106** 773–89

Johns-Krull C M and Valenti J A 1996 Detection of strong magnetic fields on M dwarfs *Astrophys. J. Lett.* **459** L95–8

Kirkpatrick J D, Henry T J and McCarthy D W Jr 1991 A standard stellar spectral sequence in the red/near-infrared: classes K5 to M9 *Astrophys. J. Suppl.* **77** 417–40

Mullan D J 1976 Flare stars: I. Physical characteristics of the stars *Irish Astron. J.* **12** 161–82

Schmitt J H M M, Fleming T A and Giampapa M S 1995 The x-ray view of the low-mass stars in the solar neighborhood *Astrophys. J.* **450** 392–400

Stauffer J R and Hartmann L W 1986 The chromospheric activity, kinematics, and metallicities of nearby M dwarfs *Astrophys. J. Suppl.* **61** 531–68

Turnshek D E, Turnshek D A, Craine E R and Boeshaar P C 1985 *An Atlas of Digital Spectra of Cool Stars* (Tucson, AZ: Western Research Company)

Mark Giampapa

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Red Giant Stars

Most stars are less massive and less luminous than the Sun, yet most of the visually bright stars are considerably more luminous than the Sun. Many of these are also cooler than the Sun, so that their higher luminosity indicates a much greater diameter, although they are not necessarily more massive. These are then known as giant stars, whereas the Sun is a dwarf. This difference in structure results from their being at different stages in their evolution (see STELLAR EVOLUTION). Dwarf stars derive their energy from hydrogen burning in the core, whereas giant stars have already exhausted hydrogen in the core, although they may still burn hydrogen in a surrounding shell, and may burn helium or even heavier elements in the core.

Stars are classified according to their surface temperature which dictates the degree of ionization and excitation of the atoms in the visible outer layers of the atmosphere. The main classes, from hot to cool, are denoted as O, B, A, F, G, K, M. Classes L and T have been added recently to describe very cool dwarf stars. There are also giant stars with peculiar surface composition. The R and N stars, which are sometimes lumped together as C, are carbon rich. The S stars are those in which the TiO molecule, which is the defining characteristic of the M-type spectra, is joined and ultimately superseded by ZrO and other molecules containing elements produced by the nuclear reaction known coincidentally as the s-process.

Giants and dwarfs may be distinguished by subtle spectroscopic differences, notably the greater strength of singly ionized metals in giants at any given temperature. The basic difference in luminosity is determined by trigonometric PARALLAXES which are based on observing the tiny angular movement relative to more distant stars which is produced by the Earth's annual motion relative to the Sun. The recent HIPPARCOS satellite measured the trigonometric parallaxes and hence distances of many stars to unprecedented accuracy. Giant and dwarf stars may also be present in a single cluster of stars all at the same distance, and then the difference in luminosity is observable directly.

Most of the giant stars which are visible to the naked eye or are seen in clusters such as the Hyades and Praesepe are of spectral type K and appear yellow to the eye. The true red giants are cooler, of spectral type M, S or N, and are much more luminous and represent the tip of the giant branch. They show a variety of interesting phenomena, including altered surface chemical composition, variability and mass loss.

The development of the technology for the efficient detection of infrared radiation has opened up a new field in recent years. Many red giant stars have been found to have extensive dust shells which, radiating at a temperature of a few hundred kelvin, are only detectable in the infrared. Some of these stars have such thick dust shells that the star itself is invisible at optical wavelengths, and the INFRARED ASTRONOMY SATELLITE (IRAS) which was equipped with a small infrared-sensitive telescope discovered thousands of stars whose existence was unknown to ground-based astronomers.

The giant branch

The recognition that the cooler stars are clearly divided into two groups, large bright giants and small faint dwarfs, was one of the principal results of the first large surveys of trigonometric parallaxes. This result is encapsulated in the HERTZSPRUNG–RUSSELL (HR) DIAGRAM, named for its independent discoverers, in which a measure of luminosity is plotted against a measure of temperature—originally spectral type but usually a photometric color—which increases to the left. Figure 1 shows

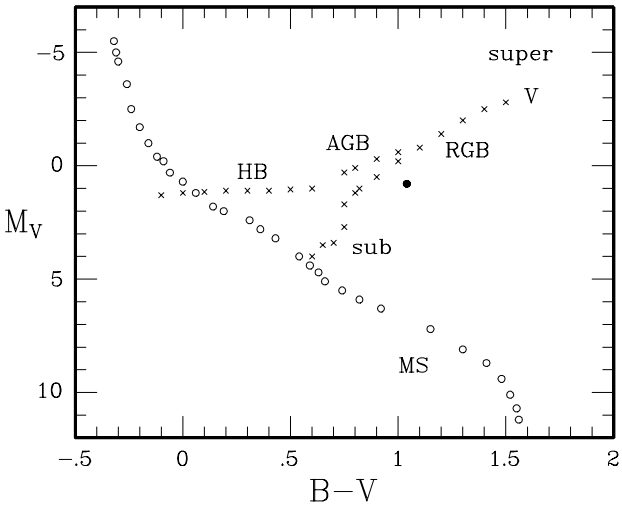


Figure 1. The HR diagram in the now usual color-magnitude form. The main sequence (MS) of H-burning dwarf stars slants from upper left to lower right (open circles). The subgiant branch (sub), red giant branch (RGB), horizontal branch (HB) and asymptotic giant branch (AGB) of a typical metal-deficient globular cluster are denoted by crosses. The AGB and RGB are only clearly separable at relatively low luminosity; it is a matter of inference that the most luminous red giants and especially those with altered surface composition belong to the AGB and not to the RGB. The V at the tip of the RGB–AGB indicates the position of red variable stars in the cluster. The filled circle to the right (red) of the RGB is the approximate position of the red giant clump in the more metal-rich stars of the solar neighborhood, while the RGB of this population lies just to the red of the clump (figure 2). Note that this diagram is based on the visible spectrum only and the hottest and coolest stars would appear relatively brighter if their entire luminous output, the bolometric luminosity, were considered. Conversely, the brightest stars of intermediate (solar) surface temperature fall above the upper limit of this diagram and the cool red variable stars of high metal content are visually much fainter than those in globular clusters denoted by V. The massive red supergiants of the younger population (super) fall in the top right of the diagram.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

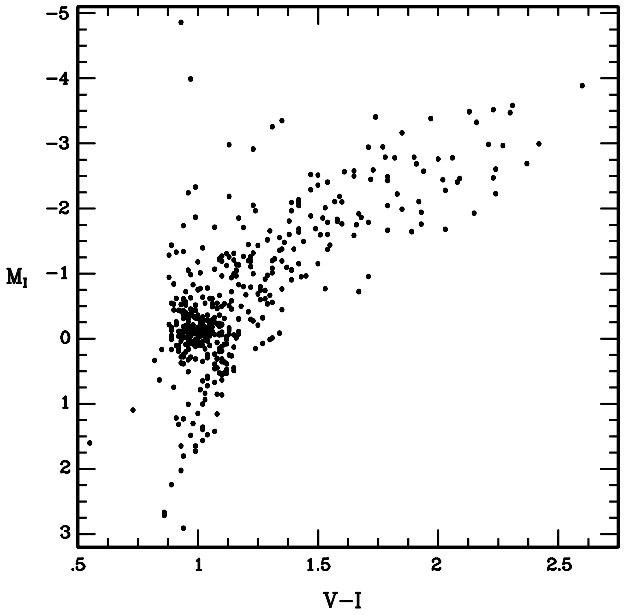


Figure 2. The color-magnitude diagram for nearby K and M0–M3 giants with well-determined parallaxes from the Hipparcos Catalog. The choice of *V-I* instead of the more usual *B–V* as abscissa uses TiO to expand the temperature scale for the cooler stars. The red giant clump is prominent near *M*I=0, *V–I*=1.0, and the red giant and asymptotic giant branches run together to higher luminosity (i.e. upward) and lower surface temperature (to the right). The few stars which lie above this are probably more massive stars. The RGB and AGB are not separable in this mixed population of stars from the solar neighborhood.

a composite color-magnitude diagram for dwarf stars near the Sun and giant stars in a typical GLOBULAR CLUSTER, using representative points along the corresponding sequences. This figure illustrates the terminology we introduce in this section and use throughout the article. It follows from the difference in luminosity that the coolest giants shown have radii which are several hundred times those of dwarfs of similar effective temperature.

Dynamical studies of BINARY SYSTEMS show that a typical giant star has a mass in the range 1–3 times that of the Sun so that the difference in radii between giants and dwarfs results from a difference in structure rather than of mass. The mean density of the Sun is close to that of water but that of a giant star may be more than a million times less. This matter is not evenly distributed, however, because while the Sun has a ratio of the central to the mean density of about 50:1 the corresponding figure for a giant can be as large as 10^8:1. The giant has a very small dense energy-generating core which contains over 10% of the mass of the star.

Observations of stars in clusters, which have the great advantage that all the stars lie at almost the same distance, so that their apparent luminosity gives their true relative luminosity, and usually have the same age and chemical composition, revealed the detailed structure of the giant branch. There is a marked concentration of stars at a luminosity about 40 times that of the Sun and slightly to the blue (hot) side of the red giant branch (RGB). This takes the form of a compact clump, the red giant clump, in open star clusters of high metal content and in the stars of the solar neighbourhood.

The red giant clump, whose location is marked in figure 1, is shown in more detail in figure 2. Figure 2 shows the color-magnitude diagram for a sample of nearby stars, based on high-quality trigonometric parallaxes obtained with the Hipparcos satellite. The early K stars are strongly concentrated in the red giant clump, while cooler K and early M stars extend to redder colors and higher luminosity; the intrinsically fainter stars are under-represented because star selection is by apparent magnitude. The clump becomes the horizontal branch, extending to much higher surface temperature, in the old metal-deficient globular clusters. The luminosity of the clump is very similar to that of the horizontal branch, as is well seen in the HR diagrams of nearby galaxies where both populations are present. The extension of the clump to the upper left in figure 2 represents stars which are substantially more massive then the majority.

The originally single giant branch becomes double at higher luminosities than that of the clump. This is not apparent in figure 2, where observational error and the range of mass and chemical composition among the nearby stars obscure the finer details. Observations of globular clusters showed that the giant branch extends as a sharp continuous locus from the subgiant branch which follows from the main sequence turnoff to the tip at a luminosity about 1000 times brighter. This is joined by a second giant branch which starts above the horizontal branch, displaced to the hot (blue) side of the RGB, and runs progressively closer to the RGB until they become indistinguishable. This second giant branch is known as the asymptotic giant branch or AGB. This is shown in figure 1 and in more detail for the giant stars in the moderately metal-poor globular cluster 47 Tucanae in figure 3.

Theoretical calculations have been able to reproduce these main features of the observed HR diagrams with some precision, yet the fundamental reason for the change in structure from the dwarf to the giant state has been hard to establish. Recent work shows that the expansion of the envelope following increasing luminosity in the core causes cooling, ion recombination and an increase in the opacity of the envelope. The inability of the envelope to transmit all the luminosity impinging on its base causes further expansion which exacerbates the situation, leading to a runaway expansion which only stops when the entire envelope becomes convective and energy can be transmitted to the surface by this more efficient process. The locus of a convective envelope in the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

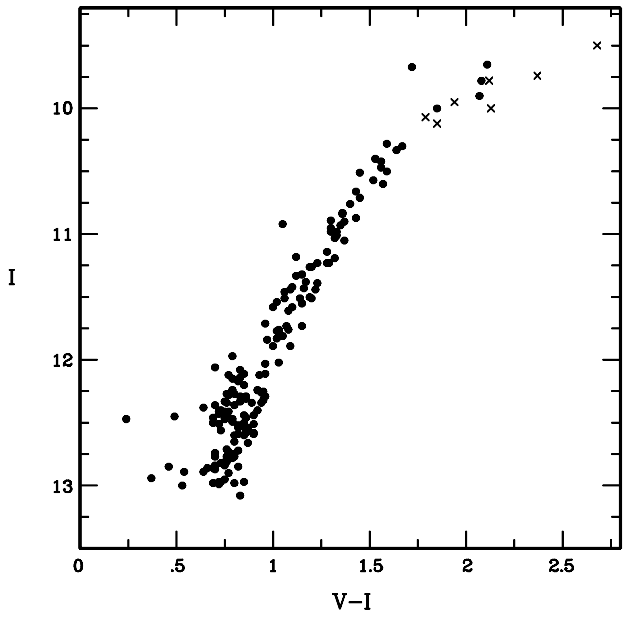


Figure 3. The color-magnitude diagram of the moderately metal-poor globular cluster 47 Tucanae. The short red horizontal branch lies below the limit of this figure, near *I*=13.4. Note the clump on the AGB for *I*=12.0–12.4, with gaps both above and below, with a resumption near *I*=11.8. The crosses denote semi-regular variable stars. This old and moderately metal-deficient cluster has a much steeper giant branch than the younger, metal-rich stars in the solar vicinity (figure 1).

HR diagram, the Hayashi track, defines the position of the giant branch.

A star leaves the main sequence and moves to the red on the subgiant branch when it has exhausted hydrogen in its core; hydrogen burning continues in a shell surrounding the currently inert helium core. This energy-generating shell gradually moves outwards, consuming hydrogen from the pristine outer layers and depositing helium and other reaction products on the core. The core in a star with a mass of not more than about two solar masses is degenerate during the ascent of the giant branch. The term red giant branch (RGB) is sometimes used specifically to refer to this stage in the evolution of a low mass star, although it is not easy to distinguish observationally between such stars and stars in the next (AGB) stage of evolution or more massive stars without degenerate cores. The H-burning shell moves outwards and at some stage it encounters the inner edge of the relatively H-rich convective envelope: the ascent of the giant branch slows for a time, causing a concentration of stars at this point on the giant branch. This is well seen in rich star clusters and is known as the RGB bump. The spectra of stars higher up the giant branch show a decline in the carbon isotope ratio; this is consistent with the mixing of an increasing quantity of the products of the CNO cycle of H-burning into the envelope. Observations of star clusters show that this process, which is not yet fully accounted for by theory, is most pronounced in metal poor stars of low mass.

The ascent of the RGB continues until ignition of the triple-alpha process of helium burning in the core. The reaction rate builds up very rapidly because of the strong temperature dependence of the nuclear reaction coupled with the inability of the degenerate material to expand and cool in response to the energy input. The degeneracy is lifted eventually and helium burning stabilizes. The rate of energy production in the star is considerably reduced during this stage, so the process described in the previous paragraph occurs in reverse and the star moves rapidly down the Hayashi track until equilibrium returns at the red giant clump or its equivalent. The theory requires mass loss to occur after the ignition of helium in the core. This is especially the case for the bluer HORIZONTAL BRANCH STARS which can only have a very small mass in the envelope. No direct evidence has been found for this mass loss so far. The process is so rapid that we would not expect to observe many stars in transit.

A star of more than about two solar masses does not have a degenerate core. The star does not increase greatly in luminosity on the giant branch before helium ignites in the core and the surface temperature increases; it spends most of its core helium burning life in a band to the blue of the giant branch. Observations of star clusters of a range of ages show the red giant clump in the oldest clusters becoming brighter and bluer as successively younger clusters are considered. These younger clusters lack the tall giant branch (RGB) of the older star clusters.

The asymptotic giant branch

The second ascent of the giant branch follows the exhaustion of helium in the core. The star now consumes hydrogen and helium in concentric shells. Instabilities occur in the energy-generating shells at high luminosity. Thermal pulses in the helium-burning shell may cause the temporary extinction of the hydrogen-burning shell, with substantial changes in the total energy production of the star and in its structure and luminosity. The observational consequences of this include changes in luminosity and in the character of the variability which is a general feature of such luminous red giant stars. This would most easily be seen by comparison with other stars in a cluster, but this evolutionary stage is quite short and well-populated clusters of the relevant age are few, so this effect is hard to establish.

These disturbances to the structure of the star lead to additional nuclear reactions as a result of material from the region outside the reaction zone being drawn into it. Deep convection in the envelope then brings reaction products to the surface of the star. This is known as the third dredge-up mechanism, as there have been earlier convective upwellings of processed material. The princi-

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

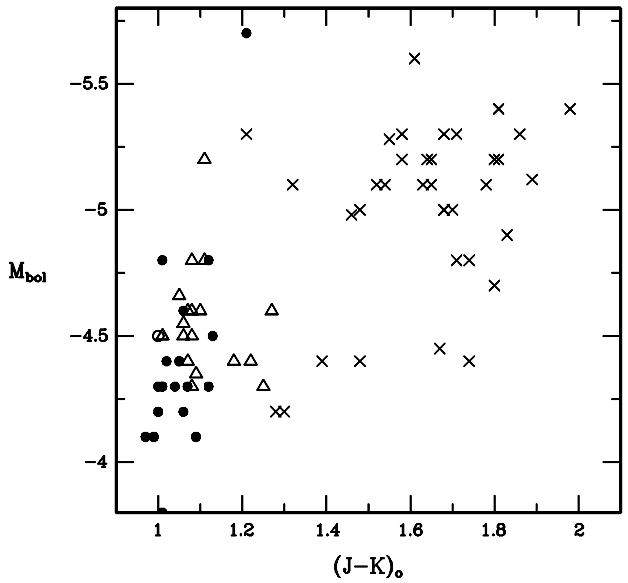


Figure 4. The composite near-infrared color-magnitude diagram of the upper asymptotic giant branch in star clusters of intermediate age in the Magellanic Clouds. Filled circles are early M stars, the open circle is a K star, the open triangles are MS or S stars and the crosses are carbon stars. The transition from M via S to carbon stars is clear.

pal reaction products brought to the surface are carbon, mainly the ^{12}C isotope produced by the triple-alpha process of helium burning, and heavy elements made by the slow neutron process (s-process).

The addition of large quantities of these materials to the STELLAR ATMOSPHERE causes spectacular changes to the spectral appearance, as these stars are cool enough to have spectra dominated by molecules and the principal molecules present depend on the chemical composition of the atmosphere.

The very low surface gravity and high luminosity of these stars are conducive to mass loss from the stellar atmosphere; this is observed and may occur at a high rate. Loss of the stellar envelope is the penultimate stage before the appearance of a WHITE DWARF star as the endpoint of the evolution of a star of small or intermediate mass; a more massive star will end its giant stage of evolution in a SUPERNOVA explosion. The ejected material from a low-mass star appears for a while as a PLANETARY NEBULA in which the material is ionized by the hot stellar remnant. The thermally pulsing AGB (TP–AGB) stage of evolution is thus one of the most dramatic in the whole evolution of a star.

The changing chemical composition of the atmosphere

Carbon produced by the triple-alpha reaction may change the normally oxygen-rich atmosphere to one in which carbon is more abundant. Most oxygen and carbon atoms are tied up in the CO molecule at the low photospheric temperature of these stars, so that only the more numerous species is available in quantity to form other molecules. Thus the spectra of the oxygen-rich M stars are dominated by oxygen-bearing molecules such as TiO and VO, while C_2, CH and CN are prominent in the spectra of the carbon-rich N stars. Stars with C/O very close to unity do not have strong bands of carbon- or oxygen-bearing molecules, although unusual molecules such as CaCl may be present; these stars are called SC or CS stars depending on whether they are marginally O or C rich. (Figure 5 shows examples of the spectra of cool giant stars of different surface composition.)

Nuclei produced in the STELLAR INTERIOR by the s-process are brought to the surface at an early stage and their presence is recorded by the appearance of ZrO, YO and LaO in oxygen-rich stars. These are classified as S stars if the additional bands are comparable with or stronger than TiO, or MS stars if these bands are weak. The lines of the s-process elements are also enhanced in the spectra of most CARBON STARS. Conclusive proof that this material is produced by an ongoing process in the star itself is provided by the presence of lines of Tc. This

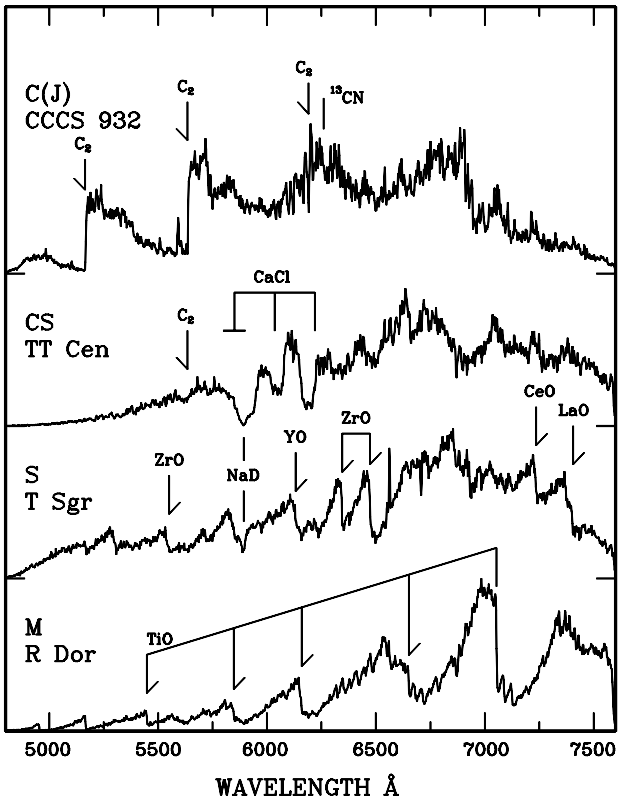


Figure 5. Optical spectra of M, S, CS and C type in the yellow-red spectral region. These show the very different sets of molecular bands which characterize these stars. These stars, except for the carbon star which is of the ^{13}C-rich J subtype, are rather cool examples or, as for TT Cen, were observed when the bands were strong. The original spectra were taken with an intensified Reticon detector at a resolution of 5 Å.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

is an s-process element with a radioactive half life of only 2×10^5 yr so that no significant quantity would remain from the time when the star was formed.

The evolutionary sequence M to S to N was proposed on theoretical grounds and ultimately proved by observations of stars of these types at successively higher luminosities on the upper giant branch of star clusters in the MAGELLANIC CLOUDS (figure 4). These clusters have ages of typically 10^9 yr and the giant stars have a mass of about 2 solar masses. Stars of type S and N are not found in the old globular clusters with evolving stars of a solar mass or less. Figure 5 illustrates the dramatic difference in spectroscopic appearance which results from the changing abundances of carbon and the s-process elements. The stars shown are galactic field stars which were selected as extreme examples with strong bands and do not represent an actual evolutionary sequence such as is seen in the Magellanic Cloud star clusters.

The spectra of M stars can be classified in terms of parameters which correlate well with surface gravity, which gives an indication of luminosity, and the photospheric temperature. This is not so for the cooler or N-type carbon stars, although the hotter R-type carbon stars are more tractable. The Na D lines seemed to offer a temperature criterion for the N stars, but the C/O ratio has a great influence on these and other resonance lines. The features which work well as luminosity criteria in ordinary giants and dwarfs, notably the Sr II lines and CN bands, are affected by the s-process and carbon enrichment and cannot be used. The CN strength also depends on the N abundance which, unlike the C abundance in the carbon stars, is essentially that with which the star was formed. Thus carbon stars in the Magellanic Clouds, which are N deficient, have weaker CN bands than most galactic carbon stars. The few galactic N stars with weak CN bands have the high space velocity typical of the old, metal-deficient stars of the galactic halo.

The transition from M to S to N occurs at a spectral type of about M4 in the Magellanic Cloud clusters. The stars near the transition are either constant in light or at most show small variations; these clusters contain a few much redder stars which are the mass-losing stars at the tip of the AGB. Thus the transition occurs well below the tip of the AGB in these moderately metal-deficient stars of two solar masses. The transition may occur at different points in different systems, for instance there are no N stars in the Central Bulge of the Galaxy although O-rich M stars are very common.

The Central Bulge of our Galaxy contains a very metal-rich old population. The quantity of carbon which must be dredged up to produce a carbon star depends on the relative quantities of oxygen and carbon which are present to start with, and it will be less if there is relatively little oxygen as in the case of a metal-deficient star. The trend of an increasing proportion of carbon stars in the sense galactic center, solar neighborhood, LMC, SMC may be dictated primarily by the metal content which decreases in this sense. It may be affected also by the distribution of stellar masses, which in turn depends on the history of star formation in each system.

There are also red giants of unusual surface composition which cannot have arisen in this way, mainly because the luminosity is far below that at which the third dredge-up occurs. The CH STARS AND BARIUM STARS, which are hotter and fainter than the stars described above, and some S stars which lack Tc lines so that their s-process material may be ancient, are found to be members of binary systems. The companions, when detected, are white dwarfs. This suggests that the companion was initially of higher mass and has already passed through the red giant stage. The processed material which gives the characteristic spectrum of the star now in the giant stage may therefore be attributed to reactions within the formerly more massive companion. This material would have been transferred when the original primary reached the mass-losing stage, or perhaps earlier if it overfilled its ROCHE LOBE as soon as processed material reached the surface layers. The existence of subdwarf and even dwarf CH and carbon stars is most easily understood in this way.

The R stars are the hotter carbon stars and are an inhomogeneous group. They include CH stars and the hydrogen-deficient carbon stars. The latter include the R Coronae Borealis stars, which undergo deep visual fading when clouds of soot are ejected. The carbon is almost pure ^{12}C and it is thought that these stars represent the final 'reborn' giant phase of the carbon core produced by the triple-alpha process of helium burning; as such they fall off the main line of evolution described above. Many of the remaining R stars form a group difficult to understand. There are unusually few binary stars among them while the spectra are characterized by a low $^{12}C/^{13}C$ ratio, suggesting that the carbon has been processed through the CNO cycle in a hydrogen-burning zone, and by the absence of an enhancement of the s-process elements. One possibility is that the helium core flash at the top of the RGB took an unusual course so that the products of hydrogen-burning were dredged up in quantity. The star would then be a carbon star at the red giant clump and throughout its ascent of the AGB. Theoretical calculations cast doubt on this: they indicate that only stars of exceptionally low metal content can become carbon stars in this way—such carbon stars have been discovered recently.

Some of the cooler N stars also have a low $^{12}C/^{13}C$ ratio: these are often called J stars. These amount to only about 15% of the local N star population, so it is surprising that many of the CH stars in both globular clusters and the field exhibit this property, which might suggest that the donor stars were not typical of carbon stars in

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

general. It has been proposed instead that the process of mass transfer caused deep convection in the stellar envelope of the receiving star, so that some of the material in the envelope has been processed by the CNO cycle in the H-burning shell.

Variability and mass loss

The coolest and most luminous stars are intrinsic VARIABLE STARS. The variability is usually at least quasi-periodic, with periods of a few hundred days in the case of the stars of larger amplitude. These are the long-period variables, divided into semiregular and MIRA VARIABLES. The latter are conventionally defined by having visual amplitudes greater than 2.5mag. They are also more regular and have longer periods. The visual amplitudes may be as great as 10 magnitudes in the case of some M and S stars. The TiO bands become very strong when the star is coolest and depress the luminosity in the *V* band, so that the amplitude is much greater than it would be in a star without absorption bands. The corresponding amplitudes at 2 μm are only 1–2 mag. The carbon stars do not have such strong molecular bands in the *V* band and generally have much smaller amplitudes in visual light.

A few stars have shown large abrupt changes in period and amplitude, indicating changes in the mode of pulsation, and the precise MACHO photometry of a large sample of stars in the Magellanic Clouds shows that several pulsation modes are possible and that more than one may be active at the same time in a given star. Several relationships between period and other properties are known. The longer-period stars have lower space velocities and so belong to a younger population, while among those in globular clusters the most metal-rich clusters have the stars of longer period. None is known in very metal-deficient clusters. The central bulge of the Galaxy, where the metal abundance is high, has relatively more Miras of very long period (around 500 days) than the solar neighborhood.

Infrared studies have revealed large quantities of dust around many of the most luminous giant stars. Most Mira variables and many semiregulars show emission from dust at 12 μm the Miras are systematically brighter at this wavelength relative to the photospheric emission at 2 μm. The few Miras in globular clusters have comparatively little circumstellar dust, which is made of the heavier elements which are deficient in globular cluster stars. The Miras of longer period all have dust shells. The most extreme cases, Miras of very long period, are invisible in the optical region because of the great extinction by the dust. The O-rich stars among them usually have strong OH maser emission in the microwave region and are known as OH/IR stars because they are observable only at these longer wavelengths. Many visible Miras and some semiregulars show weaker OH emission.

Maser emission from SiO and H_2O is also sometimes observed, but from different radial regions. SiO emission arises relatively near the photosphere and OH from very far out in the shell, while H_2O is located at an intermediate radius. The microwave emission in the form of sharp lines allows the measurement of radial velocities and it is found that the material is generally in the form of a shell with an expansion velocity of up to a few tens of kilometers per second. The combination of the expansion velocity with an estimate of the size of the shell and the density of the material allows the rate of mass loss to be estimated. The mass loss becomes observable at a rate of about 10^{-7} solar masses per year, and reaches 10^{-4} solar masses per year in the most extreme cases. This stage of evolution is necessarily shortlived, as is also apparent from the scarcity of such stars.

The dust emission has its own characteristic spectrum, dominated by silicate emission bands at 9.7 and 18 μm in the O-rich stars and by SiC emission at 11.4 μm in the carbon stars. The OH/IR stars show silicate absorption features but SiC is rarely seen in absorption in heavily obscured carbon stars. Carbon-based dust grains absorb more heavily in the optical region than does the silicate material around the oxygen-rich stars, so that a higher proportion of carbon stars are optically invisible. Many have only recently been found by infrared studies and it appears that the carbon Miras which are known from optical searches are only the less-obscured tail of a much larger population.

The ejection of material from the red giant stars is unlikely to occur at a constant rate. Long-continued observations of carbon-rich Miras show that the mean brightness varies on a time scale of decades (figure 6).

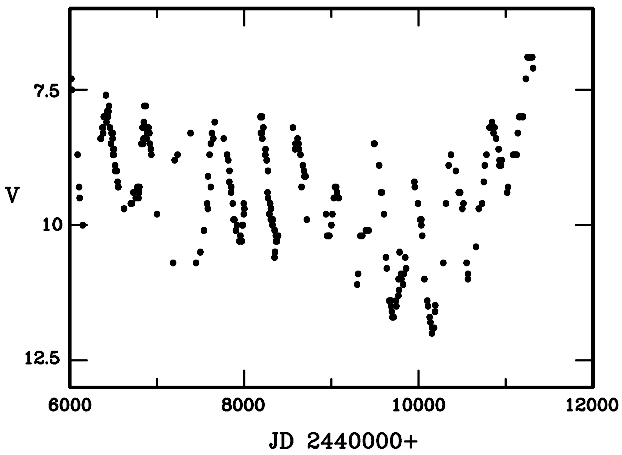


Figure 6. The light curve of the carbon Mira, R Lep, from visual observations by amateur astronomers in South Africa. The appearance of the resonance lines of Na, K and Rb and the C_2 bands in emission coincided with the marked fading of the overall light curve near JD 2 450 000 when two successive minima were exceptionally faint. These emission features disappeared rapidly when the star brightened up again. The last maximum shown is the brightest for many years.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

These stars are much redder during their faint episodes in a manner consistent with increased extinction by dust. The visible light spectrum is normally dominated by absorption bands of C_2 but during a faint episode these bands appear in emission instead; the resonance lines of Na, K and Rb appear in emission at the same time. This is suggestive of resonance fluorescence emission; the geometry remains to be established.

Hubble Space Telescope pictures of some stars with bright circumstellar material show many apparently concentric shells which suggests that mass loss has been episodic. The time scale is probably that of the helium shell instability which must cause major re-organization of the envelope and so disrupt the mass loss process.

Recent studies in which stars were selected by the IRAS and ISO infrared satellites for their far-infrared emission have revealed M stars with as much circumstellar material as the Mira variables but which are not conspicuously variable. These could be stars which alternate between Mira and small-amplitude pulsations, the dense shell being produced during the Mira episodes. The situation may alternatively imply that pulsation of large amplitude is not the only way to initiate large-scale mass loss. The flat energy distribution of some stars which are not variables of large amplitude but which have excess emission at both 3 μm and 60 μm relative to a typical Mira, may indicate that the circumstellar material is in a disk. The most extreme examples of stars with this type of energy distribution, which is not confined to very cool stars, sometimes show emission of the resonance lines of Na, K and Rb. This may be understood by a model, first put forward for the M star VY CMa, in which the star is hidden by an edge-on dust disk and is visible only by reflection from circumstellar material out of the plane of the disk. The resonance emission arises in this material.

The J-silicate carbon stars are an enigmatic group which combine the photospheric spectra of carbon stars with the circumstellar material normally associated with oxygen-rich stars. The J notation refers to the enhancement of ^{13}C in the spectra of most members of this group. These stars have a large value of the IRAS [12]–[25] index, in the range 0.8–1.8. Normal optically visible carbon stars nearly all have [12]–[25] <0.6 and redder objects are hidden by their circumstellar dust. OH and H_2O emission is seen in some cases, in agreement with the inference that the circumstellar material is O rich. The velocity structure suggests that the material is in a disk, and this is supported by the excess emission in the near infrared (around 3 μm) which is measured in some cases. A disk is a much longer-lived structure than an expanding shell and it is possible that O-rich material ejected in the previous evolution of the star, before it became a carbon star, has been retained in this way. The existence of a disk, formed from ejected material, around a single star is not easy to understand, but would be facilitated if the star is in a binary system. No direct evidence for the presence of companions has been forthcoming. Dusty disks of this type seem to be peculiar to J-type carbon stars, as similar near infrared excesses have not been found in ordinary N stars, although they are known in some M-type stars.

Giant stars in binary systems

Giant stars are not found in short-period binary systems because they would not fit. A star in a binary may evolve normally until it expands to fill its Roche lobe, following which further expansion will lead to rapid mass loss to the companion. The minimum separation of a binary system is thus such that neither star has filled its Roche lobe during its evolution to date. The present stellar radius may be much less than the maximum in the past. A classical Cepheid, a star with typically five times the mass of the Sun and in the core-helium-burning stage of evolution, would fit into a system with an orbital period of about 40 days, but the shortest periods known for binary systems containing Cepheids are about 500 days. This reflects the radius of the Cepheid when it was at the tip of the first ascent of the giant branch. Similarly, the minimum binary separation which will accommodate a star of the red giant clump is that which would accommodate the star when it was at the tip of the RGB.

Giant stars in binary systems may show evidence of interaction in the past, even though there has not been overflow of the Roche lobe. The orbits of the systems with the shortest orbital periods, with one component which is close to filling its Roche lobe, are invariably circular. A mean separation in excess of that of the smallest possible circular orbit is required to accommodate the giant star if that the minimum separation of an orbit is eccentric, as there must be space for the giant star when the stars are closest together. Tidal interaction works to reduce the eccentricity.

K giants which are members of binary systems with orbital periods less than about 100 days often show enhanced chromospheric activity, resulting in strong emission reversals of the Ca II H and K lines. Many show light variations caused by large starspots which move across the line of sight as the star rotates, so the light of the star is modulated at the rotational period of the star. There is normally close synchronism between the period of rotation of the giant star and the orbital period of the binary system. This is also a consequence of tidal interaction between the components.

The binaries where the separation is such that the evolving (giant) star will fill the Roche lobe set by the gravitational field of its companion undergo mass exchange from the evolving star to its companion. Mass loss from the system is also likely, in which case the size of the orbit will be affected. Gas in the system, whether in transit from one star to the other or in a ring or disk around one of the stars, may contribute atomic emission

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

lines to the spectrum of the system. The evolving star generally loses so much mass that it is stripped down and becomes the fainter of the pair, although it still has a large radius. This is the type of system which, seen edge-on, will undergo Algol-type eclipses. The now-fainter star is difficult to study without interference from the companion, but analyses of the light from the photospheres and coronae of several stars have shown that the chemical composition has been altered by the dredging-up of nuclear reaction products to the surface. There is increasing interest in the idea that some of the chemically-peculiar stars may owe their nature to such extreme interactions with companions which, unlike in the case of the Algol systems, are too faint to be obvious in the combined light of the system.

SYMBIOTIC STARS and some NOVAE in which the mass donating, non-degenerate component is a giant belong to the class of cataclysmic variables. The strong stellar wind of the giant may suffice to drive such activity even if the giant star does not overflow its Roche lobe. The type star of the Mira variables, Mira Ceti is a marginal example of this type. It has a companion star, at a distance of a few arcseconds, which has H emission lines and is variable. It shows both erratic rapid variability, caused by variable accretion, and long-term variability, which may result from changes in an accretion disk. These phenomena are best seen in the ultraviolet spectral region, to which the red giant contributes little.

More massive giant binaries, in which mass interchange has not yet had major evolutionary consequences, may be detected by their composite colours and spectra when the companion is a dwarf of type A or B and comparable in luminosity with the giant. Some of these systems have nearly edge-on orbits so that the small hot star acts as a probe of the atmosphere of the giant star when it passes behind it. Lines arising in the atmosphere or extended atmosphere of the giant are seen projected on the featureless spectrum of the hot star. Velocities and densities in the atmosphere of the giant star may be studied in this way. Dense clumps of gas have been found in the extended atmospheres of K supergiant stars in binaries by this means.

Stars in globular clusters

The globular clusters in our Galaxy show a wide range in chemical composition. The ratio Fe/H ranges from a value not far below solar to over 100 times less. The sharpness of the giant branch and the good agreement of the results of spectroscopic abundance analyses of the individual stars indicate that there is little spread in the iron abundance within a cluster, with the notable exception of Omega Centauri and possibly also M22. There is a relatively small range in age, of perhaps two billion years compared to an average age of over ten billion years, among the globular clusters in our Galaxy. The globular clusters were the first clusters to be formed in the Galaxy and nothing similar appears to have been formed in our Galaxy since, although some nearby galaxies contain rich young clusters.

The much smaller numbers of stars in the younger open clusters do not suffice to study the giant branch in such detail, although the populous young and intermediate-age clusters in the Magellanic Clouds as well as the DWARF SPHEROIDAL GALAXIES which appear to have a simple star formation history help to fill in the gaps. The globular clusters allow the location of the giant branch, sometimes known as the Hayashi track, to be studied as a function of chemical composition at constant age. The giant branch is displaced to the red in the color-magnitude diagrams for clusters of higher metal content and this has been calibrated spectroscopically, so that the metal content of more distant systems can be estimated from photometric data alone. Theory and observation show that mass differences have a much smaller effect on the colour of the RGB. The low mass and reduced metal content of the evolving stars in globular clusters preclude many of the phenomena we see in stars of the solar neighborhood. There are no long period Miras, no stars with thick dust shells and no carbon or S stars on the thermally pulsing asymptotic giant branch. The CH stars in several clusters are evidence that these phenomena occurred in more massive stars earlier in the life of the cluster.

The giant globular cluster Omega Centauri has long been known for having a broad giant branch, which is attributable to a range of metal content in the cluster. It has been shown recently to have stars of a range of age as well, and this has lead to the suggestion that it may be the remnant of a small galaxy cannibalised by the Milky Way. The chemical composition deviates from those found in other clusters as well as in the Galaxy at large, in that the stars of higher metal content show an exceptionally high ratio of the s-process elements to the iron group. The coolest stars show strong bands of ZrO and are classified as S stars. They are too faint to owe their overabundance of s-process elements to the operation of the third dredge-up themselves, and have presumably formed with this unusual composition as the result of incorporating material ejected by an earlier generation of more massive AGB stars.

The more metal-rich clusters such as 47 Tuc have short stubby horizontal branches, close to the giant branch, while the metal-deficient clusters have more extended or even predominantly blue horizontal branches. The discovery of clusters with red horizontal branches far out in the galactic halo suggested that, contrary to expectation, these clusters were relatively metal rich. It was soon found that the form of the horizontal branch is not solely a function of metal content and that a second parameter is involved. Age appeared as a likely contender for the second parameter but age determinations

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

from photometry of main sequence stars show that the age range among globular clusters cannot account for this. The clusters with red horizontal branches, other than those of relatively high metal content, are generally small or have a low central concentration. Stars on the blue horizontal branch are thought to have lost much of the mass of their envelopes, so it has been suggested that blue horizontal branch stars could have suffered stripping of the envelope by close stellar interaction in a densely crowded environment.

Massive stars

Much of our discussion so far has related to giant stars of low mass, typically less than about two solar masses on the RGB. Stars of higher mass differ in a number of ways. The red giant clump becomes brighter and somewhat bluer and the stage of core helium burning is now known as the blue loop. The loop extends to higher surface temperature for stars of higher mass or lower metal content. A star may pulsate as a Cepheid when it passes through the instability strip which lies well to the blue of the giant branch, and classical Cepheids, with masses between 3 and 9 solar masses, will be most common when the long-lived core-helium-burning stage falls within the instability strip. Subsequently the star will return to the giant branch. Observations of Magellanic Cloud stars suggest that stars of mass up to about 7 times that of the Sun pass through the AGB stage as described above, although the more massive stars do not become carbon stars.

Stars with moderate S characteristics (MS stars) are prominent among the most luminous identifiable AGB stars in the Magellanic Clouds. This is probably because increased CNO cycling converts carbon into nitrogen, so that there could in principle be nitrogen stars which, as C/O would be less than unity, would not be readily distinguishable from M stars. It is also possible that severe mass loss sets in to terminate evolution on the AGB before the C/O ratio exceeds unity, while the massive envelope of a star of high mass requires more processed material than that of a low-mass star to change its composition.

The mass range of these stars, 3–7 solar masses, implies that mid to late B-type main sequence stars and Cepheid variables may evolve into carbon stars or Mira variables but to date there is little direct evidence for this from the LARGE MAGELLANIC CLOUD clusters. The clusters containing carbon stars all have evolving stars within a relatively small mass range, 1.5–2.5 solar masses, but this appears to be limited on both sides by the age limits of a burst of star formation in the Large Magellanic Cloud. The failure to find carbon stars in clusters with B type main sequence stars and classical Cepheids may simply reflect a lack of large clusters of the correct age group.

Stars of still higher mass do not pass through the same evolutionary stages. Stars of mass around 10–15 times the mass of the Sun become M supergiants with luminosity class Ia, Ib or II rather than the luminosity class III giants of the lower mass stars. The Hertzsprung gap between the main sequence and the giant branch stars widens with increasing luminosity and there are few red supergiants of type earlier than M0. They are rarely of spectral type later than M3, that is they do not become as cool as the giants, and variables of large amplitude are almost unknown. One of the few exceptions is S Per, of spectral type M4–5 and with an amplitude which can reach 5 mag at times. These stars are found in several young star clusters in the Galaxy, with a concentration in the association around the Perseus Double Cluster.

The M supergiants have extensive circumstellar dust. The theoretical evolutionary tracks are nearly horizontal in the HR diagram plotted with luminosity or bolometric magnitude as ordinate. There is a loop to the blue after a red supergiant stage, then perhaps a second spell as a red supergiant before the star evolves to the blue again. The great width of the Hertzsprung gap implies very rapid evolution across the HR diagram, so that the evolved stars spend almost all their time as blue (spectral type B) or red (M) supergiants and move quickly through the intervening yellow supergiant stage. Supergiants of the intervening types, A, F, G and K, are known but are rare.

Bibliography

There is an enormous literature on red giants but several excellent reviews cover major parts of the topic. The *Annual Review of Astronomy and Astrophysics* is an excellent source and several recent reviews may be cited

Chiosi C, Bertelli G and Bressan A 1992 New developments in understanding the HR diagram *Annu. Rev. Astron. Astrophys.* **30** 235–85

Gustafsson B 1989 Chemical analyses of cool stars *Annu. Rev. Astron. Astrophys.* **27** 701–56

Renzini A and Fusi Pecci F 1988 Tests of evolutionary sequences using color-magnitude diagrams of globular clusters *Annu. Rev. Astron. Astrophys.* **26** 199–244

A detailed account of our present understanding of the ascent of the red giant branch by low mass stars is given by

Salaris M, Cassisi S and Weiss A 2002 Red Giant Branch Stars: The Theoretical Framework *Publs Astr. Soc. Pacific* **114** 375–402

A recent review of our knowledge of carbon stars is given by

Wallerstein G and Knapp G R 1998 Carbon stars *Annu. Rev. Astron. Astrophys.* **36** 369–433

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

A wide-ranging review of the field as of a decade ago is given in

Johnson H R and Zuckerman B (ed) 1989 *Evolution of Peculiar Red Giant Stars* (Cambridge: Cambridge University Press)

See also

Wing R F (ed) 2000 *The Carbon Star Phenomenon* (Dordrecht: Kluwer)

A comprehensive review of the physical processes leading to the manufacture of heavy elements in stars, containing much information on the observational properties of the stars involved, is given by

Wallerstein G *et al* 1997 *Rev. Mod. Phys.* **69** 995–1084

And also by

Sugimoto D, Fujimoto M 2000 Why Stars Become Red Giants *Astrophys. J.* **538** 837–53

The explanation for the expansion to the red giant stage is given in

Renzini A, Greggio L, Ritossa C and Ferrario L 1992 Why stars inflate to and deflate from red giant dimensions *Astrophys. J.* **200** 280

The complex field of abundance anomalies in globular cluster stars is reviewed by

Kraft R P 1994 Abundance Differences Among Globular-Cluster Giants: Primordial Versus Evolutionary Scenarios *Publs Astr. Soc. Pacific* **106** 553–565

And by

Da Costa G 1998 Abundance Anomalies in Globular Clusters *Fundamental Stellar Properties: The Interaction between Observation and Theory (IAU Symposium 189)* ed Bedding T R, Booth A J and Bavis J (Dordrecht: Kluwer) 193–202

Tom Lloyd Evans

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Redman, Richard Oliver (1905–75)

English astronomer, became director of the Cambridge Observatories. A solar astronomer and an instrumentalist, he analysed the solar chromosphere, as well as playing an active role in specifying and supervising the building of the Anglo-Australian Telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Redshift

The redshift (or blueshift) of an object is the displacement of its spectral features to longer (or shorter) wavelengths due to a combination of the gravitational redshift, Doppler motions and the general expansion of the Universe. More properly, the term RADIAL VELOCITY is used primarily for the Doppler motions, which are usually the result of gravitational interactions, while redshift is reserved for the cosmological and direct gravitational effects. Redshift, z, is defined in wavelength as

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e} = \frac{\lambda_0}{\lambda_e} - 1$$

where the subscripts 0 and e refer to observed and emitted. In the frequency domain, redshift is defined as

$$z = \frac{\nu_e - \nu_0}{\nu_0} = \frac{\nu_e}{\nu_0} - 1$$

The dichotomy between these formulae is due to the fact that redshift causes wavelengths to increase and frequencies to decrease.

The redshift any given object can be decomposed as

$$z_{\rm total} = z_{\rm grav} + z_{\rm physical} + z_{\rm H}$$

where $z_{\rm H}$ is the cosmological redshift due to Hubble expansion and $z_{\rm physical}$ is the redshift due to the physical motion of the object with respect to the observer. In an expanding universe, the cosmological redshift is always positive (objects are always redshifted). The gravitational redshift as observed from the Earth is also generally always positive, but the physical motions of stars or galaxies can produce both redshifts and blueshifts, depending on whether the induced motion is away from or towards the observer respectively.

The gravitational redshift is due to the loss of energy light suffers when leaving a deep gravitational potential. Indeed, if the potential is strong enough, light cannot escape, the gravitational redshift is infinite and and a BLACK HOLE results. The gravitational redshift of a photon emitted at radius r_1 and observed at radius r_2 from a central mass is given by

$$\frac{\Delta\nu}{\nu} = \frac{-GM}{c^2}\left(\frac{1}{r_1} - \frac{1}{r_2}\right).$$

Gravitational redshifts have been measured in the Sun (see e.g. SOLAR PHOTOSPHERE: THE LIMB EFFECT AND GRAVITATIONAL REDSHIFT) and in only a small number of dense stellar objects such as WHITE DWARF stars. For example, the gravitational redshift of light leaving the surface of the white dwarf star Sirius B and viewed at infinity (or the Earth!) is about 0.000 06 or, in velocity units, $cz \sim 19$ km s^{-1}. The gravitational redshift of light leaving the surface of the Sun, which is much less dense than Sirus B and so has a much smaller gravitational potential well, is about 0.000 002, or only 0.6 km s^{-1}.

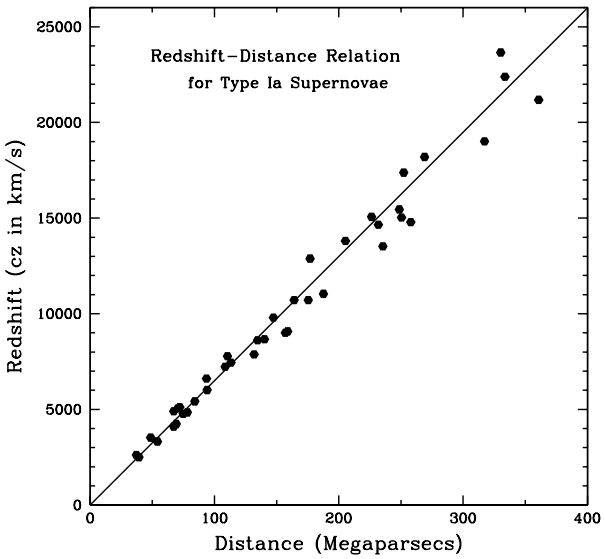


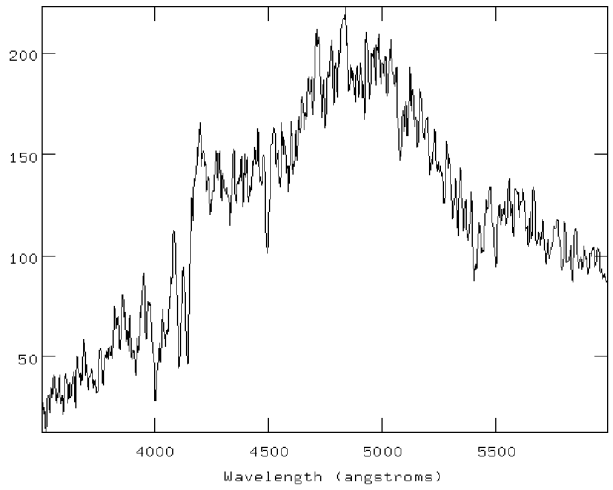
Figure 1. The redshift–distance relation for type Ia supernovae from the Calan-Tololo supernova survey (from data courtesy of S Jha, Harvard-Smithsonian). Redshift is given as cz and the distance is in megaparsecs (1 Mpc is approximately 3.26 million light years). The line through the data represents a Hubble constant of 65 km s^{-1} Mpc^{-1}, close to the value astronomers measure today.

Physical motions are most often gravitational in nature, such as the orbital motion of the Earth around the Sun. The Earth is in a nearly circular orbit around the Sun at a velocity of $\sim$ 29.7 km s^{-1}. An observer at the outskirts of the solar system and in the plane of Earth's orbit would see the velocity of the Earth relative to the Sun vary sinusoidally between +29.7 and -29.7 km s^{-1}. On cosmological scales, physical velocities are also generally caused by the orbital motions of galaxies around each other or around a common center of mass such as the deep potential well of a CLUSTER OF GALAXIES. The largest extragalactic physical velocities seen in the nearby universe are found for galaxies orbiting in clusters of galaxies ($\sim$1500 km s^{-1} or $z = 0.005$), kinematics in the central regions of QUASARS and ACTIVE GALACTIC NUCLEI ($\sim$10 000 km s^{-1} or $z = 0.035$, and our Galaxy's motion with respect to the COSMIC MICROWAVE BACKGROUND reference frame (630 km s^{-1}, $\sim z = 0.002$)

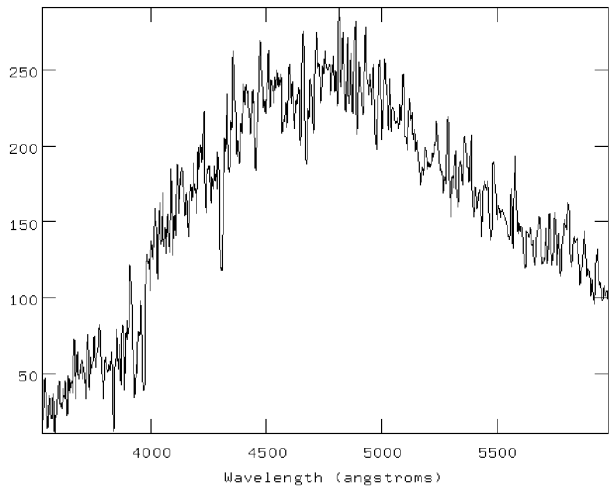
Outside the nearby Universe, redshifts or apparent radial velocities are dominated by the cosmological expansion. This expansion is more properly described as the stretching of the metric or yardstick. In the standard mathematical description of cosmology, the Friedmann–Lemaitre model, distances are defined in terms of the Robertson–Walker metric, the most general mathematical description for a uniform, homogeneous space that is expanding or contracting:

$$ds^2 = c^2\,dt^2 + R^2(t)\left(\frac{dr^2}{(1-kr^2)} + r^2\sin^2\theta\,d\phi^2\right)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*)



(*b*)

Figure 2. (*a*) The spectrum of a normal galaxy with strong absorption features due to common chemical elements, Ca, Mg, Fe, etc. The redshift of the galaxy is $z = 0.0455$ ($cz = 13\,624$ km s^{-1}). (*b*) The spectrum of a galactic star similar to the stars dominating the light of the galaxy in (*a*), but at a redshift near zero. The strongest absorption lines seen in both the star and the galaxy are the CaII doublet at 3933 + 3968 Å and the G-band (a mix of Fe, Mn, Cr and other metals) at 4300 Å. In the star, these features are seen nearly at their rest wavelengths. In the galaxy, the Ca lines have been shifted to 4110 and 4150 Å, and the G-band is nearly at 4500 Å. The 'redshift' is that shift to longer, redder wavelengths.

for spherical coordinates r, θ, ϕ and time coordinate ct. $R(t)$ is the scale factor, a function of time that is specified by the laws of gravity, and k is the curvature constant, either -1, 1 or 0 for an open, closed or flat space geometry.

In an expanding universe, the cosmological redshift of an object is directly related to its distance. Light travels on null geodesics, $ds = 0$. Therefore a burst of light emitted from a source at some coordinate distance r_1 and time t_e will be received by an observer at coordinate $r = 0$ and time t_0, and will be given by

$$\int_0^{r_1} dr/(1-kr^2)^{1/2} = \int_{t_e}^{t_0} dt/R(t).$$

The left-hand side of this equation is constant in time, while the right-hand side evolves as the photons propagate from the source to the observer. If another burst of light is emitted after an interval dt_e, then the integral on the left-hand side is unchanged, so that $dt_e/R_e = dt_0/R_0$. We can consider dt to be the time between peaks of a continuous wave, and since the frequency of the wave is the inverse of the interval between peaks, we have

$$\nu_e/\nu_0 = \lambda_0/\lambda_e = 1 + z = R(t_0)/R(t_e).$$

Thus photon wavelengths simply stretch with the universe. For small separations, the redshift z is proportional to the distance D to the galaxy, $cz = H_0 D$, where

$$H_0 = \frac{\dot{R}(t)}{R(t)}$$

is the HUBBLE CONSTANT, named for the discoverer of the expansion of the Universe. The form of this equation is the reason that redshift is often quoted as apparent radial velocity, cz. Note that there is no special relativistic $1/(1 + v^2/c^2)$ correction applied—apparent radial velocity is just cz.

There are several definitions of distance in cosmology. The proper distance to an object is the integral of the scale factor over the metric:

$$D_P = R(t) \int_0^{r_1} \frac{dr_1}{\sqrt{1-kr_1^2}} = \frac{R(t)}{\dot{R}(t)} cz$$

which again reduces to the standard redshift–distance relation

$$D_P = \frac{cz}{H_0}.$$

Other definitions of distance are defined by physical properties of objects, and include luminosity distance and angular diameter distance. These different distances are also described in any specific COSMOLOGICAL MODEL in terms of the redshift and the scale factor $R(t)$.

Another key parameter for observers and theorists alike is the lookback time to a given redshift. This depends not only on the Hubble constant but also the specific cosmological model. For variants of the Friedmann–Lemaitre models we have been considering, lookback time is given by the integral equation

$$t_0 - t_1 = H_0^{-1} \int_0^{z_1} (1+z)^{-1}[(1+z)^2(1-\Omega_M z) - z(2+z)\Omega_\Lambda]^{-1/2} dz$$

derived by Caroll, Press and Turner (1992 *Ann. Rev. Astron. Astrophys.* **30** 499). Here Ω_M is the ratio of the mean matter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

density of the universe to the closure density and Ω_Λ is the effective density in a cosmological constant

$$\Omega_\Lambda = \frac{\Lambda}{3H_0^2}.$$

The first significant extragalactic redshift was that of the ANDROMEDA GALAXY (M31) measured by Vesto Slipher in 1912. EDWIN HUBBLE discovered the empirical linear correlation between redshift and distance for nearby galaxies in the late 1920s, although we now know that Hubble's original redshift–distance relation, 600 kilometers per second per megaparsec, was nearly a factor of ten too steep—today's value lies near 65 km s^{-1} Mpc^{-1}. A modern version of the redshift–distance relation based on type Ia SUPERNOVAE is shown in figure 1.

Redshifts are measured spectroscopically by measuring the shift of spectral features of known wavelength and frequency. This is done by either measuring individual lines such as the 21 cm emission line of neutral hydrogen in the radio or the strong absorption features found in the atmospheres of cool giant stars that dominate the optical spectrum of most galaxies, or by cross-correlating the galaxy spectrum with a template spectrum of a similar star at known redshift. The spectrum of a galaxy which is at a redshift of 13 624 km s^{-1} is shown in figure 2 along with that of a star in our own galaxy at a redshift near zero (see also GALAXIES AT HIGH REDSHIFT).

In the last few decades, astronomers have begun to map the nearby Universe by measuring redshifts for samples of thousands of galaxies (see GALAXY REDSHIFT SURVEYS). Maps such as that shown in figure 3 have revealed the complex tapestry of the large scale distribution of galaxies.

John Huchra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

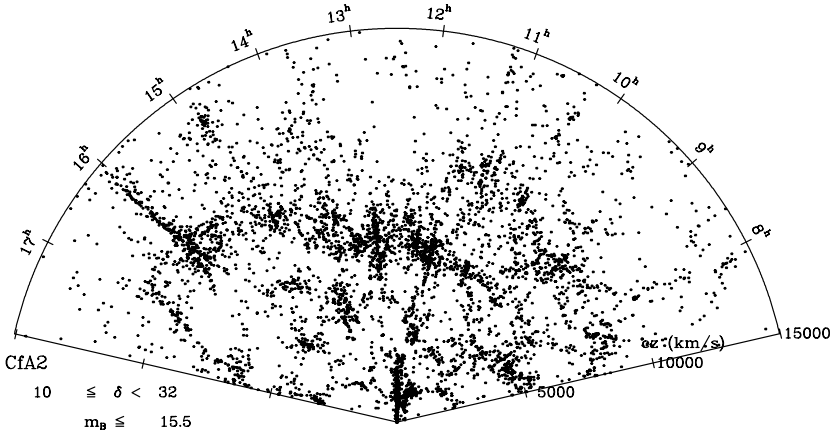


Figure 3. A redshift survey of a broad slice of the nearby Universe. The view is down on a wedge similar to a slice of an orange that is 22° thick and extends out to $z = 0.05$ ($cz = 15\,000$ km s^{-1}). Each point represents a galaxy and there are approximately 5600 galaxies in the plot. We are at the origin or apex of the wedge. The large structure crossing the wedge is the Great Wall of galaxies. The structures that point at the origin are clusters of galaxies; the individual galaxies in each cluster are actually in a relatively small volume of space but appear stretched out towards and away from the observer (the origin) when plotted in redshift space due to the physical motions of the galaxies orbiting each other in the gravtational potential well of the cluster. In real space, they would appear essentially all at the same distance, about the mid-point of the 'finger of God'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rees, Sir Martin John (1942–)

British cosmologist, became Plumian Professor of Astronomy at Cambridge and in 1995 Astronomer Royal (an honorary office by then separated from the directorship of the Royal Observatory at Greenwich). His forte is to be able to model any astrophysical phenomenon with physical insight, and he has successfully applied this technique to quasars, x-ray sources, gamma ray bursters, galaxy formation, galaxy clustering and the cosmic background radiation. Successful lobbier for and popularizer of science.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reeves, Hubert (1932–)

French astrophysicist, born in Montreal, Quebec, became Directeur De Recherche Scientifique, Centre Nationale de la Recherche Scientifique, Paris, and specialized in nuclear astrophysics, studying the reactions in the centres of stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reference Frames and Time Scales in General Relativity

In relativity one considers events to be described by points in a four-dimensional SPACE-TIME manifold. Let x^μ ($\mu = 0, 1, 2, 3$), be a set of coordinates in a certain part of the space-time manifold. $x^0 = ct$ is a time coordinate t multiplied by the speed of light, c, to gain the dimension of length. $x^i, i = 1, 2, 3$ are three spatial coordinates. If such a coordinate system is actually used in practice (e.g. for some EPHEMERIDES of solar system bodies) it is also called a reference system. A reference frame presents a materialization of a reference system. This materialization is obtained by association of some observed events with corresponding space-time coordinates. In astronomy, observations of satellites, solar system bodies, stars or extragalactic radio sources can be employed for that purpose (see SPACE AND TIME REFERENCES, CONCEPTS).

To describe the observables theoretically in a coordinate-independent manner a so-called local reference frame is often introduced. Such a local reference frame is defined by a set of four orthogonal unit vectors $e^\mu_{(\alpha)}$ ($\alpha = 0, 1, 2, 3$) moving with the observer. $e^\mu_{(0)}$ is related with the passage of time as experienced by the observer and $e^\mu_{(i)}$ $i = 1, 2, 3$ present three spatial reference directions.

Metric tensor and coordinates

The fundamental object of relativity and Einstein's theory of gravity is the so-called metric tensor $g_{\mu\nu}$ of space-time. This metric tensor describes the infinitesimal distance ds between two neighboring points x^μ and $x^\mu + \mathrm{d}x^\mu$ by

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu$$

where a summation over $\mu, \nu = 0, 1, 2, 3$ is implied. This means:

$$\mathrm{d}s^2 = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu.$$

In the absence of gravitational fields one can choose Cartesian inertial coordinates such that the metric tensor takes the simple form

$$g_{\mu\nu} \equiv f_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$$

which means:

$$g_{00} = -1$$
$$g_{ii} = 1$$
$$g_{i0} = 0.$$

This form is valid in any of such inertial coordinate systems; from this one finds that if $(ct, \mathbf{x})$ and $(ct', \mathbf{x})$ are two such coordinate systems, where the two origins move with constant speed relative to each other, the two systems are related by a *Lorentz transformation*. If the velocity is in the direction of common x-axes this transformation reads:

$$t' = \frac{t - vx/c^2}{\sqrt{1-(v/c)^2}}; \quad x' = \frac{x - vt}{\sqrt{1-(v/c)^2}}.$$

One basic property of the metric tensor is its relation with the proper time τ indicated by some (idealized) atomic clock. Let $z^i(t)$ describe the motion of the clock in the chosen coordinate system then ($\mathrm{d}z^0 = c\,\mathrm{d}t$)

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}\tau^2 = g_{\mu\nu}\,\mathrm{d}z^\mu\,\mathrm{d}z^\nu.$$

If gravitational fields are considered, the metric tensor has to be determined by Einstein's field equations of Einstein's theory of gravity (general relativity theory, GRT). The so-called post-Newtonian framework presents a slow-motion weak-field approximation to GRT. To lowest order, the gravitational potential U enters the time–time component of the metric tensor:

$$g_{00} = -1 + \frac{2U}{c^2} + \mathrm{O}(c^{-4}).$$

For the motion of solar system bodies the post-Newtonian metric must be developed to higher powers of c^{-1} and can be written in the form

$$g_{00} = -1 + \frac{2w}{c^2} - \frac{2w^2}{c^4} + \mathrm{O}(c^{-5})$$

$$g_{0i} = -\frac{4}{c^3}w_i$$

$$g_{ij} = \delta_{ij}\left(1 + \frac{2w}{c^2}\right) + \mathrm{O}(c^{-4}).$$

Here, δ_{ij} is the Kronecker symbol, the scalar potential w generalizes the Newtonian potential U and the vector potential w_i describes magnetic-type gravitational effects related for instance with the Lense–Thirring acceleration due to the gravito-magnetic effect in a rotating body.

Barycentric and geocentric coordinates

Let (ct, x^i) be barycentric coordinates with origin at the center of mass of the whole solar system. If we forget about the rest of our universe and consider our solar system to be isolated we might choose these coordinates to extend to infinity such that far from the origin the metric tensor approaches $f_{\mu\nu}$. Such coordinates are called global, asymptotically Minkowskian (barycentric) coordinates. They play an especially important role in ASTROMETRY and the definition of an international celestial reference system (ICRS).

Let (cT, X^a) be coordinates moving with the Earth. If the origin is chosen by the vanishing of the (relativistic) Earth's mass dipole moment these coordinates are called geocentric. In both barycentric and geocentric coordinates the post-Newtonian metric tensor can be written as indicated above; with (w, w_i) in the barycentric system and (W, W^a) in the geocentric one. One usually requires the external part of geocentric metric potentials resulting from the gravitational fields of solar system bodies apart from the Earth to vanish at the geocenter. This implies that the geocentric metric equals $f_{\mu\nu}$ given above (i.e. the Minkowski metric) at the geocenter if we forget about the potentials from the Earth itself.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Various time scales: TCB, TCG, TT, TAI and TDB

The barycentric coordinate time t is called TCB; the geocentric coordinate time T is called TCG. Let us consider the reading τ of some (idealized) atomic clock in the vicinity of the Earth. From the geocentric metric tensor one infers that τ is related with $T = \mathrm{TCG}$ by

$$\frac{\mathrm{d}\tau}{\mathrm{d}T} = 1 - \frac{U}{c^2} - \frac{1}{2}\frac{\mathbf{V}^2}{c^2} + \mathrm{O}(c^{-4}).$$

Here, $\mathbf{V}$ is the geocentric coordinate velocity of the clock and U is the total geocentric gravitational potential at the location of the clock (the Earth's gravitational potential plus tidal terms resulting from other solar system bodies).

The SI second is related to proper time indicated by a clock. Since the relation of proper time to coordinate time involves the coordinate velocity of the clock and the gravitational potential at the location of the clock, for the definition of an operationally defined geocentric time scale (TT = terrestrial time), one has to specify the relation of TT with the proper time of certain, carefully selected clocks.

At present the definition of TT involves the geoid: TT is a time scale that differs from $T = \mathrm{TCG}$ uniquely by a constant rate, its unit of measurement being chosen such that it agrees with the SI second on the rotating geoid. This implies

$$\mathrm{TCG} - \mathrm{TT} = L_G \times (\mathrm{JD} - 2\,443\,144.5) \times 86\,400\ \mathrm{s}$$

where JD is the Julian continuous count of days. The bracket vanishes on 0h, 1 January 1977. One has:

$$L_G \simeq 6.969\,29 \times 10^{-10}.$$

Here, the constant L_G is determined by the gravitational potential at the geoid, U_0, by $L_G = U_0/c^2$. Because of the problems with the geoid definition, the definition of TT will likely become independent of the geoid in the future. INTERNATIONAL ATOMIC TIME (TAI) is derived as a weighted mean over the readings of several atomic clocks, reduced to the geoid. At present, TAI is related to TT by

$$\mathrm{TT} = \mathrm{TAI} + 32.184\ \mathrm{s}$$

where the additive constant has purely historical origin.

The relation between barycentric and geocentric coordinates takes the form of a generalized Lorentz transformation. Let $\mathbf{z}_E, \mathbf{v}_E$ be the barycentric coordinates and coordinate velocity of the geocenter, U^{ext} the gravitational potential produced by all solar system bodies apart from the Earth then the TCB − TCG relation is given by

$$\mathrm{TCB} - \mathrm{TCG} = c^{-2}\left[\int_{t_0}^{t}\left(U^{\mathrm{ext}} + \frac{1}{2}\mathbf{v}_E^2\right)\mathrm{d}t + \mathbf{v}_E\cdot(\mathbf{x} - \mathbf{z}_E)\right] + \mathrm{O}(c^{-4}).$$

The numerical expression for TCB − TCG can be evaluated from the positions and velocities of the solar system bodies from this expression by numerical integration. Approximate formulae have been derived by Hirayama *et al* (1987) and by Fairhead and Bretagnon (1990). Note that the rates of TCB and TCG differ. This secular term is given by

$$(\mathrm{TCB} - \mathrm{TCG})_{\mathrm{secular}} = L_C \times (\mathrm{JD} - 2\,443\,144.5) \times 86\,400\ \mathrm{s}$$

with

$$L_C \simeq 1.480\,83 \times 10^{-8}.$$

Finally, another barycentric time scale is called TDB (sometimes also denoted by TB). This time has no official status, but is still used especially for solar system ephemerides and spacecraft navigation. According to IAU recommendations from 1979, TT and TDB should differ only by periodic terms leading to the serious problem that such a TDB time scale cannot be realized rigorously because of arbitrarily long periods in the motion of the solar system. Approximately we have

$$\mathrm{TCB} - \mathrm{TDB} = L_B \times (\mathrm{JD} - 2\,443\,144.5) \times 84\,400\ \mathrm{s}$$

with $L_B = L_C + L_G \simeq 1.550\,52 \times 10^{-8}$.

Dynamically and kinematically non-rotating geocentric coordinates

The relation between barycentric and geocentric spatial coordinates, x^i and X^{a} respectively, can be written in the form

$$x^i = z_E^i(T) + e_{(a)}^i(T)X^a + \cdots,$$

where quadratic and higher-order terms in X^a have been dropped. This generalizes the result from a Lorentz-transformation. Two possibilities exist:

- rotationally inertial (dynamically non-rotating) or
- fixed-star oriented (kinematically non-rotating).

A *dynamically non-rotating* reference system is a locally inertial system with relation to rotations characterized by the vanishing of Coriolis and centrifugal forces in some dynamical equation of motion for a test body. It is a remarkable fact that, due to relativistic effects, local inertial axes *precess* with respect to the fixed stars. For a geocentric reference system this precession is mainly given by the so-called *geodesic precession*, resulting from the motion of the Earth–Moon system about the Sun. The angular velocity of geodesic precession amounts to 2″ per century.

In contrast to these dynamically non-rotating systems *kinematically non-rotating systems* cannot be defined locally; instead their definition involves the globally defined coordinate system, comprising the asymptotic regime where the celestial sphere is defined. Then a (local) system is called kinematically non-rotating if the spatial coordinate lines differ from the corresponding global ones only by scale factors but not by their orientation in space, i.e. they are simply defined by $e_{(a)}^i = \delta_{ia} \times$ (scale factors).

To have a simple relation of the geocentric spatial coordinates with reference catalogs (implying a simple relation between the satellite frame and the VLBI frame) kinematically non-rotating systems are clearly preferred.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Brumberg V A 1991 *Essential Relativistic Celestial Mechanics* (Bristol: Adam Hilger)
Damour T, Soffel M and Xu C 1991 *Phys. Rev.* D **43** 3273
Fairhead L and Bretagnon P 1990 *Astron. Astrophys.* **229** 240
Hirayama T, Kinoshita H, Fujimoto M-K and Fukushima T 1987 *Proc. IAU Symp. at IUGG XIX General Assembly (Vancouver)* vol I, p 91
Soffel M 1989 *Relativity in Astrometry, Celestial Mechanics and Geodesy* (Berlin: Springer)

Michael H Soffel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reflecting Telescopes

All the earliest optical telescopes were REFRACTORS, using lenses to form an image; they date back to at least 1608. Since light of different wavelengths is bent or 'refracted' unequally as it passes through glass, a refractor always gives a certain amount of false color round an image such as that of a star. Light of whatever color is however reflected in the same direction, so a REFLECTOR does not suffer from this chromatic aberration. The first practical optical reflecting telescope was made by Isaac Newton around 1668.

In a Newtonian reflector, there is no lens to collect the light. The light passes down a tube and falls upon a concave mirror. The mirror sends the light back up the tube onto a smaller, flat mirror placed at an angle of 45°. The light is then directed into the side of the tube and is brought to focus. The image is formed outside the incoming light beam and is enlarged by an eyepiece lens—the observer's head does not obstruct the incoming beam (see NEWTONIAN TELESCOPE).

The Newtonian pattern is still widely used by amateur reflectors of the 6 inch to 20 inch (15 to 50 cm) class. The lower limit to the useful size range is set by the proportion of light to the main mirror which is obstructed by the flat mirror and its support. The upper limit is weaker and is set by the convenience of access to the eyepiece, say at the height of a standing man or woman, although larger telescopes may have a Newtonian focus accessed by ladders. The 72 inch (1.9 meter) Radcliffe Telescope, originally erected in 1948 at Pretoria in South Africa and later moved to Sutherland in 1974, was the largest Newtonian telescope constructed (equal to the size of the Rosse reflector of 1845). Its Newtonian focus is accessed from a cage attached to the dome shutter mechanism—this facility (intended for photography) is almost unused nowadays.

For larger apertures than, say, 50 cm, the CASSEGRAIN TELESCOPE, invented in 1673 is more convenient. Here the secondary mirror is convex and is placed between the main mirror and its prime focus, reflecting the light to the main focus of the telescope, through a hole in the main mirror. The combined action of the two mirrors lengthens the focal length of the telescope above the focal length of the primary mirror alone, but makes the telescope more compact. The combination of concave and convex mirrors also reduces the off-axis aberrations introduced by the primary mirror. The focus of a Cassegrain telescope is close to the observing floor, and as convenient to access as the telescope mounting will allow. The reduction in overall length of a Cassegrain telescope (compared to a Newtonian telescope) directly reduces the size of the building in which it is housed, with consequent saving of costs. Thus, Cassegrain telescopes quickly became popular for astronomical use, and still are, largely because they are much shorter and more compact than Newtonians and are therefore much easier to handle and less expensive.

The next major developments in making reflecting telescopes were due to William Herschel. He made his own mirrors and telescopes, of altazimuth design. With one of these, a reflector of 6.2 in (16 cm) aperture and 7 foot (2.1 m) focal length, he discovered the planet Uranus in 1781. His largest telescope had a mirror 49 in (1.5 m) across. It was not surpassed in size until 1845, when the third Earl of Rosse, in Ireland, produced a Newtonian telescope with a 72 in (1.9 m) mirror. Despite the limitations of its extraordinary mounting (it swung on ropes and chains between two massive stone walls on which the observing platform was mounted, to access the eyepiece), and its cylindrical tubular mounting (which trapped heat and, potentially, degraded seeing), the Rosse telescope discovered the spiral forms of the objects we now know to be external galaxies. The telescope was restored to working order (although not with its original mirror) in 1998.

The early reflectors had 'speculum' mirrors of low reflectivity and did not produce bright images. The metal mirrors tarnished quickly and had to be regularly re-polished, whereas refractors required practically no maintenance at all. It became possible to make glass mirrors, which could be coated with a highly reflective substance such as silver or aluminum. In modern times, the reflection coating can be optimized to the wavelength bands used by the telescope, e.g. the infrared.

As telescopes, whether refractors or reflectors, track a star, or point to different stars in the sky, their attitude changes relative to the vertical and thus to the direction of gravity. This produces mechanical flexure not only of the telescope structure but also of the glass lenses or mirrors, causing degradation of the images. The mechanical rigidity of glass held at the edges limits the size of lenses to about the 40 inch (1 m) size of the Yerkes refractor, made 1895. A mirror, on the other hand, can be supported over the whole of the back of the mirror. Supporting mechanisms or mirror mounts were developed to spread the support evenly and to be effective at all attitudes of the mirror.

Reflecting telescopes are typically made with a paraboloidal primary mirror, which, if perfectly formed, correctly focuses light from a star on its optical axis to a diffraction-limited point image. Off-axis, however, the images are aberrated by coma. A large reflecting telescope thus has a small field of view, perhaps only an arc minute or two, unless a correcting lens is used just ahead of the prime focus. A corrector lens combined with its 3.9 m hyperboloidal primary mirror enables the Anglo-Australian Telescope to have a field of view of up to 1°. The hyperboloidal primary mirror is combined with an ellipsoidal secondary mirror in the AAT's Ritchey–Chrétien optical design.

For survey purposes, such as to make a photographic atlas of the whole sky, an even larger field of view is a great advantage. In 1930, an entirely new optical system was developed by an Estonian researcher, Bernhard Schmidt, with a modified system invented in 1941 by Dmitri

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maksutov. In these reflecting telescopes, the incoming light is collected by a main mirror which is spherical rather than paraboloidal. Normally this would introduce serious image distortions, but these are removed by means of a glass corrector plate fixed at the upper end of the tube. The SCHMIDT TELESCOPE corrector has a complex aspheric profile, the Maksutov corrector has spherical surfaces. The detector, say a photographic plate, a photographic film, or a mosaic of CCDs, is situated at the prime focus. The main problem is that the focal plane is curved, and so the detector has to be shaped to the surface of a sphere. A special plate-holder has to be used to bend a glass photographic plate into the correct form, or the CCDs have to be mosaicked onto a spherical surface. The end result is that sharp star images can be obtained over very wide fields, often well over 10° across. The Oschin Schmidt telescope at Palomar, in California, has a 48 in (1.2 m) corrector plate feeding its 1.8 m primary mirror and was used to make the Palomar Observatory Sky Survey of the northern hemisphere. It is optically similar to the UK Schmidt Telescope in Siding Spring Observatory. The 1.35 m Schmidt at Tautenberg, in Germany, is the largest Schmidt telescope of classical design, whose size is limited by the same issue of support of the optical corrector lens as has restricted refractors to the 1 m class. However, a 4 m Schmidt telescope, LAMOST, is under construction near Beijing for multi-object spectroscopy. Its spherical mirror is fed by a celostat on whose almost flat reflecting surface the appropriate corrections are figured.

In 1908 George Ellery Hale masterminded a 60 in (1.5 m) reflector, which was set up on Mount Wilson in California. It was not so large as the Rosse 72 in, but had a silvered glass mirror and a fully maneuverable equatorial mounting. The tube was a structure of rigid, triangular, skeletal girderwork, rather than being solid walled tube. This reduced weight, improved air flow and seeing, but held all the optical components in exactly the right positions. Most modern reflectors have similar open support structures, which nevertheless are referred to as telescope 'tubes'.

Another design driver, for the tube of telescopes like the 4 m UK InfraRed Telescope on Mauna Kea, the largest designed specifically for infra-red use, is the desire to reduce the amount of structural material intruding into the infra-red beam, so as to limit the thermal radiation created as infra-red background in the telescope.

Hale made a 100 in (2.5 m) equatorially mounted reflector, again set up on Mount Wilson and financed mainly by a Los Angeles millionaire, John D Hooker, after whom it is named. Before he died, Hale planned the reflector with a 200 in (5 m) mirror on Mt Palomar named in his honor. The Hale telescope was completed in 1948 and was the largest telescope in the world for a generation, with epoch-making discoveries to its credit too many to list. The telescope incorporated the 'Serrurier Truss,' newly invented by Mark U. Serrurier of CalTech in Pasadena in 1935 as a way of controlling flexure in the tube of the telescope that he helped to design. Serrurier conceived of a structure that bent as little as possible, but if it did deflect (as was inevitable to some extent), it kept the primary and secondary mirrors in the correct relative positions. The telescope tube includes a cubical section at its center of gravity where the declination axis is attached. Eight triangular trusses are fixed to the corners of the cube, four pointing up to the top end ring which supports the secondary mirror, and four to the primary mirror cell. Four opposite pairs of the triangular trusses oppose bending by gravity, while the four interposed pairs form a parallelogram with the end-rings such that motion of the end rings is of equal magnitude and in parallel, thus keeping the optical elements collimated in spite of flexure. The Gemini telescopes use a variant of this principle, with the large cubical sections at the center of gravity eliminated because of their effect on thermal input into the light beam of the telescopes.

A telescope has two principal degrees of freedom, the axes around which it rotates to point to and follow a star. To a first approximation, an equatorially mounted telescope tracks by rotation about one axis, at constant speed. However, mechanical flexure changes the optical axis of the telescope with two additional degrees of freedom. The change of focus, whether due to mechanical flexure or thermal expansion, is another. Although the telescope mounting structure will have been made to the highest standards, nevertheless the axes will not be exactly orthogonal, nor aligned exactly on the sky. Its bearings will not be exactly circular nor exactly centered, and the gears that drive the telescope will have periodic errors introduced in their manufacture. Refraction of starlight by the curvature of the atmosphere means that the sky does not appear to rotate uniformly. All these effects limit the accuracy of telescope performance.

With the development of computers, it was possible to coordinate the motions of the telescope structure at a high enough rate to compensate for these departures from ideal. In 1974, building on the experience of radio astronomers in controlling altazimuth radio telescopes, the 3.9 m Anglo-Australian Telescope successfully implemented computer control of its equatorially mounted structure at the arc second level. This paralleled the successful computer control of altazimuthally mounted optical NASMYTH TELESCOPES, like the 6 m Bolshoi Teleskop Azimutalnyi (BTA). Virtually all the current generation of large telescopes (over 4 m) are altazimuth designs. The BTA telescope, however, revealed just how important it was to control the thermal environment of a large mirror, otherwise diurnal changes of temperature propagate into the mirror, causing an ever-shifting pattern of thermal distortion.

Symmetrically oriented in gravity, altazimuth telescopes can be made stiffer, and hence could respond better to control. The dome can be fitted more symmetrically around an altazimuth telescope, made more compact and cheaper—the less money of a telescope project spent on a building, the more that can be spent on the telescope and its instrumentation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The 30 largest reflecting telescopes

Name and abbreviation	Location	Size (m)	Date	Website URL
Keck Telescope I	Mauna Kea, Hawaii	10.0	1991	astro.caltech.edu/observatories/keck
Keck Telescope II	Mauna Kea, Hawaii	10.0	1996	
Gran Telescopio de Canarias (GTC)	La Palma, Spain	10.0	2002	www.gtc.iac.es
Hobby-Eberly Telescope (HET)	Mt Fowlkes, Texas	9.2	1999	www.as.utexas.edu/mcdonald/het/het.html
South African Large Telescope (SALT)	Sutherland, S. Africa	9.2	2003	da.saao.ac.za/∼salt/
Large Binocular Telescope (LBT)	Mt Graham, Arizona	8.4	2002/04	lbtwww.arcetri.astro.it/
Subaru	Mauna Kea, Hawaii	8.3	1999	www.naoj.org/
Antu (VLT 1)	Cerro Paranal, Chile	8.2	1998	www.eso.org/outreach/info-events/ut1fl/
Kueyen (VLT 2)	Cerro Paranal, Chile	8.2	1999	
Melipal (VLT 3)	Cerro Paranal, Chile	8.2	2000	
Yepun (VLT 4)	Cerro Paranal, Chile	8.2	2000	
Gemini North	Mauna Kea, Hawaii	8.1	1999	www.gemini.edu
Gemini South	Cerro Pachon, Chile	8.1	2000	www.gemini.edu
MMT	Mt Hopkins, Arizona	6.5	2000	sculptor.as.arizona.edu/foltz/www/mmt.html
Magellan Telescopes I and II	Las Campanas, Chile	6.5	2000	medusa.as.arizona.edu/mlab/mag.html
Large Zenith Telescope (LZT)	Vancouver, BC	6.1		www.astro.ubc.ca/LMT/lzt.html
Bolshoi Teleskop Azimutalnyi (BTA)	Mount Pastukhov, Russia	6.0	1975	www.sao.ru
200 in Hale Telescope	Mount Palomar, Ca	5.0	1948	astro.caltech.edu/observatories/palomar/
William Herschel Telescope (WHT)	La Palma, Spain	4.2	1987	ing.iac.es/WHT.html
SOAR	Cerro Pachon, Chile	4.2	2001	www.lna.br/soar/soar_e.html
Victor Blanco Telescope	Cerro Tololo, Chile	4.0	1976	www.ctio.noao.edu/telescopes/4m/base4m.html
Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)	Xinglong Station, China	4.0	2004	159.226.88.56
Anglo-Australian Telescope (AAT)	Siding Spring, Australia	3.9	1975	www.aao.gov.au
Mayall Reflector	Kitt Peak, Arizona	3.8	1973	www.noao.edu/kpno/kpno.html
UK InfraRed Telescope (UKIRT)	Mauna Kea, Hawaii	3.8	1978	www.jach.hawaii.edu/UKIRT/home.html
3.6 m Telescope	La Silla, Chile	3.6	1977	www.ls.eso.org/lasilla/Telescopes/360cat/html/tel360.html
Canada-France-Hawaii Telescope (CFTHT)	Mauna Kea, Hawaii	3.6	1979	www.cfht.hawaii.edu
Largest reflecting telescopes in space				
Next Generation Space Telescope (NGST)	Lagrangian point, L2	8	2007	ngst.gsfc.nasa.gov/index.html
Hubble Space Telescope (HST)	Low Earth orbit	2.4	1990	www.stsci.edu
Largest reflecting telescope for public education: Faulkes Telescope (FT)	Haleakala, Hawaii	2.0	2001	www.faulkes.com

The Gemini Telescopes at 8 m, the four telescopes of the VLT at 8.2 m and the Subaru Telescope at 8.3 m have probably reached the practical limit of a thin monolithic mirror, whose optical figure is computer controlled by actuator pads—the transportation of such a mirror from the factory to the site is a major issue, solved in the case of the BTA in the then Soviet Union by construction of a special railway. To transport the monolithic mirrors of the 8 m Gemini telescopes, roads were widened, the radius of curvature of bends reduced and detours constructed around low bridges.

To make larger aperture reflecting telescopes, the solution is to make a segmented mirror, in which the various smaller components are fitted together to make the correct optical curve. This technique was developed in reflecting telescopes working at wavelengths longer than light, such as the James Clerk Maxwell Telescope on Mauna Kea for sub-millimeter wave astronomy, where 276 individual panels are assembled into a mosaic, each of which can be moved by 3 motorized adjusters. Its first realization for optical astronomy was in the original 4.5 m Multi-Mirror Telescope in Arizona, before its six mosaicked mirrors were replaced by a monolithic 6.5 m.

The Keck telescopes, also on Mauna Kea, have hexagonal reflecting mirrors 12 m across (10 m diameter effective circular area). Each hyperboloidal mirror is made up of 36 hexagonal segments, 1.8 m in diameter, fitted into a honey-comb pattern like tiles. Although each telescope is more than twice the aperture, and four times the area of the Palomar 200 in (5 m) telescope, the total mass of glass in each is little more than 14 tons—about the same as the Palomar mirror. The Gran Telescopio de Canarias, under construction for La Palma, also has a similarly segmented 12 m diameter mirror, essentially replicating the Keck

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

design. As the largest telescopes in the world, the Keck telescopes are setting the pace of discovery among optical observatories.

The overall mirror figure of a segmented mirror is typically maintained through active computer control of the mirror support system, referenced to laser beams. In active optics, the shape of the main mirror is continually modified by computer-controlled pads behind it, so that the shape can be maintained very precisely when the mirror is moved around to point to different directions in the sky. Acceptance that the mirror will bend and provision of a compensating control mechanism makes it possible to reduce the thickness of a mirror, reducing its weight and the strength of the telescope support, and the thermal inertia of the mirror, with benefits in thermal distortion and cost reduction. An alternative method to reduce mirror mass is through a hollow honeycomb construction of the mirror material, used in the two telescopes of the Large Binocular Telescope with 8.4 m mirrors.

With adaptive optics, a relatively bright star in the field of view is monitored, either a naturally occurring star or an artificial star generated in the upper atmosphere by a laser beam. Reacting to the degradation of the image of the monitor star, a flexible optical component in the telescope beam (a small mirror) is continuously modified in shape, to remove distortions due to turbulence in the Earth's air. This technique is more effective at infra-red wavelengths, as optimized in the design of the 8 m Gemini telescopes, which reach the diffraction limit of resolution.

An alternative approach with large reflectors to increase angular resolution is to combine beams from separate telescopes situated tens or hundreds of meters apart. The two Keck telescopes and the Large Binocular Telescope will be used this way, as will the world's greatest telescope at Parañal in Chile; this is the VLT (Very Large Telescope) which will use four mirrors, each 8 m across, working separately as four powerful 8 m telescopes or together as an interferometer.

Seeing conditions at Parañal and other mountain-top observatories, where large reflectors are nowadays constructed, are excellent, but of course for 'perfect' seeing it is necessary to be above the Earth's atmosphere, and this has been achieved with the Hubble Space Telescope, launched from the Shuttle on 25 April 1990. It has a 2.4 m mirror. The size of the monolithic reflecting mirror of the HST was limited by the capacity of the Space Shuttle's cargo bay. In spite of its relatively small size, the HST is the reflecting telescope with the most impact of modern times. The 8 m Next Generation Space Telescope (NGST) reflecting telescope, optimized as a successor to the HST for observations in the near infrared, will be made of segments which unfold like an umbrella. Thermal distortion of the mirror due to solar, terrestrial and lunar thermal radiation, will be controlled by sunshades and passive thermal radiators. In weightlessness, gravitational deflection will not be an issue!

A completely different approach to the provision of large reflecting telescopes is to make their mirrors of a rotating reflective liquid (mercury), spinning on a vibration-free bearing such that centrifugal force builds up the correct parabolic dish shape (as first proposed by Isaac Newton). The 3 m NASA Orbital Debris Observatory (NODO) telescope in New Mexico and 2.7 m Liquid Mirror Telescope of the University of British Columbia are as far as this technology has been proved, but a 6 m Large Zenith Telescope is under construction and a 42 m equivalent aperture array of 18 10 m telescopes with combined focus is planned. Such telescopes are restricted to staring at the zenith, and can be used to sample the sky, whether for space debris or for stars or galaxies.

Another large innovative reflecting telescopes is the Hobby-Eberly Telescope, a spherical reflector of 91 interchangeable segments forming an 11×10 meter hexagon. The telescope structure is tilted at a fixed angle of 55° above the horizon (35° zenith angle). The structure rotates 360° about a vertical axis. It can be rotated to catch stars and galaxies as they rise or set across the small circle centred on the zenith at 55° altitude, and track them via a moving instrument carriage for long enough for spectroscopy (up to 2.5 hours). The South African Large Telescope (SALT) destined for the observatory at Sutherland is a copy of this design.

It is believed that there are no theoretical limitations to the manufacture of ground based reflecting telescopes up to the 50 m to 100 m class. Design studies are in progress for the OWL (Overwhelmingly Large Telescope) or MaxAT (Maximum size Astronomical Telescope); they will surely be projects with global participation.

Ever since Galileo had such spectacular success in 1608 when he first turned a telescope to the sky, astronomers have been seeking ways to make larger and more accurate telescopes in their quest for more light. The present large telescopes of the 8 to 12 meter class are fine examples of what can be done. Engineers, technologists, scientists, administrators and politicians have been inspired to work together to address questions that, in the 21st century, are the modern equivalent of the questions tackled by Galileo in the 17th century. In the words of astronomer Fred Hoyle, put into the mouth of Britain's Prince Charles at the dedication of the Anglo-Australian Telescope in 1974, 'reflecting telescopes are examples of the sorts of things our civilization does well'.

The largest reflecting telescopes operating or under construction are listed in table 1.

Paul Murdin and Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reflection Nebulae

Dust in interstellar space interacts with light primarily by scattering. When a concentration or cloud of dust happens to be near a bright source of light such as a star or group of stars, the resultant scattered light stands out against the sky background and identifies this cloud as a reflection NEBULA. Given that the source of light is almost invariably starlight, a defining characteristic of reflection nebulae is a continuous spectrum, matching that of the illuminating sources in all spectral features except for the color. The color of reflection nebulae, in most instances, is bluer than that of the illuminating radiation, because the scattering efficiency of interstellar dust particles increases with decreasing wavelength throughout the visible range of the spectrum of light.

Most commonly, reflection nebulae result from the close association of a cloud of interstellar dust and a single dominant stellar source of illumination. The cause for this association can be found in each case in one of three scenarios: most frequently, dense cloud material left over from the formation of a relatively luminous star is being illuminated by the new star. Most of the resulting reflection nebulae are therefore associated with stars known to be young and luminous and relatively abundant, i.e. stars of spectral type B. Stars even more luminous and hotter than B stars emit sufficient ultraviolet radiation to ionize the gas in their surroundings. In this case, the emission from the gas will dominate the spectrum of the resulting nebulae, which are then known as emission nebulae, although a faint scattered light continuum produced by the nebular dust is commonly observable in these objects as well.

The second most common form of reflection nebula results from the scattering of starlight by the dust produced by condensation of particles within a mass outflow from a star itself. Such stars are usually in advanced stages of their evolution and their spectral types are generally characteristic of stars cooler than a B star. In most instances, the mass outflow from such stars is preferentially directed along the axis of rotation of such stars, and the resulting nebulae can be recognized by their bipolar morphology.

The third class of single-star reflection nebulae, containing relatively few members, results from the random encounter of luminous stars and relatively dense interstellar clouds as they move through galactic space. They are distinguished from the reflection nebulae mentioned first on the basis of differences in the RADIAL VELOCITY of the nebula and that of the star, revealing their relative motion with respect to each other. The well-known reflection nebulae in the Pleiades STAR CLUSTER are examples of this third class.

The total number of known reflection nebulae is of order 200, with those located north of -33° declination cataloged by van den Bergh (1966). A fairly complete collection of photographs of reflection nebulae and their environments is shown in the *Atlas of Galactic Nebulae* by Neckel and Vehrenberg (1985–90).

Much fainter than the reflection nebulae illuminated by individual stars are extended nebulosities arising from scattering of the general interstellar radiation field of the MILKY WAY GALAXY by dust in isolated interstellar clouds. While such clouds can be observed as individual structures at high galactic latitudes, at lower galactic latitudes the scattering by multiple, overlapping clouds produces the so-called diffuse galactic light, which contributes about 20% of the overall glow of the Milky Way. In this sense, the entire Milky Way galaxy or any other dusty galaxy is a reflection nebula as well.

Reflection nebulae as astrophysical laboratories

Since the early days of the twentieth century and the work of Slipher, Hertzsprung and Hubble (1922), reflection nebulae have attracted the attention of astrophysicists. Much of observational astrophysics deals with the interaction of matter and radiation in cosmic environments, and reflection nebulae are extremely useful laboratories for such studies. The density of the illuminating radiation is typically a few hundred times higher than generally available in interstellar space, which makes the interactions more easily observable. The variable angular offset from the illuminating star provides a mechanism for varying this density over a large range; and the choice of nebulae with illuminating stars of different temperatures allows one to study the effects of radiation fields with varying fractions of ultraviolet light on the dust and gas in these nebulae. Over the years, particular attention has been focused on a number of important astrophysical processes, among these the studies of light scattering by dust; the heating of dust particles by stellar radiation, in particular the stochastic heating of nanoparticles and the resultant non-equilibrium thermal radiation; the photoluminescence by dust particles, also known as extended red emission; the excitation of the near-infrared emission bands, now attributed to small aromatic hydrocarbon particles; the excitation of fluorescence by molecular hydrogen and the photodissociation of the molecules associated with this process; and the far-infrared thermal emission by dust grains heated by starlight. In the remaining sections, these topics and their relationship to reflection nebulae will be reviewed briefly.

Dust scattering properties

The first extensive quantitative studies of reflection nebulae were carried out by HUBBLE, who discovered a close correlation between the apparent magnitudes of the illuminating stars and the angular sizes of a large sample of nebulae, measured on photographs obtained under identical conditions. This suggested that reflection nebulae are largely illumination-bounded, i.e. that their observable extent is set by their surface brightness falling below the level of detectability well before the dust boundary is reached. The constants derived from the observed correlation indicated that the dust albedo, i.e.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the fraction of the light being scattered by an illuminated dust grain, is above 0.5, a result which stands firm to this day.

Scattered light studies involving individual reflection nebulae are complicated by the largely unknown geometry, which leaves unanswered the question of where the scattering dust is located with respect to the plane containing the star. However, two factors help to alleviate this problem. Firstly, since interstellar dust grains scatter predominantly into the forward direction, nebulae in which the dust lies between the star and the observer are expected to be much brighter than nebulae where the dust lies in the same plane as the star or behind the star. The brightest nebulae, i.e. those selected most frequently for intensive study, are therefore invariably cases where the star is located behind the scattering material. Second, modern observing tools allow surface brightness measurements in nebulae over a wide range of wavelengths, from the far-ultraviolet to the near-infrared. Even if the dust/star geometry of the nebulae is not known in detail, an argument can be made that the geometry is the same at all wavelengths, whatever it may be. Important information can thus be obtained from the relative changes that can be seen as a function of wavelength. It is from studies of this kind that we now know that the dust albedo in the near- and far-ultraviolet is generally as high or higher than that derived for the visible by Hubble, and that the degree of forward scattering is increasing with decreasing wavelengths. It has also been found that scattering is still rather effective at near-infrared wavelengths (0.8–2.5 μm), leading to estimates of the dust albedo which exceeds values predicted by current dust models. This difference can be understood if reflection nebulae contain dust grains with larger upper size limits than generally accepted for the diffuse interstellar medium of the Milky Way galaxy. Thus, empirically determined scattering properties of grains in reflection nebulae provide useful constraints for dust models. These characteristics identify the scattering dust grains as particles comparable in diameter to the size of the visible wavelengths. As will be shown below, however, these relatively large grains are only part of the story.

Dust emission by very small grains

Starlight which interacts with nebular dust but is not scattered is absorbed. The absorbed energy, which is mainly in the form of stellar photons of optical and ultraviolet wavelengths, is ultimately re-emitted by dust grains by two processes taking place at longer wavelengths, namely thermal radiation and photoluminescence. Our knowledge of these processes has expanded greatly through studies of reflection nebulae.

The spectrum of the thermal radiation from dust depends on the temperature distribution of the grains and the emission characteristics of the dust materials. Given the relative proximity of the dust to the illuminating star(s) in reflection nebulae, coupled with the fact that the radiation density will decrease with the inverse square of the distance from the star, it was generally expected that grain temperatures in reflection nebulae would be somewhat higher than those found in the diffuse interstellar medium far from any star and that they would span a certain range, e.g. from 30 K to 50 K, versus 20 K in interstellar space. The thermal radiation anticipated from any grains at these temperatures occurs at far-infrared wavelengths longward of 50 μm, and it is indeed observed there. However, a remarkable discovery by Sellgren (1984) demonstrated conclusively that in addition to the expected far-infrared radiation, thermal radiation emerges from dust distributed throughout reflection nebulae which indicates grain temperatures in the range from 500 K to over 1000 K, much higher than expected from equilibrium heating of the larger grains. This additional thermal radiation emerges at near-infrared wavelengths. Grain temperatures in the 500–1000 K range cannot be generated under equilibrium conditions in the nebular environment, because the density of radiation is not nearly high enough. Such high temperatures can be explained only if the radiating INTERSTELLAR GRAINS are so small that the energy of a single stellar photon is large compared with the heat content of the grain, so that its absorption will increase the grain's temperature temporarily to values as high as 1000 K or more. Such grains must be exceedingly small and can contain only between 30 and 100 atoms, with a typical size of 1 nm. Such particles are basically equivalent to large molecules and are referred to as nanoparticles. Stochastically heated nanoparticles were subsequently found by the Infrared Astronomy Satellite (IRAS) to be a common component of the interstellar medium of the Milky Way galaxy as well of other dusty galaxies, where they generate typically about 30% of the total infrared radiation emitted by dust.

Still another unexpected radiative process involving interstellar grains was discovered first in reflection nebulae and continues to be studied in these objects (Witt *et al* 1984). This process is photoluminescence by interstellar grains, also known as extended red emission (ERE). The ERE emerges in the form of a broad, featureless emission band in the 540–950 nm spectral range. An important characteristic of this emission band is that the wavelength of its peak varies both within a given nebula and from nebula to nebula. The determining environmental parameter for this variation appears to be the density of ultraviolet photons. In reflection nebulae, this photon density is a function of the temperature and luminosity of the illuminating star and the distance from the star. One observes that the peak of the ERE band shifts to shorter wavelengths with increasing angular offset, and nebulae with somewhat cooler stars generally show the ERE band with a shorter peak wavelength than those with hotter exciting stars. Recently, the ERE has been detected in the diffuse interstellar medium of the Milky Way galaxy as well, and as expected from the observed behavior in reflection nebulae, the peak wavelength of the ERE, found near 610 nm in the diffuse interstellar medium,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is shorter than in any nebula, consistent with the fact that the radiation field in the diffuse interstellar medium is of lower density than that found in any single nebula.

Just like the stochastic heating of nanoparticles, the ERE is thought to be a single-photon process involving particles or large molecules in that same size domain. Assuming that the emission of a single ERE photon requires the excitation by a single photon of higher energy, most likely in the ultraviolet, one can estimate from the observed ERE intensities in reflection nebulae and the diffuse interstellar medium that up to 10% of absorbed ultraviolet photons are needed to generate the ERE. This suggests that the ERE is produced by a very common grain species, made from cosmically abundant elements. Current proposals point to oxidized silicon nanoparticles or partially hydrogenated carbon clusters as likely candidates. These same particles are almost certainly also contributing to the non-equilibrium thermal emission at near- and mid-infrared wavelengths, because the ERE process consumes only a fraction of the energies of the exciting photons, the remainder being available to heat the nanoparticles to high temperatures.

Additional evidence for the existence of very small carbonaceous particles with a preferential aromatic bonding arrangement comes in the form of near-infrared emission bands at wavelengths of 3.3, 6.2, 7.7, 8.7, 11.3 and 12.7 μm, which are frequently observed in reflection nebulae. Again, the emitting particles must be excited to an equivalent temperature near 1000 K in order to radiate at these wavelengths. Proposed identifications include polycyclic aromatic hydrocarbon (PAH) molecules or grains composed of amorphous, aromatically bonded hydrocarbon units. These same emission bands have also been observed at correspondingly lower intensities in the diffuse interstellar medium of the Galaxy, indicating that these nanoparticles are ubiquitous throughout space.

Emission from molecules

Many reflection nebulae are associated with star-forming regions. The gas in these environments is typically very clumpy and consists of molecules, predominantly molecular hydrogen. Reflection nebulae are thus also ideal laboratories for studying the interaction of stellar light with this most common form of interstellar molecule. Far-ultraviolet photons with wavelengths shortward of 110 nm cause electronic excitations in hydrogen molecules. The decay of an excited hydrogen molecule is associated with line and continuum luminescence in the 100–165 nm ultraviolet part of the spectrum and a host of emission lines longward of 1 μm in the near-infrared, the latter resulting from cascade transitions between vibrational states of the electronic ground state (Gatley *et al* 1987). Well over 100 of these lines can be seen in bright reflection nebulae. About 10% of the decays of the excited hydrogen molecules result in outright photodissociation of the molecules, which is the primary method by which molecular hydrogen is returned to atomic form in the interstellar medium. The observation and analysis of the ultraviolet and near-infrared emissions from molecular hydrogen in reflection nebulae lead to detailed information about the local densities and radiation field intensities in these hydrogen photodissociation regions. Models which make specific predictions for the relative intensity of the various emission features can thus be tested and used to determine the physical state of the gas in star-forming environments. Again, it was in reflection nebulae that the characteristic ultraviolet and near-infrared luminescence of excited molecular hydrogen was first observed in interstellar space, and these phenomena continue to be studied there with preference.

Bibliography

Gatley I *et al* 1987 *Astrophys. J.* **318** L73
Hubble E P 1922 *Astrophys. J.* **56** 400
Neckel Th and Vehrenberg H 1985–90 *Atlas of Galactic Nebulae* vols 1–3 (Duesseldorf: Treugesell)
Sellgren K 1984 *Astrophys. J.* **277** 623
van den Bergh S 1966 *Astron. J.* **71** 990
Witt A N, Schild R E and Kraiman J B 1984 *Astrophys. J.* **281** 708

Adolf Witt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reflector

An optical or other telescope that utilizes the reflection of light from one or more mirrors to form an image of a distant object. When reflected from the front surface of a concave mirror (the primary mirror), rays of light from a distant point source, such as a star, initially traveling parallel to the optical axis (the line through the center of the mirror, perpendicular to its surface), converge to a focus where a point-like image of the source is formed. Rays entering at an angle to the optical axis form images at the focal plane, a plane perpendicular to the optical axis that passes through the focus. The size of the image formed at the focal plane is proportional to the focal length of the mirror.

In x-ray reflecting telescopes (like XMM), the x-rays are incident at grazing reflection on to the reflectors and focused into a focal plane. The images are highly aberrated.

In most reflecting telescope systems, the converging cone of light from the primary is reflected from one or more secondary mirrors. In the Newtonian system, the secondary is a small flat mirror that reflects the focus to the side of the telescope tube, where the eyepiece is located. In the Cassegrain and Gregorian systems, and their derivatives, light from the secondary (which is convex in the former and concave in the latter) passes back down the telescope tube, through a hole in the center of the primary, to the eyepiece. Because their secondaries have curved surfaces, they change the angle at which the rays of light reflected from the primary are converging, and so increase the effective focal length and focal ratio of the instrument. A Cassegrain or Gregorian telescope, therefore, has a primary of relatively short focal length, which minimizes the overall length of the instrument, but has a long effective focal length.

The instrumentation used to detect, record or analyze the image (e.g. a photographic plate, electronic detector or spectrometer) can be attached directly at the Cassegrain focus. In large telescopes the instrumentation may be placed directly at the prime focus (the focus of the primary mirror) and, in some cases, the observer can be housed in a 'cage' at the prime focus. In either of these cases, the instrumentation has to swing round with the telescope as it points towards different parts of the sky. To minimize, or eliminate, the need for the telescope structure and mounting to shift heavy instrumentation around, some telescopes incorporate additional mirrors to direct the focus to a fixed plane (e.g. the Nasmyth focus) or a fixed position (the coudé focus).

Although the first design for a reflecting telescope (the Gregorian) was produced in 1663 by the Scottish mathematician, James Gregory (1638–75), the first working reflector was designed and built in 1668 by Sir Isaac Newton (1642–1727), this being the prototype of the Newtonian type. The Cassegrain design was produced in 1672 by the Frenchman, Guillaume Cassegrain, but did not become popular until much later because of the technical difficulties involved in making its convex hyperboloidal secondary. Early reflectors used solid metal mirrors that had relatively poor reflectivities, tarnished quickly and expanded and contracted as the temperature changed. By the late nineteenth century, these had been replaced by metal-on-glass mirrors in which a thin film of reflective metal (silver or, more recently, aluminum) is deposited on the front surface of a glass substrate. The development of low-expansion Pyrex glass in the 1930s greatly improved the performance of reflecting telescopes. Modern mirrors use ultra-low expansion materials such as quartz or ceramics as their substrates.

The largest reflector with a one-piece ('monolithic') mirror is located at the Zelenchukskaya Astrophysical Observatory, Russia. Its primary mirror has an aperture (clear diameter) of 6 m. The largest individual reflecting telescopes are the two Keck telescopes on Mauna Kea, Hawaii, each of which has a primary mirror, 10 m in diameter, which consists of 36 interlocking hexagonal segments.

See also: Cassegrain telescope, Dobsonian telescope, focal length, focal ratio, focus, Gregorian telescope, mirror, magnification, Maksutov telescope, Newtonian telescope, Schmidt telescope, Schmidt–Cassegrain telescope, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Refracting Telescopes

Nobody is sure of the date when the first telescope was made. It may be as far back as the early 1550s, when there is evidence that a telescope of some sort was constructed in England by Leonard Digges, but it is all very uncertain, and the first telescope of which we have definite proof was built by the Dutch spectacle maker H Lippershey in 1608. Others soon followed and were turned toward the skies. The first systematic telescopic series of observations was made in early 1610 by Galileo; others had preceded him, but for skill and perseverance Galileo stands alone in those very early days.

His telescope was, of course, a REFRACTOR. It was tiny, and even his most powerful telescopes magnified no more than 30 times, so that they were far inferior to modern binoculars, but they were the direct ancestors of the huge refracting telescopes of today.

In a refractor, the light from the target object is collected by a glass lens, known either as an object glass or as an objective. The light is passed down the telescope and brought to focus, where an image is formed and is enlarged by a second lens, termed an eyepiece or ocular. Note that it is the eyepiece which is responsible for the magnification; the function of the object glass is to collect the light—and, naturally, the greater the amount of light collected, the higher the magnification which can be used. The distance between the object glass and the focus is termed the focal length of the telescope. Magnification is given by the focal length of the telescope divided by the focal length of the eyepiece. The focal length of the telescope divided by the diameter of the object glass gives the focal ratio. Thus if a 3 in (7.6 cm) refractor has a focal length of 36 in (91.4 cm), its focal ratio is $36/3 = 12$ (in metric, $91.4/7.6 = 12$).

Assume now that with our 3 in refractor, with its focal length of 36 in, we use an eyepiece of focal length $\frac{1}{2}$ in. The magnification will be $36/\frac{1}{2} = 72$, often written as $\times 72$.

In theory, and usually in practice, eyepieces are made with a standard thread, so that any eyepiece can be used with any telescope. There is, however, one obvious limitation—the amount of light available depends on the aperture of the object glass (or the main mirror of REFLECTING TELESCOPES). Suppose, with our 3 in refractor, we use an aperture of $\frac{1}{8}$ in. The magnification will then be $36/\frac{1}{8} = 288$. Unfortunately, every time an image is enlarged it becomes fainter, and with this power on this telescope the image would be so faint that it would be completely useless. It is a general rule that the maximum really satisfactory magnification is $\times 50$ per inch of aperture, so that for our 3 in telescope the highest power which can be properly used is $3 \times 50 = 150$. For a higher magnification it is necessary to have a larger telescope. For example, a 6 in refractor of focal length 72 in will have a focal ratio of $f/12$. On this, an eyepiece of focal length $\frac{1}{4}$ in will give a magnification of $72/\frac{1}{4} = 288$, which is quite acceptable. If the focal length of our 6 in refractor is only 54 in, the focal ratio will be $54/6 = 9$. This makes for greater convenience, because the tube is shorter, but to make the f ratio too low will introduce other troubles, and one has to strike a happy mean.

With any telescope it is desirable to have several eyepieces: one to give low magnification and a wide field, suitable for observing objects such as star clusters; one with a moderate magnification, for views of the Moon and planets; one with high magnification, for use on really good, clear nights. With our 3 in, $f/12$ refractor, suitable eyepieces might well be of focal length 1 in ($36/1 = 36$), $\frac{1}{2}$ in ($36/\frac{1}{2} = 72$) and $\frac{1}{4}$ in ($36/\frac{1}{4} = 144$).

The main problem with a refractor is that it introduces what is known as chromatic aberration. Light is really a mixture of all the colors of the rainbow, and the color depends on the wavelength of the light; red light has the longest wavelength for visible radiations, violet the shortest, with orange, yellow, green and blue in between. When a beam of light is passed through a lens, it is split up, and the different colors are bent or refracted by different amounts—red least, violet most. They are therefore brought to focus at different distances from the object glass of a refractor, and an object such as a star is seen to be surrounded by gaudy colored rings which may look superficially attractive but which are most unwelcome to the astronomer.

The false-color trouble can be reduced by what is termed an achromatic object glass, in which there are several component lenses made of different kinds of glass; the errors then tend to cancel each other out. Some false color always remains, but with a good achromatic objective it is not really serious.

Bear in mind that an astronomical refractor will give an inverted image. In fact any refractor will do this, but in a telescope made for terrestrial use an extra lens system is put into the optical train to make the image erect. However, each time a ray of light passes through glass it is slightly weakened. This does not in the least matter when looking at birds, or ships out at sea, but it matters very much to an astronomer, who is anxious to collect every scrap of light available. Therefore, the erecting lens system is left out, although an erecting eyepiece can always be obtained.

The first known telescopes were refractors; the first REFLECTOR was not made until around 1669, when Isaac Newton believed that nothing could rid the refractor of severe chromatic aberration, and decided to develop an entirely different system (see NEWTONIAN TELESCOPES). Previously, efforts to eliminate false color meant that refractors were made with very long focal length. This does reduce chromatic aberration, but at the expense of making the telescopes very unwieldy. For example, Christiaan Huygens, probably the best observer of the early 17th century, constructed a telescope with an aperture of 2 in and a focal length of $10\frac{1}{2}$ feet and used it to discover Titan, the largest of the satellites of Saturn. He then built a telescope with a focal length of 23 feet, and with it discovered the true nature of Saturn's ring system. The telescope must have been incredibly awkward to use, and even more so were the refractors made and used by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Some large refracting telescopes.

Name	Observatory	Aperture (in)	Date of completion
Yerkes 40 in	Yerkes, Williams Bay, WI, USA	40	1897
Lick 36 in	Lick, Mount Hamilton, CA, USA	36	1888
Meudon Refractor	Meudon, Paris, France	33	1889
Potsdam Refractor	Potsdam, Germany	31	1899
Lunette Bischoffsheim	Nice, France	30	1886
Thaw Refractor	Allegheny, Pittsburgh, PA, USA	30	1985
Grosser Refraktor	Archenhold, Treptow, Germany	27	1896
McCormick Refractor	Leander McCormick, Charlottesville, VA, USA	26	1880
26 in Equatorial	US Naval Observatory, Washington, DC, USA	26	1873
Thompson Refractor	Herstmonceux, England	26	1897
Innes Telescope	Johannesburg, South Africa	26	1926
Vienna Refractor	Vienna, Austria	26	1880
Newall Refractor	Athens, Greece	25	1862
Lowell Refractor	Lowell, Flagstaff, AZ, USA	24	1895

another pioneer observer, Hevelius of Danzig (the town now known as Gdańsk). One of Hevelius' telescopes had a focal length of no less than 150 feet, but was so subject to wind disturbance that it could seldom be used to its full potential.

Next came the 'tubeless' telescope. Here, the object glass was fixed to the top of a mast, and the observer sighted it by looking along guide wires which could be used to turn the object glass to the right position. The observer then held the eyepiece by hand. One of Huygens' 'aerial telescopes' had a focal length of 210 feet, and it is said that one refractor with a focal length of 600 feet was planned, although there is no record that it was actually built.

Obviously these long-focus telescopes could never be really satisfactory, but in 1733 a wealthy amateur astronomer, Chester Moor Hall, constructed the first achromatic or compound objective, with one component made of flint glass and the other of crown glass. The idea was taken up by John Dollond, and this led on to what was probably the first 'optical firm' in history. In 1765 John's son, Peter Dollond, took one of his achromatic objectives to the Royal Observatory at Greenwich, and its performance was found to be far better than that of the Observatory's best long-focus telescope. The old aerial refractors promptly became obsolete. Moreover, Dollond's telescopes looked attractive. From about 1783 they were made with brass tubes; when set up on well-made mahogany stands they were easy to use and were fitted with slow motions.

By the first part of the 19th century it had become possible to make really good objectives. Then, in 1862, Thomas Cooke, in England, built the first of the 'great refractors'; it is known as the Newall Telescope and had an objective of 25 in across (it is still in use, at the Athens Observatory in Greece). Between 1870 and 1900 even larger refractors were built, culminating in 1897 with the 40 in telescope which was set up at the Yerkes Observatory in Wisconsin, USA. At the time, and for some years afterwards, it was the most powerful telescope in the world, and it proved to be a great success. It is still in use on every clear night.

Could it be possible to build a refractor of greater size than the Yerkes 40 in? Theoretically, yes, but there were serious practical difficulties. If a lens is too large it will begin to distort under its own weight, and this will make it useless. An attempt was made at the very end of the century, when a 49 in objective was made in France. The 180 foot tube could not be mounted in the conventional way; instead it was left horizontal, and the light was brought to it by means of movable mirrors. The telescope was shown at the Paris Exposition of 1901, but it was clearly a failure, and it was never used for any astronomical work. Before long it was dismantled, and the fate of the object glass is unknown. The fate of a 41 in, destined for the Pulkovo Observatory in Russia, was even worse; before the optics could be completed, the mounting of the telescope had rusted away.

By then it had become possible to build very powerful reflecting telescopes, and these were not subject to size limitation, because a mirror can be supported by its back; light does not have to pass through it. A few large refractors have been made in the 20th century, but not many are used by professional astronomers except for very specialized programs, and it is undeniable that the future of the telescope lies with the reflector.

Yet in amateur hands, the refractor remains ideal. It is much less delicate than a reflector and will need little maintenance; if it is treated with care, it will last a lifetime and more. Although its color correction can never be as good as that of a reflector, it will give superb, clear-cut images.

Choosing a refractor for home use is something to be undertaken with care, because a poor-quality telescope does not betray itself at a glance. Some small refractors are sold according to the power available—or so it is claimed; it one recent advertisement it was said that the telescope offered 'would magnify 750 times'. The aperture of the object glass was however only 3 in, so that the maximum power which could be used to advantage was no more than

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

150. Aperture is all-important, and a telescope advertised without giving the diameter of the objective should be avoided. Watch too for a ring fixed inside the top of the tube, masking the outer part of the objective and thereby cutting down the aperture; this is a trick designed to conceal defective optics.

Generally, it is not really sensible to spend much money on a refractor with an object glass less than 3 in in diameter. Of course a smaller instrument, such as a 2 in, is better than nothing at all, but given a choice between a very small telescope and a pair of good binoculars it is certainly better to opt for the binoculars.

A refractor of aperture up to 4 in is easily portable, although a larger instrument will be too heavy to move around, and will be best set up in a permanent observatory.

Some large refracting telescopes are listed in table 1.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Refractor

An optical telescope that utilizes the refraction of light through a lens to form an image of a distant object. In essence, a refractor consists of two lenses, an objective (or object glass) and an eyepiece. The objective is a lens of long focal length that collects light and forms an image of a distant object at its focus, or focal plane. The eyepiece is a lens of short focal length that enables an observer to view a magnified image and which may also be used to project an enlarged image onto a screen, photographic emulsion or electronic detector. In practice both the objective and the eyepiece are compound lenses, each consisting of two or more components.

The invention of the refractor, or refracting telescope, is usually ascribed to the Dutch spectacle-maker, Hans Lippershey (*c*. 1570–1619), who is reputed to have discovered the principle in 1608 when, by chance, he noticed that distant objects were seen to be enlarged when he placed one lens in front of another. The first astronomer to make serious use of the refractor was the Italian scientist, Galileo Galilei (1564–1642), who, with a telescope of his own design and construction, made a succession of epoch-making discoveries in 1609–10. The Galilean refractor used a convex ('positive') objective and a concave ('negative') eyepiece that was placed in front of the focal plane of the objective. It produced erect images, but its field of view was very small and the instrument was difficult to use. An improved design, which is the basis of the modern refractor, was published in 1611 by Johannes Kepler (1571–1630). The Keplerian, or 'astronomical', refractor uses a convex objective and a convex eyepiece that is placed behind the focal plane of the objective. Although it produces an inverted image, it has a wider field of view than the Galilean instrument. Because the eyepiece is located on the observer's side of the focal plane, it is possible to focus simultaneously on the image and on micrometer wires placed at the focal plane.

The performance of early refractors was seriously degraded by several optical defects, in particular, chromatic aberration and spherical aberration (the inability of a simple lens to focus all wavelengths and all rays from a point source at the same focal point). Because the effects of these aberrations could be reduced by using lenses of very long focal length, attempts to overcome these problems led, in the seventeenth century, to the development of extremely long and unwieldy instruments, such as the 46 m long refractor constructed in Danzig by Johannes Hevelius (1611–87). Dutch astronomer Christiaan Huygens (1629–95) dispensed altogether with the telescope tube. His 'aerial telescope' consisted of an objective, which was held in a short tube that was hoisted up a mast, linked to and aligned with an eyepiece by means of a taut cord.

The invention of the achromatic doublet (a lens consisting of two components, the combination of which greatly reduced the effects of chromatic aberration) transformed the development of the refractor and rendered the aerial telescope redundant. Following a series of experiments using different types of glass, which he had carried out between 1729 and 1733, the English amateur optician, Chester Moor Hall (1703–71), instructed professional optician George Bass to construct the first achromatic objective in 1733. From 1754 onwards, telescope-makers, notably John Dollond, began to construct good quality achromatic refractors. Greatly improved glass-making techniques, developed by the German astronomer, Joseph von Fraunhofer (1787–1826), enabled larger objectives of higher quality to be produced. The development of the refractor culminated in the construction of the 1 m ('40 inch') refractor at Yerkes Observatory, Wisconsin, USA, which was commissioned in 1897 and remains, to this day, the largest operational refractor in the world. Because reflecting telescopes are easier and cheaper to build, mount and support and do not suffer from chromatic aberration (which cannot entirely be eliminated from a refractor), large refractors are no longer constructed and all large modern instruments are of the reflecting type.

See also: chromatic aberration, eyepiece, focal length, objective lens, reflector, spherical aberration, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Regiomontanus [Müller, Johann or Johannes] (1436–76)

Born in Königsberg (latinized in his trade name), Mainz (now Germany), a pupil of PEURBACH. Succeeded him as professor of astronomy at the University of Vienna, and became astronomer to King Matthias Corvinus of Hungary. He built an observatory and instrument workshop in Nuremberg. Observed Halley's Comet in 1472 and eclipses of the Moon. This led him to suggest that eclipse timing and techniques that he had used in tracking comets could be used effectively as a clock to determine longitude. Eclipses were in fact too infrequent to be of much use as timing markers, but the general idea was picked up by JOHANN WERNER of Nuremburg and developed into the method of lunar distances (i.e. continuously measuring the distance of the Moon from the Sun). Regiomontanus wrote on calendar reform, and was summoned by Pope Sixus IV to Rome to advise on the topic but died (of the plague or poisoned by his enemies) before he could do so.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Regolith

The loose material covering the surface of a rocky planetary body. Regolith consists of dust and tiny fragments of rock produced by the impact of small meteorites and micrometeorites, and is characteristic of worlds which have no atmospheres and have been essentially geologically inactive for a few billion years. The lunar regolith is 2–8 meters deep on the maria, and as much as 15 meters deep in the highlands, which have had longer exposure to meteoritic bombardment. Its composition largely reflects that of the underlying rock. Mercury has a very similar albedo to the Moon, and is assumed on this basis to possess a similar regolith. Images showing boulders on the surface of the asteroid Ida suggest that it has a regolith.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Regulus

The star α Leonis, apparent magnitude 1.36, a blue-white main sequence star of spectral type B7V. Its name, meaning 'Little King', was given by Nicolaus Copernicus (1473–1543), marking the fact that it had been regarded as one of the leading stars in the sky for some 2000 years. It was one of the four 'Royal Stars' of the Persians, *c*. 3000 BC, who called it Venant, later Miyan ('the Center'); it marked for them the summer solstice, being prominent on June evenings (*see* Aldebaran). At a distance of 78 light-years, it has a parallax of $0.042''$ and absolute magnitude -0.5. There is an eighth-magnitude companion star $2.95'$ distant at position angle 307°.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reiche, Maria (1903–98)

German mathematician and archaeologist, born Dresden, fled the Nazi regime to Peru, identified and researched the huge figures of Nazca drawn in the desert and revealed the knowledge of astronomy of the ancient inhabitants of Peru's coastal region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reinhold, Erasmus (1511–53) and Reinhold, Erasmus (1538–92)

Father born in Saalfeld, Germany, became professor of mathematicum superiorum (astronomy) at the University of Wittenberg. Influential on BRAHE and KEPLER. Revised a work by GEORG PEURBACH about the motion of the planets and calculated the *Prussian Tables* (*Prutenicae Tabulae Coelestium Motuum*) containing calculations about the movements of celestial bodies. The first such calculations made according to the Copernican theory, they became a basis for the reform of the calendar in 1582.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Relative Sunspot Number

An index of sunspot activity originally devised in 1858 by J R Wolf (1816–93) of the Federal Observatory of Zurich and otherwise known as the Zurich sunspot number. The relative sunspot number, denoted by the letter R, is given by the formula $R = k(f + 10g)$, where g is the number of visible sunspot groups, f is the total number of individual spots visible on the face of the Sun and k is a correction factor that takes account of factors such as the size of telescope used, atmospheric conditions and the performance of the individual observer. For example, if on a particular occasion there were four sunspot groups containing 3, 7, 9 and 12 spots respectively, g would be 4 and f would be 31. Assuming $k = 1$, the value of R would be $1(31 + 40) = 71$.

The relative sunspot number places greater emphasis on the number of groups than on the number of individual spots, because the number of groups is more closely related to the number of underlying active regions. Although the relative sunspot number gives a reasonable guide to the changing level of sunspot activity throughout successive sunspot cycles, sunspot areas give a better measure of the overall level of solar activity.

See also: active region, sunspots, sunspot cycle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Relict

Two Soviet astronomical missions. Relict-1 was launched on the PROGNOZ-9 satellite in July 1983. Two small radio telescopes were used to search for any anisotropy (unevenness) in the cosmic background radiation. Relict 2 was launched into a halo orbit at the L2 Lagrangian point, 1.5 million km in the anti-sunward direction, in January 1994. It returned submillimetre wavelength astrophysical and magnetotail plasma data.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Resolving Power

A measure of the ability of a telescope to separate objects that are apparently close together. Resolving power, or resolution, is usually expressed as the minimum angular separation at which two identical point sources of light (for example, two identical stars) can just be distinguished as separate points (if the sources are closer together than this minimum angle, they will appear as a single point). The *smaller* the angular resolving power, the *better* the ability of the telescope to separate close pairs of stars and the greater the amount of detail that will be revealed in the image of an extended object (which can be regarded as consisting of large numbers of points).

The theoretical resolving power, R, of a telescope depends on its aperture (D) and the wavelength (λ) of the light that is being observed. It is given, conventionally, by the Rayleigh limit, whereby $R = 1.22\lambda/D$ (where R is expressed in radians), or $R = 2.52 \times 10^5 \lambda/D$ (where R is expressed in arcsec). For example, for light of wavelength 5×10^{-7} m (500 nm), a telescope of aperture 1 m would have a theoretical resolution of $2.52 \times 10^5 \times (5 \times 10^{-7}/1) = 0.13$ arcsec. By contrast, a telescope of aperture 0.1 m would have a resolution of 1.3 arcsec.

In practice, atmospheric effects limit the resolutions of Earth-based telescopes. Turbulence in the atmosphere smears out star images into disks ('seeing disks') that often are significantly larger than 1 arcsec in diameter. Consequently, the actual resolutions attained even by large telescopes are often poorer than 1 arcsec (the theoretical resolution of a 13 *centimeter* telescope) and seldom, even at the very best sites under the best conditions, better than 0.25 arcsec. Because it orbits clear of the atmosphere, the Hubble Space Telescope, with an aperture of 2.4 m, is able to achieve its theoretical resolving power of about 0.05 arcsec (about 5–10 times better than the largest conventional ground-based telescopes). The deleterious effects of the atmosphere can to some extent be overcome by means of active/adaptive optics or speckle interferometry.

The theoretical resolution of a radio interferometer consisting of two dishes separated by a distance D (the 'baseline') is given (in radians) by $\lambda/2D$, where λ is the wavelength of the observed radiation. For very long baseline interferometers using separations comparable to the diameter of the Earth and wavelengths of around a centimeter, resolutions as fine as 0.001 arcsec can be attained.

See also: active optics, adaptive optics, Airy disk, Dawes' limit, interferometer, radio interferometer, Rayleigh limit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Resonance

A resonance in CELESTIAL MECHANICS occurs when some of the quantities characterizing the motion of two or more celestial bodies can be considered as commensurable, i.e. their ratio is close to an integer fraction. In a simplified form, this can be expressed as

$$T_1 : T_2 = p_1 : p_2$$

where T_1 and T_2 are the actual values of the quantities involved, while p_1 and p_2 are positive integers (1, 2, 3, . . .; $p_1 \leq p_2$). Commensurabilities are frequently found in the dynamical behavior of celestial bodies of the solar system, involving widely different regimes of motion: spin, orbit, secular precessions. Although resonances in real systems usually approximate within a given accuracy of the corresponding exact value, they give important information on the existence of symmetries in the motion of the objects involved, leading to peculiar configurations and the periodic recurrence of gravitational interactions.

The resonances that can be more easily revealed among the solar system's bodies involve the orbital periods of two dynamically interrelated objects, e.g. two planets circling around the Sun or two satellites around a planet. They are usually referred to as mean motion resonances because they can be alternatively expressed by comparing the average orbital angular velocity of the bodies. Thus a mean motion resonance ensures the repetition of certain geometrical configurations: as an example, a 1:2 resonance between the orbital periods of two natural satellites implies that, while the inner body completes two orbits around the planet, the outer has made only one, and therefore conjunctions always occur along the same direction in space.

For many satellite pairs of the outer planets a resonant relation can be found at a high level of accuracy. Within the Saturnian system the pairs MIMAS–TETHYS and ENCELADUS–DIONE are both in the 1:2 resonance, TITAN and HYPERION satisfy the 3:4 resonant relation, while the ratio of the orbital periods of Titan and IAPETUS is close to 1:5. The three Galilean satellites closest to Jupiter, i.e. IO, EUROPA and GANYMEDE, represent an outstanding example of a multiple mean motion resonance. The pairs Io–Europa and Europa–Ganymede are both found in a 1:2 resonance: thus the periods of revolution of Io and Ganymede are also in the ratio 1:4. This peculiar resonance, known as the Laplace resonance, prevents the occurrence of triple conjunctions of the three satellites. As far as the planets are concerned, some are indeed found close to mean motion resonances. The so-called great inequality is a 2:5 resonance which rules the motion of the two most massive planets of our solar system, JUPITER and SATURN, while the 2:3 commensurability between NEPTUNE and PLUTO provides the basic dynamical mechanism for avoiding too close approaches between them. For the Saturn–URANUS (1:3) and Uranus–Neptune (1:2) pairs the corresponding resonant relations are satisfied with less accuracy.

The 1:1 mean motion resonance represents an exceptional case, allowing widely different dynamical behaviors depending on the relative mass and position of the bodies involved. The basic dynamical framework is provided by the so-called restricted three-body problem (one of the masses is negligible with respect to the other two) and its special solutions. In particular the two equilateral Lagrangian points, located along the unperturbed orbit of the secondary body around its primary and displaced 60° ahead or behind its current position, are of special relevance because they possess the necessary stability conditions for real systems to exist. The TROJAN ASTEROIDS are in fact grouped around the triangular Lagrangian points of the Sun–Jupiter system. Being historically the first bodies of the solar system discovered satisfying a 1:1 mean motion resonance, the attribute 'Trojan' is now commonly used to indicate the occurrence of this peculiar configuration. Trojan satellites have been also observed and many more are supposed to exist within the crowded systems of the outer planets. This is the case of two of the largest satellites of Saturn: Tethys shares its orbit with the tiny moonlets (a little more than 10 km in size) Calypso and Telesto, each located close to one of the triangular equilibrium points, while Dione is preceded by Helene, a 16 km long irregularly shaped body. When a 1:1 resonance is not exactly satisfied, it may still lead to peculiar trapping mechanisms, such as the so-called tadpole and horseshoe orbits. The motion of the already mentioned Trojan asteroids and those of Janus and Epimetheus, the two small satellites of Saturn sharing almost the same orbit, are well-known examples of this kind. The question whether the Earth itself has Trojans, small enough to have escaped direct observation, has been repeatedly posed. The advent of orbiting telescopes has allowed us to detect the existence of a thin cloud of dust dispersed along the orbital path of our planet, temporarily trapped in the vicinity of the 1:1 resonance (a possible explanation for the origin of the zodiacal light) while the recent survey of the near-Earth asteroid population has allowed us to identify (3575) Cruithne, an object in an unusual Earth horseshoe orbit.

Mean motion resonances are encountered rather often throughout the solar system, more than due to chance, suggesting that they may provide a basic mechanism for long-term stability. Yet the study of the orbital evolution of objects close to mean motion resonances has shown that a wide variety of dynamical behaviors can be expected, including regular and chaotic regimes. Both characterize the resonant structure of the asteroid main belt, the densely populated region of small-sized celestial bodies extending between the orbits of Mars and Jupiter. Opposite trends are in fact observed in the distribution of asteroidal mean distances from the Sun: regions devoid of objects ('gaps') corresponding to low-number mean motion resonances with Jupiter (e.g. 1:2, 1:3, 2:5), as well as concentrations (e.g. 2:3, 3:4, 1:1). Detailed modeling of the individual cases has shown that resonances might be closely surrounded by chaotic regions. This is the dynamical mechanism

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

responsible for the formation of the 1:3 gap: sudden chaotic jumps in the eccentricity of the orbit may have caused a near-resonant asteroid to become a Mars crosser and eventually be removed from the main belt by close encounters with that planet.

The contradictory tendency to prefer certain resonances and avoid others is observed not only in the asteroid belt but also in other densely populated systems, such as planetary rings. When the structure of Saturn's rings is analyzed in the light of resonances, it turns out that the largest relatively empty region observed, i.e. the Cassini division, is strongly related to Mimas, corresponding to the 1:2 mean motion resonance. Smaller-scale features characterizing the A-ring can be also explained by the resonant effects of the satellites Pandora, Prometheus and Janus, orbiting just beyond the main ring system (see also SATURN: RINGS).

More complex, less intuitive orbital resonances are found when the motion of a celestial body is investigated in detail. The presentation of an orbit as an ellipse fixed in space is only valid as long as two bodies alone are gravitationally interacting: distant perturbations by other bodies have the effect of inducing slow changes in the shape and the orientation of an orbit in space. This means that the eccentricity and the inclination are not constant any more, while the apsidal line (i.e. the line joining apocentre and pericentre) and the line of nodes (marking the intersection of an inclined orbit with the local reference plane) may advance, regress or librate in time. Commensurabilities in the rate of change of these quantities are called secular resonances and have been found to play a fundamental role in several problems of solar system dynamics.

The overabundance of resonant configurations can therefore be considered as the result of a long-term selection process: among the multitude of objects once crowding the solar system, only those which could rely on the strong sheltering action of one or more resonances could dynamically survive. An outstanding example of this kind involves Neptune and Pluto: the unusually large eccentricity of the latter if compared with the almost circular orbits of the other planets has the consequence of letting Pluto's perihelion distance become smaller than Neptune's mean distance from the Sun. Although at present the orbits do not cross owing to their mutual inclination and angular parameters, they might, in principle, bring the two bodies dangerously close to each other. Nevertheless, the already mentioned mean motion resonance combined with the motion of Pluto's perihelion prevents Neptune and Pluto from actually reaching their minimum achievable approach distance. The observation of several newly discovered objects sharing with Pluto the 2:3 mean motion resonance gives an indication of the importance of this mechanism in shaping the dynamical structure of the so-called Transneptunian Belt.

Hyperion, the small irregularly shaped satellite of Saturn, is also similarly sheltered from the perturbing action of Saturn's largest moon Titan. Being closer to the planet and moving faster, Titan periodically overtakes Hyperion: when this happens the two bodies experience a minimum approach distance, while their mutual gravitational perturbations reach a maximum. The existence of a resonance allows conjunctions to take place only when Hyperion is at the apocenter of its orbit, namely at the largest possible distance from Saturn and therefore also from Titan.

Exotic resonant mechanisms have also been found within the solar system. The existence of narrow arcs in Neptune's Adams ring has been explained as the observable effect of a 42:43 mean motion inclined resonance between the ring particles and the neighboring Galatea satellite, which then provides the necessary confinement (see also NEPTUNE: RINGS). Also, a peculiar high-number commensurability strongly characterizes the motion of the Moon around the Earth, as witnessed by the existence of the SAROS cycle, a period of approximately 18 yr and 10 or 11 days (depending on the number of leap years in the interval) which has allowed the prediction of solar and lunar eclipses with remarkable accuracy since ancient times. One Saros encompasses 223 lunar synodic months (the time span between two subsequent conjunctions with the Sun) which nearly equals 239 anomalistic months (the time span between two subsequent perigee passages) and 242 nodical months (the time span between two subsequent passages at the ascending node). From a strictly dynamical point of view this means that any configuration of the Earth–Moon–Sun system closely repeats after one Saros.

Resonances are also found when investigating the coupling between the rotation of a celestial body around its own axis and its orbital motion. If a commensurability exists among the corresponding periods, a spin–orbit resonance is said to occur. The Moon, whose rotation period equals its orbital period around the Earth, thus allowing only one side of its surface to be seen from our planet, is a well-known example of a 1:1 spin–orbit resonance. This particular commensurability, called synchronous resonance, is also found for the vast majority of natural satellites in the solar system, suggesting the existence of an underlying physical phenomenon in action. As a matter of fact, the synchronous resonance can be considered as a direct outcome of the long-term tidal interaction between bodies of finite dimensions (see also TIDES). The presence of a tidal bulge raised by the differential gravitational attraction on the surface of a celestial body determines the onset of internal friction, as the bulge is carried along with the rotation. Usually this leads to a slowdown of the spin rate until a 1:1 resonance is achieved, because the bulge is then permanently locked towards the attracting body. Tides are experienced by both bodies, although with different intensity and timescales depending on their relative size, mass and internal structure; thus the end state of a tidally interacting system is achieved when the revolution and the rotation periods of both bodies are equal, i.e. they revolve with their tidal bulges facing each other. This is observed in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

case of Pluto and its satellite Charon, almost comparable in size, currently trapped into a fully evolved 1:1 spin–orbit resonance. The Earth–Moon system has only partially evolved (owing to lunar tides the length of the day on our planet is at present increasing at a pace of about 2 ms per century), while the other planets have a mass much too large with respect to that of their satellites to be affected by their tides. Mercury represents an exceptional case: because of its proximity to the Sun it is subjected to strong solar tides and has therefore evolved towards a 2:3 spin–orbit resonance. Its present period of revolution is 87.97 days, while its rotation amounts to 58.65 days, showing that, for every two orbits around the Sun, Mercury completes three rotations around its axis.

Thus, resonances have provided a powerful means to explain many puzzling problems in solar system studies. The chaotic behavior of near-resonant asteroidal orbits has allowed us to discover the long-sought dynamical paths for the transport of matter from the asteroid belt to the inner solar system, since a significant fraction of near-Earth asteroids and of meteorites are thought to have originated by collisional evolution within the main belt. The Laplace resonance involving the innermost Jupiter Galilean satellite Io has been indicated as the ultimate cause for the internal tidal heating mechanism which provides the energy for the intensive volcanic activity observed on the surface of the satellite.

Resonances have proven to be a powerful tool for space mission design. Already in 1974, the US MARINER 10 spacecraft could perform repeated flybys of planet Mercury exploiting a 1:2 mean motion resonance between the spacecraft's period of revolution around the Sun and that of the target planet. The geostationary ring, located along the Earth's equatorial plane at a distance of about 36 000 km above its surface, is now widely used for telecommunication purposes because the 1:1 spin-orbit resonance between the Earth's rotation and the orbital period of a geostationary satellite allows a spacecraft to remain permanently stationed over a given longitude.

On a larger scale, when looking at the Sun and the stars in our Galaxy and the existence of extrasolar planets, resonances are also observed. In the rapidly expanding field of astroseismology, the study of solar oscillations exploits the effect of resonances to detect peculiar structures in the interior of the Sun. The confirmation of the discovery of two planets orbiting around a pulsar (PSR1257+12) has been possible thanks to the existence of a 2:3 mean motion resonance among them, which produced a characteristic signature in their mutual perturbations.

Bibliography

Murray C D and Dermott S F 2000 *Solar System Dynamics* (Cambridge: Cambridge University Press)

Roy A E 1988 *Orbital Motion* 3rd edn (Bristol: Adam Hilger)

Ettore Perozzi

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Reticulum

(the Net; abbrev. Ret, gen. Reticuli; area 114 sq. deg.) A southern constellation which lies (to the north-west of the Large Magellanic Cloud) between Dorado and Horologium, and culminates at midnight in mid-November. It was first shown as Reticulum Rhomboidalis (the Rhomboidal Net) on Isaac Habrecht II's celestial globe of 1621 and was renamed by the French astronomer Nicolas L de Lacaille (1713–62), after the reticle in the eyepiece of the telescope with which he measured the positions of the stars of the southern sky in 1751–2. The shortened form was adopted by the IAU in 1922.

A small, inconspicuous constellation, the brightest stars in Reticulum are α Reticuli, magnitude 3.3, and β Reticuli, magnitude 3.8. There are no other stars brighter than fourth magnitude. Interesting objects include ζ Reticuli, a naked-eye double star with two yellow (G1 and G2) components, magnitudes 5.2 and 5.5, separation 5.1′, that have the same proper motion (which is very similar to that of the Sun), θ Reticuli, a close binary with bluish-white (B9) and white (Am) components, magnitudes 6.0 and 7.8, separation 4.1″, and NGC 1559, an eleventh-magnitude barred spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Retrograde Motion

Angular motion in the opposite sense to the prevailing direction (as opposed to direct or prograde motion, which is in the same sense). Motion of a body in the solar system, either axial rotation or orbital revolution, is retrograde if it is in the opposite direction to that of the Sun's rotation: clockwise as viewed from the Sun's north pole. Bodies in retrograde orbits have orbital inclinations between 90° and 180°; bodies with retrograde rotation have axial inclinations between 90° and 180°. Retrograde motion is the exception to the norm, for it is in the opposite sense to the motion of the rotating disk of material from which the Sun and the planets formed, and hence is indicative of some event in a body's history having modified its motion. Of the major planets, Venus, Uranus and Pluto have retrograde rotation, possibly as a consequence of impacts they suffered at some time in their past. The outermost satellites of Jupiter, Saturn and Uranus, and also Neptune's largest satellite, Triton, have retrograde orbits, indicating that these are captured objects. Many long-period comets have retrograde orbits as a consequence of the circumstances in which they were perturbed inwards from the cometary reservoirs in the distant outer reaches of the solar system.

The term is also used for the temporary apparent motion from east to west on the celestial sphere of a superior planet or minor planet, in contrast to its usual direct eastward motion. This retrograde motion occurs around the time of a planet's opposition. As a planet nears opposition, the Earth is approaching it and travelling faster, and the planet's apparent rate of motion slows down, and then reverses direction. Its retrograde motion is fastest at opposition. The reverse occurs after opposition, when the Earth is moving away from the planet: the retrograde motion slows, halts and returns to direct. The two points at which the motion changes direction are called *stationary points*. Because the planets' orbits are inclined to the ecliptic, which defines the plane of the Earth's orbit, the direct–retrograde–direct sequence causes the planet to describe what is known as a *retrograde loop* on the celestial sphere, or sometimes an S-shaped curve. The more distant a planet, the smaller its retrograde loop.

Compare: direct motion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Revolution

The motion of one celestial body about another, for example the motion of a satellite around a planet or of a planet around the Sun. The Earth completes one revolution of the Sun in a year.

See also: axis, orbit, rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rhea

The largest of the mid-sized icy satellites of Saturn, with a diameter of 1528 km, discovered by Giovanni Cassini in 1672. It orbits at a distance of 527 000 km. The surface was well imaged by Voyager 1. Unlike the other regular mid-sized satellites of Saturn, Rhea shows little diversity of terrain—the surface is saturated with craters. The 250 km diameter crater Izanagi is the exception: the majority of Rhea's craters are around 20 km in size. It has been suggested that this even population of craters was produced by debris from the collision that left the satellite Hyperion, orbiting farther out, with its irregular shape. There are some ridges and scarps, interpreted as resulting from a global compression, as is thought to have occurred on Mercury. The lack of any orbital resonances with other satellites is what has left Rhea for the most part geologically inactive.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rheticus [Rhaeticus; Lauchen, Georg Joachim von] (1514–74)

Born in Feldkirch, Austria, and started an extraordinary life with the name Georg Iserin. After his father was executed for sorcery and his name was abandoned, Iserin took a germanized form of his mother's maiden name and then the name of Rheticus (the Roman name of the province where he had been born). Taught mathematics and astronomy at the University of Wittenberg, visited COPERNICUS for two years, on the basis of which he published and introduction to the heliocentric theory, the *Narratio Prima* or *First Report to Johann Schöner on the Books of the Revolutions of the Learned Gentleman and Distinguished Mathematician, the Reverend Doctor Nicolaus Copernicus of Torun, Canon of Warmia, by a Certain Youth Devoted to Mathematics*, which established the Copernican hypothesis with credibility. Rheticus negotiated with the authorities for the publication of Copernicus' *De Revolutionibus*, at first publishing the trigonometrical sections, with his own calculation of a table of cosines (the first published). As he neared the end of the preparation of the text of *De Revolutionibus*, he was offered a post at the University of Leipzig at greatly increased salary and left the completion of the task to OSIANDER. Rheticus had to flee a homosexual scandal in Leipzig and found himself in Cracow as a doctor, working in his spare time and with a group of assistants on trigonometric tables, making instruments and observing in astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Riccioli, Giambattista [Giovanni Battista] (1598–1671)

Italian astronomer, born in Ferrara, became a Jesuit and came into conflict with the Copernican system, like OSIANDER refuting it while acknowledging its use as a mathematical hypothesis. Mapped the Moon and introduced some of the names still used, in a chart published in the *New Almagest* in 1651. As a follower of the Ptolemaic system, he named major lunar craters after HIPPARCHUS, PTOLEMY and BRAHE (and a big crater after himself), and gave the names of COPERNICUS and ARISTARCHUS to smaller ones. He experimented with falling bodies, intending to refute GALILEO and show that the Earth was at rest. His experiments, however, confirmed Galileo's concept of uniformly accelerated motion. He made general astronomical observations—of sunspots, Jupiter and Saturn, atmospheric refraction, and the positions of the stars. His observation in 1650 that Mizar was a double star was the first discovery of such a star with a telescope. He discovered Venus' ashen light phenomenon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Richard of Wallingford (1291/2–1336(?))

Born in Wallingford, Oxfordshire, England, became abbot of St Albans, went to Avignon to have his appointment confirmed by the Pope, and contracted leprosy (or perhaps syphilis, scrofula or tuberculosis), dying young. He made and described an astronomical clock and an elaborate equitorium called Albion, a sort of analog computer that could make Ptolemaic calculations of lunar, solar and planetary positions, including eclipse predictions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Richer, Jean (1630–96)

Born in France, became an 'astronomer' of the Académie Royal des Sciences. Carried out missions for the Académie, measuring the heights of the tides at La Rochelle at the equinoxes, and on a trip to Canada tested two clocks made by HUYGENS (they stopped, to Huygens' chagrin). In 1671 he went to Cayenne, French Guyana, as part of a campaign to measure the parallax of Mars and to examine the periods of pendulums at different points on his travels so as to determine the Earth's gravity. NEWTON and Huygens used the data to determine the shape of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rigel

The star β Orionis. At apparent magnitude 0.18 it is the seventh brightest star in the sky, and though designated β is actually the brightest star in Orion. Betelgeuse (α Orionis), now an irregular variable, may have been the brighter star when the Bayer letters were allocated in 1603.

Rigel ('Left Leg of the Giant') is a blue-white supergiant of spectral type B8Ia. With an absolute magnitude of -6.7 it is the most intrinsically luminous star known. It is an α Cygni-type pulsating variable, with a period of 2.07 days, magnitude range approximately 0.0–0.3. It has a faint companion which is itself a spectroscopic binary, separated from the primary by $9.5''$; the orbital period is not measurable. The system is 770 light-years distant, with a parallax of $0.004''$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Right Ascension (RA)

The angle between the hour circles passing through the vernal equinox and a celestial body, measured eastwards (i.e. anti-clockwise) from the vernal equinox and expressed in time units (hours, minutes, and seconds) where 24 h is equivalent to 360°. In other words, it is the angle between the vernal equinox and a point on the celestial equator such that the angle between the vernal equinox, this point and a star is a right angle. The position of a star is normally expressed in terms of right ascension and declination.

See also: great circle, declination, hour angle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rille

A long narrow depression on the Moon's surface. There are several types. *Linear rilles* (or *straight rilles*), *forked rilles* and *arcuate rilles* are mostly graben—flat-bottomed strips of land that have collapsed between a pair of parallel fault lines. As their names imply, they may be straight or arc-shaped, and are often associated with large impact features. Examples are those named Rimae Mersenius and Rimae Hippalus, on opposite sides of Mare Humorum (rima is the term used in naming this type of feature). Different in origin are *sinuous rilles*. These are long, steep-sided channels, sometimes discontinuous, that wind their way across the floors of the lunar maria. They are in fact collapsed lava tubes, similar to those found on Earth. The surface of a narrow lava flow can solidify while lava beneath is still flowing, so that when the flow ceases an empty tube is left. A sinuous rille is produced when all or part of the roof of the tube collapses. An example of a lunar sinuous rille is the Hadley Rille, visited by the crew of Apollo 15. Sinuous rilles have also been identified on Venus and Mars. Some features on Uranus's satellite Ariel strongly resemble sinuous rilles, and may have a similar origin in a 'cryovolcanic' process in which melted ices can behave as lava.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ring Nebula (M57, NGC 6720)

A planetary nebula in the constellation Lyra, position RA 18^{h} 53.6^{m}, dec. $+33^{\circ}$ $02'$. It is ninth magnitude and measures $70''$ by $150''$; the central star is only fifteenth magnitude. Its telescopic appearance resembles a smoke-ring.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ritchey, George Willis (1864–1945)

At first a furniture-maker and woodworker, became an instrument-maker and especially an optician when he obtained part-time work at the observatory of the University of Cincinnati. Met GEORGE ELLERY HALE in Chicago and volunteered to assist him, preparing photographic plates, learning to use the camera to photograph stars and nebulae. Became a full time optician and supervisor of the instrument shop at Yerkes Observatory and made its 40 in refracting telescope work properly. He built the 24 in *Ritchey reflecting telescope* with which he discovered the expanding nebulosity around Nova Persei 1901. He also built the horizontal solar telescope, and the Snow horizontal telescope that replaced it. Ritchey followed Hale to the Mount Wilson Observatory, and made the 60 in and 100 in mirrors for the large Californian telescopes. He discovered a nova in a spiral galaxy, NGC 6946, showing the distances of galaxies as 'island universes'. Hale fell out with Ritchey and fired him. Ritchey went to live in France, where he implemented the design of the *Ritchey–Chrétien telescope*, a Cassegrain variant with paraboloidal concave primary mirror and ellipsoidal secondary mirror, and a curved focal plane (like the Anglo-Australian Telescope). He returned to the United States, and began to build a 40 inch Ritchey–Chrétien telescope for the US Naval Observatory in Washington, DC, but when he dropped the mirror he had been working on, the project lost confidence in him, and Ritchey retired.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rittenhouse, David (1732–96) and Rittenhouse, Benjamin (1740–1825)

David was an instrument-maker and astronomer, born in Paper Mill Run, PA, self-taught, as a young man he made clocks and telescopes, introduced the use of spider's webs as cross-hairs in transit telescopes and made orreries to show planetary motions. He used his own instruments for astronomy. He worked to survey the MASON and DIXON line and taught astronomy at the University of Pennsylvania. He became a statesman, and was director of the US Mint. Benjamin, his brother, born in Norriton, was a craftsman, who made decorated compasses and clocks.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Robinson, Thomas Romney (1792–1882)

Irish astronomer, was director of the Armagh Observatory for 59 years (possibly a world record for an observatory director). Concentrated on determining star positions with newly installed meridian and transit telescopes. For the observatory Robinson procured from THOMAS GRUBB, the instrument-maker and founder of the telescope building company that bore his name, a 15 in diameter reflector. This was the first large Cassegrain telescope built in Britain, and the first large reflecting telescope to be mounted equatorially with a clock drive. It had a novel, lever-support system to maintain the shape of the speculum primary mirror against flexure due to gravity as the mirror changed attitude. This design for the mirror cell was later used for LORD ROSSE'S six foot reflector and for the Great Southern Telescope at Melbourne, whose construction Robinson also oversaw, and for a century of later Grubb-Parsons telescopes. Robinson observed with the 'Leviathan of Parsonstown'. The duties of the observatory included meteorology and Robinson invented the cup-anemometer to measure wind speed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Roche Limit

The minimum distance from a planet at which a satellite can resist being gravitationally disrupted by tidal forces. It is given by $2.456R(\rho_p/\rho_s)^{1/3}$, where R is the planet's radius, and ρ_p and ρ_s are the densities of the planet and the satellite. This expression (without the density term) was derived in 1849 by Édouard Roche, and played a part in theories of the formation of Saturn's rings. Material within the Roche limit would be prevented from accreting into a larger body, while the break-up of a satellite that ventured inside the Roche limit would perhaps give rise to a ring system. The Roche limit as defined above applies to a fluid satellite—a hypothetical body of zero tensile strength—whereas real satellites are of course solid objects with a degree of structural integrity, able to offer some resistance to tidal disruption. Small satellites can in theory exist within the Roche limits for their bulk densities. However, it is not clear how many do, for the densities of the innermost satellites of the giant planets are mostly undetermined. Larger satellites, though, would be broken up within their Roche limit. If this happened, continued collisions between the resulting fragments, which would occupy similar orbits, would gradually produce progressively smaller and more numerous objects, and would thus create a ring system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Roche Lobe

Within the region surrounding the components of a binary system there are two competing gravity fields. The Roche lobe, named after the French mathematician Édouard Roche (1820–1883), is the volume of space within which any matter is controlled by the gravity field of a single star. It is said to be bound by an 'equipotential surface' in which the gravitational force exerted by the star is equal at all points.

The Roche lobes of each component of a binary system, which in isolation would be spheroidal, are drawn out into cone-like extensions meeting at a point between the two stars known as the *Lagrangian point*, where the gravitational attractions of the two components are exactly equal. The position of this point along the line joining the centres of the two stars depends upon their relative masses. If one of the component stars expands to fill its Roche lobe, mass transfer will occur through the Lagrangian point towards the other star.

See also: binary system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Roche, Édouard Albert (1820–83)

French astronomer and mathematician, *fl.* Montpellier, who studied the attractive force of rotating fluid bodies and discovered the *Roche limit*, the closest that a satellite can approach its planet and survive the stresses introduced when the nearer side tries to orbit faster than the further side. Empirically this closest distance is about 2.5 times the radius of the planet. Saturn's rings lie inside the Roche limit of Saturn. *Roche lobes* are the surfaces that define zones in a double-star system such that a test particle placed inside one of the surfaces will tend to gravitate towards the star near its center. If the star grows bigger than the zone in which it lies, it will expel material from its surface, possibly towards its companion star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rockets in Astronomy

Prior to the launch of the first satellite, Sputnik 1, in 1957, the only way to place scientific instruments above the atmosphere was to launch them on board rockets. In the following four decades, suborbital sounding rockets have continued to play a major role in improving our knowledge of the Earth, geospace and the universe as a whole.

Sounding rockets take their name from the nautical term 'to sound', which means to take measurements. The rockets consist of one or more stages, each with its own solid-fuel rocket motor, and a payload. They follow a parabolic trajectory, with less than 30 min of flight time. Following the launch, as the rocket motor uses its fuel, it separates from the vehicle and falls back to Earth. The payload continues into space and, once the experiment is completed, it re-enters the atmosphere. A parachute is deployed to bring the payload gently back to Earth for retrieval, a procedure that results in significant savings because all or part of the experiment can be refurbished and flown again.

The idea of using rockets to probe the upper atmosphere dates back to the 1930s, when ROBERT GODDARD proposed sending up small payloads that would parachute back to the ground from heights of up to 80 km (50 miles). Unfortunately, Goddard's plans were dashed by a lack of government interest and a shortage of funding.

As a result, rocket technology only became available for scientific research after American armed forces captured German V-2 missile factories at the end of the Second World War. Dozens of missiles and spare parts, along with blueprints and many of the key rocket designers (including WERNHER VON BRAUN), were shipped back to the United States and used to found the nation's missile program.

Between 1946 and 1951, 69 V-2s were launched from the US Army's White Sands Missile Range in New Mexico, reaching altitudes up to 185 km (116 miles). As the supply of captured missiles dwindled, a number of new rocket designs were introduced. These included the Aerobee, which was developed under the direction of space scientist JAMES VAN ALLEN from the Johns Hopkins University's Applied Physics Laboratory. Van Allen's main interest in using such vehicles was to observe cosmic rays before they struck the upper atmosphere.

Meanwhile, German engineers were also helping the Russians to build and fly their own V-2 rockets. These and their successors were launched from Kapustin Yar between 1947 and 1987. The rocket flights for atmospheric sounding were supervised by Soviet Academician A Blagonravov. The most used vehicle during the 1950s was a modified V-2, known as the R-1, which flew in five different versions.

A major impetus for rocket development and research on both sides of the Iron Curtain during the 1950s was the announcement that the 18 month period from 1 July 1957 to 31 December 1958 would be declared as International Geophysical Year (IGY). This would coincide with an expected peak in sunspot activity as well as several eclipses.

Almost 300 suborbital rocket launches were undertaken by the United States during this period, with another 175 sent up by the Soviet Union. Other countries that participated in IGY sounding rocket programmes included Australia, Canada, France, Japan and Great Britain. These were mainly designed to investigate the properties of the upper atmosphere, particularly the ionosphere.

Despite their short duration, these and subsequent suborbital flights made a number of scientific breakthroughs. The key to their success was the rocket's capability to lift detectors above the atmosphere which absorbed much of the cosmic radiation.

Attempts to study the universe at ultraviolet wavelengths from balloons during the 1920s were largely unsuccessful, since the Earth's ozone layer absorbs most of the incoming ultraviolet radiation. The first detection of ultraviolet radiation from the Sun had to wait until a spectrometer was carried to high altitude by a captured V-2 rocket in 1946. This was followed during the late 1950s by the first ultraviolet observations of stars from rockets carrying instruments with pointing capability. The breakthrough was achieved by Don Morton and Lyman Spitzer of Princeton University. Despite the failure of the rocket's parachute, their instrument survived to provide ultraviolet spectra for two stars in the constellation of Scorpius.

Also in 1946, T R Burnight from the US Naval Research Laboratory (NRL) reported fogging on a piece of photographic film carried aloft by a V-2. He claimed that the effect was caused by solar x-rays, but the cause of the blackening was questioned by other scientists. The first undisputed discovery of x-rays from the Sun had to wait until another NRL experiment, provided by HERBERT FRIEDMAN, flew on a modified V-2 rocket from White Sands in 1949. The experiment used a number of geiger counters to detect the radiation and provide basic directional information.

Then, on 18 June 1962, an Air Force experiment to search for x-rays from the Moon turned up a surprise. Three proportional counters on an Aerobee suborbital rocket found an x-ray source beyond the solar system in the constellation of Scorpius. The researchers, led by Italian physicist RICCARDO GIACCONI, called the discovery SCORPIUS X-1. Further celestial x-ray sources were discovered over the next few years by rocket-launched experiments.

INFRARED ASTRONOMY also gained an early boost from instruments on American suborbital rockets, this time provided by the US Air Force Cambridge Research Laboratories (later renamed the Air Force Geophysics Laboratory). Project HISTAR was prompted by the USAF's desire not to confuse the heat signature of an incoming enemy missile with the infrared emission from harmless astronomical sources. Accordingly, the Air Force launched a 16.5 cm (6.5 in) telescope, cooled by liquid helium, on a series of nine rockets, seven from White Sands

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

between April and December 1972, and two more from Australia in 1974. During some 30 min of accumulated observing time, more than 2000 cosmic infrared sources were detected at four wavelengths between 4 and 27 μm.

Although the limited size and mass of payload, together with the limited duration for observations, have reduced the number of such flights carried out in recent years, sounding rockets still offer a number of benefits. First, they are low cost and offer a launch flexibility not available with larger launch vehicles. They also provide the only means of making *in situ* measurements at heights between the maximum altitudes for balloons (about 48 km or 30 miles) and the minimum altitude for satellites (160 km or 100 miles).

Sounding rocket experiments provide a variety of information. This includes studies of the chemical composition and physical processes taking place in the atmosphere; measurement of the natural radiation surrounding the Earth; and data on the Sun, stars, galaxies and many other astronomical phenomena.

In addition, sounding rockets provide a reasonably economical means of conducting engineering tests for instruments and devices used on satellites and other spacecraft prior to their use in more expensive activities. Their low cost and short mission lead time mean that they are valuable scientific research tools for undergraduate and graduate students.

Although some countries, such as the UK, have phased out sounding rocket programs altogether, these launch vehicles are still widely used in the United States, Europe and Japan. NASA currently uses 15 different sounding rockets and launches about 30 such rockets each year. The launchers come in a variety of sizes from the single-stage Super Arcas, which stands 3 m (7 ft) tall, to the 20 m (65 ft) tall, four-stage Black Brant XII. These rockets can carry payloads of various weights to altitudes from 48 km (30 miles) to more than 1287 km (800 miles).

NASA sounding rockets are launched routinely from Wallops Island in Virginia, White Sands Missile Range in New Mexico and Poker Flat Research Range in Alaska, as well as sites in Canada, Norway and Sweden. They can also be launched from temporary launch ranges. In the past, NASA launch programs have been conducted from Peru, Puerto Rico, Brazil, Greenland, Australia and an aircraft carrier in the Pacific Ocean.

The Japanese space agency, ISAS, has also been building and launching sounding rockets for more than 30 yr. Many of these have been dedicated to astrophysical observations, upper atmosphere exploration, space plasma physics and other for scientific research. The main rockets used are the small MT-135, which is used primarily for middle atmosphere research, e.g. ozone layer depletion, the medium-range S-310 and the powerful S-520 which is capable of launching 100 kg of payload above 300 km (190 miles). These are usually launched from Kagoshima Space Centre on Kyushu island. Launch campaigns abroad with these sounding rockets have been made in Antarctica and at Andoya, Norway, for the *in situ* measurement of auroras.

Andoya Rocket Range (ARR) in the Svalbard archipelago is the world's northernmost permanent launch facility for sounding rockets. It provides favorable conditions for studying the ionosphere, the dayside aurora and processes in the magnetospheric boundary layer. Since 1972, ARR has been supported by several EUROPEAN SPACE AGENCY (ESA) member states. In return, the range facilities are available to the ESA member states on a marginal cost basis. By the beginning of 2000, the range had conducted 698 rocket launches.

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rooke, Lawrence (1622–62)

Born in Deptford, near London, became professor of astronomy and of geometry at Gresham College, London. Observed the satellites of Jupiter and made telescopic observations of lunar eclipses to determine longitude. But, more practical than many enthusiasts for these methods, he realized that they were too difficult to observe at sea, and could be used only to establish the longitude of cities and harbors. He proposed instead that a good lunar theory might provide an astronomical means to determine longitude at sea.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ROSAT (Röntgen Satellite)

The German–US–UK x-ray observatory ROSAT was launched into a 550 km circular orbit in June 1990. It carried two imaging telescopes operating in the soft x-ray (0.1–2.4 keV) and extreme UV (EUV) (006–0.2 keV) ranges. The x-ray telescope consisted of four nested Wolter type-I mirrors. The manufacturing of the x-ray mirrors — the largest and most accurate ones at that time — pioneered the use of Zerodur (a ceramic with zero thermal expansion) and an extremely low surface roughness (0.25 nm rms). In the telescope's focus were two position sensitive proportional counters (PSPC, 20 arcsec) and one high resolution imager (HRI, 5 arcsec). The EUV telescope consisted of three nested Wolter–Schwarzschild mirrors and two channel plate detectors (3 arcmin). The satellite and the x-ray telescope was designed, built and operated by Germany. NASA provided the Delta launch and the high resolution imager while the UK built and operated the EUV telescope. The prime ROSAT data centers were and are at the Max Planck Institute for Extraterrestrial Physics in Garching (x-rays) and at the University of Leicester (EUV) with mirror sites at the Goddard Space Flight Center and other scientific institutions.

ROSAT performed the first all-sky surveys with imaging x-ray and EUV telescopes leading to the discovery of 125 000 x-ray and 479 EUV sources. In addition the diffuse galactic x-ray emission was mapped with unprecedented angular resolution ($\leq$ 1 arcmin). Most of the mission time was devoted to pointed observations at selected targets. In total 4580 PSPC and 4482 HRI fields were covered with observation times between ~ 2000 s and ~ 1 Ms. Seven hundred scientists ('Principal Investigators') from 24 countries were involved in these pointed observations. The total number of ROSAT-based publications is 4787, with 54.9% in refereed journals (August 2001). All ROSAT data are calibrated and archived and are available through

http://wave.xray.mpe.mpg.de/rosat/catalogues

A large number of discoveries were made using both the all-sky survey and the pointed observations of ROSAT:

- x-ray shadows cast by cool interstellar clouds on the soft x-ray background in the Galaxy and nearby galaxies;
- x-rays and EUV light from the sunlit side of the Moon;
- x-rays from brown dwarfs;
- x-rays from T Tauri stars located at large distance from their birthplaces;
- underluminosity of single white dwarfs in x-rays and EUV, due to metals in their atmospheres;
- blackbody-like emission from intermediate polars;
- supersoft x-ray sources — nuclear burning on accreting white dwarfs;
- x-ray emission from the photospheres of single neutron stars;
- x-ray emission from millisecond pulsars;
- x-ray pulsations of the enigmatic γ ray source Geminga and its identification as a neutron star;
- x-ray and EUV emission from comets (e.g. Hyakutake, Levi);
- soft x-ray emission from the shock wave of SN 1987A;
- x-ray emitting explosion fragments in the Vela Supernova remnant;
- strong and highly variable x-ray emission from narrow line Seyfert galaxies;
- 80% of the cosmic x-ray background resolved into discrete source, mostly active galactic nuclei (AGN);
- superluminous flares at the centers of normal galaxies indicating episodic accretion onto supermassive black holes;
- x-ray diagnostics of merging clusters of galaxies;
- interaction between AGN jets and the intracluster medium in clusters of galaxies;
- measurement of the large-scale structure up to scales of ~ 700 Mpc using large samples of x-ray detected clusters of galaxies.

ROSAT was switched off in February 1999 after 8½ years of successful operation.

ROSAT images can be found at

http://wave.xray.mpe.mpg.de/gallery/index

Joachim Trümper

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Rosetta

The International Rosetta Mission was approved in November 1993 by ESA's Science Programme Committee (SPC) as the Planetary Cornerstone Mission in ESA's long-term program Horizon 2000. The mission goal is a rendezvous with comet 46P/Wirtanen. A subsidiary aim is the study of two asteroids during close fly-bys en route to the comet.

Background

Our knowledge of small solar-system bodies, the COMETS and ASTEROIDS, has improved dramatically over the last 20 yr. The major milestones were undoubtedly the first fly-bys of COMET HALLEY by the GIOTTO, VEGA, SAKIGAKE and SUISEI probes in 1986, the Giotto Extended Mission to comet Grigg–Skjellerup in July 1992, the first near encounters with main-belt asteroids Gaspra and Ida by the GALILEO spacecraft on its way to Jupiter, the encounters with Braille by Deep Space 1 and with near-Earth object Mathilde by the NEAR spacecraft, and finally NEAR's rendezvous with EROS in February 2000. During the same period, telescopic observations performed on the ground or in Earth orbit have greatly expanded and diversified. They constitute the basis for understanding small bodies as a population, since we can now compare observations of a large variety of objects and can undertake investigations of the variability of cometary activity. In particular, the outstanding opportunities to observe two bright comets, Hyakutake and Hale–Bopp in 1995 and 1996, provided new insights into cometary composition from ground-based observations.

Systematic observations in the visible spectrum are being performed for short-period comets as well as main-belt and near-Earth asteroids. It is now possible to observe cometary nuclei even at large heliocentric distances from the ground and with the Hubble Space Telescope. Furthermore, small bodies have now been studied systematically at all wavelengths from the ultraviolet, via the infrared to microwave and radio wavelengths. Here the International Ultraviolet Explorer (IUE) and ESA's Infrared Space Observatory (ISO) played, among others, an important role.

From this wealth of new information, it is becoming apparent that asteroids and comets constitute an almost continuous suite of progressively less evolved objects, reflecting the radial gradient in the swarm of planetesimals during the formation of the solar system. Indeed, the outermost asteroids exhibit spectral similarities to the bare cometary nuclei observed far from the Sun. A better understanding of the relationship between asteroids, comets and PLANETESIMALS throughout the solar nebula is an essential step in unravelling the first stages of the formation of the solar system.

Cometary material has been submitted to the lowest level of processing since its condensation from the protosolar nebula. It is considered likely that pre-solar grains may have been preserved in comets. As such, cometary material should constitute a unique repository of information on the sources that contributed to the protosolar nebula, as well as on the condensation processes that resulted in the formation first of planetesimals and then of larger planetary bodies. While tantalizing results have been obtained *in situ* from cometary grains and from interplanetary dust particles collected on Earth, the latter cannot be considered as fully representative, in particular in terms of their organic and volatile complement.

Two solutions to the problem of obtaining unaltered material can be considered: returning to Earth a sample of a COMETARY NUCLEUS (the original Rosetta concept of a comet-nucleus sample-return mission) or staying close to the comet and performing comprehensive *in situ* analyses of material from the surface and the coma.

The first approach had the undisputed advantage of bringing the full range of analytical techniques that are, or will be, available in the laboratory to bear on investigations of the sample material. Limiting thermal and mechanical stresses to acceptable levels during cruise and recovery, and even defining these acceptable levels, represented a significant scientific and technological challenge. The approach of a rendezvous mission with a suite of high-resolution remote sensing instruments and very sensitive *in situ* analytical investigations results in a less complex and cheaper mission and guarantees by design minimal perturbations of the cometary material, as analyses are performed *in situ*, at low temperatures and in a microgravity environment. It also provides the opportunity of observing at close range the onset and development of cometary activity, which results in the spectacular displays that have captured the imagination of humanity over the centuries.

Direct evidence on cometary volatiles is particularly difficult to obtain, as species observable from Earth and even during the Halley fly-bys result from physicochemical processes such as sublimation, interaction with solar radiation and the solar wind. The currently available information on cometary material gained from *in situ* studies and ground-based observations demonstrates the low level of evolution of cometary material. The latter's tremendous potential for providing information on the constituents and early evolution of the solar nebula has yet to be exploited.

Studying cometary material represents a major challenge, owing to the very characteristics that make it a unique repository of information about the formation of the solar system, namely its high content of volatiles and organic material.

A fundamental question that has to be addressed by the mission is, to what extent can the material accessible to analyses be considered as representative of the bulk material constituting the comet and of the early nebular condensates that constituted the cometesimal 4.57×10^9 yr ago? This representativity issue has to be addressed by first determining the global characteristics of the nucleus (mass, density, state of rotation), which can provide clues

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

concerning vertical gradients, and hence the relationship between the outer layers and underlying material.

The dust and gas activity observed around comets, as well as its rapid response to insolation, guarantees the presence of volatiles at or very close to the surface in active areas. Analyzing material from these areas will therefore provide information on both the volatile and the refractory constituents of a cometary nucleus. The selection of a proper site for surface science investigations should be relatively straightforward, given the extensive remote sensing observation phase and the advanced instrumentation, which covers a broad range of wavelengths, being provided on the Rosetta orbiter. The surface science sites can be monitored during surface activities as well as during a large fraction of the activity cycle, which should bring to light clues concerning the compositional heterogeneity of active regions.

The dust-emission processes are induced by very low-density gas outflows and should preserve the fragile texture of cometary grains. These grains can be collected at low velocities (a few tens of meters per second) by the spacecraft after short travel times (of the order of minutes), which will minimize alterations induced by the interaction with solar radiation. Similarly, gas analyzed in jets or very close to the surface should yield information on the volatile content of cometary material in each source region.

The Rosetta mission will study the nucleus of comet 46P/Wirtanen and its environment in great detail for a period of nearly two years, the near-nucleus phase starting at a heliocentric distance of about 3.25 AU, from the onset of activity through to perihelion, close to 1 AU. On its long journey to the comet, the spacecraft will pass close to two asteroids, 4979 Otawara and 140 Siwa.

Some questions to be answered

Just as the Rosetta Stone was the key that unlocked the meaning of Egyptian hieroglyphs, so the Rosetta spacecraft (figure 1) is intended to decipher the meaning of comets and their role in the origin and history of the solar system. Here are a few of the main puzzles.

- *What is the bulk density of a comet?* Guesses about the density of cometary material vary widely, and only an orbiting spacecraft can give exact measurements of the comet's volume and mass.
- *Is a comet a dirty snowball or an icy dirtball?* In other words, is it made of ices contaminated with mineral and tarry dust, or is it a consolidation of dust coated with ices?
- *Why is the nucleus of a comet so dark?* Giotto established that Halley's nucleus is like brownish-black velvet, absorbing 96% of the sunlight falling on it. Is the color due to a surface deposit of tarry dust, or is the interior dark too?
- *Why are small regions of a comet highly active when most of its surface is not?* Multiple jets of dust seen emanating from Halley's comet, and spectacularly from comet Hale–Bopp, imply that certain hot-spots differ physically or chemically from the rest of the comet's surface.

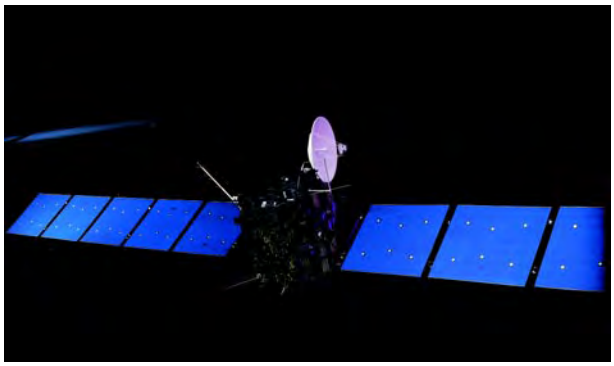


Figure 1. The Rosetta spacecraft.

- *Is a comet made as single piece, or does it consist of loosely joined blocks, as suggested by the Giotto images?* This relates to the questions of how comets are built, and why they break up into smaller fragments, as seen spectacularly with comet Shoemaker–Levy 9 which hit Jupiter in 1994.
- *Does a dying comet evaporate and disappear, or does it simply exhaust the stocks of ice that drive the emissions of gas and dust from an active comet?* If the latter answer is correct, dead comets persist long afterwards as dark, inactive masses of minerals and tar, and pose a lasting threat of collisions with the Earth.
- *What is a comet's exact composition?* Many ingredients are known, and the approximate abundances of the main constituents. Details coming from Rosetta will pin down (1) how comets were fashioned from similar constituents of interstellar dust and (2) how comets contributed to building the planets, including the Earth, and stocking their atmospheres.
- *Is the tarry, carbon-rich material in comets a jumble of every kind of chemical that inorganic processes can make from carbon, nitrogen, oxygen and hydrogen, or does it contain special compounds?* This is relevant to assessing the role of comets in the origin of life on the Earth.

Scientific objectives

The prime scientific objective of the mission as defined by the Rosetta Science Study Team is to study the origin of comets, the relationship between cometary and interstellar material and its implications with regard to the origin of the solar system. The measurements to be made in support of this objective are as follows:

- the global characterization of the nucleus, determination of dynamic properties, surface morphology and composition;
- the determination of the chemical, mineralogical and isotopic compositions of volatiles and refractories in a cometary nucleus;
- the determination of the physical properties and interrelation of volatiles and refractories in a cometary nucleus;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- the study of the development of cometary activity and the processes in the surface layer of the nucleus and the inner coma (dust–gas interaction);
- the global characterization of asteroids, including determination of dynamic properties, surface morphology and composition.

Comet 46P/Wirtanen

Comet 46P/Wirtanen ($P = 5.5$ yr), a Jupiter family comet, has been chosen as the prime target of the International Rosetta Mission. Jupiter family comets not only present a large group of the short-period comets in the solar system but also in principle are observable along their entire orbit. Thus, they are excellently suited to serve as targets for a complete description of the evolution of a comet along its orbit. Comet 46P/Wirtanen was discovered by Carl A Wirtanen at Lick Observatory on 15 January 1948 and described as a diffuse 17 mag object with a central condensation. Two close approaches to Jupiter in 1972 (0.28 AU) and 1984 (0.47 AU) changed the orbit from an initial perihelion distance of 1.63 AU and a period $P = 6.71$ yr to a perihelion distance of 1.06 AU and a period $P = 5.46$ yr.

Except in 1980, the comet was observed during all its apparitions. After the reduction of its perihelion distance, 46P/Wirtanen had come into reach of the visual observers and a first light curve of the observations in 1986 and 1991 was compiled. The coordinated observational campaign during its latest apparition in 1996–7 made this object one of the best-monitored comets. Some basic parameters of the nucleus could be derived and the evolution of its activity along the same part of the orbit that will be covered by Rosetta was closely monitored.

Assuming a geometric albedo of 0.04 a radius of about 600 m was derived with an upper limit of 700 m. Its gas production rate at perihelion is about 10^{28} molecules s^{-1}, about a factor of 100 less than 1P/Halley, which, however, is fairly active for such a small nucleus.

Scientific payload

In February 1996, ESA's SPC endorsed the Rosetta Orbiter Payload, and originally two Surface Science Packages—Champollion to be provided by NASA–JPL/CNES and Roland to be provided by a European consortium led by MPI and DLR from Germany—and five Interdisciplinary Scientists. Programmatic difficulties led to the withdrawal of NASA from Champollion in September 1996. CNES has joined the original Roland consortium to provide a single European Lander.

The Orbiter Payload has unprecedented capabilities to study the composition of both the volatile and refractory material released by the cometary nucleus with very high resolution. The remote sensing suite of instruments (see table 1) will allow for a characterization of the nucleus surface in a wide wavelength range from UV to mm with high resolution, e.g. the narrow-angle camera of OSIRIS will achieve from close orbits a resolution on the nucleus surface of better than 4 cm.

To complement these instruments and to provide for a proper monitoring of the comet environment and its interaction with the solar wind, a dust flux analyzer and a plasma instrument package have been selected.

Rosetta Lander payload

The Lander science (see table 2) will focus on the *in situ* study of the composition and structure of the nucleus material. Measurement goals include the determination of the elemental, molecular, mineralogical and isotopic composition of the cometary surface and subsurface material. Highest priority is given to the elemental and molecular determinations, as it is believed that some mineralogical and isotopic measurements can be carried out adequately by Orbiter science investigations. In addition properties such as near-surface strength, density, texture, porosity, ice phases and thermal properties will be derived. Texture characterization will include microscopic studies of individual grains.

CONSERT, with hardware on both the Lander and the Orbiter, will try to reveal the coarse structure of the nucleus through radio sounding.

Interdisciplinary Scientists

Five Interdisciplinary Scientists were nominated in support of the mission:

- M Fulchignoni, DESPA, Observatoire de Paris, France;
- M Fulle, Trieste Astronomical Observatory, Italy;
- E Grün, MPI für Kernphysik, Heidelberg, Germany;
- R Schulz, MPAe, Lindau, Germany, now ESA Space Science Department;
- P Weissman, NASA–JPL, Pasadena, CA, USA.

Mission overview

Rosetta will by launched in January 2003 by an Ariane 5 from Kourou. To gain enough orbital energy to reach its target, one Mars and two Earth gravity assists will be required (see table 3). The long mission duration required the introduction of extended hibernation periods. The mission can be subdivided into several distinct phases.

Table 3. Major mission events.

Event	Nominal date
Launch	15 January 2003
Mars gravity assist	26 August 2005
Earth gravity assist 1	26 November 2005
Otawara fly-by	11 July 2006
Earth gravity assist 2	26 November 2007
Siwa fly-by	25 July 2008
Rendezvous maneuver	24 August 2011
Start near nucleus operations at 3.25 AU (from Sun)	22 August 2012
Perihelion passage (end of mission)	10 July 2013

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Individual instruments on the Rosetta Orbiter.

Short name	Objective	Principal investigator
Remote sensing		
OSIRIS	Multicolor imaging (narrow- and wide-angle camera)	H U Keller, MPAe, Katlenburg-Lindau, Germany
ALICE	UV spectroscopy (70 nm–205 nm)	A Stern, SWRI, Boulder, CO, USA
VIRTIS	Visible and IR mapping spectroscopy (0.25 μm–5 μm)	A Coradini, IAS–CNR, Rome, Italy
MIRO	Microwave spectroscopy (1.3 mm and 0.5 mm)	S Gulkis, NASA–JPL, Pasadena, CA, USA
Composition analysis		
ROSINA	Neutral gas and ion mass spectroscopy; double focusing, 12–200 AMU, $M/\Delta M \sim 3000$ Time of flight, 12–350 AMU, $M/\Delta M \sim 1500$ including neutral dynamics monitor	H Balsiger, University of Bern, Switzerland
COSIMA	Dust mass spectrometer (SIMS, $M/\Delta M \sim 2000$)	J Kissel, MPE Garching, Germany
MIDAS	Grain morphology (atomic force microscope, nm resolution)	W Riedler, Institute for Space Research, Graz, Austria
Nucleus large-scale structure		
CONSERT	Radio sounding, nucleus tomography	W Kofman, CEPHAH, Grenoble, France
Dust flux, dust mass distribution		
GIADA	Dust velocity and impact momentum measurement, contamination monitor	L Colangeli, Osservatorio Di Capodimonte, Naples, Italy
Comet plasma environment, solar wind interaction		
RPC	Langmuir probe	R Boström, Swedish Institute of Space Physics, Uppsala, Sweden
	Ion and electron sensor	J Burch, SRI, San Antonio, TX, USA
	Flux gate magnetometer	K-H Glassmeier, TU Braunschweig, Germany
	Ion composition analyzer	R Lundin, Swedish Institute of Space Physics, Kiruna, Sweden
	Mutual impedance probe	J G Trotignon, LPCE/CNRS, Orleans, France
RSI	Radio science experiment	M Pätzold, University of Cologne, Germany

Table 2. The Lander payload.

APXS	α–p x-ray spectrometer	R Rieder, MPI Chemistry, Mainz, Germany
SD^2	Sampling and drilling device	A E Finzi, Politechnico Milano, Italy
COSAC	Evolved gas analyzer, elemental, molecular composition	H Rosenbauer, MPAe, Katlenburg-Lindau, Germany
MODULUS	Evolved gas analyzer, isotopic composition	I Wright, Open University, UK
ÇIVA	Panoramic camera, IR microscope	J P Bibring, IAS, Orsay, France
ROLIS	Descent camera	S Mottola, DLR Berlin, Germany
SESAME	Surface electrical and acoustic monitoring, dust impact monitor	D Möhlmann, DLR Cologne, Germany H Laakso, FMI, Finland I Apathy, KFKI, Hungary
MUPUS	Multipurpose sensor for surface and subsurface science	T Spohn, University of Münster, Germany
ROMAP	Roland magnetometer and plasma monitor	U Auster, TU Braunschweig, Germany I Apathy, KFKI, Hungary
CONSERT	Comet nucleus sounding	W Kofman, CEPHAG, Grenoble, France

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The science mission phases

Otawara fly-by (July 2006)
Fly-by operations last from 3 months before to 1 month after the fly-by. Parallel to the daily tracking with orbit determination and corrections, the scientific payload is checked out. The aim is to pass the asteroid at a distance of about 1000 km on the Sunward side. The relative asteroid ephemeris will be determined with a cross track accuracy of 20 km by spacecraft optical navigation. The cameras and scientific payload will be pointing in the direction of the asteroid until after the fly-by.

Siwa fly-by (July 2008)
The operations are similar to the Otawara fly-by; however, the fly-by distance will be of the order of 2500 km to be commensurate with the spacecraft angular rate capabilities which are determined by the requirements of the Otawara fly-by, which occurs at a lower relative velocity.

Global mapping and close observation phase
The preliminary survey of the surface of the comet is known as global mapping. At least 80% of the surface subject to solar illumination is required to be mapped.

Even in the gravity field of a nucleus of very irregular shape, it will be possible to choose quite stable orbits that do not impact on the cometary nucleus. Polar orbits around the comet at a distance between 5 and 25 nucleus radii will be used for mapping the surface. The orbital period will be usually greater than the spin period; horizontal swaths will cover the nucleus surface as it is presented. The semimajor axis of the mapping orbit will be chosen as a function of comet gravity and spin rate, taking into account the following constraints:

- coverage without gaps;
- safety considerations (no impact on nucleus);
- volume of data for real-time transmission;
- maximum time to complete surface mapping;
- minimum resolution and viewing angle to surface normal;
- continuous communications with Earth.

In case of slowly rotating comets, it may be necessary to use more than one revolution in the same orbit or a second different orbit by means of plane change maneuvers.

Ideally all the Orbiter payload instruments are operating during this phase. The scientific data gathered by the whole payload is buffered in the on-board memory, before transmission to the ground station.

The nucleus shape and surface properties as well as kinematics and gravitational models will be derived using optical landmark observations. Based on the mapping and remote observation data, five areas will be selected for close observation.

At the end of the close observation phase, based on the collected data, a decision will be taken to which site the Surface Science Package (SSP) will be delivered.

Surface Science Package delivery phase
Surface Science Package delivery will be from an eccentric orbit (pericenter altitude as low as possible, e.g. 1 km) with a pericenter passage near the desired landing site. The time and direction of the SSP separation will be chosen such that the package arrives with minimum vertical and horizontal velocities relative to the local (rotating) surface. An ejection mechanism will separate the SSP from the spacecraft with a maximum relative velocity of up to 1.5 m s^{-1}.

After delivery of the SSP, the spacecraft will be injected into an orbit that is best suited to receive the data transmitted from the SSP and to relay them to the Earth.

Extended monitoring phase (through perihelion)
After the end of the activities related to the SSP, the spacecraft will spend at least 200 days in orbit in the vicinity of the comet until perihelion passage. The objective of this phase is to monitor the nucleus (active regions), dust and gas jets and to analyze gas, dust and plasma in the inner coma from the onset to peak activity. Orbits between 1 and 10 cometary radii are envisaged.

Orbital design will depend on safety considerations and scientific goals taking into account that the communication relay service has to be provided for the long-life SSP. Mission planning will depend on the results of previous observations, such as the activity pattern of the comet. Extended monitoring of different regions in the vicinity of the nucleus could be performed by successive hyperbolic fly-bys configuring petal-like trajectories.

Spacecraft overview and payload accommodation

The spacecraft design is driven by key features of this very complex and challenging mission:

- the fixed and fairly short launch window for comet 46P/Wirtanen—because of the launch capability required, even for Ariane 5 there exist only very few backup scenarios within a reasonable time frame;
- the long mission duration of 10.5 yr;
- the critical gravity assists at Mars and the Earth and the close asteroid fly-bys, where the spacecraft and the payload will be active;
- the wide variation of spacecraft–Earth–Sun cycles and distances, which poses challenges for the thermal design and require large solar arrays (approximately 64 m^2) with novel cells for low-intensity low-temperature operation;
- the long operations in close proximity to the comet surface—only a few comet nucleus radii away in an environment (dust and gas emitted from the nucleus) that will not be known in great detail during development.

The long round-trip times for light of up to 90 min require an autonomous spacecraft that in addition will be able to survive a long hibernation phase without any ground contact during the cruise.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gerhard Schwehm

The payload is mounted on one spacecraft wall, which during the close comet phase will continuously be pointed towards the nucleus. All instruments are body mounted and the proper attitude is achieved by rotating the spacecraft with the high-gain antenna always pointing towards the Earth and the solar array pointing towards the Sun.

Résumé

A sophisticated payload such as the one to be flown on the Rosetta mission and especially the goal to deploy a Lander onto the nucleus surface for *in situ* investigations requires laboratory simulations of nucleus properties of which we have very little—and only indirect—knowledge.

To land the Lander and to anchor it into the nucleus surface requires basic knowledge of the material, its tensile strength, density and structure. The evidence for all of these parameters is still extremely tenuous. The presence of small-scale and possibly larger-scale active areas, such as the craters seen on Halley Multicolour Camera images of comet 1P/Halley, clearly indicates that there are differences in the near-surface properties. These in turn must be related to physical or chemical heterogeneities within the nucleus. However, from ground-based observations it is not possible to determine whether variations in activity stem from small-scale heterogeneity (such as variations in the thickness of the mantle) or from larger-scale variations of the volatile content of underlying porosity.

One of the main scientific objectives of the mission is the study of the physicochemical processes that are specific to comets. The increase in cometary activity as these bodies approach the Sun undoubtedly represents one of the most complex and fascinating processes to be observed in the solar system.

A full understanding of cometary activity involves inputs from a wide range of disciplines, including thermodynamics, phase equilibrium, gas–solid interactions, fluid dynamics, heterogeneous chemistry, photochemistry and plasma physics.

The mission will provide information on 'how a comet works' and will be crucial to our understanding of the early solar nebula and the evolution of the planetary system. The study of the chemical (isotopic) composition of the ices and the dust grains will provide us with information on the processes through which these elements formed and about prebiotic chemistry. It will answer the question what role comets played in the evolution of life on Earth.

Rosetta is a fascinating mission that will address many key questions about our origins: if we want to understand the origins of the solar system, we have to study comets.

Bibliography

For up-to-date information on the mission and its status, details of the payload, information on the science background and general information on comets, please visit

http://sci.esa.int/home/rosetta/index.cfm

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rosette Nebula (NGC 2237–39, 2246)

A complex emission nebula in the constellation Monoceros, position RA $06^{h}\ 30.3^{m}$, dec. $+05^{\circ}\ 03'$. It surrounds the star cluster NGC 2244, the stars of which energize its gas. Different parts of the nebula received separate NGC designations; the whole resembles a rosette or wreath.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rosse, 3rd Earl of [William Parsons, Lord Rosse, Lord Oxmantown] (1800–67) and Rosse, 4th Earl of [Laurence Parsons, Lord Rosse, Lord Oxmantown] (1840–1908)

Irish astronomer and landowner, the 3rd Lord Rosse was educated at Trinity College, Dublin and Oxford as a mathematician. He became interested in astronomy and made at the family castle in Birr a 36 in reflector with the same design as William Herschel's (see HERSCEL FAMILY). Mapped the Moon, and observed nebulae with the intent to resolve them into stars. He developed the technology at Birr Castle to make speculum mirrors, using the estate's blacksmiths, labourers and materials (such as peat for the fuel for the furnaces). In 1842, he successfully cast a 72 in mirror for the 'Leviathan of Parsonstown', which stood as the largest telescope in the world at the time (1845–1917), slung between two massive walls aligned on the meridian. The telescope, restored to working order in 1998, was raised in elevation, and tracked for a small angle across the meridian, by a system of ropes and pulleys. The observer had access to the eyepiece from chairs mounted on the walls and was able to observe an object for at most an hour and a half. The potato famine in Ireland delayed the astronomical work of the telescope until 1848, from when it was used to view planets, satellites and nebulae. In over a dozen nebulae Lord Rosse, his son, his astronomical assistants and visiting observers (like J L E DREYER and THOMAS ROMNEY ROBINSON) were able to resolve spiral arms, indicating that they were more than just collections of gas—galaxies in fact. They went so far as conclude that 'no real nebulae seem to exist . . . all appeared to be clusters of stars'. Lord Rosse also realized that some elliptical and lenticular nebulae were edge-on spirals. Lord Rosse's drawings of the nebulae were strikingly accurate, although there was a curious lapse in the telescope's first record of M1, where a fanciful drawing with many radial 'legs' inspired the enduring name Crab Nebula. The 3rd Earl Rosse's son, Laurence, the 4th Earl Rosse, fitted both the 72 in and the 36 in telescopes with spectroscopes. His observations proved that some nebulae had bright-line spectra, including all the planetary nebulae, and others, including the Andromeda galaxy, were made up of stars. He also measured the temperature of the Moon by focusing its light onto a thermocouple, watching the temperature drop during an eclipse.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rossi X-ray Timing Explorer (RXTE/Explorer 69)

NASA mission to study variations in x-ray emissions on time scales from microseconds to years. Named in honour of x-ray astronomy pioneer Bruno B Rossi, former Professor of Physics at Massachusetts Institute of Technology. Launched December 1995. Carries a Proportional Counter Array and High-Energy X-ray Timing Experiment to study compact objects in the energy range 2–200 KeV. Also has an All Sky Monitor which scans over 70% of the sky each orbit. Key feature is flexible operations through rapid pointing, high data rates and nearly continuous receipt of data. Discovered millisecond x-ray pulsars and high-frequency x-ray pulses from neutron stars in binary systems. (See EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rossi, Bruno Benedetto (1905–93)

Cosmic ray physicist, born in Venice, Rossi studied and worked in Italy until the Fascist regime dismissed him. He held positions in Europe and joined the Los Alamos Laboratory in New Mexico, to develop the atomic bomb. He became professor at MIT. He devised the *Rossi coincidence circuit*, which recorded the simultaneous occurrence of electric pulses in cosmic ray and other detectors to distinguish real events from noise. He investigated the nature of and energies of cosmic rays, and placed detectors in space for cosmic ray investigations, discovering with the Explorer X satellite in 1961 the magnetopause, the boundary containing the Earth's magnetic field. With RICCARDO GIACCONI he launched a rocket to detect cosmic x-rays (from the Moon), which discovered by chance, behind the Moon, the Scorpius X-1 x-ray source, the first discovered. An x-ray astronomy satellite, the Rossi X-ray Timing Explorer (RXTE), is named after him.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rotation

The motion of a body about an axis; for example, the Earth rotates on its axis in a period of 23 h 56 min. All celestial bodies, such as planets, stars and galaxies, exhibit some degree of rotation.

See also: axis, orbit, revolution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Royal Astronomical Society

The Royal Astronomical Society (RAS) represents professional astronomers and geophysicists in the UK. It was founded in 1820 as the Astronomical Society of London by John Herschel and 13 other well-known astronomers and scientists. The Society received the grant of a Royal Charter from King William IV in 1831. Its headquarters are in Piccadilly, London.

The Society's aims are 'the encouragement and promotion of astronomy and geophysics'. Its main functions are to publish the results of astronomical and geophysical research, to maintain as complete a library as possible in these subjects and to hold meetings, in London and elsewhere, at which astronomical and geophysical matters can be discussed.

The RAS is affiliated to the European Astronomical Society and to the European Geophysical Society. With the Geological Society, it has formed a Joint Association for Geophysics to promote geophysics.

The RAS library contains material for research in astronomy and geophysics as well as the histories of these sciences and of associated fields such as navigation. Collections of the Society's instruments are loaned to museums, universities and observatories throughout the UK.

The Society publishes research papers in two academic journals: *Monthly Notices of the Royal Astronomical Society* for astronomers and *Geophysical Journal International* for geophysicists. The Society's house publication, *Astronomy and Geophysics* carries reports of scientific meetings, papers on current topics, correspondence and book reviews.

Regular scientific discussion meetings are organized in London, while annual national astronomy meetings are held at different venues. The Society has a number of interactions with the Particle Physics and Astronomy Research Council (PPARC) and the Natural Environment Research Council (NERC). The Society awards the Michael Penston Astronomy Prize (formerly the RAS Astronomy Prize) and the RAS Blackwell Prize, and awards a number of medals for contributions to astronomy and geophysics.

For further information see
http://www.ras.org.uk/ras/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Royal Astronomical Society of Canada

The Royal Astronomical Society of Canada, originally known as the Toronto Astronomical Club, was founded in 1868 by eight charter members. On 25 February 1890, in Toronto, it was incorporated with a constitution and bylaws. In 1903 it became known as the Royal Astronomical Society of Canada. Today, the Society has 23 affiliated centers from coast to coast, and a membership of close to 4000.

The Society produces several widely respected publications, including the *Journal of the Royal Astronomical Society of Canada*, the *Observer's Handbook*, the *Beginner's Observing Guide*, and the *Observer's Calendar*.

For further information see the publication *Looking Up* or the Society's website at http://www.rasc.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Royal Observatory, Edinburgh

The Royal Observatory, Edinburgh (ROE) comprises the UK Astronomy Technology Centre (ATC) of the PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL, and the University of Edinburgh's Institute for Astronomy.

The Institute for Astronomy is a research and teaching group within the university's Department of Physics and Astronomy. It is one of the UK's major centers of astronomical research, particularly involving cosmology, active galaxies and star formation.

The institute has a 0.5 m telescope and a 3.5 m spectrograph/solar telescope, infrared, reticon and CCD detectors, facilities for photography, photoelectric photometry, polarimetry and a CCD spectrograph.

The ATC is the UK's national center for the design and production of state-of-the-art astronomical technology. It includes an applied optics group which designs and builds instruments for both ground-based telescopes (UKIRT, GEMINI, ING) and space telescopes (FIRST, NGST). Other groups in the ATC deal with mechanics, electronics and software. Major technology research is undertaken into CCD detectors/cameras, infrared systems, submillimeter and cryogenics.

Particular research interests of the ATC astronomy group include star and planet formation, brown dwarfs and interacting galaxies. One of the main areas of study is the remnant dust disks around main sequence stars such as Vega.

The ROE manages UK telescope sites and data archive resources. It is the home of the Plate Library for the UK Schmidt telescope, which contains over 17 000 plates, and the SuperCOSMOS advanced photographic plate digitizing machine.

The ROE was founded in 1811 by a group of amateur and professional astronomers. Originally known as the Astronomical Institution of Edinburgh, it was renamed the Royal Observatory in 1822 after a visit to the city by King George IV. In 1834, the University of Edinburgh took over its administration on condition that the government would provide the salary for a professor who would hold the title of Astronomer Royal for Scotland. The link continued until 1995, when this honorary title was separated from the post of Director of the ROE.

For further information see
http://www.roe.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Royal Observatory, Greenwich

The Royal Observatory at Greenwich, London, founded in 1675, is the location of the Airy Transit Telescope that defines the prime meridian of the world and is the home of the Harrison Chronometers. The Observatory was founded by Charles II with the ultimate purpose of providing an accurate star catalog and model of the Moon's motion, that enabled mariners to find their longitude. During the twentieth century the observatory became increasingly involved in astrophysics and the research arm, known as the Royal Greenwich Observatory, moved from the Greenwich site in the 1950s and was eventually closed in 1998. The Greenwich site is now a museum, visited by half a million visitors per year, containing clocks, astronomical instruments, including a 70 cm refractor and a planetarium. It is administratively part of the nearby National Maritime Museum. The buildings, including Flamsteed House designed by Christopher Wren, have been restored. The Observatory contains a public information group that answers public and media enquiries about modern astronomy. It will be the management center of the 2 m Faulkes telescope for public education, located in Hawaii and used by students in Hawaii and the UK.

For further information see http://www.rog.nmm.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

RR Lyrae Stars

RR Lyrae stars are members of a class of VARIABLE STAR whose variations in light, radius and temperature are known to be caused by self-sustained radial pulsations of giant A stars with masses between 0.5 and 0.8 solar masses, mean radii of 5 times the radius of the Sun and absolute magnitudes at mean light of $\langle M_V \rangle \sim +0.5$. The amplitudes of their light curves range between 0.2 and 1.8 magnitudes in B. Their periods range between 0.25 and 1.2 days.

Although early called 'short-period Cepheids' because the shapes of their light, color and radial velocity curves are so similar to those of the long-period classical Cepheids, it was soon realized that the two groups were in fact separate in other of their characteristics. They differ in their spatial distribution in the Galaxy. The long-period classical Cepheids are concentrated to the Galactic disk while the field RR Lyrae stars have an almost spherical distribution in Galactic latitude. Furthermore, most of the RR Lyraes have high space velocities relative to the local standard of rest, whereas the long-period classical Cepheids have the low space velocities characteristic of extreme population I stars (see STELLAR POPULATIONS).

By 1950 it had become evident that the two groups of variables are of different population types and are in different evolutionary states. Using the RR Lyrae stars of the extreme population II, the extent of the Galactic halo has been mapped to at least 100 kpc from the Sun and the Galactic center, showing the enormous extent of the spherical component of the Galaxy.

In addition, these RR Lyrae variables in the halo and in the thick disk are metal poor. The chemical elements heavier than He, Li, Be and B are between 5 and 150 times less abundant than in the Sun, providing important clues to events in the chemical evolution of the Galaxy. A small subset of RR-Lyrae-type variables in the disk do, however, have near-solar metallicity and population I low-velocity kinematics, providing additional clues, not yet deciphered, for Galactic evolution.

RR Lyraes as a distinct class of variable stars were discovered in 1893 when large numbers of them were found in globular clusters by Solon Bailey. Because of their high abundance in many globular clusters, they were initially called 'cluster-type variables' whether they occurred in clusters or the field.

The brightest of the class in the general field, discovered by W Fleming at Harvard in July 1899 by the technique of multiply exposed photographic plates, is the eponymous star RR Lyrae itself. Soon other similar variables in the field were discovered, and the identification was soon made between the characteristics of Fleming's short-period field variables and Bailey's cluster variables.

There are now more than 7000 RR Lyraes known in the field and 1990 in Galactic GLOBULAR CLUSTERS. The class accounts for 20% of the total of nearly 30 000 variables in the Russian standard *Fourth General Catalogue of Variable Stars* (1985 edition). However, the discovery is far from complete. A current estimate is that the Galaxy contains of the order of 100 000 RR Lyraes in the halo, with a still higher number in the thick disk and the galactic bulge.

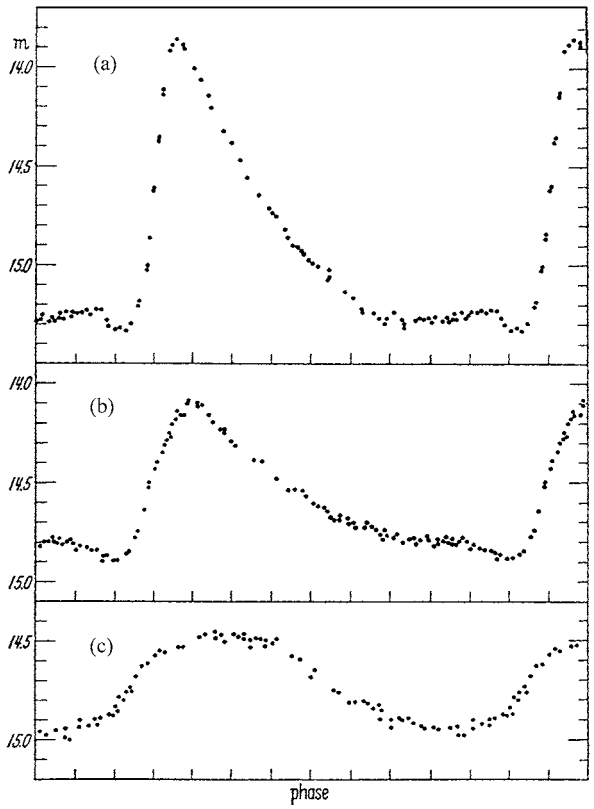


Figure 1. The light curve characteristics of the three Bailey types of RR Lyrae stars of classes a, b and c.

Shapes of the light curves

Bailey's early (1902) division into three classification groups, based on the shape of his light curves for the variables in the globular cluster omega Centauri, is still valid. We now know that this division into Bailey types, a, b and c is related to the physics of the internal forces that drive the self-sustaining pulsation according to the degree of energy dissipation that determines whether pulsation can occur.

Figure 1 shows typical light curves for the three Bailey RR Lyrae types. Type a variables (top panel) have a very steep rise to maximum light, usually taking only 10% of the period to go from minimum to maximum light. Moreover, the amplitude is largest in Bailey type a compared with the average amplitudes for types b and c. Type b variables have a more gradual rise to maximum, taking typically 20% of the period. The decline from maximum to minimum light is continuous rather than having the nearly flat light curve near minimum that is typical for Bailey type a variables.

The difference in the shapes of the light curves is undoubtedly due to a difference in the strength of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mechanism that drives the pulsation. Struve in 1947, Sanford in 1949 and Preston and Paczynski in 1964 discovered that type *a* variables all show the hydrogen lines in emission for about 10 min midway on the rising light curve from minimum to maximum, whereas none of the type *b* and *c* variables shows the emission. The emission is interpreted to be caused by a shock wave in the rising gas in the highly non-linear and stronger pulsation push of the *a* type compared with *b* and *c* type variables. The non-linearity of the driving is seen directly by the highly asymmetrical light curves of Bailey types *a* and *b*.

Type *c* variables have more symmetrical light curves that are also of smaller amplitude than either type *a* or *b*. They also are bluer than *ab* variables (see next section), occupying the left portion (hotter temperature) of the instability strip of the HR diagram, and they have shorter periods. Type *c* variables are vibrating in the first overtone of the pulsation, whereas type *a* and *b* variables are vibrating in the fundamental mode.

In recent years a type called RRd has been introduced, based on the discovery of stars that are driven by a combination of both the fundamental and first overtone periods. These occur in a very narrow range in period at the transition between the type *c* and type *ab* variables. These stars are called double-mode cluster variables in the literature.

Finally, concerning light curve shapes, it is to be noted that a few, perhaps 15%, of the RRab stars show systematic shape variations that themselves are periodic. Said differently, a given shape appears again and again as in the beat modulation in the addition of two periodic functions with slightly different periods. In RR Lyrae itself, the return of the light curve to a given shape has a period of 41 days. The phenomenon is called the Blazhko effect. Blazhko periods, as known up to 1995, range from 11 days for AH Cam to 533 days for RS Boo. Smith (1995) states that 'though a number of hypotheses have been advanced, there is no accepted explanation for the Blazhko cycles'.

Importance for theories of stellar evolution, galactic structure and cosmology

The importance of RR Lyrae variables for a large number of astronomical problems stems from their unique place in the HR diagram on the horizontal branch of globular cluster color–magnitude diagrams, showing that they have only a small spread in absolute luminosity as a group. Shapley used this fact to derive the first reliable distances to globular clusters using the cluster-type variables to calibrate the brightest 25 stars near the tip of the giant branch of the HR diagram. The calibration of the absolute magnitude of RR Lyrae stars has occupied astronomers from 1915 to the present.

The variables have been central in such diverse studies as (1) the structure and age of the Galaxy, in both the disk and the halo components, (2) the age of globular clusters, (3) study of the advanced, post-main-sequence evolution state of the variables in the HR diagram, (4) study of the astrophysical processes of stellar pulsation as a problem in stellar structure that combines the physics of classical mechanics, hydrodynamics and thermodynamics, (5) an understanding of the position of the instability strip in the HR diagram and (6) the calibration of extragalactic distances and the expansion rate of the universe.

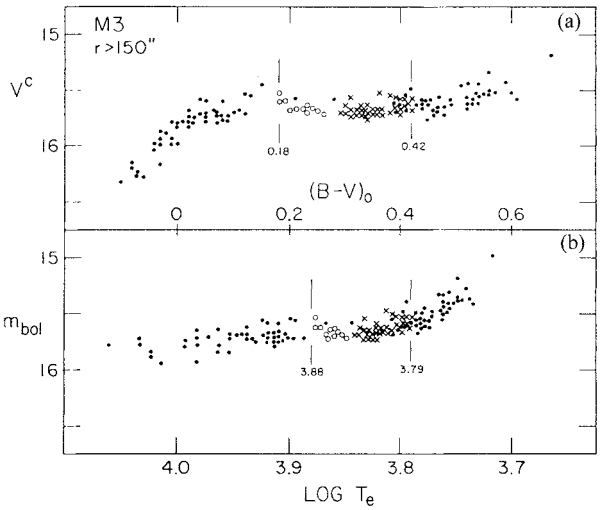


Figure 2. The observed position of the cluster variables in the globular cluster M3 along the horizontal branch. Circles in the instability strip are for Bailey type *c* variables. Crosses are for Bailey type *ab* variables. Type *c* variables are pulsating in the first overtone. Type *ab* variables pulsate in the fundamental mode.

Placing RR Lyrae variables in the HR diagram

In 1940 Martin Schwarzschild first showed that the RR Lyrae stars occur only within a small region the horizontal branch of the HR diagram of the globular cluster M3. The horizontal branch is populated by low-mass stars that have evolved from the main sequence to the tip of the giant branch, and after mass loss and the ignition of helium in the hydrogen-exhausted core have evolved from that tip at $M_V \sim -3$ to $M_V \sim +0.5$ mag The horizontal branch is a configuration of decreased radius and hotter surface temperature (see STELLAR EVOLUTION) and has only a small spread in absolute magnitude from cluster to cluster.

Schwarzschild's discovery of the cluster instability strip for the variables in M3 was soon extended as a general feature to all globular clusters whose horizontal branches covered the appropriate color range. His conclusion that 'stars which can pulsate do pulsate' was shown to be universal.

A modern mapping of the instability strip in the color–magnitude diagram of the globular cluster M3 is shown in figure 2. The top panel shows the photometric data for a complete sample of horizontal branch stars outside a radius of 2.5 arcmin from the cluster center, plotted at mean light.

The RR Lyrae domain occupies the color interval $0.18 < B{-}V < 0.42$. The Bailey type *c* variables (circles) are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

separated in color (being bluer) from type *a* and *b* variables (crosses). The bottom panel shows the conversion of the data to bolometric apparent magnitudes and logarithm of the effective temperature. There is a sharp separation color (and temperature) between type *c* and type *a*, *b* variables in M3. However, in other clusters there is an overlap region occupied by both types.

The vertical boundary lines in figure 2 define the blue and red edges of the pulsator instability strip at the luminosity of the RR Lyrae stars. It is known that the strip continues vertically in the diagram, sloping slightly to cooler temperatures at luminosities that are fully 6 magnitudes brighter over the whole of the long-period Cepheid instability strip.

Properties of the instability strip

The parameters of (a) amplitude, (b) shape of light curve and (c) period vary systematically with color within the instability strip. In a given cluster the longest-period *ab* variables are at the red side of the strip; the shortest period are the bluest. This is the period–color relation. Clearly such a relation is demanded by the pulsation model where the period must vary with the inverse square root of the density, which can be expressed as a combination of mass and radius, and therefore of temperature at a given mass and luminosity. A period–color relation also exists for the variables in the cluster M15, which, however, is shifted by a factor of 1.2 in period at a given color relative to the relation in M3. The periods in M15 are longer.

The largest-amplitude light curves occur near the center of the strip, decreasing toward the red edge of the instability region. Because there is a period–color relation, this amplitude–color relation can be changed to an amplitude–period relation. The amplitude correlation is also shifted in M15 by again a factor of 1.2 in period relative to the relation in M3. The same is true of the light curve shapes, as measured by the rise times (in units of the period) from minimum light to maximum.

These correlations with their period shifts between M3 and M15 are shown in figure 3. The top panel shows the period–amplitude relation. The relation between period and the light curve shape, measured by the rise time, is in the bottom panel.

Because the metallicity is different for M3 and M15 (M3 is down by a factor of 50 in metals from the Sun; M15 is down by a factor of 150) the period shifts were suspected early on (in 1958) to be a metallicity effect. From studies of many clusters of different metallicities, the period shifts were found to be strongly correlated with metallicity. The effect is a generalization of the Oosterhoff–Arp–Preston period–metallicity correlation that provides an important method for the RR Lyrae absolute magnitude calibration (see next section). The period–metallicity correlation is obeyed both in the globular clusters and in the general field.

Figure 2 shows that, within a given cluster, the absolute *V* magnitudes of the RR Lyrae stars are closely the same. There is a small vertical dispersion in magnitude

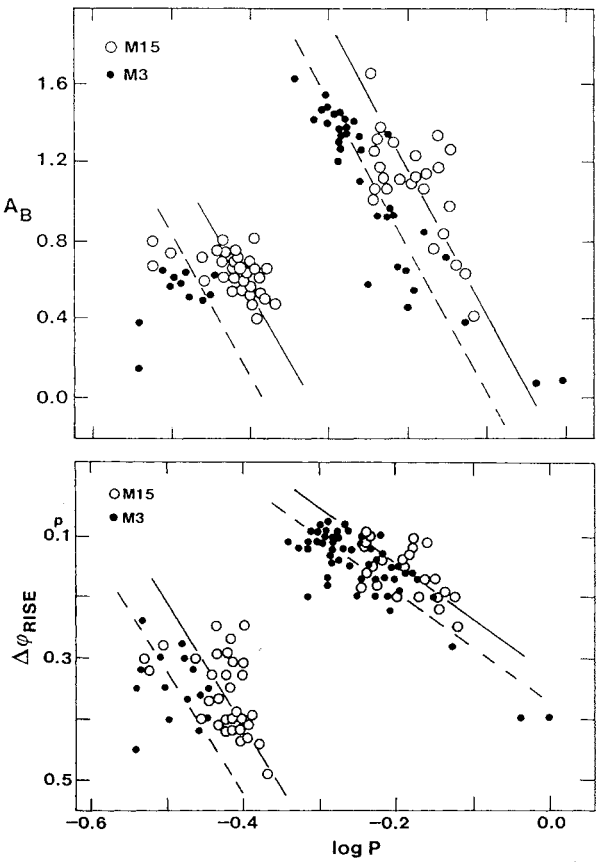


Figure 3. (top) The period shifts in the period–amplitude relation between the RR Lyrae stars in M3 and M15. The clusters have different metal abundances. (bottom) Same as the top but using the shape of the light curves as measured by the rise time (in phase units) from minimum to maximum light.

which has been shown to be real. It is caused by the evolution from the 'age zero' horizontal branch as the helium burning in the core proceeds. Proof that the dispersion in absolute magnitude is real, not due to observational error, is made by showing that the variables that are brighter have longer periods than the fainter cluster variables of the same color. This, of course, is required from the period–density relation. However, the dispersion is small, only at the 0.1 mag level. Most of the stars in a given cluster have a nearly constant *V* magnitude, showing that the RR Lyraes are excellent distance indicators once the metallicity dependence and the evolutionary effects can be calibrated out.

The expectation that there is a correlation of absolute magnitude with metallicity, required by the period shifts, is confirmed observationally. Figure 4 shows a composite color–magnitude diagram of seven globular clusters whose metallicities range from 2.5 times down from the Sun (NGC 6838) to 150 times smaller than the Sun (M15). All seven have horizontal branches, four of which thread the instability strip and have cluster variables, but at

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

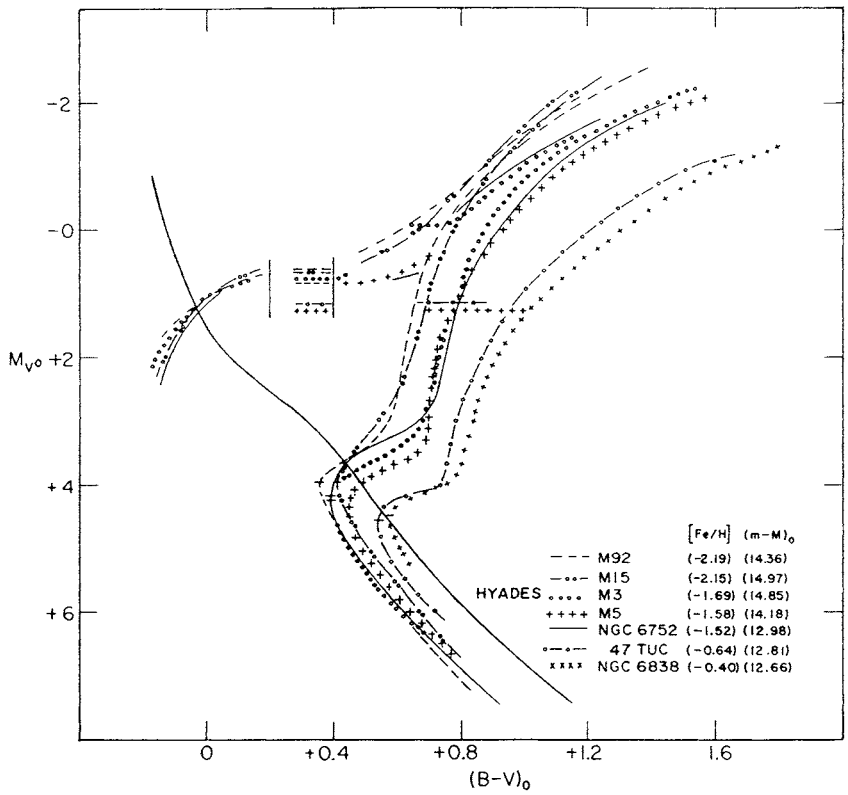


Figure 4. Composite HR diagram for seven globular clusters of different metallicity. Their absolute magnitude positions are set by main sequence fittings to the positions of fiducial main sequences as a function of metallicity. The sold curve from upper left to lower right is the position of the main sequence for stars of solar metallicity determined from trigonometric parallaxes.

different luminosities.

The cluster diagrams in figure 4 have been fitted vertically according to the expected fiducial main sequence positions as a function of metallicity (see HERTZSPRUNG–RUSSELL DIAGRAM and LARGE HOT SUBDWARFS AND EXTREME HORIZONTAL BRANCH STARS). From these main sequence normalizations, the spread in the vertical positions of the cluster variable instability strip as a function of metallicity is evident. This is the luminosity–metallicity relation that must be calibrated and accounted for in using RR Lyrae stars as distance indicators.

Calibration of the RR Lyrae absolute magnitudes

An early history

Shapley in 1918 was the first to calibrate the RR Lyrae absolute magnitudes. After calibration, he used his value of $\langle M_{\rm pg}({\rm RR})\rangle = -0.23$ to derive his famous new model of the Galaxy using his new distances to globular clusters which he had derived from the apparent brightness of the cluster variables in a few clusters. He then used these data as an intermediate step to calibrate the 25 brightest giant stars that could be more easily be measured in a large number of clusters. His radical new model of the Galaxy was that of a globular-cluster system that defined a distant center in the direction of Sagittarius, offset from the central position of the Sun that had been the main feature of the Galaxy model of Kapteyn.

His route to the RR Lyrae mean luminosity was to repeat Hertzsprung's 1913 luminosity calibration of the long-period field classical Cepheids. The method was that of statistical parallaxes to calibrate Leavitt's period–apparent luminosity curve for the long-period classical Cepheids in the Magellanic Clouds. Shapely then noted that a few globular clusters also had long-period variables which he took to be of the same type as the field Cepheids which he and Hertzsprung had just calibrated. Tying the cluster RR Lyrae to the cluster 'long-period Cepheids' and then tying this combination to the field Cepheids gave the RR Lyrae calibration quoted above as $\langle M_{\rm pg}\rangle = -0.23$.

Shapley's conclusion for a displaced center of the Galaxy and the consequent huge dimensions of the resulting Galaxy was so revolutionary that it was immediately challenged by Kapteyn and van Rhijn. These were the two greatest statistical astronomers of their day. They published what seemed to be a devastating, critical paper in 1922 in which they determined statistical parallaxes directly for the few field RR Lyraes whose proper motions were then known. Using the value of the solar motion relative to the slow-moving ordinary field stars, called now stars of population I, they derived a mean

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

absolute magnitude of $\langle M(\mathrm{RR})\rangle = +4.2$ which require a distance scale that was fully seven times smaller than Shapley's. Their scale would have returned the Galaxy back closer to Kapteyn's earlier small-scale model with the Sun near its center.

Shapley replied in a short note in *Harvard Bulletin 765* (1922), noting that his much brighter calibration was consistent with the discovery of RR-Lyrae-like variables in the SMC at $m_{\mathrm{pg}} = 16.1$. With the 'known' large distance of SMC from its classical Cepheids, which was not in contention by Kapteyn and van Rhijn, Shapley's claim that Magellanic Cloud RR Lyraes occur at $m_{\mathrm{pg}} = 16.1$ would clearly confirm his new RR Lyrae bright calibration near $M_{\mathrm{pg}} = 0$.

However, it is now known that Shapley's argument here is not correct. The true RR Lyrae stars that belong to SMC occur only at the much fainter level of $m_{\mathrm{pg}} = 19.6$. Either the variables noted but not identified by Shapley are foreground or, more likely, have longer periods than he believed. Nevertheless, Shapley's conclusion the $M(\mathrm{RR})$ is near $M(\mathrm{RR}) = 0$, not as faint as +4.4, was correct.

Shapley hints at the reason for the Kapteyn–van Rhijn incorrect conclusion by noting that the field RR Lyraes have a spherical spatial distribution, whereas the long-period Cepheids are confined to the plane, and therefore that the two types are different, requiring different precepts for the statistical parallax calculations. Nevertheless, his argument on this point is curious because Shapley himself, in his method through the cluster data calibrated with classical Cepheid field variables, had earlier combined the data for the two types as if they were related.

In a longer, much more closely reasoned reply, Shapley *(Harvard Circular* 237, 1922) points to the abnormally high radial velocity of RR Lyrae itself at -69 km s^{-1}, which is much higher than the radial velocity of any classical Cepheid. In addition, he pointed to the most unusual high proper motion of two of the Kapteyn–van Rhijn field variables, and therefore he correctly pointed out the high-velocity character of the spherically spread system of field RR Lyrae as a group. This requires that one must use a radically higher value for the solar motion for them than for the low-velocity population I objects for the statistical parallax calculation. The concept of the high-velocity stars and the nature of the Strömberg asymmetric drift was just coming into view in 1922, and Shapley again correctly identified the Strömberg drift of the RR Lyrae system, perhaps for the first time.

A number of direct determinations via statistical parallaxes, using a correctly high solar motion, were then made from 1922 to the present, all generally confirming Shapley's $\langle M(\mathrm{RR})\rangle \sim 0$ value. The two most important studies were by P F Bok and C D Boyd at Harvard in 1933 and R E Wilson at Mount Wilson in 1939 who had much more complete proper motion and radial velocity data than were available to Kapteyn and van Rhijn.

Because of the consistent results from the several new statistical parallax studies in the decades of the 1920s and 1930s, the problem was considered solved until Baade failed to resolve RR Lyraes in the disk of M31 with the Palomar 200 in telescope in the early 1950s. His result showed either that the classical Cepheid distance to M31 was incorrect (based on R E Wilson's 1939 recalibration of the classical Cepheid P–L relation) or that $\langle M(\mathrm{RR})\rangle$ had to be as faint as $\langle M(\mathrm{RR})\rangle \sim +2$. It was then that the method shown in figure 4 was introduced after the globular cluster main sequences had been identified beginning in the early 1950s showing, in fact, that $\langle M(\mathrm{RR})\rangle \sim 0$ was indeed correct.

Modern methods

Four principal modern methods have been used to determine the absolute magnitude of the RR Lyrae variables.

The three other methods besides that of statistical parallaxes are (1) the Baade–Becker–Wesselink way of combining the light and radial velocity curves to determine the radius of a pulsator, from which the absolute magnitude can be calculated from the radiation laws once the temperature is known (it must be measured independently), (2) pulsation theory together with the Oosterhoff–Arp–Preston metallicity effect to give the slope of the correlation of M_{bol} with [Fe/H], together with an independent method for the zero point, and (3) main sequence fitting of globular cluster color–magnitude diagrams to calibrate the absolute level of the horizontal branch as in figure 4.

The chief problem with the powerful main-sequence-fitting method is the lack of precise knowledge, at the 0.1 mag level, of the position of the fiducial main sequence as a function of metallicity. Eventually, precision trigonometric parallaxes from space interferometry measurements of a large sample of subdwarfs with a range of metallicity will solve this problem. Nevertheless, the method is now sufficiently accurate to show beyond doubt that the luminosities of the horizontal branches, and therefore of the cluster variables, is a function of metallicity (figure 4).

Present status of the calibration

The results of the four methods to derive $\langle M(\mathrm{RR})\rangle$ are in agreement at the 0.3 magnitude level, but there still remain (1999) systematic differences in $\langle M(\mathrm{RR})\rangle$ as a function of metallicity at that level.

Figure 5 compares the extant calibration using two of the four methods. The individual data points summarize results for individual field RR Lyraes determined by several independent groups using the Baade–Becker–Wesselink method. Details of the method that combines the light curve with the radial velocity curve to determine the stellar radius, and then the luminosity, are set out in a number of the references listed at the end, principal of which is the comprehensive review by Gautschy.

The line marked '*Oosterhoff pulsation*' in figure 5 uses the period ratios similar to those shown in figure 3 between RR Lyrae stars of different metallicities, but read

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

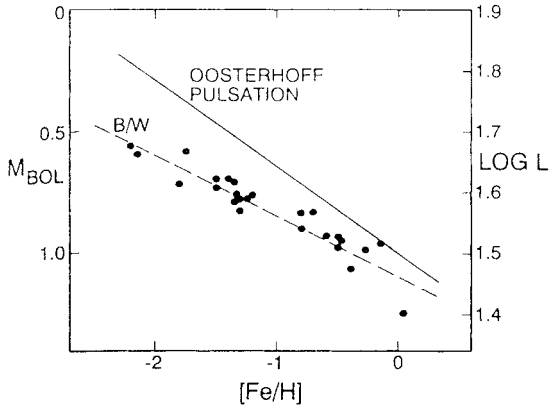


Figure 5. Comparison of the RR Lyrae star calibration (circa 1995) of the bolometric absolute magnitude determined using the Baade–Becker–Wesselink and the Oosterhoff–Arp–Preston metallicity-effect methods. The differences in the slope and zero points of the two correlations are crucial in the many applications of RR Lyrae stars to astronomical problems. The calibration of $\langle M(\mathrm{RR})\rangle$ as a function of metallicity is not yet (1999) solved at the level of the differences shown.

at the blue edge of the instability strip. The blue-edge periods are well correlated with metallicity at the rate of $\log P = -0.112[\mathrm{Fe/H}] - 0.500$, which, when put in the theoretical pulsation equation that relates the luminosity to the period, temperature and mass, predicts the pulsation relation in figure 5. Conversion to M_V gives the calibration as $M_V(\mathrm{RR}) = 0.30[\mathrm{Fe/H}] + 0.94$. A detailed account of the method can be found in the first paper in the *IAU Symposium 139* conference report cited in the bibliography.

Astronomical used of RR Lyraes in clusters and in the field

Because RR Lyrae variables are such excellent standard candles once the metallicity dependence is calibrated out, and because they are relatively bright at $\langle M\rangle \sim 0.5$, they have been used for many purposes. Among many we list five. An extensive literature exists for each.

(1) The system of field RR Lyrae stars shows the Strömberg asymmetric drift in its kinematics. This, of course must be accounted for in the calculations for statistical parallax, but simply its existence as a function of metallicity gives an important clue to the chemical and dynamic evolution of the Galaxy. The asymmetric drift velocity varies from $\langle V\rangle \sim -50$ km s^{-1} for RR Lyraes with solar metallicity to $\langle V\rangle \sim -250$ km s^{-1} for the most metal-poor variables of the extreme population II. Many of the latter are on retrograde galactic orbits. The early history of the collapse of the protogalaxy as a function of time and of the progressive chemical enrichment of the disk and halo can be studied from such data.

(2) Because of the large amplitudes of their light curves and their short periods, RR Lyraes are easy to discover. They have been extensively used in studies of Galactic structure. Among other data, they provide the identification of stellar populations in the halo, bulge, center and the thick disk components of the Galaxy. Two large surveys for RR Lyraes in the halo, one by Saha at Palomar and the other by Kinman at Lick, have mapped the extent of the halo and have determined the density distribution in various directions at high Galactic latitude. A large discovery survey in the bulge has been completed at Palomar by Plaut to determine the population characteristics at intermediate latitudes in the bulge toward the Galactic center.

(3) The ages of globular clusters have been determined to be 13 billion years by comparing the difference in apparent magnitude of the RR Lyraes and the main sequence turn-off point in globular clusters of different metallicities. The question of whether there is an age–metallicity relation for the clusters depends on which RR Lyrae absolute magnitude calibration as a function of metallicity is used. If the steep metallicity dependence is used that is required to explain the Oosterhoff–Arp–Preston effect (see previous section), then all globular clusters are closely the same age. However, if the more shallow dependence of $M_V(\mathrm{RR})$ on metallicity (figure 5) is used that is determined by Baade–Becker–Wesselink methods, then there would be an age–metallicity variation for the globular clusters. The sense would then be that metal–poor clusters would be older than the more metal-rich clusters by several billion years. The problem has not been resolved at the time of writing (1999).

(4) One way to test whether the period–luminosity relation for the long-period classical Cepheids depends on metallicity is to compare the distance to a given galaxy determined by its RR Lyrae stars and independently by its classical Cepheids using a CEPHID PERIOD–LUMINOSITY RELATION with no metallicity term. If the two distances agree, then no metallicity correction would be need for the long-period Cepheids. Of course, a correct RR Lyrae–metallicity calibration is required to make the test. At the time of writing, with the several versions of the RR Lyrae calibration extant as in figure 5, the method, although very powerful in principle, is not yet definitive. The metallicity dependence of the RR Lyrae calibration is the centerpiece here.

(5) The physics of the pulsational instability of stars and of the mechanism that drives the pulsation has been understood at a deep level by comparing the theoretical pulsation calculations with various observed aspects of light and velocity variations of both the RR Lyrae stars and the long-period classical Cepheids. The literature is vast. The theory is elegant. The agreement between theory and observations is excellent. The subject is one of the major triumphs of modern astrophysics (see STELLAR PULSATION).

Bibliography

Cox J P 1980 *Theory of Stellar Pulsation* (Princeton, N J: Princeton University Press)

Gautschy A 1987 The Baade–Wesselink method *Vistas in Astronomy* vol 30, ed A Behr and A Roy (Oxford: Pergamon) p 197

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hoffmeister C, Richter G and Wenzel W 1985 *Variable Stars* (Berlin: Springer)

Nemec J M and Matthews J M (eds) 1993 *New Perspectives on Stellar Pulsation and Pulsating Variable Stars (IAU Colloq. 139)* (Cambridge: Cambridge University Press)

Rosseland S 1949 *The Pulsation Theory of Variable Stars* (Oxford: Clarendon Press)

Schmidt E G (ed) 1989 *The Use of Pulsating Stars in Fundamental Problems of Astronomy (IAU Colloq. 139)* (Cambridge: Cambridge University Press)

Smith H A 1995 *RR Lyrae Stars* (Cambridge: Cambridge University Press)

Tsesevich V P 1975 The RR Lyrae stars *Pulsating Stars* ed B V Kukarkin (New York: Wiley)

Allan Sandage

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

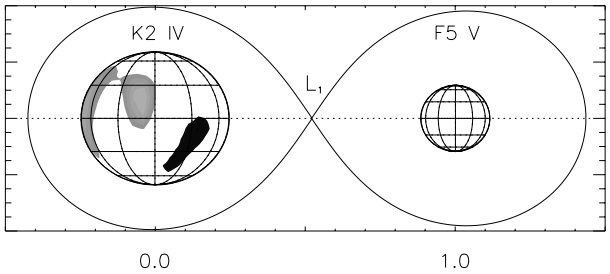


Figure 1. A three-dimensional map of the close binary system RS CVn, the prototype of a class of intrinsically variable systems showing solar-like activity phenomena of magnetic origin, as derived from cool spot modelling of the 1964 optical light curve (Rodonò *et al* 1995). The projected two-dimensional Roche lobes in the plane of sky and the inner Lagrangian point (L_1) are also shown.

RS Canum Venaticorum

RS CVn = RS Canum Venaticorum (BD +36° 2344, HD 114519; $\alpha_{1950.0} = 13^h08^m18^s$, $\delta 1950.0 = +36°12'.0$; $\Delta d = 2.79$ s yr^{-1}, $\Delta\delta = -0.319'$ yr^{-1}) is a detached CLOSE BINARY STAR (F5 V+K2 IV) in the CANES VENATICI constellation (cf Rodonò *et al* 1995 for a discussion of the spectral classification). In addition to the canonical eclipse-related light variability, early photometric observations had already shown uneven outside-of-eclipse maxima suggesting low-amplitude intrinsic variability of the component stars, which led to the classification of RS CVn as an intrinsically VARIABLE STAR in the *General Catalogue of Variable Stars*. An illustrative map of the RS CVn system is shown in figure 1.

RS CVn is considered the prototype of a class of close binaries with outside-of-eclipse variability in optical light essentially consisting of a so-called 'wave-like distortion' or 'photometric wave' (figure 2), which is characterized by an amplitude up to a few tenths of a magnitude, variable on time scales from several months to years. Moreover, this 'photometric wave' usually systematically moves towards decreasing orbital phases at a variable rate. This peculiar optical behavior can be ascribed to one or both components of the binary system by assuming that their rotation periods are slightly shorter than the orbital period, so that the dark or bright surface inhomogeneities producing the observed variability appear on the visible stellar hemisphere(s) at systematically decreasing orbital phases. Actually, the first interpretation of this phenomenon, as suggested by Rodonò (1965) soon after he discovered the systematic variability of the RS CVn light curve, was that solar-like activity phenomena, such as dark spots or bright PLAGES, form in the photosphere of one or both components, so that the visibility of such surface features, and the related effects on the optical light curve, are modulated by the star's rotation. A slightly asynchronous rotation of the spot-forming region on a differentially rotating photosphere may account for the characteristic variability and its systematic backwards drift on the light curve. Most of the RS CVn systems show a 'photometric wave' that migrates towards decreasing orbital phases, but in a few cases where the reverse situation is observed and in a few other systems (SS Boo, V 711 Tau, CG Cyg, AR Lac) changes of the migration direction have occurred (see also STELLAR PHOTOMETRY). Moreover, for SS Boo and V 711 Tau there is some evidence of cyclic changes in the migration direction. Stellar spots and flares, which are occasionally observed as well (cf Catalano 1990), appear to be orders of magnitude more powerful and extended than on the Sun.

The identification of a new class of variable stars, named after the prototype system RS CVn, was proposed by Oliver (1974) and Hall (1976), who, in addition to the already mentioned photometric wave and the well-known presence of Ca II H and K emission lines in their spectra, identified the following distinctive characteristics of RS CVn close binaries: (a) 'undersized' subgiant component, well within its ROCHE LOBE, (b) fast rotation, i.e. almost synchronized binaries with orbital periods shorter than about 15 days and (c) orbital period variation (for further details see the review articles by Rodonò 1983, Catalano 1983, Hall 1991, Rodonò 1992).

In addition to spots and flares, RS CVn stars show a rather complete set of solar-like activity signatures, such as chromospheric and transition region bright plages, as derived from synoptic studies of enhanced UV line (Rodonò *et al* 1986, 1987, Neff *et al* 1989, Linsky *et al* 1989), coronal x-ray (Walter *et al* 1983, White *et al* 1990, Rodonò *et al* 1999) and microwave (Gibson and Hijellming 1985, Trigilio *et al* 1993, Umana *et al* 1995) emissions. In non-optical wavelength domains, the quiescent flux from late-type cool stars being rather low, the characteristic high-temperature flare phenomena are best observed for contrast effects. In a rather complete compilation of magnetic activity signatures detected on stars (Linsky 1988), only two question marks concerning the detection of magnetic fields and optical flares were listed. Both question marks were positively answered within a few years (Saar 1988; Donati *et al* 1990; Foing *et al* 1991).

Activity phenomena of presumably magnetic origin have been detected also on other types of close binaries, e.g. BY Dra type (Bopp and Fekel 1977, Rodonò 1983) and UV Cet type M–K dwarfs (Rodonò 1990), W UMa type contact systems (Eaton *et al* 1980), Algols (Olson 1984) and cataclysmic variables (Bianchini 1990). These different types of binary systems share the common properties that one or both component stars have deep convection zones. Since they are tidally locked close binaries, they are also fast rotators. These same two factors (deep convection and rapid rotation) also exist in several types of single stars that display the very same activity phenomena (Hall 1991). When these two conditions are met, an α–ω dynamo can efficiently operate at the base of the convection zone (see DYNAMOS: SOLAR AND STELLAR). This leads, with a sometime substantial contribution of surface turbulence, to strong intensification of the star's magnetic fields. When the magnetic pressure surpasses the gas pressure, magnetic flux tubes can emerge at the stellar surface giving rise to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

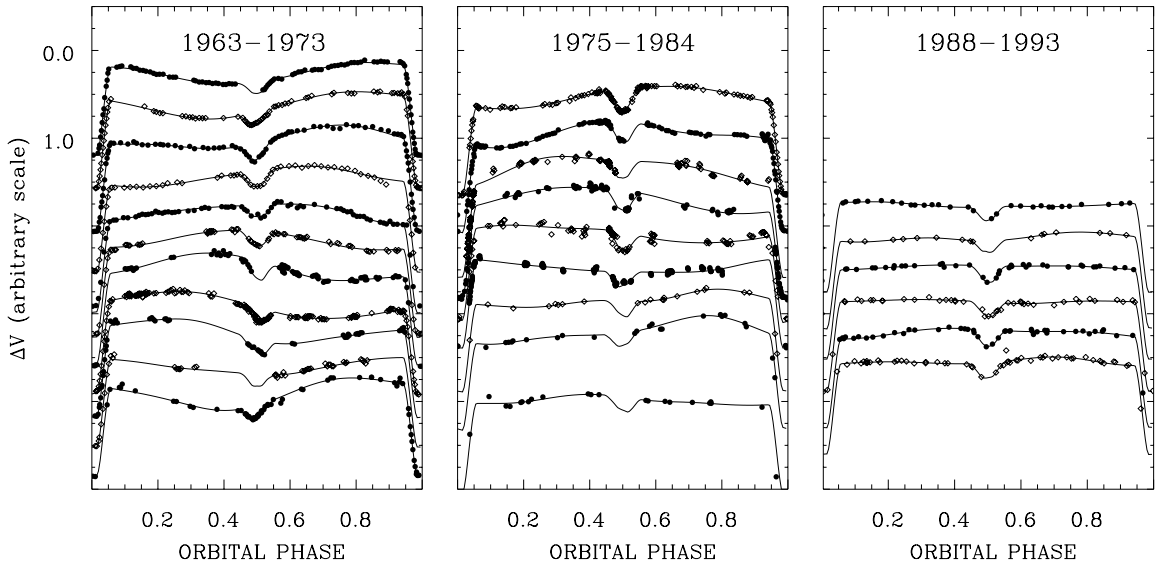


Figure 2. A sequence of RS CVn light curves over a 30 yr interval showing the characteristic 'wave-like distortion' or 'photometric wave' that migrates toward decreasing orbital phases. Each yearly light curve is shifted vertically with respect to the previous one. A blank vertical space is left if observations are not available. The thin continuous curves are maximum-entropy spot-model fits by Rodonò *et al* (1995). Most of the plotted data were obtained at Catania Astrophysical Observatory, where systematic photometry of RS CVn systems and other magnetically active stars has been carried out since 1963 (Rodonò 1965, Chisari and Lacona 1965).

whole panoply of activity phenomena that are observed on the Sun and several other late-type stars. The emerging magnetic loops interact with pre-existing ones and the subsequent magnetic line reconnections lead to energy dissipation. Therefore, magnetic fields appear to be the principal energy source of the observed phenomena.

Among magnetically active stars, RS CVn binaries offer, however, the cleanest, rather stable and understandable phenomena, when compared, for example, with the complex variability of contact binaries or binaries containing a degenerate component, where other types of phenomena simultaneously contribute to the observed variability. This makes it rather difficult to identify and disentangle pure solar-like activity phenomena from the observations and, consequently, it renders rather uncertain their study and interpretation.

In addition to allowing us to study the physical characteristics of surface activity phenomena, the monitoring of RS CVn stars and other active systems has led to the detection of activity cycles (see Strassmeier *et al* 1993) and has triggered the study of their effects on the star and system dynamics (see also STELLAR DYNAMICS): short-term variations of the orbital period appear to be compatible with changes of the quadrupole moment of the active component that are produced by the modulation of the star's internal magnetic field and angular momentum distribution throughout a magnetic cycle (Lanza and Rodonò 1999).

In conclusion, RS CVn systems offer one of the best and most complete astrophysical laboratories for studying solar-like activity phenomena in stellar photospheres and overlying atmospheric layers up to corona levels.

Bibliography

Bianchini A 1990 *Astron. J.* **99** 1941
Bopp B W and Fekel F C 1977 *Astron. J.* **82** 490
Catalano S 1983 *Activity in Red-dwarf Stars (IAU Colloq. 71)* ed P B Byrne and M Rodonò (Dordrecht: Reidel) p 343
Catalano S 1990 *Active Close Binaries* ed C Ibanoglu (Dordrecht: NATO Advanced Study Institute, Kluwer) p 411
Chisari D and Lacona G 1965 *Mem. Soc. Astron. Ital.* **36** 483
Donati J-F, Semel M, Rees D, Taylor K and Robinson R 1990 *Astron. Astrophys.* **232** L1
Eaton J A, Wu C C and Rucinski S M 1980 *Astrophys. J.* **239** 919
Foing B M, Char S, Ayres T, Neff J E and Zhai Di-S *et al* 1991 *Astron. Astrophys.* **292** 543
Gibson D M and Hijellming R M (ed) 1985 *Radio Stars* (Dordrecht: Reidel)
Hall D S 1976 *Multiple Periodic Variable Stars (IAU Colloq. 29)* ed W S Fitch p 287
Hall D S 1991 *The Sun and Cool Stars: Activity, Magnetism, Dynamos (IAU Colloq. 130)* (Berlin: Springer) p 353
Lanza A F and Rodonò M 1997 *The Earth and the Universe* ed G Asteriadis *et al* (Aristotle University of Thessaloniki: Ziti) p 161
Lanza A F and Rodonò M 1999 *Astron. Astrophys.* **349** 887
Lanza A F, Catalano S, Cutispoto G, Pagano I and Rodonò M 1998 *Astron. Astrophys.* **332** 541
Lanza A F, Rodonò M and Rosner R 1998 *Mon. Not. R. Astron. Soc.* **296** 893
Linsky J L 1988 *Multiwavelength Astrophysics* ed F Cordova (Cambridge: Cambridge University Press) p 49

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Linsky J L, Neff J E, Brown A *et al* 1989 *Astron. Astrophys.* **211** 173
Neff J E, Walter F M, Rodonò M and Linsky J L 1989 *Astron. Astrophys.* **215** 79
Oliver J P 1974 *PhD Thesis* University of California, Los Angeles CA
Olson E C 1984 *Advances in Photoelectr. Photom.* **3** 15
Rodonò M 1965 *Physics 'Laurea' Thesis* Catania University
Rodonò M 1983 *Adv. Space Res.* **2** (9) 225
Rodonò M 1990 *Flare Stars in Star Clusters, Associations and in the Solar Vicinity (IAU Symp. 137)* ed L V Mirzoyan, B R Petterson and M K Tsvetkov p 371
Rodonò M 1992 *Evolutionary Processes in Interacting Binary Stars (IAU Symp. 151)* ed Y Kondo, R F Sistero and R S Polidan (Dordrecht: Kluwer) p 71
Rodonò M, Cutispoto G, Pazzani V *et al* 1986 *Astron. Astrophys.* **165** 135
Rodonò M, Byrne P B, Neff J E *et al* 1987 *Astron. Astrophys.* **176** 267
Rodonò M, Lanza A F and Catalano S 1995 *Astron. Astrophys.* **301** 75
Rodonò M, Pagano I, Leto G *et al* 1999 *Astron. Astrophys.* **346** 811
Saar S 1988 *Astrophys. J.* **324** 411
Strassmeier K G, Hall D S, Fekel F C and Scheck M 1993 *Astron. Astrophys. Suppl.* **100** 173
Trigilio C, Umana G and Migenes V 1993 *Mon. Not. R. Astron. Soc.* **260** 903
Umana G, Trigilio C, Tumino M, Catalano S and Rodonò M 1995 *Astron. Astrophys.* **298** 143
Walter F M, Gibson D M and Basri G S 1983 *Astrophys. J.* **267** 665
White N E, Shafer R A, Horne K, Parmar A N and Culhane J L 1990 *Astrophys. J.* **350** 776

Marcello Rodonò

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rubin, Vera Cooper (1928–)

Astronomer, born in Philadelphia, worked at the Carnegie Institute on problems of extragalactic research, including the dark matter that forms a major component to galaxies and reveals its existence by its gravitational effect on their stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rue, Warren de la (1815–89)

Born in Guernsey, astronomer at the Kew Observatory, astrophotography pioneer. Showed in the total solar eclipse of 1860 that the prominences seen at the edge of the Sun and Moon were independent of the motion of the Moon and thus solar appendages.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Runaway Star

Term used to describe a star that is moving with very high velocity, typically hundreds of kilometers per second, relative to the local standard of rest. Such exceptionally high velocity presumably indicates an 'explosive' departure of the star from the location of its formation. The most likely explanation of this is that the star was a minor component of a binary system, ejected when the primary component exploded as a supernova.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Russell, Henry Norris (1877–1957)

Astronomer, born in Oyster Bay, NY, spent nearly all his life working at Princeton University. He spectroscopically studied eclipsing binary stars to determine the masses of their component stars. At first collaborating with the British astronomer Hinks at Cambridge, he started to measure stellar parallaxes and, plotting the absolute magnitudes of stars whose distance he had thus measured, against their spectral types, he found the correlation now known as the main sequence of the *Hertzsprung–Russell diagram*. He distinguished between giant stars and dwarfs. He proposed that stars cooled along this correlation. This theory of stellar evolution was superseded, but the H–R diagram continues to provide a tool with which to test theories of stellar evolution. Russell applied MEGHNAD SAHA'S theory of ionization to stellar atmospheres and determined the abundance of the elements in stars, including, after the foundation work by CECILIA PAYNE-GAPOSCHKIN, the great abundance of hydrogen. In the theory of stellar structure, he showed that the size, temperature, etc, of a star at each stage of its evolution can be found solely from its mass, chemical composition and age (the *Vogt–Russell theorem*).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Russian Astronomy

Russia is a vast multinational country whose boundaries have been changed to a great extent. The Russian Empire at the beginning of this century included many countries, which are now independent states. This article is concentrated on the history of astronomy and current activity on the territory of modern Russia with necessary brief astronomical excursions in the history of the states.

History

Archaeo-astronomical studies in Russia have brought some material evidence of astronomical knowledge of previous generations (e.g. thousand year old ring and rod inscribed with a primitive astronomical calendar found in an excavation near Achinsk in Siberia). The first astronomical observatory was established in 1690 by the archbishop of Archangel'sk for navigational purposes. In 1701 the Russian tsar Peter the Great who was highly impressed by his visit to Greenwich and Copenhagen observatories established in Moscow 'The School for Navigation and Mathematics' with an astronomical observatory, which was headed by J Bruce. In 1725 the Astronomical Observatory of the Petersburg Academy of Sciences was founded. The well-known French astronomer Joseph-Nicolas DELISLE was the first director of this well-equipped (according to the standards of that epoch) observatory.

In the 18th century most astronomical observations, both at observatories and during numerous expeditions, were aimed to maintain a network of points with accurate coordinates. Thus the 'Catalogue of astronomical coordinates' by S Rumovsky contained coordinates of 82 points with an accuracy, good for that period, of 8′. The first astrophysically important observation was carried out by a famous Russian scientist of encyclopedic knowledge, M LOMONOSOV, who in 1762 observed the passage of Venus across the solar disk and discovered the atmosphere of that planet. Growing industrial and military needs caused quick development of astrometry and practical astronomy in Russia, which led to the construction of new observational facilities.

In the 19th century new observatories at Kharkov (1808), Derpt (Tartu) (1809), Nikolaev (1827), Kiev (1845) and Odessa (1871) universities were built. In 1831 the construction of a new Moscow observatory began. However, the most prominent event in Russian astronomy of the 19th century was the construction of the Pulkovo Observatory (figure 1). This observatory was founded in 1839. The first director of the observatory was the famous astronomer Vassily (Wilhelm) STRUVE, well known for his classical studies of double stars and for his guess (1847) about the existence of interstellar absorption. Because of the outstanding organizational abilities of its first director, the Pulkovo Observatory in 30 yr after its foundation became one of the world's best astronomical institutions and had unprecedented high-precision astrometric instruments. Pulkovo fundamental stellar catalogues of 1845, 1865, 1885, 1905, 1930 and 1955 were most precise and the observations performed in Pulkovo were included into the world's popular fundamental catalogues with the highest weight. Stellar astronomy also was a field of interest to astronomers from Pulkovo. The world's best instruments were installed at Pulkovo observatory, e.g. a REFRACTOR with $D = 76$ cm (1885). This telescope was destroyed (as were almost all the instruments and buildings of the observatory) by fascist troops attacking Leningrad during World War II. In the second half of the 19th century astrophysics in Russia was developing too. The theory of cometary tails (F Bredikhin, 1862–1877) and the first observational evidence of periodic changes of the radial velocity in δ Cep (which turned out to be typical for cepheids) were the most important contributions of Russian astrophysicists of the 19th century.

The very beginning of the 20th century was relatively favorable for the development of astronomy in Russia. New observatories at the Kazan University (1901), L'vov University (1907) and Simeiz Observatory (1908) were built. World War I and the October revolution, followed by years of economic dislocation, were negative factors for science in Russia. In 1922–1924 an important discovery was made by A FRIEDMANN, who found non-stationary solutions of the Einstein equations and thus contributed to the fundamentals of cosmology.

In the period of the strengthening of the state in 1930s new observatories were built (in ABASTUMANI—1932 and Irkutsk—1931) and new big plans of scientific development were elaborated. The leading role of the Academy of Sciences in fundamental research was then consolidated. An all-Union coordination body—the Astronomical Council of the USSR Academy of Sciences—was established in 1936. Members of this body were responsible for strategic planning of astronomical research in the USSR. New approaches were advanced. World War II prevented the realization of these ambitious plans.

After the war intensive investments in fundamental science (particularly in astronomy) brought a flourishing period. Astrophysical instruments require particularly dark and clean skies, so the best places for the new observatories were found in the southern mountains. Construction of the new large CRIMEAN ASTROPHYSICAL OBSERVATORY (CrAO) (1945–1954), BYURAKAN ASTROPHYSICAL OBSERVATORY (1946–1956), observatories near Alma-Ata (1951), Ashkhabad (1946), Goloseevo (founded in 1944), Dushanbe (1958), Shemakha (1960) and a solar station near Kislovodsk were essential steps of postwar development. The Pulkovo Observatory was reconstructed. Great progress in radioastronomy was not missed in the USSR (as happened with cybernetics and genetics) and new instruments and dedicated observatories were built. The radioastronomical station of the Institute of Radio Physics near Gorky (1949), radioastronomical station of the Lebedev Physical Institute (1956), radioastronomical observatory in Latvia (1958), radio telescopes in CrAO, Kharkov etc were evidence of attention to radioastronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. Pulkovo Observatory reconstructed after World War II.

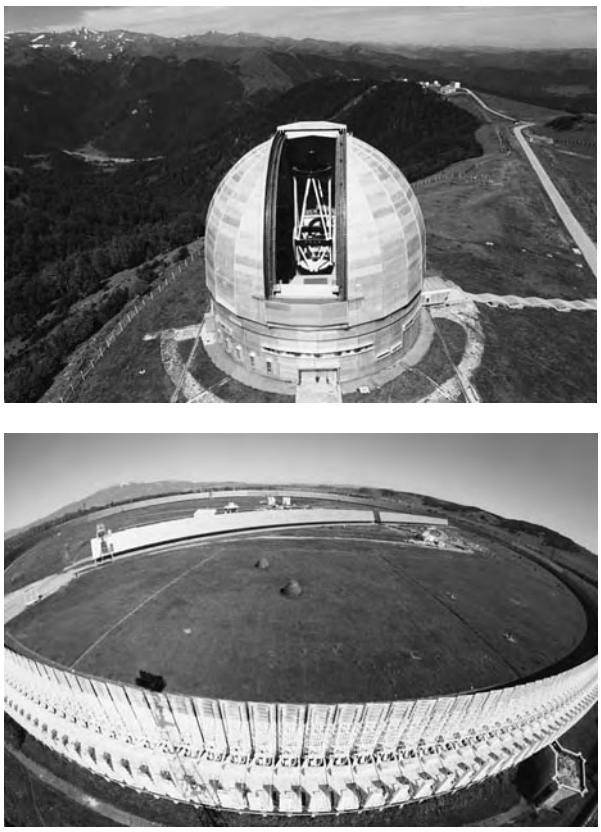


Figure 2. Special Astrophysical Observatory in the Caucasus: top—Large Alt-azimuth Telescope, bottom—RATAN-600 radio telescope.

Prominent discoveries of that period were made by V AMBARTSUMIAN (young stellar associations—1947, ideas on the important role of nuclei in active galaxies—1958) and I SHKLOVSKII (ideas on the observation of interstellar molecules in the radio range—1949, on the formation of planetary nebulae by red giants loosing their mass at the latest stages of stellar evolution—1956).

The peak of this postwar thriving of ground-based astronomy was marked by the construction of the SPECIAL ASTROPHYSICAL OBSERVATORY (SAO) in North Caucasus (1966–1976) (figure 2). This was a nationwide project which required huge resources. The 6 m optical telescope was for a long time the world's largest one. The unique 600 m radio telescope RATAN-600 even now is one of the most powerful instruments in the world. The 6 m telescope is located in a region whose astroclimate is far from ideal (although this place seems to be one of the best in the Russian Federation); that is why new astronomical centers were built in the Pamir region where places with record seeing parameters were found. All-Union observatories in Mydanak and Sanglok were the most promising for development of observational facilities in the USSR. At the same time space astronomy was quickly developing.

The serious problem was the lack of modern detectors (and computers) as a consequence of the general lag of the USSR in some high-technology fields. That substantially reduced the efficiency of large optical astronomical instrumentation. Thanks to international cooperation this situation was somewhat improved. A special program for the design and implementation of modern astronomical detectors was supported by the Academy of Sciences but the epoch was over.

During and after the 'perestroika' period the economic situation in the USSR drastically worsened. This immediately influenced scientific development. In the last 15 yr no new observatories and no large instruments have been built. All programs including prestigious space experiments were slowed down or stopped. After the disintegration of the USSR, Russia and other former USSR (fSU) republics were busy with political and social problems. Strong economic and sometimes political boundaries appeared. Most observatories are located outside Russia and this prevents effective use of them. Regular funds for astronomy as well as for other fundamental sciences in Russia were cut at least by an order of magnitude. In the process of intensive social separation the Russian astronomers (as most of population) quickly sank close to the social bottom. The various support programs were initiated by the world astronomical community to help Russian astronomers. Yet Russia remains a major astronomical country.

Astronomical institutions

Astronomy in Russia is a matter of research in academic institutions (those of the Russian Academy of Sciences—RAS) and at some universities.

The SAO (of the RAS) is Russia's main center for ground-based space research. The observatory is located in the south of Russia, in the Caucasus mountains of Karachaevo-Cherkessia. The main instruments of the observatory are the optical telescope BTA (Large Alt-azimuth Telescope) which primary mirror has a diameter of 6 m and the radio telescope RATAN-600 with an antenna diameter of 600 m. The observatory has highly experienced experts and modern detectors (some of them were provided by international cooperation). Study

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of the universe as a whole, analysis of non-stationary energetic processes in galactic and extragalactic objects of various types, investigation of evolution and chemical composition of our Galaxy and the nature of solar activity and the ecology of the upper atmosphere and nearby space are basic research activities of the observatory.

The Main Astronomical Observatory of the RAS (Pulkovo Observatory) is located in the suburb of a big city. The observatory lost its own observational facilities in the south and researchers actively seek (and develop) cooperation with observatories over the world. Stellar astronomy, astrophysics and solar research remain as prime research directions of the observatory.

The Institute of Applied Astronomy of the RAS in St Petersburg has newly built observational facilities—30 m dish radio telescopes over Russia, originally designed for the construction of national time-coordinate system.

The Institute of Astronomy of the RAS (former Astronomical Council) is situated in Moscow. Research directions of the institute include astrophysics and astronomical data processing. Zvenigorod Observatory of the institute was specialized for tracking satellites and relevant basic research (Earth rotation, geodynamics, study of near-Earth objects).

The Astro Space Center of the Lebedev Physical Institute of the RAS, whose own main observational facility is a 22 m radio telescope in Pushchino, is carrying out studies of cosmology, physics of extragalactic objects and interstellar matter.

The Radio-Astronomical Station of the Institute of Radio Physics in Gorky is involved in studies of interstellar matter.

Sternberg State Astronomical Institute of Moscow University is a large institution with multidirectional research activity: cosmology, astrophysics, celestial mechanics, astrometry etc.

There are also smaller university observatories and institutes in St Petersburg, Kazan, Ekaterinburg and research groups in some other high schools. Big groups of experts in high-energy astrophysics and planetary research work in the Space Research Institute of the RAS (Moscow). Groups for solar physics and solar–terrestrial relations work in the Institute of Terrestrial Magnetism and Propagation of Radio Waves of the RAS in Troitsk (and a similar institution in Irkutsk) and small groups for theoretical astrophysics do studies in other research centers.

Space astronomy

In the USSR the number of astronomical experiments in space was continuously increasing. Numerous experiments on board Mir and Salyut stations in the field of high-energy astrophysics and solar physics as well as many dedicated missions to the Moon, Venus, Mars, Halley's comet, missions for study of solar–terrestrial relations and the Sun and the astrophysical missions ASTRON (80 cm telescope for UV spectroscopy), GRANAT (high-energy astrophysics) and RELICT (microwave background measurement) were carried out successfully. In the last 10 yr Russia has reduced ambitions in the field but still has a rather big space program. The RAS is responsible for preparation of the program of fundamental research. The RUSSIAN AVIATION AND SPACE AGENCY provides the general control and funding, while the scientific institutions are the main contractors for the following projects: SPECTRUM-X-GAMMA—a high-energy observatory (Space Research Institute), Spectrum-R—a radio interferometer with a space 10 m aperture radio telescope and base distance 70 000 km (Astro Space Center), Spectrum-UV—a general-purpose observatory with a 170 cm aperture UV–optical telescope equipped with spectrographs, imaging camera and polarimeter (Institute of Astronomy), CORONAS-F—project for investigating solar–terrestrial relations (Institute of Terrestrial Magnetism and Propagation of Radio Waves), 'Fobos-Ground'—delivery of a sample of the Fobos soil to the Earth (Space Research Institute). In the last few years the space program has encountered serious financial difficulties and launch dates are somewhat indefinite, which worries the Russian project teams as well as international cooperations.

Organizational features

Researchers. The total number of professional astronomers as estimated at the beginning of 1999 is about 1000 (members of the IAU—300, members of the EAS—250).

Funding. Most astronomical institutions are funded by the state via RAS or the Ministry of Education. Annual expenditures for astronomy (except space projects) in 1999 are expected to be equivalent to US$3.5 million.

Coordinating bodies. General cooperation in fundamental research is carried out by the scientific councils (SCs). In 1999 the list of the councils included the SC for astronomy, SC for physics of the solar system, SC for solar–terrestrial relations and the Space Council (includes sections for space astronomy, solar system, cosmic rays etc). All four are established by the RAS. The Council for Education in Astronomy was established by the Ministry of Education. The councils include leading persons in the field and are devoted to strategic planning of research and educational policy in astronomy.

Astronomical societies in Russia. These began to appear from 1890 (Russian Astronomical Society, Petersburg). The most long-lived society was the semiprofessional All-Union Astronomical-Geodetic Society (founded in 1932). The first fully professional society (The EURO-ASIAN ASTRONOMICAL SOCIETY—EAAS) was established in 1990. It is officially legalized as an international society. The EAAS performs its activities mainly on the territory of the fSU. The main goals of the EAAS are to maintain the development of astronomy and to reinforce scientific contacts between astronomers of the fSU and their colleagues all over the world. About 700 professional astronomers from Russia and the other fSU countries are members of the society. The EAAS has been affiliated to the EAS since 1992.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astronomical journals. The main professional journal published in Russia is *Astronomicheski Zhurnal* (*Astronomy Reports*); two others are *Pis'ma in Astronomicheski Zhurnal* (*Astronomy Letters*) and *Astronomicheski Vestnik* (*Solar System Research*). Some institutions have non-regular series of publications.

Bibliography

Kolchinsky I G, Korsun A A and Rorriges M G 1986 *Astronomers* (Kiev: Naukova Dumka) (in Russian)

Ponomarev D N 1987 *Astronomical Observatories of the Soviet Union* (Moscow: Nauka) (in Russian)

Boris Shustov

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Russian Aviation and Space Agency

The Russian Space Agency (RKA) was created on 25 February 1992 by a decree issued by the President of the Russian Federation. It was formed after the break-up of the former Soviet Union and the dissolution of the Soviet space program. The RKA uses the technology and launch sites that belonged to the former Soviet space program. This includes payment to Kazakhstan for use of the Baikonur Cosmodrome.

The official responsibilities of the RKA were codified in August 1993, in the Russian Federation Law on Space. It defines the state policy in space research and exploration for peaceful purposes, and execution of the Federal space program. This includes promotion of Russia's interests in the international space arena.

RKA has centralized control of Russia's civilian space program, including all manned and unmanned non-military space flights. The agency now oversees a growing number of bilateral and multilateral accords with other national space agencies, e.g. the Shuttle–Mir program and the International Space Station.

Since 1994, the agency has gained greater power over Russian industrial concerns. Under the 1994 government decree, 38 enterprises of the rocket and space industry were transferred to RKA. In 1998 RKA was given a control over 38 more state companies and federal shares in 22 joint-stock companies. On 11 March 1999 it was announced that RKA would gain control over 350 companies of the aviation industry and be renamed the 'Russian Aviation and Space Agency'.

For further information see
http://www.rka.ru/english/eindex.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rutherford Appleton Laboratory

One of three sites operated since 1995 by the UK Central Laboratory of the Research Councils. Located in Chilton, Oxfordshire.

In association with university science groups, the Rutherford Appleton Laboratory (RAL) space science department is involved in the design, construction, test and management of space instrumentation as well as reception, processing and distribution of space data.

RAL has a significant role in many space missions, including the Infrared Space Observatory (ISO); the ERS, Envisat and Meteosat Earth observation projects; the ROSAT and Spectrum-X high-energy astronomy missions; SOHO and Cluster which study the Sun–Earth connection; the Rosetta comet mission; and the Earth Observing System.

Specialist facilities include class-100 clean rooms, a full range of environmental test facilities, optical design and test facilities, a satellite control center, geophysical and solar–terrestrial data facilities and a molecular spectroscopy facility.

In addition to the space activities, RAL is actively involved in research into particle physics, high-power lasers, computing, cryogenics and microelectronics.

RAL's principal customers are the UK research councils, the British National Space Centre, the UK Meteorological Office, ESA, NASA, institutes of higher education and industry.

The astrophysics division undertakes research and projects in space science. Research covers astronomy and solar system science in the areas of space plasma physics, solar physics and planetary science. Project work involves space instruments in these areas and the provision of science support to programs undertaken in other divisions.

For further information see
http://www.ssd.rl.ac.uk/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ryle, Sir Martin (1918–84)

British radioastronomer, Nobel prizewinner (1974) for pioneering research in radio astrophysics, in particular for his invention of the aperture synthesis technique. During the Second World War worked on the development of radar and joined the Cavendish Laboratory, Cambridge, becoming professor there. In 1972 he was appointed Astronomer Royal, the first such appointment as an honorary post unconnected with the Royal Observatory. He invented radio interferometry and determined accurate positions and fluxes for radio sources. This led to the identification of many sources with galaxies that could be studied in visible light, revealing their nature. He compiled catalogs of radio sources that provided a body of statistical information about the numbers of sources at fainter and fainter fluxes (and therefore greater and greater distances) which proved that the universe had evolved, i.e. started from a Big Bang. This was the first of several near-fatal blows to the credibility of the steady state model, championed by his disputative Cambridge colleague FRED HOYLE, before the theory was finished off by the discovery of the microwave background radiation. Ryle invented aperture synthesis, by which an interferometer, rotating on the Earth, could fill in the aperture of a virtual large radio telescope. This created the method by which radioastronomers can produce unsurpassed angular resolution, studying phenomena close to the central black holes in radio galaxies, for example.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sabine, Sir Edward (1788–1883)

Physicist, astronomer, soldier and explorer, born in Dublin. He was the astronomer on two expeditions to the Arctic regions. He carried out experiments with pendulum clocks to measure the force of gravity at Spitzbergen and in Africa, in order to determine the shape of the Earth and studied terrestrial magnetism. Inspired by the work of HEINRICH SCHWABE on the 11 year cycle of sunspots, as reported by ALEXANDER VON HUMBOLDT, he correlated hiss measurements and his compilations of magnetic data and found a similar periodicity in magnetic storms, evidence of a solar influence on the magnetic field of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sadler, Donald H (1908–87)

Mathematician, became superintendent of the Nautical Almanac Office at the Royal Greenwich Observatory. His expertise was in numerical analysis, tabulation and printing, and he applied these skills to the production of the *Nautical Almanac*, and took pride in the accuracy of computation and reproduction.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sagan, Carl Edward (1934–96)

Planetary scientist, born in Brooklyn, NY, became professor at Cornell University and director of its Laboratory for Planetary Studies. He directed programs on theoretical, laboratory and spacecraft studies of organic molecules in the atmospheres and on the surfaces of planets, satellites and comets in the solar system, and in the interstellar medium. This was a means of understanding the origin of life on Earth and the possibility of the present or past existence of life elsewhere in the solar system. He was very influential in NASA, playing key roles in the Apollo program to the Moon, and the Mariner, Viking, Voyager and Galileo expeditions to the planets. He helped identify the runaway greenhouse effect on Venus which produces its high temperature, the dust storms which cause seasonal changes on Mars, and the complex organic molecules which cause the reddish haze of Titan and are the building blocks of proteins and nucleic acids. He inspired the continuing search for extraterrestrial intelligence, and his novel on the subject, *Contact*, was made into a successful film. He was a very successful popularizer of astronomy, with a book and TV series called *Cosmos*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sagitta

(the Arrow; abbrev. Sge, gen. Sagittae; area 80 sq. deg.) A northern constellation which lies between Vulpecula and Aquila, and culminates at midnight in mid-July. Its origin dates back at least to ancient Greece, where it was identified (by different authorities) with arrows belonging to Eros, Apollo and Hercules in Greek mythology. The brightest stars of Sagitta were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small but distinctive constellation, the brightest stars in Sagitta are γ Saggitae, magnitude 3.5, and δ Sagittae, magnitude 3.7. There are no other stars brighter than fourth magnitude. ζ Sagittae is a triple system with bluish-white (A1) and bluish components, magnitudes 5.6 and 9.0, separation 8.3″, the former of which has a very close (separation 0.2″) bluish-white (A3) companion, magnitude 6.0, which revolves around it in 22.8 years. Interesting variable stars include U Sagittae (range 6.45–9.3, period 3.38 days), an eclipsing binary, and WZ Sagittae (range 7.0–15.5), a recurrent nova which is normally about fourteenth magnitude but flared up to seventh to eighth magnitude in 1913, 1946 and 1978.

The Milky Way passes through Sagitta and the constellation contains a number of star clusters and planetary nebulae, the brightest of which is M71 (NGC 6838), an eighth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sagittarius

(the Archer; abbrev. Sgr, gen. Sagittarii; area 867 sq. deg.) A southern zodiacal constellation which lies between Ophiuchus and Capricornus, and culminates at midnight in early July. Its origin dates back to Sumerian times, when it was identified with Nergal, a god of war, but today it is associated with Crotus, son of the Greek god Pan and the inventor of archery, and it is shown on early celestial charts as centaur with a bow and arrow aimed at the heart of a scorpion (the neighboring constellation Scorpius). The brightest stars of Sagittarius were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A large, conspicuous constellation, the brightest stars in Sagittarius are ε (Kaus Australis), magnitude 1.8, σ Sagittarii (Nunki), magnitude 2.1, ζ Sagittarii (Ascella), a very close binary with white (A2) components, magnitudes 3.3 and 3.5 (combined magnitude 2.6), separation 0.2–0.6″, period 21.1 years, and δ Sagittarii (Kaus Media), magnitude 2.7. There are 12 other stars of magnitude 4.0 or brighter. Other interesting stars include β Sagittarii (Arkab), a wide, naked-eye optical double with bluish-white (B9) and pale yellow (F2) components, magnitudes 4.0 and 4.3, the former of which has a white (A5) companion, magnitude 7.2, separation 28.5″, and RY Sagittarii, an R Corona Borealis variable (range 5.8–14.0). Eight of the brighter stars form an asterism known as 'the Teapot': δ, ε, ζ and φ represent the body, σ and τ the handle, γ the tip of the spout, and λ the top of the domed lid.

The Milky Way passes through the western half of the constellation, which lies in the direction of the center of the Galaxy. It is brightest just to the north of γ Sagittarii, where it is known as the Great Sagittarius Star Cloud. Among the many star clusters and nebulae are the open clusters M25 (IC 4725), which contains more than 30 stars fainter than sixth magnitude (including the Cepheid variable U Sagittarii, range 6.28–7.15, period 6.75 days), M21 (NGC 6531), which contains about 70 stars fainter than seventh magnitude, and M23 (NGC 6494), which has about 150 stars fainter than ninth magnitude, the fifth-magnitude globular cluster M22 (NGC 6656), which is the third-brightest in the sky, and the sixth-magnitude emission nebulae M8 (NGC 6523, the Lagoon Nebula), M17 (NGC 6618, the Omega Nebula) and M20 (NGC 6514, the Trifid Nebula). M24, a bright portion of the Milky Way sometimes called the Small Sagittatius Star Cloud, contains the open cluster NGC 6603, which consists of several dozen stars fainter than fourteenth magnitude.

Also in Sagittarius are the Sagittarius Dwarf Galaxy, a dwarf spheroidal galaxy in the Local Group and the closest known satellite of the Milky Way Galaxy, lying about 80 000 light-years from its center, and the strong radio source Sagittarius A, part of which (Sagittarius A*) is believed to lie at the center of our Galaxy.

See also: Lagoon Nebula, Omega Nebula, Sagittarius A, Sagittarius Dwarf Galaxy, Trifid Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sagittarius A

The brightest radio source in the constellation of Sagittarius. Located at the center of the Milky Way Galaxy, at a distance of some 25 000 light-years, Sagittarius A (abbreviation, Sgr A) consists of two principal components, Sgr A East and Sgr A West. Sgr A East, which is a source of synchrotron radiation, appears to be a bubble of ionized gas, possibly a supernova remnant. Embedded within Sgr A West, which is a cloud of hot gas, is an intense, very compact, variable radio source called Sagittarius A* (pronounced 'A-star') that has an angular diameter of less than 0.002 arcsec. At the distance of the galactic center, this angular size corresponds to a linear diameter of less than 15 astronomical units (smaller than the diameter of the orbit of the planet Saturn).

Sgr A* appears to have no orbital motion and to be completely stationary. It is therefore believed to mark the exact center of the Galaxy. It lies at the center of a three-armed mini-spiral of hot, ionized gas, some 10 light-years in diameter. The orbital velocities of gas clouds close to the galactic center imply that some 6 million solar masses must lie within a light-year of the galactic center. Since infrared observations imply that stars contribute only about 3 million solar masses to that total, many astronomers contend that the residual 3 million solar masses is contained in a central black hole and that the underlying energy source for Sgr A* is a disk of hot gas swirling round this black hole. Sgr A*, then, appears to be similar in nature to an active galactic nucleus, though on a smaller and less energetic scale.

In 2001, NASA's Chandra X-ray Observatory, detected a sudden x-ray flare while observing Sgr A*. In a few minutes, the source brightened dramatically, eventually reaching a level 45 times brighter than before the flare. After about three hours, the x-ray intensity rapidly declined to the pre-flare level. The rapid rise and fall of the x-rays from this outburst are compelling evidence that the x-ray emission is coming from matter falling into a supermassive black hole. Given the extremely accurate position, it is highly unlikely that the flare is due to an unrelated contaminating source such as an x-ray binary system.

See also: active galactic nucleus, black hole, galactic nucleus, Milky Way Galaxy, radioastronomy, synchrotron radiation.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sagittarius Dwarf Galaxy

The Sagittarius DWARF GALAXY is the closest member of the Milky Way's entourage of satellite galaxies. Discovered by chance in 1994, its presence had previously been overlooked because it is largely hidden by the most crowded regions of our own Galaxy with which it is merging.

Position and morphology

Despite being located at a mere 16 kpc behind the GALACTIC BULGE (see figure 1), and 6 kpc below the Galactic disk, the determination of the morphology of this object has been complicated by its unfavorable position. For this reason, the inferred geometrical picture is somewhat sketchy, especially at Galactic latitudes $b > -10^\circ$ where a large fraction of the dwarf remains hidden behind the Galactic disk (and is revealed in those fields only by number counts of RR Lyrae variables). The Sagittarius dwarf galaxy appears to be a prolate body with axis ratios ∼3:1:1, with its longest dimension extending some 10 kpc in the direction perpendicular to the GALACTIC PLANE. It is assumed that the center of this dwarf galaxy is coincident with the region of highest stellar surface density, which is found at Galactic coordinates $\ell = 5^\circ$, $b = -14.5^\circ$, coincident with the massive GLOBULAR CLUSTER M54. At distances greater than 5 kpc from this center, a very low surface brightness stellar stream has been detected. Because of its irregular structure, it is difficult to estimate the total luminosity of the dwarf; an upper limit from integrating the surface brightness profile (given a simple model of its structure) is $5.8 \times 10^7 L_\odot$.

Its present proximity to the Galactic center must induce huge tidal stresses on the dwarf; this ongoing tidal disruption is expected to lead to its destruction, so its stars and globular clusters will eventually become dispersed and assimilated into the Milky Way halo. Modern GALAXY FORMATION simulations suggest that large galaxies like the Milky Way formed through the repeated ACCRETION of many small galaxies; the existence of the Sagittarius dwarf provides striking evidence that the Milky Way is still growing in the way predicted by such theories. Thus, this dwarf galaxy is perhaps a prime, and it is certainly the closest, example of a small galactic building block, as envisaged in this currently popular picture of hierarchical structure formation. One of the motivations for its study is that this may shed light on the hypothesized population of primordial galaxies that merged to form the Milky Way.

Stellar populations

This dwarf galaxy contains a mix of STELLAR POPULATIONS ranging from intermediate-age stars (given the presence of carbon stars) to old stars (some RR Lyraes are observed), with the dominant stellar population being 10–14 billion years old. It also contains four globular clusters (M54, Ter 7, Ter 8 and Arp 2) which also display an age spread, with Ter 8 being as old as the oldest Galactic halo clusters. The mean chemical abundance lies between [Fe/H] = −0.8 and [Fe/H] = −1.2, while the full abundance range observed covers $\gtrsim$1 dex around this mean.

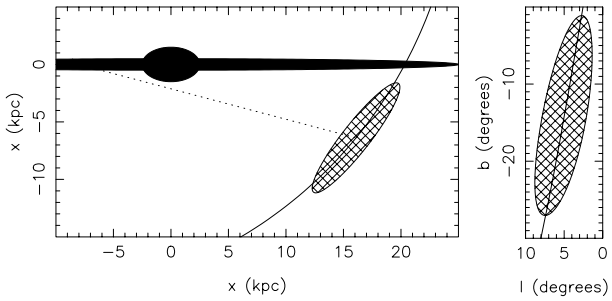


Figure 1. Schematic diagrams showing the position of the Sagittarius dwarf in relation to the Galaxy. In both diagrams, the orbital path of the dwarf is indicated with a solid curve. The left-hand panel is in the x–z plane of the Galaxy. The line of sight from the Sun to the center of the dwarf is marked with a dotted line. The right-hand panel shows the projection of Sagittarius dwarf on the sky.

Kinematics

The RADIAL VELOCITY of the central parts of the Sagittarius dwarf galaxy is $v_r = 171 \pm 1$ km s^{-1}, (after subtraction of the solar motion around the Galactic center). Normally, this would be almost all the kinematic information one could measure from such a system; however, as this galaxy is so close, it is also possible to deduce the transverse motion by measuring its PROPER MOTION. Several numerical studies have shown that dwarf satellite galaxies become elongated in the tidal field of their massive companion. The elongation, it transpires, always points parallel to the plane of the dwarf's orbit. Since the Sagittarius dwarf is almost directly behind the Galactic center, as viewed from Earth, the projected elongation must be aligned with the proper motion vector (and the projection of the orbit) to very good approximation. Given that the Sagittarius dwarf is elongated perpendicular to the Galactic plane, these considerations indicate that it must be on a nearly polar orbit. Proper motion measurements deduced from Schmidt plates show that it has a transverse motion (along the major axis direction) of $v_\perp = 250 \pm 90$ km s^{-1}, and that it is moving northwards.

Orbit and survival

With the above measurements of the three-dimensional position and three-dimensional velocity of the dwarf galaxy, it is possible to calculate its orbit around the Galaxy, assuming a model for the gravitational potential of the Milky Way. Taking a standard Galaxy model, with a double-exponential disk, and a spherical halo giving rise to an asymptotically flat rotation curve ($v_{\rm cir} = 220$ km s^{-1}), the period of the orbit is 0.7 Gyr, and the peri- and apogalactic distances are, respectively, ∼12 kpc and ∼60 kpc. This means that the Sagittarius dwarf galaxy is on a collision course with the disk of the Milky Way that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

will bring the two galaxies into contact within 25 million years.

Given the age of its dominant stellar component, this orbit implies that the Sagittarius dwarf has had many (more than 10) close interactions with the Milky Way during its lifetime. This turns out to be a very interesting constraint. Numerical N-body simulations show that the tidal stresses on the dwarf are so large each time it passes close to the Milky Way that it should have been destroyed after just a few ($\sim$2) close approaches to the Galaxy, and so it is very surprising that it has survived until the present day. A self-consistent solution to the present existence of the dwarf can be obtained if the observed stars do not trace the underlying mass. That is, the tidal disruption of the Sagittarius dwarf can be impeded if the stellar component of the dwarf galaxy is enveloped in a halo of DARK MATTER, which has a mass profile such that dark matter density is a factor of 2–3 larger than the mean Galactic density interior to its perigalacticon distance. To be consistent with the observed low velocity dispersion of the stellar component embedded therein, the core radius of the dark halo would have to extend out to the photometric edge of the system. The total mass of stars and dark matter is then $M > 5 \times 10^8 M_\odot$. With this model, the escape velocity from the center of the Sagittarius dwarf is substantial, $v_e = 90 \mathrm{km\,s^{-1}}$, which would help to explain the observed wide abundance range. The only alternative to this dark matter halo scenario is if the Sagittarius dwarf was deflected from a previously longer period orbit into its current short period orbit by a close encounter with another satellite, perhaps the LARGE MAGELLANIC CLOUD (LMC); however, the chances of such a 'bull's-eye' collision are small.

It is interesting to note that all of the numerical simulations of the disruption of this dwarf galaxy give rise to streams of tidally disrupted material that follow the orbital path of the main body of the dwarf quite closely. The disrupted fraction can be significant, so one may expect to find a sizeable fraction of the stellar component of the Sagittarius dwarf spheroidal, including perhaps several globular clusters, stretching along the locus of the orbit around the sky. It will be very fruitful to detect such material, as the kinematics of the stream will provide a very sensitive test of the Galactic potential gradients.

Interaction with the Milky Way

The models of the Sagittarius dwarf galaxy that are designed to reduce tidal disruption have sufficient mass to affect the structure of the Milky Way. Hydrodynamical calculations undertaken to simulate the collisional interaction between the Sagittarius dwarf and the Galactic outer H I disk show that a significant distortion of the Galactic H I disk will be induced by the collision if the mass of the dwarf exceeds $\sim 10^9 M_\odot$. Indeed, if $M \gtrsim 10^9 M_\odot$, the tidal forces on the Milky Way due to the Sagittarius dwarf will be substantially larger than those due to the LMC, which has also been invoked as a possible perturber of the H I disk. For model masses $\gtrsim 5 \times 10^9 M_\odot$, prominent spiral arms and a substantial lopsidedness in the outer disk are produced. Furthermore, a noticeable warp-like structure is induced in the disk. It seems that the Sagittarius dwarf may have significantly affected the star formation history and structure of the outer Galaxy.

Much interesting work remains to be done in understanding the chemical evolution of the dwarf (and the outer regions of the Galaxy) and how this is affected by the interaction (see also GALAXIES: INTERACTIONS AND MERGERS). From this, one may gain insights into the chemical evolution of galaxies during a merger event, and in turn about dominant evolutionary processes occurring during hierarchical formation of 'normal'-sized galaxy units.

Bibliography

The discovery and some subsequent studies that discuss the survivability and dark matter problem in the Sagittarius dwarf can be found in:

Ibata R, Gilmore G and Irwin M 1994 *Nature* **370** 194
Ibata R, Wyse R, Gilmore G, Irwin M and Suntzeff N 1997 *Astron. J.* **113** 634
Mateo M, Olszewski E and Morrison H 1998 *Astrophys. J.* **508** 55L
Zhao H-S 1998 *Astrophys. J.* **500** 149L

The following is a selection of the numerical attempts to model the interaction between the Sagittarius dwarf and the Milky Way:

Ibata R and Lewis G 1998 *Astrophys. J.* **500** 575
Ibata R and Razoumov A 1998 *Astron. Astrophys.* **336** 130
Johnston K V, Spergel D N and Hernquist L 1995 *Astrophys. J.* **451** 598
Velasquez H and White S 1995 *Mon. Not. R. Astron. Soc.* **275** L23

Rodrigo Ibata

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saha, Meghnad (1893–1956)

Physicist, born in Seoratali, near Dacca, India, became professor of physics at Calcutta, where he founded the Calcutta Institute of Nuclear Physics, now the Saha Institute. Saha (and Bose) translated several of Einstein's papers from German into English, leading to their greater dissemination, and he worked on the thermal ionization that occurs in the atmosphere of stars. He showed that elements in stars are ionized in relation to their temperature, as described by *Saha's equation*. This work fundamentally underpinned the interpretation of stellar spectra and led to the determination by HENRY RUSSELL of the abundances of the elements.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sakigake

Japan's first deep space mission. Launched in January 1985, it approached to within 7 million km of Halley's Comet on 11 March 1986 to study radio and plasma waves. Sakigake means 'pioneer'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sakurai's Object

Sakurai's Object (V4334 Sgr, RA = $17^h\,52^m\,32^s.69$ and Decl. = $-17^\circ\,41'\,07''.7$ at equinox 2000.0) is the central star of a faint PLANETARY NEBULA which experienced a late helium flash at, presumably, the end of 1994. It was discovered by amateur astronomer Yukio Sakurai in February 1996 as a rapidly brightening object, then at the 11th magnitude. It remained bright during two years, before dimming in mid-1998.

Sakurai's Object is the third 'final helium flash' star monitored so far. The first one was V605 Aql whose outburst and subsequent fading occurred during 1919–1923. Its characteristics were very similar to those of Sakurai's Object. The second one, FG SAGITTAE, had a longer evolution time scale. It presumably underwent its final helium flash at the beginning of the 19th century, reached its maximum visual brightness in mid-1970, and faded in 1992.

Final helium flash stars are post-asymptotic giant branch (post-AGB) stars on their way to the WHITE DWARF stage which experience a late thermal instability in their helium burning shell. The blue compact post-AGB star evolves back to a RED GIANT configuration on a time scale of a few years to a few decades. As a result, its effective temperature decreases from above 90 000 K to less than 7000 K, but its bolometric luminosity is expected to remain more or less constant (after a possible rapid initial increase at the onset of outburst). The rapid increase in the visual magnitude by several orders of magnitude then essentially results from the surface color change.

The current evolution of Sakurai's Object on its way back to the red giant configuration is accompanied by a change in its surface chemical composition. Spectroscopic monitoring during 1996 reveals an important decrease of its hydrogen abundance by a factor of ten, while the abundances of some other elements, such as lithium or some elements heavier than iron, have meanwhile increased. The exact internal mixing and NUCLEOSYNTHESIS scenarios responsible for those surface abundance variations are still a matter of debate among astrophysicists, given the complexity of the structural evolution of those stars. In this context, the observed properties of Sakurai's Object offer a unique opportunity to test our understanding of the physical and nucleosynthesis mechanisms at work in such stars. Final helium flash stars may furthermore provide important clues to understand the formation of R CORONAE BOREALIS (R CrB) STARS as well as of some white dwarf stars.

The final helium flash

The rapid evolution of the photometric and spectroscopic properties of Sakurai's Object is currently understood as resulting from the occurrence of a thermal instability in its interior. Such a thermal instability is a well established feature of AGB stars, the progenitors of final helium flash stars. It results from their double burning shell configuration, one of helium capped by one of hydrogen. Those burning shells are very thin, the hydrogen burning shell having about 0.001 solar mass ($M_\odot$) and the helium burning shell about 0.01 $M_\odot$.

During most of the AGB lifetime, the stellar luminosity is supplied by the hydrogen burning shell. The helium synthesized in that burning shell accumulates in the helium burning shell, whose temperature eventually increases until a thermal runaway is triggered. The instability, also called helium flash (not to be confused with the core helium flash which develops in the electron degenerate central regions of low-mass stars at the onset of helium burning), liberates one hundred to one million times the stellar luminosity on a time span of only several tens to hundreds of years. That energy, however, is mainly absorbed by the intershell layers, and affects the surface luminosity only to a factor of two. The intershell layers are thus subject to important thermal and structural readjustments during the after-flash phase. In particular, the convective envelope may penetrate through the now extinct hydrogen burning shell into the C-rich layers, and bring to the surface ashes of helium burning nucleosynthesis. CARBON STARS, for example, are currently understood as resulting from the operation of a series of such dredge-ups.

Those thermal instabilities (also called pulses) occur periodically in AGB stars on a time scale of 10^3 to 10^5 years, depending on the mass of the hydrogen-depleted core. When the star leaves the AGB, it evolves to the white dwarf stage on a time scale comparable to that of the interpulse phase. A 'final' helium flash may occur in that post-AGB stage if the star has not yet sufficiently cooled in the white dwarf configuration at the time the next instability is expected to occur. Current estimations predict the occurrence of such a final instability in about 10% of post-AGB stars.

The detailed structural and chemical evolution of post-AGB stars undergoing a final helium flash is not yet well understood. The current lack of model predictions for such stars is due to the difficulty of modeling adequately the chemical and energetic evolution of their intershell layers. In some cases indeed, such as presumably for Sakurai's Object, the final helium flash is triggered at a late stage of the post-AGB phase, when the star is already fading to the white dwarf stage. In those conditions, the intershell mixing triggered by the helium flash may cross the now almost extinct hydrogen burning shell and reach the envelope. Hydrogen is consequently mixed from the surface down to the high temperatures characterizing the helium burning shell, and burns. The observed depletion of hydrogen at the surface of Sakurai's Object may result from the operation of such a scenario.

The planetary nebula

The faint planetary nebula observed around Sakurai's Object supports the idea of it being a final helium flash star. The measurement of the size of the planetary nebula ($\sim 40''$) and of its expansion velocity ($\sim 31\,\mathrm{km\,s^{-1}}$) can lead to the determination of its age, or in other words of the time

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

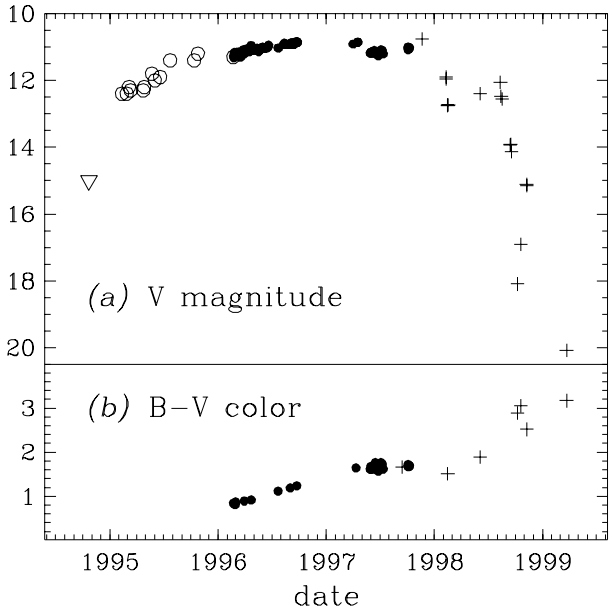


Figure 1. Evolution of the V magnitude of Sakurai's Object from 1995 to 1999. Open symbols refer to pre-discovery estimates from Takamizawa's photographic plates (http://www.kusastro.kyoto-u.ac.jp/vsnet), the triangle representing an upper limit. Filled circles are observations published in *Astron. J.* **114** 1657, and in *Astron. Lett.* **24** 248. Plus signs are measurements reported in several circulars of the International Astronomical Union.

elapsed since Sakurai's Object left the AGB, if we know the distance to the object. Based on an estimation of the absolute stellar magnitude from photoionization models of the nebula, a distance of five to eight kiloparsecs is found (see also next section). The planetary nebula would then have a radius of ~ 0.7 pc and an age of $\sim 24\,000$ years. This age is consistent with the time necessary for a post-AGB star to evolve from the tip of the AGB to the high surface temperatures ($> 50\,000$ K) required to ionize the nebula.

It is worth mentioning that the planetary nebula around Sakurai's Object was in fact searched for, and discovered, only after the object brightened, guided by the remarkable similarity of Sakurai's Object to V605 Aql, the central star of the planetary nebula A 58.

Luminosity and color variations

The light curve of Sakurai's Object in the V band is shown in figure 1(a). It has been followed in detail since Sakurai's discovery in 1996 (filled circles in the figure), and reveals a continuous brightening of the star by about half a magnitude over the following 18 months. Superimposed on that steady visual magnitude increase, small amplitude variations ($\lesssim$ 0.1 mag) are further detected on a time scale of few to 60 days (not visible in the figure). Those variabilities may be related to stellar pulsation modes triggered by the classical κ-mechanism acting on the partial ionization zones in the envelope of Sakurai's Object.

When was the outburst of Sakurai's Object initiated? The answer requires further information on the light curve of that object prior to Sakurai's discovery. Fortunately, photographic records taken by Takamizawa during 1995 enable one to extend the light curve in some detail up to one year prior to its discovery (open circles in figure 1(a). They confirm the general picture of the star's rapid brightening. Moreover, the analysis of photographic plates in 1994 puts a lower limit on its visual magnitude (upper limit on its brightness) of 15 mag (open triangle in the figure), and locates the start of the outburst at the end of 1994.

The rapid increase in visual magnitude after the outburst, by four stellar magnitudes over 500 days (figure 1(a)), is compatible with the picture of an expanding photosphere surrounding an object of constant luminosity. The radial expansion rate of the photosphere is estimated to be 0.05 solar radius ($R_\odot$) per day (i.e. 1500 km h^{-1}) at burst, and twice this value in 1996. The expansion of the compact post-AGB star to a giant configuration is accompanied by a concomitant cooling of its photosphere. This cooling is attested by the change in the B − V color index as shown in figure 1(b). Photometric and spectroscopic measurements lead to effective temperatures of 8000 and 7000 K at the beginning and end of 1996, respectively, to be compared with an estimated temperature higher than 90 000 K before burst.

A rapid increase of the luminosity may have occurred at the onset of the flash in 1995. The analysis of the remnant ionization of the planetary nebula around Sakurai's Object indeed suggests that the luminosity of the pre-flash central object, then close to a white dwarf's configuration, was of the order of one hundred times the solar luminosity ($L_\odot$). It must have rapidly reached 5000–10 000 $L_\odot$ after the outburst, and kept this value, to within a factor of three, during the subsequent expansion phase. The increase in the visual magnitude from 1996 to 1998 then mainly results from the change in color during that period.

Knowledge of the distance to Sakurai's Object would be an important ingredient in estimating its physical characteristics such as its absolute magnitude. Unfortunately, such information is difficult to derive from current observations. Several techniques have been applied to Sakurai's Object, such as those based on its PROPER MOTION in the Galaxy, on the degree of interstellar extinction suffered by the light during its travel from the star to us, on the comparison of its luminosity with that of V605 Aql, or on the comparison of the observed characteristics of the star and of its nebula with those predicted by model calculations. Most of them lead to a distance estimation between 5 and 8 kiloparsecs, the latter value putting Sakurai's Object in the bulge of our Galaxy. But even those values are still uncertain (a value of 1.1 kpc has even been derived from the extinction method), and illustrate the difficulty of deriving accurate individual distances in astronomy.

Comparison of the observed characteristics of Sakurai's Object with the few available final-helium flash model predictions suggests a stellar mass between 0.7 and 0.8 $M_\odot$, a luminosity between 5000 and 15 000 $L_\odot$, and a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

radius between 50 and 100 $R_\odot$. Those data, however, still need to be confirmed by adequate helium flash model predictions.

Sakurai's Object dimmed dramatically, though irregularly, from 1998 onward (plus signs in the figure), with M_v increasing above 20 in 1999. The rapid fading is very similar to that observed in V605 Aql at the beginning of this century. It is attributed to the formation of dust in an extended shell around the star. This is attested to by the detection, already from mid 1997, of the CN and C_2 bands in the spectrum of Sakurai's Object, and of a strong excess in the infrared emission. Photometry taken by the Infrared Space Observatory (ISO) in 1997 indeed showed an increase of the flux in the range 4–15 μm by a factor of about ten. The inferred mass loss from Sakurai's Object reaches $10^{-7} M_\odot$/yr. The extensive dust formation is very similar to that seen in R CrB stars.

Surface abundances

Most of the data currently available in the literature on the surface abundances of Sakurai's Object were obtained in 1996. Sakurai's Object is slightly metal-poor, with an iron mass fraction 0.2 to 1 dex below solar, and displays peculiar chemical composition compared to normal red giants. In general, the observed abundances confirm the double shell structure characteristic of AGB stars and the dredge-up scenario.

Sakurai's Object is first of all enriched in carbon, with [C/Fe] $\simeq$ 2.3 (the notation [X] stands for $\log X/X_\odot$, where $X_\odot$ is the solar abundance). The carbon enrichment can easily be explained by the operation of one or more dredge-up episodes when the star was on the AGB, the carbon being produced in the helium burning shell. Such a scenario is probably also the origin of the high oxygen ([O/Fe] $\simeq$ 1.5), sodium ([Na/Fe] $\simeq$ 1.4), neon ([Ne/Fe] $\simeq$ 2.4) and aluminum ([Al/Fe] $\simeq$ 1.1) abundances. Oxygen is produced directly in the helium burning shell, and sodium could be produced by successive dredge-up and hydrogen burning sequences. Reliable chemical abundance predictions, however, are still lacking and prevent a quantitative comparison with the abundances observed in Sakurai's Object.

Sakurai's Object also reveals an important nitrogen abundance enhancement ([N/Fe] $\simeq$ 2.2), which must have its origin in the CNO cycles of hydrogen burning during the AGB phase. In order to lead to significant nitrogen production, however, the synthesis of nitrogen must have occurred directly in the convective envelope, requiring temperatures at the base of the envelope exceeding 50×10^6 K. This scenario is further supported by the low $^{12}C/^{13}C$ isotopic abundance ratio of 1.5–5 measured in Sakurai's Object, which essentially points to the equilibrium value of the CNO cycle.

A similar constraint is provided by the measurement of high lithium abundances in Sakurai's Object. The most remarkable feature, however, is the rapid increase currently observed in the Li abundance, by 0.6 dex within four months in 1996. The scenario responsible for such an increase invokes the production of beryllium in the hot (temperatures above 50×10^6 K) bottom layers of the envelope and its convective transport to the surface, where it transforms to lithium by electron capture. Such a scenario requires thus a subtle coupling of mixing and nucleosynthesis, which must currently be at work in Sakurai's Object.

Rapid abundance alterations are also recorded for other elements such as Sc, Ti, Cr, Zn, Rb and Y. The abundances of those heavy elements have indeed significantly increased by 0.6 to 1 dex between April and October 1996. Those particular elements are synthesized by neutron capture reactions (the so-called s-process, which is also responsible for a similar abundance pattern of heavy elements in S and C stars). The production of the required neutrons in the helium burning shell, however, imposes specific constraints on stellar models which are still matter of research. In this respect too, Sakurai's Object provides a unique opportunity to test our understanding of those objects in particular, and of AGB stars in general.

Last but not least, the hydrogen abundance of Sakurai's Object was observed to decrease rapidly in 1996. In April of that year, it was already one hundred times lower than the hydrogen abundance in the Sun, and it decreased by another factor of ten during the next six months. The abundance pattern of Sakurai's Object is actually very similar to those of R CrB stars. Those stars are characterized by very low hydrogen abundances and by unpredictable visual dimming effects. Both of those features are now displayed by Sakurai's Object. We may thus be witnessing in Sakurai's Object the birth of a new R CrB star, providing a clue to the formation of at least a fraction of those stars.

Bibliography

Given the recent discovery of Sakurai's Object, further information should be sought in papers published in international astrophysical journals. The interested reader may look to the first papers on this object:

Duerbeck H W and Benetti S 1996 *Astrophys. J. Lett.* **468** L111

Duerbeck H W *et al* 1997 *Astron. J.* **114** 1657

More information on the planetary nebula and on the distance derivation to Sakurai's Object is presented by

Jacoby G H *et al* 1998 *Astron. J.* **116** 1367

The past of that object before its burst is discussed by

Kerber F *et al* 1999 *Astron. Astrophys.* **344** L79

Information on the on the chemical abundances is given by

Asplund M *et al* 1999 *Astron. Astrophys.* **343** 507

and on dust formation around Sakurai's Object by

Kerber F *et al* 1999 *Astron. Astrophys.* **350** L27

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Let us also mention a study of the relation between final helium flash stars and the formation of R CrB stars and hydrogen-deficient white dwarfs presented by

Iben I Jr 1984 *Astrophys. J.* **277** 333

Nami Mowlavi

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Salpeter, Edwin Ernest (1924–)

Nuclear astrophysicist, born in Austria, emigrated to Australia and was educated in Sydney, and at Birmingham University. Became professor at Cornell University. Worked with HANS BETHE on atomic physics, and in 1951 explained the 'triple alpha' reaction, also called the *Salpeter process*, which in red giant stars makes carbon-12 from three helium nuclei. This led him to investigate stellar evolution, and, from this theory and observations of stars of different luminosities, he worked back to the initial mass function, i.e. the rate of formation of stars of different mass in the Galaxy, now known as the *Salpeter function*. This important concept identifies a target for theories of star formation to explain.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sampex (Small, Anomalous and Magnetospheric Particle Explorer)

First of NASA's Small Explorer missions, launched 3 July 1992. Designed to investigate cosmic rays, solar energetic particles and magnetospheric electrons. (See EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

San Fernando Observatory

The San Fernando Observatory is located on the north side of the San Fernando Valley, approximately 30 miles northwest of downtown Los Angeles. It was built by the Aerospace Corporation in 1969 and donated to the California State University, Northridge, in 1976.

The observatory has a variety of modern instruments. The main instrument is a 61/28 cm aperture vacuum telescope feeding a vacuum spectroheliograph. At present, the spectroheliograph is configured as a Video Spectro-Spectraheliograph (VSSHG). After the VSSHG completes a scan of an active region, the output consists of two data cubes, each of which has solar north–south and east–west as the x- and y-axes and wavelength as the z-axis. The two cubes have orthogonal polarizations.

Other instruments are contained in the Little Dome. These are the full disk photometric telescopes: the Cartesian Full Disk Telescope no. 1 (CFDT1); Cartesian Full Disk Telescope no. 2 (CFDT2); and the Rotating Full Disk Photometer (RFDP). The CFDT telescopes obtain daily full-disk photometric images of the Sun.

A separate 10 m high tower contains a folded 15 cm refracting telescope designed to study individual active regions at the diffraction limit of 1 arcsec.

Research specialties are the measurement on a daily basis of solar variability due to sunspots and faculae and the study of the eruption and disappearance of solar magnetic flux, particularly in active regions.

Support comes primarily from the federal government, both from the National Science Foundation and from the National Aeronautics and Space Administration. The University provides some support for upkeep and maintenance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sandage, Allan Rex (1926–)

Astronomer, born in Iowa City, IA. Worked with WALTER BAADE and EDWIN HUBBLE, and became professor at Mount Wilson and Palomar observatories. His research has focused on the biggest problems in cosmology. He determined the ages of the stars in globular clusters (the oldest objects known, apart from the chemical elements). In 1960, with JESSE GREENSTEIN, he identified the radio source 3C48 as a blue star with an unusual, unidentifiable spectrum. After MAARTEN SCHMIDT'S breakthrough recognition of the nature of 3C273, this discovery by Sandage was recognized as the first quasar. Discovering many more quasars, Sandage worked on their properties. Realizing that the key question in cosmology was (and perhaps still is) the age and distance scale of the universe, he invested his career in the calibration of 'standard candles' (Cepheids, supernovae, ...), used to determine the distances of remote galaxies, most recently using the Hubble Space Telescope, for which he prepared an influential preparatory atlas of galaxies. This has led him to progressively more accurate values of the Hubble constant and the deceleration parameter of the universe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saros

The eclipse cycle of 18 years 11.3 days, after which the pattern of eclipses which occurred at the beginning of the cycle is quite closely repeated. The cycle was known in ancient Greek times and was used for eclipse prediction. The repetition is not perfect, however, as changes in the pattern increase with time. The cycle is almost exactly equal to the time taken for the nodes of the Moon's orbit to make a complete circle of the Earth. After one cycle, the Sun, Moon, and nodes return almost to the same relative positions and it is the alignment of these which determines the occurrence of eclipses. As the relative positions of the bodies are slightly changed after each saros, an eclipse cycle ends after a number of saroses—approximately 71 for solar eclipses and 48 for lunar. Each saros contains about 43 solar and 28 lunar eclipses.

See also: eclipse, epoch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sarpi, Paolo (1552–1623)

Born in Venice, and became a monk in the Servite order, rose to politically powerful positions in Venice. Came into conflict with the Pope in Rome over his defense of intellectual enquiry and religious freedom. Sarpi was a friend of GALILEO, discussing his work with him, and it was he who told Galileo about the reports from the Netherlands about telescopes, and brokered Galileo's offer of a telescope to the Venetian government in 1609.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Satellite (Natural)

In its most general sense, any celestial object in orbit around a similar larger object. Thus, for example, the Magellanic Clouds are satellite galaxies of our own Milky Way galaxy. Without qualification, the term is used to mean a body in orbit around a planet; an alternative term is *moon*. The term *natural satellite* distinguishes these bodies from *artificial satellites*—spacecraft placed in orbit around the Earth or another celestial body.

In the solar system, all the major planets except Mercury and Venus possess one or more satellites, and many of these are planetary bodies in their own right. Jupiter's largest satellite, Ganymede, is bigger than Mercury, and seven solar system satellites are larger than Pluto. Saturn's largest satellite, Titan, is the only one to possess a substantial atmosphere, with a surface pressure one and a half times the Earth's; other large satellites are known to have rarefied atmospheres. Many show signs of present or past geological activity, and Jupiter's Io is the most active body of all the planets and satellites in the solar system. The giant planets possess the lion's share of the major planets' complement of satellites: 60 of the 64 known for certain by 1999. Saturn and Uranus have the most, each with 18 confirmed (other minor satellites of Saturn and Uranus are suspected, and there are almost certainly minor satellites of the other giant planets awaiting discovery). A few asteroids, including (45) Eugenia and (243) Ida, have small satellites which have been directly imaged optically, while a number of others, including (4179) Toutatis, have been shown by techniques such as radar imaging to consist of two objects in close contact. The nucleus of Comet Hale–Bopp was found to have a small satellite in orbit around it.

Planetary satellites may be categorized as *regular* or *irregular*, according to their orbit. Regular satellites are those in direct orbits, in the direction of the planet's axial rotation, which are of low eccentricity and orbital inclination. Irregular satellites tend to be smaller, and follow orbits which are retrograde, or of high inclination, or show any combination of these properties. Whether satellites are regular or irregular also reflects their origin. Large regular satellites such as the Galilean satellites of Jupiter are believed to have condensed out of the same part of the solar nebula as their parent planet, from a *protosatellite disk*. The small and medium-sized regular satellites of Saturn are most likely collisional fragments of larger bodies. Uranus's satellite system is inferred to have been formed by the impact that is surmised to have given the planet its extreme axial tilt. Likewise, the Earth's Moon is believed to have accreted from the debris of a giant impact. Irregular satellites, such as Neptune's Triton, and the outermost satellites of the other three giant planets, were acquired by gravitational capture. Triton is probably a captured Kuiper Belt object; the others are probably captured asteroids or extinct comet nuclei. Jupiter's outer satellites fall into two groups, each representing a different capture mechanism.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn

Without its rings, Saturn looks like a pale JUPITER, with striking similarities to its giant sister (figure 1). As planetologists became more familiar with the structure of both planets, when they measured their thermal structure, composition and the zonal winds, this similarity appeared still more evident. However, important differences have been revealed as well. On the contrary, the other giants, URANUS and NEPTUNE, are more different, with atmospheric temperatures low enough to condense methane. It is therefore natural to list, as a starting point for the studies of Saturn's atmosphere, the similarities and differences between Saturn and Jupiter, each one telling us more about the formation of planets in the external solar system. As a model, Jupiter is better understood than Saturn, because of proximity and because it was visited by an atmospheric probe in 1995. As expected, the Jupiter results would benefit the Saturnian studies, as illustrated by examples given below.

Historical observations

The observations of the disk of Saturn at the beginning of telescopic observations, in the 17th century, were rendered difficult by the presence of the rings, which perturbed even simple observations of the shape of the planet (see SATURN: RINGS). The flattening of the planet was measured by William Herschel only in 1789, when the rings were observed on their edge, a configuration repeated only every 14 yr. The fainter contrast of atmospheric features, compared with Jupiter, precluded also the observation of any atmospheric structure before 1676, when Gian Domenico Cassini discovered the existence of an equatorial belt. The period of rotation was accurately measured later in 1790 by Herschel. Accurate telescopic observations on a regular basis have now permitted the constitution of a planetary database, used for the study of long-term evolution of atmospheric features, such as the global storms discussed below. However, only few individual atmospheric spots were observed before the space age. After the *Pioneer 11* observations in 1979, which gave important information about thermal structure and magnetospheric emissions, the *Voyager* observations in 1980–1981 provided the first accurate tracking of individual spots in the atmosphere, and the differential rotation of Saturn could be measured for the first time. The development of ultrasensitive charged coupled devices has since permitted a continuous monitoring of Saturn's long-lived individual spots with ground-based telescopes and the Hubble Space Telescope.

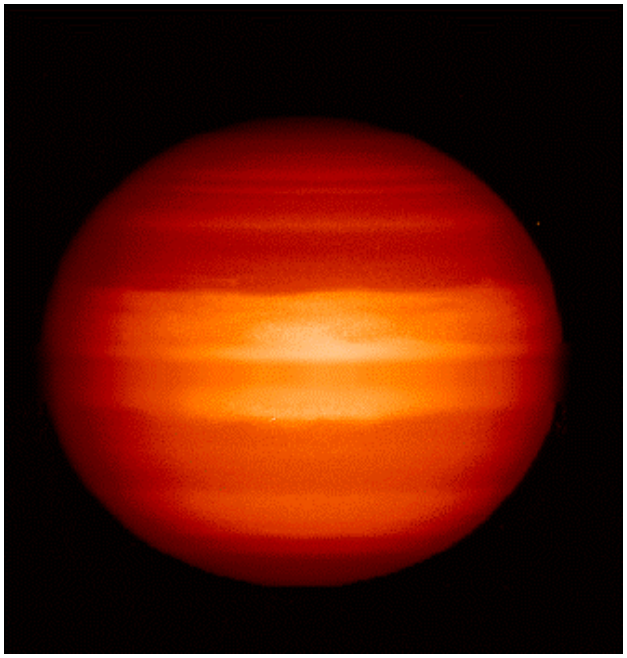


Figure 1. This unusual view of Saturn has been processed from an image taken by the HST Wide Field Planetary Camera (STScI-PRC96-16) on 17 November 1995. The rings have been artificially erased, and the equatorial region has been filled to eliminate all trace of the rings. Credit NASA.

Description

The physical and orbital parameters of Saturn are given in table 1, with comparison with Jupiter. These parameters give a first constraint to the models of formation of Saturn. Such models can predict the thermal evolution of a giant planet, given its initial mass and composition. In the list, the effective temperature is an important parameter, as it is related to the energy budget of the planet. It can be defined as the temperature of

Table 1. Jupiter and Saturn, the sister planets.

	Jupiter	Saturn
Revolution period	11.9 yr	29.6yr
Mass	1.899×10^{27} kg	0.568×10^{27}
Equatorial radius (at 1 bar)	71 492 km	60 268 km
Flattening	6.48%	9.80%
Rotation period (magnetic field)	9 h 50 mn 30 s	10 h 39 mn 22 s
Atmospheric He mass fraction (Y)	0.234[a]	0.06[b]
Mean density	1.33 g cm^{-3}	0.70 g cm^{-3}
Equilibrium temperature	109.5 K	82.4 K
Effective temperature	124.4 K	95.0 K
Emitted/received energy	1.67	1.78
Gravity at the equator (1 bar)	22.9 m s^{-2}	9.1 m s^{-2}
Obliquity	3°	27°

[a] from Galileo.
[b] from Voyager.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

a black body that would radiate on average the same amount of energy per surface unit as the planet itself. The effective temperature is higher than the equilibrium temperature, the temperature of a planet heated solely by the Sun. An important result of the Pioneer and Voyager missions was an accurate measurement of the radiative budget of Saturn and of the excess energy due to an internal source. In the case of Saturn, the internal heat magnitude is found to be comparable with the amount absorbed from the Sun.

The important fact that both Jupiter and Saturn emit more energy to space than received from the Sun has implications for their internal structure. The excess energy can come either from a residual contraction of the planet, transforming gravitational energy to heat, which is in turn radiated to the space, or from a compositional fractionation. In the latter case, some heavy elements, especially helium, fall to Saturn's core, again dissipating the gravitational energy. The last process cannot take place within the upper fluid atmosphere, where the gas or liquid phase of normal hydrogen is uniformly mixed with other gases. At high pressures (3–5 Mb) in the interior of Saturn, hydrogen changes to a metallic state, helium condenses to a liquid state and fractionates from the former.

The fact that, despite a lower mass, the diameter of Saturn is close to that of Jupiter is also understandable from the internal structure of the planet. The thermal history of Saturn is different from that of Jupiter, owing to the different mass and distance to the Sun. The difference in the abundance of helium between Saturn and Jupiter, as measured by Voyager, can be understood if the separation of helium took place earlier on Saturn than on Jupiter. Some recent measurements have questioned the accuracy of the helium abundance of Saturn. New estimates give a helium abundance closer to the original H/He of the solar nebula, although still lower than that on Jupiter. The question of the internal structure of the giant planets is therefore still largely uncertain, and further measurements are highly desirable to improve the models.

Planetary formation

The formation of a giant planet from the original solar nebula is described by modeling the accretion mechanisms in the first stages of the solar system. The scenario has been assembled from various constraints, in particular on the composition, internal structure and the shape. Many details remain to be refined, but a general agreement between various models is as follows.

The first step begins with the ACCRETION of a core of icy PLANETESIMALS and grains of dust up to about 10 Earth masses, with a composition thought to be similar to that of the KUIPER BELT objects, which form a belt of icy asteroids beyond the orbit of Neptune, or to some comets that are observed in the inner solar system. With a core of this size, the gas of the solar nebula itself begins to be accreted in the protoplanet, resulting in a fast growth of the planet until its final mass, reached when all the available material in the vicinity of the core has been accreted. The entire process could be completed in only a few million years. This scenario has been confirmed by the measurements of the heavy elements on Jupiter made by the GALILEO probe. Furthermore, the similarity between Saturn and Jupiter suggests a similar formation scenario for Saturn also. In the case of Jupiter, the measured composition gives a comparable enhancement for the elements C, S, N, Ar, Kr and Xe compared with the solar nebula. This result rules out older models that propose homogeneous accretion in the solar nebula, which would imply a global composition of the planets similar to the solar nebula. It also rules out models involving planetesimals which would have experienced a differentiation in their ice contents, the most volatile compounds, such as nitrogen, being lost during thermal processing of pristine ices. Once formed, the new planet slowly cools down, and some constituents, such as helium, begin to sediment when the temperature is cold enough. The timeline of the evolution of Saturn will be dependent on the convection in the interior and the heat release from the interior. A determination of its present state, in particular the He/H_2 ratio, will therefore help to constrain the models of evolution.

Interior

The interior of Saturn can be probed indirectly by measurements of the gravity field, the composition, thermal structure, heat flow and magnetic field. The models of the interior are based on the high-pressure thermodynamics of hydrogen–helium mixtures, which can be computed from first principles of physics, but unfortunately only with large uncertainties. Laboratory experiments are necessary to explore this poorly understood high-pressures physics. A *terra incognita* is the region of metallic hydrogen, a still hypothetical phase of high-pressure hydrogen where H_2 molecules can exchange their electrons becoming electrically conductive. Unfortunately, the experiments hardly reach the margin of this region. The interior of Saturn exhibits very large pressures, ranging between 0 and almost 50 Mbar (1 Mbar is one million Earth atmospheric pressures) at the center, and temperatures rising to about 10 000 K. Let us imagine a journey towards the interior of Saturn. Starting from the external layers, we encounter first the boundary between metallic hydrogen and molecular hydrogen envelope at about 2.8 Mbar, at about half the radius of Saturn. Is the change in density at the boundary large enough to generate giant waves, like at the surface of a monstrous ocean? Nobody knows. Deeper in the planet we find a boundary between an ice core, enriched in He, and the hydrogen-rich envelope at about three-quarters of the way into Saturn's interior.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Finally, a rock core is expected to be located at 0.88 of Saturn's radius towards the centre. The mass of this central core is estimated at between 0 and 10 Earth masses, and current models do not even rule out the possibility that it could be totally absent, the heavy elements being diluted in the other parts of the planet, although this case is not considered as the most probable.

The separation of helium from hydrogen takes place in the region where helium condenses to liquid and hydrogen turns to metallic hydrogen, according to the theoretical models of the equation of state of hydrogen. The upper atmosphere also becomes depleted in helium and relatively enriched in hydrogen, and the fall of helium droplets into the interior releases gravitational energy, contributing partly to the source of internal heat. On Jupiter the separation process started later than on Saturn, according to the model, because of the higher temperatures experienced at Jupiter. Effectively, the depletion of helium measured by Galileo is only marginally smaller than the solar nebula abundance. On Saturn, there has been a recent reevaluation of the He/H_2 ratio; new measurements seem to indicate less depletion in helium than the first Voyager measurements, listed in table 1, but they still show a substantial depletion.

Thermal structure

The thermal structure of Saturn is again very similar to that of Jupiter, with differences attributable to the larger distance to the Sun. Several techniques are used to remotely retrieve the thermal profile of Saturn. One is the use of a stellar or solar OCCULTATION, when a star (or the Sun) is eclipsed behind Saturn: the extinction curve of the light when observed through denser and denser atmospheric levels gives a measure of the vertical variation of density. From the density curve, one can retrieve the temperature with height, or pressure in the upper atmosphere. Another technique is the inversion of infrared spectra where a well-known constituent, such as CH_4, absorbs with spectral characteristics depending on the temperature. Combining these techniques, together with infrared and radio science data, it has been possible to probe the atmosphere in the pressure range from 1 bar to 0.001 mbar. An extrapolation of the profiles extends the thermal profiles to lower or higher atmospheric regions.

Like the atmosphere of Jupiter, Saturn's upper atmosphere can be separated into two domains, with opposite temperature gradients: the troposphere, where temperature is increasing downwards, and the stratosphere, with temperature increasing upwards (see PLANETARY ATMOSPHERES). The boundary between them is called the tropopause, a layer located between 60 and 100 mbar, at a temperature of 80 K (or –193 °C). The presence of a minimum temperature in the atmosphere at the tropopause is due to the absorption of solar energy by methane in the upper atmosphere, which heats the upper stratospheric layers.

The heat is dissipated mainly by two processes in a planetary atmosphere: convection transports heat through atmospheric turbulent motions and radiation transports heat by the thermal (mainly infrared) radiation through the atmosphere, if it is transparent enough. To the first approximation, convection in the troposphere and radiation in the stratosphere are the dominant mechanisms of heat transport. In the troposphere, convection maintains a constant gradient in the temperature, with a regular increase of about 0.8 K km^{-1} below the clouds. In the stratosphere, the atmosphere is stable against convection, since the upper layers are hotter than the lower, and heat is transported by radiation (or also by more complex dynamical processes like atmospheric waves, as discussed below).

The reality could be more complex than described here. In particular, the presence of a radiative layer, below the clouds, within the tropopause has been recently predicted: the presence of such a layer, where the atmosphere would be stable against convection in contradiction to the above description, as in the stratosphere, would have important implications for the transport of heat from the interior of Jupiter or Saturn to the observable layers. A better knowledge of the external composition, cloud structure and the radiative opacity of the atmosphere will be needed to complete such models.

The upper atmosphere of Saturn is dominated by other processes. A large temperature increase is observed in the upper atmosphere, at the microbar level and above, with the temperature reaching a constant value at high altitude, called the exospheric temperature. Two main mechanisms could be responsible for this heating, which is also observed in the stratosphere and thermosphere of other giant planets.

An explanation could be the deposition of energy by charged particles from the magnetosphere (see SATURN: MAGNETOSPHERE). In contrast to Jupiter, in the case of Saturn the deposited energy seems too low to explain the high exospheric temperature. It has been proposed that the ionosphere could also be traversed by high electrical currents whose dissipation would heat the atmosphere. An internal origin of the heating is also possible. Heating by internal waves, produced in the deep atmosphere and dissipating their energy in the upper atmospheric layers of the order of 1 μbar, has been proposed. This model is very attractive, as it also takes into account several measurements of the turbulence in the giant planets.

Composition

The general composition of Saturn is given in table 2, where the measurements have been taken from various ground-based and space observations, in particular from Voyager and ISO spacecraft. Even though compounds

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Composition of Saturn's atmosphere.

Molecule	Mole fraction relative to H_2
Hydrogen (H_2)	1
Helium (He)	0.034
Methane (CH_4)	4.5×10^{-3}
Ammonia[a] (NH_3)	$(1–6) \times 10^{-4}$
Water[a] (H_2O)	Still below the saturation vapor pressure at 3 bar
Phosphine[b] (PH_3)	4.5×10^{-6}
Monodeuterated methane (CH_3D)	3.25×10^{-7}
Carbon monoxide (CO)	1.65×10^{-9}
Arsine[b] (AsH_3)	35×10^{-9}
Germane[b] (GeH_4)	55×10^{-10}
Ethane (C_2H_6)	35×10^{-6} (stratosphere)
Acetylene (C_2H_2)	2.15×10^{-7} (stratosphere)
Carbon dioxide (CO_2)	3×10^{-10} (stratosphere)
Benzene (C_6H_6)	(Stratosphere)
Methyl radical (CH_3)	(Stratosphere)

For species vertically varying by [a] condensation or [b] photochemical destruction, the current estimate of the deep abundance has been listed.
From Atreya *et al* 1999 *Planet. Space Sci*. The abundance measurements are given with very large uncertainty, even for the major constituents (He, CH_4).

heavier than helium constitute less than 0.2% of the atmosphere of Saturn (by volume), they are important in the energy balance of radiation, because hydrogen and helium are poor radiative emittors. Methane, for example, controls mainly the thermal structure of the stratosphere, despite its low abundance. Other compounds, such as phosphine (PH_3) are found only at a few ppm but also have pronounced effects on the spectrum of Saturn. Access to the composition of Saturn's atmosphere is limited by remote sensing to the outer layers, at pressures lower than 10 bar. The physical processes affecting the composition are of two origins: thermochemical reactions in the troposphere and photochemical reactions in the stratosphere.

Thermochemical reactions

Thermochemical reactions equilibrate to produce an average composition. A typical example is the conversion of carbon compounds between CH_4 and CO, through the reaction

$$CH_4 + H_2O \rightleftharpoons CO + 3H_2.$$

At temperatures lower than 2000 K, CH_4 is the dominant carbon-containing species, but, as temperature increases, the production of CO becomes more important. At the low temperatures of the atmospheres of the outer planets, thermodynamics would predict the presence of CH_4 only. Nevertheless, CO was observed in the atmosphere of Saturn, in 1985. One possible explanation for this observation is the presence of an out-of-equilibrium phenomenon, where kinetics of reaction is taken into account and not only the final equilibrium. The conversion of CO to CH_4 is inhibited at the low temperatures of the upper troposphere, because the above reaction is slow. If CO can be transported rapidly enough to atmospheric regions where the reaction rate becomes insignificant, it will remain unaffected, because of a too-slow kinetics. This mechanism, called quenching, transports a compound to a region of low reactivity, even if thermodynamic equilibrium seems to be violated. Indeed, on Saturn, the observed tropospheric composition is far from equilibrium. Disequilibrium species, such as PH_3, GeH_4 and AsH_3, have been observed on Saturn. PH_3 is particularly dominant, because of its strong bands in the infrared. These compounds are tracers of the composition of heavy elements compared with hydrogen and helium. For example, the phosphorus elemental ratio is found to be 5–10 times greater than on the Sun (according to submillimeter and infrared measurements), implying a composition of Saturn's atmosphere highly enriched in heavy elements. Similar enhancements seen for several compounds favor an inhomogeneous accretion scenario for the formation of Saturn, as discussed above.

Photochemical reactions

The stratosphere is accessible to UV photons, and photochemistry becomes a dominant process, with a completely different chemistry, compared with most of the troposphere. Each molecule is sensitive to a specific range of UV wavelengths, which penetrate to different depths, and the photochemical reactions start at different levels, for different molecules such as NH_3, PH_3 and CH_4. The photochemistry of CH_4 is found at the uppermost levels in the atmosphere, with the production of heavier hydrocarbons such as C_2H_6, C_2H_4 and C_2H_2 initially. These compounds, in turn, absorb the solar UV, dissociate, react with other molecules and form more complex products. Heavier hydrocarbons can ultimately condense in the stratosphere, and they are thought to be part of the stratospheric hazes. At deeper levels, NH_3 and PH_3 compounds are also photolyzed, with a production of N_2H_4 or P_2H_4; such compounds could also be an important source of haze particles.

A recent and unexpected discovery by the ISO satellite has given new insight into the complex processes of the photochemistry of the upper layers in Saturn's atmosphere. ISO observations have shown the presence of CO_2 and H_2O in the stratosphere. Oxygen compounds coming from the interior are efficiently trapped at the tropopause, as condensation of H_2O, the main reservoir of oxygen at Saturn, prevents water vapor from reaching the stratosphere in significant amounts. ISO observations suggest that there is an influx of external oxygen, coming probably from dust particles of the interplanetary medium, falling into the atmosphere of Saturn (as well as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

into the other giant planets and Titan), or from the rings of Saturn. The presence of water and carbon dioxide strongly modifies the photochemistry, with different reactions with methane byproducts and the possibility of forming other constituents. From ISO observations, an average flux of 12 000 tons of material per day over Saturn has been estimated. The photochemical reactions between oxygen and carbon compounds can also become much richer than with pure methane in a hydrogen environment, with the possible production of many organic molecules, such as fomaldehyde (H_2CO).

Clouds

The giant planets exhibit many meteorological features and a large variety of cloud structures, apparent in telescopic observations (figure 2). The clouds appear when atmospheric gaseous compounds condense into liquids or solids at the cold temperatures found in the upper atmosphere (see CLOUDS IN PLANETARY ATMOSPHERES). As on the Earth, water clouds are expected in the troposphere, but the water clouds are hidden by upper cloud layers on Saturn. Above the water cloud, thermodynamic calculations predict the presence of two overlying layers, one cloud of ammonium hydrosulfide, NH_4SH, and another of ammonia ice, NH_3. The latter, an ammonia cirrus, is probably ubiquitous over Saturn's atmosphere, with an opacity varying with local meteorology. Individual clouds and spots seen in some images are probably the result of the lower less uniform cloud layers. The main difficulty in the interpretation of cloud structure is that the atmosphere is usually not in thermodynamic equilibrium, as we already know from the composition. To form a cloud, it is necessary not only to be at or below the condensation point but also to have nuclei of small particles to initiate the growth of cloud drops or crystals. A cloud is formed in a perpetual competition between small growing particles and large sedimenting particles.

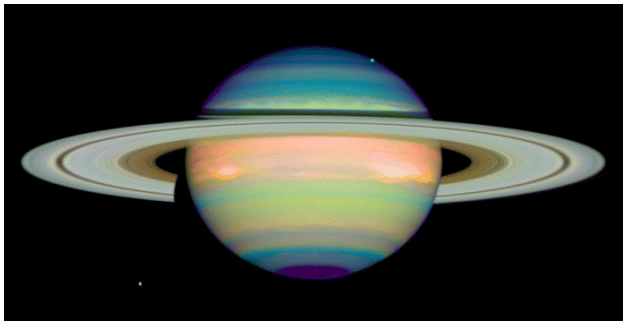


Figure 2. Saturn in false color from the NICMOS camera of the HST (STScI-PRC98-18). The false-color image shows indications of cloud layers. Green and yellow colors are sensitive to hazes above the cloud layers. Red and orange colors are sensitive to high clouds reaching the top of the atmosphere. Blue colors are an indication of a clear atmosphere down to the main cloud layer. Credit NASA.

The optical properties of the clouds largely depend on an extraordinarily complex microphysics, not always fully understood even in terrestrial water clouds. On Jupiter, the Galileo probe has shown that thermodynamically predicted clouds were probably present, but with different structure and altitude to what was thought before, because of local meteorological conditions. Above the main cloud layers, a haze of small particles is present on Saturn as well as on Jupiter.

Remote sensing observations of Jupiter and Saturn can measure with a reasonable accuracy the vertical distribution of small haze particles in the upper atmosphere, but the origin of the colors and morphology seen in the images remains mysterious. Minor constituents could possibly populate the clouds. These minor constituents (sulfur, phosphorus, etc) could be responsible for the coloration of Saturn's clouds. The structure of the clouds is also responsible for the different aspects of Saturn when seen at different wavelengths. Observations in a methane filter, where light is strongly absorbed by the atmosphere, are limited to the uppermost layers of the atmosphere. The regions of high clouds are therefore seen with better contrast. In the northern regions, an enhanced absorption is found in the UV. It is attributed to the presence of a polar haze. This haze could be formed by heavy hydrocarbons, generated by precipitating particles in the auroral regions, as on Jupiter, and condensing in the low temperatures of the lower stratosphere.

Photometric properties of clouds

The cloud structure has important effects on the radiative budget of Saturn. Variations in reflectivity govern the global absorption and reflection of the solar flux. Opacity variations also influence the reabsorption of the thermal flux inside the atmosphere, with pronounced consequences for the thermal structure.

To retrieve the atmospheric structure from photometric observations, one method is to measure center-to-limb variation: by measuring the variation of reflectivity along a parallel on Saturn's disk, vertical variations of the cloud structure are inferred. The scattering by small particles is observed at higher altitudes in the atmosphere close to the limb than at the central meridian, because of the more vertical path of the photons coming from the Sun at the center of the disk. The cloud structure of Saturn is found to be homogeneous in longitude, and variations occur mostly with latitude, as is apparent in the usual images in the visible, from ground-based or space missions (Voyager, HST).

Meteorology

As on Jupiter, the wind system on Saturn consists of a longitudinal banded structure, with strong westward winds at the equator, limited by eastward jets north and south. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturnian cloud bands correspond roughly, although not exactly, to a system of easterly–westerly jets. The equatorial wind speeds are as high as 500 m s^{-1}, 4 times larger than on Jupiter, and directed in the sense of rotation. There is still no general agreement about the origin of these winds. On Jupiter, the Galileo probe measurements have shown that strong winds persist down to pressure levels of at least 20 bar, below the levels where solar energy has been deposited. It has been proposed that the internal heat source is the most likely source of these deep winds. The banded structure, on the other hand, is believed to be caused by convective motions in the form of columnar cells, chaotically built and destroyed permanently. Even though no completely realistic models of the dynamics of giant planets have been constructed yet, laboratory measurements on convection in rotating spherical shells show some similarities to the banded structure seen on the giant planets. At first glance, these experiments tend to support models explaining the wind structure by global convection in the whole planetary interior (or at least the upper part of the atmosphere above the metallic hydrogen transition). Another class of models explain the appearance of the giant planets by dynamical forcing by solar energy deposition. Even if solar energy is deposited at the external boundary, over long time scales the entire atmosphere can be moved as well. A competition between internal energy release and solar heating is probably at work, but the hydrodynamical problems pose a challenge to our current knowledge. Without a global model of atmospheric motions, the meteorology of Saturn remains mostly confined to a more phenomenological description of atmospheric structure.

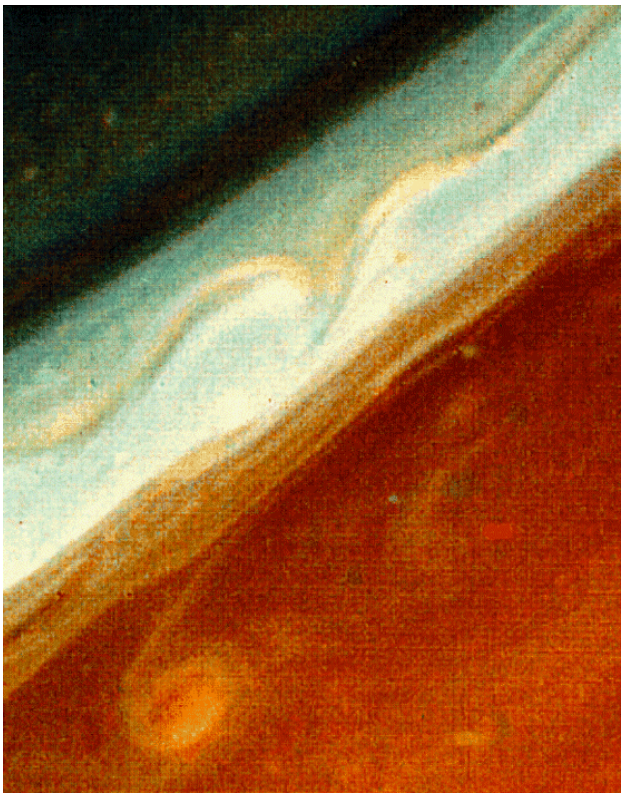


Figure 3. Spots on Saturn: image from Voyager 2 (Voyager 2, P-23919) in the northern temperate belt, showing an elongated atmospheric disturbance in Saturn's atmosphere on 20 August 1981. Credit NASA.

Spots

There is no feature on Saturn like the Great Red Spot on Jupiter, but several small ovals of various colors have been observed by Voyager (figure 3). At the time of Voyager observations, some symmetric, anticyclonic brown ovals (rotating clockwise in the northern hemisphere) were situated at 42°N, 5000 km × 3000 km in size. These spots have flows similar to the Jovian White Oval Spots. A unique structure present on Saturn during the Voyager epoch was a ribbon-like structure, located at 46°N, and described as a dark, wavy line moving at about 150 m s^{-1} eastward (figure 3). A so-called vortex street has also been observed. It represents a line of vortex with alternate direction of rotation, usually formed in turbulent flow in the wake of an obstacle. Generally, Saturn's atmosphere looks less active than Jupiter's. It must be noticed that the activity of the planet is observed through the cloud motions. Since clouds on Saturn are located deeper, owing to temperatures colder at corresponding pressure levels of Jupiter, they are more obscured by the overlying stratospheric hazes. Comparisons of the Jovian and Saturnian cloud activities should therefore be made cautiously.

Storms

One of the most striking differences in Saturn's meteorology compared with Jupiter is the periodic appearance of large storms on the former. Several events have been observed since 1876, showing an apparent periodicity related to the seasons of Saturn. Observations of such phenomena were made subsequently in 1903, 1933 and 1960. The last event was observed in 1990, and an image of Saturn soon after the beginning of the disturbance is shown in figure 4. The interpretation of these storms remains mysterious. They could correspond to a large convective cloud, originating deep in the atmosphere and redistributing a large amount of water to high altitudes. This cloud could therefore be compared to the equatorial plumes of Jupiter, which are also thought to correspond to water clouds, regularly spaced along the equator, and are present permanently. Another difference in the case of Saturn is the latitude of the storms, which varies from 5° to 58° (in 1960). If Saturn's storms are indeed convective water clouds, they would correspond to intermittent convection, with plume generation, as opposed to the more regular and permanent pattern of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

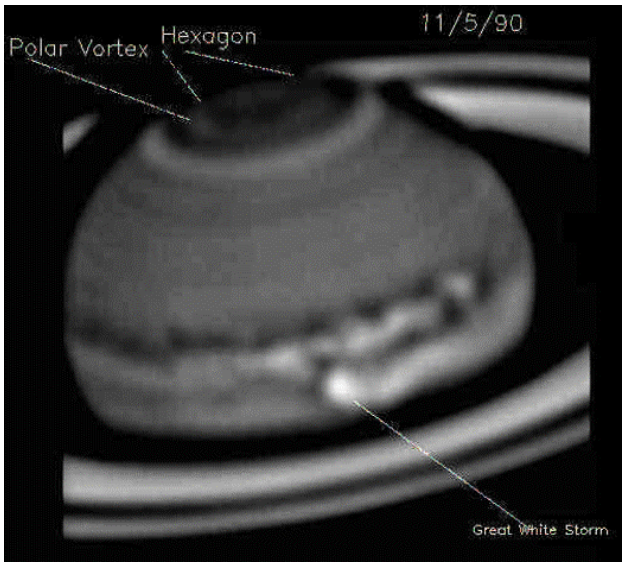


Figure 4. A giant storm on Saturn: image from Pic du Midi observatory, from October 1990, a few days after the emergence of a giant storm in equatorial regions. Several rare structures are seen in the northern regions also.

Jupiter. Such phenomena are indeed observed in the laboratory in experiments on convection. A difference in the thermal structure of Saturn, as compared with Jupiter, would result in differences in the convective activity. The observation of this storm was useful in completing the measurements of the wind field in the equatorial regions, which were obscured by the rings at time of the Voyager observations, but not in 1990 from ground-based observations.

Hot spots

One of the better-known features in Jupiter's atmosphere are the 5 μm hot spots. One of these is where the Galileo probe entered Jupiter's atmosphere. These spots are very bright in thermal emission at 5 μm. They represent regions of very low humidity, high transparency and low cloud coverage. On Saturn, there seem to be no features like Jupiter's hot spots. Water in the troposphere of Saturn has been measured from the ISO satellite, which found a very low water abundance. As for Jupiter, it has been attributed to some meteorological processes which decrease the water content of the atmosphere, as in Earth's deserts. The observation of an average low humidity despite the absence of hot spots seems therefore to argue for a different model of meteorology on Saturn than to that on Jupiter.

North polar regions

A special mention must be made about the north polar regions of Saturn. The Voyager 1 and 2 spacecraft recorded a strange hexagonal structure of clouds in the northern latitudes, and the presence of a particular spot, the North Polar Spot, named 'Big Bertha' by the Voyager imaging team. Since Voyager, this spot has been observed from the HST and from the ground, allowing planetary astronomers to follow its motion over a long period of time (figure 4).

Seasonal effects

Because of the 27° inclination of its axis, the north and south polar regions are alternatively in shadow during part of the orbital period of Saturn. Despite this long interval of 13 yr, no substantial effects are expected in the deep atmosphere, because the response time of the deep atmosphere to illumination conditions is much smaller than the orbital period, for pressures greater than several bars. The appearance of large storms seems to be periodic and could be related to Saturnian seasons, as mentioned above. Major seasonal effects are expected only in the stratosphere. The stratospheric thermal structure also depends on illumination effects and could exhibit seasonal variations. Indeed, the stratospheric temperatures measured by Voyager showed a difference in temperature at the 150 mbar level, the southern hemisphere being 6–8 K warmer than the northern, consistent with the modeling. Models of temperature variations predict a delay in the temperature response of the atmosphere subject to seasonal forcing, with maximum temperatures shifted by about one-sixth of a Saturnian year at 150 mbar. Finally, there is a possible influence of thermal effects on the underlying cloud structure, through the variation of the upper haze and the condensation of potential compounds of the stratospheric haze, such as N_2H_4 and P_2H_4, or the heavy hydrocarbons.

Aurorae

Although less intense than Jupiter's, Saturn's magnetic field also produces a large variety of magnetospheric effects. One of them affects the atmosphere. The observation of auroral emissions (see AURORA) in an atmosphere corresponds to the precipitation of charged particles from the magnetosphere, guided along the magnetic field lines and falling into the atmosphere. Saturn's strong magnetic field is, as for Jupiter, associated with auroral effects, mainly represented by the emissions of hydrogen molecules and ionospheric byproducts, in particular H_3^+. The emissions of hydrogen molecules and atoms can be seen only from space, as in figure 5 observed from the HST. These emissions are located along ovals, whose pattern is directly related to the geometry of the magnetic field. As on Jupiter, a complete oval is observed in the H_2 bands. Auroral emissions on Saturn have also been observed from the ground in the infrared through the emission of a molecular ion present in the ionosphere, H_3^+. The aurorae on Saturn correspond to a total deposition of energy estimated to be 10^{11} W, about 100–1000 times fainter than on Jupiter. Unlike Jupiter, where the internal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

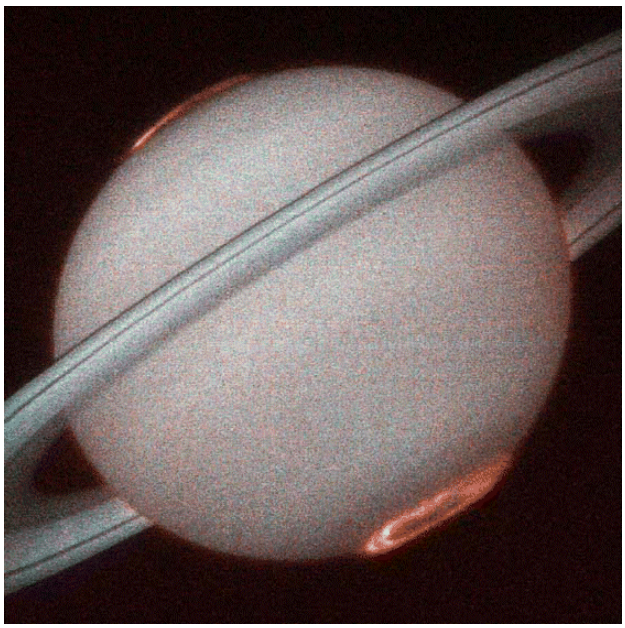


Figure 5. HST aurorae on Saturn (STScI-PRC98-05). Image of ultraviolet aurora, taken from the HST Imaging Spectrograph in October 1997. Credit NASA.

magnetosphere is populated by particles produced by the Galilean satellites, Saturn's aurorae are more dependent on the SOLAR WIND, which is 4 times fainter at Saturn because of the distance to the Sun. Saturn appears therefore as an intermediate case between Jupiter and the Earth for auroral processes.

Future observations

The Cassini spacecraft will enter the Saturnian system in 2004 and will provide extensive observations of Saturn, its moons (see SATURN: SATELLITES) and its rings. Although Titan is considered as the main objective of the Cassini–Huygens mission, the atmosphere of Saturn will also be extensively observed by remote sensing instruments during the mission. Several objectives are expected to be satisfied from this mission from the close study of composition variations and of the dynamics of the system. Unprecedented spatial resolution will be obtained on the cloud features, the cloud structure will be observed at angles inaccessible from the Earth and the night side of Saturn will reveal the thermal emission without contamination by the solar reflection which affects ground-based measurements. Many of the open questions mentioned above should find a solution after the Cassini observations, and many others will probably arise.

Web update

The discovery of twelve new moons of Saturn brings the planet's total to 30: the most in the Solar System. The new moons vary between about 6 and 30 km in diameter, and are moving in irregular, tilted orbits. They fall into several clusters, leading Brett Gladman, of the Observatoire de la Cote d'Azur, Nice, France, and colleagues to conclude that they are the remnants of larger satellites that have been fragmented by collisions. The moons probably began life as wandering bodies that were captured by Saturn's gravity. Gladman and colleagues found the moons by scanning Saturn with sophisticated electronic detectors mounted on medium-sized, ground-based telescopes.

Bibliography

Atreya S K 1986 *Atmospheres and Ionospheres of the Outer Planets and their Satellites* (Berlin: Spinger)

Gehrels (ed) 1984 *Saturn* (Tucson, AZ: University of Arizona Press)

Special issue on Galileo:
1996 *Science* **272** (10 May)

Special issue on Jupiter and Saturn (see example articles by Atreya *et al* and Encrenaz *et al*:
1999 *Planet. Space Sci.* **47** (16 December)

ISO results:
P Cox and M Kessler (ed) 1999 The Universe as Seen by ISO (ESA SP-427) (Nordwijk: ESTEC)

Web sites

Cassini mission: http://www.jpl.nasa.gov/cassini/
Galileo mission: http://www.jpl.nasa.gov/galileo/
Planetary Data System: http://www.pds.nasa.gov/pds_home.html

Pierre Drossart

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn Nebula (NGC 7009)

A planetary nebula in the constellation Aquarius, position RA $21^{h}\ 04.4^{m}$, dec. $-11^{\circ}\ 22'$. It measures $25''$ and shines with a greenish hue. It is of eighth magnitude and for a planetary nebula has a high surface brightness. Two small lobes on either side give it the appearance of the planet Saturn.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn: Magnetosphere

Perhaps the most awe-inspiring sight in our solar system, which can be appreciated by even a modest telescope, is that of the planet SATURN, with its banded atmosphere and delicate ring system. Saturn is the sixth planet away from the Sun at a distance of 9.5 AU. Saturn is the second-largest planet in our solar system after JUPITER with a radius of 60 330 km (defined as $1R_S$), and has a rapid rotation rate of 10 h and 39.4 min. This rapid rotation rate generates a centrifugal force that causes the equatorial regions of the planet to bulge outwards and the poles to flatten. Saturn is the least dense of all the planets and does not have a solid surface. Instead the planet consists of an enormous sphere of gas that gradually compresses to liquid deep beneath the clouds.

The rings of Saturn were first observed by Galileo in 1610, although because of the poor quality of his telescope he did not recognize them as such. He initially declared that Saturn consisted of three different planets, and then later he observed Saturn as first a single planet and then one that seemed to have developed side handles. It was only in 1655 that Huygens, with a telescope of far better quality, was able to propose that Saturn actually had rings in orbit around the planet. These rings we know today consist of countless individual particles that range from a fraction of a centimeter in radius to a few meters, with each particle having its own orbit around the planet.

The current count of satellites which orbit around Saturn is 18 and the biggest of these is Titan, which is unique in the solar system since it is the only moon to have its own atmosphere. Saturn has had three spacecraft fly past it, Pioneer 11 and Voyagers 1 and 2, the last taking place in 1981. The Cassini mission, which will spend 4 yr orbiting around Saturn, was successfully launched on 15 October 1997 and will reach Saturn in July 2004.

The magnetosphere is the volume around the planet that is dominated by Saturn's magnetic field. The magnetosphere of Saturn results from the interaction of the solar wind with the magnetic field of the planet, with the solar wind being deflected around the planet by the pressure of the planetary magnetic field.

This cavity is compressed on the Sunward side of the planet, under the influence of the pressure of the solar wind, and is drawn out into a long magnetotail on the nightside of the planet. Magnetospheric cavities form around any magnetized planet, for example, at the Earth and Jupiter (see MAGNETOSPHERE OF EARTH; JUPITER: MAGNETOSPHERE). Figure 1 shows a schematic of the Saturnian magnetosphere with the solar wind impinging onto the planetary magnetic field from the left. The shock wave that results, the bow shock, is shown by the dotted curve, and this results in the solar wind slowing down as well as being heated and deflected around the planet. The broken curve, the magnetopause, separates the flowing solar wind from Saturn's magnetic field. Saturn, like the Earth, has the magnetopause and bow shock at a standoff distance from the planet that scales with solar wind dynamic pressure. The region between the bow shock and magnetopause is known as the magnetosheath. The cusp region, as shown, allows magnetosheath particles entry into the magnetosphere since that is where the magnetic field is weakest. Full curves show the magnetic field lines of the planet. The cavity which forms the magnetosphere is filled with electrons, protons, ions, molecules, energetic particles and charged dust grains, with the charged particles and dust grains interacting with the electric and magnetic fields which are present. The charged particle motion generates currents within the magnetospheric plasma; the main current systems that exist are those due to the magnetopause, the equatorial ring current and a cross-tail current in the magnetotail.

The nightside of the planet shows a fully developed magnetotail that has been observed to be about $40R_S$ in diameter at $25R_S$ downtail. This magnetotail develops as a result of the solar wind flow past the magnetopause that aids in pulling the planetary field lines in an anti-Sunward direction. This magnetotail which may extend hundreds of R_S downstream acts as a reservoir for energy and plasma. Shown in figure 1 are magnetic reconnection processes which generate plasma bubbles or plasmoids. It has been postulated that these break off from the Saturn magnetotail and escape downtail just as occurs at the Earth. The outer magnetosphere (close to the magnetopause) and magnetotail are thought to be the primary regions where energy from the solar wind is transferred to the magnetospheric plasma.

TITAN, the largest moon of Saturn, orbits the planet at a radial distance of $20R_S$. Depending on solar wind conditions, Titan can be found in the magnetosphere (as shown in the figure), or in the magnetosheath or out in the solar wind (see SATURN: MAGNETOSPHERE INTERACTION WITH TITAN). During the previous planetary flybys of Saturn, a variety of solar wind conditions generated varying sizes of magnetospheric cavities, and this will be discussed later.

The magnetosphere

The magnetosphere is defined as being everywhere inside of the magnetopause, and comprises the magnetic and electric fields, plasma, energetic particles, the rings and some of the satellites.

Saturn's magnetic field

The behavior of the magnetosphere of Saturn is controlled by the magnetic field of the planet and hence in order to understand magnetospheric dynamics we need to understand the planetary magnetic field. The intrinsic magnetic field of Saturn was first measured by the PIONEER 11 spacecraft which flew past the planet in 1979. This spacecraft observed not only the intrinsic magnetic field but also the accompanying magnetosphere. The equatorial ring current was implied by the Pioneer 11 observations, and its existence was confirmed by the VOYAGER 1 and 2 flybys in 1980 and 1981.

The equatorial ring current is of the order of 10^7 A and flows eastward around the planet between radial distances

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

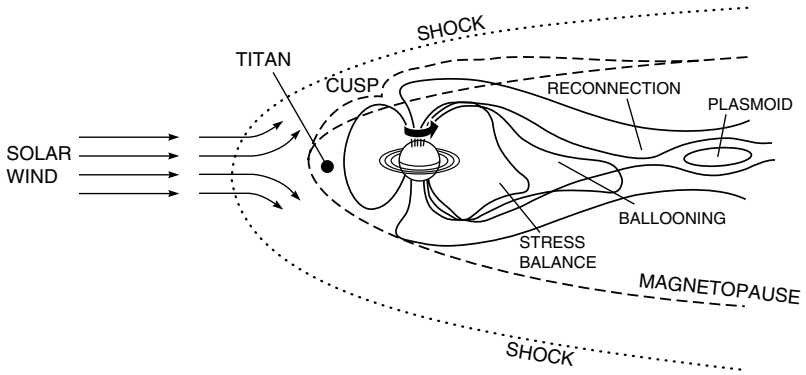


Figure 1. The magnetosphere of Saturn, with the solar wind flowing in from the left. The bow shock and magnetopause are shown, as are various processes within the magnetosphere. Titan is shown just inside the magnetopause. The large full arrow denotes the rotation direction of Saturn.

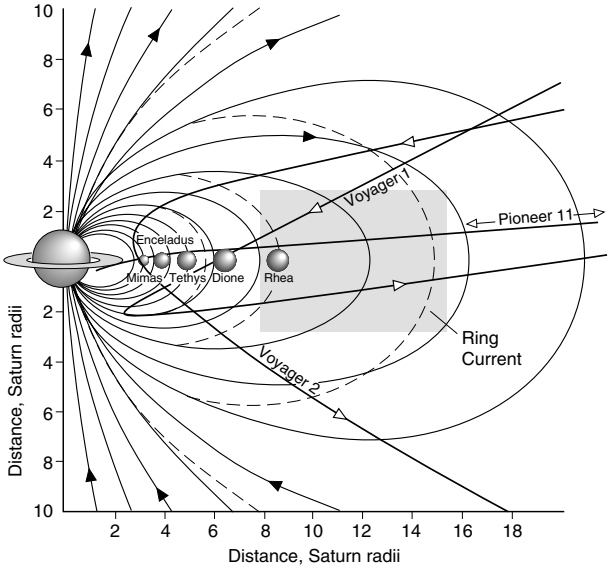


Figure 2. A meridian plane projection view of the magnetic field of Saturn (in a planet-centered cylindrical coordinate system). The broken curves represent dipole field lines and the full curves represent field lines that have been distended radially outwards as a result of the influence of the equatorial ring current. The shaded region shows the position of the ring current plasma. The flyby trajectories of the Pioneer 11 and Voyager 1 and 2 spacecraft are also shown. (Based on and courtesy of Connerney *et al* (1983) and Spilker (1997)). Figure 4 reveals a view of the flyby trajectories with respect to local time.

of $8R_S$ and $16R_S$. This ring current is generated by the drift of ions and electrons in opposite directions around the planet. Current flow in a plasma generates magnetic fields, and the ring current results in the dipole-like magnetic field lines of the planetary field being stretched outwards in a radial direction (although this radial distension is moderate when compared with the field lines at Jupiter). Figure 2 shows a view of the magnetic field at Saturn, with the broken curves denoting dipole field lines and the full curves containing not only the dipole field but also the contribution from the ring current as well. The shaded area shows the ring current and the dark full arrows show the direction of the magnetic field, with the open arrows revealing the direction of motion of the three spacecraft which have already visited Saturn.

On the nightside of the planet a well-developed magnetic tail occurs such as is found at the Earth and Jupiter. Effects of currents in the magnetotail as well as on the magnetopause need to be taken account of as far as their contribution to the magnetic field observed at Saturn is concerned. The planetary field as has been observed so far is primarily that of a dipole with a magnetic moment of $0.21R_S^3$ G. The polarity of the magnetic field is opposite to that of the Earth, that is positive from north to south, and therefore is in the same direction as that of Jupiter. The magnetometer observations from the Pioneer 11 and Voyager 1 and 2 spacecraft revealed quadrupole and octupole terms as well, although no non-axisymmetric terms could be resolved from the observations. Hence the magnetic field shows a remarkable symmetry about the rotation axis of the planet with there being a less than 1° tilt between the dipole and rotational axes. This is a unique feature in our solar system with all other magnetized planets having significant tilts between the dipole and rotational axes, for example the Earth has an 11.5° tilt and Jupiter has one of 9.6°. If, however, the internal field at Saturn has, like those of the other giant planets, been formed by dynamo action deep inside the interior of the planet, then one would expect non-axisymmetric coefficients to be present in the field (that is, an asymmetry in longitude around the planet). The circulation of conducting fluids deep in the planetary interior generates dynamo action, where the two physical processes of convection and planetary rotation in turn generate the interior magnetic field.

Not only does dynamo theory imply the existence of a non-axisymmetric magnetic field but there are also other observations that seem to confirm this: for example, rotational modulation of kilometric radiation near local

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

noon in the northern hemisphere, a periodic variation which occurs in optical spoke activity in one of the rings and auroral brightening at the poles at specific longitudes. All of these observations seem to imply that longitudinal asymmetries do in fact exist in the magnetic field close in to the planet although they have not yet been observed by spacecraft magnetometers. The most likely reason why such non-axisymmetric terms have not been observed is that they are very weak in comparison with the stronger axial terms, and the magnetometer onboard Cassini has been designed and built specifically with such measurements in mind.

Plasma

The behavior of the plasma in Saturn's magnetosphere can be separated into three distinct regions as are shown in figure 3. The left-hand side of the figure shows the magnetosphere at noon local time whereas the right-hand side reveals the behavior in the dawn magnetosphere. The inner plasma torus is characterized by the presence of dense low-energy ions and electrons at low temperatures (<10 eV, where an eV is equivalent to 11 605 K) and high densities of about 100 cm^{-3}. The inner region of the plasma torus, inside of $4R_S$ is thin, of the order of $0.4R_S$. Between $4R_S < r < 8R_S$, the plasma sheet thickens out to about $2R_S$ with a density of a few cm^{-3} and is oxygen rich owing to sputtering from the satellites Dione and Tethys. This outer region of the torus extends to just beyond Rhea's orbit.

In between the inner torus and the hot outer magnetosphere is an extended equatorial plasma sheet with enhanced levels of cold plasma. This is probably the primary contributor to the equatorial ring current. This plasma sheet extends out to about $15R_S$ and has a thickness of $8R_S$, density between 0.1 and 2 cm^{-3} and an electron temperature of 10–100 eV. The inner edge of the extended plasma sheet coincides with a vast cloud of neutral hydrogen that extends beyond $25R_S$, shown by the dot-filled circle. This neutral hydrogen probably escaped from Titan and the other satellites. At the inner edge there are very hot ions of the order of thousands of eV which are most probably new-born ions that have been heated by planetary rotation.

Beyond the extended plasma sheet lies the hot outer magnetosphere which consists of a highly variable region that coincides with the Titan-produced hydrogen–nitrogen torus. This torus maintains the density between about 0.01 and 0.1 cm^{-3} and the temperature between 50 and 800 eV (about 10^6 K). Near noon local time, cold density enhancements were observed by the Voyager 1 and 2 spacecraft which also correspond to drop-outs in the magnetic field. The Voyager 1 enhancements have been described as the spacecraft having crossed through plumes of Titan. Titan is a direct source of plasma, with the plasma streaming away from Titan in the local corotation direction and thereby generating a plume of plasma. These plumes tend to remain in the equatorial plane and will slowly over time merge with the background plasma. For the Voyager 2 flyby, Titan was not inside the magnetosphere, but was rather in the magnetosheath plasma and nowhere near the spacecraft itself. The cold density enhancements and magnetic field drop-outs which were observed have been described as plasma bubbles which have broken away from the edge of the extended plasma sheet. Similar types of plasma bubbles have also been observed at Jupiter and may be symptomatic of plasma transport processes in fast-rotating magnetospheres.

In all three regions of the magnetosphere the plasma particles tend to be concentrated in a disk in the equatorial plane where most of the plasma sources are located. Titan acts as a significant source of plasma in the Saturnian magnetosphere, as does the vast hydrogen torus, with the other satellites playing a role as well, the importance of which is yet unclear.

Energetic particles

Saturn like other planets in our solar system has populations of energetic particles, which are trapped by the planetary magnetic field. These trapped energetic particles, similar to the Van Allen radiation belts at the Earth, interact with the inner satellites and rings of Saturn. In the inner magnetosphere, inside of about $10R_S$, the ions tend to rigidly corotate with the planet, with the planet enforcing this through the magnetic field. The potential sources for these energetic particles are the solar wind, cosmic rays, ionized gas from the ionosphere of the planet and gas from the rings and satellites. The energization of these particles remains unclear and is one of the many questions that scientists with instruments onboard the Cassini spacecraft hope to answer.

The motion of charged particles in a magnetic field is described by adiabatic theory, and there are three distinct motions that take place. The particle gyrates around the field line as well as moving along it. There is also a slow drift of the energetic particles through the magnetic field, with the ions drifting one way and the electrons the other. The radiation belts are made up of the energetic particles moving in helical orbits along the magnetic field lines, with the particles being reflected or mirrored near the poles and therefore bouncing back and forth on the field lines.

The energetic particles in the inner magnetosphere are severely affected by the rings and inner satellites in a number of different ways. The energetic particles can be swept up by collisions with the satellites, the rings and satellites can act as sources of secondary particles as well as sources of gas and dust. The core of the radiation belts is made up of high-energy protons, above 1500 eV, inside of the orbits of Enceladus and Mimas. In the outer regions of the magnetosphere there is a population of low-energy electrons of tens of eV and at the magnetopause there is a large increase in the proton and electron population below 500 keV.

Rings

The rings of Saturn remain one of the great mysteries of our solar system and the more we find out about them the more questions seem to arise (see SATURN: RINGS).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

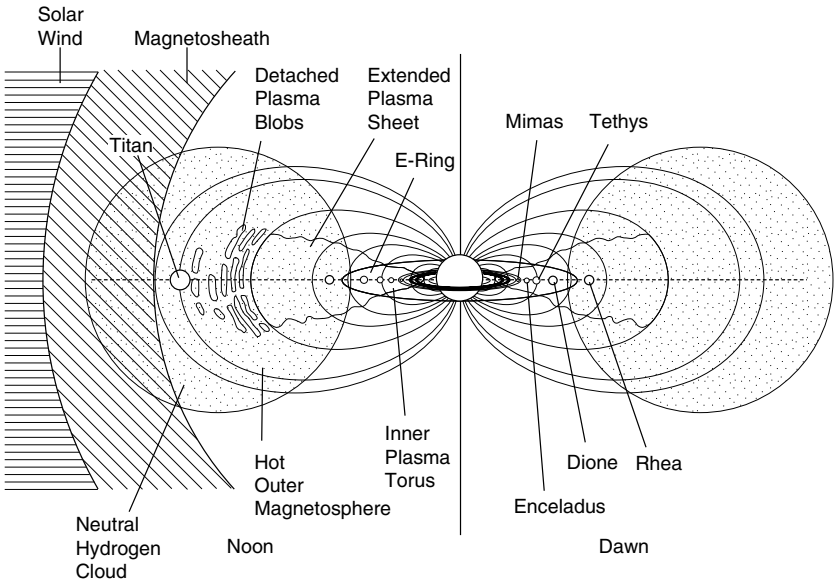


Figure 3. The different plasma regions of Saturn's magnetosphere, where the noon meridian is shown on the left of the figure and the dawn meridian on the right. The inner plasma torus, extended plasma sheet and hot outer magnetosphere are all shown with respect to the position of the rings and some of the inner satellites. The neutral hydrogen cloud is shown by the dotted circles, the magnetosheath as angled straight lines and the solar wind as straight lines. (Based on and courtesy of Sittler *et al* (1983)).

With Earth-based optical telescopes, one is able to differentiate distinct zones to the rings and with spacecraft flybys and Earth-orbiting telescopes further ring divisions have been observed. The outer zone, known as the A ring, is separated from a bright inner zone, the B ring, by a gap known as the Cassini division. Inside of the A and B rings is a semitransparent ring, the C or crepe ring. There are also a large number of smaller rings labeled through from the D through to the G ring, which were named in order of discovery and not in order of placement from the planet. The A and B rings are known to be composed of nearly pure water ice and dark, nearly radial lines or spokes have been observed in the B ring. There seems to be a link between the formation of these dark radial wedge-shaped features that appear on the ring and the rotational period of the magnetic field and Saturn kilometric radiation (SKR), and hence there must be an interaction between the electromagnetic fields of Saturn and the ring system.

Waves have been observed to propagate through the rings. The rings are made up of particles ranging in size from a few millimeters to several meters. These particles are bright and backscatter light rather strongly, behaving in a similar fashion to rough icy blocks. Moonlets have been observed within the ring system by the Voyager spacecraft, some of which seem to act as shepherds to the smaller rings. Some of the ring regions contain fine dust particles, which implies a link between the ring system and the magnetosphere of Saturn.

Even though we have been able to learn a great deal about this fascinating ring system there still remain a great deal of unanswered questions, some of which will be answered by the Cassini mission.

The icy satellites

There are at least 18 satellites orbiting about Saturn and there may well be other small yet undiscovered satellites as well (see SATURN: SATELLITES). All of these satellites are cratered to some extent and range from the biggest planet-like Titan with its own dense atmosphere to tiny irregular objects just tens of kilometers in diameter. The positions of some of the larger satellites are shown in figures 2 and 3. Mimas, which is at a distance of $3.08R_S$ from Saturn with a radius of 197 km, is heavily cratered and there is very little evidence for thermal evolution or resurfacing having taken place. Enceladus, at $3.95R_S$ from the planet, has a radius of 251 km and is a good reflector. It shows no large craters but there is some indication of large-scale surface changes having taken place which therefore implies some type of internal activity having taken place. Its orbit coincides with the densest part of the E ring and one of the questions which needs to be answered by Cassini is whether Enceladus could be the source of particles for the E ring.

Tethys, at $4.9R_S$ from Saturn with a radius of 530 km, also reflects well and is very densely cratered. Dione out at $6.3R_S$ from the planet has a radius of 560 km and a bright icy surface and there is evidence of past internal activity with numerous craters and valleys and troughs on the surface. Rhea is the outermost and largest of the inner icy satellites at $8.3R_S$ with a radius of 765 km and has well-formed craters on the surface. Before discussing Titan in some detail, Iapetus out at $59R_S$ from Saturn with a radius

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of 730 km deserves a mention. Iapetus has a great many craters on its surface and a density similar to the other icy satellites, which is consistent with water ice being the major constituent.

Titan

Titan, at $20R_S$ from Saturn and with a radius of 2575 km, is the second largest satellite in the solar system after Jupiter's moon Ganymede. It is unique in the sense that it is the only satellite in our solar system with its own atmosphere, which is dense and predominantly consists of nitrogen and methane. Until very recently the surface of Titan has remained hidden beneath an orange hydrocarbon haze, which is formed as a result of methane in the atmosphere being destroyed by sunlight. Recent images from the Hubble Space Telescope (HST), however, in near-infrared wavelengths (to which the haze is transparent) have revealed a surface feature on Titan of about 4000 km across (about the size of Australia).

An interesting issue for Titan concerns the relationship of the atmosphere, its surface and its interior. Its atmosphere is believed to be filled with complex organic molecules similar to the atmosphere of primordial Earth. Titan is the sole focus of the Huygens probe as well as a main target of the Cassini orbiter.

The Galileo orbits of Jupiter (see JUPITER: SATELLITES) have revealed that both Ganymede and Io have internal magnetic fields of their own. Titan has a size and density very similar to Ganymede and this raises the possibility that Titan and possibly some of the other Saturnian satellites may have internal fields of their own, either intrinsic like Ganymede's or remnant such as has been recently observed at Mars. The study of any induced or intrinsic fields associated with satellites is important for a variety of reasons. It allows a better understanding of the interior structure, composition and dynamics of the satellites, the way in which particles from the magnetosphere impinge on the satellite and how this affects the evolution and generation of magnetic fields in solid bodies in the solar system.

Saturn as a radio and wave source

For 20 yr before the Voyager missions reached Saturn, scientists had been searching for radio emissions at Saturn. There were no radio astronomy investigations onboard the Pioneer 11 spacecraft and it was only when Voyager 1 was within 3 AU of Saturn that radio emissions were first measured, thereby confirming Saturn as a radio source, although a much weaker source than Jupiter.

There are three distinct components to the radio spectrum, kilometric radiation, which we will discuss in some detail, together with narrowband noise and related continuum and electrostatic discharges. SKR appears from the auroral regions in both the northern and southern hemispheres. There are localized regions near noon local time that are evident with the most fundamental period observed occurring at the rotation period. This is one of the observations which tends to disprove the validity of the very symmetrical magnetic field which has been measured to date. The source energy of the SKR seems to be the solar wind, with a strong correlation being in evidence between the solar wind and SKR including a 25 day periodicity (which is the solar rotation period).

In a magnetized plasma such as a planetary magnetosphere, a whole range of different types of waves are produced which are trapped inside the magnetosphere. These radio and plasma wave emissions range from a single frequency (narrowband) through bursty and broadband (several frequencies) to continuous frequencies. The types of waves range from electromagnetic, sound and gravity waves (such as are supported in gases and fluids), to an entire range of waves unique to a magnetized plasma, such as Alfvén waves, magnetosonic waves and ion and electron cyclotron waves for example. The variety of waves trapped inside of Saturn's magnetosphere will be measured by the Cassini instruments.

Auroral processes

Saturn, like the Earth and Jupiter, has aurorae generated at its north and south poles. These aurorae are caused by charged particles depositing their energy in the upper atmosphere. The auroral emission is generated when the energetic particles interact with gas in the atmosphere; in the case of Saturn this gas is molecular and atomic hydrogen.

The Voyager spacecraft first detected the Saturn aurorae. Recent HST images of the aurorae have revealed both light and dark aurorae at the planet. The light aurorae have been observed at other magnetized planets but the dark auroral oval which was observed encircling the north magnetic pole is unique in the solar system. This seems to represent an excess of absorption of sunlight by the atmospheric particles that are a product of the auroral process.

Magnetospheric processes

Figure 4 shows a view of the Saturnian magnetosphere in the equatorial plane, with the Sun to the left of the figure, the nominal bow shock denoted by the full curve marked $\overline{S}$ and the nominal magnetopause again shown as a full curve and denoted by $\overline{MP}$. The magnetosphere is not usually static, however (as it is being shown in this figure), and for the three planetary flybys that are shown as broken curves a variety of solar wind conditions existed.

During the Pioneer 11 flyby in 1979, a fast solar wind stream inbound caused a bow shock to form upstream at $17.3R_S$ away from the planet. On the outbound pass during which time the solar wind must have been rather variable there were numerous magnetopause and bow shock crossings. Voyager 1 in 1980 observed a magnetosphere during relatively stable solar wind conditions with single bow shock crossings both inbound and outbound. The Voyager 2 flyby in 1981 occurred during extremely disturbed solar wind conditions, with initially an interplanetary shock wave being measured 3.5 days before the first bow shock encounter. An

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

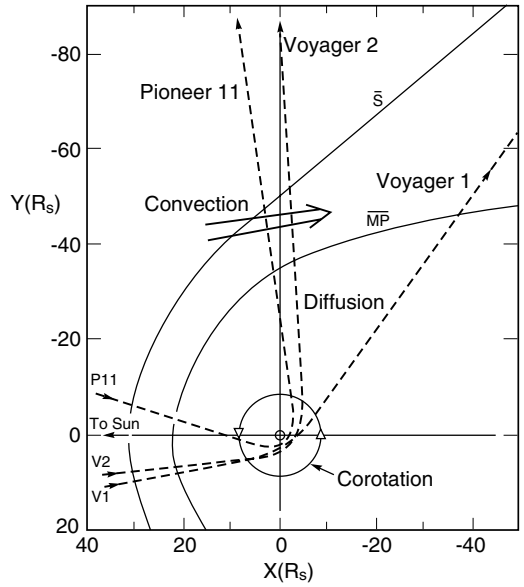


Figure 4. A view of the Saturnian magnetosphere in the equatorial plane, in a coordinate system which is Saturn centered with X positive in the direction from the planet to the Sun, Z positive along the rotation axis and Y completing the right-handed set. The nominal bow shock ($\bar{S}$) and magnetopause ($\overline{MP}$) positions are shown. The inbound and outbound trajectories of the Pioneer 11 and Voyager 1 and 2 spacecraft are shown by the broken curves, with the arrows showing direction of motion. The various regions of plasma circulation, that is corotation, diffusion and convection effects, are also shown. (Based on and courtesy of Sittler *et al* (1983)).

interplanetary current sheet was then observed some 12 h later, all of which resulted in five different bow shock crossings inbound (between $31.5R_S$ and $23.6R_S$). The magnetopause was observed at $18.5R_S$ inbound and then an expansion of the magnetosphere occurred with numerous magnetopause and bow shock crossings on the outbound pass; the final magnetopause crossing occurred at $70.4R_S$. Another explanation for this extremely disturbed behavior could be that Saturn in fact passed through an extended Jovian magnetotail during this flyby.

In planetary magnetospheres there are two main sources of energy which drive the physical processes which take place, those of the solar wind and the planetary rotation. This in turn results in two main types of large-scale plasma flow, those of convection and corotation. Convection describes the flow of plasma in a tailward direction, which results from the solar wind dragging magnetic field lines past the planet (as shown in figure 4). The term convection was first used for this process in the Earth's magnetosphere, where it was noticed that plasma flowed over the poles of the planet from noon towards midnight and then back again towards the dayside at lower latitudes. The type of flow pattern that results was very similar to flow cells that are thermally driven and hence the term convection has been used ever since. The fast rotation of Saturn and its magnetic field causes a large-scale electric field to form, which extends out into the magnetosphere. This in turn causes the magnetospheric plasma to rotate with the planet and the magnetic field, that is the plasma corotates with the planet. Now, at the Earth, solar wind processes dominate the magnetosphere, with the convection pattern usually penetrating halfway from the magnetopause to the Earth. At Jupiter, its very fast rotation rate ensures that the circulation of magnetospheric plasma is dominated by the planetary rotation. The Saturnian magnetosphere is somewhere in between the two cases, with both convection and corotation playing an important role. Corotation effects seem to dominate close to the planet, inside of Rhea's orbit out to about $9R_S$, as revealed in the figure. Convection processes seem to dominate much further out close to the bow shock and magnetopause, with the plasma in the outer magnetosphere being moved by large-scale convection. One of the goals of the Cassini mission is to better understand the role of these two plasma circulatory processes at Saturn. In addition to these rapid transport processes there is also the more uniform redistribution of plasma which occurs by means of diffusion within the magnetosphere itself.

The Cassini–Huygens mission

Cassini–Huygens is a joint NASA–ESA mission, which was successfully launched from Cape Canaveral, Florida, on 15 October 1997. Its aim is to study the Saturn–Titan system, with the Cassini orbiter orbiting Saturn for 4 yr beginning in July 2004 and the Huygens probe which will descend through the atmosphere of Titan in November 2004. This mission is named after two astronomers who pioneered observations of Saturn. The orbiter is named after Jean-Dominique Cassini who discovered the moons Iapetus, Rhea, Dione and Tethys, as well as ring features such as the Cassini division, during the period 1671–1684, and the probe is named after Christiaan Huygens who discovered Titan in 1655.

The general science targets for Cassini–Huygens can be divided into five main categories: that of Saturn itself, the rings, the icy satellites, Titan and the magnetosphere of Saturn. Although Titan is primarily the target of the Huygens probe, which will descend through the atmosphere onto its surface, it will also be studied by the Cassini orbiter during 40 or so flybys of the moon.

The Cassini–Huygens mission is described in detail in the article on the CASSINI MISSION as are the science objectives of the mission. This section will simply give a brief description of the magnetospheric objectives of the mission.

The science objectives for the magnetosphere of Saturn are to determine the configuration of the nearly axially symmetric magnetic field and its relationship to SKR. Current systems, the composition and sources and sinks of the charged particles within the magnetosphere will also be determined. Wave–particle interactions that take place within the magnetosphere will be investigated,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

as will the dynamics of the dayside magnetosphere and the magnetotail. The interaction of the magnetosphere with the solar wind, satellites and rings will also be studied.

A comparison will be made between the solar wind interaction that occurs at Saturn and those that occur at the Earth and Jupiter. At Earth, the magnetosphere is dominated by the solar wind whereas at Jupiter the magnetosphere is dominated by the rapid rotation rate and the internal plasma sources. Saturn's magnetosphere is likely to be somewhere in between these two extremes. It is more likely to mimic that of Jupiter since it too has a fast rotation rate and internal plasma sources, Titan and the icy satellites. Understanding the plasma dynamics at Saturn will allow a better understanding of the dynamics of planetary magnetospheres in general.

Bibliography

Connerney J E P, Acuna M H and Ness N F 1983 Currents in Saturn's magnetosphere *J. Geophys. Res.* **88** 8779–90

Gehrels T and Matthews M S (ed) 1984 *Saturn* (Tucson, AZ: University of Arizona Press)

Sittler E C Jr, Ogilvie K W and Scudder J D 1983 Survey of low-energy plasma electrons in Saturn's magnetosphere: Voyagers 1 and 2 *J. Geophys. Res.* **88** 8847–70

Spilker L J (ed) 1997 *Passage to a Ringed World (NASA SP-533)* (Washington, DC: NASA)

M K Dougherty

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn: Magnetosphere Interaction with Titan

Titan is the largest satellite of SATURN with an orbital distance from that planet of 20.6 Saturn radii (R_S), at which distance it will usually finds itself within Saturn's outer magnetosphere, where it interacts with the magnetospheric plasma. Titan is, after Jupiter's satellite Ganymede, the second largest satellite in our solar system and it possesses both an atmosphere and an ionosphere, which together act as an obstacle to the flow of Saturn's magnetospheric plasma that is incident on that satellite.

Introduction

Titan apparently does not have a significant intrinsic magnetic field. Consequently, the external plasma flowing nearby impinges directly on Titan's upper atmosphere and ionosphere. In contrast, at the Earth the solar wind plasma is deflected by Earth's strong intrinsic magnetic field. Titan is usually located inside Saturn's magnetosphere and the external plasma is the magnetospheric plasma, but occasionally solar wind conditions are such that Titan lies outside Saturn's magnetopause so that the external plasma encountering Titan is the solar wind.

Most of what we know about Titan and its interaction with Saturn's magnetosphere derives from data gathered by the VOYAGER 1 spacecraft when it encountered Titan on 12 November 1980. Titan was located within the Saturnian magnetosphere on this occasion. The radial distance of Voyager's closest approach to Titan's center was 6900 km, and the trajectory was downstream of the satellite with respect to the motion of the magnetospheric plasma. Titan's radius is 2575 km (R_T); hence, the radial distance of closest approach was $r = 2.67 R_T$. General information on Titan can be found in the main article TITAN.

The plasma instruments on Voyager 1 observed clear signatures of a plasma wake, in which the plasma densities were much greater than, and the temperatures much lower than, those in the ambient, external, magnetospheric plasma. The plasma wake is located downstream of Titan with respect to the external magnetospheric flow. The magnetometer on Voyager 1 observed a magnetotail coincident with the plasma wake. In this magnetotail, the measured magnetic field pattern was characteristic of field lines 'draped' around an obstacle (figure 1). Plasma wakes and magnetotails are also present at most planets and solar system bodies, and for objects such as Venus, Mars and comets, which lack significant intrinsic magnetic fields yet possess significant atmospheres. The tails are formed by a draped external magnetic field. Titan provides another example of the interaction of an externally flowing plasma with a non-magnetic body which has a significant atmosphere. A schematic diagram of the Titan interaction is shown in figure 2. In the rest of this article, a brief description of the atmosphere will be followed by a description of the ionosphere, and then the magnetospheric plasma interaction with the atmosphere and ionosphere will be revisited in more detail.

Titan's upper atmosphere

The atmospheric surface pressure of Titan is 1.45 bar and the surface temperature is a chilly 94 K. The dominant atmospheric species is molecular nitrogen (N_2), but there is a considerable abundance of methane (CH_4) with a mixing ratio in the range 2–10% in the lower atmosphere. Many other species are present at lower abundances, including ethane (C_2H_6), propane (C_3H_8), acetylene (C_2H_2) and hydrogen cyanide (HCN). In the upper atmosphere, molecular nitrogen is the most abundant neutral below an altitude of about 1700 km, but methane becomes the dominant neutral constituent above this altitude. At very high altitudes ($z > 2500$ km), molecular hydrogen (H_2) and atomic species such as N and H become more important.

A planetary exosphere is the uppermost region of a neutral atmosphere, in which the collision mean free path exceeds the scale size (or scale height) of the atmosphere (see also PLANETARY ATMOSPHERES). The exobase is the lower boundary of the exosphere, and for Titan the exobase is located at a radial distance of about $r \approx 4200$ km (or an altitude of $z \approx 1500$ km). In figure 2, the exosphere starts midway through the region labeled ionosphere. A considerable escape flux of H and N, and other species, is produced in Titan's exosphere due to the dissociation of CH_4 and N_2. The exosphere and ionosphere act as sources of ionized and neutral material for the outer magnetosphere of Saturn. The extended neutral exosphere of Titan strongly affects the interaction of Titan with the magnetosphere, giving this interaction a 'comet-like' character.

Titan's ionosphere

An ionosphere is an atmospheric region, or layer, consisting of charged particles (electrons and ions). Ionospheres form around any planet which possesses an atmosphere and are produced by the ionization of neutral atoms and molecules (see also PLANETARY IONOSPHERES). The ionosphere coexists with the neutral upper atmosphere. The main ionization sources in any planetary ionosphere, including the Earth's, are photoionization of neutrals by extreme ultraviolet (EUV) solar photons and electron-impact ionization of neutrals by fast electrons. The energy of the ionizing photon or ionizing electron must exceed the ionization potential of the relevant neutral (this energy is typically in the range 10–20 eV). The electrons are either photoelectrons produced by photoionization or energetic electrons which have traveled from the adjacent planetary magnetosphere. For example, the auroral displays observed at high latitudes on the Earth are produced by collisions of energetic electrons with oxygen and nitrogen atoms and molecules in the upper atmosphere. These electrons originate in the magnetotail region of EARTH'S MAGNETOSPHERE and follow magnetic field lines that extend from the magnetotail down to the polar ionosphere. At Titan, both the photoionization source and the electron impact ionization source contribute to the formation of the ionosphere. The ionizing electrons at Titan originate in the nearby magnetosphere of Saturn and follow the magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

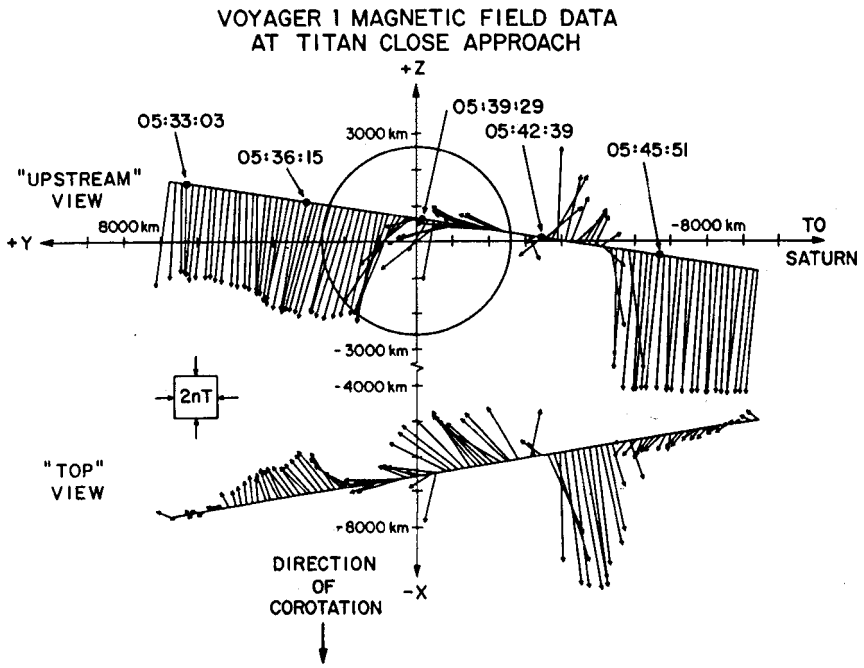


Figure 1. Projections of measured magnetic field vectors on two planes are shown along the Voyager 1 trajectory during its encounter with Titan's plasma wake. (From Ness *et al* (1982), © 1982 American Geophysical Union.)

field lines down into Titan's upper atmosphere. These electrons are hot, with thermal energies of about 200 eV. The relative proportion of the two ionization mechanisms (EUV or magnetospheric electrons) at a particular location on Titan and at a particular time is a complicated function of altitude, latitude and longitude, and orbital position of Titan.

Many different ion species are produced by ionization processes in Titan's upper atmosphere. The ion production rates are highest for the following species: N_2^+, N^+, CH_4^+ and CH_3^+. The overall ion production rate has a maximum at an altitude (above Titan's surface) of about 1100 km. This is the altitude where the optical depth for the incoming EUV radiation, or for the magnetospheric electron 'beam', is unity. Optical depth is the probability of absorption and is proportional to both the column density of absorbing atmospheric constituent and to the absorption cross section. In the main ionospheric region (altitudes between about 1000 km and 1700 km), the neutral density is high enough that ionization and chemical processes determine the distribution of ion species. The electron density at any location is equal to the sum of the ion densities of all ion species, because the plasma must remain charge neutral to an extremely high degree. N_2^+, CH_4^+ and CH_3^+ ions which are originally produced by ionization processes are quickly converted by ion–neutral chemical reactions to a variety of other ion species. These other species include hydrocarbon species as $C_2H_5^+$ and CH_5^+ and nitrile species such as $HCNH^+$. These molecular ion species then recombine with the thermal ionospheric electrons via dissociative recombination reactions. For example, $C_2H_5^+$ ions are produced from CH_3^+ ions via reaction with methane, and $C_2H_5^+$ in turn reacts with hydrogen cyanide (HCN) producing $HCNH^+$. $HCNH^+$ can further react with the neutral hydrocarbon C_4H_2, producing even heavier hydrocarbon ion species, or it can dissociatively recombine. Two of the more important chemical reactions can be represented by

$$C_2H_5^+ + HCN \longrightarrow HCNH^+ + C_2H_4$$

$$HCNH^+ + e \longrightarrow HCN + H.$$

The full ion chemistry scheme for Titan's ionosphere is quite complex and is not yet well understood.

The ionospheric electron density profile measured near the dawn terminator (i.e. the day–night boundary) by the Voyager spacecraft using the radio occultation technique had a peak density of 2400 cm^{-3} (with a possible uncertainty of 1100 cm^{-3}), and the peak was located at an altitude of 1180 km (with an uncertainty of 150 km). These values are in reasonable agreement with theoretical models which incorporate known ion sources and known chemical processes.

For altitudes above about 1700 km, where the neutral density is rather low, plasma transport associated with rapid flow becomes more important than chemical processes in controlling the distribution of ionospheric plasma. This flow is driven mainly by the interaction of the ionosphere with the external magnetospheric

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

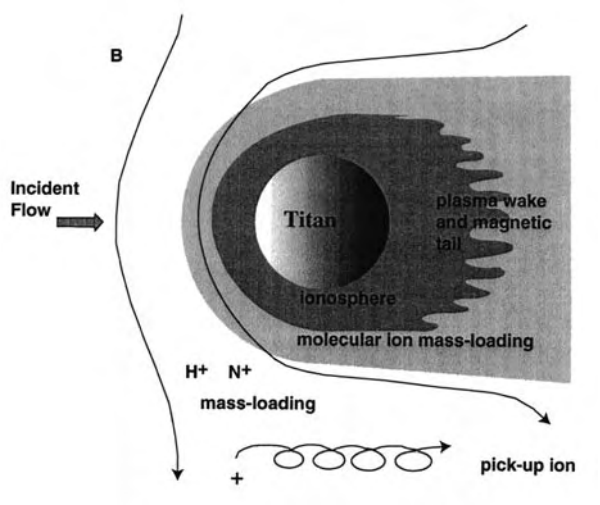


Figure 2. Schematic representation of the interaction of Titan's exosphere and ionosphere with the Saturnian magnetospheric plasma which is being convected from the left of the diagram. The external flow is mass-loaded with ions created by ionization of Titan's exospheric neutrals, such that the flow is slowed down and the magnetic field becomes draped around the satellite. (Adapted from Keller *et al* (1992).)

plasma. *In situ* measurements were made by the Voyager 1 spacecraft in the plasma wake region at a radial distance of about $2.7R_T$. The plasma seen by Voyager was much denser, colder and slower than the external magnetospheric plasma, yet was less dense, hotter and slower than the plasma deep in the ionosphere. The measured flow speeds were a few kilometers per second and the ion temperatures were such that typical thermal energies were a few eV. The total ion density was about 30 cm^{-3}. The major ion species had a mass of about 28 atomic mass units (making it either N_2^+, $HCNH^+$ or some similar species). Measurements of the upper hybrid resonance (a type of plasma wave) also indicated the existence of a cold (temperature, T_e, of a few eV) electron population with a maximum electron density of 39 cm^{-3}. The wake plasma appears to be an extension of the ionosphere and is probably transported outward from the main ionosphere.

Interaction of Titan with Saturn's magnetosphere

We now reconsider the interaction of Titan's atmosphere and ionosphere with Saturn's magnetosphere. The plasma conditions near Titan in Saturn's outer magnetosphere are quite variable (see SATURN: MAGNETOSPHERE). At the time of the Voyager 1 encounter, the magnetospheric plasma incident on Titan had H^+ and N^+ densities of $\approx$0.1 cm^{-3} and 0.2 cm^{-3} respectively, an N^+ temperature of about 3 keV (this is really a thermal energy and the corresponding temperature is about 3×10^7 K), and a flow speed of approximately 120 km s^{-1} relative to Titan (see table 1). This flow speed is less than the full co-rotation speed of about 200 km s^{-1}. The co-rotation speed is that speed, as a function of distance from Saturn, which is needed for the all the magnetospheric plasma to move around the planet with the same rotational frequency. The high temperature, the presence of N^+ and the flow speed less than the full co-rotation speed are all indications that a fair fraction of the plasma in Saturn's outer magnetosphere has come from Titan itself, no doubt from previous encounters. Plasma is apparently added to the magnetosphere of Saturn in the near vicinity of Titan from ionization of the exosphere, or from direct escape out the plasma wake, and also in a very extended region by ionization of the extensive neutral torus which exists throughout a significant part of the outer magnetosphere. The torus is formed from neutrals which have escaped from Titan.

Table 1. Properties of Saturn's magnetospheric plasma flow upstream of Titan determined by Voyager 1. (Adapted from Neubauer *et al* (1984).)

Magnetic field strength (B)	5 nT
Plasma flow speed	120 km s^{-1}
Proton number density	0.1 cm^{-3}
N^+ number density	0.2 cm^{-3}
Electron number density	0.3 cm^{-3}
Proton temperature	210 eV ($\approx 10^6$ K)†
N^+ temperature	2.9 keV
Electron temperature	200 eV
Sound speed	210 km s^{-1}
Alfvén speed	64 km s^{-1}
Magnetosonic/fast mode speed	220 km s^{-1}
Magnetosonic/fast mode Mach number	0.55
Alfvénic Mach number	1.9
Thermal proton gyroradius	413 km
Thermal N^+ gyroradius	5790 km ($\approx 2R_T$)

† Note that 1 electron volt (eV) is 1.602×10^{-19} J and 1 keV = 1000 eV. Also note that thermal energy $\approx kT$ where k is Boltzmann's constant and T is temperature.

The plasma flow incident on Titan at the time of the Voyager encounter (table 1) was subsonic ($M_s < 1$) and submagnetosonic ($M_f < 1$), yet super-Alfvénic ($M_A > 1$), which distinguishes this interaction from other examples of plasma interactions with non-magnetic bodies. For example, the solar wind is supersonic and supermagnetosonic. Jupiter's magnetospheric plasma impinges on its satellite Io as both a subsonic and a sub-Alfvénic flow (see IO: PLASMA TORUS, MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS). A Mach number is the ratio of a flow speed to a characteristic wave speed in a fluid medium. For a magnetized plasma the relevant wave modes are sound waves (propagation of pressure disturbances), Alfvén waves (propagation of magnetic disturbances) and magnetosonic waves (also called fast mode waves). Magnetosonic waves are the propagation of combined magnetic and pressure disturbances.

Putting aside for the moment the question of how Titan acts as an obstacle, one difference between Titan's interaction and the solar wind interaction with non-magnetic bodies such as Venus or comets is that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solar wind, unlike Saturn's plasma, is supersonic and supermagnetosonic. When a fluid approaches a body (that is, an obstacle) at subsonic speeds, sound waves are able to communicate the existence of the body out into the approaching fluid, so that the flow is able to adjust gradually and is able to smoothly flow around the body. However, when a supersonic flow (or supermagnetosonic flow for a plasma) encounters an obstacle, a shock wave forms upstream of the obstacle. A shock wave is a very thin layer across which the fluid properties change suddenly. The flow speed decreases and the density and temperature both increase as a fluid goes from upstream to downstream of a shock front. Bow shocks have been observed in the supersonic and super-Alfvénic solar wind flow upstream of all the planets in the solar system (except Pluto), as well as in the solar wind upstream of COMET HALLEY and COMET GIACOBINI–ZINNER. However, because Saturn's plasma encounters Titan at a submagnetosonic speed (at least at the time of the Voyager encounter), a bow shock is not expected in this case and, in fact, was not observed by any of the Voyager 1 plasma and field experiments. Numerical magnetohydrodynamic (MHD) models of this interaction confirm this feature of the magnetospheric interaction with Titan.

What is the nature of the obstacle that Titan presents to the external plasma flow? What determines the type of interaction? Why is Saturn's magnetic field draped around Titan? The atmosphere and ionosphere of any non-magnetic body present an obstacle to an external flow. The nature of this interaction depends on the characteristics of the object and on the nature of the external flow. The solar wind interactions with comets and Venus represent extreme cases of how this can happen (see SOLAR WIND: INTERACTION WITH COMETS, VENUS: INTERACTION WITH SOLAR WIND). The mass-loading process dominates the cometary interaction, whereas the ionospheric pressure force is the most important factor for Venus. Both seem to be important near Titan. First, consider the important mass-loading process.

Ions which are newly created in a plasma flow must somehow be assimilated into that flow. A charged particle moves in a circular trajectory in a uniform magnetic field if the background plasma is stationary. The radius of this circle is called the gyroradius. The gyroradius increases with increasing particle mass and with increasing particle speed. For example, the N^+ ions observed in Saturn's outer magnetosphere are fast enough that their gyroradii are roughly the size of Titan. H^+ ions near Titan have smaller gyroradii. When ions are created in a magnetized flowing plasma, their motion is a superposition of both gyromotion and drift motion in the direction of the background plasma flow. This is called ion pick-up. Associated with ion pick-up is a process called mass-loading, which appears to be operating near Titan, as well as near comets. The addition of mass to a fluid will slow down the fluid, in order for momentum to be conserved. The ion pick-up and mass-loading at Titan mainly consists of molecular ions close to that satellite and atomic ions further away (figure 2), reflecting the radial variation of the neutral composition.

The slowing down of a magnetized and flowing plasma, whether due to mass-loading or due to the other dynamical processes, affects the magnetic field in the plasma. Magnetic fields are 'frozen into' any fluid, such as most plasmas, which have high electrical conductivity. In this case, a piece of field line stays attached to a particular parcel of fluid. Hence, if the flow is slower in one region, such as near an obstacle or in a mass-loaded region, yet remains rapid elsewhere, the result is the draping of the magnetic field lines around the obstacle, with the associated formation of a magnetotail.

The interaction of the solar wind with comets is primarily determined by the mass-loading process. The nucleus of a comet is typically only a few kilometres in extent and is made of ice and dirt. The neutral gas (mainly water vapor) produced by the nucleus is not restrained by gravity and forms an atmosphere (or coma) that extends more than 10^5 km into interplanetary space. These neutrals are ionized by solar EUV radiation and the newly created ions are 'picked up' by the solar wind. On the other hand, the interaction of the solar wind with Venus takes place mainly near the planet. Mass-loading plays a relatively minor role in the Venus interaction. The ionosphere itself acts as the main obstacle to the solar wind flow. Ionospheric plasma is an excellent electrical conductor, and the solar wind induces electrical currents in the ionosphere that act to largely exclude the solar wind plasma with its imbedded interplanetary magnetic field. The magnetic field strength builds up as the flow approaches the ionosphere, resulting in the formation of a magnetic barrier located outside the ionopause. The ionopause is the boundary between the cold, dense ionospheric plasma and the hotter magnetized solar wind plasma. At the ionopause, the ionospheric thermal pressure balances the solar wind pressure, which mainly consists of magnetic pressure in the magnetic barrier. In fact, this pressure balance determines the location of the ionopause. As the solar wind pressure increases, the ionopause moves to lower altitudes where the ionospheric pressure is greater.

Ordinarily, for time periods when the solar wind dynamic pressure is significantly less than the maximum ionospheric thermal pressure, the ionosphere of Venus remains virtually field-free and the ionopause is a sharp boundary. However, at other times, when the external pressure exceeded the maximum ionospheric pressure, the ionosphere of Venus was observed by the magnetometer on board the Pioneer Venus Orbiter to be permeated by large-scale magnetic fields. In this case, the ionopause was much broader and less distinct, and was located at lower altitudes. The interaction of the solar wind with Mars seems to be similar to the Venus-like interaction for high solar wind dynamic pressure conditions, although mass-loading processes appear to make some contribution at Mars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

What about the interaction of the external plasma with Titan? The exosphere of Titan is rather extensive, resulting in some comet-like mass-loading, but there is also a substantial ionosphere, so that the Venus-like interaction scenario also applies to some extent. The Titan interaction appears to be a hybrid between a comet-type interaction and a Venus-type interaction. A magnetic barrier and a draped-field magnetotail are part of all these interaction scenarios. Will an ionopause boundary exist at Titan, just as it does at Venus? The total external plasma pressure upstream of Titan, including dynamic, thermal and magnetic pressure (which can be found from the parameters listed in table 1), is approximately $p_{\rm ext} \approx 10^{-9}$ dyn cm^{-2}. The maximum ionospheric pressure can be estimated to be $p_{\rm ionos} = 0.95 \times 10^{-9}$ dyn cm^{-2} at an altitude of 1150 km for daytime conditions. Hence the ionospheric and the external pressure are comparable, leading us to believe that, although the ionospheric pressure plays a very important role in the interaction, the external magnetic field will push its way deep into the ionosphere. The ionopause will be very broad, indistinct and located at a low altitude, and the ionosphere should be permeated with magnetic field. The proper analogy is Venus for high solar wind pressure conditions but with a significant amount of mass-loading also taking place.

Much remains to be learned about how Titan interacts with Saturn's magnetosphere. The next major advance in our understanding will come from the CASSINI MISSION, which was developed by NASA and ESA. The Cassini spacecraft was successfully launched in October 1997, and will arrive at Saturn in 2004. After insertion into orbit around Saturn, the Cassini Orbiter will encounter Titan many times.

Bibliography

Bird M K, Dutta-Roy R, Asmar S W and Rebold T A 1997 Possible detection of Titan's ionosphere from Voyager 1 radio occultation observations *Icarus* **130** 426

Hartle R C, Sittler E C Jr, Ogilvie K W, Scudder J D, Lazarus A J and Atreya S K 1982 Titan's ion exosphere observed from Voyager 1 *J. Geophys. Res.* **87** 1383

Keller C N, Cravens T E and Gan L 1992 A model of the ionosphere of Titan *J. Geophys. Res.* **97** 117

Ness N F, Acuna M H, Behannon K W and Neubauer F M 1982 The induced magnetosphere of Titan *J. Geophys. Res.* **87** 1369

Neubauer F M, Gurnett D A, Scudder J D and Hartle R F 1984 Titan's magnetospheric interaction *Saturn* ed T Gehrels and M S Matthews (Tucson, AZ: University of Arizona Press) p 760

T E Cravens

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn: Rings

Saturn's magnificent ring system is unique, with uncountable billions of icy particles in orbit around the planet. They represent a fundamental class of planetary structure. In this way, the studies of Saturn's rings may be crucial to one of the fundamental objectives of planetary science: to understand the origin and the evolution of the planetary systems. The aim of this article will be to emphasize general aspects of the structure of Saturn's rings.

The outer planet SATURN symbolizes all the planets. Its beautiful rings, discovered in the 17th century, give it a unique majesty, and whoever has seen Saturn through a moderately large telescope is not likely to forget the experience. As revealed by Pioneer and Voyager spacecraft and more recently by the Hubble Space Telescope, the rings are unbelievably dynamic and complex. The wealth of knowledge about Saturn's rings accumulated over 300 yr, with special attention paid to the cascade of data obtained by VOYAGER, is summarized according to the following steps.

The questions of the origin and of the evolutionary processes are directly correlated with the properties of individual ring particles. These properties include composition, nature of surface and particle size distribution; they will be discussed after the presentation of a brief history of the very exciting discovery of Saturn's rings.

The structural features, at whatever scale the rings were observed, posed great challenges to scientists to construct new dynamic theories. Some of these features have been explained from previous work in galactic and stellar dynamics. However, the rings remain extremely useful to astronomers, because some of the unexplained dynamic processes operating in the present ring may be similar to processes that occurred in the early ACCRETION DISKS. We will describe in a particular section the major dynamic phenomena which sculpt the rings.

The rings show many youthful features, raising the problem of their age and their origin. The problem is essentially this: calculated lifetimes for phenomena that are now occurring in the rings are considerably less than the age of the solar system. A review of the theories proposed to explain the formation of the rings will be presented. We will end with current and future developments.

History of their discovery

Considerable controversy resulted from the precise nature of the appendages of Saturn first observed by Galileo in 1610. Was Saturn grossly oblate? Were there two giant satellites? CHRISTIAAN HUYGENS revealed in 1659 that they were parts of a ring that surrounded Saturn whose diameter was more than twice Saturn's. The anagram being a common method of publishing among scientists in the seventeenth century, he announced his discovery as the following anagram: AAAAAAA CCCCC D EEEEE G H IIIIII LLLL MM NNNNNNNNN OOOO PP Q RR S TTTTT UUUUU, meaning *Annulo cingitur, tenui, plano, nusquam cohaerente, ad eclipticam inclinato* (He is surrounded by a thin flat ring, which does not touch him anywhere and is inclined to the ecliptic).

Huygen's speculation, that the ring was monolithic, became more controversial after JEAN DOMINIQUE CASSINI's observation of a gap in Huygen's single ring in 1675; this gap is now known as the *Cassini Division*. PIERRE SIMON DE LAPLACE began in 1787 dynamic studies on the ring system and concluded that the planet was surrounded by a large number of narrow solid rings. During the first part of the 19th century, Laplace's vision was generally accepted, and it inspired some observers to report faint additional divisions of the rings that could not be confirmed later.

In 1849, EDOUARD ROCHE determined the minimum distance at which a satellite influenced by its own gravitation and by its planet can exist; in the case of Saturn, if the planet and the satellites have equal density, the satellite cannot come closer than 2.44 times the radius of the planet without being broken up. He noted that this distance is 'à peine supérieure au rayon extérieur de l'anneau de Saturne' (a little farther than the external radius of Saturn's rings). This result was later known as the 'Roche limit' and began to discredit the idea of solid rings.

Finally, a series of dynamical studies followed, culminating in JAMES CLERK MAXWELL's classic essay of 1857, after which no theoretical doubt remained that Saturn's rings could only consist of an indefinite number of very small particles. Only much later (in 1895) did observations confirm the theory, when JAMES KEELER spectrographically measured the Doppler shift of the particles.

The discovery of new rings around Saturn (and around Jupiter, Uranus and Neptune (see PLANETARY RINGS)) has been a major finding of the past 30 yr, and has been the result of spacecraft- and ground-based observations. In 1967, Feibelman first reported photographic evidence of a wide outer ring (now called the E ring) encompassing the orbits of the satellites Mimas, Enceladus, Tethys and Dione (see SATURN: SATELLITES).

A narrow ring (called the F ring) lying a few thousand kilometers beyond the outer edge of the common broad Saturnian ring was detected by Pioneer 11 in 1979.

From the early 1970s, numerous astronomers attempted to identify some material within the main rings. However, only Voyager spacecraft were able to distinguish the presence of material between the C ring and the planet. This material came to be identified as the D ring.

Of all of the Saturnian rings given a letter designation, the latest and the least known is the G ring. Its presence was detected by Pioneer 11 through its absorption of several charged particles and visually confirmed by Voyager 2.

Overview of the remarkable rings

The rings fall into two main categories, depending on their optical depth τ, which measures the average number of particles that a light ray perpendicular to the ring would go through.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

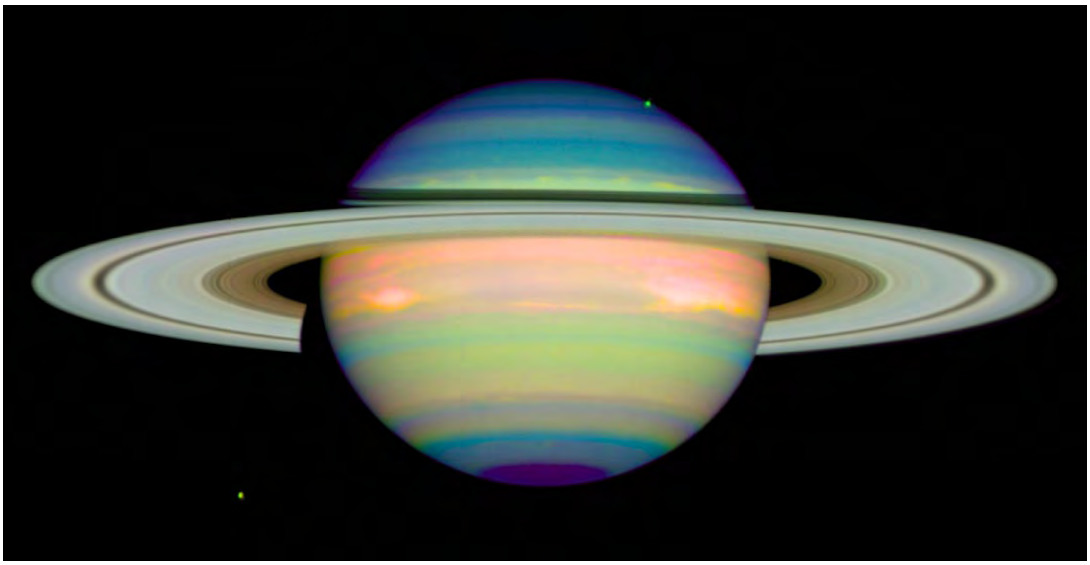


Figure 1. The main ring system. Saturn's main ring system taken by the Hubble Space Telescope. Moving radially out from Saturn, we see the general areas which are the C ring, the B ring, the Cassini Division and the A ring. The rings cast their shadow onto Saturn. The C ring and the Cassini Division are significantly darker than the A ring; the B ring is slightly brighter. The bright line seen within this shadow is sunlight shining through the Cassini Division. Two of Saturn's satellites were recorded, Dione on the lower left and Tethys on the upper right. Tethys is just ending its transit across the disk of Saturn. **This figure is reproduced as Color Plate 43.**

The main rings (C ring, B ring, Cassini Division and A ring), displayed in figure 1, are located within or near the Roche limit and have $0.05 < \tau < 2.5$. The total mass of these rings is approximately that of the satellite MIMAS, which has a radius of 200 km. All observations imply that the main rings' particles are primarily water ice, but there is evidence of some ALBEDO and therefore some compositional variation within the ring system. Most of the particles are in the 1 cm–5 m size range, but wave structures detected by Voyager spacecraft are strong indicators that small satellites with radii ∼10 km also exist. The motion of a particle is determined, to a first approximation, by the gravitational influence of the planet. To a much lesser extent, the gravity exerted by Saturn's satellites and the other ring particles and the collisions between particles disturb the orbit.

By contrast, the G and E rings (see figure 2) and the D ring are tenuous ($\tau < 10^{-4}$) and consist primarily of micrometer-sized particles. These rings are distinguished in two fundamental ways from the major Saturnian rings: single-particle dynamics (dominated by Saturn's gravity, solar radiation pressure and electromagnetic forces), rather than collective effects, probably govern their form and the majority of particles have quite limited lifetimes ($<10^4$ yr). The smaller particles, less than 1 cm in radius, would indeed be quickly removed from the Saturnian environment by falling onto the planet or ejected outwards. These particles must be replenished by some source within the rings.

The F ring ($\tau \sim 0.1$) is intermediate in properties between the main rings and tenuous rings. It is a narrow, inclined, multistranded ring lying 3400 km beyond the edge of the main ring system and shows a remarkable degree of longitudinal variation, including clumps, kinks and braided structures (see figure 3). Data on Saturn's rings are given in table 1.

Rings: a myriad of small satellites

Saturn's rings are composed of a myriad of individual particles. However, even in the post-Voyager era, no individual particle has been seen yet. The basic properties of individual particles can be constrained by reflected sunlight, thermal emission and radar reflectivity. Specifically, ground-based and spatial observations from ultraviolet (10^{-5} cm) to radio (10 cm) wavelengths contain crucial information on the ring properties. Because the rings comprise an ensemble of particles acting collectively, radiative transfer relates the ring properties to some intrinsic particle properties such as the composition, the size and the physical nature.

Composition of the particles

Near-infrared spectral data (1–2.5 μm) of the main rings were used to establish that water ice is a major component of the ring material. The decrease in reflectivity from the visible to the ultraviolet indicates that something other than pure water ice (probably silicates) is present, the non-icy component making up less than about 10% of the rings.

The tenuous rings are spectrally resolved with difficulty. However, recent observations during the 1995 Earth ring plane crossings show that the spectra of the E and F rings are clearly bluish, different from the main rings and from most of the solar system objects, which are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

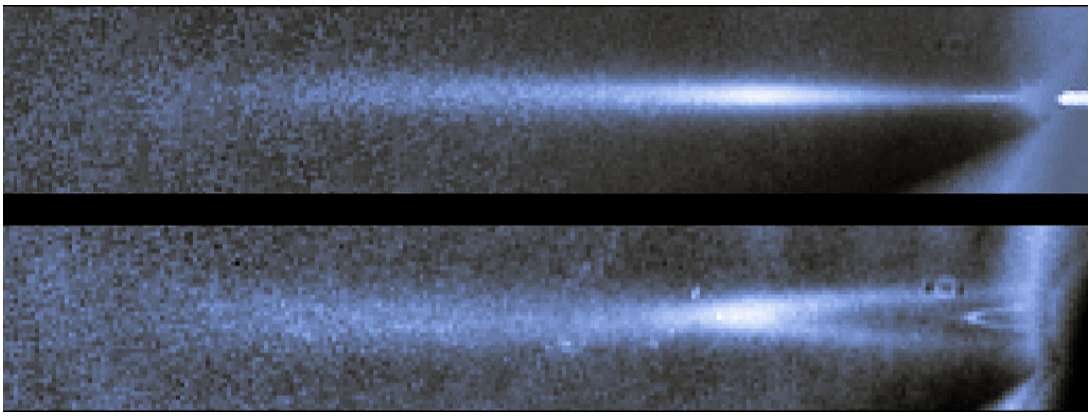


Figure 2. The tenuous E and G rings. Two images of the east ansa of the E and G rings, taken near the Earth August 1995 ring plane crossing (top) and near the Sun November 1995 ring plane crossing (bottom) with the Hubble Space Telescope. The planet Saturn, not visible, is on the right side. The two images cover the same region ranging approximately from 2.5 to 8 Saturn radii. The G ring appears as a thin bright line extending beyond the edge-on main rings in August but it is visible more distinctly in November, when the rings are inclined. The E ring is seen as a line extending beyond the G ring. The E ring peaks at the satellite Enceladus.

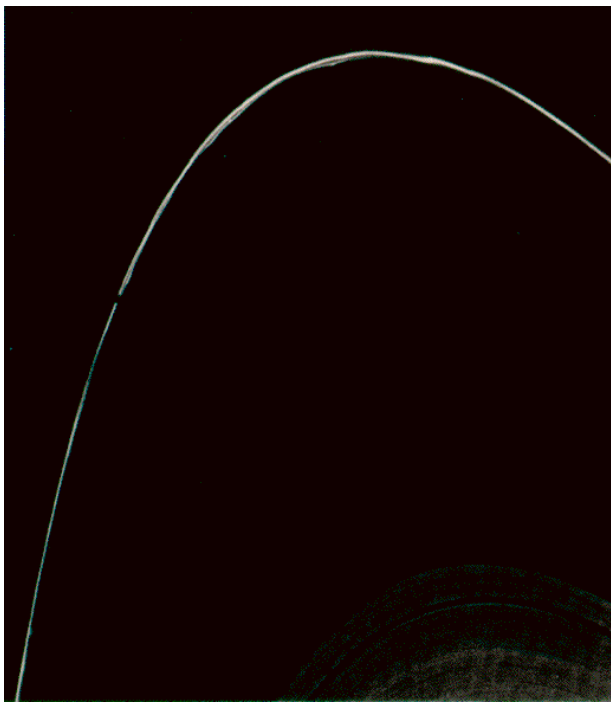


Figure 3. The strange F ring. One section of the strange F ring viewed by Voyager 1 showed braids, clumps of matter and kinks. Nine months latter, Voyager 2 revealed that the structures were gone.

Figure 4. Radial structures of the rings. Voyager 1 obtained this spectacular view of the crescent Saturn and its rings from a distance of about 1.5 million kilometers. Typical structures in the rings vary. Going outwards from the innermost visible ring, we first encounter the C ring characterized by a broad platform. Then, the B ring dominated by irregular structures. The Cassini Division is a region of low optical thickness. The A ring is almost featureless. The Encke gap can clearly be seen in the outer part of the A ring. Finally, the F ring is visible like a pencil line circumscribing the main rings.

reddish. More detailed analysis is needed to indicate their particle composition.

Size and physical nature of the particles

Our understanding of the sizes of the main rings' particles comes principally from radio measurements of the brightness temperature and the radar reflectivity of the rings. Observations at these millimeter to centimeter wavelengths indicate that most of the surface area of the ring must be filled with particles whose radii lie between a few centimeters and a few meters. Very little micrometer-sized dust is present. Small moonlets (10–100 m radius)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

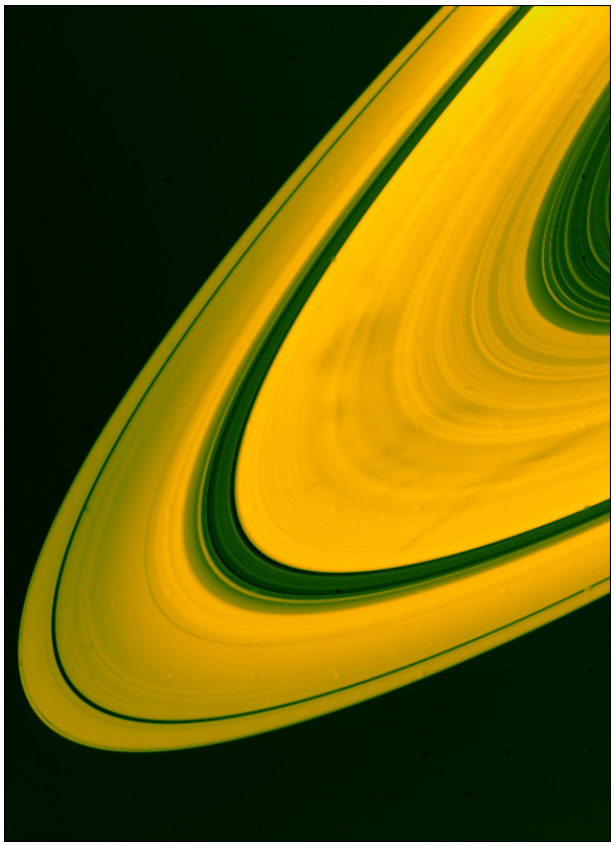


Figure 5. Spokes. The dark spokes in the B ring can be seen in this Voyager 2 picture by reflected light. They are made of small particles (about 1 micrometer), which do not reflect the light properly.

have been invoked to explain some irregular structures such as gaps and narrow ringlets. One satellite, Pan (∼10 km), has been discovered within the A ring.

Although the rings are inside the Roche limit, particles can gravitationally be bound onto the surfaces of larger ones, and particle collisions should produce accretion in Saturn's rings, especially in the A ring. The process of accretion can be counterbalanced by tidal disruption, so that complex, aggregate, fluffy particles (dynamical ephemeral bodies (DEBs)) may exist. Recent studies indicate that aggregates of microscopic particles, gathered in weak clusters of a few centimeters to a few meters, could explain the spectroscopic properties. These results seem to support the DEB model of ring particles, but further studies are required.

The E ring particle size distribution is dominated by microscopic particles: a narrow size range centered near 1 μm radius dominates its brightness. The G ring contains a large range of particle size which goes from dust particles to kilometer-sized bodies.

The very strange morphology of the F ring is not easy to interpret and could be due to its position near the Roche limit, which would imply a very broad particle size range. Voyager detected a narrow core (a few kilometers in width) of centimeter-sized particles, with an inward envelope (some 10 km in width) of micrometer-sized particles. Both the F and G rings must also contain a certain number of moonlets, roughly 1 km in size, to supply the smallest particles and to explain some azimuthal variations.

Ring structures

Disks and rings are a natural consequence of dissipation in rotating systems. A cloud of debris surrounding an oblate planet settles very quickly into a flat circular ring in the equatorial plane because of collisions between the particles. Subsequently, a spreading process occurs on a time scale much larger than that of the flattening process.

Eventually, the traditional image of a ring is that of a set of particles in orbit around a planet undergoing mutual collisions and making up a structure with perfect circular symmetry and ill-defined edges. However, Voyager probes and the most recent ground-based observations revealed several surprising structural features (see figure 4): eccentric shapes and inclined orientations; sharp edges; azimuthal variations, arcs, clumps, kinked and braided configurations; waves. We will now discuss some major structures by describing the phenomena which create them.

Gaps and narrow rings

The fact that the rings have sharp edges rather than diffuse ones, as well as the presence of thousands of narrow rings clearly separated by gaps, indicates that some mechanism confines the particles. Through gravity, a nearby satellite will alter a ring particle's orbit, making it elliptical. This effect is especially pronounced when there is a resonance between the ring particle and the satellite (a resonance occurs when the orbital periods of the ring particles are commensurate with the period of the satellite by a ratio of integers). The resonant perturbations on a population of ring particles increase the density in some places and decrease it in other places. Because of the collisions, the theory shows that there is a repulsion between part of the ring and the satellite, leading to the creation of a gap in the disk. For instance, the small moon Pan has been discovered in the Encke gap. If the forcing satellite is external to the ring system and close to it, a sharp edge will form accordingly. The process is clearly illustrated in the A ring, where a sharp edge is seen at a Janus resonance, which marks the outer edge of the ring.

Strange narrow and eccentric features are present in the C ring, the Cassini Division and the Encke gap. These ringlets are surrounded by two gaps. In some cases, such as the Encke gap, the ringlet can only be seen at some ring longitudes and may consist of dust on 'horseshoe orbits' which has been ejected from the moonlet Pan. By contrast, the formation of some structures remains unexplained.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Azimuthal variations

Narrow rings in the Encke gap exhibit strong azimuthal brightness variations. However, the most striking phenomenon is undoubtedly the F ring, located near the Roche limit. Voyager images indeed showed some condensations up to about 5 km in radius within the rings. The F ring is surrounded by two moons (Pandora and Prometheus) orbiting on either side of the ring and called the 'shepherd' satellites. The strange appearance of the F ring suggests that moons–F ring interplay has some importance, but the confinement of this ring remains unexplained.

A moon in an eccentric orbit induces variations in semimajor axes of the ring particles. As a result, the ring particles tend to crowd into clumps. Some observed condensations are indeed roughly consistent with those induced by gravitational influence of the two moons. The clumps or arcs could also be explained through resonances between particles and moons, which confine the ring particles. Unfortunately, collisions between the ring particles tend to smooth the clumps, so that it is not clear whether these features would be detectable. The existence of small (radius <10 km) moons within the F ring has been postulated by several authors, because an embedded moon may provide a source for the condensations. However, no moon has been seen yet. So, it is not clear whether the clumpy structures are individual solid objects embedded in the ring or just enhancements in the local surface density due to the perturbation of the shepherd satellites. Further theoretical and observational studies are required to answer the fundamental questions of confinement of ring particles by moons.

Waves

Among the most intriguing phenomena observed by Voyager were the spiral waves. Observations and dynamical theories have identified two kinds of spiral waveforms: density waves and bending waves. Density waves are spiral patterns of enhancement and depletion of ring materials. Bending waves are vertical undulations of the disk, formed by inclined orbits of particles.

A wave is the oscillatory response of the disk associated with a satellite resonance. Saturn has several known satellites inside the orbit of Mimas (see table 1). Satellites in eccentric and inclined orbits supply periodic horizontal and vertical perturbations to ring particles at resonance locations. The displacement of particles away from their circular orbits perturbs the gravity field in the immediate area. The self-gravity of the aggregation of ring particles allows the perturbations to propagate away from resonance locations, leading to the formation of spiral (density or bending) waves.

Waves are most common in the A ring. The waves are signatures of a number of physical properties: mass density, viscosity, vertical thickness. As an example, the total mass of Saturn's rings could be estimated to be about the same as its moon Mimas by inferring the dynamic response of the rings to the waves.

Meteoritic bombardment

Soon after the spectacular imaging observations of the ring systems at the Saturn encounter by Voyager, it was quickly recognized that the formation and structure of the rings must be closely coupled to the interplanetary meteoroid complex via collisional interactions. The basic idea is that hypervelocity meteoroid impacts create EJECTA, which are reabsorbed by the ring plane. Continuous exchange of ejecta between different ring regions causes mass transport from one region of the ring to another. In particular, the bombardment may help to sharpen the inner edges of the A and B rings, which cannot be done by resonances.

Another physical consequence of the meteoroid–ring impact would be the generation of ring atmosphere. Moreover, in some theoretical models elaborated from the rings–meteoroid interaction, the lifetime of the rings was estimated to be as short as a few million years. While such estimates have large uncertainties, they underline the importance of the meteoritic bombardement in shaping the ring dynamics.

Spokes

Thirty-seven days before its closest approach to Saturn, Voyager 1 discovered dark, nearly radial features in Saturn's B ring; these finger-like markings are called spokes, visible in figure 5. These dark clouds are about 8000 km long and 2000 km wide, develop in several minutes, follow the rotation of the rings and disappear several hours later. In fact, the spokes appear dark in reflected light and bright in scattered light, indicating that they are made of micrometer particles. These small particles are thus most easily moved by electric and magnetic forces. These properties point to a mechanism coupling the dynamics of ring particles with electromagnetic phenomena in the magnetosphere and have prompted several scenarios to explaining the spokes' formation. The most advanced one is that the spokes are the result of meteoroid impacts onto the rings, producing a plasma. Lorentz forces from Saturn's magnetic field force the plasma into a radial direction, which electrostatically levitates small ring particles, leading to the formation of spokes. In another theory, it would be an electron beam originating in the upper magnetosphere, which could hit the ring and charge the dust grains. One alternative possibility is that spokes are the manifestation of compressional waves (called also magnetosonic waves) in the partly ionized ring disk in Keplerian rotation, embedded in the vertical magnetic field of the planet. However, none of these scenarios is fully satisfactory.

Dusty rings

The dusty D, E, F and G rings tend to have very brief lifetimes owing to a variety of physical processes: drag forces, such as plasma and Poynting–Robertson drag, sweep material out of the ring, while destruction mechanisms, including magnetospheric and meteoroid impacts, eliminate dust grains in place. The dust observed must therefore be replenished by some nearby source,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Saturn's rings and inner satellites

Object	Radial distance[a] (km)	Period (hr)	Particle size[b]	Optical depth	Mass (g)
Cloud tops	60 330	10.66	–	–	–
D ring	66 000–74 500	4.9–5.6	10–100 μm	10^{-6}–10^{-5}	?
C ring	74 500–92 000	5.6–7.9	1–500 cm	0.05–0.35	1×10^{21}
B ring	92 000–117 580	7.9–11.4	1–500 cm	0.4–2.5	3×10^{22}
Cassini division	117 580–122 170	11.4–12.1	1–750 cm	0.05–0.15	0.5×10^{21}
A ring	122 170–136 780	12.1–14.3	1–500 cm	0.4–2	6×10^{21}
Pan[c]	133 580	13.8	10 km	–	4×10^{18}
Atlas	137 670	14.4	15 km	–	9×10^{18}
Prometheus	139 353	14.7	45 km	–	3×10^{20}
F ring	141 220	14.9	0.01 μm–10 cm	0.01–0.6	$1 \times 10^{17\pm1}$
Pandora	141 700	15.1	40 km	–	2×10^{20}
Epimetheus	151 420	16.7	60 km	–	5.5×10^{20}
Janus	151 470	16.7	90 km	–	1.98×10^{21}
G ring	166 000–173 000	19.9	$<$0.03 μm	10^{-6}	?
Mimas	185 540	22.6	–	–	4.5×10^{22}
E ring	181 000–483 000	$\sim$33[4]	1 μm	10^{-5}	7×10^{11}
Enceladus	238 040	33	–	–	8.5×10^{22}

[a] Radial distance is the boundaries for the rings, semimajor axis for the satellites.

[b] Particle size is range of typical radii for ring particles, mean radius for the satellites.

[c] Pan orbits inside the Encke gap.

[d] The E ring is very broad, but densest near the orbit of Saturn's moon Enceladus.

either seen or unseen. Sources of the dust are meteoroid impacts onto moons and large ring particles and collisions between ring particles which release regolith material. The dust content within the rings is then determined by a balance between collisional and meteoroid excavation of dust, as sources, and sweep-up and loss from the rings by drag and magnetic field, as sinks. The typical dusty ring model considers two distributions of particles: one includes the large particles which are responsible for the production of dust, the other the dust particles. The balance between dust loss and dust production allows us to determine the population of the large bodies. For instance, it is proposed that the G ring is formed by a dusty population of short-lived particles continuously maintained by a belt of large bodies of roughly 1 km in size.

Age and origin of the main rings

The main rings appear to result from a similar process to that which created regular satellites. Their orbits are prograde, equatorial and nearly circular. A question that immediately arises is whether they are the uncoagulated remnants of satellites which failed to form because of the tidal forces or the result of a disruption of pre-existing objects. In fact, evolutionary processes with time scales considerably smaller than that of the solar system appear to affect the ring systems. Because of these short time scales, rings would not form with their primary planets but could be created later by disruption of satellites.

Age of the rings

The high albedos of the main rings' particles seem to imply that the rings lack dark impurities. Over the age of the solar system, the rings should have been struck by roughly 0.1–1 times their mass in meteoritic material. So, if the rings existed from the origin of the solar system, they should be much darker than they are at present. In particular, the impacts of dark micrometeoroids onto the B ring should have darkened the particles to their present albedo in only 100 million yr. This darkening time is comparable with the evolution time inferred for the main rings from mass loss by erosion.

An independent short time scale argument comes from the evolution of the small moons near the rings. The moons which have resonances with rings pull angular momentum out of the ring system and, therefore, should evolve outward. For example, the transfer of angular momentum between rings and moons would move the satellite Prometheus the same distance it has already moved from the edge of the A ring in several million years. Neither the moons nor the rings could then maintain their present positions for more than about 10 million yr.

Origin scenarios

Since many workers in the field now look at rings as systems which cannot survive for longer than 10–100 million yr, the rings must be created and recreated many times over the age of the solar system. In the recent-origin models, the rings are created by the breaking up of some other body, either by cometary impact on one of the nearby moons of Saturn or by tidal disruption of a large comet which passed close to the planet.

The lifetime of a moon before a catastrophic disruption depends on its size and on the bombardment flux. The bombardment flux has varied during the age of the solar system. It is smaller in the current epoch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

than during the early phase, so that the expected lifetime for a moon 200 km in radius is close to the age of the solar system: the creation of Saturn's rings in the last 100 million yr is therefore an improbable event. A similar alternative involving tidal disruption of a large comet suffering a close encounter with the planet, has been discussed, but it appears to be equally unlikely.

Current understanding favors the destruction model in which the Saturnian ring system is derivative. However, it is especially difficult to explain, because of its large mass and compositional purity. Clearly, considerable effort needs to be devoted to this.

Prospects for the future

The complicated physical and dynamical processes involved in the generation of Saturn's rings require further studies and observations. The recent observations of the 1995 ring plane crossings and the upcoming Cassini mission to Saturn should be able to yield new knowledge on the ring system.

New results and challenges from 1995 ring plane crossings observations

The most spectacular ring system of the outer planet can also be a nuisance for the detection of faint material lying close to the main rings. Huygens was the first who realized that Saturn's ring system appears to vanish twice every orbital period of 30 yr as the Earth crosses the ring plane and that the rings are viewed edge on. The last crossing through Saturn's ring plane during 1995–1996 has provided the best opportunity to observe the faint E, F and G rings and small satellites from Earth until the year 2009.

The structure and particle properties of the E ring were constrained by new photometric observations, from the UV (0.3 μm) to the near infrared (2.2 μm). These observations continued to indicate a fairly narrow particle size centered around 1 μm radius. This predominance of such a narrow-size particle is unique in planetary rings and confirms that the ring may be self-sustained by collisions between E ring particles and the satellite ENCELADUS. A puzzle is the absence of localized peaks associated with other icy moons.

The 1995 campaign captured stunning new images of the G ring. Its intensity and inferred optical depth show the G ring to be a region where many large particles are present. This supports the idea that the G ring was formed by the disruption of a progenitor satellite. So, the observations of E and G rings are in fair agreement with previous spacecraft data, but photometric analysis and modeling should continue.

The most surprising result was that a few teams have detected several never-before-seen bodies in the vicinity of the F ring. First studies indicated that these bodies are not new satellites but more probably transient clumps or arcs within the F ring. These discoveries are more interesting than finding new satellites, because the dynamical processes sculpting these objects probably occurred when many of the planets first formed moons and rings. An explanation of these phenomena would constrain the processes of accretion in the early solar system. The other very interesting and puzzling result to come out of the ring plane crossing data is the 1 h lag of the satellite Prometheus behind its expected position. This lag actually challenges celestial mechanics and seems to reflect the importance of the moon–ring interaction.

The Cassini/Huygens mission

The CASSINI/HUYGENS MISSION is a 4 yr, close-up study of the Saturnian system. It will begin in 2004 and include the planet's atmosphere and magnetic field, several moons and of course the rings. As the best-instrumented probe ever sent to another planet, Cassini will produce the most complete information about a planet system ever obtained. For instance, the instruments will investigate the rings at UV, optical, infrared and microwave wavelengths. We will be able to map the chemical makeup and the size distribution of ring material. Some cameras will return images of the rings at resolutions better than a few hundred meters/pixel. It will be essential to observe possible changes at high resolution over a period of 25 yr between the Voyager and Cassini missions and therefore to select peculiar features to be repeatedly observed during the 4 yr of the mission. Finally, one hopes that the orbiter will remove many of the uncertainties currently besetting the problem of ring origin.

Bibliography

Greenberg R and Brahic A (eds) 1984 *Planetary Rings* (Tucson, AZ: University of Arizona Press)

Nicholson P D, Showalter M R, Dones L, French R G, Larson S M, Lissauer J J, McGhee C A, Seitzer P, Sicardy B and Danielson G E 1996 Observations of Saturn's ring-plane crossings in August and November 1995 *Science* **272** 509–15

François Poulet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Saturn: Satellites

We now know that SATURN has at least 30 satellites, forming a largely diverse set of bodies (figure 1). They range from the planet-like TITAN, surrounded by a dense atmosphere and half the Earth's size, to small, barren objects, of irregular shapes (table 1). The surfaces of the latter objects are all believed to be covered with some type of frozen volatile, primarily water ice, but also carbon dioxide, methane products, ammonia and other volatiles. The surface of Titan remains hidden underneath thick clouds. At some 1.5 billion kilometers from the Sun, nights are terrible on the surfaces of Saturn's companions: −220 ºC. However, during daytime, also, the thermometer never shows more than −170 ºC.

The first of the Saturnian moons, Titan, which is also the biggest satellite after Jupiter's Ganymede, was discovered in 1655 by Christiaan Huygens. Later on, Cassini discovered the next four largest satellites of Saturn: Rhea, Iapetus, Dione and Tethys. The family grew as Earth-based telescopes acquired more resolving power and the Voyager missions visited the system (figure 1). The eighteenth, Pan, was discovered nearly 10 years after the Voyager fly-bys, embedded in the A ring of the giant planet (see SATURN: RINGS). Saturn's ring plane crossings—that is, when the rings move to an edge-on orientation as viewed from the Earth and therefore have their bright light dimmed by satellite eclipses—represent an ideal configuration for discovering new satellites. The recent discovery, in 2001, of twelve new moons of Saturn brings the planet's total to 30: the most in the Solar System. The new moons vary between about 6 and 30 km in diameter, and are moving in irregular, tilted orbits. They fall into several clusters, leading Brett Gladman, of the Observatoire de la Cote d'Azur, Nice, France, and colleagues to conclude that they are the remnants of larger satellites that have been fragmented by collisions. The moons probably began life as wandering bodies that were captured by Saturn's gravity. Gladman and colleagues found the moons by scanning Saturn with sophisticated electronic detectors mounted on medium-sized, ground-based telescopes.

At the beginning of the 20th century, it was recognized that Titan was a special body in our Solar System, the only satellite known to possess a significant atmosphere, which was soon demonstrated to host an active organic chemistry similar, in some aspects, to that found on Earth.

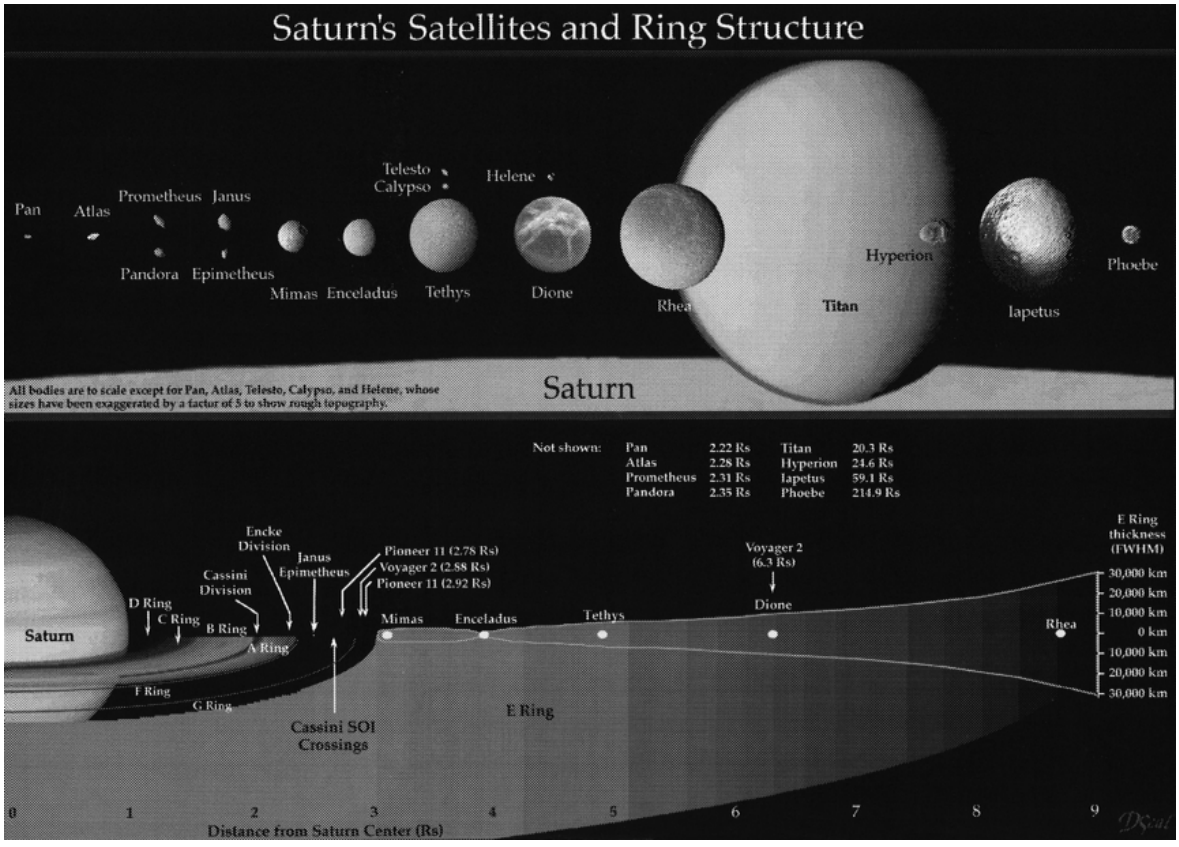


Figure 1. Saturn's satellites are a diverse set of objects, ranging from Titan, a planet-size moon with a dense atmosphere, to Hyperion, a small and dark, irregularly shaped icy body. Some of the satellites are embedded within the ring system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Saturn's satellites.

Name, year of discovery	Magnitude	Orbital radius (km)	Orbital period (days)	Density (g cm^{-3})	Diameter (km)
Pan, 1990	18.5	133 580	0.56	–	20
Atlas, 1980	18	137 670 (A-ring shepherd)	0.60	–	30
Prometheus, 1980	16.5	139 350 (inner F-ring shepherd)	0.61	–	100
Pandora, 1980	16	141 700 (outer F-ring shepherd)	0.63	–	90
Epimetheus, 1966	15.5	151 450 (co-orbital)	0.69	0.63	118
Janus, 1966	14.5	151 450 (co-orbital)	0.69	0.65	178
Mimas, 1789	12.9	185 520	0.94	1.12	398
Enceladus, 1789	11.8	238 020	1.37	1.00	498
Tethys, 1684	10.3	294 660	1.89	0.98	1060
Telesto, 1980	19	294 660 (Lagrangian: trailing)	1.89	–	30
Calypso, 1980	18.5	294 660 (Lagrangian: leading)	1.89	–	26
Dione, 1684	10.4	377 400	2.74	1.49	1118
Helene, 1980	18.5	377 400	2.74	–	32
Rhea, 1672	9.7	527 040	4.52	1.24	1528
Titan, 1655	8.4	1 221 830	15.94	1.88	5150
Hyperion, 1848	14.2	1 481 100	21.28	–	290
Iapetus, 1671	10.2–11.9	3 561 300	79.33	1.03	1436
Phoebe, 1898	16.5	12 952 000	550.40	–	220

Formation and evolution

Our solar system is believed to have formed from the condensation of a primordial cloud of gas and dust about 4.5 billion yr ago (see SOLAR SYSTEM: FORMATION). The outer regions of the protosolar nebula, where the giant planets and their satellites formed, were colder than the inner ones. As a consequence, volatile material, gases with low condensation temperatures, prevailed there and they became the building blocks of the outer bodies. Such materials include methane, ammonia (and their hydrated forms), water ice and ice silicate mixtures. The giant planets and their satellites accreted these components into successively larger lumps of matter, forming the planetesimals. The planets accreted enough mass to maintain large atmospheres and became the gas balls we know today. Their satellites, for the most part, were small blocks of mass, characterized by low density and therefore unable (with the exception of Titan) to hold onto a significant amount of gas to form an atmosphere against thermal escape. The surfaces of these smaller bodies were 'contaminated' by two sorts of carbon- and hydrocarbon-rich material: the C type and the D type respectively.

Following their formation, a long period of heating awaited the newly accreted satellites. The heating was due to the release of the gravitational potential energy and of the mechanical energy produced during the heavy bombardment of their surfaces by remaining debris. The Voyager images also showed that the bright region of IAPETUS (and especially the North Pole) is heavily cratered, as in the cases of Mercury, CALLISTO or the Moon. HYPERION is similar to Iapetus in terms of surface cratering. MIMAS and TETHYS have impact craters (named Herschel and Odysseus), caused by bodies that must have been rather large. Hyperion is believed to have disrupted in the past (as a result of an important impact?) and its debris may have fallen on Titan, which is in a 4:3 resonance with Hyperion. Tidal and resonant interactions among the satellites and among the satellites and Saturn are another source of frictional energy causing heating of the satellites.

Chemical fractionation is possible on most of the larger satellites of Saturn, which may have undergone periods of melting and active geology within a billion years of their formation, before becoming quiescent. ENCELADUS may still be geologically active.

Physical and dynamic characteristics

Most planetary satellites present the same hemisphere toward their primary as a result of tidal evolution. They are in synchronous rotation with the planet. Of Saturn's satellites, only two, Hyperion and Phoebe, are known to exhibit asynchronous rotation.

The Saturnian satellites other than Titan (the largest and most massive of Saturn's satellites, boasting 0.02% of the planet's mass) are too small to retain much of an atmosphere. When GERARD KUIPER, of Chicago

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

University, discovered in 1944 spectral signatures in the Titan spectrum at wavelengths longer than 0.6 μm, among which he identified two absorption bands of methane at 6190 Å and 7250 Å, he also searched for a similar behavior in the spectra of other Saturnian satellites. In data he had obtained in 1952, he found—by contrast—differences between Titan and the other satellites in the intensity observed in the ultraviolet and visible continuum. These differences, due to the presence of an atmosphere around Titan, hinted at the lack of such an adornment on the smaller satellites of Saturn. As a consequence, astronomers expected them to be simply icy, geologically dead bodies. However, the era of space exploration had some surprises in store for the scientists.

High-resolution photometry and spectroscopy have shown that the main constituent on the surface of the brightest among Saturn's satellites is water ice. Their densities (ranging from 1.0 to 1.9) are such that ice must be a major constituent of the interior, also. The most probable structure for each of them (and Titan too) is a rocky center surrounded by a thick shell of ice. Taking typical densities for rock and ice, and knowing the size of the satellite and its mass, a model for the relative sizes of the rocky core and icy mantle can be worked out. In the case of smaller satellites, such as Janus and Epimetheus, the densities are of the order of 0.65.

Of course, the ice and the rock will not be pure water or silicate but may contain many impurities. The icy surfaces are obviously contaminated by materials of various colors and these can be studied using spectroscopy. Contrary to what was previously believed, the Voyager and Galileo spacecraft revealed in 1980 and 1995–1998 a wide range of geological processes on the moons of the outer planets and showed that several of the larger Kronian satellites may have undergone melting of their mantles and subsequent differentiation and resurfacing of the liquid ice or ice silicate mud. Flows of these liquid components are similar to the lava flows witnessed on Earth and the Moon, owing to the melting of silicate rock mixtures. The ridged and grooved terrain found on Enceladus and Tethys may be the result of tectonic activities occurring throughout the solar system. Explosive volcanic eruptions may be occurring on Enceladus, as on Earth.

In the young solar system, meteoritic bombardment was intense and frequent causing the surfaces of icy satellites (among others) to be heavily cratered. Nowadays the bombardment continues but at a significantly lower rate. Knowing something about the flux of impacting material and by enumerating the number of CRATERS on a solid surface, geologists can recover information on the origin and epoch of formation of ground portions. Impacts have several consequences: they pulverize the rock and the icy components of a surface, forming a fine material that covers it, the REGOLITH; they tend to excavate and expose fresh material on the surface; they also cause the vaporization and escape into space of volatiles, leaving behind on the ground a deposit of opaque and dark material. As a consequence, the larger satellites of Saturn, like those of Jupiter (see JUPITER: SATELLITES), have a brighter 'leading' than 'trailing' hemisphere (a leading side is the hemisphere seen from the Earth when the satellite moves toward our planet; the trailing side is the opposite in the direction of orbital motion). This is the case for Hyperion and Titan, for instance, but not for Iapetus, as will be discussed in the following section. The hemispheric asymmetry is believed to be due to more meteoritic encounters occurring on the leading side. The fact that well-formed craters, similar to those observed in the inner solar system bodies, are also observed on the Galilean and Kronian satellites today, suggests that the surface ice is strong enough to sustain intense impacts, perhaps because it is mixed with silicate contaminants or other impurities. The low temperatures and the small surface gravity in these cases prevent viscous relaxation from eradicating the impact features. Mimas is thus covered with high-rimmed, bowl-shaped craters, Herschel being the most prominent one (130 km large).

Individual properties

With a diameter of more than 500 km, the six larger satellites of Saturn are Titan (5150 km, half our own Earth), followed by RHEA and Iapetus (less than a third of Titan's size), DIONE and Tethys (with close to 1000 km diameter) and Enceladus, the smallest of the mid-size satellites (table 1; figure 2). Hyperion, Phoebe (farthest from Saturn), Janus, Epimetheus, Mimas and Prometheus afford diameters of more than 100 km, being sizable objects. The latter four are found close to their primary and are therefore difficult to observe from the ground because of the planet's scattered light. Most among these and the even smaller icy satellites (such as Pan, Atlas, Pandora, Telesto, Calypso and Helene), were not discovered before the 1980s. Most of what is presently known about the Saturnian system is due to the extraordinary exploration of the two Voyager missions, which encountered Saturn, its rings and satellites in 1980 and 1981.

The satellites have low densities and high ALBEDOS. They were therefore expected to be largely composed of water ice, probably combined with ammonia and other volatiles. Spectroscopic measurements performed on the medium-sized Saturnian satellites confirm this hypothesis by showing the spectral signatures of ice. Even Titan's surface, eternally hidden by a thick cloud deck and only recently explored by imaging and spectroscopy in the near-infrared, is believed to be covered with some sort of 'dirty' water ice. Other areas of the satellite's surface are expected to host a variety of landscapes, ranging from hydrocarbon lakes to ice-covered mountains.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

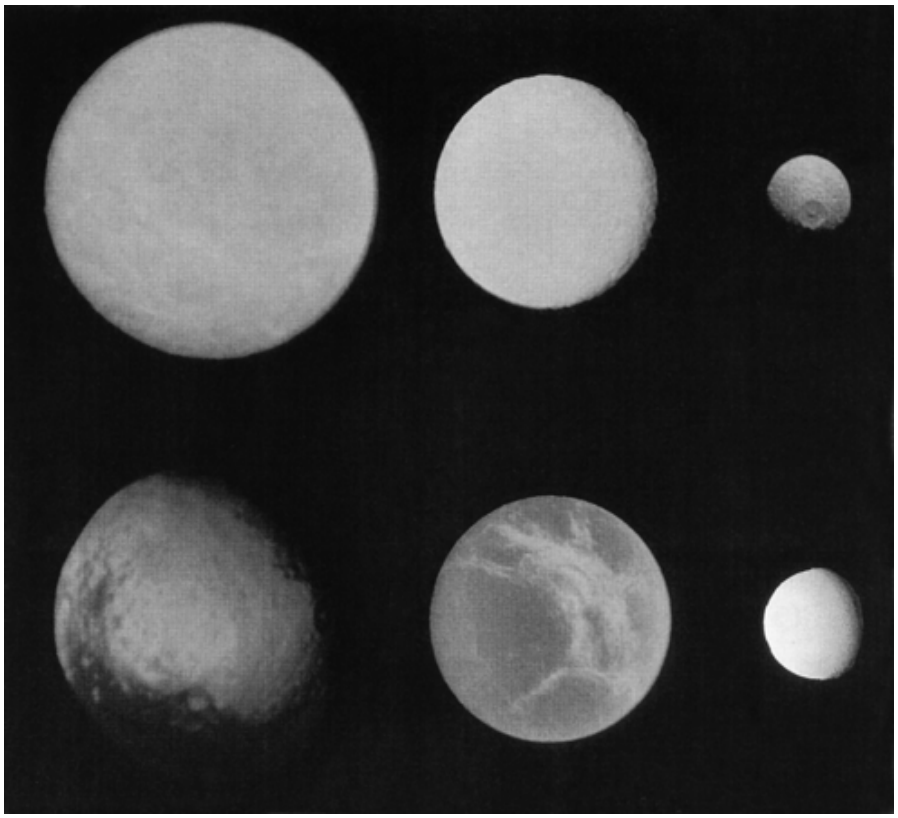


Figure 2. The six largest icy Kronian satellites are all unique worlds; from top left and clockwise, Rhea, Tethys, Mimas, Enceladus, Dione and Iapetus.

The larger satellites

After the planet-sized Titan, four large and four medium-sized satellites share the system's mass: Rhea, Iapetus, Dione and Tethys are kilometer-sized objects, while Enceladus, Mimas, Hyperion and Phoebe are less than half that size.

Iapetus is not like the other Kronian satellites in that it is the only large satellite of Saturn in a highly inclined orbit and—for no obvious reason—its two faces are radically different, one bright and the other very dark (figure 3). The surface albedo on the trailing hemisphere varies from 0.4–0.5, values typical for ice-covered objects, to as low as 0.02 in the central parts of the leading hemisphere. Already in 1672, when CASSINI discovered Iapetus, it was noticeable that, at one point on its orbit around Saturn, the satellite seemed quite bright, whereas on the opposite side of the orbit it almost disappeared. Cassini then correctly deduced that this was ample indication for a marked difference between the moon's two hemispheres, with one of them (the trailing hemisphere, in other words the side viewed from the Earth when the satellite moves away from us) bearing highly reflective material, while the other (the leading hemisphere, or the side facing the Earth when the satellite moves toward us) was of a darker nature.

Images of Iapetus obtained by Voyager confirm these observations and show that the bright region, reflecting about 50% of the incident radiation, is typical of heavily cratered icy surfaces. The other side is centered on the direction of motion around Saturn and is coated with a redder material. This gives a symmetrical distribution about the apex of the orbital motion. In the craters found near 200ºE longitude, the shape and orientation of the dark deposits suggest that they cover bright terrain and that they are, in consequence, more recent. The mean density of the satellite is 1.03 g cm^{-3}, somewhat less dense than the other icy satellites of Saturn, but still consistent with models predicting water ice as the main component of these bodies (although this may be indicative of a higher fraction of ice in its interior). The dark material on the surface must then have been deposited in a very thick layer if it can cover large quantities of ice.

Because Voyager was unable to map the surface on the whole leading hemisphere, the nature and origin of the dark material on Iapetus remain a mystery. The observed symmetry with respect to the direction of the orbital motion of the satellite strongly argues in favor of an external control, or even perhaps an external origin for this material. Studies of its composition, based on ground-based low-resolution observations performed at

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the infrared telescope facilities, the United Kingdom Infrared Telescope and the University of Hawaii telescope from 1981 to 1983, show that it is very red in the visible and near-infrared. A fit of the Iapetus spectrum was obtained at first by using a mixture of polymers (10%) and hydrated silicates (90%), simulated in the laboratory. A more recent study, however, achieves a better agreement with a mixture of HCN polymers, organic residuals and water ice.

The mechanisms suggested for the production of this dark material are (a) a layer of eroded dust from the surface of Phoebe (the dark outermost Saturnian satellite, characterized by a highly inclined, retrograde orbit and with a surface albedo similar to but a little less red than that of Iapetus) pulled inwards and (b) a leading hemisphere surface modified by impacts from Phoebe material, leading to a low-albedo material concentration on that side of the surface; in the latter case, the dark material would be of endogenic origin, but controlled by exogenic impacts. A recent study of some UV Voyager archived images, in accordance with these models, suggests impacts of exogenic material from round-shaped Phoebe on the surface of Iapetus, leading to ice evaporation and leaving a dark and red deposit on the crater floors. This deposit would be a mixture of Phoebe material and a pre-existent non-volatile constituent, uncovered by the impacts, similar to the D-type material found in the asteroids of this type, as well as on Hyperion (an example would be methane erupting from the interior and subsequently being darkened by ultraviolet radiation). The major problem with the exogenic models is that the dark material on Iapetus is generally 'redder' than that found on Phoebe. However, it is conceivable that dark particles ejected from Phoebe may have undergone chemical changes that rendered them redder. Smaller than Hyperion, Phoebe's eccentric orbit suggests it may be a captured object. With a rotational period of about 9 h, Phoebe is the only Saturnian satellite to exhibit a simple, asynchronous rotation.

Hyperion is a special satellite because of the complex aspect of its chaotic rotation and its irregular shape. Although its global composition is uncertain, we know that water ice is present on the surface. Its spectrum in the visible range and in the near-infrared is compatible with that of Titan, showing a brighter leading and a darker trailing hemisphere. Hyperion is in 4:3 resonance with Titan and it has been suggested that Hyperion material was deposited in the past on Titan. Hyperion's albedo is rather low (0.21 in the visible), lower than all satellite albedos but for the ones of Phoebe and Iapetus. Its red color and low albedo value suggest the presence of an additional component (besides water ice) on the surface of Hyperion. A partial coverage by Phoebe-originating material has been suggested in this case also, as on Iapetus. One cannot exclude, however, the possibility of the existence of a certain intrinsic dark material which would have found its way to the surface through 'pockets' located in the regolith.

Because of the differences between the albedo values of Iapetus and Hyperion, different mixtures of surface material are required. Dark organic material and hydrated silicates are likely candidates for the composition of the external objects of our solar system and could explain the albedos of both Hyperion and Iapetus. It is worth noting that Hyperion is the satellite closest to Saturn after Iapetus and, consequently, it is quite conceivable because of this proximity that its surface may be covered by a similar material, which, however, was not deposited in the case of Hyperion preferentially on one of the hemispheres, owing to its chaotic rotation.

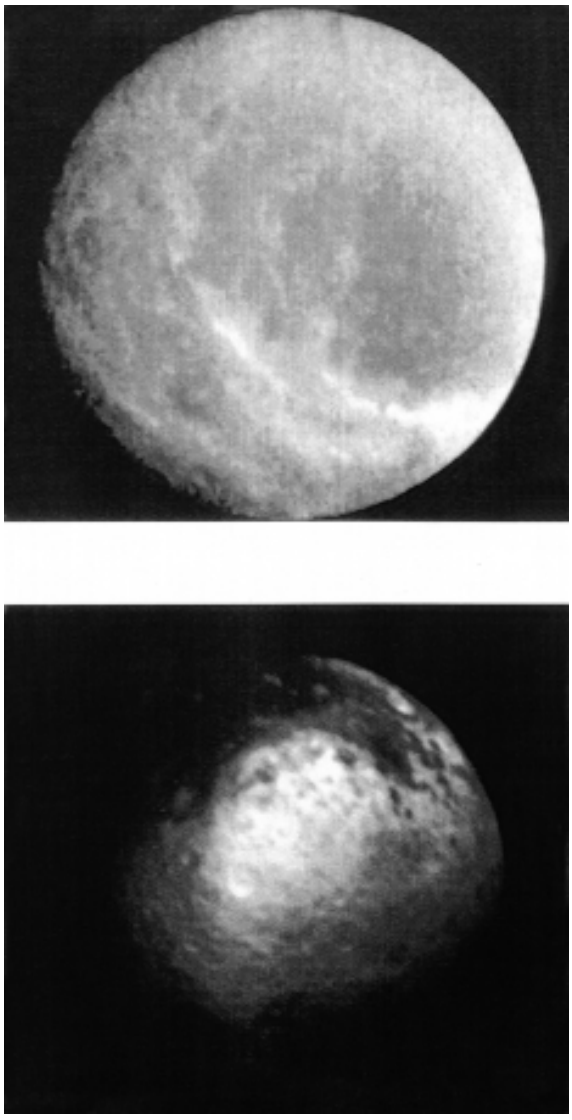


Figure 3. Rhea (upper half), the second largest satellite of Saturn, after Titan, shows an array of bright streaks, whereas Iapetus (bottom half) exhibits both bright and dark terrains on this Voyager 2 image.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thus, among the medium-sized icy satellites of Saturn, Hyperion and Iapetus are historical mysteries presenting a pronounced hemispheric asymmetry, also observed on Titan. The observational data that we possess today have not provided yet a valid explanation as to the composition of the surface of these satellites. The nature of the dark component observed is not yet discovered.

Rhea is the second most massive satellite of Saturn (figure 3), but still possesses less than 2% of Titan's mass. The surface of Rhea bears distinct signs of multiple impacts and is heavily cratered. One hemisphere shows bright wispy streaks that may have been caused by ejecta from impacts or perhaps have been caused by the condensation of volatiles leaking from the interior as a sign of internal activity. Parts of the ground are covered with large craters while other regions lack them. The larger craters are probably the signatures of early bombardment, while the smaller ones may have been more prevalent during a later episode of collisions. In many respects Rhea looks very much like the MOON or MERCURY, although the morphology of individual craters is somewhat different.

Dione appears to be quite similar to Rhea on the images, showing a wide diversity of morphology (figure 4). It also shows parts in which craters seem to be too few in number, as though many have been obliterated or covered up, indicating that several periods of resurfacing have occurred during the first billion years of its existence. It therefore appears that, although small, these bodies afforded, nevertheless, in some distant past significant internal activity. This could be the case if the relatively high densities have provided added radiogenic heat from siliceous material to spur this activity. The leading hemisphere of Dione is about 25% brighter than the other, as a result of more intense meteoritic bombardment on this side.

Icy Tethys is heavily cratered, (figure 4), similarly to Mimas, one of Saturn's inner satellites. Mimas' remarkable crater (figure 5), Herschel, is about the same size as Tethys' Odysseus (400 km in diameter, ~40% of the diameter of the satellite, the largest crater in the Saturnian system). This outstanding topographic feature is accompanied by a huge trench formation, the Ithaca (Odysseus' home country) Chasma, similar in appearance to the grooves on Enceladus. Viscous relaxation and flow in the craters of Tethys under the influence of its strong gravitational field give them an aspect flatter than for Mimas or the Moon. Regions where craters are fewer give evidence for several episodes of resurfacing in contrast to Mimas, where surface gravity is not sufficient to have caused viscous relaxation and craters tend to be high-rimmed, bowl-shaped pits.

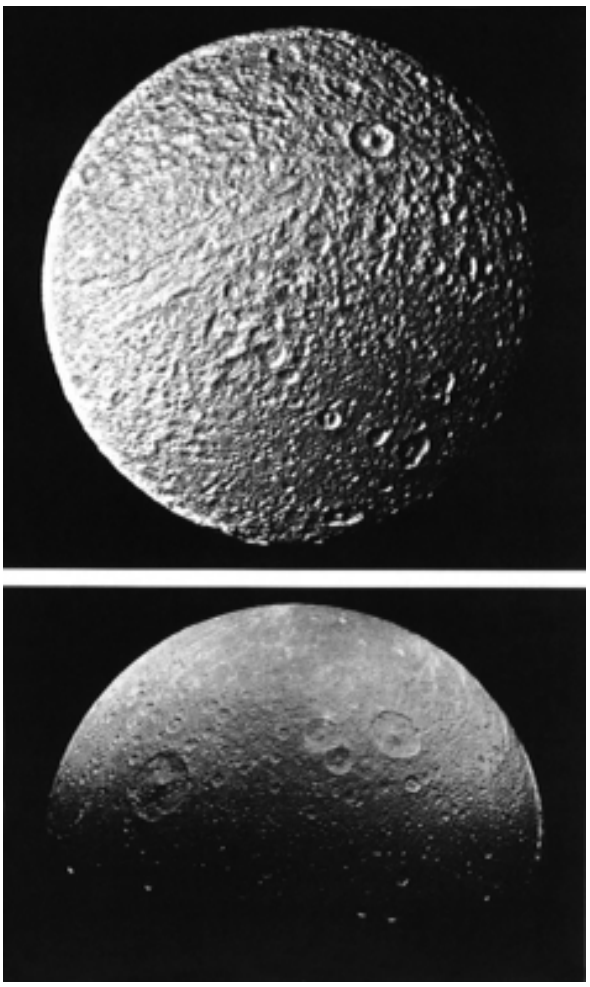


Figure 4. The surface of Tethys (upper half) shows many impressive craters, the most prominent of which is Odysseus (400 km, the largest impact structure in the Saturnian system) and Telemachus, as well as the Ithaca Chasma, a groove-like trench formation. Dione's surface (lower half) is also heavily cratered, as shown on this Voyager 1 image, with bright wispy streaks visible on the limb.

Enceladus may well be one of the most interesting and beautiful examples among the primitive, airless worlds accompanying Saturn (figure 5). Extremely close to its planet, at a distance of 238 000 km, Saturn's apparent diameter (25º) in Enceladus' sky is 50 times larger than that of the Moon seen from the Earth. Part of its surface was detected by Voyager and seemed entirely covered with vast plains of ice, an ice so pure and bright that it reflects almost all the visible radiation incident on it (by comparison, our Moon reflects only 11% of the light it receives). Enceladus shines bright: the only plausible composition for this part of its surface is pure water ice or some other very bright volatile. Here and there, on Enceladus, dark spots testify to the presence of craters, but not very many. In fact, this lack of impact craters on half of Enceladus' surface points to a relatively small age, less than a billion years, and subjection in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

recent geological past to extensive resurfacing, while some form of ice volcanism may still be occurring. This phenomenon may well be responsible for the formation of the E ring around Saturn, by providing the source of particles. The other half of the surface, as seen by Voyager 2, was a great relief for geologists because it appeared heavily cratered, and is therefore compatible with an age of 4 billion yr. The complex nature of Enceladus' surface is a delight to astronomers.

The smaller satellites

Ten of the Saturnian satellites are decent-sized objects. However, the system also has a number of unique small satellites under or around 200 km in diameter (figure 6). These are distinguished in basically three classes of objects: the shepherding satellites, Atlas, Pandora and Prometheus, believed to play a key role in defining the edges of Saturn's A and F rings; the co-orbital satellites, Janus and Epimetheus, move in almost identical orbits at about 2½ Saturn radii; finally the Lagrangian satellites, Calypso, Helene and Telesto, are so called because they orbit in the Lagrangian points of larger satellites, Dione and Tethys. Lagrangian points lie about 60º in front of and behind the larger body, and correspond to locations within an object's orbit in which a less massive body can move in an identical, stable orbit. No other satellites in the solar system are Lagrangians. The only other similar example we have is of the Trojan asteroids orbiting in two of the Lagrangian points of Jupiter.

Of the shepherds, Atlas lies several hundred kilometers from the outer edge of the A ring. The other two orbit on either side of the narrow F ring, constraining its width and causing its kinky appearance.

Janus and Prometheus, discovered in 1966 and 1978 respectively, may once have been part of a larger body that disintegrated following a major collision. Every 4 yr, the inner satellite (which orbits slightly faster than the outer one) overtakes its companion. Instead of colliding, the satellites simply exchange orbits. The 4 yr cycle then starts all over again.

Pan was discovered in 1990 in Voyager 2 images dating back to 1981. The small object lies hidden within the A ring and helps to keep the Encke division clear of particles.

Future investigations

The two Voyager spacecraft provided the first detailed reconnaissance of the Saturnian system, with Titan as one of the major poles of attraction. It seems, nevertheless, that this precious data set did but whet our appetite for our exploration of this world. Titan has become in recent years one of the main goals for the CASSINI/HUYGENS MISSION to the Saturnian system, a joint ESA–NASA mission, launched in October 1997 and which will reach its target in 2004.

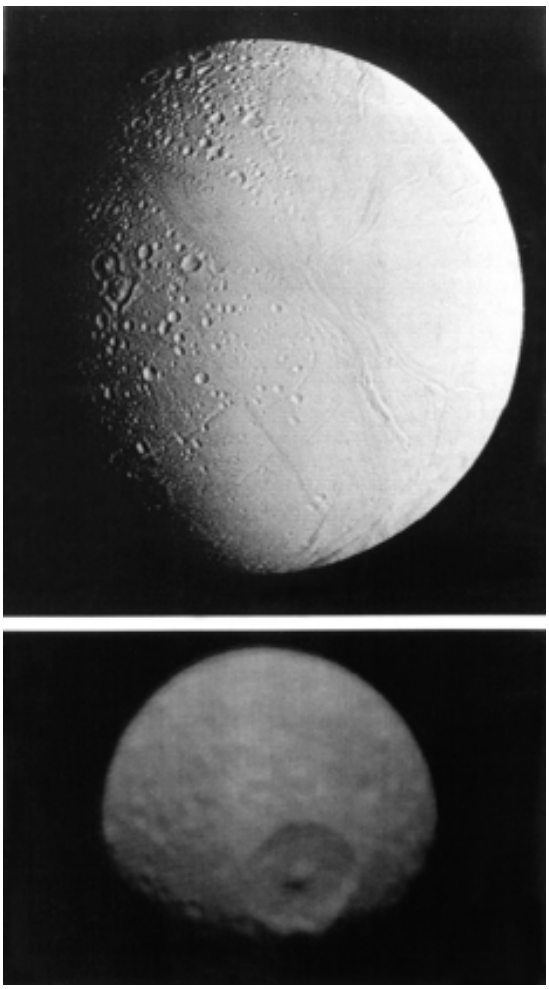


Figure 5. Saturn's icy moon Enceladus (upper half), shows a bright surface with evidence of numerous resurfacing episodes in the past. Mimas (lower half), is a small object with crater Herschel (about 130 km in diameter), nearly centered on the leading hemisphere.

Perhaps only then will some of the remaining questions regarding Titan and the other Saturnian satellites find an answer. Succinct examples are given below.

- Are there more satellites around Saturn than those known today? They could be, for instance, embedded in the ring system, and hence difficult to detect. High-resolution imaging by the Cassini Imaging Science Subsystem (ISS) will help discover even kilometer-sized objects.
- What is the detailed morphology and the exact composition of the surfaces of the satellites? For example, what is the distribution of dark material on Iapetus' craters and what does that tell us about the origin of this material? Besides ISS imaging, the three Cassini spectrometers (CIRS, UVIS and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

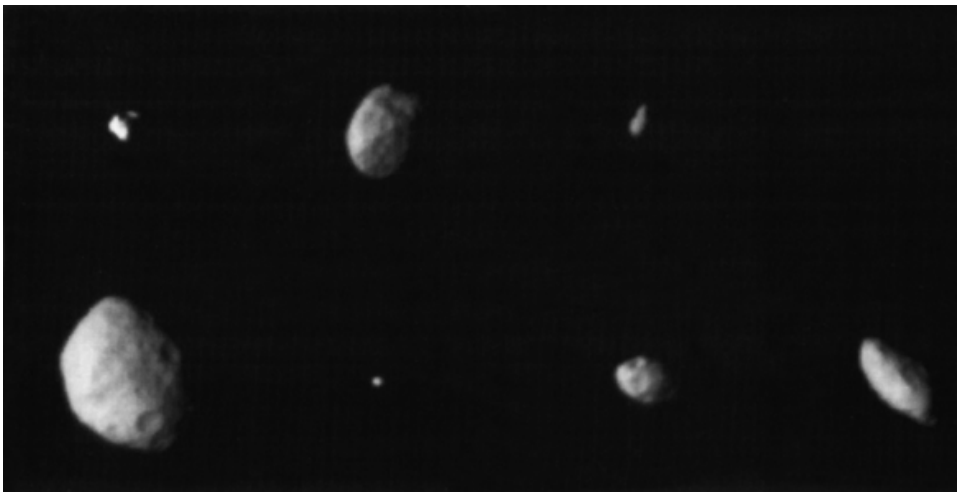


Figure 6. Seven of Saturn's small inner satellites. Clockwise from top left: Helene, Epimetheus, Calypso, Prometheus, Pandora, Telesto and Janus.

VIMS) will probe the satellites' surfaces in high-resolution mode (capable of distinguishing tens of meters on the ground) and thus allow us to correlate geophysical units with specific compositions.

- What is the connection between the rings, the dust (which, along with the icy particles, will be measured in detail by the Cosmic Dust Analyzer—CDA—on board Cassini) and some of the satellites in the Saturnian system and how does this help us understand its formation and evolution? For instance, is Enceladus the source of the E ring?
- Do any of the other Saturnian satellites, besides Titan, possess even a thin, tenuous atmosphere and what is its composition? The Cassini spectrometers will here again enable us to detect such atmospheres and their interaction with Saturn's magnetosphere (as in the case of Titan) will be studied by the Dual Technique Magnetometer (MAG), the Magnetospheric Imaging Instrument (MIMI), and the Cassini Plasma Spectrometer (CAPS). The Ion and Neutral Mass Spectrometer (INMS), the CDA and the Radio and Plasma Wave Science Instrument (RPWS) on board Cassini will on the other hand study the composition, size and mass of material eroded or expelled from the satellites' surfaces.

In the 4 yr mission of Cassini, numerous close fly-bys of the Saturnian satellites will allow us to determine their characteristics with precision and tie these precise measurements to the origins, evolution and fate of the Kronian system and furthermore to that of the solar system as a whole.

Bibliography

Beatty J K and Chaikin A (ed) 1990 *The New Solar System* 3rd edn (Cambridge, MA: Sky Publishing)

Burns J A and Matthews M S (ed) 1986 *Satellites* (Tucson, AZ: University of Arizona Press)

Coustenis A and Taylor F 1999 *Titan: the Earth-like Moon* (Singapore: World Scientific)

Athena Coustenis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scaliger, Joseph (1540–1609)

French mathematician who, in 1582, founded the system of the Julian day, a continuous reckoning of time which starts at midday GMT on 1 January 4713 BC. The system has the advantage that it is able to set aside issues of leap years, the different lengths of months, 'lost' days at changes of calendars etc. The label 'Julian' commemorates Scaliger's father Julius.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schall von Bell, Johann Adam (1592–1666)

Astronomer, born in Cologne, Germany, became a Jesuit and studied astronomy in Rome. He was one of numerous Jesuit missionaries sent to China, and was the first European ever to be a member of the court bureaucracy in Peking, becoming head of the Imperial Board of Astronomy, and adviser to the young emperor Shun-chih (ruled 1644–61). He produced a large six-part cosmological map, accompanied by pictures of astronomical instruments, he translated astronomical books and he reformed the Chinese calendar. He was falsely implicated in the death of the emperor, and escaped execution only through a fortuitous earthquake.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scheiner, Christoph (1575–1650)

Astronomer, born in Wald, near Mindelheim in Swabia (southwest Germany), became a Jesuit and professor of mathematics at the university at Ingolstadt. He made instruments, including sundials and a pantograph. On learning about GALILEO's discoveries, he obtained his own telescopes, with which he observed the Sun, and in 1611 independently discovered sunspots for himself. His publication on the subject started from the Aristotelian concept of the perfection of the Sun, so he proposed that sunspots represented transits of satellites of the Sun. He came into conflict with Galileo, who argued that sunspots are on the surface of the Sun, because they change their shapes, and they originate and perish on the solar disk. Scheiner at first persisted stubbornly against the evidence. In Rome he published *Rosa Ursina* (1630), a work on sunspots that was required reading for a century. His theories having been repeatedly dismissed by Galileo, Scheiner attacked him in the book, although by then he had actually accepted many of Galileo's opinions. Shortly after the appearance of *Rosa Ursina*, sunspot activity decreased drastically (the MAUNDER minimum, 1645–1710), and Schiener's work could not be verified or superseded. Scheiner retired to Vienna and Neisse, building a Jesuit college and writing a massive refutation of the Copernican theory (*Introductory Treatise in Favour of a Moving Sun and a Stable Earth against Galileo Galilei*), which, posthumously published, had no influence whatsoever.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schiaparelli, Giovanni Virginio (1835–1910)

Astronomer, born in Savigliano, Italy, studied in Berlin and at Pulkova, Russia, and became director of Brera Observatory, Milan. He wrote authoritative accounts of the early astronomy of many cultures. He observed comets, inferring from their tails that there was a repulsive force from the Sun (solar wind and radiation pressure). He explained the regular meteor showers as the result of the dissolution of comets and proved it for the Perseids by calculating orbital effects on the Perseid shower. From his observations, he mapped Mars, naming martian 'seas' and 'continents' and connecting features with linear features which he called 'channels' or, in Italian, *canali*. The implication that these features, mistranslated 'canals', were artificial water-distribution systems, stimulated PERCIVAL LOWELL to a sensational search for life on Mars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schickard [Schickhardt], Wilhelm (1592–1635)

Born in Herrenberg, Württemberg, Germany, became professor of astronomy in T'bingen. He corresponded with KEPLER, was one of the first to teach Kepler's discoveries, and, on the basis of his experience in making maps and other devices, proposed to Kepler the development of an analog computer to calculate planetary positions. He invented the first calculating machine, a 'hand planetarium' (like an orrery) and a machine for Hebrew grammar. He died of the plague.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schlesinger, Frank (1871–1943)

Astronomer, born in New York City, became director of the University of Pittsburgh's Allegheny Observatory and the Yale University Observatory. He developed a method to observe stellar parallaxes photographically, and compiled a catalog of 4000 stellar distances. He compiled data to produce the first and second editions of an influential and practical *Catalog of Bright Stars*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schmidt Telescope

A type of telescope, invented by the Estonian optician Bernhard Schmidt (1879–1935), that is used to photograph large areas of the sky. Because, in its original design, it was useable only for photography, the instrument is also known as the Schmidt camera. The Schmidt uses a concave spherical mirror as its light collector and corrects for the optical defect, known as spherical aberration, that is introduced by a mirror of this shape, by means of a specially shaped thin lens, or corrector plate, which is located at the front end of the telescope tube. An instrument such as this, which uses both lenses and mirrors to collect light, is called a catadioptric instrument

By contrast with a conventional telescope, where the sharply focused field of view may be only a few arc minutes or, at best, a few tens of arc minutes, the field of view of a Schmidt telescope may be six to ten degrees across. Its wide field of view makes it eminently suitable for surveying large areas of sky. A disadvantage of the system is that the surface of sharp focus (the focal 'plane') is curved rather than flat. This problem is overcome by bending the photographic plate, film, or detector to match the shape of the surface of sharp focus.

Because the corrector plate has to be situated at the center of curvature of the primary mirror, its distance from the mirror is twice as great as the focal length of that mirror. Consequently, the tube of a Schmidt telescope is at least twice as long as the focal length of its primary mirror. Because the correcting lens causes a degree of divergence in the incoming light rays, the diameter of the primary mirror has to be greater than that of the corrector plate. The maximum size of a Schmidt camera is limited by the practical problems associated with supporting the thin lens (which can be supported only round its edges). The largest Schmidt telescopes in the world, one located at Mount Palomar, California, and the other at Siding Spring, Australia, each have corrector lenses with diameters of 1.2 m.

See also: field of view, focal length, mirror, reflector, Schmidt–Cassegrain telescope, spherical aberration, telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schmidt, Bernhard Voldemar (1879–1935)

Optical designer, born on the island of Naissar, off the coast of Tallin, Estonia. In spite of an accident at the age of 15, in which he lost his right hand and forearm, he successfully made fine optics for amateur astronomers. He started an optical workshop in Mittweida, Germany, and his reputation spread. He worked at the observatory in Bergedorf (near Hamburg) as an optician and made observations of Jupiter, Saturn, the Moon and eclipses. During the eclipse expedition of 1929, Schmidt discussed a camera for wide-angle sky photography with WALTER BAADE, and invented the *Schmidt telescope* or *Schmidt camera*. The camera has a spherical mirror and an aspheric corrector plate of half the mirror's diameter. Its optics is almost identical when viewed from over a wide angle and the pictures that it takes are therefore almost uniform in quality. The focal plane is curved, but nevertheless, in the largest Schmidt telescopes (e.g. on Mount Palomar and in Coonabarabran) which have corrector plates of 48 in diameter, 14 in photographic plates can be exposed. These telescopes have been used to survey the entire sky for stars, galaxies and nebulae of different kinds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schmidt, Maarten (1929–)

Astronomer, born in Groningen, Netherlands, worked on galactic dynamics and the mass distribution of the Galaxy with JAN OORT at Leiden and went to CalTech, and became director of the Hale Observatory. In 1963 he observed 3C273, a radio source which had been identified by Cyril Hazard with a blue star by using the edge of the Moon as a scanning device when the Moon occulted the source. With a 200 in telescope picture, he discovered that there was a jet protruding from 'star'. He obtained a spectrum, which like 3C48 studied by ALLAN SANDAGE, showed emission lines at unusual positions. Schmidt's breakthrough was to recognize that the object exhibited ordinary hydrogen lines, but at a redshift far greater than those observed in stars. The distance that was implied meant that the power output of the object was very large, and connected with the jet mechanism. This was the first completed discovery of a quasar. Schmidt went on to investigate the distribution of quasars, discovering that they were much more abundant when the universe was young.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schmidt–Cassegrain Telescope

A form of telescope that combines some of the features of the Schmidt and the Cassegrain systems. The principal optical components of the Schmidt–Cassegrain are a concave primary mirror, a thin corrector lens located close to the focus of the primary, and a convex secondary (which is often attached directly to the inside of the corrector lens), which reflects light through a central hole in the primary to a focus. The curved secondary changes the angle at which light rays from the primary mirror are converging and, as with the conventional Cassegrain, increases the effective focal length of the instrument while keeping down its overall length. The corrector lens gives a wider field of sharp focus than that of the conventional Cassegrain. The Schmidt–Cassegrain is an example of a catadioptric system, one that uses both lenses and mirrors to collect light.

See also: Cassegrain telescope, field of view, focal length, focus, mirror, Schmidt telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schopf, James William (1941–)

Paleontologist, born in Urbana, IL, became professor of paleobiology at the University of California, Los Angeles. He was NASA's principal investigator of lunar samples, and discovered fossil organisms in 3.5 billion-year-old Australian rocks, one of the earliest manifestations of life on Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schröter Effect

A phenomenon in which the observed and predicted phases of Venus do not coincide. It applies in particular to the predicted and observed times of dichotomy (half-phase), at which the terminator—the boundary between the illuminated and unilluminated portions of Venus's disk—should be a straight line. At eastern elongation, when the planet is visible in the evening sky, dichotomy usually comes a day or two earlier than it theoretically should, while at western elongation, when Venus is visible before sunrise, dichotomy tends to occur a day or two later. The effect is named after Johann Schröter, who first described it in 1793. There is no satisfactory explanation for the Schröter effect. It may be purely subjective, or it may have a physical cause, perhaps the scattering of twilight in Venus's thinner upper atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schröter, Johann Hieronymous (1745–1816)

Magistrate of Lilienthal near Bremen, Germany, assiduous mapper of Mars, although he believed that all martian markings were duststorms. The *Schröter effect* in Venus is the gap of a day or two between the time of the geometrical dichotomy of Venus and the time at which Venus actually appears as a half-circle (an effect of Venus' atmosphere). His assistant K L HARDING discovered Juno, and F W BESSEL began his career at his observatory, which, with his records, his wealth and his position, was lost during the Napoleonic war.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schramm, David N (1945–97)

Cosmologist, born in St Louis, MO, studied with WILLIAM FOWLER, became professor at the University of Chicago. He was a founder of the subject of 'astro-particle physics', studying the particle-physics properties of the Big Bang at the start of the universe. When only two families of particles were known, Schramm predicted from cosmological considerations that there would be just one more family, a prediction confirmed in 1989 by experiments at particle accelerators in Stanford and Geneva. He worked on the nuclear physics by which light elements (hydrogen, deuterium, helium, lithium, beryllium and boron) were produced in the Big Bang. Schramm was a large man and a champion wrestler. He was killed piloting his own plane.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schuster, Sir Arthur (1851–1934)

Physicist, born in Frankfurt, Germany, became professor at Manchester. In his work on terrestrial magnetism he developed the *Schuster–Smith magnetometer* for measuring the Earth's magnetic force.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schwabe, Samuel Heinrich (1789–1875)

Born in Dessau, trained as a pharmacist. When he sold the family pharmacy he became financially independent and was able to turn to astronomy, erecting an astronomical observatory on top of his house, from which he made observations of the Moon, the Sun and the planets. Schwabe began a search for the planet Vulcan inside Mercury's orbit, hoping to find it amongst the sunspots that he observed and listed. After 17 years he began to suspect a regular variation in sunspot numbers, with a period of about 10 years. His under-sold article 'Solar observations during 1843' was overlooked until ALEXANDER VON HUMBOLDT drew attention to it in *Cosmos*, his encyclopedia of natural science. The 11 year cycle was confirmed by Julius Schmidt, RUDOLF WOLF, EDWARD SABINE, and J von Lamont, in both the sunspot cycle and in the Earth's magnetic field.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schwarzschild, Karl (1873–1916)

Mathematical physicist, born in Frankfurt am Main, Germany, at first worked on celestial mechanics, including POINCARÉ's theory of rotating bodies, the tidal deformation of moons and LAPLACE's origin of the solar system. He became professor at Göttingen and Potsdam. He wrote on relativity and quantum theory. He early on proposed that space was non-Euclidean, giving a lower limit for the radius of curvature of space as 2500 light years, and in his mature papers gave the first exact solution of EINSTEIN's general gravitational equations about the geometry of space near a point mass. This led to the first relativistic study of black holes. The size of the 'event horizon' within which light cannot escape from a black hole is called the *Schwarzchild radius*. His mathematical formulation of the description of the geometry of space is called the *Schwarzchild metric*. He served in the Army in the First World War, contracting an illness while in Russia from which he died on returning home.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Schwarzschild, Martin (1912–97)

Astrophysicist, born in Potsdam, Germany, the son of KARL SCHWARZSCHILD, left Germany, became professor at Princeton University. Working with John von Neumann, Schwarzschild used the powers of the newly developed electronic digital computers to work on the theory of stellar structure and evolution. He uncovered phenomena in red giant stars, including how they evolve off the main sequence in the Hertzsprung–Russell diagram, how they generate energy in hydrogen shells, the helium flash and the ages of star clusters. He summarized his work in an influential text-book, *Structure and Evolution of the Stars*. In the 1950s and 1960s Schwarzschild pioneered the use of space telescopes for precise imagery of the Sun, planets and stellar systems. Stratoscope I, a 12 in telescope lifted to 80 000 feet altitude by balloon, was the first instrument to obtain sharp photographs of the Sun from above most of the Earth's atmosphere. Stratoscope II, with a 36 in telescope, yielded infrared images and spectra of the outer planets, red giant stars and galactic nuclei.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scintillation

Rapid fluctuations in the radiation from stars due to its passage through the atmosphere. The more familiar term for the effect on visible light is 'twinkling'. The effect is most marked for a point source of light, such as a star. An extended source (covering a finite area) does not exhibit the effect so appreciably (thus a planet tends to twinkle less obviously than a star). Scintillation is caused by turbulence in layers of the atmosphere (which may have differing temperatures, refractive indices, velocities, etc). Seen from space, stars would not appear to twinkle.

At radio frequencies it is possible to detect scintillation due to the passage of radiation through the solar wind. Analogous to the twinkling of stars, the distortion of radio emissions in the ionosphere occurs at wavelengths longer than about 20 cm (1.5 GHz).

See also: seeing.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scorpius

(the Scorpion; abbrev. Sco., gen. Scorpii; area 497 sq. deg.) A southern zodiacal constellation which lies between Ophiuchus and Ara, and culminates at midnight in early June. Its origin dates back to Sumerian times, when it was called Girtab, 'the stinger', but today it is associated with the scorpion that, in Greek mythology, killed Orion the hunter—and the two constellations lie on opposite sides of the sky. Scorpius was originally a larger constellation, which was divided into two by the ancient Greeks, the western part, Chelae (the Claws—of the Scorpion), being renamed Libra by the Romans in the first century BC (*see* Libra). The brightest stars of Scorpius were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*. (An alternative form of the name for the constellation, Scorpio, is used by astrologers.)

A conspicuous constellation and one of the few that even remotely resembles the object after which it was named, Scorpius is easily recognized by α Scorpii (Antares), a red giant irregular variable (range 0.9–1.2), which marks the scorpion's heart, and an arc of second to fourth magnitude stars, ε, μ^1/μ^2, ζ, η, θ, ι^1, κ, λ and υ Scorpii, marking its tail and sting.

α Scorpii is also a close binary, it having a bluish-white (B2.5) companion, magnitude 5.4, separation 2.9″, which appears pale green in contrast with its red (M1) primary. Other bright stars include λ Scorpii (Shaula), magnitude 1.6, θ Scorpii (Girtab or Sargas), magnitude 1.9, δ Scorpii (Dschubba), magnitude 2.3, ε Scorpii, also magnitude 2.3, and κ Scorpii, magnitude 2.4. There are 14 other stars of magnitude 4.0 or brighter, including β Scorpii (Acrab or Graffias), an optical double consisting of two bluish-white (B0.5 and B2) components, magnitudes 2.6 and 4.9, separation 13.7″, the former of which has two companions, magnitudes 5.9 and 4.2, separation 0.5″ and 0.001″ (orbital period 6.83 days), and the latter a companion of magnitude 7.6, separation 0.1″. Another interesting multiple star is ν Scorpii, which has two bluish-white (B2 and B8) components, magnitudes 4.1 and 6.8, separation 41″, each of which has a close companion, magnitudes 6.2 and 7.5, separation 1.3″ and 2.3″, the primary also having two much fainter components, one at a separation of 0.06″ and the other with an orbital period of 5.55 days.

Interesting variable stars in Scorpius include RR Scorpii, a Mira-type star (range 5.0–12.4, period about 281 days), and U Scorpii (range 8.7–19.3), a recurrent nova which is normally about eighteenth magnitude but flared up to ninth magnitude in 1863, 1906, 1936 and 1979. Other interesting objects include the naked-eye open clusters NGC 6231, which contains more than 100 stars (including many high-luminosity O- and B-type supergiants) of mostly tenth to thirteenth magnitude but with a dozen or so brighter than seventh magnitude, including ζ^1 Scorpii (magnitude 4.7), M7 (NGC 6475), which has more than 80 stars, mostly between seventh and tenth magnitude, and M6 (NGC 6405, the Butterfly Cluster), which also contains more than 80 stars, mostly between magnitudes 7 and 11. There are also two bright globular clusters: M4 (NGC 6121), which is sixth magnitude and located about 1.3° west of α Scorpii, and M80 (NGC 6093), which is seventh magnitude. Another interesting object is NGC 6302 (the Bug Nebula), a tenth-magnitude, elongated planetary nebula.

Also in Scorpius is Scorpius X-1, an x-ray binary and the brightest x-ray source in the sky.

See also: Antares, Scorpius X-1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scorpius X-1

The brightest cosmic x-ray source in the constellation of Scorpius and the first cosmic x-ray source to be discovered. Detected for the first time in 1962 by instrumentation carried to an altitude of 225 km by an Aerobee rocket, Scorpius X-1 is, apart from occasional transient sources, the brightest cosmic source of x-radiation in the sky.

It is a low-mass x-ray binary with an orbital period of 0.787 days and an x-ray luminosity of about 2×10^{37} watts (about 5000 times the optical luminosity of the Sun) that is located at a distance of about 2300 light-years. The binary is believed to consist of a low-mass star and a neutron star, the x-radiation being emitted from the surface of the neutron star and from its accretion disk.

See also: accretion disks, binary system, neutron star, x-ray astronomy, x-rays.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sculptor

(the Sculptor; abbrev. Scl, gen. Sculptoris; area 475 sq. deg.) A southern constellation which lies between Cetus and Phoenix, and culminates at midnight in late September. It was introduced as Apparatus Sculptoris (the Sculptor's Workshop) by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2.

An inconspicuous constellation, the brightest star in Sculptor is α Sculptoris, magnitude 4.3. Interesting stars in Sculptor include ε Sculptoris, a binary with pale yellow (F2) and yellow (G9) components, magnitudes 5.4 and 8.8, separation 4.9″, period about 1200 years, and R Sculptoris, a very red (C6) semiregular variable star (range 9.1–12.9, period about 370 days).

Other interesting objects include NGC 288, an eighth-magnitude globular cluster, the Sculptor Dwarf Galaxy (ESO 351-G30), a widely dispersed dwarf spheroidal galaxy in the Local Group, about 260 000 light-years distant, NGC 253 and NGC 55, eighth-magnitude edge-on spiral galaxies (both members of the Sculptor group, which is the nearest collection of galaxies to the Local Group, about 12 million light-years distant), and the Cartwheel Galaxy (ESO 350-G400), a fourteenth-magnitude ring galaxy with faint, radial 'spokes' bearing witness to a former collision.

The south galactic pole is located in Sculptor.

See also: Cartwheel Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Scutum

(the Shield; abbrev. Sct, gen. Scuti; area 109 sq. deg.) A southern constellation which lies to the south-west of Aquila, and culminates at midnight in early July. It was introduced as Scutum Sobiescianum (Sobieski's Shield) by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk) in 1684 in honor of Jan Sobieski III, king of Poland.

A small, inconspicuous constellation, the brightest star in Scutum is α Scuti, magnitude 3.9. There are no other stars brighter than magnitude 4.0. Interesting objects include the variable stars R Scuti (range 4.2–8.6, period about 147 days), an RV Tauri star, and δ Scuti (range 4.6–4.8, period 0.19 day), the prototype of a class of pulsating stars, and M11 (NGC 6705, the Wild Duck Cluster), a fan-shaped open cluster which contains more than 400 stars between eighth and fourteenth magnitude.

The Milky Way passes through Scutum and is particularly bright in the north-east of the constellation, where it is known as the Scutum Star Cloud.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sdO Stars

Subdwarfs of spectral type O (sdO stars) are hot evolved stars with a lower luminosity than main sequence O stars. The sdO stars cover a wide range in the HERTZSPRUNG–RUSSELL DIAGRAM (HRD, see figure 1). At lower luminosities and temperatures they form the extension of the subdwarfs of spectral type B (sdB), bounded by the extreme blue extension of the Horizontal Branch (EHB) and the helium main sequence (He-MS). At higher luminosities the sdOs are found up to the region of the central stars of planetary nebulae (CSPNe). Some of the most luminous sdOs are indeed associated with a planetary nebula.

The existence of the sdO stars is a challenge for stellar evolution theory since their location in the HRD is usually avoided by evolution tracks. Understanding the sdOs stars and inclusion of these stars in the overall picture of stellar evolution requires knowledge of their fundamental stellar parameters, i.e., the effective temperature, the surface gravity and the chemical composition, for a reasonably large sample of sdOs. This requires a search for these rare objects, the analysis of their spectra, and an interpretation of the results in the context of stellar evolution theory.

A further introduction to the evolution of low and intermediate mass stars can be found in the article HORIZONTAL-BRANCH STARS.

Many results for sdO stars are summarized in the proceedings from the Second and Third Conferences on Faint Blue Stars (FBS2, Philip *et al* 1987; FBS3, Philip *et al* 1997). A review on subdwarfs has been given by Heber (1992).

The search for sdO stars

Blue, i.e. hot, stars in the Galactic disk are mainly young main sequence O and B stars. In surveys at high Galactic latitudes, however, hot evolved stars dominate the stellar objects. A large number of surveys have been conducted, which are described in detail in the article BLUE STARS AT HIGH GALACTIC LATITUDES. A list of about 1000 sdO stars is provided by Kilkenny, Heber and Drilling (1987, FBS2). Recently, several sdO stars in globular clusters have been detected by the Ultraviolet Imaging Telescope. In the UV the majority of the cluster stars are invisible and only the UV bright stars like sdOs can be picked out easily.

Spectral analyses of sdO stars

In the last decade many of the sdO stars were analysed and high resolution optical and UV spectra allowed the determination of the chemical composition for selected sdOs. The major uncertainty originates from systematic differences between the results of different working groups, which are not resolved up to now. The latest results for nearly 50 sdO stars were presented by Lemke *et al* (1997, FBS3). They used non-LTE model atmospheres and a χ^2-fitting technique to determine effective temperatures, gravities and H/He abundances.

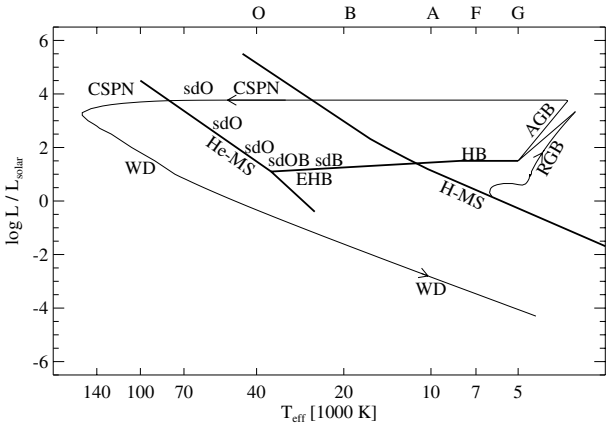


Figure 1. Schematic Hertzsprung–Russell diagram with an evolution track of a 1 $M_\odot$ star. H-MS, He-MS: hydrogen and helium main sequence; RGB, AGB: red and asymptotic giant branches; HB, EHB: horizontal and extreme horizontal branches; CSPN, WD, sdB, sdO: positions of central stars of planetary nebulae, white dwarfs, sdB and sdO stars, respectively.

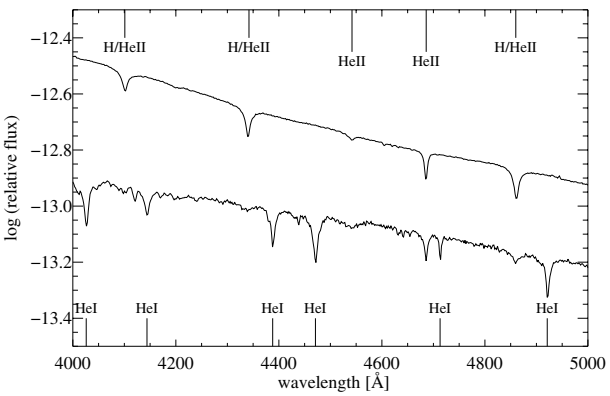


Figure 2. Sample spectra of a hot (80 000 K) sdO with normal H/He composition (top) and a cooler (40 000 K) helium-rich sdO (bottom).

Effective temperatures of sdO stars range from 40 000 K up to 100 000 K, the surface gravities range from $\log g = 4.5$ up to $\log g = 6.5$. The He/H ratio varies from about the solar value up to extreme helium-rich compositions with no hydrogen detectable. Besides these sdO stars there is also a group of helium-poor subdwarfs, mainly below 40 000 K and surface gravities around $\log g = 6$. In the literature these objects are also termed sdOB stars, emphasizing the close evolutionary connection with the sdBs.

Spectral characteristics

The spectral characteristics (see figure 2) of sdO stars in the optical spectrum are the presence of the 4686 Å line of singly ionized helium and the broad and shallow line profiles of the hydrogen Balmer series, which are superposed by the lines from the Pickering series of singly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ionized helium. The cooler members of the sdO group also display several spectral lines from neutral helium. At a spectral resolution of about 3 Å only a few lines from carbon or nitrogen can be detected in some sdO stars.

Model atmospheres for sdO stars
The progress of spectral analyses of sdO stars is very closely connected to the progress of modeling the atmospheres of hot stars. STELLAR ATMOSPHERES of solar type stars can in general be described with the assumption of local thermodynamic equilibrium (LTE) for the calculation of the population of all the atomic energy levels. In hotter stars this population is no longer determined by the local electron temperature but strongly influenced by the radiation field, which is non-local in those regions of the stars where the emergent spectrum is formed. Early analyses using LTE model atmospheres were therefore hampered by a poor reproduction of the observed spectra. The upcoming non-LTE stellar atmosphere models at the beginning of the seventies allowed for the first time a realistic modeling, however, taking into account only a basic model atom for hydrogen and helium. Greenstein and Sargent (1974) used these non-LTE models to determine effective temperatures and surface gravities for a large fraction of the sdOs known at that time. The abundances of heavier elements like carbon, nitrogen and oxygen could be determined only in calculations, where the structure of the atmosphere was kept fixed and the population of the element was iterated. This procedure cannot take into account the influence of the heavier elements on the structure of the model atmosphere. With the launch of the IUE satellite in 1978, access to the UV part of the spectra of sdOs became available and it soon became obvious that these stars exhibit an enormous number of metal lines in the UV. It was therefore clear that these numerous lines must significantly influence the atmospheric structure by blocking the transport of radiation, which can now be taken into account in state-of-the-art non-LTE model atmospheres (e.g. Haas *et al* 1996, Lanz *et al* 1997).

Chemical composition
Constraints for the evolution scenarios might be obtained from metal abundances. Several results are summarized by Dreizler (1993) and Bauer and Husfeld (1995). The latest analysis using HST spectra is presented by Lanz *et al* (1997). The metal abundances also show a large scatter. In general, the helium-rich sdO stars are also enriched in some metals, especially in carbon and nitrogen. This clearly indicates that the surface contains material burned in the CNO cycle as well as the products of the helium burning.

Elements like iron or nickel can be used as indicators for the metallicity since these elements are not produced in low mass stars. However, the Fe/Ni ratio can show deviations from the solar value, which very likely hints at diffusion processes, caused by gravitational settling and radiative levitation. In that case, the element composition would not allow conclusions on the evolutionary status since the abundance pattern can be completely altered by diffusion.

Mass loss
While main sequence O stars are known to exhibit a stellar wind, the atmospheres of most sdO stars can be regarded as hydrostatic. However, signatures of a stellar wind can be detected in the most luminous sdOs through P-Cygni profiles of UV resonance lines or through emission lines in the optical region. But even in these stars all other lines originate from the quasi-static layers of the atmosphere.

Evolutionary status

Even though sdO stars are predominantly found at high Galactic latitudes, investigation of the kinematics of these stars reveals an affiliation with the old disk population. Only a very small minority of the sdOs belongs to the POPULATION II, either through membership of a globular cluster or through a high radial velocity. No reliable scale height nor space density has been determined yet. In addition to spectroscopically determined parameters this would require knowledge of the masses to derive distances and the lifetime of the sdO stage. However, the evolutionary status must be identified beforehand.

Several evolutionary scenarios for single and binary stars exist to explain the sdO stars, which will be discussed in the following. The situation is complicated by the fact that the sdO stars cover a wide range in the HRD and comprise very different chemical compositions. In this context one question requires special attention: How did some sdOs become helium-rich?

Post-AGB scenario
The hottest and most luminous sdOs are found in the region of CSPNe. Therefore, those few sdOs with an associated planetary nebula can be identified as post-AGB stars evolving from the asymptotic giant branch (AGB) towards the white dwarfs. Heber and Hunger (1987, FBS2) suggested that the helium-rich sdOs in the region of the CSPNe may be linked to extreme helium and R CrB stars and could be products of a late helium shell flash scenario in which a star leaves the AGB, becomes a white dwarf and suffers a late helium shell flash. This takes the star back to the AGB and the post-AGB evolution repeats. The time span until the star is in the CSPNe range for the second time is sufficiently long to let the planetary nebula, ejected on first leaving the AGB, disappear.

Post-EHB scenario
The majority of the sdOs are found at the extreme end of the HB along the He-MS. The close proximity in the HRD to the sdB stars suggests an evolutionary connection between these two groups. Indeed, evolutionary tracks leaving the extreme end of the HB, the EHB (the position of the sdB stars) pass through the region in the HRD where sdOs are found. This scenario therefore is a plausible explanation, at least for many of the helium-normal sdOs. The post-EHB sdOs would have a very narrow mass range around $0.5\,M_{\odot}$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It was long discussed how helium-rich sdO could be explained in this scenario, since a transition of a helium-poor sdB into a helium-rich sdO must have occurred then. Convective mixing caused by the helium ionization does not start off as long as the helium abundance is as low as in the sdB stars. A possible solution might be that the hydrogen-rich envelope, which was inert on the EHB, re-ignites during the post-EHB phase followed by excursions towards the AGB. The luminosity is not sufficient to let the star ascend the AGB and it therefore returns back to the sdO region. This excursion might cause deep mixing processes, which would help to explain the transition from hydrogen-rich sdBs to helium-rich sdOs.

UV bright stars in globular clusters have been analysed recently by Moehler *et al* (1998) and were identified as post-EHB or post-early AGB stars, the hotter ones being sdOs with a normal H/He ratio. The detection of sdOs in globular clusters is especially interesting in connection with the lower-than-expected detection rate of CSPNe in globular clusters. This matches the significantly higher population of the EHB where the stars evolve into post-EHB or post-early AGB stars, but cannot evolve to the top of the AGB and become CSPNe.

Delayed helium core flash scenario

The post-EHB scenario requires an exact tuning of the mass loss on the red giant branch (RGB) if the sdB stars, and therefore in the above picture also the sdOs, are explained by single star evolution. Mass loss on the RGB must be sufficiently high to remove most of the hydrogen-rich layer around the helium core at the moment the helium core ignites. Hydrogen layer masses below about $0.02\,M_{\odot}$ are required to have a sufficiently hot star on the HB. If, however, the mass loss on the RGB is too high then the helium core never reaches its ignition mass and the star ends as helium white dwarf without a HB phase. The scenario of the delayed helium flash as presented by Sweigart (1997, FBS3) opens the possibility to explain the cooler and most compact sdOs, especially the cluster of helium-rich sdOs around $T_{\rm eff} \approx 45\,000\,{\rm K}$ and $\log g \approx 5.8$, not covered by post-EHB tracks. If high mass loss terminated the RGB evolution before the helium core ignites, this ignition can still happen when the star is already on the white dwarf cooling sequence. The flash takes the star into the region of the sdOs also covering those compact sdOs which cannot be explained in the post-EHB scenario. Helium enrichment would be the consequence of mixing and burning of the remaining hydrogen-rich layer during the flash.

Subdwarfs and binary star evolution

Besides the evolution of single stars, binary evolution could also produce sdO stars. If two helium white dwarfs in a close binary system merge due to angular momentum loss through gravitational waves, the result could be an sdO star. The problem for this scenario is the observational lack of possible progenitor systems. The long time scale of this evolution would also imply membership of an old population. This scenario might therefore account for individual sdOs but not for the group in total.

Subdwarf O stars in binary systems are a valuable tool to check the evolutionary status since the stellar masses can be determined independently. However, mass transfer before the pre-subdwarf stage has to be excluded in order to test single-star scenarios. As discussed by Heber (1992), this seems not to be the case even in the widest system. On the other hand, the mass transfer is a plausible explanation for the helium enrichment through removal of the progenitor's envelope.

Bibliography

Bauer F and Husfeld D 1995 Metal abundances in subdwarf O stars *Astron. Astrophys.* **300** 481–92

Dreizler S 1993 Spectral analysis of extremely helium rich subdwarf O-stars *Astron. Astrophys.* **273** 212–20

Greenstein J L and Sargent A L 1974 The nature of faint blue stars in the halo II *Astrophys. J. Suppl.* **28** 157–209

Haas S, Dreizler S, Heber U, Jeffery S and Werner K 1996 Iron and nickel abundances of subluminous O-stars. I. NLTE-model atmospheres with line blanketing by iron group elements *Astron. Astrophys.* **311** 669–79

Heber U 1992 Hot subluminous stars in *The Atmospheres of Early-Type Stars, Lecture Notes in Physics 401* (Berlin: Springer) p 233

Lanz T, Hubeny I and Heap S R 1997 Non-LTE line-blanketed model atmospheres of hot stars. III. Hot subdwarfs: The sdO star BD +75°325 *Astrophys. J.* **485** 843–58

Moehler S, Landsman W and Napiwotzki R 199, Hot UV bright stars in globular clusters *Astron. Astrophys.* **335** 510–6

Philip A G D, Hayes D S and Liebert J W 1987 *The Second Conference on Faint Blue Stars, IAU Coll. 95* (Schenectady: L Davis Press)

Philip A G D, Liebert J W and Saffer R A 1997 *The Third Conference on Faint Blue Stars* (Schenectady: L Davis Press)

Stefan Dreizler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Seares, Frederick Hanley (1873–1964)

Born in Michigan, joined the original staff of the Mount Wilson Observatory. He used astrophotography of JACOBUS KAPTEYN's selected areas as part of the worldwide effort to determine the structure of the Galaxy, standardizing the stellar magnitude system. He extended it beyond the eighteenth magnitude by photographing stars of different brightnesses through wire screens and comparing the diffracted images of bright stars with the zeroth-order images of the faint ones. He used the body of data that this work produced to investigate the brightness of the Milky Way and the distribution and properties of its interstellar dust.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Season

Natural environmental and climatic changes occurring as a planet completes one orbit of the Sun. Any planet experiences seasons as long as its rotation axis is not at 90° to the plane of its orbit (the plane of the ecliptic, in the case of the Earth). Seasonal effects are very marked on Earth and Mars, especially at the poles. On Earth the four seasons—winter, spring, summer and autumn—begin in the Northern hemisphere on the winter solstice (December 22 or 23), the vernal equinox (March 20 or 21), the summer solstice, (June 21 or 22), and the autumnal equinox (September 22 or 23). In the Southern hemisphere, summer and winter are reversed, as are spring and autumn.

At the poles, the seasons consist of a short summer and a long winter when there is continuous darkness all winter and continuous daylight or twilight all summer. In low latitudes, where the range of the annual solar radiation and temperature cycle is very small, seasonal changes are based largely on rainy and dry periods.

Outside the tropics and the polar regions, seasons are distinguished by the temperature oscillating between a single maximum and a single minimum. This oscillation results from the annual variation in the angle at which the Sun's rays reach the Earth's surface and the duration of sunlight on the Earth's surface each day. Inclined about 66° 33′ to the orbital plane, the Earth's axis maintains a nearly constant orientation in space as it orbits the Sun. During the six-month half of each orbit when the North Pole is inclined toward the Sun, a point in the Northern hemisphere receives the Sun's rays at an angle closer to 90° than does a point in the Southern hemisphere. The result is greater heating and more hours of daylight in the Northern hemisphere than in the Southern hemisphere. During the following six months the conditions are reversed.

The amount of solar radiation arriving at the surface of a planet is also affected by the eccentricity of its orbit. Because the Earth currently reaches perihelion (closest approach to the Sun) in early January, the input of solar radiation to the Northern hemisphere during Northern hemisphere winter is slightly greater than the input to the Southern hemisphere during that hemisphere's winter. Conversely, the input of solar radiation to the Northern hemisphere in summer is less than to the Southern hemisphere in summer. Furthermore, because the Earth travels slightly faster along its orbit at perihelion than at aphelion, the Southern hemisphere summer is of marginally shorter duration than the Northern hemisphere summer. Because of its greater orbital eccentricity (0.0934, compared to 0.0167 for the Earth), these effects are more marked in the case of Mars.

Because the rotational axes of planets precess (undergo slow periodic changes in their orientation), the times at which planets reach perihelion or aphelion migrate slowly through the seasons; for example, in about 13 000 years' time, the Earth will reach perihelion during the Northern hemisphere summer. Furthermore, because of the gravitational influence of other solar system bodies, the axial inclinations and orbital eccentricities of the planets undergo slow, quasi-periodic changes that alter the input of solar energy at different latitudes during the different seasons and cause climatic changes. According to M Milankovich, the combination of these three cycles leads, from time to time, to a drop in the Earth's temperature that triggers the onset of ice ages. Mars experiences greater changes in axial inclination and orbital eccentricity than does the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Secchi, Angelo (1818–78)

Born in Reggio Emilia, became a Jesuit and was exiled with the rest of the order by Mazzini's Roman Republic. After a period at Georgetown Observatory, he returned to Italy in 1849 as director of the Roman College Observatory, constructing a new observatory dome on top of the main pillars of the incomplete church of Saint Ignazio. Secchi used its telescopes for solar and stellar research, one of the first astronomers to concentrate on physical properties rather than positions. He classified the spectra of over 4000 stars into five classes based on the appearance of the spectra (this work developed and superseded by the more detailed Harvard system of EDWARD PICKERING), and studied the planets. He discovered that Jupiter was gaseous. He drew the dark lines that join areas of Mars, and used the word canali, a direction taken up by SCHIAPARELLI.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Second of Arc

One second of angular measurement. Also known as 'arc second' (abbreviation 'arcsec') and denoted by the symbol $''$, a second of arc is one-sixtieth of a minute of arc (arcmin) which, in turn, is one-sixtieth of a degree. There are 3600 (60×60) seconds of arc in 1 degree and 206 265 seconds of arc in one radian.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Secondary Crater

A small crater formed on a planetary body by a block of ejecta thrown out during the excavation of a larger impact crater, the *primary crater*. Secondary craters from recent impact events, such as those associated with the lunar crater Tycho, tend to be more easily identifiable as secondary features. Some are clustered around the primary crater, and tend to be irregular, having been formed by low-angle, low-velocity impacts. Others, produced by ejecta on higher trajectories travelling faster, are found quite far away and tend to be more circular. They can occur as crater chains, as on the Moon for example (though some crater chains are volcanic in origin) and in the so-called intercrater plains of Mercury. Other bodies on which secondary craters have been identified include Ganymede, associated with the large multiringed impact feature Gilgamesh, and Europa, where they are presumed to have been formed by the impacts that produced large craters such as Pwyll.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Seeing

The sharpness of a telescope image determined by the degree of turbulence in the Earth's atmosphere.

Temperature alters the refractive index of air, so bubbles of air at different temperatures, blown by the wind across the aperture of a telescope, cause a distortion of its image of a star, much like the effect of wrinkled bathroom window glass. Physically the effect of inhomogeneous refractive index corrugates the incoming planar wave front from the star, and its image does not focus to a point (or to a near-point-like diffraction disk). A crucial quality is the size of the bubbles of inhomogeneous air relative to the aperture of the telescope. Small aperture telescopes view the star through not many bubbles and the effect of a single bubble is to tilt the incoming light beam. As the wind blows the bubbles across the telescope, the star image dances in the focal plane. As viewed in a very small telescope such as the eye, the star may twinkle a lot. By contrast, a large telescope views a star image through many bubbles and the image is stationary but diffuse. In the intermediate case, the image shows condensations or 'speckles'—the technique of speckle interferometry constitutes a way to sharpen the image.

Seeing is quoted in arc seconds across a star image. There is no clear convention whether the number quoted is the Full Width at Half Maximum, of the light distribution in the image, or the diameter of the circle enclosing 90% (or 95%, or 99%) of the radiant flux. Eye-estimates of seeing are not very reliable because the eye is a quick-response detector and picks out the brighter core of the image, minimizing the apparent size of the seeing disk.

Seeing is crucial to the performance of a telescope. For this reason, amateurs avoid placing telescopes near sources of heat (such as chimneys and houses) and seek country locations for their observations, and professionals place their telescopes on high mountains in undisturbed air, designing the observatory buildings and telescopes to minimize seeing induced by the equipment ('dome seeing').

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Selene (Selenological and Engineering Explorer)

Japanese lunar exploration project conducted jointly by NASDA and the Institute of Space and Astronautical Science (ISAS). Launch is planned for 2003. Consists of a lunar polar orbiting satellite and a relay satellite. After one year, the polar orbiter's propulsion module will soft-land on the Moon. The major objectives are to obtain data necessary for clarifying the origin and evolution of the Moon, and to develop technologies vital for lunar exploration, such as soft landing technology.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Seleucus of Seleucia (*c.* 190 BC–?)

Greek philosopher, born in Seleucia, on the Tigris River in Babylon. He followed Aristarchus in holding that the Earth rotated on its axis and revolved around the Sun. He correlated the tides of the Indian Ocean with the position of the Moon, theorizing a causal connection from the Moon to the sea through the Earth's atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Self-Organized Criticality and Turbulence in the Solar System and Beyond

Complexity, an ubiquitous phenomenon?

Forest fires, earthquakes, avalanches, solar flares and stellar flares caused by the collapse of ACCRETION DISKS all share common observable features. These systems do not appear to be in a single equilibrium. Rather, they migrate between many metastable states, releasing energy in an irregular, bursty manner. These energy release events, or avalanches, do not have the characteristic mean size, or timescale, that those seen in an equilibrium system would have. Instead their probability density functions are inverse power laws, $P(x) \sim x^{-s}$ where s is the exponent. In consequence, 'extreme' events (with large x) are thus much more probable than they would be in systems in equilibrium which exhibit Gaussian statistics ($P(x) \sim \exp(-x^2/\sigma^2)$).

A recent paradigm is that these non-equilibrium systems are highly correlated. In such complex systems the physics occurring on different scales is strongly coupled. As a consequence we say that the observed behavior is scaling, that is to say it is statistically self-similar, and on the average looks the same on any scale. The corollary of this is inverse power law statistics, and the presence of self-similarity (and possibly fractality, i.e. a non-integer dimension) in the spatiotemporal structure of the events themselves. This key aspect of the phenomenology arises from the nature of the coupling between processes on many scales, rather than from individual processes when understood in isolation. Thus, although these individual processes (say plasma instabilities or the point of ignition in forest fires) may be very different, the way in which processes on different scales couple may share the same mathematical description.

A topical, and in principle generic, paradigm that provides elements of a spatiotemporal description for such complex, highly correlated systems is that of self-organized criticality (SOC), introduced by Bak *et al* (1987). Essentially, such a system is high dimensional, with non-linear coupling between the many nodes of the system. Characteristically this coupling takes the form of a hysteresis. A local variable characterizing, say, stress in an earthquake model, or free energy in a plasma, is driven to increase, but the system does not relax until an (arbitrary) critical threshold, that is a threshold for instability, is exceeded. Once this threshold is exceeded, the system relaxes locally on a fast timescale to a value well below critical. This fast local reconfiguration may trigger instability in its vicinity, thus initiating an avalanche of energy release. The driven system then releases energy in bursts and crucially it does not relax back to an equilibrium ground state but, rather, returns after each avalanche to one of many metastable states. The avalanches then can occur on all sizes supported by the system and have no characteristic scale, instead exhibiting inverse power law statistics.

A related class of systems is that of turbulent fluids. Again, one in principle has a system of equations that specify the dynamics, the Navier–Stokes equation in the case of FLUID DYNAMICS or (in the case of a plasma on sufficiently large spatiotemporal scales) MAGNETOHYDRODYNAMICS (MHD). In practice, once a parameter characterizing the flow is increased beyond a certain threshold, the dynamics becomes highly irregular. In the 1940s Kolmogorov showed from energy and dimensionality considerations that turbulence should have a self-similar energy cascade. The resulting power laws in, for example, the wavenumber spectrum (k spectrum) have exponents which can be obtained from Kolmogorov's theory. A different approach in which there has also been considerable progress is to model the cascade in energy using dynamical systems techniques (see for example Bohr *et al* (1998) and references therein). These seek to describe the turbulent cascade in terms of the non-linear coupling of the different k shells. The relevance of SOC to these approaches is that it may provide insights that will lead to a spatiotemporal theory of turbulence that predicts not only the formation and evolution of spatial structure but also the scaling laws and the observed correlation in both space and time.

Self-organized criticality, the renormalization group approach

An elegant analytical framework for SOC is provided by the renormalization group (RG) procedure (for an introduction see Wilson (1979) and Sethna *et al* (2001); a detailed discussion of the method is given by Cardy (1996)). The key idea here is to consider the effect of coarse graining a physical system. This may be pictured either as examining the system averaged over a progressively longer and longer spatial scale or as stepping backwards from a given system so that the available resolution is progressively reduced. The mathematics of the renormalization procedure ('group' is in some ways a misnomer as the transformation does not have an inverse) essentially embodies a recurrence relation that describes any single step in the above coarse graining of a system. Importantly, the coarse graining may be applied in an abstract system space, where the coordinates are the various values of parameters in a given theoretical model or different models, or in real (configuration) space (see figures 4–6 of Sethna *et al* (2001)). We can specify a path (the renormalization flow) of the system in the corresponding RG space which has coordinates that essentially specify the number of times that the coarse graining operation has been applied. The system may be found to possess a fixed point in this RG space;

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

that is, the dynamics appears to be unchanged as we repeatedly coarse grain. If the renormalization is in real space, the implication of such a fixed point is that the dynamics must then be the same on all spatial scales under suitable rescaling; hence the dynamics is self-similar. In this case the RG will predict the exponent of the inverse power law statistics of event size and so forth (for an example applicable to THE MAGNETOSPHERE OF THE EARTH, see Tam *et al* (2000)). If the renormalization is in system space, the fixed point provides a fundamental prediction of the universality of the system; the different values of the many available parameters produce the same dynamics in the repeatedly coarse grained system and thus, although the microphysical processes may be very different, the macroscopic physics will be the same. An example here is the identical exponents describing the self-similar fluctuations in a liquid–gas critical point and a ferromagnetic critical point (Goldenfeld 1992). The RG thus provides the theoretical basis for understanding why universality, and self-similarity, can occur. Although it may appear that the microphysics has been 'lost' in this analysis this is not the case; instead, the physics on all scales is embodied in the recurrence relation that relates the physics on one scale to that of the next, more coarse-grained scale.

Renormalization (successive coarse graining) in configuration space of SOC models and some natural systems has revealed that the dynamics is governed by a fixed point and in this sense such systems are indeed critical. As a consequence, in the vicinity of the fixed point the avalanches have no characteristic scale, exhibiting inverse power law statistics with characteristic exponents. If the renormalization flow in system space is always towards the fixed point the system self-organizes towards criticality with weak dependence on tuning parameters (see also Jensen 1998) and independently of its starting point in the system space. In such a system, details of the microphysics, such as thresholds for local instability, would not strongly affect the macroscopic physics. Thus if a given system is tractable to RG approaches it is possible to analytically establish whether the system is in SOC and to predict the characteristic exponents of event statistics.

Unfortunately, many physical systems, and in particular the hierarchy of plasma equations, are so far intractable to the RG. An approach is then to construct a simple analytical model for the system (a sandpile or 'avalanche' model, say) that preserves the conservation properties and symmetries of the physical system of interest but which is also tractable to the RG.

Numerical approaches

The intermediate step is to construct simple numerical models which mimic the physical system and which are computationally tractable, that is they can be run on a sufficiently large spatial grid, and for a sufficient time, to test for self-similarity, i.e. to establish power law statistics in events. For a discussion of numerical, as well as analytical, approaches to SOC see Jensen (1998) and for a review relevant to magnetospheric physics see Chapman and Watkins (2001). Once the dynamics of these simple systems has been understood, one can then in principle use RG to establish whether the simple model is in the same universality class as the physical system of interest. In practice a more readily achievable goal is to use real space RG to obtain the scaling predicted by the model. This can then be directly compared with observations of the physical system. Simple numerical models also allow the effect of finite-sized systems on the dynamics to be explored; the real space RG naturally only treats systems where length scales can be taken arbitrarily large (see, for example, Tam *et al* (2000)).

The numerical sandpile approach parallels that of shell models for turbulence, where numerical models (difference equations) in k space describe an energy cascade from one k mode to the next; these are constructed to preserve the symmetries and conservation properties of the Navier–Stokes equations (see Bohr *et al* (1998) for a discussion). SOC models may also be contrasted with other numerical schemes where difference equations locally couple the nodes (typically spatial coordinates at which fields are defined) on the grid. Coupled map lattices use chaotic maps to specify the local spatiotemporal evolution of the system; these systems can exhibit self-similarity but require fine tuning of the parameter of the map (for a review see Kaneko (1993)). The perhaps more familiar example is where the local difference equations are obtained directly from the governing equations of the system (say the MHD equations) essentially by Taylor expansion. Provided that the difference scheme captures the essential physics of MHD, and a sufficiently large computational grid is used, these can generate self-similar dynamics in the form of turbulent cascades (for a recent example see Müller and Biskamp (2000)). Directly physically motivated simulations of this type can provide insight into the spatiotemporal structure and in particular the role of topology in an evolving turbulent system. Unless tractable to the RG, they cannot establish whether the self-similarity found in the simulation is a universal property of turbulence or is unique to the particular conditions and geometry of the simulation.

Solar system plasmas: complicated or complex?

Observational motivation for SOC in the solar system and beyond currently centers around evidence of self-similarity and also evidence of intermittent, bursty anomalous transport (seen also in artificial plasma confinement systems such as Tokamaks, see Chapman *et al*

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

(2001)). A significant difficulty is that establishing a power law in the statistics of events requires large datasets with a dynamic range spanning several orders of magnitude. Nevertheless, self-similar statistics of bursty energy release have been found in x-ray flux from solar flares over several orders of magnitude (see, for example, Crosby *et al* (1993) and references therein); this prompts comparisons with simple avalanche models (see, for example, Lu and Hamilton (1991)). Self-similarity, and, by implication, SOC, has also been inferred from the power spectra of stellar accretion disks (e.g. Mineshige *et al* 1994). In the case of the Earth's magnetosphere, the challenge is to find an observable that unambiguously contains information on all the avalanches hypothesized to be occurring in the system and an attempt has been made with images of patches of activity in the AURORA (Lui *et al* 2000)). This followed the observation of power law power spectra of indices (AE – auroral electrojet) which characterize magnetic perturbation underneath the auroral oval. However, power laws are not a unique property of SOC (for a review of other processes see Sornette (2000)). In particular, power law power spectra are consistent with a colored noise process (for a magnetospheric example in the context of chaos see Shan *et al* (1991)).

Progress is thus needed on two fronts. First, a combination of numerical experiments and RG analysis is needed to make predictions of quantitative observables such as the exponents of inverse power law statistics of events. The theory needs to establish that non-trivial self-similarity arises from criticality as a natural emergent property of the multiscale plasma process at hand. Second, unambiguous observables with sufficient dynamic range are needed to test the prediction of scaling behavior. The promise of SOC is that it may provide a unifying theory of anomalous plasma transport in natural astrophysical plasma confinement systems.

Bibiliography

Bak P, Tang C and Wiesenfeld K 1987 *Phys. Rev. Lett.* **59** 381

Bohr T, Jensen M H, Paladin G and Vulpiani A 1998 *Dynamical Systems Approach to Turbulence* (Cambridge, UK: Cambridge University Press)

Cardy J 1996 *Scaling and Renormalization in Statistical Physics* (Cambridge: Cambridge University Press)

Chapman S C and Watkins N W 2001 Avalanching and self organised criticality: a paradigm for magnetospheric dynamics? *Space Sci. Rev.* **95** 293 (see also other reviews in this issue)

Chapman S C, Dendy R O and Hnat B 2001 *Phys. Rev. Lett.* **86** 2814

Crosby N B, Aschwanden M J and Dennis B R 1993 Frequency distributions and correlations of solar x-ray flare parameters *Sol. Phys.* **143** 275

Goldenfeld N 1992 *Lectures on Phase Transitions and the Renormalisation Group* (Perseus Publishing)

Jensen H J 1998 *Self Organized Criticality* (Cambridge: Cambridge University Press)

Kaneko K 1993 *Theory and Applications of Coupled Map Lattices* (New York: Wiley)

Lu E T and Hamilton R J 1991 Avalanches and the distribution of solar flares *Astrophys. J.* **380** 89

Lui A T Y, Chapman S C and Liou K *et al* 2000 Is the dynamic magnetosphere an avalanching system? *Geophys. Res. Lett.* **27** 911

Mineshige S, Takeuchi M and Nishimori H 1994 Is a black hole accretion disk in a self-organised critical state? *Astrophys. J.* **435** 125

Müller W C and Biskamp D 2000 Scaling properties of three dimensional MHD turbulence *Phys. Rev. Lett.* **84** 475

Sethna J P, Dahmen K A and Meyers C R 2001 Crackling noise *Nature* **410** 242

Shan L-H, Hansen P, Goertz C K and Smith R A 1991 Chaotic appearance of the AE index *Geophys. Res. Lett.* **18** 147

Sornette D 2000 *Critical Phenomena in the Natural Sciences* (Berlin: Springer)

Tam S, Chang T, Chapman S C and Watkins N W 2000 Analytical determination of power law index for the Chapman *et al* sandpile (FSOC) analog for magnetospheric activity–renormalization group analysis *Geophys. Res. Lett.* **27** 1367

Wilson K G 1979 Problems in physics on many scales of length *Sci. Am.* **241** 140

Sandra C Chapman and Nicholas W Watkins

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Separation

The angular distance between two objects, measured in arc units on the celestial sphere. This is most commonly used in measurements of double and multiple stars, and together with the position angle defines the observed form of the system. It is also used to determine the observed distance of a satellite from its primary planet.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Serpens

(the Serpent; abbrev. Ser, gen. Serpentis; total area 637 sq. deg.) A constellation that is unique in that it is divided into two: Serpens Caput (the Serpent's Head; area 429 sq. deg.), which lies mostly in the northern sky between Böotes and Hercules, and culminates at midnight in mid-May, and Serpens Cauda (the Serpent's tail; area 208 sq. deg.), which lies mostly in the southern sky between Ophiuchus and Scutum, and culminates at midnight in late June. It represents the huge snake held by Ophiuchus, the Serpent-bearer, around whose body it is coiled. The brightest stars of Serpens were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A rather inconspicuous constellation, the brightest stars in Serpens are α Serpentis (Unukalhai), magnitude 2.6, η Serpentis, magnitude 3.2, μ Serpentis, magnitude 3.5, and ξ Serpentis, also magnitude 3.5. There are four other stars brighter than magnitude 4.0. Interesting stars in Serpens include θ Serpentis (Alya), a wide double with white (A5) components, magnitudes 4.6 and 5.0, separation 22″, that have the same proper motion, δ Serpentis, a close binary with white (A9 and A7) components, magnitudes 4.2 and 5.3, separation 4.0″, R Serpentis, a Mira-type variable (range 5.2–14.4, period about 356 days), and W Serpentis, an Algol-type eclipsing binary (range 8.4–10.2, period 14.15 days).

Other interesting objects in Serpens include M5 (NGC 5904), a sixth-magnitude globular cluster, and IC 4703 (the Eagle Nebula), an emission nebula which contains the open cluster M16 (NGC 6611) of more than 60 stars fainter than eighth magnitude.

See also: Eagle Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SETI Institute

The SETI Institute, founded in 1984, is a private non-profit center for research and education related to the scientific search for extraterrestrial life.

The SETI Institute currently encompasses more than 30 research projects. The most visible of these is Project Phoenix, the world's most sensitive search for extraterrestrial intelligence ('SETI'). Begun in 1995 at the Parkes 210 ft radiotelescope in Australia, Phoenix moved to the 140 ft telescope at the National Radio Astronomy Observatory in Green Bank, West Virginia in 1996. From fall of 1998 through the summer of 2003, observations will take place at the 1000 ft Arecibo radiotelescope in Puerto Rico. Project Phoenix is a targeted search which scans billions of frequency channels for narrow-band continuous wave and pulsed signals from nearby Sun-like stars. It commands slightly more than half of the Institute's total operation, and is privately funded. The Institute is now developing and prototyping new system designs for next-generation SETI.

The balance of research projects at the SETI Institute fall within the Institute's Center for the Study of Life in the Universe. These projects investigate the origin and evolution of life on Earth and possibilities for life elsewhere. Most of this work is currently funded through competitively won peer-reviewed government research grants.

For further information see
http://www.seti.org.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Severin [Soerensen, Longomontanus], Christian (1562–1647)

Danish astronomer, born in Longberg (from which his Latin name), Jutland, studied astronomy under TYCHO BRAHE, became professor of mathematics and astronomy at Copenhagen. His work, *Astronomia Danica*, was effectively Tycho Brahe's legacy, and resolutely maintained the geocentric theory of the planets in the Tychonic version (the Moon and Sun revolving round the stationary Earth and the remaining planets revolving around the Sun). His only concession to COPERNICUS was to introduce the concept of the rotation of the Earth on its axis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sextans

(the Sextant; abbrev. Sex, gen. Sextantis; area 314 sq. deg.) An equatorial constellation which lies between Leo and Hydra, and culminates at midnight in late February. It was introduced as Sextans Uraniae (Urania's Sextant) to commemorate the astronomical sextant (which is larger than the later marine sextant) by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* of 1687.

A small, insignificant constellation, the brightest star in Sextans is of magnitude 4.5. Interesting objects include NGC 3115 (the Spindle Galaxy), a ninth-magnitude lenticular galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Seyfert Galaxies

In many respects Seyfert galaxies are similar to any other galaxy, except their nuclear regions show unusual properties: they can be much brighter and their spectra show strong emission lines of high excitation. Such nuclei are called 'active', and so Seyfert galaxies belong to the wider category of 'ACTIVE GALAXIES' or, more specifically, they have 'ACTIVE GALACTIC NUCLEI' (AGN). Seyferts were, in fact, the first type of active galaxy to be discovered.

In the zoo of AGNs the power of the nuclear activity spans a wide range. Within this zoo, Seyferts show intermediate levels of activity, being more powerful than LINER galaxies but less powerful than QUASARS. Seyferts also join the majority of active galaxies in being 'radio-quiet', with radio luminosities 10^3–10^4 times weaker than the 'radio loud' category, which includes RADIO GALAXIES and radio quasars.

Over the past 20 years much effort has gone into finding Seyfert galaxies amongst the more normal galaxy population. Efficient ways to detect Seyferts include looking specifically for galaxies with bright UV emission, bright x-ray emission, or unusual far-infrared colors, since these are common characteristics of Seyferts. The fraction of galaxies with active nuclei depends on the level of activity: moderate luminosity Seyferts make up $\sim$ 1–2% of nearby bright galaxies, lower luminosity Seyferts make up $\sim$ 10%, while LINERs make up $\sim$ 40%. Currently, about 800 Seyferts are known, though the number continues to rise.

From a wide range of studies, a picture has emerged of the structure of the central regions of Seyferts. Briefly, the gravitational field of a supermassive (10^6–10^9 $M_\odot$) black hole provides the ultimate energy source, as gas falls inwards via a luminous accretion disk (see SUPERMASSIVE BLACK HOLES IN AGN). A small region of fast moving clouds surrounds this central engine, spanning a few light days or weeks, while a larger region of slower moving clouds extends out a few hundred to a few thousand light years. In some cases, oppositely directed collimated jets of low density gas plow into the surrounding galaxy, accelerating the interstellar medium and creating jet-like or linear radio structures. In at least some Seyferts, dense gas and dust shroud the innermost regions, hiding them from our direct view, though the nuclear radiation escapes in other directions to light up gas and dust residing much further out in the galaxy.

The conditions that give rise to Seyfert activity are still under debate. While Seyferts are found in a wide range of galaxy types, they are more commonly found in reasonably luminous early-type spirals (e.g. types S0/a, Sa, Sb in the Hubble classification scheme). It seems, therefore, that prerequisites favoring Seyfert activity include a massive galaxy bulge and the presence of an interstellar medium. There is also evidence that Seyferts prefer galaxies with an asymmetric gravitational field arising from a bar or other distortion, possibly induced by a nearby companion galaxy. Such distortions are thought to be instrumental in funneling gas down to the nuclear regions where it can then fuel the central black hole.

Discovery, classification and detection

The early identification of nearby Seyfert galaxies rested on two observational properties—their unusual nuclear spectra and their unusual nuclear brightness. While studying galaxy spectra for his PhD work in 1908, Fath noted strong emission lines in the spectrum of the nucleus of NGC 1068. These observations were confirmed in 1917 by Slipher and by 1926 Hubble had added two more galaxies with similar spectra: NGC 4051 and NGC 4151. In 1932 Humason noted that NGC 1275 had a bright starlike center, but it wasn't until 1943 that CARL SEYFERT recognized a distinct class of galaxies with unusually bright and concentrated nuclei, and studied in detail six of the 12 then known cases (NGC 1068, 1275, 3516, 4051, 4151, 7469), leading to the group's name: Seyfert galaxies (see figure 1). Seyfert's spectra showed emission lines which were not only unusually strong, but also unusually wide. Interpreting the width of the spectral lines as Doppler shifts caused by motion of the ionized (line emitting) gas implied gas velocities in some objects of several thousand km s^{-1}, far higher than velocities found in normal galaxies. A further subtlety Seyfert noticed was that in some objects the hydrogen lines were broader than the other lines.

Little work was done until the 1960s, when it became clear that Seyfert galaxies shared a number of properties with the then recently discovered quasars. At this time, quasars posed a serious theoretical puzzle. While their high REDSHIFTS implied great distance and therefore high luminosity, their brightness had been found to vary from month to month, indicating that the region producing all this energy was only a few light months across—tiny compared to the rest of the galaxy. The puzzle of how so much energy could emerge from so small a volume was sufficiently acute that some astronomers even questioned whether the quasar redshifts were true indicators of distance, leading to the so-called 'redshift controversy'. Although this controversy has since been resolved in favor of large distances it might never have been as problematic if Seyfert galaxies had been better studied in the early 1960s, since they are clearly low redshift counterparts to the quasars. Not only are their spectra basically similar, but in 1968 NGC 4151 was found to vary in brightness, confirming a highly compact nuclear energy source. As more Seyferts and quasars were discovered, their separation in luminosity narrowed and finally even overlapped. Continuity had been established and, for most astronomers, the reality of the quasar distances and luminosities was no longer in doubt.

In 1974, Khachikian and Weedman identified two types of Seyfert galaxy on the basis of the widths of the nuclear emission lines. Figure 2 shows example spectra of each Seyfert type. While spectra of type 2 Seyferts have a single set of relatively narrow emission lines, in the spectra of type 1 Seyferts the hydrogen and helium

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

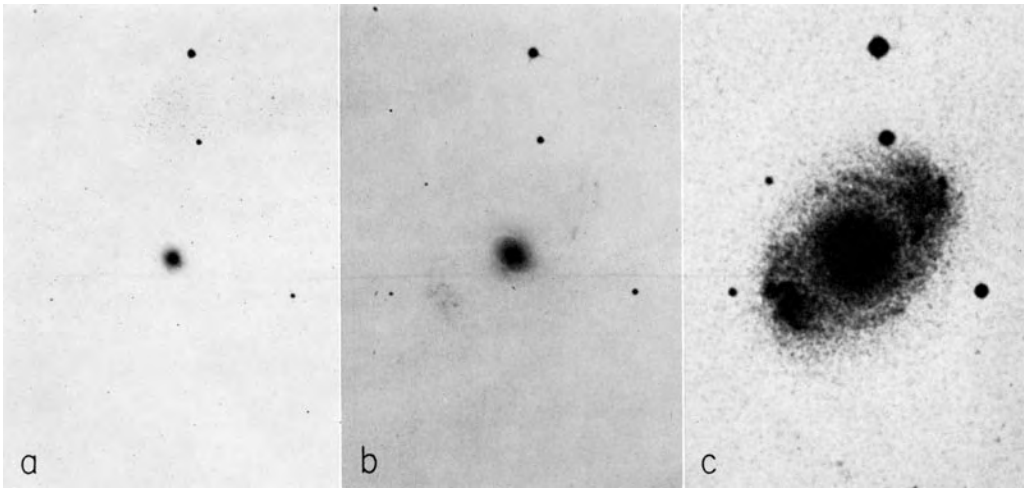


Figure 1. Three photographs of NGC 4151 taken with increasing exposures (a, b, c) showing its unusually bright nucleus—an important characteristic used in the early identification of Seyfert galaxies. (Taken from Morgan W W 1968 *Astrophys. J.* **153** 27)

lines have an additional much broader component. In the simplest cases, the broad component is either absent (Seyfert 2) or strong and dominant (Seyfert 1). With better data it became clear that there is a wide range in the relative strength of the broad and narrow emission lines, and this led Osterbrock in 1981 to refine the Seyfert classification by introducing intermediate types. A good place to establish the classification is with the hydrogen Hβ line at rest wavelength 4861 Å. As the broad component of Hβ becomes weaker relative to the narrow component, the Seyfert type changes from 1 to 1.2 to 1.5 to 1.8. For Seyfert 1.8 galaxies, weak broad wings are just visible at the base of Hβ while for Seyfert 1.9 galaxies they are only visible on the Hα emission line at 6563 Å. In practice, these Seyfert sub-types have not been formally defined but instead give an overall indication of the degree to which the broad component is present.

Although there are a number of Seyfert properties that depend on Seyfert type, perhaps the most noticeable is the 'white light' continuum coming from the nucleus. In normal galaxies the nuclear continuum is mostly starlight, but in active galaxies there can be additional emission, often blue in color and with few or no absorption features. This component is much stronger in Seyfert 1s, weaker in Seyfert 2s, and may even be undetectable in the lowest luminosity Seyfert 2s. Although we refer here to the optical continuum emission, this component extends into the UV and x-ray, and so emission in these spectral bands can also be quite different for Seyfert 1s and Seyfert 2s.

Bearing in mind the principal features which distinguish Seyfert spectra from normal galaxy spectra—strong broad emission lines and a blue continuum—we can now follow the history of how Seyferts have been identified over the past 40 years.

As outlined above, the first Seyferts were discovered essentially by accident when a galaxy spectrum was noted as being unusual. The first survey which identified many Seyferts was that of Markarian and collaborators at the Byurakan Observatory in Armenia, during the period 1962 to 1981. This survey used a technique known as Objective Prism Spectroscopy to find galaxies with unusually blue continuum emission. The telescope was a 1.3 meter Schmidt telescope with a thin prism bolted over the entrance aperture. This combination gives relatively wide field photographs in which each star and galaxy appears as a small spectrum. In this way, the spectra of many galaxies can be studied quickly, and those which appear unusual noted and listed. The survey covered $\sim$ 10 000 square degrees and yielded 1500 galaxies with blue continua, of which $\sim$ 10% were Seyferts (the remainder were mainly STARBURST GALAXIES, blue because of the presence of young massive O and B type stars). Because of the emphasis on blue continuum, the Markarian Seyferts contain a relatively high proportion of Seyfert 1s (about 50%) and rather luminous Seyfert 2s.

Since x-ray emission is common in active galaxies, x-ray surveys provide an effective means to find Seyferts. Starting in the late 1970s surveys have included those of the Ariel V, HEAO-1, Einstein and ROSAT satellites. Some of these were all-sky surveys, some included deeper pointed observations, some were sensitive to hard x-rays (2–10 keV) and others to soft x-rays ($\sim$ 0.2–4 keV). In general, x-ray surveys tend to find Seyfert 1s since their nuclear continuum emission is stronger than in Seyfert 2s.

In 1982, the IRAS satellite provided a huge list of galaxies which were relatively bright in the far infrared (25–100 μm). Although bright infrared flux by itself is no guarantee of Seyfert activity, the 'color' of the infrared emission can help distinguish between Seyfert and normal galaxies: normal galaxies have colder dust ($\sim$ 30 K) than Seyfert galaxies ($\sim$ 100–300 K) in which nuclear activity provides an additional heating source. Thus, IRAS-selected galaxies with 'warm spectra' have yielded many Seyferts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

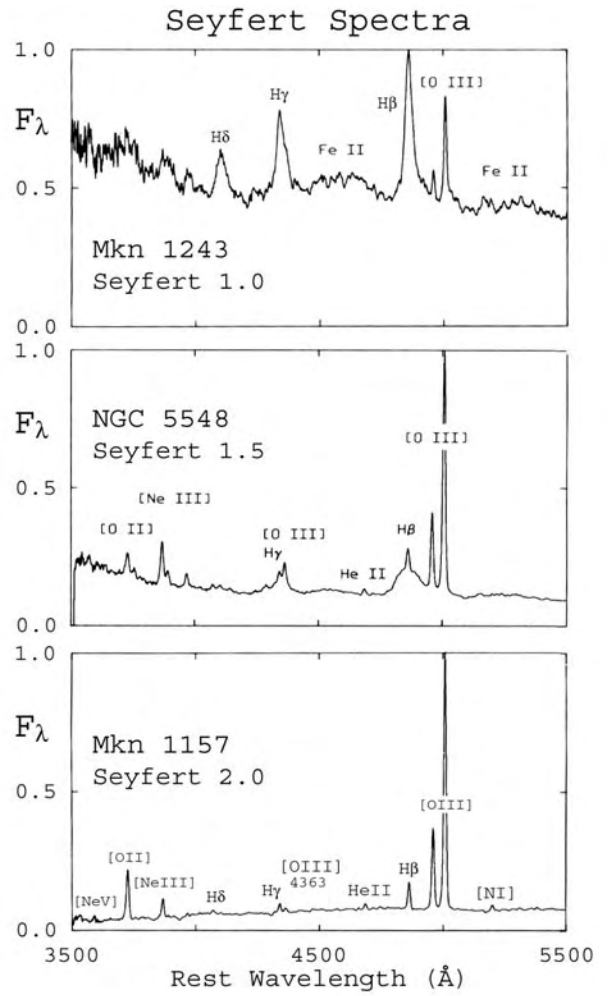


Figure 2. Spectra of three Seyfert galaxies illustrating the difference between Seyfert 1, 1.5 and 2 classes. Principal emission lines are labeled, with square brackets indicating forbidden lines. (Adapted from Osterbrock D E 1984 *Quart. J. R. Astron. Soc.* **25** 1.)

Since Seyferts belong to the radio-quiet group of active galaxies, radio surveys at high flux levels tend not to find Seyferts, but rather radio galaxies. Deeper radio surveys, however, do pick up more Seyferts but along with many star-forming galaxies. On further study, the presence of a compact core with flatter spectral index increases the chance that the galaxy is a Seyfert.

Perhaps the most reliable method of finding Seyferts, though also the least efficient, is optical spectroscopy of galactic nuclei. In the early 1990s the CfA redshift survey of about 2000 galaxies brighter than about magnitude 14–15 found about 2.5% to be Seyferts with an even mix of Seyfert 1s and 2s. Using smaller apertures and higher quality data, the deep Palomar survey of the nuclei of nearly 500 bright galaxies from the RSA catalogue yields about 10% Seyferts, including many of lower luminosity. In the near future, the SLOAN DIGITAL SKY SURVEY will surely find many more.

A catalog of all known Seyferts has been maintained by Veron-Cetty and Veron, and at its last edition (1998) about 790 Seyferts were listed with redshifts < 0.1 and brighter than 17^{m} in V.

Overall structure and geometry

In this section we give a brief outline, without justification, of the basic structures found in Seyfert galaxies; brief because similar material is given elsewhere in this encyclopedia and also because more detailed descriptions of some of the structures are given in subsequent sections.

Figure 3 gives a cartoon of the structures nested from innermost to outermost. The region sizes probably scale with luminosity, larger for quasars and smaller for low luminosity Seyferts and LINERs. The values given are rough guides for the case of moderate luminosity Seyferts.

The ultimate origin for the energy which drives nuclear activity is thought to lie in the gravitational field of a central supermassive black hole of mass $\sim 10^7$–$10^8\,M_\odot$. Gas falls into the black hole via an accretion disk, which radiates powerfully across much of the electromagnetic spectrum. The disk may span from a few Schwarzschild radii to a few thousand Schwarzschild radii, probably with x-ray emission coming from the inner parts, UV and optical from further out. It is this radiation which floods out into the galaxy and ionizes any surrounding gas, causing it to radiate emission lines. The innermost region of ionized gas resides a few light days from the center, and contains gas with density $\sim 10^9$–10^{12} atom cm^{-3} moving with speeds of a few thousand km s^{-1}. Doppler shifts therefore lead to broad emission lines—hence the name: Broad Line Region (BLR). Outside the BLR, on scales from one to a few pc lies thick dense gas, possibly in the form of a torus. This gas blocks radiation coming from the innermost regions, which can only escape out of the equatorial plane along the torus axis. This partially collimated radiation field emerges further into the galaxy to ionize gas residing in a region out to a few kpc. In this region, the gas has a lower density $\sim 10^2$–10^6 atom cm^{-3} and moves with speeds of a few hundred km s^{-1}, leading to narrower emission lines—hence the name: Narrow Line Region (NLR). Because of the partially collimated radiation field, the NLR often appears elongated or even biconical, with axis matching that of the central radiation field. When a radio source is present it is often elongated, aligned with the NLR axis, and has approximately the same size. The radio source emerges from the inner regions as a bipolar flow, accelerating and possibly shocking some of the gas in the NLR. The entire NLR and associated radio structure usually fill a region a few kpc across within the bulge of the host galaxy, but randomly aligned with respect to obvious galactic structures and orientation. If there is a more extended gaseous component, it can be illuminated by the central ionizing radiation field, leading to faint extended emission, sometimes showing conical or biconical form. This is the so-called Extended Narrow Line Region (ENLR).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

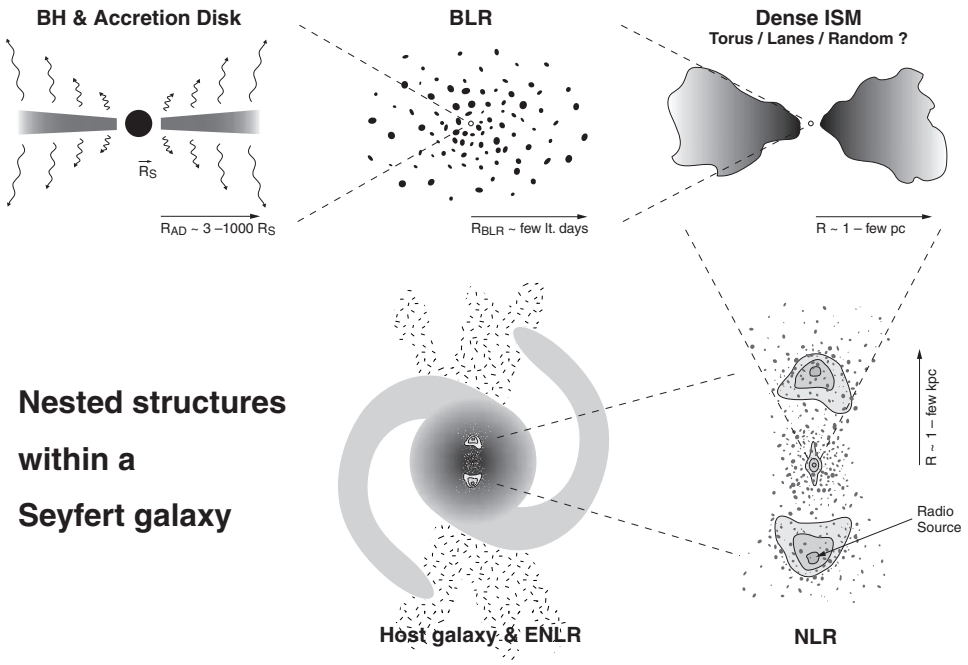


Figure 3. Cartoon illustrating the nested structures thought to be present within Seyfert galaxies.

Inherent in this picture is a lack of spherical symmetry, rooted in the anisotropic emergence of both radiation and radio emitting flows. The origin of the anisotropy might arise in either of two ways: (a) an intrinsic nuclear axis perhaps defined by the black hole spin and/or the accretion disk, and (b) partial obscuration by dense gas which blocks radiation in some directions but not in others. For Seyferts, partial obscuration is thought to be important. The simplest geometry envisioned is a thick disk or torus a few parsecs in size and aligned with the central engine axis. More recently, images from the Hubble Space Telescope (HST) have shown that dust lanes and filaments are common in the central few hundred parsecs, particularly in Seyfert 2s, raising the possibility that less coherent obscuration on these larger scales might also be important. Either way, the fact that the nuclear radiation can only emerge along particular lines of sight raises an interesting possibility which was first noted in 1978 by Osterbrock: namely, that Seyfert 2s might actually *be* Seyfert 1s but oriented in such a way that our view to the central region is blocked. Stated slightly differently, the same Seyfert galaxy might appear as a Seyfert 1 from some directions and a Seyfert 2 from others. This constitutes a 'Unification' scheme, since apparently different classes of objects are 'unified' into a single class. The important topic of unification arises in a number of contexts in the study of active galaxies and is discussed at length in ACTIVE GALAXIES: UNIFIED MODEL. Here we give a brief outline of the evidence supporting the particular case that many (though perhaps not all) Seyfert 2s are in fact hidden Seyfert 1s.

First, the emission line regions of Seyfert 2s, on both large scales (ENLR) and small scales (NLR), are often elongated, with a subset showing sharp edges and even conical or biconical form. These patterns are consistent with dense nuclear gas casting shadows in a strong central ionizing radiation field. When visible, the cone opening angles are $\sim 40°$–$80°$, suggesting $\sim 25\%$ of the surrounding sky sees the nuclear regions. Good examples of this are found in NCG 1068, NGC 5252, MKN 78, MKN 573.

Extending this topic further, one can use spectroscopy of the ionized gas to show that *it* sees a significantly brighter central radiation source than *we* see. There are two approaches. In the first, emission line strengths from extended gas are used to infer both the gas density and its degree of ionization. Knowing the distance between this gas and the central source then allows one to estimate the (ionizing) luminosity of the central source. This luminosity is often significantly greater than the luminosity we see directly, along our line of sight. The second approach uses the total Hβ luminosity coming from all regions to estimate the total ionizing luminosity of the central source. This luminosity is again often found to be significantly greater than the luminosity we infer from our own line of sight to the nucleus.

The third and perhaps most dramatic piece of evidence involves taking spectra of light that was initially moving out of the nucleus in one direction but then got reflected into our line of sight. This provides a sort of periscope, allowing one to look around a corner and into the nucleus from a different direction. The mirrors of the periscope are dust particles or electrons in the surrounding gas which scatter light in all directions, some

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of which comes towards us. Since the scattered light is so faint compared to the normal light we need a way to isolate it. Fortunately, the scattering process also induces polarization, and so a spectrum of the polarized light shows just the scattered component. Remarkably, for many of the brighter Seyfert 2s these polarized light spectra show Seyfert 1 form: strong non-stellar continuum and broad hydrogen and helium lines. The classic discovery was for NGC 1068 by Antonucci and Miller in 1983, but since then many other examples have been found.

A fourth approach is to compare the nuclear x-ray emission of Seyfert 1s and 2s. Consistent with unification, soft x-ray spectra show significant absorption towards Seyfert 2s but little or none towards Seyfert 1s. Furthermore, at higher x-ray energies where absorption isn't important, the spectra of Seyfert 1s and 2s are often quite similar, with ratios of x-ray to optical flux much higher for the Seyfert 2s, as expected if the optical emission is blocked but the hard x-rays are not.

Finally, there is some evidence for the nuclear obscuring material itself. When water masers are detected, they tend to be in Seyfert 2s. These sources form in dense warm molecular gas with long paths in our line of sight, consistent with an edge-on disk geometry. Also, infrared spectra are often dominated by thermal emission from warm dust. Although some of this originates on NLR scales, some comes from much smaller scales. The fact that the total luminosity of many Seyferts, particularly Seyfert 2s, is dominated by the thermal infrared is consistent with the obscuration picture: much of the nuclear luminosity is absorbed by gas and dust whose principal mode of cooling is in the far infrared.

Assuming, for the moment, that all Seyfert 2s contain hidden Seyfert 1s, then we can ascertain the extent of the obscuration by comparing the relative number of Seyfert 2s (obscured) with Seyfert 1s (unobscured). In practice, it is important (and difficult) to select an unbiased sample of Seyferts using a property that is *not* thought to depend on orientation. Possibilities include radio flux, far-infrared flux, host galaxy magnitude, and narrow emission line flux, though none are ideal. With some uncertainty, the ratio of Seyfert 2s to Seyfert 1s found in such samples is roughly 2–5 to 1, suggesting $\sim$ 20–30% of the sky sees an unobscured view of the nucleus, consistent with other estimates based on the opening angle of emission line cones.

Whether *all* Seyfert 2s contain hidden Seyfert 1s is still an open question. Arguments against full unification of the two classes focus either on the absence of expected properties (for example, most Seyfert 2s have lower polarization than expected), or the fact that properties which should be the same in Seyfert 1s and 2s are not (for example, host galaxy morphology, and possibly radio luminosity).

Relation of Seyferts to other AGN

To the newcomer, it may seem that there are an annoyingly large number of names for different classes of active galaxy, of which 'Seyferts' are only one class. This situation has arisen for a couple of reasons: (a) nuclear activity seems to have some genuine range of character, and (b) historical development is never clean and different approaches have naturally focused on different manifestations of activity. In hindsight, it is now possible to group the various categories in a more or less natural way which depends on two or three fundamental aspects of the activity itself: total luminosity, radio luminosity and variability. While these aspects are discussed in ACTIVE GALAXIES: OVERVIEW and ACTIVE GALACTIC NUCLEI: VARIABILITY, here we give a brief overview of the different classes of AGN, intending to locate Seyferts in their proper place.

Consider first the enormous range in total luminosity of AGN: $\sim 10^{39}$ erg s^{-1} ($M_{\rm B} \sim -10$) at the lowest detectable end, up to $\sim 10^{47}$ erg s^{-1} ($M_{\rm B} \sim -30$) at the highest. Seyferts span the low to intermediate range, overlapping somewhat on the high end with quasars. Since it now seems that Seyferts and quasars form a continuous physical sequence, any attempt at a formal separation is somewhat arbitrary. Although a morphological criterion is usually adequate (quasars have point-like appearance) close scrutiny of low redshift quasars often reveals galactic 'fuzz' in ground-based images or clear host galaxies in HST images. For this reason, a luminosity criterion is sometimes chosen to delineate Seyferts and quasars. There is no consensus, but threshold values of $M_{\rm B} \sim -23$ or $L_{\rm B} \sim 10^{44}$ erg s^{-1} are typical. While Seyferts and quasars may be physically similar, differences do seem to occur with increasing luminosity, including longer variability timescales indicating larger emission regions, the apparent loss of the 'type 2' class, and some line strength changes including the relative weakening of both broad and especially narrow lines relative to the continuum.

Turning to lower luminosities, although one of the lowest luminosity AGNs is in fact a Seyfert (NGC 4395, $M_{\rm B\ nuc} \sim -10$) most of the low luminosity AGNs belong to the 'LINER' class. This group was originally identified by Heckman in 1980 from a spectroscopic survey of a large number of bright nearby galaxies. A high fraction were found to have weak emission with low overall degree of ionization, hence the name: Low Ionization Nuclear Emission Line Regions, or LINERs. The exact nature of LINERs is still unclear, with evidence for an active nucleus in at least $\sim$ 20%: a compact nuclear UV source seen in HST images; a compact though weak nuclear radio source; or weak broad line emission visible at the base of Hα. A major uncertainty is the ionization mechanism for the gas. Possibilities include shocks or photoionization by either a weak nuclear UV source, very hot young stars or very hot evolved stars.

Shortly after the identification of LINERs, a method for classifying emission line regions was introduced by Baldwin, Phillips and Terlevich and subsequently refined by Veilleux and Osterbrock. They used a number of diagrams which plotted the ratio of one pair of emission line fluxes against another pair. Figure 4 shows one of these diagrams and nicely illustrates how three major

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

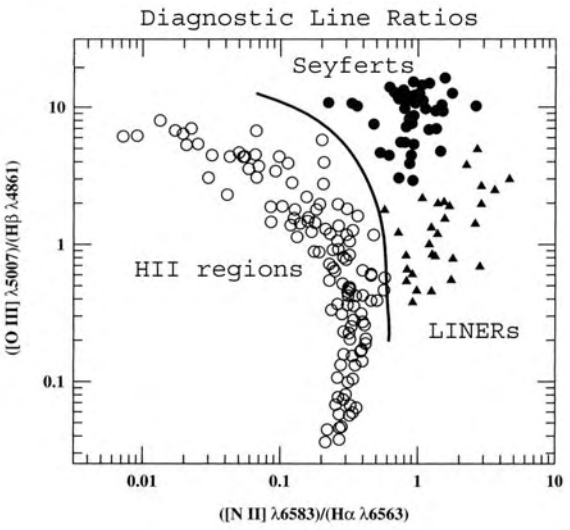


Figure 4. One of several useful diagnostic diagrams which help distinguish between different classes of emission line object. (Adapted from Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge University Press), figure 2.3, p 26).

emission region types occupy different places on the diagram; in this case Seyferts separate from LINERs around [OIII] $\lambda 5007/\mathrm{H}\beta \sim 3$ while Seyferts and LINERs separate from H II regions around [NII] $\lambda 6584/\mathrm{H}\alpha \sim 0.6$. These line ratio diagrams also provide a suitable place to compare observations with theoretical models. For example, the H II REGION sequence matches models of gas clouds with a range of metallicities photoionized by radiation from OB associations with a range of ages. Similarly, Seyferts and LINERs form a sequence that matches gas which is photoionized by a power law spectrum with a range of intensities; the highest excitation Seyferts correspond to the most intense nuclear radiation field. The apparent continuity of Seyferts and LINERs on these diagrams supports the possibility that at least some LINERs are simply low intensity Seyferts.

Turning now to radio properties, Seyferts belong to the majority ($\sim$ 90%) of AGN in being *radio quiet*—more luminous than normal galaxies but $\sim 10^3$–10^4 times fainter than radio-loud AGN. The closest radio-loud counterparts to Seyferts have similar optical luminosity and are called Broad Line Radio Galaxies (BLRG) and Narrow Line Radio Galaxies (NLRG), with optical spectra very similar (though not identical) to Seyfert 1s and Seyfert 2s respectively. At higher optical luminosity, we have radio-loud quasars, and at lower optical luminosity we have radio galaxies whose spectra are often LINER-like. Although most radio galaxies have huge radio sources which extend well outside the host galaxy, some have much smaller radio sources comparable in size to, though much more powerful than, the Seyfert sources (these are called, depending on details, Compact Steep Spectrum (CSS), Compact Double (CD) or Gigahertz Peaked Spectrum (GPS) sources). It is quite likely that similar physical phenomena are occurring in the Seyferts and these compact powerful sources—both are embedded in a dense gaseous environment and probably interacting strongly with it. Thus, studying one may help understand the other.

The fundamental reason why $\sim$ 5–10% of AGN become radio-loud is still not understood, though it is probably related in some way to the interesting fact that radio-loud and radio-quiet AGN tend to inhabit different types of host galaxy: ELLIPTICAL and SPIRAL GALAXIES respectively. It has been suggested that the radio-loud AGN have rapidly spinning (Kerr) BLACK HOLES which drive powerful jets along the spin axis, and these jets generate the powerful radio source. Conversely, slowly rotating (Schwarzschild) black holes cannot drive such jets and so remain radio-weak. The link between black hole spin and host galaxy type could arise from the fact that the elliptical hosts have formed from the merger of two spirals along with the merger of their two black holes to form a single rapidly rotating black hole.

Finally, Seyferts have little relation to the highly variable 'BLAZAR' classes of AGN: BL Lac objects, OVV (optically violently variable) quasars and HPQs (high polarization quasars). These are all radio-loud objects and, under the unified scheme for this category, are thought to be seen looking directly down a powerful jet where relativistic Doppler boosting enhances the luminosity, variability and polarization of the continuum.

Black holes and continuum emission

It is worth noting that the evidence for black holes in Seyferts and other AGN is mostly indirect: large variable luminosity; relativistic jets with a stable axis; high energy spectra. Ideally, one would study gravitational dynamics of near-nuclear stars or gas. While this has been possible in a number of normal or weakly active galaxies, it is very difficult for most Seyferts simply because of the bright nuclear light and uncertain nature of the gas motions. With many ongoing HST observations, the case is now building for black holes at the centers of most (perhaps all) luminous galaxies, with black hole mass a few per cent of the bulge mass. It would be ironic if the discovery of black holes in most normal galaxies gave the most convincing evidence, by continuity, for their presence in AGN.

Two methods might provide a better approach to detecting black holes in Seyferts. The first method is to measure the dynamics of water masers in the dense warm gas close to the center. The classic case is NGC 4258 (not, in fact, a Seyfert galaxy), which provides the best case to date for a supermassive black hole, but other possible examples are increasing in number. The second method is to study the Fe Kα x-ray emission line profile at 6.4 keV. In several cases it is very broad and double peaked, suggesting its origin in a nuclear accretion disk at a few tens of Schwarzschild radii. An important goal will be to use profile fits to infer not only the mass of the hole but also its spin and orientation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The energy released from matter falling towards the central black hole emerges in both bulk flows and photons; the latter yields, either directly or indirectly, the observed nuclear continuum emission. Perhaps the most noteworthy aspect of AGN continua is their extremely wide spectral range, with approximately equal energy per decade of frequency from far infrared to 100 keV x-rays (five decades) with extensions to radio and γ-ray energies in about 10% of AGN (not usually Seyferts). To a first approximation, the continuum shape can be described as a power law, $f_\nu \propto \nu^{-\alpha}$, with index $\alpha \sim 0$–1, with several features superposed. These features include: (a) a 'Big Blue Bump' (BBB) comprising a rise below $\sim$ 4000 Å peaking in the UV or (unobservable) EUV and dropping to rejoin the underlying continuum in the soft ($\sim$ 0.5 keV) x-ray region (the 'soft excess'); (b) an infrared peak rising from ~ 1 μm, peaking at $\sim$ 25–60 μm and falling again at $\sim$ 0.3 mm; (c) a rise at the highest energies $\sim$ 5–200 keV with cut-off at even higher energies.

The interpretation of these various components is still somewhat uncertain, though there seems to be a general consensus. The underlying power law might be generated by the synchrotron self-Compton process: a population of relativistic electrons not only generates low energy photons via the synchrotron process but then inverse-Compton scatters these same photons up to higher energies. A related model for the x-ray power law involves inverse-Compton scattering of optical–UV photons produced in the ACCRETION DISK by electrons in a hot corona above the disk. The BBB is thought to arise from thermal emission from the accretion disk, with progressively higher energy photons coming from higher temperatures at smaller radii. Although the overall BBB feature can be nicely fitted by disk models, some expected disk signatures are *not* seen: a strong Lyman edge at 912 Å; high polarization; and significant phase delays between flux variations at different wavelengths. For this reason alternatives have been suggested, such as an optically thin bremsstrahlung component. The infrared peak, which can be very strong in some objects (e.g. NGC 1068), is thought to arise in most cases from thermal emission from dust. The rise longward of ~ 1 μm suggests hot dust close to its sublimation temperature at $\sim$ 2000 K, possibly located within $\sim$ 1 pc. At longer wavelengths ($\sim$ 25–60 μm) the dust is cooler and probably located in or near the NLR. Finally, the x-ray rise might come from reflection of the underlying power law off 'cool' gas, either the accretion disk or the dense obscuring torus. A byproduct of such reflection is fluorescence of the Fe Kα emission line at $\sim$ 6.4 keV, which can be quite strong.

The observed blue–UV continuum in Seyfert 2s deserves further comment since, as we have discussed, the nuclear contribution should be hidden behind an obscuring torus. Why, then, do many Seyfert 2s have a sufficiently strong UV excess to show up in the Markarian survey? The reason is that much of the blue–UV light in Seyfert 2s is *extended*, coming from the inner few kpc. The origin of this continuum is still under some debate, but it seems there are at least several contributions. Polarization observations indicate that part of the emission is scattered from the hidden nucleus. The remainder (sometimes called 'FC2', for Featureless Continuum 2) may be a combination of near-nuclear star formation, Balmer continuum emission from the ionized gas, and continuum emission from hot post-shock gas associated with the radio source.

Broad line region

A defining characteristic of Seyfert 1 spectra is the presence of broad emission lines, including the Balmer lines, He II λ4686 and He I λ5876 in the optical, and Lyα λ1212, C IV λ1549, C III] λ1909 and Mg II λ2800 in the near UV. In the early 1980s it was found that the Balmer line strengths in Seyferts and quasars form a single tight correlation with the non-stellar continuum strength, confirming a basic connection between the two: continuum emission photoionizes the BLR gas which then produces line radiation. An estimate of the gas density in the BLR comes from the absence of any broad components on the forbidden lines. These so-called 'forbidden' emission lines come from electron transitions which cannot occur via the fast electric dipole mode (which requires a quantum number change $\Delta l = \pm 1$). With only slower modes of transition available (e.g. magnetic dipole), the ions are vulnerable to collisional deexcitation in a high density gas where the mean time between collisions is small. The 'critical density', $n_{\rm crit}$, above which collisional de-excitation becomes important varies widely for different lines, so one can bracket the gas density in the BLR by noting which lines have broad components and which do not. For example, while the [OIII] λ4363 line with $n_{\rm crit} \sim 3 \times 10^7$ cm^{-3} has no broad component the semiforbidden CIII] λ1909 line with $n_{\rm crit} \sim 3 \times 10^{10}$ cm^{-3} does, implying a gas density in the BLR somewhere in the range $10^{7.5}$–$10^{10.5}$ cm^{-3} (see below for a refinement of this estimate). In contrast to the forbidden lines, the so-called 'permitted' lines arise from fast ($\sim 10^{-8}$ s) electric dipole transitions with correspondingly high critical densities $\sim 10^{17}$ cm^{-3}. Such lines, which include those due to recombination, are therefore easily produced in the BLR gas.

In the last decade, Seyfert variability studies have significantly helped develop our picture of the BLR. In these studies, the emission lines and continuum are repeatedly measured over periods of many months. Because of the finite size of the region, it takes time for a change in the central continuum brightness to reach the BLR gas and thereby cause a change in the emission line. Allowing for light travel delays between the continuum source and BLR gas and between the BLR gas and us, it is in principle possible to reconstruct the distribution of line emission and velocity field throughout the BLR. In practice, this full 'reverberation mapping' has not yet been possible, but instead simple time delays have been measured in the response of each emission line to the continuum, and these delays indicate an approximate region size. The most well studied object is the Seyfert

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

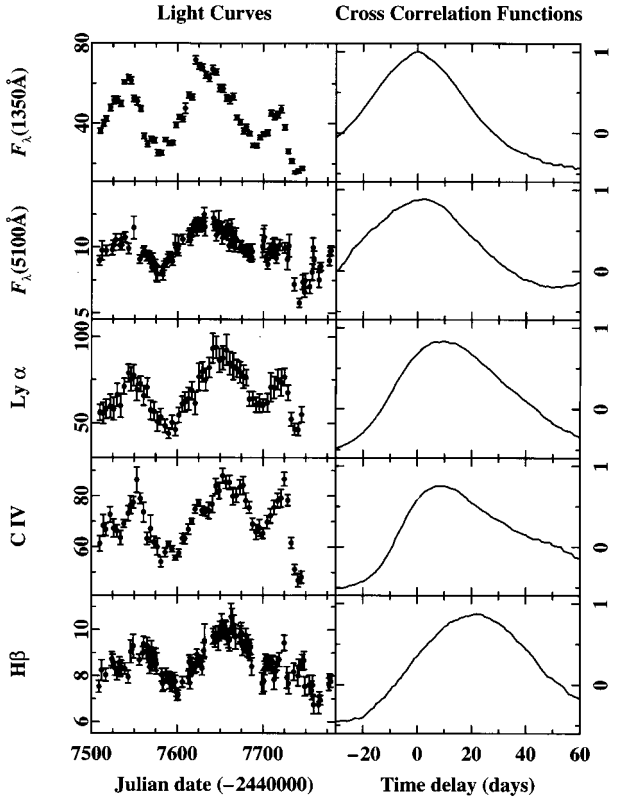


Figure 5. Variability in NGC 5548 over a period of 280 days in 1988–89. Left panels show changing fluxes in the continuum (upper two) and emission lines (lower three), while the right panels show the cross correlation of each flux curve with the UV continuum (uppermost left panel). Note that while the optical and UV continua vary together, the Lyα, C IV λ1550 and Hβ emission lines lag behind by increasing amounts, Hβ being delayed by about 20 days. These lags indicate the size of the BLR (about 20 light days) and the fact that it is radially stratified, with C IV produced preferentially at smaller radii and Hβ at larger radii. (Taken from Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge University Press), figure 5.6, p 86).

1.5 galaxy NGC 5548, for which delays of $\sim$ 2–10 days are found for the highest ionization lines (e.g. He II λ1640, C IV λ1549) and $\sim$ 10–20 days for the lower ionization lines (e.g. Lyα, Hα, Hβ, CIII] λ1909), indicating a radially stratified BLR with size $\sim$ 2–20 light days (figure 5). Surprisingly, these size estimates were a factor of ten smaller than previous estimates based on photoionization calculations (see below). The reason for the discrepancy is now understood: not only does stratification undermine the simple photoionization estimate but the densities in the BLR are probably higher than previously thought (e.g. $\sim 10^{11}$ cm^{-3} in the C IV λ1459 emitting region).

Since the ionization structure in the BLR is roughly the same from object to object (because the line ratios are similar), we infer that BLR sizes must increase for more luminous objects. Using the C IV λ1549 data from NGC 5548 as fiducial, the form of the dependence is $R_{\rm BLR} \sim 10 h_0 L_\lambda(1350\ \text{Å}, 40)^{1/2}$ light days, where h_0 is the Hubble constant in units of 100 km s^{-1} Mpc^{-1} and $L_\lambda(1350\ \text{Å}, 40)$ is the continuum luminosity at 1350 Å in units of 10^{40} erg s^{-1} Å^{-1}. Since, roughly speaking, quasars are $\sim$ 100 times more luminous than Seyferts, they have BLRs which are $\sim$ 10 times larger, perhaps a few light months to a light year across.

In the 1980s photoionization models were developed which aimed to reproduce the observed emission line strengths. In general, these calculations consider a power law (or AGN-like) continuum impinging on a thick slab of gas. At a given depth in the cloud, ionization and thermal equilibrium conditions are evaluated, together with the emitted line spectrum. Absorption of the incoming ionizing continuum is calculated and the modified continuum enters the next zone deeper in the cloud. The calculation repeats for ever deeper zones in the cloud until the gas is sufficiently cool and/or neutral that no more emission is produced. Overall, the ionization degree drops into the cloud, becoming suddenly more neutral at a depth corresponding to the classical Strömgren depth (where the total number of ionizing photons is balanced by the total number of hydrogen recombinations). However, unlike H II regions ionized by hot stars, AGN continua also have a high energy component which penetrates beyond the Strömgren depth to create an extended partially ionized zone in which many low ionization lines are produced. Usually, the emergent spectrum from a single cloud is compared to observations, though more recently cloud ensembles spanning a range of conditions have been considered in an attempt to match the stratified nature of the BLR.

Several parameters help define these photoionization calculations. The most important is the 'radiation parameter', defined as the ratio at the cloud front of the ionizing photon density to the hydrogen density: $U = N_{\rm ion}/n_{\rm H}$, where $N_{\rm ion} = Q/4\pi r^2 c$ where Q is the number of ionizing photons produced by the central source per second, and r is the distance to the cloud. Other relevant parameters are gas density, cloud thickness, ionizing spectral shape and element abundances. Optimum fits suggest that for Seyfert 1s, $U \sim 0.04$ and $n_{\rm H} \sim 10^{11}$ cm^{-3}. Estimating cloud thickness is more difficult. The fact that in variable objects the emission lines follow the continuum tells us the clouds are at least optically thick to the ionizing radiation, i.e. they are at least thicker than the Strömgren depth $S \sim Uc/n_{\rm e}\alpha_{\rm B} \sim 5 \times 10^{10}$ cm where $n_{\rm e}$ is the electron density and $\alpha_{\rm B}$ is the case-B hydrogen recombination coefficient ($\sim 2.5 \times 10^{-13}$ cm^3 s^{-1}). Matching the low ionization strengths requires the extended partially ionized region to be a few times deeper—perhaps a few times the diameter of our Sun. There is also evidence, it should be said, for an optically thin component to the BLR gas, coming from studies of line ratio and profile variability.

It is possible to estimate some other properties of the BLR line-emitting gas, starting with the basic equation for the production of Hβ by recombination: $L_{\rm H\beta} =$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$h\nu_{H\beta} n_e n_H \alpha_{H\beta} V$ where $h\nu_{H\beta}$ is the energy of each Hβ photon, n_e and n_H are the electon and proton densities, V is the volume of ionized gas, and $\alpha_{H\beta}$ is the recombination coefficient for Hβ ($\sim 3 \times 10^{-14}$ cm^3 s^{-1}). Since the mass of ionized gas is simply $M = m_p n_H V$ where m_p is the proton mass, we find an ionized gas mass and volume of $\sim 0.1\, M_\odot$ and 10^{45} cm^3 for a typical Hβ luminosity $\sim 10^{42}$ erg s^{-1} and BLR gas density $\sim 10^{11}$ cm^{-3}. Taking the BLR radius of 10 light days from the variability measurements given above, we find that the ionized gas occupies only $f_c \sim 1.5 \times 10^{-5}$ of the BLR volume (f_c is called the filling factor). If we assume the gas is in the form of clouds with diameters a few times the Strömgren depth ($d_c \sim 5 \times 10^{11}$ cm) we derive a number of clouds $N_c \sim V/d_c^3 \sim 10^{10}$ which collectively intercept a fraction of the central radiation of $C_c \sim N_c d_c^2/4\pi r^2 \sim 0.3$ (C_c is called the covering factor). In this picture, then, the BLR is sparsely filled by a large number of small clouds which collectively intercept $\sim 30\%$ of the ionizing radiation.

The velocity field in the BLR is poorly understood, despite the fact that the emission line profiles clearly have Doppler origin. Profiles have approximately logarithmic form (flux $\propto \log \Delta\lambda$) often with slight asymmetry and/or substructure, with full widths at half maximum (FWHM) anywhere from 1000 to 10 000 km s^{-1} and maximum wing widths of up to 30 000 km s^{-1}. Lines in the same object can have slightly different shape (e.g. He II λ4686 is often broader than Hβ), indicating both velocity and ionization stratification in the BLR, with higher velocities at smaller radii. One approach has been to fit each line by the sum of two components, one broad and one very broad with possible blueshift. These fits support the decrease in velocity for lower ionization gas, and point to a slight radial velocity component, although profile variability studies suggest that radial flow is *not* a dominant component of the velocity field. If the velocities are gravitational, estimates of central mass $M \sim rv^2/G \sim 10^7$–$10^8\, M_\odot$ with some evidence for a correlation between central mass and luminosity. The luminosities derived amount to $\sim 1\%$ of the Eddington luminosity, which is the natural upper limit to accretion-powered energy release when outward radiation pressure just balances inward gravitational force, limiting further accretion.

The geometry of the BLR is also poorly understood, with spherical or flattened cloud distributions as obvious possibilities, including a thin but flared disk, although the absence of very narrow lines expected for face-on disks weighs against this last possibility. The origin and evolution of the cloud system are also poorly understood. Possibilities that have been explored include a two-phase medium in which the clouds are in pressure balance with a hot intercloud medium; clouds may be confined by magnetic forces; clouds may form in cooling post-shocked gas; clouds may even comprise the extended atmospheres of giant stars. Each of these has its strengths and weaknesses, and we must await further work to clarify the true situation.

Narrow line region

In the spectra of Seyferts, the narrow forbidden and permitted emission lines come from a region well outside the BLR, spanning a few tens of pc to a few kpc called the 'Narrow Line Region' (NLR). Unlike the BLR, therefore, ground-based telescopes can often resolve the outer parts of the NLR while HST has provided some detailed images and spectra. The mere presence of forbidden lines tells us that gas densities are lower than in the BLR. At these lower densities the physics of line production is simpler and line ratios may be used to estimate physical properties of the ionized gas. Classic examples include the two line ratios: [S II] $\lambda\lambda$6717/6731 and [O III] $\lambda\lambda$5007/4363 which yield electron densities $\sim 10^3$–10^4 cm^{-3} and electron temperatures $\sim$ 1–2 $\times 10^4$ K respectively. Such simple estimates must be viewed with caution, however, because they represent weighted averages over a wide range of conditions: for example, gas densities probably rise from $\sim 10^2$ cm^{-3} on kpc scales to $\sim 10^{6-8}$ cm^{-3} in the inner regions. More sophisticated approaches follow the same strategy as for the BLR: photoionization calculations evaluate conditions within a single gas cloud and, in more recent work, integrate over a distribution of cloud properties. As with the BLR gas, much of the NLR gas is thought to be optically thick to the ionizing radiation, although studies of the He II λ4686/Hβ ratio indicate the presence of some optically thin gas. In trying to match these calculations to observations, sequences of models are evaluated with a range of radiation parameter and/or relative mix of optically thick to optically thin gas. Typical values for the radiation parameter found for the NLR are $U \sim 0.01$, with some evidence for higher values in Seyfert 1s than Seyfert 2s.

Repeating the analysis for the BLR, we can derive crude estimates for the global properties of the NLR gas. Taking as fiducial, $L_{H\beta} \sim 10^{41}$ erg s^{-1}, a gas density of $\sim 10^3$ cm^{-3} and NLR size of ~ 100 pc, we obtain a mass of ionized gas of $\sim 10^7\, M_\odot$ with filling factor $\sim 10^{-2}$. The Strömgren depth in the clouds is $\sim 10^{18}$ cm; there are $\sim 10^6$ clouds with integrated covering factor ~ 1. While the numbers are far from precise, they do indicate that the amount of gas is dynamically unimportant, that it only sparsely fills the region, and that it intercepts a significant fraction of the ionizing radiation.

As the name suggests, the emission lines from the NLR have 'narrow' Doppler widths, with FWHM $\sim$ 200–800 km s^{-1} (a few lie outside this range). The emission line shapes are almost always more peaky than Gaussian, and most have an asymmetry with the base and wing stronger and more extended on the blue side than the red (figure 6). This asymmetry tells us that, at least for the higher velocity, more nuclear gas, there is a radial flow and part of the region is obscured by dust. If the far side of the region is obscured by intervening dust, then the flow is *outwards* (the visible near side is blueshifted). If the clouds themselves contain dust, and the line emission comes from the side of the cloud facing the central source, then the flow is *inwards* (we cannot see the clouds on the near side

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

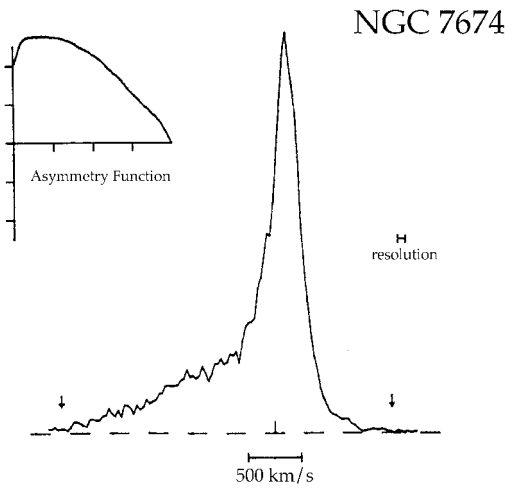


Figure 6. The [O III] λ5007 emission line profile from the nuclear region of NGC 7674 provides an extreme example of an asymmetric blue wing. A velocity scale, line median, instrument resolution and asymmetry function are also shown.

since we look into their backs). Recent observations have shown that the optical and infrared forbidden lines have similar asymmetries, showing that the dust is still opaque even at these long wavelengths. The preferred scenario is thick nuclear dust, possibly in a dense disk, blocking the far side and indicating an outflow. The outflow might be driven by a nuclear wind or be related to the jets which give rise to the radio source. For those Seyferts with relatively strong double or triple structured radio sources it is often possible to see separate red- and blueshifted emission components close to the radio lobes, indicating bipolar driven outflows. In these objects, the integrated line profile can be particularly structured and broad, with base and core widths up to $\sim$ 1000 km s^{-1} (examples include NGC 1068, MKN 3, MKN 78). For most objects, however, there is good evidence that the majority of the NLR gas, as measured by the [O III] λ5007 profile core width (FWHM), is moving in the gravitational field of the bulge of the host galaxy. The evidence includes the fact that the profile peaks and cores are close to the rest frame of the galaxy, and that there are good correlations of the [O III] λ5007 emission line width with (a) rotation velocity of the galaxy, (b) bulge luminosity and (c) bulge stellar velocity dispersion. Stated simply, galaxies with massive bulges have relatively broad forbidden lines (e.g. $\sim$ 500 km s^{-1}), while galaxies with low mass bulges have relatively narrow forbidden lines (e.g. $\sim$ 200 km s^{-1}). Galaxies with relatively strong double or triple radio sources tend to stand off these correlations with yet broader lines, due to bipolar flows. For the gravitationally dominated gas, it is still unclear whether the predominant motion is rotation or random, though probably both occur.

Not all emission lines have the same profile shape, and in general lines of higher ionization and/or higher critical density tend to be broader with a higher degree of asymmetry. This indicates that, like the BLR, the NLR is stratified with the higher velocity, more nuclear gas having higher density and ionization degree. It is conceivable, though not yet shown convincingly, that a continuous sequence of conditions connects the inner NLR to the BLR.

The spatial structure of the NLR emission has been studied using emission line imaging, both from the ground and using HST. In this technique, images are taken through two narrow pass filters, one centered on the color of the emission line (ON band) and one centered on the nearby continuum (OFF band). By subtracting the OFF band from the ON band one is left with a pure emission line image. Furthermore, by dividing two emission line images such as [O III] λ5007 and Hβ one can generate a map of the gas excitation which clearly highlights AGN-related emission ([O III]/Hβ $\sim$ 10) and distinguishes it from star formation emission ([O III]/Hβ $<$ 3). Such images indicate that the NLR is often elongated, with an axis parallel to the radio axis. Indeed, for those Seyferts with well developed radio sources the NLR emission is closely related to the radio source, with brighter emission either coincident with jets and jet bends, or surrounding and sheathing radio lobes, sometimes with a curved 'bow shock' appearance. Spectroscopic studies show higher velocities for this radio-related emission, indicating strong dynamical interaction and an overall bipolar outflow, but also including split lines near jets, possibly indicating lateral expansion away from the jet. A good example of close radio/emission line association is Markarian 78 (figure 7).

The close association between line and radio emission in the NLR raises the important possibility that the gas is ionized, at least in part, by shocks arising in the interaction. Although early calculations of slow shocks could only match LINER-like spectra, more recent calculations of fast ($\sim$ 500 km s^{-1}) shocks can match Seyfert-like spectra. A key feature of these calculations is the production of UV and x-ray radiation, both line and continuum, in the hot post-shock gas. This radiation then photoionizes gas both behind the shock and in front of it, yielding a spectrum quite similar to that observed in Seyferts. At present, 'photoionizing shocks' and standard power law photoionization (possibly including optically thin gas) give an equivalently good match at least to the optical lines, though there is some hope that UV lines will help discriminate between the models, as long as the uncertainties introduced by dust reddening are not too great.

Returning to the emission line images and excitation maps, a number show sharp edges which outline parts of a cone, or bi-cone if it occurs on both sides. As described earlier, this seems to indicate a partly shadowed nuclear source of ionizing radiation. These emission regions can extend out to considerable distances, $\sim$ 15–20 kpc, where they are termed the 'Extended Narrow Line Region' or ENLR. This emission has high excitation and narrow lines of low velocity, suggesting it is simply undisturbed ambient gas residing either in the disk of the galaxy or in the galactic environment. The fact that most

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

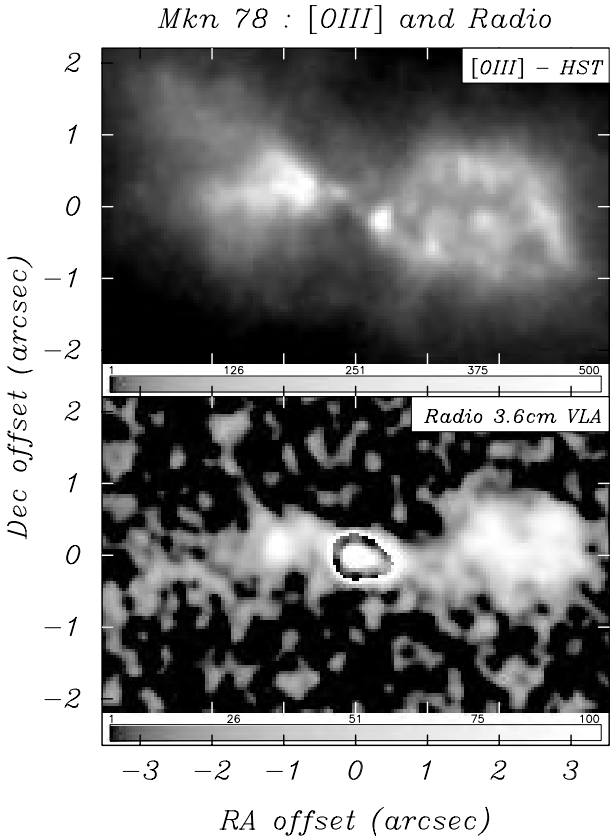


Figure 7. HST [O III] emission line image and VLA radio image of the Seyfert 2 galaxy Mkn 78 showing close spatial association. Note that a dust lane crosses the nucleus, hiding the most nuclear [O III] emission. Gray scale numbers are proportional to flux, with the radio core rescaled by a factor of five for clarity. For this galaxy 1 arcsec $\sim$ 1 kpc ($h_0 = 50$ km s^{-1} Mpc^{-1}). (Data courtesy of M Whittle and A S Wilson.)

ENLRs lie well beyond any radio source and can show conical form suggests they are ionized by the central UV radiation source. Recently, however, careful estimates of the required UV flux suggest that at larger radii there may be an additional source of ionization, though what this is is not yet known.

A final pragmatic theme is the frequent need to correct measured spectra for the effects of reddening by dust. Fortunately, the NLR Balmer line strengths are expected to be close to their theoretical values, and so any deviation can be attributed to reddening. For example, the Hα/Hβ flux ratio should be $\sim$ 2.8 (the 'Case B' value, though a more detailed calculation gives 3.1) and so a measured value of, say, R, yields an estimate for the absorption due to dust of $A_V \simeq 6.8 \log_{10}(R/2.8)$ magnitudes in the V band, where the coefficient depends on the reddening law (Whitford's law in this case). Knowing A_V allows one to correct the observed spectrum, using the same reddening law $f(\lambda)$, where $A_\lambda = A_V \times f(\lambda)$.

Hosts, environment and triggering

While the previous sections have concentrated on the anatomy of activity in Seyfert galaxies, there remains an important question: why are some galaxies active while others are not? Although we cannot yet answer this question with confidence, two or three lines of enquiry help pave the way forward. The first is to ask whether Seyfert host galaxies are unusual in any way; the second is to ask whether any active properties correlate with any host galaxy properties; and the third is to identify mechanisms which might send gas down to fuel the nuclear regions. In this area, empirical studies are difficult because of the statistical weakness of the effects and the presence of sample and control biases.

Perhaps the one clear result concerns the morphological type of the host galaxies: Seyferts favor early-type spirals, S0–Sbc, clearly avoiding late-type spirals and, to a lesser extent, ellipticals. Interestingly, LINERs show a similar distribution (with somewhat more ellipticals), while radio galaxies are almost exclusively elliptical or S0. In all these cases, more luminous galaxies are more likely to be active. It seems, therefore, that an important condition for the onset of activity is a relatively massive bulge. There is also a tendency for active galaxies to be disturbed in some way and/or have a nearby companion—both being conditions which may lead to nuclear fueling. The neutral gas content (atomic and molecular) of Seyferts is basically normal, though Seyfert 2s may have more molecular gas and associated star formation than otherwise similar non-active galaxies. Recent HST images have also shown that Seyfert 2s tend to inhabit later Hubble types and have more nuclear dust lanes than Seyfert 1s. This suggests a modification to the standard unification scheme: Seyfert 2s simply have more nuclear dust than the Seyfert 1s and are therefore more likely to be obscured, rather than being identical except for our viewing angle.

It is worth noting a few active properties which do (or don't) correlate with host galaxy properties. First is the axis of nuclear activity, as defined by the radiation field and/or radio source. Overall, there appears to be little alignment of this axis with the large scale disk or bar of the host galaxy. Recent evidence, however, suggests that for later Hubble types there *is* an alignment with the host disk, while for early Hubble types the alignment is with smaller scale nuclear gas or dust disks or lanes whose more random orientation probably reflects recent merger or infall history. Because the NLR resides within the galaxy bulge, it is perhaps not surprising that some NLR properties correlate with bulge luminosity, for example emission line and radio luminosities. The origin of these correlations is not yet clear, but might involve a number of properties which depend on bulge mass (i.e. depth of gravitational potential): emission region volume, gas and radio source pressures, mass loss rate from bulge stars, and mass of central black hole. More easy to understand is the dependence of NLR emission line width on bulge mass, since the gas velocities are largely gravitational. An interesting wrinkle on this last result is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that disturbed galaxies have even broader emission lines, confirming that such disturbances can affect gas motions in the inner regions. Perhaps more fundamental is a possible correlation between nuclear (active) luminosity and host galaxy (or more likely, bulge) luminosity—low luminosity Seyferts reside in lower mass bulges than quasars. An obvious explanation is that more massive bulges have more massive central black holes, a result supported at least in part by recent black hole mass measurements in more normal galaxies.

Perhaps the central theoretical problem has been to understand how gas moves from far out in the galaxy disk down to the innermost regions where it can accrete onto the black hole—a loss in ANGULAR MOMENTUM by a factor $\sim 10^5$! Early work focused on bars as a means of draining angular momentum from gas, via shocks and dissipation. It is at present unclear how important this particular mechanism is since, although studies have been somewhat ambiguous, there seems to be no tendency for active nuclei to prefer barred galaxies. More recently, Barnes and Hernquist have shown, using computer simulations, that galaxy encounters and mergers can be particularly effective at sending gas into the nuclear regions. The tidal fields resulting from the encounter help create a gaseous bar which lags behind a stellar bar. The gravitational torque on the gas robs it of angular momentum and it falls inwards, losing up to $\sim 99\%$ of its angular momentum. At smaller radii it is unclear what happens: gas may build up and become unstable to self gravity and/or there may be bars within bars which help funnel the gas to ever smaller radii. Once in an accretion disk, however, Balbus and Hawley have shown that magnetically driven turbulence provides enough viscosity to drive the gas inwards to the black hole.

The future

We now briefly identify several themes likely to be pursued in the coming years. (a) The hunt for black holes will continue, aiming to establish a demographic perspective—how do black hole mass and spin relate to AGN or host properties? (b) How does the inner accretion disk function, and how does it play a role in generating the continuum emission? (c) Much remains to be done to find out the structure and velocity field of the BLR—ideally, its kinematics will be dominated by black hole gravity giving another handle on the properties of the central object. (d) The root cause of the apparent unification of Seyfert 1s and 2s needs to be clarified—is the torus model too simple in the light of the recent HST observations showing more distributed dust lanes in Seyfert 2s? (e) We need to refine our picture of the NLR, and in particular the importance of the radio emitting jets—clearly the jets accelerate the gas, but do they also ionize it? (f) What is the nature of the extended UV continuum in Seyfert 2s, and is enhanced star formation associated with nuclear activity? (g) Finally, we would like to know how activity is triggered in more detail. Although interactions are now known to send gas to the central kpc, how does it get from there to the pc scale and beyond?

Clearly, these are exciting times in the study of Seyferts.

Bibliography

Some excellent graduate level texts include:

Osterbrock D E 1989 *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (University Science Books)

Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

Krolik J H 1999 *Active Galactic Nuclei, From the Central Black Hole to the Galactic Environment* (Princeton, NJ: Princeton University Press)

Kembhavi A K and Narlikar J V 1999 *Quasars and Active Galactic Nuclei, An Introduction* (Cambridge: Cambridge University Press)

Some useful review articles in approximate chronological order include:

Weedman D W 1977 Seyfert galaxies *Ann. Rev. Astron. Astrophys.* **15** 69–95

Balick B and Heckman T M 1982 Extranuclear Clues to the Origin and Evolution of Activity in Galaxies *Ann. Rev. Astron. Astrophys.* **20** 431–68

Rees M J 1984 Black hole models for active galactic nuclei *Ann. Rev. Astron. Astrophys.* **22** 471–506

Osterbrock D E and Mathews W G 1986 Emission line regions of active galaxies and QSOs *Ann. Rev. Astron. Astrophys.* **24** 171–203

Osterbrock D E 1991 Active galactic nuclei *Rep. Prog. Phys.* **54** 579–633

Barnes J E and Hernquist L 1992 Dynamics of interacting galaxies *Ann. Rev. Astron. Astrophys.* **30** 705–42

Antonucci R 1993 Unified Models for active galactic nuclei *Ann. Rev. Astron. Astrophys.* **31** 473–521

Mushotsky R F, Done C and Pounds K A 1993 X-ray spectra and time variability of active galactic nuclei *Ann. Rev. Astron. Astrophys.* **31** 717–61

Peterson B M 1993 Reverberation mapping of active galactic nuclei *Publ. Astron. Soc. Pacific* **105** 247–68

Koratkar A and Blaes O 1999 The ultraviolet and optical continuum emission in active galactic nuclei: the status of accretion disks *Publ. Astron. Soc. Pacific* **111** 1–30

Mark Whittle

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Seyfert, Carl K (1911–60)

American astronomer, studied the properties of galaxies and identified a class with condensed nuclei and broad, high-excitation emission lines. These are now known as *Seyfert galaxies*, with the central nucleus supposed to be a massive black hole.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shakerley, Jeremy (1626–*c*. 1655)

Born in Halifax, Yorkshire, self-educated. Taken into the Towneley household and encouraged in his scientific pursuits, he discovered the work of HORROCKS and was the first mathematician to recognize its significance. He emigrated to India, possibly with the East India Company, and observed a transit of Mercury in 1651, the second such transit ever observed. He died young in India.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shanghai Astronomical Observatory

The Shanghai Astronomical Observatory, Chinese Academy of Sciences (CAS), is located in Shanghai, China. It was established in 1962 combining Xujiahui and Sheshan, which were founded in 1872 and 1900 respectively. The main fields cover astrophysics and astrogeodynamics, including star clusters, our Galaxy, galaxies, Earth rotation and satellite dynamics application. The major facilities are: 1.56 m telescope, 25 m radio telescope, 60 cm SLR system and 40 cm astrograph. The staff consists of some 50 astronomers, 40 engineers and technicians.

For further information see
http://center.shao.ac.cn/english.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shapley Concentration

Although the universe is homogeneous on very large scales, it is clearly clumpy even on scales of a few million parsecs. Thousands of galaxies, together with gas and dark matter, make up gravitationally bound clusters, and several associations of clusters, known as superclusters, have been identified. The *Shapley concentration* is probably the largest concentration of galaxies in our nearby universe that forms a gravitationally interacting unit, thereby pulling itself together instead of expanding with the universe. It appears as a striking overdensity in the distribution of galaxies (figure 1) in the constellation of Centaurus.

Rich CLUSTERS OF GALAXIES are rare peaks in the distribution of matter in the universe, but their distribution can be expected to follow that of the underlying matter. The distribution of rich clusters found on photographic plates by Abell and his colleagues is shown in figure 2, and shows the Shapley concentration, also known as the *Shapley supercluster* or the *alpha concentration*, as the most obvious aggregation of clusters.

Furthermore, the Shapley concentration lies very close to the direction in which the LOCAL GROUP of galaxies (including our Galaxy) is moving with respect to the rest frame of the universe defined by the COSMIC MICROWAVE BACKGROUND radiation. This has led many to speculate that the Shapley concentration, even though it is at a distance of 140 Mpc from us, may indeed be one of the major causes of our peculiar motion.

Shapley's galaxy survey

In the late 1920s, HARLOW SHAPLEY and his team at the Harvard College Observatory started a survey of galaxies in the southern sky, using photographic plates obtained at the 24-inch Bruce telescope at Bloemfontein, South Africa. By 1932, Shapley could report the discovery of 76 000 galaxies between 14th and 18th magnitude in a third of the southern sky, for a small fraction of which actual positions and magnitudes had been measured. More than half of these galaxies were estimated to be fainter than 17th magnitude. Some of these data were later published as part of the Harvard galaxy counts, intended to map Galactic obscuration and to find the space density of galaxies.

In this catalog, Shapley could see most of the 'Coma–Virgo cloud' (now known to be a superposition of the Coma and Virgo superclusters), but found a 'cloud' in Centaurus to be the most striking concentration of galaxies. He found it particularly interesting because of its *'great linear dimension, the numerous population and distinctly elongated form'*. This $2^\circ \times 1^\circ$ cloud, centered on right ascension $\alpha = 13^{\rm h}\,26^{\rm m}$ and declination $\delta = -31^\circ\,37'$ (J2000) can be identified with what we now know as the core of the Shapley concentration. He also noted an *'indefinite extension'*, consisting of 30 'systems', from the rich cluster Shapley 8 (Abell 3558) towards the north of the cloud.

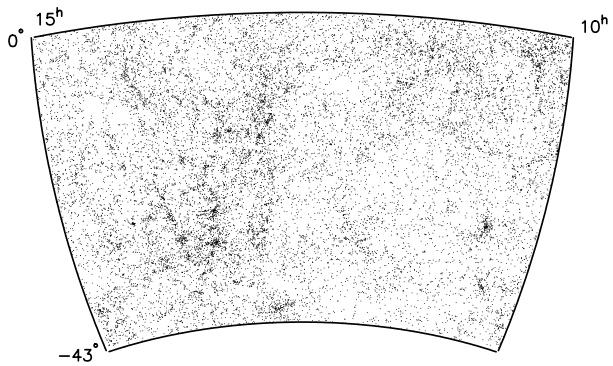


Figure 1. All galaxies brighter than $B = 17$th magnitude between right ascension $10^{\rm h}$ to $15^{\rm h}$ and declination -43° to 0° plotted in an equal-area projection. The Shapley concentration is the obvious crowd of galaxies in the lower left half.

Shapley estimated the distance to this cloud to be 14 times that to the Virgo cluster, from the average diameters and the shapes of the luminosity function of the galaxies belonging to the two concentrations. His actual distance of 45 Mpc, however, was in error, because the Hubble constant he had used was too large, but the relative value is fairly accurate. The whole system, he speculated, consists of 2000 galaxies, and 'is obviously a super-system that is important in our consideration of our own galactic system'. This was remarkably prescient, in a paper which otherwise (wrongly) proposed that we live in a supergalaxy 40–50 times the size of a normal galaxy, that the Hubble constant $H_0 = 430\ {\rm km\,s^{-1}\,Mpc^{-1}}$ and that the dark clouds in Sagittarius–Scorpio–Ophiuchus constitute a distinct single galaxy as large as M31, belonging to our supergalaxy.

How massive is the Shapley concentration?

The ∼20 rich clusters in the core of the concentration, lying between $13^{\rm h} < \alpha < 14^{\rm h}$ and $-40^\circ < \delta < -30^\circ$ is centered on the richest cluster Abell 3558 (see figures 3 and 4). The mean REDSHIFT of these clusters is $\langle cz \rangle = 14\,000\ {\rm km\,s^{-1}}$.

Its angular extent is about 10°, corresponding to about $25\,h^{-1}$ Mpc at that distance (here I use the value of the Hubble constant to be $H_0 = 100\,h\ {\rm km\,s^{-1}\,Mpc^{-1}}$). If the Shapley concentration is a bound and virialized entity, one can roughly estimate its mass to be about $10^{17}\,h^{-1}\,M_\odot$ within a sphere of diameter $25\,h^{-1}$. The crossing time of a cluster in this concentration is comparable to the age of the universe, which is why the assumption that it is virialized is not reasonable and this mass would certainly be an overestimate.

Even if the Shapley concentration is not a virialized entity, it might still be a gravitationally bound system. From the temperature of the x-ray emitting gas and the velocity dispersion of galaxies in the clusters belonging to the Shapley concentration, the mass of the Shapley supercluster has been estimated by various authors to be in the range $1.5\text{–}4 \times 10^{16}\,h^{-1}\,M_\odot$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

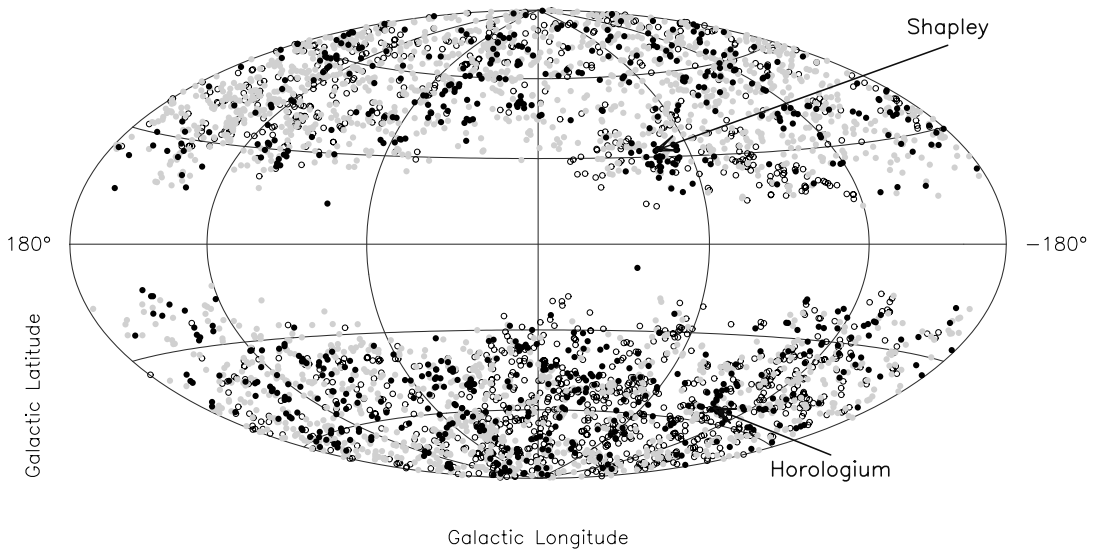


Figure 2. The distribution of 4076 optically selected Abell clusters plotted in an equal-area projection in Galactic coordinates. Clusters lying within redshift $z \leq 0.1$ are plotted in black, between $0.1 < z \leq 0.2$ in grey and those with $z > 0.2$ as open circles. The underpopulated area is the portion of the sky obscured by our own Galaxy. The two richest concentrations of clusters in the $z \leq 0.1$ universe are indicated.

Is it the 'Great Attractor'?

Several studies of the distances and motions of galaxies have revealed large systematic departures from the Hubble Flow in the nearby universe. Many attempts have been made to account for them in terms of the gravitational effects of significant overdensities and underdensities in the local distribution of matter.

The simplest such model is that of the 'Great Attractor', originally proposed to be a large concentration of galaxies of mass about $2 \times 10^{16}\, h^{-1}\, M_{\odot}$, situated at a distance of $r_{\rm GA} = 42\, h^{-1}$ Mpc from us. Subsequent studies have disagreed on whether a back-flow beyond $r_{\rm GA}$, which should be the signature of a single mass concentration causing our peculiar motion, is detected or not. Interestingly, the detailed reconstruction of the density field from peculiar velocities, within $50\, h^{-1}$ Mpc from us, finds evidence for a large concentration of mass at the distance of the proposed Great Attractor.

However, it is uncertain whether this concentration is enough to account for the peculiar velocities in the Local Supercluster, and whether the peculiar velocities converge on larger scales (see also SUPERCLUSTERS AND THE LOCAL SUPERCLUSTER). It is in this context that the Shapley concentration was rediscovered in the 1980s.

The fact that the Shapley concentration lies within 10° of the direction of motion of the Local Group, and very close to the proposed position of the Great Attractor, prompts us to consider the possibility of its being an important contributor to the peculiar motion of the Local Group. However, the Shapley concentration is situated at three times the putative distance of the Great Attractor. If it were to be responsible for the entire peculiar motion of the Local Group (600 km s^{-1} towards Hydra–Centaurus), it would have to have a mass of $M_{\rm GA} \sim 2.6 \times 10^{17}\, h^{-1}\, M_{\odot}$.

If all the galaxies brighter than 17th magnitude within 10° ($25\, h^{-1}$ Mpc) of Shapley 8 were placed at a distance of $140\, h^{-1}$ Mpc, the mean distance of the Shapley concentration, then the total blue light from these galaxies would amount to $L_B \sim 5 \times 10^{13}\, h^{-2}\, L_{B,\odot}$. This is clearly an overestimate, because a lot of these galaxies would be in the foreground, particularly in clusters belonging to the Hydra–Centaurus supercluster. So the *lower limit* to the mass-to-light ratio of the Shapley concentration would be $M/L_B \sim 4500\, h$, if it alone were responsible for the entire peculiar motion of the Local Group. This is an improbably high value, considering that a large spiral galaxy like ours has a mass-to-light ratio of about 10, and a typical cluster of galaxies has $M/L_B \sim 300\, h$. Indeed, for a critical-density universe ($\Omega_0 = 1$), the corresponding value is $M/L_B \sim 1500\, h$.

The optical dipole

Both gravity and luminosity decrease as the inverse square of the distance. Thus the dipole moment of an all-sky distribution of galaxies (individually weighted by their luminosities) can be used to estimate the peculiar acceleration of the Local Group. Since this does not require one to know the distances of individual galaxies, it can be calculated from a catalog of positions and luminosities (or sizes) of galaxies. It has been shown that the net acceleration of the Local Group estimated in this way is almost parallel to its direction of motion, supporting the notion that in the universe the distribution of matter follows the distribution of light.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

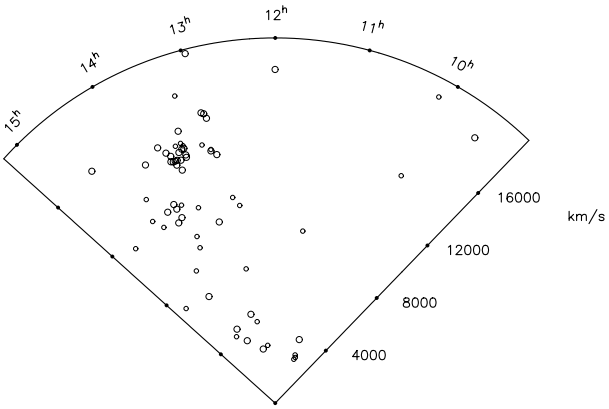


Figure 3. Galaxy clusters with known redshift in the direction of the motion of the Local Group (in a wedge of width 30° in declination). Rich clusters are plotted as larger circles, and poor clusters and groups as smaller ones. This plot of their position on the sky (right ascension) against mean redshift shows the existence of three distinct groups of clusters. The poorer clusters < 5000 km s^{-1} belong to the 'Hydra–Centaurus Supercluster'. The dense bunch of clusters at $\sim 14\,000$ km s^{-1}, between $13^{\rm h}$–$14^{\rm h}$, consisting of significantly richer clusters, is the 'Shapley Concentration'. The Abell 3571 group of clusters at $\sim 12\,000$ km s^{-1} belongs to a foreground filament-like structure.

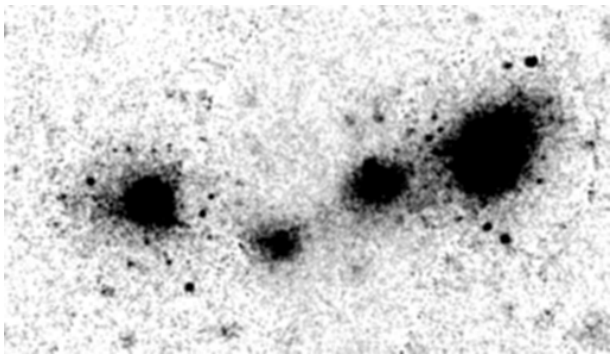


Figure 4. An x-ray map of the core of the Shapley concentration, from a mosaic of two long observations with the ROSAT PSPC. The cast of characters, from left to right, are the clusters Abell 3562, SC1329-31, SC1327-31 and Abell 3558 (Shapley 8). All of these clusters are around a redshift of 14 000 km s^{-1}. Shapley 8 is the central cluster of the Shapley concentration, and it is almost certain that the other three clusters are in the process of merging with it.

However, for galaxies in an extended region towards the direction of the motion of the Local Group, which includes the Shapley concentration, all galaxies smaller than the mean size of galaxies at about $45\,h^{-1}$ Mpc contribute only about 15% of the dipole acceleration due to those that are larger (nearer).

Its contribution to our peculiar motion

Both the values of the luminosity and the optical dipole of the Shapley concentration indicate that it could be responsible for 10–15% of the net acceleration of the Local Group, the rest being due to much nearer galaxies. This is consistent with the sum of the masses of its component clusters found from x-ray analyses.

If the Shapley concentration were a major source of the dipole acceleration of the Local Group, one would expect peculiar velocities to continue to rise beyond the nearer concentration of galaxies. At this moment there is conflicting evidence about whether this is seen or not.

This makes the Shapley concentration unlikely to be the Great Attractor, since that would imply (a) that mass does not follow light, and dark matter prefers to reside in clusters and superclusters of galaxies than in the field, and (b) that superclusters are capable of having higher mass-to-light ratios ($M/L_B \gtrsim 4000\,h$) than even the richest clusters of galaxies (e.g. for the Perseus cluster: $M/L_B \sim 350\,h$), over much larger scales. Both of these are against theories of GALAXY FORMATION from gravitational collapse.

Cosmological significance

The fraction of baryons in the universe, calculated from the abundance of primordial light elements in the universe and predictions from the standard Big Bang model, is $\Omega_{\rm b}/\Omega_{\rm m} \sim 0.014\,h^{-2}$. However, various studies have claimed that on scales of about $15h^{-1}$ Mpc within the core of the Shapley concentration, the observed baryon fraction from the visible galaxies and the x-ray emitting hot gas is about $\Omega_{\rm b}/\Omega_{\rm m} \geq 0.04\,h^{-1.5}+0.02$, where the first term comes from the gas and the second from the galaxies.

If this is indeed borne out by future research, it would suggest that the baryon fraction observed over the scales of dozens of clusters might exceed that of an $\Omega_{\rm m} = 1$ universe by a factor of about 3. Detailed hydrodynamic simulations show that baryons do not segregate into rich clusters and superclusters to such overdensities within the age of the universe. The above results would therefore imply that either the mean density of the universe is $\Omega_{\rm m} \lesssim 0.3$, or that the baryon density is much larger than predicted by NUCLEOSYNTHESIS.

It is not clear on what scales superclusters are bound, but there is evidence of merging clusters in the cores of the richest concentrations of clusters. The Shapley concentration provides the best such example (figure 4), with tenuous evidence of x-ray emitting gas being detected outside the clusters at its core. It is the best laboratory we know of in which to study the formation of nonlinear structures in a dense environment. It is also the place to look for answers to questions like: what are the largest coherent bound structures? On what scales can they have dynamical influences on each other within the lifetime of the universe? On what scales do coherent peculiar velocities exist? What is the typical morphology of these structures? From the answers to such questions it is hoped that important clues to the history of galaxy and cluster formation will emerge.

Bibliography

The Great Attractor was first proposed in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lynden-Bell D *et al* 1988 Spectroscopy and photometry of elliptical galaxies: V. Galaxy streaming towards the new Supergalactic Center *Astrophys. J.* **326** 19

For an exciting though idiosyncratic account of this collaboration, see

Dressler A M 1994 *Voyage to the Great Attractor: Exploring Intergalactic Space* (Knopf/Vintage)

The rediscovery of the Shapley concentration in the context of our peculiar motion can be found in

Raychaudhury S 1989 The distribution of galaxies in the direction of the 'Great Attractor' *Nature* **342** 251
Scaramella R *et al* 1989 A marked concentration of galaxy clusters—is this the origin of large-scale motions? *Nature* **338** 562

The concentration was named after Shapley in the first of these papers.

Harlow Shapley wrote several papers in the *Circulars* of the Harvard College Observatory in the early thirties. Notable among them in this context are

Shapley H 1930 Note on a remote cloud of galaxies in Centaurus *HCO Bull.* **874** 9
Shapley H 1930 The Super-Galaxy hypothesis *HCO Circ.* **350** 1
Shapley H 1932 Seventy-six thousand faint galaxies *HCO Bull.* **889** 1

For discussions of the cosmological implications of the Shapley concentration, see

Raychaudhury S *et al* 1991 *Mon. Not. R. Astron. Soc.* **248** 101
Ettori S *et al* 1997 *Mon. Not. R. Astron. Soc.* **289** 787
Hanami H *et al* 1999 *Astrophys. J.* **521** 90

Somak Raychaudhury

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shapley, Harlow (1885–1972)

Born in Nashville, MO, he studied with FREDERICK SEARES and at Princeton with HENRY NORRIS RUSSELL, where he took over the analysis of Russell's observations of the light curves of 90 eclipsing binary stars (ten times the number known up to then). In this work he established that Cepheid variable stars, whose spectral lines mimic the spectroscopic binary stars, are in reality pulsating stars. He joined the staff of the Mount Wilson Observatory in California, moving to the study of globular star clusters. He discovered Cepheid variable stars in some and calibrating HENRIETTA LEAVITT'S period–luminosity relation, determining their distances. The asymmetric distribution of globular clusters on the sky meant that the Sun is off the center of the distribution, some 50 000 light years away in the direction of the constellation of Sagittarius. Shapley identified this with the center of our Galaxy. Although Shapley had underestimated interstellar absorption and thus overestimated the size of the Galaxy, this established the size of our stellar system. Shapley took part in the great debate with HEBER CURTIS and successfully demonstrated that our Galaxy was a star system comparable to other galaxies. Shapley became director of the Harvard College Observatory, establishing modern outstations at the Oak Ridge (later, Agassiz) Station west of Cambridge, MA, and by moving the Arequipa Southern Station in Peru to become the Boyden Station in South Africa, each equipped with a 60 in telescope. He used the Cepheid variable stars to investigate the Magellanic clouds, to probe the spatial distribution of external galaxies. He discovered the dwarf galaxies in Sculptor and Fornax.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shen Kua (1031–95)

Born in Ch'ien-t'ang, Chekiang province (now Zhejiang), China, turned to science after a less than completely successful career as a general. He described his work on mathematics, astronomy, cartography, optics and medicine. He developed ideas on the motion of the planets which included techniques to predict retrograde motion, and used the techniques to improve predictions of the Moon. This led him to propose a new calendar, whose implementation required further observations of the sky, stimulating the development of astronomy in China. He made the first reference to a magnetic compass in his work *Meng ch'i pi t'an* (*Dream Pool Essays*). He also devised a celestial sphere, a bronze sundial and an improved clepsydra (water clock).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shepherd Moon

A minor satellite whose gravitational influence on the particles of a narrow planetary ring constrains the ring's radial extent. The ring systems of the giant planets are all shaped by gravitational perturbations of the particles that make up the rings. A lone shepherd moon can give a broad ring a sharply delineated outer edge, while pairs of shepherd moons can 'squeeze' the particles of narrow rings into well-defined orbits.

Shepherd pairs were originally hypothesized in the late 1970s by Scott Tremaine and Peter Goldreich to account for the narrow rings of Uranus. In their model, one satellite orbits just outside a ring, and one just inside it. The outer shepherd moves more slowly than the ring particles. When a particle on the outside of the ring passes it, the ring drags it back, causing it to lose energy and fall into a slightly lower orbit. The inner shepherd, in contrast, moves faster than the ring particles, so when it overtakes a particle on the inside of the ring its gravitational attraction gives an energy boost to the particle, lifting it into a slightly higher orbit. The net effect is to counteract the natural tendency of the ring particles to spread out, so maintaining the narrow ring.

The theory received support with the discovery of such pairs of shepherd moons, first in the Saturnian system in 1981, and then in the Uranian system in 1986. The minor satellites Prometheus and Pandora orbit either side of Saturn's F ring. The clumpy, braided appearance of the F ring has been ascribed to the slightly different eccentricities of the two shepherds. Another minor satellite, Atlas, orbits right at the outer edge of the bright A ring, and appears to act as a shepherd, preventing the ring's particles from spreading outwards. Another pair of shepherds, Ophelia and Cordelia, were found to be orbiting either side of the ε ring around Uranus.

What is not clear is how other narrow rings can exist without the channeling effect of shepherds. It could well be that they are in fact shepherded by as yet undiscovered small satellites. On the other hand, some other mechanism could be playing a part in constraining the particles in them.

See also: Neptune: rings, Saturn: rings, Uranus: rings.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shinsei

First Japanese scientific satellite, launched September 1971. Returned data on the ionosphere, solar wind and cosmic rays. Shinsei means 'new star'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shklovskii, Iosif Samuilovich (1916–85)

Soviet astrophysicist, educated and worked, particularly in radioastronomy, at Moscow State University and the Shternberg Astronomical Institute. He investigated the solar corona, showing that its temperature is of the order of a million degrees kelvin, and made theoretical and radio studies of supernovae and supernova remnants. In 1953 he proposed that the continuum radiation from the Crab nebula is synchrotron radiation, because of its broad spectral distribution, and predicted that the radio emission would be polarized by the magnetic fields causing the synchrotron radiation. This was demonstrated by Dombrovsky in 1954. From a very general theoretical model of how a supernova remnant expands into the interstellar medium he predicted the 1% per annum decrease of emission from the Cassiopeia A radio source, confirmed by Högbom and Shakeshaft in 1961. Shklovskii's book on *Intelligent Life in the Universe* was translated and expanded by CARL SAGAN.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shoemaker, Eugene ['Gene'] Merle (1928–97) and Shoemaker, Carolyn [née Spellman]

Astrogeologist Gene Shoemaker was born in Los Angeles, California and became chief scientist at the USGS Center of Astrogeology (Flagstaff) and professor of geology at CalTech. He worked with his wife Carolyn on cratering, both on the Moon and the Earth, by volcanoes and by meteor and cometary impact. He discovered, with Edward Chao, coesite, a type of silica produced in a violent impact and a signature that a terrestrial crater is meteoritic. With Eleanor Helin, Carolyn and he searched for potential Earth-impacting asteroids. As a result of this search (through which they found 13% of all known asteroids), they also discovered 32 comets which now bear the Shoemaker name, including Comet Shoemaker-Levy-9 which impacted on Jupiter in 1996. Gene died in a car crash in Australia while visiting meteor craters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Shoemaker–Levy 9–Jupiter Collision

The collision of comet Shoemaker–Levy 9 (SL9) with JUPITER, in July 1994, was an exceptional event for astronomers who were able to watch in real time the response of a planetary atmosphere to a large meteoritic impact. A wide variety of phenomena were observed such as temperature increase in the stratosphere, formation of aerosols and new gaseous species, and changes in the magnetosphere. Some of them remained visible in the following months and years. The observations were performed at all wavelengths, using all possible means from Earth and from space. Their interpretation has given an estimate of the size of the various impactors, their penetration level in the Jovian atmosphere, the energy released in the explosions, and the mass of the newly formed molecules. There is still uncertainty, however, about the chemical composition of comet SL9 itself.

The comet's history

Comet Shoemaker–Levy 9 was discovered in March 1993 by EUGENE AND CAROLYN SHOEMAKER and by David Levy, as a trail of about 20 individual fragments (figure 1). The study of its trajectory showed that it was a Jupiter-family comet (see COMETS) which was disrupted by tidal forces in July 1992 at its previous perijove passage, and that it would collide with the planet at its next perijove passage, in July 1994 (figure 2).

As a result of the discovery of comet SL9, a huge international effort took place using a wide variety of ground-based telescopes and instruments, working at all wavelengths from the ultraviolet to the radio range. Observations were also planned from spacecraft, in particular the HUBBLE SPACE TELESCOPE (HST) and GALILEO, but also the INTERNATIONAL ULTRAVIOLET EXPLORER (IUE) and the x-ray satellite ROSAT. Questions to be addressed were the amount of released energy, the temperature evolution, the size of the impactors, their physical nature and their chemical composition. In the meantime, the SL9 fragments, labeled from A to W, were repeatedly monitored from the the HST and ground-based telescopes between March 1993 and July 1994.

A collision of this type is very rare: according to the models of orbital dynamics, such an event might be expected once every few hundred years. A similar event might indeed have taken place about three centuries ago: as reported by the archives of Paris Observatory, Jean-Dominique Cassini reported in 1690 an unusual feature on the disk of Jupiter and his drawings show a remarkable similarity with the craters observed after the SL9 collision in 1994.

The fragments of comet SL9 entered the Jovian atmosphere between 16 July and 22 July 1994, at a latitude of -44°, only a few minutes within the dates predicted by astrometry. The impacts took place just behind the limb, and came into direct viewing from Earth about 10 minutes later, due to Jupiter's rotation. The Galileo spacecraft, then at a distance of 1.6 AU from Jupiter, benefited from a direct viewing of the impacts. During the week of impacts, all fragments collided on the $l = -44^\circ$ parallel, at different longitudes, as an effect of the planet's rotation. All explosions were extensively studied from almost all observatories; a huge temperature increase was measured in the stratosphere of Jupiter, dust was formed and new molecules were detected; changes were reported in the magnetosphere (see JUPITER: MAGNETOSPHERE).

The impacts

Three types of data have been recorded during the impacts: (a) the lightcurves, at different visible and infrared wavelengths, have given the temperature evolution and the energy budget; (b) images taken from the ground, from the HST and from Galileo, have shown the meteor entry, the ejecta trajectories and crater evolution; (c) spectra, recorded in the ultraviolet, visible, infrared and radio range have shown evidence for new species. For all observed impacts, from the recorded lightcurves, events can be interpreted as a sequence of three main phases (figure 3): the 'entry' phase, the 'explosion' phase and the 'splash' phase. The entry phase, which took place behind the limb as seen from Earth, was mostly observed by Galileo and we have little information about it; it was sometime detected as a first weak emission in the ground-based lightcurves. The explosion phase, about 1 min later, was observed as a strong emission, especially in the visible range, corresponding to the explosion of the impactor within the atmosphere and the ascension of the fireball; here again, most of our knowledge about this phase comes from Galileo. The splash phase, which starts about 6 min after the explosion and lasts for about 15 min, corresponds to the heating of the Jovian stratosphere by the ejecta as they fall back on the planet; this phase was extensively studied from Earth. The intermediate period between these two last phases was mostly observed by the HST which was able to measure the maximum altitude reached by the EJECTA (over 3000 km).

The entry phase

The lightcurve of the entry phase or 'meteor phase' has been mostly observed by the imaging system (SSI) and the photopolarimeter (PPR) of the Galileo spacecraft. Several ground-based observatories also recorded a weak, early event, about 1 min before the explosion itself; it probably corresponds to a meteor storm accompanying the main fragment.

Several attempts have been made to detect light echoes on Galilean satellites, but they were all unsuccessful; the corresponding upper limits for the radiated energy are compatible with the estimates derived from the Galileo lightcurves.

The explosion phase

From a modeling of the lightcurves by Crawford, McLow and Zahnle, and with the use of the plume altitudes observed with the HST, estimates of the fragment diameters have been obtained. They range between 150 m

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

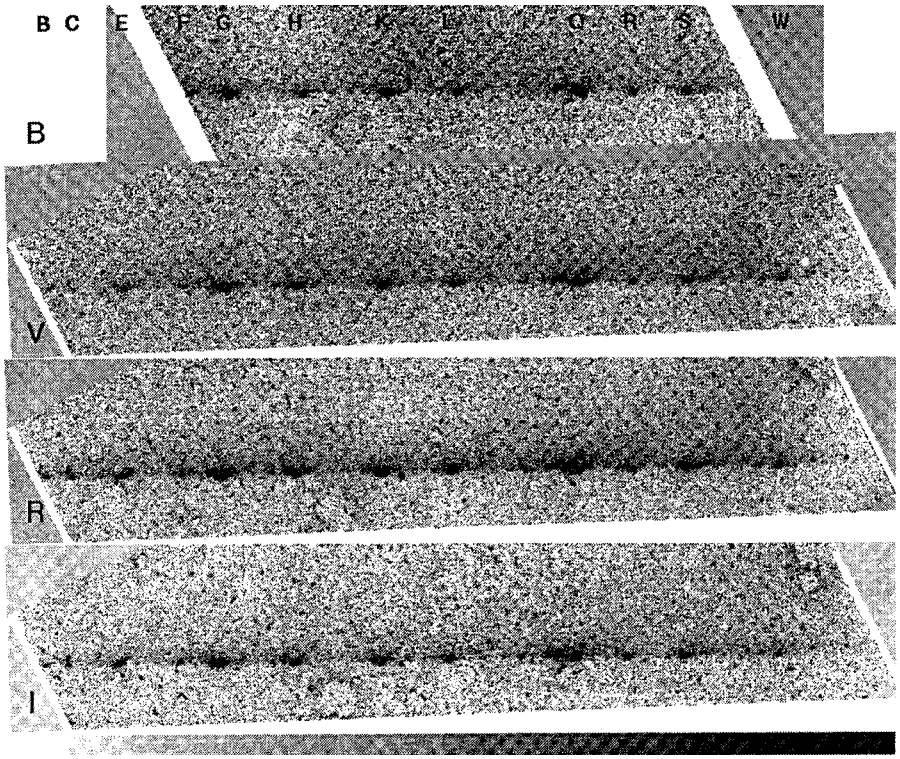


Figure 1. The fragments of comet Shoemaker–Levy 9 as observed from the 3.5 m New Technology Telescope (NTT) at ESO, La Silla, on 1 and 2 July 1994. About 12 of the 20 identified fragments, labeled from A to W, are visible in theses pictures (from top to bottom: filters B, V, R and I). The figure is taken from Stüwe *et al* in West and Bohnhardt (1995, pp 17–22).

and 600 m, which corresponds to H_2O masses of 2×10^{12} and 10^{14} g respectively. These values also agree with the estimate deduced by Asphaug and Benz from the trajectories of the cometary fragments after disruption. Calculations show that the altitude reached by the ejecta is about the same (about 3000 km) for all fragments whatever their size is, as the smaller fragments explode at a higher level than the more massive ones. The total energy released is about 3×10^{27} erg for the biggest impactors (K and G impacts), and about 10 times less for the smaller ones. However, even in the case of the big impacts, the energy is not sufficient to lead to detectable seismic waves, for which only an upper limit was obtained by Galdemar and Mosser.

The temperature of the fireball resulting from the explosion has been estimated by Martin and Orton from the visible and near-IR lightcurves measured by the Galileo SSI, NIMS and PPR experiments. From over 10 000 K at the beginning, it decreased to about 2000 K after about 15 s. According to the NIMS data, the size of the fireball increased adiabatically from 15 km, 10 s after the explosion, to 100 km after 40 s. There is still some uncertainty about the penetration level of the fragments, as some observations (including infrared emission lines of NH_3) suggest that the explosion of the biggest fragments took place below the NH_3 cloud, at a pressure higher than 0.5 bar, while the interpretation of the Galileo NIMS observations favors a higher penetration level, at about 0.2 bar. In any case, the explosions are likely to have taken place in the upper troposphere or near the tropopause, in the pressure range 0.1–1 bar.

The splash phase

Information upon the temperature evolution in this third phase comes from the lightcurves, but also from the spectra of the emission bands of CH_4, especially in the strong ν_3 band around 3.3 μm. Methane, a spectroscopically very active gas, is one of the main minor constituents of the Jovian atmosphere (see JUPITER: ATMOSPHERE) and its mixing ratio, with respect to the main constituent H_2, is well known (about 2×10^{-3}). From its emission spectrum, one can derive the temperature of the emitting region. Measurements obtained by this method indicate a decrease from about 1000 K at $T = 10$ min down to about 600 K at $T = 25$ min.

Apart from the information about the temperature, spectra obtained during the splash phase have led to the discovery of new species, both gaseous and solid. The presence of dust was visible in the early part of the splash phase, with the formation of dark craters at the center of the impacts, surrounded by crescent-shaped areas due to infalling ejectas (figure 4). Infrared spectra have shown the presence of silicates, which suggests, for the impactor, an O/C ratio larger than 1, and a cometary origin. The dust

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

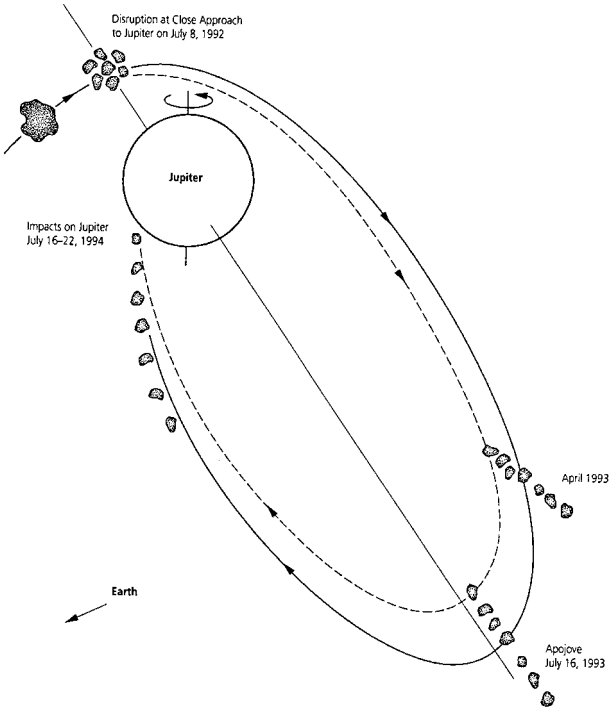


Figure 2. The orbit of comet SL9 between July 1992 (time of tidal break-up) and July 1994 (time of collision with Jupiter). This schematic diagram is not drawn to scale. (Prepared by Z Sekanina, P Chodas and D Yeomans.)

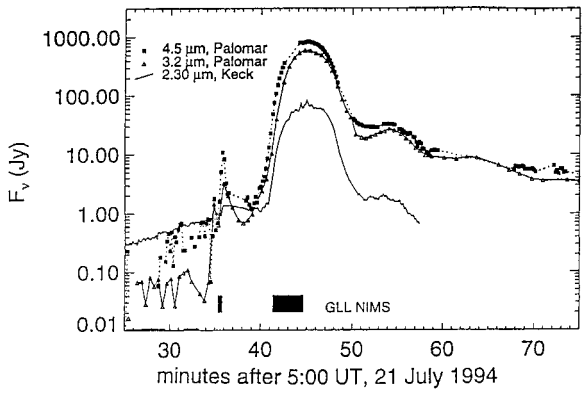


Figure 3. Lightcurves of impact R, observed with the Keck (10 m) and Palomar (6 m) telescopes at three different infrared wavelengths. The black bars indicate the time intervals at which NIMS–Galileo observations were recorded. The time of explosion is 05:34:57 (Chodas and Yeomans 1996). The first weak maximum corresponds to the entry phase ($t = -1$ min). The second maximum corresponds to the explosion phase ($t = 0$–2 min) and the third maximum, brighter and extended in time, corresponds to the splash phase ($t = 6$–25 min). The figure is taken from Nicholson (1996).

probably also contains a carbonaceous component, but no firm identification could be derived from spectroscopy. The total volume of dust injected in the stratosphere has been estimated by West to make up a sphere of about 1 km diameter, with particles of 0.25 μm diameter. From their infrared signature, the total mass of the silicates has been estimated to 8×10^{13} g, or 10–20% of the total volume of the impactor.

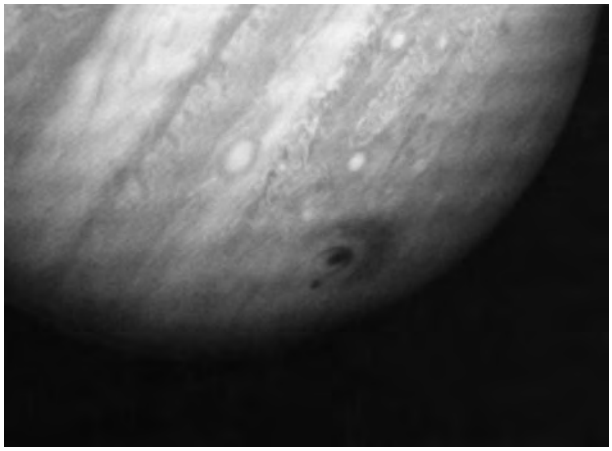


Figure 4. This image, taken with the wide-field camera of the Hubble Space Telescope, shows the impact craters for fragments D and G. The larger crater was created by impact G on 18 July 1994. The central hole is distinctly surrounded by two concentric rings. The smaller crater on the left of the G impact site is due to impact D which took place a day earlier. This image was taken 1 h 45 min after impact G (Hammel, NASA, ESA).

Many new gaseous species have been detected in the stratosphere during the splash phase: CO, H_2O, S_2, CS_2, CS, OCS, HCN, C_2H_4, NH_3 (previously present mostly in the Jovian troposphere), and possibly PH_3 and H_2S. These discoveries and their implications have been reviewed by Lellouch. The most abundant of the newly formed species was CO, detected both in the infrared and in the radio range. A remarkable result was the high excitation temperature deduced from CO infrared observations, which increased from 2000 to 5000 K during the first 10 min of the splash phase. In contrast with other molecules, CO is not dissociated at these temperatures. The temperature increase has been explained by Zahnle on the basis of a ballistic model, in which the CO molecules having the highest excitation temperature are the ones which, being lifted at highest altitudes, fall down later. H_2O was detected in the beginning of the splash phase; it was observed both by the near-infrared spectrometer of Galileo (NIMS) and from the Kuiper Airborne Observatory; its excitation temperature was about 1000 K. S_2, CS_2, CS and possibly H_2S (although this latter detection is tentative) were found from ultraviolet spectroscopy with the HST. In addition to CO, ground-based millimeter heterodyne spectroscopy provided the detection of HCN, CS and OCS, while C_2H_4 and HCN, together with NH_3 emission lines, were found in the infrared region, at 10–13 μm. The infrared and millimeter spectroscopic observations have also provided key information about the evolution of the stratospheric thermal profile.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Abundances (in g) of detected species and corresponding elemental masses following the impacts of SL9 on Jupiter. Elemental abundances expected from a comet nucleus and from a carbonaceous chondrite are shown for comparison (from Crovisier 1996).

Molecule	Observation†	Molecular mass‡ in an SL9 impact	Elemental mass in an SL9 impact	Elemental mass in a 10^{15} g fragment	
				Comet	CI chrondrite
H_2O	**IR**, radio	$>2 \times 10^{12}$	$>1.5 \times 10^{14}$ [O]	5×10^{14} [O]	5×10^{14} [O]
CO	IR, **radio**	2.5×10^{14}			
NH_3	UV, **IR**	1×10^{13}	1×10^{13} [N]	2×10^{13} [N]	3×10^{13} [N]
HCN	IR, radio	6×10^{11}			
S_2	**UV**	1.5×10^{12} ?			
CS_2	**UV**	1.5×10^{11}			
CS	UV, **radio**	5×10^{11}	5×10^{12} [S]	4×10^{13} [S]	6×10^{13} [S]
H_2S	UV	?			
OCS	**Radio**	3×10^{12}			
			1×10^{14} [C]	2×10^{14} [C]	3×10^{13} [C]
PH_3	IR	?	3×10^{11} [P]	1×10^{12} [P]	
Silicates	**IR**	6×10^{12}	8×10^{13} [Si]	1×10^{14} [Si]	
				4×10^{13} [Mg]	1×10^{14} [Mg]

† In bold are the observations from which the abundances are taken.
‡ Some of these values are preliminary and subject to revision.

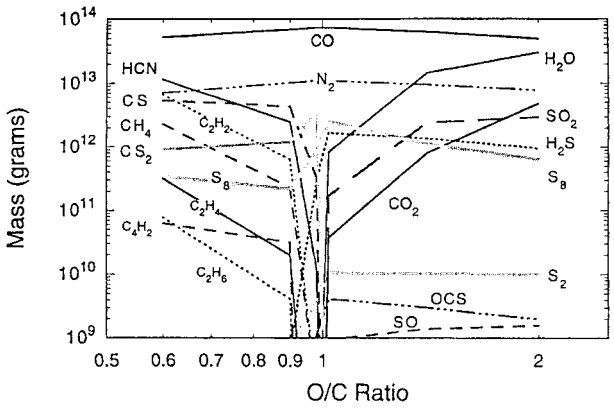


Figure 5. Total production of a mass of gas, 50% Jovian and 50% cometary, produced by a comet of 10^{14} g, as a function of the cometary O/C ratio. This model assumes gaseous phase chemistry. The production corresponds to temperatures ranging from 1500 to 5000 K. Comparison with table 1 shows that the global composition of the newly formed species cannot be explained by a single value of the O/C ratio. The figure is taken from Zahnle (1996).

The relative abundances of the newly formed species can provide information about the chemical nature of the impactors. Assuming that the emitting region is formed with material containing 50% of Jovian material and 50% of cometary matter, Zahnle has calculated the expected relative abundances of the expected new molecules, as a function of the O/C ratio (figure 5). If oxygen dominates, the expected new species are CO, N_2 and H_2O, then SO_2, SO and CO_2. In contrast, if carbon dominates, then CO, N_2, CS, HCN, C_2H_2, CH_4 and CS_2 are expected. The puzzling result is that the newly formed species belong to the two lists, which makes the interpretation difficult. This might imply that the chemical composition was different for the various impacts, but the origin of such differentiation would remain to be explained. Inversely, if a cometary O/C ratio is assumed, then the relative abundances observed for the new species imply that the emitting region must have been mainly formed of Jovian gas.

Another surprising result was the detection, in the visible range, of a large number of atomic and ionized lines at the beginning of the splash phase. In particular, lithium was observed for the first time in an outer solar system body. These emissions have to be of cometary origin; however, as discussed below, it is difficult to derive the cometary abundances of these species, because their excitation mechanism is not fully understood at present.

The long-term evolution of impact phenomena

An extensive monitoring of the Jovian disk, both by imaging and spectroscopy, took place during the months following the impacts, and some of the observations have been periodically repeated during the three following years.

Evolution of the thermal profile

The thermal evolution of the Jovian stratosphere has been studied using the spectral profiles of the various emission lines observed during and after the impacts: CH_4 at 8 μm, C_2H_2 at 13 μm, CO at 4.7 μm, and CO, CS and HCN in the millimeter range. The stratosphere cooled rapidly (figure 6), and had returned to its nominal temperature within about a month; this rapid cooling has been explained by Bezard, taking into account the presence of the silicate particles injected into the stratosphere at the time of the explosion, which efficiently contribute to its radiative cooling.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

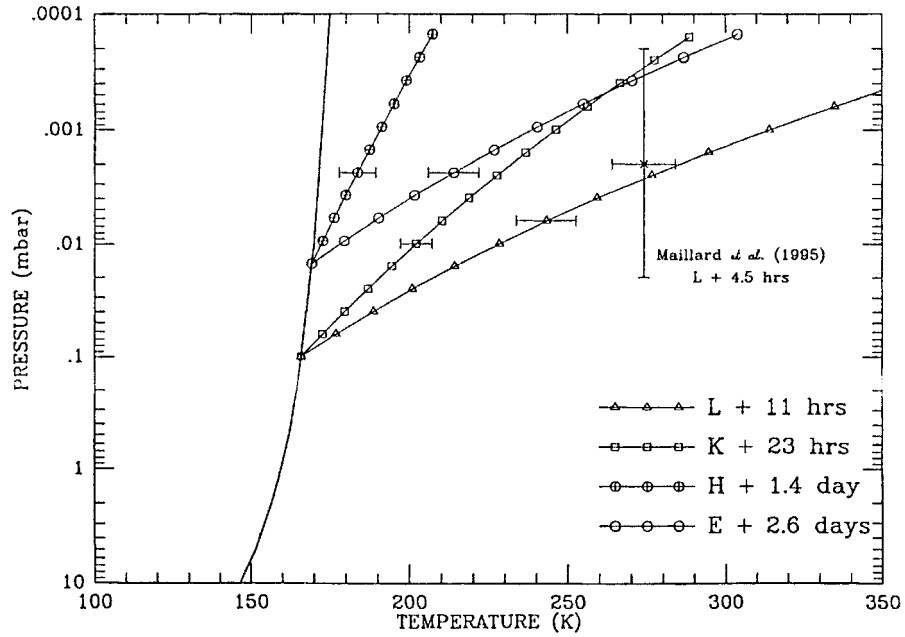


Figure 6. Evolution of the Jovian stratospheric temperature after impacts L, K, H and E. Calculations of radiative cooling show a faster cooling than anticipated by the Jovian gases only, which can be explained by the effect of newly formed silicates. The figure is taken from Bézard *et al* (1997).

Evolution of the newly formed molecules

The newly formed species had very variable lifetimes. H_2O disappeared after a few hours, while S_2, OCS and NH_3 lasted for a few weeks. CS_2 and CO were detectable for several months; CS and HCN lasted several years. A good general agreement was found between the observed evolution of these species and the predictions provided by the photochemical evolutionary models developed by Moses.

Evolution of the dust

The clouds formed at the time of impacts were extensively monitored in the weeks and months following the impacts. The horizontal dust distribution was observed with UV HST images and ground-based images within the CH_4 bands, probing the high stratosphere. The clouds first spread in longitude during the first few days, as an effect of the fast rotation of Jupiter. Then they elongated to form a continuous dark band along the -44° parallel. During the year following the collision, they spread in latitude, between -20° and -80°. The density of aerosols was reduced by a factor 4 to 10 after 10 months; the aerosol layer was still weakly detectable after one year.

Dust particles were originally located at high stratospheric levels, stable against convection, and they were sufficiently light to have a residence time of about a year, as estimated by West. Their vertical motion was monitored through the evolution of the stratospheric profile, as they are mostly responsible for its cooling. After a year, the areosol layer had reached the tropopause level at 0.1 bar.

The morphology of the ejecta showed two concentric rings which expanded at constant velocities of about 450 m s^{-1} for the outer ring and 210 m s^{-1} for the inner ring; it was suggested by Ingersoll that they possibly originated from waves. The radial speed of the outer ring, however, is too slow for an acoustic wave and too fast for a gravity wave, so that the origin of the rings is still an open question.

Magnetospheric effects

The radio emission

The radio emission of Jupiter has three types of components: the thermal millimeter radiation, the synchrotron emission of the electrons along the magnetic field lines, in the centimeter–decimeter range, and the outbursts from localized radio sources, in the decametric range.

The Jovian magnetosphere has also been perturbed by the collision. If little change was observed at decimetric or decametric wavelengths, the synchrotron radiation, observed in the centimeter range, showed a significant enhancement (by 10 to 40%) during the week of impacts, as well as some modification of its properties. Most of the emission was confined to one hemisphere, between 100 and 240° longitude. The observations apparently imply the existence of a population of newly excited electrons which are confined in a particular region of the magnetic field. The hemisphere of maximum radio emission correlates with the side where the magnetic field lines crossing the radiation belts intercept the Jovian disk in the region of latitude -44°, i.e. where the impacts took place. Several mechanisms have been proposed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for the origin of the new population of electrons: radial diffusion enhancement, shock acceleration and pitch angle scattering; as discussed by Bolton, more than one mechanism is probably required to account for the observed data.

Auroral phenomena

Changes have been also recorded in the AURORA of the planet. The first sign of auroral activity may have been an Mg^+ flash fortuitously recorded from the HST at the G-condensation, on 14 July 1994, four days prior to the entry, presumably corresponding to the G-entry in the Jovian magnetosphere. Other UV signatures were recorded from the HST in the H and H_2 lines a few hours before the impacts; they were probably emitted in auroral flares when the fragments crossed high L-shell magnetic lines. A strong infrared H_3^+ emission was recorded a few minutes before the C impact, implying a strong heating of the upper stratosphere.

A remarkable result was the detection of emissions, on several occasions, at the location of the northern counterparts of the impacts (the footprints of connecting magnetic field lines): such an x-ray emission was observed by the ROSAT satellite at the time of the K and P2 impacts; UV emissions of norther counterparts were recorded by the HST just after the splash phase (t = 30–60 min), and infrared H_3^+ emissions of northern counterparts were reported a few days after the impacts. As reviewed by Ip, all these results, together with the observed enhancement of the synchrotron radiation, suggest that field-aligned current systems established a connection between the impact sites and their magnetic conjugate points.

The auroral activity of Jupiter was monitored in the UV with the IUE satellite during the impacts, and in the IR from ground-based H_3^+ observations during the week following the impacts. An enhancement of the northern auroral emission was recorded a few days after the last impacts: its ratio to the southern emission increased to a value of 7–8 on 25 July, and then returned to its normal value of 1–3. The decrease of the southern aurora may have been associated with chemical effects induced by newly formed material reaching high southern latitudes.

What have we learnt about comet SL9?

History of SL9

On the basis of more than 3000 astrometric measurements, Chodas and Yeomans have been able to predict with high accuracy the impact times of all fragments. The set of predicted impact times was systematically early by about 7 min, probably due to systematic errors in the reference star catalogs used for the data reduction. The timing of the impacts as measured by Galileo was found to be accurate within 10 s of the predictions, corrected from the systematic effect, while the impact timing derived from ground-based observations was accurate to within a few minutes. Using the observed impact times, new orbit solutions were generated, leading to the best orbital fits ever obtained for a comet.

On the basis of its retrieved orbit, the pre-break-up history of comet SL9 has been inferred by Chodas and Yeomans, using a Monte Carlo analysis. The progenitor of comet SL9 was most likely captured by Jupiter around 1929 $\pm$ 9 years. Before capture, the comet was in a low-eccentricity, low-inclination heliocentric orbit, probably inside Jupiter's orbit, consistent with a group of Jupiter-family comets called the quasi-Hildas.

Evolution and activity of SL9

Studying the possible break-up scenarios of comet SL9 can provide important constraints upon its internal strength. Two extreme models have been proposed, with on the one hand a coherent body subject to sequential crack formation, and on the other side a rubble pile/swarm model. Neither of them is fully satisfactory, which led Rickman to propose a new scenario involving processes from both models. He estimated the tensile strength of the comet to be about 100 Pa, a very low value consistent with a fluffy aggregate of sub-micron particles.

Asphaug and Benz have modelled the elongation of the swarm, on the basis of tidal forces and self-gravity, and they have calculated the expected chain length as a function of the comet density and the size progenitor. Comparison with the actual chain length favors a 1.5 km progenitor with a density of 0.5 g cm^{-3}. This estimate is consistent with the mean size of the impactors derived from the dynamical simulations.

The activity of the comet prior to the impact has been studied from the HST and ground-based images. The break-up of the comet apparently triggered the activity of the fragments which were continuously surrounded by a dust coma from July 1992 to July 1994; estimates of dust production rate are rather low (between 1 kg s^{-1} and 5 kg s^{-1} for the largest components). All attempts to detect gaseous activity, in the visible, near-UV and radio range, were unsuccessful. Upper limits of production rates were in the range of 10^{27} s^{-1} for OH, and 3×10^{23} for CN. Taking into account the small size and the low dust production rate of SL9, the absence of a gaseous signature in its spectrum is not surprising.

Chemical composition of SL9

A spectacular result was the ground-based detection of the atomic lines in the visible range, in the splash phase of impacts L, Q1 and Q2. The observed species were Fe, K, Ca, H, Na, Mn, Mg, Cr and, for the first time, Li. Other atomic and ionic transitions were also detected in the UV from the HST: H, He, S, Si, Mg^+, Fe^+, Si^+, Al^+ .... All these elements, absent from the Jovian outer atmosphere, are indubitably coming from the impactor; their signatures are also found in the spectra of sungrazing comets. These emissions are usually excited by fluorescence. However, as discussed by Crovisier, this mechanism does not seem to be responsible for the SL9 emissions, or at least it cannot be the only one: the observed lines are heavily saturated, and their intensities and widths are too large to be explained by fluorescence. Other possible mechanisms

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

could include thermal excitation by collisions, electronic recombinations, or prompt emissions following chemical reactions; all these time-dependent excitation phenomena are unfortunately very difficult to model. No firm conclusion can thus be derived at present about the abundances of the observed atom and ion species.

The chemical composition of the comet is also difficult to retrieve from the newly formed molecules. Indeed, all cometary elements must have been vaporized at the time of the explosion; the new molecules observed are not coming directly from the comet, but are recombination products, most likely formed in the early part of the splash phase. In addition, the newly formed species can be partially of planetary origin, so that even the atomic abundances derived from their mass may not necessarily provide reliable determination of the comet elemental masses.

Some information, however, can be retrieved from the relative abundances of the newly formed species; this study has been performed, in particular, by Lellouch and Crovisier (table 1). The oxygen- and sulfur-bearing species are known to be absent from the Jovian stratosphere; their abundance can thus be used to infer the cometary mass of oxygen and sulfur. The derived values are 5×10^{12} g for S and $(1–3) \times 10^{14}$ g for O; the derived S/O ratio is compatible with both cometary and asteroidal abundances, and does not allow us to discriminate between the two origins. The cometary abundance of carbon cannot be retrieved from newly formed carbon-bearing species, as carbon might come from Jovian CH_4. The case of nitrogen is more dubious: some newly formed NH_3 was probably of cometary origin, but some may have come from the NH_3 cloud if this cloud was hit by the meteor. The derived abundance of stratospheric nitrogen was 10^{13} g. Assuming an impactor origin for N, the inferred N/O ratio would clearly favor a cometary origin.

Another important parameter is the composition of aerosols. The signature of silicates was identified at 10 μm, leading to an estimate of 8×10^{13} g for the Si abundance. The inferred S/O ratio is in good agreement with a cometary composition.

Was SL9 a comet or an asteroid? Although no definite conclusion can be drawn on the basis of its chemical composition, several arguments strongly favor the cometary origin: the small size, the activity level, weak but real, the low density and the low tensile force, and the silicate signature. SL9 was most probably a small, very common Jupiter-family comet which would never have caught astronomers' attention without its premature death in a crash with Jupiter.

Conclusions

Several years after the collision, there is still work to be done on the modeling and interpretation of the huge database collected at the time of impacts and in the following weeks and months.

The coordinated observation of the entire impact has allowed a repeated intercomparison of all observations, and their interpretation in the light of theoretical impact models and fireball evolutionary models. There is a good overall understanding of the general phenomenon, which has allowed a determination, at least to first order, of the main physical parameters involved in the collisions: energetic budget, size of the fragments, penetration levels. The abundance of newly formed species has been analysed in the light of shock chemistry models, and their temporal evolution is in good agreement with the predictions of photochemical evolutionary models.

From the global set of astrometrical measurements, the history of the comet has been determined with high precision. However, the physical and chemical nature of the comet itself is still poorly known. In summary, one could say that the collision of comet Shoemaker–Levy 9 with Jupiter has not allowed us to fully understand the nature of the comet; in contrast, it has provided a huge amount of information about the physics of large meteoritic impacts, as well as on the behavior of a planetary atmosphere after such impact.

Bibliography

Articles of the authors quoted above can be found in the following references:

Special journal issues

Earth, Moon Planets 1994 **66** (1)
Science 1995 **267**
Geophys. Res. Lett. 1995 **22** (12, 13, 17)
Icarus 1996 **121** (2)
Planetary and Space Science 1997 **45** (10)

Noll K S, Weaver H A and Feldman P D (eds) 1996 *The Collision of Comet Shoemaker–Levy 9 and Jupiter* (Cambridge: Cambridge University Press)

Spencer J R and Mitton J (eds) 1995 *The Great Comet Crash* (Cambridge: Cambridge University Press)

West R M and Bohnhardt H (eds) 1995 *Proc. European Workshop on the SL9–Jupiter Collision* no 52 (Garching: ESO)

Thérèse Encrenaz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Short-period Comet

A comet whose period is less than 200 years, also known as a *periodic comet*. Comets with longer periods are termed *long-period comets*; these are also 'periodic', but a comet can be incontrovertibly classified as periodic only when it has been observed on at least two returns. The choice of 200 years as a cut-off is a historical convenience, corresponding roughly to the length of time for which it has been possible to calculate orbits with sufficient accuracy to predict a return; also, no comet with a calculated period of more than 200 years has yet been observed to return since its discovery. In the system of cometary designations introduced in 1995, a short-period comet is assigned the prefix 'P/' and a number indicating the order in which the periodicity of its orbit was established, in a sequence beginning with 1P/Halley. As of April 2000 there were 143 known short-period comets. In addition there are over 40 comets with calculated periods of less than 200 years that have been observed at only one return, plus some two dozen short-period comets that have failed to show up when expected or have been observed to disintegrate.

Most short-period comets have orbital inclinations of less than 30° (15° is typical) and move in direct orbits. The shortest period, 3.3 years, is possessed by 2P/Encke; the longest, 155 years, by 35P/Herschel–Rigollet. Many short-period comets have been perturbed into their present orbits by the gravitational influence of Jupiter and have aphelia near 5.2 AU, Jupiter's distance from the Sun, and orbital periods of 6 to 12 years—in fact, three-quarters of all short-period comets have periods that fall within this range. These so-called *Jupiter-family comets* are believed to have originated in the Kuiper Belt. Those with periods of 20 to 200 years, known as *random* or *Halley-type comets*, are, like the long-period comets, thought to come from the Oort Cloud.

See also: comets, comet family, Kuiper Belt, long-period comet, Oort Cloud, solar system: nomenclature.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sidereal Period

A period of time measured relative to the 'fixed' stars. In normal usage the term refers to the time taken by a planet to complete one orbit of the Sun, returning to its original position relative to the position of the stars, or the time taken for a satellite to complete one orbit of its parent planet, again measured relative to the background stars. In the case of the Earth, this period of time is referred to as the sidereal year and is equal to 365.2564 mean solar days.

See also: synodic period.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sidereal Time

A time system based on the rotation of the Earth measured relative to the background stars, which for this purpose are regarded as fixed in position. Relative to the stars, the Earth rotates on its axis in a period of 23 h 56 min 04.1 s of mean time (i.e. ordinary civil time), and this period is called the sidereal day, which is, in turn, divided into sidereal hours.

For an observer located at a particular longitude on the Earth, the sidereal day is equivalent to the apparent rotation period of the celestial sphere. This period may be determined by measuring the interval between two successive upper transits of a given star across his meridian. The value of the sidereal time at any instant is defined to be the hour angle, i.e. the angle measured clockwise from the meridian, of the vernal equinox (a fixed point on the celestial sphere). Thus, when the vernal equinox is on the meridian, its hour angle is zero, and the sidereal time is zero hours. By the time that the rotation of the celestial sphere has increased this angle to 90°, 6 h have elapsed, and the sidereal time is 6 h, and so on. After 24 h, the vernal equinox returns to the meridian.

The hour angle measured by this observer relative to his meridian is the local hour angle, and the sidereal time so obtained, the local sidereal time (LST). Greenwich sidereal time (GST) is the hour angle of the vernal equinox measured from the Greenwich meridian. The difference between LST and GST corresponds to the longitude of the observer expressed in time units (where 1 h is equivalent to 15°, since the Earth rotates through 15° per hour). If the observer is east of Greenwich, LST is greater than GST, if he is west of Greenwich his LST is less than GST.

See also: equinox, hour angle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Siderite

An obsolete term for an iron meteorite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Siderolite

An obsolete term for a stony-iron meteorite.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Siderostat/heliostat/coelostat

A plane mirror that is driven about an axis so as to reflect light from a particular celestial object along a fixed direction. In practice, a coelostat usually comprises two plane mirrors. One mirror is driven around an axis parallel to that of the Earth at half the Earth's rotation rate in order to counteract the apparent rotation of the celestial sphere and ensure that rays from a particular object are reflected along a constant direction. The other mirror is orientated so as to reflect the beam of light from the first mirror into a fixed telescope. A siderostat uses a mirror driven at the same angular rate as the rotation of the Earth but suffers from the drawback that although a star located at the center of the field of view remains stationary, the rest of the field of view rotates about that central point.

Where the mirror, or mirrors, is (are) driven to follow the apparent motion of the Sun, the device is called a heliostat. Systems of this kind are used in a number of specialist solar telescopes. For example, the McMath solar telescope at Kitt Peak, Arizona, uses a 2 m diameter heliostat, mounted at the top of a 30 m tower, to reflect light down a cooled inclined tunnel, about 150 m long, to a reflecting telescope system that produces a fixed image of the Sun, some 76 cm in diameter, in a laboratory.

See also: celestial sphere, diurnal motion, solar feeds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sigüenza y Gongora, Carlos de (1645–1700)

Spanish colonial astronomer, born in Mexico City, became professor of astrology at the university but seems strongly to have opposed astrology and, in 1681, wrote on comets to calm fears aroused by the one of 1680–1. Appointed by decree of Charles II of Spain as royal cosmographer, he mapped New Spain, Mexico and Pensacola Bay and determined the longitude of Mexico City. He published almanacs and observed the solar eclipse of 1691.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sikhote Alin Meteorite

An iron meteorite that fell in the Sikhote Alin mountain range in southeast Siberia on 12 February 1947. It was preceded by a brilliant fireball. The incoming body fragmented at a height estimated at about 5 km, and the fragments produced nearly 400 small impact craters, the largest 27 m across. The total of 23 tonnes that was collected included a fragment with a mass of 1.7 tonnes. According to some estimates there may be as much as 75 tonnes of material uncollected.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sirius

Sirius is the brightest star in the night sky and lies in the constellation CANIS MAJOR. In reality, Sirius is a BINARY SYSTEM consisting of a main sequence star (Sirius A) and a WHITE DWARF companion (Sirius B). Other common names for this star system include Canicula, Aschere and the Dog Star.

Canis Major, which translates to the greater dog, is one of the hunting dogs of Orion, with CANIS MINOR being the other. The Sirius binary star system represents the nose of the greater dog, and can be easily found by following the line made by the three stars comprising the belt of ORION.

Several centuries ago, Sirius was believed to be a single star. In 1862, nearly 20 years after a companion star to Sirius A was first suspected to exist, ALVAN CLARK discovered Sirius B while testing a new lens he had made for Dearborn Observatory. The location of this binary system in the sky is at right ascension $06^h\,45^m\,09^s$ and declination $-16^\circ\,42'\,57''$. The distance to these two stars is 8.6 light years, or just over 80 trillion kilometers. Sirius A and Sirius B constitute a VISUAL BINARY, which means that each of the two stars can be individually resolved through the use of a telescope.

Sirius A, which has a spectral type A1 V, is considerably larger than our own Sun. Sirius B, spectral type DA, is roughly four times the size of the Earth. Although the ratio of the sizes of these two stars is roughly similar to the same ratio for the Earth–Sun system, the mass ratios of the two pairs are quite different. It would not be unreasonable to approximate the Earth–Sun system as the Earth moving in a nearly circular path about a fixed Sun. Sirius A and Sirius B, however, must be considered as two stars that each orbit about their common center of mass. It should be noted that the Earth and Sun also orbit about their own common center of mass, but that lies very close to the center of the Sun while center of mass of the Sirius system is relatively well removed from either star.

The apparent magnitude for Sirius A is -1.5, which makes it appear to be the brightest star in the night sky; its absolute magnitude is 1.4. The apparent and absolute magnitudes for Sirius B are 8.3 and 11.2, respectively. Note that the difference between absolute and apparent magnitudes for each star is the same. This is a direct result of the fact that they are at the same distance.

The evolutionary state of the Sirius binary system today is quite different from what it must have been in the past. Sirius B, which is now a white dwarf, was most likely a RED GIANT. When the old Sirius B exploded, its core remained behind to become the Sirius B we are familiar with today. Estimates for the earlier Sirius binary system place the mass of Sirius B near four solar masses, the separation between the two stars at seven astronomical units, and the orbital period about the center of mass at seven years.

Sirius A recently provided the first proposed detection of a stellar wind from a spectral type A star. There is no currently known mechanism capable of initiating a stellar wind in A-type stars. However, if such a mechanism does exist, it is believed that radiation pressure would be capable of fueling such a wind. The nearby proximity of Sirius A, along with its white dwarf companion, provides a unique environment in which to search for such winds and their origin.

An intriguing puzzle has surrounded Sirius A for nearly two thousand years. Specifically, observations by PTOLEMY and others suggest that Sirius A was considerably redder during their time. Several theories suggesting mechanisms capable of producing the observed color differences within a time frame of 2000 years have been investigated. One of the more viable explanations is that Sirius A appeared redder due to interference caused by the Earth's atmosphere, given the position from which Ptolemy was most likely to have made his observations.

Because Sirius B has a mass very close to that of our Sun confined within a sphere whose radius is less than five times the radius of the Earth, the gravitational attraction at the surface of Sirius B is extremely high (more than two million times that felt on the surface of the Earth). It is so high, actually, that it causes a REDSHIFT in the light emitted from Sirius B.

Sirius B is also intriguing in that it is one of the most massive white dwarfs detected to date. Although it is very close as far as astronomical distances are concerned, observations are made quite difficult due to its association with Sirius A. Newer, more accurate measurements for the basic properties of Sirius B can help us to understand not only the Sirius binary system, but also our universe as a whole.

Bibliography

Bertin P, Lamers H J G L M, Vidal-Madjar A, Ferlet R and Lallement R 1995 HST-GHRS observations of Sirius A III. Detection of a stellar wind from Sirius A *Astron. Astrophys.* **302** 899–906

Holberg J B, Barstow M A, Bruhweiler F C, Cruise A M and Penny A J 1998 Sirius B: a new, more accurate view *Astrophys. J.* **497** 935–42

Whittet D C B 1999 A physical interpretation of the 'red Sirius' anomaly *Mon. Not. R. Astron. Soc.* **310** 355–9

Robert Mohr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sitterly, Charlotte Emma Moore (1898–1990)

Astrophysicist and atomic physicist, worked with HENRY NORRIS RUSSELL at Princeton on binary stars and their masses. She worked at Mount Wilson with Charles E St John and HAROLD BABCOCK analysing the atomic lines in the sunspot spectrum. At the National Bureau of Standards and the Naval Research Laboratory she analysed laboratory data on the solar spectrum and the atomic data by which spectral lines are characterized; towards the end of her long life she was still working extending the tables into the ultraviolet for use with data from space instruments.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Skylab

The NASA Skylab space station was launched from Cape Kennedy, Florida, on 14 May 1973 with the Saturn V rocket. It went into a 93 min low-Earth orbit with an altitude of 433 km and an inclination of 50°. The Apollo telescope mount (ATM) of the Skylab space station carried a suite of instruments of unprecedented spatial resolution that made solar observations leading to significant advances in solar physics. These instruments were originally developed in 1965 for a NASA mission known as the Advanced Orbiting Solar Observatory (AOSO), which was intended to provide high-spatial-resolution solar observations as a successor of the ORBITING SOLAR OBSERVATORY series of solar missions. Cancellation of the AOSO program in late 1965 led to a plan to fly the experiments on the ATM by 1969, near the maximum of solar activity. Changes and delays in the Skylab program resulted in the 1973 launch during the decline in the solar activity cycle.

The Skylab space station (figure 1) consisted of the 15 m long orbiting workshop (OWS), which provided storage space and crew quarters. That was joined to the multiple docking adapter, which housed the ATM control and display panel (figure 2) and provided a docking port for the Command and Service Module, which ferried the astronauts to and from Skylab. The ATM consisted of an outer rack structure attached by a gimbal system to an inner 3 m long canister that could rotate, pitch and yaw relative to the rack. The canister housed the solar instruments. The ATM optical bench for the instruments was a stiff cruciform structure dividing the canister length into four quadrants. The pointing capability of the ATM, critical for the success of the solar observations, was achieved by a pointing control system that sensed Sun center and sent error signals to the torque motors controlling the ATM canister gimbals. The yaw and pitch stability over minutes of time was about 0.5 arcsec, considerably exceeding the nominal 2.5 arcsec specification. Offset pointing was accomplished by the astronaut with a panel joystick.

ATM observations were conducted over three manned missions (SL-2, 3 and 4), each consisting of a crew of three astronauts who lived and worked on the station. Following Skylab launch problems that threatened the mission, the SL-2 crew carried out the first major repair work in space by erecting a parasol that replaced a lost thermal shield and by deploying the closed solar array on the OWS. The SL-2 mission lasted 28 days, from 25 May to 22 June 1973; SL-3 59 days, from 28 July to 25 September 1973; and SL-4 84 days, from 16 November 1973 to 8 February 1974. These missions permitted about 9 months of solar observations. After the SL-4 mission, Skylab was inoperative and re-entered the atmosphere 11 July 1979.

ATM instruments and operations

The ATM suite of solar instruments consisted of six primary and seven auxiliary instruments. The Skylab instrument numbers, names, sponsoring institutions and brief descriptions were as follows.

Figure 1. View of Skylab from the Sun. The OWS is at the top, covered with a parasol and sunshade. The one OWS solar array is at the upper left; the diagonal arrays supplied power to ATM. The ATM doors and experiment access ports are shown in the circular structure at the lower center. From Eddy (1979).

S052: White Light Coronagraph. The HIGH ALTITUDE OBSERVATORY *(HAO):* The instrument provided white-light images of the full CORONA in the altitude range of about 1.5 to 6 solar radii with a spatial resolution of 8 arcsec. Light from the solar disk was blocked by a series of three external disks and an additional adjustable internal disk behind the primary objective lens. The light passed through several lenses and mirrors and a polaroid wheel to the focus on a removable film camera. The polaroid wheel had one clear position and three polaroid positions with 60° differences in orientations. A movable mirror permitted the beam to be sent to a TV camera which provided a real-time image for crew observation and ground transmission. More than 35 000 exposures were obtained on 35 mm film from both manned and unmanned missions.

S054: X-ray Spectrographic Telescope. American Science and Engineering (AS&E): The instrument used a grazing-incidence Wolter system of concentric and confocal mirrors to form a full-disk solar x-ray image at the focus on a removable film camera. A prefilter and six-element filter wheel provided various broad wavebands in the 2 to 60 Å

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

range with 3 arcsec of spatial resolution and a 48 arcmin field of view. A transmission diffraction grating could be moved into the field of view behind the mirrors to provide spectrographic images. Nearly 35 000 exposures were obtained on 70 mm film from both manned and unmanned missions.

S056: X-ray Telescope. MARSHALL SPACE FLIGHT CENTER *(MSFC):* The instrument was very similar to the S054 x-ray spectrographic telescope, but the metallic filters provided a generally shorter wavelength range of about 6 to 30 Å, corresponding to somewhat higher coronal source temperatures than those of the S054 instrument. The light-gathering area of S056 was about one third that of S054, the field of view was 38 arcmin, and there was no diffraction grating in the S056 instrument. More than 27 000 exposures were obtained on 35 mm film from the manned missions only.

S055: Ultraviolet Spectrometer-Spectroheliometer. Harvard College Observatory (HCO): The instrument was a seven-channel photoelectric spectrometer operating in the 280 to 1340 Å range. Solar rays were reflected off a parabolic mirror onto a reflection grating and then into channel electron multipliers. Seven wavelengths could be simultaneously observed with a spatial resolution element of 5×5 arcsec^2 which could be scanned in a raster pattern of 5×5 arcmin^2 in 5.5 min. Other instrument modes permitted spectral scans over a spatial resolution element or scans at seven wavelengths along a line 5 arcmin long and 5 arcsec wide. Photoelectric data were obtained and stored on a tape recorder on both manned and unmanned missions.

S082A: Extreme Ultraviolet Spectroheliograph. US NAVAL RESEARCH LABORATORY *(NRL):* The instrument consisted of a Wadsworth concave grating which formed spectrally dispersed full-disk images on a film strip. The field of view was 57 arcmin, the wavelength range was 171 to 630 Å, and the spatial resolution was about 2 arcsec. A total of 1023 images were obtained on 35 mm film from the manned missions only.

S082B: Ultraviolet Spectrograph. NRL: The instrument consisted of a parabolic mirror reflecting light through a 2×60 arcsec entrance slit to a predispersing grating, which then reflected it to a main grating and on to a film strip. One setting of the predisperser grating covered the wavelength range 970 to 1970 Å with 0.06 Å resolution and the other 1940 to 3940 Å with 0.12 Å resolution. A slit imaging system allowed the astronaut to observe the position of the slit relative to the white light image of the Sun. Over 6400 spectra were obtained on 35 mm film from the manned missions only.

Of the seven auxiliary instruments, one class permitted the astronaut to do precise solar pointings with coordinates which could be displayed on board and telemetered to the ground. Two separate Hα telescopes with zoom capability and 2 arcsec resolution were used, images of which could be displayed on one of the two 6″ diameter CRT monitors on the ATM control and display panel. The Hα #1 telescope, provided by HCO, was equipped with a film camera which provided pointing information via a superposed reticle that indicated the positions of the coaligned S055 and S082B instruments. Over 68 000 photographs were obtained from Hα #1 on 35 mm film from manned and unmanned missions. Hα #2, provided by MSFC, had a wider field of view but no film camera.

The XUV monitor was provided by NRL and attached to S082B to provide CRT images for the astronauts and ground images for planning solar observing programs. The full-disk images were of 15 arcsec resolution in the 170 to 500 Å range. They were valuable for identifying coronal features and SOLAR FLARES.

Among the instruments used to alert the astronauts and solar mission scientists at JOHNSON SPACE CENTER (JSC) in Houston to solar flares was the x-ray event analyser (XREA), a proportional counter x-ray experiment provided by MSFC which measured x-rays in the 3 to 33 Å range. AS&E provided the photomultiplier exposure counter (PMEC) to detect x-rays in the 0 to 8 Å range and an image disector telescope (IDT) to provide a 3 to 10 Å coarse full-disk image for the ATM CRT. Finally, the radio noise burst monitor (RNBM) provided a continuous measurement of the solar 6 cm flux. In practice the PMEC was used as the flare-warning device, with a threshold set each day for astronaut initiation of flare observing modes.

The ATM scientists were forced by the limited time for astronaut ATM operations and by the number and complexity of the instruments to develop plans for simultaneous operations of ATM instruments. Each scientific problem resulted in a joint observing program (JOP), and each JOP was composed of building blocks (BB) describing the operational sequences of each of the instruments (figure 3). The 13 JOPs of SL-2 evolved into 27 JOPs for SL-4.

Extensive time was required before the mission to train the astronauts in solar physics and the requirements and operations of the ATM experiments. In addition, it was necessary for the scientists and engineers to become familiar with flight operations procedures. An ATM mission control room was set up at JSC which was manned by a scientist and an engineer from each of the five principal institutions around the clock in three shifts. An ATM planning session began each day at noon to plan the next day's operations. Forecasts of solar activity from National Oceanic and Atmospheric Administration (NOAA) personnel and results of current and previous ATM observations were important for the planning. Detailed observing plans were sent up to Skylab that evening. During the night (12 a.m. to 08 a.m.) shift Skylab summary planning was done which resulted in determining how much time (or how many orbits) would be allotted for ATM astronaut operations the following day. Both the summary planning and ATM planning sessions frequently involved contentious negotiations for resources and priorities. Far less planning and support was required for the unmanned periods when only S052, S054, S055 and Hα #1 were operational.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 2. SL-4 astronaut Bill Pogue at the ATM control console, showing instrument switches and viewing monitors. From Eddy (1979).

JOP 2B - ACTIVE REGIONS: Long Term Evolution - Disc

Step	BB	Target
(1)	4A	• Roll 82A XUV dispersion clear of activ • Point 55 OFFSET so that 55 MAR covers • Record pointing coordinates RL+/- ____
	B	• Point Hα1 at a Hαbright point. • Roll slit for uniform emission. • Maximize DET 1 or 3.
	C	• Repeat B at a second Hα1 bright point.
	D	• Point Hα1 at a relatively dark Hα area • Roll 82B slit for uniform emission.
	E	• Pointing and roll same as 4A. • Repoint 82B slit for uniform emission.

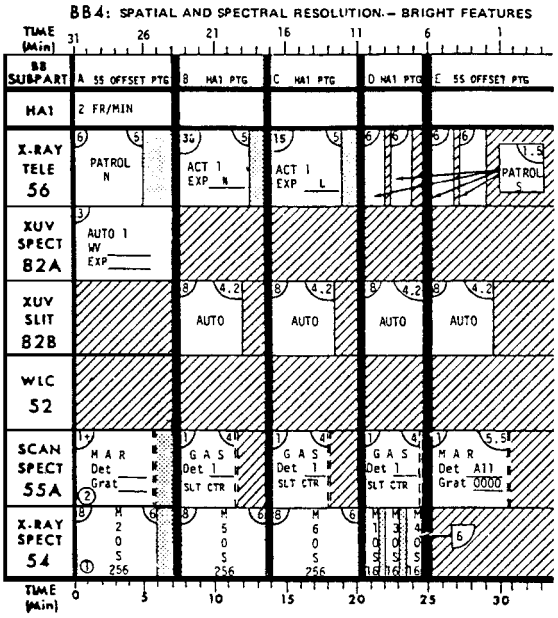


Figure 3. Part of a JOP (top) for active region study and part of one of the BBs (bottom) called for in the JOP. ATM experiments are listed vertically in the BB while time runs to the right, showing various modes and settings for the experiments. From Tousey (1977).

The solar science results

The high-spatial-resolution observations made possible by the ATM pointing system provided significant advances in a number of areas of solar physics. Only after ATM was it clear that coronal structures consisted of magnetic loops. Some phenomena, such as bright points, were completely new. Others, such as coronal holes and coronal mass ejections, had been observed previously, but the clarity and extent of the ATM observations established their basic properties and importance for coronal physics. Earlier ideas, based on observations with coarse spatial resolution, of a relatively homogeneous and slowly varying corona were demolished.

The new results were intensively investigated in a series of three Skylab Solar Workshops in Boulder, Colorado, in the 1970s. The first was devoted to CORONAL HOLES, the second to flares and the last to ACTIVE REGIONS. Each workshop was followed by a published monograph discussing the new results.

Historical perspective on instruments and operations

Solar imaging observations today are obtained from unmanned spacecraft using CHARGE-COUPLED DEVICES (CCDs). Prior to Skylab, solar images from spacecraft were obtained by photoelectrically scanning or rastering the Sun to build up an image. The Skylab mission was a labor intensive transition through a photographic detector stage involving considerable astronaut and ground support involvement. Photographic film, with its instantaneous images over wide fields of view with high spatial resolution, provided a big leap in information content over previous spacecraft observations. Film had, however, several drawbacks. It was a finite consumable which required careful budgeting throughout the mission. It was vulnerable to fogging by heat and radiation. It had a low dynamic range, in contrast to that of the intensities of coronal features being observed. This required that sequences of increasing exposure times be employed to observe both faint and bright features. The S054 x-ray instrument typically used a seven

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or eight frame sequence of exposures ranging by factors of 4 from 1/64 s to 64 or 256 s in duration. The film roll had to be carefully developed on the ground, and calibration was always a challenge. CCDs have proven their clear superiority to film for solar observations.

After the Skylab mission, it was generally concluded that the astronauts had played key roles in the success of the mission. The reasons were twofold: (a) they carried out solar observations, and (b) they performed repairs on instruments. Astronauts removed cloth threads from the front occulting disk of the S052 coronagraph and moved a jammed filter wheel into a filter position in the S054 x-ray telescope. They also pinned open experiment doors that had failed shut and replaced failed or troublesome film transport assemblies.

Astronauts functioned well as observers, but they carried out solar observations for only about 23% of the total manned mission periods. The presence of limited film budgets and the need to set the hard-wired switches for experiments on the ATM control panel also mandated a critical role for the crew that photoelectric detectors and computer control of instruments render unnecessary. A similar critical role was accorded the astronauts in the eight-day Spacelab 2 mission in July 1985, but the limited success of that mission indicated the problems of a multidiscipline, limited-duration solar mission with untried instruments. Perhaps the most significant factor in the demise of the crew-operated solar experiments is the great improvement in communications that allows a high degree of control of space experiments from the ground. That capability has obviated the expensive cost of launching and supporting human solar observers.

Skylab did provide us with some visions of the future. Among the viable innovations associated with the Skylab program, we can count the following:

(a) The development of coordinated observing programs among the experiment groups.
(b) A guest investigator program for scientists not part of the PI institutions.
(c) The use of thematic workshops to advance the science from the mission.
(d) The use of calibration rocket flights (calrocs) as a means of instrumentation calibration.
(e) The use of a supporting coordinated ground-based observing program.

Bibliography

Eddy J A 1979 *A New Sun: the Solar Results from Skylab* NASA SP-402 (Washington, DC: NASA)

Tousey R 1977 Apollo telescope mount of Skylab: an overview *Appl. Opt.* **16** 825 (several ATM instrument descriptions follow this introductory article)

Stephen Kahler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Slipher, Vesto Melvin (1875–1969)

Astronomer, born in Mulberry, IN, became director at the Lowell Observatory. He used spectroscopy to determine the rotation periods of several planets, and identified the constituents of their atmospheres. He recorded the first radial velocities of galaxies. The exposure times for these early photographs were as long as a week. The data were extended by EDWIN HUBBLE to discover the expansion of the universe. With the data, Slipher measured the rotations of the spiral galaxies. He supervised the search for the ninth planet, Pluto.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sloan Digital Sky Survey

The Sloan Digital Sky Survey employs a special purpose 2.5 m telescope located in New Mexico to create a five-color digital map of 10 000 square degrees of sky, to be complete by about 2005. The images will be used to autonomously select one million objects for spectroscopy, via a multi-object spectrograph mounted on the same telescope, yielding redshifts, and thus distances, for each galaxy. A huge number of previously unrecognized quasi-stellar objects and peculiar stars will also be revealed. Although originally motivated by problems concerning large scale structure of the universe, the Survey will constitute a permanent public data bank applicable to a large variety of problems in astronomy, ranging from nearby asteroids to the most distant quasars. The Survey is conducted by an international consortium of ten universities and research laboratories.

For further information see http://www.sdss.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Small Astronomy Satellite (SAS)

Series of three pioneering NASA scientific satellites, launched 1970–75. SAS-1, also known as UHURU (Freedom), was the first dedicated x-ray astronomy satellite. SAS-2 was the first satellite to detect gamma rays, and it discovered the enigmatic Geminga source. SAS-3 also carried x-ray experiments.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Small Magellanic Cloud

The smaller of two nearby companions of the Milky Way Galaxy that can be seen with the naked eye in the southern hemisphere sky and which are named after the Portuguese navigator, Ferdinand Magellan. Located in the constellation of Tucana, at a distance of about 190 000 light-years, the Small Magellanic Cloud (SMC) has an angular diameter of about three degrees, about half the apparent diameter of its larger sibling, the Large Magellanic Cloud. It contains about 2 billion stars and has a mass of about one-fiftieth of the mass of the Milky Way Galaxy (about one-fifth of the mass of the LMC). An irregular galaxy (Hubble class Irr I), it has been stretched into an elongated cylinder, some 60 000 light-years long by about 15 000 light-years wide, with its long axis pointing towards our Galaxy, presumably as a result of gravitational tidal forces exerted on it by the Milky Way Galaxy. Although it contains substantially less gas and dust than the LMC, it nevertheless contains significant numbers of HII regions (luminous nebulae), young clusters and hot young stars. The presence of these objects suggests that it has experienced a recent bout of star formation, possibly triggered by a close encounter with the Milky Way that took place about 200 million years ago.

See also: interacting galaxies, irregular galaxy, Large Magellanic Cloud, Local Group, Milky Way Galaxy, nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Smart-1

First European Space Agency advanced technology demonstration mission proposed under the Horizons 2000 Science plan. Intended to demonstrate innovative and key technologies for scientific deep-space missions such as solar electric propulsion. Launch is planned for late 2001. Possible targets include the Moon, near-Earth asteroids and comets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Smith, Harlan James (1924–91)

Astronomer, born in Wheeling, WV, became director of the McDonald Observatory of the University of Texas at Austin. He discovered (with Dorrit Hoffleit) that quasars were variable, making it possible to set limits to their size and probe the mechanisms of their emission. He identified the dwarf Cepheid variable stars. He studied the radio emissions from Jupiter and the other planets, and the influence of the solar wind on them. He was one of the first astronomers to plan ground-based observations in conjunction with space missions, constructing a 107 in telescope at the McDonald Observatory to support planetary space missions. He was an advocate of space exploration both by orbiting astronomical telescopes and by humans on the Moon, using lunar bases for astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Smith, Sir Francis Graham (1923–)

British radioastronomer, worked on radar during the Second World War, joined Martin Ryle's group at Cambridge and together they discovered the strongest radio source Cassiopeia A in 1948. For his PhD in 1951 he refined the positions of Cas A and Cygnus A. David Dewhirst at Cambridge pointed out that there was nebulosity at the position of Cas A, which proved to be a supernova remnant, and WALTER BAADE and RUDOLPH MINKOWSKI used the 200 in telescope to identify Cyg A as a galaxy. He showed that the fluctuations in brightness of radio sources were due to local effects in the ionosphere and used the Long Michelson Interferometer at Cambridge to measure its electron density. Became professor at Jodrell Bank where he worked on pulsars, discovering the high polarization of their radio emission and worked on their optical identification and period changes. He directed the Royal Greenwich Observatory and built the UK's telescopes on La Palma in the Canary Islands. He was given the honorary title of Astronomer Royal.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Smyth, Charles Piazzi [also Charles Piazzi-Smyth] (1819–1900)

Astronomer, son of an amateur astronomer, Admiral William Henry Smyth, and named after the Italian astronomer GIUSEPPE PIAZZI. Became director of what was then called the Astronomical Institution on Calton Hill in Edinburgh, which evolved into the Royal Observatory, Edinburgh, in a new building on Blackford Hill in the same city. He made an estimate of the amount of heat radiation received from the Moon and studied the zodiacal light. However, he did not distinguish himself in the role of director, being responsible for the neglect of the instruments, and commissioning a telescope that never worked. Moreover, he was obsessed with pyramidology, founded in the mistaken belief that the measurements of the Great Pyramid are of mystical and mathematical significance. He published very successful popular books on this topic but the Royal Society refused to publish his cranky papers and he became the only person ever to have resigned his fellowship, apparently making a ill-judged threat to do so without calculating that his offer would be accepted. After a negative official report on the state of the Calton Hill observatory, he resigned the directorship and devoted the rest of his life to cloud photography in the Lake District. But although this is basically a sorry tale, Piazzi Smyth made one outstanding contribution to the astronomy of Britain and the world. In 1856, he travelled, on his honeymoon with his wife Jessica, to 'Teneriffe' in the Canary Islands and camped on its high volcano, El Teide, investigating Isaac Newton's suggestion that optical observatories would be better sited at altitude. He discovered that he could indeed see fainter stars than with the same telescope in Edinburgh, and resolve closer double stars. Mount Hamilton Observatory was the first mountain observatory set up on the basis of Piazzi Smyth's site-testing expedition, and the Observatorio del Teide and the Observatorio del Roque de los Muchachos on the sister Canary Island of La Palma are direct consequences.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SNC Meteorite

One of a small number of achondritic meteorites believed to have originated on Mars. The letters SNC (pronounced 'snick') stand for the three main classes: *shergottites*, *nakhlites* and *chassignites*. As of April 2000, 14 SNCs had been discovered. The largest at 40 kg is the first known nakhlite, which fell at Nakhla in Egypt on 28 June 1911.

See also: METEORITES FROM MARS

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Snel [Snell or Snellius or Snel van Royen], Willebrord (1580–1626)

Born in Leiden, Netherlands, at first studied law but eventually succeeded his father as professor of mathematics at the University of Leiden. He founded the science of geodesy in 1617, when he published *Eratosthenes Batavus*, which developed the method of triangulation for surveying. In 1621 he discovered what became known as *Snell's sine law* of refraction; it was first published by HUYGENS in his book *Dioptrica*. He published on astronomy including on comets; he remained committed to the Ptolemaic theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Société Française d'Astronomie et d'Astrophysique

The French Society of Astronomy and Astrophysics (SF2A) (previously SFSA for Société Française des Spécialistes d'Astronomie), officially constituted on 18 November, 1978, includes about 500 French astronomers and astrophysicists. Its basic aim is to induce a better cohesion among the professionals in astronomy and to defend the interest of their research activity.

Since November 1980 a journal has been regularly published in French: *Le Journal des Astronomes Français*. It is sent to more than 800 scientists. Also a regular letter is sent by e-mail, in which information of all kinds concerning French astronomers is given.

The society organizes a meeting every year, in a different French town. In 1999, the meeting was organized in Toulouse, together with the general meeting of the EUROPEAN ASTRONOMICAL SOCIETY (JENAM'99). In 2000 the SF2A meeting was held in Marseille.

Every year, during the meeting, a scientific prize is given to a young researcher who contributed in a major way to the advances of research in astronomy. The prize is sponsored by the society Compaq-France. In 1999 it was given to Francis Bernardeau, for his work about the formation of structures in the primordial universe.

The society also organizes a school every year on a different specialized subject, in the town of Goutelas. The proceedings of these Goutelas Schools are regularly published.

For further information see
http://www.iap.fr/sf2a/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Software in Astronomy

'Software' is the term used to refer to the programmed sequences of instructions that are executed by a digital computer. Software of some sort is required by any task that involves a computer. Computers first became generally available in the years following World War II, and since then they have become progressively more powerful, progressively cheaper and progressively more ubiquitous. Computers now influence almost all aspects of astronomy: applying for observing time; the scheduling of observations; the acquiring and reducing of data; the final dissemination of data across networks, particularly the internet; even the use of presentation graphics systems when giving talks at conferences.

Software is needed for all of this. In some cases, general-purpose software, such as Microsoft's PowerPoint for presentations, or LaTeX for typesetting, can be used. In other cases, for example the control of a specific instrument, special-purpose software is needed. This has to be written, either by astronomers or by specialist programmers employed by an astronomical institution or university. The production of specialized software is a labor-intensive undertaking, and this provides a continuous incentive to re-use existing software if at all possible. However, there are a number of factors conspiring against such re-use, and how these are overcome may be one of the most important challenges facing astronomical software.

Acquisition and reduction of astronomical data

Most of the specialized software written especially for astronomy is used either for the actual obtaining of data or its subsequent reduction and analysis.

Data acquisition is a question of the control of telescopes, instruments and detectors. Telescopes need to be pointed at their targets and kept pointing at them despite structural deformation, atmospheric refraction and other effects. Their control software can attempt to calculate these effects in advance and allow for them. It can also make use of feedback systems such as autoguiders and—in recent years—active optical systems. Detectors and instruments need to be configured and the data read and recorded, generally with some feedback to the astronomer as to the quality of the data being obtained. Radio observations may involve the control of hardware correlators. All of these aspects need to be coordinated by some higher-level system that may even allow the order of pre-planned observations to be changed in response to atmospheric and other conditions.

Most data reduction has an initial stage, which is essentially the removal of the instrumental effects from the data. In many cases the ideal result will be a calibrated image—a two-dimensional image each element of which is at a known sky coordinate and whose value is an absolute intensity. Alternatively, it may be a calibrated spectrum or set of spectra—each a set of data values covering a known wavelength or frequency range and whose value is again an absolute intensity measurement. There are other types of data, such as time-resolved data, and some measurements may include POLARIZATION and other parameters, but in all cases the ideal is that all instrumental effects have been removed and the result is as the target object produced it. Subsequent processing will depend on the problem the data was taken to solve.

Optical data are usually in the form of images taken by a two-dimensional detector. These may be direct images of an area of sky or may be the images seen by a spectrograph CAMERA, in which case they will contain one or more spectra. These spectra may be distorted, especially when echelle spectrographs are used. Long-exposure CCD images often contain cosmic ray events which can be identified by comparing one or more exposures of the same target field. Images usually require division by a flat field to take out pixel-by-pixel instrumental variations, and linearity corrections if the detector is non-linear. It may be necessary to subtract detector bias values, determined by exposures with no light on the detector, or sky background. For images, techniques such as the construction of median sky frames may be used. For spectra, extraction techniques attempt to trace the spectra and any corresponding sky data and subtract the two.

Radio data, particularly when produced by synthesis telescopes, need more complex processing even to get to a two-dimensional image of a target. CORRELATORS provide the fringe phases from baselines between pairs of antennae, which, once calibrated, are then samples of the Fourier transform of the required image. As a result Fourier transforms are required to obtain an image. The uneven and incomplete sampling in the Fourier domain means that even a perfect point source gives a messy picture with significant side-lobes, effectively convolved with a 'dirty beam' caused by this imperfect sampling. An iterative algorithm due to Hogböm, called CLEAN, is usually used to tidy this up, estimating the dirty beam and successively subtracting it from each source. Further calibration steps, making use of redundant information in the data, can reduce the residual errors.

Data and file formats

Data files are usually passed, on tape, disk or a network, from the data acquisition system at the telescope to the data reduction system to be used by the astronomer.

Being able to transfer information between computer systems is of fundamental importance, but the differences between machines can complicate this. If a 32-bit number is held in four 8-bit 'bytes', is the first byte the most or least significant? Unfortunately, even today both conventions, termed 'big-endian' and 'little-endian' (after the petty squabblers over boiled eggs in *Gulliver's Travels*) are commonly used. Personal computers based on Intel processors are little-endian, SUN Workstations are big-endian. Floating-point formats—the way non-integer values are represented—have also varied in the past, but there is now an IEEE standard for floating-point data and most machines have adopted this.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

These considerations make it difficult even to write out a single number from one machine in the certainty that another machine can read it. However, for astronomy a much more complicated problem is the question of the higher-level file formats used. A 'file format' refers to the conventions adopted when writing data to a file. An instrument may create a data array of *nx* by *ny* data values, but if all its control software does is write this to a disk file as an image of how it appears in the instrument computer's memory, this will be hard to make use of. At the very least, a program that will read it needs to be able to start by determining the dimensions of the data, so this information needs to be included and some convention is needed for how this is to be done. For practical purposes a great deal of additional information is needed—the name of the object under study, instrumental parameters, time of observation, coordinates of the target object, etc. In the early days of astronomical instrumentation, *ad hoc* formats proliferated, and in many cases the only software that could read data written by a given instrument was the software package that wrote it. The search for a universal file format has elements of the search for the Holy Grail, but it is also an area in which a great deal of progress has been made.

FITS format

It is still the case that most astronomical software packages have their own internal formats. However, in 1981 Wells, Greisen and Harten published a description of FITS (Flexible Image Transport System). FITS defined a set of conventions based on the use of a header prefacing the actual data. This header contained pairs of keywords and values describing the actual data. Additional keywords could be added to provide the ancillary information associated with the image. To get the most out of a FITS file one had to have a reading program that understood the meaning of all the keywords in the header. However, it was possible to write a program that could extract the main data array from a FITS format file even in the presence of ancillary information that the programmer had not anticipated. FITS became and has remained the *de facto* standard for the interchange of astronomical data. Over the years many groups have proposed and implemented and used extensions to the original FITS standard, and some of these have been incorporated into the evolving standard. FITS can now handle multiple images (so can associate error and quality information with a base image, or can handle a set of related images) and can include an arbitrary amount of tabulated data. A standard library of routines for accessing FITS data (CFITSIO) is available from GODDARD SPACE FLIGHT CENTER. FITS provides a good example of the advantages of standardization and the CFITSIO library is a good example of software re-use.

For some purposes it is natural to arrange data items in a hierarchical structure. Hierarchies provide an obvious way of arranging large amounts of related data, and systems exist that provide such an arrangement within a disk file. The Starlink project in the UK uses a hierarchical format called HDS, while the National Center for Supercomputing Applications (NCSA) supports a format called HDF that has a similar philosophy. These formats can handle a greater variety of information than can FITS, but have not attained the almost universal acceptance that FITS has. Most large astronomical software packages can import and export data in FITS format.

Evolution of astronomical software

Until the 1960s, computers were of only limited application in astronomy. In the 1950s early programmable calculating punches and similar devices were used by almanac offices. Although these were programmable, they were programmed using plug boards and the term 'software' was not used in this context. Mainframe computers such as the IBM 7090, introduced in 1959, could run FORTRAN programs and these began to be used for ephemeris calculations. They could also be used to run model atmosphere codes, which involved intense mathematical calculation but which did not require connection to instruments. At that time, observational astronomy made little use of computers and software.

Mini-computers

The emergence of the mini-computer, notably the PDP-8 and PDP-11 series, in the late 1960s and early 1970s meant that it was feasible to provide computerized control systems for the 4 m class optical telescopes then being built. Earlier telescopes, in the main, had relied on the characteristics of the equatorial mount to keep them tracking their targets and on offset guiders to provide a fine correction. Introducing computers made it feasible to model the telescope errors, introduced for example by mechanical flexure, and to correct for them. Computers could also allow for atmospheric refraction, could precess coordinates, could slew the telescope in the most efficient way and could control raster scans. With the advent of computer control systems telescopes began to be judged by how accurately they could set on an object of given coordinates, and how accurately they could track that object without the use of guiders. The Anglo-Australian Telescope (AAT) set the standard for such systems, being able to set with an accuracy of better than 1.5 arcsec when it came on line in 1974. Computerized control system could also handle the more complex control requirements of telescopes with alt-az mounts.

At about the same time, optical detectors were moving into the digital era. Before then, most optical observations had been photographic in nature. While spectra might be scanned by microdensitometers, the output would go to chart recorders rather than being fed to a computer. However, the new electronic detectors such as Wampler's Image Dissector scanner or Boksenberg's Image Photon Counting System, both of which fed data directly into computer memory, could not operate without computers.

On the mini-computers available then, 64 Kbytes represented a substantial amount of memory and hard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

disk drives were expensive options that held only a few megabytes (2.5 Mbytes was typical). These were '16-bit' machines, using only 16 binary bits to represent memory addresses, giving them a natural address limit of 64 Kbytes, although this could be worked around in some cases. Programming was often done using assembler languages, although FORTRAN was available. Minicomputers could be used for instrumentation because they were sufficiently affordable (for the time) that such a machine could be dedicated to an instrument or telescope. The amount of memory available for such machines was a serious restriction, but they did have very capable real-time operating systems. Portability of code was rarely an issue, and programmers generally felt that their only object was to produce code that would do the required job with the available equipment. Proponents of methodologies such as 'structured programming', which emphasized a clear structure to a program above absolute performance, did not always find a receptive audience until machines became rather faster.

Memory limitations were one reason for the popularity of the Forth programming language. A program written in Forth's somewhat unconventional syntax would be converted into instructions for a 'Forth machine' that had an extensible instruction set—new instructions could be defined in terms of sequences of other instructions. Since this extensibility encouraged the re-use of existing instruction sequences, this made for a remarkably efficient use of memory. The first complete, standalone Forth system was developed by Moore for the Kitt Peak 11 m radio telescope in 1971. Forth became very popular and was used widely for telescope and instrument control and data reduction. As machines gained more memory, however, Forth's efficient use of memory became less compelling an attraction, the freedom it provided could be seen as a liability, and its reputation for making it easy to write code that was very hard to understand began to tell against it.

Data reduction for RADIO TELESCOPES became more complex with the advent of the APERTURE SYNTHESIS telescopes such as the VERY LARGE ARRAY (VLA), because of the need for Fourier transform processing to produce images. At this point, the need for radio-astronomical computing increased dramatically.

Software frameworks

In 1977 Digital Equipment Corporation (DEC) introduced a significant machine for astronomy: the VAX 11/780 was an outgrowth of the PDP-11 line, but it used 32 bits for memory addresses and could swap memory blocks to and from disk to give the impression it had more memory available than was in fact the case (a technique called 'virtual memory'). Such a machine can handle data arrays of many megabytes directly, without having to resort to devious programming techniques. With a 1024 by 1024 pixel CCD producing data arrays of 4 Mbytes per readout (if converted to floating-point form for data reduction purposes), this capability was sorely needed. For most of the 1980s the VAX and its VMS operating system became a *de facto* standard for astronomical software.

The ability of the VAX and similar machines to run much more complicated software allowed astronomical software designers to think more about the 'systems' aspects of what they were writing. An ideal had always been to provide a framework that could be used by any programmer as a basis for a new program, rather than requiring each new program to be written from scratch. For example, a data reduction framework could define the file formats to be used, could define a common command language that could invoke a new program—provided that the new program was written to the framework's conventions—and could allow that new program to interact with other programs. The way parameter values could be specified by the user and passed to a program would be laid down, so that all programs run within the framework would provide the same 'look and feel'. By providing subroutine libraries that could be called by programs written for the framework, it would be made easy to write new programs that would provide new facilities while integrating seamlessly with the existing system.

Data reduction frameworks

Some work on framework systems had started earlier, but the advent of the more powerful machines allowed it to accelerate. Data reduction systems were the obvious applications for such frameworks. Almost all astronomers want to process their data in some standard ways (everyone wants to display data, divide by flat fields, do wavelength calibrations or Fourier transforms, etc) and most then want to apply some specialized processing as well. Ideally, they would use the standard tools provided by a framework—the existing application programs—for the former. Then not only would they write their own programs, using the framework, for the latter, they would make these generally available so that the system would grow. Software re-use of this sort is potentially one of the best ways of reducing the cost of astronomical software. The AIPS system, first developed for the reduction of radio-astronomical data in 1978, was a successful early example of this sort of framework system.

AIPS was written in the first instance to support the new VLA radio telescope. It was designed to be easy to port to different computer systems. It was implemented first on a ModComp Classic and on a VAX 11/780. Written in a standard language (the long-established FORTRAN variant called FORTRAN-66), it had different implementations of the same low-level subroutine interfaces used to conceal differences between machines. AIPS emphasized the need to handle coordinates properly and provided general-purpose image manipulation routines together with the radio-specific operations such as CLEAN.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Data acquisition frameworks

Data acquisition systems were less obvious candidates for the framework approach. An observatory would usually have only a limited number of instruments, and these would be programmed in house. However, even these needed a significant amount of infrastructure. An optical observation might require coordination between the telescope, a detector, a spectrograph, a data display subsystem and a data recording subsystem. Early control systems were often large monolithic single programs that handled all aspects of an observation. In principle, a more manageable structure results from splitting such a system into a number of cooperating tasks each handling one aspect, such as the telescope. It also allows standard parts such as the telescope task to be re-used unchanged in a different configuration of tasks designed to control a quite different instrument.

Although such an approach has compelling advantages in terms of modularity and structural elegance, it requires an underlying framework that can coordinate a number of concurrent real-time tasks (tasks responding within rigid time constraints to external events generated by the instruments they control). This is generally a more complex framework than is needed by a system designed purely for data reduction. An example was the ADAM system designed originally at the Royal Greenwich Observatory in the UK and ported to the VAX at the Royal Observatory Edinburgh. It can be argued that what is required for data reduction is a subset of that required for data acquisition, and a data acquisition framework should be ideally suited for data reduction. The UK Starlink project took this line, and uses ADAM as the basis for its data reduction software.

Workstations, microprocessors and UNIX

The VAX was a multi-user machine, but it was possible to use it for dedicated instrumentation control; its VMS operating system provided, at least nominally, the facilities needed by real-time tasks, such as the ability to insist that physical rather than virtual memory be used for critical program sections, and a task scheduler that should prevent high-priority tasks being blocked at critical times by the more mundane requirements of lower priority tasks. The 4.2 m William Herschel telescope (WHT) in La Palma was controlled by a dedicated VAX system, as was the 2.3 m at Siding Spring in Australia.

At the same time, the microprocessor had started to appear. Microprocessors allowed engineers to design hardware subsystems around dedicated programmable computers. These were often run with 'home-built' operating systems written for the immediate purpose—sometimes without operating systems as such at all. These would usually be programmed in assembler, sometimes in a higher-level language, C. A small microprocessor-based system could be used to provide direct real-time control of a single instrument or part of an instrument. Data acquisition systems could now be built with a machine such as a VAX handling the overall control and processing the resulting data, but using a bevy of microprocessors to provide the tight real-time control needed by the actual instruments.

It is unusual for one computer system to remain dominant for any length of time, and the challenge to the VAX came from the emerging UNIX workstations. The UNIX operating system, as it appeared at first in 1967 on a PDP-7, made little pretence at being a real-time system. Its great advantage was its portability; it was an operating system written not in assembler (as most were, for efficiency) but in C. Moreover, it was made freely available and so became very popular, particularly in university computing departments. This made it an ideal match for the workstations that were being built around the more powerful of the new microprocessors, such as SUN's SPARC and the MIPS processors. These reduced instruction set computers (RISCs) provided a limited set of basic instructions, but were able to run them very quickly. Computers based around RISC processors, almost all of which now run UNIX, are the dominant systems in astronomy at the turn of the century. Unfortunately, different implementations of UNIX still manage to differ, sometimes in quite significant ways.

Recent developments

There are now some common elements to astronomical computing systems. In recent years the trend has been for workstations, almost all running UNIX, to become faster, and almost all are now networked. Instrument control systems generally involve workstations for overall control and on-line data reduction, with dedicated real-time systems based on microprocessors in direct control of the instruments and the telescopes themselves. These microprocessors, often as powerful as the workstations, usually run some real-time kernel system; examples are VxWorks and pSOS. These generally try to provide a programming environment as close to UNIX as possible, but with additional features to support their real-time nature—and with additional limitations imposed by that real-time nature.

Data reduction

A number of data reduction frameworks exist. AIPS continues as the mainstay of radio astronomical data reduction, but is in the process of being superseded by AIPS++. AIPS++ is written using the object-oriented language C++, a descendant of C that supports and encourages an approach to programming where one creates software 'objects' that are capable of performing a number of related functions. This can often be a more natural way to program, possibly because it has analogies to the way tasks in the real world are accomplished by people or groups each specializing in certain tasks, which they perform as requested. For optical data reduction, the dominant framework is IRAF, the Image Reduction and Analysis Facility developed and maintained at NOAO since the mid-1980s (see SOFTWARE: IRAF). IRAF provides a programmable command language and a large and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

increasing number of applications. Originally, IRAF applications had to be written in its own programming language, SPP, provided in the interests of portability; more recently interfaces to more standard languages have been provided. The UK Starlink project continues to support the use of ADAM for data reduction, and ESO provides MIDAS. However, no such system has gained the level of world-wide acceptance that IRAF has achieved.

There are numerous other data reduction packages in use. A number of programs are built around the commercial IDL system, which provides interactive facilities for manipulating and displaying data arrays. Interactive languages such as Perl and Tcl/Tk are used to produce very flexible programs with easily programmed user interfaces, and Tcl/Tk in particular is used by a number of the larger frameworks. The Glish interactive language, used by AIPS++, is also finding uses in instrument and telescope control.

Data acquisition

The data acquisition software environments used at major observatories are now almost all based on a framework of some sort. This is partly because of the increasing trend to have instruments for telescopes built not by the observatory but by outside institutions, often university departments. In these circumstances, the observatory can get a homogeneous set of instrument control packages only by imposing the use of such a framework—which also makes it easier to write the instrument control software, once the often steep learning curve for the framework has been overcome. Moreover, this allows for much better integration of instrumental software with the overall observatory system, which now tends to take control of aspects such as archiving of data and even observation scheduling. Stand-alone instruments, with isolated one-off control systems, are no longer an acceptable or efficient way to use a large telescope.

There is, however, no dominant instrumentation framework as yet. A number of existing systems are based around a 'database' approach to instrument control, a paradigm where instrumental settings are viewed as database entries. Such entries can be monitored easily to provide user interfaces, especially if a change in a database entry triggers the corresponding actual change in the hardware. This style of operation originated in systems such as accelerators where there are a large number of small items to control. It has more difficulty handing hierarchical systems where there is a natural flow of control with a number of stages overlapping in time. A number of observatories, such as Gemini, have adopted the EPICS database system. A similar system, the commercial RTAP system from Hewlett-Packard, is used by the ESO VLT project. Systems differ in the extent to which this database paradigm (as opposed to more conventional command-based systems where relatively large-scale tasks respond to commands sent by other tasks) pervades the whole system. The VLT has a more conventional command-based layer than does the Gemini system, where commands are implemented in terms of the underlying EPICS database. Other systems, such as the AAO's DRAMA, a descendant of ADAM used on a number of UK-affiliated telescopes, are entirely command based. The Keck telescopes use a system called KTL (Keck Task Library) which is a command-based system that provides interfaces to both EPICS-based and non-EPICS instruments. IRAF is beginning to provide some of the communications facilities needed by data acquisition environments, and these are being used by some instrumentation projects.

Variety and standards

The variety of astronomical software can be a problem for the astronomer looking for nothing more than a way to reduce her data. However, it can be seen as a necessary consequence of a continual search for improved ways of doing things. When there is no perfect solution, and the best solution on yesterday's hardware fails to take full advantage of what is available today, new ideas have to be pursued. Only if these are pursued simultaneously by a number of separate developers will the best solutions emerge. The real problem is to provide ways for such separate developers to take advantage of what others are doing where possible and to avoid unnecessary duplication of effort. The number of *de facto* standards that have emerged is encouraging. For example, although there are a large number of different image display packages available, both Gemini and ESO are using the ESO real-time display package (RTD). Also, almost every astronomical package can read and write FITS format files. The production of standard libraries of routines—a long-lasting example might be the PGPLOT library of routines used for plotting data—can assist here, and this should extend to the re-use of C++ objects, even eventually the production of re-usable software components, such as Active-X controls or Java beans.

Personal computers

Personal computers (both Intel-based systems running a version of Microsoft Windows and to a lesser extent Apple Macintoshes) are used extensively by astronomers, since astronomers have to write papers and give presentations. There are a number of packages aimed at amateurs that do, for example, CCD image processing on PCs, and there are also some excellent 'planetarium' programs available. However, there is little professional astronomical data reduction software available for Windows or Macintosh. There is some use of component software, such as that provided by Active X controls, for small-scale instrument control, but this can be hard to integrate into an overall observatory system. However, the emergence of Linux (a UNIX-like operating system that is freely available and which runs on PCs of both kinds) has made it possible for astronomers to run astronomical UNIX-based packages on their home computers. This has a number of implications for the way astronomers will work in future—many find they have a more powerful machine available at home than they do at work.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Archives and the internet

Software is becoming increasingly important both before observations are made and well after the data are reduced and published. For the VLT, for example, observations need to be planned well in advance using software that generates observation 'blocks' that will be executed much later at the telescope by the control system. The information introduced at that planning phase is associated with the data as they become part of the data archive. That data archive can be thought of as an instrument in its own right, and its store of data can be 'mined' years later, either reducing the need for observations that replicate earlier ones or to provide a long-term study of particular objects (see e.g. SOFTWARE: RADIO PULSAR SOFTWARE).

Making proper use of the huge data archives, particularly those from survey telescopes, is a new challenge for astronomical software (see also SOFTWARE: DIGITAL SKY SURVEYS). A large number of internet web sites now provide access to archival data, survey data, preprints of articles and even journals that are completely electronic in nature. NASA is a good example of an organization that disseminates information widely through the internet, both to professional astronomers and to members of the public. This becomes circular: observation preparation software can obtain from the internet a digitized picture of the sky around a target object and observers can use this to choose the guide stars needed for a new observation, which in turn will eventually be archived and made available.

Bibliography

Few facts are so transient as those relating to computers and software, and there are few conventional publications on the subject. All the various software packages are documented and in most cases both the documentation and the code can be found on the internet. Most of the major observatories have web pages with links to the packages they use and develop. Each year there is a conference on Astronomical Data Analysis Software and Systems (ADASS) whose proceedings are published in the *Astronomical Society of the Pacific Conference Series.*

Keith Shortridge

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Software: Digital Sky Surveys

Large digital sky surveys, over a broad range of wavelengths, both from the ground and from space observatories, are becoming a major source of astronomical data. Some examples include the SLOAN DIGITAL SKY SURVEY (SDSS) and the Digital Palomar Observatory Sky Survey (DPOSS) in the visible, the Two-Micron All-Sky Survey (2MASS) in the near-infrared, the NRAO VLA Sky Survey (NVSS) and the Faint Images of the Radio Sky at Twenty Centimeters (FIRST) in the radio. Many other surveys are planned or expected (see also ASTROMETRIC SURVEYS). While most surveys are exclusively imaging, large-scale spectroscopic surveys also exist. In addition, a number of experiments with specific scientific goals, e.g. microlensing surveys for MACHOs, searches for near-Earth asteroids, are generating comparable volumes of data. Typical sizes of resulting data sets (as of the late 1990s) are in the range of tens of terabytes of digital information, with detections of many millions or even billions of sources, and several tens of parameters measured for each detected source.

This vast amount of new information presents both a great scientific opportunity and a great technological challenge: how to process, and calibrate the raw data; how to store, combine and access them using modern computing hardware and networks; and how to visualize, explore and analyse these great data sets quickly and efficiently. This is a rapidly developing field, which often entails collaborative efforts between astronomers and computer scientists.

Broadly speaking, the steps along the way include: obtaining the data (telescope/observatory and instrument control software), processing the data into catalogs of detected sources and their observed properties, calibrating the data, archiving the results and providing the tools for their scientific analysis and exploration.

Survey pipeline software

The actual gathering and processing of the raw survey data encompasses many steps, which can often be performed using a software pipeline. The first step involves hardware-specific data acquisition software, used to operate the telescopes and the instruments themselves. In principle this is not very different from the general astronomical software used for such purposes (see SOFTWARE IN ASTRONOMY), except that the sky surveying modes tend to require a larger data throughput, a very stable and reliable operation over long stretches of time, and greater data flows than is the case for most astronomical observing. In most cases, additional flux calibration data are taken, possibly with separate instruments or at different times. Due to the long timescale required to complete a survey (often many years), a great deal of care must be exercised to monitor the overall performance of the survey in order to ensure uniform data quality.

The next step is the removal of instrumental effects, e.g. flat-fielding of CCD images and subtraction of dark current for infrared detectors. Other than the sheer volume of the data, this process is extremely similar to the usual astronomical observational data reduction techniques.

At this point one does some kind of automated source detection on individual survey images, be it CCD frames, drift scans, photographic plate scans, their subsections, or whatever other 'raw' image format is being generated by the survey. This process requires a good understanding of the noise properties, which determines some kind of a detection significance threshold: one wants to go as deep as possible, but not count the noise peaks. In other words, maximizing the completeness (the fraction of real sources detected) while minimizing the contamination (the fraction of noise peaks mistaken for real sources). Typically one aims for completeness levels of at least 90%, and contamination of less than 10% in the first pass, and typically the source catalogs are purified further at subsequent processing steps.

Most source detection algorithms require a certain minimum number of adjacent or connected pixels above some signal-to-noise threshold for detection. The optimal choice of such a threshold depends on the power spectrum of the noise. In many cases, the detection process involves some type of smoothing or optimal filtering, e.g. with a Gaussian whose width approximates that of an unresolved point source. Unfortunately, this also builds in a preferred scale for source detection, usually optimized for the unresolved sources (e.g. stars) or the barely-resolved ones (e.g. faint galaxies), which are the majority (see DETECTION OF FAINT OBJECTS). This is a practical solution, but with the obvious selection biases, with the detection of sources depending not only on their flux but also on their shape or contrast: there is almost always a limiting surface brightness (averaged over some specific angular scale) in addition to the limiting flux. The subject of possible missing large populations of low surface brightness galaxies has been debated extensively in the literature. The truth is that a surface brightness bias is always present at some level, whether it is actually important or not. Novel approaches to source, or more accurately structure, detection involve so-called multiscale techniques.

Once individual sources are detected, a number of structural parameters are measured for them, including fluxes in a range of apertures, various diameters, radial moments of the light distribution, etc, from which a suitably defined, intensity-weighted centroid is computed. In most cases, the sky background intensity level is determined locally, e.g. in a large aperture surrounding each source; crowding and contamination by other nearby sources can present problems and create detection and measurement biases. Another difficult problem is unblending or splitting of adjacent sources, typically defined as a number of distinct, adjacent intensity peaks connected above the detection surface brightness threshold. A proper approach keeps track of the hierarchy of split objects, usually called the parent object (the blended composite), the daughter objects (the first level

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

splits), etc. Dividing the total flux between them and assigning other structural parameters to them are non-trivial issues, and depend on the nature of the data and the intended scientific applications.

Object detection and parameter measurement modules in survey processing systems often use (or are based on) some standard astronomical program intended for such applications, e.g. FOCAS, SExtractor or DAOPHOT, to mention just a few of the programs often used in 1990s. Such programs are well documented in the literature. Even if customized software is developed for these tasks, the technical issues are very similar. It is generally true that all such systems are built with certain assumptions about the properties of sources to be detected and measured, and optimized for a particular purpose, e.g. detection of faint galaxies or accurate STELLAR PHOTOMETRY. Such data may serve most users well, but there is always a possibility that customized reprocessing for a given scientific purpose may be needed.

At this point (or further down the line) astrometric and flux calibrations are applied to the data, using the measured source positions and instrumental fluxes. Most surveys are designed so that improved calibrations can be reapplied at any stage. In some cases, it is better to apply such calibration after the object classification (see below), as the transformations may be different for the unresolved and the resolved sources. Once the astrometric solutions are applied, catalogs from adjacent or overlapping survey images can be stitched together.

Object classification, e.g. as stars or galaxies in the visible and near-IR surveys, but more generally as resolved and unresolved sources, is one of the key issues. Classification of objects is an important aspect of characterizing the astrophysical content of a given sky survey, and for many scientific applications one wants either stars (i.e. unresolved objects) or galaxies; consider for example studies of Galactic structure and studies of the large-scale structure in the universe (see UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION). More detailed morphological classification, e.g. Hubble types of detected galaxies, may be also performed if the data contain sufficient discriminating information to enable it. Given the large volumes of data involved in digital sky surveys, object classification must be automated, and in order to make it really useful, it has to be as reliable and objective as possible, and homogeneous over the entire survey.

In most cases, object classification is based on some quantitative measurements of the image morphology for the detected sources. For example, star–galaxy separation in optical and near-IR surveys uses the fact that all stars (and also QUASARS) would be unresolved point sources, and that the observed shape of the light distribution would be given by the point-spread function, whereas galaxies would be more extended. This may be quantified through various measures of the object radial shape, e.g. moments of the light distribution in various combinations. The problem of star–galaxy separation thus becomes a problem of defining a boundary in some parameter space of observed object properties, which would divide the two classes. In the simplest approaches such a dividing line or surface is set empirically, but more sophisticated techniques use artificial intelligence methods, such as artificial neural nets or decision trees (artificial induction software). They require a training data set of objects for which the classification is known accurately from some independent observations. Because of this additional information input, such techniques can outperform the methods where survey data alone are used to decide on the correct object classifications.

There are several practical problems in this task. First, fainter galaxies are smaller in angular extent, thus approaching stars in their appearance. At the fainter flux levels the measurements are noisier, and thus the two types of object become indistinguishable. This sets a classification limit to most optical and near-IR surveys, which is typically at a flux level a few times higher than the detection limit. Second, the shape of the point-spread function may vary over the span of the survey, e.g. due to the inevitable SEEING variations. This may be partly overcome by defining the point-spread function locally, and normalizing the structural parameters of objects so that the unresolved sources are the same over the entire survey. In other words, one must define the unresolved source template which would be true locally, but may (and usually does) vary globally. Furthermore, this has to be done automatically and reliably over the entire survey data domain, which may be very heterogeneous in depth and intrinsic resolution. Additional problems include object blending, saturation of signal at bright flux levels, detector nonlinearities, etc.

The net result is that the automated object classification process is always stochastic in nature. Accuracies better than 90% are usually required, but accuracies higher than about 95% are generally hard to achieve, especially at faint flux levels.

In other situations, e.g. where the angular resolution of the data is poor, or where non-thermal processes are dominant generators of the observed flux, morphology of the objects may have little meaning, and other approaches are necessary. Flux ratios in different bandpasses, i.e. the spectrum shape, may be useful in separating different physical classes of objects.

Survey archive software

Once all of the data have been extracted from the image pixels by the survey pipeline software, it must be archived in order to facilitate scientific exploration. Basically, this data archiving can take one of two forms: raw data access through the Web or some suitable storage and distribution medium (e.g. CD-ROMs or data tapes), or utilization of a database system designed specifically to handle a given digital sky survey archive. The former storage technique, while considerably less expensive, is certainly not optimal for any real analysis, especially for very large volumes of data. On the other hand, using a database system provides

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

significant advantages (e.g. powerful query expressions) limited primarily by how much a survey can afford. Currently, most surveys are accessible from the Internet, whether it is a simple ftp location or a full featured, Web-accessible, query engine.

While certainly flexible and generally easy to establish, Web accessible data are subject to limitations on bandwidth. Despite technological advances which promise to ease such concerns, the rapid growth in available data will continue to swamp available resources. To illustrate this bandwidth problem, for a typical building network (e.g. shared Ethernet at 10 Mbit s^{-1}), it would take nearly a week and half to sift through a terabyte of data (which is the current benchmark for large sky surveys), and this assumes a dedicated bandwidth. Even at fast SCSI speeds (e.g. 100 Mbit s^{-1}), it would take approximately one day to merely move the data of interest.

Each archival center is presented with the challenging problem of storing and serving vast amounts of complex data. Currently the majority of the software written for these applications is in either C++ or Java, while the actual data are transferred via ASCII, FITS or XML. Recently, many of the major data centers have begun to work on sharing knowledge and expertise in order to simplify the development process, as well as improve the overall efficacy of astronomical archives. This work is leading to standards which dictate how archives can communicate with each other, how archives can describe themselves (i.e. their metadata, or data that describes the data), how archives can transfer large amounts of dynamic information, and how sources in different archives can be cross-identified.

In general, the data processing flow is from the pixel (image) domain to the catalog domain (detected sources with measured parameters). This usually results in a reduction of the data volume by about an order of magnitude (this factor varies considerably, depending on the survey or the data set), since most pixels do not contain statistically significant signal from resolved sources. However, the ability to store large amounts of digital image information on-line opens up interesting new possibilities, whereby one may want to go back to the pixels and remeasure fluxes or other parameters, on the basis of the catalog information. For example, if a source was detected (i.e. cataloged) in one bandpass, but not in another, it is worth checking if a marginal detection is present even if it did not make it past the statistical significance cut the first time; even the absence of flux is sometimes useful information!

The two most common types of database management systems used within astronomy are relational-based (where data are manipulated as tables), and object-based (where data are manipulated individually as objects). Each of the two methods have working sites which demonstrate the technology in action (e.g. the 2MASS project uses Informix, while the GSCII project uses Objectivity). In general, relational systems offer more features including third-party add-ons, and powerful query mechanisms due to their dominant position in the business world. On the other hand, object-based systems have shown higher performance and better potential for scaling to extremely large data sets (e.g. CERN is developing an object-based persistence solution to multi-petabyte archives).

Regardless of how the data are actually archived, with the advent of the Internet, essentially all archives are Web accessible. This trend away from tabular or media based archiving produces data sets which are distributed in nature, allowing users and tools equal access to vast amounts of data, which in the past were nearly impossible to efficiently query. As a result, astronomical data are now able to be utilized in a more democratic fashion resulting in uses which the survey institutions did not even imagine.

Scientific analysis and exploration

The end goal of digital sky surveys is to enable scientific investigations. In some cases, the goals are so specific that the best approach is to use analysis software designed and optimized to tackle them, e.g. as in the studies of the cosmic microwave background, in gravitational microlensing experiments, etc. However, most of the modern survey data sets are so information-rich, that a wide variety of different scientific studies can be done with the same data. This entails some general tools for the exploration, visualization and analysis of large survey data sets.

The tools which will be utilized with the new data sets will need to employ cutting-edge techniques from computer science. For example, clustering techniques to detect rare, anomalous or somehow unusual objects, e.g. as outliers in the parameter space, to be selected for further investigation (i.e. follow-up spectroscopy). Other examples include genetic algorithms to improve current detection and supervised classification methods, new data visualization and presentation techniques, which can convey most of the multidimensional information in a way more easily grasped by a human user, and the use of semi-autonomous AI or software agents to explore the large data parameter spaces and report on the occurrences of unusual instances or classes of objects.

The nature of sky surveys and the resulting data sets dictates the kinds of science one may wish to do with them. For example, most surveys would cover a large solid angle, but not go as deep as the typical 'pointed observations'. This, in turn, drives the kinds of data analysis software needed for their scientific exploration. Broadly speaking, the kinds of astronomical investigations for which surveys are well suited include the following.

Multiwavelength astronomy

Combining surveys done at different wavelengths, for a more panchromatic view of the universe. Typically, optical IDs are needed for most of the follow-up work. An obvious example is the optical identification of radio or x-ray sources. A typical basic software requirement is efficient and accurate matching of sources detected in different wavelengths. If positional accuracies and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

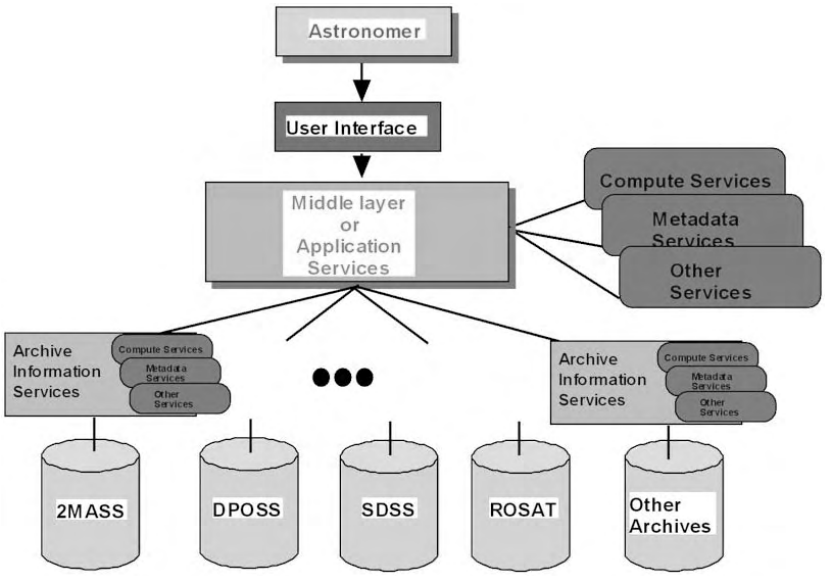


Figure 1. A sample prototype blueprint for the system architecture of a virtual observatory, connecting the scientific users with a multitude of separate digital sky survey archives (schematically represented at the bottom of the picture). The basic task of such a system is to combine large volumes of data from different archives, e.g. through the positional matching of detected sources. Once this is achieved, various other computing services provided by such a virtual observatory may include visualization of data, data-mining, cluster analyis, etc. This system model is predicated on the universal adoption of standards dictating everything from how archives communicate with each other to how data are transferred between archives, services and users.

source densities in matched surveys are comparable this is a relatively straightforward task; however, if either of these two conditions is not satisfied, more sophisticated algorithms may be needed to assign matching probabilities to multiple candidates. These techniques can generally be optimized for certain classes of sources by utilizing as much *a priori* knowledge as possible.

Statistical astronomy

Studies of the large-scale structure (if one detects galaxies in copious numbers), or studies of the Galactic structure (if one detects stars in copious numbers). The sheer numbers of detected sources make the Poissonian fluctuations unimportant, but systematic errors may limit the results. Accurate and uniform flux calibrations and source classifications are essential for such applications, and possible biases should be modeled. In addition, the sheer size of most surveys allows for further subdivision of various analysis (e.g. by morphological or spectral type) for more specific results.

Rare object searches

Searches for rare types of objects or groups of objects, whose frequency of occurrence is so low that huge input source catalogs are necessary in order to find them. This 'needle in a haystack' search is facilitated by the vast number of sources in current surveys (i.e. a one in a million object would appear a thousand times in a billion source survey). Typically such rare objects are found as outliers in some parameter space, e.g. colors: the bulk of 'ordinary' sources (normal stars, galaxies, etc) would form well-defined clusters in the parameter space of observables, and some rare or peculiar types of objects may be found away from them. Unsupervised classification and cluster analysis techniques may be especially useful for this task. This data-mining also facilitates significantly higher efficiencies in performing follow-up spectroscopy of unusual objects. We note that rare objects may belong to known classes (e.g. distant quasars (see QUASISTELLAR OBJECTS: SURVEYS), brown dwarfs, etc), which can be known to occupy some *a priori* determined portion of the given parameter space. Alternatively, there may be previously unknown or unexpected types of objects discovered in this manner.

Future directions

Astronomy finds itself at an interesting time, as several, large-area digital sky surveys are currently under way. In the future, the number of such surveys, as well as the wavelength coverage, will only continue to increase, providing temporal as well as spatial coverage of the sky. This flood of data necessitates a new approach to data handling—a virtual observatory (see figure 1). Individual surveys are important in their own right, but the federation of multiple, cross-wavelength digital surveys provides a tremendous opportunity to truly quantify the origins of stars, galaxies and the universe itself.

This future situation provides the opportunity to adopt a new research paradigm for studying the heavens, as the ability to perform cutting edge research will not be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

restricted to those fortunate enough to have access to the best facilities. Instead, anyone who has the diligence and ability to sift through the avalanche of data can perform novel science.

Bibliography

The field of large digital sky surveys, the related software methodologies, and the new astronomy they enable are developing very rapidly as of this writing (late 1999). There are thus no standard texts or reviews, and whatever articles do exist tend to become obsolete very quickly. Probably the best way to find more up to date information is through the World Wide Web (and the interested reader will know how). While many Web sites can be somewhat ephemeral, we suggest one which will provide a number of useful links in this field, and which we hope will be active throughout the useful shelf life of this encyclopedia: `http://www.digital-sky.org`

Another possibility is conference proceedings on this and related subjects, e.g. the series of volumes on 'Astronomical Data and Software Systems', published by the Astronomical Society of the Pacific in their conference series. A number of useful papers can also be found in

McLean B J *et al* (ed) 1998 *New Horizons From Multi-Wavelength Sky Surveys (IAU Symp. 179)* (Dordrecht: Kluwer)

An example of a particular software system for processing of digital sky surveys is described by

Weir N *et al* 1995 *Publ. Astron. Soc. Pacific* **107** 1243

S George Djorgovski and Robert J Brunner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Software: IRAF

IRAF, the Image Reduction and Analysis Facility, is a large software system used primarily for the reduction and analysis of optical and infrared data from ground- and space-based observatories. The principal IRAF software distributions provide an integrated collection of standard data reduction packages for all instruments operated by the National Optical Astronomy Observatories (NOAO) and the Hubble Space Telescope (HST). Many other observatories use IRAF for processing data from their own telescopes, often providing customized 'layered packages' tailored to reduce data from their instruments. IRAF is also used extensively at the telescope for data acquisition and quick look and by many individuals around the world for general astronomical data reduction and analysis.

Historical perspective

In the years leading up to 1980, data processing at KITT PEAK NATIONAL OBSERVATORY (KPNO) was performed on a CDC Cyber-series computer, a classical mainframe system with computer operators and batch submission of programs using decks of punched cards. To this was connected the pioneering IPPS interactive picture processing system. This consisted of some custom in-house-developed hardware and software used to connect a Tektronix storage scope terminal and a Comtal image display to the Cyber mainframe. Revolutionary for its time, this combination permitted interactive use of the Cyber mainframe for program development and image processing. The Tektronix storage scope terminals and Comtal image display provided interactive vector graphics and image display capabilities. A menu-based system provided the main user interface.

Applications for the old IPPS system were written in Cyber Forth, a locally developed variant of the Forth programming language, and CDC Fortran, used for numerically intensive functions. The IPPS system operated for approximately 10 yr, and was one of the most advanced image processing systems in the world at the time.

In the 1970s the DEC PDP-11 minicomputer appeared on the scene. This was followed shortly by the legendary VAX minicomputer. The VAX was quickly adopted for data processing by astronomy worldwide. These new minicomputers were interactive systems, putting the end-user in control, and would soon doom the batch-oriented mainframe to oblivion. The dominant operating system for the VAX at the time was VAX VMS, which commanded a very strong following from most of the scientific users at the time. A handful of software developers, however, including a small group at Kitt Peak, were excited about a new operating system called Unix, and the new language it was written in, the C programming language.

The transition from the mainframe to the minicomputer sealed the fate of the old IPPS system. The combination of custom hardware and custom software used to build the IPPS ensured that the IPPS was a one-of-a-kind system, the only IPPS installation in the world, and that virtually all of the effort which went into building this system would be lost when the technology it was based on became obsolete. With the coming of the minicomputer, it was time to build a new image processing system to replace the IPPS.

Around 1980 Kitt Peak undertook a study of the existing astronomical data analysis systems of the time (AIPS, Starlink and others) and eventually concluded that a new system was needed to serve the needs of KPNO. The developers were determined that this time around the system would be put together in such a way that the many years of effort spent developing applications software would survive the technology it was based on. To ensure this the system would be both portable and device independent, with controlled interfaces to all external services. Everything learned with the IPPS about interactive graphics and image processing would be preserved in the new system.

The little-known (at the time) Unix operating system would be the primary development platform for the new image processing facility, but the software would be portable and would be supported on VMS and other operating systems from the beginning (in addition to VMS there were subsequent ports to other now extinct operating systems, e.g. from Data General and Perkin Elmer). Two new VAX minicomputers, one running Berkeley Unix and the other VMS, were purchased for developing the new image processing system. Graphics and image display capabilities were provided by DEC VT100/VT640 graphics terminals and an IIS image display.

Development of the IRAF system began at KPNO in Tucson in 1981. By 1984 the basic system framework, command language and a few of the core applications packages were complete. The SPACE TELESCOPE SCIENCE INSTITUTE, after its own internal review of existing systems, adopted IRAF as their data processing framework in 1984 and completed the initial VMS port within the following year. This began a major collaboration between ground- and space-based astronomy which continues to this day, providing users of both communities with a consistent, integrated data processing capability which can be used for both types of data. By the mid-1990s about half of all IRAF software had been contributed by the space science community.

In the late 1980s workstations appeared on the scene, led by Sun, which was just another small workstation vendor in the early days. Following a trip to Silicon Valley to visit Sun, MIPS and other early workstation vendors, Kitt Peak chose to go with Sun. This eventually proved to be another in a series of fortuitous choices for IRAF and NOAO.

The workstation was a major development because it provided a high-resolution bit-mapped screen which could be used to run a windowing system. This was a great boon to scientific data processing because one could use this to implement custom graphics and image processing visualization tools in software (see also SOFTWARE IN

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ASTRONOMY), and every workstation could have its own integrated image display capability. IRAF developed one of the earliest software image displays during this period, the imtool display server for the now obsolete Sunview window system (which pre-dated X). A companion program gterm provided interactive vector graphics. With the subsequent wide adoption of the workstation and the development of the X window system, this was followed by the development of SAOimage by SAO, the X11IRAF tools from IRAF and many other data visualization tools from various sources in the years to follow.

The appearance of the workstation on the world astronomy scene had another big impact: these new systems ran the Unix operating system almost exclusively. Although DEC would later develop VMS-based workstations such as the VAXstation, this meant the beginning of the end for the VMS operating system as a major platform for astronomy software. Some years later the PC would evolve to the point where it was the equal of the workstation. With the PC came the open source Unix-based operating systems such as FreeBSD and Linux, further sealing the fate of proprietary systems like VMS.

By the late 1990s the IRAF system, now approaching 20 yr in age, was a mature system in wide use throughout the world. IRAF had survived a very turbulent period in the evolution of computers and computer software spanning at least three major eras from minicomputer to workstation to PC. Approximately 200 man-years of software development had been invested in IRAF and its applications. By the end of the decade attention was turning to the evolution of the IRAF system itself. A new system framework was needed to take full advantage of new software technology and to permit better use of the wealth of non-astronomy software now available. Research and development leading to this new framework got underway in the late 1990s, aided by several grants from NASA.

The challenge facing IRAF in 2000 as we start into the new decade is to upgrade the foundations of the system even while it is still in use, to bring the system up to date with the latest technology, while preserving the many years of work invested in developing and refining the current wide range of science applications.

System structure

IRAF is not a program, but rather a large integrated software system. Included are a system framework upon which all the other software is based, an integrated programming environment, standard user interface software such as a command language and graphics and visualization tools, and a great variety of applications software. The applications software is organized into packages of logically related tasks.

An IRAF task is an individual program, designed to perform some specific data processing function. A parameter set controls the operation of each task. Tasks may be either simple batch-oriented tasks or interactive tasks which use keystroke, text, vector graphics or image display capabilities to interact with the user during execution. IRAF tasks are normally run interactively from within the IRAF environment, but most tasks may also be called at the host level, e.g. for integration into an external data processing system. IRAF tasks are pure computational modules; where user interaction occurs, the user interface is provided externally by separate user interface modules in the system framework which communicate with the IRAF task at runtime using messaging.

All the IRAF software is integrated into a single system, but all of the applications packages and tasks, and most of the system modules such as the command language interpreter and the visualization tools, are modules which can easily be replaced or substituted. This flexibility is the chief reason that so many institutions have been able to extend IRAF to process their own data, or otherwise integrate IRAF into their local data processing systems. Every IRAF installation is unique, consisting of only those packages which the user has obtained and installed in the system.

From an architectural point of view IRAF consists of a host system interface (HSI), which includes all the system dependent code, and the portable IRAF system, which includes most of the system framework and all applications. The portable code is identical on all IRAF platforms, including all build files; there are no 'ifdefs' or other system-dependent features in the source code or build files. In the runtime system all binary object code and executables are segregated from the rest of the system into 'bin' (binary) directories. A single IRAF installation can simultaneously support any number of architectures, or target platforms; every major class of platform must, however, have a separate HSI.

Applications

An IRAF installation consists of the system framework and the applications packages and tasks installed in the system. As we noted earlier, IRAF tasks (programs) are organized into packages of logically related tasks. Above this there is one additional level of structure, collections of logically related packages. Chief among these are the core IRAF packages and the standard astronomical data reduction and analysis applications developed by NOAO and included in the standard IRAF release. In addition there are many external or layered packages which may be separately obtained and installed in IRAF. Finally there are many individual programs, including user-written Fortran or C (IMFORT) programs, which have been developed by users primarily for their own use. It is this broad collection of IRAF science applications, many of which were contributed by the community using IRAF, which more than anything else has made IRAF such a successful and widely used system.

Core system

The core IRAF packages are generic system packages providing standard facilities for importing and exporting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

data, general image processing, graphics and image display and data visualization, and user interaction including the IRAF command language (CL), as well as facilities for software development.

Table 1. Core IRAF packages.

dataio	Data format conversion package
images	General image processing package
language	The command language itself
lists	List processing package
local	The template local package
obsolete	Obsolete tasks
plot	Plot package
proto	Prototype or interim tasks
softools	Software tools package
system	System utilities package
utilities	Miscellaneous utilities package

The major packages in the core IRAF system are listed in table 1. Each of these packages contains many tasks. The images package, for example, contains approximately 100 tasks which provide facilities for image filtering, image matching, fitting functions and surfaces, geometric corrections, coordinate manipulation, and general image manipulation. Altogether the core system contains three or four hundred tasks.

Astronomical applications

The standard IRAF distribution also includes a suite of packages for the reduction and analysis of optical and infrared astronomical data called the 'NOAO' packages (table 2). Although these packages include special support for reducing data from the NOAO instruments, most of the packages therein are general enough to be applied to data from similar instruments outside NOAO. Packages are provided for reducing both image and spectral data, as well as standard astrometric, photometric and spectroscopic data analysis. All told the NOAO packages contain about 1200 individual tasks.

Layered packages

Layered or 'add-on' packages are optional applications packages which may be separately obtained and installed into IRAF. Once installed, layered packages are no different from the standard packages included in the IRAF distribution. The NOAO package is actually a layered package itself, even though it is included in the standard IRAF distribution.

Many layered packages have been developed for IRAF, enough so that it has been difficult to track them all. It is hard to be fair in listing layered packages, but a few of those in use at NOAO are listed in table 3.

Table 2. NOAO packages.

artdata	Artificial data generation
astrometry	Astrometry package
astutil	Astronomical utilities package
digiphot	Digital stellar photometry
focas	Faint object classification
imred	Image reductions package
nobsolete	Obsolete tasks to be phased out
nproto	Prototype–interim tasks
observatory	Observatory database
onedspec	One-dimensional spectral analysis
rv	Radial velocity analysis package
surfphot	Galaxy isophotal analysis package
twodspec	Two-dimensional spectral analysis

Table 3. Miscellaneous layered packages.

arcon	CTIO Arcon CCD data acquisition
color	RGB color image composition
ctio	CTIO miscellaneous tasks
dimsum	IR image reductions
euv	EUVE extreme UV reductions
finder	ST guide star catalog astrometry
ftools	HEASARC FITS tools package
grasp	GONG helioseismology package
ice	KPNO/Steward ICE CCD data acquisition
mscred	CCD Mosaic image reductions
nso	National Solar Observatory reductions
rgo	Royal Greenwich Observatory reductions
steward	Steward Observatory reductions
stsdas	Hubble Space Telescope reductions
tables	STScI tabular data format package
xray	SAO x-ray data analysis package

The most notable of the layered packages are the tables and stsdas packages from STScI, which provide extensive facilities for reducing HST data as well as many other generally useful IRAF tasks. The xray and ftools packages also provide many generally useful IRAF tasks. The ice (IRAF Control Environment) package is widely used for CCD data acquisition into IRAF. The mscred package from NOAO is used to reduce data from wide-field CCD mosaic imagers. Other major packages are available from outside groups such as Starlink in the UK and ESO/ECF in Europe.

If we were to add them all up, there are probably a few hundred IRAF packages in all, containing several thousand tasks in total. New IRAF packages are constantly under development.

Graphic and image display

Interactive graphics and image display has been an integral part of IRAF since the earliest days. Although these capabilities are built into IRAF, the actual graphics and image display servers are replaceable components which are distributed separately from the main IRAF distribution. NOAO currently supports a separate product called x11iraf which includes the graphics and image display servers xgterm (basically xterm but with vastly more powerful graphics) and ximtool, a powerful multiframe image display server. The Smithsonian Observatory (SAO) provides additional image display servers which can be used with IRAF, most notably SAOimage and SAOtng. Data interaction is a vital

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

part of IRAF and new facilities are continually under development.

In addition to basic graphics and image display, IRAF includes capabilities for integrated application graphical user interfaces, or GUIs. The system support for these GUIs is included in x11iraf. The actual GUI applications are essentially standard IRAF tasks, with additional messaging capabilities used to link the computational application with the external GUI. This allows the GUI technology to evolve without greatly affecting the science software at the heart of IRAF.

Data formats

The fundamental data objects in IRAF are the image and the table. Logically, images are N-dimensional pixel arrays with an associated image header containing keyword = value attributes describing the image. Images may be complex objects, including associated information such as pixel masks or world coordinate systems describing the mapping of an image onto the sky, or a spectral dispersion. Examples of subclasses of images might include direct images of the sky, spectra or variance arrays. Observations are composed of one or more logically associated images and auxiliary data objects such as tables or other associated information.

IRAF uses FITS for all archival data. Internally IRAF supports a number of runtime image formats used for various types of data. Each image format has a corresponding image kernel in IRAF, which maps the physical or external data format to the logical image data model used in IRAF.

The currently supported IRAF runtime image formats are summarized in table 4. All of these are pixel (raster) image formats except the qpf format, which is a sparse event list used for event-counting detectors. The pixel list image format stores typical image masks in a highly compressed internal format. Most IRAF applications support only the abstract image model defined by IRAF and will work with any of these image formats.

Table 4. IRAF runtime image formats.

oif	Native (original) IRAF image format
fxf	FITS image format
plf	Pixel list image format
qpf	Event data format
stf	HST GEIS image format

Many additional image formats can be used, but a separate image format conversion is required before the image can be used within IRAF. For example, the PC image formats such as GIF and TIFF can be used with an extra data conversion step (see the dataio package).

Special applications

Data acquisition

In addition to conventional data reduction and analysis, IRAF is also used at the telescope for data acquisition and quick look. Using IRAF for data acquisition allows a single familiar environment to be used for all data interaction both at the telescope and at the observer's home institution.

The ice package (IRAF Control Environment), an IRAF layered package, provides integrated control of a detector, instrument and telescope, allowing various types of exposures to be taken with the observation written directly to an IRAF image file at the end of the exposure. Automatic display, postprocessing and archiving of newly acquired images are supported. A more specialized package ccdphot is available for digital aperture photometry with CHARGE-COUPLED DEVICES (CCDs).

More recently the Mosaic Data Handling System has been developed for the acquisition and quick-look processing of CCD Mosaic images (or single-frame CCD images). The companion IRAF package mscred provides for the reduction of Mosaic data. The Mosaic DHS system is component based, using messaging to connect the components and a data feed component to interface to an external detector subsystem and feed data to the DHS. A data capture agent captures Mosaic data and writes them to disk as multi-extension FITS files. Real-time image display and quick-look data interaction capabilities are also provided.

Pipeline processing and archiving

IRAF has long provided standard data reduction packages for a wide range of instruments. In the classic IRAF system data is reduced by running a reduction task or imred (image reduction) CL script on a list of unprocessed images. Some interaction is typically required to reduce the data, especially for spectral data.

More recently an automated pipeline processing facility has been added. A pipeline manager executes the pipeline, driven by an externally defined rule-based pipeline script. External pipeline modules (IRAF tasks or other programs) perform all the actual processing operations. The pipeline manager continually loops over the data set, classifying objects by type and applying processing rules to move data through the stages of processing defined by the user-defined pipeline being executed. The pipeline manager has the ability to run a number of tasks concurrently, on the same host or on a set of cooperating hosts. The parallel processing capability allows this scheme to be scaled up to handle very large data sets.

The automatic pipeline processing capability can be combined with a database system to generate a data archive. In general, whenever an automated pipeline is set up to routinely process data for a particular instrument, it is desirable to go one step further and create an online archive of the data products. If a pipeline is set up and the data are put online, IRAF provides an integrated framework for all data handling from acquisition and quick look at the telescope, through pipeline processing, to interaction with the processed data from the online archives.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Future developments

The framework for a new IRAF system is being prototyped in the Mosaic Data Handling System and the automated pipeline processing facilities discussed in the previous sections. A related NASA-funded project called Open IRAF is adding a multilanguage programming development capability to IRAF, allowing applications to be written in a number of standard compiled or interpreted languages. Other projects are underway to add advanced GUI capabilities and to interface modern scripting languages such as Python and Tcl to IRAF.

The new framework will be component based, using a message bus to connect the components. The system framework will provide standard components for application scripting, GUI presentation, data visualization, parameter editing, online help and the like. Standard data service components based on distributed shared object technology will be used to provide efficient shared concurrent access to data objects such as images or tables. Applications will be assembled from custom compute components which implement the numerical or science algorithms, combined with mostly standard user interface components for user and data interaction. An interpreted script written in one of the supported scripting languages, executing in a script language component on the message bus, will tie the components together.

The new system framework will be inherently distributed since the message bus will allow components on different hosts to interoperate and since distributed shared objects will allow data objects such as images to be simultaneously accessed by multiple components or applications. Including a capability to directly access images and tables stored in local or remote data archives will allow applications written within the new framework to be used for data mining of archival data.

Ultimately the technology developed by these projects will be used to rebuild the framework of the existing IRAF system. Although the IRAF framework and the major user interfaces will be almost completely new, the existing science applications (most compiled tasks) are expected to migrate to the new framework almost unchanged. This will give us a modern new IRAF system while preserving most of the 200 or so man-years invested in the current science applications.

Availability

IRAF is available for most major Unix-based platforms, including both workstations and PCs. A Windows port is possible but does not yet exist. VMS has been supported for many years, but support will be dropped in versions of IRAF beyond V2.11. The set of supported platforms varies from year to year; check the IRAF web pages (http://iraf.noao.edu, or any mirror) to determine whether support is available for a particular platform.

A typical IRAF installation requires a minimum of 50–150 MB of disk space. If some of the larger layered packages are installed, or an installation is configured to support multiple target architectures, the space required can be several times larger. The largest installations, including the most common layered packages and supporting several target architectures, can require a gigabyte or more of disk space.

IRAF is freely available via download over the internet, or a mailed CD-ROM distribution can be obtained from NOAO for a nominal fee.

Doug Tody

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Software: Radio Pulsar Software

Software used to process radio pulsar data is almost all written in Fortran, C and C++ and runs mainly on UNIX platforms. The size of the user base is such that most of the major groups have their own home-grown solutions to what are very similar problems. However, there are generic features in the software used in the discovery, timing and understanding of the statistical properties of PULSARS.

When the first pulsars were discovered in 1967, computers were not common in radio observatories. Now they play an essential role in the discovery, timing and analysis of pulsars as well as in Monte Carlo simulations that enable us to understand the population. In the 32 yr since the discovery of pulsars we have witnessed an exponential growth in the power of computers with the speed at which they are able to perform calculations doubling every 18 months or so. This has made it possible to undertake large-scale surveys of the Galaxy for pulsars where over a terabyte of information has been recorded and processed. In the future real-time processing of the Nyquist-sampled electric field will be possible for ultra high-precision observations of radio pulsars. At the end of this article a number of WWW (world-wide web) links are given to assist the reader to download software and use online resources (see also SOFTWARE IN ASTRONOMY).

Pulsar searching software

There is no standard pulsar search software used by groups around the world. This is due in part to the fact that most observatories have their own data acquisition systems and so the format of pulsar search data differs. There is also the fact that most astronomers are trained in physics, not computer science, and the flexibility of the software is not always what it could be. Search codes often take a long time to perfect, and, once this is achieved, the source code is rarely distributed. This would mean providing adequate documentation, which is a chore, and there is an implied responsibility that the code is bug free. In my experience, most groups are more than happy to provide mature codes which address most aspects of pulsar astronomy provided that they are not expected to provide a 24 h support line and that the final responsibility for published material lies with the authors of the paper, not the authors of the software.

Almost all pulsar search codes operate on a two-dimensional (2D) data array of time and frequency samples. Because of interstellar pulse dispersion, a large number of frequency channels are required to prevent dispersion reducing sensitivity to radio pulsars. Multichannel filterbanks are the most commonly used data acquisition system, but in the past autocorrelators have been used to form filterbank data and in the future it is possible that searches will record the electric field directly to obtain the ultimate frequency resolution to improve sensitivity.

There are two basic methods of processing these data. The first is to dedisperse at several hundred different trial dispersion measures, perform a fast Fourier transform (FFT), look for spikes, and progressively sum harmonics until 32 or so are added, looking for significant peaks after each harmonic sum. A list of candidates can then be formed, and either the raw data folded at the period and dispersion measure corresponding to the peak to explore the temporal and frequency structure of the candidate, or, as some groups do, just taking note of candidates well above the noise. The alternative is the 2D FFT method, which is faster computationally, but has a more complicated harmonic summing algorithm. Pulsars appear as a series of harmonics equally spaced in the 2D FFT. To be efficient, this requires at least enough random access memory (RAM) to accommodate the 2D FFT. For the Parkes 70 cm all-sky survey this would have been $256 \times 512 \times 1024 \times 4 = 512$ megabytes (MB) which is too large for most typical workstations (in early 1999). At Parkes, larger bandwidths are possible in contrast to most other observatories and the number of channels is particularly large. At Arecibo and Jodrell Bank, it is becoming much easier to fit the search data into RAM and process the search data very quickly. In a recent survey, the Princeton group used the 2D FFT method and reported an increase in processing speed of a factor of 3; see Camilo *et al* (1996).

Unfortunately, the major headache for pulsar astronomers is the eradication of periodic interference from their datasets. Almost all of the recent software developments in search software, at least in our group, have been aimed at eliminating man-made signals which mimic pulsars. One can do this in a number of ways, but it is customary to run through all of the data in a quick pass and find periodic signals that are present in a very large number of pointings and could not possibly be pulsars. Very effective filters can then be implemented in software.

It is wise to fold the original data at the period of the candidate pulsar using the best dispersion measure (DM) and view the pulse profile as a function of both time and frequency. This can eliminate a lot of interference which is often narrow band in character or is not a pure tone, sometimes referred to as having a low Q. The only way to confirm that a source is really a pulsar, however, is to re-observe it.

Pulsar timing software

Once found, a great deal of science can often be deduced from regular observations of pulsars. By cross-correlating the average profile with a standard template, topocentric arrival times can be deduced and fitted to a simple timing model of a pulsar. The model includes such basic parameters as the position in the sky, DM, period and period derivative, from which most of the physical parameters of a pulsar are derived.

When a binary pulsar is discovered it soon becomes apparent that additional parameters are necessary to adequately describe the arrival times. The normal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

procedure is to obtain as accurate a measure of the pulsar's apparent period as possible at several different epochs and then to perform a least-squares fit to the periods for a Keplerian orbit using a program such as fitorbit, originally written by Brinklow and Lyne at Jodrell Bank. The mathematical foundation of fitorbit can be found in Smart (1977). A similar program 'binfit' was developed by the Princeton pulsar group and is probably more portable. Once several orbits have been observed a first-order fit is usually sufficient to attempt a timing solution with a timing program such as TEMPO.

TEMPO is the most widely used pulsar timing program, currently maintained by the Princeton group and the Australia Telescope National Facility and freely available from the WWW. Numerous authors have made contributions to this program over the years and it has recently undergone extensive revisions. The program runs under LINUX, as well as on SUN workstations and COMPAQ's Digital Unix, and ports have been made to SGI and HP. The only issues to address when attempting a new port are usually byte swapping and the record length of Fortran records.

TEMPO is capable of taking an input pulsar model and list of arrival times and fitting for almost any reasonable parameter. Taylor has introduced detailed support for a variety of models of relativistic gravity and TEMPO was used to time the celebrated 'Hulse–Taylor' (Hulse and Taylor 1975) binary pulsar to a high degree of accuracy which ultimately led to the awarding of the 1993 Nobel prize for physics (see RADIOASTRONOMY).

TEMPO was originally written in the era of punch cards, and for a long time it demanded input parameters in a very strict format in the first six lines of the input file. The top line was a series of 1s and spaces and it was necessary to learn which parameter occupied which column if you wanted to fit for it. Although black belts in TEMPO would not have it any other way, to beginners this proved an impediment to mastery of the program. In the latest revision (version 11) there is a new user-friendly input structure, which is free format and much easier to use.

TEMPO also has the ability to produce an ephemeris suitable for reducing gamma-ray observations made with the COMPTON GAMMA-RAY OBSERVATORY (CGRO) and has a very useful facility to create a polynomial description of a pulsar's topocentric period and phase at any given epoch. This is often used when observing to fold incoming pulsar data at the correct rate to form a nice sharp profile with no smearing.

Pulsar database programs

The new TEMPO input format is identical to the most widely used pulsar database program PSRINFO, developed by Harrison at Jodrell Bank in the early 1990s. The source code now resides at Swinburne University of Technology and has been expanded. This is one program which is freely available with a web interface, and it is possible to interrogate the catalogue online and also create publication quality plots of the pulsar population using PSRPLOT, a program written by Hughes at Swinburne. PSRPLOT uses PSRINFO to retrieve information from the catalogue and has the ability to select populations of pulsars based on several different criteria. The user can request a gif or Postscript version of the plot and a list of pulsars and parameters is provided to the user.

PSRINFO can also create an ephemeris file suitable for input straight into TEMPO, and it is therefore easy to generate polynomial coefficients for observing with a script. PSRINFO can also access the database for specific pulsars or provide lists of parameters. The source code is available on request from Swinburne.

PSRCAT is another catalog program which is less flexible than PSRINFO but has nice formatting abilities when preparing catalogs for the literature and is maintained at Princeton and the Australia Telescope National Facility by Taylor and Manchester.

The European pulsar community provided one of the most sensible advances in pulsar astronomy by standardizing their data into one format. Details of the format are freely available on the WWW. This is probably a sensible standard to follow for new groups as the software to read this format and manipulate it is available from the web site. One of the most useful resources on the web is the European pulsar network's database of pulsar profiles. There are literally hundreds of pulsar profiles nicely indexed and these form a valuable resource. There are numerous mirror sites around the world.

Coherent dedispersion

Pulsar signals are dispersed by the ionized interstellar medium and this smears the observed profiles of radio pulsars. The most common method to minimize this dispersion is to build a multichannel filterbank or CORRELATOR system as the smearing is proportional to the filter width. However, for the fastest pulsars, this is often insufficient to prevent a large amount of pulse dispersion within a single frequency channel.

In their 1975 paper, Hankins and Rickett describe in some detail a method for removing the effects of dispersion exactly by recording the electric field in the antenna at the Nyquist rate and 'coherently dedispersing' the data. This involves filtering the complex data using effectively the inverse filter to that of the interstellar medium in a deconvolution process that gives the highest time resolution possible. At several major observatories around the world coherent dedispersion software is being written. Even in 1999 coherent dedispersion is difficult, because, with any reasonable bandwidth, the amount of data is immense. For a 10 MHz bandwidth and two-bit sampling, observers are required to record 10 MB s^{-1} of raw data and take this away for processing. The Caltech–Parkes–Swinburne Recorder (CPSR) is located at the Parkes observatory and it is capable of recording 20 MB s^{-1} continuously using four DLT7000 tape drives and six 9 GB disks for destriping the data for convenient post-processing. Similar systems are in place at Arecibo, Jodrell Bank and Westerbork but with more modest bandwidths.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The software used for this processing has to be able to unpack the two-bit numbers, filter them and fold the data at the topocentric pulsar period. Although this sounds easy, the major problem is that in a week-long run the amount of data recorded can be 10 MB s^{-1} for 7 days, which is 6 TB. Just to read these data back on a single DLT7000 takes 14 days. On the Swinburne supercomputer, our coherent dedispersion software takes 64 times real time to process a 20 MHz band on a single processor. Hence it is imperative to ensure that the code has a high degree of parallelization in it or that one can run on multiple processors. Groups thinking seriously about commencing coherent dedispersion timing programs should think carefully about the computer power required. As computers become more powerful this will be less of a consideration.

Coherently dedispersed data have a time resolution of $1/B$, where B is the observing bandwidth. It therefore makes it possible to study the pulsar magnetosphere on submicrosecond timescales—provided that the DM is accurately known.

Software for understanding the pulsar population

When we consider the pulsar population as a whole, we can learn interesting things about their birthrate, kinematics and evolution. The best way to do this is to perform Monte Carlo simulations of the Galactic population of pulsars. By modeling the major pulsar surveys it is possible to tweak the parameters of the input population until it mirrors what we have discovered. There are at least a half a dozen different groups that have attempted such studies, but none of the codes are publicly available.

These studies often vary in their approach and complexity. The most complicated generate each pulsar from birth and follow the trajectory through the Galaxy by modeling the Galactic potential. Simpler models just assume that, by the time a pulsar dies, the effect of acceleration in the Galactic potential is small and ignore it. The right approach depends on the issue you wish to address. Millisecond pulsars live for many oscillations in the Galactic potential whereas magnetars spin down very rapidly. Short-cuts invoked to save computer time a decade ago are becoming less important as computers grow in processing power.

A modern approach might be to model the Galactic potential and generate pulsars every N years with some initial Galactocentric and Z distribution with a variable velocity distribution. By defining an initial period and magnetic field distribution it is easy to determine the period and magnetic field distribution of the population at any instant in time. By assigning a luminosity law as a function of period and magnetic field strength, it becomes possible to determine the radio flux density of a given pulsar at the position of the Sun. A model of the distribution of free electrons and scattering can then determine the dispersion measure and scattering of the pulsar and its detectability in various surveys can then be determined by comparison with a formula such as

$$S_{\min} = \frac{10(T_{\rm rec} + T_{\rm sky})}{G(BN_{\rm p}t)^{1/2}} \left(\frac{P}{P-w}\right)^{1/2}$$

where $S_{\min}$ is the minimum detectable flux density, $T_{\rm rec}$ and $T_{\rm sky}$ are the temperatures in kelvin of the receiver and sky respectively, G is the antenna gain in K Jy^{-1}, B the bandwidth in hertz, $N_{\rm p}$ is the number of polarizations of light (maximum 2), t the integration time, P the period of the pulsar and w the observed width of the pulsar, a combination of the intrinsic width, scattering, DM smearing and the sampling time of the data. If the width is larger than one pulse period the pulsar is invisible in a standard pulsar search.

Although such codes are popular, in my opinion they often tell us more about what we do not know or cannot really say about the pulsar population than what we do know about it. Any estimate of the pulsar birthrate should not be trusted to better than a factor of 2. Anyone who has taken part in a pulsar survey will tell you that (especially in recent times) the effects of interference on a survey's sensitivity are impossible to accurately estimate.

Bibliography

A database of pulsar profiles often with full polarization information and information on the European Pulsar Network's (EPN's) data format and software can be found at

http://www.mpifr-bonn.mpg.de/div/pulsar/epn/epn.html

The source code and help on TEMPO can be found at

http://pulsar.princeton.edu/tempo

and also at

http://www.atnf.csiro.au/research/pulsar/psr/timing/tempo

The pulsar catalog and plotting program can be found at

http://www.swin.edu.au/astronomy

A paper which outlines how to search for pulsars is

Manchester R N *et al* 1996 *Mon. Not. R. Astron. Soc.* **279** 1235

Two papers describing how to do population synthesis are

Lorimer D R, Bailes M, Harrison P A and Dewey R J 1993 *Mon. Not. R. Astron. Soc.* **263** 403

Narayan R 1987 *Astrophys. J.* **319** 162

TEMPO is described in

Weisberg J and Taylor J H 1989 *Astrophys. J.* **345** 434

Coherent dedispersion is described in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hankins T H and Rickett B J 1975 *Methods Comput. Phys.* **14** 55

Other references:

Camilo F, Nice D J, Shrauner J A and Taylor J H 1996 *Astropys. J.* **469** 819–27
Hulse R A and Taylor J H 1975 *Astropys. J.* **195** L51
Smart W M 1977 *Textbook on Spherical Astronomy* (Cambridge: Cambridge University Press)

Matthew Bailes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Software: Stellar Photometry

The term 'photometry' refers to the accurate determination of the apparent brightness of an astronomical object. Differences in brightness as measured through several colored-glass or interference FILTERS (for example) permit the definition of color indices, from which the temperature, chemical composition or other properties of the object can be approximately inferred. Changes in the apparent brightness of certain types of astronomical object over time also allow interesting conclusions to be drawn about their intrinsic nature. Photometry plays a key role in the measurement of the ages of star clusters and the distances of external galaxies.

Until roughly 1980, nearly all astronomical photometry was done by means of analog measurements of photographic plates (for instance, with an optomechanical device called an 'iris photometer'), or by analog or digital (photon-counting) techniques with photomultipliers. These photometers produced brightness readings which were typically displayed on dials, plotted on strip charts or printed on strips of paper, and it was often quite practical to analyse these raw data with pencil, paper and a slide rule or table of logarithms. However, during the late 1970s electronic area detectors for astronomy became more advanced: first, for a brief period, television-type cameras were employed, but these were soon supplanted by CHARGE-COUPLED DEVICES (CCDs), which remain overwhelmingly the detector of choice to the present day. CAMERAS employing these devices can produce vast quantities of data directly in electronic form, and the only realistic way to convert these bit strings into human knowledge is by computerized data analysis.

There are two properties of the older types of photometric measurement that distinguish them from modern CCD photometry: previously, (1) one observation produced a measurement consisting of one number and (2) the astronomer attempted, at the time of the observation, to measure *only* the particular object of interest, isolating its light from that of other, nearby astronomical objects. Two or more astronomical objects which lie too close together on the sky cannot be effectively measured either with a photoelectric photometer, owing to the large measuring aperture needed (see below), or on a photographic plate, owing to its nonlinear response to the incoming light. In a modern area detector, which automatically produces a linear brightness map of a given patch of sky, the image of any given astronomical object will usually be spread over several picture elements ('pixels') in the data array, and its brightness will therefore be divided among several simultaneous but separate measurements. Furthermore, the linearity of the detector combined with the fact that each object's image is divided among several pixels allows the *ex post facto* disentangling of the individual brightnesses of numerous different astronomical objects whose images were recorded in the original observation.

There is also one extremely important fact that distinguishes modern photometry of *stars*, in particular, from photometry of other astronomical objects such as galaxies and gaseous nebulae: the apparent angular sizes of all stars as viewed from the Earth (except, of course, the Sun) are very much smaller than the angular resolution permitted by the terrestrial atmosphere and the optical elements of the telescope. Because in any given observation the atmosphere and telescope affect all stars comparably, the perceived images of all stars in a data frame have almost exactly the same shape. (In fact, some camera systems and some techniques—such as observing with ADAPTIVE OPTICS—can produce stellar images whose shape does vary somewhat with position within the frame. Except in the worst cases these variations are small and easily accounted for.) As will be seen below, this property is routinely exploited by modern software to extract stellar photometry from digital images.

The entire process of STELLAR PHOTOMETRY can be broken down into three stages which, for want of better terms, can be identified as follows.

1. *Preprocessing*. The raw digital frames from the telescope are rectified onto a linear intensity scale by operations such as bias subtraction and flat fielding.
2 *Processing*. Each rectified digital frame is searched for the images of individual stars, whose positions and relative brightnesses on an instrumental scale are then determined with the highest possible precision.
3. *Postprocessing*. The results of stage 2 from numerous different frames are combined and calibrated to fundamental magnitude scales by comparison with similar observations of other stars having known brightnesses. Measurements made with different filters can then be compared to form observational colors and/or observations made at different times can be assembled into light curves for variable stars.

The preprocessing stage will not be discussed here. Stage 2 is by far the subtlest part of the whole process, and a great deal of creativity has gone into the development of specialized software for extracting the desired quatities; a few specific software packages for this purpose have become quite widely used (see SOFTWARE IN ASTRONOMY). The postprocessing that occurs in step 3 is conceptually the same as the calibration process for photoelectric photometry. For the most part, general statistical modules or home-grown software have been used for this purpose, so postprocessing will not be discussed in detail here.

The simplest way to extract a brightness measurement for a star which is well isolated from other astronomical objects in its digital data frame simulates the type of measurement that is made with photomultipliers. This older type of photometer contains within it a flat piece of metal with a small hole, or 'aperture', drilled through it, whose diameter generally corresponds to 10 arcsec or more projected onto the sky. The telescope is used to focus an image of a star onto this aperture, which permits the light of the star and some light from the foreground

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

night-time sky to pass through and be measured by the photomultiplier. With the same aperture, a separate measurement is made for an empty patch of sky with *no* star visible in it, permitting the brightness of the night-time sky alone to be determined. When subtracted from the other measurement, this yields an estimate of the brightness of the star alone. An analogous process may be carried out with stellar images in a digital data frame: one simply defines a synthetic 'aperture' consisting of all the pixels lying within a certain distance from the center of the star image in question, and sums the brightnesses measured in all of those pixels. One also determines the average brightness of the night-time sky from the values measured in nearby pixels which do *not* lie within the images of perceptible stars in the frame. The average sky brightness per pixel is scaled up to the area of the synthetic star aperture and is subtracted from the stellar measurement. This difference represents the brightness measurement for the star alone.

The size of the best digital 'aperture' to use in synthetic-aperture photometry depends on several factors. To provide a reliable measurement, the aperture must be large enough that a constant *fraction* of the light of any given star will be measured, regardless of where the star falls with respect to the pixel grid of the detector. On the other hand, a large aperture produces a measurement with more observational uncertainty, resulting from the readout noise of the detector, random noise in the contaminating light of the night-time sky, the increased possibility of including other astronomical objects in the 'aperture' and the greater likelihood of encountering a detector blemish or cosmic-ray event in the data. For comparing brightnesses of different stars measured in the *same* data frame, a fairly small aperture (perhaps of order 1 arcsec in diameter) can usually be used, since to the extent that all stellar profiles have the same shape, a fairly small aperture can include a constant fraction of the profile, thus allowing precise *relative* brightness determinations. When it is necessary to compare brightnesses among different frames, however, it is usually necessary to employ a comparatively large aperture (at least several arcseconds in diameter), because differences in SEEING and telescope guiding will cause star images in different frames to have different profile shapes: one must then use a synthetic aperture that is large enough to contain all of the profile variations produced by these causes. In other words, measurements through small apertures have high precision (i.e. the random measuring noise is low) but poor accuracy (i.e. there may be large systematic differences from one data frame to the next due to differences in the stellar profile), whereas large-aperture measurements have low precision but high accuracy (they are noisy, but they mean the same thing from one frame to another).

In cases where numerous stars are measured in a given data frame, it is possible to combine measurements through both small and large apertures to yield results with the desirable properties of both. One can measure all the stars in the frame through a small aperture of fixed size, and one can then remeasure selected bright, well-isolated stars through a larger aperture, also of fixed size. By comparing these two results for the stars that were measured twice, one can derive an accurate mean value for the *ratio* of brightnesses as measured through the two apertures for that frame. A constant brightness ratio corresponds to a constant magnitude *difference*. This difference can then be added to the small-aperture magnitude measurements for faint and crowded stars, to yield the magnitude that *would have been measured* for them through the larger aperture, had it been practical to do so. These magnitudes have nearly as high a precision as the small-aperture measurements (because the correction, if determined from the average of many stars, can have comparatively little uncertainty) and the accuracy of the large-aperture measurements. The actual *or* inferred large-aperture magnitudes for stars in any one frame can then be compared with similarly derived large-aperture magnitudes for stars measured in other frames, even though the image profiles may have been rather different.

In general, while measurements through small apertures contain the least possible noise, the comparatively large corrections required to calculate the corresponding large-aperture magnitudes still do have some uncertainty. If a star is very faint, it is best measured through a very small aperture because in such a case the uncertainty associated with the aperture measurement is more critical than the uncertainty of the correction to the larger aperture (see also DETECTION OF FAINT OBJECTS). Conversely, if the star is bright it is better measured through a large aperture because otherwise the uncertainty of the correction would dominate over the uncertainty of the aperture magnitude. A more elaborate scheme of measurement through apertures of many different sizes can be undertaken, which would allow one to choose for each star the 'best' aperture which offers the optimum combination of uncertainty in the aperture magnitude and uncertainty in the correction. The way in which the instrumental magnitudes measured in a given data frame vary with the radius of the synthetic aperture used is called that frame's 'growth curve' (not to be confused with the spectroscopic 'curve of growth').

Even more precise measurements can be made at the expense of much more computational work by using the technique of 'profile-fitting' photometry. In this method, one exploits the fact that any given star image is spread over several pixels, and the internal shape of the image, broadened by the atmospheric seeing and the telescope optics, can be mapped out. To exploit this approach, one must generate a numerical model of the actual profile shape for star images in each data frame. This model profile is determined from the brightest, best-isolated star images in the frame and must be encoded in such a way that the computer can then predict the actual local brightness that should be measured at any point within the profile of a given star image. This profile, or 'point-spread function' (PSF) as it is called (since it describes the way the image of a point source—a star—is spread

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

out by the atmosphere and optics) can be encoded as a simple mathematical formula with parameters describing the variation of the local brightness in the profile as a function of position relative to the center of the star or as a table of actual numerical values extracted from the original data frame, from which specific brightness values are subsequently estimated by interpolation techniques.

An analytic model profile has the advantage of simplicity and computational speed but is limited in its ability to represent complex profiles produced by aberrated optics or guiding errors. A PSF encoded as an empirical look-up table can represent any arbitrary profile shape but is subject to the difficulty of achieving accurate interpolations for stars that fall differently on the underlying pixel grid. Interpolation errors are especially critical when the diameter of a stellar image is less than a few pixels. Again, there are advantages to be achieved by combining the two types of encoding. An analytic function fitted by least-squares techniques to the profiles of several bright, isolated stars in a data frame is generally found to represent the actual form of those profiles with 95–98% fidelity. The numerical differences between the actual observed profiles and the fitted analytic models can then be computed and stored as a look-up table. In using this hybrid PSF to predict the brightness value for a pixel located somewhere in the profile of an arbitrary target star, first a value corresponding to that pixel's location is calculated from the analytic approximation to the PSF, and then a correction to the actual empirical profile is estimated from the look-up table. The analytic approximation to the PSF is simply a mathematical formula, so its contribution to the total profile can be interpolated with arbitrary precision, while the amplitudes of the profile corrections stored in the look-up table represent only of order 2–5% of the total amplitude of the PSF, so that errors in the tabular interpolations are proportionately reduced.

Once the PSF for a given data frame has been defined in suitable computer-accessible form, it can be fitted to the images of all detected stars in that frame using nonlinear fitting techniques. For each star, four fundamental parameters must be determined: the position, in x and y, of the centroid of the star within the frame, the local diffuse sky brightness in the frame (which may vary from position to position because of underlying light from other astronomical objects, such as unresolved stars or nebulosity) and the amplitude of that star's profile above the local diffuse sky brightness. This amplitude is linearly related to the relative brightness of the star, and brightness ratios between stars are readily converted to magnitude differences. Separate positions and magnitudes can be estimated even for stars that are so close together that their images are somewhat blended together in the data frame, by suitable simultaneous fits of overlapping model profiles.

The scaling of the model PSF to match the amplitudes of stellar profiles provides a very precise measure of the relative brightnesses among stars in any given data frame. However, since each data frame has its own PSF that may be quite different in form from those found in other frames, profile-fitting photometry does not necessarily guarantee accurate comparisons from one exposure to another. As a safeguard against the possibility of systematic errors in frame-to-frame comparisons, it is common to compare the relative, profile-fitting magnitudes for a number of comparatively bright, isolated stars in each frame with large-aperture magnitudes for the same stars which have been either measured directly or inferred from smaller-aperture magnitudes with growth-curve corrections. The average difference between aperture magnitudes and profile-fitting magnitudes measured for these stars can be added to the profile-fitting magnitudes for all stars in the frame, resulting in a magnitude scale that is both highly precise and fundamentally accurate.

Bibliography

The rationale behind and statistical properties of synthetic-aperture photometry and magnitude growth curves are discussed extensively in these two papers:

Howell S B 1989 *Publ. Astron. Soc. Pac.* **101** 616
Stetson P B 1990 *Publ. Astron. Soc. Pac.* **102** 932

Two of the currently most popular software packages for performing profile-fitting photometry were first described in detail in the following papers:

Mateo M and Schechter P 1989 *1st ESO Data Analysis Workshop (Garching)* ed P Grosbøl
Stetson P B 1987 *Publ. Astron. Soc. Pac.* **99** 191

P B Stetson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SOHO (Solar and Heliospheric Observatory)

The Solar and Heliospheric Observatory (SOHO) is a project of international cooperation between ESA (the EUROPEAN SPACE AGENCY) and NASA to study the Sun, from its deep core to the outer corona, and the solar wind. SOHO was launched on 2 December 1995, on top of an Atlas/Centaur combination, from Cape Canaveral Air-Force Base in Florida. It reached its operating orbit around the L1 Sun–Earth Lagrangian point in mid-February 1996. For more than two years this unique vantage position has allowed nearly uninterrupted observations of the Sun, 24 hours a day, 365 days a year.

The main scientific objectives of the SOHO mission are

- to study the SOLAR INTERIOR, using the techniques of HELIOSEISMOLOGY,
- to study the heating mechanisms of the solar CORONA through data obtained by imaging telescopes and spectrometers, and
- to investigate the SOLAR WIND and its acceleration processes, again by remote sensing, and by *in situ* particle measurements.

To achieve these objectives SOHO carries onboard a complement of 12 state-of-the-art instruments, developed and furnished by 12 international Principal Investigator (PI) consortia involving 39 institutes from 15 countries. Nine of these consortia are led by European PIs, and the three others by US PIs.

During the nominal mission, from 1 May 1996 to 30 April 1998, the SOHO spacecraft operated extremely well, exceeding its specifications in many instances, while the SOHO instruments collected a large number of data, allowing SOHO scientists to make substantial progress in all three science objectives. Scientific highlights from the nominal mission will be presented below. Because of SOHO's scientific success, ESA and NASA have both committed themselves to a mission extension of several years. On 24 June 1998, during routine maintenance operations, ground controllers lost contact with the spacecraft. Efforts to re-establish nominal communication succeeded seven anguish-filled weeks later, and a team of engineers from ESA and Matra-Marconi, the spacecraft contractor, succeeded in returning the spacecraft to full nominal status. All experiments onboard have been restored to the same operational status as before the accident, despite the extremes in heat and cold endured. In the aftermath of these events the third and last gyroscope onboard the spacecraft failed, but after some initial difficulties ground controllers succeeded in operating SOHO in a gyro-less mode, with the function of the gyros taken over by the reaction wheels.

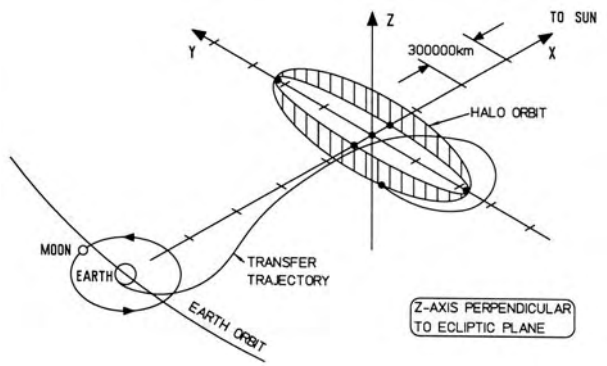


Figure 1. The SOHO orbit around the first Lagrangian point, 1.6 million km sunward from the Earth.

Development and launch

SOHO was first proposed in November 1982, in response to a call for mission proposals by ESA, as a high-resolution spectroscopic investigation of the upper solar atmosphere. In 1983 helioseismology was included as one of the science objectives. After a joint study by European and US scientists, the Science Program Committee of ESA approved SOHO as part of the Solar Terrestrial Science Program (STSP), to be implemented in collaboration with NASA.

ESA and NASA agreed that ESA would take the responsibility for developing the spacecraft, and NASA for the SOHO launch and operations. The 12 SOHO experiments were selected in 1988 after a combined ESA–NASA announcement of opportunity. The SOHO spacecraft was build by a consortium of European industries, led by Matra-Marconi.

SOHO is large, $4.3 \times 2.7 \times 3.7$ m^3, and weighs 1861 kg, with 655 kg for the payload and 250 kg of propellants. The maximum power supply from the solar cells is 1400 W, with a maximum payload consumption of 625 W. SOHO's reaction wheels are designed to provide short-term pointing stability better than 1 arcsec over 15 min, while the roll angle around the Sun is stabilized with the use of a star sensor unit to under 1.5 arcmin over 15 min. Two Gbits of solid state memory, backed up by a one Gbit tape recorder, can store the continuous scientific data rate of 40 kbit s^{-1}, plus 1.3 kbit s^{-1}, for more than 13 h. During continuous contact with the stations of NASA's Deep Space Network (DSN), another 160 kbit s^{-1} of MDI high-rate data are downlinked.

SOHO was launched on top of a Centaur-Atlas combination on 2 December 1995 at 8:08 UT. After one orbit around the Earth the Centaur engine was reignited to set SOHO on course towards the first Lagrangian point, about 1 million miles sunward from the Earth. The launch was so accurate, and the orbital maneuvers so efficient, that SOHO reached its halo orbit six weeks ahead of schedule, on 14 February 1996, with sufficient fuel left to maintain the (unstable) halo orbit for more than a decade, at least twice the time foreseen prior to launch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The first four months of SOHO operations were devoted to commissioning and performance verification of the spacecraft and the scientific instruments. By the end of this period the spacecraft was shown to be in excellent shape, with a pointing accuracy of 3 arcmin, better by a factor of 2 than the specification, and a short-term (15 min) stability of 0.3 arcsec, exceeding the specification by a factor of 3. With some minor exceptions, the experiments were all found to operate nominally or better than anticipated.

Experiments and science investigations

An overview of the twelve science experiments onboard SOHO is presented in table 1. The experiments are grouped according to their science objectives. The three interior and surface experiments (GOLF, VIRGO and MDI) use the techniques of helioseismology to measure the properties of the solar interior. GOLF measures the net velocity fluctuations of the entire Sun, VIRGO the intensity variations of the Sun in a dozen zones, and MDI makes maps of the entire surface in velocity, intensity and magnetic field on a 1024 × 1024 CCD. The GOLF and VIRGO instruments are designed to achieve high absolute precision in order to detect the Sun's internal gravity waves (g-modes). MDI is geared towards higher order acoustic modes (p-modes), in order to probe the detailed structure of the SOLAR INTERIOR CONVECTION ZONE. MDI surface maps also allow us to follow the generation and evolution of magnetic fields on a GRANULATION to global scale, with a spatial and temporal resolution sufficient to follow the life-cycle of individual magnetic structures. The continuity of SOHO observations is essential for achieving SOHO's helioseismology objectives. To obtain a signal-to-noise ratio necessary to detect the low-amplitude g-modes, and to measure the frequency of the p-modes with the desired accuracy of a few tens of nanohertz, years of virtually uninterrupted data are required.

SOHO's remote sensing experiments study the outer atmosphere and solar wind. Three of those, SUMER, CDS and EIT, concentrate on the transition region and corona at moderate to high spatial and spectral resolution. The others, UVCS, LASCO and SWAN, avoid the disk to provide data on the outer corona, including the SOLAR WIND ACCELERATION region. The CELIAS solar extreme-ultraviolet monitor (SEM) is a very stable photodiode spectrometer that measures the full disk solar flux in He II at 304 Å, as well as the absolute integral flux between 170 and 700 Å.

SUMER is a UV telescope with a normal incidence spectrometer with high spatial and temporal resolution (∼10 s). It measures line profiles and intensities of UV lines in the range from 500 to 1600 Å, with a resolution of 18 800 to 40 000. Thus SUMER can measure velocities down to 1 km s^{-1}. CDS operates at shorter wavelengths (150 to 800 Å), and consists of a Walter II telescope equipped with both a normal and a grazing incidence spectrometer. It determines line intensities in selected EUV lines to study the temperature and density of coronal structures. EIT obtains full-Sun high-resolution images, limited only by the 2.6 arcsec pixel size, in four emission lines, and corresponding to temperatures of 80 000, 1.3, 1.6 and 2.0 $\times 10^6$ K. EIT's observations provide the morphological context for the spectrographic data from CDS and SUMER.

LASCO is a triple coronagraph with nested concentric, and increasing annular fields of view. It also has a Fabry–Pérot interferometer for spectroscopic measurements with a spectral resolution of about 700 mÅ, in Fe XIV and X, Ca XV, Na D_2 and Hα. UVCS is an occulted telescope equipped with high-resolution spectrometers for spectroscopic observations of the solar corona out to 10 solar radii to determine the source regions, the acceleration and the heating of the solar wind. One UVCS grating is optimized for line profile measurements of Lyα, another for line measurements in the range from 944 to 1070 Å. SWAN, finally, maps the intensity of interplanetary Lyα, with two identical periscope systems, each of which observes one hemisphere at a resolution of 1 arcmin. Its goal is to determine the latitude distribution of the solar wind flux, and its variation. In addition it can measure the shape of the Lyα line at a resolution of about 3×10^5.

Of the three *in situ* particle experiments onboard SOHO, CELIAS studies the solar wind at L1, pristine in comparison with the data affected by the Earth's foreshock, which are collected by near-Earth spacecraft. The CELIAS proton monitor measures ions in the range 0.3 to 6 keV/e, and generates values of the solar wind proton bulk speed, density and north/south flow direction with a 30 s temporal resolution. The CELIAS mass determining time-of-flight (MTOF) sensor is a high-resolution mass spectrometer, which measures the solar wind's elemental and isotopic composition. CELIAS/CTOF, the charge determining time-of-flight sensor has obtained high time resolution charge state spectra for solar wind iron ions, but is no longer fully operational. Finally, CELIAS' suprathermal time-of-flight sensor (STOF) measures ionic charge states of particles with suprathermal energies from just above the solar wind up to low-energy solar flare particles (20–4000 keV amu^{-1}; see SOLAR FLARE OBSERVATIONS), thus bridging the gap between the solar wind instruments and the high-energy particle instruments ERNE and COSTEP.

ERNE and COSTEP are designed to detect Earthward traveling energetic particles ejected from the Sun in the course of flares and coronal mass ejections (CMEs; see SOLAR CORONAL MASS EJECTION: OBSERVATIONS). Electrons are recorded at energies from 44 keV to 50 MeV, and protons from 44 keV to over 100 MeV. Alpha particles are observed from a few MeV/nucleon to over 100 MeV/nucleon. Heavier elements, up to iron, can be identified up to 500 MeV/nucleon.

Operations and data archive

In nominal science operations each SOHO experiment carries out an individual synoptic or baseline science program, complemented for the remote sensing experiments and MDI by a coordinated mode, involving frequent joint

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. An overview of the twelve science experiments on SOHO.

Experiment	Investigation	Technique	Bit rate (kb s^{-1})
Helioseismology			
GOLF (Global Oscillations at Low Frequency)	Global Sun velocity and magnetic field oscillations, harmonic degree $l = 0$–4	Na vapor resonant scattering cell, Doppler shift and circular polarization	0.160
VIRGO (Variability of Solar Irradiance and Gravity Oscillations)	Low degree ($l = 0$–7) irradiance oscillations and solar constant	Global Sun and low-resolution (12 pixels) photometers, active cavity radiometers	0.1
SOI/MDI (Solar Oscillations Investigation/ Michelson Doppler Imager)	Velocity oscillations with harmonic degree up to 4500	Doppler shift with Fourier tachometer, 4 and 1.5″ resolution	5 and 160
Solar atmosphere remote sensing			
SUMER (Solar Ultraviolet Measurements of Emitted Radiation)	Plasma flow characteristics (temperature, density, velocity) chromosphere through corona	Normal incidence spectrometer, 50–160 nm, spectral resolution 20 000–40 000, angular res. 1.2–1.5″	10.7 or 21
CDS (Coronal Diagnostic Spectrometer)	Temperatures and density: transition region and corona	Normal and grazing incidence spectrometers, 17–80 nm, spectr. res. 1000–10 000, angular res. 2″	11.5 or 23
EIT (Extreme-Ultraviolet Imaging Telescope)	Evolution of chromospheric and coronal structures	Full disk images (42′ × 42′ with 1024 × 1024 pixels) in He I, Fe IX, Fe XII and Fe XV	2.0 or 6.6 or 27
UVCS (Ultraviolet Coronagraph Spectrometer)	Electron and ion temperature densities, velocities in corona (1.3–10$R_\odot$)	Profiles and/or intensity of spectral EUV lines (Lyα, O VI, etc)	5
LASCO (Large Angle and Spectrometric Coronagraph)	Coronal evolution, mass, momentum, and energy transport (1.1–30 $R_\odot$)	One internal and two externally occulted coronagraphs, Spectrometer for 1.1–3$R_\odot$	5.9 or 9.2 or 27
SWAN (Solar Wind Anisotropies Study)	Solar wind mass flux anisotropies and temporal variations	Scanning telescopes with hydrogen absorption cell for Lyα light	0.2
Solar wind 'in situ'			
CELIAS (Charge, Element, and Isotope Analysis System)	Energy distribution and mass, charge and charge state of ions 0.1–1000 keV/e	Electrostatic deflection, time-of-flight measurements, and solid state detectors	1.5
COSTEP (Comprehensive Suprathermal and Energetic Particle Analyser)	Energy spectrum of protons, α-particles, and electrons, composition of ions	Solid state detectors and electrostatic analysers	0.3
ERNE (Energetic and Relativistic Nuclei and Electron experiment)	Energy distribution and composition of ions (p, Ni), 1.4–540 MeV/n, and electrons 1–25 MeV	Solid state, and plastic and crystal scintillator detectors	0.71

operations programs (JOPs), occasional targets of opportunity, and regular campaigns with other spacecraft and ground-based observatories (GBOs).

The SOHO experiments were designed to complement each other, and through intense and innovative use of the Internet, SOHO has lifted the standard, frequency and documentation of coordinated solar observations to a new level. Images, observing sequences and target coordinates are exchanged instantaneously, while the data are automatically linked through the SOHO catalogs. Hundreds of campaigns have been carried out during the nominal mission. These vary from large-scale campaigns of several hours daily for one or two weeks, involving GBOs all over the world as well as other spacecraft (YOHKOH, TRACE, ULYSSES), to instrument intercalibrations of an hour or two, occasionally supported by rocket underflights.

The center of SOHO operations is the Experiment Operations Facility (EOF) at NASA's GODDARD SPACE FLIGHT CENTER, where experiment commanding, real-time coordination and long-term planning are executed. Representatives of the six interactive experiments (SUMER, CDS, EIT, MDI, UVCS and LASCO) are permanently present at the EOF to operate the mission in collaboration with the NASA Flight Operations Team and the ESA/NASA Project Sci-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

entist Team. The science observing programs are planned through a regular, nested series of meetings, that lead to a schedule that is both predictable enough to prevent overburdening operations scientists, and flexible enough to allow rapid reaction to observing opportunities.

SOHO is unique among solar physics missions in that data are received 'live', and the experiments commanded in near real time directly from the EOF for about 8 h a day. Scientists, from behind their workstations, can retarget and reprogram their experiments in a matter of minutes in response to events on the Sun.

Four SOHO data archives are being constructed, one in the US, at Goddard Space Flight Center, and three in Europe, at Medoc in Orsay, RUTHERFORD–APPLETON LABORATORY in the UK, and in Torino. These archives contain all experimental data, except for the MDI helioseismology data, which are stored in a separate facility at Stanford. To make the data easily accessible through the Internet for the widest possible group of scientific users, and to facilitate multi-experiment data analysis, the SOHO archive has a uniform data format (FITS), uniform access to all experimental data, a campaign catalog that cross-links experimental data of coordinated observations, a collection of synoptic data from observatories world wide, and a complete set of analysis software. The main SOHO catalog, an event catalog and the experiment catalogs round out the archive.

Science highlights

Exploitation of the enormous number of data from the SOHO nominal mission is only in its initial phase, but already a large number of exciting new results have been announced. A brief synopsis is given below.

Solar surface and interior

MDI medium angular degree data have provided the most precise measurements of the sound speed profile and the differential rotation profile within the Sun, thus setting new boundaries for the solar model

Application of a new technique, called time–distance helioseismology, to high-resolution MDI data has provided the first images of the convection zone of a star; maps of vertical and horizontal flow velocities, as well as sound speed variations in the convection zone, just below the visible surface.

MDI data have revealed a jet-like flow near the poles, which is totally inside the Sun and cannot be seen at the surface. Circling the Sun at about 75 deg latitude, this flow consists of a flattened oval region about 30 000 km across where material moves about 10% faster than its surroundings.

MDI data have also revealed at least six differentially rotating belts, three in each hemisphere. These belts are more than 65 000 km across and they move about 15 km h^{-1} faster than their surroundings.

The use of the new time–distance helioseismology technique in analysing MDI data has yielded that the entire outer layer of the Sun, to a depth of at least 25 000 km, is slowly but steadily flowing from the equator to the poles at about 80 km h^{-1}, fast enough to transport an object from the equator to the pole in just over a year.

High-precision MDI measurements of the Sun's shape and brightness obtained during two special 360° roll maneuvers of SOHO have produced the most precise determination of solar oblateness ever. There is no excess oblateness. These measurements unambiguously rule out the possibility of a rapidly rotating core and any significant solar cycle variation in the oblateness.

Results from the radiometers and photometers of VIRGO confirm that active regions modulate total and spectral irradiance on time scales of days to weeks.

Solar corona and solar wind

The UVCS coronagraph has revealed dramatic differences in the line widths of the strong hydrogen Lyα line and even more so in the resonance lines of O VI in coronal streamers and CORONAL HOLES at heliocentric heights from 1.25 out to 3.5 solar radii. Preferential acceleration of high mass ions affords one possible explanation for generating the observed velocities in the O VI ions.

Combined Doppler-dimming measurements of UVCS, LASCO white-light images of the solar corona, and radio scintillation measurements by the Galileo spacecraft have yielded new insight into the origins of the slow and fast solar wind. While the prevailing view is that the fast solar wind originates only in polar coronal holes and their equatorial extensions, the new measurements are interpreted as evidence that the slow solar wind is limited to the axes (or 'stalks') of coronal streamers while the fast solar wind dominates the corona.

A UVCS spectroscopic study of the composition in coronal streamers has revealed an ionization temperature of 1.6×10^6 K. The elemental abundances in the leg of the streamers and in ACTIVE REGION streamers are similar to those observed in the slow solar wind, indicating that the slow solar wind originates at the legs of the coronal streamers.

The CDS and SUMER spectrometers have observed extensive evidence of explosive events in transition-region lines. These could be the signature of MAGNETIC RECONNECTION events in nearby small magnetic dipoles, observed in MDI magnetograms. These events seem to account for a substantial part of the energy that is needed to heat the corona.

Coordinated measurements by SUMER and CDS have provided new observations of the temperature structure of coronal holes. The temperature measured at the base has a higher value than anticipated, of the same order as that of closed field regions (about 10^6 K), but it decreases much more rapidly with height than had been expected, falling to about 0.3×10^6 K at 1.3 solar radii. This is not consistent with a thermally driven fast solar wind.

Time-lapse sequences of white-light images from the LASCO coronagraph, obtained during sunspot minimum conditions in 1996, have been used to measure, for the first time, the acceleration of the solar wind. The outflow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of about 150 features in the streamer belt has been tracked between 3 and 30 solar radii. The speed versus radial distance profiles cluster around a parabolic path characterized by a constant acceleration of about 4 m^{-2} throughout the field of view. That profile is consistent with an isothermal solar wind expansion at a temperature of $\approx 1.1 \times 10^6$ K and a sonic point around 5 solar radii.

EIT and LASCO have recorded the very onset and evolution of several Earth-directed CMEs, a few of which actually hit the Earth's magnetosphere (see MAGNETOSPHERE OF EARTH). This type of observation holds considerable promise for the further development of space weather forecasting methods. The LASCO instrument is collecting an extensive database for establishing firm statistics on CMEs and their geomagnetic effects.

The LASCO coronagraphs have discovered more than 50 sungrazing comets, several of which were observed crashing into the Sun (see COMETS: KREUTZ SUNGRAZING).

The *in situ* particle detectors (ERNE and COSTEP) regularly detect high-energy protons, electrons and alpha-particles, in the aftermath of prominence eruptions and CMEs, while CELIAS, SOHO's solar wind analyser, has doubled the number of elements and isotopes previously recorded in the solar wind. During a Venus disk passage CELIAS also observed, for the first time from such a large distance, pick-up ions from the Venusian ionosphere, finding a much smaller diffusion cone than anticipated.

SWAN, looking away from the Sun, has recorded full-sky Lyα maps at a resolution and signal-to-noise ratio of at least a factor 10 better than previous observations. From these maps, the latitude distribution of the solar wind is determined.

Bibliography

SOHO Web Site with the latest information on mission status and observing programs, and a picture and movie gallery: `http://sohowww.nascom.nasa.gov/index.html`

Fleck B and Svestka Z (ed) 1997 The first results from SOHO (reprinted from *Solar Phys.* **170**(1), **175**(2))

Fleck B, Domingo V, and Poland A (ed) 1995 The SOHO Mission (reprinted from *Solar Phys.* **162**(1,2))

Lang K R 1997 SOHO reveals the secrets of the Sun *Sci. Am.* March 40

Wilson A (ed) 1997 The corona and solar wind near minimum activity *ESA Report* SP-404

Petrus C Martens

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Abundances

What is the Sun made of? This fundamental question only received an answer about 70 years ago. H N RUSSELL, whose name is associated with much pioneering research in astrophysics during the first half of this century as well as with a series of basic work in atomic spectroscopy, made the first quantitative analysis of the chemical composition of the SOLAR PHOTOSPHERE in 1929. Using eye estimates of solar line intensities, he succeeded in deriving the abundances of 56 elements. He also showed that the Sun and, finally, the universe were essentially made of hydrogen; this observation took some time to be accepted by the whole astronomical community. Many of the remarkable features correlated with nuclear properties (see figure 1), giving clues to the origin of the different elements, were already present in Russell's results.

Is the solar composition unique or does it vary with time and/or from one solar layer to another? The only obvious expected variation is in the central layers, where, because of thermonuclear reactions, the hydrogen content decreases, leading to an increase in the helium content (see NUCLEAR REACTION RATES). More subtle and unforeseen variations in the chemical composition have, however, been observed recently. During the solar lifetime, the convective zone reservoir which fills the outer solar layers is slowly enriched in hydrogen from the radiative zone below it, whereas it slowly loses, through its lower boundary layer, about 10% of all the heavier elements (see SOLAR INTERIOR: CONVECTION ZONE). In the outer atmospheric layers, and particularly in the very heterogeneous solar CORONA, the observed composition is variable in different types of solar matter and a fractionation occurs in the low chromospheric layers: elements of low first ionization potential ($\leq$10 eV) show abundances larger than in the photosphere.

In spite of these subtle variations, it is possible to define a standard solar chemical composition, essentially derived from spectroscopic analysis of the solar photosphere.

Interest of solar abundances

The chemical composition of the Sun is obviously a key set of data for modeling the interior as well as the atmosphere of the Sun.

The Sun, being the best-known star, has always been considered as the typical star, the reference to which the abundance analyses of other stars are compared.

The standard abundance distribution as shown in figure 1 is also the basic set of data that have to be reproduced by nucleosynthesis theories; it also plays a key role in modeling the chemical evolution of galaxies and of the universe.

The Sun is also unique because chemical composition data can be measured in other objects of the solar system such as the Earth, Moon, planets, comets, meteorites. These data are very important for modeling the evolution of the solar system (see SOLAR SYSTEM: FORMATION).

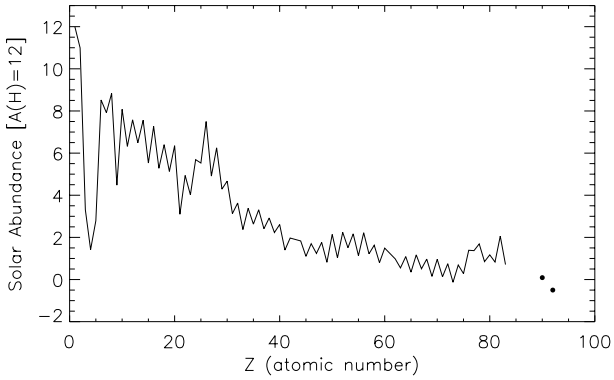


Figure 1. Logarithmic distribution of the abundance of the elements relative to hydrogen as a function of the atomic number, Z.

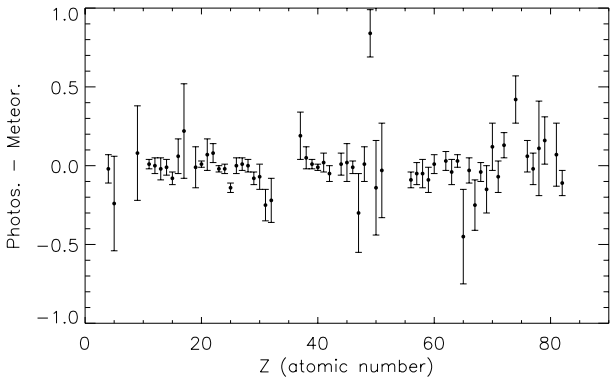


Figure 2. Difference between solar and meteoritic abundances of elements, in the usual logarithmic scale, as a function of Z. Error bars represent the uncertainty of the solar abundance determinations. The point representing Li falls outside of this figure (−2.21).

Sources of solar abundances

Solar abundances can be derived by very different techniques and for very different types of solar matter, from the interior to the outermost coronal layers. Using spectroscopy in a very large wavelength range, we can derive the chemical composition of the photosphere, CHROMOSPHERE and CORONA and also of SUNSPOTS. Particle collection techniques from space allow us to measure the chemical composition of the SOLAR WIND and of SOLAR WIND ENERGETIC PARTICLES, emitted when the Sun is active. We can also obtain information on solar flares from GAMMA-RAY ASTRONOMY. Finally, calibration of theoretical solar models and direct inversion of solar oscillation data allow us to derive the solar abundance of helium. Note also that lunar soils contain a record of the past chemical history of the Sun.

For different reasons and because of the variations in chemical composition, all of these sources of abundance results do not have the same weight. The solar photosphere is actually the layer from which we have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Element abundances in the solar photosphere and in meteorites.

Element	Photosphere[a]	Meteorites	Photosphere–meteorite difference	Element	Photosphere[a]	Meteorites	Photosphere–meteorite difference
1 H	12.00	–	–	42 Mo	1.92 ± 0.05	1.97 ± 0.02	-0.05
2 He	$[10.93 \pm 0.004]$	–	–	44 Ru	1.84 ± 0.07	1.83 ± 0.04	$+0.01$
3 Li	1.10 ± 0.10	3.31 ± 0.04	-2.21	45 Rh	1.12 ± 0.12	1.10 ± 0.04	$+0.02$
4 Be	1.40 ± 0.09	1.42 ± 0.04	-0.02	46 Pd	1.69 ± 0.04	1.70 ± 0.04	-0.01
5 B	(2.55 ± 0.30)	2.79 ± 0.05	-0.24	47 Ag	(0.94 ± 0.25)	1.24 ± 0.04	-0.30
6 C	8.52 ± 0.06	–	–	48 Cd	1.77 ± 0.11	1.76 ± 0.04	$+0.01$
7 N	7.92 ± 0.06	–	–	49 In	(1.66 ± 0.15)	0.82 ± 0.04	$+0.84$
8 O	8.83 ± 0.06	–	–	50 Sn	2.0 ± 0.3	2.14 ± 0.04	-0.14
9 F	$[4.56 \pm 0.30]$	4.48 ± 0.06	$+0.08$	51 Sb	1.0 ± 0.3	1.03 ± 0.07	-0.03
10 Ne	$[8.08 \pm 0.06]$	–	–	52 Te	–	2.24 ± 0.04	–
11 Na	6.33 ± 0.03	6.32 ± 0.02	$+0.01$	53 I	–	1.51 ± 0.08	–
12 Mg	7.58 ± 0.05	7.58 ± 0.01	$+0.00$	54 Xe	–	2.17 ± 0.08	–
13 Al	6.47 ± 0.07	6.49 ± 0.01	-0.02	55 Cs	–	1.13 ± 0.02	–
14 Si	7.55 ± 0.05	7.56 ± 0.01	-0.01	56 Ba	2.13 ± 0.05	2.22 ± 0.02	-0.09
15 P	5.45 ± 0.04	5.56 ± 0.06	-0.11	57 La	1.17 ± 0.07	1.22 ± 0.02	-0.05
16 S	7.33 ± 0.11	7.20 ± 0.06	$+0.13$	58 Ce	1.58 ± 0.09	1.63 ± 0.02	-0.05
17 Cl	$[5.50 \pm 0.30]$	5.28 ± 0.06	$+0.22$	59 Pr	0.71 ± 0.08	0.80 ± 0.02	-0.09
18 Ar	$[6.40 \pm 0.06]$	–	–	60 Nd	1.50 ± 0.06	1.49 ± 0.02	$+0.01$
19 K	5.12 ± 0.13	5.13 ± 0.02	-0.01	62 Sm	1.01 ± 0.06	0.98 ± 0.02	$+0.03$
20 Ca	6.36 ± 0.02	6.35 ± 0.01	$+0.01$	63 Eu	0.51 ± 0.08	0.55 ± 0.02	-0.04
21 Sc	3.17 ± 0.10	3.10 ± 0.01	$+0.07$	64 Gd	1.12 ± 0.04	1.09 ± 0.02	$+0.03$
22 Ti	5.02 ± 0.06	4.94 ± 0.02	$+0.08$	65 Tb	(-0.1 ± 0.3)	0.35 ± 0.02	-0.45
23 V	4.00 ± 0.02	4.02 ± 0.02	-0.02	66 Dy	1.14 ± 0.08	1.17 ± 0.02	-0.03
24 Cr	5.67 ± 0.03	5.69 ± 0.01	-0.02	67 Ho	(0.26 ± 0.16)	0.51 ± 0.02	-0.25
25 Mn	5.39 ± 0.03	5.53 ± 0.01	-0.14	68 Er	0.93 ± 0.06	0.97 ± 0.02	-0.04
26 Fe	7.50 ± 0.05	7.50 ± 0.01	$+0.00$	69 Tm	(0.00 ± 0.15)	0.15 ± 0.02	-0.15
27 Co	4.92 ± 0.04	4.91 ± 0.01	$+0.01$	70 Yb	1.08 ± 0.15	0.96 ± 0.02	$+0.12$
28 Ni	6.25 ± 0.04	6.25 ± 0.01	$+0.00$	71 Lu	0.06 ± 0.10	0.13 ± 0.02	-0.07
29 Cu	4.21 ± 0.04	4.29 ± 0.04	-0.08	72 Hf	0.88 ± 0.08	0.75 ± 0.02	$+0.13$
30 Zn	4.60 ± 0.08	4.67 ± 0.04	-0.07	73 Ta	–	-0.13 ± 0.02	–
31 Ga	2.88 ± 0.10	3.13 ± 0.02	-0.25	74 W	(1.11 ± 0.15)	0.69 ± 0.03	$+0.42$
32 Ge	3.41 ± 0.14	3.63 ± 0.04	-0.22	75 Re	–	0.28 ± 0.03	–
33 As	–	2.37 ± 0.02	–	76 Os	1.45 ± 0.10	1.39 ± 0.02	$+0.06$
34 Se	–	3.41 ± 0.03	–	77 Ir	1.35 ± 0.10	1.37 ± 0.02	-0.02
35 Br	–	2.63 ± 0.04	–	78 Pt	1.8 ± 0.3	1.69 ± 0.04	$+0.11$
36 Kr	–	3.31 ± 0.08	–	79 Au	(1.01 ± 0.15)	0.85 ± 0.04	$+0.16$
37 Rb	2.60 ± 0.15	2.41 ± 0.02	$+0.19$	80 Hg	–	1.13 ± 0.08	–
38 Sr	2.97 ± 0.07	2.92 ± 0.02	$+0.05$	81 Tl	(0.9 ± 0.2)	0.83 ± 0.04	$+0.07$
39 Y	2.24 ± 0.03	2.23 ± 0.02	$+0.01$	82 Pb	1.95 ± 0.08	2.06 ± 0.04	-0.11
40 Zr	2.60 ± 0.02	2.61 ± 0.02	-0.01	83 Bi	–	0.71 ± 0.04	–
41 Nb	1.42 ± 0.06	1.40 ± 0.02	$+0.02$	90 Th	–	0.09 ± 0.02	–
				92 U	(<-0.47)	-0.50 ± 0.04	–

Abundances are given in the logarithmic scale usually adopted by astronomers $A_{\rm el} = \log N_{\rm el}/N_{\rm H} + 12.0$, where $N_{\rm el}$ is the abundance by number. By definition the abundance of hydrogen is exactly $A_{\rm H} = 12.0$.
[a] Values between square brackets are not derived from the photosphere, but from sunspots, solar corona and solar wind particles. Values between parentheses are less accurate results.

the largest number of data. Actually, 65 elements, out of 83 stable elements, are present in the photospheric spectrum; other sources of solar abundances only concern a limited number of elements. The photosphere, just above the convection zone, is a well-mixed region whereas the outer solar layers show a very heterogeneous and changing structure. Furthermore, it is also the layer that was studied quite a long time before the other layers for an obvious reason: the solar photospheric spectrum has been recorded for quite a long time with high resolution and high signal-to-noise ratio. For all these reasons, photospheric abundances are adopted as a reference for all the other solar (and stellar) data.

Solar photospheric abundances

The basic data for deriving photospheric abundances are from the photospheric spectrum. Solar photospheric spectra with very high resolution and very high signal-to-noise ratio, obtained from the ground and from space, are now available for quite a large wavelength range, from the ultraviolet to the far infrared.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The strength of a line in the ABSORPTION SPECTRUM is directly related to the abundance of the element producing the line. The exact relation between these two quantities can be obtained if the physical conditions and physical processes of the layers where the line is formed are known. Empirical modelling of the photosphere has now reached a rather high degree of accuracy. Physical processes are also well known. Actually, in the photosphere, collisional processes dominate, which allows us to simplify the treatment. From the line strength, now measured with high accuracy, or directly from the fit of the predicted and observed spectra, the so-called spectral synthesis technique, it is possible to derive rather accurate values for the product of the abundance of the element producing the line, relative to the hydrogen abundance, times the transition probability of the line (see also SOLAR SPECTROSCOPY AND DIAGNOSTICS). The ratio $N_{\rm el}/N_{\rm H}$ is a natural measurement of abundance because the continuum is essentially produced by the opacity due to hydrogen species whereas the absorption line is the result of an additional opacity caused locally by a given element.

One of the largest sources of uncertainty of solar photospheric abundances comes from the inaccuracies of the transition probabilities. These data really play a key role in solar spectroscopy. Most of the progress in our knowledge of the solar photospheric chemical composition during the last decades has been mainly, if not uniquely, due to the use of more accurate transition probabilities. Actually, we believe that the dispersion of solar photospheric abundance results reflects the internal accuracy of the transition probabilities used to derive the abundances. Also, unfortunately, too few very accurate transition probabilities are available for the atomic and molecular lines present in the photospheric spectrum.

In table 1, we give the latest data for the solar photospheric abundances relative to hydrogen. The values are given in the logarithmic scale usually adopted by astronomers $A_{\rm el} = \log N_{\rm el}/N_{\rm H} + 12.0$, where $N_{\rm el}$ is the abundance by number. It is to be noted that such a scale was originally adopted in order to avoid negative numbers for the least abundant elements. It is still used nowadays even if two values are now negative numbers i.e. their abundance is less than one million millionth that of hydrogen—rather rare!

As may be seen from table 1, the photospheric abundances of a few elements cannot be measured in the photosphere, because these elements do not show lines in the photospheric spectrum, the most important of them being helium. This does not at all mean that these elements are absent in the Sun. Actually, this is purely for basic spectroscopic reasons: under the physical conditions in the photosphere, with a typical temperature around 5000 K, no line of these elements (neutral or once-ionized species) falls within the wavelength range covered by the photospheric spectrum.

Comparison with meteorites

Although most of the matter of the solar system is contained in the Sun, very useful information concerning the chemical composition of the initial solar nebula and its evolution can be obtained from the observations of other bodies of the solar system. Early terrestrial data very rapidly showed that the Earth suffered from heavy chemical fractionation since its formation but results of measurements made on METEORITES, during the first half of this century, soon showed many regularities, as seen in figure 1.

Most meteorites suffer from differentiation due to melting after condensation in the parent body. However, the so-called CI carbonaceous chondrites, has preserved the bulk composition of the stellar nebula from which they condensed, except for the highly volatile elements (H, C, N, O and the rare gases) which partly escaped.

Meteoritic chemical composition can now be measured with very high accuracy. When compared with photospheric results, there were past discrepancies, for several elements, but the discrepancies have gone away as the photospheric values have become more accurate. Now, as seen from figure 2, the agreement between the photosphere and CI chondrites is remarkably good. The uncertainties essentially come from the photospheric results which cannot be obtained to better than 10–20%. The few discrepant points can be accounted for by the larger uncertainty of some photospheric results, especially for elements which only show very few blended lines in the photospheric spectrum.

Table 1 allows us to define a 'standard abundance distribution' (also sometimes called 'cosmic abundance' or 'local galactic abundance') by combining photospheric results, for elements partly lost by meteorites, with meteoritic results for the other elements.

A few special cases

For different reasons that will become clear in the following, a few elements deserve particular comments.

Helium

In 1868, a new element was discovered in the solar spectrum obtained during an eclipse. The name of the Sun was given to this unknown element, namely helium. It was only discovered on Earth in 1895.

Despite its name and its very high abundance, about one-tenth of hydrogen, the most abundant element, He is not present in the photospheric spectrum, for purely spectroscopic reasons, as explained previously, and it is largely absent from the meteorites. Solar wind and solar energetic particles show a very variable and rather low value when compared with values observed in hot stars and in the interstellar medium from H II regions around us. Coronal values derived from spectroscopy still have too large uncertainties, $N_{\rm He}/N_{\rm H} \approx 8 \pm 1.5\%$. Among the giant planets, values are still too uncertain and fractionation might have been at work in some of them.

Progress in our knowledge of the solar helium content has recently come from standard solar models as well as non-standard models and also from the direct inversion of HELIOSEISMIC OBSERVATIONS. The calibration of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solar theoretical models, i.e. the fit of a theoretical model to the presently observed characteristics of the Sun (age, mass, diameter, luminosity), leads to the abundance of helium in the protosolar cloud from which the Sun formed ($N_{He}/N_H = 9.8 \pm 0.4\%$). Inversion of helioseismic data leads to a very accurate, but smaller, value, $N_{He}/N_H = 8.5 \pm 0.07\%$; this is the present solar abundance of helium in the outer convection zone. The difference between these two values is now explained by slow element diffusion at the base of the convection zone during the solar lifetime, as already mentioned in the introduction.

Astronomers often use the abundances by mass of hydrogen (X), helium (Y) and all the other elements together, the metallicity (Z), in a scale where $X + Y + Z = 1$. The numbers cited above become $Y = 0.275 \pm 0.01$ or $A_{He} = 10.99 \pm 0.02$ in our logarithmic scale, at the birth of the Sun, and $Y = 0.248 \pm 0.02$ or $A_{He} = 10.93 \pm 0.004$, as the present value in the solar outer layers.

Lithium, beryllium, boron

The strikingly low abundances of these light elements are a direct consequence of their low nuclear binding energy; at temperatures of the order of a few million kelvins, these nuclei are destroyed by proton collisions.

In the Sun, meteoritic and photospheric abundances of Be and B agree within the error bars, whereas Li is 160 times less abundant in the (present) photosphere than in the meteorites, i.e. at birth of the Sun. Lithium is the most fragile of these three elements and its solar destruction rate could be explained if the solar convection zone reached a little deeper layers than now deduced from helioseismology. So additional mixing, just below the convection zone, has to be added in order to explain the observed solar lithium depletion.

Carbon, nitrogen, oxygen

These elements which have largely escaped from meteorites are key elements. Because of their large abundances, they are the main contributors to the metallicity (O 47%, C 17%, N 5%) and they are also very important contributors to the opacity which plays a crucial role in modeling the solar interior.

Although coronal measurements of the CNO abundances are certainly very helpful, the solar abundances of these three elements rely heavily on photospheric analyses.

In spite of the fact that the solar photospheric abundances of these elements can be derived from a large number of indicators, atoms as well as diatomic molecules made of C, N, O and H, the uncertainties of the solar results are uncomfortably too large, partly because of the lack of very accurate transition probabilities.

Neon, argon

These two noble gases do not appear in the solar photospheric spectrum, for the reasons already described, and are largely lost by meteorites. It is hoped that values obtained for the outermost solar layers, from coronal spectra, solar wind and solar energetic particle measurements as well as gamma ray spectroscopy of solar flares, do not suffer from the fractionation mentioned in the introduction and can be used as representative of the present solar photosphere.

Isotopic abundances

Because isotopic shifts in atomic lines are extremely small, it is impossible to obtain isotopic ratios for most of the elements. For a few elements, however, such as C, O and Mg, which appear in diatomic molecules, we can derive estimates of the solar isotopic composition, with rather large uncertainties. Isotopic data can be obtained from solar wind measurements for He, Ne and Mg.

Isotopic ratios for all the elements can be measured with high accuracies for terrestrial material. They agree with the results obtained from meteorites although 'isotopic anomalies' are observed which are, however, confined to a very small fraction of the meteorites. Thus the method is to take solar abundances for the elements together with isotopic ratios as derived on earth in order to obtain the solar abundances of individual nuclides.

Conclusions

With the solar chemical composition presented in table 1, the usual mass abundances are $X = 0.735$, $Y = 0.248$ and $Z = 0.017$ with $Z/X = 0.023$.

The solar abundances are in excellent agreement with the meteoritic abundances derived from CI carbonaceous chondrites. The effects of element migration at the bottom of the convection zone are, however, not observed. This is puzzling although not surprising. Actually, the uncertainties of the photospheric abundance results are still too large to allow such a faint effect to be detected.

Bibliography

Anders E and Grevesse N 1989 Abundances of the elements: meteoritic and solar *Geochim. Cosmochim. Acta* **53** 197–214

Frohlich C, Huber M C E, Solanki S and von Steiger R (ed) 1998 *Solar Composition and its Evolution—From Core to Corona, Space Science Series of ISSI* (Dordrecht: Kluwer) (also 1998 *Space Sci. Rev.* **85**)

Grevesse N and Sauval A J 1998 Standard solar composition *Solar Composition and its Evolution - From Core to Corona, Space Science Series of ISSI* ed C Frohlich, M C E Huber, S Solanki and R von Steiger (Dordrecht: Kluwer) pp 161–74 (also 1998 *Space Sci. Rev.* **85** 161)

Palme H and Beer H 1993 Abundances of the elements in the solar system *Landolt-Börnstein, Group VI, Astronomy and Astrophysics* Extension and supplement to vol 2, subvol a, ed H H Voigt (Berlin: Springer) pp 196–221

Russell H N 1929 *Astrophys. J.* **70** 11–82

N Grevesse and A J Sauval

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Active Regions

Solar active regions form where the tops of loops of magnetic flux, shaped like the Greek letter omega (Ω), emerge into the solar atmosphere where they can be seen. These Ω loops are thought to be generated by the solar dynamo, located deep within the Sun (see DYNAMOS: SOLAR AND STELLAR). The appearance of active regions depends greatly on the radiation used to image them, as well as the angle at which they are observed. When seen projected against the dark sky at the edge of the Sun (called the LIMB), the magnetic field of an active region is traced by glowing arches of gas confined within CORONAL LOOPS (figure 1). When projected against the solar disk the appearance of active regions is much different. In hydrogen Hα, at low telescopic resolution, one sees bright SOLAR CHROMOSPHERIC PLAGE (figure 2). At higher resolution, the relatively cool (chromospheric) loops of the active region appear (figure 7). In x-rays, the much hotter (coronal) loops of the active region are seen (figure 3).

All of these active region loops, whatever their temperature, trace out magnetic field lines. Plage appears at their footpoints, where the loops pass through the chromosphere and photosphere. At one footpoint of each loop the magnetic field comes out of the Sun (north polarity), and at the other footpoint it goes back into the Sun (south polarity). Active regions are formed of many such bipolar pairs of footpoints connected by coronal loops, as if a large magnet was buried beneath the visible solar surface at the site where the active region appears.

The high density at which magnetic loops are packed together at their photospheric footpoints is the distinctive feature that makes active regions different from the surrounding *quiet* atmosphere. The effects of the magnetic field on the temperature and density structure of an active region depend on its strength, the relative fraction of the active region it fills (compared with non-magnetic regions) and its topology, i.e. how each point is connected to another. In chromospheric plage—the bright areas that can be seen in figure 2 both scattered around the quiet Sun (two examples shown) and within the active region itself—magnetic loops are packed 10–100 times more densely than in the quiet Sun. In SUNSPOTS, loops are packed so tightly that there is no significant amount of unmagnetized gas, which inhibits the flow of heat by convective motions and makes sunspots cooler and darker than their surroundings. Finally, irregular dark strands of gas called solar FILAMENTS form where the magnetic field lines dip to form valleys in the corona that can hold ionized gas (which cannot easily move across field lines) in clouds above the photosphere. Several such filaments can be seen in figure 2, both inside the active region and outside it, in the quiet solar atmosphere. When filaments are seen in profile at the solar limb, they are called SOLAR PROMINENCES.

Active regions come and go, and their lifetime depends to a large extent on their size. Active regions have no preferred size; they form a continuous distribution, with many more small ones than large ones. The smallest

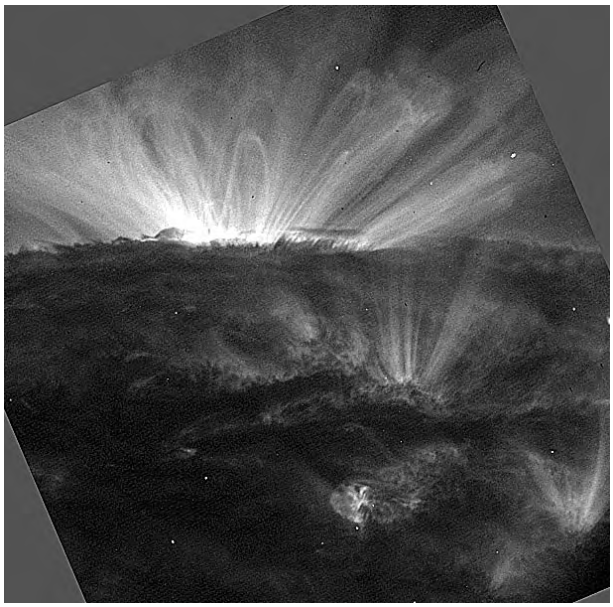


Figure 1. A small solar active region near the edge of the solar disk, seen in emission from a highly ionized Fe line emitted by gas at about 1.5×10^6 K. Photo from the NASA TRACE satellite courtesy Lockheed Martin.

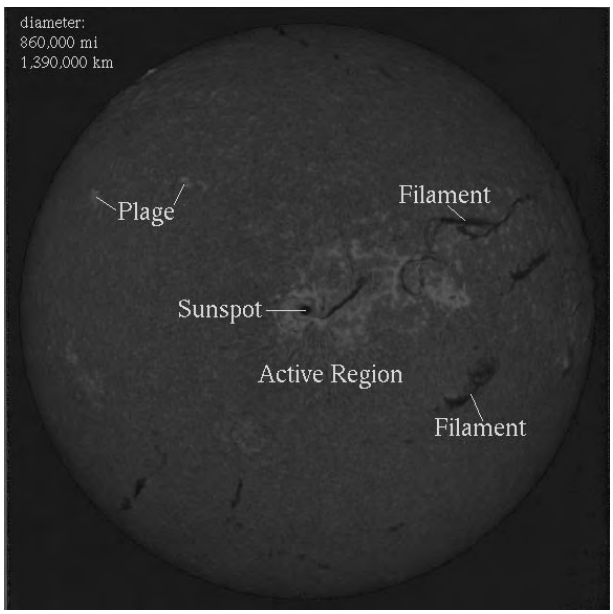


Figure 2. An image of the Sun's chromosphere in the neutral hydrogen Hα line emitted at about 10^4 K, showing a large active region near disk center, containing sunspots, plage and filaments (labeled elsewhere on the disk). Photo courtesy the National Solar Observatory, which is supported by the US National Science Foundation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

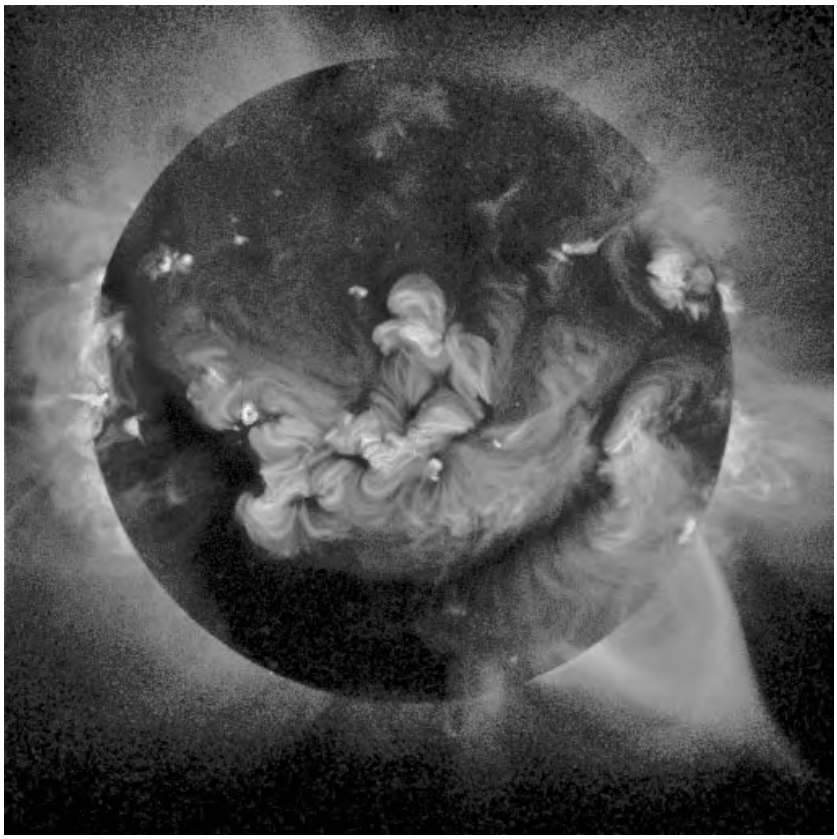


Figure 3. An x-ray image of the Sun's corona on 25 January 1992, a few years after the 1989 solar cycle maximum. Such x-rays are emitted by gas at about (3–5) $\times$ 10^6 K. At mid-latitudes a huge solar activity complex, made of many solar active regions, is present. The tiny x-ray bright points seen most easily in the northern (upper) hemisphere mark the sites of ephemeral solar active regions. The prominent cusp in the lower right signals a recent coronal mass ejection. Yohkoh image courtesy L Acton, Montana State University Bozeman.

ones that can be observed depend on telescope resolution, and presumably there are many still smaller ones that have not been resolved. In the photosphere the largest and most long-lived active regions have areas in the range 50–100 000 Mm^2, which is only 1–2% of the surface area of the Sun, but about 100–200 times greater than that of the Earth. The x-ray image shown in figure 3 illustrates the extreme range of sizes seen in the solar corona. The smallest active regions, and also the most short lived, are called ephemeral solar active regions, which are often only a few thousand km across. Several tiny x-ray bright points, which mark the sites of ephemeral solar active regions, can be seen the the northern hemisphere of the image in figure 3 (see SOLAR CORONA: X-RAY BRIGHT POINTS; SOLAR ACTIVE REGIONS: EPHEMERAL). The active regions in which phenomena such as SOLAR FLARES and sunspots occur typically fall in the size range from 10 000 to 100 000 km, and live for a month or two (one or two solar rotations). Many such active regions can be seen in figure 3. It happens that, at the time of this image, more were located in the southern hemisphere than in the northern, but that is not a general rule. At the extreme end of the distribution of sizes and lifetimes are SOLAR ACTIVITY COMPLEXES, which are made up of several active regions. The interconnected active regions at and below the center of the solar image in figure 3 collectively form such a complex of activity.

Active regions have several interesting properties, which vary either on the 22 yr SOLAR CYCLE or the 11 yr sunspot cycle. Very few active regions, or the sunspots they contain, are present at sunspot minimum—sometimes, none at all. In contrast, near sunspot maximum, it is not usual to find five to ten active regions on the hemisphere visible from Earth. Their emergence frequency varies by about a factor of 8, in phase with the sunspot cycle. On the other hand, the shape of the size distribution of regions does not change significantly from minimum to maximum.

The orientation of the magnetic polarity pair in active regions, with respect to the direction of solar rotation (from left to right, solar east to west, if north is up), follows what is called the Hale–Nicholson polarity law during the solar cycle. As can be seen in figure 4 one magnetic polarity tends to predominate at the leading (right) edge of all active regions in one hemisphere, whereas the opposite polarity predominates at the leading edge in the other

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

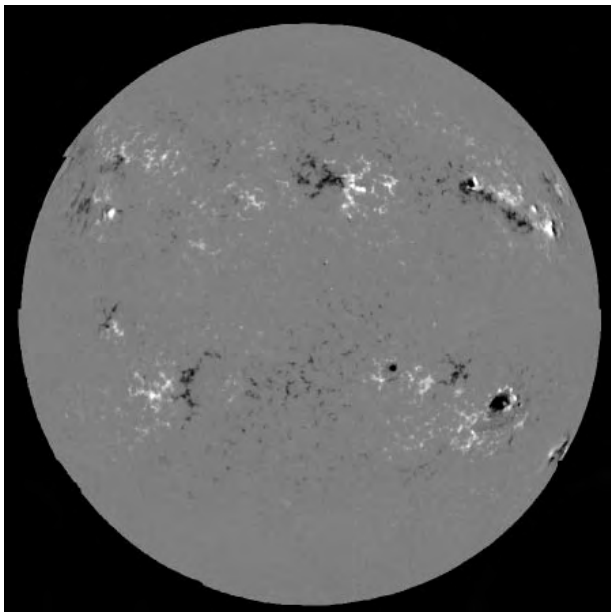


Figure 4. Solar magnetogram, 19 February 1999 at 15:17 UT. Photo courtesy the National Solar Observatory, Kitt Peak, which is supported by the US National Science Foundation.

hemisphere. The polarity that dominates in a given hemisphere reverses every 11 yr, repeating with the 22 yr period of the solar cycle. Interestingly, the Hale–Nicholson polarity law is only an approximation, and many regions are exceptions. These exceptions are attributed to kinks in Ω loops and buffeting of them by convective motions.

The solar latitude at which active regions emerge varies from about 60° at the start of each sunspot cycle to very close to the equator at its end almost 11 yr later. The regions that emerge at high latitudes early in the solar cycle tend to be very small—not large enough to contain sunspots, which tend not to be seen in regions above 30° or so. As the cycle goes on, the latitude of emergence proceeds toward the equator, and the size of the active regions tend to increase. This tendency for active regions to emerge at progressively lower latitudes as the sunspot cycle progresses is called Spoerer's law. There is also a tendency for the smaller regions to be spread out more in latitude of emergence, which may be attributed to greater buffeting by convection.

The spots on the leading edge of active regions tend to be closer to the equator than those that follow, so the axes of the active regions appear to be tilted with respect to the equator. Several clearly tilted regions appear in figure 4; in the north, white polarity leads and tends to be closer to the equator, and in the south, black polarity plays that role. When enough regions are studied, it becomes clear that this tilt depends on latitude, being only about 1–2° near the equator, but increasing to 5–8° at latitude 25° The flux in the leading part of active regions also tends to be concentrated in fewer larger spots than the following part (see figure 4, lower right quadrant, and figure 5). This

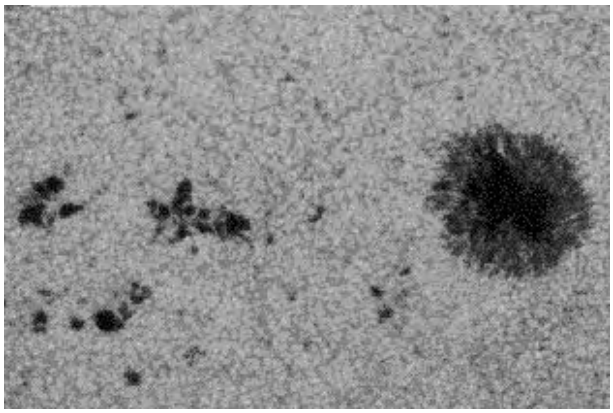


Figure 5. Sunspots are the most prominent feature of active regions in the photosphere, as seen in ordinary visible light. Photo courtesy the National Solar Observatory, Kitt Peak, which is supported by the US National Science Foundation.

morphology has been ascribed to the dynamics of gases in flux loops as they rise through the convection zone and expand. Gas tends to flow toward the following leg of the Ω loop, increasing its cross-section. This reduces the field strength in the following leg, making it more easily shredded into smaller strands by convection.

The rope-like magnetic field structures of active regions in the northern hemisphere tend to be twisted weakly in a left-handed sense, and those in the southern, right-handed. As a result, coronal loops in the south tend to be shaped like an S, and those in the north, the mirror image of an S. These properties, and some of the other properties described above, are thought to be determined by the forces felt by convection zone flux tubes (see SOLAR INTERIOR: CONVECTION ZONE FLUX TUBES). The fields of active regions are twisted by motions of the rising gas, which tend to rotate in opposite directions in the two hemispheres under the influence of the Coriolis force. As well, they are buffeted by strong turbulence. The same tendencies of tilt and asymmetry seen in active regions appear in emerging magnetic flux tubes (see SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES) that have just appeared in the atmosphere after their journey through the convection zone.

Active regions are sites of strong heating of the outermost solar atmosphere (corona). Hence, active regions are much brighter than their surroundings in coronal x-ray images (figure 3). Active regions are filled with coronal loops, which outline magnetic field lines along which the ionized coronal plasma is forced to move. In active regions, magnetic fields are very strong, ranging from roughly 1000 to 3000 G. These fields are about 10 000 times stronger than the 0.3 G field of the Earth at the equator. In a mature active region, the amount of magnetic flux (the product of the strength of the magnetic field and the area it covers) ranges from 10^{19} to 5×10^{22} Mx. These magnetic fields carry currents—as much as 10^{12} A. In comparison, a 100 W light bulb carries a current of about 1 A. Nevertheless, it is not understood just how the vast

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

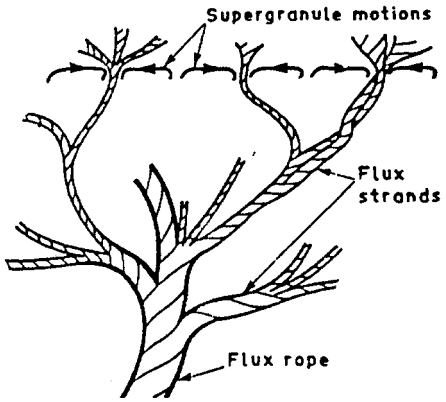


Figure 6. The tree-like structure of an active region thought to exist below the photosphere. Adapted from a figure by J H Piddington.

amounts of energy contained in these currents is converted into coronal heating, solar flares or coronal mass ejections (see SOLAR CORONAL MASS EJECTION).

Active regions—photosphere to corona

Active regions extend from below photosphere far into the corona. The physical conditions change dramatically over this range. In and below the photosphere, the gas is so dense that its turbulent motions push the magnetic fields around. Such motions serve to break up the magnetic fields into bits and pieces outside active regions, as well as to concentrate magnetid fields in individual dark sunspots and SUNSPOT PORES within them. Active regions are surrounded by brighter granular areas which are almost field free (figure 5). As one goes higher in the solar atmosphere, the density of the gas falls off more rapidly than the strength of the magnetic field. In and above the chromosphere, the magnetic field can no longer be forced to move by the flow of the gas, and it spreads out to fill all space. Hence, the appearance of active regions is much different in the photosphere, where kinetic and thermal terms dominate the energy budget, from in the chromosphere and corona, where the magnetic field is more important.

At some depth below the visible surface convection zone flux ropes are shredded into smaller strands, as envisioned in figure 6. Motions within granules, supergranules and perhaps even larger structures (see SOLAR PHOTOSPHERE: GRANULATION; SUPERGRANULATION) push these strands around to form the sunspots and pores of an active region, such as those shown in figure 5. Our present knowledge of the subphotospheric structure of active regions is fuzzy, since the techniques of helioseismology that tell us about this structure do not have high enough spatial resolution. The main reason for believing that the structure has such form is the appearance of the active region in the photosphere shown in figure 5. This appearance typically evolves from many tiny pores early

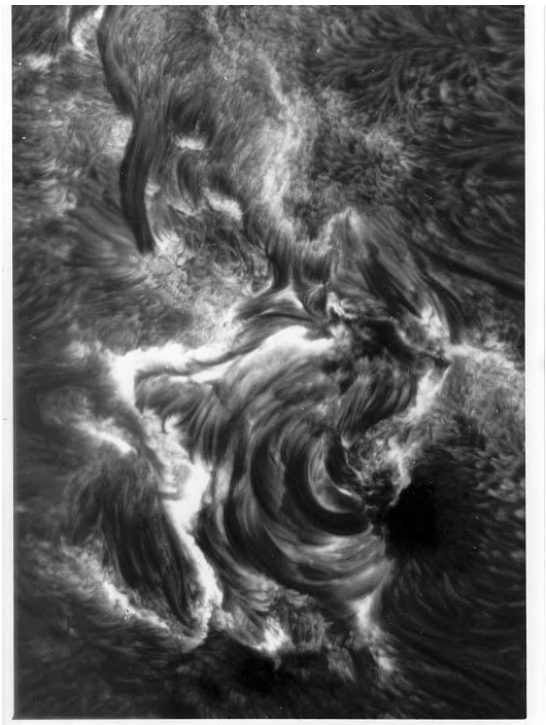


Figure 7. The graceful structures of an active region in the chromosphere, consisting of loops and swirls connecting magnetic features in the underlying photosphere. Sunspots, as they appear in the chromosphere, can be seen in the lower part of the active region. Photo courtesy R Muller, Pic du Midi Observatory.

in the region's life to large spots (in the leading polarity) when it is fully developed. This is understood as the consequence of the gradual rising and emergence of the tree-like structure of figure 6.

In the chromosphere, imaged in the hydrogen Hα line in figure 7, the structure of the active region consists primarily of magnetic loops, which are the continuation of the flux strands into the upper atmosphere from below. Such loops show up in a hydrogen Hα image only where the atmospheric conditions are just right for hydrogen gas to be partially ionized. The temperature in this part of an active region is typically about 10^4 K, and extends a few thousand km above the photosphere. The important aspect of such Hα images is that the structures seen in them outline the magnetic field, since the ionized gas must follow magnetic field lines. Generally, Hα images are more useful for revealing the connections of the magnetic field than for measuring the temperature or density of the gas—brightness in an Hα image has little do do with these thermodynamic variables. An exception is at the footpoints of hot coronal loops, visible in x-rays and not in Hα. Their presence in the corona, overlying the chromosphere, is reflected in the very bright areas that wind through the lower left-hand part of figure 7. In these regions the chromosphere is unusually hot and dense

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

because the overlying arcade of coronal loops that has its footpoints there is also hot and dense. Such active region coronal loops, those with temperatures of several times 10^6 K, appear prominently bright in each of the many active regions in figure 3. Other loops, somewhat cooler, appear in images made in light from spectral lines radiated by ions formed at lower temperatures around 10^6 K, such as that of figure 1.

Evolution of active regions

Active regions are born when strands of magnetic flux emerge into the visible atmosphere from the solar interior. Sometimes flux continues to emerge for as long as 5–7 days, and the active region grows to large size; at the other extreme, flux stops emerging so quickly that observers never even notice the small active region before it vanishes. The shape and brightness of the region depend on both the rate and location of emergence of this new flux and the manner in which it reconnects with the pre-existing magnetic fields. After a lifetime that varies up to a few months, the active region is gradually dispersed by convective motions differential rotation, meridional motion, and magnetic reconnection.

The first indication of an emerging active region is the appearance of bright chromospheric plage in Hα and tiny sunspot pores in visible light. Pores of opposite magnetic polarity are soon seen to be connected by chromospheric loops called arch filaments, which appear as dark linear features in line-center Hα images. The curvature of these features, which are the projections of three-dimensional arches, depends on the angle with which they are viewed. An arcade of several such arch filaments, with bright chromospheric plage at their ends, can be seen in figure 8 (top). These arch filaments are the portions of strands of magnetic flux that are just now emerging into the visible atmosphere, and the pores seen in figure 8 (middle) are the cross-section of these strands where they pass through the photosphere. The image in figure 8 (middle) is made in the blue wing of the Hα line. The chromosphere at this wavelength is transparent when no line-of-sight motions are present, allowing the pores in the underlying photosphere to be seen clearly. However, at the tops of arch filaments, absorption can be seen, as a consequence of Doppler blueshift from their rising motion.

As flux continues to emerge in an active region, pores and small spots tend to coalesce into larger spots, implying that the thinnest flux strands in figure 6 are now above the photosphere, and larger strands of the lower part of the flux rope are now emerging. As long as flux emergence continues, arch filaments can be seen. Often new flux emerges amidst old flux, so only part of the region shows obvious arch filaments, as in the lower part of the region shown in figure 8 (top, middle). The distance separating the opposite-polarity spots of the active region increases. This growth of the active region shown in figure 8 can clearly be seen by comparing the bottom panel of the figure with the two above it, taken a day earlier. Bright loops of comparable size (not shown) can often be seen in x-rays,

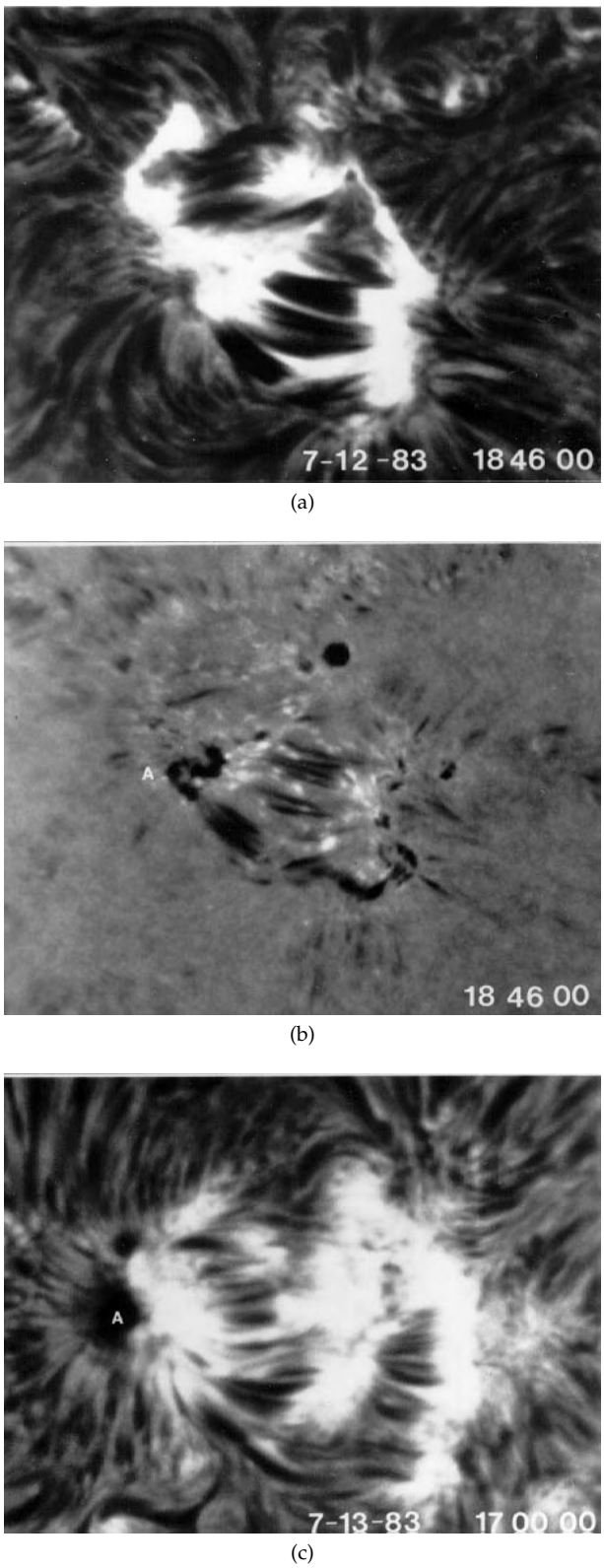


Figure 8. An emerging flux region: (top) early stages, Hα line center; (middle) early stages, Hα blue wing; (bottom) 1 day later, Hα line center. Courtesy H Zirin, Big Bear Solar Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

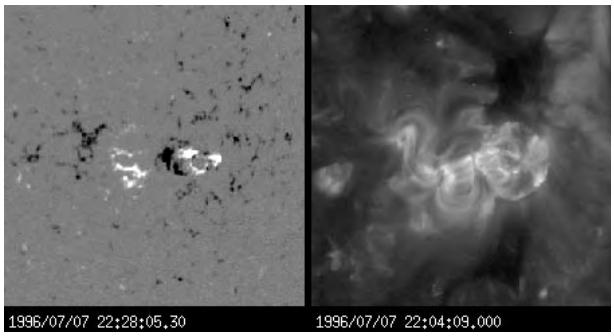


Figure 9. Co-aligned images of the photospheric magnetic field (left) and coronal loops (right) in an active region that has reconnected with its surroundings. Courtesy B Thompson, NASA/ESA Solar and Heliospheric Observatory.

overlying the chromospheric arch filament system seen in Hα.

The next phase of the development of an active region is the reconnection of the magnetic fields of the bipole with the ambient fields into which it has emerged. The structure that forms as a result of this reconnection is shown in figure 9, in which a small active region is seen in projection on the solar disk. To understand that structure, it is useful to compare the co-aligned photospheric magnetic field map (left) and projected coronal loop image (right). The polarities in the magnetic field map (toward and away from the viewer) are indicated by black and white; where the field is weak, the map is gray. Note that a strong and compact bipole is present just to the right of the center of the map—black on the left, white on the right. The coronal image shows that the black and white regions of the magnetogram are connected to one another, as one would expect from an emerging bipole. However, there are also prominent connections of the bipole to surrounding regions, forming a structure variously called a pinwheel, anemone or fountain. These loops are formed by reconnection of the bipole, after its emergence, with its surroundings.

The lifetime of active regions, the period for which they can be seen to be bipolar, has no unique or typical value. The most short-lived regions, called ephemeral active regions, and the most long-lived complexes, in which one active region may emerge within another, are discussed elsewhere. Ordinary active regions, as seen on these pages, live from a few days to a few months. The end of the lives is determined by the dispersal of their magnetic flux by convective motions, differential rotation, meridional flows and magnetic reconnection. Bits of magnetic flux break off from the edge of a decaying active region, like icebergs from a glacier. Such moving magnetic features often produce surges as they reconnect with surrounding flux of opposite polarity. Those bits of flux that do not reconnect nearby may be transported great distances. Typically an irregular diagonal streak of such magnetic debris may be seen poleward of decaying active regions. Evidence of such dispersed flux can be seen faintly above some of the northern-hemisphere regions, and below some of the southern-hemisphere regions, in figure 4. Ultimately all such flux reconnects and flows out in the solar wind; new flux in emerging active regions replaces it, and the Sun moves on to its next 22 yr cycle.

Bibliography

Harvey K L and Zwaan C 1993 Properties and emergence patterns of bipolar active regions *Sol. Phys.* **148** 85–118

Howard R F 1996 Solar active regions as diagnostics of subsurface conditions *Ann. Rev. Astron. Astrophys.* **34** 75–107

Lang K R 1995 *Sun, Earth and Sky* (Berlin: Springer)

Zirin H 1989 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

Richard C Canfield

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Active Regions: Ephemeral

Magnetic flux emerges through the surface of the Sun as bipolar regions with a wide range of sizes, such as shown in figure 1. Larger magnetic bipoles develop SUNSPOTS and are identified as ACTIVE REGIONS. Small bipolar regions, such as indicated in the lower panels of figure 1, are called ephemeral regions, a name given by H Dodson in 1953 because of their short lifetimes. Ephemeral regions have sizes typically less than 20 000 km along their long axis, lifetimes of a few hours, and no sunspots. Their numbers vary in phase with the SOLAR CYCLE. Unlike active regions, which are confined within a latitude band extending from 40°N to 40°S, ephemeral regions occur over the entire solar surface. The bipolar axes of ephemeral regions show all possible orientations, but there is a preference for a proper, low-inclination orientation as seen in larger active regions. During the declining phase of a cycle, ephemeral regions at high latitude (poleward of 30°) develop with a preferential orientation reversed from the active regions and ephemeral regions at lower latitudes. These regions are identified with the new cycle and are seen at least 3 years *before* sunspot minimum and 2–3 years *before* the first recognized sunspot regions of the new cycle.

The amount of magnetic flux that ephemeral regions bring to the surface in one day is equivalent to that in one or more major active regions. Their flux does not accumulate at the surface, however, as the magnetic poles of ephemeral regions quickly interact with nearby magnetic fields leading to the disappearance (also termed cancellation) of one or both of the poles of these ephemeral regions. This interaction results in the complete replacement of the network magnetic fields in the quiet Sun (although not the large-scale pattern of magnetic polarities) within about 40 h, perhaps playing a key role in the heating of the solar CORONA.

General properties of ephemeral regions

Ephemeral regions were first identified in Ca II K images by H Dodson in 1953. Our knowledge about most of the properties of ephemeral regions, however, comes from studies of high-spatial-resolution observations of the magnetic fields on the solar surface during the last two decades. Ephemeral regions are most easily identified in time-sequence magnetograms, where they first appear as two closely spaced positive and negative polarity elements (poles) that strengthen and separate with time. The average flux density in ephemeral regions increases with time, indicating that the magnetic flux in ephemeral regions emerges from below the surface.

Table 1 summarizes the range and average values of the total magnetic flux (i.e. the sum of the absolute values of the positive and negative flux), area, separation of the centroids of the two opposite polarity poles, and lifetime of ephemeral regions.

The lifetime of ephemeral regions, defined as the period both poles can be distinguished as separate elements, strongly depends on the proximity, strength and polarity of the surrounding magnetic elements. During their emergence, the separating poles of an ephemeral region often encounter and 'cancel' with magnetic elements of opposite polarity or merge with elements of the same polarity.

Table 1. Summary of ephemeral region properties.

Property	Average	Range
Total magnetic flux (10^{18} Mx)	9.3	0.1–67
Area (10^{18} cm^2)	0.46	0.1–75
Separation (km)	6910	2200–12 000
Lifetimes (h)	4.4	0.5–140

Orientations

The magnetic axis of an ephemeral region is defined by a line connecting the centroids of its two opposite polarity poles. The bipole orientation of the region is the angle of its magnetic axis with the equator. Ephemeral regions, although they have a wide range in the orientations of their bipolar axes, do show a preference to emerge with a proper, low-inclination orientation. This preference persists throughout a cycle.

The preferential orientation of ephemeral regions follows that of active regions. For the periods when there are two concurrent cycles, the orientations of ephemeral regions are consistent with the preferred active region orientations for those cycles, as discussed in more detail below. The latitude at which ephemeral regions change their preferential bipole orientation coincides with the boundaries separating active regions belonging to different cycles.

Frequency of emergence

The frequency with which new ephemeral regions emerge at the surface ranges from 0.02 to 1 $h^{-1}(10^{10}\ km^2)^{-1}$. These emergence frequencies suggest that thousands of ephemeral regions emerge over the entire surface of the Sun within a 24 h period. With an average flux of 9.3×10^{18} Mx, ephemeral regions can bring to the surface each day from 0.3 to 5×10^{22} Mx, the equivalent of one or more large active regions.

Time variation of ephemeral regions

The number of ephemeral regions present on the visible hemisphere of the Sun at any given time clearly varies nearly in phase with solar activity, as shown in figure 2. The minima in ephemeral region counts precede those defined by sunspots. For example, the minimum in the number of ephemeral regions in 1975 occured about one year prior to the sunspot minimum between cycles 20 and 21 in June 1976; between cycles 21 and 22, the ephemeral region minimum in early 1985 was about 1.5 years before the sunspot minimum in September 1986. By the time of the sunspot minima, the counts of ephemeral regions have increased 40% to 50% above their minimum levels of more

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

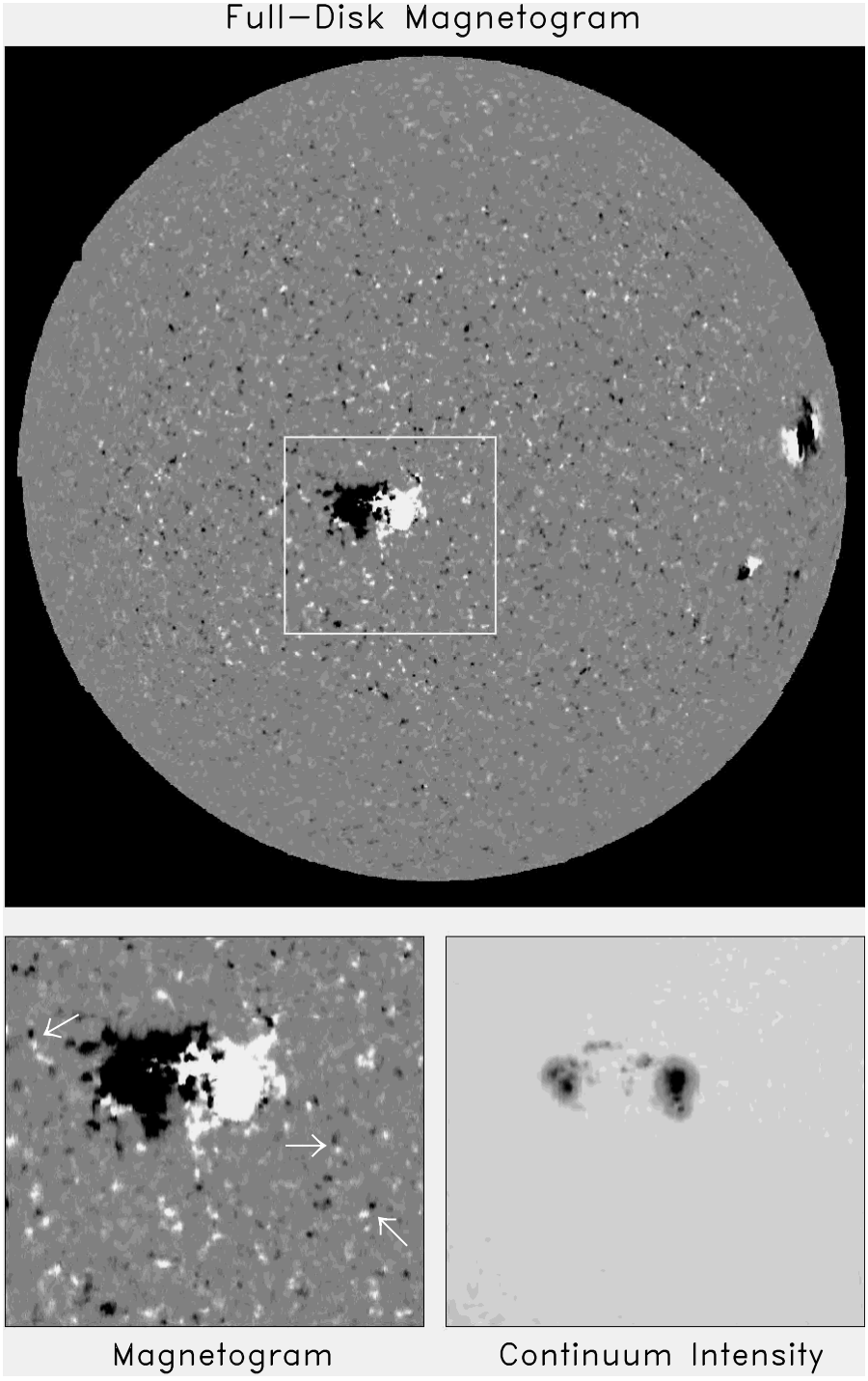


Figure 1. Full-disk magnetogram of the Sun on 25 November 1996 comparing a large bipolar active region and nearby ephemeral regions. The area outlined in white on the full-disk magnetogram is expanded in the lower panels to better show the size comparison and the association of sunspots with a large active region. The horizontal distance of the expanded panels is 3.6×10^5 km. Magnetic fields coming out of the Sun are shown as white and those going into the Sun are black. North is at the top and east to the left.

than a year earlier. In figure 2, the maximum phase for ephemeral regions is reached in early 1978, more than a year earlier than for other parameters of solar activity (e.g. sunspot number) and persists into early 1983. The counts of ephemeral regions also appear to vary on time scales of several rotations in some cases coincident with the pulses

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

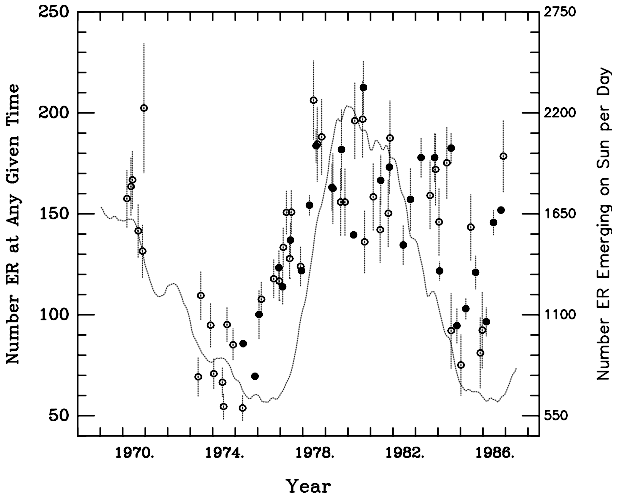


Figure 2. The cycle variation in the number of ephemeral regions observed at any given time (left scale). These numbers are converted to the number of regions per day (right scale) assuming an average lifetime of 4.4 h. The estimated errors, shown by the vertical bars, are derived from the square roots of the number of regions, normalized as per the data points. The smoothed sunspot number (light dotted curve) is shown for comparison.

of activity that are seen in other measures of solar activity, such as the 10.7 cm radio flux.

From the minimum to the maximum phase of cycle 21, the ephemeral region counts increase by a factor of 2.0 to 2.5. This is substantially smaller than the factor of 8 observed for numbers of larger active regions.

If it is assumed that the average lifetime of 4.4 h is valid for any time in the cycle, then the number of ephemeral regions emerging on the Sun per day increases from 870 during their minimum to about 1900 during maximum phase. This range is consistent with the number of ephemeral regions emerging per day determined from their frequency of emergence.

Spatial distribution of ephemeral regions

Ephemeral regions can be observed at any solar latitude. Figure 3 shows the variation in their distribution both in latitude and in relation to larger active regions during a solar cycle. The distribution in latitude of ephemeral regions roughly corresponds to a function that describes how easily ephemeral regions are observed as a function of their distance from the disk center. This correspondence indicates that outside the latitudes where active regions occur, the distribution of ephemeral regions is nearly uniform. There are, however, significant departures that are related to active region belts. The low-latitude peaks in the distribution for 1975 to 1976, for example, coincide with the activity of outgoing cycle 20. At higher latitudes during 1975 to 1978 are peaks in the distributions related to the increase of cycle-21 active regions; these peaks increase in amplitude and shift systematically toward the equator

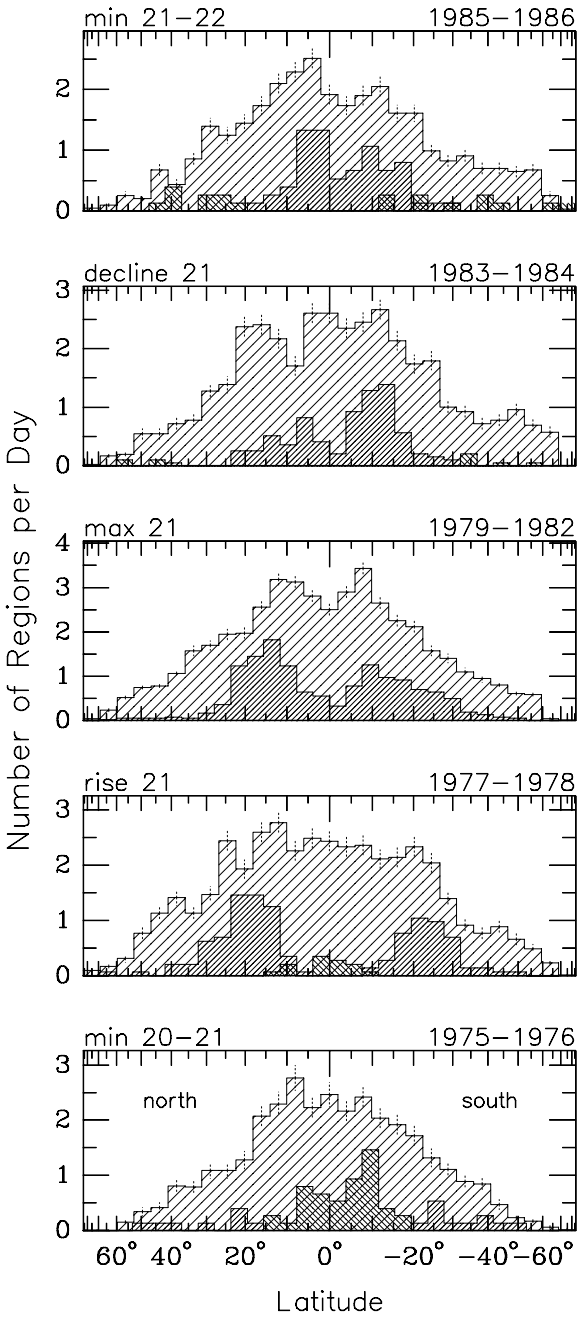


Figure 3. Distribution of ephemeral regions in heliographic latitude during five phases in the rise and decline of cycle 21. The lightly hatched histograms are in equal surface area bins in sine latitude. The more closely hatched histograms, rescaled to fit the plot, are the latitude distributions of active regions for cycles 20, 21 and 22. Cycle-21 active regions have their hatching in the same direction as the ephemeral regions and cycle 20 and 22 regions are perpendicular.

over the cycle, following the emergence pattern of active regions during a cycle.

Two to three years prior to the first appearance of sunspot regions belonging to a new cycle, a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

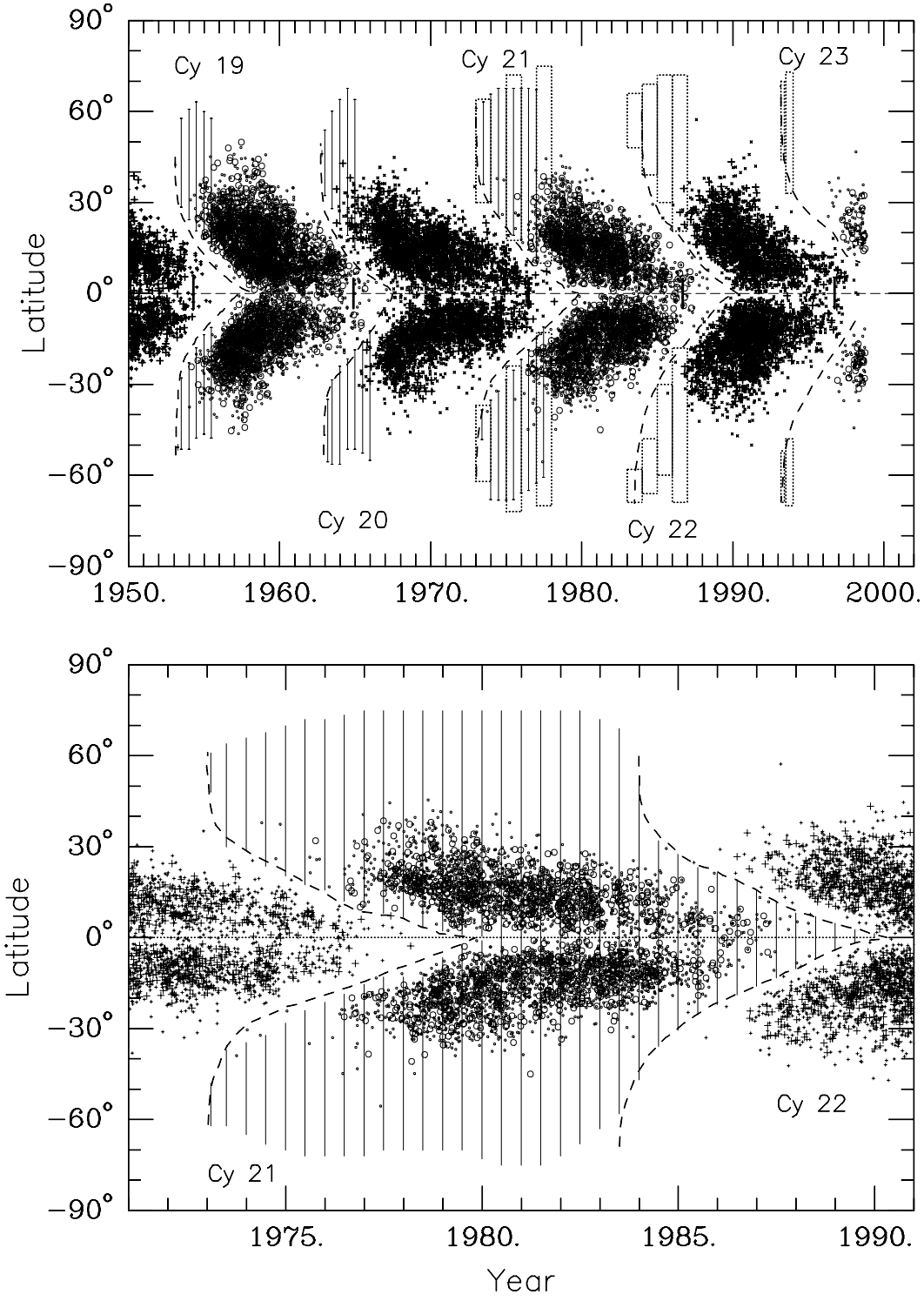


Figure 4. Superposition of the butterfly diagram of sunspot regions with the latitude extent of ephemeral regions having the same preferential orientation. Top panel: Ephemeral regions identified as Ca II K regions are indicated by vertical bars and as magnetic bipoles by boxes. Sunspot regions in even-numbered cycles are shown as ◦ and in odd-number cycles as ×. Bottom panel: An overlay of the latitudinal extent of ephemeral regions (vertical bars) relative to sunspot regions during cycle 21. The heavy dashed curves indicate the boundaries between activity of two concurrent cycles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

persistent and expanding pattern with preferential reverse orientation ephemeral regions appears at latitudes in both hemispheres poleward of 30°. As shown in both panels of figure 4, this pattern persists and merges with the pattern of new-cycle sunspot regions. As this pattern develops, ephemeral regions at lower latitudes still have a clear tendency for bipole orientations that are proper for the older cycle. The preferentially reverse-orientation, high-latitude ephemeral regions are interpreted to be the first emergence of bipolar regions associated with the onset of a new cycle. The early appearance of new cycle activity in the form of small bipoles (top panel of figure 4) was observed in cycles 19 to 21 as small Ca II K plage regions; these small plage regions are spatially and temporally consistent with the pattern of small reverse-orientation magnetic bipoles in cycles 21 to 23. The high-latitude patterns of ephemeral regions suggest that they are a common feature of solar activity, and they mark the onset of a new solar cycle substantially earlier than would be indicated by sunspot regions. The latitudes and timing of ephemeral regions extends the sunspot butterfly pattern to an earlier time and higher latitudes.

Are ephemeral regions active regions?

There are many similarities between the properties of ephemeral regions and larger active regions strongly suggesting that a substantial fraction of these small bipoles are in fact the small end of a wide size distribution of active regions. The similar preferential orientations, temporal variations and co-spatial peaks in the latitude distribution of ephemeral regions and the larger active regions indicate that these small-scale regions are related to the large-scale emergence pattern of solar activity. Their wider latitude distribution and lower amplitude variation from minimum to maximum suggests that the population of ephemeral regions is a composite of activity related small bipoles and what appears to be a nearly uniform component.

Importance of ephemeral regions to the global magnetic field pattern

Recent observations by SOHO/MDI (Solar and Heliospheric Observatory/Michelson Doppler Imager) indicate that ephemeral regions play a significant role in maintaining the magnetic fields on the solar surface. As ephemeral regions emerge and evolve, their magnetic poles separate and encounter existing magnetic flux. During these encounters, the magnetic fields of most of the ephemeral regions will 'cancel' with opposite polarity magnetic networks, leaving the pole having the same polarity as the network somewhat displaced from the cancellation site. As a result of this process, the magnetic field of an ephemeral region has effectively replaced some of the existing magnetic flux. It is found that ephemeral regions emerge in sufficient numbers to completely replace the magnetic fields in the quiet Sun regions within about 40 h, leaving the large-scale pattern of magnetic polarities unchanged. This 'cancellation' process and the resulting reconfiguration of the magnetic fields (*magnetic reconnection*) can result in localized heating in the chromosphere and corona (see SOLAR CORONA: X-RAY BRIGHT POINTS), and therefore, may be an important source for CORONAL HEATING.

Bibliography

Harvey K L and Martin S F 1973 Ephemeral active regions *Solar Phys.* **32** 389–402

Harvey K L, Harvey J W and Martin S F 1975 Ephemeral active regions in 1970 and 1973 *Solar Phys.* **40** 87–102

Martin S F and Harvey K L 1979 Ephemeral active regions during solar minimum *Solar Phys.* **64** 93–108

Schrijver C J, Title A M, Harvey K L, Sheeley N R Jr, Wang J-M, van den Oord G H J, Shine R A, Tarbell T D and Hurlburt N E 1998 Large-scale coronal heating by the dynamic, small-scale magnetic field of the Sun *Nature* **394** 152–3

Schrijver C J and Zwaan C 1999 *Solar and Stellar Magnetic Activity* (Cambridge: Cambridge University Press)

Karen Harvey

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Activity Complexes

SOLAR ACTIVE REGIONS tend to cluster in space and time by emerging in or beside an existing active region or at the site of a previous one. Whereas an individual bipolar sunspot region (see SUNSPOT CLASSIFICATION) typically lasts from a few days to a few weeks—rarely as long as a few solar rotations—a cluster may survive for over a year, kept alive by fresh bipolar regions repeatedly emerging within the small area which delimits the cluster. This tendency has been recognized since the early 18th century but it is only in recent decades that its implications have had an effect on DYNAMO THEORY and on studies of interacting magnetic fields at the solar surface (see MAGNETIC RECONNECTION).

Identifying clusters of sunspot regions

Various names have been invented to describe this phenomenon. Compilers of the *Greenwich Photoheliographic Results* in the 19th century recorded clusters as 'revivals', not mere 'recurrences', of sunspot regions on successive 27 day solar rotations. Recently introduced terms include 'Sonnenfleckenherde' (U Becker), 'complex of activity' (V Bumba and R Howard) and 'sunspot nests' (M Castenmiller, C Zwaan and E van der Zalm). The last of these names was chosen to imply closeness for the emerging bipolar regions and persistence of their rejuvenation far beyond any single region's lifetime. Below, 'sunspot nest' is used chiefly for discussing statistical properties of the clusters and their bipolar sunspot contents, while 'activity complex' is reserved for issues concerning the distribution of magnetic flux in and around clusters.

Statistical methods

Sunspot nests can be identified from tabulations of sunspot positions such as the *Greenwich Photoheliographic Results* by defining selection rules and then applying objective statistical methods to distinguish between real clusters and those produced by chance. Mathematical techniques at different levels of sophistication have been applied to selected epochs drawn from this data base. Several studies, reviewed by C Zwaan and K Harvey, confirm the non-random nature of the space–time distribution of sunspot groups. Basic properties found for nests include compactness (areas much smaller than a medium–large sunspot group, provided that due allowance is made for differential rotation (see SUN: BASIC PROPERTIES) of the nests), ubiquity (at least 30% of all sunspot groups), a steady rate of rotation about the Sun (but unique to each nest) and lifetimes up to about 15 solar rotations (allowing for gaps up to 2.5 rotations). If substantially longer interruptions are allowed, then some nests may even last a few years. There is also a strong tendency for nests to cluster yet again in still larger nests: more than a third of the compact nests are components of composite nests. The components can be quite close in latitude ($<2.5^\circ$) and longitude ($<25^\circ$), although pairs of nests sometimes differ in longitude by as much as 60° to form 'superclusters'.

Other aspects of the clustering phenomenon are revealed from magnetic data of the line-of-sight component of the photospheric magnetic flux, $B_\parallel$ (see SOLAR MAGNETIC FIELD: POLARIMETRY). The daily full-disk magnetograms (figure 1(*a*)) produced at the Vacuum Telescope of the National Solar Observatory (NSO) at Kitt Peak, Arizona (see SOLAR TELESCOPES AND INSTRUMENTS: GROUND), are especially useful in this regard because of their high resolution (1" pixels), high dynamic range (3 orders of magnitude in the magnetic flux density) and high continuity in the flow of data. White patches in figure 1 denote positive (outward-directed) magnetic fields, black patches denote negative polarity. This bipolar signal has an overriding advantage for defining the boundaries of active regions: it makes their evolution and their tendency to cluster much easier to follow from the critical moment of their birth than do, for example, white-light photographs of sunspots.

Thus the 'great complex of activity' enclosed by a box in figure 1(*a*) can be readily decomposed during this particular disk passsage into five individual bipolar regions (A–E), of which three are nested inside one another (figure 1(*b*)). This example also serves to raise a cautionary note about the significance of the word 'complex'. These multiply nested emergences reveal a complicated pattern of polarity reversals. Yet it took a long time after the last emergence to develop *magnetic complexity* of the kind associated with intense current sheets and MHD instabilities resulting in large flares (see SOLAR FLARE MAGNETIC FIELDS). The evidence to date does not favor clustering as the sole process for building up the extreme magnetic complexity of spots classified as a δ configuration. Large δ spots are relatively rare and shorter lived than normal spots of the same size; they seldom last more than one passage across the solar disk. This means that their complexity has, on average, to be produced very quickly, at or even before their emergence.

K Harvey surveyed nearly 1000 newly emerged bipolar active regions larger than 2.5 square heliographic degrees (373 Mm^2) chosen from 29 intervals of above-average continuity in the daily flow of magnetograms at NSO between 1975 and 1986. She found that sunspot nests were rejuvenated by the emergence of new bipolar regions at intervals of, typically, 4–5 days. The number of regions that emerge in nests is comparable with emergences elsewhere in the activity belts: from 40% to possibly 55% of all active regions emerge inside nests. However, the rates of emergence, defined as the number of emergences per day per unit available surface area, differ drastically: new regions are at least 22 times more likely to emerge within existing sunspot regions than elsewhere.

Pictorial methods

A convincing illustration of the clustering phenomenon is made with a stackplot (figure 2), a method first adopted for the NSO synoptic magnetic maps by V Gaizauskas and co-workers. This figure assembles data, separately for each hemisphere, for 100 Carrington rotations (SOLAR ACTIVITY: LONG-TERM RECORDS) during solar cycle 21. It was

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

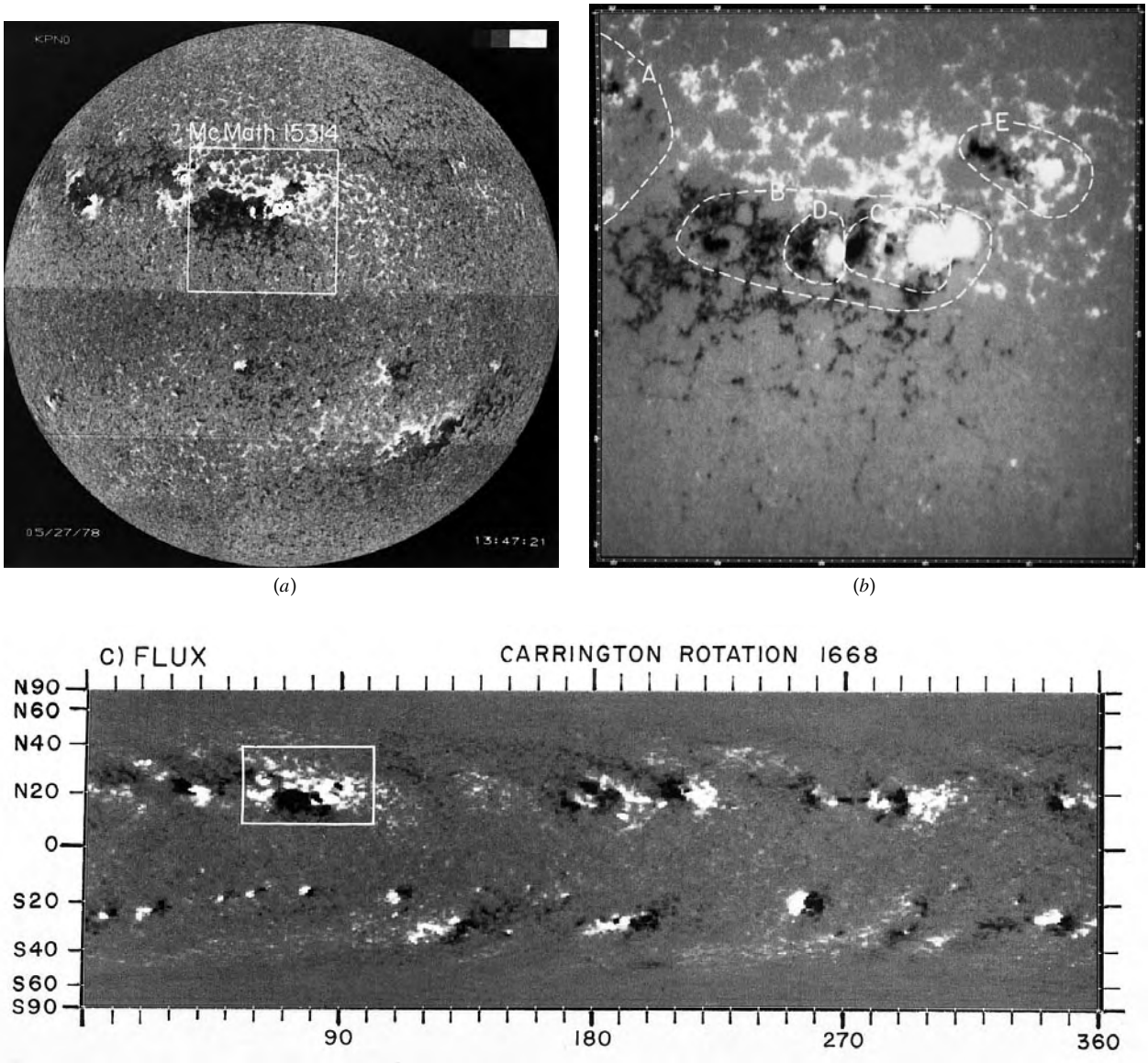


Figure 1. NSO (Kitt Peak) magnetograms and synoptic maps (courtesy of J W and K L Harvey): (*a*) full disk on 27 May 1978 with the 'great complex of activity' boxed in; (*b*) a close-up view of the boxed area ($8' \times 8'$) and with individual bipolar active regions outlined with dashed curves; (*c*) a synoptic map of the magnetic flux density for Carrington rotation 1668; the horizontal scale is the Carrington longitude increasing westward (right), while the vertical scale is heliographic latitude; the rotation begins on the extreme right on 7 May 1978 and ends at the extreme right on 2 June 1978; the white box encloses the same portion of the great complex as panel (*a*) of this figure.

constructed by stacking 30° wide latitudinal strips, one per Carrington rotation, extracted from successive synoptic maps of $B_{\parallel}$ (e.g. figure 1(*c*)). Longitude (L) increases from left (east) to right and extends for another one-third of a rotation to provide continuity for clusters formed near $L = 0°$; time (t) advances downward. In the ~7.5 yr period sampled here the activity belts migrate equatorward from the latitude of 40° used as the maximum of the strips at the beginning of the sample. The central latitude of each 30° wide strip was therefore shifted according to Spörer's law (SOLAR CYCLE) before the data were extracted from the full synoptic maps. The maps are uncompensated for the differential rotation of the Sun; their frame of reference rotates rigidly at the Carrington synodic period of 27.275 days.

In the L–t plane note the linear tracks of different lengths, spacings and slopes which are sometimes double, often diverging, sometimes converging. These tracks locate activity complexes; they share many properties in common with sunspot nests. Thus a nest rotating around

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

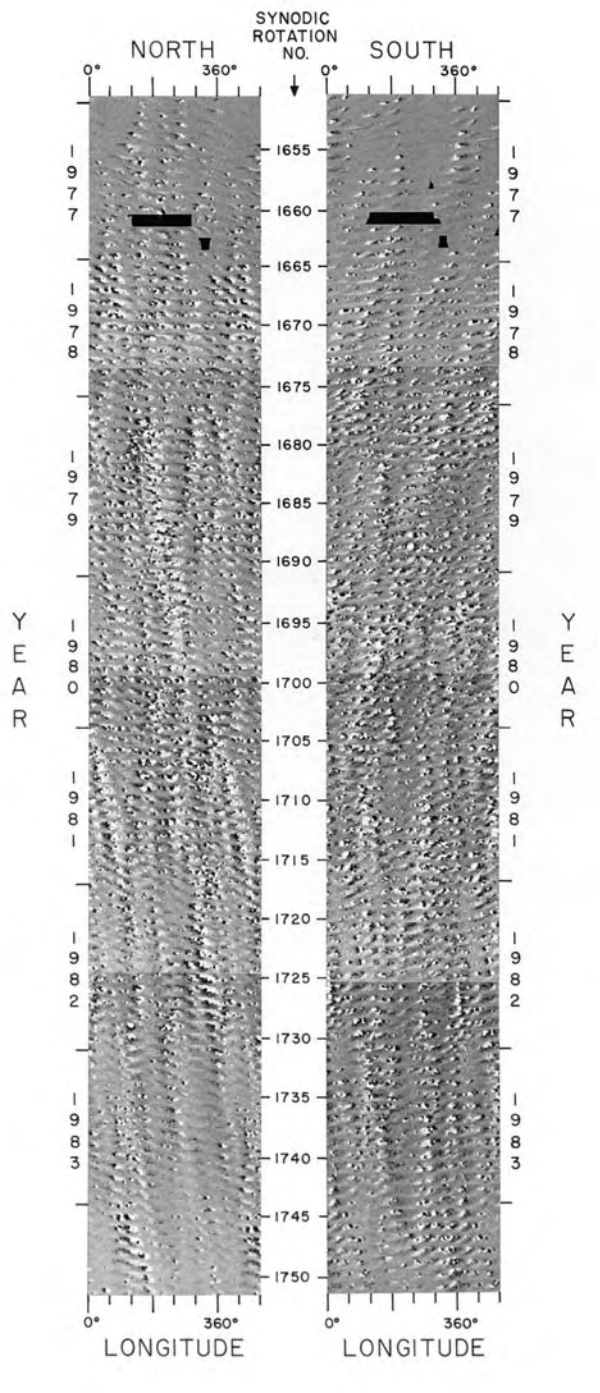


Figure 2. Active belts in both hemispheres of the Sun for 100 Carrington rotations during solar cycle 21. Construction, courtesy of J W Harvey, is described in the text.

the Sun at the Carrington rate produces a vertical track. Nests rotating faster than the Carrington rate produce tracks sloping down to the west (right); nests lagging behind the Carrington rate produce tracks sloping down to the east (left).

The pattern of tracks changes with the phase of the solar cycle—widely spaced in longitude during the onset (1977–1978) and the decay (1982–1983), but more closely packed (1979, N hemisphere) or even chaotic (1979–1980, S hemisphere) during the maximum phase. Even during highly active periods there can be longitudinal zones over 90° wide which are relatively free of activity (centered near 0°, N hemisphere, early in 1980, and near 190°, S hemisphere, in mid-1981). Some complexes are fully active for an entire year (1982, N hemisphere, near $L = 330°$); others fade out after a brief period of recurrent activity only to revive a few rotations later for further lively rotations (1982, N hemisphere, near $L = 120°$).

The existence of so many cohesive tracks enables a number of important inferences to be drawn about activity complexes. They usually appear and end quickly—within 1 solar rotation. The linearity of most tracks means that activity complexes are unaffected by differential rotation over many rotations; only when the injection of fresh flux stops are the weak remnants differentially sheared. Magnetic flux does not accumulate in an activity complex. The 'great complex' depicted in figures 1(*a*) and 1(*b*) had at least 24 additional major bipolar regions emerge within it, for a total of 29, over a period of 6 solar rotations, yet the total flux in the complex remained near a steady level of 10^{23} Mx. During its active lifetime a complex loses as much flux as it gains. The greater part of the flux disappears locally, within the complex; only a minor portion is carried poleward by meridional circulation (see SOLAR PHOTOSPHERE). At times a wavelike pattern of eight regularly spaced bands of alternating polarity is noticeable around the entire Sun.

Very little drift in latitude is evident in complexes or in the tightly bounded sunspot nests. In particular, there is no meridional spreading of magnetic flux across the solar equator during the active lifetime of an activity complex. Hence the term 'active solar longitude', a popular one employed in solar–terrestrial studies, is open to misinterpretation and should be avoided.

Interpretation

The flux densities and spacing of activity complexes have important quantitative consequences for theoretical dynamo models of stars (DYNAMOS: SOLAR AND STELLAR). E N Parker has pointed out that the gaps in longitude between activity complexes are so wide as to limit severely the reconnection between them. So small a portion of the emerging flux is able to reconnect and escape that the surface of the Sun resembles an impenetrable barrier rather than an open surface. The localized disappearance of the surface magnetic flux suggests an active dynamical control by subsurface convection.

The arrival of intense magnetic fields in bursts at a few localized sites has been modeled by Parker as a thermal relaxation oscillator. He conceives an intense azimuthal magnetic field (>3 kG) pressed down into the lower convective zone (SOLAR INTERIOR: CONVECTION

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ZONE) by its thermal shadow. The accumulation of heat beneath the field causes the gas heated below to penetrate intermittently through the field, sending thermal plumes of gas with entrained fields to the surface at irregular intervals ∼1 week.

D Wentzel and P Seiden reproduced the quasi-regular distribution of active regions in clusters by using percolation theory. Their model bundles complicated MHD and turbulent processes into two dimensionless parameters: one to express the probability for the release and rise of one flux tube to stimulate the subsequent release and rise of its neighbors; the other to measure the lifetime of flux once it arrives at the surface. Wentzel and Seiden argue that if thermal plumes were the only cause of rising flux tubes, then the resulting surface patterns would not exhibit any of the correlations apparent in active regions. They suggest that flux tubes are released instead by magnetic dissipation of turbulently distorted magnetic fields.

Further progress in elucidating the phenomenon of sunspot clusters awaits helioseismic tomography of subsurface layers of the Sun (see HELIOSEISMIC OBSERVATIONS) with instruments on board the space satellite SOHO (Solar and Heliospheric Observatory) and at the ground-based observatories of GONG (GLOBAL OSCILLATION NETWORK GROUP).

Bibliography

Gaizauskas V, Harvey K L, Harvey J W and Zwaan C 1983 Large-scale patterns formed by solar active regions during the ascending phase of cycle 21 *Astrophys. J.* **265** 1056–65

Harvey-Angle K L 1993 Magnetic bipoles on the Sun *PhD Thesis* Astronomical Institute, Utrecht

Parker E N 1987 The dynamical oscillation and propulsion of magnetic fields in the convective zone of a star: I. General considerations *Astrophys. J.* **312** 868–79

Wentzel D G and Seiden P E 1992 Solar active regions as a percolation phenomenon *Astrophys. J.* **390** 280–9

Zwaan C and Harvey K L 1994 Patterns in the solar magnetic field *Solar Magnetic Fields* ed M Schüssler and W Schmidt (Cambridge: Cambridge University Press) pp 27–48

Victor Gaizauskas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Activity Indices

Solar activity is driven by the temporally and spatially varying distribution of magnetic flux in the photosphere, chromosphere and corona. It covers a range of phenomena at all levels in the solar atmosphere and time-scales ranging from seconds and minutes (SOLAR FLARES and SOLAR CORONAL MASS EJECTIONS), through months (the evolution of ACTIVE REGIONS and SOLAR ACTIVITY COMPLEXES), to the 11 (or 22) yr solar activity cycle. Being able to express aspects of that activity in terms of single indices is useful for studying the Sun's long-term behavior, comparing the Sun with other stars or investigating its role as a driver for various space and terrestrial phenomena. Lang (1995) gives a useful overview of the Sun, its activity and its connection with the Earth. Zirin (1988) provides a more focused discussion on solar astrophysics. Some indices, such as sunspot number, require imaging the solar disk, and, until we can similarly map other stars, are confined to solar applications. Others, such as the Ca II H and K indices or the 10.7 cm solar (radio) flux, are measurements of emission integrated over the Sun's visible disk and are therefore measurable in other stars.

The longest continuous record of solar activity is provided by sunspot counts (*sunspot number*). In 1946 measurements began of another index, the 10.7 cm solar flux. This is more consistent and objective than sunspot number, and so may be more applicable to studies covering the last few decades. More recently, other indices of activity have become available. In this article we will discuss indices of solar activity and the relationships between them. A single index cannot fully describe the level of solar activity. Although the indices discussed here are those which are most widely used, more specialized studies might involve using other manifestations of solar activity as indices. Therefore this discussion should not be regarded as complete.

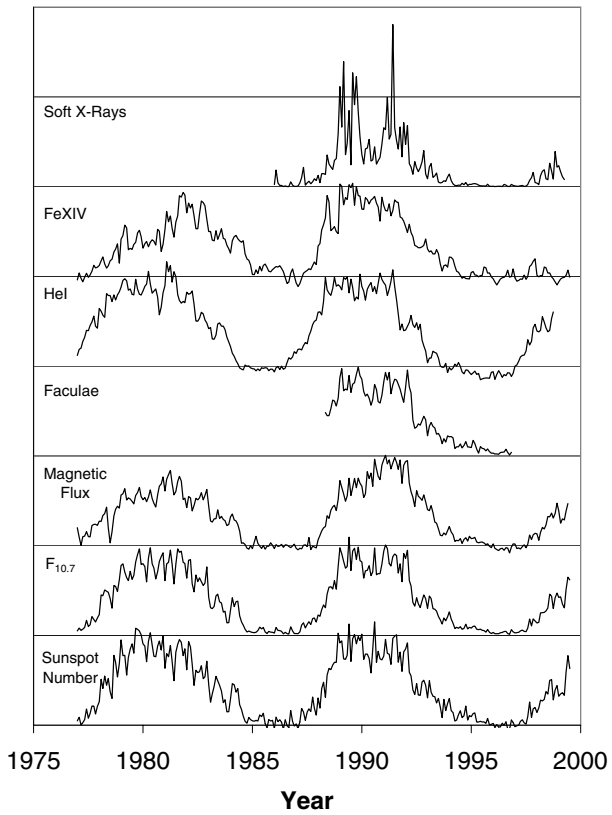


Figure 1. Monthly averages of various indices of solar activity plotted against time for the period 1975–99. Sunspot number, magnetic flux and facular area are indices that are photospheric in origin. The 10.7 cm solar flux is coronal, but is highly correlated with sunspot number, and is therefore plotted adjacent to it. The He I plot is a chromospheric index, and the upper plots (Fe XIV and soft x-rays) describe activity in the corona.

Activity indices

An index of solar activity is a quantity intended to describe some aspect of activity for the Sun as a whole. Although the Sun is highly complex, and we are far from understanding all the processes taking place in the centers where activity takes place, or even the small-scale magnetic and plasma structures within them, similarities and large-scale patterns are present in almost all aspects of solar activity. This is because all manifestations of solar activity are driven by the same 'engine': the temporally and spatially varying distribution of magnetic flux in the SOLAR PHOTOSPHERE, CHROMOSPHERE and CORONA. In principle, determinations of magnetic flux would provide the best general-purpose single index of solar activity. However, this cannot be measured directly. Instead we use as indices of solar activity quantities which are driven or strongly modulated by magnetic fields. By appropriate choices of index we can describe activity at different levels in the solar atmosphere. Figure 1 shows the cyclic variability of several activity indices over the period 1975–99. Sunspot number is a photospheric phenomenon. The 10.7 cm solar flux is coronal, but correlates very strongly with sunspot number. Most active regions contain FACULAE, which are visible in white light, and stand out more clearly with increasing angular distance from disk center. The total facular area is related the total magnetic flux and is therefore a useful index of solar activity. The temperatures (of the order of 10^4 K) and densities of plasma concentrations supported in chromospheric magnetic fields give rise to some useful spectral lines. The lines most widely used as activity indices are the H and K lines of Ca II and Mg II and the 1083 nm line from He I. In the corona, temperatures are of the order of 10^6 K and densities are much lower than they are in the chromosphere. Useful indices of activity are provided by emission lines such as Fe XIV, the 530.2 nm coronal green line and soft x-rays, which are produced by thermal bremsstrahlung (free–free) emission from concentrations of hot plasma. All these indices are strongly modulated by the ebb and flow of magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

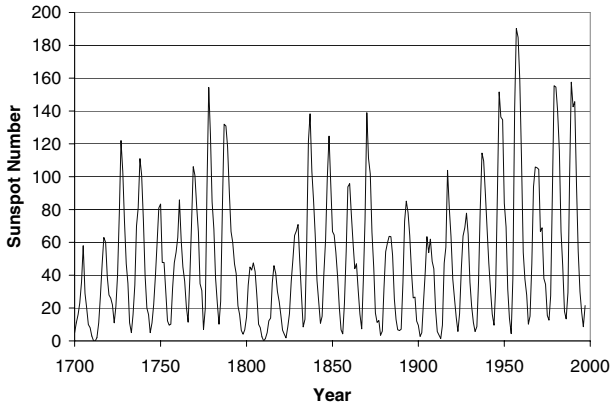


Figure 2. Monthly means of sunspot number between 1700 and present. This is the longest, consistent record of solar activity available. Note the variations in the amplitudes of the maxima and minima from cycle to cycle, and that the cycles vary in length.

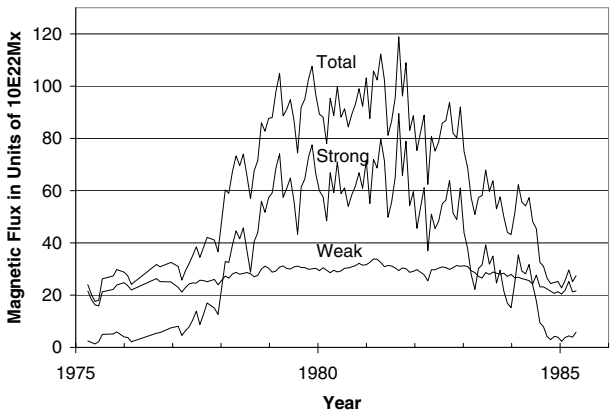


Figure 3. Monthly means of the photospheric magnetic flux over a solar cycle. The upper plot shows the total flux contained in all flux elements, the middle curve the total magnetic flux contained in *strong-field elements*, with average magnetic field strengths stronger than 25 G, and the lower curve, which shows little variation over the solar cycle, the total magnetic flux in *weak-field elements*, where the average magnetic field strengths are less than 25 G ($\text{total}_i = \text{strong}_i + \text{weak}_i$).

flux over the solar activity cycle and remarkably highly correlated, which emphasizes the strong role of magnetic fields in all observable manifestations of solar activity.

Photosphere

SUNSPOTS have been observed since the earliest days of astronomy, probably for more than three millennia. In the Western world they were rediscovered in the 17th century by Galileo and Scheiner. Systematic observations of sunspot number were well under way by the beginning of the 18th century. Schwabe (1789–1875) was the first to identify the 11 yr periodicity in sunspot number. However, these data, available through various sources, were put on a more systematic basis by Wolf (1816–93), who defined the partly empirical 'Zurich sunspot number', given by $R_z = K(10g + f)$, where R_z is the Zurich sunspot number, g is the number of sunspot groups, which are always assumed to contain 10 members, and f the number of single spots. The parameter K, known as the *personal reduction coefficient*, is a scaling factor taking into account the instrumentation, site quality and observer skills and practices for each source of data. In practice, K ranges between about 0.55 and unity. Ideally, there should be no need for parameters specific to individual observatories or observers, i.e. $K = 1$, which is the value Wolf used for his own observations. However, in order to relate modern data obtained using more sensitive instrumentation to older data, lower values of K need to be used. For more information on the history and the scientific background of the sunspot number index, see Zirin (1988), chapter 10, and Wilson (1994), chapters 1, 2 *et seq.*

Despite its rather empirical definition, the Zurich sunspot number (usually simply referred to as *sunspot number*) is a useful index of solar activity. It is monitored at several observatories and yields a continuous record of solar activity dating back at least to the early 18th century (see SOLAR ACTIVITY: LONG-TERM RECORDS). Such timescales are needed for studies of trends in solar activity and in terrestrial CLIMATE studies. Figure 2 shows a plot of monthly means of the sunspot number dating back to 1700. This is the longest, most consistent record of solar activity observations available. It has been possible to extend the sunspot number record back to at least AD 1600. Total sunspot *area* would probably provide a better indicator of total magnetic flux, as would total plage or facular areas. However, these data are not as available as sunspot number. Zirin (1988) reports that, over many years of observations, the solar community has established that a usable estimate of total sunspot area can be obtained by multiplying the sunspot number by 16.7.

Measurements of magnetic field strength at the photosphere are now made routinely by determining the degree of Zeeman splitting of selected photospheric spectral lines. Magnetograms showing the distribution of magnetic flux over the solar disk have been made regularly for roughly two solar activity cycles. Harvey (1994) shows that magnetic elements in the photosphere fall into two broad classes: *weak-field elements*, having average magnetic field strengths smaller than 25 G and *strong-field elements*, where the average magnetic field strength exceeds that value. Figure 3 shows a plot of the total magnetic flux over the solar sphere plotted over a solar cycle, together with the totals of magnetic flux strong-field elements and weak-field elements.

The variation of total magnetic flux over the solar activity cycle shows clearly the ebb and flow of activity and the formation and decay of active regions and activity complexes (these entities are described in Gaizauskas *et al* 1983). However, this occurs almost entirely in elements with magnetic fields stronger than 25 G. When the total

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

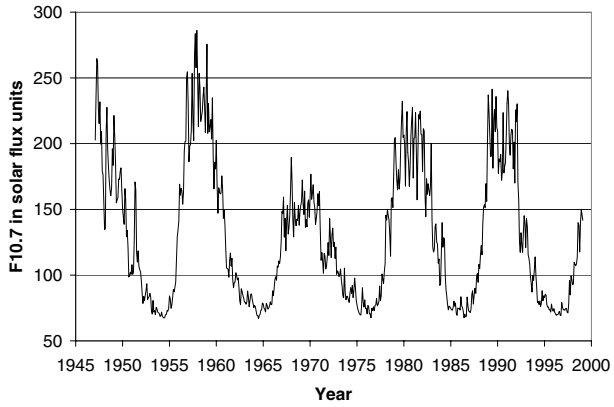


Figure 4. Monthly means of the 10.7 cm solar flux between 1946 and present. With this index the shape of the cycle is much more apparent, and it is possible to see the fluctuations of up to a year or so due to the appearance and decay of complexes of activity. There is a constant background component due to thermal emission from the quiet solar disc, which has a flux density of about 64 solar flux units.

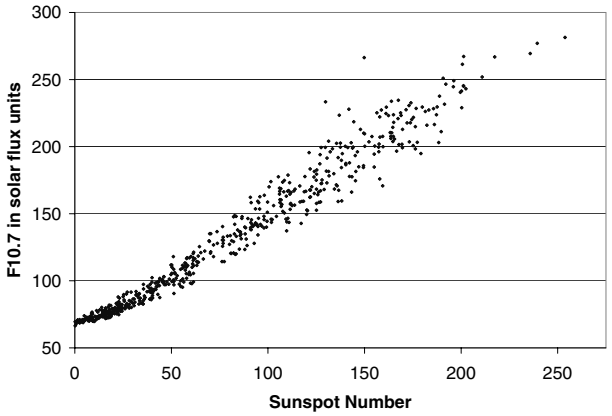


Figure 5. Plot of monthly means of the 10.7 cm flux plotted against similarly averaged values of the sunspot number. The two indices are highly correlated. The reduction in slope for sunspot numbers less than about 30 may be due to low levels of photospheric activity not generating much activity in the overlying corona.

magnetic flux in these strong-field elements is subtracted from the total magnetic flux in all elements, the remainder, which is the total magnetic flux in weak-field elements, shows little variation over the solar activity cycle. Harvey (1994) and Schrijver (1986) note that the total magnetic flux seems to correlate with the total area of active regions, as represented by total PLAGE area and other indices, and that it can be broken into contributions from individual active regions on the basis of their areas.

Chromosphere

Chromospheric activity can be monitored through measurements of the millimeter-wavelength emission integrated over the solar disk. These emissions are purely free–free thermal emission from plasmas trapped in active region magnetic fields. However, because the chromosphere is not much hotter than the photosphere, millimeter-wavelength emission from active regions is only slightly hotter than the background emission from the solar disk. Moreover, atmospheric attenuation makes these measurements difficult to make from the ground on any regular basis. They need to be made from space platforms. The most widely used indices of chromospheric activity are derived from spectral lines. The main ones are the H and K lines of Ca II and Mg II. These have the great additional advantage that they can be used to study chromospheric activities in other stars. Another spectral line widely used for monitoring chromospheric variability is the 1083 nm line produced by He I. This line is purely chromospheric in origin, although there is evidence that this line is 'pumped' to some extent by ultraviolet emission from the corona.

Corona

The slowly varying component of solar radio emission provides an excellent activity index. This emission is best observed in the wavelength range 2–50 cm but shows the largest variability over the solar cycle at wavelengths around 10 cm. It is thermal in origin and is dominated by free–free emission from plasma concentrations trapped over active regions. It also contains a contribution from gyroresonance at low harmonics of the electron gyrofrequency, originating in the corona over sunspots (Krüger *et al* 1985; Tapping 1987; Tapping and Harvey 1994), and maybe some non-thermal radio emission. The 10.7 cm solar flux index, also referred to as F10.7, $F_{10.7}$ or F_{10}, is a measurement of the slowly varying component at 10.7 cm wavelength. Its advantage over sunspot number is that it is an objective measurement that can be made in almost any weather. The monthly averages of $F_{10.7}$ from 1947 to the present are shown in figure 4. It is expressed in *solar flux units*, where one solar flux unit (or sfu) is 10^{-22} W m^{-2} Hz^{-1}. It now forms a continuous, consistent record covering more than 50 yr. In the optical part of the spectrum, the *coronal green line* at 530.2 nm wavelength is a very useful index. This strong emission line is produced by Fe XIV. Other useful spectral lines include Fe X (637.3 nm), Fe XI (789.2 nm) and Fe XIII (1076.8 nm).

Thermal emission in the x-ray wavelength region is significant for plasmas with electron temperatures of 10^6 K or more. This radiation is produced entirely by free–free transitions and is probably the 'purest' index of coronal activity. Its disadvantage is that it has to be measured from space platforms, where it is difficult to obtain long-term, consistent, continuous records. The data here were obtained from the GOES satellite.

Comparing indices

The plots in figure 1 show significant correlations between the indices, despite their being produced at different locations in the solar atmosphere and by different processes. The most striking correspondences are between

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

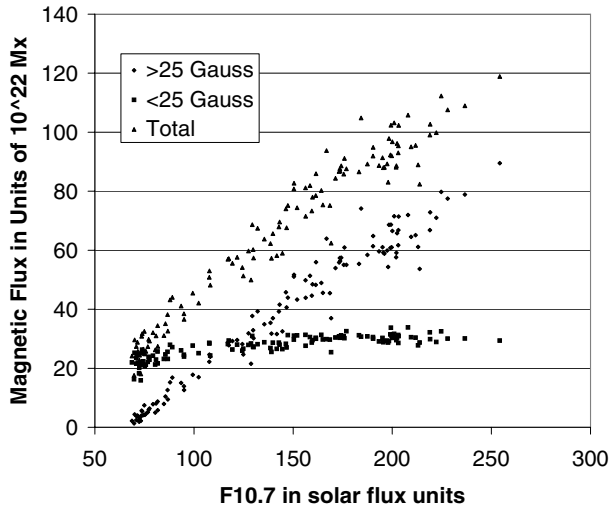


Figure 6. Monthly means of the *weak-field* and *strong-field* magnetic elements plotted against the 10.7 cm solar flux. The *weak-field elements* show no relationship with the 10.7 cm solar flux. However, the correlation with the *strong-field elements* is very strong.

sunspot number, $F_{10.7}$, the He I (1083 nm) line and total facular area. These indices are also highly correlated with total magnetic flux, although rather less so than they are with each other. The Fe XIV line correlates slightly less well (but still significantly), and the soft x-ray flux a little less again. When the short-duration, high peaks are removed, the remaining emission correlates well with $F_{10.7}$ and sunspot number when activity is high, but not when it is low.

Figure 5 shows a plot of sunspot number against $F_{10.7}$. The correlation is high. However, there is a significant decline in the degree of correlation for sunspot numbers smaller than about 30. Figure 6 shows a similar plot of total magnetic flux against $F_{10.7}$. In this case elements of photospheric magnetic flux are divided into two classes: strong-field elements with average magnetic fluxes exceeding 25 G and weak-field elements with average field strengths less than this value. The total magnetic flux from all elements correlates well with $F_{10.7}$, but this variation comes almost entirely from the strong-field elements.

These varying but usually high degrees of correlation can be largely understood when we consider the various levels in the solar atmosphere at which these emissions originate. If magnetic fields are the engine driving all manifestations of solar activity, it is intriguing that the total photospheric flux does not correlate perfectly with sunspot number, $F_{10.7}$ and He I. The not-perfect correlation with sunspot number could be due to not all active regions having sunspots, and the maxima of plage area and sunspot number not occurring at the same time in the evolution of an active region. In the case of indices using chromospheric and coronal emissions, it is possible that not all elements of photospheric flux stronger than 25 G are associated with the footpoints of loops extending into the upper chromosphere and corona and which contain significant amounts of trapped plasma. The indices arise in these trapped plasmas. However, the strong magnetic fields in faculae and especially sunspots form loops extending into the overlying corona over their host active regions. These are good candidates for being plasma traps. This could account for there being a good correspondence to He I and $F_{10.7}$, which originates primarily in the low corona. The soft x-ray flux is an exclusively coronal emission. This emission would only increase when activity extends to those heights. Therefore, we would expect poor correspondence when activity is low, as is observed. The Fe XIV line probably originates lower in the corona, hence the better correlation with other indices over the whole SOLAR CYCLE. The high degree of correlation between the indices, particularly sunspot number, $F_{10.7}$, facular area and He I, indicates that we can use one of these indices as a proxy for the others. For example, for sunspot numbers bigger than about 30, the sunspot number and $F_{10.7}$ are related by $R_z = 1.14 \times F_{10.7} - 73.21$, and $F_{10.7}$ and total magnetic flux can be related through the equation $\Phi = 0.5 \times F_{10.7} - 10$, where Φ is the total (strong-field) magnetic flux in units of 10^{22} Mx.

Discussion

The situation we have with solar activity indices is extremely fortunate for solar researchers. There are indices more-or-less specific to particular levels in the solar atmosphere, which are highly correlated with one another over (with the possible exception of the soft x-ray flux) the full variation in activity during the solar cycle. This would suggest that the emission processes (or mixes of processes) do not change suddenly as activity changes. The other implication is that all the activity indices are strongly modulated by the varying distribution of magnetic flux at and above the photosphere, in a more-or-less simple, linear way, probably in that the magnetic fields support the plasma concentrations producing the emissions forming the indices. If the temperatures ($\sim 10^4$ K in the chromosphere and $\sim 10^6$ K in the corona) vary less than does the density, there is a fairly simple relationship: the more magnetic flux—the higher the density—the brighter the emission.

The ability to summarize the various aspects of solar activity into a single quantity is highly useful. It makes it possible to study the long-term behavior of the Sun, to compare it with other stars and to investigate its role as a driver for processes on the Earth and in space. Considering the rich menagerie of phenomena embraced in the term *solar activity*, is is surprising and fortunate that single-parameter indices have proved valuable in many areas of solar and solar-related research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Acknowledgments
The magnetic flux and He I (1083 nm) data used here were kindly provided by the National Solar Observatory at Kitt Peak. I would like to thank Karen Harvey particularly for providing some of the results of her research into solar magnetic activity. The sunspot number data and GOES soft x-ray observations were obtained from the National Geophysical Data Center, which is operated by the US National Oceanic and Atmospheric Administration. The 10.7 cm solar flux data are provided as a service by the National Research Council of Canada. The Fe XIV observations were made by the Czech Academy of Sciences.

Bibliography
Gaizauskas V, Harvey K L, Harvey J W and Zwaan C 1983 Large-scale patterns formed by solar active regions during the ascending phase of cycle 21 *Astrophys. J.* **265** 1056–65

Harvey K L 1994 Irradiance models based on solar magnetic fields *The Sun as a Variable Star (IAU Colloq. 143)* ed Pap, Frölich, Hudson and Solanki (Cambridge: Cambridge University Press) pp 220–5

Krüger A, Hildebrandt J and Furstenberg F 1985 A working model of the solar s-component radio emission *Astron. Astrophys.* **143** 72–6

Lang K R 1995 *The Sun, Earth and Sky* (Berlin: Springer)

Schrijver C J 1986 Stellar magnetic activity *PhD Thesis* University of Utrecht

Tapping K F 1987 Recent solar radio astronomy at centimeter wavelengths. The temporal variability of the 10.7 cm flux *J. Geophys. Res.* **92** 829–38

Tapping K F and Harvey K L Slowly-varying microwave emissions from the solar corona *The Sun as a Variable Star (IAU Colloq. 143)* ed Pap, Frölich, Hudson and Solanki (Cambridge: Cambridge University Press) pp 182–95

Wilson P R 1994 Solar and stellar activity cycles *Cambridge Astrophysics Series* No 24 (Cambridge: Cambridge University Press)

Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

K F Tapping

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Activity: Long-Term Records

Long-term records of solar activity show several interesting features. The most prominent one is the 11 yr cycle that can be traced back in the sunspot record to the 17th century. Indications of longer cycles have also been found, but are difficult to confirm due to the limited length of the observational period. Another interesting feature is the increasing trend of the solar activity between 1700 AD and today. This trend is interrupted by short periods of reduced solar activity around 1900, 1800, and 1700. The last period, called the Maunder minimum (after E W MAUNDER), is by far the most pronounced one and lasted from 1645 to 1715 AD. It is characterized by an almost complete absence of sunspots.

Cosmogenic nuclides measured in tree rings and polar ice indicate that similar minima occurred several times during the past 10 000 years.

When analysing long-term records of solar activity we distinguish between direct and indirect proxies. Direct proxies are continuously or regularly recorded from ground-based as well as from space-borne instruments. In general their quality is high, but limited to the last few decades or centuries. Going back in time the number of continuous direct observations of phenomena related to solar activity decreases. Table 1 provides a list of the main proxies of solar activity that were recorded continuously for at least 50 yr with a time resolution of 1 yr or better.

Table 1. Direct proxies of long-term solar activity.

Solar activity proxy	Records since
Sunspots	∼1600
Aurorae	∼1600
Aa index	1844
Neutrons	1937

Direct proxies of solar activity

SUNSPOTS are the most prominent proxies of solar activity. After the discovery of the 11 yr SOLAR CYCLE by Schwabe in 1843 a well-defined observational program was set up by Wolf leading to the Zurich RELATIVE SUNSPOT NUMBER. Collecting all available information about earlier sunspot observations, Wolf reconstructed the sunspot record back to about AD 1610, the time when the telescope was invented. Pretelescopic observations are only sporadic and limited to relatively large spots. A re-examination based on an extended database of observations agrees well with the Zurich sunspot numbers since 1750, but differs somewhat for earlier times (Hoyt and Schatten 1998).

AURORAE are spectacular colorful phenomena high up in the atmosphere which are mainly visible at high latitudes in both hemispheres (aurora borealis and aurora australis). They are due to ionization processes caused by the interaction of the SOLAR WIND with the atmosphere. Although well known since ancient times their connection to solar activity was only discovered in 1896 by Birkeland. To establish a consistent reliable record is difficult because their appearance is restricted mainly to two oval bands around the magnetic poles that are only sparsely populated (Legrand and Simon 1987). However, many strong individual events are reported that could also be seen in highly populated low-latitudinal areas.

The aa index is a measure of the global disturbance of the geomagnetic field. It is derived from averaged two-hourly measurements at two antipodal stations (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS). The aa record goes back to 1844, about the time of the invention of the magnetometer (Mayaud 1973).

The neutron flux in the atmosphere is produced by the interaction of galactic COSMIC RAY particles (protons and alpha particles) with the atmosphere. The primary cosmic ray flux is modulated by the magnetic properties of the solar wind which prevent the low-energy particles from penetrating into the atmosphere. As a consequence the neutron flux recorded by neutron monitors is low during periods of high solar activity and vice versa. The longest record of a neutron monitor goes back to 1951 (Climax, CO, USA, operated by the University of Chicago). It can be combined with the record of the ionization chamber from Cheltenham, which was in operation between 1937 and 1973. In figure 1 sunspots, aurorae and aa index are plotted with an annual resolution. All of them show clearly the 11 yr Schwabe cycle and a generally increasing trend. However, a detailed comparison reveals differences in amplitudes and phases that are due to the fact that, beside observational uncertainties, each solar proxy reflects a different aspect of solar activity.

Indirect proxies of solar activity

High-resolution direct observations of proxies for solar activity are restricted basically to the telescopic era starting in AD 1610. Earlier observations are sparse and usually not very reliable. At present, the only way to extend significantly the history of solar activity is to rely on indirect proxies. Cosmogenic radionuclides such as ^{10}Be and ^{14}C proved to be useful tools for this purpose (Beer *et al* 1990, Stuiver and Braziunas 1993).

Cosmogenic radionuclides are produced by the interaction of cosmic ray particles with the atmosphere. ^{10}Be (half-life: 1.5×10^6 yr) is mainly produced at high energy by neutron-induced spallation reactions on nitrogen and oxygen, whereas ^{14}C (half-life: 5730 yr) is produced by thermal neutrons through the reaction ^{14}N(n, p)^{14}C. The production rate for a particular cosmogenic nuclide j can be calculated for any point D within the atmosphere according to the formula:

$$P_j(D) = \sum_i N_i \sum_k \int_0^\infty \sigma_{ijk}(E_k) J_k(E_k, D)\, \mathrm{d}E_k$$

where $P_j(D)$ is the production rate of nuclide j at position D, N_i the density of target element i, $\sigma_{ijk}(E_k)$ the cross section for the production of nuclide j by the particle k with energy E_k interacting with the target element i and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

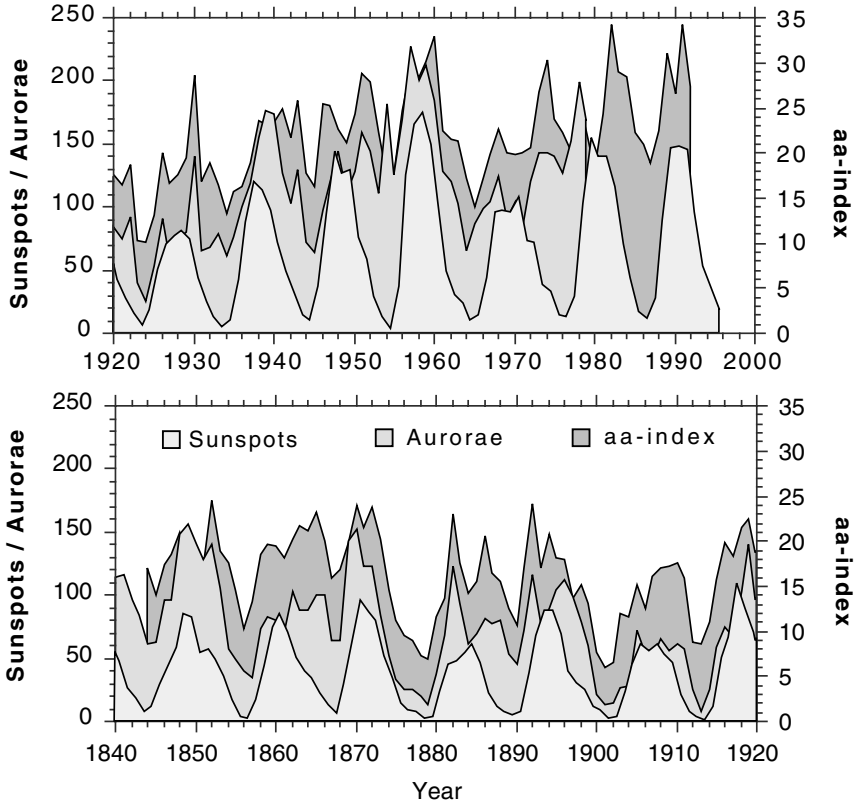


Figure 1. Comparison of the three longest records of direct solar activity proxies.

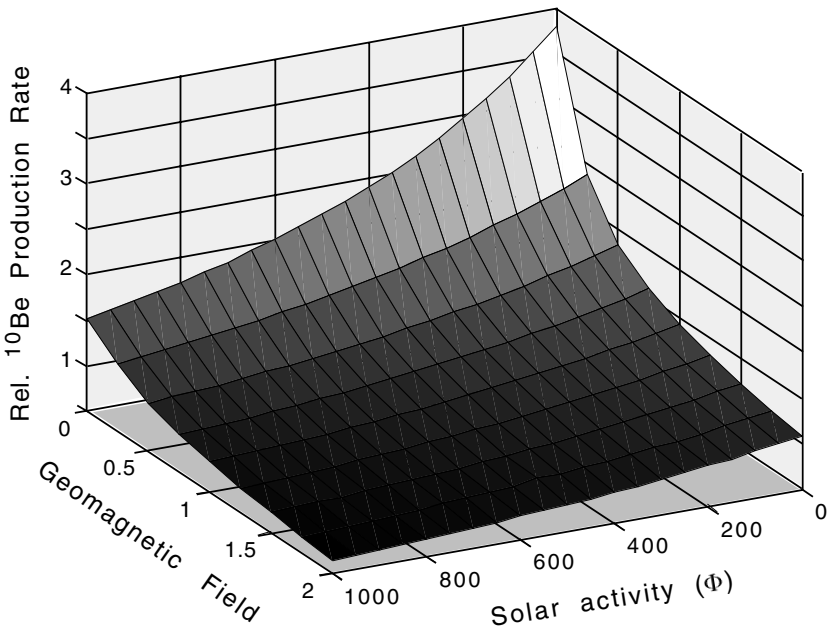


Figure 2. Calculated relative mean global production rate of ^{10}Be in the atmosphere as a function of solar activity expressed by the parameter ϕ and the geomagnetic field intensity relative to its present value. The production rate 1 corresponds to 0.018 ^{10}Be atoms cm^{-2} s^{-1}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

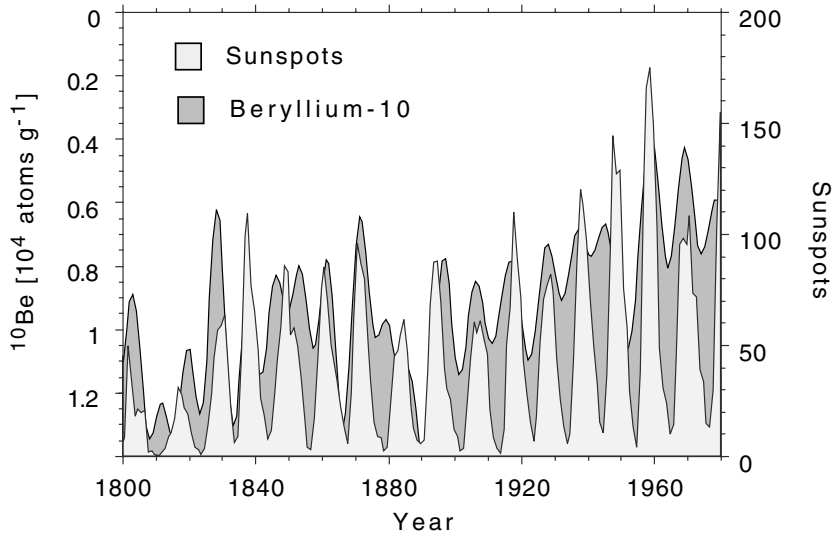


Figure 3. Comparison of the ^{10}Be record from Dye 3, Greenland, with the sunspot record. The time scale of the ^{10}Be record has been shifted by 1 yr to account for the atmospheric residence time.

$J_k(E_k, D)$ the differential flux of particle k with energy E_k at position D.

$J_k(E_k, D)$ can be calculated using a physical model simulating the interaction of primary cosmic ray particles with the atmosphere and the subsequent production and transport of secondary particles in the atmosphere (Masarik and Beer 1999). To describe the dependence of the primary cosmic ray proton flux on the solar activity the following parameterization is usually applied:

$$J(E_p, \phi) = C_p \frac{E_p(E_p + 2m_pc^2)(E_p + x + \phi)^{-2.5}}{(E_p + \phi)(E_p + 2m_pc^2 + \phi)}$$

where $x = 780 \exp(-2.5 \times 10^{-4} E_p)$, $J(E_p, \phi)$ is the proton flux with energy E_p for the solar modulation parameter ϕ, E_p the proton energy, M_pc^2 the proton rest mass (938 MeV) and C_p a normalization factor C (1.244×10^6 cm^{-2} s^{-1} MeV^{-1}). $\phi = 1000$ MeV corresponds to a very active Sun, $\phi = 0$ MeV to a completely quiet Sun. The production rate P depends not only on the solar modulation parameter ϕ but also on the altitude in the atmosphere, the geomagnetic latitude and the geomagnetic dipole moment. In figure 2 the mean global production rate for ^{10}Be is depicted as a function of the solar parameter ϕ and the geomagnetic field intensity relative to its present value. The production rate P is normalized to 1 for $\phi = 550$ MeV and the geomagnetic field 1. During a typical 11 yr Schwabe cycle ϕ fluctuates between 450 and 750 MeV, leading to a mean global peak-to-trough ratio of 1.3 for ^{10}Be and 1.4 for ^{14}C.

After production ^{10}Be atoms in the atmosphere become attached to aerosols and are removed from the atmosphere after a mean residence time of about 1 yr mainly by wet deposition. ^{14}C oxidizes to CO_2 and exchanges between atmosphere, ocean and biosphere with residence times in the last two of about 10 and 60 yr, respectively. To reconstruct the history of production variations natural archives can be used: ice sheets record continuously the ^{10}Be concentration in precipitation and tree rings the ^{14}C concentration in the atmosphere.

Measurements of cosmogenic nuclides in natural archives provide therefore information on several parameters.

1. Changes in the production rate caused either by solar or geomagnetic modulation.
2. Changes in the system itself. Climate changes affect the system behavior. In the case of ^{10}Be they influence the atmospheric transport and deposition processes. In the case of ^{14}C changes in the size of the carbon reservoirs (atmosphere, biosphere, ocean) and in the exchange fluxes between them can occur, affecting the atmospheric ^{14}C concentration.

An interpretation of cosmogenic radionuclide records in terms of solar modulation needs therefore a detailed analysis of all possible involved processes in order to exclude geomagnetic modulation and system effects.

System effects can relatively easily be identified by comparing ^{10}Be with ^{14}C variations. The two radionuclides are very similar regarding production but completely different regarding system effects. Therefore, a comparison using a carbon cycle model allows us to distinguish between production and system effects. The results of several comparisons confirm that both nuclides basically reflect changes of the production rate during the past centuries to millennia.

Information on solar activity derived from cosmogenic radionuclides

The 11 yr solar modulation of the ^{10}Be production rate was found for the first time in an ice core from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

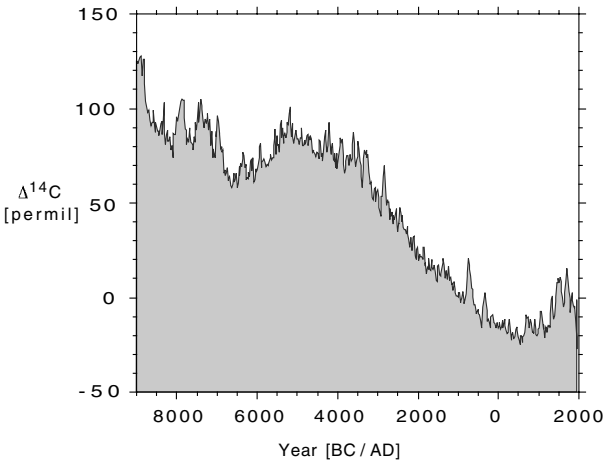


Figure 4. $\Delta^{14}C$ record measured in tree rings for the last 11 000 years. The atmospheric $^{14}C/^{12}C$ ratio was most of the time higher than today owing to a higher production rate. Note the peaks of typically 70–120 yr duration that indicate periods of low solar activity (e.g. Maunder minimum AD 1645–1715).

Dye 3, Greenland, covering the time interval 1423–1985 (Beer *et al* 1994) (figure 3). The comparison of the Dye 3 record with the sunspot record shows in general the expected anticorrelation with a phase lag of 1 yr. The periodicity over the time interval 1423–1985 is 10.83 yr. The Dye 3 record also shows the 88 yr Gleissberg cycle which is difficult to derive from the sunspot record because its length is less than 400 yr. As a consequence of the strong damping effect of the carbon system on short-term variations the most prominent periodicity in the ^{14}C record is the 205 yr De Vries cycle. Longer periodicities, which show up during certain time intervals in cosmogenic radionuclide records, cannot yet be attributed unambiguously to solar activity because they could also be due to changes in the geomagnetic dipole field.

An especially interesting period in the sunspot record is the so-called Maunder minimum (AD 1645–1715). During this time interval only very few sunspots could be observed and the solar activity was very low, leading to an enhanced production rate of ^{10}Be and ^{14}C. Similar minima periods occurred earlier often at intervals of about 200 yr (Spoerer around 1500, Wolf around 1300) and several others are visible in the $\Delta^{14}C$ curve during the last 11 000 years (figure 4). $\Delta^{14}C$ is defined as depletion or enhancement (in per mil) of the atmospheric $^{14}C/^{12}C$ ratio relative to the standard ratio.

The absence of sunspots during solar activity minima does not mean that the solar dynamo was switched off. On the contrary, ^{10}Be shows a clear 11 yr Schwabe cycle during the Maunder minimum (Beer *et al* 1998).

The results obtained so far show that, in spite of the difficulty in separating solar, geomagnetic and climatic effects, cosmogenic radionuclides do have the potential to extend the solar activity record back to several 10 000 yr with a time resolution of 1 yr.

Bibliography

Beer J, Baumgartner S *et al* 1994 Solar variability traced by cosmogenic isotopes *The Sun as a Variable Star: Solar and Stellar Irradiance Variations* ed J M Pap, C Fröhlich, H S Hudson and S K Solanki (Cambridge: Cambridge University Press) pp 291–300

Beer J, Blinov A *et al* 1990 Use of ^{10}Be in polar ice to trace the 11-year cycle of solar activity *Nature* **347** 164–6

Beer J, Tobias S M *et al* 1998 An active Sun throughout the Maunder minimum *Sol. Phys.* **181** 237–49

Hoyt D V and Schatten K H 1998 Group sunspot numbers: a new solar activity reconstruction *Sol. Phys.* **179** 189–219

Legrand J P and Simon P A 1987 Two hundred years of auroral activity (1780–1979) *Ann. Geophys.* **5a** (3) 161–8

Masarik J and Beer J 1999 Simulation of particle fluxes and cosmogenic nuclide production in the Earth's atmosphere *J. Geophys. Res.* **104** 12 099–13 012

Mayaud S N 1973 A hundred year series of geomagnetic data 1868–1967, indices aa, storm sudden commencements *Bull. Int. Assoc. Geomagn. Aeron.* **33** 1–251

Stuiver M and Braziunas T F 1993 Sun, ocean, climate and atmospheric $^{14}CO_2$, an evaluation of causal and spectral relationships *Holocene* **3** 289–305

J Beer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Atmosphere: Empirical Models

The basic atmospheric stratification of the Sun can be described as having (1) an outwardly decreasing temperature with height in the photospheric layers which emit radiation at visible wavelengths, (2) a temperature minimum region where the temperature no longer decreases but begins to increase with height, (3) an extended chromospheric region where the temperature increases to values at which hydrogen starts to become partially ionized and the atmosphere starts to lose energy by Lyman line radiation, (4) a narrow transition region separating the chromosphere and the fully ionized corona, having a large initial temperature gradient near the top of the chromosphere and a smaller gradient at higher temperatures, and (5) the hot ($T > 10^6$ K) corona where hydrogen is fully ionized. See figure 1.

The temperature distribution shown in this figure is for a model of the average quiet Sun and does not represent regions of the solar disk that are brighter and darker than the average, but similar models can be constructed for these other regions. The observed inhomogeneous features on the solar disk include the granular cell and network pattern throughout quiet, relatively undisturbed areas, active regions of varying brightness, as well as sunspots and flares. Significant variations with time occur in such structures, particularly in flares (see PHOTOSPHERE: GRANULATION, SOLAR CHROMOSPHERIC PLAGE, SOLAR ACTIVE REGIONS, SUNSPOT MODELS and SOLAR FLARE MODELS).

To a first approximation, such different types of features can be treated separately, assuming little interaction between the different features, since their horizontal extent is usually much larger than the vertical distance between the photospheric and coronal layers in each case. Hence, it is reasonable as a first approximation to construct separate one-dimensional models that describe the vertical structure of each horizontally distinct region.

Consider a given type of brightness feature in a quiet or active region, such as typical centers of granulation cells or bright network structures, and suppose we have observations of the spectrum of such a feature over a wide range of wavelengths, from the radio region to x-rays. Because the opacity of the solar atmosphere varies with wavelength, we can probe different atmospheric layers by observing radiation at different wavelengths. At wavelengths where the opacity is small the observed radiation originates deep in the atmosphere, and at wavelengths where the opacity is greater the observed radiation originates higher in the atmosphere.

The lowest opacity in the solar atmosphere occurs at $\lambda = 1.6$ μm, the threshold for H^- bound–free absorption. At that wavelength the light originates from the deepest observable layers, and the disk-center brightness is observed to be about 6800 K. The average brightness of the integrated disk at 1.6 μm is about 500 K lower because of limb darkening (see SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION). For infrared wavelengths longer than 1.6 μm the opacity increases as λ increases so that the observed radiation is emitted from higher, cooler layers in the atmosphere, and the observed brightness temperature decreases, reaching a minimum value of about 4500 K at $\lambda = 150$ μm. The opacity continues to increase with increasing wavelength for $\lambda > 150$ μm, and the observed brightness temperature now increases as a result of the increasing temperature in the chromosphere.

Now consider the wavelength region $\lambda < 1.6$ μm. The opacity increases with decreasing wavelength and the brightness temperature decreases, since unit optical depth occurs at successively greater heights. Near 160 nm the brightness temperature passes through a minimum value of about 4400 K and then increases with decreasing λ, again because of the increasing temperature in the chromosphere.

Similar indications of the temperature stratification are evident in the profiles of spectral lines, since the opacity is much greater at line center than in the line wings or the nearby continuum. The Ca II resonance lines have weak emission features near the line centers owing to the chromospheric temperature rise, but this emission varies with time and spatial position and is sensitive to small-scale kinematic effects. The clearest evidence for the chromospheric temperature rise is the increase in continuum brightness temperature in the $\lambda > 150$ μm and $\lambda < 160$ nm ranges and the UV emission lines shortward of 160 nm. These emission lines vary with position and time, but observations at high spatial and temporal resolution (Carlsson *et al* 1997) show that, while the UV line emission varies with time and position, it never vanishes, implying that the chromospheric temperature rise is persistent and not the time average of transitory increases in temperature (see SOLAR SPECTROSCOPY: ULTRAVIOLET AND EUV EMISSION).

However, a problem that is still under investigation is why the strongest lines of carbon monoxide (CO) in the infrared show no central emission. Unit optical depth in these line cores is estimated to occur above the temperature mimimum region (based on results from one-dimensional time-independent, hydrostatic-equilibrium modeling), and these lines are dominated by collisional processes and should be formed in local thermodynamic equilibrium (LTE). Thus the brightness temperature at line center should be the kinetic temperature at unit optical depth. The lowest disk-center brightness temperature in the strongest infrared CO line is about 4100 K (Avrett 1995), but values down to 3800 K are seen near the solar limb (Noyes and Hall 1972, Ayres 1995) (see SOLAR SPECTROSCOPY: INFRARED EMISSION).

An important characteristic of the CO lines is their sensitivity to low temperatures. High-temperature regions along the line of sight tend to be transparent, because CO molecules readily dissociate with increasing temperature. CO observations might be reconciled with the observations of persistent UV line emission if the cores of the strong CO lines are formed no higher than the temperature minimum region, which shows temporal and spatial variations as a result of granulation

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

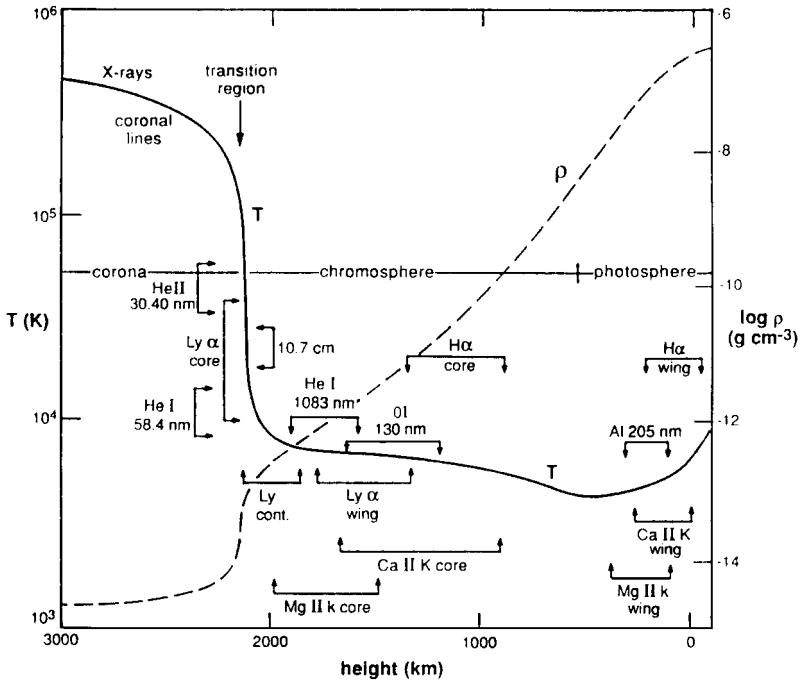


Figure 1. Temperature and density as functions of height for a model of the average quiet Sun (model C). The calculated height intervals where the various continua and line features originate are indicated. (See Avrett (1998) for details.)

and wave motions. See Uitenbroek *et al* (1994). The 3800 K values seen near the limb are more difficult to interpret, however, indicating that the line of sight near the limb passes through cool regions that do not affect the UV chromospheric lines (see CHROMOSPHERE: THERMAL BIFURCATION).

Apart from the unsolved difficulties presented by these CO lines, models corresponding to temperature distributions such as in figure 1 can be constructed for the various brightness components, i.e. for different types of observed features, that roughly account for the entire observed spectrum from each component. Such models are given by Fontenla *et al* (1993) for the average quiet Sun (as in figure 1), for faint and bright components of the cell-network pattern of quiet areas, and for a bright active region. These are shown as models C, A, F and P, respectively, in figure 2.

The temperature as a function of height for each brightness component model has been adjusted so that the calculated spectrum agrees as well as possible with observations. The corresponding density distribution is determined from a balance between gas (and turbulent) pressure and solar surface gravity. The intensity of radiation as a function of wavelength, position and direction and the number densities of atoms, ions, and molecules in various energy states are determined by solving the coupled equations of radiative transfer and statistical equilibrium, accounting for arbitrarily large departures from LTE and arbitrarily large optical depths. The calculations also depend on the use of accurate

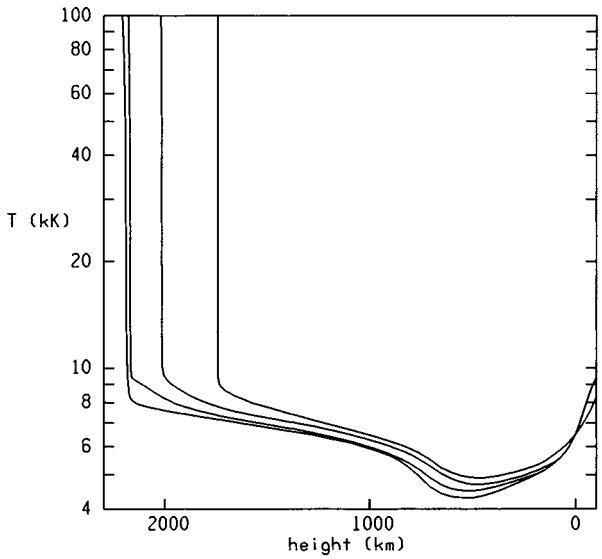


Figure 2. Temperature as a function of height for quiet-Sun models A, C and F and for plage model P (see Fontenla *et al* 1993).

opacities for the millions of atomic and molecular lines throughout the spectrum. See Kurucz (1996, 1997) and SOLAR SPECTROSCOPY: VISIBLE EMISSION.

In addition to the temperature distribution, we assume a microvelocity distribution needed to account for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the observed nonthermal broadening of lines formed at various heights. The same values are used for a turbulent or kinetic pressure contribution to the total pressure. It appears to be sufficient in the various component model calculations to assume that the microvelocity is the same function of density for all models (Fontenla *et al* 1993). Thus the temperature distribution essentially defines each model.

To the extent that the calculated spectrum agrees with observations, we can have confidence that the various models represent the structure of the various components of the solar atmosphere, and we can use the models to understand the physical processes at work. For example, Anderson and Athay (1989) derived the distribution of mechanical heating needed to account for the chromospheric temperature rise in model C (see CHROMOSPHERE: HEATING MECHANISMS).

Spectra computed from component models A, C, F and P agree reasonably well with time-averaged observations of faint, average and bright components of quiet regions and of bright plage regions. As stated earlier, observations of UV emission lines at high spatial and temporal resolutions show that, while there are significant time variations at each position, the UV emission lines never vanish, indicating that the chromospheric temperature rise is persistent rather than transitory.

The narrow transition region separating the chromosphere and corona can be interpreted as the result of (1) sufficient mechanical heating at low densities to heat the corona to the high temperatures at which hydrogen is fully ionized and (2) strong radiative cooling by the hydrogen and helium resonance lines in the upper chromosphere that balances chromospheric mechanical heating, so that the temperature remains in the 8000 K range until the transition region is reached where hydrogen abruptly changes from being mostly neutral to mostly ionized (see TRANSITION REGION MODELS).

The energy required to produce these outwardly increasing temperatures is relatively small. About 10^{-4} of the total energy output of the Sun is needed to heat the chromosphere. The energy required to produce the corona is an order of magnitude smaller, and most of this energy is transported back through the upper transition region by thermal conduction and particle diffusion to be radiated away by the hydrogen and helium resonance lines. The transition region for each of the component models can be determined theoretically by solving the energy balance equations, making use of upper-chromosphere boundary conditions that are determined with reference to observational data. See Fontenla *et al* (1993).

Each empirical model considered here is based on an assumed temperature versus height extending from the deep photophere to the base of the transition region. In principle, an entire spectrum can be computed and compared with observations, and the disparities used to adjust the temperature distribution to get optimum agreement with observations for a particular component region. Given sufficient agreement, such models can provide much useful information about the physical properties of the solar atmosphere.

Time-independent, one-dimensional component models based on observations with low time resolution can only provide information about time-averaged behavior. It is not difficult to construct a series of 'snapshot' models describing the time evolution of the atmospheric region in response to waves or other disturbances. The assumption that different brightness components can be independently modeled, ignoring their effect on one another, is clearly not always valid, as indicated, e.g. by the central brightness temperature of 3800 K observed in strong CO lines near the limb discussed above.

The models in figure 2 indicate that the transition region is located at a height of approximately 2000 km above the photosphere (above the level where the radial optical depth in the continuum at 500 nm is unity). This 2000 km height of the transition region is inferred from the properties of the disk-center spectrum used to define the models. It is important to determine whether limb observations give the same result.

Observations at the limb in the 1083 nm line of He I show a narrow band of enhanced emission at approximately this height. This band of emission can be explained as the result of coronal line radiation at wavelengths below 50 nm, which can ionize He I, illuminating the top layers of the chromosphere. The resulting ionization of He I followed by recombination populates the levels of He I responsible for the 1083 nm emission from these upper chromospheric layers. This interpretation is consistent with the observed property that 1083 emission is diminished in coronal hole regions where the coronal line radiation is reduced. See Zirin (1975) and Avrett *et al* (1994).

For a uniform one-dimensional atmosphere with the transition region at a fixed height, the 1083 nm emission above the limb would abruptly decrease with height after reaching the peak value near 2000 km. However, observations show a gradual decrease above 2000 km (Schmidt *et al* 1994) indicating the presence of chromospheric material at heights above 2000 km.

The 1083 nm line observations at the limb give particularly useful information because this line is strong enough to produce observable emission, but weak enough so that the total optical thickness along a tangent ray passing just above the limb is less than unity, so that this ray can sample the entire line of sight. In stronger lines, such as the hydrogen Hα line, or in the continuum at centimeter wavelengths, the transition region would appear to be located higher owing to the optical thickness of relatively small amounts of chromospheric material projecting above 2000 km in the form of spicules and relatively cool prominence-like gas (see CHROMOSPHERE: SPICULES and SOLAR PROMINENCE MODELS).

Solar atmospheric modeling gives us direct insight into the physical processes that occur in the atmosphere. The spectra calculated from the models also have a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

practical application. The calculated intensity as a function of wavelength for any particular quiet or active region can be used to fill in the gaps between the relatively few portions of the spectrum that have been observed with high spatial and spectral resolution with good absolute calibration. To the extent that, for a well-defined component region (such as cell centers or bright network features), the calculated intensities generally agree with the available observations over a sufficient wavelength range, the model calculations for the given quiet or active feature provide the absolute intensity versus wavelenth over the entire spectrum at high spectral resolution for that feature.

Variations of solar irradiance are due mostly to changes in the fraction of bright active regions present on the solar disk facing the Earth as a function of time (see SOLAR IRRADIANCE). This fraction varies as the Sun's rotation brings active regions into and out of the hemisphere facing the Earth. Active regions also are responsible for most of the long-term variations of both the total irradiance and the spectral irradiance over the solar cycle. Solar images at a given time at one or more wavelengths (e.g. in the visible) can be used to define the pattern of bright and dark regions. Using the calculated intensity versus wavelength for each region, the disk image can be constructed for any given wavelength (e.g. in the ultraviolet), and we can integrate over the disk to get the spectral irradiance at that wavelength. See Avrett (1998). Such results can provide spectral irradiances on solar-rotation and solar-cycle time scales at wavelengths for which observations are not available (see SOLAR CYCLE).

Bibliography

Anderson L S and Athay R G 1989 Model solar chromosphere with prescribed heating *Astrophys. J.* **346** 1010–8

Avrett E H 1995 Two-component modeling of the solar IR CO lines *Infrared Tools for Solar Astrophysics: What's Next* ed J Kuhn and M Penn (Singapore: World Scientific) pp 303–11

Avrett E H 1998 Modeling solar variability—Synthetic models *Solar Electromagnetic Radiation Study for Solar Cycle 22* ed J M Pap, C Frölich and R K Ulrich (Dordrecht: Kluwer) pp 449–69

Avrett E H, Fontenla J M and Loeser R 1994 Formation of the solar 10830 Å line *Infrared Solar Physics* ed D M Rabin, J T Jefferies and C Lindsey (Dordrecht: Kluwer) pp 35–47

Ayres T R 1995 Thermal bifurcation revisited *Infrared Tools for Solar Astrophysics: What's Next* ed J Kuhn and M Penn (Singapore: World Scientific) pp 289–302

Carlsson M, Judge P G and Wilhelm K 1997 SUMER observations confirm the dynamic nature of the quiet solar outer atmosphere: the internetwork chromosphere *Astrophys. J.* **486** L63–6

Fontenla J M, Avrett E H and Loeser R 1993 Energy balance in the solar transition region. III. Helium emission in hydrostatic constant-abundance models with diffusion *Astrophys. J.* **406** 319–45

Kurucz R L 1996 Model stellar atmospheres and real stellar atmospheres *Model Atmospheres and Spectrum Synthesis (ASP Conf. Ser. 108)* ed S Adelman, F Kupka and W W Weiss pp 2–18

Kurucz R L 1997 Progress on model atmospheres and line data *Fundamental Stellar Properties* ed T R Bedding, A J Booth and J Davis (Dordrecht: Kluwer) pp 217–26

Noyes R W and Hall D N B 1972 Thermal oscillations in the high solar photosphere *Astrophys. J.* **176** L89–92

Schmidt W, Knoelker M and Westendorp Plaza C 1994 Limb observations of the He I 1083.0 nm line *Astron. Astrophys.* **287** 229–32

Uitenbroek H, Noyes R W and Rabin D 1994 Imaging spectroscopy of the solar CO lines at 4.67 microns *Astrophys. J.* **432** L67–70

Zirin H 1975 The helium chromosphere, coronal holes, and stellar x-rays *Astrophys. J.* **199** L63–6

Eugene H Avrett

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Chromosphere: Ellerman Bombs

In 1917, Ellerman discovered short-lived, intense brightenings at tiny (<1 arcsec) points in the lower solar CHROMOSPHERE. A typical Ellerman bomb lasts less than 10 min. Its spectrum is characterized by very broad emission wings on the hydrogen Balmer lines. On a photographic negative, the pattern of a bomb's emission resembles a mustache or whisker. Reinforcing the allusion to mustaches is the fact that a bomb's enhanced emission does not extend into the hydrogen line cores. The spectral line then resembles a stripe of shaved skin flanked by two tufts of dark hair. Emission at line center is lacking, presumably because the radiation is absorbed in the overlying atmosphere or in the bomb itself.

Mustaches are about 10% brighter than the continuum at 1 Å from the Hα line center. Because of the strong absorption at the line center, Ellerman bombs must be low in the chromosphere. This is an important difference from flares: all flares are brightest in the cores of the affected spectral lines. Flares are bright because the chromosphere is heated from the overlying corona. The topmost layers are affected first, whereas in bombs, the mustache shape of the spectrum means that the heating must originate in the chromosphere.

Bombs always occur along the boundaries of magnetic field concentrations, so they might be evidence for magnetic field reconnection. They are also interesting because of the resemblance of their extended emission wings to those found in some solar flares. Bombs, which are almost always present on the Sun, may shed light on the elusive mechanism of solar flares.

Bomb sites

Bombs are almost always seen in emerging flux regions (see CHROMOSPHERE: EMERGING FLUX REGIONS) (figure 1) and at the outer edges of SUNSPOT PENUMBRAE (figure 2). There may be as many as 100 bombs per spot group. Very high resolution pictures, such as figure 2, also reveal that Ellerman bombs occur on the network of photospheric faculae (see SOLAR PHOTOSPHERE: FACULAE). The network of bright points extends into the low chromosphere as seen in the figure, and it corresponds to the network of SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES in the photosphere. Elements of the network brighten about 3 min before Ellerman bomb onset. Thus the true energy source for the explosion may lie in the photosphere.

Ellerman bombs appear before and during the formation of sunspots. In the best images, bright features composed of elements smaller than 700 km diameter can be seen drifting outward from sunspot pores. Bombs probably always occur where magnetic features of one polarity meet other, opposite polarity, features. Roy found that bombs occur only where the magnetic fields are changing and estimated the average rate of flux change at 3×10^{15} Mx s^{-1} (where Mx s^{-1} $= 3 \times 10^{7}$ weber s^{-1}). Bomb frequency and number increase with growth rate of the magnetic field. When bombs occur in clusters, they attain maximum brightness within 2–3 min of each other.

Bomb visibility does not depend on distance from the solar limb, that is, they seem to have the same brightness when viewed from any angle. But near the solar limb bombs are elongated in the radial direction. One study gave the mean length as 800 km and the mean width as 350 km. The maximum brightness and maximum length were achieved at nearly the same time, but the bombs grew only in length at first.

The average lifetime of a bomb is probably less than 10 min, but some bombs last much longer. The longer-lived ones pulsate, so they might be a succession of explosions. Recurrence of bombs in the same location is common. Rust suggested that bombs occur at magnetic field reconnection points, while Kitai suggested instead that bombs occur in isolated flux tubes. An important clue to resolving the issue may be found in the association between bombs and surges.

Hα surges and bombs

Hα SURGES originate in bombs, and surge length (amount of material) is proportional to bomb size times the lifetime, according to Roy and Leparskas. Surges are clusters of very fine dark (sometimes bright) filaments connected to bombs (see SURGES). They consist of dense material apparently ejected from the chromosphere. The velocity averages about 100 km s^{-1}. The trajectories of surge threads above 5000 km match the magnetic lines of force calculated from models of current-free fields in sunspot regions.

No x-ray emission has been clearly associated with Ellerman bombs, but x-ray jets do accompany surges (see SOLAR: X-RAY JETS).

The link between surges and bombs is consistent with a model (figure 3) in which the bomb site lies above a *satellite sunspot*. Satellite sunspots are generally not visible in white light, but they stand out in maps of the magnetic fields surrounding major sunspots. Their polarity is opposite to that of the spot and most of its surroundings. In figure 3, the X marks the supposed site of magnetic field reconnection, and presumably of the bomb itself. One effect of the reconnection is to transfer chromospheric material from lower-lying fields into the higher families of field lines that stretch away from the sunspot. This would explain the surge of material from the bomb site. Numerical models suggest that reconnection should accelerate both hot (x-ray emitting) and cold (Hα emitting) plasmas.

Spectral properties

Severny was the first to study the spectra of mustaches extensively, and he showed that Stark broadening cannot explain the extraordinary extended line wings. Engvold and Maltby measured mustache profiles systematically, finding spectral line widths at Hα ranging from 3 to 10 Å. The true widths could have been greater, having escaped detection because of instrumental dilution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

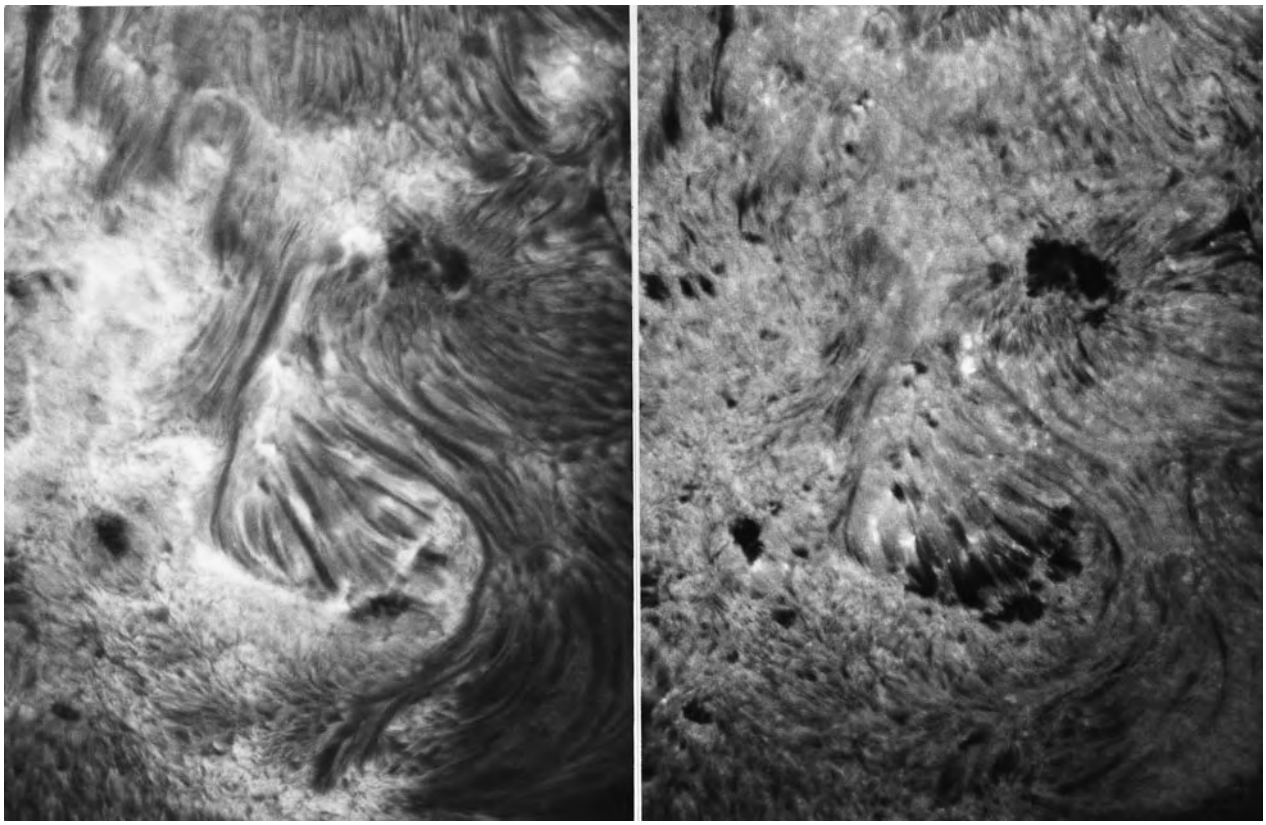


Figure 1. Ellerman bombs in a region where magnetic flux is emerging through the chromosphere. The left and right images were obtained simultaneously through a birefringent filter. The left image shows the upper chromosphere. The dark fibrils in the center are the telltale signatures of emerging flux. The lower chromosphere, on the right, has many bright point-like Ellerman bombs that do not extend into the upper chromosphere. (Sacramento Peak Observatory photograph.)

Bomb spectra are usually asymmetrical, with 60% being brighter in the blue and 40% brighter in the red. But the asymmetry could be due to the effect of overlying material, if there are steady upward or downward flows. During the first 2 min of development, the mean upward velocity deduced from mustache spectra is 8 km s^{-1}. This could account for the blue asymmetry. The mean upward velocity in the underlying photosphere is 1 km s^{-1}, which may indicate that the matter in surges is being supplied from the photosphere.

The spectra of bombs are weakly polarized. Babin and Koval at the Crimean Astrophysical Observatory found 24% linear polarization in the enhanced emission. They subtracted an estimated background from the bomb brightness before computing the polarization degree. From a study of Hα spectra taken simultaneously at four linear polarizations, Babin and Koval also concluded that the degree of linear polarization in any given mustache is the same over the entire emission contour but the orientation of the plane of polarization may be different for adjacent mustaches, and some large mustaches encompass differently polarized elements. Generally, they say the electric vector is perpendicular to the radial direction on the solar disk, but no physical process that could produce such a high degree of polarization was suggested. Another study, at the Sacramento Peak Vacuum Tower Telescope, did not confirm polarization greater than 2%. If a systematic pattern of polarization could be found, it might provide a crucial clue to the physical nature of bombs.

What are Ellerman bombs?

Although the close association of Ellerman bombs with magnetic boundaries suggests strongly that magnetic reconnection is the energy source for bombs and surges, many questions remain unanswered. It is very difficult to explain the great width of the spectral lines, especially since Severny showed that it cannot be due to Stark broadening. That would seem to rule out electric fields, and it casts doubt on the reconnection model.

Shklovskii proposed that the wide wings are Doppler broadened, which implies velocities of several hundred km s^{-1} for the hydrogen atoms. In Shklovskii's model, the radiating matter must be renewed several hundred times. Shklovskii proposed to achieve that with a fusillade of plasmoids from below the chromosphere. His model provides an effective excitation process and explains

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

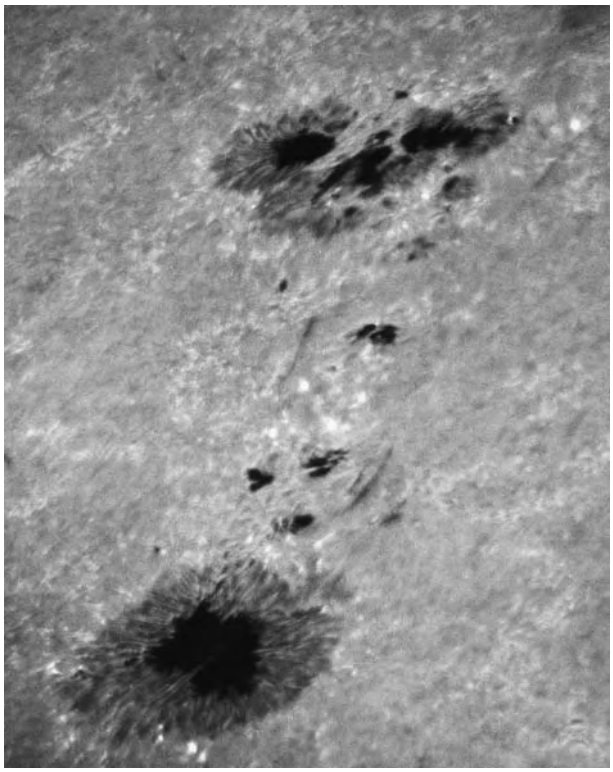


Figure 2. Ellerman bombs at the outer boundaries of sunspots and at the boundary between spots of opposite magnetic polarity (center).

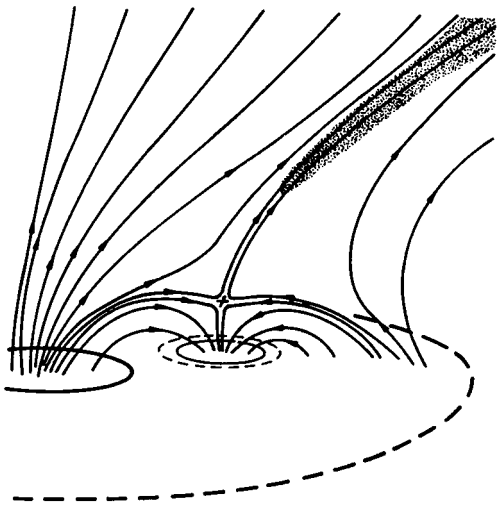


Figure 3. A model of the magnetic field lines above a satellite sunspot. The X marks the point where magnetic reconnection is supposed to heat and/or compress material to give the intense brightening of Ellerman bombs. The shaded feature represents a dark surge of material ejected from the reconnection region.

the broadening. The difficulty is in accounting for the acceleration of the plasmoids.

Another possibility is that the erupting material forming a bomb is driven initially by a Kelvin–Helmholtz fluid instability (see MAGNETOHYDRODYNAMIC INSTABILITIES) at the interface between the chromosphere and photosphere. The focus of this model is more on the steep density gradients and fluid flows at the outer boundaries of sunspots than on the magnetic field configuration. A fluid instability would raise a blob of dense material into the chromosphere where the magnetic fields in it might be pinched off, thus isolating the blob and heating it. The wide spectral line wings would be explained easily because an isolated gas with nearly photospheric density would have very large optical depth in the hydrogen line wings. The problem is to defy gravity and hold a heavy blob up above the photosphere for 5–10 min.

Although they are still not well understood, Ellerman bombs are not much studied today. At one time, it was believed that the extraordinarily wide wings are evidence of previously undiscovered physical processes, and that they are formed at very high densities below an optical depth of 0.4. Today it is understood that a variety of now-familiar phenomena might produce the conditions for broad line wings. Kitai developed a mustache model based on solar flare-like excess heating of 1500 K at heights of 700 to 1200 km above the level of optical depth unity. The principal distinction between his model and flare models is that the disturbance is confined between two chromospheric layers. To reproduce mustache line profiles Kitai had to assume a density enhancement of five in the affected layers, in addition to the heating. The magnetic fields where bombs occur are strong enough to contain the gas at the added pressure, but the problem is to find the heating mechanism. Magnetic reconnection is the best candidate, but no one has worked out a model of reconnection at chromospheric densities.

Bibliography

Canfield R C, Reardon K P, Leka K D, Shibata K, Yokoyama T and Shimojo M 1996 Hα surges and x-ray jets in AR 7260 *Astrophys. J.* **464** 1016

Kitai R 1983 On the mass motions and the atmospheric states of mustaches *Solar Phys.* **87** 135

Rust D M and Keil S L 1992 A search for polarization in Ellerman bombs *Solar Phys.* **140** 55

Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

D Rust

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Chromospheric Oscillations

By the term solar chromospheric oscillations we here refer to any periodic or quasi-periodic dynamic phenomena seen in the solar CHROMOSPHERE. We thus include both propagating and standing waves.

Information about chromospheric oscillations is given by the Doppler shift of spectral lines formed there. To avoid strong temperature sensitivity in the opacity it is also advantageous to study lines originating from the ground state of an element (called resonance lines). Unfortunately, there are only two such lines opaque enough to be formed in the mid chromosphere that are visible from ground based observatories—the two resonance lines from singly ionized calcium called the H and K lines. Much of our knowledge about chromospheric dynamics has therefore been derived from studies of these lines. The H and K lines appear as very strong absorption lines with a slight emission in the line center. This emission has a central reversal such that there are two slight emission peaks with a dark core in between. The peak on the violet (shortward) side of line center is called H_{2V}, the peak on the red (longward) side H_{2R} (and correspondingly for the K line). The calcium H and K emission outlines a pattern strongly correlated with the photospheric magnetic field, the so-called network (CHROMOSPHERE: NETWORK). The dynamic behavior is very different in the network patches and in the inter-network areas and these are therefore treated separately in the following.

Inter-network oscillations

In the inter-network the calcium emission is absent most of the time. When the emission appears it is almost always only the violet peak that appears and in a small area with a diameter of a few arcseconds. Jensen and Orral found in 1963 that the brightening in the emission peak is preceded by a brightening in the line wing that is formed in the photosphere and the phenomenon is often repeated a number of times in the same location with about a three minute period. The localized brightenings are called K_{2V} (or H_{2V}) bright points or preferably bright grains (to indicate that they have a resolvable structure). The quasi-periodic behavior is known as the chromospheric three minute oscillation.

The calcium bright point behavior has been explained as resulting from propagating acoustic waves that form shocks about 1000 km above the visible surface of the photosphere (see figure 1 and Carlsson and Stein 1997). Some aspects of the observed behavior turn out to be very robust diagnostic indicators of the dynamics of the chromosphere. The strongly asymmetric intensity profile with only the violet peak appearing is only possible if there are very strong velocity gradients around the formation height of 1000 km. Such strong velocity gradients typical of shock formation thus have to exist in the solar chromosphere. Acoustic waves steepening into shocks is also the expected behavior from the fact that the gas density falls off exponentially with height in the solar atmosphere forcing a disturbance to grow in amplitude with height in order to conserve energy.

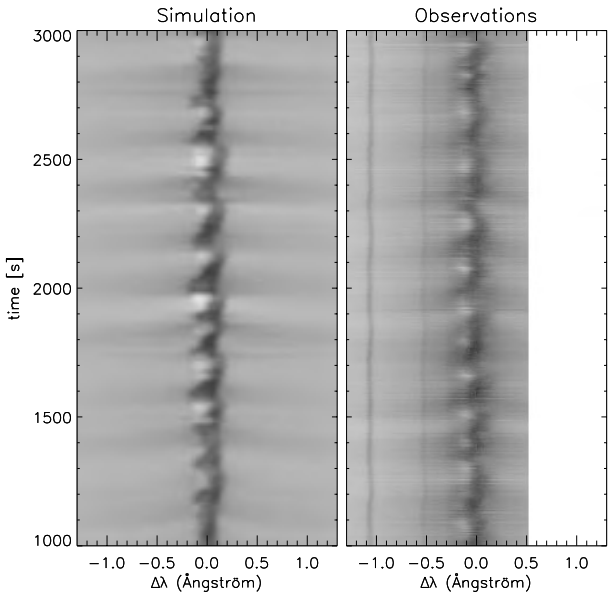


Figure 1. The computed Ca II H line intensity as a function of wavelength and time compared with observations. There is close correspondence between the simulated and observed grains.

The strongly dynamic behavior of the solar chromosphere manifested in the calcium bright grains has important consequences for our concept of the chromosphere. Processes like line formation, line emission and radiative cooling are strongly nonlinear. In a dynamic medium this means that such quantities have to be calculated from dynamic atmospheres; a semi-empirical chromospheric 'average' model (see SOLAR ATMOSPHERE: EMPIRICAL MODELS) will then be misleading.

How are the chromospheric oscillations excited? In a stratified atmosphere like the solar chromosphere acoustic waves can only propagate if they have a period shorter than a characteristic period called the acoustic cut-off period. This period is about 200 s in the solar chromosphere. The acoustic cut-off period is thus close to the three minute period of the dominant chromospheric oscillations. Waves with periods close to the cut-off period can be excited in several different ways and which one is most important in the solar case has been much debated. Long-period evanescent waves will excite wakes with the acoustic cut-off period. Merging of shocked waves with shorter periods will also preferentially favor the same period waves. A sudden pulse with a time scale smaller than three minutes will excite such waves as well. Lastly, in a spectrum of waves with many periods with most power at longer periods, all the evanescent modes will decrease in importance with height leaving the ones close to the cut-off period as the most important ones.

The issue can only be resolved using numerical simulations coupled with observations. From detailed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

simulations it is clear that the grain behavior is set by acoustic waves that at a height of 280 km above the visible surface have a period of around three minutes. At that height, the velocity field is dominated by the five minute oscillations (see HELIOSEISMOLOGY) but these waves are evanescent and their relative importance decreases with height. These simulations do not, however, give the excitation mechanism for the photospheric three minute period waves. Recently it has been shown (Skartlien 1998) that small collapsing granules, where upflow is turned into downflow in a time scale smaller than three minutes, may be the photospheric excitation mechanism for the calcium grains.

Many more chromospheric diagnostics than the resonance lines from singly ionized calcium are now available thanks to access to the ultraviolet (UV) part of the electromagnetic spectrum from instruments on board satellites and spacecraft. With the SUMER (Solar Ultraviolet Measurements of Emitted Radiation) spectrograph on board the SOlar and Heliospheric Observatory (SOHO) spacecraft much more observational material on solar chromospheric oscillations now exists. The very dynamic behavior of the lower inter-network chromosphere has been confirmed with intensity oscillations and Doppler shifts in general agreement with numerical simulations. The spectral lines formed in the upper chromosphere, however, are not well reproduced by one-dimensional dynamic models. The emission lines in the UV exist in emission all of the time at all locations indicating a mean temperature rise. This may be in the form of magnetic hotter elements overlying the non-magnetic inter-network regions. The UV continua and emission lines from neutral elements show grain-like behavior at all inter-network locations with dominating periods around three minutes. Lines from singly ionized carbon and silicon formed in the upper chromosphere show grain-line variations mainly in Doppler shift with intensity variations of much smaller amplitude than would be the case in a one-dimensional picture. Even a high-temperature indicator like five times ionized oxygen formed in the transition region often shows Doppler shifts well correlated with the oscillations in the lower chromosphere but again with the intensity controlled mainly by processes other than compressive acoustic waves. It is thus clear that the upper chromosphere contains influences from several physical processes. This is also expected from the fact that magnetic fields must start to play a dominant role in the upper chromosphere where the magnetic pressure is larger than the gas pressure. In the upper chromosphere it is thus probably not meaningful to speak about non-magnetic regions any more. The dynamic behavior observable in the calcium resonance lines is, however, evident throughout the chromosphere into the transition region.

The calcium H and K lines are the ideal diagnostic lines available from the ground but there are also other spectral features formed in the solar chromosphere that have received a lot of attention. Examples are the hydrogen Hα line, the calcium infrared triplet lines, the sodium D resonance lines and the helium line at 1083 nm. They all suffer from line formation complexities (like very temperature sensitive opacity, non-locally determined source function or influence from incident radiation field) making the diagnostic inference complicated. Velocities inferred from the calcium infrared lines, the calcium resonance lines and the helium 1083 line seem to be all in phase. In a traditional atmosphere these lines are formed at different heights and the zero phase difference then indicates the presence of standing waves. This has been interpreted as an indication of a chromospheric 'cavity' where reflection off the boundaries set up a standing wave pattern. This picture is, however, in conflict with other observations (like the clear signature of propagating waves seen in the calcium resonance lines).

In addition to the three minute oscillations, recent SUMER observations show oscillatory behavior at shorter periods. Waves with periods around 150 s are seen in upper chromosphere/lower transition region lines. These oscillations can be coherent over thousands of kilometers. They may be correlated with phase jumps observed in the lower chromosphere in the sodium D-line by Deubner and co-workers. The photospheric temperature drop is rather steep and may cause the reflection of waves. A steep temperature rise from the chromosphere to coronal temperatures may also cause partial reflections (around 70% reflection of planar waves in a one-dimensional simplified model). A chromospheric cavity between these partially reflecting may support a velocity field with modal structure even when the transmission coefficient is rather high and the chromosphere is dominated by propagating shocks. Due to the high phase speed of waves with periods just below the acoustic cut-off period of the chromosphere there is less sensitivity to the exact height of the reflecting layer. Such a 'partial' cavity may explain phaseshifts as seen in lower chromospheric lines, the existence of standing wave signatures and preferred periods around 150 s. Modes may also exist due to reflection of waves from shocks.

Network oscillations

The network shows a dynamical behavior very different from that of the inter-network regions. The H and K lines from singly ionized calcium show two emission peaks most of the time. The emission varies with time, but both the red and the violet peak may be missing and the time-averaged profile is quite symmetric. No pronounced three minute oscillations can be found but slower variations around a five minute period are present (Lites *et al* 1993).

SUMER observations give the same picture of mainly slower variations in the network. There are no reported phase relations that would support a picture with upward propagating acoustic waves. It is unclear whether the slow variations are manifestations of waves or of the buffeting of the network magnetic flux tubes by the granular motion. There are some SUMER observations showing three minute periods in network structures but much more rarely than in the inter-network.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Carlsson M and Stein R F 1997 Formation of solar calcium H and K bright grains *Astrophys. J.* **481** 500–14

Lites B W, Rutten R J and Kalkofen W 1993 Dynamics of the solar chromosphere I. Long-period network oscillations *Astrophys. J.* **414** 345–56

Skartlien R 1998 3D Modeling of solar convection and atmosphere dynamics *PhD Thesis* University of Oslo

Mats Carlsson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Chromospheric Plage

Chromospheric plage (from the French 19th century wording: 'plage brillante', bright plage or bright area) or facular plage (facula, Latin word for small torch) is a bright area seen in chromospheric lines most of the time surrounding one or several SUNSPOTS. Chromospheric plage is one of the main components of the solar activity (see SOLAR ACTIVE REGIONS).

Plages are the results of very complex magnetic processes. After a general description, the magnetic field properties of plages introduce the plage structure. Although very general and simplified, some insight into plage modelization is given.

General description

Chromospheric plages are the prolongation in the CHROMOSPHERE and the TRANSITION REGION of bright faculae seen in the photosphere (see SOLAR PHOTOSPHERE: FACULAE). The linear extension of a plage varies between a few tens of arcseconds to 3–4 arcmin. While the area covered by one plage may vary between less than 1 arcmin^2 and 16 arcmin^2, the total area covered by plages is between less than 1% and 17% of the solar disk area and is strongly related to the solar activity cycle.

In the visible, chromospheric plages are best seen in the strongest lines such as the red H I Hα 656.3 nm line and the blue Ca II H and K 393.4–396.9 nm lines. In the ultraviolet the chromospheric H I Lα 121.6 nm dominates the spectrum, but plages produce enhanced intensities in the entire set of chromospheric emission lines (optically thin or thick). In all the ultraviolet transition region lines the intensity in plages is also strongly enhanced by a factor of 2–20. Chromospheric plages can also be observed in the submillimeter wavelength range (100 μm–1000 μm). In this range, if a 5400 K temperature is assigned to the quiet Sun brightness in the low chromosphere, the corresponding plage brightness can attain up to 6500 K.

Chomospheric plage is the bright part of the chromospheric active region. In the low transition region, where sunspots are not seen, plage fills the whole space of the active region as loops of different temperatures. So, the plage represents an intermediate stage of the magnetic field density, where the flux tube density is not high enough to block the transfer of radiation but large enough to be seen as emissive loops.

The complexity of intermixing of plages with sunspots and filaments in the chromosphere and the transition region is illustrated in figure 1. The fine structure of chromospheric plages observed on high-resolution spectroheliograms shows a fibril-like distribution (e.g. in Hα, Lα) or a distribution of fine grains (faculae) in Ca II K1 (low chromosphere).

The plage area in Ca II K is used as an index of activity and as a proxy component for the modeling of the solar irradiance variation (three components are used to model solar activity variation: sunspots, plages, network).

Magnetic field

Direct measurements of the magnetic field in chromospheric plages have been performed in the H I Hβ 486.1 nm line by using the Zeeman effect. The longitudinal field can reach 80–100 G (1 G = 10^{-4} T) near disk center. New measuring techniques using Stokes polarimetry in Ca II K lines are beginning to provide chromospheric magnetic field values.

The magnetic field of the plage is not homogeneous. It is structured in tubes with a mean flux of 2.5×10^{18} Mx. The filling factor of flux tubes is within the 5–15% range in photospheric layers; this value can be compared with the less than 1% filling factor in a quiet area. From vector magnetogram measurements a gradient of around -3 G km^{-1} with height has been inferred, starting from 1400 G in the middle photosphere. Higher in the atmosphere the magnetic fields from an individual tube spread rapidly with height, producing a 'canopy' of much lower field strengths above the photosphere.

At any time on the Sun a significant part of the magnetic flux defined as the weak fields (flux density below 25 G) comes from a dispersing active region.

There is a strong correlation between the brightness of the Ca II chromospheric plages and the magnetic field strength measured in the photosphere: the Ca II plage bright patches overlie areas of strong magnetic field. The relation between Ca II K intensity and magnetic flux density (measured at photospheric level) holds from the quiet Sun cell network up to the brightest plage. Statistically, the mean trend between Ca II K excess flux density and the magnetic flux density is a power-law relation with an exponent smaller than unity (about 0.6).

The total magnetic flux in a plage is of the order of 10^{22} to 10^{23} Mx. With or without spots the plage magnetic field is bipolar.

The fibril-like structure is related to photospheric magnetic field tubes, filled with matter at relatively low pressure (inside each tube the gas pressure is added to the magnetic pressure to balance the external pressure) and expands and bends with height in a canopy-type structure (see CHROMOSPHERE: MAGNETIC CANOPY).

Plage structure

The plage is the basic element of an active region: plage appears before the sunspots and may disappear long after the sunspots' disappearance. The overall structure of an active region is determined by both the structure of the emerging flux loops and the sign and strength of the fields already present on the surface when it emerges. A plage appears as an emerging flux region sometimes within or near an already existing active region and sometimes in a quiet region. A few hours or a day later, in most of the plages, sunspots emerge within the bright bipolar region. After 3–4 days, the plage has reached its maximum extent (up to 5 arcmin or about 200 000 km across) and intensity in Hα. During the following days (5–15 days) the activity in the plage grows and reaches its maximum. In most cases, after one solar rotation the sunspots disappear and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

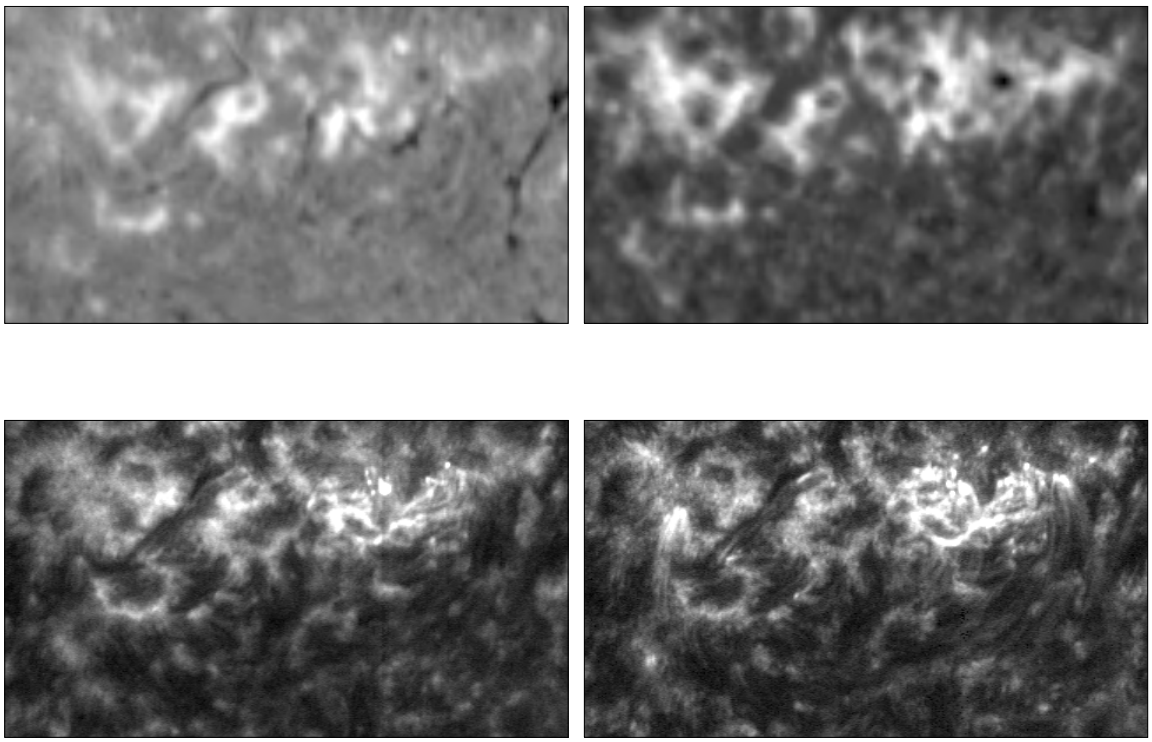


Figure 1. The same active region seen in several lines (8×4.5 arcmin^2). The top left and right are Hα and Ca K spectroheliograms from Sacramento Peak synoptics data (3.8 arcsec pixel size) taken on 13 May 1996 near 13:40 UT. The bottom left and right are simultaneous hydrogen Lϵ and S VI (93.3 nm, 2×10^5 K) spectroheliograms (1 arcsec pixel size) recorded by the SUMER spectrometer on the SOHO ESA/NASA spacecraft taken on the same day near 3:00 UT.

the plage brightness begins to decrease while the size may continue to increase. If the plage is not reactivated by the emergence of a new flux region, the brightness continues to decrease and progressively the plage dissolves in the surrounding network after a few solar rotations (2–4 rotations).

While it is not clear why a plage (or emerging active region) appears in a specific location on the solar surface (although preferred longitudes seem to exist where the plage may emerge), detailed observations of the disappearance phenomenon have been performed. The characteristic diameter of a plage network (or the center-to-center supergranular cell distance) is typically 10 000 km–15 000 km, a factor of 2 smaller than the corresponding quiet Sun network diameter. The following scenario, which combines several processes, has been proposed to explain the decay of magnetic plages:

- from observations using magnetograms and Dopplergrams it has been inferred that flux cancellation often occurs along the cell's boundary when two elements of opposite polarities have approached each other;
- other observations of plage decay suggest that motions along supergranular lanes contribute strongly to the random walk of magnetic features;
- it has also been suggested that retraction, a process that is the reverse of the emergence of magnetic flux, controlled by subsurface motions, contributes substantially to the decay of plages.

At the edge of the plage, where the magnetic flux density is lower than 50–80 G, because of random-walk motion of flux tubes, the plage flux leaks into the quiet region while some flux tubes coming from the quiet region are captured at the plage periphery. The combination of these two processes may explain the slow decay of most of the plages. Some data suggest that less than half of the plage flux leaves the active area; the largest part disappears *in situ* by reconnection, cancellation or submergence.

Plage flux tubes are filled with matter in motion. From the chromosphere to the low transition region (6000 K $\leq T \leq 2 \times 10^5$ K) the velocities increase from a few km s^{-1} to 25–30 km s^{-1}. There is a systematic redshift detected over plages in the chromosphere transition region. It is more important than in the quiet Sun and the velocity differential, compared with the quiet region, peaks at 7 km s^{-1} near 10^5 K. In the transition region, in lines with a maximum of the ionization temperature near 10^5 K, along individual loops, velocities reaching up to 60 km s^{-1} have been measured (e.g. from SOHO, the Solar and Heliospheric Observatory). These results must be confirmed by several other measurements to evaluate the importance of flows

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in plage loops. There are still uncertainties in the interpretation of such Doppler velocities:

- either the flow is coming up from one loop leg and returning down to other loop leg,
- or the apparent Doppler shift is due to the propagation of sound waves, Alfvén waves, or magneto-acoustic waves along the loop magnetic field.

Few measurements of chromospheric plage oscillations have been done. Although 5 min intensity and velocity oscillations in phase over an extent of 10–20 arcsec have been reported, the amplitudes are smaller than in the quiet area and an increase of power has been noticed for periods shorter than 100 s.

The complexity of active regions, a mixture of very compact clusters of magnetic flux tubes (sunspots) and more dispersed flux tubes (plage) of several polarities, arises from the mutual interaction of new flux loops rising within and near old ones. Within a few hours the whole magnetic flux tube distribution is renewed.

Modeling of plage

The solar abundance in the quiet solar photospheric region is used as a reference for abundance determination in different structures at several temperatures. Measurements made in using the intensity of lines formed in the 2×10^4–10^5 K temperature range indicate a possible enhancement of low first ionization potential (FIP) (smaller than about 10 eV) elements compared with the photospheric abundance in plage. The presence of electric fields in this highly dynamic atmosphere has been proposed as the cause of the enrichment of low-FIP elements.

Semiempirical models of plage in one or two dimensions have been built. Such models provide a first estimate of temperature and density distribution as a function of an average height. From the top of the photosphere the plage temperature is higher than the quiet Sun temperature and, above the minimum the temperature gradient is increasing faster, giving a more compact chromosphere. In the chromosphere the electronic density is higher than the corresponding quiet Sun density. While in the quiet Sun the chromospheric plateau has an electronic density of a few 10^{10} cm^{-3}, the chromospheric plage plateau reaches a few 10^{11} cm^{-3}. At the top of the chromosphere the total pressure is greater than 1 dyn cm^{-2}, nearly 10 times the pressure of the quiet Sun. On average the plage chromospheric radiative cooling is about 5 times that of the quiet Sun.

Several dynamical models have being proposed to explain the emergence and decay of active regions (see CHROMOSPHERE: EMERGING FLUX REGIONS).

Bibliography

Article in review journal:

Zwaan C 1987 Elements and patterns in the solar magnetic field *Ann. Rev. Astron. Astrophys.* **25** 83–111

Journal articles:

Schrijver C J 1989 The effect of an interaction of magnetic flux and supergranulation on the decay of magnetic plages *Sol. Phys.* **122** 193–208

Seiden P E and Seiden P E 1996 Solar active region as a percolation phenomenon *Astrophys. J.* **460** 522–9

Philippe Lemaire

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Constant

The amount of radiant solar energy per second arriving perpendicularly on an area of one square meter, at a distance of 1 astronomical unit from the Sun. It is equivalent to the average flux of solar radiation at the top of the Earth's atmosphere. The measured value of the solar constant is 1368 W m^{-2}.

Because the Earth travels round the Sun in an elliptical orbit, its distance from the Sun ranges from 1.0167 AU to 0.9833 AU. Consequently, over the course of a year, the flux of solar radiation at the top of the atmosphere varies about its mean value by about $\pm 3\%$. Precise measurements of the solar constant have shown that the luminosity of the Sun fluctuates on time-scales of days and weeks by up to 0.2%, mainly as a result of changes in the numbers of dark spots or bright faculae on its surface. Long-term measurements suggest that there is a general variation in solar luminosity that appears to follow the solar cycle, the Sun being about 0.08% more luminous at solar maximum (when there are more active regions on its surface) than at minimum.

See also: active region, faculae, sunspot cycle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Corona: X-ray Bright Points

X-ray (or coronal) bright points are small knots of intense x-ray emission scattered randomly throughout the solar CORONA. They are associated with pairs of opposite-polarity magnetic flux seen below in the SOLAR PHOTOSPHERE and are most likely to be heated by the motion of the magnetic fluxes causing reconnection in the corona.

Observations

In the late 1960s and early 1970s rocket-borne X-RAY TELESCOPES capable of resolution and pointing stability in the arcsecond range imaged the Sun for between 5 and 15 min per flight. These images taken using very similar grazing incidence techniques to those used today revealed the complex beauty of the solar corona and revolutionized our understanding of it. One particular rocket flight on 8 April 1969 took longer time exposures than previous flights and revealed for the first time hundreds of small faint spots of emission scattered randomly across the disk; they were named x-ray bright points. However, it was not until the SKYLAB mission provided consecutive images over a period of hours and days that the characteristics of bright points could be studied. From these images it was found that x-ray bright points are made of a diffuse cloud of emission which grows at about 1 km s^{-1} up to approximately 22×10^3 km in diameter and contains an intense bright core about $(3–4) \times 10^3$ km across. Their sizes were determined to range from 5×10^3 to 20×10^3 km, however, the lower bound to this range was limited by telescope resolution and the upper bound was arbitrarily chosen so that the bright points could be distinguished from ACTIVE REGIONS with which they appeared identical from coronal images. It was found that the lifetimes of bright points varied widely from 2 h to 2 days which is considerably shorter than for active regions. Recent high-cadence observations, however, have discovered x-ray bright points with lifetimes as short as a few minutes. These bright points are normally just a few arcseconds across (2×10^3 km) and are known as micro bright points.

X-ray bright points have fairly low temperatures of about 2×10^6 K but high electron densities of around 5×10^9 cm^{-3}. They appear randomly distributed throughout the corona in coronal holes, quiet regions and active regions. Approximately 400 are present at any one time at solar minimum with approximately 1200 appearing each day, although these numbers vary depending on the wavelength, cadence and resolution of the telescope used.

Comparing the first rocket images with magnetograms it was found that bright points were associated with pairs of opposite-polarity magnetic fragments. Since ephemeral regions, pairs of newly emerged opposite-polarity flux, had just been discovered it was natural to assume that bright points were caused by the new emergence of a small loop in the corona. However, studies of the long-term variation of bright points revealed that their number decreased near solar maximum, suggesting that they were out of phase with the solar cycle. This was in complete contradiction to the observations of ephemeral regions which were found to be in phase with the solar cycle.

This discrepancy remained a mystery for many years basically because after Skylab soft x-ray images could only be taken on board rocket flights and so further comparative studies could not be made until it was discovered that He 10830 dark points observed by ground-based telescopes were fairly good proxies for the 30% largest bright points. A ground-based study comparing He 10830 images and magnetograms then came up with an explanation in 1984. This detailed study of time sequences of magnetograms revealed that most pairs of opposite-polarity flux below dark points were in fact converging not diverging and that the pairs of magnetic fragments were mutually losing flux. These new photospheric phenomena were called cancelling magnetic features and are believed to be related to the amount of mixed-polarity flux in the photosphere which decreases by a factor of 6 from solar maximum to solar minimum. Thus, like x-ray bright points, cancelling magnetic features were also found to be out of phase with the solar cycle. Indeed, recent studies show that about 80% of bright points are associated with cancelling magnetic features and just 10% with emerging flux regions. The remaining 10% are mysteriously related to unipolar regions of flux, although with higher-resolution instruments it may become apparent that these unipolar regions actually contain weak opposite-polarity regions of flux after all.

Interestingly these recent studies also suggest that bright points are in fact neither in nor out of phase with the solar cycle, but that their number remains approximately constant over the whole cycle. The change in opinion has come about through results from soft x-ray telescopes with much less scatter than Skylab which reveal that the intense emission from active regions can obscure the faint bright points during periods of increased activity on the Sun and so make it appear that at solar maximum there are far fewer bright points.

A study in the late 1970s using a time series of Fe XV images from the S-082A instrument revealed that x-ray bright points consist of small loops which vary in intensity on a timescale of about 6 min. The loop structures were found to be quite complex in nature and had loops of different temperatures. This was believed to be consistent with the repeated impulsive heating of the loops followed by their subsequent cooling before another pulse of heating. It was also found that sometimes these fluctuations could increase by a factor of 10–100. In such cases the bright point was said to have flared. These microflaring events tend to occur in about 10% of bright points and generally last from 5 to 10 min. Occasionally when x-ray bright points flare they produce long hot loops that grow up to hundreds of megameters long at speeds of 1000 km s^{-1}. These events are known as SOLAR X-RAY JETS. In bright points the jets are mostly two winged.

X-ray bright points and bright point flares have typical energy emissions between 10^{26} and 10^{29} ergs and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

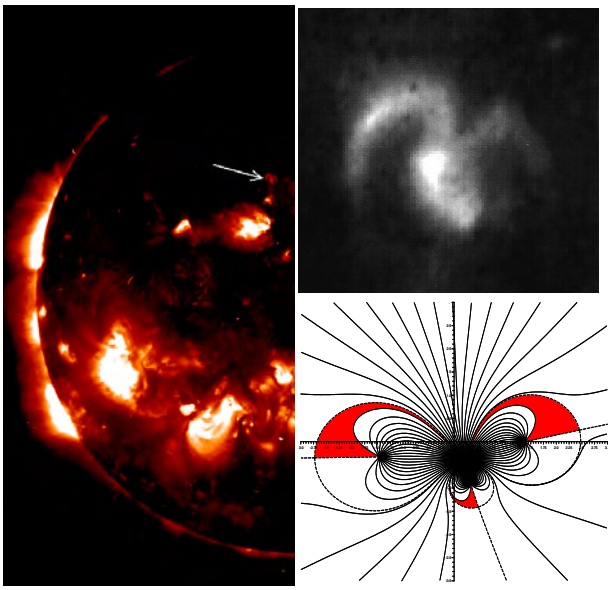


Figure 1. Left: half of a full-disk soft x-ray image taken by NIXT. The arrow indicates a bright point which is enlarged at the top right. Bottom right: the magnetic field due to the fragments below the bright point. The shaded regions are the newly reconnected field lines that give rise to the bright point.

so are the size of microflares. Indeed, signatures of microflares can be identified in the GOES data coincident in time with some of the larger bright point flares. Bright points also produce radio emission that can be both thermal (e.g. 17 GHz, 20 cm, etc) and non-thermal (e.g. type III bursts) suggesting that there are both open and closed field lines and that whatever powers bright points is not just heating them but is also accelerating particles.

An interesting phenomenon relating to cancelling magnetic features has been observed in Hα images: tiny filaments form above the cancelling magnetic feature. They typically last between 10 min and a few hours before they erupt. It is estimated that about 10% of cancelling magnetic features may have associated filaments, but there have, as yet, been no studies relating bright points or bright point flares to these filaments or their eruption.

Further studies of x-ray bright points using high-resolution soft x-ray telescopes (for instance, NIXT and TRACE) have managed to resolve the structure of bright points. Many are just single loops, but others are more complex and have a multiple loop structure (figure 1). In general, the complexity of the x-ray emission is related to the complex mixture of magnetic polarities below.

Theory

It is believed that x-ray bright points are heated by driven reconnection: the motion of footpoints of magnetic loops that are pushed together, interact and reconnect in the corona. The generally accepted model is the converging flux model proposed in 1994 which predicts in a natural way the key observations relating to x-ray bright points and their associated cancelling magnetic feature. In this model the interaction of two opposite-polarity magnetic fragments in an overlying ambient field are modelled in three stages: the pre-interaction phase, the interaction phase and the cancellation phase (Figure 2). First, in the *pre-interaction phase* of the model, a positive and a negative fragment are situated in a uniform overlying field such that they are unconnected magnetically, analogously to the observation that before opposite polarity flux pairs in a cancelling magnetic feature start to cancel no fibrils are observed linking them.

These opposite-polarity flux regions are seen to converge slowly at a typical speed of 0.3 km s^{-1}, which compared with the Alfvén speed of 100–1000 km s^{-1} is very slow; thus in the model the magnetic field is assumed to be evolving through a series of quasi-static equilibria as the fragments converge. The observations show that at some point during the convergence a bright point will appear (not necessarily directly) above the pair of opposite-polarity flux fragments. This sudden appearance of the bright point marks the start of the *interaction phase* and is explained naturally in the model as the point at which the fragments first interact magnetically at a neutral point which forms in the photosphere. The field is assumed to continue to evolve through a series of equilibrium states as the fragments converge even more closely; thus reconnection must take place at the neutral point. The reconnection causes field lines to be 'broken' and new field lines to be formed. Although initially the neutral point and therefore the energy released in the model start out at the level of the photosphere as the fragments continue to converge the neutral point rises up into the corona. Newly reconnected field lines are believed to be injected with hot dense plasma which creates the increased emission observed in soft x-rays as the bright point.

Observations show that bright points start before cancellation starts and may even end before cancellation starts. In the model this is also true: at the point where the fragments first interact magnetically they are not actually losing flux, merely changing their flux connections. The bright point will continue until the two fragments are completely connected magnetically. This point may be reached before or after the two fragments actually touch one another in the photosphere. At this point the *cancellation phase* begins even though the interaction phase may not have ended. Flux is lost mutually from each fragment through reconnection in the photosphere. This process can continue until either the smallest fragment has been completely cancelled or the flows driving the two fragments stop driving them together.

The converging flux model clearly explains the 80% of bright points that are associated with cancelling magnetic features. The 10% associated with the emerging flux regions can also be explained by just considering the stages of the converging flux model in reverse (i.e. considering a diverging flux model). The cancellation phase if run backwards is analogous to an emerging phase where two new fragments grow in strength and diverge, as observed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

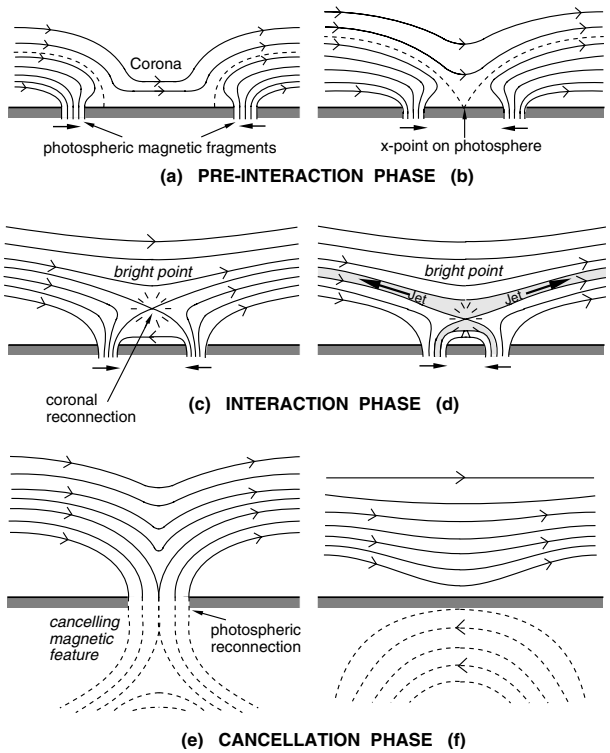


Figure 2. Converging flux model. (*a*) and (*b*) show the *pre-interaction phase* where the fragments are initially unconnected. (*c*) and (*d*) show the *interaction phase* in which reconnection creates a bright point in the corona. (*e*) and (*f*) show the *cancellation phase* where photospheric reconnection forms a cancelling magnetic feature.

This may or may not, depending on the circumstances, overlap with the interaction phase which as before gives rise to reconnection in the corona that creates the bright point. This time, however, the reconnection is driven by the divergence of the fragments. The bright point ends when the two fragments are completely unconnected magnetically.

The reconnection that powers the bright points is driven by the convergence or divergence of the two opposite-polarity magnetic fragments and is likely to be of an impulsive bursty nature. This means that the reconnection will not continue in a steady manner but is likely to take place in frequent, but random, spurts that occur every few minutes. It is unclear whether the flaring stage of a bright point is just a very large sudden random release of energy or whether it is some completely separate event like the eruption of a mini filament.

In figure 1(*a*) an arrow indicates a three-looped bright point that resembles a bird in flight (figure 1(*b*)). The philosophy of the converging flux model was used to model this bright point by comparing it with images of the magnetogram below to discover the magnetic flux regions associated with this structure. These revealed an imbalance of flux with one positive and three negative fragments. The corresponding magnetic field topology of these four fragments was modelled as it evolved according to the observations through a series of equi-potential states. The newly reconnected field lines that formed are those shaded in figure 1(*c*) and compare very favorably with the structure of the bird bright point observed. Because of the imbalance of flux in this situation there would also have been newly reconnected open field lines that would not brighten since the heat flux along very extended loops would be small. They may, however, have enabled non-thermal radiation to be observed from particles accelerated along them.

Bibliography

Forbes T G 1994 'Fireflies' on the Sun *Nature* **369** 278

Golub L and Pasachoff J M 1997 *The Solar Corona* (Cambridge: Cambridge University Press)

Priest E R, Parnell C E and Martin S F 1994 A converging flux model of an X-ray bright point and an associated cancelling magnetic feature *Astrophys. J.* **427** 459–74

Clare E Parnell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Coronal Mass Ejection: Observations

The Sun ejects matter into space. There appear to be two basic ejection classes, the first leading to the SOLAR WIND outflow, and the second leading to the occasional ejection of huge, discrete clouds known as coronal mass ejections (CMEs). First identified in the early 1970s, a typical CME can carry 1 000 000 000 t of plasma into space as part of the Sun's atmosphere erupts. The CME source regions within the Sun's atmosphere can be hundreds of thousands of km across and the resulting CMEs can expand into space at many hundreds of km s^{-1}. The rate of CME eruptions varies with the SOLAR CYCLE, but not by a large factor; one might expect about one CME event per day. CMEs represent a very significant disturbance to the solar wind. Given their size and mass, combined with the fact that the expanding clouds carry a frozen-in magnetic field, such events can engulf the Earth system and their arrival at the Earth can generate significant geomagnetic storms.

Observation of activity in the corona

In visible light the Sun's CORONA, or outer atmosphere, at its brightest, is a million times less intense than the solar disk. With such a huge contrast only rather special techniques can be used to observe the corona, effectively by removing the glaring solar disk emission. Indeed, one such technique is provided by nature.

During a total solar ECLIPSE the Moon occults perfectly the Sun's disk and the corona can be seen with the naked eye, as a system of milky, diffuse closed-loop features and streamers extending well into space. However, at best we see just a few minutes of totality and thus this only gives a brief glance. We may view the structure of the Sun's corona during the eclipse, but we cannot rely on such observations to observe the dynamic nature of the corona. Having said that, we can make use of an eclipse-like technique using an instrument called a CORONAGRAPH.

The bulk of our knowledge of CMEs comes from visible or 'white light' coronagraphs. Such a coronagraph is basically a telescope which contains an externally mounted or internally mounted device for occulting the solar disk. Obscuring such a bright, extended source does introduce difficulties with scattered light within the instrument; one can never produce as good an occulting disk as the Moon. However, modern coronagraphs can observe to within a tenth of a solar radius of the edge of the solar disk. In addition, because of scattered light from the Sun's disk in the Earth's atmosphere, one must operate such instruments at high altitude or from spacecraft, that is, above most of the atmosphere or outside the Earth's atmosphere.

Given the ability to produce an artificial eclipse one can monitor the Sun's atmosphere and study its structure and evolution at will, and this allows the study of activity in the Sun's corona, such as the CME eruptions.

The plasma in the solar corona is optically thin. This means that any observation along a line of sight is the integrated intensity of light emitted along that line of sight. Thus, we observe three-dimensional features projected onto a flat image—the plane of the sky.

The observation of CMEs is further complicated. The Sun's corona radiates in visible light because the light from the photosphere (the visible solar 'surface') is scattered off free electrons which are concentrated in the magnetic structures of the corona, and this is true for CMEs. The geometrical factors relating to the scattering process mean that coronagraphs are best suited to the observation of features over the limb or edge of the Sun. Thus, a CME ejected above the limb would, in general, appear to be brighter than one ejected out of the plane of the sky.

Another feature of the Sun's atmosphere allows further refinements to the coronal observing techniques. The Sun has a so-called K-corona and F-corona. The K-corona is the 'electron' corona; this is the true solar corona, i.e. radiation from the electrons contained within the structures in the Sun's atmosphere. The F-corona is effectively the near-Sun dust environment which is illuminated by solar radiation. For the analysis of the solar coronal structure and evolution, including CME activity, we wish to observe the K-corona. The K-corona is brighter than the F-corona close to the Sun, but, beyond about 2 solar radii, the F-coronal radiation begins to dominate. Thus, one has to separate the two and here we can make use of the fact that emission from the K-corona is heavily polarized. Indeed, this fact can help us to observe the coronal features against unpolarized scattered light such as in the Earth's atmosphere.

The first space-borne 'white light' coronagraph was flown in ORBITING SOLAR OBSERVATORY (OSO) 7 in 1973. With this device, the first CME observations were made, though it has been suggested that eclipse photographs from 1860 demonstrate that a CME was in progress, although no-one at the time could have identified it as an ejection. After OSO 7, extensive observations of CMEs were made using space-borne coronagraphs aboard SKYLAB (1973–4), P 78-1 (1979–85), SOLAR MAXIMUM MISSION (1980 and 1984–9) and SOHO, the Solar and Heliospheric Observatory (1995 to date). In addition, ground-based observations of CMEs have been made for many years with a similar device, called a K-coronameter, which measures the polarized brightness from the K-corona, at the HIGH ALTITUDE OBSERVATORY's Mauna Loa Solar Observatory in Hawaii.

The anatomy of a coronal mass ejection

The most basic definition of a CME is that it represents a discrete ejection of mass from the solar atmosphere detected as a transient feature expanding outward through a coronagraph field of view. CMEs represent a significant, rapid restructuring of a large portion of the solar atmosphere as the mass is ejected. CMEs occur in all shapes and sizes, and with many topological variations, but some basic characteristics are quite common. In this section we try to describe a 'typical' CME event, but draw attention to some features that vary from event to event.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

First, to put the CME story in perspective we must remember that the solar atmosphere is an environment where the interplay between the solar plasmas and magnetic fields produces a complex hierarchy of magnetic loops containing plasmas at temperatures up to a few million kelvins. The loops are rooted in the body of the Sun and move with the surface plasma, which is subject to differential rotation as well as the turbulent effects of the convection patterns below the surface. The net result is an environment where magnetic field systems can be driven to a great deal of complexity; the motions of the solar surface literally weaving the magnetic field lines into highly stressed patterns. Eventually, one could imagine some kind of breakdown or rapid expansion of magnetic fields, where the response to further motion actually drives magnetic fields outwards and this becomes the CME, i.e. the ultimate result is the ejection of mass from the Sun. We know that SUNSPOTS, flares and other forms of solar activity are all driven by the Sun's magnetic fields and the CME is another form of solar magnetic activity. In short, one can state that it would be a logical process for the Sun to expel magnetic fields in the way described, and that is the CME. Having said that, the actual process by which the eruption occurs, the reasons for the topology we see and the prediction of mass ejections are all beyond our understanding at the moment.

The first view one might get of a CME is the expansion of a bright loop system rising into the solar corona from behind a coronagraph occulting disk. This can appear as a sharply defined single loop or a highly structured loop system. As the loop system expands outward one often sees a cavity in the region behind the loop, that is, the loop system is leading a region of depleted density. Within the cavity, for many CME events, one might find the ascending features of a prominence or filament which erupts as part of the CME event.

An example is shown in figure 1, taken using the coronagraph aboard the Solar Maximum Mission. This CME occurred on 14 April 1980 and it is shown here in a sequence of four images shown in negative form. The images were taken at 04:48, 05:44, 06:10 and 07:09 UNIVERSAL TIME (top left to bottom right) and in each image one can see streamers extending from the disk to the top and left of the image. These persist throughout the ejection event. In the images the Sun is tilted at 45°, with the north pole to the upper left, and the occulting disk (white in the negative) is 3 solar radii across.

The first image (top left) shows the CME loop just ascending above the occulting disc, apparently over the northern polar region. In the second image (top right), 56 min later, the CME loop has ascended further, to half way to the top left corner of the image, and a second loop can be seen. The second loop is an ascending prominence and it actually lies in a density 'cavity' which occupies the volume behind the outer loop. The third image (bottom left), taken only 26 min later, shows significant outward expansion, and, in the last image (bottom right), the CME outer loop has left the field of view and the prominence can just be made out.

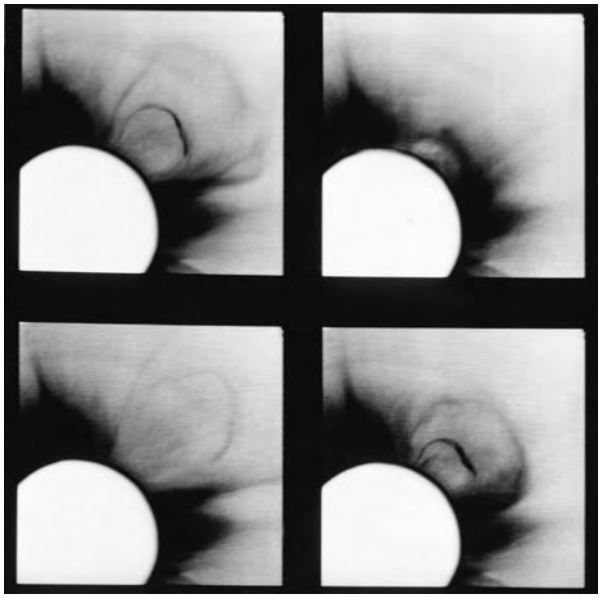


Figure 1. Four images of the CME of 14 April 1980 detected by the Solar Maximum Mission coronagraph (courtesy of High Altitude Observatory).

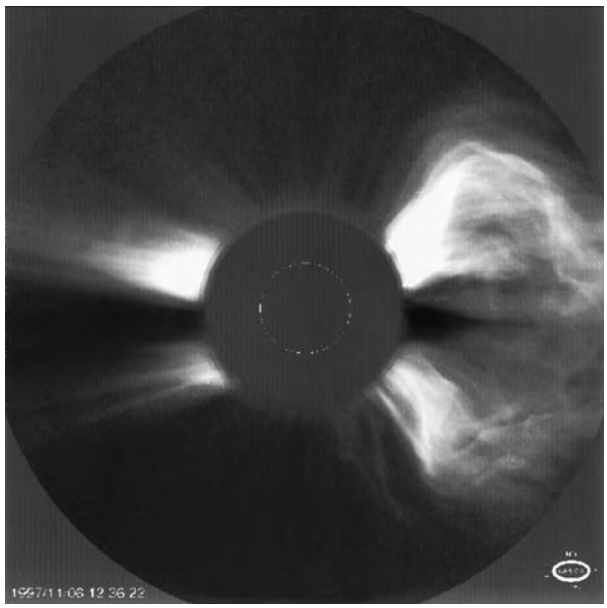


Figure 2. The CME of 6 November 1997 detected by the coronagraph on the Solar and Heliospheric Observatory (courtesy of Naval Research Laboratory).

A second CME event is shown in figure 2, this time taken by the Solar and Heliospheric Observatory coronagraph on 6 November 1997. This event shows a more complex outer loop system expanding out from the Sun, and no clear sign of a following prominence. In

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

this event the legs of the CME are particularly bright and the outer portions show a dramatic amount of hair-like structure. In this case, the occulting disk is also 3 solar radii across, and the solar disk is denoted by the white circle.

Typically a CME spans some 45 heliographic degrees of the solar disc, with clear legs which remain fixed over the same location of the Sun as the CME expands outward. However, CMEs have been reported with widths of less than 10°, and CMEs which are apparently directed toward or away from the Earth can appear to expand like a halo around the entire solar disk. Thus, although the average CME span is about 45°, and about 60% of CMEs have spans between 30° and 60°, CMEs of almost all spans are found.

Measurements of the speed of the outer loops of CMEs give typical outward speeds of 300–400 km s^{-1}, although CMEs have been detected with speeds as high as 2000 km s^{-1} and as low as just a few tens of km s^{-1}. The majority of CMEs show a modest acceleration with altitude from the solar surface, or a constant velocity. Very few display deceleration with height. The ascending prominence or filament material always follows the outer loops at a slower velocity. Since the escape velocity at the surface of the Sun is over 600 km s^{-1} the ejection process, at least in most cases, cannot involve the ballistic flight of a cloud which is launched by an impulse at a single point in time. The constant velocity or modest acceleration properties of CMEs, combined with the fact that they leave the Sun and can travel out as slowly as a few tens of km s^{-1}, suggest that the ejection expands outwards under the influence of a continuous process.

Statistical analyses indicate that 75% of CMEs contain prominence or filamentary material as part of the ejection. Figures 1 and 2 show an example of a CME with and without a prominence eruption. The prominence or filament can sit symmetrically within the CME loops (as with figure 1) or can appear to lie asymmetrically within the CME. The prominence can rise as a single entity within the CME but many events show partial prominence eruption under the CME.

The scattering process which allows the observation of CMEs also allows us to estimate the amount of mass contained in the CME, by measuring the intensity of the CME material, and many estimates have put typical CME masses at about 10^{12}–10^{13} kg, i.e. 1 000 000 000–10 000 000 000 t, although once again there must be a huge spread of masses.

Many CMEs occur underneath or within pre-existing coronal helmet streamers (see CORONAL STREAMERS). For a period which may last for some days the streamer can show signs of the impending ejection by displaying a gradual 'swelling'. The ejection itself can disrupt the streamer which may form again afterwards, although in the aftermath of a CME, for some days, the coronal region from which the CME erupted can appear to be swept out. Whether or not the CME 'cloud' is magnetically connected to the ejection site for some days after the ejection itself is open to debate.

We know that prominences can erupt as part of the CME process, but other forms of solar activity may accompany the event. For example, many CMEs involve SOLAR ACTIVE REGIONS; either the CME can appear to erupt from an active region or an active region lies at one footpoint of the CME event. For such active region related CMEs the ejection process can be associated with flare activity. However, detailed studies of the relative timing of flare onsets and CME onsets show quite clearly that the flare onset can occur well before, coincident with or after the CME onset. The only thing which is clear is that there is an association. The location of a CME associated flare can be anywhere under or near the span of the CME, and frequently in one leg. From these associations it is clear that CMEs and flares do not cause one another, but they often occur in close association.

Here, we have described a 'typical' CME. However, some CME events do not fit nicely into the scenario described. Some events are more 'tongue-like', appearing as an ascending mound. Some are more spikey and narrow. A newly discovered class of events are relatively common, small ejecta which appear to ascend within the streamer systems and their relationship to their larger brothers is not at all clear at this stage. Thus, while describing CME activity it should be noted that there may actually be different event types relating to different ejection processes, or there may be a wide event 'spectrum' with huge eruptions at one extreme, and small ascending clouds at the other extreme. Of course, all of this is subject to the sensitivity of our coronagraphs and as we develop better techniques we are likely to identify events or event types which were not previously visible.

Finally, one feature which we have not addressed here is the fact that the CME is a projection of a three-dimensional object projected onto a flat image, in the plane of the sky. This provides a few very basic questions. Is the CME loop representing the projection of a bubble, a loop or an arcade? If the CME loop was identified as the blast or coronal response to a sudden impulsive event from below, such as a flare, then a bubble-like expansion may be expected. However, no such cause can be identified. This, combined with the fact that the CME appears to be a signal of the eruption of a magnetic loop system, suggests quite strongly that we are witnessing the eruption of a magnetic loop or (more likely) arcade. The latter would fit nicely with the multiloop structures often seen in CMEs and is consistent with the idea that prominences lie within arcade systems in the solar atmosphere. It should also be noted that observations of prominences on the limb, particularly in eclipse images, reveal that the prominence arcade systems appear to enclose density depletions or cavities, which is also consistent with the basic structure of many CMEs.

Coronal mass ejections and the solar cycle

Solar activity adheres to an 11 yr magnetic cycle and CMEs are no exception to this. The rate of CMEs varies in harmony with the numbers of sunspots and rates of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

flares. However, whereas the sunspot number increases by a factor of over 100 between the minimum and maximum of solar activity the number of CMEs appears to vary by a factor of only 10 or so, from about 0.1 per day to over 1 per day at solar maximum.

As well as variations in the CME rate, the solar cycle sees a variation in the latitude of CME source regions. During solar minimum CMEs tend to come from regions focused at or near the solar equator. At solar maximum, CMEs come from a larger spread of solar latitudes. This appears to be a direct association with other coronal features. At solar minimum the helmet streamers are mainly confined to a single belt around the Sun which lies at low latitudes. At solar maximum the corona is far more complex with systems of streamers at all but the highest latitudes. This appears to confirm that the majority of CMEs are associated with streamers.

The two CMEs shown in figures 1 and 2 show two extremes. Figure 1 shows a CME at the maximum of solar activity. Its apparent ejection over the polar region is most likely a projection effect. However, the ascending prominence was seen to erupt on the disk and it was formerly a high-latitude polar crown filament. On the other hand, figure 2 is a CME at solar minimum which erupted through a pre-existing equatorial helmet streamer.

The CME onset question

CMEs are detected using coronagraphs. Thus, the source regions of CMEs are obscured by the occulting disks of the very instruments which detect them. CMEs are seen clearly to rise up from behind the occulting disks of Skylab, Solar Maximum Mission and the intermediate coronagraph of the Solar and Heliospheric Observatory—i.e. they rise up from below half a solar radius above the limb of the Sun. Coronagraphs which are able to observe nearer to the disk, such as the Mauna Loa coronameter and the inner coronagraph of the Solar and Heliospheric Observatory, may be able to view portions of the CME source regions, although the observations thus far are inconclusive. To observe the onset of CMEs one must observe the lowest levels of the solar atmosphere and for this many researchers employ different techniques.

For reasons given above, coronagraphs are best suited to observations in the plane of the sky, that is, they are best suited to detecting CMEs which erupt over the edges of the Sun. Thus, even if one could observe the source region of the CME it would be near or even beyond the limb, or edge, of the Sun. Even if the source region of the CME is on the near side of the Sun, it will be near the limb and thus, because of the curvature of the Sun's disk, the features in the Sun's atmosphere associated with the CME onset will be foreshortened.

Before considering activity which we might observe on the solar disk which could be associated with a CME onset, we can make some comments on the CME source region. The average CME is 45° in span; some are much larger. CMEs often have well-defined legs which must indicate where they are rooted at the Sun, and those legs do not move during the eruption. These observations imply that the CME comes typically from a rather large source region, much larger than a flare or even an active region.

Flares and prominence eruptions are the most obvious forms of solar activity whose relationship with CMEs can be investigated. They are both observed using ground-based techniques. In basic terms, we can identify a CME using a coronagraph and then examine the underlying portion of the Sun to look for flare or prominence activity. Some discussion about the flare–CME and prominence–CME relationship are given above, but here we consider their possible role in the onset.

We already know that SOLAR PROMINENCES can erupt as part of the CME process, but is that the result of some kind of secondary process or could the prominence signal the driving force behind the entire CME event? Over the years the fact that the observing techniques used to detect CMEs and to observe prominence eruptions are so different has spawned many lines of research which consider each in isolation from the other. This has hindered our understanding of the full picture. However, whereas CMEs can be seen without obvious prominence material within them, coronagraphs never detect prominences ascending through the corona without an associated CME. Thus, the current thinking is that the prominence eruption is part of the CME process, but it does not drive the overlying CME loops.

As described above, the association between flares and CMEs is not straightforward. The association was recognized in the early years of CME observation. Again, flares could be identified from ground-based observations and their locations and timings noted with respect to any overlying CME activity (see SOLAR FLARE OBSERVATIONS). Quite logically, many researchers at that time suggested that the CME was some kind of blast or coronal response to an underlying flare. However, many CMEs occur without underlying flares and the timing and location of flares which are associated with CMEs imply that flares do not drive CMEs.

One has to conclude that the associated flare and prominence activity occur as a result of processes which also led to the mass ejection, but that neither the flare nor the prominence eruption cause the CME. For the same reasons, the CME cannot generate the flare.

So, far from making the picture clear, the examination of flare and prominence associations with CME onsets reveals a rather confusing picture.

The Sun's corona radiates ultraviolet and x-ray light. This is due to the fact that the plasma in the corona is so hot—rising from 6000 K at the photosphere to a few million K. However, as well as thermal emission, ultraviolet observations in particular reveal patterns—so-called emission lines—which come from trace elements in the Sun's atmosphere, such as iron, silicon, oxygen, magnesium and so on. We are aware of this because the wavelength or color of the emission characteristics is known from laboratory measurements. By making observations in specific emission lines one can view

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

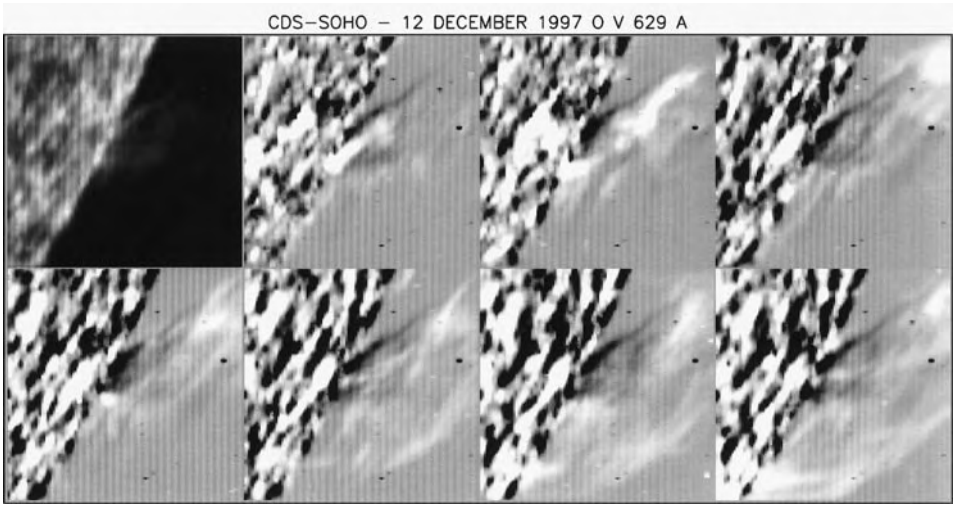


Figure 3. An ejection event detected in ultraviolet light from oxygen showing ascending plasma at a temperature of 250 000 K (courtesy of Rutherford Appleton Laboratory).

different temperature layers in the Sun's atmosphere and derive plasma characteristics of the solar plasmas, such as density, temperature, flow velocities and plasma constituents.

Employing these spectroscopic techniques, recent spacecraft campaigns, involving the Solar and Heliospheric Observatory in particular, have been used to examine in detail the solar atmosphere at the time of a CME onset. By making spectroscopic observations near the solar limb and awaiting an ejection a new view of the ejection process is being formed.

Figure 3 is an example of the fruits of one of these observing sequences, where, in this case, an ultraviolet spectrometer has been used to scan a region of the solar limb where an ejection is underway. Each image covers the same 200 000 km × 200 000 km area on the edge of the Sun at 16 min intervals from top left to bottom right. Although the event was detected in emission lines which allow the analysis of plasmas at 20 000 K–2 000 000 K, we show here radiation from a form of ionized oxygen which has a characteristic temperature of 250 000 K. The first image is a raw image which shows the solar limb and a weak, diffuse cloud extending into the corona. The subsequent images are differenced, i.e. the intensity of the first image is subtracted from each. Thus, for each image, white indicates an increase in intensity with time, and black shows a depletion with time. The heavy black and white mottled patterns of the solar surface reflect the activity of the supergranulation. However, the dramatic feature is the clear outward expansion of a 250 000 K cloud as part of a CME onset. The ultraviolet spectroscopy reveals clear ejection of material associated with a prominence eruption, as well as the outward expansion of plasmas up to several hundred thousand kelvins. Hotter plasmas (around 1 million kelvins and upwards) show little evidence for the ejection at all, and certainly confirm that there was no hot, impulsive event at the onset. These spectral analyses are continuing and, at this time, we conclude that the ejection process is a very subtle event, i.e. the events which are so dramatic in the high corona are really rather insignificant lower in the atmosphere.

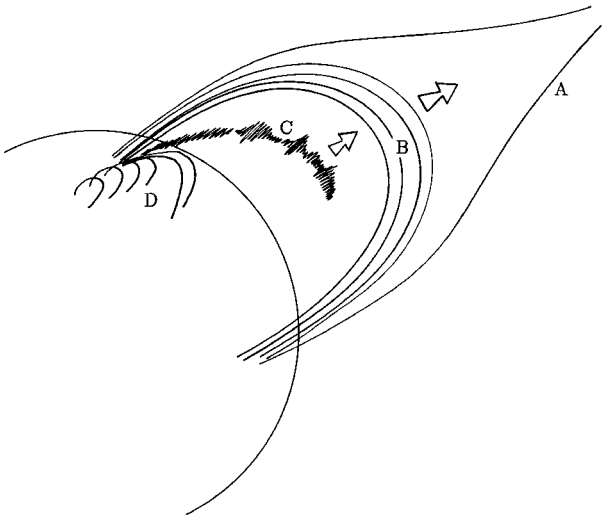


Figure 4. A schematic model of the CME process indicating (A) a pre-existing coronal streamer, (B) the ascending CME loops, (C) an erupting prominence and (D) an underlying active region.

Conclusion

Figure 4 shows a schematic of the CME process as many would describe it today. The salient components of the cartoon are (A) a pre-existing coronal streamer, (B) the ascending CME loop structure, (C) a prominence erupting behind the CME loops and (D) an active region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In this case, we include the pre-existing streamer (A) although for some events such a feature may not be apparent. Under the streamer are closed-loop systems spanning some tens of heliographic degrees; this is a typical streamer form and size. These loops may be rooted in quiet Sun regions but in this case we show an active region (D). Under one portion of the loop systems within the span of the streamer is a magnetic arcade containing a filament or prominence.

The magnetic fields are rooted in the body of the Sun. As a result of the differential motion of the solar 'surface' the loop footpoints may shear, i.e. slide in different directions with the flow of the Sun. Such motion can cause the loops to begin to rise as the magnetic system tries to achieve an equilibrium. This shear activity, combined with magnetic complexities associated with the associated active region, can cause the large-scale fields to achieve a non-equilibrium situation or a lack of stability. The loops may expand outward. As they begin to expand the streamer above shows signs of some swelling. The continued ascent of the loops heralds the ascent of the outer loop of the CME as we know it.

As the magnetic loops expand out, either the arcade enclosing the filament or prominence ascends with it or quite separately reaches a non-equilibrium because of the removal of the overlying magnetic fields which would have acted to anchor the system. Thus, the eruption of a prominence (C) follows the CME. The exact nature of the ascending prominence will depend on many factors such as the state of development of the prominence prior to the event.

In this picture part of the ascending loop system is rooted in an active region (D). One might expect that the effect of this opening of field lines could in itself drive secondary changes in the active region. This may result in the onset of a solar flare if the conditions are right. Such an event may occur at any time in the sequence and, thus, may occur before the CME, at the same time as the CME onset, or after the CME onset. Thus, the CME onset and the flare are associated but do not cause one another. They are both driven by the same basic process.

Bibliography

Crooker N, Joselyn J-A and Feynman J 1997 *Coronal Mass Ejections (Geophysical Monograph Series, vol 99)* (American Geophysical Union)

Hundhausen A J 1988 The origin and propagation of coronal mass ejections *Proc. 6th Int. Solar Wind Conf. (Technical Note TN-306)* ed V J Pizzo, T E Holzer and D G Sime (Boulder, CO: National Center for Atmospheric Research) pp 181–214

Richard A Harrison

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Coronal Mass Ejection: Theory

Solar coronal mass ejections (CMEs) are episodic expulsions of mass out of the solar CORONA involving large-scale reconfigurations of the corona. In a typical event, several 10^{15-16} g of plasma pre-existing in the low corona is expelled into interplanetary space, with a rate of occurrence varying from one event every few days at activity minimum to as high as three events per day at activity maximum in the course of a SOLAR CYCLE. The CME and the FLARE are the two most energetic phenomena, liberating some 10^{31-32} erg of energy in each case, among the myriad of time-dependent phenomena in the corona. Since it involves length scales of the order of the solar radius and time scales characteristic of sound and Alfvén speeds in the corona, the CME is appropriately described by MAGNETOHYDRODYNAMICS (MHD) and is thus one of few known MHD processes in astrophysics which are accessible to direct imaging as a spatially resolved time-dependent event (see SOLAR CORONAL MASS EJECTION: OBSERVATIONS).

Discovered in the 1970s, CMEs have been a subject of intensive research, with several outstanding questions to which contending theoretical answers have been proposed. The rest of this article treats some of these questions: those dealing with the causes and consequences of CMEs, the origin of the CME energy, the nonlinear, time-dependent MHD flows of CMEs in the corona, and the characteristic structural form of the CMEs and their manifestations in the SOLAR WIND.

Causes and consequences of CMEs

The observed close association between CMEs and the other two major forms of coronal eruptions, namely flares and eruptive prominences, was the basis for the theory, once seriously considered, that CMEs are the responses of the corona to these other eruptions occurring lower down in the corona. This suggestion was attractive because of its physical simplicity. It was also in keeping with the reasonable idea prevalent before CMEs were discovered that the impulsive flare might produce a blast wave traveling out of the corona. This blast wave was identified with type II traveling radio noise and with shock wave structures observed in the solar wind by satellites at 1 AU. In this theory the CME would be identified with the blast wave generated by the flare. This scenario was pursued with many simulations with time-dependent MHD numerical codes treating CMEs as the results of impulsive inputs of flare-like energies at the base of static atmospheres embedding a potential magnetic field.

The flare initiation of CMEs as a theoretical idea was shown to be not tenable by the observational results of R Harrison and by the follow-up work of A Hundhausen. Although the question is still being debated by some theorists, it is widely accepted that where observation is of sufficiently good quality, when a CME is associated with an observed flare, the flare does not as a rule precede the CME onset. Actually, events were found which showed that the CME preceded the associated flare. This rules out the flare as a cause of CMEs. From the theoretical point of view, there are several other aspects of the CME phenomenon which suggest that flares are not likely to be the cause of CMEs, notable of which are the following two points. Firstly, an impulsive ejection of the corona would involve the highest speeds of the MHD medium whereas the observed CME speeds fall in a huge range, from 10 to 2000 km s^{-1}, with a median of about 400 km s^{-1}. Secondly, CMEs tend to show a characteristic three-part structure in white-light coronagraph observation—a bright loop leading a dark cavity containing a bright core. This structure is reasonably interpreted to be corresponding to that of the pre-existing helmet streamer—a bright dome overlying a dark cavity containing a quiescent prominence (see CORONAL STREAMERS). The breaking up of a coronal helmet has been commonly observed to be the origin of a CME. Thus the CME is a nonlinear but not necessarily an impulsive 'breaking loose' of a large-scale coronal structure which preserves the coherence of the three-part structure.

The rejection of a flare as the cause of the CME is an important theoretical development, for it brings about a new or different way of viewing the large-scale corona. This view has a strong theoretical appeal because it integrates various theoretical and observational results in terms of simple MHD principles.

The solar corona is ordered by an interplay between two competing effects as the result of heating and the presence of magnetic fields. The hitherto still poorly understood heating maintains the corona at temperatures of the order of 1 to 2 million degrees with such a degree of relentlessness that at no time is the corona in any large bulk part able to cool down significantly below these temperatures. One of the consequences of its high temperature is the high thermal conductivity of the corona. Combined with the inverse square fall-off in solar gravity, the hot corona can extend so far out from the solar surface that the natural state of the corona is one of outward expansion into the solar wind, as first pointed out by E N Parker. This expansion is not ballistic or forced from below. The expansion results from a failure of gravity to confine the upper atmosphere. The upper atmosphere continuously expands to supersonic speeds, during times of more or less steady conditions, while the lower atmosphere heaves continuously outward, departing only weakly from hydrostatic equilibrium, to make up for the expansion.

The presence of magnetic fields in the electrically highly conducting corona introduces the Lorentz force, whose action may be decomposed into that of a magnetic pressure force combined with that of a magnetic tension force. The magnetic pressure can only enhance the tendency of the corona to expand. The magnetic tension force provides a means of restraining the expansion provided it acts towards the Sun and provided the magnetic field is sufficiently intense. This occurs low in the corona where magnetic fields are bipolar with the magnetic feet anchored to the solar surface on the two sides

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of a polarity inversion line. The low corona is dominated by the typically 10 G magnetic field, with a plasma β of the order of less than 0.1, that is, with the plasma pressure typically of the order of a tenth of the magnetic pressure. In such a region, the closed magnetic fields may trap a parcel of plasma in more or less static equilibrium against the global tendency to expand. This is the origin of the ubiquitous coronal helmet streamer straddling a photospheric magnetic inversion line. Outside of the helmet, the field is strong but, because it extends too far out where the solar wind dominates, it is kept open into interplanetary space by the solar wind.

Another way of seeing the effect of the magnetic tension force is to note that, under the condition of high electrical conductivity, a bipolar magnetic field anchored to the surface of the Sun has a tendency to remain closed because this is a state of minimum magnetic energy. The competition alluded to above is then between this magnetic tendency and the opposing expansion of the global corona. The balance between these competing effects determines how much mass and magnetic flux is trapped in a coronal helmet surrounded by open magnetic fields where the wind flows. The CMEs are then the consequence of the failure of this balance when the helmet would break loose to move out along with the global solar wind.

In this view, the term 'loss of equilibrium or non-equilibrium' is invoked to make the point that it is not just an instability. The remarkable aspect of a CME is that, when initiated, it expands out in a global nonlinear flow with a total mass of ejection characteristically of the order of a few times 10^{14-16} g. The idea of an instability merely implies that the system would evolve away from an initial equilibrium when perturbed, but such a behavior does not *a priori* require the ejection of a huge amount of material out of the corona, a basic feature of the CME phenomenon. It is the teetering between confinement and expansion of plasmas trapped in the coronal helmets to result in a global loss of equilibrium that underlies the CME phenomenon, in this theoretical interpretation. Depending on physical circumstances under which this loss of equilibrium develops, it is theoretically conceivable that the structure breaking up may maintain its internal form with coherence and may move out with a broad range of speeds as opposed to the case of an impulsive blast-off.

The coronal magnetic field that produces the CME is also the agent for producing the PROMINENCE and its eruption, and the solar flare. Thus it is not a simple causal relationship one needs to find among these major eruptions making up the activity in the corona. This is an important point for theory, and it is reflected in the facts that flares, as defined as a sudden localized heating of the corona, have several different varieties and that the rate of flare occurrence increases by orders of magnitude over a solar cycle as opposed to an increase of a factor not larger than 10 in the CME rate over the same period of time.

There is, however, a particular relationship between CMEs and flares of a special kind, often referred to as the two-ribbon flares, which is of theoretical significance. The adjective 'two-ribbon' came originally from observations of the Hα part of this type of flare, referring to the characteristic heating occurring at the chromosphere along two bands which separate in time. This type of flare, known long before the discovery of CMEs, has been well explained in terms of magnetic reconnection taking place above the flaring chromosphere which recloses a magnetic field that has been caused to open up. The two flare ribbons in the chromosphere are then reasonably interpreted to be marking the footpoints of newly reconnected fields containing heated plasma. As more magnetic fields reconnect, newly heated footpoints light up while footpoints heated earlier cool out of visibility to produce the effect of the separating ribbons. With soft x-ray observations available from space these days, the two-ribbon flares are observed as long duration outputs of soft x-rays (hours to a fraction of a day) rising above the full-disk output, or as arcades of flaring loops in soft x-ray images.

Before CMEs were known, this flare mechanism lacked a physical reason why the magnetic field should open up on its own to produce a two-ribbon flare. The idea that a CME is the opening up of the magnetic field in a coronal helmet provides the missing ingredient. If the two-ribbon flare is indeed the result of the reclosing of opened magnetic fields by reconnection, and the CME is indeed responsible for the opening up of an initially closed magnetic field, the CME necessarily has to precede the two-ribbon flare associated with it. Theoretically, the reconnection could begin as soon as the CME takes off with the field opening up, or the reconnection could be delayed until after the field has completely opened up and the CME is well on its way out of the upper corona. The latter is supported by some observations and the former leaves the temporal sequence ambiguous because the flare would have an onset very close to, or, in practical terms, almost simultaneously occurring with the CMEs. What this theory would not accommodate would be the observation of a two-ribbon flare preceding the associated CME. If it can be shown that such an event occurs and that the association between the flare and CME is not by chance, then the event cannot be explained by the theory, pointing to a much more complex physical behavior than the one envisioned in the theory.

CME energy

The mechanical energy of a CME is impressively large. The often quoted CME energy of 10^{31-32} erg is the sum of its kinetic energy and the work done against solar gravity to lift its mass of 10^{15-16} g out of the corona. The gravitational potential energy typically is about a half of the sum, as can be seen from the fact that the gravitational escape speed in the low corona is about 500 km s^{-1} whereas the observed median speed of CMEs is about 400 km s^{-1} with the large range mentioned above. The sound speed of the 2 million degree corona is about 120 km s^{-1}. The Alfvén speed in the low corona can be estimated to be in excess of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

700 km s^{-1}. Therefore, a striking feature of the CME is that a great many CME speeds are supersonic, sub-Alfvénic, and below the low-coronal gravitational escape speed.

It is widely held that the CME energy originates from energy contained within the corona previously stored in the magnetic field. This view is based on noting that the CME moves at speeds near the limit or in excess of the characteristic MHD speeds of wave propagation in the corona. In this case, it is not possible for the CME to be driven by photospheric processes which involve typical speeds of 0.5 km s^{-1}. It is possible that the photosphere may stress the coronal magnetic field over a long period of time to result in a build up of energy in the corona, but it is not likely in this view that the photospheric motion, given its small magnitude, is relevant during the course of the CME moving out of the corona at the characteristic high speed of the latter. This point of view is supported by the fact that to date no surface photospheric motions of a particular type have been shown to be indicative of a CME leaving the solar corona directly above the photospheric region of observation.

A minority view has persisted, hinting at some possible large-scale non-MHD effects in which the CME is driven by currents generated at the photosphere or in the solar convection region. This view has remained speculative and non-quantitative.

Accounting for the CME energy in terms of magnetic energy stored in the corona has turned out to be a deeper problem than previously appreciated. Taking the simplifying limit of the corona as a perfect plasma conductor with negligible inertia, it was an attractive idea to try to account for the energy in the field-aligned currents of a FORCE-FREE MAGNETIC FIELD anchored to the surface of the Sun. These currents were taken to be induced by the stressing of the field by slow but persistent photospheric motion taking the embedded magnetic footpoints along with it. It was proposed in the 1970s that the spontaneous opening up of a closed force-free magnetic field may result from a build-up of magnetic energy by surface stressing to a level above that of an open state. This idea was shown to be incorrect by J J Aly in the 1980s, who proposed a now widely accepted conjecture that the build-up of energy in a closed and anchored force-free magnetic field cannot exceed the energy of any open magnetic field having the same anchored footpoints on the solar surface.

The energy demand of a CME is especially impressive if one has to account not only for its energy but also for the energy of the open field it leaves behind, in the light of Aly's conjecture. The large amount of energy in the open field left behind by the CME would neatly account for the post-CME flare that results from the reclosing of the opened magnetic field. Therefore, the theoretical question on the energy storage for the CME is one of demonstrating that there could be coronal structures with closed magnetic fields in possession of two huge amounts of free energy, that carried away in the gravitational and kinetic energy of the CME and that left behind in the open magnetic field to fuel the CME-associated flare. This, by Aly's conjecture, cannot be accounted for by the stressing of force-free magnetic fields.

A first step in the resolution of this problem is to realize that the force-free assumption is not valid over the large scale of the corona. In the thermally highly conducting corona, pressure is almost isothermal and it falls with height less rapidly than the magnetic pressure, especially with gravity dropping inversely with distance squared. Both the pressure and the solar wind dominate over the magnetic field at increasing distance from the Sun and the force-free assumption is invalid over such scales, for example, as characterized by the radial extent of the helmet streamer. Over these scales, the Lorentz force is not zero everywhere but interacts with pressure gradient and gravitational forces, and all these forces also produce the solar wind in the open-field regions.

Once this is recognized, models of the corona in static equilibrium can be constructed to show that the necessary amount of magnetic energy can be stored to fuel ejection of a CME and its associated flare. Of particular note is that the magnetic energy is most efficiently stored if a part of the magnetic field is not in the form of simple bipolar anchored field but is in the form of a flux rope detached in its main part from the coronal base, that can be identified with the cavity under the helmet-streamer. The confinement of this flux rope in equilibrium requires the weight of the large mass in the helmet dome. In this model, the large mass of the CME plays the very role of trapping the magnetic energy in the pre-eruption stage that eventually drives the CME.

This promising resolution of the energy problem has also raised a new question. If the magnetic energy needed to drive the CME requires interaction between the plasma and the field for its storage, it must follow that plasma pressure forces must also play a significant role in the CME outflow. This implication is not surprising when it is realized that the Lorentz force cannot work against gravity in the direction of the magnetic field, only the component of the pressure force in that direction can. And, working against gravity is an essential aspect in the energetics of the CME. Ultimately, to understand the effect of pressure would require an understanding of the nature of coronal heating under the time-dependent circumstance of the moving CME.

The time-dependent MHD flow of a CME

The simplest description of the CME in fully developed motion uses the one-fluid MHD equations including the presence of solar gravity. To avoid the complication of the energy and heating of the plasma, a polytropic assumption with an index taken smaller than 5/3 is usually made to simulate the heating of the coronal, largely proton, gas in expansive motion.

The only general approach to treat the MHD equations is by the use of numerical codes with the capability of dealing with at least two dimensions in space. Previous work had treated the flare-initiated CME theory and the shearing of magnetic footpoints to force

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a spontaneous opening up of the coronal magnetic field. These works had been instructive in their respective ways but have not dealt with the CME nonlinear flows such as observed in the low corona. Recent numerical modeling efforts have begun to address the idea that the CME cavity is a magnetic flux rope and morphological agreement between two-dimensional models and the observed three-part structures of CMEs have been obtained, lending credence to the suggestion of magnetic flux ropes playing a role in the CME phenomenon.

The MHD equations for a polytrope of index 4/3 admit a family of self-similar solutions both in two- and in three-dimensional space. These solutions have been useful in capturing the essence of the MHD time-dependent expulsion of mass out of a gravitational potential well. Among the properties directly demonstrated with these solutions is that the transport of magnetic field, mass and the 4/3 polytropic pressure may proceed in a mode such that gravity can be countered throughout the expulsion process. The result is that the modeled CME could in this mode of behavior travel out of the corona with mildly accelerated or constant speeds in an enormous range of speeds, such as seen in the observed CMEs. The possibility of generating solutions in this class with fully three-dimensional variations was also exploited to produce a CME geometrically sophisticated enough for direct (favorable) comparison with white-light observations.

The large range of observed CME speeds implies that a rich variety of MHD shocks may be associated with this highly nonlinear flow. The fast MHD shock has behaviors which are intuitively simple modifications of the hydrodynamic shock by the presence of magnetic fields. The slow and intermediate MHD shocks have properties which can be counter-intuitive, such as the possibility of having a shock surface shaped with a concavity away from the shock driver. The leading fronts of observed CMEs have shown a variety of shapes suggestive of such unconventional shock-surface geometry. Recent time-dependent MHD models have demonstrated that these observations may be explained in terms of the MHD shocks formed by the CMEs plowing into the ambient corona at speeds in the supersonic but sub-Alfvénic range. Much work remains to be done in this classical area of MHD shock theory, especially in the discovery of the role of the intermediate shock in the CME outflow.

In interplanetary space, the signatures for recognizing a CME in the solar wind are being developed. Among these signatures are the bidirectional streaming electrons indicative of magnetic lines of forces with both ends anchored to the low corona such as might be expected from the stretching out of bipolar lines of force by the CMEs. The three-part structure of the CMEs seen in the corona has not been unambiguously identified with their counterparts in interplanetary space but some promising results have been obtained, including the identification of the CME cavity with the so-called interplanetary magnetic clouds. Both the cavity and the magnetic cloud may be the manifestation of a magnetic flux rope. It is conceivable that the CME will undergo significant evolution as it travels along with the solar wind. Much theoretical work and modeling remain to be done in parallel with observational studies of coronal and interplanetary observations to address these issues.

Conclusion

The discovery of CMEs in the 1970s has ushered in a new chapter for coronal physics, taking us from the concern with the solar wind structures in the synoptic regime to the time dependence of episodic mass expulsions, and taking us from the preoccupation with flares to a new form of eruption involving just as much energy. Worthy of note is the fact that the flare is dissipative in nature, heating of the coronal plasma; whereas, the CME involves liberation of energy in the ordered forms, gravitational potential and kinetic motion. The two forms of eruption are related in ways that are not simple and the physics of the corona is richer than many would have realized when CMEs were first discovered. Much remains to be done, and there is a confusion of ideas. These have been partially resolved and the view will probably continue to evolve.

Bibliography

Gibson S E and Low B C 1998 *Astrophys. J.* **493** 460
Low B C 1996 *Solar Phys.* **167** 217

Boon Chye Low

CC: <cmw@hao.ucar.edu>

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Cycle

The term *solar cycle* refers to a quasi-periodic variation with a period of about 11 years, visible in many of the Sun's observables. The solar cycle is most easily observed in indices directly related to the Sun's magnetic field, such as SUNSPOTS. During the last decades, solar-cycle variations have also been found in many other aspects of the Sun (irradiance, surface flows, coronal shape, oscillation frequencies, etc). The modulation amplitudes may vary widely between different indices. It is minute in visible light, and this explains why the solar cycle is not visible to a casual naked-eye observer. Elsewhere (in the far ultraviolet, x-rays, radio frequencies) the modulation amplitude is large (table 1). The prime cause of the solar cycle is a quasi-periodic oscillation of the SOLAR MAGNETIC FIELD.

Magnetic-field indices

The oldest and best known index of the solar cycle is provided by sunspots, whose discovery for Europe around 1610 is associated with the names of FABRICIUS, GALILEO and SCHEINER. The commonest sunspot index is the Zürich sunspot number $R = K(10g + f)$, where g is the number of sunspot groups, f is the number of individual spots and K is a coefficient that corrects for differences in the quality of the observations. Usually one derives an annual average $\overline{R}$, shown in figure 1. The most striking feature of the sunspot record is the Schwabe cycle of about 11 years, discovered by Samuel Heinrich SCHWABE in 1843. By convention, the cycle that began in 1755 is assigned number 1; hence cycle 23 began in 1997.

Cycle shape

A typical sunspot cycle in figure 1 is characterized by a sharp rise from minimum to maximum, lasting 3–6 years (on average 4.8). The duration of the rise phase is anticorrelated with the height of the maximum (Waldmeier's rule). The maximum is followed by a gradual decline lasting 5–8 years (on average 6.2). Whereas the maxima of the Zürich number are typically well-defined, this is not the case for some other SOLAR ACTIVITY INDICES. For example, it has been suggested that the number of *large* sunspots has a double-peaked maximum.

Most indices vary roughly in phase with the sunspot number, and therefore the sunspot extrema are also referred to as the solar minimum and the solar maximum. On the other hand, some coronal indices are clearly out of phase with the sunspot cycle, and they peak at the sunspot minimum (see below).

Table 1. Selected indices of the solar cycle.

	Period	Minimum†	Maximum	Unit
Magnetic indices				
Zürich annual mean sunspot number $\overline{R}$	1610–1997	0–11	38–201‡	
Weak field (B < 25 G) flux	1975–92	15–20	30–33	10^{22} Mx
Strong field (B > 25 G) flux	1975–92	2–3	60–80	10^{22} Mx
Full-disk spectral indices				
10.7 cm radio flux	1947–95	70	160–280	10^{-22} W m^{-2} Hz^{-1}
393.4 nm Ca II K 1 Å index (UV)	1975–90	0.085	0.105	
121.6 nm Lα UV flux (SME§)	1982–9	2.5	4	10^{11} photons cm^{-2} s^{-1}
0.1–0.8 nm x-ray flux (GOES‖)	1974–88	<0.01	3	10^{-6} W m^{-2}
Other indices				
Total irradiance (ACRIM¶)	1979–97	1365.6+	1366.9+	W m^{-2}
530.3 nm green coronal index	1939–94	2	10–20	10^{16} W sr^{-1}
Solar wind speed at 1 AU (NSSDC*)[1]	1965–87	400	500–530	km s^{-1}

†Minimum occurs at solar minimum, except for solar wind speed, whose minimum occurs at solar maximum.

‡Excluding the Maunder minimum (1645–1715) and the scarce data before 1645.

§Solar Mesosphere Explorer.

‖Geostationary Operational Environmental Satellites.

Active Cavity Radiometer Irradiance Monitor on board the Solar Maximum Mission.

+A calibration error of about 3 W m^{-2} must be allowed for.

*National Space Science Data Center.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

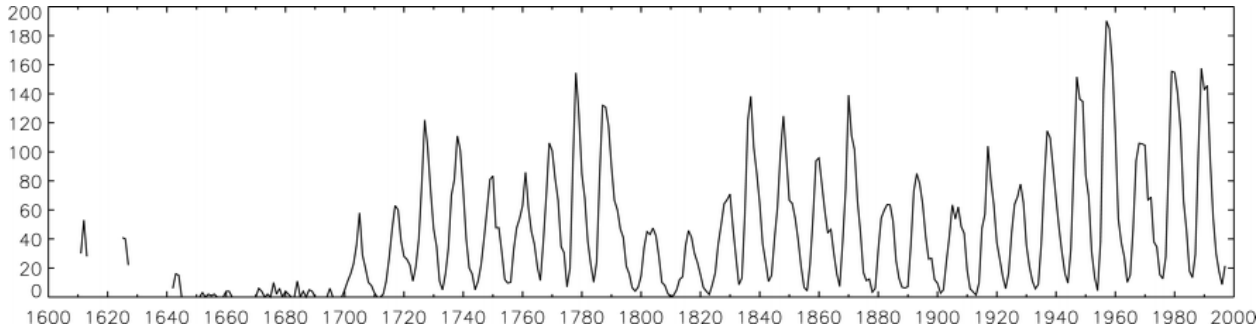


Figure 1. The Zürich annual-mean sunspot number (data from the Geophysical Data Center, Boulder, CO, USA).

Cycle length and amplitude

Individual sunspot cycles since 1710 lasted between 7 and 14 years (on average 11) and had amplitudes in the range $38 \leq \overline{R} \leq 201$ (on average 105). No significant correlation between length and amplitude of individual cycles has been found. For nearly all of the cycles 0–21, the amplitude of an odd-numbered cycle exceeds that of the preceding even-numbered cycle (the Gnevyshev–Ohl rule).

Magnetic bipoles

Sunspots are sites of strong magnetic fields ($B \approx 2$–3 kG). Frequently, they form bipolar pairs, consisting of a leading spot and a trailing spot with respect to the solar rotation. Virtually all bipolar pairs obey the Hale–Nicholson polarity rules. (a) The magnetic polarities of leading and trailing spots are opposite, and those of the leading spots in one hemisphere are opposite to those of the leading spots in the other hemisphere. (b) The predominant sunspot polarities reverse after the solar minimum. Taking into account this polarity reversal, the magnetic field has a period of about 22 years, which is referred to as the Hale cycle.

The axes of bipolar pairs are slightly tilted with respect to the equator, leading spots being closest to the equator (Joy's rule). Tilt angles (on average 4º) increase with latitude, but show no significant variation with the solar cycle. Also the relative size distribution of SOLAR ACTIVE REGIONS is almost invariant throughout the cycle. The only exception is a period of 2–3 years before solar minimum, during which small bipolar regions are relatively more frequent. This is due mainly to the emergence of opposite-polarity ephemeral regions at high latitudes. They follow the Hale cycle with a minimum preceding the solar minimum by about 1 year.

Butterfly diagram

The latitudinal distribution of sunspots as a function of time can be viewed from the butterfly diagram (figure 3). Sunspots, indicated by black dots, appear in two belts parallel to the equator, whose midpoints migrate equatorward from about ±27º to about ±8º during the course of a cycle (Spörer's law), and whose widths attain a maximum of about 36º during the sunspot maximum. The polarities of the northern and southern wings of the butterfly diagram are opposite, and they alternate from one Schwabe cycle to the next. Sunspots exist for at most a few months, and they hardly migrate themselves, but in the course of a cycle every new sunspot appears on average at a lower latitude.

Sunspots and active regions have a tendency to emerge near existing active regions. These loci of flux emergence, referred to as active nests or active longitudes, may live up to 6 months (see SOLAR ACTIVITY COMPLEXES). The magnetic field of sunspots probably has its origin near the base of the SOLAR INTERIOR CONVECTION ZONE, at a depth of about 200 000 km, and the butterfly diagram traces this deep-rooted, large-scale magnetic field.

Magnetic network

Magnetograms (figure 2) reveal an extended pattern of surface fields, referred to as the magnetic network (see SOLAR PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS). Around the solar maximum, the magnetic network up to a latitude of about ±60º consists mainly of large unipolar regions, in which the bipolar active regions are embedded. For about 5 years around the solar minimum, the network is much weaker and has mainly mixed polarities. From the magnetograms it is also concluded that the solar-cycle modulation decreases with decreasing field strength (table 1). Thus the total unsigned flux of strong-field regions ($B > 25$ G) varies by a factor 30–40, compared with a factor of only 1.5–2 for weak-field regions ($B < 25$ G).

Polar field

The weak polar magnetic field has mainly one polarity at each pole, and the two poles have opposite polarities. Between the solar minimum and the solar maximum, its polarity on each hemisphere agrees with that of the leading spots of bipolar pairs. Around the solar maximum, the polar field changes its sign.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

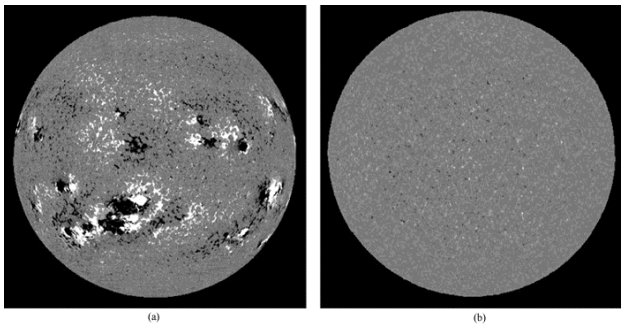


Figure 2. Magnetograms of the solar disk taken respectively at 23 August 1990, near the solar maximum, and at 15 October 1996, near the solar minimum. Black and white indicate magnetic regions of, respectively, negative and positive polarity (National Solar Observatory, USA.)

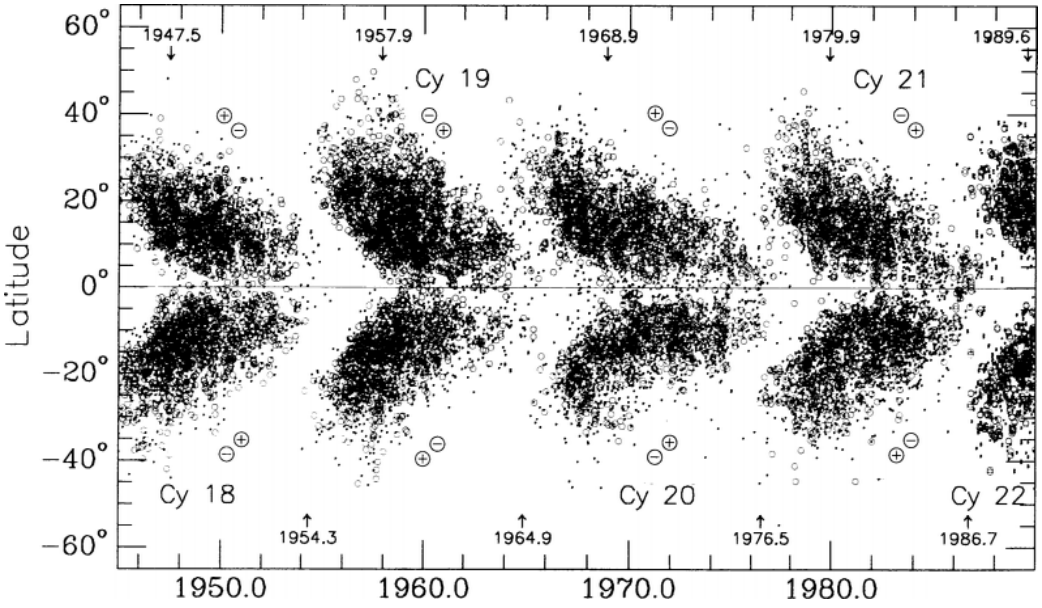


Figure 3. Butterfly diagram. Arrows denote sunspot maxima (top) and minima (bottom); ● and ○ indicate sunspot regions smaller and larger than 80×10^{-6} of a hemisphere, respectively; ⊕ and ⊖ indicate whether the magnetic field of leading and trailing spots is directed outward or inward, respectively. (Adapted from K L Harvey (1992).)

Other periods besides the Hale cycle

If the Sun's surface field as a function of time is decomposed in terms of spherical harmonics, one finds very different frequency spectra for axisymmetric modes (i.e. $m = 0$) with odd ℓ (i.e. antisymmetric with respect to the equator) and axisymmetric modes with even ℓ. Whereas odd-ℓ modes are dominated by the Hale cycle, even-ℓ modes are weaker and have no single, well-defined frequency. No such difference between even- and odd-ℓ modes is observed for non-axisymmetric modes ($m \neq 0$), which have a main peak at the Hale cycle, and smaller peaks at its higher harmonics, almost irrespective of ℓ. Altogether, though, the reality of magnetic cycles shorter than the Hale cycle is still uncertain.

Magnetic cycles longer than the Hale cycle have also been proposed. The sunspot record appears to be modulated by a period of 80–90 years, referred to as the Gleissberg cycle, but the number of oscillations is very small (about three can be seen in figure 1). There is some evidence for this and longer periodicities in the records of ^{14}C and ^{10}Be, but their existence remains uncertain (see SOLAR ACTIVITY: LONG-TERM RECORDS).

Grand minima

During the so-called Maunder minimum (1645–1715) sunspots were very few and, prior to 1700, no cycles are apparent. Coverage of the sunspot record in this period is estimated to be about two-thirds, and the reality of the Maunder minimum is undisputed. This and earlier grand minima such as the Spörer minimum (1420–1530) are clearly visible in the ^{14}C and ^{10}Be data. Timing and duration of the known grand minima are irregular, but the Sun may have spent as much as a third of its time in grand minima. Surprisingly, these records also suggest that the solar cycle continued throughout the Maunder minimum at a reduced level, an issue which is of particular interest for solar dynamo theory (see DYNAMOS: SOLAR AND STELLAR). More recent but less impressive prolonged minima, visible in figure 1, are the Dalton minimum (1800–30) and the Modern minimum (1880–1910).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Grand minima are but the most conspicuous manifestations of solar variability, which is also apparent for example in length and amplitude variations of the sunspot cycle. The question of whether solar variability should be ascribed to (quasi-) periodic or chaotic nonlinear behavior on the one hand, or to stochastic processes on the other hand remains as yet unresolved due to the lack of a sufficiently long dataset (see DYNAMO THEORY).

North–south asymmetry

Asymmetry between sunspot activity in the northern and southern hemispheres is larger than average during solar minima and was particularly strong during the Maunder minimum, when the few observed sunspots seem to have been concentrated on the southern hemisphere. Smaller but nevertheless clear north–south asymmetries can be seen in the regular sunspot cycle as well, for example during the years 1955–65 (figure 3).

Irradiance and spectral properties

Solar-cycle modulations are apparent throughout most of the Sun's electromagnetic spectrum, which extends from radio to gamma ray energies (see SOLAR ACTIVITY INDICES). Only in the far infrared, a technically difficult spectral range, has no modulation been demonstrated yet. Data for a selection of full-disk spectral indices are included in table 1. They exhibit a wide range of minimum-to-maximum variations, from 0.1% for total SOLAR IRRADIANCE up to factors >100 for the x-ray flux at 0.1–0.8 nm. As a general rule, the larger the solar-cycle modulation of an index is, the more it is dominated by phenomena related to strong magnetic fields.

Solar irradiance, like most spectral indices, is largest at the solar maximum, in spite of the darkening by sunspots. This net luminosity increase is attributed largely to an overcompensation by photospheric faculae, the number of which peaks at solar maximum (see SOLAR PHOTOSPHERE: FACULAE).

Solar oscillations and internal structure

The frequencies of solar p-mode oscillations increase by up to 0.1% from solar minimum to solar maximum (see HELIOSEISMOLOGY). The shift increases with frequency, and is largest for oscillations with a period of about 4 min. Such a frequency dependence is interpreted as evidence that the shifts are caused by magnetic fields near the solar surface, rather than by a deep-rooted field. At the solar minimum, p-mode amplitudes are about 35% larger than at the solar maximum.

The rotational splittings of p-mode frequencies also depend on solar-cycle phase. They are possibly related to solar-cycle variations in the Sun's internal rotation, but attempts to derive angular-velocity changes from the measured frequency splittings have so far yielded inconclusive results.

A solar-cycle modulation of the neutrino flux originating in the Sun's core is unproven, and seems unlikely on theoretical grounds (see NEUTRINOS). No significant solar-cycle modulation of the solar diameter and oblateness has yet been found (see SUN: BASIC PROPERTIES).

Velocity fields

Solar-cycle changes in the surface rotation rate of the Sun are seen in Dopplergrams, and are traced by sunspots and other surface features. The Doppler data reveal a pattern of parallel belts of faster and slower rotation known as torsional oscillations, reminiscent of the butterfly diagram. Their amplitude is 5–10 m s^{-1} (i.e. 0.25–0.5% of the surface rotation), and the regions of maximal shear coincide more or less with sunspot latitudes. Whereas these belts were once thought of as being uninterrupted during the 16–20 years of their migration from pole to equator, persistent gaps at mid latitudes now suggest that their pattern is in fact more complex. Solar-cycle changes in the rotation rate of sunspots are larger (4–5%) and do not show evidence of latitude migration. The largest value of the sunspot rotation rate occurs at the sunspot minimum, and a smaller peak occurs at the sunspot maximum.

Doppler measurements also reveal a general poleward meridional flow of 10–20 m s^{-1} at all latitudes, with no visible solar-cycle modulation. A poleward drift of similar magnitude is seen in the motion of unipolar magnetic regions at high latitudes. Their velocity is largest around the solar maximum. The latitudinal drift of sunspots is typically 10 times smaller. It is not generally poleward, and can be different for leading and trailing spots. There is some evidence for a solar-cycle dependence.

Finally, there are indications for solar-cycle changes of the GRANULATION. At the solar maximum, size and flow speed of the granules seem somewhat reduced, but this modulation needs further confirmation.

Coronal indices and solar wind

The shape of the CORONA changes during the solar cycle. At the solar minimum it is elongated along the equatorial direction, and POLAR PLUMES can be seen. At the solar maximum, most of the disk is surrounded by CORONAL LOOPS and CORONAL STREAMERS, giving the corona a rounder shape. This increase of coronal activity at solar maximum is also evidenced by the number of coronal mass ejections and by the green coronal index (table 1; see SOLAR CORONAL MASS EJECTION: OBSERVATIONS). Some coronal indices vary out of phase with the sunspot cycle, and peak at the sunspot minimum. These include the number of polar FACULAE, SOLAR CORONA X-RAY BRIGHT POINTS and CORONAL HOLES. Coronal holes exist during most of the solar cycle and occur less frequently only during a few years near solar maximum.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SOLAR PROMINENCES and SOLAR FLARES are phenomena related to SOLAR ACTIVE REGIONS, and their numbers vary in phase with the solar cycle. Solar wind speed and the spiral angle of the interplanetary magnetic field are slightly reduced during the solar maximum (table 1). This near-constancy of the SOLAR WIND may in part arise because it is fed with particles from both flares and coronal holes, the numbers of which vary out of phase with one another. However, the flux of high-energy solar particles is more strongly modulated because they are produced mainly by flares.

Extended solar cycle

The extended solar cycle is a concept based on the observation that several branches that are assigned to consecutive cycles can coexist in the butterfly diagram, and seem to be continuous across the solar minimum. The main example is the emergence of opposite-polarity bipolar ephemeral regions at high latitudes, 2–3 years before the solar minimum, apparently precursors of a new cycle. Also observations of the green coronal line, filaments, torsional oscillations, and the reversal of the polar field near the solar maximum may be mentioned. Dynamo theory suggests that such overlapping bands are a consequence of traveling wave fronts.

Terrestrial indices

The SOLAR–TERRESTRIAL CONNECTION is a wide and complex subject, and the implications of the solar cycle for the Earth's climate are controversial due to uncertainties of atmospheric physics and climate modeling. Out of the many solar–terrestrial correlations that have been proposed, only a selection is given here.

It is well established that the strength of the Earth's magnetic field, as measured by the AA index, varies in phase with the solar cycle. The field increase is attributed to compression of the Earth's magnetopause (see MAGNETOSPHERE OF EARTH: MAGNETOPAUSE) due to coronal mass ejections, the number of which peaks at solar maximum. The solar wind carries a magnetic field into interplanetary space, which provides enhanced shielding of COSMIC RAYS during solar maximum. The resulting solar-cycle modulation of the cosmic-ray flux at Earth is typically 50%. This is subsequently passed on to various other indices, including the rate at which radio isotopes are produced in the upper atmosphere. Measurements of the concentration of ^{10}Be in arctic ice cores and of ^{14}C in tree rings thus provide a unique tool for studying the history of solar activity. Due to different atmospheric storage times, the solar-cycle variations in these records are reduced to about 25% and 2% respectively. The 11-year signal in the ^{10}Be concentration follows the solar cycle with a delay of 0.5–1.5 years.

The number of AURORAS is related to the total particle flux emitted by the Sun, and it is weakly modulated by the solar cycle. Low-latitude auroras occur only around the solar maximum; they are attributed to flare-related high-energy particles, whose number peaks at the solar maximum.

Climate models indicate that the solar-cycle variation of the Sun's irradiation has only a marginal effect on terrestrial temperatures (about 0.05 K). A more likely solar-cycle influence on the Earth's climate, if present, would be chemical effects due to changes in the solar UV flux. Another possible cause is variations in cloud formation. They may arise from the solar-cycle modulation of the cosmic-ray flux, which affects the availability of condensation nuclei.

Examples of climate indices that show some evidence of a solar-cycle modulation are pressure scale height and temperature at low latitudes. The evidence for such modulations is, however, mostly based on a small number of cycles (often 1–2), which is probably insufficient. Also a remarkable anticorrelation between temperature on the northern hemisphere and cycle *length* has been found, but no mechanism is known that may account for it.

Stellar cycles

Magnetic cycles have been detected in other stars besides the Sun, as the result of a monitoring project conducted at Mount Wilson since 1966 (see SOLAR–STELLAR CONNECTION). For some 100 lower main-sequence stars regular measurements are made of the chromospheric Ca II H and K emission cores, which are known to have a magnetic origin from solar observations. Cyclic variations are found predominantly in old, slowly rotating stars of stellar types G-K, of which the Sun is one. Stars with well-defined activity cycles have cycle periods ranging from 7 to 14 years. Some stars have long-term activity trends that may turn out to be cyclic as the observations continue. Stars with a low and flat activity level may be in a grand minimum. They make up about a third of the total number of solar-type stars, which may be indicative of the fraction of time spent by the Sun in grand minima.

Origin of the solar cycle

The solar cycle is a magnetic phenomenon, and its explanation must come from dynamo theory, which concerns the generation of magnetic fields in conducting fluids. Theoretical difficulties have as yet prevented a satisfying, comprehensive explanation of the solar cycle. Nevertheless, several elements of the solar dynamo mechanism may be mentioned, that play an important role in most models.

Magnetic-field storage

As suggested by the Hale–Nicholson polarity rules, the prime cause of the solar cycle is an oscillating strong magnetic field (10^4–10^5 G) of predominantly toroidal orientation. This field can be stably stored not in the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

convection zone itself, but in the overshoot layer at the base of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE). For sufficiently large field strengths a magnetic buoyancy instability sets in, which leads to the eruption of magnetic loops from the overshoot layer, their subsequent rise through the convection zone and, as a by-product, the emergence of bipolar pairs of sunspots at the solar surface (figure 4(*a*)).

Alpha effect

Convection or magnetic instabilities (see MAGNETO-HYDRODYNAMIC INSTABILITIES) can cause a magnetic flux element to rise or to sink. Due to the stratification of the atmosphere, a rising gas parcel, say, will generally expand. Through the Coriolis force, rotation provides this diverging flow pattern with a twist that is clockwise on the northern hemisphere and anticlockwise on the southern hemisphere. A sinking, contracting parcel acquires a twist of the opposite sense. The embedded magnetic field is also twisted, and thus poloidal loops may be formed around an initially toroidal flux element (figure 4(*a*)). This amounts to the creation of a net electric current parallel to the magnetic field, and is referred to as the alpha effect. Through the alpha effect, a net poloidal magnetic field is generated from the toroidal magnetic field.

Differential rotation

In the solar overshoot layer, rotation varies strongly with depth. This differential rotation, in turn, stretches the poloidal field into a strong toroidal field.

Dynamo mechanism

Both the alpha effect and the existence of an overshoot layer point to the crucial role played by convection in the solar dynamo. Together, the alpha effect and differential rotation maintain the solar magnetic field. They enable the formation of new toroidal flux elements of both polarities adjacent to an initial toroidal flux element. If this flux merges in a suitable manner, and if the differential rotation and the alpha effect have the proper sign, the net effect can be magnetic belts that migrate toward the equator. As the northern belt and the southern belt, which has the opposite polarity, meet at the equator, they both annihilate. At high latitudes, magnetic fields of opposite polarity, marking the beginning of a new cycle, may be formed a in this way (figure 4(*b*)). Thus, the solar cycle may be a traveling-wave phenomenon of the deep-rooted magnetic field.

The spatial distribution of magnetic flux at the solar surface may be largely explained by an additional flux transport model based on flux emergence in active regions, and surface transport by turbulent diffusion and meridional flow. Part of the magnetic energy can be released through various conversion mechanisms, and this leads to the broad range of solar-cycle modulations.

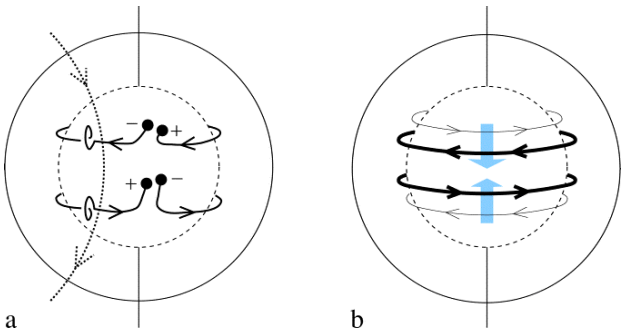


Figure 4. Diagram of the solar-cycle mechanism. The dashed circle represents the top of the radiative interior. (*a*) Generation of the poloidal field (dotted curve) from the toroidal field (drawn curve) by an alpha effect, and the formation of bipolar pairs of sunspots. (*b*) Migration and reversal of the toroidal field.

Breaking news update (30 April 2002)

Solar activity varies on a rough 11-year cycle with some cycles having more than one peak. The current activity cycle appears to be reaching its second peak. The most recent (and ongoing) solar max crested in mid-2000. Sunspot counts were higher than they had been in 10 years, and solar activity was intense. After that, sunspot counts slowly declined and the Sun was relatively quiet for month-long stretches. Solar max was subsiding. But now, activity has once again increased: the Sun is again peppered with spots, and eruptions are frequent.

Bibliography

Harvey K L (ed) 1992 *The Solar Cycle* (Astronomical Society of the Pacific Conference Series 12)

Proctor M R E and Gilbert A D (eds) 1994 *Lectures on Solar and Planetary Dynamos* (Cambridge: Cambridge University Press)

Stephenson F R and Wolfendale A W (eds) 1988 *Secular Solar and Geomagnetic Variations in the Last 10,000 Years* (NATO ASI Series C 236) (Dordrecht: Kluwer)

Willson R C and Hudson H S 1991 The Sun's luminosity over a complete solar cycle *Nature* **351** 42–4

Wilson P R 1994 *Solar and Stellar Activity Cycles (Cambridge Astrophysics Series 24)* (Cambridge: Cambridge University Press)

Mathieu Ossendrijver

Peter Hoyng

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Evolution

It is generally accepted that the formation of the Sun arose from the passage of a shock wave through an interstellar cloud (see, for example, INTERSTELLAR MATTER), causing part of the cloud to become gravitationally bound. The major component of the cloud was hydrogen, together with about 25% helium, and more massive elements in minor proportions.

On contraction of the cloud, half the kinetic energy so produced caused an increase in its temperature and half was lost as radiation. The stages in solar evolution are effectively illustrated by the HERTZSPRUNG–RUSSELL DIAGRAM, a plot of the logarithm of the surface temperature against the logarithm of the luminosity. The pre-main sequence phase, in which the contracting cloud began to emit as a cool, luminous star, lasted some 10^7 years (see PRE-MAIN-SEQUENCE STARS). With further contraction and a consequent increase in temperature, fusion of hydrogen nuclei to helium became the dominant process, the energy released halting contraction and maintaining the core temperature at approximately 12×10^6 K. This stage in the evolution of the Sun is referred to as the zero-age main sequence, the effective point at which the Sun can be considered as a star.

The subsequent main sequence phase, during which the Sun continues the fusion process of hydrogen to helium, has an estimated duration of 10^{10} years, and is characterized by a slow expansion of the Sun together with a steady increase in luminosity. The main sequence phase will cease when the hydrogen in the core has been exhausted, and the Sun will evolve into a red giant, then a red supergiant and finally to a white dwarf star (see RED GIANT STARS, WHITE DWARFS and STELLAR EVOLUTION).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Feeds

Observations of the Sun typically involve modest aperture and a large and heavy spectrograph. In the field at ECLIPSES the telescope itself may need to have a long focal length. These requirements are conveniently met by having a tracking flat mirror which feeds a stationary telescope and its instrumentation. Magnetograph observations, which involve the polarization analysis of spectrum lines, place an additional constraint by demanding that the imaging system introduce a minimum of polarization. Geographic latitude of the site is another factor. These various considerations have led to a variety of what may be called 'solar feeds'.

A mirror which tracks a celestial object and delivers the light in an arbitrary but fixed direction is called a *siderostat*. In general, the mirror must move at a non-uniform rate, the image field rotates non-uniformly and linear polarization results from the oblique mirror reflection. Percentage polarization increases with the angle of reflection.

If the optical path is confined to the polar axis, either by placing the telescope there or by introducing a second mirror, conditions simplify and we have a *heliostat*. For the heliostat the polar axis of the mirror turns at the diurnal rate and the change of declination for the Sun is accommodated as in a conventional EQUATORIAL MOUNTING. The image plane unfortunately also rotates at the diurnal rate. This means that for other than short exposures the instrument package (spectrograph) must co-rotate. Also the amplitude of introduced feed polarization is constant during the day (ignoring any second flat), but its plane rotates.

If the mirror surface is parallel to the polar axis we have a *coelostat*, figure 1. Now either the telescope must move in angle in the meridian plane or, more commonly, a second mirror is introduced which must translate for objects at different declinations to intercept the beam and direct it to the stationary telescope. The coelostat mirror rotates at half the diurnal rate. Most important the image plane does not rotate, although it does change if the second mirror is translated. Compared with a heliostat, the per cent polarization may be less depending on mirror positions, although it will vary through the day.

Because of the deleterious effects of feed polarization, particularly for full vector magnetographs, and taking into account the availability of CHARGE-COUPLED DEVICE (CCD) arrays which mean smaller image scales, the present trend is toward compact systems in which the entire telescope points at the Sun. This avoids oblique reflections and is achieved by *altazimuth* or *equatorial* mounts.

Solar mirror feeds remain popular for eclipse work. For a site near the equator a heliostat is convenient since the exposures are short and the telescope nearly horizontal, figure 2. A coelostat, with its fixed image, is well suited for observations near the solstice, since declination change is then at a minimum. If an altazimuth mounting is considered, remember that the Sun disappears at totality where no guiding is possible.

Figure 1. Large coelostat feed for the Kitt Peak Vacuum Telescope. Here the number 2 mirror is adjustable in height and angle only. The coelostat itself swings along a circular path to account for seasonal and diurnal change. This scheme was first used at the Potsdam Observatory in Germany.

Figure 2. Eclipse site near the equator with a heliostat arrangement.

Other solar feeds include the *polar siderostat*, which is the same as a heliostat except the beam is directed toward the pole, and the *uranostat*, which is a siderostat which operates or tracks correctly only near the meridian (see also SIDEROSTAT/HELIOSTAT/COELOSTAT).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Manley P L 1991 *Unusual Telescopes* (Cambridge: Cambridge University Press)

William Livingston

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Filament Channels

A filament channel is a magnetic rift that runs through the chromosphere along the base of a solar filament and beyond either end. Because the filament and its channel separate magnetic fields of opposite magnetic polarity at the solar surface, field lines will cross over them at some height as an arcade of closed loops. A channel is a rift in the sense that it partitions positive and negative fields by directing its own magnetic field predominantly along, rather than across, a horizontal, axial direction. A channel develops before its filament forms and can outlive it by a large margin. Channels are therefore an essential ingredient of the prominence-forming environment (see SOLAR PROMINENCE FORMATION). Although filament channels were first recognized 30 years ago, investigations into their origins are recent. We now believe that both filaments and their channels derive from the evolution of surrounding magnetic fields at the solar surface.

The clearest signature of a channel is a pattern of chromospheric fibrils (see CHROMOSPHERE: FIBRILS) aligned in a band along a magnetic polarity boundary. This pattern is not present at every polarity boundary; the bands of fibrils appear only where overlying magnetic fields are suitably configured to form and support a solar filament. Not every filament channel contains a filament, nor does every polarity inversion develop an associated channel. The term polarity inversion line (PIL) signifies a line traced along the photospheric surface where the normal component of the magnetic field ($B_{\|}$) reverses sign. The visibility of a channel depends upon the strength of the fields adjacent to its PIL.

Historical background

Dark chromospheric filaments have been familiar for over a century as bright prominences at the limb carried onto the solar disk by the Sun's rotation (see SOLAR PROMINENCES). Yet channels were, until recently, overlooked, chiefly because they are often veiled by optically thick matter in their overlying filaments. They need to be viewed in large-scale images (resolution $\sim$1") at wavelengths where filaments become optically thin. Furthermore, the significance of filament channels was not appreciated until chromospheric images could be compared with maps of the underlying photospheric magnetic fields.

Solar astronomers have long observed the entire chromospheric surface by making monochromatic images in the core of the strongest Fraunhofer absorption lines, usually the Balmer α line of hydrogen and the H and K resonance lines of ionized calcium. Instruments used for this purpose are either the spectroheliograph or the tunable birefringent monochromatic filter (see SOLAR TELESCOPES AND INSTRUMENTS: GROUND). The resulting images are recorded as spectroheliograms and filtergrams, respectively. Observers frequently liken extensive patterns of short dark fibrils in Hα images to patterns of iron filings sprinkled over a sheet of paper placed above a bar magnet. The mapping of magnetic fields anywhere on the Sun thus became a motivating factor for the invention of the solar magnetograph in the 1950s by H W and H D Babcock (see SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY). The magnetograph's mapping capacity led to their discovery that solar filaments are located along PILs.

Where a PIL passes between bright chromospheric plages of opposite magnetic polarity (see SOLAR CHROMOSPHERIC: PLAGES), its corresponding filament is shorter and narrower than a quiescent prominence (see SOLAR PROMINENCE OBSERVATIONS) surrounded by weak fields. When observed at wavelengths with lower optical thickness in the chromosphere than the core of Hα, a channel surrounded by strong fields appears as an elongated void conspicuously free of absorbing structures. It is also called a plage corridor, or plage couloir, following M-J Martres and her colleagues at Meudon Observatory, who first described the phenomenon in the 1960s.

Solar magnetographs were few and far between in the early years of space exploration; even fewer could measure all the Stokes parameters for deriving vector magnetic fields to test SOLAR FLARE MODELS. Pressing demands to forecast trends in solar activity, both in the long and short term, stimulated attempts in the 1970s to use monochromatic Hα images in conjunction with maps of $B_{\|}$ as a proxy for a chromospheric vector magnetograph. H Zirin and co-workers at Big Bear Solar Observatory (BBSO) coined the term *chromospheric magnetograph* and championed the method as a viable option for inferring field directions in some magnetic configurations from Hα filtergrams. The Hα line has the combined advantages of large opacity and large scale height at a geometrical height where magnetic flux tubes, penetrating the surface, rapidly expand laterally to form a magnetic canopy (see CHROMOSPHERE: MAGNETIC CANOPY). Accordingly, the coarser fine structures in Hα, being nearly horizontal wherever conditions favor canopies such as the borders of plages or the outer chromospheric reaches of sunspot penumbrae (the *superpenumbra*), indicate the location and direction of the transverse component of the local magnetic field in the chromosphere.

Because large-scale patterns of solar magnetic activity evolve slowly, i.e. often lasting many days, Hα filaments and patterns of Hα fibrils follow suit even though they may dissolve or erupt and then reappear several times in response to small local changes. Evolutionary changes are slow enough that a panoramic, or 'synoptic', view can be constructed from the daily Hα filamentary structures to summarize the distribution of magnetic activity at each longitude during a full rotation of the Sun. P S McIntosh pioneered the use of *synoptic maps* based on Hα observations to forecast trends in solar activity from one solar rotation to the next. He recognized magnetic polarity boundaries as key structural elements for describing how global surface activity is organized. To infer their locations he used filaments, plage corridors, emerging flux regions with arch filament systems (see CHROMOSPHERE: EMERGING FLUX REGIONS), and alignments of Hα fibrils. Patterns of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

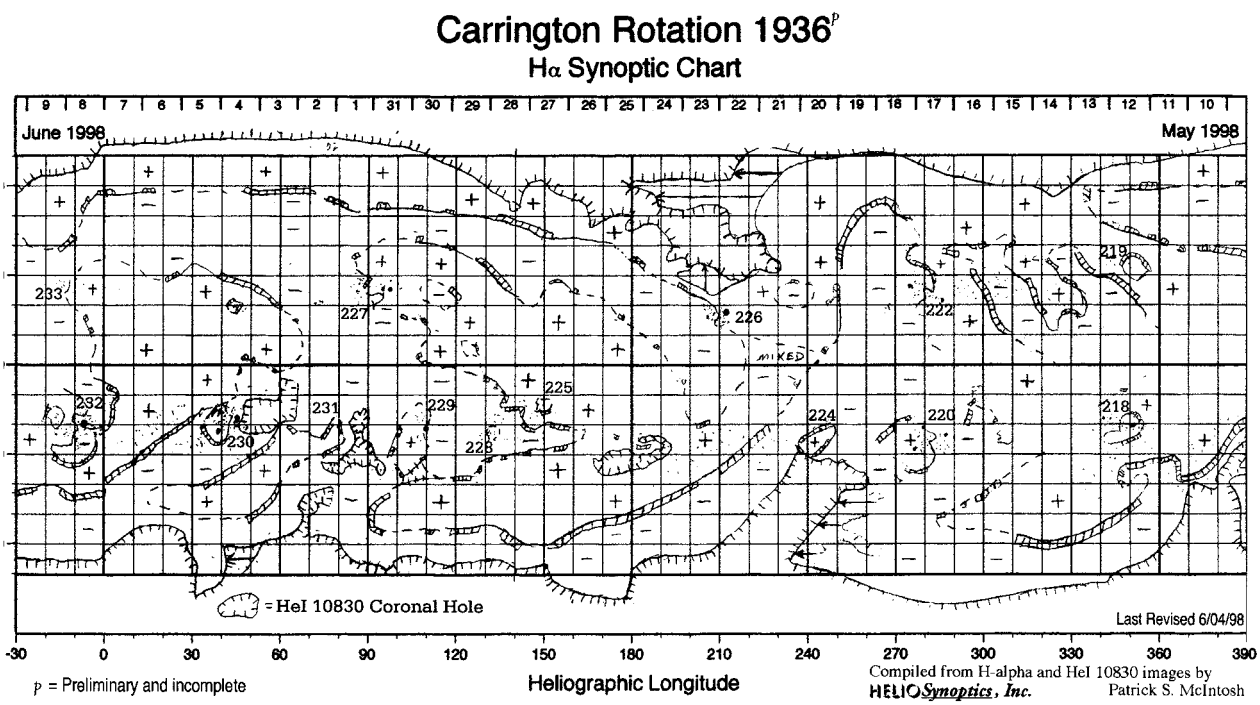


Figure 1. This synoptic map is constructed from chromospheric elements: filaments, plage corridors, emerging flux regions with arch filament systems , and alignments of Hα fibrils. It also provides the boundaries of coronal holes based on spectroheliograms made in the λ 1083 nm line of He I. (Courtesy of P S McIntosh.)

aligned fibrils extending axially far beyond either end of a filament, or patterns of axially aligned fibrils without a filament forming among them, were called 'filament channels' by McIntosh. In his accompanying synoptic map (figure 1) channels are shown as dashed lines joining Hα filaments. They are extrapolated to form proxy PILs which wind their way through, around, and between solar activity complexes, over vast distances.

Chromospheric properties of filament channels in strong background fields

The earliest descriptions noted an inverse relation between the width of chromospheric corridors and the strength of the field gradient across their corresponding PILs. This effect is illustrated by three channels in the small activity complex of figure 2. Two sunspots, S1 and S2, of following (black) magnetic polarity (see SUNSPOT CLASSIFICATIONS) in the magnetogram in panel C mark the locations of the strongest concentrations of flux in this complex. The remaining flux is distributed among plages of both leading (white) and following magnetic polarities: bright Hα plages in panel A coincide with strong flux concentrations outside of the sunspots in C, faint plages with the weaker flux concentrations. Three Hα filaments (F1, F2, F3 in panels A and C) and three corresponding channels (C1, C2, C3 in panel B) separate plages, or plages and sunspots, of like polarity from plages of opposite polarity. The narrowest channel, C2, is a narrow crack about 1 Mm wide among the optically thick absorbing structures just above the characters 'C2' in panel B; it is located close to spot S1 at a PIL within a steep gradient of $B_{\parallel}$. The widest channel, C3, appears as a broad void between absorbing features in panel B, and over a PIL with a shallow gradient between widely spaced flux concentrations. A channel of intermediate width, C1 in panel B, separates two bright plages.

A more exact description of a chromospheric channel includes the orientation of Hα fibrils along its fringes. In figure 2, fibrils rooted in the bright plage above channel C3 (panel B) are oppositely inclined to the fibrils rooted in the fainter plages of opposite magnetic polarity below C3, as indicated by the curved arrows. P Foukal first drew attention to such patterns where Hα fibrils are oppositely oriented on either side of a filament channel. He concluded that this characteristic pattern arises from a magnetic field with a predominantly horizontal component along, and not across, the axial direction of the channel.

S Martin and associates at BBSO applied Foukal's result to many quiescent filaments and their underlying channels. They discovered a hemispheric asymmetry for filaments based on the patterns of oppositely oriented fibrils. When channels are viewed from the positive polarity side, a majority in the northern (southern) hemisphere have fields directed to the right (left) as determined from the orientation of the fibrils fringing the channel. Thus the channel field has a definite orientation—either dextral or sinistral—which predominates in the northern and southern hemispheres

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

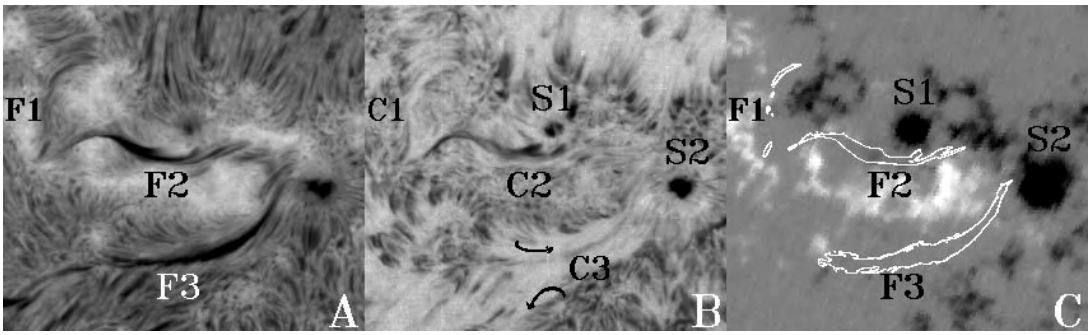


Figure 2. Filament channels in McMath 14726 observed at the Ottawa River Solar Observatory (ORSO) on 18 April 1977: (*a*) Hα filtergram; (*b*) cospatial filtergram at Hα + 0.6 Å, with voids between plages at channels C1, C2, C3; (*c*) cospatial NSO/KP $B_{\|}$ magnetogram with contours of three filaments superposed. S1 and S2 are both p-polarity spots (black is negative). Top N; right W.

respectively. Channels with a common orientation can join to form very long and very persistent structures. Channels can exceed in length and duration any filament which forms within them.

Chromospheric properties of filament channels in weak background fields

Filament channels, seen as aligned bands of Hα fibrils, are seldom found alongside quiescent filaments which border the Sun's polar caps or, more generally, in weak magnetic fields. The example in figure 3 (middle panel) contains Hα filaments formed in segments along a single outstretched PIL which winds from high to low latitudes around an extended area of predominantly negative (black) polarity flux (figure 3, bottom panel). Because this PIL reverses direction in an east–west sense by bending around a centre of curvature located to the west of the reversal, it is called a *switchback*. Fibrils next to the filaments in the polar arm (polar crown filaments, or PCFs) are not preferentially aligned with respect to the PIL; fibrils in the arm at lower latitudes do align strongly with the PIL in the dextral orientation, which is characteristic of a filament channel in the northern hemisphere. White contours in each panel of figure 3 outline narrow arteries following several PILs (with and without filaments) of reduced absorption in the He I line at λ1083 nm (upper panel). Brightness inside these arteries is comparable to that of the small CORONAL HOLE outlined in black at the right-hand edge of each panel. The bright He I arteries, considered equivalent to Hα filament channels because they correspond closely in width and location in stronger fields at lower latitudes, have the advantage of remaining visible among weak magnetic fields elsewhere.

Yet the Hα and He I signatures of channels differ in essential characteristics. The Hα signatures have a radiative and a morphological component: a void observed at wavelengths where the Hα line is optically thin informs us about chromospheric excitation conditions along a PIL, while the orientations of fibrils inform us about the geometry of the magnetic field in the chromosphere at a PIL. The He I signature is primarily a radiative one from a more complex regime of densities and temperatures characteristic of both the chromosphere and the corona. The same can be said for other diagnostics which have been applied in the recent past to the detection of filament channels at wavelengths ranging from soft x-rays to sub-millimetre radio waves.

The absence of an Hα signature for a filament channel at a PCF is probably due to weaker ambient magnetic flux densities and to their greater height relative to the corresponding factors for a quiescent filament at lower latitudes. But an elevated axial field of a PCF can leave its imprint on the chromosphere in other ways: the horizontal field inhibits downward (i.e. cross-field) heat conduction from the corona, thereby affecting the balance and transport of energy directly beneath it. For example, the λ1083 nm line of He I forms through an interplay between direct collisional excitation of the relevant energy levels of He I within the ambient plasma and photoionization–recombination of those same levels driven by overlying coronal sources (see SOLAR SPECTROSCOPY AND DIAGNOSTICS). Obstruction of downward heat conduction could, in principle, contribute to the reduced absorption seen as a bright He I artery at a PCF. The actual mechanism that dominates the formation of the λ1083 nm line can be determined from models, provided that the local circumstances are properly defined. Here atmospheric inhomogeneities play significant roles, the most relevant being the CORONAL CAVITY—plasma at coronal heights enveloping a prominence where density rather than the coronal temperature is reduced. H S Hudson, K L Harvey and associates have recently used simultaneous observations in soft x-rays (Yohkoh/Soft X-ray Telescope) and the λ1083 nm line of He I (National Solar Observatory, Kitt Peak (NSO/KP)) to show that the outer boundaries of a coronal cavity coincide with the boundaries of a bright He I artery at a filament channel. The chief factor responsible for locally weakening the λ1083 nm absorption and thereby creating a bright artery is, according to this evidence, the reduced coronal illumination from the rarefied cavity overlying the filament channel. Further simultaneous, multiwavelength

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

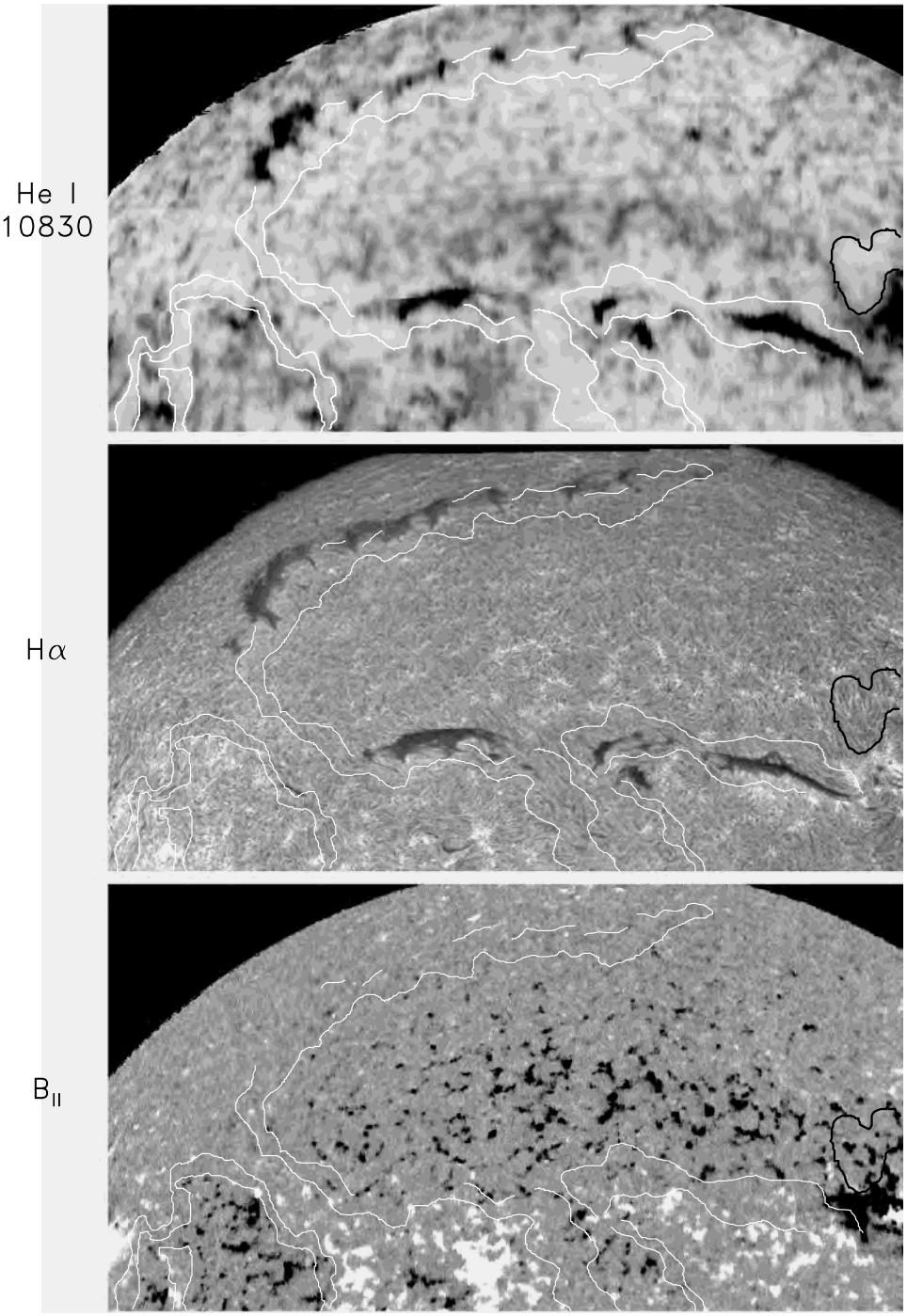


Figure 3. Top: NSO/KP He I λ1083 nm spectroheliogram; middle: ORSO Hα filtergram; bottom NSO/KP magnetogram, all of the same portion of the northern hemisphere of the Sun on 14 July 1979. The bright arteries in He I λ1083 nm are outlined in white and a small coronal hole is outlined in black in all three panels. Top N; right W. (Courtesy of K L Harvey).

observations with space- and ground-based instruments will soon be forthcoming to clarify the link between chromospheric filament channels and coronal cavities.

Creation of filament channels

The initial conditions for creating a filament channel—based on empirical criteria identified in the 1980s by S F Martin—are the same as those for solar prominence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

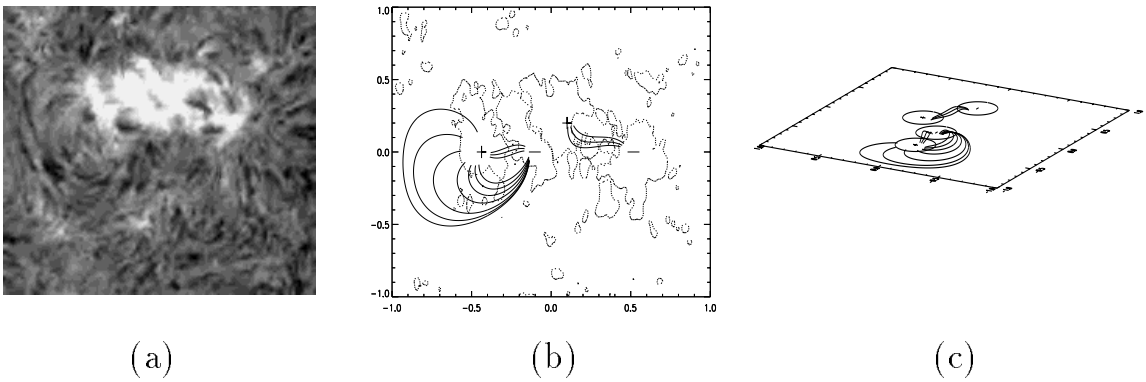


Figure 4. Field lines in a filament channel. (*a*) Hα fibrils around an activity complex (ORSO image). (*b*) Field lines deduced from force-free model superposed on a contour of the NSO/KP magnetogram. (*c*) Oblique view of field lines in (*b*). (Courtesy of D H Mackay.)

formation: mass flows transporting magnetic flux of opposite polarities to a PIL, cancellation of opposite polarity flux in these converging flows, and a magnetic arcade overlying the PIL. During the process of formation, magnetic shear is created above the PIL, i.e. the magnetic field in a vertical plane above a PIL departs from an ideal potential (current-free) configuration (see MAGNETOHYDRODYNAMICS). This departure is manifestly obvious in the axial alignment, described above, of chromospheric fibrils in a filament channel.

Models of solar prominence formation have been put forward over many years; not until the past decade have any addressed the formation of filament channels. Opinion about the formation of channels is divided between two main schools of thought (see SOLAR PROMINENCE MODELS). One school creates a channel at the solar surface by converging mass flows of already emerged magnetic elements onto a PIL so that oppositely directed field lines of the surrounding arcade reconnect but only the axial component survives. The other school produces an axial component of the field below the surface by making differential rotation act on subsurface magnetic fields (see SOLAR INTERIOR: ROTATION). In the latter case the filament channel forms when the subsurface axial field emerges at a PIL. Of the two hypotheses, the first is more amenable to direct observational tests.

The most complete observational description to date of a channel in the act of formation, by V Gaizauskas, J B Zirker, C Sweetland and A Kovacs, supports instead a process by which an existing magnetic topology at the surface readjusts to the arrival and rapid evolution of new magnetic flux from below. In this scenario, an existing slowly decaying bipolar plage trails behind a large field-free area in which new flux emerges not as a single bipolar region but as two, arriving in quick succession and almost on top of one another, thereby forming a compact activity complex. The placement of the bipoles, old and new, results in a regular alternation of magnetic polarity along almost the same parallel of latitude, thus ensuring a PIL between the existing plage and the new cluster of sunspots. The nearly superposed emerging bipoles create a highly non-potential magnetic configuration with nearly horizontal field lines low in the atmosphere. Normal evolutionary expansion of the new active regions then drives this configuration with its strong horizontal field component towards the PIL.

The visible result of these actions appears at the extreme left of figure 4(*a*) as an alignment of many Hα fibrils which happen to be superposed on a PIL. That alignment develops within a few hours which coincide with rapid growth of the second bipole in the compact activity complex. These Hα fibrils stream in opposite directions on opposite sides of that PIL—thus confirming that the aligned pattern is a filament channel. According to the Hα observations the channel is maintained on succeeding days. But it takes three more days before a filament forms in it. By using a force-free approximation (see FORCE-FREE MAGNETIC FIELDS) to model the magnetic field due to this distribution of magnetic sources, D H Mackay was able to fit the pattern of fibrils with a system of horizontal loops (figure 4(*b*)). The flatness of the magnetic loops is emphasized in the oblique view of figure 4(*c*). Mackay also applied the force-free model to represent the filament itself, and showed that the representation would fail without the presence of the pre-existing plage. This combination of observations and models shows that it is possible to attribute the formation, at least among active regions, of a filament channel and a filament entirely to normal evolutionary processes acting on flux after it has emerged. No recourse is needed to a hypothetical submerged flux rope.

Although this channel and its filament are small their environment is a typical one within the belts of active latitudes—new flux emerging in an activity complex next to decaying flux. They are special only in the fact that emergence occurred in an area free of any magnetic activity which could have complicated and confused the evolving patterns of chromospheric fibrils. The real merit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of this example lies in hinting at an alternative source for providing magnetic fields with a strong horizontal component: currents associated with emerging activity complexes that distort the magnetic topology into a highly non-potential configuration. Horizontal fields are naturally resistant to decay or to submergence, although they will disrupt due to plasma instabilities of various kinds (see SOLAR PROMINENCE ERUPTION). When confined as channels at a PIL they can also be replenished by reconnections between numerous canceling magnetic elements converging onto the PIL.

So one way to produce channels on the scale and duration of the ones outlining the switchback of figure 3 is to link up channels with common orientation that have been created in the belts of active latitudes by flux emergence of many neighboring activity complexes and then transport them poleward. The push can come from migrating flux driven poleward by large-scale emergences of new flux at lower latitudes combined with meridional flows. In this scenario filaments at the polar crowns are created when the migrating flux reconnects with fields already present in the polar caps. At polar latitudes an arcade overlying the PIL between the polar cap and an approaching surge of migrating photospheric flux arches well up into the corona. Field lines anchored in converging, sheared photospheric flows of opposite polarity flux near the polar caps are thus more likely, on reconnection, to leave an axial field suspended in the corona well above the chromosphere. In effect, the 'channel' for a polar crown filament as defined by the axial field leaves little or no imprint on the chromosphere, which is consistent with the situation illustrated in figure 3.

The elegant model of global flux transport by van Ballegooijen, Cartledge and Priest contains some of these features but does not include any role for reconnection of field lines in arcades overlying PILs to produce the axial horizontal field. The combined effects in this model of differential rotation, magnetic diffusion (random dispersal of magnetic elements by supergranulation (see SOLAR PHOTOSPHERE: SUPERGRANULATION)), and meridional flows acting on distributions of flux emerging at low latitudes result in axial fields located around switchbacks as observed. The orientation of the axial fields in the low-latitude diagonal arms of switchbacks derived from their model is consistent with the global organization of the orientation of channels found by Martin, Bilimoria and Tracadas. However, the orientation of channels in the arms of switchbacks at the polar crown predicted by this model of global flux transport are in conflict with observations. Some essential aspects are still missing in the physics of formation of the axial field in polar crown filaments. New observations made simultaneously from spacecraft and the ground of polar crown filaments are already planned to identify the missing ingredients.

Bibliography

Gaizauskas V, Zirker J B, Sweetland C and Kovacs A 1997 Formation of a solar filament channel *Astrophys. J.* **479** 448–57

Martin S F, Bilimoria R and Tracadas P W 1994 Magnetic field configurations basic to filament channels and filaments *Solar Surface Magnetism (NATO ASI Series C 433)* ed R J Rutten and C J Schrijver (Dordrecht: Kluwer) pp 303–38

Ruzdjak V and Tandberg-Hanssen E (ed) 1990 *Dynamics of Quiescent Prominences (Lecture Notes in Physics 363)* (Berlin: Springer)

van Ballegooijen A A, Cartledge N P and Priest E R 1998 Magnetic flux transport and the formation of filament channels on the Sun *Astrophys. J.* **501** 866–81

Webb D F, Schmieder B and Rust D M (ed) 1998 *New Perspectives on Solar Prominences (IAU Coll. 167)* (San Francisco: Astronomical Society of the Pacific)

Victor Gaizauskas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flare Classification

SOLAR FLARES are classified according to their size, duration, morphology or magnetic topology and the composition of their associated energetic particles. This article will cover the more enduring and comprehensive of these various classification schemes.

Solar flares are wonderfully complex phenomena and are observed across the electromagnetic spectrum as well as via energetic particles in space. The first reports of a flare, observed in white light by Carrington and Hodgson in 1859 (see SOLAR FLARES: WHITE LIGHT), captured their essence as transient energy releases in SUNSPOT regions. Following Hale's invention of the spectrohelioscope during the 1920s, flares were extensively studied in the Hα emission line. The 1940s saw the first observations of flare-associated radio and particle emissions and during the space age the full gamut of flare x-ray, extreme-ultraviolet and gamma-ray data (SOLAR FLARES: GAMMA RAYS) became available along with observations of accompanying transient coronal activity.

Not surprisingly, the successive opening of new observational windows to the flare phenomenon has resulted in the development of classification schemes particular to given wavelengths. These classification schemes encompass flare size and morphology, and, in the case of flare-associated energetic particles, the particle composition. If there is one overarching theme, it is that solar flares, despite their seemingly infinite variety, come in two basic flavors—impulsive and gradual. Impulsive flares have time scales of the order of a few minutes or less while gradual flare durations range from tens of minutes to several hours. These two basic types are combined in fully developed flares in which an impulsive phase is followed by a gradual main phase (see SOLAR FLARES: IMPULSIVE PHASE; SOLAR FLARES: MAIN PHASE). The time scale (gradual versus impulsive, short versus long-duration) distinction between the two basic kinds of flares has been noted by investigators working in virtually every wavelength and energy range. During the 1970s, it became appreciated that these time scales were indicative of magnetic topology; long-duration flares were linked to SOLAR CORONAL MASS EJECTIONS (CMEs) and impulsive flares were thought to be completely confined within the Sun's lower atmosphere. Today it is recognized that some short-duration flares may also have ejecta. Nevertheless, the concept of two fundamental types of flares (impulsive/short duration/confined versus gradual/long duration/eruptive) frames current thinking on this subject.

Size classification of solar flares: Hα and soft x-ray

There are only two widely used classification systems that address flare 'size' or 'importance'. The first of these is the Hα classification scheme (table 1) that was developed during the 1930s. It consists of a character (S = subflare, or 1, 2, 3, or 4 for successively larger flares) that denotes the flare size and a letter (f = faint, n = normal, b = bright) corresponding to a subjective estimate of the intensity of the emission. Thus the most outstanding flares are classified as 4b and the smallest and faintest as Sf. In general, the larger the optical classification of a flare, the more likely that it is a gradual (long-duration) or eruptive flare, but this is only a rule of thumb.

Table 1. Optical classification of solar flares.

Importance class	Area (A) at disk center (millionths of a solar hemisphere)
S	$A < 100$
1	$100 \leq A < 250$
2	$250 \leq A < 600$
3	$600 \leq A < 1200$
4	$A \geq 1200$

A second size classification that has come into common usage since about 1970 is based on soft x-ray observations of the Sun in the 1–8 Å band by Earth-orbiting satellites. The size of the flare is given by the peak intensity (on a logarithmic scale) of the emission. Flares are classified with a letter (A, B, C, M or X) corresponding to the power of 10 (-8, -7, -6, -5, -4, respectively) of the peak 1–8 Å flux in units of W m^{-2} and a number (1–9) that acts as a multiplier. For example, a B3 flare has a peak flux of 3×10^{-7} W m^{-2} and an M8 flare has a peak flux of 8×10^{-5} W m^{-2}. Generally, flares smaller than C1 can only be detected during a SOLAR CYCLE minimum when the x-ray background is low. Flares occasionally exceed class X9 in intensity; they are simply referred to as X10, X11 etc events. Note that the soft x-ray classification gives no indication as to whether the flare is eruptive or compact. While there is a general tendency for intense flares to be eruptive, X-class flares can be confined and B-class flares can have CME associations.

Classification of flares in terms of magnetic topology: confined versus eruptive

The discovery of coronal mass ejections in about 1970 gradually led to a profound change in the way that solar astronomers think about flares. The word 'flare' implies an event that is localized within an active region while CMEs leave the Sun at speeds of up to 2000 km s^{-1} and can have angular spans that encompass several active regions. At first it was thought that CMEs were blown out by the explosive release of energy during the flare, whereas today the prevailing view is that eruptive flares are a consequence of CMEs. Figure 1 captures much of the rich phenomenology associated with eruptive flares. The essential aspect from a flare classification point of view is that magnetic field lines are opened by CMEs in eruptive events. The closing down or reconnection of these newly opened field lines typically occurs on a time scale of hours and provides the prolonged energy release characteristic of gradual or eruptive flares, although it must be noted that CMEs can also be accompanied by short duration flares. As can be seen in figure 1, the intersection of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

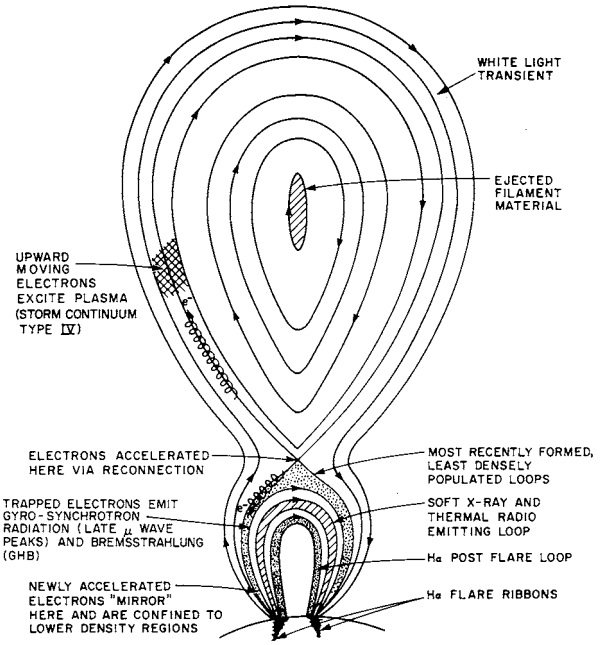


Figure 1. Magnetic topology of a gradual or eruptive flare (from Cliver *et al* 1986 *Astrophys. J.* **305** 920).

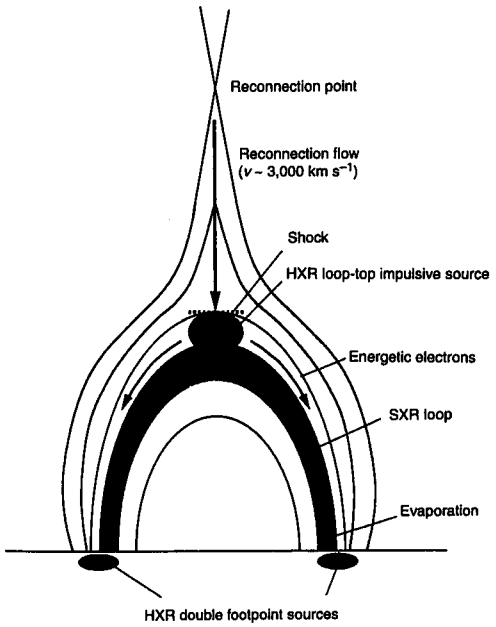


Figure 2. Magnetic topology of an impulsive or confined flare (from Masuda *et al* 1994 *Nature* **371** 495).

newly formed flare loops (or 'post-flare' loops as they were previously called) with the solar surface is marked by two parallel ribbons in Hα. Thus, before the role of the CME was fully appreciated, eruptive flares were referred to as two-ribbon or double-ribbon events. This type of flare morphology is also observed in association with eruptive prominences outside active regions (such events are often referred to by the French term 'disparition brusque'). Today it is generally thought that the basic physics of eruptive prominences and large two ribbon flares is identical and both phenomena are referred to as eruptive flares.

While eruptive flares attract much attention because of their size, complexity and association with GEOMAGNETIC STORMS and Forbush decreases of cosmic ray intensity, they comprise only a small fraction of all flares. The vast majority of flares are short-lived or impulsive events that lack CMEs. Such flares were traditionally modeled in terms of energy release within a single static magnetic loop and were thus referred to as single-loop, point, compact or confined flares. However, recent images of flares in energetic x-rays have indicated that confined or impulsive events may also result from loop top magnetic reconnection as shown in figure 2. In that case, the difference between the eruptive and confined classes of flares may be more one of degree than of kind, since the reconnection event depicted in figure 2 will result in the detachment of a narrow magnetic loop into the SOLAR WIND. The idea that the particle acceleration process is similar in impulsive and eruptive flares (or in the impulsive and gradual phases of fully developed flares) finds some support from gamma-ray observations that indicate similar compositions of interacting particles for both types of events.

The most important flares in terms of their geophysical consequences are large fully developed events in which an impulsive phase is followed by a gradual main phase. Such 'hybrid' flares are classified according to their end phases as eruptive flares. Occasionally, disparition brusque events, which represent a 'pure' form of eruptive flares, can also result in severe magnetic storms.

Classification of meter wavelength radio bursts associated with flares

Solar flares have associated radio events, or 'bursts', at wavelengths ranging from millimeters to kilometers. The radio burst classification scheme developed during the 1950s by Australian and French solar radio astronomers has proven to be particularly useful in terms of the current division of flares into confined and eruptive categories (see SOLAR FLARES: RADIO BURSTS). A schematic of the meterwave dynamic spectrum corresponding to a large eruptive flare is shown in figure 3. In terms of the eruptive/confined classification scheme, burst types III and V are characteristic phenomena of impulsive flares (or the impulsive or initial phase of fully developed eruptive flares). Type III bursts and their associated type V continua are attributed to flare-accelerated electrons moving along open field lines into the corona. Type II and type IV bursts are most commonly identified with eruptive flares. Type II or 'slow-drift' emission indicates a shock wave moving out through the corona with a characteristic speed $\sim$500 km s^{-1}. The shock exciter is either a flare-initiated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

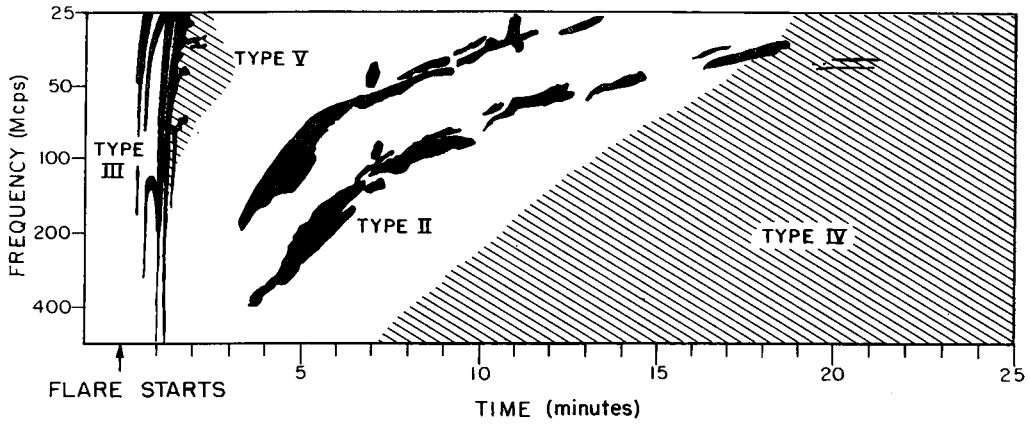


Figure 3. Metric radio emission associated with a fully developed eruptive flare.

blast wave or a CME. Type IV emission is thought to result from magnetic reconnection in the wake of the CME (figure 1).

Classification of solar energetic particle events

The energetic particle ($E > 1$ MeV/nucleon) events observed in space following flares were one of the last aspects of flares to be brought under the two-class picture (see SOLAR WIND: ENERGETIC PARTICLES). Prior to about 1985, there were a number of abundance anomalies associated with solar particle events that defied organization. These included a wide range of variation in electron (e) to proton (p) ratios, ^{3}He/^{4}He ratios and Fe/O ratios. In the mid-1980s it was recognized that flare time scale was a useful parameter that could order the data. In general, it was found that particles from impulsive flares have higher e/p, ^{3}He/^{4}He and Fe/O ratios than those recorded for 'normal' particle events (which include virtually all large events) that are associated with long-duration flares. In addition, heavy ions in large events have relatively low charge states corresponding to 2 MK temperatures, indicating acceleration high in the CORONA and in the solar wind rather than in hot (10 MK) flare plasma. Energetic particle events associated with eruptive flares can be observed over a much wider range of heliolongitudes than is the case for confined flares. Thus these large events are attributed to acceleration by coronal/interplanetary shocks driven by CMEs. Distinctions between impulsive and gradual events may be blurred in hybrid particle events for which both flare and shock components are observed. The current two-class picture of SEP events is summarized in table 2. The usage of 'impulsive' and 'gradual' in this table refers to both the flare and particle event duration.

Table 2. Properties of impulsive and gradual energetic particle events and their associated flares.

	Impulsive	Gradual
Particles:		
e/p	Electron-rich	Proton-rich
^{3}He/^{4}He	∼1	∼0.0005
Fe/O	∼1	∼0.1
H/He	∼10	∼100
Fe charge state	∼20	∼14
Duration	Hours	Days
Longitude cone	<30°	180°
Flare:		
Type	Confined	Eruptive
Radio bursts	III/V	II/IV
Soft x-ray duration	<1 h	>1 h
CME?	—	Yes
Interplanetary shock?	—	Yes
Events/year	∼1000	∼10

Open questions and ongoing research

The separation of flares into two distinct classes has proven to be a useful framework that organizes a great deal of disparate data. At the same time, many key questions remain unanswered. A sampler of some of these open questions follows:

(a) While the relationship of CMEs to long-duration flares is generally interpreted in terms of magnetic reconnection, the role of CMEs in impulsive flares remains unclear. Nor is the relationship of the impulsive phase to the gradual phase in large fully developed flares understood. Does the impulsive flare serve to trigger a larger-scale eruption?

(b) There is some evidence that CMEs associated with prominence eruption events have different acceleration profiles from CMEs that accompany eruptive flares in active regions. Specifically, flare-associated CMEs undergo much more rapid acceleration than CMEs associated with disparition brusques. The basis for this difference is not understood and requires further investigation.

(c) The origin of the coronal shocks manifested as metric type II bursts remains a matter of controversy, with advocates for both flare and CME causes. The source of the chromospheric Moreton waves (or the waves imaged in extreme-ultraviolet wavelengths in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

corona) that have been identified with type II shocks is similarly unclear.

(d) Does the magnetic topology for gradual or eruptive flares depicted in figure 1 apply to all such events? In particular, there is a controversy as to whether magnetic reconnection is required for the formation of the magnetic flux ropes (corresponding to the upper 'disconnected' structure seen in cross section in figure 2) that are observed in interplanetary space.

(e) Does the magnetic topology for impulsive flares depicted in figure 2 apply to all such events? Intense impulsive flares are frequently observed that lack type III emission; this is contrary to expectation, given the open field lines at the acceleration site in the model.

(f) Can coronal/interplanetary shocks accelerate particles to the tens of GeV energies occasionally observed in solar particle events?

Bibliography

Cliver E W 1995 Solar flare nomenclature *Solar Phys.* **157** 285–93

——1996 Solar flare gamma-ray emission and energetic particles in space *High Energy Solar Physics* ed R Ramaty, N Mandzhavidze and X-M Hua (Woodbury, NY: American Institute of Physics) pp 45–60

Crooker N, Joselyn, J A and Feynman J (ed) 1997 *Coronal Mass Ejections* (Washington, DC: American Geophysical Union)

McLean D J 1985 *Solar Radiophysics* (Cambridge: Cambridge University Press)

Shibata K 1996 New observational facts about solar flares from Yohkoh studies-evidence of magnetic reconnection and a unified model of flares *Adv. Space Res.* **17** (4–5) 9–18

Švestka Z, Jackson B V and Machado M E 1992 *Eruptive Solar Flares* (Berlin: Springer)

Edward Cliver

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flare Forecasting

Like weather forecasting, solar flare forecasting (or forecasting solar activity in general) is motivated by pragmatic needs. Solar flares, coronal mass ejections, solar winds and other solar activity intimately influence the near-Earth space environment. All kinds of spacecraft including weather and communication satellites are orbiting Earth, and their performance and lifetimes are greatly influenced by solar activity. High solar activity increases the density of the stratosphere, which increases the atmospheric drag of low-Earth-orbiting satellites. In planning manned space activity such as space station fabrication and interplanetary travels, one must consider harmful effects of solar energetic particles and radiations.

Even on the ground, we are not free from the influence of solar activity. The Great Quebec Blackout of 13 March 1989 was caused by a big solar flare and its associated phenomena. When the geomagnetic field is abruptly changed by a magnetic cloud crashing onto it or by an interplanetary shock wave, strong electric fields are induced in large-scale electric power distribution systems. The blackout was a result of such a process. Additionally, geomagnetic storms make a compass useless for navigation. Short-wave communication is hampered when the ionosphere is disturbed by energetic UV and x-radiations.

In deploying solar experiments, one should also consider the level of solar activity. Spacecraft containing instruments for solar activity are usually launched near solar cycle maxima. For the same reason, it is better to launch a balloon carrying instruments for investigating solar flares during or just before an episode of high solar flare activity, if possible.

Even long before the advancement of meteorology, people began to forecast weather on the basis of phenomenology. Likewise, even though we do not fully understand the underlying mechanisms for various manifestations of solar activity, we should make best of our current knowledge in order to forecast solar activity more reliably.

Long-term forecasting

We utilize our knowledge of the local climate and seasons in order to predict the local weather far in the future. Likewise we can utilize the 11 yr sunspot cycle in order to predict the level of solar activity far into the future. The phase and the strength of the sunspot cycle are relevant information. However, the length of the 11 yr cycle, unlike the length of a year, is not fixed but can vary abruptly in a range from 7.3 to 17.1 yr. The maximum sunspot number is also highly variable from one solar cycle to the next. In spite of these, various empirical rules, although not strict, have been utilized to predict the upcoming sunspot cycle. For example, it has been noticed that an odd-numbered cycle is usually stronger than the preceding even cycle. (The sunspot cycle which peaked in the year 1760 is defined as cycle 1.) Inspecting the strengths of cycles 1–21, we find that this rule has been observed by all the even–odd pairs except two pairs (cycles 4 and 5, cycles 8 and 9). It remains to be seen whether cycle 23, which is expected to reach its maximum in the year 2000, is going to be stronger than cycle 22. However, considering the chaotic nature of sunspot cycles, which is inherent to many nonlinear systems, we cannot be sure. (See SOLAR ACTIVITY: LONG-TERM RECORDS.)

Another empirical rule is that a cycle of larger-than-average size usually follows a cycle of shorter-than-average period, while a cycle of smaller-than-average size usually follows a cycle of longer-than-average period. By the end of one sunspot cycle, one can guess the strength of the next sunspot cycle by using this rule.

These empirical rules have no obvious scientific justifications. On the other hand, the 'precursor technique' is widely accepted by the scientific community because it has a scientific basis. This technique is based on the concept of extended solar cycles. According to this concept, the true length of a solar cycle is longer than 11 yr (perhaps 18 yr), and a new solar cycle does not start near the end of the preceding solar cycle but starts much earlier in the declining phase of the preceding cycle. Therefore, for some years, two cycles coexist: the signatures of a new cycle are manifested in high latitudes, while sunspots belonging to the declining cycle are found in low latitudes.

According to the precursor technique, each of the geomagnetic indices *Ap* and *aa* is made of two components: one is due to the current-cycle solar activity and the other is due to the precursor of the next solar cycle. The second component is due to stable coronal holes, giving rise to a 27 day recurrent pattern. A stable coronal hole is a unipolar region with magnetic field lines open to interplanetary space. The magnetic fields of unipolar regions are thought to become seeds for the next sunspot cycle. The second component of the geomagnetic indices, therefore, can reveal the strength of the next cycle. On this basis, some solar physicists successfully retrodicted the peak strengths of cycles 12–22. Using this technique, solar physicists formed a consensus that cycle 23 is going to be much stronger than the average.

From the early days of solar activity study, it was attempted to express the sunspot cycle with analytical equations involving a few parameters. Recently, some scientists found that the solar activity curve for each cycle can be fit by a single equation with three free parameters, which are adjustable for each cycle. By fitting such an equation to the rising phase of a sunspot cycle, one can deduce the best values of the free parameters, which in turn determine the activity curve of the remaining part of the cycle. In this method, one has to wait until a new cycle has progressed considerably.

Mid-term forecasting

The prospect of mid-term forecasting of solar activity has not been studied in recent years. Here timescales from months to several years are regarded as mid terms. Scientists and forecasters have not paid much attention

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to mid-term forecasting because of the lack of useful tools. Recently, it has been found that the occurrence rate of solar flares episodically exhibits periodic modulations. The best-known periodicity is a 152 day periodicity, which was in operation during the four years from 1980 through 1983. Other well-known periods are 51 day and 76 day periods. These periods are close to integral multiples of 25.5 days (2, 3 and 6 times, respectively). Even though these periodicities are not persistent but episodic, it can be used for mid-term flare forecasting when one of them is found to be in operation.

By the end of 1981, one of the scientists in charge of the gamma-ray detector aboard the Solar Maximum Mission found that four episodes of high flare activity were roughly separated by 5 months. Suspecting the existence of 5 month periodicity, he predicted that the next high flare activity would occur in February 1982. Not only was this prediction realized, but also high flare activity was later observed with 5 month intervals (in June and July 1982, December 1982, May 1983 and October 1983). After it was recognized in March 1992 that the 51 day periodicity was in operation, four episodes of high flare activity occurred in accordance with this periodicity. Therefore, one can say that flare periodicities were, in fact, used for mid-term flare forecasting. In order to use flare periodicities as a practical tool for forecasting, however, one has to investigate the reliability and limitation of this method, analyzing the past data.

Short-term forecasting

Studies on flare forecasting have been mostly concentrated in short-term flare forecasting for timescales of a few days. The Space Environmental Lab of the National Oceanic and Atmospheric Administration of the United States of America mainly performs short-term flare forecasts rather than long-term forecasts. All the major flares occur within sunspot groups; hence, the term 'active region' is often used for a sunspot group. Therefore, investigation of the morphology and the development of a sunspot group is a useful tool for making short-term flare prediction. (See SUNSPOT CLASSIFICATION.)

Short-term flare prediction is based on both theoretical understanding of solar flares and hands-on experience of observers. Because a rather small number of highly flare-productive active regions account for the majority of energetic flares, it is important to recognize the signatures of flare-productive active regions. Well-known signatures of flare-productive active regions are as follows.

(i) Complex active regions in which many sunspots of opposite polarities are mixed tend to produce many flares. Even if an active region is made of many sunspots, it is not flare productive if sunspots of one polarity are neatly separated from those of opposite polarity. An active region with a delta configuration, which is defined as two umbrae of different polarities surrounded by a single penumbra, is flare productive.

(ii) If, in a delta configuration, a bright H-alpha plage obscures a part of an umbra or runs along the neutral line, it is very flare productive.

(iii) If an emerging sunspot bumps into an existing sunspot of opposite polarity, flares are likely to occur.

(iv) If magnetic field lines connecting the sunspots of opposite polarities are highly sheared (in other words, the angles between these field lines and the neutral line are small), flares are likely to occur.

(v) A rapidly growing active region is flare productive. In general, an active region in the decaying phase does not produce energetic flares unless new spots emerge within it.

Very-short-term forecasting

Flare prediction with timescales of tens of minutes is based on flare precursors. X-ray bright points, bright EUV knots and high-temperature emission of 6–20 GHz microwaves are recognized as precursors of flares. (See SOLAR FLARES: PREFLARE PHASE.) Such very-short-term flare forecasts are not usually valuable because of short lead times. However, it is useful to recognize early a flare that is likely to be associated with solar energetic particles and/or coronal mass ejections because energetic particles take hours to propagate to Earth and coronal mass ejections take days to do so. Therefore, even if we identify them while they are in progress on the Sun, we can have some preparation time before their impacts arrive here.

Flares associated with large fluxes of solar energetic particles, coronal mass ejections, and interplanetary shocks have been recognized as constituting a distinct class of flares, and they have been called 'proton flares', 'two-ribbon flares', 'gradual flares', 'extended flares', 'long-decay events', 'gradual gamma-ray–proton flares' and so forth. (See SOLAR CORONAL MASS EJECTIONS and SOLAR FLARE CLASSIFICATION.) This kind of flare has several characteristics: (1) long durations (of order of minutes) of individual hard x-ray spike bursts; (2) a long total duration (of tens of minutes) of hard x-ray emission; (3) spectral hardening of hard x-rays during individual spike bursts; (4) a gradual rise and very gradual decay of soft x-ray emission, (5) Two H-alpha ribbons separating gradually. One can use the first three characteristics for forecasting proton flares by analyzing hard x-ray data in real time. Although soft x-ray emission from the Sun is monitored continuously, the fourth characteristic has limited utility because the distribution of soft x-ray emission durations has no recognizable demarcation line. H-alpha observations are routinely available, but H-alpha ribbons begin to separate in the declining phase of a flare, limiting its usefulness as a timely identification method of proton flares.

Future prospects

Identification of flare-productive active regions discussed earlier is mainly based on high-resolution, ground-based observations of magnetic fields and H-alpha emissions of active regions. Now high-resolution instruments aboard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

various spacecraft are observing the Sun. The Michelson Doppler Imager (MDI) on the Solar and Heliospheric Observatory is measuring photospheric Doppler shifts with a spatial resolution of 2 arcsec and a good cadence. The Transition Region and Coronal Explorer (TRACE) is observing the transition region and the corona with high spatial resolution. Yohkoh, which has been taking soft x-ray pictures of the Sun since 1991, is still in good condition. When we analyze Yohkoh, MDI and TRACE data appropriately, we may be able to learn new methods of forecasting solar flares.

Using local helioseismology, one can detect sunspots before they emerge to the surface because they absorb p-mode oscillations and modify sound speeds. (See SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY .) This technique has not been perfected yet, but we can see its potential. It is impractical to apply local helioseismology to see whether sunspots are buried below the photosphere at random places and times. However, one can apply this technique to a developing active region to see whether a large reservoir of strong magnetic fields is beneath the surface.

With the abundance of data on solar phenomena recently accumulated and to be obtained in the near future, scientists will be able to better understand various solar phenomena influencing the Earth environment. It will naturally improve the ability to forecast solar flares and space weather in general.

Bibliography

Sawyer C, Warwick J W and Dennet J T 1987 *Solar Flare Predictions* (Boulder, CO: University of Colorado)

Thompson R J *et al* (eds) 1990 *Proc. Leura Solar Terrestrial Predictions Workshop* (Boulder, CO: NOAA)

Taeil Bai

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flare Magnetic Fields

This article describes the relationship between magnetic fields and solar FLARES.

Like most dynamic phenomena on the Sun, the occurrence of solar flares is closely related to the presence and evolution of solar magnetic fields, especially the complicated, non-potential magnetic configuration. As a simple introduction, we compare flares with earthquakes. The occurrence of earthquakes involves two steps: energy build-up and energy release. The stress and energy are built up by relative motion of two tectonic plates along fault lines. When the stress reaches a critical point, two plates cannot slide further, the equilibrium breaks down and a section of the fault line is restored to its original position (minimum energy state). The excess energy is released in the form of kinetic energy and propagated as seismic waves. A simple flare can be described by a very similar scenario. We start with two magnetic regions of opposite polarities. The line dividing two polarities is called the neutral line. If the field lines are connected directly, i.e. perpendicular to the magnetic neutral line, this is called the potential configuration. If, for some reason, magnetic regions slide, the field lines will start to move away from perpendicular to the neutral line. This is called the sheared configuration. In an extremely sheared configuration, the magnetic field lines are nearly parallel to the neutral line. A sheared magnetic field has more magnetic energy than the corresponding potential field configuration and the extra energy is called the free magnetic energy. When the fields are extremely sheared, an instability occurs. Fields tend to be restored to the potential configuration and magnetic free energy is released by the process of MAGNETIC RECONNECTION. The free magnetic energy is converted to thermal and non-thermal energy to power solar flares, which create thermal emissions and acceleration of non-thermal particles. The total energy of a flare depends on the total volume and complexity of the sheared magnetic region.

The determination of the magnetic field configuration in flare-productive regions can be done directly and indirectly. Most of the methods will be addressed throughout the article. Here is the summary of them. The most direct way is by the vector magnetograph observations, at present all from ground-based solar observatories. However, the Japanese–US mission SOLAR-B (launch date: 2004) will provide the first space measurements of this kind. The line-of-sight magnetograms can provide certain information of ACTIVE REGION magnetic structure too: data are available from space (e.g. MDI on board SOHO) as well as ground-based observatories. Indirect measurements of magnetic field structure include Hα, white-light and D3 images from ground-based observations and soft X-RAY (YOHKOH), UV/EUV (TRACE, EIT/SOHO) images from space observations.

In addition to the presence of sheared magnetic neutral lines, the complexity of active regions is also vitally important in producing solar flares. The close association between δ sunspots and major solar flares provides evidence that flares are associated with complexity of active regions (Zirin and Liggett 1987). The δ sunspots are SUNSPOTS having two or more umbrae of opposite polarities within a common penumbra. Figure 1 shows an example of a δ sunspot. It is a magnetogram of active region NOAA 6659, taken on 10 June 1991 by Big Bear Solar Observatory; white and black gray scales represent line-of-sight magnetic field of north and south polarities respectively. The dominant negative spots are surrounded by opposite-polarity satellite spots. This active region was one of the most flare-productive regions in solar cycle 22, which produced many flares including six of the strongest flares (classified as X-class flares). As a matter of fact, almost all of the more than 100 X-class solar flares that occurred in solar cycle 22 were associated with δ sunspots.

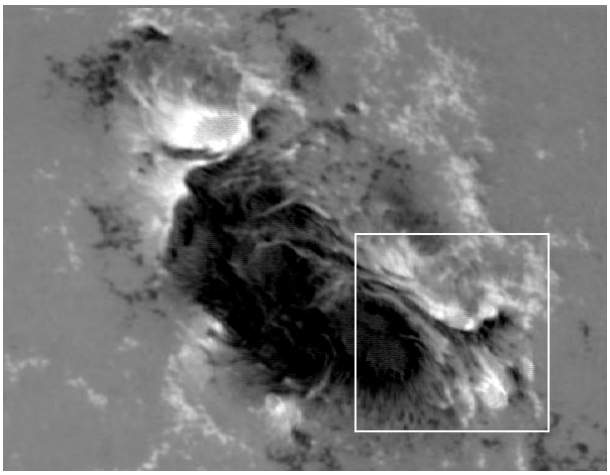


Figure 1. A line-of-sight magnetogram of 10 June 1991 active region NOAA 6659 obtained at Big Bear Solar Observatory. It is a δ sunspot group. The field of view is 5 arcmin $\times$ 4 arcmin. White and black gray scales represent field strength of positive and negative magnetic polarities, respectively. The dominant negative spots are surrounded by opposite-polarity satellite spots. In the lower-right corner (marked by a white box), a magnetic channel structure due to new flux emergence is visible.

Another prominent common feature in the flare-productive δ sunspot groups is the unbalanced magnetic flux inside the δ configuration. Again, figure 1 gives a very good example of an unbalanced magnetic flux configuration. Inside the δ group, the negative magnetic flux (field strength multiplied by area) is about 6 times greater than the positive flux. From theory, of course, we know that the overall fluxes must be balanced when added up over a large enough area. It is not clear why the flare-productive δ groups always show such very unbalanced fluxes. Statistical studies show that the average ratio of dominant flux versus minor flux is around 5 in all the δ groups.

Another very important aspect in flare magnetic field studies is the electric currents in flare-producing regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

When the field is potential, it is called the current free field. Sheared and twisted magnetic fields carry electric currents. The existence of currents is important for flares, because an electric field is most efficient in both heating (Joulean dissipation) and acceleration of charged particles.

The magnetic fields in flare-productive active regions have been observed by various vector magnetograph systems. The most notable ones are systems at Marshall Space Flight Center, Big Bear Solar Observatory, Huairou Solar Observatory, Mees Solar Observatory and Advanced Stokes Polarimeter at National Solar Observatory. Studies of the magnetic structure of flare-producing regions have been concentrated in the following areas.

Relationship between magnetic shear and solar flares

Study on this subject was first carried out extensively by the solar group at Marshall Space Flight Center/NASA. The magnetic shear is defined as the angular difference between the azimuth of the potential transverse field calculated from the observed longitudinal magnetic field and the observed transverse magnetic field (Hagyard *et al* 1984). Statistically speaking, flares occur in areas where the local photospheric magnetic field departs the most from a potential field. These regions are characterized by (1) high shear angle, (2) strong transverse magnetic field along the neutral line and (3) long sheared magnetic neutral line. Figure 2 shows a transverse magnetogram of 10 June 1991, obtained at Marshall Space Flight Center. It contains the top-left part of the active region shown in figure 1. Magnetic neutral lines are shown as contours and the transverse fields are represented as bars. These field components are superposed on a broad-band Hα image which shows that the flare kernels were on opposite sides of the neutral line in an area where the transverse field was highly sheared along that line.

However, more recently, it has been found that not all the flares occur in sheared neutral lines. If the flare ribbons start near the neutral line, the strong shear conditions are usually satisfied. If the flare ribbons start further apart, the neutral line may not have large shear (Chen *et al* 1994). The former situation may represent a reconnection closer to the photosphere, while the latter may imply a reconnection higher up.

Flux emergence and formation of magnetic shear

The picture of magnetic shear due to relative motion of magnetic regions that we gave in the introduction is oversimplified. There are several ways that a sheared magnetic configuration can form. A few examples are given here and have been discussed by Wang (1993, 1997): (a) shear formed by relative shear motion of magnetic regions with opposite polarities; (b) shear formed by head-on collision between two magnetic regions of opposite polarities; (c) complicated magnetic structure, perhaps formed as a result of new flux emergence, in which opposite magnetic polarities can be closely intertwined. The three most famous superactive regions in solar cycle 22 (NOAA 5395, March 1989; NOAA 6659, June 1991; NOAA 6891, October 1991) all show very similar new flux emergence inside the already existing penumbrae, and new fluxes subsequently formed magnetic channels (elongated mixed-polarity magnetic regions) with high magnetic shears. These new flux emergence regions are the sites of many major solar flares (see CHROMOSPHERE: EMERGING FLUX REGIONS). In the lower-right part of figure 1 (marked by a box), the magnetic channel structure is demonstrated for the June 1991 region, which was a result of new flux emergence inside the sunspot penumbra.

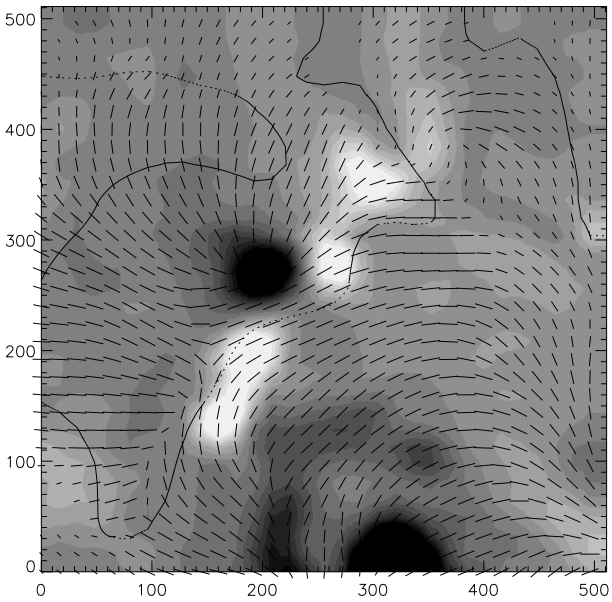


Figure 2. AR6659 on 10 June 1991, at the start of an M3 flare at 16:53 UT. The 1.5 arcmin square field of view shows the north-east portion (top-left) of the same active region as in Figure 1. The solid contour marks the magnetic neutral line separating the positive polarity of the umbra near the center of the picture from the negative polarity of the larger umbra at the bottom. The line segments indicate the magnitude and direction of the transverse component of the magnetic field and are scaled from 125 to 500 G. These field components are superposed on a broad-band Hα image which shows that the flare kernels were on opposite sides of the neutral line in an area where the transverse field was highly sheared along that line. This figure is provided by Dr M Hagyard at Marshall Space Flight Center.

It has also been brought to light that magnetic shears may not be caused by surface flows; they may instead be formed beneath the surface by convective motion. When flux emerges, the twisted flux tubes are brought up above the surface, which provides free energy to power flares. Such an idea was originally proposed by Tanaka (1991) based on observations from Big Bear Solar Observatory. In a recent paper, Ishii *et al* (1998) studied active region 5395 (March 1989) in detail. To explain the apparent complicated flux emergence and flow motions, they proposed a schematic model of the successive emergence of twisted and winding magnetic flux ropes coiling around a trunk of a big magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

flux tube. The location of strongest flare activity was found to coincide with the vortex-like apparent surface motion of sunspots which are actually a consequence of twisted flux emergence. Data obtained with the Advanced Stokes Polarimeter confirmed such an idea of emergence of twisted flux tubes (Lites and Low, 1997).

Changes of magnetic shear after solar flares

Before vector magnetograms were available, people used the Hα fibril structure to study the shear of magnetic fields and its relation to flares. Sheared magnetic neutral lines can be easily visible as filaments which are dark elongated structures in the solar chromosphere, as seen in Hα images. Evidence exists to show that the field lines are along the major axis of filaments—in the same direction as neutral lines. Many flares are related to eruptions of filaments. Eruptive filaments are also known to be associated with a type of large-scale coronal activity called coronal mass ejections. Relaxation from a sheared to a potential magnetic configuration is evidenced by the disappearance of sheared filaments and formation of potential post-flare loops which connect two magnetic polarities directly (Zirin 1988). However, the Hα structure may only represent the field in the chromosphere. Although in some cases it matches the direction of the transverse field in the photosphere, it is not guaranteed that this match is still valid for complex flare-producing regions. Furthermore, Hα fibrils provide only the direction; transverse field measurements provide field strength as well. X-ray loop structures and their evolution during flares, as observed by the Japanese Satellite Yohkoh, have demonstrated magnetic reconnection and the change from a sheared to a more potential configuration. However, studies of the evolution of photospheric magnetic fields in flare-producing active regions is still very important, because they are the only way to provide a quantitative description of magnetic fields in active regions. Even though observed fields are measured along a two-dimensional surface, force-free magnetic field extrapolations based on measured photospheric magnetic fields as a boundary condition provide a powerful tool to obtain the three-dimensional field structure.

Study of evolution of vector magnetic fields during solar flares requires accurate vector field measurements, with sufficient spatial resolution (1–2 arcsec, the solar diameter is about 2000 arcsec), high cadence (5 min or less) and long duration (several hours before and after flares). Only sufficiently strong flares may have a measurable effect on photospheric magnetic fields. Giving the observing requirements, it is not surprising that such observing sequences are rare. At the time of writing, the change of photospheric magnetic fields before and after solar flares is still an unsettled issue. Wang (1997) has reported shear increases for five X-class flares, while Chen *et al* (1994) reported that there were no detectable changes associated with about 20 M-class flares. Observations at Marshall Space Flight Center demonstrated a similar inconsistency: some events show a shear increase, some a decrease and some were unchanged. Theoretical predictions are also mixed, as shear increase, decrease or undetectable changes in photospheric magnetic fields can each be explained by different models. Flares are mainly considered as coronal events; in the corona, the magnetic shear has to be released to produce solar flares, so shear there is definitely released after a flare. It may not be required for its photospheric boundary to have similar shear relaxation.

The study of magnetic fields in flare-productive regions is still a very active research area. The most outstanding problems include the following. (1) How is magnetic shear built below or above the surface? (2) Do photospheric magnetic fields change when flares occur mainly in the chromosphere and corona? (3) Can we observe direct evidence of magnetic reconnection as a result of flares? To answer these questions, a new generation of instruments and new models are needed. New observations would require high spatial and temporal resolutions as well as accuracy in the polarization measurements. They also need to be multiwavelength so information from the photosphere to corona can be combined to provide a three-dimensional view of this physical problem. Observations from new space missions such as TRACE and planned missions such as Solar-B and improved ground-based instruments such as SOLIS may shed light on this research area.

Bibliography

Chen J, Wang H, Zirin H and Ai G 1994 *Sol. Phys.* **154** 261
Hagyard M J, Smith J B, Teuber D and West E A 1984 *Sol. Phys.* **91** 115
Ishii T, Kurokawa H and Takeuchi T 1998 *Astrophys. J.* **499** 898
Lites B W and Low B C 1997 *Sol. Phys.* **174** 91
Tanaka K 1991 *Sol. Phys.* **136** 133
Wang H 1993 *The Magnetic and Velocity Fields of Solar Active Regions (IAU Colloquium 141(Beijing))* ed H Zirin, G Ai and H Wang
Wang H 1997 *Sol. Phys.* **174** 163
Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)
Zirin H and Liggett M 1987 *Sol. Phys.* **132** 173

Haimin Wang

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flare Models

Even though FLARES have been observed on the Sun for nearly 150 years, their origin remains a mystery. At the present time there is no generally accepted model which explains why they occur, but there do exist models which successfully explain certain limited aspects such as the formation of flare loops and ribbons. Before discussing particular models, we review the constraints imposed on models by the observed properties of flares.

Principal observational constraints

In constructing models, theorists are constrained by several basic properties of flares which have been firmly established by observations (see also SOLAR FLARE OBSERVATIONS). One of the most important is that they occur in a PLASMA environment dominated by magnetic fields. Magnetic energy is the only source of energy in the solar atmosphere which is capable of producing the radiative and kinetic energy output of large flares. Prior to the onset of a flare the magnetic energy density ($B^2/2\mu_0$, where μ_0 is the permeability of free space) of a 100 G (10^{-2} T) coronal field is about 40 J m^{-3}. By comparison, the thermal energy density (nkT, where k is Boltzman's constant) is about 0.01 J m^{-3} since the coronal particle density, n, and temperature, T, are about 10^{15} m^{-3} and 10^{-6} K respectively. Finally, the kinetic energy density ($m_p n v^2/2$, where m_p is the proton rest mass) in the corona is about 10^6 J m^{-3} assuming that the velocity, v, is of the order of 1 km s^{-1}—the convective velocity imparted by flows at the photospheric level. The gravitational energy density ($m_p n g h$, where g is the solar surface gravity of 247 m s^{-2}) is of the order of 0.04 J m^{-3} assuming that the average mass height, h, is about 10^8 m. Thus, the magnetic energy density is about three orders of magnitude greater than any of the other types. Since large flares typically have an energy of 10^{25} J (10^{32} erg) and a volume in the range from 10^{24} m^3 to 10^{25} m^3, an average energy density of 1–10 J m^{-3} is required. Only the magnetic energy density is in this range, so non-magnetic models of flares are ruled out.

Another important constraint is that the slow movements of SUNSPOTS and other magnetic features in the SOLAR PHOTOSPHERE are unaffected by the occurrence of flares. This is because the plasma in the photosphere is almost 10^9 times denser than the plasma in the corona where flares originate. Consequently, it is very difficult for disturbances in the tenuous corona to affect the extremely massive plasma of the photospheric layer. Field lines mapping from the corona to the photosphere are said to be 'line-tied' to the photosphere, meaning that the footpoints of coronal field lines are effectively stationary over the timescale of a flare.

Perhaps the most important observational constraint is the rapid timescale of the initial flare development. From observations it has been estimated that as much as 10^{25} J (10^{32} erg) is generated in the form of thermal and kinetic energy during the first few minutes after onset. This is an extremely short timescale for an organized process to be occurring over a region whose size is of the order of 10^5 km (about 10 times larger than the diameter of the Earth), and it implies dynamic velocities of the order of 100–1000 km s^{-1}—close to the speed at which magnetic and acoustic waves propagate in the corona. The implication of such a large energy output in so short a time is that about 10% of the available magnetic energy in a volume of 10^{15} km^3 is converted to thermal and kinetic energy within a few wave-travel times. In other words, flares require a process which converts magnetic energy with an efficiency of the order of 10%, or greater, on a timescale which is only a few times greater than the wave travel time. This combination of efficiency and speed is not easy to explain theoretically.

There are some additional constraints that are especially important for the class of flares known as large, two-ribbon flares or long duration events (see SOLAR FLARE CLASSIFICATION). After the initial rapid phase, these flares continue to produce a sustained, low level emission for over 10 h. They stand in stark contrast to the more numerous, short-lived events (<30 min), known as compact flares. Long duration events give rise to two bands of chromospheric emissions called flare ribbons which slowly propagate away from one another. (Compact flares may also produce ribbons which are usually called 'knots' because they are small and short-lived.) The flare ribbons lie at the feet of dense, x-ray emitting loops often referred to as 'post'-flare loops with the word 'post' in quotes to indicate that nowadays they are considered an integral part of the flare process, rather than simply an aftereffect. High resolution images and Doppler measurements have established that the motion of both ribbons and loops are only apparent and are caused by the progressive lighting up and dimming of the regions through which they propagate. Models for large, two-ribbon flares must be able to account for these apparent motions as well as the long lifetime of the events themselves.

Difficulties of constructing flare models

It is sometimes thought that the only requirement for a successful theoretical model is that its predictions match observations. What is forgotten in this simplistic view is that any model can be made to match a particular set of observations to any degree of accuracy desired simply by including enough adjustable, free parameters. When the number of free parameters reaches the point where there is one for each observational measurement, the model becomes simply a reorganization of the observational data. Thus a successful model should not only match observations, but also have as few free parameters as possible.

Virtually all quantitative flare models are formulated as solutions of initial boundary-value problems. These problems are comprised of a system of differential equations, a set of boundary conditions, and an initial state. Since the underlying system of differential equations for all models are those of electromagnetism (i.e. Maxwell's

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

equations) and Newtonian mechanics (for continuous media), models differ from one another in their choice of boundary and initial conditions and in the approximations which are made to solve the equations.

Most flare models use the MAGNETOHYDRODYNAMIC (MHD) approximation. Although these equations do not provide a description of kinetic processes, they nevertheless provide a complete and self-consistent description of the fluid behavior of the coronal plasma, and what is equally important, they can be solved in many circumstances where more complete descriptions cannot. Used alone, the MHD equations do not provide any information on the production of radio waves or energetic particles, both of which constitute a major database for the flare phenomenon. Thus, several flare models attempt to augment the MHD system by adding additional equations. For example, some energetic particle models use an MHD solution to provide electric and magnetic fields as inputs into the exact equation which describes the motion of the particles in electromagnetic fields (i.e. the Lorentz force equation).

Boundary conditions for flare models are fairly well constrained by observations. Observations have already established that ACTIVE REGIONS, with complex magnetic fields, are most likely to flare. Particularly favored are active regions containing multiple sunspots with opposite magnetic polarities in close proximity to one another and undergoing motions which act to shear the magnetic field. These observations imply that currents are built up in the corona prior to a flare by the transport of magnetic flux from the convection zone into the corona and by the stresses exerted on the existing coronal field by the shear flows in the photosphere.

The greatest uncertainty in the development of quantitative flare models is caused, not by the boundary conditions, but by the initial conditions because these are very poorly constrained by observations. Since the flare is a magnetic phenomenon, the magnetic vector field is the most important variable which needs to be specified in the initial state. Unfortunately, the magnetic field is the one quantity which cannot be directly observed in the corona at the present time. The problem of determining the coronal magnetic field is further exacerbated by the fact that it is only the component of the magnetic field generated by currents flowing in the corona that contributes to the energy release. The component of the field due to photospheric sources remains invariant during a flare and so cannot contribute to the energy release. The contribution to the magnetic field from the coronal currents is estimated to be anywhere between 10–50% of the total field. Thus, even if the coronal magnetic field could be directly observed, it would be necessary to measure its strength and its orientation to an accuracy of about 10% everywhere in the flare region before such measurements could be used to formulate initial conditions. The best that can be done at the present time is to infer the coronal magnetic structure from variations in the density and temperature of the coronal plasma. This is done by assuming that such variations occur only across field lines, so that contours of constant density or temperature follow field lines. However, there are many instances where this assumption is known to be violated (e.g. density and temperature structures in flare loops).

Even if the system of equations, the boundary conditions, and the initial conditions could be accurately established by observations, there would still be the enormous mathematical difficulty of actually solving the governing equations. Fluid equations, with or without the addition of Maxwell's equations, are notoriously difficult to solve because of their high degree of nonlinearity. For example, there is still no generally accepted analytical description for ordinary fluid turbulence, let alone MHD turbulence. Some progress has been made using numerical methods, but these methods cannot handle the tremendous range of scales involved in flares. To provide a complete description of all phenomena, a flare model should include scales as small as 10^{-2} m (e.g. an electron gyro radius) and as large as 10^{8} m (or even greater) if the flare produces a SOLAR CORONAL MASS EJECTION.

Most modelers, whether they use analytical or numerical techniques, are forced to simplify the geometry of their initial configuration to the point where it is far removed from the complexity of the actual solar field. However, such models are still valuable because they provide a quantitative way to test whether a proposed flare mechanism is really consistent with the laws of physics. It will be many years before it is technically possible to solve the complete set of plasma equations using initial and boundary conditions which actually match reality.

Storage models

Most flare models are based on mechanisms which release magnetic energy stored in the corona. The energy is stored in the form of coronal currents, and it is released when these currents become unstable due to the slow evolution of the normal magnetic field component at the photosphere. The continual emergence of new flux from the convection zone and the shuffling of the footpoints of closed coronal field lines causes stresses to build up in the coronal field. Eventually, the stress exceeds a threshold beyond which a stable equilibrium cannot be maintained, and the field undergoes a rapid transition to a configuration with a lower magnetic energy.

Because the magnetic pressure greatly exceeds the gas pressure in the corona, the currents associated with the magnetic energy stored there must either be force-free (i.e. field aligned) or confined to thin layers (i.e. current sheets). In other words, the energy cannot be stored as a distributed, non-force-free current since such a current cannot be in equilibrium in the absence of any significant gas pressure. Models for magnetic energy storage in the corona can be divided into those based on force-free currents and those based on current sheets. A well-known example of the latter is the emerging flux model illustrated in figure 1. As new flux emerges from the photosphere it

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

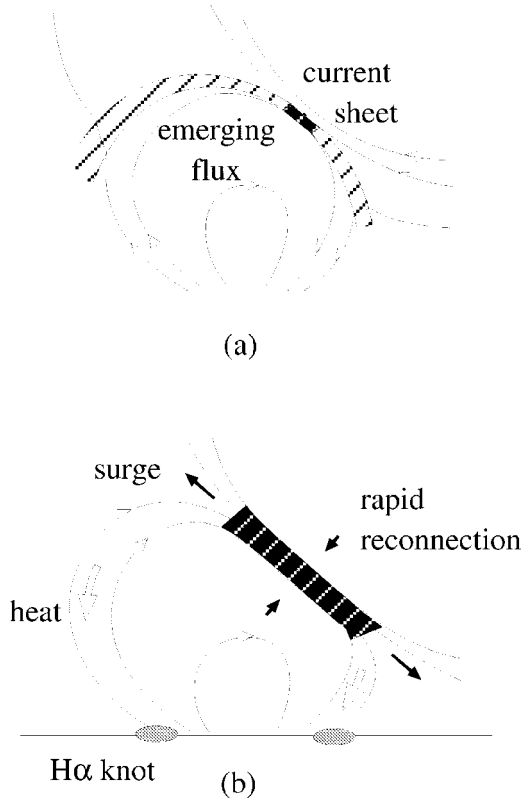


Figure 1. The emerging flux model of impulsive flares. In (a) a current sheet forms as a newly emerging loop collides with a pre-existing field. In (b) the onset of a resistive microinstability produces rapid reconnection of the field lines.

forms a current sheet as it presses against field structures that are already present. As the current in the sheet grows, it may then reach a critical threshold for the onset of a micro-instability. Once such an instability occurs, it is possible that the local resistivity of the plasma will increase so much that the current sheet is rapidly dissipated by means of magnetic reconnection. This process changes the connectivity of the field lines and releases most of the stored magnetic energy, but in order to be effective in flares it must occur very rapidly. How such rapid reconnection occurs is still a matter of debate.

In order for the emerging flux model to work for large flares, the pre-existing current sheet must be stable for a period of the order of a day or more. However, numerical simulations suggest that coronal current sheets reconnect so rapidly that a stable sheet cannot be sustained for such a long period of time. The success of x-ray bright-point models, which assume that current sheets are dissipated as soon as they try to form, reinforces this view (see SOLAR CORONA: X-RAY BRIGHT POINTS). Also, since no flux is ejected in the emerging flux model, it cannot explain the coronal mass ejections that are associated with very large flares. Thus, the emerging flux model is limited to small flares.

One type of force-free model that has received much attention is based on trying to create an eruption by shearing the footpoints of an arcade of loops. In two-dimensional force-free configurations with translation symmetry (i.e. sometimes referred to as $2\frac{1}{2}$ D configurations) shearing causes the arcade to smoothly expand outwards towards a fully opened state without ever producing an eruption. It is not yet known for certain whether an arcade will maintain a stable equilibrium as its footpoints are sheared in less symmetric systems. Even if no loss of mechanical equilibrium or stability occurs, it is still possible to create an eruption of the arcade by invoking a resistive instability as is done in the emerging flux model. Shearing an arcade leads to the formation of a current sheet which can undergo reconnection, but again there must exist a critical threshold for the sudden onset of a micro-instability.

Another way for an arcade to form is for part of a flux rope to emerge from the convection zone, through the photosphere and into the corona. If the axis of the flux rope remains below the surface, then a simple arcade exists. However, if the axis of the flux rope emerges into the corona, the arcade will contain a flux rope. The presence of the flux rope may make it possible for the arcade to lose equilibrium and erupt even when no current sheet or x line is present.

Figure 2 shows a quantitative flux-rope model which demonstrates the underlying principle of this process. The model has the drawback that it is two dimensional and, therefore, does not include the effect of anchoring the end of the flux ropes in the photosphere. The evolution of the configuration takes place in two phases. During the storage phase, the evolution is assumed to be slow relative to all other timescales, so that at any given moment the configuration is very well approximated by a static equilibrium. The second phase occurs when the sources reach a critical distance, which is of the order of the initial height of the flux rope. At this point the equilibrium vanishes, and the flux rope jumps to a higher altitude. This process releases a small fraction ($\approx$5%) of the magnetic energy stored in the system, and creates a large-scale current sheet as shown. The extent and lifetime of this current sheet depends on how fast reconnection occurs. Without reconnection, there is no further release of energy, and the jump of the flux rope is finite, but if reconnection occurs, the flux rope can continue to move upwards indefinitely.

At least two important issues still remain to be investigated before the viability of flux-rope models can be satisfactorily determined. First of all, there is the unknown effect of anchoring the ends of the flux rope in the photosphere. Although anchoring the ends might make an eruption more difficult, it is not clear that it would prevent a loss of equilibrium from occurring. The motion of the middle of the flux rope relative to the ends would constitute a type of external kink, and it is known that kink modes are not stabilized by anchoring the ends of a flux rope if the rope is sufficiently long. Second, there is the question of whether the two-dimensional flux-rope equilibria will remain stable to the kink instability as the system approaches the equilibrium point.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

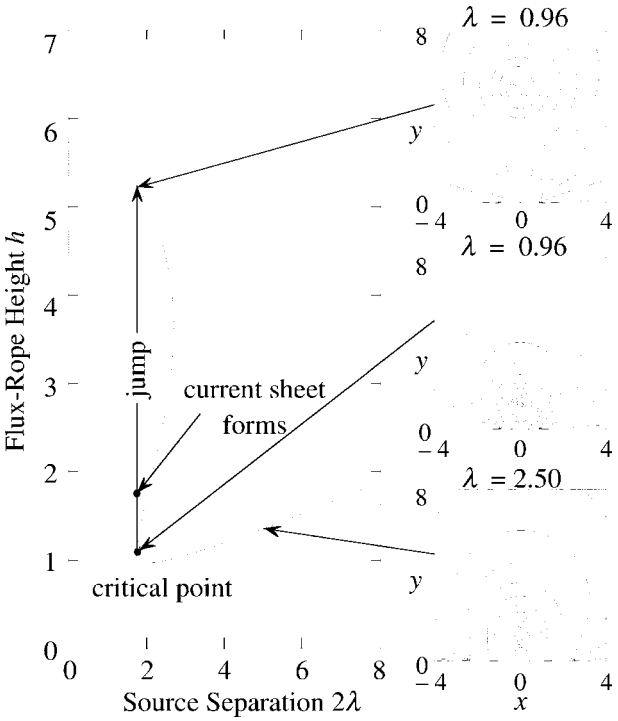


Figure 2. An arcade model containing a flux rope. Eruption occurs when the source regions in the photosphere are pushed together. A current sheet forms only after the eruption begins.

Models for flare loops and ribbons

Flare loops range from temperatures of 10^4 K to 3×10^7 K with the cooler loops nested below the hotter ones. Remarkably, the loops at the low temperature are two orders of magnitude cooler than the ambient coronal plasma even though they lie underneath a region which is an order of magnitude hotter than the ambient corona. It is generally accepted that the cool loops are formed from the hot loops by a thermal instability. Such an instability is possible in the solar atmosphere because of the non-black-body behavior of the corona. The coronal radiative loss function decreases with temperature in the range from 10^5 K to 10^7 K, so in this range a temperature perturbation can lead to runaway cooling.

The footpoints of the flare loops map to Hα emission features in the CHROMOSPHERE that are known as flare ribbons. (Hα refers to the red light emitted in the Balmer transition of the hydrogen atom.) Very roughly, the outermost edge of the hot x-ray loops maps to the outer edge of the ribbons while the innermost edge of the cool Hα loops maps to the inner edge of the ribbons. During the course of the flare the separation between the ribbons increases, while the loops grow larger.

At the present time, MAGNETIC RECONNECTION is thought to be the only mechanism which can account for the apparent motion of the flare loops and ribbons. Although some alternate mechanisms have been suggested in the past, these mechanisms have become less tenable as observations have improved. For example, simple expansion of the loops due to outward motion of the plasma from the flare site has been ruled out by Doppler-shift measurements of Hα loops. These measurements show that the plasma in the Hα loops flows downward at speeds of 100–500 km s^{-1} during the time that the loops appear to be expanding upwards. Thus the loop motions are not due to mass motions of the plasma, but rather to the continual propagation of an energy source onto new field lines.

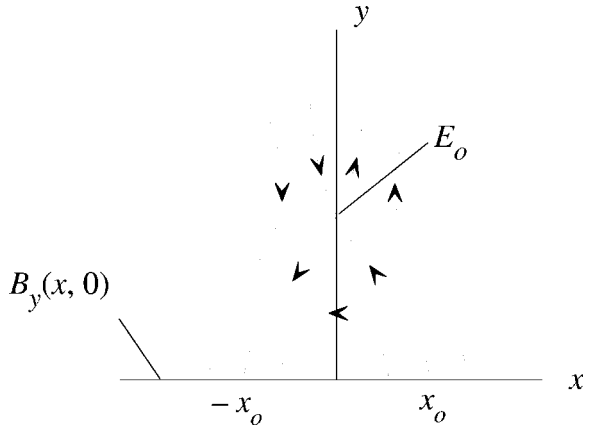


Figure 3. Relation between the rate of reconnection, as measured by the electric field at an x line, and the motion of the flare ribbons.

In a two-dimensional configuration there is a simple relation between the rate of reconnection and the motion of the photospheric footpoints of the field lines mapping to the x line. Consider the configuration shown in figure 3. Using Faraday's equation, the electric field E_0, at the x line can be written as

$$E_0 = V_0 B_y(x_0, 0) \tag{1}$$

where x_0 is the location of the outer edge of the ribbon, $B_y(x_0, 0)$ is the vertical magnetic field located at x_0, and V_0 is the apparent velocity of the ribbon edge. This relation is model independent except for the assumptions that an x line exists, that the configuration is two dimensional, and that the field lines are tied to a stationary photosphere.

If one assumes that the outer edges of the flare ribbons lie at the footpoints of the field lines mapping to the x line, then it is possible to use equation (1) to determine the electric field in the corona as a function of time, as shown in figure 4 for a flare observed by the SKYLAB space station in 29 July 1973. A few minutes after flare onset the electric field for this event reaches a peak value of about 2 V cm^{-1}, but then declines until it is an order of magnitude weaker a few hours later.

Although one can argue that the loop and ribbon motions are strong evidence of reconnection , they are, nevertheless, indirect evidence. In principle, x-ray observations of sufficient spatial resolution and sensitivity can provide direct evidence. However, the ability to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

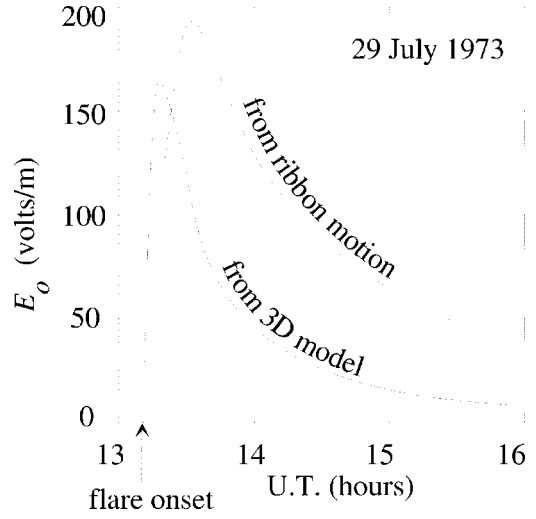


Figure 4. Coronal electric field inferred for the flare on 29 July 1973. The upper curve is obtained using formula (1) in the text, while the lower curve is based on a more elaborate three-dimensional potential extrapolation of the surface field.

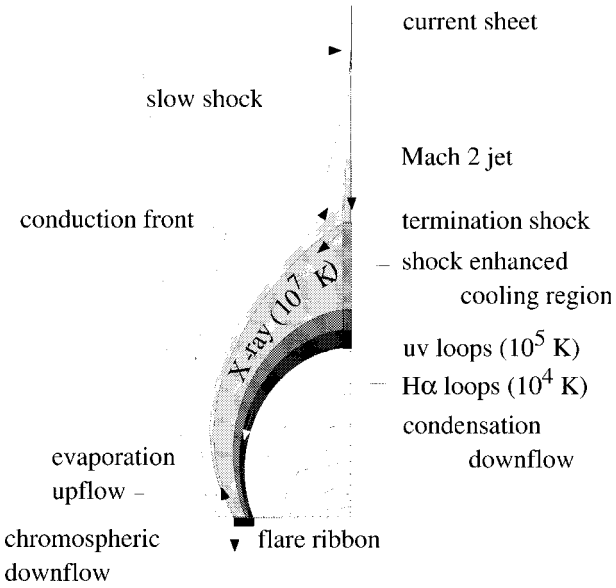


Figure 5. Model diagram for a flare loop system formed by reconnection in the corona (only the left half is shown). In this particular model, the outflow from the reconnection region creates upward and downward supermagnetosonic jets, but only the downward jet is shown in the figure. Solid curves indicate boundaries between various plasma regions, while dashed ones indicate magnetic field lines.

determine whether or not there is a reconnection site in the corona depends very much on theoretical expectations of what such a site should look like. High resolution images obtained from x-ray telescopes on the Japanese satellite YOHKOH show several features that are suggestive of a reconnection site in the corona. These features include: (i) a hard x-ray source located above the soft x-ray loops; (ii) cusp structures suggestive of either an x type or a y type neutral line; (iii) bright features at the top of the soft x-ray loops; and (iv) high temperature plasma along the field lines mapping to the tip of the cusp.

Some of the cusp-shaped loops observed by Yohkoh have a linear, trunk-like feature which extends from the top of the cusp all the way down to the innermost arch of the flare loop system. The hottest regions in the loop system do not lie in the trunk feature but along the edges of the cusp formed by the outermost loop. The trunk feature is both cooler and denser than the plasma surrounding it.

Figure 5 shows a theoretical model which explains the loop structures and evolution in terms of reconnection and evaporation. According to this model, flare loops are created by chromospheric evaporation on field lines mapping to slow-mode shocks in the vicinity of the neutral line (sometimes referred to as Petschek shocks). These slow shocks are somewhat different than the standard MHD slow-mode shocks because the conduction of heat along the field lines causes them to dissociate into isothermal subshocks and conduction fronts. The shocks annihilate the magnetic field in the plasma flowing through them, and the thermal energy which is thus liberated is conducted along the field to the chromosphere. This in turn drives an upward flow of dense, heated plasma back towards the shocks, and compresses the lower regions of the chromosphere downward.

In order for strong slow shocks to form on the field lines below the X-line, the outflow from it must be supermagnetosonic with respect to the fast-mode wave speed. If the magnetic fields are sufficiently strong, the outflow from the X-line produces two supermagnetosonic jets—one directed upward and the other downward. Because of the obstacle presented by the closed field lines attached to the photosphere, the lower jet terminates at a fast-mode shock after traveling a short distance. Below this termination shock the flow is deflected along the field, and only weak, field aligned slow-mode shocks are present. Consequently, the magnetic energy released below the termination shock is relatively small.

When the magnetic field is relatively strong, the fast-mode Mach number of the jets is about two, but as the field decreases, the Mach number decreases and the jets eventually become submagnetosonic. The transition occurs when the ratio of gas to magnetic pressure (plasma β) exceeds $(3-\gamma)/\gamma$ in the x-ray loops where γ is the ratio of specific heats. For $\gamma = \frac{5}{3}$ this gives $\beta = \frac{4}{5}$, which for typical loop parameters corresponds to a coronal magnetic field strength of a few gauss.

When the fast-mode Mach number becomes less than one, the lower jet disappears and almost all of the outflow from the neutral line is directed upwards. Only a relatively slow flow, aligned along the separatrices, remains below the neutral line. An upward reconnection jet is still present, but it is located at the upper end of a current sheet which may extend far out into the corona. Slow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

shocks are still present below the neutral line at the lower tip of the current sheet, but they are aligned along the separatrices and are much weaker than the Petschek-type shocks. Because of the weaker shocks, the evaporation process is greatly reduced and the plasma density in the loops becomes too low to trigger a thermal condensation. For a while, condensations remain on the loops at lower altitude, but, eventually, as the older loops continue to cool and disappear, the trunk feature vanishes.

The ribbon width, W, is approximately determined by

$$W = V_0 t_{\rm cr} \tag{2}$$

where V_0 is the ribbon velocity and $t_{\rm cr}$ is the cooling time due to conduction and radiation acting together. For the 29 July 1973 flare observed by Skylab, the observed densities and temperatures imply cooling times ranging from $t_{\rm cr} = 15$ min ($n = 10^{17}$ m^{-3}) shortly after flare onset to $t_{\rm cr} = 240$ min ($n = 10^{16}$ m^{-3}) 10 h after onset. When used in equation (2), these values predict ribbon widths which agree well with the observed widths.

The thickness, $W_{\rm j}$, of the region of the ribbon which maps to the reconnection jet, is roughly

$$W_{\rm j} = (V_{\rm R}/V_{\rm j}) L_{\rm j}$$

where $V_{\rm j}$ is the speed of the plasma flow in the jet. Assuming $V_{\rm j} = 10^3$ km s^{-1}, $V_{\rm R} = 10$ km s^{-1}, and $L_{\rm j} = 10^5$ km, gives $W_{\rm j} = 10^3$ km which is a little more than an arcsecond in angular size. This value of $W_{\rm j}$ is consistent with the dimensions of the thin ($\approx 1''$) region of redshifted chromospheric plasma observed at the outer edge of flare ribbons. The redshift is thought to be produced by the bombardment of the chromosphere by energetic particles produced in the region of the jet.

One possible explanation for the lower part of the trunk feature is that it is created by the onset of the thermal instability which produces the cool Hα loops lower down. A very similar feature has been found in numerical simulations which include radiation but not conduction. In these simulations the termination shock triggers a thermal instability in the downstream region, and as the reconnection site travels upward, the condensing region forms a trunk-like feature on the static loops.

Just as for reconnection in the geomagnetic tail (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL), reconnection in flare loops leads to a relaxation of stretched field lines. In the geomagnetic tail, this relaxation is referred to as dipolarization, but in flare loops it is referred to as shrinkage. What causes the shrinkage is the upward retreat of the neutral line. When a field line is connected to the neutral line it is cusp-shaped, but as the neutral line moves further up, the field line becomes rounded. The expected shrinkage from this effect is about 20–30% for typical flare loops, and this matches the observed values fairly closely.

Bibliography

Priest E R 1982 *Solar Magnetohydrodynamics* (Dodrecht: Reidel)

Somov B V 1992 *Physical Processes in Solar Flares* (Dodrecht: Kluwer)

Svestka Z 1976 *Solar Flares* (Dodrecht: Reidel)

Tandberg-Hanssen E and Emslie A G 1988 *The Physics of Solar Flares* (Cambridge: Cambridge University Press)

Terry G Forbes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flare Observations

Flares are caused by the release of magnetic energy up to some 10^{27} J in the solar atmosphere within a few minutes (figure 1). The energy input causes a myriad of phenomena including the flaring up of the region and its surroundings at all wavelengths from radio to gamma rays, the acceleration of elementary particles to relativistic energies and the launch of a shock wave. Flares may be responsible for the heating of the solar corona and for large-scale restructuring of the corona. The largest ones influence interplanetary space and substantially affect the Earth's lower ionosphere. Solar flares have been a major, unexplained problem of astrophysics for more than a hundred years.

History of solar flare observations

The history of flare observations is a fascinating unfolding of a persisting enigma and is full of surprises. The first solar flare was recorded independently by R C CARRINGTON and R Hodgson on 1 September 1859. They recognized intensely bright patches close to a group of dark sunspots. In retrospect the patches can be identified as a white light flare of the photospheric continuum emission, consistent with the great, terrestrial magnetic storm that occurred a few hours later. With the visual technique at that time, flares were extremely rare phenomena. That changed a few years later with the newly developed spectroscope. C A Young, C Secchi and J N LOCKYER noted much more frequent brightenings near sunspots in certain spectral lines, particularly in the hydrogen Hα line originating at a higher altitude, in the chromosphere. G E HALE's spectrohelioscope made this method sensitive enough in the early 1930s to record several such Hα flares per active day.

The gradual progress in observations and understanding changed pace in 1944, when J S HEY serendipitously discovered solar radio emissions and associated some of them with flares. At about the same time, S E Forbush noticed ground-level cosmic ray enhancements associated with major solar flares. These discoveries could only mean that the flare phenomenon is not confined to thermal plasma, but includes high-energy particles and involves the corona. Rocket flights in 1959 by T A Chubb, H FRIEDMAN and R W Krepin detected flare x-rays up to 125 keV and again opened a new dimension. From x-ray emissions, the fundamental conclusion was drawn later that accelerated electrons were not a peripheral phenomenon, but recipients of a considerable fraction of the primary energy released in flares. At the same time soft x-rays from flares were discovered. They usually peak when the hard x-rays have already disappeared and most of the flare energy resides in the more than 10 million degree soft x-ray emitting plasma. With the discovery of gamma-ray line emission in 1972 by E L Chupp and collaborators, the >1 MeV ion component of flare particle acceleration became accessible. There is now observational evidence for the suggestion that ions get a similar share of the primary energy release as electrons.

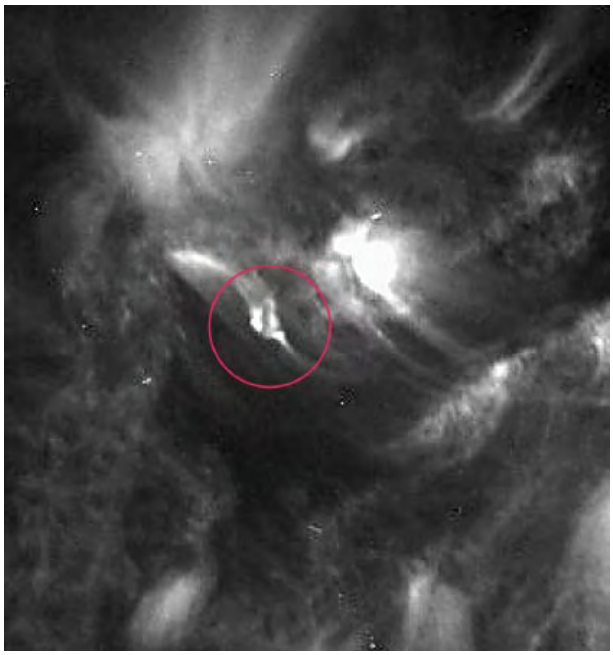


Figure 1. Encircled is the ultraviolet emission of a small flare observed by TRACE. It has resulted from the collision of two expanding magnetic loops. The two loops brightened, in particular the point of contact, where the magnetic field was observed to reconnect. The size of the image is 175 × 175 Mm.

Recent space missions, including SKYLAB, SOLAR MAXIMUM MISSION, YOHKOH and SOHO, as well as new ground-based observatories, both for optical and radio waves, have been essential in the detection of flares outside of active regions. First, CORONAL BRIGHT POINTS in thermal soft x-ray emission were noted, then smaller events in the network of the quiet Sun granulation and finally even in the interior of supergranular cells were discovered. These microflares are about a million times smaller than the one Carrington and Hodgson discovered, and they occur at a rate of thousands per second on the Sun at any time.

A general flare scenario has emerged in which the flare energy is released in the corona and accelerates or heats particles to an energy of 20 keV and more. The particles precipitate into the dense chromosphere, where they collide and deposit their energy. The chromospheric material is heated to the temperatures seen in soft x-rays up to tens of millions of degrees. It expands and adds new high-temperature plasma to the atmosphere above the transition region. The upward motion usually fills up coronal loops, but the motion may continue in an expansion of these loops.

In some cases, the above flare scenario may be only a minor part of a much larger destabilization of the corona, when the magnetic confinement of a considerable part of the corona is removed. It expands and is expelled by magnetic forces in a coronal mass ejection (CME).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

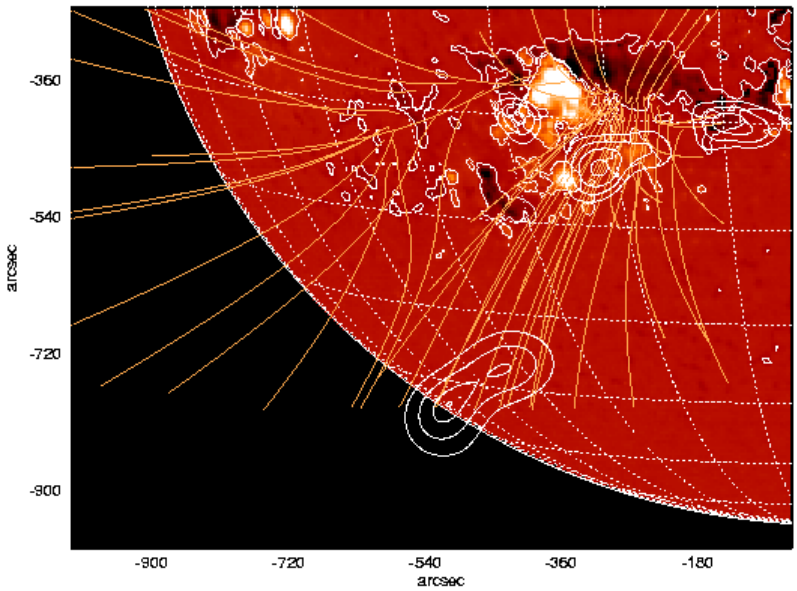


Figure 2. The source region of narrowband spikes, identified with the Phoenix radio spectrometer, has been observed with the Very Large Array at 0.33 GHz. The intensity is presented with contours and is apparently at high altitude, in projection near the solar limb. The photospheric magnetic field is shown in gray scale together with the extrapolated potential field lines. Enhanced thermal radio emission at 1.44 GHz (contours) is also present near the active regions.

The shock front associated with this motion is also a site of particle acceleration, particularly of high-energy solar cosmic rays observed on Earth at ground level.

Types and phases of flares

Flares vary greatly in their spatial shapes and in the relative intensities at different wavelengths. Considering this observed variety, it is not surprising that many attempts have been made to classify flares.

(a) The archetypical flare is impulsive in time and 'compact' in space. It occurs in loops of an active region, which confine most of the energy, i.e. the accelerated particles and the heated plasma. Thus it is also referred to as a 'contained' flare.

(b) A very spectacular class combines large flares occurring at higher altitude. They are often preceded by the disappearance of a prominence and sometimes related to a coronal mass ejection. In view of their longer duration they are called gradual or 'long-duration events', and in Hα they are often seen as 'two-ribbon flares' and 'eruptive'.

(c) High coronal flares occur at even higher altitude. They are responsible for metric type III radio bursts (cf SOLAR SPECTROSCOPY: COHERENT PLASMA EMISSION) and interplanetary electron events, both caused by electron beams propagating into interplanetary space. The energy release may occur at altitudes of half a solar radius or more above the photosphere (figure 2).

(d) In quiet regions of the Sun, flares observed in the emissions of the coronal plasma are very small in energy (some 10^{16} to 10^{20} J) and reach heights of only a few thousand kilometers. They have been termed network flares because of their preferred location above the magnetic network of the chromosphere (figure 3) and may be related to 'explosive events' and 'blinkers' observed previously in the transition region. The largest network flares have been known since the 1970s as coronal bright points (see SOLAR CORONA: X-RAY BRIGHT POINTS).

There is little doubt that the primary cause of the large variety originates in the wide range of possible plasma parameters and unstable magnetic configurations in the energy release region of flares. In active regions of the corona, the magnetic field is generally observed to have a larger scale height than in quiet regions, although the actual energy release may take place on equally small scales and by similar processes. Differences in altitude reflect the change of the global field properties from predominantly closed at low altitudes to partially open at high altitudes.

Flare emissions at different wavelengths do not peak at the same time. The general order of these emissions is shown in figure 4. Flares often start with a brightening in soft x-rays of a few keV due to chromospheric material freshly heated to coronal temperatures. Metric radio emission caused by escaping electron beams indicates that this pre-flare phase is not entirely thermal. The bulk of the flare energy is released in the impulsive phase (or non-thermal phase; see SOLAR FLARES: IMPULSIVE PHASE). Its start is characterized by the onset of hard x-ray (HXR) and centimetric gyrosynchrotron radiation of the energetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

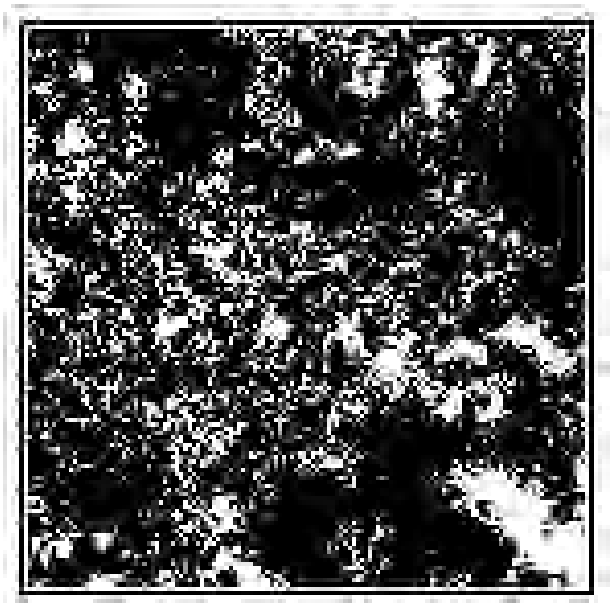


Figure 3. Microflare activity in a field of 280 × 280 solar radii of the quiet corona observed by EIT on SoHO. All pixels that have a $>3\sigma$ enhancement of the coronal emission measure (density) in at least one time step during a 42 min observation are indicated white. The large-scale structures in white follow the magnetic network in the chromosphere below.

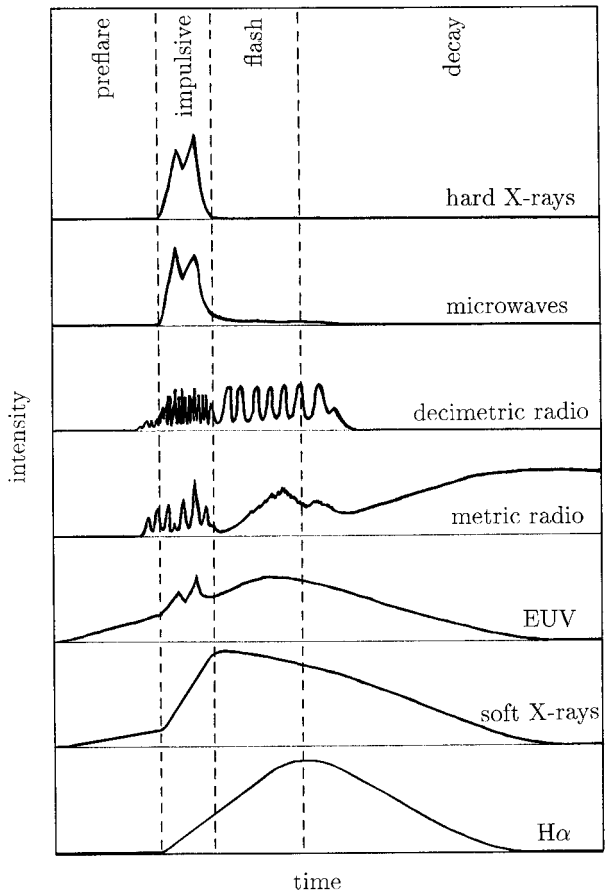


Figure 4. A schematic evolution of flare intensities at different wavelengths. The phases indicated at the top vary greatly in duration. In large events, the preflare phase lasts 10 min, the impulsive phase 1 min, the flash phase 5 min and the decay phase 1 h.

electrons (see SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS and SOLAR FLARES: RADIO BURSTS) as well as gamma-ray line emission caused by energetic ions (see SOLAR FLARES: GAMMA RAYS). Throughout the impulsive phase, the soft x-ray emission increases in size and total flux. Also the lower chromosphere is increasingly affected as evidenced by growing Hα emission. In the flash phase (or thermal or late phase) the non-thermal particles and their emissions have mostly disappeared and the impact of the energy release becomes fully visible. The temperature of the hot plasma decreases. Sometimes, however, gigantic soft x-ray emitting loops cool only slowly and indicate that energy is still released in this phase. Then the flare develops gradually into the decay phase, during which the thermal flare emissions disappear. In the dense flare plasma evaporated into the corona, hydrogen recombines and emits spectacular post-flare loops in Hα. They often show down-moving material. It is the signature of a cooling loop, the density scale height of which is decreasing. Thus the plasma evaporated by the flare is 'raining' back to the chromosphere.

General properties of flares

The typical timing of the observed flare emissions suggests that the major energy release in the impulsive phase transfers the free energy into coronal particles, both electrons and ions. They reach energies far beyond the mean thermal values of the corona. A considerable fraction of this energy has been found in 10–30 keV electrons. A characteristic of this initial particle acceleration is that it produces a power-law in energy distribution of the form $E^{-\delta}$, where δ is between 2 and 5. Electrons reach energies up to 300 MeV, ions up to 10 GeV.

The energetic particles, generally called 'non-thermal', have a much reduced collisional cross section. They spiral along magnetic field lines due to the Lorentz force. A few escape into interplanetary space, but most follow loops and eventually penetrate the denser chromosphere where they quickly lose their energy by collisions.

An impulsive flare of say 10^{24} J is typically spread over an area of several 10^{14} m^2 as seen by its Hα imprint on the chromosphere. Limb flares indicate that such a flare may have a vertical extension of several 10^7 m, and thus a volume of about 10^{22} m^3.

The energization of particles far beyond thermal energies is no surprise since the corona is a plasma with a low ratio of thermal energy density to magnetic energy density. Any relevant release of magnetic energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

must therefore produce particles beyond thermal energies. If all the magnetic energy were released in the above example, an average field of 160 G must be annihilated. If this energy is taken up by plasma with an electron density of 10^{16} m^{-3}, each electron and ion receives 30 keV on average. The process that transfers electromagnetic energy into kinetic particle energy is not clear (see SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS). Moreover, the number of electrons required to produce the observed HXR bremsstrahlung is astonishing. The actual value is model dependent and may be of the order of 10^{36} s^{-1} or more.

In an extreme scenario, the flare energy builds up in the corona and all of the flare energy is released there into accelerating non-thermal particles. The electrons are observable in their radio and hard x-ray emissions, the ions in gamma-ray lines. These particles precipitate after a few seconds and dump their energy into small spots of the chromosphere, which heat up to temperatures of several tens of millions of degrees. The larger the flare, the higher is the observed temperature. The hot material, i.e. the flare plasma, explodes due to overpressure and expands along field lines into the corona with a velocity of several hundred km s^{-1}. This phenomenon is observed in transition region lines and is termed 'evaporation'.

The above scenario does not fit all observations. It does not explain why flares usually start with thermal soft x-ray emission, when very few non-thermal particles are observed in the preflare phase. Similarly, the extended heating in the flash phase does not seem to be associated with particle acceleration. Furthermore, the ratio of gyrosynchrotron to soft x-ray emissions decreases to smaller flares, indicating relatively fewer relativistic electrons or smaller lifetimes. For these reasons, occasional or partial direct thermalization without acceleration cannot be completely excluded.

In spite of the large variety, homologous flares occurring at the same site of an active region have also been reported. Homologous flares have similar energy and shape. They occur within hours and suggest a continuous energy build-up that is occasionally released.

Flare sites

Magnetic fields can best be measured in the photosphere, and thus flares can easily be located relative to photospheric magnetic features. As a rule, flares occur above the places in the photosphere with largest $\nabla \times \boldsymbol{B}$, thus where the electric current has a maximum (figure 5). Most productive flare sites are regions in sunspots or groups of sunspots where new and oppositely directed magnetic field emerges from below.

Large, 'gradual' flares often occur above 'neutral' lines in the photosphere, separating regions with opposite magnetic polarity. Neutral lines are bridged by arcades of loops. The Hα picture of flares in such structures shows two bright ribbons formed by the footpoints on each side of the neutral line. Flares usually appear above the part of the neutral line which has experienced most shear by different surface motions on both sides. The first bright Hα point appears near the neutral line. In the course of the impulsive phase, the maximum brightness drifts away to both sides and remains stable during the flash phase.

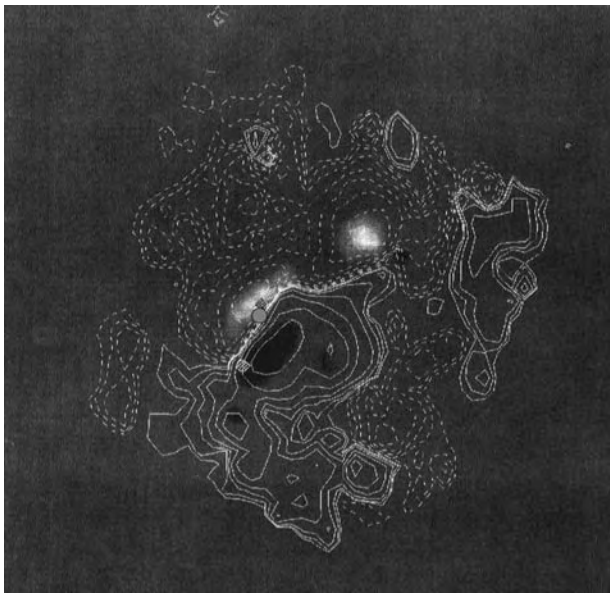


Figure 5. Map of the vertical current density (gray scale), vertical magnetic field (contours), and sites of Hα signatures of non-thermal electron precipitation (gray circle) and high coronal pressure (cross-hatched triangles) (courtesy of R C Canfield).

In quiet regions, the most powerful microflares occur at the boundary of supergranular cells. The cellular structures are formed by large-scale convection below, sweeping frozen-in magnetic field lines to the down-draft region at the surface. There a complex, small-scaled entangled field builds up, the magnetic network. Time scales are a few tens of minutes, over which magnetic elements can be observed to appear and disappear. Enhanced coronal emissions and microflares in soft x-rays and high-temperature iron lines are predominantly associated with bipolar magnetic elements, consisting of a patch of up-going and adjacent down-going magnetic field.

Photospheric observations clearly support the scenario that flare energy originates from electric currents in the corona or, which is equivalent, from free magnetic energy in excess of the potential value defined by the photospheric footpoints. Such currents may arise from the emergence of new magnetic loops penetrating into an old coronal magnetic field configuration. Another possible site is the separatrix between oppositely directed field lines forming a current sheet, e.g. above a cusp-shaped coronal structure. Currents can also build up by shearing or twisting footpoints in the photosphere.

The unifying theme over the past few decades for releasing the energy stored in electric currents has been reconnection. Enormous current densities are necessary to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

make it happen fast enough. Such high current densities exist only on the scale of a few tens of meters. Thus it is not surprising that there is little direct evidence for the process. In fact, the actual reconnection region cannot have the above volume derived from the energy and particle requirement. Thus it is often assumed that the bulk energy release does not occur in the reconnection region, but in the outflows from it, where shocks and high-frequency turbulence release most of the energy and accelerate particles.

It is not clear whether actual signatures of the outflow from flare acceleration sites have been observed. Potential candidates are hard x-ray sources observed above the soft x-ray loops and occasional narrowband spikes in decimeter radio waves. The narrowband spikes emitted by tens of thousands of sources smaller than about 100 km having electron densities between 10^{15} and 10^{17} m^{-3} are still speculative diagnostics.

Large flares do not consist of a single process continuously releasing energy during the impulsive phase. A fragmentation into 'elementary flares' has been noted first in hard x-rays involving time scales of some ten seconds. The total flux of hard x-ray flare emission often consists of a sequence of such peaks. They are sometimes resolved into shorter structures at radio waves of short centimeter and millimeter wavelength. Also, they loosely correlate with electron beams propagating upward in the corona and visible by their radio emission in meter waves. Both radio emissions strongly suggest an even higher degree of fragmentation. Narrowband radio spikes exceed even this number by another order of magnitude or two. Their duration is only a few tens of microseconds.

The observational evidence indicates impulsive acceleration in a large number of small-scale regions. The flare process thus involves scales from a solar radius for the energy build-up and large-scale instability to about the ion gyroradius of few centimeters for acceleration and energy release. It involves both magnetohydrodynamic and kinetic plasma phenomena.

Effects and consequences of solar flares

The sudden release of energy in the tenuous corona disturbs the fragile equilibrium with the underlaying chromosphere. The transition region that couples the two layers adjusts to the new situation, and these transitory changes are observable in ultraviolet lines and jet motions. In the late phase, the flare energy input into the corona becomes well observable in all thermal coronal emissions.

The role of this impulsive, localized injection of hot plasma for the heating of the corona is not yet fully understood. It is clear, however, that the flares observed in active regions contribute only about 20% of the energy loss by thermal radiations from ultraviolet to x-rays. The flare frequency distribution with energy is too flat to make flares below the sensitivity threshold obvious candidates for the rest. Microflares in quiet regions, on the other hand, seem to be relatively more important contributors and have a steep distribution suggesting a dominant role for the smaller events. The actual value of impulsive energy input is model dependent. At present, microflares are an attractive possibility to satisfy the requirement for coronal heating in quiet regions and coronal holes (see CORONAL HEATING MECHANISMS).

As a rule of thumb, about one particle in a thousand accelerated by a flare escapes into interplanetary space. This number varies over orders of magnitude for different flares, and some particles are accelerated not by the flare process releasing magnetic energy, but by the shock front of the CME. Nevertheless, flares are frequent sources of energetic particles, both electrons and ions, in the solar wind, even though CMEs generally have a more dramatic impact on interplanetary space.

Major flares considerably influence the Earth's upper atmosphere. The prompt effects by flares must be distinguished from impinging CMEs, delayed several days by propagation. The enhanced x-ray emission of flares increases the ionization in the ionospheric D region at heights around 80 km. The increase in electron concentration leads to several phenomena grouped together under the name sudden ionospheric disturbance. The disturbance has a profound effect on the propagation of radio waves. A more dramatic ionization is caused by flare protons in the polar regions of the Earth, where the magnetosphere does not shield against solar cosmic rays.

The average rate of major flares has been up to thousand times higher when the Sun was younger and rotating faster as can be observed for example today on the 70 million year old star EK Dra. The solar activity has markedly decreased after the age of about one billion years. It may increase again for a short time in some 6 billion years, when the Sun will develop into the subgiant branch and the convection layer will temporarily deepen.

Bibliography

Benz A O 1993 *Plasma Astrophysics: Kinetic Processes in Solar and Stellar Coronae* (Dordrecht: Kluwer)

Güdel M, Schmitt J H M M and Benz A O 1995 Microwave emission from x-ray bright solar-like stars: the main sequence and beyond *Astron. Astrophys.* **302** 775–87

Svestka Z, Jackson B V and Machado M E (eds) 1992 *Eruptive solar flares (Lecture Notes in Physics 399)* (Berlin: Springer)

Tandberg-Hanssen E and Emslie A G 1988 *The Physics of Solar Flares* (Cambridge: Cambridge University Press)

Watanabe T, Kosugi T and Sterling A C (eds) 1998 *Observational Plasma Astrophysics: Five Years of Yohkoh and Beyond* (Dordrecht: Kluwer)

Arnold O Benz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Gamma Rays

Overview

Electromagnetic radiation extends over a very broad range of wavelengths, from gamma rays at the shortest wavelengths to radio waves at the long-wavelength end of the spectrum. In terms of photon energies, gamma rays are at the high end of the spectrum, starting at a few tenths of an MeV. This unit of energy (1 MeV $= 1.6 \times 10^{-13}$ J) is used throughout this article. In solar FLARES, as at many other astrophysical sites, gamma-ray emission results from interactions of fast particles with an underlying ambient medium. These fast particles, protons, α particles, heavier nuclei and electrons, are accelerated from the ambient plasma by the electric fields associated with the complex and varying magnetic fields in the flaring solar atmosphere. Thus, solar flare gamma rays can teach us about the mechanisms that accelerate the particles, in particular those which yield particles with energies in excess of about 1 MeV, the minimum energy needed to produce gamma rays. This is quite important for the understanding of flare mechanisms, because such protons and α particles, along with lower-energy electrons, contain the bulk of the energy released in flares. In addition, the solar flare gamma-ray emission exhibits characteristic spectral lines which provide information on the elemental composition of the ambient solar atmosphere.

Gamma-ray production mechanisms

Solar flare gamma-ray emission exhibits both lines and continuum (figure 1). This theoretical spectrum extends over the entire energy range in which gamma rays from flares were observed. The lines appear at energies from about 0.5 to 8 MeV, whereas the continuum extends up to at least 1000 MeV. Up to about 1 MeV, and again from about 10 to 50 MeV, the continuum is dominated by bremsstrahlung produced by the braking of the accelerated electrons in the Coulomb fields of the ambient nuclei and electrons. The bremsstrahlung produced by ultrarelativistic electrons is strongly collimated along the direction of motion of the electrons. The lines result from the deexcitation of nuclei, from the capture of neutrons and from the annihilation of positrons. The relevant nuclear cross sections are available from accelerator measurements. Deexcitation lines are either narrow or broad. Narrow lines result from the bombardment of ambient nuclei by accelerated protons and α particles, while broad lines result from the inverse reactions in which accelerated C and heavier nuclei collide with ambient H and He. The strongest narrow deexcitation lines are at 6.129 MeV from ^{16}O, 4.438 MeV from ^{12}C, 1.779 MeV from ^{28}Si, 1.634 MeV from ^{20}Ne, 1.369 MeV from ^{24}Mg and 0.847 MeV from ^{56}Fe. The broad lines merge into a quasi-continuum above the bremsstrahlung between about 1 and 8 MeV. The broadening of the deexcitation lines is the consequence of the Doppler shifting of the essentially monochromatic radiation produced in the rest frame of the excited nuclei. In the case of the narrow lines the broadening is due to the recoil velocity of the excited nuclei which is quite small. The widths of the broad lines are much larger because the excited nuclei continue to move rapidly after their excitation.

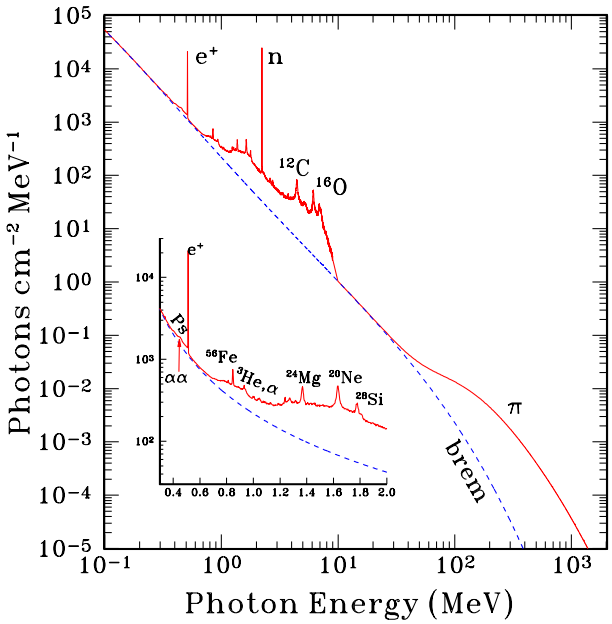


Figure 1. Theoretical solar flare gamma-ray spectrum constructed to fit the observations.

The strong, very narrow line at 2.223 MeV is due to neutron capture. All accelerated ions (protons and heavier nuclei) produce neutrons. The dominant neutron production mode in solar flares is the breakup of He nuclei, both in the accelerated particles and the ambient medium. Along with the deexcitation lines, the neutrons are produced at sites most likely located in the chromospheric portions of magnetic loops. The neutrons propagate both upward, away from the Sun, and downward into the photosphere where they are first thermalized by elastic collisions with protons and subsequently captured mostly by protons to produce deuterium and essentially monoenergetic photons at 2.223 MeV, the binding energy of deuterium. The neutrons moving away from the Sun can reach Earth where they were detected with ground-based and Earth-orbiting instruments. The protons resulting from neutron decay in the interplanetary medium were also detected.

The 2.223 MeV neutron capture line is very narrow because it is broadened by only the relatively low photospheric temperature of about 6000 K. Because the production site of the 2.223 MeV line is situated much deeper than that of the nuclear deexcitation lines, the 2.223 MeV line can be attenuated resulting in limb darkening. This means that, for flares located at or near the solar LIMB, the intensity of the line is much weaker than that of the deexcitation lines, in contrast with disk flares located far from the limb for which the 2.223 MeV line is the strongest.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

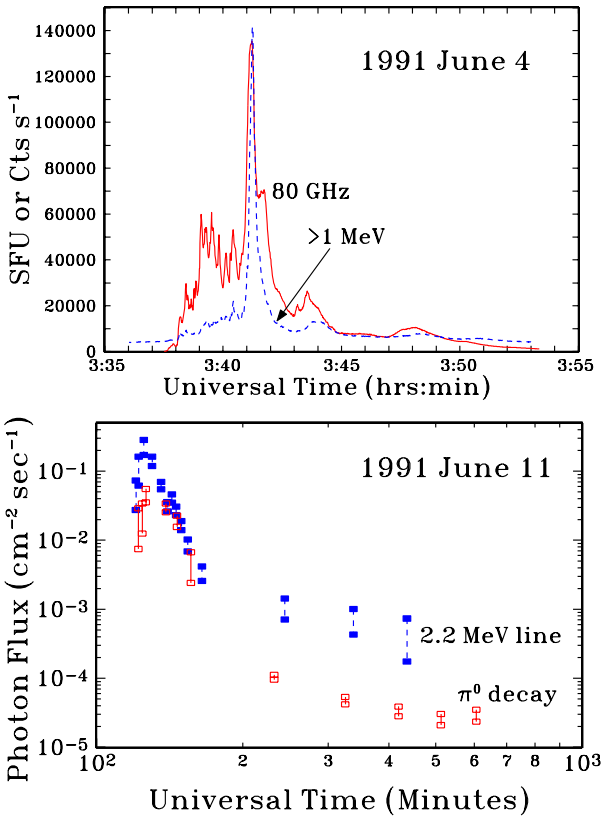


Figure 2. Time dependences. The upper panel exhibits the impulsive nature of the gamma-ray emission, along with the associated high-frequency radio emission. In contrast, the lower shows gamma-ray emission extending over many hours.

A competing mode of neutron capture in the photosphere is that on ^{3}He. This has been used to obtain information on the photospheric ^{3}He abundance.

Another strong narrow line is that at 0.511 MeV from positron annihilation (figure 1). The positrons result mainly from the decay of various short-lived radioactive nuclei, for example ^{11}C, ^{13}N and ^{15}O, which are also produced by interactions of the accelerated ions. The positrons subsequently either annihilate directly into 0.511 MeV gamma rays or form positronium (an atom analogous to hydrogen with the nuclear proton replaced by a positron), which also annihilates into gamma rays. Positronium annihilation leads to both line emission at 0.511 MeV and continuum below this energy. The positronium continuum (denoted by Ps) can be seen in figure 1 at energies just below the e^+ (0.511 MeV) line. The width of this line is very sensitive to the temperature of the medium in which the positrons annihilate. For the calculations of figure 1 it was assumed that the positrons annihilate in the CHROMOSPHERE.

Two strong lines result from the interactions of α particles with He. Fast and ambient ^{4}He nuclei fuse into ^{7}Li and ^{7}Be which are born either in their ground states or in their respective excited states at 0.429 and 0.478 MeV. Because of Doppler broadening, the ensuing deexcitations produce a relatively broad emission feature centered around 0.45 MeV (except under conditions of strong accelerated particle anisotropy when the two lines are narrowed into separate distinguishable features). The combined feature, generally referred to as the $\alpha\alpha$ line, can be seen in figure 1, superposed on the Ps and bremsstrahlung continua. Along with the $\alpha\alpha$ line, there are several other lines which can only be excited by accelerated α particles, as well as lines which are excited exclusively by accelerated ^{3}He nuclei. The latter are of interest because of the very large ^{3}He abundances observed in accelerated particles from impulsive flares (see below). The lines in question are at 1.00, 1.05 and 1.19 MeV and at 0.937, 1.04 and 1.08 MeV from α particle and ^{3}He induced reactions, respectively.

At high energies the continuum in some flares is dominated by pion decay radiation. Neutral and charged pions are produced mostly in high energy (greater than hundreds of MeV) proton–proton, proton–α particle and α–α interactions. Neutral pions decay directly into two photons, while charged pions decay (via muons) into secondary electrons and positrons which produce gamma rays via bremsstrahlung and annihilation in flight. The combined pion decay radiation is shown in figure 1.

The data and their implications

Gamma-ray lines from solar flares were first observed in 1972 with a detector flown on spacecraft. However, it was not until 1980 that routine observations of gamma-ray lines and continuum became possible with the much more sensitive spectrometer on the SOLAR MAXIMUM MISSION (SMM), a spacecraft that carried out successful solar observations for almost a decade. During that period, gamma-ray observations were also carried out with a smaller instrument on the Japanese spacecraft HINOTORI. During the 1990s, solar flare gamma rays have been detected with instruments on the COMPTON GAMMA RAY OBSERVATORY (CGRO). This observatory was launched in 1991, and it is expected that it will continue to operate well into the first decade of the 21st century. Additional solar gamma-ray observations during this period were carried out with instruments on the GRANAT and Gamma-1 spacecraft, which are no longer operational, as well as with a small detector on the YOHKOH spacecraft which continues functioning. Starting in 2000, a new spacecraft, the High Energy Solar Spectroscopic Imager (HESSI), will carry out solar flare x-ray and gamma-ray observations (see also SOLAR FLARE OBSERVATIONS). The main implications of the already available data are the following.

Flare energy release

Hard x-ray observations of solar flares demonstrated that a major fraction of the released flare energy resides in subrelativistic electrons of energies above 0.02 MeV. This, together with the observed impulsiveness of the hard x-rays, strongly suggested that electron acceleration to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

these subrelativistic energies is closely associated with the process that releases the flare energy initially stored in magnetic fields. However, prior to the availability of gamma-ray data, the accepted paradigm was that ion acceleration is only a secondary manifestation of the flare energy release process. The gamma-ray emission, however, turned out also to be very impulsive. Moreover, recent studies based on the relative intensities of gamma-ray lines (in particular the ^{20}Ne line at 1.634 MeV) provided new information on the energy distribution of the accelerated ions, requiring very large particle fluxes near 1 MeV. It was shown that the energy contained in such ions is comparable with that contained in the subrelativistic electrons. It thus appears that a large fraction of the released flare energy (approximately 10^{32} erg for large flares) is indeed in accelerated particles, but equipartitioned between ions and electrons. The top panel in figure 2 shows the impulsive time profile of gamma-ray emission produced by ions and electrons of MeV energies compared with very high-frequency radio emission produced by electrons of similar energies gyrating in solar magnetic fields of hundreds of gauss.

Particle acceleration and transport at the Sun

Particles accelerated at or near the Sun are also observed by detectors on spacecraft in interplanetary space. These observations have led to the identification of two classes of acceleration events, impulsive and gradual. Among the various characteristics of the two classes, the composition of the accelerated particles is perhaps the most important. The impulsive events exhibit large enhancements of relativistic electrons relative to MeV protons, of ^{3}He relative to ^{4}He and of heavy ions (particularly Fe) relative to the C and O. In contrast, the gradual events have smaller electron-to-proton ratios (e/p), and their heavy ion abundances and ^{3}He-to-^{4}He ratios are similar to coronal values. The strong association of the gradual events with SOLAR CORONAL MASS EJECTION (CME) suggests that the particles in these events are accelerated by CME driven shocks. On the other hand, the electron, ^{3}He and heavy ion enrichments in impulsive events require selective acceleration which is most likely achieved by gyroresonant interactions with plasma waves.

The gamma-ray observations have independently revealed the characteristics of impulsive acceleration. In particular, high e/p ratios are required by the observed continuum-to-line flux ratios and high ^{3}He abundances are suggested both by the 2.223 MeV line observations, which require enhanced neutron production, and very recent findings in SMM data of ^{3}He induced lines. In addition, there is evidence in the Granat data for highly enhanced heavy ion abundances based on broad lines from Ne, Mg, Si and Fe.

The increased sensitivity of the CGRO detectors and the occurrence of large flares while these instruments were observing the Sun have shown that flares can produce gamma rays for very long periods of time. The data, showing gamma ray emission lasting for up to 8 h after the impulsive phase of the flare, can be seen in the bottom panel of figure 2. It is still not known whether the particles were accelerated in the impulsive phase of the flare and subsequently trapped in magnetic loops at the Sun or accelerated continuously over the duration of the emission.

Additional information on accelerated particle transport at the Sun was obtained from observations of the 2.223 MeV line and electron bremsstrahlung above 10 MeV. The limb darkening of the 2.223 MeV was observed from many solar flares, and it was demonstrated most dramatically by gamma-ray observations of a flare located 10 degrees behind the limb for which the 2.223 MeV line was absent while the deexcitation lines were still seen. Evidently, a considerable fraction of the interactions occurred in the CORONA, at a site which was visible from Earth orbit while the neutron capture site in the SOLAR PHOTOSPHERE was occulted. On the other hand, there is one observation of a behind-the-limb flare from which the 2.223 MeV line was seen. Because of the very strong expected attenuation, the observed 2.223 MeV line must have been produced by charged particles interacting on the visible hemisphere of the Sun. These particles were either accelerated by a coronal shock over a large volume, thereby producing an extended gamma-ray-emitting region, or accelerated locally at the flare site whence they propagated along large loops to the visible hemisphere.

In contrast to the limb darkening of the 2.223 MeV line emission, the bremsstrahlung above 10 MeV was observed to be limb brightened. This means that the flares from which such emission was observed were preferentially located close to the solar limb. This effect is most likely the consequence of particle motion in magnetic loops which converge toward the footpoints with the particles radiating most efficiently when they move parallel to the photosphere near the mirror points.

Ambient medium abundances

SMM and CGRO data on narrow gamma-ray lines have provided information on solar atmospheric elemental abundances (see SOLAR ABUNDANCES). While the C-to-O abundance ratio was found to be consistent with both photospheric and coronal values, the Mg-to-O, Si-to-O and Fe-to-O ratios turned out to be enhanced relative to the photospheric abundances but consistent with those of the corona. The first ionization potentials (FIPs) of Mg, Si and Fe are lower than those of C and O. The enhancement of the abundances of low-FIP elements in the corona relative to the photosphere has been known from both atomic spectroscopy and particle observations of gradual events, but the origin of this fractionation is still only poorly understood. The gamma-ray results, and the fact that the gamma-ray lines are most likely produced in the chromosphere, indicate that the FIP bias sets in already at relatively low heights in the solar atmosphere. Ongoing research on the $\alpha\alpha$ line indicates that in the gamma-ray production region either the α particle or the ambient He abundance is enhanced, exceeding the standard He/H value of 0.1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The early development of the field is summarized in the reviews of Chupp and of Ramaty and Murphy. Much of the recent observations and theory, the relationship of the gamma-ray studies to other solar flare investigations, as well as a detailed historical review (by Chupp), are given in the HESP conference proceeding edited by Ramaty, Mandzhavidze and Hua. The Murphy *et al* paper provides details on ongoing research.

Chupp E L 1984 High-energy neutral radiations from the Sun *Ann. Rev. Astron. Astrophys.* **22** 359–87

Murphy R J *et al* 1997 Accelerated particle composition and energetics and ambient abundances from gamma-ray spectroscopy of the 1991 June 4 solar flare *Astrophys. J.* **490** 883–900

Ramaty R and Murphy R J 1987 Nuclear processes and accelerated particles in solar flares *Space Sci. Rev.* **45** 213–68

Ramaty R, Mandzhavidze N and Hua X M (ed) 1996 *High Energy Solar Physics* (New York: American Institute of Physics)

R Ramaty and N Mandzhavidze

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Impulsive Phase

Although impulsive phenomena are commonplace in the universe, solar FLARES are unique in astrophysical research because they provide the largest diversity of observational data, which means diagnostic information available to the observer. Solar flares are truly unique, insofar as they provide the vantage point of seeing the development of impulsive phenomena with high temporal as well as spatial resolution. This also includes the possibility of direct *in situ* measurements of its byproducts, such as accelerated particles in interplanetary space. As we shall demonstrate below, the analysis of all this information has led to the inevitable conclusion that flares are very efficient particle accelerators. A significant fraction of the 10^{28}–10^{33} erg of magnetic free energy that are liberated during an event is released as energetic particles. Since such release occurs in the few seconds to few tens of minutes that it lasts, it is no wonder that the impulsive phase can be considered as the 'core' of the flare phenomenon.

Observations

The impulsive phase of a flare is characterized by a rapid increase in radiation intensity at all wavelengths, spanning from radio frequencies to the gamma rays. Its main characteristic is the ubiquitous presence of rapidly fluctuating bursts of hard x-ray (HXR) radiation, as well as similarly structured and cotemporal emission in the ultraviolet (UV) and extreme ultraviolet (EUV), decimeter and centimeter radiation, and the so-called type III and U radio bursts. In UV and soft x-rays (SXR), the spectra show redshifts and blueshifts in lines formed at 10^5 and 10^7 K temperatures respectively. When looking at its spatial characteristics, the impulsive phase can be adequately described as composed of 'kernels' or 'knots' at all wavelengths, whose area usually represents only a fraction of the, ensuing, 'main phase' of the flare (see SOLAR FLARES: MAIN PHASE). These kernels were first detected, and can easily be seen, in ground-based off-band H-alpha observations as regions of enhanced and highly broadened Balmer line emission (and sometimes continuum enhancements). These characteristics indicate that the emission originates deep in the solar chromosphere or upper photosphere, at layers where the electron density exceeds 10^{13} cm^{-3}. With the advent of space-borne solar observatories, first SKYLAB, then the SOLAR MAXIMUM MISSION (SMM) and HINOTORI and more recently YOHKOH, the concept of impulsive kernels was extended to the EUV, UV and x-ray domains. However, the most definite indication of what, as we shall see, is the basic physical property of the impulsive phase came with the advent of HXR imaging from SMM, Hinotori and Yohkoh. Observations from the SMM spacecraft showed the presence of HXR kernels spatially coincident with the UV and H-alpha knots at the feet of bipolar magnetic loops.

The observations showed, when combined with vector magnetic field observations, that the bulk of the HXR emission was concentrated in loops rooted in highly nonpotential field regions, thus indicating that impulsive phenomena are associated with current-carrying bipoles. The observations also show that, even though the basic structure of a flare consists of a bipolar loop or arcade of loops, the observed impulsive phenomena generally encompass two or more bipolar structures. This suggests that their interaction leads to the impulsive release of magnetic free energy.

Interpretation of the observations

To explain all the phenomena that characterize the impulsive phase, solar physicists have developed over the years a most useful paradigm, the so-called nonthermal thick-target model (TTM). In the TTM, initially developed to explain HXR burst observations, the dissipation of energy, presumably occurring at the top of the loop, first goes into the acceleration of electron beams to energies $\geq$ 20 keV with a power law distribution $AE^{-\delta}$. As these electrons stream down the legs of the loop, they lose energy primarily through electron–electron collisions and produce HXR via a bremsstrahlung process in electron–ion interactions. As long as the ambient loop density is small ($\leq 10^{10}$ cm^{-3}), the beam electrons lose little energy and also produce little HXR emission. However, on reaching the dense layers of the chromosphere ($\geq 10^{12}$ cm^{-3}), both processes increase dramatically and the electrons are stopped, heating the ambient plasma and more efficiently producing HXR photons. It should be noted that only a small fraction, of order 10^{-5}, of the beam energy, is 'used' in bremsstrahlung emission, the bulk is spent in heating ambient electrons with characteristic preflare temperatures $\leq 10^4$ K (see also SOLAR FLARES: PREFLARE PHASE). As with any theoretical model, the power of the TTM resides not only in its ability to explain the data from which it is deduced (in this case HXR burst intensity and spectra), but also in successfully predicting other observational signatures. The TTM predicts the following.

- The HXR emission should be concentrated at the chromospheric feet of coronal loops.
- The upper chromosphere is heated and driven upwards, owing to convective flows caused by the high pressure of the previously cool and dense plasma.
- This hot plasma upflow increases the coronal loop density, causing an increase in its emission measure and, thus, a rise in SXR emission.
- The beam heating also produces a cotemporal increase in transition zone lines' emission (characteristic $T \cong 10^5$ K), which are seen at UV and EUV wavelengths.
- Momentum balance at the transition layer, between the upflowing SXR emitting plasma and the cooler and denser transition zone and chromospheric material, leads to the condensation and downflow of the cooler plasma, causing red-shifted UV lines and asymmetric H-alpha profiles.
- The energy deposition in the deep layers of the chromosphere, due mostly to electrons with energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

50 $\geq$ keV, causes an increase in visible line emission and the broadening of hydrogen line profiles (such as H-alpha and other Balmer lines), owing to the Stark effect caused by high ($>10^{13}$ cm^{-3}) electron densities.
- Relativistic electrons streaming along the loops produce microwave bursts, because of gyrosynchrotron emission as they spiral along the magnetic field and precipitate to the footpoints.
- Those electrons that escape from the closed bipolar loop, either by getting into open field structures or into neighboring longer-scale bipoles, produce type III and/or reverse U radio bursts respectively. In many cases, H-alpha brightenings are observed at the far end of such loops, in such a close time correspondence to the main flare site that any exciting agent other than energetic electrons is ruled out.

No other model, with perhaps the sole exception of *ad hoc* versions of the so-called thermal model (TM), to which we shall refer in the next section, can account for so many observational constraints. Furthermore, although not explicitly stated in the formulation of the TTM, the acceleration of ions during the impulsive phase can also explain observed gamma-ray phenomena. The interaction of accelerated ($\geq$ 1 MeV per nucleon) ions with ambient nuclei leads to the production of 1–7 MeV nuclear deexcitation gamma-ray lines, the positron annihilation line at 0.511 MeV and the neutron capture line at 2.2 MeV. Reactions of relativistic ions with ambient particles also produce pions and high-energy neutrons. Neutrons with energies between 50 and 500 MeV are produced by protons with up to GeV energies and have been directly observed in space. The pions decay into gamma rays or into ultrarelativistic secondary electrons and positrons, which also contribute to the observed >10 MeV continuum. For a long time it was believed that only a few of the strongest flares produced gamma rays. However, the advent of more sensitive space-borne detectors has shown that the lack of detected gamma-ray emission in many flares may be due to just its weakness rather than its absence. Forthcoming observations, such as those to be obtained with detectors to be flown aboard the HESSI mission, will show whether the great majority of impulsive events are accompanied by high-energy photon and neutron emission. They will also contribute to determining how important, to overall flare energetics, ions are.

The energy and number problem

As we mentioned above, the efficiency of the bremsstrahlung process is very low, so that $\approx 10^5$ erg of electron kinetic energy are necessary to produce just 1 erg of HXR emission that is radiated. We also mentioned that the observed photon spectrum of HXR can be reasonably well fitted by a power law distribution $aE^{-\gamma}$. As shown by J C Brown, R P Lin and H S Hudson and others, in the TTM the energy differential rate at which the electrons are injected into the flaring loop is also a power law $AE^{-\delta}$, where $\delta \approx \gamma + 1$. Then, from the observed intensity or flux of HXRs at 1 AU, one can estimate the total number of electrons accelerated per second at the source. Typical HXR fluxes show that some 10^{35}–10^{37} electrons s^{-1} have to be accelerated to energies $\geq$20 keV in order to reproduce the observed impulsive phase HXR photon spectra. To illustrate the relative magnitude of these numbers, we can mention that a 'typical' quiescent active region loop contains $\approx 10^{36}$ electrons, meaning that the whole electron content of a loop is being accelerated per second. It is also very easy to calculate that the power dissipated in the accelerated electrons' distribution is, for typical impulsive phase parameters, some 10^{28}–10^{30} erg s^{-1} over the 10^2 s duration of the burst. To complicate matters even further, we note that in the largest events the electron number, power and duration of the burst can all be up to two orders of magnitude larger than these 'typical' values. Thanks to the orders of magnitude denser chromospheric reservoir, that can provide a continuous supply of 'fresh' thermal electrons, near-real-time replenishment of the acceleration region may not be a problem. The power issue instead is difficult to avoid. It implies that a substantial fraction of the total energy of the flare resides in accelerated electrons. In view of this, considering that the bulk of the power resides in the low-energy end of the power law distribution, an alternative TM has been widely discussed over the past 20 years. In the TM concept, the lower-energy photons are predominantly emitted by hot (10^8 K) thermal electrons, confined by ion-acoustic turbulent fronts. Since the confined plasma electrons collide among themselves, the ratio of bremsstrahlung to collisional energy loss could, in principle, be orders of magnitude higher than in the TTM. However, the turbulent fronts are 'transparent' to the higher-energy electrons (those with speeds >3 times the thermal speed), so that there is a leakage of high energy electrons that precipitate to the footpoints in a thick target fashion. Considering the aforementioned characteristics of impulsive phase phenomena, such resemblance might be considered a favorable feature, since energetic particle precipitation is needed to explain many of the observations. However, it also means that, in order to explain HXR burst observations, the escaping tail needs to be regenerated quasi-continuously. This reduces the theoretical high efficiency of the TM, and turns it into a hybrid, where acceleration again plays an important role.

New perspectives and the future

It is noteworthy that the first observations ever recorded of a solar flare, by Carrington and Hogdson in 1859, can also be recognized as the first observation of the impulsive phase of a flare. One hundred and forty years later, we are able to understand their white light continuum observations. They are the effect of precipitating high-energy particles into deep layers of the solar chromosphere and upper photosphere. Carrington immediately noticed the close temporal correspondence of the event he had observed to terrestrial disturbances but judiciously refrained from speculating about a cause–effect relationship. Many decades of dedicated research

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and improvements in observational methods have shown that such a relationship does indeed exist, and this has developed as one of the most dynamic areas of solar–terrestrial research.

We thus find it worthwhile to conclude by pointing out some recent results obtained from ground- and space-borne instrumentation that may still be a subject of debate. First of all, new observational evidence seems to converge towards a view that impulsive phenomena originate at relatively high coronal levels. Radio emission, due to upward and downward directed electron beams, gives an indication that the source of these accelerated electrons is in the corona, at levels where the electron density is of order 10^9–10^{10} cm^{-3} (see SOLAR WIND: ENERGETIC PARTICLES). Observations from the Yohkoh spacecraft have also shown cases of high-altitude coronal HXR emission regions. These sources, detected in flares occurring near the solar limb, are cotemporal with footpoint HXR emission, although considerably weaker, and located above the SXR-emitting loops seen by the spacecraft's SXR telescope. These sources have been interpreted as evidence of magnetic reconnection and energy release above the, already cooling, SXR-emitting loops (see SOLAR FLARE OBSERVATIONS and SOLAR FLARE MODELS). Although their interpretation may still be a matter of debate, the mere existence of the coronal HXR sources signals that suprathermal particles are being generated or injected there, traveling downwards and upwards, as the radio observations suggest. The downward-traveling particles cause the footpoint HXR emission, while those traveling upwards either escape or lead to remote brightenings, when they access large-scale closed field structures.

The most intriguing aspect of impulsive phase physics resides in the mechanism that leads to the release of, say, 10^{30} erg in 10^2 s. This is equivalent (e.g.) to the annihilation of a 100 G field over a $(3 \times 10^8$ cm$)^3$ cube in such a time interval. This is particularly difficult to accomplish under coronal conditions of low resistivity, unless the actual dissipation occurs at spatially unresolvable microscale levels. For many years, radio observations have given an indication of a 'fragmentation' of impulsive phase emission, as evidence of microstructure in the energy release region. This property has been confirmed by observations of prompt and optically thin bremsstrahlung HXR emission, which give direct evidence of the timescales associated with the energization of suprathermal electrons. Observations obtained with BATSE (Burst and Transient Source Experiment) aboard the Compton Gamma Ray Observatory (CGRO) have shown that HXR bursts are composed of both spiky structures due to precipitating electrons and a gradual component with particle trapping characteristics (figure 1). The range of timescales associated with the spiky component ranges from 200 to 1000 ms, which could be associated with dissipation regions with characteristic lengths $\lambda \approx 10^7$ cm.

The resolution of the impulsive phase enigmas will be addressed through a worldwide coordinated program of flare observations to start in the year 2000. Among the observations obtained by ground observatories and spacecraft, a major contribution to impulsive phase studies will be provided by the High Energy Solar Spectroscopic Imager (HESSI) mission. HESSI will obtain HXR images and spectra with an unprecedented combination of spatial, temporal and spectral resolution (2 arcsec to 300 keV, tens of milliseconds, and less that 1 keV FWHM increasing to 5 keV at 20 MeV). With its fine resolution and sensitivity, it will be possible to locate both coronal and chromospheric HXR sources and carry out a numerical inversion of the spectral data. This will help determine the parent electron spectrum as a function of time and energy. Furthermore, HESSI's gamma-ray spectroscopic imaging will provide the first information ever obtained on the spatial distribution of gamma-ray lines and on the spatial distribution of energetic (>1 MeV) protons.

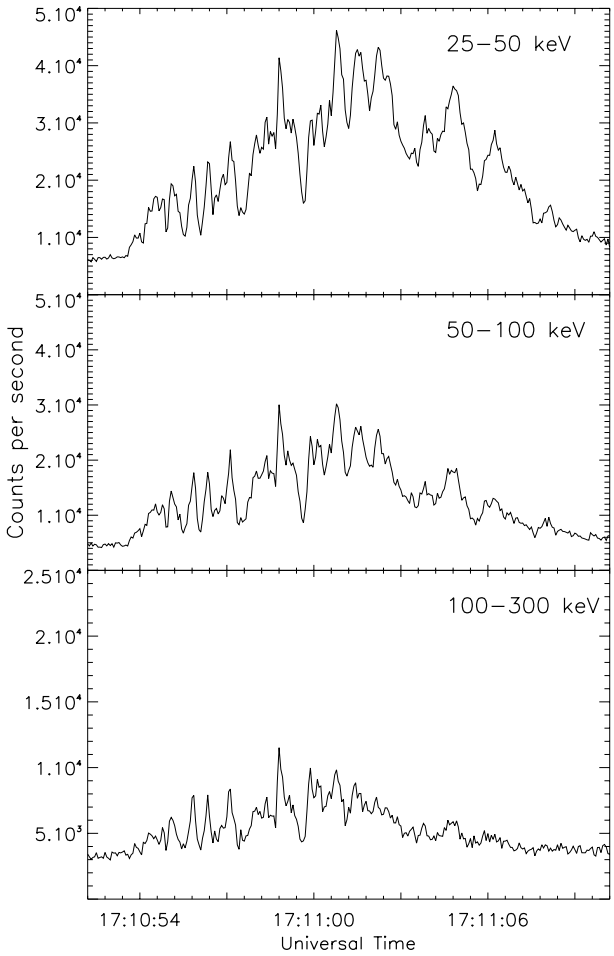


Figure 1. Observations of an HXR burst detected by BATSE in three energy intervals. The temporal resolution is 64 ms, the bin size, and the shortest statistically significant fluctuations observed in this event are in the 200–500 ms range. The burst was associated with an x-ray class C3 flare of 26 June 1991.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Priest E R (ed) 1981 *Solar Flare Magnetohydrodynamics* (New York: Gordon and Breach)

Somov B V 1992 *Physical Processes in Solar Flares* (Dordrecht: Kluwer)

Svestka Z 1976 *Solar Flares* (Dordrecht: Reidel)

Tandberg-Hanssen E and Emslie A G 1988 *The Physics of Solar Flares* (Cambridge: Cambridge University Press)

Marcos E Machado

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Main Phase

After the impulsive phase of a flare is over (see SOLAR FLARES: IMPULSIVE PHASE), we talk about the main phase or gradual phase. Most FLARES (often called compact or confined flares) are short lived and they simply cool during the main phase, decaying within minutes or tens of minutes. However, there is also another kind of flares (eruptive flares) which continue to release energy during their main phase, and in those flares the decay is much longer: in some events it can continue for many hours. These eruptive flares are much more interesting, because they are often associated with other active phenomena in the solar CORONA which may be sources of disturbances in interplanetary space and in the Earth environment.

Confined flares

The majority of flares (and essentially all subflares) are confined flares—relatively simple structures consisting of one or several brightened loops. It is likely that even those flares which look like a brightened single loop actually consist of more loops which our telescopes are unable to resolve. After the heating process, accomplished during the impulsive phase, these loops begin to cool. This cooling is due partly to heat conduction from the hot flare loop to the cooler CHROMOSPHERE (conductive cooling) and partly to radiation emitted by the loop (radiative cooling) and in some cases also due to mass flow to the loop footpoints. The conductive cooling time increases with the length of the loop, while the radiative cooling time decreases with increasing density. As a flare can consist of a set of loops of various lengths and densities, its cooling is a complex process: short and dense flare loops disappear first, while long and less dense loops survive longer.

In addition to these internal differences in the flare structure, there occur sometimes secondary phenomena that accompany confined flares and eventually may become more impressive, and last longer, than the flare itself. For example, energetic electrons accelerated during the impulsive phase of a flare can stream into a pre-existing larger coronal loop which is rooted close to the flare region, thus forming a less bright, but much larger, structure which, because of its large size and low density, can be seen much longer than the flare which caused its brightening. Another example are flaring arches, in which material of widely different temperatures and densities flows along magnetic field lines from the flare site to another footpoint which sometimes can be located more than 100 000 km afar. In some cases, this material does not reach any secondary footpoint, but just rises and falls back along magnetic field lines, thus creating a flare-associated SURGE. Thus even the 'simple' cooling main phase of a confined flare may be sometimes very complex and not easy to understand.

Eruptive flares

More interesting flares on the Sun belong to the other kind: events in which energy release continues during the main phase. This class comprises a wide energetic range of phenomena which all can be interpreted in the same way: the originally closed magnetic field is disrupted so that its field lines open, and subsequently the open field lines begin to reconnect. The field disruption is often indicated by an activated dark filament which eventually erupts and rises into the corona, but this need not be observed in all events of this kind: sometimes the disrupted configuration is young, or it already disrupted once shortly before, so that there has not been enough time left for plasma to condense and cool to temperatures which make a filament visible in the Hα line.

After the field lines open, mass begins to flow from the dense lower layers to the corona. Therefore, the magnetic pressure begins to prevail over the gas pressure and this imbalance causes reconnection of the open field lines: first fast, low in the atmosphere, where the magnetic field is strongest, and thereafter with progressively slower speed at higher altitudes. As the field lines reconnect, they create new loops at progressively higher altitudes and release energy at the reconnection site, i.e. at the tops of the loops. The loops are formed at very high temperature (close to 20 million K and still higher in the early phase of the loop-system creation) and are first visible in x-rays. Thereafter they gradually cool, so that one can see them in images corresponding to lower temperatures, until they become visible in the Hα line at temperatures below 20 000 K and eventually disappear. As the result of this process we see a growing system of flare loops, in which the lowest loops cool and decay while new ones are formed above them. The growth of this loop system proceeds with speeds of a few, up to a few tens of, kilometers per second at the beginning of the loop system formation and decreases to less than 1 km s^{-1} late in the main phase. Because this process of loop formation and cooling can continue for hours, long after the impulsive phase is over, these newly formed loops are often called post-flare loops (incorrectly, as they actually form the long-lasting main phase of the flare).

The released energy at the reconnection site heats the chromosphere at the loop footpoints which begins to evaporate into the loops, thus enhancing the loop density and making the loops visible. The heating also produces chromospheric emission in the Hα line which is aligned at both sides of the neutral line of the magnetic field, thus creating two bright flare ribbons. As the loops rise in the corona, their footpoints move farther from the neutral line and the two ribbons gradually separate from each other.

The amount of energy released, and thus also the intensity of this chromospheric emission, depends on the strength of the magnetic field in which the process of field opening and reconnection has been accomplished. In very strong fields, i.e. in fully developed ACTIVE REGIONS, we find enormously bright two-ribbon flares, an example of which is shown in figure 1. Essentially all major flares on the Sun belong to this category. In weaker fields the Hα ribbons are less bright, and they may be completely invisible when the process occurs in a very old decaying active region. Such events are then indicated in Hα only by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

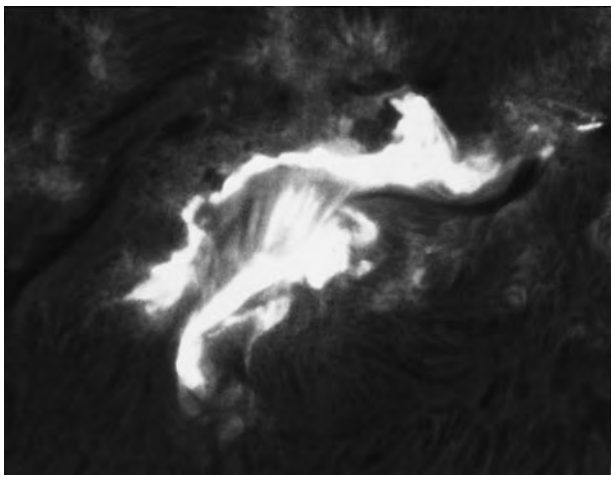


Figure 1. Main phase of a major eruptive (two-ribbon) flare observed in the Hα line at Big Bear Solar Observatory. Two bright flare ribbons extend along the neutral line, marked in the preflare state by a dark filament—its remnant is still visible at the right-hand side of the figure. Inner edges of the ribbons are connected by loops, some in emission and other in absorption.

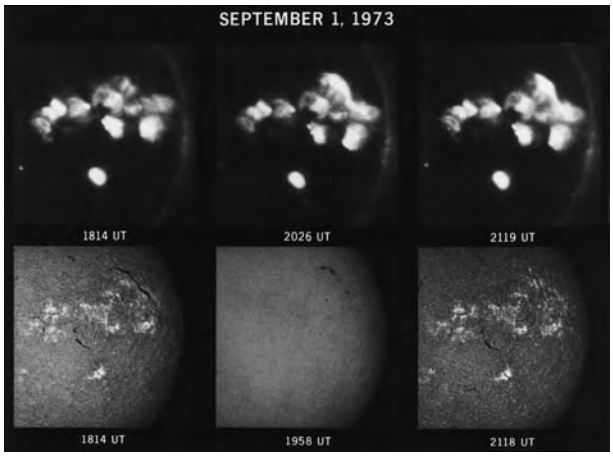


Figure 2. 'Disparition brusque' of a quiescent filament as seen in Hα (below) and in soft x-rays (above, aboard Skylab). The central Hα frame is out of the line center and shows the moving (Doppler-shifted) material of the filament.

the disappearance of an activated dark filament, and have been called disparitions brusques (figure 2). However, all these phenomena, from the disparitions brusques up to the major two-ribbon flares, although accomplished in regions with different magnetic characteristics, represent the same process on the Sun, and, therefore, have been called, all together, eruptive flares.

While in Hα some these events do not look like flares (and, therefore, the term 'eruptive flares' may be somewhat misleading), in x-rays all eruptive flares have one characteristic common: a bright x-ray ribbon-like emission extending along the neutral line at the top of the reconnecting loops (figure 2) and, in consequence of

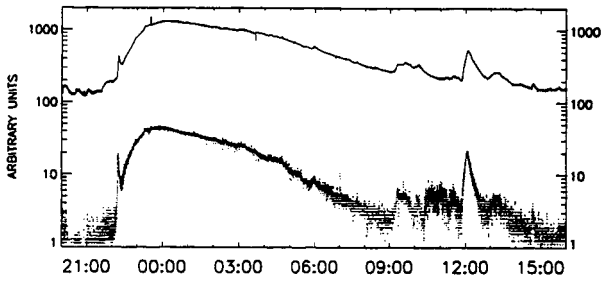


Figure 3. A long-lasting enhancement in soft x-rays (1–8 Å above and 0.5–4 Å below) recorded by the GOES satellite during an eruptive (LDE) flare. The horizontal scale gives time from 21:00 UT on 28 August 1992 through 15:00 UT next day. X-ray enhancements corresponding to confined flares are short lived, like that around 12:00 UT.

it, they all produce long-lasting enhancements of x-ray emission in x-ray records of the Sun (figure 3). For this reason this kind of flares is also called long-decay events, often abbreviated as LDE. Other terms, sometime used for the same phenomena, are dynamic flares or ejective flares.

Complex events

The classification into two classes of flares, confined and eruptive, somewhat simplifies the real situation, because there are also complex flares on the Sun which do not fit exactly any of these two classes. Let us mention the most common cases.

Most (and perhaps all) eruptive flares are related to SOLAR CORONAL MASS EJECTIONS (CMEs). One would expect that one observes the CME first, when the magnetic field disrupts and opens, and only then, through reconnection, the flare itself begins to be formed, but that is not always the case. In some events one begins to observe the flare emission already before any CME is recorded. A likely explanation is that we see first a confined flare, which triggers the field disruption and thus initiates the eruptive flare that follows. In such a case the main phase of the confined flare that triggers the whole process merges with the early phase of the triggered eruptive flare (which, as it seems, needs not always to be impulsive). In some cases a flare is followed in x-rays by an extensive arcade of growing loops which may extend over very large distances on the Sun, far from the original site of the flare. In such cases apparently magnetic field lines opened along a very long portion of the neutral line and subsequently reconnected. However, it is indicated that a substantial part of the field lines stays open for a long period of time and only some of them reconnect. The same may be true, although perhaps to a lesser extent, in all eruptive flares. Then the flaring active regions in which open field lines do not fully reconnect or in which the process of full reconnection takes a long time may become temporary sources of SOLAR WIND.

There are also eruptive flares in which the speed of growth of the loop system does not decrease with time:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

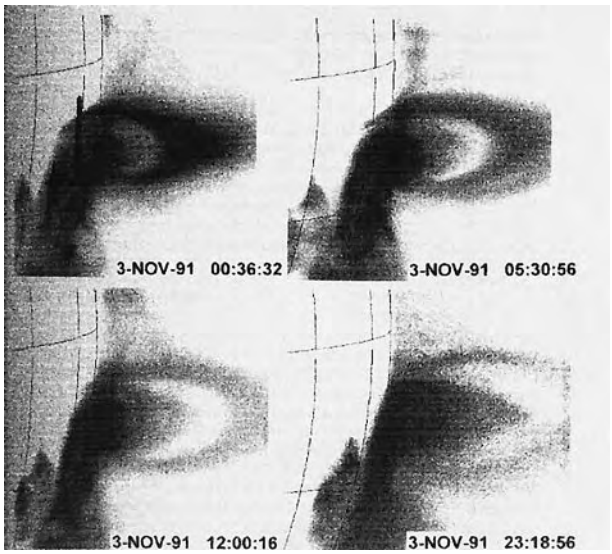


Figure 4. Soft x-ray images of a rising giant arch obtained aboard Yohkoh. The arch followed an eruptive flare of 2 November 1991.

during most of the main phase the loops grow with a constant speed of a few km s^{-1} for many hours. This indicates that the loops are continuously heated, do not cool, and expand.

Coronal main phase

Confined flares may be associated with radio type III or type V bursts (see SOLAR FLARES: RADIO BURSTS) which are related to the flare impulsive phase, but the main phase usually has no response in higher coronal layers. On the other hand, above eruptive flares the solar corona sometimes shows striking changes, and energetic particles (see SOLAR WIND: ENERGETIC PARTICLES), accelerated during the flare, are often trapped in the corona above the flaring active region, where they manifest themselves in several different forms. One can consider these phenomena as the main phase of the eruptive flare in the corona.

On metric radio waves one often observes a type IV burst which is a continuum radio emission, usually consisting of two parts: a moving type IV starting shortly after the impulsive phase and rising in the corona for several tens of minutes and a stationary type IV that appears later below the rising structure. There are also events in which only the stationary type IV is present and, on the other hand, this component may be missing in events observed close to the solar limb. The stationary type IV burst may be very long lived: in some events the continuum gradually becomes bursty and eventually can change into a type I noise storm, which then stays above the flaring active region for tens of hours or even several days.

Soft X-RAY TELESCOPES in space record after some eruptive limb flares coronal phenomena which have been called giant arches. Some of them are stationary, corresponding to lower and denser parts of a stationary type IV burst or a noise storm. While the radio emission is due to gyro-SYNCHROTRON RADIATION or SOLAR WIND PLASMA WAVES, the x-ray emission is produced by bremsstrahlung of the energetic electrons. Other giant arches rise in the corona with constant speeds of 1 or a few km s^{-1} for tens of hours (figure 4). It is likely that these arches are rising loop-like structures, left behind a CME, which begin to expand into the coronal 'vacuum' above them. Although these flare-associated coronal arches can be observed in x-rays only above flares that occur close to the solar limb, they are probably common phenomena characterizing the coronal main phase of eruptive flares.

Bibliography

Schmieder B 1992 'Post'-flare loops *Eruptive Solar Flares (Lecture Notes in Physics 399)* ed Z Švestka, B V Jackson and M E Machado (Berlin: Springer) pp 124–33

Švestka Z 1989 Solar flares: the gradual phase *Solar Phys.* **121** 399–417

Švestka Z, Poletto G, Fontenla J, Hick P, Kopp R A, Sylwester B and Sylwester J 1999 The gradual phase *The Many Faces of the Sun: a Summary of the Results from NASA's Solar Maximum Mission* ed K T Strong, J R L Saba, B M Haisch and J T Schmelz (Berlin: Springer)

Tandberg-Hanssen E and Emslie A G 1988 *The Physics of Solar Flares* (Cambridge: Cambridge University Press) pp 206–28

Z Švestka

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Nonthermal Electrons

Solar flares energize particles to different levels, usually subdivided into (1) thermal, (2) nonthermal or energetic and (3) high energetic (figure 1). The lowest energetic level is a thermal particle distribution (where particles have a Maxwellian velocity distribution), produced by Coulomb collisions between the particles of the heated flare plasma. Because the coronal plasma is heated to temperatures of 10–30 million K during a solar flare, thermal electrons have a mean kinetic energy of 0.4–1.3 keV, with tails reaching up to a maximum of about 20–30 keV. The next higher energetic level is referred to as nonthermal particles. Particles acquire nonthermal energies when they become accelerated above the thermal level. Accelerated electrons maintain their nonthermal nature as long as they propagate in a sufficiently low-density (collisionless) plasma so that they do not suffer Coulomb collisions and become thermalized again. Nonthermal electrons with typical energies of 20–100 keV are often observed in hard x-ray and radio wavelengths during flares. In strong flares, particles become accelerated up to 1–10 MeV, referred to as high energetic, and can be detected in gamma ray wavelengths (see SOLAR FLARES: GAMMA RAYS).

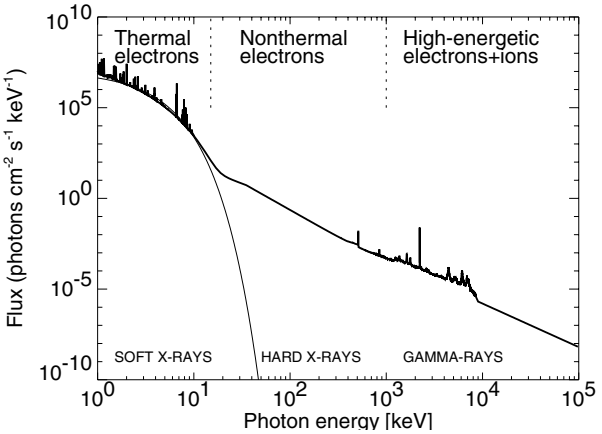


Figure 1. Composite photon spectrum of a large flare, extending from soft x-rays (1–10 keV), through hard x-rays (10 keV–1 MeV), to gamma rays (1 MeV–10 GeV), mostly produced by thermal, nonthermal (energetic), or high-energetic electrons. Gamma-ray line emission and parts of the gamma-ray continuum are produced by interactions of protons, neutrons, ions and pion decay.

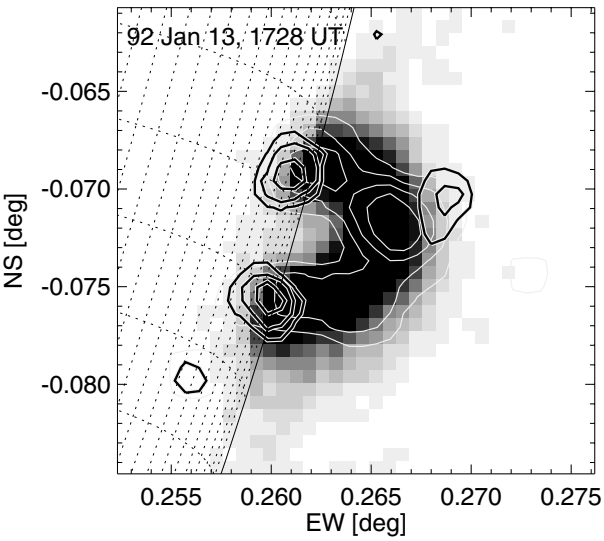


Figure 2. Discovery of the acceleration site of nonthermal electrons by Masuda in the flare of 13 January 1992. The flare loop is bright in soft x-rays (indicated with white contours and grayscale), and shows an above-the-loop-top hard x-ray source (black contours) where electrons are accelerated and two hard x-ray footpoints (black contours), where electrons precipitate to the chromosphere and produce the most intense hard x-ray bremsstrahlung there.

Particle acceleration

Particle accelerators were built in large physics laboratories to study the elementary structure of particles, by analysing the kinematics of high-energetic particles and their decay products in collision experiments. In solar flares, such high-energy experiments take place too, when particles accelerated in the coronal plasma propagate towards the solar surface and collide there with the dense chromospheric plasma. However, how the particles become accelerated in the coronal flare plasma is difficult to probe. There are a number of theoretical models for the accelerations process, such as (1) acceleration by electric fields, (2) stochastic acceleration or (3) shock acceleration (see SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS), but their diagnostics can only be accomplished by remote sensing methods. It was therefore until recently not even clear where the acceleration sites are located in a solar flare. Only the most recent observations by the YOHKOH hard x-ray telescope (HXT) and the COMPTON GAMMA RAY OBSERVATORY (GRO) enabled a breakthrough in the localization of particle acceleration sites.

The most significant discovery was made by Satoshi Masuda, who identified weak hard x-ray sources above flare loops that appear bright in soft x-rays (figure 2). Because the weak above-the-loop-top hard x-ray sources were detected up to 50 keV and their spectrum was identified as nonthermal, they provided for the first time a glimpse of the actual acceleration site. Because Yohkoh revealed also cusp-like plasma structures above the soft x-ray flare loops, the above-the-loop-top hard x-ray sources were brought into context with MAGNETIC RECONNECTION processes. Accurate electron time-of-flight measurements with Compton GRO confirmed that the propagation distance between the acceleration site and the chromospheric footpoints is consistent with the location of the acceleration site found by Masuda. A third piece of evidence for this location of the acceleration site was found from bi-directional electron beams detected at decimetric radio wavelengths, which show electron

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

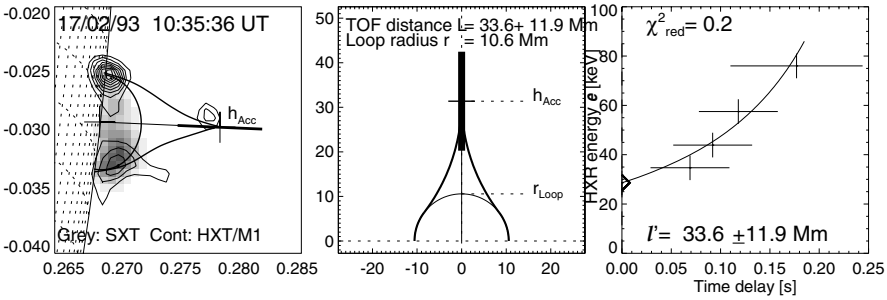


Figure 3. Electron time-of-flight difference measurement between electrons that produce 20–80 keV hard x-ray emission (right panel). From this time difference and a soft x-ray image that shows the loop footpoint separation (left panel), the height of the acceleration site can be reconstructed (middle panel), which is here found at an altitude of 30 000 km.

streams propagating in upward and downward directions, both starting in a common region that has an electron density of $n_e = 10^9$–10^{10} cm s^{-1}, and is thus consistent with the electron densities measured in above-the-loop-top locations. Given this overwhelming evidence, theoreticians can now work with more specific physical parameters to model the acceleration process itself. The above-the-loop-top location is likely to be located in the downward reconnection outflow from an x-point reconnection process, where slow and fast shocks occur and plasma turbulence is created, effects that faciliate both shock and stochastic acceleration. Some new observations with Yohkoh soft x-ray telescope (SXT) and the Nobeyama Radio Heliograph reveal also sometimes large-scale secondary flare loops that have one footpoint in common with the compact soft x-ray flare loop (with hard x-ray emission at the footpoints), suggesting that more complicated loop–loop interactions take place at the acceleration site.

Particle propagation

Once particles are accelerated to nonthermal energies, the ambient plasma becomes collisionless and they can propagate freely along magnetic field lines. The Lorentz force caused by the magnetic field and the particle velocity forces charged particles to spiral along the field lines. The magnetic topology of flares can often be visualized from soft x-ray images, where the heated plasma traces out the magnetic field lines. From such soft x-ray images, e.g. obtained from Yohkoh SXT, it becomes apparent that magnetic field lines connect the acceleration site with flare loop footpoints and thus reveal to us the likely propagation path of accelerated particles. Above the acceleration site, it is likely that the magnetic field lines are open ended and connect with interplanetary space, or close over large-scale arches back to the solar surface at a remote location from the flare site.

The propagation of downward-ejected electrons from the acceleration site has recently been studied by means of time-of-flight differences. Electrons with a kinetic energy of 25 keV have a relativistic velocity of $v/c = 0.3$, while electrons with an energy of 200 keV have a relativistic speed of $v/c = 0.7$. This velocity difference makes the high-energy electrons arrive earlier at the CHROMOSPHERE than the low-energy electrons, if they start simultaneously at the coronal acceleration site. Because the acceleration site has typically a height of 20 000–50 000 km above the solar surface, the resulting time-of-flight differences amount to 100–200 ms. This timing can be measured from energy-dependent time delays of hard x-ray emission, which is produced when the precipitating electrons enter the chromosphere and lose their energy by collisional bremsstrahlung. This time-of-flight method represents therefore an important tool to localize the acceleration site and to probe particle propagation. An example of such a measurement is shown in figure 3.

Particle trapping

Particles injected from the acceleration site onto a magnetic field line can have different velocities and different pitch angles. Because the magnetic field generally increases towards the chromosphere, propagating particles see a converging magnetic field when they approach the flare loop footpoints. Because of the conservation of the

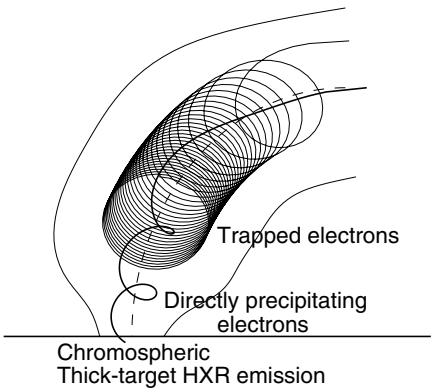


Figure 4. Electrons with small pitch angles precipitate directly, while electrons with large pitch angles become trapped and precipitate after many bounces and eventual scattering into the loss cone.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

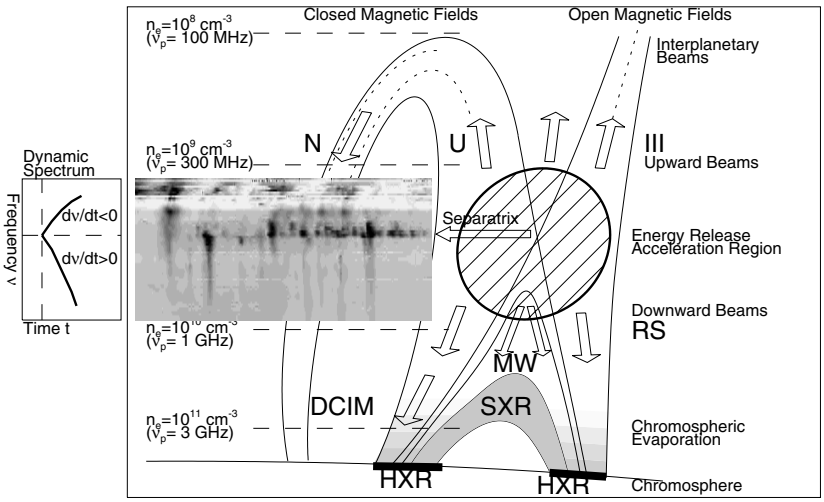


Figure 5. A schematic flare scenario that indicates various trajectories of energetic particles and their possible radiation signatures: radio emission (type III, RS, U, N, decimetric (DCIM) bursts), microwave emission (MW) and hard x-ray emission (HXR).

magnetic moment, the perpendicular velocity or the pitch angle to the magnetic field line increases in response to the increasing magnetic field strength (figure 4). A consequence of this magnetic convergence is that particles bounce back at the bottleneck above the footpoints, called mirror points. This phenomenon of magnetic mirroring is also observed in other converging magnetic fields, e.g. in the Earth's polar cusp magnetic field or in tokomaks. In solar flare loops, mainly electrons with high initial pitch angles become trapped this way and are forced to bounce forth and back between the opposite flare loop footpoints until they suffer a sufficient number of collisions that their pitch angle becomes parallel, which allows the electrons eventually to precipitate. This collisional deflection time τ increases with higher kinetic energy E, according to $\tau \propto E^{3/2}$. This energy dependence of the trapping time was recently confirmed from detailed measurements of hard x-ray delays with Compton GRO.

Hard x-ray and radioemission

Our diagnostics of particle kinematics in solar flares is almost entirely based on their radiative signatures we observe at radio or hard x-ray wavelengths. As long as particles propagate in the collisionless plasma (as is the case in the corona for >30 keV electrons), no hard x-rays are produced. However, when they precipitate towards the dense chromospheric layers, their velocity is slowed down by Coulomb collisions, in which hard x-ray photons are produced by bremsstrahlung. This process illuminates the two footpoints of a flare loop. Flare models that simulate the bombardment of the chromosphere by energetic particles (both electrons and ions) predict a strong local heating of the plasma, resulting in an overpressure that causes the heated plasma to evaporate and to flow upwards into the flare loop. This hot plasma upflow eventually fills the entire flare loop, and thus explains the common appearance of soft x-ray-bright flare loops (marked with SXR in figure 5). This thermal plasma is also responsible for the bulk of hard x-ray emission at energies of 10–30 keV near the loop tops.

Electrons can escape from the acceleration site in upward and downward directions. Because of their anisotropic velocity distribution, propagating electron beams are unstable toward growth of Langmuir waves, which excite plasma oscillations and are converted into electromagnetic waves observable at radio wavelengths. From the frequency–time drift rate of such propagating electron beams, upward (type III bursts) and downward (reverse-slope (RS) bursts) electron beams have been identified in flares. The common start frequency of upward- and downward-injected electron beams is sometimes visible as a separatrix in radio dynamic spectra (figure 5), from which the plasma frequency can be used as a diagnostic of the electron density in the acceleration site. Some upward-moving beams are guided along large-scale arches, which show up in the form of inverted-U bursts in dynamic spectra, or N-shaped bursts, if they are subject to a subsequent mirroring (figure 5). The dominance of closed field lines in active regions explains why a much smaller number of energetic electrons escape into interplanetary space than those that precipitate to the photosphere, as estimated from the hard x-ray flux.

High-energetic (relativistic) electrons, say with energies > 300 keV, emit also gyrosynchrotron emission because of the interaction of their motion with the magnetic field, which is observable in microwaves. Such microwave emission is often seen cospatially with flare loops, with high frequencies closer to the footpoints (because of the higher magnetic field). In summary, most coherent and incoherent radio emission observed in flares is caused in one way or the other by nonthermal electrons and thus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

represents an important diagnostic of particle dynamics in flare plasmas (see also SOLAR FLARES: RADIO BURSTS).

Bibliography

Aschwanden M J 1998 What did Yohkoh and Compton change in our perception of particle acceleration in solar flares? *Observational Plasma Astrophysics: Five Years of Yohkoh and Beyond* T Watanabe *et al* (ed) (Dordrecht: Kluwer) pp 285–94

Benz A O 1993 *Plasma Astrophysics, Kinetic Processes in Solar and Stellar Coronae* (Dordrecht: Kluwer)

M J Aschwanden

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Author
Markus J. Aschwanden
Lockheed-Martin ATC
Solar & Astrophysics Lab.
Dept. H1-12, Bldg.252
3251 Hanover St.
Palo Alto, CA 94304, USA
phone: 650-424-4001
fax: 650-424-3994
e-mail: aschwanden@sag.lmsal.com

Abbreviations
K = Kelvin
keV = kilo electron Volt = 10^3 eV
MeV = Mega electron Volt = 10^6 eV
Yohkoh = Japanese solar satellite (launched in 1991)
SXT = Soft X-ray Telescope on Yohkoh
HXT = Hard X-ray Telescope on Yohkoh
Compton GRO = Compton Gamma Ray Observatory (NASA satellite launched in 1991)
Nobeyama = radio observatory in Nobeyama, Japan

Abbreviations in figure 5:
SXR = soft x-rays
HXR = hard x-rays
MW = microwaves
DCIM = decimetric radio bursts
RS = Reverse-slope bursts
III = type III bursts
U = type U bursts
N = type N bursts

Definitions defined in text
NONTHERMAL PARTICLES - YOHKOH HXT - YOHKOH SXT - COMPTON GRO - TIME-OF-FLIGHT - BI-DIRECTIONAL ELECTRON BEAMS - MAGNETIC MIR-RORING - TRAPPING TIME

Definitions to be defined elsewhere
MAXWELLIAN VELOCITY DISTRIBUTION - COULOMB COLLISIONS - STOCHASTIC ACCELERATION - SHOCK ACCELERATION - MAGNETIC RECONNECTION - RE-CONNECTION OUTFLOW - X-POINT RECONNEC-TION - SLOW SHOCKS - FAST SHOCKS - PLASMA TURBULENCE - NOBEYAMA RADIO HELIOGRAPH - LOOP-LOOP INTERACTIONS - LORENTZ FORCE - MAGNETIC MOMENT - PITCH ANGLE - COLLI-SIONAL DEFLECTION TIME - BREMSSTRAHLUNG - LANGMUIR WAVES - PLASMA OSCILLATIONS - FREQUENCY-TIME DRIFT RATE - TYPE III BURSTS - REVERSE SLOPE BURSTS - INVERTED U-BURSTS - GY-ROSYNCHROTRON EMISSION - COHERENT RADIO EMISSION - INCOHERENT RADIO EMISSION

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Particle Acceleration Mechanisms

Electrons and ions with energies well into the relativistic regime are a universal feature of solar flares (see SOLAR FLARES: NONTHERMAL ELECTRONS and SOLAR WIND: ENERGETIC PARTICLES). The attainment of these energies poses severe constraints on particle acceleration mechanisms from the viewpoint of timescales, energies, total number of accelerated particles, spectra and the energy of different isotopes and charge states.

Electron energies well in excess of 100 keV, and occasionally up to tens of MeV, are inferred. Electrons are accelerated up to 100 keV in <400 ms and to higher energies in a few seconds. The number of electrons required above 20 keV is large—for x-class flares 10^{37} electrons s^{-1} are inferred, although this number is model dependent and suffers from inadequate hard x-ray data. Information on spectra is sparse. One balloon flight in 1980 with germanium detectors obtained high-resolution (1 keV) spectra in hard x-rays. The data were fitted with a combination of power laws and Maxwellians.

Ions (especially protons) are inferred to have energies in excess of 1 MeV up to $\geq$1 GeV per nucleon. The $\approx$ 10 MeV ions are accelerated on the same timescale as the electrons, with the higher-energy ions being delayed by perhaps 10 s. The total number of accelerated ions is unclear. Traditionally the total energy associated with the accelerated ions was assumed to be much smaller than contained in the electrons, but recent evidence suggests that the energetic ion and electron energy contents may be similar in some flares. The γ-ray spectra are usually power laws. Another important result is the massive enhancement of energetic $^3_2$He in very impulsive flares. This is discussed in the section SOLAR WIND: ENERGETIC PARTICLES.

Efforts to account for these energetic particles are hampered by very limited observations. The poor spectral resolution in hard x-rays and the lack of information on ions with energies below $\approx$ 2 MeV are particularly debilitating. There is also minimal information on the strength and topology of the coronal magnetic field, whose associated energy must be tapped to produce the energetic particles. Nonetheless, in the past decade hard x-ray imaging, in particular from the YOHKOH spacecraft, has begun to shed light on what processes may be responsible for particle acceleration in solar flares.

A variety of possible mechanisms exist that could produce the energetic particles that are inferred to exist in solar flares. The most prominent among them are acceleration by shock waves, magnetohydrodynamic (MHD) turbulence and direct electric fields.

Shock acceleration

Shock acceleration has been invoked widely to account for energetic particles in the solar wind, at the termination shock and at a host of galactic and extragalactic locations. Two types of shock acceleration exist. In one, referred to as shock drift acceleration, particles move along the surface of the shock, gaining energy from the shock electric field. The mechanism is only effective when the angle between the magnetic field and shock normal approaches 90°, and even then the energy gain and particle flux are very limited. Since one would expect shocks to propagate at arbitrary angles, it is unlikely to be the source of the large number of particles accelerated in flares. However, the process may be very relevant in situations on the Sun where small numbers of accelerated electrons are required, such as type II radio bursts (see SOLAR FLARES: RADIO BURSTS).

The second process, called shock diffusive acceleration, occurs when a particle interacts with so-called scattering centers on either side of the shock. If, in the frame of the shock, the scattering centers approach each other, the particle systematically gains energy. This process is similar to that suggested by Fermi in 1949, except that, in the solar flare context, the scattering centers are low-frequency MHD waves generated by energetic particles streaming away from the shock. The mechanism is the same for electrons and ions, with the important restriction that each species must reach a critical velocity prior to acceleration becoming effective. Ions need to have velocities greater than the Alfvén speed ($V_A = B/(\mu_0\rho)^{1/2}$) while electrons must have velocities exceeding at least $(m_i/m_e)^{1/2}V_A$. This is the so-called injection problem. Numerical modelling has shown that it is easily overcome for ions. For electrons it is unsolved at this time.

Diffusive acceleration has been used to account for energetic interplanetary particles but has been less widely applied to impulsive flare acceleration. There are a number of reasons for this. Doubts have persisted that shocks do not form quickly enough to accelerate particles. These doubts are spurious since shocks form on kinetic timescales ($\ll$1 s). Others have questioned whether a flare gives the plasma a big enough impulse to form a shock. Although a flare may appear as a large explosion, the coronal magnetic field is very resistant to compression, and hence to shock formation. There is little doubt that if shocks do form, and particles can reach energies above which diffusive acceleration is effective, they are capable of generating the required particle energies very quickly. Whether the required flux can be produced is unknown at this time.

One particular location where shocks almost certainly do occur is in the vicinity of sites of magnetic reconnection (see SOLAR FLARE MODELS). However, some of these shocks may not be suitable for acceleration of particles. Figure 1 shows a sketch of a flare energy release region (based on suggestive evidence from the Yohkoh satellite) with a pair of shocks trailing away from the reconnection site. The problem is that the shocks are slow mode, corresponding to the steepened slow mode wave. At such slow shocks, the scattering centers do not approach each other, so that no energy gain results. Fast mode shocks have also been proposed in this geometry, as shown in figure 1. It is argued that the high-speed plasma jet that arises during magnetic reconnection interacts with stationary magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

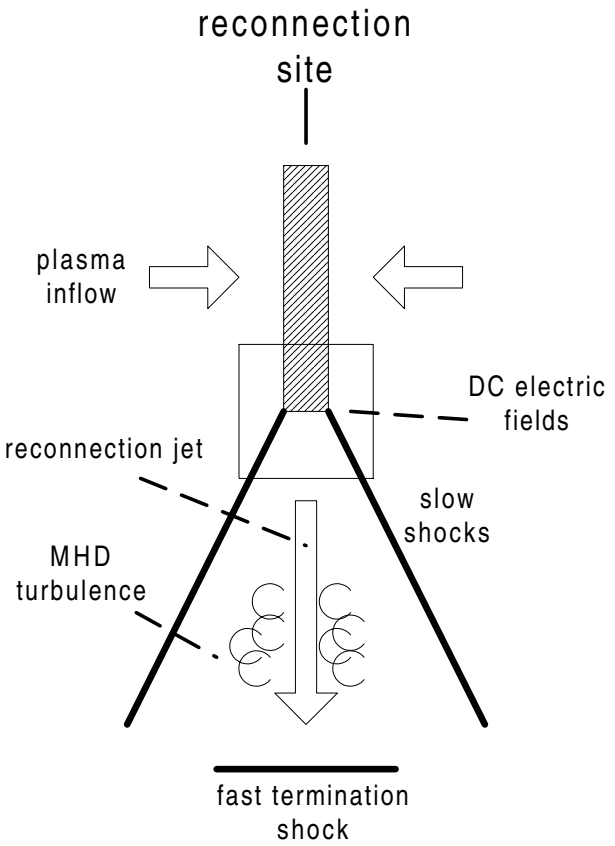


Figure 1. A sketch of a flare energy release region in the solar corona, with possible locations of the particle acceleration mechanisms discussed in this article indicated. The sketch is based on interpretation of hard and soft x-ray data from the Yohkoh satellite. A magnetic reconnection site is assumed to be present, with plasma flowing into it from the left and right sides. The reconnection leads to the generation of a pair of slow mode shock waves, a high-speed plasma jet that may lead to the generation of MHD turbulence and a possible fast mode shock that may terminate the jet. Direct electric fields in the vicinity of the reconnection site itself can also accelerate particles.

fields, resulting in a fast mode termination shock. The role of such a shock in accelerating particles is only beginning to be assessed.

MHD turbulence

MHD turbulent acceleration occurs when particles interact many times with randomly propagating MHD waves. The particles undergo slightly more head-on than overtaking collisions with the waves, so that an energy gain results. The rate of energy gain is slower than for diffusive shock acceleration. MHD turbulence has suffered from the perception that there was an injection problem. In addition, the solution proposed for ion injection at shocks does not work here. The difficulty is that for typical parameters in the solar corona prior to a flare ($T \approx 3 \times 10^6$ K, $n \approx 10^{16}$ m^{-3} and $B \approx 500$ G), a thermal ion or electron cannot resonate with MHD waves. The resolution of this difficulty is similar in concept for ions and electrons, yet different in detail. It should also be noted that all of the models discussed here assume that pitch-angle scattering is efficient enough to maintain an isotropic distribution throughout the time that acceleration is proceeding.

For ion acceleration, the most complete argument proposes that if a fully developed turbulent spectrum of Alfvén waves develops, rather than only long-wavelength waves, thermal ions can interact with waves with frequencies at roughly the ion gyrofrequency. As an ion gains energy, it is continually scattered by longer-wavelength waves, until, when it has a high energy, it resonates with MHD waves. Calculations have been performed indicating that energies of 1 MeV can be produced in 0.1 s, energies of 500 MeV in 2 s for a turbulence level that has magnetic field fluctuations 1% of the ambient field. For the low-energy ions, the majority of the time required to reach these energies is taken up with waves cascading to high frequency.

A similar process of a turbulent cascade, this time of fast mode waves, could operate for electron acceleration, but needs high levels of turbulence. An alternative idea stems from the fact that MHD fast mode waves can accelerate electrons directly by the process of transit time damping. This leads to an increase in the parallel energy of the electron, and so requires ancillary pitch-angle scattering for it to be effective. In fact, for the process to operate, a cascade of energy up the fast mode branch is also required since Coulomb collisions prevent low-energy electrons being energized. Energization of the electrons thus occurs initially through high-frequency waves, with the MHD waves gradually becoming important as the particle energy increases. Initial estimates indicate that the acceleration is prompt (relativistic energies obtained in $\ll 1$ s) and that small volumes (10^{19} m^3) can produce roughly 10^{34} electrons with energies above 20 keV. A flare is then the sum of many of these small energy release regions.

While MHD turbulence is clearly an attractive acceleration mechanism, its possible origin in solar flares is not fully clear. By examining the spectral profiles of coronal emission lines, it is clear that there are significant random motions during flares that may correspond to MHD turbulence. However, these observations have no spatial resolution so that it is unclear where the turbulence is located. One speculation is that the turbulence is the result of an instability in the high-speed plasma flows associated with magnetic reconnection in, or at the top of, flare loops as shown in figure 1. The latter hypothesis is especially attractive given the Yohkoh observations of hard x-ray emission from high in the corona.

Direct electric fields

Direct electric field acceleration may occur in flares in two different forms depending on the strength of the electric field. The key parameter is the field strength that leads to an electron moving at the electron thermal speed gaining energy indefinitely. This is defined as the Dreicer electric field ($E_D = (e/4\pi\epsilon_0)(\omega_{pe}/V_{Te})^2 \ln\Lambda$ V m^{-1}). For typical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

preflare parameters ($T_e = 3 \times 10^6$ K, $n = 10^{16}$ m^{-3}, $\ln \Lambda = 20$), $E_D \approx 10^{-2}$ V m^{-1}. If $E < E_D$, then only the tails of the electron distribution functions are accelerated. If $E > E_D$ then most electrons will gain energy. In the magnetic reconnection model shown in figure 1, $E \gg E_D$, with 100 V m^{-1} being a plausible value, but other scenarios have been proposed with $E < E_D$.

When $E < E_D$ the energy gain that particles attain is limited by the flare geometry, i.e. the length of the potential drop. For a flare loop length of 5×10^7 m, one is limited to 500 keV, lower than the maximum electron energies inferred from γ-ray emission. The ion energies are much lower than required. However, the process may be capable of reproducing hard x-ray spectra below 100 keV. The acceleration time is short, and significant fluxes of electrons can be generated. For $E > E_D$, high energies are readily attainable. Particles of energy 1 MeV can be produced in 10^4 m, a much smaller distance than the resolution of current observations. Ions with energies >1 GeV may pose more significant problems to the mechanism.

The limitation on both mechanisms is the need to ensure overall quasi-neutrality of charge and current in the acceleration regions. Since the inferred number of energetic particles requires that a significant fraction of the coronal population be accelerated, an efficient means of repopulating the acceleration site is required. Also, the accelerated particles correspond to a significant electric current, that in turn will generate large magnetic fields if the current is unneutralized. The resolution of these problems is straightforward for acceleration by shocks and MHD turbulence and is accomplished by a return current of low-energy thermal particles travelling from chromosphere to corona. The return current is oppositely directed to the current due to the accelerated particles, so that only weak magnetic fields are generated. The low-energy electrons also permit overall neutrality to be maintained in the acceleration region. Here the return current flows unimpeded, since no direct electric field acts to oppose it. However, when a direct electric field is present, it will accelerate particles in one direction, and so a return current cannot flow cospatially in the same direction. The problem is especially acute when $E < E_D$ since long potential drops are required. The resolution is to propose that the acceleration site is composed of millions of oppositely directed electric fields such that in an average sense there is approximately no net current or charge deficit. The problem is less severe when $E \gg E_D$ provided that the particles can escape from the electric field before a large current builds up, but precise calculations have yet to be performed.

The future

In recent years, interpretation of satellite data and theoretical modelling have led to major advances in our understanding of particle acceleration mechanisms in flares. While the models still do not address particle acceleration in the global flare context in an adequate manner, the understanding of the plasma physics of the acceleration mechanisms has improved greatly for all of the competing mechanisms. While it is to be expected that progress will be made on the theoretical understanding of mechanisms in the next few years, the major advance in this field is likely to come from the high spectral and spatial resolution at hard x-ray and γ-ray wavelengths made possible by the instruments to be launched on the NASA High Energy Solar Spectroscopic Imager (HESSI) spacecraft in mid-2000.

Bibliography

Cargill P J, Simnett G M and Emslie A G 1996 Do protons or electrons dominate particle acceleration in solar flares? *EOS Trans. Am. Geophys. Union* **77** (37) 353–7

Heyvaerts J 1981 Particle acceleration in solar flares *Solar Flare Magnetohydrodynamics* ed E R Priest (New York: Gordon and Breach) pp 429–555

Melrose D B 1992 The Sun as a lab for particle acceleration mechanisms *Particle Acceleration in Cosmic Plasmas* ed G P Zank and T K Gaisser (New York: American Institute of Physics) pp 3–14

Miller J A, Cargill P J, Emslie A G, Holman G D, Dennis B R, LaRosa T N, Winglee R M, Benka S G and Tsuneta S 1997 Critical issues for understanding particle acceleration in impulsive solar flares *J. Geophys. Res.* **102** 14 631–60

Peter J Cargill

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Preflare Phase

Solar flares are believed to be the result of a sudden conversion of a large amount of free magnetic energy, previously stored in the solar CORONA. The release takes place in a matter of minutes to half an hour, and the amount of energy involved can be up to 10^{26} J (10^{33} erg).

It is generally accepted that in the preflare phase this magnetic energy is slowly built up in the corona, in a matter of hours to days. The sudden release may then be triggered by the onset of a MAGNETOHYDRODYNAMIC (MHD) instability, by 'loss of equilibrium', or by other triggers. Solar physicists have for a long time studied the preflare phase to try to understand the mechanism for the supposed build-up of of magnetic energy, and for a clue to the trigger mechanism. The purpose of these studies is to try to develop the tools to identify regions that are likely to flare and some day to be able to predict the location, time, type and intensity of flares. Moreover, a good physical understanding of the preflare build-up and the trigger mechanism is at the heart of understanding the physics of solar flares as a whole.

Multiwavelength observations of the preflare regions are clearly needed for studying the preflare phase. These observations suffer from the fundamental difficulty that the defining event—the flare—takes place at the end of the observing sequence. A preflare observing campaign therefore often consists of the identification of a 'promising' solar region and then leaving the participating instruments and observatories in a 'sit and stare' mode for as long as it takes. The day–night cycle for ground-based observatories and for earth-orbiting spacecraft very often frustrates this endeavor. Experiments on the recent SOHO mission (which orbits the Sun and therefore has an uninterrupted view) have been quite successful in following regions with a flare promise for more than a day. Future observations from a fleet of spacecraft, in solar orbit and positioned at different angles from the Sun (the 'Stereo' mission) will further eliminate the two week limit of visibility of any promising region, due to the rotation of the Sun itself, and further allow a three-dimensional view of the structures involved.

In the following section the observational results to date on flare precursors will be reviewed. After that the observational signatures for the magnetic energy build-up and changes in the magnetic field structure in preflare regions will be summarized. The article concludes with a description of the basic requirements for theoretical models of preflare energy build-up and the triggering of a flare.

Flare precursors

The categorization of precursors by Gaizauskas (1989) will be adhered to here. His definition of a precursor is a transient event preceding the impulsive phase, possibly even before the onset, and not necessarily at the site of the flare itself (figure 1). Sometimes the frequency of occurrence of minor flares reaches a maximum just before a major flare. In that case the minor flares may qualify as precursors.

Homologous flares

Homologous flares are earlier flares in the same location with similar emission patterns. They occur most often in periods of frequent flare activity. The rate of repetition in a homologous series ranges from a few per hour to several days. There is no consistent relationship between repetition rate and brightness in x-ray, optical or microwave emissions. It is expected that the same flare-producing stress acts on the regions for hours and days, and indeed it has been found that homology is clearly correlated with the development of magnetic shear in an active region, but strict homology is never realized because of the ongoing evolution of small-scale magnetic fields.

Sympathetic flares

Sympathetic flares are earlier flares in different locations, but erupting in near synchronism. Soft x-ray images of the solar corona clearly demonstrate the existence of multiple links between adjacent, and even remote, active regions. Along these links a single excitation could, in principle, be transmitted to other sites and trigger flares. Statistical tests of large samples of flares have been inconclusive, but the analysis of individual cases has shown unequivocally that one flare can trigger another.

Soft x-ray precursors

Soft x-ray precursors are transient enhancements in soft x-rays, lasting for several minutes, that occur in loops or unresolved kernels at, or close to, flare sites. Their shape and location may coincide with those of the subsequent impulsive phase, but that coincidence is usually only partial. Weak soft x-ray bursts are often observed at the time of the projected onset of a SOLAR CORONAL MASS EJECTION (CME) associated with a flare. This may be tens of minutes prior to the impulsive phase of the flare, and the location is usually at one foot of a large pre-existing coronal arch. The rest of the arch brightens weakly within minutes, and the outward motion of the x-ray emitting plasma indicates that a CME has been launched. One interpretation of this sequence of events is that a small magnetic structure interacts with the large coronal arch at one of its footpoints (seen as the precursor), destabilizing the whole structure. This global instability then causes the launch of the CME and the onset of a flare at one of its footpoints.

Radio precursors

Radio precursors, most often observed in microwaves, consist of changes in intensity and/or polarization of radiowaves emitted from an active region, tens of minutes before the onset of a flare. Statistical studies indicate that radio precursors do not occur for the majority of flares, although the correlation between radio bursts and flares is significant. Also, preimpulsive bursts may have very narrow spectral bands, so that most are missed by single-frequency radio telescopes. Individual

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

radio precursors observed at high resolution show morphological transitions of the emitting structures for tens of minutes through the preflare and impulsive phase, while there are also dramatic changes in polarization in centimeter wavelengths from minutes up to an hour before the flare, continuing through the impulsive phase. These results indicate a preflare reorganization of the magnetic field, but the exact nature of this reorganization (flux emergence, reconnection, eruption, etc) cannot be determined from the radio data alone.

Ultraviolet precursors

Ultraviolet (UV) precursors are mostly small-scale transient brightenings above active regions, and they exhibit a broad range of amplitudes. Some of the UV kernels coincide with the later flare, but in other cases the flare may be concentrated in nearby new kernels. Repeated transient UV brightenings have been observed in association with the upwelling, twisting and disruption of an active region prominence just prior to a two-ribbon flare. However, there is not enough statistical evidence to reliably identify a UV flare precursor against the varying UV background.

Surging arches

A surging arch is a transient absorbing feature visible at wavelengths displaced from the central core of Hα. Simultaneous red- and blueshifted components are also visible from the first appearance of the arch, but they are not cospatial. The arch is initially straight in shape (seen from above), expands rapidly and unravels in multiple strands by the time the associated flare erupts in the same active region. In a statistical study of 58 flares it was found that just over half were preceded by a surging arch.

The common inference from the variety of flare precursors described above is that up to tens of minutes prior to the impulsive phase of the flare coronal plasma is heated and set in motion by a rearrangement of the coronal magnetic field. This rearrangement accelerates from gradual to explosive at the onset of the impulsive phase of the flare.

Prominence eruptions and coronal mass ejections

SOLAR PROMINENCE ERUPTIONS very often, but not always, precede two-ribbon flares, the most energetic and long-lived flare type. CMEs are also often associated with a two-ribbon flare. In fact, the extrapolated launch of the CME precedes the prominence eruption and flare in most events, but this launch is difficult to observe. The time delay between the onset of the prominence eruption and the impulsive phase is of the order of minutes, but activity such as enhanced mass motions inside the prominence, a slow rise of the prominence and untwisting can precede the main flare by hours. Enhanced mass motions in active region prominences are the earliest sign of an impending flare and form one of the best flare predictors.

However, not all flares, not even all large ones, have associated prominence eruptions, and not all prominence eruptions lead to flares. While some observers claim there are indications for the eruption of a magnetic flux tube in all flares, other soft x-ray observers claim there is ample evidence for a class of flares that involves the interaction of two or more magnetic loops without an eruption. What seems clear is that at least for two-ribbon flares the prominence eruption is direct evidence for a preflare rearrangement of coronal magnetic field, setting the stage for the sudden energy release that causes the flare emissions.

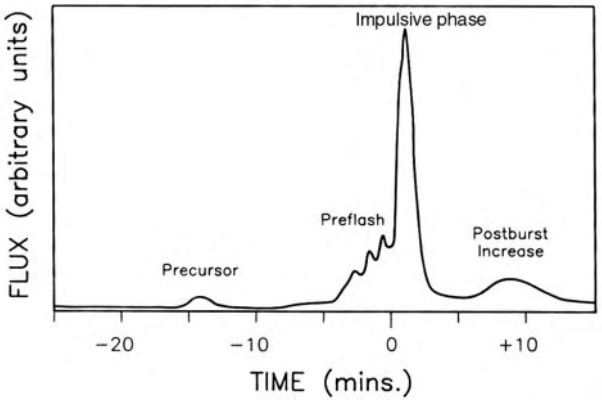


Figure 1. Idealized flux profile of a flare detected in an energy band that senses a distinct impulsive phase (after Gaizauskas 1989).

Magnetic field structure and evolution

The precursors described above are mostly indirect indicators of changes in the coronal magnetic field prior to a flare. A direct signature of these changes comes from observation of the magnetic field in the corona and photosphere. Because of the high coronal temperature only infrared lines are really suitable to measure the magnetic field in the corona, and those have not been explored sufficiently yet. Observations in soft x-rays and the UV show CORONAL LOOPS, which outline the magnetic field morphology. Hα observations show cool matter in filaments or arches that is either streaming along the field or resting at the bottom of slings in the magnetic field, thus yielding additional clues to the magnetic field structure.

Vector magnetic field measurements

More quantitative information can be extracted from the photospheric magnetic field. Here both the line-of-sight and the transverse magnetic field component can be measured in a usually thin layer where the specific line is formed (see the article on SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY). The three components of the magnetic field then can be deduced after removing the sign ambiguity in the transverse magnetic field component, which can be done in several ways, all based upon plausibility arguments. These vector magnetic field measurements have been carried out since the early 1980s at several ground-based solar observatories, and their resolution has increased with newer equipment becoming available.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

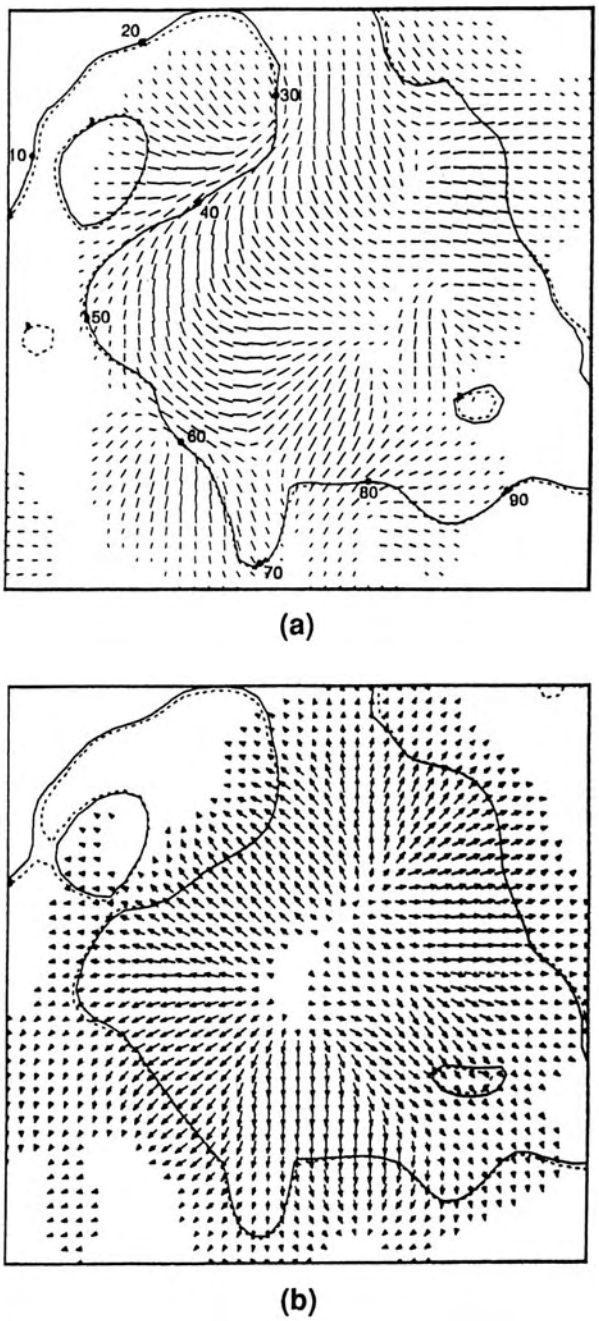


Figure 2. The transverse magnetic field of AR 4711 (*a*) measured by the Marshall vector magnetograph on 3 February 1986, at 16:42 UT, compared with the potential field direction (*b*) calculated from a line-of-sight magnetogram. The region of the largest shear along the polarity inversion line, between points 30 and 50, will be the location of a flare about 4 h later. The field of view is 110 by 110 arcsec, and north is up (from Hagyard 1990).

The magnetic field in the corona can be calculated from the vector magnetic field measured at the lower coronal boundary if one assumes a force-free state for the coronal magnetic field. This assumption is very sensible for two reasons: (1) the observed velocities in the corona are very low compared with the Alfvén velocity, and hence the field is quasi-static, and (2) the gas pressure in the corona is much lower than the inferred magnetic pressure. Therefore the Lorentz force cannot be balanced by either gas or flow pressure gradients over large areas, and the magnetic field must be close to force free. Practical methods have been devised to calculate a force-free field with known vector magnetic boundary conditions, although neither the existence nor the uniqueness of the solutions is guaranteed.

Coronal magnetic fields calculated from the measurements of vector magnetographs show satisfactory agreement with the above-mentioned observed coronal field morphology. Observations of flare sites prior to the impulsive phase have confirmed the scenario of a gradual build-up of free magnetic energy, and yielded several additional clues to the cause of the sudden release of that energy. Hagyard (1990) emphasizes the importance of angular shear, i.e. the angle between the observed transverse field and the potential field that can be derived from the line-of-sight component of the magnetic field alone (figure 2). From the analysis of a sample of only four flares she found that these flares occurred in an area where the local photospheric field departed the most from a potential field. For all flares the magnetic field was strong and the maximum angular shear was at least 85°, while over an extended interval of 8000-10 000 km the shear angle was larger than 80°. It has been claimed that these are sufficient conditions for a flare, but the exact timing for the ensuing flares cannot be determined, so exact flare prediction remains an elusive goal. Clearly, however, vector magnetic field measurements represent the most direct and physical approach towards SOLAR FLARE FORECASTING.

Magnetic field complexity and evolution

Extreme magnetic complexity has long been recognized as very advantageous for producing great flares. Complexity in the magnetic field can be present from the moment the magnetic flux emerges, while in other cases it is fabricated when several sunspot pairs collide and interact.

Numerous flares have been observed to erupt in the vicinity of emerging flux, while the opposite process, flux cancellation, has also been shown to lead to the formation of prominences and to trigger their eruption and the subsequent flare. Large flares often erupt in areas where old and new magnetic flux interacts. An empirical rule here is that prior to a flare the magnetic features of one polarity are increasing, while nearby features of opposite polarity are decreasing.

Chromospheric fibrils (see CHROMOSPHERE: FIBRILS) adjacent to the magnetic polarity inversion line usually stretch nearly parallel to it, indicating a strong magnetic shear in flare-prone regions, in agreement with vector magnetograph observations. Rapid SUNSPOT motions and high inferred velocity shear (i.e. a rapid variation of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

P C Martens

magnitude of the horizontal velocity in the horizontal direction perpendicular to the velocity) are known to correlate with increased flare activity.

There is no evidence that complexity, flux emergence or cancellation are sufficient conditions for flares, since all can occur without ensuing flares. What the above observational results indicate is the importance of considering the evolution of the magnetic field. Free (i.e. non-potential) magnetic energy can be transferred from the photosphere into the corona during the evolution of the magnetic field, both through shearing motions of existing magnetic fields and through emergence of new current-carrying field. At the same time the cancellation of photospheric flux elements can reconnect coronal magnetic loops at their footpoints, changing the connectivity of the coronal field and destabilizing it.

Modeling the preflare phase

The task for the theorist in modeling the preflare phase that emerges from the description of the observational results above is clear. In the preflare phase the magnetic field evolves quasi-statically through a series of force-free states in response to slow changes in the photospheric boundary field. Because of the low electrical resistivity in the corona ideal MHD may be assumed, so no large scale reconnection takes place and the field line connectivity is maintained. The footpoints of field lines are moved around by the photospheric motions, and the coronal magnetic field adjusts by evolving through adjacent force-free states with identical connectivity. New flux elements may be introduced to account for the process of flux emergence. Flux cancellation may represent a deviation from ideal MHD at the photospheric boundary and can be modeled as a boundary change in the fieldline connectivity and magnetic topology.

The simplest model for flare energy build-up therefore consists of calculating a series of topologically connected force-free equilibria that corresponds to a physically meaningful series of photospheric boundary conditions (e.g. continuous motions, finite velocities). At some point in the series the magnetic field may become linearly unstable, and it may be assumed that at that point a dynamical evolution on a much more rapid time scale will set in; the onset of the flare. The same response may occur at a point where no adjacent force-free equilibrium is available, a so-called catastrophe point. A detailed discussion of several of such currently prevailing models is presented in the article on SOLAR FLARE MODELS.

Bibliography

Gaizauskas V 1989 Preflare activity *Sol. Phys.* **121** 135–52

Hagyard M J 1990 The significance of vector magnetic field measurements *Mem. Soc. Astron. Ital.* **61** (2) 337–57

Rust D M, Sakurai T, Gaizauskas V, Hoffman A, Martin S M, Priest E R and Wang J 1994 Preflare state *Sol. Phys.* **153** 1–17

Svestka Z, Jackson, B V and Machado M E (eds) 1991 Preflare situation and flare onset *Eruptive Solar Flares (Lecture Notes in Physics 399)* (Berlin: Springer) ch 1

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Radio Bursts

Energy release in solar flares heats plasma and accelerates electrons and ions to high energies. Radio bursts of various types are produced by energetic electrons interacting with the ambient plasma or with the magnetic field. They may be divided into two broad classes according to frequency range and the dominant emission mechanisms. (1) Coherent plasma radiation plays a dominant role for burst phenomena that occur at meter and decimeter wavelengths (frequencies $\nu \lesssim 3$ GHz). Plasma radiation results from the nonlinear conversion of electron energy to plasma waves and from thence to electromagnetic waves at a frequency near the local plasma frequency or its harmonic (see SOLAR SPECTROSCOPY: CONTINUUM RADIO EMISSION AND DIAGNOSTICS). Plasma radiation is rarely a factor at higher frequencies, where it experiences strong free–free absorption. (2) Incoherent gyrosynchrotron radiation dominates burst activity at centimeter and millimeter wavelengths ($\nu \gtrsim 3$ GHz). Gyrosynchrotron radiation is due to electrons with energies of tens of keV to a few MeV gyrating in a magnetic field at the local electron gyrofrequency. Thermal or nonthermal distributions of such electrons emit a broadband continuum of radiation. Gyrosynchrotron radiation is relatively unimportant at frequencies $\lesssim$1 GHz because the radiation is strongly self-absorbed and/or masked by free–free absorption in overlying plasma.

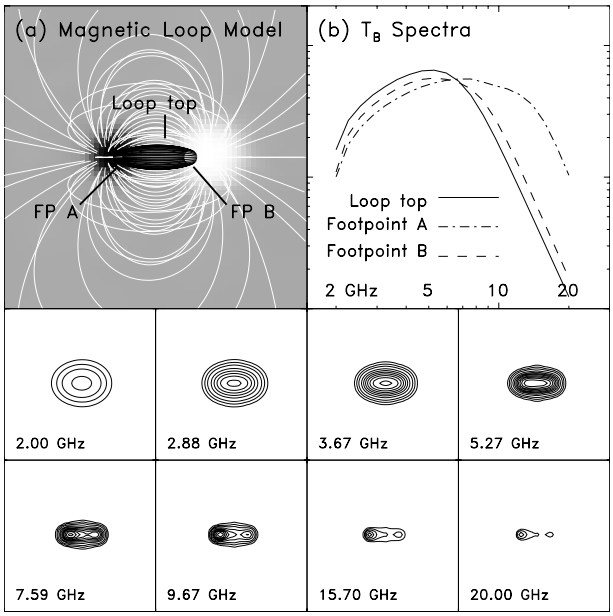


Figure 1. A schematic model of gyrosynchrotron emission from a single asymmetric magnetic loop. Top left: the magnetic loop model. Top right: brightness temperature spectra at the two footpoints and the loop top. Bottom panels: the brightness distribution in total intensity at the frequencies indicated.

Bursts at meter and decimeter wavelengths

At meter wavelengths, radio bursts of types II, III and IV accompany flares (see SOLAR FLARES: RADIO BURSTS). Type II radio bursts, the result of plasma radiation associated with a magnetohydrodynamic shock propagating through the corona, are well correlated with flares. More than 90% of type II bursts have an associated flare. Not all flares produce type II bursts, however. They are rarely seen in association with flares with an Hα importance less than 2 but they accompany 30% of flares with an Hα importance of 2 and 3. However, because small flares are more common than large flares, type II radio bursts are nevertheless more commonly associated with small flares. Type IV bursts are a broadband continuum usually seen in association with type II bursts. Both type II and type IV bursts have a close association with coronal mass ejections (CMEs). Roughly 70% of all type II bursts are associated with a CME. Whether coronal type II and/or type IV radio bursts are produced by CMEs or by the blast wave from an associated flare remains controversial (see SOLAR FLARES: RELATION TO CORONAL MASS EJECTIONS).

Type III bursts occur whenever active regions are present on the Sun. Hence most occur during times when flares are absent. Type III bursts are due to beams of suprathermal electrons propagating upward in the corona. Dynamic spectra of type III bursts are characterized by frequency drifts from high to low frequencies as the beam excites plasma waves at lower densities with increasing height in the corona. More interesting as a flare phenomenon are the decimeter wavelength type III and type-III-like bursts. Type IIIdm bursts have durations and drift rates that follow the same frequency dependence as meter wavelength type III bursts. Other than the fact that type IIIdm bursts occur lower in the corona in a higher density environment ($n_e \sim 10^{15}$–10^{16} m^{-3}), an important difference between meter wavelength type III and type IIIdm bursts is in the sense of frequency drift. The bulk (∼90%) of type IIIdm bursts drift from low to high frequencies, implying that they are excited by downward-directed electron beams. Roughly half of these reverse-slope type IIIdm bursts are correlated with impulsive hard x-ray (HXR) emission. Studies of strong flares show that a detailed correspondence exists between type III, reverse-slope type IIIdm, or type III–reverse-slope type IIIdm pairs and discrete HXR pulses in flares. This is taken to be strong evidence that bidirectional beams of electrons are produced during the impulsive phase of flares. The upward directed beams can produce decimeter and meter wavelength type III bursts (or type J and type U bursts in large-scale, closed magnetic loops) while downward-directed beams can produce type IIIdm and/or HXR pulses via thick-target bremsstrahlung emission in the chromosphere.

Bursts at centimeter and millimeter wavelengths

Bursts at centimeter wavelength accompany flares of all sizes. Electrons are commonly accelerated to energies of several ×100 keV during the impulsive phase on a timescale of seconds. These electrons are injected

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

into coronal magnetic loops where they emit incoherent gyrosynchrotron emission. The spectral maximum of gyrosynchrotron emission is typically in the range of 5–10 GHz ($\lambda \approx$ 3–6 cm). The low-frequency ($\nu \lesssim$ 5 GHz) source is therefore optically thick while the high-frequency ($\nu \gtrsim$ 10 GHz) source is optically thin. The morphology of a radio burst depends on the details of the energy distribution of energetic electrons, on the magnetic field configuration and on the ambient plasma. At low frequencies the entire radio source, defined by those coronal magnetic loops to which energetic electrons have access, is typically optically thick. The optical depth of the emission strongly depends on the magnetic field strength. Since the footpoints of a given coronal magnetic loop are more strongly magnetized than the loop top, footpoint emission dominates in optically thin sources (figure 1).

Radio bursts are highly dynamic. Time sequences of images of the evolving radio source show that, as the number of energetic electrons in the source increases with time, parts of the source change from being optically thin to optically thick. The radio source morphology can also vary as the magnetic connectivity between coronal loops changes in time. As a consequence of optical depth effects and/or changes in the magnetic topology, the morphology, flux and polarization of radio sources can change radically as a function of time and frequency. Figure 2 shows an example of an observation made by the Very Large Array at 4.9 GHz (λ = 6.1 cm) compared with the underlying chromospheric Hα emission. The Hα ribbons of emission demarcate the footpoints of an arcade of magnetic loops. Early in the event, footpoint emission is seen. Later, the Hα ribbons are bridged, the radio source becomes optically thick and the source maximum lies between the magnetic footpoints.

Flares are commonly accompanied by millimeter wavelength emission. Radiation at millimeter wavelengths is of particular interest because, when due to gyrosynchrotron emission, it is produced by electrons with energies of a few MeV. Hence, millimeter wavelength observations offer access to some of the most energetic electrons accelerated in flares, those which are also responsible for electron bremsstrahlung γ-ray emission (see SOLAR FLARES: GAMMA RAYS). Fixed-frequency observations of the spatially integrated flux from flares with polarimeters show that many have an apparent spectral flattening at millimeter wavelengths. Most impulsive flares show evidence for a post-burst increase in flux due to optically thin, thermal free–free emission from warm plasma ablated from the chromosphere by nonthermal electrons. This may account for the spectral flattening in some bursts. Other flares, however, clearly involve a distinct population of nonthermal electrons at high energies, separate from that producing centimeter wavelength and HXR radiation, and due, perhaps, to an additional acceleration mechanism or mechanisms. Spectral hardening at high energies has also been noted in photon spectra observed by spectrometers on board the Solar Maximum Mission, Hinotori and GRANAT satellites. Joint millimeter wavelength–γ-ray observations have confirmed that at least some flares clearly show the presence of a distinct, hard, nonthermal component of energetic electrons.

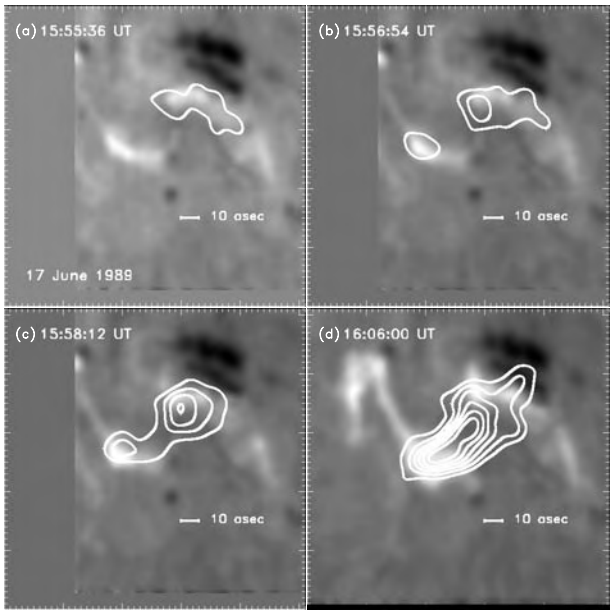


Figure 2. Example of the time evolution of a flaring source observed by the Very Large Array at 4.9 GHz (λ = 6.1 cm; contours) on 17 June 1989. The grayscale shows the underlying Hα emission.

Relation between radio and x-ray bursts

The fact that time profiles of centimeter wavelength and HXR emission are quite similar during flares was recognized long ago, as was the correlation between flux levels in the two emissions. The close correlation between centimeter wavelength and HXR emission during flares has often been cited as evidence that the 'same' population of energetic electrons is responsible for both types of emission. Comparisons of HXR and centimeter wavelength bursts have established that this is essentially the case to the extent that the HXR- and centimeter-wavelength-emitting electrons are drawn from the same nonthermal distribution of electrons. Centimeter-wavelength-emitting electrons tend to be of higher energy than the electrons responsible for 30–100 keV HXR emission. For impulsive flares, electrons with energies of order a few $\times$100 keV gyrating in magnetic fields of 300–900 G are typically responsible for centimeter wavelength emission. In long-duration events, higher-energy electrons ($\sim$1 MeV) gyrating in weaker magnetic fields ($\leq$100 G) are responsible for the observed centimeter wavelength emission (see SOLAR FLARE CLASSIFICATION).

While there is often a close correlation between HXR and centimeter wavelength emission flux and time variation, they differ in detail. Timing comparisons made with spatially unresolved HXR counting rates and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

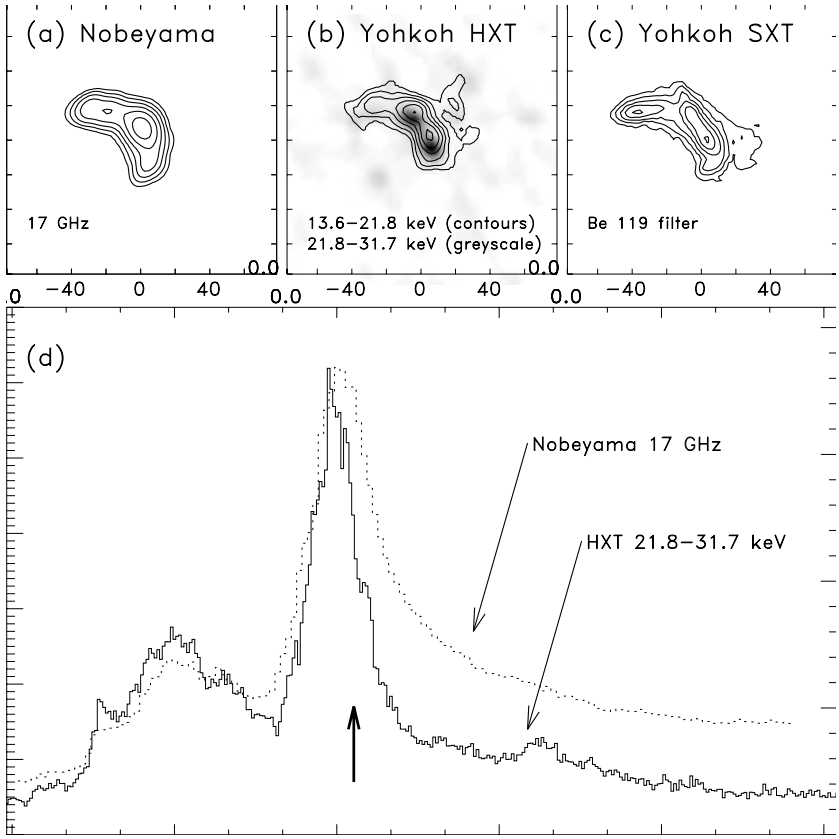


Figure 3. Observations of a flare by the Nobeyama radioheliograph and Yohkoh. (*a*) A snapshot map of the 17 GHz source. (*b*) Snapshot maps of the 13.9–22.7 keV (contours) and the 22.7–32.7 keV (grayscale) emission. (*c*) A snapshot of the SXR emission. (*d*) Time variation of the total 17 GHz flux and the 22.7–32.7 keV HXR counts. The time of the images shown in (*a*), (*b*) and (*c*) is indicated by a vertical arrow. Note the delay of the 17 GHz maximum relative to that in HXRs and the more extended decay.

radio fluxes find that temporal features in centimeter wavelength emission (e.g. the time of maximum flux) lag those in HXRs by up to a few seconds. Furthermore, the centimeter wavelength emission typically decays more slowly than the HXR count rate (figure 3). While transport effects might be relevant to the smallest delays observed, electron trapping may account for the full range of delays observed. Electron trapping is a result of electrons being injected into a magnetic field topology that converges near the base of the corona. Only those electrons with trajectories at sufficiently small angles relative to the local magnetic field vector (the pitch angle) can escape, or precipitate, from the trap. The cone of pitch angles that enable electrons to escape from the trap is called the loss cone. Various mechanisms can scatter electrons into the loss cone, including collisions of the energetic electrons with background electrons and ions (Coulomb collisions) or scattering on a spectrum of waves (e.g. whistler-mode waves). In the case of Coulomb collisions, high-energy electrons have a longer lifetime against scattering into the loss cone than low-energy electrons and therefore remain in the magnetic trap for a longer time. Consequently the radiation they emit peaks later and decays more slowly than emission from lower-energy electrons. Another possible reason for the delay between centimeter wavelength and HXR emission is that higher-energy electrons are simply accelerated somewhat later than lower-energy electrons in a 'second-step' process such as stochastic acceleration.

The spatial relationship between centimeter wavelength and x-ray bursts has recently been clarified. Soft x-ray (SXR) emission demarcates coronal magnetic loops containing hot, dense, thermal plasma. The dominant sources of HXR emission are in conjugate magnetic footpoints in the upper chromosphere and low corona where nonthermal electrons impact dense, relatively cool material and produce thick-target bremsstrahlung radiation. Centimeter wavelength emission traces out the entire volume accessible to nonthermal electrons. Hence, the optically thick emission defines its spatial extent. At frequencies where the centimeter wavelength emission is optically thin, regions characterized by strong magnetic fields dominate, e.g. magnetic footpoints. Recent studies indicate that, in asymmetric magnetic loops, HXR emission is emitted preferentially from the magnetically weaker footpoint, where electrons precipitate from the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

trap in greater numbers, while centimeter wavelength emission is stronger in the magnetically stronger footpoint, leading to a spatial displacement between the dominant HXR and centimeter wavelength sources. Recent studies of centimeter wavelength and HXR source morphologies suggest that impulsive flares commonly involve loop–loop interactions—both a compact and a larger-scale magnetic loop structure interact owing to new flux emerging into pre-existing magnetic structures. Magnetic cusp morphologies also appear to play an important role in flares (see SOLAR WIND: ENERGETIC PARTICLES).

Radio bursts as diagnostics of physical processes

Both low- and high-frequency radio bursts offer diagnostics of a variety of physical parameters and processes during flares. The multitudes of type IIIdm bursts that accompany impulsive flares yield insight into the nature of energy release. With each type-III-like event corresponding to a discrete electron energization event, studies of the temporal and spatial distribution of type IIIdm bursts offer a means of constraining local conditions in the energy release site and the means by which energy release proceeds. Recent work in this area has concentrated on 'avalanche' models of flares, wherein the fragmentary nature of energy release plays a central role.

High-frequency bursts also yield insights into local conditions in the flaring source and into electron acceleration and transport. Unlike HXR emission, gyrosynchrotron emission has a sensitive dependence on the magnetic field strength and orientation in the source. Spatially and temporally resolved spectral observations offer a means of constraining the magnetic field in the flaring source. Such observations also allow the electron distribution function to be measured as a function of space and time. Current research is directed toward making such measurements in order to clarify which, of the many possible, electron acceleration mechanisms operates during flares—quasi-static dc electric fields, weak double layers, stochastic acceleration or shocks—and under what circumstances (see SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS). Similarly, the details of the evolution of the emitted spectrum of radiation shed light on the transport of electrons throughout the flaring volume: the details of electron trapping, electron scattering, electron precipitation and chromospheric ablation.

Bibliography

Bastian T S, Benz A O and Gary D E 1998 *Ann. Rev. Astron. Astrophys.* **36** 131–87

Benz A O 1993, *Plasma Astrophysics* (Dordrecht: Kluwer)

McLean D J and Labrum N R ed 1985 *Solar Radiophysics* (Cambridge: Cambridge University Press)

T S Bastian

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: Relation to Coronal Mass Ejections

SOLAR FLARES and SOLAR CORONAL MASS EJECTIONS have both been associated with geomagnetic activity such as the AURORA, radio interference and 'geomagnetic storms' caused by large currents flowing through the Earth's magnetosphere. They are also both associated with high fluxes of energetic particles ('proton events') in near-Earth space. However, they are different types of event.

Solar flares are sudden, intense increases in brightness of particular solar features. They have been observed for over a century in visible light, and are detectable with radio telescopes. With the advent of space-based observations, flares have been observed throughout the short end of the electromagnetic spectrum, from ultraviolet through to gamma rays. Flares occur when small, intense regions of the solar magnetic field undergo reorganization, suddenly releasing energy into the solar coronal and chromospheric plasma. In the near-Earth space environment, flares cause immediate increases in the flux of ionizing radiation such as highly energetic electrons (β rays), protons and photons (γ rays). They are also correlated with delayed geomagnetic effects at Earth: the aurora and temporary motion of the Earth's magnetic field are more likely to be observed three days after a solar flare than at other times. This delay time corresponds to the transit time of the slow SOLAR WIND that propagates out from the Sun at $\sim$400 km s^{-1}. However, not all flares are associated with aurorae or geomagnetic storms; and some aurorae occur even without a precursor solar flare.

Coronal mass ejections (CMEs) are eruptions of large, massive bodies of plasma from the corona. CMEs can span 120° of solar latitude or longitude, and involve up to 10^{16} grams (10^{10} metric tons). They have been discovered only since the advent of the CORONAGRAPH, and have been studied in detail only since the advent of SKYLAB and other space-based coronagraphs such as the Solar Maximum Mission coronagraph and, more recently, the LASCO coronagraph aboard the SOHO spacecraft. Coronal mass ejections are correlated with flares, but appear to be more closely associated with the lift-off of FILAMENTS from the surface of the Sun. Because only half of the solar surface is visible at any one time, half of all CMEs originate on the far side of the Sun. Approximately half of the observed CMEs can be traced to the disappearance of a filament or prominence on the visible disk of the Sun; these statistics indicate that all CMEs originate from filaments. High resolution ultraviolet image sequences from the Transition Region and Coronal Explorer (TRACE) satellite and from the SOHO/EIT instrument have shown that filament 'lift-off' is central to the formation of CMEs.

The association between flares and CMEs is not coincidental. Coronal mass ejections appear to be caused by large-scale instabilities in magnetic structures (filaments) up to 300 000 km in size, while flares are caused by smaller scale instabilities that cause sudden reconnection of field lines within very small (<1000 km) volumes. Filamentary lift-off, and its associated field reorganization, causes additional local stress in the magnetic field, which may sometimes trigger a flare; and flaring causes field reconfigurations that can trigger large scale instabilities. Which (if either) of these mechanisms is correct is currently unknown.

Craig DeForest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Flares: White Light

Characteristics of white light solar flares

A white light flare is a short-lived increase of the solar continuum emission, with a duration of 1–10 min, that takes place during some solar FLARES. The first white light flare was detected by Carrington in 1859 as a local and short-duration brightening on a white light picture of the Sun. Since spectral observations indicate that the continuum emission enhancement extends towards the UV part of the spectrum, the term 'White light flare' means more generally an enhancement of the continuum spectrum.

There is no world-wide flare patrol organization to record flares in white light such as exists for recording flares in the hydrogen Hα line. Consequently, white light flares appear as events rarer than they presumably are, and only about a hundred of them (fewer than 90 in 1993) have been reported.

The typical appearance of the spatial distribution of the white light emission is one, two or sometimes more bright patches. The white light brightenings are most of the time moving very fast or jumping from point to point. Some observations made with a passband filter centered at 4308 Å displayed a double footpoint character, suggesting that the white light flares observed were located at the base of magnetic flux tubes (see also SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: OBSERVATIONS).

Filtergram measurements, made with broad-band FILTERS, are affected by the enhancement of the intensities and of the number of the atomic lines that occurs during the flare process. This enhancement is particularly high in the blue part of the spectrum. Intensities up to twice the quiet Sun continuum intensity in the visible have then been reported using filters, when the continuum intensity enhancements measured on the spectrum did not exceed 10–20%. Spectral observations of white light flares, when not made with broad-band filters, require the spectrograph slit to be located at the right position and at the right time. Therefore, the wavelength intensity spectra of no more than a few tens of flares have been observed, making the physical origin of white light flares not yet fully understood.

Two extreme typical spectra have been reported in observed white light flare spectra. They differ by the presence or not, near 3646 Å, of a discontinuity of the continuum intensity. The presence of a discontinuity implies that the emission is hydrogen free–bound emission resulting from the recombination between a proton and an electron. At visible wavelengths, the main contribution comes from the Paschen continuum associated with recombination leaving the recombined electron on an energy level described by the main quantum number $n = 3$. Recombination to the level $n = 2$ produces Balmer continuum emission at wavelengths shorter than 3646 Å, introducing a discontinuity at this wavelength. Contrary to electron–proton recombination, recombination of electrons with neutral hydrogen atoms to form an H^- ion or modifications of the electron velocities in the field of this ion do not produce a discontinuity in the UV.

The two kinds of continuum emission, due respectively to hydrogen and H^-, originate from different depths in the solar atmosphere. The Balmer continuum is formed at the chromospheric level, while the H^- emission is formed deeper at the photospheric level where the local hydrogen and electron densities are higher. Therefore, the amplitude of the Balmer discontinuity provides information on the depth of formation of the white light flare emission, i.e. on the atmospheric layers from which the radiated energy originates.

It is widely accepted that solar flares come from the fast dissipation of magnetic energy, and most flare models put the location of this dissipation somewhere in the solar CORONA. This hypothesis has been supported by the possibility of explaining the main properties of chromospheric Hα flares by a transport of energy from the corona down to the lower atmosphere by various processes, such as particle beam propagation, heat conduction or irradiation. The total energy radiated in the excess of continuum emission of a white light flare can reach 10^{30} erg; this is comparable to the energy involved in chromospheric flares. Therefore, flare models must take into account the white light flare phenomenon.

The possibility of explaining white light flares by energy transport from the corona must be checked. This requires the estimation of the depth of formation of the observed continuum emission, in order to put constraints on the depth of penetration of the energy coming from coronal heights, and on the amount of energy that must be deposited in order to compensate for the radiative loss.

Origin of white light flares

Apart from local heating, the origin of white light flares was looked for in the few processes likely to deposit energy into the solar CHROMOSPHERE or high SOLAR PHOTOSPHERE, i.e. heat conduction, particle bombardment and radiative heating. The decrease of the thermal gradient with height suggests that the heat conduction flux is too low to explain white light flares, most of the heat flux being evacuated in UV lines in the transition zone between the corona and the chromosphere. The remaining energy transport processes are particle bombardment and radiative heating.

Electron bombardment is present in most energetic flares. It is detected via the bremsstrahlung hard x-ray emission associated with the change in the orientation of the velocity vector of energetic electrons in the electric field of the ions. During their propagation to the lower atmosphere, electrons also collide with local electrons and lose most of their energy—nearly five orders of magnitude more than in brehmsstrahlung emission—in these Coulomb collisions. The rate of deposition of energy by an electron beam at the maximum of hard x-ray emission can exceed 10^{11} erg cm^{-2} s^{-1} and be one order of magnitude higher than the rate of energy lost in the enhancement of the continuum emission.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The beam electrons lose their energy in inelastic collisions with electrons and in elastic and inelastic collisions with neutral hydrogen. The depth of penetration, expressed by the column mass m crossed by high-energy electrons, moving vertically, is related to their energy E_m by

$$E_m \approx \left(\frac{K' 2\pi e^4}{m_{\rm H}} m \right)^{1/2}$$

where $m_{\rm H}$ is the hydrogen atom mass and K' is a function of the target plasma degree of ionization and of various parameters describing the effect of collisions on the particle energy. Entering the numerical values of the constants, and expressing m in g cm^{-2} and E_m in keV, leads to $E_m \approx 1.25 \times 10^3 \; m^{1/2}$. Collisions with local electrons heat the upper part of the chromosphere to coronal temperatures and shift down the region of minimum temperature and the chromosphere. The formation of the hot 10^7 K coronal plasma is due to the lower part of the electron energy spectrum. For beams carrying a high energy flux, the electrons that penetrate to layers dense enough for their temperature to stay, after heating, below 10^4 K are those of energy higher than about 50 keV. Only electrons of about 120 keV penetrate down to the base of the preflare chromospheric plateau, at m equal to 0.01 g cm^{-2}.

Inelastic collisions with neutral hydrogen, by increasing the hydrogen ionization rate by four to six orders of magnitude, increase accordingly the recombination rate and consequently lead, in the heated and deeper chromosphere, to an enhancement of both the bound–free Paschen and Balmer continua. Prediction of the resulting continuum emission requires us to compute the depth variation of temperature and density that results from beam bombardment.

Depending on the amplitude of the particle energy flux, chromospheric condensations may or may not be generated. They are formed in explosive chromospheric heating, i.e. when the heating rate of the chromospheric plasma is beyond its ability to radiate, leading to a very rapid temperature increase (see also CHROMOSPHERE: HEATING MECHANISM). If the time scale for this temperature change is shorter than the hydrodynamic expansion time, a strong pressure increase follows that drives matter both upward and downward. The compression of the plasma below increases its density. That, in turn, leads to a decrease of its temperature in order to reduce its radiative loss. A chromospheric condensation moving downward is formed, with higher density and lower temperature than the surrounding atmosphere, and such condensation was suggested to be at the origin of white light flares.

Computations of the white light emission of chromospheric condensations have been limited to empirical models, where the temperature and density of the condensation were prescribed. These computations showed that an increase of the white light continuum intensity requires the condensation to be hotter than its surroundings, while dynamical models of the response of the atmosphere to beam bombardment lead to a condensation cooler than its surroundings.

Observations of white light flares during hard x-ray emission in the absence of chromospheric condensation have been reported. In a flaring ACTIVE REGION, the presence of electron bombardment was established by the characteristic shape of the Hα line profiles that exhibited a strong self-reversal of the emission core. The selected locations, where this signature was the strongest, did not show the typical asymmetry due to the depth dependent Doppler redshift, that results from the downward motion of chromospheric condensations. However, they were associated with a strong enhancement of the white light continuum emission. This enhancement was not present at other locations that showed the signatures of chromospheric condensation.

Comparison of the time variation of the white light and hard x-ray flare continuum emission generally shows a good time correlation during the few minutes' duration of the impulsive phase (see SOLAR FLARES: IMPULSIVE PHASE). Since hard x-ray emission is measured without any spatial resolution, comparison is made between the total hard x-ray power of an unresolved hard x-ray source and the white light flare intensity at the location of maximum brightness. The total power deposited into the solar atmosphere by an electron beam, the electron energy flux, and the resulting white light flare intensity, depends on the beam cross section. Therefore, not all hard x-ray power peaks are expected to be associated with white light flare emission.

The overall time correlation between white light and hard x-ray emission varies with the hard x-ray photon energy. On average, a good time correlation is found, that increases with the photon energy. Time delays of the white light flare emission peak relative to the hard x-ray power peak above 50 keV, close to the observation time resolution of about 1 s, were observed. Indeed, this strongly suggests that energetic electrons bombarding the solar atmosphere are at the origin of the observed white light emission.

Impulsive hard x-ray emission lasts only a few minutes and is generally followed by a more gradual rise and fall of the soft x-ray emission coming from the hot plasma generated presumably by electron bombardment. White light flare emission has been reported to continue during this gradual phase and to peak at the time of maximum of the 10 keV soft x-ray emission. Therefore, heating of the lower atmosphere by soft x-ray photons was proposed as the origin of the gradual phase of white light flares.

For an event observed on 1 July 1980 the soft x-ray peak power was within a factor two of the white light power, supporting the hypothesis of soft X-ray irradiation as at the origin of the progressive phase of white light flare emission. In other events, at the time of the peak of white light emission the power radiated over 2π by 1–8 Å soft x-ray photons was reported to be an order of magnitude smaller than the white light flare power. However, the assumption of isotropic emission in soft

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

x-rays, used in all comparisons, is not valid if the x-ray source is optically thin and not spherical. The x-ray intensity is then proportional to the extension of the soft x-ray source along the direction of photon propagation, and a source more extended in height that laterally would irradiate the lower atmosphere more than expected from measurements made in a direction other than vertical.

The wavelength dependence of white light emission provides additional constraints on white light flare models. The number of events for which simultaneous observations of the white light spectrum and of the hard x-ray emission are available is still very limited. However, it has been reported that the wavelength spectrum of events for which a good time correlation between hard x-ray and white light flare emission is observed shows a significant Balmer jump, not observed in events for which there is no such time correlation.

Atmospheric heating by electron beams is not restrained to chromospheric layers and occurs also in the upper photospheric layers. At the higher hydrogen number density of the temperature minimum region, electron bombardment, via an increase in the electron number density, enhances the H^- number density. As a result, atmospheric backwarming by the enhanced chromospheric radiation takes place, increasing the temperature of the upper photosphere. As an example, a beam of electrons with an energy distribution of the energy flux, above a 70 keV low-energy cut-off, given by a power law, carrying an energy flux of about 8×10^{10} erg cm^{-2} s^{-1}, produces a temperature rise of 240 K at the temperature minimum. Consequently, the intensity of the H^- white light continuum emission increases, which adds a significant contribution to the chromospheric hydrogen free–bound emission and reduces the amplitude of the Balmer jump. However, even when reduced in amplitude the Balmer discontinuity is still expected to be present in the spectrum.

In a very few observed spectra, the Balmer discontinuity is not present at all. Therefore, soft x-ray irradiation has been proposed as possibly releasing energy deep enough in the solar atmosphere to generate a dominant H^- continuum emission. Such a model requires that most of the x-ray energy is deposited deep enough in the solar atmosphere, at a column mass higher than 0.1 g cm^{-2}. 10 keV photons satisfy this condition. As reported above, the observational checks of the validity of this hypothesis are still contradictory.

Like electron beams, proton beams deposit energy in the solar atmosphere. The ratio of proton and electron energy needed to reach a given atmospheric depth is the square root of their mass ratio. Consequently, protons must have an energy higher than about 150 keV at the beginning of the flare to interact with the chromosphere, with an energy requirement increasing to the MeV range with the downward shift of the chromosphere that occurs during the flare. Above about 1.5 MeV, protons can be detected via the gamma ray lines they produce (see also SOLAR FLARES: GAMMA RAYS). Below this limit they are not detectable, except by measurements of the linear polarization in chromospheric lines, like the hydrogen Hα line, that they generate by impact.

By non-thermal ionization of the background hydrogen, protons could generate white light continuum emission. The resulting white light flare would be time correlated with the evolution of the hot, soft x-ray emitting thermal plasma resulting from their bombardment. Depending on the proton beam characteristics, the power radiated in the white light flare could be higher than that radiated in soft x-rays. There will be no associated hard x-ray emission.

In conclusion, in the limit of the accuracy of existing measurements, energy transfer from coronal heights down to the low atmosphere, either by energetic particles or by energetic soft x-ray photons, is able to explain the white light flare characteristics without requiring *in situ* energy deposition. However, white light flare characteristics depend on the energy transport processes, which could be either particle bombardment or x-ray irradiation, and on their characteristics. Consequently, observations made from UV to infra-red, in hard and soft x-rays, with high space and time resolution, are still needed in order to progress in our understanding of the white light flares.

Bibliography

Canfield R C *et al* 1986 Impulsive phase transport *Energetic Phenomena on the Sun (NASA Conf. Publ. 2439)* ed M Kundu and B Woodgate pp 26–31

Carrington R C 1859 *Mon. Not. R. Astron. Soc.* **20** 13

Švestka 1976 *Solar Flares* (Dordrecht: Reidel)

J C Hénoux

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior

The interior of the Sun is hidden from our sight, because it is opaque to electromagnetic waves: the radiation we receive from it on Earth is emitted in the outermost layers. Our knowledge of the solar interior is based solely on *theoretical models* which are built with some assumptions about the physical conditions and processes that are likely to prevail there, and on *helioseismology*, a very powerful technique which permits us to explore the interior directly with acoustic waves.

General properties

Until recently, all that we knew about the Sun came from observing it in visible light (see SUN: BASIC PROPERTIES). Its distance and radius $R_\odot$ were determined long ago by triangulation of the solar system, although the precision was increased significantly by using the modern technique of radar echoes. In the early nineteenth century, spectroscopy of the solar light revealed the gaseous nature of the emitting layers and the presence of many chemical elements known on Earth, with the notable exception of helium, which was detected on the Sun before it was discovered on Earth. Later, with the advent of atomic physics, it became possible to decipher the spectral lines, and to assess the temperature and composition of the surface layers (see SOLAR SPECTROSCOPY AND DIAGNOSTICS): hydrogen turned out to be most abundant species, followed by helium (≈25% in mass), the other elements representing only 2% altogether. The mass of the Sun, $M_\odot$, is deduced from the orbital motions of the planets and from precise laboratory measurements of the gravitational attraction between two bodies. Finally, the heat flux received on Earth gives the total amount of thermal energy radiated by the Sun into space, i.e. its luminosity $L_\odot$ (see SOLAR IRRADIANCE).

With a few 'reasonable' assumptions, these global parameters permit us to derive some important properties of the solar interior.

Central temperature

The first of these assumptions is that of *hydrostatic equilibrium*, which states that everywhere in the solar interior there is an exact balance between gravity and the pressure forces. This allows us to evaluate the central pressure, which is the weight of a column of matter of unit area; it is given by $P_c = \overline{\rho}\,\overline{g} R_\odot$ where $\overline{\rho}$ and $\overline{g}$ are suitable averages of the density and the gravity along the radius. As crude estimates, one may take the median value of all the quantities involved:

$$\overline{\rho} \approx \frac{\rho_c}{2} \qquad \overline{g} \approx \mathcal{G}\frac{M_\odot/2}{(R_\odot/2)^2}$$

with ρ_c being the central density and $\mathcal{G}$ the gravitational constant, which leads to

$$P_c \approx \rho_c \frac{\mathcal{G}M_\odot}{R_\odot}.$$

One may further assume that the solar material consists mostly of ionized hydrogen, as suggested by its surface composition, the molecular mass thus being $\mu = 0.5$ g, and that it obeys the perfect gas law $P = \mathcal{R}\rho T/\mu$, with $\mathcal{R}$ being the gas constant. This yields an estimate of the *central temperature* which depends only on the mass and the radius:

$$T_{\text{central}} \approx \frac{\mu}{\mathcal{R}}\frac{\mathcal{G}M_\odot}{R_\odot} \approx 10^7 \text{ K} \qquad (1)$$

a crude figure which is not that far from the value predicted by the best models of the present Sun.

Heat transport

This high temperature, compared with that of the surface layers (see SOLAR PHOTOSPHERE) whose effective temperature is $T_{\text{eff}} \approx 6000$ K, tells us that the thermal energy radiated into space is produced in the deep interior. One way to transport this energy to the surface is by radiation: the photons travel a short distance, are either scattered in another direction, or absorbed and re-emitted, and so on, until they finally reach the outermost layers from which they escape. Their mean-free path is small enough to guarantee that the solar interior is locally in *thermal equilibrium* and thus that the radiation is close to that of the ideal black body, meaning that the energy density only depends on temperature, and that it scales as T^4. This allows us to estimate the *optical thickness* τ of the Sun along its radius

$$\tau \approx \left(\frac{T_{\text{central}}}{T_{\text{eff}}}\right)^4 \approx 10^{13}.$$

This huge figure indicates that the Sun is extremely opaque: on the average, the mean free path of a photon is of order $R_\odot/\tau \approx 10^{-2}$ cm.

If it did not interact so strongly with matter, a photon emitted in the central region would cross the Sun in 2 s, but this very opaque medium forces it into a random walk: instead of τ interactions with matter, it will undergo τ^2 and therefore it will take τ times longer to escape, which amounts to 10^6 years.

But closer inspection of the solar surface shows that it is the seat of turbulent motions which also contribute to the heat transport. Indeed, near the surface, where hydrogen and helium are partly ionized, the opacity is so high that the thermal stratification becomes unstable: a bubble of solar matter which is lifted there from its equilibrium position becomes lighter than its environment, and hence it continues to rise. This *convective instability* occurs whenever the temperature gradient exceeds a critical value, which is the adiabatic gradient (i.e. the gradient that would be experienced by the bubble if it did not exchange heat with the surrounding medium). The instability condition is called the Schwarzschild criterion; it is usually expressed in terms of logarithmic temperature gradients with respect to pressure:

$$\frac{\mathrm{d}\log T}{\mathrm{d}\log P} > \left(\frac{\partial \log T}{\partial \log P}\right)_{\text{adiabatic}} \text{ or in short } \nabla > \nabla_{\text{ad}}.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Detailed models indicate that this criterion is satisfied in the outer third of the Sun (in radius). This turbulent region is the cause of the various phenomena which characterize the solar activity. Unfortunately, the description of the convective motions is still in a very crude state, and it remains one of the weakest points of modeling the solar interior (see SOLAR INTERIOR: CONVECTION ZONE). The inner two-thirds of the Sun are in stable equilibrium, and there the thermal energy is transported by radiation.

The energy source

The source of solar energy long remained a puzzle. All energy sources which were considered in the nineteenth century predicted a life expectancy for the Sun which was much smaller than the estimated age of some rocks on Earth. The problem was solved only in the 1920s, as scientists became aware that nuclear reactions could provide the necessary energy. When hydrogen is transformed into helium, 0.7% of the mass is changed into energy, according to Einstein's formula $E = mc^2$. One can easily estimate the Sun's life expectancy $t_{\rm exp}$ by equating the rate at which energy is emitted with the total amount which is available:

$$t_{\rm exp} = 0.1 \times 0.007 M_\odot c^2 / L_\odot \approx 10^{10} \text{ years}$$

where we have assumed that only 1/10 of the solar material will be changed into energy, namely that which is located in the innermost region, where the temperature is high enough to sustain nuclear reactions. This figure is compatible with the age of the solar system, which is now determined by the precise radiochemistry of its oldest components which can be analysed, namely meteorites.

The specific nuclear reactions which are responsible for the energy release in the Sun were identified later. The main contribution comes from the p–p chain, which starts with the fusion of two protons into a nucleus of deuterium: $\mathrm{p} + \mathrm{p} \rightarrow^2 \mathrm{H} + \mathrm{e}^+ + \nu_e$. This reaction involves the weak interaction and therefore emits a neutrino; since it is very slow, it regulates the whole energy production process. Other reactions follow, until ^{4}He is produced (see SOLAR INTERIOR: ENERGY GENERATION).

The next step of nuclear burning would transform helium into carbon, but with an efficiency ten times less per unit mass processed. This explains why most stars are observed in the hydrogen burning phase, on the so-called main sequence.

The rate of production of nuclear energy in the solar interior is very sensitive to temperature, because the fusion of two nuclei can only occur when their electrostatic repulsion is overcome by their kinetic energy: the higher the temperature, the higher the release of nuclear energy. This has an important consequence: assume that for some reason the production of nuclear energy slightly increases; additional heat is then deposited in the outer layers, which for this reason expand, but through hydrostatic balance that increase in radius leads to a decrease of the central temperature (see equation (1)), which has the effect of reducing the nuclear reaction rate. This self-regulation ensures that main-sequence stars, such as the Sun, are in stable thermal equilibrium.

Modeling the solar interior

A stellar model describes the internal constitution of a star, namely its pressure and temperature stratification, and also its chemical composition, as a function of depth, at a given age (see SOLAR INTERIOR: STANDARD MODELS). Usually such models are based on the assumption of hydrostatic equilibrium, although small departures from it are easily accounted for. Most models also assume spherical symmetry, a simplification which is well justified for the Sun, where the rotation is so slow that the centrifugal force may indeed be neglected compared with gravity.

The conservation laws for mass, momentum and energy yield a differential system which is of fourth order in space and of first order in time. Four boundary conditions are available for the main dependent variables: the obvious $m = 0$ for the mass and $L = 0$ for the luminosity at center $r = 0$, with the pressure P and the temperature T at the surface being borrowed from an atmospheric model. This would be sufficient to build a model if the chemical composition were known at each depth, but it too must be calculated in a consistent way. For that purpose one has to add a set of n differential equations, of first order in time and second order in space, which follow the evolution of n chemical species under the effect of nuclear reactions and/or microscopic diffusion; the associated boundary conditions simply state that the flux of each species vanishes at center and surface. In the first stellar models built in the 1950s, one was able only to track the transformation of hydrogen into helium, but nowadays n is no longer limited by computer performance. Finally one has to provide an equation of state to express the pressure as a function of density, temperature and composition, $P = P(\rho, T, X_i)$, the nuclear energy production rate $\varepsilon = \varepsilon(\rho, T, X_i)$ and the opacity $\kappa = \kappa(\rho, T, X_i)$ which determines the temperature gradient in the radiative regions. The time integration of this differential system is started at an early stage, when the star is still entirely convective, with specified uniform composition.

The need to calibrate

Mathematically, the problem is well defined: the differential system is of order $4+2n$ in space with an equal number of imposed boundary conditions. Its numerical integration presents no fundamental difficulty. Thus it would seem that it suffices to run a stellar structure code with specified mass $M_\odot$ to the present age of the Sun in order to obtain an acceptable solar model. Such a model could then be validated by comparing the predicted radius and luminosity with the observed ones. This would indeed be the case if the initial composition were known with high precision and if the physical processes coming into play were all translated into perfect numerical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

prescriptions. Unfortunately, these two conditions are not yet satisfied.

The abundance of helium determined by spectroscopy is rather inaccurate: most of its spectral lines are hard to interpret because they are produced in the chromosphere and in the corona, where there are large deviations from local thermodynamic equilibrium. Furthermore, what one needs to start the time integration is the initial abundance Y_0, which may depart significantly from the present one because helium, being heavier than hydrogen, may have settled below the observed layers. For this reason Y_0 is taken as an arbitrary parameter (chosen in some reasonable range), to be adjusted by comparing the model with the actual Sun.

Another major uncertainty arises from the poor description of the convective heat transport. The recipe which is still commonly used, for lack of something better, also involves an arbitrary parameter, which is the ratio α between the mean-free path of the convective eddies (the so-called mixing length) and the pressure scale height.

For these reasons, the time integration of the constitutive equations is started with trial values for the two parameters Y_0 and α, and the resulting radius and luminosity, at the age of the Sun, are compared with the actual values $R_\odot$ and $L_\odot$. This procedure is iterated until the calibration is judged satisfactory, considering the other uncertainties.

The standard model

The standard model may be defined as the model which uses the best physical data that are available (equation of state, opacities, nuclear reaction rates) and which agrees best with the observations, while including the minimum of physical processes to achieve this result. As the observational constraints become more severe, the standard model is bound to include physical processes which have been neglected before. At the time of writing, the last process implemented in the standard model was the microscopic diffusion of helium and the heavy elements in the radiation zone, and this addition was required by the helioseismic profiles which clearly indicated that helium has settled out of the convection zone (see below).

Probing the solar interior

Since the observed values of the solar luminosity and radius are needed to calibrate the initial helium abundance Y_0 and the mixing length parameter α, one has to use other means to check the accuracy of solar models. Indirect evidence can be gleaned from observing the properties of other stars, for which models have been built with the same prescriptions (see SOLAR–STELLAR CONNECTION). But fortunately we dispose of means which allow us to probe the solar interior directly.

Solar neutrinos

Unlike photons, neutrinos show extremely little interaction with matter and therefore most of those emitted by the fusion of hydrogen in the central region of the Sun escape and a very small fraction of these can be detected on Earth. Indeed, no fewer than six different experiments have measured the flux of solar neutrinos, thus proving that the solar energy arises from nuclear fusion (see NEUTRINOS). Originally, the expectation was that these experiments would provide precious constraints on the central conditions of the Sun. However, the measured flux turned out to be systematically smaller than that predicted by the standard model, by a margin which cannot be reconciled by adjusting the sensitive parameters; we shall see below that non-standard models do not fare much better, although they include additional physical processes (see below).

The reason for this discrepancy has been the matter of long-standing debate. Is it due to serious imperfections of the standard model? Or is there something wrong with the experiments and their interpretation, as suggested by the incompatibility of their results? Or is the observed deficit due to some unknown property of the neutrino, beyond the electroweak theory? Recent results obtained at SuperKamiokande seem to favor the latter explanation; their straightforward explanation is that neutrinos have a non-zero mass; the neutrinos emitted in the solar core would then oscillate between two states, of which only one is detected by the present experiments. Thus it is quite possible that in the not too distant future the neutrino flux can be used to probe the solar core.

Helioseismology

Fortunately another technique has been developed in the meantime for the same purpose of sounding the solar interior, and it has turned out to be extremely powerful: called helioseismology, it makes use of the elastic properties of the Sun which allow for the propagation of waves. The sound waves observed at the surface of the Sun penetrate to a depth which depends on their frequency and on their wavelength; through an appropriate treatment inspired by seismology on Earth, they reveal some properties of the layers which they traverse (see HELIOSEISMIC OBSERVATIONS and HELIOSEISMOLOGY THEORY).

The vertical profile of the sound velocity determined by this technique has now reached a precision of the order of 10^{-4}, thus calling for models built with at least the same numerical accuracy, and also for a refined evaluation of solar age. Helioseismology permits us to locate the base of the convection zone within a thousandth of the radius, and it proves unquestionably that helium has settled below it. The characteristics of high-degree modes allow us to evaluate the helium content of the convection zone. Moreover, the rotational splitting of the mode frequencies yields the internal rotation velocity of the Sun as a function of depth and latitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rotation of the solar interior

Helioseismology has revealed that the angular velocity varies with latitude in the solar convection zone, the equator rotating faster than the poles as observed at the surface, but that it depends little on depth (see SOLAR INTERIOR: ROTATION). Deeper down, the radiation zone appears to rotate almost uniformly, with the possible exception of the innermost core. The shallow boundary layer between the two rotation regimes, which has been called the *tachocline*, is a region of strong vertical shear, and it certainly plays a crucial role in the generation of the toroidal component of the magnetic field (see SOLAR ACTIVITY).

Non-standard models

The present standard model fits the helioseismic data rather well: within a few thousandths, when comparing the observed and predicted sound velocities. Many solar physicists believe that this excellent agreement will further improve with better opacities, better nuclear reaction rates, some refinements of the equation of state and more accurate determinations of the chemical composition. However, there are some observations which call for the implementation, in the model, of additional physical processes which have been neglected until now.

One is the *depletion of lithium* with age which is observed in solar-type stars: the lithium abundance is 100 times higher at the surface of young stars of solar mass (members of the Pleiades cluster, for example) than it is in the Sun. The reason is certainly that this fragile element is destroyed inside the Sun by the nuclear reaction $^{7}\mathrm{Li+p}\rightarrow{}^{4}\mathrm{He+}{}^{4}\mathrm{He}$, but the temperature at the base of the convection zone is not high enough to explain such a large depletion. Therefore some mixing must occur within the radiation zone to carry lithium deeper down, and non-standard models have been built to account for this mixing (see the review by Pinsonneault 1997). Some invoke macoscopic motions, such as the large-scale flow which is induced by the spin-down of the star and the weak turbulence caused by the shear of differential rotation. Others ascribe this extra transport to the gravity waves which are emitted by the convective motions at the base of the convection zone.

Some mixing is also needed to soften the profile of molecular weight below the convection zone; microscopic diffusion and gravitational settling alone would establish a gradient whose steepness is incompatible with the helioseismic data. The fact that this mixing ought to occur in the strong shear of the tachocline is hardly a coincidence.

Non-standard models are also required to correctly predict the observed *internal rotation* of the Sun: in particular, they have to explain why the radiative interior appears to rotate so slowly and so uniformly. Before this property was established by helioseismology, the solar core was expected to rotate faster than the convective envelope, which is spun down by the solar wind. Magnetic torques can easily achieve that uniform rotation, but other mechanisms may also contribute, such as the transport of angular momentum by gravity waves.

Other non-standard models have been built to produce a smaller flux of neutrinos, in an effort to match the observed values; in general they introduce some extra mixing in the solar core, which modifies chemical composition and therefore the network of nuclear reactions and the emission of neutrinos. But so far these models have failed to comply with the helioseismic constraints. Work progresses very actively on such non-standard models.

Bibliography

Hansen C J and Kawaler S D 1994 *Stellar Interiors; Physical Principles, Structure, and Evolution* (Berlin: Springer)

Kippenhahn R and Weigert A 1990 *Stellar Structure and Evolution* (Berlin: Springer)

Pinsonneault M 1997 *Ann. Rev. Astron. Astrophys.* **35** 557

Stix M 1989 *The Sun* (Berlin: Springer)

Jean-Paul Zahn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Convection Theory

Convection is the transport of energy by hot fluid moving upwards and cold fluid moving downwards. A familiar example is the upwards and downwards motions one may observe (best just before the boiling point) when heating water in a casserole. In stars, convection is often responsible for transporting heat up to the visible surface, or for transporting heat away from the central parts of the star, where energy is being generated by nuclear processes.

Whether or not convection occurs in a certain region of a star depends on how transparent or opaque the region is to radiation. The transparency (or its inverse, the opacity) determines the temperature gradient necessary to transport the total stellar energy output; the stellar luminosity. If the temperature gradient exceeds a certain critical threshold (which depends on the detailed properties of the gas), then the gas is convectively unstable, and convective motions occur spontaneously. The convective motions assist in the transportation of the stellar luminosity towards the surface, where the energy is finally radiated away freely into space in the form of light.

In the Sun, conditions are such that convection is the dominant energy transport mechanism in the outer 30% (by radius), while there is no convective transport of energy associated with the energy-generating central region.

Convective motions are directly visible on the solar surface; energy-carrying scales give rise to a cellular brightness pattern called GRANULATION (sizes from a few hundred to several thousand km), while larger, more deeply rooted scales are visible as patterns of primarily horizontal motions called meso-granulation (up to about 10 000 km) and super-granulation (up to about 100 000 km) (see SOLAR PHOTOSPHERE: MESOGRANULATION; SUPERGRANULATION).

Density stratification

One of the most decisive conditions influencing the way convection in the Sun has to organize itself is the enormous range of mass densities in the SOLAR INTERIOR. Figure 1 shows the density and pressure stratification, illustrating that the density (pressure) increases by over ten (fourteen) orders of magnitude from the surface down to the center of the Sun, where nuclear processes are generating the energy.

Note that, even though the solar convection zone extends over only 30% of the radius, eight orders of magnitude of density change occur there. These eight orders of magnitude have again been split into half, to illustrate that the top four orders of magnitude of density change occur within only 3000 km of the surface (corresponding to about 4% of the radius). The reason why one order of magnitude of density (or pressure) change near the surface covers a much smaller depth interval than further down is that the density scale height, defined as $H_\rho = \mathrm{d}z/\mathrm{d}\ln\rho$, is roughly proportional to temperature, and the temperature increases with depth below the solar surface. (See also SOLAR INTERIOR: CONVECTION ZONE.)

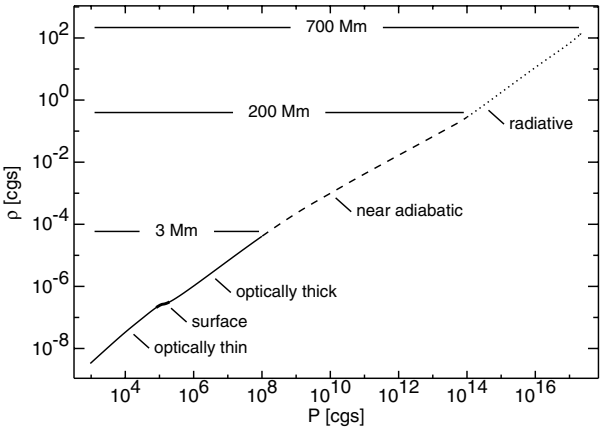


Figure 1. The density and pressure stratification of the Sun. The figure extends from the top of the photosphere to the center of the Sun. The properties of the thin transition region between the optically thin layers of the solar photosphere and the convection zone are crucial for the structure of the whole convection zone.

The small 'plateau' in the density marks the visible surface of the Sun. There, the temperature drops so rapidly that the density (which is roughly proportional to the pressure divided by the temperature) stays nearly constant, or even decreases slightly with depth. The pressure must necessarily increase monotonically with depth, since the pressure at any one depth is determined by the weight of all the mass at smaller depths (larger radii). But there is relatively little mass in the thin surface layer where the temperature drops rapidly (from about 10 000 K to about 5000 K over less than 100 km), and thus the pressure does not change much, and therefore the temperature drop is reflected in the density stratification as that little plateau.

The occurrence of that plateau, where the density scale height is radically different, and possibly even goes to infinity (this happens if the density has a local minimum there), may have significantly influenced the development of the theory of stellar convection. It was clear to researchers early on that the rapid change of density with depth must have important consequences for the flow patterns.

Consider, for example, two depth levels, where the vertical velocities are about the same, but the density differs by a factor of ten. The amount of matter transported upwards (as well as downwards) then is ten times larger through the deeper lying level, and all of the upward moving mass clearly cannot be accommodated by the higher lying level. The downward moving mass at that higher level, on the other hand, can easily be accommodated by the lower lying level, taking up only 10% of the downward mass transport capacity there. One concludes that most of the ascending gas from the lower lying level has to turn over into descending motion in between the two levels (thus making up the remaining 90% of downflow there).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It would then have been natural to incorporate the density scale height H_ρ, mentioned earlier, into a theoretical description of convection. However, because the density scale height may become infinite, whereas the pressure scale height by necessity is regular (while still being comparable to the density scale height where that is regular) one chose to instead use the pressure scale height as the geometrical scaling parameter of the theory.

The theory, that was developed for stellar convection by Erika Vitense (1953), thus has as one of its main ingredients the assumption that ascending hot gas must necessarily mix with descending, cooler gas over a distance that is roughly proportional to the pressure scale height, $H_P = \mathrm{d}z/\mathrm{d}\ln P$, which for an ideal gas is equal to $RT/g\mu$, where R is the gas constant, T is the temperature, g is the acceleration of gravity and μ is the mean molecular weight.

The distance over which mixing occurs is thus written as $\ell = \alpha H_P$, where α is a free parameter of the theory. In analogy with similar early theories for laboratory turbulence, the distance ℓ was called a 'mixing length', the parameter α was called the 'mixing length parameter', and the theory has been referred to as the 'mixing length theory'ever since.

Convective heat transport

Heat transport by convection occurs because upward moving gas, which is hotter, transports more energy upwards than cooler, downward moving gas transports downwards. The rate of convective energy transport, the convective energy flux, is therefore proportional to the product of the velocity amplitude and the amplitude of the temperature fluctuations. Pressure is roughly constant in horizontal planes, and one may thus invoke the specific heat capacity at constant pressure, C_P, and write the convective energy flux as

$$F_{\mathrm{conv}} \approx \langle \rho u_z C_P \Delta T \rangle \tag{1}$$

where u_z is the vertical velocity (positive upwards), and ΔT is the temperature fluctuation.

The convective instability occurs when vertical motions spontaneously generate positively correlated temperature and velocity perturbations. A parcel of gas that is displaced upwards expands and cools, because of the drop in pressure with height. The cooling is associated with adiabatic expansion, i.e. expansion without exchange of heat. Cooling by expansion is the process at work in a refrigerator, and the inverse, heating by compression, is the process at work in the cylinders of an engine (before ignition).

If the mean temperature drop with height is larger than the adiabatic cooling, then the displaced parcel finds itself hotter than the surroundings. Conversely, a downward displaced parcel of gas from above becomes cooler than the surroundings. Since hotter gas is lighter, and cooler gas is heavier, such perturbations are self-amplifying; the larger weight accelerates the cooler gas further downwards, while the lighter gas is accelerated upwards.

The mixing of ascending and descending gas mentioned earlier implies mixing of hot and cold gas, and mixing of upwards and downwards motions, and hence a loss of organized heat transport. It is the balance between the spontaneous growth associated with the convective instability and the loss associated with the mixing that ultimately determines the level at which convection saturates, for any given temperature gradient.

Balancing the loss of vertical momentum against the gain from the acceleration associated with the temperature difference ΔT one obtains the estimate

$$u_z^2 \sim \alpha_1 H_P g \Delta T / T \tag{2}$$

where g is the acceleration of gravity and α_1 is a parameter of the order of unity. Balancing the growth of temperature differences from vertical motions against the loss from mixing one obtains the estimate

$$\frac{\Delta T}{T} \sim \alpha_2 \left[\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P} - \left(\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P} \right)_{\mathrm{adiabatic}} \right] \tag{3}$$

where α_2 again is a parameter of the order of unity, and where the bracket is the difference between the actual temperature gradient and the one that corresponds to adiabatic expansion/contraction with height—this temperature gradient difference is often denoted $\Delta\nabla$.

When taken together, the equations above imply that the convective energy flux may be expected to scale roughly as

$$F_{\mathrm{conv}} \sim \alpha^{3/2} \rho c_s C_P T (\Delta\nabla)^{3/2} \tag{4}$$

where α is a parameter of the order of unity that satisfies $\alpha^{3/2} = \alpha_1^{1/2} \alpha_2 / \Gamma_1^{1/2}$, where Γ_1 is the adiabatic gas gamma, relating P/ρ to sound speed c_s. The quantity $\rho c_s C_P T$ has the dimension of an energy flux and in the interior of the Sun it is much larger than the actual solar energy flux. This means that the difference between the actual temperature gradient and the adiabatic one, $\Delta\nabla$, is small, and that consequently the relative temperature fluctuations $\Delta T/T$ and the ratio of velocity to sound speed v/c also are small.

One may therefore conclude that in the interior of the Sun convection is able to carry practically all of the required energy flux with relatively weak convective velocities, maintained by a temperature gradient that differs only little from the adiabatic one.

Close to the surface of the Sun this is no longer true; there the mass densities ρ, the sound speed c_s and the temperature T have all dropped so much that convective velocities reach a good fraction of the sound speed, the relative temperature fluctuations are large, and the temperature gradient deviates significantly from the adiabatic one.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Near-surface convection

The visible surface of the Sun occurs where the dimensionless optical depth τ_λ is equal to unity. The optical depth increment $\mathrm{d}\tau_\lambda$ is equal to $\rho\kappa_\lambda\,\mathrm{d}z$, so it is proportional to the mass density and the absorption coefficient κ_λ (that has dimension $\mathrm{m}^2\ \mathrm{kg}^{-1}$ in SI units). At temperatures of the order of the solar surface temperature, κ varies extremely rapidly with temperature ($\kappa \sim T^{10}$). In addition, a large number of atomic and molecular spectral lines also make the wavelength dependence of κ_λ very complicated.

The ultimate outcome of the decreasing efficiency of convective energy transport and the rapid decrease of opacity with decreasing temperature is that, near the solar surface, radiative energy transport again takes over as the dominant energy transport mechanism, after having turned over the responsibility for energy transport to convective motions in the solar convection zone. The transition occurs in an extremely narrow boundary layer, the thickness of which is significantly smaller than a pressure scale height, and much smaller than the horizontal size of the energy-carrying convection cells.

Because of the relatively low efficiency of convection in the crucial layers near the solar surface, the transition from predominantly convective to predominantly radiative energy transport is accompanied by a significant drop in entropy; there is a thin but very important superadiabatic layer that separates the optical surface of the Sun from the almost adiabatic interior. It is this layer that gives rise to the near-surface 'plateau' in density in figure 1, and it is also this layer that ultimately determines the radius of the Sun; if the entropy jump were different, the whole convection zone would have to adjust, so as to maintain the correct surface luminosity.

Numerical modeling

It is in practice impossible to determine the quantitative properties of the transition from convective to radiative convective energy transport by analytical means; there are too many, too complicated processes that interact.

There are two compounding reasons in particular that complicate the physics of the solar surface layers: on the one hand the optical properties, and on the other hand the very strong density and pressure stratification of the solar surface layers. The crucial transition from convective to radiative energy transport thus takes place inside convection cells, across a thin surface layer that is itself buckling and folding in response to local changes of the velocity and temperature patterns. At the same time, across a height extent of the order of characteristic horizontal scales of the energy carrying cells, the density varies by several orders of magnitude (cf figure 1).

Thus, whereas the scaling analysis leading to the 'mixing length' estimates above is useful to obtain order of magnitude estimates (and for many years have been used even for quantitative modeling, after 'calibrating' the mixing length parameter α empirically), one cannot hope to obtain self-consistent, parameter-free prediction without actually including all relevant physical processes in sufficient detail. For that, one must resort to accurate numerical modeling. (See NUMERICAL SIMULATION OF CONVECTION: COMPARISON WITH MIXING LENGTH THEORY.)

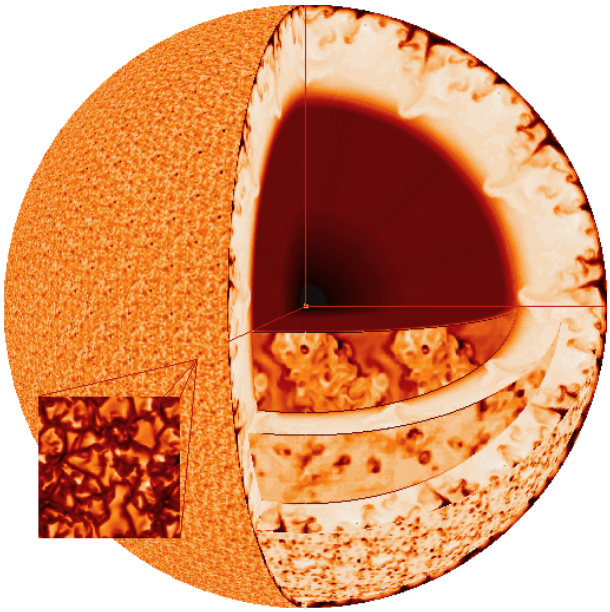


Figure 2. A composite illustration obtained by combining warped images from numerical simulations onto the faces of a cut-out view of the Sun. The quantity displayed is the entropy, which is almost constant in the convection zone, and is lower in the solar interior.

The circumstance that a shallow surface layer is almost all that matters, in so far as both observable surface properties and the influence of convection on the solar structure is concerned, is then indeed fortunate. It means that detailed 3D models may be restricted to a region only somewhat larger than that layer, thus avoiding most of the complications of covering multiple scales (in space and time). Also, for the very same reason that numerical modeling is really required, i.e. large-amplitude temperature fluctuations and ratios of velocity to sound speed v/c_s no longer small, fully compressible 3D simulations become feasible, because turn-over times are comparable to the sound travel times.

Deeper down, where the convective velocities are much smaller than the sound speed, and where the actual solar flux is much smaller than $\rho c_s C_P T$, direct numerical simulations are still prohibitively costly, because too many time steps are required to follow the slow motions. There, one may instead make scaled models that have convective fluxes much larger than the actual solar flux. From equations (2) and (4) one expects the velocity to scale as $v \sim F_{\mathrm{conv}}^{1/3}$, and the temperature to scale as $\delta T \sim F_{\mathrm{conv}}^{2/3}$. It is thus possible to rescale the results of such simulations to solar conditions. Properties such as the topology of the flow are expected to be invariant under such scalings.

Figure 2 shows results from such scaled models,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

wrapped onto a spherical model of the Sun. Also shown is an inset with a blow-up of the surface radiation intensity of a model of the solar surface layers. Typical geometrical scales and evolution times in the surface model are a few thousand km, and a few minutes, respectively. The corresponding numbers for the lower part of the convection zone are of the order of a few hundred thousand km, and a few weeks, respectively. This illustrates why it is not possible to cover all geometrical scales and all time scales in a single numerical model. Nevertheless, numerical models are very helpful, both as a guide to the development of theoretical concepts, and as a source of synthetic observational data that may be compared in a direct, forward manner with observations.

Bibliography

Spruit H C, Nordlund Å and Title A 1990 *Ann. Rev. Astron. Astrophys.* **28** 263–301
Stein R F and Nordlund Å 1998 *Astrophys. J.* **499** 914
Vitense E 1953 *Z. Astrophys.* **32** 135

Åke Nordlund

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Convection Zone

The Sun, like all main sequence stars of moderate mass, possesses a deep convective envelope just below its surface in which turbulent fluid motions serve to carry outward the energy flux that results from nuclear burning within its core. Observations of the Sun reveal complex flows and magnetic structures that are evidence of highly turbulent convection just below the surface of this rotating star. Theoretical models of stellar structure and evolution indicate that such convection must extend from the surface well into the interior within the present Sun, forming a convection zone occupying about the outer 30% by radius (or a depth of nearly 200 Mm). Within this outer envelope, convective motions will transport nearly all of the emerging flux of energy, whereas below it radiation suffices to carry the flux outward from the core. The energy transport by fluid motion is very efficient throughout the convection zone, except in a thin superadiabatic boundary layer immediately beneath the photosphere in which there is a transition to transport by radiation in the solar atmosphere. There is another complex transition layer at the bottom of the zone where penetrative convective motions extend a small distance into the very stable radiative interior.

Intricate turbulent dynamics

Intense turbulence within the convection zone not only transports heat but also continually redistributes angular momentum. This leads to a pronounced DIFFERENTIAL ROTATION within the convection zone. It has long been known from observation of features such as SUNSPOTS that the surface of the Sun rotates differentially: there is a smooth poleward decline in the angular velocity Ω, the sidereal rotation period being about 25 days in equatorial regions and about 33 days near the poles. More recently, helioseismology has revealed how Ω varies with both radius and latitude throughout much of the solar interior. The convection also appears to build magnetic fields by some form of DYNAMO action, the details of which are likely to depend sensitively on the rotation profile that is established both within the convection zone and in a region of overshooting immediately beneath its base. Yet how the turbulence couples to the rotation and the magnetic field is not understood in detail. It is very difficult to describe theoretically the dynamics of such highly nonlinear compressible fluid flows that extend over many decades in scale. One of the novel aspects of such dynamics is that amidst the apparent chaos there also appears to be distinct order. This is suggested by the relatively smooth variation of angular velocity observed at the solar surface, and by the nearly cyclic temporal variation of magnetic activity, involving sunspot eruptions with very well defined rules for field parity and emergence latitudes as the cycle evolves.

The convection zone of the Sun thus exhibits some striking properties that have largely defied theoretical explanation. Most prominent are issues concerning its rotation profile with latitude and depth (the 'solar differential rotation problem'), and the manner in which the Sun achieves its 22-year cycles of magnetic activity ('solar dynamo problem'). Both global issues touch on the seeming inconsistency that turbulence can be both highly intermittent and chaotic on smaller spatial and temporal scales, and yet achieve a large-scale order that is robust in character. Observations of the solar surface show that the Sun possesses a wide range of scales of convection, ranging from granules ($\sim 10^3$ km or 1 Mm in horizontal size), to mesogranules ($\sim$5 Mm), to supergranules ($\sim$30 Mm), to possible patterns of giant cells comparable to the overall depth of that zone ($\sim$200 Mm) (see SOLAR PHOTOSPHERE: GRANULATION, SOLAR PHOTOSPHERE: MESOGRANULATION and SOLAR PHOTOSPHERE: SUPERGRANULATION). Similar scale separation is seen temporally. Small-scale magnetic fields rekindle with the turnover of the granulation ($\sim$10 min), and yet somehow the Sun achieves 22-year cycles of magnetic activity. The wide range of dynamical scales involved with turbulence in the Sun presents severe challenges to both theory and numerical simulation: in the convection zone of the Sun these scales range from about 10^5 km (about the depth of the convective layer) to 0.1 km or smaller (dissipation scales), encompassing at least six orders of magnitude for each physical dimension. The largest current turbulence simulations can capture about three orders of magnitude in each dimension. Theoretical progress in dealing with the turbulent dynamics of the solar convection zone thus requires trade-offs in either studying a highly turbulent but localized portion of a convecting spherical shell or a less turbulent but full shell with correct account of geometry.

The magnetic fields that emerge into the solar atmosphere are also diverse in character, and this must reflect how they are produced and shuffled within the depths of the convection zone. The cyclic eruption of strong magnetic fields through the solar surface in the form of sunspots and ACTIVE REGIONS has long been a center of attention. In the active Sun, large sunspots have central dark umbras with field strengths of 2000 to 3000 G and diameters as great as 10 Mm, surrounded by a brighter penumbra of up to three times that size. The largest of such active region complexes can span a good fraction of the disk. Yet observations have shown that magnetic flux is also emerging at a comparable rate during the quieter phases of the 22-year cycles, but does so through far smaller structures. In the quiet Sun, much of the magnetic energy resides in compact flux elements or tubes with strengths of the order of 1000 G to 1500 G, with these occupying only a very small fraction of the surface area to yield an average field strength of a few gauss. The smallest of the detected compact magnetic structures are the bright points, about 0.1 Mm in size, visible in the intergranular lanes. The magnetic fields also appear as somewhat larger structures, possibly composed of many individual flux tubes assembled into flux ropes that have broken through the solar surface, appearing as dark magnetic pores about 1 Mm across. Such strong field components in the quiet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun are augmented by a weaker component made up of many small dipole loop pairs with field strengths possibly as low as 50 G. The gradual sweeping of magnetic fields to mesogranule and supergranule boundaries yields an ordering of these fields into networks of larger scales, with such relatively narrow but widely spaced network lanes occupied by compact field elements.

Solar differential rotation

The advent of helioseismology, which involves the study of the acoustic p-mode oscillations of the solar interior that are observable at the surface, has permitted probing of both structure and dynamics below the solar surface. The profiles of angular velocity (Ω) revealed by helioseismology have turned out to be revolutionary, for they are unlike any anticipated by theory prior to such probing of the interior of a star. The convection in early numerical simulations of rotating convection in spherical shells was dominated by columnar roll-like cells (or 'banana cells') oriented in a north–south direction, the tilting of which yield nonlinear interactions or 'Reynolds stresses' that sustain the differential rotation. Such convection models yielded an angular velocity Ω nearly constant on cylinders, aligned with the rotation axis and decreasing with depth in the equatorial plane. The Sun does not appear to comply.

Helioseismology has shown that the rotation profiles obtained by inversion of frequency splittings of the *p* modes have a decidedly different character: the angular velocity observed near the surface, where the rotation is faster at the equator than near the poles, extends through much of the convection zone with little radial dependence. In particular, Ω at high latitudes in the convection zone appears to increase slowly with depth, at mid-latitudes it is nearly constant on radial lines, and near the equator it first increases and then gently decreases with depth. Another striking feature is the region of strong shear at the base of the convection zone, known as the *tachocline*, where Ω adjusts to apparent solid body rotation in the deeper radiative interior. Thus whereas the convection zone exhibits prominent differential rotation, the deep radiative interior does not; these two regions are joined by the complex shear of the tachocline. More detailed deductions about subsurface solar flows and rotation, and possible changes as the Sun enters active phases of its magnetic cycle, are becoming available from the nearly continuous helioseismic observations with the ground-based six-station Global Oscillation Network Group (GONG) project, and from the SOI–MDI experiment on the SOHO spacecraft.

Fully developed turbulence at Reynolds numbers R_e (ratio of inertial to viscous forces) of order 10^{12}, as estimated for the Sun, appears to redistribute angular momentum quite differently from what was thought on the basis of laminar and mildly turbulent convection studies, thereby yielding mean flows and rotation profiles of a special and different kind. Recent three-dimensional simulations have begun studying full spherical shells of rotating turbulent convection using spherical harmomics Y_ℓ^m (for horizontal expansion) up to degree $\ell \sim 400$. The models are a highly simplified description of the solar convection zone: solar values are taken for the heat flux, rotation rate, mass and radius, and a perfect gas is assumed since the upper boundary of the shell lies well below the H and He ionization zones. The computational domains include a region of stable stratification below the primary unstable zone in which the effects of penetrative convection can also be studied.

The resulting convection is highly time-dependent and the flows are intricate. The time-averaged angular velocity Ω is nearly constant on radial lines throughout much of the convection zone at mid latitudes, and there is a systematic decrease of rotation rate with latitude in going from the equator to the poles. The convection is now dominated by intermittent plumes of upflow and stronger downflow, some possessing a distinctive cyclonic swirl; the dominant role of coherent plumes is now becoming apparent. A tachocline at the base of the convection zone is not yet realized, since the simulations still operate at R_e where viscous coupling can imprint the local rotation rate of the unstable zone upon the stable region. The development of a tachocline may involve anisotropic diffusivities and confining magnetic fields. Some theoretical discussions concerning the tachocline invoke strong horizontal viscosity due to anisotropic turbulence in the stable layer to circumvent the diffusive spreading. Other models invoke magnetic fields to enforce solid body rotation in the radiative interior. Simulations at higher R_e (requiring greater resolution) and including the effects of magnetic fields can test such theories and provide hope for modeling the tachocline. It will be essential to have at least four-fold higher spatial resolution (effectively dealing with 2048^3 modes) in such simulations, now becoming feasible on the latest parallel machines, to study a fully developed convection zone bounded below by a radiative layer.

Solar dynamo and cyclic field emergence

Magnetic flux emergence (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES) at the solar photosphere occurs on a wide range of scales, from large active regions (sunspot groups) to small fibril concentrations in the intergranular lanes. A central problem of magnetic field generation in the Sun is the origin of the coherent magnetic field that emerges at the solar surface to form the active regions. Although it was once widely believed that all the smaller-scale flux was derived from the break-up of the larger regions, current observations suggest that flux emerges at a rate which is inconsistent with this simple fragmention model, and therefore flux must be actually generated at the smaller scales. The solar dynamo problem can thus be thought of as two problems rather than one, i.e. a *large-scale dynamo* responsible for global properties of the magnetic field including cyclic behavior and active regions, and a *small-scale dynamo*, contributing to the abundance of short-lived mixed-polarity flux found outside the active regions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It is now commonly thought that the global dynamo behavior is derived from strong organized toroidal magnetic fields generated by the shearing flow of the tachocline just below the base of the solar convection zone. The weaker poloidal field is then regenerated throughout the convection zone by the action of cyclonic turbulence. This separation of the generation of toroidal field and the regeneration of poloidal field has been termed an *interface dynamo*. Recent *mean-field* dynamo models have shown that such an interface model can circumvent the problem of strong dynamo suppression feedback (α quenching) by mean magnetic fields and so produce field strengths similar to those inferred from observations. They can also give rise to modulation of the basic cycle and recurrent so-called grand minima.

The interface paradigm is based on a number of underlying processes, crucial to the overall operation of the dynamo, which are parametrized in mean-field models and yet should be modelled explicitly. The building blocks of the full mechanism are (a) the actual generation of field by cyclonic turbulence within the convection zone (small-scale dynamo, parametrized by the α effect in mean-field models), (b) turbulent transport of this field across the interface (enhanced turbulent diffusion in the mean-field sense), (c) its subsequent amplification by large-scale shear in the overshoot region below (large-scale dynamo, represented by the ω effect in mean-field models), and (d) transport of the resultant large-scale field through the convection zone to be either shredded and recycled or to emerge in the photosphere. Currently, computational resources are insufficient to model a complete system incorporating both small- and large-scale dynamo effects. Efforts are under way to study separately these individual building blocks, with the aim of using the intuition so developed toward attaining a fully nonlinear interface-type solar dynamo model as resources become available.

Bibliography

Brummell N, Cattaneo F and Toomre J 1995 *Science* **269** 1370–9

Elliott J R, Miesch M S and Toomre J 2000 *Astrophys. J.* **533** 546–56

Schou J *et al* 1998 *Astrophys. J.* **505** 390–417

Thompson M J *et al* 1996 *Science* **272** 1300–5

Juri Toomre

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Convection Zone Flux Tubes

The magnetic field at the solar surface reflects the presence of isolated flux tubes below the photosphere. The magnetic field is not diffusively distributed over the entire surface but is concentrated into regions of rather intense field, ranging from SUNSPOTS—with diameters of approximately 20 000 km and field strengths of 3000 G—down to smaller magnetic elements (100–200 km and 1000 G). The existence of this hierarchy of concentrated magnetic structures in the observable layers is indicative of the intermittent nature of magnetic fields in the underlying convection zone (see SOLAR INTERIOR: CONVECTION ZONE). The equilibrium and dynamics of flux tubes are basically different from those of a diffuse field. This has important consequences for the storage of magnetic flux, the onset and development of instabilities, the rise of magnetic flux towards the surface to form the observed active regions, and also for the understanding of the dynamo mechanism (see DYNAMO THEORY). Magnetic buoyancy due to the increased effective pressure inside the tubes poses a dilemma: on the one hand, it is necessary to bring the magnetic flux to the solar surface; on the other it leads to a rapid loss of magnetic flux from the convection zone, thus preventing the storage of magnetic flux long enough for the operation of a hydromagnetic dynamo. It is nowadays widely accepted that the magnetic field emerging at the surface is stored within a stable layer at the bottom of the convection zone prior to eruption.

Filamentary nature of the magnetic field

High-resolution observations show that the magnetic field at the surface of the Sun is not diffuse, but extremely concentrated into discrete structures in the form of magnetic flux tubes, of which sunspots are the most prominent manifestation (e.g. Zwaan 1992). More than 90% of the magnetic flux on the solar surface outside of sunspots and their associated active regions is concentrated into intense flux tubes (1–2 kG) of diameter 100–500 km. These small magnetic elements are surrounded by plasma with substantially lower field strength. This fibril state of the observed magnetic field is a consequence of the very large magnetic Reynolds number of the solar plasma and its unstable (superadiabatic) stratification in the layer below the photospheric surface. The idea that the discrete appearance of the magnetic field extends further through the convection zone is also supported by numerical simulations of magnetoconvection (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION).

Apart from observations of the solar surface, there are indirect hints for the existence of a strong toroidal system of magnetic flux at the bottom of the solar convection zone or below. Bipolar active regions form by the emergence of large bundles of toroidal magnetic flux from the convection zone into the photosphere. In spite of the turbulent convection zone in which they are immersed, active regions show a large number of regularities (Moreno-Insertis 1994). For example, Hale's polarity rule, the tilt of active region's main axis (by typically 10°) with respect to the equator and the systematic asymmetry between the preceding and following parts of the active region (as regards morphology, stability and proper motion) point towards the presence of a strong toroidal magnetic field at the bottom of the convection zone or just below.

The concentration of magnetic flux into flux tubes has important consequences for its storage. Since a magnetic field gives rise to an effective 'magnetic pressure', a flux tube surrounded by almost field-free plasma can become buoyant. Buoyancy can be a very efficient mechanism in expelling magnetic flux to the solar surface. Some simplified estimates as well as detailed numerical simulations show that flux tubes with a field strength of the order of equipartition with the energy density of convective motions (or larger) rise to the solar surface within one month or less; this time scale is much shorter than the 11-year characteristic time for field generation and amplification by the dynamo process. In order to circumvent this problem, it has been suggested that most of the magnetic flux is stored within the overshoot layer, a slightly subadiabatically stratified layer below the convection zone proper.

Overshoot region

Overshooting is the penetration of convective motions from an unstable layer into an adjacent stable layer. The existence of convective penetration is known from both observation and experiment in the laboratory and in the Earth's atmosphere. A convective eddy is driven in the unstable region and arrives at the boundary of that region with some inertia sufficient to penetrate into the surrounding stable layers where it is decelerated by the stable stratification. This yields an enlargement of the region effectively mixed by convection.

Near the bottom of the solar convection zone, a steep gradient of the differential rotation profile can be inferred from helioseismological data (see HELIOSEISMIC OBSERVATIONS); this zone is called the 'solar tachocline'. Many astrophysicists think that this zone plays an important role in the mixing processes between the radiative and the convective zones. It could also be the principal source of solar magnetic flux. In order to investigate the problem of the storage of magnetic flux tubes in the overshoot region one exploits the fact that they are *thin* as compared with the local pressure scale-height.

Thin flux tubes

The study of the structure and dynamics of magnetic fields in stars is a complex mathematical problem due to the nonlinear nature of the magnetohydrodynamic (MHD) equations, to the presence of boundary layers and to the inhomogeneities involved. The filamentary nature of magnetic fields in the convection zone allows for a considerable simplification of the MHD equations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The *thin flux-tube approximation* permits the reduction of the full set of MHD equations to a mathematically more tractable form without making any restriction on the compressibility of the plasma, the magnetic Lorentz force or gravity. In physical terms, the approximation amounts to assuming that the variation of the physical properties over the cross section of the tube is *small* as compared with the variation along it (with the exception of a jump at the boundary). Since for the stability and the first stages of the rise one is interested in perturbations that affect the flux tube as a whole, and given the large value of the pressure scale-height at the bottom of the convection zone (about 60 000 km), the use of the thin flux-tube approximation for the dynamics of the magnetic flux concentrations is perfectly justified.

With the aid of the thin flux tube approximation it is possible to obtain a consistent model of magnetic field storage, instability, dynamics and eruption not only for the Sun, but also for other stars with convection zones.

Stability of toroidal magnetic flux tubes

For a linear stability analysis of flux tubes in a stellar convection zone, one considers a toroidal flux tube symmetric with respect to the solar rotation axis and lying in a plane parallel to the equator, at an arbitrary latitude. Figure 1 shows a sketch of the adjustment of an initially buoyant flux tube with equal internal and external temperatures and angular velocities to a mechanical equilibrium characterized by neutral buoyancy and a longitudinal internal flow in the direction of stellar rotation: since the curvature force and the rotationally induced forces are both perpendicular to the axis of rotation, there always remains an unbalanced axial component of the buoyancy force (except for flux tubes lying exactly at the equatorial plane). Consequently, if the tube is to be at mechanical equilibrium, the matter inside the tube must be slightly cooler than its environment and rotate slightly faster. Since the motions causing the perturbation of the flux tube have a time-scale that is much shorter than the characteristic time for thermal relaxation, isentropic perturbations may be assumed (i.e. the entropy of each fluid element remains unchanged). Stability against isentropic perturbations can be examined by means of a normal-mode analysis of both axisymmetric and non-axisymmetric displacements of the equilibrium shape of the tube. The stability properties depend on the various parameters (e.g. latitude, field strength, superadiabaticity of the stratification, angular velocity and its gradients).

The stability analysis shows that the unstable stratification of the convection zone and the strong convective motions render the storage of magnetic flux in the convection zone proper impossible. The alternative is that the magnetic flux is stored in the overshoot layer underneath the convection zone, some 200 000 km below the solar surface, and with a thickness of only a fraction of the local pressure scale-height. Independent support for this idea comes from numerical simulations showing

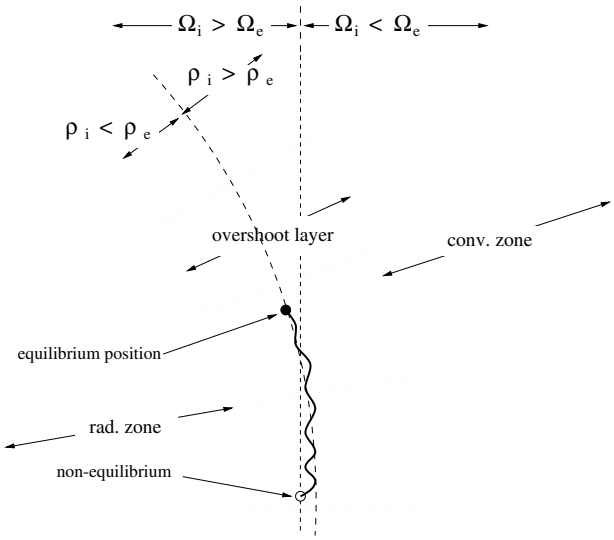


Figure 1. Approach of an initially buoyant flux tube (open circle) with $T_i = T_e$ and $\Omega_i = \Omega_e$ to a position of mechanical (force) equilibrium (full circle) characterized by neutral buoyancy ($\rho_i = \rho_e$) and faster internal rotation, $\Omega_i > \Omega_e$ (from Caligari *et al* 1998). Ω is angular velocity, ρ density, and T temperature. The subscripts 'i' and 'e' denote physical quantities inside and outside the flux tube, respectively.

that magnetic flux with a field strength of the order of the equipartition value is pumped downwards below the convection zone by the strong sinking plumes of overshooting convection.

Instability and rise of magnetic flux tubes

Due to the subadiabatic (stable) stratification of the overshoot layer, flux tubes can achieve a stable configuration in which buoyancy vanishes. Probably, the flux tubes evolve through a sequence of equilibria while being continuously amplified by the dynamo mechanism which is supposed to operate within this layer. Once the field has become sufficiently strong, instabilities set in; the most unstable perturbations (i.e. with the shortest growth time) are non-axisymmetric (i.e. undulatory) and give rise to the formation of rising loops. The critical magnetic field strength for the onset of the instability lies around 10^5 G for conditions prevailing at the bottom of the solar convection zone (e.g. Schüssler *et al* 1994). If we assume that a typical tube which later gives rise to an active region has a magnetic flux between 10^{21} and 10^{22} maxwell (i.e. at least the total flux of the active region), then the typical size of the tube at the onset of the instability must be of the order of 1000 km.

A linear stability analysis can provide the proper initial conditions for numerical simulations of the emergence of magnetic flux loops through the convection zone. Once an unstable loop has entered the superadiabatic part of the convection zone, the subsequent evolution becomes very fast and nonlinear, so that numerical simulations are necessary to follow its rise towards the surface. In the last two decades, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

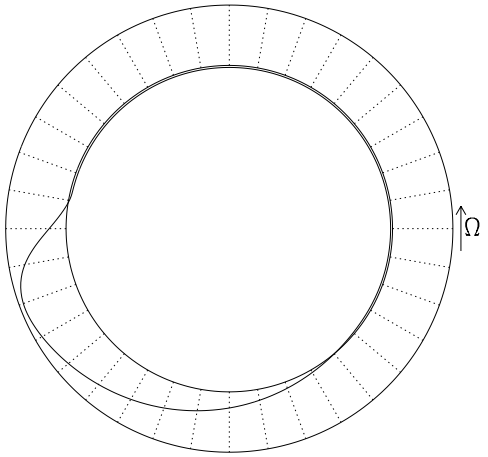


Figure 2. Snapshot of an unstable magnetic flux tube that started at the bottom of the convection zone. The direction of solar rotation is towards increasing azimuthal angle (from Caligari *et al* 1995).

nonlinear dynamic evolution of unstable flux tubes has been studied in detail by numerical integration of the thin flux-tube equations (see e.g. Schüssler *et al* 1994, Fisher *et al* 2000). For initial field strengths of the order of equipartition (10^4 G) or a few times this value, the Coriolis force plays a dominant role and flux tubes starting from low latitudes are deflected polewards along the solar rotation axis and emerge at unrealistically high latitudes, contrary to sunspot obervations. With a field strength of approximately 10^5 G, the results of the numerical simulations are in agreement with a number of observable features. For instance, the rise is not significantly deflected by the Coriolis force, so that the flux tubes emerge at low latitudes, the deviation of the eruption latitude from the starting latitude being less than 5°. Other features that are reproduced by the numerical simulations are the tilt angle (inclination with respect to the East–West direction) of sunspot groups or the asymmetry between the two legs of the rising tubes (figure 2).

Flux tubes and the dynamo problem

It is widely accepted that the solar activity cycle, with its various manifestations and regularities, is the result of a hydromagnetic dynamo operating at the bottom of the solar convection zone. The principle of dynamo action was first suggested by Larmor in 1919. Basically, it requires that the plasma moves in such a way as to induce electric currents capable of maintaining and amplifying a 'seed' field against Ohmic decay. In the case of stars like the Sun, one of the main ingredients for dynamo action is differential rotation, which generates the toroidal field component from the poloidal component. The generation of the poloidal field from the toroidal field is explained in terms of an induced electric current parallel to the mean magnetic field (α-effect), an idea originally put forward by Parker in the context of a turbulent solar dynamo.

The concentration of magnetic flux into intense tubes also has important consequences for the operation of the dynamo. The conventional kinematic approach—in which the back-reaction of the Lorentz force on the flow field is neglected and in which turbulent flows play a key role in regenerating the poloidal field from the toroidal component—is not applicable. Nevertheless, although the α-effect was originally formulated for rotating, turbulent convective systems, the basic idea behind Parker's topological argument is that the generation of an α-current is due to the lack of symmetry of the flow, and this can also be achieved by the combination of rotation and buoyancy instability. The picture of a dynamo working with strong (super-equipartition) fields without invoking convective turbulence has been outlined recently. In this dynamo model the flux tubes are stored in mechanical equilibrium in a layer of overshooting convection at the boundary with the radiative region. Differential rotation (or other mechanisms) intensifies the field strength until a critical value is reached above which buoyancy instabilities set in. The non-axisymmetric perturbations of the flux tubes combined with the Coriolis force give rise to helical waves of growing amplitude which produce an electric current in the direction of the mean ('unperturbed') magnetic field. The resulting α-effect is a 'dynamic' one in the sense that the Lorentz force and its feedback on the velocity field are taken into account (as well as the compressibility of the plasma). Only unstable, non-axisymmetric perturbations contribute to the mean electric current, which is antisymmetric with respect to the equator.

Conclusions and some open questions

From the properties of active regions it can be inferred that there must exist a strong toroidal flux system in the solar convection zone. Magnetic buoyancy poses a dilemma for the efficient operation of a hydromagnetic dynamo: on the one hand, it is necessary for flux emergence to take place, but on the other hand it is so efficient that it leads to a rapid flux loss from the convection zone. A way out from the dilemma is to assume that prior to eruption the magnetic field is stored in a stably stratified layer of convective penetration for a sufficiently long time for the dynamo to operate on it. Whether this storage is already in the form of an ensemble of flux tubes or as a magnetic layer of 'diffuse field' from which the flux tubes thread out by some sort of MHD instability is still an open question.

In any case, the field is already in the form of intense flux tubes when it enters the convection zone and buoyantly rises through it to emerge as sunspots at the solar surface. A stability analysis yields values of the order of 10^5 G prior to eruption. There are independent arguments in favor of the intense (super-equipartition) fields. One is that a flux tube of equipartition strength at the bottom of the convection zone would become extremely weak when rising across several scale-heights through the turbulent convection zone, contrary to the coherent and concentrated form of bipolar active regions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observed at the solar surface. Another is that the results of the numerical simulations are in agreement with the observed basic properties of active regions only for field strengths of the order of 10^5 G.

Although the production of the strong toroidal field is localized at the lower boundary of the convection zone, the exact process of intensification is not quite understood. It is possible that the explosion process or the formation of Ω-loops operate in conjunction with differential rotation (Moreno-Insertis *et al* 1995). In any case, the existence of super-equipartition fields requires a modified dynamo approach. A possible scenario is the one depicted above: a dynamo based on the instability of strong magnetic flux tubes. Alternatively, one could imagine a dynamo with spatial separation of the generation processes for the poloidal and toroidal components.

A complete theory encompassing in a self-consistent way the generation, structure, dynamics and evolution of magnetic fields in stellar convection zones does not exist; this lack of self-consistency is basically due to our insufficient comprehension of turbulent flows.

Bibliography

Caligari P, Moreno-Insertis F and Schüssler M 1995 Emerging flux tubes in the solar convection zone. I: Asymmetry, tilt, and emergence latitude *Astrophys. J.* **441** 886–902

Caligari P, Schüssler M and Moreno-Insertis F 1998 Emerging flux tubes in the solar convection zone. II: The influence of the initial conditions *Astrophys. J.* **502** 481–92

Fisher G H, Fan Y, Longcope D W, Linton M G and Pevtsov A A 2000 The solar dynamo and emerging flux *Solar Phys.* **192** 119

Moreno-Insertis F 1994 The magnetic field in the convection zone as a link between the active regions on the surface and the field in the solar interior *Solar Magnetic Fields* ed M Schüssler and W Schmidt (Cambridge: Cambridge University Press) pp 117–35

Moreno-Insertis F, Caligari P and Schüssler M 1995 'Explosion' and intensification of magnetic flux tubes *Astrophys. J.* **452** 894–900

Schüssler M, Caligari P, Ferriz-Mas A and Moreno-Insertis F 1994 Instability and eruption of magnetic flux tubes in the solar convection zone *Astron. Astrophys.* **281** 69–72

Zwaan C 1992 The evolution of sunspots *Sunspots—Theory and Observations (NATO Advanced Research Workshop)* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) pp 75–100

Antonio Ferriz-Mas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Emerging Magnetic Flux Tubes

The magnetic field observed in the photosphere and at higher levels of the solar atmosphere is generated in the convective envelope that extends along the first 200 000 km below the solar surface (see DYNAMOS: SOLAR AND STELLAR and SOLAR INTERIOR). To properly understand the solar magnetic cycle, it is necessary to explain the processes whereby the magnetic field, once generated, emerges across the convection zone and then appears at the surface. In fact, when we observe the birth at the solar surface of sunspots and, more generally, active regions, which are structures with an intense magnetic field, we are witnessing the final stages of the emergence of magnetic flux from the solar interior. Looking at the solar disk face-on, what we see after some hours of eruption of magnetic flux is a simple bipolar structure consisting of two patches of mutually opposite magnetic polarity, each containing one or more sunspots (see SOLAR ACTIVE REGIONS and SOLAR ACTIVITY COMPLEXES). The imaginary line linking the two patches is generally oriented in an approximate east–west direction. After the first stages of formation, the two patches separate from each other, reaching a mutual distance of up to 150 000 km. This pattern of behavior seems to be followed by most incipient active regions in the Sun. A simple interpretation of this process is that, when an active region is formed, we are witnessing the emergence at the surface of a large magnetic tube (or, rather, a bunch of them) oriented in a direction roughly parallel to the equator (figure 1). The *roots* of this magnetic tube are located at some deep level below the surface. By studying the dynamics of magnetic tubes in the convection zone including, in particular, why and how they rise and appear at the surface and comparing the conclusions with the actual observations of the birth of active regions, we can therefore gain precious insights into the structure and dynamics of the solar interior.

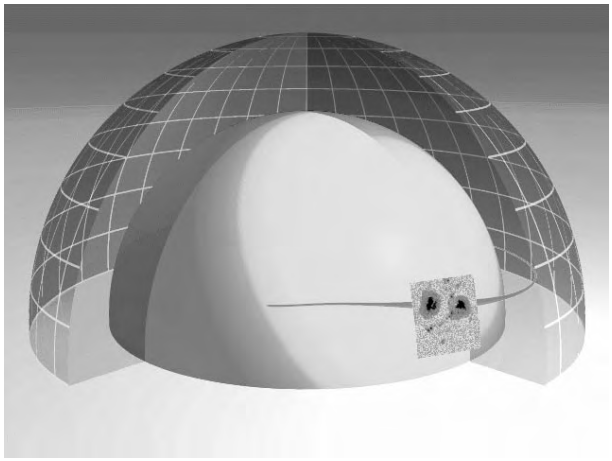


Figure 1. Illustration of a magnetic tube rooted in the deep solar interior whose apex has emerged through the convection zone producing an active region at the surface. The transparent spherical surface with a grid represents the solar surface and the inner solid sphere corresponds to the solar interior below the convection zone. (Figure courtesy of P Caligari.)

The magnetic field in the deep convection zone and the instabilities that trigger its eruption

According to current understanding, the magnetic field that we see in active regions is originated and stored in a magnetized sheet at the bottom of the convective envelope, possibly with the shape of a spherical shell or a spherical band extending on both sides of the equator roughly coinciding with the region known as the *overshoot layer* right beneath the convection zone. A simple possibility (favored by many theoreticians, who are looking for simplicity wherever possible) is that the magnetic field, instead of permeating the whole sheet, adopts the form of a collection of magnetic rings or *hoops* around the Sun, the axes of which roughly coincide with the solar rotation axis; in between the rings the plasma has a much weaker magnetic field. Another possibility, on the opposite extreme, is that the field is continuous and fills up a large fraction of the sheet instead of being concentrated into tubes or rings. In either case, the magnetic field in the sheet does not stay put forever: at some point, it becomes unstable through a process of a kind collectively known as *buoyant instabilities*. The basic idea is that, under a variety of physical conditions, a portion of matter in the solar interior or atmosphere, if given an initial kick upward, will not stop rising because it becomes increasingly light with respect to its surroundings (and thus, also increasingly buoyant).

In the presence of an ordered magnetic field, a classical example of buoyant instability is the *Parker instability*. To understand how it operates, imagine a magnetic tube lying horizontally in the deep convection zone (and, for a graphical image, compare it with a bunch of flexible plastic tubes filled with gas lying at the bottom of a gas tank). If a certain stretch of the magnetic tube is moved upward, then, under given conditions for the temperature distribution in the region, the matter within the tube moves in such a way that the apex of the deformed stretch becomes increasingly light with respect to its surroundings. In this way, the initial upward movement becomes amplified and does not stop. The conditions under which this unstable process takes place are fulfilled for isolated magnetic tubes in the deep solar convection zone if their field strength is above a threshold of roughly 100 000 G (some 300 000 times the strength of the Earth's surface magnetic field). The field strength may indeed be this high at the bottom of the solar convection zone; the Parker instability is thus one of the fundamental mechanisms that can explain why the magnetic field leaves the deep solar interior and erupts at the solar surface. In the alternative view, with the magnetic field permeating a whole sheet instead of being in the form of individual tubes, there is also a mechanism for triggering the eruption of magnetic flux, called the magnetic Rayleigh–Taylor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

instability. This instability (akin to the instability that ensues when heavier fluid rests on a less dense one) may lead to the disruption of the magnetic structure and the formation of rising buoyant magnetic tubes.

Modeling the inner workings of the Sun in a computer

As a result of the buoyant movement of the magnetic tubes, magnetic flux is brought all the way from the bottom of the convection zone to the surface, where we see it appearing as sunspots and other magnetic elements. If we could contemplate the solar interior, then we would see how the magnetic tube stretches from the bottom of the convection zone, where it is rooted, all the way to the surface and into the solar atmosphere. Sunspots are, in this sense, just a *cut* across the flanks of an erupted magnetic tube.

In the past two decades there has been a true *revolution* in our understanding of the detailed physics of the process of rise of the magnetized plasma, thanks, among other things, to the possibility of detailed numerical simulation using the impressive increase in computing power and the development of simplified models. A first generation of models (the so-called *slender flux tube models*) use the fact that the radius of the magnetic tube progenitors of active regions (possibly some 1000 km only) is very small with respect to the size of the solar structures at the bottom of the convection zone. The slender flux tube models use a drastically simplifying approach; given that the tubes are thin, one can, in a first go at the problem, neglect the complicated transverse structure of the tubes in each cross section. There results a self-consistent set of equations, the *thin flux tube equations*, which contain the most essential ingredients to describe why the tubes rise and what the main features of the resulting structure are. The solution of those equations yields a simple model for the process of emergence of flux.

One such model can be seen in figures 2 and 3. The plasma in the rising loop of the tube experiences a huge expansion as it adapts to the new environments it crosses in its upward journey: the pressure of the solar plasma is a billion times higher at the bottom of the convection zone than at its topmost levels. This causes a marked weakening of the field intensity along the rise. On the other hand, as a result of the unstable evolution, the lower reaches of the tube descend to somewhat deeper levels and become anchored in the overshoot region right beneath the convection zone. The matter in the *overshoot region* is being stirred because of the intrusion of convection cells from the overlying layers, but, in itself, is stable to convection. A magnetic tube in sufficiently low levels in that region is stable to the buoyant instabilities and remains there while the rising sections of the tube are on their way to the surface.

The thin tube models have provided explanations for some of the observed large-scale features of active regions. One of them concerns an old observational result: Hale, Ellerman, Nicholson and Joy observed in 1919 that the *main axis* of an active region (i.e. the imaginary line linking the centers of the two patches of opposite polarity) and the solar equator were not strictly parallel but, rather, subtended an angle of, typically, several degrees with opposite sign in the northern and southern hemisphere. The magnitude of this angle was larger the higher the latitude at which the active region appeared. The thin tube models explain the origin of this empirical law as a consequence of the rotation of the Sun or, more precisely, through the action of the Coriolis force. This can be understood by comparing this phenomenon with the formation of an anticyclone in the Earth's atmosphere when air moves away from a high-pressure area (which is also a consequence of the Coriolis force). Studying the motion of the individual mass elements of the tube along the rise, one can see that those along the flanks move sideways away from the mass element at the apex. The Coriolis force thus brings about a rotation of the vertical plane containing the magnetic tube. When the tube reaches the surface, this is seen as a *tilt* of the main axis of the active region (clearly visible in figures 1 and 2).

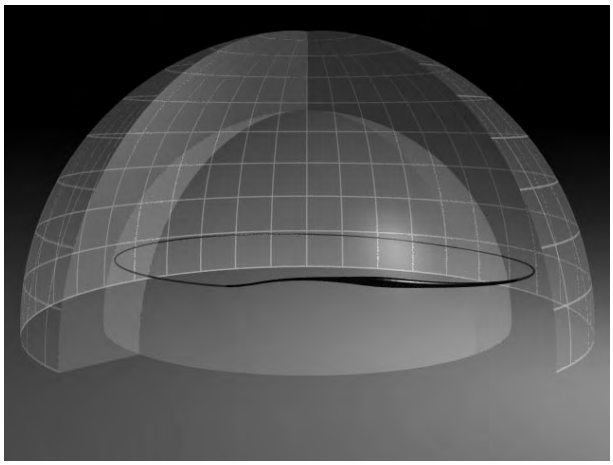


Figure 2. Result of a time-dependent calculation of emergence of magnetic flux from the bottom of the convection zone. The equations that govern the time evolution of a thin magnetic tube have been solved starting with a Parker-unstable configuration.

A further empirical law (in this case, a much more recent one) concerning active regions is the fact that the mutual separation of two patches of opposite polarity does not proceed in a symmetrical way: it appears that one side (the one closest to the western limb of the Sun) moves away from the line separating the opposite polarities much faster than the other side. This phenomenon can also be explained by the thin tube models and is, again, a consequence of the solar rotation: because of the conservation of angular momentum, the two flanks of the rising loop become differently inclined with respect to the vertical direction (figure 3). When the tube breaks out at the photosphere, this difference in inclination translates into a higher speed of separation of one of the branches of the rising loop from the apex than the other branch.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

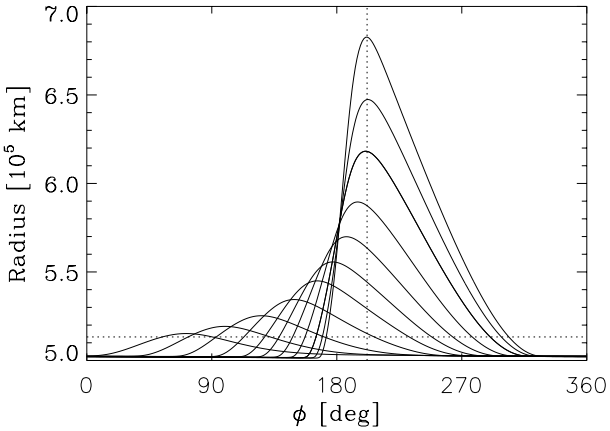


Figure 3. Several stages of the evolution of a rising magnetic tube with the horizontal axis representing solar longitude and the vertical axis height (measured by the radial distance to the solar center). Note the asymmetry in inclination of the flanks of the tube.

Perhaps the most interesting result of the thin flux tube models is an estimate of the strength of the magnetic field at the bottom of the convection zone. For the thin tube models to yield a good agreement with the empirical observations of active regions discussed above, the field intensity in the bottom layers must be of the order of 100 000 G. This result came as a surprise when it was first clearly stated since most people thought it to be much weaker, some 1000 to 10 000 G. That the field must be as strong as 100 000 G can be understood in different ways: first, if the tubes in the deep convection zone had weaker fields, then, upon the huge expansion taking place along the rise, they would yield very weakly magnetized regions rather than concentrated tubes. Second, the risetime of magnetic tubes with weaker field strength is much longer, so that the Coriolis force has a much larger impact on the rising tube. Consequently, the resulting active regions would have features not seen in the actual Sun like: (a) the eruption at the surface would take place at high solar latitudes, (b) their main axes would systematically subtend a large angle to the equator and (c) the asymmetric motion of the polarities discussed in the previous paragraph would be extreme.

Recent research: twists, buckles, inner structure and aerodynamic properties of a rising magnetic tube

In recent years, the research on the emergence of magnetic flux has focused on issues concerning the transverse structure and the three-dimensional stability of the rising magnetized regions. One of the issues deals with the question of whether a magnetic tube that is rising across the convection zone can maintain its unity, i.e. if it can rise without splitting into smaller strands. To understand the problem, we can recall the case of rising columns of hot air in the Earth's atmosphere, typically occurring on sunny summer days on a plain: the rising column of air typically turns into two horizontal rolls (also called vortex tubes) which rotate in opposite directions. Similarly, a buoyantly rising magnetic tube can easily split into two horizontal vortex tubes. Vortex tubes suffer aerodynamic lift forces, so that their path may deviate significantly from the simple path calculated with the standard thin flux tube model. A branch of research has therefore been devoted to studying the details of the internal structure of the tube and of the surrounding boundary layer. This includes the generation of vorticity, especially around the tube boundary, the formation of a trailing wake, the aerodynamic lift and drag experienced by the tube, etc. A major result of these studies is that the magnetic field lines must be *twisted* for the tube to preserve its unity during the rise, i.e. they must be wound around the tube's main axis instead of running more or less parallel to it. The magnetic field thus has both a component parallel to the tube's axis *and* a component contained within the cross section of the tube. It is the Lorentz force associated with the latter that gives the tube the possibility to withstand the forces that tend to break it into vortex rolls.

Another related issue subject to scrutiny in recent years is the development of kink instabilities (well known from the plasma fusion research in laboratories) in magnetic tubes with twisted field lines in the solar interior. The instability has to do with the fact that the magnetic field exerts a tension along its lines. As a consequence, if the tube is sufficiently twisted it may *buckle* and produce *kinks*, like a simple rope subject to tension and twist. This is being studied with the help of computer simulations. An interesting result of this research is a possible explanation for the peculiar behavior of a special class of active regions, the so-called *δ-spots*. *δ*-spots are exceptional active regions with the two patches of opposite polarity close together and with the main axis of the region inclined at high angles with respect to the solar equator. These peculiar features may be explained as the arrival at the surface of a strongly twisted magnetic tube in an advanced stage of development of the kink instability. An easy pictorial image for this is a highly twisted rope which is quickly developing buckles: the direction of the rope in the buckles may deviate considerably from the original direction of the unkinked rope.

In parallel to the research into the behavior of individual magnetic tubes, there have been several studies of the physics of continuous magnetic field distributions. The buoyant instabilities in the latter are in several ways similar to the magnetic tube counterparts. However, the techniques for studying them are different. For instance, there is no equivalent to the thin tube approximation, so that there are to date no full simulations of the eruption of a continuous magnetic layer in the lower convection zone all the way until the up-going elements reach the solar surface. However, these studies have provided interesting insights into how magnetic flux tubes may form out of a continuum of field, the subsequent three-dimensional instabilities, etc.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

All of the foregoing studies require the use of the largest and most advanced computers available: if the pace of increase in computing power continues at the current rate, we can expect substantial advances in our understanding of the physics of the magnetic field in the solar interior and its emergence to the surface.

Bibliography

Caligari P, Moreno-Insertis F and Schüssler M 1995 Emerging flux tubes in the solar convection zone. I Asymmetry, tilt and emergence latitude *Astrophys. J.* **441** 886–902

D'Silva S and Choudhuri A R 1993 A theoretical model for tilts of bipolar magnetic regions *Astron. Astrophys.* **272** 621–33

Emonet T and Moreno-Insertis F 1998 The physics of twisted magnetic tubes rising in a stratified medium: two-dimensional results *Astrophys. J.* **492** 804–21

Fan Y, Zweibel E G, Linton M G and Fisher G H 1999 The rise of kink-unstable magnetic flux tubes and the origin of δ-configuration spots *Astrophys. J.* **521** 460–77

Moreno-Insertis F 1997 Emergence of magnetic flux from the solar interior *The Inconstant Sun* (*Mem. Soc. Astron. It.* **68**) ed G Cauzzi and C Marmolino pp 429–47

Fernando Moreno-Insertis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Energy Generation

Energy generated deep in the SOLAR INTERIOR via the nuclear fusion of hydrogen into helium provides the 3.9×10^{26} J s^{-1} radiated by the surface of the Sun. Beginning in the 1920s the foundations of nuclear astrophysics were laid on the basis of (a) the recognition that the solar system was several billion years old (based on the decay of radioactive elements such as uranium), (b) the recognition that, in contrast with inadequate chemical and gravitational energy sources, the then newly discovered nuclear reactions could readily supply the energy required to account for the luminosity of the Sun over this lifetime and (c) the recognition by Gamow in the late 1920s of the role of quantum mechanical tunneling in making it possible for these reactions to occur in the solar interior where kT was most probably equivalent to only a few keV, well below the repulsive Coulomb barrier. Building on that foundation, in the late 1930s, Weizsäcker and Bethe independently laid out the energetics and many of the theoretical details for converting hydrogen to helium in the solar interior via the pp chain and the CN cycle.

At the temperatures and densities characteristic of the solar interior ($T \approx 16 \times 10^6$ K and $\rho \approx 150$ g cm^{-3}), interacting nuclei reach a Maxwellian equilibrium distribution in a time that is infinitesimal compared with the mean lifetime of these nuclear reactions and decays. Therefore it is possible to determine the thermally averaged nuclear reaction rate by folding its energy-dependent cross section $\sigma(E)$ with the Maxwell–Boltzmann distribution as

$$R_{12} = \frac{n_1 n_2}{1+\delta_{12}} \langle \sigma \nu \rangle_{12}$$

where

$$\langle \sigma \nu \rangle_{12} = \left(\frac{8}{\pi \mu (kT)^3} \right)^{1/2} f_0 \int_0^\infty E \sigma(E) \exp(-E/kT)\, \mathrm{d}E$$

and where f_0 is a correction to the laboratory (bare-nucleus) cross section measurements to take into account the effects of plasma screening in STELLAR INTERIORS. (This correction is the order of 1.20–1.25 for pp-chain reactions in the solar interior.) As noted above, in the solar interior nuclear reactions take place via tunneling at energies which are far below the Coulomb barrier. One of the major sources of the energy dependence of the cross sections arises from this tunneling whose principal effects can be factored out by reexpressing the cross section in terms of the astrophysical S factor,

$$S(E) = \sigma(E) E / \exp(2\pi\eta)$$

where $\exp(-2\pi\eta)$ is the usual quantum mechanical penetration term associated with tunneling under the Coulomb barrier, in which $\eta = Z_1 Z_2 e^2/\hbar\nu$, so that

$$\langle \sigma \nu \rangle_{12} = \left(\frac{8}{\pi \mu (kT)^3} \right)^{1/2} f_0 \int_0^\infty S(E) \exp(-2\pi\eta) \times \exp(-E/kT)\, \mathrm{d}E. \quad (1)$$

An example of the comparison between the energy dependences of $\sigma(E)$ and $S(E)$ is shown in figure 1, for the case of the ^{3}He(^{3}He,2p)^{4}He reaction; in the energy range plotted in this figure, 20 keV$\leq E_{\mathrm{cm}} \leq$ 350 keV, $\sigma(E)$ varies by nearly a factor of 10^{10} while $S(E)$ varies by only about 20%. For reactions in which there are no important low-energy resonances so that $S(E)$ is only a slowly varying function of energy (which is the case for all the pp-chain reactions), the dominant energy dependence in the integrand in equation (1) is then given by the product of barrier-penetration and Maxwell–Boltzmann terms, shown schematically in figure 2. The centroid E_0 of the peak of the integrand (the 'Gamow peak') and ΔE_0 (its width at 1/e of its maximum value) are given by

$$E_0 = [\pi (Z_1 Z_2 e^2/hc) kT (\mu c^2/2)^{1/2}]^{2/3}$$
$$\text{and } \Delta E_0 = 4(E_0 kT/3)^{1/2}.$$

Because of the effects of the Coulomb barrier, laboratory measurements of the cross sections for these reactions can typically be carried out only at energies well above E_0. Because of its much smaller energy dependence, the $S(E)$ function is then frequently a convenient way to extrapolate these measurements to the lower energies of interest in the $\langle \sigma \nu \rangle_{12}$ integral in equation (1). It is usually convenient to characterize the energy dependence of $S(E)$ in terms of its intercept and derivatives at $E = 0$: $S(0)$, $S'(0)$ and $S''(0)$.

$$S(E) = S(0) \left(1 + \frac{S'(0)}{S(0)} E + \frac{1}{2} \frac{S''(0)}{S(0)} E^2 \right). \quad (2)$$

For a more detailed discussion of these rate equations, the reader should consult one of the standard reviews listed in the bibliography at the end of this article.

The proton–proton chain

The pp chain (figure 3) is responsible for ≈98% of the $4\,^1\mathrm{H} \rightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e$ nuclear-fusion energy generation in the Sun. The critical reactions in the chain are (i) the $^1\mathrm{H} + {}^1\mathrm{H} \rightarrow {}^2\mathrm{D} + e^+ + \nu_e$ reaction which is the slowest link in the pp chain and whose rate therefore determines the central temperature of the Sun, (ii) the ^{3}He(^{3}He,2p)^{4}He and ^{3}He(α, γ)^{7}Be reactions whose relative rates determine the fraction of pp-chain terminations which produce neutrinos with energies greater than the $E_{\nu\,\mathrm{max}} = 420$ keV from the pp reaction and (iii) the ^{7}Be(e^-, ν_ε)^{7}Li and ^{7}Be(p,γ)^{8}B reactions whose relative rates determine the fraction of terminations which result in the production of high-energy neutrinos, $E_{\nu\,\mathrm{max}} \approx 14$ MeV. The three terminations for the pp chain (as well as the various permutations of the CN cycle) all generate the same total energy (25.731 MeV, which is simply the mass difference between $4\,^1$H and ^{4}He). However, the amount of that energy lost directly from the center of the Sun via the two neutrinos can be quite different, varying from 0.420 MeV for the ^{3}He(^{3}He,2p) termination to ∼7 MeV for the ^{7}Be(p,γ) termination. (For a more complete discussion see SOLAR INTERIOR: NEUTRINOS.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

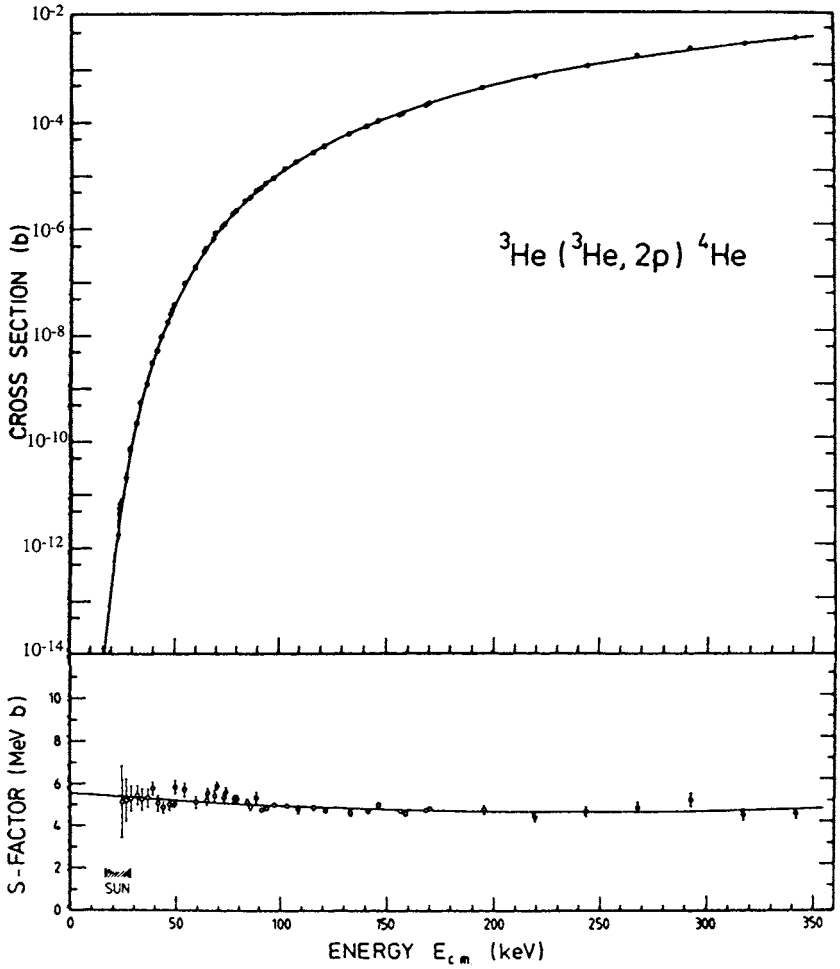


Figure 1. Comparison of the energy dependences of $\sigma(E)$ and $S(E)$ for the ^{3}He(^{3}He,2p)^{4}He reaction.

The ^1H+$^1H\rightarrow$ ^2D+e^+ + ν_e reaction

The ^{1}H +^{1} H $\rightarrow$ ^{2}D + e$^+$ + ν_e reaction takes place via the weak interaction and is by far the slowest link in the pp chain; its $S_{11}(0)$ intercept is roughly 10^{20} times smaller than the intercepts for the other reactions in the chain. Its cross section is not currently measurable in laboratory experiments and must instead be determined on the basis of measurements of the properties of the weak interaction—for example, measurements of the beta decay of the neutron and of superallowed $0^+ \rightarrow 0^+$ beta decays. On this basis, the astrophysical S factor for this reaction can be written as

$$S_{11}(E) = 6\pi^2 m_p c\alpha \ln 2 \frac{\Lambda^2}{\gamma^3}\left(\frac{G_A}{G_V}\right)^2 \frac{f_{pp}^R}{(ft)_{0^+\rightarrow 0^+}}(1+\delta)^2$$

where Λ is the overlap integral between the pp and deuteron wavefunctions, γ is the deuteron binding wavenumber, G_A and G_V are the usual axial-vector and vector weak coupling constants, f_{pp}^R is the phase space factor for the pp reaction, $(ft)_{0^+\rightarrow 0^+}$ is the ft value for superallowed $0^+ \rightarrow 0^+$ beta decay and δ is the fractional correction to the nuclear matrix element due to exchanges of π and ρ mesons. Utilizing the current best values for these parameters yields the following values for the $S_{11}(0)$ intercept and its energy derivative, $S'_{11}(0)$:

$$S_{11}(0) = 4.00\times10^{-22}\,\text{keV b} \qquad \text{and} \qquad S'_{11}(0) = 4.5\times10^{-24}\,\text{b}.$$

Currently the largest source of uncertainty in this reaction rate is in the parameter δ.

The ^{3}He(^{3}He,2p)^{4}He reaction

Since ^{4}Li is particle unstable, one important aspect of the ^{3}He(^{3}He,2p)^{4}He reaction is the fact that it provides a termination for the pp chain without requiring the presence of previously synthesized ^{4}He. Recently, the LUNA facility (Laboratory for Underground Nuclear Astrophysics) has been used to carry out measurements of the ^{3}He(^{3}He,2p)^{4}He reaction down to the region of the Gamow peak. (The LUNA Collaboration includes scientists from Italy, Germany and the US.) Using the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

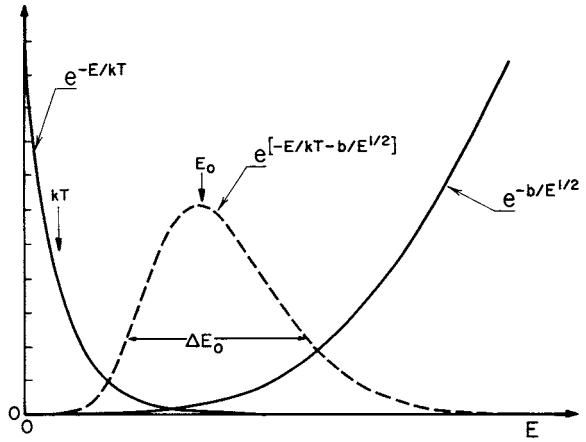


Figure 2. Energy dependence of the integrand in equation (1).

Gran Sasso tunnel in the the Appennine Mountains in central Italy, as a cosmic ray shield, it has been possible to measure cross sections as small as 0.76 pb and to thereby carry out direct measurements of this reaction down to $E_{\rm cm} = 20.8$ keV. Together with earlier above-ground measurements of the Bochum group, these measurements cover more that half of the actual Gamow peak for this reaction, $E_0 \approx 22$ keV at the center of the Sun.

In measurements made for $E_{\rm cm} \geq 100$ keV the distance of closest approach for these nuclei is inside the electron cloud so that their interaction is as essentially bare nuclei. In contrast, at the low energies used in the LUNA measurements described above, the interacting nuclei have a distance of closest approach that is near or outside the atomic radius $R_{\rm a}$, so that electron screening corrections are necessary in order to relate these laboratory measurements $\sigma_{\rm sc}$ to the cross sections that would be measured for bare nuclei $\sigma_{\rm b}$. This screening correction is given by

$$\sigma_{\rm b} = \sigma_{\rm sc} \exp(-\pi\eta U_{\rm e}/E_{\rm cm}) \qquad \text{where } U_{\rm e} = Z_1 Z_2 e^2/R_{\rm a} \tag{3}$$

and where η is the usual Coulomb parameter. For the ^{3}He(^{3}He,2p)^{4}He reaction ($U_{\rm e} = 240$ eV), at $E_{\rm cm} = 22$ keV this corresponds to a 20% correction, while at $E_{\rm cm} > 150$ keV the correction is $\leq$1%. Note that this correction should not be confused with the plasma screening correction which must be used to relate laboratory measurements of bare-nuclear cross sections to reaction rates in plasmas.

A fit (see figure 4) to the LUNA and Bochum data sets (which are consistent with each other and with earlier higher-energy measurements), including the screening correction of equation (3), determines the following bare-nucleus values for the $S_{33}(0)$ intercept and for $S_{33}(E_0)$ at the center of the Gamow peak:

$$S_{33}(0) = 5.4 \pm 0.4 \text{ MeV b}$$
$$\text{and} \qquad S_{33}(E_0) = 5.3 \pm 0.4 \text{ MeV b}.$$

The ^{3}He(α, γ)^{7}Be reaction

The relative rates of the ^{3}He(^{3}He,2p)^{4}He and ^{3}He(α, γ)^{7}Be reactions determine the fraction of pp-chain terminations which produce neutrinos with energies greater than the $E_{\nu\,\rm max} = 420$ keV available from the pp reaction. While there have been nine independent sets of measurements of the cross section for the ^{3}He(α, γ)^{7}Be reaction, because the cross section for this reaction is approximately 10 000 times smaller than the cross section for the ^{3}He(^{3}He,2p)^{4}He reaction none of these measurements has been carried out close to the Gamow peak for this reaction. Therefore, the rate of this reaction in the solar interior must be determined from an extrapolation of the higher-energy measurements. Extrapolation is always dangerous, but fortunately the energy dependence of this reaction is well described by a variety of nuclear models (ranging from ^{3}He$\otimes^4$He cluster models to microscopic potential models) which agree within a per cent or two and which can then be used to carry out the extrapolation with confidence. The good agreement between these calculations arises from the fact that over the energy range 0 keV$\leq E_{\rm cm} \leq$ 2500 keV this reaction is a non-resonant direct-capture process for which almost all the contributions to the matrix element come from the region outside the nuclear potential well. Figure 5 displays the agreement between such calculations and the measured data sets from the four experiments which cover significant energy ranges; for one of the experiments the data extend up to $E_{\rm cm} = 2500$ keV (still well below the first unbound ^{7}Be state at $E_{\rm cm} = 2983$ keV) with the same quality of agreement between the model calculations and the measurements.

A weighted mean of the extrapolation of the nine independent experiments determines the intercept $S_{34}(0)$, while the theoretical calculations determine the slope $S'_{34}(0)$:

$$S_{34}(0) = 0.53 \pm 0.05 \text{ keV b}$$
$$\text{and} \qquad S'_{34}(0) = -3.0 \times 10^{-4} \text{ b}.$$

In spite of the fact that $S_{34}(0)$ is $\approx 10^4$ times smaller than the corresponding $S_{33}(0)$, the ^{3}He(α, γ)^{7}Be reaction can compete with the ^{3}He(^{3}He,2p)^{4}He reaction in the solar interior because of the large abundance of ^{4}He in the solar interior relative to ^{3}He; ^{4}He/^{3}He$\sim 10^3$–10^5, depending on the location. Integrated over the solar interior, solar model calculations find that 16% of the pp-chain terminations occur via the ^{3}He(α, γ)^{7}Be reaction while the remaining 84% occur via the ^{3}He(^{3}He,2p)^{4}He reaction.

The ^{7}Be(e^-,ν_e)^{7}Li decay

For the 16% of the pp-chain terminations that proceed through the ^{3}He(α, γ)^{7}Be reaction, the branch between ^{7}Be electron capture and ^{7}Be proton capture determines the fraction of terminations which produce high-energy neutrinos ($E_{\nu\,\rm max} \approx 14$ MeV) via the beta-decay ^{8}B. In this way, the relative rates of the ^{7}Be(e^-, γ_e)^{7}Li decay and the ^{7}Be(p,γ)^{8}B reaction are critical in determining both

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$4\,^1\mathrm{H} \rightarrow\ ^4\mathrm{He} + 2e^+ + 2\nu_e$	$Q = 25.731$ MeV	
$^1\mathrm{H} +\,^1\mathrm{H} \rightarrow\ ^2\mathrm{D} + e^+ + \nu_e$	1.442 MeV	$E_{\nu\,\max} = 0.420$ MeV
$^1\mathrm{H} + e^- +\,^1\mathrm{H} \rightarrow\ ^2\mathrm{D} + \nu_e$	1.442 MeV	$E_\nu = 1.442$ MeV
$^2\mathrm{D} +\,^1\mathrm{H} \rightarrow\ ^3\mathrm{He} + \gamma$	5.493 MeV	
(a) $^3\mathrm{He} +\,^3\mathrm{He} \rightarrow 2\,^1\mathrm{H} +\,^4\mathrm{He}$	12.859 MeV	
(b) $^3\mathrm{He} +\,^4\mathrm{He} \rightarrow\ ^7\mathrm{Be} + \gamma$	1.587 MeV	
(b1) $^7\mathrm{Be} + e^- \rightarrow\ ^7\mathrm{Li} + \nu_e$	0.862 MeV	$E_\nu = 0.862$ MeV (89.5%)
$^7\mathrm{Li} + \mathrm{p} \rightarrow 2\,^4\mathrm{He}$	17.346 MeV	$= 0.384$ MeV (10.5%)
(b2) $^7\mathrm{Be} +\,^1\mathrm{H} \rightarrow\ ^8\mathrm{B} + \gamma$	0.135 MeV	
$^8\mathrm{B} \rightarrow\ ^8\mathrm{Be}^* + e^+ + \nu_e$	14.939 MeV	$E_{\nu\,\max} \approx 13.9$ MeV
$^8\mathrm{Be}^* \rightarrow 2\,^4\mathrm{He}$	3.132 MeV	

Figure 3. Proton–proton chain.

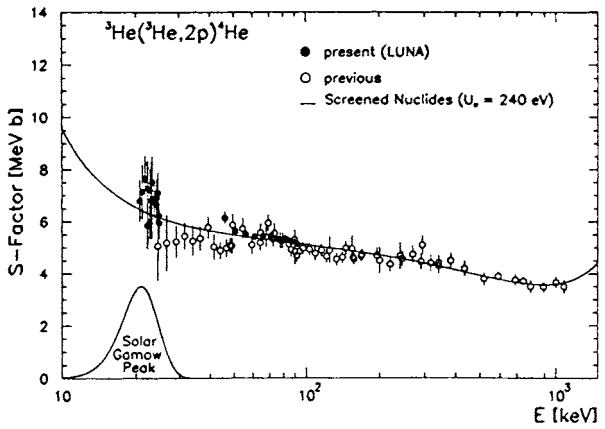


Figure 4. Low-energy measurements of $S(E)$ for ^{3}He(^{3}He,2p)^{4}He, showing the effects of the electron screening described in equation (3).

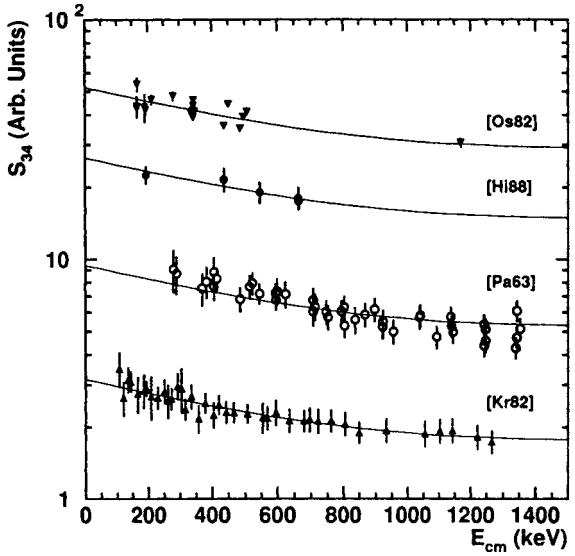


Figure 5. Comparison of the energy dependence of a theoretical calculation of $S(E)$ with the measured data sets from four independent ^{3}He(α, γ)^{7}Be experiments.

the effective energy generation of the pp chain and the resulting neutrino energy spectrum.

The decay of ^{7}Be via the capture of its own atomic electrons has been well studied in the laboratory; it has a measured half-life of 53.29 ± 0.07 days, with a branching ratio of 89.5% for decays to the ground state of ^{7}Li ($E_\nu = 0.862$ MeV) and 10.5% to the first exited state of ^{7}Li ($E_\nu = 0.384$ MeV). However, in the solar interior ^{7}Be atoms will be mostly ionized, so that any ^{7}Be(e^-, ν_e)^{7}Li decays will occur mostly via the capture of continuum electrons from the plasma, and the resulting decay rate will be proportional to the electron density in the plasma. Calculations of the continuum decay rate combined with calculations of the bound-state electron populations and their contributions to the decay rate determine

$$1/\tau\ (\mathrm{s}^{-1}) = (1/\tau)_\mathrm{c} + (1/\tau)_\mathrm{b} = 5.60 \times 10^{-9}\,\frac{\rho(1+X_\mathrm{H})}{2}\,T_6^{-1/2}[1 + 0.004(T_6 - 16)].$$

Integrated over the solar interior, this corresponds to 82% of the decays coming from the capture of continuum electrons from the plasma and 18% from bound atomic electrons. Under conditions typical of the hydrogen burning region in the solar interior ($\rho = 100$ g cm^{-3}, $X_\mathrm{H} = 0.50$, $T_6 = 13.8$), this corresponds to a ^{7}Be half-life of 70 days (compared with the laboratory value of 53 days).

The ^{7}Be(p,γ)^{8}B reaction

The fraction of pp-chain terminations proceeding through the ^{7}Be(p,γ)^{8}B reaction is especially significant (i) because of the production of high-energy (^{8}B) neutrinos which are crucial for a number of solar neutrino detectors (see NEUTRINO ASTRONOMY) and (ii) because of the ≈27% energy loss associated with the escape of those neutrinos from the center of the Sun. The rate of the ^{7}Be(p,γ)^{8}B reaction is therefore particularly important for probing and understanding energy generation in the solar interior. In spite of the significant added complications associated with the radioactive nature of ^{7}Be, seven such experiments have been carried out using ^{7}Be targets to measure the rate of this reaction. This is another case in which the $S_{17}(0)$ intercept must be determined by extrapolation from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

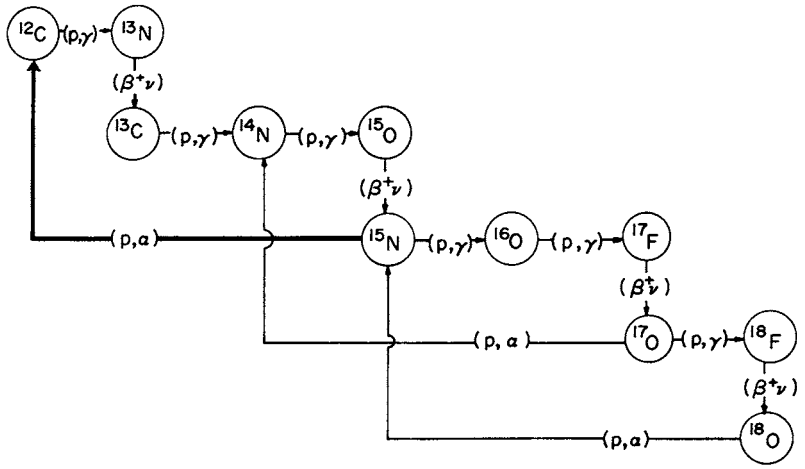


Figure 6. The CNO tri-cycle.

values measured at energies well above the Gamow peak. In this case the low-energy cross section is again non-resonant, and there are a number of consistent theoretical calculations which provide a basis for the extrapolation. The full range of the $S_{17}(0)$ results from these experiments covers a variation of a factor of $\approx$2 between the highest and lowest values. Following a critical review of all of these experiments, the 1997 INT Workshop on Solar Fusion Cross Sections recommended the adoption of

$$S_{17}(0) = 19^{+4}_{-2}\ \text{eV b}. \qquad (4)$$

A quadratic fit to the theoretical energy dependence over the range 0 keV $\leq E_{\rm cm} \leq$ 300 keV, using the form of equation (2), determines

$$\frac{S'(0)}{S(0)} = -0.7\ \text{MeV}^{-1} \qquad \text{and} \qquad \frac{S''(0)}{S(0)} = +3.8\ \text{MeV}^{-2}.$$

These results, integrated over the solar interior, determine that the ^{7}Be(p,γ)^{8}B termination accounts for only 0.02% of the pp-chain terminations. While this is clearly not an important consideration as far as solar energy generation is concerned, because of its high-energy neutrinos this branch is still the most important termination for a number of solar neutrino experiments. The interpretation of the results of those experiments therefore requires that $S_{17}(0)$ be more accurately measured.

We appear to be on the threshold of a series of promising new experiments to measure the rate of this reaction, using a variety of new techniques, such as ^{7}Be beams, new ^{7}Be–target strategies and the Coulomb dissociation of ^{8}B. These new techniques will have the advantages of avoiding some of the systematic uncertainties associated with the previous experiments, but at the same time these new techniques will have systematic uncertainties of their own which will need to be understood in order to interpret the results of such experiments together with the previous (p,γ) experiments. Initial measurements utilizing a ^{7}Be beam have recently been reported by the NABONA collaboration, and initial Coulomb dissociation measurements have provided preliminary results which are not inconsistent with equation (4) and are encouraging for further studies to provide an independent determination of $S_{17}(0)$.

The carbon–nitrogen cycle

The CN cycle uses ^{12}C as a nuclear catalyst in the conversion of $4\,^1\text{H} \rightarrow {}^4\text{He} + 2e^+ + 2\nu_e$. Because of the roles of various branch points available in the CN cycle, the whole process (see figure 6) is now described as a 'tri-cycle'. Because of the high Coulomb barriers of the nuclear reactions involved in the CNO cycles (compared with those in the pp chain) and the relatively low temperature in the solar core (compared with more massive stars), the CNO tri-cycle is responsible for only a small fraction ($\leq$2%) of the energy generated in the Sun. Based on a comparison of the measured cross sections for the ^{15}N(p,γ)^{16}O and ^{15}N(p,α)^{12}C reactions, the main cycle, terminating via the ^{15}N(p,α)^{12}C reaction, is responsible for $\geq$99.5% of this small fraction.

By far the slowest reaction in the CN cycle is the ^{14}N(p,γ)^{15}O reaction—typically more than 100$\times$ slower than the other reactions for temperatures between 15×10^6 and 20×10^6 K. It is the rate of this reaction which then determines the rate of the CN cycle. Recent, extensive measurements of the ^{14}N(p,γ)^{15}O reaction show evidence for important interference effects involving the high-energy tail of a subthreshold resonance ($E_{\rm cm} = -504$ keV; $J^\pi = 3/2^+$). However, the magnitude of these interference effects at low energies is not well determined by these recent data. A combination of these data (0.20 MeV $\leq E_{\rm cm} \leq$ 3.6 MeV) and earlier, lower-energy data (0.100 MeV $\leq E_{\rm cm} \leq$ 0.135 MeV) has been used by the 1997 INT Workshop on Solar Fusion Cross Sections to determine a current 'best' value of $S(0)$ for this reaction,

$$S_{^{14}\text{N(p,}\gamma\text{)}}(0) = 3.5^{+0.4}_{-1.6}\ \text{keV b}.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

New experiments are clearly needed in order to better determine the cross section of this reaction and, thereby, the rate of the CN cycle.

Summary

Energy generation in the solar interior represents an intertwining of the effects of the gravitational and nuclear forces. The total gravitational potential energy available from the collapse of the proto-solar nebula to the size of the Sun would be only sufficient to generate the solar luminosity for the order of 50 million years (≈1% of the current age of the Sun). However, the contraction of the proto-solar nebula converted gravitational potential energy to thermal kinetic energy which heated the solar interior until the $4\,^1\mathrm{H} \rightarrow {}^4\mathrm{He}$ nuclear fusion reactions became fast enough to generate the solar luminosity; at that point an equilibrium was established between the nuclear fusion energy supporting the star and the attractive gravitational force containing this fusion device.

Based on the measurements and extrapolations of nuclear cross sections discussed above and their integration over the solar interior in a standard solar model (see SOLAR INTERIOR: STANDARD MODELS), we find the following percentages for each of the significant $4\,^1\mathrm{H} \rightarrow {}^4\mathrm{He}$ processes:

pp chain	98.2%
^{3}He(^{3}He,2p) termination	82.5%
^{7}Be(e^-,ν_e) termination	15.7%
^{7}Be(p,γ) termination	0.02%
CN cycle	1.8%

Bibliography

Adelberger E G *et al* 1998 Solar fusion cross sections *Rev. Mod. Phys.* **70** 1265–91

Burbidge E M, Burbidge G R, Fowler W A and Hoyle F 1957 Synthesis of the elements in stars *Rev. Mod. Phys.* **29** 547–650

Clayton D D 1983 *Principles of Stellar Evolution and Nucleosynthesis* (Chicago, IL: University of Chicago Press)

Parker P D MacD and Rolfs C E 1991 Nuclear energy generation in the solar interior *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M S Matthews (Tucson, AZ: University of Arizona Press) pp 31–50

Rolfs C E and Rodney W S 1988 *Cauldrons in the Cosmos* (Chicago, IL: University of Chicago Press)

Peter Parker

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Equation of State and Opacity

In stellar models, the equation of state and opacity are, together with nuclear reaction rates, the fundamental material properties. The structure of a star is a result of (i) a balance of forces, (ii) a balance between the energy loss at the stellar surface and energy generation in the core and (iii) stationary energy transport between the core and the surface (see SOLAR INTERIOR: STANDARD MODELS). The balance of forces, described as hydrostatic equilibrium, results in a relation between the pressure gradient and the gravitational acceleration. The force of gravity is determined by the density distribution in the star: thus stellar modeling requires a relation between density and pressure through the specific properties of the matter. More precisely, the relevant properties of stellar matter are expressed by the *equation of state*, connecting pressure p, density ρ, temperature T and composition. A simple example is the perfect gas law for a fully ionized gas, which may be written as $p = \mathcal{R}\rho T/\mu$; here $\mathcal{R}$ is the gas constant and μ is the mean molecular weight of the various species contained in the gas. Clearly μ depends on the composition, which is often characterized by the fractional mass abundances X, Y and Z of hydrogen, helium and heavier elements.

The temperature of the stellar interior is determined by the energy balance. In much of the Sun, energy transport takes place through radiation and depends on absorption coefficients, obtained from atomic physics, which determine the *opacity* of stellar matter. It was also realized early on that the requirements of radiative transport could result in a temperature gradient so steep that a star would become convectively unstable: convection, where hot elements of gas rise and then generally dominate the energy transport. In the Sun this occurs in about the outer 30% of the radius. Because of the efficiency of convective energy transport, it generally requires a temperature gradient only slightly in excess of the adiabatic gradient, a thermodynamic quantity derived from the equation of state. Another consequence of the equation of state, adiabatic sound speed, plays a crucial role in helioseismology (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY; HELIOSEISMOLOGY: THEORY).

The equation of state appears as a necessary part of stellar modeling as well as of any opacity calculation. For the latter, it has to provide ionization equilibrium concentrations and level populations. However, the interest in the stellar equation of state is not merely motivated by astrophysics. It has turned out that one star—the Sun—is very special in two respects. First, the methods of helioseismology allow us to infer conditions in the solar interior very accurately (in particular sound, speed and density). Second, in the solar convection zone, helioseismology presents an opportunity to isolate the question of the equation of state from opacity and nuclear reaction rates, since the stratification is essentially adiabatic and thus determined by thermodynamics. Thus, the Sun has become an astrophysicsal laboratory to study subtle thermodynamic properties of a Coulomb system under conditions that cannot be acheived on Earth.

One of the two major recent equation-of-state and opacity efforts is the international Opacity Project (OP; see the books by Seaton 1995, Berrington 1997). A necessary part of the OP is the so-called Mihalas–Hummer–Däppen (MHD) equation of state (see below). It deals with *heuristic* concepts about the modification of atoms and ions in a plasma. The principal effort of OP has been to construct a tremendous atomic database, comprising state-of-the-art theoretical and experimental data. The other major recent equation-of-state and opacity effort is being pursued by the OPAL group at Lawrence Livermore National Laboratory (Iglesias and Rogers 1996; Rogers *et al* 1996). Its equation of state is based on a sophisticated *systematic* method for including density effects in a plasma. A further effort is ongoing at Los Alamos National Laboratory, which has been an important opacity provider since the 1960s.

The equation of state

Chemical and physical picture

Most realistic equations of state that have appeared in the last 30 yr belong to the *chemical picture* and are based on the free-energy-minimization method. This method uses approximate statistical mechanical models (for example the nonrelativistic electron gas). From these models a macroscopic free energy is constructed as a function of temperature T, volume V and the particle numbers $N_1, \ldots, N_m$ of the m components of the plasma. This free energy is minimized subject to the stoichiometric constraint. The solution of this minimum problem then gives both the equilibrium concentrations and, if inserted in the free energy and its derivatives, the equation of state and the thermodynamic quantities.

In the absence of interaction terms, the free-energy minimization method is equivalent to the familiar Saha equation. For more realistic systems, interactions can be modeled in many ways. The MHD equation of state, *e.g.*, is based on an occupation probability formalism. Specifically, the internal partition functions $Z_s^{\rm int}$ of species s adopted by MHD are sums that are made convergent by weights

$$Z_s^{\rm int} = \sum_i w_{is} g_{is} \exp\left(-\frac{E_{is}}{k_B T}\right). \qquad (1)$$

Here, the index is labels the state i of species s. E_{is} and g_{is} are their energies (in MHD assumed to be unshifted) and statistical weights, respectively, k_B is the Boltzmann constant, T temperature and the weights w_{is} are the occupation probabilities that take into account charged and neutral surrounding particles. In physical terms, w_{is} gives the fraction of all particles of species s that can exist in state i with an electron bound to the atom or ion, and $1-w_{is}$ gives the fraction of those that are so heavily perturbed by nearby neighbors that their states are effectively destroyed.

In the alternative *physical picture*, on which the OPAL equation of state is based, no assumptions about

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

energy-level shifts or the convergence of internal partition functions have to be made. On the contrary, properties of energy levels and the partition functions come out from the formalism. The physical-picture approach is illustrated by the activity expansion for gaseous hydrogen. Pressure is determined from a self-consistent solution of the equations

$$\frac{p}{k_B T} = z + z^2 b_2 + z^3 b_3 + \cdots \tag{2}$$

$$\rho = \frac{z}{k_B T}\left(\frac{\partial p}{\partial z}\right). \tag{3}$$

In these equations, $z = \lambda^{-3}\exp(\mu/k_B T)$ is the activity, $\lambda \equiv h/(2\pi m_e k_B T)^{1/2}$ is the thermal (de Broglie) wavelength of electrons and μ is the chemical potential. The b_n are cluster coefficients such that b_2 includes all two particle states, b_3 includes all three particle-states, etc. Although the bound-state part of b_2 is divergent, the scattering-state part, which is omitted in the chemical-picture approach, has a compensating divergence. Consequently the total b_2 does not contain a divergence of this type.

Solar applications

In the solar convection zone, helioseismology presents an opportunity to isolate the question of the equation of state from opacity and nuclear reaction rates, since the stratification is essentially adiabatic and thus determined by thermodynamics. Accurate analysis of the observations requires use of the full, nonasymptotic behavior of the oscillations (see SOLAR INTERIOR: INVERSIONS OF HELIOSEISMIC DATA). Figure 1 shows a typical result of a numerical inversion of solar oscillation data. It gives, as a function of distance r from the center, the relative difference (in the sense Sun − model) between the squared sound speed obtained from inversion of oscillation data and that of two standard models. If a model were perfect, it would lead to a zero line. The two models used for figure 1 are in all respects identical except that they use a different equation of state (MHD and OPAL).

In a simplifying spirit we can look at results such as shown in figure 1 as the *data* of helioseismology, disregarding how they were obtained from solar oscillation frequencies. The most important result of the helioseismic equation-of-state analyses was that it is essential to include the leading Coulomb correction to ideal-gas thermodynamics. Under solar conditions, the size of the relative Coulomb pressure correction is largest in the outer part of the convection zone (about 8%) and it has another local maximum in the core (about −1%). Figure 1, however, hides the fact that the Coulomb correction is the most important one, since both MHD and OPAL already contain it. Note that in figure 1 the most significant information about the equation of state regards the convection zone ($r > 0.71 R_\odot$), since, beneath it, one cannot disentangle the influence from the equation of state from other effects. Figure 1 contains the evidence that in the region $0.90 R_\odot < r < 0.97 R_\odot$ the OPAL equation of state is a better fit to reality than MHD.

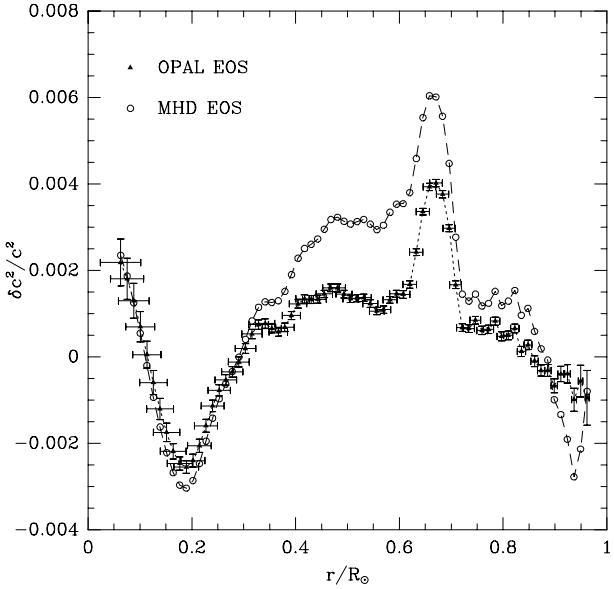


Figure 1. Difference between squared sound speed from inversion of oscillation data and that of a standard model based on the MHD (circles) and OPAL (triangles) equation of state. Figure provided by S Basu (Princeton).

Two very recent inversions have had further implications for the equation of state. First, the strong constraints from helioseismology now force us to include *relativistic effects* of electrons. Neither MHD nor OPAL has so far included relativistic effects, unlike the earlier and simpler EFF and its more recent sibling SIREFF (see below). Second, there are indications that, for $r > 0.97 R_\odot$, MHD is favored over OPAL. If this result is confirmed, it would mean that the treatment of excited states in equation (1) is validated. Such a result might appear somewhat surprising, because in the upper part of the Sun, with density decreasing, OPAL by its nature of a systematic expansion is becoming inevitably more accurate; however, by its heuristic approach MHD would have incorporated even finer, higher-order effects. Since the excited states are effective in the hydrogen and helium ionization zones, this result would have implications for the solar helium abundance determination.

Opacity

History

Opacity has long been an issue in understanding stars. As long ago as 1926, Eddington identified opacity as one of two clouds obscuring stellar model calculations (the other being the source of stellar energy). At that time it was thought that bound–bound absorption was not a significant source of opacity. It was another 40 yr before a group at Los Alamos included bound–bound transitions and obtained increases in the Rosseland mean opacity[1]

[1] The Rosseland mean is a weighted frequency average over the inverse opacity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

exceeding a factor of 3 in some cases. The Cox–Stewart opacities greatly improved the quality of stellar models and remained the standard for more than a quarter century.

Even though the Los Alamos opacities helped elucidate many features of stars, a number of observations continued to resist explanation. For example, period ratios in classical Cepheid models were too low, the mechanism for pulsation in β-Cephei stars could not be identified, the calculated Li abundance in dwarf stars of the Hyades cluster was much less than observed and simulations underestimated wind-driven mass loss in classical novae. A number of studies found that these problems are sensitive to changes in opacity. However, the opacity increases needed seemed unrealistically large, as much as 300% in the case of the classical Cepheids and the β-Cephei stars. This situation has spawned the major opacity recomputation efforts by OP and OPAL, which have now been ongoing for 15 yr. The new calculations of both groups have resulted in the expected large opacity increases, and they have helped resolve a number of long-standing puzzles.

Solar applications

The opacity affects the thermal structure of a star and therefore the sound speed and density stratification. Thus observations of solar and stellar pulsations have the potential to constrain and test opacity calculations. It is important to note, however, that helioseismology alone cannot determine the thermal stratification and hence the opacity, without additional assumptions. The oscillation frequencies are sensitive only to the mechanical structure of the star, particularly the sound speed c. The helioseismic inferences can only provide a combination of the opacity and the composition profile, but not each of these properties separately. Only if the composition profile is otherwise determined, e.g. from a stellar evolution calculation, can the opacity be properly constrained from the helioseismic data. Such successes were indeed possible; already with early inversions for the solar sound speed done in the mid 1980s, it resulted that a modest increase in opacity, around 20% at temperatures between about 2×10^6 K and 5×10^6 K, was required relative to the Los Alamos opacity tables used then. Such an increase has indeed resulted from the more recent (OPAL) opacity calculations. Interestingly, this trend now appears to be reversing: the latest updates to the OPAL opacities and the solar initial composition have tended to increase the discrepancies between the Sun and the model.

Practical considerations

Opacity and equation of state data can be in the form of either analytical or in-line routines that calculate the desired opacities or thermodynamic quantities, given as input the temperature, density and chemical composition. Alternatively, pre-calculated opacity and equation of state tables can be provided, with appropriate look-up and interpolation routines. The hydrostatic structure calculation requires fewer thermodynamic quantities or derivatives than a pulsation calculation, for which smooth derivatives of the various quantities are required.

There are advantages and disadvantages of using tables versus in-line analytical formulations. Tables are often limited in coverage of T, ρ and composition space, and so may not be useful for stellar models over a wide range of masses, compositions and evolutionary states. Interpolation within and between tables or extrapolation off the edges of tables can be difficult and risky, and introduces significant uncertainty into the calculations. On the other hand, tables often include superior physics compared with simple analytical schemes and can be less computationally expensive than a sophisticated analytical routine.

Analytical procedures can be more flexible, for example in taking into account variations in element mixtures and composition or in providing smooth derivatives. They also may be modified easily to generate additional thermodynamic quantities. Analytical procedures can also be useful to explore the importance of various physical processes, since these processes can easily be turned on and off, exaggerated with multipliers, etc.

Selected equations of state

Table 1 lists the specifications of four of the more popular equations of state. For more information, see the article by Däppen and Guzik (1999), which contains a more detailed description of these, and other equations of state. The equations of state of table 1 are the Eggleton, Faulkner and Flannery (EFF) formalism, the EFF extension by Swenson, Irwin and Rogers (SIREFF), MHD and OPAL. Documentation of the MHD equation of state is found in the book by Seaton (1995). The newest OPAL equation of state tables are described in the article by Rogers *et al* (1996).

Mid- to high-temperature opacities

The three currently most popular opacity sources are listed and the reader is again referred to the article by Däppen and Guzik (1999) for further details and other opacities.

- The updated OPAL tables (Iglesias and Rogers 1996) are based on a 21-element solar mixture. The tables span the range $3.75 \leq \log T \leq 8.7$, and $-8 \leq \log R \leq +1$, where $R = T_6^3/\rho$ and $T_6 = 10^{-6}T$. These opacities and interpolation routines are available on the web site `http://www-phys.llnl.gov/V_Div/OPAL/`.
- The OP provides monochromatic opacities for 17 astrophysically relevant elements up through Ni, on grids of T and N_e (electron density). Tables of Rosseland and Planck[2] mean opacities for 213 compositions are available. Interpolation codes and tables are available on the web site of the OP `http://visier.u-strasbg.fr/OP.html`.
- Los Alamos has recently generated new opacities, using the so-called LEDCOP (Los Alamos Light Element

[2] The Planck mean is a weighted frequency average over the opacity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Fine effects included in the EFF, MHD, OPAL and SIREFF equations of state.

Effect	EFF	MHD	OPAL	SIREFF
Coulomb correction	No	Yes[a]	Yes	Yes[a]
Pressure ionization	Yes[b]	Yes	Yes	Yes[c]
Partial ionization	Yes	Yes	Yes	Yes
Molecules	No	H_2, H_2^+, H^-	No[d]	H_2
Classical ions	Yes	Yes	Yes	Yes
Electron degeneracy	Relativistic	Nonrelativistic	Nonrelativistic	Relativistic
Excited states	No	Yes	Yes	No
Electron exchange	No	No	Yes	Yes
Quantum diffraction	No	No	Yes	No

[a] Debye-Hückel approximation.
[b] *Ad hoc*.
[c] *Ad hoc*, but more general than in EFF.
[d] Molecules will be included in future OPAL tables.

Detailed Configuration Opacity) code. The new Los Alamos Astrophysical Opacity database contains elements from hydrogen to zinc (atomic number $Z = 1$–30). The individual element data can be manipulated to generate Rosseland or Planck mean opacities for any desired mixture. The ranges of validity in T and ρ are 0.5 eV $< T <$ 100 keV, and 10^{-10} g cm^{-3} $\leq \rho \leq 10^9$ g cm^{-3}. These codes and tables are available on the web site `http://t4.lanl.gov`. These new Los Alamos opacities have not yet been tested extensively for astrophysical applications, but a direct comparison with OPAL opacities reveals that they are much closer to OPAL than to the older Los Alamos opacities. Figure 2 shows a selected comparison of OPAL, LEDCOP and OP opacities for stellar interior conditions. It emerges that the results of these three modern efforts agree quite well with each other, but small differences remain.

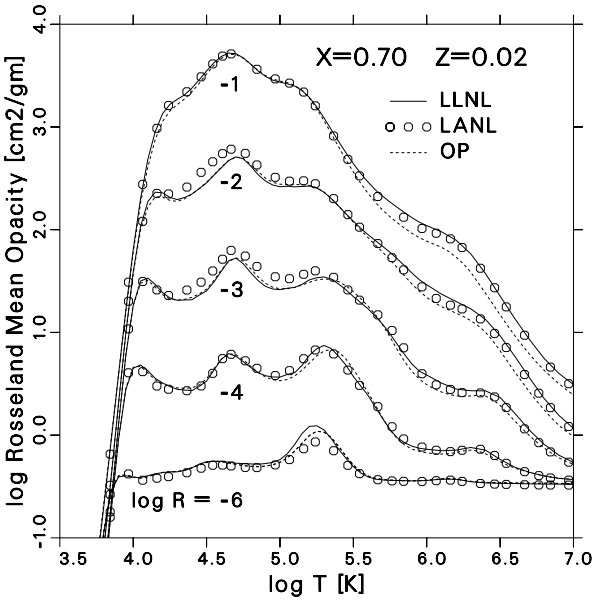


Figure 2. OPAL (Livermore), LEDCOP (latest Los Alamos) and OP Rosseland mean opacities versus temperature for $X = 0.7$ and $Z = 0.02$, for several log R values, where $R = \rho/T_6^3$. Figure provided by J J Keady (Los Alamos).

Low-temperature opacities

Note that neither the OPAL nor OP opacities extend to low temperatures or include molecules. Also, it may be more correct to use Planck mean opacities for optically thin outer layers. For this reason, modelers of the Sun and low-mass stars must supplement the OPAL or OP tables with low-temperature opacities. The article by Däppen and Guzik (1999) lists the current options. The two most important are as follows.

- Kurucz provides Rosseland mean opacities including 58 million lines and diatomic molecules for $1 > \log(Z) > -3$ relative to solar, based on the Grevesse solar mixture (data available at the web site `http://cfaku5.harvard.edu/OPACITIES`).
- Alexander and Ferguson provide Rosseland and Planck mean opacity tables valid for $T = 700$–12 500 K including atomic and molecular line absorption, grain absorption and scattering by silicates, iron, carbon and SiC (contact Alexander, e-mail: `dra@twsuvm.uc.twsu.edu`).

Bibliography

Berrington K A 1997 *The Opacity Project* vol II (Bristol: Institute of Physics Publishing)
Däppen W and Guzik J A 2000 Astrophysical equation of state and opacity *Variable Stars as Essential Astrophysical Tools (NATO-ASI, Çeşme, 1998)* ed C Ibanoglu (Dordrecht: Kluwer) pp 177–212
Iglesias C A and Rogers F J 1996 Updated OPAL opacities *Astrophys. J.* **464** 943–53
Rogers F J, Swenson F J and Iglesias C A 1996 OPAL equation-of-state tables for astrophysical applications *Astrophys. J.* **456** 902–8
Seaton M J 1995 *The Opacity Project* vol I (Bristol: Institute of Physics Publishing)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Werner Däppen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Influence of Magnetic Fields

The p-modes that propagate in the solar interior are sound waves in a gravitationally stratified medium (see SOLAR INTERIOR; HELIOSEISMOLOGY: THEORY). Sound waves involve compressions and rarefactions of the medium and as such are influenced by other effects that cause pressure variations. Magnetic fields impart an extra degree of elasticity to a medium. In the plasma that models the Sun and its atmosphere, magnetic elasticity adds to the compressibility of the medium. These general considerations lead us to expect a coupling between the Sun's global oscillations (p-modes) and magnetism. What is the nature of that coupling and how may it be used to further our understanding of both oscillations and solar magnetism? This is the topic of magnetic helioseismology; still at an early stage of observational and theoretical development, the subject promises to provide insights not only into the physics of the Sun's global oscillations but also into the nature of some of its magnetic structures.

In what ways, then, may magnetism influence p-mode oscillations? An equilibrium magnetic field carries with it a magnetic pressure and a magnetic tension (see MAGNETOHYDRODYNAMICS); when perturbed, the field generates restoring forces through the bending or compressing of its field lines, which lead to magnetohydrodynamic waves (see MAGNETOHYDRODYNAMIC WAVES). Magnetoacoustic waves are produced whenever perturbations compress or expand the medium. Thus p-modes, being compressible, are quite naturally influenced by magnetism.

Structured magnetism

Now the Sun's magnetic field is not distributed uniformly through its interior, surface and atmosphere; rather, the field tends to be accumulated in certain locations favored by stability considerations or the influence of convective flows. In the surface layers of the Sun, magnetism is most apparent in sunspots where the field conspires to produce a 20% or 30% local cooling of the plasma when compared with surroundings (see SUNSPOTS). Aside from sunspots, magnetism is also present in small magnetic flux tubes which border the network between supergranules, preferring the downdraught regions (see SOLAR PHOTOSPHERE). Thus, the field of the photospheric layers of the Sun is present in concentrated forms, rather than spread uniformly throughout the medium: flux tubes or ensembles of flux tubes are the preferred forms.

Above the photosphere, the fall off in plasma pressure brought about by gravitational stratification allows the confined fields of flux tubes to spread out and occupy the whole of the atmosphere. The expansion of the field is accomplished by some 1000–1500 km above the solar surface. The result is a magnetic canopy. The canopy overlies the photosphere and separates magnetized plasma (in the overlying chromospheric and coronal atmosphere) from the convection zone and below (see CHROMOSPHERE: MAGNETIC CANOPY).

Below the photosphere, the roots of the magnetic field lie within the solar interior where various effects bring about the manipulation and transport of magnetic flux through the convection zone and into the visible layers (see SOLAR INTERIOR: CONVECTION ZONE FLUX TUBES). The base of the convection zone, some 200 000 km below the solar surface, is a favored location for the storage and manipulation of magnetic field.

Thus, it is convenient to view the magnetic field of the Sun in three separate locations: the solar interior, the vertical tubes of the photosphere, and the expanded field of the solar atmosphere.

These general considerations suggest that p-modes are influenced by magnetism, and that influence depends upon the local magnetism of the Sun. Thus there arises the possibility of a local seismology (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY) in which p-modes interact with the magnetic fields of a particular region, producing a distinctive signature that may be analysed for information about the magnetic structure in question.

The influence of a magnetic field on p-modes may be thought of in several ways. First, the field modifies the equilibrium structure of a plasma, simply because its magnetic pressure has to be balanced along with the fluid or plasma pressure. Gradients in the total pressure (plasma plus magnetic) are balanced by gravitational effects: the medium is stratified according to magnetohydrostatic balance. Secondly, the presence of a magnetic field within a medium (plasma) means that additional forces act on the system when it is perturbed, and these forces depend upon the strength of the magnetic field. The effect is most simply characterized by the Alfvén speed, the basic speed with which a magnetic disturbance propagates, to be compared with the sound speed (the basic speed with which sound propagates in a medium). Together, these two speeds may be combined to describe a fast magnetoacoustic speed which is both supersonic and super-Alfvénic. Finally, there is an indirect effect possible because of magnetism: the temperature of the medium may be changed because of the presence of a magnetic field, thus modifying the sound speed. This effect is most clearly apparent in sunspots with their reduction in temperature (and thus sound speed), but may also occur elsewhere in more subtle ways.

Thus, the influence of magnetism on p-modes is different in different regions of the Sun. Sunspots offer locations where magnetism is especially significant; a 3000 G field in the visible layers of a sunspot produces a magnetic pressure that is broadly comparable with the atmospheric pressure of the plasma within the spot. Consequently, we can expect p-modes to be appreciably influenced by sunspots. Indeed, observations show that p-modes are strongly absorbed (up to 50%) in the vicinity of a spot (see SUNSPOT OSCILLATIONS AND SEISMOLOGY). Also, p-modes are influenced by the active regions that overlie sunspots; the amplitudes of oscillations are found to be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

suppressed in the vicinity of active regions. Several theoretical interpretations have been advanced as possible explanations of the effect.

Elsewhere in the Sun, away from sunspots, the effects of magnetic fields on p-modes are likely to be less pronounced, either because the magnetic pressure in the field is considerably lower than the plasma pressure—as within the solar interior—or because the modes themselves do not reach that particular magnetic region (as with the solar corona). Generally, p-modes occupy the solar interior to different depths, depending upon the degree (a measure of horizontal wavelength) of the mode. Low-degree modes penetrate deeply below the solar surface, intermediate-degree modes are largely confined to within the convection zone, and high-degree modes occupy the upper reaches of the convection zone.

Now the solar interior is the store-house of all the magnetic field that is ultimately seen in the Sun's atmosphere. But, despite the obvious importance of the interior in the process of generating and manipulating magnetic field (see DYNAMOS: SOLAR AND STELLAR), the effects of subsurface magnetic fields on p-modes are not likely to be large. This is simply because stratification leads to a high plasma pressure in the interior, dominating any magnetic pressure that may be present. In fact, the field strength is uncertain in these layers. For example, estimates of the field at the base of the convection zone range anywhere from 10^3 G to 10^6 G; only field strengths towards the upper end of this range are likely to produce detectable effects on the oscillations. If fields of this order are present—and magnetic helioseismology can be expected to resolve this issue—then slight but significant changes in p-modes are to be expected as a consequence of such magnetism. Oscillations of low degree are perhaps most likely to be influenced by internal magnetic fields, since these modes penetrate deep within the Sun; modes of intermediate or high degree are confined mainly to the upper layers of the convection zone, and so are not directly subject to the fields stored deeper within the Sun. Instead, high-degree modes are likely to be influenced more by the fields that fill the chromosphere or are in photospheric flux tubes and sunspots.

Oscillations have also been used to probe active regions in the process of emerging through the convection zone. This is the developing subject of solar tomography. By looking at travel times of modes that occupy the upper 30% of the convection zone, it has proved possible to infer the existence of 'acoustic perturbations' at a depth of 10–20 Mm below the solar surface. The perturbations are in the mode propagation speed and amount to about ± 1.5 km s^{-1} in the magnetoacoustic speed, and are interpreted as corresponding either to local temperature changes or magnetism. At a depth of 10 Mm, the perturbation in propagation speed would correspond to a temperature variation of the order of 2000 K or a magnetic field of strength 15 000 G (or some combination of the two effects).

Solar cycle variations

Of particular interest in magnetic helioseismology is the observed variation of p-mode frequencies with the Sun's activity cycle (see SOLAR ACTIVITY). The effect was first suggested by data for low-degree modes (degrees 0–4) but the smallness of the effect (a frequency shift of about 1 part in 10^4) meant that it was not at first universally accepted. Later, detailed studies of both low- and intermediate-degree (5–140) modes showed unequivocally that oscillation frequencies vary with the solar cycle. Accurate ground-based measurements of low-degree modes are now available over an almost complete solar cycle of 22 years, and the period 1990 to 1996 (during which solar activity was falling) has been recently studied in detail. Also, observations from space, using instruments on the Solar and Heliospheric Observatory (SOHO) satellite, have monitored p-modes since 1996 (when the Sun was at a minimum in activity). Considerable interest attaches to expected developments in these programs as the Sun moves through solar maximum.

The analysis of frequency changes in time is done by comparing mode frequencies for one year with those in another year; systematic shifts are found. For example, frequencies measured when the Sun was at or close to solar minimum, such as in the year 1986, may be compared with frequencies measured in later years when the Sun was more active (e.g. 1988 and 1989). The shift in frequency may then be displayed as a function of frequency (taking, for example, the frequency at solar minimum as a base). The results show (see figure 1) that frequency shifts are very small for oscillations of frequency below about 2 mHz, but rise rapidly thereafter and reach a peak shift of about 0.8 μHz (for intermediate-degree modes) at a frequency just below 4 mHz. Thereafter, at higher frequencies the shift falls rapidly and may switch in sign. A similar trend is seen also in low-degree modes.

Theoretical explanations of the observed frequency shifts with the solar cycle have been based primarily upon a combination of temporal changes in chromospheric magnetism coupled with simultaneous changes in chromospheric temperature (i.e. changes in the magnetic canopy) or in changes in the magnetism and thermal structure of photospheric magnetic flux tubes. Such models give frequency shifts similar to those observed, but only at the expense of invoking thermal changes that are larger than expected. It may be that other effects, masquerading as temperature changes, are also involved.

Coronal seismology

Since p-modes are mainly confined to the solar interior they are not expected to be greatly influenced by the solar corona. Consequently, if we are to study the corona through seismology, it is unlikely that p-modes will be involved. Instead, a coronal seismology must be developed using waves other than p-modes as diagnostic probes. A number of such attempts are currently under way, using magnetohydrodynamic waves to probe the magnetism of the corona. Both propagating waves and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

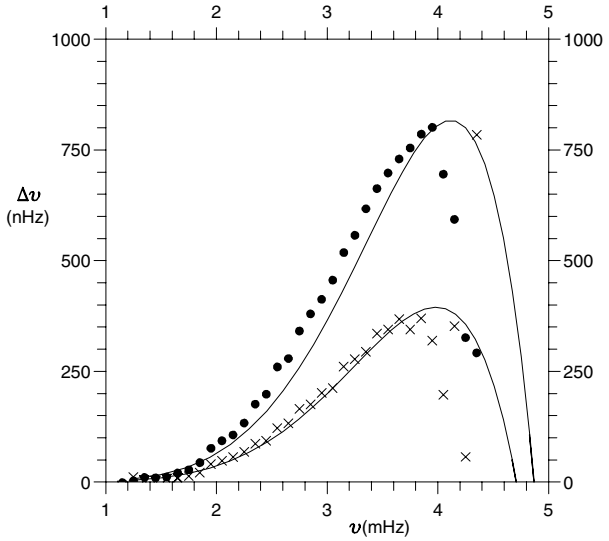


Figure 1. The observationally determined frequency shifts $\Delta\nu$ (in nHz) of intermediate p-modes as a function of frequency ν (in mHz) for the years 1988 (crosses) and 1989 (full circles), compared with 1986. The observations were made by K G Libbrecht and M F Woodard. Also shown are the results from model calculations by R Jain and B Roberts, for oscillations of degree 75. The theoretically determined frequency shifts (shown as full curves) arise from an increase in chromospheric magnetic field strength (from 30 G to 50 G (lower curve) or 61 G (upper curve)) combined with a simultaneous rise in chromospheric temperature (from 4170 K to 6370 K (lower curve) or 7770 K (upper curve)).

standing oscillations have been used in a variety of magnetic objects. Magnetic loops (see CORONAL LOOPS), plumes (see POLAR PLUMES), and prominences (see SOLAR PROMINENCE OSCILLATIONS) are all being investigated in an attempt to discover their properties through the use of a local seismology.

Bibliography

Kosovichev A G 1999 Oscillations in active regions—diagnostics and seismology *Magnetic Fields and Oscillations (Astron. Soc. Pacific Conf. Ser. 184)* ed B Schmeider, A Hofmann and J Staude (San Francisco: Astronomical Society of the Pacific) p 151

Pallé P L 1995 Solar cycle frequency shifts at low *l GONG '94: Helio- and Astero-Seismology from the Earth and Space (Astron. Soc. Pacific Conf. Ser. 76)* ed R K Ulrich, E J Rhodes and W Däppen (San Francisco: Astron Soc Pacific) p 239

Rhodes E J *et al* 1995 Solar cycle dependence of p-mode frequencies at intermediate and high degrees *GONG '94: Helio- and Astero-Seismology from the Earth and Space (Astron. Soc. Pacific Conf. Ser. 76)* ed R K Ulrich, E J Rhodes and W Däppen (San Francisco: Astron Soc Pacific) p 227

Roberts B 1996 The influence of chromospheric magnetism on oscillation frequencies *Bull. Astron. Soc. India* **24** 199

Thompson M J 1994 Recent results from helioseismology *Advances in Solar Physics* ed G Belvedere, M Rodono and G M Simnett (Berlin: Springer) p 19

B Roberts

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Inversions of Helioseismic Data

Once the eigenfrequencies of the solar oscillations are measured (see HELIOSEISMIC OBSERVATIONS), they are analysed further to investigate the solar interior. For example, we can use these frequencies to reveal how the adiabatic soundspeed varies with depth, or how the interior of the Sun rotates differentially, both in depth and in latitude. This procedure is called *inversion*, because the procedure of calculating eigenfrequencies for a given solar structure, or a given rotation rate, is conventionally considered 'forward' (see HELIOSEISMOLOGY: THEORY). Inversions play the central role in our investigation of the solar interior, and indeed most of what we know about the internal structure of the Sun and its rotation (see SOLAR INTERIOR, SOLAR INTERIOR: ROTATION) have been obtained through inversions.

General properties

Generally, when there are many measurements (such as eigenfrequencies) which reflect a certain property (such as soundspeed distribution) of the system observed (such as the Sun), we can hope to infer this property in detail. How well we can do this depends on the actual information content of the entire dataset.

First of all, the information content of each measurement depends on how well the particular property is reflected in this measurement. The m-averaged eigenfrequencies ω_{nl} (n and l are the radial order and the degree of the mode) are determined by the (spherically averaged) distribution of soundspeed and density, but largely by the former in the case of p-mode oscillations. As a result, there is more information about the soundspeed distribution in the p-mode frequencies than about the density distribution. But a mode frequency can reflect only soundspeed distribution within the propagation zone of the mode; the frequency carries essentially no information about the soundspeed below the lower turning point. The statistical error in the measurement is also important; as the measurement error decreases the actual information content increases.

Another very important issue is how the information content of each measurement adds up; if the data are highly redundant, even if there are a large number of them, they do not add much independent information to each other. Consequently, although the effective measurement error will be lowered, the actual information content may not be rich enough for us to obtain much detail about the property we are investigating.

In case of inversions of solar oscillation frequencies, a certain degree of non-redundancy is guaranteed by the variation of the penetration depths from mode to mode, which ensures that the different modes sample different regions of the Sun, and therefore carry independent information.

Method of linear inversions

Helioseismic inversions are often linearized. In a typical linear inverse problem in one dimension ($a \le x \le b$), we seek to estimate a function $q(x)$ from a set of integral constraints

$$d_i = \int_a^b K_i(x) q(x)\,\mathrm{d}x + e_i \quad (i = 1, \ldots, M) \qquad (1)$$

where the measurement d_i, associated with error e_i, is linearly related to $q(x)$ with a kernel (or sensitivity function) $K_i(x)$, and M is the number of such measurements.

In fact, it is not possible to determine the function $q(x)$ uniquely if M is finite, even when the data are completely free from measurement error. This is because for any finite M, one can find an infinite number of functions that are orthogonal to $K_1(x) \ldots K_M(x)$, which, when added to $q(x)$, have no effect on the measurement. The dataset has no information about the component of $q(x)$ that is a linear combination of these functions, so this component has to remain undetermined. Fortunately, in helioseismic inversions these unknown components are generally highly oscillatory functions in space, so our lack of knowledge about their magnitude is unlikely to affect our global view of the Sun.

There are two important classes of inversion methods: optimally localized averaging (OLA) and regularized least-squares fitting (RLSF). Both are linear methods comprising only linear processes, and therefore produce an estimate of q at $x = x_0$, $\hat{q}(x_0)$, that is a linear combination of the data:

$$\begin{aligned}\hat{q}(x_0) &= \sum_i c_i(x_0) d_i \\ &= \int_a^b D(x; x_0) q(x)\,\mathrm{d}x + \sum_i c_i(x_0) e_i\end{aligned}$$

where $c_i(x_0)$ is inversion coefficient and

$$D(x; x_0) \equiv \sum_i c_i(x_0) K_i(x)$$

is the averaging kernel, the width of which gives the resolution of the estimate; if $D(x; x_0)$ is sharply peaked at $x = x_0$, then $\hat{q}(x_0)$ will be close to $q(x_0)$ but for the error term. The standard deviation of the estimate $\hat{q}(x_0)$ is given by

$$\delta\hat{q}(x_0) = \left(\sum_i c_i(x_0) c_j(x_0) E_{ij} \right)^{1/2}$$

where E_{ij} is the covariance of e_i and e_j. It would be desired that the resolution be as high as possible and that the error be as low as possible, but in fact there is a trade-off between the two; a higher resolution is generally associated with greater amplitudes of the coefficients $c_i(x_0)$, and hence with greater error (figure 1).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

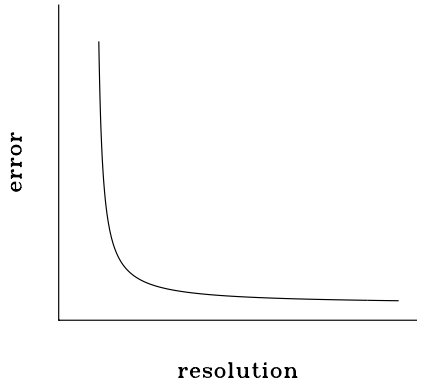


Figure 1. A schematic figure of a trade-off curve.

Optimally localized averaging (OLA)
The optimally localized averaging technique directly seek to balance resolution and error by minimizing, for example,

$$S_{\rm OLA} = \int_a^b D(x;x_0)^2 (x-x_0)^2 \,{\rm d}x + \alpha[\delta\hat{q}(x_0)]^2$$

for each target position x_0, subject to the unimodular condition

$$\int_a^b D(x;x_0)\,{\rm d}x = 1$$

by varying $c_1(x_0)\ldots c_M(x_0)$. Significant amplitudes in $D(x;x_0)$ far from $x = x_0$ are penalized by the integral, a measure of spread of the averaging kernel, whereas a large error is penalized by the second term. The trade-off parameter α determines the compromise between the two. As the data error increases, a higher value of α will have to be chosen and as a result, $D(x;x_0)$ will be more poorly localized. When the localization is poor, often the center x_1 of $D(x;x_0)$ is calculated under a proper definition (e.g. the center of gravity), and $\hat{q}(x_0)$ is re-interpreted as $\hat{q}(x_1)$.

Using a different measure of spread will lead to a different formalism. An essential and important alternative, however, is so-called subtractive optimally localized averaging (SOLA), in which the averaging kernel is fit to a target function (e.g. Gaussian) centered at the target position.

Regularized least-squares fitting (RLSF)
To many people, probably the most natural way to solve the set of equation (1) is to use least-squares fitting; to find an estimate $\hat{q}(x)$ one minimizes

$$S_{\rm LSF} = \sum_i \frac{1}{e_i^2}\left(d_i - \int_a^b K_i(x)\hat{q}(x)\,{\rm d}x\right)^2.$$

In practice, $\hat{q}(x)$ is expanded in a series of functions and then the minimization problem is reduced to a matrix equation. It is often found, however, that the resulting matrix is nearly singular, producing an 'ill-posed' problem. This is normally caused by the redundancy of the data. The most common way to avoid this is to introduce a 'regularization' term that imposes some kind of smoothness to the estimates $\hat{q}(x)$. For example, minimization of

$$S_{\rm RLSF} = S_{\rm LSF} + \alpha \int_a^b [\mathcal{L}q(x)]^2\,{\rm d}x,$$

where α is a regularization parameter and $\mathcal{L}$ is a differential operator, leads to a linear system of equations that is not singular, provided α is large enough. Unlike OLA, it is not necessary to obtain inversion coefficients explicitly, but they are often calculated for diagnostic purposes.

Structure inversion

Since the spherically averaged structure of the Sun determines the m-averaged eigenfrequencies ω_{nl}, in *structure inversion* the observed spectrum of the m-averaged frequency is inverted to recover the internal structure of the Sun.

The m-averaged frequencies of a solar model are calculated by solving an eigenvalue problem. Although this is almost routine work today, its complexity, and nonlinear nature in particular, does not permit direct inversions. Therefore we have to introduce some approximation to simplify the forward problem, so that the problem becomes directly invertible.

One strategy is to use asymptotic analysis. From the asymptotic analysis of solar p-mode oscillations, we see that the m-averaged frequency ω is related to $c(r)$, the soundspeed as a function of radial coordinate r, in the following way:

$$\int_{r_t}^R \left(1 - \frac{L^2c^2}{r^2\omega^2}\right)^{1/2} \frac{{\rm d}r}{c} = \frac{\pi(n+\alpha)}{\omega}$$

where r_t is the radius of the lower turning point determined by the relation $c(r_t)/r_t = \omega/L$, R is the solar radius and $L = \sqrt{l(l+1)}$. The phaseshift α is generally a function of frequency, but in the simplest treatment it is regarded as a constant. We note that, for a given soundspeed distribution, the left-hand side is a function of $w \equiv \omega/L$ alone, and denote this function by $F(w)$, which is identified to the observed quantity in the right-hand side. Then the relation

$$\int_{r_t}^R \left(1 - \frac{L^2c^2}{r^2\omega^2}\right)^{1/2} \frac{{\rm d}r}{c} = F(w)$$

is analytically inverted to yield

$$r = R\exp\left[-\frac{2}{\pi}\int_{a(R)}^{a(r)} \left(\frac{1}{w^2} - \frac{1}{a^2}\right)^{-1/2} \frac{{\rm d}F}{{\rm d}w}\,{\rm d}w\right]$$

where $a(r) \equiv c(r)/r$ is a monotonically decreasing function of r in the solar case. This formula gives the depth as a function of a, which can easily be converted to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the soundspeed distribution $c(r)$. In reality, $F(w)$ can be observationally constructed only in a limited range: even for the highest degree modes, w is larger than $a(R)$, the surface value, and therefore $F(w)$ cannot be known down to $a(R)$. A minor correction term is usually introduced to amend this. Likewise, the highest observed $w = \omega/L$ determines the deepest layer of the Sun which is accessible by this method.

Although the accuracy of the asymptotic method is limited by the accuracy of asymptotic analysis itself, the method proved to be very successful. More sophisticated asymptotic analyses, which can also be used for inversion, have been developed.

Another strategy, which is more commonly adopted nowadays, is to abandon the idea of dealing with frequencies themselves. Instead, we consider the difference between solar frequencies $\omega_{\rm sun}$ and the frequencies of our reference model, $\omega_{\rm ref}$ (the suffixes n and l are omitted). If we assume that our reference model is close enough to the Sun, then we can use linear perturbation theory to relate the difference in structures to difference in frequencies:

$$\frac{\delta\omega}{\omega} = \int_0^R \left(K_{c,\rho}(r)\frac{\delta c}{c}(r) + K_{\rho,c}(r)\frac{\delta\rho}{\rho}(r) \right) \mathrm{d}r + E^{-1}\mathcal{F}_{\rm surf}(\omega)$$

where $\delta\omega = \omega_{\rm sun} - \omega_{\rm ref}$, ω is normally taken to be $\omega_{\rm ref}$, $K_{c,\rho}(r)$ and $K_{\rho,c}(r)$ are the kernels calculated from the reference model (figure 2), δc and $\delta\rho$ denote the excesses in soundspeed c and density ρ in the Sun, respectively, with respect to the reference model. The extra term is present because of our poor knowledge of the surface layers of the Sun; E is the normalized mode inertia and $\mathcal{F}_{\rm surf}$ is the surface term. Choices of quantities other than c and ρ are possible too.

Once the forward problem is linearized in this way, it can be inverted by using either OLA or RLSF. The formulation in the previous section requires some modifications because the number of variables is two in this case, and also because of the presence of the surface term. In the case of RLSF, having two variables requires that the two functions have to be fitted at the same time. In the case of OLA, one of the variables (say, soundspeed excess) is chosen as the target, and to $S_{\rm OLA}$ is added an extra term that penalizes contributions from the other variable.

The surface term has to be evaluated from the comparison between $\omega_{\rm sun}$ and $\omega_{\rm ref}$ and then subtracted from the left-hand side. This can also be done implicitly in the case of OLA, by choosing the coefficients in such a way that the contribution from the surface term vanishes.

Rotation inversion

The internal rotation of the Sun manifests itself in the variation of eigenfrequency over azimuthal order m. In *rotation inversion*, we invert the rotational splitting, the shift of frequency from the m-averaged value $\omega_{nlm} - \omega_{nl}$, for internal differential rotation of the Sun. The rotational splitting is usually calculated by using linear perturbation theory, noting that the Sun's rotation rate is slow compared with its oscillation frequencies. The result is a linear relation between solar internal rotation and rotational splittings:

$$\frac{\omega_{nlm} - \omega_{nl}}{m} = \int_0^R r\,\mathrm{d}r \int_0^\pi \mathrm{d}\theta\, K_{nlm}(r,\theta)\Omega(r,\theta)$$

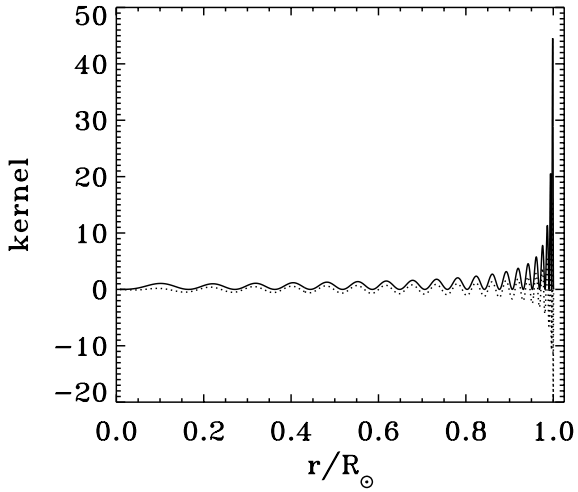


Figure 2. The kernels $K_{c,\rho}(r)$ (full curve) and $K_{\rho,c}(r)$ (dotted curve) calculated from a standard model for $l = 1$, $n = 20$ mode.

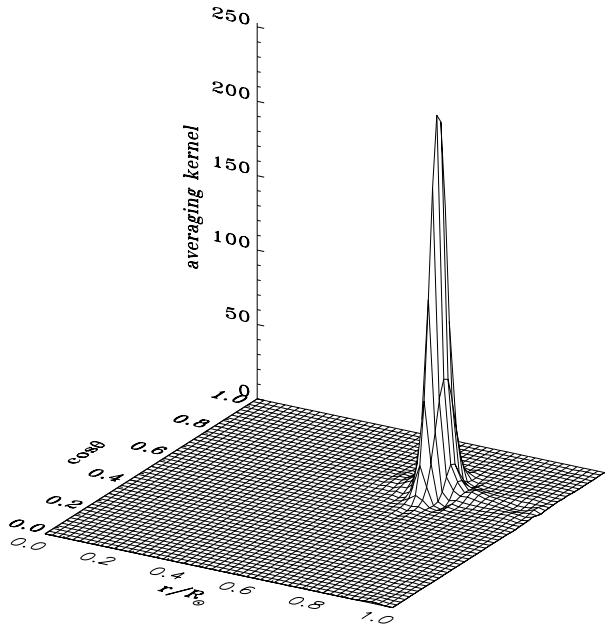


Figure 3. An example of a well localized averaging kernel for two-dimensional rotation inversion. The target position is $r/R_\odot = 0.75$, colatitude $\theta = 45^\circ$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which is numerically invertible. This is similar to the second approach of structure inversion: here we use the linearized difference between eigenfrequencies of a rotating Sun and with frequencies of a non-rotating Sun to infer the difference (namely the rotation). In this case, the linear inversion methods discussed before are more readily applicable except for the difference in dimensions. Figure 3 illustrates an averaging kernel constructed by an OLA method.

Bibliography

Gough D O 1993 Linear adiabatic stellar pulsation *Astrophysical Fluid Dynamics, Les Houches Session XLVII, 1987* ed J-P Zahn and J Zinn-Justin (Amsterdam: Elsevier) pp 399–560

Gough D O and Thompson M J 1991 The inversion problem *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M S Matthews (Tucson, AZ: University of Arizona Press) pp 519–61

Unno W, Osaki Y, Ando H, Saio H and Shibahashi H 1989 *Nonradial Oscillations of Stars* (Tokyo: University of Tokyo Press)

Takahashi Sekii

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Local Helioseismology

This article describes how we use acoustic waves to study the subsurface properties of particular areas of the Sun. Local helioseismology is not as well developed as the more mature area of study, global HELIOSEISMOLOGY. New techniques are still being suggested and developed.This article will discuss some of these techniques and the most solid of the inferences about the SOLAR INTERIOR derived from them.

When one observes a spectral line formed in the SOLAR PHOTOSPHERE, one finds that the wavelength shifts back and forth with a period of about 5 min. This wavelength shift is caused by the DOPPLER EFFECT and tells us there is an approximately vertical mass motion in the photosphere of several hundred m s^{-1}. Theoretically it was calculated some years ago that acoustic waves with 5 min period should not propagate in the sun's photosphere. In fact, by measuring the Doppler shift at different heights in the photosphere, it was found that the waves are not propagating, but that the whole atmosphere is lifting up and down.

From calculations we know that acoustic waves of 5 min period propagate throughout most of the interior of the Sun, but that near the photosphere, where the density is falling off rapidly, the waves reflect. So the interior of the Sun can be considered as a cavity allowing acoustic wave propagation. As part of this reflection process, the atmosphere is raised and lowered with the period of the wave, leading to our observations of Doppler shifts in the photosphere.

By observing different locations in the photosphere, we obtain a picture of apparently independent oscillators, separated by roughly the acoustic wavelength in the photosphere, 2 Mm. The oscillators are not very good ones, with the amplitude rising and falling and the frequency varying.

Below the surface, waves propagate away from a source near the surface. The waves follow ray paths, as seen in figure 1. For a wave propagating into the interior obliquely, the increasing temperature (and hence increasing sound speed) causes the ray paths to bend, or refract. At some point the rays are bent back toward the surface which can be called total internal refraction, similar to the optical phenomenon of total internal reflection. The bottom of the ray path, or turning point, occurs at a characteristic depth. From a straightforward examination of the purely acoustic dispersion relation, we find that the sound speed at the bottom of the ray path is the same as the horizontal phase speed seen at the surface (ω/k_{h}), where ω is the frequency and k_{h} is the horizontal wavenumber. An important point to remember is that, when we observe a wave with phase speed ω/k_{h}, that wave propagates to a depth at which the sound speed is the same as the phase speed.

From a variety of evidence, we believe that the acoustic waves are excited by convective motion just below the photosphere.The lifetime of this convection, visible at the surface and called GRANULATION, is about 15 min. One possible picture of the wave field is a random array of sources sending off wave packets with 5 min period (the turnover time of the convection) and lasting for about 15 min. This random array of sources could very densely cover the solar surface or not—we do not know yet. The basic problem of local helioseismology is how to use these waves as a probe of solar structure.

One way to use these waves is to somehow isolate the waves travelling from one surface location to another. Two methods have been studied that perform this isolation, termed time–distance helioseismology and acoustic imaging.

In time–distance helioseismology, the signals at two surface locations are cross correlated, a mathematical operation consisting of shifting in time the two signals relative to each other and measuring how similar the two signals are at each value of the shift.

If a wave packet was emitted at surface location 1 at time t and it takes an interval T to travel along the subsurface path to location 2, we expect that the signal at location 2 at time $t + T$ would be most similar to the signal at location 1. The time interval T, called the travel time, depends on the distance between the locations, with T increasing with the separation. T therefore depends on the normal increase of sound speed with depth into the Sun. It also depends on temperature perturbations, flows and magnetic fields along the way. If there is a 'hot spot' along the waves' path, the wave will temporarily speed up and the resultant interval T will be reduced. If there is a material flow, say from a convective cell, with the flow directed towards location 2, the wave will be advected by the flow and the resultant interval T will be reduced. Conversely, for waves travelling from location 2 to location 1, the interval T_{21} will be increased. So, by measuring the difference in travel time for the two directions, the direction and magnitude of flows can be isolated.

The second method of isolating the waves is called acoustic imaging. An acoustic signal present at a certain location and a certain time will act like a source. The signal will propagate out from the origin and at a certain time interval later will reach the surface. In principle, all surface points can be used and it is only necessary to know how long it takes to propagate to each surface point. Then, by shifting the signals in time and adding them up over all the surface points, we can reproduce the signal at the origin. What we have described here is the generation of the origin signal from the outgoing waves. However, the process can be reversed and the signal can be reconstructed from the ingoing waves. That is, we look earlier in time at waves that will propagate to the origin point.

The reconstructed signals, ingoing and outgoing, contain information about flows, magnetic fields, temperature perturbations and absorption. To date, information about absorption and phase changes has been derived. By comparing the strength of the ingoing signal, which is unaffected by whatever is at the origin, with the outgoing signal, which will be reduced in amplitude by an absorber

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

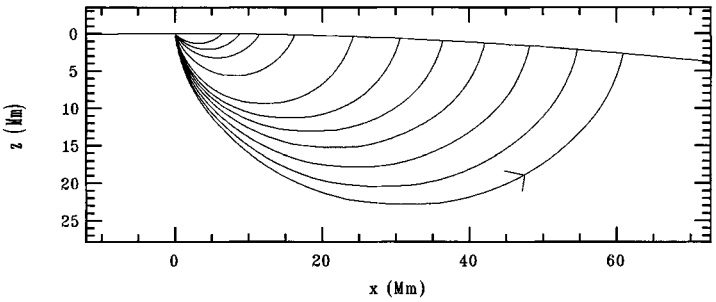


Figure 1. Acoustic ray paths below the solar surface. An acoustic source at the surface would send out energy along these paths that would return to the surface. It takes a longer time for waves to travel the longer paths.

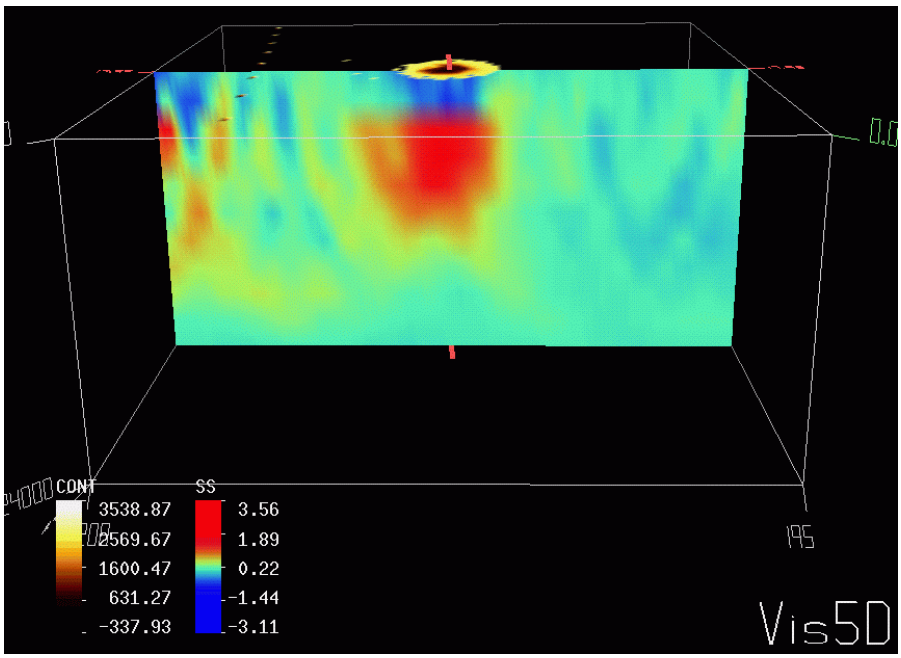


Figure 2. Subsurface wave speed map below a sunspot. One plane from a three-dimensional inversion is shown from the surface to 24 Mm below the surface. The top plane is the intensity of the sunspot at the surface with the normal surface intensity transparent. The wave speed perturbations are shown in the blue–red scale in km s^{-1}. **This figure is reproduced as Color Plate 51.**

near the origin, it has been possible to detect the presence of absorbers, mostly in the neighborhood of SUNSPOTS.

By cross correlating the ingoing signal with the outgoing signal, it is possible to find phase shifts that occur near the origin. These have been used to study the high-wave-speed region below sunspots. The high wave speeds are presumably caused by a combination of higher than normal temperature and magnetic field.

Both time–distance helioseismology and acoustic imaging can benefit from a further step of modelling or inversion. In this step the measurements, which are smoothed averages or integrals of interior solar properties, are used as input to a calculation that attempts to isolate the inhomogeneities at and below the solar surface. This is an important step that can better summarize our knowledge and uncertainties of the interior. To date, more of the modelling effort has centered on the time–distance technique, which has led to some interesting results as seen in the next section (see also SUNSPOT MODELS).

Examples

Sunspots have been observed with telescopes for nearly 400 yr. Our knowledge of their properties below the Sun's surface, or photosphere, is almost completely theoretical. We know that the gas in a sunspot is cooler than the surroundings, leading to the relative darkness and hence the name sunspot. We also know from our surface observations that a strong magnetic field permeates the sunspot (see SUNSPOT MAGNETIC FIELDS). An example of the intensity of a spot is shown in figure 2 in the top plane in which only the region of reduced brightness is shown. The black region, or umbra, is the region of maximum

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

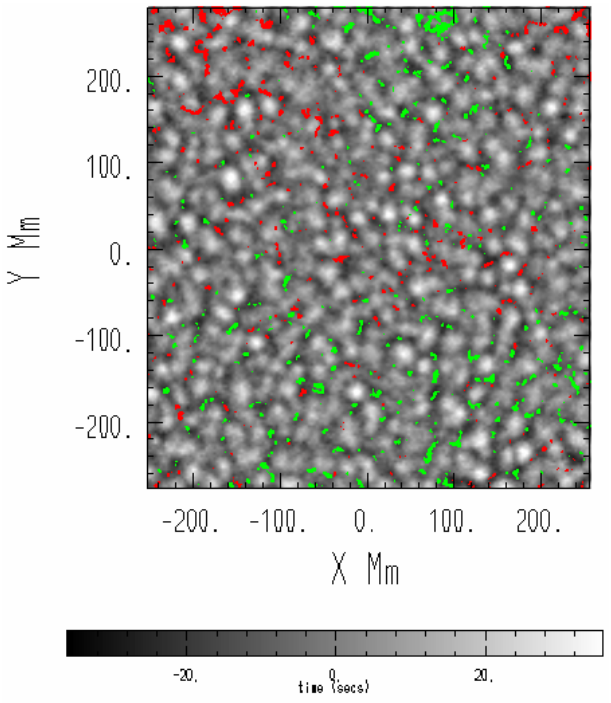


Figure 3. Divergence and magnetic field near the solar surface. The gray scale shows travel time measurements that are proportional to the velocity divergence, with white indicating outflow, black converging flow. The green and red are the two polarities of the magnetic field, which are only shown when the field is larger than 15 G. **This figure is reproduced as Color Plate 52.**

reduction of brightness. The region of intermediate brightness, the penumbra, is shown in red surrounding the umbra.

Below the intensity plane is shown a vertical cut from a three-dimensional wave-speed inversion using time–distance data. The most striking feature is the red region extending about 12 Mm below the sunspot indicating a region of higher than normal wave speed. Although in principle we may be able to separate temperature changes from magnetic fields, for this work it was not possible to distinguish the two. So, at a depth of 10 Mm, we have either a magnetic field of strength 10 kG or a temperature rise of 3% of the temperature.

The Sun's radiant energy, generated in the core by nuclear fusion, is carried outwards in the outer 30% of the radius by convective motions. These motions near the surface have been studied by time–distance helioseismic techniques. In figure 3, we show near-surface measurements of the flow of the supergranules, a cellular flow with cell sizes of 30 Mm. The gray scale in the figure is showing the flow divergence, positive or white for outflow and negative or black for converging flows. This signal was measured using surface-gravity waves concentrated in the 2 Mm immediately below the surface. Superposed on the divergence signal in green and red are areas where the magnetic field strength is larger than 15 G. The two colors are for the two polarities of the field. We see that the magnetic fields are concentrated in the regions of converging flows, as if the flows were pushing them around. The time–distance measurements provide a way to study this flow pattern in detail over most of the visible hemisphere, a definite improvement over other techniques.

Thomas L Duvall Jr

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Neutrinos

Six experiments have observed solar NEUTRINOS, three of which are radiochemical experiments: Homestake (chlorine detector), GALLEX (gallium detector) and SAGE (gallium detectors). These radiochemical detectors register all neutrinos above a fixed threshold energy (0.8 MeV for chlorine, 0.2 MeV for gallium), with no further information about energy. Three other experiments, Kamiokande, SUPER-KAMIOKANDE, and SUDBURY NEUTRINO OBSERVATORY (SNO), measure the energies in real time of electrons produced in neutrino interactions in water. Kamiokande and Super-Kamiokande use neutrino–electron scattering in ordinary (but very pure) water and SNO uses neutrino absorption and neutrino disassociation of deuterium in heavy water. These water detectors are designed to measure electrons, from neutrino interactions, with more than 5 MeV of energy.

I will discuss the results of each of these experiments after I first answer a series of background questions. Why study solar neutrinos? What does the combined standard model (solar plus electroweak) predict for solar neutrinos? Why are the calculations of neutrino fluxes robust? What are the three solar neutrino problems? I will conclude by summarizing what have we learned in the first 35 yr of solar neutrino research and what we hope to learn in the next decade.

Why Study Solar Neutrinos?

Astronomers study solar neutrinos for different reasons than physicists. For astronomers, solar neutrino observations offer an opportunity to test directly the theories of STELLAR EVOLUTION and of nuclear energy generation. With neutrinos one can look into the interior of a main sequence star and observe the nuclear fusion reactions that are ultimately responsible for starlight via the general reaction

$$4\,^1\mathrm{H} \longrightarrow {}^4\mathrm{He} + 2\mathrm{e}^+ + 2\nu_\mathrm{e}. \qquad (1)$$

The optical depth of the Sun for a typical neutrino produced by nuclear fusion is $\sim 10^{-9}$, about 20 orders of magnitude smaller than the optical depth for a typical optical photon. As we shall see, solar neutrino experiments constitute quantitative, well-defined tests of the theory of stellar evolution. Since stellar evolution theory is used widely in interpreting astronomical observations, direct tests of this theory are of importance to astronomers.

Table 1 shows the principal nuclear reactions that accomplish equation (1) via the proton–proton chain of reactions. In what follows, we shall refer often to the reactions listed in this table.

Solar neutrinos are of interest to physicists because they can be used to perform unique particle physics experiments. Many physicists believe that solar neutrino experiments may have already provided evidence for a non-zero neutrino mass and that electron flavor (or the number of electron-type neutrinos) may not be conserved.

Table 1. The principal reactions of the p–p chain.

Reaction number	Reaction	Neutrino energy (MeV)
1	$\mathrm{p} + \mathrm{p} \rightarrow {}^2\mathrm{H} + \mathrm{e}^+ + \nu_\mathrm{e}$	0.0–0.4
2	$\mathrm{p} + \mathrm{e}^- + \mathrm{p} \rightarrow {}^2\mathrm{H} + \nu_\mathrm{e}$	1.4
3	${}^2\mathrm{H} + \mathrm{p} \rightarrow {}^3\mathrm{He} + \gamma$	
4	${}^3\mathrm{He} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + 2\mathrm{p}$	
	or	
5	${}^3\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^7\mathrm{Be} + \gamma$	
	then	
6	$\mathrm{e}^- + {}^7\mathrm{Be} \rightarrow {}^7\mathrm{Li} + \nu_\mathrm{e}$	0.86, 0.38
7	${}^7\mathrm{Li} + \mathrm{p} \rightarrow {}^4\mathrm{He} + {}^4\mathrm{He}$	
	or	
8	${}^7\mathrm{Be} + \mathrm{p} \rightarrow {}^8\mathrm{B} + \gamma$	
9	${}^8\mathrm{B} \rightarrow {}^8\mathrm{Be} + \mathrm{e}^+ + \nu_\mathrm{e}$	0–15

For some of the theoretically most interesting neutrino parameters, solar neutrino experiments are more sensitive tests for neutrino transformations in flight than experiments that can be carried out with laboratory sources. The reasons for this exquisite sensitivity are (1) the great distance between the accelerator (the solar interior) and the detector (on Earth), (2) the relatively low energy (MeV) of solar neutrinos and (3) the enormous path length of matter ($\sim 10^{11}$ g cm^{-2}) that neutrinos must pass through on their way out of the Sun.

One can quantify the sensitivity of solar neutrinos relative to laboratory experiments by considering the proper time that would elapse for a finite-mass neutrino in flight between the point of production and the point of detection. The elapsed proper time is a measure of the opportunity that a neutrino has to transform its state and is proportional to the ratio, R, of path length divided by energy:

$$\text{proper time} \propto R = \frac{\text{path length}}{\text{energy}}. \qquad (2)$$

Future accelerator experiments with multi-GeV neutrinos are expected to reach a sensitivity of order $R = 10^2$ km GeV^{-1}. Reactor experiments are planned that will reach a level of sensitivity of $R = 10^5$ km GeV^{-1} for neutrinos with MeV energies. Solar neutrino experiments, because of the enormous distance between the source (the center of the Sun) and the detector (on Earth) and the relatively low energies (1 MeV to 10 MeV) of solar neutrinos involve much larger values of neutrino proper time,

$$R(\text{solar}) = \frac{10^8}{10^{-3}} \left(\frac{\text{km}}{\text{GeV}}\right) \sim 10^{11} \left(\frac{\text{km}}{\text{GeV}}\right). \qquad (3)$$

Because of the long proper time that is available to a neutrino to transform its state, solar neutrino experiments are sensitive to very small neutrino masses that can cause neutrino oscillations (changes between different neutrino states) in vacuum. Quantitatively, vacuum neutrino oscillations are sensitive to masses as small as

$$m_\nu(\text{solar level of sensitivity}) \sim 10^{-6}\ \text{eV}–10^{-5}\ \text{eV} \qquad (4)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

provided that the electron-type neutrino that is created by beta decay contains appreciable portions of at least two different neutrino mass states. Technically, this amounts to saying that the mixing angle between different neutrino states is relatively large. Neutrino mixing is analogous to photon polarization in that photons can be in different polarization states (e.g. linear or circular) and there are mixing angles which describe the relations between the different states of the photon.

Laboratory experiments have achieved a sensitivity to electron-type neutrino masses of order 1 eV. Observations of neutrinos produced by COSMIC RAY interactions in the Earth's atmosphere have suggested the existence of neutrino oscillations involving neutrino mass differences of order ~0.07 eV.

Resonant neutrino oscillations, which may be induced by neutrino interactions with electrons in the Sun (the famous Mikheyev–Smirnov–Wolfenstein (MSW) effect), can occur even if the electron neutrino produced in the Sun is not well mixed with any of the other neutrinos (i.e. even if the mixing angles between e and μ neutrinos and between e and τ neutrinos are tiny). Standard solar models indicate that the Sun has a high central density, $\rho(\text{central}) \sim 1.5 \times 10^2$ g cm^{-3}, which allows electron-type neutrinos to be resonantly converted to the more difficult to detect μ or τ neutrinos by the MSW effect, provided that the difference in neutrino mass satisfies a resonance condition involving the local electron density and the neutrino energy.

Given the solar parameters from standard models, the planned and operating solar neutrino experiments are sensitive to neutrino masses in the range

$$10^{-4}\ \text{eV} \lesssim m_\nu \lesssim 10^{-2}\ \text{eV} \qquad (5)$$

via matter-induced resonant oscillations (MSW effect).

The range of neutrino masses given by equations (4) and (5) is included in the range of neutrino masses that are suggested by attractive particle-physics generalizations of the standard electroweak model.

Both vacuum neutrino oscillations and matter-enhanced neutrino oscillations can change electron-type neutrinos to the more difficult to detect μ or τ neutrinos. In addition, the likelihood that a neutrino will have its type changed may depend on its energy, affecting the shape of the energy spectrum of the surviving electron-type neutrinos. The shape of the energy spectrum for solar neutrinos produced by any particular nuclear process must be the same as the shape measured in the laboratory unless some new physics (e.g. neutrino oscillations) is occurring. Currently operating solar neutrino experiments will measure the shape of the energy spectrum.

What does the combined standard model tell us about solar neutrinos?

In this section, I will describe the combined standard model (standard solar model and standard electroweak theory) that is used to decide whether solar neutrino experiments have revealed something unexpected. Then I will present the calculated solar neutrino spectrum as predicted by the standard model.

The combined standard model

In order to interpret solar neutrino experiments, one must have a quantitative solar model. Unlike many other areas of astronomy in which one can make important discoveries by identifying new classes of objects (such as quasars, or x-ray sources, or γ-ray sources), solar neutrino research requires a reliable theoretical model for comparison with the observations in order to determine whether one has found something surprising. Our physical intuition is not yet sufficiently advanced to know whether we should be surprised by, for example, 10^{-2}, by 10^0 or by 10^2 neutrino-induced events per day in a chlorine tank the size of an olympic swimming pool.

I will use the most conservative model for comparison with experiments, the combined standard model, unless explicitly stated otherwise. The combined standard model is the standard model of solar structure and evolution and the standard electroweak model of particle physics (see also SOLAR EVOLUTION, STANDARD MODEL OF PARTICLE PHYSICS).

A solar model is required in order to predict the number of neutrinos created in a given energy range per unit time. On a fundamental level, a solar model is required in order to predict the rate of nuclear fusion by the p–p chain (shown in table 1 and discussed below) and the rate of fusion by the CNO reactions (originally favored by H BETHE in his epochal 1939 study of nuclear fusion reactions).

In our discussion, I will assume a result common to all modern solar models, namely that the CNO reactions contribute only a very small fraction of the luminosity of the Sun. Although the dominance of the p–p chain is often taken for granted in theoretical analyses of solar neutrino experiments, it is not an *a priori* obvious result.

A precise solar model is required to calculate accurately which nuclear reaction occurs more often at the two principal branching points of the p–p fusion chain. Referring to table 1, the branching points occur between reactions 4 and 5 and between reactions 6 and 8. If the p–p chain is terminated by reaction 4, only low-energy (<0.4 MeV) p–p neutrinos are produced, but if the termination occurs via reaction 5, then higher-energy ^{7}Be and ^{8}B neutrinos are created.

Many authors have claimed that the p–p neutrino flux is essentially determined by the solar luminosity. This is wrong. Depending on whether reaction 4 or reaction 5 is dominant, the flux of p–p neutrinos can vary by a factor 2. If reaction 4 is dominant, two p–p neutrinos are produced each time four protons are burned to produce an alpha particle (see equation (1)). If reaction 5 is dominant, only one p–p neutrino plus a ^{7}Be neutrino (or rarely a ^{8}B neutrino) is produced for each alpha particle that is formed. A solar model is required to decide which reaction, 4 or 5, is dominant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The ratio of the rates for reaction 6 and reaction 8 determines how often ^{7}Be neutrinos (two lines: 0.86 MeV and 0.38 MeV) are produced rather than the rare, but more easily detected, ^{8}B neutrinos (maximum energy ~14 MeV) are produced. The predicted rates in solar neutrino experiments depend sensitively on the relative frequencies of these crucial reactions. Fortunately, the theoretical uncertainties in the predicted neutrino fluxes are not very large. For the important fluxes, the uncertainties vary from ~1% to ~20%, depending on the neutrino source in question. The rates for individual detectors are determined by the energy spectrum, by the neutrino type of the incoming solar neutrinos and by the interaction cross sections of the different detectors.

A particle physics model is required to predict what happens to the neutrinos after they are created. I will use the simplest version of the standard electroweak model, according to which nothing happens to the neutrinos after they are created in the interior of the Sun. In this theory, neutrinos are massless and neutrino flavor (electron type) is conserved. The standard electroweak model has had many successes in precision laboratory tests; modifications of this theory will be accepted only if incontrovertible experimental evidence forces a change.

The solar neutrino spectrum

Figure 1 shows the calculated neutrino spectrum for the most important neutrino sources from the Sun. I will discuss briefly the p–p (and pep) neutrinos, the ^{7}Be neutrinos, and the ^{8}B neutrinos. I will concentrate on the reliability of the predictions and will indicate the role of each of these neutrinos in the ongoing experiments.

The dominant source of solar neutrinos is the first reaction listed in table 1, the basic p–p reaction ($p + p \longrightarrow D + e^+ + \nu_e$), which creates neutrinos with energies less than 0.4 MeV. Most of the nuclear energy that emerges as sunlight begins with this reaction. The theoretical uncertainty in the p–p neutrino flux is about 1%. Among the solar neutrino experiments that are currently operating or that are being constructed, only the GALLEX and SAGE gallium experiments have energy thresholds low enough to detect the p–p neutrinos. About 0.2% of the p–p fusions are believed to occur via the pep reaction, the second reaction in table 1. This reaction produces a neutrino with a fixed energy, a 'neutrino line' that contributes a small part of the calculated event rate in the chlorine and gallium experiments.

The next most important source of neutrinos is from the ^{7}Be neutrino line at 0.86 MeV, which is produced by reaction 6 of table 1. About 15% of the solar luminosity is produced by reactions which go through this channel; the uncertainty in the neutrino flux is ~9%. The ^{7}Be neutrinos contribute significantly, according to standard model calculations, to the chlorine and the gallium experiments, but are too low in energy to be detected in the Kamiokande experiment. In an experiment under development called BOREXINO, ^{7}Be neutrinos will be detected by the unique signature they produce in scintillation light caused by neutrino–electron scattering.

The ^{8}B neutrino flux, produced by reaction 9 of table 1, is tiny, $\sim 10^{-4}$ of the flux of p–p neutrinos. However, the ^{8}B neutrinos are crucial for solar neutrino physics and astronomy. Because of their high energy (~10 MeV, which takes advantage of a superallowed transition from the nuclear ground state of chlorine to an excited state of argon), ^{8}B neutrinos dominate the predicted capture rate for the chlorine experiment. They are also the only significant source of neutrinos above the energy threshold in the water Cherenkov experiments, Kamiokande, Super-Kamiokande and SNO. Unfortunately, the theoretical uncertainty in the predicted ^{8}B neutrino flux is relatively large, ~19%.

Why are the predicted neutrino fluxes robust?

The predicted event rates in the different solar neutrino experiments have been remarkably stable over the past 30 yr. Our published estimate in 1968, which accompanied the first report by Davis and his collaborators of measurements with the chlorine experiment, was 7.5 ± 1.0 solar neutrino units (SNU); the most recent and detailed calculation yielded in 1998 a predicted rate of $7.7^{+1.2}_{-1.0}$ SNU. An SNU is a convenient unit to describe the measured rates of solar neutrino experiments: 10^{-36} interactions per target atom per second. The theoretical errors are intended to be as close as possible to effective 1σ errors; they are obtained by carrying out detailed calculations using 1σ uncertainties on all the measured input data and, for the theoretical errors (which are generally less important), by taking the extreme range of theoretical calculations to be 3σ uncertainties.

In the intervening three decades since the first experimental report on solar neutrinos, my colleagues and I have calculated many different models with steady improvements in measured and calculated input data and in the physics used to describe the solar interior. The best-estimate predictions have bounced around every few years, as different improvements were included, but the answers have always remained within ± 2 SNU of the 1968 best estimate.

There are three reasons that the theoretical calculations of the neutrino fluxes are robust: (1) the availability of precision measurements and precision calculations of input data; (2) the connection between neutrino fluxes and the measured solar luminosity; (3) the measurement of the helioseismological frequencies of the solar pressure-mode (p-mode) eigenfrequencies.

Over the past three decades, many hundreds of researchers have performed precision measurements of crucial input data including nuclear reaction cross sections and the abundances of the chemical elements on the solar surface. Many other researchers have calculated accurate opacities, equations of state and weak interaction cross sections. By now, these input data are relatively precise and their uncertainties are quantifiable.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

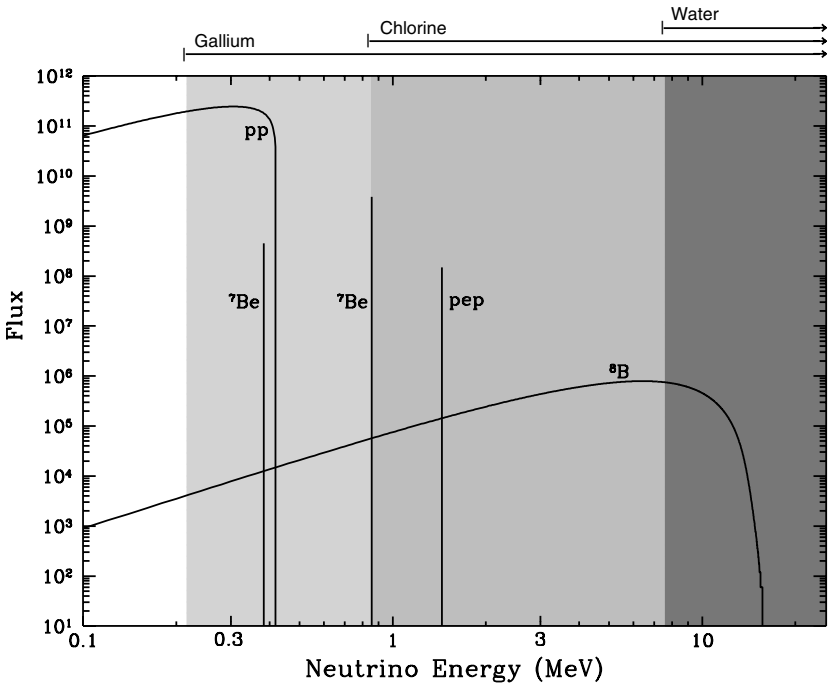


Figure 1. Solar neutrino spectrum. This figure shows the energy spectrum of neutrinos from the p–p chain that is predicted by the standard solar model. The neutrino fluxes from continuum sources (p–p and ^{8}B) are given in the units of number per cm^2 per second per MeV at 1 astronomical unit. The line fluxes (pep and ^{7}Be) are given in number per cm^2 per second. The arrows at the top of the figure indicate the energy thresholds for the ongoing neutrino experiments. The higher-energy ^{7}Be line is just above threshold in the chlorine experiment. For simplicity, CNO neutrinos are omitted.

The solar neutrino fluxes and the solar luminosity both depend on the rates of the nuclear fusion reactions in the solar interior. Since we know experimentally the solar luminosity (to an accuracy of ∼0.4%), the calculated neutrino fluxes are strongly constrained by the fact that the standard solar models must yield precisely the measured solar luminosity.

Thousands of p-mode helioseismological frequencies have been measured to an accuracy of 1 part in 10^4. The standard solar models discussed here reproduce these p-mode frequencies to an rms accuracy of better than 1 part in 1000. In fact, the standard solar models are in agreement with the measured helioseismological frequencies to a high level of precision without any adjustment of parameters.

The calculated solar neutrino fluxes are, after 30 yr of intense study, known to reasonable accuracy because of the many precise measurements and calculations of input data, because of the strong constraint imposed on the models by the measured total solar luminosity and because of the important tests of solar structure that are provided by helioseismological measurements (see HELIOSEISMIC OBSERVATIONS).

What are the three solar neutrino problems?

I will compare in this section the predictions of the combined standard model with the results of the operating solar neutrino experiments. We will see that this comparison leads to three different discrepancies between the calculations and the observations, which I will refer to as the three solar neutrino problems.

Figure 2 shows the measured and the calculated event rates in the five solar neutrino experiments (Homestake, GALLEX, SAGE, Kamiokande and Super-Kamiokande) for which quantitative results are available at the time of this writing. The figure reveals three discrepancies between the experimental results and the expectations based on the combined standard model. As we shall see, only the first of these discrepancies depends at all sensitively on predictions of the standard solar model.

Calculated versus observed chlorine rate

The first solar neutrino experiment to be performed was the chlorine radiochemical experiment, which detects electron-type neutrinos that are more energetic than 0.81 MeV. After more than 30 yr of the operation of this experiment, the measured event rate is 2.56 ± 0.23 SNU, which is a factor 3 less than is predicted by the most detailed theoretical calculations. Most of the predicted rate in the chlorine experiment is from the rare, high-energy ^{8}B neutrinos, although the ^{7}Be neutrinos are also expected to contribute significantly. According to standard model calculations, the pep neutrinos and the CNO neutrinos (for simplicity not discussed here) are expected to contribute less than 1 SNU to the total event rate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

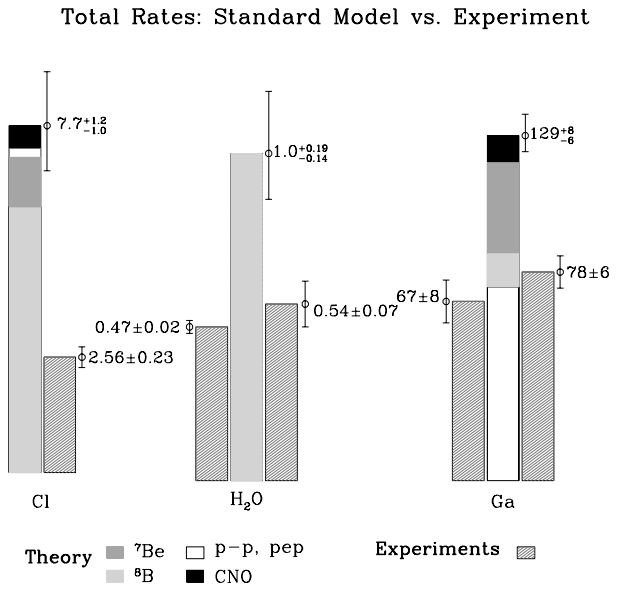


Figure 2. Comparison of measured rates and standard-model predictions for four solar neutrino experiments.

This discrepancy between the calculations and the observations for the chlorine experiment was, for more than two decades, the only solar neutrino problem. I shall refer to the chlorine disagreement as the 'first' solar neutrino problem. (It used to be called 'the' solar neutrino problem.)

Incompatibility of chlorine and water experiments

The second solar neutrino problem results from a comparison of the measured event rates in the chlorine experiment and in the Japanese pure-water experiments, Kamiokande and Super-Kamiokande. The water experiments detect higher-energy neutrinos, those with energies above 7.5 MeV (Kamiokande) or 6.5 MeV (Super-Kamiokande, so far), by neutrino–electron scattering: $\nu + e \longrightarrow \nu' + e'$. According to the standard solar model (see also table 1), ^{8}B beta decay is the only important source of these higher-energy neutrinos.

The Kamiokande and Super-Kamiokande experiments show that the observed neutrinos come from the Sun. The electrons that are scattered by the incoming neutrinos recoil predominantly in the direction of the Sun–Earth vector; the relativistic electrons are observed by the Cherenkov radiation they produce in the water detector.

In addition, the water experiments measure the energies of individual scattered electrons and therefore provide information about the energy spectrum of the incident solar neutrinos. The currently most popular parameters for neutrino oscillations predict only relatively small deviations from the spectrum expected with no oscillations. Indeed, Super-Kamiokande does observe a deviation that can be explained by neutrino oscillations but as of the time of this writing the statistical significance is not very high.

The event rate in the Kamiokande experiment is determined by the same high-energy ^{8}B neutrinos that are expected, on the basis of the combined standard model, to dominate the event rate in the chlorine experiment. I have shown elsewhere that solar physics changes the shape of the ^{8}B neutrino spectrum by less than 1 part in 10^5. Therefore, we can calculate the rate in the chlorine experiment that is produced by the ^{8}B neutrinos observed in the Super-Kamiokande experiment (6.5 MeV threshold energy). This partial (^{8}B) rate in the chlorine experiment is 2.78 ± 0.10 SNU, which exceeds by about one standard deviation the total observed chlorine rate from all neutrino sources, of 2.56 ± 0.23 SNU.

Comparing the rates of the Kamiokande and the chlorine experiments, one finds that the net contribution to the chlorine experiment from the pep, ^{7}Be and CNO neutrino sources is negative: -0.22 ± 0.10 SNU. The standard model calculated rate from pep, ^{7}Be and CNO neutrinos is 1.9 SNU.

The apparent incompatibility of the chlorine and the Kamiokande experiments is the 'second' solar neutrino problem. The inference that is often made from this comparison is that the energy spectrum of ^{8}B neutrinos is changed from the standard shape by physics not included in the simplest version of the standard electroweak model.

Gallium experiments: no room for ^{7}Be neutrinos

The results of the gallium experiments, GALLEX and SAGE, constitute the third solar neutrino problem. The average observed rate in these two experiments is 73 ± 5 SNU, which is essentially fully accounted for in the standard model by the theoretical rate of 72 SNU that is calculated to come from the basic p–p and pep neutrinos (with only a 1% uncertainty in the standard solar model p–p flux). The ^{8}B neutrinos, which are observed above 6.5 MeV in the Super-Kamiokande experiment, must also contribute to the gallium event rate. Using the standard shape for the spectrum of ^{8}B neutrinos and normalizing to the rate observed in Kamiokande, ^{8}B contributes another 6 SNU, unless something happens to the lower-energy neutrinos after they are created in the Sun. (The predicted contribution is 12 SNU on the basis of the standard model.) Given the measured rates in the gallium experiments, there is no room for the additional 34 ± 4 SNU that is expected from ^{7}Be neutrinos.

The seeming exclusion of everything but p–p neutrinos in the gallium experiments is the 'third' solar neutrino problem. This problem is essentially independent of the previously discussed solar neutrino problems, since it depends on the p–p neutrinos that are not observed in the other experiments and whose calculated flux is approximately model independent (if the general scheme of the p–p chain shown in table 1 is correct).

The missing ^{7}Be neutrinos cannot be explained away by any change in solar physics. The ^{8}B neutrinos that are observed in the Kamiokande experiment are produced in competition with the missing ^{7}Be neutrinos;

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the competition is between reaction 6 and reaction 8 in table 1. Solar model explanations that reduce the predicted ^{7}Be flux reduce much more (too much) the predictions for the observed ^{8}B flux.

I conclude that either (1) at least three of the four operating solar neutrino experiments (the two gallium experiments plus either chlorine or Kamiokande) have yielded misleading results or (2) physics beyond the standard electroweak model is required to change the neutrino energy spectrum (or flavor content) after the neutrinos are produced in the center of the Sun.

What have we learned?

No solar-model solution has been found that explains the results of the five existing solar neutrino experiments. Many particle physics solutions have been proposed that can explain the existing data. In this section, I will summarize the main astronomical lessons that have been learned from the first 30 yr of solar neutrino research.

The chlorine solar neutrino experiment was proposed in 1964 as a practical test of solar model calculations. The only motivation presented in the theoretical and experimental papers was to use neutrinos 'to see into the interior of a star and thus verify directly the hypothesis of nuclear energy generation in stars'.

What have we learned by direct experiments about nuclear energy generation in stars? How does our 1964 understanding compare with the results of the solar neutrino experiments? Table 2 summarizes the six principal predictions that were made (or which were implicit in the theory) in 1964 and compares those predictions with the results of the four ongoing solar neutrino experiments.

The neutrinos were predicted to originate in the solar interior; the direction of origin of the neutrinos has been verified by detecting neutrino–electron scattering (as was also suggested in 1964) in the Kamiokande and Super-Kamiokande experiments.

The rates of the four operating experiments are in semiquantitative agreement with the predictions; the ratios of the observed to the predicted rates are 0.3 (chlorine), 0.5 (neutrino–electron scattering in water), and 0.6 (gallium, average). This agreement is better than any of us dared hope for in 1964, especially since the dominant neutrino flux (from ^{8}B beta decay) for the first two experiments depends on the central temperature of the Sun as approximately the 24th power of the central temperature.

The energy range of the dominant neutrinos was predicted to be from 0 MeV to 14 MeV, which is consistent with the observations from the Kamiokande and Super-Kamiokande experiments.

Standard models predict that the neutrino fluxes are constant in time except for a small seasonal variation. (The Kelvin–Helmholtz cooling time for the solar interior is $\sim 10^7$ yr.) The consensus view of the experimentalists is that there are no statistically significant variations in the available data. Small deviations from the constant flux prediction are predicted by some of the popular neutrino oscillation scenarios, but improved statistics are required to test these predictions.

Standard solar models predict that the Sun shines almost entirely via the p–p chain of nuclear fusion reactions, rather than the CNO reactions originally emphasized by Bethe. If CNO reactions were dominant, the event rates in solar neutrino experiments could be calculated precisely. These 'all-CNO' rates, as shown in table 2, differ from the observed rates by more than an order of magnitude. Finally, if we crudely characterize the rate of the ^{8}B neutrino emission by its approximate dependence on the central temperature of the solar model, then the central temperature of the solar model agrees with the value obtained from the experimental rates to an accuracy of $\sim$2% or better.

The pioneering solar neutrino experiments have shown directly that the Sun shines by nuclear fusion reactions, thus achieving the original goal proposed in 1964. Quantitative improvements in the tests shown in table 2 will occur with the next generation of experiments, which should also refine our knowledge of the physical properties of neutrinos. However, the most important qualitative result has been established: *neutrinos have been observed from the interior of the Sun in approximately the number and with the energies expected*.

In the more than three-decade long struggle to improve solar models in order to calculate more accurate solar neutrino fluxes, we have obtained a greater understanding of solar structure. The theoretical models have gradually been refined as improved input data, more accurate physical descriptions and more precise numerical techniques have been employed. Perhaps most importantly, the complementary field of helioseismology has been developed and now provides precise data that determine the sound velocity over most of the solar interior; these beautiful measurements are used to test and to refine the standard solar model. Further improvements in the solar model are desirable and important, but the quantitative agreement, typically better than 1 part in 1000, between the calculated eigenfrequencies of pressure modes and the measured (helioseismological) frequencies provides strong evidence for the basic correctness of the standard solar model.

What next?

In this section, I will summarize the goals of solar neutrino research, first in physics and then in astronomy, during the next decade or two.

Physics goals

The fundamental goal of physics research with solar neutrinos is to measure the energy spectrum and flavor content as a function of time of the solar neutrino flux. We want to know how many neutrinos reach the Earth with a given energy and with a given flavor (i.e. neutrino type: e, μ or τ), all as a function of time. Because of some exotic particle physics possibilities, we also want to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Predictions versus observations: 1964 versus 1995.

Property	Predicted	Observed
Direction	From the Sun	OK
Rates	Measurable	~ Predicted rates (within factor of few)
Neutrino energy	0–15 MeV	<15 MeV
Time dependence	Constant (except seasonal)	OK
p–p not CNO	If CNO: Cl : 28 SNU	2.6 SNU
	Ga : 610 SNU	73 SNU
	H_2O : 0.0	0.44 standard model
Central temperature	16×10^6 K	$T(^8B)/T_{model} \gtrsim 0.98$

know whether the solar neutrino flux contains any anti-neutrinos.

The standard model predicts that the energy spectrum of neutrinos from any given neutrino source, e.g. from ^{8}B beta decay, will be the same to high accuracy as the energy spectrum inferred from terrestrial laboratory measurements. In the standard electroweak theory, only massless electron-type neutrinos are created in nuclear beta decay or nuclear fusion reactions. Standard electroweak theory predicts that the solar neutrinos produced by nuclear fusion reactions are all ν_e, not ν_μ or ν_τ. (According to MSW and vacuum oscillation theories, neutrinos created in nuclear beta decay or nuclear fusion reactions are linear combinations of different neutrino types and at least one neutrino type has a non-zero mass.) Finally, the total amount of thermal energy in the solar interior implies that the neutrino fluxes will be constant in time (for time scales less than 10^7 yr) except for the seasonal dependences caused by the Earth's orbital eccentricity. Any departure from these expectations will be a signal of physics beyond the standard electroweak model.

Physicists want to use solar neutrino experiments to measure, or to set stringent limits on, the elementary properties of neutrinos, especially their masses and mixing angles. It seems likely that we will make important progress toward this goal in the next decade.

Astronomy goals

The fundamental goal of solar neutrino astronomy is to determine the rates of different nuclear fusion reactions in the solar interior. Neutrino fluxes created by the different nuclear sources are the signatures of the fusion reactions. We must know what happens to the neutrinos after they are created in order to infer the created neutrino energy spectrum from the measured neutrino energy spectrum.

Progress in solar neutrino astronomy is held hostage to progress in particle physics. As discussed in the previous subsection, it seems likely that we will learn enough about the particle physics in the next decade to permit accurate inferences about the rates of neutrino creation in the Sun from the observed rates of neutrino arrival at the Earth. The discussion in this subsection presumes that the required progress in understanding the properties of the neutrino will be achieved.

Completing hydrogen fusion Table 1 shows that the two principal ways of completing nuclear fusion in the sun are reactions 4 and 5, the so-called ^{3}He–^{3}He and ^{3}He–^{4}He reactions. Because of the slightly smaller reduced mass that exists for the ^{3}He–^{3}He reaction, Coulomb barrier penetration favors this reaction over the ^{3}He–^{4}He reaction at lower temperatures. According to the standard solar model, the ^{3}He–^{4}He reaction is dominant in the innermost region of the Sun (where it is 1.5 times faster than the ^{3}He–^{3}He reaction), but overall occurs in only ~15% of the fusion terminations that are described by equation (1). That is, in the most detailed solar models, the ^{3}He–^{3}He reaction is on average more than 6 times faster in completing the nuclear fusion of protons into α-particles than the competing ^{3}He–^{4}He reaction.

Is this prediction of the standard solar model correct? A determination of the p–p and ^{7}Be neutrino fluxes (corrected for what non-standard particle physics has done to them after they were created in the Sun) can answer this important question. The average ratio of the total number of ^{3}He–^{4}He reactions per unit time in the Sun to the total number of ^{3}He–^{3}He reactions per unit time in the Sun is

$$\frac{\langle {}^3\text{He}\text{–}{}^4\text{He}\rangle}{\langle {}^3\text{He}\text{–}{}^3\text{He}\rangle} = \frac{2\phi(^7\text{Be})}{\phi(\text{p–p}) - \phi(^7\text{Be})} \qquad (6)$$

where ϕ(p–p) and $\phi(^7$Be) are, respectively, the fluxes from the p–p and ^{7}Be neutrinos.

Equation (6) is the most precisely testable prediction that I know of that follows directly from the theory of stellar energy generation. The known theoretical uncertainties in the calculation of the average solar ratio of ^{3}He–^{4}He to ^{3}He–^{3}He reactions is 9%.

The ^{8}B neutrino flux The flux of neutrinos from ^{8}B beta decay in the Sun (see reaction 9 of table 1) is, in principle, the simplest solar neutrino flux to measure. The higher energies of the ^{8}B neutrinos make them easiest to detect. For this reason, the Kamiokande, Super-Kamiokande and SNO neutrino experiments will all concentrate on the ^{8}B neutrinos.

However, one must determine the total flux of ^{8}B neutrinos, including the more difficult to detect μ or τ neutrinos that may have been produced by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

neutrino oscillations from the originally created electron-type neutrinos. The total number of neutrinos of all types will be measured directly in the SNO experiment via the neutral-current disintegration of deuterium and, less directly, via electron–neutrino scattering in Super-Kamiokande. (This statement presumes there are no sterile neutrinos, i.e. neutrinos that do not interact with matter.)

The magnitude of the ^{8}B flux (all neutrino flavors), which is a sensitive probe of the temperature of the solar interior, varies approximately as $T_{\rm central}^{24}$. Therefore, it is important to determine experimentally the total ^{8}B solar neutrino flux.

The temperature profile of the solar interior A precision test of the theory of stellar structure and stellar evolution can be performed by measuring the average difference in energy between the neutrino line produced by ^{7}Be electron capture in the solar interior and the corresponding neutrino line produced in a terrestrial laboratory. This energy shift is calculated to be 1.29 keV. The energy shift is approximately equal to the average temperature of the solar core, computed by integrating the temperature over the interior of a standard solar model with a weighting factor equal to the locally produced ^{7}Be neutrino emission. The total range of values for the shift, calculated for a number of modern solar models (going back to 1982), is 0.06 keV.

A measurement of the energy shift is equivalent to a measurement of the central temperature distribution of the Sun.

The calculated energy profile of the ^{7}Be line contains, analogous to line-broadening in classical (photon) astronomy, information about the distribution of solar interior temperatures. The theoretical shape of the ^{7}Be neutrino line is asymmetric: on the low-energy side, the line shape is Gaussian with a half-width at half-maximum of 0.6 keV, and, on the high-energy side, the line shape is exponential, with a half-width at half-maximum of 1.1 keV.

The calculated shape of the ^{7}Be neutrino line is not affected significantly by vacuum neutrino oscillations, the MSW effect or other frequently discussed weak interaction solutions to the solar neutrino problems. This is a key result: it implies that the astronomical information contained in the line shift and in the line profile is not dependent on further progress in neutrino physics.

Detectors are available that have the resolution to measure the line shift. Unfortunately, their current sizes are too small to permit a full-scale solar neutrino experiment. However, proposals have been made in the literature for developing detectors that are sufficiently large to be able to measure well the average shift in energy of the solar neutrino line.

More complete models of the Sun The accuracy of the physical description that is currently achieved with one-dimensional (spherically symmetric) models of the Sun that include diffusion is sufficient to permit excellent quantitative agreement with the measured p-mode oscillation frequencies. Numerical experiments and theoretical arguments also suggest that further improvements are unlikely to affect significantly the calculated neutrino fluxes.

Nevertheless, current models of the Sun are incomplete. They are spherically symmetric and do not take account of the two-dimensional (or three-dimensional) nature of solar structure. They do not contain a self-consistent dynamical treatment of the effects of rotation, of magnetic fields, of mass loss or of other possible effects that may violate the currently used approximations of spherical symmetry and quasi-static evolution. We know observationally that the Sun (at least near its surface) contains magnetic fields, that it is losing mass, and that it departs from spherical symmetry by $\sim$1 part in 10^5.

There are both analytic and calculational challenges in including these complicated processes in a more complete physical description in the next generation of solar models. New self-consistent methods of calculating solar models (and stellar models) must be developed, and then the appropriate numerical techniques must be worked out, tested and applied.

The goal of developing a more complete solar model is a challenge for the next decade and beyond. Fortunately, it is a challenge that could lead to important progress since computing power is much greater than it was in the past and there is an abundance of precision data with which to make detailed comparisons.

Summary

The first 30 yr of solar neutrino research have verified experimentally the fundamental predictions of nuclear energy generation in stars. The next 10 or 20 yr of research will, I think, concentrate on using solar neutrinos to learn more about weak interaction physics. As the weak interaction questions are being resolved, it will be possible to carry out progressively more accurate tests of the theory of nuclear energy generation and of stellar structure.

In retrospect, the history of solar neutrino research seems ironic. It began with an effort to use neutrinos, whose properties were assumed to be well known, to study the interior of the nearest star. The project was an unconventional application of microscopic physics that was designed to carry out a unique investigation of a massive, macroscopic body, the Sun. It now appears likely that a large community of physicists, chemists, astrophysicists, astronomers and engineers working together may have stumbled across the first observed manifestation of physics beyond the standard electroweak model.

We may have been incredibly lucky.

Bibliography

Bahcall J N 1989 *Neutrino Astrophysics* (Cambridge: Cambridge University Press)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chlorine collaboration:

Cleveland B T *et al* 1998 Measurement of the solar electron neutrino flux with the Homestake chlorine detector *Astrophys. J.* **496** 505–26

Kamiokande collaboration:

Fukuda Y *et al* 1996 Solar neutrino data covering solar cycle 22 *Phys. Rev. Lett.* **77** 1683–6

GALLEX collaboration:

Hampel W *et al* 1996 GALLEX solar neutrino observations: results for GALLEX III *Phys. Lett.* B **388** 384–96

SAGE collaboration:

Gavrin V *et al* 1997 Solar neutrino and ^{51}Cr results from SAGE *Neutrino 96: Proc. XVIIth Int. Conf. on Neutrino Physics and Astrophysics (Helsinki)* ed K Huitu, K Enqvist and J Maalampi (Singapore: World Scientific) pp 14–24

Super-Kamiokande collaboration:

Suzuki Y 1999 *Neutrino 98: Proc. XVIIIth Int. Conf. on Neutrino Physics and Astrophysics (Takayama, 4–9 June 1998)* ed Y Suzuki and Y Totsuka *Nucl. Phys.* B (Proc. Suppl.) **77** 35–42

Bahcall J N, Basu S and Pinsonneault M H 1998 How uncertain are solar neutrino predictions? *Phys. Lett.* B **433** 1–8

Recent results on standard solar models and on neutrino oscillations are available at

http://www.sns.ias.edu/~jnb

John N Bahcall

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

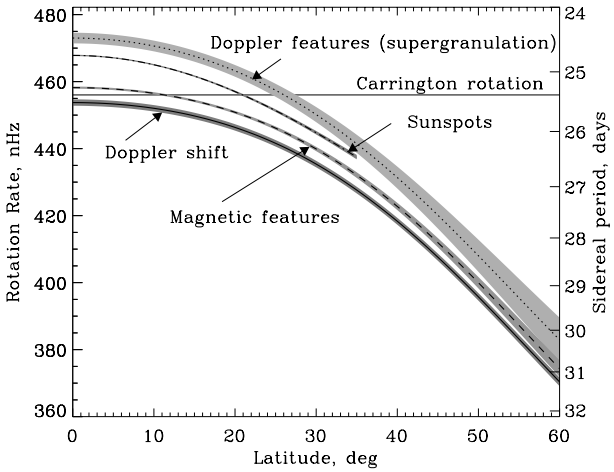


Figure 1. Rotation rate, $\Omega/2\pi$, and period of various tracers on the Sun's surface: recurrent (old) sunspots (dashed curve), magnetic features (dot–dash), and Doppler features (dots). The rotation rate and period determined spectroscopically through the Doppler shift are shown by the full curve. The shaded areas show the 1σ error estimates.

Solar Interior: Rotation

The Sun rotates differentially, i.e. the solar rotation rate varies with both latitude and radius. The differential rotation is particularly prominent in the convection zone where equatorial zones rotate almost 30% faster than near-polar regions. There are also strong variations of the rotation rate with radius near the surface and at the bottom of the convection zone (rotational shear layers). In contrast to the convection zone, the radiative zone rotates almost uniformly, although the rotation rate of the inner solar core has not been determined reliably. The mechanisms of the differential rotation, which are probably due to the interaction between convection and rotation, are not fully understood. The differential rotation is believed to play a major role in the dynamo mechanism of solar activity. It also affects the internal structure of the Sun because of rotational instabilities related to rotational shear layers. Knowledge of the internal rotation (in particular, whether the inner core rotates rapidly) is important for determining the quadrupole gravitational moment of the Sun and testing Einstein's theory of relativity (SOLAR INTERIOR: SOLAR DIAMETER, OBLATENESS, AND TEMPERATURE).

Measurements of solar rotation

It has been known since the first observations of solar rotation around 1610 by GALILEO GALILEI, Johannes FABRICIUS and Christoph SCHEINER that the solar equator rotates faster than the regions closer to the poles (latitudinal differential rotation). However, the accurate determination of the differential rotation on the solar surface and in the deep interior is still one of the major problems of solar physics. A unique tool for determining the rotation of the Sun's interior is provided by helioseismology.

Differential rotation

There are two basic approaches for measuring the surface rotation: spectroscopic and tracer motion. The spectroscopic method is based on measuring the Doppler line shift. It provides the angular velocity in those layers of the solar atmosphere where the spectral lines chosen for the measurements are formed. In principle, by observing the spectroscopic shift in several lines it is possible to determine the variation of solar rotation with height in the solar atmosphere because different spectral lines are formed in different layers of the atmosphere. Such measurements of the radial variation of the rotation in the surface layers are not particularly accurate, because of the significant noise caused by large-scale convective motions—supergranulation. Nevertheless, these measurements have provided evidence that the lower layers of the solar atmosphere, and thus presumably the subsurface layers, rotate faster. This was later confirmed by helioseismology.

The tracer measurements are carried out either by tracking the rotation rate of individual resolved features on the surface (such as sunspots, active regions, supergranular cells etc) as they pass across the disk, or by cross-correlating magnetograms and Dopplergrams obtained at different times.

Representations of the surface differential rotation

Angular velocity Ω of the surface differential rotation as a function of co-latitude θ is usually represented in a parametric form. Historically, it was given in terms of trigonometric functions:

$$\Omega(\theta) = A + B\cos^2\theta + C\cos^4\theta \qquad (1)$$

where the coefficient A gives the equatorial rotation rate ($\theta = 90^\circ$), and coefficients B and C give the differential rotation. However, the basis functions in this expansion are not orthogonal. This results in correlated errors ('cross-talk') among the coefficients. Therefore, more recently a presentation in terms of orthogonal polynomials has been used:

$$\Omega(\theta) = \bar{A} + \bar{B}T_2^1(\theta) + \bar{C}T_4^1(\theta) \qquad (2)$$

where $T_{2k}^1(\theta) \equiv a_k P_{2k+1}^1(\theta)/\sin\theta$, $a_k = k!2^k/(2k+1)!!$ and $P_\lambda^1(\theta)$ are the associate Legendre functions of order 1 and degree λ. For the first two terms, $T_2^1(\theta) = 5\cos^2\theta - 1$ and $T_4^1(\theta) = 21\cos^4\theta - 14\cos^2\theta + 1, a_0 = 1, a_1 = 2/3, a_2 = 8/15$. Coefficients $\bar{A}$, $\bar{B}$ and $\bar{C}$ are linear combinations of A, B and C of equation (1):

$$\bar{A} = A + \tfrac{1}{5}B + \tfrac{3}{35}C \qquad \bar{B} = \tfrac{1}{5}B + \tfrac{2}{15}C \qquad \bar{C} = \tfrac{1}{21}C.$$

The equatorial angular velocity in this formulation is given by $\bar{A} - \bar{B} + \bar{C}$. The coefficients of the orthogonal representation (equation (2)) are usually obtained with higher accuracy that the coefficients of equation (1).

The surface angular velocity is usually measured in deg day^{-1} or μrad s^{-1}. However, for internal rotation obtained by helioseismology the typical measure

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Coefficients of the solar rotation rate in nHz.

	Tracers			
Coefficients	Sunspots	Doppler features	Magnetic features	Doppler Shift
A	469.39 ± 0.194	473.02 ± 1.59	458.22 ± 0.32	453.76 ± 0.95
B	-92.58 ± 1.94	-77.03 ± 6.05	-53.96 ± 2.07	-54.59 ± 0.80
C		-57.46 ± 8.12	-77.19 ± 3.34	-75.44 ± 1.11
$\bar{A}$	450.87 ± 0.43	452.65 ± 0.95	440.87 ± 0.32	433.71 ± 0.80
$\bar{B}$	-18.52 ± 0.39	-23.08 ± 0.64	-21.17 ± 0.32	-21.17 ± 0.08
$\bar{C}$		-2.71 ± 0.32	-3.66 ± 0.16	-3.50 ± 0.06

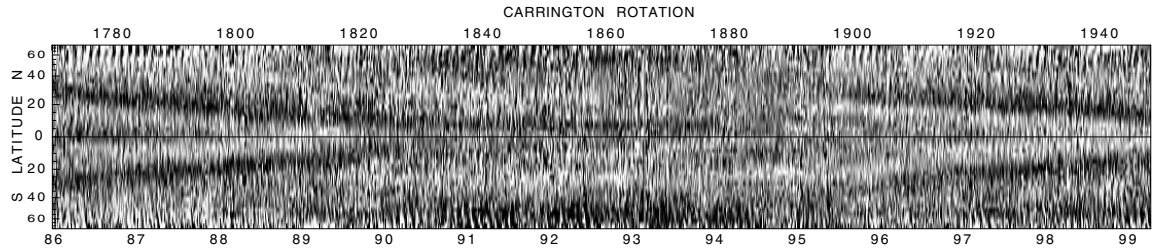


Figure 2. Migrating zonal flows ('torsional oscillations') in 1986–99, determined from the Mt Wilson Observatory data. The vertical scale is latitude in degrees, the lower horizontal axis shows time in years starting from 1986, and the upper scale shows the corresponding Carrington rotation numbers. The dark areas correspond to a rotational velocity 7.5 m s^{-1} faster than the average velocity, and the light area shows rotation 7.5 m s^{-1} slower than the average (Ulrich 1998).

is the rotation rate, $\Omega/2\pi$, expressed in nanohertz (nHz). Therefore, we shall use this measure in this article. The rotation rate can converted in the angular velocity units using the following relations: 1 nHz $= 6.283 \times 10^{-3}$ μrad $s^{-1} = 0.031$ deg day^{-1}. The typical rotation rate at the solar equator is about 460 nHz (2.89 μrad s^{-1} or 14.31 deg day^{-1}), and corresponds to a linear velocity of 2012 m s^{-1}.

Measurements of the surface rotation

The coefficients of the sidereal rotation rate of different features on the solar surface are shown in table 1 and figure 1 (Snodgrass 1992). The surface features (tracers) related to magnetic field structures rotate faster than the solar plasma, rotation of which is measured from the Doppler shift of spectral lines. The magnetic structures are likely to be anchored in the solar interior. Therefore, the faster rotation of these structures probably can be related to higher angular velocity of the interior. The averaged period of rotation of sunspots is close to the 'Carrington period', 25.38 days (the corresponding synodic period is 27.275 days). The Carrington period corresponds to the differential rotation of recurrent (i.e. old) sunspots at approximately 20° latitude and is often used as a reference. The surface rotational velocity at this latitude is approximately 1.9×10^3 m s^{-1}. It is found that the rotation rate of individual sunspots may vary depending their size, age and relative position to other sunspots of the same group.

Observations of solar rotation have also revealed temporal variations related to the 11 yr activity cycle. These include an increase in the equatorial rotation rate at solar minimum and a migrating pattern of zonal flow bands, known as 'torsional oscillations' (figure 2). The bands start at high latitudes near solar maximum and approach the solar equator at solar minimum.

Helioseismic measurements

Helioseismology provides a picture of rotation inside the Sun as a function of both depth and latitude. The helioseismic measurements are based on the frequency shift of resonant solar oscillations (modes) due to rotation. The modes which propagate in the direction of solar rotation have higher frequencies than the modes with the same resonant properties propagating in the opposite direction. This effect called 'rotational frequency splitting' is somewhat analogous to the Zeeman or Stark splitting of energy levels in atoms. The amount of splitting depends on the rotation rate in the mode resonant cavity and their azimuthal order (HELIOSEISMIC OBSERVATIONS, HELIOSEISMOLOGY: THEORY).

Rotational frequency splitting

The mode frequencies, ν_{nlm}, are characterized by three 'quantum numbers': radial order n which is essentially the number of radial nodes in the mode eigenfunction and the angular degree l and azimuthal order m ($-m \leq l \leq m$) which come from the spherical harmonic part of the eigenfunctions, $\Xi_n(r)Y_l^m(\theta, \phi)$. The modes with $m \neq 0$ represent azimuthally propagating waves. The modes with $m > 0$ propagate in the direction of solar rotation and, thus, have higher frequencies in the inertial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

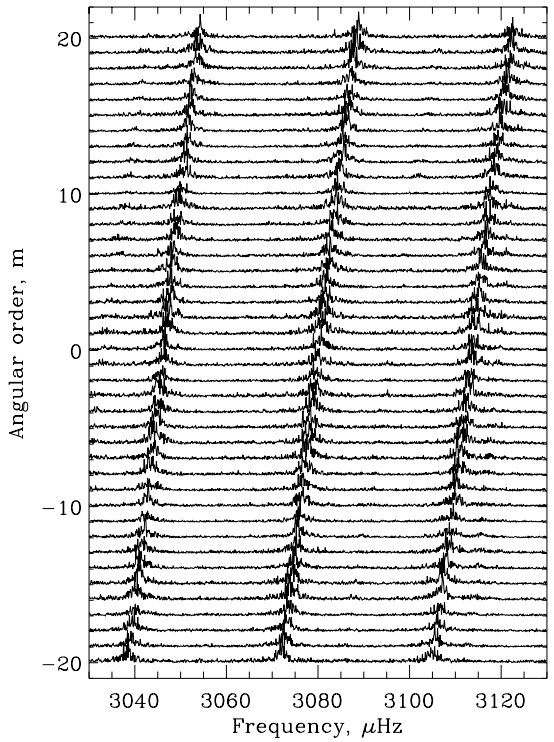


Figure 3. Typical power spectra of solar oscillation data from the MDI instrument on SOHO. Each horizontal curve shows three lines of the power spectrum for different azimuthal order m with radial order $n = 15$ and angular degree $l = 19$, 20, and 21 (from left to right). The slope of the modal lines is due to the rotational frequency shift: prograde modes with positive m have higher frequencies than retrograde modes with negative m.

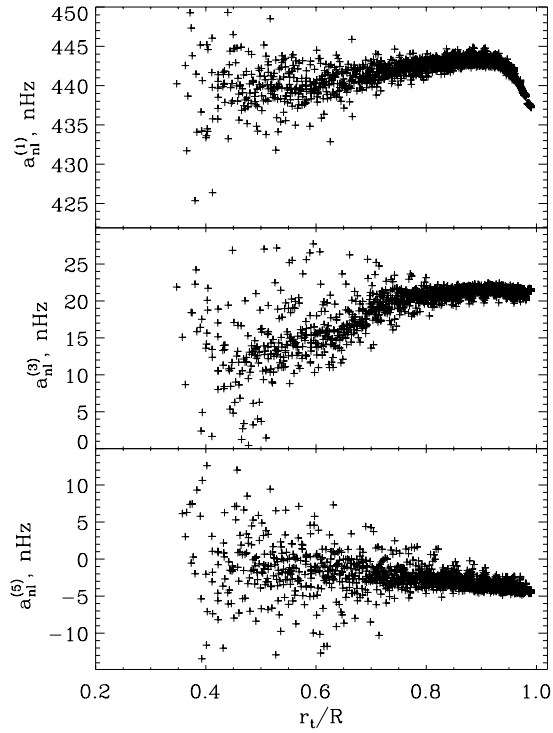


Figure 4. Rotational splitting coefficients obtained from a 144 day observing run from SOHO/MDI. The coefficients are plotted as a function of r_t, the radius of the mode turning point, which is determined for each mode from the relation: $r_t/c(r_t) = L/2\pi\nu_{nl0}$, where $c(r)$ is the sound speed.

frame than the modes with $m < 0$ which propagate in opposite direction. As a result the modes with fixed n and l are split in frequency: $\Delta\nu_{nlm} = \nu_{nlm} - \nu_{nl0}$. Thus, the internal rotation is inferred from splitting of normal mode frequencies with respect to the azimuthal order, m.

An example of the frequency splitting is shown in figure 3, in which the horizontal curves are oscillation power spectra for $m = -20, \ldots, 20$ of modes with $n = 15$ and $l = 19$, 20 and 21. The slope formed by the modal peaks, $\Delta\nu_{nlm}/\Delta m$, is approximately the solar synodic rotation rate 0.43 μHz (or 430 nHz).

The frequency splitting is proportional to m and the internal angular velocity $\Omega(r, \theta)$ averaged over radius r and colatitude θ with splitting kernels $K_{nlm}(r, \theta)$:

$$\Delta\nu_{nlm} = m \int_0^R \int_0^\pi K_{nlm}(r, \theta)\Omega(r, \theta)\, \mathrm{d}\theta\, \mathrm{d}r$$

where R is the solar radius. The kernels K_{nlm} are calculated from mode eigenfunctions.

The determination of $\Omega(r, \theta)$ from a set of observed $\Delta\nu_{nlm}$ represents a helioseismic inverse problem. This problem is solved by a regularized least-squares method or by estimating optimally localized averages $\bar{\Omega}(r_0, \theta_0)$ from a linear combination of the data:

$$\bar{\Omega}(r_0, \theta_0) = \sum_{nlm} c_{nlm}(r_0, \theta_0)\Delta\nu_{nlm}$$

where the coefficients c_{nlm} are chosen to localize the corresponding averaging kernel $A(r, r_0) \equiv \sum_{nlm} c_{nlm}(r_0, \theta_0)K_{nlm}(r, \theta)$ at the target position (r_0, θ_0). When the localization attempt is successful the optimally localized kernels resemble a Gaussian, and $\bar{\Omega}(r_0, \theta_0)$ provides a good estimate of the angular velocity.

The measurements of $\Delta\nu_{nlm}$ are often noisy. Therefore, it is useful to determine the frequency splitting in terms of Legendre polynomials in m/L, where $L = [l(l+1)]^{1/2}$:

$$\Delta\nu_{nlm} = L \sum_{i=1}^{N} a_{nl}^{(i)} P_i\left(\frac{m}{L}\right)$$

where N is typically 5–36. (In some recent measurements the Legendre polynomials are replaced with orthogonal polynomials in m/L.) The odd terms in this equation represent the rotational splitting; the even terms are due to asphericity of the solar structure.

If the internal angular velocity is represented in the form of equation (2) with the coefficients depending on radius r,

$$\Omega(r, \theta) = \bar{A}(r) + \bar{B}(r)T_2^1(\theta) + \bar{C}(r)T_4^1(\theta)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

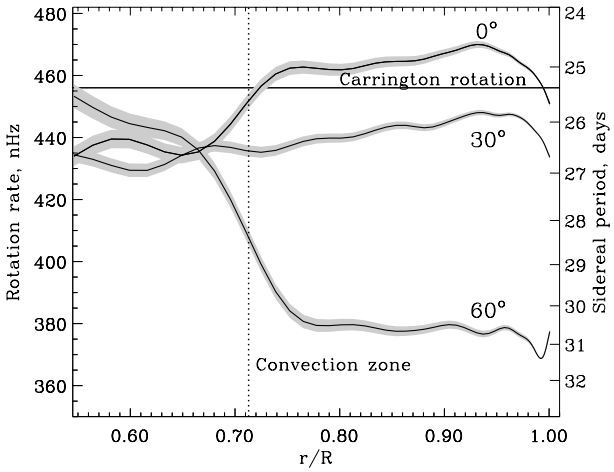


Figure 5. Solar rotation rate inferred from 144 day series of MDI medium-*l* data as a function of radius at three latitudes, 0°, 30° and 60°. The formal errors are indicated by the shaded regions. The vertical straight line shows the lower boundary of the convection zone (Schou *et al* 1998).

then each of the coefficients in this form is primarily determined from the individual $a_{nl}^{(i)}$ coefficients:

$$a_{nl}^{(1)} = \int_0^R K_{nl}^{(1)}(r)\bar{A}(r)\,\mathrm{d}r$$
$$a_{nl}^{(3)} = -\int_0^R K_{nl}^{(3)}(r)\bar{B}(r)\,\mathrm{d}r$$
$$a_{nl}^{(5)} = \int_0^R K_{nl}^{(5)}(r)\bar{C}(r)\,\mathrm{d}r. \qquad (3)$$

The radial functions $\bar{A}(r)$, $\bar{B}(r)$ and $\bar{C}(r)$ can be determined by solving the one-dimensional integral equations (3).

For acoustic (p) modes the seismic kernels $K_{nl}^{(i)}(r)$ are significant mainly inside the mode propagation region, (r_t, R), where r_t is the radius of the inner turning point of acoustic modes and R is the solar radius. Therefore, it is useful to plot the a coefficients as a function of r_t (figure 4). Each point in this plot represent an average of $\bar{A}(r)$, $\bar{B}(r)$ and $\bar{C}(r)$ in the interval (r_t, R). The first coefficient, $a^{(1)}$, which corresponds to the spherically symmetric component of the angular velocity, $\bar{A}(r)$, shows a steep decrease with r_t near the surface. This corresponds to the increase of the angular velocity with depth in the subsurface layers.

The coefficient $a^{(3)}$, which determines the prime component of the latitudinal differential rotation, shows a decrease below $0.7R$, the lower boundary of the convection zone. Therefore, one may expect a decrease of the differential rotation in the radiative zone. In this way, investigating seismic data versus the radius of the mode turning points helps us to make qualitative inferences without solving the inverse problem.

Results of helioseismic measurements

The rotation rate inferred by inversion as a function of radius at the equator and at 30° and 60° latitude is shown in figure 5. It shows that the latitudinal differential rotation extends through the convection zone and is sharply reduced in the radiative zone.

The layer of transition from the differential rotation in the convection zone to almost uniform rotation in the radiative zone is characterized by a strong radial gradient of the angular velocity at low and high latitudes. This is probably the place where the global magnetic field of the Sun is generated in the course of the 11 yr cycle. The part of the transition layer which is located in the transition layer is called the 'tachocline'. Because the internal rotation matches the Carrington rotation only in low-latitude zones, it is likely the magnetic field which constitutes sunspots is generated at low latitudes. Most of the transition layer is located beneath the lower boundary of adiabatic convection (shown in figure 5 by the vertical dashed line) which is located at 0.713 solar radii. The estimates of the width of the transition layer range from 0.05 to 0.1 solar radii.

Another region with a strong radial gradient of the angular velocity is located near the upper boundary of the convection zone above 0.95 solar radii.

Contour lines of the rotation rate in a cross-section of the interior are shown in figure 6. The lines of constant angular velocity are quite complicated and follow neither radial nor vertical (cylindrical) lines as predicted by models of the differential rotation.

Rotation in the central and near-polar regions (shaded areas) has not been determined reliably. However, the current helioseismic data are consistent with the assumption that the central core rotates with the same rate as the outer part of the radiative zone, which is approximately 440 nHz (corresponding period 26.3 days).

The data also show that the rotation rate in the near-polar region of the upper convection zone is substantially (about 5%) lower than the rate estimated from the three-term formula (equation (1) or (2)) (figure 7). There is an indication that the polar rotation varies with the solar cycle.

The results in figure 7 also show variations of the rotation rate of ≈0.5% (2 nHz), or variations in the rotational velocity of 10 m s^{-1}, at low and mid-latitudes. These variations correspond to the migrating zonal flows ('torsional oscillations') associated with the solar activity. These flows extend to, at least, 5% of the solar radius (35 Mm) in depth.

North–south asymmetry

The observation of rotation of sunspots reveals the asymmetry in the differential rotation between the northern and southern hemispheres. The asymmetry correlates with the magnetic activity—the hemisphere with lower activity (fewer active regions) rotates faster than that with higher activity. The difference in the average rotation rate may reach 0.5% (10 m s^{-1}).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

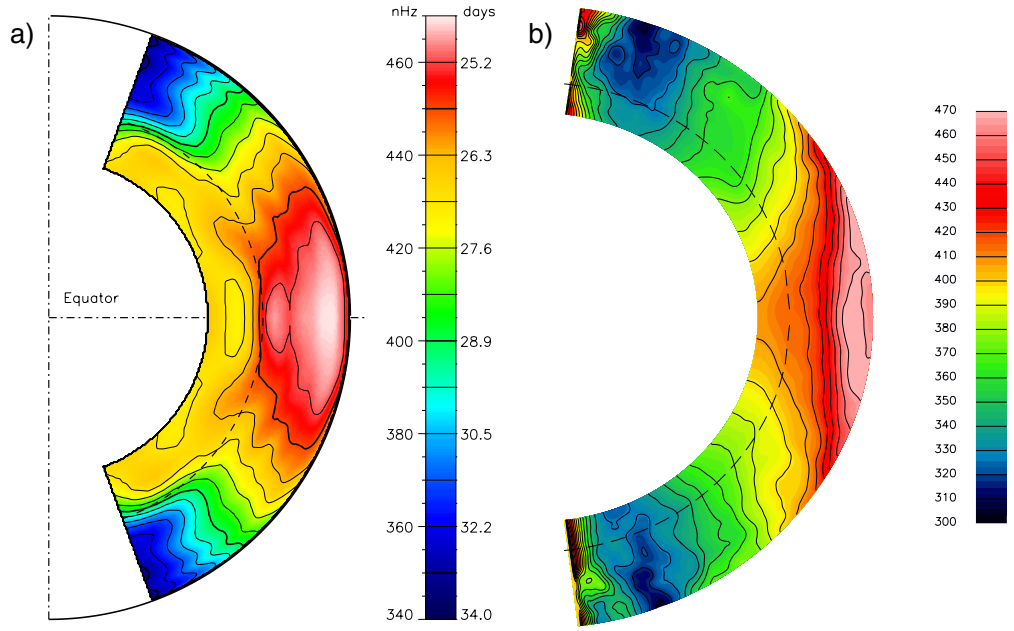


Figure 6. (*a*) Angular velocity profile $\Omega(r, \theta)$ as deduced from helioseismology, using the GONG data. (*b*) Time-averaged angular velocity profile obtained by a 3D simulation. The dashed lines indicate the lower boundary of the convection zone (Elliott *et al* 1998). **This figure is reproduced as Color Plate 55.**

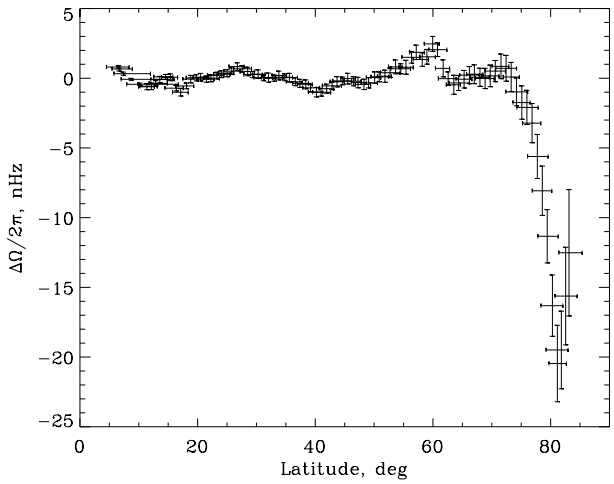


Figure 7. The deviation of the solar rotation rate from the three-term law (equation (1)) in the upper 4% of the convection zone. The horizontal bars indicated the half-width at half-maximum to the left and right of the peak of the localized averaging kernels, $A(r, r_0)$. The vertical bars indicate the standard error estimate (Birch and Kosovichev 1998).

The rotational frequency splitting is not sensitive to the asymmetrical component of rotation to first order. However, the north–south asymmetry of the internal rotation can be studied by methods of local helioseismology, e.g. ring-diagram and time–distance techniques (SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY).

Theories of solar rotation

Evolution of solar rotation

The Sun rotates more slowly, by a factor of 50, than early main sequence stars. This suggests that the Sun initially rotated much faster and lost most of its angular momentum over its lifetime. The angular momentum of the Sun is believed to be lost as a result of 'magnetic braking' through the magnetized solar wind.

The magnetic braking is mainly applied to the surface layers of the Sun. However, it also slows down the interior because the deep layers of the Sun are coupled with the surface because of convective flows and internal magnetic field. In fact, the convection zone reacts to the external torque almost like a solid body because the angular momentum transport provided by turbulent friction is very efficient.

In the radiative zone the angular momentum transport is much less efficient. It primarily depends on the diffusion transport coefficient. This coefficient may be increased by various instabilities due to strong shearing motions or excitation of internal gravity waves in the convective overshoot layer or by instability arising from the accumulation of the ^{3}He at the outer boundary of the energy-generating core. The internal magnetic field provides probably the most efficient mechanism of angular momentum transport. Even a very weak magnetic field ($\simeq 10^{-6}$ G) is sufficient to slow down the solar core.

The theories of the evolution of solar rotation based on these mechanisms predicted that the central core might still rotate much faster than the surface. However, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

recent helioseismic data have provided strong evidence that the whole radiative zone rotates almost uniformly. Which of the mechanisms of angular momentum transport provide most of the braking is not yet understood.

Mechanisms of the latitudinal differential rotation

The angular momentum transport which causes the differential rotation of the convection zone results from thermal stresses and the anisotropic Reynolds stresses, $Q_{ij} = \langle u_i u_j \rangle$, where u_i and u_j are components of the convective velocity and the brackets mean an average over longitude ϕ. The anisotropy arises from the distortion of convection motions by the Coriolis force and other effects of rotation such as the meridional circulation.

The evolution of the angular momentum in the convection zone is described by non-linear Navier–Stokes equations. The coupling between the turbulent solar convection and rotation can be modeled in the 'mean-field' approximation by averaging over an ensemble of convective elements or by solving the equations numerically on a high-resolution spatial grid. Both approaches are still far from the real situation. The mean-field theory employs various approximations and assumptions for the mean anisotropic properties of convection, whereas the numerical models do not provide sufficient spatial resolution to model the fully developed turbulent convection accurately.

The basic mechanisms of the differential rotation can be understood from the conservation of the angular momentum in the convection zone (for more details see Stix 1989). Consider the velocity field, $\boldsymbol{v}$, in the convection zone as a superposition of the mean rotational velocity, $\langle v_\phi \rangle$, a meridional circulation, $\boldsymbol{v}_m = (\langle v_r \rangle, \langle v_\theta \rangle)$, and the convective velocity, $\boldsymbol{u}$:

$$\boldsymbol{v} = \langle \boldsymbol{v} \rangle + \boldsymbol{u}$$

From the azimuthal component of the equation of conservation of momentum,

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\frac{1}{\rho} \nabla P - \nabla \Phi$$

where ρ is the density, P is the pressure and Φ is the gravitational potential, and the conservation of mass,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0$$

we obtain the equation of the conservation of the angular momentum:

$$\frac{\partial}{\partial t} \left(\rho s^2 \Omega \right) + \nabla \cdot (\rho s^2 \Omega \boldsymbol{v}_m + \rho s \langle u_\phi \boldsymbol{u} \rangle) = 0 \qquad (4)$$

where $s = r \sin\theta$, and $\Omega = \langle v_\phi \rangle / s$. (Fluctuations of density which may also be important are not taken into account in equation (4).) This equation shows that the density of angular momentum, $\rho s^2 \Omega$, changes either because angular momentum is transported by the meridional circulation, $\boldsymbol{v}_m$, or because of the Reynolds stresses, $\langle u_\phi \boldsymbol{u} \rangle$. These two processes play important roles in the mechanism of the differential rotation, and, therefore, it is very important to study them. The meridional circulation and convective flows are studied by measuring the Doppler velocity of the solar plasma on the surface and inside the convection zone by time–distance helioseismology.

The numerical simulations have been able to reproduce the differential rotation qualitatively (figure 6(*b*)). However, there are significant differences. The simulations provide an angular velocity profile with the contour lines parallel to the rotation axis in the plot, meaning that the angular velocity is constant on cylindrical shells, whereas the observations reveal a more complicated profile. The 'cylindrical' distribution of the angular momentum is attributed to large-scale convective cells oriented north–south, so-called 'banana cells'. This type of convection results from the effect of the Coriolis force. Evidently, some other important physical processes are missing in the simulations. Perhaps the small-scale turbulent convection which is not resolved in the current simulations transports angular momentum quite differently, yielding other classes of mean flows and rotation profile.

The future efforts in our understanding of solar rotation will be focused on the precise determination of the rotation rate of the solar core, tachocline, near-polar regions and the upper convective boundary layer. It is very important to study variations of the solar rotation, both temporal and spatial, which are associated with the solar cycle, and, in particular, the variations in the tachocline and zonal flows ('torsional oscillations') because these studies will help us to understand the mechanism of solar activity.

Bibliography

Birch A C and Kosovichev A G 1998 Latitudinal variation of solar subsurface rotation inferred from p-mode frequency splittings measured with SOI–MDI and GONG *Astrophys. J.* **503** L187–90

Elliott J, Miesch M S, Toomre J, Clune T and Glatzmaier G A 1998 Turbulent convection and its coupling with rotation *Structure and Dynamics of the Interior of the Sun and Sun-like Stars* ed S Korzennik and A Wilson (Noordwijk: European Space Agency) pp 765–70

Schou J *et al* 1998 Helioseismic studies of differential rotation in the solar envelope by the solar oscillations investigation using the Michelson Doppler Imager *Astrophys. J.* **505** 390–417

Snodgrass H 1992 Synoptic observations of large scale velocity patterns on the Sun *The Solar Cycle (ASP Conf. Ser. vol 27)* ed K L Harvey (San Francisco: Astronomical Society of the Pacific) pp 205–40

Stix M 1989 *The Sun: an Introduction* (Berlin: Springer)

Ulrich R K 1998 Identification of very large scale velocity structures on the solar surface using Mt Wilson synoptic observations *Structure and Dynamics of the Interior of the Sun and Sun-like Stars* ed S Korzennik

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and A Wilson (Noordwijk: European Space Agency)
pp 851–6

Alexander Kosovichev

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Solar Diameter, Oblateness and Temperature

The visible edge of the Sun is called the LIMB and is almost circular, but very small deviations from a spherical Sun are measurable. The largest deviation is a 'flattening' of the polar radius in comparison to the equatorial solar radius. The Sun's surface temperature is also not spherically symmetric but has a tiny asphericity in the form of a latitudinal temperature variation. The radius, shape and surface temperature may also change on SOLAR CYCLE (yearly) time scales and the measurement of these properties yields important clues about the Sun's interior, and the physical origins of the 11-year solar cycle.

Solar shape and temperature measurements

The limb of the Sun is not a sharply defined edge but declines in brightness smoothly as we measure from the solar disk to off of the limb. This limb brightness profile changes most rapidly within about an arcsecond of the edge, or visible photosphere of the Sun. Here the effects of the Earth's atmospheric blurring, and magnetic features in the Sun's photosphere and the higher solar atmosphere (like SUNSPOTS or FACULAE), tend to cause the limb profile to change with time and limb position. Real changes in the brightness or temperature of the photosphere will also affect the apparent limb profile. These effects make it difficult to measure the true angular position of the limb of the Sun against the sky, since variations in the form of the limb profile can be confused with limb position changes (see SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION). In general, solar shape and temperature variations are difficult to disentangle and accurate shape measurements must often depend on a simultaneous study of the Sun's brightness or temperature.

Measurements of the shape of the Sun, and in particular the solar oblateness (which can be expressed as the relative difference between the equatorial and polar diameters), have been published and debated for nearly a century. Interest in the solar oblateness over the last 35 years was rekindled by measurements from Princeton which suggested that general relativity may be seriously 'incomplete'. Subsequent data did not confirm these results, although the ensuing experimental activity has generated a better understanding of the SOLAR INTERIOR and atmosphere. Figure 1 shows oblateness measurements since the mid-1960s. The most recent observations have the smallest error bars and were obtained from space, above the blurring effect of the atmosphere. The dotted line, which the measurements seem to converge toward, describes the approximate 'flattening' or oblateness expected due to the nominal solar surface rotation rate. Although the rotational flattening of the Sun amounts to an equatorial–polar diameter difference of only about 12 km, it seems to be reliably observed in the modern experiments and if it does vary with time, the temporal changes are very small

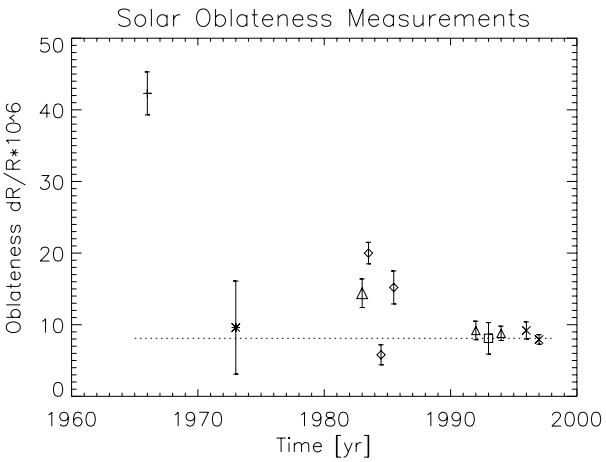


Figure 1. Recent measurements of the solar oblateness (10^6 times the fractional difference between equatorial and polar solar radii) are plotted versus time. The dotted line shows the expected solar flattening due to the nominal surface rotation rate.

(corresponding to less than about 1 km in the radius difference).

More complicated departures from a spherical limb shape have also been observed. Beyond the oblateness the next largest equator-symmetric departure from circularity is called the hexadecapole shape term and resembles a four-leaf clover pattern. The solar hexadecapole shape implies a solar diameter which is about 2 km smaller at mid-latitudes than near the equator or poles.

Limb shape measurements also reveal that the Sun's surface temperature varies with latitude and with the 11-year solar cycle by as much as a few kelvin. The poles and equator are always hotter than solar mid-latitudes by 1–2 K while the low-latitude regions, where sunspots and faculae appear on the surface, vary in brightness temperature throughout the 11-year solar cycle, but are usually a few degrees hotter than the mid-latitude photosphere. It is possible that the variability of the Sun's surface temperature has been mistaken as a variable solar diameter and oblateness in some solar size and shape measurements. This may help to explain the large variations in the early data points in figure 1.

Solar diameter measurements

Archival records of solar diameter measurements extend back to the mid-seventeenth century. By using the Earth's rotation to transit a solar image created by a long focal length telescope objective, the solar diameter could be deduced subject to the accuracy and repeatability of an observer's definition of the solar limb, and instrumental, timing and atmospheric distortion errors. Other historical diameter measurements have been inferred by timing solar eclipses, and by timing transits of Mercury across the solar disk. Although the many observations are not entirely consistent, there is no evidence that the Sun's diameter has changed by more than 2″ over the last

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

350 years. Modern optical measurements yield a solar diameter of 1919.36 ± 0.02 and imply that solar cycle (yearly) variations in the diameter are less than a few tenths of an arcsecond. The intercomparison of different diameter determinations is difficult because the solar limb profile is not sharply defined. The outer part of the Sun varies continuously, so that over a height difference of about 400 km the solar gas density changes by a factor of 10—there is no unique definition of the Sun's limb. A 1 arcsecond change in the solar diameter corresponds to a solar radius change of 370 km.

Measurements of solar acoustic wave modes allow the possibility of detecting very small changes in solar radius. We can expect to observe changes in helioseismic mode frequencies caused by a changing solar diameter, just as the pitch of an organ pipe shifts when the length of the pipe changes. So far, helioseismic acoustic measurements yield no evidence of solar diameter changes larger than about 10 km over 1 year interval (see HELIOSEISMIC OBSERVATIONS).

Mechanisms for solar shape and size changes

The fluid outer regions of the Sun respond to gravitational, centrifugal and gas and turbulent pressure forces, so that the shape of the Sun yields a direct measure of the balance of these forces at the visible photosphere. For example, the solar oblateness can be directly related to the non-spherical solar quadrupole gravitational potential and the Sun's internal and surface rotation rates. The observed oblateness is entirely consistent with a Sun which rotates at approximately the observed surface rate, and which does not have a rapidly spinning inner core. The solar hexadecapole shape probably results from the variation in rotation rate with solar latitude.

We expect the Sun's diameter to change very slowly over millions of years because of variations in the internal solar structure caused by the Sun's nuclear evolution. While these diameter changes are too small for us to measure there have been inconsistent observational hints that the solar radius could be changing, perhaps with the solar cycle. Such changes, should they be detected, would be quite interesting because even a small change in the solar radius requires an enormous energy. For example, the energy required to lift just the outer convection zone of the Sun by 1 arcsecond (or 730 km) is equivalent to the total luminous energy of the Sun summed over 100 years. A physical mechanism that could cause such changes in diameter over a few years would also have profound consequences for our understanding of the internal solar structure.

Bibliography

Kuhn J R *et al* 1998 The Sun's shape and brightness *Nature* **392** 155–7

Ribes E *et al* 1991 The variability of the solar diameter *The Sun in Time* ed C P Sonett, M S Giampapa and M S Matthews (Tucson, AZ: University of Arizona Press) pp 59–97

Toulmonde M 1997 The diameter of the Sun over the past three centuries *Astron. Astrophys.* **325** 1174–8

Jeffrey R Kuhn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Interior: Standard Models

Due to its proximity, there are many properties of the Sun which we can observe. Consequently, it plays a key role in our understanding of the physics of the internal structure of stars. Standard models of the Sun are theoretical representations of its present structure. These models include the best physical description of the fundamental processes which contribute to solar evolution and allow theoretical predictions of all the solar observables. Testing against observations brings about theoretical improvements which are regularly implemented in standard models and new ideas for laboratory experiments (e.g. nuclear reaction rates, opacity measurements) or solar investigation (e.g. detailed photospheric abundances, acoustic modes, gravity modes, neutrino fluxes).

From the age of Earth to stellar evolution

If gravity were the main actor in the structure of the stars, one could easily calculate how long it would take for the Sun's own gravity to cause a significant change in its radius. The corresponding time scale is the free-fall time, $t_{\rm ff}$, which is a function of the mass M of the star, its radius R and the gravitational constant G:

$$t_{\rm ff} \propto \left(\frac{R^3}{GM}\right)^{1/2}. \quad (1)$$

This time is of the order of half an hour for the Sun, and is incompatible with its present extreme stability (see SUN: BASIC PROPERTIES). Another way of determining the age of the Sun is to correlate the rate of stellar evolution with the available energy. The gravitational potential energy has the following dimensional form: $\Omega \propto -M^2G/R$. For the Sun, $\Omega = 4 \times 10^{48}$ erg and the present luminosity is 4×10^{33} erg s^{-1}. The time scale over which the Sun could maintain more or less the present luminosity is called the Kelvin–Helmholtz time:

$$t_{\rm K} \propto \frac{\Omega}{L} \approx 30 \text{ Myr}. \quad (2)$$

This would give an order of magnitude for the solar age if one assumed a constant luminosity and no other energy losses. Until early in this century, people believed that the solar energy came from gravitational contraction, and that the Sun, and presumably the Earth, were a few million years old. However, the time scale required for biological evolution is considerably greater than $t_{\rm K}$, and this fact was at the origin of public disagreements between Darwin and Lord Kelvin. At about the same time, the geologist Sir Charles Lyell established that the ocean floor should be at least a billion years old in order to be compatible with features of sediments.

Later, Rutherford developed the radioactive dating of rocks, and it became evident that the Earth had to be several billion years old. However, the Sun should be at least as old as the Earth!

The thermonuclear source of energy stabilizes the Sun

Considering this contradiction, Eddington finally suggested in 1920 that perhaps a subatomic source of energy existed in stars and he introduced another concept, the Einstein time, which corresponds to the transformation into radiation of the total mass:

$$t_E = \frac{Mc^2}{L} = 14 \times 10^{12} \text{ yr}. \quad (3)$$

It then became evident that thermonuclear fusion had to play the fundamental role.

In fact, anticipating the result that the duration of hydrogen burning is about 10 billion years and that fusion occurs only in the inner part of the Sun (30–40% in mass fraction), one can deduce from the Einstein time that the efficiency of the conversion of mass to energy in the Sun is of the order of some per cent.

The present quiet Sun and most of the stars cannot be understood without their having an internal energy source that compensates for the energy lost at the surface. During stellar formation, nevertheless, gravity really is the main actor and its effect leads to an increase in the central temperature until nuclear fusion can deliver the energy required for a stationary equilibrium. After the first stage of deuterium burning, the Sun is limited in its contraction by the beginning of hydrogen burning; the beginning of this new phase is often called 'arrival on the main sequence' in stellar evolution.

Hydrogen, the most abundant element in stars, is also the first one to be converted because of its lower Coulomb barrier. Gamow, von Weizsacker and Bethe have shown clearly that two different sets of reactions, the p–p chain and the CNO cycles, could provide the energy needed to understand the present solar luminosity (see SOLAR INTERIOR: ENERGY GENERATION). The reactions involved in the conversion of hydrogen to helium depend on the temperature and density of the stellar interior. In fact, for the present solar central temperature, the p–p chain is the most energetic (more than 98% of the total luminosity), but when temperature increases, the CNO cycles quickly become the most efficient. One of the main consequences is that low-mass stars ($<1.5M\odot$), governed by p–p reactions, have a slower evolution than the others, governed by CNO cycles.

The stellar structure: the basic equations

Today, the origin of the Sun's stability is clearly established and it seems reasonable to treat the evolution of a star as a succession of static equilibrium models. Of course, the rapid contraction of the protostar which ignites the nuclear reactions must be examined with more caution. The main hypotheses are the following: the star is spherical, without the effect of rotation or a magnetic field. The internal structure of the star is described by four structure equations. The first assumes hydrostatic equilibrium (each gas shell is balanced by the competition between the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

downward gravitational force and the outward pressure force):

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{M(r)G}{r^2}\rho \tag{4}$$

where P and ρ are the pressure and the density. $M(r)$ represents the mass enclosed within a sphere of radius r:

$$\frac{\mathrm{d}M}{\mathrm{d}r} = 4\pi r^2 \rho. \tag{5}$$

One further assumes thermal equilibrium. The energy ($4\pi r^2 \rho\epsilon$) produced by nuclear reactions balances the energy flux $L(r)$ emerging from the sphere of radius r. To this energy flux must be added the energy loss by neutrinos (which is small in hydrogen burning but not in a supernova explosion). Taking into account quasistatic gravitational readjustment and composition variation, one must incorporate a heat transfer term $T\mathrm{d}S$ where S is the total entropy per gram of the gas and the energy bookkeeping yields:

$$\frac{\mathrm{d}L}{\mathrm{d}r} = 4\pi r^2 \rho \left(\epsilon_{\mathrm{nucl}} - \frac{T\mathrm{d}S}{\mathrm{d}t}\right). \tag{6}$$

At the end of hydrogen burning, the nuclear energy vanishes and again contraction produces the energy for the luminosity.

Finally, the temperature gradient depends on the luminosity and the physical process of energy transport. In a radiative region of a star, the diffusion approximation is appropriate, and the relation between temperature gradient and luminosity is

$$\frac{\mathrm{d}T}{\mathrm{d}r} = \frac{-3}{4ac}\frac{\kappa\rho}{T^3}\frac{L(r)}{4\pi r^2}. \tag{7a}$$

When the opacity coefficient κ increases too much, as in the external part of smaller stars with $M \leq 1.5M_\odot$ or when the luminosity is very high (in the internal part of stars with mass $>1.5M_\odot$), the radiative gradient increases so much that matter becomes convectively unstable. The resulting temperature gradient is then nearly adiabatic:

$$\frac{\mathrm{d}T}{\mathrm{d}r} = \left(\frac{\mathrm{d}T}{\mathrm{d}r}\right)_{\mathrm{ad}} = \frac{\Gamma_2 - 1}{\Gamma_2}\frac{T}{P}\frac{\mathrm{d}P}{\mathrm{d}r}. \tag{7b}$$

The adiabatic exponent Γ_2 is defined by

$$P^{1-\Gamma_2} T^{\Gamma_2} = \text{const.}$$

Boundary conditions (typically for pressure and temperature at the surface and for mass and luminosity at the center of the star) are added to complete the problem of stellar structure and evolution.

The physical quantities

The initial composition

The first ingredient of the calculation is the initial detailed composition of the Sun deduced from the observation of the present photospheric abundances. The determination of this composition is described in the article SOLAR ABUNDANCES. As it is now clear that there is some gravitational settling of elements during the lifetime of the Sun from the surface down to the center (see below), this composition is chosen to get the detailed photospheric composition at the present solar age and the whole calculation is iterated to get the observed present Z/X ratio at the surface (mass fraction of the elements greater than helium over hydrogen). Helium is not measured in the photosphere but helioseismology gives serious constraints on this abundance.

Recent calculations consider the 21 most important elements from hydrogen up to nickel. All elements up to ^{16}O contribute to calculation of the nuclear energy of the present Sun. The 21 elements are useful for precise opacity calculations.

The nuclear reaction rates

The nuclear reaction network is presented in the article on SOLAR INTERIOR: ENERGY GENERATION. That article includes the complete proton–proton chain and CNO cycles.

One can consider three types of species:

- Those which are only slowly produced or destroyed during stellar evolution (e.g. H, ^{4}He, ^{12}C). The evolution of the abundances is computed using a linearization method, semi-implicit or implicit.
- The very rapidly reacting minor species are assumed to be in local equilibrium and calculated separately (D, ^{7}Li, ^{7}Be).
- ^{3}He, an intermediate case, is treated analytically, assuming that deuterium has reached equilibrium and that the hydrogen abundance remains constant during a small nuclear time step. The assumption of ^{3}He in equilibrium is completely justified in the core of the Sun and becomes a rough approximation at $0.20R_\odot$. The treatment of ^{3}He is crucial for energy generation because the production of this element and its destruction by the reaction ^{3}He+^{3}He contributes to 80% of the total luminosity.

All nuclear reactions involving nuclei have been measured in the laboratory, except the p–p and pep reactions which are inaccessible due to the weak interaction. The most important cross sections have been generally measured in a higher energy range than needed for the Sun, therefore a theoretical extrapolation is often necessary. This effort is pursued nowadays at low energy in the laboratory to approach solar plasma conditions (see Adelberger *et al* 1998). Specific attention is paid to the Coulomb enhancement effect of the interaction, mainly due the presence of the negative cloud of the free electrons.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The equation of state

The relationship between pressure, temperature, density and composition called the *equation of state* must be known for very different conditions of matter (plasma, neutral atoms, molecules). The total pressure P is the sum of the gas pressure and radiation pressure.

The radiation pressure $P_R = \frac{1}{3}aT^4$, where a is the radiation density constant. is negligible in the solar center (1/1000 of the gas pressure), but not near the photosphere ($0.6P_G$). When the temperature is high and the density low enough, interactions between the particles are negligible and the gas can be reasonably well considered as perfect:

$$P_G = \frac{N}{V}kT = \frac{\rho kT}{\mu(r)} \tag{8}$$

where V is the volume of the gas. This approximation is relatively correct in the central part of the Sun for ions (see below for the treatment for electrons). But the large range of solar temperature (5×10^3 to 10^7 K) and density (10^{-12} to 10^2 g cm^{-3}) requires a more detailed description of the equation of state. As the Sun is made up of a mixture of different species, one introduces the notion of mean molecular weight μ, defined by

$$\mu^{-1} = \sum_i n_i(m_H/m_i). \tag{9}$$

In this expression, n_i is the number of free particles: Z electrons + 1 nucleus for each atom of atomic number Z. More generally, this number depends on the degree of ionization of the considered species. In astrophysics, X, Y and Z represent the mass fraction number (normalized to 1) of hydrogen, helium and other elements (heavier than helium) respectively. In the central conditions of the Sun, all the species (except iron and nickel) are completely ionized. The solar value μ decreases from the central region where helium is produced down to the nuclear edge. In the intermediate region of the Sun, the heavy elements become progressively partially ionized but this has little impact on μ which increases outward due to the partially ionized elements and the appearance of molecules.

Even though the correct treatment of the heavy elements affects the total pressure very little due to their small fraction number, this knowledge is important for opacity calculations (see below). The appropriate treatment of the degree of ionization of hydrogen and helium, together with a correct treatment of atomic and molecular hydrogen, is crucial to determine the pressure below 20 000 K.

The central solar density is sufficiently high to cause partial degeneracy of the electrons, so electron number and electron pressure must be expressed as functions of the degeneracy parameter η:

$$n_e = \frac{4\pi}{h^3}(2mkT)^{3/2}F_{1/2}(-\eta) \tag{10}$$

$$P_e = \frac{8\pi kT}{3h^3}(2mkT)^{3/2}F_{3/2}(-\eta)$$

where $F_{1/2}$ and $F_{3/2}$ are the Fermi–Dirac functions. In the case of the Sun, the effect of degeneracy is small (η is of the order of -1) but when stars continue to evolve, this effect becomes dominant, leading to an increase in the pressure and a quasi independence of the pressure on the temperature.

Relativistic and Coulomb effects must be taken into account to correct the gas pressure for electrostatic attraction. With the development of helioseismology, a sophisticated equation of state including all these processes and a large number of species may be checked in great detail, transforming the Sun into a real laboratory (cf SOLAR INTERIOR: EQUATION OF STATE AND OPACITY).

Energy transport

Two processes dominate the transport of energy in a solar-type star: radiation and convection. Conductivity plays no role in the Sun and acts only in higher-density stars (for example in a supernova core). In the same way, the energy transported by neutrinos is very small. In the central part of the Sun, radiation is the most efficient process, contrary to the case in more massive stars where energy is transported by convection due to the extremely high temperature produced by the CNO cycle.

Radiative transport

In the Sun's interior, the average energy per photon is in the keV range, which is characteristic of x-rays, whereas the light escaping from the surface is in the visible spectrum, corresponding to an energy 10^5 times smaller. The cause of this degradation in photon energy is the coupling between radiation and matter. The effect of the interaction of photons with electrons, atoms, ions and molecules is evaluated and is summarized in what one calls the 'Rosseland mean opacity' κ, which is a function of temperature, density and composition. The equation of radiative transfer adapted to stellar interiors is given by equation (7*a*). The opacity of the gas really governs the rate at which energy escapes, and plays a crucial role in determining the age of a star and the evolved central temperature. The sources of opacity are the following:

- electron scattering (Thompson and Compton scattering): scattering of the photon on free electrons;
- bound–bound transitions: the photon induces a modification of level at a bound electron;
- bound–free (photoionization) transitions: a bound electron becomes free after interaction with the photon or the reverse situation;
- free–free transitions (inverse bremsstrahlung): the electron stays free after the interaction with the photon;
- molecular transitions.

Except for the first one, the computation of all these processes necessitates detailed calculations of atomic physics. Frequency-dependent opacities are needed for detailed stellar models: $\kappa_{tot}(\nu) = \sum_i \kappa_i(\nu)$, where i

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

represents the different processes. It is important to notice that the mean Rosseland opacity is not obtained by averaging the frequency-dependent opacity, but by averaging its reciprocal, that is the radiative conductivity (the weight function $f(\nu)$ is the temperature derivative of the Planck function):

$$\frac{1}{\bar{\kappa}} = \sum_i \left(\int \frac{f(\nu)}{\kappa_i(\nu)} \mathrm{d}\nu \right). \tag{11}$$

The mean Rosseland value clearly favours peaks around $4\text{–}6h\nu/kT$ where the Rosseland weighting is maximum. This adds to the already nonlinear character of the processes and gives a supplementary reason to perform computations with an extremely refined grid well adapted to the problem. Moreover, due to the non-additive character of the opacity coefficients, the spectrum must be calculated for the correct mixture before determining the Rosseland mean value. This leads to extensive calculations which must take into account the detailed abundance of the 21 elements considered.

The main reason for the difficulty of such calculations is that one single source of opacity may exceed all the others. For example the bound–bound processes may introduce cross sections which are one or two orders of magnitude greater than the competing contributions.

So, even a trace element may produce a non-negligible contribution to the mean opacity. In the case of the Sun or solar-like stars, knowledge of radiative processes in detail is crucial. Indeed, in low-mass stars, the contribution of heavy elements (about 2% in mass fraction and 0.14% in number fraction) to the equation of state does not exceed 1 or 2%, but their contribution to the opacity is about 40% in the central part of the Sun, and 80% in the intermediate region. When oxygen (the most abundant heavy element) becomes partially recombined, the temperature gradient largely increases. Then nitrogen and carbon have the same behavior and the increase of κ is so important that the radiative temperature gradient exceeds the adiabatic one:

$$\left| \frac{\mathrm{d}T}{\mathrm{d}r} \right|_{\mathrm{rad}} > \left| \frac{\mathrm{d}T}{\mathrm{d}r} \right|_{\mathrm{ad}}. \tag{12}$$

This is the criterion for the onset of convective instability, called the Schwarzschild criterion. In the Sun, there is a large convective solar region below the photosphere which corresponds to such conditions, that oxygen then nitrogen, carbon and finally helium and hydrogen have important bound–bound contributions (figure 1). But at about $0.995R_\odot$, hydrogen and helium become neutral, the opacity decreases, radiative flux comes in competition with convective flux and a superadiabatic regime must be considered.

Convective transport

In any case, the energy must escape from the star. When the opacity increases too much, one would require a very steep temperature gradient to maintain the energy flow,

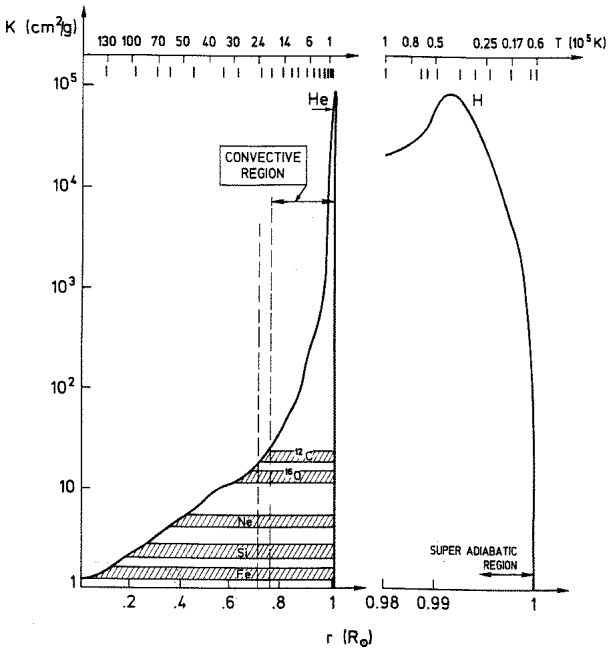


Figure 1. Present solar radial dependence of the total opacity κ ($\mathrm{cm}^2\ \mathrm{g}^{-1}$). The Sun has a large radiation region from the center down to $0.71R_\odot$. κ increases from the center to the photosphere with bound–bound processes. In this figure we show the radial position where the elements begin to be partly ionized. One notices that the base of the convective zone (broken lines) appears when oxygen (the third most abundant element) begins to have bound–bound contributors. A zoom of the external layers shows the role of partial ionization of helium then hydrogen; finally the formation of molecules induces a reduction of the opacity coefficients and new competition between convection and radiation in the superadiabatic region. The outermost layers are totally radiative.

leading to an unstable situation. When the instability criterion is fulfilled, the real temperature gradient exceeds the adiabatic gradient and the difference is called the superadiabatic gradient:

$$\frac{\Delta \mathrm{d}T}{\mathrm{d}r} \equiv \frac{\mathrm{d}T}{\mathrm{d}r} - \left(\frac{\mathrm{d}T}{\mathrm{d}r} \right)_{\mathrm{ad}} \tag{13}$$

Convection is a very efficient means of energy transport. In fact, near the base of the convective envelope, the radiative flux is so small that we can ignore the superadiabatic gradient and use the adiabatic one as the temperature gradient (this is also the case in the central part of massive stars). Unfortunately, this is not the case in the outer layers of the Sun. Evaluating this difference would require a theory of convection which is yet lacking; the simulation of such effects for stellar conditions is extremely complex (see SOLAR INTERIOR: CONVECTION THEORY), the size of the cells and the associated turbulence is not well established; so one usually restricts the calculation to an order of magnitude calculation in the framework of the 'mixing length theory'. This approximation conceptually

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

replaces the real situation of 'plumes', convective eddies of different sizes, by an average situation where each convective element travels, on average, a distance Λ, before mixing with the surrounding matter. This distance is generally scaled to the pressure scale height[1] by a parameter α called *the mixing length parameter*:

$$\Lambda = \alpha\lambda \equiv \alpha \left(\frac{\mathrm{d} \ln P}{\mathrm{d}r} \right)^{-1} . \qquad (14)$$

In this calculation, one must estimate the relative amount of energy carried by convection and by radiation. Two different cases must be studied: the convection in the central part of the star and in the outer layers. These two regions differ by their density values, respectively about 100 g cm^{-3} and 10^{-6} g cm^{-3}, and the typical lengths of Λ, 700 000 km and 200 km, so the difference between the temperature of a convective element and the average temperature of the surrounding matter is really very different in the two situations. In the central part of stars, the temperature gradient is about 10^{-4} K cm^{-1} to be compared with the absolute value of 10^7 K and a convective time of about 1 to 100 days. Therefore if such an instability exists (for stars greater than $1.5 M_\odot$), convection is an extremely efficient process of energy transport and the temperature gradient can safely be approximated by the adiabatic one. In the outer layers, the temperature variation between shells is of the order of 100 to 1000 K, to be compared with the gas temperature of 5000 to 10 000 K, the convective time is shorter, 1 to 100 min, and we really need an adapted treatment of the convection. In that case, due to partial ionization of hydrogen or helium, Γ_2 drops to a value near unity and the adiabatic gradient becomes relatively small. This favors convective transfer but the radiative flux is still important and cannot be neglected.

In the solar case, the value of the mixing length parameter is in fact calibrated in solar models by the knowledge of the solar radius at the Sun's present age. In this specific case, therefore, one may consider that this small superadiabatic region just below the photosphere is rather correctly treated. The extension of this solar value to other stars with rather different surface temperatures is more uncertain due to the fact that it is, amongst other things, largely dependent on the opacity in the superadiabatic part of the convection zone. This is why theoretical efforts to model convection, largely encouraged by the development of solar local helioseismology measurements are extremely promising.

The evolution of the Sun with time

The Sun is approximately in the middle of the first, longest (about 10 Gyr) and calm stage of stellar evolution, corresponding to the burning of hydrogen. A theoretical description of the evolution of the Sun with time is obtained by considering a succession of static equilibrium models described by the basic equations (4) to (7). The time evolution of each chemical species X_i is given by three terms:

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \frac{\mathrm{d}X_i}{\mathrm{d}t}_{\text{ nucl}} + \frac{\mathrm{d}X_i}{\mathrm{d}t}_{\text{ diff}} + \frac{\mathrm{d}X_i}{\mathrm{d}t}_{\text{ turb}} . \qquad (15)$$

The first term is the major one for the evolution of a star. It determines the evolved composition due to the differing destruction and creation of elements through nuclear reactions. It allows also us to determine the nuclear energy ϵ_{nucl} of equation (6). Starting from an initial mass fraction of 0.71, the present relative abundance of hydrogen in the central region is 0.34 and the helium has increased in consequence (from 0.27 to 0.63). The main important elements and their radial distributions are shown in figure 2. One notices very different compositions and largely different radial distributions due to nuclear processes. The second term has recently been introduced into the equations to take into account microscopic diffusion. This very slow process of elemental diffusion with a characteristic time-scale of 10^{10} years, is theoretically studied as a partial solution to the observed deficit of some 'fragile' element such as lithium, which is destroyed at about 2.4×10^6 K in typical solar mass conditions.

The introduction of such a process has been largely validated by the capability of helioseismology to determine the photospheric solar helium. The value of 0.249 ± 0.003 in *mass fraction* is very near the cosmological value (0.23–0.24) but rather incompatible with the assumption that the initial solar composition is the present photospheric one. Effectively the initial helium abundance of 0.27, obtained by the standard solar model, shows a reasonable galaxy enrichment between the Big Bang and the formation of the solar system. Consequently this second term is definitively introduced in the equations describing the standard solar model.

The diffusive velocity V_i of an element i, typically some 10^{-8} cm s^{-1}, is the result of several processes with opposing directions, including gravitational settling, radiative and thermal terms. The net result is a slight migration of all elements heavier than hydrogen towards the center with time. At the present time, the photospheric abundances are depleted by about 10% in comparison with their initial value due to these processes.

The last term in equation (15) represents a turbulent term which simulates macroscopic motion inside the Sun. This term is generally ignored in the standard model but begins to be considered at the basis of the convective zone. In this region of transition between radiation and convection, macroscopic motions and mixing could be induced by the differential rotation, which is now observed rather precisely by seismology. Such a process partly inhibits the microscopic diffusion.

Confrontation of standard models with observations

The stellar evolution codes were developed in the 1960s and early 1970s when computers were expanding in power, memory and speed. The codes solve the basic

[1] The pressure scale height is the distance corresponding to a pressure decrease of a factor e: 2.718.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

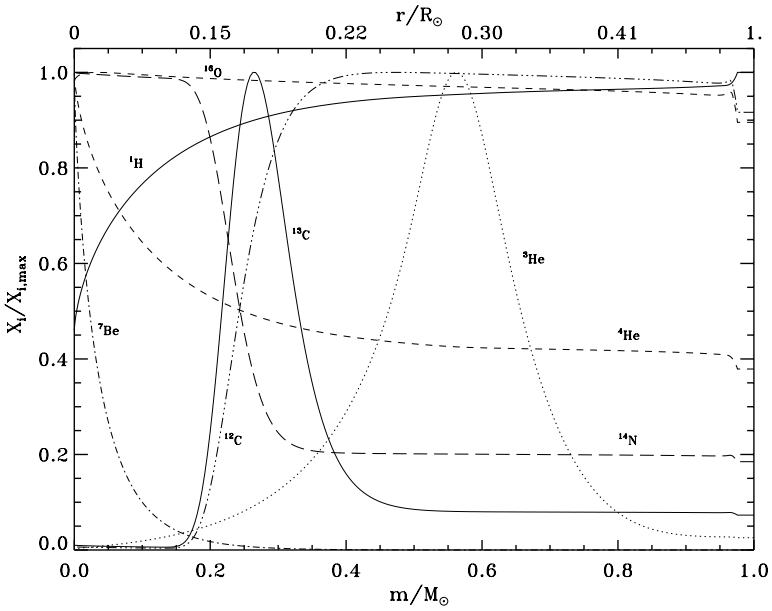


Figure 2. Radial dependence of the abundances of the most crucial elements which have been modified by nuclear reactions. The maximum values are the following: H = 0.7204, ^{3}He = 0.003 09, ^{4}He = 0.6350, ^{7}Be = 1.718×10^{-11}, ^{12}C = 0.0030, ^{13}C = 0.000 44, ^{14}N = 0.0045, ^{16}O = 0.0085.

equations (4) to (7) generally expressed in Lagrangian coordinates following the motion of mass elements. The starting point is usually an isolated non-rotating, non-magnetic spherical mass of gas. In that early period, methods and codes were developed by Iben, Paczyński and Kippenhahn. The most recent ones have introduced numerical methods which guarantee a high accuracy (typically 10^{-7}), useful to avoid any numerical effect in the interpretation of the present solar observables.

At each time step, the spatial stellar structure is described with the help of 300 to 600 mesh points the nonlinear differential equations are solved with the five structural variables, pressure $P = P(\rho, T, X_i)$, opacity coefficient $\kappa = \kappa(\rho, T, X_i)$, energy production $\epsilon = \epsilon(\rho, T, X_i)$, entropy $S = S(\rho, T, X_i)$ and adiabatic exponent $\Gamma_2 = \Gamma_2(\rho, T, X_i)$. The initial model is built in such a way that it simulates the stellar equilibrium corresponding to the moment when the contraction of the star is stopped by the onset of hydrogen burning; it may also be built to include the gravitational contraction of the pre-main sequence (see PRE-MAIN SEQUENCE STARS). One then follows stellar time evolution through a succession of hydrostatic equilibria. The time step is automatically adjusted and determined by the maximum acceptable variation of the principal variables.

The simulation of the present Sun is based on the previous concepts. The peculiarity of the Sun as a star is due to direct comparison with a variety of observed phenomena. The known present luminosity $L_\odot$ and radius $R_\odot$, as well as the age, avoid the two main difficulties of the theory of stellar evolution: the absence of a precise determination of helium abundance and of an *ab initio* theoretical formalism for convection.

Table 1 summarizes the results obtained by standard model calculations, compared with present observations. Starting with a star of $1M_\odot$, just on the arrival of the main sequence or in the pre-main sequence contracting phase, with an initial composition provided by present observation, we solve the structure equations (4) to (7) and (15) and progress in time by successive steps of the equilibrium state, including nuclear and diffusion processes, up to the present time estimated to be 4.52–4.55 Gyr. The computation is done *ab initio*, without any free parameters. The computation of the reference model is iterated with small modifications of the initial helium content Y, the initial heavy element composition, and the parameter α in order to obtain the observed values $L_\odot$, $R_\odot$ and $Z/X_\odot$ at $t_\odot$. The solar structure is more and more under control, from a standard model without diffusion to a standard model including microscopic diffusion and some turbulent mixing at the base of the convective zone, with very good agreement for the photospheric helium and the position of the base of the convective zone between the latest model and the results obtained by inversion of the sound speed and the adiabatic exponent Γ_1 ($c^2 = \Gamma_1 P/\rho$) (see HELIOSEISMOLOGY: THEORY).

Figures 3(*a*) and (*b*) present the corresponding squared sound speed difference between the Sun and these models. The present errors on the solar sound speed vary from about a few parts in 10^5 near the surface to a few parts in 10^4 in the solar core. Such an accuracy allows us to check some reaction rates ((p,p) or ^{3}He,^{4}He) or opacity questions at the level of some 1% or equation of state ones

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Standard solar models. Thermodynamical quantities and neutrino predictions obtained in the computation of the standard model: the present $R_\odot$ and $L_\odot$ are adjusted to $(6.9599 \pm 0.0002) \times 10^{10}$ cm and $(3.846 \pm 0.004) \times 10^{33}$ ergs s^{-1} respectively, α is the mixing length parameter; Y_i, Z_i, $(Z/X)_i$ are initial helium, initial heavy element and initial ratio heavy element to hydrogen; Y_s, Z_S, $(Z/X)_S$ are the same for photospheric compositions; R_{bcz}, T_{bcz} are the radius and temperature at the base of the convective zone; Y_c, Z_c, T_c, ρ_c are central helium, heavy element contents, central temperature and density; ^{37}Cl, ^{71}Ga, ^{8}B respective neutrino predictions for the chlorine, gallium and water detectors.

Parameters	Observations	Without diffusion	With diffusion	With diff. + turb
α		1.71	1.77	1.76
Y_i		0.265	0.272	0.272
Z_i		1.757×10^{-2}	1.898×10^{-2}	1.898×10^{-2}
$(Z/X)_i$		0.0245	0.0277	0.0277
Y_s	0.249 ± 0.003	0.265	0.243	0.251
Z_s		1.757×10^{-2}	1.803×10^{-2}	1.858×10^{-2}
$(Z/X)_s$	$0.0245 \times (1 \pm 0.1)$	0.0245	0.0245	0.0255
$R_{bcz}/R_\odot$	0.713 ± 0.003	0.729	0.713	0.714
$T_{bcz} \times 10^6$ (K)		2.055	2.190	2.194
Y_c		0.614	0.640	0.641
Z_c		1.807×10^{-2}	2.094×10^{-2}	2.094×10^{-2}
$T_c \times 10^6$ (K)		15.44	15.71	15.71
ρ_c (g cm^{-3})		147.80	153.1	153.1
^{37}Cl (SNU)	2.55 ± 0.25 SNU	5.65	7.04 ± 1.8	7.04 ± 1.8
^{71}Ga (SNU)	76 ± 8 SNU (cal $= 0.91 \pm 0.08$)	119.4	127.1 ± 9	127.1 ± 9
^{8}B (10^6 cm^{-2} s^{-1})	2.44 ± 0.26	3.66	$5. \pm 1.25$	$5. \pm 1.25$

at the level of some 0.1%. So, comparisons with precise helioseismological data push astrophysicists to use the best physical developments available in atomic, nuclear and plasma physics, and the results are representative of this knowledge.

Nevertheless, figures 3(*a*) and (*b*) show regions of mismatch where the differences on the squared sound speed between model and Sun is about 10 times the experimental determination of such a quantity. These regions are useful to correct the physics used or to introduce a new phenomenon, as has been done for the two processes explicitly discussed here. The residual difference may be attributed to uncertainties on abundances, opacities or reaction rates.

The results are reproducible, except to some extent in the solar nuclear core where solar results are sensitive to the data used and the model results depend on the details of most of the physical inputs. The core of the Sun is not totally under control, and this justifies complementary efforts to try to detect gravity waves which will help us to better investigate this region of the Sun and to disentangle different physical effects.

The neutrino fluxes obtained by the standard models are not in agreement with the observations. Helioseismology constrains the situation offered by the stellar structural problem even if it does not at present sufficiently constrain the region of interest located below $0.15R_\odot$. Nuclear reaction rates in the secondary channels (as the ppIII chain) or absorption cross sections play a crucial role in these predictions. Experimental and theoretical efforts in these disciplines continue to be useful as they are potentially able to explain some of the discrepancy. In parallel the study of the properties of neutrinos is also required to get a definitive understanding of solar neutrino fluxes on the ground (see SOLAR INTERIOR: NEUTRINOS).

Toward a dynamical representation of the stars

The present theoretical description of the Sun is a static monodimensional representation of our star where rotation and magnetic field are not yet introduced explicitly. The development of local seismology and three-dimensional simulation of convection contribute to progress towards a dynamical representation of the Sun. Nowadays the differential rotation coupled with turbulent convection improve the agreement with observations and encourages three-dimensional representation of the Sun in order to disentangle the internal competitive physical processes.

Some complementary indicators are specific abundances such as photospheric ^{3}He, ^{7}Li and ^{9}Be abundances, which have a certain sensitivity to internal mixing as they are destroyed outside the nuclear core (see figure 2). One may notice that the destruction of lithium occurs partly during the pre-main sequence and the main sequence. The understanding of the observed deficit of present photospheric lithium is largely dependent on the competition of photon-matter absorption, element segregation and turbulent diffusion.

The development of asterosismology (seismic study of solar-like stars) may contribute to refining our representation of the Sun by properly introducing the dynamical processes of the young Sun.

The internal magnetic field is still rather poorly known; a global history of the angular momentum of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

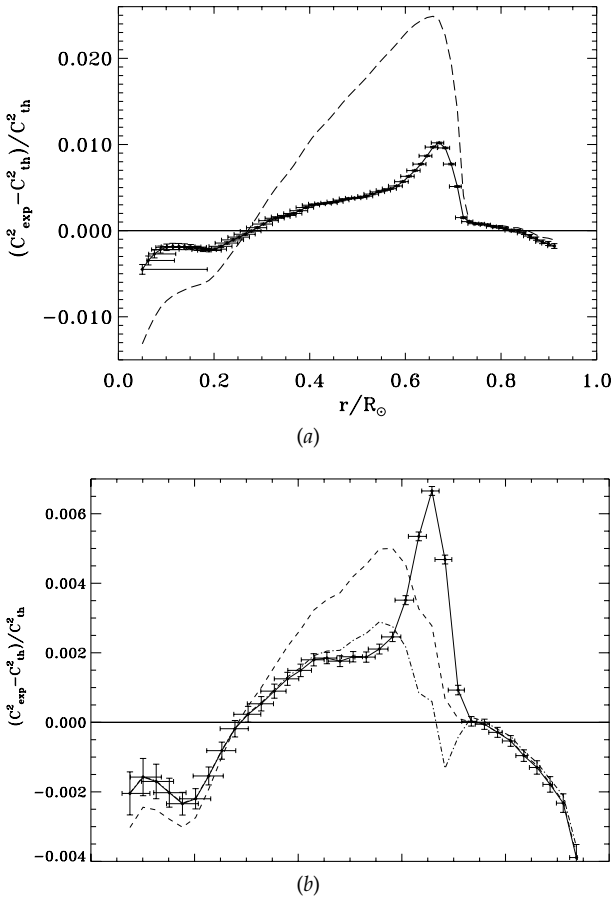


Figure 3. Sound speed squared difference between the Sun measured by GOLF+LOWL experiments and standard models. The seismic uncertainties are reported on the difference itself. No theoretical uncertainty is included. (*a*) corresponds to standard models without (full) (model 1, table 1) and with microscopic diffusion (full line with experimental error bars, model 2, table 1). (*b*) corresponds to the model including microscopic diffusion (model 2, table 1) and to the model where a turbulent term is added at the base of the convective zone (model 3, table 1).

Sun coupled to temporal seismic studies along the solar cycle could help progress in this direction.

The very specific study of the solar interior is crucial for the understanding of general astrophysics problems such as the age of the universe, galactic evolution, the role of convection etc. It also plays a crucial role in understanding the properties of neutrinos.

Bibliography

Adelberger E *et al* 1998 Solar fusion cross sections *Rev. Mod. Phys.* **70** 1265

Brun S, Turck-Chièze S and Morel P 1998 Standard solar models in the light of new helioseismic constraints: I The solar core *Astrophys. J.* **506** 913–25

Brun A S, Turck-Chièze S and Zahn J P 1999 Standard solar models in the light of new helioseismic constraints: II Mixing below the convective zone *Astrophys. J.* **525** 1032–41

Proffitt C R and Michaud G 1991 Microscopic diffusion *Astrophys. J.* **380** 238

Turck-Chièze S, Däppen W, Fossat E, Provost J, Schatzman E and Vignaud D 1993 The solar interior *Phys. Rep.* **230** 57–235

Turck-Chièze S, Nghiem P, Couridat S and Turcotte S 2000 *Solar Physics* to be published

Zahn J P 1998 Macrophysics: large-scale advection, turbulent diffusion, wave transport *Space Sci. Rev.* **85** 79–90

Sylvaine Turck-Chièze

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Irradiance

Irradiance is the power from a source of radiant energy that is incident on unit area of a surface located at some distance from the source. The Sun's total irradiance is 1365 ± 2 W m^{-2}. This is the total radiant energy per unit time (i.e. power) at all wavelengths that the Earth receives on unit area of its surface from the entire solar disk when separated from the Sun by one astronomical unit (1 AU $= 1.49 \times 10^{13}$ cm).

Spectral irradiance is the power per unit area attributable to radiation within a specified wavelength interval. The Sun emits radiation that ranges in wavelength across the entire electromagnetic spectrum, from very short-wavelength x-rays, to ultraviolet (UV), visible, infrared (IR) and very long-wavelength radio waves. Figure 1 illustrates the dependence of solar spectral irradiance on wavelength, a dependence that covers more than six orders of magnitude and peaks in the visible spectrum.

Solar irradiance is not constant. Although the total irradiance has in the past been termed the solar 'constant', levels of total and spectral irradiance fluctuate continuously in concert with solar activity (see SOLAR ACTIVITY). During recent epochs of high solar activity, near maxima of the 11 year solar cycle (see SOLAR CYCLE), mean levels of total irradiance increased by about 0.1% (1.3 W m^{-2}) relative to solar cycle minima levels. Spectral irradiance also increases with solar activity, by different amounts depending on wavelength, as shown in figure 1 for a recent 11 year cycle. Small increases of less than a few tenths of a per cent occur at visible and near-infrared wavelengths, whereas radiation at the shortest x-ray and longest radio wavelengths can change by factors of two and more. Note that the actual radiant energy falling on the Earth varies by $\pm 3\%$ annually because of regular changes in the distance of the Earth from the Sun. This wavelength-independent fluctuation is not intrinsically solar but, rather, a geometrical effect. Solar irradiance variability refers to changes arising in the Sun itself, specified at a fixed Sun–Earth distance of 1 AU.

Acquiring reliable observations of solar irradiance has proven a challenging technological task. The longest continuous irradiance database—that of total (i.e. spectrally integrated) irradiance—has yet to cover two 11 year solar activity cycles. Since the late 1970s, more than five total irradiance instruments on various space platforms have utilized electrical substitution cavity pyrheliometers to achieve total irradiance measurements with uncertainties approaching $\pm 0.2\%$ and long-term repeatabilities of the order of $\pm 0.01\%$. Spectral irradiance observations are less reliable and less frequent than are total irradiance observations. Even in the UV spectrum, where most spectral irradiance measurements have been made, uninterrupted monitoring has yet to be accomplished for a complete solar cycle. Using combinations of redundant gratings and filters (to achieve spectral dispersion), photoelectric detectors and on-board calibration tracking, solar UV spectroradiometers that have operated continuously in space since late 1991 are at present achieving uncertainties in the range ± 6–10%, and long-term repeatabilites approaching $\pm 1\%$. Regular monitoring of the Sun's visible and infrared spectrum remains to be undertaken.

Close physical relationships between the Sun's radiative output and magnetic fields in the solar atmosphere (see SOLAR MAGNETIC FIELD) facilitate the primary mechanisms of solar irradiance variability. Magnetic fields, when present in the Sun's atmosphere, perturb its radiative processes by altering temperature and density (see, for example, SOLAR SPECTROSCOPY AND DIAGNOSTICS). Thus, magnetic active regions (see ACTIVE REGIONS) are sites of local enhancement or depletion of solar radiation. Particularly important for their effects on solar irradiance are dark sunspots—in which radiation is depleted (see SUNSPOTS)—and bright faculae—which are sources of enhanced radiation (see SOLAR PHOTOSPHERE). The variable occurrence and inhomogeneous distribution of both sunspots and faculae on the solar disk throughout the solar cycle produce net fluctuations in irradiance (i.e. the emission from all radiative sources integrated over the entire disk), with strong wavelength dependences. Particularly significant perturbations ($>100\%$) occur in emissions at the shortest (x-ray) and longest (radio) wavelengths that are formed in the highest and hottest layers of the solar atmosphere.

Past and future solar irradiance variability potentially exceed the range detected by space-based solar irradiance observations. Compared with historical proxy records of solar activity, including sunspot and geomagnetic records, and cosmogenic archives in ice cores and tree rings, current solar monitoring has been accomplished during an epoch of overall high levels of solar activity and larger than average 11 year cycle amplitudes. Associations of modern magnetic sources with historical solar activity proxies suggest that total irradiance may have increased a few tenths of a per cent since an epoch of anomalously low solar activity during the seventeenth century. Called the Maunder minimum, this episode of solar quiescence appears to be but the most recent of similar episodes that have occurred semiregularly in the Sun's multicentennial past, and are expected in its future.

Solar irradiance is the primary energy source for the Earth. Thus it is a crucial geophysical quantity for modeling and understanding a myriad of terrestrial processes including global climate change and space weather (see SOLAR–TERRESTRIAL CONNECTION). Radiation at wavelengths longer than 300 nm penetrates though most of the Earth's atmosphere to within 15 km of the surface. This radiation comprises 99% of total irradiance and its fluctuations have long been speculated to alter surface temperatures and other climatological phenomena on decadal and centennial time scales, vying with greenhouse gases and sulphate aerosols in the industrial epoch. The Earth's atmosphere absorbs the remaining 1% of the Sun's irradiance, including the high-energy UV, extreme

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

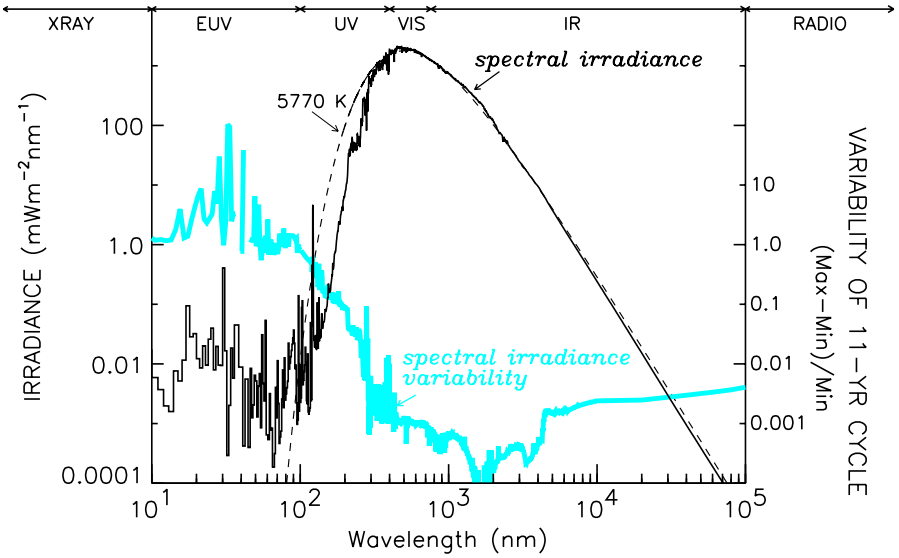


Figure 1. The spectrum of the Sun's radiation incident on top of the Earth's atmosphere (thin dark curve), and of a blackbody at 5770 K (thin dashed curve). The Sun's total irradiance is the integral over all wavelengths of this spectral irradiance. The broad spectral bands identified along the top of this figure are designated the extreme ultraviolet (EUV), ultraviolet (UV), visible (VIS) and infrared (IR). The x-ray spectrum lies shortward of the EUV spectrum and at wavelengths longward of the IR is the microwave or radio portion of the solar spectrum. Also shown (thick gray cuvre) are estimates of the variations in solar spectral irradiance for a recent 11 year activity cycle. At wavelengths longer than 400 nm these estimates are speculative because very limited observational data exist.

ultraviolet (EUV) and x-rays. These radiations would have detrimental biological effects were they able to penetrate to the Earth's surface. Rather, short-wavelength solar radiation provides the primary energy input to the atmosphere, where it initiates and controls the ozone layer in the stratosphere and, at higher altitudes, the ionosphere and thermosphere. More variable by orders of magnitude than the total irradiance, solar ultraviolet emissions may impact global change indirectly through middle atmosphere coupling with the biosphere. Solar heating and ionization perturbations of the upper atmosphere by highly variable extreme ultraviolet radiation impact communications and satellite drag.

Spectrum of the Sun's radiation

The spectrum of the Sun's radiation—its distribution of irradiance with wavelength, shown in figure 1—is a combination of continuum, emission and absorption features. Solar continuum radiation has a smoothly varying spectral shape similar to that of a black body near 5770 K (also shown in figure 1), which is the approximate temperature of the Sun's visible surface. The Sun's surface is defined as the layer of its atmosphere from which radiation at 500 nm (near peak irradiance levels of the blackbody spectrum) emerges with unit optical depth. Most (>99%) of the Sun's irradiance—that at wavelengths from 160 nm to 5 μm—also emerges from the vicinity of the solar surface, in a layer of the Sun's atmosphere a few hundred kilometers thick called the photosphere.

Numerous emission and absorption lines are superimposed on the dominant solar continuum. Gases in the Sun's atmosphere cause these spectral features by either absorbing the underlying continuum radiation (producing, for example, Fraunhofer lines) or by augmenting the continuum radiation with local emission processes. Least absorption by the solar atmosphere of the underlying photospheric continuum emission occurs in the vicinity of 1.6 μm (1600 nm). This radiation thus emerges from the deepest observable layers of the Sun and is the least variable region of the solar spectrum (see figure 1).

Many species in the solar atmosphere, for example Al, Mg, Ca, O, Fe, He and H, absorb radiation at UV, visible and IR wavelengths. In the visible and infrared spectrum their absorption lines are superimposed on the dominant 5770 K blackbody continuum. In the UV spectrum, however, this line absorption is so strong that it depletes significantly the underlying continuum radiation. At about 500 km above the Sun's visible surface the solar atmosphere temperature has a minimum value of 4500 K that delineates the upper boundary of the solar photosphere from the overlying chromosphere and corona. The solar spectra near 160 nm and from 75 μm to 300 μm originate near this temperature minimum region.

Radiation at wavelengths increasing shortward of 160 nm and longward of 300 μm emerges from increasingly higher, and generally hotter, layers of the solar atmosphere. In the solar chromosphere and corona temperatures in the range of 10^4 to 10^6 K produce excited and ionized states of atoms. Emission processes of these species emit radiation with characteristic spectral signatures whose strengths depend on the local temperature and composition of the solar atmosphere. Of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

these features, the emission by Hydrogen I produces the strongest line, at 121.57 nm, clearly evident in figure 1. Although the net irradiance of line emission at EUV and x-ray wavelengths is many orders of magnitude less than the visible continuum irradiance, it nevertheless exceeds by orders of magnitude the blackbody radiation in this short-wavelength spectral regime.

Observations of total irradiance

Attaining reliable observations of the Sun's total irradiance—the energy radiated at all wavelengths—has been a goal for more than a century, motivated by the need to specify fluctuations in the radiative energy input to the Earth. On the basis of observations made with ground-based instrumentation from 1874 to 1976, the total irradiance of the Sun was considered to be invariant with time, and hence designated the solar 'constant'. But the measurements lacked the accuracy and repeatability to establish either the correct absolute value of the irradiance, or its variability, to better than about half a per cent.

During the past two decades, solar radiometers flown on orbiting spacecraft (to eliminate interference by the Earth's atmosphere) have detected real total irradiance variability at levels of tenths of a per cent. These observations have achieved uncertainties in absolute calibration of $\pm 0.2\%$ and long-term repeatabilities better than 0.01% by utilizing the technique of electrical substitution radiometry, combined with detailed characterization and monitoring of in-flight sensitivity changes in the radiometers. A cavity in the instrument collects solar photons independently of their wavelength. The inside surface of the cavity is blackened and shaped so that every absorbed photon contributes thermal energy when a shutter permits solar radiation to reach the cavity. With the shutter closed, electrical power substitutes for the radiant power to maintain the cavity at a constant temperature. Knowing the aperture of the cavity then provides a direct measure of the incident total irradiance in absolute units of power per unit area (W m^{-2}), at the known distance of the cavity from the Sun.

The five or more radiometric instruments that have observed the Sun's total irradiance at various times during the past two decades have each realized a slightly different absolute scale, uncertainty and long-term stability. The various instruments have different cavity, baffle and aperture geometries, thermal losses, drift compensations and electronic readouts, and have flown on space platforms with different thermal stabilization and pointing capabilities. Figure 2 shows a composite record of daily mean total irradiance obtained from the cross-calibration of overlapping datasets to adjust for scale differences, taking into account drifts in the sensitivities of the individual radiometers during different phases of their missions. Whereas the time series in figure 2 presents the most probable observational record of the Sun's total irradiance, individual radiometric data at times deviate notably from this composite record, and from each other, because of instrumental sensitivity drifts. A primary cause of these drifts is solar exposure itself, which alters the blackness of the cavity surface and affects the thermal energy conversion. Drifts in temperature-sensitive electrical components and in the orientation of the cavity to the Sun also contribute. The most successful observations of total irradiance are achieved from instruments that utilize a number of redundant cavities with widely different duty cycles to effect in-flight calibration tracking, and that are mounted on thermal- and view-stabilized platforms.

Time scales of variability present in the observational record of total irradiance range from decades to minutes. Prominent in the composite record in figure 2 is an 11 year cycle that tracks closely the solar activity cycle, represented in this figure by the smoothed sunspot number. Monthly to yearly irradiance changes reflect the impact of the evolution of magnetic complexes that comprise sunspots and faculae. Significant irradiance dips of a few tenths of a per cent on a monthly and semimonthly cadence are associated with the impact on solar irradiance of the Sun's rotation on its axis. This rotation modulates irradiance by presenting different populations of magnetic active regions to the Earth's view. Peaks in the power spectrum shown in figure 2 identify these effects near 13 and 30 days and in additional longer period peaks. Total irradiance responds as well to the 5 min pressure mode oscillation of the solar interior (see HELIOSEISMOLOGY: THEORY, HELIOSEISMIC OBSERVATIONS), and possibly to the hourly time scales of gravity modes as well, but these fluctuations are not resolved in the daily mean data in figure 2.

At present two instruments are monitoring the Sun's total irradiance from space: the Active Cavity Radiometer Irradiance Monitor (ACRIM II) on the Upper Atmosphere Research Satellite (UARS, launched in 1991) and the Variability of Irradiance and Gravity Oscillations (VIRGO) instrument on the Solar Heliospheric Observatory (SOHO, launched in 1995). NASA's Earth Observing System (EOS), the European Space Agency (ESA) attached pallet to the International Space Station (ISS), France's CNES and the US National Polar-orbiting Operational Environmental Satellite System (NPOESS) each have programs for subsequent measurements. Acquiring reliable knowledge of the Sun's total irradiance variability on decadal time scales, independent of instrumental drifts, will require the indefinite continuation of solar irradiance monitoring using radiometers with greater stability than those that have flown in the past two decades, mounted on space platforms with highly stabilized thermal and pointing environments.

Observations of spectral irradiance

Reliable observations of the Sun's spectral irradiance must, like those of total irradiance, be made from space to avoid contamination by the Earth's overlying atmosphere. In particular, the solar spectrum at wavelengths shorter than 310 nm can only be observed from space as the Earth's atmosphere absorbs this radiation entirely. Because of their importance for understanding processes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

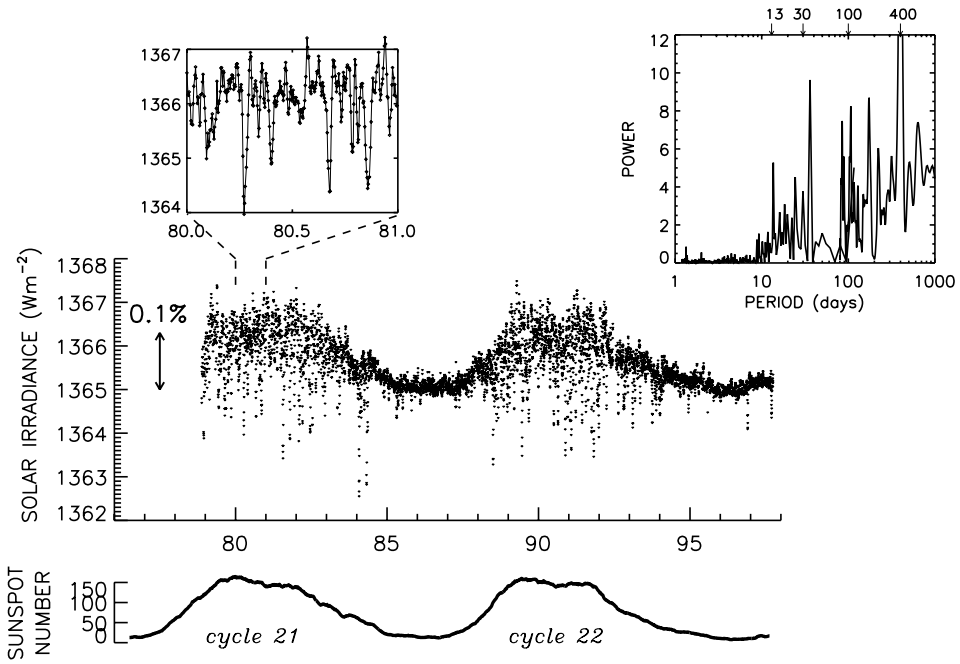


Figure 2. Daily mean values of the Sun's total irradiance are shown from 1979 to 1997. This composite irradiance record was constructed from observations made by four different solar radiometers at various times throughout the epoch. Evident is a pronounced 11 year cycle with an approximate peak-to-peak magnitude of 0.1%, on which are superimposed larger weekly to monthly fluctuations of a few tenths of a per cent. The 11 year total irradiance cycle tracks the general level of solar activity, indicated by the sunspot numbers in the lowest plot. The subset of data shown in the upper left hand plot provides details of rotational modulation, and the power spectrum in the upper right hand plot summarizes various apparent cycles in the composite total irradiance record.

in the Earth's atmosphere, the UV, EUV and x-ray regions of the solar spectrum have been monitored far more frequently than has any other part of the solar spectrum. But even these shorter-wavelength measurements have, over the past two decades, been made only intermittently. Except for two narrow x-ray bands, which Geostationary Operational Environmental Satellites (GOES) have monitored operationally since the 1970s, continuous reliable time series of solar spectral irradiance have not been procured for even one 11 year solar cycle. Figure 3 provides a sample of UV spectral irradiance monitoring—that made of the upper photospheric emission at 200 nm by instruments on the UARS and Solar Mesosphere Explorer—and of the coronal soft x-ray 0.1–0.8 nm fluxes.

Solar spectral irradiance instruments typically operate only over a limited wavelength range. They achieve wavelength discrimination by employing optical elements, usually gratings or spectral band filters, to disperse or select specified bands of the solar spectrum. Since narrow wavelength bands comprise significantly less solar power than the 1365 W m^{-2} total irradiance, spectral irradiance instruments utilize detectors, typically photomultipliers or diodes, which have sensitivities higher by many orders of magnitude than that of an electrical substitution cavity. These detectors amplify the small signal incident on them from the dispersing element when it is positioned at a known wavelength setting. Unlike total irradiance radiometers, spectral irradiance instruments are not electrically self-calibrating. Rather, pre-flight laboratory calibrations and characterizations establish the absolute irradiance scale at each wavelength by using standard irradiance sources and detectors. For all but the most recent UV spectral irradiance measurements, whose uncertainties approach ±6%, pre-flight calibrations of instruments typically achieve uncertainties in the range ±10–50%. On-orbit sensitivity degradation can introduce additional large uncertainties to long-term monitoring of spectral irradiance variability. Only the most recent UV spectral irradiance instruments, those flying on the UARS, have the capability for on-orbit calibration tracking of these changes, and achieve long-term repeatabilities approaching ±1% at some UV wavelengths.

Significant instrumental uncertainties are present in the historical database of solar spectral irradiance observations, arising from different realizations of absolute spectral irradiance scales combined with in-flight sensitivity drifts of the instruments. Exposure to solar radiation degrades the reflectivities and transmittances of optical elements in solar spectroradiometers; slippage in grating drive mechanisms contributes uncertainties in the wavelength of the radiation. At most wavelengths in the solar spectrum, the uncertainties in the historical irradiance database are comparable to, or exceed, the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

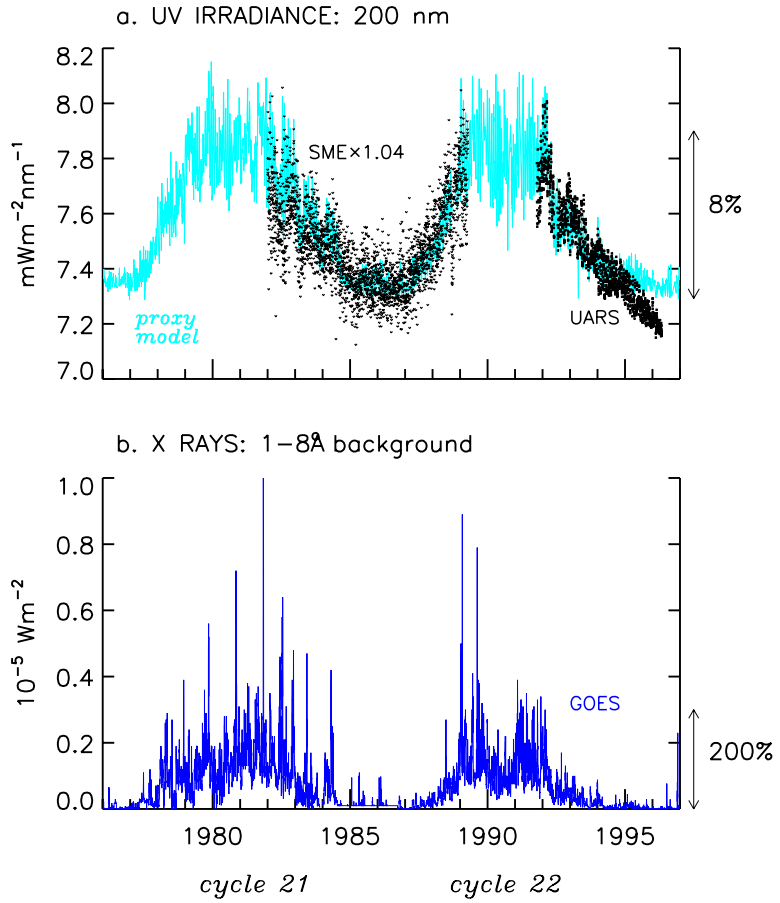


Figure 3. Variations occur in the Sun's radiation at wavelengths throughout the spectrum, and are shown in (*a*) the UV 200 nm irradiance measured by instruments on the SME and UARS spacecraft and (*b*) the GOES 1–8 Å background flux measured by a series of operational spacecraft. These spectral irradiance variations occur in concert with solar activity, whose long-term changes produced the variable sunspot record in figure 2. An empirical solar variability model that utilizes a facular proxy to estimate the 200 nm irradiance is also shown (gray line) for comparison with the SME and UARS 200 nm irradiance observations, which do not overlap in time and which cannot therefore be cross-calibrated to remove instrumental calibration scale differences.

amplitude of the solar cycle variability itself. In the visible and infrared solar spectrum, monitored by only a very few filter radiometers, instrument sensitivity changes in the range 5–50% far exceed the real solar variations which are thought to be in the range of a few tenths of a per cent; reliable observational determination of solar visible and infrared spectrum variability has thus far not been achieved. Knowledge of the irradiance and variability of even the strongest line in the entire solar spectrum—the emission of hydrogen Lyman α (121.6 nm)—remains uncertain by 30% to 50%, as does that of most EUV emission lines. Only in the spectral region from 250 to 120 nm have instrument sensitivity drifts been tracked with a repeatability of a few per cent, sufficient to estimates solar cycle changes of order of 3–30%, but these data exist thus far for less than one 11 year cycle.

In spite of their limitations, the available spectral irradiance measurements have established beyond doubt that the solar spectrum varies in concert with solar activity over multiple time scales. As shown in figure 1 for a typical 11 year activity cycle, the amplitude of the variability is strongly wavelength dependent. Spectrum variability occurs on time scales shorter than the solar cycle because active region evolution, solar rotation and interior oscillation also modulate spectral irradiance, in different degrees depending on the source of the irradiance within the solar atmosphere. In general, the emissions formed in higher, hotter solar atmospheric layers vary the most. Coronal x-rays are seen in figure 3(*b*) to vary by two orders of magnitude; these emissions are highly sensitive to magnetic activity, especially eruptive instabilities that produce flares, but they are insensitive to interior oscillations. EUV radiation from the solar chromosphere and transition region can vary by factors of two or more and UV radiation by a few to 50%; the 200 nm fluxes shown in figure 3(*a*) decreased by 8% from the maximum to minimum of the most recent 11 year cycle. Radiation formed in the vicinity of the Sun's surface is the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

least variable with solar cycle amplitudes of no more than a few tenths of a per cent. This radiation is insensitive to flares and sensitive to interior oscillations, which are detectable in irradiance time series made with sufficiently high time resolution.

At present, two instruments also on board the UARS are measuring solar spectral irradiance at wavelengths from 120 to 420 nm in 1 nm bands. These are the Solar Stellar Irradiance Comparison Experiment (SOLSTICE) and the Solar Ultraviolet Spectral Irradiance Monitor (SUSIM). GOES spacecraft continue operational solar x-ray monitoring, augmented since mid 1991 by soft x-ray observations made using an imaging (rather than full disk flux) instrument on board the Yohkoh spacecraft. Apart from VIRGO's filter radiometers on SOHO, which cover three selected near-UV, visible and IR narrow (5 nm) bands, the solar visible and IR spectrum is not monitored at present. Future programs include observations of the total and spectral irradiances by EOS, ESA and NPOESS, and of the EUV spectral irradiance by the Thermosphere Ionosphere Mesosphere Energetics and Dynamics spacecraft. Apart from NPOESS, these programs are of insufficient duration to determine long-term spectral irradiance trends, but they will advance significantly knowledge of shorter-term spectral irradiance variability.

Mechanisms of present irradiance variability

Mechanisms of solar irradiance variability arise from the fundamental connections between radiation and magnetism in the Sun's atmosphere. The prominent 11 year cycles in the total and spectral irradiance in figures 2 and 3 are simply global manifestations of these connections. So too are irradiance fluctuations that track multimonthly evolution of active regions, the 27-day rotation of the Sun on its axis, and a range of shorter time scale variability related to meso- and super granulation (see SOLAR PHOTOSPHERE: MESOGRANULATION; SUPERGRANULATION). In each case, the eruption, rearrangement and disappearance of magnetic fields in the solar atmosphere modify the Sun's radiative output. Magnetic fields perpetrate distinctive features in different solar atmosphere regimes, as displayed in the images in figure 4. These features, which include dark sunspots and bright faculae, affect solar irradiance at different wavelengths in different ways. Thus, the three time series in figure 4, of total, UV and x-ray irradiances emitted from respectively the photosphere, upper photosphere and corona, exhibit markedly distinct temporal variability even though solar magnetism is ultimately their common origin.

In the photosphere near the Sun's surface, from whence emerges near-UV, visible and IR radiation, sunspots are the pre-eminent signature of solar magnetism. Formed by compact clumps of magnetic fields, sunspots are darker and cooler than the surrounding photosphere because the magnetic fields inhibit the upward flow of energy from the Sun's convection zone to its visible surface. Less compact aggregates of magnetic field lines also abound in the photosphere. Being only slightly brighter than the surrounding photosphere, and much less compact than sunspots, these features, the faculae, are barely detectable in images of the solar disk recorded with visible continuum radiation, except near the limb. However, because of their greater dispersion over the entire solar surface they are equally important in modifying the net radiative output of the Sun (see FACULAE, SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES).

Hundreds of kilometers above the Sun's surface the primary evidence of magnetic fields are bright, rather than dark, activity features. In the largest of these regions, called plage, magnetic fields clump together to form complexes of activity that extend over much more of the solar disk than do the compact sunspots. Bright emission is also present in a network that covers most of the solar surface, and in a range of interspersed magnetic features that have smaller spatial scales than active regions. Plage and network are evident as regions of enhanced emission in the image of the Sun's chromosphere in figure 4 made in the core of the Ca K Fraunhofer line. These bright magnetic features are the chromospheric counterparts of photospheric faculae, and are sources of significantly enhanced UV and EUV emission from the upper photosphere, chromosphere and transition region. Thus the fluctuations in solar UV and EUV spectral irradiance track closely the evolution of bright faculae, with little discernible sunspot influence. Higher still, in the Sun's corona, magnetic fields form bright loops and extended magnetic complexes that obliterate distinct connections to sunspots, plage and network below. As seen in the Yohkoh soft x-ray images in figure 4, coronal magnetic structures at times extend over much of the solar disk; they emit prodigious levels of x-ray, EUV and radio flux and control irradiance variability at these wavelengths.

Specific knowledge of the areas and brightnesses of the various magnetic features present in different solar atmospheric regimes permits the empirical modelling of solar irradiance variability for comparison with directly observed irradiance data at different wavelengths. Ground based, visible light solar images record areas and locations of sunspots that are used to calculate sunspot darkening contributions. Other spectroheliograms, such as the Ca K images in figure 4, record the locations and areas of enhanced emission in plage, faculae and the surrounding bright network. Solar variability models that combine, in different amounts, global parametrizations of these two components can explain much of the observed irradiance variability from the photosphere and chromosphere. A model with significant contributions from both sunspot darkening and facular brightening accounts for almost 90% of the variance in the daily mean total irradiance data over nearly two decades, shown in figure 2. In comparison, the model estimates shown in figure 3(*a*), used to cross-calibrate two on-overlapping

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

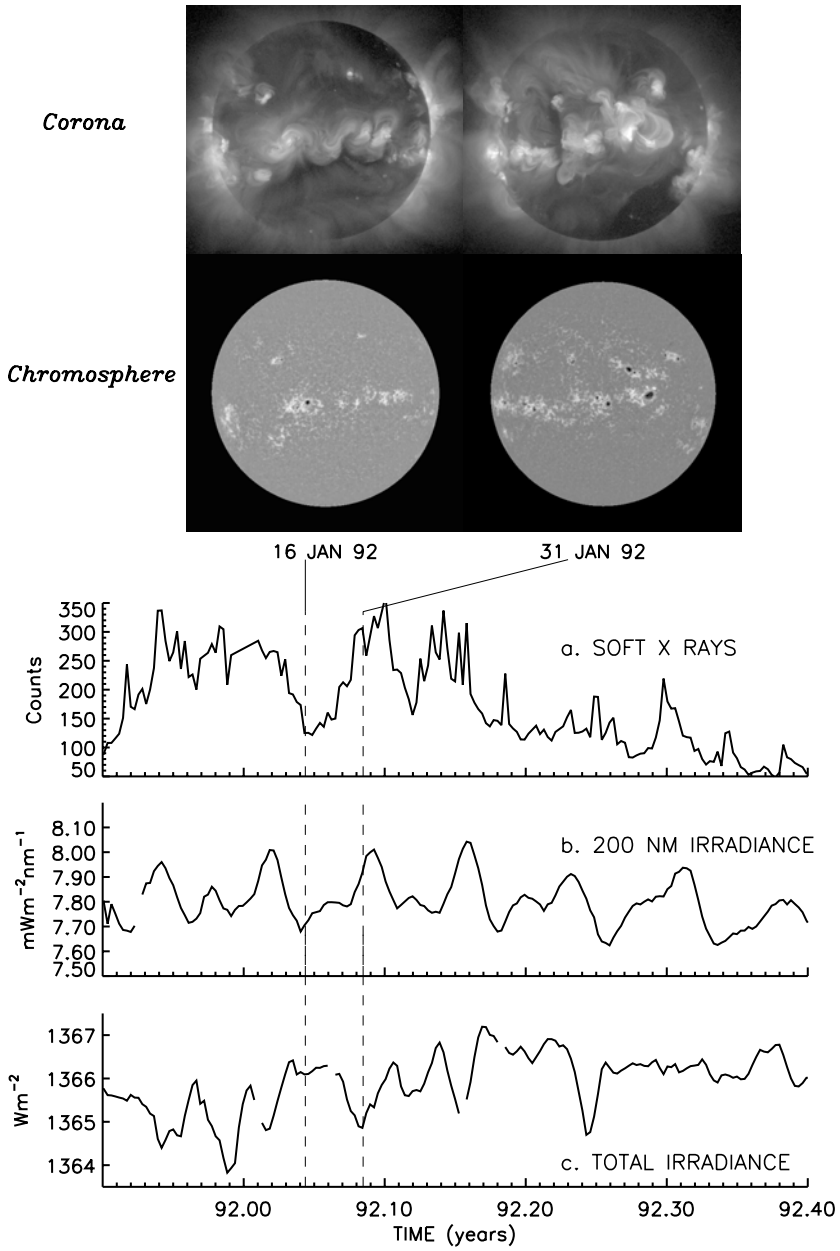


Figure 4. Solar activity and solar rotation alter the distribution of magnetic features in the solar atmosphere, viewed from Earth, as shown in two types of solar images made in 1992 near maximum levels of solar activity. These changes generate irradiance variability. Soft x-ray (upper) images measured by the Yohkoh spacecraft display the changing occurrence of global-scale magnetic features in the corona. Ca II K (lower) images measured at the Big Bear Solar Observatory show related but less extended magnetic features present in the chromosphere on the same days. Soft x-rays fluxes measured by Yohkoh (*a*), the UV 200 nm irradiance measured by UARS (*b*) and total irradiance (*c*) measured by ACRIM on UARS track the changing distributions of active regions projected toward Earth during solar rotation and over the solar cycle.

datasets of UV 200 nm irradiance, utilizes the facular component alone.

Irradiance variability on historical time scales

How might solar irradiance have varied in the past, and what are the prospects for future variability? Astrophysics and geophysics both motivate the need for this understanding. Examining long-term variability of other stars like the Sun can provide clues about broader aspects of stellar evolution (see SOLAR–STELLAR CONNECTION); and long-term solar variability has implications for terrestrial climate and global change on centennial time scales and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

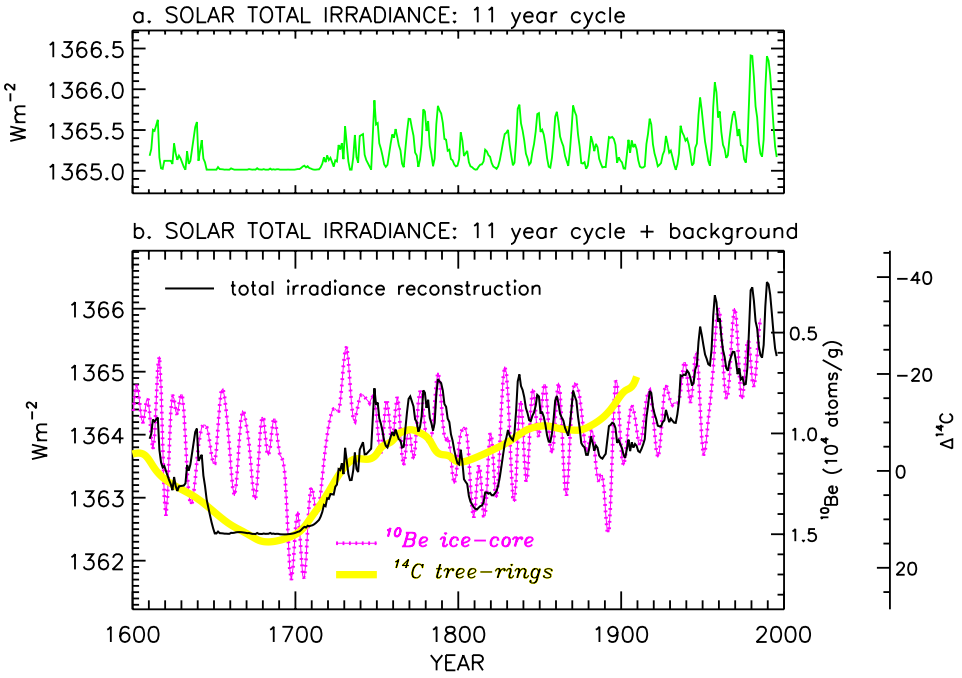


Figure 5. Reconstructions of the Sun's total irradiance, assuming in (*a*) an 11 year cycle alone and in (*b*) the 11 year activity cycle and a longer-term component based on the average amplitude of each sunspot cycle. This latter irradiance reconstruction (dark solid curve) is compared with ^{10}Be (small squares) and ^{14}C (thick gray curve) cosmogenic isotope records.

longer, thereby aiding in the specification of natural climate variability for comparison with anthropogenic influences (see SOLAR–TERRESTRIAL CONNECTION: LONG-TERM AND SHORT-TERM CLIMATE VARIABILITY).

A variety of solar, geomagnetic and geological records contain information about long-term solar activity fluctuations and the irradiance changes that presumably accompanied these fluctuations. Images of the Sun that depict sunspots have been made from ground based observatories since the seventeenth century, including during the Maunder minimum when for many years sunspots occurred only very infrequently. Conceptually, the sunspot record indicates that solar activity increased from anomalously low levels during the Maunder minimum to the present epoch of frequent sunspot occurrence with numbers as high as any witnessed during the past 400 years. Furthermore, indices of geomagnetic activity recorded at the Earth during the past 150 years indicate an overall upward activity trend concurrent with increasing solar cycle amplitudes. Geological records of ^{14}C and ^{10}Be cosmogenic isotopes in, respectively, tree rings and ice cores confirm that overall levels of solar activity have increased during the past three centuries. The isotope levels reflect variations in the terrestrial flux of galactic cosmic rays that arise from solar activity modulation of the magnetic coupling of the Sun and the Earth. Consistent with the evidence from sunspots, geomagnetic data and cosmogenic isotopes, the Sun's present activity is in the high range of levels seen in a collection of solar-like stars.

Two decades of space-based solar irradiance monitoring has thus been conducted during an era of overall high solar activity compared with levels during the past few centuries. This suggests the plausibility of solar irradiance levels lower than any yet observed in recent solar cycle minima, for example during the Maunder minimum. Speculated longer-term changes are not yet detectable in the limited-duration observational irradiance record. However, estimates of historical solar irradiances are possible by using solar activity proxies to parametrize irradiance variability mechanisms—such as the influences of sunspots, faculae and network as quantified for the present-day irradiance record. Figure 5(*a*) shows a reconstruction of the 11-year total irradiance cycle, alone, since 1610, whereas in figure 5(*b*) the 11 year irradiance cycle is combined with an assumed longer-term irradiance variability component. Adopting this longer-term component permits better agreement between the irradiance reconstruction and both the geomagnetic and cosmogenic isotope records, the latter of which are also shown in figure 5(*b*).

Total irradiance increased by about 0.25% from 1650 to the present, according to the historical reconstruction shown in figure 5. A variety of similar techniques yield estimates of long term total irradiance variability in the range of 0.2 to 0.5%. At UV wavelengths, long term spectral irradiance changes are estimated to be factors of two larger than the solar cycle changes. But these estimates of the Sun's irradiance variability on historical time scales

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

are by necessity highly speculative; and they will remain so until the observational record is of sufficient length to establish that long-term irradiance variability is in fact present, in addition to the confirmed variability during the 11 year cycle and on shorter time scales.

Bibliography

Cox A N, Livingston W C and Matthews M S 1991 *Solar Interior and Atmosphere* (Tucson, AZ: University of Arizona Press)

Foukal P V 1990 *Solar Astrophysics* (New York: Wiley)

Hufbauer K 1991 *Exploring the Sun: Solar Science since Galileo* (Baltimore, MD: Johns Hopkins University Press)

Lean J 1997 The Sun's variable radiation and its relevance for Earth *Ann. Rev. Astron. Astrophys.* **35** 33–67

Sonnet C P, Giampapa M S and Matthews M S 1991 *The Sun in Time* (Tucson, AZ: University of Arizona Press)

Judith Lean

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Magnetic Field

Electrical currents flowing in the solar plasma generate a magnetic field, which is detected in the SOLAR ATMOSPHERE by spectroscopic and polarization measurements (SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY). The SOLAR WIND carries the magnetic field into interplanetary space where it can be measured directly by instruments on space probes.

The structure of the solar magnetic field comprises a wide spectrum of spatial and temporal scales. A global order is apparent in the periodic reversals of the magnetic polarity in connection with the 11 yr cycle of SOLAR ACTIVITY. In the SOLAR PHOTOSPHERE the magnetic field is highly inhomogeneous with small-scale filaments formed through its interaction with the convective flows. In the upper layers of the solar atmosphere, the magnetic force dominates over all other forces, so that the magnetic field acquires a less intermittent structure and fills all available volume. Magnetically 'open' regions with diverging field lines reaching into interplanetary space and 'closed' regions pervaded by loop-like structures characterize the magnetic field in the CORONA. Further out, the solar wind dominates the dynamics and, in concert with solar rotation, imprints a spiral structure onto the magnetic field.

It is generally assumed that the magnetic flux observed in the solar atmosphere is generated by a hydromagnetic DYNAMO mechanism working near the bottom of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE). Due to MAGNETIC BUOYANCY, the field rises through the convection zone in the form of magnetic flux tubes. In the solar atmosphere, the emerging magnetic flux forms magnetically bipolar SOLAR ACTIVE REGIONS and sunspot groups. The various manifestations of solar activity are all intimately related to the magnetic field. Apart from its variability on timescales determined by the dominant convective patterns and the modulation of flux emergence in the course of the activity cycle, a long-term variability of the solar magnetic field is indicated by a marked variation of the intensity of the 11 yr cycles and by extended periods of very low solar activity, like the second half of the 17th century when sunspots were almost absent (see SOLAR ACTIVITY: LONG-TERM RECORDS).

Large-scale structure and active regions

A magnetogram (i.e. a map of the magnetic field on the solar surface, figure 1) exhibits bipolar magnetic regions (active regions) as well as large areas of apparently unipolar field in a network pattern, which outlines the boundaries of large convection cells (SOLAR PHOTOSPHERE: SUPERGRANULATION). In particular, such unipolar regions can nearly always be found around the solar poles. They suggest the existence of a global dipole component of the solar magnetic field, which is nearly aligned with the rotation axis. The orientation of this dipole field reverses in phase with the 11 yr cycle of solar activity (SOLAR CYCLE).

The network structure of the magnetic field outside active regions continuously rearranges itself in response to the birth and decay of supergranules with a timescale of several days. Together with an analogous process on the much smaller scale of granulation (see SOLAR PHOTOSPHERE: GRANULATION), this effectively represents a diffusion process, spreading magnetic flux over large areas and thereby allowing flux elements of opposite polarity to cancel. Besides such turbulent diffusion, two other processes contribute to the distribution and transport of magnetic flux at the solar surface. Firstly, a meridional flow of about 10 m s^{-1} directed from the equator to the poles and, secondly, latitude-dependent DIFFERENTIAL ROTATION. Computer simulations incorporating local flux emergence and the various transport processes reproduce the evolution of the magnetic field on large spatial scales remarkably well, including the reversals of the global dipole field.

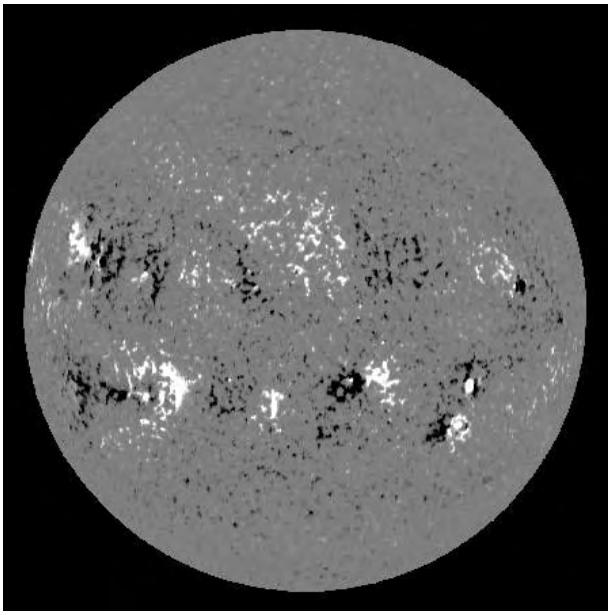


Figure 1. Map of the magnetic field in the solar photosphere on 25 July 1993. Black and white indicate positive and negative magnetic polarity, respectively. The dominant features are bipolar active regions and extended unipolar domains with a network structure outlining the convective flow pattern of supergranulation. (Courtesy National Solar Observatory, USA.)

The most conspicuous magnetic phenomena in the visible layers of the solar atmosphere are SUNSPOTS. Their magnetism was discovered in 1908 when G E Hale derived magnetic field strengths of a few 100 mT from the Zeeman splitting of spectral lines. Sunspots are found in active regions and typically appear in groups comprising spots of both magnetic polarities. The magnetic polarities in sunspot groups and active regions are roughly oriented in the east–west direction of solar rotation. They obey systematic polarity rules: during one 11 yr cycle, all active regions have the same east–west polarity orientation in the northern hemisphere of the Sun, while the opposite orientation is found in the southern hemisphere. All

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

polarities are reversed in the subsequent activity cycle. These properties reveal the magnetic nature of the 11 yr cycle and show that a full magnetic cycle has a duration of 22 yr. The polarity rules can be understood in terms of a large-scale, east–west oriented (toroidal) magnetic field below the visible surface. This field is antisymmetric about the equator and gives rise to active regions and sunspots at the surface through the emergence of magnetic loops (SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES).

Active regions and sunspots appear on the Sun within broad activity belts on both sides of the equator. In the course of the 11 yr cycle, these zones migrate from about 35 degrees latitude toward the equator. Upon dissolution of an active region, its magnetic flux spreads over an increasing area until it merges with the general network pattern of the background field.

The magnetic flux emerging at the solar surface shows a broad spectrum in both area and amount of magnetic flux of bipolar active regions: from large sunspot groups with a total flux (sum of both polarities) of more than 10^{14} Wb and with dimensions exceeding 100 Mm down to the small ephemeral active regions with diameters of some tens of Mm, containing a few times 10^{10} Wb. The number of active regions emerging per unit time on the solar surface increases strongly with decreasing total flux of the region: ephemeral active regions are capable of replenishing the magnetic flux of the background field within about 2 d. Since the total magnetic flux is not observed to accumulate in time, similarly effective processes of flux removal or flux cancellation must be at work. Apparently due to their small intrinsic length scales, however, these processes are difficult to observe and not well understood. While the ephemeral active regions obey the magnetic polarity rules (at least in a statistical sense) and so are part of the global solar cycle, at smaller length scales (below 1 Mm) magnetic flux is continuously reprocessed by fragmentation, clustering, collision and advection of small flux elements by convective flows, mainly granulation.

The polarity rules of active regions, the latitude migration of the activity belts, and the reversals of the dipole field demonstrate the global nature of the solar magnetic field. On the other hand, this large-scale organization is accompanied by strong fluctuations on a wide range of spatial scales.

Photospheric flux tubes

The inhomogeneous distribution of magnetic flux on large spatial scales (active regions and network structure) illustrated in figure 1 continues down to the smallest scales that are accessible to observation (see figure 2 as an example of a medium-resolution magnetogram): the major part of the magnetic flux permeating the solar photosphere is filamented into a spectrum of flux tubes (SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES) with field strength above 100 mT and mainly vertical orientation. These flux tubes contain almost all of the magnetic energy. The largest photospheric flux tubes are sunspots with diameters between a few Mm and 50 Mm and central

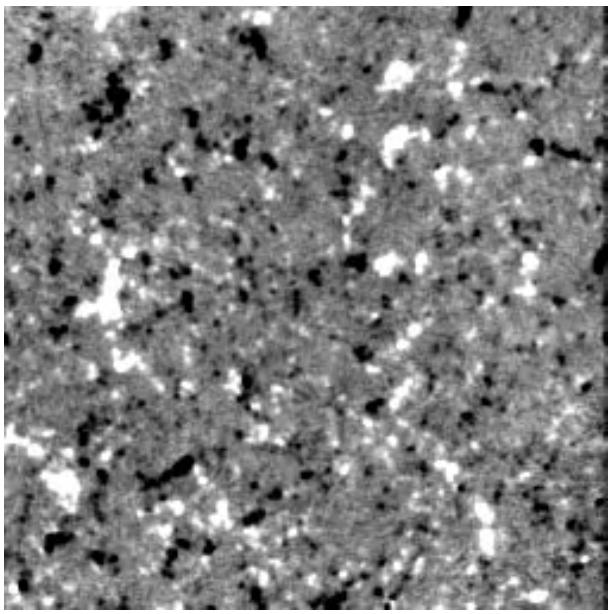


Figure 2. Magnetogram showing the intermittent magnetic field distribution in a 'quiet' region of the solar photosphere (size $\simeq 160 \times 160$ Mm^2). Black and white indicate negative and positive magnetic polarity, respectively. Most of the magnetic flux is contained in patches of strong field appearing in a broad range of sizes down to the resolution limit of the observation; there is no dominant polarity. (Courtesy C Denker, Big Bear Solar Observatory, New Jersey Institute of Technology.)

magnetic field strengths between $\simeq$200 mT and 350 mT. Sunspots represent the biggest accumulations of magnetic flux in the photosphere. Large sunspots are composed of a dark core (see SUNSPOT UMBRA) surrounded by a striated annular region called penumbra (see SUNSPOT PENUMBRA). Small sunspots without a penumbra are called pores (see SUNSPOT PORES); they comprise the size range between about 0.5 Mm and 5 Mm. Flux tubes smaller than pores appear bright in white light and in selected spectral bands containing many spectral lines; they are called magnetic elements (figure 3).

The (relative) darkness of sunspots is caused by the suppression of convective energy transport due to their strong magnetic field (MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION). On the other hand, a small flux tube appears brighter than the average solar surface because its partially evacuated interior is efficiently heated by radiation from its surroundings in the deeper layers and, presumably, also by dissipation of mechanical energy in its higher atmospheric layers. Note that, for a cylindrical flux tube, the ratio of the heating surface area of its 'wall' and the internal volume to be heated is inversely proportional to the tube radius. The hot walls of small flux tubes become best visible near the limb of the Sun, where the combined effect of many flux tubes gives rise to bright FACULAE. Averaged over the whole Sun, the enhanced brightness of the magnetic elements dominates over the reduced energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

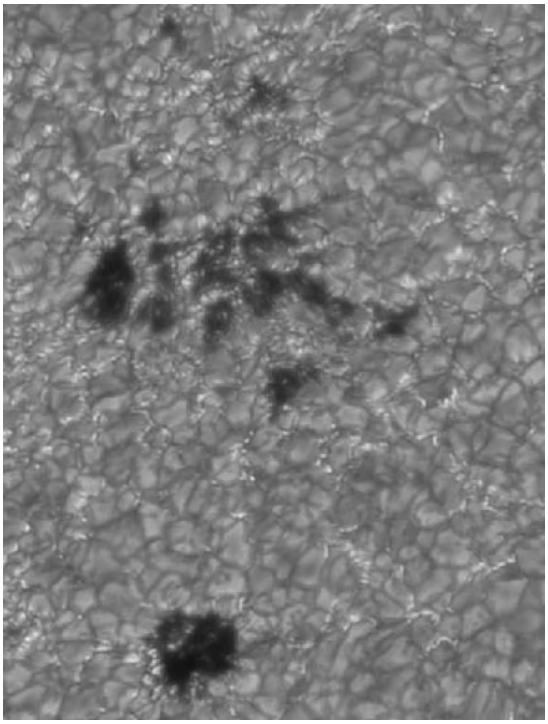


Figure 3. Size-dependent brightness signature of magnetic flux tubes in the solar photosphere: large tubes (sunspots and pores) are darker than the average solar surface, while small flux tubes (magnetic elements) appear as bright dots, often strung together like chains. The roughly polygonal pattern is due to convective motions (granulation). The darker intergranular lanes are cool downflow regions while the brighter granules are formed by hot upflows. The magnetic elements are predominantly located in the intergranular lanes. The image has been taken with a 10 nm wide filter around 468 nm wavelength. (Courtesy G Scharmer, Stockholm Observatory.)

flux in the sunspots, so that the total solar radiation output increases with growing magnetic flux in the photosphere. As a consequence, the brightness of the Sun increases by about 0.1% from minimum to maximum of the activity cycle (SOLAR IRRADIANCE).

In contrast to the rather strong variation in brightness, the field strength inside flux tubes depends only weakly on their size. In fact, the average over their full cross section remains nearly constant at values between 150 mT and 200 mT. Only the very smallest flux tubes show indications of a significantly reduced field strength. Note that the magnetic pressure (MAGNETOHYDRODYNAMICS) of a field of 150 mT is comparable to the gas pressure at the base of the photosphere, so that the magnetic flux tubes must be nearly evacuated at that height in order to attain force balance.

Observations with low spatial resolution average over the intermittent structure of the magnetic field and thus yield a mean magnetic flux density. The range of measured values between roughly 0.1 mT in magnetically 'quiet' areas and up to 30 mT in active regions (outside sunspots) can be understood in terms of a varying area fraction occupied by strong flux tubes: values between less than 0.1% in quiet areas and 20% in active regions are indicated.

Spectroscopic observations in circularly polarized light, which originates exclusively from the magnetic regions in the photosphere, reveal details of the internal structure of magnetic flux tubes. This has led to semi-empirical models for the thermodynamic and magnetic structure of flux tubes in the network and in plage regions. These results indicate that the photospheric layers of small flux tubes are hotter than their environment and that the plasma within a tube is in rapid motion caused by flows, waves, oscillations and shocks. The tubes are embedded in downflowing gas and often show rapid horizontal displacements. This dynamical nature of the tubes is probably relevant for transporting mechanical and magnetic energy into the upper atmosphere, where it contributes to the heating of these layers (CHROMOSPHERE: HEATING MECHANISMS; CORONAL HEATING MECHANISMS).

If the area fraction occupied by flux tubes is not too large, the distribution of the magnetic flux tubes over the photospheric 'surface' of the Sun reflects the convective flow patterns: magnetic flux tends to accumulate in the downflow lanes. On the scale of supergranulation, this leads to the large-scale network pattern with a typical mesh size of 30 Mm. A similar structure is repeated on the scale of granulation. Such a distribution indicates the operation of flux expulsion, i.e. the sweeping of vertically oriented magnetic field lines into the downflow regions by the horizontal convective flows. This process is capable of enhancing the field to about equipartition of the magnetic energy density with the kinetic energy density of the flow, leading to a field strength of about 50 mT. In fact, freshly emerging magnetic flux appears in the photosphere with about this equipartition field strength. Theoretical studies and numerical simulations suggest that further intensification of the field strength up to 150 mT is caused by the combined effects of the strongly superadiabatic stratification of the convective region just below the photosphere, the suppression of convective motions by a magnetic field of sufficient strength, and the efficient cooling of the photospheric region due to radiative losses. These processes conspire in exciting an instability of convective type. This results in a strong downflow and cooling within the magnetic region, leading to a partial evacuation of its photospheric layers. The pressure force exerted by the external plasma causes a contraction of the tube and thereby amplifies the field. This process continues until the magnetic field becomes sufficiently strong to suppress the instability and maintain lateral force balance. The principle of this convective collapse mechanism is schematically sketched in figure 4.

The collapse process loses its efficiency for flux tubes with a diameter smaller than the photon mean free path of about 100 km at the base of the photosphere. At that limit, the tube interior is kept at the temperature of its environment and the evacuating downflow is throttled, so that the equipartition limit cannot be exceeded. In

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

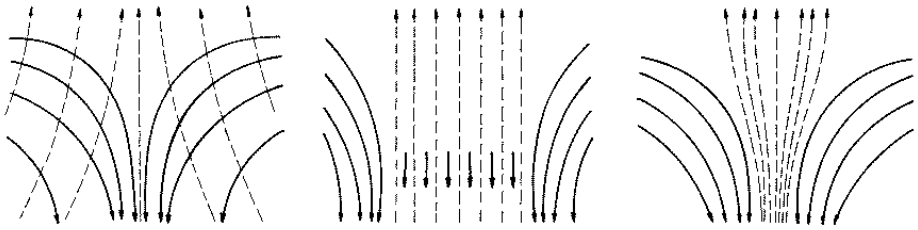


Figure 4. Schematic sketch of the processes that lead to intensification of magnetic field in the solar (sub)photosphere. Left: the horizontal flows of granular convection sweep magnetic flux towards the downflow regions (full: streamlines, dashed: field lines). Middle: magnetic forces suppress the convective motions when the magnetic energy density becomes comparable to the kinetic energy density. This throttles the energy supply into the magnetic region, the gas cools off due to radiative losses and the internal downflow is enhanced. The superadiabatic stratification further amplifies the flow. Right: the downflow has evacuated the upper layers and the magnetic field in the quenched tube has increased accordingly to maintain lateral force balance.

fact, observations in the infrared spectral range reveal the existence of small magnetic features with about equipartition field strength in the photosphere. This relatively weak field is dominated by the convective motions; the corresponding field direction is found to be predominantly horizontal (PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS).

An illustration of the coexistence of strong vertical flux tubes and weaker horizontal field in the solar photosphere is given in figure 5, which shows a snapshot from a numerical simulation of the interaction between magnetic field and convective flow in the photosphere. A vertical flux tube with a strong field has formed by the convective collapse process in a downflow region. Magnetic flux that has not been swept to the flux tube is wrapped up by the convective pattern and brought to the surface by upflows, where it forms patches of mainly horizontal field with about equipartition field strength. Such simulations are also consistent with the observed dynamic nature of the flux tubes: they show strong internal flows, wave and shock propagation, as well as a horizontal swaying of the flux tube.

Besides the flux tubes and the equipartition-field features, a third component of the photospheric magnetic field is indicated by measurements of the Hanle depolarization in spectral lines. In contrast to the measurements based on the circular polarization, this method is most sensitive to weak fields and the signal does not vanish for mixed magnetic polarity within the observational resolution element. Measurements based upon this technique suggest the existence of a small-scale turbulent magnetic field with no preferred direction and a field strength in the range of 0.4–4 mT.

Upper atmosphere

In the atmospheric layers above the photosphere, magnetic field measurements are difficult. Since the temperature rises to values of about 2×10^4 K in the chromosphere and up to a few times 10^6 K in the corona, the atoms become highly ionized and most of their spectral lines shift into the (extreme) ultraviolet range, where the Zeeman splitting is very small while the lines are thermally

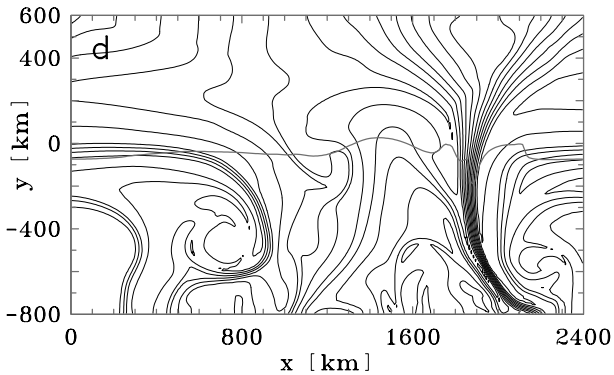


Figure 5. Magnetic field distribution arising during numerical simulation of the interaction between magnetic field and convective flows under conditions prevailing in the solar photosphere. Shown are magnetic field lines in a 'box' spanning a horizontal distance of 2400 km on the Sun, while the vertical coordinate ranges from 800 km below the surface (optical depth unity at 500 nm wavelength for a mean solar atmosphere model) to 600 km above it. The mainly horizontally running line located around $y = 0$ indicates optical depth unity (from Grossmann-Doerth *et al* 1998).

broadened due to the large temperature of the gas. Moreover, the magnetic field strength in the flux tubes decreases with height due to the diminishing external gas pressure. Alternatives to the Zeeman effect are the Hanle effect (SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS) and gyro-synchrotron emission (SOLAR SPECTROSCOPY: CONTINUUM RADIO EMISSION AND DIAGNOSTICS), but the precision of the measurements and their spatial resolution are much lower than for the magnetic field in the photosphere. Field strengths obtained by these methods range from 1–10 mT in PROMINENCES and reach 10 mT at the base of the corona in active regions; above sunspots, values up to 100 mT are found. Further out, a typical field strength at two solar radii is 0.1 mT.

For increasing height in the solar atmosphere, the magnetic field becomes dynamically as well as energetically dominant: while the gas pressure decreases

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

exponentially with height, the magnetic field strength follows a radial power law above the height of the magnetic canopy (CHROMOSPHERE: MAGNETIC CANOPY), where the photospheric flux tubes merge; above this level, the magnetic field occupies all available space. As a consequence, in quiet regions the magnetic pressure becomes comparable to the gas pressure in the chromosphere and lower corona, while it exceeds the gas pressure in active regions. Apart from very dynamic phenomena like SOLAR FLARES or SOLAR CORONAL MASS EJECTIONS, the magnetic energy density also exceeds the kinetic energy density contained in the bulk motion of the plasma.

Since there is no component of the magnetic Lorentz force along the magnetic field lines, material motion in the upper atmosphere is strongly guided by the magnetic field. Moreover, the particle collision rate decreases with height and thermal conduction in the corona becomes anisotropic, i.e., heat is conducted much easier parallel to the field lines than in the perpendicular direction. Under these circumstances, the magnetic field channels the mass and energy flows in the upper atmosphere, so that any structure visible in absorbed or emitted radiation naturally tends to outline the pattern of magnetic field lines. The overall arrangement of the field can thus be inferred from the geometry of intensity structures. Examples are the myriads of CORONAL LOOPS, which are best observable in soft x-rays (SOLAR SPECTROSCOPY AND DIAGNOSTICS: X-RAY EMISSION), the chromospheric fibrils and spicules visible in the hydrogen H_α line (CHROMOSPHERE: FIBRILS; SPICULES), and the white-light morphology of the corona as observed during total eclipses or with coronographic telescopes (SOLAR TELESCOPES AND INSTRUMENTS: GROUND). All this structure owes its existence to the magnetic field since the magnetic forces can easily balance the corresponding pressure gradients.

The magnetic topology of the corona is divided into 'closed' regions dominated by coronal loops connecting regions of opposite magnetic polarity in the photosphere (like active regions) and large unipolar 'open' regions with roughly radial field lines. While the closed regions are bright in x-ray and EUV emission, the open regions are more tenuous and, therefore, show much less emission (CORONAL HOLES). The magnetic force hampers the flow of the solar wind in the closed regions, while matter can freely stream outward along the field lines in open regions. This is consistent with the observation that the coronal holes are the source regions of the fast solar wind while only a relatively low-speed and unsteady outflow emanates from the closed regions. Because the field lines are rooted in the rotating solar body, the field develops a spiral pattern in interplanetary space (SOLAR WIND: MAGNETIC FIELD).

A relatively simple configuration develops around the phases of minimal solar activity: the polar regions are open while closed fields dominate near the equator; further out, the kinetic energy density of the solar wind dominates also in the equatorial regions and draws the field open, so that coronal streamers form. During solar activity maximum, many bipolar regions dominate the magnetic field distribution so that the closed regions cover a larger part of the corona and the magnetic configuration becomes more complicated.

Generation of magnetic flux

What is the origin of the solar magnetism? The electrical conductivity of the solar plasma is sufficiently large to prevent the dissipation of a large-scale magnetic field permeating the whole Sun over its lifetime. Consequently, if the Sun inherited a magnetic field from the interstellar cloud that it was formed from, the large-scale part of this field could still remain as a slowly decaying fossil field. However, if such a field exists, it must be buried in the solar interior since no indication for its existence is found in the solar atmosphere. On the other hand, the nearly rigid rotation of the solar radiative core below the convection zone found by HELIOSEISMOLOGY can be most easily understood in terms of magnetic stresses exerted by a weak fossil field of the order of 0.1 mT (SOLAR INTERIOR: ROTATION).

The magnetic field observed at the surface varies on a broad range of timescales and even the large-scale dipole component reverses during each 11 yr cycle. This indicates that induction processes due to plasma motion play a crucial role in the generation and evolution of the solar magnetic field (DYNAMO THEORY). Two kinds of such motion are of particular relevance: convection and differential rotation. The latter is capable of generating a toroidal (east–west directed) magnetic field from a field in meridional planes (poloidal field). The existence of a strong toroidal field component is inferred from the east–west orientation and the polarity rules of sunspot groups. The dynamo cycle is completed by the generation of a poloidal field from the toroidal field through the systematic twist of convective motions under the influence of the Coriolis force. If the new poloidal field has reverse polarity with respect to the original field, the toroidal field generated, in turn, from that field is reversed as well. The full dynamo cycle is closed when a poloidal field of the same polarity as the original field has been generated. This sequence is in accordance with the characteristics of the 22 yr magnetic cycle of the Sun.

Theoretical considerations show that indeed differential rotation and convective flows are capable of sustaining a self-excited hydromagnetic dynamo that generates a periodically oscillating magnetic field. Models agree with many observed properties of the solar cycle provided the adjustable parameters of the models (typically related to the turbulent part of the velocity field) are properly chosen. Therefore, it is widely accepted that a hydromagnetic dynamo working in the convection zone generates the solar magnetic field. On the other hand, there is as yet no consensus about its detailed modus operandi. Currently discussed models propose that the dynamo mechanism operates near the lower boundary of the convection zone, some 200 Mm below the solar surface. There are two main reasons for this conjecture: (1) helioseismology has shown

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that in a layer at the bottom of the convection zone the rotational velocity varies rapidly with radius thus producing a strong local shear flow; (2) magnetic flux must be prevented from rapidly rising to the surface due to magnetic buoyancy before it could be sufficiently amplified by the dynamo process. Buoyancy can be suppressed in a layer of overshooting convection at the bottom of the convection zone. It is thought that where the field has grown sufficiently strong it becomes unstable and magnetic flux tubes start to rise towards the surface due to the stratification of the convection zone. Emerging at the photosphere, a magnetic flux tube forms an active region. In order to be compatible with observed properties of active regions (emergence latitude, orientation, asymmetry) the field strength at the bottom of the convection zone must be in the range of 3–10 T. This puts a tight constraint on the operation of the dynamo since the associated magnetic energy density is 1–2 orders of magnitude larger than the kinetic energy density of the convective motions.

Bibliography

Grossmann-Doerth U, Schüssler M and Steiner O 1998 Convective intensification of solar surface magnetic fields: Results of numerical experiments *Astron. Astrophys.* **337** 928–39

Priest E 1982 *Solar Magnetohydrodynamics* (Doredrecht: Reidel)

Schmelz J T and Brown J C (eds) 1992 *The Sun: A Laboratory for Astrophysics* (NATO ASI Series **C 373**) (Amsterdam: Kluwer)

Schüssler M and Schmidt W 1994 *Solar Magnetic Fields* (Cambridge: Cambridge University Press)

Solanki S K 1993 Small-scale solar magnetic fields: an overview *Space Sci. Rev.* **63** 1–188

Manfred Schüssler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Magnetic Field: Inference by Polarimetry

Most solar structure and phenomena occurring in the solar atmosphere arise from the presence of magnetic fields. This article describes techniques for remote sensing of the solar magnetic field, primarily via analysis of the polarization it imparts to the radiation emitted by the Sun.

The ZEEMAN EFFECT provides by far the most useful diagnostic of the solar magnetic field. Accordingly, most of this article is devoted to polarimetry of spectral lines sensitive to the Zeeman effect. As a result of recent advances both in technology and in techniques for interpretation of solar polarimetry, it is now possible to infer routinely not only the strength of the field in the solar photosphere but also its orientation. These vector magnetic field measurements provide powerful new diagnostics of solar phenomena; accordingly, much of this article is devoted to these techniques. A few different polarimetric diagnostics are described which provide different information about the magnetic field. This article does not embrace *in situ* measurements of the interplanetary magnetic field.

Importance of the magnetic field

Solar magnetic fields arise within the Sun, and also within its atmosphere, as a result of fluid motions interacting with the magnetized ionized gas (plasma) of which the Sun is composed. Because the gas is ionized, it is an excellent conductor of electricity. Electric currents are quickly established, which in turn induce magnetic fields, which in turn may then react back on the plasma, sustaining a dynamo (see DYNAMOS: SOLAR AND STELLAR). A glance through this volume will quickly convince the reader of the extraordinary, pervasive importance of magnetic fields in solar physics. They give rise to many phenomena such as SUNSPOTS, SOLAR FLARES, PROMINENCES, coronal structure and SOLAR CORONAL MASS EJECTIONS and the intrinsic variability of the Sun itself. Inference of this fundamental property of the solar plasma is central to advancement in our understanding of these phenomena.

Over many decades, qualitative measurements of solar magnetic fields have given us a glimpse of the evolution of the global SOLAR MAGNETIC FIELD as well as insight into solar activity. Recent advances in instrumentation and analysis techniques have placed the measurement of solar magnetic fields on a quantitative footing, so that it is now possible to provide strong observational guidance for the rapidly advancing theoretical understanding of solar processes.

Polarimetry and the Zeeman effect

The presence of a magnetic field in the solar atmosphere breaks the degeneracy of atomic levels, causing wavelength shifts of transitions between levels. The orientation of the magnetic field also establishes a reference frame for preferred interaction of these atomic states with polarization states of electromagnetic radiation. The strengths of the magnetic fields encountered in the solar atmosphere are such that bound–bound transitions may be characterized by the ZEEMAN EFFECT, the article on which in this volume describes its origin and its influence on the polarization of atomic transitions observed in the solar spectrum. Figure 1 illustrates the polarization of spectral lines in and around a sunspot. It is this polarization, along with the observed intensity, that permits one to infer both the strength and the orientation of the magnetic field.

Polarized light may be described by four quantities known as the Stokes 4-vector $\boldsymbol{S} = [I, Q, U, V]^T$ (the superscript ^{T} indicates transpose of the 4-vector). The components of the Stokes vector may be generated by a set of ideal polarization measurements: I is the intensity, Q is the difference of intensity between measurements with an ideal linear polarizer with polarization axis oriented vertically minus the measurement with the polarizer rotated by 90°, U is obtained in a similar fashion to Q except that the orientations of the polarizer are rotated by 45° with respect to those defining Q, and V is the difference measurement of ideal circular polarizers: right- minus left-handed. Figure 1 illustrates that the measures of linear polarization (Stokes Q, U) are fundamentally *symmetric* about the center of these absorption lines and Stokes V is fundamentally *antisymmetric*. Furthermore, the magnitude of the splitting is proportional to the strength of the magnetic field; linear polarization is present only if the field vector $\boldsymbol{B}$ is inclined to the line of sight, and circular polarization is present only if the field has a component along the line of sight. The complete polarization information represented by the Stokes 4-vector, when provided for the spectral variation of solar absorption lines, contains *almost* (except for an ambiguity of π in the azimuth angle of the field: see below) enough information for a complete description of both the strength and the orientation of the magnetic field in the solar atmosphere. Indeed, the magnetic field is a physical quantity that now may be measured remotely with considerably higher accuracy than many of the other thermodynamic properties of the solar atmosphere. A device that provides a map of the magnetic field at the solar surface is known as a magnetograph.

'Longitudinal' magnetometry

At wavelengths of visible radiation, even those atomic transitions with the greatest Zeeman splitting show atomic states fully separated in wavelength (i.e. greater than the thermal broadening of the lines) only for the strongest fields (>2000 G) present in the solar atmosphere. Note this splitting in spectra from the dark sunspot umbra in figure 1, where the field strength typically is in excess of 2000 G. Weaker fields produce splitting that is incomplete, and the degree of circular polarization (Stokes V) then approaches a proportionality with the wavelength splitting and hence with the field strength $|\boldsymbol{B}|$. Additionally, V depends on the angle of the field with respect to the line of sight γ. In the limit of weak magnetic fields, the peak amplitude of Stokes V is given

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

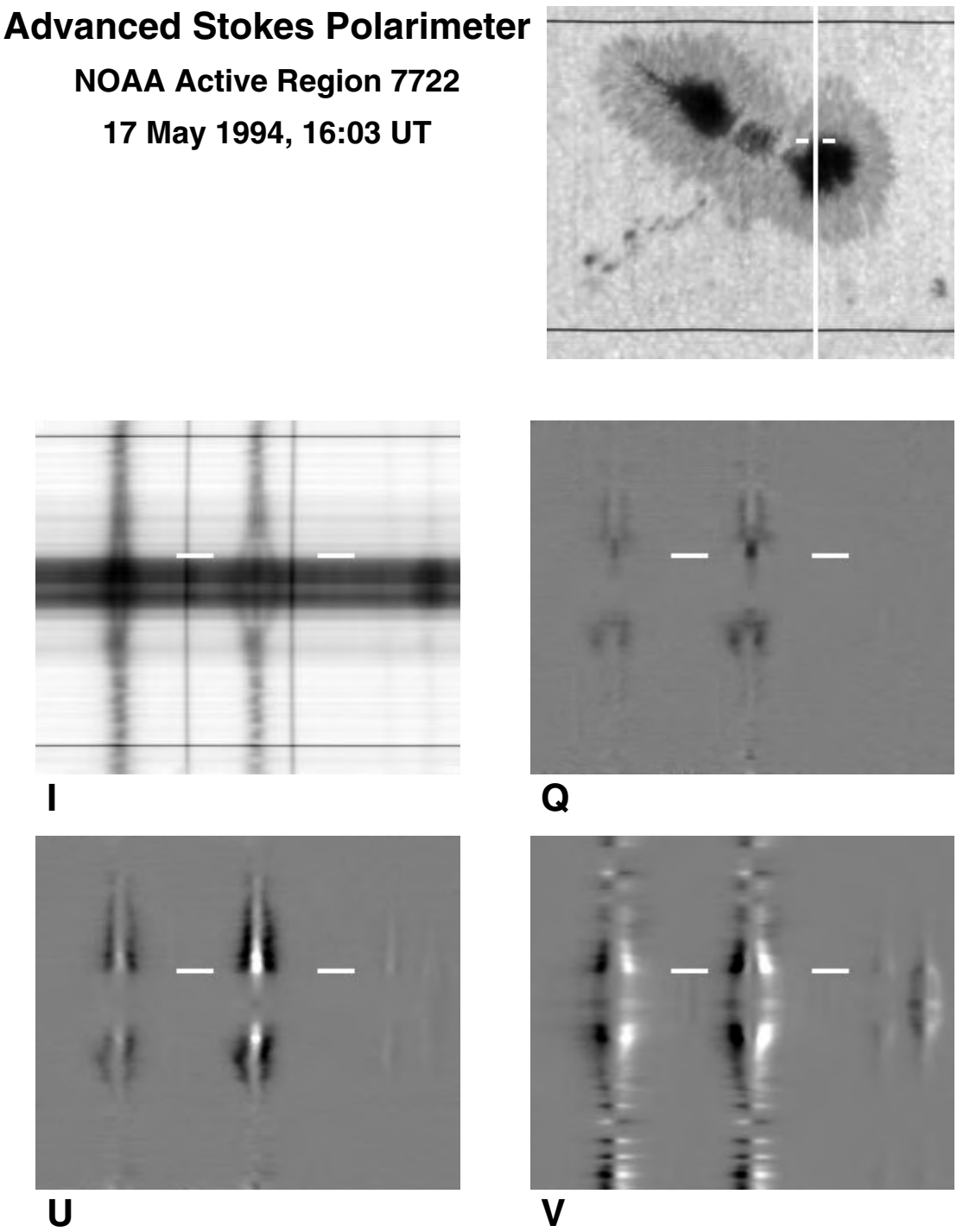


Figure 1. Polarization spectra of a sunspot. Spectral observations of the Stokes parameters I, Q, U, V characterizing the complete state of polarization of a narrow range of a sunspot spectrum near 630 nm are shown in the lower four panels. The upper-right panel shows the location of the spectrograph slit over the sunspot. The two broader spectrum lines due to absorption by neutral iron atoms form in the solar atmosphere and show both solar structure and polarization. The two narrow, polarization-free lines arise from oxygen molecules in the Earth's atmosphere. The horizontal lines near the top and bottom are present for calibration purposes. The white lines indicate the location of the spectral profiles shown in figure 4.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

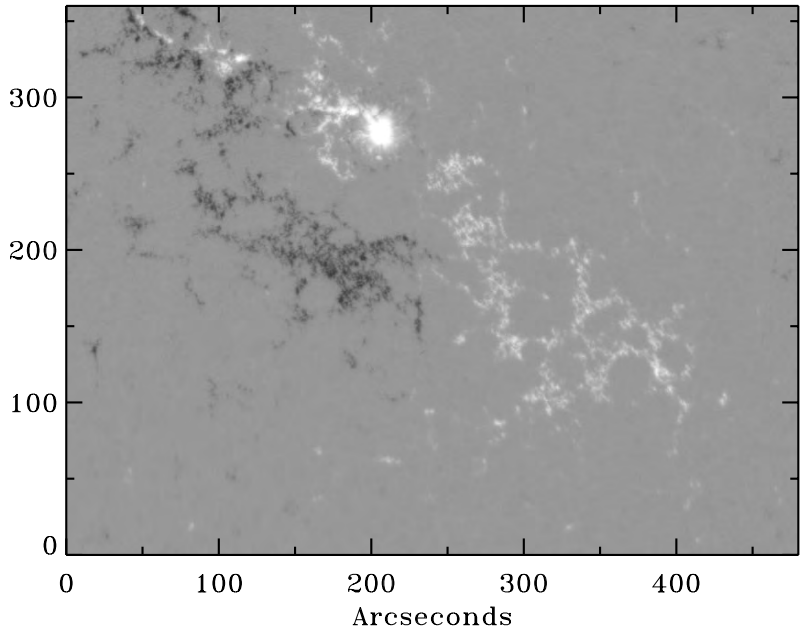


Figure 2. Longitudinal magnetogram from SOHO/MDI measured on 30 August 1997.

by $V_{max} \propto |\boldsymbol{B}| \cos\gamma$: the component of the field along the line of sight. Hence, magnetographs which measure the amplitude of the circular polarization signal are called longitudinal magnetographs. The measured difference between right- and left-handed circular polarization at one (or more) wavelengths in a Zeeman-sensitive spectral line (see figure 1) constitutes a magnetographic measurement. Systematic longitudinal magnetography of solar magnetic fields began with the Babcock magnetograph at the Mt Wilson Observatory in 1951, and it has continued and improved up to the present, as embodied by the superb data from the Michelson Doppler Imager (MDI) aboard the NASA/ESA Solar Heliospheric Observatory mission (SOHO), an example of which is shown in figure 2.

The great advantage of magnetograms like the one shown in figure 2 lies in their ability to reveal the temporal evolution of the fields. Operating continuously outside of the Earth's atmosphere, the MDI instrument is giving the first steady view of this evolution, ranging from large time and spatial scales corresponding to the disk passage of magnetic regions to the rapid evolution of small-scale flux elements.

Longitudinal magnetographic measurements provide approximate, nonetheless extremely useful, measures of the solar magnetic field. The magnetic field in the solar photosphere is highly structured at small scales. Outside of sunspot umbrae, magnetographic measurements are very far from resolving this structure. Longitudinal magnetographs in fact *do not* present us with a measure of the line-of-sight component of the field, but rather they provide some indicator of the magnetic flux $(\boldsymbol{B}\cdot\mathrm{d}\boldsymbol{A})$ within the spatial resolution element $\mathrm{d}\boldsymbol{A}$ of the instrument. Note also that $\mathrm{d}\boldsymbol{A}$ is defined in the plane of the sky, not the local normal to the solar surface. Calibration of magnetographic data to a quantitative scale of magnetic flux is often hindered both by incomplete resolution of the spectral line shapes and by approximate 'calibration constants' that do not account for variation in the structure of the solar atmosphere from point to point on the solar disk.

Vector magnetometry

The magnetic field is a vector quantity. Many solar processes involving the magnetic field will be best understood with knowledge of the full magnetic vector. There are numerous examples of the utility of vector magnetometry, of which the following is only one. Distinct modes of wave and oscillatory motions are possible in a magnetic field, and certain of these modes may transport mechanical energy of convection from below the surface into the upper layers where they heat the atmosphere (MAGNETOHYDRODYNAMIC WAVES). One may distinguish these modes by the direction of the motions with respect to the field vector, so it is important in this example, as in many other problems, to precisely establish the orientation of the field. When one needs to specify individual components of the magnetic vector, the field measurement becomes a highly quantitative endeavor. Vector magnetometry thus demands high precision from both the polarimetric measurements and the analysis.

The circular polarization from the Zeeman effect is commonly much greater than the linear polarization because, in the limit of small $|\boldsymbol{B}|$, the total linear polarization is approximated by $P_L = (Q^2 + U^2)^{1/2} \propto (|\boldsymbol{B}| \sin\gamma)^2$. Offsetting this difficulty is the great

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

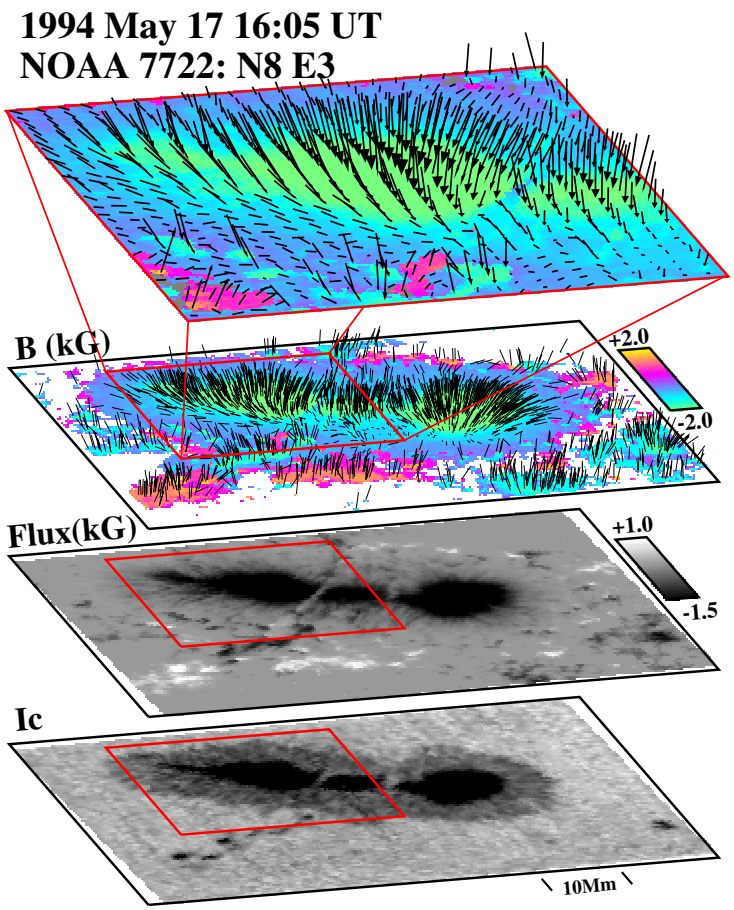


Figure 3. Vector field map. **This figure is reproduced as Color Plate 53.**

advantage that linear polarization is sensitive to the component of the field *transverse* to the line of sight. Advances in technology for both detectors and data acquisition, coupled with parallel developments in techniques for analysis of polarimetric data, have made possible routine vector magnetometry of the Sun.

Modern spectropolarimeters provide low-noise polarimetry from which it is possible to infer the components of the field vector in the solar atmosphere within a few per cent accuracy. This accuracy is achieved through a combination of the following attributes of the measurement:

- measurement of the complete, spectrally resolved absorption profiles of the solar spectrum lines;
- simultaneous measurements of two or more lines having similar thermodynamic sensitivity but differing sensitivity to the Zeeman effect;
- high signal-to-noise ratio of the measurements—typically $S/N = 10^3$ in the spectral continuum, or better;
- high-cadence polarization measurements to minimize image distortion from the Earth's atmosphere.

The measurements shown in figure 1 satisfy these requirements. The two broad solar absorption lines arise from neutral iron (Fe I) in the solar atmosphere. They are members of the same multiplet (816) of Fe I, and hence they respond similarly to point-by-point variations in temperature, pressure and ionizing radiation in the solar atmosphere. They differ principally in their sensitivity to the Zeeman effect: this difference is evident in the sunspot umbra where the splitting is noticeably smaller in the spectrum line on the left.

A sequence of such measurements, where the spectrograph slit is stepped in small increments in the direction perpendicular to its length, allow one, after involved analysis, to construct the vector field map shown in figure 3.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Analysis of polarization data

The fundamental data for vector magnetic field measurement are the spectral profiles of the Stokes parameters, as shown in figure 1 and also in figure 4 for the single spectrum highlighted in figure 1. Three quantities are needed to completely specify a vector. From the perspective of analysis of polarimetric data, a useful representation of the vector is the magnetic field strength (the length of the field vector) $|\boldsymbol{B}|$, the azimuth angle ϕ of the field vector in the plane of the sky relative to some coordinate frame (usually measured counterclockwise from the solar West direction) and the inclination angle γ of the field relative to the line of sight.

The field strength, or flux density $|\boldsymbol{B}|$, is normally presented in *cgs–emu* units of *gauss*. Even when the Zeeman splitting of the lines shown in figures 1 and 4 is less than the intrinsic (thermal + nonthermal) Doppler broadening of the solar plasma, subtle differences in the line shapes of the spectral profiles of the two lines permit an accurate measurement of the field strength when the S/N ratio of the observations is adequate. Thus, it is not necessary to rely solely on the amplitude of the ratio of Stokes V/I for a field strength measurement, as that quantity often leads to severe underestimate of $|\boldsymbol{B}|$ owing to the magnetized plasma filling only a small fraction of each spatial resolution element. Because the multiple-line analysis is independent of this magnetic fill fraction f, it is then possible to extract a crude measure of f from the data as well.

The linear polarization Stokes parameters Q, U are related to field azimuth ϕ through a simple rotation of the coordinate system for the linear polarization. Neglecting the magneto-optical effect[1], ϕ is determined by a simple rotation of the coordinate system to minimize the Stokes U profile from the solar lines. When precision linear polarization measurements are available, ϕ is well determined.

The ratio $P_L^{1/2}/V$ provides an approximate measure of the inclination angle γ. The differing sensitivity of P_L and Stokes V to $|\boldsymbol{B}|$ demands that γ be determined in concert with $|\boldsymbol{B}|$. Underestimates of $|\boldsymbol{B}|$ that result from ignoring the possibility of $f < 1$ may lead to spurious values of γ which are far more inclined to the line of sight than reality. This potential source of error is circumvented by the full line profile analysis described below.

At present, the most practical way to extract measures of the vector magnetic field from spectropolarimetric measurements is by numerical fits to the Stokes spectral profiles. The result of such a fit is indicated by the solid curves in figure 4. In this particular case, an analytic solution to the transfer of polarized radiation in a somewhat simple model solar atmosphere provides the basis for the fitting procedure. Of order nine parameters are needed to obtain reliable fits to the Stokes profiles: the vector field (three parameters), the fill fraction (one), the Doppler shift (one) and the thermodynamics of the atmosphere (four). This technique gives an accurate indicator of the field vector averaged along the line of sight within the range of heights in the solar atmosphere where the Zeeman-sensitive spectral lines form. Even though this analysis technique uses a simple model of the magnetized solar atmosphere, it usually gives a good fit to the Stokes profiles, and much experience with the technique shows that the results of this fitting are robust and accurate.

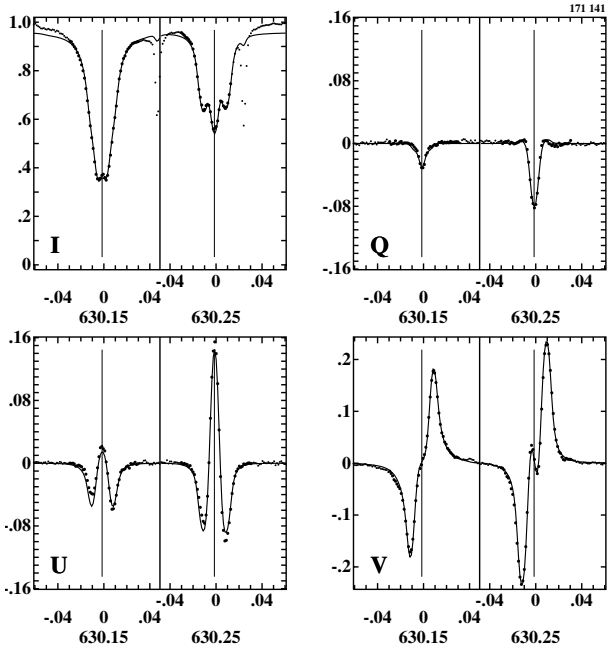


Figure 4. Polarization profiles and least-squares fits. A set of individual Stokes spectra from the location indicated in figure 1 are shown (dots), along with least-squares fits to the two solar spectrum lines (solid curves). The wavelength scale is indicated in nanometers. The extracted magnetic field parameters for this location in the sunspot are $|\boldsymbol{B}| = 2206$ G, $\phi = 140^\circ$, $\gamma = 134^\circ$.

The standard analysis technique allows only the temperature to vary along the line of sight; it does not permit estimates of line-of-sight *gradients* of the field, the line-of-sight velocity component or the other important thermodynamic properties. Such gradients may break the symmetries of the Stokes profile shapes about the central wavelength of the line, and the degree of asymmetry carries important information about the sign and magnitude of gradients. Recently, sophisticated fitting procedures have been developed which are based on more realistic models of the solar atmosphere, including gradients. The knowledge of local gradients of all components of $\boldsymbol{B}$ allows one to extract the electric current

[1] When the optical thickness of the spectrum lines becomes considerably larger than unity, Faraday rotation, also known as the magneto-optical effect, causes the direction of linear polarization to rotate as a function of wavelength near the centers of spectral lines. It also causes an anomalous 'reversal' in the core of the Stokes V profile to the right in figure 4. This effect may be accounted for in the least-squares fitting described below, and it actually helps to establish the optical depth of the lines in that analysis.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

density ($\boldsymbol{J} = (c/4\pi)\boldsymbol{\nabla} \times \boldsymbol{B}$) vector: an important diagnostic of the solar plasma. Furthermore, it provides a glimpse of the 3-dimensional structure of the atmosphere, albeit limited in extent to the few scale heights in the atmosphere over which the diagnostic spectral lines form.

The Zeeman effect carries no information on the *parity* of the component of the vector field transverse to the line of sight $\boldsymbol{B}_t$. That is, $\boldsymbol{B}_t$ is said to be *ambiguous* to within a multiple of π rad because the linear polarization introduced by the Zeeman effect responds identically to either $\boldsymbol{B}_t$ or $-\boldsymbol{B}_t$. Resolution of this ambiguity in the field azimuth angle ϕ requires additional information about the field or about its origin. Such information may take the form of an assumption. As examples, the azimuth may be chosen such that it is closest to that of the potential field defined uniquely by the locally vertical component of the field B_z, or one may choose the azimuth such that discontinuities of the field vector from one point in the magnetogram are minimized. A few techniques have been proposed to add different information to the problem, but none has yet seen practical use. In large measure, the resolution of the azimuth ambiguity remains an art, and, in special cases, errors of π may persist in the transverse field component of vector magnetograms.

Infrared field diagnostics

The Zeeman effect causes a splitting in the energy of atomic states proportional to $|\boldsymbol{B}|$. The splitting in wavelength is thus proportional to $|\boldsymbol{B}|\lambda^2$, but the Doppler broadening of an atomic transition is proportional only to λ. As a result, the Zeeman splitting relative to the intrinsic width of spectrum lines varies as $|\boldsymbol{B}|\lambda$. Spectrum lines forming in intrinsically weak solar magnetic fields of less than 1000 G do not show a separation of the Zeeman components at visible wavelengths, but the splitting of a line in the infrared (see INFRARED ASTRONOMY) can be nearly complete (i.e., comparable with or larger than the Doppler broadening; see figure 5).

Infrared measurements then offer a clear advantage for the study of weak-field regions in the solar atmosphere. The large splitting in the infrared relaxes the requirement for precision polarimetry. One drawback of solar magnetometry in the infrared is that the diffraction limit of the telescope is also proportional to λ. This limitation is a practical concern, particularly for measurements in the thermal infrared. Countering the reduced resolution are the improvements in atmospheric seeing and the polarizing properties of optics at infrared wavelengths. More importantly, when the Zeeman components of the line are fully or nearly split, magnetic elements unresolved within the observing element which differ only in $|\boldsymbol{B}|$ are easily identified based on their differing Zeeman splitting. Detailed analysis of the separate components is possible even without resolving them spatially. Furthermore, the size, efficiency and availability of infrared array detectors has increased in recent years. Infrared vector magnetometry has become progressively more useful as it provides unique tools to study weak solar magnetic fields.

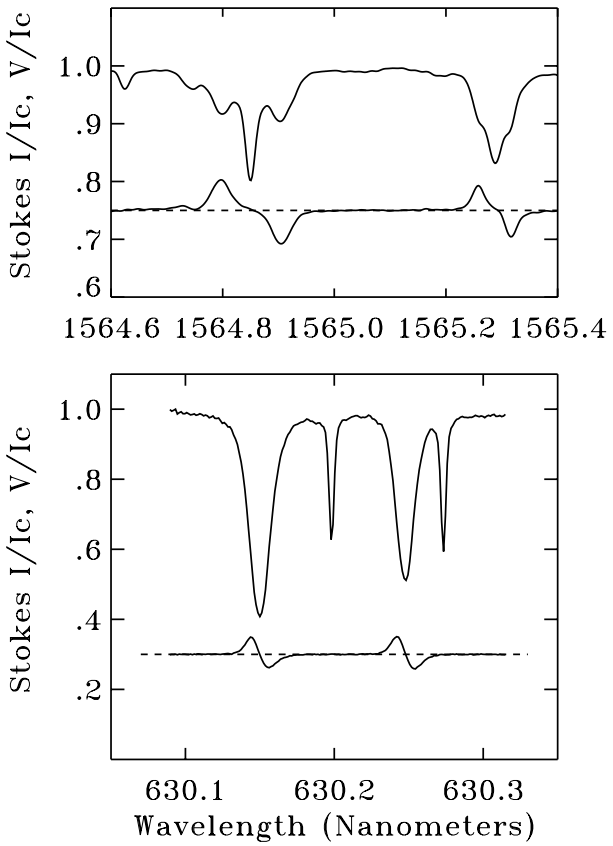


Figure 5. Comparison of Zeeman splitting in visible and infrared. The bottom panel shows Stokes I (top) and V (bottom) spectral profile shapes for a typical plage region as seen in the neutral iron line at 630 nm. The top panel shows analogous spectra, and correspondingly much larger Zeeman splitting, for a plage region observed in the neutral iron lines at 1565 nm (courtesy of I Rüedi). Dashed lines indicate the level of $V = 0$.

Prospects for magnetic field measurements above the photosphere

Most solar magnetograms sample magnetic fields in the photosphere—the thin layer of the solar atmosphere near the visible surface of the Sun—but many of the interesting physical challenges of solar physics arise from phenomena and processes in the higher layers: the chromosphere and corona. Thus, it is of great scientific interest to pursue analagous diagnostics of the chromospheric and coronal magnetic fields.

Several circumstances hinder efforts to measure fields above the photosphere. There are only a handful of spectral lines accessible to ground-based observations that form significantly above the photosphere, and these lines are not as sensitive to the Zeeman effect as those lines commonly used as photospheric field diagnostics. Owing to the larger nonthermal velocities, the chromospheric lines are intrinsically broader than photospheric lines and thus they produce smaller polarization signals. Most importantly, however, the magnetic fields of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

chromosphere and corona are commonly very much weaker than photospheric fields.

Although there is as yet no fully satisfactory means of measuring magnetic fields above the photosphere, routine longitudinal measurements are carried out daily for a line forming in the upper photosphere and low chromosphere (figure 6). These measurements graphically illustrate that the field has 'fanned out' considerably from the concentrated fields of the photosphere, mostly because of the combination of highly localized fields at the photospheric level and the rapidly decreasing density with height in the atmosphere.

Several research programs are aimed at improving chromospheric field measurements. For chromospheric lines forming in the visible and near-ultraviolet, it will be necessary to increase substantially the polarization sensitivity of the measurements over that commonly used for the photosphere. Infrared measurements would be ideal for this purpose, but few suitable spectral lines have yet been identified. The chromospheric neutral helium line at 1083 nm shows promise as a diagnostic of chromospheric fields.

Under most circumstances, the magnetic forces acting on the tenuous corona are far greater than the pressure forces of the gas itself. Magnetic fields thus play a dominant role in coronal structure and dynamics, and therefore any measurement of coronal magnetic fields is highly desirable. Because of the intrinsic weakness of coronal fields and the difficulty of observing the weak coronal emission outside of eclipse, there are few prospects for coronal field measurements. This situation may change in the future, as a few emission lines in the infrared could prove to be useful coronal field diagnostics: the Stokes V signature arising from the Zeeman effect should be measurable.

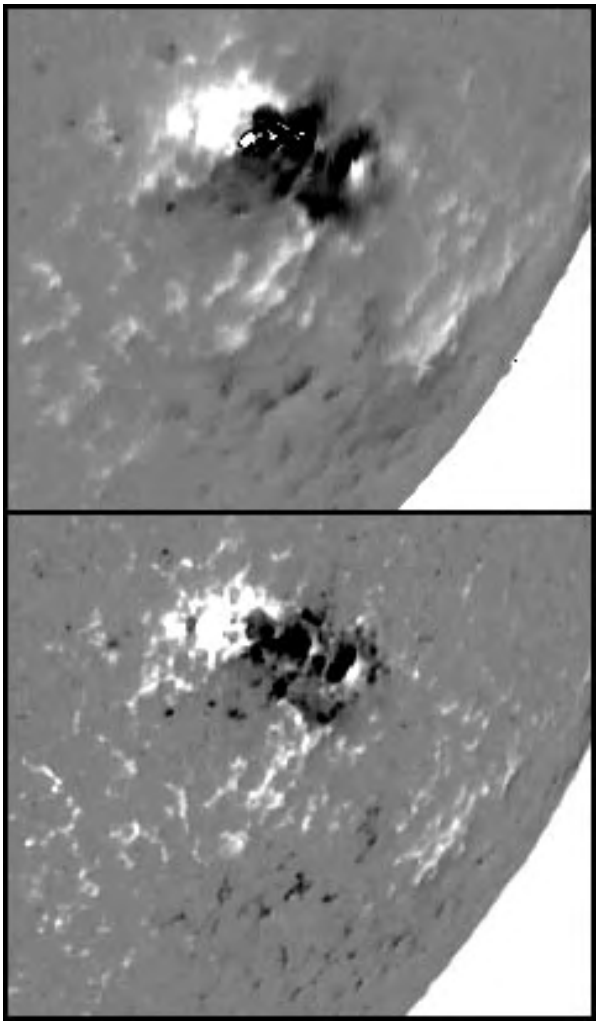


Figure 6. Expanding fields with height. The bottom image shows a photospheric longitudinal magnetogram of a typical active region near the southwest solar limb on 15 May 1998. The top image is a magnetogram of the same active region as observed in the singly ionized chromospheric calcium line at 854.2 nm, showing a considerably more diffuse appearance as a result of the expansion of the magnetic field with height in the solar atmosphere. Images supplied courtesy of J Harvey.

Other diagnostics of the solar magnetic field

The Zeeman effect introduces polarization into the solar spectrum. Another valuable diagnostic is the Hanle effect (see SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). When a weak magnetic field is present, the Hanle effect causes a reduction of the degree of linear polarization, accompanied by a rotation of the polarization direction, in spectrum lines which scatter an anisotropic radiation field. The degree of depolarization is a sensitive diagnostic of intrinsically weak magnetic fields in the solar atmosphere, and it has been used to reveal the presence of a weak, ubiquitous 'turbulent' field in the photosphere, presumably arising from the turbulent convection. Anisotropy of the radiation is most commonly encountered near the edge of the solar disk, owing to the variation of the brightness of the Sun with angle from the normal to the solar surface (limb darkening). Similar radiative processes acting in SOLAR PROMINENCES make the Hanle effect very useful for sensing the orientation of magnetic fields in these structures, thus helping to constrain theoretical models of the magnetic fields which suspend the cool, dense prominences in the hot, tenuous solar corona.

Anisotropy of the photospheric radiation also induces polarization at certain wavelengths corresponding to forbidden transitions of highly ionized atomic species, just as it does for transitions subject to the Hanle effect. Here, however, the Zeeman effect modifies the atomic level populations and thus alters the linear polarization of light scattered by these atoms. This emission line polarization may be used as a diagnostic of the orientation of the magnetic field, interpretation of the polarization is complicated by the presence of many magnetic structures along the line of sight through the optically thin solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

corona. If the usual circular polarization from the Zeeman effect can be measured at the same time, it will provide an indication of the strength of the field.

The microwave emission of the solar plasma provides another magnetic field diagnostic, based on rapid changes of the opacity of the chromosphere and low corona near the gyroresonance frequency for electrons precessing in the magnetic field. This technique can yield an indication of those temperatures of the upper atmosphere at which specific field strengths are encountered, but so far it seems to be limited to the very strong fields encountered over sunspots.

Future instrumentation

The magnetic field in the solar photosphere is structured on very small scales, and many processes of importance to heating and dynamics of the upper solar atmosphere occur principally at these small scales. It is therefore of utmost importance to describe the dynamical behavior of magnetic fields at these scales, accompanied by observations of the consequences in the upper layers at equally high angular resolution. The upcoming Japanese–USA mission SOLAR-B will provide continuous, high-resolution imaging polarimetry and spectroscopic polarimetry, along with imaging and spectroscopy of the upper layers. When this mission is launched in early 2004, it will provide the first space-based high-resolution quantitative vector field measurements. Efforts are also underway to develop space instrumentation of even higher resolution which should be able to actually resolve the tiny, intense photospheric flux elements and their accompanying dynamical interaction with the solar convective motions. In the meantime, one continues to improve the resolution of solar magnetometry from the ground using various image reconstruction and real-time image correction techniques.

The recent advances in techniques for vector magnetometry reveal the need for more comprehensive observations of this nature in order to better understand the workings of the solar cycle. To this end, the National Solar Observatory has begun development of instrumentation to provide frequent vector magnetograms of the entire solar disk, based on quantitative spectral measurements of the type shown in figure 1. This instrument is part of the SOLIS package for long-term, synoptic observations of the Sun.

Bibliography

Jefferies J, Lites B W and Skumanich A 1989 Transfer of line radiation in a magnetic field *Astrophys. J.* **343** 920–35

Lites B W and Low B C 1997 Flux emergence and prominences: a new scenario for 3-dimensional field geometry based on observations with the Advanced Stokes Polarimeter *Solar Physics* **174** 91–8

Skumanich A, Lites B W and Martinez Pillet V 1994 Vector spectropolarimetry with the Advanced Stokes Polarimeter (ASP) for quantitative solar magnetometry *Solar Surface Magnetism* ed R Rutten and C J Schrijver (Dordrecht: Kluwer) pp 99–125

Stenflo J O 1994 *Solar Magnetic Fields* (Dordrecht: Kluwer)

Bruce Lites

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Magnetographs

The Sun's disk is covered with magnetic fields, not only the strong kilogauss concentrations of SUNSPOTS, but also as slightly weaker flux in PLAGES, the network, and probably as truly weaker flux everywhere (see also CHROMOSPHERE: NETWORK). These fields may be mapped using the Zeeman effect: a spectral line formed in the presence of a field splits into components, the separation of which is proportional to the field strength, and the components are polarized depending on the orientation of the field (see also SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). An instrument that measures magnetic fields using the Zeeman effect is called a magnetograph; it is a tool central to solar research.

History

A Sun without magnetism would be a featureless sphere and would exhibit no activity. That the Sun possesses surface magnetism had been suspected already for two centuries in 1900, since the structure of the corona at a total eclipse mimics the patterns of iron filings on a paper over a bar magnet (figure 1). In 1896 Pieter ZEEMAN discovered in the laboratory the effect that bears his name. Soon after, in 1908, George Ellery HALE demonstrated Zeeman splitting in the spectrum of sunspots. His observations were both visual and by photography (figure 2). This was the first cosmic magnetism ever detected. For the next 3 decades the Sun's 'general field' outside of sunspots was pursued without any being found.

Finally, in 1953, Horace BABCOCK at the MOUNT WILSON OBSERVATORY, taking advantage of WWII technology, invented the solar magnetograph (although credit should also be given to G Thiessen and K O Kiepenheuer for earlier developments). Using a large spectrograph, Babcock arranged for a double exit slit to bisect a Zeeman-sensitive spectral line. The light from the line wings was directed to a pair of photomultiplier tubes (PMTs). An electro-optic modulator (EOM) at the entrance slit, actually an image slicer, alternated the transmission of first right-circular and then left-circular incident light. Output from the two PMTs was connected to a difference amplifier which yielded a signal at the EOM switching frequency of 120 Hz. Even minute Zeeman shifts produced a clear signal. Although the splitting in the quiet Sun was a small fraction of the line width, the signal had a low noise owing to the high efficiency of the PMTs. Some polarization is introduced by oblique reflections at the telescope feed mirrors, but this unwanted component is common to both input channels and cancels at the difference amplifier.

Fields were mapped by slowly scanning in a raster pattern the entire solar disk. Spatial resolution was sacrificed, being only about 20 arcsec × 20 arcsec, in order to have enough light to detect weak fields the order of a gauss. Magnetism was found over the entire disk. What Babcock and other contemporary workers did not appreciate was that in reality the fields are strong, mostly at least 1000 G, but highly fragmented into sub-arcsec elements except in sunspots. A 1000 G element whose size is 0.5 arcsec × 0.5 arcsec will appear as 0.6 G at 20 arcsec × 20 arcsec resolution. Solar magnetographs based on Babcock's design were built at observatories world-wide.

Figure 1. Corona at total eclipse.

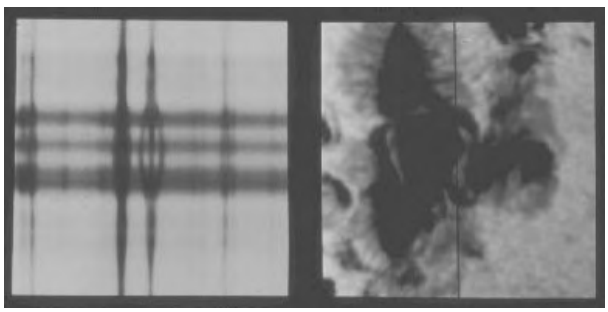


Figure 2. Zeeman splitting in the spectrum of a sunspot.

Today most magnetographs have a spatial resolution more like an arcsec and the fields they measure are accordingly much stronger. High resolution is achieved by using diode arrays or CHARGE-COUPLED DEVICES (CCDs) instead of PMTs. For space applications the spectrograph is replaced with an optical filter whose passband is comparable with line width. A few such instruments are described below.

Spectral lines for magnetographs

Depending on the height of the solar atmosphere to be probed, there are a few especially favorable lines for magnetometery. Zeeman splitting is proportional to $g\lambda^2$, where g is the Landé factor, which depends on the term values of the spectroscopic transition, and λ is the wavelength. Landé g ranges from 0 to 3, but is typically about 1. As spectral line width increases linearly with λ, the effective splitting is actually $g\lambda$. Obviously there is an advantage to the red and infrared wavelengths. Other considerations include the height of line origin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Some useful lines for magnetographs.

Species	λ (μm)	Landé g	$g\lambda$	Origin	Problems
Fe	0.5250	3.00	1.57	Photosphere	Temperature
Fe	0.6302	2.5	1.57	Photosphere	
Hα	0.6562	1	0.65	Chromosphere	Broad
Ni	0.6768	1.5	1.02	Photosphere	
Fe	0.8688	1.66	1.44	Photosphere	
Fe XIII	1.0747	1	1.07	Corona	
Fe	1.5648	3.00	4.69	Photosphere	
Ti	2.2310	2.5	5.6	Photosphere	Umbral only
Fe	4.1364	1.7	7.0	Photosphere	Blends
Mg	12.32	1	12.32	High photosphere	No quiet Sun

(photosphere, chromosphere or corona), line strengths and telluric or other blends.

Babcock chose Fe 0.5250 μm which has $g = 3$ and is unblended. Also, its wavelength is well matched to the maximum of PMT response. Because this line arises from an excitation state near ground level, it is somewhat temperature sensitive and weakens in plages which are hotter than the average quiet Sun. Less affected by temperature is Fe 0.6302 μm, $g = 2.5$. The availability of InSb near-IR arrays makes Fe 1.5648 μm, $g = 3$, a winner. By a wide margin Mg 12.32 μm, $g = 1$, is the best, but this line is found only in active regions and near the Sun's limb; it is absent on the quiet disk. Table 1 lists a few useful lines.

Modern magnetographs

The Kitt Peak vacuum system

This is basically a Babcock-type magnetograph except that the detector is a 512 × 242 CCD chip. Anamorphic optics demagnify the 10.7 m Littrow spectrograph focal plane by 1/6 in the dispersion (wavelength) direction and 1/18 in the spatial. The 242 pixel width of the CCD means that information on line shape and intensity can be recorded for some observations. Circular polarization modulation is done with a liquid crystal device. A full disk observation using Fe 0.8688 μm with a 1 arc-sec pixel size requires about 45 min. This is done daily and the results are distributed around the world by the US Government Space Environment Laboratory in Boulder, Colorado. Figure 3 is an example of an active Sun magnetogram; figure 4 shows a quiet Sun magnetogram (see also KITT PEAK NATIONAL OBSERVATORY).

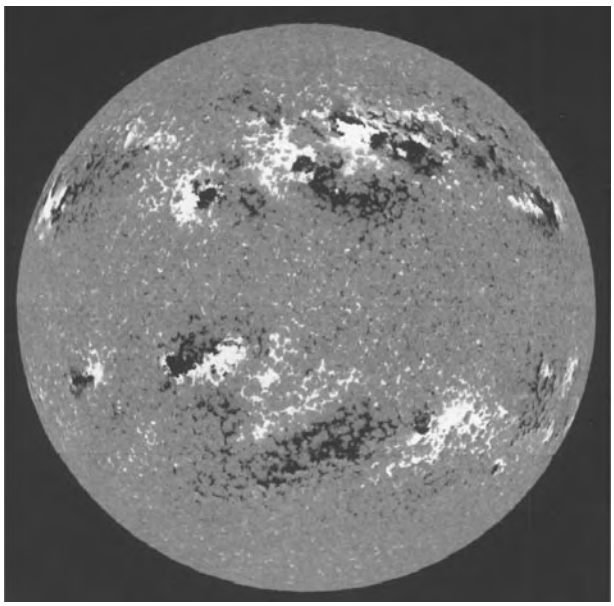


Figure 3. Active Sun magnetogram.

The Marshall Space Flight Center vector magnetograph

Fields directed toward or away from the observer are called longitudinal and give rise to oppositely circularly polarized light in the line wings. This is easy to detect by the means described above. Transverse fields result in an undisplaced linearly polarized line and split components, which are both linearly polarized in the same direction and at right-angles to the central part. Unless the splitting is comparable with line width, as in a sunspot umbra, there will be no 'difference' signal as in the longitudinal case and the field is difficult to measure and separate from any telescope polarization. Nevertheless, many instruments have been constructed that propose to measure vector fields (see Bray and Loughhead (1965) for examples).

One vector magnetograph that has seen much use is that at MSFC. This uses a beautifully built Zeiss birefringent filter of 0.125 Å bandwidth centered at 0.5250 μm. Modulation is by a combination of rotating quartz waveplates and a potassium dideuterium phosphate (KD*P) EOM. To avoid telescope polarization the instrument points directly at the Sun without any oblique reflections from feed mirrors. Sensitivity is about 4 G for longitudinal fields and 60 G for transverse. The main use of this system has been to try to understand the role of transverse fields in the production of solar flares (see also MARSHALL SPACE FLIGHT CENTER).

The Michelson Doppler Imager magnetograph on the satellite SOHO

The MDI instrument has a tunable Michelson interferometer with a 0.094 Å bandpass that can be tuned across the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

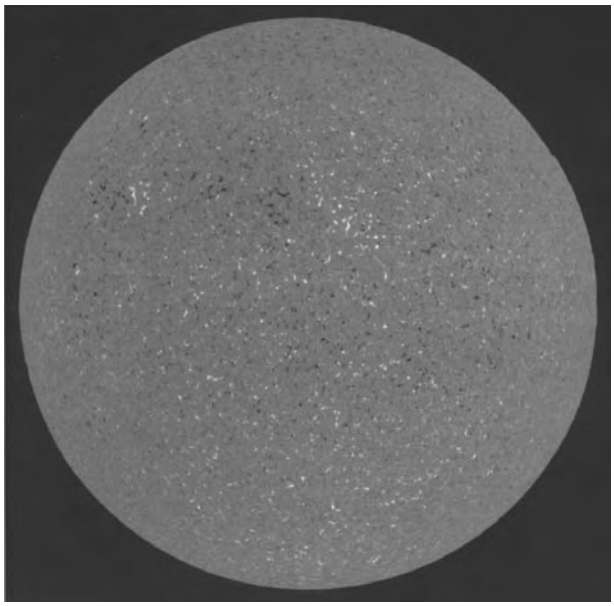


Figure 4. Quiet Sun magnetogram.

Ni 0.6768 μm, $g = 1.5$ line. Rotating wave plates are satisfactory in the absence of (atmospheric) seeing. Also many magnetograms can be averaged to reduce noise without a loss of resolution for the same reason (see also SOHO).

The vector spectromagnetograph for SOLIS

Synoptic Optical Long-term Investigations of the Sun (SOLIS) represents a next generation program to monitor the Sun. At the heart of this telescope cluster of instruments is a vector magnetograph. Using a ferroelectric liquid crystal as the polarization modulator, a 2048 $\times$ 1 pixel (1 arcsec) CCD detector, and using the favorable Fe 0.6302 μm line, full-disk vectorgrams are the end product. The aim is not one but several full disk observations per day. Much of the system is automated and may be remotely controlled.

Bibliography

Babcock H W 1953 *Astrophys. J.* **118** 387

Bray R J and Loughhead R E 1965 *Sunspots* (New York: Wiley)

Jones H P, Duvall T L Jr, Harvey J W, Mahaffey C T, Schwitters J D and Simmons J E 1992 The NASA/NSO spectromagnetograph *Sol. Phys.* **139** 211

Scherrer P H and the MDI team 1995 The Solar Oscillations Investigation–Michelson Doppler Imager *Sol. Phys.* **162** 129

William Livingston

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Maximum Mission (SMM)

NASA satellite designed to study solar flares during the peak of the solar cycle. Launched in February 1980. Carried seven instruments to record flares in visible, x-ray, ultraviolet and gamma ray wavelengths. Also measured the solar constant. Discovered x-rays originating from the bases of loop structures in flares. Failed after nine months, but was repaired by Shuttle astronauts in 1984. Re-entered the atmosphere in 1989.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere

The photosphere (Gr $\phi\omega\varsigma$, *light*) of the Sun are those layers in its atmosphere that one observes in white light (broad band visible light). An image in such light (figure 1) shows a sharp edge, as if the light were emitted from a single layer, the 'visible surface' of the Sun. Since stars are gaseous, they do not have sharp surfaces, however. Moreover, the radiation at different wavelengths is generally emitted from different heights in the atmosphere of a star. A more precise and generally applicable definition of the term 'photosphere' is often needed. The observed photons are produced in the atmosphere, which also absorbs these same photons. The probability that a photon, newly produced at some depth in the atmosphere, escapes from the star depends on the optical depth. If a_λ is the extinction coefficient (cm^{-1}) (see SOLAR SPECTROSCOPY AND DIAGNOSTICS) at a wavelength λ, and if z is a depth coordinate in the star such that $z \to -\infty$ outside the star, then the optical depth at this geometrical depth z is defined as

$$\tau_\lambda = \int_{-\infty}^{z_P} a_\lambda \, \mathrm{d}z. \tag{1}$$

The probability that a photon created at depth τ escapes from the star without being absorbed or scattered is $\mathrm{e}^{-\tau}$. In the deeper layers where τ is large, this probability is small: the photons are trapped, diffusing outward by repeated scattering or absorption and re-emission processes. These regions are not directly observable. From higher layers where the optical depth is small, photons can escape freely, but the atmosphere is also tenuous there, and produces only few photons. A common definition of the photosphere at wavelength λ is now the layer in the atmosphere from which the bulk of the observed photons are emitted, roughly from $\tau_\lambda = 0.1$ to $\tau_\lambda = 3$. Since the optical depth is a function of the wavelength of the photons, photons of different energy each escape from their own photosphere. For this reason, one often speaks of a visible photosphere, an infrared photosphere, a radio photosphere, etc. If the term photosphere is used without reference to a particular wavelength, the wavelength range is implied in which the bulk of the energy is radiated. The Sun emits its energy mostly in the visible range of the electromagnetic spectrum; a nominal photospheric surface is sometimes defined for the Sun as the surface $\tau_{500} = 2/3$, i.e. at a wavelength of 500 nm (green light).

For stars, the photospheric surface so defined is roughly spherical, but the definition allows it to be corrugated (for example by GRANULATION, below) or deformed (for example flattened by rotation of the star).

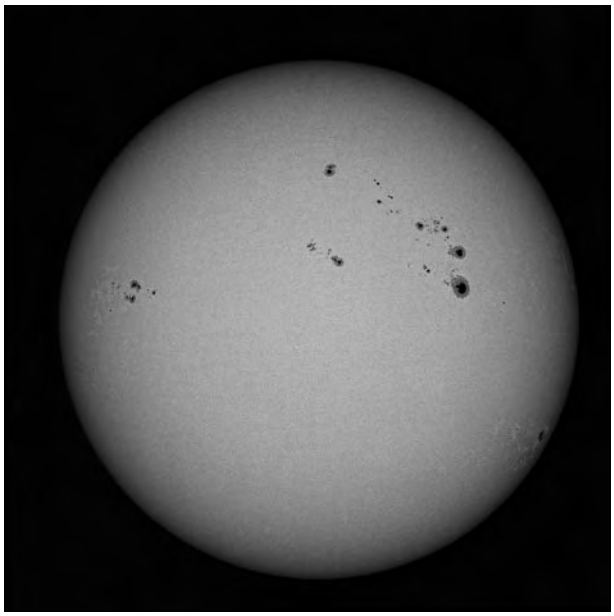


Figure 1. Image of the Sun in visible light, showing the photosphere with sunspots (dark) and plages (irregular brighter regions visible near the limb). Note general darkening towards the limb. Courtesy T Pettauer (Sonnenobservatorium Kanzelhöhe, Graz, Austria) and J A Bonet (Instituto de Astrofísica de Canarias, Tenerife, Spain).

Equilibrium structure

The type of radiation emitted reflects the temperature and density of the plasma at the photosphere as well as the physical processes producing the radiation. The bulk of the energy emitted by normal stars is in the UV to near-infrared range. The processes producing this radiation are known with high accuracy (see SOLAR SPECTROSCOPY), so a detailed comparison between theory and observed radiation is possible. The situation is complicated, however, by the presence of flows and magnetic fields in the atmospheres of stars. In particular the magnetic processes cannot as yet be computed from first principles, and their manifestations are noticeable in the entire electromagnetic spectrum. They even dominate the radiation in the radio and x-ray ranges. Nevertheless, important aspects of the physics of the photosphere can be understood by pretending that the atmosphere is static and without magnetic fields (see STELLAR ATMOSPHERES). It is then in *hydrostatic* and *radiative* equilibrium, and the structure of the atmosphere is determined by the basic parameters of the star: its luminosity L, radius R and mass M.

In hydrostatic equilibrium, the pressure drops with height in the atmosphere according to the barometric height formula, as in the Earth's atmosphere. The pressure P increases with depth z as $\mathrm{d}\ln P/\mathrm{d}\ln z = 1/H_P$, where the pressure scale height H_P is

$$H_P = \frac{\mathcal{R}T}{\mu g}. \tag{2}$$

Here $\mathcal{R}$ is the gas constant (8.134×10^7 erg K^{-1} mol^{-1}), μ the mean atomic weight per particle (1.28 for un-ionized gas of solar composition), g the acceleration of gravity at the surface,

$$g = GM/R^2 \tag{3}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and G is the constant of gravity.

If the energy radiated by the star is transported from the interior entirely by photons, the star is in radiative equilibrium. This is not the case in the Sun, since the energy is actually carried to the photosphere by convection, but for the observable layers it is still a fair approximation which is also of significant historical interest. The energy flux per unit surface area through the atmosphere is $F = L/4\pi R^2$. This flux has to be carried by photons diffusing outward; hence the density of photons has to decrease outward. Where photons are created and destroyed rapidly, the temperature of the gas is in equilibrum with the radiation (local thermodynamic equilibrium or LTE). Under this assumption, a decrease of the photon density requires a decrease of the temperature. The photosphere is cooled by radiation into the 3 K vacuum of space, and heat flows towards it from the hotter interior in the form of photons. A simple but useful estimate of the temperature variation through the atmosphere is obtained by assuming the extinction coefficient to be independent of photon energy (setting it equal to the value at the mean energy of the emitted photons, for example; this is called a grey atmosphere). Then the temperature T varies with the optical depth τ as:

$$\sigma T^4(\tau) \approx \tfrac{3}{4} F(\tau + \tfrac{2}{3}). \tag{4}$$

Here $\sigma = 5.669 \times 10^{-5}$ erg s^{-1} cm^{-2} K^{-4} is the radiation constant appearing in Stefan–Boltzmann's law, which says that a black body of temperature T radiates an energy flux $F = \sigma T^4$ per unit area of its surface. At $\tau = 2/3$, the temperature in this simple model is equal to the so-called effective temperature of the star, about 5780 K for the Sun.

In reality the extinction is not independent of photon energy. At wavelengths where it is low, one looks deeper into the atmosphere, where the temperature is higher, than at wavelengths where the extinction is high. If the dominant radiation processes are absorption and emission of photons (rather than scattering), the radiation generated in the atmosphere is close to that of a black body (i.e. the Planck function at the temperature of the plasma). Since the temperature in the atmosphere depends on depth, the temperature seen at each wavelength is different. This reflects itself in the spectral energy distribution of the star (the wavelength dependence of the emitted radiation), which can differ strongly from a black-body spectrum. This effect is responsible, in part, for the appearance of dark lines in the spectrum of a star at the wavelengths of atomic transitions. Through these deviations from a black-body spectrum, the spectrum can be used to make inferences about the composition and the thermal structure of the atmosphere.

Although the photosphere extends over a depth range in the atmosphere of a few hundred km, this is only a tiny fraction of the solar radius ($R_\odot \approx 700\,000$ km, see SUN: BASIC PROPERTIES). Even when seen with the best existing telescopes, which reach angular resolutions of about 0.25 arcsec (corresponding to 180 km), the Sun (figure 1) has a sharp edge. This sharpness is related to the temperature T and the acceleration of gravity g at the surface. Neglecting for the present purpose the deviations from hydrostatic equilibrium caused by oscillations (see HELIOSEISMIC OBSERVATIONS) and convection, we can apply the barometric height formula (2). This gives $H_\mathrm{p} \sim 150$ km for the scale height of the photosphere, only 1/5000 of the solar radius, and too small (0″2) to be resolved with existing telescopes. That this is indeed the scale height of the photosphere can be verified, however, during a solar eclipse. The moon moves about 0.5 arsec s^{-1} across the sky. Projected at the distance of the Sun, the lunar limb therefore occults the atmosphere at a rate of 300 km s^{-1}, or about 2 scale heights s^{-1}. By measuring the brightness variation in the last second before totality of the eclipse, one can accurately measure the extent of the solar atmosphere.

With the isothermal sound speed $c_\mathrm{i} = (\mathcal{R}T/\mu)^{1/2}$ (about 6 km s^{-1} for the solar photosphere), the ratio of pressure scale height to stellar radius can also be written as

$$H_\mathrm{p}/R = c_\mathrm{i}^2/v_\mathrm{e}^2 \tag{5}$$

where v_e is the escape speed from the surface, $v_\mathrm{e} = (GM/R)^{1/2}$, about 440 km s^{-1} for the Sun. For red supergiants such as Betelgeuse or Mira, which are about 200 times larger than the Sun but have similar masses and surface temperatures, this ratio is about 0.1. Such stars will therefore look much fuzzier when seen close up (techniques for such close-ups are becoming available through OPTICAL INTERFEROMETERS).

Determination of the temperature structure

Moving up into the atmosphere, the temperature declines from a value around 6400 K at $\tau_{500} = 1$ to around 4100 K at a height of 500 km ($\tau_{500} \sim 10^{-3}$). Above this height, the so-called temperature minimum, the average temperature starts inceasing again owing to the mechanical heating processes that produce the CHROMOSPHERE and CORONA. This variation of temperature with height in the photosphere can be deduced directly from the center-to-limb variation and from the strengths of *spectral lines* (see SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION; THE LIMB EFFECT AND GRAVITATIONAL REDSHIFT). An image of the Sun in the visible (figure 1) is noticeably darker towards the limb; the degree of this darkening depends on wavelength. For parts of the surface observed near the limb, the optical depth between the observer and a given level in the atmosphere is larger than for parts seen near the center of the disk, since the line of sight passes through a larger amount of mass. Near the limb, one therefore receives radiation from somewhat higher layers in the atmosphere. A limb darkening results if the temperature decreases with height at the observing wavelength, a limb brightening if the temperature increases. This effect is one of the classical ways for deriving empirical models for the solar atmosphere.

The second, equally classical, method can also be used for stars, where the disk cannot be resolved. It

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

uses the shape and strength of spectral lines. Quiet stellar atmospheres show spectral lines in *absorption*, i.e. dark, because the opacity in a spectral line is higher than in the nearby continuum. The known variation of the opacity with wavelength through the profile of the line can be used to invert the observed profile into a variation of (brightness) temperature with height, assuming that the line is formed under conditions of LTE. Such direct inversion of limb darkening and line profiles is not used in practice any more, since forward modeling by *spectral synthesis* (see SOLAR ATMOSPHERE: EMPIRICAL MODELS and SOLAR SPECTROSCOPY: VISIBLE EMISSION) is necessary to properly account for the radiative transfer process, in particular deviations from LTE.

Since the solar surface is inhomogeneous and not quite static, there are limits to what can be deduced from average limb darkening and line profiles. For example, line profiles are strongly affected by the Doppler shifts due to the convective flows. These distortions are an important test and diagnostic of the radiation–hydrodynamic simulations of granulation (see SOLAR INTERIOR: CONVECTION THEORY).

Physical conditions

With a temperature around 6400 K, the photospheric plasma is only partially ionized. At $\tau_{500} = 1$ there are less than 0.001 electrons per hydrogen atom. These few electrons come mostly from less abundant elements with low ionization potentials such as magnesium, while hydrogen and helium are almost completely atomic. In spite of this low degree of ionization, the plasma is almost infinitely conducting, for most purposes. The electron–hydrogen collision rate (the dominant rate at photospheric conditions) is about 10^{11} s^{-1}, and the mean free path of the electrons carrying the electric currents about 10^{-3} cm. With these values, the behavior of magnetic fields embedded in the plasma is described accurately by the ideal MHD approximation, at the length scales of interest (see MAGNETOHYDRODYNAMICS). In particular, the magnetic field and the gas are intimately tied to each other: the gas can move along the field lines without changing it, but in the direction perpendicular to the field either the gas is carried with the field, or the gas flow carries the field with it. One says that the field is *frozen in* to the gas.

The degree of ionization increases very steeply with depth. At a depth of only 50 km below the surface $\tau_{500} = 1$ of a granule, the temperature is 11 000 K. At this temperature some 10% of the hydrogen is ionized, 100 times more than at the surface. (In the cooler intergranular lanes, the increase is less strong.) With this increasing ionization, the opacity of the plasma increases by similar factors. This is mostly because the abundance of the H^- ion, the main source of continuum opacity in the solar photosphere, is proportional to the electron density. At the same time, the population of excited levels in all atomic species increases when partial ionization sets in, so that a great variety of atomic bound–free and bound–bound transitions starts adding to the opacity. Under these conditions, the horizontal mean free path of photons (in the visible continuum) is only 1 km, down by a factor of 50 from its value at $\tau = 1$.

Convection

The high opacity of a partially ionized plasma impedes the flow of heat through the outer envelope of a star like the Sun. To transport the heat emitted at the surface by radiation alone, through this opaque layer, a very steep temperature gradient would be needed. However, a stratification with a temperature gradient steeper than an adiabatic gradient (the 'adiabatic lapse rate') is unstable to convection (as in the earth's troposphere). The partially ionized envelopes of low-mass stars ($M < 2M_\odot$) are therefore *convective*. By overturning fluid motions, heat is carried bodily to the surface of the star, bypassing the inefficient transport by radiation, and is radiated away at the photosphere. The convective flows in the solar envelope are the cause of all its activity, including its magnetic fields (see SOLAR MAGNETIC FIELD), the corona, SOLAR FLARES, SUNSPOTS and its p-mode oscillations (see HELIOSEISMOLOGY).

The convection process can be observed in detail in the photosphere in the form of its granulation. A granule forms where the hot opaque gas from the envelope approaches the photosphere. As it is exposed to the interstellar vacuum it cools suddenly and becomes transparent. By cooling, the gas also becomes denser than its surroundings, and starts dropping down, a process observed in the form of intergranular lanes. In response to this disappearing cool gas, hot gas flows up from below to replace it. In this way the entire flow field making up the granulation is driven by the cooling at the photosphere. In fact, the flow field in the entire convection zone is probably a consequence mostly of this surface process. This is unlike the case of, for example, daytime convection in the troposphere, where thermals form by gas rising from the heated surface of the Earth. It is also different from laboratory convection where top and bottom surfaces of the container contribute equally to the production of convective flows. Heating from below plays such a minor role in the case of a stellar envelope because of the extreme stratification of density. At the base of the solar convection zone (at 200 000 km, or 30% of a solar radius below the surface), the density ρ is a million times higher than at the surface. The relative temperature excess $\Delta T/T$ in upward plumes generated at the base is also roughly a million times lower than at the surface. The acceleration of a convective flow, which is proportional to $g\Delta\rho/\rho \sim g\Delta T/T$, is therefore roughly a million times weaker at the base of the convection zone than at the surface.

In the granulation one thus observes the primary process of stellar envelope convection. The solar granulation can presently be reproduced with high fidelity with *ab initio* numerical simulations, which use only the equations of hydrodynamics, radiative transfer and the (large amount of) atomic physics of ionization and radiative processes. Other manifestations of convection

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

visible at the surface are flows on larger scales than granules: the mesogranulation, supergranulation and large scale flows (see SOLAR PHOTOSPHERE: MESOGRANULATION; SUPERGRANULATION).

Magnetic fields

Apart from the granulation, the main sources of structure seen in the photosphere are magnetic fields. In addition to the conspicuous sunspots, there are structures on smaller scales (see SOLAR PHOTOSPHERE: FILIGREE; FACULAE, and SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES), down to and probably well below to the resolution limit of current telescopes. The field strength in these structures can be measured through the Zeeman effect on spectral lines (see SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY). This is done most accurately in the infrared, where the Zeeman splitting is larger relative to the line width than in the visible. The field strengths found range from about 1000 G in the smallest structures to 3000 G in sunspots. At 1000 G, the magnetic pressure $P_{\rm m} = B^2/8\pi$ is 4×10^4 erg cm^{-3}, a significant fraction of the gas pressure P at $\tau = 1$, $P \approx 10^5$. Magnetic structures with such field strengths are therefore called 'strong fields'. (In addition to these strong fields, there is evidence for a less conspicuous weaker field component in the photosphere with strength of 300 G or less; see SOLAR PHOTOSPHERE INTRANETWORK AND TURBULENT MAGNETIC FIELDS.) The structures in this strong field component are sharply bounded: the normal photosphere in which they are embedded is only weakly magnetic, and the transition between the structure and its surroundings usually is sharper than can be resolved with current instruments.

Although all magnetic structures evolve in size and shape, the time scales for these changes are long compared with the time it takes a sound wave to cross their width. These structures are therefore approximately in pressure balance with their environment. This is possible only if the gas pressure inside the structures is less than in the surrounding photosphere, since the magnetic field exerts an additional pressure that has to be balanced. At the photospheric level, magnetic structures are thus confined by the gas pressure in their surroundings. The reduced internal pressure also makes the plasma more transparent, so that the photosphere is locally depressed inside the magnetic structures: they are dimples in the surface. This depression ranges from about 150 km in the smallest structures to 600 km in sunspots. In spots, the effect is large enough to cause an easily observable distortion (called *Wilson depression*, see figure 1 in SUNSPOTS: WILSON EFFECT) when they are seen near the solar limb. This effect also applies to small magnetic structures; they appear as little dimples in the surface. In the smallest elements of the magnetic field, the side walls of these dimples radiate extra heat from the hot subsurface layers, making them bright when seen near the limb of the Sun. Clusters of these elements make up the faculae.

Magnetic fields are most easily observed, through the Zeeman effect, in the photosphere. Field lines have no ends, however, and extend both above into the chromosphere and corona, and below into the convection zone. The magnetic patches seen as spots at the surface are only the intersections of this surface with three-dimensional magnetic structures. The rapid decrease of the gas pressure with height in the atmosphere means that, at a height of a only few hundred km above $\tau = 1$, the gas pressure is insufficient to confine a magnetic field of 1000 G. The magnetic structures adapt to this by expanding or 'fanning out' with height (figure 2). At a height of 500–1000 km, they have expanded so much that the fields of neighboring structures meet. Above this level (which depends on the average field strength, or 'degree of magnetic activity' of the photosphere), the entire atmosphere is magnetic. This situation applies to the higher atmospheric levels which one observes in the cores of strong spectral lines in the visible, such as H_α, as well as in all wavelengths longer than a few cm. These wavelengths probe the chromosphere and corona.

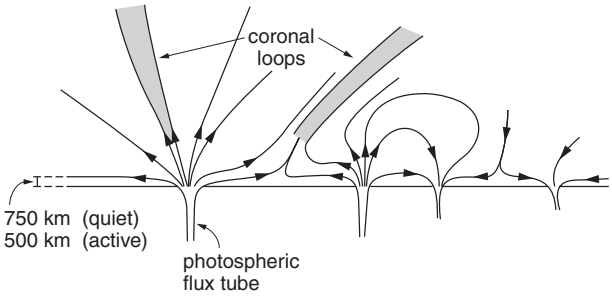


Figure 2. Vertical cross section through the solar atmosphere showing its magnetic structure (sketch). Below the surface, field lines are bundled into magnetic ropes with field-free convecting gas between; above the surface, the field lines fan out to occupy all available space. Sunspots are the locations where large magnetic ropes cross the surface.

The atmosphere as seen, for example, in H_α (figure 1 in CHROMOSPHERE: FIBRILS) looks very different from an image in the visible continuum: instead of granules the surface is covered with filamentary structures. These so-called fibrils trace field lines, much like iron filings around a bar magnet. In this magnetically dominated atmosphere the gas pressure is generally less than the magnetic pressure. Rather than being confined by the gas, the field here is in equilibrium with itself, with the magnetic pressure and tension forces balancing each other. The presence of the gas causes only small changes in the magnetic field: it is approximately force free. A familiar example of such a force-free field is the vacuum field (current-free field) around a bar magnet. A force-free field cannot exist entirely on its own: its internal stresses must be taken up at some boundary. In the case of the bar magnet, these stresses are taken up by the solid-state forces in the bar. In the case of the solar atmospheric magnetic field, the stress is taken up by the dense plasma in the convection zone into which the field lines are anchored (figure 2). In the atmospheric magnetic field, on the other hand, the motion of the gas is restricted. Flows along

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the field lines are unimpeded, but strong Lorentz forces limit displacements perpendicular to the field lines. The transition from the gas pressure dominated interior to the magnetically dominated atmosphere of the Sun happens to lie roughly at the photosphere. This is a coincidence that applies to most stars with convective envelopes but not to some other stars, such as strongly magnetic A stars and white dwarfs.

Although the convection zone is of course far from solid, it is able to take up the magnetic stresses acting in the atmosphere by its large inertia. This inertia is conveniently expressed in terms of the Alfvén speed $v_A = B/(4\pi\rho)^{1/2}$, where ρ is the gas density. This is the speed at which bends in the field, for example imparted by convective flows, propagate along the field lines. At 1000 G (a typical field strength in small-scale magnetic structures), the Alfvén speed at photospheric densities is of the order 5 km s^{-1}, while the typical convective velocity is around 2 km s^{-1}. At a depth of only 10 000 km, the density is 2500 times higher, and the Alfvén speed at the same field strength is only 100 m s^{-1}. This is of the order of the convective velocities, or smaller. At these high densities, the convective flows can displace field lines faster than an Alfvén wave can react to restore the shape of the field lines. The flow then effectively carries the field around as an inclusion. At the low gas densities in the photosphere, on the other hand, the field lines at the same field strength of 1000 G behave much more 'stiffly' and to a large extent resist bending by flows.

Since the gas density is also lower inside the photospheric magnetic structures (because of the lower pressure), they are *buoyant*. The lower density makes them behave like 'tethered balloons': they follow the motion of their feet in the deeper layers but tend to stay vertical in the atmosphere. Small structures, and those with weaker fields, are more easily bent and displaced by flows than large structures such as spots. The strongest flows in the photosphere are those due to the granulation. Accordingly, the small-scale field is found almost exclusively in the intergranular lanes and the vertices between granules, where the horizontal flows converge. Large structures such as spots behave much more independently. For example, they are observed to form by coagulation of many small fragments, each of which finds its way to the spot under construction independently of, and often opposite to, the photospheric flow field (see SUNSPOT EVOLUTION).

In this way, the magnetic fields seen at the photosphere and their evolution are only very partially governed by local processes at the photospheric level. In order to interpret their observed evolution, one has to envisage the entire three-dimensional structure. This immediately raises the question at what depth inside the Sun one should locate the 'roots' of a spot or a small scale structure. Current understanding of the solar magnetic cycle puts the source of the magnetic field near the base of the convection zone. In this picture, the photospheric magnetic structures are connected to the base of the convection zone by approximately vertical fields. At the base of the convection zone, there is believed to be a layer of horizontal field which is the source of the magnetic fields seen at the surface. The appearance of a new ACTIVE REGION at the surface is an event that has lifted a bundle of fields lines from this layer and has made them loop up through the photosphere, like a stitch of yarn pulled from a woollen sweater. After the eruption of the active region, in this view, the connection between the surface manifestations, such as spots, and source layer is by approximately vertical field lines, as sketched in SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES. Apart from disturbances due to the kicking and shoving around by flows in the convection zone, the evolution of the surface field might actually be an image of the evolution of its roots in the source layer.

The magnetic fields at the photosphere have a small effect on the brightness of the Sun (see SOLAR IRRADIANCE). Sunspots, being dark, cause a decrease in brightness, plage areas an increase. The areas outside plages and spots remain unaffected: the darkness of spots is not, for example, compensated by increased brightness of the nonmagnetic parts of the surface. This is understood theoretically as a consequence of the heat transport and heat storage properties of the concvection zone.

Bibliography

Foukal P V 1990 *Solar Astrophysics* (New York: Wiley)

Priest E R 1982 *Solar Magneto-hydrodynamics* (Dordrecht: Reidel)

Spruit H C and Roberts B 1983 Magnetic flux tubes on the sun *Nature* **304** 401

Stix M 1989 *The Sun, an Introduction* (Berlin: Springer)

Henk Spruit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Center-to-limb Variation

Many phenomena which are observable on the Sun's 'surface' change in characteristic ways between the center and LIMB of the apparent disk (e.g. the shapes of sunspots, the profiles of Fraunhofer lines, etc). The most obvious and easily noticeable (e.g. at sunrise or sunset, or on a good photograph) center-to-limb variation is the decrease of brightness from the disk center towards the limb. This so-called limb darkening is treated in the present article, in particular for wavelengths representing the continuous spectrum between the FRAUNHOFER LINES.

Limb darkening and angular distribution of radiation

Actually, limb darkening is a direct consequence of the fact that the intensity (I in W cm^{-2} ster^{-1}) of the radiation emerging from the photosphere depends on the angular distance (θ) from the direction of the solar radius (see figure 1). Ignoring any small- or large-scale structures of the photosphere (e.g. granulation, spots, faculae, etc), and a possible (but not yet clearly proved) dependence on heliographic latitude, we suppose here that the intensity I depends on r/R (or θ) only and exhibits at any time an exact rotational symmetry. In other words, in this article the intensity distribution across the solar disk is treated as an ideal smooth function $I = I(r/R)$ (see figure 2), which can be approximated by one or other mathematical formulae and which represents then the average for the whole disk and for an appropriate interval of time (e.g. near the minimum of solar activity).

The wavelength dependence of solar limb darkening

The decrease of intensity towards the Sun's limb depends on wavelength λ (unit nm). If special exceptions are ignored, the dependence follows the rule that the shorter the wavelength the stronger the decrease. This rule is illustrated in figure 2, which displays the center-to-limb variation of the intensity I_λ (unit W cm^{-2} ster^{-1} nm^{-1}) for two wavelengths, one in the UV (λ = 350 nm) and one in the infrared (λ = 950 nm). For both wavelengths, I_λ is shown as a function of r/R (limb-darkening 'curves') and of θ as well. As a consequence of the λ rule, also the color of the Sun's disk exhibits a center-to-limb variation: it reddens from the center towards the limb (see also SOLAR PHOTOSPHERE: THE LIMB EFFECT AND GRAVITATIONAL REDSHIFT).

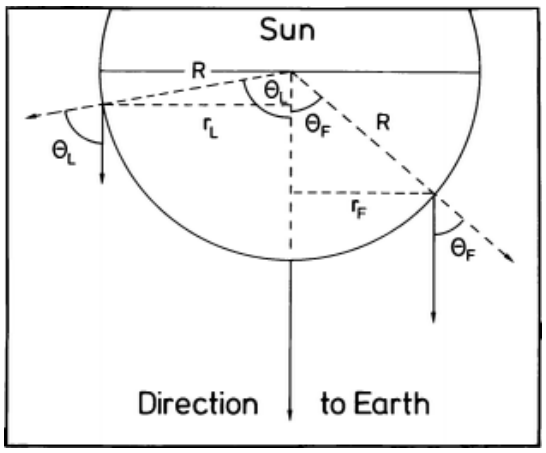


Figure 1. Geometric relations. R = angular radius of apparent solar disk as viewed from Earth (16'16" in early January, 15'44" in early July), r = angular distance of observed surface element from center of disk, θ = angle at solar 'surface' between the normal at position r and the direction towards the Earth. Two special cases are shown: suffix F refers to the position where the intensity equals the average disk intensity ($\theta_F \simeq 50°, r_F/R \simeq 0.76$), suffix L to a position about 14" from the limb ($\theta_L = 80°, r_L/R \simeq 0.985$). One notes easily that $r/R = \sin\theta$.

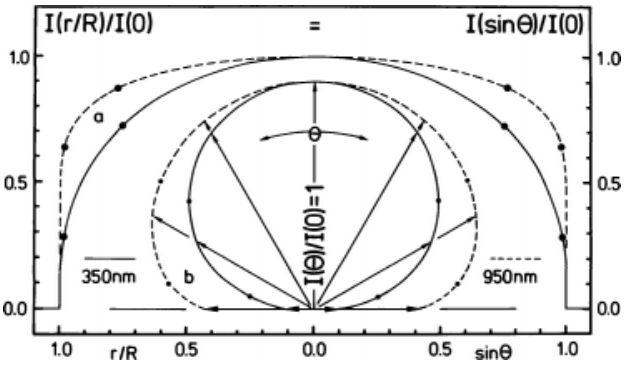


Figure 2. Solar limb darkening for λ = 350 and 950 nm from observations made with the McMath–Pierce Solar Telescope of Kitt Peak National Solar Observatory in 1975/76 and in 1986/87. (*a*) $I_\lambda = I_\lambda(r/R) = I_\lambda(\sin\theta)$; (*b*) $I_\lambda = I_\lambda(\theta)$ (for θ = 0°, 30°, 60°, 90°). In both cases the intensities are normalized to the disk-center intensity: $I_{350}(0) = 3.35$, $I_{950}(0) = 1.41$ W m^{-2} ster^{-1} nm^{-1}. The full circles refer to the positions specified in figure 1 ($\theta_F \simeq 50°$, $\theta_L = 80°$).

The reason for solar limb darkening and limb reddening

Usually, the intensity (I_λ) of the radiation from any homogeneous, hot, solid or fluid surface of uniform temperature T (unit K) does not depend on the angular distance (θ) from the direction perpendicular to the surface: the brightness of such a surface does not depend on the angle from which it is viewed. In this case, the intensity (I_λ) for any direction (θ) is simply given by the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

product $e_\lambda B_\lambda(T)$, where e_λ denotes the (dimensionless) emissivity of the surface material ($e_\lambda \leq 1$) and $B_\lambda(T)$ the intensity given by the Planck law of radiation; $B_\lambda(T)$ is a universal function with λ and T as the only essential parameters (unit W cm^{-2} ster^{-1} nm^{-1}); actually, $B_\lambda(T)$ also depends on the refractive index of (or on the velocity of light in) the medium (vacuum, air) where the radiation is observed or analysed.

But the radiation from the Sun is not emitted by a well-defined surface; instead, it emerges from a gaseous layer about 300 km deep, the so-called SOLAR PHOTOSPHERE. Inside this photosphere, almost all physical parameters depend on the depth t (unit m; counted from an arbitrary level 'zero'). For limb darkening two of these parameters are essential: the temperature T and the (continuous) absorption coefficient κ_λ (unit m^{-1}) (= fraction of radiation lost along a unit of length; roughly, the complement to the transparency). Note that κ_λ refers to the radiation between the Fraunhofer lines. Both parameters increase with increasing depth. κ_λ defines, together with the angle θ, the photospheric layers contributing the essential part of the emerging radiation, and the temperatures T define the amount of radiation emitted by these layers. The role of the angle θ may be inferred from figure 1: the larger θ, the smaller are the depths of the layers being relevant for the observed radiation. Noting that these depths vary (approximately) with cos θ, one realizes easily that limb darkening just reflects the decrease of temperature from the 'bottom' to 'top' of the photosphere. At the limb (r/R = 1, cos θ = 0), we receive of course only radiation from its highest layer (at level 'zero').

Recording the solar limb darkening

Measurements of solar limb darkening need a telescope yielding a solar image that is as large as possible (e.g. the McMath–Pierce Telescope of Kitt Peak National Solar Observatory; focal length $\simeq$87 m, diameter of solar image $\simeq$81 cm), a spectrometer to select the desired wavelength λ and a suitable detector (e.g. photomultiplier) plus electronic equipment. The observations are usually made by sampling the spectrometer output as the solar image drifts across the spectrometer entrance slit (telescope drive off, center of disk must pass the slit center!). The resulting 'drift curves', which then yield the intensity distribution along the solar diameter defined by diurnal motion, resemble of course the curves shown in figure 2(*a*); they yield the intensities of the Sun and of the (preceding and subsequent) skies in arbitrary units ('counts per second'), as a function of time (daily motion: 1 solar diameter in about 2.2 min).

However: the actual, individual drift curves differ from any 'true' mean distribution in two important aspects:

(a) The Sun's GRANULATION (and rare phenomena of solar activity) causes an appreciable 'noise' (standard deviation at 350 nm about 1.5%, at 950 nm 0.6%).
(b) Phenomena of the EARTH'S ATMOSPHERE (seeing, blurring, scattered light) and imperfections of the telescope (stray light) lead to systematic errors in the intensity signals, in particular near the Sun's limb. Also, imperfect electronics may cause small errors, but these can be avoided with special care (e.g. multiplier hysteresis).

For the reduction of the actual drift curves also the determination of the 'limb' positions is problematic. Even the true intensity cut-off at the limb is not perfectly sharp: the intensity drops exponentially within about 1". Thereby, the position of the steepest gradient (by definition the limb) depends on wavelength: at the local maximum of the continuous absorption coefficient near λ950 nm, the true solar radius is about 0.1" larger than at the minimum near 350 nm. Seeing and blurring not only cause considerably larger 'smearing' of the limb, but also lead to noticeable displacements of the 'limb inflection points'. Even for the averages of many drift curves these displacements don't cancel out completely: depending on both the size of the seeing (blurring) parameter and the shape of the drift curve, the solar radii are shortened systematically by amounts between about 0.05" (seeing parameter 1", $\lambda \simeq$ 950 nm) and more than 0.5" (3", $\lambda \simeq$ 350 nm).

Mathematical treatment of limb darkening observations

We don't discuss here the rather complex treatment of the intensity signals to account for the errors caused by the phenomena of the Earth's atmosphere and the instrumental imperfections. Supposing that the necessary corrections are made in one way or the other, the next and final step is then to fit a mathematical expression representing the limb darkening 'law' to the observed signals. Thereby one must realise that a theoretically reasoned, mathematical 'law' (comparable to, for example Planck's law of radiation, or the shape of a planetary orbit), does not exist *a priori*. Therefore, one has to find empirically a solution which fits the observations as well as possible, and, if possible, also fulfils special physical requirements.

In figure 3(*a*) the same two intensity distributions are plotted as in figure 2, but here versus μ = cos θ instead of r/R = sin θ. Both curves can be considered as being composed of two terms: a linear term $M(\mu) = m + (1 - m)\mu$ (note $m_{350} < m_{950}$), and a term $D(\mu)$ which represents the deviations from the linear relation. The characteristic behavior ($D(\mu) > 0$ except for $D(0) = D(1) = 0$; maxima between μ = 0.3 and 0.5) is found for all wavelengths

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

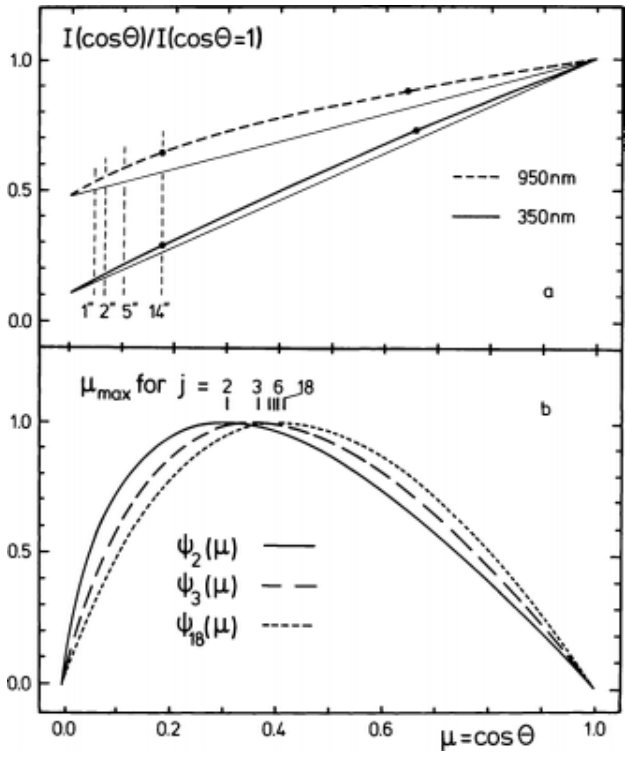


Figure 3. (*a*) Observed $I_\lambda(r/R)/I_\lambda(0)$ plotted versus $\mu = \cos\theta$ (heavy curves), from same data as in figure 2. Again, the full circles refer to the positions specified in figure 1. Vertical dashed lines indicate distances from the 'limb' of 1", 2", 5" and 14". Weak, straight lines indicate the linear relations $M(\mu) = m + (1-m)\mu$ defined by the intensities at the center ($\mu = 1$: $M = 1$) and 'limb' ($\mu = 0$: $M = m$) of the solar disk; thereby, the deviations $D(\mu)$ of the observed limb darkening from linearity are clearly noticeable. (*b*) Functions $\psi_j(\mu)$ which are the physically most reasonable functions to represent the deviations $D(\mu)$, shown for $j = 2$, 3 and 18; maximum positions are marked for $j = 2$ to 6 and 18. Note that maximum(ψ_j) = 1.00 for all j, and that μ(max) increases with increasing j from $\mu = 0.302$ for $j = 2$ to a limit near $\mu = 0.42$.

$\lambda > 330$ nm, while one finds for $\lambda < 320$ nm $D(\mu) \leq 0$, with shallow minima instead of lowly maxima. So, searching for the appropriate, mathematical relation which fits the observed limb-darkening curves, one must 'just' find a general relation which represents the terms $D(\mu)$.

The most reasonable relation approximates $D(\mu)$ by a series of very special (and more or less complex) mathematical functions, $\psi_j(\mu)$ ($j \geq 2$), which all have the same characteristics: $\psi_j(0) = \psi_j(1) = 0$, maximum ($\psi_j$) = 1.00. Three of these functions are shown in figure 3(*b*).

Accordingly, the most appropriate, empirical limb darkening 'law' is:

$$\frac{I_\lambda(\mu)}{I_\lambda(\mu=1)} = L_n(\mu) = c_0 + (1-c_0)\mu + \sum_{j=2}^{n} c_j \psi_j(\mu). \tag{1}$$

The meaning of the coefficients is clear: c_0 is a measure for the 'limb'-intensity and defines the linear term of the limb darkening function, and the coefficients c_j ($j \geq 2$) measure the maximum contributions of the functions $\psi_j(\mu)$ to the deviations $D(\mu)$. In practice it is sufficient to choose $n = 4$.

The functions $L_n(\mu)$ yield of course also the ratios between average disk intensity F_λ and disk center intensity $I_\lambda(\theta = 0)$, which are defined by equation (2):

$$F_\lambda / I_\lambda(\theta=0) = \frac{\int 2\pi r I_\lambda(r)\,\mathrm{d}r}{I_\lambda(0)\int 2\pi r\,\mathrm{d}r} = 2\int_0^{\pi/2} \frac{I_\lambda(\theta)}{I_\lambda(0)} \cos\theta \sin\theta\,\mathrm{d}\theta. \tag{2}$$

Specifying $n = 4$ and presenting just numerical values instead of the explicit, rather complex mathematical formula, the result following from relations (1) and (2) is:

$$\{F_\lambda/I_\lambda(0)\}_{L_4} = \tfrac{2}{3} + \tfrac{1}{3}(c_0 + 1.665c_2 + 1.764c_3 + 1.798c_4). \tag{3}$$

Now we recall that the intensity distribution across the Sun's disk depends on the vertical distribution of temperature (T) and (continuous) absorption coefficient (κ_λ) inside the Sun's photosphere. This dependency can be treated even exactly by using equation (4):

$$I_\lambda(\theta) = \{I_\lambda(\theta)/I_\lambda(0)\} I_\lambda(0) = \int_0^\infty B_\lambda(T(\tau_\lambda)) \exp(-\tau_\lambda/\mu)\,\mathrm{d}(\tau_\lambda/\mu). \tag{4}$$

Here, τ_λ is the optical depth of the layer with the geometrical depth t (m) below the level zero:

$$\tau_\lambda = \int_0^t \kappa_\lambda\,\mathrm{d}t. \tag{5}$$

Note that relation (4) describes the intensity $I_\lambda(\theta)$ of the observed radiation as the sum of all contributions from the 'visible' photospheric layers below the 'surface': implying local thermodynamic equilibrium, the contribution from the layer with temperature T is given by the intensity following from Planck's law of radiation, multiplied by a factor (<1) which accounts for the absorption along the light-path from that layer to the 'surface'; the length of this path is proportional to $1/\mu$ (further details are found in almost any relevant text-book; see, e.g. Minnaert (1954)).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In case of the limb darkening law presented by equation (1) it can be stated that it corresponds precisely to a well defined 'source function', $B_\lambda(\tau_\lambda)$:

$$B_\lambda(\tau_\lambda) = I_\lambda(0)\left\{k_0 + k_1\tau_\lambda + \sum_{j=2}^{n} k_j E_j(\tau_\lambda)\right\}. \tag{6}$$

where the E_js are the so-called exponential integral functions defined by

$$E_j(\tau) = \int_1^\infty [\exp(-t\tau)/t^j]\,dt \tag{7}$$

while the coefficients k_j are well defined functions of the coefficients c_j. Although the basic functions entering the solution ($\psi_j(\mu)$,$E_j(\tau)$) are rather complex (but present no problems in the PC era!), compared to other solutions they provide the important advantage that they don't lead to physically unrealistic structures in the limb darkening curves and—in particular—in the source function, which could be introduced by the natural scatter in the observations.

For example, if observed limb darkening curves are fitted by polynomials P_n in μ

$$I_\lambda(\mu)/I_\lambda(\mu = 1) = P_n(\mu) = \sum_{i=0}^{n} a_i \mu^i \tag{8}$$

as was usually done in the pre-computer era, then even slight errors in the coefficients a_i may cause large, unrealistic fluctuations in the corresponding polynomials representing then the source function:

$$B_\lambda(\tau_\lambda) = I_\lambda(0)\sum_{i=0}^{n} A_i \tau_\lambda^i \qquad \text{with } A_i = a_i/(i!). \tag{9}$$

In particular, in this case the source function also does not have the correct asymptotic behavior for both $\tau\to\infty$ and $\tau\to 0$. But observational scatter does not harm the resulting ratios of average to disk center intensity:

$$\{F_\lambda/I_\lambda(0)\}_{P_n} = 2\sum_{i=0}^{n} a_i/(i+2). \tag{10}$$

Wavelength dependence of the limb darkening coefficients (c_j)

Once any functions like those defined by equations (1) or (8) have been fitted to limb darkening observations made at many different wavelengths, the λ dependence of the coefficients (c_j,a_i) can be analysed. However, it seems that reliable results are obtained only if (a) the observations were made in a relatively short time interval (about 1 month; possibly near the minimum of solar activity), and (b) the data were fitted by curves defined by equation (1). Only very recently (1997) could it be shown (by means of data obtained in 1986/87 in two four-week observing seasons), that in this case the coefficients (c_j) of the limb darkening functions (L_4) and the corresponding function $\{F_\lambda/I_\lambda(0)\}_{L4}$ depend on λ in a very obvious, simple way:

$$c_0 = c_{00} + c_{01}\lambda^{-1} \tag{11a}$$

$$c_j = c_{j0} + c_{j1}\lambda^{-5} \qquad \text{for } j = 2,\ 3,\ \text{and } 4 \tag{11b}$$

$$F_\lambda/I_\lambda(0) = f_0 + f_1\lambda^{-1} + f_2\lambda^{-5} \tag{12}$$

with—according to relations (3) and (11*a*,*b*) —

$$f_0 = \tfrac{2}{3} + \tfrac{1}{3}(c_{00} + 1.665c_{20} + 1.764c_{30} + 1.798c_{40}) \tag{13}$$

$$f_1 = \tfrac{1}{3}c_{01} \tag{14}$$

$$f_2 = \tfrac{1}{3}(1.665c_{21} + 1.764c_{31} + 1.798c_{41}). \tag{15}$$

The good definition of these functions is demonstrated in figure 4, which exhibits the λ dependence of c_0 and of $F/I(0)$. Different sets of coefficients must of course be used for the spectral regions 'left' or 'right' from the Balmer 'jump' (at $\log\lambda \simeq 2.6$). Another obvious peculiarity is the 'bump' at $\log\lambda \simeq 3.2$ ($\lambda \simeq 1650$ nm); it is due to the minimum of H^- absorption between the spectral regions with bound–free and free–free transitions.

Possible variations of solar limb darkening

Usually, the coefficients of published limb darkening functions show—when plotted against λ—a considerable scatter, which prevented realistic fits by handy functions in the past. In this respect one should realize that there are no sets of limb darkening data which come from just one short-term 'snapshot'.

The usual observing procedure (wavelength after wavelength, day after day, some observing campaigns extending even over months or years) necessarily yields results which are influenced by all sorts of variable phenomena. In particular, in the case of the polynomials

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

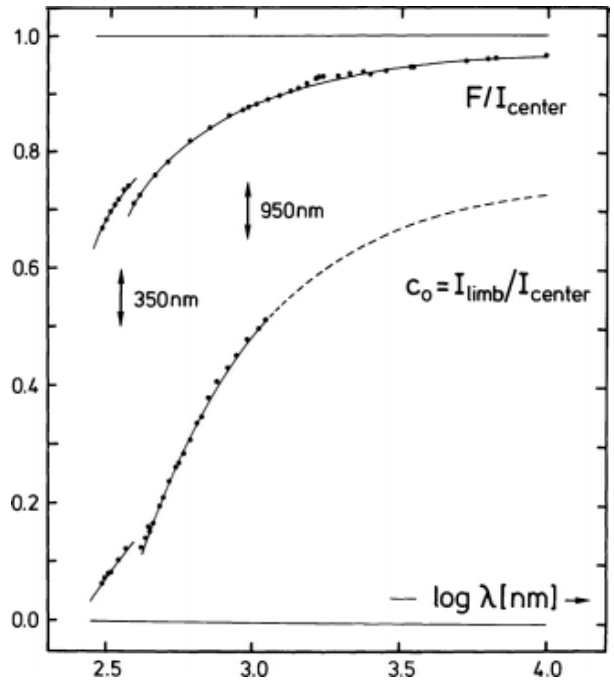


Figure 4. λ dependence of the coefficient c_0 and of the ratio $F/I(0)$ in the case of the limb darkening function L_4; c_0 from observations made in two 1 month seasons in 1986 and 1987; the smoothing functions are: $c_0 = 0.400-103\lambda^{-1}$ ($\log \lambda < 2.6$), and $c_0 = 0.756-268\lambda^{-1}$ ($\log \lambda > 2.6$, for $\log \lambda > 3.04$ extrapolated). $F/I(0)$ adopted from a compilation of 14 independent observations published from 1946 to 1983 ($\log \lambda \leq 3.38$), and from a recent paper published in 1996 (log;$\lambda > 3.38$); the smoothing functions are: $F/I(0) = 0.861-34\lambda^{-1}-18.10^{10}\lambda^{-5}$ ($\log \lambda < 2.58$) and $F/I(0) = 0.975-91\lambda^{-1}-24.10^{10}\lambda^{-5}$, ($\log \lambda > 2.58$). Arrows mark the two wavelengths with data in figures 2 and 3(*a*).

$P_n(\mu)$ (e.g. with $n = 5$), even minor variations in the true intensity distribution (granulation!) may cause large responses in the coefficients a_i. But different signs lead here to an almost complete compensation in the corresponding λ plots of $I(\mu)/I(0)$ and $F/I(0)$.

However, a significant scatter was also found in λ plots of the coefficients c_j of the functions L_3, which resulted from observations obtained over long periods of time (months, years). This is possibly a hint that even the 'average' solar limb darkening may be slightly variable in time, with time-scales larger than a month. The fact that also in this case the scatter of the coefficients c_j is not transferred to the λ plots of the corresponding ratios $F/I(0)$ would then indicate that limb darkening variations occur in such a way as to leave these ratios nearly constant.

Other hints for possible limb darkening variations come from special, systematic searches. But the question of whether the observed effects just reflect well-known characteristics of the solar photosphere (random fluctuations in distribution, size and brightness of the granules, variable structure of the supergranulation network, all sorts of oscillation patterns, intensity excesses of deficits connected to active regions), or do indicate occasional variations of the temperature distribution in the upper photosphere, cannot yet be answered definitively.

Bibliography

For general background see, e.g.,

Minnaert M 1954 The photosphere *The Sun* ed G Kuiper (Chicago: University of Chicago Press) pp 88–106
Gibson E G 1973 *The Quiet Sun (NASA Special Publication 303)* (Washington, DC: NASA) pp 93–106
Stix M 1989 *The Sun* (Berlin: Springer) pp 133–5

For numerical data see

Pierce K 2000 Limb darkening *Allen's Astrophysical Quantities* ed A N Cox (Berlin: Springer) pp 335–7

For detailed information on limb darkening functions, the constants involved and further literature, see

Neckel H 1996 *Sol. Phys.* **167** 9–23
Neckel H 1997 *Sol. Phys.* **171** 257–68

Heinz Neckel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Faculae

Photospheric faculae, often called white-light faculae, are best seen away from the center of the solar disk beginning at a heliocentric angle of about 60°. They appear as irregular bright patches whose contrast increases towards the solar LIMB, and are more extensive when they accompany sunspot groups. Near the limb they tend to appear as facular granules. The facular granules consist of aggregations of mostly vertical magnetic flux tubes.

Figure 1 shows half of a photometric image obtained on 1 July 1998 at a wavelength of 672.3 nm using a filter with a bandwidth of 10 nm. The limb darkening of the quiet sun has been removed to show more clearly the increasing contrast of the faculae as the limb is approached. The patchy or intermittent nature is obvious.

Faculae consist of magnetic flux tubes with a typical magnetic field strength of approximately 1.5 kG. Magnetograms of facular regions usually do not resolve these flux tubes and the observed magnetic field strengths depend on the filling factor and SEEING conditions at the time of observation.

The increasing continuum contrast of faculae with decreasing distance from the limb suggests an increasing temperature difference between faculae and the quiet sun. Models of facular flux tubes are unable to completely account for this steadily increasing contrast of faculae as the limb is approached. The contrast of faculae increases toward the limb for continuum wavelengths as well as for many absorption lines that are formed well below the temperature minimum. When viewed in the center of absorption lines, faculae are seen all across the solar disk in a photospheric network. Faculae seen near the limb in white light are morphologically identical to those seen in absorption lines. This behavior suggests that the appearance of faculae at different wavelengths and different positions is due to an interplay of geometry and thermal structure of the individual flux tubes.

Figure 2 shows the observed contrast of faculae from several sources (see Chapman 1994 for references). The scales are affected by the differing spatial resolution of the observations. Some progress in untangling the effects of filling factor on the contrast has been made by the Lockheed solar group by sorting their observations of facular contrast against the observed magnetic flux.

The source of the excess emission is not well understood. The emission appears to be a combination of the 'hot-wall' effect and an elevated temperature in the upper photosphere. The hot-wall effect is caused by the deeper layers of the photosphere shining through the partially evacuated flux tube (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: OBSERVATIONS for more details). It has even been suggested by Stenflo and others that electric currents could be responsible for heating of the flux tubes.

Faculae and sunspots both contribute significantly to variations in the total solar irradiance which is the amount of energy reaching the Earth. Sunspots produce the largest short-term decrease because of their high contrast.

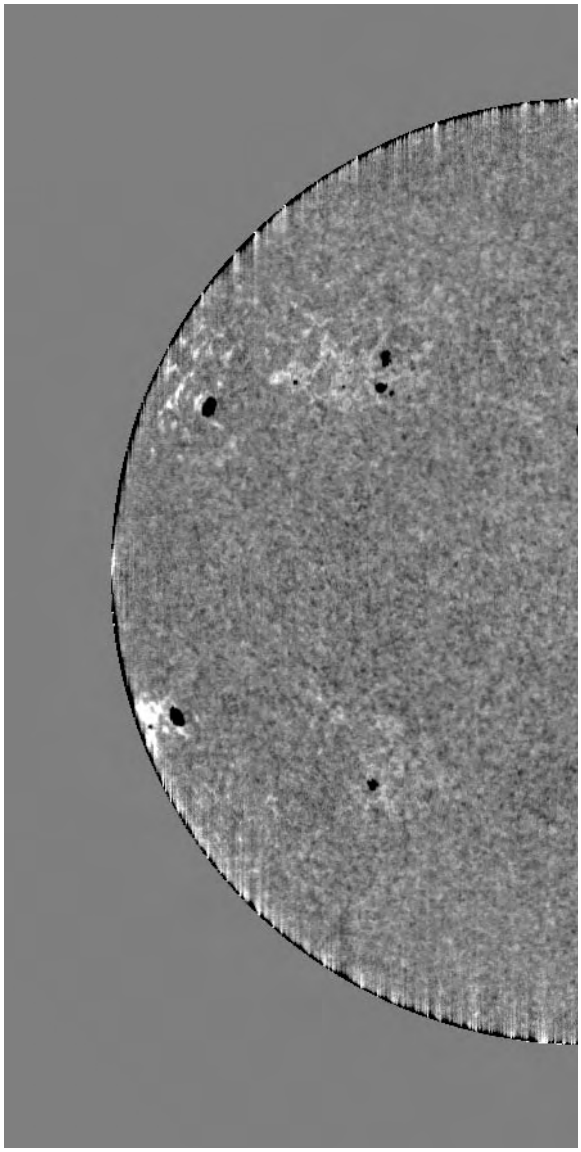


Figure 1. Photometric half-image of the Sun at 672.3 nm with a 10 nm bandpass filter on 1 July 1998. Facular contrast, seen with broad-band filters, increases towards the limb. This increase is especially notable between the limb and the sunspot near the lower left limb. Limb darkening has been removed to more clearly reveal the presence of faculae. Pixel size is 2.5 arcsec.

However, because faculae have a much larger area (at least a factor of 15 larger than sunspots) they make a positive contribution to solar irradiance variability that is approximately the same magnitude as the negative contribution of sunspots (Chapman *et al* 1996).

Figure 3 shows the change in facular area during the descending part of solar cycle 22 and the beginning phase of solar cycle 23. The data are from the Cartesian Full Disk Photometric Telescope No 2 (CFDT2), which has square pixels with a size of 2.5 arcsec. Both filters are near the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

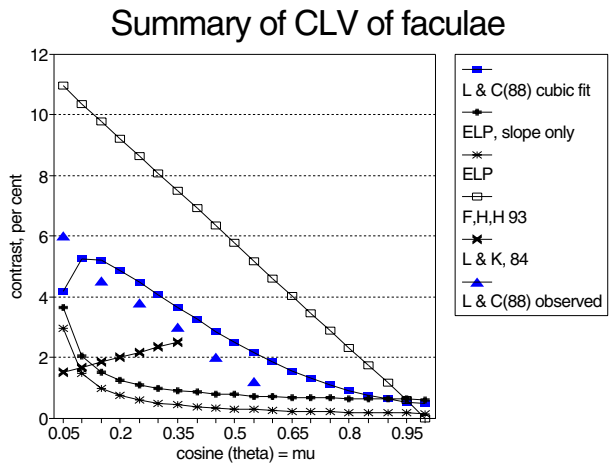


Figure 2. Facular contrast from a variety of sources versus μ. The vertical scale has not been adjusted for the different filling factors of the observations.

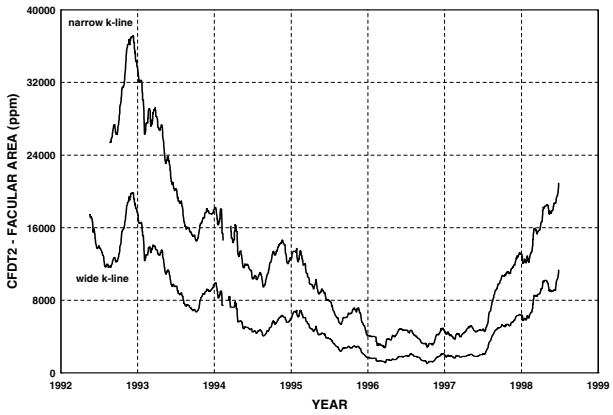


Figure 3. Facular area as measured with the Cartesian Full Disk Photometric Telescope, no 2. The narrow K line area was measured with a 0.3 nm bandpass filter. The wide K line area was measured with a 1 nm bandpass filter.

center of the calcium K line at 393.5 nm. The narrow filter has a bandwidth of 0.3 nm and the wide filter has a bandwith of 1 nm. One can see that the detected area of faculae is strongly dependent on the bandwidth of the observations. The areas shown here are much greater than one would detect in white light.

White-light faculae near the limb are morphologically identical to the photospheric network seen in absorption lines. However, the faculae seen in absorption lines have a higher contrast than white-light faculae. The identical morphology near the limb shown by Wilson supports the view that both thermal structure and geometry determine the appearance of faculae at different positions and different wavelengths. Above the photosphere, the field lines expand in the chromosphere, and the contrast increases in selected wavelengths. If there is a large concentration of flux, such as in an ACTIVE REGION, the area is called a PLAGE when seen in the light of strong FRAUNHOFER LINES.

The lifetime of an individual facula is not well established as it depends strongly on the spatial resolution of the observation as well as the evolution of the individual magnetic flux tubes that make up the facula. At moderate resolution, the lifetime is of the order of tens of hours but at higher resolution away from the solar limb (where foreshortening is less) the lifetime is less, from minutes to tens of minutes as smaller elements such as filigree and bright points are seen (see SOLAR PHOTOSPHERE: FILIGREE). Unfortunately, the determination of the lifetimes of individual faculae is difficult because the flux tubes can move and merge as the granulation buffets them (see SOLAR PHOTOSPHERE: GRANULATION).

Models of facular flux tubes have been reviewed by Spruit *et al* and have been compared with observations by Grossman-Doerth *et al* (1994) and, more recently, by Topka *et al* (1997) who conclude that a high contrast very near the solar limb requires something besides emission from hot walls.

Bibliography

Chapman G A 1994 Photometric observations of the Sun *The Sun as a Variable Star: Solar and Stellar Irradiance Variations* ed J M Pap, C Fröhlich, H S Hudson and S K Solanki (Cambridge: Cambridge University Press) pp 117–29

Chapman G A and Ziegler B 1996 An analysis of 1983 observations of facular contrast with an extreme limb photometer *Sol. Phys.* **168** 259–72

Chapman G A, Cookson A M and Dobias J J 1996 Variations in total solar irradiance during solar cycle 22 *J. Geophys. Res.* **101** 13 541–8

Grossmann-Doerth U, Knölker M, Schüssler M and Solanki S K 1994 Solar magnetic elements: models compared with observations *Solar Active Region Evolution: Comparing Models with Observations (ASP Conf. Ser. 68)* ed K S Balasubramaniam and G W Simon (San Francisco, CA: ASP) pp 96–101

Lawrence J K and Chapman G A 1988 Photometric observations of facular contrasts near the solar limb *Astrophys. J.* **335** 996–1004

Topka K P, Tarbell T D and Title A M 1997 Properties of the smallest solar magnetic elements: II. Observations versus hot wall models of faculae *Astrophys. J.* **484** 479–86

Gary A Chapman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Filigree

The SOLAR FILIGREE (figure 1) is a pattern of strings of small bright points, visible in the SOLAR PHOTOSPHERE located in the dark lanes which separate granules. Its origin is magnetic, and each bright point is the site of a magnetic flux tube (see the fourth section). The solar filigree was first observed in the wings of the Hα line at 6563 Å, at the Sacramento Peak Observatory, in 1972, by Dunn and Zirker (1973).

The pattern is found in the quiet Sun (at the boundaries of the supergranular cells, where it forms the so-called photospheric network), as well as in ACTIVE REGIONS, where it corresponds to the photospheric faculae, visible near the LIMB as bright extended PLAGES. In active regions, they often form rings, 2–3″ in diameter, which surround a single or a few granules.

The bright points forming the filigree are very clearly visible in wideband (about 10 Å) filtergrams taken in the Fraunhofer G-band lines of the CH molecule, centered at 4307 Å, where they appear much brighter than the surrounding granules (Muller 1985). They can also be seen in white-light images, but with a lower contrast. The difference of contrast observed in G-band (or Hα) filtergrams and in white-light images is due to the fact that the light comes from different layers in the photosphere: G-band filtergrams are formed in upper layers, near the middle of the photosphere, where the difference of temperature between the filigree elements and the surrounding atmosphere is larger than in deeper layers.

Size and shape

Although a majority of bright points forming the filigree are nearly circular, many of them are significantly elongated. One does not find bright points larger than 0"6 (450 km). Their size distribution has a peak at 0"3 (220 km) at λ 4307 Å and at 0"33 (240 km) at λ 5750. It must be noted, however, that the size distributions have been obtained with the finite resolution of 50 cm telescopes, which means that a much larger number of bright points of size smaller than 0"3 could be revealed by larger telescopes, and that the lower size limit is still unknown.

Brightness

The brightness of the real bright points is not well known either, because they are so small that they are strongly blurred by the finite size of the telescope, even when the seeing degradation is negligible. The fact that they are embedded in dark lanes makes the situation even worse: their contrast drops to zero for a resolution of 0"4. In white light, at 5750 Å, their brightness is, on average, 1.08 times the brightness of the mean photosphere, when the observation is made with a 50 cm telescope; maxima of 1.25 are measured for the brightest points. The mean relative brightness increases toward the limb, and reaches a maximum of 1.25 at $\mu = \cos\theta = 0.25$; closer to the limb, it decreases. The brightness, corrected for the finite size of the telescope, is 1.25 at the solar disk centre and 1.40 at $\mu = 0.25$.

In the G-band, which is formed higher in the atmosphere, the contrast is larger: it is 1.30 at the disk center, relative to the surrounding photosphere; this value is not corrected for blurring. For more details about the filigree bright points contrast, see the review by Muller (1994).

Dynamics of filigree bright points

The filigree bright points are always constrained to stay in the intergranular lanes, by converging granular flows. Because of this strong interaction with the dynamic and energetic granules, they are very unstable: their timescale is of the order of the lifetime of granulation (typically 10 min) and they often split and merge (see SOLAR PHOTOSPHERE: GRANULATION). However, they usually first appear in a relatively wide dark space at the junction of several converging granules. It thus seems that converging granules concentrate the magnetic flux into a flux tube of small cross section and high field strength, giving rise to a bright point. The maximum brightness is reached after 4 to 5 min, on average. The mean lifetime is about 10 min, but a few bright points can be identified for nearly one hour.

They are continuously buffeted by the moving expanding granules to speeds reaching several kilometers per second. Such large velocities reflect, in fact, the expansion rate of granules.

The magnetic origin of filigree bright points

The filigree bright points are closely associated with the small-size magnetic elements visible in high resolution magnetograms: each point is spatially associated to a magnetic element, which has nearly the same size. However, the magnetic elements do not always show up as a brightening, suggesting that the magnetic flux is more stable than the bright points, which may correspond to occasional flashes. The bright points are believed to be cross sections of nearly vertical magnetic flux tubes of field strength 1–1.5 kG, which extend from the top of the convection zone to the CORONA, through the CHROMOSPHERE and the TRANSITION REGION. Flux tube diameters are smaller than 400 km at the photospheric level; they expand upward. In the models, simplified flux tubes are considered, with sharp walls separating the magnetic plasma inside, from a magnetic-free atmosphere outside. The magnetic configuration of a real flux tube is certainly more complex (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES).

The flux tubes are in quasi-static equilibrium; consequently, the existence of a magnetic pressure makes the gas pressure inside the field smaller than outside: the tube is partly evacuated. Because of the lower density inside, the gas is heated by the hotter atmosphere outside, through the walls, on the one hand, and one sees deeper layers on the other hand: the combination of these two effects makes the magnetic elements appear brighter than the surrounding atmosphere. The more the field is concentrated, the more it is shining. In fact,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

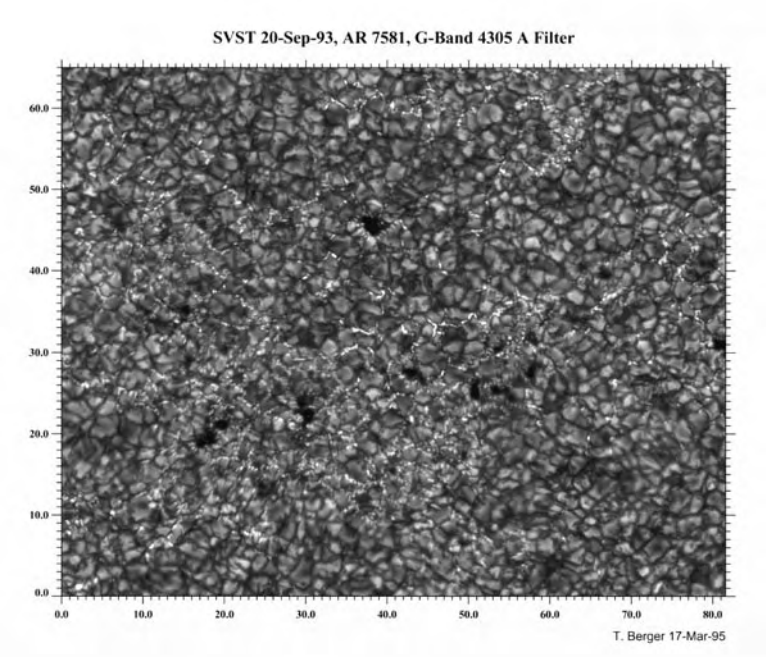


Figure 1. The solar filigree visible against the granulation pattern in a 4305 Å G-band filtergram of an active region; it is formed with more or less elongated bright grains located in intergranular lanes. Tickmarks are arcseconds. The filtergram was taken with the Swedish Vacuum Solar Telescope at La Palma, Canary Islands (courtesy of T Berger and the Lockheed Palo Alto Research Laboratory).

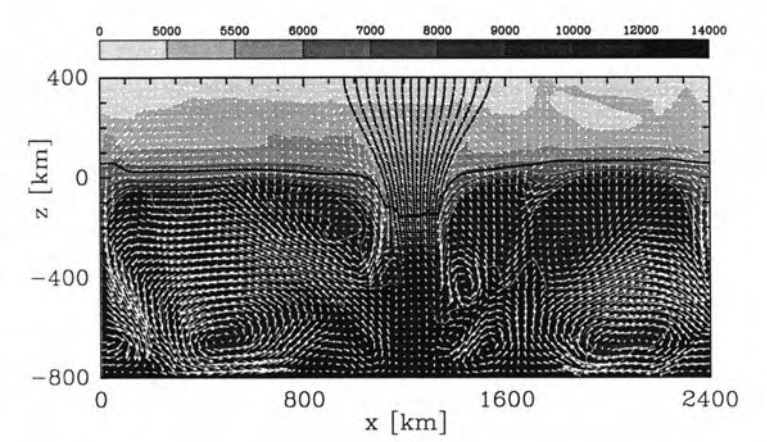


Figure 2. A schematic presentation of a nearly vertical magnetic flux tube, crossing the surface of the Sun (from a two-dimensional theoretical model computed by Steiner *et al* (1998)). The vertical lines represent the magnetic field lines. The horizontal line is the surface of the Sun, separating the convection zone below from the photosphere above. The white short vectors represent the convective flows, corresponding to granular cells. The gas pressure is lower inside the tube than outside; it is heated by the radiation from the hot, non-magnetic, atmosphere surrounding the tube. A filigree bright point corresponds to the cross section near the surface of the Sun, of a nearly vertical magnetic flux tube.

some observational details like the non-exact coincidence between the location of bright points and magnetic elements, or like the non-exact coincidence between the time of maximum brightness and of maximum field strength, is better explained with dynamical rather than static models (Steiner *et al* 1998). The rapid evolution and structural changes also demonstrate, of course, that the magnetic flux tubes are not static.

The physical origin of the magnetic flux tubes is still unknown: are they formed locally by some mechanism at work near the surface, and destroyed within a short time interval, or are they small fragments of SUNSPOTS, pulled out by convective flows, and spread over the surface of the Sun by large-scale convective motions ? But, whatever

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the formation process may be, it seems that the magnetic flux, which is of the order of 10^{17}–10^{18} Mx, is temporarily concentrated into a flux tube, visible as a small bright point in high resolution filtergrams, by the compression of converging granules; during this phase it is visible as a filigree bright point.

Filigree bright points, magnetic flux tubes and the heating of the solar corona and chromospheric features

The bright points visible in high-resolution filtergrams are used to track the chaotic motion of magnetic flux at the surface of the Sun. Magnetic waves, transverse to the magnetic field, are generated by these motions in the top of the convection zone (see SOLAR INTERIOR: CONVECTION ZONE), and propagate upward, through the chromosphere, the transition region and the corona, with an exponential growth of the amplitude. The dissipation of this energy (various kinds of dissipation processes have been proposed) may heat the solar corona and some chromospheric features as well.

Distribution at the surface of the Sun

Filigree bright points are found everywhere, in SOLAR ACTIVE REGIONS as well as in the quiet sun, where they are concentrated at the boundaries of supergranules by the convective supergranular flows, forming the photospheric network. Further, the density (number of bright points by surface unit) of the network is not uniform at the surface of the Sun: it is minimum at the sunspot belt latitudes (where the filigree is present mainly in active regions) and maximum at the poles. Polar faculae consist of bright point concentrations (see SOLAR PHOTOSPHERE: FACULAE). The surface distribution is not constant in time: it varies with an 11 y period. More precisely, the density at the equator and at the poles varies almost in antiphase to the sunspot number, with a 1–2 y shift.

The size and field strength histograms of the observed bright points in the network are narrow; this means that each bright point carries nearly the same amount of magnetic flux, of the order of 2.5×10^{17} Mx. The spatial density of bright points is thus equivalent to a magnetic flux, and their spatial and temporal variations reflect magnetic flux variations in the quiet atmosphere of the Sun. Neither this weak network flux, corresponding to the very quiet Sun, nor its time variation are detectable with conventional full disk SOLAR MAGNETOGRAPHS.

Bibliography

Dunn R B and Zirker J B 1973 *Solar Phys.* **33** 281
Muller R 1985 *Solar Phys.* **100** 237
Muller R 1994 *Solar Surface Magnetism* ed R J Rutten and C J Schrijver (Dordrecht: Kluwer) p 73
Steiner O, Grossmann-Doerth U, Knölker M and Schüssler M 1998 *Astrophys. J.* **495** 468

Richard Muller

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Granulation

Under good atmospheric seeing conditions observations of the Sun's surface with a telescope of at least 30 cm aperture reveal a rugged or polygonal pattern of bright elements separated by darker lanes that form an interconnected network; this pattern is called GRANULATION and was first described by Sir William Herschel in 1801. The granulation represents the top of the solar convection zone (see SOLAR PHOTOSPHERE), where hot gas rises to the surface, cools by radiation and sinks back into deeper layers. The observed size of granules ranges from approximately 150 km (the observational limit of ground-based telescopes) to approximately 2500 km, with 1.3×10^6 km^2 mean cell area of the granular elements including one-half of the surrounding dark lanes. The root-mean-square contrast of the granular pattern, after correction for degradation by telescope, stray light and seeing, is about 13% at a wavelength of 550 nm. Horizontal and vertical motions measured in granules and intergranular lanes are of the order of 2 km s^{-1} peak to peak. Granules change their shape and identity on time scales of 8–16 min; they evolve mainly by expansion, fragmentation, merging or dissolution. Numerical simulations of convection at the solar surface are capable of predicting many details of the granular structure (horizontal and vertical velocity, temperature and size distribution, evolution, vorticity etc) that agree rather well with the observations—one has to keep in mind, however, that the smallest structure of the models is not accessible to observation.

Observations of the granulation

Ground-based observations of the solar granulation (see SOLAR TELESCOPES AND INSTRUMENTS: GROUND) are hampered by two basic effects: the limited resolution of the telescopes employed and the image degradation due to refractive index inhomogeneities in the terrestrial atmosphere called SEEING. An order-of-magnitude consideration illustrates the difficulties. A telescope of $D = 60$ cm aperture, working at $\lambda = 500$ nm under diffraction-limited conditions, offers a resolution, defined as $r = 1.22\lambda D^{-1}$, of 0.21 arcsec or 150 km on the Sun. This means that, because of the telescope's optical modulation transfer function, the contrast of features of 0.2 arcsec extent is reduced by a factor of order 10. Any optical aberration in the telescope increases this factor. Moreover, daytime atmospheric conditions usually limit the attainable resolution to values larger than 1 arcsec. Only the combination of carefully selected sites with optically nearly perfect telescopes and refined observing techniques, such as real-time frame selection or speckle interferometry (see INTERFEROMETRY: GROUND), makes quantitative studies of the granulation down to the smallest scales possible. Only a few telescopes in the US (Sacramento Peak), France (Pic du Midi) and Spain (Canary Islands) are capable of doing such work. Figure 1 shows the granular pattern observed with speckle interferometric techniques; here details of approximately 150 km diameter can be studied.

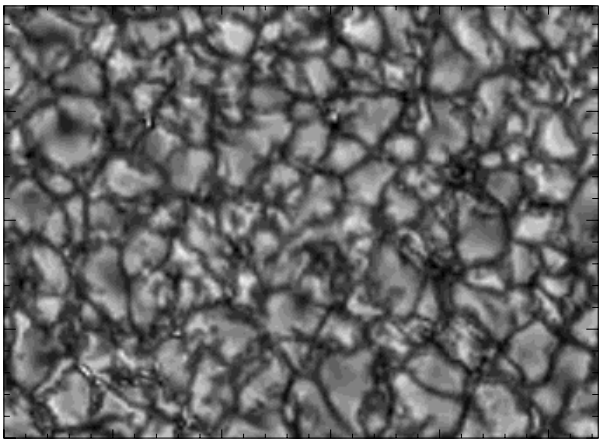


Figure 1. Solar granulation at disk center, observed with the German Vacuum Tower Telescope (VTT) at the Observatorio del Teide (Tenerife), and restored with speckle interferometric techniques by V Wilken and coworkers (Göttingen) 1997. Field of view 20×14.5 Mm^2.

Size distribution, shape and structure of the granulation

The solar granulation represents a continually changing pattern on the Sun. Nevertheless, most observers select individual granules for further study; to this end different algorithms are applied (cf Muller 1999). One consists of a combination of high-pass and low-pass filtering in the spatial frequency domain with subsequent thresholding close to the mean photospheric intensity; this defines iso-intensity contour lines and thus segments the image into granules and intergranular lanes. Alternative procedures are 'lane-finding' schemes (cf Spruit *et al* 1990), algorithms based on local gradients of the intensity, or the determination of the intensity inflection points which mark the edges of granules. As an example, figure 2 gives the size distribution of granules derived with the latter technique. The smallest granules, near the observation limit, here 0.4 arcsec or 290 km on the Sun, are the most frequent ones with a slow decline in number up to a diameter near 2 arcsec, and a subsequent rapid decline towards the largest granules. Counting the number of granules per unit area on the quiet Sun, different authors arrive at similar values: on average a granule with the surrounding half of the intergranular lane covers 1.3×10^6 km^2; this is called the mean cell area.

Intensity contrast and temperature fluctuations

The best images taken from the ground exhibit a root-mean-square intensity contrast (without correction for the telescope point spread function and scattered light) of $\simeq$10.7% at $\lambda = 468 \pm 5$ nm. Contrasts of up to 8–9% rms have been measured at 570 nm. After correction for the above-mentioned effects, values between 11% and 20% were obtained. Conservative estimates, e.g. from speckle-reconstructed images like figure 1, arrive at a contrast around 13% at 550 nm. This would correspond to rms

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

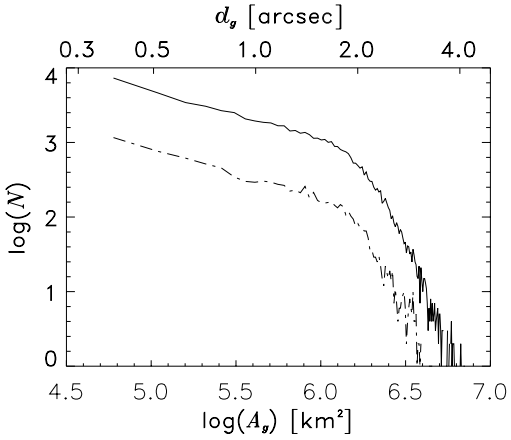


Figure 2. Plots of the granular areas and diameters published in 1997 by J Hirzberger and coworkers (Graz, La Laguna, Ondřejov). Full line, whole sample of 61 138 granules; dash-dotted line, 9656 granules from images with granular contrasts $\Delta I_{\rm rms}/I > 5.5\%$.

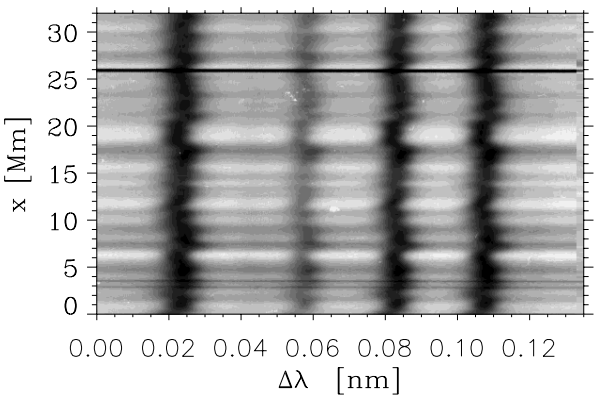


Figure 3. High-resolution slit spectrogram taken with the German VTT at Izaña (Tenerife). The weak line near $\Delta\lambda = 0.057$ nm is the Fe I line at $\lambda = 491.15$ nm. Note the strong blueshifts in the granules (bright horizontal streaks) and the redshifts, frequently connected with line broadenings, in the intergranular lanes. First published in 1996 by A Nesis and coworkers (Freiburg).

temperature fluctuations of order 200 K; according to a rule of thumb the peak-to-peak differences are 3 times larger, i.e. 600 K (see also figure 4).

On the other hand, numerical models predict temperatures around 10 000 K for the granular upflows and of 6000 K for the intergranular downflows, if one refers to the same geometrical depth level near the surface. Why does one not observe this? The reason is the very high temperature sensitivity of the H^- opacity ($\propto T^{10}$): at the same optical depth ($\tau = 1$) the temperature fluctuations are much smaller than at the same geometrical levels. Additionally, the limited spatial resolution of the observations plays a role.

Evolution, lifetime and horizontal proper motions

The study of granular evolution and lifetime requires time series of high spatial resolution that are available from space, e.g. from the Solar Optical Universal Polarimeter (SOUP) instrument on Spacelab 2, or from exceptional ground-based observations (cf Brandt and Wöhl 1998). Granules evolve in several different ways. Frequently they expand and split into smaller components that drift apart, thus forming an exploding granule. The fragments may in turn grow and fragment, or merge with others; shrinking and dissolution of granules are also observed. Again, the numerical simulations help us to understand the basic physics behind this: when a fluid parcel approaches the solar surface, radiative cooling carries away its energy; in a few tens of seconds the density increases above that of its surroundings. Then the vertical motion is decelerated and diverted horizontally. Further cooling lets the density grow further, and finally gravity pulls the parcel back down in the lanes between granules. This whole process takes several minutes.

Experimentally determined values of the granular lifetime depend on the method used. If the intensity pattern is autocorrelated, then (after removal of oscillatory intensity variations) lifetimes around 8 min are obtained; here, lifetime is defined as drop to $1/e$ of the autocorrelation function. If individual granules are followed throughout their evolution, then visual identification schemes yield a mean lifetime of approximately 16 min (with a standard deviation $\simeq$5 min), whereas objective algorithms again give values of 8–10 min. The reason for this large discrepancy is not clear; on the other hand it should not be taken too seriously, since each identification scheme involves some arbitrariness of the definitions: e.g. when two granules merge, which of the predecessors 'continues to live'?

An important feature of granular evolution is the horizontal proper motion of granules. It can be determined by local correlation tracking, a method that searches for the best spatial correlation in small subfields of consecutive granulation pictures. With this method a larger pattern of 4000–8000 km mesh width was found, the mesogranulation (see SOLAR PHOTOSPHERE: MESOGRANULATION). In certain parts of the mesogranulation granules consistently move horizontally towards a common center (converging region) while in other parts diverging motions prevail; the typical speed of the horizontal motion is 0.5 km s^{-1}. The lifetime of the mesogranulation appears to be of the order of hours—but is not well established yet. There are indications that some properties of granules, such as area, brightness, lifetime and expansion or shrinking rate, depend on their position in the mesogranular field.

Vertical and horizontal velocities

At photospheric levels and above the whole Sun is covered by a pattern of oscillatory elements with vertical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

amplitudes of several 100 m s^{-1} and a wide range of sizes (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY); these also show up as intensity oscillations. For studies of the granulation the effects of these oscillations have to be separated from the intensity and velocity variations of the granulation pattern. From high-resolution slit spectrograms taken at disk center (like the one shown in figure 3) one can measure the vertical velocity related to granular motion and obtain peak-to-peak values in the range 1.5–2 km s^{-1}, after removal of the velocity components due to the oscillation. The horizontal velocity components can be investigated by observations near the solar limb where they show up as line-of-sight velocities. In this way values are found that are a factor of $\simeq$1.5 larger than the vertical components.

The vertically ascending material does not stop its motion at the very surface but penetrates into higher layers owing to excess momentum; this is known as convective overshooting. Velocity measurements in lines of different strength reveal that the convective overshooting at granular scales extends several hundred km above the photosphere but the exact height dependence of the velocities is still not well established. Figure 4 shows an empirical model of an average granule derived from 2-dimensional spectroscopic observations of a large number of granules. An inversion technique was used to obtain the height dependence of temperature, density and line-of-sight velocity; horizontal velocities were derived from mass conservation considerations. Between the upflow in the center of the granule and the downflow in the intergranular lane there is a velocity difference of $\simeq$2 km s^{-1}; also the decrease of velocity with height is clearly visible from this figure. At height levels below 140 km, upward velocity is correlated with higher temperature near the granule center, which implies upward transport of energy, and lower temperature with downward motion above the intergranular lane. Above 140 km the temperature stratification reverses.

Granulation in active regions and near sunspots

The granular pattern in magnetic regions near sunspots and in ACTIVE REGIONS (see SOLAR CHROMOSPHERIC PLAGE) exhibits lower contrast and reduced vertical velocity as well as smaller mean cell area; it is called abnormal granulation. Also the temporal evolution is slower, and horizontal flow speeds as measured by local correlation tracking are one-half those in the quiet photosphere. When observed in the wings of strong lines, such as Hα, or in the band head of the CH radical at 430 $\pm$ 5 nm (the G-band) these regions show small bright features of $\simeq$200 km extent that frequently sit in the spaces between granules. They were first seen in 1973 by R Dunn and J Zirker and were called filigree (see SOLAR PHOTOSPHERE: FILIGREE). Apparently, the presence of a magnetic field (see SOLAR PHOTOSPHERE: INTRANETWORK AND MAGNETIC FIELDS) modifies the convective energy transport in these regions.

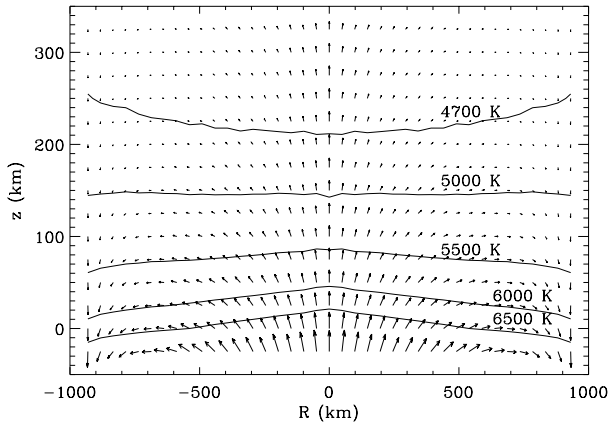


Figure 4. Velocity and temperature field measured for an average granule. The arrow length is proportional to the velocity, with a maximum vertical upward velocity of 1.6 km s^{-1}. The distance from granule center to the intergranular lane center has been scaled to 940 km. Published in 1996 by B Ruiz Cobo and coworkers (La Laguna).

Numerical simulations—agreement with observations

In the recent decade numerical simulations of the solar convection, particularly of the granulation, have attained a high degree of sophistication (SOLAR INTERIOR: CONVECTION THEORY). Despite their limitations, e.g. in the very small volume that can be modelled, some ambiguity in the boundary conditions, in the treatment of viscosity etc, the simulations are capable of describing many aspects that are confirmed by observations. For example, spatial intensity distribution and size spectrum, amplitudes of horizontal and vertical motion, evolution and topology of the granular pattern are rather well reproduced—for comparison the numerical results have to be slightly degraded spatially in order to account for the limited spatial resolution of the observations. Simulations help us to arrive at a basic understanding of the top of the solar convection zone and to infer details that are not accessible by observation. We learn that solar convection is driven by both buoyancy forces at subphotospheric levels and by radiative cooling at the surface thermal boundary layer, that the cooled plasma turns over and converges into the dark intergranular lanes and further into the vertices between granulation cells forming narrow 'fingers' of high-velocity downdrafts, which penetrate many scale heights. Recent observations have shown that some intergranular vertices are much more stable in space and time than the rest of the granular pattern. Such persistent intergranular features may mark at the photospheric level the locations of the high velocity downdrafts.

Moreover, the numerical simulations suggest that downdrafts are surrounded by high-vorticity regions while upflows are essentially laminar. Spectrograms of very high spatial resolution (cf figure 3) frequently exhibit increased linewidth near the edges of granules

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where vertical velocity gradients are large. Here shear flows are surmised to create turbulence—more cautiously, some authors use the term unresolved velocities. When measured in lines that originate 70–200 km above the photosphere values of the rms turbulent velocity $w_{\rm turb} \approx 0.5$ km s^{-1} are found, which increase with decreasing height; extrapolation to photospheric levels yields $w_{\rm turb} \approx 1.5$ km s^{-1}. In this case the turbulent pressure associated with the turbulent velocity: $P_{\rm turb} = (1/2)\rho w_{\rm turb}^2$, where ρ is the density, may locally attain a significant fraction (of order 0.3) of the gas pressure, and thus play an important role in the dynamics of granular evolution.

Bibliography

Brandt P N and Wöhl H 1998 *The Solar Granulation, Video Movie Tape* (Göttingen: Institut für den Wissenschaftlichen Film)

Bray R J, Loughhead R E and Durrant C J 1984 *The Solar Granulation* (Cambridge: Cambridge University Press)

Muller R 1999 The solar granulation *Motions in the Solar Atmosphere* ed A Hanslmeier and M Messerotti (Dordrecht: Kluwer)

Rutten R J and Severino G (ed) 1989 *Solar and Stellar Granulation* (Dordrecht: Kluwer). Contains extensive bibliography up to 1989

Spruit H C, Nordlund Å and Title A M 1990 Solar convection *Ann. Rev. Astron. Astrophys.* **28** 263–301

Peter N Brandt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Intranetwork and Turbulent Magnetic Fields

The names 'intranetwork' and 'turbulent magnetic fields' are used to represent the solar magnetic fields of mixed polarities at the smallest scales of the spatial spectrum. Since the spatial separation of the opposite polarities is small, and since the magnetic flux of each small-scale magnetic element is tiny, they can only be made partly visible in 'deep' magnetograms obtained with both high spatial resolution and long integration times. If the magnetic elements are much smaller than the photon mean free path (optically thin regime) they may not be spatially resolved even in principle. Yet such optically thin magnetic structures may be diagnosed by indirect methods, in particular with the Hanle effect (see SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS).

While the word 'intranetwork' generally refers to the spatially resolved mixed-polarity field, the phrase 'turbulent field', when used in the context of interpreting spectro-polarimetric observations, usually refers to the spatially unresolved, 'hidden' mixed-polarity field, which remains practically invisible in solar magnetograms owing to cancellation of the contributions from the opposite-polarity elements within the spatial resolution element.

Figure 1 illustrates the typical spatial distribution of the resolved mixed-polarity magnetic fields on the quiet Sun in the form of a 'deep' magnetogram of the line-of-sight component of the magnetic field. Such a magnetogram is a map of the circular polarization that is produced by the longitudinal Zeeman effect in a selected spectral line. The opposite polarities are represented by the brighter and darker patches against a neutral background, giving a 'salt and pepper' appearance when the polarities are mixed.

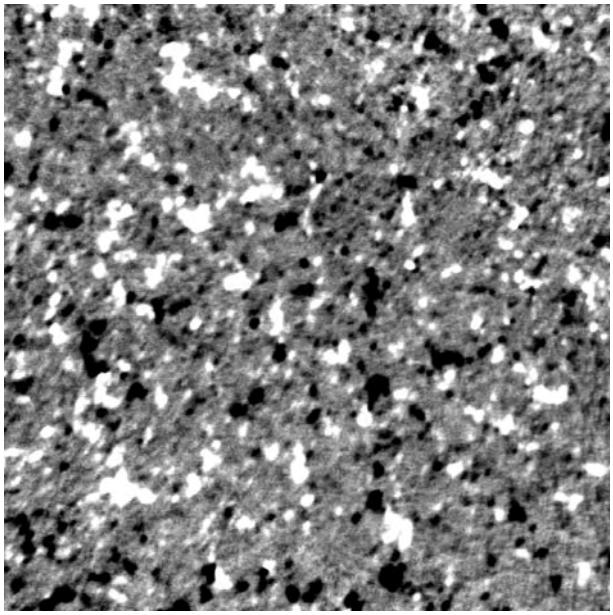


Figure 1. 'Deep' magnetogram of a 280 × 280 arcsec^2 portion of the quiet Sun at the center of the solar disk, obtained on 1 July 1998, by Jongchul Chae at BBSO. The opposite-polarity magnetic fluxes are represented by the brighter and darker patches against a neutral grey background. 4096 separate frames have been added to enhance the sensitivity and bring out the weaker magnetic fluxes. Courtesy of Big Bear Solar Observatory/New Jersey Institute of Technology.

Intranetwork fields

The choice of the word 'intranetwork' is made to contrast these magnetic fields with those that are related to the emission network that is seen as a brightness pattern in chromospheric and TRANSITION REGION spectral lines (see CHROMOSPHERE: NETWORK). This emission is located at the boundaries of the velocity cells of the supergranulation (with a cell diameter of 30 000 km) (see SOLAR PHOTOSPHERE: SUPERGRANULATION). Since the magnetic field is frozen into the plasma, it is carried by the flow pattern of the supergranulation and converges to the downdraft regions at the cell boundaries. The supergranulation flow thus organizes the large-scale pattern of magnetic flux and emission.

Because of the organizing effect of the supergranulation flow, most of the magnetic flux becomes concentrated in the network, and it is this network flux that is the source of the large-scale magnetic field that reaches up in the CORONA and beyond. The magnetic fluxes in the network are predominantly of one polarity over scales of the supergranulation size. This flux occurs in the form of highly intermittent magnetic flux tubes, which occupy typically 1% of the volume in the photosphere and have field strengths on the order of 1 kG.

In magnetograms with insufficient polarimetric sensitivity and spatial resolution one may get the impression that *all* the photospheric magnetic flux is in the form of network flux with a typical magnetic filling factor of 1% , while the remaining 99% appears as if it were field free. No portions of the solar plasma can, however, be strictly field free, since the flux concentration by the supergranular and granular flows is of finite efficiency (see SOLAR PHOTOSPHERE: GRANULATION), and any left-over, unconcentrated weak magnetic flux will be acted upon by the convective and turbulent motions and tangled up until some magnetoturbulent equilibrium between magnetic and kinetic energies is reached. Many of these magnetoturbulent processes may also take place deep in the convection zone before the fluxes emerge at the solar surface (see SOLAR INTERIOR: CONVECTION ZONE).

What we see in figure 1 is actually a mixture of network and intranetwork fields. The more prominent flux patches are due to the network, while the rather inconspicuous dot-like features represent the intranetwork field. These features can only be seen in 'deep' magnetograms under good seeing conditions.

If an observed magnetic element carries a small magnetic flux, this does not necessarily imply that the field strength is also small, since the measured flux is a product

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

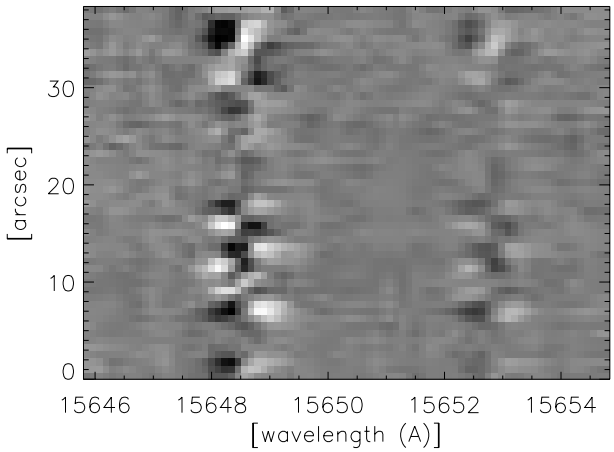


Figure 2. Spectral image in the near-infrared of the circular polarization, recorded in 1997 at the National Solar Observatory (Sunspot, New Mexico) with the spectrograph slit positioned at the center of the solar disk. The two spectral lines of neutral iron at 15 648 and 15 652 Å show the characteristic Zeeman-effect signatures of a magnetic field that exhibits variations of sign or polarity on scales of a few arcsec along the spectrograph slit. At each spatial position the circularly polarized line profile is antisymmetric, but the sense of the antisymmetry depends on the polarity. Courtesy H Lin and T Rimmele, NSO/Sacramento Peak.

of field strength and magnetic filling factor. For fluxes larger than about 2×10^{17} Mx (corresponding to 20 G Mm^2; 1 Mm = 1000 km, corresponding to 1.4 arcsec on the Sun) the field strength depends very weakly on flux (and is about 1.0–1.5 kG), which may be understood in terms of the convective collapse mechanism for the formation of kG flux tubes. For smaller fluxes the field strength decreases rapidly with decreasing flux, since the diameter of kG flux tubes with such small flux would be optically thin, which would prevent the collapse mechanism from working (owing to damping by the effective heat exchange between flux tube interior and exterior).

In the visible spectral range the Zeeman splitting is generally incomplete (smaller than the spectral line width), which has the consequence that the observed circular polarization mainly carries information on the magnetic *flux* but not directly on the field strength or filling factor. To extract information on the field strength one has had to develop indirect methods by forming ratios between the circular-polarization amplitudes measured in spectral lines with different sensitivities to the Zeeman effect. In the infrared, however, the Zeeman splitting is much larger and becomes complete even for moderately strong fields. This allows us to extract direct information on both flux and field strength (and thus also on magnetic filling factor) without having to use tricks such as polarized line ratios. Figure 2 gives an example of a spectro-polarimetric recording of intranetwork magnetic fields in the wavelength region around 1.565 μm. The spectrograph slit has been placed at the center of the solar disk, and the spectrogram shows the characteristic antisymmetric circularly polarized line profiles due to the longitudinal Zeeman effect. The polarity reversals occur on a scale of a few arcsec, close to the scale of the solar granulation. Such recordings allow the magnetic flux, field strength, and filling factor to be determined for each spatial location along the spectrograph slit.

The relation between the network and intranetwork magnetic fields is not clear. It is generally believed that the magnetic flux that is the source of solar activity originates by dynamo processes inside the Sun (see DYNAMOS: SOLAR AND STELLAR), preferentially near the bottom of the Sun's convection zone. New magnetic flux emerges from the interior in bipolar form on different scales. The bipolar regions with the largest amount of flux form the ACTIVE REGIONS with SUNSPOTS and appear at low heliographic latitudes. Smaller bipolar regions without sunspots on typical scales of 10–20 arcsec are called ephemeral active regions because of their shorter lifetimes. They are more abundant with a larger latitudinal spread. It is possible, although controversial, that the still smaller-scale and ubiquitous intranetwork fields with scales of a few arcsec may be considered as the small-scale portion of a continuous spectrum of bipolar magnetic regions that includes active regions in its large-scale portion. While the flux emergence rate for active regions is 10^{20} Mx day^{-1}, it is 10^{22} Mx day^{-1} for ephemeral regions and 10^{24} Mx day^{-1} for intranetwork fields (Zirin 1987). The smallest scales thus dominate the flux budget.

The large bipolar magnetic regions (active regions) are highly organized, with the magnetic polarities being oriented in the east–west direction and evolving with the 22 year solar magnetic cycle according to Hale's polarity law (see SOLAR CYCLE). The ephemeral active regions also obey Hale's polarity law in a statistical sense, but with a much larger spread in their orientations. It is not yet known whether the orientations of the intranetwork fields are entirely random or not. If their directional distribution contains a non-random component, they could even be the dominating contributor to the global magnetic flux pattern owing to their extremely high flux emergence rate (which is 10 000 times that of active regions). Their possible role in the operation of the solar dynamo is still a controversial subject.

The large emergence rate needs to be balanced by a similar rate of flux removal, otherwise the photosphere would become 'choked' by all the emerging flux. Flux removal can occur mainly either by (i) submergence, whereby flux ropes are retracted down below the solar surface, or by (ii) annihilation, when opposite-polarity fields meet and reconnect on the Ohmic diffusion scale (on the order of 1 km). Reconnection can also release field lines from being anchored to the solar body and allow the liberated flux loops to be convected into the heliosphere. It has not yet been possible to determine the relative role of these various mechanisms for the overall flux budget.

Large active regions as well as ephemeral active regions are localized bipolar structures that only cover a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

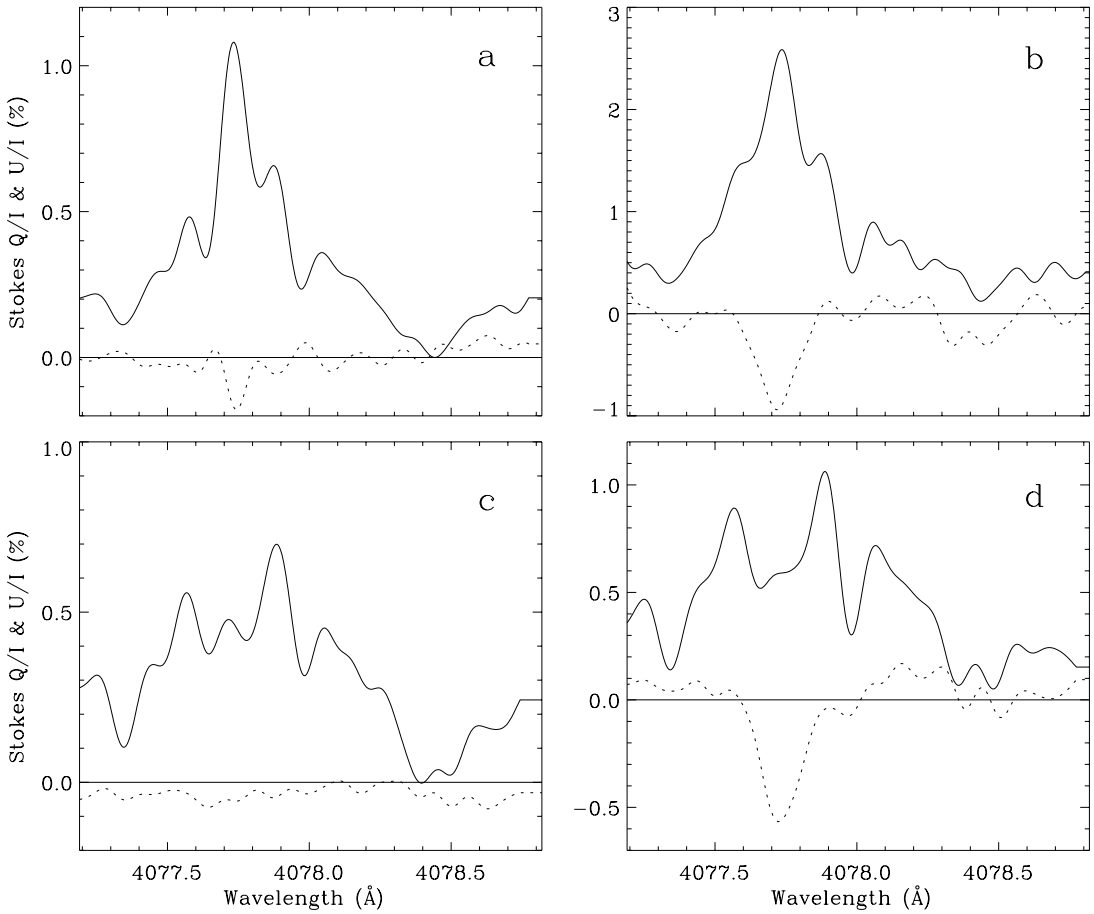


Figure 3. Illustration of spectral signatures of the Hanle effect in the Sr II 4077.7 Å line at four locations on the solar disk. The curves represent the degree of linear polarization oriented parallel to the nearest solar limb (Stokes Q/I, full curves) and at 45° to the limb (Stokes U/I, dotted curves). The shapes of the polarized line profiles change from place to place on the Sun because of varying Hanle depolarization in the line core of the Q/I profiles (while the line wings are almost unaffected). Partially resolved magnetic fields with a net orientation cause a rotation of the plane of linear polarization, which shows up as a U/I signal in the line core. From Bianda *et al* (1998).

small fraction of the solar surface. Likewise the network magnetic fields, which represent the footpoints of the large-scale coronal field, are highly intermittent in the photosphere with kG fields and a magnetic filling factor on the order of 1%. The intranetwork magnetic fields retain some degree of intermittency, although much less pronounced, and have substantially lower field strengths (in the range of hundreds of G or less). Other properties of the small-scale and intranetwork fields are discussed in Stenflo (1989), Martin (1990) and Wang *et al* (1995).

Turbulent fields: beyond the spatial resolution limit

It is not known to what extent the magnetic field retains some degree of intermittency when going to still smaller scales. The theory of magnetoconvection (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION) in a highly stratified medium is not in a position to make any clear prediction about this, and the scales that we are now considering are too small to be spatially resolved. Still it is possible to obtain information on the spatially unresolved turbulent magnetic fields via spectro-polarimetric methods involving the Hanle effect.

Spectral lines are linearly polarized by coherent scattering processes, but in the presence of a turbulent magnetic field the amount of polarization is reduced depending on the strength and degree of intermittency of the turbulent field. To apply this method one first has to describe the spatially unresolved field with an idealized model that is characterized by some free parameters, the values of which are constrained or determined by the amount of observed Hanle depolarization. With a one-parameter model, characterized by a volume-filling field with a random orientation of the field vectors (note, however, that the spatial topology is not constrained to be random) and a single-valued field strength as the free parameter, one finds field strengths in the range of 4–40 G, depending on position on the solar disk. If the field would

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

be intermittent with a filling factor smaller than unity, the field strength would be correspondingly larger.

Figure 3 illustrates the kinds of spectral signatures used to extract information on both spatially unresolved and partially resolved magnetic fields via the Hanle effect. The figure shows four recordings at four different locations on the solar disk of the linear polarization produced by coherent scattering in the spectral line Sr II 4077.7 Å. The Hanle effect is only effective in the line core but not in the wings, and it manifests itself in the form of depolarization and rotation of the plane of polarization. The four full curves show various degrees of depolarization, while the dotted curves show a polarization signal in the line core if the plane of polarization has been rotated. The interpretation of the observations is complicated, but such data give access to a different parameter regime than can be studied via the ordinary Zeeman effect.

It is natural to expect that the field would become ubiquitously tangled by the turbulent motions until some equipartition between the magnetic and kinetic energies is reached at the smallest scales, where the convective collapse mechanism is inoperative and magnetic diffusion plays an increasing role. The nature of the magnetoturbulent spectrum, however, remains unknown. With future multiline applications of Hanle diagnostics it should be possible to make the interpretative models more sophisticated and realistic.

If the turbulent field strength is determined by the kinetic energy spectrum of the granulation and smaller-scale turbulence, and if this kinetic spectrum is the same at all places on the quiet solar disk, then one would expect the turbulent field strength to be independent of position on the quiet Sun (outside active regions). Hanle observations, however, show (Stenflo *et al* 1998) that the turbulent field strength is not spatially invariant but exhibits variations by an order of magnitude. It is a quantity that needs to be mapped across the solar disk.

Outlook

Progress in understanding the intranetwork and turbulent magnetic fields requires improved spatial resolution combined with high polarimetric accuracy and spectral resolution. The near-infrared spectral region is particularly suited for Zeeman diagnostics of the spatially resolved fields, since the large Zeeman splitting of infrared lines allows a determination of both field strength and magnetic filling factor (cf Solanki 1995). The Hanle effect as a diagnostic of spatially unresolved turbulent magnetic fields is best applied with lines at shorter wavelengths in the visible domain, where scattering polarization occurs with larger amplitudes.

Small-scale structures evolve on small time scales, so high spatial resolution also demands moderately high temporal resolution. The four observational parameters of spatial, temporal and spectral resolution as well as polarimetric accuracy form a 4D parameter space that characterizes the observations. The observations will always be photon starved, since it will always be necessary to make trade-offs between these four parameters, even with the largest foreseeable future telescopes. The chosen trade-offs depend on the emphases and objectives of the particular observing program. With adaptive optics it will become possible to combine high spatial resolution with high-precision spectro-polarimetry requiring relatively long integration times. This, in combination with larger telescope apertures, will provide access to unexplored territory in the 4D parameter space, allowing for major advances in understanding magnetoconvection and the operation of the solar dynamo.

Bibliography

Bianda M, Stenflo J O and Solanki S K 1998 Hanle diagnostics of solar magnetic fields: the Sr II 4078 Å line *Astron. Astrophys.* **337** 565–78

Martin S F 1990 Small-scale magnetic features observed in the photosphere *IAU Symp.* **138** 129–46

Solanki S K 1995 Magnetic field measurements in the infrared *Infrared Tools for Solar Astrophysics: What's Next?* ed J R Kuhn and M J Penn (Singapore: World Scientific) pp 341–53

Stenflo J O 1989 Small-scale magnetic structures on the Sun *Astron. Astrophys. Rev.* **1** 3–48

Stenflo J O, Keller C U and Gandorfer A 1998 Differential Hanle effect and the spatial variation of turbulent magnetic fields on the Sun *Astron. Astrophys.* **329** 319–28

Wang J, Wang H, Tang F, Lee J W and Zirin H 1995 Flux distribution of solar intranetwork magnetic fields *Solar Phys.* **160** 277–88

Zirin H 1987 Weak solar fields and their connection to the solar cycle *Sol. Phys.* **110** 101–7

J O Stenflo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Mesogranulation

Mesogranulation is a horizontal cellular pattern in the convective motions at the surface of the Sun with sizes of 5–10 Mm. Convection is the transport of energy by rising warm fluid and sinking cool fluid. How convection appears at the solar surface depends on how it is viewed. In the emergent intensity, solar convection is visible as a cellular pattern called granulation (see SOLAR PHOTOSPHERE: GRANULATION). Observed as quasi-steady fluid velocity averaged over times of order 1 h, a larger-size cellular pattern of diverging horizontal flows appears, called mesogranulation, with diameters of 5–10 Mm. Horizontal velocity flows also reveal an even larger pattern of supergranulation (see SOLAR PHOTOSPHERE: SUPERGRANULATION).

Mesogranulation was first observed as a pattern in the persistent (time-averaged) vertical velocity observed as a Doppler shift in photospheric and chromospheric spectral lines. It is most clearly seen in the quasi-steady divergence of the horizontal surface velocity, which can be determined by following the motion across the surface of features in the granulation using local correlation tracking. This technique calculates the cross correlation between images at two different times as a function of their displacement. The local motion between the two times is taken to be the displacement that maximizes the cross correlation at each location in the image. Using the quasi-steady horizontal motions to calculate the trajectories of markers in the flow shows that they end up after a few hours in thin lanes surrounding the mesogranule cells (figure 1). Magnetic fields are shuffled about in a similar fashion by the fluid motions and as a result are observed to collect on the mesogranule boundaries and then be swept to the supergranule boundaries where the magnetic field produces the chromospheric network (see CHROMOSPHERE: NETWORK).

The mesogranular component of solar convection affects the properties of the smaller-scale granules. By conservation of mass, regions of horizontally diverging flow are regions of upflowing fluid and regions of horizontally converging flow are regions of downflowing fluid. Bright, long-lived, rapidly expanding granules are found preferentially in upflowing, diverging interiors of mesogranule cells. Small, faint, short-lived granules are found preferentially in the downflow lanes of converging flow at the edges of the mesogranule cells.

Solar convection is driven by radiative cooling in a very thin layer at the solar surface of that small fraction of fluid that reaches the surface. This cool fluid forms the low-entropy cores of downdrafts that descend into the interior and is the primary site of the buoyancy work that produces the convective motions. The convective flow pattern is determined by mass conservation. Warm fluid ascending into lower-density layers is like a fountain (figure 2). Most of it must turn over and become entrained in the downdrafts within a density scale height (the distance in which the density decreases by a factor of $e = 2.7$). With increasing depth, the temperature increases and the scale height increases, which leads to larger-size cells of ascending fluid (figure 3). Thus mesogranular-scale upflows become dominant about 2 Mm below the surface and supergranular-scale upflows dominate at about 10 Mm below the surface.

Bibliography

Brandt P N, Ferguson S, Scharmer G B, Shine R A, Tarbell T D, Title A M and Topka K 1991 Variation of granulation properties on a mesogranular scale *Astron. Astrophys.* **241** 219–26

November L J 1989 The vertical component of the supergranular convection *Astrophys. J.* **344** 494–503

November L J and Simon G W 1988 Precise proper-motion measurement of solar granulation *Astrophys. J.* **333** 427–42

Topka K P and Title A M 1991 High resolution observations of solar granulation *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M S Matthews (Tucson, AZ: University of Arizona Press) pp 727–47

Robert Stein

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

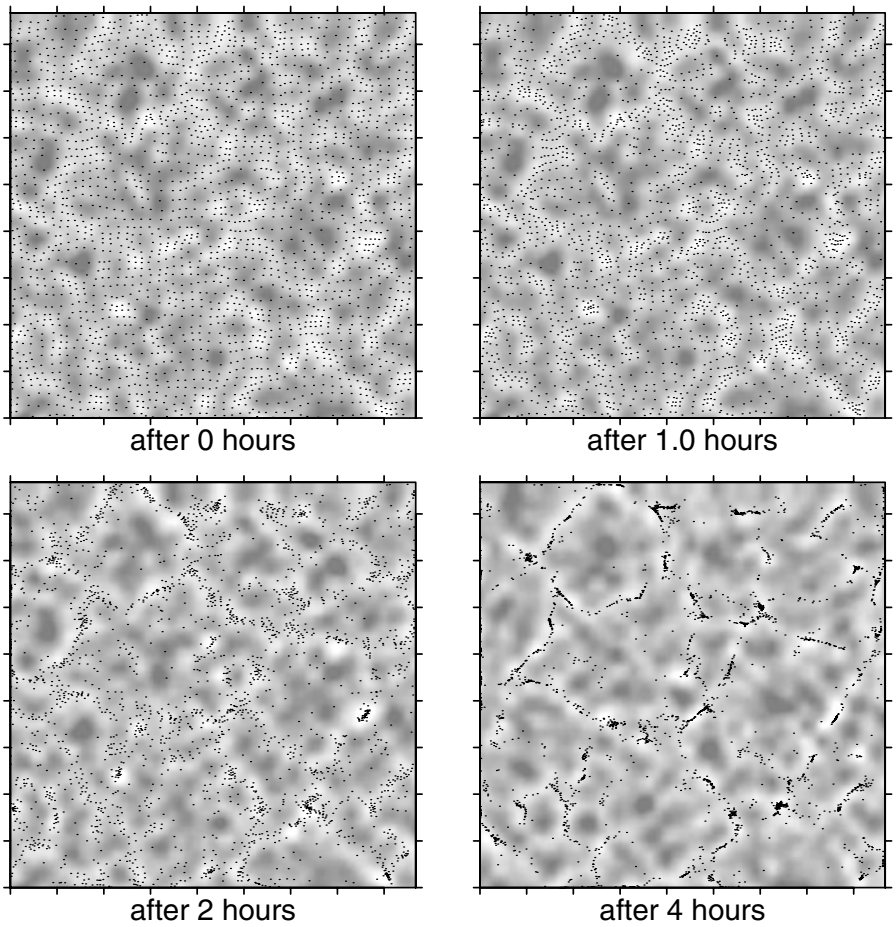


Figure 1. Sequence of images showing the velocity divergence (dark is diverging) and the motion of corks (fluid parcel tracers that move with the horizontal fluid velocity as determined by local correlation tracking). Tick marks denote 10 Mm intervals. After 1 h the corks have collected at the boundaries of the diverging flow cells which mark the mesogranules. After a longer time (4 h), the corks become concentrated on the larger supergranule boundaries.

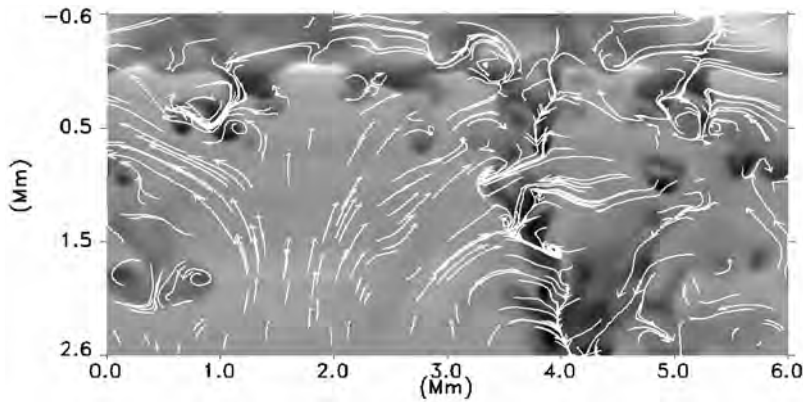


Figure 2. Image of entropy fluctuations and velocity arrows in the x–z-plane. A broad, high-entropy, upflow at depth (a mesogranule) breaks up into smaller upflows near the surface (the granules). A downdraft of low-entropy fluid penetrates the entire computational domain near 4.0 Mm horizontal location. The flow topology is similar to a fountain, with most rising fluid turning over and going back down within a scale height.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

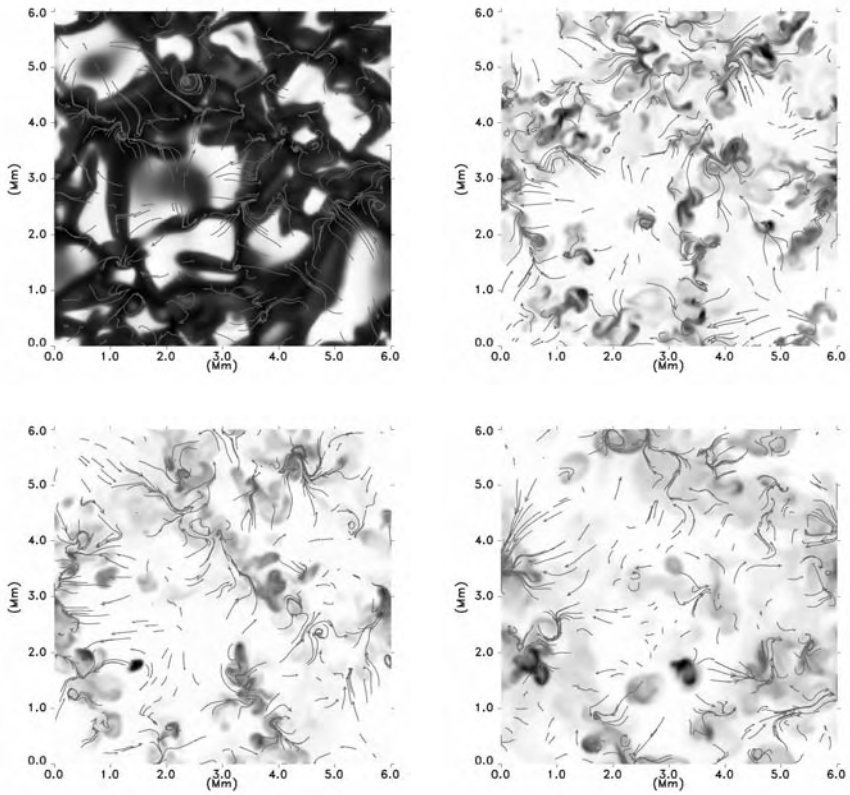


Figure 3. Image of entropy fluctuations and horizontal velocity arrows at four depths: surface, 0.5 Mm, 1.0 Mm, 2.0 Mm below surface (from left to right, top to bottom). Dark is low-entropy descending fluid and light is high-entropy ascending fluid. The scale of the horizontal cells increases with depth from granulation to mesogranulation size.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: Supergranulation

The surface of the Sun is covered by convection cells called GRANULATION, which can easily be seen in white light through a telescope under good SEEING conditions. These cells, typically 1 Mm in diameter, are visible because their centers are hotter (brighter) than their cooler (darker) boundaries. Granules float on top of a much larger convection pattern called supergranulation. Supergranules, with diameters of about 20 Mm, are the tops of plumes (bubbles) which probably form 5–10 Mm below the solar surface in the Sun's convection zone (see SOLAR INTERIOR: CONVECTION ZONE). Supergranulation is invisible in ordinary photographs, but can be seen in Dopplergrams. In a supergranule gas flows radially outward from its center to its periphery. This motion drags with it magnetic flux tubes, which pile up at the supergranule boundary to form a magnetic network (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES) and define the locus of the Sun's chromospheric network (see CHROMOSPHERE: NETWORK).

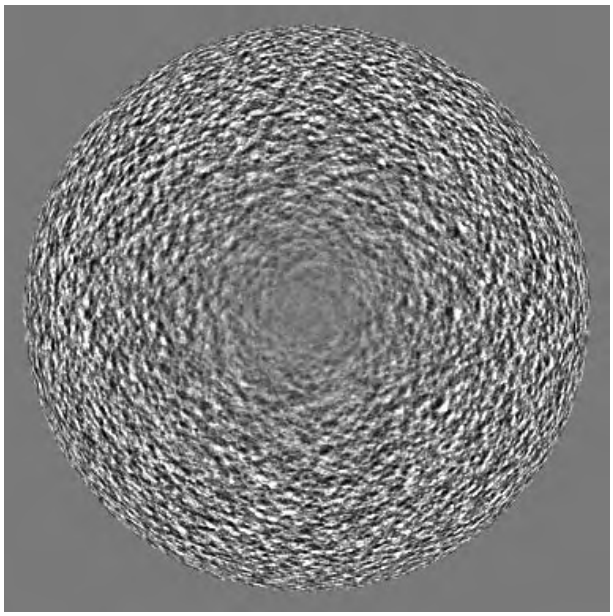


Figure 1. A full-disk Dopplergram of the Sun obtained by MDI. Dark and bright features indicate line-of-sight approaching and receding velocities. Supergranules, not visible at disk center, appear as elliptical structures, each with a dark and a bright half. (Courtesy of R Bogart, Stanford University, and G Simon, Air Force Research Lab.)

Solar convection and mixing-length theory

In 1801 the great English astronomer William Herschel first observed that the Sun's visible surface, the SOLAR PHOTOSPHERE, has a granular appearance. Seventy-five years later the first high-resolution photographs of the Sun revealed that this granulation was composed of closely packed cells having bright (hot) centers surrounded by dark (cool) lanes. In the 1930s the German theoretician Ludwig Biermann realized that, in the outer 30% of the Sun, heat from the interior is transported to the surface by convection, and granules are the surface manifestation (the boiling bubbles) of this convective process. However, theory was incompatible with the observations. According to then-accepted mixing-length theory, convective bubbles should have sizes proportional to the local scale height H (the distance in which density or pressure changes by a factor e $\approx$ 2.7). Thus a large range of cell sizes was expected. Why was it then that only granular-sized bubbles ($\approx$ 1 Mm) were seen, while H ranged over three orders of magnitude, from 100 km at the Sun's surface to 100 Mm at the base of the convection zone?

Discovery of supergranulation

Not until the 1960s was this mystery finally solved, thanks first to an invention by the brilliant American physicist Robert Leighton and second to a simple theory of heat transport through the convection zone. Leighton modified the well-known SPECTROHELIOGRAPH so that it could not only take ordinary brightness photographs of the Sun but also obtain Dopplergrams and Zeemangrams. That is, with the new instrument, he was able to visualize line-of-sight velocities and magnetic fields at the solar surface. Immediately thereafter (in 1961) he and his colleagues made two major discoveries that spawned two new disciplines of solar research. The first was the Sun's five-minute oscillations which opened the field of HELIOSEISMOLOGY. The second was supergranulation, which started the study of solar magnetoconvection. Supergranules were the second scale of solar convection to be discovered. Their area is 400 times that of granules. Mixing-length theory, already seriously weakened because only one scale of convection (granulation) had been observed, was further destroyed by the discovery of supergranulation. Instead of many convective size scales, now two discrete scales of widely disparate sizes had been seen. In 1968 Nigel Weiss and George Simon suggested a model of rising plumes which predicted the existence of three or four cell sizes, the two that had been observed, a giant cell formed at the base of the convection zone and perhaps an intermediate scale. Lawrence November discovered the last of these scales in 1981, calling it mesogranulation, and giant cells were first observed in 1997.

Properties of supergranules

Convection cells typically have depths from 1/4 to 1/2 the size of their diameters; they look like very thick pancakes. So supergranules, which have mean diameters of 20 Mm, are generally believed to originate 5–10 Mm below the solar surface. By the time these hot bubbles reach the surface they have given up their excess heat to the surrounding gas. Hence they show so little brightness contrast between their warmer centers and cooler boundaries that they are invisible in ordinary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

photographs. Only their motions, revealed in the Dopplergrams, can be observed. Recent helioseismology measurements, by the MDI instrument on the SOHO spacecraft (launched in 1995), indicate that supergranules extend about 8 Mm below the surface, exactly in the expected depth range. Helioseismology is permitting the first three-dimensional view of the Sun's subsurface layers and holds much promise for understanding convection processes below the Sun's visible surface. In stark contrast to pancake observations, one theoretical model suggests that supergranules may be tall vertical cylinders, extending from the bottom to the top of the convection zone. Future helioseismological data may be able to discriminate between pancake and cylinder models of supergranules.

Hot gas rises (50–100 m s^{-1}) at a supergranule's center, flows horizontally outwards (200–500 m s^{-1}) and then sinks (100–200 m s^{-1}) at the cell boundary. Most of the downflow occurs at vertices of the irregular polygonal structure of supergranulation. At the Sun's surface one sees primarily the horizontal flow pattern. In a full-disk Dopplergram (figure 1), supergranules are elliptical features, half-bright, half-dark. None is seen at disk center. The round supergranules appear elongated owing to foreshortening. This photograph is explained schematically in figure 2. Since Dopplergrams show only line-of-sight motions, horizontal motions will not be visible at disk center. Away from center, one observes supergranules at a slant angle, so a component of the horizontal motion projects into the line of sight, and appears either dark (approaching) or light (receding). The slant angle of view produces the geometrical foreshortening such that, when viewed from Earth, the supergranule looks smaller (broken arrows) than at disk center. The gas circulation within the supergranulation is shown schematically by the solid arrows. In figure 3 we show a high-resolution Dopplergram, taken at the Sun's south limb by MDI. Here we have enclosed eight typical supergranules in boxes. Each has its darker half closer to disk center (located above the image, at x, $y = 300$ arcsec, 975 arcsec). If the supergranule has a circular cross-section and a horizontal radial outflow from its center, the Doppler effect and foreshortening give it an elliptical shape, with its long axis dividing the dark and bright halves. This long axis is parallel to the solar limb and perpendicular to the line we have drawn along a solar radius from the supergranule toward disk center.

Until recently it has been impossible to measure accurately the lifetime of supergranules, since they were known to live 1 day or longer. Because of the day–night cycle, ground-based observations of supergranules were interrupted at sunset, and by the following sunrise the cell had either died or changed shape sufficiently so that identification was uncertain or impossible. Now the space-based MDI instrument has, on several occasions, provided uninterrupted Dopplergrams (at a 1 min cadence) for 60–90 days. Special feature-tracking algorithms have been applied to these data sets. From measurements of many thousands of supergranules it has been learned that, while most supergranules live between 15 and 30 h, a small number have lifetimes exceeding 2 days, a few even over 4 days.

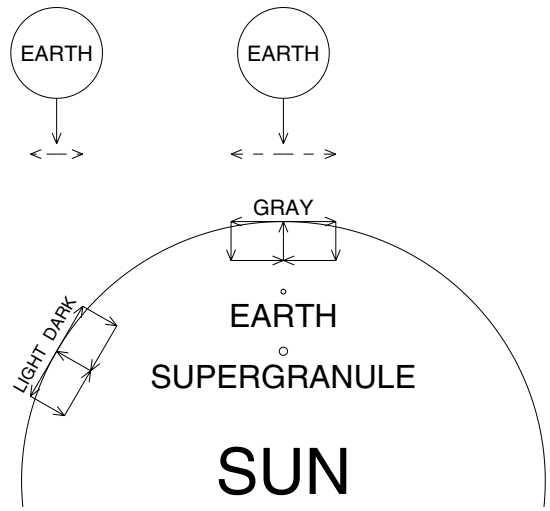


Figure 2. Schematic diagram (enlarged) of supergranules, at and away from disk center, to explain features seen in figure 1. and figure 3. The actual sizes of supergranules and the Earth, relative to the Sun, are shown as small circles.

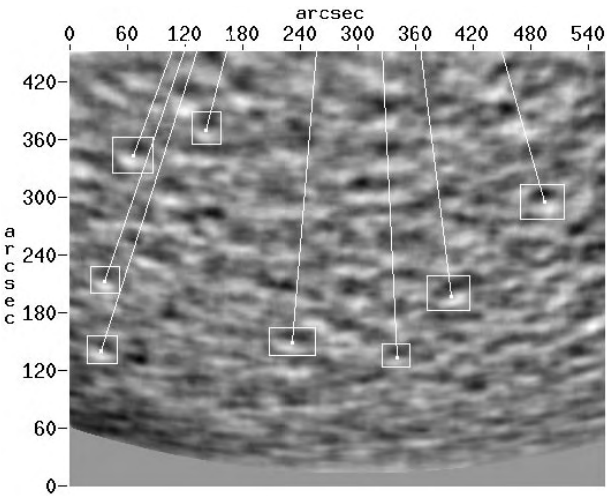


Figure 3. Appearance of supergranules near the solar limb. Eight are enclosed in boxes. Lines connect the centers of these supergranules with Sun center. Note that each line is approximately perpendicular to the line (not drawn) that separates the dark and bright halves of that supergranule. (Courtesy R Shine, Lockheed Martin Solar and Astrophysics Lab, and G Simon, Air Force Research Lab.)

Supergranular advection of granules

The supergranular flows carry granules, much like flotsam is transported in a river. Thus it is possible to observe these flows indirectly, by measuring the motions of the granules,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

using the technique of local correlation tracking (LCT), first developed in the 1980s and since then a powerful tool for describing the Sun's surface velocity structure. Since granules are such small features, LCT only works with ground-based observations under very good seeing conditions (times of low atmospheric turbulence) or with space-based telescopes. Otherwise the granule shapes are distorted or even disappear and cannot be used as supergranular tracers. LCT wonderfully complements Doppler measurements, since it works best at disk center, where Dopplergrams are ineffective. Because LCT can also be used far from disk center where the Doppler signal is strong, it has been possible to confirm that the two techniques give consistent values for supergranular velocities.

The magnetic network

Not only granules but also small magnetic flux tubes are carried by the supergranular flow field. Since supergranules cover the solar surface in a close-packed polygonal-shaped cellular structure, this topology requires that the outflows from adjacent supergranules meet at their common boundary. Some of this gas, which had risen to the surface at the supergranule centers, falls back down along this boundary line. The remainder moves along the boundary to vertices in the pattern, where the downflow is concentrated into small plumes, and often shows a vortex (twisting) motion. The magnetic flux is forced to the cell boundaries by the flows, where it gathers into a network pattern, which is especially concentrated at the vertices. The study of such interactions between convective gas motions and magnetic field is called magnetoconvection. Models of solar magnetoconvection attempt to describe how magnetic flux is carried, concentrated, dispersed, diffused, twisted and stretched by motions of the ionized solar gas and how, at the same time, the magnetic field reacts against this motion, thus altering the nature of the gas flow. While the supergranular interactions described here occur probably at and within 10 Mm of the solar surface, similar forces acting far below the surface, near the base of the convection zone (200 Mm depth), are probably responsible for the formation of SUNSPOTS and the solar dynamo (see DYNAMOS: SOLAR AND STELLAR) which produces the solar activity cycle (see SOLAR CYCLE).

The magnetic network causes localized heating at the solar surface and directly above it in the chromosphere. A chromospheric network, consisting of rising and falling jets of hot gas called spicules, forms at the loci of the magnetic flux. In figure 4, a filtergram taken in the light of Hα, a spectral line of hydrogen that originates in the chromosphere, the spicules correspond to the dark elongated structures. The distinct network pattern resembles hedgerows that separate farm fields. (The complex bright and dark region in the foreground is an ACTIVE REGION containing two small sunspots. This is not part of the network pattern which is visible in the central and upper portions of the filtergram.)

Figure 4. The chromospheric network (seen in this Hα filtergram) is coincident with a network of magnetic flux elements and marks the boundaries of supergranules. The network consists of dark fibrillar vertical structures (spicules). Black crossed lines indicate the long and short axes of two foreshortened supergranules. (Courtesy of R Dunn, National Solar Observatory.)

Magnetic diffusion and kinematic modeling

Knowledge of the supergranule lifetime is important for studies of solar magnetoconvection. It is believed that large-scale long-term transport of magnetic field across the Sun is related to the solar activity cycle. This transport may be due to a diffusion mechanism controlled by the supergranulation. As we have pointed out, the supergranular flow concentrates magnetic flux into a network pattern. This pattern undergoes little change during the supergranule's life, but when the cell disappears it is replaced by a new supergranule 5, 10 or 20 Mm removed from the former cell. The magnetic network must now rearrange itself to conform to the flow pattern of this new supergranule. The continual birth and death of supergranules creates a random-walk diffusion that steps the magnetic field across the Sun. This diffusion is directly proportional to the square of the step size, and inversely to the time between steps, i.e. to the area of a supergranule divided by its lifetime.

Kinematic modeling of solar magnetoconvection (in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

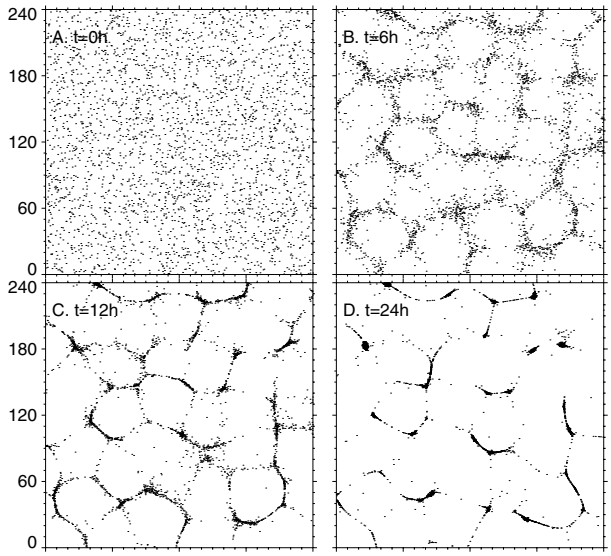


Figure 5. Formation of the magnetic network. In this numerical simulation, supergranular flows carry an initially uniform distribution (A) of tracer particles (representing magnetic flux tubes) to the boundaries of the supergranules. After 6 h (B) the evolving network can easily be recognized, and it is almost fully formed in 12 h (C). After 24 h (D) the network is very thin, and beginning to concentrate into small clumps at vertices of the pattern.

which the magnetic field does *not* react back against the velocity field, and thus the magnetic flux tubes act as simple tracers of the flow) can be used to illustrate the concentration, dispersal and diffusion of magnetic flux at the solar surface. If one uses the observed sizes, lifetimes and velocity field of supergranules to create a slowly changing artificial grid of convection cells, one can observe the effect of this evolving flow pattern on an initially uniform randomly distributed set of tracer particles (called *corks*) and compute the rate of diffusion. Results from such a model calculation by Simon, Title and Weiss are shown in figure 5. The visualization shows the evolution of the cork pattern over 1 day, from a uniform random distribution to a well-defined network structure, with concentrations of particles forming at vertices of the flow pattern. While they are not correct representations of magnetoconvection of strong magnetic fields, these models are sufficiently accurate for weak fields. They can be quickly computed for large areas of the solar surface and thus provide useful descriptions of magnetoconvection and magnetic diffusion on the quiet Sun, away from sunspots and active regions.

Bibliography

Leighton R B, Noyes R W and Simon G W 1962 Velocity fields in the solar atmosphere. I. Preliminary report *Astrophys. J.* **135** 474–99

Simon G W and Weiss N O 1968 Supergranules and the hydrogen convection zone Z. *Astrophys.* **69** 435–50

Simon G W, Title A M, Topka K P, Tarbell T D, Shine R A, Ferguson S H, Zirin, H and the SOUP Team 1988 On the relation between photospheric flow fields and the magnetic field distribution on the solar surface *Astrophys. J.* **327** 964–7

Simon G W, Title A M and Weiss N O 1995 Kinematic models of supergranular diffusion on the sun *Astrophys. J.* **442** 886–97

George Simon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photosphere: The Limb Effect and Gravitational Redshift

The 'limb effect' is the observational finding that many spectral absorption lines formed in the photosphere (FRAUNHOFER LINES) show a REDSHIFT of their mean wavelength when comparing limb observations with observations from the solar disk center. The magnitude of this effect is different for each spectral line and equals some 100 m s^{-1} when explained as a DOPPLER EFFECT.

The 'gravitational redshift' is the redshift of spectral lines originating near a mass concentration as described by Einstein's theory of GENERAL RELATIVITY.

Both historic terms are related to wavelength shifts and spectral line shape changes when analysing the center-to-limb variation of the spectral absorption lines formed in the photosphere (see SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION), which is the topic of this article.

Wavelength shifts and line shapes

Using present knowledge of the physical conditions within the SOLAR PHOTOSPHERE one expects several effects on the wavelengths and line shapes of the spectral lines formed there, which depend on each other:

(a) Due to the gravitational force on the Sun, which is much larger than that on the Earth, the wavelengths of spectral lines are shifted to longer wavelengths as compared with the same spectral lines on Earth. The effect amounts to an equivalent Doppler shift of 636 m s^{-1}. This amount is the same at the disk center and at the solar limb because the gravitational redshift depends on the gravitational potential energy at the emitting location. This energy is the same everywhere on the surface of the Sun (which is an equipotential surface).

(b) The convective motions, especially those of the GRANULATION, which exhibit upward motions of about 2 km s^{-1} in their centers and downflows in the intergranulation, cause spectral line wavelength shifts. A shift towards shorter wavelengths due to the Doppler effect can be expected for spectral lines originating in the central region of the disk, because most of the light originates from bright granules moving upward.

(c) Each solar absorption line is formed within a certain layer of the photosphere. The line centers are formed higher in the atmosphere than the wings. Since the pattern of convective motions in the photosphere depends on the depth in this atmosphere and the velocity components depend on the position on the solar disk, besides wavelength shifts spectral line asymmetries also occur. These effects depend on the selected spectral line and in addition show a center-to-limb variation, which means that the spectral line shape is most often different in the central disk region as compared with that observed near the limb.

To investigate the line shifts, those caused by large-scale motions, like solar rotation or supergranulation, have to be removed, either by correction or by averaging in space or time.

Several methods are used to determine spectral line positions from digital line profile data, only two of which are mentioned here: (i) a parabola can be fitted to the lower third of the spectral line and the deepest point of that parabola represents the position of the spectral line; (ii) the bisector points (a definition and examples are given in this article) can be used to fit the spectral line position.

Historic observations

The first detection related to these effects was published by J Halm in 1907, who found that the wavelengths of two iron lines in the solar spectrum were larger at the solar limb as compared with the disk center. The shifts found by Halm amounted to Doppler velocities of about 600 m s^{-1} and had the same magnitude at the poles and at the equator of the Sun. This finding is nowadays called the 'limb effect'. It was confirmed in 1907 by W S Adams, who later measured the limb effect for many other spectral lines.

After the successful testing of Einstein's theory of general relativity, attempts were made to measure the redshift of solar spectral lines as compared with the same lines produced on Earth. Indeed C E St John found in 1928 that the spectral lines observed at the solar limb showed about the expected redshift due to the gravitation, but the redshift in the disk center was only about half of that amount. He was the first to explain the observed limb effect as compared with a reference wavelength on Earth as a combination of convective motions and gravitational redshift.

After 1950 when the first models, which included convective motion, were created to compute solar spectral lines, it became evident that besides the upward motion of hot gas within the granules, downward motion of cooler gas also had to be taken into account. At that time these different motions could not yet be detected by spectroscopic methods, but the models suggested that another effect should be the result of the mixture of upward and downward motions, namely the asymmetry of most solar spectral lines and in addition a center-to-limb variation of the asymmetry. When the detection of the asymmetry of solar spectral lines was reported, it was first doubted. Only after 1960 was the observational material good enough to determine more precisely the amount of asymmetry: the bisector of a spectral line, which is the average of the wavelengths at which the line flanks on the blue and the red side have the same measured intensities, exhibits for many spectral lines a shape like the character 'C'. Therefore it was later called the 'C shape' of a spectral line. When taking the deepest point of the spectral line, where the bisector crosses the spectral line profile, as the wavelength reference, the bisector shows a maximum difference of less than 1 ppm towards the shorter wavelength. Such small deviations from the symmetric profile could only be measured by electronic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

methods and taking into account possible falsifications of the spectrometer used. In 1975 a first collection of line shapes and their center-to-limb variation for about 100 different solar spectral lines in the wavelength range from 575 nm to 660 nm was published by E A Gurtovenko *et al*.

Historic models

The Einstein shift of solar spectral lines is given by

$$\Delta\lambda\,\lambda^{-1} = GMR^{-1}c^{-2} = 2.12 \times 10^{-6}$$

with G the constant of gravitation, M the solar mass, R the solar radius and c the velocity of light. For all absolute wavelength comparisons this effect has to be included in the computations. Since the absolute comparison of wavelengths with a precision of better than 1 ppm is still a problem, in general only relative comparisons of wavelengths are performed.

In 1954 K H Böhm included the convection in a model for computing solar spectral lines in the photosphere, as Doppler shifts of upward moving gas, downward moving gas and static gas. H H Voigt in 1956 added a higher layer without vertical motions to this model. This model allowed spectral line asymmetries to be computed. In 1957 E H Schröter published a model, which could explain the limb effect with only two moving components of gases. The static component was not necessary in his model.

All these models support the basic idea that the observed limb effect as compared to laboratory reference wavelengths is caused by the gravitational redshift, which is reduced on the disk, especially near its center, by the vertical upward motion of the hot granules. For strong Fraunhofer lines, like for example Na D, Mg b and the K line at 770 nm, no limb effect is found, because they are formed in higher layers, where no vertical motions are present. But these lines exhibit the gravitational redshift.

Although the explanation of the solar spectral line asymmetry and the relative limb effect by mainly convective motions seems to be rather successful, other mechanisms have been discussed. A few of them will be given in historical order:

(a) When J Halm published the detection of the limb effect in 1907 he also gave his explanation that pressure differences should be the cause of the limb effect and that these pressure differences should change during the SOLAR CYCLE. He explicitly argued against the possibility of explaining the limb effect by gas motions.
(b) In 1967 G Erikson and P Maltby discussed the influence of sound waves on the spectral line wavelengths, but they could only reduce the gravitational redshift at solar disk center and did not obtain the observed blueshifts for any line.
(c) In 1974 M H Hart tried to include a Lennard-Jones potential for the pressure to explain the limb effect. It was only possible to explain some sodium lines using his model, but it failed for all other spectral lines. Directly related to Hart's publication was a publication by J M Beckers and P de Vegvar in 1978. They showed that pressure effects could not be used to explain the limb effect of iron lines.
(d) In 1977 H H Plaskett combined several mechanisms: the motions were the most important, but he included also a pressure effect and correction for a gravitational effect.

Further investigations

Besides the vertical motions of the solar granular gas, horizontal motions were also detected. By including these horizontal motions of granular gas in their model, J M Beckers and G D Nelson could explain in 1978 two special details of the limb effect of spectral lines:

(a) The limb effect shows for many lines a minimum of the wavelength not in the disk center but near to the center-to-limb position of $\cos\theta = 0.8$, where θ is the angle between the line-of-sight and the vertical direction on the Sun. This will be discussed again later and is also shown in figure 3.
(b) Often a redshift of the spectral lines is found at the solar limb, which is larger than the Einstein shift. This effect is called supergravitational redshift.

Only after 1980 was it possible, as for example in work by Å Nordlund, to solve numerically the hydrodynamic equations under conditions typical of the solar photosphere and compute the convective motions self-consistently, including their depth dependence. Using these models it was possible to compute spectral lines which show the observed C shapes and also their center-to-limb variation. Some details of these models and an investigation of the C shape of many iron lines from a digital solar spectral atlas were published by Dravins *et al* in 1981 and also presented in a review on spectral line asymmetries and wavelengths shifts by Dravins in 1982.

Within more recent investigations, the number of spectral lines used from different elements and of different excitation potentials increased. In addition, the investigations concentrated on the formation heights of the spectral lines. Balthasar (1984) published results from analyses of 143 selected lines, which were observed with a Fourier transform spectrometer (FTS) in the wavelength region 500 nm to 630 nm: he described three different classes of C shapes and center-to-limb variations. In figures 1 and 2 examples of typical spectral line C shapes of the three classes are given for the disk center and a limb position respectively.

Although the number of spectral lines within each class was quite different, he found a convincing dependence on their formation heights and their lower excitation potentials. Some general findings are:

(a) There are a few spectral lines formed in deep layers, which show a red asymmetry (right part of figure 2).
(b) Lines formed in higher layers show larger asymmetries in the disk center.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

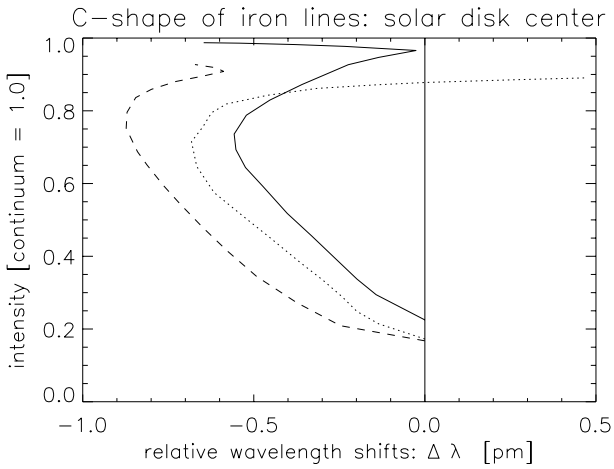


Figure 1. The C shapes of typical iron lines in the disk center: Fe 524.25 nm (full curve), Fe 507.47 nm (dotted curve), and Fe 507.97 nm (dashed curve). The wavelength positions of the deepest points in the spectral lines are taken as reference. The spectral lines are typical examples of classes I, II and III of Balthasar (1984). The plot was produced using data obtained from Balthasar.

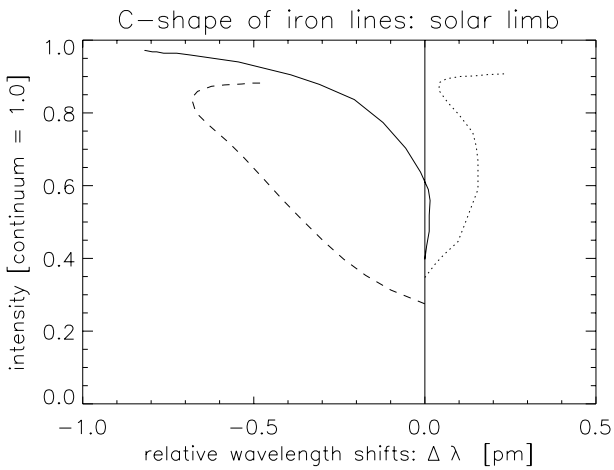


Figure 2. The C shapes of typical iron lines at the solar limb (at a position of $\cos\theta = 0.112$): Fe 524.25 nm (full curve), Fe 507.47 nm (dotted curve), and Fe 507.97 nm (dashed curve). The wavelength positions of the deepest points in the spectral lines are taken as reference. The spectral lines are typical examples of classes I, II, and III of Balthasar (1984). The plot was produced using data obtained from Balthasar.

(c) When reducing all wavelength shifts to 500 nm it was possible to determine a limb effect curve for 96 lines of neutral iron. The best fit was possible for a variable $(1-\cos\theta)$:

$$\Delta\lambda(\theta) = -0.027 - 0.320(1-\cos\theta) + 1.212(1-\cos\theta)^2,$$

with the shifts given in pm. This relation and that of subsamples of spectral lines formed deeper and higher in the photosphere is given in figure 3. (The fit procedure allowed small deviations from a zero value in the disk center.)

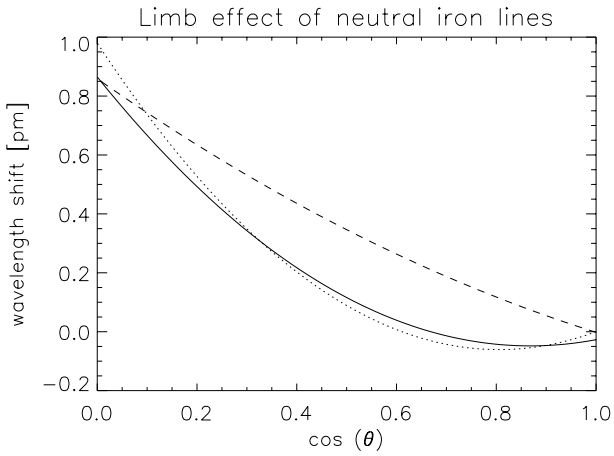


Figure 3. The mean limb effect for the 96 spectral lines of neutral iron is given as the full curve. The mean limb effect of a subsample of 21 lines formed in deep layers ($\log\tau = 0$ to -1) is given as the dotted curve. The mean limb effect of another subsample of 10 lines formed in high layers ($\log\tau = -3$ to -5) is given by the dashed curve. The plot was produced using data from Balthasar (1984).

In 1985 Balthasar published another paper, where he used a subsample of 20 iron lines from the FTS spectra at disk center, which are formed at different depths. He corrected their wavelengths for a laboratory reference wavelength. For this sample he found the following numerical value of the vertical velocities, depending on their formation depths:

$$v_{\rm ver} = 684 - 1922.8\exp(1.3577\log(\tau))({\rm m\ s^{-1}})$$

It is interesting that there are already some spectral lines showing a redshift larger than the gravitational redshift. For the horizontal velocities, depending on their formation depths, he found from this sample

$$v_{\rm hor} = (510 \pm 23) - (199 \pm 8)\log(\tau)({\rm m\ s^{-1}}).$$

The contribution of the horizontal motions to the line asymmetries is explained by a geometrical model, taking the roughness of the atmosphere into account.

Further problems

Although the occurrence of solar spectral line shifts and their C shapes is now quite well understood, there are several findings which need further investigation:

(a) For the non-split iron line at 557.61 nm P N Brandt and E H Schröter in 1982 found a more pronounced limb effect on the east–west diameter as compared with the north–south diameter of the Sun, while J M Beckers and W R Taylor had found in 1980

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

just the opposite relation—which they explained as a meridional motion towards the poles.

(b) Since reports exist about changes of the granulation during the solar activity cycle, it can be speculated that there should also be changes of the limb effect, the C shapes of spectral lines and their center-to-limb variation during the activity cycle. Maybe the finding described under (a) is an example.

(c) For the modeling of C shapes the real formation heights of spectral lines in the photosphere are still not well enough known. The atomic parameters, the physical conditions of the photosphere and their changes have to be investigated further. These uncertainties also influence the results of 3D modeling.

(d) Using the best spatial resolution available, it has been possible for some years to determine line profiles of different parts of the granules and the intergranular space separately. This topic is a very promising tool for understanding both the granulation and the spectral line formation within it much better.

Acknowledgments

The author thanks H Balthasar and M Stix for discussions of the text and H Balthasar for data supplied to produce figures 1 and 2.

Bibliography

Balthasar H 1984 Asymmetries and wavelengths of solar spectral lines and the solar rotation determined from Fourier-transform spectra *Solar Phys.* **93** 219–41

——1988 The centre-to-limb variation of solar spectral lines *Astron. Astrophys. Suppl. Ser.* **72** 473–95

Dravins D 1982 Photospheric spectrum line asymmetries and wavelength shifts *Ann. Rev. Astron. Astrophys.* **20** 61–89

Dravins D, Lindegren L and Nordlund Å 1981 Solar granulation: influence of convection on spectral line asymmetries and wavelength shifts *Astron. Astrophys.* **96** 345–64

Nordlund Å 1984 Modelling of small-scale dynamical processes: convection and wave generation *Small-scale Dynamical Processes in Quiet Stellar Atmospheres* ed S Keil (Sunspot, NM: National Solar Observatory) pp 181–221

Hubertus Wöhl

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photospheric Magnetic Flux Tubes: Observations

The magnetic field on the solar surface (SOLAR MAGNETIC FIELD) is highly filamentary, with a considerable fraction being in the form of flux tubes (roughly speaking bundles of concentrated field lines). Small-scale flux tubes have diameters below approximately 400 km. They appear bright and are the basic magnetic features underlying the magnetic network and active region plages (SOLAR CHROMOSPHERIC PLAGE).

Averaged over the solar disk the magnetic field strength varies between a few gauss and a few tens of gauss, whereas the field strength inside the flux tubes, B, is around 1500 G at the solar surface. Consequently, flux tubes cover only of the order of 1% of the solar surface. Although a weak and probably turbulent field (SOLAR PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS) is expected to fill the space between the flux tubes, they nevertheless contain the dominant fraction of the magnetic energy, since it is proportional to B^2. In addition, flux tubes, in particular the small-scale ones, are the main source of energy transfer from the solar interior to the Sun's outer atmosphere and hence are important players in the heating of the solar CHROMOSPHERE and solar CORONA.

Magnetic flux tubes with a wide range of sizes or cross-sectional areas are present on the Sun. The largest and best known are SUNSPOTS (diameters 4000–50000 km). Somewhat smaller flux tubes are called pores (diameters roughly in the range 400–5000 km; SUNSPOT PORES), with the smallest pores often also referred to as magnetic knots. Finally, the smallest magnetic flux tubes (with diameters less than approximately 400 km) are called magnetic elements and are the main subject of this article. Whereas sunspots are clearly distinguished from pores by the presence of a fully developed penumbra in the former, the distinction between knots and normal pores or between knots and magnetic elements is not so clear-cut. In addition to the size this distinction rests on differences in brightness. Magnetic elements are brighter than knots, which in turn are brighter than pores. Magnetic elements emit a larger radiative flux than the quiet Sun (i.e. the field-free part of the solar atmosphere), while larger flux tubes are at least slightly darker. A high-resolution magnetogram is plotted in figure 1, together with the white-light image of the corresponding region. Examples of magnetic features that are bright (e.g. at 8″ horizontal and 4″ vertical) and dark (at 11″ and 11″) are present in white light in the images.

The basic theoretical picture of each of these types of feature is that of a flux tube (or possibly a tight bundle of very small flux strands) which passes more or less vertically from the solar interior into its atmosphere, so that it is seen in cross section in magnetograms or in other images of the solar surface. In principle, the magnetic field of the flux tube may be twisted, in which case it is called a magnetic flux rope.

Flux tubes are not directly accessible to observations in the interior of the Sun, are seen in cross section at the surface at wavelengths ranging from the ultraviolet to the infrared and are outlined by x-ray emitting gas in the corona (SOLAR SPECTROSCOPY: ULTRAVIOLET AND EXTREME ULTRAVIOLET EMISSION, VISIBLE EMISSION, INFRARED EMISSION, SOLAR SPECTROSCOPY AND DIAGNOSTICS: X-RAY EMISSION). The magnetic field of flux tubes can be best studied in the photosphere due to the significant Zeeman splitting of spectral lines in the photospheric spectrum (SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY), the relatively simple formation physics of these spectral lines and the high spatial resolution achievable in the visible. Also, at greater heights the field from the flux tubes has expanded, so that neighboring flux tubes have generally merged, making it difficult to separate and identify individual flux tubes in these layers.

Sunspots, pores and magnetic elements differ in numerous respects. The largest sunspots possess 10^5–10^6 times as much magnetic flux as magnetic elements. Sunspots and pores are dark relative to the quiet Sun, whereas magnetic elements are on average bright, particularly when observed near the solar limb. Sunspots, and to a large extent also pores, are present only in SOLAR ACTIVE REGIONS, which are restricted to the activity belts (roughly within 30° of the solar equator), whereas magnetic elements are present at all latitudes. They are not completely homogeneously distributed, however. The number density of magnetic elements is considerably larger in active regions (where they cluster to form plages or faculae) than in the quiet Sun. Even outside active regions the magnetic elements are concentrated at the boundaries of supergranule cells (SOLAR PHOTOSPHERE: SUPERGRANULATION) forming the so-called magnetic network (CHROMOSPHERE: NETWORK). Flux tubes can also be present inside the supergranule cells, but these intranetwork fields generally have a smaller magnetic flux per flux tube and also lower intrinsic field strengths than their counterparts in the network. At a smaller spatial scale magnetic elements are observed to prefer the dark downflow lanes of the granulation (SOLAR PHOTOSPHERE: GRANULATION) and to avoid the bright, upflowing granules.

The photospheric layers of magnetic flux tubes ranging from sunspots to magnetic elements possess remarkably similar field strengths (averaged over the flux-tube cross section). At the solar surface the field strength lies between 1200 and 1800 G, with magnetic elements having a field strength of around 1500 G.

The brightness of magnetic elements makes them important contributors to the variable component of SOLAR IRRADIANCE. The magnetic network and active region faculae or plages owe their excess brightness at all atmospheric layers to excess radiation from the magnetic elements. Numerous lines of evidence suggest that they are dominantly responsible for the fact that the Sun is brighter at the maximum of magnetic activity, although in that phase of the SOLAR CYCLE the number of dark sunspots on the solar surface is largest.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

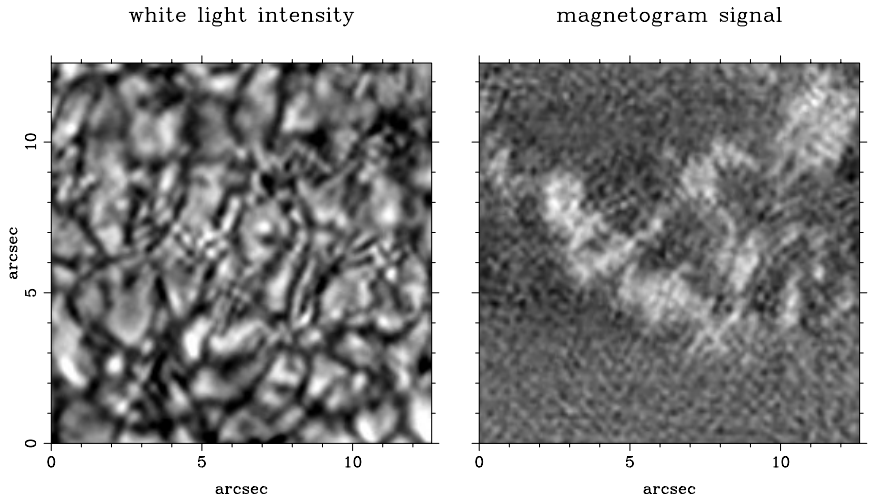


Figure 1. High-resolution white-light image (left) and magnetogram (right). Regions with a magnetic field (i.e. with significant net circular polarization) appear bright in the magnetogram. The pattern of relatively bright cells surrounded by dark lanes in the white-light image are granules. Note the increased small-scale structure at the location of the magnetic field. (Figure kindly provided by C U Keller.)

Two-component model of magnetic elements

An important parameter determining the way in which a solar feature is described by empirical modelers is its size relative to the spatial resolution element of typical observations. If the feature is spatially resolved and does not show significant internal structure then one-component, i.e. purely plane-parallel, modeling is sufficient. Solar magnetic elements, however, are usually not completely resolved, so that two components are required. The first component describes the magnetic flux tubes, the second their (relatively) field-free environment.

Since the Zeeman effect (SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS) not only splits spectral lines, but also polarizes them, the polarization in spectral lines, described by the Stokes parameters, serves as a diagnostic of flux tubes. For a two-component model the observed Stokes profiles (which are averaged over the spatial resolution element) can be written as

$$I_{\rm obs} = \alpha I_{\rm m} + (1-\alpha) I_{\rm s} \quad (1)$$

$$P_{\rm obs} = \alpha P_{\rm m} \quad (2)$$

where $I_{\rm m}$ is the intensity profile arising from the magnetic feature, $I_{\rm s}$ is the intensity from the surroundings, α is the magnetic filling factor (i.e. the fraction of the surface area within the spatial resolution element covered by magnetic field) and $P = Q$, U or V represents in turn the two linearly and the circularly polarized Stokes parameters. It is clear from these equations that in the context of this simple model $Q_{\rm obs}$, $U_{\rm obs}$ and $V_{\rm obs}$ are formed in the magnetic element only. Hence it is possible to obtain information on unresolved flux tubes by observing one or more of the polarized Stokes parameters.

Since the magnetic filling factor is generally far smaller than unity, the intrinsic field strength B of magnetic elements is correspondingly larger than the spatially averaged field strength $\langle B \rangle$. The latter quantity is a measure of the magnetic flux in the spatial resolution element. Typical values of the filling factor α are 1% or less in the quiet Sun, 5–10% in the network and 10–30% in active-region plage.

Flux tube sizes and magnetic fluxes

The true sizes of magnetic elements are highly controversial, since they are either close to or below the size of the spatial resolution element, which can, in the best cases, be as small as 150 km on the Sun, corresponding to an angular resolution of roughly 0.2″. One problem is that achieving this high a resolution in magnetograms is extremely difficult. Standard techniques such as freezing the distortions due to the Earth's turbulent atmosphere by making short exposures fail due to the low signal level in magnetograms (these receive orders of magnitude fewer photons than the brightness images, which can have a much broader wavelength band). Hence most estimates of the sizes of magnetic elements are indirect, based on the sizes of features in an image taken in some proxy of the magnetic field (e.g. bright points in a filtergram taken in the core of a spectral line) under the assumption that there is a one-to-one correspondence between the two sizes. This assumption is difficult to test and is, due to the strong dependence of the contrast of magnetic features on their size, at the most only partly correct.

Magnetic elements probably have no unique size. For example, the sizes of bright points in an active region have been found to be log-normally distributed (just as the sizes of sunspots), with an average value of 220 km. Also, the center-to-limb variation (SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION) of the continuum contrast of faculae (SOLAR PHOTOSPHERE: FACULAE) is best reproduced if they are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

composed of a mixture of flux tubes of different sizes (see the section on thermal structure). Direct observations with as high a spatial resolution as possible as well as the comparison of spectral diagnostics with theoretical models suggest that on average magnetic elements are somewhat larger in regions with a greater magnetic filling factor (e.g. in active-region plage) than in regions with a lower average field strength (e.g. in the quiet Sun network). In other words the average diameter of magnetic elements increases with α or $\langle B \rangle$. This relation may break down for the smallest $\langle B \rangle$ values, where the intrinsic field strength, B, instead depends on $\langle B \rangle$.

An open question is whether individual magnetic elements are themselves formed of smaller flux strands and, if yes, what the size of these flux strands is. It has been argued that such flux strands are indeed present and that they are far below the photon mean-free-path in diameter, making them less than 10–20 km broad. At present there is insufficient evidence to confirm or reject this hypothesis.

The magnetic flux per element varies between typically 5×10^{16} Mx for intranetwork elements (i.e. weak magnetic features present within supergranules) over 5×10^{17}–5×10^{18} Mx for typical (bright) magnetic elements to increasingly larger values for points, knots, pores and sunspots. Note that because the intrinsic field strength is almost the same for differently sized flux tubes there is an almost linear relation between flux tube size and magnetic flux. Therefore, the flux per magnetic element also slowly increases with $\langle B \rangle$.

Magnetic field strength and orientation

The most accurate field strength values are probably provided by observations in the infrared (at 1.5648 μm) due to the linear dependence of the Zeeman sensitivity of spectral lines on wavelength. (The Zeeman sensitivity is given by the ratio of Zeeman splitting to Doppler width of the line.) Recent inversions of spectral lines in the visible should also give relatively accurate results. The greatest currently achievable accuracy is roughly 20 G at a given geometrical height in the atmosphere, corresponding to a relative uncertainty of 1–2%. This accuracy allows the small variations in field strength as a function of magnetic filling factor α to be measured. In figure 2 the intrinsic field strength B of magnetic features is plotted versus the spatially averaged field strength $\langle B \rangle$, which is proportional to α over most of the plotted range.

In active regions, i.e. for $\langle B \rangle \gtrsim 100$ G, B increases very slowly with $\langle B \rangle$. At significantly smaller fluxes (below approximately 25 G, corresponding to the network and intranetwork), however, B depends strongly on $\langle B \rangle$. The features with smallest $\langle B \rangle$ are most likely intranetwork magnetic elements. They have field strengths of only a few hundred gauss.

This observed dependence of B on $\langle B \rangle$ is predicted relatively accurately by models describing the formation of intense photospheric flux tubes by the convective collapse mechanism (SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: THEORY).

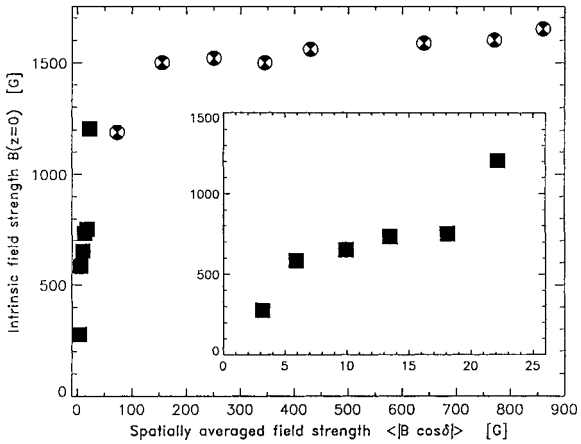


Figure 2. Intrinsic magnetic field strength at the solar surface $B(z = 0)$ versus the unsigned, spatially averaged longitudinal field strength $\langle |B \cos \delta| \rangle$, where δ is the angle between the magnetic vector and the line-of-sight. Plotted are binned values resulting from quiet Sun (squares) and active region plage observations (circles). The inset shows only the leftmost part of the main figure, with the squares being identical to those in the main figure.

Due to the requirement of horizontal pressure balance (which is a consequence of the so-called thin-tube approximation) field strengths of 1500–1700 G at the solar surface imply that the flux tubes are significantly evacuated, i.e. that the gas pressure and density is less than roughly 20–30% of that in the surroundings (at the same geometrical height). Hence, like sunspots, magnetic elements are also thought to possess a Wilson depression, although this has not been directly confirmed (SUNSPOTS: WILSON EFFECT).

This large evacuation implies that the magnetic energy density, $B^2/8\pi$, in the flux tubes is considerably larger than the energy density of the gas inside the flux tube. Equivalently, the plasma beta $\beta = 8\pi p/B^2 < 1$ in the flux tube (where p is the gas pressure). The measured B implies $\beta \approx 0.3$. On the other hand, the high evacuation of the flux tubes also means that the magnetic energy density in the flux tubes is similar to the energy density of the gas in the surroundings and much larger than the average kinetic energy density of photospheric motions. For the intranetwork elements the situation is quite different. The intranetwork fields are in equipartition with the kinetic energy density (i.e. equality of magnetic and kinetic energy density) and thus only weakly evacuated.

The 1.56 μm lines of neutral iron, on which figure 2 is based, are formed deep in the solar atmosphere, near the continuum-forming layer. When spectral lines formed at greater heights are observed then lower values of B are obtained. Consider lines formed at increasingly greater heights in active region plage: the Fe I 6302 Å line gives on average 1400 G, Fe I 5250.2 Å gives 1000–1200 G and the Mg I 12.32 μm line (formed in the upper photosphere just

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

below the traditional temperature minimum) gives 200–500 G. This decrease of B with height is consistent with the roughly exponential height dependence predicted by horizontal pressure balance, suggesting that this is indeed the confining mechanism of the field. The observations therefore suggest that the thin-tube approximation is a good description of magnetic elements.

Theory also predicts that the magnetic field expands with height. This follows from the decrease of the field strength with height and magnetic flux conservation. Various observations confirm this prediction. For example, magnetograms recorded in lines formed at chromospheric heights are far more homogeneous than those recorded in lower-lying, photospheric lines. Also, the spectral shapes of the polarized Stokes profiles of Zeeman sensitive lines formed in the upper photosphere or over a large height range reveal the expansion of the field.

All in all, measurements of the magnetic field strength support the description of magnetic elements as flux tubes shaped by pressure equilibrium.

By using all four Stokes parameters it is possible to determine the orientation of the magnetic vector (with some remaining ambiguity in the orientation of the field component perpendicular to the line of sight). Such observations reveal that magnetic elements have an average inclination of approximately 10° or less to the vertical in both active region plages and the network. Isolated, compact groups of magnetic elements, particularly in active regions, appear to form rosettes, i.e. the flux tubes near the center of the group are vertical, while with increasing distance from the center the flux tubes become increasingly inclined to the vertical. These rosettes have been called azimuth centers. In active regions they are often associated with considerably larger filling factors than the average plage.

There is also evidence that flux tubes with weaker magnetic fields are often considerably more inclined than the kG flux tubes. Thus the weaker intranetwork fields appear to be directed almost isotropically.

A special case of weak fields that are predominantly horizontal are freshly emerging fields. The tops of loops just emerging from the convection zone into the atmosphere are horizontal and have a strength of a few hundred gauss. These freshly emerged fields exhibit a nearly linear relationship between B and inclination. It is speculated that the field intensifies and becomes increasingly vertical as time elapses after its emergence.

Thermal structure

The following three points summarize the main observed thermal properties of magnetic elements.

(a) Small flux tubes are hotter and brighter than larger flux tubes. Equivalently, magnetic elements in regions with small magnetic filling factor are hotter than in regions with large filling factor.
(b) In the continuum-forming layers magnetic elements are cooler than the average quiet Sun *at the same geometrical depth*, but are hotter in the higher layers of the photosphere and in the chromosphere. When converted to equal optical depth, however, the smaller magnetic elements turn out to be brighter than the quiet Sun in all observable layers.
(c) The chromospheres of magnetic elements are particularly hot and bright. The chromospheric temperature rise starts 200–300 km below where it is located in standard models of the average quiet Sun atmosphere.

In the following the above points are discussed in greater detail. The thermal structure of magnetic elements depends sensitively on their size, particularly in the lower photosphere. The continuum-forming layer in small flux tubes is generally hotter than the average quiet Sun. Consequently, these flux tubes appear bright. This is the case, for example, in the magnetic network. As the magnetic flux and hence the average size of the magnetic features increases they become increasingly darker (examples of bright and dark magnetic features are visible in figure 1). Although this qualitative dependence of continuum or white-light brightness on size has been confirmed by different types of observations the absolute value of the brightness is far more controversial. Values of the continuum contrast (i.e. the ratio of the continuum brightness of flux tubes to that of the quiet Sun) range between 1.8 and 0.9 for typical flux tubes. This is because the contrast also depends strongly on other properties such as wavelength, position on the disk and spatial resolution. The highest contrast values are usually obtained at high spatial resolution in broader and bluer wavelength bands, which are more contaminated by spectral lines and thus do not entirely reflect the continuum contrast. A critical assessment of the different determinations in the literature suggests that continuum contrast values of magnetic elements at 500 nm probably lie between 0.9 and 1.3.

Spectral lines are formed at greater height in the atmosphere than the continuum and they clearly indicate an increased contrast. This is illustrated by the G-band (430 nm) image shown in figure 3. At least some of the flux tubes are visible as small bright points. This increased contrast in spectral lines is due to three main reasons. Firstly, it is due to an increasing temperature difference between flux tubes and surroundings with height. A second contributing factor is the large temperature sensitivity of some lines, for example lines of minor ions or molecular lines. Finally, the increased contrast at greater height is due to the expansion of the flux tube with height, so that the atmosphere is increasingly filled with hot gas at greater heights. Since individual flux tubes are usually not resolved this increased filling enhances the brightness.

Figure 4 shows the empirically derived temperature of active region plage and network flux tubes as a function of $\log(\tau_{500})$, the logarithmic continuum optical depth at 500 nm. In this representation the network flux tube model is hotter than the quiet Sun in all layers, whereas the plage

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

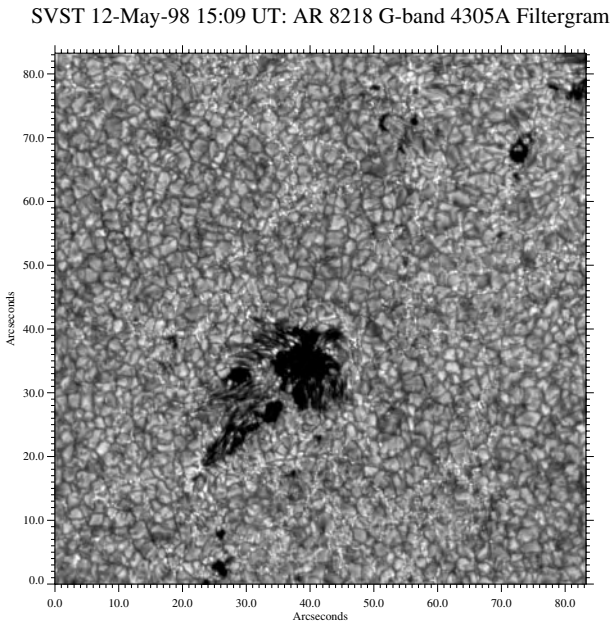


Figure 3. Image of a part of the solar surface obtained in the G-band at 430 nm. In addition to the granulation and the dark sunspot, numerous small bright points are visible. These correspond to magnetic elements or small flux tubes. The high contrast in this image is due to the presence of spectral lines due to the CH molecule in the imaged wavelength band. (Figure kindly provided by T E Berger.)

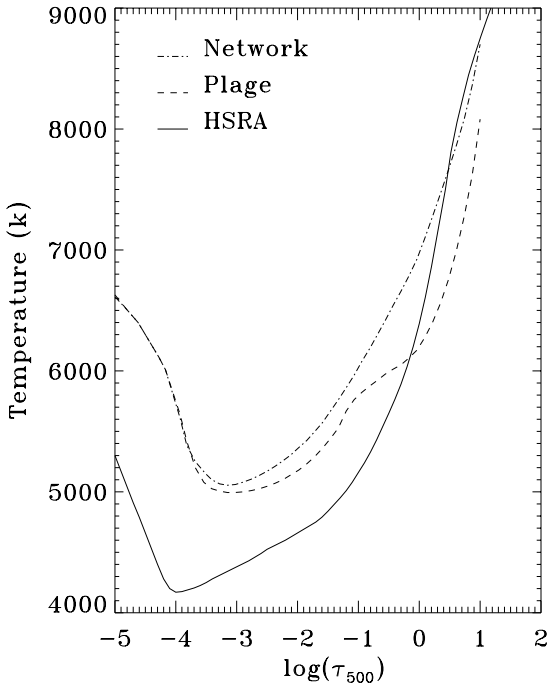


Figure 4. Temperature versus logarithmic continuum optical depth (log τ_{500}) of three empirical model atmospheres. Solid curve: quiet Sun model HSRA; dashed curve: model describing the magnetic elements found in active region plage; dot-dashed curve: magnetic elements forming the network.

flux tube model, which describes somewhat larger flux tubes, is cooler in the lowest layers. If geometrical height were to be used instead of optical depth, then the flux tube curves would have to be moved to the right due to the Wilson depression, so that in the lower photosphere the temperatures of both flux tubes would lie below that of the quiet Sun.

The difference between the two flux tube curves is mainly restricted to the lower photosphere. In the upper photosphere the models are relatively similar. In particular, they show almost identical chromospheric temperature rises. Other observations also suggest that the size dependence of the thermal structure is primarily a property of the lower layers, as long as we don't consider very large flux tubes and extremely high filling factors. In sunspots, for example, the chromosphere starts decidedly higher in the atmosphere than in magnetic elements.

The chromospheric temperature rise seen in figure 4 within the flux tubes sets in 200–300 km lower than in the quiet Sun. This explains the high contrasts seen, for example in Ca II K. The contrast continues to increase with height, so that the contrast in transition region lines (observed in the EUV) is larger than in the visible.

One observation that needs to be reproduced by theory is that whereas the flux tube magnetic field expands rapidly with height in chromospheric layers (forming a magnetic canopy), the brightness structures associated with the field in the chromosphere and transition region are much more compact than expected from such an expansion, partly lying at the resolution limit of 1″ of EUV observations (see CHROMOSPHERE: MAGNETIC CANOPY). Only the central parts of the flux tubes appear to get significantly heated.

Magnetic regions exhibit a definite center-to-limb variation of the continuum intensity. At low and intermediate resolution the continuum contrast is small at the center of the solar disk, but increases rapidly towards the limb. This dependence can be explained by a hot cloud model, in which the temperature difference at a given optical depth increases rapidly with height in the lower and middle photosphere. This, however, does not agree with the temperature stratification derived from Stokes profiles, in particular for network flux tubes (indicated by the dot-dashed curve in figure 4). The center-to-limb variation can also be explained by a geometrical effect, such as the change in visibility of the hot walls of a partially evacuated flux tube as it moves across the disk. In order to reproduce the observations in this manner a mixture of flux tubes of different sizes needs to be present, with smaller flux tubes being bright closer to the center of the disk, while larger flux tubes exhibit the largest contrast near the limb. High-resolution filtergrams support this view. Whereas near the center of the disk the bright features associated with network or facular flux tubes are mainly bright points, often lying at the limit of the spatial resolution, the bright

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

structures forming faculae near the limb are dominantly the considerably larger facular granules.

Lifetimes

The lifetime of a magnetic flux tube can be defined as the time that it exists as a coherent structure before it either splits, merges with others, submerges, etc. Note that the question of a flux tube's lifetime is distinct from the question of how long magnetic flux remains on the solar surface before it is expelled or submerges again.

Determining lifetimes of individual flux tubes is even more difficult than determining their sizes, since the lifetime of a magnetic feature can only be determined if it can be resolved, or at least separated from its neighbors over its whole lifetime. As pointed out in the section on flux tube sizes and magnetic fluxes this is generally not the case, even on individual frames of magnetograms, not to say for whole time series.

Therefore, mainly two types of observations are used to estimate flux tube lifetimes; those which determine the lifetime of a group of flux tubes (e.g. a part of the magnetic network) since only the whole group can be resolved, and those which determine the lifetime of some proxy of the magnetic field (such as bright points seen in filtergrams).

Both methods have their problems, since neither the lifetimes of flux-tube groups nor of proxies need correspond to the actual lifetimes of individual flux tubes (not even by an order of magnitude). For example, the brightness of a magnetic element can fluctuate, due to dynamic processes, in the course of its lifetime (only some fraction of the flux appears bright at any given time).

The former method generally gives relatively long lifetimes for stable structures such as network features (50–100 h), whereas the latter gives extremely short durations (5–20 min). Therefore there is an uncertainty of almost three orders of magnitude in our knowledge of flux tube lifetimes.

Time series of proxies and, in particular, images in the G-band at 430 nm, show a continual motion of bright points along intergranular lanes, together with splittings and mergings, apparently driven by the evolution of the neighboring granules.

Dynamics

Flux tubes harbour a wide variety of dynamic phenomena, such as lateral motions of the whole flux tube, steady flows along the field lines and waves supported and guided by the flux tubes.

Stationary flows in flux tubes

A direct indicator of flows within magnetic elements is the zero-crossing wavelength λ_V of Stokes V. Stokes V is the difference between right and left circularly polarized light and λ_V corresponds to its central wavelength, at which Stokes V becomes zero.

In active region plage no average λ_V shifts greater than 250 m s^{-1} are observed. This is true for all lines formed between the low photosphere and the lower to middle chromosphere. There is, however, also evidence for a small redshift in the magnetic features of approximately 100 m s^{-1} at large μ, with indications of a small blueshift at small μ. Evidence also exists for an increase of the average λ_V shift with decreasing magnetic filling factor.

In contrast to spatially averaged values of λ_V, the spatially localized snapshots of λ_V often show significant blue and redshifts (up to 1 km s^{-1}).

The largest line shifts in the photospheric layers of small-scale magnetic features are, however, associated with magnetic neutral lines. Simultaneous observations of the field strength and flow speed on both sides of the neutral line suggest that these shifts are due to siphon flows along small magnetic loops connecting the opposite polarities, with the flow becoming supersonic near the top of the loop. Siphon flows along magnetic loops are driven by the difference in gas pressure in the two footpoints of the loop, which in turn is coupled to a difference in field strength.

The origin of the line shifts (of both signs) seen in high-resolution spectra is less clear. One explanation also invokes siphon flows along many thin (but probably long) loops. Since matter flows from the footpoint of the weaker field to the more intense field, a distribution of field strengths among neighboring flux tubes could produce up- and downflows in the neighboring flux tubes. Another possibility, namely that these shifts represent snapshots of non-stationary motions, is discussed in the next subsection. Finally, at present we also do not know the cause of the small shifts ($\approx$100 m s^{-1}) of the spatially averaged V profiles. Even such small velocities would drain the corona on a time scale of hours to days in open-field regions unless sufficient matter can diffuse across the field lines into the flux tube, which appears unlikely according to current understanding. An alternative explanation involves a net Doppler shift produced by non-stationary velocities within the flux tube.

Another diagnostic is the observed strong blue–red asymmetry of Stokes V profiles. Both the absolute value of the area, A_b, and amplitude, a_b, of the blue Stokes V wing differ from the absolute value of the red wing area, A_r, and amplitude, a_r, respectively. An asymmetric Stokes V profile is illustrated in figure 5. Quantitatively, this asymmetry may be expressed by the relative amplitude, δa, and area, δA, asymmetry defined as

$$\delta a = \frac{a_b - a_r}{a_b + a_r} \quad \text{and} \quad \delta A = \frac{A_b - A_r}{A_b + A_r}. \tag{3}$$

Close to solar disk center both δa and δA are positive for photospheric lines, although δA almost disappears for the strongest lines. In general, $\delta a > \delta A$. Closer to the limb both δa and δA change sign, i.e. the red Stokes V wing becomes stronger than the blue wing.

It is now generally accepted that for photospheric lines cospatial gradients of the line-of-sight velocity and magnetic vector produce the Stokes V asymmetry. Downflows of 1–2 km s^{-1} in the immediate surroundings

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

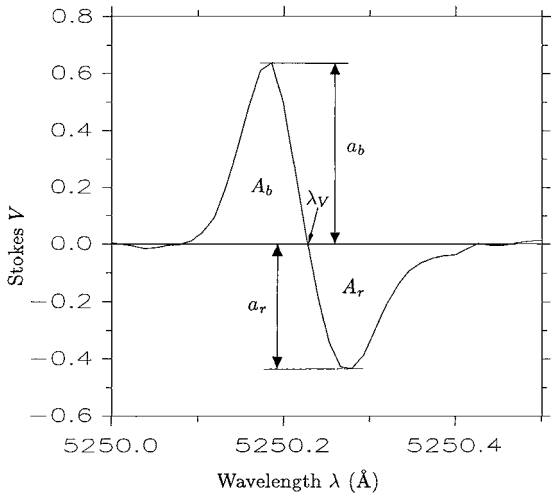


Figure 5. Stokes V profile, i.e. net circularly polarized light, of the Fe I line at 525.2 nm observed in the magnetic network near the center of the solar disk. The zero-crossing wavelength, λ_V, the amplitude, a_b, and area, A_b, of the blue wing of Stokes V and the corresponding red wing parameters, a_r and A_r, are indicated. Note the larger amplitude and area of the blue lobe than of the red lobe.

of the magnetic elements can reproduce the observed δA values without producing any zero-crossing shift if the surroundings are also approximately 200–300 K cooler than the average quiet Sun, highly suggestive of intergranular lanes. This picture of magnetic elements located in the downflowing intergranular lanes is confirmed by direct imaging.

A different view of the production of Stokes V asymmetry is based on the picture that magnetic features are composed of many narrow strands of field, each thinner than the horizontal photon mean-free path. These strands are interleaved with flowing material. Such a model can also reproduce δA and even δa, if two different sets of velocities are assumed to be present in the surroundings.

Traditional flux-tube models with downflows in the surroundings can also simultaneously reproduce δA and δa if flows inside the flux tubes are also allowed. The internal flows may be either a downflow increasing rapidly with depth, or a combination of up- and downflow, suggestive of oscillatory or wavelike motion. Both of these models reproduce the observations equally well, but the latter is to be preferred since it fulfils mass conservation in the flux tube.

In summary, the δA observed in magnetic elements appears to be largely a product of the granulation surrounding the magnetic elements. Conversely, δA can constrain the velocity and the temperature of the convection in active regions, so-called abnormal granulation.

Non-stationary velocities inside flux tubes

The detection of MHD waves in flux tubes is of particular interest, since these are one prime candidate for transporting the energy which heats the chromosphere and which may contribute to the heating of the corona.

A number of diagnostics support the presence of non-stationary velocities in magnetic elements, for example flux tube waves. The most direct of these, time-series of λ_V, have so far mainly uncovered oscillations or waves with a period close to 5 min and an amplitude of 0.2–0.3 km s^{-1} in photospheric spectral lines that increases slowly with height. There are indications that the waves are propagating upwards with a phase velocity of approximately 100 km s^{-1}. The results appear to be relatively independent of the spatial resolution of the observations, which varies between 1″ and 8″.

There is one significant exception to the 5 min periods, however. In an isolated magnetic feature near disk center 100 s periods in λ_V, with a 280 m s^{-1} amplitude have been detected, corresponding to an estimated energy flux of $(1.6–2.3)\times10^7$ erg cm^{-2} s^{-1}, sufficient to heat the associated chromospheric network if the wave is not significantly radiatively or acoustically damped on the way. According to linear theory this period corresponds to propagating longitudinal tube waves, i.e. acoustic waves modified by the magnetic field propagating along the flux tube. Unfortunately, only a single magnetic feature was found to exhibit such oscillations and confirmation is important.

The widths of the V profiles of relatively Zeeman-insensitive lines suggest that non-stationary velocities with much larger amplitudes than those deduced from time-series of λ_V are present in magnetic elements. Typical rms values of such velocities lie around 1.5–2.5 km s^{-1} for photospheric and lower chromospheric lines. These values are equivalent to or larger than those obtained from fits to quiet Sun Stokes I profiles. In the quiet Sun, however, most of the line broadening comes from the granulation, which is, of course, not present within the magnetic features. Therefore, the broadening must be due to an oscillatory or wave-like velocity field, or due to the presence of stationary up- and downflows within different spatially unresolved magnetic elements (e.g. siphon flows on small scales). A fraction of the broadening velocity (10–20%) is detected as the rms of the λ_V spatial fluctuations seen in high-resolution observations (see the section on stationary flows in flux tubes).

The velocity amplitudes deduced from line broadening near the limb are almost the same as near disk center. Similarly, the rms of the λ_V of high-resolution observations does not decrease towards the limb. These two observations suggest that transverse wave modes (e.g. kink mode, torsional Alfvén mode) are just as strongly excited as longitudinal modes (e.g. sausage or tube mode, overstable oscillations). Some evidence of quasi-periodic transverse motions of small-scale magnetic features has indeed been determined (although only for a single feature).

Another possible source of horizontal velocity visible in Stokes V is the motion of complete magnetic elements

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that are continually being shuffled around by the evolving granulation. Velocities of 0.5–5 km s^{-1} are mentioned in the literature. These have been derived by tracking features in time series of filtergrams or magnetograms. Of interest in this respect are the observations of short-duration rapid horizontal movements of bright points. If bright points are identified with flux tubes the amount of energy generated in the form of an upward-propagating kink-wave pulse along a slender flux tube by the observed rapid horizontal movements may be estimated. It turns out to be sufficient to heat the quiet solar corona, even in the presence of reflections at the SOLAR TRANSITION REGION.

The final parameter indicating non-stationary velocities in magnetic features is the Stokes V amplitude asymmetry, δa. It can be reproduced along with δA and λ_V by a simple representation of a nonlinear wave with different up- and downflow velocity amplitudes, along with a downflow outside the flux tube (see the section on stationary flows in flux tubes).

In summary, some diagnostics indicate the presence of broadening velocities up to 3 km s^{-1} (both Stokes V line broadening and δa give similar velocity amplitudes), but λ_V time series only show rather low-amplitude perturbations, while snapshots at high spatial resolution yield λ_V with an intermediate rms. These results suggest that (a) many flux tubes with tube waves having different phases are present in the resolution element of a given observation, or (b) the tube waves have a very short wavelength and consequently also a short period (if it becomes smaller than the half-width of the Stokes V contribution function of a particular line, the waves become invisible in the λ_V time series of that line), or (c) the dynamical thin-tube approximation breaks down and most of the power is present in higher-order modes having multiple nodes or at least different phases across the cross section of the tube. Such waves are extremely difficult to detect, due to the cancellation of phases when averaging over even a single flux tube.

Bibliography

Photospheric flux tubes are fascinating and heavily investigated objects, so that only some of the many observational results on them could be included in the present article. There are a number of reviews which provide a good introduction to the subject and the literature, some of which also go into greater detail then the present article. General introductory reviews covering both magnetic elements and sunspots have been written by Stenflo (1989, 1994) and Solanki (1998), while reviews dealing with magnetic elements alone (but including both observational and theoretical aspects) have been published by Spruit *et al* (1992), Schüssler (1992) and Solanki (1993), with the last being the most comprehensive in its literature coverage.

Schüssler M 1992 Small-scale photospheric magnetic fields *The Sun—a Laboratory for Astrophysics* ed J T Schmelz and J C Brown (Dordrecht: Kluwer) pp 191–220

Solanki S K 1993 Small-scale solar magnetic fields: an overview *Space Sci. Rev.* **61** 1–188

Solanki S K 1998 Solar magnetic fields *Space Solar Physics* (Berlin: Springer) pp 41–76

Spruit H C, Schüssler M and Solanki S K 1992 Filigree and flux tube physics *Solar Interior and Atmosphere* ed A N Cox, W Livingston and M S Matthews (Tucson, AZ: University of Arizona Press) pp 890–910

Stenflo J O 1989 Small scale magnetic structures on the Sun *Astron. Astrophys. Rev.* **1** 3–48

Stenflo J O 1994 *Solar Magnetic Fields: Polarized Radiation Diagnostics* (Dordrecht: Kluwer)

Sami K Solanki

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Photospheric Magnetic Flux Tubes: Theory

The magnetic field in the photospheric layers of the Sun is found to occur not in a homogeneous form but in discrete concentrations of intense field. The most obvious form of magnetic flux (magnetic field strength times surface area occupied by the field) is seen in SUNSPOTS but it turns out that much smaller arrangements of magnetic field are to be found in the lanes between granules where downdraughts occur (see SOLAR PHOTOSPHERE: GRANULATION). These are the photospheric flux tubes: small-scale concentrations of intense magnetic field. The tubes have diameters of a few hundred kilometers or smaller and field strengths of 1–2 kG. Generally found in intergranular lanes, they are subject to the dynamical nature of the photospheric environment with its sound waves and flows. The tubes are isolated flux tubes in that their immediate surroundings are essentially field-free. See SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: OBSERVATIONS.

The small-scale tubes of the photosphere may be thought of as the elemental building blocks of solar magnetism, so an understanding of intense flux tubes has implications for a wider understanding of solar and stellar magnetic phenomena. Theoretical progress in modelling small-scale flux tubes builds on the fact that in such tubes variations across the interior of a tube are less important than variations along the tube.

The elemental photospheric flux tube is taken to be a cylinder of magnetic flux embedded in a field-free environment; the photospheric flux tube is thus a region of locally high Alfvén speed. Two basic speeds prove central to an understanding of tube dynamics: the slow tube speed c_t and the kink speed c_k. These speeds are in turn defined in terms of the sound speed and Alfvén speed. For a tube of magnetic field strength B_0 with plasma density ρ_0 and pressure p_0, the sound speed c_s and Alfvén speed v_A of the tube are defined by

$$c_s = \left(\frac{\gamma p_0}{\rho_0}\right)^{1/2} \qquad v_A = \left(\frac{B_0^2}{\mu\rho_0}\right)^{1/2} \tag{1}$$

where γ (generally taken to be 5/3) is the ratio of specific heats of the plasma and μ is its magnetic permeability.

From the speeds c_s and v_A we may construct the slow magnetoacoustic speed c_t and the kink speed c_k of a tube:

$$c_t = \frac{c_s v_A}{(c_s^2 + v_A^2)^{1/2}} \qquad c_k = \left(\frac{\rho_0}{\rho_0 + \rho_e}\right)^{1/2} v_A \tag{2}$$

here ρ_e denotes the plasma density of the field-free environment of the tube. The slow tube speed and the kink speed are important in coronal flux tubes too (see MAGNETOHYDRODYNAMIC WAVES).

A photospheric flux tube is typically a region of plasma density depletion: the presence of the magnetic field leads to a plasma pressure reduction inside the tube, in keeping with the requirement of transverse pressure balance, and unless there are substantial temperature differences between the tube and its environment then this reduction in plasma pressure is accompanied by a reduction in plasma density. In other words, the magnetic field of a photospheric flux tube partially evaculates the plasma. This is in contrast to the coronal flux tube which is generally a region of plasma enhancement and reduced Alfvén speed, achieved without violation of pressure balance because the strong and pervasive magnetic field of the solar corona can easily accommodate any enhancements or reductions in the plasma pressure.

The thermal structure of a photospheric flux tube is complicated. The tube may exchange heat with its surroundings, through radiative transport and convective energy transport. Detailed numerical studies suggest that photospheric magnetic flux tubes are hotter than their surroundings in their upper layers but cooler than their surroundings in their deeper layers.

The speeds c_t and c_k of a photospheric flux tube are sub-Alfvénic, and the slow speed is also sub-sonic; whether the kink speed is sub-sonic depends on the specific parameters of the tube. A numerical illustration is of interest. In a photospheric flux tube with $B_0 = 2$ kG and plasma density $\rho_0 = 2.2 \times 10^{-4}$ kg m^{-3}, the Alfvén speed is $v_A = 12$ km s^{-1}. Taking the sound speed c_s of the photosphere to be 8 km s^{-1} then gives a slow speed of $c_t = 6.7$ km s^{-1}. For a flux tube that has depleted its plasma density to some 50% of its environment (so that $\rho_0 = \rho_e/2$, consistent with an environment sound speed of 9.6 km s^{-1}), we obtain $c_k = 6.9$ km s^{-1}. Thus the slow speed c_t is some 56% of the Alfvén speed and the kink speed c_k is about 57% of the Alfvén speed. The precise ordering of the two speeds depends on the details of the modelling.

The speeds c_t and c_k are associated with the dynamics of flux tubes, arising in both wave phenomena and flows in tubes. The speed c_t, made up in equal measure of the sound speed and the Alfvén speed, is evidentally associated with the compressibility of the tube. Any squeezing of a tube results in increases in the magnetic pressure and plasma pressure in the tube, which act to restore the undisturbed state. This is a feature shared by all elastic tubes and consequently the equivalent of the speed c_t arises in a variety of other physical situations, with the role of the Alfvén speed being played by the appropriate elastic speed of an elastic tube. For example, in the case of a blood vessel the equivalent of the speed v_A is simply the elastic speed in the membrane of the blood vessel. The speed of sound c_s in blood is much larger than the elastic speed and so the slow speed is close to the elastic speed of the blood vessel. For water in a pipe, the relative magnitudes of the two basic speeds depend on the material of the pipe. In a metal pipe, the elastic speed is much larger than the sound speed in water and so the effective slow speed is close to the sound speed in water (about 1.4 km s^{-1}). By contrast, in a plastic pipe the orderings in the two speeds are reversed and the effective propagation speed is close to the elastic speed in plastic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(about 10 m s^{-1}), lying far below the speed of sound in water.

The kink speed involves both the tube and its environment. The kink wave which propagates with this speed (see below) involves the magnetic tension force in the tube and the kink wave displaces both the tube and its surroundings by about equal measures, hence the combined occurrence of the densities ρ_0 and ρ_e. The speed also arises in the description of magnetoacoustic surface waves on an interface between two plasmas of different properties.

An important question in the physics of photospheric flux tubes is how are they formed. A flux tube with kilogauss field strength requires strong forces to bring about its intense state, with the magnetic pressure $B_0^2/2\mu$ being far in excess of the pressure of converging flows that granulation may provide; the tube finds equilibrium through evacuation of its interior, reducing the plasma pressure and thus allowing magnetostatic pressure balance between the flux tube and its environment to occur. It is envisaged that granulation acts to 'shepherd' flux tubes into the regions where downdraughts between the granules occur. The process is an example of flux expulsion (by advection), whereby the convective flows of the photosphere move magnetic field into intergranular lanes, in keeping with the frozen flux concept of ideal MAGNETOHYDRODYNAMICS. The process is not of itself sufficient to account for the observed field strengths; it brings about the 'seedlings' of photospheric flux tubes but other effects are needed to produce the observationally determined intense field strengths. The additional process that comes into play is driven by the super-adiabatic temperature gradient of the solar convection zone (see SOLAR INTERIOR: CONVECTION ZONE), which in field-free regions drives the convective flows (granules and supergranules; see SOLAR PHOTOSPHERE: SUPERGRANULATION) that distinguish the Sun's convection. In flux tubes, convectively driven flows are channelled by the tubes. The fact that photospheric flux tubes reside in a plasma that is stratified by gravity means that any downdraught generated within a flux tube brings less dense plasma from the higher reaches of the stratified tube to the deeper layers within the tube, allowing the high plasma pressure of the environment to squeeze the tube further. An increase in magnetic field strength results. The process is referred to as convective collapse. It may be described in mathematical detail for a thin flux tube (see below).

A flow in a magnetic flux tube may also develop because of pressure differences between the footpoints of the tube. For example, if the footpoints of a tube are rooted in regions that are magnetically distinct from one another, such as one footpoint in a sunspot and the other in a field-free region, then plasma pressure differences arise and a flow may be driven in the tube. Such motions of the plasma within the tube are called siphon flows.

Waves in photospheric flux tubes

To give a detailed description of the magnetohydrodyanmic waves that an isolated photospheric flux tube supports we consider the equations of ideal MHD. Consider an equilibrium magnetic field $\boldsymbol{B}_0 = B_0(r)\hat{\boldsymbol{z}}$ aligned with the z-axis of a cylindrical polar coordinate system (r, θ, z). To begin with we will ignore the effects of gravity. Then the equilibrium plasma pressure $p_0(r)$ of a radially structured magnetic atmosphere is such as to maintain total pressure balance: the sum of the plasma pressure $p_0(r)$ and the magnetic pressure $B_0^2(r)/2\mu$ is a constant. Where the field is strong, as in the center of a flux tube, the plasma pressure is correspondingly reduced.

Small-amplitude motions $\boldsymbol{v} = (v_r, v_\theta, v_z)$ about a structured equilibrium state satisfy the coupled wave equations

$$\rho_0(r)\left(\frac{\partial^2}{\partial t^2} - v_A^2(r)\frac{\partial^2}{\partial z^2}\right)v_r + \frac{\partial^2 p_T}{\partial r \partial t} = 0 \qquad (3)$$

$$\rho_0(r)\left(\frac{\partial^2}{\partial t^2} - v_A^2(r)\frac{\partial^2}{\partial z^2}\right)v_\theta + \frac{1}{r}\frac{\partial^2 p_T}{\partial \theta \partial t} = 0 \qquad (4)$$

$$\left(\frac{\partial^2}{\partial t^2} - c_t^2(r)\frac{\partial^2}{\partial z^2}\right)v_z = -\frac{c_s^2(r)}{c_s^2(r) + v_A^2(r)}\frac{1}{\rho_0(r)}\frac{\partial^2 p_T}{\partial z \partial t} \qquad (5)$$

with the evolution of $p_T(r, \theta, z)$ described by

$$\frac{\partial p_T}{\partial t} = \rho_0(r)v_A^2(r)\frac{\partial v_z}{\partial z} - \rho_0(r)[c_s^2(r) + v_A^2(r)]\mathrm{div}\,\boldsymbol{v}. \qquad (6)$$

Here $c_t(r)$, $c_s(r)$ and $v_A(r)$ denote the slow speed, sound speed and Alfvén speed within the radially structured plasma, of density $\rho_0(r)$. Equations (3)–(5) come from the components of the momentum equation in which the magnetic force has been expressed in terms of the perturbation in total pressure p_T, defined as the sum of the perturbation p in the dynamical plasma pressure and the magnetic pressure perturbation $B_0(r)B_z/\mu$, with B_z being the component of the perturbed magnetic field in the direction of the applied magnetic field $\boldsymbol{B}_0$. Equation (6) results from a combination of the isentropic (adiabatic) equation and the induction equation of ideal MHD.

One solution of the above equations is the torsional Alfvén wave, which propagates motions v_θ, independent of θ, according to the wave equation

$$\frac{\partial^2 v_\theta}{\partial t^2} = v_A^2(r)\frac{\partial^2 v_\theta}{\partial z^2}. \qquad (7)$$

Torsional oscillations, which have $v_r = 0$, $v_z = 0$ and $p_T = 0$, have a radial dependence determined by the manner in which the oscillations are generated. The oscillations exhibit phase mixing by which radial gradients of the motion grow rapidly in regions where the Alfvén speed $v_A(r)$ varies sharply in r.

In addition to torsional Alfvén waves there are motions that are compressive, with $p \neq 0$ and $p_T \neq 0$. We examine the elemental magnetic flux tube that models photospheric conditions, taking the applied magnetic field

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$B_0(r)\hat{z}$ to be confined within a cylindrical tube of radius a. The plasma pressure and density within the tube are p_0 and ρ_0; the corresponding plasma pressure and density in the tube's field-free environment are p_e and ρ_e. The tube is in magnetostatic pressure balance with its surroundings: $p_e = p_0 + B_0^2/2\mu$. With c_s and v_A now denoting the sound speed and Alfvén speed within the tube and c_e the sound speed in the tube's environment, pressure balance combined with the ideal gas law then shows that the plasma density ρ_0 within the tube is related to the density ρ_e in the tube's environment through

$$\frac{\rho_0}{\rho_e} = \frac{c_e^2}{c_s^2 + \frac{1}{2}\gamma v_A^2}. \tag{8}$$

Thus, for sound speeds c_s and c_e that are roughly comparable, a photospheric flux tube is a region of density depletion (i.e. $\rho_0 < \rho_e$).

Now the media inside and outside the elemental flux tube are uniform, and this affords a considerable simplification. In a uniform medium, the wave equations (3)–(6) may be combined to yield an equation for p_T. Writing

$$p_T(r, \theta, z, t) = p_T(r) \exp \mathrm{i}(\omega t + n\theta - k_z z)$$

provides a Fourier description of the motions, of frequency ω, longitudinal wavenumber k_z and azimuthal number n. The integer $n(= 0, 1, 2, \ldots)$ describes the geometrical form of the perturbations. The case $n = 0$ gives the sausage wave and corresponds to symmetric squeezings and rarefactions of the tube; these are compressional ($p_T \neq 0$) oscillations. The case $n = 1$ describes the kink wave of the tube; in such waves the motion of the tube resembles a wriggling snake, with the whole tube being displaced without changing its cross-sectional shape. Finally, there are also waves with $n \geq 2$; these are the fluting waves.

For a uniform tube, p_T satisfies

$$r^2 \frac{\mathrm{d}^2 p_T}{\mathrm{d}r^2} + r \frac{\mathrm{d}p_T}{\mathrm{d}r} - (m_0^2 r^2 + n^2) p_T = 0 \tag{9}$$

where

$$m_0^2 = \frac{(k_z^2 c_s^2 - \omega^2)(k_z^2 v_A^2 - \omega^2)}{(c_s^2 + v_A^2)(k_z^2 c_t^2 - \omega^2)}.$$

This is a form of Bessel's equation with solution $p_T = I_n(m_0 r)$ inside ($r < a$) the tube; I_n is a modified Bessel function. Similarly, in the field-free environment ($r > a$) of the flux tube, where the sound speed is c_e, a solution p_T proportional to the modified Bessel function $K_n(m_e r)$ arises, where

$$m_e^2 = k_z^2 - \frac{\omega^2}{c_e^2}.$$

It is usual to examine modes for which m_e^2 is positive, corresponding to selecting waves that are confined to near the tube. Then, matching p_T and v_r across $r = a$ results in

$$\frac{1}{\rho_0(k_z^2 v_A^2 - \omega^2)} m_0 \frac{I_n'(m_0 a)}{I_n(m_0 a)} + \frac{1}{\rho_e \omega^2} m_e \frac{K_n'(m_e a)}{K_n(m_e a)} = 0 \tag{10}$$

where a prime ($'$) denotes the derivative of a modified Bessel function (e.g. $I_n'(m_0 a) \equiv \mathrm{d}I_n(x)/\mathrm{d}x$ evaluated at $x = m_0 a$, etc). This is the dispersion relation for tube waves in an isolated magnetic flux tube, requiring that $m_e > 0$; surface waves arise whenever $m_0^2 > 0$, and body waves occur if $m_0^2 < 0$. It is evident that tube waves are dispersive, their phase speed $c(\equiv \omega/k_z)$ depending on the wavenumber (in the combination $k_z a$).

Solution of the dispersion relation (10), for conditions appropriate for an isolated photospheric flux tube, show that there are slow body waves (both sausage and kink) with phase speeds c that lie between c_t and c_s and slow surface waves which have phase speeds that are less than c_t. Also, there is a surface wave with a phase speed close to the kink speed c_k and another surface wave with phase speed near c_e. Fluting modes also have speeds close to c_k.

There is particular interest in waves that are much longer than a tube diameter, corresponding to $k_z a \ll 1$. This is the thin flux tube limit, of special interest for photospheric flux tubes. Since generally $\omega \sim k_z v_A$, the condition $k_z a \ll 1$ corresponds to requiring that periods $\tau(\equiv 2\pi/\omega)$ are such that $\tau \gg 2\pi a/v_A$, giving $\tau \gg 50$ s for a tube of radius $a = 100$ km. P-modes, for example, have periods around 300 s, roughly meeting this criterion. With $k_z a \ll 1$, the slow surface wave has phase speed close to c_t.

Dispersion relations such as equation (10) are important in the development of a nonlinear wave theory for a tube. Indeed, it has been shown that slow (sausage) surface waves have motions $v(z, t)$ along a thin tube which satisfy the nonlinear integrodifferential equation

$$\frac{\partial v}{\partial t} + c_t \frac{\partial v}{\partial z} + \beta_0 v \frac{\partial v}{\partial z} + \alpha_0 \frac{\partial^3}{\partial z^3} \int_{-\infty}^{\infty} \frac{v(s, t)\, \mathrm{d}s}{[\lambda^2 a^2 + (z - s)^2]^{1/2}} = 0. \tag{11}$$

The constants α_0, β_0 and λ depend on the various parameters of the tube.

Equation (11) is sometimes referred to as the Leibovich–Roberts equation. An explicit form of its solution is not known although it is amenable to numerical solution. It is of particular interest because it is associated with soliton behavior. Solitons are nonlinear waves with specific interaction properties that make them worthy of special study. For the related problem of a magnetic slab—the Cartesian equivalent of a flux tube—the equation describing the weakly nonlinear slow surface wave is the Benjamin–Ono equation, i.e.

$$\frac{\partial v}{\partial t} + c_t \frac{\partial v}{\partial z} + \beta_0 v \frac{\partial v}{\partial z} + \alpha_1 \frac{\partial^2}{\partial z^2} \int_{-\infty}^{\infty} \frac{v(s, t)}{s - z}\, \mathrm{d}s = 0 \tag{12}$$

for constant α_1. The Benjamin–Ono equation has been studied extensively and its soliton solution is known explicitly. Equations (11) and (12) may be derived using thin flux tube theory (see below) or from the full equations of ideal MHD, assuming that motions are weakly nonlinear and weakly dispersive.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thin flux tube theory

The restriction to thin tubes, in the sense that waves are much longer than a tube radius (so that $k_z a \ll 1$), leads to a considerable simplification in the description of the dynamics of photospheric flux tubes, so much so that it becomes possible to include a number of other effects ignored in the discussion of an elemental magnetic flux tube. Of particular importance is the influence of stratification brought about by gravity. The pressure scale height in the photosphere is comparable with a tube radius so the effects of gravity are particularly significant in the photospheric layers of a tube. Other effects, such as radiative transport, may also in principle be included in thin flux tube theory.

To describe motions in a thin tube in the presence of gravity the so-called thin flux tube equations are used. These equations have been used extensively for both analytical and numerical investigations of flux tube dynamics. The sausage and kink modes are treated separately. The thin tube equations for the *sausage mode* (in ideal MHD) are

$$\frac{\partial}{\partial t}\rho A + \frac{\partial}{\partial z}\rho v A = 0 \tag{13}$$

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial z} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g \tag{14}$$

$$\frac{\partial p}{\partial t} + v\frac{\partial p}{\partial z} = \frac{\gamma p}{\rho}\left(\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial z}\right) \tag{15}$$

$$BA = \text{constant} \tag{16}$$

$$p + \frac{B^2}{2\mu} = p_\mathrm{e}. \tag{17}$$

In these equations, $B(z,t)$ is the field strength of a thin tube with cross-sectional area $A(z,t)$, and $v(z,t)$ is the longitudinal flow speed within the tube, where the plasma pressure and density are $p(z,t)$ and $\rho(z,t)$. The gravitational acceleration is $g(=274\ \mathrm{m\ s^{-2}})$, acting in the negative z-direction. The external gas pressure $p_\mathrm{e}(z,t)$ is calculated on the boundary of the tube and may vary in response to motions in the tube or its environment; pressure balance across the tube is maintained at all times (as expressed by equation (17)).

In equilibrium ($v = 0, \partial/\partial t = 0$) the thin tube equations for a magnetic flux tube in temperature balance with its surroundings (so that $c_\mathrm{s} = c_\mathrm{e}$) yield

$$p_0(z) = p_0(0)\,\mathrm{e}^{-N} \qquad \rho_0(z) = \rho_0(0)\frac{\Lambda_0(0)}{\Lambda_0(z)}\,\mathrm{e}^{-N}$$

$$A_0(z) = A_0(0)\,\mathrm{e}^{N/2} \qquad B_0(z) = B_0(0)\,\mathrm{e}^{-N/2} \tag{18}$$

where

$$N(z) = \int_0^z \frac{\mathrm{d}z}{\Lambda_0(z)}$$

is the integrated pressure scale height; $\Lambda_0(z)(\equiv p_0(z)g\rho_0(z))$ is the pressure scale height inside the tube. See figure 1.

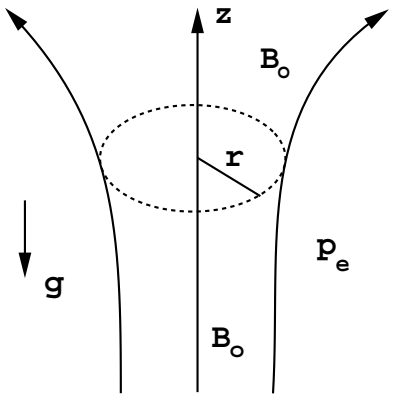


Figure 1. Equilibrium state of a thin magnetic flux tube in a stratified plasma. The tube is confined by the external plasma pressure p_e in the field-free environment of the tube. The region within the tube is generally of lower plasma density than the environment.

The linear form of the thin tube equations is readily found for the equilibrium (18). Ignoring contributions from *variations* in p_e, longitudinal motions are found to satisfy the Klein–Gordon equation

$$\frac{\partial^2 Q}{\partial t^2} - c_\mathrm{t}^2(z)\frac{\partial^2 Q}{\partial z^2} + \Omega_\mathrm{S}^2(z)Q = 0 \tag{19}$$

where $Q(z,t)$ is related to the flow $v(z,t)$ and $\Omega_\mathrm{S}^2(z)$ depends on the details of the equilibrium state.

In an isothermal atmosphere, Λ_0, c_t and Ω_S^2 are all constants, with

$$\Omega_\mathrm{S}^2 = \frac{c_\mathrm{t}^2}{4\Lambda_0^2}\left[\left(\frac{9}{4} - \frac{2}{\gamma}\right) + \frac{4c_\mathrm{s}^2}{\gamma v_\mathrm{A}^2}\left(1 - \frac{1}{\gamma}\right)\right]. \tag{20}$$

The Klein–Gordon equation (19) then yields the dispersion relation

$$\omega^2 = k_z^2 c_\mathrm{t}^2 + \Omega_\mathrm{S}^2 \tag{21}$$

which shows that Ω_S is a cutoff frequency (for sausage waves in a thin tube). Ω_S^2 may be viewed as made up of two contributions, the first (corresponding to the first term on the right-hand side of equation (20)) arising from the geometrical shape of the undisturbed tube and the second (corresponding to the second term on the right of equation (20)) being determined by the tube's elasticity. A rigid tube with exponential cross-sectional area (determined according to the equilibrium (18) with Λ_0 constant) has cutoff frequency $(9/4 - 2/\gamma)^{1/2}c_\mathrm{s}/2\Lambda_0$, whereas a straight and vertical rigid tube has cutoff frequency $c_\mathrm{s}/2\Lambda_0$. In general the cutoff frequency for the sausage mode in a tube is less than the cutoff frequency of a rigid tube.

The Klein–Gordon equation also describes the kink mode in a thin tube. Thin tube equations for the kink mode may be used to describe the transverse displacements $\xi(z,t)$ of a vertical tube, leading to the wave equation

$$\frac{\partial^2 \xi}{\partial t^2} = c_\mathrm{k}^2(z)\frac{\partial^2 \xi}{\partial z^2} + g\left(\frac{\rho_0 - \rho_\mathrm{e}}{\rho_0 + \rho_\mathrm{e}}\right)\frac{\partial \xi}{\partial z}. \tag{22}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Equation (22) may be cast into an equation of the Klein–Gordon form, just as for the sausage mode. For an isothermal medium, the dispersion relation for the kink mode is

$$\omega^2 = k_z^2 c_k^2 + \Omega_K^2 \tag{23}$$

where the cutoff frequency for the kink mode is

$$\Omega_K = \frac{c_k}{4\Lambda_0}. \tag{24}$$

The forms (19)–(24) allow a direct comparison between sausage and kink modes. For example, with again $v_A = 12$ km s^{-1}, $c_s = 8$ km s^{-1} and $\Lambda_0 = 140$ km, sausage and kink modes both grow in amplitude by a factor of e in propagating a distance of 560 km (four scale heights). The sausage wave, propagating with a speed $c_t = 6.7$ km s^{-1}, has a cyclic cutoff frequency, $\Omega_S/2\pi$, of 4.2 mHz (period 240 s). The kink wave has a speed of $c_k = 6.9$ km s^{-1}, close to the sausage mode's, but its cutoff frequency at 2 mHz (period 500 s) is very different. The sausage mode's behavior is similar to that of a vertically propagating sound wave, which has a propagation speed of 8 km s^{-1} and cutoff frequency of 4.5 mHz (period 220 s), although it e-folds in just 280 km (two scale heights); consequently, sound waves are expected to form shocks lower in the atmosphere than tube waves. The fact that the cutoff frequency of the kink mode is considerably smaller than the cutoff frequency of the sausage wave or a sound wave suggests that kink waves may survive to higher levels in the solar atmosphere than sausage waves or sound waves.

The dispersion relations (21) and (23) imply that an impulsively generated wave (sausage or kink) results in a wavefront which propagates with the speed c_t or c_k, the wavefront supporting an oscillating wake which rises and falls with the frequency Ω_S or Ω_K.

Convective collapse

Convective collapse, the process by which a photospheric flux tube may be formed with kilogauss field strengths, may be demonstrated from the thin flux tube equations for the sausage mode. The differential equation satisfied by the longitudinal flow, Fourier analyzed by writing $v(z, t) = u(z)\exp(i\omega t)$, is

$$\frac{B_0}{\rho_0}\frac{d}{dz}\left(\frac{\rho_0 c_t^2}{B_0}\frac{du}{dz}\right) + \left[\omega^2 - \omega_g^2\left(\frac{c_t^2}{v_A^2} + \frac{\gamma c_t^2}{2c_s^2}\right)\right]u = 0 \tag{25}$$

where

$$\omega_g^2 = \frac{g}{\Lambda_0}\left(\Lambda_0' + \frac{\gamma - 1}{\gamma}\right)$$

is the square of the Brunt–Väisäla (or buoyancy) frequency of a plasma element in a stratified atmosphere. In the atmosphere of a flux tube embedded within the solar convection zone, ω_g^2 is negative and this drives an instability within the tube. The magnetic field may quench the instability if it is sufficiently strong. This is readily illustrated by an approximate solution of equation (25). With the illustrative boundary conditions that $u = 0$ at levels $z = 0$ and $z = -d$, we obtain the approximate solution

$$u(z) = u_0 \exp\left(\frac{z}{4\Lambda_0}\right)\sin\left(\frac{\pi z}{d}\right) \tag{26}$$

with

$$\omega^2 = c_t^2\left[\frac{\pi^2}{d^2} + \frac{1}{16\Lambda_0^2} + \left(\frac{1}{v_A^2} + \frac{\gamma}{2c_s^2}\right)\omega_g^2\right].$$

Hence, if $\omega_g^2 < 0$ then a mode with $\omega^2 < 0$ may arise, corresponding to an instability. The instability amounts to a flow of plasma within the stratified tube; a downdraught brings the more tenuous plasma in the higher levels of the tube to layers deeper within the tube, allowing the external plasma pressure to squeeze the tube. A stronger field results. This is convective collapse. The process is halted when the field strength B_0 is sufficiently large to resist the compression of the tube, and so quench the instability (reducing ω^2 to zero). A flux tube with only moderate field strength finds itself subject to the instability until it has strengthened itself, through convective collapse, to a state where it may resist the process. Consequently, intense and isolated magnetic flux tubes are the expected norm in any medium such as the solar photosphere where convective effects operate in the presence of magnetic field.

A determination of the field strength necessary to quench the instability depends on the precise details of the modelling. Studies of the process under conditions appropriate for the solar convection zone show that kilogauss field strengths are required before the collapse is halted. However, what precisely is the ensuing state of a collapsed flux tube remains uncertain, mainly as a result of the need to take into account a number of detailed and complicated effects ignored in the above treatment. In particular, it is necessary to include the effects of RADIATIVE TRANSFER, nonlinearity and a more realistic choice of flow boundary conditions.

Numerical simulations

Numerical simulations allow the inclusion of a variety of effects that complicate the description of the basic modes of behavior of photospheric flux tubes, and this promotes a more direct comparison with solar observations. By following the motion of a parcel of magnetized fluid (using a so-called Lagrangian approach), it proves possible, for example, to explore the linear and nonlinear sausage and kink modes in a vertical tube or to simulate the convective collapse process. These problems are usually addressed through numerical solution of thin flux tube equations, allowing for the important effects of stratification. Broadly, analytic theory for an ideal plasma has provided valuable insights for the more complex modelling of flux tubes under conditions representative of the solar photosphere. Thin flux tube theory for the kink mode has also been applied to modelling the behaviour of a nonvertical flux tube, addressing in particular the important question of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

how magnetic field stored below the convection zone is transported to the solar surface.

An important question still under active consideration is how to incorporate the environment's back-reaction on a moving flux tube. The effect has been explored for the sausage mode and leads to the formation of solitons; the influence of stratification on flux tube solitons has not so far been explored. The back-reaction of the environment on the kink mode is somewhat uncertain, with several alternative possibilities suggested in the literature.

The behaviour of shocks excited within tubes and also the formation of shocks and downflows as a result of the convective collapse of a tube have been modelled in two-dimensional simulations. Numerical simulations have also permitted realistic estimates of the energy flux carried by the various waves, allowing applications to the Sun and stellar atmospheres more generally to be explored.

Bibliography

Hollweg J V 1990 MHD waves on solar magnetic flux tubes *Physics of Magnetic Flux Ropes (Geophys. Monogr. 58)* ed C T Russell, E R Priest and L C Lee (Washington, DC: AGU) p 23

Parker E N 1979 *Cosmical Magnetic Fields* (Oxford: Oxford University Press) pp 841

Roberts B 1992 Magnetohydrodynamic waves in structured magnetic fields *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) p 303

Roberts B and Ulmschneider P 1997 Dynamics of flux tubes in the solar atmosphere: theory *Solar and Heliospheric Plasma Physics* ed G M Simnett, C E Alissandrakis and L Vlahos (Berlin: Springer) p 75

Ryutova M P 1990 Waves and oscillations in magnetic flux tubes *Solar Photosphere: Structure, Convection, and Magnetic Fields (IAU Symp. 138)* ed J O Stenflo (Dordrecht: Reidel) p 229

Schüssler M 1990 Theoretical aspects of small-scale photospheric magnetic fields *Solar Photosphere: Structure, Convection, and Magnetic Fields (IAU Symp. 138)* ed J O Stenflo (Dordrecht: Reidel) p 161

Spruit H C 1981 Magnetic flux tubes *The Sun as a Star (NASA SP-450)* ed S Jordan (Washington, DC: NASA) p 385

Spruit H C and Roberts B 1983 Magnetic flux tubes on the Sun *Nature* **304** 401

Spruit H C, Schüssler M and Solanki S K 1991 Filigree and flux tube physics *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M S Matthews (Tucson, AZ: University of Arizona Press) p 890

B Roberts

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Probe

Proposed NASA mission to fly through the Sun's corona. Planned launch date 2007. Inserted into an eccentric orbit, it would approach within 3 solar radii of the surface to take *in situ* measurements of the corona, flying from pole to pole. A second flyby may occur in 2015.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominence Chirality

A prominence is a relatively cool sheet of dense plasma embedded in the hot tenuous corona. They are easily seen as bright clouds at the 'limb' or edge of the solar disk. When viewed on the solar disk, prominences are long narrow features, i.e. 'FILAMENTS'. Prominences are permeated by weak magnetic fields that are thought to support the plasma against the pull of gravity, and to guide its flows.

In recent years, a global pattern has been recognized in the orientation of magnetic fields inside prominences, especially those 'quiescents' located far away from sunspots. The pattern is characterized by a preferred chirality or handedness in the northern and southern solar hemispheres. The origin of the pattern is controversial at the time of writing, but may shed some light on the large-scale processes involved in the 11-year SOLAR CYCLE.

D Rust discovered the pattern in 1967 from his measurements of the Zeeman effect in prominences that form an east–west chain or crown at high latitudes around the poles of the Sun. (In a magnetic field some spectral lines split into two or more circularly polarized components. This is the ZEEMAN EFFECT.)

Those prominences at the northeast limb had negative field components along the line of sight, those at the northwest limb, positive fields. So both suggested that the component of the field along the long axis of the prominences pointed from east to west. At the same time axial fields in the southern hemisphere pointed east. Rust predicted this pattern would reverse in the next sunspot cycle, cycle 21.

In 1970, J-L Leroy, V Bommier and S Sahal-Brechot, at the Pic du Midi Observatory, began a long series of observations of prominence magnetic fields. They measured the Hanle effect in a strong spectral line of neutral helium, the so-called D3 line at 587.6 nm. This line is emitted by the chromosphere that surrounds and lies under a prominence. When the light in the line is scattered by the helium atoms in the prominence, it becomes linearly polarized. But the magnetic field in the prominence tends to *depolarize* the scattered light. By measuring the amount of depolarization, the French group was able to determine the direction and strength of the magnetic fields.

Leroy's group confirmed Rust's finding of a hemispheric pattern for the axial component of the field in a large sample of quiescent prominences and also confirmed his prediction that the directions of the axial fields would reverse with the onset of cycle 21. Figure 1 is an example of the French group's results. It is a map that shows all the filaments visible on the solar disk during a solar rotation. The arrows show axial field components pointing to the west in the northern hemisphere and to the east in the southern hemisphere. The + and − signs designate the polarity of the weak magnetic fields around the filaments, and the shadings indicate bright Hα active regions.

S F Martin and her associates at Big Bear Observatory extended the Rust–Leroy pattern to lower-latitude prominences. In 1992 they were able to infer the direction of the axial field in a filament from the orientation of small thread-like features ('fibrils') that cluster around a filament in the neighboring chromosphere (see CHROMOSPHERE: FIBRILS).

They discovered a pattern which is implicit in figure 1. The majority of quiescent filaments are either left-handed ('sinistral') or right-handed ('dextral') depending on the direction of their axial fields as seen by a hypothetical observer who stands in the positive magnetic region adjacent to the filament (figure 2). Moreover, most dextral filaments are found in the northern hemisphere, most sinistrals in the southern hemisphere. This relationship remains unchanged, cycle after cycle. Although the absolute direction (eastward or westward) of the axial fields reverses from one cycle to the next, so does the relative position of the neighboring positive magnetic zone.

Filament channels, in which filaments form, are also either dextral or sinistral and also display this hemispheric preference (see SOLAR FILAMENT CHANNELS).

Martin and her colleagues also pointed out a *structural* pattern of chirality in filaments. As seen from above, the 'feet' or 'barbs' of a filament project from the axis in only one of two possible ways, a fact that P Foukal had noticed in 1971. Sinistral filaments are always 'left-bearing', dextral filaments are always 'right-bearing' (figure 2)

Interpretations

D Rust and A Kumar have extended a conceptual model of the solar cycle, due to H Babcock, to explain the chirality of filaments. In Babcock's model (see DYNAMO THEORY) the subsurface magnetic field at the beginning of a cycle extends north–south to connect the poles. As the cycle progresses, differential rotation of the plasma (see SOLAR INTERIOR: ROTATION) wraps the meridianal field lines around the Sun, and forms a system of subsurface magnetic 'ropes' that are inclined from northeast to southwest in the northern hemisphere and oppositely in the southern hemisphere. In the same way that Babcock used these ropes to explain Hale's laws of sunspot polarity, Rust and Kumar use them to explain the handedness of filaments. They picture a filament as a twisted magnetic rope that erupts into the atmosphere, properly oriented with the correct handedness (but see SOLAR PROMINENCE FORMATION).

In an elaboration of their filament model, Rust and Kumar discuss the generation and loss of helicity, the 'twist' of solar magnetic fields. Filaments often disappear by erupting bodily from the Sun. In many cases a helical structure in the filament is seen to unwind during the eruption. Rust and Kumar propose that this helical field is created somehow during the subsurface formation of a filament. Moreover they claim the sense of the filament's twist (clockwise or anticlockwise) is correlated with the direction of its axial field: sinistral filaments are right-handed helices and dextrals are left-handed. Rust and Kumar support their proposal by showing that interplanetary magnetic clouds, which originate in the eruption of a filament, have the same helicity as their

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

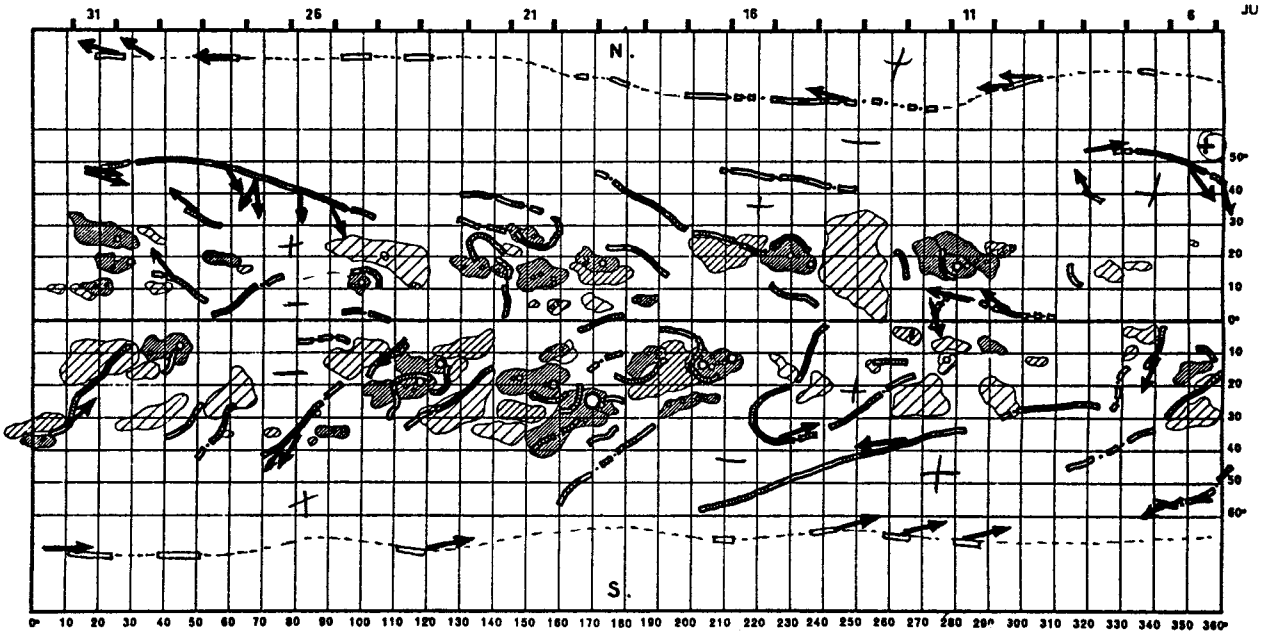


Figure 1. A map of the axial magnetic fields in filaments, over the whole Sun. (Courtesy of J-L Leroy.)

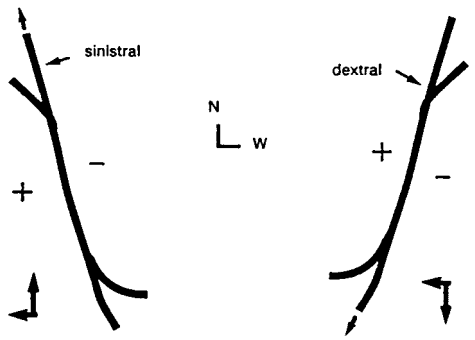


Figure 2. A diagram of sinistral and dextral filaments, showing their axial fields. The small appendages ('barbs') also indicate chirality: as seen from either side of the filament, they bear to the left in the sinistral and to the right in the dextral.

associated filaments. So not only the axial fields of filaments, but also the *twist* in these fields, may show a hemispheric preference.

E Priest and his colleagues also invoke differential rotation to form subsurface magnetic ropes in each hemisphere that have the correct orientations to explain filament handedness. But instead of postulating that the ropes emerge as a whole, they suggest that small loops bob up to the surface and link up, end to end, to form the axial field in a filament (figure 3).

Recently A Van Ballegooijen and his collaborators have explored another scheme, in which only *surface* fields and surface motions are involved. They made numerical simulations of the solar cycle, allowing bipolar active regions to emerge at the observed rates, sizes and latitudes. Simultaneously, three well-known surface flows act on these regions: differential rotation distorts the regions, supergranule cells diffuse their magnetic flux and poleward drifts carry their flux to the poles. The result is the formation of so-called switchbacks, hairpin turns in the curves that divide large areas of opposite magnetic polarity in the photosphere. The authors then examine the coronal fields these processes produce. They find that filaments forming on the preceding branches of such switchbacks would obey the hemispheric rule, but those on the following branches would not, in contradiction to the observations.

It does seem therefore that subphotospheric fields and flows are necessary to explain the observed hemispheric pattern of the chirality of filaments. Other magnetic chirality patterns have been recognized in recent years. A global view of the patterns seems to be emerging.

Bibliography

Leroy J-L, Bommier V and Sahal-Brechot S 1984 New data on the magnetic structure of quiescent prominences *Astron. Astrophys.* **131** 33–44

Martin S F, Billamoria R and Tracadas P W 1994 Magnetic configurations basic to filament channels and filaments *Solar Surface Magnetism* ed R J Rutten and C Schrijver (Berlin: Springer) pp 303–38

Priest E R, Van Ballegooijen A A and MacKay D H 1996 A model for dextral and sinistral prominences *Astrophys. J.* **460** 530–43

Rust D M and Kumar A 1994 Helical fields in solar prominences *Solar Phys.* **155** 69–97

Van Ballegooijen A A, Cartledge N and Priest E R 1998 *Astrophys. J* **501** 866–81

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

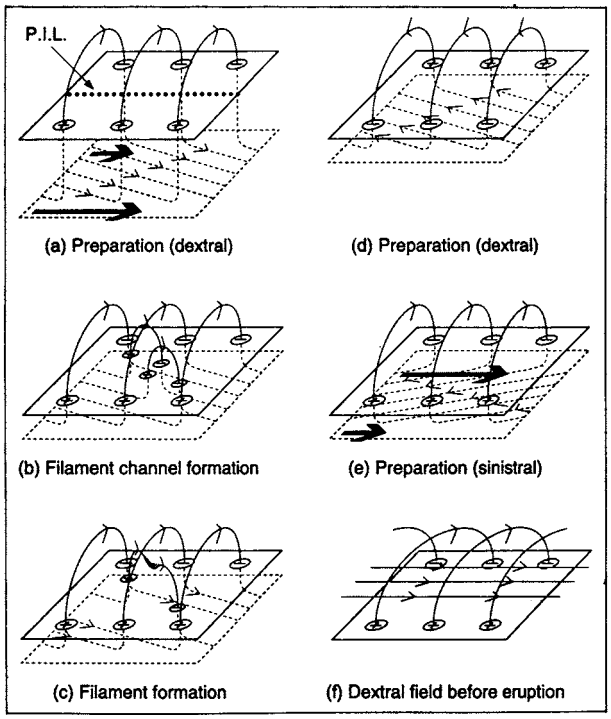


Figure 3. The formation of either a sinistral or dextral filament, according to Priest *et al* 1996: (*a*) differential rotation (heavy arrows) shears the subsurface magnetic field (dotted lines); (*b*) small loops erupt to form a filament channel; (*c*) small loops link to form a filament. The coronal arcade does not participate in the formation. (Courtesy of E Priest.)

Jack Zirker

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

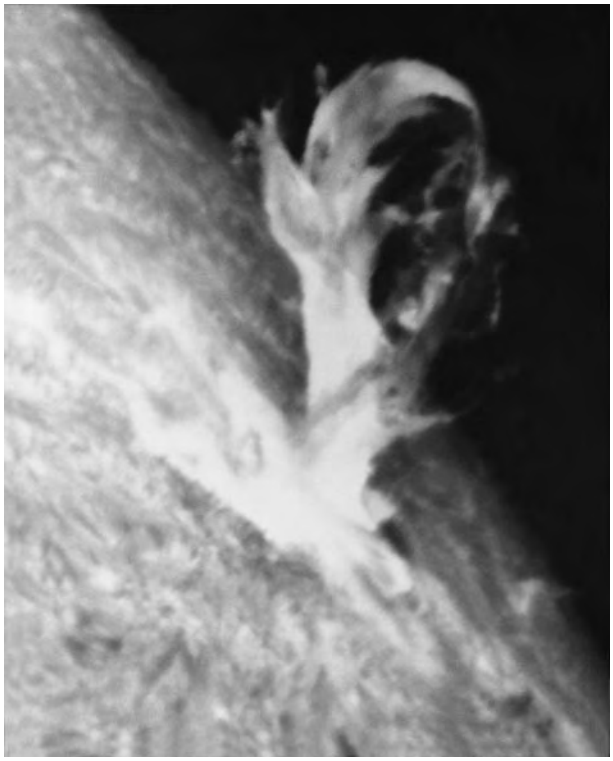


Figure 1. An ejective filament eruption exploding out of an active region near the limb. The expanding filament magnetic flux rope is seen to have overall twist and complex substructure. At this point, about 10 min into the explosion, the filament flux rope has increased in volume by a factor of ∼1000. This photograph is an Hα filtergram (from the Big Bear Solar Observatory); it shows plasma in the erupting filament that is at chromospheric temperatures (5000–10 000 K).

Solar Prominence Eruption

The prominence in a solar prominence eruption is a magnetic structure in the chromosphere and corona (see SOLAR PROMINENCES; SOLAR PROMINENCE: ACTIVE). Prior to its eruption, the prominence is visible in chromospheric images by virtue of chromospheric-temperature plasma suspended in the magnetic field, and belongs to that large class of solar magnetic structures appropriately called solar filaments because of their characteristic sinewy sigmoidal form. Hence, the term 'filament eruption' is used interchangeably with the term 'prominence eruption'. The magnetic field holding a filament is prone to undergo explosive changes in configuration. In these upheavals, because the filament material is compelled by its high conductivity to ride with the magnetic field that threads it, this material is a visible tracer of the field motion. The part of the magnetic explosion displayed by the entrained filament material is the phenomenon known as a solar filament eruption, the topic of this article. Figure 1 shows a filament eruption in progress, exploding up into the corona.

This article begins with a description of basic observed characteristics of filament eruptions, with attention to the magnetic fields, flares, and coronal mass ejections in which erupting filaments are embedded (see SOLAR FLARE OBSERVATIONS; SOLAR CORONAL MASS EJECTION: OBSERVATIONS). The present understanding of these characteristics in terms of the form and action of the magnetic field is then laid out by means of a rudimentary three-dimensional model of the field (see SOLAR FLARE MODELS; SOLAR CORONAL MASS EJECTION: THEORY). The article ends with basic questions that this picture leaves unresolved and with remarks on the observations needed to probe these questions.

Observed characteristics

Every stationary filament, whether or not it eventually erupts, stands over and tracks a neutral line (i.e. a polarity dividing line) in the photospheric magnetic flux (see SOLAR FLARE MAGNETIC FIELDS). The width of the opposite polarity flux areas bracketing the neutral line is typically considerably greater than the height of the filament, and coronal images show a magnetic arcade rooted in the opposite polarities and arching high over the filament. That is, the filament is centered low in the tunnel of the coronal arch of the overall magnetic bipole, suspended in the *core magnetic field*, the field that is rooted near the neutral line and closely envelops the neutral line. The filament and substructure striations within the filament run nearly parallel to the neutral line. This shows that the core field threading the filament runs nearly parallel to the neutral line rather than arching right over the neutral line as it would if it were near its relaxed, minimum-energy potential configuration. In contrast to the core field, the loops of the overall arch of the bipole are more nearly orthogonal to the neutral line. Thus, the form of the filament and its location within the bipole show that the core field is greatly deformed from the potential field configuration, being sheared so that it stretches along the neutral line, and that the core field is more strongly sheared than is the field farther out in the envelope of the bipole.

In a filament eruption, the active agent is the core field in and around the filament; it is the core field that explodes, exploding both itself and the filament that it carries. Thus, the size of a filament eruption is set by the length of the *sheared core field* traced by the filament at eruption onset. This length is a measure of the overall size of the closed magnetic bipole in which the filament and sheared core field are seated. Accordingly, the range of sizes of filament eruptions is more or less the range of the lengths of filaments found on the Sun, and this is roughly the range in the size of magnetic bipoles that occur on the Sun. This range extends from as large as half a solar radius (∼300 000 km) to as small or smaller than the larger bipoles found in the magnetic network ($\lesssim$10 000 km). The magnetic bipoles forming active regions with sunspots, and the filament eruptions in these regions, are usually medium sized, typically ∼30 000 km, but ranging up to as large as 100 000 km at eruption onset. The filament eruption in figure 1 is seated in an active region and is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

larger than average: the erupting filament and sheared core field were initially ∼50 000 km long. Although the largest (in span) of all filament eruptions do not occur in active regions, the larger of the filament eruptions in active regions, in particular filament eruptions from between large impacted opposite-polarity sunspots, mark the Sun's most powerful magnetic explosions. This is so because sunspots have the strongest magnetic fields found on the Sun, and consequently the sheared core fields between the largest impacted sunspots have the largest store of non-potential magnetic energy. The very largest filaments and filament eruptions are in quiet regions, on the neutral lines of the bipolar arcades under large coronal streamers. The strength of the filament core field in these is only ∼10 G, a hundred times less than in the most powerful filament eruptions from large sunspot regions. The eruption of one of these largest quiet-region filaments is spectacular, as in the example shown in figure 2, even though the total energy of the explosion is 10 to 100 times less than for the strongest ones in active regions. While the largest filament eruptions occur in only the largest bipoles, which are all in quiet regions, small filament eruptions can occur wherever small bipoles occur, which can be within active regions as well as in quiet regions. However, most of the smallest filament eruptions (from bipoles $\lesssim$10 000 km across) occur in quiet regions, due to the ubiquity of such bipoles in the magnetic network (see CHROMOSPHERE: NETWORK). Because many macrospicules look like small versions of the filament eruptions in figures 1 and 2, it is likely that many macrospicules are small (∼10 000 km) magnetic explosions similar to filament eruptions, and this raises the possibility that many spicules (which are rooted in the network and are similar to macrospicules but smaller) are still smaller analogues of filament eruptions (see CHROMOSPHERE: SPICULES).

There is enough magnetic flux all over the Sun, in all quiet regions as well as in active regions, that the entire solar atmosphere above the photosphere is permeated with magnetic field strong enough to impose its structure on the chromosphere and corona. Consequently, none of the bipoles in which filaments and filament eruptions are seated is totally isolated; instead each is hemmed in to some degree by adjacent magnetic fields. When these adjacent fields are weak relative to the eruption bipole, the eruption proceeds about as it would if there were no adjacent field. Often, some of the adjacent fields, weak or strong, are in other bipoles that have their own sheared core fields. In these cases, a filament eruption in one bipole can trigger and interact with another eruption in one or more adjacent bipoles. The morphology and development of these multiple-bipole eruptions is naturally more complex than in single-bipole events. All of the figures in this article show examples or models of single-bipole events.

The filament eruption in any one bipole, whether or not it is part of a multiple-bipole eruption, is either ejective or confined. In an *ejective filament eruption*, the erupting core field becomes an expanding, twisted, arched magnetic flux rope, the top of which escapes far out of the originally closed bipole while the ends remain rooted in the surface. This requires that the envelope of the bipole also eject along with the filament flux rope, the filament rope riding inside the distending bubble of envelope field. The filament eruptions in figures 1 and 2 are ejective. In a *confined filament eruption*, there is no distinct formation of a monolithic rising flux rope. Instead, there is complex eruptive untwisting and restructuring of the core field as the erupting core field undergoes an overall expansion. The whole eruption is confined and arrested within the enveloping bipole, and the overall magnetic arch of the bipole remains closed throughout the eruption. Figure 3 shows a confined filament eruption from beginning to end. This eruption spans the interior of an active region; so it is of average size and is seated in strong field. As do ejective eruptions, confined filament eruptions occur in quiet regions as well as in active regions, over the same wide range of size and field strength. The factors that dictate whether an eruption is to be confined or ejective are probably specifics of the pre-eruption field configuration and the strength of the envelope field relative to the core field.

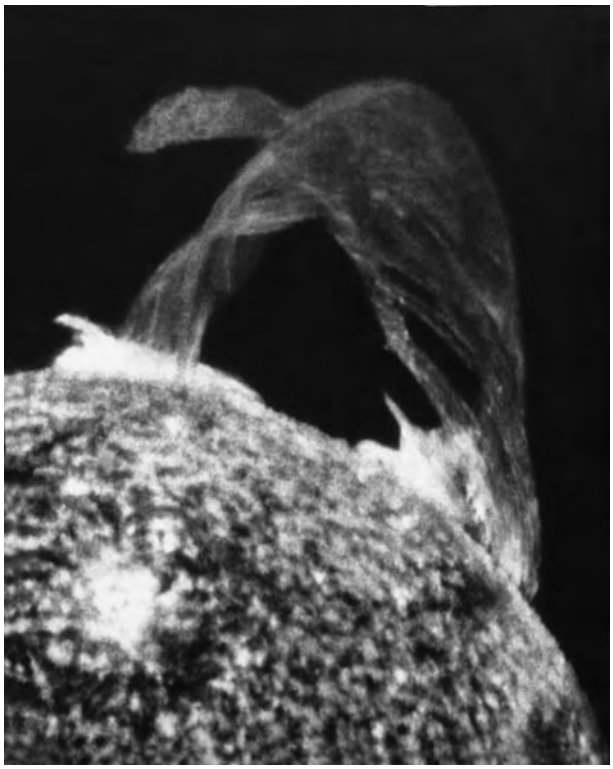


Figure 2. A large filament ejectively erupting from a quiet region on the limb. The many long striations show that the filament material is strung in a twisted magnetic flux rope. This image is an He II 304 A spectroheliogram (from the Naval Research Laboratory's extreme ultraviolet spectroheliograph on Skylab); it shows plasma at transition-region temperatures (∼80 000 K).

Every filament eruption, confined or ejective, is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

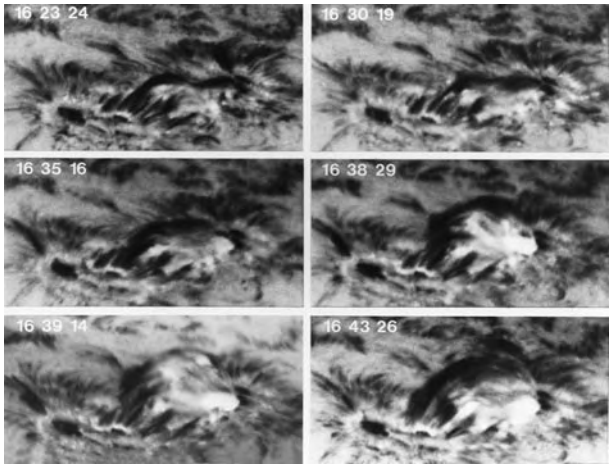


Figure 3. A confined filament eruption in an active region. In this 20 min sequence, the filament core field untwists and expands as it erupts to become a magnetic arch in the interior of this overall bipolar active region. The active-region's opposite-polarity sunspots are seen at opposite ends of the erupting filament and final arch. These images are blue-wing Hα filtergrams from the Big Bear Solar Observatory.

accompanied by increased coronal heating (see CORONAL HEATING MECHANISMS) in the bipole in which it is seated, and every ejective filament eruption, in addition, accompanies the ejection of the corona in and above the bipole, i.e. is an intrinsic part of a larger surrounding coronal mass ejection, the core of which is the ejected filament. Figure 4 shows the coronal mass ejection bubble around an ejected quiet-region filament. The bright twisted loop in the core of the bubble is the filament flux rope from a quiet-region filament eruption like that in figure 2, but at a later phase, farther out in the corona. The corresponding coronal mass ejections for ejective filament eruptions from active regions are similar in overall form to those from quiet regions, but usually accelerate more rapidly to final speeds that are two to five times faster (as fast as 1000–2000 km s^{-1}). Because these fast ejective filament eruptions are always accompanied by a strong flare in the ejective bipole, they are aptly called flare sprays. The burst of heating in any large strong-field filament eruption in an active region produces a major flare in coronal emission, increasing the Sun's total x-ray luminosity by a factor of 10 to 1000. In these eruptions and in many quiet-region eruptions, the enhanced coronal heating begins in the erupting core field at eruption onset, rendering this field visible in coronal images. Coronal images during the onset phase often show the whole core field to have the overall shape of an S or backward S. Later in ejective eruptions, after the filament flux rope has been ejected far out of the bipole, the coronal images show a bright, growing coronal arcade straddling the neutral line. In large weak-field ejective eruptions like those in figures 2 and 4, the coronal heating can be so weak as to be hardly noticeable in the global x-ray luminosity, but is usually strong enough to be

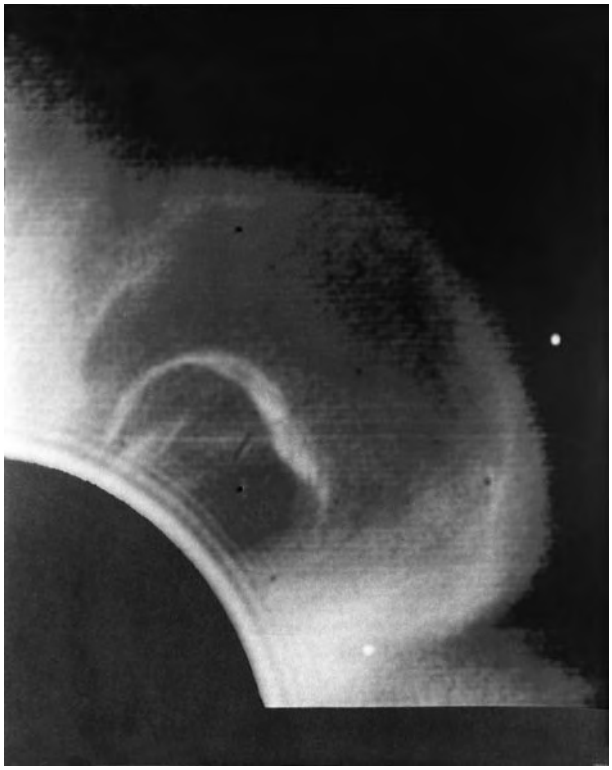


Figure 4. The filament flux rope in the core of the coronal mass ejection from a large quiet-region ejective filament eruption. The ejecting magnetic field, which is traced by and carries the ejecting plasma within it, is seen in the act of exploding out through the corona and into the solar wind. The occulting disk extends half a solar radius above the limb. This white-light image is from the High Altitude Observatory's coronagraph on the Solar Maximum Mission.

seen in coronal x-ray images, at least in the post-ejection coronal arcade. In active-region eruptions the heating also produces flare brightening of the chromosphere at the feet of the field lines threading the coronal heating. In confined eruptions and early in ejective eruptions, this chromospheric brightening occurs close along the neutral line under the filament, as the filament erupts (as in figures 1 and 3). In confined eruptions, coronal images show the filament core field brightening and expanding as it transforms into an arch standing on the chromospheric brightening, as does the chromospheric arch that appears in figure 3 as the eruption ends. In ejective eruptions after the filament is ejected, the chromospheric flare brightening is in two separating ribbons at the feet of the coronal arcade.

Form, energy and action of the magnetic field

The filament eruptions in flares and coronal mass ejections, together with magnetograms (maps of the strength and distribution of the photospheric roots of the magnetic field) reveal that these events are fundamentally magnetic explosions, and that the exploding magnetic field has

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

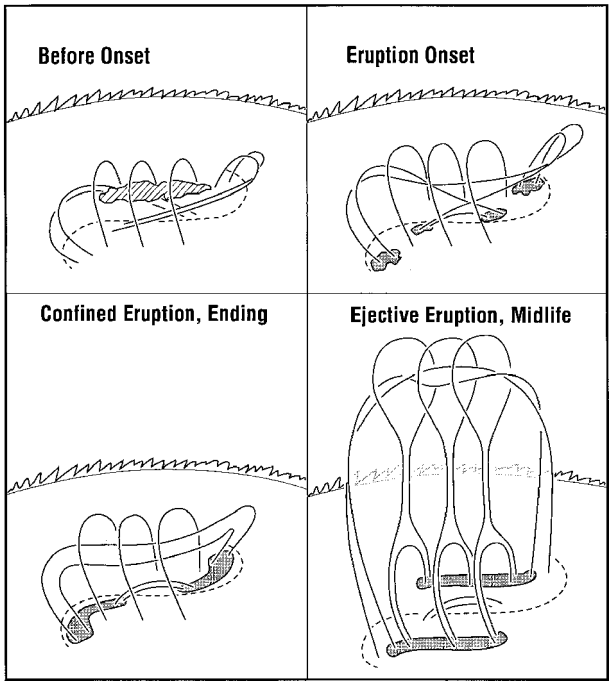


Figure 5. Three-dimensional sketches of the magnetic field before and during filament eruptions. The dashed curve is the polarity neutral line in the photosphere. The diagonally hatched object in the upper left panel is the pre-eruption filament; for clarity of the eruption of the core field that carries the erupting filament material, this material is not shown in the other panels. The gray areas are chromospheric flare brightenings. The ragged arc in the background in the chromospheric limb. Early in the eruption, reconnection occurs in the sheared core field in or under the filament, as in the upper right panel. In ejective eruptions, after the filament flux rope forms and ejects, the ejected envelope field of the pre-eruption arcade is wrapped around the filament core flux rope and forms much of the surrounding expanding coronal mass ejection bubble that envelops the filament core rope, and the stretched legs of the envelope field reconnect to form the growing coronal arcade rooted in the separating chromospheric flare ribbons, as in the lower right panel.

the following characteristics: (a) The pre-eruption field is a closed bipole. (b) The field along the neutral line through the core of the bipole is greatly sheared and has an overall sigmoidal form, so that it traces the neutral line through the bipole and has oppositely curved elbows on opposite ends. (c) The pre-eruption core field has a large store of non-potential magnetic energy. (d) Much of this energy is released by the untwisting and expansion of the filament core field as it erupts. (e) The released energy is comparable to the total energy expended in the accompanying flare and/or coronal mass ejection (up to $\sim 10^{32}$ erg in the largest active-region events).

The evidence for the field configuration and its transformation in the filament-eruption explosion is outlined in the previous section. The sketches in figure 5 show the three-dimensional form, action and reconnection of the magnetic field implied by the observed characteristics of filament eruptions and their observed involvement in flares and coronal mass ejections. In confined eruptions, the overall bipole remains closed, containing and arresting the eruption within it. In ejective eruptions, the entire bipole eruptively opens and simultaneously recloses via the sequence and topology of the reconnection. The expanding ejected filament flux rope and envelope field can continue on to become a coronal mass ejection that blasts out into the solar wind.

Current research

The model sketched in figure 5 displays, for single-bipole filament eruptions, the typical three-dimensional form of the magnetic field and its change in topology as it explodes, to the extent that these have been revealed by the best magnetograms and chromospheric images from the ground together with the best transition-region and coronal images from space. These observations indicate that reconnection occurs, beginning low in the sheared core field early in the eruption, but it is not clear why the reconnection starts or whether the reconnection is essential for the explosion. Basic questions left unresolved by present observations are:

(a) How does the metastable pre-eruption field configuration originate?
(b) How does the configuration destabilize to explode?
(c) Is reconnection necessary for either the triggering or the growth of the explosion?

The observations most needed for answering these questions are vector magnetograms of substantially better sensitivity, spatial resolution and continuity than are now possible from the ground. Vector magnetograms provide a direct measure of the shear in the magnetic field in the photosphere. To probe why and how filaments erupt, the vector magnetograms need to track the development of the entire bipole in which the sheared core field builds up and explodes. It is anticipated that the required improved vector magnetograms, along with simultaneous high-resolution images of the chromosphere, transition region and corona, will be provided from space by the Solar-B Mission currently being developed by Japan, jointly with the United States and the United Kingdom, for launch in 2004.

Bibliography

Kahler S W 1992 Solar flares and coronal mass ejections *Ann. Rev. Astron. Astrophys.* **30** 113–41

Moore R L 1988 Evidence that magnetic energy shedding in solar filament eruptions is the drive in accompanying flares and coronal mass ejections *Astrophys. J.* **324** 1132–7

Moore R L, Schmieder B, Hathaway D H and Tarbell T D 1997 3D magnetic field configuration late in a large two-ribbon flare *Solar Phys.* **176** 153–69

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Rust D M *et al* 1980 Mass ejections *Solar Flares* ed P A Sturrock (Boulder, Co: Colorado Associated University Press) pp 273–339
Svestka Z, Jackson B V and Machado M E 1992 *Eruptive Solar Flares* (Berlin: Springer)

Ron Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominence Fine Structure

A clear feature of high-resolution observations of quiescent PROMINENCES, both on the disk and at the limb, is the presence of fine-scale fibril structures. The fine structure of prominences provides vital clues to questions such as how prominences form and how they are supported. Magnetic fields permeate the prominence plasma and play a decisive role in shaping the fine structure and its dynamics. Conversely, the shape, orientation and possibly the strength of the magnetic field may be inferred from the observed substructures.

Observations

Limb observations leave little doubt that prominences consist of thin threads and knots. The threads are typically 5000–35 000 km long, and from 200 km to 2000 km thick. The thread thickness appears to increase with height above the solar limb. Threads and bright knots are of similar sizes (figure 1). Since the observed widths of the thinnest threads are of the same size as the instrumental resolution limit, their true widths may be a lot less.

In filaments viewed from above against the disk many threads are inclined to the direction of the long axis of the filament by about 30° (figure 2). Observations near or above the limb reveal a lot of fine-structure inclined by similar amount relative to vertical.

Characteristic properties of prominence threads are listed in table 1.

Threads and the overall structure of prominences

The thin threads (fibrils) of prominences constitute characteristic medium-scale structures.

At some distance from the disk center, quiescent filaments often display a well-defined, sharp upper edge, which is called the filament *spine*. A spine often consists of several to many thin, long absorbing threads. When observed at the solar limb the threads which are associated with prominence spines are not always conspicuous, depending on the perspective from which the prominence is viewed. But overall, the threads then appear to run more or less parallel to the solar surface. They also run parallel to a polarity reversal boundary in the photospheric magnetic fields below. In addition to the threads along the top of the spine, there can be low or intermediate horizontal threads, as well.

Often the ends of a filament fan out in the shape of a thin 'shroud'. This is where the magnetic fields of the threads are thought to connect to the chromosphere; the splayed threads at ends are associated with opposite polarity magnetic fields in the photosphere below. In other cases, at the ends of filaments, the fine threads become vanishingly thin and only an occasional flowing knot of mass reveals the connection of the body of the filament to the chromosphere.

At semiregular intervals along a filament, some threads converge towards one or more points in the chromosphere to the side of the spine. These structures are identified as major 'barbs'. However, amid the multiple-thread barbs, single-thread barbs also join the spine to the chromosphere. These barbs typically curve downward to the chromosphere and, hence, have a large vertical component. Because prominences are dynamic, at any moment in time, some threads can be in transit and might not appear to connect to the chromosphere.

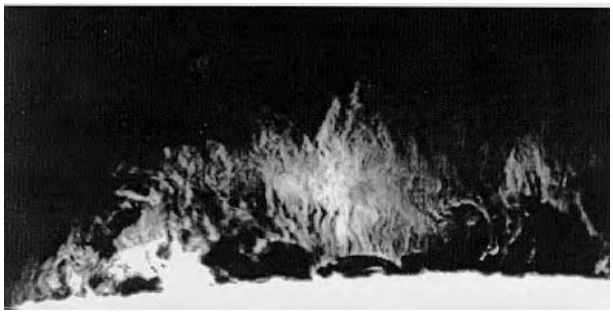


Figure 1. A quiescent prominence showing numerous vertical treads observed in the Ca II K line with the Vacuum Tower Telescope of National Solar Observatories. (Courtesy of Dr R B Dunn.)

The visibility of prominence threads

The observability of prominence threads is restricted by their contrast against the background and the resolution of the observing instrument. High-resolution Hα filtergrams from La Palma have resolved threads with thickness down to about $\frac{1}{3}$ arcsec (250 km on the Sun). However, the true visibility of a prominence thread in hydrogen Balmer lines is a function of temperature and density of the 'cold' plasma. The full length of a magnetic flux tube that forms a thread may escape detection if the plasma parameters vary along the tube.

Number density and filling factor

The number of thin and barely resolvable threads in the line of sight may be estimated by assuming that all threads have the same brightness and that the distribution of velocities is Gaussian. The observed intensity will then increase in proportion to the number of threads. Few threads in the line of sight give rise to large variations in line shift and in line widths, whereas averaging over many threads will produce less variation and wider line widths. Multithread models that are based on Hα spectral data give numbers of threads ranging from a very few up to 100 per resolution element (a few arcsec^2). A recent study based on N V 1238 Å line spectra of a large prominence observed with the SOHO/SUMER instrument gives 15–20 unresolved threads.

It is common practice in solar physics to specify the fraction of the overall volume filled with emitting matter, i.e. emitting and/or absorbing threads, as the filling factor f. Values of f between 0.01 and 0.1 are commonly quoted for prominences.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

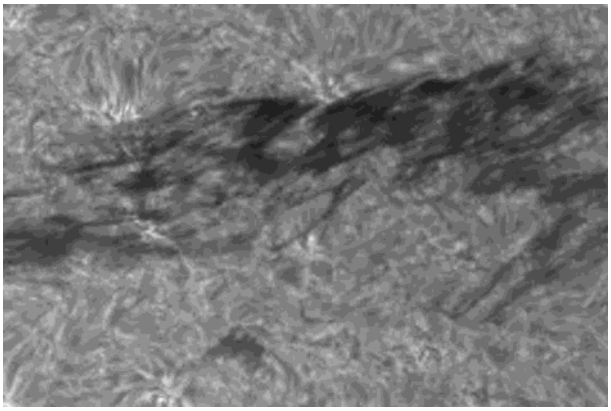


Figure 2. A large polar crown prominences observed with the Swedish Vacuum Solar Telescope on La Palma on 19 June 1998. The Hα filtergram shows that the prominence consists of numerous fine threads which are oriented at an angle of 20°–30° relative to the long axis of the filament.

Lifetimes

In hedgerow prominences (Tandberg-Hanssen 1995) the lifetime of vertical threads is $\leq$10 min. High-resolution Hα filtergrams obtained with the SVST, La Palma (see figure 2), show individual thin filament threads appearing and disappearing in the course of 1 min. This suggests that prominence plasma may be undergoing rapid variations in temperature and/or density. Such short lifetimes might imply that the prominence plasma is not in a stationary radiative state.

Dynamics

The term quiescent is somewhat misleading in describing stable, large prominences, since their small-scale structure is strikingly dynamic. One sees small-scale line-of-sight velocities of a few km s^{-1}, both in horizontal and vertical directions. Limb observations show an apparent random motion of the threads at velocities up to about 8 km s^{-1}. Doppler measurements in disk filaments show both upwards and downwards mass motion (cf Schmieder 1988). The dominant vertical motion in the main body of filaments appears to be oscillatory. SOLAR PROMINENCE OSCILLATIONS in velocity and intensity have periods ranging from less than 1 min to hours; the shorter periods are presumably associated with the small-scale structure.

Time series of Hα filtergrams of large quiescent prominences show bidirectional motions lasting for up to 7 h. Such motions, which are noticed throughout the full body of prominences, in the spine as well as in the more vertical bundles of threads in barbs, suggest that the prominence plasma is flowing along the magnetic field lines.

Interpretation of observations

The magnetic nature of prominence threads

It is generally assumed that prominence threads outline a fine-scale structure of the magnetic fields of solar prominences. This view is supported by the fact that threads typically are inclined by 20°–30° relative to the prominence long axis, which is similar to the general orientation of the magnetic field.

Since the PLASMA β $<$ 1 in solar prominences, i.e. the ratio of gas pressure to magnetic pressure, it is somewhat surprising that their magnetic fields are so highly structured. Therefore, could the fine threads seen in Hα rather be 'veins' of denser and 'cooler' plasma in a relatively uniform magnetic field, instead of individual flux ropes? A uniform field would seem to be at variance with the observed fluctuation in line-of-sight velocity of individual neighboring threads, which seems to imply that the threads represent separate flux tubes. A possible weak interaction between neighboring threads is indicated from observations showing that the *velocity* threads seem to be slightly thicker than the *density–intensity* threads.

Table 1. Characteristic values of prominence threads.

	Value
Width	200–2000 km
Length	5000–35 000 km
Lifetime	1–10 min
Velocity along spine	5–15 km s^{-1}
Velocity variation	
$\perp$ threads	$\pm$5–12 km s^{-1}
$\parallel$ threads	$\pm$1–5 km s^{-1}
Magnetic flux density	4–20 G

Temperatures

The strong EUV line emissions (see SOLAR SPECTROSCOPY: ULTRAVIOLET AND EXTREME ULTRAVIOLET EMISSION) show that solar prominences contain gas at all temperatures between 10^4 K and 10^6 K. Recent observations with the SUMER instrument on board SOHO show that the thread-structure prevails up to temperature $\sim 10^5$ K. Emission measure studies of EUV line emission from quiescent prominences show that the emitting structures have subresolution geometries. The nature of these geometries is still unknown, but the following two scenarios are being considered:

1. Being situated in a hot coronal cavity with cold cores, the threads are 'wrapped' in a thin transition region in a direction perpendicular to the magnetic field.
2. The individual threads are essentially isothermal, but a given thread may attain various temperatures between 10^4 K and 10^6 K.

The tenuous interthread medium could be either hotter or cooler than the threads themselves. It is generally assumed, however, that prominences and their substructures are embedded in hot CORONAL CAVITIES.

Relation to photosphere below

Assuming that single fibrils or threads form on individual thin magnetic flux tubes, these must evidently tie back

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

down onto the chromosphere and photosphere on both sides of the prominence. This tying is most conspicuous where several threads cluster and form barbs. Recent observational results suggest that the magnetic fields of many barbs do *not* connect to the well-known magnetic network that is located at the boundaries of the supergranulation cells below (see SOLAR PHOTOSPHERE: SUPERGRANULATION), but rather to the weaker intra-cell field. Additionally, the lack of merging or embedding of the barb footpoint into the network implies that many barb ends are rooted in the half of the bipolar intra-cell fields that are opposite in polarity to the network magnetic fields. This would be consistent with the few examples of barbs whose rooting is associated with the relatively strong magnetic fields of ephemeral ACTIVE REGIONS or very small active regions. In these cases, the rooting has been identified with the polarity opposite to the network instead of to the polarity that is the same as the network (Martin, 1998).

It is likely that an exchange of mass between the prominence and the photosphere below takes place via these magnetic anchor points.

Modeling of threads

The physical cause of threads or fibrils and their effects on the MAGNETOHYDRODYNAMICS of prominences are not well understood.

Various static models have been put forward to explain the thread-like substructure of prominences. These models all assume a basically horizontal magnetic field to support the plasma.

It is likely that the thermal instability which might be triggering the formation of a prominence as a whole does so preferentially in a filamentary manner. Equilibrium solutions are found that seem to reproduce well the horizontal threads of prominence spines.

On the other hand, modeling of vertical threads that are perpendicular to a horizontal magnetic field meets severe difficulties. At a temperature of 8000 K the hydrostatic scale height in the lower solar corona would be about 200 km. Observations show, on the other hand, that vertical threads are much longer by at least two orders of magnitude. The length-scale associated with normal thermal conductivity perpendicular to the magnetic field becomes very short compared with the length of a typical thread. It has therefore been suggested that anomalous perpendicular thermal conduction, possibly due to turbulence, might produce length scales similar to observed thread lengths. Another idea is that vertical threads could be the result of gravitational Rayleigh–Taylor instability, but such a mechanism would require an extremely high viscosity in order to produce the observed width of a thread.

Other studies of stationary prominence thread models include ambipolar diffusion perpendicular to a horizontal magnetic field. Furthermore, from magnetostatic theory one may derive equilibrium solutions that give long vertical fibrils whose weight may be supported by locally bowed magnetic fields (Lorentz force), which implies strong localized electric currents.

However, none of the static models seems able to explain the dynamic character of prominence threads. The inferred field-aligned flows would suggest that the magnetic fields in prominences barbs are nearly vertical, in which case one would expect to see the cool prominence plasma in the threads sliding down towards the photosphere at 'free fall' velocities. Neither Doppler shifts in prominences seen against the solar disk nor movies of prominences at the solar limb show evidence for 'free fall' velocities at locations where vertical thread structure is clearly developed. The propelling force exerted by weakly damped, high-frequency Alfvén waves is currently being investigated in a search for mechanisms that may counteract gravity in nearly vertical magnetic flux tubes.

Future prospects

Today's solar physicists do not yet fully understand how solar prominences are formed, supported and finally destroyed. More high-angular-resolution, high-cadence studies of the enigmatic fine structure will provide better insight in the basic physics of prominences. Lifetimes of only a few minutes and flow velocities of some km s^{-1} along horizontal and vertical prominence threads, as determined from the best observations of today, could imply that the mass contained in an individual fine thread is continuously being drained and replenished. The notable dynamical and small-scale character of the substructure is an important feature and consideration in modeling of prominence support in the field of gravity. Likewise, the mechanism by which mass is supplied to prominences, through either condensation or inflow from the photosphere below, awaits further observational and theoretical investigations.

Bibliography

Engvold O 1998 *New Perspectives on Solar Prominences: IAU Colloq. 167 (ASP Conf. Ser. 150)* ed D Webb, D Rust and B Schmieder, p 23

Martin S F 1998 *Solar Physics* **182** 107

Ruzdjak V and Tandberg-Hanssen E (eds) 1989 *Dynamics of Quiescent Prominences: Proc. IAU Colloq. 117 (Hvar, SR Croatia, Yugoslavia, 1989)*

Schmieder B 1989 *Dynamics and Structure of Quiescent Solar Prominences* ed E R Priest (Dordrecht: Kluwer) p 15

Tandberg-Hanssen E 1995 *The Nature of Solar Prominences* (Dordrecht: Kluwer)

Oddbjørn Engvold

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominence Formation

Most aspects of PROMINENCE formation are only observable against the solar disk where prominences are called 'FILAMENTS'. The formation of filaments can be described as a two-stage process: (1) the development of an environment necessary for filaments and (2) the formation of the filament. After the filament environment is established, the physics of formation of the filament involves two interrelated processes: the formation of the filament magnetic field and the filling of a volume of space within or around that magnetic field with the mass of the filament. Although prominences have been frequently modeled as if they were static structures, there is both old and recent evidence that the formation is a continuous process. Mass is continuously entering and exiting the filament magnetic field throughout its lifetime. It is likely, but not proven, that the magnetic field is also continuously increasing, decreasing or changing. If so, it is necessary to know the properties of the filament environment in order to understand the source of the magnetic field and mass flows in filaments.

The features of a filament environment

The photospheric and chromospheric filament channel

Characteristic features of the filament environment are revealed in the photosphere, the chromosphere and corona. Filaments only form in close association with photospheric magnetic field boundaries where there is a sharp reversal in the direction of the photospheric magnetic field from outward to inward or vice versa. While filaments are not found along all such boundaries, the presence of an apparent 'polarity reversal boundary' is a necessary condition for prominence formation (reviewed in Martin 1990, 1998).

Vector magnetograms have been successfully used to further identify the photospheric magnetic field environment of filaments in active regions. Filaments selectively form only at those boundaries where the magnetic field is both along the polarity reversal boundary and nearly parallel to the photospheric solar surface. When this condition exists, the magnetic field at the polarity reversal boundary does not have a loop configuration across this boundary at either the photosphere or the chromosphere or within the volume of the filament. Instead, in the dimension across the boundary, the magnetic field has a rotational component; precisely at the polarity reversal boundary, the magnetic field, at the photosphere, chromosphere and within the filament above, is parallel or nearly parallel to the boundary (Martin *et al* 1994). This field configuration across the polarity reversal boundary is also described as 'magnetic shear'.

In magnetograms at the photosphere, the polarity reversal boundary is identified by a series of sites where positive and negative network or active region magnetic fields come into contact. Thus, at any given time, the polarity reversal boundary has a precise location; the change in direction of the field at the photosphere occurs in a volume smaller than the spatial resolution of current-day magnetograms. However, this polarity inversion is also a flexible boundary; its exact location at any time is dependent on the slow migration of active region and network magnetic fields of either polarity toward or away from the opposite polarity.

In and around active regions, the chromosphere is largely composed of ubiquitous fine structure called 'fibrils'. They are aligned with the magnetic field. When the fibrils are aligned along a polarity reversal boundary, one of the conditions of filament formation has been satisfied; a chromospheric 'filament channel' has formed and its magnetic field pattern is like the photospheric field pattern (see SOLAR FILAMENT CHANNELS). The channel is more clearly seen when not occupied by a filament. An example of an empty filament channel near an active region with sunspots is shown in figure 1.

In active regions the channel is often extremely narrow, sometimes only 1000 km wide. The length is typically a few to many tens of thousands of kilometers. As the magnetic fields of active regions disperse with age, the magnetic flux density decreases and the scale of the filament channel increases proportionately. In intermediate cases, filament channels are more typically 10 000–30 000 km wide and several hundred thousand kilometers long.

Filament channels also exist along polarity reversal boundaries where the magnetic flux density of the network is very low such as in the polar areas. In these cases, the polarity reversal boundary is identified in magnetograms only by a series of discrete points at the boundaries of the supergranule solar convection cells where opposite-polarity network magnetic fields have become abutted. In these circumstances of low magnetic flux density, the fibril orientations are nearly random. However, the channel (excluding the filament) is characterized nevertheless by an absence of vertical coronal structures and a dominance of horizontal fields. The horizontal fibril structure simply lacks the organization found in filament channels of higher magnetic flux density.

The overlying arcade

Every filament forms beneath a coronal loop system which resembles a tunnel. This system of CORONAL LOOPS is also known as the 'overlying arcade' as it arches over the filament and connects opposite-polarity magnetic fields in the photosphere on the two sides of the filament. These loops and photospheric magnetic fields are the fields of active regions or decaying active regions, fields between adjacent active regions (as in figure 1) or those between merged remnant fields of many active regions. More filaments form beneath coronal arcades joining adjacent active regions than between the loops within an original bipolar active region. There is a relatively large separation between the overlying arcade and the filament both in height and laterally.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

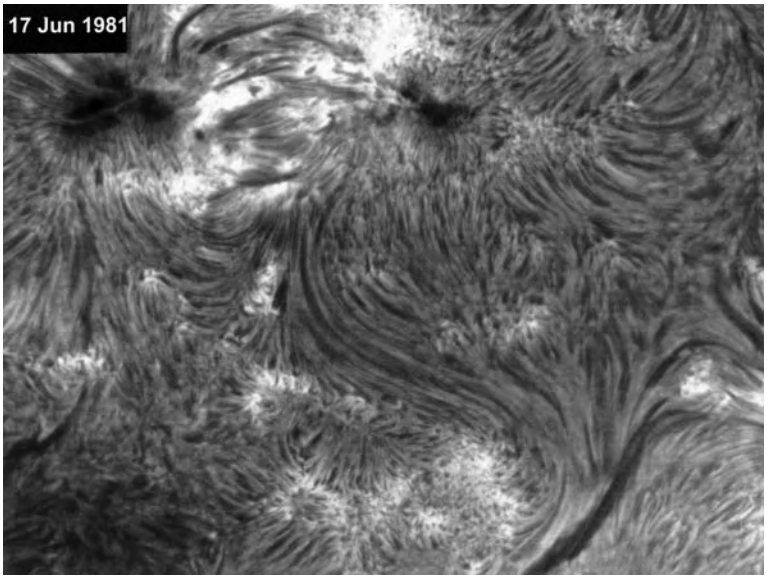


Figure 1. The S-shaped trail of fibrils in the middle of this image is a filament channel. A filament channel typically lies between opposite-polarity magnetic fields at the sides of the channel. Its ends connect to other opposite-polarity magnetic fields. In this case, the channel joins the active region with sunspots in the upper part of the image and another filament channel containing a filament in the lower right. When a filament forms, its axis is aligned with the channel. This H alpha image was recorded at the Big Bear Solar Observatory on 17 June 1981.

The cavity around prominences

Another key feature of the filament environment is the 'filament cavity'. It is a volume of space above and to the sides of prominences where the coronal density is very low at all wavelengths. The physical nature of the cavity has been a mystery. On the quiet Sun, the magnetic flux density at the photosphere beneath the cavity is much less than at the base of the overlying arcade; additionally, the aligned fibrils indicate that the network field becomes mostly horizontal at the base of the cavity. This means that there is very little photospheric magnetic field threading through this part of the corona between the filament and the coronal arcade which defines the outer boundary of the cavity. For pressure balance, the low plasma density implies a slightly higher magnetic field in the cavity rather than a lower magnetic field. This apparent discrepancy of implied higher magnetic field without any associated source of observable photospheric magnetic flux might be resolved by further consideration of new information on the opposing chiralities of the filament channel and the filament (see SOLAR PROMINENCE CHIRALITY). Currently, there is no consensus on the physics of the cavity.

Chirality of filaments and their environment

The chromospheric and photospheric filament channel and the overlying coronal arcade have a property known as 'chirality' which means 'handedness'. The chirality of the pre-filament environment is a new key to filament formation. A filament only forms with the same sign of chirality as its channel. For active-region and intermediate filaments, this leads to the partial alignment of some of the filament structure with the chromospheric fibrils as seen in figure 2 (see also CHROMOSPHERE: FIBRILS). The sinistral (left-handed) filament and channel are on the left in figure 2 and the dextral (right-handed) filament and channel are on the right. The handedness is defined for each solar feature by the orientation of its fine structure relative to the polarity reversal boundary or to the axis of the filament. Looking along the long axis of the filaments in figure 2, the filament fine structures form acute angles to the right of the long axis on the dextral filament but to the left on the sinistral filament.

The coronal arcade is the first condition to be established before a filament can form. The coronal loops within the arcade are either left skewed or right skewed relative to the polarity reversal boundary which it straddles (Martin and McAllister 1995). Again, looking along the polarity boundary, if a coronal loop makes an acute angle to the right, it is defined as right skewed and similarly, if it makes an acute angle to the left, it is left skewed.

The filament channel begins its existence with the reorientation of the chromospheric fibrils; the direction of reorientation determines the chirality of the channel. The fibrils become skewed relative to the polarity reversal boundary in the opposite sense to the overlying arcade. Hence the channel develops the opposite chirality relative to the arcade. When the fibrils above the polarity reversal boundary are completely aligned with the boundary, the channel is fully formed. At this stage, the environmental conditions are suitable for the formation of a filament of the same chirality as the filament channel and opposite

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

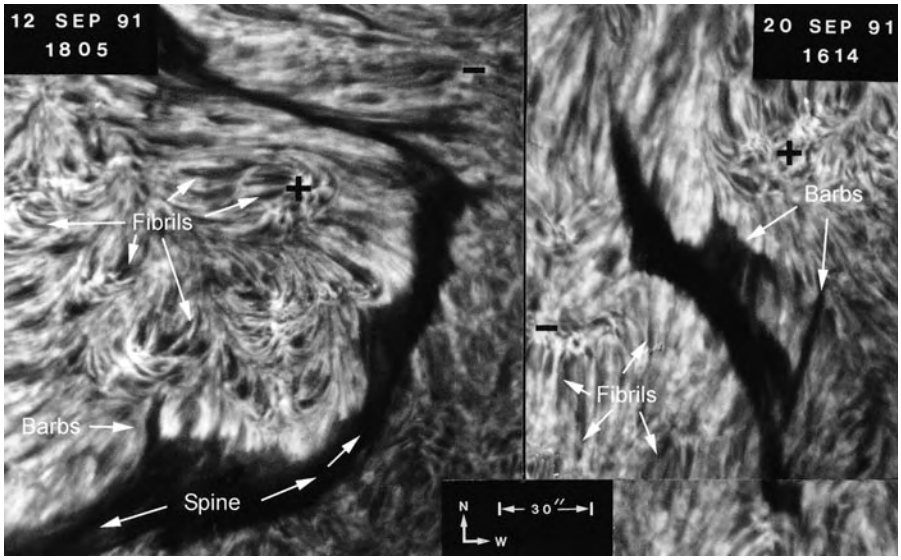


Figure 2. Example of a sinistral filament (left) and a dextral filament (right). This perspective, mostly looking at the filament from above, reveals the approximate alignment of the fine structure of the filament barbs with the chromospheric fibrils. From photographs recorded though an H alpha filter at the Big Bear Solar Observatory.

in chirality to the overlying arcade. The seminal example of channel formation followed by filament formation is illustrated by Gaizauskas *et al* (1997).

Features of the same chirality merge with one another. Examples are the merging of filaments and filament channels on the Sun to make much longer channels and filaments. Filaments and other solar features of opposite chirality remain separate and do not merge (Martin *et al* 1994). Because filaments and their overlying arcades are of opposite chirality, they too should be expected to remain isolated from each other; indeed, they are observed to be separated by the filament cavity. For this reason, it now appears that filaments cannot be formed from the magnetic fields of their overlying arcades. This is a significant new constraint not previously imposed on filament models.

Chirality can be directly interpreted as helicity of the same sign. The significance of helicity in the development of filament channels and filaments and their overlying arcades is not yet fully understood. Filaments and filament channels might develop the opposite sign of helicity from their overlying coronal arcades as a requirement of the conservation of helicity. Another idea is that the eruption of filaments (Rust 1994; see also SOLAR PROMINENCE ERUPTION) and their overlying arcades is a means by which the build-up of excess helicity is removed from the Sun. It has been shown that the chirality of filaments (Rust 1994) and coronal arcades (Martin and McAllister 1997) is directly related to the helicity of interplanetary clouds. Interplanetary magnetic clouds are the expelled parts of coronal arcades and their embedded filament magnetic fields.

Filament formation

Definition of terms

The formation of the magnetic field of a filament can be considered as two stages. The first is the formation of the axial magnetic field along the spine. The spine is the highest horizontal part of the filament. It is thought to be common to all filaments. The mass which forms the spine of filaments only accumulates precisely along and above the boundary between the opposite polarity magnetic fields at the photosphere. The spine lies beneath a coronal loop system but is separated from the loop system by the volume of space called the prominence cavity.

Barbs are appendages along the sides of filaments which extend from the spine to the chromosphere (see figure 2). The pattern of deviation of the barbs from the spine is related to their chirality while the average angle of deviation appears to be inversely related to the magnetic flux density of the surrounding active region or network magnetic fields.

Formation of the spine

There are many alternative models for the structure and formation of the magnetic field of filaments (see SOLAR PROMINENCE MODELS). Until recently, most models have only attempted to explain the spine. There is no consensus on a model for magnetic field structure or mode of formation of spines or barbs. However, there are specific observed properties and dynamics which are relevant to the formation.

The dynamics of filaments, as well as their location along polarity reversal boundaries, indicates that changes in the magnetic field precisely at the polarity inversion are important to the formation of the spine. Filaments

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

only form when opposite polarity network magnetic fields converge and encounter one another. On encounter, the opposite polarity patches of magnetic field disappear or 'cancel'. Cancellation is defined as the mutual disappearance of magnetic flux of both polarities at a common boundary. The convergence and cancellation of active region or network magnetic fields is a necessary condition for the formation of a filament spine (Martin 1990) but the cancelling magnetic fields can be interpreted in a number of different ways. There is yet no consensus on a single interpretation. Hence prominence models differ widely.

After the specific environment is established, and convergence and cancellation are occurring, pre-filament mass begins to accumulate along the polarity inversion within the cavity of the coronal arcade. The mass can appear as condensations along the polarity inversion above the chromospheric fibrils. It can also appear to be injected along the polarity inversion by surges associated with small flares. In other cases, as in figure 3, the filament appears to grow as an extension of the spine of an initial small filament. The initial development by apparent injection is common in active regions while apparent condensations are more characteristic of filaments forming on the quiet Sun. However, even when apparent condensation is observed, the filament mass is thought to come from the chromosphere. It was previously assumed that the injection was usually not seen because the temperature is too high or the density too low for the mass to be visible. However, high-resolution observations of filaments at diverse wavelengths now reveal some mass inflows from the chromosphere.

The example of filament formation in figure 3 shows a continuity between the classic active region filament and a quiescent filament; the upper end is characteristic of active region filaments and the lower end is characteristic of quiescent filaments. Such examples demonstrate that separate models are not required for filament formation in active regions and on the quiet Sun.

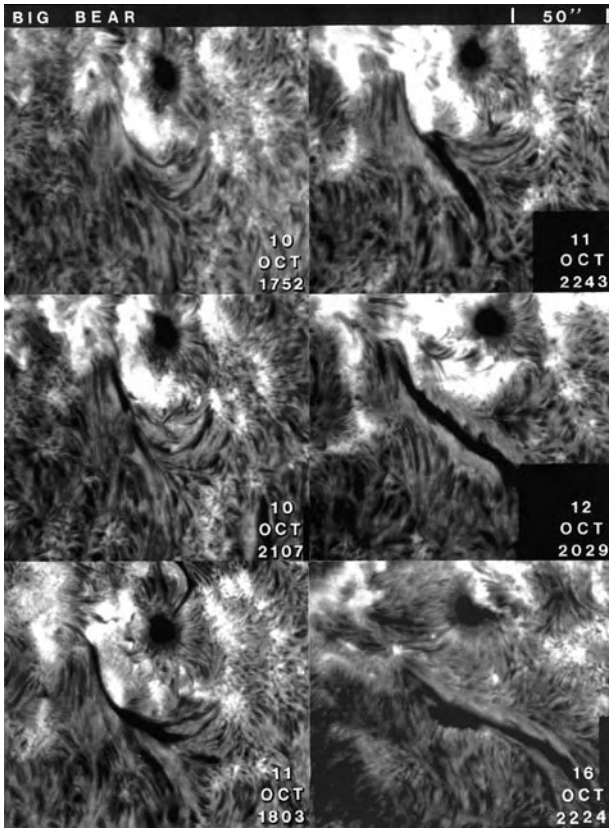


Figure 3. The formation and growth are seen over a 3 day interval from 10 to 12 October. The growth appears as an extension of an initial small filament in the active region. As the filament extends away from the active region, it begins to form conspicuous barbs characteristic of intermediate and quiescent filaments. From a photographic series recorded in H alpha at the Big Bear Solar Observatory.

Formation of barbs

If barbs are seen on active-region filaments within bright plage, they are nearly aligned with the spine. Such filaments are low and narrow and are bounded by regions of high magnetic flux density. They do not necessarily have barbs. Therefore, the barbs are regarded as secondary to the formation of the spine. Barb formation is prevalent at the stage when the active-region fields are dispersed by the large convection cells known as supergranules.

The complete nature of barbs is not known but a number of properties have been identified. There is an anticorrelation of barb ends with network magnetic fields (Engvold 1998). Many chromospheric barb ends are associated with extremely weak magnetic fields. However, recent observations of the formation of some barbs, in concert with the evolution of ephemeral active regions, confirm that some barbs are also associated with strong magnetic fields (Martin 1998). In these cases the barb ends appear to be only linked to the polarity that is opposite to the dominant network magnetic polarity on each side of the filament. This observed linkage implies that the magnetic fields in these barbs has a component opposite in direction to the magnetic component in the overlying arcade that is perpendicular to the filament axis.

The formation of some barbs in conjunction with the small, short-lived, bipolar fields of ephemeral active regions is additional evidence that the filament barbs are secondary to the spine. That is, without the spine, barbs would not form. However, mass is not necessarily visible along the entire magnetic field of the spine in observations of filaments in the light of H alpha. The mass of filaments tends to vary with the magnetic flux density in the adjacent photosphere and with the age of the filament. Among magnetic fields of low flux density, filaments can consist only of a series of tufts of mass along a polarity reversal boundary. These tufts are not necessarily perpendicular to the chromosphere but most have large vertical components and semiregular spacing. They can erupt in unison, which implies that there is a magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

field composing the spine of the filament which links the tufts together even though the density of mass along the spine can be too low to be visible.

In regions of high to intermediate magnetic flux density, the barbs also maintain a specific pattern in relation to the fibril pattern in the filament channel. While the barbs, except for their endpoints, are above the chromosphere, they have an axis with a horizontal component that is nearly aligned with the fibril structure below. This relationship is readily seen when intermediate filaments are observed from above (as in figure 2). This pattern of the barbs is related to the local magnetic field and identifies their chirality or handedness. The pattern is such that the barb magnetic fields on both sides of a filament always have a component in common with the direction of the magnetic field along the spine of the filament and a component perpendicular to the spine.

Maintenance of a filament

With the discovery of bi-directional streaming in the spines and barbs of filaments (Zirker *et al* 1998), the formation process is more clearly recognized as a continuous process of mass transport into and out of the filament magnetic field. The relative rates of mass inflow and outflow into a prominence magnetic field are important to filament visibility. A filament cannot form unless the rate of plasma inflow exceeds the rate of outflow. To be sustained, the inflow must be equal to or greater than the outflow. The time that mass resides or is recycled along the filament magnetic field is still unknown owing to the difficulty of tracking discrete packets of mass observed in filaments. The duration of any mass in filaments could be only minutes to a few hours. This would mean that a continuous recycling of mass from the chromosphere into the filament magnetic structure and back to the chromosphere is always a normal part of their formation and existence.

The extent to which the magnetic field changes in filaments, along with mass input, is not known. However, many factors indicate the filament magnetic fields are continuously changing by small amounts. These are intermittent cancellation of opposite-polarity magnetic fields along the polarity reversal boundary, the continuous formation and dissolution of barbs, the migration of the barb endpoints, the merging of two filaments to create a single longer one and the formation of a new filament at the location where one has erupted. New filaments commonly develop immediately after the eruption of a filament in an active region. On the quiet Sun, new filaments sometimes form within a few days after the eruption of a quiescent filament. However, reconfiguring of the magnetic field environment, in either case, can also eliminate the conditions necessary for filament reformation.

High-resolution velocity data and magnetic field observations are still needed to answer many remaining questions about filament formation. One of the outstanding questions is how mass is continuously guided into and out of filament magnetic fields

Bibliography

Engvold O 1988 Observations of filament structures and dynamics *New Perspectives on Solar Prominences(ASP Conf. Ser. 150)* ed D Webb, D M Rust and B Schmieder, pp 23–31

Gaizauskas V, Zirker J B, Sweetland C and Kovacs A 1997 Formation of a solar filament channel *Atrophys. J.* **479** 448

Martin S F 1990 Observational conditions for the formation of prominences *Dynamics of Quiescent Prominences (Lecture Notes in Physics, vol 363)* ed V Ruzdjak and E Tandberg-Hanssen (Berlin: Springer) pp 1–48

Martin S F 1998 Conditions for the formation and maintenance of filaments *Sol. Phys.* **182** 107–37

Martin S F and McAllister A H 1995 The skew of x-ray coronal loops overlying H alpha filaments *Magnetodynamic Phenomena in the Solar Atmosphere (IAU Colloq. 153)* ed Y Uchida, H S Hudson and T Kosugi, pp 497–8

Martin S F and McAllister A H 1997 Predicting the sign of magnetic helicity in erupting filaments and coronal mass ejections *Coronal Mass Ejections (AGU Geophysics Monogr. Ser. 99)* ed N Crooker, J A Joselyn and J Feynman pp 127–38

Martin S F, Bilimoria R and Tracadas P W 1994 Magnetic field configurations basic to filament channels and filaments *Solar Surface Magnetism* ed R J Rutten and C J Shrijver (Dordrecht: Kluwer) pp 339–46

Rust D M 1994 Spawning and shedding helical magnetic fields in the solar atmosphere *Geophys. Res. Lett* **21** 241

Zirker J B, Engvold O and Martin S F 1998 Counterstreaming flows as evidence for vertical magnetic fields in prominences *Nature* **396** 440

Sara F Martin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominence Models

SOLAR PROMINENCES are ribbons of cool dense gas embedded in the hot tenuous corona, which forms the outer atmosphere of the Sun. Prominence models aim to describe the physical conditions in and near prominences and the physical processes involved in their formation, maintenance and disappearance.

PROMINENCES are located tens of thousands of kilometers above the visible 'surface' of the Sun (the photosphere) and have temperatures $\sim 10^4$ K, a hundred times lower than the temperature of the surrounding corona. The particle densities in prominences range from 10^{16} to 10^{17} m^{-3}, a hundred times greater than coronal values. When viewed above the solar limb, prominences appear as bright features against the dark background. They can also be observed on the solar disk by taking images of the Sun in certain narrow wavelength bands corresponding to strong Fraunhofer lines in the solar spectrum. The higher opacity of the solar atmosphere in a spectral line allows observers to see structures higher up in the atmosphere which are not visible in broad-band 'white' light. One spectral line often used for such studies is the hydrogen Hα line at a wavelength of 6563 Å, which is formed in the solar chromosphere. When observing in Hα on the disk, prominences show up as dark filaments overlying the chromosphere. In the following we refer to these filaments as 'prominences' since they really are one and the same phenomenon.

Prominences are always located at the polarity inversion lines separating regions with opposite magnetic polarity in the photosphere (i.e. lines where the radial component of the photospheric field changes sign). They form in so-called filament channels, regions where the chromospheric fibrils (thread-like fine structures in the chromosphere) are aligned parallel to the inversion line (see SOLAR FILAMENT CHANNELS and SOLAR PROMINENCE FORMATION). The magnetic field in a filament channel is mainly horizontal and directed *along* the polarity inversion line. Filament channels can be classified as *dextral* or *sinistral*, depending on the direction of the axial field as seen by an observer standing on the positive-polarity side. Quiescent prominences in the northern hemisphere are predominantly dextral, while those in the south are predominantly sinistral (see SOLAR PROMINENCE CHIRALITY).

Prominences are large, long-lived structures which can persist for many days, but they usually contain thread-like fine structures which last only a few minutes (see SOLAR PROMINENCE FINE STRUCTURE). Some of these threads are clearly aligned with the local magnetic field, but this is not so obvious for the vertical threads frequently seen in quiescent prominences above the limb. It is unclear, for example, whether the magnetic field in these vertical threads is vertical or horizontal.

Magnetic fields play a key role in many aspects of prominence physics, including their support against gravity and thermal insulation from the surrounding corona. It the following we discuss some of these aspects in more detail.

Prominence support

The outer atmosphere of the Sun is permeated by magnetic field, and this field is also present within prominences, where it provides the magnetic force necessary to support the dense prominence gas against gravity. This magnetic force is made possible by the fact that prominences (and the surrounding corona) consist of partially ionized gas, or *plasma*, in which there is a high concentration of unbound electrons. The electrical conductivity of this plasma is very high, and electric currents can easily flow through the plasma. In such a highly conducting medium the plasma is forced to move *along* magnetic field lines and cannot easily cross from one field line to another. Therefore, a parcel of dense prominence plasma will stay on the field line on which it was originally located. Under the influence of gravity, the parcel will slide down along the field line until it reaches the chromosphere or encounters a 'dip' in the field line where the field is locally horizontal and curved upward. Various models of magnetic configurations with dipped field lines have been proposed (see figure 1). Prominences are thought to be cool plasmas which have come to rest in such dipped (or nearly horizontal) field lines.

Kippenhahn and Schlüter (1957) were the first to develop a model of the equilibrium and stability of prominence plasma in a magnetic configuration with dipped field lines. They assumed a so-called *normal polarity* configuration in which the field lines pass through the prominence from the region of positive polarity (radially outward magnetic field in the neighboring photosphere) to the region of negative polarity (figure 1(*a*)). Kuperus and Raadu (1974) later proposed a different model in which the prominence has *inverse polarity* compared with the neighboring fields (figure 1(*b*)). These figures show vertical cross-sections of a prominence in a plane perpendicular to the long axis of the prominence. It should be kept in mind that in general there is also a component of magnetic field *along* the prominence (into or out of the plane), which often is stronger than the component within the plane. However, as we will see below, this axial component of the field is not essential for prominence support.

The magnetic force in a plasma is given by the Lorentz force, $\boldsymbol{j} \times \boldsymbol{B}$, where $\boldsymbol{B}(\boldsymbol{r})$ is the *magnetic induction* and $\boldsymbol{j}(\boldsymbol{r})$ is the *electric current density*, i.e. the electric current per unit area perpendicular to the current. Hence, to support the weight of the prominence in the dips of field lines, there must exist an electric current which flows through the prominence in a horizontal direction which crosses the magnetic field lines. An important difference between the Kippenhahn–Schlüter and Kuperus–Raadu models is the direction of this electric current: in figure 1(*a*) the current flows out of the plane of the figure (toward the observer), whereas in figure 1(*b*) the current flows into the figure (away from the observer). Moreover, in the Kuperus–Raadu model the current is present not only inside the prominence, but also in the 'magnetic island' just above the prominence, as indicated by the cross in figure 1(*b*).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

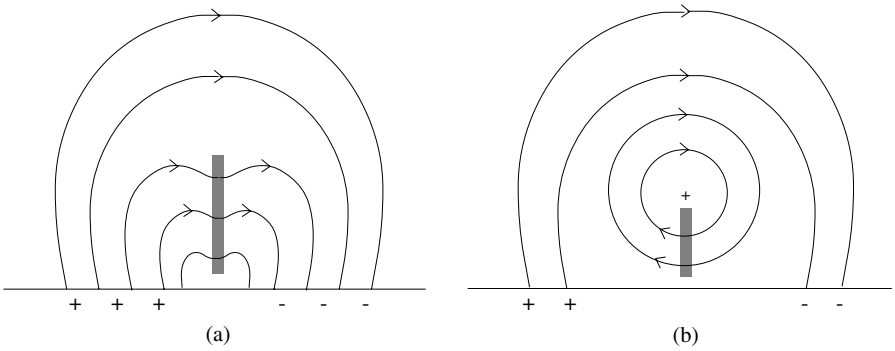


Figure 1. Two models of the magnetic field supporting a solar prominence: (*a*) Kippenhahn–Schlüter model; (*b*) Kuperus–Raadu model. These figures show the projection of the field lines onto the plane perpendicular to the long axis of the prominence (shaded region). The horizontal line at the base indicates the solar photosphere.

The plasma within the prominence is subject to three forces: (1) the Lorentz force, $\boldsymbol{j} \times \boldsymbol{B}$; (2) the gravitational force, $-\rho g \hat{\boldsymbol{z}}$, where $\rho(\boldsymbol{r})$ is the mass density, g is the acceleration of gravity and $\hat{\boldsymbol{z}}$ is the radially outward direction on the Sun; (3) the force due to the gradient of gas pressure, $p(\boldsymbol{r})$. In equilibrium these forces must balance each other:

$$-\boldsymbol{\nabla} p - \rho g \hat{\boldsymbol{z}} + \boldsymbol{j} \times \mathbf{B} = \mathbf{0}. \tag{1}$$

Using Ampère's law, the electric current density can be written as $\boldsymbol{j} = \mu^{-1} \nabla \times \boldsymbol{B}$, where μ is the magnetic permeability. The mass density can be written as

$$\rho = \frac{mp}{kT} \tag{2}$$

where $T(\boldsymbol{r})$ is the temperature, k is the Boltzmann constant and m is the mean mass per particle (m lies between 0.6 and 1.3 times the proton mass, depending on composition and ionization state of the plasma). Kippenhahn and Schlüter (1957) modeled the prominence as a thin vertical sheet in which the pressure $p(x)$, density $\rho(x)$ and vertical magnetic field $B_z(x)$ depend only on the horizontal coordinate x perpendicular to the sheet. The temperature T and horizontal field components B_x and B_y are assumed to be constant. Then the x- and z-components of equation (1) reduce to

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p + \frac{B_z^2}{2\mu}\right) = 0 \tag{3}$$

$$-\rho g + \frac{B_x}{\mu}\frac{\mathrm{d}B_z}{\mathrm{d}x} = 0. \tag{4}$$

The boundary conditions far away from the sheet ($x \to \pm\infty$) are $p \to 0$ and $B_z \to \pm B_{z\infty}$. Integrating equation (3) yields

$$p = \frac{B_{z\infty}^2 - B_z^2}{2\mu} \tag{5}$$

and, using equations (2) and (5), equation (4) yields a differential equation for B_z which has the following solution:

$$B_z(x) = B_{z\infty} \tanh \frac{B_{z\infty} x}{2 B_x H_p}$$

$$p(x) = \frac{B_{z\infty}^2}{2\mu}\left(\cosh \frac{B_{z\infty} x}{2 B_x H_p}\right)^{-2}.$$

Here $H_p \equiv kT/mg$ is the so-called *pressure scale height* which describes how rapidly the pressure and density fall off with height along the field lines ($H_p \sim 200$ km within the prominence). Note that the plasma pressure at the center of the prominence sheet is equal to the external magnetic pressure associated with the *vertical* component of the magnetic field and that the width of the sheet is of order $4(B_x/B_{z\infty})H_p$. The observed widths of quiescent prominences (~8000 km) can be reproduced with $B_{z\infty} \sim 0.1 B_x$, in other words, the support of the prominence requires only a minor perturbation of the surrounding magnetic field.

In general there is also a component of magnetic field along the prominence (B_y), so the normal polarity field shown in figure 1(*a*) is actually a sheared arcade (with dips at the loop tops), and the circular field lines in figure 1(*b*) are actually helical windings which are wound around a horizontal axis that runs parallel to and above the prominence. Therefore, to obtain a more accurate picture of the prominence magnetic field we must consider its full three-dimensional structure.

Prominence magnetic structure

There exist a variety of methods for measuring the magnetic fields in and around prominences, most of which are based on the *Zeeman effect* (the splitting of atomic energy levels in the presence of a magnetic field). The Zeeman effect causes the light emitted in a spectral line to become circularly polarized. By measuring the degree of circular polarization it is possible to deduce the component of magnetic field along the line of sight in a prominence. Another method is based on the fact that much of the light emitted by a prominence above the solar limb is actually scattered light which originates from the chromosphere

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

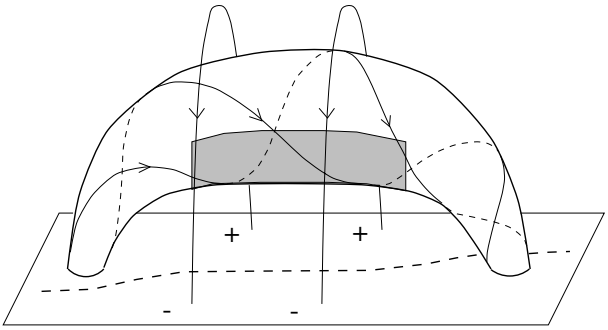


Figure 2. Twisted flux tube model for solar prominences. The prominence sheet is indicated by the shaded region.

below. The magnetic field affects the scattering properties of the atoms, and manifests itself as a change in the state of linear polarization of the scattered light in certain spectral lines. This so-called *Hanle effect* allows observers to deduce both the strength and direction of magnetic fields in prominences. Observational studies have shown that the magnetic fields in quiescent prominences are in the range (3–30) $\times 10^{-4}$ T and that the field is mainly directed along the length of the prominence: the magnetic vector is inclined to the prominence axis at an average angle of about 25° (Leroy 1989). Most quiescent prominences have *inverse* polarity, i.e. the magnetic field traverses the prominence from the region of negative polarity to the region of positive polarity, opposite to what would be expected for a coronal arcade. The field strength increases with height in the prominence.

These observations are consistent with the idea that a quiescent prominence is located within a large, twisted flux tube or flux rope (see figure 2). A number of authors have developed prominence models based on this idea (e.g. Priest *et al* 1989). According to these models, an arched flux tube is anchored in the photosphere at two ends, and the field lines in the coronal portion of the tube make one or two revolutions about the tube axis, forming a helical field. Overlying this helical field is a coronal arcade, which probably plays an important role in the equilibrium and stability of the field. Cool prominence plasma is supported at the troughs of the helical windings where the field lines are curved concave-upward. This twisted flux tube model has many features which agree with observations, including the inverse polarity of the magnetic field in the prominence and the fact that when a prominence erupts it sometimes looks like a twisted tube (see SOLAR PROMINENCE ERUPTION).

How does the Sun produce such helical fields? One possibility is that the twist is produced by vortical motions of the photospheric footpoints of the tube, but this would require persistent twisting over several revolutions, which is not observed. Another possibility is that the twisted flux rope is created in the convection zone below the photosphere and emerges through the photosphere with its twisted structure already formed (e.g. Rust and Kumar 1995). A third possibility is that the helical field is produced by magnetic reconnection, i.e. the reconfiguration of magnetic field lines due to plasma resistivity effects (see MAGNETIC RECONNECTION). For example, Pneuman (1983) proposed that the radial outward distension of a bipolar region by gas pressure gradients could lead to an inward collapse of the region, causing the field lines to reconnect. If the initial field is significantly sheared along the polarity inversion line, reconnection produces helical field lines in the region above the reconnection site.

Observations indicate that prominences tend to form in regions where opposite-polarity flux is being cancelled at the polarity inversion line. In a sheared coronal arcade magnetic flux cancellation can proceed only if reconnection occurs in the region just above the inversion line (van Ballegooijen and Martens 1989). The process is illustrated in figure 3, which shows the evolution of a sheared arcade in response to converging motions of the photospheric footpoints. The initial field is assumed to be sheared (figure 3(*a*)), which may have been caused by flows along the inversion line or by some other effect. When the footpoints of the coronal loops are pushed to the inversion line, the loops become more and more aligned with the inversion line, eventually causing different loop systems to reconnect (figure 3(*b*)). The model assumes that the short, highly curved loops produced by such reconnection are pulled below the photosphere by magnetic curvature forces, causing magnetic flux to disappear from the photosphere (flux cancellation). However, the reconnection also produces longer loops which remain in the corona because their curvature radii are too large to overcome the buoyancy forces at and below the photosphere. As overlying loops are pushed to the inversion line, further reconnection produces helical loops in which a prominence can form (figures 3(*c*) and 3(*d*)). As more and more flux is forced to reconnect at the inversion line, the width of the helical flux tube gradually increases and the axis of the tube slowly rises. Eventually the magnetic structure becomes unstable, causing the helical field to erupt.

Origin of prominence plasma

How do prominences acquire their mass? One viewpoint is that prominences are formed by the cooling and 'condensation' of plasmas from the surrounding corona (see Priest 1982, chapter 11). To understand how this happens, we must consider the energetics of the coronal plasma. A hot gas such as the corona is subject to radiative cooling, i.e. the loss of thermal energy due to collisional excitation of ions by electrons and the subsequent emission of radiation at extreme ultraviolet and x-ray wavelengths (the corona is optically thin, so this radiation can freely escape into space). The rate of energy loss per unit volume is approximately given by $n_e^2 \Lambda(T)$, where n_e is the electron density, T is the temperature and $\Lambda(T)$ is the so-called radiative loss function, which depends on atomic parameters and the chemical composition of the plasma. In a coronal loop

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

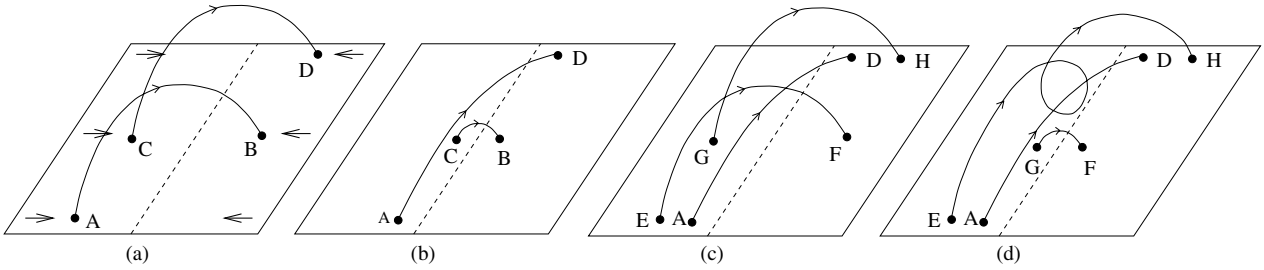


Figure 3. Flux cancellation in a sheared coronal arcade. The rectangle represents the solar photosphere, and the broken line is the polarity inversion line. (*a*) Initial sheared field subject to converging flows. (*b*) Reconnection produces a long loop AD and a shorter loop CB, which subsequently disappears below the photosphere. (*c*) Overlying loops EF and GH are pushed to the inversion line. (*d*) Reconnection produces the helical loop EH and a short loop GF, which again submerges.

these radiative losses are balanced by heating due to dissipation of magnetohydrodynamic (MHD) waves or other disturbances. Electron thermal conduction plays an important role in redistributing this heat along the coronal loop. However, as coronal conditions change, the density and/or loop length may slowly increase with time, causing radiative losses to become more important relative to conduction. Eventually, the coronal loop may reach a point where a stable equilibrium between energy gains and losses is no longer possible. The plasma then rapidly cools and settles into a new equilibrium with much lower temperature but higher density. This condensation process occurs for a wide range of forms of the heating function.

A difficulty with the above scenario is that the mass of a quiescent prominence ($\sim 10^{12}$ kg) typically exceeds the mass available in the surrounding corona before the prominence is formed. Therefore, additional mass must somehow be supplied to the prominence. One possibility is that the mass is injected along the field lines that connect the prominence with the chromosphere below. This could be in the form of a *siphon flow* driven by a possible pressure difference between the chromosphere and the prominence. Another possibility is that cool plasma is lifted up by magnetic fields as they rise through the chromosphere. For example, in the flux rope model of Rust and Kumar (1995), the twisted flux rope emerges from the convection zone and sheds most of its mass on it way up into the corona (the density in the corona is much less than that in the convection zone). However, a small fraction of the initial mass remains trapped in the troughs of the helical windings. This remnant is believed to form the prominence.

Observations suggest that prominences are very dynamical structures, with plasma continually draining downward and new material being injected into the prominence. Priest *et al* (1996) propose a model in which the prominence is maintained by a continual input of mass and magnetic flux from below. In this model the correct dextral and sinistral patterns for high-latitude, east–west prominences are produced by the combined effects of differential rotation acting on subphotospheric flux, its subsequent emergence by magnetic buoyancy and its rearrangement by reconnection to form a filament channel with magnetic flux oriented along its axis. Continual emergence and reconnection creates a prominence as a flux tube along the filament channel and filled with cool plasma which is lifted up from the photosphere and chromosphere by the reconnection process. According to this model, reconnection occurring in the chromosphere yields prominence densities in rough agreement with observations.

Thermodynamic modeling

The temperature within a prominence is determined by the balance between heating and cooling of the prominence plasma. There are several possible contributions to the heating: (1) energy may be transported into the prominence by thermal conduction from the hot corona, but this is not very effective in the low-temperature prominence material; (2) the plasma can be heated by dissipation of MHD waves or other disturbances which propagate into the prominence from the sides; (3) the prominence may be heated by absorption of ultraviolet radiation from the chromosphere. The energy losses occur mostly in the form of radiation in hydrogen spectral lines and continua (Lyman and Balmer series). The radiation transport is complicated by the fact that prominences are optically thick at these wavelengths, which means that the ultraviolet radiation is reabsorbed and reemitted many times before it finally escapes. Another complication is that prominences have a filamentary structure, with thin (sometimes vertical) threads of dense plasma embedded in a much more tenuous medium. This allows the ultraviolet radiation to penetrate deep into the prominence, greatly enhancing the excitation rate compared with models without such fine-scale structures (Heasley and Mihalas 1976). The cause of these fine structures is not yet understood.

Bibliography

Heasley J N and Mihalas D 1976 Structure and spectrum of quiescent prominences: energy balance and hydrogen spectrum *Astrophys. J.* **205** 273–85

Kippenhahn R and Schlüter A 1957 Eine Theorie der solaren Filamente *Z. Astrophys.* **43** 36–62

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Kuperus M and Raadu M A 1974 The support of prominences formed in neutral sheets *Astron. Astrophys.* **31** 189–93

Leroy J L 1989 Observation of prominence magnetic fields *Dynamics and Structure of Quiescent Solar Prominences* ed E R Priest (Dordrecht: Kluwer) pp 77–113

Pneuman G W 1983 The formation of solar prominences by magnetic reconnection and condensation *Sol. Phys.* **88** 219–39

Priest E R 1982 *Solar Magnetohydrodynamics* (Dordrecht: Reidel)

Priest E R, Hood A W and Anzer U 1989 A twisted flux tube model for solar prominences: I general properties *Astrophys. J.* **344** 1010–25

Priest E R, van Ballegooijen A A and Mackay D H 1996 A model for dextral and sinistral prominences *Astrophys. J.* **460** 530–43

Rust D M and Kumar A 1995 Helical magnetic fields in filaments *Sol. Phys.* **155** 69–97

van Ballegooijen A A and Martens P C H 1989 Formation and eruption of solar prominences *Astrophys. J.* **343** 971–84

A A van Ballegooijen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominence Oscillations

SOLAR PROMINENCES are continuously subject to oscillations in response to ambient atmospheric perturbations. These vibrations can be of global nature, for example when a large-scale disturbance, carrying considerable energy, impinges on the broad side of the prominence body and shakes its whole structure, or of local nature, when the prominence material is perturbed by motions in the SOLAR PHOTOSPHERE, CHROMOSPHERE or CORONA. In any case, the response of the PLASMA is directly related to its physical parameters, so the investigation of this type of phenomenon can provide information on the physics of prominences.

The research of prominence oscillations, also known as prominence seismology, seeks to achieve a success similar to that of HELIOSEISMOLOGY from the comparison between observations and the theoretical analysis of prominence vibrations. This task is not an easy one because of the complicated three-dimensional structuring of these objects, whose theoretical modelling is extremely complex (see SOLAR PROMINENCE MODELS). With the investigation of prominence oscillations we expect to improve our knowledge on the nature of these exotic objects.

Flare-initiated filament oscillations

Observations of solar FILAMENTS through a narrow Hα FILTER made during the first half of the 20th century revealed that prominences can be activated by FLARES as far away as 500 000 km, a considerable distance compared with the 4.4×10^6 km length of the solar circumference. This activation often leads to the disappearance of the filament, or the appearance of a previously undetected filament, in the Hα image. The associated optical effect gave rise to the term 'winking filament' for this kind of filament behavior.

The first reports on prominence oscillations came from the detection of 'winking filaments' that disappear and reappear at regular intervals after being disturbed by a wave emanating from a distant flare. In 1966 H Ramsey and S Smith made an in-depth study of this type of filament activation by using three narrow filters, one centered in the Hα line and the other two in the red and blue wings of the line, ± 0.5 Å from the centre. An unperturbed prominence is only visible in Hα, but invisible at the other two wavelengths; however, Ramsey and Smith discovered that, on the passage of the flare-induced wave, the absorption wavelength of the prominence material oscillates about the Hα line center, with an amplitude of at least 0.5 Å, for a few periods before the oscillations are damped. They also found that this process is more or less periodic, with periods between 6 and 40 min for different filaments. In addition, they were able to study the perturbations induced by four flares on a particular filament in three consecutive days and found that its frequency of oscillation remained almost unchanged.

A few years later, J Kleczek and M Kuperus gave a theoretical explanation of these flare-induced oscillations. The complicated prominence magnetic field was represented by an 'effective' uniform magnetic field in the direction of the prominence main axis and with its endpoints tied at a fixed position (simulating the photospheric line-tying effect). The magnetic tension in this kind of configuration provides the force required to counteract the pull of gravity and also a restoring force that tends to bring the prominence to its equilibrium state after suffering a disturbance. Then, the passage of a large-amplitude wave induces horizontal prominence motions which result in a periodic variation of the line-of-sight velocity component (of course, this is not true when the filament is near the center of the solar disk and the line of sight is almost perpendicular to the oscillatory motions, but this was not the case with Ramsey and Smith's observations). The consequence of these periodic mass motions is a periodic Doppler shift of the filament absorption wavelength about the Hα line centre, alternatively towards the red and blue wings, in agreement with observations. The Doppler displacements detected by Ramsey and Smith, of at least 0.5 Å, point to an oscillatory amplitude in excess of 20 km s^{-1}, a rather large value.

Kleczek and Kuperus also gave an expression for the oscillatory period in terms of the prominence's mean magnetic field and density. The fact that in this model different filaments have different effective magnetic fields and densities implies that a given prominence will oscillate at its own characteristic frequency as long as its overall properties do not change in time (as is the case with Quiescent prominences over a few days). Using values of density and magnetic field typically found in these objects, these authors obtained periods and damping times around 20 min, also in reasonable agreement with the observations.

Small-amplitude oscillations: observations

In contrast with the previous large-amplitude, global oscillations triggered by a flare, quiescent prominences are also subject to small-amplitude, small-scale periodic vibrations not related to flare activity. In 1969 J Harvey examined in his PhD thesis the Doppler velocities of a sample of 68 non-ACTIVE REGION prominence oscillations. He concluded that 31% of the objects presented no significant velocity change along the line of sight, 28% showed apparently random line-of-sight velocity variations and 41% displayed an oscillatory behavior. The measured periods were distributed between 1 and 13 min, with a tendency of periods between 3 and 7 min to occur more frequently. In addition, the line-of-sight velocity amplitude in most cases ranged from 0.1 km s^{-1} (the signal noise level) to 2 km s^{-1}.

Following this preliminary investigation the subject received little attention in the next decade, after which systematic observations were undertaken. The analyses made at the beginning of the 1980s served to confirm Harvey's results, with regard to periods and amplitudes, but they also indicated the presence of an apparently new kind of small-amplitude prominence vibrations. Velocity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

oscillations with periods between 40 and 80 min and amplitudes around 1–2 km s^{-1} were abundantly reported. The large values of these periods led to the classification of prominence vibrations into short- and long-period oscillations and to the belief that they are of different nature. Nevertheless, according to theoretical models (see below), long and short periods are the consequence of the same physical processes and intermediate values do also exist, in agreement with later observations.

The instrumental setup normally used to study prominence oscillations involves placing a SPECTROGRAPH slit on a quiescent prominence or filament so that the emission or absorption spectrum of the plasma along the slit can be investigated. A selected spectral line, such as Hα or Ca II K, can then be used to deduce the temporal dependence of the line-of-sight velocity for the points of the prominence along the slit (see figure 1). Some instruments, such as the MSDP spectrograph, allow observers to gain spectral information of a two-dimensional field of view, thus making feasible a better understanding of the spatial distribution of oscillations. Subsequent analysis of the data, e.g. using the FFT, yields information on the presence of periodic vibrations at each of the prominence locations. Sometimes more than one period will be present in a given Doppler signal at a given spatial point. On the other hand, just as Harvey found, it is common to obtain velocity signals which do not change periodically in time. The reason for this is either that the velocity amplitude or its projection along the line of sight is too small to be detected or that the prominence material does not actually oscillate, which must be understood as an indication that prominence oscillations are episodic phenomena that are continuously excited by external agents and last for a moderate time before being damped by some dissipative processes.

Regarding the sizes of regions where oscillations with a given period are detected, they range from the telescope resolution limit (i.e. a few hundred km) to as much as 20 000 km, a considerable length compared with typical quiescent prominence dimensions. Therefore, periodic vibrations seem to be excited both at local and global scales. Some authors have even been able to associate the oscillating portions of prominences with the fibril structure of these objects (see SOLAR PROMINENCE FINE STRUCTURE). This is not surprising since, after motions are excited in a fibril, its enhanced magnetic field can act as a waveguide (see MAGNETOHYDRODYNAMIC WAVES) and the perturbations will spread over the whole fibril.

Approximate or limiting values for other parameters have also been determined from the study of the vibrating areas. Wavelengths are typically found to lie between 10 000 and 100 000 km, while phase speeds lie between 25 and 100 km s^{-1}, although values outside this range have also been reported.

Researchers face a number of problems when trying to detect periodic changes in quiescent prominences. First of all, there are the inevitable consequences of the Earth's atmosphere, mostly by means of SEEING conditions,

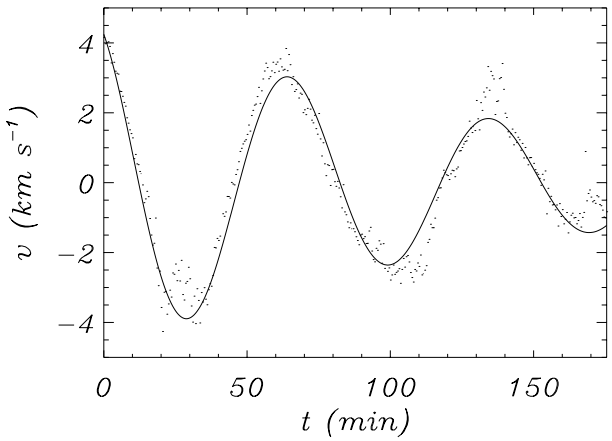


Figure 1. Time variation of the Doppler velocity observed in a quiescent solar prominence. The data can be fitted by a function of the form $v_0 \cos(2\pi t/T + \phi)\ \exp(-t/\tau)$, with amplitude $v_0 = 4.8$ km s^{-1}, period $T = 70$ min and damping time $\tau = 140$ min (solid curve). The well-defined periodic variation readily recognizable in this figure provides good support for the existence of prominence oscillations, although in most data sets they are usually harder to detect visually because of their smaller amplitude and/or higher noise level. Courtesy R Molowny-Horas.

imperfect telescope guiding and limited spatial resolution, which generally result in a slight degradation of the data. A more fundamental source of data corruption, however, may arise from the fact that light emitted or absorbed by various plasma elements along the line of sight is collected together by the detector, so the observed spectral line contains contributions from all the elements and incorrect Doppler velocities follow. This is probably the reason why oscillations seem easier to identify where the prominence looks fainter.

Small-amplitude oscillations: theory

Theoretical investigations of prominence oscillations are based on the assumption that an external perturbation excites eigenmodes of the prominence structure. At present very little is known about the mechanism responsible for generating small-amplitude waves in prominences, although it has often been claimed that some short-period oscillations can be the response of the prominence to a forcing produced by the chromospheric 3 min oscillations and photospheric 5 min oscillations (see SOLAR CHROMOSPHERIC OSCILLATIONS), whose influence could propagate along magnetic field lines and induce motions of the prominence plasma. Nevertheless, this explanation leaves out most of the detected periods, whose origin could be in small-scale, impulsive disturbances coming from the surrounding corona that would shake the prominence and would excite its normal modes of vibration.

To illustrate the kind of theoretical progress accomplished so far let us consider a basic prominence equilibrium made of a cold and dense plasma slab embedded in a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

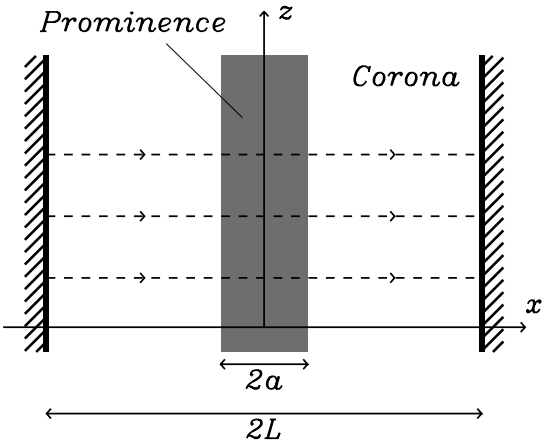


Figure 2. Sketch of an elementary equilibrium used in modelling prominence oscillations. The drawing represents a cut across the prominence, devised as an infinitely high and long plasma slab of width $2a$ embedded in the solar corona. Dashed lines represent the horizontal magnetic field threading the system and anchored at $x = \pm L$.

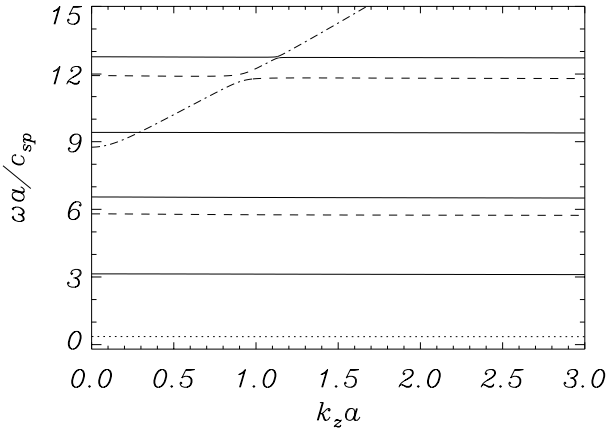


Figure 3. Frequency versus vertical wavenumber for slow and fast magnetoacoustic modes with v_x even about $x = 0$. The quantity $c_{\rm sp}$, coming into the dimensionless frequency, is the adiabatic sound speed in the isothermal prominence. Solid lines, internal slow modes; dashed lines, external slow modes; dotted line, hybrid slow mode; dot-dashed line, internal fast mode. A similar dispersion diagram is obtained for slow and fast odd modes, the central difference being that the hybrid slow mode is replaced by the hybrid fast mode. For a description of the properties of MHD modes see MAGNETOHYDRODYNAMIC WAVES.

hotter and rarer corona (figure 2). The system is threaded by a uniform, horizontal magnetic field and the photospheric line-tying is simulated by placing two vertical boundaries symmetrically with respect to the prominence center ($x = 0$). A Cartesian coordinate system is used in which the prominence long axis lies in the y-direction, the vertical lies in the z-direction and the x-axis is transverse to the prominence body. Gravity is neglected in this simple equilibrium and, to ensure total pressure (magnetic pressure plus gas pressure) balance at the prominence-corona interface, the product of density and temperature must be equal in the two uniform regions. This equilibrium model is very idealistic since the physical parameters are assumed uniform, although it is well known that quiescent prominences are highly structured. Nevertheless, this model will be helpful to introduce the basic concepts in theoretical studies of prominence oscillations.

To fully understand the dynamics of this structure one must first investigate its normal modes, which in our particular case have a velocity of the form $\boldsymbol{v}(x)\exp(-\mathrm{i}k_y y)\exp(-\mathrm{i}k_z z)\exp(\mathrm{i}\omega t)$. This expression represents a standing wave in the x direction, with frequency ω and wavenumbers k_y and k_z in the prominence long direction and in the vertical direction, respectively. After making the simplifying assumption $k_y = 0$, one is left with the task of solving a pair of ordinary differential equations for v_x and v_z decoupled from a third equation for v_y only. The large photospheric density prevents coronal motions from displacing magnetic field lines or perturbing the photosphere, so the velocity vector must vanish at $x = \pm L$. With these boundary conditions, the velocity $\boldsymbol{v}(x)$ and frequency of normal modes can be calculated. Owing to the symmetry of the equilibrium configuration and the boundary conditions, solutions to the first two equations are called even or odd depending on whether v_x is even or odd about $x = 0$. The dispersion diagrams of even and odd modes are thus drawn separately (figure 3). The two types of curves in this diagram (straight and parabolic lines) betray the different physical nature of slow magnetohydrodynamic modes and fast magnetohydrodynamic modes (magnetohydrodynamic waves). The first one is essentially non-dispersive (i.e. ω independent of k_z) and gives rise to motions practically parallel to magnetic field lines (that is, $v_x \gg v_z$) which are driven by the pressure gradient force. On the other hand, fast modes present a wavelength-dependent frequency, have basically vertical motions ($v_z \gg v_x$) and are driven by the magnetic Lorentz force.

A more detailed inspection of figure 3 and of the velocity perturbations associated to each of the modes reveals they can belong to an internal, external or hybrid class. The main features of internal (external) modes are determined from the characteristics of the prominence (corona) but are little influenced by the existence of the other region. According to this classification scheme, hybrid modes are internal and external at the same time. Thus, for example, modifying the width $2L$ of the system results in a variation of the frequency of external and hybrid modes, while internal modes remain unscathed. The eigenfunctions (v_x for slow modes and v_z for fast modes) are also useful in determining the character of modes: hybrid modes are the ones with the simplest possible structure in the x-direction, showing two nodes at $x = \pm L$ and a single maximum (at $x = 0$), external modes achieve a much larger amplitude inside the corona than in the prominence and, finally, the velocity amplitude

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of internal modes is as large in the prominence as in the corona.

The third possible orientation of gas motions, in the y-direction, corresponds to solutions driven by the magnetic tension in which the density remains unperturbed. These are basic features of the Alfvén mode, for which the subdivision into internal, external and hybrid is also pertinent in the present case.

At this point the obvious question is how well the periods of the three magnetohydrodynamic modes compare with the observed ones. Using reasonable values of the equilibrium parameters one recovers long periods (accounted for by the hybrid slow mode), intermediate periods (hybrid fast and Alfvén modes) and short periods (slow, fast and Alfvén internal modes). External modes are of no interest because of their presumed minute velocity inside the prominence.

Apart from reproducing the values of periods, this naive theoretical work has important implications for the detectability of periodic Doppler variations caused by the three magnetohydrodynamic modes. For example, the chances of observing the vertical motions generated by the fast mode are largest for a filament at the solar disk center, whereas they are practically null for a limb prominence. The velocity polarization associated with each mode is then an essential result of theoretical investigations. This study was later extended to include other basic ingredients of quiescent prominences, such as the curvature of magnetic field lines, the magnetic field shear with respect to the prominence long axis (see SOLAR PROMINENCE CHIRALITY), the prominence–corona TRANSITION REGION and the internal fibril structure. The addition of these complexities is fundamental for progress in this area as they substantially modify the picture given here with respect to the frequency and velocity polarization of the magnetohydrodynamic modes.

Bibliography

A theoretical description of flare-initiated prominence oscillations can be found in

Kleczek J and Kuperus M 1969 Oscillatory phenomena in quiescent prominences *Sol. Phys.* **6** 72–9

Examples of the techniques used in obtaining and analysing observational data can be found in

Suematsu Y, Yoshinaga R, Terao N and Tsubaki T 1990 Oscillatory and transient features detected simultaneously in the CaII K and Hβ line spectra of a quiescent prominence *Publ. Astron. Soc. Japan* **42** 187–203

and in

Molowny-Horas R, Oliver R, Ballester J L and Baudin F 1997 Observations of Doppler oscillations in a solar prominence *Sol. Phys.* **172** 181–8

A detailed study of a more complex prominence equilibrium can be found in

Oliver R and Ballester J L 1996 The influence of the temperature profile on the magnetohydrodynamic modes of a prominence–corona system *Astrophys. J.* **456** 393–8

Theoretical work on the influence of prominence fine structure on the normal modes was developed by

Joarder P S, Nakariakov V M and Roberts B 1997 Oscillations in prominence fine structures *Sol. Phys.* **173** 81–101

Ramón Oliver

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Prominences

Prominences are clouds of relatively cool and dense gas in the solar atmosphere. These clouds change shape and come and go, but they do not drift like terrestrial clouds. Prominences are suspended above magnetic channels (see SOLAR FILAMENT CHANNELS) that change little from day to day. They are trapped in magnetic fields in the solar atmosphere, but they churn slowly, at heights of up to 100 000 km above the photosphere (see SOLAR PHOTOSPHERE). They peek brilliantly over the Moon's limb during eclipses. The magnetic fields thread them through and balance the force of gravity, and some prominences live, thus suspended, for months. The magnetic fields also inhibit heat flow from the surrounding corona (see CORONA), where the temperature is 200 times higher than their 6000 K. Thus, prominences emit the visible radiation of hydrogen and helium.

Prominences take on many forms and their frequency of appearance generally parallels the sunspot cycle. Also, while they have long been distinguished for their comeliness (e.g. figure 1) and for the challenges they pose to theorists, they have recently been the focus of heightened interest because they appear to play a fundamental role in the 11 yr solar cycle of magnetic field generation and dispersal (see SOLAR CYCLE).

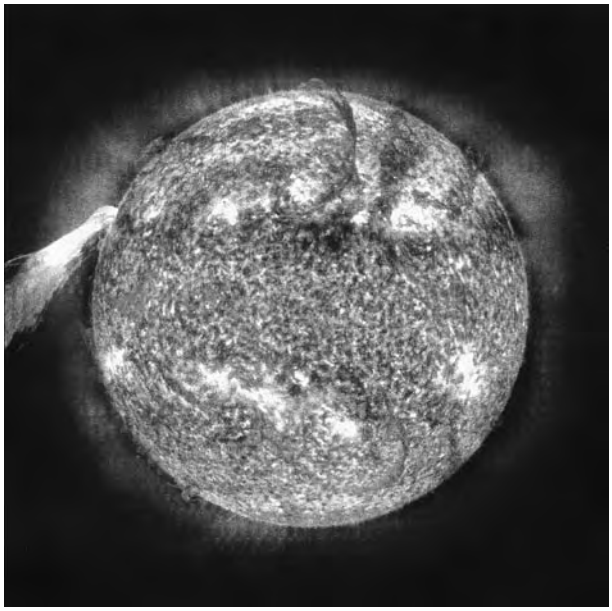


Figure 2. An eruptive prominence (left) and a quiescent filament (upper center) photographed on 15 June 1998 with a SOHO telescope.

History of prominence research

Prominences, or *protuberances*, were recorded during eclipses as early as 1239, but their true character was not understood until the solar eclipse of 1868. It was then that spectroscopy revealed the distinctive emissions of hydrogen and calcium, which had earlier been detected in the solar spectrum. At the same eclipse, the great French astronomer, Pierre-Jules-César Janssen, first obtained evidence from prominence spectra for an unknown element. The element was later named helium for its solar origins, after it was identified in terrestrial minerals in 1895.

The hydrogen spectral lines in prominences are so bright that they can be seen without an eclipse and, at the turn of the century, G Hale in the United States and H Deslandres in France independently invented the spectroheliograph to study them. In the spectroheliograph, the entrance slit of a spectrograph moves across a solar image formed with a telescope. In unison, an exit slit passing only a single spectral line moves over a photographic plate which records a monochromatic image of the Sun. Spectroheliograms of hydrogen and calcium emission revealed that prominences are almost always present at the sun's limb (edge). They also revealed dark threads, *filaments*, on the disk. L d'Azambuja at the Meudon Observatory outside Paris and Hale in California showed that filaments are just prominences seen with the chromosphere (see CHROMOSPHERE) instead of the dark sky as background.

In the first half of the 19th century, E Petit, D Menzel and others classified prominences according to their shapes and motions. The names convey the variety of forms seen—tornado, spray, surge, loop, coronal rain, hedgerow—but little physical insight was achieved until the spectroscopic analyses of the 1950's and 1960's. H Zirin and E Tandberg-Hanssen and others established that prominences can have a range of temperatures and densities with some parts as cool as 6000 K and other parts above 50 000 K.

Two advances in instrumentation by B Lyot greatly aided the study of prominences: the CORONAGRAPH in which a disk inside the telescope serves the function of the moon and allows viewing of prominences and the surrounding corona as in an eclipse, and the Lyot filter, which replaced the spectroheliograph and gives monochromatic images of the whole sun with exposure times as short as 0.2 s. These instruments allowed films of prominences and filaments to be made. The most interesting motions are the sudden eruptions or *disparitions brusques* by which prominences leap from the Sun. Many eruptive prominences have been recorded since the International Geophysical Year (1958–1959) by a worldwide network of telescopes with Lyot filters. Eruptive prominences (figure 2) leave the Sun at 300–1000 km s^{-1} and are associated with SOLAR FLARES.

Shortly after H Babcock developed the solar magnetograph in 1953, H Zirin and A Severny made the first prominence magnetic field measurements. D Rust established that the magnetic field strength in quiescent prominences averages about 10 G and that there is a global pattern to their fields (see SOLAR PROMINENCE CHIRALITY). J Harvey found fields up to 100 G in active prominences (see SOLAR PROMINENCES: ACTIVE), and J–L Leroy and his coworkers at Pic-du-Midi in France established that the fields are predomi-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

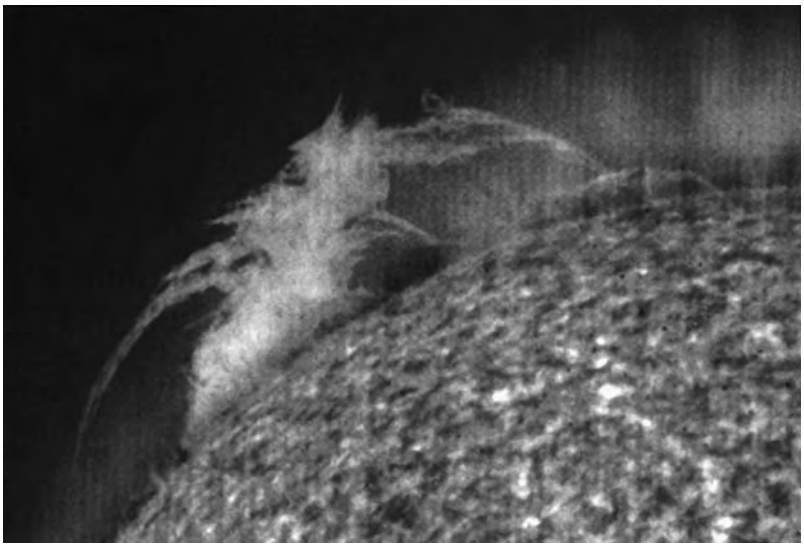


Figure 1. Hedgerow prominence photographed in light emitted by helium ions by the Extreme Ultraviolet Imaging Telescope aboard the SOHO spacecraft.

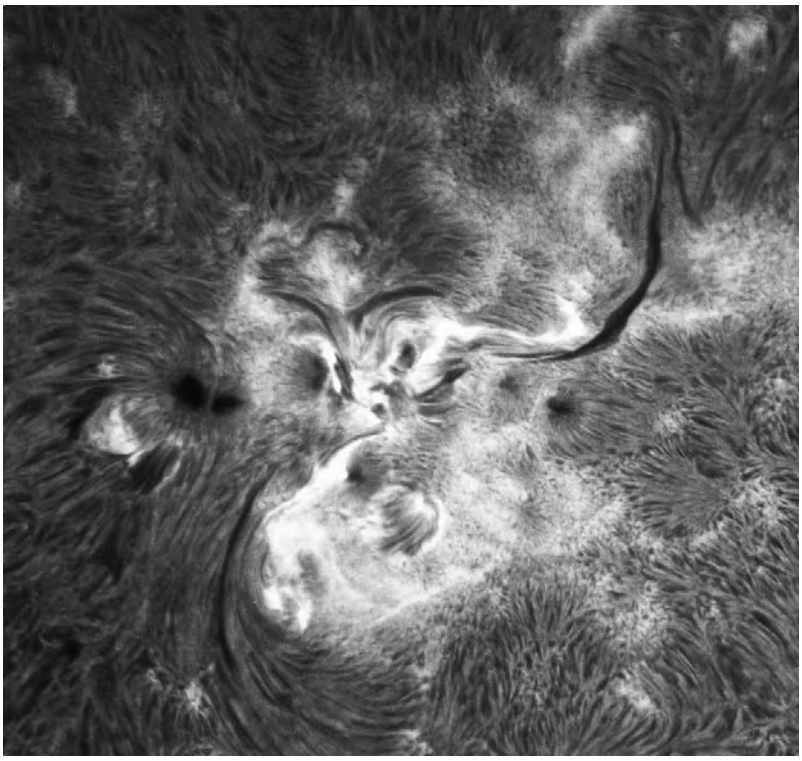


Figure 3. A solar active region with many dark filaments. The arcades of short, curved filaments near the center are called arch filament systems. The very bright ribbons show that a solar flare is taking place (19 September 1979 picture from the Holoman Solar Observatory).

nantly horizontal, a result that confirmed that prominence gases are supported by the magnetic fields against the pull of gravity.

With the advent of observations from space, R Munro and others established a statistical link between prominence eruptions and coronal mass ejections (CMEs) (see SOLAR CORONAL MASS EJECTION: OBSERVATIONS). L Burlaga detected cool gas clouds in the solar wind near Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Typically, one of his magnetic clouds will arrive at Earth about 4 days after a prominence eruption from the center of the visible solar disk. Many such clouds have elevated levels of helium and this may be a sign that they originated in an eruptive prominence. Furthermore, they are embedded in the hot interplanetary magnetic clouds that fly past the Earth and which are associated with CMEs. As solar telescopes in space improve and as more interplanetary plasma detectors are distributed about the heliosphere, it should be possible to sample eruptive prominences systematically and link the sampled plasmas to their precise origins on the Sun.

Figure 4. Loop prominences photographed in the emission of atomic hydrogen with a Lyot filter at the Sacrament Peak Observatory.

Morphology and dynamics

Solar prominences take on a bewildering variety of shapes because of their complex fine structure and their large size. The best observations show structure down to a few hundred kilometers, and prominences (or filaments; the terms are used interchangeably) may stretch to nearly a solar radius (700 000 km) in length across the face of the Sun. Eruptive prominences are frequently 500 000 km long or longer because they expand radially as they stream outward (figure 2).

Prominences that appear in regions well removed from SOLAR ACTIVE REGIONS are called quiescent. They belie their name because films and spectra frequently show motions of 10–20 km s^{-1} at their edges, especially at the top. Nevertheless, quiescent prominences usually maintain their overall shape for days. They often resemble leafy trees or hedgerows on the limb and long (10 000–500 000 km), thin (500–10 000 km) threads on the disk. Quiescent filaments invariably overlie the borders between oppositely directed magnetic fields in the underlying photosphere and chromosphere. They seem confined to 10 000 km wide zones where magnetic fields of opposite polarities mingle in a sort of no man's land between large patches of positive and negative fields. The stability of these large field patterns on the Sun no doubt accounts for the stability of the quiescent filaments. However, the mingling magnetic fields underneath may be just as important for their maintenance as for their stability. The underlying fields are thought to be magnetic flux ropes which merge in the filament channel and expand upward, possibly carrying mass and twisted magnetic fields upward from the chromosphere. Maintaining the 10^{15} g of mass in a typical filament may therefore be a matter of dynamic balance between updrafts in expanding magnetic fields and downdrafts where the horizontal fields are too weak or disordered and material falls into the chromosphere.

Active prominences (figure 3) are those that occur near active regions, of course. They are thinner, darker and more changeable than quiescent prominences and have names that convey it (surges, sprays, loops). Quiescent and active prominences can be distinguished not only by their morphology and motions but also by their spectra. The neutral hydrogen spectral lines and the so-called H and K lines of ionized calcium dominate quiescent prominence spectra, but there are many cool spectral lines, such as those of neutral and singly ionized iron, titanium and chromium. These are also characteristic of the photospheric spectrum, which is emitted by a 5800 K gas. The spectra of active prominences have much brighter and wider hydrogen and helium lines and fainter photospheric lines. Sometimes very highly ionized lines such as the yellow line of Ca XV appear. The latter occurs only in loop prominences (figure 4), which are really a component of solar flares. The presence of Ca XV reveals temperatures above 2 million kelvins. Active prominence spectra also frequently reveal highly broadened lines that vary in brightness, possibly indicative of ion beam heating. Quiescent prominences have remarkably repeatable and stable spectra, which suggests that one of the explanations for their stability may be that the radiation rate is well balanced with the heat input.

Flares, flare waves, and filaments appear in monochromatic images obtained at the wavelength, 6563 Å, of the hydrogen Balmer line, which is commonly called Hα by astronomers. The hydrogen line sometimes shows red and blue Doppler shifts starting when a flare wave reaches a filament. Vertical oscillations can be induced. Some observers suggest that a particularly strong flare wave can cause a filament to erupt from the Sun, but definitive observations are lacking. Whether a filament erupts or not appears to depend mostly on the magnetic fields in and around it.

Quiescent prominences oscillate even in the absence of flares. The oscillations are horizontal and have periods anywhere from less than 5 min up to 90 min, with 7.5 min being most frequently reported. They last only a few cycles and are not well understood theoretically. Considering, however, that prominences are held in place by magnetic tension and that the underlying atmosphere is always oscillating (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY), oscillations are practically inevitable.

Prominence plasma and environment

While most quiescent prominences have very similar spectra, monochromatic images obtained in 'hot' and 'cold' emission lines of the same prominence may differ

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

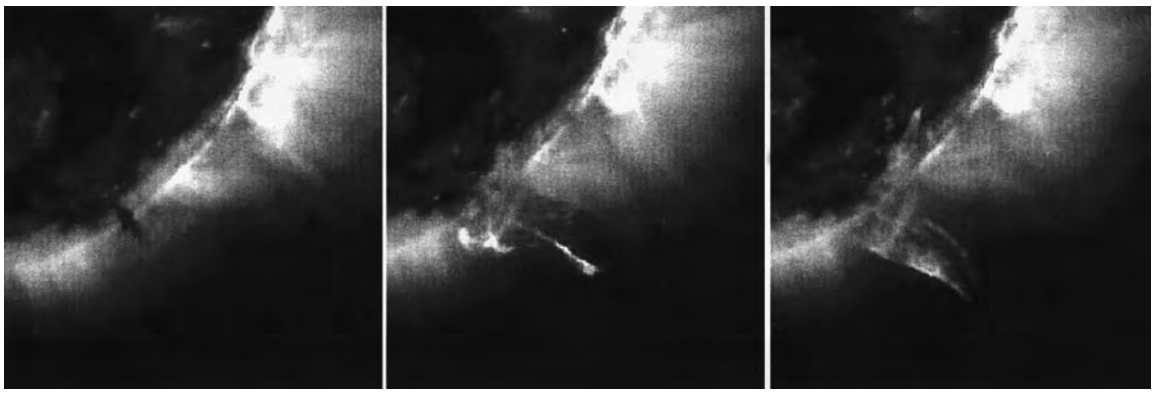


Figure 5. An eruptive prominence showing the transformation of cool, absorbing material (dark feature, left center, left image) into a bright spray (center and right images) emitting the radiation of Fe XII, which exists only at temperatures above one million kelvins.

considerably in appearance. For example, an image formed from the emissions of ionized helium ('hot') may show a high, diaphanous cloud, while an Hα ('cold') image would show a low, dense hedgerow. This means that, although the average properties of prominences are fairly stable, the temperature may vary considerably from 6000 K to 50 000 K from point to point within a single entity. There is some evidence that the local heating rate may change and a prominence may then just 'evaporate' when the delicate normal balance among heat input, radiation rate, density and gravity is lost.

Prominences literally reflect their environment because most of the radiation they emit originated in photons absorbed from the underlying chromosphere. Filaments appear darker against the chromosphere because they absorb most of the upward-directed photons and reemit them in all directions. They appear bright against the dark sky for the same reason. Prominences also radiate from their interfaces with the million kelvin corona, showing that some of the intense heat there is penetrating the magnetic barriers. These radiations are in the ultraviolet region of the spectrum and are indicative of a 100 000 K gas.

Images from the Solar and Heliosphere Observatory (see SOLAR TELESCOPES AND INSTRUMENTS: SPACE) provide dramatic evidence of how efficient the hydrogen and helium of prominences are at absorbing radiation (figure 5). The first image (left) of figure 5 shows a quiescent prominence as a *dark* feature in the corona. This is because the helium atoms in it are absorbing the extreme UV emission at 195 Å from the corona behind it. However, in the other two images, we see the prominence erupt and begin to emit strongly the same high-temperature radiation that comes from the million kelvin corona. Thus, this sequence of images traces the dramatic heating of prominence material at the same time as it is ejected from the coronal cavity in which it rested. Theory suggests that turbulence and reorganization of the magnetic fields heat the material. The heating is too rapid and localized to be explained by thermal conduction.

Figure 6. The S-shaped feature on the right is the signature of an eruptive prominence in this image of the Sun in x-rays, from the Yohkoh satellite.

Magnetic fields

In the presence of a magnetic field, the spectral line emission of hydrogen and other elements becomes partially polarized and broadened, so a sensitive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

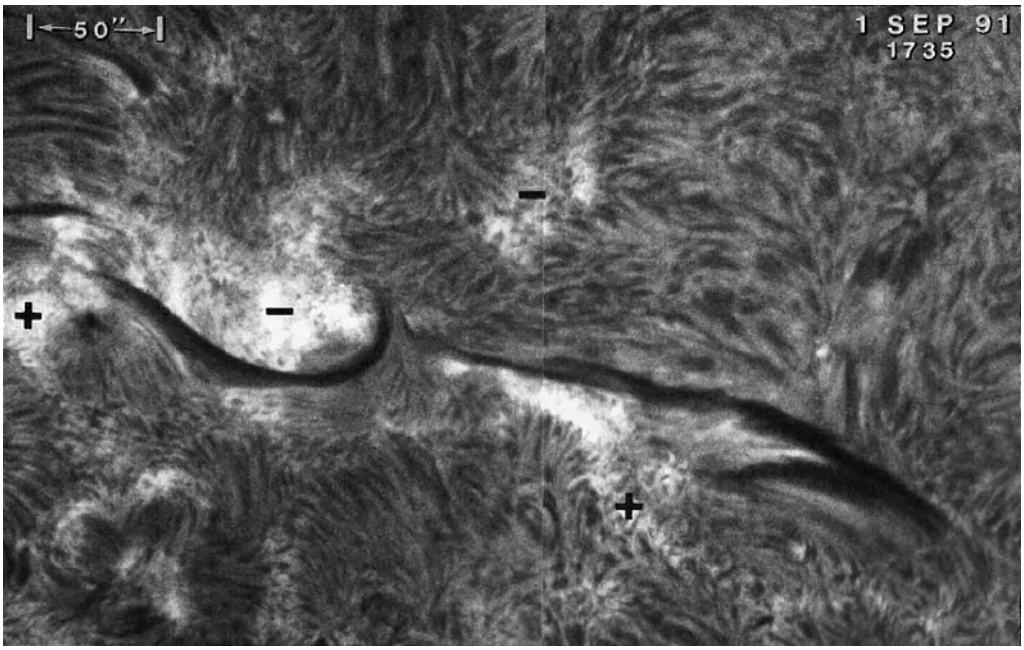


Figure 7. The two filaments shown here have opposite chiralities (sinistral and dextral), yet they spring from nearly the same point (center) in the chromosphere. The sinistral filament on the right clearly resembles a left-hand screw.

polarimeter attached to a telescope will reveal much about the magnetic fields inside prominences. The Zeeman effect, for example, causes circular polarization when the magnetic field is parallel to the line of sight. It produces linear polarization when the field is perpendicular. The sense of polarization changes with increasing wavelength through the spectral line, so a magnetograph measures the rate of change of polarization with wavelength. There is no net polarization in prominence lines because of the Zeeman effect.

The light emitted by prominences is scattered incident light that comes only from the solar disk (i.e. the light is anisotropic). This scattering of anisotropic light causes weak linear polarization parallel to the limb. In the presence of a magnetic field, the plane of polarization will rotate. This is called the Hanle effect (see SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). It depends in a complicated way on the direction and strength of the field and the height of the prominence, but it has been carefully calculated and measured by J–L Leroy and his colleagues, and it is the most useful diagnostic of the magnetic fields in prominences.

The Zeeman and Hanle effects reveal that a prominence's magnetic field vector will be generally within 25° of its long-axis direction. The fields are predominantly horizontal and actually strengthen slightly with height. Theory shows that these characteristics are required for prominence stability and support against gravity. The fields in quiescent prominences most likely resemble twisted ropes called magnetic flux tubes (see SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES). However, the resolving power of present-day magnetographs is considerably less than the smallest features in prominences, so there is some uncertainty about the field's fine structure.

Magnetic fields in active prominences are much less well understood than in quiescent prominences, of which hundreds were measured over three solar cycles. Active prominences are much more difficult to catch. One measurement of fields in a loop prominence gave 60 G fields, which is about 10 times the field in quiescent prominences. However, the morphology of the fields in loop prominences is possibly one of the least controversial issues, because loops (see SOLAR PROMINENCES: ACTIVE, figure 5) look so much like the loops of magnetic field lines seen in textbooks. The smooth slippage of plasma down the legs of a loop prominence and the ease with which the loops' shapes can be matched by elementary models suggests that the magnetic fields in loop prominences have been drained of most of their energy. This is not surprising considering that loop prominences appear only after solar flares. Presumably most of the available energy, in the form of electric currents, has been drained from the fields to produce the flare.

The magnetic fields in eruptive prominences are of great interest because they may be a vital clue to the instability that causes eruptions. However, from its onset to its disappearance from the field of view of most telescopes, an eruptive prominence is visible for only about 20 min. This is hardly enough time to obtain a good measurement of its magnetic fields, so solar physicists have to rely on theoretical models. The models are helped

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

along by a curious feature of eruptives, however. The first stage of many eruptions is an apparent simplification of its structure. What had looked like a hedgerow for a week may, in the course of a few hours, turn into a well-defined dark thread. As it rises off the Sun's limb, an eruptive prominence may resemble a coiled spring. The pitch of the coil is sometimes easy to measure and it decreases as the prominence lengthens. The total number of turns from end to end remains the same even as the length increases.

How can we explain the observations of magnetic fields in prominences? See the article on SOLAR PROMINENCE MODELS for details, but from observations we are drawn to eruptive prominences which often have a characteristic twist. They become unstable and expand outward, and they still have enough internal energy to overcome solar gravity and accelerate 10^{15} g of plasma up to speeds as high as 1000 km s^{-1}. Many theorists have pointed out that fields twisted beyond a certain point, usually about one full turn from end to end, will become unstable. The best evidence that prominences erupt as the result of MAGNETOHYDRODYNAMIC INSTABILITIES was obtained with the x-ray telescope aboard the Yohkoh satellite. As shown in figure 6, a sigmoid (S-shaped) brightening appears at the onset of filament eruption. The ratio of sigmoid length to width in a large sample of eruptives peaks at a value of 5, and this is exactly the shape of the most likely helical kink instability of a twisted magnetic field.

Birth and death

The origin of the cool plasmas of prominences is a controversial subject. It is almost impossible to discover from limb observations how a prominence forms. Observations usually show that a small bright feature grows upward from the Sun's limb. However, this could sometimes be explained simply as a filament rotating onto view, because the Sun rotates completely every 27 days. Other possible explanations are condensation from the surrounding corona or injection of chromospheric material from below. There do not seem to be any convincing observations of clouds condensing from the corona, except in the special case of loop prominences. On the disk, filaments appear to form from the coalescence of thin chromospheric threads that lie parallel to the border separating positive and negative magnetic fields. This may indicate operation of a kind of siphon, or it may suggest that magnetic fields are emerging from beneath the surface parallel to the boundary and carrying material with them as they expand into the corona. Doppler observations of filaments usually show blueshifts, indicative of rising material.

Prominences nestle inside so-called coronal cavities which are so dark that none of the usual emissions from the corona is detected there. This suggests an explanation for the origin of prominence material. Maybe material could somehow be squeezed or condensed from the surrounding corona, in some way that has so far eluded detection. There are many theories for condensation, and a key element of them all is that the million kelvin coronal gases must cool to 10 000 K and even less. The process seems to occur in loop prominences where a great quantity of hot, dense gas is created in the corona and conditions are right for condensation by thermal instability. A hot gas radiates energy, which cools it. It then becomes more dense and then radiates more, because the radiation rate is proportional to the square of the density. This instability nicely explains loop prominences where stages of ionization indicative of temperatures from 10 million kelvins down to 10 000 K are present. In normal prominence formation, however, there are no high-temperature emissions. Another problem is that the mass deficiency of a typical coronal cavity is much smaller than the mass of a typical prominence.

Global patterns

Filaments form along the magnetic polarity borders in active regions, between adjacent active region remnants and between active region remnants and the polar fields. These three families of filaments occupy three tiers of latitude in each hemisphere. The active ones lie between the equator and 30°, the next between 30° and 50°, and, finally, the polar crown filaments between 50° and 80°.

The orientation of the axial magnetic fields in active region filaments follows Hale's law of sunspot polarities. A filament's axial field will have its western end in or near the leader spots. The eastern end will be near the follower spots. The axial fields flow from near the leader polarity spots to near the follower polarity spots. Of course, the direction changes each 11 yr, and it is opposite in the opposite hemispheres, consistent with Hale's law.

As active regions age and spread out, new filaments form along their high-latitude borders. In these filaments the axial fields flow from the *follower* fields of one active region remnant eastward to the *leader* fields of the next remnant. The average latitude of these filaments increases as the active region remnants expand toward the poles. Shortly after the beginning of a new cycle, the filaments of this 'first tier' begin to appear. By the beginning of the next cycle, these filaments form two crowns around the Sun at about 50° north and south latitude. Individual filaments will come and go, but the loci of active region, first tier and polar crown filaments are stable features on synoptic maps of the chromosphere.

The polar crowns persist until about 1 yr before the maximum of the sunspot cycle. Then, in the course of 1 yr, they shrink toward the poles. The average latitude increases quickly from 50° to 80° and at about sunspot maximum all the polar crown filaments disappear as the sign of the polar fields changes. To summarize, the axial fields of filaments in the low-latitude sunspot zones point to the west when negative spots follow (are westward of) positive spots. At the same time, the axial fields in the first tier of filaments point eastward, and those in the polar crown point westward. It is as though sections of the global toroidal fields of three sunspot cycles were manifested piecemeal in the corona by solar filaments.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Until 1992 no one suspected that filaments in the north were any different from those in the south. Then, S Martin noticed that barbs along the sides of polar crown filaments and those in the first tier are consistently slanted one way in the north and another way in the south. Northern hemisphere filaments somewhat resemble right-handed screws. Martin called them dextral filaments. Those in the south resemble left-handed screws and are called sinistral filaments. The pattern is less evident in active region filaments, but even they seem to differ statistically between north and south. Sometimes a dextral filament and a sinistral filament each have one end in a common point in the chromosphere (see figure 7), but, otherwise, the two filament types are never connected.

It now appears, although there is still some controversy, that most magnetic fields in the northern hemisphere are composed of flux ropes with a left-handed twist. Those in the south have a right-handed twist. Why filaments threaded by fields with a right-handed twist should resemble left-handed screws is not clear.

The importance of the chirality segregation by hemisphere rests on theories of the solar cycle, which predict that the magnetic helicity (tendency to twist) of northern and southern fields should be opposite. This is just what has been inferred from solar filament observations, and it has been confirmed by measurements of the fields in space plasmas and in active regions.

Eruptive prominences preserve their sense of twist. Recent studies indicate that those erupting from the north become left-handed magnetic clouds in space. Similarly, measurements of the magnetic fields in the eruptive prominences from the southern hemisphere, as they pass by Earth, reveal magnetic flux ropes with a right-handed twist. Quite a lot of magnetic flux and helicity leave the Sun each solar cycle in eruptive prominences and coronal mass ejections. The total flux and helicity measured in space during each 11 yr solar cycle is equal to the total flux and helicity of the eruptives and CMEs. Also, these values agree with the total flux and helicity that the solar cycle generates, according to theoretical calculations. It appears, then, that eruptive prominences and CMEs carry off very nearly all of the magnetic fields generated in each solar cycle.

Future research

Solar prominences are much better understood today than a century ago, of course, but many fundamental questions remain. There is no consensus on how prominences form. Dark threads develop in filament channels, but it has been impossible from the images at hand to determine where the material comes from. Observations eventually will show whether it comes up from the chromosphere or condenses from the surrounding corona.

Although some puzzling details remain unresolved, the stability of prominences and their resistance to heat conduction from the corona are generally understood. Magnetic tension supports the material, and the extremely low rate of thermal conduction across magnetic field lines prevents temperature changes. When prominences do become unstable, it is probably because of the helical kink instability. With sufficient magnetic helicity, any prominence will become unstable. However, we do not know how the magnetic helicity accumulates.

No theory explains the fine vertical and horizontal threads of a hedgerow prominence. Perhaps three-dimensional images will reveal them to be part of a simple pattern of concentric coiled fields. Future observations of prominences will be carried out with matched telescopes operating from two or more widely separated points in the heliosphere. The arrival of stereoscopy will add a new dimension to prominence research.

Bibliography

Foukal P V 1990 *Solar Astrophysics* (New York: Wiley)

Tandberg-Hanssen E 1995 *The Nature of Solar Prominences* (Dordrecht: Kluwer)

Webb D, Rust D and Schmieder B 1998 *New Perspectives on Solar Prominences* (San Francisco, CA: Astronomical Society of the Pacific)

Zirin H 1988 *Astrophysics of the Sun* (Cambridge: Cambridge University Press)

D Rust

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

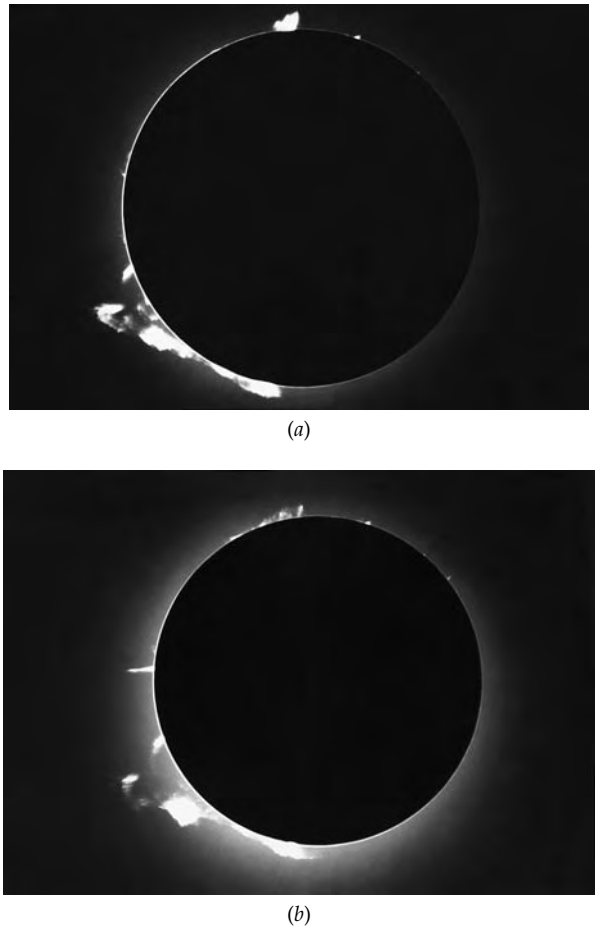
Solar Prominences: Active

Solar prominences may be divided into two classes, quiescent and active objects (see SOLAR PROMINENCES). While quiescent prominences form a fairly homogeneous group, active prominences comprise a bewildering array of objects, from filament-like prominences in active regions to surges, sprays and loop structures associated with solar flares (see SOLAR FLARES and flare SURGES). The many morphological classifications published in the 1920s–1960s that tried to describe the many aspects of active prominences now mainly have historical interest. Nevertheless, several of these classifications invoked crucial physical parameters, e.g. velocity, association with active regions or flares, and helped to establish the nature of these active objects.

In the following we shall discuss in some detail some common and fairly easily discernable classes of active prominences, i.e. active region filaments and associated sprays as well as surges and loops. Quiescent prominences often ascend in a spectacular eruption called a disparition brusque. These eruptive, quiescent prominences are treated separately (see SOLAR PROMINENCE ERUPTION) and are not to be considered active prominences *per se*. Note that when both quiescent and active region prominences are seen in absorption, e.g. in the H-alpha line, against the solar disk, they are often referred to as filaments.

Both surges and loops occur in conjunction with flare activity and may be considered part of an overall flare phenomenon. However, they are also *bona fide* prominences, i.e. contain 'cool plasma' in the hot corona, and as such will be discussed below. A number of descriptive names are given to other active prominences, e.g. coronal clouds, caps and fast ejections, but we shall concentrate on the classes mentioned above that constitute the large majority of active objects. Figure 1 illustrates the activity that can be seen in the solar atmosphere when the Sun's prominences put on their impressive show.

As a general rule, the magnetic field in active prominences is considerably stronger than the roughly 1D G field that characterizes quiescent prominences. In surges and loops the field may search values of 100–200 G.



(*a*)

(*b*)

Figure 1. Development of active prominences observed in H-alpha on 30 July 1990: (*a*) at 1400 UT; (*b*) at 1644 UT. Notice eruption of prominence on south-east limb, appearance of surge at east limb, and major change in prominence near north pole. (Courtesy Observatoire du Pic-du-Midi and J-C Noens.)

Active region filaments

These prominences always occur in, or very close to, active regions, normally with sunspots, and are also referred to as sunspot filaments. Figure 2 shows, in addition to several large, quiescent prominences, a number of small active region filaments.

Filaments, both quiescent and active, are found in regions between opposite magnetic polarities. These regions form narrow lanes between the extended plages of magnetic field, and are referred to as the polarity-inversion zone where we find the filament channel (see SOLAR FILAMENT CHANNELS). The occurrence on filament-type prominences seems to be intimately linked to the concept of chromospheric fibrils (see CHROMOSPHERE: FIBRILS). These are one of the most conspicuous features seen in the H-alpha line in and around active regions. In the central part of active regions the fibrils are arranged in patterns connecting spots and plages of opposite magnetic polarity. However, before a filament will form, the fibrils will no longer connect these areas of opposite polarities, but will curve into the polarity- inversion zone. In the middle of this channel the fibrils will thereby be more or less aligned along the direction that will become the long axis of the forming filament.

The long, thin fibrils typically have lengths of 11 000 km, a width of from 700 to 2200 km and a lifetime of 10–20 min. Active filaments come in nearly all sizes, from small structures barely larger than fibrils, to long—always low-lying—filaments measuring 100 000 km or more. While the two ends of the filament dip down into chromospheric plages of opposite magnetic polarities, the main body of the filament forms a long, nearly horizontal, magnetic flux tube.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

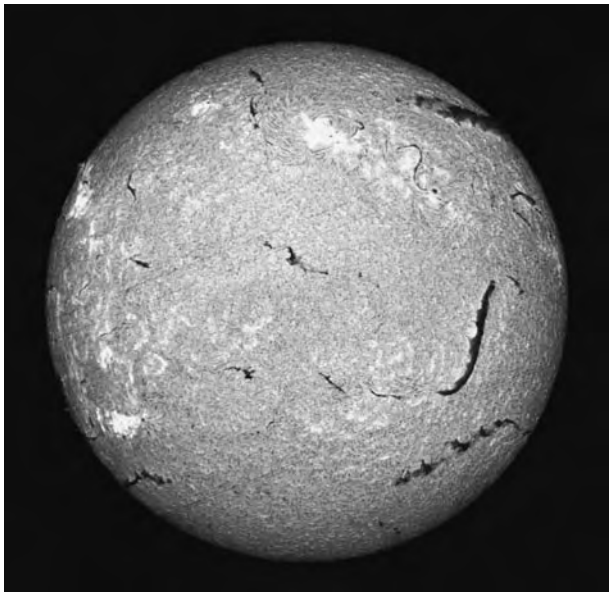


Figure 2. Spectroheliogram in H-alpha taken 16 March 1990 showing three large quiescent prominences on western hemisphere (right) and several active region prominences, in particular inverse S-shaped filament in north-west quadrant (upper right). (Courtesy Observatoire de Paris-Meudon and P Mein.)

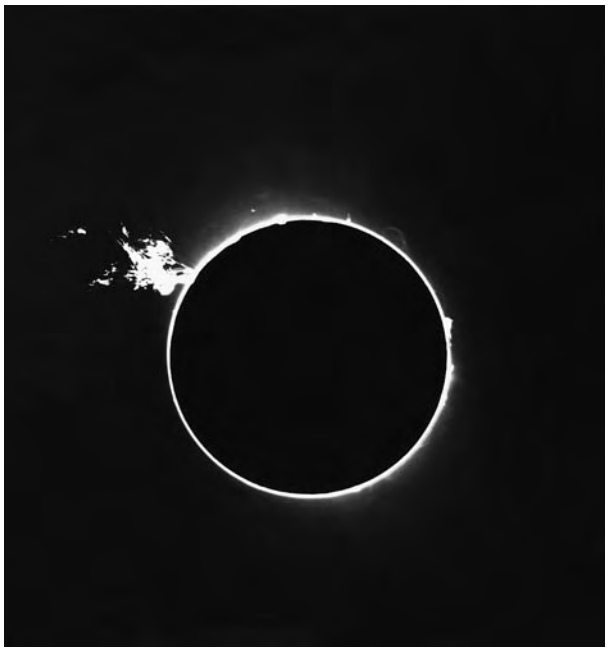


Figure 3. Spray prominence observed in H-alpha, 28 October 1972. Climax Station of the High Altitude Observatory, Boulder, CO.

No generally accepted theory exists for the formation of these prominences. Models range from invoking a coalescence of fibrils, pre-existing in the filament channel, to picturing the emergence from subphotospheric layers of the already complete magnetic flux tube anchored in plages of opposite polarity.

Since the magnetic field is oriented along the long axis of the filament, material motion along this axis will not be impeded, and under favorable conditions a continuous mass flow is seen in most active filaments. This flow is one of the strongest descriptive differences between active and quiescent filaments, and it demands a completely dynamic model for the former.

Active region filaments may last from hours to several days. Their disappearance is preceded by an activation that may manifest itself either by increased internal motion, up to 30 km s^{-1}, along the axis of the filament, or—in spectacular cases—by a genuine disparition brusque. These activities normally precede flares and hence furnish a good flare precursor (see SOLAR FLARE FORECASTING). Particularly violent expulsions of active filaments may result in sprays.

Sprays

Sprays are violent, flare-associated ejections of plasma which are frequently disrupted into clumps. They reach velocities of 500–1200 km s^{-1} in a few minutes owing to an initial very high acceleration of a few km s^{-2}. Figure 3 shows an example of a spray prominence.

When sprays originate on the visible disk one sees that the ejected material comes from a pre-existing active region filament which undergoes increased absorption some tens of minutes prior to the abrupt chromospheric brightening at the start of the spray. Often most of the spray material from the filament is confined within a steadily expanding loop-shaped envelope with parts of the material draining down both legs of the loop. In other instances the material is sprayed out in a seemingly chaotic fashion. Some material may escape the Sun as its velocity exceeds the velocity of escape. The nature of the expelling force is not known, but energy freed in magnetic reconnection seems a likely candidate.

Surges

Surges are prominences that seem to be shot out of active regions as long straight or curved columns and return along the same trajectory; see figure 4. As the material falls back, it often triggers a new surge. Surges may reach to great heights (several hundreds of thousands of km), and their velocity may exceed several hundred km s^{-1}. Some active regions produce nearly identical surges during part of their life (homologous surges). As the material is ejected, it often performs a spiraling motion on its way out. A flare is often seen at the foot point of a surge, and the phenomenon may be considered as part of the flare complex.

The trajectory of the moving plasma and its collimation indicate that surges are confined by more or less radial magnetic fields. From combined H-alpha, UV and x-ray data one can study both the cool and the hotter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

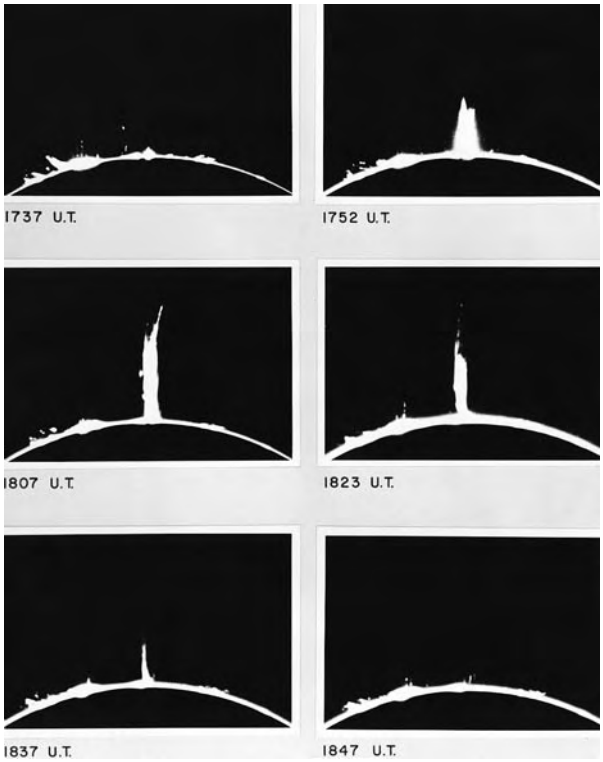


Figure 4. Development of surge prominence, 12 June 1946. Climax Station of the High Altitude Observatory, Boulder, CO.

components of the surge plasma, and one finds that cool and hot material follow parallel but separate paths. It seems that different parts of the surge may be ejected into adjacent, but separate, flux tubes. The nature of the explosive force that propels material into a surge is not known, but magnetic reconnection in the active region field again has been suggested.

Loops

Some of the most beautiful examples of gracious motion of material in the corona are provided by loop prominences and coronal rain; see figure 5. In these objects matter is observed to flow down along curved trajectories into active regions, seemingly after condensing out of the corona, often at great heights. These prominences are intimately connected with the late phase of the flare phenomenon and are, therefore, referred to as post-flare loops. They occur after the disparition brusque of a filament (quiescent or active), forming an arcade of loops connecting the two strands of emission appearing as a two-ribbon flare. The systematic motion in these loops reveals material streaming down the two legs of the loops from their apexes. The streaming motion follows a single arc; one does not observe the spiraling motion often seen in surges and in erupting quiescent prominences. During the next several hours the loop system expands and reaches, typically, a height of 50 000 km. It is important to note that

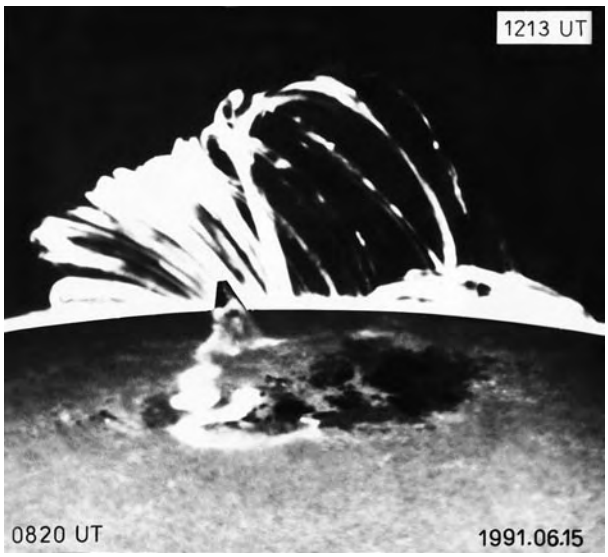


Figure 5. Composite H-alpha picture showing loop–prominence system in the corona above the solar limb (at 1213 UT) and the flares near the loops' feet in active with sunspots (at 0820 UT), 14 June 1991. (Courtesy Astronomical Observatory, Wrocław, Poland, and B Rompolt.)

the individual loops do not grow or expand much; rather, the system expands by generating higher and higher loops, while the lower ones fade away. The resulting apparent velocity of expansion is quite small, $\sim$5 km s^{-1}. The real downward flow motion in the two legs of the loop is considerably faster. (Free fall of 8–10 min. would give 130–160 km s^{-1}, but much lower values have been observed.) The density and the magnetic field in the loops strongly influence the velocity of the downward flow.

Oscillations of loop-type prominences have been reported and may be associated with the flaring activity. These post-flare loops should not be confused with the loop-like structures that at times are seen during a flare display or other plage activity, when material is ejected into radial flux tubes, giving surges, or into loop-shaped flux tubes that bend over toward the surface. In the latter case the motion is along the loop, moving up one leg and down the other. In particularly large loop structures that reach high into the corona, the ejected material may form so-called flaring arches.

In coronal rain material is often seen to rain down into active regions from coronal clouds suspended in the corona. The trajectories followed by the streaming material generally are strongly curved, apparently following the lines of force of coronal magnetic fields. The velocities along these paths are comparable with the flow velocities observed in loop prominences.

Models for post-flare loops have been worked out (see SOLAR PROMINENCE MODELS and SOLAR FLARE MODELS) and imply magnetic reconnection in the arcade field that surrounds the pre-existing quiescent prominence involved.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Tandberg-Hanssen E 1995 *The Nature of Solar Prominences* (Dordrecht: Kluwer)

Einar Tandberg-Hanssen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy and Diagnostics

Solar spectroscopy and diagnostics provide the opportunity for determining the physical parameters in different parts of the solar atmosphere.

The story of solar spectroscopy really began in 1666, when Isaac NEWTON studied the *phenomenon of colors* using a triangular glass prism which he bought in a local market. In his early experiments, he split sunlight into the colors of the rainbow. The radiation from the photosphere is similar to that of a black body with a characteristic temperature of 6000 K, which peaks in the visible wavelength range. It was only much later, in 1814, when FRAUNHOFER determined that the hundreds of dark lines in the solar visible spectrum were actually a property of the sunlight, rather than any earthly phenomenon (see FRAUNHOFER LINES). During a total eclipse of the Sun, the Fraunhofer spectrum, normally seen as absorption lines, is briefly replaced by an emission-line or flash spectrum from the chromosphere, the thin layer just above the photosphere.

QUANTUM THEORY, developed in the 1920s and 1930s, provided an explanation for spectral emission and absorption lines in terms of changes of energy state of atoms by discrete amounts—*quanta*, related to the wavelength of the radiation. Each atom has characteristic *fingerprints* of spectral lines which enable us to dissect the spectrum and determine the nature of the solar plasma where it originated. A plasma is a gas in which the atoms have been ionized.

At the time of totality, the CORONA is visible in all its glory—the *crown* around the Sun. Normally, the visible emission from the corona cannot be seen, since it is very weak, only a millionth of the intensity of the SOLAR PHOTOSPHERE or surface of the Sun. During an eclipse, the Moon blocks out exactly the photosphere and allows us a brief glance at the corona. Skilled work by a gifted French astronomer called LYOT, provided in 1930 an instrument, a coronagraph, which could simulate a total eclipse, allowing the visible coronal spectrum to be continuously observed. The coronal visible spectrum comprises emission lines (the brightest are the coronal green, red and yellow lines). These lines provided an intriguing puzzle for many years, indeed, a new element called *coronium* was invented to explain them. Although Lyot noted that the width of the green line indicated a characteristic temperature much higher than the photosphere, it was the painstaking scientific investigations by Grotian and Edlen around 1940 which eventually led to the identification of these lines from ions which only exist at very high temperatures (above 10^6 K). The green line comes from Fe XIV (iron with 13 electrons stripped off), the red line from Fe X and the hottest line, the yellow line, is from Ca XV. Thus it was established that the corona is very much hotter than the photosphere—a fascinating phenomenon, which solar physicists are still seeking to explain.

Modern technology has enabled us to move far beyond the visible wavelength range into the ultraviolet, x-rays, infrared and radio wavelength ranges. From space observatories, we are able to monitor the corona continuously in the UV and x-rays. At these wavelengths we see hundreds of emission lines, covering a wide range of temperatures. Spectroscopic diagnostics encompasses the study of techniques which have been developed to deduce plasma parameters, such as electron density and temperature, elemental abundances and mass motions from spectral lines. A comprehensive review of spectroscopic diagnostics for solar and stellar plasmas in the VUV (100–2000 Å) was published by Mason and Monsignori Fossi (1994). To use these techniques it is necessary to build accurate atomic models, including all the important processes and using the best available atomic data.

Electron temperature

As the temperature increases in a gas, the particles become very energetic, so more and more electrons get stripped off the ions. For example, neutral oxygen has eight electrons, but at a temperature of 10^5 K, four electrons have been removed. The ion formed is O V (or O^{4+} in notation used by atomic physicists). So, if we make images of the solar atmosphere in different spectral lines, we are in fact able to take 'slices' at different temperatures. For example, the Coronal Diagnostic Spectrometer (CDS) on the Solar and Heliospheric Observatory (SOHO) was designed to do just that. Simultaneous rasters, up to $4' \times 4'$ in size, can be obtained at different wavelengths, with a high spatial resolution (approximately $3''$) (see figure 1). The SOHO-CDS instrument covers lines from ions formed over a wide temperature range (2×10^4–6×10^6 K), including the low-temperature emission from He I at 584 Å (2×10^4 K), transition region emission from O V 630 Å (2.5×10^5 K); the Mg IX 368 Å line at coronal temperatures (10^6 K) and the Fe XVI 335 and 361 Å lines observed in active regions (2×10^6 K). From the CDS rasters it is evident how different the solar atmosphere looks at various temperatures.

During solar flares, the temperature can exceed 10^7 K, and many lines from highly ionized iron ions can be observed in the x-ray wavelength range. Figure 2 is an observation made with the Bent Crystal Spectrometer (BCS) instrument on the YOHKOH satellite (ISAS, Japan). The presence of the Fe XXVI lines indicates a very high temperature of around 30×10^6 K.

In the simplest coronal model approximation, we find that the emission line intensity from a volume of plasma V can be expressed as

$$I(\lambda_{ij}) = \frac{1}{4\pi} Ab(X) \int_V G(T) N_e^2 \, \mathrm{d}V \, (\text{photons}\,\text{s}^{-1}\,\text{sr}^{-1})$$

where λ_{ij} is the wavelength for the transition between energy level j and i in the ion, $Ab(X)$ is the element abundance and N_e is the electron number density (cm^{-3}). The *contribution function*, $G(T)$, contains all of the relevant atomic physics parameters—it is strongly peaked in temperature (see figure 3).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

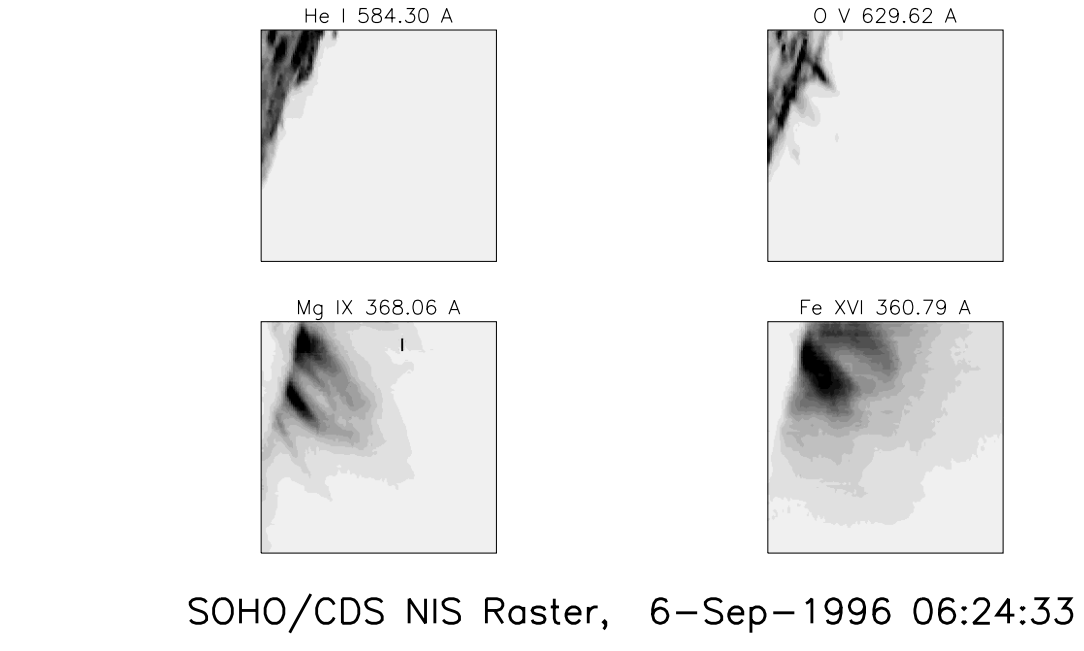


Figure 1. SOHO-CDS rasters of an active region, showing emission from He I, O V, Mg IX and Fe XVI.

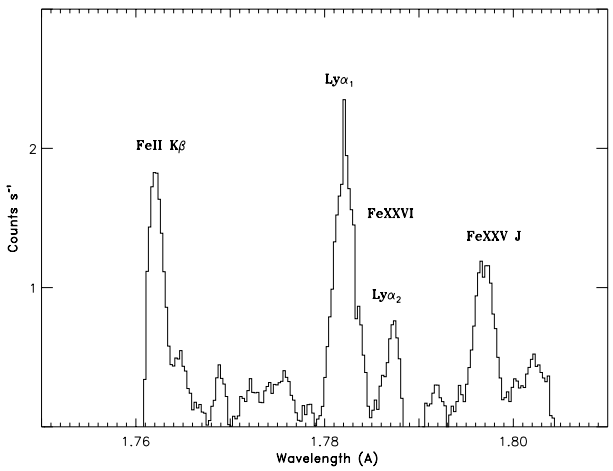


Figure 2. Fe XXVI lines observed with the YOHKOH BCS instrument. (Courtesy of Pike.)

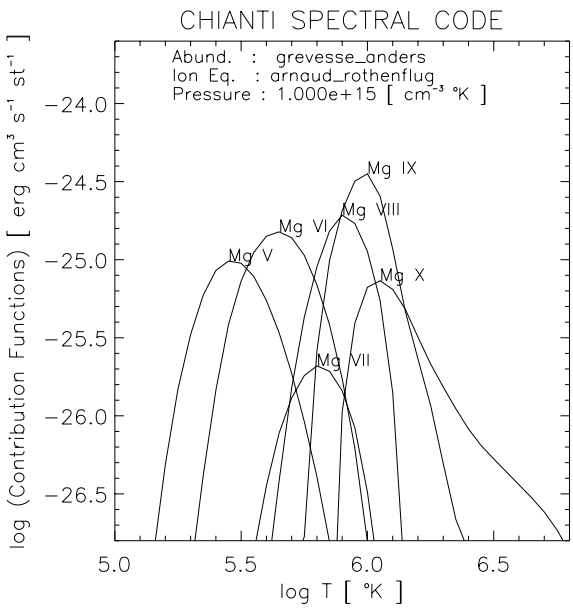


Figure 3. The contribution functions $(G(T)Ab(X)/4\pi)$ for various ionization stages of Mg: Mg V (353.09 Å), Mg VI (349.17 Å multiplet), Mg VII (363.77 Å), Mg VIII (315.04 Å), Mg IX (368.07 Å), Mg X (624.94 Å).

From a study of the intensity of spectral lines formed at different temperatures, it is possible to deduce the amount of material in each temperature range, which is called the *emission measure distribution*. With further assumptions about the geometry of the region, one can directly compare the observations with theoretical models.

Vast quantities of atomic data are required to simulate the observed spectra. For example, the CHIANTI atomic database and analysis software (Dere *et al* 1997), has been developed as a collaboration between the USA, Italy and the UK to provide a comprehensive dataset for ions of astrophysical interest. A comparison of the simulated and observed CDS spectrum is given in figure 4. Several other atomic databases and packages are available such as the Atomic Data and Analysis Structure (ADAS) developed at Strathclyde University.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

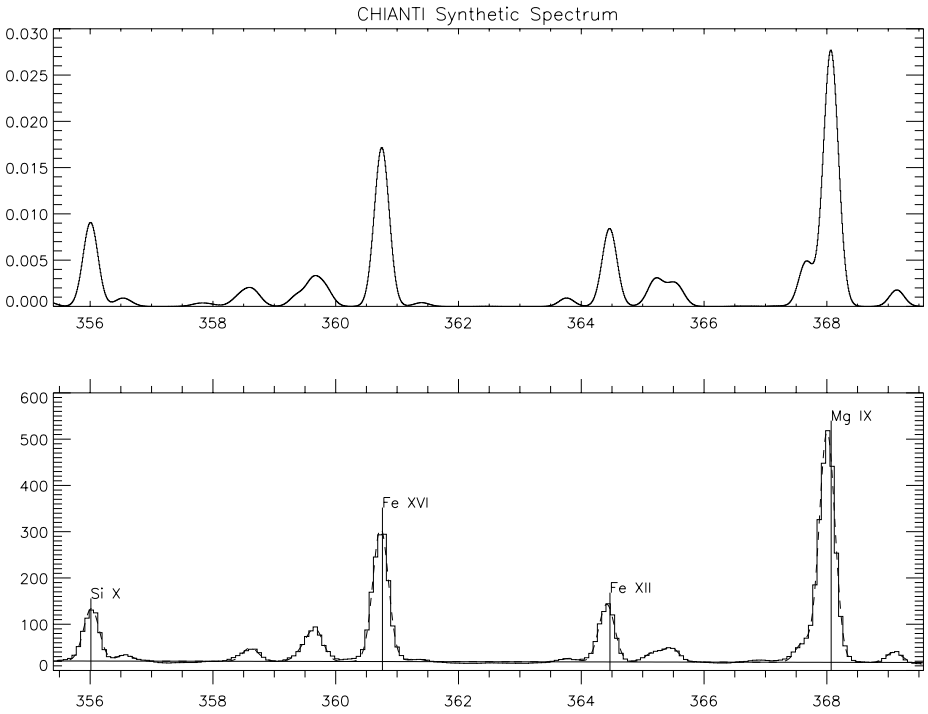


Figure 4. A sample CDS-NIS active region spectrum (355–370 Å) is shown in the lower plot—the dashed curve is a multi-Gaussian fit with background. The upper plot is a standard active region spectrum from CHIANTI.

Electron density

The electron pressure is an important parameter in any theoretical model for the plasma. This is proportional to the product of the electron density and temperature. Experience from solar observations is that the plasma often exists in the form of unresolved filamentary structures, even down to the best spatial resolution which has yet been obtained. At one extreme is the solar transition region, where only a very small fraction of the observed emitting volume is actually filled with plasma.

An estimate for the electron density can be deduced from the absolute intensity of one emission line, if the relevant atomic parameters are known. However, this method depends on several assumptions and breaks down if any filamentary structure exists.

Techniques have been developed to determine electron density from spectral line intensity ratios for the same ion. These methods make no assumption about the size of the emitting volume or the element abundance value. They therefore provide a powerful and important diagnostic for the solar plasma. Spectral lines may be grouped into different categories according to the way in which they are produced. In the coronal model approximation the spectral line intensity is a function of N_e^2; however, for some transitions, the dependency on N_e is different. The atomic processes within each ion determine the way in which the spectral line intensities vary with N_e.

For example, Si X, formed at around 1.3×10^6 K, provides a useful electron density diagnostic with lines at 347.4 Å and 356.0 Å. The intensity ratio (356.0/347.4) varies with electron density as shown in figure 5.

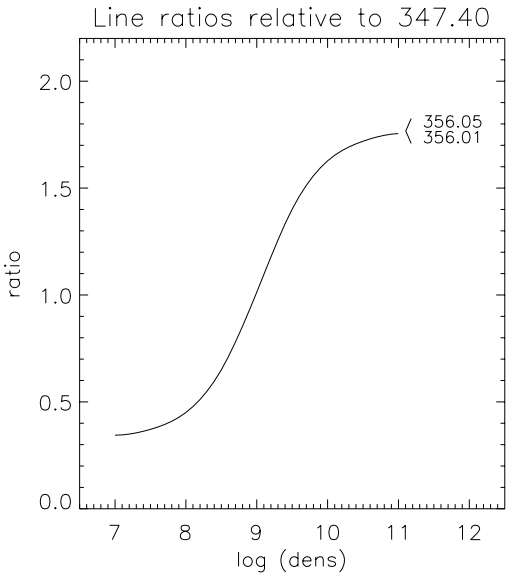


Figure 5. The Si X density sensitive 356.03/347.40 ratio (the observed 356.03 Å line is a blend of two Si X lines).

In the hottest parts of active regions the 356.0 Å line can be found to be around twice as intense as the 347.4 Å

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

line, implying a high electron density $>10^{10}$ cm^{-3}. In contrast the characteristic electron density for the quiet Sun is around 10^8 cm^{-3} and in coronal holes it is even less.

Elemental composition

There has been a great deal of discussion and controversy about variations in the elemental abundances in the solar atmosphere (see SOLAR ABUNDANCES). The element abundances measured in the solar wind differ from those in the photosphere. The coronal abundances also differ from those in the photosphere, in fact they seem to vary in different solar features. This behavior depends on the value for the first ionization potential (FIP)—which is the energy required to ionize the neutral atom. The ions with FIP greater than 10 eV appear to behave differently from those with FIPs less than 10 eV. It is believed that this could reflect the ionization and acceleration processes for the solar corona, low down in the solar atmosphere.

The intensity of a spectral line is directly dependent on the elemental abundance. One approach to determining element abundances is to use the detailed shape of the emission measure distribution for ions from the same element and apply an iterative procedure to normalizing the curves for different elements. Another procedure is to use the intensity ratios for individual spectral lines which have very similar $G(T)$ functions, but different FIPs. For example, neutral neon, with an FIP of 21.6 eV, has a closed shell structure which is difficult to ionize, whereas it is much easier to ionize magnesium, with an FIP of 7.6 eV. Studies of the Mg/Ne relative abundances indicate that some low-lying dense structures in active regions have photospheric abundances, whilst other larger more open-loop structures have coronal abundances. This could indicate the emergence and heating of dense flux tubes from beneath the solar suface.

Elemental abundance determinations in a CORONAL STREAMER have recently been made with the Ultraviolet Coronagraph Spectrometer (UVSP) on SOHO. The UVSP team find a depletion of the high-FIP elements in the center of a quiescent streamer and suggest that gravitational settling may be partly responsible. It is important to attempt to relate the elemental abundances determined in the solar atmosphere with those measured further out in the solar wind.

Spectral line profiles

Line shifts and broadenings give information about the dynamic nature of the solar atmosphere. The transition region lines, formed at around 10^5 K, are characterized by broadened and irregular line profiles, which put constraints on possible heating processes. These profiles show evidence for small explosive events which occur around the edges of the magnetic network. They could be linked to coronal heating processes in the quiet Sun. Transient brightenings in transition region lines, called *blinkers*, have been observed by SOHO-CDS.

Many detailed studies have been carried out for a variety of solar features. For example, figure 6 shows a macrospicule-like feature at the limb, which was recorded by SOHO-CDS (Pike and Mason 1998). The left-hand side shows the OV intensity raster and the right-hand side shows an OV velocity map. If an axis is defined through the center of the OV feature and extended above the limb from the footpoint region, the emission is apparently blueshifted (black) on one side of the axis and redshifted (white) on the other. It is likely that a combination of both rotating and accelerating plasma would explain these observations.

Outflows of coronal material have been correlated with CORONAL HOLES, the source of the fast SOLAR WIND. The excess broadening of the transition region and coronal lines could be due to macroturbulence and provides valuable information on the heating of the solar atmosphere and the SOLAR WIND ACCELERATION.

The SOHO-UVCS instrument is designed to carry out diagnostic studies of the extended corona, in particular to measure coronal line intensities and profiles. For the O VI lines, a process called Doppler dimming enables a determination to be made of the solar wind outflow velocity. In the equatorial streamers, the outflow velocity reaches a value of around 100 km s^{-1} at four solar radii. Bright ray-like structures have been observed in coronal holes which could be related to POLAR PLUMES. The SOHO-LASCO (Large Angle Spectroscopic Coronagraph) instrument comprises three coronagraphs which cover different distances out into the extended corona. The inner coronagraph has the capability of observing in the green, red and yellow lines as well as in white light. Dynamic features can be tracked from close to the solar surface right out to 30 solar radii. In particular LASCO has observed many CMEs (coronal mass ejections), streamers, polar plumes and comets.

Conclusion

Spectroscopy provides a powerful opportunity for probing the nature of the solar atmosphere. This is illustrated by recent advances in UV observations of the transition region and corona by SOHO. In particular, examples have been given for diagnostic techniques involving optically thin emission lines. Lower down in the solar atmosphere (photosphere and chromosphere), it is necessary to solve the radiative transfer equations to study emission or absorption features.

The x-ray spectrum has only briefly been mentioned. The SOLAR MAXIMUM MISSION (SMM) satellite flown in the 1970s and the more recent YOHKOH satellite have provided extensive observations of active regions and solar flares in the x-ray wavelength range. A multitude of diagnostic possibilities exist and have been explored with these data.

From diagnostic studies of the solar atmosphere, we now know that the transition region is dynamic and filamentary in nature. MAGNETIC RECONNECTION at the network boundaries is correlated with enhanced transition region emission, indicative of heating. The corona is confined by small- and large-scale magnetic features. In active regions, the temperature is high and the measured

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

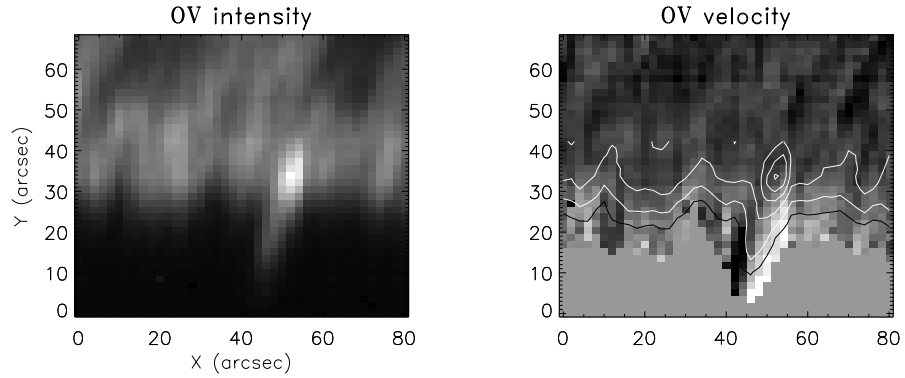


Figure 6. OV intensity (lhs) and velocity (rhs) maps for a solar tornado. The contours on the rhs velocity map are OV intensity. (From Pike and Mason 1998.)

electron densities are at least an order of magnitude higher than the quiet Sun. However, the densest, hottest parts of active regions are often low down, close to the regions of greatest magnetic shear. Large, cool (around 10^5 K) loops are a common feature of active regions. These are not stable and large velocities (around 100 km s^{-1}) have been recorded. In coronal holes, diagnostic studies give a low electron density and a characteristic temperature below 10^6 K. A heated debate is in progress about whether it is the polar plume or inter-plume regions which are the source of the high-speed solar wind. Initial velocity measurements from SUMER indicate that it is the latter, but further studies are in progress.

One has to stop and ponder what Newton would have thought about the advances which have been made in spectroscopy. Space technology has provided the opportunity to go far beyond the colors of the rainbow. Would he have shared our enthusiasm and excitement for studying the intricate features of the Sun?

Bibliography

Dere K P, Landi E, Mason H E, Fossi B C and Young P R 1997 CHIANTI—an atomic database for emission lines *Astron. Astrophys. Suppl. Ser.* **125** 149–73

Fleck B and Svestka Z (ed) 1997 *The First Results from SOHO* (Dordrecht: Kluwer)

Mason H E and Monsignori-Fossi B C 1994 Spectroscopic diagnostics in the VUV for solar and stellar plasmas *Astron. Astrophys. Rev.* **6** 123–79

Phillips K J H 1992 *Guide to the Sun* (Cambridge: Cambridge University Press)

Pike C D and Mason H E 1998 Rotating transition region features observed with SOHO-CDS *Solar Physics* **182** 333–48

Vial J-C, Brocchialini K and Boumier P (eds) 1998 *Space Solar Physics* (Berlin: Springer)

Helen Mason

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy and Diagnostics: X-ray Emission

X-ray emission from the Sun arises primarily from the solar outer atmosphere, particularly the hot corona. During transient events, such as flares and coronal mass ejections, there is also enhanced emission of both soft and hard x-rays from plasma above ten million kelvins (see SOLAR FLARES: RELATION TO CORONAL MASS EJECTIONS). Outside of flares, the bulk of the coronal plasma is at temperatures in the range 1–10 MK, leading to radiation primarily in spectral lines from ions of heavy elements in the atmosphere formed at these temperatures. Above 10 MK, the atoms of the coronal plasma are stripped of most of their outer electrons and continuum emission gains in importance.

Figure 1 shows a calculation of the soft x-ray emission (to 100 Å) of the coronal plasma at a temperature of 3 MK, which is typical of SOLAR ACTIVE REGIONS. The quantity plotted is the spectral emissivity $P(T, \lambda)$, which is the plasma emission summed over all spectral lines and continuum processes, but kept as a function of wavelength. $P(T, \lambda)$ is therefore the spectrum emitted by a unit volume of the coronal plasma; it has units erg cm^3 s^{-1} Å^{-1}.

The emission spectrum is seen to be dominated by individual strong emission lines, especially at wavelengths 10–20 Å; there are also some strong emission lines in the extreme ultraviolet (EUV) between 170 and 600 Å, not discussed in this article (see SOLAR SPECTROSCOPY: ULTRAVIOLET AND EXTREME ULTRAVIOLET EMISSION). At lower temperatures the spectrum shifts to longer wavelengths, while at higher temperatures it is dominated by shorter-wavelength lines. This is determined by the ionization stages attained by the ions in the coronal gas at each temperature, as described in the main article for this section. In the following we discuss the mechanics of emission specifically with reference to x-ray emission from the solar atmosphere.

Abundances and ionization balance

If we assume local thermodynamic equilibrium, then the x-ray emission from a given region of the solar atmosphere is determined by the composition of the plasma and its temperature. The abundances of all elements heavier than helium in the solar atmosphere are very small relative to hydrogen, in the range 10^{-4}–10^{-6} as a proportion by number. Table 1 gives estimates of abundances for the most important elements (see also SOLAR ABUNDANCES).

Coronal abundances differ from photospheric values for some elements. Known as the first ionization potential (FIP) effect, elements such as C, N and O having FIP $>$ 10 eV are underabundant compared with elements such as Mg, Si and Fe. The difference can amount to nearly an order of magnitude in some cases, and the size of the effect also varies from one place to another in the corona. A discussion of this effect and references to literature on the subject are found in Golub and Pasachoff (1997, ch 3.3).

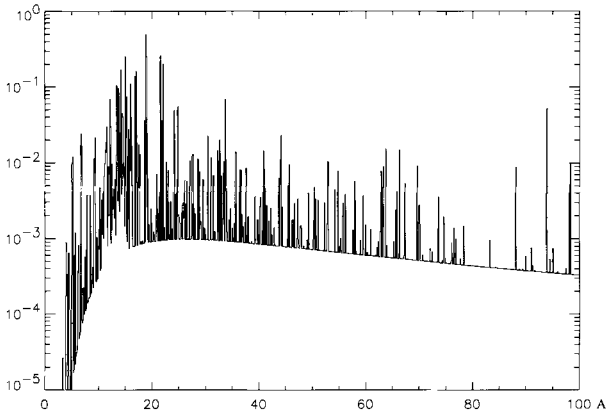


Figure 1. The spectral emissivity of the solar corona for a unit of plasma at $T = 3$ MK. The vertical scale is in units of 10^{-23} erg cm^3 s^{-1} Å^{-1}. Courtesy E DeLuca, SAO.

Table 1. Coronal abundances, normalized to hydrogen.

Element	Coronal abundance
H	1
He	0.10
C	2.3×10^{-4}
N	3.9×10^{-5}
O	2.4×10^{-4}
Ne	3.5×10^{-5}
Na	2.7×10^{-6}
Mg	3.7×10^{-5}
Al	2.7×10^{-6}
Si	3.9×10^{-5}
S	8.6×10^{-6}
Ar	2.1×10^{-6}
Ca	2.9×10^{-6}
Fe	3.9×10^{-5}
Ni	2.2×10^{-6}

The most abundant element, hydrogen, has no strong emission lines in the x-ray region; the strongest line of helium arises from the Lyman-α transition of He^+ at 304 A, in the EUV. At temperatures between 100 000 K and 1 MK the most important contributors to the total power emitted by the solar atmosphere, taking into account the combined emission from all of the relevant ionization stages of each element, are C and O; above 1 MK, the most important contributor is Fe. The fraction, f, of an element that is in a given ionization stage is a function of temperature, determined by the balance between the formation and removal processes: ionization from a lower stage, ionization to a higher stage, recombination from a higher stage and recombination to a lower stage; see main article for further details. A list of the most prominent emission lines in the corona and the temperatures at which they are maximally present is given in table 2.

If the energy balance at a given location in the corona is variable, the abundance of a given ionization stage will change. Timescales for such changes show considerable

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Prominent coronal emission lines.

Line identifier	Wavelength (Å)	$\log(T_{max})$
Fe XVII	15	6.6
Fe XVII	17	6.6
O VIII	19	6.5
O VII	21.6	6.3
C VI	33.7	6.1
Fe IX	171	5.9
Fe X	174.5	6.0
Fe XI	180	6.1
Fe XII	192.4, 193.5	6.2
Fe XIV	211	6.3
Fe XV	284	6.3
He II	304	4.7
Fe XVI	335, 361	6.4
Ne VII	465	5.8
Mg X	610	6.0
Flare lines		
Fe XXVI	1.51, 1.78	8.0
Fe XXV	1.85, 1.86	7.8
Ca XX	3.02	7.7
Ca XIX	3.17, 3.19, 3.21	7.5
Si XIV	6.18	7.1
Si XIII	6.65, 6.68, 6.74	6.9
Mg XII	8.42	6.9
Mg XI	9.17, 9.23, 9.32	6.8
Fe XVII	13.8, 13.9	6.7
Fe XVIII	14.3, 14.4, 14.5	6.8
Fe XVII	15.0, 15.3	6.7
O VIII	19.0	6.5
O VII	21.6, 21.8, 22.1	6.3

Table 3. Variability of the timescale τ for selected coronal ions.

Ion	$\log(T_{max})$	$-\log(f)$[a]	τ (s)
Ne^{6+}	5.7	0.5	1.6
Fe^{8+}	5.9	0.5	0.4
Mg^{9+}	6.1	0.7	14
O^{6+}	6.3	0.3	0.4
O^{7+}	6.4	0.3	103
Fe^{16+}	6.4	0.1	10
Fe^{23+}	7.2	0.5	206

[a] f is the fraction of the element in the given ionization stage at the stated temperature.

differences among the various elements and ionization species. Table 3 gives some examples of the timescale for the appearance of a given ion from the next lower ionization stage, at pressure $N_eT = 3 \times 10^{15}$ cm^{-3} K, in response to a sudden increase in temperature. These values are calculated at the temperature of maximum formation for the listed species. This subject is discussed in Golub and Pasachoff (1997, pp 232–3).

Emitted power

The power emitted by a unit volume of coronal plasma at a given wavelength is $N_eN_pP(T,\lambda)$, where $N_{e,p}$ are the electron and proton number densities, respectively.

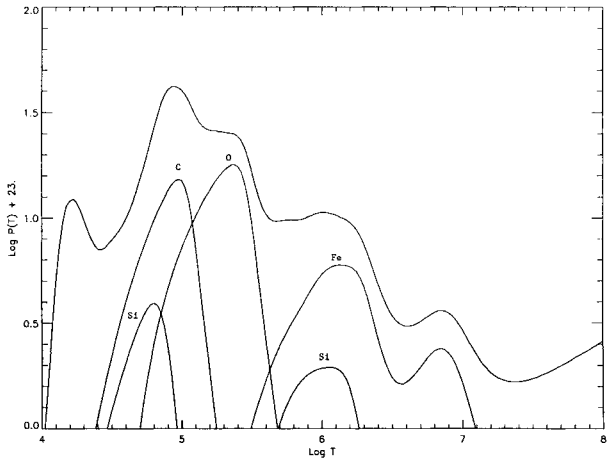


Figure 2. The radiative loss function $P(T)$ for the coronal plasma. Top curve is the total, bottom curves are individual sums for O, C, Fe and Si. Courtesy J Raymond, SAO.

The total radiative loss rate is found by integrating over all wavelengths, $E_R = N_eN_pP(T)$. The quantity $P(T)$ is known as the radiative loss function; it has units erg cm^3 s^{-1}, so that the radiative loss rate E_R has units erg cm^{-3} s^{-1}. At typical coronal temperatures $P(T)$ is of the order of 10^{-23}; a calculation of $P(T)$ from $T = 10^4$ K to $T = 10^8$ K is shown in figure 2.

Bibliography

Golub L and Pasachoff 1997 *The Solar Corona* (Cambridge: Cambridge University Press)
Meyer J-P 1985 *Astrophys. J. Suppl.* **57** 173

Leon Golub

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Atomic Processes

A Greek philosopher called DEMOCRITUS (*c*. 460–370 BC) first introduced the concept of atoms (which means indivisible). His atoms do not precisely correspond to our atoms of today, which are not indivisible, but made up of a nucleus (protons with positive charge and neutrons which have no charge) and orbiting electrons (with negative charge). Indeed, in the solar atmosphere, the temperature is such that atoms are often ionized (to form ions), that is they have lost one electron or two or more. It is the interaction of the atoms and ions with electrons, protons and the radiation field which we refer to as atomic processes. These atomic processes in a PLASMA (ionized gas) determine the intensity or brightness of the radiation which we observe at all wavelengths from the SOLAR ATMOSPHERE. In particular, the transitions give rise to emission and absorption spectral lines.

The birth of quantum mechanics

In 1900, MAX PLANCK presented a derivation of the black-body law in Germany, and the theory of quantum physics was born. In his derivation for the intensity of radiation as a function of wavelength and temperature, Planck set aside classical physics and made an *ad hoc* assumption that light energy, E, was emitted and absorbed in packets (photons) ($E = h\nu$) by oscillators with a natural frequency, ν, where h is a fundamental constant of nature, now known as Planck's constant. Planck himself was not at all happy with his own idea, and he tried for many years to disprove it and understand black-body radiation with classical physics. The radiation from the SOLAR PHOTOSPHERE can be approximated to a black body of temperature just below 6000 K.

ALBERT EINSTEIN, in 1905, went on to explain the photoelectric effect. When light is incident on a metal surface, electrons are ejected. He proposed that the energy in the beam of monochromatic light comes in parcels, $h\nu$. This quantum of energy, a photon, could be transferred completely to the electron. This explanation of the photoelectric effect added substance to Planck's earlier supposition. QUANTUM MECHANICS was born.

Solar and stellar spectral lines

The first SPECTROSCOPE, created by JOSEPH VON FRAUNHOFER in 1814, combined a prism with a small viewing telescope focused on a narrow slit. He used this instrument to view the Sun's spectrum and saw not a continuous spectrum of light, but many, many dark lines. It was later found that any chemical compound, gas or vapor, which emitted light produces its own unique spectrum.

The pattern in emission lines was established by BALMER in 1885, from studies of the hydrogen lines in the spectrum of stars. It was the theory put forward by Bohr in 1913, using Planck's ideas, which accounted for the Balmer analysis of the hydrogen spectrum. Bohr's model for the atom was based on Rutherford's model comprising a positive core, nucleus, surrounded by negatively charged electrons. However, Bohr proposed that electrons could only exist in discrete orbits. In these orbits, they did not radiate energy, but when they jumped from one orbit to another, they emitted a quantum of light. With this hypothesis, everything fell into place.

The FRAUNHOFER LINES are absorption lines from many different elements. The radiation passing through the layers in the solar atmosphere is absorbed by the atoms and ions to excite electrons to higher orbits.

Bright emission lines from PROMINENCES were recorded by Sir JOSEPH NORMAN LOCKYER in 1868. Laboratory experiments were carried out to try and reproduce these lines, but without success. He suggested that the lines were due to an element named *helium* after the Greek Sun god *Helios*. It was 25 years before the existence of helium was confirmed on Earth.

We now know that the Sun is composed mainly of hydrogen, with some helium and traces of many other elements. The spectral line patterns from these atoms and ions can be very complex, reflecting the intricacy of their atomic structure.

Atomic processes

There has been a close association between the study of the Sun and developments in atomic collision physics. The formulation of the so-called *coronal equilibrium* equations over fifty years ago pointed to the importance of excitation, ionization and recombination collisions in determining the state of the plasma and the nature of the spectral emission lines. In earlier *local thermodynamic equilibrium* models, the plasma state was determined by its temperature and the laws of statistical physics.

In the solar atmosphere, the ionization and recombination processes can usually be solved separately from the statistical equilibrium equations for the atomic processes amongst the low lying levels in the ion.

In a hot ($T > 2 \times 10^4$ K) and low electron density ($N_e < 10^{12}$ cm^{-3}) plasma, such as the outer atmosphere of the Sun and stars, it can be assumed that the spectral lines are optically thin. Low down in the atmosphere, it is necessary to solve the RADIATIVE TRANSFER equations.

The notation used is such that Fe XIV (or Fe^{13+}) is the element iron with thirteen electrons stripped off. Taking Fe XIV as an example, the ground configuration $3s^23p$ has two levels—$^2P_{1/2}$ and $^2P_{3/2}$—the transition between these two levels gives rise to the coronal green line at 5303 Å. The transitions between the excited configurations $3s3p^2$ and $3s^23p$ are at around 300–400 Å and those between $3s^23d$ and $3s^23p$ are at shorter wavelengths around 200 Å.

Spectral emission line

An ion in an excited state can spontaneously emit radiation:

$$X_j^{m+} \Longrightarrow X_i^{m+} + h\nu_{i,j} \qquad (1)$$

where an atom X of charge state m, in a bound state (orbit) j emits a photon of energy $\Delta E_{i,j}$ ($= h\nu_{i,j} = hc/\lambda_{i,j}$) to arrive at a lower energy state i.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

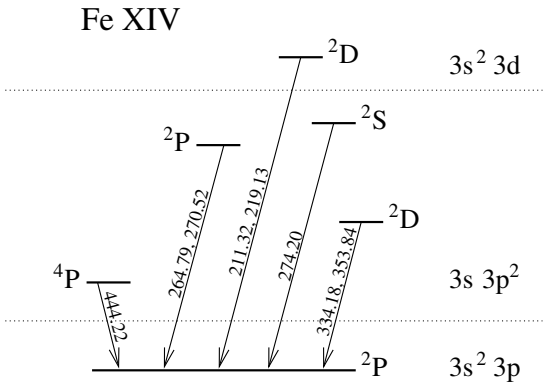


Figure 1. A diagram illustrating the strongest EUV transitions. Wavelengths are given in ångströms.

Ionization and recombination

The degree of ionization of an element is obtained by equating the ionization and recombination rates that relate successive stages of ionization.

$$N^{m+}(q_{\rm col} + q_{\rm au}) = N^{(m+1)+}(\alpha_{\rm r} + \alpha_{\rm d}). \tag{2}$$

The dominant processes in optically thin plasmas are *direct electron impact ionization* ($q_{\rm col}$)and excitation followed by *autoionization*($q_{\rm au}$); *radiative recombination* ($\alpha_{\rm r}$) and *dielectronic recombination*($\alpha_{\rm d}$). The fractional ionization ratio for each ion

$$R_m = N(X^{m+})/N(X) \tag{3}$$

is significant over a small range of temperatures and peaks at a different $T_{\rm max}$ for each ion stage. As the temperature increases, so does the ionization stage, as more and more electrons are stripped off (Arnaud and Raymond, 1992).

Direct electron impact ionization from the inner and outer shells of the ground configuration can be expressed as:

$$X_n^{m+} + {\rm e} \Rightarrow X_{n'}^{(m+1)+} + {\rm e}' + {\rm e}'' \tag{4}$$

with n and n' being the quantum state of the ions.

The process of *radiative recombination* is:

$$X_n^{(m+1)+} + {\rm e} \Rightarrow X_{n'}^{m+} + h\nu. \tag{5}$$

The inverse process is *photoionization*, which is a dominant process for many low-density astrophysical plasmas, but not for the outer atmospheres of the Sun and stars.

For the solar CORONA, the main ionization and recombination processes were originally thought to be direct electron impact ionization and radiative recombination. However, early calculations gave values of $T_{\rm max}$ for the coronal forbidden lines (due to Fe X and Fe XIV) much lower than the temperatures deduced from their spectral line profiles. This discrepancy was eventually resolved by Burgess, who showed the importance of dielectronic recombination, a process which can proceed via doubly excited autoionizing states. The process of *dielectronic recombination* is:

$$X_n^{(m+1)+} + {\rm e} \Rightarrow (X_{n'}^{m+})^{**} \Rightarrow X_{n''}^{m+} + h\nu \tag{6}$$

where an electron is captured by an ion with charge $(m + 1)+$ to form a doubly excited state $(\ )^{**}$ of an ion X with charge $m+$. This ion can then either *autoionize* back again or undergo a spontaneous radiative transition of the inner excited electron to a state below the first ionization limit. Dielectronic recombination is now known to be the dominant recombination mechanism at high temperatures, for example in the solar corona. It is at least a factor of twenty higher than radiative recombination.

The inverse process to dielectronic recombination is *autoionization*:

$$(X_n^{m+})^{**} \Rightarrow X_{n'}^{(m+1)+} + {\rm e}'. \tag{7}$$

The coronal model approximation

The atomic processes which determine the populations of the low-lying levels in an atom or ion in the solar corona are *excitation by electron (e) or proton (H^+) impact*:

$$X_i^{m+} + {\rm e}\ ({\rm H}^+) \Rightarrow X_j^{m+} + {\rm e}'\ ({\rm H}^+) \tag{8}$$

and radiative processes, spontaneous ratiative decay and photoexcitation (the excitation by absorption of photospheric radiation)

For optically allowed (electric dipole transitions) which give rise to spectral lines in the UV, EUV and x-ray wavelength ranges, the *coronal model* approximation is usually valid. The population, $N_j(X^{m+})$, of the upper level, j, is determined by electron collisional excitation from the ground state i and radiative decay from j back down to i.

$$N_i(X^{m+})N_{\rm e}C_{i,j}^{\rm e} = N_j(X^{m+})A_{j,i} \tag{9}$$

where the spontaneous radiative decay rate is $A_{j,i}$ and the electron number density is $N_{\rm e}$. The electron collisional excitation rate coefficient, $C_{i,j}^{\rm e}$, is obtained by integrating the electron impact collision cross section over a Maxwellian electron velocity distribution with a temperature $T_{\rm e}$(K).

For a typical EUV transition, for example from Fe XIV, at coronal densities and temperatures, we find that $A_{j,i}$ is approximately 10^{10} s^{-1}, whereas $N_{\rm e}C_{i,j}^{\rm e}$ is around unity. The population of the upper level j is negligible in comparison with that of the ground level i.

The solution of the electron–ion scattering problem is complex and takes a great deal of computing resources. The accuracy of a particular calculation depends on two main factors. The first is the representation which is used for the target, the atom or ion, and the second is the type of scattering approximation chosen. The main approximations used for electron–ion scattering are *distorted wave* (DW), *Coulomb–Bethe* (CBe) and the more elaborate *close-coupling* (CC) approximation. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

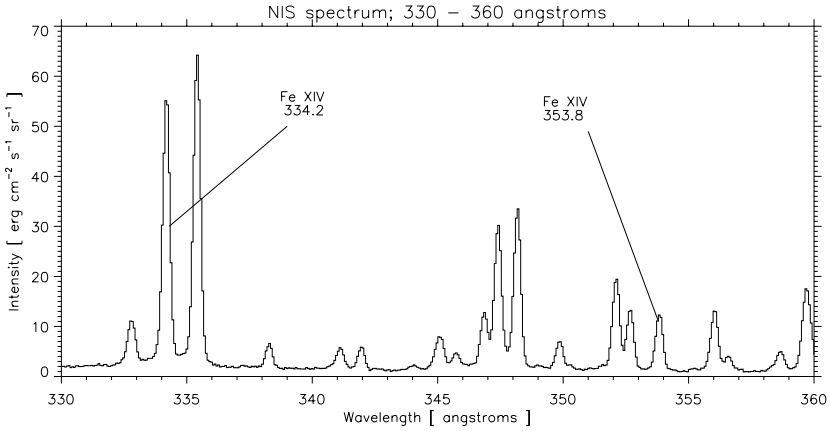


Figure 2. NIS spectrum in the range 330–360 Å taken above an active region on the solar limb. The Fe XIV 334.2 and 353.8 Å lines are clearly seen.

DW approximation neglects the coupling of the channels (target + scattering electron). Since the scattering electron sees a central field potential, the DW approximation is only valid for systems which are a few times ionized. For high partial wave values of the incoming electron, the CBe approximation is valid, when it is assumed that the scattering electron does not penetrate the target. In the CC approximation, the scattering electron sees individual target electrons, the channels are coupled and a set of integro-differential equations are solved. The CC approximation is the most accurate (better than 5%) but it is also the most expensive in terms of computing resources.

The proton collisional excitation and de-excitation rates become comparable with electron collisional processes for transitions where $\Delta E_{i,j} \ll kT_e$, where k is the Boltzmann constant. This happens for transitions between fine structure levels at high temperatures, for example the Fe XIV transition in the ground configuration: $3s^23p$ ($^2P_{1/2}$ - $^2P_{3/2}$).

Solar EUV spectral emission lines

The EUV wavelength range provides a wealth of spectral emission lines from many different elements and ion stages (see SOLAR SPECTROSCOPY: ULTRAVIOLET AND EXTREME ULTRAVIOLET EMISSION). Fe XIV is one of the most important diagnostic ions in the solar corona. It is abundant at a temperature of about 2×10^6 K. Transitions within Fe XIV give rise to spectral lines in the visible (green line, 5303 Å) and extreme ultraviolet (EUV) wavelength ranges. The transitions between the ground configuration ($3s^23p$) and excited configurations ($3s3p^2$, $3s^23d$) give rise to strong lines in the EUV wavelength range. These have been extensively observed with the Coronal Diagnostic Spectrometer on SOHO and can be used to determine electron density in the solar atmosphere. The transitions and observations are illustrated in figures 1 and 2.

New CC atomic calculations have recently been carried out as part of the IRON project for Fe XIV and many other coronal ions. The results provide a significant advance over previous work. In particular, many of the persistent discrepancies between observed and theoretical intensity ratios now seem to have been resolved.

Summary

The story of solar spectroscopy and the corresponding study of atomic processes has been fascinating to follow. The requirement for high accuracy atomic data to interpret astrophysical spectra has provided the stimulus for the development of new techniques in atomic physics. These new atomic data have provided the means to probe and define the physical parameters in the solar atmosphere. The two fields of study have been closely interwoven, with a very fruitful return for each. From recent solar observations, we know that the solar atmosphere is dynamic, constantly changing on all spatial scales. This poses an exciting challenge for atomic physics, to interface the theoretical solar models with atomic processes. The strong possibility exists that the solar plasma is not in equilibrium, that assumptions such as Maxwellian velocity distributions and ionization equilibrium may no longer be tenable.

Bibliography

Arnaud M and Raymond J C 1992 *Astrophys. J.* **398** 39

Fleck B and Svestka Z. (eds) 1997 *The First Results from SOHO* (Amsterdam: Kluwer)

Gabriel A H and Mason H E 1982 Solar physics *Applied Atomic Physics Theory* vol 1, ed H S W Massey and D R Bates (New York: Academic) pp 345–97

Mason H E and Monsignori-Fossi B C 1994 Spectroscopic diagnostics in the VUV for solar and stellar plasmas *Astron. Astrophys. Rev.* **6** 123–79

Vial J-C, Brocchialini K and Boumier P (eds) 1998 *Space Solar Physics* (Berlin: Springer)

Helen Mason

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Coherent Plasma Emission

Radio emission from the Sun is highly variable, and often much brighter than can be explained in terms of thermal radiation from the solar CORONA. The various kinds of non-thermal radio emission are called solar radio bursts.

Frequency ranges

Solar radio bursts are classified first according to the (cyclic) frequency of observation, ν, or wavelength, $\lambda = c/\nu$. The bursts discussed in this section are at decimeter-λ ($\nu \sim 0.3$–3 GHz) and meter-λ ($\nu \lesssim 300$ MHz) or longer. Decimeter-λ bursts are generated low in the corona, meter-λ bursts at heights $\sim 1 R_\odot$ and kilometer-λ bursts (observed from spacecraft) originate in the interplanetary medium (IPM). Emission at higher frequencies in the radio range (mm-λ and microwaves at $\gtrsim 3$ GHz) is discussed in SOLAR FLARES: RADIO BURSTS.

Dynamic spectra

The classification of a radio burst is based primarily on its appearance on a dynamic spectrum, which shows ν versus time t with the intensity of emission represented by a gray scale, cf figure 1. Important characteristics of bursts are their drift rate, $\dot{\nu} = \mathrm{d}\nu/\mathrm{d}t$, their bandwidth $\Delta\nu$ (at fixed t), their duration Δt (at fixed ν), harmonic structure and other fine structures.

Polarization

Solar radio bursts are usually partially circularly polarized. This is interpreted in terms of the birefringence of the coronal plasma, whose ordinary mode (o-mode) and extraordinary mode (x-mode) have their electric vectors rotating in the opposite and same senses, respectively, as electrons gyrate in the magnetic field.

Brightness temperature

The brightness temperature, T_b, is related to the specific intensity, $I(\nu)$ (the power per unit area, per unit solid angle per unit frequency), by $I(\nu) = 2\nu^2 k T_\mathrm{b}/c^2$, where k is Boltzmann's constant. A RADIOTELESCOPE measures the flux density, and this is divided by the relevant solid angle to find $I(\nu)$. A radioheliograph uses interferometry to measure the angular size of sources directly. Emission is non-thermal if T_b exceeds the temperature of the plasma in the source region. For incoherent emission kT_b cannot exceed the energy of the emitting particles. A coherent emission mechanism is required when this limit is exceeded.

Emission mechanisms

Most meter-λ and decimeter-λ solar radio bursts require a coherent emission mechanism and most are interpreted in terms of *plasma emission*, which results in either fundamental (F) emission near the plasma frequency, $\nu_\mathrm{p} = 90(n_\mathrm{e}/10^8\ \mathrm{cm}^{-3})^{1/2}$ MHz, or second harmonic (H) emission at $\nu \approx 2\nu_\mathrm{p}$. Theory predicts that the F component should be highly polarized in the sense of the o-mode, and the H component weakly polarized in the same sense. This contrasts with the x-mode polarization typical of *gyromagnetic emission*, which is due to electrons gyrating in a magnetic field, B. Incoherent gyromagnetic emission by mildly relativistic electrons, called *gyrosynchrotron emission*, is the emission mechanism in solar microwave bursts. Gyromagnetic emission can lead to coherent emission through *electron cyclotron maser emission* (ECME), which occurs when gyromagnetic absorption by non-relativistic particles is negative. ECME causes x-mode radiation to grow at just above the cyclotron frequency, $\nu_B = 2.8 \times 10^{10} B$ Hz, where B is in gauss. Plasma emission is effective in weakly magnetized plasmas ($\nu_B \lesssim \nu_\mathrm{p}$), and ECME requires $\nu_\mathrm{p} \lesssim 0.3\nu_B$. The coherent emission mechanisms are discussed in more detail below.

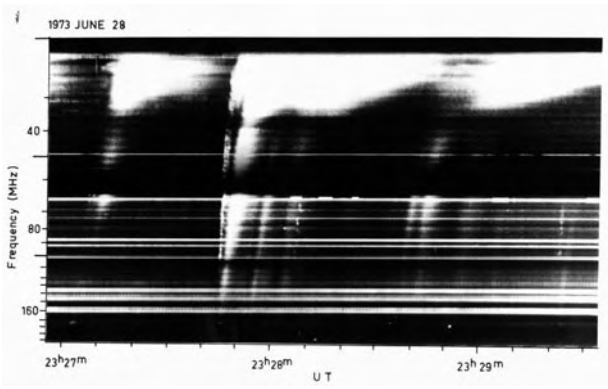


Figure 1. Example of a dynamic spectrum showing type III bursts drifting rapidly from higher to lower frequencies. The horizontal lines are interference.

Meter-λ and decimeter-λ bursts

Meter-λ bursts were originally classified (in 1950) into three types, denoted type I, type II and type III, according to their appearance on a dynamic spectrum. Later (*circa* 1960) this classification was extended to include types IV and V bursts. Detailed investigation of bursts at decimeter-λ started in the late 1970s.

Type III bursts and related bursts

Type III bursts were originally identified due to their rapid drift ($\dot{\nu} \sim -0.01\nu^{1.84}$, with ν in MHz), short duration ($\Delta t \sim 220/\nu$, with ν in MHz) and a relatively broad bandwidth ($\Delta\nu \sim \Delta t\,|\dot{\nu}|$) at a given time due to their high drift rate. A significant fraction of type III bursts exhibit harmonic structure: the F and H components are seen simultaneously with their frequency ratio, ~1:1.8. There is a broad range of T_b for type III bursts in the corona, from the lowest detectable value $\sim 10^6$ K to a maximum value that increases with decreasing frequency from $\sim 10^{11}$ K at ~100 MHz to $\sim 10^{16}$ K at ~30 kHz. Proportionally higher values of T_b are implied if the apparent sizes are scatter images of much smaller actual sources; a strong indication that they are scatter images is that the height

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the corona of an apparent source is determined by its frequency, rather than by whether it is an F or a H component. The bursts at ∼100 MHz are usually weakly circularly polarized (<0.15) in the sense of the o-mode, and occasionally the polarization is moderately high (∼0.5) for the F component. The H component is more weakly polarized in the same sense.

There are various subclasses of type III bursts. Inverted-U bursts turn over at some minimum frequency and then drift back from lower to higher frequency (interpreted as an electron beam propagating along a magnetic loop that closes very high in the corona). J bursts are similar to U bursts without the return stroke. In type V bursts the turnover in frequency is replaced by a long-duration continuum. About 10% of F components of type III bursts are type IIIB bursts, which consist of *stria bursts* (fine structures with double or triple splitting drifting more slowly than a type III burst), with an envelope similar to the profile of a type III burst.

Type III bursts are characteristic of the impulsive phase of solar flares: groups of several to several tens of type III bursts are associated with small to moderate-sized flares. There are also non-flare associated storm type III bursts. The exciting agency for a type III burst is a beam of energetic electrons propagating outward through the corona. These beams often continue through the IPM past the orbit of the Earth ($\nu_p \sim 30$ kHz), and to beyond the orbits of the GIANT PLANETS. The properties of the electron beams can be measured *in situ* at the orbit of the Earth, where a typical electron beam contains $\sim 10^{33}$ electrons over an area ~ 1 AU^2. The fastest electrons ($v \approx 0.6c$) arrive first, after which a power-law spectrum builds up and joins on smoothly to the ambient thermal distribution when the slower electrons arrive. A single interplanetary event, which lasts for ∼1 h, corresponds to many overlapping groups of type III bursts at ∼100 MHz. The mean energy of the electrons in storm type III events is significantly lower, ∼2 keV, than in flare-associated type III events, ∼10 keV.

Type II bursts

Type II bursts are slowly drifting (∼MHz s^{-1}) and narrow-band (a few MHz at ∼100 MHz). They are excited by flare-generated shock waves propagating outward through the corona. Both F and H components may be present, and they have polarization similar to type III bursts. The bursts sometimes exhibit *band splitting* in which the F and/or H component splits in two ($\Delta\nu/\nu \sim 0.1$) or *multiple-lane structures* in which multiple splitting occurs. Some bursts exhibit *herringbone structure* in which type-III-like bursts emanate from a 'backbone' of emission, and propagate either to lower or to higher frequencies. Type II bursts are sometimes seen in the IPM, although most fade from view at $\lesssim 1$ MHz, and most shocks in the IPM are not associated with type II bursts. Type II bursts are statistically associated with coronal transients, but the relation is not one-to-one.

Storms: type I emission

Type I emission is not associated with flares, but occurs in storms that can last for hours to days. Storms can include type I bursts, a type I continuum and type III bursts (type I–III storm). Type I emission is restricted to $\gtrsim 100$ MHz, and storm type III bursts to $\lesssim 100$ MHz; it appears that these originate from, respectively, closed and open magnetic structures in the solar corona, with the type I–III transition indicating the boundary between the closed and open regions. Type I emission is restricted to a few tens of megahertz, suggesting that the source is high in the corona, just below this transition.

Type I emission differs from types II and III emission in that it is strongly (∼100%) circularly polarized, and in that there is only F emission with no H component. Type I bursts have $\Delta\nu \sim 5$ MHz, $\Delta t \sim 0.3$ s at 200 MHz and $T_b \lesssim 10^9$ K. The bursts can drift to either higher or lower frequencies, with most bursts having negligible drift. Sometimes bursts line up on a dynamic spectrum to form a chain which drifts to either higher or lower frequencies at ∼0.3 MHz s^{-1}. Type I sources are directional, and are rarely visible near the solar limb. The type I continuum, which may occur with or without type I bursts, appears to be a distinct phenomenon rather than many unresolved bursts.

Continuum emission: type IV bursts

There are several different classes of radio continuum, that is, broad emission in frequency and time with little structure on a dynamic spectrum (see SOLAR SPECTROSCOPY: CONTINUUM RADIO EMISSION AND DIAGNOSTICS). The name type IV bursts was originally defined to describe continuum emission, but it is appropriate to use more specific descriptive names, such as flare continuum which is directly associated with flares, type I continuum which occurs in storms, and moving type IV bursts which drift to lower frequencies and are interpreted as gyrosynchrotron emission from plasmoids drifting out through the corona.

Decimeter-λ bursts

Type IIIDM bursts are type-III-like bursts at decimeter-λ. They are related to flare-associated meter-λ type III bursts.

Spike bursts are short-duration ($\lesssim 0.1$ s) narrow-band ($\Delta\nu/\nu \lesssim 0.03$), highly polarized bursts which may have very short (∼10 ms) rise or decay times. They can occur in the range ∼300 MHz–8 GHz, and at the highest frequencies can occur in clusters of up to 10^4 bursts. They are strongly correlated with flares, but are observed in only a few per cent of flares. The favored (but still uncertain) interpretation is: (a) that they are due to ECME generated by beams of electrons, and (b) that they represent the basic fragmentation of the energy release in flares.

Theory of solar radio emission

The basic concepts in the theory of the coherent emission mechanisms for solar radio bursts are well established but uncertainties remain in the details of the applications.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

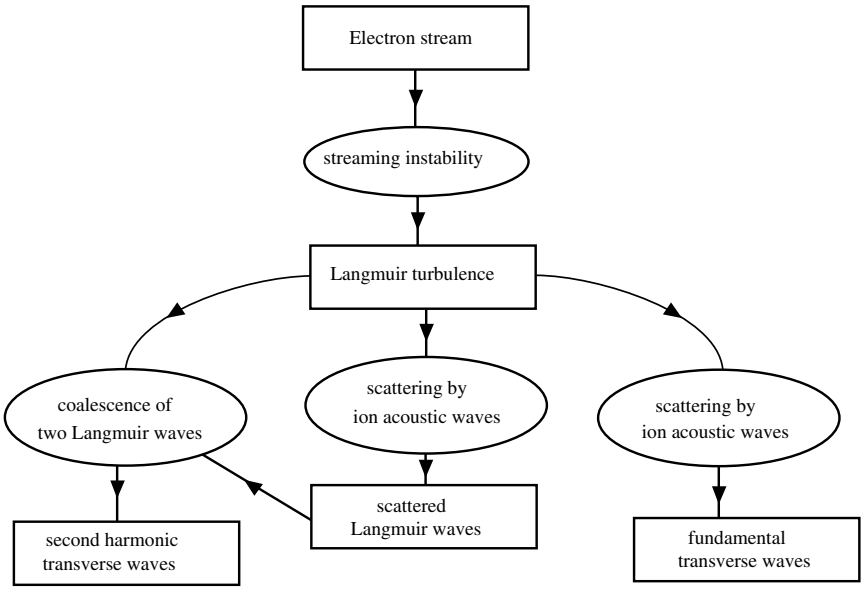


Figure 2. Flow diagram for the generation of plasma emission. In this variant of the theory the nonlinear processes involve ion acoustic waves.

Plasma emission

The first detailed theory for plasma emission was due to Ginzburg and Zheleznyakov in 1958, and although the details of the plasma physics have since been updated, the basic ideas are unchanged, cf figure 2. The first stage involves generation of Langmuir waves, which are longitudinal electron plasma oscillations just above the plasma frequency. These waves grow in a plasma instability, which in the simplest case is the bump-in-tail instability due to a beam of suprathermal electrons with distribution function, $f(v)$, with a bump, $\mathrm{d}f(v)/\mathrm{d}v > 0$, at the beaming speed, v_b. Growth of Langmuir waves is a maximum for waves propagating along the beam direction ($\theta = 0$) with phase speed, $v_\phi \lesssim v_\mathrm{b}$, and growth is effective only for small ranges of θ and v_ϕ. The evolution of the waves and of the particles is described by a pair of quasilinear equations. The back reaction on the distribution of particles, referred to as quasilinear relaxation, involves a smoothing out of the bump to form a plateau, $\mathrm{d}f(v)/\mathrm{d}v \approx 0$. In an inhomogeneous beam model, faster electrons outpace slower electrons tending to increase $\mathrm{d}f(v)/\mathrm{d}v$ near the front of the beam, and quasilinear relaxation drives the distribution function back toward marginal stability where the growth rate ($\propto \mathrm{d}f(v)/\mathrm{d}v$ at $v_\phi = v$) is just large enough to overcome any damping or losses of the waves.

The other stages involve nonlinear processes that result in partial conversion of the energy in Langmuir waves into either F or H transverse waves. The nonlinear processes can be either weak turbulence effects, involving three-wave interactions, or strong turbulence effects. The weak turbulence effects indicated in figure 2 require that low-frequency (e.g. ion-acoustic) waves be present, and these must be generated independently of the Langmuir waves. The three-wave processes saturate when the T_b of the F or H emission reaches approximately the effective temperature of the Langmuir waves, which can be estimated from the theory of the bump-in-tail instability. The required low-frequency waves have wavenumbers approximately equal to those of the Langmuir waves generated in the instability, and the three-wave interactions favor producing either transverse waves, resulting in F emission, or backward Langmuir waves, which can coalesce with the initial Langmuir waves to produce H emission. The evolution of the spectrum of Langmuir waves due to these three-wave interactions is toward higher phase speeds until they collect in a condensate at very high v_ϕ. Three-wave interactions involving even lower frequency waves, which are a common feature in the IPM, tend to scatter the Langmuir waves with little change in v_ϕ, so that they diffuse in θ. A Langmuir wave spends only a small fraction of its time in the small cone of angles about the beam direction where it grows, and a much larger fraction of its time in other directions where it damps. As a consequence, the growth rate has to be quite large to result in net wave growth.

Strong turbulence effects include parametric instabilities that are counterparts of weak-turbulence effects. One example is a decay instability in which the beam-generated Langmuir waves act as a pump that produce daughter ion-acoustic waves and either (backward) Langmuir waves or (F) transverse waves. Another example is a modulational instability which involves a four-wave interaction that tends to drive Langmuir waves to lower v_ϕ until they approach the thermal speed of electrons when Landau

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

damping by thermal electrons acts as an energy sink. An intrinsically strong turbulence effect is Langmuir collapse, in which the Langmuir wave energy collects in localized regions called *cavitons*. A caviton grows due to refraction of Langmuir waves into the underdense region increasing the local energy density in the waves, and the ponderomotive force due to the gradient in this energy density drives plasma out of the caviton, further decreasing its density.

Interpretation of in situ data

Direct confirmation of the theory of plasma emission comes from *in situ* data on type III events in the IPM and from analogous emission from planetary bow shocks. The Langmuir waves observed in association with the type III electrons are highly spatially inhomogeneous, in the form of localized clumps. The data have not proved definitive in deciding between weak and strong turbulence effects, with evidence supporting both. The shape of the distribution function is consistent with the inhomogeneous beam model, and the observed values of T_b are of the same order as those expected for the saturation model. There is direct evidence for the decay instability: ion-acoustic waves with the expected properties are observed, especially in planetary bow shocks. However, there is no compelling evidence for Langmuir collapse: the clumps of Langmuir waves are too large to be cavitons. A statistical distribution of the electric fields in the clumps is consistent with a stochastic growth theory in which an individual burst of growth increases the energy density from W_0 to $W = e^G W_0$, with G a random variable.

Electron cyclotron maser emission (ECME)

ECME results when gyromagnetic absorption by nonrelativistic particles is negative. The most favored case is for a loss-cone instability, in which there is a deficiency of electrons with small pitch angles. A loss cone is present in the reflected electrons when precipitation preferentially removes the electrons with small pitch angles at a footpoint of a magnetic loop. ECME causes x-mode radiation to grow at just above ν_B. A serious problem with application of ECME to solar radio emissions is that it should be reabsorbed by thermal electrons at the second ($\nu = 2\nu_B$) harmonic layer. Suggestions as to how this difficulty might be overcome or avoided include (a) emission at higher harmonics above the second harmonic absorption layer, (b) tunneling through this layer and (c) escape through windows in θ. However, each of these introduces other difficulties, and interpretation of any solar emissions in terms of ECME should be regarded as tentative until it can be shown how the radiation can escape. An interpretation in terms of plasma emission should not be excluded.

Propagation through the solar corona

Local inhomogeneities in the coronal plasma cause local variations in the refractive index $n(\nu) = (1 - \nu_p^2/\nu^2)^{1/2}$ of transverse waves, and hence lead to coronal scattering. This scattering tends to increase the angular size of the source, which opposes the systematic refraction into the radial direction due to the average coronal density gradient. In addition to scattering, there is strong evidence for ducting for some bursts, in which the radiation is guided from its actual source to a much greater height where it is scattered to produce an apparent source. Observations suggest that the observed meter-λ sources are scatter images, that are much higher and larger than the actual sources. Propagation effects also modify the polarization. In particular, although the ~100% o-mode POLARIZATION of many type I bursts confirms the prediction of the theory of plasma emission for plausible values of B in the source region, the much weaker polarization in type III and type II emission, and for type I sources as they approach the solar limb, imply that depolarization occurs. Depolarization is probably due to reflections off overdense structures in the corona converting pure o-mode radiation into a mixture of o-mode and x-mode radiation.

Bibliography

Bastian T S, Benz A O and Gary D E 1998 Radio emission from solar flares *Ann. Rev. Astron. Astrophys.* **36** 131–87

McLean D J and Labrum N R 1985 *Solar Radiophysics* (Cambridge: Cambridge University Press)

Melrose D B 1986 *Plasma Astrophysics* vols I and II (New York: Gordon and Breach)

Melrose D B 1991 Collective plasma radiation processes *Ann. Rev. Astron. Astrophys.* **29** 31–57

D B Melrose

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Continuum Radio Emission and Diagnostics

The Sun is a strong radio source (one of the first objects detected by radio telescopes) and radio observations can provide information on structures throughout the solar atmosphere. Radio techniques allow high-quality images with arcsecond resolution to be achieved, and different radio frequencies provide access to different layers of the solar atmosphere. This section discusses incoherent emission from thermal plasma in the non-flaring solar atmosphere; other relevant material may be found in SOLAR SPECTROSCOPY: COHERENT PLASMA EMISSION and in SOLAR FLARES: RADIO BURSTS.

Emission mechanisms

Three different emission mechanisms are the dominant sources of continuum radio opacity in the non-flaring solar atmosphere, and their contrasting properties allow them to be exploited in different ways. An important feature of radio emission from the solar atmosphere is that the presence of magnetized plasma breaks the degeneracy between the two electromagnetic modes and leads in general to very different opacities for them. In particular, the mode in which the electric vector rotates with the same sense as the gyration motion of an electron about a magnetic field (known as the *extraordinary* or X mode) interacts more strongly with radiating electrons than the mode with the opposite sense of rotation of the electric vector (the *ordinary* or O mode). RADIO TELESCOPES typically make images in the two oppositely circularly polarized electromagnetic modes which correspond to the natural radiation modes of the plasma: differences between the two modes result in polarized emission which reflects the conditions in the radio source and is often a valuable diagnostic.

Free–free or bremsstrahlung emission

This is the name given to the electromagnetic interaction occurring when a free electron collides with a positively charged ion. In the solar atmosphere protons and helium make up the bulk of the ions: at temperatures of 10^4 K and higher, all hydrogen atoms are ionized and the atmosphere is a plasma dominated by free electrons and protons. The opacity of this mechanism is proportional to the product of the electron and ion charge densities and inversely proportional to the square of frequency and to temperature to the power 1.5. It is therefore particularly effective at high densities, low frequencies and low temperatures. This mechanism can produce mildly polarized emission.

Thermal gyroresonance emission

An electron moving in a magnetic field gyrates about the field direction with a characteristic frequency called the electron gyrofrequency, $\nu_B = 2.8 \times 10^6 B$ Hz, where the magnetic field B is measured in gauss (G). The acceleration associated with the gyromotion provides opacity in the radio regime at frequencies which are integer multiples (*harmonics*) of ν_B for magnetic field strengths in the range 10–3000 G, which corresponds well to the range of magnetic fields seen in the solar atmosphere above active regions. Gyroresonance opacity increases strongly with temperature, and is much larger when viewed nearly orthogonal to the magnetic field direction in the source than when viewed parallel to the magnetic field. This is the dominant source of opacity in regions of strong magnetic fields in the CORONA, such as above SUNSPOTS. This mechanism can produce highly polarized emission because of the large difference in the opacities of the X and O modes. Strictly speaking this is a resonant mechanism (i.e. a given radio frequency is associated with a particular value of magnetic field strength), but may be regarded as a source of continuum opacity since B varies continuously throughout the atmosphere.

Gyroresonance opacity drops dramatically as harmonic number increases. In conditions typical of the solar corona (electron densities of order 10^9–10^{10} cm^{-3}, temperatures of order 10^6 K, length scales 10^9 cm), the third (and lower) harmonic of the gyrofrequency will be optically thick in the more strongly interacting polarization (the X mode) as long as the line of sight is not nearly parallel to the magnetic field. Since magnetic field strength generally decreases with height, emission from the first and second harmonic layers originates below the opaque third harmonic layer and is not seen, although in the O mode the third harmonic layer may be optically thin and then the second harmonic layer dominates. Thus a radio image of optically thick gyroresonance emission at a given frequency ν GHz in the polarization corresponding to the X mode can be regarded as a map of the temperature on the layer corresponding to a constant magnetic field strength of $\nu/8.4 \times 10^6$ G.

H^- opacity

As one goes deeper into the solar atmosphere the temperature drops below 10^4 K and the dominant species (H and He) become neutral. The total electron and proton densities can remain high in the chromosphere (temperatures between 5000 and 10 000 K) because the total density increases rapidly with depth while the fractional ionization decreases, but below 5000 K essentially all hydrogen is neutral. Free electrons are still present at these temperatures due to ionization of metals such as sodium, but there are insufficient free protons to provide significant free–free opacity. In this situation a mechanism involving free electrons and neutral H dominates the radio opacity: the electron polarizes an H atom and the interaction between the electron and the dipolar atom provides opacity known (somewhat confusingly) as H^- opacity. From the relevant temperature range it is clear that this mechanism is only important in the lower atmosphere near the temperature minimum, probed by very high frequencies (millimeter wavelengths). Circular polarization in the radiation from this mechanism is expected to be weak.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The appearance of the Sun at different radio frequencies

The Sun can usefully be studied at all radio frequencies, from the ionospheric cutoff of the order of 10 MHz (wavelength $\lambda = 30$ m) up to the submillimeter domain above 300 GHz ($\lambda \leq 1$ mm). Different frequencies penetrate to different depths in the solar atmosphere, depending on local conditions, so that the appearance of the Sun changes dramatically over the radio frequency range.

An important feature of solar radio emission is that every frequency becomes optically thick (i.e. opaque) at some point in the solar atmosphere. Just as we cannot see below the SOLAR PHOTOSPHERE in visible light, in a radio image at a given frequency we cannot see material below the optically thick layer. Radio wavelengths lie in the Rayleigh–Jeans limit, $h\nu \ll kT$, and therefore the radio flux is proportional to the brightness temperature. In the optically thick layer the brightness temperature is equal to the local electron temperature so that (unlike most other diagnostics of the solar atmosphere) radio images can often be interpreted directly as temperature maps. Since temperature in the solar atmosphere generally increases outwards from the temperature minimum (just above the photosphere), it is also possible for a radio image to show both the emission from the low-lying cooler optically-thick layer as well as contributions from hotter optically-thin plasma in the overlying atmosphere.

As noted above, free–free opacity increases rapidly as frequency decreases and it dominates the quiet radio Sun at low frequencies (plasma emission is also prevalent at low frequencies, and is discussed in SOLAR SPECTROSCOPY: COHERENT PLASMA EMISSION). The whole solar corona is optically thick at frequencies below 500 MHz, so that the Sun appears as a large relatively featureless source with a dimension larger than the solar radius, usually elongated in the equatorial direction due to the fact that the solar atmosphere is generally of higher density at the equator than at the poles. The fact that the low-frequency radio Sun had a brightness temperature of the order of 10^6 K was one of the important pieces of evidence which led to the discovery of the solar corona. Propagation effects can make the radio Sun appear to be larger than the actual height of the optically thick surface. Scattering off small-scale turbulence in the outer atmosphere and solar wind (see SOLAR WIND TURBULENCE) can also supply a frequency-dependent lower limit to the spatial resolution which can be achieved in an observation. Since the whole corona is at a temperature of the order of 1–2 million K, there is little contrast between quiet features on the solar disk at low frequencies. As is often the case at low frequencies, the brightest radio feature in the 0.3 GHz image shown in figure 1 (the feature at the right edge of the 0.3 GHz disk) is not due to free–free opacity at all, but rather is a noise storm (also called type I continuum, a form of broadband coherent emission commonly found over very active regions in the frequency range 200–500 MHz).

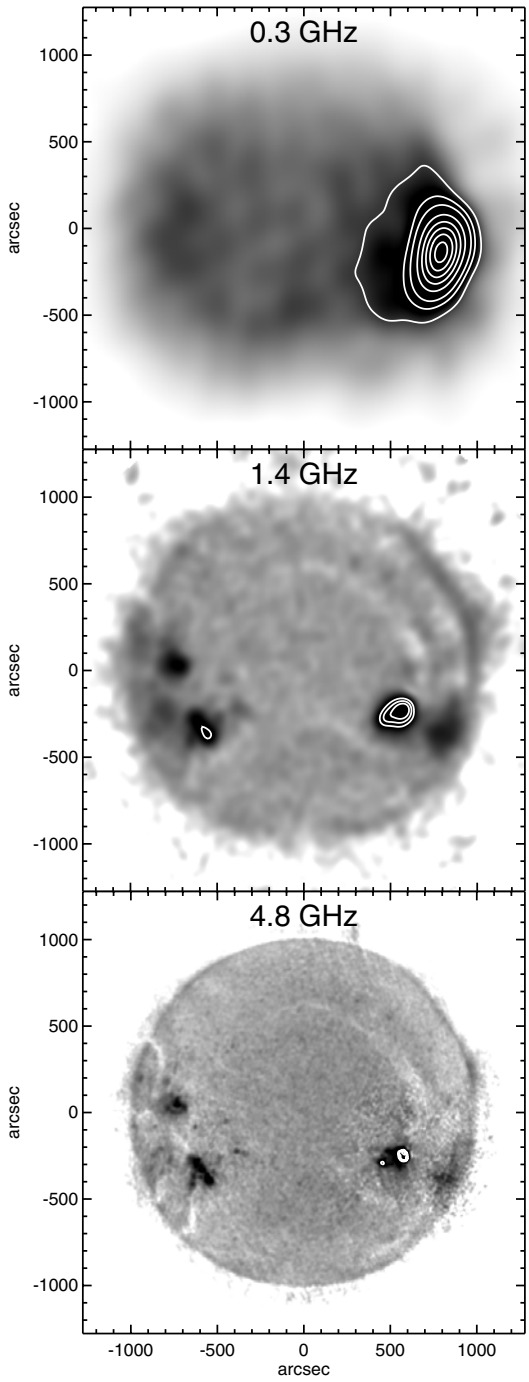


Figure 1. The appearance of the Sun at three different radio frequencies. The intrinsic resolution is different in each image: it is 220″ at 0.3 GHz (top), 50″ at 1.4 GHz (middle) and 12″ at 4.8 GHz (lower panel). These images were made with the Very Large Array radiotelescope on 7 November 1993 (courtesy S White). White contours highlighting the brightest features are plotted at brightness temperatures of 0.8, 1.2, 1.6, 2.4, 3.2, 4.0, 4.8 and 5.6×10^6 K. The color table saturates (is black) at 0.8×10^6 K at 0.3 GHz, 0.6×10^6 K at 1.4 GHz and 0.1×10^6 K at 4.8 GHz. Solar west is to the right.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

As frequency increases above 500 MHz the quiet solar atmosphere starts to become optically thin and its brightness temperature drops, while the enhanced density in the corona above active regions maintains a high optical depth and hence high brightness temperatures. Thus at 1.4 GHz radio images of the Sun show much more contrast between features, depending on their density. The background disk has a temperature of the order of 10^5 K at this frequency. In CORONAL HOLES and FILAMENT channels where the density is low, the brightness temperature is below this value, while in ACTIVE REGIONS and PLAGE the brightness temperature remains above 10^6 K. Virtually all the non-flaring radio emission seen at 1.4 GHz is believed to be produced by free–free emission. Gyroresonance emission is only rarely seen at this frequency, possibly because the gyroresonance-emitting layers lie below optically thick free–free sources.

At 4.8 GHz, on the other hand, gyroresonance emission produces the brightest features observed in radio images. Optically thick free–free emission has never been clearly identified at frequencies this high, so that any feature with a brightness temperature in excess of 10^6 K at frequencies of 4 GHz or higher is almost certainly a gyroresonance source. This is true in the image shown, where two sunspots in an active region southwest of disk center show up as the only features on the disk with coronal brightness temperatures. These are the only two locations in the solar corona on this particular day where the magnetic field strength exceeded the value 600 G needed to produce gyroresonance emission at 4.8 GHz. The other bright features, such as those near the east limb, are due to optically thin free–free emission from dense structures in the corona (these also show up prominently as bright sets of loops in x-ray images), and have brightness temperatures of the order of 10^5 K, compared with 10^6 K over the sunspots. Low-density features clearly show up as regions of reduced brightness, e.g. within the active-region complex near the east limb, and in the filament channel stretching across the northwest quadrant. The disk at this frequency is at a brightness temperature of the order of 20 000 K, corresponding to the lower TRANSITION REGION/upper CHROMOSPHERE. Optically thin coronal emission from features such as active regions appears as an additional contribution to the brightness temperature, but since it is transparent the underlying optically thick chromospheric contribution can still be seen. The size of the radio disk at 4.8 GHz is of the order of 30″ larger than the optical photosphere.

The brightness temperature contribution of optically thin free–free emission scales as ν^{-2}, so that at yet higher frequencies the optically thin contributions from hot dense coronal material diminish rapidly, while the solar disk component, being due to optically thick emission from the chromosphere, diminishes less rapidly: at 17 GHz it is 10 000 K. Gyroresonance emission, on the other hand, can continue to produce features with brightness temperatures in excess of 10^6 K as long as sufficiently strong magnetic fields are present in the corona. The upper limit to coronal magnetic field strengths is not known, but may be of the order of 3000 G (the largest observed field strengths in the photosphere are of the order of 5000 G), which could produce gyroresonance emission up to 25 GHz. This results in contrasts of up to 300 between the brightest radio features and the background 'disk' brightness temperature at frequencies such as 17 GHz. Above the highest frequency at which gyroresonance emission is effective, there is no form of opacity available which can make coronal features optically thick in the quiet solar atmosphere and consequently there are no features with coronal brightness temperatures: the contrast between the disk brightness temperature and the brightest features in the radio image is greatly reduced. The height of the layer in which a given frequency becomes optically thick continues to drop through the chromosphere as frequency increases: the brightness temperature of the disk component is of the order of 6700 K at 100 GHz and 5800 K (corresponding to the vicinity of the temperature minimum) at 300 GHz, with contrasts only of the order of 10% being seen in the images. At these temperatures, free–free emission is probably still the dominant source of opacity, although H^- opacity may also be contributing.

Radio continuum diagnostics

The value of radio diagnostics lies in a number of unique characteristics which complement those available in other wavelength ranges: (i) radio data provide a direct measurement of electron temperature; (ii) again, in contrast to most other diagnostics which are optically thin above the chromosphere and thus show the emission integrated along the line of sight, radio emission is optically thick in the solar atmosphere and by using the fact that different frequencies become optically thick at different heights, multifrequency observations can be used to 'peel away' successive layers of the atmosphere; and (iii) the influence of magnetic fields on radio emission allows coronal magnetic fields to be measured directly. At present radio observations are the best technique available for studying magnetic fields in the corona.

Free–free emission can be used to study densities and temperatures in the solar corona. Radio observations of free–free emission complement x-ray and EUV observations of the corona well: both are sensitive to the dense loops found in the corona above active regions, but the radio data are often optically thick whereas the x-ray/EUV data are optically thin. The fact that the radio emission is optically thick makes it particularly sensitive to temperature gradients. The radio emission is weighted towards cooler plasma due to the $(\text{temperature})^{-1.5}$ dependence of free–free opacity. Information on abundances can be derived from a comparison of the different wavelength regimes, since x-ray fluxes depend on the number densities of ions such as Fe in specific charge states, while the radio flux depends on the number density of protons.

Observations of gyroresonance emission have wide usage in the study of coronal properties above active

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

regions, where magnetic fields are strong enough for gyroresonance emission to dominate. The fact that it is a resonant mechanism results in emission at a given frequency coming from a very thin layer of constant magnetic field strength: typically B varies by less than 2% across a gyroresonant layer. The optical depth may vary by a factor of 1000 from one harmonic layer to the next, which simplifies interpretation since it means that only one layer is likely to be contributing to the observed flux. Thus observations of gyroresonance emission pick out physical conditions in a highly localized region of the corona; changing the observing frequency picks out a different layer, so that in principle a three-dimensional picture of the corona may be developed. There are some complications, however, in interpreting gyroresonance emission: (i) the highest optically-thick harmonic layer is not known in advance: for typical conditions it occurs where $\nu = 3\nu_B$ in the X mode and $\nu = 2\nu_B$ in the O mode, but where the line of sight is parallel to the magnetic field the opacity is greatly reduced and one sees down to a lower harmonic layer; and (ii) the radio data do not contain any independent information on spatial scales, so that one does not know what the height of each gyroresonance layer is from the radio data alone.

Solar radio telescopes

The Sun was first identified as a radio source during World War II. In 1942 J S Hey was investigating bursty interference in meter-wavelength radar receivers in Britain when he noticed that it coincided with a solar flare. Almost simultaneously G C Southworth in the U.S. looked for and detected steady emission from the Sun at centimeter wavelengths. After the war years many veterans of the radar effort turned to RADIO ASTRONOMY. A group in Australia led by Joe Pawsey developed radio instruments for the study of the Sun and carried out pioneering work on the classification of solar radio bursts.

A number of instruments are routinely used to make radio images of the Sun at the time of writing. Solar-dedicated telescopes include the Nançay radioheliograph in France operating at frequencies in the range 150–450 MHz, the Owens Valley frequency-agile array in the USA in the range 1–18 GHz, and the Nobeyama radioheliograph in Japan at 17 and 34 GHz. In Russia the RATAN-600 telescope observes in the range 1–20 GHz and the Siberian Solar Radio Telescope at 5 GHz. Low-frequency facilities previously operated at Culgoora (40–320 MHz) in Australia and at Clark Lake (20–110 MHz) in the USA. The Very Large Array radiotelescope in the USA (0.3–22 GHz) is used for solar observing on an occasional basis, and provides a powerful combination of sensitivity, spatial resolution and frequency coverage.

Bibliography

Bastian T S, Ewell M W Jr and Zirin H 1993 The center-to-limb brightness variation of the Sun at λ = 850 microns *Astrophys. J.* **415** 364

Dulk G A 1985 Radio emission from the Sun and stars *Ann. Rev. Astron. Astrophys.* **23** 169

Kundu M R 1965 *Solar Radio Astronomy* (New York: Interscience)

White S M and Kundu M R 1997 Radio observations of gyroresonance emission from coronal magnetic fields *Solar Phys.* **174** 31

Zirin H, Baumert B M and Hurford G J 1991 The microwave brightness temperature spectrum of the quiet sun *Astrophys. J.* **370** 779

Stephen White

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Infrared Emission

The solar infrared spectrum extends from 1 μm to 1 mm, nestled between the well-studied visible and radio bands. The infrared is rich in spectral proxies of temperature, velocity and magnetic field. These range from high-excitation He I 1.083 μm, to low-excitation molecules like carbon monoxide, hydroxyl, and even water vapor (in sunspots), to magnetically sensitive atomic iron absorptions near 1.5 μm, and to the curious high-Rydberg emissions of Mg I near 12 μm. Local thermodynamic equilibrium (LTE) applies to many of the line formation processes; infrared features are generally better tracers of physical conditions than the scattering lines of the visible and ultraviolet intervals. Seeing (distortions of the telescope images by turbulence in the Earth's atmosphere) also tends to improve toward longer wavelengths. However, even the world's largest optical solar telescope—the 1.5 m McMath–Pierce on Kitt Peak—is limited by diffraction in and beyond the key 5 μm region; atmospheric absorption blocks large stretches of the thermal IR; and infrared sensor technology lags behind its counterparts in the visible and ultraviolet. Nevertheless, the fast pace of discoveries makes IR solar physics one of the exciting areas of modern astronomy.

Background

The infrared generally is considered to extend from 1 μm on the short-wavelength side, to perhaps 1 mm at the longer wavelengths. The lower boundary is about where optical CCD camera technology gives way to specialized infrared sensors, and the long-wavelength side corresponds to the highest frequencies of the radio band. Both of these boundaries are somewhat vague, particularly the upper side where sub-mm radio technology is making serious inroads. Indeed, much of the region between 20 μm and 1 mm is blocked by the Earth's atmosphere, and the Sun's disk intensity distribution in that interval appears to be a featureless continuum; more akin to the radio regime than to the highly structured ABSORPTION SPECTRUM seen at shorter wavelengths.

Aside from the 1.083 μm triplet of atomic helium, there are no prominent high-excitation disk features in the spectrum longward of 1 μm. Instead, the infrared is dominated by low-excitation species: neutral atoms and molecules.

By their nature, the low-excitation species must arise in relatively quiescent gas that is not strongly heated (by the processes that cause, for example, the chromospheric temperature inversion). Such conditions are found in the outer SOLAR PHOTOSPHERE in the supergranulation cell interiors (see SOLAR PHOTOSPHERE: SUPERGRANULATION), in the interstices between the magnetic network elements (see CHROMOSPHERE: NETWORK), and in the hearts of SUNSPOT UMBRAE (cold because intense vertical magnetic fields suppress the convective heat flux). One of the key low-excitation species—carbon monoxide—has given birth to the current dominant controversy of infrared solar physics.

On the one hand, a stigmatic spectrogram of the violet Ca II K line (0.393 μm) recorded over a quiet area on the solar disk shows unmistakable core emission—classic signature of the chromospheric temperature rise—over much of the slit length. On the other hand, extreme-limb observations of the $\Delta v = 1$ fundamental vibration–rotation bands of CO near 4.7 μm reveal unexpectedly cool material ($T < 4000$ K) at high altitudes where one traditionally would place the hot chromosphere ($T \sim 7000$ K). That behavior is difficult to reconcile with the best available one-dimensional models of the solar chromosphere, and forces one to consider a highly inhomogeneous description of those 'layers'.

Infrared diagnostics

Several general characteristics of infrared diagnostics will be mentioned, followed by more detailed descriptions of specific cases. Figure 1 provides examples of some of the spectral regions that will be described later, from the quiet Sun as well as a mature sunspot umbra.

General characteristics

Low-excitation species that form close to LTE. Aside from the true 'chromospheric' tracer He I λ1.083 μm, most of the important spectral features longward of 1 μm are sensitive to *cool* gas in the stellar photosphere. Many of the infrared features form quite close to local thermodynamic equilibrium, particularly the molecules (a very desirable attribute from a diagnostic point of view: in LTE, the brightness temperature of the core of an absorption line corresponds closely to the kinetic temperature in the layers in which the line becomes optically thick).

Furthermore, between 1 μm and 1 mm the continuous opacity is dominated by H^- in the photosphere, and by $p–e^-$ collisions ('H free–free' (f–f)) at higher temperatures in the chromosphere. The negative hydrogen ion has a lone bound–free (b–f) edge at 1.64 μm. There, the background opacity of the solar atmosphere reaches its minimum value, and $\tau_C = 1$ penetrates about 20 km deeper than at visible wavelengths. Longward of the H^- b–f edge, $\tau_C = 1$ ascends rapidly in altitude owing to the λ^2 dependence of the f–f cross section. Beyond about 300 μm, the continuum arises at chromospheric altitudes. For all intents and purposes, the infrared continuous opacities are completely in LTE.

Linear temperature averaging. An often-cited benefit of the spectrum beyond 1 μm is the linear temperature averaging property of the infrared Planck function. In other words, the amount of thermal emission is directly proportional to the local temperature. That means the average intensity of a scene containing a jumble of discrete structures should be an evenly weighted mean of the surface temperatures of those features. In the ultraviolet, on the other hand, the thermal emission is *exponentially* sensitive to temperature ('Wien limit' of Planck function). In that situation,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

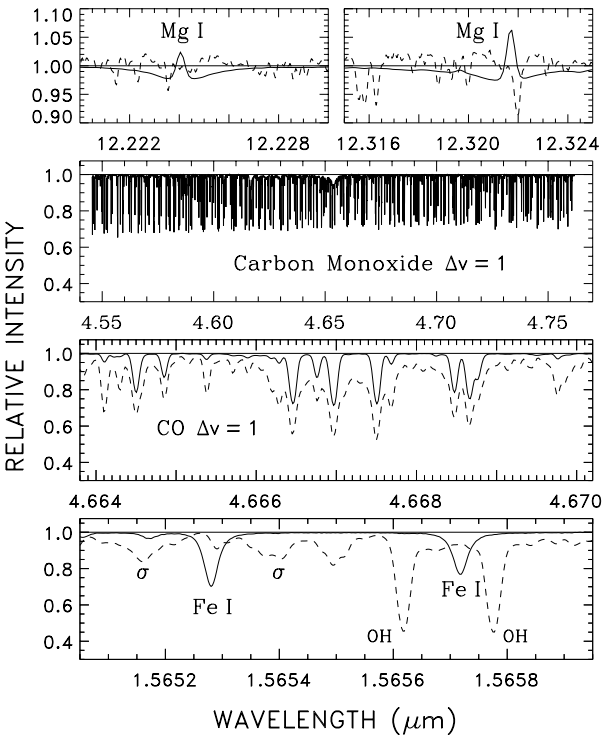


Figure 1. Selected wavelength intervals from the IR spectrum of the quiet photosphere (solid curves), and a sunspot (dashed). The sunspot shows generally greater spectral complexity than the quiet photosphere, owing to enhancement of low-excitation species in the cold umbra, and to Zeeman splitting (the 'σ' components of Fe I λ1.565 in the bottom panel) of magnetically sensitive lines in the multi-kilogauss fields. (Courtesy National Solar Observatory Digital Archive.)

the average intensity of a scene can be completely dominated by a single hot spot, and the deduced average temperature could be quite misleading. Of course, the best approach (although technically challenging) is to view a region simultaneously in the ultraviolet and infrared, to utilize the complementary temperature responses to better constrain the mix of thermal structures present. The full benefit of the linear temperature averaging is achieved for diagnostics that form close to LTE, like the CO $\Delta v = 1$ bands or the f–f continuum. However, one must be cautious in practical applications. In particular, one views the thermal emission at the 'surface' where the gas becomes optically thick (opaque). Where that surface occurs, and thus what temperature is sampled, depends on the opacity of the medium at the wavelength of observation. Even though the infrared thermal emissivity depends linearly on temperature, there is no guarantee that the opacity does as well. For example, the f–f optical depth depends on the electron and proton densities, which vary *exponentially* with temperature over the narrow range 5000–8000 K owing to the ionization of atomic hydrogen. Thus, in a jumble of structures, the $\tau_C \approx 1$ surfaces could occur at a wide range of physical heights and temperatures: the hotter structures become opaque high up, while the colder ones are more transparent, allowing one to view conditions deeper in. The temperature averaging might well be linear, but it could be an uncontrolled average over a wide range of heights, and thus of little diagnostic value.

Magnetic leverage. Magnetic fields are pivotal in shaping the discrete thermal features of the SOLAR PHOTOSPHERE, CHROMOSPHERE, and CORONA, and are probably responsible for a large proportion of the mechanical heating, particularly at high altitudes. The central role of magnetic fields in the solar atmosphere is probably replayed among all late-type convective stars in the cool half of the HERTZSPRUNG–RUSSELL DIAGRAM. The Zeeman splitting of a spectral line in the presence of a magnetic field goes as

$$\Delta\lambda_B \sim g_L B \lambda^2$$

where g_L is a factor that depends on properties of the radiating atom, B is the strength of the longitudinal component of the surface magnetic field, and λ is the wavelength of the spectral line. The intrinsic, non-magnetic width of the line usually is controlled by random thermal and turbulent motions in the gas, and scales as

$$\Delta\lambda_L \sim (v/c)\lambda$$

where v is the velocity dispersion and c is the speed of light. Thus, the magnetic splitting becomes more prominent with respect to the non-magnetic line width as one moves to longer wavelengths, and correspondingly easier to measure. Accordingly, the infrared is well-suited for magnetic investigations, particularly of intrinsically weak fields. Furthermore, cool surface features like sunspot umbrae show much less intensity contrast against the surrounding hotter photosphere in the infrared than in the visible (again, due to linear temperature averaging at the longer wavelengths, versus exponential at the shorter). Consequently, dilution of the umbral light by scattered photospheric radiation is much less of a problem. Thus, the infrared is a good place to measure fine structure in the strong fields of sunspots.

Specific examples

He I λ1.083 μm. The He I λ1.083 μm triplet is the most prominent high-excitation feature that falls in the infrared, and one of the most widely used in ground-based monitoring of chromospheric and coronal structure. The importance of the He I absorption is in how it forms. The high-lying metastable state in the He I term system is populated primarily by recombination of He^+ ions. The latter are produced by photoionization of chromospheric atomic helium by coronal XUV radiation that penetrates into the lower atmosphere. The absorption of the outward streaming photospheric continuum by the highly excited chromospheric He I strengthens in areas that are strongly illuminated by the hard coronal radiation and weakens in areas where the coronal back-radiation is less. High

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

contrast x-ray features like CORONAL HOLES (which play a central role in the SOLAR WIND and are thus prime targets for SPACE WEATHER PREDICTIONS) leave an imprint on He I λ1.083 absorption maps, and can be identified routinely from the ground. He I also is used in studies of SOLAR FLARES and SOLAR CORONAL MASS EJECTION. High-resolution spectroscopy yields the total absorption strength, the line depth and the Doppler shift. Relating these empirical quantities to the dynamics, heating, and XUV back-illumination of the upper chromosphere requires sophisticated numerical simulations.

Other high-excitation lines. A few features formed under truly coronal conditions (temperatures in excess of a million K) are known in the infrared; mostly forbidden ground-state fine structure transitions in coronal ions, such as a pair of Fe XII lines near 1.08 μm and Si X λ1.43 μm. Although the lines are quite faint, and must be recorded off-limb away from the glare of the photosphere, they have diagnostic potential for coronal densities and perhaps magnetic fields. The infrared coronal spectrum is still largely unexplored at high sensitivity, however, owing to the lack of a suitable large-aperture coronagraphic facility, although the McMath–Pierce telescope undoubtedly will see more use in this area over the coming years.

The opacity minimum. The H^- opacity minimum at 1.63 μm allows one to 'see' deeper into the solar atmosphere than at any other wavelength. By comparing infrared opacity minimum pictures with simultaneous, co-spatial filtergrams taken at shorter continuum wavelengths and in Ca K, one can follow the vertical divergence, for example, of hot gas entrained in flux tube fields from the deep photosphere out to the chromosphere itself (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES). The recent availability of large-format high-sensitivity infrared cameras has made such experiments practical. The diffraction limit of the McMath–Pierce at 1.6 μm is about 0.3″, comparable to the *optical* resolution of the (smaller) windowed evacuated solar telescopes typically used for visible imaging. (The vacuum telescopes are often suitable for 1.6 μm work, but with a correspondingly poorer diffraction limit.)

Magnetic field measurements beyond 1 μm. Solar infrared magnetography was pioneered by J Harvey and D N B Hall in the mid-1970s. Hall's 1974 *Atlas of the Infrared Sunspot Spectrum* contains many low-excitation absorption lines that are completely split in a ∼2 kG longitudinal field: Ti I λ2.23 μm (a^5P_1–$z^5D^o_0$) is the premier example. Solanki, Biémont and Mürset have published a list of interesting lines in the interval 1.5–1.8 μm, including many that exhibit large Stokes V amplitudes in the quiet network. A popular target of recent magnetic studies has been the simple Zeeman triplet Fe I λ1.565 μm. The high-Rydberg transitions of Mg I at 12 μm (see below) also have substantial magnetic sensitivity, although the observations are challenging and the spatial resolution is limited by diffraction. POLARIMETRY in the He I λ1.083 μm line allows magnetic fields in the chromospheric layers to be mapped.

Molecules. A wide variety of diatomic molecules, including many isotopic variants, are found in the infrared spectrum longward of 1 μm. The only triatomic species known is water vapor, and it is found exclusively in the deep high-density layers of cold sunspot umbrae (the formation of polyatomic species generally requires high pressures, in addition to cool temperatures). Many of the abundant diatomics have their principal vibration–rotation bands (arising within the electronic ground state of the molecule) in the 1–15 μm interval.

The low-energy rotationally split vibrational transitions are collisionally excited (and quenched) by the abundant moderate-speed hydrogen atoms, and the 'collisionally-controlled' lines thus form quite close to LTE. Typical atoms and ions, on the other hand, have higher-energy visible/ultraviolet transitions that are collisionally coupled mostly to the much rarer high-speed electrons in the gas. For such lines, photon scattering delocalizes the radiation formation, and degrades the diagnostic potential.

Another advantage of the molecular bands is that they typically contain thousands of very similar lines. The redundancy allows statistical techniques to be applied, and is valuable in the infrared where the spectrum is fragmented by molecular absorptions in the Earth's atmosphere (see §58 of C W Allen's *Astrophysical Quantities*). Furthermore, the populations in the vibrational ladder of a typical molecule are very sensitive to temperature, and the molecular spectra are thus nearly ideal 'thermometers', at least for the cooler zones of the solar atmosphere.

The infrared contains strong absorption bands from CO, CH, OH and NH. Of these, carbon monoxide completely dominates the spectral character of the 1–10 μm region, contributing about half of the strong lines of solar origin. CO is the most abundant molecule at the relatively warm temperatures of the high solar photosphere ($T < 5000$ K), owing to its large dissociation potential (11 eV) and the fact that its photoionization and photodissociation thresholds fall shortward of H I Lyα (121 nm), the strongest emission feature of the ultraviolet chromospheric spectrum. Molecular hydrogen also is quite abundant, but is a homonuclear species and thus lacks the significant dipole moment required for a (permitted) vibration–rotation spectrum.

Near limb observations (for which the extreme slant angle causes the $\tau \sim 1$ emitting surface to move upward to high altitudes) of the $\Delta v = 1$ fundamental bands of CO at 4.7 μm show surprisingly cool gas, with temperatures as low as ∼3700 K, at heights where one ordinarily would place the chromosphere (with its ∼7000 K temperatures, too hot for molecules to form). The existence of CO at high altitudes is completely at odds with the best available one-dimensional reference models (e.g., of the 'Harvard school'). The dichotomy—hot chromosphere seen in Ca II K and cold 'comosphere' deduced from the CO infrared bands—highlights the present uncertainties

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

concerning the crucial interface where the radiation-dominated photosphere gives way to the mechanically heated chromosphere.

Further discussion of the CO fundamental bands can be found in the entry on CHROMOSPHERE: THERMAL BIFURCATION.

The 12 μm High-Rydberg emission lines. In the early 1980s, F Murcray and his collaborators, and independently J Brault and R W Noyes, called attention to several curious emission features in the 12 μm region of the solar spectrum. These quickly were identified as high-Rydberg transitions of magnesium and aluminum. Mg 7h–6g (12.22 μm) and Mg 7i–6h (12.33 μm), in particular, exhibit striking Zeeman sensitivity.

Subsequently, it was found that Mg I λ12.33 μm showed a large velocity response to the photospheric p-mode oscillations at disk center, with a peak-to-peak amplitude of more than 1000 m s^{-1}, comparable with what is seen in strong CO $\Delta v = 1$ lines. Curiously, however, intensity oscillations like those in CO were lacking. The velocity response pegged the line formation altitude in the high photosphere (rather than the chromosphere where traditional 'emission' lines arise), and the lack of intensity oscillations pointed to a non-LTE formation mechanism.

The latter has been clarified recently by a number of authors, who have demonstrated that population inversions can happen naturally in the high-Rydberg systems of species for whom the next ionization stage up is the most abundant under the conditions of the upper photosphere (e.g. Mg^+, here). The uppermost levels of the neutral atom near the ionization limit are strongly tied by collisions to the (fixed) ion ground state population, but levels further down the ladder are progressively more coupled to the lowest-lying states of the atom by radiative transitions; the resulting 'photon suction' depletes the lower levels of the high-Rydberg transitions more than the upper levels, producing a population inversion that in turn drives the line cores into emission, without the usual agency of a chromospheric temperature rise.

The large Zeeman sensitivity of the 12 μm lines makes them appealing for magnetic studies, although substantial improvements in sensor technology and telescope aperture will be required to fully exploit their diagnostic promise.

Far-infrared continuum. Sub-millimeter continua provide a unique view of the solar chromosphere at a wide range of altitudes, complementary to the more familiar visible tracers like Ca K and Hα. 850 μm maps of the solar chromosphere have been obtained by raster-scanning the 15 m James Clerk Maxwell Telescope (JCMT) on Mauna Kea across an ACTIVE REGION. The resulting sub-millimeter image closely resembles Ca K pictures: the dense hot PLAGE surrounding the spot groups appears strongly enhanced in sub-millimeter brightness temperature. Counterintuitively, the sunspots themselves also are bright, with intensities comparable with their surroundings. Thus, while the heat flux through the lower atmosphere of the sunspot is strongly suppressed, producing the dark umbra at visible wavelengths, the outer atmosphere apparently receives a normal dose of mechanical heating, leading to more uniform chromosphere at the high altitudes where the 850 μm radiation arises.

As with the 12 μm Mg I lines, significant advances in instrumentation will be required to fully utilize the 150–850 μm continuum region (the lower end of which also is mostly blocked by terrestrial absorption and must be recorded from high-altitude balloons). The most pressing need is for higher angular resolution: the JCMT can achieve a diffraction-limited beam of only about 15″.

Summary

The infrared spectrum contains valuable diagnostics of physical conditions in the solar photosphere, chromosphere and corona. Infrared solar studies have a long history, although the pace of discovery has speeded up considerably in the 1990s with the development of sensitive large format cameras for the 1–5 μm region. Detection of ultra-cool gas in the hot chromosphere, traced by carbon monoxide, remains the major controversy that the infrared has brought to present-day solar physics. The infrared undoubtedly will yield additional puzzles, and perhaps new contradictions, as it is further explored; and exploited as a complement to traditional visible and ultraviolet remote sensing of solar atmospheric structure. Nevertheless, a major impediment is the lack of a large-aperture infrared-optimized solar telescope to achieve the high spatial resolution that is mandatory for modern investigations of our nearby star.

Bibliography

Kuhn J R and Penn M J (ed) 1995 *Infrared Tools for Solar Astrophysics: What's Next?* (Singapore: World Scientific)

Rabin D M, Jefferies J T and Lindsey C (ed) 1994 *Infrared Solar Physics* (Dordrecht: Kluwer)

T R Ayres

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Ultraviolet and Extreme Ultraviolet Emission

Vacuum ultraviolet (VUV) emission is defined as that electromagnetic radiation with wavelengths shorter than 2000 Å. Its name comes from the fact that light shorter than 2000 Å is strongly absorbed by most gases and, in particular, the atmosphere. VUV is usually divided into far ultraviolet (FUV) extending from 1200 to 2000 Å and the extreme ultraviolet (EUV) in the range 300 to 1200 Å. The spectral features of these wavelengths will be discussed with the focus on the spectral emission lines. Continuum radiation will only be discussed briefly. One can ask why is it interesting to observe the VUV emission from the Sun? There are several important reasons for this as it provides important information on:

- physical properties of the solar and stellar atmospheres
- fundamental mechanisms responsible for heating the corona and accelerating the solar wind
- how VUV variability alters the dynamics and chemistry of the Earth's upper atmosphere
- possible influence on the Earth's climate due to variations in the VUV irradiance from the Sun.

In this article the main focus will be on the first two items while the latter two are discussed in more detail elsewhere (see EARTH'S ATMOSPHERE, SOLAR IRRADIANCE, SOLAR–TERRESTRIAL CONNECTION: LONG-TERM AND SHORT-TERM CLIMATE VARIABILITY.

Historical background of VUV observations

The solar ultraviolet (FUV) and extreme ultraviolet (EUV) emission contain a number of strong emission lines and continua well suited for quantitative plasma diagnostics of the solar atmosphere. Our entire knowledge of temperatures, densities, emission measures, mass motions and elemental abundances comes from high-resolution spectral observations in the vacuum ultraviolet.

The VUV wavelength range is absorbed by oxygen and ozone in the Earth's atmosphere and hence is totally inaccessible to even the largest ground-based telescopes. Below 2000 Å the radiation dissociates molecular oxygen in the upper atmosphere, indirectly resulting in the formation of ozone (O_3). To observe the Sun, stars and other celestial objects in this wavelength range, the instruments must be carried above the absorbing atmosphere by means of sounding rockets, Earth satellites or space vehicles. Space astronomy started in 1946 in the United States when captured V-2 rockets became available to navy scientists for use as free-flying, high-altitude observatories. It was not surprising, under these circumstances, that a group from the US Naval Research Laboratory (NRL) in Washington, DC, became the first to observe the Sun's ultraviolet radiation with a spectrograph mounted on the tail fin of a V-2 rocket. Since then much progress has been made in solar high-resolution spectroscopy from space (see SPACE INSTRUMENTATION, SPECTROGRAPHS: HIGH-RESOLUTION SPECTROGRAPHS). A large number of experiments have been launched to observe the EUV and UV portions of the solar spectrum.

Measurements of the solar UV spectral energy distribution are often divided into two groups. Some instruments use only a spectrometer and observe the spectral irradiance of the integrated solar disk. The irradiance measurements are designed to study the Sun's radiant energy output and its variability, and thus require very high photometric accuracy and instrument stability. Other instruments employ an imaging device followed by the spectrometer and observe the radiance from a limited area of the solar disk. The primary goal with the radiance measurements is to differentiate between various emitting regions on the Sun. However, it is much more difficult to achieve high photometric accuracy because of the more complex optics and the often larger number of reflecting surfaces.

Calibration and instrument degradation

Quantitative measurements of the properties of the solar gas should be carried out from intensity calibrated spectra. Even without an absolute calibration one can extract useful information from the spectral data such as the shapes of spectral lines, their wavelength positions as well as line ratios. However, to determine how many photons actually entered the instrument from the Sun an absolute calibration needs to be applied where the reflectivity of each optical element of the instrument needs to be accounted for. The fact that the reflectivity of the coatings used in UV/EUV instruments often degrades with time complicates the calibration effort.

Ultraviolet intensity measurements are particularly difficult because the very solar radiation to be measured with high precision is itself the main cause of instrument degradation. The effect is attributed to contamination by organic material, which is outgassing from materials used in the instrument design and is subsequently photoactivated and deposited on irradiated surfaces. The process, although now well understood, is very difficult to prevent in complex optical instruments for space flight. Comprehensive cleanliness control programs during integration of the instruments are used to reduce degradation. Most EUV/UV instruments still suffer from some sensitivity losses during their operation. Thus, intercalibration between different instruments, both on satellites and rocket payloads, is being used to determine the absolute level of emission from single features on the Sun.

VUV emission lines

The high-temperature solar atmosphere is the only astrophysical plasma source that can be studied with high spatial resolution. Much of our understanding of STELLAR ATMOSPHERES is based on the understanding of plasma processes that occur in the upper solar atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

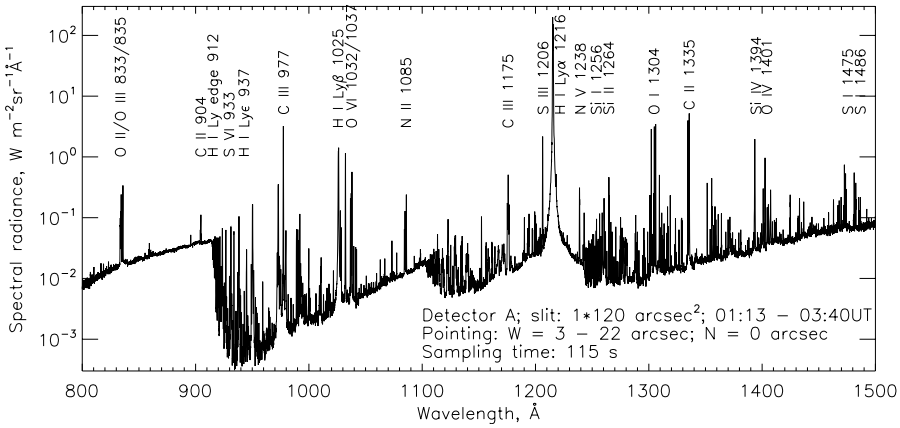


Figure 1. Quiet Sun spectrum from 800 Å to 1500 Å shown as spectral radiance. Selected prominent emission features are identified.

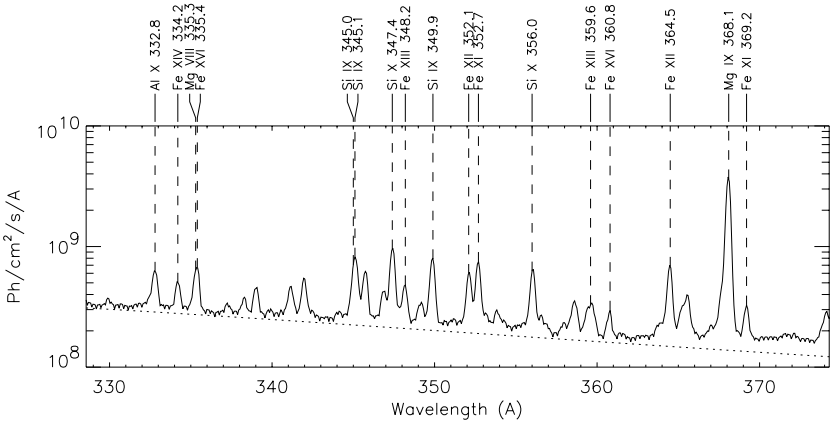


Figure 2. Integrated spectrum of the Sun observed with the normal incidence spectrometer (NIS) on CDS showing a number of highly ionized emission lines. The irradiance spectrum was derived by adding the emission from 690 different exposures distributed over the solar disk.

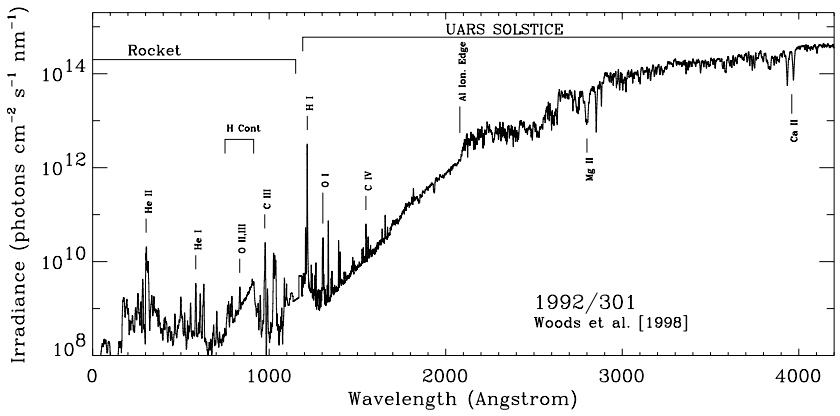


Figure 3. Solar VUV irradiance spectrum obtained with a LASP/NASA rocket instrument ($\lambda \leq 1200$ Å) and by the UARS/SOLSTICE instrument ($\lambda \geq 1200$ Å). At longer wavelengths the spectrum is dominated by the continuum emission and absorption lines while at shorter wavelengths the spectrum is dominated by emission lines.

Although some of the plasma processes can be studied by high-resolution images of the solar atmosphere much

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of our knowledge on temperatures, densities, emission measures, mass motions and elemental abundances comes from high-resolution spectral observations in the far ultraviolet (VUV) to the x-ray wavelength range (2000–1.7 Å).

The solar spectrum in the range 1150–1700 Å contains a large number of bright lines, mainly from plasmas of the CHROMOSPHERE and the lower TRANSITION REGION, i.e. at electron temperatures $T_e \leq 2.5 \times 10^5$ K. However, no lines from the upper transition region ($2.5 \times 10^5 \leq T_e \leq 1 \times 10^6$) are strong enough to be observed, and only a few lines originating at coronal temperatures appear mainly in PLAGES and LIMB spectra. In comparison, the solar spectrum in the range 100–1150 Å includes a large number of lines from the upper transition region and CORONA in addition to lines from the chromosphere and lower transition region.

At wavelengths longward of 1150 Å high spectral and spatial resolutions were achieved by several instruments. In the UV wavelength range there are in particular two instruments that made great impact on solar spectroscopy: the NRL/SO82B EUV spectrograph on Skylab ATM (Apollo Telescope Mount) and the NRL/High Resolution Telescope and Spectrograph (HRTS). More than 3000 emission lines have been observed in the 1150–1700 Å spectral range by these instruments and most of them have been identified. At shorter wavelengths, however, the spectral and spatial resolution was relatively poor and only the strongest lines could be identified. Recent observations with EUV spectrometers on the Solar and Heliospheric Observatory (SOHO) have greatly improved our knowledge about the emission at shorter wavelengths down to 150 Å. The Coronal Diagnostic Spectrometer (CDS) and the Solar Ultraviolet Measurement of Emitted Radiation (SUMER) complement each other since CDS relies heavily on diagnostics to determine the physical parameters in the solar corona and SUMER is designed to study the dynamical aspects of the chromosphere, transition region and corona in more detail. Selected line ratios give the electron temperature and the electron density. Absolute line intensities provide the differential emission line (DEM) distribution, the ion and elemental abundances. The line shift and broadening give information about dynamical phenomena of the emission plasma.

The SUMER spectrometer can observe emission lines in the spectral range 500–1600 Å. Figure 1 shows the quiet Sun spectral radiance spectrum from 800 Å to 1500 Å shown as spectral radiance. Selected prominent emission features are identified. In particular the wavelength range below 1175 Å has never before been observed with such high spectral resolution, and it contains a wealth of spectroscopic detail. More than 1000 emission lines have been observed and many of them have been recorded for the first time. About 98% of the observed spectral features have been identified. SUMER provided for the first time detailed information of high members of the hydrogen Lyman series which are important not only for understanding properties of the solar chromosphere but also of the O I excitation process in the Earth's atmosphere. SUMER is also the first instrument to provide comprehensive spectroscopic information on the solar corona above the limb out to two solar radii. In contrast to past observations, where only a very small number of lines were observed above the limb, more than 600 coronal emission lines have been observed in the quiet corona by SUMER, from which about 400 are still unidentified.

CDS is a dual extreme ultraviolet (EUV) spectrometer covering most of the wavelength range 150–780 Å. It has given the first detailed observations of this wavelength range with higher spectral, spatial and temporal resolution than previous instruments. The different wavelength bands have been carefully selected to cover useful spectroscopic diagnostic lines. An example of a quiet Sun spectrum observed with the normal incidence spectrometer on CDS is illustrated in figure 2. The wavelength range (310–380 Å) is dominated by Mg IX 368 Å and a number of lines from highly ionized ions, in particular from iron. A number of line pairs in this band are useful for density diagnostics using ratios such as the Si X 347/356, Si IX 345/349 and Fe XIII 359/348 lines.

Continuum emission

The UV radiation emitted by the Sun between 1400 Å and 1680 Å originates from the temperature minimum region and the low chromosphere according to calculated model atmospheres. The absolute value of the continuum intensity in the spectral range around 1600 Å is important since it reflects the value of the temperature minimum used in model calculations of the solar atmosphere.

The quiet solar radiation shortward of 1680 Å is primarily due to free–bound transitions from lower lying energy levels of neutral silicon in the temperature minimum region, about 500 km above $\tau_{5000} = 1$ according to model calculations. At wavelengths shorter than 1521 Å the emission is almost entirely caused by recombination to the ground state ($3p^2\ ^3P$) with smaller contribution to the emission from C I and from the Ly_α wing. At wavelengths longer than 1521 Å recombination to the first excited level ($3p^2\ ^1D$) of Si I is the main source of emission together with Fe I and Mg I. Ultraviolet line emission from higher temperatures (and heights) also contributes to the photoionization rate of neutral silicon in the temperature-minimum region. For quiet Sun conditions this contribution is small but is greatly increased during flares.

The continuum emission at shorter wavelengths is also a useful diagnostic tool. The C I, H I and He I continua have upper bounds of 1100 Å, 911 Å and 504 Å respectively. However, few observations of this weak continuum emission have been made until recently when the improved instrument sensitivity on the SOHO satellite became available. The H-continuum (also called the Lyman-continuum) is fairly free from emission lines as can be seen in figure 1. The C I continuum shortward of 1008 Å is contaminated by the He I continuum which tends to appear in the spectral second order of the grating.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectral irradiance and solar variability

Only about 1% of the Sun's total energy is emitted at VUV wavelengths less than 3000 Å, while approximately 50% is emitted between 4000 and 8000 Å. However, the variability in the VUV portion of the Sun's spectrum exceeds that at visible wavelengths and contributes significantly to the solar irradiance variability in spite of its minor contribution to the total irradiance itself (see SOLAR IRRADIANCE for more details). The VUV radiation is furthermore the dominant source of energy for heating and ionization in the terrestrial upper atmosphere at altitudes above 90 km. Thus, a good knowledge of the solar EUV spectral irradiance is of critical importance for many analyses of the photochemistry and energy balance of the ionosphere and the thermosphere. Solar UV light is primarily responsible for both creation and destruction of ozone in the Earth's stratosphere and mesosphere (see OZONE HOLE). Stratospheric ozone densities are known to vary with the 11-year SOLAR CYCLE. Solar variability over the solar cycle also causes expansion and contraction of the outward extension of the Earth's atmosphere into space. Thus, satellites in low orbit will feel an increased drag when the Sun is active and the result is reduced lifetime. Another important issue is that changes in the Earth's upper atmosphere induced by variations in solar UV radiation could affect the surface climate through feedback mechanisms. It is therefore important to obtain a reliable specification of the Sun's radiative output variability in order to isolate anthropogenic global changes from natural variability.

Bibliography

For further reading about the VUV emission a few selected pointers to key books and papers are listed below.

Mariska J T 1992 *The Solar Transition Region* (Cambridge: Cambridge University Press)

Mason H E and Monsignori Fossi B C 1994 *Astron. Astrophys. Rev.* **6** 123

Vernazza J E, Avrett E H and Loeser R 1981 *Astrophys. J. Suppl.* **45** 635

Woods T N, Rottman G J, Bailey S M, Solomon S C and Worden J 1998 *Solar Phys.* **177** 133

P Brekke

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Spectroscopy: Visible Emission

There are several perspectives from which to consider the solar spectrum. One is obviously as a tool for studying the Sun itself. The temperature and pressure variation with depth, the composition, the magnetic fields, the changing structure and velocity fields on the surface can be derived from high-resolution, high signal-to-noise spectra that also have high spatial and temporal resolution.

A second perspective is that the Sun is the brightest star; we can study it in the same way that we study other stars, but at higher resolution and higher signal-to-noise. The spectrum must be integrated or averaged over the solar disk to produce the flux spectrum. This is the spectrum seen by all the bodies in the solar system and that determines their atmospheric chemistry and weather. The flux spectrum at the mean distance of the Earth from the Sun is called the irradiance spectrum because it irradiates the top of our atmosphere (see SOLAR IRRADIANCE). However, the high-resolution flux spectrum has not been observed above the atmosphere. It is apparently too basic and fundamental to interest ESA or NASA. The available flux atlases have been made through the Earth's atmosphere and include the telluric absorption by O_2, H_2O, and other molecules. Figure 1 (color plate xx) shows the spectrum observed through the atmosphere from KITT PEAK in Arizona at 2000 m. It is plotted as a 'residual' spectrum with the upper envelope defined to be unity. In this form the overall energy information has been factored out of the spectrum. The actual energy spectrum is discussed below. The strong features in the spectrum were given letter designations in the nineteenth century that are still in use and are marked on the plot. Wavelength regions with strong telluric absorption are indicated. Our eyes have evolved to be sensitive only to the light that passes through the window of maximum transmission which we see as the colors shown in the plot.

The third perspective for studying the solar spectrum is that it serves as a spectroscopic data source for atoms and molecules that goes to higher temperatures and is more stable than what is available in the laboratory. There are many cases where lines can be seen in the Sun that have been difficult or impossible to see in the laboratory. Comparisons of predicted spectra to observed spectra provide a test of theoretical model atmosphere calculations (see also SOLAR ATMOSPHERE: EMPIRICAL MODELS).

In every case we want high signal-to-noise and high resolution, high enough to allow recovery of the true spectrum with instrumental degradation removed. This is resolving power, wavelength/Δwavelength, of one half million or more (0.001 nm at 500 nm) if there are no telluric lines, or resolving power of two million or more (0.000 25 nnm at 500 nm) if there are. Telluric lines are generally much narrower than solar lines because the Earth's atmosphere is much cooler than the Sun's atmosphere such that thermal line widths are about one fourth as great. Physical processes, such as RADIATIVE TRANSFER, always take place at high spectral resolution.

Reviews by the author that list the atlases of observed solar spectra in the visible are given in the bibliography. The reviews include very compressed plots of the spectra like that shown in figure 1 (color plate xx), which is from the Kitt Peak Solar Flux Atlas. Since the Sun is rotating at about 2 km s^{-1}, one limb of the solar disk is moving away from us and is redshifted while the opposite limb is moving toward us and is blueshifted by 2 km s^{-1}. Since the flux spectrum is an average over the disk, it is broadened by this rotation. A spectrum taken from a small area of the disk does not have this broadening so has more detail and reflects local physical conditions. For this reason there are also atlases of intensity spectra taken at the center of the disk, such as the Jungfraujoch atlas, and near the limb (see also SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION). The central intensity spectrum is a better spectroscopic source than the flux spectrum. Most of the atlases are available as large scale plots on paper. The atlases can be found in digital form on the World Wide Web. All existing atlases need much improvement in quality, both in signal-to-noise and in resolution.

Line identifications

The reviews also list published tables of line identifications. However, except for the far infrared, where some lines are isolated, lists of line identifications are not very useful because the features we see in the solar spectrum are not single lines. They are blends of many lines, from atoms and molecules of various species, some with isotopic components or hyperfine components, and both solar and telluric lines. Only a computed spectrum can indicate in any meaningful way what actually makes up each feature. Figure 2 shows a small section of spectrum between 599.0 and 599.2 nm that is relatively free of lines. It shows a theoretically computed solar flux spectrum, a theoretically computed transmission spectrum for Kitt Peak at the time the observation was made, the computed flux spectrum transmitted through the atmosphere to the top of Kitt Peak, and the flux spectrum observed at Kitt Peak. The stronger lines used in the calculation are labelled. These labels provide the indentifications for the observed spectrum (see also SOLAR SPECTROSCOPY AND DIAGNOSTICS. The computed and observed spectra do not match because of missing or inaccurate line data. At present one-half of the lines in the solar spectrum are not identified. Corrections to the line data can be determined by iterating on the spectrum calculations to improve the agreement. The computed spectra in figure 2 have been iterated in this way. Major improvements await new laboratory spectrum analyses and new line strength data for the atomic and molecular species that appear in the spectra.

In progress

New atlases for the flux, central intensity and limb intensity are in preparation with spectra taken by James Brault at Kitt Peak. These atlases will show the computed spectra and line identifications together with the observed spectrum as in figure 2. The irradiance spectrum, the true

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

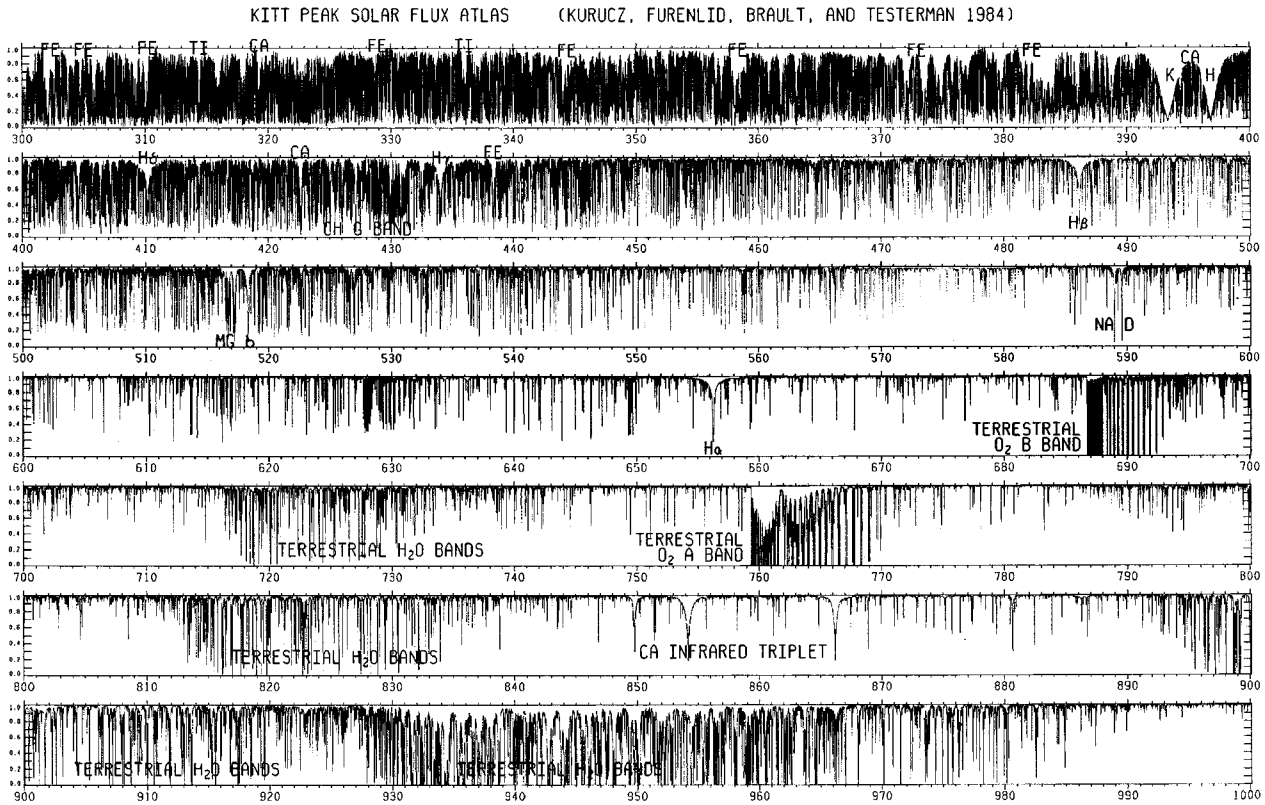


Figure 1. The solar flux spectrum observed through the atmosphere in true color from 300 to 1000 nm. It is plotted as a 'residual' spectrum with the upper envelope defined to be one. The strong features in the spectrum were given letter designations in the nineteenth century that are still in use and are marked on the plot. Wavelength regions with strong telluric absorption are indicated. **This figure is reproduced as Color Plate 49.**

solar spectrum, must still be determined. As the quality of the calculations improves it will be possible to remove many of the telluric features from the observed spectra.

Absolute calibration

Much of the actual irradiance energy distribution has been determined in 2 nm bands by Neckel and Labs. The residual flux spectrum can be integrated in those bands and normalized to produce the corrrect energy scale. Figure 3 shows the absolute irradiance spectrum derived from the Kitt Peak Solar Flux Atlas. In sections where there are telluric lines, the upper envelope of the spectrum is the upper bound to the true irradiance and the spectrum is the lower bound.

Bibliography

Kurucz R L 1991 The solar spectrum *The Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M Matthews (Tucson, AZ: University of Arizona Press) pp 663–9

Kurucz R L 1992 Atomic and molecular data for opacity calculations *Rev. Mex. Astron. Astrofís.* **23** 45–8

Kurucz R L 1993 Atomic data for interpreting stellar spectra: isotopic and hyperfine data *Phys. Scr.* **T47** 110–17

Kurucz R L 1995 The solar spectrum: atlases and line identifications *Laboratory and Astronomical High Resolution Spectra* ed A J Sauval, R Blomme and N Grevesse (San Francisco: Astronomical Society of the Pacific) pp 17–31

Neckel H and Labs D 1984 The solar radiation between 3300 and 12 500 Å *Solar Phys.* **90** 205–58

Robert Kurucz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

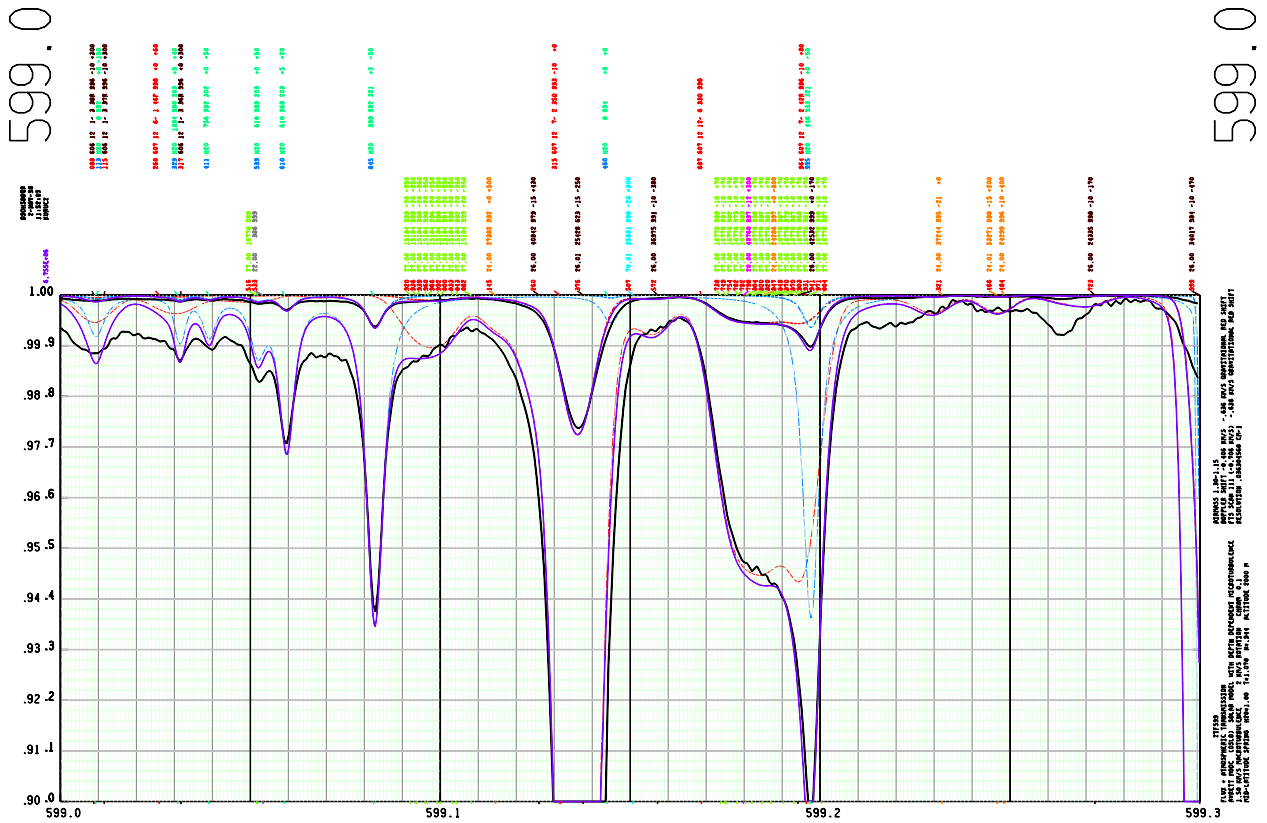


Figure 2. A small section of the solar flux spectrum at 599 nm plotted at full scale and at 10× scale. The heavy black lines show the observed spectrum from James Brault at Kitt Peak. The resolving power is 523 000 and the signal-to-noise is about 3000 which is poor for this work. The thin lines are the computed spectra: in red for the solar flux, in blue for the telluric atmospheric transmission and in purple for their product which should be compared with the observed heavy black lines. There are solar lines of Ca I, Ti I, Cr I, Cr II, Fe I, Fe II, Co I, Yb II, C_2, CN, and telluric lines of H_2O. The first number in each line label is the last three digits of the wavelength and the fourth number is the per mil line depth if the line were computed in isolation. The middle numbers are either the lower energy level in cm^{-1} for atoms or quantum numbers for molecules. Two of the Co I lines have been divided into hyperfine components. The hyperfine and isotopic splittings have not yet been determined for the other lines. Some splittings may be negligible. There are many missing lines. The lines that are present have been adjusted to improve agreement between calculation and observation. That process is continuing. This sample will be part of a forthcoming atlas of Brault's solar spectra. **This figure is reproduced as Color Plate 50.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

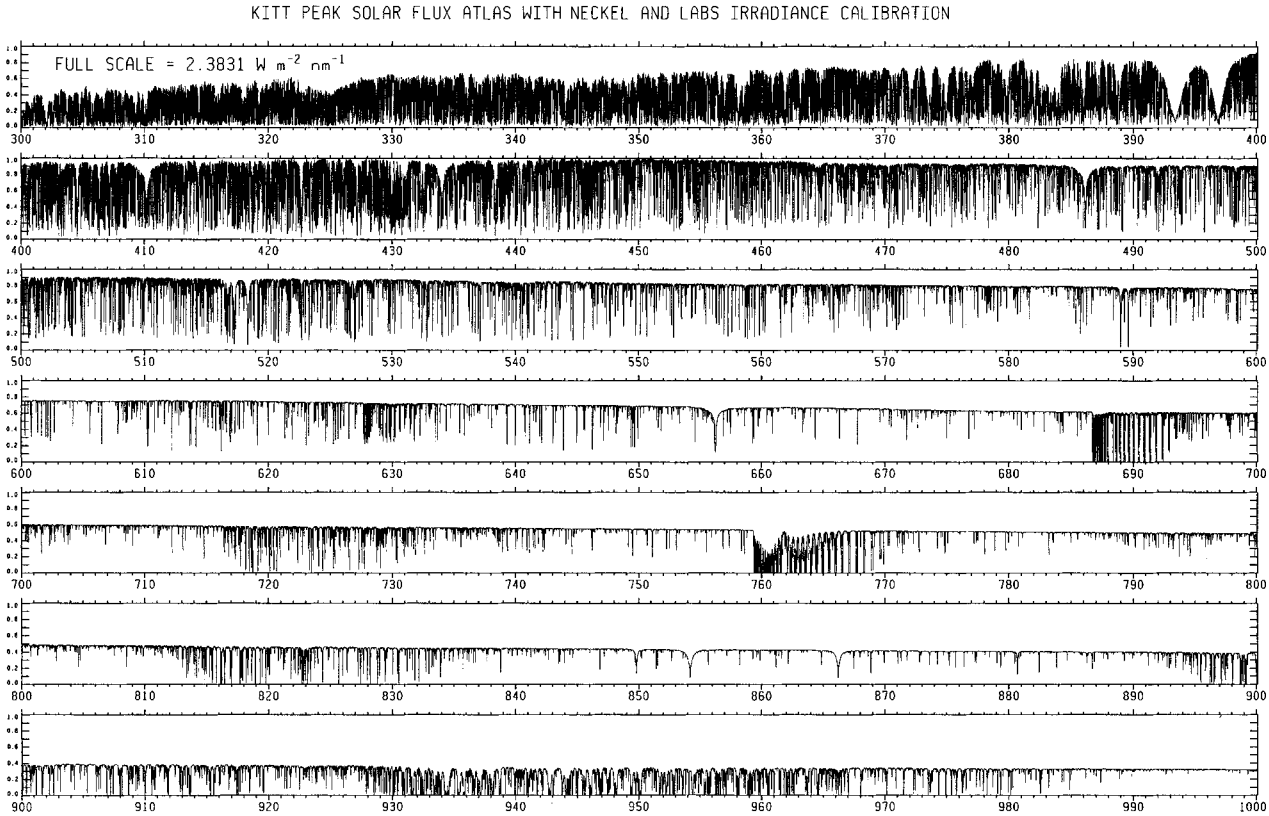


Figure 3. The absolute irradiance spectrum, the solar flux at the top of the atmosphere, derived from the Kitt Peak Solar Flux Atlas by normalizing it to the low-resolution irradiance measurements of Neckel and Labs. The wavelength range is 300 to 1000 nm. All the telluric oxygen and water lines must be edited out to determine the true irradiance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar System

The region of the universe in which we live, the Solar System, is but one of many clusters of stars and planets in space. Our Solar System lies in a particular region of the universe called the 'Milky Way Galaxy'.

The Solar System originated some 4.6 billion years ago through the accretion of gases and dust grains. Four groups of objects constitute the Solar System: the Sun, the terrestrial planets (including Earth's moon), the Jovian planets and their satellites, and the small bodies (asteroids and comets) (figure 1). Recent discoveries prove that our Solar System is not alone. Astronomers have discovered planets orbiting several other stars.

All the planets move around the Sun in the same direction, and nearly in the same orbital plane, the ecliptic plane. The ecliptic plane is roughly in the equatorial plane of the Sun's rotation. The planets' axes of rotation are nearly perpendicular to this plane, with the exceptions of Uranus and Pluto, which are tilted on their sides (table 1). Other small bodies, such as comets, are in elliptical orbits. Some travel as far out as Jupiter and return close to the Sun, repeating these orbits until disturbed by the gravitational attraction of a passing planet.

The Sun

Most of the mass of the Solar System is contained within the Sun which mass is 330 000 times Earth's mass. It is composed primarily of hydrogen and helium with tiny amounts of other elements, including carbon, nitrogen, oxygen, silicon, and iron. The Sun is more than 1390 000 km across. Like the planets, the Sun rotates on its axis; its rotation period is 27 d. At its center temperatures are thought to reach nearly 15 000 000 °C. The Sun's energy comes mainly from the thermonuclear fusion of hydrogen to form helium. This thermonuclear reaction takes place in the Sun's core (see also SUN: BASIC PROPERTIES).

This energy works its way gradually outward through the Sun's inner layers (i.e. the radiative and convective zones) until it reaches the photosphere, the visible surface of the Sun and the site of the sunspots and (boiling) convective motions. Above the photosphere is the chromosphere into which huge flares and loops of hot gas often shoot, extending tens of thousands of kilometers above the Sun's surface; the flares shoot electrically charged particles into the Solar System, causing aurorae when they reach the Earth. The CHROMOSPHERE is the transition from photosphere to the Sun's outermost region, the corona, which consists of a very thin layer of extremely hot gas.

The SOLAR WIND is a gale of relatively low-energy atomic particles, protons and electrons, that form a hot, magnetized gas, a plasma. Solar wind comes from great rips in the Sun's corona, called coronal holes. These holes appear to be predominant in the polar regions and are fairly small. The fast wind blows at a steady 750 km s^{-1}, but the typical solar wind emerging from the Sun's equatorial zone is variable but relatively slow, at 350–400 km s^{-1}. The solar wind flows straight out in all directions and carries the magnetic field lines with it.

The Sun's magnetic field is generated by plasma motions below the Sun's surface and extends out to shape and control the solar atmosphere and the entire heliosphere. It is uniform at all latitudes in the heliosphere. Close to the visible surface of the Sun, the magnetic field is strongest over the poles. The magnetic field lines carry other particles that have greater energy than those in the solar wind (see also SOLAR MAGNETIC FIELD). Each of the high-energy particles has an energy of up to millions of electron volts. The highest-energy particles are few in number, so their density is low compared with that of the solar wind.

The terrestrial planets

The terrestrial planets consist of Mercury, Venus, Earth, Mars, and the Earth's moon. Although by strict definition Earth's MOON is not a planet, it is usually included in the terrestrial planets' category because of its large size, its rock composition, and the similarity of its early geologic history to the other terrestrial planets.

The terrestrial planets, also called the inner planets because of their place in the Solar System, show several common characteristics. Their relatively high densities suggest that, like Earth, they are composed of elements such as silicon, iron and nickel. These dense elements would have been retained close to the Sun, while lighter elements, such as hydrogen and helium, would have escaped to the outer reaches of the Solar System during its formation and early evolution (see SOLAR SYSTEM: FORMATION). The individual planets evolved by the collection of smaller bodies, planetesimals, through accretion. As they grew in size, their increasing mass attracted still more bodies. Within the first half-billion years of Solar System history much of the debris had been swept up by the planets. The terrestrial planets began to cool and form crusts as elements combined and crystallized as rocks and minerals. Various combinations of silicon, oxygen, iron, magnesium, sodium, calcium, potassium and aluminum produced the silicate minerals of most rocks composing the Earth and lunar crusts, and probably most of MERCURY, VENUS and MARS. With the formation of a crust, the geologic evolution was initiated, and each developed along different paths.

Venus, Earth, and Mars all have significant atmospheres. Probably some gases were accumulated by each protoplanet from the original solar nebula, but these original atmospheres have been lost to space. Secondary atmospheres were outgassed from their hot interiors (with a possible contribution also coming from meteoritic impacts) and subsequently modified by chemical interactions with their surfaces. Mercury and the Moon are too small to retain anything but the most tenuous atmospheres.

Mercury is the planet closest to the Sun. It has the most elliptical orbit of any planet except Pluto. Because of its elliptical orbit, Mercury's closest distance to the Sun is only 46 million km while its greatest distance is 70 million km. Because Mercury is so close to the Sun, noontime

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Solar system planet data.

	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
Mean distance from the Sun (10^6 km)	57.9	108	150	228	778	1426	2868	4494	5900
Mean distance from the Sun (AU)	0.39	0.72	1.00	1.52	5.20	9.54	19.18	30.06	39.44
Revolution period (yr)	0.24	0.62	1.00	1.88	11.86	29.46	84.07	164.82	248.6
Orbital eccentricity	0.206	0.007	0.017	0.093	0.048	0.056	0.047	0.009	0.246
Inclination to ecliptic (deg)	7.00	3.40	0	1.85	1.30	2.49	0.77	1.77	17.17
Rotation† (d)	58.65	243.0 (R)	1.00	1.03	0.41	0.44	0.65 (R)	0.76	6.39 (R)
Diameter (km)	4878	12 102	12 756	6788	142 984	120 536	51 118	50 538	2300
Mass (Earth = 1)	0.06	0.82	1	0.11	317.89	95.18	14.54	17.15	0.002
Density (g cm^{-3})	5.4	5.2	5.5	3.9	1.3	0.7	1.3	1.7	2.0
Escape velocity (km s^{-1})	4	10	11	5	60	36	21	23	1
Surface (main materials)	Silicates	Basalt, granite?	Basalt, granite, water	Basalt, ice	None	None	None	None	CH_4 ice
Atmosphere (main components)	trace Na	CO_2 (98%)	N_2 (78%), O_2 (21%)	CO_2 (95%)	H_2, He, CH_4, NH_3	H_2, He, CH_4, NH_3	H_2, He, CH_4, NH_3	H_2, He, CH_4, NH_3	trace CH_4

† (R) = retrograde

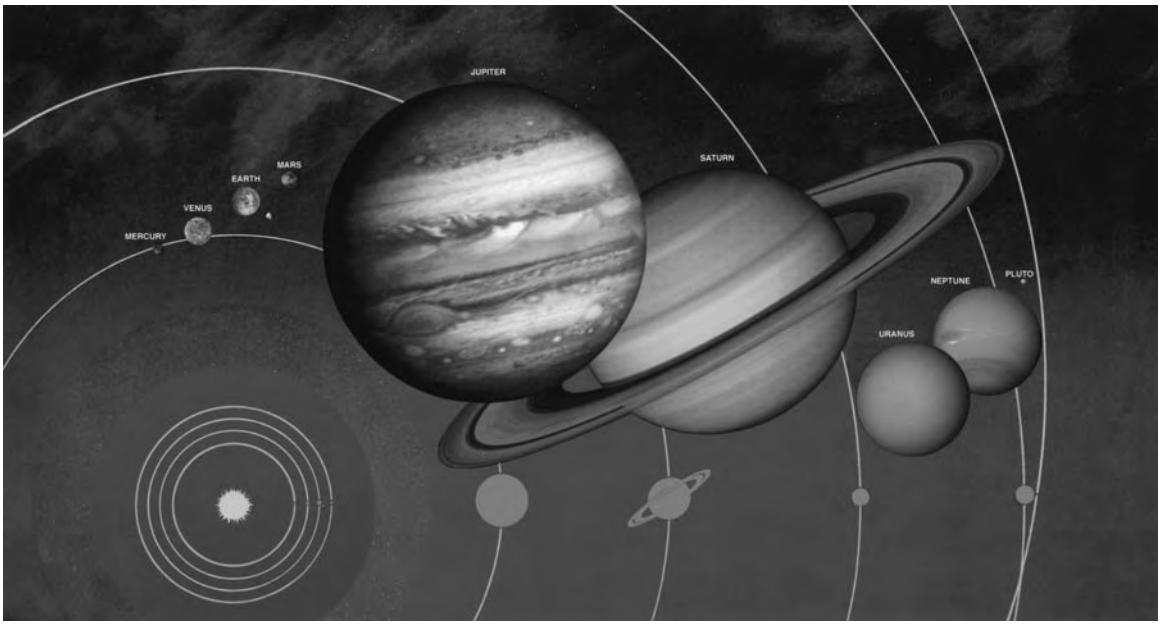


Figure 1. Our Solar System (NASA lithograph LG-1997-12-475-HQ).

temperatures soar to 370 °C. But because it has almost no atmosphere to hold in the heat, temperatures plummet to almost −185 °C at night. Mercury rotates on its axis once every 58.6 d and circles the Sun once every 87.9 d. Mercury is smaller than any other planet except Pluto, and is about one-third the size of Earth. However, Mercury's density (5.4 g cm^{-3}) is about the same as Earth's and, therefore, it is assumed to have an enormous iron and nickel core some 75% of the planet's diameter (42% of the volume), surrounded by a rocky mantle and crust only 600 km thick. As Mercury spins on its axis, the iron in its core generates a weak magnetic field that surrounds the planet.

Although the surface of Mercury superficially resembles that of the Moon, there are significant and geologically important differences. Like the Moon, it has heavily cratered upland regions and large areas of smooth plains that surround and fill impact basins. The heavily cratered terrain records the period of heavy meteoroid bombardment that occurred throughout the Solar System during its early history and ended about 3.8 billion years ago. Unlike the Moon, however, Mercury's heavily cratered terrain contains large regions of gently rolling intercrater plains, the major type of terrain on the planet. Both the smooth plains and intercrater plains are attributed to volcanism early in Mercury's history. Mercury also has experienced a unique history of crustal deformation which has resulted in a global system of compressive structures. The general geologic history of Mercury is that soon after the planet formed it nearly melted by heating from the decay of radioactive elements and the inward migration of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the large amount of iron to form its enormous core. This led to expansion of the planet and extensional fracturing of the surface that provided an exit for lavas to reach the surface and form the intercrater plains. As the core and mantle began to cool, Mercury's radius decreased by about 2–4 km, and the crust was subjected to compressive stresses. At about this time, a large basin was formed by a gigantic impact. Further eruptions of lava within and surrounding this large basin and other large impact craters formed the smooth plains. Volcanism finally ceased when lithospheric compressive stresses became strong enough to close off magma sources. All these events probably took place very early, perhaps in the first 700 or 800 million years, in Mercury's history.

Venus is the hottest world in the Solar System. It is blanketed by a thick layer of clouds that generate a greenhouse effect and consequently heat its surface up to 480 °C. The atmospheric pressure at the surface is 90 times more than that at Earth's surface. The atmosphere of Venus is 98% carbon dioxide and contains virtually no water vapor. The upper clouds are composed of sulfuric acid droplets. Astronomers have called Venus Earth's twin because of their apparent similarities: their mass and volume (and hence their density and gravity) are remakably alike. However, Venus is very different in other respects. Venus has no satellite. A magnetic field has not been detected. A day on Venus (243 Earth days) is longer than its year (225 Earth days). Venus rotates in a direction opposite that of Earth and most other planets.

Venus' surface consists of about 27% lowlands, 65% rolling plains, and 8% highlands. The density of craters on Venus is much lower (about two craters per million square km) than densities on the Moon and Mars, which suggests a relatively young average surface age. Volcanism (vast lava plains, thousands of volcanoes) and tectonism (highly fractured elevated terrains, mountains over 11 km high above the plains) are widespread and fundamentally important processes in the formation and evolution of Venus' crust. Models of the rate of volcanic resurfacing based on crater inventory suggest that local surfaces can have an age from very young to 800 million years old.

Like the other terrestrial planets, Venus was probably formed by accretion of small bodies about 4.6 billion years ago. With time and cooling, a crust, mantle, and core developed. The lithosphere could be 100 km thick, and the core could be about the same size as Earth's core. The lack of abundant impact craters on Venus suggests that most of the crust that formed early in its history has been destroyed by burial by lava flows. The formation of both large and small volcanoes and the presence of extensive lava flows over hundreds of millions of years on Venus were accompanied by crustal deformation that led to the development of large tectonic structures (e.g. faults and rifts).

Earth is surrounded by a gaseous atmosphere composed primarily of nitrogen and oxygen (see EARTH'S ATMOSPHERE). Beneath the atmosphere, the solid surface is nearly (70%) covered by a global, 4 km deep ocean. The atmosphere and ocean balance incoming solar radiation and outgoing terrestrial radiation to maintain surface temperatures whithin the small range which permits life. Earth is special in many respects. Of all the planets in our Solar System, Earth is the only one that can support life. No other planet has oceans of water and an oxygen-rich atmosphere. It is also the largest body in the Solar System with a solid surface and the only planet with ongoing plate tectonics. However, Earth is just an average planet. It is bigger than four planets in the Solar System and smaller than the other four (table 1).

Earth formed about 4.6 billion years ago, and it is still evolving. Earthquakes and volcanoes are violent proofs of Earth's shifting crust. In the early stages of its history (4.6–4.0 billion years ago), heat from planetary accretion melted most, if not all, of the mass. It is usually thought that volatile elements such as oxygen, carbon, and nitrogen were implanted in the accreting Earth by planetesimals. Dense elements, such as iron, settled into the interior of the molten protoplanet to form the core, while lighter elements rose to the surface, leading to chemical differentiation of our planet (see EARTH'S INTERIOR). Cooling of the surface of the protoplanet led to the formation of a thin crust. However, continued impact bombardment fractured the crust, allowing parts of the still molten interior to flow to the surface. Gases trapped in the interior were released to the surface from these volcanic processes and formed Earth's early atmosphere. With time, rainfall accompanied the evolution of the early atmosphere and led to the accumulation of liquid water on the surface. The constant weathering and renewal of Earth's crust have obliterated nearly all of the craters which were formed during this period of time.

Tectonic deformation of the crust and volcanism contributed to the formation of mountains and valleys, generating topographic relief. Erosional processes, including running water, glaciers and weathering, have continually reduced the high areas, and deposition of transported sediments have filled up the low areas to form plains. However, much of the Earth's surface results from internal processes, i.e. convection within the mantle. The mantle that surrounds the iron–nickel core is composed mostly of hot, plastic, silicate materials and is subdivided into the mesosphere, the astenosphere (a plastic layer about 250 km thick), and the lower part of the lithosphere (a cool rigid layer about 100 km thick). The upper part of the lithosphere corresponds to the crust whose average thickness ranges from 8 km (the oceanic crust composed of silicate rocks rich in iron and magnesium) up to 70 km (the continental crust composed of silicate rocks rich in aluminum, silica, and calcium). The lithosphere consists of seven major plates, and many smaller plates, sliding about on the underlying astenosphere. The energy responsible for the plate motion comes from the heat generated by the decay of radioactive elements, plus residual heat resulting from planetary accretion some 4.6 billion years ago. Because of plate motion, most of the record earlier than 600 million years ago has been lost through plate

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

tectonics and crustal recycling. Consequently, some 80% of Earth's history is not well understood.

Because of synchronous rotation, the *Moon* revolves once on its axis each time it orbits the Earth, thus always presenting the same face, the nearside, toward our planet. The lunar nearside shows two contrasting types of landscape: dark, plains-like areas of low relief, and brighter, more rugged regions which cover about two-thirds of the surface. The smooth dark areas were named *maria*, and the name *terrae* was given to the bright upland regions. The *terrae* are seen to be densely cratered highlands. Craters of all sizes up to about 250 km across are scattered over the surface in great profusion, frequently overlapping one another. The *maria* vary from almost circular features to more irregular-shaped regions. Their dark surfaces are apparently smooth, broken only by a few large craters. The *maria* are younger than the *terrae* whose cratered surfaces may be relics of a primordial landscape.

Two different processes have been responsible for the present-day landscape: crater-forming impact of countless meteorites, and igneous and volcanic activity from within the Moon. Early in its history the Moon was subjected to intensive meteorite bombardment, including catastrophic basin-excavating impacts of massive asteroidal bodies. The crust probably has been intensely fractured by those basin-excavating impacts to a depth of 20 km or more. These huge basins, which range in size from 300 to more than 2500 km across, were subsequently filled by tremendous surface outpourings of volcanic lava (i.e. basalts) to form the *maria*. Although lacking the large, lava-filled basins so characteristic of the nearside, the farside of the Moon displays a mostly terra-type landscape, formed almost entirely by heavily cratered highlands, and several large crater-like basins. Since large impact basins are evenly distributed over the Moon's surface, impact histories for the two sides are about the same. The absence of extensive spreads of lava in farside basins may be related to the tidal forces exerted by the Earth on a partly molten Moon in its history. It may also reflect the farside presence of a much thicker layer of crustal rocks through which there were no large-scale outpourings of lava like those through the comparatively thin-crusted nearside. The lunar crust appears variable in thickness, ranging from a few tens of km beneath some basins to more than 100 km under some highland regions.

The Moon may have formed during a collision between the early Earth and a Mars-sized rocky planet about 4.6 billion years ago. The collision of this object with the proto-Earth threw a cloud of debris into orbit around Earth. This debris, which probably formed a disk-like body in Earth orbit, quickly collected into a single, large body, our Moon. The rapid rate of lunar formation released large amounts of heat which produced near-global melting, a stage called the magma ocean. As minerals began to crystallize in this magma ocean, the low-density, aluminum-rich minerals floated upward, forming the lunar crust (about 70 km thick). High-density, iron-rich minerals crystallized in the lower levels of the magma ocean, forming the lunar mantle, which represents most of the mass of the Moon, and probably an iron-rich core that may be partly molten and is perhaps 800 km in diameter. This differentiation of the Moon was accomplished around 4.5 to 4.4 billion years ago. The early Moon was very hot, and magma continued to permeate the solidified crust for at least the next 300 million years. Some magmas produced by the melting of the lunar mantle flowed out onto the Moon's surface as volcanic eruptions that began about 4.3 billion years ago. Most surfaces of these lava flows were obliterated by large impacts. The dark plains that dominate the lunar nearside are lava flows that erupted mostly between 3.8 and 3.0 billion years ago. The principal process operating on the Moon for the last 3 billion years has been impact cratering, dominated by bodies in the microscopic size range. This bombardment of micrometeorites has ground up the lunar bedrock into a chaotic mass of powdery, fine debris and rock fragments, several metres thick. This material, called regolith, is directly exposed to both cosmic radiation and the solar wind. Thus, the regolith preserves a record of the history of the Sun's output for the past 3 billion years.

Mars is the outermost of the terrestrial planets. At 6788 km, its diameter is intermediate in size between Earth and Mercury. Mars is generally redddish-orange in color due to abundant iron oxide in its surface materials. Mars moves in a distinctly elliptical orbit, in contrast to the Earth's nearly circular orbit. Its orbit is so elliptical that Mars varies in distance from the Sun by 20%, which results in a 45% variation in solar radiation received at the surface and a 30 °C range in temperature at the subsolar point. This effect, combined with the effects from the Martian seasons and the absence of an ocean that acts as a heat sink, produces an extraordinarily wide range in surface temperature. The surface warms to about 22 °C in summer at mid-latitude and cools to about −125 °C at the south pole during mid-winter. Mars has a thin, cool atmosphere made primarily of carbon dioxide. The average pressure at the surface is about one hundredth that of Earth's atmosphere. The global circulation of the atmosphere is dominated by the seasonal exchange of volatiles as they migrate from the summer pole to the winter pole. Occasionally winds on Mars are strong enough to raise dust and initiate a dust storm.

The surface of Mars has recorded evidence of a wide range of geological and atmospheric processes that have operated over the history of the planet. On a global scale the Martian surface is divided into roughly two hemispheres separated in places by a kilometre-high scarp. This escarpment extends around the planet at a 35° angle from the equator. To the north of the escarpment are sparsely cratered lowlands, the northern plains, and to the south are regions of elevated and heavily cratered terrain. On a regional scale the northern plains are featureless with only a few impact craters. But locally these plains exhibit a wide variety of surface features indicative of wind action, water erosion, ground ice action, and volcanism. The high density of impact craters in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the southern terrain suggests that this terrain is ancient (about 4 billion years old). Channel systems are common in the old cratered terrain, suggesting fluvial erosion and dramatic climate change early in Martian history, from a warmer wetter environment to the present cool dry conditions. Because liquid water cannot exist on the surface of Mars under present conditions (the atmosphere is too thin and cold), Mars probably had a warmer, denser atmosphere at the point in its history when the channels were eroded. However, frozen water still might exist in Martian subsurface layers.

Located just north of the equator is the Tharsis dome, a gigantic volcanic and tectonic bulge, 5000 km across and 7 km high. Three enormous volcanos, the Tharsis Montes, are located in line near the top of the bulge. They are all 300–400 km across. Olympus Mons, the tallest volcano on Mars, is located on the northwest flank of the Tharsis bulge. It is about 600 km across, and like the other large volcanos its absolute elevation is about 27 km. An extensive array of faults radiates from the Tharsis bulge. A 5000 km long interconnecting canyon system, called Valles Marineris, radiates outward on the eastern flank of the Tharsis bulge. In places, the single canyon is 700 km across and 7 km deep. Enormous landslides are common along the canyon's wall. This canyon system is one of the major tectonic features of the planet and may have resulted from tectonic extension, or rifting. Valles Marineris merges eastward with canyons that contain arrays of jumbled blocks that appear to have formed by collapse. Large outflow channels originate in this area and extend northeast into the lowlands. These channels are thought to be produced by the catastrophic release of subsurface ground water.

Mars probably has a central, high-density core composed mostly of iron with small amounts of sulfur, oxygen, and nickel. Outward from the core is a thick iron-rich mantle. The outermost shell is the low density crust, a few tens of km thick, and probably composed of basaltic rocks. Though Mars is made of generally the same materials as the other terrestrial planets, its mean density is 3.9 g cm^{-3}.

Two small moons, PHOBOS AND DEIMOS, orbit the planet. They are thought to be captured asteroids composed of dark, carbon-rich rock. Each satellite has been heavily cratered.

The Jovian planets

JUPITER, SATURN, URANUS, and NEPTUNE are referred to as the Jovian planets. All are gaseous objects and lack solid surfaces. Unlike the terrestrial planets, they are enormous and contain most of the mass in the Solar System outside the Sun.

The Jovian planets formed by the accretion of smaller solid bodies to form a nucleus of sufficient size to capture gravitationally the lighter elements that escaped from inner parts to the outer parts of the Solar System. As this process continued, the Jovian planets grew to consist mostly of hydrogen and helium. They grew and evolved to have a family of smaller bodies in orbit about them. These moons are typically composed of ices as well as rocky materials. One of them, Titan, a satellite of Saturn, has a thick atmosphere of nitrogen, along with methane, ethane, acetylene, ethylene, hydrogen cyanide, and other carbon–nitrogen components. The satellites of the Jovian planets represent a collection of objects of different sizes, compositions, and geologic histories. Some, such as the Jupiter moons Ganymede and Callisto, are about the size of Mercury. Others, such as the Saturnian satellite Enceladus and the Uranian satellite Miranda, have experienced spectacular crustal deformation. Others appear to have remained relatively unaltered since their initial formation. Two of them, Jupiter's Io and Neptune's Triton, are currently volcanically active.

Jupiter is the largest planet in the Solar System. The characteristics of its interior are highly speculative. Toward the center of the planet, the pressure is estimated to be 100 million times greater than the atmospheric pressure on the Earth's surface, and the temperature may reach 30 000 °C. Under this conditions, hydrogen would have a density of 4 g cm^{-3} and would form a metallic phase. Some models suggest a central solid core as large as 20 000 km across, which may contain small amounts of rocky material. The planet completes one orbit of the Sun each 11.8 yr and its day is 9 h 55 min. The rapid spin rate may explain the huge magnetic field that surrounds Jupiter. The cloud patterns (i.e. zonal jets, eddies, storms, Great Red Spot) apparently result from convection cells that are driven by the escape of heat from Jupiter. A faint dusty ring of material encircles Jupiter. Its outer edge is 129 000 km from the center of the planet, and it extends inward about 30 000 km.

Jupiter is now known to possess at least 16 satellites (see JUPITER: SATELLITES). The Galilean satellites, Io, Europa, Ganymede, and Callisto, are the largest. Io is the most famous satellite because active volcanos dot its surface. They erupt frequently, spewing sulfur dust up to 300 km high. Io's volcanos are apparently due to heating of the satellite as it is perturbed in its orbit by Europa and Ganymede then pulled back again by Jupiter. It appears that volcanism on Io affects the entire Jovian system, in that it is the primary source of matter that pervades Jupiter's magnetosphere. Europa, the next farthest moon from Jupiter, is covered by a thin crust of water ice (30 km thick), possibly floating on a H_2O ocean which might be about 50 km deep. The surface of the next satellite, Ganymede, shows two distinct types of terrain, cratered and grooved, suggesting that its entire icy crust has been under tension from global tectonic processes. Callisto, the Galilean farthest from Jupiter, has a very old, heavily cratered surface showing remnant rings of enormous impact craters.

Jupiter's rings and satellites exist within an intense radiation belt of electrons and ions trapped in the planet's magnetic field. These particles and fields comprise the Jovian magnetosphere, which extends 3–7 million km toward the Sun, and stretches in a windsock shape at least as far as Saturn's orbit, a distance of 750 million km. As the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

magnetosphere rotates with Jupiter, it sweeps past Io and strips away about 1000 kg s^{-1} of material. The material forms a torus, a doughnut-shaped cloud of ions. The torus's heavy ions migrate outward, and their pressure inflates the Jovian magnetosphere to more than twice its expected size.

Saturn is the second largest planet in the Solar System and takes 29.5 Earth years to complete one orbit of the Sun. Saturn has at least 18 satellites and a complex ring system (see SATURN: SATELLITES; RINGS). Measured from edge to edge, Saturn's rings span about one million km. Saturn's rings are made of ice and rock. They are not solid, but consist of small chunks of frozen material in a relatively thin layer. Some of the rings show elaborate structures such as kinks and spokes that are due to the gravitational effects of nearby satellites. Radial, spokelike features are believed to be composed of fine, dust-size particles. Electrostatic charging may create spokes by levitating dust particles.

Saturn shows broad atmospheric banding similar to but much fainter than that found on Jupiter, and ovals and other atmospheric features are generally smaller than those on Jupiter. An enormous white cloud has been discovered near the equator that grows and spreads rapidly, and rises higher and higher in the atmosphere. This cloud is believed to be composed mainly of ammonia ice crystals.

Saturn holds a wide variety of satellites in its orbit, ranging from Phoebe, a small satellite that travels in an orbit in the opposite direction from the other satellites (probably a captured asteroid), to Titan, the planet-sized satellite with a thick nitrogen–methane atmosphere which may strongly resemble the Earth's atmosphere before life evolved. The most active surface of any of Saturn's satellites is that of Enceladus. Its bright surface is marked by faults and valleys, showing evidence of tectonic motions of the icy crust. One other satellite, Mimas, is scarred with a crater so huge that the impact that caused it nearly broke the satellite apart.

Uranus has a cloudy atmosphere of primarily hydrogen and helium with a small amount of methane that causes its bluish-green color. Uranus is unusual in several ways. Unlike any other planet in the Solar System, it spins sideways. Uranus also rotates in the opposite direction from most of the other planets. Its magnetic field is also tipped by about 50° with respect to the planet's axis, making the magnetic field of Uranus, along with that of Neptune, unique among the planets. Surrounding Uranus are several very thin dark rings, similar to those around Jupiter.

Five large moons, OBERON, TITANIA, MIRANDA, UMBRIEL, and ARIEL, and 10 smaller ones orbit Uranus. Their orbital paths are sideways compared to other planets and moons because of Uranus' sideways spin. The largest moon, Titania, is 1610 km in diameter and its icy surface shows a lack of large craters that may have been erased by some internal process. Miranda is the smallest of the five large satellites of Uranus. Its icy crust displays a puzzling array of geologic forms, including huge oval features and complex systems of faults and ridges that may result from either internal processes, or the random reassembly of Miranda's fragments after it was broken apart by impact with an asteroid-sized body.

Neptune is the fourth, outermost, and smallest gas giant of our Solar System. Even though Neptune receives only 3% as much sunlight as Jupiter does, it is a dynamic planet. Neptune's atmosphere, mostly composed of hydrogen with methane in its upper part, exhibits several large, dark spots reminiscent of Jupiter's Great Red Spot. The magnetic field of Neptune, like that of Uranus, is highly tilted (47° from the axis of rotation).

Neptune has at least eight satellites. TRITON, the largest, is one-third as massive as Earth's Moon. Triton is also the most unusual moon, because it orbits Neptune in the opposite direction from all the other satellites (see also NEPTUNE: MINOR SATELLITES). Triton has an extremely thin atmosphere extending to about 800 km above its surface. The atmospheric pressure at the surface is about 1/70 000th the surface pressure on Earth, and the surface temperature is about −235 °C. Triton's surface, mostly frozen methane and nitrogen, shows evidence of a remarkable geologic history, with active geyser-like eruptions spewing invisible nitrogen gas and dark dust particles several kilometres into its tenuous atmosphere.

Small bodies and Pluto

ASTEROIDS, COMETS, and the smaller planetary satellites are often called small bodies. However, some are not so small; the largest asteroid, Ceres, is more than 1000 km across.

Most *asteroids* occur in the zone between the orbits of the planets Mars and Jupiter, the 'asteroid belt'. In addition to the asteroid belt, asteroids are found elsewhere in the Solar System. Two sets of asteroids, named Trojan asteroids, are found in the same orbit as Jupiter. Three populations of asteroids, referred to as near-Earth asteroids, reside in the inner Solar System. Among these three populations, two of them, the Atens and the Apollos, cross Earth's orbit and occasionally collide with the Earth and the Moon. The third population, the Amors, is found between Earth and Mars. More than 18 000 asteroids have been identified, of which the orbits of some 5000 have been determined. It is estimated that there are at least 1000 asteroids larger than 30 km, of which some 200 are larger than 100 km. Asteroids are classified both by type, related to their inferred composition, and by their location in the solar system. Because many meteorites seem to be derived from asteroids, a classification of asteroids is based on the comparison of their spectral reflectance measurements with compositional signatures of meteorites. This classification includes three groups: C-type for carbonaceous asteroids, rich in carbon and complex organic compounds, considered to be the most primitive, with unaltered material similar in composition to the terrestrial planets; S-type for siliceous asteroids, composed of silicate materials lacking the carbon-rich compounds, considered to be slightly processed; M-type for metallic asteroids, considered to be chemically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

differentiated as the result of extreme heating. In addition, a unique asteroid, Vesta, appears to consist of rocks similar to the basaltic lava flows found on Earth and the Moon.

Comets are probably primordial material left over from the early stages of the Solar System's formation. Commonly described as 'dirty snowballs', comets appear to be composed of solid dust grains embedded in water ice. The main component of a comet is the nucleus, a mass of ice and dust that typically is 1–10 km across. As comets travel toward the Sun, at about the distance of the orbit of Mars they begin to release gases in response to solar energy. The released gases form a glowing mass, the coma, surrounding the nucleus, 100 000 km across. Streaming away from the coma is the cometary tail which extends over 1–10 million km and is always oriented away from the Sun, regardless of the comet's trajectory. The cometary tail includes two components: the plasma tail composed of carbon dioxide, nitrogen, and water ions and a broad, slightly curved tail composed of dust grains. Most comets reside in the 'Oort cloud', and the 'Kuiper belt', beyond the orbit of Pluto. It is assumed that perturbations from passing stars may reset the orbits of some comets residing in the Oort cloud. Once travelling on an elliptical path through the inner solar system, comets may not survive in terms of Solar System history. Many of them are thought to collide with planets (e.g. the collision of Comet Shoemaker–Levy 9 into Jupiter in July 1994), satellites, the Sun, or other objects. With each fly-by of the Sun, a comet loses material into deep space by gas jetting at a typical rate of about a million tons per day. At this rate, comets would shrink by several meters with each orbit and would have lifetimes of only thousands of years.

Pluto does not fit into the subdivision of planets as well as terrestrial and Jovian objects. Little is known of its properties; almost nothing is known of its surface. Although Pluto is the outermost of the Sun's known planets, its elliptical orbit occasionally brings Pluto inside the orbit of Neptune. Pluto has a satellite, Charon, which is 1190 km in diameter, or about half the size of its parent planet. Thus, the Pluto system is considered to be a double planet. Moreover, PLUTO AND CHARON are dynamically locked, always showing the same face to each other. Ground-based observations and from the Hubble Space Telescope show that Pluto's surface contains methane, while Charon is mostly water ice. Pluto also has a thin atmosphere of methane and may contain carbon dioxide or molecular nitrogen. Both Pluto and Charon are dense objects (>2 g cm^{-3}), and are thought to consist of a rocky core surrounded by a water-ice mantle. Because of their peculiar orbits and high densities, the origin of Pluto and Charon remains speculative: they could be either satellites of Neptune that escaped, or remnants of a large satellite of Neptune that was broken apart by a collisional impact, and knocked out from Neptune's orbit.

Bibliography

Beatty J K and Chaikin A 1990 *The New Solar System* 3rd edn (Cambridge: Sky Publishing Corporation and Cambridge: Cambridge University Press)

Carr M H 1984 *The Geology of the Terrestrial Planets* (Washington, DC: National Aeronautics and Space Administration) Special Paper 469

Greeley R and Batson R M 1997 *The NASA Atlas of the Solar System* (Cambridge: Cambridge University Press)

Taylor S R 1992 *Solar System Evolution—A New Perspective* (Cambridge: Cambridge University Press)

Philippe Masson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar System: Dating of Events

The formation of the solar system began with the collapse of a fragment of an interstellar molecular cloud to a protoplanetary disk (the solar nebula) and continued through a complex process of accretion, coagulation, agglomeration, melting, differentiation and solidification to build the planets. Added to this were the secondary influences of aqueous alteration and thermal metamorphism and the tertiary effects of bombardment, collision, break-up and re-formation. Since all traces of the original materials that formed the Earth have been obliterated by bombardment and by geological processing, the rocks that are accessible for study at the Earth's crust are not representative of material that aggregated from the solar nebula. Thus the chronology of events taking place during the early solar system can only be determined by the study of meteorites and lunar material brought back to Earth by the Apollo and Luna missions.

Chondrites

Meteorites are pieces of rock and metal that fall to the Earth — fragments broken from asteroids — with varying compositions that span a wide range of planetary materials, from completely unmelted and unfractionated stony chondrites to highly fractionated and differentiated iron meteorites. These materials and the components within them carry records of all stages of solar system history.

The most significant meteorites for early solar system chronology are the chondrites, which are the most primitive of all meteorites, having experienced only mild thermal or hydrothermal metamorphism since accretion into parent bodies. Chondrites are composed of high temperature components — CAIs (calcium, aluminum-rich inclusions) and chondrules (spherical to subspherical assemblages, up to ~5 mm in diameter, of iron- and magnesium-rich silicate minerals that were partially or totally melted prior to parent-body accretion) — set in a matrix of fragmented chondrules mixed with sulfides, metal and minerals formed at lower temperatures (clay minerals, carbonates, sulfates, organic matter). The CAIs are refractory inclusions (up to ~1 cm in size) of spinel, hibonite, melilite, etc.

Radioactive isotope dating

The principle behind radioactive isotope dating is the fixed rate with which an unstable radioactive isotope (the parent) decays to a stable isotope (its daughter). The time taken for a radionuclide to decay to half its initial abundance is the half-life ($T_{1/2}$) of the system. The half-lives of several radioisotope systems used for chronology are given in table 1. In practice, absolute abundances are not measured; rather, the isotopes are normalized to the amount of a related stable isotope, and isotope ratios measured.

Table 1. Radionuclide decay systems used for solar system chronology.

Parent	Daughter	Half-life
^{147}Sm	^{143}Nd	106 Gyr
^{87}Rb	^{87}Sr	48.8 Gyr
^{238}U	^{206}Pb	4.45 Gyr
^{40}K	^{40}Ar	1.25 Gyr
^{235}U	^{207}Pb	0.70 Gyr
^{41}Ca	^{41}K	0.15 Myr
^{26}Al	^{26}Mg	0.73 Myr
^{10}Be	^{10}B	1.5 Myr
^{60}Fe	^{60}Ni	1.5 Myr
^{53}Mn	^{53}Cr	3.7 Myr
^{107}Pd	^{107}Ag	6.5 Myr
^{182}Hf	^{182}W	9 Myr
^{129}I	^{129}Xe	16 Myr

Several isotope systems with different half-lives are used to measure different events in solar system history. Long-lived radionuclides (given in the top section of table 1) are used to date the age of formation of meteorites. Short-lived radionuclides (lower half of table 1) date specific processes that meteoritic material has experienced. So, for example, the Al–Mg and Ca–K systems are used to determine the formation interval between stellar processing of material and its incorporation into meteoritic components, while the I–Xe chronometer is used to investigate the formation of secondary salts produced by aqueous processes, and the Hf–W system is used to trace planetary core formation.

Stages in Solar System formation

The main stages resulting in formation of the Earth and the solar system can be summarized as follows in figure 1.

Aggregation of planets

Gravitational instability within an interstellar molecular cloud resulted in the collapse of a fragment of the cloud to form a protoplanetary disk. The mechanism that triggered cloud collapse is not clear but, whatever the mechanism, the collapse of the cloud and the subsequent aggregation of material must have been sufficiently fast to carry a complement of 'live' short-lived radionuclides into the protoplanetary disk. Evidence for the speed of this process comes from the presence of ^{26}Mg (from the decay of ^{26}Al) within CAIs in chondritic meteorites. It has been generally accepted that freshly synthesized ^{26}Al and other short-lived radionuclides (such as ^{41}Ca) were injected into the molecular cloud fragment from a

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

supernova just prior to solar system formation (indeed, the supernova is assumed to have triggered cloud collapse). The occurrence of ^{26}Mg in the inclusions shows that the CAIs formed while ^{26}Al was still 'live' in the solar nebula, that is agglomeration took place over a very short timescale, < 3 Myr. The ^{41}Ca–^{41}K chronometer, with even shorter half-life, implies even more rapid formation of CAIs, with an interval between nucleosynthesis and agglomeration of < 0.3 Myr. Absolute dating of the components within meteorites uses the U–Pb isotope system, which is very precise, and places the time of formation of CAIs as 4566 Myr ago. Chondrules, which are the most abundant component within chondrites, show little evidence for live ^{26}Al, implying that the chondrule-forming process took place ~2–3 Myr after the formation of CAIs. The aggregation of interstellar dust ($\leq$ 0.1 μm in diameter) into increasingly large bodies, eventually forming kilometer-sized planetesimals and culminating in the asteroids and planets, took place over a time interval of some 8 Myr following formation of the CAIs. The chronology of these different events is shown schematically in figure 1.

Note, however, that use as a chronometer of isotope abundance differences between chondritic components is dependent on the models of short-lived isotope formation. Over the past few years, an alternative mechanism for short-lived radionuclide formation has been proposed, whereby the species are formed by nuclear reactions in the energetic X-wind of a young stellar object. In this scenario, no chronological significance can be attached to the lack of daughter products from short-lived isotopes in chondrules — the effect is spatial, rather than temporal.

Differentiation

Igneous activity on differentiated stony asteroids is traced by data from ^{53}Mn–^{53}Cr chronometry. For the basaltic achondrites (a group of meteorites whose parent body is believed to be the asteroid 4 Vesta), core formation apparently occurred < 4 Myr after CAI formation; subsequent volcanic activity on the asteroid continued for a further 4 Myr or so (see figure 1). This compressed timescale indicates the relatively short period that elapsed between accretion of primordial dust into small protoplanets and the onset of volcanism, melting and differentiation.

Core Formation

Once the proto-Earth had aggregated, internal heat from radioactive decay, combined with gravitational energy and collisional energy from planetesimal bombardment, kept the planet molten. As the proto-Earth cooled, reduction reactions within the convecting system resulted in the production of a metal-rich core and silicate-rich crust–mantle structure. The timescale over which core formation occurred can be deduced using several radiometric decay schemes, one of the most telling of which is the newly established ^{182}Hf–^{182}W chronometer. The strongly lithophilic ^{182}Hf is partitioned into silicates, relative to the more siderophilic W, during differentiation, and subsequent variations in Hf/W are caused by decay of ^{182}Hf to ^{182}W. Models based on the ^{182}Hf–^{182}W chronometer indicate that formation of the Earth's core

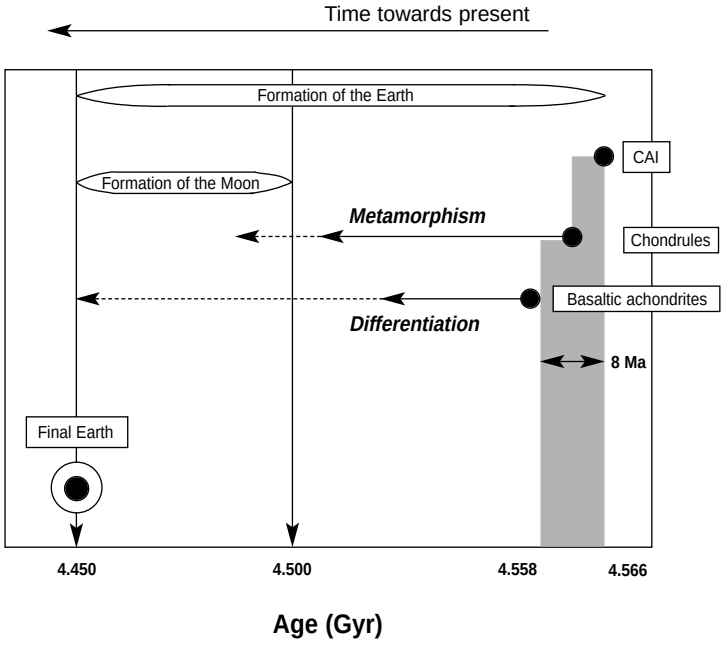


Figure 1. Schematic chronology of early solar system history (after Allègre *et al*, 1995).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

took place gradually, some 50 Myr or so after the differentiation of iron meteorite parent bodies.

Formation of the Moon

The last major event in the history of the proto-Earth was the formation of the Moon. The most widely accepted hypothesis is that in which a Mars-sized body collided with the Earth. Refinements of this hypothesis indicate that the impact occurred after the Earth's core had formed and that the crust–mantle regions of both the impactor and the proto-Earth were vaporized and then mixed during the impact. This hypothesis accounts satisfactorily for the geochemical, mineralogical and isotopic compositions of the lunar samples as represented by material returned by the Apollo and Luna missions. The timing of the giant impact event is set at around 4.51 Gyr ago.

Bibliography

Allègre C J, Manhès G and Göpel C 1995 The age of the Earth *Geochim. Cosmochim. Acta.* **59** 1445–1456

Lugmair G W and ShukolyukovA 2001 Early solar system events and timescales *Meteorit. Planet. Sci.* **36** 1017–1026

Shu F H, Shang H, Glassgold A E and Lee T 1997 X-rays and fluctuating X-winds from protostars *Science* **277** 1475–1479

Wadhwa M and Russell S S 2000 Timescales of accretion and differentiation in the early solar system *Protostars and Planets IV* ed Boss A, Mannings V and Russell S S (Tucson, AZ: University of Arizona Press)

Monica M. Grady

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Solar System: Exploration

We are now close to the 40th anniversary of the first missions to the MOON: the LUNA 1 fly-by, which set a very low upper limit on the magnetic field of the Moon, the suicide mission Luna 2, which impacted the nearside, and the Luna 3 fly-by, which imaged the far side for the first time were launched by the Soviet Union in 1959, less than two years after the first Sputnik. These 40 yr of exploration of the solar system by spacecraft have represented one of the major scientific and technical adventures of the 20th century. Our view of the solar system is now vastly different from that provided by ground-based planetary astronomy in the first half of this century. There is no example of a major planetary mission which did not discover quite unexpected features. Each of the dots of light or marginally resolved objects which could be observed from the ground is now revealed as a diverse and complex world in its own right.

This long-term effort was initiated as a spin-off of the development of intercontinental missiles at the peak of the cold war. The commitment of the US to the first manned landing on the Moon was mainly motivated as a demonstration of the American leadership in launchers, which was challenged by the launch of Sputnik 1 in 1956 by the Soviet Union. These political and military undertones justified budgets of more than 1% of the GNP in the US during the 1960s and even more for the Soviet Union. After the APOLLO program, space budgets sharply decreased, reaching a plateau which has remained nearly level (with wide fluctuations) until recently. A new technological impetus was given by the American 'star wars' program in the early 1980s, which triggered major advances in the miniaturization of spacecraft subsystems. The 10 yr since 1989 can be considered as a transition period, with the demise of the Soviet planetary program and budgetary limitations in the US and Europe, Japan constituting an exception with its expanding planetary program.

The exploration of the solar system can therefore be separated in three phases. From 1956 to 1972, programs are dominated by political considerations and are mainly focused on the race to the Moon. From 1972 to the mid-1990s, the missions are mainly driven by scientific goals, and the emphasis is on very ambitious programs such as Viking and Voyager for NASA, or Phobos 1 and 2 on the Soviet side. The third phase is characterized by much smaller missions with a handful of experiments, the major missions of ESA's program such as CASSINI (in collaboration with NASA) and ROSETTA constituting an exception. Considerations of public interest ('outreach' is the present catch-word) became increasingly important, in particular in the US, with emphasis on such topics as extraterrestrial life, 'resources' (another catch-word) or potential catastrophes (asteroidal or cometary impacts on the Earth). The 4 m class telescopes equipped with increasingly sensitive instrumentation, the development of interferometry and adaptive optics have completely renewed ground-based planetary astronomy which plays an increasing role in the exploration of the solar system.

From the point of view of science, there is a logical sequence of investigations of solar system bodies, which will be presented in the next section. The political and programmatic hazards have strongly biased the mission definition process, so that there are significant inconsistencies in the present status of knowledge on the three major types of solar system bodies: the inner planets, the giant planets and their environments (satellites, rings, magnetospheres) and the small bodies, comets and asteroids.

The solar system is the only stellar environment which can be studied *in situ*. This unique status is unlikely to change until well into the next century (or millenium). However, with the recent discoveries of planetary companions around more than 30 stars, the solar system is now revealed as one of possibly billions of planetary systems in our galaxy alone. Its exploration can therefore be put in this more general context, so as to define what is generic and what is specific about our star and its companions. The characteristics of other planetary systems and the information from the new generation of missions to small bodies will much improve our understanding of the formation of planetary systems, one of the major goals of the quest for knowledge since the origins of science.

Physical constraints for the exploration of the solar system

The major breakthrough provided by space exploration for planetary science has been the capability to get very close to solar system bodies. Up to now, planetary missions have all relied on chemical propulsion systems derived from the missile technology developed in the years following the Second World War. The hot gases formed in the combustion chamber leave the rocket at a speed of 3–4.5 km s^{-1}. The highest velocities are obtained with liquid hydrogen and oxygen as reactants, a technology which requires cryogenic temperatures and very large systems such as the Shuttle (NASA) or Ariane 4 and 5 (ESA). The higher the velocity increment, the larger the proportion of the mass in fuel. As a rule of thumb, a single rocket stage can provide a velocity change similar to the ejection velocity. This is why three-stage rockets are needed to leave Earth, as its escape velocity is 11.2 km s^{-1}.

Even higher velocities are needed to reach directly solar system bodies. The optimum transfer is not possible all the time. These 'planetary windows' occur when the Earth and the planetary body return to the optimum relative position. After the body is reached, more fuel is needed to get into orbit, and even more to land, unless one can take advantage of the atmosphere (if any). The total velocity changes give a measure of the difficulty of the missions. These parameters are listed in table 1 for the different classes of planetary bodies, together with the repeat interval between mission opportunities, which is a major constraint for space programs. The values given

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Constraints on missions to various bodies.

	Launch velocity − escape velocity (km s^{-1})	Orbit insertion (3 day period) (km s^{-1})	Window repetition (yr)
Moon	−0.17	0.24	Daily
Venus	0.3	0.6	1.6
Mars	0.35	0.73	2.1
Mercury	3.5	4.5	1.0
Jupiter	3.0	1.7	1.1
Saturn	4.1	1.4	1.03
Asteroid (main belt)	1.3–1.7	3–5	1.25–1.5
Comet (short period)	3–3.3	0.9–1.5	5.5–6

do not take into account the inclination and eccentricity of actual orbits except for Mercury and comets, for which these parameters are so large that they severely constrain the transfer trajectories.

Table 1 clearly separates two classes of objects. The Moon, Mars and Venus are relatively easy targets, for which the launcher requirements are close to the escape threshold, and a single chemical stage is adequate even for an orbiter mission. Mercury, outer planets and small bodies require either a very large launcher or a more efficient mission scenario. A direct rendezvous with Mercury or a main belt asteroid is particularly difficult as it requires the largest on-board propulsion capabilities. A lander or entry probe mission is much easier if there is an atmosphere (Venus, Mars, giant planets, Titan), in which case it is less demanding than an orbiter mission. The return of a sample from a planetary body is the most demanding of all missions, requiring a launch capability from the object back to the Earth.

With chemical propulsion systems, the only solution for such ambitious missions is to trade mass against time using planetary swing-bys. Such 'gravity assists' can provide large velocity changes at little or no cost in fuel. The first mission of this type was MARINER 10, which used VENUS to perform three fly-bys of MERCURY in 1974. The strategies can be quite surprising: recent missions to the giant planets first visited Venus, which is closer to the Sun than the Earth. The drawback is the very long mission time: 8 yr of travel are required to rendezvous with Saturn or a comet with large spacecraft, which has an impact on mission costs and sets an interval of 15 years between the definition of instruments and the scientific results. The brilliant successes and promises of the gravity assist strategy will be described later. The limits of this approach are being reached, and new technologies such as ion propulsion must be developed for the future.

These dynamical constraints have defined to a large extent the sequence of missions to solar system bodies, from the easiest to the most difficult. Fortunately, this is also what is needed for a logical sequence of scientific investigations.

Experimental techniques

Once a spacecraft has succeeded in getting close to a planetary body, the range of possible scientific investigations is quite broad. They can be separated into two major categories: remote sensing techniques and *in situ* or laboratory techniques.

Remote sensing techniques take advantage of the whole electromagnetic spectrum to obtain information on a solar system body at distances of hundreds or thousands of km. High-energy photons such as gamma-rays and x-rays give information on individual nuclei and atoms, and hence the chemical composition. Electronic transitions dominate the UV region, with the best signatures of radicals and ions. The visible part of the spectrum provides the best spatial resolution, giving it a privileged role for photogeology and topography. At longer wavelengths, molecular vibrations and rotations dominate the infrared and submillimeter parts of the spectrum, and hence information on the molecular composition of atmospheres and the mineralogical composition of surfaces. Finally, the bulk physical characterisistics (porosity, conductivity and dielectric constant) are critical for radar observations which, together with laser echos, also provide accurate altimetric information. Radio techniques are extremely useful for characterizing the dynamics of the ionized environment, which can be studied *in situ* by particle detectors. Apart from photons, the gravitational and magnetic fields can also be measured, providing critical information on the internal structure and the state of the core.

Remote sensing techniques have been extensively used on space missions, from Earth orbit (Hubble Space Telescope, IUE, ISO) and from the ground at observable wavelengths. They can be considered relatively mature, as such experiments are now close to the limit in performance set by the statistical noise on collected photons over most of the electromagnetic spectrum. This was not the case even 10 yr ago for infrared detectors. Technological advances can still be expected in terms of mass and power budgets, but the photon noise links directly the size of the collecting optics to the science return, so that going too far down in size and mass does jeopardize the science.

In situ techniques require direct contact to the surface, the atmosphere or the ionized environment of the object to be studied. The whole range of Earth, atmosphere and plasma sciences can then be considered, but they have to be adapted to very stringent constraints in mass and power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Among the most important techniques, one can quote seismology and electromagnetic sounding for the internal structure, gas chromatography and mass spectrometry of neutral species and ions for analyses of the surface, atmosphere and environment. These experiments are improving very rapidly in terms of capabilities, making possible within a few kg investigations which required quite large laboratory experiments in the early 1980s.

On the basis of these experimental capabilities and the constraints of table 1, it is possible to define a logical sequence of scientific missions for the exploration of a solar system body.

- *The discovery mission (or missions).* This first step is a fly-by mission which provides the first high-resolution remote sensing information on the body. It reveals its lot of unexpected features and helps to define the specific scientific goals of future missions. The Luna 3 image of the far side, in 1959, was the first to demonstrate the discovery potential of such missions (figure 1), as it revealed a hemisphere drastically different from the near side, with little or no dark flat regions (the mare). Other examples are the PIONEER and VOYAGER fly-bys of the systems of giant planets, from 1977 to 1989, which discovered the bewildering diversity of the satellites and ring structures, or the GIOTTO and VEGA fly-bys of comet P/Halley in 1986.
- *The in-depth exploration missions.* The second step is a series of orbiter missions which make possible a comprehensive mapping of the planetary bodies by a wide range of remote sensing techniques. These missions provide the first global view of the body and make possible the selection of areas of interest for *in situ* studies. VIKING orbiter, in 1976, around MARS, and GALILEO, which has orbited JUPITER since 1995, are good examples of such missions.
- *The in-situ studies missions.* This third step is provided by modules which land on the planet, providing information on its atmosphere, if any, during the descent. Such missions are required to obtain detailed and direct information on the internal structure and composition of the body. Only a few points can be characterized (eventually traverses with a rover or balloon). However, the *in situ* results provide 'ground truth' which calibrates the remote sensing techniques, hence improving our knowledge for all the regions of the body.
- *The sample return mission.* Because of the mass and power constraints, many critical investigations cannot be done *in situ*, even accounting for technological advances. The high-precision determination of isotopic abundances is particularly critical, as it provides the only absolute chronometers and information on solar system formation processes (isotopic anomalies). The return of samples to the laboratory is then required. The potential of such missions has been demonstrated by the wealth of data obtained from the lunar samples returned from 1969 to 1983.

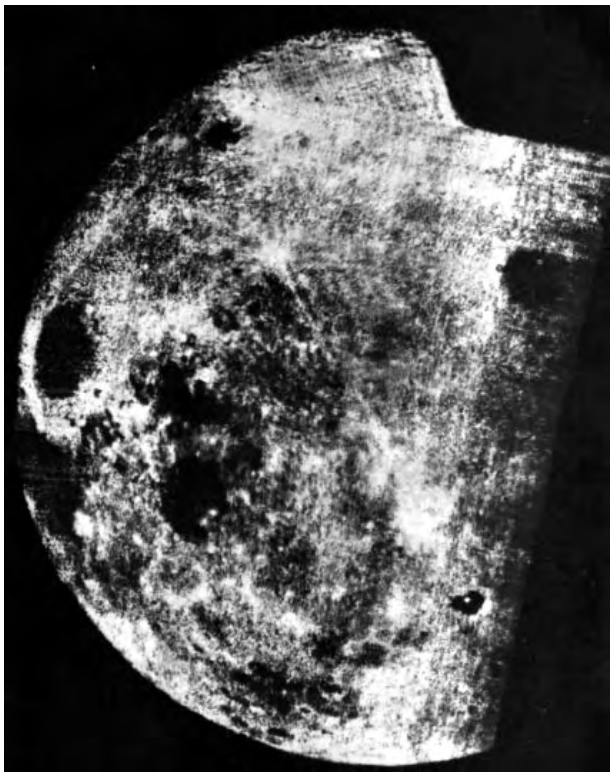


Figure 1. This image of the far side of the Moon obtained by Luna 3 is the first example of the discovery potential of space missions for solar system exploration. The far side, which can never be observed from the Earth, appeared completely dominated by the bright continents which constitute only 60% of the near side.

- *The field geology mission.* Given the enormous cost impact of a manned mission, such a step cannot be proposed on the basis of science alone. If a manned mission is decided for political or technological reasons (the Apollo program to the Moon, possibly a future lunar base or manned Mars mission), there are significant advantages in having a field geologist on site, such as improved procedures for the selection and documentation of samples.

In this sequence, each step is required as a stepping stone for the next. The comprehensive coverage defines the regions of interest for *in situ* studies, which in turn prepare the selection of sample return sites. This sequence also ranges from the easiest (fast fly-by of discovery missions) to the most demanding in terms of mission capabilities (orbiter, then landers, then sample return, then a huge gap to manned missions). It is therefore not surprising that it has been followed for most solar system objects, the exceptions resulting from political considerations or from a specific favorable condition such as the presence of an atmosphere, which greatly helps for landers or entry probes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The inner planetary bodies

Apart from the Earth, there are four large inner planetary bodies: Mercury, Venus, Mars and the Moon. As we have seen, the Moon, Venus and Mars are the easiest to reach in the solar system, in particular the Moon which is only 400 000 km away. Therefore all planetary missions during the first 20 yr of the space era have been dedicated to inner planets.

The Moon is unique in the history of PLANETARY EXPLORATION. With its proximity and modest mission requirements, it was the obvious stepping stone to other planets. In 1963, J F Kennedy defined the landing of an American on the Moon before the end of the 1960s as a national goal, demonstrating the US leadership in rocket technology (and hence nuclear strike capabilities) as a response to the Soviet achievements in space (the first artificial satellite, in 1956, then the first man in space, Yuri Gagarin, in 1961). This greatly accelerated the sequence of missions: after the first fly-bys, in 1959, only 3 yr separated the first lander (Luna 9) and orbiter (Luna 10) from the first manned landing and sample return, Apollo 11, in July 1969. The sixth and last Apollo landing, in 1972, and then the last automated Soviet sample return, in 1976, marked the end of the first phase of lunar exploration. The 'lunar jeeps' implemented by the Apollo program and two soviet automatic rovers have traveled several tens of km on the lunar surface. 400 kg of rocks and soil samples have been brought back from nine sites, six from Apollo and three with Soviet robotic missions.

The wealth of science results from this hectic sequence of missions is impressive, and the Moon is indeed in many respects the best-known solar system body apart from the Earth. However, the 'race to the Moon' left much on the roadside. Indeed, the orbiter missions were mainly focused on the characterization of potential landing sites and the remote sensing techniques of the 1960s were much less sophisticated than what is now available. Only a few per cent of the surface were covered at a resolution of less than 10 m, and there were big gaps in information on the mineralogy, chemistry and internal structure (the Apollo seismometers were quite close to one another). The sample distribution is extremely biased, with very little continental material and no sites at high latitudes or on the far side. These gaps are now being filled by a new generation of lunar missions. Two small American missions, CLEMENTINE in 1994 and LUNAR PROSPECTOR in 1998, have much improved the global characterization of the surface, with altimetry, color imaging, chemical mapping, determination of the magnetic and gravity fields. Lunar Prospector has provided convincing evidence for significant ice crystal deposits in the very cold regions near the lunar poles. The next step is the Japanese mission LUNAR-A, in 2002, with two antipodal seismometers which should set strong constraints on the size of a possible lunar iron core. The Japanese have programmed for 2003 a major remote sensing and landing mission, SELENE, which will complete the global characterization phase for the Moon. These missions will help to define a new program of lunar sample returns. A manned lunar base is unlikely to be built until at least 2010.

Venus was a high-priority target during the early phases of planetary exploration. Its size and mass make it a near twin to our Earth, and hence it has a great interest for comparative planetology. As shown in table 1, it is an easy target for fly-by and orbiter missions, and the dense atmosphere (90 bar) can be used to deploy entry probes and balloons. Indeed, the first planetary fly-by, the American mission Mariner 2, was performed as early as 1962, and the first succesful *in situ* mission, the Soviet VENERA 4 descent probe, was in 1967. Venus has been the main target of the Soviet planetary exploration effort, with 15 successful missions from 1967 to 1984: entry probes, landers with cameras and chemical analyzers, and finally balloons deployed by the VEGA missions on the way to comet P/Halley. The two major American missions have been PIONEER VENUS in 1979 (an elliptical orbiter associated with a set of 4 entry probes) and MAGELLAN, in 1989, a radar mapping mission. The thick atmosphere is nearly opaque to all but the longest wavelengths, severely complicating the remote sensing characterization of the surface. The altimetric information of PIONEER VENUS revealed a surface radically different from that of the Earth, with smaller altitude variations. The radar images obtained in 1983 by Venera 13 and 14 of the northern hemisphere and then the global coverage of Magellan in 1989 (with a higher resolution of a few hundred meters) showed extensive signatures of a volcanic activity driven mainly by hot spots (similarly to Hawaii) and not by plate tectonics as on the Earth.

The major problem for the next steps in the exploration of Venus is the extremely high surface temperature (700 K) which has restricted the probe lifetime on the surface to less than 1 h. High-temperature technologies are required for the *in situ* phases, seismic stations and rovers, or the sample return. Until they become available, the exploration of Venus will be limited to the middle and upper regions of the atmosphere, such as the Venus multiple-probe mission which is now being considered by NASA.

The exploration of Mars is mainly driven by comparative planetology objectives. Mars is nearly twice as small as the Earth, but its surface and atmosphere processes are much more similar in many respects than in the case of Venus. In terms of mission requirements, Mars is equivalent to Venus. Missions to Mars are unfortunately unique in the history of planetary exploration for their failure rate. 20 spacecraft were launched to this planet from 1964 to 1996. None of the 11 Soviet missions and only six of the nine American missions were complete successes. The discovery phase, with the two first fly-bys of Mariner 6 and 7 in 1969, revealed a surface pockmarked with craters, so that Mars was then considered as a larger version of the Moon. This view was completely reversed by Mariner 9, in 1971, an orbiter which observed for the first time the northern hemisphere and revealed the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

giant volcanoes and canyon systems which are now the trademark of the planet. This episode underlines the importance of global coverage before drawing conclusions on a given planetary body. The Moon, Mars and even the Earth, with its Pacific Ocean exhibit such a planetary dichotomy. The first major step for the global characterization phase has been the American mission Viking Orbiter, in 1976, which obtained a global coverage in the visible at medium resolution (100 m), very high resolution images and broad band spectrometry in the near and thermal infrared with a resolution of about 10 km. The Soviet mission PHOBOS 2, in 1989 (Phobos 1 failed one month after launch) was mainly dedicated to a very close rendezvous with Phobos, the largest satellite of Mars. It was lost during approach, but it obtained significant information on Mars during its two months in orbit, such as the first spectral images in the near infrared.

The recent failure of two ambitious global characterization missions had a major impact on Mars exploration. In 1992, the American Mars Observer mission was lost during orbit insertion. Its extensive set of instruments included a very-high-resolution camera, several spectrometers dedicated to the study of the surface and the atmosphere, a gamma-ray chemical mapper and a plasma package. The MARS GLOBAL SURVEYOR recovery program is based on small orbiters each carrying part of the spares of Mars Observer. The first orbiter was launched in 1996 and started close-range observations in the fall of 1998, after a long phase of progressive orbit reduction by the atmosphere of Mars. This recovery program will be completed in 2002. The Russian MARS 96 mission which failed during launch in November 1996 was arguably the most ambitious mission to Mars, with a 400 kg orbiter payload as well as four entry modules. There was a large international collaboration involving most European countries. A smaller mission, MARS EXPRESS, has been programmed by the European Space Agency for a launch in 2003. It will include five instruments derived from Mars 96 and tackle new scientific goals, such as the detection of subsurface ices with a radar.

The *in situ* phase was initiated by the Viking landers in 1976, which obtained panoramas of the surface, chemical analyses of the surface and atmosphere, climatological information on winds and seasonal cycles at two sites of the northern hemisphere. After a gap of 20 yr, an intensive phase of surface missions was initiated by NASA's MARS PATHFINDER, which deployed a small rover and obtained information on the chemical composition of soils and rocks. A landing mission is programmed by NASA for each Mars window, every 2 yr (see table 1). That launched in 1998 is targeted for the polar caps. A sample return is scheduled for a launch in 2005. Such an ambitious program requires a high level of international collaboration, with major contributions now being considered from Europe (among others, a French-provided Ariane 5 launcher and ESA's Mars Express as a relay satellite).

Mercury is a very difficult target. As the innermost planetary body, it is critical to our understanding of the planetary formation sequence in the solar system. The Mariner 10 mission, in 1974, imaged nearly half of the surface. The smooth plains and ubiquitous craters show many similarities and significant differences to lunar features. The other hemisphere is expected to be similar, although some caution is required (as shown by the example of Mars). Orbiter missions are now considered by ESA as part of its long-term program, NASA and Japan. Given its proximity to the Sun, Mercury is a very good target for a mission using solar energy to drive a plasma or ion engine.

The outer planets

The four outer planets, Jupiter, Saturn, Uranus and Neptune, have complex satellite and ring systems and extensive magnetospheres (with the possible exception of Uranus). The discovery phase has been accomplished by two pairs of missions: Pioneer 10 and 11, launched in 1973, and Voyager 1 and 2, launched in 1977. After encountering Jupiter, Pioneer 10 was set on a trajectory leaving the solar system, while Pioneer 11 was retargeted to SATURN, which it reached in 1979. The more capable Voyager spacecraft used the gravity assist strategy and a unique configuration of the giant planets (which repeats only every 200 yr) to visit all four outer planets from 1979 to 1989 (fly-by of NEPTUNE by Voyager 2).

The Pioneer missions obtained the first information on the interplanetary medium beyond the orbit of Mars. They also obtained high-quality images during the planetary encounters. These results have been superseded by the outstanding scientific results of Voyager, one of the most successful missions in the history of planetary exploration. Possibly the most remarkable discovery is the bewildering variety of the satellites of giant planets, from the lunar-like, frozen surface of icy CALLISTO to the tidally driven sulfur volcanism of IO, the innermost satellite of Jupiter, the billiard ball look of EUROPA, the tectonic processes on icy satellites of Saturn and URANUS and the thick nitrogen atmosphere of TITAN, the largest satellite of Saturn. One can also mention the fractal structure of Saturn's ring systems, the discovery of rings around Jupiter, the demonstration of the relationship of the great red spot of Jupiter to large-scale cyclonic and anticyclonic structures and the observation of similar features on Saturn and Neptune.

Two global characterization and *in situ* missions have been defined to address the wide range of scientific issues opened by Voyager for surface, atmosphere and plasma sciences: Galileo for Jupiter and CASSINI–Huygens for Saturn. The launch of Galileo, arguably the most complex spacecraft ever built, was delayed owing to Shuttle development delays and the Challenger disaster from 1983 to 1989. Galileo reached Jupiter late in 1995 and was injected in a highly elliptical orbit after releasing an entry probe in the atmosphere. Orbital changes using satellite swing-bys have been programmed to explore as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

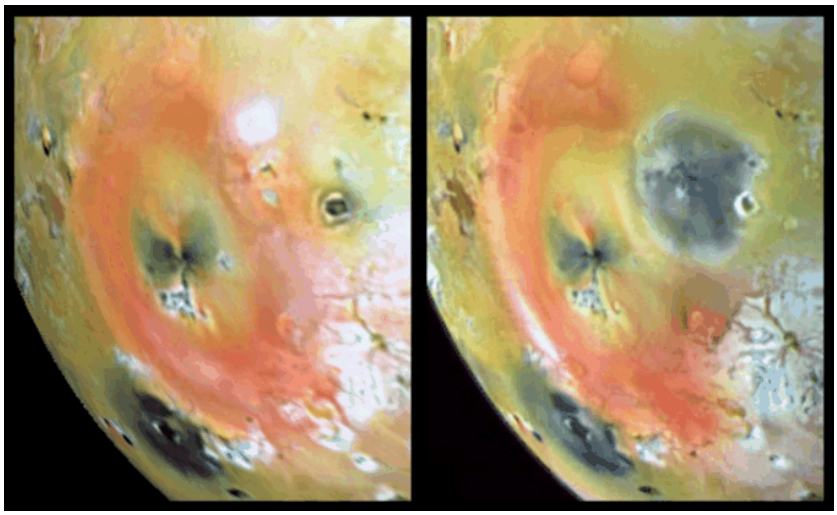


Figure 2. The surface of Io as observed by Galileo is extremely young. No impact craters are observed and sulfur volcanoes completely dominate the landscape. Major features have markedly evolved from Voyager (1981) to Galileo (1996).

much as possible of the Jovian environment. The failure of the main antenna to deploy resulted in a reduction of the available data rate from 100 kbits s^{-1} to a few 10 bits s^{-1}. In spite of this severe limitation, an outstanding science return has been obtained, such as the first profile of temperatures and chemical composition in the atmosphere of Jupiter, the first high-resolution images of the satellites (figure 2) and the discovery of an internal magnetic field for GANYMEDE, the largest of the four satellites discovered by Galileo. These successes have led to an extension by 3 yr of the mission until the end of 2000.

The next major step in the exploration of the giant planets is Cassini–Huygens, a collaborative mission between NASA, which has the responsibility for the orbiter, and ESA, which provides an entry probe in the atmosphere of Titan. This mission was successfully launched in 1996 and will reach Saturn in 2004 after swing-bys of Venus, the Earth and Jupiter. A orbital tour of four years will visit all the major satellites, explore the magnetosphere at various latitudes, characterize the ring systems and provide continuous coverage of the meteorology of Saturn. Titan, Saturn's largest satellite, constitutes a major target of the mission. The European entry probe will perform a descent of several hours in the dense nitrogen atmosphere, rich in hydrocarbons and aerosols, and is likely to reach the surface itself. Similarly to Venus, Titan's atmosphere is nearly opaque in the visible. Radar and near-infrared observations will map the surface during the 40 scheduled fly-bys of Titan.

The next steps will include small missions focused on specific scientific questions, such as the Europa orbiter mission dedicated to the characterization of a possible ocean buried under the smooth ice shell of Europa, which has been selected by NASA for a launch in 2004. Global characterization missions to Uranus or Neptune similar to Galileo and Cassini will eventually be programmed. Because of the very long travel times using classical propulsion (already 7 yr for Saturn), it is likely that such missions will require advances in technology, such as an ion drive powered by a nuclear reactor. The huge gravity well of giant planets precludes the implementation of sample return missions until well into the next century.

The small bodies of the solar system

The small bodies, COMETS and ASTEROIDS, are late comers on the scene of solar system exploration. The main science driver is to obtain information on the formation of the solar system, as these small bodies are the last survivors of the vast swarm of planetesimals which participated in the accretion of planetary bodies.

The first discovery missions were the five fly-by missions to comet P/Halley in 1986: two distant Japanese fly-bys dedicated to plasma measurements, the two Soviet VEGA spacecraft which came within 10 000 km of the nucleus and the kamikaze GIOTTO mission of ESA which approached the nucleus at a distance of 600 km. The nucleus as revealed by these missions (figure 3) was much darker and hotter than expected, showing that these objects are covered by a refractory mantle.

For main belt asteroids, the discovery phase was initiated by Galileo on its way to Jupiter, which observed two medium-size objects, GASPRA and IDA, in 1991 and 1993. The highlights of these fly-bys were the discovery of a satellite a few km in size around Ida and the observation of a magnetic field around Gaspra. These two main belt asteroids were rocky ('S' type). These results were complemented by the fly-by of Mathilde, a carbonaceous asteroid, by the American NEAR mission in 1997. The evolution of the surface of asteroids appears to be dominated by collision processes, the surfaces being saturated by craters at all scales.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3. The nucleus of Halley as imaged by the camera of the European mission GIOTTO. The very low albedo (3%) and small proportion of active areas were quite unexpected and led to the revision of the 'dirty ice ball' model of cometary nuclei in favor of a much higher proportion of refractory material.

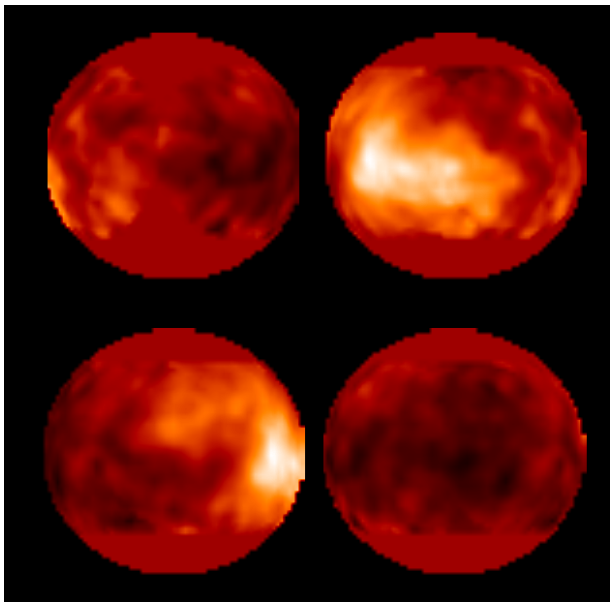


Figure 4. The surface of Titan as imaged by the Hubble Space Telescope in the near-infrared. The dark and light features reveal a significant surface differentiation. The relationship to topography and much higher resolution will become available with the Cassini–Huygens mission in 2004.

This discovery phase is far from complete. There is of course no chance of visiting each of the hundred of comets and thousands of asteroids. This comprehensive coverage is a key objective of Earth-based observations. However, the close-range observations of a few short-period comets and of the major missing types of asteroids (metallic and basaltic) would provide extremely useful information. Most missions to the outer solar system will include as many asteroid fly-bys as possible. Two cometary fly-by missions are programmed by NASA before 2008, STARDUST and CONTOUR.

The global characterization of asteroids has started with the first rendezvous with asteroid 433 EROS by the NEAR mission in early 2000. The major global characterization and *in situ* mission for comets is the ROSETTA mission of ESA, which will carry a comprehensive set of remote sensing instruments and two analyzers for dust and gas in the cometary environment. A lander provided by a collaboration of European countries will perform analyses of the physical and chemical characteristics of the cometary material and serve as a beacon for a radar sounding of the internal structure. This ambitious mission will be launched in 2003 for a rendezvous in 2011.

The ground-based discovery of a second belt of small bodies, the KUIPER BELT, beyond the orbit of Neptune, has significantly modified our view of the outer solar system. There are now only four outer planets, PLUTO AND CHARON, its companion, being considered by most planetary scientists as the largest members of this new family. A discovery mission to one or several Kuiper belt objects is now being considered by NASA, so as to obtain information on objects dynamically related to short-period comets.

The role of Earth-based observations

Earth-based observations have made major progress during the last 10 yr with the development of adaptive optics, which correct most of the atmospheric perturbations, the large ground-based interferometers (VLA, IRAM), the launch (and repair) of the HUBBLE SPACE TELESCOPE of NASA and the launch of the Infrared Space Observatory of ESA. Apart from direct imaging, much can be learned from occultations of stars about the environment of solar system bodies. Among the most important discoveries made with Earth-based observations, one can mention the rings of Uranus and Neptune, the Kuiper belt (already more than 30 objects larger than 200 km on orbits similar to that of Pluto), the detailed composition of the atmosphere of giant planets or the recent images of the surface of Titan obtained by Hubble in the near-infrared (figure 4). The discovery in 1995 of planets around other stars by ground-based telescopes sets the exploration of the solar system in the broader context of the characterization of stellar environments.

Three main objectives can be set for Earth-based observations: continuity, diversity and very-high-resolution spectrometry. They can fill the long gaps between space missions for observing variable phenomena, such as meteorology. For small bodies, Earth observations extend the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

knowledge obtained from the few visited targets to hundreds of objects, providing statistical information and criteria for selecting the next targets of space missions. The highest spectral resolutions still require large instruments incompatible with spacecraft resources. The recent observations of solar system objects by ISO have demonstrated the interest of such methods, in particular for planetary atmospheres and comets.

The future of planetary exploration

All the major classes of solar system objects have been observed at close range, with the exception of Kuiper belt objects. Past, present and programmed missions should go a long way to complete the global characterization phase: SELENE for the Moon, the MGS series and Mars Express for Mars, Galileo for the system of Jupiter, Cassini–Huyghens for the system of Saturn, ROSETTA for short-period comets and then a Mercury orbiter mission. After the turn of the century, the emphasis will progressively shift to *in situ* and sample return missions. After 30 yr of very large programs, it is however likely that much will be done by smaller missions (the 'faster, better, cheaper' approach first implemented by NASA with its Discovery program) with a limited set of instruments. This new type of mission has been made possible by the technological developments triggered by the Strategic Defense Initiative (miniaturization of subsystems) and the growing industrial capabilities worldwide for building cheap spacecraft. The short development and implementation time (4–5 yr) provides more flexibility, increases the frequency of missions within a given budget and gives a chance for the latest instrumental developments to be used. The reduction of travel times is a major constraint for such missions. It requires the development of high-capability propulsion systems, such as the ion and plasma drives, with ejection speeds 10 times higher than those of classical rockets. Small test missions of this type have been approved by NASA, by Japan and by ESA, so that this technology should be fully demonstrated by 2005. A balance will eventually be reached, as the ambitious science objectives of the *in situ* and sample return phase cannot all be reached with such missions, which are also more risky as demonstrated by a series of minor and major problems. There will still be a need for large-scale international missions such as that building up for a Mars sample return before 2010.

The discovery of other planetary systems (EXOPLANETS) will eventually link solar system exploration to interferometric missions able to image other planetary systems. Indeed, once this is achieved, it is likely that exploration missions to the closest stars with planetary systems will be considered.

Yves Langevin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar System: Formation

Cosmogony tells us about the origin and evolution of our SOLAR SYSTEM in general, and of the planets in particular. Planet formation seems to be a by-product of star birth. Current models have indicated the births of stars observed now occur in almost the same conditions as the formation of the Sun and its planets 4.55 billion years ago.

Recent observations of what we interpret as PROTOPLANETARY NEBULAS (the first has been seen around β Pictoris, see figure 1), together with the detection—although indirectly—of planets orbiting other stars (more than 20 already known; see EXOPLANETS) confirm our current ideas. Very young stars have been observed by the Hubble Space Telescope, the new VLT (the European Very Large Telescope) and space infrared telescopes (like ISO). Some young stars seem to be naked while others are dressed by disks of gas and dust which probably live for only a few million years; this is the most interesting period for us here, because we think that our own planetary system formed from such a disk.

The study of meteorites, of lunar rocks (brought back by the Apollo astronauts and Soviet automatic machines) and of interplanetary dust (analysed by space probes which have encountered comets, e.g. Giotto when it visited the nucleus of Comet Halley in March 1986) has allowed us to give a precise value for the age of the solar system and to better understand the processes which lead to formation of planets from an interstellar cloud. Among the difficulties of modelling such a process a particular problem is how to separate fundamental general mechanisms from fortuitous local ones; comparison with other planetary systems would be very useful. This is a very rapidly developing field of study.

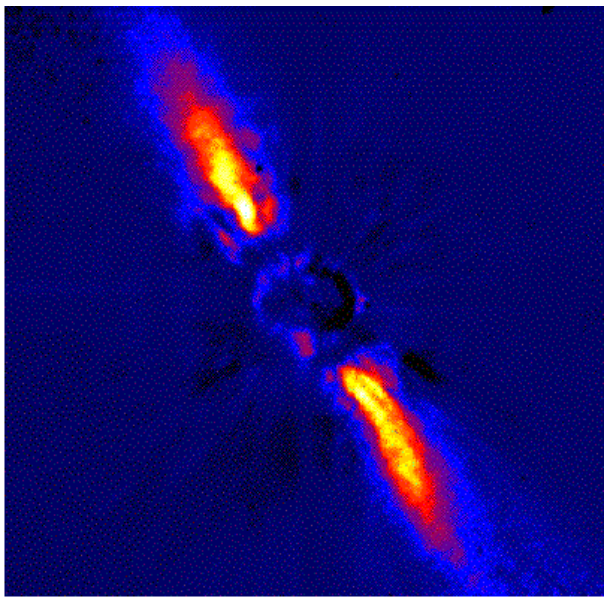


Figure 1. The beta Pictoris protoplanetary disk. Image taken with the 3.6 m ESO telescope using the adaptative optics system Adonis by David Mouillet (Observatoire de Grenoble) and colleagues; high resolution allows us to observe the disk at less than 25 AU from the star, and to distinguish the warping of the disk which could be induced by gravitational perturbations of a massive planet on an inclined orbit. (See http://www.iap.fr/LaboEtActivites/ThemesRecherche/BetaPic/images) **This figure is reproduced as Color Plate 58.**

The solar nebula

General ideas about the formation of the solar system arose in the seventeenth century, when Copernicus' heliocentric model became credible after Galileo's observations and Kepler's laws of planetary motion; the origin of this planetary system then became a subject of scientific research. One of the firsts models was proposed by René Descartes in 1644: primordial matter, rotating globally as a fluid, shows vortices; most of the matter may then accumulate in the center of a large central vortex and there the Sun forms; at the other vortices planets form and, due to the global rotation, they orbit the Sun. Although this explanation is obsolete, it contains a fundamental element which will be found in almost all subsequent models: the formation of the planets is linked to the formation of the Sun.

In the following century, Emmanuel Kant, in his *Natural Philosophy* (1755), proposed an explanation of the origin of the solar system by condensations in a flat rotating cloud, i.e. a disk of gas and dust rotating around the center where the Sun was forming; condensations in the disk gave birth to the planets. Although some basic elements, such as the condensation mechanism, were then still unexplained, the concept of a rotating disk, what we call the *solar nebula*, is the basis of modern theories. A little later, Pierre Simon de Laplace, in his *Exposition du Système du Monde* (1796), proposed a mechanism to explain this structure where a few large planets rotate in almost circular orbits largely separated from each other: contraction of the disk increases the rotation velocity so that the centrifugal force separates matter into several concentric rings, from which planets may then form.

But this model did not explain convincingly why matter should separate into a vast amount which contracts (the Sun) and a minor quantity in which contraction is stopped, or why the latter separates into rotating rings (the planets) instead of staying as a disk with continuously varying density. In particular the Sun should rotate considerably faster than it does, because its current rotation rate is much too slow for centrifugal force to counterbalance gravity. These (temporary) failures prevented the solar nebula hypothesis from receiving serious consideration for more than a century.

There have been attempts to explain planetary formation independently from that of the Sun; the most well known is the catastrophe thesis, where a comet—following Buffon—(or, more impressively, another star—following Jeans) passed very near the Sun (already formed) and pulled out some of the solar matter which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

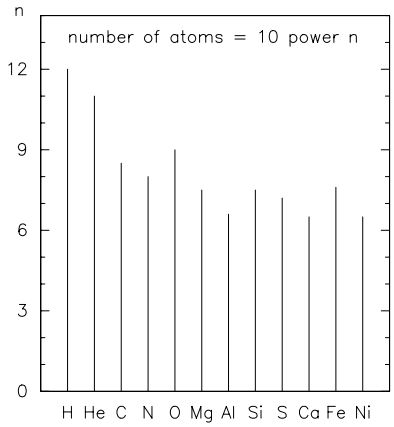


Figure 2. Cosmic abundances of the elements.

they orbited the Sun and finally formed the planets. These ideas have now been abandoned.

The nebula hypothesis, indirect descendant of the Cartesian theory, came back into favour at the end of the first half of the twentieth century, and eventually came to be the 'standard model'. Theories have been constructed to describe how gravitation could bring about the collapse of an interstellar cloud; in particular a fundamental criterion has been defined with regard to the limiting mass necessary to trigger gravitational instability, called the *Jeans' mass*. Astronomers now think that formation of the solar system began with the collapse of such an interstellar cloud, a very dense, very cold cloud where molecules might form—an INTERSTELLAR MOLECULAR CLOUD. The collapse was followed by the fast structuring of a rotating disk and the fall of the remaining matter of the cloud down this ACCRETION DISK, at the heart of which the Sun was forming.

Some properties of the nebula

Chemical elements can be placed in the well-known Periodic Table, where their position depends upon the number of protons in the atom's nucleus. The increase in this number, from the light elements (essentially hydrogen (H) and helium (He)), produced soon after the 'Big Bang', to heavier elements made through nuclear reactions in the center of stars and then ejected into the interstellar medium, is reflected in the so-called COSMIC ABUNDANCES which, at least in the vicinity of the Sun, do not vary much through space, particularly for the INTERSTELLAR MATTER (gas and grains) which will give birth to the next generation of stars.

Hydrogen and helium are very much more abundant than the other elements (figure 2); moreover, they are still gaseous down to temperatures very near 0 K (0 K = $-273.16\,^{\circ}$C, 'absolute zero'), which implies that almost 98% of the solar nebula remain gaseous even at very low temperatures. The behavior of the other elements is different; they are often assembled into molecules which can condense at much higher temperatures, up to tens or even a hundred kelvin; these molecules are called ices and their constitutive elements are essentially hydrogen, carbon (C), nitrogen (N) and oxygen (O), the latter three being the most abundant (almost 2%) apart from H and He; for example, CO (carbon monoxide), although very volatile, condenses at around 25 to 30 K, and H_2O (water) condenses at more than 110 K; it could be deduced that ices may be very frequent in all sufficiently cold parts of the nebula, i.e. the outer regions. All other elements (less than 0.3%) react above all with oxygen to form oxides; an important process is formation of silicon dioxide (SiO_2, silica) with which other elements (such as magnesium (Mg), sulphur (S), iron (Fe), aluminum (Al) or nickel (Ni)) react to form silicates. These molecules condense at high temperatures, up to 1000 degrees, and are called refractories (which are therefore essentially silicates and metals), and which therefore could exist as solids even in the innermost parts of the nebula.

Let us compare the planetary compositions with the nebula. Telluric planets are essentially refractories: a more or less large metallic nucleus covered with a thick mantle of silicates, the amount of ices and volatiles being almost negligible. Therefore, the Earth could form in the solar nebula only if an Earth's mass of refractories was available. However, refractories contribute only 0.3% to the nebula's mass (i.e. around 1/300); then we need 300 Earth's masses (M_E) of the nebula to form the Earth; the same estimates can be done for Mercury ($15M_E$), Venus ($300M_E$), Mars ($30M_E$) and the asteroids ($0.15M_E$). Jovian planets are different and contain essentially hydrogen and helium, with a small proportion of ices and refractories (probably more in the case of Uranus and Neptune than for Jupiter and Saturn), which is not far from nebula abundances; this implies that the nebula mass in the region of formation of the Jovian planets can be estimated respectively as $1000M_E$ (Jupiter) and $500M_E$ for Saturn, Uranus and Neptune. When we add all these values, we obtain around $3000M_E$, which is 1% of the mass of the Sun. Even if the nebula did not end steeply at Neptune's orbit, taking into account the nebula mass corresponding to the KUIPER BELT would not significantly increase this value. Such a nebula mass of even several per cent of the solar mass is a model of a light NEBULA; there also exists a model where this mass can increase up to several tenths of a solar mass and even one solar mass; this is a massive nebula. Until now, observations have not allowed us to clearly decide between the two models, but here we will discuss the light nebula model.

Observations of nebulae around other stars

To our eyes, regions of star birth are very dark: they are interstellar clouds of high density (high compared with the common interstellar medium) where dust and grains are sufficiently numerous to hide the stars which lie behind.

Fortunately, radio telescopes allow us to observe inside these clouds and, particularly, to verify that these dark clouds are very rich in molecules. The most easily observed is CO, but the most abundant is dihydrogen,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

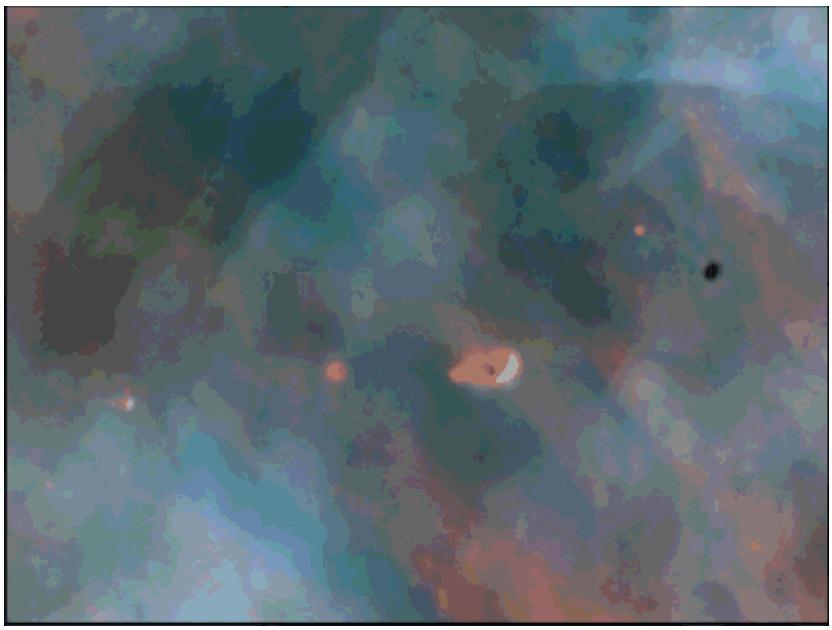


Figure 3. Proplyds in Orion. A Hubble Space Telescope view of a small portion of the Orion Nebula star-birth region; several very young stars are surrounded by gas and dust which are possibly protoplanetary disks, or 'proplyds', that might evolve on to agglomerate planets; the field of view is only 0.14 light-years across; image taken on 29 December 1993 with the HST's Wide Field and Planetary Camera 2. (STScI-PR94-24b; credit: C.R. O'Dell/Rice University—NASA; see also http://oposite.stsci.edu/pubinfo/) **This figure is reproduced as Color Plate 59.**

H_2. The measured temperatures are very low, down to 10 K. This extreme cold has two origins: first, the infrared radiation emitted by the molecules escapes easily out of the clouds, an efficient mechanism for exhaust of heat; second, dust prevents star light and heat from penetrating. We observe that gas is concentrated in a disk rotating around a central mass, an indication in favor of Kant- and Laplace-like models. Moreover, the rotation of the disk obeys Kepler's laws, so that we can estimate the value of this central mass. In the internal regions, which have almost the dimension of our planetary system, this value barely exceeds one solar mass; this looks good for the light nebula hypothesis, nevertheless it is possible that the actual disk extends beyond the visible part.

In the beginning the molecular clouds were very dark and cold. At a certain time, the gravitational instability was triggered so that a central source of energy 'switches on', surrounded by dark dust which, when warmed, emits infrared radiation.

Later, the newborn star and its disk become progressively clear of the cloud, allowing visible light emitted by the star to travel very large distances, to the Earth for example. The Hubble Space Telescope, as well as the first unit of the VLT, due to their very high angular resolution, have observed such disks, for example in the great molecular cloud in Orion where numerous structures have been identified as accretion disks (named *proplyds*, for 'proto-planetary disks'; see figure 3), although classical observations showed only simple stars, known for a long time as T TAURI STARS from the variable star of this name in Taurus. The luminosity of this class of stars varies because of their very turbulent environment, and they are often identified through spectral lines characteristic of the surrounding gas.

During this early stage of a star's life, the star itself ejects large quantities of gas in every direction. This stellar wind progressively blows the disk away, and we observe a class of T Tauri stars which have no disk: we call them 'naked' T Tauri. A chronology has been established; emerging from the cloud to the blowing away of the disk (i.e. the lifetime of the solar nebula) occupies a time interval of only some million years—a mean estimation is around 5 million. This corresponds to only one-thousandth of the age of our solar system, and this is during this very small fraction of time that the formation of the planets occurs.

From dust to planetesimals

The composition of the solar nebula was almost 99% gas (hydrogen, helium and volatiles) and 1% solid grains; these latter are very tiny, of the order of 0.1 μm, but they are the most interesting objects from the point of view of the history of solid matter in the solar system.

These grains already had a long history; born in the atmospheres of giant red stars by condensation of the chemical species (depending on the conditions of temperature and pressure), they progressively acquired a three-layer structure (figure 4): a nucleus of refractories (silicates and metals), a mantle of carbonaceous materials,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

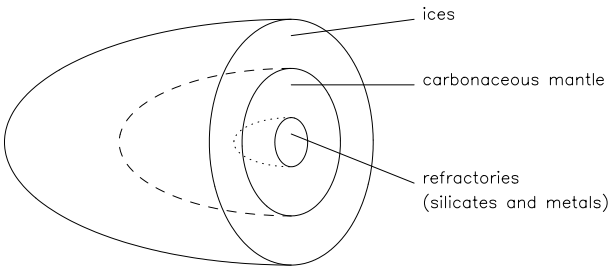


Figure 4. Model of an interstellar grain.

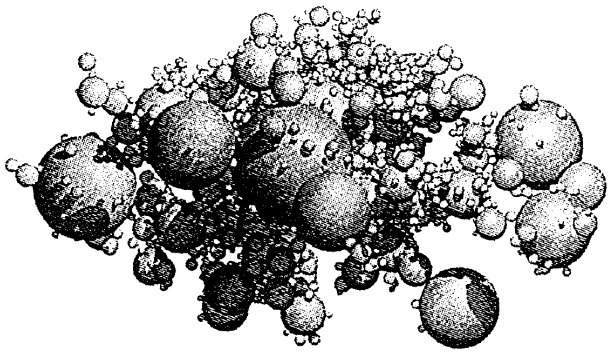


Figure 5. Numerical simulation of an aggregate (from Blum *et al* 1994).

which can also be considered as refractories but condensing at lower temperatures (around 500 to 400 K), and which have suffered physico-chemical transformations due to irradiation by cosmic rays, for example; and finally this mantle is covered by ices. The main ice species is amorphous H_2O, a form different from the crystals we know on Earth and which exists only at very low temperatures; several other molecules can be found—like CO, carbon dioxide (CO_2), methane (CH_4), ammonia (NH_3) and so on—in proportions which depend on the place where ices form because the ability of H_2O ice to incorporate them varies with temperature.

The hydrogen and helium molecules of the cloud move randomly and collide with each other due to Brownian motion. Grains, although very tiny, are real giants in comparison with these simple molecules; they are nevertheless slightly knocked and then gain some thermal velocity. These grains therefore encounter each other from time to time and, as we assume that their surface is somewhat sticky, aggregate into larger structures. Because the mass of these aggregates is of course heavier than the masses of individual grains, their thermal velocity decreases when they grow; when their size increases to several μm the influence of shocks becomes negligible and the process stops.

At least it would stop if there was not the turbulence of the gas, because the solar nebula in which grains and aggregates exist and grow is a very turbulent medium. Micrometer-sized particles are indeed sufficiently small to stay strongly coupled with the motion of the gas in which they live, and are pulled by vortices of the turbulence. Then, when two vortices encounter each other, pulled particles can come into contact and, as long as the relative velocity is sufficiently low, form larger and larger aggregates. Nevertheless, such growing aggregates uncouple progressively from the motions of the gas, because more and more powerful vortices are necessary to move them. Their width may then attain several centimeters. Figure 5 presents an example of such a particle, which can be described as a kind of porous fluffy aggregate, and is therefore of low density. Note that the porosity of the flakes may decrease during collisions when the relative velocity becomes sufficiently high to induce an increase in compactness and density; moreover, the velocity may attain a critical value where particles begin to break instead of sticking together. At this point, turbulent aggregation is no longer efficient.

Fortunately, another mechanism takes over. Centimeter sized particles, once decoupled from the turbulence of the gas, fall down to the equatorial plane of the nebula, and their velocities follow Kepler's laws. But gas moves more slowly because the effect of gravity is partially compensated by its internal pressure, which increases toward the interior. Therefore the flakes have to move against a kind of contrary wind, thus they lose energy and velocity, and their motion becomes a slow spiral toward the center. This spiral motion is the last step where the biggest flakes catch and 'eat' the others. The resulting size of the particles is estimated to be around several tens of meters.

These 'super-grains' form a very flat disk. It has been shown that such a disk is not stable if its constituent particles do not have sufficiently high relative velocities: otherwise, it fragments into condensates which orbit around the center. However, this is precisely what we need, at least in the light nebula model proposed by the Russian physicist V S Safronov and, later, the Americans P Goldreich and W R Ward. This is called a gravitational instability, which results in a new disk composed of the 'lumps' formed during the condensation: each lump has an approximate size of 5 to 10 km; these are the PLANETESIMALS, really the bricks for building the planets.

As has been said, the scenario of the formation of the planetesimals is linked to the light nebula hypothesis and the corresponding relatively thin disk. The massive nebula model gives a different evolution: there is neither flat dust disk nor gravitational instability of this disk, instead it is probably the entire nebula (not only its dust) which becomes unstable and fragments into large condensations of gas and dust from which planets form. But observations seem more and more to indicate that the four giant planets could have a rock-ice core and that these central masses each have about the same quantity of refractories (around one-tenth of Earth mass). This would be relatively easy to explain with the planetesimals, i.e. in the light nebula hypothesis, but much more difficult with the great condensates of the massive nebula model.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This is one of the arguments which convinces a majority of astronomers of the validity of the planetesimal scenario; and this is the scenario which is considered here. But how can we date its different steps?

The age of the chondrites and isotopic anomalies

When searching for the time-scale of the origin of the solar system, it would be useful to have small bodies whose composition and internal structure have not been modified by geological processes, such as occur inside the main planets by differentiation of minerals. On Earth, the most ancient rocks are 3.8 billion years old (sedimentary rocks, Isua, Greenland), and the most ancient minerals (zircons, 4.2 to 4.3 billion years) are the only remnants of continents from this epoch; but both had already lost the memory of the organization of primitive matter from which our planet formed. Some of the lunar rocks brought back from the Apollo missions date from 4.4 billion years, but they are transformed too. We hope that samples from comets and asteroids will be at our disposal relatively soon; fortunately every day Earth receives extraterrestrial matter in the form of meteorites, among which CHONDRITES are the most frequent (almost 85%) and the most primitive and therefore the most interesting for us here. Chemical and isotopic analysis of chondrites has led to two conclusions: (a) their age is almost 4.6 billion years, given with a good precision; (b) their composition is almost the same as in the cosmic abundances and then in the solar nebula, at least for refractories (because they contain few ices and very little gas; nevertheless, variations of elementary ratios among chondrites are very much less compared with the other types of meteorites, whose minerals come from differentiated parent bodies).

Sometimes, chondrites have very refractory inclusions, in which we can find particularly calcium (Ca) and aluminum, and even the most refractory solids (for example Al_2O_3). Mineralogical analysis of chondrites reveals the presence of unstable isotopes, some of which have a very long half-life, for example, rubidium-87 which decomposes into strontium-87 with a half-life of 49 billion years. This allows to measure the ages of very ancient minerals—in fact, this is the 'solidification age', i.e. the time elapsed since the agglomeration of the minerals. By this method, and other analogous ones (like lead–lead), ages of refractory inclusions in the Allende meteorite (which fell at Pueblito de Allende, in Mexico, in 1968) have been measured: these ages are the greatest ever found in the solar system, and these inclusions indeed represent the first bodies which formed at the beginning of our history; the age of the solar system is then 4.560 billion years. On the contrary, the ages of chondrites without refractory inclusions is around 4.555 billion years, that is 5 million years after the formation of the refractory inclusions. But 5 million years is precisely the lifetime of the solar nebula we have discussed previously: the inclusions would therefore have formed during the initial process of the fall of the cloud downward into the nebula, while the chondrites themselves formed at the end of the life of the nebula.

Finally, carbonaceous chondrites (one of the three most important groups of chondrites, the composition of which seems the nearest to the solar one) comprise minerals which survived even the thermal processes inside the nebula; this brings us back to the interstellar grains wandering through the Galaxy before the birth of our solar system. This is the case of microscopic grains of silicon carbide and diamond which have been studied only recently using new techniques: although quantification is not yet possible, it is hoped that the abundances of the different isotopes will allow us to ascertain their great age.

ISOTOPIC ANOMALIES can give us other important information. Let us consider for example the three stable isotopes of oxygen: ^{16}O, ^{17}O and ^{18}O; ^{16}O is the most abundant one on Earth and, more generally, in the universe; however, ^{17}O and ^{18}O ratios with ^{16}O vary enormously from place to place in the Galaxy, because isotopes come from nuclear reactions inside stars and each star produces its own isotopic ratios depending on its mass and history. Let us take as reference the isotopic ratios in water of the Earth's oceans: a plot of $^{17}O/^{16}O$ versus $^{18}O/^{16}O$ shows clearly that all terrestrial matter lies on a line of slope 1/2 (because in the Earth's gravitational field, any fractionation of the ^{17}O and ^{18}O isotopes with respect to ^{16}O is proportional to the difference of their masses, that is $^{17}O/^{16}O = 1/2\ ^{18}O/^{16}O$), called the *fractional line*, and the global composition of the Earth plus the Moon—as for ordinary chondrites and Martian SNC meteorites—is not much different (figure 6). Besides, the refractory inclusions of the carbonaceous chondrites have a clearly different behavior: they lie along a line of slope 1, which implies the presence of a pure ^{16}O component in these inclusions. There were therefore several oxygen reservoirs in the nebula, coming from several pre-solar sources; among which one, maybe a supernova, had furnished almost pure ^{16}O. But what should have been oxygen's fate if all the matter of the nebula were entirely gaseous (i.e. have been sufficiently hot everywhere) at a given time? In this case, all isotopes would have been mixed into a 'hot soup', and we could not now distinguish between the initial reservoirs. An important conclusion can then be made: the refractory minerals of carbonaceous chondrites never did evaporate, and so kept the memory of their pre-solar history in the interstellar grains.

The carbon isotopes ^{12}C and ^{13}C tell the same story; in comets, it is even clearer.

The final step to planets

The basis of the scenario is that any planetary mineral passed through the planetesimal stage during the formation process of the Sun and the solar system in the solar nebula. It is highly probable that the planetesimals had primary compositions varying with distance from the center. In the internal regions, where the terrestrial planets formed, planetesimals consisted mainly of rocks and metals but, farther, where we now find the giant Jovian planets, they were 'icy', i.e. they contained much ice and volatiles.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

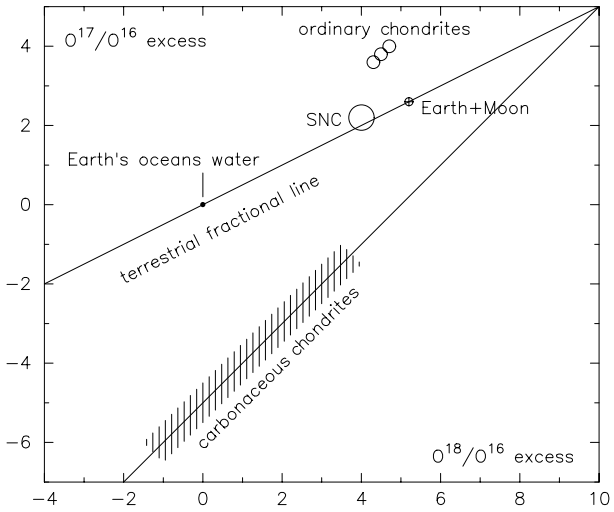


Figure 6. Isotopic anomalies for oxygen in Earth's materials and in meteorites.

The sizes of the planetesimals are estimated to be of the order of some tens of kilometres, their masses were therefore of the order of 10^{15} kg (a thousand billion tons). To build the Earth with such bricks, we would need about 10^{10} of them, i.e. ten billion. Here is the disk of planetesimals: several tens of billions of bodies only in the region of the terrestrial planets, orbiting the proto-Sun nearly circularly almost in the same plane. In each region, they move on almost concentric orbits with almost the same velocity but, given their enormous number, collisions frequently occur and, when they collide, they encounter each other with a slow relative velocity, of the order of some meters per second. During such low-velocity collisions, a kind of 'gluing' occured and therefore formation of greater and greater aggregates. This is how a population of aggregates of different sizes was progressively established, where the size distribution is simply the result of a game of probabilities: the greater the aggregate, the smaller the probability of forming it in a given time. But an important process will perturb this simple game, when the gravity of the aggregates becomes sufficiently high to attract other aggregates and planetesimals: an aggregate which initially has the luck to become large increases its efficiency to attract others; then the more it grows, more its gravity increases, and the more it grows, and so on ....

We have here an unstable process, which led to runaway ACCRETION, where the largest body in a given part of the disk enlarges at the expense of the others and so forms embryonic planets. When they attain the size of the Moon or of Mercury (i.e. around 1/4 to 1/2 of the Earth), they have exhausted the building blocks in their immediate neighborhood and their growth stops. This seems to have been the case in the region of the terrestrial planets, but how many Moons or Mercuries do we need to form these latter? Probably around a hundred. Then, after the runaway stage, which lasted maybe several million years, a *sparse stage* follows where a hundred planetary embryos revolve around the proto-Sun on orbits much more widely spaced than the orbits of the previous planetesimals. In this scenario, Jupiter has grown faster and accreted much of the gas of the nebula, so that its gravity perturbs the orbits of these embryonic terrestrial planets which then will cross each other; we are then in the final stage of the formation of the terrestrial planets where giant impacts occur between bodies width sizes of hundreds to thousands of kilometres. These impacts may be hard indeed, and we can see their scars on the surface of many bodies, for example on the Moon and Mercury (the impact craters on Earth and on other planets with an atmosphere have been mostly eroded away); moreover, such an impact can completely destroy an embryonic planet or, at least, fragment it into two bodies and so give birth to a 'binary planet' as, for example, Earth and the Moon (see the entry on the MOON for more details about its origin).

We know that the total mass in the asteroid belt is very much less than the masses contained in the other planetary regions, i.e. the region of the terrestrial planets as well as the region of the Jovian planets; the Kuiper Belt shows the same lack of mass. It does not seem likely that the density of the solar nebula should have decreased so much either in the asteroid zone or just beyond Neptune; so we may suppose that matter in these regions was blown away during stages after the formation of the planetesimals (asteroids and Kuiper Belt objects are probably 'fossils' of the early formation of the solar system). And indeed a mechanism still operates nowadays; for example, the resonances which cross the asteroid belt (mean motion resonances between asteroidal orbits and the orbit of Jupiter, as well as the so-called 'secular resonances' between some orbital elements of asteroids and the two largest Jovian planets) are like traps where any body inside is blown away in a relatively short time because of the planetary perturbations. The result is that the asteroidal orbits become very eccentric and so cross the orbits of the terrestrial planets (or even fall into the Sun), which may lead to catastrophic collisions with these latter and explains—at least in part—the lack of mass in the asteroidal region. In the Kuiper Belt, too, many objects have probably been ejected by planetary perturbations since the beginning of the solar system.

But stars are not all alone ...

Up to now we have considered only, although implicitly, processes occurring around the Sun, which is a single star. Our star, although accompanied by several bodies like planets or comets, is indeed alone, because its nearest neighboring star, Proxima Centauri, is more than 4 light-years away, i.e. almost 8000 times farther than Neptune. But more than half of all stars are not alone: they exist mainly in pairs, the 'binary stars', and even sometimes in threes ... or more. Two questions then arise: can

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

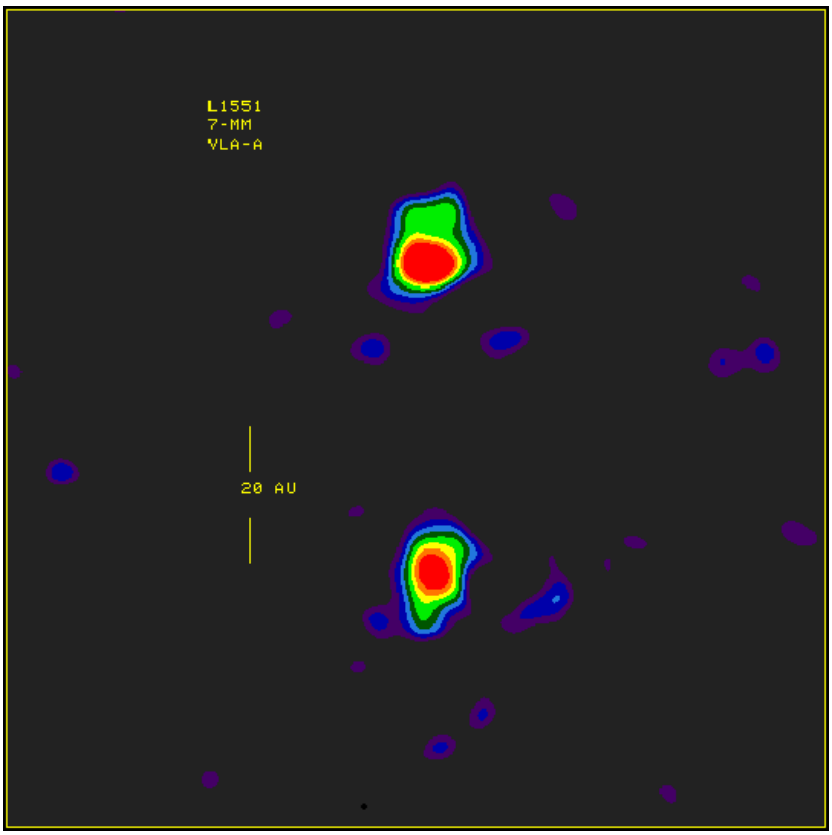


Figure 7. Protoplanetary disks around each component of a young double star in a molecular cloud designated L1551, in the constellation Taurus. VLA image: the colors represent relative intensity, or brightness, of the radio emission coming from these disks, red is strong emission and blue is weak emission; the scale bar corresponds to a distance of 20 AU; for comparison, in our own solar system, the planet Uranus is about 19 AU from the Sun. (Credit: L.F. Rodriguez *et al* National Radio Astronomical Observatory.) **This figure is reproduced as Color Plate 60.**

planets form in such environments and, if yes, do long-term stable orbits exist for such planets? Recently, more and more theoretical studies seem to show that the answers to these questions may be positive. Moreover, recent observations of what we may interpret as protoplanetary disks in binary systems (see for example figure 7), as well as several examples among the indirectly detected exo-planets being in double systems (e.g. around 16 Cyg B or Gl 86 A, detected by the radial velocity method, and the gravitational microlensing event MACHO-97-BLG-41), are very encouraging.

Some open problems

Since late 1995 the discovery, although indirect, of dozens of giant gaseous planets (the first around 51 Pegasi, discovered by Michel Mayor and Didier Queloz from Geneva Observatory in Switzerland with the 1.93 m telescope of Saint-Michel de Haute-Provence Observatory in South France) very near their 'sun' (i.e. very much nearer than Mercury's orbit), such a position for the formation of such a Jovian planet previously being ruled out by the theoreticians, led to the idea that these planets actually formed at large distance then migrated to their present place through interactions with either the solar nebula (not yet evaporated) or the disk of planetesimals. Even our own solar system, which seemed so perfectly constructed (let us recall Bode's 'law' ...) that the actual positions of the planetary orbits correspond to the location of their birth, may too have been affected by this kind of mechanism. For example, some new theoretical models describing the distribution of the Kuiper Belt objects imply that the structure of the Kuiper Belt give clues about the history of the Jovian planets and, particularly, indicates that the latter's orbits would had progressively changed after their formation, e.g. Neptune moved outwards by several AU over some tens of millions of years, Uranus and Saturn moved outwards too but by a smaller amplitude, while Jupiter moved inwards only a little. Nevertheless, note that there exist rival explanations about the distribution of the Kuiper Belt objects.

Moreover, we may have an example of another planetary system, around the star υ And (upsilon Andromedae), studied by Paul Butler and co-workers, where three Jovian planets (of masses respectively, at least,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

0.7, 2 and 4 Jupiter masses) orbit their star at respectively 0.06, 0.8 and 2.5 AU; but here theoretical models of distant formation followed by migration have great difficulties in explaining the actual structure of this planetary system.

As a conclusion, astronomers need much more time before they can confirm a theory which presents a convincing explanation of all details of the formation and the evolution of planetary systems in general and of our solar system in particular.

Daniel Benest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar System: Nomenclature

This article describes the various schemes used for assigning names to solar-system objects with visible surfaces, other than the Earth, and their surface features. The schemes are known collectively as planetary nomenclature.

Planets and satellites

The outer planets that were discovered in modern times were named after principal figures from Roman mythology, following the tradition for the planets known in antiquity (though the first two letters of 'Pluto' also commemorate Percival Lowell, whose calculations led to its discovery). Planetary satellites, discovered from the seventeenth century onwards, were given names of figures from mythology and literature associated with that of the parent planet, a practice which has been adhered to for the many minor satellites discovered in the late twentieth century. For example, the Galilean satellites of Jupiter were named by Simon Mayr (or Marius) after four paramours of the god Jupiter; names of other female figures in the Jupiter myth have been assigned to the smaller satellites. Likewise, most of the minor satellites of Uranus are named after characters from William Shakespeare's plays, the precedent having been set when John Herschel gave names to the planet's four largest satellites in the nineteenth century.

Satellites are also given numerical designations. For example, the first two outer satellites of Uranus to be discovered were temporarily referred to as S1997 U1 and S1997 U2, indicating that they were respectively the first and second new Uranian satellites to be detected in 1997. Once their orbits were confirmed they received the permanent designations Uranus XVI and Uranus XVII, labeling them as the sixteenth and seventeenth of all the planet's satellites to be discovered.

Comets

Comets are identified by name and number. In the great majority of cases a comet's name commemorates its discoverer(s). The numbering system, similar to that used for asteroids, specifies the year and month in which a comet was discovered. In addition, a prefix identifies the comet as of short or long period, and short-period comets are given permanent numbers, in order of their discovery.

The practice of naming comets dates from the eighteenth century, after Edmond Halley demonstrated in 1705 that the comets of 1531, 1607 and 1682 were one and the same object, which became known as Halley's Comet. Similarly, in 1819 Johann Encke showed that four apparitions between 1786 and 1818 were of the same comet, now known as Encke's Comet. ('Comet Halley', 'Comet Encke', and so on, are preferred in more formal usage.) Other comets were named after their discoverers. Some comets bear two or even three names. Comet Grigg–Skjellerup, for example, was found by John Grigg in 1902, and by Frank Skjellerup at a later apparition in 1922. On the night of 23 July 1995, Alan Hale and Thomas Bopp independently discovered the comet subsequently named after them. Comet IRAS–Iraki–Alcock was discovered independently by Genichi Iraki and George Alcock, and also detected by the IRAS satellite, in 1983. Since 1995 such triple-barreled names have not been admitted. Different discoveries by the same individual(s) are distinguished by the addition of an arabic numeral: Comet Shoemaker–Levy 9, for example, was the ninth discovery by the team of Carolyn and Eugene Shoemaker and David Levy.

Numerical designations for comets were introduced by the editors of the *Astronomische Nachrichten* (AN) in 1846. In the original scheme, comets were assigned roman numbers according to the order of their perihelion passages in a particular year; in 1870 the AN editors introduced subsidiary designations in which comets were assigned a lower-case roman letter reflecting the order of discovery announcements in a particular year. Both schemes had their problems: some perihelion times turned out to be inconsistent with the numbers assigned; comets received a new designation with each apparition; some comets mistakenly had more than one letter applied to them.

Despite these problems, the systems persisted until the 1990s. In 1994 the International Astronomical Union announced (in *Minor Planet Circulars* numbers 23803 and 23804) new designations, to take effect from 1 January 1995. A comet's designation consists of the year of its discovery, a letter for the half-month in which it was discovered, and an arabic numeral indicating its order of discovery in that half-month. So A is January 1–15, B is January 16–31, C is February 1–15, D is February 16–28/29, and so on (I and Z are not used). Prefixes are assigned as follows: 'P/' for a short-period comet, 'C/' for a long-period comet, 'D/' for defunct comets (such as Shoemaker–Levy 9), 'X/' for comets for which orbits cannot be computed (or whose existence is doubtful) and 'A/' for any comet that might turn out to be an asteroid. Once the orbit of a short-period comet is well established, it receives an arabic number indicating its order of discovery; the first in this sequence is 1P/Halley.

One example will illustrate the old and new systems. Comet Pons–Brooks, with a period of 70 years, was discovered by Jean-Louis Pons in 1812, before numerical designations were introduced. When it was rediscovered in 1883 by William Brooks, it received the designations 1884 I (the first comet to perihelion in 1884) and 1883b (the second comet discovery announced in 1883). On its return in 1954, it was given the additional designations 1954 VII and 1953c. Under the new system the three corresponding designations are 1812 O1, 1883 R1, and 1953 M1; the first of these indicates that the comet was first seen in the second half of July 1812. The comet has the prefix 12P/, putting it twelfth in the sequence of established short-period comets. It is therefore referred to as '12P/Pons–Brooks'; if one wanted to specify its most recent apparition, 'Comet 1953 M1' would be the form to use. Correspondences between all the old- and new-style designations are give in Marsden and Williams (1999). See also METEOR SHOWER.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Asteroids

Like comets, asteroids are identified by name and number. The system of provisional and permanent numerical designations is similar to that used for comets (see above). Unlike comets, asteroids are given names chosen by their discoverers. The sheer number of asteroids discovered in recent times has seen a steady increase in the number of categories from which asteroid names are chosen. Names of people (from all walks of life, famous or otherwise), mythological figures and geographical names are the main sources.

When the first asteroids were discovered in the early 1800s, nomenclature was not an issue. As 'minor planets' they were simply named for classical figures, following the established practice for the major planets. In 1850 the naming of the twelfth asteroid as Victoria in honor of the reigning British monarch (though also the name of the Roman goddess of victory) sparked a controversy that persisted for decades. At the end of the century the classical rule was relaxed, and any female name was deemed acceptable. This rule too was stretched, feminine endings '-a' or '-ia' being added to geographical names, giving for example 'Washingtonia'. After the Second World War the female rule too was relaxed.

Naming of asteroids is now under the control of the IAU's Minor Planet Names Committee. The present rules actually proscribe very little. Names must be pronounceable, of sixteen letters or fewer, and should not commemorate military or political figures who died less than a hundred years previously. Other than that, discoverers have free rein; they must, however, justify the significance of the name they choose. The largest category is personal names. As well as astronomers (e.g. Hubble) and other scientists (e.g. Turing) there are composers and musicians, both ancient and modern (e.g. Mozartia, Zappafrank); other names commemorate relatives, mostly wives and children, of discoverers. Geographical names include Dresden and Hanoi, and the fictitious Neverland (from *Peter Pan*). Observatories (e.g. Goethe Link), and telescopes (e.g. Spacewatch) are also represented. Some choices are unusual or even whimsical. Swissair is named for the discoverer's favorite airline. The name Tooting, a district of London, was inspired by the asteroid's provisional designation, 1992 SW_{17} (SW17 being Tooting's postal code).

There are special rules for two groups of asteroids. The Trojan asteroids, which orbit at the L_4 and L_5 Lagrangian points of Jupiter's orbit, take their names from figures from the Trojan War legend, such as Achilles (in the L_4 group) and Patroclus (in the L_5 group). The Centaurs, which orbit in the outer solar system, are given appropriate names: the largest two are Chiron and Pholus.

The numbering of asteroids was begun by Benjamin Valz, who named the asteroid Massalia in 1852, and added the number 20, written in a circle, to indicate that it was the twentieth to be discovered. This system of permanent numbering is still used today, but with the number in parentheses rather than a circle. Thus the first asteroid to be discovered is designated (1) Ceres; the ten-thousandth is (10000) Myriostos. In the nineteenth century, the editors of the AN assumed responsibility for numbering, assigning asteroids a number on the announcement of their discovery. Confusion over the order of discovery arose as the pace of discovery quickened. Also, as some 'discoveries' failed to be confirmed, 'spare' numbers began to accumulate.

A solution was proposed in 1892 by Adalbert Krüger, then editor of the AN: henceforth new discoveries would first be assigned a provisional designation consisting of the year of discovery plus a capital letter to indicate the order of the discovery in that year. Once the orbit had been satisfactorily established, the editors of the *Berliner Astronomisches Jahrbuch* would assign a permanent number. Krüger's system, which was soon expanded to accommodate increasing numbers of discoveries each year, forms the basis of the system currently used by the IAU's Minor Planet Center. A provisional designation consists of the year of discovery plus two capital letters and, as is usually necessary, a number. The first letter indicates the half-month of discovery, in the same way as for comets (see above). The second letter indicates the order of discovery within that half-month. The letter I is not used here, allowing for 25 discoveries per half-month. For the 26th discovery in a given half-month the second letter returns to A, but with '1' appended. For discoveries 51 to 75, a '2' is appended, and so on. Where typography allows, it is recommended that these numbers be written as subscripts.

Thus, for example, the sequence of discoveries in the first 15 days of January 2000 are assigned provisional designations in the following order: 2000 AA, 2000 AB, ..., 2000 AZ, 2000 AA_1, 2000 AB_1, ..., 2000 AZ_1, 2000 AA_2, and so on. The asteroid with the provisional designation 1992 SW_{17} mentioned above—now with the permanent designation (8380) Tooting—was the 542nd to be discovered in the second half of September 1992.

Surface features

In the seventeenth century the first selenographers (makers of Moon maps) all devised their own schemes for naming lunar features, and the same thing happened in the nineteenth century when observers first mapped Mars. In the twentieth century the IAU assumed the task of standardizing planetary nomenclature, which became all the more important with the mapping of increasing numbers of planets and their satellites by space probes from the 1960s onwards. Assigned names are binomial, consisting of a generic term describing the type of feature, plus a specific name. Terrestrial placenames, names of famous people, and names from the world's mythologies and literature are the main sources of specific names. For each body, appropriate themes are chosen for each category of surface feature. See also MOON MAPS.

With the advent of the telescope, selenographers began to chart the surface features of the Moon. They assumed that the surface they saw was Earthlike, and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

that the dark areas were water, to which they gave Latin names representing among other things terrestrial weather and states of mind, for example Mare Serenitatis (the Sea of Serenity) and Lacus Timoris (the Lake of Fear). Craters were named after prominent people, a practice begun by Michael Van Langren with his map of 1645; in 1651 Giovanni Riccioli set the precedent for naming them after ancient and modern astronomers and scientists, such as Plato and Copernicus. A consistent nomenclature for the most prominent features, based on Riccioli's scheme, emerged during the nineteenth century, but for the smaller features there was no consensus and selenographers frequently assigned names of their own choosing.

A similar situation had arisen with Mars. A combination of improved telescopes and favorable oppositions allowed the first good maps to be prepared in the late nineteenth century. Several astronomers had mapped the planet's elusive features, each applying their own nomenclature. The system that would eventually prevail was that adopted by Giovanni Schiaparelli on his map of 1877, for which he used names derived from classical literature and the Bible.

The chaotic state of lunar and Martian nomenclature was a concern of the IAU from the time of its foundation in 1919. Building on preliminary work by Mary Blagg, the first IAU planetary nomenclature committee tackled the problem of lunar nomenclature. The result was *Named Lunar Formations* (1935) by Blagg and Karl Müller, the first systematic nomenclature for the Moon. This work was greatly expanded in the 1960s with the publication of *The System of Lunar Quadrants* (in four parts) by D W G Arthur *et al*. In 1958 the nomenclature for Mars was put on a firm basis when an IAU committee under Audouin Dollfus adopted a standard list of albedo features (areas appearing light or dark) based on Eugène Antoniadi's expansion of the system introduced by Schiaparelli.

With the dawn of the space age came the realization that a systematic approach to planetary nomenclature was necessary. In the 1960s, *ad hoc* committees dealt with naming the host of new lunar features imaged by the Zond, Lunar Orbiter and Apollo missions, and also with the Martian features photographed by the Mariner probes. In 1973 the IAU set up the Working Group for Planetary System Nomenclature (WGPSN), which continues to be the body with responsibility for assigning new names. It operates through six task groups, covering the Moon, Mercury, Venus, Mars, asteroids and comets, and the outer solar system.

When a body is first imaged, a few prominent features are assigned names based on an appropriate theme. Initially names are suggested by, for example, project scientists, but thereafter anyone can propose a name to the appropriate task group. When the IAU was faced with the task of naming several hundred newly mapped features of Venus, for example, members of the public were invited to submit names of women whose achievements they felt deserving of honor. The task groups are guided by various rules and conventions. Nearly all names are binomial in form: one part is a 'descriptor' (see table 1), a generic term derived from Latin or Greek that characterizes the type of feature—thus continuing the scheme used by the early selenographers; the other, which is the specific name, goes before the descriptor in some cases, after it in others. Exceptions, which bear only a specific name, are craters, albedo features (on Mercury and Mars), volcanic eruptions and plumes (on Io and Triton), and some large ringed structures (e.g. Valhalla on Callisto). Other rules prohibit the use of names of political or military leaders, and encourage the task groups to draw on a broad selection of the world's cultures.

For each body, the appropriate IAU task group is left to develop interesting and appropriate conventions for naming surface features. Some examples will suffice.

Craters on Mercury are named after celebrated figures in the arts. Examples are Beethoven, Chekhov, Ōkyo (Japanese painter), and Vyāsa (Indian poet). Planitiae take their names from words for 'Mercury' in different languages, for example Budh Planitia (Hindu) and Tir Planitia (Norwegian). The names of rupēs are those of famous ships of discovery: Discovery Rupes, for example, commemorates James Cook's flagship on his last Pacific voyage. Radio telescopes, with which important discoveries about the planet have been made, lend their names to valles; an example is Arecibo Vallis.

For Mars the lunar practice of naming craters after scientists is followed, and many large craters bear the names of those who have contributed to the study of the planet, such as Lowell, and writers who have used Mars as a setting for their fiction, such as Verne. After the Moon, Mars has more charted craters than any other body, and to avoid duplication of lunar names, smaller craters are named after towns and villages with fewer than 100 000 inhabitants. Valles are given the names for 'Mars' or 'star' in various languages if they are large (e.g. Bahram Vallis, Persian) or terrestrial rivers if they are small (e.g. Mad Vallis, in Vermont). An exception is the huge canyon system Valles Marineris, which takes the name of the Mariner probes which returned the first close-up images of the planet. The names of albedo features charted by Schiaparelli are not lost: they are transferred to nearby or coincident topological features. Thus, for example, the light-colored albedo feature Hellas becomes the impact basin Hellas Planitia.

Satellites also have their own appropriate themes for nomenclature. In the Uranian system, for example, surface features take their themes from the satellites' names: on Oberon, from tragic Shakespearean heroes (e.g. the crater Romeo); on Titania, from female Shakespearean characters (e.g. the crater Elinor); on Umbriel, from malevolent spirits (e.g. the crater Gob, king of the gnomes); on Ariel, from benevolent spirits (e.g. Pixie Chasma); and on Miranda, from characters and settings from Shakespearean drama (e.g. Verona Rupes).

Asteroids too have themes assigned to them. Gaspra, the first to be imaged from close hand, by the Galileo probe,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Descriptors for planetary nomenclature used by WGPSN. A blank in the descriptor column indicates that only the plural form is in use, and a blank in the plural column that only the singular form is used.

Descriptor	Plural	Meaning	Description	Example
Catena	Catenae	Chain	Chain of craters	Kraken Catena (Triton)
Cavus	Cavi	Hole	Hollow, irregular depression	Sisyphi Cavi (Mars)
Chaos	—	Chaos	Distinctive area of broken terrain	Arsinoes Chaos (Mars)
Chasma	Chasmata	Chasm	Deep, elongated, steep-sided depression	Avalon Chasma (Mimas)
—	Colles	Hills	Small hills or knobs	Scandia Colles (Mars)
Corona	Coronae	Circle	Ovoid-shaped feature	Heng-o Corona (Venus)
Dorsum	Dorsa	Back	Ridge	Kepler Dorsum (Phobos)
Facula	Faculae	Little torch	Bright spot	Lyctos Facula (Amalthea)
Farrum	Farra	Farro (variety of grain)	Pancake-like structure, or a row of such structures	Seoritsu Farra (Venus)
Flexus	Flexūs	Bend	Very low curvilinear ridge with a scalloped pattern	Sidon Flexus (Europa)
Fluctus	Fluctūs	Flow	Flow terrain	Tung Yo Fluctus (Io)
Fossa	Fossae	Ditch	Long, narrow, shallow depression	Nike Fossae (Venus)
Labes	Labēs	Blemish	Landslide	Ius Labes (Mars)
Labyrinthus	—	Maze	Complex network of intersecting valleys	Noctis Labyrinthus (Mars)
Lacus	Lacūs	Lake	Small plain on the Moon	Lacus Oblivionis
Lenticula	Lenticulae	Lentil	Small dark spot on Europa	(not yet used)
Linea	Lineae	Thread	Dark or bright elongate marking, may be curved or straight	Telephassa Linea (Europa)
Macula	Maculae	Spot	Dark spot, may be irregular	Akapura Maculae (Triton)
Mare	Maria	Sea	Large circular plain	Mare Ingenii
Mensa	Mensae	Table	Flat-topped prominence with cliff-like edges	Nepenthes Mensae (Mars)
Mons	Montes	Mountain	Mountain	Maat Mons (Venus)
Oceanus	—	Ocean	Very large dark area on the Moon	One instance only: Oceanus Procellarum
Palus	Palūs	Swamp	Small plain on the Moon	Palus Putredinis
Patera	Paterae	Shallow dish	Irregular crater, or a complex one with scalloped edges	Shakuru Patera (Io)
Planitia	Planitiae	Plain	Low plain	Sipapu Planitia (Triton)
Planum	Plana	Plain	Plateau or high plain	Planum Angustum (Mars)
Promontorium	Promontoria	Cape	Headland on the Moon	
Regio	Regiones	Region	Large area marked by reflectivity or color distinctions from adjacent areas	Nicholson Regio (Ganymede)
Reticulum	Reticula	Net	Reticular (net-like) pattern on Venus	(not yet used)
Rima	Rimae	Crack	Fissure on the Moon	Rima Hyginus
Rupes	Rupēs	Cliff	Scarp	Pourquoi-Pas Rupes (Mercury)
Scopulus	Scopuli	Crag	Lobate or irregular scarp	Xanthe Scopulus (Mars)
Sinus	Sinūs	Bay	Small plain	Sinus Iridum
Sulcus	Sulci	Furrow	Subparallel grooves and ridges	Samarkand Sulci (Enceladus)
Terra	Terrae	Land	Extensive land mass	Lada Terra (Venus)
Tessera	Tesserae	Mosaic tile	Tile-like, polygonal terrain	Kutue Tessera (Venus)
Tholus	Tholi	Cupola	Small domed mountain or hill	Apis Tholus (Io)
—	Undae	Waves	Dunes	Al-Uzza Undae (Venus)
Vallis	Valles	Valley	Valley	Niger Vallis (Mars)
Vastitas	Vastitates	Wasteland	Extensive plain	Vastitas Borealis (Mars)

Source: Based on Batson and Russell (1995), on-line version.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is named after a spa resort in the Crimea. Its craters take their names from other spa towns; examples are Aix, and Spa itself. Regiones, such as Yeates Regio, are named after Galileo project scientists.

Bibliography

Batson R M and Russell J F (ed) 1995 *Gazetteer of Planetary Nomenclature 1994* (US Geological Survey Bulletin 2129) (Washington, DC: United States Government Printing Office) (For an updated version, see http://wwwflag.wr.usgs.gov/USGSFlag/Space/nomen/nomen.html)

Blagg M and Müller K 1935 *Named Lunar Formations* (London: Lund, Humphries)

Marsden B G and Williams G V 1999 *Catalogue of Cometary Orbits* 13th edn (Cambridge, MA: IAU Central Bureau for Astronomical Telegrams/Minor Planet Center)

Schmadel L D 1999 *Dictionary of Minor Planet Names* 4th edn (Berlin: Springer) (available with CD-ROM) (See also the lists of numbered asteroids maintained by the Minor Planet Center at http://cfa-www.harvard.edu/iau/lists/NumberedMPs.html)

Whitaker E A 1982 *NASA Catalogue of Lunar Nomenclature* (NASA Reference Publication 1097)

Whitaker E A 1999 *Mapping and Naming the Moon* (Cambridge: Cambridge University Press)

John Woodruff

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar System: Stability

The question of the stability of the SOLAR SYSTEM was raised with the discovery of Newton's law of gravitation, published in 1687. Since then, for 300 yr mathematicians and astronomers have searched for evidence for the stability of the planetary motions, and this quest led to many mathematical developments, and several successive 'proofs' of stability for the solar system: first by Laplace and Lagrange, in the 18th century, then Poisson and more recently by the Russian mathematician V I Arnold in 1963. With Poincaré (1898), one can be surprised by this succession of proofs. 'Were the old demonstrations insufficient, or the new ones unnecessary?' In fact, all this evidence of stability was based on approximations, and it was shown recently, using numerical simulations, that the motion of the solar system is chaotic: although over 10 million years the motion of the planets can be forecast with great accuracy, there is no hope of extending these predictions beyond about 100 million years owing to the exponential divergence of the orbits.

Historical introduction

The problem of the stability of the solar system has fascinated astronomers and mathematicians since Antiquity, when it was observed that, among the seemingly fixed stars, there were also 'wandering stars'—the planets. Efforts were first focused on finding a regularity in the motion of these wanderers, so their movement among the fixed stars could be predicted. For Hipparchus and Ptolemy, the ideal model was a combination of uniform circular motions, the epicycles, which were continually adjusted over the centuries to conform to the observed course of the planets. Astronomy had become predictive, even if its models were in continual need of adjustment.

From 1609 to 1618, Kepler fixed the planets' trajectories: having assimilated the lessons of Copernicus, he placed the Sun at the center of the universe and, based on the observations of Tycho Brahe, showed that the planets described ellipses around the Sun. At the end of a revolution, each planet found itself back where it started and so retraced the same ellipse. Although seductive in its simplicity, this vision of a perfectly stable solar system in which all orbits were periodic would not remain unchallenged for long.

In 1687 Newton announced the law of universal gravitation. By restricting this law to the interactions of planets with the Sun alone, one obtains Kepler's phenomenology. However, Newton's law applies to all interactions: Jupiter is attracted by the Sun, as is Saturn, but Jupiter and Saturn also attract each other. There is no reason to assume that the planets' orbits are fixed invariant ellipses, and Kepler's beautiful regularity is destroyed.

In Newton's view, the perturbations among the planets were strong enough to destroy the stability of the solar system, and divine intervention was required from time to time to restore planets' orbits to their place. Moreover, Newton's law did not yet enjoy its present status, and astronomers wondered whether it was truly enough to account for the observed movements of bodies in the solar system.

The problem of solar system stability was a real one, since, after Kepler, Halley was able to show, by analyzing the Chaldean observations transmitted by Ptolemy, that Saturn was moving away from the Sun while Jupiter was moving closer. By crudely extrapolating these observations, one finds that 6 million years ago Jupiter and Saturn were at the same distance from the Sun. In the 18th century, Laplace took up one of these observations, which he dated 1 March 228 BC: at 4:23 am, mean Paris time, Saturn was observed 'two fingers' under Gamma in Virgo. Starting from contemporary observations, Laplace hoped to calculate backward in time using Newton's equations to arrive to this 2000 yr old observation.

The variations of planetary orbits were such that, in order to predict the planets' positions in the sky, de LALANDE was required to introduce artificial 'secular' terms in his ephemeris tables. Could these terms be accounted for by Newton's law?

Laplace–Lagrange stability of the solar system

The problem remained open until the end of the 18th century, when Lagrange and Laplace correctly formulated the equations of motion. Lagrange started from the fact that the motion of a planet remains close, over a short duration, to a Keplerian ellipse, and so had the notion to use this ellipse as the basis for a coordinate system (figure 1). Lagrange then wrote the differential equations that govern the variations in this elliptic motion under the effect of perturbations from other planets, thus inaugurating the methods of classical CELESTIAL MECHANICS. Laplace and Lagrange, whose work converged on this point, calculated secular variations, in other words long-term variations in the planets' semi-major axes under the effects of perturbations by the other planets. Their calculations showed that, up to first order in the masses of the planets, these variations vanish (Poisson and Poincaré later showed that this result remains true through second order in the masses of the planets, but not through third order).

This result seemed to contradict Ptolemy's observations from Antiquity but, by examining the periodic perturbations between Jupiter and Saturn, Laplace discovered a quasi-resonant term $(2\lambda_{\text{Jupiter}} - 5\lambda_{\text{Saturn}})$ in their longitudes. This term has an amplitude of 46′50″ in Saturn's longitude, and a period of about 900 yr. This explains why observations taken in 228 BC and then in 1590 and 1650 could give the impression of a secular term.

Laplace then calculated many other periodic terms, and established a theory of motion for Jupiter and Saturn in very good agreement with 18th century observations. Above all, using the same theory, he was able to account for Ptolemy's observations to within 1 minute of arc, without additional terms in his calculations. He thus showed that Newton's law was in itself sufficient to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

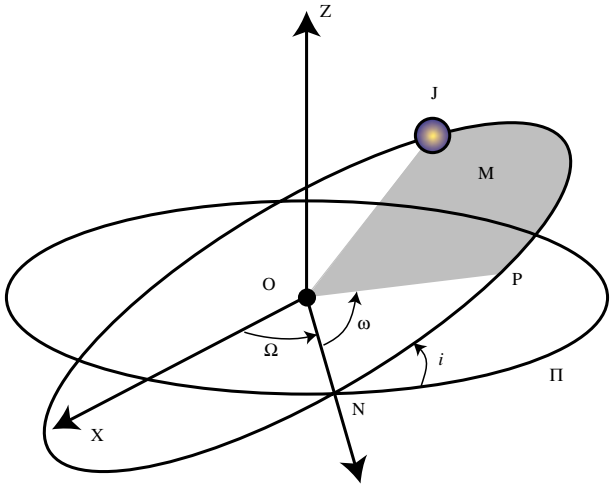


Figure 1. Elliptical elements. At any given time, a planet (J) can be considered to move on an elliptical orbit, with semimajor axis a and eccentricity e, with the Sun at one focus (O). The orientation of this ellipse with respect to a fixed plane Π, and a direction of reference OX, is given by three angles: the inclination i, the longitude of the node Ω, the longitude of perihelion $\varpi = \Omega + \omega$, where ω is the argument of perihelion (P). The position of the planet on this ellipse is given by the mean longitude $\lambda = M + \varpi$, where M (mean anomaly) is an angle which is proportional to the area OPJ (Kepler's third law).

explain the movement of the planets throughout known history, and this exploit no doubt partly accounted for Laplace's determinism.

This result, where Laplace and Lagrange demonstrated that the planets' semimajor axes undergo only small oscillations and do not have secular terms, was the first major result of stability for the solar system. At the same time, Laplace firmly established Newton's law as the universal explanation for the motion of the celestial bodies.

The problem of the eccentricities

The stability of the semimajor axes of the planets is not sufficient to ensure the stability of the solar system. Indeed, if the eccentricity of the Earth becomes larger than 0.1 and the eccentricity of Mars becomes larger than 0.3, then collisions between these two planets can occur. The problem of the stability of the eccentricities and inclination of the planets was addressed by Laplace and Lagrange in an additional set of papers.

Taking into account only terms of first order in the perturbation series, they showed that the system of equations describing the mean motions of eccentricities and inclinations may be reduced to a system of linear differential equations with constant coefficients depending on the planetary masses and semimajor axes:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} z_1 \\ \vdots \\ z_k \\ \zeta_1 \\ \vdots \\ \zeta_k \end{bmatrix} = \sqrt{-1}\begin{bmatrix} A_k & 0_k \\ 0_k & B_k \end{bmatrix}\begin{bmatrix} z_1 \\ \vdots \\ z_k \\ \zeta_1 \\ \vdots \\ \zeta_k \end{bmatrix}$$

where for each planet j, $z_j = e_j \exp\sqrt{-1}\varpi_j$, $\zeta_j = \sin(i_j/2)\exp\sqrt{-1}\Omega_j$, A_k and B_k are (k,k) matrices with real coefficients depending on the values of the planetary masses and semimajor axes. 0_k is the (k,k) zero matrix and $(a_j, e_j, i_j, \lambda_j, \varpi_j, \Omega_j)$ are classical elliptical elements (figure 1).

Using the conservation of ANGULAR MOMENTUM, Laplace demonstrated that, provided that all the planets rotate around the Sun in the same direction, polynomial or exponential solutions cannot exist for this system. He concluded that the eigenvalues g_i, s_i of A_k and B_k are real and that the solutions for this differential system are quasiperiodic expressions of the form

$$z_j = \sum_{i=1}^{k} \alpha_{ij}\, \mathrm{e}^{\mathrm{i}g_i t} \qquad \zeta_j = \sum_{i=1}^{k} \beta_{ij}\, \mathrm{e}^{\mathrm{i}s_i t}$$

where α_{ij} and β_{ij} are complex quantities. The frequencies g_i, s_i are called the secular frequencies of the solar system, and their values, as computed with a more complete model (Laskar 1990), are given in table 1 (see also figure 2).

Table 1. Fundamental frequencies of the precession motion of the solar system (excluding Pluto). These values are taken as the mean values over 20 million years from the recent solution La90. For the inner planets, because of chaotic diffusion, the frequencies can change significantly with time (Laskar 1990).

	ν ($''$ yr^{-1})	Period (yr)
g_1	5.596	231 000
g_2	7.456	174 000
g_3	17.365	74 600
g_4	17.916	72 300
g_5	4.249	305 000
g_6	28.221	45 900
g_7	3.089	419 000
g_8	0.667	1 940 000
s_1	−5.618	230 000
s_2	−7.080	183 000
s_3	−18.851	68 700
s_4	−17.748	73 000
s_5	0.000	
s_6	−26.330	49 200
s_7	−3.005	431 000
s_8	−0.692	1 870 000

The inclinations and eccentricities of the orbits are therefore subject to only small variations about their mean values. However, it must be stressed that Laplace's solutions are very different from Kepler's, because the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

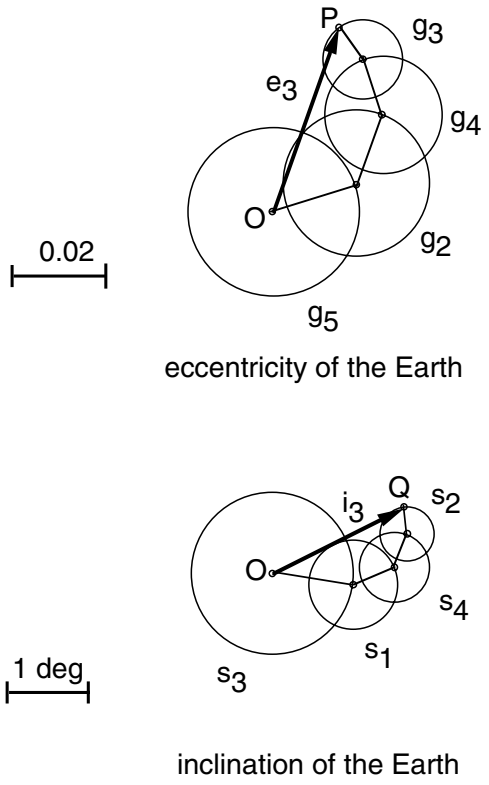


Figure 2. The solutions of Laplace–Lagrange for the motion of the planets are combinations of circular and uniform motions with the precession frequencies g_i and s_i of the solar system (table 1). The eccentricity e_3 of the Earth is given by OP, while the inclination (i_3) of the Earth with respect to the invariant plane of the solar system is OQ (Laskar 1992).

orbits are no longer fixed. They are subject to a double precessional motion with periods ranging from 45 000 to several million years: precession of the perihelion, which is the slow rotation of the orbit in its plane, and precession of the nodes, which is the rotation of the plane of the orbit in space.

Later, Leverrier, famed for his discovery in 1846 of the planet Neptune through calculations based on observations of irregularities in the movement of Uranus, took up Laplace and Lagrange's calculations and considered the effects of higher-order terms in the series (Leverrier 1856). He showed that these terms produced significant corrections and that Laplace's and Lagrange's calculations 'could not be used for an indefinite length of time'. He then challenged future mathematicians to find exact solutions, without approximations. The difficulty posed by 'small divisors' showed that the convergence of the series depended on initial conditions, and the proof of the stability of the solar system remained an open problem (see Laskar 1992 for more details on this point).

Between 1892 and 1899, Poincaré formulated a negative response to Leverrier's question. In so doing he rethought the methods of celestial mechanics along the lines of Jacobi's and Hamilton's work. In his memoir 'On the three body problem and the equations of dynamics', Poincaré showed that it is not possible to integrate the equations of motion of three bodies subject to mutual interaction and not possible to find an analytic solution representing the movement of the planets valid over an infinite time interval, since the series used by astronomers to calculate the movement of the planets were not convergent.

In the 1950s and 1960s, the mathematicians Kolmogorov and Arnold took up Poincaré's work and showed that, for certain values of the initial conditions, it was nonetheless possible to obtain convergent series. If the masses, eccentricities and inclinations of the planets are small enough, then many initial conditions lead to quasiperiodic planetary trajectories, similar to the Laplace–Lagrange solutions. However, the actual masses of the planets are much too large for this result (known as the KAM theorem) to apply directly to the solar system and thereby prove its stability.

Although the constants required for the application of Arnold's theorem correspond to extremely small values of the planetary masses, this result reinforced one more time the idea that the solar system was stable, by any reasonable acceptance of this term, on a time comparable with its age.

The results obtained through numerical integration in the past decade show the contrary.

Recent studies

In very recent years, the problem of solar system stability has advanced considerably, largely as a result of computers which allow extensive analytic calculations and numerical integrations over model time scales approaching the age of the solar system.

One part of these efforts consists of direct numerical integration of the equations of motion (Newton's equations, sometimes with additional relativistic corrections or perturbations due to the Moon). Initial studies were limited to the motion of the outer planets, from Jupiter to Pluto. In fact, the more rapid the orbital movement of a planet, the more difficult it is to numerically integrate its motion. To integrate the orbit of Jupiter with a conventional integrator, a step size of 40 days will suffice, while a step size of 0.5 days is required to integrate the motion of the whole solar system (Cohen *et al* 1973; Kinoshita and Nakai 1984; Carpino *et al* 1987). These studies, reaching 100 million years, essentially confirmed the stability of the system and the validity of the old perturbative approach of Laplace and Lagrange. At about the same time, calculations of the same system were carried out at MIT over even longer periods, corresponding to times of 210 and 875 million years. These calculations were carried out on 'Orrery', a vectorized computer specially designed for the task (Applegate *et al* 1986; Sussman and Wisdom 1988). This integration showed that the motion of Pluto is chaotic, exhibiting exponential divergence with respect to the initial conditions, with a characteristic (Lyapunov) time of 20 Myr. However, since the mass of Pluto is very small

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(1/130 000 000 the mass of the Sun), this does not induce macroscopic instabilities in the rest of the solar system, which appeared relatively stable in these numerical studies.

The other possibility, in order to overcome some of the limitations of numerical integrations, consists in a semi-analytical approach. Using perturbation methods developed by Lagrange, Laplace and Leverrier, Laskar (1989) derived an extended averaged system for the whole solar system except Pluto, including all contributions up to second order with respect to the masses, and through degree 5 in eccentricity and inclination. For the outer planets, some estimated corrections of third order were also included. The system of equations thus obtained comprises some 150 000 terms and does not model the motion of the planets, but rather the averaged motion of their orbits. It thus can be integrated numerically on a computer using a very large step size, on the order of 500 yr. An integration over 200 million years showed that the solar system, and more particularly the system of inner planets (Mercury, Venus, Earth and Mars), is chaotic, with a Lyapunov time of 5 million years (Laskar 1989). An error of 15 m in the Earth's initial position gives rise to an error of about 150 m after 10 Myr, but this same error grows to 150 million km after 100 Myr. It is thus possible to construct EPHEMERIDES over a 10 million year period, but it becomes essentially impossible to predict the motion of the planets with precision beyond 100 million years.

This chaotic behavior essentially originates in the presence of two secular resonances among the planets: $\theta = 2(g_4 - g_3) - (s_4 - s_3)$, which is related to Mars and the Earth, and $\sigma = (g_1 - g_5) - (s_1 - s_2)$, related to Mercury, Venus and Jupiter (the g_i are the secular frequencies related to the perihelia of the planets, while the s_i are the secular frequencies of the nodes) (Laskar 1990). The two corresponding arguments change several times from libration to circulation over 200 million years, which is also a characteristic of chaotic behavior.

The improvement of computer speed and the development of new methods for numerical integration allowed most of these results to be confirmed by the direct integration of Newton's equations (figure 3) (Quinn *et al* 1991; Laskar *et al* 1992; Sussman and Wisdom 1992).

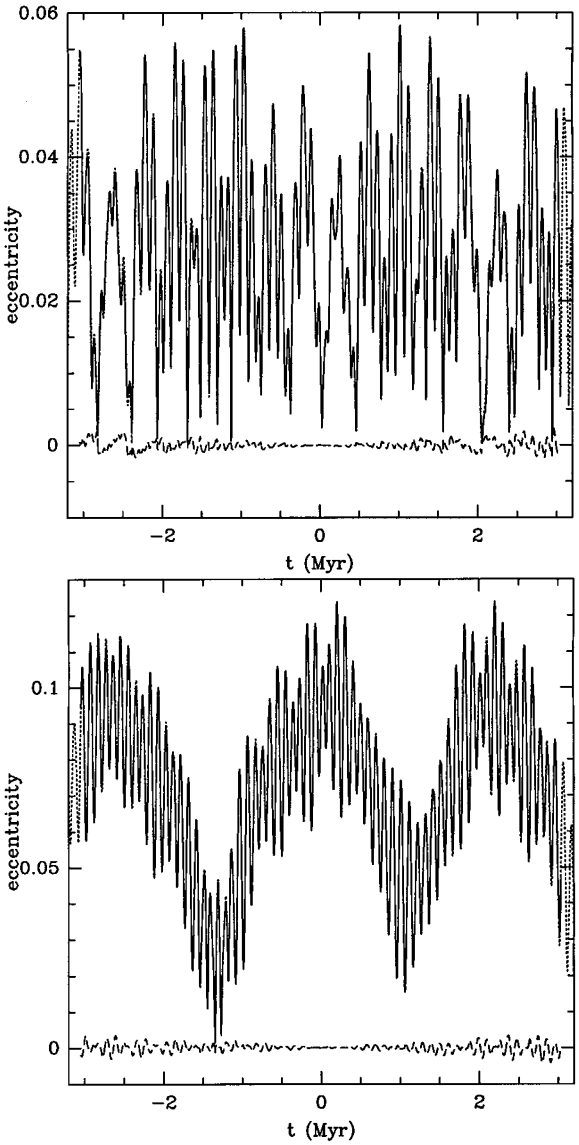


Figure 3. The eccentricity of the Earth (top) and Mars (bottom) during a 6 Myr timespan centered at the present. The solid curve is the numerical solution from Quinn *et al* (1991), and the dotted curve the integration of the secular equations (Laskar 1990). For clarity, the difference between the two solutions is also plotted (from Laskar *et al* 1992).

Evolution of planetary orbits

Over less than 1 million years, a quasiperiodic model such as the one of Laplace–Lagrange (figure 2) gives a fair representation of the evolution of the planetary orbits. This linear model, although not very precise, provides in particular a good account of the variations of the eccentricity and inclination of the Earth which will be at the origin of the variation of the orientation of its axis of rotation, and thus of the insolation on its surface. Indeed, a similar model derived by Leverrier (1856) was used by Milankovitch for the establishment of his astronomical theory of paleoclimates (see CLIMATE).

Over a longer period, of a few millions of years, a quasiperiodic approximation of the solution is still possible, but it should take into account the effect of the resonances between the secular motion of inner planets. It will thus not be possible to obtain it with the classical perturbative method of Leverrier and its successors. On the other hand, such approximation can be obtained by some refined Fourier techniques, after the numerical integration of the averaged equations (Laskar 1988, 1990).

The question of the maximum possible variations of the planetary orbits over the age of the solar system becomes now even more difficult to answer, as, because of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

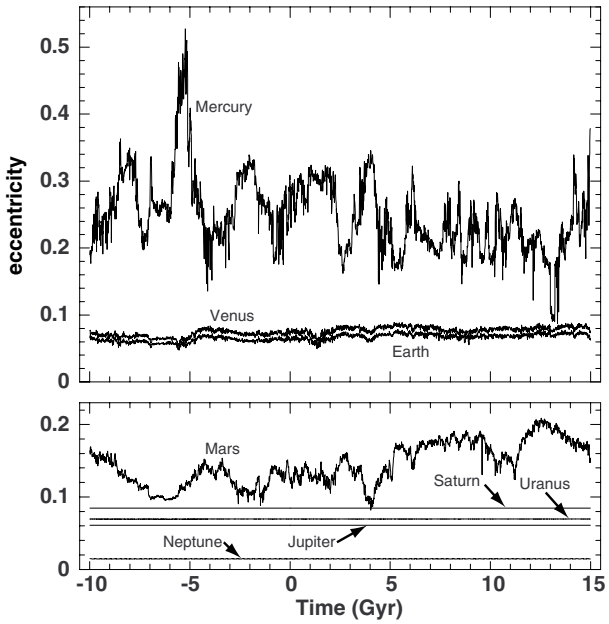


Figure 4. Numerical integration of the averaged equations of motion of the solar system 10 Gyr backward and 15 Gyr forward. For each planet, the maximum value obtained over intervals of 10 Myr for the eccentricity is plotted versus time. For clarity, Mercury, Venus and the Earth are plotted separately from Mars, Jupiter, Saturn, Uranus and Neptune. The large planets' behavior is so regular that all the curves of maximum eccentricity appear as straight lines. On the contrary the corresponding curves of the inner planets show very large and irregular variations, which attest to their diffusion in the chaotic zone (Laskar 1994).

the exponential divergence of the orbits, we know that it will not be possible to obtain precisely the orbital evolution of the solar system after more than 100 Myr.

The computation of the evolution of the solar system over 5 billion years may thus appear illusory, but one does not seek here to predict the precise evolution of the system, but to look only for its possible behavior. With this intention, the integration of the orbits was even pushed over durations going well beyond the age of the solar system (Laskar 1994, 1995). The results (figure 4) provide a very clear vision of the stability of the planetary orbits. In this figure is represented the computed evolution of the eccentricity of the orbits of planets of the solar system over a duration of 25 billion years (from −10 to +15 billion years).

In fact, for better clarity, the plotted curve represents only the variation of the maximum eccentricity reached by the planetary orbits over intervals of 10 million years. Indeed, the oscillations of the eccentricity resulting from the linear coupling of the solutions (figure 3) are removed by this procedure. In doing so, the only variations which appear in figure 4 are thus the variations due to the chaotic diffusion of the orbits.

For all external planets, the maximum eccentricity is almost constant. That reflects the fact that these trajectories are very close to regular and quasiperiodic trajectories; possible instabilities are insensitive with the scale of the the drawing.

For Venus and the Earth, one observes moderated variations, but still significant. The maximum eccentricity of the Earth acquired through chaotic diffusion reaches about 0.08, whereas its current variations are approximately 0.06. This situation is about the same for Venus.

The two curves of the maximum eccentricity of the Earth and Venus are very similar, because of the linear coupling between these two planets. The evolutions of the orbits of Mars and Mercury are very spectacular. The diffusion of the eccentricity of Mars can bring this to 0.2 in a few billion years, whereas the variations of Mercury's orbit can lead its eccentricity to values exceeding 0.5.

In fact, the system is still constrained by angular momentum conservation, which constrains strongly the most massive planets, and it is remarkable to note that, in the system of interior planets, the less massive a planet is, the larger is the possible diffusion of its orbit. The behaviors of the inclinations are very similar to those of the eccentricities.

Because of the chaotic character of the orbits, a very small modification of the initial conditions will lead to a solution different from the preceding one after a few hundreds of million years, but the general aspect of the solutions will undoubtedly remain the same. To evaluate which are the possible maximum variations for the orbits of planets over 5 billion years, one can seek, by very small modifications of the initial conditions, the trajectory which leads to the strongest variations of the orbits.

More systematically, Laskar (1994) calculated five trajectories of very close initial conditions over 500 Myr. The trajectory leading to the strongest Mercury eccentricity is then retained, and, in the vicinity of this maximum, five new trajectories are computed again for a new duration of 500 Myr. For Mercury, this method then makes it possible to obtain in ten such stages an orbit which comes to cut Venus's orbit in less than 3.5 billion years. Let us note that, brought back to the initial position, because of the exponential divergence of the orbits, the displacements of initial conditions correspond to a displacement of the position of the Earth smaller than Planck's length ($\approx 10^{-33}$ cm).

It should, however, be noted that, to arrive at this possible collision between Mercury and Venus, the model was used beyond its rigorous field of validity, which does not includes the vicinity of collisions. In addition, the solution was carefully chosen, so in any case it is surely not a very probable one, and the majority of the solutions of close initial conditions will not lead to this possible collision.

The same method was applied to all other planets, but the chaotic diffusion of their orbits did not allow for a collision in less than 5 billion years. The planet which, with Mercury, has the most unstable orbit is the planet Mars, whose eccentricity can, by this same method, reach approximately 0.25 in less than 5 billion years, while the Earth's eccentricity barely reaches 0.1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

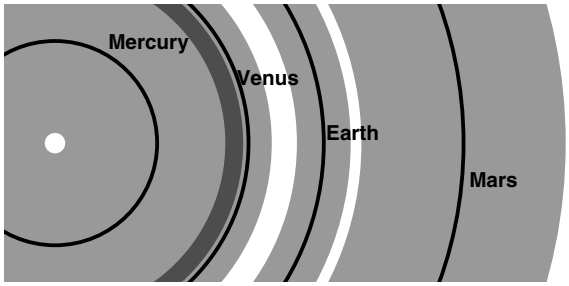


Figure 5. Estimates of the zones possibly occupied by the inner planets of the solar system over 5 Gyr. The circular orbits correspond to the bold curves, and the zones visited by each planet resulting from the possible increase of eccentricity are the shaded zones. In the case of Mercury and Venus, these shaded zones overlap. Mars can go as far as 1.9 AU, which roughly corresponds to the inner limit of the asteroid belt (Laskar 1995).

Marginal stability of the solar system

If one summarizes these results on a plane graph representing the zone swept by the planetary orbits for the maximum values of their eccentricity (figure 5), one notes that the solar system is 'full': there is no place for an additional body. One needs at least 3.5 billion years to allow a collision between Mercury and Venus, but an additional body placed in this system will probably collide more rapidly with one of the already existing planets.

This observation leads then to the concept of marginal stability for the solar system: the solar system is unstable, but catastrophic phenomena leading to the destruction of the system in its current form can take place only in a time comparable with its age, that is to say approximately 5 billion years. The observation of this present state then makes it possible to suppose that the chaotic evolution of the orbits was a determinant process for the structuring of the system after the early stages of its formation (see SOLAR SYSTEM: FORMATION). At that time, some bodies other than the current planets could have remained, but in this case the system would have been much more unstable, and a collision or an ejection could have taken place (an example could be the impactor of the Earth which was at the origin of the formation of the Moon). After this event, the remaining system becomes much more stable. We thus obtain a self-organization of the system towards increasingly stable states. This vision is in agreement with the models of formation of planets by accretion of PLANETESIMALS (Safronov 1969), because it shows how the residual bodies could disappear, in particular in the internal solar system. It is remarkable that the zone swept by the orbit of Mars to its maximum eccentricity reaches the limits of the asteroid belt.

Concerning the system of the outer planets, things are appreciably different, because the direct gravitational short-period perturbations are more significant. The recent numerical simulations show that particles placed among the outer planets do not remain beyond a few hundreds of million years, apart for some particular zones of stability or beyond Neptune, in the Kuiper belt, where objects explicitly were found.

Finally, these observations also make it possible to have an idea of the general aspect of a planetary system around a star (see EXOPLANETS). Indeed, if the process of planetary formation from planetesimals is correct, it becomes possible that, in their final states, planetary systems will always be in a state of marginal stability like our own solar system. At the end of the phase of formation of the system a great number of bodies can remain, but in this case the system is strongly unstable, and is led to a collision or an ejection. After this event, the system becomes more stable. In particular, a system with only one planet, or even with two planets like Jupiter and Saturn, will not be able to exist, because this system would be too stable so that gravitational instabilities could not evacuate all the other bodies initially present. More precisely, if such a system exists, a multitude of small bodies will have to also remain, which will not have been evacuated by these gravitational instabilities.

Bibliography

Laskar J 1989 A numerical experiment on the chaotic behavior of the solar system *Nature* **338** 237–8

Laskar J 1990 The chaotic motion of the solar system. A numerical estimate of the size of the chaotic zones *Icarus* **88** 266–91

Laskar J 1992 La stabilité du système solaire *Chaos et Déterminisme* ed A Dahan *et al* (Paris: Seuil) (partially translated and reprinted as Laskar J 1995 The stability of the solar system from Laplace to the present *General History of Astronomy* ed R Taton and C Wilson, vol 2B, pp 240–8)

Laskar J 1995 Large scale chaos and marginal stability of the solar system *XIth ICMP Meeting (Paris, July 1994)* (International Press) pp 75–120 (also in *Celest. Mech.* **64** 115–62)

Quinn T R, Tremaine S and Duncan M 1991 A three million year integration of the Earth's orbit *Astron. J.* **101** 2287–305

Sussman G J and Wisdom J 1988 Numerical evidence that the motion of Pluto is chaotic *Science* **241** 433–7

Sussman G J and Wisdom J 1992 Chaotic evolution of the solar system *Science* **257** 56–62

Jacques Laskar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Telescopes and Instruments: Ground

Solar observations require special telescopes and instruments. They share the basic optical principles with their nighttime companions, but there are also some substantial and characteristic differences between solar and stellar telescopes.

The most obvious distinction is caused by the target of observation: the Sun. It irradiates the Earth's surface with a power of about 1 kW m^{-2}, and a solar telescope has to withstand this heat input while maintaining its optical quality.

The Sun, as an extended object, allows the study of details on its surface. In order to achieve the highest spatial resolution possible, the solar image scale has to be adapted to the pixel size of the detector in the focal plane. This requires a large focal length and a large f-number for a given entrance aperture diameter, and that again distinguishes a high-resolution telescope from any other astronomical instrument. The properties of the telescope site are important for the achievable image quality. In addition to the high-resolution telescopes there are special-purpose solar telescopes which investigate global properties, such as the solar oscillations or the large-scale effects of the Sun's magnetism.

Resolution aspects

The most important characteristics of a telescope are its resolution and light collection power. Since the Sun appears at least 10^{10} times brighter than any star in the sky, one may think that the second parameter is rather irrelevant for a solar telescope, but we will see later that this is indeed not the case.

The image of a point source is completely described by the point spread function (PSF) for a perfect telescope. For a circular clear (unobstructed) aperture we obtain the well-known Airy pattern $\mathrm{PSF} = (1/\pi)[J_1(Br)/Br]^2$, where $B = \pi D/\lambda f$ and J_1 is the Bessel function of order 1; D and f are the aperture and the focal length of the telescope and λ is the wavelength used for the observation. An important quantity derived from the PSF is the resolution angle $a_1 = r_1/f = 1.22\lambda/D$, or $a_s = 2.52 \times 10^5 \lambda/D$, where r_1 is the radius of the first dark Airy ring and a_s expresses the resolution in arcseconds. The imaging of extended objects, such as the Sun, with its low-contrast structures depends strongly on factors in addition to the PSF. These factors include the imperfection of the optical elements of the telescope, atmospheric distortion and image motion, as well as the degradation effects of the detector.

The modulation transfer function (MTF) describes the reduction in contrast between the object and the image, as a function of spatial frequency: $\mathrm{MTF}(k) = C_i(k)/C_o(k)$. The MTF is the modulus of the optical transfer function (OTF), and the OTF and the PSF are Fourier pairs, so the MTF is easily computed from the PSF. For a circular clear aperture the MTF is

$$\mathrm{MTF_T}(k) = \frac{2}{\pi}\left\{\arccos\left(\frac{k}{k_c}\right) - \frac{k}{k_c}\left[1 - \left(\frac{k}{k_c}\right)^2\right]^{1/2}\right\}$$

where k_c is the cutoff wavenumber of the telescope ($\mathrm{MTF}(k_c) = 0$). The resolution angle corresponds to a wavenumber $k_m = 2\pi {r_1}^{-1}$ and the contrast reduction factor at this wavenumber is 0.09 (figure 1). In a complex optical system each component is described with its own transfer function and the total MTF is just the product of the individual functions:

$$\mathrm{MTF_{tot}} = \mathrm{MTF_D} \times \mathrm{MTF_T} \times \mathrm{MTF_S} \times \mathrm{MTF_C}$$

where the indices denote diffraction, telescope aberrations, seeing and detector.

Telescope aberrations

Deviations from a perfect figure of an optical surface are specified as wavefront error, σ, expressed in units of the incident wavelength. The PSF provides a simple criterion for the overall quality of a telescope: if the central intensity is at least 80% of the value for a perfect system, the instrument is called diffraction limited. The ratio $S_R = {I_0}^{\mathrm{real}}/{I_0}^{\mathrm{ideal}}$ is called the Strehl ratio. For small aberrations the relation $S_R = 1 - k^2\sigma^2$ holds and we find the well-known criterion for the acceptable amount of wavefront errors for a diffraction-limited system, $\sigma = 0.0712\lambda = \lambda/14$.

The disturbing Earth's atmosphere

The image degradation caused by the Earth's atmosphere is due to thermal inhomogeneities causing local fluctuations in the index of refraction. The most commonly used parameter to describe this effect is the Fried parameter $r_0 = 0.299D[(k\sigma)^2]^{-5/3}$, where σ^2 is the averaged mean square wavefront error (in rad). The value of r_0 depends on the telescope site, the time of the day and on wavelength. Even under best conditions r_0 does not exceed 20 cm at visible wavelengths. Good sites for solar telescopes are high mountain peaks, sites on islands or places surrounded by a lake. Water surfaces, both nearby and more distant ones, produce a smooth temperature profile of the overlying air—an important ingredient of good seeing. Another important factor is wind, which prevents the buildup of turbulence in the immediate surroundings of a telescope caused by heating of the soil. Qualitatively, seeing conditions corresponding to a certain value of r_0 permit (on average) the same spatial resolution as a telescope with a diameter of r_0. The Fried parameter is a time-averaged value of the statistically varying seeing, and a sequence of many short-time exposures (with exposure times of a few milliseconds) therefore will contain some frames with a quality much better than the limit given by r_0. Frame selection techniques are used to extract these high-quality frames from the sequence. The spatial resolution of solar telescopes is usually seeing-limited for spectroscopic measurements with exposure times longer than some 10 ms, and ADAPTIVE OPTICS is needed to overcome this limitation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

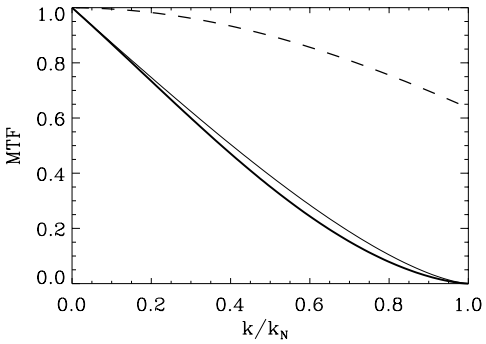


Figure 1. Modulation transfer function of an ideal telescope (thin full curve), a CCD camera (with two pixels per resolution element of the telescope, dashed curve) and their product (thick curve) plotted as a function of normalized spatial wavenumber.

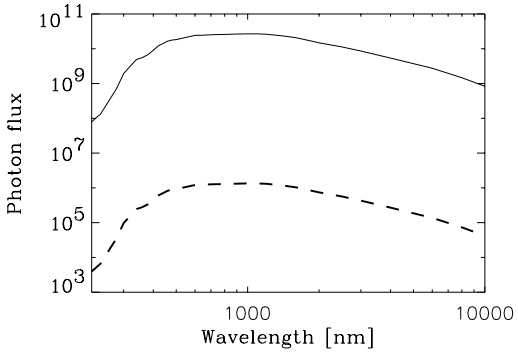


Figure 2. Photons per resolution element of a telescope, for a wavelength interval of 1 nm and perfect telescope transmission. The dashed curve shows the same function for a telescope efficiency of 0.05 and a wavelength band of 1 pm.

Detector transfer function

Solid-state digital detectors nowadays have replaced analog media as recording devices. The commonly used CCD CAMERAS have smaller pixels and a much higher sensitivity (quantum efficiency) compared with photographic film. From sampling theory it is known that correct sampling is required in order to avoid spurious signals in the recorded image. The smallest structure, $x_c = \lambda f D^{-1}$, in a telescope image is given by the diffraction limit of the instrument. Critical sampling requires two pixels per resolution element, $p = 0.5x_c$, where p is the pixel size. The length x_c corresponds to the cutoff wavenumber $k_c = 2\pi/x_c$. The pixel size defines the maximum frequency of the digitized data, $k_N = 2\pi/2p$, so $k_c = k_N$ for critical sampling. In addition to the transfer effects of pixel size, the transfer function of a CCD camera has a second component, which describes diffusion processes in the photosensitive layer: $\mathrm{MTF_C} = \mathrm{MTF_P} \times \mathrm{MTF_D}$. The pixel contribution is $\mathrm{MTF_P} = \sin(kp/2)/(kp/2)$, where p is the pixel pitch, which is assumed to equal the pixel size (modern CCDs have indeed 100% fill factor, i.e. no gaps between pixels). At the Nyquist frequency, k_N, the pixel MTF has dropped to 0.637, so critical sampling leads to a significant contrast reduction at high spatial frequencies (figure 1). The (usually small) diffusion contribution depends on wavelength and on pixel size. For a given image scale, smaller pixels increase the contrast transfer, i.e. the resolution, but this improvement has to be traded against the concomitant decrease in dynamical range, which is inversely proportional to the pixel size.

Correlation tracking

Atmospheric seeing causes image degradation due to blurring and also image motion within the field of view which leads to distortion. The local image motion can be compensated by means of a high-speed correlation tracker. A fast camera records images of the solar granulation at a rate of at least 1000 frames per second. The image displacement is computed with a suitable algorithm, e.g. cross-correlation, and a fast tip-tilt mirror finally corrects the image motion in a servo-loop. Correlation trackers need a closed-loop bandwidth of more than 50 Hz for proper functioning.

How bright is the Sun?

The spectral flux of the Sun measured outside the Earth's atmosphere peaks in the green wavelength region. At a wavelength of 550 nm it is $f_\lambda = 1.96$ J m^{-2} s^{-1}. This corresponds to a photon flux of $f_p = 5.4 \times 10^{18}$ photons m^{-2} nm^{-1} s^{-1}. Compared with the canonical value of 10^8 photons m^{-2} nm^{-1} s^{-1} for a zero-magnitude star, the Sun is indeed a bright source. Very seldom, however, is the Sun observed as a star, and high-resolution measurements are better characterized by the photon flux per resolution element of a telescope:

$$f_{\mathrm{res}} = \frac{\pi}{4} \frac{f_p \lambda^2}{R_\odot^2}$$

where f_p is the photon flux, λ the wavelength and $R_\odot$ the angular radius of the Sun (4.65 mrad). This quantity is a property of the Sun and is independent of the telescope size. Any increase in light-collecting area is exactly compensated by a corresponding decrease of the resolution element. The wavelength dependence of f_{res} is shown in figure 2, together with a curve computed for a realistic telescope efficiency of 0.05 and a wavelength bandwidth of 1 pm, a typical value for narrow-band spectroscopy.

Small-scale objects on the Sun are highly dynamic or transient; typical time-scales range between several minutes to just a few seconds. Obvious examples are the formation or disappearance of a thin magnetic flux tube or a flare event. The temporal and the angular resolution of an observation are coupled: the smaller the feature, the faster is the cadence needed for adequate sampling of the evolution. Figure 2 demonstrates the lack of photons

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

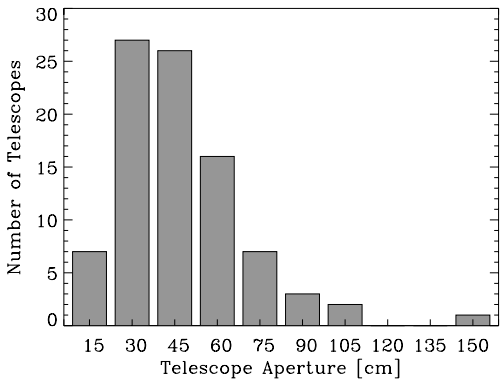


Figure 3. Distribution of telescope apertures, excluding helioseismic instruments.

especially for narrow-band measurements in the blue part of the visible spectrum.

Telescope types

Optical solar telescopes can be divided in three groups, according to their main purpose: high-resolution, synoptic and helioseismic instruments. About 100 such telescopes are installed world-wide. Figure 3 shows the distribution of apertures of the first two groups. Most of these telescopes have diameters of less than 50 cm, and there is only a single one (the 1.5 m McMath–Pierce Telescope at KPNO) with an aperture of more than 1 m. It may be worthwhile to mention that most high-resolution telescopes are older than 15 yr. This is related to the dominating effect of day-time atmospheric SEEING on the achievable spatial resolution. Even at the best sites, narrow-band spectroscopic observations are almost always seeing limited. A Large Earth-based Solar Telescope (LEST) had been thoroughly studied and discussed for many years; its realization has finally been indefinitely postponed owing to lack of sufficient funding.

The breakthrough in solar adaptive optics in 1999 will revolutionize high-resolution solar observations: diffraction-limited observations will become possible in the near future even for large-aperture telescopes and for a significant fraction of the time (whenever the seeing quality exceeds a certain threshold value), and this will eventually lead to the development of a new generation of telescopes with apertures of 2 m and beyond.

High-resolution telescopes

This class of telescopes is designed for diffraction-limited observations and therefore their image scale is adapted to the resolution (pixel size) of the focal plane detectors. In order to image one resolution element onto two pixels, a focal length of $f = 2\Delta x A_u a_s^{-1}$ is needed, where A_u is the astronomical unit, a_s the size of a resolution element and Δx the pixel size. Using the above relation for a_s we obtain

$$\frac{f}{D} = \frac{2\Delta x}{\lambda}$$

for the focal ratio ($\approx$60 for a pixel size of 20 μm). The field of view is only a small fraction of the solar disk, with a diameter of typically just 2 or 3 arcmin.

Tower telescopes have been the classical design of a solar instrument, and the main advantages are (i) simple optics (one mirror with optical power suffices), (ii) large focal length of primary mirror (no hot focal plane, large field of view), (iii) slender, tall building (telescope entrance at large height, above local turbulence layer) and (iv) non-moving telescope and focal plane instrumentation, only few moving parts (feed optics). In order to avoid image degradation caused by internal turbulence, many of these telescopes have been built as evacuated systems, with high-quality entrance and exit windows. Either coelostat systems or alt-azimuthal two-mirror configurations are used as feed optics. The main advantage of the former is the non-rotating solar image, while a turret is much more compact and can be included in the evacuated light path.

The 70 cm Vacuum Tower Telescope on Tenerife is shown as an example in figure 4. Most of the vacuum tube is indeed empty and only needed to accommodate the large focal length of 45 m. The size of the entrance window is one of the limiting factors for the aperture of such vacuum telescopes: at diameters beyond 1 m either the window becomes too thick or bending stress near the circumference, caused by the atmospheric pressure force (8 tons-force for a 1 m window), leads to unwanted polarization effects. Therefore the LEST had been planned as a helium-filled system, with a very thin window to seal the helium from the ambient air. Helium has a smaller index of refraction, (3.6×10^{-5}, compared with 2.93×10^{-4}) than air, higher thermal conductivity and larger kinematic viscosity. These properties result in very small temperature variations and consequently little wavefront disturbance in a helium-filled telescope. However, helium technology is costly and the fabrication of a thin, optical-quality entrance window with a diameter larger than 2 m is a challenging task.

Compact designs with two or more image-forming mirrors are also used for solar telescopes. These telescopes can be installed in much smaller buildings and have much less steelwork. Different optical designs have been used, including Newton-, Gregory- and Ritchey–Chrétien-type configurations. The focal length of the primary mirror is rather short, producing a 'hot' primary image, and therefore all optical elements between the primary and the first image need to be cooled. A modern example is the French THEMIS telescope on Tenerife, a 90 cm Ritchey–Crétien-type telescope, evacuated, and with very low instrumental polarization.

Open-air versus closed telescopes

The evacuation of a telescope completely eliminates internal telescope seeing, and filling such a tube with helium would produce the same effect. Future large telescopes of the 3–4 m class will be neither evacuated nor helium filled, owing to insurmountable problems with the necessary entrance window of any closed system. In

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

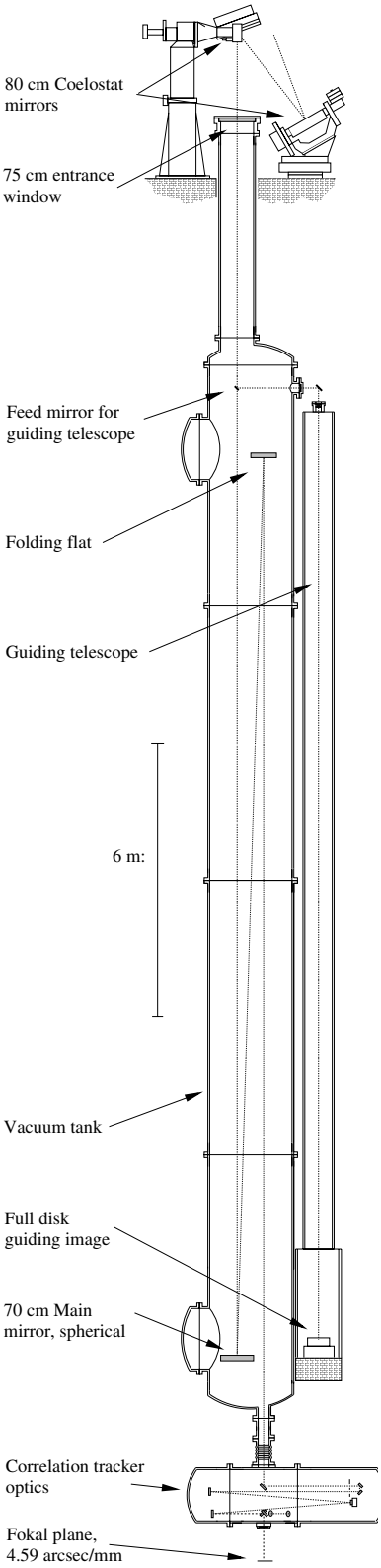


Figure 4. Sectional cut through the 70 cm German Vacuum Tower Telescope on Tenerife. The image shows the main optical components at the correct geometrical scale (T Kentischer, KIS).

addition, closed telescope are limited in spectral range to wavelengths shorter than 2.7 μm, where the window material becomes opaque.

The main source of internal telescope seeing is not the direct heating of the air inside the telescope tube; it is rather caused by the heating of the front side of the primary mirror which in turn heats the air in front of it.

Using a value of 1000 W m^{-2} for the solar irradiance at a high-mountain site, the absorbed heat amounts to about 100 W m^{-2}. From the known properties of Zerodur, the commonly used material for mirror blanks, and of aluminum one can estimate a temperature increase of about 20 K above ambient temperature. The elimination of mirror seeing in an open telescope therefore will require active cooling of the main mirror. The use of mirror materials with high thermal conductivity would make mirror cooling much easier.

Synoptic telescopes

During the first half of this century many solar observatories have been dedicated to routine or patrol observations of the Sun, and a wealth of phenomena have been described. Recently, the variability of the Sun and the possible dependence on the magnetic activity cycle of its luminosity and other quantities have initiated a new generation of synoptic telescopes, both space-borne and ground-based ones.

Synoptic telescopes have a large field of view and most of them observe the full Sun. Their main purpose is the investigation of large-scale properties and the observation of long-term variations of the Sun. For full-disk observations apertures of 10–20 cm are normally sufficient, and the focal lengths f are in the range 1–2 m, producing images of the Sun with diameters of $d = 0.01f$. With an image size of 20 mm, a large-format CCD camera with 2000 pixels $\times$ 2000 pixels and a pixel size of 10 μm, a spatial resolution of 2 arcsec can be achieved (assuming critical sampling). Synoptic observations monitor the solar atmosphere at different layers, from the photosphere through the chromosphere to the transition layer (the corona is not accessible from ground, but synoptic observations are carried out from space, e.g. by SOHO or TRACE; see also SOLAR TELESCOPES AND INSTRUMENTS: SPACE). Medium-bandwidth filtergrams (FWHM $\approx$ 50 pm) are produced using Lyot filters or interferometers, e.g. Fabry–Pérot etalons.

NASA, NOAA and other data centers provide (incomplete) databases of a variety of daily synoptic images from different observatories, including

- photospheric images in 'white light' (sunspots),
- magnetograms (photospheric magnetic field),
- Hα images (chromosphere),
- Ca II K images (chromospheric network),
- He 1083 images (chromosphere, transition region),
- coronographic images (corona).

Daily images are sufficient to study the long-term behavior of the Sun, on the time-scale of months or years, including

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

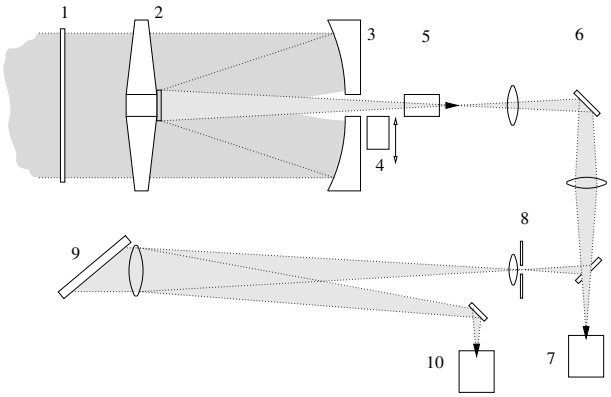


Figure 5. Schematic layout of the SOLIS vector-spectromagnetograph: 1, window; 2, 3, 50 cm Ritchey–Chrétien; 4, calibration optics; 5, polarization modulation and calibration; 6, agile mirror; 7, guider; 8, spectrograph slit; 9, grating; 10, polarizing beam splitter and detector (C Keller, NSO).

variations due to the 11 yr activity cycle. In addition, there are observing programs that study large-scale flows in the photosphere or the emergence of magnetic flux. These measurements require cadences in the order of minutes and high accuracy. Such a program is envisaged e.g. within the SOLIS (Synoptic Optical Long-term Investigations of the Sun) project. SOLIS consists of three telescopes with apertures between 8 mm and 50 cm and it will also observe integrated sunlight ('Sun as a star') and investigate the possible cycle dependence of spectral lines. Figure 5 shows the schematic layout of one of the instruments, the 50 cm vector-spectromagnetograph.

Atmospheric seeing, image quality and photon shortage are normally not an issue for synoptic observations, because diffraction-limited imaging is not required. Photometric accuracy is critical for the measurement of the vector magnetic field, and SOLIS will therefore use an 'oversized' 50 cm telescope for that purpose. Important criteria for the quality of such long-term observations are completeness, time coverage and consistency of the data over the full time interval. Completeness and time coverage could be achieved by organizing networks of similar telescopes across the globe and by automated operation of the instruments.

Helioseismology instruments

Helioseismic experiments are probing the dynamics and structure of the solar interior with very high precision and they allow us to test the standard model of stellar evolution by measuring the global solar oscillations. These consist of a superposition of millions of modes with velocity amplitudes of a few cm s^{-1} per mode, and frequencies between 2 and 4 mHz. Frequency resolution Δf is equivalent to the total duration T of a measurement ($\Delta f = T^{-1}$), and, therefore, long-term observations are necessary. Different observing strategies are possible: (1) repeated observations at a single site, (2) continuous observations from the south pole or (3) network observations from stations around the Earth. The first method is certainly the simplest one and it has been used for years, e.g. at the Observatorio del Teide, where the first station of the BISON (Birmingham Izaña Solar Oscillations Network) NETWORK has been operated since 1975. The obvious disadvantage is a strong contribution to the power spectrum at a frequency of 1 per day (=11.57 μHz) and overtones, which corrupt the spectrum. The south pole measurements have no such effect, but for practical reasons measurement cycles are limited to 3 months and bad weather interruptions may occur at any time. The network approach is the most promising one, since it offers basically unlimited observing time. It is especially important to have virtually identical instruments. At present, three major networks are operated around the Earth. With six stations, more or less equally distributed in longitude, there is sufficient overlap in operating time to compensate for technical failures or bad weather conditions. With improving reliability and increasing degree of automation, duty cycles of more than 90% have been reached.

Different approaches are used to achieve the necessary sensitivity for the velocity measurements: some network telescopes observe integrated sunlight and use atomic resonance spectrometers for the wavelength measurements, reaching an accuracy of a few mm s^{-1}. These experiments measure only the lowermost modes with $l \leq 4$. The GLOBAL OSCILLATION NETWORK GROUP (GONG) instruments use a Michelson interferometer and a Fourier tachometer technique to measure the position of the line of neutral nickel at 676.8 nm. GONG has a spatial resolution of about 8 arcsec (an upgrade to 5 arcsec is planned) and can resolve modes up to $l = 250$. In a 30 day data set, frequencies can be measured with an accuracy of about 50 nHz (about one-seventh of the nominal frequency resolution) and a 1 yr observation will provide an accuracy of 15 nHz. As a result, the frequencies of the solar oscillations are by far the most accurately known parameter of the Sun.

Focal plane instruments

Focal plane equipment can be divided in two major subgroups: (1) broad-band imaging and (2) spectroscopic instruments. In the simplest case, an imaging instrument is just a camera and a filter to select the wavelength band of interest. Imaging instruments play an important role for synoptic observations and morphological studies of the solar atmosphere. With high-speed cameras and suitable *post facto* image reconstruction techniques diffraction-limited images can be obtained. Sequences of images are used to study horizontal flows using a local correlation tracking algorithm. For the measurement of physical parameters of the Sun (e.g. temperature, element abundance, magnetic field strength) spectroscopic instruments are needed. Strictly speaking, an instrument can only collect photons from the Sun, i.e. measure intensity,

$$I = I(x, y, \lambda, t, \boldsymbol{p})$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and all physical information is derived therefrom (x, y are coordinates on the Sun, λ, t and $\boldsymbol{p}$ denote wavelength, time and polarization state, respectively). Focal plane instruments are distinguished by the corresponding measurement techniques and the range and resolution limit of one or more of the aforementioned variables.

Grating spectrographs measure simultaneously one spatial dimension, defined by length of the entrance slit, and wavelength. The basic characteristics are derived from the grating equation

$$\sin\alpha + \sin\beta = \frac{m\lambda}{g}$$

where α and β are the angles between the grating normal and the incoming and outgoing beam, m is the diffraction order and g the grating constant. For the Littrow configuration the two angles are nearly identical, so the grating equation simplifies to $m\lambda = 2g\sin\beta$. According to the Rayleigh criterion two neighboring spectral lines are 'resolved' if the diffraction maxima of one line coincides with the diffraction minima of the other line. This eventually leads to the equation for the resolving power

$$R := \frac{\lambda}{\Delta\lambda} = \frac{2W\sin\beta}{\lambda}$$

where W denotes the width of the wavefront at the grating. The resolving power is independent of the grating constant. The angular dispersion is derived from the grating equation: $\delta = 2\tan\beta\lambda^{-1}$, and is also independent of the grating constant. The linear dispersion $D(=f\delta$, where f is the focal length of the imaging system) is of high practical importance, since it defines the image scale in the focal plane of a spectrograph and, together with the pixel size of the detector, the actual spectral resolution of a measurement.

From the above equation we conclude that high spectral resolution requires large gratings used at high (and nearly equal) angles of incidence and diffraction. The resolution is sometimes written as $R = nm$, where n denotes the total number of illuminated grooves and m the diffraction order. This is mathematically correct, but it may lead to the erroneous conclusion that the number of grooves or the diffraction order contributes to the resolution power of a grating spectrograph.

The observation of a certain area on the Sun is done by scanning the solar image across the spectrograph slit. The quality of such a measurement depends mainly on the precision of the scanning. This kind of measurement is rather sensitive to image distortion, which in this case is nearly uncorrectable. With successful adaptive optics, this problem will more or less disappear.

In contrast to prisms, gratings produce overlapping spectral orders, so that additional filters are required to select a specific wavelength. Echelle spectrographs make best use of this multi-line capability and are therefore very versatile for diagnostic spectroscopy. A typical echelle grating has 79 grooves mm^{-1} and is used at (or near) an angle β of 63.4° ($\tan\beta = 2$). These gratings are blazed to maximize the grating efficiency at that angle. Visible wavelengths correspond to diffraction orders between 30 and 50 with a free spectral range FSR $= \lambda/m$ (= 13 nm for $m = 40$). The spectral order(s) of interest are chosen via a pre-disperser in front of the main spectrograph or with medium-band interference filters in the focal plane. This allows to simultaneously record several wavelength bands from different parts of the spectrum. In contrast to stellar echelle spectrographs, where a cross-disperser displaces overlapping orders perpendicular to the dispersion axis, the pre-disperser selects short wavelength bands that are imaged in the focal plane along the dispersion axis. This imposes a serious limit on the choice of spectral regions: 'degenerate' wavelengths, with the same value of $m\lambda$, cannot be observed simultaneously.

Filter spectrometers are basically narrow-band imaging devices with wavelength tuning capability. They provide spatial information in two dimensions for a specific wavelength and a complete spectrum is obtained by taking filtergrams at different wavelength positions. The spectral resolution and coverage are determined by the type of filter used. In the past, tuneable Lyot filters (or a universal birefringent filter, UBF) have been used for that purpose. The main disadvantage of this filter type is its low transmission. The spectral resolution ranges between some 10 pm and 50 pm, and the main application has been the spectroscopy of strong chromospheric lines.

The first Fabry–Pérot interferometer (FPI) had been built in 1899, long before the first Lyot filter. The commonly used name for these devices, *étalon* (=standard of measure) reminds to its original purpose. Etalons are characterized by the following (simplified) equations for the free spectral range, FSR, and the resolving power, R, which both depend on the spacing t between the interferometer plates:

$$\mathrm{FSR} = \frac{\lambda^2}{2t} \qquad R := \frac{\lambda}{\Delta\lambda} = \frac{2Ft}{\lambda}.$$

The finesse F is the ratio between the FSR and the width of the instrumental profile. It is independent of the spacing and is a measure of the quality of an FPI (reflectivity, parallelism and smoothness of the interferometer plates). It is obvious from the above relations that for a reasonably high spectral resolution of, say, 250 000 at a wavelength of 550 nm, the FSR is very small, of the order of 0.1 nm and hence a very narrow filter is needed to suppress adjacent passbands. As a consequence, several spectrometers have been built as a combination of a UBF, acting as prefilter, with an FPI. The main drawback of that solution is again the low transmission and difficulties in coherent tuning of the two components.

Another solution is the combination of two or more etalons with different spacings. Coarsely speaking, the largest spacing determines the resolution, while the narrowest etalon increases the FSR. The efficiency of a triple-etalon system is at least one order of magnitude higher compared with a UBF.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Filter spectrometers and grating spectrographs complement each other in many respects: a filter spectrometer observes a fixed field of view and the time needed to obtain a full data set $I(x, y, \lambda)$ depends on the number of wavelength steps, while a grating spectrograph always records a fixed wavelength range and the field of view is coupled to the measurement time through the number of scan positions. Measurements which require a very fast cadence can easily be made with a filter instrument, at the cost of wavelength information, but without loss of spatial field. Grating instruments are especially valuable, when precise wavelength information, e.g. the shape of line profiles, is needed. Filter and grating spectrographs are used for POLARIMETRY by adding the capability to discriminate between the polarization states of the incoming light. SOLAR MAGNETOGRAPHS usually measure the line-of-sight component of the magnetic field on the Sun. The investigation of the strength and orientation of the magnetic field vector requires complicated spectropolarimeters (see POLARIMETRY: SPECTROPOLARIMETRY) and a rather involved inversion technique to deduce the relevant physical quantities from the polarization measurements.

Bibliography

Durrant C J 1981 *Landolt Börnstein, New Series* **2** 13

Rimmele T J *et al* (ed) 1999 *High Resolution Solar Physics (ASP Conf. Ser. vol 183)*

Schroeder D J 1987 *Astronomical Optics* (New York: Academic Press)

Stix M 2000 *The Sun, An Introduction* 2nd edn (Berlin: Springer)

Wolfgang Schmidt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Telescopes and Instruments: Space

The first solar space observations

Fifty years of development in solar space instrumentation have demonstrated the close relation between technical and scientific progress.

Solar physics from space, and indeed space astronomy, started in 1946 in the United States, when captured V-2 rockets, capable of rising 160 km above the ground, became available for use as free-flying, high-altitude laboratories. It was not surprising under these circumstances that a group from the US Naval Research Laboratory (NRL) in Washington, DC under Richard Tousey became the first to make use of the fortunate opportunity.

An initial attempt was made on 28 June 1946 to observe radiation in the hydrogen Lyman alpha line from the Sun. However, the camera was never retrieved from the crater made when the V-2 impacted on the desert floor of the White Sands Missile Range in New Mexico. The next attempt, on 10 October 1946, was a success, giving us the first ultraviolet spectrum of the Sun, 220–340 nm (nm—nanometer, i.e. 10^{-9} m), shown in figure 1. At altitudes above 55 km, rising through the ozone layer, the solar spectrum appeared all the way down to 220 nm. The first observations of solar x-rays followed in 1949 when Herbert Friedman, also of NRL, flew a payload of Geiger counters on a V-2 rocket (see also ROCKETS IN ASTRONOMY).

Further rocket observations in the 1950s established that the Lyman alpha line of hydrogen was emitted from the solar CHROMOSPHERE. The radiation in this line was measured quantitatively in 1949 on the same flight that detected solar x-rays and also measured the Schumann ultraviolet continuum at 142.5–160 nm. The line profile of Lyman alpha was measured later, and the narrow absorption core at its center, detected in 1959, established the presence of an extended geo-corona. The first solar spectrum below Lyman alpha was recorded in 1960. Emission lines at these short wavelengths were classified as lines from high ionization stages, mainly of iron, and shown to come from the hot coronal plasma at 1.5–2.5 MK (MK—megakelvin, i.e. million kelvins). The strong enhancement of x-ray emission during solar flares was detected in 1956. When a rocket was flown during a solar eclipse in 1958, it was determined that solar x-rays originated in the CORONA, particularly above ACTIVE REGIONS. The first image of the Sun in x-rays was recorded in 1960, using a pinhole camera on board a rocket. Finally, it may be mentioned that the first CORONAGRAPH was flown in space in 1963.

Looking back it is amazing that these early days of rocket experiments and our first established, simple knowledge about the ultraviolet and x-ray emission from the Sun are only 40–50 years behind us. The early discoveries are now so familiar that we hardly reflect on how difficult they were to obtain and on the fact that they were not obvious.

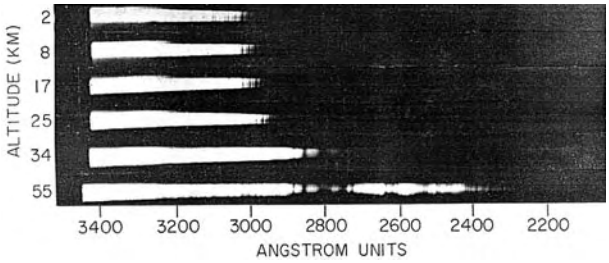


Figure 1. The first ultraviolet spectrum of the Sun. Recorded on a V-2 rocket flight on 10 October 1946.

Early technical developments

Technical developments in many fields conditioned the early progress. Contributions to these developments came from universities and research laboratories in the US, France and the UK.

Magnesium fluoride lenses, transparent to 110 nm, were used in the first rocket observations. Development of mirror coatings with high reflectivity in the far ultraviolet was, however, crucial for progress beyond the first simple experiments. Oxidized aluminum, used on mirrors in the visual wavelength range, does not reflect radiation at far-ultraviolet wavelengths. However, fresh unoxidized aluminum is an efficient reflector down to 100 nm. The solution was to coat the fresh aluminum with a layer of magnesium fluoride. This stopped oxidation and preserved the high reflectivity. Below 110 nm other reflecting materials were found, notably gold, used in instruments on ORBITING SOLAR OBSERVATORIES (OSOs) (see below) and Skylab, but also osmium and silicon carbide. Today, multilayer mirror coatings can be produced with reflectivities up to 30% in narrow ultraviolet and x-ray wavelength bands.

Another important concern is to avoid contamination by stray light. The Sun radiates the overwhelming part of its energy in the visual spectral region. Even a small fraction of this light scattered off the instrument surfaces will completely swamp the ultraviolet signal that we want to study. This problem was solved when an aluminum filter, consisting of a freely suspended aluminum foil of 100 nm thickness, was developed in France and the US. The filter reflected visual light but was transparent to ultraviolet radiation.

Detector development was no less essential. In the 1950s and 1960s Kodak increased the sensitivity of their ultraviolet-sensitive Schumann emulsions, originally developed around 1900. In these emulsions the silver halide grains are sticking out of the gelatin layer to avoid absorption of ultraviolet radiation in the gelatin. This makes it sensitive to mechanical pressure and a roll film camera that could hold large amounts of this film was not made until the 1980s for the High Resolution Telescope and Spectrograph (HRTS), on Spacelab 2. Indeed, since photographic film quickly records vast amounts of image data, it has continued to be used almost until this day.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Photoelectric detectors were used in the near ultraviolet in instruments flown on rockets already in 1952. Channel electronic multipliers came into use with the first satellite instruments. Array detectors in two dimensions proved difficult to make for ultraviolet wavelengths. The first spectral and image scans were therefore made with pinhole devices. This meant slow scans and low time cadence for images and spectra. A modern solution is to convert the ultraviolet radiation to visual light before it is registered with a conventional CCD.

The years of early rocket and satellite experiments also saw the development of increasingly accurate pointing systems. The first rockets flew without pointing and for some time instrument resolution did not put any stringent demands on pointing accuracy. However, it became obvious that much of the extreme ultraviolet and x-ray emission is concentrated in active regions a few arcminutes in extent consisting of even smaller structures. Thus, good instrument resolution and pointing became decisive in order to make any sense of the observations.

A biaxial pointing system, stable in pitch and yaw, was designed by the University of Colorado in 1954 and improved versions were used for all ultraviolet and x-ray observations in the subsequent decade. A triaxial pointing system, developed at the Atomic Energy Research Establishment in England in 1965, prevented variations in roll. Pointing could now be stabilized to 2 arcsec. Today pointing stability is better than 1 arcsec on modern satellite platforms.

In this perspective satellite observatories may be regarded as a final step among the early technical developments. The first decade of satellite observatories saw increased sophistication in satellite platforms as well as in instrumentation. Space solar physics outgrew its beginnings and reached maturity with the instruments on the Apollo Telescope Mount on Skylab in 1973–4.

Satellites, balloons and rockets in the 1960s

The first registration of solar ultraviolet and x-ray radiation using instruments on a satellite came on the Soviet Sputnik 2 in 1957. The first US satellite with solar instruments, Vanguard 3 in 1958, was a failure. However, close observations of its orbit led to the discovery of the expansion and contraction of the Earth's ionosphere caused by diurnal and long-term variations in the solar ultraviolet and x-ray emission.

Solrad 1, launched in 1960, was the first in a series of satellites observing various wavelength bands in ultraviolet and x-rays. These satellites provided long-term coverage of wavelength and intensity changes of the solar radiation until 1976. Solrad demonstrated the strong solar variability in x-rays compared with the much lower variability in the ultraviolet. The satellite instruments did not have imaging capabilities in x-rays, but observations during the solar eclipse in 1966 showed that the solar x-ray emission was concentrated in small regions, less than 1 arcmin in diameter, i.e. the hot cores of active regions.

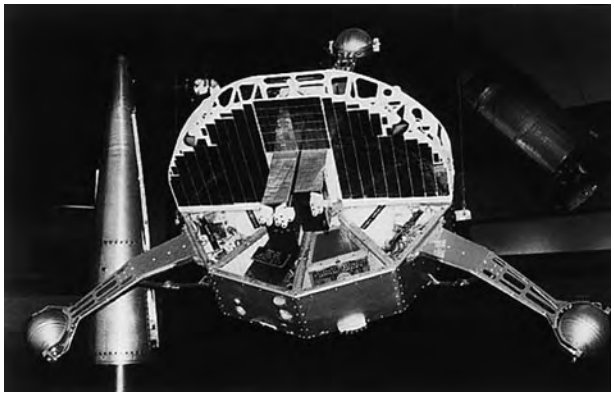


Figure 2. Model of the OSO satellites. OSOs 1–7 all looked similar.

1962 saw the beginning of a new series of solar satellites, the OSO missions (figure 2). The intent was to follow the Sun through an entire 11 year cycle with nearly identical ultraviolet and x-ray instruments. The OSO satellites had three axis-stabilized platforms. Pointing stability exceeded the imaging quality of the scientific instruments, but solar imaging was possible with a resolution of about 1 arcmin for the earliest OSOs. Studies could be made of coarse active region structures and of emission associated with the solar supergranulation in several spectral lines formed at different temperatures. OSO-8 (1975–78), the last in the series, came after Skylab (1973–4), differed from the earlier OSOs and is mentioned below.

Rocket experiments continued after 1960 in the US and in other countries. In France we may note the rocket and balloon flights in the 1960s and early 1970s that carefully measured the solar ultraviolet radiation from ∼200 nm down to Lyman alpha at 121.6 nm and its variation from Sun center to limb. The solar spectrum at these wavelengths comes from the upper photosphere, the lower part of the chromosphere and the temperature minimum layer between them. The strong discrepancy, as much as a factor of 10 at wavelengths of 150–180 nm, between the observed and theoretically calculated intensities inspired improved efforts at modeling these layers in the solar atmosphere. The lowest observed radiation temperature in the ultraviolet continuum at 150 nm was much discussed. Successive experiments found different results. This was not trivial since the value of the minimum temperature on top of the photosphere was considered to be connected to the amount of non-thermal energy passing from the photosphere to the corona and heating the corona.

The UK launched its first satellite for solar research, ARIEL 1, in 1962. Ariel 1 carried instruments that measured the solar spectrum in x-ray wavelengths, 0.4–1.4 nm. The enhancement of the solar x-ray intensity by factors of 10 or more during SOLAR FLARES was recorded on many occasions. This was of great interest at the time for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

understanding the enhanced ionization in the D-layer in the Earth's ionosphere during flares, leading to disruption of radio communications. Groups in the UK also took part with far-ultraviolet and x-ray instruments in the OSO program, i.e. OSO-4, -5 and -6, leading up to their later strong participation in SMM, Yohkoh and the Solar and Heliospheric Observatory (SOHO).

The UK Skylark rockets launched several ultraviolet instruments starting in 1964. In 1968 they obtained for the first time spectra of both the solar limb and the disk in the spectral range 150–220 nm. These observations were aimed at deriving the temperature structure of the outer solar atmosphere at ~100 000 K using a method that did not rely on uncertain instrument calibration, atomic cross sections or element abundances. Grazing incidence observations at shorter wavelengths, 15–80 nm and 1.5–5 nm, were added to study the corona at temperatures of 1 MK or more.

Slitless spectrographs launched into the total solar eclipse on 7 March 1970 by British and US investigators revealed that the emission from spectral lines emitting at temperatures of 10 000–300 000 K was not limited to a range in altitude of a few tens of kilometers as predicted by theoretical models, but extended over more than 1000 km. The true significance of this observation did not, however, make a strong impression at that time.

Finally, one might mention two tests of the NRL ultraviolet spectrometers, SO82A and SO82B, on Skylab. The test rocket for SO82A was launched during a solar flare on 4 November 1969. For the first time the small hot kernel of a flare was revealed. It was only a few arcseconds across and it showed up in highly ionized iron lines emitting at temperatures of several million kelvins. The corresponding test flight for SO82B in August 1970 led to the first published measurement of the large non-thermal widths of lines formed around 100 000 K, pointing to the presence of strong dynamics and waves in the transition region between the solar chromosphere and corona.

Early observations of the spectral region 20–150 nm showed that it contained a number of emission lines. Many of these lines had been identified and absolute intensities had been established. Analysis techniques had been developed in response to the new observations, to determine plasma densities and temperature structure of the emitting layers. In the resulting models the temperature above the chromosphere rose sharply from 10 000 K to 300 000 K in only about 100 km. It was at first not recognized that this result disagreed strongly with the observation that solar plasma in the 100 000 K range extended several thousands of kilometers above the chromosphere.

Skylab

SKYLAB is covered in a separate article in this encyclopedia and will only be described briefly here. The Skylab instruments were much larger than the corresponding OSO instruments. Angular resolutions in x-ray and extreme ultraviolet approached 2–4 arcsec. These high-resolution pictures showed that the solar transition region and corona are built up by magnetic loops containing the hot solar plasma. Solar flares occurred in the loops and the high-temperature components of flares were detected, having temperatures up to 20 MK. Loops and prominences were seen to tear loose from the solar surface to form what have later been termed coronal mass ejections (CMEs). CORONAL HOLES, extended regions where the emission in spectral lines from the corona was strongly depressed, were discovered and the occasional fast solar wind near the Earth, with wind speeds twice as high as the 'normal' solar wind, was found to be coming from the equatorial coronal holes. It was recognized that the Sun might have more permanent polar coronal holes.

Skylab missed out on dynamics and rapid time variability. Observations could be made only on a limited supply of photographic film or the photoelectric scans of wide fields of view took too long a time. Thus, high-cadence observations of loop structures were few. The Skylab spectrometers also lacked combined good spectral and spatial resolution. Thus, they could not detect the high velocities, 40–100 km s^{-1}, commonly occurring in the 100 000–500 000 K temperature range now observed with the ultraviolet spectrometers on the Solar and Heliospheric Observatory (SOHO).

Modern rockets and satellites

Many advanced instruments have flown in the more than 25 years, since Skylab. The Soviet Union had solar instruments on several satellites, an early example being OST-1, on Salyut 4 in 1975. This far-ultraviolet telescope, constructed by the Crimean Astrophysical Observatory, obtained new characteristics of plages and solar flares.

The Japanese Astro A, also known as HINOTORI, was launched in 1981. Its main objective was the detailed study of solar flares. One of its instruments imaged flares in x-rays, at energies of 10–40 keV or around 0.1 nm wavelength. Another instrument performed spectroscopy of x-ray flares in the wavelength range 0.17–0.2 nm, using a Bragg spectrometer. Investigations recorded the time profile and spectrum of x-ray flares and looked for gamma rays from flares at energies from 0.2 to 9.0 MeV. Hinotori was a forerunner for Japan's highly successful solar x-ray satellite, YOHKOH, launched in August 1991 and still in operation.

Orbiting Solar Observatory 8

OSO-8 was launched in June 1975 and operated until September 1978. It was the first solar satellite to attempt observations of the solar atmosphere with simultaneous high spatial and spectral resolution. It was also the first satellite to operate on a near-real-time basis with the science teams on the ground. It had two pointed instruments: an ultraviolet spectrometer, wavelengths 120–200 nm, and a multichannel ultraviolet and visible polychromator with six wavelength channels. These channels registered the strong resonance lines from singly

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ionized calcium and magnesium and the Lyman alpha and beta lines from hydrogen. The wavelength border regions of the channels contained visual and ultraviolet continua and the lines from doubly ionized silicon at 120.6 nm and five-times ionized oxygen at 103.2 nm. Thus, the instruments could study the solar plasma at temperatures from 6000 K to 250 000 K, i.e. the chromosphere and lower transition region. Spectral resolutions ranged from 2 to 10 pm (pm—picometer, i.e. 10^{-12} m). The best possible angular resolution of both instruments was 1–2 arcsec. However, the slit spectrometer took most of its data with a 2 arcsec × 20 arcsec slit and the multichannel instrument was mainly run in a coarse mode with 10 arcsec resolution. These two instruments were built and operated by the University of Colorado and the French National Space Agency, CNRS, respectively.

OSO-8 demonstrated that the energy in the 150–300 s oscillations was several orders of magnitude lower than the radiative losses from the upper solar atmosphere. Thus, the solar corona could not be heated by acoustic waves, the most commonly held view at the time. Persistent redshifts in lines emitted from the transition region, equivalent to down-flow velocities of $\sim$10 km s^{-1}, were also recorded for the first time. Since flows of this magnitude will drain the corona of gas in a few minutes, alternative explanations have been sought. The most promising involve disturbances generated near the top of magnetic loops by processes heating the corona, then progressing downward along the loop legs. Similar redshifts have been noted in stars and may be a general property of stellar atmospheres.

Figure 3. The HRTS, flying on the space shuttle as part of Spacelab 2, July–August 1985. HRTS is the long tube to the right in the figure.

High Resolution Telescope and Spectrograph

The High Resolution Telescope and Spectrograph (HRTS) (figure 3) had the high angular resolution not realized with OSO-8. HRTS was built and operated by the US Naval Research Laboratory and flew on 10 rocket flights between 1975 and 1997 and was a part of Spacelab 2 in 1985. On most flights the spectrograph operated at wavelengths between 117 and 170 nm, covering the full range or registering selectable narrow bands, 1.4 nm wide, centered on strong emission lines. HRTS was the first ultraviolet spectrograph with stigmatic imaging of an extended slit, 1000 arcsec long, at high angular resolution (1 arcsec).

HRTS made several important discoveries. Notable are the detection of explosive events on the first HRTS flight and the investigation into the possibly small filling factors for the emission in the solar atmosphere. Explosive events are small regions, with diameters of $\sim$2 arcsec, with a velocity dispersion amounting to $\pm$100 km s^{-1}. They are seen in the 100 000 K plasma and last typically 1 min. They could be caused by magnetic reconnection in the solar atmosphere and may thus be the observational signature of one of the main mechanisms proposed for heating of the solar corona.

The study of filling factors with HRTS was made possible by the good spatial resolution of the instrument and its spectral coverage that included density sensitive line pairs. From the derived densities, the total emission, and the overall size of the emitting structures it was concluded that only a small fraction of the volume is filled with plasma, i.e. filling factors amount to between 1% and 0.01% of the total volume of the structures.

HRTS also discovered that rapid down-flows of plasma in the transition region over sunspots are common, with in-flow velocities of 50–100 km s^{-1} or more. Another discovery was that transition region line profiles frequently have a multiple structure as if several distinct and different velocities exist inside the 1 arcsec resolution element of HRTS. This could be connected to the low filling factors if dynamic conditions are different in the small substructures that the transition region may consist of. Finally, clear connections were discovered between line intensities, wavelength shifts and line widths, and the underlying photospheric magnetic fields. The relation between fields and line shifts is asymmetric for red- or blueshifted profiles. A probable reason for an asymmetry would be if waves running in one predominant direction are present in the transition region.

Solar Maximum Mission

The SOLAR MAXIMUM MISSION (SMM) satellite (figure 4) was launched in February 1980. The primary goal was to make coordinated studies of solar activity, particularly solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

flares and eruptions, at a period of high solar activity. Over less than 10 years SMM observed more than 12 000 flares and over 1200 eruptions, called CMEs. The history of SMM was at times dramatic. The attitude control system of the satellite malfunctioned in January 1981 but was repaired in orbit in April 1984 and SMM continued observing until November 1989.

Four of the instruments on SMM registered the energetic solar radiation from gamma ray energies to soft x-rays. These instruments measured the spectral intensity of flares continuously over several years. A comprehensive set of coordinated data on this phenomenon was collected, throwing new light on all kinds of flare processes and on solar activity in general.

The Gamma Ray Spectrometer (GRS) provided gamma ray flare intensities as a function of time as well as the spectral distribution of the radiation. The Hard X-ray Burst Spectrometer (HXRBS) observed time series of hard x-ray bursts from flares in 15 energy channels between 20 keV and 260 keV (1 keV corresponds to a wavelength of 1.23 nm; higher-energy photons have proportionally shorter wavelengths). Thus spectra of the bursts could be built up. Continuous observations were made with a time resolution of 128 ms, but shorter intervals were possible. The soft X-ray Polychromator (XRP) monitored individual emission lines that are strong in active regions and flares. The selected lines give information on temperature, density, velocity, element abundance and non-equilibrium states in the flare and active region plasma. The Hard X-ray Imaging Spectrometer (HXIS) delivered simultaneous images of solar flares in six energy bands between 3.5 keV and 30 keV. The instrument had a coarse field of view corresponding to the size of an active region, with a spatial pixel of 32 arcsec, and a high-resolution field of view centered in the coarse field, with a spatial pixel of 8 arcsec. Time resolution could be automatically varied from 1.5 s in the early stages of a flare to 7 s in the decaying phase. Except for HXIS, which ended its life in November 1980, the high-energy instruments lasted until the end of the mission.

SMM also had three low-energy instruments that supported the studies of flares and solar activity. The Ultraviolet Spectrometer and Polarimeter (UVSP) produced monochromatic raster images at any wavelength between 115 nm and 360 nm with selectable wavelength bandwidth and angular pixel size down to 1 arcsec × 1 arcsec. Images in four lines could be observed simultaneously, or wavelength bands might be placed in the opposite wings of lines to measure Doppler shifts (i.e. velocities) or make polarization measurements. Time cadence for observations with a single pixel could be a fraction of a second. However, in most practical situations it took several minutes to build up a raster image. In April 1985 the UVSP grating drive failed. Observations were still possible but only at a fixed wavelength around 138 nm.

A main objective of UVSP was to study the flare plasma at temperatures below 200 000 K in the transition region lines available in its spectral range. However, all types of solar features were studied: prominences, solar active regions, sunspots, the quiet Sun. An interesting attempt was made to observe explosive events and relate them to solar magnetic fields. Since they have short lifetimes, explosive events might come and go in less time than it took to record a UVSP raster. Explosive events have a considerable velocity dispersion and could be detected by registering strong intensity variations in the far wings of the 154.8 nm line emitted at 100 000 K. Explosive events and other micro-flaring activity were found to be located in areas where magnetic fields of opposite polarity came close together and could well be caused by magnetic reconnection.

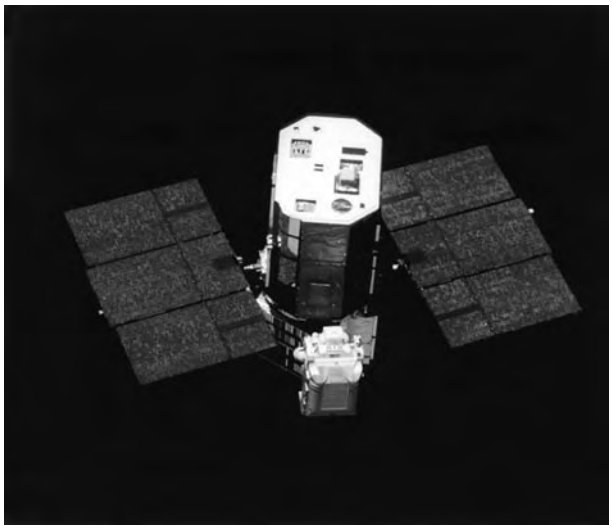


Figure 4. The SMM satellite. The aperture openings for the instruments are visible in the front plate. An astronaut is working on the satellite.

The High Altitude Observatory Coronagraph/Polarimeter on SMM produced images of a selected quadrant of the corona in the range from 1.6 to 6 solar radii with a spatial resolution of 10 arcsec set by the detector pixel size. Observations in seven wavelength bands in the visual spectral range made it possible to distinguish between various features of the solar corona and discriminate between ejected plasma at coronal and chromospheric temperatures. A major outcome of this instrument was a long-term study of CMEs and their relation to solar activity. SOLWIND, operating from March 1979 to September 1985, carried a similar coronagraph, also adding to our knowledge of coronal structure and dynamics.

CMEs were first detected with the coronagraph on OSO-7 and were routinely observed on Skylab. It soon became obvious that CMEs were the cause of strong gusts in the solar wind affecting the Earth's magnetosphere, causing beautiful polar auroras, but also potentially damaging effects to telecommunications and electric power transmissions. The SMM coronagraph had much better time coverage than Skylab, 87% for SMM

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

against 38% for Skylab. This allowed thorough statistical investigations of the connection between CMEs and solar activity, such as flares and eruptive prominences. Studies found that only half of the CMEs were clearly related to recognized active phenomena. Of the CMEs with such relations 40% were associated with flares, but more than 70% were associated with eruptive prominences. The measured average CME velocities were furthermore considerably lower than those found previously from Skylab. These new results led to a lively discussion on the origins of CMEs and the mechanisms causing them, a discussion that is still going on.

SMM also carried the Active Cavity Radiometer Irradiance Monitor (ACRIM). It measured the solar radiation from the entire solar disk integrated over all wavelengths (total irradiance). For the first time it became possible to measure variations in the solar radiation with time and thus with solar activity. This objective had been pursued from the ground for several decades, without reaching any definitive results (see also SOLAR TELESCOPES AND INSTRUMENTS: GROUND). However, with ACRIM the required accuracy of 0.1% was more than achieved. ACRIM later flew on several space shuttle flights and on the UARS satellite, where it is still in operation. Instruments supplementing ACRIM and giving similar results have been the less accurate ERB on the Nimbus 7 satellite and its successor ERBE.

ACRIM on SMM and in later flights have produced several exciting new results. These include the first unambiguous detection of a decrease in total solar irradiance when a large sunspot crosses the visible solar surface, the so-called 'sunspot deficit'. A corresponding 'excess' effect has also been demonstrated for active regions with large facular areas. Finally, ACRIM has found that the total solar irradiance varies with solar activity, showing a maximum in the period of highest solar activity around year 1990 and minima in 1986 and 1996. Typical rate of change is 0.015% per year. A possible real increase of 0.0036% per year between irradiance minima in 1986 and 1996 is less certain.

The 1990s

The 1990s have seen two extensive and very successful solar space observatories, the Japanese Yohkoh satellite, with US and UK collaboration, and SOHO, a collaboration between the European and US space organizations, ESA and NASA, with ESA as the main partner. The objective of Yohkoh is to study the high-energy radiation from solar flares as well as quiet structures and pre-flare conditions. The instruments on SOHO investigate physical conditions from the innermost core of the Sun to its outer corona and the heliospheric space. Two factors have particularly added to the quality of solar physics from Yohkoh and SOHO data. The first is the open data access and the collaborative spirit of the investigators. Secondly, the real improvement of SOHO and Yohkoh over earlier missions comes more from the high time resolution and continuous coverage than from the modest increase in instrument resolutions.

The Upper Atmosphere Research Satellite (UARS), launched in September 1991 and still in operation, is intended for studies of the Earth's atmosphere but contains relevant solar instruments. ACRIM has been mentioned. SOLSTICE and SUSIM on UARS observe the solar ultraviolet spectrum from 112 nm to 440 nm. The emphasis is on highly accurate absolute calibration that is stable over several years. The record from SOLSTICE and SUSIM constitutes a reference spectrum of the full Sun in ultraviolet and its variation with solar activity.

The main instruments on Spartan 201 are an Ultraviolet Coronal Spectrometer (UVCS) of the type flown on SOHO, and a coronagraph observing in the visual wavelengths. Spartan 201 has flown on five occasions between April 1993 and October 1998, partly supporting observations with Ulysses and SOHO.

Ulysses was launched in October 1990 (see SOLAR WIND: ULYSSES). It is the first satellite to investigate solar and heliospheric conditions at high solar latitudes. It carries no instruments for remote sensing of the Sun, as had been the original plan, but registers heliospheric magnetic fields, plasma waves, dust and solar wind particles, their speed, composition and stage of ionization. The trajectory of Ulysses first took it to Jupiter, where the gravity of the planet accelerated the satellite out of the ecliptic plane and into an orbit at nearly 90° inclination with the ecliptic. Ulysses passed over the south solar pole in June–November 1994 and over the north pole a year later, near solar activity minimum. The measurements showed the well-known slow solar wind, velocity 400 km s^{-1}, at low latitudes. However, at latitudes above 30° this was replaced by the fast solar wind, with a speed of 750 km s^{-1} or more. To the fast winds from equatorial coronal holes had now been added a fast wind streaming out of the coronal holes in the polar caps. Furthermore, this wind fanned out quickly above the solar surface and extended to latitudes much lower than the borderlines of the polar coronal holes, lying near 60° latitude. The magnetic fields from the Sun must similarly fan out in the heliosphere. It will be interesting to see what Ulysses will register on its second pass over the poles in 2000–1 at maximum solar activity when polar coronal holes will be more 'filled in'.

A final mission in the 1990s was the Transition Region and Coronal Explorer (TRACE), launched in April 1998. TRACE observes the solar corona with an unprecedented angular resolution, 1 arcsec in the temperature range from 1 MK to 2.5 MK. The solar emission is observed in EUV wavelength bands centered at 17.1 nm, 19.5 nm and 28.4 nm. Other wavelength channels isolate emission from the chromosphere (10 000–30 000 K) in Lyman alpha from hydrogen, in the resonance lines from three-times ionized carbon at 155 nm (100 000 K) and in the near ultraviolet and white light continuum.

The future

Plans for the future are many and for most part uncertain. The Japanese SOLAR-B satellite is a likely project, scheduled

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to be launched in 2004. This satellite will contain a visual telescope to measure photospheric magnetic fields with extremely high spatial resolution and ultraviolet and x-ray telescopes to measure and monitor the coronal emission. The mission aims at understanding the detailed relationship between magnetic fields and coronal heating and emission in the Sun.

NASA has sketched an entire 'road-map' of satellites. Only a few of these are likely to fly and at present none is securely financially confirmed. Interesting possibilities include a solar probe, going to 4 solar radii to make *in situ* observation of the acceleration region of the solar wind, and Stereo, which will be a pair of satellites flying in the Earth's orbit around the Sun at a significant distance in front of and behind the Earth. The steroscopic view of solar and heliospheric phenomena will allow a three-dimensional mapping of CMEs.

Europe and ESA are considering a solar orbiter, flying at 45 solar radii from the Sun, and in an inclined orbit to the ecliptic reaching heliographic latitudes of 40°. These plans are still at a preliminary stage, and joining the US in a continuation of SOHO with instruments having super-high resolution should also be considered.

Bibliography

Lemaire P 1987 La physique solaire *Atlas de L'espace* (Encyclopaedia Universalis) pp 216–19

Liller W (ed) 1961 *Space Astrophysics* (New York: McGraw-Hill)

Massie H and Robins M O 1986 *History of British Space Science* (Cambridge: Cambridge University Press) pp 90, 200–2, 339–53

Strong K T, Saba J L R, Haisch B M and Schmelz J T (ed) 1999 *The Many Faces of the Sun, a Summary of the Results from NASA's Solar Maximum Mission* (New York: Springer)

Tousey R 1961 Solar spectroscopy in the far ultraviolet *J. Opt. Soc. Am.* **51** 384–95

Olav Kjeldseth-Moe

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Transition Region

The solar transition region can generally be defined as those plasmas at temperatures between 2×10^4 and 1×10^6 K, a temperature regime bridging the CHROMOSPHERE and CORONA. Prior to the space age, the transition region was unobserved although the existence of the higher-temperature corona was known. Spectral lines formed at transition region temperatures are found primarily below 2000 Å where they are strongly absorbed by the Earth's atmosphere. It was termed the transition region since it was thought to be a thin region of the atmosphere where an abrupt transition from a relatively cool, dense plasma to a hot, highly ionized, tenuous extended plasma occurred. Understanding the solar transition region is an important piece of the great puzzle of determining how the outer solar atmosphere is heated. More recently, interest in the transition region has increased as it is recognized that these plasmas are highly dynamic, displaying such events as jets, explosive events and high-speed flows.

Spectroscopy of the transition region

Observation of the solar transition region requires the use of spectroscopic techniques to isolate spectral lines formed at transition region temperatures. For example, helium has three ionic species: He I, neutral helium, He II, singly ionized helium, and He III, completely ionized helium. In the solar transition region, all three species are present but He II is the dominant species. The ionization state, the relative fractions of the three species, is determined by a balance between the competing processes of ionization and recombination of free electrons with the ions. At 6×10^4 K, He II is the most populous helium ion since at that temperature the free electrons have sufficient energy to ionize neutral helium (He I) but not enough energy to strip away the remaining bound electron to transform He II into completely ionized He III. The emission of He II spectral lines is produced by the collisional excitation of the bound electron to excited levels which then spontaneously decay to lower levels. The Lyman α line (the 1s–2p transition) of He II at 304 Å is a very strong line. It is much stronger than other lines at nearby wavelengths so that it can be fairly easily isolated and observed. Images of the Sun in the light of He II λ304 are shown in figures 1 and 2. These images were obtained with the Extreme-Ultraviolet Imaging Telescope (EIT) instrument on the Solar and Heliospheric Observatory (SOHO) which uses narrow-bandpass coatings on the telescope mirrors to isolate strong spectral lines. Figure 1 was obtained near the minimum of the solar activity cycle and figure 2 was obtained when the Sun was more active.

Temperatures, Densities, Emission Measures, Elemental Abundances

Essentially all of our information about the solar transition region comes from the analysis of spectral lines. Figures 1 and 2 show that much can be learned from images of the intensity of spectral lines. More detailed observations of the transition region are made with spectrographs which can disperse the solar spectrum in order to pick out spectral lines of particular interest.

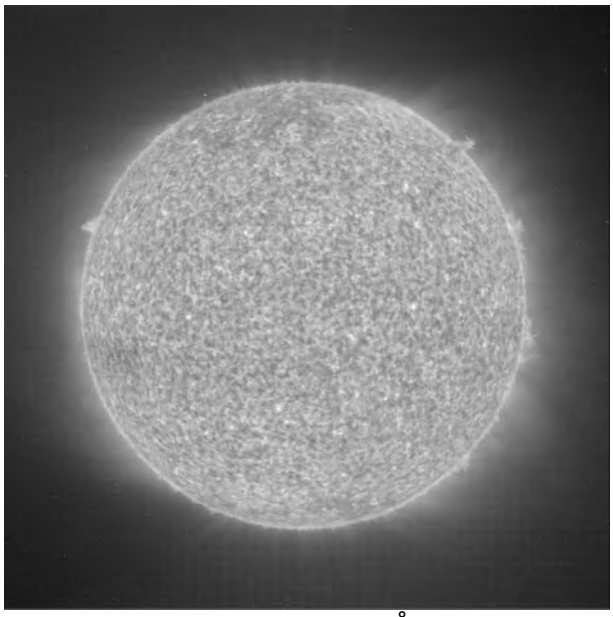


Figure 1. The solar transition region near the minimum of the solar activity cycle seen in He II λ304.

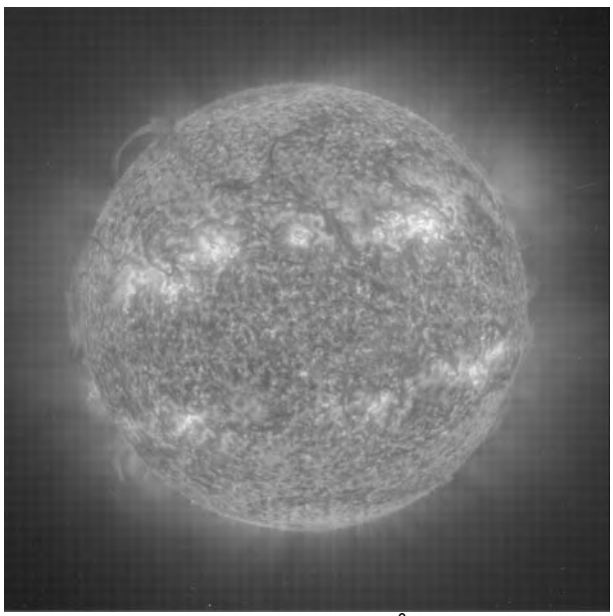


Figure 2. The solar transition region as the Sun approached the maximum of the solar activity cycle seen in He II λ304.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The most direct piece of information that can be obtained about a spectral line is its intensity. The intensity of a line depends on how efficiently the line is produced and how many of the ions that produce the line are in the observed plasma. In the transition region and corona, most lines are produced by collisional excitation of the ion by free electrons. The emission rate is the product of two factors, the electron flux (the electron density multiplied by the electron velocity) and the collision cross section. The cross section is typically large for allowed or resonance transitions and small for forbidden transitions. The number of ions producing the line in the observed plasma is given by the products of the relative population of that ion with respect to all other ions of that element, the density of that element with respect to hydrogen (the elemental abundance) and the density of hydrogen (essentially equal to the electron density) in the observed plasma. Consequently, if we know the temperature, from the calculated ionization balance, the elemental abundance (from other measurements) and the collision cross section, from calculations, we can derive the emission measure ($\int N_e^2 \, dV$) at a given temperature directly from the line intensity. The emission measure provides considerable information about the temperature and density structure of the solar plasma.

A variety of techniques for determining the temperatures and densities of transition region and coronal plasmas are discussed in the article SOLAR SPECTROSCOPY AND DIAGNOSTICS.

Velocity measurements

The spectroscopy of the solar transition region can be taken one further step further to look at spectral line profiles. The object of this is to use the DOPPLER EFFECT to provide information about the velocities in the regions producing the spectral line. If a plasma is moving along the line of sight to the observer, the spectrum of the emitting plasma is shifted in wavelength by the amount $\Delta\lambda = (v/c)\lambda$ where v is the velocity of the observed plasma along the line of sight, and c is the speed of light (in the same units as v) and λ is the rest wavelength of the spectral line (in the same units as $\Delta\lambda$). Plasmas moving toward the observer are shifted to shorter wavelength (blueshift) and plasmas moving away from the observer are shifted to longer wavelength (redshift). If these shifts are observed near disk center, a blue shift corresponds to an outflow and a redshift corresponds to a downflow. At the limb, these shifts indicate horizontal flows.

Figure 3 shows profiles of C IV λ1548 and λ1550 in an active region and sunspot. The profiles are dispersed in wavelength from left to right and position in the quiet Sun runs from top to bottom. Large redshifts can be seen in the sunspot and indicate downflow velocities of 100–150 km s^{-1}. Above the sunspot spectra, one can also see a very small region with exceptionally wide line profiles. These are signatures of explosive events which will be discussed later. These spectra were obtained with the High Resolution Telescope and Spectrograph (HRTS) during a rocket flight.

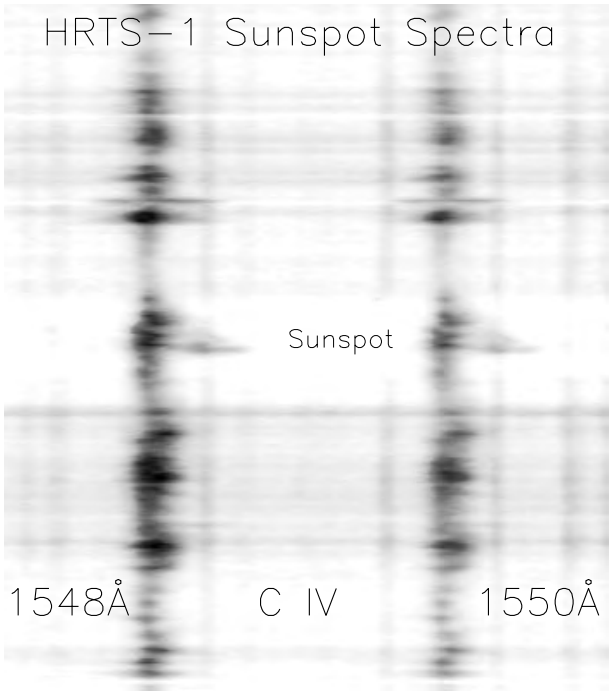


Figure 3. HRTS spectral line profiles of C IV, formed at 1×10^5 K, in an active region and sunspot.

Spectral observations often show that the line has a width that is wider than the spectral resolution of the spectrometer. Part of this excess line width is caused by the motions of the ions which follow a Maxwellian velocity distribution specified by the temperature of the plasma. On examination, one finds that transition region profiles are even broader than would be predicted from the temperature of the ion that emits the line. This excess broadening is called the nonthermal width of the line and is apparently due to motions in the plasmas on spatial scales below the spatial resolution of ultraviolet spectrographs currently in operation. The thermal and nonthermal broadening results in line profiles that have a shape that is nearly Gaussian. Also evident is the net Doppler shift of spectral line profiles of the order of 10 km s^{-1} caused by plasma flows in spicules, loops or other structures. Even more interesting are the occasional line profiles that are non-Gaussian and highly Doppler shifted. These are found in more spectacular events such as jets, ejecta and explosive events.

Elemental abundance variations

For many years, it has been assumed that the relative abundance of elements in solar plasmas does not vary. For example, the standard abundance of iron is about 4×10^{-5} that of hydrogen. The abundances are the result of nuclear processes that ended many years ago. For every parcel of solar plasma, this ratio is now expected

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to be the same. Once the variability of relative elemental abundances was considered a possibility, it was found that relative abundances in the corona were not the same as in the photosphere, that abundances seemed to change during flares and structures with nonstandard abundances were found fairly commonly. The key ingredient in the elemental abundance variations is the first ionization potential (FIP) of the element, or, the energy needed to ionize the neutral species of the element. Elements with low FIP values, such as Mg, Ca, Si and Fe, are enhanced in coronal plasmas relative to photospheric plasmas. The fact that the effect is correlated with the FIP indicates that the separation occurs at fairly low temperatures.

A more detailed discussion of this topic is given in SOLAR TRANSITION REGION.

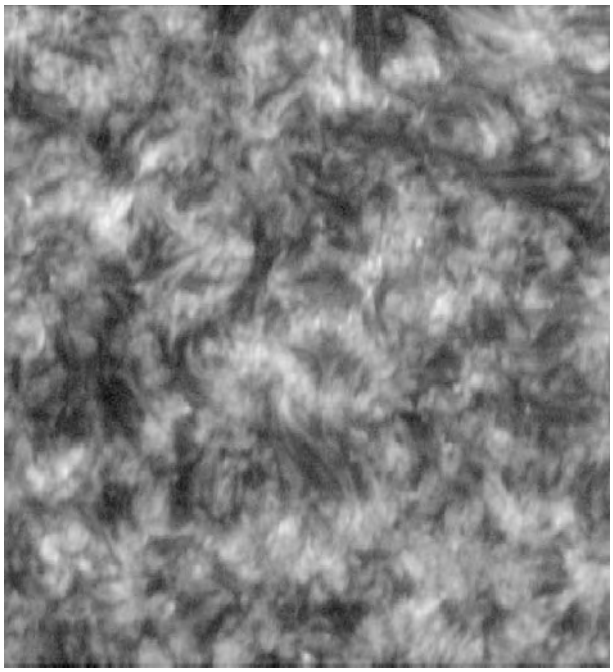
Structure

Large-scale structure of the transition region

Lines formed at transition region temperatures are emitted in most of the large-scale structures that appear on the Sun: the quiet Sun, coronal holes, active regions, prominences, sunspots and solar flares. All of these structures are a result of the interplay between magnetic fields and plasmas.

In the quiet Sun, the magnetic fields are dominated by fields associated with the supergranulation. Here, the supergranular motions in the photosphere sweep the magnetic flux elements from the supergranular cell centers to the cell boundaries. The new magnetic flux that emerges in the cell centers appear to include both positive and negative polarities (upward- and downward-pointed magnetic fields), as if they were the emergence of small kinks in a magnetic flux rope. Consequently, the magnetic fields at the cell boundaries are also bipolar. As a result of the continuous reconnection of magnetic fields, the net result is a complex web of magnetic field lines connecting opposite magnetic field polarities. This magnetic field structure is reflected in the patterns of transition region line emissions. At relatively low temperatures, mostly the base of these short loops is outlined. The corresponding structures in the chromosphere are spicules. At 10^5 K, for example in C IV lines, structures that are the extensions of the cooler chromospheric models are typical. At higher transition region temperature, longer and more complete loops become apparent. A high-resolution image of the quiet transition region obtained with the SUMER instrument on SOHO is shown in figure 4.

The magnetic field in coronal holes is predominantly of one polarity so that the large-scale field extends outward from the Sun and participates in the high-speed solar wind streams. The unipolar nature of coronal hole fields is the result of remnants of active region fields that are swept to the poles during the declining phase of the solar cycle. Nevertheless, the fields in the coronal hole are still pushed around by a supergranular flow pattern that arranges the fields into a supergranular pattern similar to that found in the quiet Sun, except that the fields are mostly of a single polarity. Bipolar fields continue to emerge in the cell centers but, since these have no net polarity, the polarity of the coronal hole fields is unchanged. In transition region lines, the intensity pattern is still that of the supergranular network although it is somewhat less intense than in the quiet Sun. A characteristic structure of the transition region in coronal holes is the macrospicule. These are similar to chromospheric spicules but are about 10 times larger. They typically jet out above the limb and either fade in place or fall back towards the photosphere. Macrospicules can be seen at the top and bottom of the solar images shown in figure 1.



SUMER Image of the Quiet Sun in O VI

Figure 4. The quiet transition region seen in O VI, formed at 3×10^5 K, obtained with the SUMER instrument on SOHO.

SOLAR ACTIVE REGIONS, consisting of sunspots and plage regions, contain the most concentrated large-scale magnetic fields. The strongest fields are in the sunspots but these have no clear signature in transition region lines. Sunspots on average are neither particularly bright nor dark and cannot be located simply from an image of the transition region. Active region plages are often simply regions of relatively bright emission, particularly at low transition region temperatures. For increasing temperatures, the bases of coronal loops and extended portions of loops are outlined. As temperatures approach 1×10^6 K, the loops become more complete as is seen in coronal lines.

SOLAR PROMINENCES are relatively cool plasmas that are typically situated along magnetic 'neutral' lines where the dominant photospheric magnetic polarity reverses. Prominence neutral lines are found in active regions and in the quiet Sun. A number of prominences can be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

seen in figure 2 in He II. On the disk they are relatively dark and above the limb they are relatively bright. The apparent reason for their darkness on the disk is that they contain cool material with neutral hydrogen and helium with absorbs the emission of extreme-ultraviolet lines. However, this is a matter of current research. Prominences also often erupt from time to time as a component of coronal mass ejections.

Transition region loops and coronal heating

The heating of the transition region and corona is a central, long-standing problem in solar physics. The radiative losses of transition region and coronal plasma occur fast enough that the outer atmosphere of the Sun would collapse in minutes if there was not a constant supply of energy deposited into these plasmas to maintain their high temperatures. The solar transition region is a component of a complex magnetohydrodynamic system. In other words, the temperatures, densities, velocities and magnetic fields of the corona and their evolution are all governed by the equations of magnetohydrodynamics. A complete physical understanding of the solar transition region is only possible if the complete system is understood. The problem of coronal heating encompasses this complete system including the photosphere, chromosphere, transition region and corona (see CORONAL HEATING MECHANISMS). Perhaps the greatest difficulty lies in our inability to measure the Sun's magnetic field above the chromosphere where much of this heating takes place.

The simplest models of the transition region and corona consider the magnetic field as a passive element that constrains plasma flow and heat conduction along the magnetic field lines. The static energy balance models of coronal loops assume that the plasma located along a magnetic field line resembling a coronal loop is uniformly heated along its length by some unspecified mechanism. At the top of the loop where the maximum coronal temperature is reached, the energy input balances the radiative losses and the thermal conduction losses down the field lines to the cooler parts of the atmosphere. In the transition region, the radiative losses are much larger than the local heating rate and are balanced by a dissipation of the thermal conduction flux from the hot corona. These models generally produce an extended corona, resembling the observed CORONAL LOOP structures, and a thin transition that supports the high conductive flux needed. Emission measures of transition region lines, derived from observations, indicate that the volume of the emission measure is small and, in some sense, consistent with a thin transition region.

However there are a number of difficulties with this model. The observed emission measure distribution of the transition region reaches a minimum around 10^5 K with increasing values toward higher and lower temperatures. The static energy balance models are unable to explain the transition region emission measure distribution at temperatures below about 3×10^5 K, in other words, most of the transition region. All of the conductive flux from the corona is radiated away before reaching the middle of the transition region, much less the chromosphere. Further, high-resolution observations of the transition region indicated that it is considerably more extended than predicted by these models. This indicates that it cannot maintain the necessary conductive flux and must itself be locally heated.

Other problems indicate the need for more sophisticated models. For example, the general downflow of transition region plasmas cannot be addressed by the static models but are the basis of steady flow models. These models are capable of reproducing the observed downflows in only the most artificial way and still produce a very thin transition region. The solution to problems of this sort led to models that could include time-dependent flows and heating.

Many of the one-dimensional models continue to predict a very thin transition region. One consequence of this is that hot electrons from the corona are able to penetrate to considerably lower temperatures. In this case, the electron velocity distribution is no longer Maxwellian and greatly complicates the construction of physical models and the spectroscopic diagnostics of these plasmas. The observation that the transition region has a rather shallow temperature gradient indicates that this may not be a real problem but it cannot be totally discounted at this point.

A further discussion of TRANSITION REGION MODELS is presented by V Hansteen.

Very-fine-scale structure and coronal heating

One of the more intriguing properties of the solar transition is the need to invoke the existence of very-fine-scale structures that only sparsely fill the observed transition region structures. In other words, when the transition region is observed at high spatial resolution, it is possible to discern such discrete structures as loops, macrospicules etc. We feel confident that we can measure the volume of such an object. Density-sensitive line ratios are then used to derive electron densities. With the electron density and volume, the emission measure ($\int N_e^2 \, dV$) can be calculated and compared with the emission measure determined simply from line intensities. This comparison generally indicates that the actual volume of emitting material must be much less than what is observed in images of these structures and the actual volume must be only about 1% of the 'observed' volume or less. This can be visualized as the break-up of transition material into thin filamentary strands that follow the magnetic field lines but only fill a fraction of the transition region volume. When a similar analysis is applied to coronal loops, the coronal material seems to completely fill the observed structure. However, the coronal measurements are perhaps not so definitive as in the transition region.

As we will see below, there is evidence for flow patterns in the transition region that are also on very fine spatial scales. These fine-scale structures and flows may be significant in trying to understand how the solar transition region is heated.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

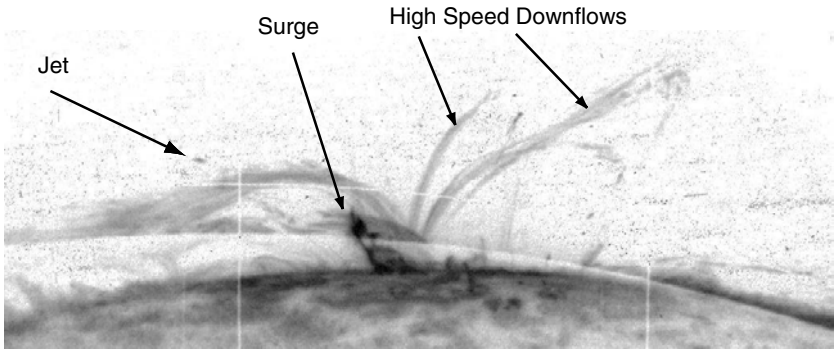


Figure 5. HRTS C IV image of the transition region in an active region at the limb.

Dynamics

As the spectroscopic capabilities of solar ultraviolet instrumentation have improved, it has become clearer that the plasmas at transition region temperatures are highly dynamic. Time sequences of images of transition region structures in the quiet Sun show a continual rearrangement of the network elements and the continual intensity changes in these elements, called 'blinkers'. In the quiet corona, there seems to be a more gradual evolution of the hot structures. Doppler information in transition region spectra also reveals a variety of flows and ejecta, indicating a truly dynamic state.

Flows

Smooth laminar flows of transition region plasmas apparently along magnetic field lines are commonly observed. The speed of sound in solar plasmas varies roughly as the square root of the temperature and has a value of about 40 km s^{-1} at 10^5 K. Flow speeds in the transition are usually subsonic and, in fact, are usually downflows. The typical downflow velocity is about 6 km s^{-1} both in the quiet Sun and in active regions and may approach about 20 km s^{-1}. The net transition region downflow is difficult to understand. There is a net outflow of solar plasma into the solar wind, but the flow velocity associated with this mass loss would be quite small in the transition because of the relatively high densities there. Otherwise we would expect to see a general balance between upflowing and downflowing material but this is not the case.

Persistent supersonic downflows with velocities of 100 km s^{-1} or greater are often seen in and near sunspots, as seen in figure 3. The suspected source for these downflows is the large-scale filamentary transition region structures seen above active regions at the limb. Figure 5 shows filamentary C IV structures above an active region where a surge is in progress.

Further information on TRANSITION REGION FLOWS can be found in the article by P Brekke.

Explosive events

Explosive events have one of the most dramatic signatures in transition region spectra, as seen in figure 6. These

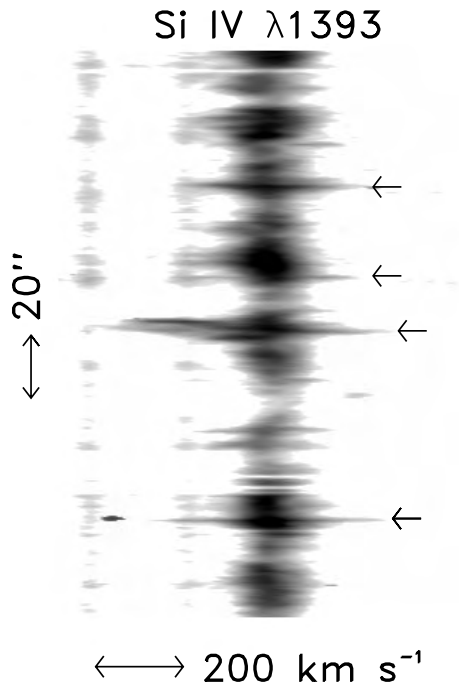


Figure 6. HRTS profiles of Si IV, formed at 7×10^4 K, showing signatures of explosive events in the quiet Sun (arrows).

are most noticeable from their Doppler shifts of about 100 km s^{-1} to the red and/or blue wing of the line in small regions only about 1–2 Mm in extent. Their typical lifetime is about 60 s. The origin of these events can be attributed to magnetic reconnection in a variety of situations. They occur in association with rapidly emerging magnetic flux in active regions, suggesting that the plasma flows are the RECONNECTION jets predicted by the emerging flux flare model. When the positions of explosive events in the quiet Sun are mapped, it is clear that they tend to occur adjacent to magnetic flux elements in the supergranular network boundary. This suggests that explosive events are the result of magnetic reconnection as intranetwork fields are driven by the supergranular flows into the network

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

boundaries. The details of this scenario are now being examined with the SUMER data from SOHO which has been able to directly observe explosive events in conjuction with magnetic flux reconnection.

Further information on EXPLOSIVE EVENTS can be found in the article TRANSITION REGION: EXPLOSIVE EVENTS by D Innes.

Ejections

One of the more puzzling aspects of explosive events is that the accelerated plasma is not observed to travel very far. Nevertheless, there are many examples of transition region plasmas that are ejected outward into the corona. Doppler shifts of small 7 Mm loops that are repeatedly accelerated to 500 km s^{-1} are observed. Doppler blueshifts (outflows) are commonly found at the top of the supergranular network indicating a continual process of plasma ejection taking place there. Macrospicules, with outward velocities as high as 100 km s^{-1}, are a common feature of coronal holes.

In active regions, SURGES are often observed in chromospheric lines such as Hα as well as in the transition region. These are tongues of plasma that are shot out of active region at velocities as high as 100–200 km s^{-1}. Small plasmoids have been observed to traverse large portions of an active region at velocities of 100 km s^{-1}. Figure 5 shows examples of a surge, a jet and prominence material, all in a single observation. Prominences often erupt and form an essential part of a coronal mass ejection. In some cases, the prominence eruption is the first sign of an incipient mass ejection.

Very-fine-scale dynamics

The nonthermal component of spectral line broadening is a signature of random plasma motions with a characteristic velocity of about 20 km s^{-1}. The typical variations in the velocities of the observed larger-scale flows are only on the order of 6 km s^{-1}. Consequently, the fine-scale motions are not simply the continuation of a distribution of velocities over a wide range of spatial scales, as might be expected for a turbulent plasma. The large-scale velocity variations do have the characteristic velocity spectrum of a turbulent Kolmogorov plasma but the fine-scale velocity variations predicted by this spectrum are much less than the observed value of 20 km s^{-1}.

The evidence for fine-scale motions and structures appears to offer a clue for solving the problem of coronal heating. Where this energy comes from and how it is dissipated is a complex problem with no certain solution at this time. For most coronal heating models the source of the energy is the fluid motions of plasmas in the solar convection zone which contort the magnetic fields in a way that amplifies their field strength and creates magnetic topologies that are at a higher energy level than the minimum energy potential field state. As a result, there are currents and mass motions in the outer atmosphere of the Sun that can be dissipated into thermal energy to replace the losses due to radiation. For example, the currents could be dissipated by Joule (resistive) heating and the mass motions through viscous dissipation. The main problem is that the corona is highly conducting and not very viscous. The electrons are quite free to respond to electric fields so that a potential drop is very difficult to maintain. The main source of resistivity is collisions with the heavier ions but this is not very effective. The induced magnetic field from a current sufficient to maintain the corona would result in magnetic loop structures that are extremely twisted, much more so than observed coronal loops which, at most, display only a mild degree of twist.

One way around this problem is to invoke small-scale currents and flows. One suggestion is that coronal heating occurs in many fine-scale magnetic reconnection events called 'nanoflares'. The small-scale motions deduced from transition region line profiles may be evidence for these nanoflares. The nanoflare theory has also been studied with computer simulations which seem to confirm the idea that motions induced by the large-scale photospheric driving motions result in small-scale stochastic energy release events that resemble nanoflares. The small-scale motions and structures in the solar transition region are probably consistent with the nanoflare theory of coronal heating but, as with many of these issues, this is a topic of current research.

Bibliography

Cook J W and Brueckner G E 1991 Fine structure of the solar transition region—observations and interpretation *Solar Interior and Atmosphere* ed A N Cox, W C Livingston and M S Mathews (Tucson, AZ: University of Arizona Press) pp 996–1028

Mariska J T 1992 *The Solar Transition Region* (Cambridge: Cambridge University Press)

Parker E N 1987 Why do stars emit X-rays? *Phys. Today* **40** (7) 36

Kenneth Dere

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind

The continuous outward flow of charged particles (predominantly electrons and protons) from the Sun into interplanetary space. This outflow of plasma 'blows' past the planets at speeds that fluctuate between 200 km s^{-1} and 900 km s^{-1}, and causes the Sun to lose about one million tonnes of mass per second. At a distance of 1 AU from the Sun, the mean density of the solar wind is about 5×10^6 particles per cubic meter and its temperature (a measure of the random kinetic motions of the solar wind particles relative to each other) is about 10^5 K.

Close to the plane of the ecliptic, the average wind speed is about 400 km s^{-1}. The particles that comprise this 'slow' wind emerge predominantly from streamer-like structures in the solar corona that tend to be located relatively close to the solar equator. A fast wind, with a mean speed of about 750 km s^{-1}, flows out of coronal holes that are centered permanently on the solar poles. When coronal holes extend down to, or across, the solar equator, high-speed streams flow out past the Earth and planets. The combination of the Sun's axial rotation and the Earth's orbital motion causes these streams to recur at intervals of 26 to 27 days as viewed from the Earth. Results obtained by the SOHO spacecraft indicate that solar wind particles originate at the boundaries of supergranular cells on the solar surface, where magnetic fields are concentrated.

The solar wind carries with it lines of force of the solar wind, which spread out to form the weak interplanetary field. The Sun's rotation, combined with the radial outward flow of the solar wind, causes the interplanetary field lines to take up a spiral form. The solar wind and interplanetary field interact with the magnetospheres of planets and the tails of comets. The pressure exerted by the solar wind squeezes a planetary magnetosphere inwards, towards the planet, on the Sun-facing side, and drags it out into a tail on the 'downstream' side. The wind drags ions out of the head of a comet to form a long ion tail. Fluctuations in the wind and interplanetary field, caused, for example, by bursts of particles from flares and by coronal mass ejections, distort planetary magnetospheres and sometimes cause ion tails to disconnect from their parent comets.

The solar wind continues to flow outward until it is halted by the weak pressure exerted by the interstellar medium. The boundary of the region within which the solar wind and interplanetary field predominate (the heliosphere) is called the heliopause and is believed to lie at a distance of about 100 AU. The radius of the heliosphere is expected to vary with the solar cycle.

See also: corona, coronal holes, flare, granulation, magnetosphere of Earth, SOHO (Solar and Heliospheric Observatory), solar coronal mass ejection.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

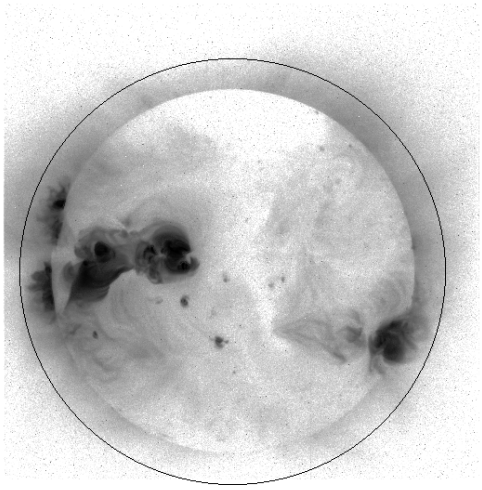


Figure 1. Soft x-ray Yohkoh image from 12 April 1993. The image shows a very distinct low-emissivity coronal hole region (large white area surrounding the north pole and stretching all the way to the equator). There is also a coronal hole present at the south pole, but this is veiled by denser material from surrounding regions and is therefore not visible at the time of the observation (see YOHKOH.)

Solar Wind Acceleration

Solar wind acceleration is the increase of the flow speed of the solar wind plasma from close to zero in the lower SOLAR ATMOSPHERE to the values observed in interplanetary space, which can range from 200 km s^{-1} for the slowest speeds in the ecliptic plane to 750 km s^{-1} for streams originating in the large polar CORONAL HOLES. The exact height in the solar atmosphere at which the plasma starts its outflow is still being debated. Until recently it was thought that the solar atmosphere was more or less static out to distances of 5 to 10 solar radii. With new observations from both ground and space based instruments, it has become clear that at least the fast solar wind streams accelerate at much lower heights in the solar atmosphere than previously thought, starting their outflows probably even below the TRANSITION REGION.

Properties of solar wind streams

Solar wind streams that differ significantly in speed will also differ in other plasma properties such as density and temperature. Different parts of the global solar magnetic field seem to give rise to streams with different properties. The large-scale magnetic field on the Sun can be approximated to first order by a dipole. Near the dipole equator inside of 1.5 to 3 solar radii the magnetic field is strong enough to withstand the outward plasma pressure, the magnetic field lines are therefore closed, and the plasma is confined in these regions called CORONAL STREAMERS. Further away from the Sun the magnetic field becomes weak compared with the plasma pressure, and the plasma forces the field lines to open, permitting a steady outflow of plasma into interplanetary space. These open field lines fill the space between the streamer boundaries and the solar poles. Plasma conditions in the open field regions differ from the conditions in the streamers. Observations show that electron densities and temperatures as well as emission at visible, UV and x-ray wavelengths are lower in the open field regions. Regions of extremely low emission are called coronal holes. The fastest solar wind originates from the large polar coronal holes which, particularly during solar minimum activity, might have extensions all the way to the solar equator (see figure 1, white areas on the solar disk). Measurements in interplanetary space show that this high-speed solar wind has rather constant plasma properties. At the Earth's orbit the flow speed is 750 km s^{-1} (which is 2.7×10^6 km h^{-1}), the proton density is about 2×10^6 m^{-3}, and the electron and proton temperatures are $(1–2.5) \times 10^5$ K respectively. The abundance of alpha particles relative to protons is about 4.5%. The alpha particles are almost six times hotter than the protons and exceed their speed by about 50 km s^{-1}, which is close to the local Alfvén speed. The small traces of other heavy ions, such as O^{6+}, C^{5+} and Mg^{10+}, that are also found in the solar wind, flow with the same speed as the alpha particles. The increase in the flow speed between 0.3 and 1 AU was determined during the HELIOS I and II era to be of the order of 1% in the high-speed wind. Generally plasma properties of the fast solar wind originating from the polar coronal holes vary by less than 10%. (For a more detailed description of the HELIOS I and II solar wind probes see Schwenn and Marsch (1990, 1991).)

In reality the large-scale magnetic field is more complicated than a pure dipole field and several closed field regions can be present at the same time (see SOLAR WIND: MAGNETIC FIELD). The slower solar wind originates from the open magnetic field regions on top of the different streamers, the edges of the streamers and the regions between streamers and coronal holes. These slower streams have plasma properties that are extremely variable. Their speed ranges from 200 to 600 km s^{-1}, their density from 2×10^6 to more than 10×10^6 m^{-3}, the electron and proton temperatures range from $(10–20) \times 10^4$ and $(2.5–15) \times 10^4$ K respectively. Heavy ions in these streams flow either with the same speed as the protons or are slightly slower. The ratio of the alpha to proton temperature is reduced to 3, and the abundance of alpha particles relative to protons ranges from close to 0% to almost 20%. The increase of the flow speed between 0.3 and 1 AU is of the order of 10%, thus larger than in the high-speed wind.

Flow speeds in the near Sun regions

The sound speed in the solar wind plasma is about 40 to 70 km s^{-1} at the Earth's orbit. In solar wind, even the slowest streams are therefore highly supersonic. The transition from sub- to supersonic occurs somewhere close to the Sun, probably between 1.5 to 3.5 solar radii in the high-speed polar wind, and between 3.5 and 8 solar radii in the slower streams. The exact distance from the Sun at which the solar wind speed reaches the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

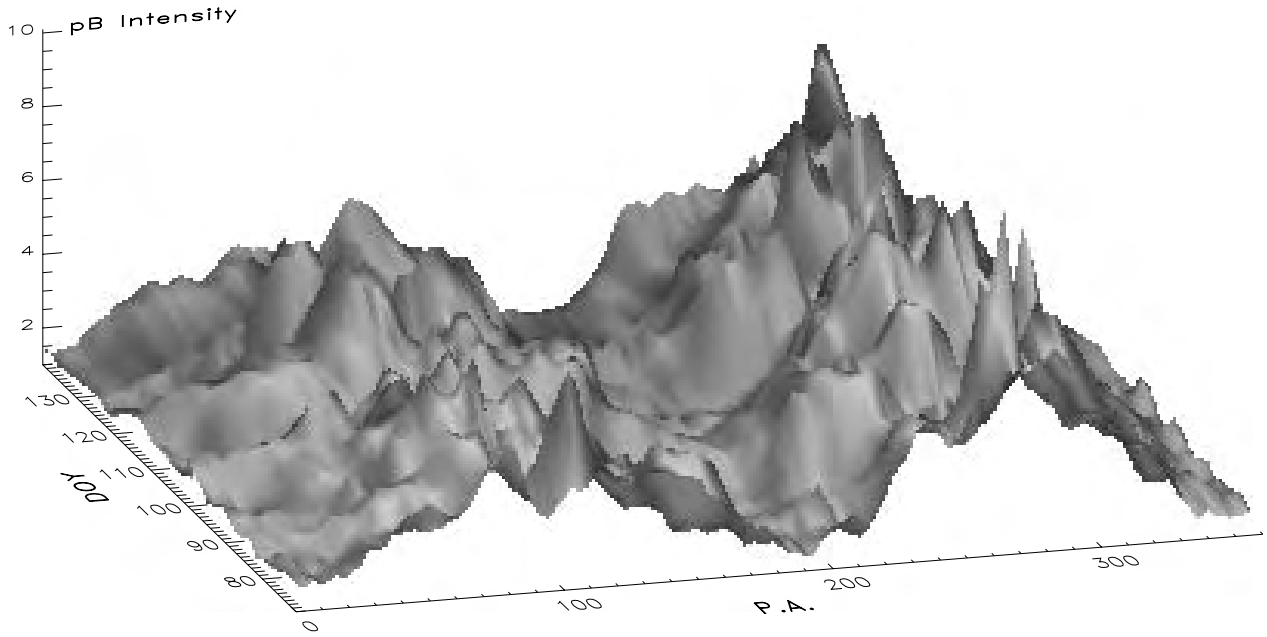


Figure 2. Polarized white light intensity measurements carried out every 3° along the circle shown in figure 1. These ground based measurements are made daily with the Mauna Loa K-coronameter, operated by the High Altitude Observatory (NCAR/HAO). The intensities are normalized to the lowest coronal hole intensity. Position angle 0 corresponds to heliographic north. Shown here are the measurements from 12 March to 19 May 1993. The low intensities from position angle 120° to 220° correspond to the southern coronal hole. Note the large daily intensity variations both inside the hole and in the surrounding regions.

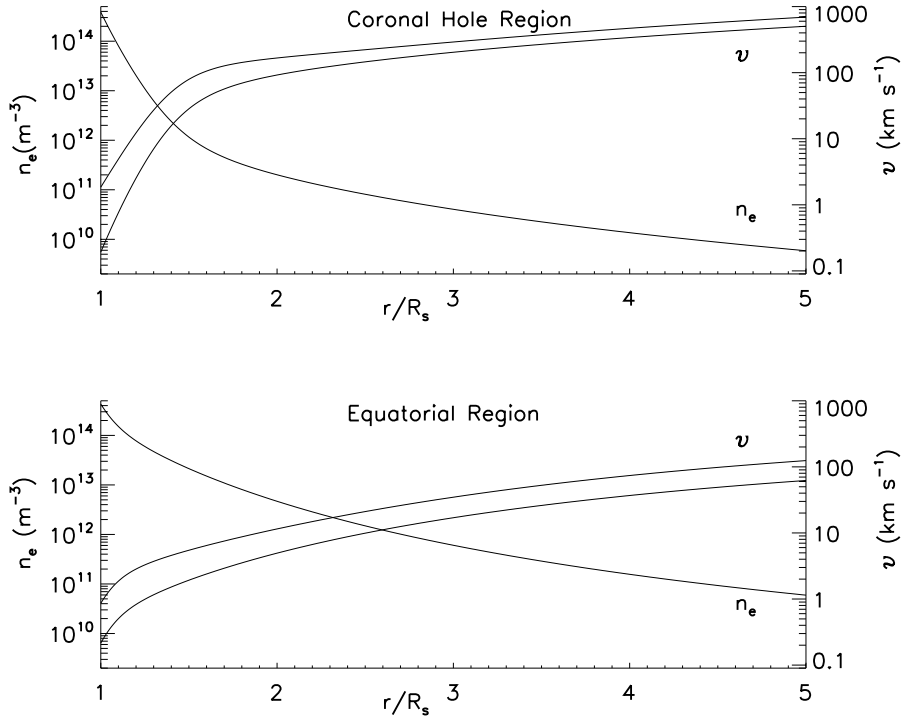


Figure 3. Electron densities derived from polarization brightness measurements in the inner corona (n_e) and estimates of the flow speed (v) derived from the mass flux measured *in situ*, these densities and the law of mass flux conservation for a radial expansion of the flow tubes (lower limits) and an expansion seven times more than radially.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

sound speed is not yet known, partly because we are not sure of the flow speeds in the inner corona, but also because we do not know the particle temperatures (sound speed) very accurately. Given the different properties of streams in interplanetary space one has to assume that their acceleration, as well as other characteristics in the inner corona, are different. The only means to determine these characteristics is by remote observations since no spacecraft has approached the Sun closer than about 60 solar radii. (A spacecraft to probe the solar wind plasma at distances of 5 to 10 solar radii is being planned.) The analysis of remote observations always requires a series of assumptions and approximations, and the results are therefore often a question of interpretation. For example, remote observations of the coronal plasma are always intergrated measurements along a line-of-sight. An example is given in figure 2. This figure shows Thompson scattered polarized white light intensity measured clockwise along the circle shown in figure 1, as a function of position angle, where 0 corresponds to the Sun's north pole. These measurements were carried out daily from day 75 to 135, 1993. The southern polar coronal hole extends roughly from position angle 120° to 220°. It can be seen that the intensity measured inside that region varies significantly both along the time axis and along the position angle, even though the plasma properties measured *in situ* in the high-speed solar wind are rather constant. The intensity changes seen when observing coronal holes remotely are mostly due to changes in the regions surrounding them. At times these surrounding denser regions can veil the coronal holes completely or partially, as in figure 1 at the southern pole. Parameters derived from remote observations represent the plasma properties averaged along the line-of-sight. Thus, particularly measurements carried out in coronal holes will be biased by surrounding regions, and the measurements can only serve as guidelines. Figure 3 shows the electron densities, n_e, derived for a coronal hole region (upper panel) and an equatorial region (lower panel) using polarized white light observations (e.g. Fisher and Guhathakurta 1995).

An estimate of the lower limit on the outflow speed of the electron–proton plasma close to the Sun can be obtained from the mass flux measured *in situ*; these electron densities and the law of mass conservation which states that the flux of protons must be conserved from the solar surface into interplanetary space where *in situ* measurements can be carried out. (The proton flux is defined as the number of protons per m^3 times the proton flow speed times the area expansion of the flow tubes, which is the ratio between the flow tube area on the solar surface, where the stream originates, and the area that this stream covers at a given distance from the Sun.)

The lower limits of the flow speed, v, shown in figure 3, are for the smallest possible flow tube expansion (radial), and the upper limits are for a very large expansion (seven times more than radially). The flow speeds derived from this estimate show that the solar wind accelerates very rapidly in the inner corona, and achieves its almost final flow speeds between 5 and 10 solar radii. This is in agreement with the *in situ* Helios observations which show that there is basically no acceleration of the flow at large distances from the sun. Observations from SOHO indicate that the heavy O^{5+} ions already flow faster than the protons in the inner corona. It seems that, at least in the high-speed wind, the differences between minor ions and protons, seen in the *in situ* measurements, originate in the inner corona with subsequent modification in interplanetary space.

Solar wind speeds of 200 km s^{-1} can be achieved by the pressure gradient force alone. To accelerate the wind to higher speeds, additional energy is needed that has to be deposited in the corona. At present it is not well understood where this additional energy comes from. It is usually assumed that plasma waves play a significant role (see SOLAR WIND: THEORY and CORONAL HEATING MECHANISMS).

Summary

The acceleration of the solar wind plasma to the speeds measured in interplanetary space happens close to the Sun, below 5 solar radii in the fastest streams and at slightly larger distances in the low-speed wind. During this acceleration process the solar wind passes from subsonic to supersonic speeds. The energy source necessary for the acceleration to happen is not yet known but is thought to be high-frequency plasma waves originating either in the lower solar atmosphere or higher up in the corona via a cascade from lower-frequency waves or via locally generated microinstabilities.

Bibliography

Fisher R and Guhathakurta M 1995 Physical properties of the polar coronal rays and holes as observed with SPARTAN 201-01 coronagraph *Astrophys. J.* **447** L139–L143

Schwenn R and Marsch E 1990 *Physics of the Inner Heliospehere* vol 1 (Berlin: Springer)

——1991 *Physics of the Inner Heliospehere* vol 2 (Berlin: Springer)

Ruth Esser

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind Composition

To first order the solar wind composition reflects the composition of the source material, which is photospheric (SOLAR ABUNDANCES). However, there are some important distinctions between the solar wind and solar abundances which will be discussed in the following paragraphs. Since solar wind particles feed the CORONA, and solar energetic particles (*SEPs*; see SOLAR WIND: ENERGETIC PARTICLES) largely originate in the *corona*, elemental and isotopic abundances of matter in the corona, in the solar wind and in solar energetic particles have much in common.

The first solar wind composition measurements were carried out soon after the experimental discovery of the solar wind in the early 1960s. Gradually improved techniques have revealed many details about the acceleration and heating processes operating in the solar atmosphere and corona. With recent improvements of resolution in mass spectrometers carried on the WIND, SOHO and ACE missions, it has also become possible to study abundances of the major isotopic species and hence to gain direct information about the isotopic composition of solar matter. This information is not accessible with optical methods because the natural line widths are much wider than isotopic shifts in spectra. Thus the interest in solar wind composition is twofold and the topics of interest are interrelated. Depending on the viewpoint and the application, studying the solar wind composition and its variability provides on the one hand knowledge about the ways solar matter propagates into the interplanetary space. On the other hand, it provides access to the nuclidic composition of solar matter, information which is relevant for a multitude of astrophysical applications since solar matter is the most important reference point for the isotopic evolution of the solar system. An example of a high-resolution solar wind mass spectrum is shown in figure 1.

An interesting minority within the solar wind population are the so-called 'pick-up ions' which mainly originate from interstellar gas flowing through the inner heliosphere. The elements helium and neon, having the highest first ionization potentials (FIPs), penetrate as neutrals to radial distances of less than 1 AU before they become ionized and can be picked up by the magnetic field which is traveling with the solar wind. Depending on their location of ionization, they relax more or less efficiently to local conditions in the plasma. However, they generally remain in low ionization states which leaves them distinguishable from the rest of the solar wind population. Freshly picked-up particles perform cycloidal motions about the outwards propagating magnetic field; hence their energy distributions reach beyond typical solar wind energies. They are thus particularly susceptible to acceleration in the fields associated with interplanetary shock fronts which are formed e.g. from coronal mass ejections (CMEs) or co-rotating interaction regions (CIRs). In addition to interstellar pick-up ions, other species with low FIPs have been found in the pick-up ion population. These particles originate from evaporating dust grains released from cometary debris and from interstellar dust particles which also can penetrate deep into the inner solar system.

Feeding of the corona with solar matter

Coronal structure and solar wind composition

The corona exhibits a complex spatial and temporally variable structure which becomes visible to the naked eye during solar eclipses. The coronal structure with many details and its dynamic evolution are essentially determined by the configuration of the surface SOLAR MAGNETIC FIELD. Coronal structure and solar wind composition are interrelated in many ways. For a rough characterization two terms are frequently used: '*interstream solar wind*' and '*coronal-hole-associated streams*'. Interstream solar wind is believed to be supplied from the fringes of the equatorial closed loop system. Solar wind particles travelling along open structures on the borderline of magnetic loops exhibit strong variations in composition. This observation has been taken as evidence for multiple sources rapidly connecting and disconnecting thereby producing expanding plasma blobs with complex spatial and magnetic topology which move outwards through the corona. On the other hand, *coronal-hole-associated fast streams* are generally much more stable in their composition.

The balance between the gravitational attraction of the Sun and the Coulomb friction force exerted from outwards travelling protons on minor species seems to play a crucial role in establishing solar wind abundances. In a corona at thermal equilibrium with a temperature of typically 10^6 K, only protons (together with electrons to maintain charge neutrality) reach the escape velocity and move into the interplanetary space. One possibility to visualize the interaction between coronal structures and solar wind composition, which has been extensively studied, uses coronal magnetic flux tubes as elements that guide the solar wind flow. Depending on the flux density or—more precisely—on the phase space density (m^{-6} s^{-3}) in a given flux tube, and depending on the divergence of the flux tube in the solar wind acceleration region, protons couple to minor ions via Coulomb collisions and tend to drag heavy species more or less efficiently according to the individual Coulomb drag factors of each species. An alternative to interplay between coronal structures and the solar wind plasma to populate the solar wind with minor ions involves magnetic fields as carriers of magnetohydrodynamic waves which interact with charged particles. This seems particularly important for coronal-hole-associated solar wind where Coulomb friction seems insufficient to drag minor species. Nevertheless, in this type of solar wind, minor ions are at least as abundant as in the interstream wind, and even helium—although still depleted relative to its solar surface abundance—is better represented in coronal-hole-associated wind than in the typical interstream wind. Hence, momentum added through waves to protons, to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

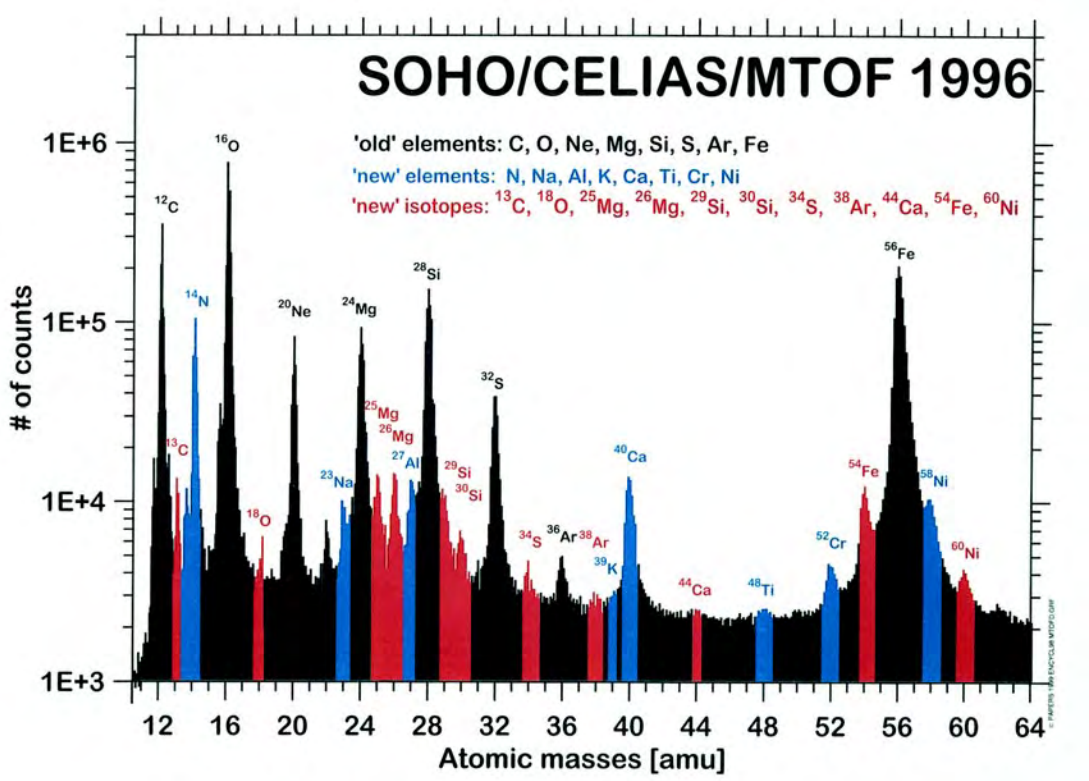


Figure 1. Solar wind elements and isotopes as observed with the MTOF sensor of the CELIAS experiment on SOHO. Elements and isotopes which have been identified for the first time in the solar wind thanks to isochronous mass spectrometry are shown in grey.

helium and to minor species seems crucial to obtain the typical properties of coronal-hole-associated solar wind. This has recently been confirmed with optical observations from the SOHO spacecraft, indicating that ion-cyclotron resonance wave–particle interaction brings minor ions to temperatures as much as a hundred million kelvins, thus exceeding the main gas temperature by one or more orders of magnitude.

Magnetic reconnection is another process to link the magnetic field structures to solar wind abundances. It occurs in the innermost corona and this link to abundances has not yet been extensively explored. Rapidly changing magnetic topologies in reconnection regions around the equatorial streamer belt, which supplies matter to the interstream solar wind, lead to dissipative heating, generation of EUV and soft x-ray radiation, and electric fields which could selectively accelerate particles according to their charge per mass ratio. The EUV radiation originating in the dissipation region and in shock fronts controls the ionization of matter in the underlying chromosphere and thus selectively regulates the replenishment of fresh ions in the corona.

Fractionation effects

Elements

As discussed in more detail elsewhere in this encyclopedia (see TRANSITION REGION: FIRST IONIZATION POTENTIAL EFFECT), the most important elemental fractionation process occurs in the ionization layer situated in the upper chromosphere where species with short ionization times (or low FIPs) are enriched over species with long ionization times typically by factors of 3–5 over photospheric abundances in the interstream solar wind. This effect seems much weaker in coronal-hole-associated wind. Further up in the atmosphere, it appears that wave–particle interaction, which plays a dominant role in the acceleration of coronal-hole-type solar wind, acts in a much more democratic manner than Coulomb drag—including also weakly ionized species—which would otherwise hopelessly remain behind. A summary of solar wind elemental abundances is given in table 1. One conspicuous feature of abundances, in both types of solar wind, is the depletion of helium by typically a factor of 2 relative to hydrogen. Since this feature of helium depletion is also found in the otherwise 'democratically elected' solar abundance representation in coronal holes, it is generally ascribed to the ion–neutral separation process which leaves its imprint on all types of solar wind.

Isotopes

Table 2 is a compilation of recent data on the isotopic composition of the solar wind. The refractory elements which have uniform isotopic composition throughout the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Solar wind elemental abundances.

				Abundance (logarithmic scale), [O] ≡ 8.78		Abundance (linear scale), [O] ≡ 1000	
Element	Z	FIP (eV)	FIT (s)	Interstream	Coronal hole	Interstream	Coronal hole
H	1	13.6	70	12.21 ± 0.13	11.89 ± 0.12	$2\,700\,000 \pm 600\,000$	$1\,280\,000 \pm 320\,000$
He	4	24.5	260	10.68 ± 0.10	10.52 ± 0.10	$80\,000 \pm 20\,000$	$55\,000 \pm 20\,000$
C	6	11.2	20	8.60 ± 0.07		660 ± 100	
O	8	13.6	74	≡8.78		≡1000	
Ne	10	21.5	81	8.01 ± 0.06		170 ± 20	
Mg	12	7.6	0.3	8.01 ± 0.10	7.69 ± 0.10	171 ± 35	81 ± 18
Si	14	8.1	0.6	8.08 ± 0.10	7.49 ± 0.08	200 ± 30	52 ± 7
S	16	10.3	11.6	7.55 ± 0.05	7.50 ± 0.06	59 ± 7	53 ± 6
Ar	18	15.7	50	6.35 ± 0.13		3.7 ± 1.0	
Fe	26	7.8	2	7.84 ± 0.10	7.58 ± 0.10	115 ± 15	63 ± 8
Kr	36	14.1	20.3	3.25		3.0×10^{-3}	
Xe	54	12.1	10.1	2.63		7.1×10^{-4}	

This is a compilation of results obtained with the Apollo/SWC experiment, from lunar soil investigations, from ISEE3/ICI, the SWICS Instrument and the Plasma Ion Spectrometer on Ulysses, and from CELIAS on SOHO (see Bochsler 2000 for more details and a complete list of references). [O] ≡ 8.78 is used as a normalized value for both types of solar wind. It reflects the photospheric abundance of oxygen in dex units as recommended by Holweger (private communication). The standard first ionization times (FITs) are taken from Geiss (1998) and references therein.

solar system and whose solar isotopic composition is presumably well known (e.g. Mg and Si) show no evidence of fractionation between the solar surface and the solar wind to the level of precision which has been achieved to date. In contrast, the volatile elements, i.e. helium, neon and nitrogen, show rather strong variations in the isotopic composition in different solar system samples, such as terrestrial and solar wind. In view of the absence of fractionation effects in the case of the refractories, it can be concluded that the solar wind is generally a faithful representation of the isotopic composition of the solar surface. For the case of the very light element helium with a large relative mass difference between the two stable isotopes, variations of the $^4He/^3He$ ratio of the order of 10% have to be taken into account under normal conditions. Hence, the observed differences between isotopic abundances of *solar wind* volatiles from planetary volatiles are differences which originate at the source; they reflect different histories of *solar* volatiles and volatiles in the rest of the solar system. Isotopic solar wind data represent important benchmarks for models of the early geochemical evolution of the solar system and also put rigid constraints for models describing the acceleration of heavy species in the inner corona.

Since all quantitative models on the FIP effect published heretofore ascribe the depletion of high-FIP elements in the corona essentially to atomic properties, it is no surprise that they predict virtually no isotopic fractionation concomitant with the ion–neutral separation. In view, however, of the sometimes strong variability of the He/H ratio it is possible than an overall isotopic fractionation of the order of 1% per mass unit results for medium-mass elements between the solar surface and the solar wind due to selective acceleration in the corona. This effect is at present, however, still below the detection limit.

Table 2. Isotopic abundances.

	Solar wind	Terrestrial
$^4He/^3He$	2350 ± 70	740 000
$^{14}N/^{15}N$	200 ± 55	272
$^{16}O/^{18}O$	450	498.8
$^{20}Ne/^{22}Ne$	13.7 ± 0.3	9.80
$^{22}Ne/^{21}Ne$	31 ± 4	34.5
$^{24}Mg/Mg$	0.7857 ± 0.0076	0.7899
$^{25}Mg/Mg$	0.1020 ± 0.0048	0.1000
$^{26}Mg/Mg$	0.1123 ± 0.0071	0.1101
$^{28}Si/Si$	0.9220 ± 0.0026	0.9223
$^{29}Si/Si$	0.0454 ± 0.0020	0.0467
$^{30}Si/Si$	0.0326 ± 0.0021	0.0310
$^{36}Ar/^{38}Ar$	5.58 ± 0.03	5.32
$^{40}Ca/^{42}Ca$	128 ± 47	149.8
$^{40}Ca/^{44}Ca$	50 ± 8	46.47
$^{54}Fe/Fe$	0.0604 ± 0.0053	0.0580
$^{56}Fe/Fe$	0.9148 ± 0.0053	0.9172
$^{78}Kr/^{86}Kr$	$0.019\,45 \pm 0.000\,15$	$0.019\,95 \pm 0.000\,08$
$^{80}Kr/^{86}Kr$	0.1274 ± 0.0006	0.1296 ± 0.0004
$^{82}Kr/^{86}Kr$	0.6573 ± 0.0020	0.6617 ± 0.0016
$^{83}Kr/^{86}Kr$	0.6586 ± 0.0017	0.6600 ± 0.0014
$^{84}Kr/^{86}Kr$	3.279 ± 0.007	3.273 ± 0.007
$^{124}Xe/^{130}Xe$	0.0290 ± 0.0007	$0.023\,35 \pm 0.0014$
$^{126}Xe/^{130}Xe$	0.0259 ± 0.0009	$0.021\,76 \pm 0.0014$
$^{128}Xe/^{130}Xe$	0.5038 ± 0.0028	0.4708 ± 0.0017
$^{129}Xe/^{130}Xe$	6.354 ± 0.0017	6.505 ± 0.0015
$^{131}Xe/^{130}Xe$	4.988 ± 0.0011	5.224 ± 0.0012
$^{132}Xe/^{130}Xe$	6.062 ± 0.0016	6.614 ± 0.0012
$^{134}Xe/^{130}Xe$	2.239 ± 0.0008	2.567 ± 0.0007
$^{136}Xe/^{130}Xe$	1.818 ± 0.0006	2.182 ± 0.0006

For details and a complete list of references see Bochsler (2000).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ionic composition

Solar wind particles (as well as SEPs), after passage through the coronal temperature maximum, are usually strongly ionized. Figure 2 is an illustration showing a mass versus mass/charge distribution as obtained with the SWICS instrument of ACE. Protons are suppressed in this instrument to avoid counting overflows; helium appears exclusively as $^4He^{2+}$ together with some traces of $^3He^{2+}$. Carbon and oxygen are almost fully ionized whereas the heavier species usually conserve their innermost electron shells, e.g. Fe is most conspicuous in the mass per charge range from 4.67 to 7.0 (Fe^{12+} through Fe^{8+}). The ionization balance of minor species is predominantly established via electron collisions in the corona at about a distance of $1R_\odot$ from the solar surface. The ionization rates, and to some degree also the recombination rates, depend on the local electron temperatures. In the dense, inner parts of the corona, collisions are sufficiently frequent to establish a quasi-equilibrium distribution of charge states. As the ions move further out, the ambient electron density decreases, collisional reactions are gradually supressed and the result is a relatively wide distribution of charge states which 'freezes' and remains unaltered throughout the heliosphere. Charge states of minor species measured far from the Sun provide a useful means to deduce the electron temperatures in the corona. Another application is reminiscent of hydrological methods: because of their conservative properties, charge states of minor ions are now frequently used as reliable tracers to map distorted solar wind flux tubes back to their source in the solar corona.

Coronal mass ejection related solar wind composition

Only recently has it become possible to reliably determine the rapidly varying composition of minor ions in CME-related solar wind. A long-known feature is the sometimes enhanced He/H ratio which has been taken as evidence for gravitational stratification in the undisturbed corona. Similar effects might be responsible for the sometimes observed mass-dependent fractionation. The most surprising recent discovery, however, is a temporary strong enrichment of ^{3}He, probably due to the same mechanism which produces the even more dramatic enrichments of ^{3}He in impulsive flare particles. With the improved time resolution of modern instrumentation, less surprisingly, extremely unequilibrated charge state distributions, e.g. large amounts of $O^{2+,3+,4+}$ together with otherwise common solar wind species O^{6+} and O^{7+}, have been observed with ACE and WIND.

Historical solar wind

Lunar soil and also the soil in the surface layers of asteroids are constantly exposed to the solar corpuscular irradiation. Particles at solar wind energies penetrate typically to depths of 0.1 μm into dust grains and are stored within these grains for extended periods depending on their diffusive behavior within the mineral and depending on the further history of the grains. Lunar soil undergoes continuous sputtering from solar and cosmic ray particles, combined with erosion from micrometeorite impacts. Because of the occasional covering of soil layers with freshly excavated material from larger impacts, it is possible that some solar particle records are buried within the asteroidal and lunar regolith for billions of years, thus storing an archive of most of the solar wind history in the near-Earth environment and the inner solar system. Some tracers such as ^{40}Ar, which is a decay product from ^{40}K, and volatile fission products from decaying ^{235}U emanating from the lunar interior and re-implanted with the solar wind into surface material make it possible to reliably date the irradiation epochs. The lunar record has been intensively investigated. The results are somewhat controversial. One major problem is the allocation of suprathermal particles and SEPs which exhibit somewhat different isotopic abundances and seem overabundant with respect to the solar wind population. There is general agreement that the isotopic composition of the solar wind has not changed significantly during the last few billion years, except for the case of the ^{15}N/^{14}N abundance ratio which, however, must be attributed to extrasolar contributions. The apparent invariance of the ^{3}He/^{4}He abundance ratio in the solar wind over this extended period bears some significance for the history of the Sun, since it indicates that the outer part of the Sun (from $M_r \geq 0.5M_\odot$ to the base of the outer convective zone) has never undergone mixing since the ignition of the nuclear fuel.

The elemental abundances of the heavy noble gases indicate, however, that there has been a secular modification in solar wind composition over the past few billion years which is possibly related to a change in solar activity (i.e. that the ancient, in-ecliptic solar particle fluence contained significantly more CME-related, transient solar wind than the more recent record). This provides an additional boost to the study of CME-related composition which has only recently become possible owing to the considerably improved time resolution of modern particle instrumentation.

Bibliography

Bochsler P 2000 Charge states and abundances of particles in the solar wind *Space Sci. Rev.* (in preparation)

Geiss J 1998 Constraints on the FIP mechanisms from solar wind abundance data *Solar Composition and its Evolution—from Core to Corona* ed C Fröhlich, M C E Huber, S K Solanki and R von Steiger (Dordrecht: Kluwer) pp 241–52

Von Steiger R, Geiss and Gloeckler G 1997 Composition of the solar wind *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: University of Arizona Press) pp 581–616

Wieler R, Kehm K, Meshik A and Hohenberg Ch 1996 Secular changes in the xenon and krypton abundances in the solar wind recorded in single lunar grains *Nature* **384** 46–9

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

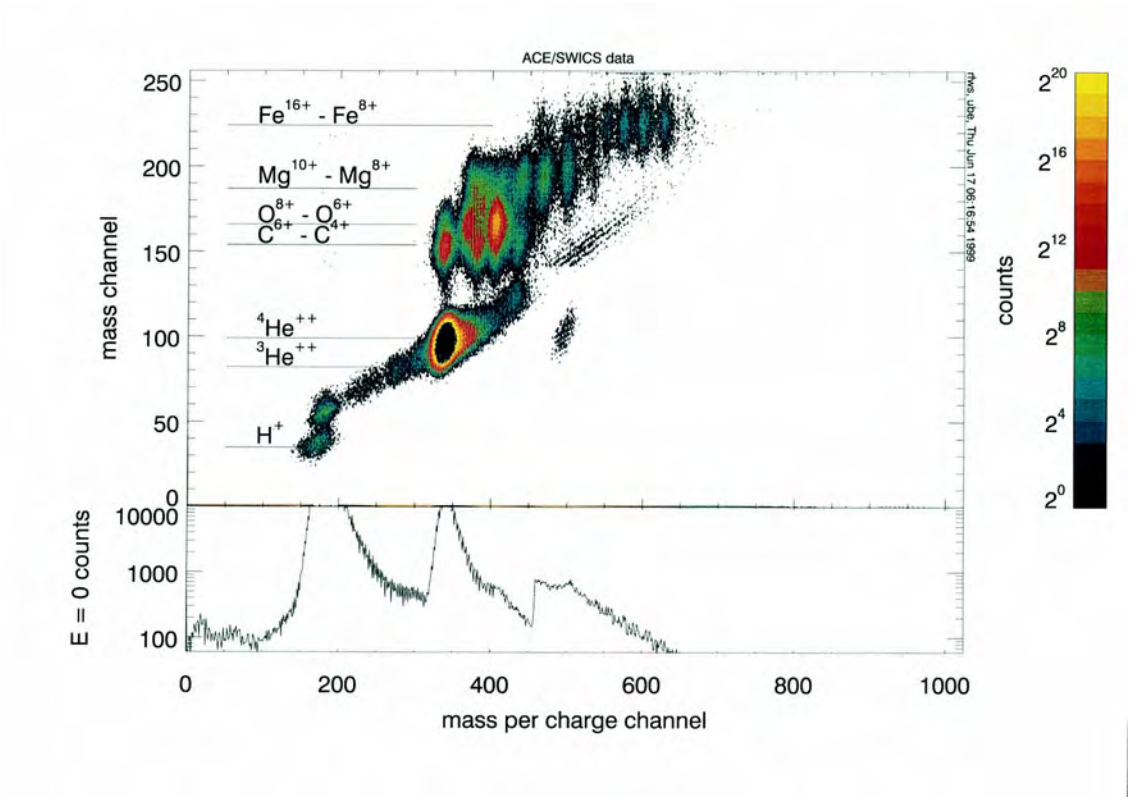


Figure 2. Mass versus mass per charge distribution in the solar wind as observed on day 29, 1998, with ACE/SWICS. **This figure is reproduced as Color Plate 56.**

Peter Bochsler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind Plasma Waves

Numerous types of waves exist in the SOLAR WIND. Since the solar wind is a plasma (i.e. an ionized gas) these waves are usually called plasma waves. Plasma waves play a fundamental role in determining the physical properties of the solar wind. In an ordinary gas, the gas is maintained in a state of thermal equilibrium by collisions. In thermal equilibrium the particle velocities have a universal isotropic form called a Maxwellian velocity distribution (see Haliday and Resnick 1981). Since the solar wind is very tenuous, collisions are extremely rare. An energetic particle from the Sun can reach the Earth without suffering a single large-angle collision. Under these conditions waves play a role similar to collisions. As the solar wind streams outward from the Sun, various dynamical effects cause the particle velocity distributions to deviate from a Maxwellian, often in the form of beams aligned along the magnetic field or various types of anisotropies. These non-equilibrium features eventually cause the growth of waves via various types of plasma instabilities. As the waves grow to large amplitudes, the waves interact with the particles in such a way as to eliminate the non-equilibrium feature that is responsible for the wave growth, thereby driving the velocity distributions toward thermal equilibrium, i.e. a role very similar to collisions. Waves also carry energy and momentum over long distances, and are believed to play a fundamental role in heating the solar CORONA and accelerating the solar wind. In this article we will describe the various types of waves that occur in the solar wind and discuss the role these waves play in determining the properties of the solar wind (see also SOLAR WIND: GLOBAL PROPERTIES; KINETIC PROPERTIES).

Plasma wave modes

Many types of waves can exist in a plasma. Figure 1 shows a plot of the propagation velocity, v, as a function of frequency, f, for the most commonly observed waves in the solar wind. This plot is representative of conditions at a radial distance of 1 AU (astronomical unit), near the orbit of the Earth. The propagation velocity is determined by a number of parameters, including the speed of light, c, the magnetic field strength, B, the particle number densities, N_s, of the various species in the plasma (s = e for electrons, and s = i for ions) and the temperatures, T_s, of these species. It is convenient to convert these parameters into a series of characteristic frequencies and speeds. The most important characteristic frequencies of a plasma are the electron cyclotron frequency, $f_{ce} = (1/2\pi)eB/m_e$, the ion cyclotron frequency, $f_{ci} = (m_e/m_i)\,f_{ce}$, the electron plasma frequency, $f_{pe} = (1/2\pi)(N_e e^2/\epsilon_0 m_e)^{1/2}$, and the ion plasma frequency, $f_{pi}(m_e/m_i)^{1/2} f_{pe}$. In these formulas e is the electronic charge, ϵ_0 is the permittivity of free space, m_e is the electron mass and m_i is the ion mass. The most important characteristic speeds of a plasma are the Alfvén speed, $V_A = B/(\mu_0\rho_m)^{1/2}$, the electron sound speed $V_e = (\gamma kT_e/m_e)^{1/2}$ and the ion sound speed $V_s = (\gamma kT_e/m_i)^{1/2}$, where μ_0 is the permeability of free space, ρ_m is the mass density and γ is the adiabatic compression factor. For a further discussion of these various quantities, see Stix (1992).

As can be seen in figure 1 the propagation velocity breaks up into a number of clearly defined branches, each of which corresponds to a well-defined mode of propagation. These modes are given specific names. Starting at the lowest frequencies, there are three modes of propagation called the magnetohydrodynamic (MHD) modes. These modes can be ordered according to the propagation velocity and are often called the fast magnetosonic mode, the Alfvén mode and the slow magnetosonic mode. The propagation velocities are given by

$$v_f^2 = \tfrac{1}{2}(V_A^2 + V_S^2) + \tfrac{1}{2}[(V_A^2 - V_S^2)^2 + 4V_A^2 V_S^2 \sin^2\theta]^{1/2} \quad (1)$$

$$v_A^2 = V_A^2 \cos^2\theta \quad (2)$$

$$v_s^2 = \tfrac{1}{2}(V_A^2 + V_S^2) - \tfrac{1}{2}[(V_A^2 - V_S^2)^2 + 4V_A^2 V_S^2 \sin^2\theta]^{1/2} \quad (3)$$

where v_f, v_A and v_s are the propagation velocities of the fast, Alfvén and slow modes. The angle θ, called the wave normal angle, is the angle between the direction of propagation and the magnetic field. As the name MHD implies, these modes involve a mixture of both magnetic and hydrodynamic (fluid) effects. The magnetic effects are represented by the Alfvén speed, V_A, which is proportional to the magnetic field strength, and the hydrodynamic effects are represented by the sound speed, V_s, which is proportional to the square root of the temperature. All three modes have electric and magnetic fields of the type that are usually associated with an electromagnetic wave. The fast and slow magnetosonic modes both have compressional motions of the type normally associated with sound waves (hence the dependence on the sound speed, V_S) and transverse motions of the type usually associated with electromagnetic waves (hence the dependence on V_A). The Alfvén mode on the other hand is a purely transverse mode and has no compressional motions (hence no dependence on the sound speed, V_S). The Alfvén mode was first discovered by Alfvén (1942) and is often called the shear Alfvén mode. For a discussion of MHD WAVES, see also Alfvén and Falthammer (1963). Under normal conditions the propagation velocities of all three MHD modes are substantially less than the solar wind velocity, V_{SW}, which is usually about 400 km s^{-1} (see figure 1).

Proceeding upward in frequency, one can see from figure 1 that as the frequency approaches the ion cyclotron frequency, f_{ci}, the propagation velocities of the fast magnetosonic mode and the Alfvén mode begin to deviate from the constant values given by equations (1) and (2). As the frequency approaches the ion cyclotron frequency the propagation velocity of the Alfvén mode begins to decrease rapidly, and it goes to zero at a frequency slightly below the ion cyclotron frequency. The exact frequency of the propagation cutoff depends on the wave

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

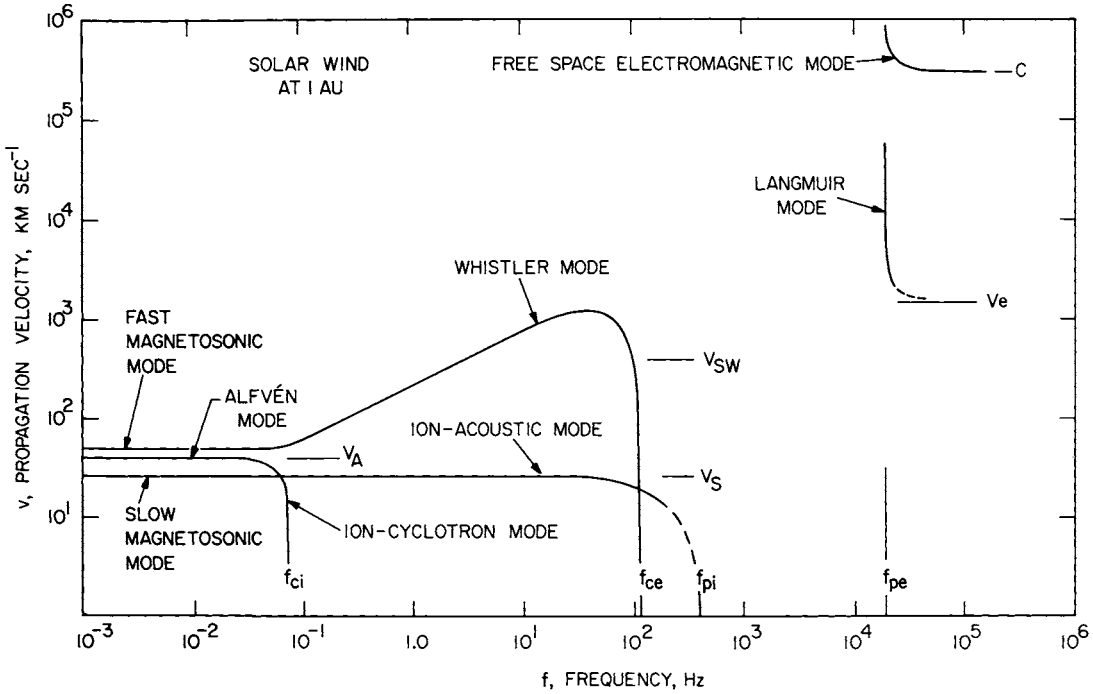


Figure 1. The propagation velocity of various plasma wave modes in the solar wind as a function of frequency for typical conditions at 1 AU.

normal angle. In this frequency range, near the ion cyclotron frequency, the Alfvén mode is usually called the ion cyclotron mode. Above the ion cyclotron frequency the Alfvén mode cannot propagate at any wave normal angle. For the fast magnetosonic mode, the propagation speed increases as the frequency crosses the ion cyclotron frequency, eventually varying as the square root of the frequency. This increase continues until the propagation velocity reaches a peak and then decreases rapidly, going to zero at a frequency slightly below the electron cyclotron frequency, f_{ce}. The exact frequency of this propagation cutoff depends on the wave normal angle. Above the electron cyclotron frequency the whistler mode cannot propagate at any wave normal angle. In the frequency range between the ion cyclotron frequency and the electron cyclotron frequency, the fast magnetosonic mode is usually called the whistler mode, after whistling signals generated by lightning that were discovered by Storey (1953). As the frequency of the slow magnetosonic mode increases the propagation velocity remains essentially unchanged as the frequency crosses the ion cyclotron frequency. Although the slow magnetosonic mode is an electromagnetic wave, at high frequencies, above the ion cyclotron frequency, the wave magnetic field becomes very weak and the wave takes on the properties of an electrostatic wave, with an electric field but no magnetic field. In this high-frequency regime the slow magnetosonic wave has many of the properties of an acoustic wave and is often called an ion acoustic wave. As the frequency approaches the ion plasma frequency, the propagation velocity of the ion acoustic mode goes to zero at the ion plasma frequency. The ion acoustic mode cannot propagate at frequencies above the ion plasma frequency. The ion acoustic mode is strongly damped by a collisionless damping process called Landau damping (Landau 1946). This damping arises because the propagation velocity is comparable with the ion thermal speed, which allows some of the wave energy to be absorbed by resonant interactions with ions moving at approximately the same velocity as the wave. The damping becomes particularly strong when the frequency is near the ion plasma frequency. This region of strong damping is shown by a dashed curve in figure 1. The damping is weak only if the electron temperature is substantially greater than the ion temperature, $T_e \gg T_i$, which raises the propagation speed of the wave well above the ion thermal speed (note that the ion sound speed, V_S, depends on the electron temperature and not on the ion temperature). The condition $T_e \gg T_i$ is only rarely satisfied in the solar wind.

At frequencies above the ion plasma frequency, no waves can propagate until the frequency reaches the electron plasma frequency, f_{pe}. At frequencies near and above the electron plasma frequency two modes of propagation occur, the free space electromagnetic mode and the Langmuir mode. The free space electromagnetic mode is a purely transverse wave with the electric and magnetic fields oriented perpendicular to each other and to the direction of propagation. The free space electromagnetic mode propagates at a speed given by $v = c/[1 - (f_{pe}/f)^2]^{1/2}$. Note that the propagation velocity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

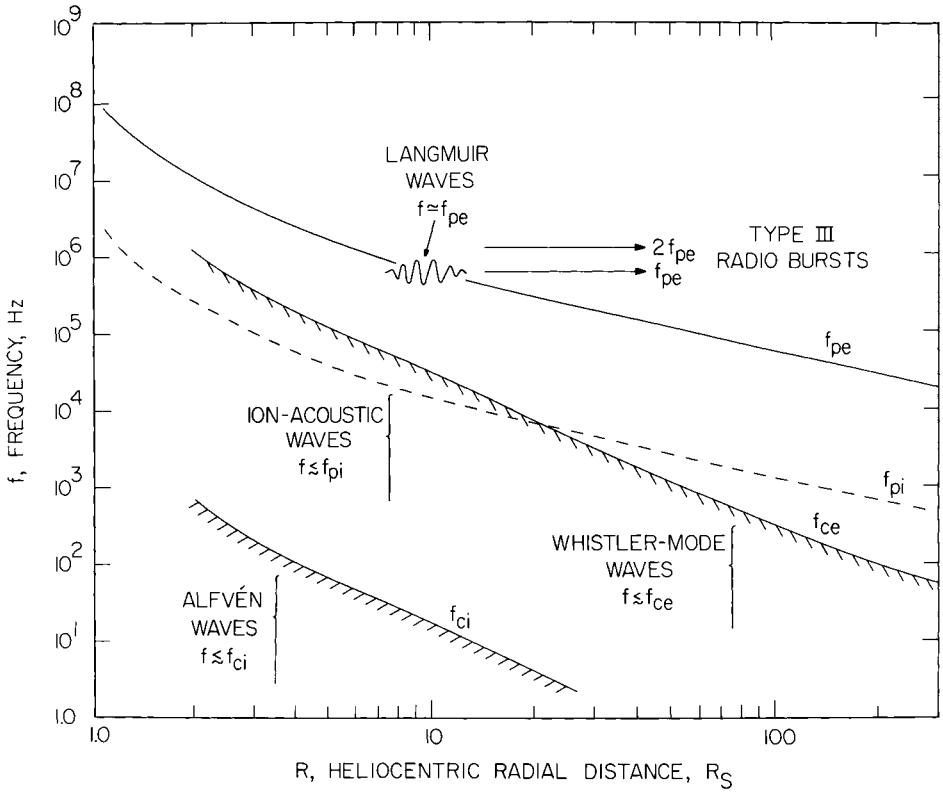


Figure 2. A plot of various characteristic frequencies of the solar wind plasma as a function of the heliocentric radial distance from the Sun in solar radii, R_S. The Earth is at a heliocentric radial distance (1 AU) of about $200R_S$.

of the free space electromagnetic mode goes to infinity at the electron plasma frequency and approaches the speed of light, c, at high frequencies. Since the propagation velocity is always greater than the velocity of any of the particles in the plasma, this mode cannot interact resonantly with the plasma particles and, therefore, does not experience Landau damping. The Langmuir mode is a purely longitudinal electrostatic wave (i.e. it has an electric field, but no magnetic field) and propagates at a speed given by $v = V_e/[1 - (f_{pe}/f)^2]^{1/2}$. Note that the propagation velocity of the Langmuir mode goes to infinity at the electron plasma frequency, and approaches the electron sound speed, V_e, at high frequencies. Since the characteristic propagation speed of this wave is near the electron thermal speed this wave is subject to strong Landau damping. The region of strong damping is indicated by a dashed curve in figure 1. The damping is small only for wave frequencies very close to the electron plasma frequency, $f \approx f_{pe}$, where the propagation speed goes to infinity. Thus, the Langmuir wave propagates with low damping only in a narrow frequency band around the electron plasma frequency, essentially a pure oscillation at the electron plasma frequency. The wave is also sometimes called an electron plasma oscillation.

To develop a complete understanding of the propagation of plasma waves in the solar wind we must consider the radial variation of the various characteristic frequencies of the plasma. Figure 2 shows a plot of the ion cyclotron frequency, the electron cyclotron frequency, the ion plasma frequency and the electron plasma frequency as a function of heliocentric radial distance. As one can see, all of the characteristic frequencies decrease with increasing radial distance from the Sun. To understand the detailed radial dependence we must discuss the radial variation of the magnetic field strength and the electron density. In the radial distance range from about $2R_S$ to $100R_S$, the magnetic field is directed almost radially outward from the Sun (Hundhausen 1972). Under these conditions conservation of magnetic flux shows that the magnetic field strength must vary approximately as $1/R^2$. Therefore, the cyclotron frequencies, f_{ci} and f_{ce}, vary as approximately as $1/R^2$ over this radial distance range. Because of strong multipole magnetic fields near the Sun, inside of about $2R_S$ it is not possible to give a specific prediction for the radial variation, except to say that it is likely to be steeper than $1/R^2$. Beyond about $100R_S$ the magnetic field develops a significant azimuthal component due to solar rotation, eventually varying as $1/R$ at large distances from the Sun. At radial distances beyond about $7R_S$ (Hundhausen 1972), where the solar wind is supersonic, conservation of particle flux shows that the electron density must vary as $1/R^2$. The ion and electron plasma frequencies, f_{pi} and f_{pe}, which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

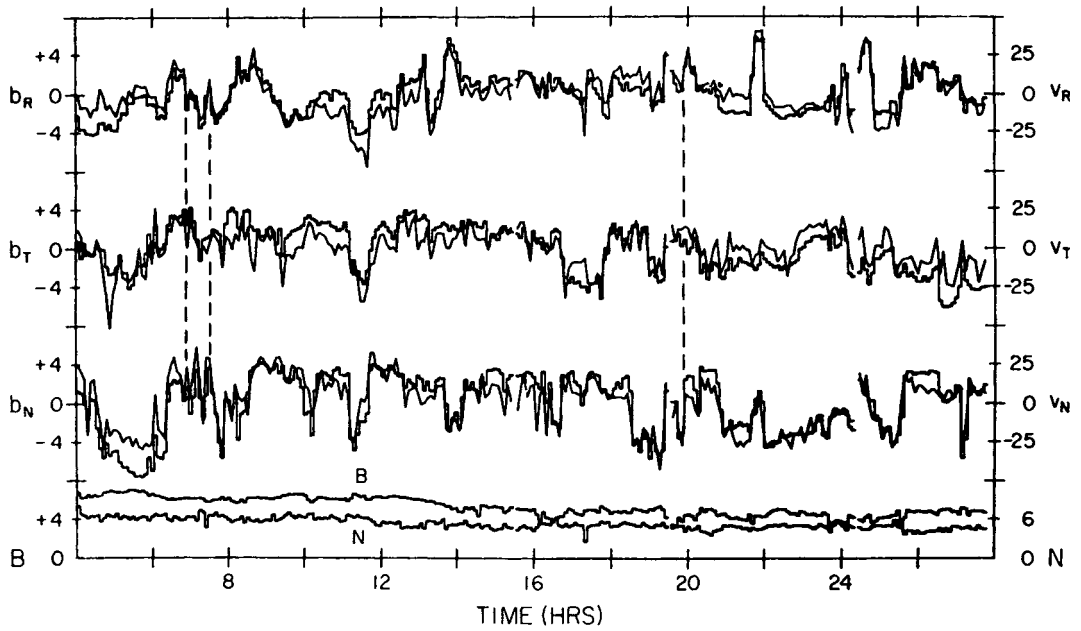


Figure 3. Alfvén waves in the solar wind as detected by the magnetic field and plasma instruments on the Mariner 5 spacecraft. From Davis (1972) with kind permission from Kluwer Academic Publishers.

are proportional to the square root of the electron density, must then vary as $1/R$ in this region. Inside of about $7R_S$, where the solar wind flow is subsonic, it is not possible to give a simple prediction for the radial variation of the plasma frequency, except to say that it is likely to be steeper than $1/R$.

Observations

To illustrate the types of plasma waves that are observed in the solar wind, we next show examples of each of the wave modes discussed in the previous section. To be consistent with the previous discussion, we start with the lowest frequencies and proceed to the highest frequencies.

Alfvén waves

Alfvén waves were first discovered in the solar wind by Unti and Neugebauer (1968) and Belcher *et al* (1969) using data from the Mariner 2 and 5 spacecraft. These waves are illustrated in figure 3, which is from Davis (1972). The top three plots shows three orthogonal components of the wave magnetic field, b_R, b_T and b_N. These components were obtained by subtracting the average background magnetic field from the measured field. Superposed on the same plots are the corresponding components of the plasma flow velocity fluctuation, u_R, u_T and u_N. The bottom plot shows the magnitude of the magnetic field, B, and the plasma density, N. The highly irregular variations in the magnetic field and flow velocity, on time scales ranging from a few hours to a few minutes, are Alfvén waves. Note that there is no evidence of corresponding variations in either the magnetic field magnitude or the plasma density. The absence of fluctuations in the magnetic field magnitude and in the plasma density indicates that the waves are non-compressional, thereby uniquely identifying the mode of propagation as the Alfvén mode. Note that the flow velocity variations are in phase (positively correlated) with variations in the corresponding magnetic field components. From the direction of the average background magnetic field during this event (directed inward, toward the Sun) and the positive correlation between the wave magnetic field and the flow velocity fluctuation, one can show that the waves are propagating outward, away from the Sun.

It is now widely believed that the Alfvén waves observed in the solar wind are excited by turbulent fluctuations near the Sun and that the waves propagate outward to great distances from the Sun, probably several AU or more. Although fast and slow magnetosonic waves are sometimes detected, the dominant wave energy is almost always in the shear Alfvén mode. The small amounts of energy in the fast and slow magnetosonic modes are most likely due to the fact that these modes are compressional, which leads to strong damping. In contrast, the Alfvén mode has very little damping, primarily because of the transverse (i.e. non-compressional) nature of this mode.

Ion cyclotron and whistler mode waves

As the Alfvén wave propagates outward from the Sun the ion cyclotron frequency decreases, eventually approaching the frequency of the wave (see figure 2). As the ion cyclotron frequency approaches the wave frequency the propagation velocity begins to decrease (see figure 2), and the wave becomes strongly damped

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by a process called ion cyclotron damping. Ion cyclotron damping occurs when the Doppler-shifted wave frequency seen in the frame of reference of an ion moving along the magnetic field matches the cyclotron frequency of the ion. The resulting interaction accelerates the ion and causes a loss of wave energy. The net effect is that the wave energy is absorbed as the ion cyclotron frequency approaches the wave frequency, thereby transferring the wave energy to the solar wind. This cyclotron resonant absorption process is believed to provide a significant heat input into the solar wind.

Evidence of this ion cyclotron absorption process is illustrated in figure 4, which shows a magnetic field spectrum observed by the Helios 2 spacecraft at a radial distance of about 0.3 AU. At the lowest frequencies, below about 10^{-2} Hz, the spectrum is mainly due to Alfvén waves. The intensity can be seen to vary as approximately f^{-1}. Proceeding upward in frequency a noticeable change in the spectral slope, from about f^{-1} to $f^{-1.7}$, can be seen as the frequency approaches the ion cyclotron frequency, f_{ci}, followed by a very abrupt decrease in intensity above the ion cyclotron frequency. The change in slope and the rapid decrease in the intensity near the ion cyclotron frequency are believed to be due to absorption of the Alfvén wave energy by either ion cyclotron damping or possibly Landau damping (Leamon *et al* 1998). Proceeding to higher frequencies, above the ion cyclotron frequency, a weak but persistent level of magnetic noise can be seen at frequencies extending up to about the electron cyclotron frequency. These frequencies are much too high to be caused by MHD waves and are almost certainly due to whistler mode waves (see figure 1). An alternative possibility, that they might be ion acoustic waves, is ruled out by the fact that, at these frequencies, the ion acoustic wave is almost purely electrostatic and would have no magnetic field. The origin of these whistler mode waves is poorly understood. Since whistler mode waves experience a significant level of cyclotron damping due to interactions with the relatively hot electrons ($T_e \simeq 10^5$ K) present in the solar wind, it seems unlikely that the waves could reach the spacecraft from a source near the Sun. The most likely explanation is that they are generated locally by anisotropies in the solar wind electron distribution. Anisotropic velocity distributions are known to be a free energy source for whistler mode waves (Kennel and Petschek 1966).

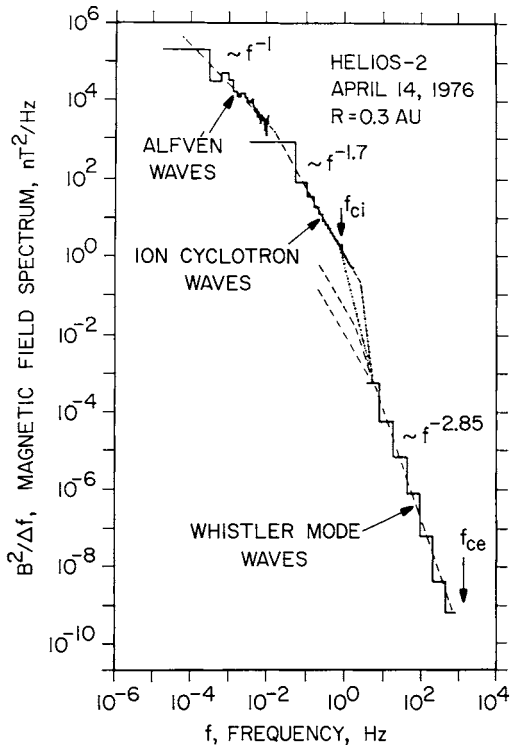


Figure 4. A magnetic field spectrum showing Alfvén waves, ion cyclotron waves and whistler mode waves detected by the Helios 2 spacecraft at a heliocentric radial distance of 0.3 AU. Illustration adapted from Denskat *et al* (1983).

Ion acoustic waves

Since the ion acoustic mode is normally very heavily damped by Landau damping, one would not expect to detect ion acoustic waves in the solar wind. Nevertheless, electrostatic waves have been detected in the solar wind that are believed to be ion acoustic waves. A frequency–time spectrogram showing a series of electrostatic bursts extending up to about 6 kHz that are believed to be ion acoustic waves is shown in figure 5. These waves were detected by the plasma wave electric field antenna on the Voyager 2 spacecraft at a heliocentric radial distance of 1.66 AU. The ion plasma frequency at this time is estimated to be about 300 Hz. Since the ion acoustic mode cannot propagate at frequencies above the ion plasma frequency, one may wonder how the frequencies could extend upward to frequencies as high as 6 kHz. The reason is that ion acoustic waves have very short wavelengths, which cause very large Doppler shifts. Under the conditions present in the solar wind near 1 AU, the frequency is almost entirely determined by Doppler shift, so there is no simple relationship to the ion plasma frequency. The Doppler shift is approximately $\Delta f = V_{SW}/\lambda$, where λ is the wavelength. The minimum wavelength of an ion acoustic wave is $2\pi\lambda_D$, where λ_D is the Debye length (Stix 1992), so the maximum frequency is approximately $f_{Max} = V_{SW}/2\pi\lambda_D$. For typical conditions at 1.66 AU the Debye length is about 10 m, which for a solar wind velocity of 400 km s^{-1} gives a maximum frequency of about 6 kHz, in good agreement with the observed upper frequency limit of the emissions in figure 5. Since the Debye length varies inversely with the square root of the electron density, the maximum frequency is predicted to increase inversely with radial distance from the Sun, in agreement with observations (Gurnett *et al* 1979).

At present the mechanism by which ion acoustic waves are produced in the solar wind is poorly understood. The intensity of these waves has been shown

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

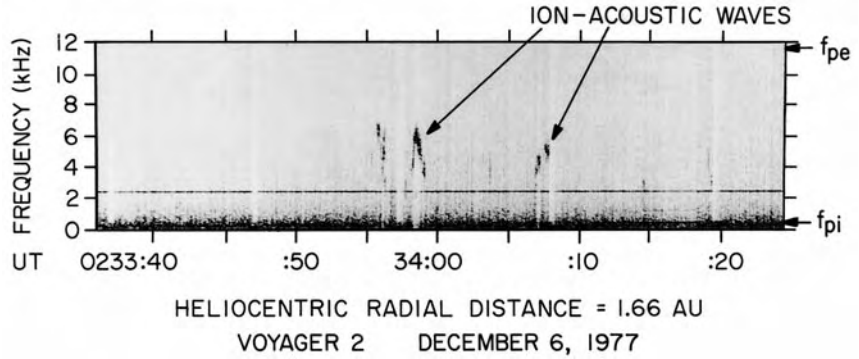


Figure 5. An electric field frequency–time spectrogram of ion acoustic waves detected in the solar wind by the Voyager 2 spacecraft at 1.66 AU.

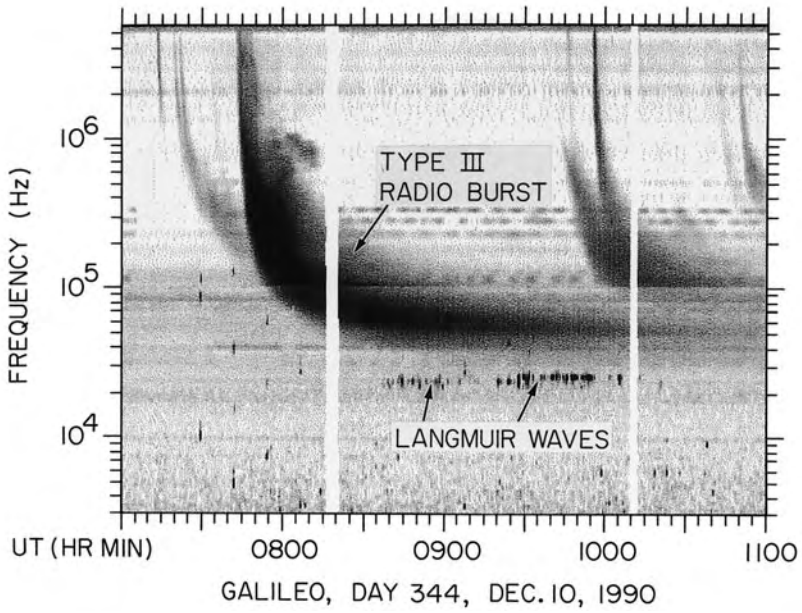


Figure 6. An electric field frequency–time spectrogram of a type III solar radio burst and associated Langmuir waves detected in the solar wind by the Galileo spacecraft at 0.98 AU.

by Gurnett *et al* (1979) to increase rapidly with decreasing distance from the Sun, which would seem to imply a solar origin. However, since the ion acoustic mode is strongly damped, it is highly unlikely that the waves could have propagated to the spacecraft from a source near the Sun. It is much more likely that they are produced locally. Both laboratory and theoretical studies show that the ion acoustic mode can be driven unstable by a shift between the average electron and ion velocities, such as occurs when a strong electrical current is present. The threshold current for the onset of the current-driven ion acoustic instability is very high when the electrons and ions have comparable temperatures, as it often is in the solar wind, but decreases rapidly as T_e/T_i increases. Although evidence exists that ion acoustic waves are more likely to be observed when T_e/T_i is large, electrical currents sufficiently strong to trigger the ion acoustic instability are usually not present in the solar wind. However, the heat flux carried outward from the Sun by solar wind electrons causes a similar shift between the core electron and ion velocities. Our present view is that the ion acoustic waves are probably driven by the solar wind heat flux. If so, ion acoustic waves could play a significant role in regulating the heat flux carried outward from the Sun. For a further discussion of ion acoustic waves in the solar wind, see Gurnett (1991).

Langmuir waves and type III solar radio bursts

For many years it has been known that a class of solar radio emissions called the type III radio bursts occurs in which the frequency drifts rapidly downward with increasing time (Wild 1950; see also SOLAR FLARES: RADIO

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

BURSTS). The time scale of the frequency drift varies from a few seconds at frequencies in the hundred MHz range to tens of minutes in the hundred kHz range. In a classic paper, Ginzburg and Zheleznyakov (1958) proposed that these radio bursts are produced by a two-step process in which (1) Langmuir waves are first excited by the energetic electrons emitted by a solar flare and (2) the Langmuir waves are converted to electromagnetic radiation at f_p and $2f_p$ by non-linear mode conversion processes. The downward frequency drift is caused by the decreasing electron plasma frequency encountered by the solar flare electrons as they stream outward from the Sun. The radiation at f_p is believed to be generated by a non-linear interaction between the Langmuir wave and a low-frequency wave such as an ion acoustic wave, and the radiation at $2f_p$ is believed to be generated by a non-linear interaction between two oppositely propagating Langmuir waves. The two-step generation process was confirmed by Gurnett and Anderson (1976) who made the first direct *in situ* observations of the Langmuir waves responsible for a type III radio burst using wave electric field measurements on the Helios 2 spacecraft.

A frequency–time spectrogram of a type III radio burst and its associated Langmuir wave emissions is shown in figure 6. This event was detected by wave electric field measurements on the Galileo spacecraft at a heliocentric radial distance of 0.98 AU. Although several type III radio bursts can be seen, the one of interest starts at 0745 UT (UNIVERSAL TIME). This event is associated with a solar flare that occurred at 0730 UT. As can be seen the type III radio burst associated with this flare was first detected at a frequency of about 6 MHz. The radio emission rapidly sweeps downward in frequency, eventually reaching a frequency of about 40 kHz after about 1 h. The Langmuir waves responsible for this radio emission can be seen from about 0835 to 1010 UT at a frequency of 23 kHz, which is the local electron plasma frequency. This onset time corresponds almost exactly to the arrival time of electrons with energies of about 100 keV from the solar flare, as determined from the energetic particle detector on Galileo; see Gurnett *et al* (1993). For this event the primary emission is believed to be at $2f_p$, which would produce radiation at a frequency of about 46 kHz. Note that the Langmuir waves responsible for radiation at frequencies higher (or lower) than 46 kHz cannot be detected, since these waves must occur closer to (or farther from) the Sun. Other studies have shown that both the intensity of the Langmuir waves and the emissivity of the type III radio events decrease rapidly with increasing radial distance from the Sun (Gurnett *et al* 1980), more or less in agreement with expectation for the $2f_p$ mode conversion process. However, the details of the mode conversion process, and the relative importance of emission at f_p and $2f_p$, remain poorly understood. For a discussion of the dynamics of Langmuir waves and the possible mode conversion processes that could be involved, see Robinson *et al* (1993).

Future research

In the sections above we have reviewed the basic characteristics of plasma waves in the solar wind. Although much is known about these waves at radial distances in the vicinity of the Earth's orbit, very little is known in the near vicinity of the Sun. The intensities of almost all the waves discussed increase rapidly with decreasing distance from Sun. Since many of the waves, such as whistler mode waves, ion acoustic waves and Langmuir waves, are locally generated, it is important that direct *in situ* measurements be made in the region close to the Sun, where the primary heating and acceleration of the solar wind are occurring. At present the closest measurements to the Sun are from the Helios 2 spacecraft, which approached to within 0.29 AU of the Sun. Hopefully, in the not-too-distant future a mission called the SOLAR PROBE will be sent much closer to the Sun, possibly to as close as $4R_S$, and will be instrumented to study plasma waves.

Acknowledgment

The research at the University of Iowa has been supported by NASA through contracts 959193 and 958779 with the Jet Propulsion Laboratory, Pasadena, CA.

Bibliography

Alfvén H 1942 Existence of electromagnetic–hydrodynamic waves *Nature* **150** 405–6

Alfvén H and Falthammer C G 1963 *Cosmical Electrodynamics, Fundamental Principles* 2nd edn (Oxford: Oxford University Press)

Belcher J W, Davis L Jr and Smith E J 1969 Large-amplitude Alfvén waves in the interplanetary medium: Mariner 5 *J. Geophys. Res.* **74** 2302–8

Davis L Jr 1972 The configuration of the interplanetary magnetic field *Solar Terrestrial Physics, 1970* ed E R Dyer, J G Roederer and A J Hundhausen (Dordrecht: Reidel) pp 32–48

Denskat K U, Beinroth H J and Neubauer F M 1983 Interplanetary magnetic field power spectra with frequencies from 2.4×10^{-5} Hz to 470 Hz from Helios observations during solar minimum conditions *J. Geophys. Res.* **54** 60–7

Ginzburg V L and Zheleznyakov V V 1958 On the possible mechanism of sporadic solar radio emission (radiation in an isotropic plasma) *Sov. Astron.–AJ* **2** 653–8

Gurnett D A 1991 Waves and instabilities *Physics of the Inner Heliosphere* ed R Schwenn and E Marsch (Berlin: Springer) pp 135–57

Gurnett D A, Anderson R R and Tokar R L 1980 Plasma oscillations and the emissivity of type II radio bursts *Radio Physics of the Sun* ed M R Kundu and T E Gergely (International Astronomical Union) pp 369–79

Gurnett D A and Anderson R R 1976 Electron plasma oscillations associated with type III radio bursts *Science* **194** 1159–62

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Gurnett D A, Hospodarsky G B and Kurth W S 1993 Fine structure of Langmuir waves produced by a solar electron event *J. Geophys. Res.* **98** 5631–7

Gurnett D A, Marsch E, Pilipp W, Schwenn R and Rosenbauer H 1979 Ion acoustic waves and related plasma observations in the solar wind *J. Geophys. Res.* **84** 2029–38

Haliday D and Resnick R 1981 *Fundamentals of Physics* (New York: Wiley) p 391

Hundhausen A J 1972 *Coronal Expansion and Solar Wind* ed J G Roederer (Berlin: Springer)

Kennel C F and Petschek H E 1966 Limit on stably trapped particle fluxes *J. Geophys. Res.* **71** 1–28

Landau L 1946 On the vibrations of the electronic plasma *J. Phys. USSR* **10** (1) 85–94

Leamon R J, Smith C W, Ness N F, Matthaeus W H and Wong H K 1998 Observational constraints on the dynamics of the interplanetary magnetic field dissipation range *J. Geophys. Res.* **103** 4775–87

Lin R P 1970 The emission and propagation of 40 keV solar flare electrons *Sol. Phys.* **12** 266

Robinson P A, Willies A J and Cairns I H 1993 Dynamics of Langmuir and ion-sound waves in type III solar radio sources *Astrophys. J.* **408** 720–34

Stix T 1992 *Waves in Plasmas* (New York: American Institute of Physics)

Storey L R O 1953 An investigation of whistling atmospherics *Phil. Trans. R. Soc.* A **246** 113–41

Unti T W J and Neugebauer M 1968 Alfvén waves in the solar wind *Phys. Fluids* **11** 563–8

Wild J P 1950 Observations of the spectrum of high-intensity solar radiation at metre wavelengths, III. Isolated bursts *Aust. J. Sci. Res.* A **3** 541

Donald A Gurnett

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind Shock Waves and Discontinuities

Shock waves in a fluid are observed as abrupt changes in all physical parameters of the fluid, such as flow speed, density, pressure and magnetic field, but, most important, in temperature, which ensures that entropy increases from the upstream to downstream side of the shock ramp. For an ordinary shock wave to exist in a fluid, its speed through the fluid must be faster than the speed of sound in the fluid. However, when a magnetic field is present, other characteristic signal speeds, which depend on the magnetic field's strength and direction, are also important in characterizing the shock. Shock waves in the solar wind, which is an ionized and magnetized fluid or PLASMA, are often referred to as interplanetary shocks. These rich events are of interest in themselves, for example, in studying steepened nonlinear waves in a so-called collisionless plasma, and as solar wind consequences of solar events, such as ejected solar mass or strong magnetic fields ramming into the upstream solar wind (see COLLISIONLESS PROCESSES IN ASTROPHYSICAL PLASMAS). They are also important as accelerators of energetic particles, as generators of radio waves and plasma waves and as triggers of geomagnetic phenomena.

In general, relatively abrupt changes in the intensity and/or direction of the magnetic field of the solar wind, i.e. in the interplanetary magnetic field (IMF), are called directional discontinuities, or simply discontinuities. In a sense shock ramps are a special class of discontinuities but will be treated separately here. Non-shock discontinuities are also important in the solar wind and are of interest in understanding its characteristics as a whole, especially in defining the boundaries of its various specific internal structures. The boundary of the Earth's magnetosphere is a directional discontinuity (see MAGNETOSPHERE OF EARTH).

The observations of interplanetary shocks and discontinuities have been the source of special interest to the space science community for over three decades, starting with their first observations by MARINER MISSION 2 and the early IMP spacecraft in the middle 1960s. In particular, it has been shown that generally these structures obey MAGNETOHYDRODYNAMIC (MHD) theory, but their very fine-scale features, which lie outside of the range of applicability for MHD theory, are not as well understood. The ability to make rapid measurements of magnetic field and plasma quantities by state of the art instruments on spacecraft provides an opportunity to investigate interplanetary shock waves and discontinuities on a fine scale for a broad variety of types of events and conditions. By contrast, widely separated multiple spacecraft allow study of the poorly known large-scale shape, evolution and coherence of these structures. The latter studies can be done even for distances encompassing the outer reaches of the solar system by observations from the Voyager 1 and 2 spacecraft, which are now (1999) farther than 70 AU from the Sun. We know, for example, that a shock wave may travel from the Sun past Earth's orbit for many 10s of AU. Hence, a very broad range of spatial scales are being investigated, i.e. from ~40 km to ~80 AU (12×10^9 km).

Shock waves in the solar wind

Definition and types of shocks

To qualify as a shock wave in a magnetized fluid, the changes in the fluid quantities (i.e. N, density; T, temperature; $\boldsymbol{B}$, magnetic field; $\boldsymbol{V}$, flow velocity) across its typically abrupt ramp must satisfy special equations of MHD, called the Rankine–Hugoniot equations, which are expressions for the conservation of energy, mass flux, and momentum flux, and Maxwell's equations of electricity and magnetism, including magnetic flux conservation. The shock parameters of general interest that help to characterize the shock include the following: upstream Mach number (which is proportional to the shock's speed relative to that of the upstream fluid and is also a measure of its 'strength'), mode or type (signifying its wave nature: fast, slow, or intermediate), obliquity (parallel, perpendicular, oblique, depending on the angle between the upstream magnetic field direction and the shock surface normal), sense of travel (forward or reverse, i.e. moving generally 'along with' or 'against' the solar wind flow direction, respectively), and sometimes shocks are separated according to their cause.

What separates the nature of solar wind shocks from those we experience on Earth (that are caused, say, by a fast-moving jet airplane) is the very low density of the solar wind and the existence of a magnetic field 'frozen into' the ionized solar wind's gas. The magnetic field's presence complicates the study of shocks in space. For example, for ordinary gasdynamic shocks in our atmosphere there is only one 'type', the sound wave mode, and obliquity has no meaning. The magnetic field, however, plays a crucial role in the solar wind by coupling its constituent particles, so that this otherwise low-density collisionless plasma acts as if it were a collisional ionized gas. Both ordinary gasdynamic shocks and solar wind shocks are characterized by Mach numbers. The ordinary one is defined as the ratio of the relative speed of the shock (i.e. with respect to the fluid's upstream flow speed) to the speed of sound, and is called the sonic Mach number. The relative speed is the speed of the fast 'piston' that is responsible for causing the shock, such as the speed of the fast jet plane. This Mach number must be greater than 1.0. Hence, high Mach number shocks are 'strong' ones that move fast. However, for a magnetized fluid several Mach numbers can be defined. These are characterized by the strength and direction of the magnetic field, as well as depending on the speed of sound as ordinary shocks do. An example of such an MHD Mach number is the Alfvén Mach number, which is the ratio of the shock speed (relative to the fluid) to the ALFVÉN SPEED, a characteristic signal speed in a plasma. The Alfvén speed depends directly on the magnetic field strength and inversely on the square root of the solar wind density. Other MHD wave speeds are called the slow- and fast-mode speeds

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and the related Mach numbers depend on combinations of the Alfvén and sound speeds.

The most common types of shocks in the solar wind at 1 AU are fast forward shocks. In the case of fast shocks generally both the upstream sonic and Alfvén Mach numbers must be greater than 1.0. Modern spacecraft observations are revealing many slow shocks, apparently because of the increased sensitivity of the advanced measuring instruments; for these the upstream Alfvén Mach numbers are less than 1.0, but the upstream sonic Mach numbers must be greater than 1.0. So-called intermediate shocks are very rare and, in fact, are believed by some not to be stable enough to be observed at all. However, a team of scientists in 1993 using Voyager data are believed to have identified such a shock. Of further interest is whether the shock is traveling in a forward or reverse sense with respect to the Sun in a solar wind frame of reference. Hence, four categories of shocks are generally of interest in solar wind studies: fast forward, fast reverse, slow forward and slow reverse, each having its unique signature. Figure 1 shows the sense of change of physical quantities for these four types of shocks. The Earth's bow shock is a fast reverse shock, for example, because the shock travels against the solar wind flow direction, and it is of a 'fast-mode wave' nature. Even though a bow shock of a planet occurs in the solar wind, and is therefore an interplanetary shock, it is not always classified as one, because it is viewed as a standing shock associated with a planet. All solar system planets that have been visited by spacecraft (i.e. all but Pluto) have been observed to have bow shocks. Figure 2 shows how the magnetic field ($\boldsymbol{B}$) and solar wind velocity ($\boldsymbol{V}$) change in strength and direction across the ramp for various kinds of interplanetary shocks, as denoted. A perpendicular shock signature (in N, T, B, V) appears to be very abrupt and squared-off like those shown in figure 1. However, a parallel shock, often referred to as a pulsation shock, has an oscillatory field nature, and the average change in B across it is ideally zero, but N, T and V all change across it.

Shock ramp thickness can vary considerably depending on the upstream conditions and on the shock type and obliquity; the abrupt vertical change of the physical quantities in figure 1 denotes the ramp. In actuality there is often a shock 'foot' before the ramp and an 'overshoot' region after the ramp whose specific signatures depend on the surrounding plasma's properties, specifically the values of N, T and B. A typical oblique shock ramp thickness is about 100 km–150 km at 1 AU from the Sun, i.e. at most about the distance between Baltimore, MD, and Philadelphia, PA. Very rapid sampling magnetometers in space allow the field measurements of such a thin ramp. Parallel shocks are much thicker.

Sources of shocks

Blast waves at the Sun, solar wind corotating stream interactions, caused by fast streams ramming into slow ones, rapid gas clouds (called driver gas or coronal mass ejections) from the Sun and/or fast-moving, strong, twisted magnetic field structures in the solar wind acting like huge driving magnetospheres (called magnetic clouds or magnetic flux ropes), plowing into the upstream solar wind, are all possible causes of interplanetary shock waves (see CORONAL MASS EJECTIONS: OBSERVATIONS, CORONAL MASS EJECTIONS: THEORY). Most of the interplanetary shocks seen in the solar wind near Earth, especially during the active part of the SOLAR CYCLE, arise near the Sun and travel to the Earth, usually driven by some fast-moving interplanetary structure. Bow shocks are a separate type again. These are standing shocks which arise from the solar wind ramming, at a fast enough speed, into a planet's outer environment, which is the magnetic field in Earth's case and the ionosphere in Venus' case (see MAGNETOSPHERE OF EARTH: BOW SHOCK, VENUS: INTERACTION WITH SOLAR WIND). The large-scale shapes of interplanetary shock surfaces are only approximately known, mainly from models, because there have not been many spacecraft mission opportunities to provide vastly separated and properly positioned observing points in the solar wind. However, some studies have shown that the radius of curvature of an interplanetary shock surface, at the Earth's distance from the Sun, is close to, but probably smaller than, 1 AU. Hence, these surfaces appear to move out almost uniformly from the Sun on a large scale. Undoubtedly on smaller scales these shocks have surface wrinkles, and other physical variations, that are caused by irregularities

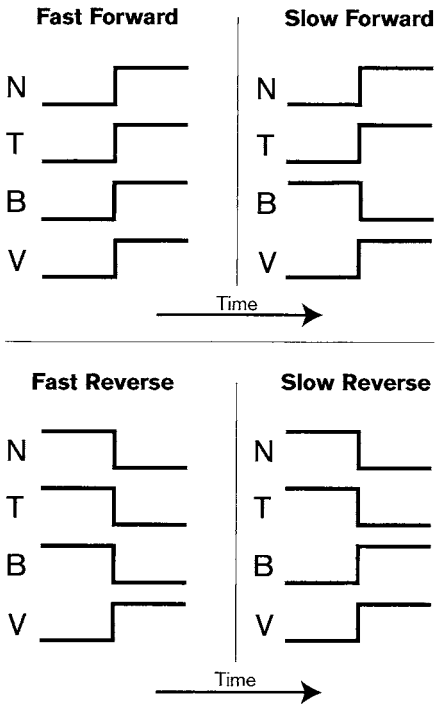


Figure 1. Profiles of solar wind density (N), temperature (T), magnetic field intensity (B) and flow speed (V) across an interplanetary shock ramp, as seen in a spacecraft frame of reference.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

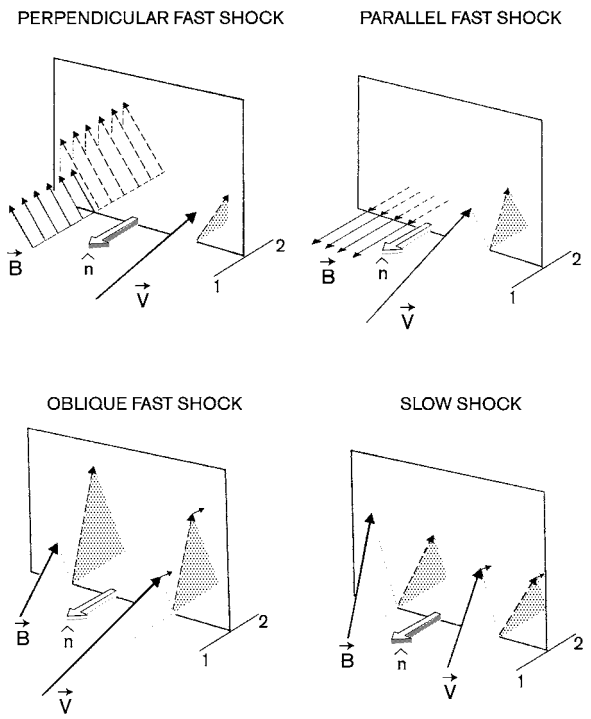


Figure 2. A vector representation of how the magnetic field ($\boldsymbol{B}$) and velocity ($\boldsymbol{V}$) change in strength and direction across a shock ramp for four various kinds of shocks, as denoted. The numbers '1' and '2' represent the up- and downstream states around the shock ramp, respectively, and $\boldsymbol{n}$ is the vector normal to the shock surface. When $\boldsymbol{n}$ is perpendicular (parallel) to the upstream IMF, the shock is referred to as a perpendicular (parallel) shock, otherwise it is called oblique.

in the solar wind as the shocks travel through it. Some of these variations are due to waves and discontinuities that the shock wave encounters, causing new discontinuous features to be generated, in turn.

Discontinuities in the solar wind

Definition and types of discontinuities

An interplanetary directional discontinuity is a relatively abrupt change in direction of the magnetic field 'carried' by the solar wind and sometimes accompanied by a change in magnitude (or intensity) of the field. Even though, by definition, a discontinuity requires a relatively rapid change in magnetic field, there is almost always a measurable thickness to the transition region in which the field's direction and other physical parameters change. This usually thin region is often called the 'discontinuity plane or sheet'. Within this sheet flows an electrical current composed of moving electrons, protons and to a lesser extent other positive ions. The state of the solar wind plasma on each side of the discontinuity plane must satisfy the fundamental MHD conservation conditions listed above for shock waves.

There are various kinds of directional discontinuities in the solar wind. They are usually classified according to the solar wind states on either side of its transition plane. In particular, there are tangential discontinuities in which ideally the component of the field perpendicular to the discontinuity plane, i.e. the 'normal component', is of zero magnitude, but the field's total magnitude does change from one side to the other. This kind of discontinuity is 'frozen into' the solar wind and travels along with it, because of its very high electrical conductivity. Such a discontinuity can have any size angle change of the field across its plane, including the special case of zero. A rotational discontinuity, on the other hand, under typical conditions, has no change in field intensity across the discontinuity plane but has a non-zero normal magnetic field component to that plane, and it propagates with respect to the solar wind. The magnitude of the normal component of the field is directly proportional to the propagation speed, which also depends on the local density of the solar wind, which is constant across the discontinuity. The propagation speed is the normal Alfvén speed. This kind of discontinuity is sometimes referred to as an Alfvén 'shock', but is not a real shock across which entropy always changes. Marked thermal anisotropy of the plasma on either side of a rotational discontinuity will cause a magnitude change of the field across it. A contact discontinuity is one in which there is a nonzero normal component of the field across the discontinuity surface (like a rotational discontinuity), is frozen into the solar wind and travels with it (like a tangential discontinuity), and has a constant field intensity across it. Both tangential and contact discontinuities have a density change across the discontinuity surface. Contact discontinuities are not as commonly observed in the solar wind as tangential or rotational discontinuities. Figure 3 provides a representation of how $\boldsymbol{B}$ and $\boldsymbol{V}$ change across various kinds of directional discontinuities.

The newly discovered Whang double discontinuity (1998) is a combination of an abrupt slow shock intimately accompanied by a much broader rotational discontinuity; such compound structures are not observed very often. Another structure in the solar wind, sometimes empirically referred to as a discontinuity, is a usually abrupt change in the magnetic field's magnitude, appearing as a sharp dip in intensity with rapid recovery reaching a minimum value only a small fraction of the intensities on either side; these are called magnetic holes and may or may not possess accompanying field directional changes.

Discontinuity thickness and rate of occurrence

The thickness of a tangential discontinuity plane in the solar wind, i.e. the thickness of its electrical current layer, depends primarily on the gyroradius (r_P) of the ions in the layer, mainly the protons. Hence, its thickness depends mainly on the ions' temperature and the ambient magnetic field. The temperature is important because it provides the ions' random (or thermal) speeds with which they revolve around the field lines penetrating from one side of the discontinuity plane to the other according to some

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

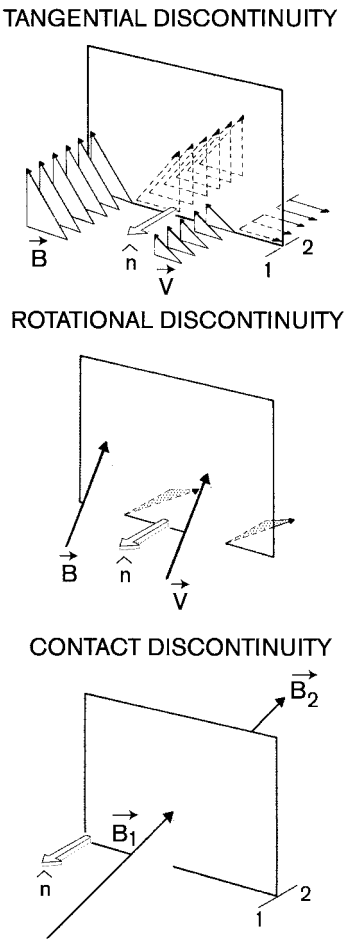


Figure 3. A vector representation of $\boldsymbol{B}$ and $\boldsymbol{V}$ changes across three different kinds of discontinuities, tangential, rotational, and contact, in the same format as figure 2.

measure of r_P. It has been shown that it usually takes many times a single gyroradius to account for the full thickness of a tangential discontinuity, about $30r_P$, which is not fully understood. Hence, on average, a discontinuity thickness at 1 AU from the Sun is about 3000 km. Data from the Mariner 10, HELIOS and VOYAGER spacecraft show that discontinuities generally become thinner as regions closer to the Sun are considered, because the IMF becomes stronger, causing r_P to become smaller. Observations of rotational discontinuity thicknesses are quite comparable with those of the tangential discontinuities on average. Hence, at 1 AU either kind of discontinuity is much thicker, and usually by more than an order of magnitude, than almost all observed interplanetary shock ramps. The estimate of the rate of occurrence of discontinuities in the solar wind depends on how strictly they are defined, i.e. on how large an angle is required across the discontinuity in some unit of time. Generally, however, the rate decreases with distance from the Sun.

Examples of discontinuities

The very large-scale structure in the solar wind called the heliospheric current sheet which separates the solar magnetic sectors is observed locally by spacecraft to be either a rotational or tangential discontinuity. Similarly, large magnetic flux rope structures from the Sun, apparently caused by erupting SOLAR PROMINENCES, called magnetic clouds, as seen at 1 AU very often have a sharp front boundary that is a directional discontinuity. However, numerous discontinuities exist all throughout the solar wind that are not associated with any other obvious solar wind structures.

Effects of shocks and discontinuities on Earth's magnetosphere

A forward interplanetary shock wave may cause a sudden commencement of a magnetic storm at the Earth, because it causes a dramatic increase in the solar wind's dynamic pressure, which depends on the wind's speed and density, exerted on the boundary of the Earth's magnetosphere. On the other hand, a magnetic field directional discontinuity in the solar wind may rapidly alter the solar wind's electric and magnetic connection to the Earth's magnetosphere by allowing (or disallowing) the IMF, depending on its direction, to interact with the Earth's magnetic field. When the dramatic process called MAGNETIC RECONNECTION (between the IMF and the Earth's field) occurs, the related discontinuity at the interface is of a rotational type. This interaction may lead to a magnetic substorm or major magnetic storm depending on the strength and duration of the magnetic field reconnection process (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS, MAGNETOSPHERE OF EARTH: SUBSTORMS). These processes of magnetic reconnection and shock wave impingement at the Earth's magnetosphere's boundary are prominent components in the study of space weather (see SOLAR–TERRESTRIAL CONNECTION: SPACE WEATHER PREDICTIONS).

Bibliography

Balogh A and Riley P 1997 Overview of heliospheric shocks *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonnett and M S Gianpapa (Tucson, AZ: University of Arizona Press) p 359

Burgess D 1995 Collisionless shocks *Introduction to Space Physics* ed M G Kivelson and C T Russell (Cambridge: Cambridge University Press) ch 5

Burlaga L F 1995 *Interplanetary Magnetohydrodynamics* (Oxford: Oxford University Press)

Colburn D S and Sonett C P 1966 Discontinuities in the solar wind *Space Sci. Rev.* **5** 439

Tidman D A and Krall N A 1971 *Shock Waves in Collisionless Plasmas* (New York: Wiley)

Whang Y C, Zhou J, Lepping R P, Szabo A, Fairfield D, Kokubun S, Ogilvie K O and Fitzenreiter R 1998 Double discontinuity: a compound structure of slow shock and rotational discontinuity *J. Geophys. Res.* **103** 6513

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

R P Lepping

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind Turbulence

There are fluctuations in all quantities that describe the local plasma state of the solar wind. For example, fluctuations in plasma velocity, magnetic field and plasma density are superposed upon the large-scale relatively smooth flow of the wind at speed $\boldsymbol{V}$, the relatively organized large-scale interplanetary magnetic field $\boldsymbol{B}_0$, and slow radial variation of the density ρ. These fluctuations, known as solar wind turbulence in anticipation of their likely connection to complex nonlinear dynamical activity, have been routinely observed by instruments aboard spacecraft for several decades, starting with early observations by spacecraft such as Mariner 2 in 1962 and Pioneer 6 in 1965 (see MARINER MISSIONS, PIONEER MISSIONS). Because of the large amount of *in situ* data that have been accumulated from spacecraft, solar wind turbulence may be the most extensively probed of all turbulent plasmas in nature or in the laboratory. Certain spacecraft missions stand out in terms of their contribution to knowledge of interplanetary turbulence, notably Mariner 10 and HELIOS (exploration of the inner heliosphere), ISEE-3 and IMP (near-Earth orbit), Pioneer 10 and 11 and Voyager 1 and 2 (outer heliosphere) and Ulysses (high solar latitude (see SOLAR WIND: ULYSSES, VOYAGER MISSION)). Advances are continuing through recent missions such as ACE and WIND. Missions that are currently in the planning stages, such as SOLAR PROBE, may answer fundamental questions regarding the origin of solar wind turbulence and its many influences in the heliosphere.

Scales of turbulence

Solar wind turbulence occurs over a wide range of length and time scales, but most of the energy resides in fluctuations for which time and length scales are long compared with the gyromotions of thermal particles about the magnetic field. In this regime, the solar wind plasma can be reasonably well approximated by the equations of MAGNETOHYDRODYNAMICS (MHD). The complex dynamical activity of such a magnetofluid, MHD turbulence, is closely related to hydrodynamic turbulence, which is well studied in a variety of terrestrial contexts (see MAGNETOHYDRODYNAMICS: MAGNETIC RECONNECTION AND TURBULENCE). It is in this perspective that low-frequency solar wind turbulence, as well as the large-scale solar wind flow itself, is most often studied.

A convenient approximation, called frozen in flow, is often invoked to interpret the time series of observations at a single spacecraft as a snapshot of the plasma fluctuations at nearby spatial locations. The mean solar wind flows at speeds ranging from 250 km s^{-1} to 750 km s^{-1}, but the characteristic speeds associated with dynamical activity of MHD fluctuations (sound speed, Alfvén speed, typical fluctuation speed) are much less (10 to 50 km s^{-1}). Therefore structures described by MHD cannot significantly distort during their time of transit past the spacecraft instruments. This frozen-in property allows the structure and the statistical properties of MHD turbulence to be probed in the solar wind for scales ranging from tenths of an astronomical unit (1 AU $= 1.5 \times 10^{13}$ cm) or more down to scales associated with the thermal proton gyroradius (about 50 km). Discussion of the distribution of energy over scales usually refers to the fluctuation energy spectrum, defined in terms of a Fourier transform of the two-point correlation function. For example, the magnetic energy spectrum is $E_B(\boldsymbol{k}) = (2\pi)^{-3} \int \mathrm{d}^3 r \langle \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{b}(\boldsymbol{x} + \boldsymbol{r}) \rangle \, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$. Single-spacecraft data give information only about the reduced spectrum, a function of radial wavenumber k_r, obtained from the spectrum by integrating over the components of $\boldsymbol{k}$ transverse to the radial direction.

Energy cascade

Observed near Earth orbit at 1 AU, solar wind turbulence exhibits features that are suggestive of fully developed MHD turbulence, and also properties that are reminiscent of MHD waves. The power spectra of MHD fluctuations exhibit a characteristic appearance. Due to the frozen-in flow property, the frequency (f) spectra correspond to reduced wavenumber spectra through the relation $f = V_{\mathrm{sw}} k_r / 2\pi$ for solar wind speed V_{sw}. Magnetic energy spectra, for example, typically display a power law, $E_B(k_r) \sim k_r^{-\alpha}$, for a range of frequencies (f) and wavenumbers k_r corresponding to spacecraft frame periods from about 10 s up to several hours. The spectral index α frequently has a value near $\alpha = 5/3$ corresponding to an inertial range or Kolmogoroff spectrum in hydrodynamic turbulence, and can be understood by an analogous argument: beginning with a reservoir of MHD energy at large energy-containing scales, turbulent interactions produce a cascade of energy through eddies (vortices), or magnetic structures of progressively smaller size. At sufficiently small scale, the dissipative effects become significant and this direct cascade terminates by damping of fluctuation energy and a corresponding production of heat. For scales that are much smaller than the energy-containing size, and much larger than the dissipative scales, the dynamics of the cascade may be self-similar, that is, invariant under a multiplicative transformation of length scale, thus producing a distinctive power-law distribution. The occurrence of the inertial range in a typical solar wind magnetic field spectrum near 1 AU is illustrated in figure 1.

There continues to be some debate concerning whether there is an active energy cascade involving solar wind fluctuations. However, there are features that seem to require its presence. The proton temperature at 1 AU is difficult to explain unless heat is added in transit from the solar wind source region. The temperature profile in the solar wind from the inner Helios orbit (0.29 AU) to several tens of AU exhibits a non-adiabatic profile, again suggesting deposition of heat. Turbulent energy can be injected by instabilities (e.g. Kelvin–Helmholtz instability) associated with regions of large velocity shear near edges of high-speed streams (velocity change $\Delta V \approx 200$ km s^{-1}),

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

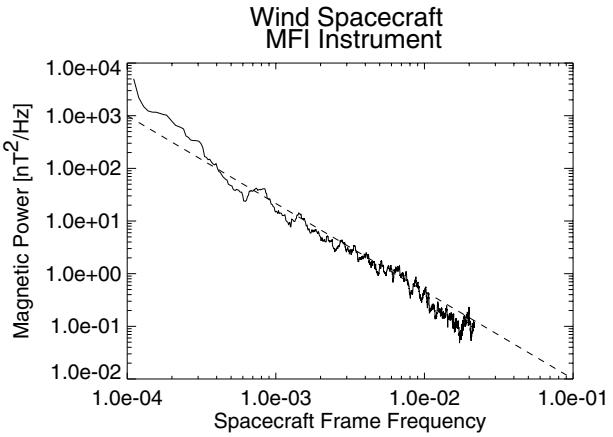


Figure 1. Reduced magnetic energy spectrum from several days of wind data at 1 AU, beginning on 13th October 1997, showing power-law inertial range with index close to the 'Kolmogoroff' value of $-5/3$ (broken line). (Wind data courtesy of Wind/MFI team.)

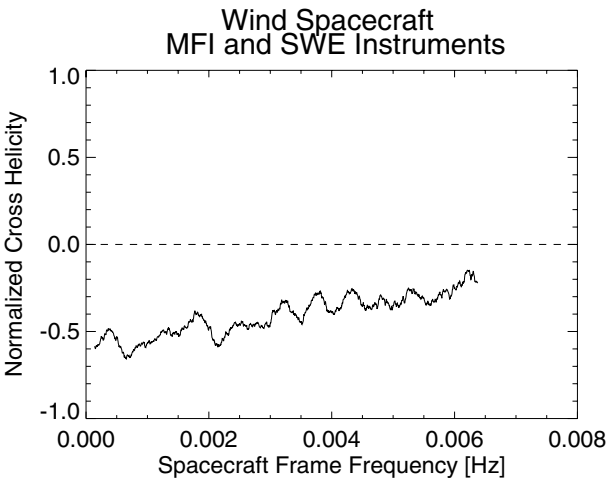


Figure 2. Normalized cross-helicity spectrum $\sigma_c(k_r) = 2\,\mathrm{Re}\langle \boldsymbol{v}_A(k_r) \cdot \boldsymbol{v}(k_r)\rangle / \langle v_A^2(k_r) + v^2(k_r)\rangle$. Values near -1 indicate outward traveling Alfvénic fluctuations. (Wind data courtesy Wind MFI and SWE teams.)

or near microstreams observed at higher latitudes ($\Delta V \approx$ 40 km s^{-1}). In the outer heliosphere, turbulence may be driven by injection of energy due to the 'pick-up' process that occurs in connection with freshly ionized hydrogen or helium of interstellar origin. Taking these sources into account, a cascade and subsequent dissipative heating would be required to prevent build-up of fluctuation energy to levels higher than what is observed.

Alfvénic fluctuations and turbulence

Wave-like properties are also evident in typical observations of solar wind fluctuations. The so-called Alfvénic correlation between the fluctuating components of velocity and magnetic fields are often seen in data at 1 AU. When the correlation is strong, a near proportionality is seen involving the fluctuating magnetic field $\boldsymbol{b}$ and the fluctuating velocity field $\boldsymbol{v}$, in the sense that $\boldsymbol{v}_A \equiv \boldsymbol{b}/\sqrt{4\pi\rho} \approx \pm\boldsymbol{v}$ where ρ is the local plasma mass density. This relation corresponds to an MHD Alfvén wave and the sign of the correlation usually indicates a direction of propagation away from the Sun. It is frequently possible to identify intervals of data that possess a substantial broad band correlation of this type by visually inspecting the $\boldsymbol{v}_A$ and $\boldsymbol{v}$ time series. Figure 2 shows evidence of the presence of these large-amplitude Alfvénic fluctuations in spacecraft data, using a more sensitive technique that extracts the degree of correlation at various scales. In MHD turbulence theory, the quantity $\langle \boldsymbol{v}_A \cdot \boldsymbol{v}\rangle$, called the cross helicity, provides a quantitative measure of the Alfvénic correlation, where $\langle \ldots \rangle$ denotes an appropriate ensemble average. For pure Alfvénic correlation, the cross helicity takes on the maximum value compatible with the fluctuation energy per unit mass, so that $\sigma_c = 2\langle \boldsymbol{v}_A \cdot \boldsymbol{v}\rangle / \langle v_A^2 + v^2\rangle = \pm 1$. Figure 2 shows, for a particular data interval, the Alfvénic correlation over a broad range of frequencies.

High cross helicity intervals are most frequently seen closer to the Sun (Helios) and at higher latitudes (Ulysses). Voyager observations suggest a decrease in Alfvénic correlation in the more distant solar wind. The most obvious interpretation of the Alfvénic fluctuations is that they represent outward propagating waves that escape from the inner corona at the Alfvénic critical point (see SOLAR WIND: CORONAL ORIGINS). In a purely wave picture, the wave spectrum is determined at this inner source surface (at about 10–15 R_s) and propagates outwards as 'fossil' turbulence in accordance with geometrical optics (also called WKB theory). However, in this view, since there is no active turbulence in the super-Alfvénic wind, it is difficult to explain the proton temperature at 1 AU, and the non-adiabatic temperature radial profile. Several effects are routinely observed that are in apparent violation of WKB transport theory, such as systematic radial decrease of cross helicity and decrease of the ratio of turbulent velocity and magnetic field energies. For these reasons the Alfvénic property has also been studied as a property of active MHD turbulence. Theoretical studies have shown, for example, that the normalized cross helicity tends to increase in time for undisturbed freely evolving homogeneous dissipative turbulence. Therefore the decrease of the cross helicity with increasing heliocentric distance points to the existence of some additional process or disturbance. Two such effects have been identified, associated with either non-WKB transport effects, or the injection of turbulent energy near regions of large shear near the boundaries of high-speed streams. Solar wind Alfvénic fluctuations are actively studied using nonlinear models that appear to be able to explain a variety of other features including the presence of discontinuities and distinctive 'arc-polarization,' in which over time the field vector traces a section of a sphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Density fluctuations

Averaged over a suitably large, statistically stationary, data interval, interplanetary density fluctuations are relatively small. Typically the root mean square fluctuation normalized by the local mean density is $\delta\rho/\rho \sim 1/10$. Wavenumber spectra of density fluctuations frequently have a 'Kolmogoroff' $k^{-5/3}$ form. This can be explained if the medium is weakly compressible, and the density fluctuations are a linear response to pressure fluctuations associated with the energy cascade. In this perspective, solar wind turbulence is a form of 'nearly incompressible' (NI) MHD turbulence, which requires low turbulent Mach number M, and that the flow not be dominated by longitudinal compressive magnetosonic waves. Such compressive waves are expected to be heavily damped in the inner heliosphere, while one typically finds that $M \approx 1/3$ for stationary samples. Other features, including the frequent appearance of anti-correlated magnetic pressure fluctuations and density fluctuations, sometimes called pressure balanced structures, emerge naturally from this picture. While some of the same features can be interpreted in terms of magnetosonic waves, the NI picture provides a conceptual link between the statistical properties of the density and other MHD turbulence parameters.

Anisotropic turbulence

In contrast to classical hydrodynamic turbulence which is usually assumed to be statistically isotropic, solar wind turbulence is typically anisotropic. Since early observations, it has been known that the variances of the Cartesian components of magnetic and velocity fields are not rotationally symmetric. Various studies also have concluded that the spectral distribution of fluctuation over wavevectors is also anisotropic. Radio scintillation studies, analysis of *in situ* spectral data, as well as inferential studies based upon analysis of cosmic ray scattering, have all concluded that the spectrum of fluctuations (velocity, magnetic field or density) might contain a spectral ingredient in which a significant fraction of the energy is localized in wavevectors that are nearly perpendicular to the local mean magnetic field. An apparently useful model is one in which the fluctuations consist of two components: a 'slab' ingredient with variations essentially only along the mean field $\boldsymbol{B}_0$ and a 'quasi-two-dimensional' component that varies mainly in the two directions transverse to $\boldsymbol{B}_0$. This model is useful in cosmic ray scattering theory, transport theories and heating theories, in which the rotational asymmetry of the fluctuations has substantial influence. In addition to the mean field direction, another important preferred direction for solar wind turbulence is the radial direction, which can also induce anisotropy, especially through expansion and transport effects.

Dissipation, magnetic helicity and scattering of charged particles

Many features of MHD turbulence in the solar wind, notably in the energy-containing range and inertial range, can be understood principally in terms of MHD theory. At shorter scales, however, kinetic processes must come to dominate the dynamical interactions and ultimately must be responsible for converting fluid-scale energy to heat (see SOLAR WIND: KINETIC PROPERTIES). The standard view is that fluctuations that have the character of magnetosonic wave modes are damped rapidly by Landau damping. At scales approaching the gyroradius of thermal protons, significant damping of incompressive MHD fluctuations must occur, presumably through cyclotron resonance damping, which depends upon the polarization or MAGNETIC HELICITY of the fluctuations, along with mechanisms akin to Landau damping, which are of the non-resonant type, and which are therefore insensitive to helicity. These crucial links of fluid-scale processes to kinetic processes are incompletely understood at present.

It is noteworthy that couplings of interplanetary MHD turbulence to non-thermal particle populations (cosmic rays, and suprathermal particles such as pick-up ions) can also be classified as either resonant interactions, such as pitch angle scattering, or non-resonant interactions, such as stochastic acceleration by induced turbulent electric fields (see COSMIC RAYS, DYNAMO THEORY). Resonant interaction of turbulence with charged particles is also sensitive to the spectrum of magnetic helicity, i.e. the spectrum of $\langle \boldsymbol{b} \cdot \boldsymbol{a} \rangle$ where $\boldsymbol{b} = \nabla \times \boldsymbol{a}$ for vector potential $\boldsymbol{a}$. Of the spectral quantities typically extracted from solar wind data, the helicity spectrum is least well understood, and is characterized by random helicity fluctuations, of both right- and left-handed types throughout the inertial range.

Bibliography

Barnes A 1979 Hydromagnetic waves and turbulence in the solar wind *Solar System Plasma Physics* ed C Kennel, L Lanzerotti and E Parker (Amsterdam: North-Holland) pp 249–319

Goldstein M, Roberts D and Matthaeus W 1995 Magneto-hydrodynamic turbulence in the solar wind *Ann. Rev. Astron. Astrophys.* **33** 283–325

Jokipii J R, Sonnett C P and Giampapa M S (eds) 1997 *Cosmic Winds and the Heliosphere* (Tucson, AZ: University of Arizona Press)

Matthaeus W, Bieber J and Zank G 1995 Unquiet on any front: anisotropic turbulence in the solar wind (*Rev. Geophys. Suppl.*) 609

Tu C Y and Marsch E 1995 *MHD Structures, Waves and Turbulence in the Solar Wind* (Dordrecht: Kluwer)

William Matthaeus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Coronal Origins

To establish the coronal origins of the solar wind is to find its source regions at the Sun. Dominated by electrons and protons, this outflow of fully ionized atoms also carries traces of neutral hydrogen and heavier elements, such as alpha particles, oxygen and iron ions. There are two types of winds detected so far: the fast and the slow (see SOLAR WIND: GLOBAL PROPERTIES). Since they are distinguishable not only by their associated speeds but more importantly by their thermodynamic and compositional properties, their corresponding sources at the Sun must have different characteristics.

The complex magnetic structure of the solar atmosphere is revealed by the electromagnetic spectrum (see SOLAR SPECTROSCOPY AND DIAGNOSTICS) of the solar corona (see CORONA). This complexity contributes to the ambiguity in establishing the source regions of the fast and the slow solar wind. For the PLASMA to escape from the solar surface, it must be connected to magnetic field lines (see SOLAR MAGNETIC FIELD) that stretch out into interplanetary space. How to identify these 'open' field lines is the key to finding the coronal origins of the solar wind.

At present, two distinct and complementary observational techniques have been used to characterize the properties of the solar wind: remote sensing observations and *in situ* measurements. In remote sensing observations, the characteristics are inferred from the electromagnetic spectrum originating from the solar disk and its outer atmosphere to within $30R_s$ (where R_s represents a solar radius) from the solar surface. The field of view is two dimensional, and the information gathered from an extended region of space along the line of sight is heavily weighted by the contributions in the plane of the sky. *In situ* measurements, on the other hand, are local or point measurements made by spacecraft typically beyond $60R_s$. They pertain to the parcel of plasma that intercepts the detector, and hence are limited to a very small volume of space. To establish the coronal origins of the solar wind, the connection between these two types of measurements is essential.

Coronal observations

The corona is revealed in a broad range of wavelengths, and its complex magnetic and density structure is best apparent in the ultraviolet, extreme ultraviolet and x-ray emission which can be achieved from space observations. Because of the intense disk emission in the visible wavelength range, the visible corona can be seen only when the disk radiation is blocked. The invention of the coronagraph (see CORONAGRAPHS) by Lyot in 1930 enabled daily observations of the broad band coronal white light emission, which for centuries was limited to total solar ECLIPSES. Since the solar wind has no detectable emission, simultaneous density and velocity measurements are necessary to establish its coronal sources.

Spectral line observations are weighted by the temperature at which the line is optimally formed. Space observations in the ultraviolet, extreme ultraviolet and x-rays, characteristic of plasmas over one million kelvins in temperature, reveal a complex magnetic structure, with three distinct regions. The most evident are arch-like structures or active regions characterized by intense emission. The other extreme are the polar caps, dubbed CORONAL HOLES, where the emission is practically absent. Coronal holes can also appear on the solar disk at low latitudes. The fraction of the solar surface that is neither coronal holes nor active regions, and appears as diffuse emission with faint large-scale arches, is called the quiet Sun or quiet regions (although far from being quiescent). The intensity of emission from spectral lines varies typically with the square of the density. As the density decreases rapidly as a function of distance away from the Sun, spectral line observations are of value only in the inner corona, within approximately half a solar radius from the solar surface.

The white light or broad band visible emission whose dominant component below $2R_s$ is the polarized K-corona is produced by the photospheric radiation scattered by free electrons traveling along magnetic field lines threading and shaping the corona. Since the K-corona is due to electron-scattered radiation it varies linearly with density as a function of distance. Consequently, polarized brightness (pB) measurements trace the coronal magnetic field to much larger distances than spectral line observations and have currently reached $30R_s$ with space-borne coronagraphs such as the Large Angle Spectroscopic Coronagraph (LASCO) onboard the Solar and Heliospheric Observatory (SOHO) (see SOLAR TELESCOPES AND INSTRUMENTS: SPACE).

The broad band white light emission has proven to be critical in establishing the coronal source of the solar wind. Shown in figure 1 is a sketch of the large-scale coronal magnetic field configuration based on coronagraph images of the Sun and the assumption that the field is basically that of a dipole. The magnetic field lines shown stretching outwards are those forming an envelope around the large-scale arches, also known as streamers (see CORONAL STREAMERS) with the exception of the example of figure 1(*a*) where a few radially extending field lines originate at low latitudes. Coronal holes shown as dark areas are assumed to have only open field lines (not drawn here) and to be the source of the fast solar wind. If coronal holes are the only source of the fast solar wind then this wind can reach spacecraft in the ecliptic plane along the field lines drawn stretching outwards, either from an equatorial extension of the coronal hole, such as in figure 1(*b*), or from a tilt in the dipole, such as in figure 1(*c*). Recent observations, however, have modified this picture.

Comparable with the broad band white light emission, remote sensing radio occultation measurements have played a pivotal role in establishing the coronal density structures associated with the fast and the slow solar wind. Such measurements are made when the radio signal from either a natural source or a spacecraft intercepts coronal density structures and produces a characteristic scintillation pattern such as shown in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

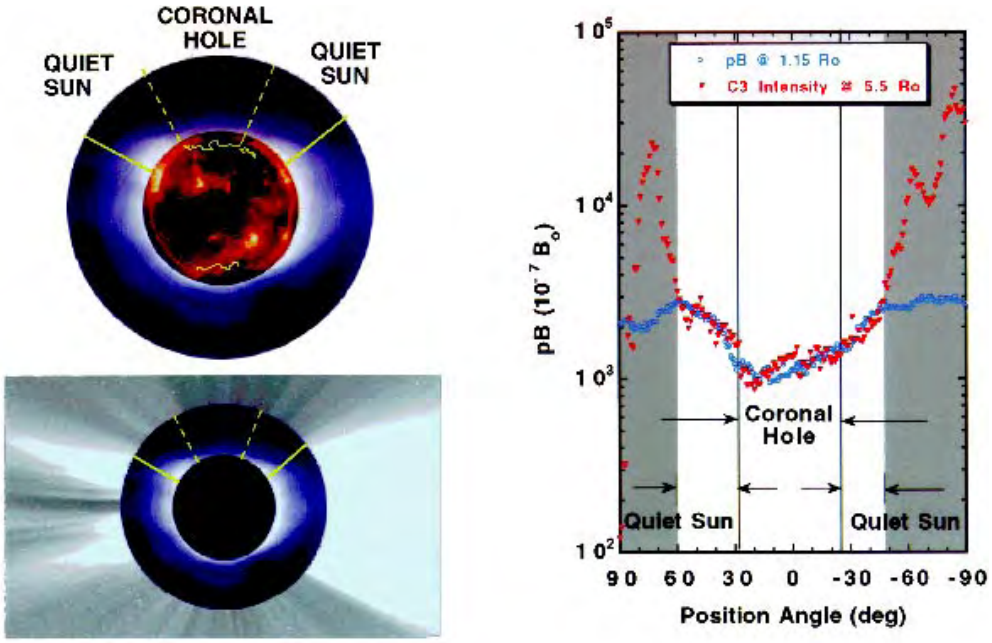


Figure 3. Extension of the coronal hole and the quiet Sun outwards from the solar surface. Top left: composite soft x-ray and pB image, with the coronal hole boundaries shown as derived from He I 18 030 Å disk observations. The dashed lines trace the radial boundaries of the northern coronal hole. The solid line gives the radial boundaries of the quiet Sun. Lower left: composite pB and C2 white light intensity images. The dashed and solid lines are the same as in the top panel. Right: profiles of the pB intensity at $1.15R_s$ (circles) and LASCO C3 white light intensities at $5.5R_s$ (triangles) as a function of position angle.

figure 2. The Doppler scintillation is directly proportional to the electron density fluctuation along the line of sight. While these measurements lack the two-dimensional imaging characteristics of the broad band white light emission, they are more like point measurements. Because of their high sensitivity to changes in electron density over small distances they provide high spatial resolution measurements.

Using Doppler scintillation measurements together with *in situ* velocity measurements, Woo and Martin showed that the slowest wind coincided with conspicuously high levels of density fluctuation associated with coronal streamer stalks. (Streamer stalks are the narrowing of the streamers that extend into interplanetary space.) On the other hand, low levels of density fluctuations were found to be characteristic of the fast solar wind (see figure 2). By comparing radio ranging measurements with simultaneous broad band white light observations, Woo and Habbal found that low levels of density fluctuations could be traced back not only to coronal holes but also to quiet Sun regions.

Further evidence for the origination of the fast solar wind from the quiet Sun as well as from coronal holes emerged from the comparison of density profiles in the corona measured at different radial distances, as shown in the right panel of figure 3. A density profile is obtained by measuring the pB or white light intensity as a function of latitude, or position angle (PA) (measured counterclockwise from 0° north) for a given radial distance. As seen in figure 3, the two profiles at $1.15R_s$ and $5.5R_s$ match over a wide angular extent. The separation of the two profiles occurs at the stalks of the streamers. The angular extent of the coronal hole was established from the radial extension of the boundaries of the coronal hole from disk observations in the He I 10 830 Å line which is commonly used to determine coronal hole boundaries on the solar disk. Beyond the radial boundaries of the coronal hole lies the quiet Sun. The outer radial boundaries of the quiet Sun are determined at the position angle where the two profiles no longer match. That the coronal density profile can be preserved out to several solar radii implies that magnetic field lines tied to the solar wind plasma extend radially outwards from both coronal holes and the quiet Sun.

Velocity measurements provide further support for this view. The only outflow velocity measurements in the corona that are spatially resolved are those recently obtained by the Ultraviolet Coronagraph Spectrometer (UVCS) on SOHO. (Doppler shifts of spectral lines observed on the solar disk do not yield unambiguous results, since the derived downflows or upflows can be connected with closed as well as open magnetic structures.) These measurements extend to $3.5R_s$ in the tenuous polar regions and to $10R_s$ in denser coronal plasmas. One of the unique advantages of this instrument is the measurement of coronal spectral lines, in particular doublets, such as the O VI 1032 Å and 1037.6 Å lines, formed primarily by the resonance scattering of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

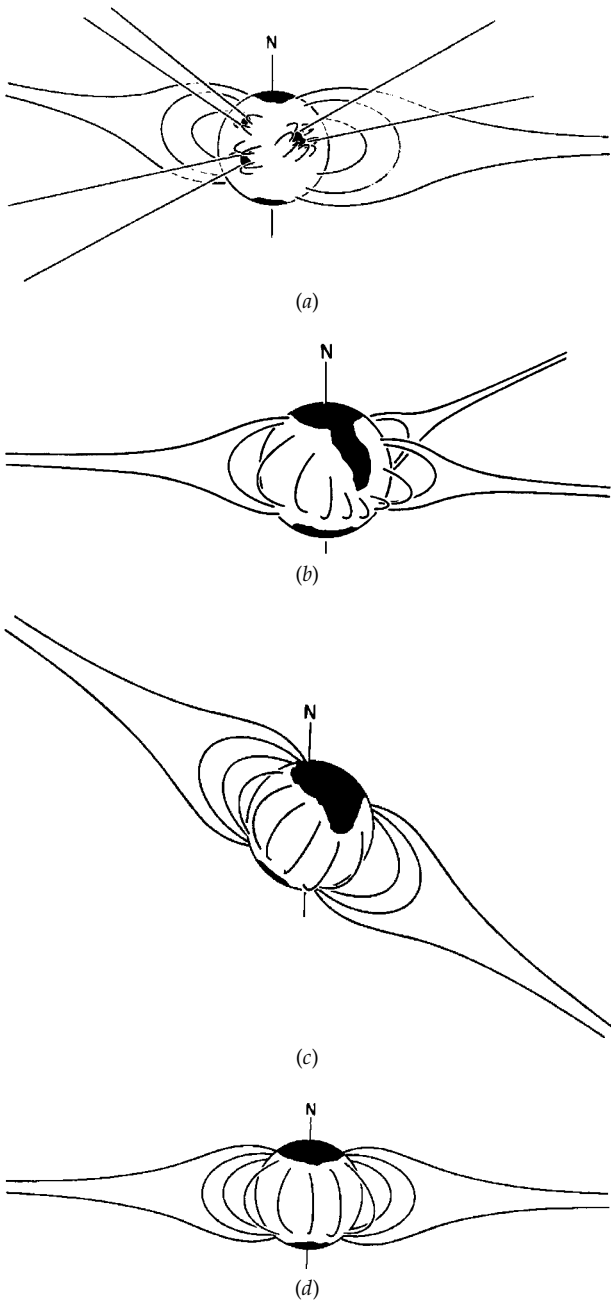


Figure 1. A sketch by Hundhausen of the boundaries between closed and open magnetic field lines, and large-scale closed magnetic field lines, at different times of the solar activity cycle. The sketch is based on visual impressions from the white light emission of the corona. The black shaded areas represent coronal holes. In (*a*) the smaller polar caps and small isolated coronal holes are representative of the magnetic field configuration around the peak of the activity cycle. Panels (*b*) and (*c*) are likely to mimic the configuration between minimum (*d*) and maximum (*a*). The case of a tilted dipole is sketched in (*c*). N points to heliographic north.

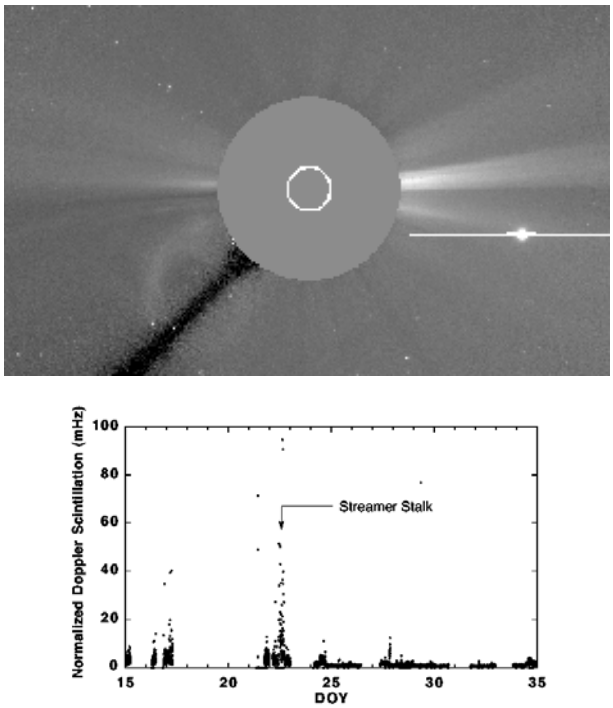


Figure 2. Top: white light image from the LASCO C3 coronagraph taken on 22 January 1997. The innermost circle traces the solar disk. The next circle is the occulter of the C3 coronagraph. The field of view extends from $3.7R_S$ to $30R_S$. The dark region in the lower left quadrant is the support arm of the coronagraph. Jupiter, the bright saturated object on the right, is at the point of closest approach of the line-of-sight radio path from the GALILEO SPACECRAFT which was orbiting Jupiter. North is up. Bottom: radio Doppler scintillation measurements by Galileo during its solar conjunction from 15 January to 4 February 1997, or day of year (DOY) 15–35, when it traversed the plane of the sky from east to west, almost parallel to the equator. The enhanced scintillation from Galileo coincides with the location of Jupiter shown in the top panel, when the radio path intercepted a streamer stalk.

chromospheric (see CHROMOSPHERE) or transition region (see SOLAR TRANSITION REGION) radiation by ions flowing in the corona. The diagnostic which places limits on solar wind velocities is based on the Doppler dimming effect first described by Hyder and Lytes in 1970. As ions flow outwards in the corona, the fraction of the spectral line formed by resonance scattering becomes Doppler shifted out of resonance with the disk emission. Subsequently, the relative ratio between the intensity of the lines forming a doublet changes drastically. In the case of the O VI lines, the intensity ratio of 1032 to 1037 times equals 4 when resonant scattering is dominant and reduces to 2 when only the collisional components are left. The intensity ratio of 1032 to 1037 lines of 2 occurs for an outflow velocity of 94 km s^{-1} and depends only on the velocity distribution along the flow direction.

Shown in figure 4 is an example of UVCS observations taken during the Galileo solar conjunction shown in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

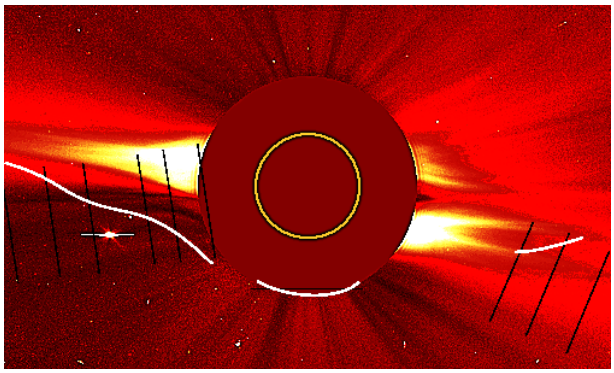


Figure 4. White light image of the corona taken with LASCO C2 on 17 January 1997. The bright object in the east below the equator is Jupiter and indicates the approximate location of the Galileo spacecraft. The field of view spans (2–6)R_S. Shown as black vertical lines are the slit positions of the UVCS detector which is approximately $2R_S$ wide. They are located at 1.9, 2.5, 3, 4, 4.7 and 5.5 solar radii on the east limb at PA= 97°, at $1.9R_S$ in the south polar coronal hole at PA= 180°, and at 4, 4.7 and 5.5 solar radii at the west limb at PA= 247°. The white contours mark the velocity contour value of 94 km s^{-1}.

figure 2. The slit positions of the UVCS detector were chosen to coincide with the passage of the radio signal from Galileo through the corona. By measuring the intensity of the two oxygen lines and their ratio along the slits for different heliocentric distances, contours of the intensity ratio equal 2, shown as white lines in figure 4, were obtained. This contour showed a sharp latitudinal gradient in wind speed close to the stalk of the streamers and a more uniform distribution as a function of heliocentric distance over a large fraction of the inner corona.

The inverse distance of this contour is shown in figure 5(*a*) together with the corresponding latitudinal profile of density from the same time. The density or pB profile again carries the distinct imprint of the coronal hole and the quiet Sun. The corresponding latitudinal velocity profile clearly indicates how the fast wind coincides with the angular extent of the coronal hole and the quiet Sun as determined from the coronal density profile.

Connection between coronal and interplanetary measurements

The connection between coronal velocity and density measurements and their interplanetary counterparts is the only approach currently available to establish the source regions of the different solar wind streams without ambiguity. The advent of Ulysses (see SOLAR WIND: ULYSSES), the first spacecraft to orbit the Sun in a plane perpendicular to the ecliptic and to reach 80° latitude, has made possible the comparison of observations from the same latitudes in the corona and the heliosphere.

Shown in figure 5(*b*) are the Ulysses density and velocity measurements of the south polar region above −30° latitude when the wind speed was consistently

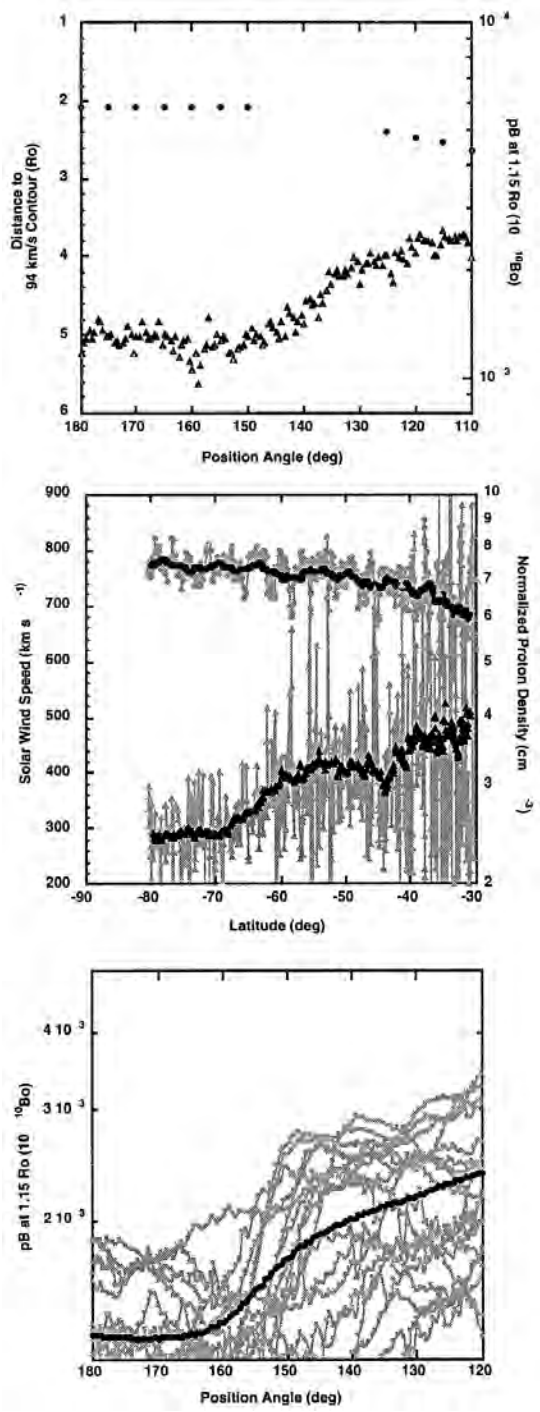


Figure 5. (*a*) pB at $1.15R_S$ and the inverse heliocentric distance of the 94 km s^{-1} velocity contour level versus PA from data corresponding to figure 4. (*b*) Daily averages of Ulysses velocity and normalized density measurements during the slow scans of the south polar region. The larger symbols represent 27-day sliding window averages. (*c*) 19 representative pB profiles at $1.15R_S$ for PA= −30 to −90° latitude for the same time period of the data in (*b*). The black solid line is the average of 253 profiles covering the whole time period.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

higher than 650 km s^{-1} characteristic of the fast solar wind. The data were collected during the slow scans of the south polar region for latitudes from -30° to -80.2° from 3 May 1993 to 15 September 1994. The daily averaged data are shown in grey and the 27-day averages are given as bold symbols. The corresponding averaged coronal pB profile from the same time period is shown in figure 5(*c*). Representative individual pB profiles from which the average was made are shown in grey.

Comparison of the coronal and interplanetary density profiles shows a remarkable resemblance indicating a preservation of the density profile from the corona out to interplanetary space. While the density is minimal between -68° and -80°, and from the corresponding PA= 158°–180°, the subsequent gradients also match in both measurements. X-ray and He I 10 830 Å observations of the solar disk indicate that the latitude at which this rise occurs corresponds to the coronal hole–quiet Sun boundary at the Sun. Its appearance in the Ulysses measurements indicates that this boundary too is preserved. Given that the corresponding velocity measured in interplanetary space is typical of the fast wind, this provides direct evidence that the fast wind originates from both coronal holes and the quiet Sun.

The anticorrelation between density and velocity in the Ulysses measurements and the slow decline of the solar wind speed at the boundaries of the polar coronal holes are also remarkably similar to the pattern observed in the inner corona which is shown in figure 5(*a*). The remarkable resemblance of the density and velocity profiles in the corona (figure 5(*a*)) and in interplanetary space (figure 5(*b*)) indicates not only that the density is anticorrelated with velocity in both regions but that the fast solar wind can be traced back to both coronal holes and the quiet Sun. In particular, a clearly defined transition between the radial extension of the polar coronal holes, where the density is lowest, and the surrounding quiet Sun is detected in both measurements, with the angular extent of both regions being preserved from the Sun out into interplanetary space. The distinction between coronal holes and quiet Sun is also evident in the latitudinal profiles of velocity. Anticorrelated with density, the solar wind speed is highest within the angular extent of the polar coronal holes and decreases gradually beyond their boundaries. The connection between coronal and interplanetary measurements provides unambiguous evidence that the fast solar wind detected at mid latitudes by Ulysses around solar minimum originates from the quiet Sun, while the faster wind at latitudes above 60° has its origin in the polar coronal holes.

That the fast solar wind can be traced back to the quiet Sun as well as polar coronal holes implies that it originates from a large fraction of the solar surface. The coronal source regions of the slow solar wind remain more elusive. Radio occultation measurements clearly identify streamer stalks with the slow wind and with magnetic field polarity reversal, while the slowest wind is systematically connected to the heliospheric current sheet in interplanetary space. The heliospheric current sheet is then undoubtedly identified with the extension of these stalks. At present, the coronal source of the slow solar wind is not established. Given that the composition of the slow solar wind is different from that of the fast, it is likely to be associated with the interaction of the closed arch-like magnetic structures and the magnetic field lines that extend into interplanetary space.

Future prospects

Despite the wealth of space- and ground-based observations, the coronal origins at the Sun of both the slow and fast wind remain a puzzle. While present-day observations indicate that the fastest winds measured so far originate from polar coronal holes, and that the quiet Sun is also a source of the fast wind, the exact origin at the Sun is unknown. In particular, what defines an open versus a closed magnetic field line is still a mystery since both coronal holes and quiet regions have a very intricate magnetic field pattern and are indistinguishable when observed in photospheric or chromospheric spectral lines. Future space missions which can reach to within $30R_s$ of the Sun and which combine remote sensing and *in situ* observations, such as the proposed SOLAR PROBE, are the key to unraveling some of these mysteries.

Bibliography

Habbal S R, Woo R, Fineschi S, O'Neal R, Kohl J L, Noci G and Korendyke C 1997 Origins of the slow and the ubiquitous fast solar wind *Astrophys. J.* **489** L103–6

Hundhausen A J 1977 An interplanetary view of coronal holes *Coronal Holes and High Speed Wind Streams* ed J B Zirker (Colorado University Press) pp 225–329

Woo R and Habbal S R 1999 Radial evolution of density structure in the solar corona *Geophys. Res. Lett.* **26** 1793–6

Shadia Rifai Habbal and Richard Woo

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Corotating Interaction Regions

For several years around each solar activity minimum corotating interaction regions are the dominant and most extensive feature in the heliosphere except for the termination shock. At this time the solar atmosphere shows a well developed and long-lasting stable structure of polar CORONAL HOLES, often extending far down to low latitudes, and a well-developed narrow streamer belt region encircling the solar equator. A quasi-stationary flow of high-speed SOLAR WIND originates from well above the coronal holes and from above a certain height propagates radially out into the heliosphere at about 800 km s^{-1}. From above the streamer belt region a much slower solar wind originates and starts propagating into the heliosphere. Due to the ∼26 day rotation of the Sun and depending on the shape of the coronal holes the faster wind starts pushing the slow wind ahead. This is the origin of corotating interaction regions (CIRs) ahead of fast solar wind streams which develop high-pressure regions and eventually shocks bounding them. As the solar wind is carrying out frozen-in solar magnetic field lines the different plasma packets cannot mingle with or penetrate into each other. CIRs, however, undergo substantial developments with solar distance and latitude which are not yet fully investigated due to the limited simultaneous observations throughout the heliosphere. They are able to accelerate charged particles up to the MeV region and cause recurrent decreases of the galactic and anomalous COSMIC RAYS. The origin, development and effects of CIRs are discussed in this article.

Origin and formation of corotating interaction regions

The solar wind (see SOLAR WIND: GLOBAL PROPERTIES) results from the supersonic expansion of the solar CORONA, the Sun's hot outer atmosphere. It consists of ionized gas or plasma that propagates out into interplanetary space with velocities between about 250 and 850 km s^{-1}. The coronal expansion is modulated by the solar magnetic field that originates from the SOLAR PHOTOSPHERE and penetrates the corona. The interaction between the field and the expansion of the corona results in a highly structured solar corona which varies considerably with the solar activity cycle (see figure 3 of SOLAR WIND: GLOBAL PROPERTIES). Near solar minimum the solar corona generates basically two very different types of solar wind: the fast solar wind originating from coronal holes and the slow solar wind related to the streamer belt region.

Coronal holes and the fast solar wind

Coronal holes appear dark on images of the corona because they have lower than average temperatures. They are regions of open magnetic fields and the source of fast solar wind streams with speeds above 600 km s^{-1}, low plasma densities (3 cm^{-3}) and high temperatures (3×10^5 K). During the decline of the solar activity cycle large-scale coronal holes develop in the polar regions of the Sun and often create extensions towards the equator. They rotate rigidly with the equatorial rotation rate of the Sun (in contrast to the photosphere which rotates more slowly at higher latitudes than at the equator). High-speed streams generally originate from unipolar magnetic field regions and often develop into stable large-scale structures observed for many months at decreasing and low solar activity. The high-speed flow is a remarkably stable equilibrium state of the solar wind which often lasts for many months.

The streamer belt region and the slow solar wind

The CORONAL STREAMER belt is a region of closed magnetic loops containing dense and hot plasma. The magnetic equator is usually completely covered by magnetic arcades. The fields are sufficiently strong to constrain the plasma from expanding outward. At larger coronal heights the magnetic field weakens and the arcades can be opened by the pressure of the coronal plasma which is then free to expand outward. This outflow is observed as helmet-like streamers in the corona and as slow and dense solar wind far from the Sun. The slow solar wind with speeds of about 250 to 400 km s^{-1}, is generally less ordered, denser with ∼10 cm^{-3} and has a lower helium content indicating a higher release location in the corona. A magnetic field polarity reversal is embedded in the slow solar wind, which reflects magnetic control of the expansion. It is possible that the slow solar wind is purely transient and associated with intermittent reconfigurations of the coronal magnetic fields in the streamer belt region.

At solar minimum, the CORONAL LOOPS and the streamers are mainly observed in a narrow belt around the solar equator. The coronal magnetic field can be approximated by a dipole which is slightly tilted against the rotation axis of the Sun with the tilt increasing with solar activity. At higher solar activity the tilted streamer belt reaches higher heliolatitudes until at solar maximum most of the corona is covered by loops and appears bright so that a dipole field approximation is no longer applicable.

The formation of a corotating interaction region

A corotating interaction region develops in the heliosphere when a quasi-stationary high-speed solar wind stream follows and compresses a slow solar wind steam along the same radius-vector from the Sun. This happens due to the solar rotation when a coronal hole (or a coronal hole extension), as a quasi-stationary source of fast solar wind, is situated to the east of a source region of slow solar wind at the same heliographic latitude. The radially aligned parcels of plasma have different solar origin and carry different magnetic field lines inhibiting interpenetration. At the stream interface a region of compressed plasma starts to build up. The interaction region develops to a corotating pattern aligned with the Parker spiral (Parker 1958). While traveling out, the compression region

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

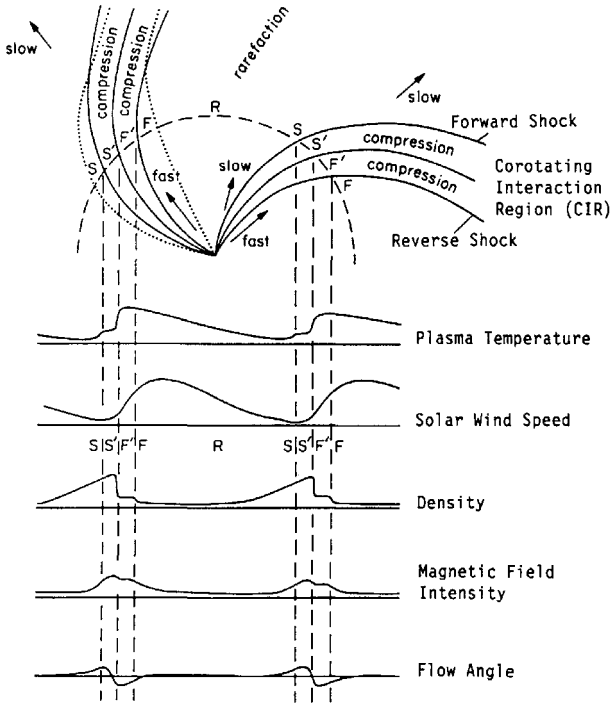


Figure 1. Schematic diagram of two CIRs corotating with the Sun and solar wind and magnetic field signatures associated as observed for example at 1 AU. (From Richardson I *et al* 1993 *J. Geophys. Res.* **98** 13.)

widens, becomes more pronounced, and develops shock waves at the leading and trailing edges.

Figure 1 shows schematically the solar wind plasma and magnetic field properties in the ecliptic plane associated with CIRs. The circular dashed arc represents the orbit of a 1 AU spacecraft. Solar wind streams are denoted S (slow wind), S′ (compressed, accelerated slow wind), F′ (compressed, decelerated fast wind) and F (unperturbed fast wind). The S′ and F′ regions together, separated by a stream interface (SI), constitute the corotating interaction region (CIR). In the outer heliosphere it is bounded by a forward and a reverse shock. Interplanetary magnetic field lines (dotted) link the shocks to the observer, e.g. at 1 AU where the lower parts of figure 1 show schematic changes in plasma temperature, speed, density and flow angle (compared to radial).

CIRs are well developed at 1 AU except for the shocks. A prominent feature is the stream interface (SI) separating the originally slow dense plasma from the originally fast thin plasma. Stream interfaces are identified by abrupt simultaneous drops in plasma density and rises in the proton temperature. The pressure peaks close to the interface. In the process of pressure release the slow and dense plasma ahead of the interface is deflected in the sense of the corotation motion while the less dense plasma behind the interface is oppositely deflected.

Coronal origin of the stream interface and the corotating interaction region in general is the western border of a coronal hole. Details of this border structure are often propagated out into interplanetary space, where changes observed within the corotating interaction region, despite temporal evolution and dynamics, can often be traced back to the source region with remarkable accuracy of about 2°. This is confirmed by measurements of the elemental and charge state composition of the solar wind. Compositional signatures (e.g. Mg/O ratios) are the best indicators of coronal origin and can be used to define the complex interfaces between different solar wind types (see SOLAR WIND COMPOSITION).

Interplanetary development of CIRs

The formation and evolution of CIRs is a strong function of coronal morphology and temporal characteristics of the source regions and the transition region between closed and open magnetic field lines.

In the inner heliosphere the transitions between slow and high velocities are commonly steeper at 0.3 AU than beyond 0.5 AU. When the plasma streams and the interface propagate outward the continuing interaction between the streams increases the pressure in the region considerably. The interface becomes more azimuthal. The velocity difference across the interface exceeds the local wave speeds and two shocks bounding the interaction region develop. As the wave speed is lower in the slower, cooler stream the forward shock is generally formed first at the downstream border of the interaction region whereas the reverse shock at the up-stream edge develops at larger distances. Forward–reverse shock pairs are commonly observed beyond 2 to 3 AU.

Beyond 10–18 AU the forward shocks start to encounter the reverse shock of a CIR released about half (in the case of two CIRs per rotation) or one solar rotation earlier (which is actually the same CIR). At these distances merged interaction regions (MIR) develop and all solar wind material has at least been shocked once.

CIRs and their successors continue to evolve while convected to greater heliocentric distances. Shocks decline in frequency and strength and MIRs are replaced by corotating pressure enhancements.

CIRs and energetic particles

Origin, injection and acceleration of energetic CIR particles

The first hints of CIRs in the early 1970s were 27-day recurrent particle events with specific features, for example slow rise times and no obvious velocity dispersion due to particle propagation. Beyond 1 AU corotating energetic particle events, as they are now called, usually consist of two reasonably well separated particle intensity peaks centered approximately on the locations of the forward and reverse shock. Above 0.5 MeV the intensities at the reverse shock are usually larger than at the forward shock. This is shown in figure 2 for an event in January 1993 observed by Ulysses (see SOLAR WIND: ULYSSES) at 5 AU and at 24°S. It is obvious that around 100 keV the peaks of the proton intensities are about equal at the two shocks F and R whereas above 3 MeV there is hardly any

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

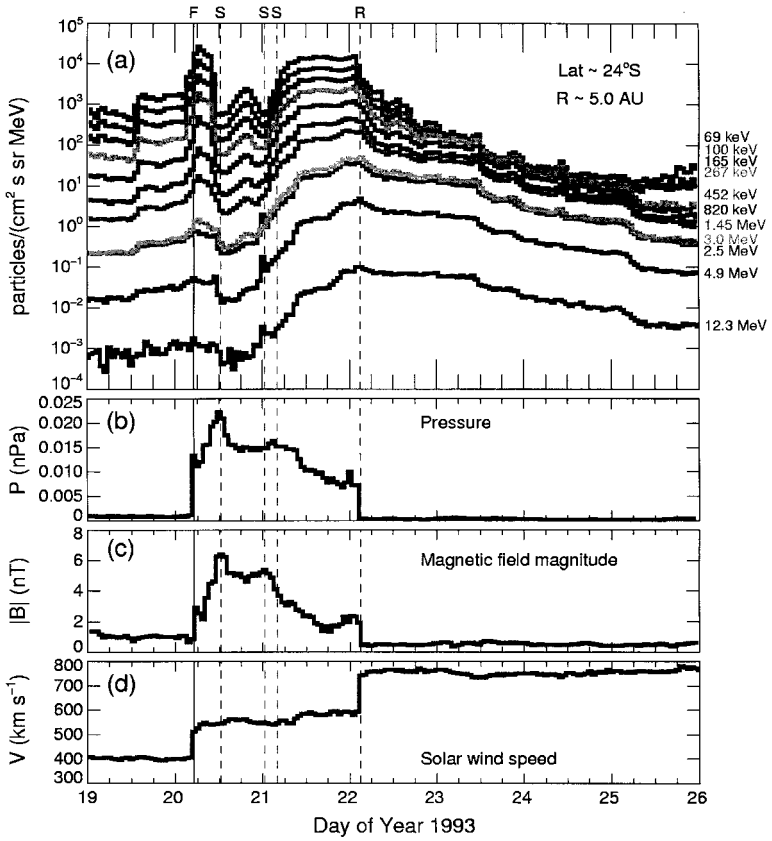


Figure 2. Hourly averages of (*a*) the 50 keV to 20 MeV proton intensities measured by two Ulysses instruments, (*b*) the pressure, P, (*c*) the magnetic field magnitude, $|B|$, and (*d*) the solar wind speed, V, from day 19, 1993, to day 26, 1993. The intensity measured in each of the 11 discrete energy channels is denoted by the geometric mean of the upper and lower energy bounds of that channel. The solid (marked F) and dashed (marked R) vertical lines denote the times of arrival of the forward and the reverse shock, respectively. The dashed vertical traces, marked S, denote the arrival times of the stream interfaces. (From Mason G *et al* 1999 *Space Sci. Rev.* **89** 327.)

increase left at the forward shock. The intensities decrease considerably, especially at low energies near the stream interfaces marked S in figure 2.

It took several years until CIRs were established as a cause for the recurrent particle increases. We now know that CIRs are capable of accelerating particles from the solar wind or suprathermal particles to energies between tens of keV up to ∼10 MeV at a wide range of solar distances and latitudes (see SOLAR WIND: ENERGETIC PARTICLES). Highest intensities are observed at ∼4 AU with a $1/r^2$ fall-off at large distances and a decay rate in the direction of the Sun of ∼300%/AU. Highest intensities versus latitude are observed around 20° with an e-folding angle of ∼8°. Ulysses observed the transport of CIR ions to latitudes well above the shock locations up to the highest latitudes of 80° in certain cases. CIR particles propagate away from source region along the magnetic field lines connecting the shock with the observer as shown in figure 1. At 1 AU a sunward flow is generally observed. However, non-field-aligned flows at peak intensities were reported at 1 AU as well. Generally, a strong inhibition of particle transport across stream interfaces is observed.

The spectral forms of CIR ions can be described by a power law spectrum from 30 keV/n to 1 MeV/n and a steepening spectrum above. At the high-energy end the spectrum merges with the ambient cosmic ray and anomalous cosmic ray (ACR) spectra. At the reverse shock the spectrum is generally harder than at the forward shock, but the forms do not change out to distances of several tens of AU.

The ion intensity peaks in the streamer belt region coincide with forward and reverse shock observations. Low-energy ions (tens of keV), however, are observed at and below 1 AU even in the absence of shocks. The time profiles are the same from He through Fe. Compositions are similar to the solar system observations except for factors of two to three enhancements of He and C relative to oxygen. The He/O and Ne/O ratios increase with the solar wind speed. The He/H abundance increases from 1 to 5 AU. The Mg/O ratio is the same for both forward and reverse shock periods and close to the average in solar wind slow and fast streams. Electrons are observed from 50 to hundreds of keV at several AU; however, only small intensities are observed at 1 AU.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3 shows the time history of CIRs and related particle increases for 12 solar rotations observed by Ulysses near 5 AU in 1993. A major restructuring of the corona is reflected in the time profiles of the particle intensities as well, e.g. by the shift in longitude between events 9–12 and 13–14. Sequences like this one are typical for the period prior to and at solar minimum displaying either one or two CIRs per rotation.

The huge three-dimensional extent of CIRs and its importance in structuring the quiet heliosphere became obvious only from Ulysses observations at higher latitudes in 1994. The maximum latitude extent of CIRs correlates well with the highest latitudes of the coronal streamer belt (which can also been described as tilt of the heliospheric current sheet) with the tendency of the reverse shock extending to somewhat higher latitudes and the forward shock extending only to lower latitudes than the streamer belt at larger radial distances.

The observation of CIR related accelerated particles (and modulation effects on galactic and ACR particles) almost up to the poles even in the absence of local shocks is surprising. Observed effects can best be ordered by field line connection to remote shocks and particle transport along the field lines with transit-time-dependent loss rates.

It is generally agreed that diffusive or first-order Fermi acceleration at the shocks is responsible for corotating energetic particle events. Charged particles originally from the high-speed solar wind are multiply scattered between the shock front and magnetic irregularities upstream. Acceleration is most efficient in the region of well developed shocks at about 3–5 AU. Energetic ion increases have also been observed at trailing edges of compression regions even when no reverse shock was detected, suggesting that the presence of the shock is not mandatory for moderate acceleration.

Pick-up ions are also accelerated within the corotating interaction regions, e.g. by stochastic acceleration in the magnetosonic turbulence. A possible injection/acceleration mechanism for pick-up ions at (quasi-)perpendicular shocks is shock surfing where particles are trapped between the electrostatic shock potential and the upstream Lorentz force. Some of the pick-up ions already constitute a suprathermal particle population. These ions becomes further accelerated by shock drift acceleration or in a Fermi type shock acceleration process if they are able to diffuse across field lines.

Modulation of cosmic rays and ACRs by CIRs

In addition to the CIR related intensity increases (from the acceleration of charged particles) above about 30 MeV/n recurrent decreases are also observed. These particles originate from the local interstellar space outside the heliosphere (galactic cosmic rays) or from local sources within the heliosphere (ANOMALOUS COSMIC RAYS or ACRs). The cosmic ray flux is lowest close to the reverse shock. This is in good agreement with 1970s observations from the PIONEER spacecraft. Ulysses found at high latitudes that in contrast to the solar wind speed galactic cosmic

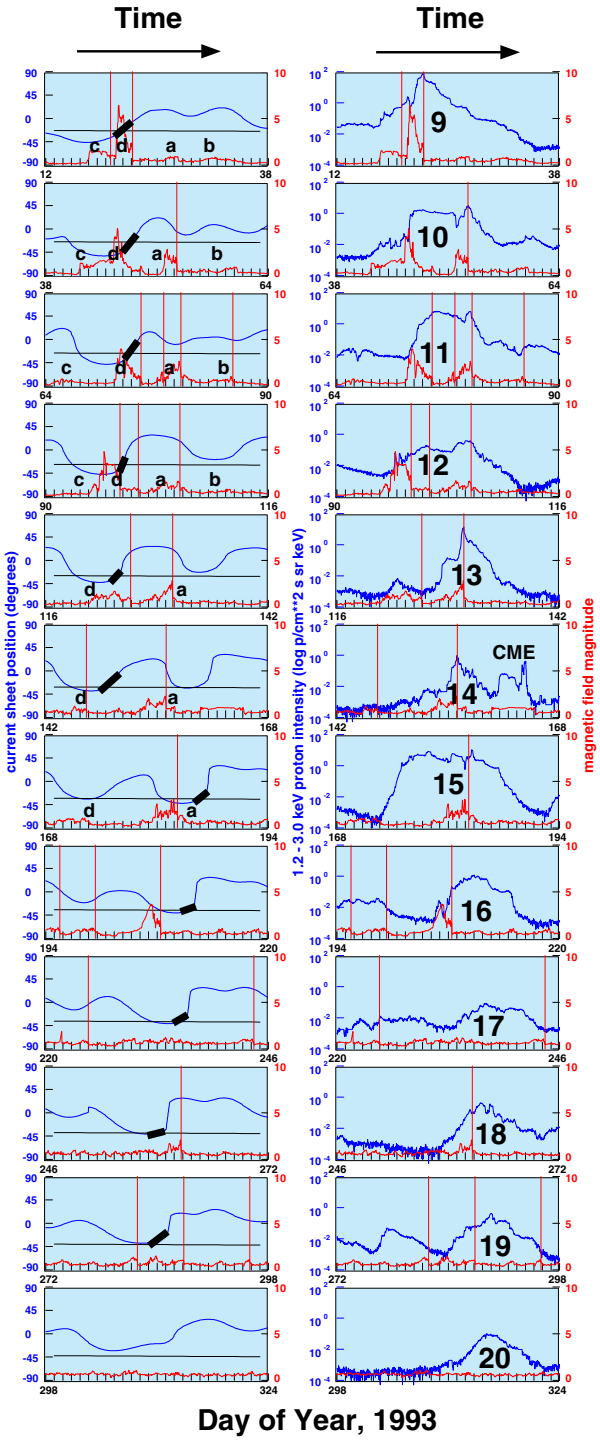


Figure 3. Data from 12 solar rotations, each of 26 days, starting on 12 January 1993. Time runs from left to right. The left column shows, in blue, the expected latitude of the current sheet at the position of the spacecraft. The black trace shows the position of the spacecraft, while the red trace shows the magnitude of the magnetic field. The right column shows, in blue, the 1.8–3.8 MeV proton intensity. The vertical lines show the observed times of reverse shocks, whilst the solid bars show the expected positions of the interaction regions. (From Kunow H *et al* 1999 *Space Sci. Rev.* **89** 221.) **This figure is reproduced as Color Plate 57.**

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rays continued to be modulated on the time-scale of one solar rotation. The amplitude of the recurrent cosmic ray decreases show a maximum at 35° and is highest between 1 and 2 GeV for galactic cosmic ray protons.

Bibliography

Balogh A, Gosling J, Jokipii R, Kallenbach R and Kunow H (ed) 1999 *Corotating Interaction Regions* (Dordrecht: Kluwer) (see also *Space Sci. Rev.* **89** 1–411)

Parker E N 1958 Dynamics of the interplanetary gas and magnetic fields *Astrophys. J.* **128** 664–75

Schwenn R and Marsch E (ed) 1990/1991 *Physics of the Inner Heliosphere* vols I and II (Berlin: Springer)

Horst Kunow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Energetic Particles

The energetic particle populations that we observe in the SOLAR WIND are accelerated in a variety of local and distant sources. We distinguish different populations by their temporal, spatial and angular distributions, energy spectra, abundances and ionization states. They tell us the physics of acceleration mechanisms and the properties of remote sites that are otherwise invisible or inaccessible.

Sources

The acceleration of particles to high energies is a remarkably common occurrence in the ionized plasmas of the heliosphere and presumably throughout the Galaxy. Often the presence of energetic particles in distant sources is inferred from observations of the photons they produce as they collide with matter. Radio emission, hard x-rays and gamma rays are produced when energetic electrons and ions interact with magnetic fields or with material in the ambient plasma. However, at times it is also possible to observe directly a sample of the accelerated material that has propagated to us at high speed along the often-tangled magnetic fields in space. In the heliosphere these energetic particles have disclosed a variety of new acceleration sites where the matter is too tenuous for photon production.

Particles are known to be accelerated to MeV and even GeV energies in solar FLARES, at the shock waves driven out from the Sun by SOLAR CORONAL MASS EJECTIONS (CMEs), in planetary MAGNETOSPHERES and at planetary bow shocks. They are also accelerated at co-rotating interaction regions (CIRs) between high- and low-speed streams in the solar wind (see SOLAR WIND: COROTATING INTERACTION REGIONS) and at the solar wind termination shock at the outer edge of the heliosphere (see SOLAR WIND SHOCK WAVES AND DISCONTINUITIES). In addition, we observe the galactic cosmic rays that have probably been accelerated by shock waves from supernovae. We distinguish these different populations of energetic particles and identify their sources by the particle arrival timing and associations with other phenomena, by their spatial distribution and arrival directions, by their energy spectra, and by the abundances of elements and the ionization states of the ions in these populations.

As measurements have become more sensitive and complete, it has become possible to distinguish particle sources that were previously unclear. One important example has been the identification of different sources for the impulsive and gradual solar energetic particle (SEP) events as flares and CME-driven shocks, respectively. Once it was believed that all of these particles came from flares, the so-called 'flare myth'. Another example has been the new evidence that particles upstream of the Earth's bow shock are actually accelerated by that shock and are not merely escapees from the magnetosphere (see MAGNETOSPHERE OF EARTH: BOW SHOCK).

Impulsive solar flares

Impulsive solar flares are sources that provide us with considerable information on particle acceleration. We can observe the accelerated particles in space and the photons produced near the Sun. Most of the acceleration in flares takes place on closed magnetic loops in the CORONA. As these loops become tangled from circulation at their photospheric footpoints or as new magnetic flux emerges from the photosphere, energy can be suddenly released by magnetic reconnection. The associated wave turbulence is the probable source of energy for stochastic acceleration (see SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS). As particles scatter into the loss cone, they plunge into increasingly dense material near the footpoints of the loops where they lose energy by Coulomb interactions. A small fraction of the electron's energy is lost to x-ray bremsstrahlung and a small fraction of the ions interact to produce γ-ray lines from excited nuclei in the beam or the ambient material. Electrons escaping along open field lines produce type III radio bursts as they stream out along the magnetic field through plasma of decreasing density. Finally, satellites in space can observe both the energetic electrons and ions directly.

As the particles stream out along magnetic field lines they are scattered somewhat by Alfvénic fluctuations of the magnetic field, although the scattering mean free path of ~1 AU is comparable with the distance to the Sun. This, together with different particle velocities and pitch angles with respect to the field, can spread the particle arrival times over a period of several hours. Typical profiles of particle intensity versus time for a series of impulsive events are shown in figure 1(a). Different particle speeds produce 'velocity dispersion' in the earliest arrival times, a sophisticated way of saying that the fastest particles arrive first. The propagation time, $t = L/\mu v$, where L is the distance along the field line, μ is the average cosine of the pitch angle and v is the particle velocity. Photons and relativistic particles take at least 8 min to propagate 1 AU, 50 keV electrons or 100 MeV amu^{-1} ions take 18 min, and 1 MeV amu^{-1} ions take 2.9 h. This relationship can be used to associate all particle species and energies with a particular flare at the Sun, within measurement errors of a few minutes. Type III radio observations provide additional information that can track the 10–100 keV electron population in direction and distance as it moves out from the Sun along the field line. Particles from impulsive events are only seen over a solar-longitude interval of ~30° where field lines are fairly well connected to the flare (see SOLAR FLARES: IMPULSIVE PHASE). Although poorly measured, this longitude interval is of interest because it is controlled by the random walk of magnetic field lines.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

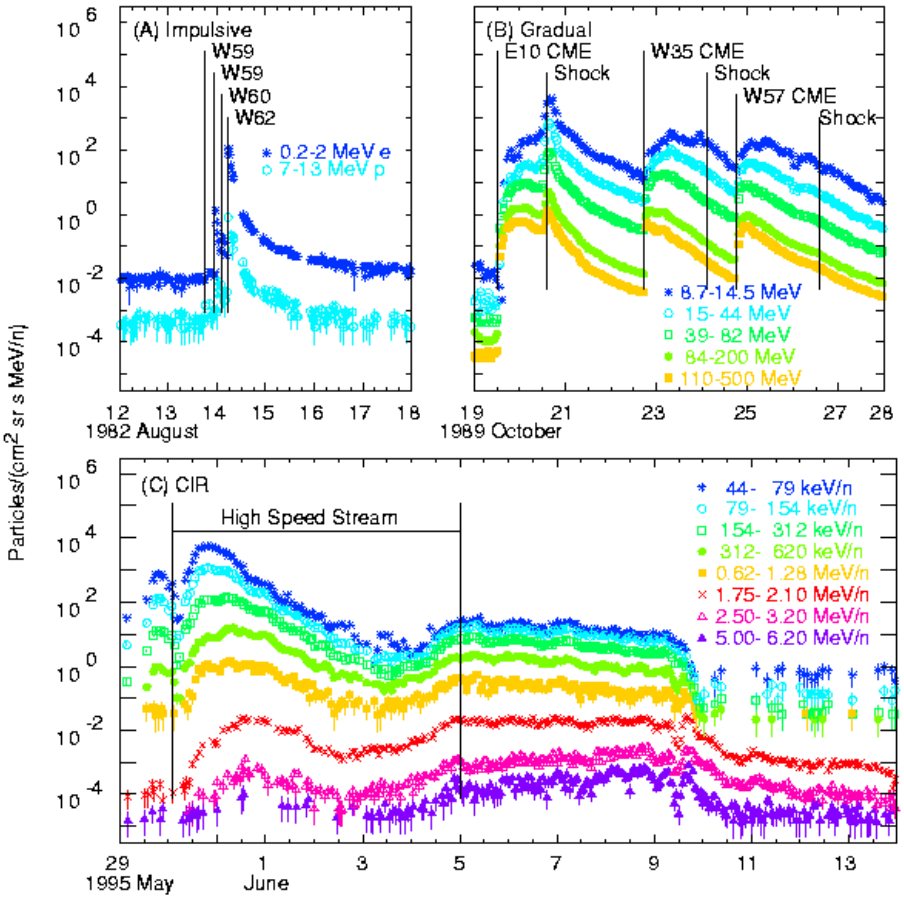


Figure 1. Intensity–time profiles are shown for various sources using the same time and intensity scales for each. (a) Electron and proton data from ISEE 3 are shown for a series of flares at the times and solar longitudes indicated. (b) Proton intensities from the NOAA/GOES spacecraft are plotted for a series of gradual events at the energies shown. Times and longitudes of CME launch at the Sun and the shock arrival at Earth are indicated. (c) Intensities of He at various energies are shown during a CIR event as observed on the WIND spacecraft.

Particle abundances are one of the most distinctive features of impulsive-flare events. We often find the ratio of abundances of isotopes of He, ^{3}He/^{4}He~1 in these events while this ratio is ~5 × 10^{-4} in the solar atmosphere or the solar wind. For this reason, these impulsive-flare events are often called ^{3}He-rich events. High electron abundances and more modest enhancements of heavy ions also occur. On average, Ne/O, Mg/O, and Si/O are enhanced by a factor of ~3 and Fe/O by a factor of ~7 in these events, relative to the corresponding coronal or solar wind abundances (see table 1 and SOLAR ABUNDANCES). Element abundances derived from the intensities of broad γ-ray lines emitted by energetic particles inside the flares are consistent with those measured for the energetic particles in space. These abundances are taken as evidence of resonant wave–particle interactions during stochastic acceleration. The huge enhancement of ^{3}He may occur because intense beams of streaming electrons excite electromagnetic ion cyclotron waves between the gyrofrequencies of the dominant species, H and ^{4}He, in the flare plasma. The rare isotope ^{3}He, the only species whose gyrofrequency lies in this region, is preferentially accelerated as it absorbs energy from these waves. Heavy ions may be enhanced by other wave modes or by second-harmonic interaction with the same waves. Acceleration of particles by resonant interaction with waves also occurs in other regions of high magnetic field energy, such as the auroral region of the Earth where both the particles and resonant waves have been observed together in the phenomenon known as ‘ion conics’.

Measurements of ionization states of the energetic ions show that all elements up to Si are fully ionized, while Fe has ~20 of its 26 electrons removed, corresponding to an electron temperature of ~15 MK (1.5 × 10^7 K). This is not surprising for a flare temperature; atomic spectral lines of highly ionized Fe are often seen from flares. However, if the elements C, N, O, Ne, Mg,

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and Si were fully ionized before acceleration, they would all have the same charge-to-mass ratio, $Q/A = 0.5$, hence the same magnetic gyrofrequency. In this case, it would be impossible to produce relative enhancements since all ions would resonate with the same waves. Actually, the pattern of enhancement is consistent with the values of Q/A that would exist in a 3–5 MK plasma, a typical active-region temperature. This suggests that the ions are drawn from the ambient plasma and accelerated early in flare; stripping then occurs later as the electron temperature rises.

Gradual events and CME-driven shocks

The most intense energetic-particle events we see near Earth are the 'gradual' events produced by particles accelerated at shock waves driven out from the Sun by fast CMEs. Essentially all (>96%) large proton events are associate with CMEs. Since most CMEs have speeds that are near or slightly above that of the solar wind, they do not produce shocks, nor do they accelerate particles. Particle acceleration only occurs for the fastest ~1% of CMEs. Shocks with speeds >500 km s^{-1} usually accelerate particles while those with speeds >750 km s^{-1} nearly always do. The largest SEP events of a solar cycle are produced by shocks with speeds of 1500–2000 km s^{-1}. Shock speeds can decrease by as much as a factor of 2 between the Sun and Earth. Acceleration of protons up to ~10 MeV usually continues to 1 AU and beyond in most moderate events. Acceleration of 100 MeV–1 GeV protons occurs primarily near the Sun in most events but continues out to 1 AU in the largest events. In the large 19 October 1989 event shown in figure 1(b), particle intensities peak at the time of shock passage (on 20 October) even at energies of ~1 GeV.

Acceleration occurs at shock waves as particles are scattered back and forth across the shock surface, gaining an increment in speed on each transit. The upstream and downstream scattering centers serve as approaching 'walls' from which the particles are reflected. The distribution of particles streaming away from the shock is itself unstable to generation of Alfvén waves that add to the scattering of particles that follow behind. This nonlinear process serves to trap particles near the shock and increase the efficiency of acceleration. However, in shock acceleration, the energy of the particles comes from the kinetic energy of the shock, not just from the energy in the waves.

Ionization-state measurements have provided the clearest and most compelling evidence on the origin of the ions in the large gradual events. The ionization states are characteristic of a 1–2 MK plasma. Even C and O are not fully ionized and the mean ionization state of Fe, Q_{Fe} ~ 14. These ions could *not* come from the hot plasma of a flare; they represent ambient unheated coronal plasma. Ionization states of Fe have now been measured over an energy interval of 0.3–600 MeV amu^{-1} by 6 different experiments on 4 spacecraft. The highest-energy measurements were made in some of the largest events of the last solar cycle. Not only do these ions place an upper limit on the source temperature, but also the high-energy ions would be stripped of electrons in seconds at densities of 10^{10} cm^{-3} found in the low corona where flares occur. The ions must have been accelerated from ambient coronal plasma at low density. This conclusion is in agreement with observations that the peak intensities of GeV protons are produced when the leading edge of the CME is at 5–10 solar radii.

Table 1. Element abundances in energetic particle sources.

	Z	FIP	Photosphere	Gradual events (SEP corona)	CIR events (Coronal hole)	Impulsive flares	Anomalous CR	Galactic CR source
H	1	13.53	1.35×10^6	$(1.57\pm0.22) \times 10^6$	$(1.81\pm0.24) \times 10^6$	$\sim 1 \times 10^6$	~1000±500	$(1.74\pm0.43) \times 10^5$
He	2	24.46	132 000±11 000	57 000±3000	159 000±1000	46 000±4000	5000±1000	23 500±1600
C	6	11.22	479±55	465±9	890±36	434±30	<20	843±67
N	7	14.48	126±20	124±3	140±14	157±18	120±10	61±24
O	8	13.55	1000±161	1000±10	1000±37	1000±45	1000±10	1000±40
F	9	17.34	0.05±0.03	<0.1		<2	<2	<5
Ne	10	21.47	162±22	152±4	170±16	400±28	70±10	125±16
Na	11	5.12	2.9±0.2	10.4±1.1		34±8	<0.2	11±7
Mg	12	7.61	51±6	196±4	140±14	408±29	1.2±0.3	207±12
Al	13	5.96	4.0±0.6	15.7±1.6		68±12	<0.1	20±8
Si	14	8.12	48±5	152±4	100±12	352±27	1.7±0.3	196±12
P	15	10.9	0.38±0.04	0.65±0.17		4±3	<0.1	<5
S	16	10.3	29±7	31.8±0.7	50±8	117±15	0.6±0.2	25±4
Cl	17	12.95	0.4±0.3	0.24±0.1		<2	<0.1	<3
Ar	18	15.68	4.5±1.0	3.3±0.2		30±8	4.2±0.5	5.9±1.4
K	19	4.32	0.18±0.05	0.55±0.15		2±2		<3.7
Ca	20	6.09	3.09±0.14	10.6±0.4		88±13	<0.1	1.7±3.5
Ti	22	6.81	0.14±0.02	0.34±0.1		<2	<0.1	<4.7
Cr	24	6.74	0.63±0.04	2.1±0.3		12±5	<0.1	<5.7
Fe	26	7.83	42.7±3.9	134±4	97±11	1078±46	<0.9	182±11
Ni	28	7.61	2.4±0.05	6.4±0.6		42±9	<0.05	10±1.7
Zn	30	9.36	0.054±0.010	0.11±0.04		6±4	<0.05	0.13±0.02

FIP: first ionization potential.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

At energies of a few MeV amu^{-1}, most element abundances in gradual SEP events do not vary significantly from event to event and the average abundances seem to directly reflect the abundances of the source plasma in the solar corona. In fact, SEP abundances provide the most complete information we have on coronal abundances on a total of 22 elements (see table 1). Element abundances in the corona differ from those in the photosphere in a way that depends upon their *first* ionization potential (FIP), or perhaps upon their first ionization time scale. Elements with FIP < 10 eV (e.g. K, Na, Al, Mg, Si and Fe) have abundances that are enhanced by a factor of ~4 relative to high-FIP elements (e.g. C, N, O, Ne, Ar). Low-FIP elements are ionized in the photosphere, while high-FIP elements are neutral. An ion–neutral fractionation occurs because ions and neutral atoms are transported differently from the photosphere into the corona. Once elements reach the 1–2 MK corona they all become highly ionized long before they are accelerated.

A characteristic feature of gradual SEP events is their large and complex spatial structure that produces a wide variety of time profiles when viewed by spacecraft at different longitudes relative to the CME. Figure 2 shows time profiles of proton intensities seen by 3 spacecraft whose configuration relative to the CME is shown in the inset. As the spiral interplanetary field lines are swept over a spacecraft, the longitude of the magnetic connection point to the shock swings eastward with time. In this event, the 3 spacecraft become better connected to the intense acceleration region near the nose of the shock as time increases. HELIOS 1 is best connected. Here the intensity rises to a flat profile where the intensity is controlled by limits on the rate particles can stream away from the shock. A sharp peak in intensity is seen as the shock passes the spacecraft. The other two spacecraft see lower intensities around the flank of the shock; intensities increase as their connection point moves toward the nose of the shock. Behind the shock all 3 spacecraft enter an invariant-spectral region where particle intensities decrease slowly with time as the containment volume for the particles slowly expands.

Energy spectra of ions from the 20 October 1995 event are shown in figure 3(b). These spectra were obtained during the invariant spectral period in this relatively small event by the WIND spacecraft. The proton spectrum from the plateau region in the large event of 19 October 1989 is also shown in the figure for comparison. This event, shown in figure 1(b), was the largest event that occurred during the last solar maximum.

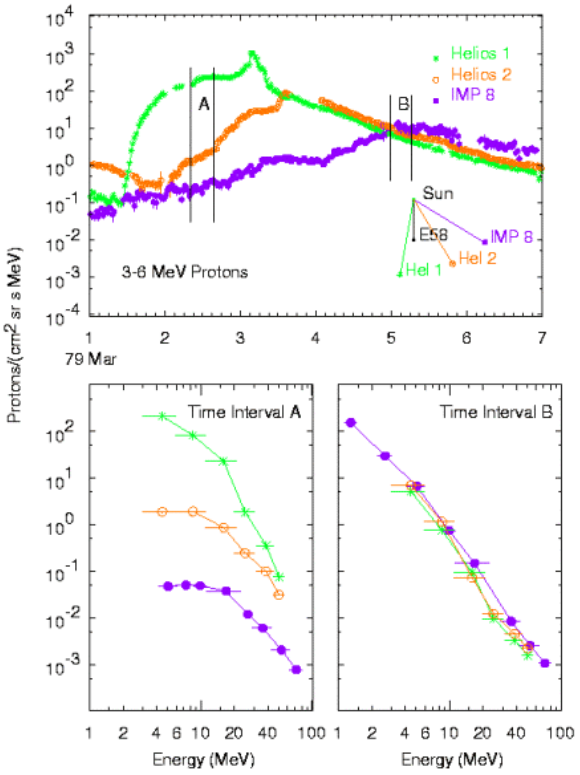


Figure 2. Intensity–time profiles of 3–6 MeV protons observed by 3 spacecraft are shown for a gradual event of modest extent. Spacecraft locations around the CME are shown in the inset. Lower panels show proton energy spectra taken early in the event (a) and during the 'invariant' spectral period (b).

Acceleration at the Earth's bow shock

The bow shock produced when the solar wind encounters the Earth's magnetosphere is also a source of accelerated particles. These were first observed as intense bursts of particles (and resonant waves) seen by the ISEE spacecraft in the region sunward of Earth. Since magnetic flux tubes contact the Earth's bow shock for a short time, the energies of the accelerated ions are usually below 100 keV amu^{-1}. Bursts were observed only during periods when the magnetic field connected the spacecraft to the bow shock. As the solar wind sweeps field lines past the Earth, they remain in contact with the bow shock for a limited period. For the nominal field direction, this contact point is usually swept across the sunward face of the shock from the dusk to the dawn side; the field direction becomes increasingly parallel to the shock normal. However, it was not possible in the early measurements to distinguish which elements comprised the 'ions', or to clearly establish that the particles were accelerated at the shock rather than simply leaking out from the Earth's radiation belts.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

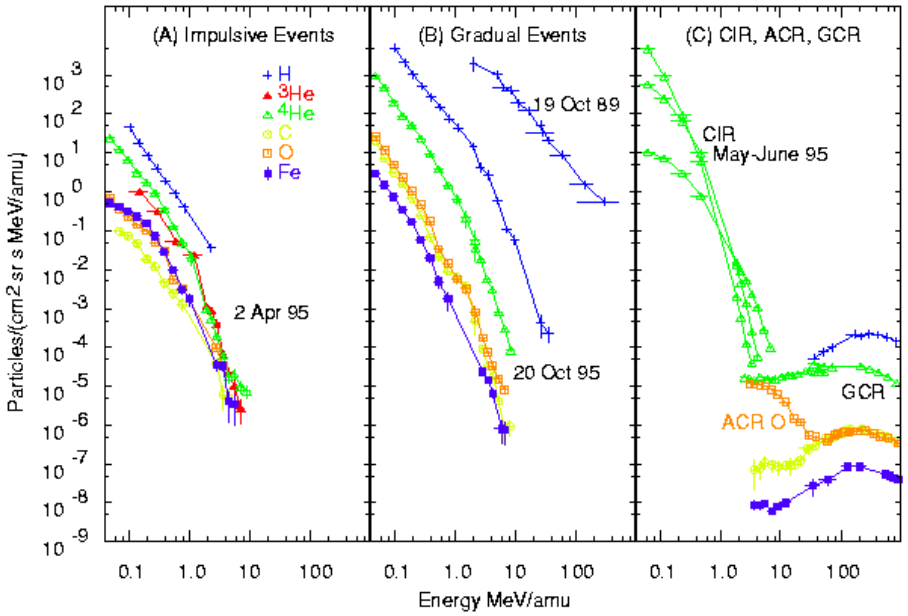


Figure 3. Energy spectra of various particle species indicated are shown for different particle populations. (a) Spectra for a small impulsive-flare event on 2 April 1995 observed on the WIND spacecraft are shown. (b) Spectra for various species in the small 20 October 1995 event observed on the WIND spacecraft are shown and a proton spectrum observed by NOAA/GOES during the large 19 October 1989 event (see figure 1(b)) is shown for comparison. (c) Spectra of He during the May–June 1995 CIR event (see figure 1(c)) show variation with time. The low-energy GCR spectra of various species during solar minimum are shown, and the low-energy rise in O from the presence of ACRs is noted on the figure.

The first measurements of element abundances in these 'upstream events' were made aboard the WIND spacecraft. It was found that the ions had the same element abundances as the high-speed solar wind and that events were much more common during high-speed solar-wind streams when the bow shock was strongest. In one case an upstream event even occurred during a small solar ^{3}He-rich event so that the intense burst of ions flowing upstream from the shock was also ^{3}He-rich. Clearly the source of the accelerated ions was these available particles from the upstream solar wind, not those from the magnetosphere.

Co-rotating interaction regions (CIRs)

CIRs are formed when high-speed solar-wind streams, emitted from coronal holes on the Sun, overtake the low-speed solar wind emitted earlier in the solar rotation. This interaction produces a pair of shocks: a forward shock propagates outward into the slow solar wind and a reverse shock propagates sunward into the high-speed solar wind. The shocks usually begin to form outside 1 AU and they continue to strengthen out to several AU. Particles are accelerated at both shocks, although the reverse shock seems to produce greater intensities and harder energy spectra. Thus particles from the reverse shock are most easily observed near Earth flowing sunward through the high-speed stream.

The time evolution of He intensities from a large CIR is shown in figure 1(c). Particles from the forward shock are seen on 29 May. As the Sun rotates, the Earth crosses the stream interface from the slow to the fast solar wind on 30 May. At first, low-energy particles predominate from the newly formed shock. Later, as we become connected to the strengthening shock farther away from us, high-energy particle intensities increase while those at low energies are attenuated by the greater transport distances. This behavior is the inverse of the velocity dispersion seen in both impulsive and gradual SEP events. For the unusual event of May–June 1995, particles were seen for 17 d while the sun rotated through an angle of ~225° and the reverse shock propagated across the high-speed stream and far into the following low-speed solar wind. The evolution of the energy spectra of He in this event is shown in figure 3(c).

Averaged element abundances of energetic particles from CIRs are included in table 1. These abundances are most similar to those of the high-speed solar wind that serves as the source. The enhancement of low-FIP elements relative to high is only a factor of ~2, indicating that abundances of the material from coronal holes suffers less fractionation than that in the corona and in active regions. However, departures from solar-wind abundances are also seen. Elevated abundances of H and He, for example, probably result from the acceleration of interstellar pickup ions (see below) although O, the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

talisman of this source, is not enhanced relative to C, N or Ne, for example.

Anomalous cosmic rays (ACRs)

ACRs are yet another shock-accelerated particle population with unique element abundances. In the interstellar material outside the heliosphere, elements with FIP less than the ionization potential of H, 13.6 eV, are ionized while those of higher FIP are neutral. Interstellar H absorbs the photons of higher energy. As the solar system moves through this gas, neutral atoms easily penetrate, but ions are largely excluded by the magnetic fields. As these interstellar neutrals approach the Sun they are ionized by solar photons or, in the case of H, by charge exchange with the solar wind. These new ions are suddenly able to sense the presence of the magnetic fields of the solar wind. Thus they are 'picked up' by the solar wind and they acquire a velocity distribution that extends to twice the solar-wind speed. The ions are eventually convected out to the solar-wind termination shock that is estimated to be formed ≥100 AU from the Sun. Here, the higher initial injection speed of the pickup ions leads to their preferential acceleration. The energetic ions are modulated as they propagate back to 1 AU where they are observed (see COSMIC RAYS).

The ACRs were first observed by the anomalously high intensities of O (relative to C, for example) at ~10 MeV amu^{-1} near solar minimum (see figure 3(c)). O, as well as He, N and Ne, persisted during solar minimum but were then modulated with the 11 yr solar cycle in a manner similar to the galactic cosmic rays. The theory of the singly ionized interstellar pickup ions and of their acceleration at the termination shock was advanced 20 years before pickup ions were first observed in the solar wind. Recently, ionization states of ACRs have also been directly measured, confirming that most are singly ionized while a small fraction have become more highly ionized during acceleration and transport. Approximate abundances of ACRs at the termination shock are included in table 1; these abundances depend upon assumptions made to correct for solar modulation.

Galactic cosmic rays (GCRs)

Cosmic rays, probably accelerated at SUPERNOVA shock waves in the Galaxy, have a constant presence throughout the heliosphere (see COSMIC RAYS: PROPAGATION IN THE HELIOSPHERE). GCRs are easily distinguished from other particle populations by their high energies and their unusual element abundances. As observed, GCRs contain unusually high abundances of Li, Be and B that are produced in nuclear reactions with interstellar matter during the ~10^7 yr lifetime of the GCRs. It is possible to correct for this matter traversal to derive GCR source abundances that are shown in table 1. The first suggestion of the existence of an ion–neutral fractionation process based upon FIP was made for the GCR source abundances by comparing them with the 'universal' abundances of elements derived from meteorites (essentially equivalent to solar photospheric abundances). Similarities between the abundances of GCRs and those of large gradual SEP events were noted many years ago, suggesting that stellar coronae or stellar energetic particles provide the seed population for the GCRs. Competing theories explain the enhancement in low-FIP elements in GCRs by the preferential acceleration of INTERSTELLAR GRAINS that are rich in those elements. However, we have seen that the process of separating ions and neutral atoms across electromagnetic fields, seen in the SEP, CIR and ACR abundances, is quite common in the heliosphere. It would be surprising if this process were confined to the heliosphere while an entirely different mechanism governed the GCR abundances.

The low-energy portion of the spectra of the dominant species in the GCRs is shown in figure 3(c). Above ~1 GeV amu^{-1} the differential intensity spectra attain an $E^{-2.6}$ dependence that continues to ~1 TeV amu^{-1} where the spectrum steepens to form a 'knee'. The low intensities of particles here and at higher energies can only be measured indirectly by large detector arrays at ground level. Hence, element abundances and their evolution as a function of energy have not been measured here.

Solar cycle variations

Different epochs in the 11 yr solar cycle bring an enormous change in the character of the energetic particles we see in the solar wind near Earth. The most favorable 'season' for observing ACRs and low-energy GCRs is at solar minimum when their intensities are highest. These particles are suppressed as the solar cycle increases. In fact, ACRs have never been observed during the years around solar maximum.

It is not surprising that the numbers of both impulsive and gradual SEP events increase as solar activity increases and more flares and CMEs occur. The number of impulsive-flare-related events varies from a few per year at solar minimum to a few hundred per year at solar maximum. The number of gradual events varies from about 5 yr^{-1} at minimum to 30 yr^{-1} at maximum. Of course these numbers depend upon the intensity threshold and the particle species and energy used to count events. However, the numbers are not the whole story. The events at solar minimum are also smaller and have softer energy spectra than those at solar maximum (see the proton spectra for the two gradual events in figure 3(b)). These are the products of less-energetic flares or of weaker shocks.

For CIR events, the variation is more complicated. Energetic particles from CIRs can be seen at most phases of the solar cycle. However, they are difficult to observe

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(or to count) during solar maximum because they are often obscured by SEP events and the CIRs themselves are disrupted by CMEs. Thus the events are best studied near solar minimum. However, at times *very* near solar minimum, such as during 1996, the heliospheric current sheet becomes very flat and lies near the solar equatorial plane. Thus, the polar high-speed streams and the equatorial low-speed streams can flow out radially with no regions of interaction. During these times there are few CIRs formed and few particles accelerated.

Bibliography

Fisk L, Kozlovsky B and Ramaty R 1974 An interpretation of the observed oxygen and nitrogen enhancements in the low energy cosmic rays *Astrophys. J.* **190** L35

Gosling J T 1993 The solar flare myth *J. Geophys. Res.* **98** 18 949

Reames D V 1990 Energetic particles from impulsive solar flares *Astrophys. J. Suppl.* **73** 235

Reames D V 1995 Coronal abundances determined from energetic particles *Adv. Space Res.* **15** (7) 41

Reames D V 1999 Particle acceleration at the sun and in the heliosphere *Space Science Reviews* **90** 413. See http://kapis1.wkap.nl/oasis.htm/231960

Donald V Reames

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Global Properties

The most fundamental problem in solar system research is still unsolved: how can the Sun with a surface temperature of only 5800 K heat up its atmosphere to more than a million K? In fact, the solar atmosphere is so hot that not even the Sun's enormous gravity can contain it. Part of it is continuously evaporating into interplanetary space: the SOLAR WIND. As a highly ionized magnetized plasma it dominates a huge volume around the Sun which is called the heliosphere. Only far beyond the outermost planets is a transition into the interstellar gas flowing through our Galaxy expected to occur.

On the rare occasions of solar eclipses, every observer is stunned by the high degree of structure in the then uncovered CORONA. No wonder that the solar wind emerging from there is similarly inhomogeneous and creates a complicated three-dimensional shape of the plasma heliosphere. Interactions between outflowing streams of different speeds and solar transient phenomena cause further complications. Thus, the solar wind as we see it from the Earth's orbit is characterized by an enormous variability in all its basic properties. It is this very variability that allows the solar wind to have a surprisingly large impact on the Earth. Some effects even make it down to the ground!

The solar wind proves to be one key link between the solar atmosphere and the Earth. Although the energy transferred by the solar wind is miniscule compared to both sunlight and those energies involved in the Earth's atmosphere, the solar wind is capable of pin-pricking the Earth system until it eventually reacts in a highly nonlinear way.

In this article the chief characteristics of the solar wind in context with its sources in the corona will be presented. The next section first describes the basic properties of the solar wind as observed in the comparatively simple situation around solar activity minimum. Next the corona and inner heliosphere are discussed in their full 3D extent. The fate of the stream structure as it evolves with increasing distance from the Sun is then dealt with. Finally we summarize the current knowledge of the different types of solar wind, ending with the admission that none of them is really explained yet.

The solar wind at 1 AU

Basic properties

The existence of the solar wind was theoretically modeled by E Parker in 1958 and experimentally verified in 1962. Since then, it has been observed throughout wide parts of the heliosphere, as close as 0.29 AU and as far as (by now) 70 AU from the Sun, mainly in the ecliptic plane, but also above both solar poles. Some typical parameters as measured at the Earth's orbit (i.e. at 1 AU distance from the Sun) are given in table 1.

Table 1. Typical parameters of the slow solar wind at 1 AU.

Flow speed v_p	350 km s^{-1}
Proton density n_p	9 cm^{-3}
Flux density $n_p v_p$	3×10^8 cm^{-2} s^{-1}
Composition	96% protons, 4% He^+ ions, minor constituents, plus an adequate number of electrons to maintain nearly perfect charge neutrality
Proton temperature T_p	4×10^4 K
Electron temperature T_e	1.5×10^5 K
Magnetic field B	4 nT

The solar wind flow speed is usually much higher than the local sound and Alfvén speeds, and typical Mach numbers are around 10. This implies that the plasma dynamic pressure is much higher than both the magnetic pressure and the thermal pressure. It also means that the magnetic field is frozen-in in the expanding flow. The field lines can be regarded as 'stream lines' of the flow. Generally, they maintain their identity on their way through the entire heliosphere since, due to the near absence of collisions in the tenuous plasma, the particles can hardly ever leave their original field lines. All particles move, on the average, radially away from the Sun. Therefore, the stream lines interconnecting particles emerging from the same source on the rotating Sun are curved like Archimedean spirals. The curvature is determined by both the flow speed and the distance from the Sun. The average Parker angle between the stream lines and the radial direction to the Sun at 1 AU amounts to about 45°. With increasing distance to the Sun the Parker spiral winds up further and further: from Jupiter's orbit in the direction of the local magnetic field can be considered almost perpendicular to the solar wind flow. Actually, the plasma keeps moving radially, in analogy to the needle in the spiral grooves of an old-fashioned record.

Solar wind stream structure

The solar wind exhibits a remarkable variability, even at times of low solar activity. As a typical example, figure 1 shows the main solar wind parameters obtained near the Earth's orbit during a whole solar rotation. They were recorded in early 1974, i.e. close to a solar activity minimum, and the basic pattern seen then remained almost unchanged through several months:

- For two intervals of some 60° in longitude (corresponding to a durations of several days) the flow speed exceeds 600 km s^{-1}.
- In both these high-speed streams the plasma density is lowest and the proton temperature is highest (while the electrons are somewhat cooler), just opposite to the low-speed regions in between.
- Both high-speed streams are regions of unipolar magnetic fields (indicated by the + and − signs),
- In a short interval in the middle of the figure, interplanetary shock waves went by, characterized by abrupt increases in speed, density, temperature and magnetic field strength (not shown here). This type of shock wave is caused by transient activity on the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

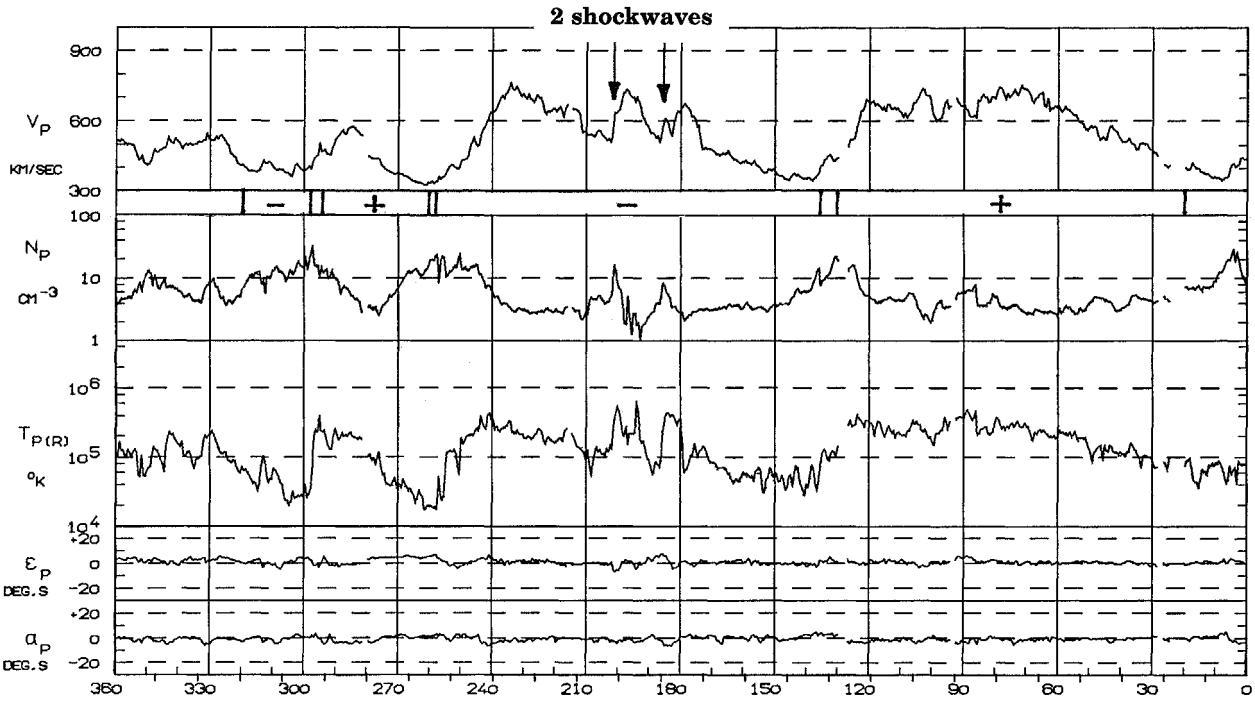


Figure 1. The solar wind proton parameters (flow speed, density, temperature and flow angles relative to the radial direction) during a full solar rotation in early 1975, as observed by the Helios 1 solar probe close to 1 AU. Typical for this phase of minimum solar activity are the two large high-speed streams with the extended compression regions in front. Note also the two transient-related shock waves.

Sun, e.g. SOLAR CORONAL MASS EJECTIONS (CMES). Usually, not more than one such event occurs per solar rotation at solar minimum.

- The left part of the figure shows a less well-ordered pattern of flow speeds and polarities.
- The flow angles ε_P (out of the ecliptic) and α_P (azimuthal) fluctuate within about 10° around the radial direction. More pronounced deflections are observed in compression regions ahead of high-speed streams and at shock waves.

Observations as shown in figure 1 revealed that solar wind variability cannot be attributed to just transient disturbances of an otherwise quiet or structureless solar wind. In particular the persistent occurrence of high-speed streams with flow speeds up to 800 km s^{-1} was realized to be an intrinsic feature of the solar wind phenomenon. With decreasing solar activity in the years 1973–75, these high-speed streams developed into stable large-scale structures. Some of them were observed at the same solar longitude for several months, i.e. they were stationarily corotating with the Sun through several solar rotations.

In the Skylab era the coronal sources of the high-speed streams could be identified: these are regions of diminished brightness at soft x-ray and EUV wavelengths (so-called CORONAL HOLES), indicating significantly lower electron densities and temperatures. The boundaries are remarkably sharp and coincide in detail with those of the high-speed streams properly mapped back to the Sun. Coronal holes mark the regions in which coronal magnetic field lines open freely into interplanetary space (see articles by Kohl and de Forest). At times of low solar activity they cover mainly both polar caps. In contrast, closed loop-like structures are always related with active centers such as SUNSPOTS which most often are located at lower heliographic latitudes. Coronal holes and the high-speed streams emerging from them have to be regarded as features of the quiet Sun, not the active Sun.

The solar wind in three dimensions

The 3D heliosphere in terms of the 'ballerina model'

On the basis of these early discoveries, a three-dimensional model of the heliosphere and the stream-structured solar wind emerged. It is most adequately visualized in terms of the ballerina model first proposed by H Alfvén in 1977.

Figure 2 is an artist's view of the inner heliosphere as it may appear immediately before a typical solar activity minimum, e.g. in 1975. We find the Sun's poles to be covered by large coronal holes. They are areas of open magnetic field lines, the northern hole being of positive (outward directed) polarity, the southern hole being negative. One solar cycle later the polarity reverses, of course. Some tongue-like extensions of the coronal holes reach well into the equatorial regions. The Sun's equatorial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

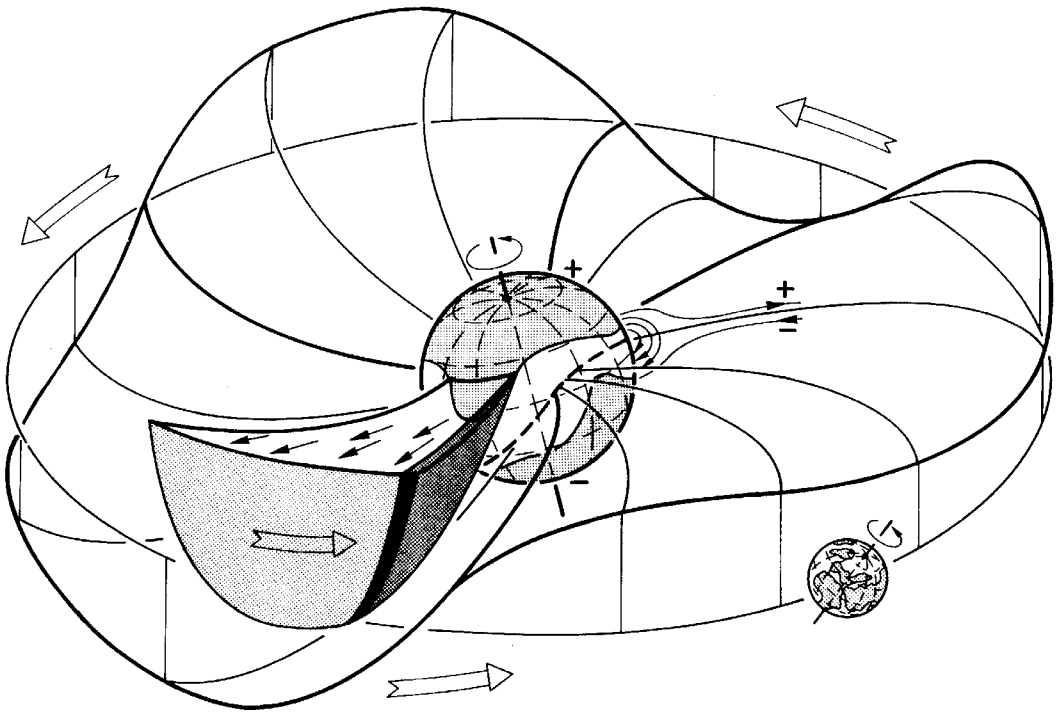


Figure 2. The 'ballerina model' of the 3D heliosphere, according to H Alfvén.

region is governed by bright active centers (including some sunspots left over from the past cycle) and their loop-like and mainly closed magnetic structures.

What looks like the skirt of a spinning ballerina is the warped separatrix between positive and negative solar magnetic field lines dragged out into interplanetary space by the radially outflowing solar wind plasma. This heliospheric current sheet is formed on top of the closed magnetic structures at the transition between closed and open flux tubes, i.e. generally in the middle of the near-equatorial 'belt' of activity. If the spinning skirt passes an observer sitting, say, at the earth, he would notice a polarity switch and call it a crossing of a magnetic sector boundary. The size and number of magnetic sectors is closely related to the structure of the underlying corona. i.e. the shape of the activity belt and the coronal holes, respectively (for more details see the article SOLAR WIND: MAGNETIC FIELD).

The heliospheric current sheet: shape and origin

The coronal holes are the sources of high-speed solar wind. The emission of slow solar wind is sharply confined to a belt centered at the warped current sheet of about 40° width in latitude. The warps of both the current sheet (which can be taken to be the heliomagnetic equator) and the coronal hole boundaries with respect to the heliographic equator allow some high-speed streams to extend to low latitudes so that they become observable at times even in the plane of the ecliptic. This occurs preferentially in the 2 years before activity minimum, when the large-scale coronal structure is rather stable, and high-speed streams reappear at the same heliographic longitudes for many solar rotations.

At times of minimum solar activity there are almost no warps left in the current sheet which is then planar like a disk lying very close to the plane of the heliographic equator. This is demonstrated in figure 3 where two images registered by the LASCO coronagraphs on board SOHO in early 1996 were merged. They give a very typical example of the appearance of the extended corona at the most recent activity minimum. The inner part was taken in the light of the green corona line at 530.3 nm produced by Fe XIV ions at temperatures of 2×10^6 K. This spectral line is particularly well-suited to outline hot magnetic structures in the inner corona. The outer part taken in white light shows the larger-scale electron density distribution.

Images of this type taken during the months around solar activity minimum have revealed that:

- the appearance is nearly symmetric about the solar axis, with a flat high density sheet right in the equatorial plane,
- the green (inner) patterns merge well into the white (outer) patterns, indicating that hot magnetic loops are closely associated with high density streamers on top,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

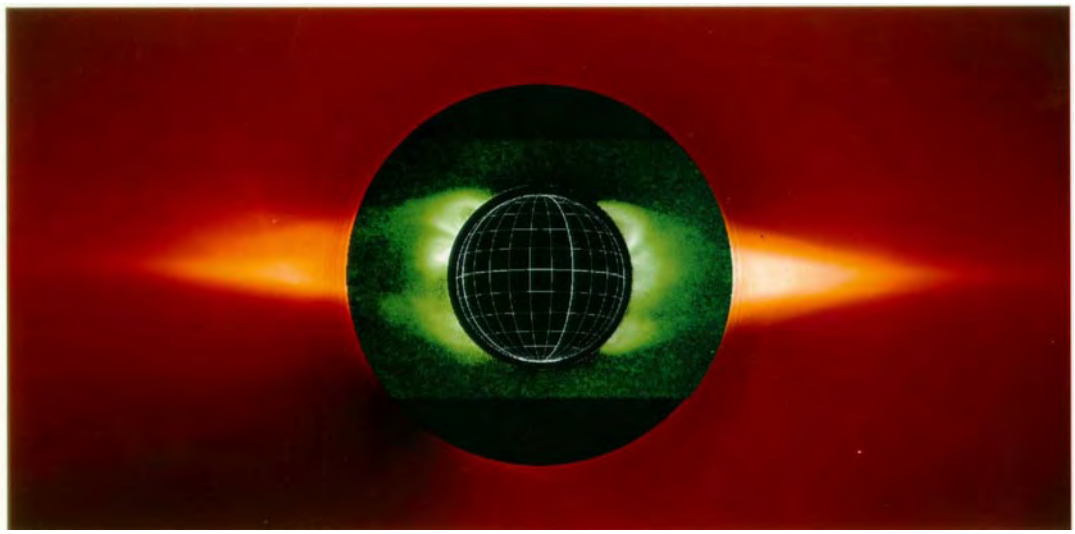


Figure 3. A coronagraph view of the extended minimum corona (on 1st February 1996), composed by a green-line emission image and a white-light image taken by the LASCO coronagraphs on-board SOHO. **This figure is reproduced as Color Plate 54.**

- the dense sheet emerges from bright apparently closed loop systems (called helmet streamers) centered at latitudes of 30–45° in both hemispheres; the loop tops usually reach out to about 1.5 R_s, in some cases to well beyond 2 R_s; their helmet-like outer extensions are clearly bent towards the solar equator,
- around the equator, there is a more diffuse bright pattern, clearly separated from the high latitude loops and strongly varying with time,
- there is no detectable green-line emission above either of the poles, confirming that the density and temperature in polar coronal holes is rather low; their edges appear to be sharp and well-defined and remain stable on time scales of several days.

The structure of the corona is governed by the solar magnetic field. The magnetic topology is not just that of a simple magnetic dipole, not even at the comparatively quiet minimum phase. Higher-order multipole components are always involved. A coronal image like figure 3 leads to the immediate impression that there are several magnetic loop systems anchored at the Sun. As a consequence, there must be a series of polarity changes around the limb, apart from the global polarity switch from 'positive' at the northern coronal hole to 'negative' in the south. The radial extent of these multipole moments is much more limited than that of the overall magnetic dipole centered in the polar coronal holes.

Comparison of coronal images like figure 3 with photospheric magnetograms shows that the mid-latitude loop systems are located well on top of magnetic neutral lines which often reach all around the visible half of the Sun. This is also the location of polar crown filaments well known from H-alpha observations. The near-equatorial magnetic features are more variable, both with space and time, due to the dynamics of some remaining active regions. They cause the green-line corona to vary in intensity by a factor of 10 and more. This variability in conjunction with the smaller scale and often inclined magnetic structures may be the reason for the more diffuse appearance in the equatorial regions.

The formation of antiparallel magnetic fields beyond the closed loops would require the existence of current sheets. Since large streamers coexist at different latitudes, multiple current sheets must also exist, unless they merge somewhere. They form a streamer sheet of initially 1 solar diameter width. Processes still to be explored may eventually lead to the formation of the large-scale interplanetary heliospheric current sheet imbedded in a flow of slow solar wind. Note that the transition from closed to open magnetic topology at the top of a streamer is generally not yet understood, nor is the release of slow solar wind from these regions.

The flatness and stability of the distant heliospheric current sheet(s) is probably due to the persistence of the mid-latitude streamers. In other words, the current sheet is determined by mid-latitude phenomena rather than by streamers above the near-equatorial activity belt. The mid-latitude streamers are apparently bent towards the equator by the more than radially expanding polar coronal hole magnetic field and high-speed plasma flow. After all, that would mean that the heliospheric current sheet close to the Sun's equator is being shaped by forces originating in the Sun's polar regions!

More information can be found in the articles CORONAL LOOPS, CORONAL STREAMERS, and FILAMENTS.

Solar wind stream boundaries

Coronal holes attract attention not only by their surprisingly sharp boundaries. The high-speed solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

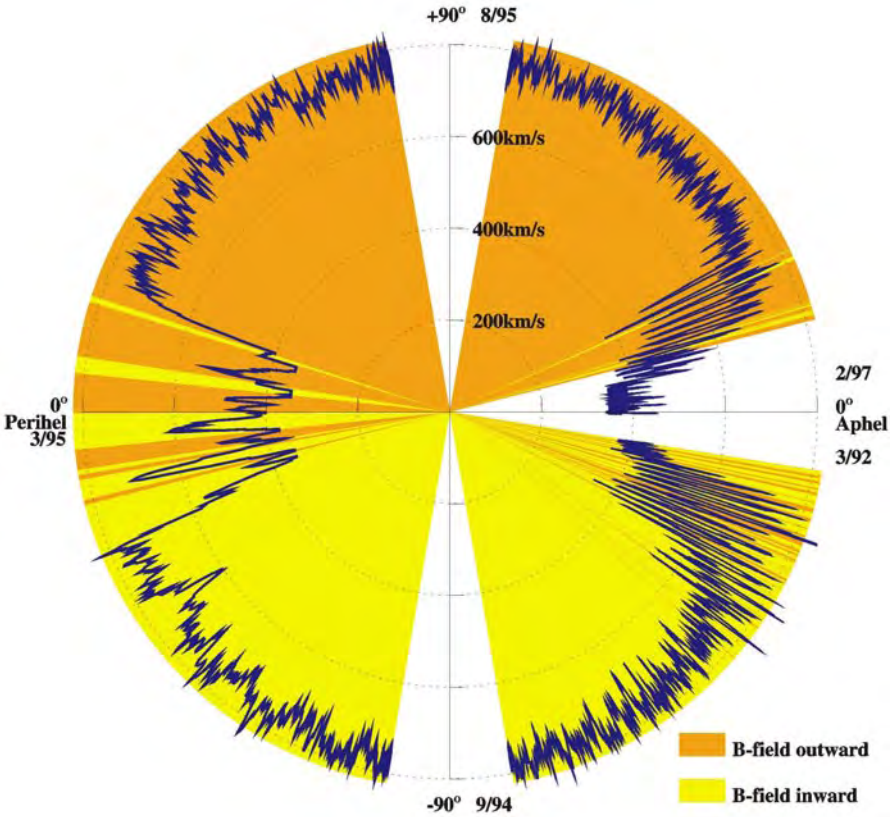


Figure 4. Solar wind and magnetic field polarity observations during the passage of Ulysses across both solar poles. (Courtesy of J Woch, Max-Planck-Institute für Aeronomie.)

wind streams emerging from them are also very distinctly separated from adjacent slow flow. The boundary layer between high-speed wind at 600 km s^{-1} and slow wind at 300 km s^{-1} is often found to be less than 1.5° with respect to heliographic latitude! In longitude, mutual interactions between flows of different speeds tend to smooth the originally steep profiles with increasing distance to the Sun. Only when the Helios solar probes approached the Sun as close as 0.29 AU were similarly thin boundary longitudinal boundary layers confirmed and a 'mesa-like' shape of high-speed streams was recognized. In high-speed wind, remnant signatures of the coronal network could be identified, proving its originally filamentary nature. Close to the Sun, all the transition from slow to fast wind was observed to occur apparently as the transition from one flux tube to the next! These discoveries led to the conclusion that high- and slow-speed wind are very different in nature, emanate from well-separated sources in the corona, and are due to basically different acceleration mechanisms.

Figure 4 is an impressive experimental confirmation of what has been said in this subsection. This unusual polar plot shows the solar wind speed as measured from the Ulysses space probe during its passage across both solar poles in the years 1992–97. The magnetic polarity is also indicated. Clearly, high-speed wind is encountered predominantly at latitudes beyond 20° in either hemisphere. During the fast passage through the ecliptic plane (left part of figure 4) abrupt changes from fast to slow and wind and back occurred. In contrast, the rather slow orbital motion of Ulysses away from and towards the ecliptic (right part) allowed mid-latitude fast streams and interstream slow flow to be observed alternately for many solar rotations. Beyond about 35° no slow wind and no 'wrong' sector polarity were found any more. For more details about Ulysses see the article SOLAR WIND: ULYSSES.

Further signatures of stream structure

There are some more signatures indicating the basically different nature of high-speed wind from coronal holes and low-speed flow in the streamer belt. They are mainly associated with constituents other than protons and their ionization states (for details see the article SOLAR WIND: COMPOSITION).

- The helium abundance is 3.6% in high-speed wind, very constant in time and almost identical for all streams. In slow wind the abundance is only 2.5% and highly variable. This is one strong indication that high-speed wind is released at lower altitudes in a gravitationally stratified corona.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- The ionization state of heavy ions is significantly lower in fast flow, indicating a lower coronal temperature in its source region.
- In the slow solar wind there is a much more pronounced first ionization potential (FIP) effect, i.e. the elements with a FIP of less than 10 eV are significantly overabundant in comparison to photospheric values (see the article by von Steiger).
- The angular momentum carried away by the solar wind from the rotating Sun is almost completely contained in the slow flow. This is another strong indication that the fast wind starts from rather close to the solar rotation axis, while the slow wind is released only beyond 30 R_s.

On the other hand, there are also some stunning similarities between the two types of solar wind:

- The momentum flux density $n_p\, v_p^2\, m_p$ does not differ by more than 5%. Using this quantity alone no stream structure at all would be discernible! That means that both types of wind put identical dynamic pressure on any obstacle, e.g. planetary magnetospheres or the pouring up interstellar gas, implying important consequences for the 3D shape of the heliosphere as a whole.
- Similarly, the total energy flux density summoned by the Sun in order to release the solar wind is equal all around the Sun. Of course, the fast flow carries more kinetic energy $n_p\, v_p^3\, m_p$, with it, but the Sun has to bring up much more energy $n_p\, v_p\, Gm_p\, M_S/R_s$, in order to lift the high density slow wind out of its gravitational field, and the sum of both remains constant.

There is no conclusive answer yet as to whether these invariances happen by chance, whether they are intrinsic solar phenomena, or are the result of global rearrangement of the different flows beyond the corona. The solution of this problem might hold a key to the understanding of the corona and the generation of the solar wind .

Solar wind streams on their way out into the heliosphere

Solar wind streams and their interactions

A fast stream may follow a slow one if the coronal hole boundary is inclined by some angle to the Sun's rotation axis which is almost perpendicular to the equatorial plane. These two types of plasmas flow side by side in the first place, but start interacting with each other with increasing distance from the Sun. The Parker spiral of a fast plasma stream is less tightly wound up and thus pushes the slow plasma with the more strongly curved field lines. Compression and deflection of the plasma on both sides of the stream interface finally yields the typical profiles found at 1 AU as shown in figure 1. The transition from slow to fast flow in these corotating interaction regions (CIRS) now extends over some 10–30° in longitude (see SOLAR WIND: COROTATING INTERACTION REGIONS). In many cases, the stream interface survives as a tangential discontinuity which allows us to still recognize what was originally fast and slow flow. The total range filled with plasma affected (i.e. compressed and deflected) by these interactions amounts to some 30° at 1 AU. However, the plasma packed into that range stems from coronal source regions originally spanning some 90° in longitude! Generally, at about 1 AU distance from the Sun all solar wind in the ecliptic plane will have undergone some processing due to stream–stream interactions. In other words: no longer will there be any solar plasma in its original state, and primordial coronal imprints on the plasma will be lost by then. For this reason, one may draw a borderline between the inner and the outer heliosphere near 1 AU.

At the backside of fast streams no interaction of the flows occurs and no CIRs develops since here the different Parker angles lead to a separation of the flows rather than a compression. The stream boundaries are found to extend over some 60° at 1 AU (see, e.g., the stream in figure 1 on the right-hand side). However, mapping back the flow to the Sun assuming a strictly radial flow at the locally measured speed, the original rectangular profile at the Sun is nicely reconstructed!

At even larger distances from the Sun, the compression waves at the CIRs steepen to finally form corotating shock waves. There are fast forward shocks at the frontside, i.e. traveling to the slow wind side, and fast reverse shocks traveling seemingly backward, i.e. into the fast wind coming from behind. Thus, the originally steep stream profiles are further eroded (see also the article SOLAR WIND: COROTATING INTERACTION REGIONS). The corotating shocks at CIRs, similar to transient-related interplanetary shock waves, can accelerate ionized particles to considerable nonthermal energies, as explained in the article SOLAR WIND: ENERGETIC PARTICLES.

Beyond several AU, the CIRs themselves interact and merge by a variety of processes to form merged interaction regions (MIRS). They may be purely corotating MIRs (CMIRs), or include ejecta and shocks from CMEs to form global MIRs (GMIRs). After all, there is an increasing loss of detail of the solar source conditions with further increasing distance from the Sun, such that even the signatures of the solar rotation are lost beyond about 40 AU. At much larger distances, by interaction with the local interstellar medium the solar wind undergoes the termination shock of the heliosphere where it is diverted and decelerated. This region of subsonic flow, the heliosheath, is separated from the impinging interstellar plasma at the heliopause. Inside, the solar wind flow is diverted into the heliotail which is many thousands of AU long. On the upstream side, outside the heliopause, a bow shock may form to divert the interstellar material around the heliospheric obstacle. For details see the article SOLAR WIND: INTERACTION WITH THE LOCAL INTERSTELLAR MEDIUM.

Radial variations of average solar wind parameters

The radial profiles of the basic solar wind plasma parameters are of special importance in that they narrow

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

down the degrees of freedom available for modeling solar wind expansion. In-situ measurements as close as 0.29 AU (i.e. 60 R_s from Sun center) have recently been complemented by optical observations from close to the Sun's limb out to 32 R_s. On the other hand, the Pioneer and Voyager spaceprobes are currently expanding our knowledge of heliospheric signatures to beyond 70 AU. In the inner heliosphere, the local properties of the flow and its radial expansion are basically determined by the coronal sources and their topology. From about 1 AU distance on, stream-stream interactions and local plasma processes begin playing a dominant role in the solar wind's radial evolution.

- The flow speed profile of the slow wind close to the sun has been inferred from the coronagraphs on SOHO: small density blobs could be traced which are thought to float along with the flow. It begins to accelerate from about 3 R_s only, and by 10 R_s it is barely faster than 250 km s^{-1}. This profile is consistent with an isothermal expansion at a temperature of about 1.1 MK and a sonic point near 5 R_s, in good agreement with the simple Parker model. There is a further speed increase of about 50 km s^{-1} until about 1 AU.
- The UVCS instrument on SOHO by applying the Doppler dimming technique was able to determine the fast wind acceleration profile. This type of flow starts in the very low corona and attains a speed of 300 km s^{-1} at 3 R_s and 800 km s^{-1} at 10 R_s. From there on it travels at constant speed. Further details of solar wind acceleration and expansion in the sun's vicinity are found in SOLAR WIND: CORONAL ORIGINS.
- In the outer heliosphere, the speed differences between fast and slow flows are wiped out, and the heliospheric plasma expands at an average constant speed of 400 to 500 km s^{-1}. A slowdown of the solar wind by action of the interstellar gas has not yet been noted.
- The density is consistently found to drop off as R^{-2} as would be expected for a medium which expands spherically at constant speed. However, inside 1 AU the average gradient is $R^{-2.1}$, according to the mentioned speed increase of the slow wind. Also, there is a slight relative over-expansion of the slow wind (by some 10%), at the expense of a commensurate fast wind compression.
- The proton flux density goes most accurately as R^{-2}, on the grand average, from at least 0.3 AU on outward. However, inside 1 AU, the mass flux in fast streams grows by some 15% and decreases accordingly in slow wind. The conclusion is that flux tubes carrying the fast, low-density wind are compressed by some 15% in cross section by the flux tubes carrying the slow, high-density plasma.
- UVCS measured strikingly large line widths of O^{5+} lines in fast flow from polar coronal holes. If interpreted in terms of kinetic ion temperatures, values of more than 100 MK would be obtained! Also, there is an extreme temperature anisotropy, i.e. the temperature perpendicular to the magnetic field directions is much higher than that parallel to it. These findings support the concept of ion cyclotron heating in this type of solar wind inferred from in-situ observations at 0.3 AU from the Helios solar probes
- The proton temperature in slow wind drops with distance from the sun as $R^{-1.33}$. That is exactly what would be expected for an adiabatic expansion. Extrapolation back to the corona leads right to the observed coronal temperatures of about 2×10^{-6} K. That means that this type of solar wind escapes wave heating by originating from sources with closed magnetic fields without intense MHD-wave emission. In contrast, the fast wind cools off as $R^{-0.78}$, thus indicating substantial additional heat input beyond the corona due to heat conduction and local dissipation of convected wave energy. In the range between 1 and 20 AU temperature variations of $R^{-0.5}$ to $R^{-0.7}$ were reported. Apparently, that is a result of compression heating caused by stream-stream interactions.
- The electron temperature falls off, by and large, significantly slower than expected for adiabatic and isotropic expansion. For details the reader is referred to SOLAR WIND: KINETIC PROPERTIES and the specialized literature.

Signatures of what is affecting the earth system

Generally, the solar wind flow is diverted around the Earth by its magnetosphere which is maintained by the Earth's intrinsic magnetic field (see MAGNETOSPHERE OF EARTH). Solar wind particles cannot enter, unless there occurs a MAGNETIC RECONNECTION of interplanetary and planetary field lines. That may happen if the northward pointing Earth field on the front of the magnetosphere is hit by solar wind carrying a southward pointing interplanetary field. In such cases, significant geomagnetic disturbances of various kinds are initiated (see, e.g. the article SOLAR–TERRESTRIAL CONNECTION: COUPLING BETWEEN SOLAR WIND, MAGNETOSPHERE, IONOSPHERE, AND NEUTRAL ATMOSPHERE, MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS and others).

In the present context it is important to discuss by which means the usually radially pointing interplanetary field can be tilted in such an abnormal way.

- Compression and deflection of the solar wind flow in CIRs has important consequences for its magnetic field, too. Naturally, it undergoes the same compression and participates in the deflection process. Thus, there may arise enhanced out-of the-ecliptic field components, particularly in the vicinity of magnetic sector boundaries which often happen to be embedded in CIRs at 1 AU. The action of potential magnetic reconnection is enhanced by the pressure pulse from the compressed plasma.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- High-speed streams are also characterized by the occurrence of large amplitude transverse Alfvén fluctuations. They are easily recognizable by large excursions in the directions of both the interplanetary magnetic field (including geo-effective southward excursions) and the plasma flow. The field magnitude and the plasma density remain about constant. Typical fluctuation periods at 1 AU range from fractions of a minute to a few hours. Probably, these waves are remnants from coronal activity since they travel outward only, without evidence for local damping (for details see the chapters by Roberts and Gurnett). There is a close correspondence between the southward pointing of the interplanetary magnetic field in the course of Alfvénic fluctuations and geomagnetic activity, leading to clear association between medium-level geomagnetic activity and high-speed solar wind.

The recurrence of this particular type of geomagnetic activity every 27 d, i.e. exactly in the rhythm of solar rotation, had led J Bartels to postulate the existence of M regions on the Sun already in the 1930s. He thought they were long-lived stable regions on the Sun which emit certain particles capable of stirring geomagnetism. After all, he was strikingly right except for one aspect: these M regions are not to be sought in active regions on the Sun, as he thought, but rather in the inactive parts: the M regions are to be associated with the coronal holes representative of the inactive sun, and the geomagnetism is stirred by the streams of high-speed plasma emanating from them.

The major geomagnetic storms are also due to large southward excursions of the interplanetary field, but these are caused by transient processes on the Sun, e.g. coronal mass ejections and their interplanetary counterparts. The interested reader is referred to the articles on CMEs and geomagnetic effects, and SOLAR–TERRESTRIAL CONNECTION: SPACE WEATHER PREDICTIONS.

Slow variations during the solar activity cycle

What does the interplanetary medium sense from changes in solar activity during the solar cycle? That can best be described using the ballerina model (figure 2). It shows the slightly warped heliospheric current sheet separating inward and outward going magnetic field lines around activity minimum. Starting from this state the warps in the skirt begin to grow with increasing activity. They finally break up and turn over at activity maximum. Then, the current sheet begins reorganizing itself according to the overall magnetic field of the Sun which has just undergone a global polarity reversal. The warping diminishes, and at the following minimum the ballerina skirt is flat again; just the magnetic polarity is reversed. Thus, a full magnetic cycle of the Sun (called the Hale cycle) takes 22 yr, i.e. two activity cycles. This is described in more detail in the article SOLAR WIND: MAGNETIC FIELD.

According to the structural changes of the corona, the solar wind also changes. The remarkably stable shapes of both coronal holes and solar wind high-speed streams in the years before minimum activity are replaced by increasingly irregular patterns. The few remaining fast streams are smaller in size and amplitude, and they come from small, isolated, transient coronal holes.

Almost all around the Sun now emits slow wind. That applies to the poles as well, as indirect measurements using interplanetary scintillations of radio pulsar signals have shown. It is hard to conceive that this can be the same type of slow flow as is found in the streamer sheet at solar minimum. This 'maximum type' slow solar wind is found to emerge from substantially larger areas often located far from any current sheet. It is highly variable and usually contains some 4% of helium, thus indicating its release at lower altitudes in the corona than the 'minimum type' slow wind..

The whole structure is additionally confused by abrupt disturbances due to interplanetary shock waves driven by CMEs. (see articles on CMEs and SOLAR WIND SHOCK WAVES AND DISCONTINUITIES). Naturally, the occurrence rate of big solar transients follows solar activity very closely, and one would expect to find considerable variations of the average solar wind's properties going in parallel. It turns out, though, that this is actually not the case, at least not for the plane of the ecliptic where most long-term observations were made so far. There is not much change: the average speed drops from some 500 km s^{-1} at minimum to 400 km s^{-1} at maximum. Other quantities such as particle densities, and the fluxes of mass, momentum, and energy are modulated by not more than about 20%. It is interesting to note that that they all reach their minima at activity maximum, despite the many CME and shock events at that time!

Outside the ecliptic plane and especially at very high solar latitudes the solar cycle modulation can be expected to be completely different. At solar minimum there is only fast wind characterized by its high speed and very low density. With increasing activity and the associated disappearance of the polar coronal holes the share of slow type solar wind grows, with its high values of particle and mass flux densities. Thus, the modulation of the average solar wind properties at high latitudes is expected to be much more pronounced than in the ecliptic plane. We will certainly learn more details once the spaceprobe Ulysses has performed its second run across the Sun's poles.

Summary

In conclusion, it can be stated that the ballerina model (see figure 2) as conceived in the mid-1970s describes the 3D heliosphere at activity minimum in connection with the solar corona fairly well.

It is now well established that there are two basic types of quasi-steady solar wind which differ markedly by their main properties, the location and magnetic topology of their sources in the corona, and probably in their acceleration mechanism. Therefore, in table 2 the more specific values are given separately for both types:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Average solar wind parameters at 1 AU, for the time around solar activity minimum.

	Slow wind	Fast wind
Flow speed v_P	250–400 km s^{-1}	400–800 km s^{-1}
Proton density n_P	10.7 cm^{-3}	3.0 cm^{-3}
Proton flux density $n_P v_P$	3.7×10^8 cm^{-2} s^{-1}	2.0×10^8 cm^{-2} s^{-1}
Proton temperature T_P	3.4×10^4 K	2.3×10^5 K
Electron temperature T_e	1.3×10^5 K	1×10^5 K
Momentum flux density	2.12×10^8 dyn cm^{-2}	2.26×10^8 dyn cm^{-2}
Total energy flux density	1.55 erg cm^{-2} s^{-1}	1.43 erg cm^{-2} s^{-1}
Helium content	2.5%, variable	3.6%, stationary
Sources	Streamer belt	Coronal holes

Some further characteristics of different solar wind types at solar activity minimum are:

- The *fast solar wind* emerges from magnetically open coronal holes which are representative of the inactive Sun, i.e. the 'quiet' Sun. Consistently, the 'quiet' type solar wind is found in high-speed streams. Although the major coronal holes usually cover the polar caps at latitudes beyond 40–60°, the solar wind emerging there over-expands significantly and fills up all heliosphere except for the 40° wide streamer belt close to the heliomagnetic equator.
- The *slow solar wind of the minimum type* originates from above the more active regions on the Sun. It is constrained to the warped streamer belt and comprises the large-scale heliospheric current sheet. The low helium content in this type of solar wind indicates a larger release height. It is not clear yet how slow solar wind can be released at all. The transition from closed loops down below to an open field topology of the free solar wind flow above requires some kind of reconnection, be it in a steady-state mode or involving transient phenomena. It is conceivable that the slow solar wind might be released in a transient mode, in analogy to water drops dripping from the ceiling of a limestone cavern. (A thorough discussion of solar wind acceleration processes can be found in the articles SOLAR WIND: THEORY and SOLAR WIND ACCELERATION.)

Note that there is no continuous transition between these two types. Rather, the boundary layers between them are very thin close to the Sun and correspond in detail to the boundaries of the coronal holes.

There are two additional types of solar wind which can be considered products of the active Sun rather than the quiet Sun:

- the *slow wind of the maximum type* emerges from substantially larger areas distributed all over the Sun and often located far from any current sheet. It contains twice as much helium than the minimum type,
- *plasma clouds* released in huge eruptions at the sun can often be discerned by the unusually high helium percentages (up to some 30%) and other signatures (see the article SOLAR WIND: MANIFESTATIONS OF SOLAR ACTIVITY). Around activity maximum, these clouds contribute about 10% to the total solar wind flow, on the average.

In conclusion, it is only fair to state that for none of the mentioned types of solar wind is there a unique and generally accepted theoretical model. At present, existing theories are highly controversial. For example, an exciting alternative model for generating the fast or even all solar wind is acceleration in the form of numerous smallest-scale high-velocity jets (other terms in use: microflares, nanoflares, coronal bullets) that were discovered in very high resolution UV observations of the Sun (see TRANSITION REGION: EXPLOSIVE EVENTS). Apparently, such a model is in extreme contrast to Parker's classical thermal expansion model and its derivatives. This clearly illustrates that the understanding of solar wind acceleration is not yet on firm ground, and further research is needed.

Bibliography

Burlaga L F 1995 *Interplanetary Magnetohydrodynamics* (New York: Oxford University Press)

Cravens T E 1977 *Physics of Solar System Plasmas* (Cambridge: Cambridge University Press)

Golub L and Pasachoff J M 1997 *The Solar Corona* (Cambridge: Cambridge University Press)

Schwenn R and Marsch E (ed) 1990/1991 *Physics of the Inner Heliosphere* vols I and II (Berlin: Springer)

Vial J C, Bocchialini K and Boumier P (ed) 1998 *Space Solar Physics (Lecture Notes in Physics)* (Berlin: Springer)

Rainer Schwenn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Interaction with Comets

Cometary plasma tails are formed by the interaction of solar wind plasma with ionized gaseous material from a cometary nucleus over large volumes of space. The electromagnetic interaction proceeds as mass is added into, and slows, the solar wind flow near the comet.

The concept of a continuously flowing, but gusty, solar wind was postulated in the 1950s by Biermann to explain observations of cometary plasma tails. These spectacular tails are one of two effects of the solar wind that can be seen from the Earth's surface, the other being Earth's aurorae. Plasma tails were observed to point away from the Sun, with a small aberration angle due to the comet's orbital speed. Early theories included only the effects of solar wind particles, which proved to be inadequate, but later the crucial role of the magnetic field, frozen into the flow, was realized by Alfvén. The field should drape around an active cometary coma, ultimately forming the tail.

The earliest space probes confirmed the existence of both the solar wind itself and its magnetic field, frozen in to the highly conducting, radially flowing plasma at the Sun, but direct cometary exploration by spacecraft did not occur until the mid-1980s. So far four COMETS have been explored at close range with instrumentation to study the solar wind interaction: Giacobini–Zinner (by NASA's ICE in 1985), Halley (by ESA's GIOTTO, Russia's Vega 1 and 2 and the Japanese SUISEI and SAKIGAKE in 1986), Grigg-Skjellerup (by ESA's Giotto in 1992) and Borrelly (by NASA's DEEP SPACE 1 (DS1) in 2001). In addition, the Ulysses spacecraft encountered the distant (3.8 AU) tail of COMET HYAKUTAKE in 1996.

We now know that the interaction of a comet with the solar wind is quite different to that of a magnetized planet. As a comet approaches the Sun in its orbit, neutral atoms and molecules sublime away from the icy cometary nucleus, taking dust with the expanding flow of gas. The neutral atoms and molecules are ionized, on a timescale longer than a week, by sunlight or by charge exchange with the solar wind. Following ionization they interact with the electric and magnetic fields in the solar wind, slowing the flow. The solar wind drapes around the cometary nucleus, ultimately forming the comet's plasma tail. The solar wind interaction is thus quite different to that of a magnetized planet, where a magnetopause forms a blunt obstacle in the supersonic solar wind flow. These general concepts, which were predicted prior to cometary encounters by spacecraft, have been confirmed by spacecraft observations. Although many aspects of the interaction process have been elucidated, others are not yet understood and many unexpected results were gained. These are left as problems for future space missions.

Here we describe the evolution of the solar wind interaction with a comet as it nears the Sun, how particles are ionized and how they interact with the highly conducting solar wind plasma. We consider the expected boundaries and plasma regions in the solar wind–comet interaction together with the physical processes which are involved and the results from the cometary space missions of the mid-1990s. We then look at the outstanding problems for future space missions in this area.

Gas emission from comets

As well as the properties of the solar wind itself, the gas production rate of the comet is a vital parameter which determines the size and nature of the comet–solar wind interaction region. During the approach of a comet towards the Sun, the gas production rate increases owing to the increased surface temperature and the interaction changes from that of a bare nucleus to the fully developed coma structure. Also, different comets have different gas production rates, for example the production rates of Grigg–Skjellerup, Giacobini–Zinner and Halley were in a ratio of approximately 1:10:100 at the times of the encounters; for Halley the rate was 6.9×10^{29} molecules s^{-1}, corresponding to about 20 t s^{-1}. The Borrelly production rate was comparable with that of Giacobini– Zinner.

At distances greater than approximately 3 AU from the Sun the surface of a cometary nucleus is cold and no gas emission occurs. The bare nucleus interacts with the solar wind in the same way as unmagnetized, inactive asteroids or moons. A wake is formed behind the nucleus devoid of ions, assuming that the nucleus size is larger than or comparable with the solar wind ion gyroradius. The more mobile electrons set up potential structures which close the wake and may charge the night side of the nucleus, producing potentials of volts to kilovolts. The solar wind particles ram straight into the cometary nucleus, perhaps affecting the surface layers, while the solar wind magnetic field diffuses rapidly through the object. On the side of the nucleus facing the Sun, sputtering processes due to the incoming solar wind ions may produce neutral particles which move away from the comet perhaps to be ionized and picked up by the solar wind.

As the comet approaches the Sun more closely, sublimation processes begin to produce an outflow of gas. The species driven away and the rates at which this happens depend critically on the surface temperature and therefore the position in the orbit. A dense 'ionosphere' forms near to the comet, bounded by a contact surface. New, 'pick-up' ions are produced on a much larger distance scale of ~106 km, forming an enormous region over which the comet–solar wind interaction occurs. The new ions form an electromagnetically unstable population in the flowing solar wind. The solar wind is slowed, forming a bow shock and several other features. The interaction of a comet with the solar wind at large and small heliocentric distances is shown in figure 1.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

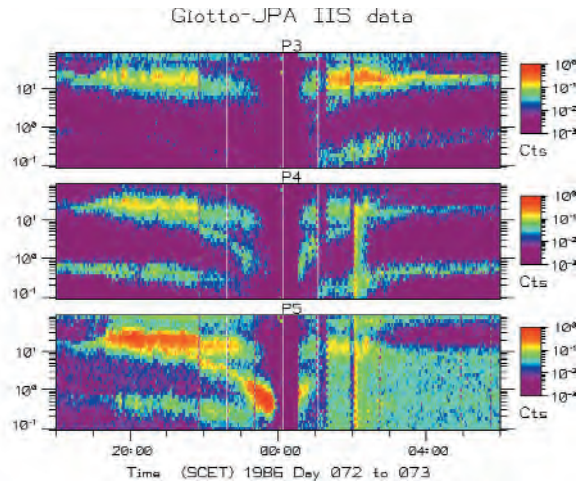


Figure 2. Summary of the Johnstone Plasma Analyser (JPA) Implanted Ion Spectrometer (IIS) results from Giotto covering 1.5 million kilometers each side of comet Halley, with closest approach at 00:03 spacecraft event time (SCET). The three panels show different look directions of the instrument, and in each panel the color scale shows number of particles. The solar wind is the lower trace in each panel, below 1 keV and decreasing in energy closer to the comet. The upper traces, at over 10 keV but decreasing closer, are the cometary water group ions. The number of cometary ions increases nearer to the comet and the solar wind slows, illustrating the interaction process. The bow shock is seen at ~19:35 inbound and ~02:45–03:15 outbound.

The Borrelly results in particular contain what may be evidence for an extended although well-defined source of neutral particles in the cometary coma, and an association with dust/gas jets from the nucleus is being sought. The observation in the plasma data, which leads to this tentative conclusion, was of a significant asymmetry in the cometary ion spectra in the interaction region. However, it is not clear why this should be much more pronounced at Borrelly than at the earlier cometary encounters.

Ionization and interaction with the solar wind

Before the encounters with any comets by spacecraft, the process of ion pick-up and implantation into the solar wind flow, which loads the flow with additional mass, was expected to be the key to the comet–solar wind interaction. Following the spacecraft encounters, the importance of pick-up ions has been confirmed (e.g. figure 2) and we now have a better understanding of the wave–particle interactions, particle distributions and wave properties which are key features of the deceleration of the solar wind in the cometary environment.

When a cometary neutral atom or molecule is ionized, it immediately responds to the electric and magnetic fields in the solar wind. In real space its path follows a cycloid; the exact form depends on the relative orientation between the solar wind velocity and the magnetic field. In its simplest form, where the electric and magnetic fields are perpendicular, the path is analogous to that of the valve on a bicycle wheel: rotation around the magnetic field (axle) superimposed with a drift speed (along the ground) given by $\boldsymbol{E}\times\boldsymbol{B}/B^2$. In velocity space, which uses velocities rather than distances as coordinates, this corresponds to a ring distribution. Because the ring distribution is unstable, plasma waves are produced.

Prior to the spacecraft encounters, it was expected that particles would scatter in pitch angle to form a shell distribution from the ring as the instabilities developed, and a first look at the data from Vega, Suisei and Giotto confirmed this. For a ring distribution, the maximum energy of a pick-up ion in the comet frame is $2mv_{sw}^2\sin^2\tau$, where m is the picup ion mass in amu, v_{sw} is the solar wind speed and τ is the angle between the solar wind and magnetic field vectors. This compares with a maximum of $2mv_{sw}^2$ for a shell distribution. Note that, for an average solar wind energy of 1 keV, and for ions of mass 18 (H_2O^+), the maximum shell energy is 72

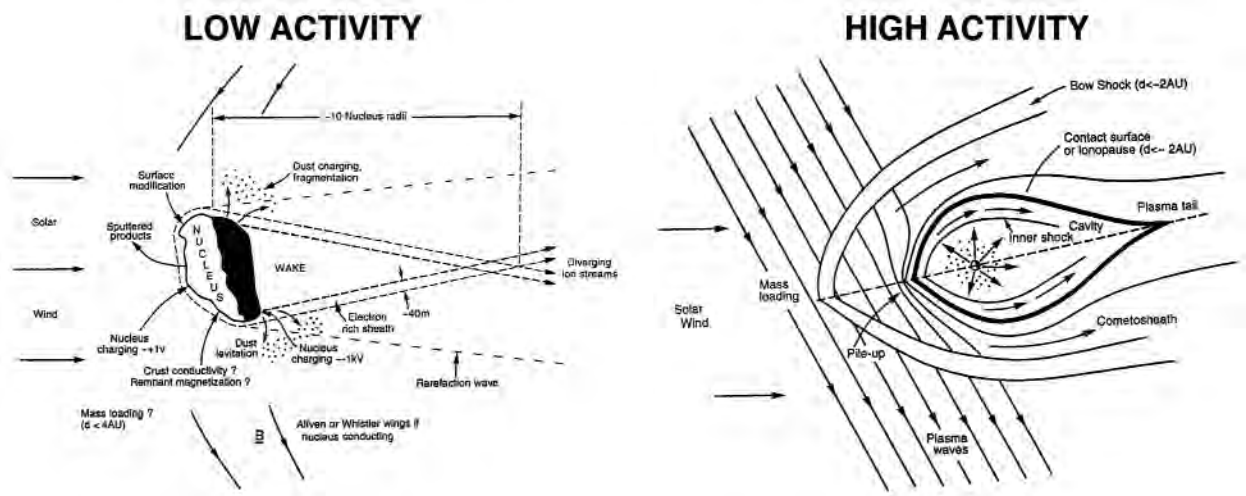


Figure 1. Comet–solar wind interaction (*a*) at large heliospheric distances (low activity) and (*b*) closer to the Sun (high activity).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

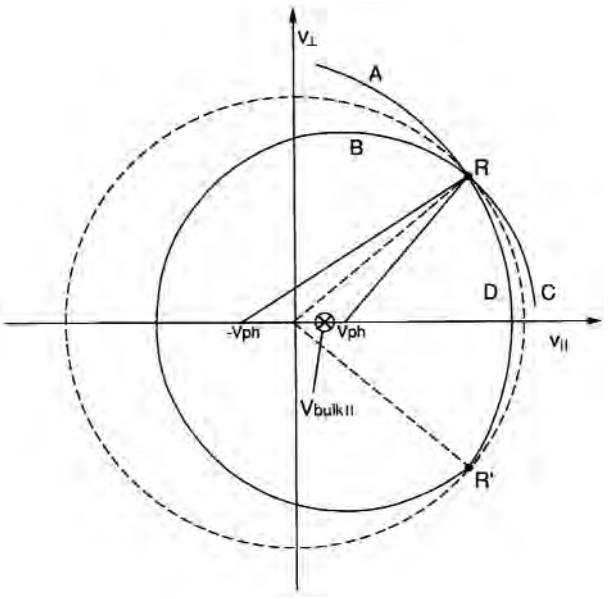


Figure 3. Velocity space diagram illustrating the ion pick-up process. Particles scatter from the original ring RR (onto the bispherical shell BD, due to interactions with upstream (B) and downstream (D) propagating waves.

keV. The Giotto IIS covered from 70 to 85 keV in order to cover fully the bulk of the pick-up ion distribution. Sensors on Suisei, Vega and ICE could only measure part of the pick-up ion distribution, in the last case only the high-energy tail.

From the shell, velocity diffusion was expected to broaden the shell in energy. Given enough time, a Maxwellian would finally be produced. Results from Giotto show that this later stage may only be approached in the very inner regions of COMET HALLEY.

More accurate theory reveals that simple shells are an approximation because the energy of the waves and the energy lost by the particles must be taken into account and energy must be conserved. This led to the idea of a 'bispherical shell' distribution for the cometary ions, which was subsequently confirmed in the data from Halley. The situation is illustrated in figure 3, where the scattering of particles in pitch angle from the ring R onto the spherical arcs RB and RD can be seen. In these cases, energy from the particle distribution goes into plasma waves. These ideas have been developed further in simulations and models; in particular, a bispherical model was used to predict the wave spectrum successfully and to compare the energy available

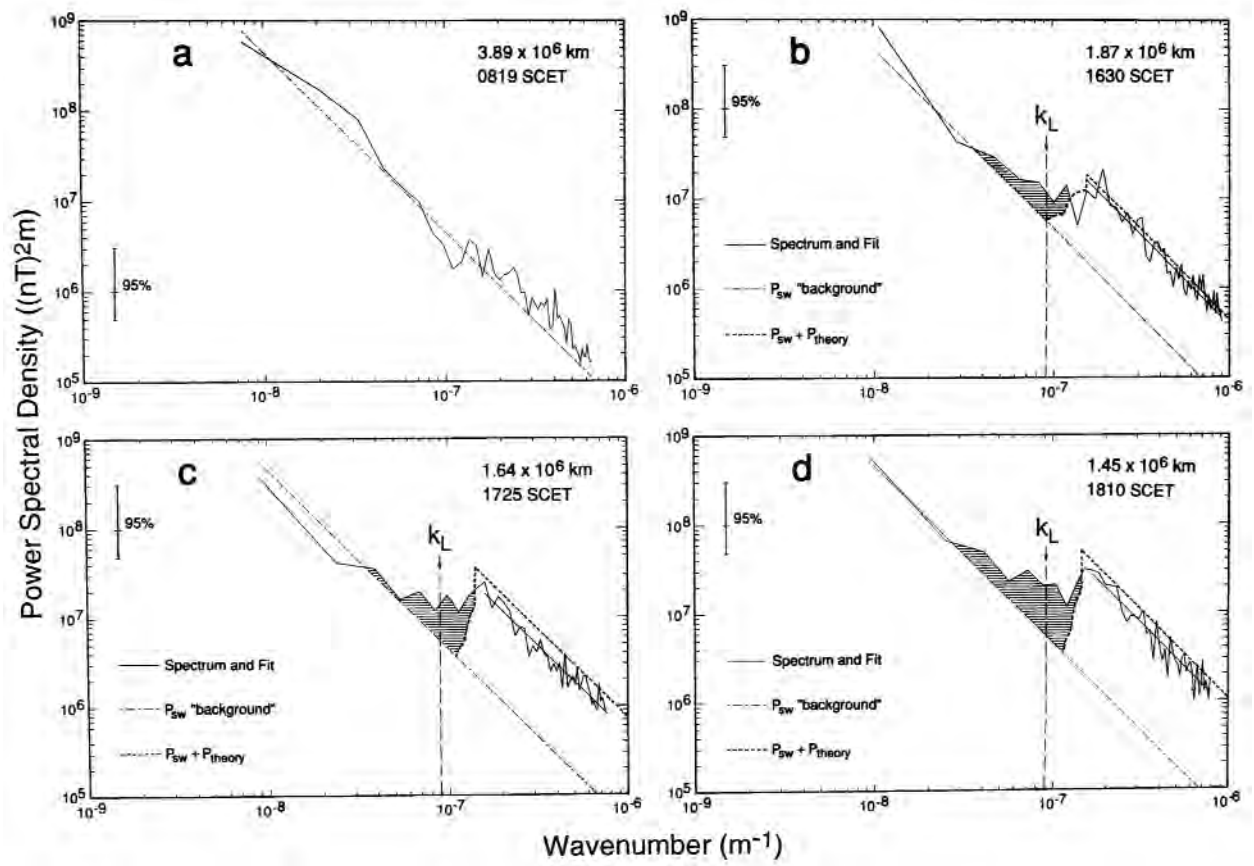


Figure 4. Observed (full curve) and predicted (broken curves) magnetic wave power spectra at different distances from comet Halley. The predictions are made using a bispherical shell model and the resonance of particles with different wave frequencies depending on their position on the shell, with no parameters.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Table 1.. Summary of observed boundary crossings at comets.

Comet	*Giacobini–Zinner*	*Halley*				*Grigg–Skjellerup*
	ICE	*V1*	*V2*	*Suisei*	*Giotto*	*Giotto*
Bow shock	Y	Y	Y	Y	Y	Y
Mystery boundary	?	?	?	?	Y	Y
Cometopause	—	—	Y	—	—	—
Magnetic pileup boundary	—	N	N	—	Y	Y
Discontinuities	—	—	—	—	Y	—
Ion pileup boundary	—	Y	Y	—	Y	?
Cavity boundary	—	—	—	—	Y	—

from the ring distribution with the energy released and the energy in the waves. In this elegant model, produced by Huddleston and Johnstone in 1992, particles at different positions on the shells resonate with specific frequencies of plasma waves and this can explain the observations without the need for further parameters in the theory (see figure 4).

From the Giotto data at comet Halley, well upstream of Halley the cometary ion distributions were found to follow the ring prediction, while closer to the comet the distributions were more shell like. The velocity distributions in the solar wind frame show a small amount of acceleration in the upstream region, significant enough to be seen by the energetic particle instrument, but the most significant acceleration was seen downstream of the cometary bow shock. The bulk velocity of the pick-up heavy ion population was found to be consistent with bispherical distributions in the region of the shock, despite the particle acceleration seen in that region.

In the case of comet Grigg–Skjellerup, it was shown that the ion distributions were ring like almost up to the bow shock, except that one-sided distributions were seen for a time outbound. Non-gyrotropic distributions were found particularly in the inner regions at Grigg–Skjellerup, owing to the different ion implantation rates between adjacent cusps of the cycloid in those regions. This can cause additional instabilities to operate.

One of the most dramatic features in the Grigg–Skjellerup results was the existence of large amplitude waves in the magnetic field and modulations of the particle distributions. It is thought that this is due to the textbook conditions of the encounter, as the solar wind velocity and magnetic field were mutually perpendicular at the time of encounter.

Mass loading by cometary ions was predicted and observed at comets. Models for the deceleration were compared with the measurements at comet Halley and good agreement was found.

Comet-associated energetic ions beyond the pick-up prediction had not been expected at comets, except in one paper immediately before the 1980s encounters. However, they were seen at all of the comets visited so far (except Borrelly as DS1 had no energetic ion detector) and mechanisms have been suggested for their acceleration. These included adiabatic acceleration, the interaction of ions with waves across a sharp gradient in solar wind flow speed (Fermi I) or the interaction with upstream and downstream propagating waves (Fermi II). The form of the bulk of the cometary ion distribution function depends on the point in the mass-loaded flow at which the observation is made. Well upstream of the comet the one-dimensional distribution function is a delta function modified on the low energy side owing to the bispherical distributions. Closer to the comet, where the solar wind speed is lower and some mass loading has occurred, the distribution is modified both by the slowing solar wind and by adiabatic acceleration. Downstream of the bow shock, two populations are found owing to the velocity jump across the shock. All of these distributions are further modified by the Fermi I or II processes. Comparison of models with data showed that Fermi II alone may not be sufficient to explain the Halley observation. The proportion of downstream propagating waves was larger than expected, which may account for some additional acceleration. Additional energy may go into particle acceleration from the solar wind deceleration.

Boundaries and regions in the comet–solar wind interaction

Two boundaries in the mass-loaded flow were expected in the 1970s and 1980s by workers including Wallis, Biermann and Schmidt, separating plasmas with different properties. These were the bow shock and the contact surface. The reality turned out to be much more complex with several new boundaries seen by the spacecraft.

Figure 5 is a schematic diagram which shows the trajectories and boundary observations of all the cometary encounters so far (except Borrelly as no peer-reviewed papers have appeared thus far). As mentioned above, the bow shock and cavity boundary were the only boundaries predicted to form prior to the spacecraft encounters. The data have added at least four additional features: the mystery boundary, the cometopause, the magnetic pile-up boundary (MPB) and the ion pileup

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

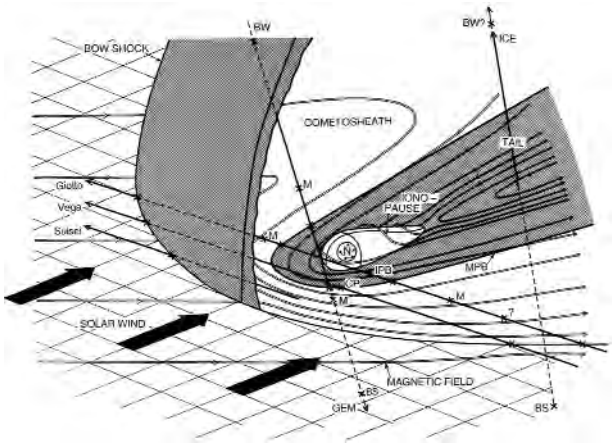


Figure 5. Schematic of the comet–solar wind interaction, with spacecraft trajectories and boundary crossings. For comparison, the DS1 Borrelly encounter was similar to the GEM (Grigg–Skjellerup) geometry.

boundary (IPB). The observations are summarized in table 1. We will now discuss each of the boundaries in turn.

Bow shock

he bow shock was one of the predicted boundaries in the comet–solar wind interaction region. In the predictions, the mean molecular weight of the flow reached a critical value of 4/3 in the first simple one-dimensional model before a strong singularity, the bow shock, occurred. In more realistic two- and three-dimensional simulations a weak bow shock forms at a value lower than 4/3, dependent on the gas production rate. The Mach numbers of these weak shocks were approximately 2.

Observations at comets showed that this boundary was present at all of the spacecraft encounters but its properties changed dramatically between crossings at the same comet depending on the interplanetary magnetic field orientation. In all cases the cometary ions dominated the dynamics in terms of pressure. The width of the feature varied from a few water group ion gyroradii in the quasi-perpendicular case to many gyroradii in the quasi-parallel case. Where sufficient data were available, the bow shock was shown to occur at the predicted point in the mass-loaded flow. Also, the Mach number of the features between 1 and 2 supported the interpretation of a weak bow shock. The various particle populations (electrons, protons, alpha particles, cometary water ions) also increased in temperature across the boundary. While there is some evidence that the GZ data can be explained in terms of a reforming shocklet model and the Suisei and Giotto–Halley inbound data could have been affected by interplanetary disturbances, there is broad agreement that the boundary seen in the various cases was a bow shock.

Mystery boundary

One of the dramatic plasma discoveries at comet Halley was the clear bifurcation of the cometary ion peak in the 'cometosheath' region between the bow shock and contact surface. The explanation for the bifurcation has already been alluded to, namely the different velocities either side of the bow shock which give different pick-up shell radii. However, this does not explain the very suddenly enhanced splitting of the water group ion population which occurs at approximately 500 000 km from the nucleus caused by a change in solar wind velocity there. This boundary, called the mystery boundary as its origin is still not understood, is also characterized as the end of a region of significantly enhanced densities of hot (0.8–3.6 keV) electrons. The mystery boundary also ends a region of higher solar wind density and velocity, and following the boundary the number of ions in the ram direction increases sharply. There is no strong effect in the magnetic field to match these sharp and significant features in the plasma data.

The boundary is still present on the outbound pass (although suitable electron data are not available) and also appears to be at comet Grigg–Skjellerup. Also, it has been argued that comparable boundaries were also observed by Vega, Suisei and ICE. However, it is not present in any of the MHD or multifluid models and its origin has still not been explained. Although the boundary itself is still a mystery, it is conceivable that the increased numbers of hot electrons in the region it bounds could play a role in the x-ray emission from comets discovered in the 1990s.

Cometopause

Analysis of the data from Vega-2 has shown the existence of a relatively sharp boundary some 160 000 km from comet Halley separating regions of different chemical composition, according to measurements in the ram direction. In these data the protons showed a sudden decrease in density and the cometary ion density also suddenly increased, as seen in the ram direction. In addition, the proton velocity and water group velocity was quite different each side of the transition. It was suggested that the cometopause is a permanent boundary in the comet-solar wind interaction. The analysis of Giotto data from a similar region revealed a much broader transition between a mass-loaded solar wind region and a heavy ion dominated region with no significant velocity difference.

In this region it is certainly clear that charge exchange collisions are becoming more important, which would give a shift from a solar wind dominated flow to cometary ion domination. This is the physics which has been put forward to explain the cometopause transition but it remains to be explained why the transition was seen to occur more rapidly on Vega than Giotto. A sharp

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

boundary is not predicted by models and one possibility is that a convected change of interplanetary conditions could explain the Vega observations.

Magnetic Pile-up Boundary (MPB)

In the Giotto data, it is more widely accepted that a separate MPB is seen some 45 000 km or 11 min later than the cometopause. The main characteristics of this boundary on the inbound pass were a sudden magnetic field jump (tangential discontinuity), a slowing and reduction of electron density and a decrease in the electron perpendicular pressure. A dramatic increase in the rate of charge exchange and a proton density decrease was also seen. There are signs of a similar boundary outbound in the magnetometer data. This boundary, inside which the magnetic field drapes strongly around the contact surface, was not predicted by models and, interestingly, was absent in the Vega data at comet Halley. This, and the observation that the cometary ion density shows no significant discontinuity, has led some to speculate a solar wind origin while others have postulated that on the night side it becomes the plasma tail boundary.

An MPB was also observed at Grigg–Skjellerup; in the magnetic field, electron and ion data.

Discontinuities X, X′, Y and Z

Three additional boundaries have been pointed out in the Giotto data between the MPB and the IPB, called X, Y and Z. Discontinuity X is in a region where cometary ions are gradually being lost by charge exchange; there is also a sudden, significant drop in proton density there. A similar boundary X′ occurs a little later. Transition Y is the start of a sharp decrease in ion temperature. Discontinuity Z, 4 min before closest approach, marks the transition from radial outward flow of cometary ions away from the nucleus to stagnant plasma.

Ion Pile-up Boundary (IPB)

In the Giotto data at approximately 10 000 km, a significant unexpected peak in counts was found. The gradient of the log count rate changed from r^{-1} (inside) to r^{-2} (outside). This can be interpreted as the limit of ionospheric plasma. A magnetic field maximum also appeared at this point. In this region, significant chemical changes are occurring, for example the fraction of the H_3O^+ ion became dominant close to this point. This boundary was also seen by the Vega probes at comet Halley. It was not predicted in models and it remains the most significant deviation from these models in the inner regions.

The early interpretations of this region, including temporal effects in the neutral gas, change of upstream solar wind parameters and local electron impact ionization, were all rejected in favor of a reduction of the rate of recombination of ions and electrons caused by increased electron temperature. Observations and theory now back up this explanation.

Cavity boundary

Giotto was the only spacecraft to cross the expected cavity boundary; nevertheless, it is reasonable to suppose that this is one of the permanent boundaries in the comet–solar wind interaction. The Giotto observations showed that the magnetic field was excluded from the cavity, and the boundary itself was approximately 25 km wide. Near the cavity boundary, bursts of ions were seen. The position of the boundary can be understood by a balance between inward magnetic gradient forces and outward ion–neutral drag. The pile-up of ions just inside the cavity boundary was predicted by models and this perhaps plays the role of the inner shock, a long-predicted structure in the inner coma. Inside the cavity boundary the cometary plasma is stagnant.

One interesting feature discovered in the inner regions was the existence of negative ions. The density variation with distance is steeper inside the cavity boundary compared with outside, and the existence of these ions was not predicted by models. What is now clear is that their existence is an important clue to the photochemical processes at work.

Remote observations

While spacecraft encounters have added immensely to our knowledge of the comet–solar wind interaction, there is still a vital role for remote studies of comets from the ground and from space. From the ground it is possible to use the plasma tail of any comet as a probe of the solar wind. Study of tail features such as kinks and knots can reveal the solar wind speed. In particular, disconnection events, where the plasma tail is removed from a comet and a new one forms, can be used to infer the presence of discontinuities in the solar wind. One of the curious features is the formation of rays in the cometary coma. It is still not clear precisely what relation this bears to tail formation and to the *in situ* results.

Composition may also be studied from the ground; indeed, one of the early discoveries was that the blue color of a plasma tail is due to emission mainly from the CO^+ ion. Many other species were found from telescopic measurements and confirmed by *in situ* measurements.

The use of different wavelengths has also revealed new features. The discovery of x-ray emission from Comet Hyukatake, and subsequently also from Hale–Bopp and Encke, using the Rosat satellite, has shown that unexpected discoveries are still possible from remote observations. Several mechanisms have been suggested for the emission but the most likely appears to be that the x-rays are from the decay of excited states formed as a result of charge exchange between heavy solar wind ions and cometary ions.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The 1996 discovery, made using the Ulysses spacecraft, that the ion tail of comet Hyukatake was at least 3.8 AU long opens up a new area of remote observations of comets using particles and fields instruments on interplanetary spacecraft. Associations of magnetic signatures in Pioneer Venus Orbiter (PVO) data with an unknown comet had in fact already been made using data from 1982. This was the largest of other similar signatures from PVO and other spacecraft, and searches of data from other interplanetary spacecraft are in progress for this type of event. However, the Ulysses–Hyakutake event is unique in that there was a clear association with a comet and there were also indications of cometary ions even at that large distance.

Outstanding problems

It is clear that the solar wind interaction with a comet is radically different to that with a magnetized planet. An induced magnetosphere forms, owing to the implanted cometary ions loading the flow over vast regions of space and causing the draping of magnetic field within the extended 'obstacle'. The unstable implanted ion distributions cause plasma waves, which scatter the ions to form a bispherical distribution in velocity space. Particle acceleration is caused by wave–particle interactions in the disturbed cometary environment. Non-gyrotropy of the ion distributions is significant where the ion implantation rate varies significantly over the cycloid. This is particularly relevant at small, or weakly outgassing, comets.

Of the various boundaries observed in the interaction region, only the bow shock and contact surface were predicted. These were confirmed as permanent features of the developed solar wind interaction. Other permanent boundaries are the mystery boundary and the IPB. Other boundaries, which may not be permanent, are the cometopause and the MPBs. Although the latter was seen both at Halley and Grigg–Skjellerup it was absent during the Vega flybys of Halley and hence it may not be permanent.

The encounters thus far have left us with a number of intriguing questions about the areas they visited, including the following. How important is the Fermi I mechanism in particle acceleration? Is the inner shock really a layer just inside the cavity boundary? How permanent are the boundaries in the comet–solar wind interaction, and what is their formation mechanism? How does the cometary plasma tail form and what is the importance of tail rays? Could the mystery region hot electrons cause x-ray emission in the cometary environment?

There are also significant questions about the regions not visited, or only fleetingly sampled, so far. These include many features of the cometary ionosphere, photochemical interactions and the formation of negative ions, the exploration of the tail formation region, the dependence of features on gas production rate as the comet approaches the Sun, the role of reconnection in the interaction processes and the dependence on different solar wind conditions. Some of these questions should be answered by ESA's ROSETTA mission which will rendezvous with comet Wirtanen in 2011 and fly with the comet as it nears the Sun and its activity develops. At present this is the only planned mission with suitable instrumentation to explore further the solar wind–comet interaction.

Bibliography

1991 *Cometary Plasma Processes (Geophysical Monograph 61)* (Washington, DC: American Geophysical Union)

1993 Giotto's encounter with comet Grigg–Skjellerup *J. Geophys. Res. Space Physics* **98** (A12) (special issue)

Grewing M, Praderie F and Reinhard R (ed) 1987 *Exploration of Halley's comet* (Heidelberg: Springer) (reprinted from 1987 *Astron. Astrophys.* **187** (1 and 2)

Neugebauer M 1990 Spacecraft observations of the interaction of active comets with the solar wind *Rev. Geophys.* **28** 231–52

Newburn R L Jr, Neugebauer M and Rahe J (ed) 1991 *Comets in the Post-Halley Era* (Dodrecht: Kluwer)

Andrew Coates

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Solar Wind: Interaction with Local Interstellar Medium

The solar wind forms a bubble, called the heliosphere, in the local interstellar medium (LISM), within which the solar system resides and whose size and properties are determined by the manner in which the solar wind plasma and the partially ionized LISM are coupled. In the last decade, great progress has been made in our understanding of the physical processes thought to describe the outer heliosphere. Fundamental to these advances has been the recognition that the interstellar medium is coupled intimately to the heliosphere itself and that much of outer heliospheric physics cannot be understood independently of the local interstellar medium. With the possibility that the aging spacecraft Voyager 1, 2 and Pioneer 10, 11 might encounter the heliospheric boundaries in the not too distant future, interest in the far outer reaches of our solar system and the LISM has been rekindled (see also PIONEER MISSIONS, VOYAGER MISSION). A convenient, if slightly vague, definition of the outer heliosphere, and one that is adopted here, is that it is that region of the solar wind influenced dynamically by physical processes associated with the LISM. Thus, for example, neutral interstellar hydrogen is the dominant (by mass) constituent of the solar wind beyond an ionization cavity of ∼6–10 astronomical units (AU) in the upstream direction (the direction antiparallel to the incident interstellar wind). The neutral hydrogen is coupled weakly to the solar wind plasma via resonant charge exchange—a coupling which leads to the production of pick-up ions that come to dominate the internal energy of the solar wind. The solar wind changes then from a small plasma beta β_{SW} (the ratio of plasma pressure to magnetic field pressure) environment to one in which $\beta_{SW} > 3$.

Figure 1 shows the complex interaction between solar wind plasma and the LISM. While clearly simplistic, it serves to illustrate the primary physical processes. The interstellar plasma may or may not impinge supersonically on the heliosphere. The LISM flow is diverted about the heliospheric obstacle either adiabatically or by a bow shock (BS). The LISM plasma is coupled to the neutral interstellar hydrogen (H) primarily through charge exchange. Charge exchange corresponds to a neutral atom surrendering an electron to a passing ion (typically a proton), so creating a new charged particle and (a hydrogen) atom. Thus, when neutral H charge exchanges with LISM plasma that has been decelerated, heated and diverted about the heliosphere, the neutral H acquires a subpopulation that too is heated, decelerated and diverted. This neutral hydrogen, together with the LISM H, is called component 1 and is created in region 1 of figure 1. If the charge exchange mean free path is sufficiently small in the region of decelerated LISM flow, a 'wall' of neutral hydrogen will form in the upstream direction (the hydrogen wall). This is because resonant charge exchange is essentially a scattering rather than an extinction process.

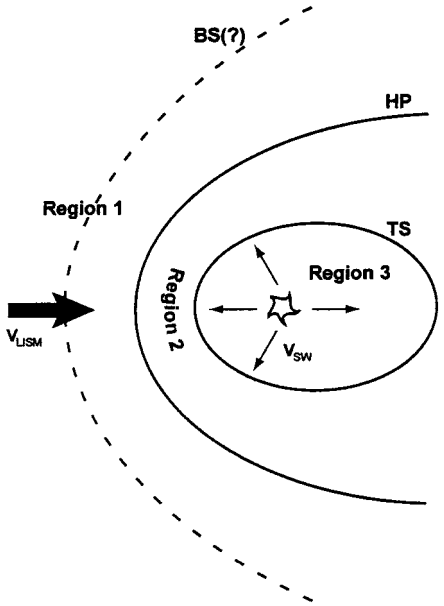


Figure 1. Schematic of the solar wind–local interstellar medium (LISM) boundary regions which act as neutral hydrogen sources whose characteristics are identifiably distinct. The solar wind is enclosed by the termination shock (the curve labelled TS) and is identified as region 3, the shocked solar wind, the heliosheath, is bounded by the termination shock (TS) and the heliopause (HP) and is identified as region 2, while the interstellar medium is found beyond the heliopause. The interstellar medium may or may not form a bow shock (BS) depending on whether the interstellar flow velocity V_{LISM} is supersonic. The radial solar wind flow vector is denoted by V_{SW}.

A boundary, called the heliopause, separates the heliosphere and the LISM plasma, and is either a contact or a tangential discontinuity. The supersonic solar wind is decelerated, diverted and heated by a reverse shock called the termination shock (TS). (See the accompanying article on SOLAR WIND SHOCK WAVES AND DISCONTINUITIES by R. Lepping for definitions of terms related to hydrodynamic and MHD discontinuities.) The location of the termination shock is determined by the steady and temporal solar wind pressure exerted on the LISM.

Neutral interstellar hydrogen that crosses the heliopause into the heliosphere can also experience charge exchange. Neutral H atoms that are created by charge exchange with the very hot subsonic plasma downstream of the heliospheric termination shock possess thermodynamic attributes that correspond to their origin, i.e. very hot ($\sim 10^6$ K) with an outwardly directed velocity component. This population of neutral atoms is called component 2, is created in region 2 (figure 1), and it streams away from the heliosphere and is quite distinguishable from component 1. By means of secondary charge exchange with the LISM plasma, component 2 acts to transport heat from the heliosheath (the region located between the heliospheric termination shock and the heliopause) into the LISM and this leads to an increase in the LISM tempera-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ture in the immediate vicinity of the heliopause. The final neutral component, component 3, possesses characteristics quite distinct from the other two components and has its origin in the supersonic solar wind (region 3, figure 1). Component 3 has large outward radial velocities and relatively low temperatures but is dynamically unimportant owing to its very low energy density.

The ions born of charge exchange in the supersonic solar wind, the pick-up ions, are, however, of considerable dynamical importance since they remove both momentum and energy from the bulk solar wind flow. The solar wind is decelerated, so reducing the ram pressure and hence the expected size of the heliospheric cavity, and the solar wind acquires a tenuous non-thermalized population of suprathermal ions (typical energies ∼1 keV). The initial pick-up ion population is unstable, generating low-frequency magnetic turbulence which can scatter both pick-up ions and cosmic rays. Some fraction of the pick-up ions is further energized, possibly by shock acceleration, to form the anomalous cosmic ray component, although the precise injection mechanism awaits clarification.

Clearly, a heliosphere mediated by neutral interstellar hydrogen possesses qualities that are likely to be very different from a heliosphere in which the role of neutral hydrogen is neglected. Below, we shall concentrate primarily on global models of the heliosphere in which interstellar neutral hydrogen is accounted for self-consistently. It should be noted that LISM conditions were very different in the past than they are at present (and indeed will be in the future) and thus the structure of the heliosphere may have been very different from the models discussed here.

The material presented here has been reviewed extensively by Holzer (1979) and Zank (1999) and the reader is referred to these works for more detail and an extended bibliography.

The local interstellar medium

Many of the pertinent physical parameters of the LISM are poorly constrained and so, by implication, the detailed solar wind–LISM interaction is not yet well understood. Our best inferences about the LISM flow velocity and temperature have been made from observations of neutral interstellar hydrogen (H) and helium (He) within the heliosphere, using either resonantly scattered solar UV light techniques (since H atoms in the heliosphere are illuminated by the Sun in the HI resonance line, i.e. the Lyman-α line) or *in situ* spacecraft measurements of the He distribution. The bulk velocity of the LISM flow is ∼26 km s^{-1} and the plasma temperature is ∼8000 K. The H and H$^+$ number densities are not well determined, although approximate limits can be derived. For H I (i.e. neutral H only), number densities lie in the range of ∼0.15–0.34 cm^{-3}, while for H II (ionized H), the range is ∼0.03–0.14 cm^{-3}, depending on whether the EUV radiation field is included or not. This leads to a limit on the total number density of both neutral and ionized interstellar hydrogen in the range ∼0.15–0.34 cm^{-3}.

Although most models assume that the relative motion of the heliosphere through the interstellar medium is supersonic, it is by no means clear that such an assumption is completely warranted. Our knowledge of both the local interstellar magnetic field strength and orientation and the energy density in interstellar cosmic rays is somewhat rudimentary. Cosmic rays with energies of 300 MeV/nucleon and less may contribute $\sim(3 \pm 2) \times 10^{-13}$ dyn cm^{-2}, which can yield an interstellar flow that is subsonic. The magnetic field orientation appears to be perpendicular to the interstellar flow velocity vector and estimates for the magnetic field strength range from ∼1.3–2 μG.

Evidently, our knowledge about important LISM plasma and neutral H parameters is rather poor and this must therefore be taken into account when evaluating models of the global heliosphere. Nonetheless, the basic underlying physical processes appear to be reasonably well understood.

Global heliospheric structure

The dynamical or ram pressure (given by ρu^2, ρ being the fluid mass density and u the flow speed) and thermal pressure p of the solar wind decrease with increasing heliocentric distance and must reach a value which eventually balances the pressure exerted by the LISM. The relaxation towards pressure equilibrium between the solar and interstellar plasmas is characterized by (a) a transition of the supersonic solar wind flow to a subsonic state, and (b) a divergence of the interstellar flow about the heliospheric obstacle. The transition of the supersonic solar wind is most likely accomplished by means of a shock transition, the termination shock (denoted by TS), and it is anticipated that at least the Voyager spacecraft will encounter this boundary in the early twenty-first century. The divergence of the LISM flow about the heliosphere may be accomplished either adiabatically (if the relative motion between the Sun and the LISM is subsonic) or by means of a bow shock in the case of supersonic relative motion. If we neglect both the deceleration of the solar wind by resonant charge exchange and temporal variations in the solar wind ram pressure, the minimum radius $R_{\rm TS}$ of the solar wind shock transition can be estimated from

$$\frac{R_{\rm TS}}{r_0} = \left(\frac{\gamma + 3}{2(\gamma + 1)} \frac{\rho_0 u_0^2}{p_\infty} \right)^{1/2} \qquad (1)$$

where $\rho_0 u_0^2$ is the solar wind ram pressure at $r_0 =$ 1 AU, γ the solar wind adiabatic index and p_∞ the total LISM pressure. The LISM pressure term can include the thermal gas, cosmic rays, the interstellar ram pressure, the magnetic field pressure and MHD turbulence, and may be expressed as

$$p_\infty = \rho u^2 + p_{\rm th} + \eta B^2/8\pi + p_{\rm CR} + p_{\rm dust} + p(\delta B^2)$$

where the terms are all evaluated in the LISM and the factor η attempts to include the effects of interstellar magnetic field obliquity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

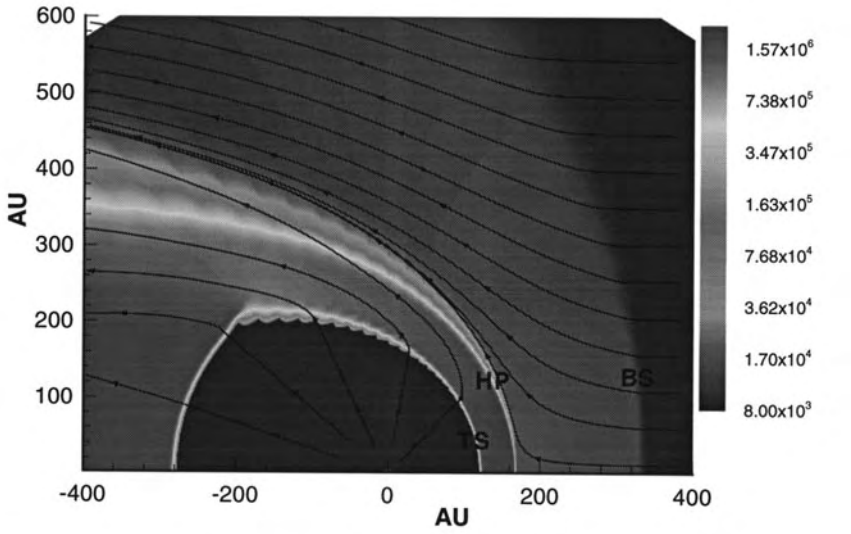


Figure 2. Gray-scale plot of the $\log_{10}$ (temperature) and streamlines for a 2D two-shock model of the solar wind–LISM interaction in the absence of neutral interstellar hydrogen. The termination shock (TS), heliopause (HP) and bow shock (BS) are identified.

Although one can estimate the location of the TS and the heliopause, the discontinuity separating solar wind material from the interstellar plasma (a contact discontinuity in the case of gas dynamics), using simple pressure balance arguments, the interaction of the solar wind with the LISM is fundamentally multi-dimensional. Thus, the main advances in our understanding of global heliospheric structure since the original analytic models have been more recent and based largely on computer simulations. The initial simulations were of purely one-fluid gas dynamic models and only now has the inclusion of neutral interstellar hydrogen been considered self-consistently.

To model the interaction of the solar wind with a partially ionized LISM, the following set of 3D MHD equations must be solved

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \boldsymbol{u} = Q_\rho \tag{2}$$

$$\rho \frac{\partial \boldsymbol{u}}{\partial t} + \rho \boldsymbol{u} \cdot \nabla \boldsymbol{u} + (\gamma - 1)\nabla e + (\nabla \times \boldsymbol{B}) \times \boldsymbol{B} = \boldsymbol{Q}_m \tag{3}$$

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\rho u^2 + e + \frac{B^2}{8\pi}\right) + \nabla \cdot \left[\left(\frac{1}{2}\rho \boldsymbol{u}^2 + \gamma e\right)\boldsymbol{u} + \frac{1}{4\pi}\boldsymbol{B} \times (\boldsymbol{u} \times \boldsymbol{B})\right] = Q_e \tag{4}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} - \nabla \times (\boldsymbol{u} \times \boldsymbol{B}) = 0 \tag{5}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{6}$$

together with the equation of state $e = \alpha n k_B T/(\gamma - 1) = p/(\gamma - 1)$. Here $\alpha = 2$ if only electrons and protons are considered as part of the LISM plasma, otherwise α can be larger when, for example, cosmic rays are included. The charge exchange terms Q_ρ, $\boldsymbol{Q}_m$ and Q_e represent pick-up ion source terms for the density, momentum and energy equations; ρ, $\boldsymbol{u}$ and p denote plasma density, velocity, and total pressure, and the electrons and ion temperatures are assumed to be equal. The adiabatic index $\gamma = 5/3$. The source terms couple the plasma and neutral H distribution and, in principle, one needs to compute the evolution of both distributions simultaneously.

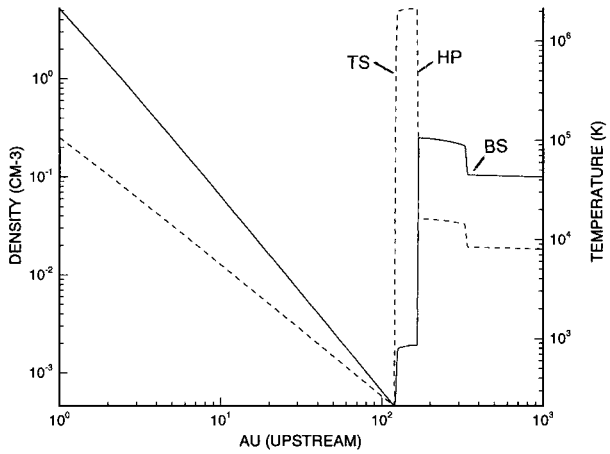
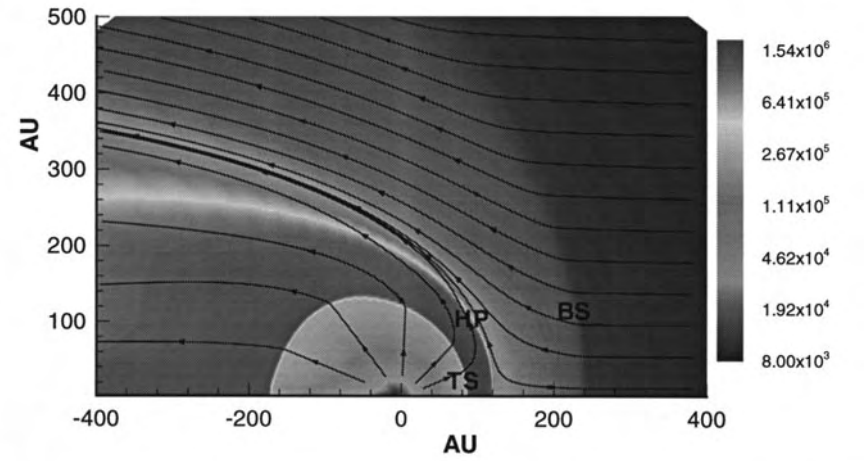


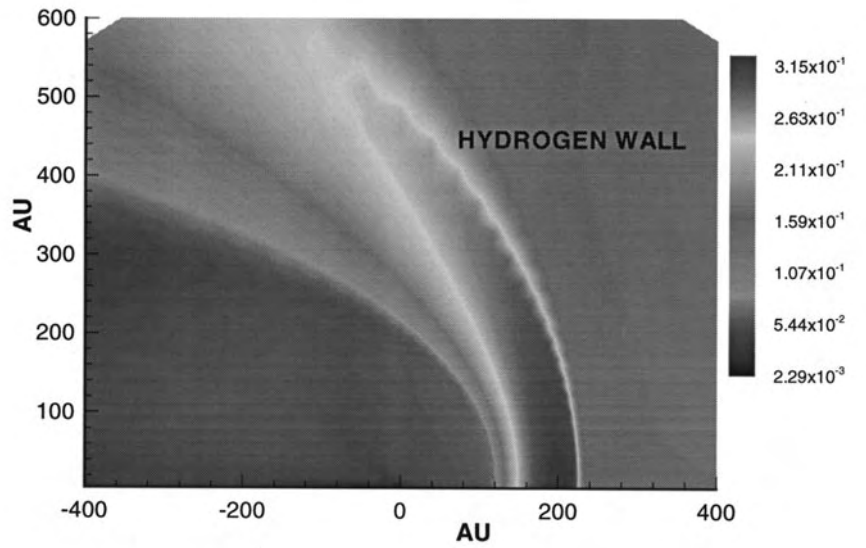
Figure 3. One-dimensional cuts along the stagnation axis in the upstream direction of the plasma-only 2D simulation shown in figure 2. The solid curve shows the plasma density and the dashed curve the plasma temperature profile. The termination shock (TS), heliopause (HP) and bow shock (BS) are marked.

Two approaches have been taken to evaluate the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*)



(*b*)

Figure 4. (*a*) A 2D plot of the global plasma structure for the two-shock model when neutral H is included self-consistently. The shading refers to $\log_{10}$ (temperature). The streamlines are included. (*b*) A 2D plot of the corresponding component 1 neutral distribution, where the shading now refers to the number density. The hydrogen wall is clearly visible between the bow shock (BS) and the heliopause (HP).

neutral H distribution. One approach is to solve the Boltzmann equation

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla f + \left(\frac{\boldsymbol{F}}{m} \cdot \nabla_v \right) f = P - L \qquad (7)$$

for neutral H directly. Here, $f(\boldsymbol{x}, \boldsymbol{v}, t)$ is the neutral hydrogen distribution function expressed in terms of position $\boldsymbol{x}$, velocity $\boldsymbol{v}$ and time t. $\boldsymbol{F}$ is the force acting on a particle of mass m, typically gravity and radiation pressure. The terms P and L describe the production and loss of neutral particles at $(\boldsymbol{x}, \boldsymbol{v}, t)$ and both terms are functions of the plasma and neutral distributions.

An alternative approach, and one that is much less prohibitive computationally, is to recognize that the heliosphere–LISM environment comprises three thermodynamically distinct regions: the supersonic solar wind (region 3), the very hot subsonic solar wind (region 2), and the LISM itself (region 1) (figure 1). Each region acts as a source of neutral H atoms whose distribution reflects that of the plasma distribution in the region. One may identify neutral components 1, 2 and 3

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with neutral atoms originating from regions 1, 2 and 3. Each of these three neutral components is approximated by a distinct Maxwellian distribution function appropriate to the characteristics of the source distribution. This observation allows the use of simpler production and loss terms for each neutral component. The complete highly non-Maxwellian H distribution function is then the sum over the three components, i.e.

$$f(\boldsymbol{x}, \boldsymbol{v}, t) = \sum_{i=1}^{3} f_i(\boldsymbol{x}, \boldsymbol{v}, t).$$

Under the assumption that each of the neutral component distributions is approximated adequately by a Maxwellian, one obtains an isotropic hydrodynamic description for each neutral component, and thus a closed multifluid model describing self-consistently the interaction of the solar wind with the partially ionized LISM.

Before discussing the models which include interstellar neutrals self-consistently, consider briefly the 2D structure of the heliosphere in their absence. Figure 2 illustrates the 2D plasma temperature and streamlines for the H^+ fluid at steady state. This particular model does not include either interplanetary or interstellar magnetic fields and assumes a supersonic interstellar wind. Three major boundaries or interfaces are present: the bow shock which compresses and diverts the interstellar flow about the heliosphere, the heliopause which separates the interstellar medium and the solar wind, and the termination shock at which the supersonic solar wind is decelerated abruptly. The termination shock possesses a characteristic bullet shape, quite unlike the approximately spherically symmetric heliosphere that is assumed typically. The supersonic solar wind velocity is constant until the very strong shock (Mach number $\simeq$ 170), and figure 3 illustrates the radial density and temperature profiles along the stagnation line in the upstream direction. The termination shock is located at ~120 AU for this simulation. The solar wind temperature decays adiabatically with increasing heliocentric distance. An extended and very hot tail of subsonic solar wind is formed (called the heliotail) in the downstream direction.

A major effect of charge exchange on the heliospheric interfaces is to decrease the distances to the TS, HP and BS, illustrated in figure 4(*a*). The decrease in distance results primarily from the reduction in solar wind ram pressure, due to the mediation of the wind by charge exchange.

Besides the distance to the various heliospheric boundaries, charge exchange affects the shape of the termination shock, making it more spherical than the purely ionized gas dynamic description. This is a consequence of charge exchange in the heliosheath, which ensures that the shocked solar wind plasma remains subsonic in this region.

Owing to the deposit of interstellar protons in the solar wind when charge exchange is included, the solar wind flow now departs slightly from simple spherical symmetry, and the contribution to the internal energy of

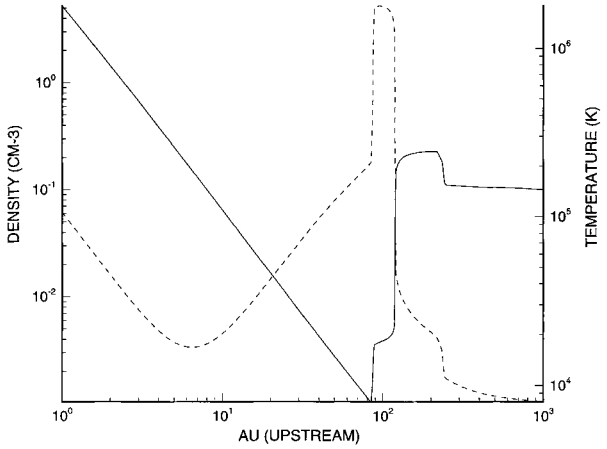


Figure 5. One-dimensional cuts along the stagnation axis in the upstream direction of the coupled plasma-neutral 2D simulation shown in figure 4(*a*). The solid curve shows the plasma density and the dashed curve the plasma temperature profile.

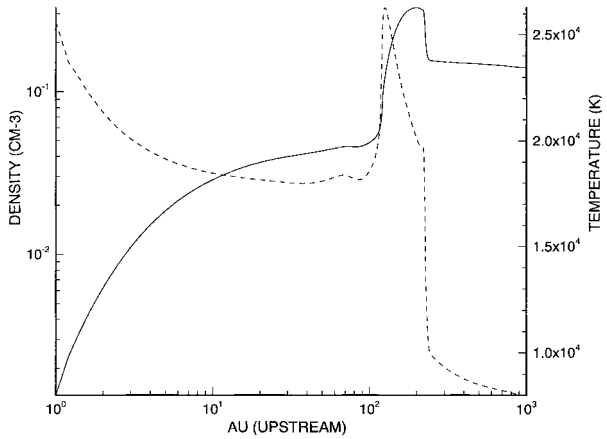


Figure 6. One-dimensional cuts along the stagnation axis in the upstream direction of the coupled plasma-neutral 2D simulation shown in figure 4(*b*). The solid curve shows the component 1 density and the dashed curve the component 1 temperature profile.

the supersonic solar wind by pick-up ions is substantial. This is illustrated in the 1D radial plots of the plasma density and temperature along the upstream stagnation axis (figure 5). The tail region is also cooler when the proton fluid and neutral fluid are coupled compared with no-charge exchange models. The cooling is a consequence of the very hot heliotail being cooled by charge exchange with cooler component 1 neutrals.

To enter the heliosphere, a hydrogen atom has to cross three thermodynamically distinct regions. Through resonant charge exchange, the diverging LISM flow and the shocked solar wind flow act to divert some fraction of the incident interstellar neutral H flux away from the heliosphere. This effective swapping of trajectories by an interstellar neutral and a proton can also lead

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to an enhancement in the neutral interstellar density and temperature in the vicinity of the solar wind–LISM stagnation region.

Figure 4(*b*), a 2D plot of the component 1 neutral H density distribution, shows that inflowing component 1 neutrals are substantially decelerated and filtered by charge exchange with the interstellar plasma between the BS and HP in the upstream direction. This leads to the formation of a hydrogen wall with maximum densities ∼0.2–0.3 cm^{-3}, column densities ∼10^{14} cm^{-2} and temperatures ranging from 18 000 to 28 000 K (depending on the assumed LISM parameters). The pile-up in the neutral gas results from the deceleration and deflection of the neutral flow by charge exchange with the interstellar plasma, which is itself decelerated and diverted due to the presence of the heliosphere. Note that the charge exchange mean free path is typically less than the separation distance between the heliopause and bow shock and so a large part of the incident interstellar neutrals experience charge exchange.

Component 2, produced via charge exchange between component 1 and hot shocked solar wind plasma between the TS and HP, streams across the HP into the cooler shocked interstellar gas and heats the plasma through a second charge exchange. This leads to an extended thermal foot abutting the outside edge of the HP. This heating of the plasma by component 2 serves to broaden the region between the BS and HP, as well as to (indirectly) further heat the component 1 interstellar neutrals after subsequent charge exchanges. Some minor heating of the unshocked LISM also occurs upstream of the BS, thereby marginally reducing the Mach number of the incident interstellar wind.

The temperature of component 1 neutrals once inside the heliosphere remains fairly constant in the upstream region, and represents a substantial increase over the assumed LISM temperature. A further increase in the component 1 temperature occurs in the downstream region.

The number density of component 1 crossing the TS is approximately half the assumed incident LISM number density, an effect termed 'filtration'. The density varies only weakly between the TS and ∼10 AU from the Sun in the upstream region. In the downstream direction, component 1 densities are lower within the heliosphere. The 1D component 1 density and temperature profiles are plotted in figure 6, and the heated hydrogen wall is exhibited clearly.

A further effect of filtration is to decelerate the upstream neutral gas from 26 km s^{-1} in the LISM to ∼19 km s^{-1} at the TS in the region of the nose. Deflection of the flow also reduces the radial velocity component at angles away from the nose. Such a deceleration is in accord with observations which measure the Lyman-α resonance line scattered by neutral H in the heliosphere.

Models which assume a subsonic LISM flow do not have a bow shock. Instead, upstream of the heliopause, the LISM flow is compressed adiabatically. This more gradual compression leads to the formation of a lower-amplitude hydrogen wall that is more extended in the radial direction. It is also less extended in the transverse direction because of the localized nature of the adiabatic compression. The maximum density of the wall in the upstream direction is smaller than that of the two-shock counterpart (though still larger than the incident LISM number density). However, because it is wider, the column density is comparable with the two-shock case.

The heliosphere is less distorted along the axis of symmetry than for the two-shock case, and is smaller due to the higher assumed LISM pressure.

Three-dimensional models of the solar wind during solar minimum introduce some interesting differences from the 2D models discussed above. The presence of an anisotropic solar wind with a high-velocity, low-density, high-temperature polar stream acts to divert both the subsonic solar wind and the LISM plasma flow into the ecliptic region of the heliosphere. The presence of a high-density band of plasma about the ecliptic plane increases the filtration of neutrals compared with the higher polar regions. With an increased neutral flux of H over the poles, the polar wind experiences comparatively greater deceleration than the ecliptic wind, so reducing the degree of anisotropy seen in 3D gas dynamic simulations. In addition, the global distribution of neutral H is anisotropic in heliolatitude. Structurally, the heliosphere at low latitudes can revert to a 'bullet' shape, typically seen only in models which neglect to include interstellar neutral hydrogen. It should be borne in mind that these remarks pertain primarily to the period of solar minimum observed by the Ulysses spacecraft.

Very few simulations include either the interplanetary or interstellar magnetic fields dynamically, and those that do neglect neutral H completely or use an extremely simplified description. Nonetheless, it appears that the magnetic field in the heliosheath can acquire a very interesting structure. Along the stagnation line in the heliosheath, the flow velocity decreases approximately as r^{-2} (r the heliocentric radial distance), leading to an amplification in the azimuthal magnetic field in the region. This, together with the $\boldsymbol{J} \times \boldsymbol{B}$ force causes the equatorial current sheet to bend either upward or downward (depending on solar cycle) away from the equatorial plane. This leads to the formation of an asymmetric 3D magnetic shell or wall in the outer region of the upstream heliosheath with a gap in the neighborhood of the displaced current sheet. The region between the termination shock and the magnetic wall is occupied by solar wind originating from the middle and high latitudes whereas, thanks to the gap in the magnetic wall, the solar wind in the region between the magnetic wall and heliopause originates from the low-latitude equatorial neutral sheet region.

Inferred structure of the heliosphere

Until Voyager 2 encounters the termination shock and heliopause, we are forced to rely on remote observations to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

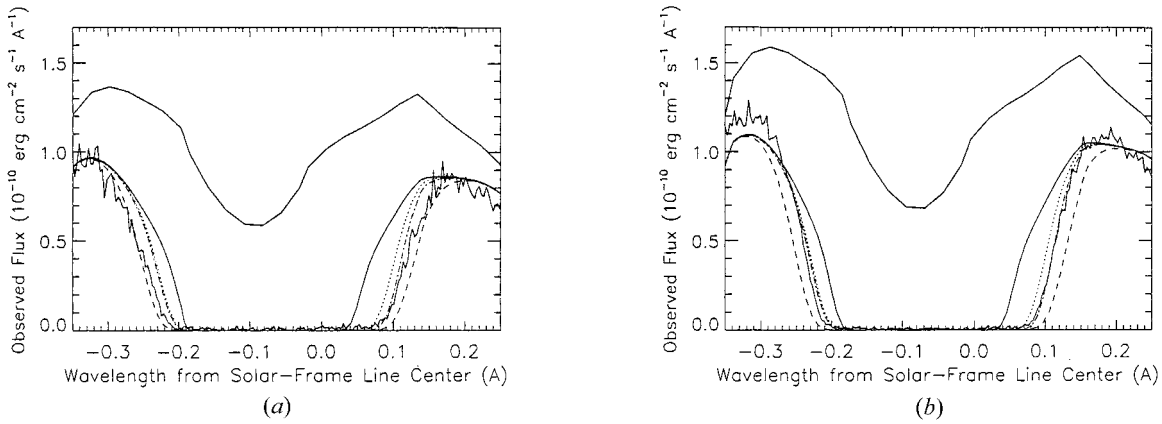


Figure 7. The solid jagged curves are GHRS Lyman-α absorption profiles towards (*a*) α-Cen A and (*b*) α-Cen B. The uppermost solid curve is the assumed intrinsic stellar Lyman-α emission profile. The smoothed solid curve which exhibits saturated absorption corresponds to the intrinsic stellar emission line after absorption from the LISM only. Corresponding theoretical absorption profiles are plotted for an $M = 1.5$ (M the Mach number) LISM (the dashed curve), an $M = 0.7$ LISM (the dotted curve) and an $M = 0.9$ LISM. All three theoretical models show that additional Lyman-α absorption on the red side results from the presence of a heliospheric hydrogen wall and a marginally subsonic model appears to fit the data best.

infer the structure of the heliosphere. The most promising approach presently is to use the Lyman-α absorption line in the direction of nearby stars. A neutral hydrogen pile-up or wall may be detectable in directions where the decelerated H is red-shifted out of the shadow of the interstellar absorption if the interstellar column density is sufficiently low. Red-shifted excess absorption in HUBBLE SPACE TELESCOPE GHRS Lyman-α observations towards α-Cen (seen previously by Copernicus and IUE) have been interpreted as evidence for the existence of the solar hydrogen wall.

Shown in figure 7 are Lyman-α observations towards α-Cen A and B (the jagged solid curves), and the wavelength scale is relative to the Lyman-α line center in the heliocentric rest frame. The upper solid curve is the assumed intrinsic stellar Lyman-α emission profile. The accurate representation of the intrinsic stellar profile is unimportant since the absorption features of interest vary sharply. The remaining smooth solid curve (the saturated absorption curve) in figure 7 gives the attenuation of the stellar Lyman-α emission by interstellar H. The column density for H was fixed by scaling to the deuterium column density, assuming the commonly accepted value of D/H = 1.6×10^{-5}. Figure 7 shows very clearly that additional absorption is required both redward and blueward of the interstellar feature if the fit is to be completed. Furthermore, the fit must be applied preferentially to the redward side, so arbitrarily changing the D/H ratio is unacceptable. The additional redward absorption has been interpreted as evidence for the detection of the hydrogen wall. The blueward absorption suggests the possibility of a hydrogen wall about α-Cen A and B.

Since the column depth of the hydrogen wall is three orders of magnitude smaller than the column depth in the LISM toward α-Cen, it may be surprising at first glance that the heliospheric optical depth at +0.1 Å is of the same order as the LISM optical depth at that wavelength. The key difference is that the hydrogen wall is heated and decelerated, which both broadens and redshifts the heliospheric component away from the −0.07 Å centroid of the LISM absorption and toward the +0.1 Å wavelength of interest.

Figure 7 also shows the Lyman-α absorption at the red edge of the LISM feature for three heliospheric models—a subsonic LISM model (dotted line), a barely subsonic LISM model (dash-dot line), and a supersonic model (dashed curve). The primary and very important conclusion to emerge from such modelling is that the synthetic Lyman-α profiles support the interpretation of the observed additional redward Lyman-α towards α-Cen as evidence for the remote detection of the hydrogen wall. Thus, it appears that the hydrogen wall has indeed been observed! The unprecedented accuracy of the Hubble GHRS observations suggests further that, with the refinement of solar wind–LISM interaction models, the question of whether the heliosphere has a one- or a two-shock structure may be answered in the foreseeable future.

The future

Evidently, many more physical processes remain to be incorporated in models describing the interaction of the solar wind with the LISM. This will undoubtedly be accomplished in the next decade as computational power increases and the underlying physics is better understood. However, the most important limitation that the field currently faces concerns the paucity of information about the state of the local interstellar environment.

Ingenious efforts have been made to obtain *in situ* and remote sensing measurements of interstellar pick-up ions and neutrals, which have advanced our theoretical understanding of the solar wind–LISM interaction considerably. Possible future missions dedicated to solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

wind/LISM questions, such as Interstellar Probe, will advance the field (and others) tremendously and in directions that we may not yet even fully appreciate.

In concluding, we should not lose sight of the fact that the interaction of the solar wind with the LISM represents only one, admittedly important, astrophysical example of the interaction of a stellar wind with a partially ionized interstellar medium. This subject, although only cursorily studied so far, may well prove fundamental to our understanding of the evolution of stellar winds in the warm interstellar medium and to the evolution of the interstellar medium itself. In this respect, the solar wind interaction with the LISM is truly a subject at the interface of space physics and astrophysics.

Acknowledgements
This work is supported in part by an NSF Presidential Young Investigator Award ATM-9357861, an NSF-DOE grant ATM-9713223, a NASA grant NAG5-6469, a Jet Propulsion Laboratory contract 959167, and a NASA Space College Grant. The NSF San Diego Supercomputer Center provided computational support.

Bibliography
Holzer T E 1979 The solar wind and related astrophysical phenomena *Solar System Plasma Physics* ed E N Parker, C F Kennel and L J Lanzerotti (Amsterdam: North-Holland) pp 101–76

Zank G P 1999 Interaction of the solar wind with the local interstellar medium: a theoretical perspective *Space Sci. Rev.* **89** 1–275

Gary P Zank

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Interplanetary Radio Bursts

Solar radio emissions at long wavelengths (λ greater than 10 m or frequencies smaller than 30 MHz) are produced in the upper CORONA and in the interplanetary medium, sometimes at very large distances from the Sun, up to the orbit of Jupiter and beyond. They are essentially produced by local instabilities, resulting from the interaction of the ambient SOLAR WIND with disturbances: streams of subrelativistic particles, and shock waves, whose origin is mostly in the solar corona; they provide a unique means to detect perturbations originating from SOLAR ACTIVE REGIONS and to track their progression through the solar wind. These emissions are produced in a frequency range over which the Earth's ionosphere transmits either poorly or not at all; a spacecraft in orbit well above the ionosphere must carry the instruments.

Solar radio astronomy at long wavelengths: specificities and problems

Electromagnetic waves at long wavelengths cannot reach the Earth because the ionosphere stops them. The cut-off frequency depends on the ionosphere peak electron density and is about 20 MHz by day and 10 MHz or less by night. Observations at long wavelengths thus require putting the radio instrument on a spacecraft well above the ionosphere. Table 1 shows a few typical frequencies and the corresponding wavelength regimes.

Table 1. The frequency/wavelength regimes.

Frequency	Wavelength	Band	Distance range from center of Sun
300 MHz	1 m	metric	lower corona up to 2–3 $R_\odot$
30 MHz	10 m	decametric	2–3 to ∼10 $R_\odot$
3 MHz	100 m	hectometric	∼10 to ∼60 $R_\odot$
300 kHz	1 km	kilometric	beyond ∼60 $R_\odot$
30 kHz	10 km		

Figure 1 shows the observed radio frequency of interplanetary radio bursts as a function of heliocentric distance (distance from the center of the Sun). A strong dependence is clearly seen, which is simply due to the fact that most radio sources consist of a narrow band emitted near the local plasma frequency and/or its harmonic. Since the plasma frequency is directly proportional to $\sqrt{N_e}$ (N_e is the local electron density), the distribution of frequencies as a function of distance closely follows that of the densities, with the usual dispersion. In a way radio observations in a large band provide a radial imaging of solar sources. This interpretation is very important, since most phenomena considered here are traveling phenomena that propagate outward from the Sun (streams of subrelativistic particles, shock waves, see below), mostly in the radial direction.

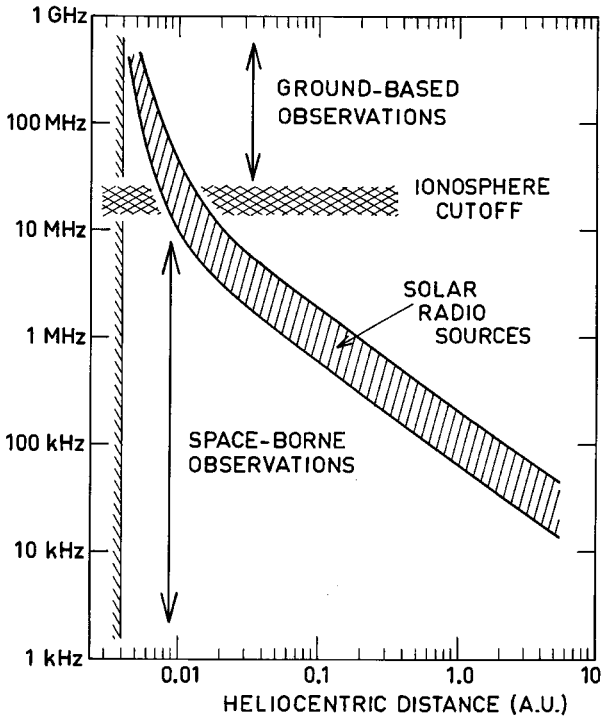


Figure 1. Heliocentric distance of radio bursts in the solar corona and interplanetary medium. The two regimes of observation—ground-based and space-borne—complement each other in order to study the physical processes involved in the radio bursts.

An important feature of radio observations from space is that they give access to sources in the interplanetary medium. In some cases, the spacecraft is even inside the radio source itself and local conditions of the solar wind plasma and of its perturbations can be measured. In the latter case, the observation of interplanetary radio bursts actually links phenomena on the Sun to *in situ* phenomena.

The antennas on spacecraft are usually wire antennas or booms from a few meters long to a few hundred meters long. Wire antennas require spinning spacecraft. The antenna length is generally short compared with the wavelength, so that they can be considered as pure dipoles, with no spatial resolution. Imaging radio sources is very difficult at low frequency (very long baseline arrays are required and have never been used as yet). Direction finding techniques are currently used on spinning spacecraft: from the analysis of the modulation of the radio signal over a spin, it is possible to deduce the source direction (direction of the source centroid), an estimate of its angular size, and its degree of circular POLARIZATION. Thus, the present radio instruments on spacecraft are essentially SPECTROGRAPHS, some of which have the direction finding capability. We note here that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

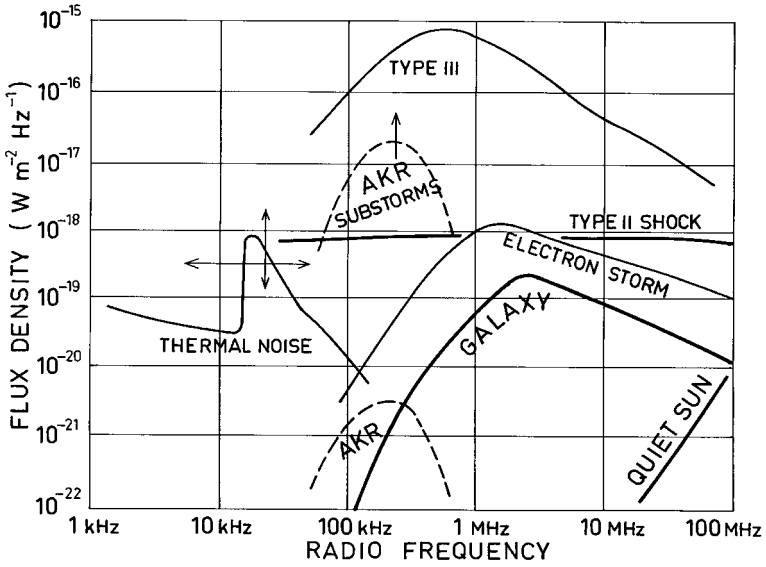


Figure 2. Flux densities of a variety of common radio sources in the range 1 kHz to 100 MHz.

the radio astronomy receivers on spacecraft are often part of more extensive radio and plasma wave investigations yielding information on *in situ* waves as well. These are observed by the same instruments and displayed on the same 'dynamic spectrum' and are sometimes essential to the understanding of radio wave production mechanisms.

Figure 2 shows instantaneous spectra of various radio events. Intense, sporadic sources are easily identified, sometimes more than 50 dB above the background. They originate from solar disturbances in the interplanetary medium and from radiation near magnetized planets, including the Earth. Above a few 100 kHz, all other sources, Galactic or extragalactic, merge into a background, which is well observed but not resolved; it is called the 'Galactic background'. Below 100 kHz, the background is determined by the electron quasithermal noise near the local plasma frequency. The radio intensity or flux density (in $\mathrm{W\,m^{-2}\,Hz^{-1}}$) is often displayed on a 'dynamic spectrum' where it is coded with a gray (or color) scale as a function of frequency and time. If time is usually displayed on the x-axis, frequency is represented on the y-axis, either increasing or decreasing upwards. In the latter case, the y-axis suggests the increasing distance from the Sun, which varies as $1/f$, since $N_\mathrm{e} \sim R^{-2}$ in the solar wind, and $f \sim f_\mathrm{p} \sim N_\mathrm{e}^{1/2}$.

The intrinsic properties of the radio source may be masked by propagation effects, on both large and small scales. Stereoscopic observations are required to assess these effects; they have shown that the radiation can be strongly beamed, sometimes in a non-radial direction. These effects provide additional information on the medium between the source and the observer. Such observations have been made in the past, either with dedicated stereoscopic missions or by combining simultaneous observations from different spacecraft on widely different orbits.

Radio signatures in the solar wind

Radio sources in the solar wind are sporadic and highly variable. A smaller variety of sources are seen than in the lower corona: only the type II (shock wave) and type III (electron stream) bursts are currently observed. Recent observations from the WIND spacecraft suggest the presence of type IV bursts and of fine structures associated with energetic particles.

A type II burst is detected in the dynamic spectrum plots as one or a few narrow bands of emission slowly drifting from high to low frequencies. The type II bursts are usually faint but can sometimes be tracked for several hours up to a day or more. Their observation proves the existence of a shock wave. Indeed, it is believed that a shock is needed to accelerate electrons to about 10 keV, where the distribution becomes unstable to the production of Langmuir waves, which are converted into radio waves through processes similar to the type III radiation. Thus, type II emission provides a unique means of tracking shocks remotely. It is particularly important in the interplanetary medium, where it can help track the progression of interplanetary shocks from the upper corona to the Earth (see table 1).

Type III bursts are produced by beams of subrelativistic electrons (1–100 keV) originating from active regions; they propagate along open magnetic field lines from the solar corona throughout the interplanetary medium. They produce the most frequent solar radio bursts in the interplanetary medium, which certainly are the most exhaustively studied and best understood. Radiation is emitted near the plasma frequency and/or its harmonic. As

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the electrons move out from the Sun, they encounter decreasing plasma densities and therefore emit at lower and lower frequencies. The radio signature shows a frequency drift seen on spectrograms—the dynamic spectrum. At a given frequency, the duration of the burst is about the time it takes for the beam to cross the corresponding 'plasma level'. The duration thus increases as the beam expands and spreads out in the interplanetary medium.

Type III bursts sometimes occur in storms which consist of thousands of overlapping individual type III bursts. These interplanetary radio storms are the extension through the solar wind of the type III storms which are seen at metric and decametric wavelengths and are themselves often associated with the type I storms seen in the metric and decimetric regime. These radio events trace long-lasting storms of suprathermal electrons injected near active regions. They may help trace the expansion of the slow solar wind above active regions.

Generation mechanisms and source structure

Radio bursts in the interplanetary medium are caused by the interaction of beams of energetic electrons—produced during solar flares and at shock fronts—with the ambient solar wind plasma. The beam free energy is converted into Langmuir waves, and then some Langmuir wave energy is converted to radio waves at the fundamental and/or the harmonic of the local plasma frequency. This is known as the plasma mechanism and can account for almost all radio bursts in the interplanetary medium. On rare occasions type IV-like events have been observed up to 10 or 15 solar radii and on one occasion a group of very intense type III-like bursts was convincingly interpreted as being due to synchrotron emission caused by energetic electrons confined in a magnetic trap created behind an interplanetary shock wave.

The size of an interplanetary radio burst may be very large. Typically, the extent of the radio sources grows as the distance to the Sun, with a factor of proportionality close to 1. As already mentioned, propagation effects may be important and may account for a significant portion of the observed source sizes—up to 50%; the source position, beaming and polarization may also be modified. Stereoscopic observations have been shown to provide essential clues in order to understand and correct those effects and even gain information from them.

It is accepted, both on observational and theoretical grounds, that the radio emission is generated in localized, unresolved regions of the interplanetary medium, rather than uniformly along the path of the electron beam for the type IIIs, or along the shock front for the type IIs. In general, interplanetary radio sources can be modeled by taking into account two essential factors: the known finite extent of the radio source and the directivity of the radio emission.

Large scale structure and dynamics of the solar wind from radio burst observations

Type III bursts and the large-scale magnetic field

The interplanetary medium is structured by the flow of the solar wind, a fully ionized, magnetized plasma. Radio bursts can provide unique information on the structure of this medium because they reflect mostly local conditions at their source.

The interplanetary magnetic field assumes an Archimedean spiral shape, due to the combination of the solar wind expansion with the solar rotation. The energetic electrons, because of their small gyro-radius, have to follow (or spiral around) the average field lines. Thus, the analysis of the source location of interplanetary type III radio bursts should, in principle, yield a large-scale mapping of the interplanetary magnetic field, with an indication of electron density along it, and a tracking of the evolution of beams of electrons over large distances. When one spacecraft only can determine the direction of the centroid of the radio source, a hypothesis has to be made on the radiation mode and on the average density model. This works quite well up to 0.5 Astronomical Units (AU) but the deduced trajectories tend to show anomalous behavior beyond this distance, because of propagation effects, increasing source sizes and proximity effects. The use of a second spacecraft helps resolve some of the latter effects and eliminate the hypothesis of a density model and radiation mode.

Radio storms and the slow solar wind

The long-lasting interplanetary radio storms (type III storms) can be seen in the interplanetary medium at a variety of radio frequencies, rotating from east to west, as the Sun rotates. The analysis of these structures yields information on the solar wind density and velocity above active regions (where they are observed). The derived velocities tend to be smaller than the average solar wind velocity, consistent with a slow solar wind above solar active regions.

Type II shocks and coronal mass ejections

The observation of a type II burst is believed to demonstrate the passage of a collisionless shock wave, which cannot generally be witnessed by any other means. The first extensive studies of interplanetary type II radio shocks were made with the ISEE-3 data (essentially 1978–1984) and now with the WIND spacecraft (launched in November 1994). Many questions related to the nature of the type II source and its association with the shock front remain unsolved, although they often are basic questions if one desires to really use the type II observations as a tool for diagnosing the shock progression and evolution in the solar corona and interplanetary medium. It is generally believed that the interplanetary shock is (piston-) driven by the SOLAR CORONAL MASS EJECTION (CME), but there is no direct observational proof as yet. This is generally supported by timing considerations. However, the estimation of the shock velocity depends strongly on the choice of the density model, which remains an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

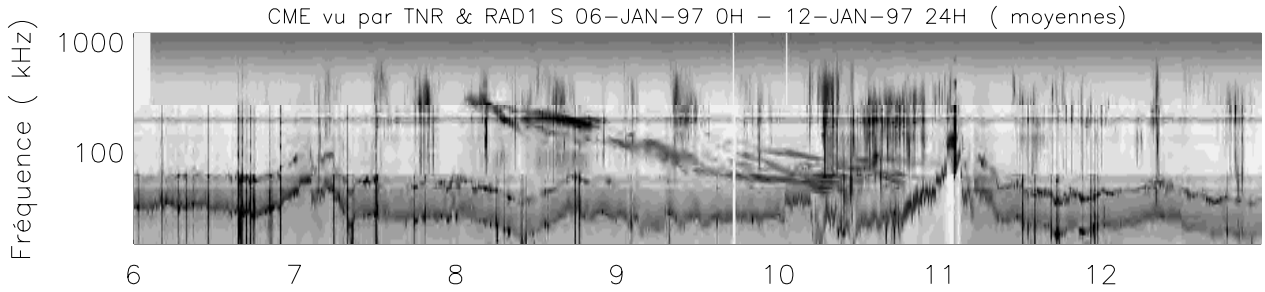


Figure 3. Dynamic spectrum obtained by the WAVES investigation on the WIND spacecraft, for the period 6–12 January 1997, and for the range from 4 kHz (*in situ* waves, bottom of *y*-axis) to 1 MHz (about 15 $R_\odot$, top of axis). The fluctuating line which runs through the plot (between about 10 and 60 kHz) is the plasma line which follows the fluctuations of the electron density at the spacecraft. The type II-like emission is seen as slowly drifting emission bands from 8 to 11 January.

unknown, and of the radiated mode, which is sometimes hard to identify. The situation is rather unclear for coronal shocks, even up to several $R_\odot$ from the Sun: some type II shocks could trace pure blast waves. Observations have not yet determined the precise location of the source with respect to the shock front, radially (upstream or downstream) or along the shock front: it is not known whether the shock radiates uniformly or if the radiation is only produced in regions where special conditions are met (for instance, parallel or perpendicular shock front, special range of Mach number, etc). There is sometimes evidence for multiple shocks, associated with single flare or CME events. The absence of spatial resolution often leads to oversimple models. Stereoscopic observations are expected to help resolve a number of these issues, in particular the location (centroid) of the radio source and the shock velocity.

Radio emissions may also reflect complex situations when the CME interacts with large-scale structures in the interplanetary medium, such as corotating interaction regions (CIR) or the heliospheric current sheet (HCS). This again may provide unique information on events that are otherwise difficult to detect but may have a strong terrestrial impact. As an example, figure 3 shows the dynamic spectrum of a very unusually strong type II-like radiation associated with the 6–11 January 1997 coronal mass ejection (CME). This event started on the Sun as a halo CME and was associated with a (faint) disappearing FILAMENT on the solar disk. Even though it is complex and hard to model, it yields definitive information on the strength of the interplanetary shock and its progression through the solar wind.

Shock-associated (SA) events

A class of large interplanetary type III bursts closely associated with meter-wave (coronal) type II bursts was proposed to be produced by electrons accelerated by shock waves in the solar corona: some events appeared to be the extension of 'herringbone structures', which are generally accepted to show evidence for shock acceleration. It was found that the shock-associated (SA) events are well correlated with major proton events, which are known to be mostly associated with interplanetary shocks.

This interpretation was challenged by some authors who argued that, in many cases, the good correlation between the SA time profiles and the concurrent microwave burst and/or hard x-ray burst strongly suggested that the SA events were caused by electrons accelerated in or near the microwave or x-ray source region, i.e. near the chromosphere rather than in coronal shock waves. Thus, there is an ongoing controversy on the origin of the electrons responsible for the SA events. Global observations of the full range of radio emissions (observed from both ground and space) associated with x-ray, EUV and optical observations are certainly necessary to fully assess the origin of the energetic particles, whose radio signature appears in the decametric and hectometric range.

In conclusion, interplanetary radio bursts provide a unique signature of energetic phenomena in the solar wind, including shock wave tracking. The difficulties encountered in their observation and interpretation testify to the complexity and richness of the radio diagnostic. The field of radio astronomy at long wavelengths is very active and much remains to be done. Stereoscopic observation is a first step towards a better tracking of disturbances in the solar wind and a better modeling of beaming and propagation effects: NASA's Solar TErrestrial RElations Observatory (STEREO) will have a major role in this step. A second step would be to do imaging at long wavelengths, which would require an array of small spacecraft. The long wavelength range is indeed the last window of the electromagnetic spectrum never to have been observed with spatial resolution.

Bibliography

Extensive information on this topic can be found in the American Geophysical Union (AGU) monograph:

Stone R G *et al* (ed) 2000 *Space-Based Radio Observations at Long Wavelengths* (AGU) to be published

which also includes papers and reviews on astronomy at long wavelengths. Propagation effects and the infor-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mation they yield on the structure of the interplanetary medium were discussed by

Hoang S and Steinberg J L 1987 Radio wave propagation in the heliosphere *Phys. Scripta* **T18** 45

For the association with CMEs and a discussion on the SA events, see the articles in AGU Monograph 99:

Crooker N *et al* (ed) 1997 *Coronal Mass Ejection* (AGU)

Jean-Louis Bougeret

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Kinetic Properties

The SOLAR WIND consists of electrons, protons, alpha particles and heavy ions. Its kinetic physics deals with the collective behavior of these particles in a statistical ensemble. Space-borne particle spectrometers enable us to measure the composition and three-dimensional velocity distribution functions (VDFs) of the particles. Vlasov kinetic plasma theory provides the adequate means for their theoretical description. Key issues of kinetic physics are to address the coronal origin and acceleration of the wind and the spatial and temporal evolution of the particles' VDFs (see SOLAR WIND: CORONAL ORIGINS, SOLAR WIND ACCELERATION). They are shaped through the forces of the Sun's gravitational field, the average-macroscale and fluctuating-mesoscale electric and magnetic fields of interplanetary space, and through multiple microscale kinetic processes such as binary Coulomb collisions and collective wave–particle interactions.

Velocity distribution functions (VDFs)

General considerations

The global solar magnetic field varies over the activity cycle, but the coronal field always consists of three main components: long-term open CORONAL HOLES (CHs), closed CORONAL STREAMERS and transiently open or closed CORONAL LOOPS. These components are respectively associated with uniform fast solar wind, filamentary slow wind, and transient variable-speed wind such as that related to SOLAR CORONAL MASS EJECTIONS. These three types differ substantially in their kinetic properties because of different solar boundary conditions and interplanetary plasma dynamics. The radial evolution of the internal state of the wind resembles a complicated relaxation process, in which the particles' free (as compared with local kinetic and magnetohydrodynamic equilibrium conditions) energy is converted into thermal, wave and turbulence energy on a wide range of scales. Spherical expansion and large-scale inhomogeneity compel the wind plasma to attain a variable state of dynamic statistical equilibrium between the particles and field fluctuations. In principle, these processes are described by the full Boltzmann–Vlasov equation for the phase-space distribution, $f_j(\boldsymbol{x}, \boldsymbol{v}, t)$, which is a measure of the number of particles in a volume surrounding position $\boldsymbol{x}$ and with velocities in a certain range around $\boldsymbol{v}$, and which evolves according to

$$\left\{\frac{\partial}{\partial t} + \boldsymbol{v}\cdot\frac{\partial}{\partial \boldsymbol{x}} + \left[\boldsymbol{g} + \frac{e_j}{m_j}\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{v}\times\boldsymbol{B}\right)\right]\cdot\frac{\partial}{\partial \boldsymbol{v}}\right\} f_j = \left(\frac{\mathrm{d}}{\mathrm{d}t} f_j\right)_{c,w} \quad (1)$$

with the interplanetary magnetic field, $\boldsymbol{B}(\boldsymbol{x}, t)$, electric field, $\boldsymbol{E}(\boldsymbol{x}, t)$ and the Sun's gravitational acceleration, $\boldsymbol{g}(\boldsymbol{x})$. Coulomb collisions or wave–particle interactions are also included and described by the Fokker–Planck collision integral or the quasilinear diffusion operator on the right-hand side of equation (1). Here the particle charge is e_j, its mass m_j, speed $\boldsymbol{v}$ and space coordinate $\boldsymbol{x}$. The speed of light is c. In addition to equation (1), Maxwell's equations have to be solved with the self-consistent current and charge density of the multicomponent wind, as calculated from the velocity moments of this equation. This complex problem has not been solved, but solutions of simplified versions of this problem, restricted to a single particle species, do exist.

Ion velocity distribution functions

The observed distributions of ions and electrons exhibit various shapes and change widely with the local *in situ* conditions, heliographic coordinates and the phase of the solar cycle. The proton VDFs range from Maxwellians in slow wind, embedding the heliospheric current sheet (HCS), to highly non-thermal ones in fast streams emanating from CHs. In fast solar wind the proton temperatures are anisotropic, with $T_{p\perp} > T_{p\parallel}$. Frequently, strong field-aligned proton beams with a drift speed larger than the local Alfvén speed occur.

The electrons are cooler than the protons in fast wind (T_e = 0.1–0.2 MK and T_p = 0.5–0.8 MK at 0.3 AU), but hotter in slow wind, which is more variable in abundance, more compressive and comparatively cold, with all particle temperatures becoming minimal at the HCS ($T_p = 5 \times 10^4$ K at 1 AU). The fast wind is permeated by Alfvén waves from coronal holes, which are broad-band in frequency and play a significant role in maintaining, through their dissipation, the ion temperatures above the level expected for adiabatic cooling. Whereas high-energy extensions are a universal property of the protons, they are less frequently seen in the alpha particles. Some typical proton distributions in fast wind are presented in figure 1, together with the radial profile of the average magnetic moment, $\mu_p = T_{p\perp}/B$, of the particles, which is displayed as a function of radial distance from the Sun. That μ_p increases radially indicates continuous ion heating perpendicular to the magnetic field in the interplanetary medium. The solid line in figure 1 shows results of a turbulence model which explains the heating by wave dissipation.

The heavy minor ions seem to trace wave effects in the wind. They stream faster than protons by the local Alfvén velocity, $\boldsymbol{V}_i \leq \boldsymbol{V}_p + \boldsymbol{V}_A$, and tend to surf on the ubiquitous Alfvén waves. This differential streaming originates in the CORONA and is observed by HELIOS to fade away with increasing heliocentric distance. Speed differences between protons and alpha particles are seen also by Ulysses beyond 1 AU and appear to be generated locally at shocks and stream interaction regions (see SOLAR WIND: ULYSSES). In fast wind, heavy ions have a high kinetic temperature, with $T_i \approx A_i T_p$ and A_i being the atomic mass number of the ion. They come in various ionization stages and have about coronal abundances in fast wind, but show element fractionation in slow wind. Helium is an exception. Its sizable relative abundance of about 3–5% on the average differentiates it from a minor constituent. In slow wind, the higher density and lower temperature enforce a state near collisional equilibrium.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

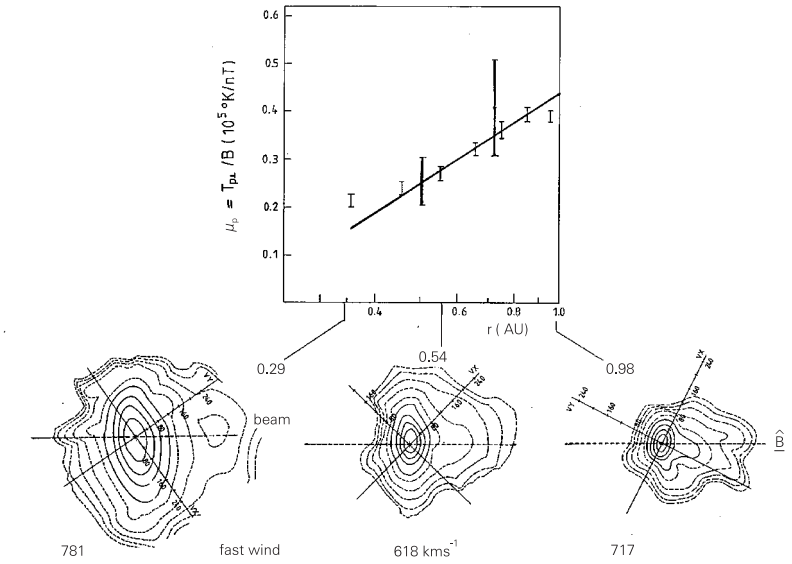


Figure 1. The observed proton magnetic moment versus heliocentric distance (top) and selected velocity distribution functions measured in high-speed wind. The solid isodensity contours correspond to 20% steps of the maximum, and the last broken contour is at 0.1%. Note the large temperature anisotropy in the core and the strong beam along the magnetic field.

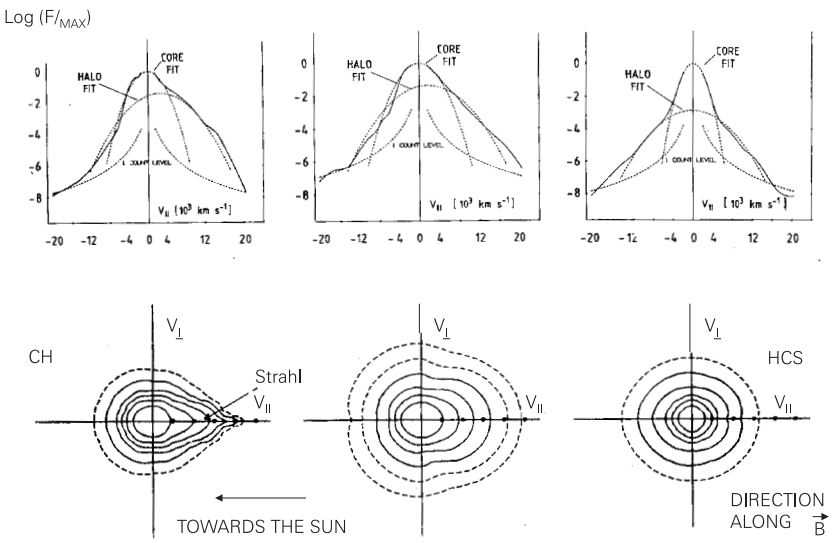


Figure 2. Electron velocity distribution functions as energy spectra (top) and velocity-space contours (bottom) for fast (left), intermediate and slow (right) wind. Isodensity contours are in 10% steps. Note the core–halo structure and the strahl of suprathermal electrons in fast solar wind.

Electron velocity distribution functions

The electrons are, because of their small masses, dynamically less important for the solar wind. Yet they ensure quasineutrality, constitute an electric field through their thermal pressure gradient and carry heat in the skewness of the bulk and the suprathermal tail of their VDFs. The distribution functions are determined mainly by the large-scale interplanetary magnetic field and the self-generated electrostatic potential, by Coulomb collisions in the thermal energy range at a few 10 eV and by various kinds of wave–particle interactions. The electrons are subsonic, i.e. their mean thermal speed considerably exceeds the solar wind (ion) bulk speed. Suprathermal electrons (at several 100 eV) may be considered as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

test particles that quickly explore the global structure of the heliospheric magnetic field. It may attain the form of open field lines anchored in CHs, of bottles and loops originating around the solar equator or of flux ropes. This topology shapes the distributions, which can observationally be decomposed into two main components, a cold and almost isotropic collisional core and a hot variably skewed halo population. The VDFs are often modeled by two convecting bi-Maxwellians as illustrated in the top part of figure 2. On open field lines in the fast wind, the halo usually develops a high-energy extension with a very narrow pitch-angle distribution only 10°–20° wide. This electron population has been named the 'strahl' and responds sensitively to the local field orientation.

Common observations of the same plasma parcel of the wind by instruments on different spacecraft, when being radially aligned, allow one to characterize the radial gradients of electron thermal parameters. The core temperature is found to vary widely between isothermal and adiabatic, while the halo temperature behaves more isothermally. The halo density falls off more steeply in dense plasma. Electron parameters have been studied by Ulysses in the distance range from 1 to 4 AU. The halo is found to represent always about 4% of the total electron number. Since there is no reason for this ratio to be constant if the halo and core particles were completely decoupled, it appears that halo particles are not entirely decoupled from the core. The break-point energy in the electron spectra of figure 2 scales on average like 7 times the core temperature, a result which is predicted by a kinetic theory for electrons mediated by Coulomb collisions. Such a value of the break-point energy is also consistent with its interpretation as being equal to the electrostatic interplanetary potential trapping the thermal electrons.

Significant progress has been made, with the Ulysses electron measurements, in the understanding of the electron heat flux regulation. An empirical heat conduction law has been suggested, in which the heat flux is carried mainly by the halo electrons and scales like the halo particle flux times the halo thermal energy. This proportionality has experimentally been verified, demonstrating that solar wind heat conduction has nothing to do with the local temperature gradient but with the thermal energy convected by halo electrons. Their drift speed is closely tied to the Alfvén velocity. This supports the evidence for a regulation of the heat flux by whistler-mode waves.

Wave–particle interactions

It is widely recognized that plasma waves in the collisionless solar corona and wind play a role similar to collisions in ordinary fluids (see also SOLAR WIND PLASMA WAVES). In the expanding inhomogeneous solar wind particle distributions develop velocity-space gradients and deviations from Maxwellians, which drive plasma instabilities and lead to wave growth or damping. The kinetic wave modes of primary importance are the ion-cyclotron, ion-acoustic and whistler-mode waves, which are the high-frequency extensions of such fluid modes as the Alfvén, slow and fast magnetoacoustic waves.

Usually, the multicomponent plasma is analysed locally for stability, by Fourier decomposing linear fluctuations of the fields into plane waves with wave vector $\boldsymbol{k}$. Each species contributes a distinct dispersion branch for kinetic waves that are weakly damped normal modes or unstable in the presence of free energy. As an example, the dispersion equation for parallel propagating left (− sign) and right (+ sign) handed circularly polarized electromagnetic waves with wavevector component $k_\parallel$ along the background magnetic field, $\boldsymbol{B}_0$, reads

$$(k_\parallel V_A)^2 = z_{k_\parallel}^2 \left(\frac{V_A}{c}\right)^2 + \sum_j \hat{\rho}_j \Omega_j^2 \, \hat{\varepsilon}_j^{\pm}(z_{k_\parallel}, k_\parallel) \qquad (2)$$

where the fractional mass density of species j has been used, defined as $\hat{\rho}_j = n_j m_j / \sum_\ell n_\ell m_\ell$. The Alfvén velocity based on the total mass density, $V_A^2 = B_0^2/4\pi\rho$, is used to normalize the phase speed properly. The plasma frequency is given by $\omega_j^2 = 4\pi e_j^2 n_j / m_j$, and the gyrofrequency, carrying the sign of the charge, by $\Omega_j = e_j B_0 / m_j c$. For a given real wavevector, $k_\parallel$, complex solutions of the frequency $z_{k\parallel} = \omega_{k_\parallel} + \mathrm{i}\gamma_{k_\parallel}$ are sought, whereby a positive $\gamma_{k_\parallel}$ signifies a plasma microinstability. The dielectric constant involves a resonance integral over velocity-space derivatives of the VDF, corresponding to the pitch-angle gradient for a particle in the wave frame as defined by the phase speed, and reads

$$\hat{\varepsilon}_j^{\pm}(z_{k_\parallel}, k_\parallel) = 2\pi \int_0^\infty \mathrm{d}w_\perp \, w_\perp \int_{-\infty}^\infty \mathrm{d}w_\parallel \frac{w_\perp/2}{w_\parallel - w_j^{\pm}} \times \left[\left(w_\parallel - \frac{z'_{k_\parallel}}{k_\parallel} \right) \frac{\partial}{\partial w_\perp} - w_\perp \frac{\partial}{\partial w_\parallel} \right] f_j(w_\perp, w_\parallel). \qquad (3)$$

The symbols $w_{\perp,\parallel}$ refer to velocity components perpendicular and parallel to $\hat{\boldsymbol{B}}_0$ in the species j proper frame, with $\boldsymbol{w} = \boldsymbol{v} - U_j \hat{\boldsymbol{B}}_0$. The prime at $z_{k_\parallel}$ indicates the Doppler shift into the rest frame of species j, drifting at bulk velocity U_j, i.e. $z'_{k_\parallel} = z_{k_\parallel} - kU_j$. The VDF here is understood to be normalized to the particle number density n_j. Numerical investigations of equation (2) have been carried out in the literature under various conditions. The multicomponent solar wind plasma is usually at the margin of microscopic stability. Interacting with a wave, a particle sees a stationary electric field if its velocity meets the condition for cyclotron resonance, where $w_\parallel = w_j^{\pm} = (z'_{k_\parallel} \pm \Omega_j)/k_\parallel$. Energy and momentum between particle and wave are exchanged as a result of this wave–particle interaction. The velocity distributions are reshaped, until the free energy in the form of temperature anisotropy, beam drift or ion differential motion, and skewness or heat flux are reduced or removed. These processes are analytically described by quasilinear theory, based on equation (1) with the fields being decomposed into means

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and fluctations, and are visualized by direct numerical simulation.

In addition to electromagnetic modes, electrostatic waves frequently occur with variable intensity in various regions of the heliosphere. Below the proton plasma frequency one has the ion acoustic modes and above the electron plasma frequency the electrostatic Langmuir oscillations and the free-space electromagnetic waves. These waves are derived from the simple dispersion equation

$$k^2 = \sum_j k_j^2 \int_{-\infty}^{+\infty} \mathrm{d}^3 w \frac{V_j^2}{\boldsymbol{k}\cdot\boldsymbol{w} - z_{\boldsymbol{k}}'} \boldsymbol{k}\cdot\frac{\partial}{\partial \boldsymbol{w}} f_j(\boldsymbol{w}) \qquad (4)$$

where k_j is the Debye wavenumber of species j, given by $k_j = \omega_j / V_j$. The thermal speed is $V_j = (k_B T_j / m_j)^{1/2}$, with the Boltzmann constant k_B. The dielectric function involves the Landau-resonance integral over the derivative of the VDF in the $\boldsymbol{k}$-direction. The dispersion relation (4) gives the wavevector k in dependence on the propagation direction $\hat{\boldsymbol{k}}$ and phase speed $z_{\boldsymbol{k}}/k$. The wavelength may vary between infinity and the very short electron Debye length $\lambda_e = k_e^{-1}$. Observationally, the occurrence of ion-acoustic waves is determined by the electron to proton temperature ratio, T_e/T_p, and the magnetic field direction. Highest wave intensities are observed around the HCS, where usually $T_e/T_p > 1$, implying weak Landau damping.

Most of the stability analyses carried out are based on idealized model distributions such as drifting bi-Maxwellians or Lorentzians with high-energy tails. Even the measured VDFs are implemented in dispersion codes to diagnose their stability or predict the associated wave activity. In summarizing our understanding of the role of microinstabilities in the solar wind, we see that the VDFs are generally found to be marginally stable but prone to instability. The four salient modes (and free energy sources) are (1) ion acoustic wave (ion beam, electron heat flux), (2) electromagnetic ion cyclotron wave (proton core temperature anisotropy), (3) magnetosonic wave (proton beam, ion differential streaming) and (4) whistler-mode and lower-hybrid wave (core–halo drift, electron heat flux). The quasilinear evolution of these wave instabilities, let alone their nonlinear evolution or possible saturation, and the associated spatial evolution of the VDFs in the nonuniform interplanetary medium have not yet been explored.

Bibliography

Feldman W C and Marsch E 1997 Kinetic phenomena in the solar wind *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: Arizona University Press) pp 615–76

Marsch E 1991 Kinetic physics of the solar wind *Physics of the Inner Heliosphere* vol II, ed R Schwenn and E Marsch (Heidelberg: Springer) pp 45–133

Marsden R G (ed) 1995 *The High Latitude Heliosphere* (Dordrecht: Kluwer) (reprinted from *Space Sci. Rev.* **72**)

Neugebauer M 1981 Observations of solar wind helium *Fundam. Cosm. Phys.* **7** 131–99

E Marsch

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Magnetic Field

The 'heliosphere' is the region surrounding the Sun that contains material and magnetic fields carried away from the Sun by the SOLAR WIND. It extends from the Sun, to the orbit of Pluto and out to the boundary of the solar wind with the interstellar medium. The 'heliospheric magnetic field' is the magnetic field that pervades the heliosphere. The 'interplanetary magnetic field' is that part of the heliospheric magnetic field which occupies the region between the planets, near to the ECLIPTIC plane.

The heliospheric magnetic field has been measured directly by instruments carried on several spacecraft which explored the region from 0.3 AU to beyond 65 AU (1 AU is the distance from the Sun to Earth). The heliospheric magnetic field has also been measured over a range of solar latitudes from $\approx 80^\circ$ S to $\approx 80^\circ$N in the region between 1 AU and 5 AU.

The magnetic field and the velocity of a fluid are vector fields, specifying a vector at each point in the region of interest. At a given instant, a vector field has a direction and a magnitude (magnetic field strength and speed respectively) at each point of the region. The direction of the vector field at any point is specified by two angles relative to a coordinate system whose origin is at that point. The location of a point in the heliosphere (e.g. a spacecraft position) is specified by another coordinate system, whose origin is at the center of the Sun.

When one places a piece of paper with a thin layer of iron filings on top of a bar magnetic, each filing lines up with the direction of the magnetic field at the filing's location. One can see a pattern in the set of iron filings, an imaginary set of curves that reveals a dipole pattern for the magnetic field lines.

Parker spiral

E N PARKER introduced a model for the basic structure of the heliospheric magnetic field based on the following assumptions:

(a) The solar wind moves radially away from the Sun at a constant speed V.
(b) The Sun rotates at an angular speed Ω, given by 2π divided by the solar rotation period.
(c) The solar wind is azimuthally symmetric about the solar rotation axis.
(d) The interplanetary magnetic field is frozen-in the solar wind and anchored at the Sun. This means that the solar wind, which flows radially relative to an inertial frame, flows along the magnetic field lines when viewed from the rotating Sun.

Consider a point in the heliosphere given by the coordinates R (the radial distance from the Sun), θ (the elevation angle above the solar equatorial plane) and ϕ (the azimuthal angle) in a Sun-centered coordinate system. At that point one can define a local coordinate system with unit basis vectors $\boldsymbol{r}$, θ, ϕ pointing in the radial direction, the meridional direction and the azimuthal direction respectively. Parker showed that, for a stationary magnetic field, the components of the magnetic field along these directions are given by the following expressions:

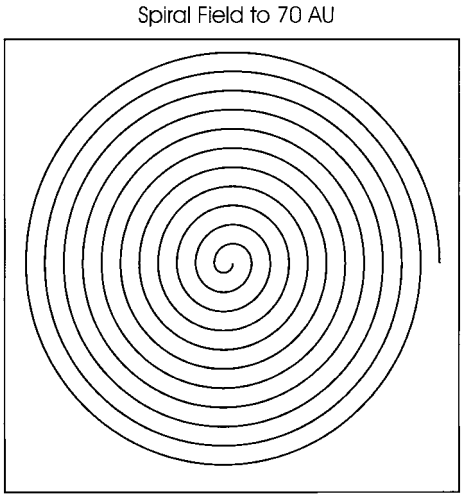


Figure 1. A spiral magnetic field line from the Sun to 70 AU according to Parker's model of the interplanetary magnetic field.

$$B_r(R,\theta,\phi) = B_r(R_1,\theta,\phi - R\Omega/V)(R_1/R)^2 \quad (1)$$

$$B_\theta(R,\theta,\phi) = 0 \quad (2)$$

$$B_\phi(R,\theta,\phi) = B_r(R_1,\theta,\phi - R\Omega/V)(R_1\Omega/V)(R_1/R)\cos\theta. \quad (3)$$

Here R_1 is a reference radius (typically either 1 AU or a few solar radii), and B_r is related to the strength of the magnetic field near the Sun. Thus, the radial component of $\boldsymbol{B}$ decreases as R^{-2}, the azimuthal component decreases as R^{-1}, and the meridional component is zero.

From the components of the magnetic field vector given above, one can calculate that a magnetic field line in the solar equatorial plane has the form of an Archimedes spiral, which is the shape of a rope coiled around itself on a surface. This is illustrated in figure 1 for a magnetic field line extending from the Sun to 70 AU in the solar equatorial plane. The angle between the magnetic field and the radial direction in the equatorial plane is determined by $\arctan(B_\varphi/B_r)$. Near the Sun, the heliospheric magnetic field is approximately radial. At the orbit of Earth (1 AU) the magnetic field line makes an angle of 45° with the radial direction. Far from the Sun the spiral is tightly wound and the magnetic field is nearly orthogonal to the radial direction. A magnetic field line at a heliographic latitude θ has the form of a spiral wrapped on a cone whose axis is the solar rotation axis and whose angular half-width is $90^\circ - \theta$. Over the poles of the Sun, $B_\phi \ll B_r$, unless R becomes very large. The magnetic field wraps around the same degrees of longitude regardless of the latitude.

Observations in the interplanetary medium from the Sun to Pluto and within $\approx \pm 30^\circ$ of the ecliptic have confirmed the spiral pattern predicted by Parker.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

However, the observations by the Ulysses spacecraft (which surveyed latitudes between $\pm 80^\circ$ at relatively low levels of solar activity) suggest that the magnetic field lines might not be wrapped on cones of constant latitude, but rather extend from the polar regions toward the equatorial regions (see SOLAR WIND: ULYSSES). This model was introduced to explain the observations that (a) the radial component of the magnetic field did not vary with latitude along the trajectory of Ulysses, and (b) energetic particles that were presumably accelerated by shocks at low latitudes and large distances were observed at high latitudes closer to the Sun. The new model should be regarded as preliminary, and its applicability to periods of high solar activity and to large distances remains to be determined.

Sector structure and the heliospheric current sheet

Measurements of the direction of the magnetic field in the solar wind at a point near the orbit of Earth often show the field pointing away from the Sun (positive magnetic polarity) for approximately half a solar rotation and pointing toward the Sun (negative polarity) for the remaining part of the solar rotation. Each interval of uniform (positive or negative) magnetic polarity is called a 'sector', and the magnetic polarity pattern observed during one or more solar rotations is called the 'sector pattern'. The 'two-sector pattern' just described is typically observed during the period of declining solar activity. Near solar minimum, a 'four-sector pattern' is frequently measured during a solar rotation. The four-sector pattern consists of four sectors of alternating positive and negative polarity, each passing the spacecraft in approximately one-quarter solar rotation. Since the solar rotation period is approximately 27 days as seen by a spacecraft near the orbit of Earth, the duration of a sector is approximately 14 days and 7 days for a two-sector and four-sector pattern respectively. Near solar maximum the polarity pattern of the magnetic field is complex, and a sector pattern is not always present.

Assuming that the solar wind is stationary, the temporal pattern of magnetic polarities observed at a given location during a solar rotation is equivalent to a spatial pattern of magnetic polarities. Figure 2 illustrates the spatial configuration of a four-sector pattern. Each sector is a region of uniform polarity in the ecliptic plane (the plane in which the planets move), extending from the Sun to at least the orbit of Earth and bounded by Archimedes spirals. Thus, each sector in the ecliptic is topologically equivalent to the shape of a piece of a pie, hence the name sector.

The origin of the sector pattern is ultimately related to the large-scale pattern of the magnetic field at the Sun. The two-sector pattern is related to a dipolar magnetic field configuration and the four-sector pattern is related to a quadrupole component. Consider the singular case of a magnetic dipole whose axis is exactly along the solar rotation axis. The solar wind draws the field lines of the dipole into spiral field lines, but the polarity of the field

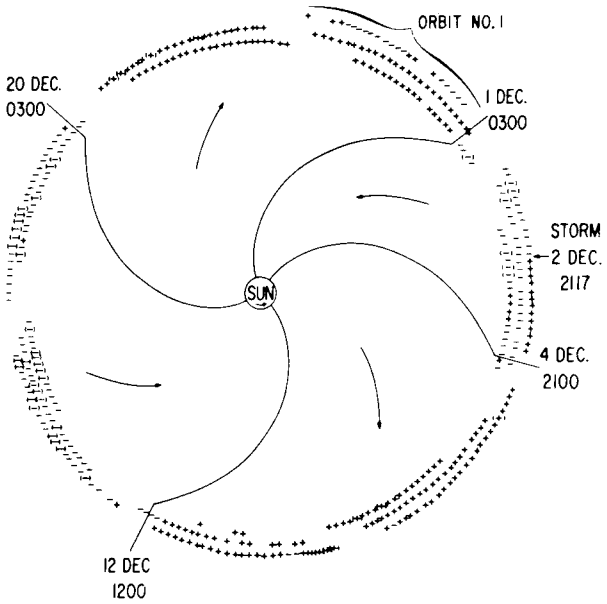


Figure 2. The four-sector pattern of the polarity of the interplanetary magnetic field.

lines is preserved in this process. Consider a SOLAR CYCLE for which the magnetic fields in the northern hemisphere have positive magnetic polarity. Then the heliospheric magnetic fields above the solar equatorial plane would have positive magnetic polarity and those below the solar equatorial plane would have negative polarity. The solar equatorial plane in this idealization is a singular surface at which the magnetic field strength is zero and across which the magnetic field polarity changes.

The Sun's rotation axis is tilted 7.5° with respect to the ecliptic, so that the solar equatorial plane is inclined relative to the ecliptic plane. Since the solar equatorial plane rotates with the Sun, an observer in the ecliptic plane will be below the solar equatorial plane for half a solar rotation and above it during the second half of the solar rotation. When the dipole axis coincides with the rotation axis, an observer in the ecliptic will alternately observe positive and negative magnetic polarities; he will measure a two-sector pattern. An observer located at heliographic latitudes $>7.5^\circ$ or $<-7.5^\circ$ would observe a uniform magnetic polarity. The sector structure would not be observed above or below the zone within $\pm 7.5^\circ$ of the solar equator, the 'sector zone'.

A second model for the sector pattern was discussed in the early literature on sectors. In this model, the sectors extend from pole to pole, like the pieces of an orange. According to this model, an observer would see a sector pattern at all latitudes.

In both models described above, the sectors are bounded by surfaces, which are planes in the region near the Sun. In the dipole model, the plane is the solar equatorial plane. In the orange model, the plane passes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

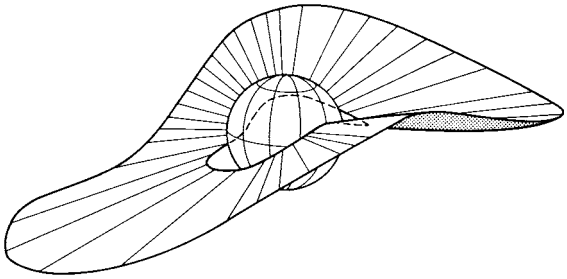


Figure 3. The shape of the heliospheric current sheet in the region within several solar radii of the Sun.

through the dipole axis. Obviously, these two models are diametrically opposed and mutually exclusive.

A third model of the heliospheric sector structure, which avoids the extreme and singular assumptions of the two models described above while retaining some features of each, was proposed many years ago. This model focuses on the form of the surface separating the positive and negative sectors, the 'heliospheric current sheet' or 'HCS'. According to this 'HCS model' the structure of the HCS can vary with the phase of the solar cycle. Near solar minimum and close to the Sun, the surface is approximately a plane that is basically the equatorial plane of the solar magnetic dipole. However, the surface can be distorted by an additional quadrupole component of the solar magnetic field such that it has folds near the Sun (figure 3). A four-sector pattern (see figure 2) can be produced by such a warped surface. The northern and southern extent of the sector pattern can exceed 7.5° when the dipole axis is tilted with respect to the solar rotation axis and when the quadrupole distortion of the HCS is large. The northern extent can differ from the southern extent; i.e. an asymmetry in the sector pattern can also be produced by quadrupole components of the solar magnetic field.

The heliospheric current sheet model has been confirmed by measurements at various latitudes. When a spacecraft is at sufficiently high latitude during the period of decreasing solar activity and near solar minimum, it will observe a single magnetic polarity, while a spacecraft in the sector zone at lower latitudes will observe a sector pattern at the same time. This result has been demonstrated by observations from HELIOS 1 and 2, PIONEER 11, VOYAGER 1 and 2, and Ulysses. For example, figure 4 shows a situation in which Voyager 1 was above the sector zone and Voyager 2 was in it. Voyager 1 observed a single magnetic polarity above the sector zone, while Voyager 2 observed both positive and negative magnetic polarities within the sector zone. This figure illustrates a hypothetical meridional cross section of the HCS, which appears as a wavy line. In reality periodic crossings of the HCS are not observed at large distances, owing to dynamical interactions in the outer heliosphere that distort the HCS. However, the latitudinal extent of the HCS, marked by the dotted lines in figure 4, does not change greatly with increasing distance, since the solar wind expansion is approximately

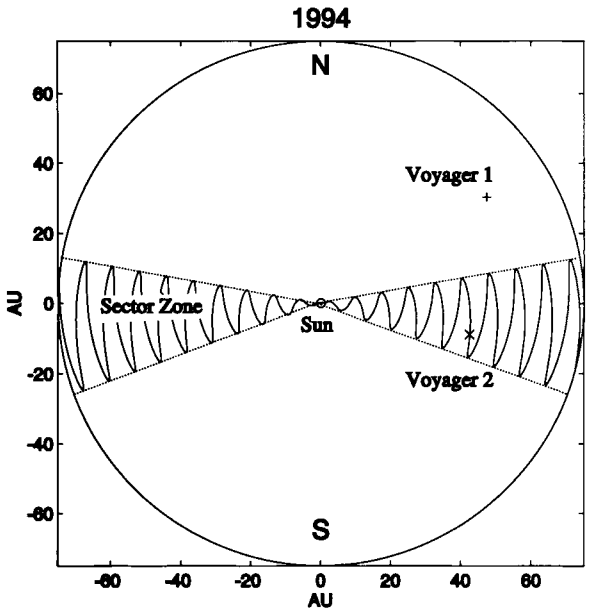


Figure 4. The sector zone illustrated here is the region in which both positive and negative polarity regions (sectors) are observed, not necessarily periodically as suggested in the idealized sketch here. Above the sector zone the polarity is positive (negative), and below the sector zone the polarity is negative (positive).

radial. Thus, the sector zone is well defined even when a periodicity of the sector pattern is not observed.

The foot-points of the heliospheric current sheet on an imaginary surface centered on the Sun and with a radius of a few solar radii (the 'source surface') can be calculated by extending the measurements of the photospheric field to the source surface. There are various ways to do this, but the simplest procedure assumes that the magnetic field is radial on a spherical source surface, and that there are no electric currents in the region between the photosphere and the source surface. Then one can use the photospheric magnetic field measurements to calculate a curve on this surface where the magnetic polarity changes from positive to negative. This curve is called the 'neutral line', and it corresponds to the foot-points of the HCS on the source surface.

Solar cycle dependence

Magnetic field strength

The strength of the interplanetary magnetic field varies with the solar activity, being weakest near solar minimum and strongest near solar maximum. This is illustrated in figure 5. The middle panel of figure 5 shows the SUNSPOT number, which is a measure of solar activity, as a function of time from 1977 to 1996. The bottom panel of figure 5 shows yearly averages of the magnetic field strength measured by spacecraft at 1 AU as a function of time. The magnetic field strength varies approximately sinusoidally (as illustrated by the dashed curve), from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

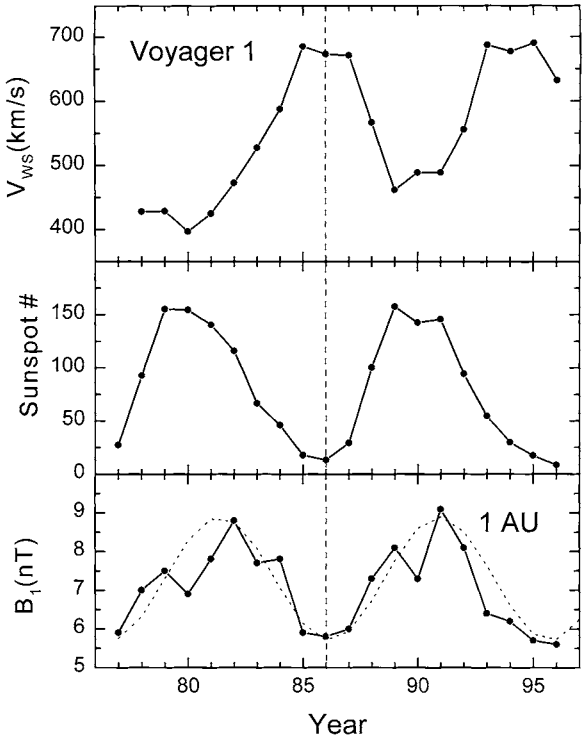


Figure 5. The Sunspot number versus time (middle panel) illustrating the solar activity cycle. The magnetic field strength at 1 AU (bottom panel), which is a measure of the solar magnetic field strength, varies with the solar activity cycle. The estimated solar wind speed along the trajectory of Voyager 1 (top panel) also shows a solar cycle variation.

≈6 nT near solar minimum to ≈9 nT near solar maximum. Thus, the strength of the magnetic field varies by 50% relative to its minimum value during the course of the solar cycle.

The solar cycle variation of the heliospheric magnetic field could be a consequence of two factors. First the quasi-stationary, large-scale magnetic field (e.g. the dipole and quadrupole components) might vary with solar activity. Second, transient ejecta have relatively strong magnetic fields and are more frequent when the Sun is more active, so that they will contribute to the variation in the magnetic field strength over the solar cycle.

Sector structure

Both the calculations based on the source surface model and the *in situ* observations show that the amplitude of the excursions of the neutral line on the source surface, and hence that of the HCS, increase with solar activity. Near solar minimum, the HCS is close to the solar equatorial plane. The maximum deviation of the HCS from the equatorial plane increases as solar activity increases. In other words, the width of the sector zone increases with increasing solar activity.

The sectors and the sector structure are less well defined and less persistent when solar activity is more intense. Intrusions of 'foreign' magnetic polarity are produced by ejecta from the Sun, which are associated with dynamical solar processes such as SOLAR FLARES and CORONAL MASS EJECTIONS. For example, there are often regions of negative polarity in positive sectors when the Sun is active, often making it difficult to identify a given sector or sector boundary. Nevertheless, an observer in the sector zone will typically measure a polarity distribution with two distinct peaks corresponding to positive and negative polarities at the spiral field directions even when the Sun is active.

Two possible sources for the magnetic fields in the sectors were considered in the early literature: (*a*) compact local sources (nozzles) and (*b*) the entire surface of the Sun. It is now known that the truth is somewhere between these extremes. The quasi-steady streams of uniform magnetic polarity, which often fill most of the heliosphere, originate in CORONAL HOLES, which are regions where the magnetic field lines are open to the heliosphere and the density is low. The coronal holes do not occupy the entire surface of the Sun, and the fraction of the solar surface that they occupy decreases with increasing solar activity. However, coronal holes do cover the polar caps during most of the solar cycle, they can extend from the polar regions to the equatorial regions over a limited range of longitudes in each hemisphere, and isolated coronal holes are occasionally observed near the solar equator. The origin of the magnetic field lines in the slow solar wind is not completely understood.

Elevation angles of $\boldsymbol{B}$

The elevation angle $\delta = \arctan[B_\theta/(B_r^2+B_\phi^2)^{1/2}]$ is the angle between the magnetic field vector and its projection on the $\boldsymbol{r}-\varphi$ plane. For a spacecraft in the solar equatorial plane, δ is the elevation angle of the magnetic field with respect to that plane, positive for a northward directed magnetic field. According to Parker's model, $\delta = 0$ at all latitudes because $B_\theta = 0$. However, there are always fluctuations in the magnetic field direction. The elevation angles of $\boldsymbol{B}$ have a Gaussian distribution, but the average and the most probable values of the distribution are zero at any epoch of the solar cycle.

Transient ejecta from the Sun carry magnetic fields whose configuration is not that of a Parker spiral. There is a class of ejecta called magnetic clouds for which the magnetic field direction rotates smoothly through a large angle, and in many cases the rotation tends to be from either south to north or vice versa. For most ejecta, the magnetic field configuration is more complex, but there is generally a relatively large meridional component B_θ of the magnetic field. When the Sun is active, there are many ejecta in the heliosphere and the distribution of elevation angles is relatively broad. The width of the distribution of elevation angles varies with the solar cycle, being smallest at solar minimum and largest at solar maximum. It is likely that this variation is primarily the result of a solar cycle variation in the frequency with which ejecta leave the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Magnetic field strength

Parker's model

According to Parker's model, the strength B of the magnetic field at any point in the heliosphere should be the square root of the sum of the squares of the components of $\boldsymbol{B}$ given by equations (1) to (3) above. Thus, Parker's model can be used to describe the large-scale heliospheric magnetic field strength to lowest order. In this case, one considers averages of the magnetic field over one or more solar rotations, and the ϕ dependence drops out of the equations. The source-field strength might change with both the time t throughout the solar cycle and the heliographic latitude θ. Similarly, the solar wind speed V changes with time t and θ. Thus, according to Parker's model, the strength of the magnetic field varies as

$$B_p(R,t,\theta) = B_{r1}(t,\theta)R^{-2}[1 + (419.5R\cos(\theta)/V(t,\theta))^2]^{1/2} \quad (4)$$

when the rotation period of the Sun is assumed to be 26 days independent of latitude, and the units of θ, V and R are degrees, km s^{-1} and AU respectively. Here, the factor $B_{r1}(t,\theta)$ is the radial component of the magnetic field at 1 AU as a function of θ and time, which we measure in units of a year. The first term in the brackets on the right-hand side of (4) is from the radial component of the magnetic field, which decreases as R^{-2}, and the second term is from the tangential component B_ϕ, which varies as R^{-1}.

Since the Ulysses data indicate no latitude dependence of the radial component of the magnetic field during the high-latitude passes, we shall assume that there is no latitude dependence in the source term at all phases of the solar cycle, i.e. we take $B_{r1}(t,\theta) = B_{r1}(t)$. From (4) with $R = 1$ and $\theta = 0$ we obtain $B_{r1} = B_1(t)\alpha(t)$, where $\alpha(t) \equiv [1+(419.5/V_1(t))^2]^{-1/2}$. $V_1(t)$ and $B_1(t)$ are the speed and magnetic field strength respectively, measured at 1 AU over the course of a solar cycle. Thus, in the absence of a latitude dependence of B_{r1}, Parker's equation (4) becomes

$$B_p(R,t,\theta) = \{B_1(t)[1 + (419.5/V_1(t))^2]^{-1/2}\} \times R^{-2}[1 + (419.5R\cos(\theta)/V(t,\theta))^2]^{1/2}. \quad (5)$$

Note that (5) has no free parameters, since $B_1(t)$, $V_1(t)$ and $V(t,\theta)$ can be determined from measurements, as discussed below.

Beyond a few AU, (5) has the simple form

$$B_p(R,t,\theta) = B_1(t)\alpha R(t)^{-1}\cos(\theta)(419.5/V(t,\theta)). \quad (6)$$

In particular, beyond 5 AU the difference between (6) and (5) is only 2% at $\theta = 0$. This means that beyond several AU the radial component of the magnetic field is negligible and the magnetic field strength is determined by the azimuthal component of the field. The magnetic field strength would vary with heliocentric distance as R^{-1}, if all other factors were constant. In general, however, the magnetic field strength at a spacecraft such as Voyager 1 will vary with time t (in particular, with the solar cycle) and with θ, because it is proportional to the source term $B_1(t)$ and inversely proportional to the solar wind speed $V(t,\theta)$ at the spacecraft.

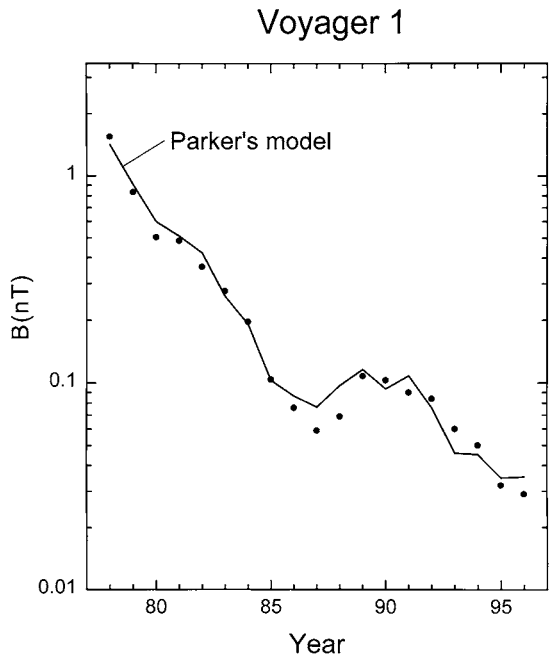


Figure 6. The variation of the heliospheric field strength measured by Voyager 1 (full circles) between 3 AU and 65 AU from 1978 to 1996 is described by Parker's model (solid curve).

The results for B_p presented in the following paragraph are based on equation (5), but essentially the same results were obtained from (6) with $\alpha = 0.72 \pm 0.02$. This value of α is obtained with $\langle V1\rangle = 446 \pm 32$ km s^{-1} which is computed using measurements of speed at 1 AU from 1978 through 1996. The value $\alpha = 0.72$ is consistent with a spiral angle of 45° at 1 AU. The small uncertainty in α indicates that it is not sensitive to variations of $V_1(t)$, which is why (6) is an excellent approximation to (5).

Measurements of the variation of the magnetic field strength by Voyager 1 as it moved from ≈3 AU to ≈65 AU are shown in figure 6. The measurements were made from 1978 to 1996. If the magnetic field strength at 1 AU did not vary with the solar cycle, and if the speed were constant with time and latitude, then one would expect from (6) that B would decrease monotonically with increasing time as the spacecraft moved away from the Sun. The bottom panel of figure 5 shows that the magnetic field at 1 AU does vary with the solar cycle. The top panel of figure 5 shows that the speed at Voyager 1 also varied with time, as a result of the increasing latitude of the spacecraft and the solar cycle variation of the speed. When these variations are considered, Parker's model (6) predicts the result shown by the solid curve in figure 6. Parker's model, which was introduced 40 years ago, provides an explanation for the observed magnetic field strength out to at least 60 AU.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

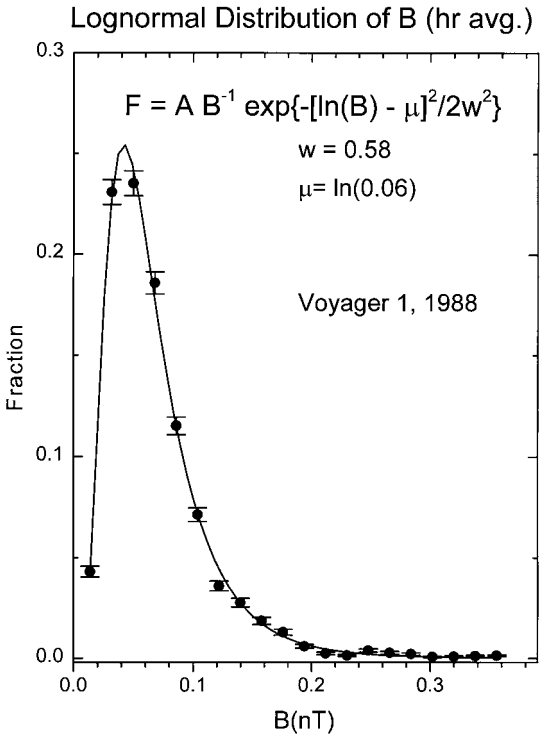


Figure 7. The distribution of hour averages of the magnetic field strength (full circles) is generally a log-normal distribution (solid curve).

Magnetic field strength fluctuations

The heliospheric magnetic field strength is highly variable over a large range of scales. To a first approximation the distribution of hour averages of the magnetic field strengths observed during a given year is log-normal, i.e. the distribution of $\ln B$ is a normal distribution. This result is illustrated for the distribution of hour averages measured by Voyager 1 during 1988, shown in figure 7, where the points are the observations and the curve is a best fit to a log-normal distribution. A log-normal distribution of B implies that relatively strong magnetic fields are observed far more frequently than one would expect for a normal distribution. The strongest magnetic fields are produced by the compressions of the front of streams, by shocks and by the ejection of plasmas carrying strong fields from solar disturbances. Log-normal distributions of the hour averages of B have been observed from 1 AU to 65 AU, from -20° to 30° in heliographic latitude, and at all phases of the solar cycle. The variance of the log-normal distribution of B relative to the mean magnetic field for a given year tends to be constant, independent of solar activity and position in the heliosphere. This means that near solar maximum, when the heliospheric magnetic field strength tends to be high, the variance in the log-normal distribution of B is also high. It implies that the intervals with the strongest magnetic fields occur near solar maximum, when transient ejecta are most abundant and when departures of the elevation angle of the magnetic field from Parker's model are greatest.

The distribution of magnetic field strength generally shows a scaling symmetry that is described by a multifractal distribution over a large range of scales. This means that the magnetic field strength has a fractal pattern in various parts of the structures carried by the solar wind, but the fractal dimension can vary within these structures. The situation is analogous to a landscape containing mountains and river basins. The fluctuations in altitude can vary greatly from place to place, much more so than one would expect from a Gaussian distribution of elevations such as one sees on the surface of a lake. Yet the shape of some large regions resembles the shape of smaller regions, etc. The dynamical causes for these invariant properties of the magnetic field strength distributions throughout the heliosphere are not yet known.

Bibliography

Burlaga L F 1995 *Interplanetary Magnetohydrodynamics* (New York: Oxford University Press)

Hundhausen A J 1972 *Coronal Expansion and Solar Wind* (New York: Springer)

Hundhausen A J 1977 An interplanetary view of coronal holes *Coronal Holes and High Speed Wind Streams* ed J B Zirker (Boulder, CO: Colorado Associated University Press) pp 225–329

Parker E N 1963 *Interplanetary Dynamical Processes* (New York: Wiley-Interscience)

Sheeley N R Jr, Wang Y-M and Phillips J L 1997 Near-sun magnetic fields and the solar wind *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: University of Arizona Press) pp 459–84

Smith E J 1997 Solar wind magnetic fields *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: University of Arizona Press) pp 425–58

Leonard F Burlaga

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Manifestations of Solar Activity

The Sun's outer atmosphere, the corona, is continually heated and expands to create the solar wind. Solar activity waxes and wanes with the 11 yr cycle of the Sun's magnetic field, and these activity changes also affect the solar wind. The brightest structures in the corona are a bright belt of streamers, the source of the slowest solar wind. These streamers are occasionally disrupted as coronal mass ejections (CMEs) which spew large amounts of gas and magnetic fields into the solar wind. CMEs appear to be a fundamental way in which the Sun loses mass and releases the magnetic flux and field twist built up during the cycle.

The solar magnetic field and the solar wind

The Sun's magnetic field controls much of its surface activity which evolves with an 11 yr period. The field tends to simplify with height in the corona, but the corona, and its outward expansion as the solar wind, also evolves on time scales related to the evolution of the surface field. The simplest magnetic configuration of the Sun occurs during activity minimum. At this time the Sun's magnetic field can be approximated as a dipole whose axis is tilted slightly with respect to the axis of rotation. Large CORONAL HOLES cover the Sun's polar regions and extend toward the equator. These open-field regions are of opposite magnetic polarity. The polar holes are separated by a wide band centered over the heliomagnetic equator within which lie the magnetically closed structures near the surface, i.e. active regions and sunspots. CORONAL STREAMERS extend over these regions forming a bright band around the Sun. The high-speed wind flowing from the polar holes constricts the oppositely directed fields over the streamers to a narrow current sheet which has the appearance of a ballerina's skirt (figure 1). This rotating, warped heliospheric current sheet (HCS) appears as a sector boundary crossing at Earth, such that during one solar rotation the Earth will be immersed in a minimum of two large sectors of opposite polarity each with the relatively high-speed wind of its parent hole. During the solar cycle new magnetic flux emerges at higher solar latitudes leading to a more complex and disturbed corona. The streamer belt and heliomagnetic equator become more complicated and more inclined to the rotation axis, and the magnetic sector structure observed in the solar wind is more chaotic. Near maximum activity a simple, single current sheet surrounding the Sun may not even exist.

On a global scale the slow solar wind is confined to the streamer belt and the HCS (see SOLAR WIND: GLOBAL PROPERTIES). It now appears that this band is the source of most transient activity which affects the heliosphere, except possibly near sunspot maximum. Near the solar surface most activity occurs on smaller scales within regions of strong magnetic fields (see ACTIVE REGIONS) in the form of flares and active region loop expansions. However, in the corona the most rapid, dramatic and energetic form of evolution occurs in large events called CMEs. Most CMEs tend to cluster within $\sim 30^\circ$ of the coronal base of the HCS over the cycle. Thus, there is a close relationship among coronal streamers, the HCS and CMEs and their evolution. Since CMEs dominate the transient activity in the solar wind, they are the focus of this article.

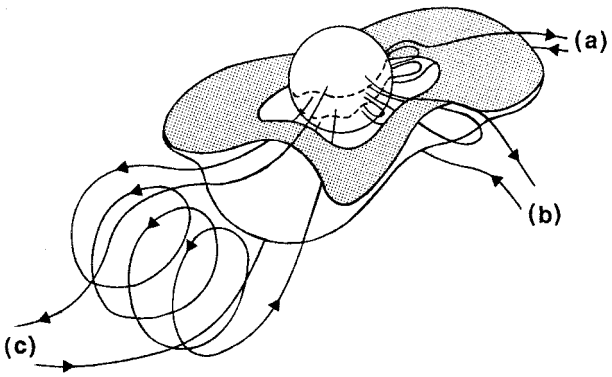


Figure 1. Schematic model of the coronal streamer belt as a disk of variable thickness forming a conduit for transient outflow from (*a*) steady-state helmet streamers, (*b*) small-scale ejections and (*c*) large-scale CMEs, some with flux rope structure. After Crooker *et al* (1993).

CMEs are immense structures of plasma and magnetic fields that are expelled from the Sun into the heliosphere. As such they are the nearest example on a large scale of the common process of magnetohydrodynamic (MHD) expulsion of material from objects throughout the universe. Although the range of CME speeds is large, even the slowest CMEs (slower than the in-ecliptic solar wind) never fall back to the Sun. This suggests that their internal magnetic fields control the dynamical evolution of CMEs near the Sun. CMEs can cause large geomagnetic storms at Earth and their attendant effects, such as auroral displays. Faster CMEs drive transient interplanetary shocks, which in turn can accelerate solar energetic particle (SEP) events. At the Sun the onset of CMEs can be associated with both flares and filament eruptions, but most flares occur without CMEs and some CMEs cannot be associated with any surface activity. Other articles discussing the origins and early evolution of CMEs are SOLAR CORONAL MASS EJECTION: OBSERVATIONS, SOLAR CORONAL MASS EJECTION: THEORY, SOLAR PROMINENCE ERUPTION and SOLAR FLARES: RELATION TO CORONAL MASS EJECTIONS.

CMEs near the Sun

CMEs are best viewed in white light from space-borne coronagraphs; these reveal that, even near the Sun, a CME can dwarf the solar disk. CMEs often appear to have a bright leading looplike structure followed by a dark cavity and a bright core of denser material, suggesting the eruption of a pre-existing prominence, its overlying coronal cavity and the ambient corona (figure 2). Prominences are elongated regions of dense, cooler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

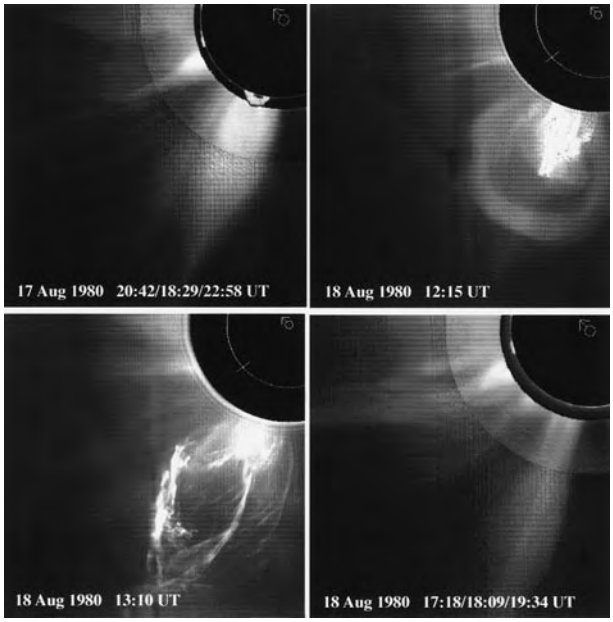


Figure 2. A CME and prominence eruption on 17–18 August 1980. The first and fourth panels are composites of an Hα (inner) image from Mauna Loa Observatory showing the prominence, an MLO K-coronameter image of the mid white light corona from 1.2 to 1.8 solar radii, and a solar maximum mission (SMM) coronagraph image of the white light corona from 2 to 5 solar radii. The second and third panels show only SMM data. The inner circle on each panel marks the solar limb and the arrow points to solar north. From Hundhausen (1997).

material suspended in the low corona in helical magnetic fields (see SOLAR PROMINENCE MODELS). Large prominences are surrounded by dark voids called cavities which have less coronal material but may be shaped by strong, helical magnetic fields (discussed later). There is a large range in the basic properties of CMEs. Their speeds, masses and energies range over 2–3 orders of magnitude, and their widths exceed by factors of 3–10 the sizes of flares and active regions. Occurrence rates of CMEs vary in phase and amplitude with the sunspot cycle.

CMEs arise in large-scale, closed structures, such as coronal streamers, where the magnetic field is strong enough to prevent the plasma from expanding outward. The temporal and latitudinal distributions of streamers and prominences are similar to those of CMEs, being confined to low latitudes about the current sheet near sunspot minimum and becoming broadly distributed near maximum. This evolution is very different from that of active regions, flares or sunspots, which migrate from mid to low latitudes during the solar cycle. Many energetic CMEs result from the disruption of a pre-existing streamer, which increases in brightness and size for days before erupting as a CME (figure 2). Possible causes of such disruptions include the emergence through the surface of new magnetic flux, the dynamical evolution of arcades and the shearing of field lines.

Signatures of CMEs in the solar wind

The CME's plasma is entrained on magnetic field lines and transported into the solar wind, where it can be detected by remotely measuring the plasma along the line of sight or by measuring directly its properties as the material passes over a spacecraft. CME plasma in the solar wind has been remotely detected by white light photometers on the Helios spacecraft and by measurements of interplanetary scintillation (IPS) of galactic radio sources from the ground.

The two Helios spacecraft were placed into highly elliptical solar orbits ranging from 0.3 to 1 AU from the Sun and operated from 1974 to 1983. The on-board photometers were able to produce crude images of $\approx$200 CMEs out to the orbit of Earth. The occurrence rates, widths and speeds of these CMEs were consistent with those of the coronagraph observations within 30° of the Sun. In these views of CMEs in the inner heliosphere, the CMEs are obviously expanding but otherwise their shapes are unchanged, with two brighter 'legs' behind and flanking a tenuous leading front. The interplanetary (IP) masses and energies of the Helios CMEs were factors of 2–10 higher than those determined by the earlier coronagraphs. However, the more sensitive measurements from the SOHO LASCO coronagraphs now indicate that mass flows out of the corona behind the leading edges of CMEs for a prolonged period of time, yielding higher masses consistent with the Helios observations. CMEs can contribute a sizable fraction of the total mass and energy flux of the solar wind, especially at solar maximum and in or near the ecliptic. The mass contribution ranges from a few percent at minimum to possibly as much as 40% at maximum.

The IPS technique detects transient disturbances in the solar wind as enhanced scintillation of distant astronomical sources in the sky along the line of sight. The scintillation is caused by small-scale turbulence in the wind due to the enhanced density of a transient such as a CME. By observing daily changes in the scintillation of a large number of sources distributed over the sky, crude maps of IP transients can be produced. A related technique involves the tracking and analysis of signals from deep-space spacecraft such as Pioneer Venus, Ulysses and Galileo as the extended solar corona transits between the spacecraft and the Earth. This technique is particularly sensitive to the enhanced density and density fluctuations in the compressed plasma between a CME and a shock if there is one and in the CME itself. However, these techniques have poor spatial and/or temporal coverage and are most useful when combined with other data sets on a particular event.

Most of our understanding of CMEs in the solar wind has come from direct measurements of the plasma and magnetic fields as the CME material passes over a spacecraft which is situated outside of the Earth's magnetosphere. A variety of signatures have been proposed as identifiers of CME ejecta in the solar wind. Table 1 summarizes those signatures which

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Signatures of CMEs in the solar wind.

Transient interplanetary shock waves
He abundance enhancements
Unusual ionization states (e.g. He^+)
Brief density enhancements and longer-duration density decreases
Proton and electron temperature depressions
Bidirectional field-aligned flows of halo electrons and/or low-energy protons
Magnetic field variations usually associated with magnetic 'clouds' or flux ropes:
Strong magnetic field
Smooth field rotation
Low plasma β
Low field strength variance

are different from the normal solar wind and usually observed for at least several hours. These include shock waves, density changes, depressed temperatures, flows with enhanced helium abundances and magnetic field structures consistent with looplike topologies. Although the front and sides of a CME are visible near the Sun because they are denser than the ambient solar wind, expansion of the CME can yield lower than average densities further from the Sun. Many of these signatures were first identified in the plasma which followed an IP shock by several hours and was considered to be the piston (CME) driving the shock (see figure 3). Some signatures can also be observed elsewhere in the solar wind where they may identify relatively slower CMEs not driving shocks.

More sensitive instruments show that the abundances and charge state compositions of elements and ions are systematically different in CME flows as compared with other kinds of solar wind (see SOLAR WIND COMPOSITION). As the solar wind–corona expands outward from the Sun, the electron density decreases so rapidly that the plasma becomes collisionless and the relative ionization states become constant, thus reflecting the conditions in the corona where this occurs. The charge states of so-called 'minor ions' ($Z > 2$) in CME flows usually suggest slightly hotter than normal coronal conditions at their origin (i.e. >2 MK). However, mixtures of charge states and, therefore, ionization temperatures are also found. In addition, transient flows often exhibit element and ion abundances (especially He and Fe) that are enhanced relative to the nominal solar wind. The coronal composition of transient flows are also seen in the systematic enhancement of elements with low first ionization potentials (FIPs) (see TRANSITION REGION: FIP EFFECT). Abundances of elements with FIPs <10 eV are enhanced by a factor of ∼4 relative to those high-FIP elements, an effect which is typical of coronal, not photospheric material. Finally, the Ulysses spacecraft (see SOLAR WIND: ULYSSES) showed not only that is there a latitude dependence on the speed of IP CMEs but also that their compositional dependence is different. The He and minor

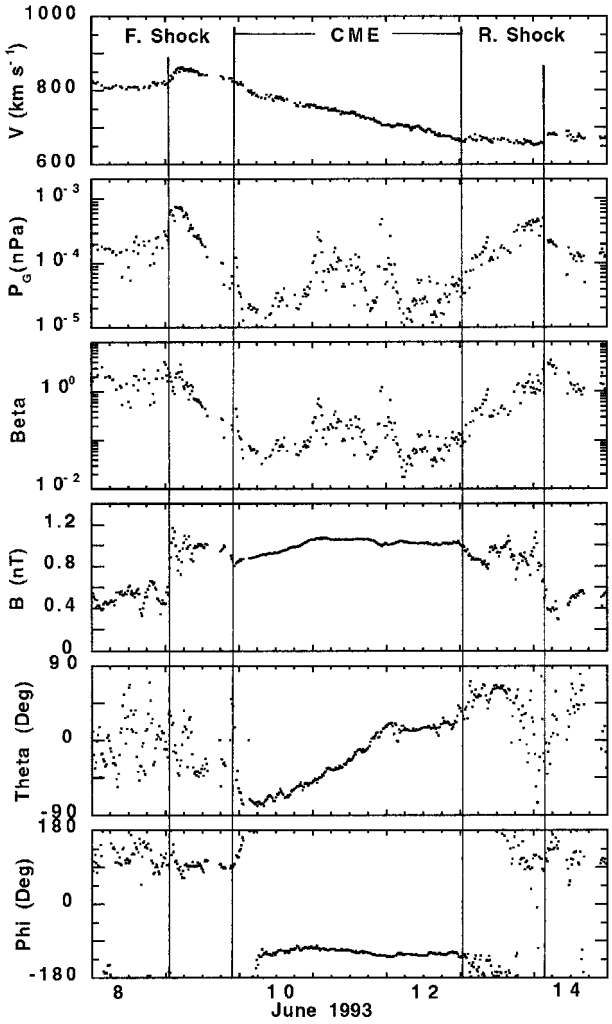


Figure 3. Ulysses plasma and magnetic field observations of an expanding CME–magnetic flux rope in the solar wind. From top to bottom the parameters plotted are bulk flow speed, plasma pressure (ions plus electrons), plasma β, magnetic field strength, field polar angle and field azimuthal angle. The bar marked 'CME' defines the region of bidirectional electron streaming. Ulysses was at 4.6 AU from the Sun and at a heliographic latitude of S33°. From Gosling (1996).

ion abundances and charge states become more and more similar to coronal hole flows the higher the ecliptic latitude.

Infrequently, unusually low ionization states of He and minor ions such as O, Fe and C are detected in CME flows. Although rarely observed before this decade, enhanced He^+ flows have been detected in at least four Earth-directed CMEs with the SOHO and WIND spacecraft during 1997–1998. In each of these events an erupting filament–CME could be associated with an unusually dense and narrow 'plug' of cool plasma at WIND in the trailing edge of a magnetic cloud. It is likely that such material is the solar wind signature of the filament itself and, therefore, is chromospheric in origin.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Although the speeds of the leading edges of CMEs near the Sun exhibit a wide range from ∼50 to >2000 km s^{-1}, the average speed is slightly less than the typical solar wind speed in the ecliptic. The speeds of most CMEs beyond several solar radii appear constant, indicating that they are rapidly accelerated low in the atmosphere. In the ecliptic within 1 AU of the Sun it appears that the fastest CMEs decelerate as they interact with the slower ambient plasma, while the slowest CMEs are accelerated, possibly by the same forces acting on the normal solar wind. At high heliographic latitudes the speed profiles of CMEs are different. Using bidirectionally streaming events observed at Ulysses as CME proxies, the speeds of high-latitude CMEs (730 km s^{-1} on average) were much higher than in the ecliptic plane and comparable to the surrounding solar wind flow characteristic of polar coronal holes (figure 3). This again suggests that CMEs are accelerated along with the surrounding solar wind.

In a hydrodynamic plasma regime, a dense region moving through a uniform, background flow at a sufficiently faster speed will produce a shock wave which stands out ahead of the driver (see SOLAR WIND SHOCK WAVES AND DISCONTINUITIES). Such shocks are frequently observed throughout the universe, but our detailed understanding of them comes from direct measurements of their properties made in the interplanetary medium and in Earth's bow shock. Within 1 AU of the Sun the strongest shocks are driven by CMEs whose leading edges are often a few hundred km s^{-1} faster than the normal wind speed. A spacecraft sees a forward-moving shock as a discontinuous increase, or jump, in most plasma and magnetic field parameters. In figure 3 a shock at Ulysses is encountered first (left) and stands off from the leading edge of the CME driver by about 20 h. Between the shock and the CME lies a turbulent region where the ambient solar wind plasma and interplanetary magnetic field (IMF) are compressed ahead of the CME. Depending on the signatures used, approximately 1/3–1/2 of all ejecta in the solar wind within 1 AU follow IP shock waves. When these shocks first develop is unclear, but there is evidence that most CMEs moving outward in the low corona with speeds greater than the local Alfvén speed (∼400 km s^{-1}) are associated with coronal shock waves. Coronal shocks are remotely detected as Moreton waves at the surface and metric radio type II bursts in the corona. At heights and densities typical of the solar wind, shocks can be tracked outward at kilometric wavelengths (see SOLAR WIND: INTERPLANETARY RADIO BURSTS).

Simple hydrodynamic model simulations have been successful in reproducing the basic speed and pressure profiles of shocks and CMEs even out to large distances from the Sun. Usually a simple pressure pulse of appropriate duration and amplitude is initiated into a uniform ambient flow with characteristic solar wind parameters. As the front of the faster CME overtakes the slower wind, a strong gradient develops and pressure waves eventually steepen into a forward shock propagating into the ambient wind ahead and, occasionally, a reverse shock that propagates back through the CME towards the source (the Sun). Because of its interaction with the ambient plasma, the CME slows until it eventually 'rides along' at the solar wind speed. This may not occur until several AU from the Sun. A different class of forward–reverse shock pairs was observed by Ulysses at high latitudes. In the event shown in figure 3, the speed declined from the front to the rear of the CME and the pressure was maximum immediately downstream from the shocks. Since the CME was not traveling faster than the solar wind ahead of it, the shocks were not caused by the relative motions between the CME and ambient wind but by expansion of the CME due to its high internal pressure. Simulations show that such features can be reproduced by the ejection of a dense CME into a uniform, high-speed solar wind. Initial propagation pressure waves steepen into shocks at a distance of about 3 AU.

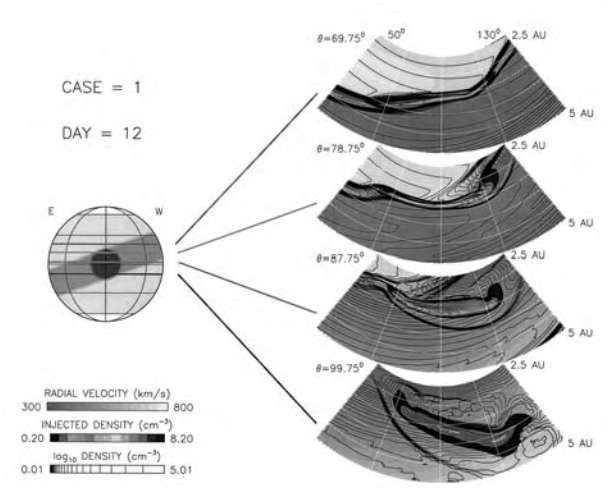


Figure 4. Results of a 3D simulation of a CME moving through a typical ambient solar wind. Here the CME is injected in the center of the HCS–streamer belt (left), specified as a high-density, low-temperature, slow-speed structure tilted to the solar rotation axis. The CME is represented as a 30° wide, 12 h duration pressure pulse. The resulting CME interaction with the solar wind flow is shown as the slices in heliolongitude and at distances 2.5–5 AU from the Sun 12 days after launch. The slices are through four different heliolatitudes and show that the CME's shape, pressure and speed will vary significantly depending on the characteristics of the ambient medium it encounters. Courtesy V Pizzo, NOAA Space Environment Center.

In recent two-dimensional (2D) and 3D hydrodynamic simulations attempts are being made to include more realistic solar wind conditions in which to inject the transient. For example, if an initially large but homogeneous CME is injected into a more realistic, latitudinally inhomogeneous solar wind, it will evolve very differently in the fast and slow wind regimes. Figure 4 demonstrates one such 3D simulation wherein a pressure-pulse CME is injected into a bimodal solar wind which is slow and dense

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

at low latitudes and fast and tenuous at higher latitudes. The CME extends across the boundary between the two regimes and is distorted into two sections. The parts separate radially owing to the strong velocity shear between the slow and fast ambient wind and latitudinally as a result of pressure gradients associated with rarefaction regions as the CME moves outward. This simulation also incorporates features of the wavy 'ballerina skirt' HCS which adds to the complexity of the interaction in the longitudinal direction. Full 3D MHD simulations using the observed photospheric magnetic field as an inner boundary condition have been successful in replicating gross coronal structure near the Sun, as well as the initiation of transients by shearing of the magnetic footpoints. However, none of these efforts is as yet sufficiently developed to reproduce the complex magnetic and plasma structure of the actual solar wind or the transients that flow through it.

Adding to the complexity of the real solar wind are interactions which occur between the transient flows, including their shock waves, and pre-existing faster or slower flows and the interaction regions bounding them. For example, a fast CME can overtake a slower CME or a fast flow can overtake a slow CME. In the inner heliosphere these interactions produce compound streams. These systems can continue to evolve and merge with other CMEs and shocks as they move outward. Eventually, well beyond the 5 AU orbit of Jupiter, such structures can form what are called merged interaction regions. These regions can become very extensive, sometimes essentially encircling the Sun like a distant belt. Such regions can diminish the flux of high-energy particles which continually stream into the heliosphere, called galactic COSMIC RAYS, causing solar-cycle-dependent decreases in their flux.

Energetic particle events associated with solar activity (SEPs) are also commonly associated with IP shocks (see SOLAR WIND: ENERGETIC PARTICLES). Solar activity produces two classes of SEPs, gradual and impulsive, depending on their energy versus time profiles. About 10 gradual events occur per year near maximum and each can last for several days. Although flares were once considered the source of SEP events, we now know that most of these particles are accelerated directly from the coronal and solar wind population by CME-driven IP shocks. Impulsive events are much more frequent, ∼100 per year near maximum, but last for only a few hours. They are generally much weaker than the gradual events and are only observed in space when a spacecraft or Earth intersects the pre-existing field lines which connect back to the flare–active region site. These are also called ^{3}He-rich events because the ^{3}He/^{4}He ratio in them is much higher than in the normal solar wind. During impulsive events is about the only time we can sample flare material and its associated accelerated particles in the solar wind.

The magnetic field topology of CMEs

Recent research has focused on those solar wind signatures indicative of the topology of the ejected magnetic fields.

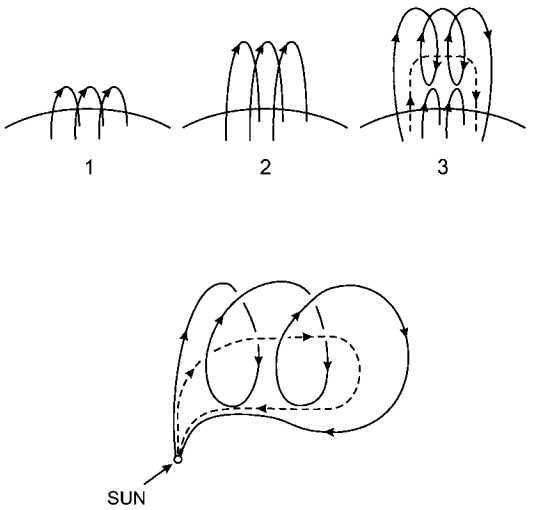


Figure 5. Schematic diagram of a reconnection model of magnetic flux rope formation at the Sun. In the top row from left to right rising, sheared field lines reconnect to form a rising flux rope and a closed arcade at the surface. Bottom: the flux rope structure as it would appear later and further out in the solar wind. Adaptation courtesy of J Gosling.

Most such studies involve observations and modeling of magnetic clouds and bidirectionally streaming particle flows (see below). Magnetic clouds are long-lived solar wind flows having enhanced field strengths which exhibit smooth, coherent rotations. Figure 3 shows an example of such a cloud embedded in an expanding CME; note the strong and smooth magnetic field strength, B, the smooth rotation of the polar field angle, θ and the low plasma β (ratio of plasma to magnetic pressure). Such a structure can be modeled as a force-free flux rope, which is a series of helical field lines, like the coils of a spring (figures $1(c), 5$) with pitch angles increasing towards the outer edge.

Some magnetic clouds have been associated with solar filament disappearances. Since filament plasma is embedded in helical, horizontal magnetic fields, the close association of CMEs with filament eruptions and shearing fields near the surface has led to the modeling of CMEs as flux ropes. One idea is that the interior fields of a rising, sheared CME reconnect, resulting in an ejected flux rope and new, closed coronal loops at the Sun (figure 5). (Magnetic reconnection is a dynamic plasma process whereby field line systems interact, resulting in a change in their topology.) At least 1/3 of all CMEs in the solar wind appear to be flux ropes but this fraction may be higher considering the limited sampling available from a single spacecraft. In several studies magnetic clouds have been found to have the same orientation and polarity as associated erupting filaments at the Sun. Furthermore, larger filaments always have twist in the same sense in a given hemisphere, even though the hemispherical polarity reverses every solar cycle. These suggest that the sign of the magnetic helicity (or sense of twist) of the erupted structure in the solar wind can be predicted from a given

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solar filament eruption. Filament eruptions and CMEs may be important ways that the Sun sheds magnetic helicity as well as flux built up over the solar magnetic cycle.

Although CMEs clearly involve the ejection of magnetic fields from the Sun, our understanding of the strength and topology of these structures is poor. Partly this is because instruments such as coronagraphs view the optically thin coronal material in projection against the sky. The circular CME structures we see (figure 2) are likely to be 2D projections of 3D structures such as arcades, flux ropes or shells. A prominence and its associated helmet streamer can be modeled as a dual flux system, one part of which is a flux rope and its surrounding coronal cavity which may help to drive the eruption once the streamer is disrupted. It appears that the prominence itself may be only a small part of this system, lagging the bulk of the CME–flux rope in the solar wind. This seems to be supported by the recent WIND observations.

In the normal solar wind field lines are 'open', i.e. connected to the hot corona at only one end. This results in an outward flux of hot electrons from the Sun streaming along the IMF. Bidirectional flows of these electrons are occasionally observed in which the electrons flow in both directions along the field lines. These are interpreted as evidence that the associated transient field lines are closed with both footpoints connected either to the Sun (a bottle) or to each other (a plasmoid) and, thus, a good proxy for CMEs in the solar wind. Bidirectional proton flows have also been detected, often but not always in association with shocks and ejecta flows. Although both magnetic clouds–flux ropes and bidirectional flows are usually considered good evidence for closed structures and CMEs, it is still difficult to ascertain the field topology of a given structure, i.e. entirely closed (bottle or plasmoid), entirely open or a hybrid (flux rope).

Reconnection of magnetic field lines is thought to be an important energy release and heating process near the surface of the Sun, especially in flares. The disconnection from the Sun of the field lines stretched outward during CMEs and their associated reconnection in the low corona may also be fundamental to the ejection process. In a popular eruptive flare model, field lines stretched open during the eruption of a prominence and CME reconnect near the surface to form a magnetic arcade, accounting for the long-duration optical and x-ray events commonly associated with CMEs (e.g. figure 5). Other observations, such as the ejection of plasmoid-like structures and evidence for newly forming streamers beneath CMEs, support this concept. Some degree of disconnection–reconnection of transient field lines seems necessary to prevent a continual increase of the net IP magnetic flux, which is not observed. For instance, it has been estimated that, in the absence of any reconnection, CMEs should double the field magnitude in the ecliptic every 9 months. There are several ways of alleviating this problem. First, recent LASCO observations suggest that the occurrence of structures within CMEs having a concave outward appearance is much greater than recorded with older, less sensitive coronagraphs. Nearly half of LASCO CMEs show evidence for such features, which are prime candidates for loops disconnecting from the Sun. (However, in some CMEs these structures could be the projection of a rising pre-existing cavity–flux rope in the skyplane which would, therefore, not require reconnection–disconnection.) Second, there are a few observations and a model of outward-moving 'U-shaped' structures suggesting the disconnection of *previously open* field regions at the top of streamers, resulting in the expulsion of a detached structure open to the heliosphere. These might be associated with so-called 'heat flux dropout' events, intervals in the solar wind devoid of the normal halo electron flux. Third, it is evident that the bidirectional streaming signature is not always present throughout all parts of an IP CME, or can be missing entirely from CMEs identified by other signatures such as a magnetic cloud. This suggests that some field lines within the CME may be 'open' in the sense that they do not have both ends connected back to the Sun. Although the CME may begin at the Sun as a pre-existing flux rope connected to the surface at two ends, continual small-scale reconnection associated with the CME may produce a mixture of closed, open and disconnected field lines as it moves through the solar wind.

As mentioned earlier, CMEs tend to arise in coronal streamers which form a belt that encircles the Sun and is the base of the HCS. The HCS is essentially the heliomagnetic equator of the Sun, separating the opposite-polarity fields in the northern and southern hemispheres. A recent model suggests that the base of the HCS may often be broad, encompassing multiple helmet streamers and their current sheets. Most CMEs might then be spatially associated with the HCS (figure 1). In this view the HCS and its associated plasma are more dynamic than previously thought, acting as a conduit for a range of activity from slowly evolving streamers to large CMEs.

Indeed, it appears that in many, if not most, cases, CME–flux ropes in the solar wind expand through the current sheet in such a way that they carry the sector boundary, or polarity change, in the form of a large-scale rotation rather than a radially expanding current sheet (figure 1(*c*)). Reconnection near the Sun then re-establishes the current sheet behind the CME. In the solar wind following transients at sector boundaries there is evidence of small intertwined flux ropes which might result from disconnection and distension of the fields. In addition, quiet, radially aligned fields are often found in the trailing portions of CMEs and may be the legs of magnetic structures carried out by the CME. The Helios photometer images and the bidirectional electron data also indicate that the closed-field regions of CMEs can extend well beyond the magnetic cloud portion.

Finally, such transient activity can also compress, amplify and align pre-existing magnetic discontinuities in the HCS and at the leading edge of high-speed streams. In addition, since the closed fields of a CME effectively

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

isolate its interior from the surrounding solar wind, the ambient IP field and its associated plasma will be deflected around the oncoming CME. This interaction causes the ambient field to be compressed and to drape about the CME, much as a moving body in a fluid pushes the ambient medium aside. Compression will increase the ambient field strength and draping can reorient its direction, especially ahead of a fast CME.

The effects of CMEs on the solar wind also help explain why CMEs cause the largest geomagnetic storms. Geomagnetic storm indices are well correlated with the speed of the solar wind and the strength and direction (southward) of the IMF (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS) parameters, which are also enhanced during the passage of CMEs. For example, compression and draping of magnetic fields at the leading edge of CMEs and of the ambient IP field are prime causes of strong southward fields. Good associations are also found between storms and other IP proxies of CMEs, in particular strong southward fields in magnetic clouds and compound streams (formed by faster streams overtaking slower ones), rapid decreases in the galactic cosmic ray flux, enhanced He abundance events and filament disappearances which can be isolated from flares. Storms are often preceded by sudden commencements which are well correlated with the IP shocks driven by faster CMEs. In general, CMEs which are associated with storms are among the largest and fastest structures in the solar wind. Since the occurrence rate of CMEs follows the solar activity cycle so that they are much more frequent around activity maximum, it is not surprising that geomagnetic storms are also more frequent and stronger during maximum.

Other sources of transient activity in the solar wind

Finally, we briefly mention several other kinds of solar activity that may have an influence on the solar wind. There is some evidence, although controversial, from global IPS measurements that solar active regions provide a source of small-scale density variations in the quiet solar wind. If they directly add denser-than-ambient plasma to the flow, then active regions can contribute to the mass flux of the solar wind. Previously, active regions had been thought to be small-scale, entirely closed magnetic structures with no direct access to the solar wind. The revisionist view is supported by x-ray images from the YOHKOH satellite showing that loops overlying active regions are continually expanding outward in the corona. There is also some evidence that open field lines overlie some active regions and could act as conduits for mass flow into the wind. Although it is well known that most of the material in the slow component of the quiescent solar wind originates from the vicinity of the coronal streamer belt, the details of where this material comes from and how it is transported outward are unknown. Recent results from LASCO, IPS and total solar eclipse data suggest that the streamer belt is very filamentary and highly structured. Blobs of material can now be observed being ejected from the cusps of streamers, probably through a quasi-steady reconnection process. Although the fast component of the solar wind, known to originate from the open fields of coronal holes, has low density, it does contain small-scale plasma structures whose source is not known. It has long been conjectured that the source of this material might be POLAR PLUMES, bright, radially elongated structures that emanate over polar coronal holes. Plumes often, but not always, have bipolar magnetic fields at their base, suggesting that reconnections with the prevailing open fields drive plasmoid-like structures into the wind. Recent observations from SOHO show upward motions of plasma in plumes, supporting this view.

Bibliography

Burlaga L F E 1991 Magnetic clouds *Physics of the Inner Heliosphere* vol 2 ed R Schwenn and E Marsch (Berlin: Springer) pp 1–22

Crooker N U, Siscoe G L , Shodan S, Webb D F, Gostling J T and Smith E J 1993 Multiple heliospheric current sheets and coronal streamer belt dynamics *J. Geophys. Res.* **98** 9371–81

Crooker N, Joselyn J and Feynman J (ed) 1997 *Coronal Mass Ejections (Geophysical Monograph 99)* (Washington, DC: American Geophysical Union)

Gosling J T 1996 Corotating and transient solar wind flows in three dimensions *Annual Review of Astronomy and Astrophysics* vol 34 ed G Burbidge and A Sandage (Palo Alto, CA: Annual Reviews) pp 35–74

Hundhausen A J 1997 Coronal mass ejections *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: University of Arizona Press) pp 259–96

Webb D F 1995 Coronal mass ejections: the key to major interplanetary and geomagnetic disturbances *Rev. Geophys., Suppl.* 577–83

D F Webb

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Radio Techniques for Probing

The solar wind is a complex magnetized plasma containing large-scale magnetohydrodynamic (MHD) structures, waves and turbulence (see SOLAR WIND PLASMA WAVES and SOLAR WIND TURBULENCE). The structure of the solar wind is modulated in both time and space by solar variability. The solar activity cycle modulates the structure of the solar wind over a time scale of years while transient energetic phenomena such as coronal mass ejections (see SOLAR CORONAL MASS EJECTION: OBSERVATIONS) can strongly perturb the solar wind on short time scales.

Many direct and remote sensing techniques are available for exploring the basic structure of the solar wind and its variation in time and space. Direct sensing techniques require *in situ* measurements by instrumentation on board spacecraft located outside of the Earth's magnetosphere. Remote sensing techniques employ a variety of ground-based and/or space-based instrumentation. A powerful class of remote sensing techniques exploits the use of radio waves to probe the solar wind. Beginning in the early-1960s, observations of radio waves from natural radio sources (e.g. quasars), spacecraft radio beacons and radar signals have been used to deduce the solar wind velocity and constrain its acceleration profile, to map three-dimensional MHD structures in the heliosphere, to determine the spatial spectrum of electron density inhomogeneities and to place constraints on the magnetic field in the inner heliosphere.

Scattering of radio waves in the solar wind

The solar wind is a plasma composed of electrons and ions, the bulk of the ion population being protons. The propagation of radio waves in the solar wind is determined by the refractive index of the medium, μ. For rarefied plasmas such as the solar wind and the wavelengths typical of radio experiments designed to study the solar wind, the refractive index $\mu(\boldsymbol{r})$ at position $\boldsymbol{r}$ is $\mu(\boldsymbol{r}) = 1 - r_e\lambda^2 N_e(\boldsymbol{r})/2\pi$, where $N_e(\boldsymbol{r})$ is the electron number density, λ is the wavelength of the radio radiation and r_e is the classical electron radius. In addition to the overall decline in electron density with distance from the Sun, there are variations in the plasma density due to large-scale MHD disturbances and turbulence which cause fluctuations in the refractive index: $\delta\mu(\boldsymbol{r}) \propto \lambda^2\delta N_e(\boldsymbol{r})$. The mean radial gradient in N_e can cause significant refraction of radio waves. Fluctuations in the refractive index $\delta\mu$ cause a variety of additional effects that are generally referred to as 'scattering'.

Scattering phenomena depend on $\delta\mu^2$. The mean density of the solar wind declines nearly as $N_e(\boldsymbol{r}) \propto r^{-2}$. With $\delta N_e(\boldsymbol{r}) \propto N_e(\boldsymbol{r})$ we have $\delta\mu^2 \propto r^{-4}$. Hence radio waves propagating through the solar wind are most strongly affected when they are closest to the Sun, as illustrated in figure 1. It is therefore useful to think of the solar wind as a thin scattering screen between the radio source and the observer. Consider a distant background source of radio waves—a QUASAR, for example. The radio waves incident on the thin screen are plane waves. The effect of density inhomogeneities in the plasma screen is to change the phase ϕ of the incident wave in a position-dependent manner: $\phi(x, y) = k\int \mu(x, y, z)\,\mathrm{d}z$, where $k = 2\pi/\lambda$. The magnitude of the phase change therefore depends on the nature of the electron density inhomogeneities from location to location in the screen. The wave emerging from such a screen is no longer planar, but is instead 'rumpled' owing to the phase changes imposed by the plasma screen. In other words, the wave now propagates into a range of directions—it has been scattered into an *angular spectrum*.

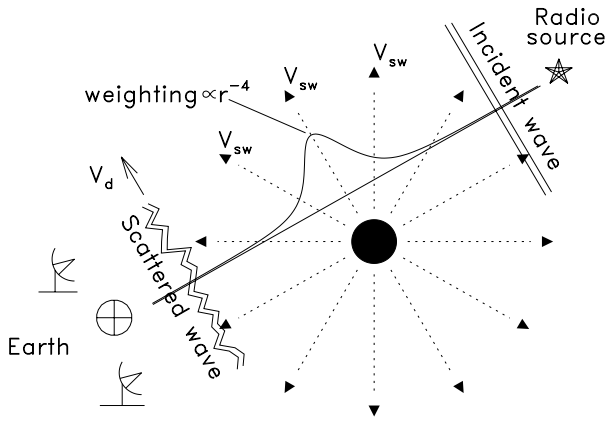


Figure 1. Radio emission from a natural or man-made source passes through the solar wind and is detected by ground-based radio telescopes. The spectrum of electron density inhomogeneities impresses position-dependent phase changes in the incident wave front. These yield both angular broadening and scintillation.

A useful statistic of the scattering medium is the *wave structure function*. Consider a Cartesian coordinate system where the scattering screen lies in the x–y plane at $z = 0$ and an observer is located at a distance L along the z-axis in the x–y plane. Let $\boldsymbol{r} = (x, y, L)$ and $\boldsymbol{s} = (x, y, 0)$. The wave structure function is $D(\boldsymbol{r}) = \langle[\phi(\boldsymbol{r}) - \phi(\boldsymbol{r} + \boldsymbol{s})]^2\rangle$, where $\phi(\boldsymbol{r}) = r_e\lambda\int N_e(\boldsymbol{r})\,\mathrm{d}z$ is the phase deviation along a straight-line path from the source to the observer. The *spatial spectrum* of electron density fluctuations $\Phi_{N_e}(q)$ (where $q = 2\pi/s$) is the Fourier transform of the spatial correlation function of the electron number density fluctuations $\langle\delta N_e(\boldsymbol{r})\delta N_e(\boldsymbol{r} + \boldsymbol{s})\rangle$. Generally speaking, the spatial spectrum of electron density fluctuations describes the nature of electron density inhomogeneities in the solar wind. Under certain simplifying assumptions, it can often be inferred from the wave structure function. A practical advantage of the wave structure function is that it is straightforward to measure.

The *diffractive scale*, r_{diff}, sometimes referred to as the *coherence scale*, is defined by $D(r_{diff}) = 1$, i.e. the transverse scale over which the phase varies by a radian or less. The angle into which the incident radiation is scattered by the solar wind is $\theta_c = 1/kr_{diff}$. The projected size of the source

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on the scattering screen is a second spatial scale, commonly referred to as the *refractive scale*: $r_{ref} = L\theta_c = \lambda L/2\pi r_{diff}$. Note that a third scale can be identified. The Fresnel scale is $r_F = (\lambda L/2\pi)^{1/2}$ and we have $r_{diff} r_{ref} = r_F^2$. The Fresnel scale provides a convenient demarcation between two scattering regimes. When $r_F \ll r_{diff}$, we refer to the *weak scattering* regime; when $r_F \gg r_{diff}$, we refer to the *strong scattering* regime.

In weak scattering, inhomogeneities of a size comparable with r_F cause weak focusing and defocusing of the wave front, causing weak flux variations, or SCINTILLATIONS, on the ground. In strong scattering the wave will constructively and destructively interfere with itself as it propagates from the screen to the observer, producing a diffraction pattern on the ground. If the plasma screen were fixed, the diffraction pattern on the ground would be a fixed, random distribution of bright and dark patches. However, the solar wind moves away from the Sun with a speed of several hundred km s^{-1}. The electron density inhomogeneities being effectively 'frozen' into the solar wind plasma, the diffraction pattern sweeps across the ground at a similar speed and direction to the solar wind. An observer measuring the intensity of the background source at a fixed location on the ground therefore observes strong scintillations. Scintillations are often characterized by the scintillation index $m = \sigma_I/I$, the rms fractional intensity fluctuation. In weak scattering $m \ll 1$; in strong scattering $m \sim 1$. Radio scintillations have an analog in the visible part of the spectrum: stars twinkle as a result of their light being scattered in the Earth's atmosphere.

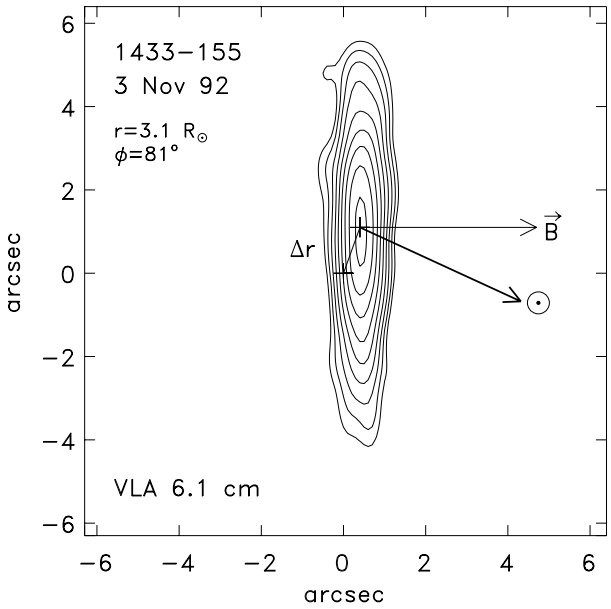


Figure 2. An observation of 1433–155 with the Very Large Array radio telescope at a wavelength $\lambda = 6.1$ cm. The radial distance of the source from the Sun and its position angle are as indicated. Note that the source is displaced by a fraction of an arcsec from its nominal position due to refraction. The heavy arrow indicates the direction toward the center of the Sun, while the lighter arrow indicates the direction of the electron density inhomogeneities, presumed to be aligned with the local magnetic field.

The use of cosmic radio sources

There are many cosmic sources of radio emission (e.g. quasars, radio galaxies, pulsars; see also RADIO EMISSION FROM BINARY STARS) that are spatially coherent—that is, point like—which lend themselves as probes of the foreground solar wind. As the Earth orbits the Sun, these sources are viewed at various elongations from the Sun. A variety of experiments can be performed which exploit natural sources of radio emission. One of the simplest is to measure the apparent angular size of the source. As discussed in the previous section, radio waves are scattered into an angular spectrum. As measured on the ground, the source is broadened in its angular extent. Since the source is, in the absence of the foreground solar wind, a point source, any angular broadening is solely the result of scattering in the solar wind. The spatial coherence of radio sources is measured using interferometric techniques. A RADIO INTERFEROMETER measures the *mutual coherence function*, $\Gamma(\boldsymbol{s}) = \langle E(\boldsymbol{r})E^*(\boldsymbol{r}+\boldsymbol{s})\rangle/\langle|E^2|\rangle$, of the incident radiation, where $E(\boldsymbol{r})$ is the electric field at location $\boldsymbol{r}$. The mutual coherence function is related in a particularly simple way to the wave structure function when observing a distant, point-like source of radio emission: $\Gamma(\boldsymbol{s}) = \exp[-D(\boldsymbol{s})/2]$. As discussed in the previous section, the angular width of the scatter-broadened source is conveniently expressed as $\theta_c = 1/kr_{diff}$, i.e. where $\Gamma(s) = \Gamma(r_{diff}) = \exp(-1)$. It is worth mentioning that the Van Cittert–Zernicke theorem shows that the inverse Fourier transform of the mutual coherence function (suitably normalized) yields the radio brightness distribution on the sky. This is the underlying basis of Fourier synthesis telescopes such as the VERY LARGE ARRAY NATIONAL RADIO ASTRONOMY OBSERVATORY. An example of an observation of a scatter-broadened source made by the Very Large Array is shown in figure 2. The image is anisotropic, and hence the scattering is anisotropic, implying that the electron density inhomogeneities are themselves highly anisotropic near the Sun.

Other experiments exploit the fact that, since the radio flux of many cosmic sources is constant over timescales of days to months, any short-term variation in the radio flux density is due to the passage of the radiation through the solar wind, i.e. scintillations. Scintillations are commonly exploited in three ways. First, scintillation is caused by the movement of the diffraction pattern of the incident radiation across the ground as the solar wind travels outward from the Sun. The intensity fluctuations can be measured by a single antenna and the autocorrelation of a time series of such measurements can be formed. The Fourier transform of the autocorrelation function of the intensity fluctuations is the power spectrum $P(f)$ of the intensity fluctuations at the location of the antenna. In the weak scattering regime, $P(f) \propto \Phi_{N_e}(q)$, the spatial spectrum of electron density fluctuations, where

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$q = 2\pi f/V_{SW}$. Hence, single-station measurements of scintillations of background radio sources can be used to constrain $\Phi_{N_e}(q)$ at various elongations from the Sun.

A second way in which scintillations are exploited is to use them as a passive tracer of solar wind velocity. In a procedure similar to that outlined above for one antenna, the temporal cross-spectrum $P_c(f)$ can be measured by each pair of antennas and the diffraction pattern velocity vector can be extracted from the cross-correlation function (the Fourier transform of the cross-spectrum), or the cross-correlation function can be measured directly. As the background source changes in solar elongation, V_{sw} can be measured as a function of elongation and its acceleration profile can be inferred. While conceptually simple, complications enter into such measurements in practice. One of the main complications is the fact that the solar wind is composed of both slow-speed and high-speed streams. In addition, the electron density inhomogeneities are anisotropic.

A third use of scintillation in the weak scattering regime is to observe a grid of background cosmic sources and to compute the scintillation index m and its mean value $\langle m \rangle$ in each case. The Cambridge disturbance factor is then computed as $G = m/\langle m \rangle$. Typically, hundreds of measurements are made each day to provide a dense sampling of the scintillation properties of the IPM in the form of a 'G map'. Lines of sight along which enhanced regions of electron density occur yield enhanced values of G. Coherent density structures in the IPM can be mapped out using G maps and their evolution in time can be tracked by forming daily G maps. For example, heliospheric structures that corotate with the Sun have been observed using this technique. Recently, tomographic techniques have been exploited to infer the three-dimensional structure of such disturbances (figure 3).

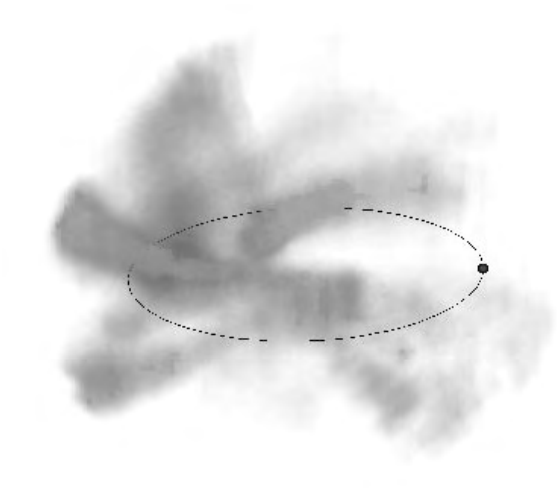


Figure 3. An example of a tomographic reconstruction of the inner heliosphere, to a distance of 1.5 AU, which was derived from a combination of IPS intensity level data and velocity data obtained over one solar rotation period (23 June 1994 to 20 July 1994, Carrington Rotation 1884). Earth's orbit is as indicated. The technique is described in detail by Jackson *et al* (1998).

The use of artificial radio beacons

Radio beacons on the ground (radar) or on board spacecraft offer another means of probing the solar wind. Temporally coherent radio signals—that is, those which are sufficiently narrow band to be regarded as monochromatic—have been used to exploit the phenomena of spectral broadening and phase scintillation to study the solar wind. When a spacecraft is in opposition to the Earth, onboard transmitters can be employed as beacons to study the effect of the solar wind on the signal as it passes through the solar wind. Radar experiments have used VENUS as a passive reflector. As Venus moves into opposition behind the Sun, a radio signal can be transmitted from Earth to Venus. There, the signal is reflected and returns to Earth, where it is detected and spectrum analysed. Both spacecraft signals and reflections from Venus are sufficiently local that the wavefronts involved are no longer planar. Furthermore, in the case of radar experiments, the wave makes a double pass through the solar wind (see VENUS: INTERACTION WITH SOLAR WIND). These complications aside, radar and spacecraft signals have been successfully employed for a variety of studies.

Spectral broadening of a monochromatic signal emitted by a spacecraft is caused by differential Doppler shifts from angular scattering on electron density inhomogeneities in the solar wind. A ground-based receiver can measure the power spectrum of the signal emitted by spacecraft. The Fourier transform of the power spectrum can then be related to the spatial spectrum, $\Phi_{N_e}(q)$, of the solar wind turbulence. In addition, such measurements can be used to constrain the solar wind velocity. Similar analyses are employed for a radar signal reflected from a planet.

It is possible to infer $\Phi_{N_e}(q)$ by measuring the phase variations (phase scintillation) of the incoming radio signal directly. In practice, observations are made at two separate radio frequencies in order to recover both the free space and plasma contributions to the phase. The temporal structure function of the phase is then constructed as $D(\tau) = \langle [\phi(t) - \phi(t+\tau)]^2 \rangle$. The temporal phase structure function is then transformed to the spatial structure to the spatial structure structure function using $D(s = \tau V_{SW})$.

Future prospects

While *in situ* measurements of the solar wind plasma have been made in the ecliptic plane for decades, and more recently at high latitudes by *Ulysses* (see SOLAR WIND: ULYSSES), such measurements have only been possible at distances of ∼1 AU or more from the Sun. Remote sensing

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

techniques at radio wavelengths offer the only means of probing the inner heliosphere at present. Although a future satellite mission designed to probe the outer corona and inner heliosphere directly is in the planning stages, high-quality synoptic measurements of the solar wind at critical distances from the Sun will continue to rely on remote sensing techniques. With improvements to the sensitivity of groundbased RADIO TELESCOPES, greater numbers of cosmic background sources will be available for routine monitoring of the solar wind.

Bibliography

Jackson B V, Hick P L, Kojima M and Yokobe A 1998 *J. Geophys. Res.* **103** 12 049

Narayan R 1992 *Phil. Trans. R. Soc. London* A **341** 151

Narayan R, Anantharamaiah K R and Cornwell T J 1989 *Mon. Not. R. Astron. Soc.* **241** 403

Rickett B J 1990 *Ann. Rev. Astron. Astrophys.* **28** 561

Tim Bastian

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Theory

The supersonic outflow of electrically charged particles, mainly electrons and protons from the solar CORONA, is called the SOLAR WIND. The solar wind was described theoretically by E N PARKER, in 1958. Parker's theory was verified experimentally by *in situ* observations by Soviet and American spaceprobes. On its way to Venus, in 1962, the MARINER II spacecraft observed the solar wind for 104 days. The average flow speed was more than 500 km s^{-1}. This observation showed that the coronal plasma expands into a supersonic solar wind as Parker had predicted.

Parker's solar wind theory

Parker was familiar with the work on comet tails that had been carried out in Germany in the late 1940s and early 1950s under the leadership of LUDWIG BIERMANN (see COMETARY TAILS). The 'ionic' comet tails were observed to be pointing radially out from the Sun. This required a radial force much larger than the force that the photons from the Sun can provide, and Biermann concluded that the 'corpuscular' radiation from the Sun may play an important role in forming the radially pointing comet tails.

Parker based his solar wind theory on the fact that the solar corona has a temperature of more than a million kelvin. He argued that the electron density and the pressure in such a hot atmosphere decrease rather slowly, and the pressure, far from the Sun, is orders of magnitude larger than the pressure of the interstellar gas surrounding the solar system. According to Parker, this imbalance between the pressure in the outer corona and the local interstellar medium would lead to an expansion of the coronal gas into a supersonic solar wind.

The coronal gas, at a temperature of one million kelvin or more, is fully ionized; the number of neutral atoms is very small. However, the thermal energy of the plasma is not large enough to overcome the gravitation field and escape from the Sun. In the inner corona the pressure decrease is determined by gravity, so the gas is in static equilibrium. However, an ionized gas at a temperature of more than a million kelvin is a good conductor of heat. Hence, the coronal temperature does not decrease significantly with increasing distance from the Sun. The thermal energy of the plasma is almost constant whereas the energy needed to overcome the gravitational field decreases. In the outer corona the thermal energy is larger than the escape energy, and the gas can escape.

In Parker's solar wind model, the energy transport (in the form of heat conduction) from the inner corona plays an important role for supplying the energy needed to bring the plasma out of the solar gravitational field. The electrons are much better at conducting this heat than are the ions, owing to their smaller mass, but most of the energy flux goes into increasing the energy of the ions. This transfer of energy from electrons to ions is achieved through an electric field. This electric field must be consistent with the force and energy balance in the flow.

Parker described the coronal gas as a fluid. This allowed him to take into account the coupling of electrons and ions, without calculating the electric field explicitly. In his first study he assumed that the heat conductivity is so high that the coronal plasma has a constant temperature. On the basis of this assumption he could illustrate how the coronal plasma reaches supersonic speed around 5 solar radii from the Sun, and expands into interplanetary space with a steadily increasing flow speed. The equations also allow for subsonic flow, but the pressure far from the Sun in such solutions is almost the same as the pressure in a static, isothermal corona. Hence, the subsonic solutions do not describe an outflow that is in force balance with the local interstellar medium.

In the supersonic flow the flow speed is larger than the thermal motion in the gas, and the dynamic pressure (associated with the directed motion) is larger than the thermal pressure (associated with the random motion). The dynamic pressure of the spherically expanding solar wind decreases with distance from the Sun. At a distance where the dynamic solar wind pressure is equal to the pressure of the interstellar gas a shock is formed: the flow speed is reduced, the density increases and most of the flow energy is transferred into thermal energy. The solar wind termination shock, at 100 AU or so (1 AU = Sun–Earth distance), has not yet been observed, but the Voyager 1 and 2 spacecraft may cross this boundary in a not too distant future.

Parker's solar wind theory has formed the basis for our understanding of the expanding solar corona as well as the outflow of ionized gas from galaxies and stars, and other celestial bodies, and the outflow of light ions from the polar regions of the Earth (the polar wind).

History

Studies of the Sun have a long history in many cultures, but it was the studies of aurora and geomagnetic storms that led to the studies of the solar wind.

The AURORA was part of the mythology and the life of peoples living in the Arctic, but when Celsius and Hiorter in Uppsala, Sweden, started systematic observations of the 'the magnetic needle', in the 1740s, the correlation between aurora and fluctuations in the geomagnetic field was first found. Their collaboration with Graham in London showed that geomagnetic fluctuations occurred both in London and in Uppsala when aurora was seen in Sweden. These findings seem to be the first to establish the link between aurora and GEOMAGNETIC STORMS.

It took some time to find the common cause of aurora and geomagnetic disturbances. The 11 year SUNSPOT CYCLE was noticed by the German amateur astronomer Schwabe in 1843, and several investigations showed a correlation between auroral and geomagnetic activity, and sunspot number. In the 1850s Broun found that geomagnetic storms had a tendency to recur after 27 days, a time close to the rotation period of the Sun, seen from the Earth. On 1 September 1859, a large solar flare was observed by Carrington and by Hodgson, and approximately 18 h later

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

there was a large geomagnetic storm with aurora at very low latitudes. This event served as an indication that there is a connection between solar 'storms' and geomagnetic storms (see MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS). Throughout the last part of the 19th century there were several studies of the possible link between solar activity and geomagnetic storms. 'At the end of the century it was established that aurora and geomagnetic storms should be regarded as manifestations of an unknown cosmic agent of solar origin', Kristian Birkeland wrote in 1908.

One hundred years ago Birkeland set up an impressive research program in Norway to investigate the effects of charged particles from the Sun on the near Earth environment. Together with his assistants he carried out laboratory experiments, observations from field stations around the world and theoretical studies. This was the first comprehensive research effort in solar–terrestrial physics. (Birkeland financed the research program with money he earned from working with industry.)

Birkeland is best known for his Terella experiments, where cathode rays interact with a magnetized sphere, placed in a low-density gas. The experiment showed that the cathode rays impacted on the sphere in regions around the two poles, much like the auroral zones at high latitudes. However, what is less known is that Birkeland also used his magnetized sphere as a cathode to study the emission of 'electric corpuscules' from the Sun. On the basis of observations of continuous geomagnetic activity at stations in the Arctic, Birkeland concluded that there is a continuous flow of charged particles from the Sun. These particles interact with all bodies in interplanetary space, and the interaction with comets leads to the formation of comet tails. To study this process in the laboratory he let cathode rays impinge upon an anode of coal. On the basis of many years of studies of the Sun and of geomagnetic activity he was convinced that all stars, in the course of their evolution, emit electrons and ions into space, and he went on to speculate that most of the mass in the universe is not in stars and nebulae but in 'empty' space.

At the time when Birkeland carried out his work it was not known that the outer solar atmosphere, the corona, is hot. It was first in the 1930s that the picture of a corona with a temperature of more than a million kelvin began to emerge. This was made possible by the development of the CORONAGRAPH by BERNARD LYOT. By shielding the solar disk with a circular plate it was possible to carry out observations of the corona on a regular basis. Previously, such observations could be made only during eclipses. Lyot measured the width of the green line (5303 Å) to 0.9 Å (1 Å $= 10^{-10}$ m). He suggested that the broadening could be due to thermal motions, but the element emitting the line was not known, so a temperature could not be determined.

Grotrian argued, from the early 1930s, that the corona is hot, but it was the identification of coronal lines, as emission lines from highly ionized elements, that established that the corona is hot. Edlén identified the green line (5303 Å) as an emission line from iron atoms that have lost 13 electrons, Fe XIV, and he showed that many of the other coronal lines were emitted by highly ionized elements. These elements cannot exist in the corona unless the temperature is a million kelvin or more. With such a high temperature one could describe the relatively slow electron density fall-off and the widths of the coronal spectral lines. However, it was not easy to understand how it is possible to maintain a hot corona overlaying a 'cold' CHROMOSPHERE. The rapid increase in temperature from the chromosphere to the corona is consistent with a heat conductive energy loss from the corona. This energy loss must be balanced by coronal heating. Biermann and Schwarzschild claimed that acoustic waves could transport energy into the corona from the lower layers in the solar atmosphere, whereas ALFVÉN suggested that MAGNETOHYDRODYNAMIC WAVES (later called Alfvén waves) may carry the energy necessary to heat the corona.

Already in 1941, Alfvén had published a model of a hot static corona, extending out to 10 solar radii, and in the early 1950s there were indications that the solar corona extends even further into interplanetary space. Hewish and Vitkevich observed fluctuations in radio signals that pass through interplanetary space relatively close to the Sun. They found electron density irregularities extending out to at least 20 solar radii and that the 'super corona' changed over the sunspot cycle. Latitude variations were also observed. These observations, together with Biermann's observations of ionic comet tails and Forbush's observations of the variations of the low energy cosmic ray intensity over the sunspot cycle, were known when Parker formulated his solar wind theory.

The solar wind

In 1958 Parker published a paper on the 'dynamics of the interplanetary gas and magnetic field' in the *Astrophysical Journal*. In this paper he presents a model of an expanding solar corona; the coronal gas is allowed to flow out from the Sun in the form of a solar wind. With this model Parker could describe the transition from a quasi-static inner corona to a supersonic solar wind with speeds of several hundred kilometers per second at the orbit of Earth. Parker used hydrodynamic equations, and he considered an isothermal, spherically symmetric, radial flow. This model suffers from several shortcomings, where the most severe may be that the energy balance in the flow is not addressed. However, with this very simple model, Parker could illustrate how the hot coronal gas expands and expels the interstellar gas from interplanetary space. Parker's solar wind model bears some similarities to Bondi's and McCrea's models of accretion of interstellar gas onto a central object. Formally, the two problems are identical, but the physics of Parker's solar wind model is more difficult to understand than the physics of the accretion flow.

The simplicity of the model invited criticism. Chamberlain pointed out that the temperature in the flow is determined by degradation of the heat conductive flux

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and adiabatic cooling, whereas in the Parker model it was taken to be constant. This also implied that the energy per unit mass in Parker's model is infinite. Chamberlain argued that the energy per unit mass should be set to zero. This assumption led to a low-speed 'breeze' solution of the fluid equations. Parker included the energy balance in the flow in his model and showed how heat is conducted outward from the inner corona and converted into flow energy by the pressure gradient force. This model could also describe the supersonic solar wind.

Opponents argued that a fluid treatment of the coronal gas is not valid, and that a kinetic approach should be used. During the 1960s there were several attempts to construct exospheric solar wind models. Many of these studies gave results in better agreement with the breeze solution than with the supersonic wind solution. One reason some of the exospheric models gave low flow speeds is that the electric field in the models was too small. This electric field is set up by the plasma to balance the pressure gradient force in the electron gas. In the fluid model the electric field does not appear explicitly. In the kinetic models this is not the case. Many of the exospheric models used the electric field of a static corona. This is smaller than the field in a subsonic–supersonic solar wind and gives rise to a smaller acceleration of the protons than is found in the hydrodynamic model.

In Parker's model the solar magnetic field is 'dragged out' by the solar wind. Because of the high electric conductivity the electric field in the expanding gas is small. As a consequence, the magnetic field is connected to the source region at the Sun. When the Sun rotates, the gas emitted from one region in the corona is situated on a spiralling field line, and the direction of the field depends on the polarity in the corona. In the ECLIPTIC, the average magnetic field is in the ecliptic plane, and the angle between the average field and the radial direction is around 45° at the orbit of Earth.

It took only a few years before the solar wind 'controversy' could be settled. *In situ* observations of the interplanetary plasma were carried out by the Soviet spacecraft Lunik 2 and the American spaceprobe Explorer 10. However, it was not until the fall of 1962, after the Mariner II observations of the solar wind, that Parker's solar wind model was accepted. On route to Venus, Mariner II obtained 104 days of observations. It measured an average solar wind flow speed of 504 km s^{-1}. The proton density was around 5 cm^{-3}, much lower than the density in a static corona. The average interplanetary magnetic field showed a spiral structure very similar to the model proposed by Parker.

The Mariner II flight took place during the declining phase of the sunspot cycle. There were several high-speed solar wind streams, with a recurrence period of 27 days (seen from the Earth), during the mission, that gave rise to recurrent geomagnetic storms with the same period. Thus, the source of recurrent geomagnetic storms could be identified as high-speed solar wind streams, but the solar source regions of these streams, often called M-regions, were not identified. It was 10 years later that 'coronal holes', were first identified as the source regions of quasi-steady high-speed solar wind; coronal holes are regions in the corona, with unipolar magnetic field and low electron density. During the declining part of the sunspot cycle the polar regions develop into large coronal holes, and the high-speed solar wind from these regions fills up a large fraction of interplanetary space.

Solar wind from a given corona

When Parker formulated his solar wind theory he took the inner solar corona to be a reservoir of particles and energy for the outflowing solar wind. Therefore it can be argued that he made the assumption that the energy balance in the corona is between coronal heating and inward heat conductive energy loss and that the energy loss in the solar wind is not large enough to significantly change the structure of the inner corona.

This model can describe the basic dynamics of the corona–solar wind system quite well, i.e. the expansion of the coronal gas into a supersonic solar wind. However, Parker realized that the model could not describe the quasi-steady high-speed solar wind streams that were observed, mainly during the declining phase of the sunspot cycle. He found that energy has to be added beyond the inner boundary to speed up the flow. Many years later Leer and Holzer showed that this energy has to be deposited in the supersonic region of the flow; energy deposition close to the Sun increases the solar wind particle flux whereas energy deposition in the supersonic flow increases the energy per unit mass in the flow and therefore the flow speed.

There are several problems with a model where the electron density and temperature in the inner corona are taken to be independent of the solar wind outflow. If we assume that the electron (proton) density at the inner boundary is fixed, and let the temperature increase, the solar wind proton flux increases rapidly. By varying the temperature from one to two million kelvin the solar wind proton flux changes by a factor of 100 or so. However, the solar wind proton flux is observed to be fairly constant whereas the coronal temperature shows considerable variation. This inconsistency cannot be resolved in a reasonable manner within the framework of a traditional solar wind model. In order to obtain a more complete description of the solar wind we must include the coronal energy balance in the model.

Formation of the corona and acceleration of the solar wind

In Parker's theory the solar wind is a consequence of the coronal heating process, so the inclusion of the coronal energy balance would be a natural extension of his model. This can be done by moving the inner boundary from the corona into the upper chromosphere. A significant energy flux is needed to balance the radiative losses from the chromosphere, but the energy flux deposited in the outermost part of the solar atmosphere, where the electron

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

density is very small, is not radiated away locally. The temperature increases until some other loss mechanism balances the heat input. In magnetically closed regions, where the coronal plasma is trapped by the magnetic field, inward heat conduction is the most significant energy loss, whereas in regions where the magnetic field extends into interplanetary space, and the coronal plasma is free to escape, energy may also be lost in the solar wind. How much of the energy that is lost as inward heat flux and as solar wind energy flux may depend on the amplitude of the energy flux and how and where this energy flux is deposited in the corona.

Energy balance in a static corona

Let us first consider the energy balance in magnetically closed regions, where the energy deposited as heat is lost as inward heat conduction. To make the problem as simple as possible we can consider a spherically symmetric corona, with a 'lid' on it. This outer boundary does not allow transport of either plasma or energy.

If this static corona is heated by an energy flux from the Sun, say 100 W m^{-2} at the solar surface, and this energy flux is deposited in the corona over a length scale comparable with a solar radius, we find a coronal temperature in the range 1.5–2.0 million K; a corona at this temperature will lose 100 W m^{-2} into the TRANSITION REGION in the form of an electron heat flux. However, the heat conductive loss from the corona depends sensitively on the coronal temperature; a corona at 0.5 million K loses only 1 W m^{-2} as inward heat flux. Because of the very strong temperature dependence of the heat conduction in an ionized gas we find that a small heat input is sufficient to maintain a rather hot corona.

In a corona where the electron (proton) density is so low that the electrons and protons are not thermally coupled, the proton heating must be balanced by heat conduction in the proton gas. As there is no reason to expect that the protons are heated less than the electrons, the lower heat conductivity in the protons leads to a proton temperature that is higher than the electron temperature.

The electron density in the inner corona is determined by the pressure in the chromosphere–corona transition region, and this pressure is determined by the heat conductive flux from the corona. If we assume that the radiative losses from the transition region are balanced by the heat flux from the corona, we find that an inward heat flux of 100 W m^{-2} corresponds to a transition region pressure of $p_{TR} = 7 \times 10^{-3}$ N m^{-2}. The pressure is proportional to the heat flux, so a heat flux of 10 W m^{-2} corresponds to 7×10^{-4} N m^{-2}. A large inward heat flux leads to a high transition region pressure, a large electron density in the inner corona and strong collisional coupling between electrons and protons. A small inward heat flux is consistent with a low pressure in the transition region and a low electron density in the corona.

Energy balance in coronal holes

In magnetically open regions the energy that is deposited in the corona can be lost as inward heat flux and as solar wind energy flux. Which of these two loss mechanisms is the most important one may depend on how and where energy is transferred to the coronal plasma. If most of the energy is added to the electrons, as heat, a large fraction may be lost as inward heat flux. However, if most of the energy is added to the ions the inward heat flux may be reduced significantly.

In order to investigate the energy balance of coronal hole regions we can construct mathematical models extending from the upper chromosphere, through the transition region and corona, and far into interplanetary space. The heating of the corona is specified through the amplitude of the energy flux and how and where this energy flux is deposited. These types of studies show that heating of the very inner corona leads to a large inward heat flux, whereas only a small fraction of the energy flux is lost in the solar wind. Heating further out from the Sun leads to a larger solar wind energy loss, but there is a significant difference between models with electron and proton heating.

In models with extended electron heating the inward heat flux is a significant fraction of the energy flux, and because of the high heat conductivity in the electrons the temperature in the corona does not exceed 1.5 million K for a reasonable heat input. The large inward heat flux is consistent with a large transition region pressure and a quite large electron density in the inner corona. The solar wind emanating from such a corona does not reach a high flow speed far from the Sun. A typical value is 300 km s^{-1}. In these models, where the electrons are heated, a large fraction of the outward energy flux from the corona is carried as electron heat conduction. This energy is transferred to the ions via the polarization electric field. This model has many similarities to the 'classical' solar wind model, where electron heat conduction from the inner corona supplies the energy flux that is needed to drive the solar wind.

If most of the energy is added to the protons, the inward heat flux is much smaller, the transition region pressure is low, the electron density in the inner corona is low, the thermal coupling may be weak and the coronal proton temperature may be quite high. In models with extended proton heating most of the energy deposited in the corona is lost in the solar wind. The density in the corona and the solar wind proton flux are small, so the energy per unit mass in the flow, and the flow speed far from the Sun, can be quite large. In order to obtain flow speeds of 800 km s^{-1}, measured in the high-speed wind by the Ulysses spacecraft, a large fraction of the energy must be deposited in the outer corona where the proton density is so low that the heat cannot be conducted away. Then, the heating leads to a high proton temperature and rapid acceleration of the flow.

This type of model study of the corona–solar wind system is limited by the assumptions made, but some

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

general results seem to emerge. High-speed solar wind streams cannot be found in models where most of the energy is deposited in the very inner corona. In models with significant electron heating the inward heat flux is too large, the coronal electron density is too high and the coronal temperature is too low to generate high-speed wind. The high-speed streams can be obtained in models where most of the energy is deposited in the protons. The energy may be deposited as heat or it may go into direct acceleration of the flow, but none of these processes will lead to a large flow speed unless the energy is deposited sufficiently far out in the corona. Most of the energy flux deposited in the corona, in these models, is lost in the solar wind. As the asymptotic flow speed is comparable with the escape speed of the Sun, the solar wind mass flux is proportional to the the energy flux deposited in the corona.

Ulysses and SOHO

In the 1990s there have been two successful space missions that have contributed significantly to our understanding of the solar wind. The Ulysses spacecraft has observed the solar wind at all solar latitudes (see SOLAR WIND: ULYSSES). The first orbital period was during the declining phase of the sunspot cycle. The observations showed a fast and steady solar wind at latitudes larger than 20°. Close to the ecliptic the flow speed was lower and more variable. The solar wind mass flux showed little variation with latitude.

The fast solar wind streams originate in large coronal holes. Observations of spectral lines from large polar coronal holes, with instruments on board the Solar and Heliospheric Observatory (SOHO), show that heavy ions are warmer than protons which themselves are warmer than the electrons. These observations support a model of high-speed wind, with significant energy deposition in the protons and not in electrons. The fact that we observe heavy ions in the corona may be taken as an indication that their temperature is high. If the temperature of heavy ions were equal to the proton temperature, the density of heavy ions would fall off so rapidly with heliocentric distance that it would be difficult to see them in the corona. However, does a high temperature tell us that a heavy ion is preferentially heated?

Particle escape seems to be the most important energy loss process for protons and heavy ions in coronal holes. As the flux of escaping particles is an exponential function of the thermal energy over the escape energy a modest heating of heavy ions leads to a high temperature. If the energy input per unit mass is the same for protons and heavy ions the temperature should be proportional to mass. However, if the ion temperature increases with mass, more rapidly than proportional to mass, we can conclude that the heavy ions are preferentially heated. So far, only a few ion temperatures have been measured in large coronal holes with the Ultraviolet Coronagraph and Spectrograph (UVCS) on board SOHO, but observations of spectral lines from oxygen atoms that have lost 5 electrons, O VI, seem to indicate that there are heavy ions in the corona that are preferentially heated.

Summary

Parker's solar wind theory forms the basis for our understanding of the interplanetary plasma. The dynamics of the solar wind is described by models developed more than 40 years ago. The extension of these models, which includes the coronal energy balance, has led to a fairly good understanding of the acceleration of the solar wind.

The solar wind is driven by the energy deposited in the corona. This process is not understood. In the future the emphasis will therefore be on trying to understand coronal heating. A better understanding of how energy is transported into the corona and transferred to the gas will also give us a deeper understanding of how the solar wind is accelerated.

Bibliography

Eather R H 1980 *Majestic Lights* (American Geophysical Union)

Golub L and Pasachoff J M 1997 *The Solar Corona* (Cambridge: Cambridge University Press)

Jokipii J R, Sonett C P and Giampapa M S (eds) 1997 *Cosmic Winds and the Heliosphere* (Tucson, AZ: University of Arizona Press)

Parker E N 1958 Dynamics of the interplanetary gas and magnetic fields *Astrophys. J.* **128** 664–76

Egil Leer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar Wind: Ulysses

Ulysses is the first space mission in a polar orbit around the Sun, and the first to make observations in the previously unvisited regions of the heliosphere over the polar regions at high heliolatitudes. The objective of the Ulysses space mission, launched on 6 October 1990, is the exploration of the dependence on heliolatitude of the heliospheric medium and its relationship with the Sun's coronal structure and dynamics. The heliosphere is the large, three-dimensional region of space around the Sun filled with the SOLAR WIND and the remnant of the SOLAR MAGNETIC FIELD carried in it. The outer boundary of the heliosphere is determined by the interaction between the solar wind and the local interstellar medium.

The properties of the heliospheric medium and their dependence on solar and coronal phenomena have been studied since the early 1960s by numerous spacecraft in Earth orbit, by planetary probes during their cruise to the target planets and by dedicated interplanetary missions. Most of the direct, *in situ* observations have been confined to the vicinity of the ecliptic plane (the plane of the Earth's orbital motion around the Sun). In heliographic coordinates, this corresponds to a range of $\pm 7.25^\circ$ in heliolatitude, the angle between the ecliptic plane and the solar equatorial plane. Solar observations from the ground clearly show that solar and coronal phenomena strongly depend on heliolatitude. In particular, observations during solar ECLIPSES make clear that the CORONA usually has an appearance and structure in the polar regions very different from those near the Sun's equator. As the solar wind, which forms the heliospheric medium, is a result of the expansion of the solar corona into space, its properties are also expected to depend strongly on heliolatitude. This dependence of the heliospheric medium on heliolatitude had been recognized, but, prior to Ulysses, no space mission had been undertaken to study the heliospheric medium in three dimensions.

This delay was due, in part, to the difficulties involved in placing a spacecraft into an orbit out of the ecliptic plane. Ulysses' unique orbit was achieved by first targeting the spacecraft to Jupiter and using gravity assist (the maneuver which allows, during the close flyby of a planet, use of the gravitational potential of the planet to modify the orbit of a spacecraft) to deflect the trajectory of Ulysses out of the ecliptic plane. The Jupiter flyby took place in February 1992, and since then Ulysses has been in a nearly polar orbit around the Sun. Its orbital plane makes an angle of about 80° with the solar equatorial plane, its aphelion (the furthest distance from the Sun) is at 5.4 astronomical units (AU, the mean distance between the Sun and the Earth), its perihelion (closest approach to the Sun) is at 1.4 AU. The orbital period of Ulysses is 6.2 years; its first complete polar orbit is shown in figure 1. Ulysses is equipped with a comprehensive payload to make observations in the solar wind. Instruments to measure the speed, density, temperature and composition of the solar wind plasma are complemented by instruments to measure the magnetic field, radio and plasma waves, as well as energetic particles from the Sun and COSMIC RAYS.

Following the Jupiter flyby, the spacecraft first explored the region of the heliosphere over the south pole of the Sun, in 1993–4. It then moved relatively rapidly, between September 1994 and August 1995, from the southern polar regions to the northern polar regions. This passage from south to north allowed a unique survey of the latitudinal dependence of the solar wind. Ulysses completed its first full solar polar orbit in March 1998. The timing of the mission during the declining and minimum phases of the SOLAR CYCLE, with the polar passes taking place in 1994 and 1995, has yielded a unique set of observations of the configuration of the heliosphere and the properties of the heliospheric medium at solar minimum activity.

The key results of Ulysses are:

(a) The confirmation of the very nearly uniformly high solar wind velocities (>750 km s^{-1}) from the polar coronal holes.
(b) The significant over-expansion of the polar solar wind when compared with the areas occupied by coronal holes.
(c) The clear differentiation, as determined by ion composition and freezing in temperature measurements, between the characteristics of fast and slow solar wind streams.
(d) The latitudinal independence of the radial component of the magnetic field, constraining solar wind models.
(e) The high levels of mostly Alfvénic, transverse fluctuations associated with high-speed wind streams.
(f) The slower evolution of the turbulence in the uniform solar wind in the polar regions when compared with the near-ecliptic regions of the heliosphere.
(g) The less than expected latitudinal gradient of cosmic rays, calling into question established cosmic ray modulation models.

These findings of Ulysses are described in more detail in the following paragraphs.

It has been known from previous solar and interplanetary observations that there are two kinds of solar wind. The slow solar wind, with speeds not exceeding about 450 km s^{-1}, originates from above coronal regions in which magnetic fields lines are generally closed in large-scale loops. The fast solar wind, with speeds in excess of 550 km s^{-1}, but usually about 750 km s^{-1} or more, originates from regions of the solar corona, the so-called CORONAL HOLES, in which magnetic field lines rooted in the solar surface at one end spread out into the heliosphere. The configuration of the solar corona around solar minimum is dominated by two large coronal holes in the polar regions and a belt of closed magnetic fields in the equatorial region. This was the coronal configuration during the polar passes of Ulysses in 1994 to 1996. The observations of the solar wind onboard Ulysses confirmed that the polar regions of the heliosphere were filled with fast solar wind, carrying magnetic field lines of opposite

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

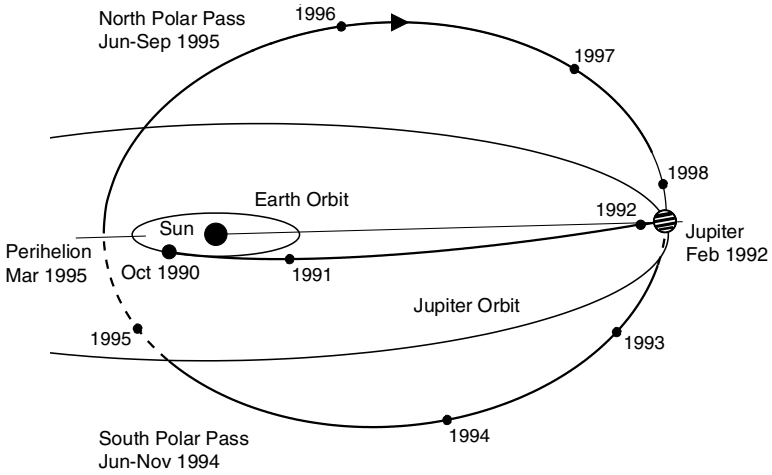


Figure 1. The first solar polar orbit of the Ulysses mission.

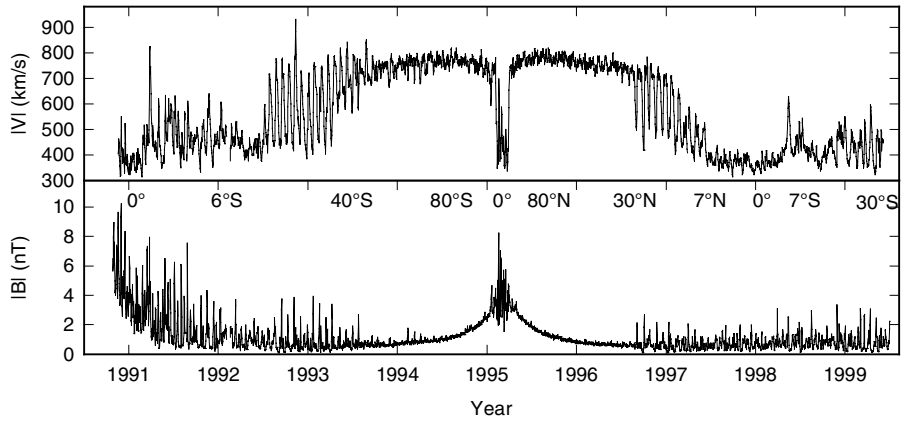


Figure 2. The solar wind speed (upper panel) and the magnitude of the magnetic (lower panel) observed by Ulysses from launch through its first orbit.

polarity from the two polar coronal holes. When Ulysses was close to the solar equator, at latitudes less than 35° south in 1993 and about 25° north in 1996, the solar wind reached the spacecraft both from the equatorial band of solar streamers (the heliospheric extension of the regions above the closed magnetic regions in the corona) and from the equatorward extensions of the polar coronal holes. This was due to the small inclination and warping of the band of slow solar wind with respect to the equatorial plane, so that as the Sun rotated, Ulysses was immersed alternately in slow and fast solar wind in each period of solar rotation. When Ulysses passed poleward of the equatorial region, the speed of the solar wind was found to be uniform, at about 750 km s^{-1}, in a very large volume of space over the two solar poles. Only small variations in this high speed were found, explained by small-scale dynamic effects ('microstreams') and by the effect of very large-amplitude, long-wavelength Alfvén waves which dominated the large-scale fluctuations in the magnetic field. The overview of the two key parameters, the solar wind speed and the magnitude of the magnetic field measured by Ulysses since launch, is shown in figure 2.

The areas covered by coronal holes near the Sun's poles are relatively small, only about 20% of the solar surface at the base of the corona, yet Ulysses observed that more than 75% of space (measured in terms of solid angle) around the Sun was filled with solar wind originating from those coronal holes. This means that the solar wind does not simply expand radially in the corona but considerable expansion takes place in latitude, from the poles towards the equatorial regions, as the solar wind is accelerated away from the Sun. Several previous models of coronal expansions had considered this possibility, but only Ulysses' observations were able to establish this important aspect of the origin of the solar wind.

The composition of the solar wind is different in slow and fast streams, a result well established in observations made in the ecliptic plane. Partly because of its polar orbit, partly because of the instrumentation carried, Ulysses provided the first very strong and comprehensive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

evidence that the compositional differences between the two kinds of solar wind represented the coronal conditions in which they originated. Slow solar wind originates in regions where the coronal plasma is very hot, about 1.5 MK or so, so that elements such as oxygen, iron or magnesium are highly ionized. But coronal holes are cooler and the lower temperature is reflected in the degree of ionization of the elements. Using ratios of the ionization states of elements (for example, the abundance ratio of O^{7+} and O^{6+}, respectively oxygen ions with 7 and 6 electrons stripped away) measured in the solar wind, it is possible to determine the temperature of the corona at the height where the ionization/recombination time scale of the ions (which is increasing with distance from the solar surface due to the decreasing density) overcomes the expansion time scale of the accelerating solar wind. This is the 'freezing-in' temperature which can be used to differentiate between the regions of origin of the solar wind. The major result from Ulysses is that the boundaries between slow and fast solar wind streams are extremely sharp in the corona; in addition, it appears that the fast solar wind is accelerated much lower in the corona (at heights less than four solar radii) than previous models suggested. This, together with more detailed elemental abundance data, provides strong constraints on the major unsolved question as to how the solar wind is accelerated.

Observations of the magnetic field carried in the solar wind, along the high-latitude orbit of Ulysses, yielded three results which had not been predicted. One of these is that the magnetic field does not depend on latitude, but remains constant over the whole polar region of the Sun. It had been assumed that however complex the coronal magnetic field may be, at large distances it will appear to be approximately like a dipole, with stronger fields near its axis. As this is not the case, the implication is that magnetic fields become more evenly distributed in latitude close to the Sun than the dipole model implies. This finding is related to the non-radial expansion of the fast solar wind from the coronal holes to lower latitudes and confirms the overall structure of the heliosphere at solar minimum.

Two other properties of the high-latitude heliospheric magnetic field differentiate it from the interplanetary magnetic field observed near the ecliptic. Fluctuations in the direction of the magnetic field in the high-speed solar winds from the polar coronal holes were found to be very much greater than at low latitudes; it is important to note that the magnitude (or strength) of the magnetic field was relatively constant, only varying as a function of distance from the Sun. The directional fluctuations were most pronounced perpendicular to the radial direction from the Sun. These large-scale directional fluctuations, transverse to the flow of the solar wind, were identified as Alfvén waves, driven by the tension of the magnetic field lines. Such waves had been first identified in low-latitude data in fast solar wind streams, and attributed to a population of waves which travelled in the solar wind from the corona far into interplanetary space. At high latitudes, however, the waves had much larger amplitudes and covered a significantly wider range of periods than previously seen. This property of the high-latitude heliosphere is the result of a considerably reduced rate at which long-wavelength waves decay with distance from the Sun. The heliospheric medium is fundamentally a turbulent flow, in which lower-frequency (longer-wavelength) waves decay to give rise to waves of higher frequency. Turbulent processes in the solar wind are not well understood. The results of Ulysses at high latitudes indicate, when compared with the turbulent decay of waves in the equatorial region, that the large speed differentials in the solar wind found at low latitudes are responsible for the faster decay rate observed there. In the more uniform (and faster) flow speed of the solar wind in the polar regions large-amplitude magnetic fluctuations remain a prominent feature of the medium over larger distances than near the ecliptic.

This result leads directly to another unexpected observation made by Ulysses. Before these observations were made, it had been thought that the intensity of cosmic rays would be significantly higher over the poles of the Sun than in the ecliptic. As cosmic rays (highly energetic ions and electrons originating elsewhere in the galaxy) follow the lines of force of the magnetic field, they were expected to find access to the inner solar system easier over the solar poles where the magnetic field lines are less wound up into spirals by the rotation of the Sun. However, the intensity of cosmic rays showed a much less pronounced increase than expected as Ulysses reached high heliolatitudes. It appears that the large amplitude magnetic fluctuations at high latitudes effectively act as barriers which impede the propagation of cosmic rays into the heliosphere. This conclusion was well supported by the cosmic ray observations during the period when Ulysses moved rapidly from the south to the north pole in 1994–5, when the dependence on heliolatitude was not significantly complicated by either temporal changes or by a dependence on heliocentric distance. A summary of the cosmic ray intensity, as a function of heliolatitude, measured on Ulysses, is shown in figure 3, together with the model predictions based on our knowledge before Ulysses. As a result of Ulysses' observations, models of cosmic ray propagation in the heliosphere need significant revisions (see COSMIC RAYS: PROPAGATION IN THE HELIOSPHERE).

Overall, the observations made at high heliolatitudes by Ulysses have provided the basis for a better understanding of the solar wind and the heliosphere. Many of the results were seen as welcome confirmation of models extrapolated from the observations near the ecliptic plane. However, there were enough unexpected results to provide an *a posteriori* justification for the mission. Ulysses remains in orbit around the Sun to explore the same regions that it visited during solar minimum, to observe the changes in the solar wind caused by the increased solar activity at high latitudes during the next solar maximum in 2000–1. It is likely that these observations, which are likely to remain unique as no similar space mission is being planned at present, will provide a data base for future, possibly more refined,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

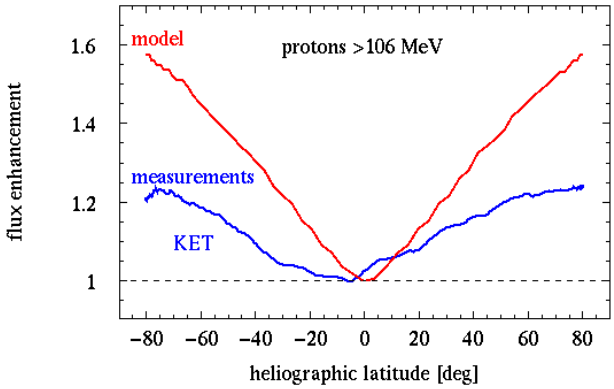


Figure 3. The enhancement of the intensity of cosmic rays observed by Ulysses with respect to observations in Earth orbit during the passage of the spacecraft from the south to the north pole of the Sun in 1994–5, and, for comparison, the model prediction which had assumed that cosmic ray access over the poles would be significantly easier than was found to be the case. (Figure courtesy of H Kunow, University of Kiel, Germany.)

models of the solar wind and the three-dimensional heliosphere.

A Balogh

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar X-ray Jets

X-ray jets have been discovered on the Sun by the soft x-ray telescope aboard the Japanese YOHKOH satellite as transitory x-ray enhancements in the solar CORONA with apparent collimated motion. They are associated with small flares (microflares or subflares), which occur in x-ray bright points (XBPs, see SOLAR CORONA: X-RAY BRIGHT POINTS), emerging flux regions (EFRs, see CHROMOSPHERE: EMERGING FLUX REGIONS) and ACTIVE REGIONS (ARs). Whereas many XBPs occur in quiet regions and CORONAL HOLES, x-ray jets occur throughout the solar corona. As the spatial resolution and exposure time of soft x-ray observations increased, more and more x-ray jets were found. Hence, it could be that the solar corona is full of tiny unresolved x-ray (coronal) jets which are responsible for coronal heating and acceleration of the high speed SOLAR WIND.

There is a lot of evidence for MAGNETIC RECONNECTION in x-ray jets. In this sense, x-ray jets are also useful for studying the basic physics of magnetic reconnection.

Figure 1 shows a typical example of an x-ray jet that occurred in active region NOAA 7001 on 11 January 1992. This x-ray jet is one of the largest observed so far and could be a prototype of many other, smaller jets.

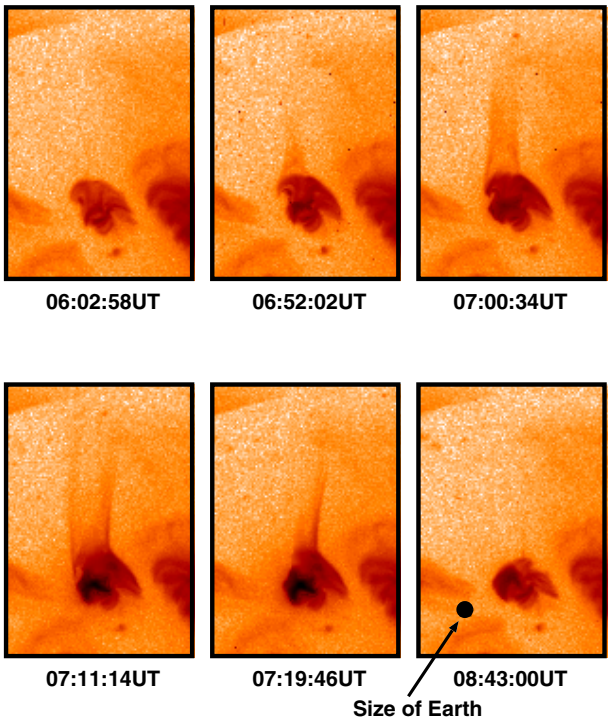


Figure 1. An x-ray jet occurred in NOAA7001 on 11 January 1992 (Shibata *et al* 1994 *Astrophys. J.* **431** L51). The size of each frame is 1200″ × 1040″ (≃ 8.6×10^5 km by 7.5×10^5 km). These are negative images. North is at the top and west is to the right. The maximum length of the jet was greater than 3×10^5 km and the width was about 7.5×10^4 km. The apparent velocity changed with time and was 90–230 km s^{-1}. The active region at the footpoint of the jet looks like a sea anemone, and so is called an *anemone-AR*. The surrounding region is a coronal hole. During the ejection of the jet, a bright loop appeared (as a result of a small flare) about 3×10^4 km south of the footpoint of the jet, while the loop system just below the jet disappeared at 0721 UT. The boundaries of the jet moved apart at a speed of ∼ 20 km s^{-1}, increasing the apparent width of the jet. The west boundary of the jet continued to move even after the east boundary had disappeared, giving an impression of whip-like motion.

Observed properties

Structure/morphology and dynamics

Most jets are associated with small SOLAR FLARES (microflares, subflares) at their footpoints. The length of the jets is a few 10^4–4×10^5 km (average ≃ 1.5×10^5 km), their width is 5×10^3–10^5 km (average 1.7×10^4 km), their aspect ratio (= length/width) is 3–27 (average ∼ 10), their apparent velocity is 10–1000 km s^{-1} (average ≃ 200 km s^{-1}) and their lifetime ranges from a few minutes to more than a few hours. The number of jets decreases as the length, the velocity or the lifetime increases. The histograms (figure 2) are similar to those of flares and EUV explosive events. Many jets (∼ 68%) appear in or near to active regions (ARs). Among the jets ejected from bright-point like features in ARs, most (∼ 86%) are observed to the *west* of the active region. The x-ray intensity distribution along an x-ray jet often shows an *exponential* decrease with distance from the footpoint.

Temperatures, densities and energies

The average temperatures of jets are ∼ 3–7 MK, which is comparable to that of the microflares at the footpoints of the jets, and the average electron densities are 0.3–4×10^9 cm^{-3}. The total thermal energy is 10^{26}–10^{28} erg, which is about 1/10 of the total (released) thermal energy of the footpoint flare. The masses of the jets are 10^{11}–10^{13} g and their kinetic energy is estimated to be 10^{25}–10^{27} erg.

Magnetic fields and evidence of reconnecting fields

Co-alignment of soft x-ray images with magnetograms (taken at ground-based observatories) revealed that 72% of jets occur in mixed polarity or satellite polarity regions. This gives direct evidence for the presence of neutral points (or current sheets) near the footpoint of jets. X-ray jets favor regions of evolving magnetic flux (increasing or decreasing). This suggests that x-ray jets are produced by magnetic reconnection.

There is much direct and indirect evidence for magnetic reconnection in jets:

(a) *Two types of interaction between emerging flux and coronal field*: There are two types of interaction (reconnection) between emerging flux and coronal field: the *anemone-jet* type and the *two-sided-loop/jets* type (figure 3). The former occurs when emerging flux appears in coronal holes. In this case, a jet is ejected in a vertical direction. On the other hand, the latter occurs when emerging flux appears in quiet regions, and two-sided loop brightenings (or jets) occur in the horizontal direction on both sides of the emerging flux.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

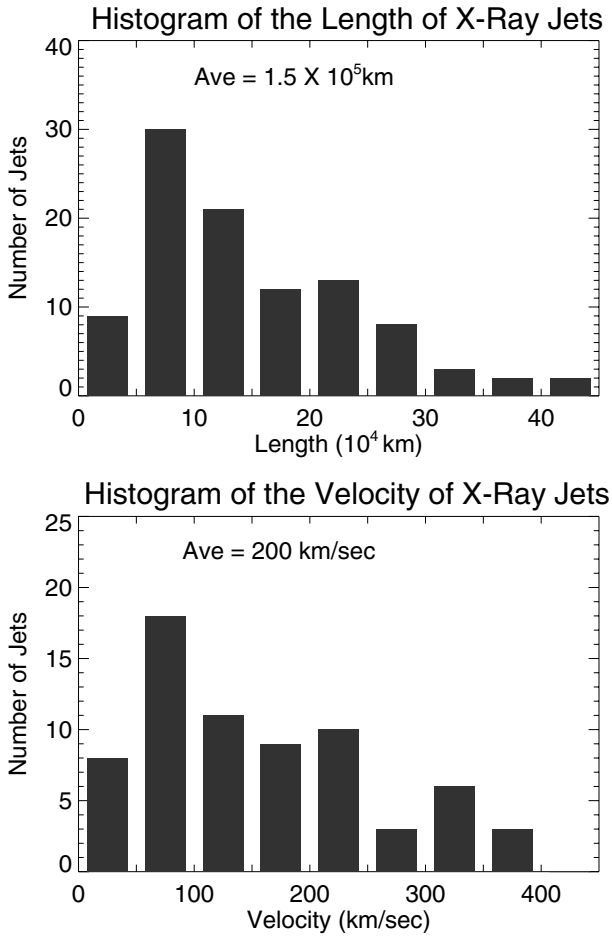


Figure 2. Histograms of velocity and length of x-ray jets (Shimojo *et al* 1996 *Publ. Astron. Soc. Japan* **48** 123).

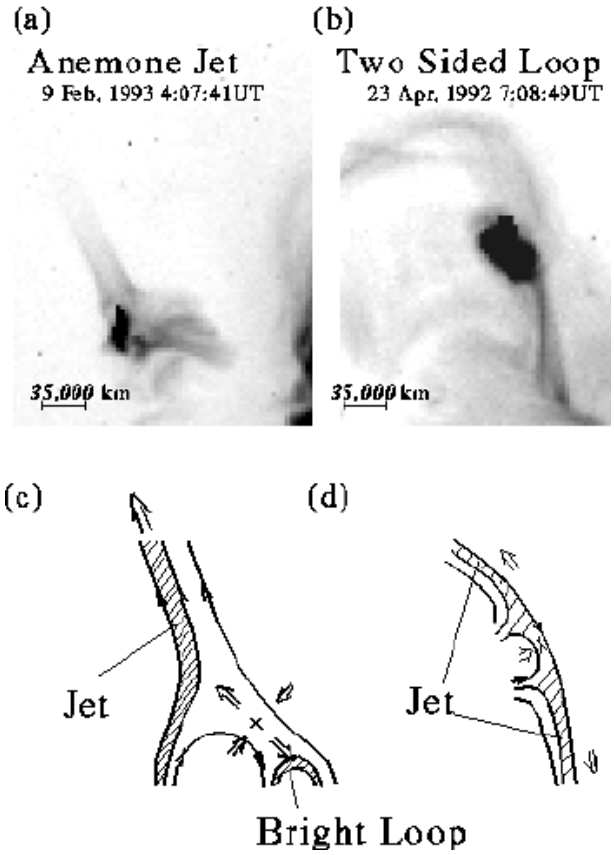


Figure 3. Soft X-ray Telescope images (from Yohkoh) of two types of interaction of emerging flux with overlying coronal magnetic fields. (*a*) *Anemone-Jet* type observed on 9 February 1993. (*b*) *Two-sided-loop* type observed on 23 April 1992. Schematic illustration of (*c*) *anemone-jet* type and (*d*) *two-sided-loop* type (Yokoyama and Shibata 1995 *Nature* **375** 42).

(b) *Converging shape of jets*: The width of the jets often decreases with height (i.e., *converging* shape), which is similar to the shape of Hα surges observed in emerging flux regions and EUV macrospicules. This shape suggests that the cross-section of the flux tube decreases with height, i.e., the field strength increases with height. Such situations arise if there is a neutral point near the footpoint of a jet as in the *anemone-jet* model in figure 3. This situation is expected also when the satellite spots appear in an opposite polarity region, and in fact such magnetic field properties have been confirmed by comparing magnetograms with soft x-ray images of many jets as discussed above (figure 4).

(c) *A gap between the footpoint of a jet and the brightest part of a footpoint flare*: Though the footpoints of jets roughly correspond to small flares, close examination of the footpoints has revealed that often small flares (or loop brightenings) occur separately (by more than a few thousand km) from the exact footpoints of jets. This characteristic is also seen in tiny XBP jets. Such a gap is expected for the magnetic reconnection mechanism, because the heated reconnected field lines are quickly ejected in opposite directions to form one bright loop and a separate jet in the other direction (figure 3).

(d) *Change of topology of a footpoint active region*: When the ARs at the footpoints of jets can be resolved well, their morphology changes substantially during the occurrence of a jet. For example, a loop system disappeared during the jet of 11 January 1992 (figure 1).

(e) *Whip-like motion of jet*: In some cases, a jet moves perpendicularly to the jet axis at a few 10 km s^{-1} during the ejection of the jet, which has been referred to as *whip-like motion* (figure 1). This might be evidence of a dynamical rearrangement of the magnetic field configuration as a result of reconnection.

Relation to other dynamic phenomena

Hα surges: Sometimes, x-ray jets are observed to be associated with Hα SURGES, nearly simultaneously and in nearly the same direction. There are also cases with no x-ray jet when Hα surges are observed. This is probably

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

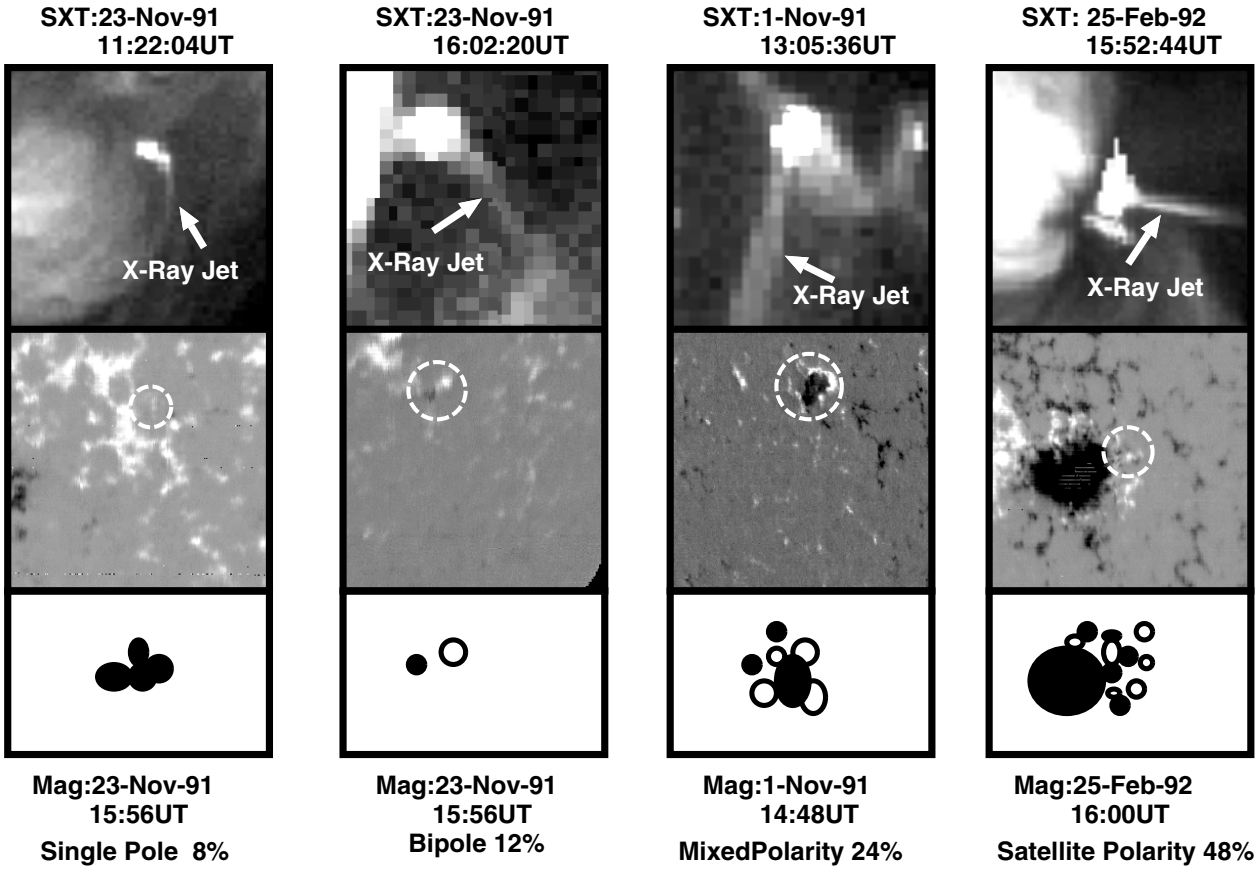


Figure 4. Examples of the magnetic field properties of x-ray jets (Shimojo *et al* 1998 *Solar Phys.* **178** 379). The first row shows x-ray jet images taken by the *Yohkoh* Soft X-ray Telescope. The second row are the corresponding magnetograms with white circles indicating the footpoint of the jets. The third row shows schematic illustrations of magnetic field configuration types.

because the density of the hot jet is too low to be observed by a soft x-ray telescope.

Type III bursts: Type III radio bursts (see SOLAR FLARES: RADIO BURSTS) are often associated with x-ray jets. This implies the existence of high-energy electrons in these small flares and jets, and supports the view that the generation mechanism of x-ray jets and microflares may be physically similar to that for larger flares.

Nanoflares: It was found that very faint transient soft x-ray brightenings occur in polar regions. The absolute soft x-ray intensity of these brightenings is two orders of magnitude smaller than those of active region transient brightenings, and fluctuates on a timescale of a few minutes. The total (released) energy is probably comparable to 10^{24} erg, i.e., that of nanoflares. The polar coronal holes are found to be very active and full of nanoflares, and tiny x-ray jets often occur from these nanoflares.

X-ray plasma ejections: The *Yohkoh* soft x-ray telescope has discovered another new ejection feature, *x-ray plasmoids* (or *x-ray plasma ejections*). X-ray plasmoids are blob-like hot plasma ejections, and are sometimes seen as an erupting x-ray loop. They are ejected not only from long duration flares but also from compact impulsive flares, so that they may be evidence for magnetic reconnection occurring above the soft x-ray loop. Although x-ray plasmoids are morphologically very different from x-ray jets, there are common properties in both hot plasma ejections. Table 1 summarizes the physical conditions of both x-ray jets and x-ray plasmoid ejections.

Coronal mass ejections: Some large jets have shapes intermediate between well-collimated jets and loop-like ejections which resemble a SOLAR CORONAL MASS EJECTION (CME).

The physics of x-ray jets/mechanisms of formation

Various observations suggest that magnetic reconnection is actually occurring during the ejection of x-ray jets. Hence the magnetic reconnection mechanism is a most promising energy release mechanism for x-ray jets. The current sheet for the magnetic reconnection mechanism is created in the following situations: (1) the current sheet is created between emerging flux and the coronal field (emerging flux model; see figure 5), (2) the current sheet is created between a horizontally moving bipole and the network field (converging flux model), (3) the current

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Physical quantities of x-ray jets and x-ray plasmoids (based on Ohyama *et al* 1997 *Adv. Space Res.* **19** 1849).

	X-ray jets	X-ray plasmoids
Temperature (10^6 K)	4–6	5–16
Density (cm^{-3})	3×10^8–4×10^9	10^9–10^{10}
Mass (g)	10^{11}–10^{13}	10^{13}–10^{14}
Thermal energy content (erg)	10^{26}–10^{28}	10^{28}–10^{29}
Kinetic energy (erg)	10^{25}–10^{27}	10^{27}–10^{28}
Thermal energy of flare (erg)	10^{27}–10^{29}	10^{29}–10^{30}
GOES class of flare	B–C class	M class

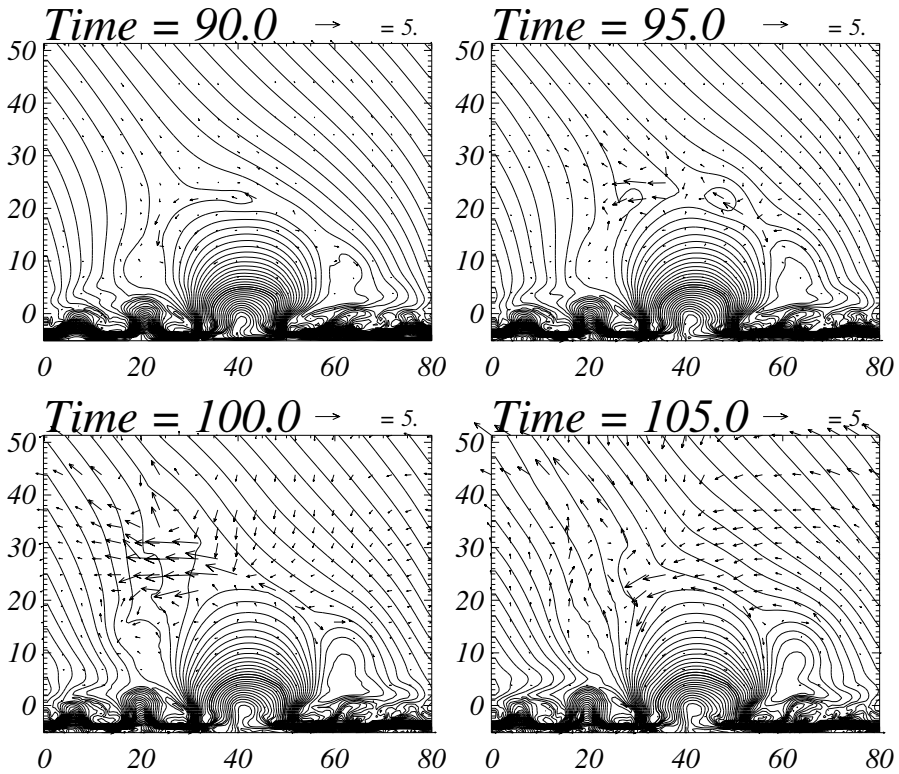


Figure 5. Emerging flux reconnection model of x-ray jets (from a 2D MHD numerical simulation of reconnection driven by emerging flux (Yokoyama and Shibata 1996 *Publ. Astron. Soc. Japan* **48** 353)).

sheet is created between sheared flux and pre-existing flux (sheared flux model).

There is no doubt that x-ray jets exist along magnetic field lines, and hence are collimated by a magnetic force. As for the acceleration mechanism of x-ray jets, there still remains a fundamental question: How are jets accelerated in the magnetic reconnection process described above? There are three possible processes.

Evaporation flow (jet)
An evaporation flow (jet) occurs as a result of a flare that is produced by the reconnection. In this case the jet is accelerated by the *gas pressure gradient force*, which is enhanced by sudden heating of plasmas in the upper chromosphere due to heat conduction flux or high energy electron beams propagating from the flare energy release site. The maximum velocity of an evaporation jet is of the order of

$$V_{\text{ev-jet,max}} \sim 3C_s \simeq 1500 \left(\frac{T}{10^7\ \text{K}}\right)^{1/2} \ \text{km s}^{-1} \qquad (1)$$

where C_s and T are the sound speed and temperature of the flaring plasma, respectively. The apparent velocity of the constant density part of the jet is about 30–300 km s^{-1}, if the density of the jet is 20–40% of that of the flare, consistent with the apparent velocities of many x-ray jets.

Magnetic twist
A magnetic-twist jet is accelerated by the *magnetic pressure force* in a relaxing magnetic twist as a result of reconnection

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

between a twisted loop and an untwisted loop. If the density of the jet is much higher than that of the ambient corona, the maximum velocity of the magnetic-twist jet becomes

$$V_{\text{mt-jet,max}} \sim 2v_A \simeq 2000 \frac{B}{10\ \text{G}} \left(\frac{n}{10^9\ \text{cm}^{-3}} \right)^{-1/2}\ \text{km s}^{-1} \quad (2)$$

where v_A is the Alfvén speed, B is the magnetic flux density and n is the hydrogen number density, but again the apparent velocity of the constant-density part of the jet is of the order of 15–300 km s^{-1} if the density of the jet is 20–40% of that of the flare.

Reconnection jet
A reconnection jet is accelerated (like a slingshot) by the *magnetic tension force* in the reconnection process. The velocity of the jet is of the order of the local Alfvén speed

$$V_{\text{rec-jet}} \sim v_A \simeq 1000 \frac{B}{10\ \text{G}} \left(\frac{n}{10^9\ \text{cm}^{-3}} \right)^{-1/2}\ \text{km s}^{-1}. \quad (3)$$

The velocity of magnetic islands (or plasmoids or expanding helical loops ejected from current sheet) is also comparable to the local Alfvén speed.

Since the observed velocities of many jets are within the range of possible speeds from the above mechanism, at present it is not possible to identify which jets belong to which type. We need more detailed theoretical modeling as well as finer observations such as Doppler shift measurements.

Bibliography
Shibata K, Ishido Y, Acton L *et al* 1992 *Publ. Astron. Soc. Japan* **44** L173
Shibata K 1996 *Adv. Space Res.* **17** (4/5) 9
Shibata K 1998 *Proc. Solar Jets and Coronal Plumes* (ESA SP-421) p 137
Shimojo M *et al* 1996 *Publ. Astron. Soc. Japan* **48** 123
Yokoyama T and Shibata K 1995 *Nature* **375** 42

Kazunari Shibata

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Stellar Connection

The solar–stellar connection refers to the complementarity of solar and stellar studies in general, and to those concerning dynamic phenomena in stellar outer atmospheres in particular. The most important agent in these phenomena is the stellar magnetic field, although wave motions associated with convection also play a role.

The Sun is the only star that offers the potential to observe its internal structure (through helioseismology) and the dynamic geometry of its (outer) atmosphere (SOLAR TELESCOPES AND INSTRUMENTS: GROUND, SPACE) in great detail. On the other hand, observations of stars like the Sun provide tests of our understanding of the physical processes that determine solar structure and dynamics because we can use the diversity of the stellar population to probe dependences on, for example, chemical composition, mass, rotation rate or binarity. In traditional astrophysics, the combined study of the quasi-steady properties of Sun and stars in general have deepened our understanding of the solar interior, of solar evolution and of radiative transfer in stellar atmospheres (SOLAR SPECTROSCOPY AND DIAGNOSTICS).

Whereas effects of solar magnetic activity have been studied ever since the discovery of sunspots by European observers in the early 17th century (although the phenomenon was already reported on at least 23 centuries ago), the study of the stellar equivalent had to wait until the second half of this century. Pioneering studies of stellar chromospheres, that is of the relatively warm layers immediately above the stellar surface, in the 1950s by Olin C WILSON at MOUNT WILSON OBSERVATORY, gave the study of stellar magnetic activity its initial momentum, which has continued to grow over the years. By now we know that stellar outer atmospheric processes that occur on time scales of up to centuries rather than millions of years are the result of the fact that most stars like the Sun have magnetic fields threading their surfaces.

Stellar magnetic activity is characterized by highly structured magnetic fields at the stellar surface that persist for times much longer than the local time scales for convection. The fields at the stellar surface lie beneath volumes within the outer atmosphere in which the temperatures significantly exceed that of the surface and which show up as bright at selected wavelengths. The details of the mechanisms heating those domains are still unclear, but models either rely on the dissipation of electric currents or on the damping of a variety of waves traveling through the magnetic field. The amount of energy involved and the way it is distributed over the different temperature domains in stellar outer atmospheres depends on the magnetic flux density at the surface, and appears to adhere to the same rules regardless of stellar structure: the Sun and almost all other cool stars define simple relationships between radiative losses from different domains in their outer atmospheres that are independent of stellar properties. The energies involved in any outer-atmospheric layer extend well beyond the range seen on the Sun, however.

Through the solar–stellar connection we have learned that the Sun is a rather inactive star, typical of its class. We now know that some stars show activity cycles, like the Sun, but that these actually form a minority. The levels of activity in cool stars cover an enormous range. The most active stars have more than 50% of a visible hemisphere covered by dark starspots, suggesting that most of their surface is in fact covered by a strong magnetic field, only part of which 'condenses' into dark spots. In inactive stars, the coverage by spots is insufficient to be detectable by current white-light instruments, but there the associated emissions from the hot outer atmosphere can be used as a measure of activity. Inactive giant stars have been observed for which the entire stellar corona, with temperatures exceeding 1 million kelvin, is dimmer than the darkest coronal regions on the Sun (the so-called coronal holes).

We have also learned that tidal interaction between companion stars in compact binary systems enhances activity, that stellar rotation is slowed over time by stellar winds flowing away from cool stars, that heating of the outer atmosphere occurs both because of the action of the magnetic field and because of the damping of purely acoustic waves, and that convection requires rotation to excite a globally functioning dynamo that can generate and maintain the stellar magnetic field. The following sections provide more information on each of these topics.

A brief historical perspective

Signatures of stellar magnetic activity have been observed for almost a century now through the white-light modulation caused by the stellar equivalent of large sunspots, or clusters of these, rotating onto and off stellar disks. Early this century, emission components were discovered in the cores of the chromospheric H and K resonance lines of singly ionized calcium in the spectra of many stars of spectral type G and cooler (the solar spectral type is G2V). The discovery paper by Karl SCHWARZSCHILD and Eberhard, published in 1913, was followed by a stream of usually brief papers publishing results on one or a few stars. In the early 1950s, a vigorous study of the Ca II H and K resonance lines in stellar spectra was started by Olin C Wilson at Mount Wilson Observatory. This research has continued to date, and has resulted in records of stellar chromospheric activity, with frequent if not daily observations spanning more than two decades for about 100 stars in the solar neighborhood.

In the last two decades of the twentieth century, a series of spacecraft were launched to observe the (extreme) ultraviolet and (soft) x-ray signatures of stellar activity. The HEAO-B EINSTEIN Observatory and the INTERNATIONAL ULTRAVIOLET EXPLORER were the first to observe extensive samples of cool stars at these short wavelengths with low but useful spectral resolution. Later the spectroscopic capabilities of the EXTREME ULTRAVIOLET EXPLORER and of instruments on the HUBBLE SPACE TELESCOPE

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

provided much more detailed spectral information from which atmospheric densities and temperatures could be derived.

Convection and acoustic phenomena

As we work down in mass from heavier (hotter) to lighter (cooler) stars on the main sequence (generally referred to as dwarf stars or even simply as dwarfs), near-surface convection sets in around a surface effective temperature of ≈9000 K (Sun: 5770 K) and persists down to the coolest red dwarf. For evolved (sub-)giant stars, in which the hydrogen in the core is exhausted, the convection sets in at progressively lower surface temperatures with decreasing surface gravity. Stars with a surface temperature below this onset temperature are referred to as cool stars. For main-sequence stars, the ratio of the depth of the convective envelope to the stellar radius increases with decreasing effective temperature; the coolest main-sequence stars are entirely convective. In absolute terms, however, the depth of the convective envelope is about the same for all of these stars cooler than the Sun, but the stellar radius decreases with surface temperature. For giant stars, the convective envelope deepens steadily with decreasing temperature, and therefore with age, at least during the initial phases of post-main-sequence evolution; for these evolved stars, the convective envelopes become very deep as the stars expand, reading depths up to ten times the solar radius.

Stars with masses $M_* \lesssim 1.3\,M_\odot$, where $M_\odot$ is the solar mass, maintain a convective envelope throughout their main-sequence and post-main-sequence evolution up to the giant branch, i.e. as long as nuclear fusion persists. In contrast, more massive stars have no convective envelope on the main sequence or in the early post-main-sequence phase, but develop one after crossing the dividing line for convective envelopes in the HERTZSPRUNG–RUSSELL DIAGRAM.

Convective motions generate sound waves. This happens most efficiently just below the stellar surface where the plasma velocities are largest in order to be able to transport the stellar luminosity despite the low specific heat of the plasma at the low densities there. With the velocity, the wave power increases with decreasing surface gravity, and of course with increasing luminosity. Waves that are reflected near the surface (either directly or after being deflected upward after traveling inward) are used in helioseismology. Waves with a frequency above the acoustic cut-off frequency penetrate into the outer atmosphere, i.e. into the part of the atmosphere above the layers emitting the bulk of the stellar luminosity.

In the outer atmosphere, the density decreases rapidly within the first few hundred kilometers above the surface as a consequence of the strong stratification: the pressure scale height is only of the order of 100 km in main-sequence stars. Energy conservation results in a dramatic increase in the amplitude of the sound waves as they travel upward, until the waves develop into rapidly dissipated shock waves. The chain of processes involved, from radiation damping in the stellar photosphere to the dispersive and dissipative behavior of weak shocks, strongly reduces the dependence of acoustic energy deposition on surface gravity. Complementary solar and stellar observations suggest that sound waves result in a so-called basal emission from, primarily, the chromosphere. The basal emission depends on surface temperature, but little on gravity or chemical composition. The basal atmosphere consists of rapidly evolving, localized, fluctuating brightenings that recur a few times as trains of shock waves travel by.

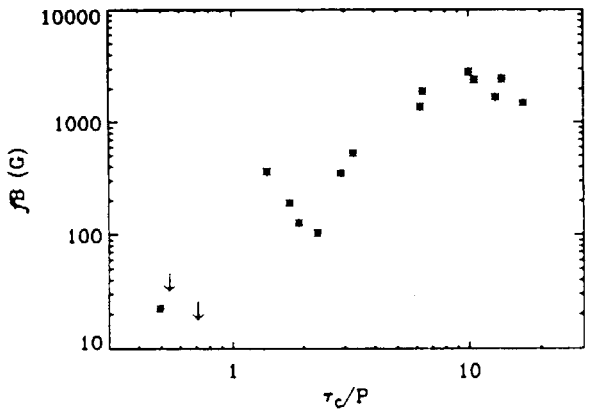


Figure 1. The measured average surface flux density of magnetic field versus the Rossby number in a sample of main-sequence stars. The Rossby number is the ratio of the convective turnover time scale τ_c for convection at the bottom of the convective envelope to the stellar rotation period P, and thus a measure for the effectiveness of Coriolis forces in affecting the circulation of mass. The figure shows how magnetic activity increases with rotation rate.

The basal emission was discovered initially in stellar data through the comparison of radiative losses from different temperature domains in the outer atmospheres of cool stars. It found subsequent empirical support in solar data with spatial and temporal resolution that demonstrated the absence of intrinsically strong magnetic fields at those locations from which this basal, pulsating chromospheric emission originates.

Solar and stellar magnetic fields

Magnetic field at the stellar surface

Cool stars other than the Sun have similar magnetic fields, but attempts to measure these fields by means of the polarization of Zeeman components in spectral lines necessarily fail, because the polarities of the field are strongly mixed, so the net polarization signals from neighboring unipolar patches cancel almost entirely (see SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). Strong fields with sufficiently large filling factors are detectable, however, by a broadening of Zeeman-sensitive spectral lines even in the unpolarized intensity spectrum. Note that measurements in the visual spectrum severely underestimate the contributions of starspots to the total magnetic flux because of their relative darkness.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Moderately active stars of spectral types G and early K have intrinsic field strengths B ranging from 1.0 to 1.9 kG, and fractional surface coverage (or surface filling factor) f ranging from about 1.5% to 35% (see figure 1). The very active Ke and Me stars (i.e. stars with chromospheric lines in emission over the nearby continuum) exhibit B up to 4.2 kG and f up to 70%. These field strengths are compatible with predictions from stellar atmospheric models based on the solar example, which have the field strength dictated by a balance between gas and magnetic pressures. For evolved giant stars, B decreases from $B \simeq 650$ G for early G-type giants to 400 G at K5 III, in accord with the decreasing gas pressure at the surface.

Starspots and stellar active regions

As stars rotate, their outer atmospheric signal as detected on Earth is modulated, because of the inhomogeneous coverage of their surface by magnetic field. This causes the observed signal to change as regions rotate into and out of view. The resulting signal is too confusing to interpret for stars with rotation rates like that of the Sun. For much more active stars, such as the rapidly rotating tidally interacting binaries, the signal is less ambiguous. Their active regions, the regions of strong, bipolar magnetic activity, can be very different from solar ones. Observations of one such binary, AR Lac, for example, show that the surface filling factor for active regions varied from 2% to 20% over the years. The associated mean radiative flux density from its active regions in the chromospheric Mg II resonance lines somewhat exceeds the very top of the flux range observed for stars as a whole, as one would expect. It is, however, three times larger than the flux from the brightest non-flaring parts of SOLAR ACTIVE REGIONS, while for the areas surrounding these bright regions, a flux density was inferred comparable to the flux density for an average solar active region. Clearly, the activity of this binary star is not reflected by a solar counterpart.

Modeling light curves of spot-studded stars is often the only way to obtain information on active regions on stars, but the derived information can be ambiguous. The observed signals may show reasonably large amplitudes, with long-term coherence, so that simple light-curve analyses require only a few, typically two, spot groups to match much of the modulation curve. Yet many more spots or spot regions may exist that would result in a similar light curve. What is abundantly clear, however, is that these spots or spot complexes can be very substantial. XX Tri and II Peg are extreme examples, with amplitudes in the visible of 0.6 and 0.5 magnitudes, or factors of 1.7 and 1.6 respectively. For II Peg this amounts to a spot coverage of 60% of one hemisphere, whereas for the Sun the spot filling factor never exceeds 0.2%. Spot temperatures (generally taken to be the same for all spots on a star, averaging over the dark umbrae and the less dark, more horizontally structured penumbrae) appear to lie some 500 K to 2000 K below the ambient surface temperature, with an average temperature difference of the order of 1000 K.

For sufficiently rapidly rotating stars, emissions from various features on stellar disks are subject to Doppler shifts associated with the line-of-sight component of the rotational velocity. These shifts can be used to make crude Doppler maps of the surface. In two thirds of the active stars for which such surface maps have been made, indications have been found for spots at high latitude or even at the rotational poles. This set of stars is dominated by tidally interacting RS CVn binaries, but also includes young T Tauri stars, single rapidly rotating FK Comae stars, and W Uma (near-)contact systems. No clear correlation has been found between stellar class and the occurrence of polar starspots. The existence of high-latitude starspots in rapidly rotating stars may be the consequence of strong Coriolis forces on flux rising from deep within the star, which are expected to deflect rising flux bundles towards latitudes of up to 60°. Truly polar spots could result either from flux eruption originating very deep in stars with relatively small radiative interiors, or by a pole-ward slip of the previously anchored deep segments of flux rings following an earlier eruption of flux at mid latitudes elsewhere on that ring.

The distribution of sizes and lifetimes of stellar active regions have not yet been determined, leaving the Sun as the only star for which this information is known. A lower limit to the lifetimes of stellar active regions is probably given by the lifetimes of (clusters of) starspots, but even these are difficult to determine. The most active stars, the tidally interacting RS CVn binaries, have had spots reported that appear to live from a few months up to several years, although significant changes from one rotation to the next have been reported, suggesting that small-scale changes to the spots, or more likely to spot groups or nests, occur on a time scale of no more than weeks.

For the Sun, the size distribution of active regions is a monotonically decreasing function of region size, extending from the smallest ephemeral regions up to the largest active regions. For regions larger than ephemeral regions, that distribution is independent of the phase of the solar cycle, only changing by a multiplicative factor. Lifetimes increase from only hours for ephemeral regions to about two weeks for the largest active regions, before the flux is effectively distributed into the surrounding quiet regions.

Atmospheric structure

Diagrams relating the radiative losses from the different temperature domains in the outer atmospheres of cool stars show a remarkably high degree of correlation (figure 2). It appears that the radiative flux densities are related through power-laws, such that two flux densities $F_{i,j}$ are related through $F_j \propto F_j^{\beta_{ij}}$. These power-law relationships are independent of the stellar effective temperature or surface gravity, provided that the basal emission component is removed. The relationships then hold for stars ranging from early F-type where convection sets in, to mid M-type stars, and from main-sequence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

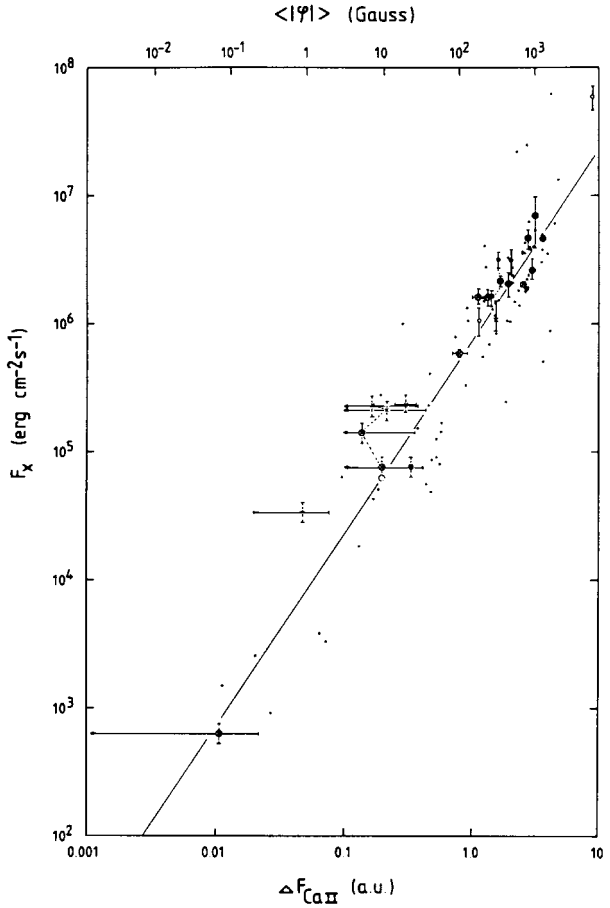


Figure 2. Relationship between the soft x-ray flux density, F_X (0.05–2 keV; observed with EXOSAT's CMA/L1 assembly), and the Ca II H+K excess flux density ΔF_{CaII} (one unit equals 1.3×10^6 erg cm^{-2} s^{-1}). The line is a power-law with slope 1.5. EXOSAT and Mount Wilson Ca II H+K measurements were obtained within three days for 13 of the stars, and within three months for the others. Small dots are used for the comparison of Einstein IPC and Mount Wilson Ca II H + K flux densities that were measured up to several years apart. The scale at the top of the diagram translates the Ca II H + K emission to the equivalent average magnetic flux density.

stars up to at least luminosity class II–III of the semi-bright giants; the only exception is formed by the M-type main-sequence stars beyond $B - V \approx 1.3$. The scatter about the mean relationships is caused by measurement uncertainties, intrinsic stellar variability, and differential limb darkening. The steepness of these relationships increases as the difference in the formation temperature of the two diagnostics used increases, with the most extreme cases being close to quadratic relationships.

Solar data observed with moderate spatial resolution of several thousand km up to about 10 000 km exhibit the same relationships as the disk-integrated signal for stars. Surprisingly, stellar power-law relationships extend beyond the range of flux densities seen in spatially resolved solar data: in the coronal soft x-ray emission, the stellar relationship spans more than five orders of magnitude, ranging from an order of magnitude dimmer than a solar coronal hole, to more than two orders of magnitude brighter than an average solar active region. The combination with solar data has allowed a calibration of the stellar radiative losses in terms of magnetic fields (figure 2), showing that chromospheric losses scale roughly as the square root of the magnetic flux density, and coronal soft x-rays scale roughly linearly.

The agreement of the solar and stellar relationships between radiative losses is surprising given their nonlinear character. After all, simply increasing the number of active features on a stellar disk should have resulted in linear scaling relationships for stars, even if the locally valid relationships are nonlinear. The analysis of solar information has shown that the distribution of magnetic flux over the solar surface is the key behind this invariance of relationships: the histogram of flux densities over the solar and (by inference) stellar surfaces is such that it preserves the nonlinear character upon integration over the visible hemisphere. The requirements this puts on the dynamo and the surface dispersal of flux have yet to be explored.

Chromospheric modeling for stars is complicated because of the high degree of structure and the effects this has on radiative transfer. A comprehensive atmospheric model is hampered by the fact that for a star like the Sun, the chromospheric emission is dominated by the relatively quiet regions, while the coronal emission is dominated by the active regions, thus making it impossible to develop a consistent model in the absence of knowledge of the detailed surface distribution of magnetic field.

Main-sequence stars of moderate activity show a dominant coronal emission component at about 2–4 MK, which is generally assumed to be directly comparable to the most common component in the quiescent solar corona, i.e. a combination of relatively stable active-region and quiet-region magnetic arches or loops. A hotter component exists on more active dwarf stars, and on evolved stars in general, even in the absence of obvious large-scale flaring. The peak contribution of this component lies between about 7 MK and 15 MK.

The determination of dimensions of coronal magnetic loops in stellar coronae can be made if the characteristic mean coronal electron density within the loops contributing most to the observed spectrum is known. For the handful of stars for which coronal density diagnostics have been measured, the characteristic densities estimated from diagnostics for coronal components around 3 MK are typically about three orders of magnitude lower than those found for the hotter components around 10–15 MK. The associated high volume emission measure (proportional to the square of the electron density) of the hotter component implies that the volume filling factors of the hotter components are several orders of magnitude smaller than those associated with the 3 MK component. Perhaps we

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

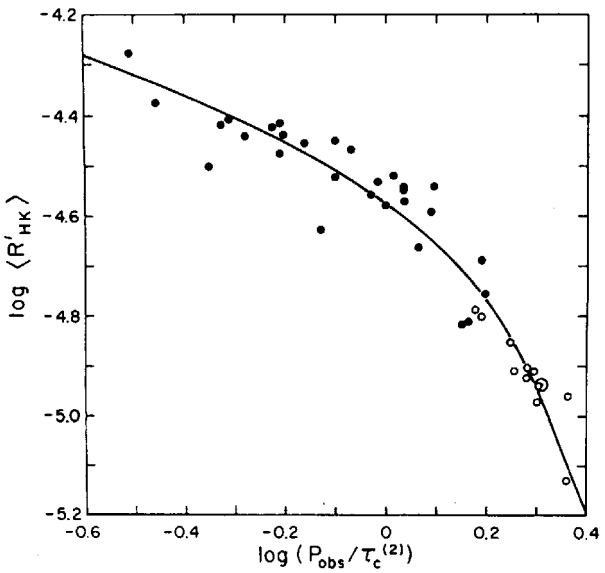


Figure 3. The ratio $R'_{\rm HK}$ of the emission from the chromosphere in the cores of the Ca II H and K lines, after subtraction of a photospheric contribution (including the basal flux), to the total bolometric emission of these main-sequences stars versus the ratio of the rotation period to a color-dependent scaling function $\tau_{\rm c}^{(2)}$ that resembles the convective turnover time at the bottom of the convective envelope for a mixing-length model with a mixing length of two pressure scale heights. Open and closed circles indicate stars below and above the so-called Vaughan–Preston gap, respectively, separating stars with predominantly regular cycles below the gap from those with irregular cycles or non-cyclic behavior above it.

are looking at the signature of a multitude of small, compact, short-lived brightenings occurring over a substantial part of the stellar surface but with a small volume filling factor.

The magnetic dynamo

The outer-atmospheric activity in cool stars has its roots in what is referred to as the dynamo. No comprehensive theoretical model for stellar dynamos has yet been developed (see DYNAMO THEORY). Empirically, however, we have inferred a substantial body of information. For example, we know that a dynamo resulting in magnetic activity appears to operate in all rotating stars with a convection zone directly beneath the photosphere. In single stars, the dynamo strength is a continuous, and largely monotonic, function of rotation rate (figure 3), at least down to the intrinsic scatter associated with stellar variability. It also depends on some other stellar property or properties, but that dependence has not yet been unambiguously determined.

Atmospheric radiative losses associated with magnetic activity observed for main-sequence stars of a given rotation period—and by inference also the dynamo efficiency—increase rapidly with increasing depth of the convective envelope from the onset of convection up to about $B - V \approx 0.6$. A similar increase is inferred for (sub)giants.

A globally functioning dynamo requires only vigorous envelope convection and rotation, but not necessarily a boundary layer between a radiative interior and a convective envelope, because even the fully convective M-type dwarf stars exhibit magnetic activity.

Among the cool main-sequence stars, there is a variety of patterns of variability in their activity; only a minority fraction of these stars show cyclic variations that resemble that of the Sun. For main-sequences stars with moderate to low rotation rates, activity tends to be cyclic, but no clear trend of cycle period with stellar parameters has been found. For truly active stars, various variability patterns exist, but generally no unambiguous activity cycle is seen. The relatively active stars never reach low states of activity on the timescale of years to decades; perhaps multiple sunspot cycles are present on the surface (i.e. a more extreme overlap of successive cycles than seen for the solar cycle), or perhaps the dynamo merely varies around an average state of quasi-permanent activity.

The atmospheric radiative losses appear to 'saturate' for stars with rotation periods below about one or two days, but it remains unclear whether this reflects (i) a saturation in the dynamo action, (ii) in the heating mechanism, or (iii) in the photospheric flux content through increased annihilation of opposite polarities.

The presence of a companion star enhances the dynamo action in cool components of tidally interacting binaries not only by enforcing rapid rotation through tidal interaction (see below), but presumably also by affecting the pattern of differential rotation.

Stellar winds

Measuring the properties of STELLAR WINDS that, like the solar wind, are at coronal temperatures of around a million degrees is difficult, because the tenuous material has little effect on the x-ray, UV and visible spectra emitted by the central star and its outer atmosphere. Winds may be detected through radio and near-IR free–free emission, but only with difficulty even in favorable cases. That mass loss occurs is primarily an inference based on the solar example and the observed effects on the stellar rotation rates.

Despite the existence of at least some potential diagnostics, only upper limits have been derived for mass loss rates of cool stars on the main sequence or in the early phases of post-main-sequence evolution. For late-A and early-F type stars and for a few M-type dwarfs, upper limits to the mass loss have been found of 10^{-10}–$10^{-9}\ M_\odot\ {\rm yr}^{-1}$, which may be far above actual values (the solar mass loss is $\approx 2 \times 10^{-14}\ M_\odot\ {\rm yr}^{-1}$).

For most stars well away from the main sequence, the mass loss rates are large, and the winds cool; in late-K and M-type (super-)giants, (much of) the wind is at chromospheric temperatures, which makes detecting signatures of the wind easier. One case for a cool main-sequence star is V741 Tauri, in which a white dwarf revolves around a K2V star. It has a wind at a velocity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of 600–800 km s^{-1}, a wind temperature of no more than a few times 10^4 K, and a mass loss rate of a few times $10^{-11}\,M_\odot\ \mathrm{yr}^{-1}$, i.e. a thousand times the solar mass loss rate, possibly dominated by coronal mass ejections rather than a steady outflow.

For the coolest giants and supergiants that have no detectable coronal emission in their outer atmospheres, the mass loss rates scale roughly with the surface area of the star, $\dot{M}_*/M_\odot \approx 2.5 \times 10^{-14}(R_*/R_\odot)^2$.

Indirect effects of winds: stellar spin down

Outer-atmospheric activity is accompanied by loss of stellar angular momentum, which in all stars up to and including at least giant stars is an important process in the deceleration of stellar rotation.

The rotational velocities of subgiants decrease strongly within a period of roughly 50 million years following the onset of deep envelope convection. During this phase of the stellar evolution, the moment of inertia of the envelope, I_{CE}, is less than 1% of the stellar value, I_*. As the stars evolve, the convective envelope deepens. If the wind torque would act only on the envelope with little interaction between envelope and interior, then the interior would be left rotating substantially faster than the envelope, and the deepening envelope in the evolving star would dredge up angular momentum from the radiative interior, and then spin up. Such a resurgence of rotation rate and associated activity is not observed, and so much of the interior's angular momentum must have been removed before I_{CE} reaches a substantial fraction of I_*, i.e. well before an effective temperature of about 5250 K (at about G2 III), requiring at least moderately strong coupling between most layers in the interior of the stars.

Stars are born with a broad range of initial rotation velocities. The rotational evolution of young stars can be studied particularly well using data on open clusters. By the age of the α Persei cluster, some 50 Myr, there is a substantial spread in rotation velocities, which narrows remarkably by the age of the Pleiades cluster, at only 70 Myr for the G and K-type stars. The swift change implies a spin-down time scale of the order of the age difference between these two clusters, i.e. only $\sim$20 Myr. By the age of the Hyades, some 600 Myr, the spread has been reduced strongly, and a reasonably well-defined, monotonically decreasing relationship exists between rotation rate and effective temperature, extending from about F0 down to late of K.

An important property of stellar rotational history is the convergence of rotation rates that occurs for ages beyond $\sim$1 Gyr: for stars of solar age, the details of the initial dispersion in rotation rates have been effectively lost, both where the surface rotation rate and where internal differential rotation are concerned. It is this reduction of the initial dispersion that allowed Skumanich to find the age–activity relationship for mature main-sequence stars $v_{\mathrm{eq}}(t) \propto t^{-1/2}$. This convergence now appears to be the consequence of the nonlinear response of angular-momentum loss to angular velocity.

An interesting side effect of the loss of stellar angular momentum through the magnetized stellar wind occurs in tidally interacting binaries. Let us start with a system in which the two component stars have been synchronized with the orbital motion. As one or both of the stars lose angular momentum, the tidal coupling will tend to keep the stars rotating at the orbital period. But that means that part of the orbital momentum is transferred into the rotational momentum. As a result the orbit shrinks. This leads to a higher rotation rate, higher activity and thus a more effective loss of angular momentum through the wind. The result is that the stars spiral in towards each other, until they ultimately merge. This has been proposed as the explanation for rapidly rotating, apparently single evolved FK Comae stars that should, had their evolution been that of a single star all the time, have only a very low amount of angular momentum left.

Bibliography

Jokipii J R, Sonnett C P and Giampapa M S 1997 *Cosmic Winds and the Heliosphere* (Tucson, AZ: University of Arizona Press)

Schrijver C J and Zwaan C 1999 *Solar and Stellar Magnetic Activity* (Cambridge: Cambridge University Press)

Strassmeier K G and Linsky J L 1996 *Stellar Surface Structure: Proc. IAU Symp. 176* (Dordrecht: Kluwer)

Ulmschneider P, Priest E R and Rosner R 1991 *Mechanisms of Chromospheric and Coronal Heating* (New York: Springer)

Carolus J Schrijver

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Stellar Connection: Activity and Brightness Changes of Solar-Type Stars

The surface and outer atmosphere of the Sun are dominated by the presence of localized magnetic fields. They influence the structure of and energy transport through the solar atmosphere and wind, outward to the distance of the heliopause (see SOLAR WIND: MAGNETIC FIELD). In addition, the occurrence of the magnetic features varies on characteristic time scales. Chief among them is the roughly 11 yr cycle of the number of sunspots, which has been recorded telescopically by Galileo and contemporaries since the early 17th century.

With the advent of satellite observations, changes of the total SOLAR IRRADIANCE on time scales of years have been recorded. The total solar irradiance varies with the 11-year activity cycle and by as much as 0.14% from minimum to maximum of the activity cycle; the irradiance is highest at solar maximum.

Any successful theory of the origin and nature of solar surface magnetism would require that similar phenomena be present on stars that are close in mass and age to the Sun, i.e. the solar-type or Sun-like stars. In recent decades it has become possible to measure changes in spectrophotometric and photometric fluxes that indicate changes in both surface magnetism and brightness of Sun-like stars over time scales of years. The characteristics of the observed variations in surface magnetism and brightness resemble those of the Sun and appear to be a universal feature of Sun-like stars.

Surface magnetism

On the sun, the disk-integrated Ca II H (396.8 nm) and K (393.4 nm) emission fluxes increase as coverage by surface magnetic features such as sunspots increases. The magnetic fields provide energy to the CHROMOSPHERE and other outer layers of the solar atmosphere; that energy is shed by emission lines such as Ca II formed in those regions. With the benefit of satellite-based observatories, many other emission lines similarly energized by the magnetic fields are available for study in the ultraviolet and x-ray spectrum regions. However, the Ca II H and K fluxes have historically provided much information on magnetic activity in lower main sequence stars because the spectrum features are accessible to ground-based instruments.

Measurements of Ca II H and K emission fluxes of stars are necessarily integrated over a star's visible hemisphere. It is assumed that the observed variations of the Ca II emission fluxes in Sun-like stars are caused by changes in surface coverage or magnetic field strength on the unresolved stellar disks, and therefore the Ca II fluxes contain information on stellar surface magnetism.

Two aspects of surface magnetism derived from spectroscopic indicators of magnetic activity in solar-type stars are the fluxes averaged on evolutionary time scales (i.e. a billion years or so) and their year-to-year variability. Both cross-sectional (time-averaged) and time serial studies have been made in order to investigate the origin and nature of magnetic activity in Sun-like stars.

Cross-sectional studies of surface magnetism

Olin C Wilson, working at the 100 in Hooker telescope at Mount Wilson Observatory, and other astronomers established four decades ago the basis of cross-sectional studies of average Ca II fluxes in lower main sequence stars.

The average surface magnetism, indicated by the time-average value of the Ca II fluxes, shows a dependence on axial rotation. In addition, rotation is observed to slow on long time scales, i.e. with increasing main sequence age.

The body of evidence suggests that, shortly after a solar-mass star arrives on the main sequence, its rotation rate and time-averaged flux of Ca II emission show similar values among stars close to it in age, and the values of those parameters decrease gradually with age. That means that the observed parameters of age, rotation and Ca II fluxes are closely coupled in Sun-like stars. A similar relation can be found among lower main sequence stars of different masses. In other words, the average Ca II emission and the rotation rate of a lower main sequence star, including the Sun, are determined primarily by the star's main sequence age and mass.

The problem of creating, maintaining and transforming magnetic fields by means of plasma fluid motions, which is relevant to planetary, solar, stellar and galactic magnetic fields, can be studied with the magnetohydrodynamic dynamo. Briefly, DYNAMO THEORY suggests that surface magnetism can be described by the dynamo number, D, which depends on mean helicity and differential rotation shear. In a lower main sequence star, those plasma motions may be linked to the observable parameters of mass (depth of the convection zone) and age (or axial rotation, which may be related to internal differential rotation). It is the concept of the dynamo number that leads to the expectation that stars close in mass and age (i.e. rotation) would show similar attributes in surface magnetic activity. In contrast, stars differing greatly in mass and age (i.e. rotation) should exhibit magnetic activity that is dissimilar.

The interpretation of the observational results is that the production of magnetic fields causes a star's surface angular momentum to decrease with age. The slowing of rotation with time is a sign of a gradual loss in angular momentum caused by the presence of a magnetized stellar wind, which presumably obtains throughout the star's main sequence lifetime. As a consequence of the production of magnetic activity, the dynamo's efficacy decreases with increasing age, and angular momentum is lost. The slowing rotation weakens the production of magnetic fields and the fluxes of emission lines that are driven by the presence of magnetic fields.

The coupling between Ca II emission, age and rotation is so close among lower main sequence stars that knowledge of any one parameter, along with an estimate

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

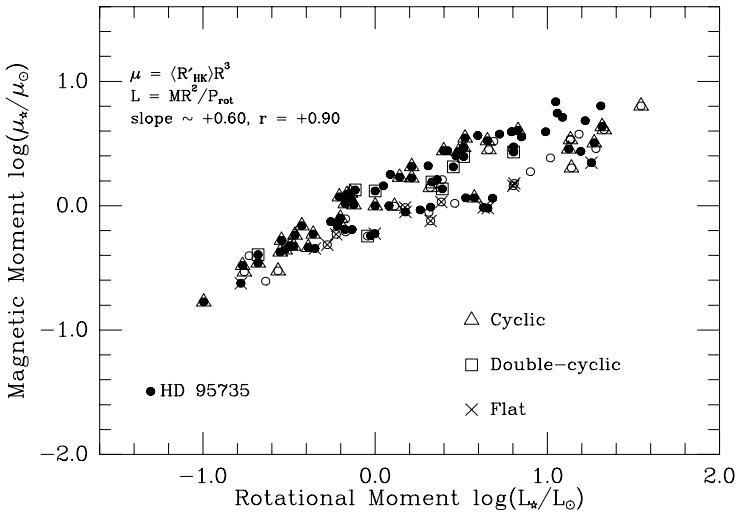


Figure 1. The magnetic moment is plotted against rotational moment, both values normalized to those of the Sun, for the Sun and 112 stars from Wilson's sample of stars on or near the lower main sequence. Measured rotational moments are denoted by filled circles; estimated rotational moments are denoted by open circles. The slope for the $\log(\mu)$–$\log(L)$ relation is $d \sim +0.60$ and the Pearson correlation coefficient is ∼+0.90. The three classes of variability in Ca II fluxes based on the 25 yr records discussed in Baliunas *et al* (1995) are also indicated. Points enclosed by triangles denote stars with a periodic cycle; points enclosed by squares denote stars with two distinct periods ('doubly cyclic'); points with crosses denote stars with little Ca II variability. (Reproduced from *The Astrophysical Journal* by permission of the authors.)

of a star's mass, leads to good estimates of the other two parameters. One of several possible depictions of the relation among those parameters in a sample of lower main sequence stars, whose spectral types extend from mid-F to late K, is illustrated in figure 1. There the magnetic moment, calculated from the Ca II emission fluxes, is seen to depend on the rotational moment. The existence of the close relation is tantamount to saying that the mechanism generating magnetic activity is affected by and affects a star's rotation. The process by which magnetic fields are produced influences angular momentum to the extent of erasing differences in rotation among lower main sequence stars of similar mass and age that surely must have held in the pre-main-sequence phase.

Decade-scale variations in Ca II fluxes

Between 1966 and 1978, Wilson carried out a program at Mount Wilson Observatory of monthly monitoring of the Ca II H and K emission fluxes in 91 lower main sequence stars. Variations of Ca II emission fluxes on time scales of years seen in the records of Sun-like stars are assumed to be produced by a mechanism similar to that causing the 11 yr solar activity cycle, e.g. the magnetohydrodynamic dynamo.

Wilson's purpose was to provide records on the counterpart of the solar activity cycle in lower main sequence stars. Since 1978, not only have the records of Wilson's sample been continued, but also the program has expanded in scope. The longest records now stretch back more than three decades. Monitoring of a much larger sample of nearly 150 Sun-like stars began in 1991.

At times, stars in the program have been observed as often as nearly every night. Such densely sampled records reveal axial rotation from the modulation of Ca II fluxes when magnetic surface features, asymmetrically distributed on the surface, persist for several weeks. Analysis of densely sampled records suggests the presence not only of surface differential rotation but also of features stable on time scales as long as several years, although not particularly for the stars close in mass and age to the Sun.

The three classes of magnetic variability

Across the sample of lower main sequence stars, Wilson found in the Ca II H and K records three broad classes of variations on time frames of years: (1) cyclic; (2) variable; (3) non-variable. The cyclic, or periodic, variations show a pronounced periodicity like the 11 yr solar cycle. The variable records display major fluctuations but no persistent periodicity. The non-variable records are essentially constant with time, to the limits of detection. Examples of each class of magnetic variability in Sun-like stars are shown in figure 2.

The cyclic records can be found among lower main sequence stars of nearly any age and mass. The variable records tend to occur in the younger stars. The flat records may be occurring only in the Sun-like stars.

Figure 2 may be interpreted as a crude evolution of surface magnetism in Sun-like stars. A young star (with an age less than 2 billion years or so) of about one solar mass is rapidly rotating and often presents a variable record (top panel). An older star (with an age beyond two billion years or so) near one solar mass, i.e. a sun-like star, shows either

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

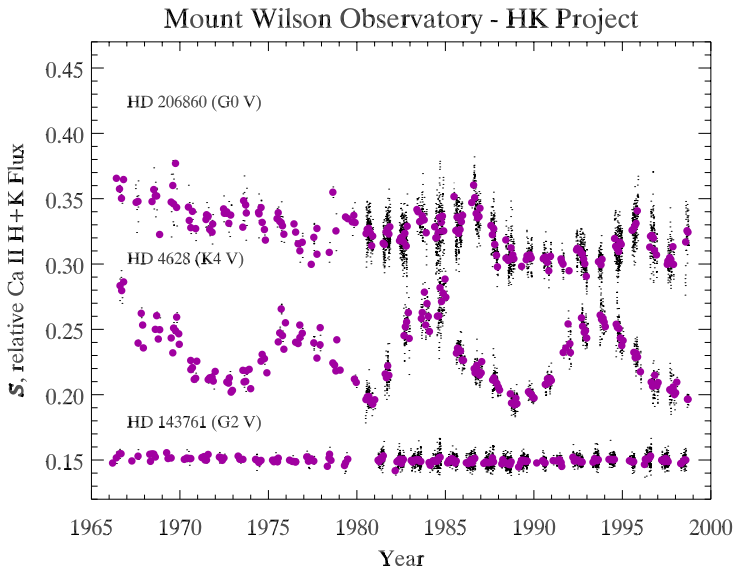


Figure 2. Records of Ca II H and K fluxes in three stars whose masses are similar to the Sun's. The record of a star whose age is about 1 billion years (upper panel) is usually variable with little indication of a preferred periodicity. The records of Sun-like stars—those close in mass and age to the Sun—are generally either cyclic (middle panel) or essentially non-variable (lower panel) to within the precision of the observations (∼1% rms over the length of the record).

a cyclic (middle panel) or a flat (bottom panel) record. The two patterns of records for the Sun-like stars may be the result of sampling relatively short intervals of solar-type variability on time scales of centuries. That is, the Sun shows cyclic magnetic changes a large part of the time. At other times, the Sun's magnetism decreases to a sustained state of little variability, similar to the Maunder minimum (ca 1640–1720 CE) and other periods of low surface magnetism revealed by the cosmogeochemical records of radiocarbon (^{14}C) and ^{10}Be.

At present the Sun is in the cyclic phase and exhibits an 11 yr period. The cyclic phase lasts roughly several centuries. Every few centuries or so, the Sun enters and remains in a low-magnetic state, with a low-amplitude, long-period cycle, if any, and low average flux. This state roughly lasts a century. The low-magnetic states are assumed to be similar to the essentially non-variable records of Ca II fluxes in Sun-like stars. An analysis of the statistics of the number of cyclic and flat records in the relatively short records (∼30 yr) of Sun-like stars suggests a comparable frequency and duration of cyclic and flat phases of surface magnetic variability.

It is still unclear whether, for main sequence stars much older than the Sun, further and more subtle changes in magnetic activity records would occur, for example an increase in the frequency or duration of the Maunder-minimum phase, along with a lengthening of the cycle period during the cyclic phase.

Explanation by the dynamo theory

The evolution of the Sun's magnetism may be further illuminated by dynamo theory. As mentioned above, the dynamo number, D, depends on the product of differential rotation shear and mean helicity. For the youngest solar-mass stars, which are rapidly rotating, the dynamo number is large, with the consequence that several dynamo modes may be simultaneously excited, producing surface magnetism that is variable, with no clear periodicity. As a solar-mass star ages and its rotation slows, the dynamo number diminishes and a characteristic periodicity, indicating a single excited mode, persists. At smaller dynamo numbers, a star's surface magnetism may have a cyclic periodicity combined with occasional periods of low surface magnetism and little variability. At greater ages, and slower rotation, still smaller dynamo numbers suggest that the dynamo may barely operate, and a star may show little variability in surface magnetism. For the smallest dynamo numbers, the dynamo may not be globally sustainable.

It remains difficult to compare results from dynamo theory with empirical results. One reason why comes from consideration of the common argument that the observed cycle period depends on the dynamo number through a power law, $P_{\rm cyc} \propto D^i$, where i is estimated from theoretical considerations. A comparison of various predictions yields a range of possible values for the parameter, i, from $-1.0 < i < +0.5$; that is, not even the sign of the dependence is reliably known. Furthermore, it is difficult to know which theoretical parameters are to be compared with which empirical parameters.

Correlation between magnetic changes and brightness changes

Satellite-based radiometry measurements of the Sun made over the last 20 yr (see SOLAR TELESCOPES AND INSTRUMENTS:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SPACE) show that, over the course of the 11 yr cycle, changes in magnetic activity are positively correlated with changes in total solar irradiance. The total solar irradiance variations arise from changes in the sum of the irradiances of individual surface features, each with a different irradiance, whose surface coverage varies with time. For example, consider just two varying components, visibly dark sunspots and bright faculae, at one phase, activity maximum. Sunspots are dark compared with the magnetically quieter surface and so subtract from the the total irradiance that would result in the absence of sunspots. However, faculae are bright. At activity maximum the area coverage by faculae is more extensive than by sunspots, and the positive irradiance contributions of the faculae more than compensate for the loss by sunspots. As a result, the net irradiance change over the sunspot cycle shows a positive correlation with surface magnetism.

Roughly analogous observations of Sun-like stars reveal a similarly positive correlation over their activity cycles. Techniques that emphasize highly precise differential photometry, e.g. rapid measurement of close-by, non-variable comparison stars, have been used to measure variations of Sun-like stars in the visible (e.g. *b* and *y* Stromgren passbands). The corresponding amplitude of solar brightness change over recent activity cycles is estimated to be roughly 0.1% in the visible spectrum. In Sun-like stars, the photometric brightness is observed to change with the Ca II fluxes over the activity cycle, i.e. the photometric brightness changes are positively correlated with surface magnetism through the activity cycle. Amplitudes of photometric brightness change over the activity cycle as low as the Sun's have been observed for Sun-like stars. Amplitudes several times larger than the Sun's have also been observed.

The existence of the positive correlation between brightness change and activity change over the activity cycle is another important aspect of the similarity of surface magnetism in Sun-like stars. The different amplitudes of observed photometric amplitudes seen among the Sun-like stars suggests that at times the Sun can vary substantially more during an activity cycle than it has during the limited period of recent satellite measurements. In other words, there may exist components of total solar irradiance change that occur on time scales of several decades or longer. Sustained observations of the total solar irradiance should reveal any long-term components of solar irradiance change. It may also be possible to study decades-long changes in one star, the Sun, by considering the cycles observed in a large sample of Sun-like stars, despite their relatively short records.

Applications to studies of extra-solar planets and terrestrial climate change

Studies of the variability of sun-like stars have been used in two topics removed from the study of the theory of surface magnetism. First, stars with little variability, because they are in a temporary or chronic state of low variability, are good candidates for searching for EXOPLANETS by techniques using Doppler or photometric variations. The reason is that the condition of little activity variability produces photometrically and spectrally quiet backgrounds against which to see the subtle planetary effects, if present. Second, if the records of a large sample of Sun-like stars can be considered as different phases of changes in the Sun on time scales of decades to centuries, then information can be obtained on the long-term variability of the Sun and its potential impact on terrestrial CLIMATE change.

Bibliography

Baliunas S L and Soon W 1995 *Astrophys. J* **450** 896–901
Baliunas S L *et al* 1995 *Astrophys. J* **438** 269–87
Baliunas S L, Sokoloff D and Soon W 1996 *Astrophys J.* **457** L99–L102
Radick R R, Lockwood G W, Skiff B A and Baliunas S L 1998 *Astrophys. J. Suppl. Ser.* **118** 239–58
Wilson O C 1978 *Astrophys. J* **226** 379–96

Sallie Baliunas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Terrestrial Connection

Radiation from the Sun and the energy from the SOLAR WIND that crosses the magnetopause have a profound influence on the Earth and its surrounding atmosphere (see also MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). The consequences range from the formation of ozone to the flickering aurorae in both polar regions. The most significant solar variations cause long-term changes in the Earth's atmosphere and intense storms in space. Such storms can damage spacecraft and power generation networks.

The Earth's atmosphere

The EARTH'S ATMOSPHERE is, to a first order, horizontally stratified largely through the effects of gravity. Its structure is best described by temperature (figure 1), which results from a balance between absorption of solar radiation and re-radiation, and horizontal and vertical transport. In the troposphere (0–12 km altitude), the temperature decreases by $\sim$7 K km^{-1}, the key radiatively active molecules being water vapor (H_2O), carbon dioxide (CO_2) and methane (CH_4). Above the troposphere lies the stratosphere where there is a modest rise in temperature owing to the absorption of ultraviolet radiation in the range 290–360 nm by ozone (O_3). Radiative cooling, particularly by nitric oxide (NO), causes a sharp reduction in temperature around 85 km altitude, which lies near the center of the mesosphere. This is the coldest part of the Earth's environment with temperatures of about 150 K. Above this altitude, extreme ultraviolet and x-rays are absorbed leading to temperatures that rise to about 1500 K at 300 km altitude, a region known as the thermosphere. Above 1000 km, the atmosphere remains at a constant temperature with altitude, because diffusion processes dominate. This regime is called the exosphere.

The pressure of the atmosphere, p, at an altitude z, is given by the following equation, where p_0 is the pressure at a reference height z_0, m is the mean molecular mass, T is the temperature, g is gravitational acceleration and k is Boltzmann's constant:

$$p = p_0 \exp\left[-\frac{mg}{kT}(z - z_0)\right]. \qquad (1)$$

The parameter kT/mg is known as the scale height. Typical values are about 7 km at the Earth's surface, 5 km in the mesosphere and 50 km at 300 km altitude.

The mean molecular mass of the atmosphere is constant below about 70 km, because the atmosphere is turbulent and mixing occurs. Above about 70 km, the scale height begins to change, because solar radiation causes molecules, especially oxygen (O_2), to dissociate. Then above about 115 km altitude, turbulence no longer dominates so that the gases begin to separate by diffusion, i.e. each gas responds according to its own atomic mass, as given in equation (1). Thus for a given temperature the scale height of hydrogen (H) is 28 times greater than that of molecular nitrogen (N_2). Above about 1500 km altitude, hydrogen is the dominant gas species.

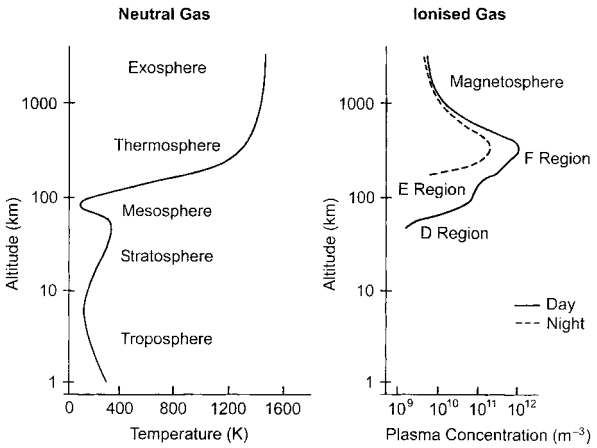


Figure 1. Left: The temperature of the Earth's atmosphere as a function of altitude, identifying the various spheres of the neutral atmosphere. Right: The vertical distribution of ionization both for daytime (full curve) and at night (broken curve). The D, E and F regions are identified. The region above about 1000 km altitude is the magnetosphere.

The ionized atmosphere

Solar radiation ionizes some of the neutral atmospheric constituents in the mesosphere and thermosphere leading to the formation of the ionosphere, which may be defined as the region where there are sufficient free electrons and ions to affect the propagation of radio waves. The ionosphere is embedded within the thermosphere, and it is only above 1000 km altitude that the concentration of the ionized atmosphere exceeds that of the neutral component, and this upper region is known as the magnetosphere (see MAGNETOSPHERE OF EARTH).

The plasma concentration in the ionosphere and magnetosphere may be described by the continuity equation (2), where q is the ionization rate of free electrons (N). The ionosphere and magnetosphere is an electrically neutral plasma, and hence the concentrations of electrons and ions are the same. The parameter, l, is loss rate due to recombination. The final term in equation (2) is a transport term due to the divergence of the flux ($N\boldsymbol{V}$) where $\boldsymbol{V}$ is the plasma velocity.

$$\partial N/\partial t = q - l(N) - \mathrm{div}(N\boldsymbol{V}). \qquad (2)$$

The ionosphere comprises three main regions, the D region (70–95 km), the E region (95–150 km) and the F region (>150 km). These names were chosen early in the 20th century in case additional layers were found at lower altitudes.

In the D and E regions, the electron concentration is dominated by local production and loss. The collision frequency of ions with the neutral particles is so high in these regions that, to a first order, the ions move with the neutral particles. The collision frequency of the electrons is much smaller, and thus they are free to move under the actions of electric fields. Therefore this leads to the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

possibility of ions and electrons moving with respect to each other, which is in essence a current.

The production of plasma depends on the flux of x-rays and EUV, the solar zenith angle, the atmospheric composition and density, the last two parameters combining to form what is termed the optical depth. The height of maximum production is a balance between the increasing density of the atmosphere with decreasing altitude and the diminishing intensity as the ionizing fluxes are absorbed as they penetrate further into the atmosphere. The maximum photo-ionization occurs at different altitude for each wavelength. For example, the peak ion production occurs at about 105 km for $\lambda = 1$–14 nm, 115 km for $\lambda = 80$–103 nm and 150 km for $\lambda = 14$–80 nm. Integrating across all wavelengths the peak production occurs about 110 km for an overhead Sun. In general, the ionization in the D and E regions is dominated by molecular species (NO^+ and O_2^+), whereas in the F region O^+ is the dominant species.

Molecular ions are lost through dissociative recombination, a process whereby molecules break apart during recombination. The atomic ions are lost through radiative recombination, a process that requires the emission of a photon to conserve both momentum and energy. Radiative recombination is a much slower process than dissociative recombination, by a factor of 10^3. The typical lifetimes of ions (and hence electrons) at D, E and F region altitudes are approximately 1 s, 1 min and 1 h respectively. Because there is only a minute amount of photoionization at night from scattered Lyman α and β from the Earth's geocorona the plasma concentration in the D and E region is very low (figure 1).

The F region peak plasma concentration occurs where diffusion of plasma along the magnetic field line is approximately equal to the loss rate of plasma. Above the F region peak, diffusion dominates plasma processes. Equation (1) can be used to describe the plasma distribution but using the average ion and electron temperatures, $(T_i + T_e)/2$, rather than the neutral temperature, T. Above about 110 km for electrons and 300 km for ions, there are insufficient collisions to maintain thermal equilibrium with the neutral particles. Near the F region peak, the ion temperatures are 5–20% greater than the neutral temperatures, but the electrons are between 2 and 5 times greater. Therefore the scale height of the plasma in the ionosphere is considerably larger than that of the neutral atmosphere.

By chance, the ionization potentials of O^+ and H^+ are almost identical, and thus the charge exchange readily occurs ($O^+ + H \leftrightharpoons O + H^+$). The formation of H^+ contributes significantly to the increase of the plasma scale height and is a major reason for the ionosphere acting as a major source of plasma for the magnetosphere. During daytime, the upward fluxes through 1000 km altitude are $\sim 10^{12}$ m^{-2} s^{-1}. During the night, cooling occurs and some of the magnetospheric plasma falls back into the ionosphere, acting to maintain the F region concentration through the night.

During the day, the absorption of solar radiation causes the thermospheric temperature to rise by ~400 K, resulting in a large pressure bulge on the dayside of the Earth at F region altitudes compared with nighttime values. A major consequence is that winds blow at speeds of 100–400 m s^{-1} from the dayside to the nightside thermosphere. The expansion of the atmosphere at E region heights causes differential motion of the electrons and the ions, which causes current to flow, as discussed above. This current changes the intensity of the magnetic field at the Earth's surface by ~50 nT, and the associated dynamo electric field is about 10 mV m^{-1}.

Solar-induced variations in the neutral atmosphere

The total solar irradiance is remarkably constant, although there is about an 8% increase in the flux reaching the ground in January compared with June owing to an eccentricity in the Earth's orbit. Since space-based measurements of solar irradiance began in the 1970s, the variation has been less than 0.2%. Such measurements must be made from space to ensure that observations are not affected by the Earth's atmosphere. It is estimated that a 1% increase of solar irradiance is required to cause a 1 K increase in global temperature. Therefore the solar irradiance variation (~0.2%) is of minor importance in recent global climate change. However, there are over 2000 papers which show statistically significant correlations between tropospheric parameters (e.g. temperature and rainfall) and the 11 yr solar cycle. Therefore there must be less direct physical processes that have an impact on the troposphere (see SOLAR–TERRESTRIAL CONNECTION: LONG-TERM AND SHORT-TERM CLIMATE VARIABILITY).

The daily and seasonal changes in the temperature of the atmosphere are very large compared with those resulting from the increase of greenhouse gases. In the troposphere, models predict that a doubling of CO_2 will lead to a few K rise in surface temperature. The corresponding variation through a year at a given site is of the order of 20–30 K. The mesosphere would cool by ~10 K as a result of increased CO_2 owing to the additional radiative cooling. The natural seasonal variation is about 20 K. In the thermosphere, a 50 K reduction in temperature would occur, with a corresponding variability over a season of about 400 K. Therefore the effects of changes in greenhouse gas concentration are much smaller than the natural changes caused by normal diurnal and seasonal changes at each level in the atmosphere. These comparisons also show that the mesosphere is the best level in which to search for global change effects, because the predicted changes are comparable with the natural variability, whereas in other levels the predicted changes are a small fraction of the natural variations. Alas, mesospheric measurements are difficult as the mesosphere is too high for aircraft and balloon-borne measurements and too low for satellites, but new techniques using optical methods for remote sensing the mesosphere are being developed. The longest data series of the ionosphere

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

extend for only a few decades, short in comparison with meteorological data series. However, even these data sets are indicating a reduction in the altitude of the peak concentration in the F region of up to 1 km yr^{-1}. There is still considerable regional variability in the level of change, but the sense is consistent with increased CO_2 concentrations (Jarvis *et al* 1998).

Solar-radiation-induced variations in the ionized atmosphere

The intensity of the solar radiation in the wavelength range critical for the ionosphere and magnetosphere ($\sim$1–140 nm) varies on average by a factor of about 2.5 between sunspot maximum and minimum, although some specific lines may be enhanced by a factor of $>$100. The consequence for the ionosphere is that the electron concentration increases by approximately the same factor between sunspot minimum and maximum. Similarly, a 27 day modulation of solar flux can be detected in the ionospheric electron concentration both at E and F region altitudes. The 27 day modulation is difficult to identify owing to the large day-to-day variability in the ionosphere caused by geomagnetic storms and wave energy propagating upwards from the troposphere and stratosphere. Measurements of sunspot number and the intensity of the solar radio flux at 2800 MHz (λ = 10.7 cm) are often used in ionospheric models because these parameters are easy to determine at the Earth's surface. However, they are at best only a guide to the general level of activity, rather than a reliable indicator of EUV and x-ray fluxes. These fluxes must be known accurately if ionospheric concentrations are to be predicted reliably.

Solar flares last 1–100 min and release $\sim 10^{22}$–10^{25} J of energy. The largest flares, C class, give an energy flux at the Earth (1 AU) of 10^{-6}–10^{-5} W m^{-2} near 1 nm. This produces a corresponding increase of ionization in the D and E regions of the dayside ionosphere that has profound effects on radiowave propagation, as discussed later.

Solar-wind-induced variations in the ionized atmosphere

Typically 1% of the 10^{13} W of solar wind power that impinges upon the magnetopause is transferred into the magnetosphere when the interplanetary magnetic field (IMF) has a southward component. Under such conditions, reconnection is likely to occur, this being the major physical mechanism by which energy enters the magnetosphere–ionosphere system. When the IMF is northward the energy crossing the magnetopause falls by 2 orders of magnitude. The solar wind energy is redistributed in the magnetosphere by the effects of electric fields, currents, energetic particles and waves, which in turn have significant effects on the stratosphere, mesosphere, thermosphere, ionosphere and magnetosphere (see SOLAR–TERRESTRIAL CONNECTION: COUPLING BETWEEN SOLAR WIND, MAGNETOSPHERE, IONOSPHERE AND ATMOSPHERE). The resultant changes also have a major and increasing effect on human activities (see SPACE WEATHER PREDICTIONS).

The heliospheric current sheet is the boundary between the northern and southern lobes of the Sun's magnetic field. At the Earth, which lies close to the heliospheric equator, the current sheet is distorted and is often likened to a ballerina's skirt. These two factors combined mean that the north–south component of the IMF varies from northward to southward on average every $\frac{1}{2}$ h. Therefore reconnection at the dayside magnetopause is switched on and off frequently, thus the energy flow from the solar wind to the magnetosphere is highly irregular both in space and time.

There is still considerable uncertainty over some of the detailed response of the coupled magnetosphere–ionosphere system to southward turnings of the IMF and thus increased energy transfer. There is general agreement that the electric field across the magnetosphere and the currents that couple the ionosphere to the magnetosphere both increase within minutes. After about 1 h, there is a major reconfiguration of the geomagnetic tail, known as a substorm, which leads to majestic displays of the AURORA in the nightside ionosphere, among other phenomena. Intense currents (1 MA) are driven from the magnetosphere through the ionosphere, which results in enormous Joule (Ohmic) heating. These currents deposit $\sim 10^{11}$–10^{12} W in the ionosphere and thermosphere which is about twice as much as the energy deposited by the energetic particles that cause the aurora. Substorms typically last about $\frac{1}{2}$ h and recur on average every 3–5 h. Hence there are about 1500 substorms per annum. Energetic ions and electrons with high pitch angles injected during the substorm process drift around the Earth at (2–3.5) $\times$ 10^4 km altitude. Their motions are controled by the gradient and the curvature of the Earth's magnetic field and the azimuthal drift are in opposite directions for ions and electrons, hence the formation of the ring current (see MAGNETOSPHERE OF EARTH: RING CURRENT). Typically about 10^{11} W power is stored in the ring current, the energetic particles eventually being scattered by waves and ending up precipitated in the stratosphere, mesosphere and thermosphere.

The major exception to the continually oscillating IMF occurs during the passage of a coronal mass ejection (CME) event. Although CME events occur about twice per day on the Sun, only a small fraction of these impact upon the Earth's magnetosphere. When they occur, the IMF magnitude and direction change very slowly with time. Thus the IMF may have a northward component for 1 day, the consequence of which would be that the aurorae diminish markedly in intensity and move progressively poleward. Similarly the southward component of the IMF may last for 1 day also, and hence considerable energy is transferred to the magnetosphere. During these intervals, frequent and intense substorms usually occur, and their cumulative effect is known as a magnetic storm.

The occurrence of magnetic storms varies over the solar cycle. In general, storms are most frequent up to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

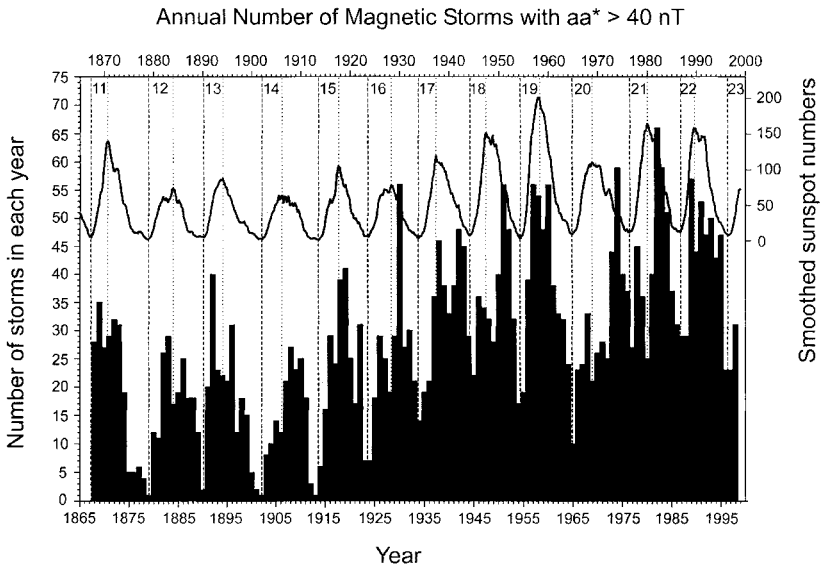


Figure 2. The sunspot number and the number of magnetic storms since 1865. A storm is defined as an occasion when the aa magnetic index exceeds 40 nT. The number of storms at all epochs of the solar cycle has increased markedly in the last 100 yr. (From Clilverd *et al* 1998.)

2 yr after sunspot maximum (see figure 2). However, the number of storms at each phase of the solar cycle has increased markedly over the last 100 yr. Figure 2 shows the number of storms per annum, storms being defined as when the geomagnetic aa index exceeded 40 nT. The aa index is a measure of the deviation of the Earth's magnetic field from its quiet level in each 3 h epoch of the day, as determined from two mid-latitude observatories. The broken vertical lines show the times of sunspot maximum and minimum and the second curve shows the sunspot number.

One of the effects of storms and substorms is to increase the temperature of the thermosphere by up to 400 K. This causes the atmosphere to expand and the density at a fixed altitude near 400 km altitude to increase by an order of magnitude. Also the atmosphere becomes less molecular, i.e. the O/N_2 ratio decreases. The recombination rate increases and the plasma concentration falls. This is most marked at F region altitudes where the concentration falls by up to an order of magnitude. Winds in the thermosphere also increase by a factor of 2–3, reaching up to 1 km s^{-1}. Although the initial energy from substorms is deposited in the thermosphere near the auroral oval, all latitudes are eventually affected through the transport of energy through winds and atmospheric gravity, or buoyancy waves. The latter propagate away from the auroral oval at close to the speed of sound in the thermosphere (750 m s^{-1}).

Solar proton events are when protons with energies of 1–500 MeV are ejected from the Sun. Such events, although infrequent, cause substantial changes to the minor species chemistry, especially O_3 and NO_x, in the atmosphere between 20 and 80 km altitude. They last from a few hours up to 10 days. Relativistic electron events (energies 1–20 MeV) are also observed inside the magnetosphere. The sources of these very energetic particles are still unclear. There are plausible arguments suggesting that the Sun, Jupiter, the dayside cusp or the geomagnetic tail may be the source of such particles. Like the solar proton events, their atmospheric consequence may be very significant in the aerosol chemistry. Both energetic ion and electron precipitation may also affect the global electric circuit. This is the cumulative effect of all the thunderstorms which charge the ionosphere to a potential of several 100 kV with respect to the Earth's surface. This potential difference drives vertical electric currents downward from the ionosphere to the ground in regions where thunderstorms do not occur. The fair weather electric current, as it is called, varies spatially and temporally according to the ionospheric potential and the total column resistance between the ionosphere and the ground. This column resistance may vary when very energetic proton or electron events are in progress.

The human impact of solar variability

All Earth-orbiting satellites fly in the ionosphere and magnetosphere, and satellites are used for a wide variety of purposes now. Modern communications use satellites to an ever-increasing extent. Geostationary satellites are used for television transmissions, and there is a massive increase of global mobile telephone systems that use satellites in a variety of orbits. Most navigation now uses the Global Positioning System (GPS), an extensive constellation of satellites in highly elliptical orbits. Earth observations, e.g. meteorological forecasting, hurricane track predictions and agricultural crop assessments,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

extensively use space-based data. Marconi first used high-frequency (HF) (3–30 MHz) radio communications via the ionosphere a hundred years ago. Despite the advent of many new means of communication, HF communications are still of great importance in many areas of the world, being both cheap and usually reliable. All of these features underpin much of civilization today but they are adversely affected by solar variations that causes perturbations in the ionosphere and magnetosphere. Even the distribution of electricity by the power industry, and corrosion of long pipelines at high latitudes is affected by solar-induced changes, which are now known as space weather, the consequences for humanity being outlined below.

Spacecraft environment

Many satellites operate at geostationary orbit (∼40 000 km), a region where the effects of energetic particles on spacecraft operations are very important. During storms, the particle fluxes of energetic electrons and protons increase by many orders of magnitude. This is particularly prevalent in the outer radiation belts (see MAGNETOSPHERE OF EARTH: RADIATION BELTS). The high fluxes cause severe damage to spacecraft. The most energetic particles (energies ∼MeV) penetrate deep into the body of the spacecraft and into coaxial cables. This can lead to deep dielectric charging which results in damage to spacecraft components, and in turn leads to spurious signals and occasionally to complete failure of some or all of the spacecraft. The less energetic particles (∼keV) tend to cause surface charging and discharging which in turn can cause anomalous operation of the spacecraft.

Solar cells are also vulnerable. A modest solar flare can reduce their efficiency by ∼3% per event, ultimately affecting the lifetime of the satellite itself.

There are a variety of mitigating actions that can be taken to reduce the effects, many of which can be incorporated into the design of the satellite. Shielding of key elements and cables is the primary mechanism. Also during the operational phases avoidance of sensitive maneuvers during the storms is attempted.

The human perspective

The exposure of humans in space to the various radiations described above is a very significant hazard. US Shuttle flights have been curtailed to prevent astronauts receiving high dosages of radiation. Although the fluxes of particles at ∼15 km are tiny compared with the ionosphere, there is still a modest threat to the operation of high-flying aircraft, such as Concorde. Again electronics may be vulnerable and the total radiation dosage, especially to the crew, is an issue. The aeroplanes carry radiation monitors. If fluxes increase, the aircraft lowers its cruising altitude to reduce the damaging effects. The next generation of passenger aircraft will fly near 15 km and thus monitoring of the fluxes of energetic particles from geospace will be required.

Radio propagation

The presence of free electrons in the ionosphere and magnetosphere affect the propagation of radio waves from very low frequencies (3–30 kHz) up to the frequencies used to communicate with satellites (10^{12} Hz). Much of the knowledge of the ionosphere and magnetosphere has been determined by observing and then understanding the effects of radio waves that are refracted by the plasma. To a first order, the attenuation of radio signals is proportional to the inverse of the square of their frequency.

Propagation conditions change markedly as a result of the effects of the Sun on all time scales from seconds to a solar cycle. On a time scale of seconds, solar flares cause enhanced ionization in the D and E regions which results in additional attenuation of the radio signal. The normal diurnal variation of solar zenith angle causes the maximum electron concentration at F region altitudes to increase by up to 200% at sunspot maximum from midnight to noon. Thus the ionosphere can refract much higher frequencies at noon than midnight. As described above, the effect of storms is to reduce the electron concentration compared with quiet times owing to the composition changes, thus reducing the bandwidth for effective radio propagation.

At frequencies used in satellite communications, it is not the attenuation of the ionosphere that is important but the scattering of radio waves by small-scale (1 cm–10 km) plasma irregularities. These irregularities cause rapid fading of the signal. In practical terms, they can introduce errors into global positioning and cause distortion of images collected with synthetic aperture radars.

Spacecraft orbital decay

About 5000 spacecraft have been placed in Earth orbit, and about half of these remain there. The rest have re-entered the atmosphere. Of the remaining satellites, only 25% are still operational. In addition, many missions leave space junk, including booster rockets and explosive bolts. Satellites have blown apart and there are now approaching 10 000 pieces of spacecraft that can be tracked from the ground, i.e. items larger than ∼1 m, although the size that can be tracked depends on distance from the Earth. In addition, there are millions more pieces of space debris <1 m. There is a natural decay of spacecraft orbit due to atmospheric drag for all satellites below about 1500 km altitude. The atmospheric density increases by over an order of magnitude at F region altitudes at sunspot maximum compared with sunspot minimum. Thus much greater orbital mass falls to Earth during periods of high solar flux. This is illustrated in figure 3. In the years centered about the 1989–90 solar maximum, over 550 t of material re-entered the Earth's atmosphere, causing a noticeable fall in the satellite population.

Ground-based effects of solar variability

During substorms, very large currents (1 MA) flow horizontally in the E region of the ionosphere in the auroral oval. The magnetic variation from such

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

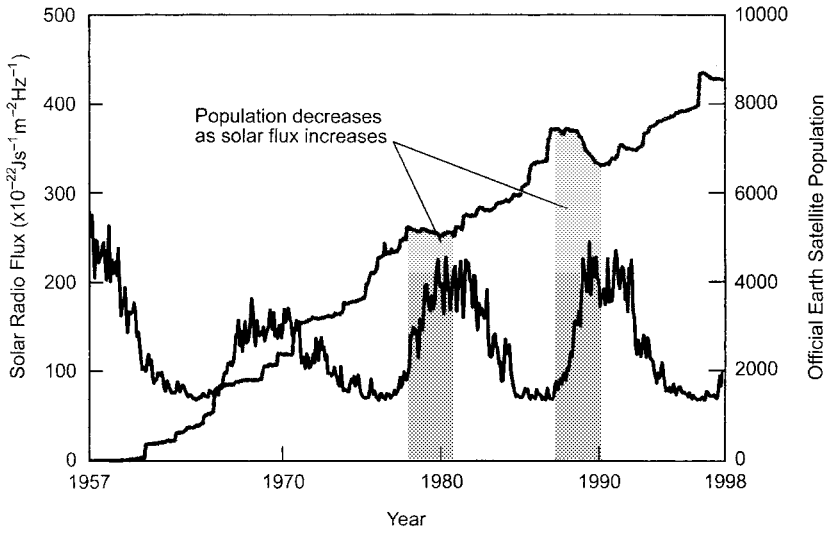


Figure 3. The 10.7 cm solar radio flux over the last 40 yr together with the satellite population. Here satellite is defined as any piece of spacecraft or debris >1 m in diameter. The shaded regions are about sunspot maximum. During these periods, the density of the atmosphere increased, resulting in enhanced satellite drag, faster decay to the orbit, and thus more debris burning up in the Earth's atmosphere.

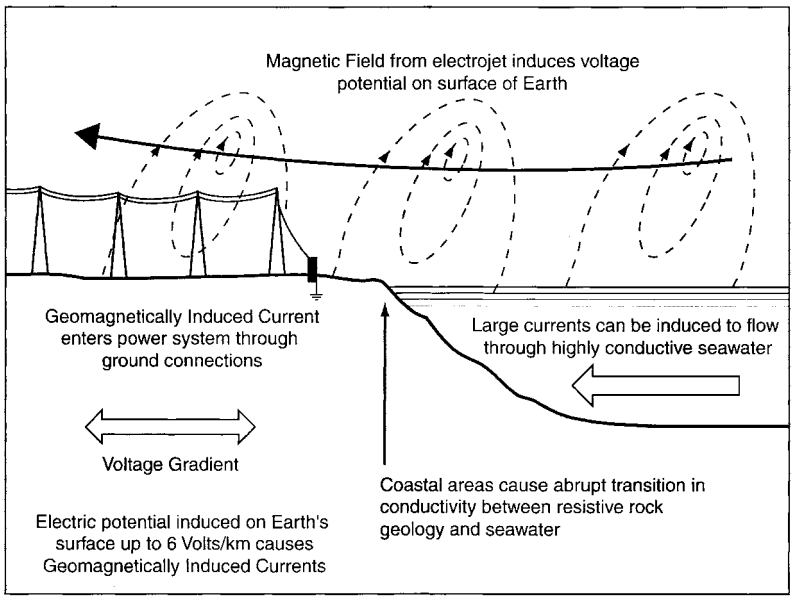


Figure 4. The electromagnetic coupling that occurs between the auroral electrojet in the ionosphere (110 km), the underlying geological features and the power systems. In regions of high resistivity (igneous rock) the Earth's potentials over large regions can cause large induced currents in powerlines. Also there may be induced currents in coastal areas when currents induced in sea water encounter the land.

currents, which are called the auroral electrojets, is $\leq$4000 nT. This represents about 10% of the Earth's field at the latitude where these currents occur. These ionospheric currents induce currents in the sea, on land and in power lines (figure 4). The main process for producing geomagnetically induced currents in electric power systems is induced Earth-surface potential due to geomagnetic field fluctuations during storms. The potential ($\sim$5 V km^{-1}) is imposed across grounded neutral points of three-phase transformers. This potential essentially acts as a voltage source and has a frequency range of a few mHz. The induced current biases the excitation characteristics of transformers causing half-cycle saturation, and hence distortion, and is thus

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

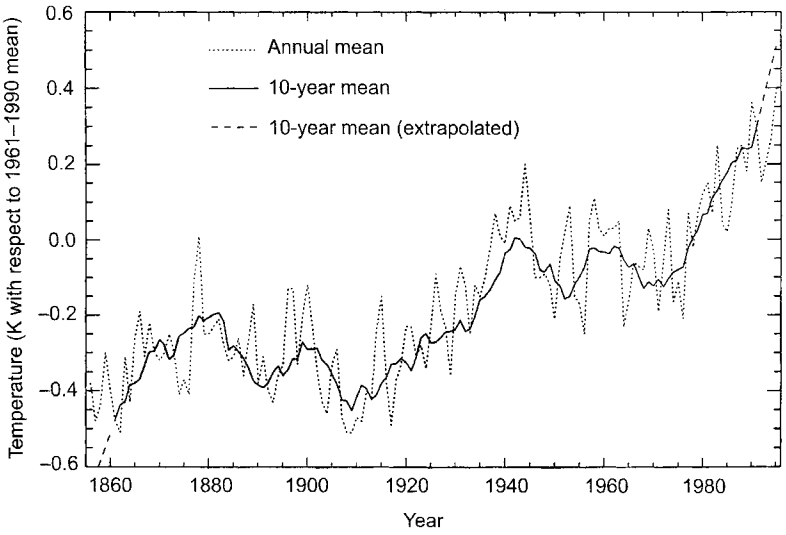


Figure 5. The mean annual global surface temperature (K) from 1856 to 1995 with respect to the mean for the period 1961–90 (dotted curve). The full curve is a 10 yr running mean. The broken curve is an extrapolation of the 10 yr mean.

rich in harmonics. The consequence is additional dissipation within the transformer itself which can lead to overheating and occasionally transformer failure. The best-documented event occurred in March 1989 when the entire province of Quebec in Canada was blacked out for more than 9 h. The economic effect of the failure was put at ∼$5 billion, i.e. similar to that of a hurricane.

Traditionally the Earth's magnetic field has been used for navigation, but with the development of GPS this is much less common nowadays. Variations in the apparent location of the magnetic pole can be several degrees during a storm, depending on the latitude of the observer. This is clearly an important effect for accurate navigation, but these large deviations last the duration of a substorm (<1 h). Ground, aircraft and, to a lesser extent, satellite measurements of the Earth's magnetic field are still important in geological survey work. This needs to be carried out during periods of quiet solar conditions, otherwise the required signals are corrupted by a combination of the ionospheric and induced currents, which are driven ultimately by solar variations.

The Sun and weather

There are >2000 papers published in the refereed scientific literature describing meteorological phenomena that have a statistically significant relationship with solar variability, and especially the 11 yr solar cycle. Over the last 30 yr during which accurate measurements have been possible, the output of the Sun has been remarkably constant (1460 W $m^{-2} \pm 0.5\%$ at the Earth's surface). The small uncertainties cannot account for the observed Sun–weather relationships directly. It is also interesting to note that the increase of global mean surface temperature (figure 5) and the occurrence of magnetic storms (figure 2) both increased markedly from about 1915. There are no generally accepted physical processes that can explain the data; however, in recent years several promising theories have been developed. On a longer time scale, the Earth's climate is known to be directly affected by variations in the Earth's orbit. These are known as the Milankovich cycles.

Predicting space weather

The requirement for accurate prediction of space weather is likely to become increasingly important with time. There is greater reliance on satellite-based communications and navigation systems, together with the prospect of much increased high-flying commercial passenger air transport and long-duration manned space flight to the International Space Station and other planets. The prediction of space weather is in its infancy, with the first comprehensive models of geospace under construction. There are two key elements where further scientific understanding of the coupled system is essential before accurate predictions can be a reality. The first is identification of the features on the Sun that lead to the severe storms. CME events that impact the Earth are certainly one important factor, but not the only one. Second, severe storms occur when the IMF is southward. There is at present no way of predicting the IMF remotely. The best input parameters for the present forecasting models are solar wind and IMF conditions from satellite measurements at the L1 libration point. (The L1 point is where the gravitational forces of the Sun and the Earth are equal, about 1.5 million km upstream in the solar wind.) These data give about a 1 h warning of the impacts on Earth, which is insufficient to take all the actions required to minimize the possible damaging consequences of magnetic storms which are caused by variations in solar processes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Clilverd M A, Clark T D G, Clarke E and Rishbeth H 1998 Increased magnetic storm activity from 1868–1995 *J. Atmos. Sol.-Terr. Phys.* **60** 1047–56

Jarvis M J, Jenkins B and Rodgers G A 1998 Southern hemisphere observations of a long-term decrease in F-region altitude and thermospheric wind providing possible evidence of global thermosphere cooling *J. Geophys. Res.* **103** 20 774–87

A S Rodger

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Terrestrial Connection: Coupling Between Solar Wind, Magnetosphere, Ionosphere, and Neutral Atmosphere

The particle flux from the Sun and the magnetosphere represents a large source of energy and ionization for the lower thermosphere and ionosphere. The energy flux, which ranges from $<10^{-3}$ J m^{-2} s^{-1} to over 0.1 J m^{-2} s^{-1}, is deposited by particles with energies that range from hundreds of eV to several hundred MeV. The intensity, spectrum, and localization of the precipitation are functions of solar and geomagnetic activity (MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS). The mean values also show longer-term dependence on the SOLAR CYCLE. The particle-induced ionization leads to dissociation of tightly bound N_2 molecules and to the formation of the reactive nitrogen compounds NO and NO_2. These compounds, when transported to lower altitudes, participate in a catalytic cycle leading to the destruction of ozone. Continuing research seeks to understand the impact of magnetospheric and solar energetic particles on the chemistry and electrodynamics of the middle atmosphere.

Atmospheric processes

During the last 25 years intensive efforts have been underway to develop an understanding of variations in middle-atmospheric ozone (O_3) due to natural processes and due to the effects of mankind (see OZONE HOLE). It is now reasonably well understood that the global balance of O_3 is governed in part by the balance between the production of O_3 and its destruction by reactions within the NO_y, Cl_y, HO_y, and O_y chemical families. One of the most important catalytic cycles in this global balance is that due to the following reactions:

$$\begin{aligned} NO + O_3 &\rightarrow NO_2 + O_2 \\ NO_2 + O &\rightarrow NO + O_2 \\ \hline O_3 + O &\rightarrow 2O_2 \end{aligned}$$

The bulk of the odd nitrogen (i.e. NO_y, oxides of nitrogen) in the stratosphere is formed from the oxidation of N_2O by $O(^1D)$ (excited atomic oxygen) forming NO. It is known, however, that middle-atmospheric NO_y is formed by ion chemistry initiated by the precipitation of energetic particles into the upper atmosphere (see SOLAR WIND: ENERGETIC PARTICLES).

The first step in determining the production of NO_y due to the precipitation of energetic particles is the calculation of the deposition of energy from particle penetration. The particle energy is absorbed by the atmosphere through collisions with, and scattering by, atmospheric constituents which result in either dissociation or ionization of the atmospheric constituents. Production of each electron pair within the atmosphere requires an energy deposition of approximately 35 eV. In addition to the energy imparted directly to the atmosphere during collisions, a certain amount of energy is converted to x-rays by the bremsstrahlung process. This is associated with the rapid deceleration of energetic electrons during their penetration into the atmosphere, providing additional ionization down to altitudes of 20 km.

The ion and neutral odd-nitrogen chemistry initiated by energetic particles produces secondary electrons, e^*, which in turn ionize and dissociate the major atmospheric species. This ionization is followed by a series of interchange and recombination reactions involving nitrogen and its ions which produce additional atomic nitrogen. The resulting atomic nitrogen may either be in the ground ^4S level or excited to the ^2D level. Nitric oxide is formed by the reaction of atomic nitrogen with O_2. Reactions involving excited nitrogen are faster than those with the ground-state N atom. The destruction of odd nitrogen proceeds through the reaction of N atoms with NO to produce molecular nitrogen and atomic oxygen. In the sunlit atmosphere the photodissociation of NO is important; however, in the fall, winter, and spring in the polar region, the photolytic reaction is negligible and the resulting NO lifetime is sufficient for transport to bring the NO into the mesosphere and stratosphere.

Current numerical models take into account the details of the photochemical complexities and are capable of calculating the odd-nitrogen production rates as functions of the energy-dependent ionization and deposition rates. Thus, if energetic particle precipitation can be measured, the auroral-induced emissions and the NO can be specified. As the NO produced in the thermosphere and mesosphere is transported to lower altitudes it reacts with ozone to form NO_2 and remains as this form of odd-nitrogen for the remaining time of transport.

There have been several studies of these effects suggesting that precipitating electrons could lead to the formation of oxides of nitrogen and hydrogen. These could affect O_3 near 60 km and the ion formation rates due to these electrons could dominate those of other processes down to approximately 40 km. Baker *et al* (1987) found ion formation rates using electron measurements from geostationary orbit that showed that the fluxes of energetic electrons and hence the ion formation rates were strongly modulated by solar activity. Callis *et al* (1991), using estimates of the precipitating electron fluxes derived from geostationary data and 2D simulations, concluded that the precipitating electrons could have a significant effect on the budgets of both NO_y and O_3 within the stratosphere. This coupling mechanism between solar activity, solar wind structures, the energetic particle populations within the magnetosphere, and the chemical state of the middle atmosphere represents an important linkage which must be understood to assess the natural variations in the global chemical state of the middle atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Energetic particle sources

The energetic particle flux into the atmosphere has characteristic spectral features and temporal and spatial scales depending on latitude and local time. The most important sources include the solar energetic particles within the polar cap, the diffuse and discrete auroral precipitation within the auroral zone, the energetic magnetospheric electrons extending from midlatitudes throughout the auroral zone, and the ring current particles precipitating at midlatitudes. Table 1 summarizes the known particle input characteristics; in the following sections we discuss each type of particle input separately.

Solar energetic particle influx

Solar energetic particle (SEP) events extending in energy up to hundreds of MeV can produce dramatic enhancements in the ionization and NO_y production rates in the polar regions. Figure 1 shows instantaneous production rates of NO calculated for several solar particle events. These particles enter via the open magnetic field line regions. Also shown are the average production rates due to galactic COSMIC RAYS (GCRs) at solar maximum, and estimated production rates due to precipitating electrons (Brasseur and Solomon 1984). Note that solar protons can have a large effect between the altitudes of ~20 and 100 km.

Of the solar proton events shown in figure 1, that of 4 August 1972 was one of the largest on record, the kind that occurs roughly once a decade. However, SEP events of the size of the other two typically occur several times per year during solar maximum. Reid *et al* (1991) have calculated the ionization rates and resulting effects on NO_y and O_3 for the series of large solar proton events in late 1989 that were comparable in intensity to those of August 1972. They predicted NO increases as large as a factor of 20 near 60 km altitude, with corresponding peak O_3 depletion as large as 20% near 40 km in late October 1989. Furthermore, calculation of the latitudinal range over which the deposition occurs requires information of the geomagnetic cutoff. Because the cutoff can be lowered by several degrees at the time of these events due to associated GEOMAGNETIC STORM activity, the cutoff needs to be monitored at the same time. The scale on the right of figure 1 indicates the minimum energy proton that can penetrate to a given altitude, which shows that calculation of the ionization rate due to solar protons from 20 to 80 km requires information on the proton energy spectrum from about 3 to 500 MeV.

Relativistic electron precipitation

Baker *et al* (1986, 1987) reported observations of the fluxes of energetic (1–15 MeV) electrons at geostationary orbit (6.6 R_E) for the period 1979 to 1986, from solar maximum conditions to solar minimum. Their results indicated that large enhancements in the electron population were sporadic in the 1979–1981 solar maximum period. However, during the declining phase of the solar cycle, from late 1981 until mid-1986, frequent large increases in this population were observed. The average intensity and

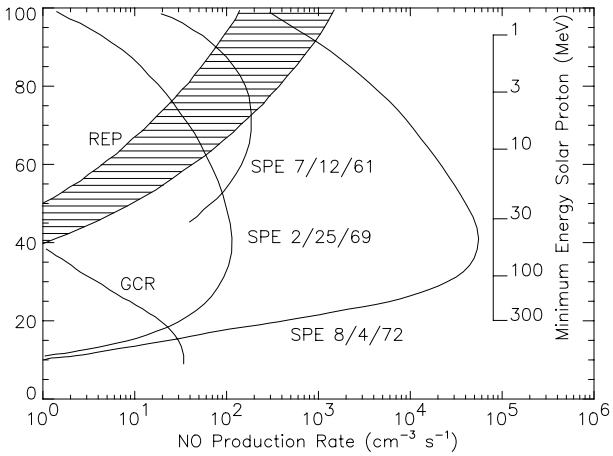


Figure 1. NO production from solar proton events (SPE), relativistic electron precipitation (REP), and galactic cosmic rays (GCR).

frequency of these events increased dramatically so that the 6-month average of the flux of these electrons was larger by a factor of 8 or 9 than a similar average observed during 1979–1980.

The population of relativistic electrons is known to be closely related to the occurrence of high-speed ($V_{SW} \geq 600$ km s^{-1}) solar wind streams and often shows a periodicity of 27 d. Since high-speed solar wind can be associated both with SOLAR CORONAL MASS EJECTIONS and with the cyclical development of solar CORONAL HOLES, their occurrence is temporally associated with the 11 y solar cycle. The maximum frequency of energetic electrons develops during the declining SUN-SPOT phase, but even near solar maximum conditions, very strong relativistic electron events appear in association with coronal mass ejections and the related magnetic clouds interacting with the Earth's magnetosphere. A dramatic example is the event of 24 March 1991, when a strong interplanetary shock struck the Earth's magnetosphere and generated several orders of magnitude increases in the fluxes of multi-MeV electrons and protons in the inner magnetosphere (see MAGNETOSPHERE OF EARTH). These results provide support for and strengthen the suggestion that relativistic electron precipitation may significantly influence the photochemical state of the middle atmosphere.

Auroral particle influx

The particle flux precipitating from the magnetospheric plasma sheet represents a large source of energy and ionization for the lower thermosphere and ionosphere (see MAGNETOSPHERE OF EARTH: PLASMA SHEET). The energy flux, which ranges from less than 10^{-3} J m^{-2} s^{-1} to over 0.1 J m^{-2} s^{-1}, is mainly deposited by particles with energies that vary from hundreds of eV to several hundred keV. The intensity, energy, and localization of the precipitation are functions of geomagnetic activity: higher localized fluxes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Particle input characteristics.

	Latitude range (deg)	Longitude range (MLT)	Time scale (d)	Flux range (J m^{-2} s^{-1})	Energy range (keV)	Total energy input (10^{13} J)	Tangent column NO production (cm^{-2})
SEP	>60–>70	24	1–3	10^{-5}–10^{-4}	10^3–5×10^5	3–80	6.2×10^{14}–1.2×10^{16}
REP	55–72	24	2–5	10^{-6}–10^{-4}	10^3–1.5×10^4	1–300	1.2×10^{14}–3.1×10^{16}
Diffuse aurora	65–72	24	—	5×10^{-4}–0.5	0.5–50	3.6–3600 per h	1.3×10^{15}–1.2×10^{17} per h
Discrete aurora	1–10	2–4	0.02–0.125	2.5×10^{-4}–0.1	0.5–30	0.005–2600	3.1×10^{14}–7.8×10^{17}
Ring current	55–65	24	1–3	10^{-5}–10^{-3}	10–500	4–1000	6.2×10^{14}–1.2×10^{17}

and higher mean energies are found during substorms, and the mean values also show longer-term dependence on the solar cycle.

The discrete AURORAS are highly structured and localized, and their occurrence frequency maximizes in the premidnight sector at auroral latitudes. Individual auroral arc widths vary from a few km to 100 km in latitude, but during substorms the discrete auroras can cover a region of about 1000 km in latitude and several thousand km in longitude. The discrete precipitation consists mainly of electrons in the energy range from a few hundred eV to a few tens of keV, often accelerated by field-aligned potential structures which produce a characteristic monoenergetic peak in their spectrum. Electrons in the 0.5–30 keV energy range make a significant contribution to the energy input to the lower ionosphere, and are important for maintaining ionization levels in the ∼100 km altitude region during the night when photoionization stops and particle impact ionization is dominant. Their energy and ionization inputs are thus important for the study of external influences on the mesosphere.

The diffuse auroras are less structured and persist in the auroral latitudes almost continuously. This more uniform precipitation consists mostly of electrons and ions with energies from a few hundred eV to many tens of keV. In particular, during the substorm recovery phase, energetic (tens of keV) electron precipitation is often found in the morning sector of the auroral oval affecting the chemistry at about 80 km altitude.

Ring current ion precipitation

At mid-latitudes the precipitation from the ring current just inside the plasmapause (see MAGNETOSPHERE OF EARTH: RING CURRENT) consists mainly of ions in the 10–200 keV energy range. The ions are anisotropic, with a peak near 90° pitch angle, and have been observed down to an altitude of about 120 km. The energy flux increases with magnetic activity from 10^{-7} to 10^{-5} J m^{-2} s^{-1}. For disturbed conditions, the precipitating ions and neutrals are the primary source of nighttime ionization at these altitudes and latitudes. Ring current ions interact with the outer edge of the plasmasphere, where they excite ion cyclotron waves (see MAGNETOSPHERE OF EARTH: PLASMASPHERE). These waves lead to scattering of the ring current ions, which most likely produces the observed precipitation. Outside the plasmapause there is another region of ion precipitation, where the observed precipitating flux is more isotropic. The filled loss cone suggests that a mechanism for strong pitch angle diffusion exists in the outer magnetosphere with energy deposition near 10^{-3} J m^{-2} s^{-1}.

Precipitation of energetic ions and neutrals ionizes and heats the thermosphere. Good agreement is found between the ionization rate based on the energetic ion and neutral precipitation with measurements of the electron density profile. The ionization profiles show a nearly constant rate of ionization from 120 to 200 km, and the ionization rates increase with magnetic activity from 10^{-1} to 10^2 cm^{-3} s^{-1}.

Present evidence of energetic particle coupling to atmospheric chemistry

The extensive thermospheric NO data set acquired by the Solar Mesosphere Explorer (SME) satellite has quite clearly shown that NO is produced above 100 km by impact of auroral particles with energies less than 10 keV. On the other hand, the observational evidence of the significance of NO_y production in the middle atmosphere and its subsequent downward transport has been limited: regular mesospheric NO measurements have become available only after the 1991 launch of the Halogen Occultation Experiment (HALOE) onboard the UARS satellite. These data have provided evidence that large amounts of NO are being produced in the mesosphere. Further evidence suggests transport to the stratosphere.

Callis *et al* (1996) looked at specific events and compared the observed NO changes with concurrent electron measurements by the SAMPEX satellite. They found significant NO increases (up to a factor of ten from 60 to 120 km) during longlasting electron events. A good correlation was found between the $E > 30$ keV electron measurement and the NO response measured by HALOE. Backed by 2D model calculations of ozone decreases due to enhanced NO_y, Callis *et al* (1996) made the provocative suggestion that NO_y enhancements such as discussed above were responsible for the ozone depletions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

D N Baker

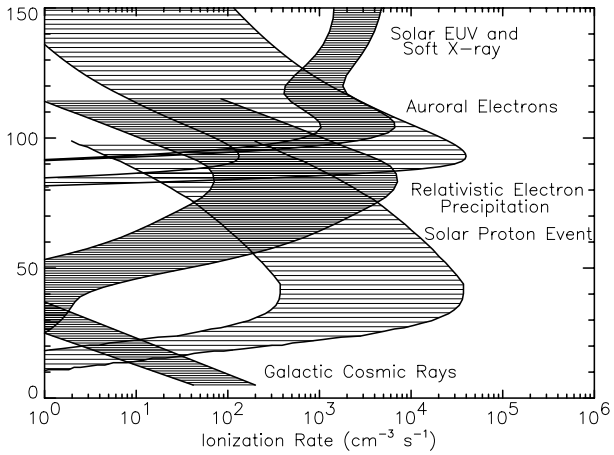


Figure 2. Comparison of energy inputs and subsequent ionization rates in the atmosphere due to energetic particle sources, solar EUV, x-rays, and galactic cosmic rays.

in the lower stratosphere. Thus, the current observational database suggests that the influence of energetic particle precipitation on atmospheric chemistry extends to as low as 25 km. Dedicated measurements of both NO_y and O_3 throughout the polar winter are clearly required to prove this contention.

In summary, as shown in figure 2, there are a wide variety of coupling processes between the Sun, solar wind, and magnetosphere on the one hand and the various layers of the atmosphere on the other. This coupling affects both the chemical and the electrodynamic properties of the middle and upper atmosphere. Thus, it is quite possible that solar wind and magnetospheric coupling to the lower layers of the atmosphere can have substantial effects on NO_y and O_3 concentrations.

Bibliography

Baker D N, Blake J B, Klebesadel R W and Higbie P R 1986 Highly relativistic electrons in the Earth's outer magnetosphere, 1, Lifetimes and temporal history 1979–1984 *J. Geophys. Res.* **91** 4265

Baker D N *et al* 1987 Highly relativistic electrons: A role in coupling to the middle atmosphere? *Geophys. Res. Lett.* **14** 1027

Brasseur G and Solomon S 1984 *Aeronomy of the Middle Atmosphere* Atmospheric Sciences Library (Hingham, MA: D Reidel)

Callis L B *et al* 1991 Precipitating relativistic electrons: their long term effect on stratospheric odd nitrogen levels *Geophys. Res. Lett.* **96** 2939

Callis L B *et al* 1996 Precipitating electrons: Evidence for effects on mesospheric odd nitrogen *Geophys. Res. Lett.* **23** 1901

Reid G C, Solomon S and Garcia R R 1991 Response of the middle atmosphere to the solar proton events of August-December 1989 *Geophys. Res. Lett.* **18** 1019

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Terrestrial Connection: Long-Term and Short-Term Climate Variability

The Earth's CLIMATE expresses the combined response of the atmosphere, the OCEANS and the continents to the energy that is received from the Sun. Any variation in the energy received from the Sun or radiated away from Earth and any change in the distribution of energy over the Earth's surface will therefore have an effect on climate (see also EARTH'S ATMOSPHERE).

By climate we usually understand the average state of the atmosphere over several years. Superposed on that is the large seasonal variation caused by the annual change of the solar zenith angle and the daily variation due to changes in the energy distribution over the Earth caused by the rotation of the Earth around its axis.

The observed climate is a mixture of many different atmospheric processes including a number of complex feedback mechanisms. Some of these processes are directly affected by human activity such as burning of fossil fuel. However, there is also a large natural variability, which it is necessary to understand in order to assess the possible contribution of humanity to climate.

Causes of climate changes

Causes of climate change can be divided into two main categories: those that are due to external forcing of the climate system and those that are due to an internal redistribution of energy and matter within the global climate system.

The internal redistribution has a significant effect on short time scales ranging from years to decades. The most prominent example of the interannual variability in the climate is the El-Niño-Southern oscillation (ENSO). It is due to a complicated interaction between various air masses in the tropics and anomalous warming of the eastern and central tropical Pacific Ocean. The heat capacity of the oceans and the nature of the ocean currents play a major role in controlling the state of the atmosphere and hence the climate. Ocean currents are in general not yet known with sufficient accuracy to enable us to predict their effect on climate. In addition, their effect is probably associated with chaotic behavior that may make it impossible to predict such changes. Indications of the probable range of such variations in a statistical sense may, however, be acquired from numerical climate models.

Changes in the energy budget of the Earth will result in persistent climate changes. The Sun's radiation is the source of the energy that drives the weather and climate. The Sun emits radiation at various frequencies. As a first approximation the Sun radiates energy corresponding to a black body with a temperature of 5770 K, the solar surface temperature. Integrated over all wavelengths the total irradiance from the Sun at the position of the Earth is approximately 1368 W m^{-2}, the solar constant. The cross section exposed to the solar energy corresponds to an area of πR^2 whereas the total surface area of the Earth including the night side is $4\pi R^2$. Distributed evenly over the Earth's surface we therefore need to divide the number by 4 to yield an average power of 342 W m^{-2} at the top of the atmosphere. In a steady state with no climate change the Earth would reradiate the same amount of energy as a black body at the global average temperature, approximately 288 K. This corresponds to a longer wavelength with a peak in the infrared part of the spectrum. The short-wave radiation from the Sun reaching the Earth's surface and the long-wave radiation from Earth to space have to pass through the Earth's atmosphere, which acts like a filter. The effect of the filter depends on the wavelength of the radiation and it has therefore a different effect on the incoming and outgoing radiation. The net effect of these processes, also called the Earth's radiation budget, is therefore highly dependent on the composition of the atmosphere. At present, the composition of the atmosphere is such that it keeps the global temperature about 31 K higher than the Earth would be without an atmosphere. The dominant gases contributing to this warming effect, the natural GREENHOUSE EFFECT, are water vapor and carbon dioxide (CO_2).

The composition of the atmosphere may change for a number of reasons. Volcanic activity is often accompanied by emission of considerable amounts of sulfur in the stratosphere. These emissions create aerosols that reflect sunlight and the effect is therefore a cooling of the Earth. Since the aerosols remain in the atmosphere for several years the effect on the Earth's average surface temperature may be quite significant. Energy balance calculations indicate that volcanic activity may have contributed to the observed global average temperature variations during the past 150 yr. In the stratosphere another gas, ozone (O_3), is varying because of the solar cycle variation in the ultraviolet radiation that produces O_3. However, the ozone layer is also depleted owing to the accumulating amount of anthropogenic chlorofluorocarbons (CFCs). Both CFCs and O_3 are very effective greenhouse gases but their combined effect on climate is significantly reduced because of the destructive effect of CFCs on O_3 (see OZONE HOLE).

Most concern is, however, associated with the effect of the atmospheric concentrations of greenhouse gases such as CO_2 and methane (CH_4). We know that these concentrations have increased considerably over the past 150 yr, after the start of industrialization and the rapid increase of the population of our planet. The gases absorb the long-wave radiation from the Earth's surface and are supposed to be responsible for at least part of the global warming that has characterized this century. The effect of human activity is rather complex, because the burning of fossil fuel that causes the increased concentrations of CO_2 has also caused an increase in the contents of sulfate aerosols in the troposphere. The effect of this may be a cooling similar to the effect of volcanic aerosols. Whereas the production of CO_2 is proportional to the amount of fossil fuel used, the emission of sulfate aerosols may

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

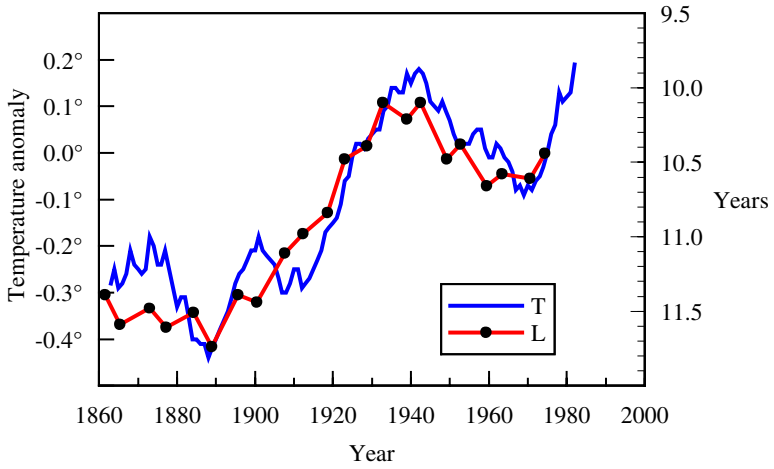


Figure 1. Smoothed length of the solar cycle (L) and the 11 yr running average of the northern hemisphere land temperature (T). (Figure taken from Friis-Christensen E and Lassen K 1991 Length of the solar cycle: an indicator of solar activity closely associated with climate *Science* **254** 698–700.)

be affected considerably by various measures to clean the exhaust, for instance the use of filters. The relative importance of these two atmospheric constituents that are products of fossil fuel burning has therefore changed in time and so has their expected impact on climate.

Long-term changes of climate

On very long time scales, billions of years, we know that the solar luminosity has been increasing monotonically with time. At the time of the formation of the Earth the Sun's output was only 70% of its present output. According to paleoclimatic records this huge change in luminosity has not, however, been associated with an expected corresponding increase in the average temperature of the Earth. This 'faint young Sun' paradox may be explained in several ways. Some of the most plausible ones are related to changes in the atmosphere and its composition. In particular, the large amounts of carbon trapped in carbonate sediments indicate that the abundance of CO_2 in the atmosphere must have been much larger during times in the past. An interesting theory by James Lovelock, called the Gaia hypothesis, even suggests that on very long time scales life itself is capable of modifying the Earth's environment, including the atmosphere, in order to optimize conditions for its own continued existence.

The history of Earth shows that climate has been varying on many time scales. Most prominent during the last two million years has been the repeated occurrence of glaciation with a period of approximately 100 000 yr. Reconstruction of past climate variations shows that, on top of this dominant quasi-cycle, shorter quasi-cycles of roughly 41 000 and 21 000 yr are superimposed. The apparent regularity of this phenomenon and the calculated periods has called for an astronomical explanation that was first quantitatively formulated by MILANKOVITCH in the 1920s. According to the Milankovitch theory the climatic cycles are related to the variations in the received solar energy and in particular its distribution over the Earth's surface. These variations can be predicted by changes in the rotational and orbital parameters of the Earth. The major period of about 100 000 yr is related to the variation of the eccentricity of the Earth's orbit around the Sun. The other parameters are related to the variation of the tilt of the Earth's axis between about 22° and 25° (~41 000 yr) and the precession of the equinoxes and solstices around the Earth's orbit (~21 000 yr). Although the changes may seem rather small, they have apparently been sufficient to initiate large oscillations in the Earth's climate in a manner still poorly understood. Model calculations show that the response to the orbital forcing is non-linear and that it involves internal processes and feedback mechanisms. These feedback mechanisms may involve changes in the atmospheric composition, in the amount of snow and ice, and in the biological cycle. It is still an open question, however, whether the external orbital forcing drives the internal processes or whether they act as a pacemaker for the free oscillations in the internally driven system. However, the apparent large climate sensitivity to small changes in the amount and latitudinal distribution of solar radiation has contributed further to the concern of possible effects of a humanly induced enhanced greenhouse effect.

Climate variations on shorter time scales

In addition to these crucial variations in the Earth's climate our records of past climate do show that climate has been varying also on time scales that are too short to be possible manifestations of orbital parameters. Some of these oscillations may be the result of internal oscillations but during the last century there has been an increasing number of scientific reports that indicate that some of them may be attributed to variations in the Sun.

Usually we regard the Sun as just an eternal source of energy that sustains life on Earth. Until recently the energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

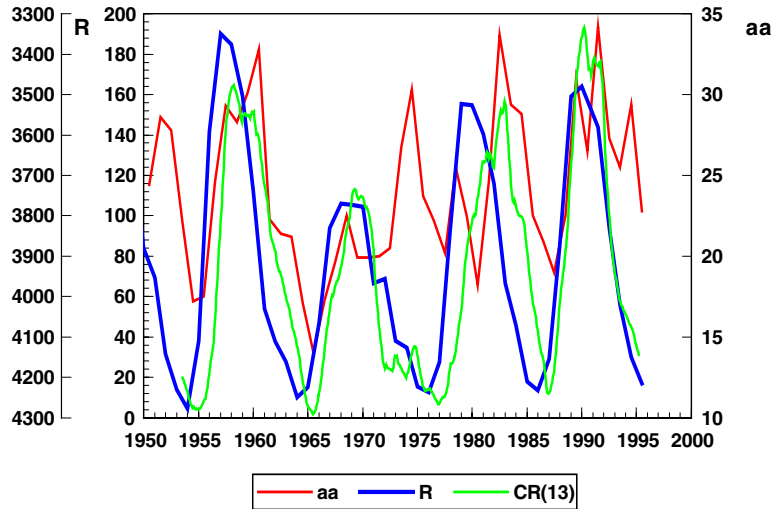


Figure 3. Variation of the cosmic ray flux (CR) observed at Climax station, USA, the yearly sunspot number (R) and the geomagnetic activity index *aa*. Note the reversed scale of the cosmic ray flux.

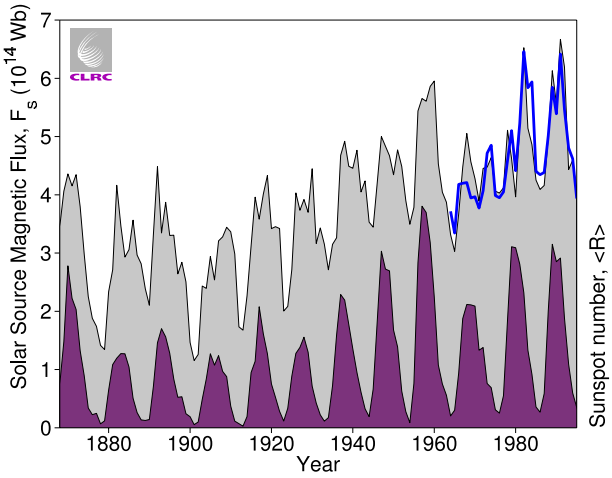


Figure 2. This figure shows estimates back to 1868 of the total solar magnetic flux emanating through the coronal source sphere. (Figure 3 in Lockwood M, Stamper R and Wild M N 1999 A doubling of the Sun's coronal magnetic field during the past 100 years *Nature* **399** 437.)

received from the Sun was believed to be constant, as the term 'SOLAR CONSTANT' for this value indicates. However, in astronomical terms the Sun is a variable star. This has been known for many centuries during which scientists have performed systematic observations of the Sun's surface. The observations demonstrate periodic variations on different time scales. Most pronounced among these is the nearly 11 yr cycle in the number of SUNSPOTS (see also SOLAR CYCLE). Direct indices for solar activity are not really available before the start of the use of telescopes around 1610. However, recently a promising tool has been developed which may enable us to derive indices of solar activity on much longer time scales. This method is based on indirect proxies, for example the cosmogenic radionuclides ^{14}C and ^{10}Be. These nuclides are caused by cosmic ray particles and, because they are electrically charged, both the geomagnetic field and the magnetic field in the heliosphere modulate their production rate. The magnetic field in the heliosphere has its origin in the solar surface magnetic activity. Therefore the inferred production rate of radionuclides in the atmosphere may be used to provide an index of solar magnetic activity back in time (see SOLAR ACTIVITY: LONG-TERM RECORDS).

Such estimates of solar activity variations in the past have made it possible to examine a possible relationship with past climate. A number of studies showing a striking correlation have been reported. During the last millennium the major historically recorded deviations from the present climate are the Medieval warm period during the 11th and 12th centuries, when the Vikings settled in Greenland, and the cool period in Greenland and Europe in the 17th century known as 'the Little Ice Age'. The warm period is associated with a period of high solar activity as deduced from the estimated very low production of radionuclides. On the contrary, the Little Ice Age is associated with a prominent minimum in solar activity, known as the Maunder minimum, during which the production of radionuclides was very high.

The occurrence of sunspots indicates only one manifestation of solar activity. However, it is by no way obvious that the sunspot number is the most relevant parameter for a possible direct effect on climate. Another index of solar activity is the duration of the solar cycle. On the average a solar cycle is about 11 yr but high-activity cycles are generally shorter whereas solar cycles with low peak activity are longer. During this century solar cycles were long at the beginning of the century whereas they were short at the end. In figure 1 the change in solar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

cycle length during the last 130 yr is compared with the 11 yr running average of the northern hemisphere land temperature. The striking similarity between the two curves indicates that the Sun may have played a major role in global temperature change during this century.

Possible physical mechanisms for solar activity induced effects on climate

In order to claim a direct cause-and-effect relationship it is necessary to point at a plausible physical mechanism. An obvious candidate is the possibility that the total solar irradiance (the solar 'constant') varies in accord with solar activity. Accurate observations of the total irradiance are not possible from the ground because variations in the atmospheric composition by far dominate any changes that might be attributed to variations in the solar 'constant'. It was therefore not until direct measurements from satellites became available in the 1980s that the hypothesis of a solar cycle variation in total irradiance could be tested. These measurements did in fact demonstrate that changes exist, both on short time scales (days to weeks) and on longer time scales (decades). The competing effects of sunspot darkening and faculae brightening could account for the short-term variations. The variations on the decadal time scale did unambiguously show a variation in phase with the sunspot cycle but these variations only amounted to 0.1% between sunspot maximum and sunspot minimum.

The effect on the global temperature of such a change is, however, small compared with other radiative climate forcing parameters associated with changes in the composition of the atmosphere. Even if changes in total irradiance may amount to ∼0.25% over longer periods as has been suggested to be the case for the Maunder minimum, they may only be of minor importance for climate. Whereas the global warming during this century could be mainly due to the enhanced greenhouse effect this cannot be the case for the significant past climate variations. Since many of the past climate changes have been demonstrated to be correlated with changes in solar activity variations, people have looked for indirect mechanisms by which solar activity variations could modulate climate.

One possibility concerns the ultraviolet part of the solar irradiance. Solar cycle variations in this spectral band are much larger than for the visible light. Attempts have therefore been made to model the climatic effect of solar cycle variations taking into account also the modulating effect of the varying ultraviolet radiation on the lower stratospheric ozone. The models show an increased tropospheric Hadley circulation during high solar activity consistent with observations, although the observations also indicate a larger effect than the models predict. The effect may contribute to the observed very clear solar cycle signature in the height of constant pressure levels in the stratosphere, seen consistently during the last four solar cycles in particular between 20°N and 45°N in the Pacific–Atlantic area.

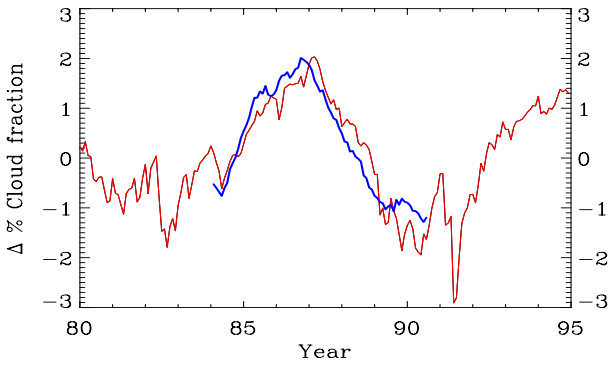


Figure 4. Variation of total cloud cover and cosmic ray flux observed at Climax. (Figure taken from Svensmark H and Friis-Christensen E 1997 Variation of cosmic ray flux and global cloud coverage—a missing link in solar–climate relationships *J. Atmos. Sol.-Terr. Phys.* **59** 1225–32.)

Even more pronounced manifestations of solar cycle effects in the solar energy output are associated with the solar wind, the emission of particles and fields from the solar surface. Although the energy in the solar wind is negligible compared with the energy in the ultraviolet and visible spectral bands, the relative variations are much larger. A direct effect of the varying solar wind is seen in the fluctuations in the geomagnetic field due to electric currents in the magnetosphere and the ionosphere. The *aa* index has been derived from such fluctuations recorded at two geomagnetic observatories in England and Australia since 1868. From the time series of this index it has been inferred the total magnetic field leaving the Sun may have increased by a factor of more than 2 during this century (see figure 2). Such an increase is most likely the cause of the decreasing cosmic ray flux during this century, which has been observed from the ground since the 1930s and which has also been inferred from the ^{10}Be data from ice cores in Greenland. Figure 3 shows the cosmic ray flux observed at the Climax Neutron Monitor station in Colorado. Note the reversed scale indicating a clear solar cycle modulation nearly in antiphase with the variation of the sunspot number, R. The figure also shows the geomagnetic activity index, *aa*, which has a solar cycle variation quite different from the sunspot cycle.

One of the major uncertainties in climate models is the role of clouds (see also CLOUDS IN PLANETARY ATMOSPHERES). In particular there are large difficulties associated with the parameterization of these effects in general circulation models used in climate studies. Recent studies indicate that cloud formation may be influenced by Galactic cosmic rays through ionization changes that cause microphysical changes in the atmosphere. Hereby nucleation and growth of ice particles may be affected.

A change in cloud cover would indeed be a very effective amplifying mechanism for climate forcing because the energy necessary to condense water vapor is small compared with the resulting changes in energy of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solar radiation received at the Earth's surface. Figure 4 shows the result of an examination of the compiled International Satellite Cloud Climatology Project (ISCCP) data. The 12 months running mean of the total cloud cover (thick curve) is shown together with the 12 months running mean values of cosmic ray intensity measured at the Climax Neutron Monitor station, Colorado. The correlation between the cosmic ray flux and the global cloud cover is very high and indicates a direct relationship although a detailed mechanism is still lacking.

Summary

The Earth's climate is a very complex system with a number of elements and with many feedback mechanisms. Various processes including variations in solar activity can cause climate changes. The detailed mechanisms and the relative importance of the different processes are still not very well known but increased understanding of our environment including the processes on the solar surface and in the heliosphere will considerably improve our ability to estimate the effect on climate of human activity.

E Friis-Christensen

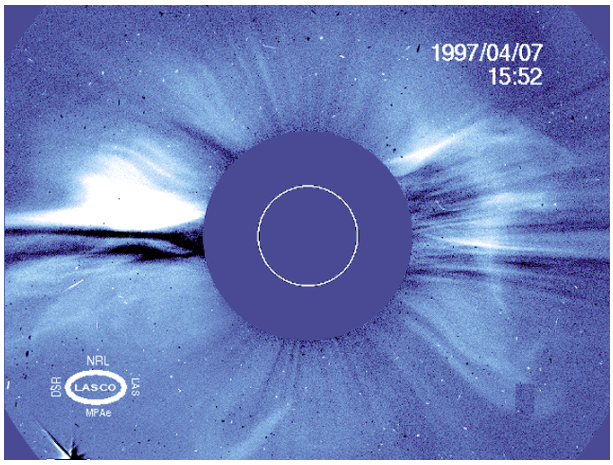
Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar–Terrestrial Connection: Space Weather Predictions

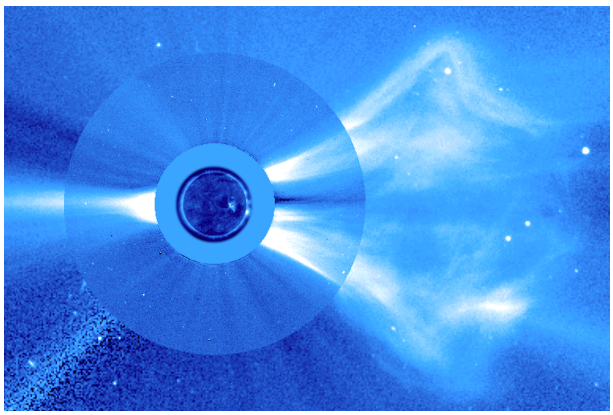
'Space weather' refers to the dynamic state of the magnetosphere and ionosphere, which is determined by the SOLAR WIND. Interest in the ability to predict space weather, and its effects on Earth, really began during World War II. For the first time, electronic technologies, such as radars and radio communications, were heavily relied on, and it soon became clear that these technologies could be seriously disrupted by solar activity. After the war, more uses were developed for electronic technologies, including the start of the space program. Since then, our need to be able to forecast solar activity has continued to increase as technology becomes more widespread. Under disturbed conditions, satellite- and ground-based technological systems, e.g. communications networks, electric power grids and satellites, can suffer harmful effects. Such systems are particularly vulnerable during severe GEOMAGNETIC STORMS. Large storms are relatively infrequent but, when they occur, they can stress the susceptible systems for prolonged periods of time over large geographic areas. Secure operation of systems can still be maintained and hazards can be minimized if the occurrence, duration and severity of impending storms can be accurately predicted in a timely manner. Thus, space weather forecasting is important for protecting national assets in both the commercial and military sectors.

In the early days of solar forecasting, it was assumed that when a large FLARE occurred on the Sun there would be a very predictable geomagnetic disturbance on Earth within a few hours or days. Later it was realized that it was the SOLAR CORONAL MASS EJECTIONS (CMEs) that had a greater effect on Earth. It was therefore believed that improved forecasting was just a matter of making better observations of the Sun so that flares and CMEs could be detected immediately after they occurred. Experience, however, soon showed that the effects on Earth did not correlate so simply with events on the Sun and, moreover, not all mass ejections had a noticeable effect on Earth. Sometimes geomagnetic storms occurred when there was no apparent eruptive activity on the Sun. We now know that how the Sun's magnetic field connects with the geomagnetic field makes a big difference in how solar activity affects Earth. When a mass of PLASMA is ejected from the Sun, the plasma travels outward in the solar wind. These plasma bursts have their own magnetic fields, which are carried along with the plasma. How these fields are oriented when they arrive at Earth determines whether or not the event will be effective. When the direction of the solar wind field is opposite the direction of Earth's field, MAGNETIC RECONNECTION occurs, and the magnetosphere essentially becomes joined to the solar magnetic field. In this condition, Earth is much more prone to the effects of the solar wind. Solar wind particles can enter the magnetosphere more easily, and those already within the magnetosphere are energized. If the magnetic field of the solar wind is in the same direction as the Earth's field, then magnetic reconnection does not occur and the magnetosphere is much more separated and protected from the solar wind. Under these conditions, the effects of CMEs are much less significant. In order to know what is going to happen on Earth it is important to know not only what happened on the Sun but also the characteristics of the magnetic fields that are carried along with the solar wind. Accurate forecasts of large storms are difficult to achieve because the propagation of solar disturbances to the Earth and magnetic field characteristics are difficult to predict with high accuracy. Using the solar wind and interplanetary magnetic field quantities measured upstream from the Earth as input, however, can warn of the impending arrival of solar wind structures and, therefore, predict their geoeffectiveness, that is to say, their effect on the Earth. Forecasters have reviewed large amounts of historical solar wind data and have found that they can identify and predict the occurrence of large storms, with accuracy in the range 70–80%.

Recently, a number of spacecraft have been launched



(*a*)



(*b*)

Figure 1. Coronagraph observations of the solar corona from the SOHO spacecraft: (*a*) shows a clear halo CME propagating away from the Sun and (*b*) shows a limb event.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

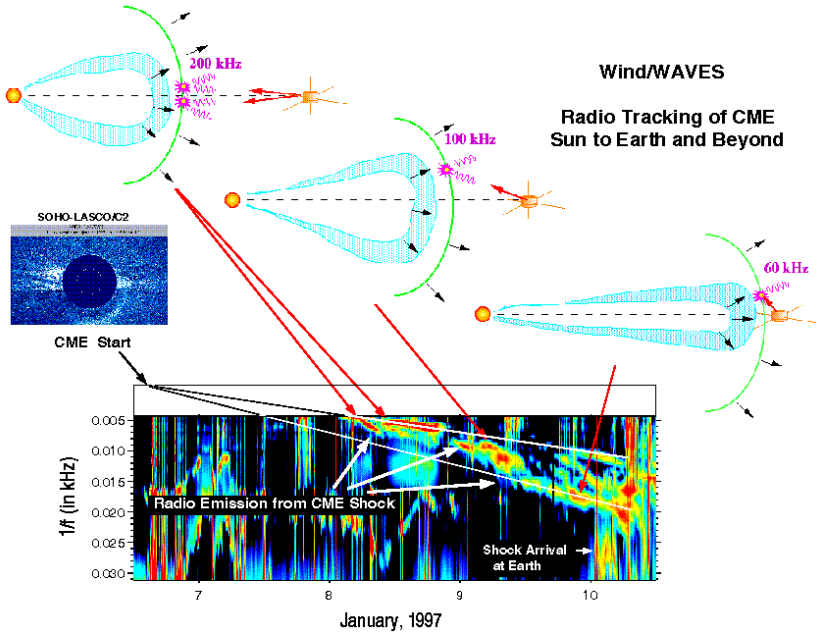


Figure 2. Dynamic spectra of radio data taken by the WIND spacecraft during 6–11 January 1997 in the frequency range 23–245 Hz. Radio emissions from the CME shock are marked on the figure and decrease in frequency as the CME moves further from the Sun. The cartoon above the spectrogram represents the progress of the CME as it expands away from the Sun towards the WIND spacecraft located about half a million miles in front of the Earth.

which have improved our ability to forecast space weather disruptions. SOHO (Solar and Heliospheric Observatory), a joint NASA (National Aeronautical and Space Administration) and ESA (European Space Agency) mission, provides near-real-time images of the Sun and its CORONA, clearly showing ACTIVE REGIONS, flares and CMEs. Events occurring on the side of the Sun—the limb—are regularly seen which will not affect the Earth. Sometimes, however, CMEs occur on the side of the Sun facing Earth. These events appear to be very different when viewed from Earth. Instead of looking like a 'bubble' of plasma, they form a circle of light around the Sun. This light is much dimmer than the Sun itself, so it is necessary to put an occulting disk in front of the Sun in order to see what goes on around it—like an artificial solar eclipse. The instrument used to do this is called a CORONAGRAPH. An example of such a 'halo' event is shown in figure 1, together with one occurring on the limb for comparison.

The faster CMEs have outward speeds of up to 2000 km s^{-1}, considerably greater than the normal solar wind speeds of about 400 km s^{-1}. These produce large shock waves in the solar wind as they plow through it, much like a jet breaking the sound barrier drives a sonic boom into the atmosphere. In the case of the CME shock, however, the sonic boom takes the form of very intense, low-frequency radio emissions known as solar type II bursts. (Instruments on Earth cannot detect the type II bursts because the ionosphere blocks the waves. Only a radio receiver in space can detect them.) As the shock approaches Earth, these type II bursts drift down in frequency (see figure 2) because the electron density of the interplanetary medium to which the solar bursts are 'tied' is also constantly decreasing. Space-borne instruments are capable of detecting these radio bursts from a great distance, long before the CME arrives at Earth. These radio receivers can also determine the direction from which the waves are arriving. Combined with the known propagation speed of the shock, this information allows forecasters to track the CME shock front nearly from the time that it is no longer visible to the solar telescopes, all the way to Earth. Finally, as mentioned above, it is critical to have a satellite between the Sun and Earth so that the nature of the solar wind can be observed before it arrives. In August of 1997, NASA launched ACE, the Advanced Composition Explorer satellite, which is located a million miles upstream of the Earth. This location allows ACE to detect the Earth-directed solar wind about 1 h before it impacts us.

The NOAA Space Environment Services Center (SESC) in Boulder is one of the world centers making forecasts of solar and geomagnetic activity (see figure 3). Daily predictions are issued for the likelihood of solar flares, proton flares, x-ray events and magnetic storms. Longer-range forecasts are also made so that the launches of manned space flights can be planned with more safety. The SESC receives about 1400 data streams daily, including x-ray and particle flux data from the GOES satellites, Hα images and magnetograms from observatories around the world, measurements of the geomagnetic field at many locations, and 10.7 cm radio levels from several

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3. The new NOAA Space Environment Forecast Center. In a similar manner to meteorologists checking temperatures, winds and pressure to predict the weather on Earth, scientists monitor the Sun and the space environment to help forecast space weather.

radio telescopes. Each day the features of the solar disk are mapped by hand so that the evolution of active regions, coronal holes, filaments, and neutral lines may be carefully studied. Forecasters attempt to consider all of this information when making their daily forecasts of solar effects on Earth. At the present time, these forecasts are not very reliable; major flares are sometimes not forecast and predictions that are made often do not come true. Even though forecasters have a large amount of data to work with, the physics of the Sun, the magnetosphere and the interplanetary medium is not well understood. At the present time, many partial mathematical models have been developed, but there is no comprehensive model of the solar–terrestrial environment.

In most cases, the ability to predict the behavior of nature comes from a mathematical model. Earth weather forecasters have been trying for the last 30 yr to construct a mathematical model of the global weather using the very complex equations of fluid dynamics to describe the circulation of the oceans and atmosphere. Even with the best supercomputers to run these models, it has proven impossible to precisely model Earth weather. Modeling the solar–terrestrial environment is vastly more complex. The physics necessary to do this includes not only fluid dynamics but also Maxwell's equations. This combination is known as MAGNETOHYDRODYNAMICS (MHD), and at the present time the equations of MHD cannot be completely solved analytically.

Research to improve solar forecasting is occurring in three major areas. The *first* is improving the basic understanding of the way in which eruptions from the Sun occur, namely of prominence eruptions and CMEs. In particular, a workable two-dimensional model of the way a coronal arcade loses equilibrium and erupts as an MHD catastrophe is being developed into a three-dimensional model. Also, models are being developed for the three-dimensional evolution of coronal structures before they erupt. The *second* area is the correlation of observable phenomena with effects on Earth. For example, a strong correlation has been observed between sunspot cycles and disturbances on Earth. However, this correlation is very coarse; it is known that during a certain period of years there will be high levels of solar activity and accompanying disturbances on Earth. It is not yet possible to accurately predict these disturbances as happening over specific days or hours. Many researchers are trying to refine the correlations between observable symptoms, such as increased radio emission, and subsequent eruptions of mass. Some of the best correlations yet are those that have been found between the evolution of sunspot groups and eruptions.

The *third* area of work is that of constructing a model for the solar–terrestrial environment. In addition to the complexities of MHD, the problem is difficult because there are three different domains involved, which all couple together. The first domain is that of the Sun; to simply construct a mathematical model of the Sun is far beyond researchers at the present time. There are still many mysteries about what processes occur inside the Sun, what triggers flares and even why sunspots form. The second domain is the interplanetary medium, once thought of as empty space. This space is filled with the solar wind plasma, which is not fully understood. The third domain is the Earth's magnetosphere, with its many regions and currents. The magnetotail, extending for millions of kilometers out from Earth, has been difficult to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

study directly and remains poorly understood. Models for any one of these domains by themselves are not even close to completion, but the final complication arises because these three domains are not at all separate. A change in one of these domains can have major consequences on the Earth. Scientists and forecasters alike continue to analyze misses and false alarms in detail to better understand the performance characteristics of the prediction methods, and a number of improvements have been developed to deal with the known failure modes. It is hoped that one day there will be a comprehensive model for the entire solar–terrestrial environment but this is certainly a problem for physicists of the future. Forecasters will continue to make forecasts based on present knowledge and available data, and hopefully the forecasting ability will continue to improve. The physics of the solar–terrestrial environment is still one of the great frontiers, awaiting new generations of scientific explorers.

Nicola J Fox

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solar-B

Japanese–US–UK mission to study the Sun's corona, scheduled for launch in 2004. Expected to carry a 50 cm optical telescope, an imaging x-ray telescope and an extreme-ultraviolet imaging spectrometer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SolarSoft

The SolarSoft system (SSW) is a set of integrated software libraries, databases, and system utilities which provide a common programming and data analysis environment for solar physics. The SolarSoft environment provides a consistent look and feel at widely distributed co-investigator institutions to facilitate data exchange and to stimulate coordinated analysis. Commonalities and overlap in solar data and analysis goals are exploited to permit application of fundamental utilities to the data from many different solar instruments. The use of common libraries, utilities, techniques and interfaces minimizes the learning curve for investigators who are analyzing new solar data sets, correlating results from multiple experiments or performing research away from their home institution.

SolarSoft is primarily written in the Interactive Data Language (IDL) which is ideal for manipulating and visualizing image, spectral, time series and other data types common in solar physics. The extensibility of IDL and large number of SSW contributors, coupled with SolarSoft automated upgrade and software exchange utilities, stimulates rapid evolution of analysis capabilities. The SolarSoft system was originally built from the Yohkoh, SOHO, Solar Data Analysis Center (SDAC), and the Astronomy IDL software libraries and it draws upon contributions from many members of those projects. More recent participants in SolarSoft include TRACE, HESSI, SXI, and many ground-based optical and radio observatories.

Principal goals of SolarSoft

The primary goals of SolarSoft include:

- *Provide a large reusable software library.* The shared, general purpose SSW library provides a large and growing collection of observatory and instrument independent software. This permits instrument Principal Investigator (PI) teams to concentrate software efforts on detailed analysis and calibration issues rather than reinventing general use development and analysis software.
- *Promote an evolutionary environment.* Feedback and contributions from the many SSW institutes and users are strongly encouraged. Utilities to simplify remote SSW upgrades and accept user contribution are designed to make the system evolutionary instead of stagnant and divergent.
- *Promote the use of standards that will facilitate coordinated data analysis.* Through use of a *minimal* set of suggested standards, many powerful applications are immediately available for the analysis of new data sets.
- *Provide easy access to supporting ancillary databases.* Access to supporting data sets from ground-based and satellite-based solar observatories are provided with consistent interfaces to facilitate coordinated analysis.
- *Provide a file-format-independent analysis environment.* To accommodate the largest number of solar data sets, file I/O is isolated and analysis applications are designed to operate on objects which are independent of file type.
- *Provide a hardware- and site-independent analysis system.* Fundamental shared software is written in a hardware-independent and site-configuration-independent form. The SSW setup utilities support local configuration files to tailor the site and user's environment.
- *Permit simplified and customized installation and environment definition.* Due to the large number of contributing instruments and libraries, integrated function should be preserved when any desired subset of libraries is included in a local installation or individual analysis session.
- *Maximize Principal Investigator team control of instrument level software.* Within the integrated system and analysis environment, allow PI teams full control of the instrument specific analysis software and provide system utilities and suggested guidelines to faciliate integration.
- *Provide integrated access to other IDL packages.* When IDL libraries that provide a useful extension to SSW capabilities are identified, they are integrated into the SSW system.

The reusable software library

The reusable software library is the heart of the SolarSoft system. Routines in this library are instrument and mission independent, and available to all users for analysis and application development. This library represents an evolutionary system, dating back to SMM analysis at the Solar Data Analysis Center at Goddard Spaceflight Center (GSFC). It has since evolved through the Yohkoh and SoHO missions, and now includes contributions from TRACE, HESSI, and from many other solar instrument PI team members.

A very small sample of existing library capabilities include:

- time series analysis, time conversions, time series plotting
- spectral fitting
- image and image cube (movie) processing and display
- solar image data routines (limb fitting, image and grid overlay, coordinate transformations, feature tracking, co-alignment, etc)
- file I/O (generic binary, ASCII, FITS, etc)
- IDL data manipulation (structure, string, array, mathematics, etc)
- WWW interface (HTML conversion, file conversions, FORM handling, WWW server mode)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

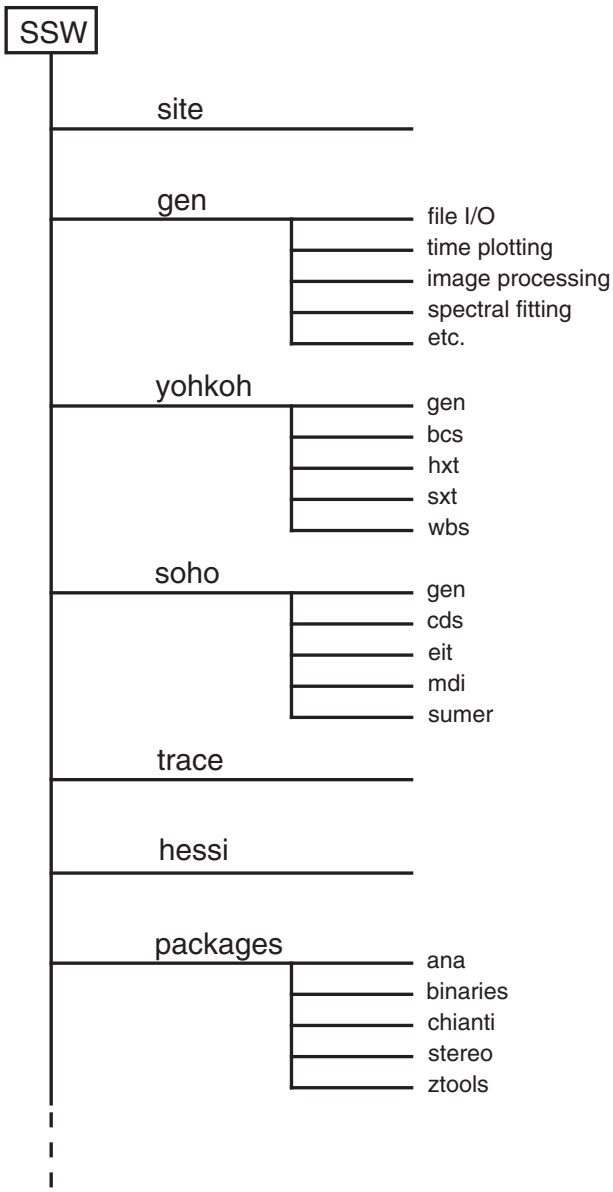


Figure 1. SolarSoft directory tree structure. The software libraries for each instrument are organized into a number of functional subdirectories which are not shown.

SolarSoft system overview

Library and path organization

A schematic of the upper level SolarSoft library and directory hierarchy is illustrated in figure 1.

The topmost level includes a SITE branch for local configuration files, the GEN branch which contains the SSW general purpose library and an extensible number of mission level branches. For clarity, only a representative subset of missions are included in the figure. In this context, a mission is a set of one or more related instruments or observatories. Mission branches include single instrument observatories, such as TRACE and HESSI, multi-instrument observatories, such as Yohkoh and SoHO, and groupings of related ground based observatories, such as optical and radio.

This intermediate level of organization by mission facilitates access to shared libraries among the related instruments or observatories. The contents of those libraries are mission dependent but instrument independent. For example, all of the instruments on a multi-instrument orbital observatory desire access to spacecraft ephemeris software and data bases. The gen branches under the mission level provide common, single source software to all of the related instruments (Yohkoh-gen and SoHO-gen for example).

The instrument level branches contain the calibration files and software which are required for detailed or specific instrument data analysis. With the exception of very minimal library and setup file naming conventions, the contents and organization of the branches under the instrument levels is (and *must be*) defined by and controlled by the the instrument PI teams.

Mastering, integration, and verification

The master copy of SolarSoft and support utilities is maintained at the Solar Data Analysis Center at GSFC. This is also the master site for the general purpose SSW library. Only a few core SSW members are authorized to update the general library. Since this is automatically distributed to all remote SSW sites, routines must remain backwardly compatible and continued support for older IDL versions must be considered. All additions and modifications are logged, and backup copies of modified modules are maintained by the *SSW online* utility. Additional protection is provide by the SDAC system facilities, which include multiple levels of automated backup and disk recovery.

Principal Investigator (PI) teams from each participating instrument define and control the master libraries and databases for the individual instruments. These instrument and mission level master sites are widely distributed, and currently include institutes in Japan, England, France, Switzerland, and many institutes within the United States.

SolarSoft utilities running on the SSW master site at GSFC update from the various instrument master sites into the SSW master for centralized integration and distribution. Due to the very large number of contributing sites and routines, background utilities (written in SSW) monitor incoming additions and flag potential conflicts within the integrated system. Additional SSW utilities monitor modifications and sizes of software libraries and data bases to generate system files required by the installation and upgrade utilities.

Now that the SolarSoft organization, monitoring, and distribution system is in place, adding new instrument libraries into the infrastructure approaches 'plug and play' simplicity, assuming that the PI teams are willing to address any conflicts which are flagged by the system. There is an intermediate level of integration available which allows new libraries to take advantage of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

conflicts checking and distribution portions of SSW without requiring more rigorous, conflict-free standards suggested for fully integrated libraries.

Remote installation and upgrades

SolarSoft is easily installed through a web-based interface. Although the number of instrument libraries and databases under SSW is extensive, a primary design goal was to permit remote sites to install any desired subset of SSW while still preserving the integrated function. The SSW installation FORM allows the WWW client to select the desired instrument libraries and generates a customized installation script. When executed on the local machine, the script creates the necessary directories, automatically transfers the required installation files from the designated SSW host, and installs them to produce the local customized SolarSoft tree. Due to the evolutionary philosophy of SolarSoft, frequent periodic upgrades are strongly suggested (daily is typical) and utilities are included in the SSW distribution to automate upgrades of the local SSW system. The installation and upgrade utilities use a programmable SSW host internet address for file transfers to permit use of the SSW 'master' with the best internet connection to the local site. Following the initial installation and configuration, it is sufficient to add instrument or mission names to a SSW site configuration file to automatically include new branches in the local SSW system. The upgrade process is incremental so that only new or modified software is transferred.

The history of SolarSoft

The overall organization and philosophy of the SSW system structure draws heavily from the Yohkoh data analysis system. Even though Yohkoh observatory included four very different science instruments, it was obvious that there were a large subset of common analysis goals and that the use of some common definitions would permit use of fundamental cross instrument software. For example, all four instrument teams required access to Yohkoh ephemeris and supporting ancillary databases—those commonly desired software functions should be written once, not rewritten four times. The analysis of data from each Yohkoh instrument is greatly enhanced by crucial context provided by the other instruments. A minimal set of parameters, such as observation time, were defined in common formats to simplify cross-instrument access and facilitate coordinated analysis.

A key requirement and innovative feature of the Yohkoh system was a simple, internet-based installation, upgrade and feedback infrastructure. Since the Yohkoh co-investigator institutes included widely distributed sites in Japan, the UK, and several states in the US, it was important to propagate changes and incorporate contributions quickly enough to avoid divergent libraries at the remote institutes. Installation and upgrade utilities developed prior to Yohkoh launch (and predating the WWW) greatly simplified this task.

Another design goal was platform independence. At the time of Yohkoh design, many different computer platforms were in use at the Institute for Space and Astronautical Science (ISAS) in Japan and at many co-investigator institutes in the UK and USA. Systems designated for Yohkoh analysis included DEC VMS, DEC Ultrix, and Sun and RISC workstations running Unix.

Selection of IDL

The Interactive Data Language (IDL) software package written by Research Systems Inc. (RSI), Boulder, Colorado was ideal for analysis of the disparate science data sets on Yohkoh, which include images, spectra and time series. The use of IDL also moved the majority of the system dependency burden from Yohkoh programmers to the RSI/IDL programmers. Some of the Yohkoh team members had previously worked at the SDAC on the Solar Maximum Mission (SMM)—several of the SMM instrument teams and SDAC members were using IDL for data analysis and this provided an experience base and the starting seed for a general purpose IDL library.

Initially, some Yohkoh teams tried to maintain portions of the software using heritage Fortran programs. They rapidly discovered that, in order to support the different platforms used for Yohkoh analysis, five or more executables needed to be built for each task. Updating and propagating an application to the distributed investigator sites in a timely fashion was nearly impossible, since the multiple executables required verification on each of the architectures. It was therefore decided to translate the bulk of the remaining Fortran software into IDL. There are still a very few tasks written in Fortran or C, but these are primarily applications which have specialized use and only need to be run at a restricted number of sites.

By following a few simple programming rules and extending the language as required with a few key interface routines, the Yohkoh IDL-based system was easily transportable. For example, filenames are never specified with an explicit directory or path—environmental variables, or their equivalent are used to indicate file locations located within the computer file system. The mapping of several key environmentals to the local, hardware specific pathnames is accomplished by simple modification to setup files in the site branch. Relative pathnames are then constructed dynamically by interface routines which use the context of the current operating system. Yohkoh contributors were encouraged to prefix instrument-specific routine filenames with their instrument acronym in order to minimize potential collisions in routine names.

From Yohkoh to SolarSoft

Prior to the launch of the Solar and Heliospheric Observatory (SoHO) it was decided to review the Yohkoh utilities, identify all of the truly mission independent routines and place them into a shared library. A prime motivator for this was that quite a few of the Yohkoh science and programming team were now members of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

various SoHO instrument or analysis teams and they desired to continue using and extending familiar analysis and programming utilities. A common library was then available for Yohkoh, SoHO and future missions for applications development and analysis and offered future potential for coordinated analysis. A mission level layer was added to the existing Yohkoh instrument level library and path infrastructure and SolarSoft was created.

Recent trends—SolarSoft and the World Wide Web

An extensive set of SolarSoft routines and techniques have evolved which leverage WWW technologies and offer great potential for distributed solar physics analysis. In addition to extensive file converters, movie making tools and HTML generators written in IDL, SolarSoft also includes an innovative set of utilities for executing SolarSoft applications through any WWW client (browser). This system uses a very simple (about five line) CGI script to pass information from an WWW FORM to an IDL server which is running the SolarSoft environment. An SSW IDL interface function converts the FORM stream directly into an IDL data structure. This infrastructure implies that all data base searching, access, analysis, and display functions which are available within an interactive SolarSoft session are also available through a WWW client. Examples of this server mode SSW application include the remote SolarSoft software and data base installation FORMS, customized Yohkoh/SXT movie maker, the TRACE WWW catalog search and data request system and many others. The HESSI team and Max Millennium Program is using this SSW server technology to provide WWW client access to the ancillary data bases available at the SDAC.

Instruments supported by SolarSoft

By 1999, analysis capabilities for a large number of instruments from many different solar missions were fully integrated into SSW, including:

Yohkoh: BCS, HXT, SXT, WBS
SoHO: MDI, EIT, CDS, SUMER, LASCO
TRACE: TRACE
CGRO: BATSE
GOES: Soft X-ray Monitor
Spartan: Spartan
SMM: XRP, HXIS, HXRBS, CP UVSP, GRS
Radio: ETHZ, Nancay, Nobeyama
Optical: SOON, La Palma

Descriptions of the space-based instruments may be found in SOLAR TELESCOPES and INSTRUMENTS: SPACE. The ground-based radio and optical observatories are described in SOLAR TELESCOPES and INSTRUMENTS: GROUND.

Other missions which are expected to participate, with the planned launch dates indicated, include:

HESSI: Hessi (2000)
GOES: Soft X-ray Imagers M,N,O (2000-2005)
Solar-B: SOT, XRT, EIS (2004)

Software packages by other individuals or groups which provide useful extensions are integrated and distributed with SSW. These SSW *packages* include:

Chianti: UV and EUV diagnostics; K Dere
ANA: public domain data analysis; R Shine
Stereo: dynamic stereoscopy; M Aschwand
Ztools: vector and solar tools; C DeForrest
SPEX: spectral analysis; R Schwartz

Acknowledgments

The success and evolution of solarsoft depends upon contributions from its many users and SSW instrument team members. Special thanks are due to members of the Yohkoh and SoHO co-investigator institutes, the Solar Data Analysis Center at GSFC, and the Yohkoh Data Archive Centre (YDAC), at the Mullard Space Science Laboratory. A portion of the work for design and administration of SolarSoft is supported under NASA contracts NAS8-40108, NAS5-32966, and NAS5-38099.

Bibliography

Freeland S *http://www.lmsal.com/SolarSoft/* SolarSoft description and links to SSW installation, analysis concepts, standards, and other related documents (World Wide Web)

Svestka Z and Uchida Y (eds) 1991 *The Yohkoh (Solar-A) Mission* (Dordrecht: Kluwer)

Fleck B, Domingo V and Poland A (eds) 1995 *The SOHO Mission* (Dordrecht: Kluwer)

Samuel L Freeland and Robert D Bentley

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Solstices

The point on the ecliptic at which the Sun reaches its maximum declination north or south of the celestial equator. The greatest northerly declination corresponds to the summer solstice, the greatest southerly declination to the winter solstice. In the Northern Hemisphere the summer solstice occurs on 21 or 22 June and the winter solstice on 21 or 22 December. In the Southern Hemisphere, where the seasons are reversed, the situation is exactly the opposite.

In the Northern Hemisphere at the time of the summer solstice, the North Pole is tilted 23.45° towards the Sun. Because the Sun's rays are shifted northward by the same amount, the vertical noon rays are directly overhead at the Tropic of Cancer. Six months later, the south polar end of the Earth is inclined 23.45° towards the Sun. On this day of the summer solstice in the Southern Hemisphere, the Sun's vertical overhead rays progress to their southernmost position, the Tropic of Capricorn.

See also: celestial coordinates.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sombrero Galaxy

Popular name for M104, a spiral galaxy in the constellation of Virgo which is almost edge-on as seen from the Earth. The disk that surrounds its unusually large central bulge contains a very prominent lane of obscuring dust which, because the disk is tilted to the line of sight by a small angle (about 6°), cuts across the face of the bulge. These features cause the galaxy to look rather like a wide-brimmed sombrero hat.

Located at a distance of about 60 million light-years, the Sombrero Galaxy is a massive spiral with a bright compact nucleus that is star-like in appearance. Spectroscopic observations reveal high rotational velocities inside this nucleus, and imply that about a billion solar masses of material is concentrated very closely together, possibly within a supermassive black hole, at the center of M104

See also: active galactic nucleus, black hole, Doppler effect, Messier Catalog, spectrum, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Somerville, Mary [née Fairfax] (1780–1872)

Physicist and mathematician, born in Jedburgh, Roxburghshire, Scotland, self-educated, and educated with the aid of her brother's tutor. With the early death of her husband, she found herself a widow of independent means and took up mathematics, studying astronomy and dynamics. She carried out experiments on magnetism and the solar spectrum. She wrote a popularized rendition of Laplace's *Mecanique Celeste* and Newton's *Principia* called *The Mechanism of the Heavens*, and an account of *The Connection of the Physical Sciences*. Her discussion of a hypothetical planet causing the orbital perturbations of Uranus inspired JOHN ADAMS to make calculations of the position of Neptune. Mary Somerville and Caroline Herschel (see HERSCHEL FAMILY) were the first two women to write papers read to the Royal Society. *Somerville College* in the University of Oxford was named after her.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sosigenes of Alexandria (*c.* 50 or 90 BC–?)

Peripatetic Greek philosopher, born in Alexandria, astronomer and mathematician who advised Julius Caesar on calendar reform. He recommended a year of 365.25 days, and inserted extra days into the year 46 BC to bring the calendar back in register with the seasons. His account of Eudoxus's planetary system is the most complete surviving account.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

South African Astronomical Observatory

The South African Astronomical Observatory (SAAO) is the national facility for optical and infrared astronomy in South Africa. Its prime function is to further fundamental research in astronomy and astrophysics at a national and international level through the provision and use of a world-class astronomical facility. There are strong international links to other countries worldwide through scientific collaboration and technological exchange. The SAAO contributes to the future development of South Africa by providing training in a scientific and high-tech environment; by stimulating young people to follow careers in science and technology through a school and teacher science education program; and by helping to create a culture of science and technology amongst all communities by a vigorous science awareness program.

The SAAO is one of a number of research facilities in South Africa run by the National Research Foundation and it is financed by a government grant. SAAO's headquarters are in Cape Town (at the Royal Observatory, Cape of Good Hope, founded in 1820) and the telescope facilities are located at Sutherland in the Northern Cape. The four main telescopes range in aperture from 1.9 m to 0.5 m. The major new facility will be the SOUTHERN AFRICAN LARGE TELESCOPE (SALT) to be constructed by South Africa and international partners. This 10 m class telescope will be based on the Hobby–Eberly Telescope at MCDONALD OBSERVATORY, Texas and is of tilted-Arecibo design.

For further information see
http://www.saao.ac.za.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

South African Astronomy

Prior to the middle of the eighteenth century almost all of the astronomical observations made in South Africa were those necessary for the determination of position—particularly that of the Cape of Good Hope. Some exceptions were the independent discovery of bright comets, of which van Riebeeck's comet of 1652, named after the first Governor of the Cape, is an example.

Because of the need for a comprehensive survey and catalog of the southern sky, in 1751 the French Academy of Sciences sent Nicolas Louis de la Caille (1713–62; usually known as LACAILLE) to the Cape for two years. Establishing a small observatory in Strand Street, Cape Town, during the 12 month interval starting in August 1751 Lacaille measured the positions of over 9800 stars. With this came knowledge of about 40 nebulae and clusters. In order to give classification numbers to the brightest stars, Lacaille invented many of the southern constellations that are in use to this day: Horologium, Octans, Microscopium etc.

During his second year Lacaille measured an arc of meridian in order to determine the radius of the Earth at the southern latitude of 34°.

The Royal Observatory, Cape of Good Hope

No further significant astronomical observations were made in South Africa until the founding of the Royal Observatory, Cape of Good Hope (RO CGH, figure 1) in the third decade of the nineteenth century. The foundation of this observatory, located about three miles east of Cape Town, was motivated by the fact that British astronomy was falling behind that on the European continent. The two responses were the founding of the Astronomical Society (later the Royal Astronomical Society) and the observatory at the Cape.

The RO CGH was built during 1825–8 and was directed by Fearon FALLOWS (1788–1831) who had been a Cambridge mathematician and exact contemporary of John HERSCHEL. The observatory was equipped for meridian astronomy—with a transit instrument and meridian circle identical to those employed at the ROYAL OBSERVATORY, GREENWICH.

Fallows died in 1831 from the effects of scarlet fever before he could prove the worth of the instruments. He was succeeded in 1831 by Thomas HENDERSON (1798–1845) who only remained a year, but his extensive use of the meridian instruments resulted in his detection and measurements of the parallax of ALPHA CENTAURI—one of the first of measured distances to stars. In 1833 Henderson left to become the first Astronomer Royal for Scotland, at the ROYAL OBSERVATORY, EDINBURGH.

Henderson was succeeded in 1834 by Thomas MACLEAR (1794–1879), who labored successfully to make the RO CGH into a world-class institution. Maclear and his principal assistant Charles Piazzi SMYTH (1819–1900), who succeeded Henderson in Scotland in 1845, were deeply engaged from 1837 to 1850 in re-measuring and extending Lacaille's arc of meridian with the result

Figure 1. The Royal Observatory, Cape of Good Hope.

that fewer stellar positional catalogs were produced than expected. When Maclear's principal work, the results of the arc of meridian, was published in 1866 it was considered the finest of its kind, and Maclear is known more as the founder of trigonometric survey in South Africa than for his contributions to positional astronomy.

During his first four years at the RO CGH Maclear worked closely with Sir John Herschel, who resided at the Cape from 1834 to 1838, providing him with accurate positions of the bright stars needed by the latter for reducing his 'sweeps' of the sky for nebulae, clusters and double stars.

Edward James Stone (1831–1897), who had been Chief Assistant to the Astronomer Royal, George Biddell AIRY, at Greenwich, became director of the RO CGH in 1870 and until he left in 1879 concentrated on preparing for publication all of the positional observations made by Maclear and his assistants. His 'Cape Catalogue' of 12 441 stars, published in 1880, included the extensive work of himself and his predecessor.

The RO CGH continued into the twentieth century as the only professional observatory in South Africa. Stone was replaced in 1879 by David GILL (1843–1914) who was later described (by Sir Robert BALL) as 'the most distinguished practical British astronomer since BRADLEY who has presided over one of our national observatories'. Gill's exceptional talents in design and application of astronomical instrumentation raised the RO CGH to the equal of any observatory in the world. His measurement of stellar parallaxes and (among many other examples) of the positions of Jupiter's satellites, using a heliometer of his own design, gave fundamental data of the highest available precision. His invention of wide-field photography for use in construction of catalogs of positions of stars led directly to the 'Cape Photographic Durchmusterung' and to the 'Astrographic Catalogue'.

David Gill's organization of an international campaign of observation of the minor planets Iris, Victoria and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sappho, led to his deducing a solar PARALLAX and (indirectly) a mass determination for the Moon that were both accurate and not improved upon for many decades. As with Maclear, Gill participated in and supervised a major trigonometric survey, extending into what is now Zimbabwe. After retirement in 1906, Gill (knighted in 1900) returned to Britain; his principal occupation was the production of his book *History and Description of the Royal Observatory, Cape of Good Hope*, which was modeled on F G W Struve's similarly monumental description of the Pulkova Observatory, which had strongly influenced Gill in his development of the RO CGH.

Among the new instruments that Gill had designed and had built at the Cape Observatory were a unique Reversible Transit Circle, a twin 18/24 in refractor and the astrographic telescope. These continued to be employed in the first half of the twentieth century by Gill's successors: Sydney Samuel Hough (1870–1923; director 1907–23), Harold SPENCER-JONES (1890–1960; 1923–33), John Jackson (1887–1958; 1933–50) and Richard Hugh Stoy (1910–1994; 1950–68). During this time the work of the RO CGH was largely to produce another multivolume star catalogue—the *Cape Photographic Catalogue*, but new areas of astrophysical measurement were introduced in support of the catalog: in particular photometry and spectral types for most of the brighter stars.

A large effort was made in production of stellar parallaxes, using the photographic 24 in refractor, and Spencer-Jones organized an international campaign of observing the minor planet EROS from which he deduced another value for the solar parallax.

As, in effect, the national observatory, although it continued to be run by the British Admiralty until 1965, the RO CGH provided the time service for South Africa—initially through time balls, then electric telegraph and from 1925 by radio transmission. To this day the noonday cannon on Signal Hill in Cape Town is fired by electric impulse from the observatory.

At the end of 1971, through an agreement between the British and South African governments, the RO CGH ceased to exist and was reorganized at the beginning of 1972 as the SOUTH AFRICAN ASTRONOMICAL OBSERVATORY (SAAO).

The Union (later Republic) Observatory

The first truly South African observatory was opened in Johannesburg on 17 January 1905 as the Government Meteorological Observatory. Its director was Robert Thorburn Innes (1861–1933), who had worked at the RO CGH under Gill. When Gill was asked to recommend a director for the meteorological observatory he undoubtedly knew that Innes would soon expand that institution's work to include astronomical observations—Innes was a dedicated observer of visual DOUBLE STARS. By 1907 he had installed a 9 in refractor and in 1909 the South African government agreed to the purchase of a $26\frac{1}{2}$ in refractor from the Dublin optical firm of Sir Howard GRUBB. This was only installed in 1925. However, as a gift from the English amateur astronomer John FRANKLIN-ADAMS, in 1909 a 10 in photographic refractor was acquired and the resulting Franklin-Adams photographic atlas of the southern sky was the standard reference for the subsequent 50 years.

In 1912 what had been known as the Transvaal Observatory was officially designated the Union Observatory, with its director entitled the Union Astronomer. Altogether there were four Union Astronomers in the working life of the observatory: Innes (1912–27), H E Wood (1927–41), W H van den Bos (1941–56) and W S Finsen (1957–65). The work of the observatory (renamed the Republic Observatory in 1961) was largely the measurement and computation of double-star orbits—in later years the measurements of very close double stars were made with an ingenious 'eyepiece' interferometer invented by Finsen. Some 3000 southern double stars were found with the large refractor. However, much photographic survey work was also carried out, leading to the discovery of PROXIMA CENTAURI by Innes, the discovery of asteroids by Wood—up to the end of Wood's directorship about 600 asteroids had been found—and tens of thousands of photographs of Mars by Finsen.

The Republic Observatory was effectively closed in 1972 when the recently (1968) acquired 20 in reflector was moved to the site of the new national observatory in Sutherland. The $26\frac{1}{2}$ in refractor remains at its original location and is used occasionally for public demonstrations.

The Boyden Observatory

In 1927 Harvard College moved its Boyden Observatory from Arequipa in Peru to the resort of Mazelspoort, near Bloemfontein. Equipped with the Bruce astrograph, the 60 in Rockefeller reflector and a 13 in Alvan Clark refractor, a variety of observational programs were possible. The first director, John Stefanos Paraskevopoulos (1889–1951) carried out all-sky photographic patrols of variable stars and made detailed studies of variable stars, particularly in the LARGE and SMALL MAGELLANIC CLOUDS.

Later collaboration with other northern observations resulted in the acquisition of the ADH (Armagh–Dunsink–Harvard) Schmidt telescope of 36 in aperture.

The observatory was given to the University of the Orange Free State in 1972, after which it operated at only a very low level. The 60 in reflector is currently being modernized and should soon be brought into full-time operation.

The Leiden Southern Station

In 1923 Innes at the Union Observatory negotiated an agreement that enabled astronomers from the LEIDEN OBSERVATORY in Holland to make use of the Union Observatory's facilities. In 1938 the South African government provided funds for a building in the Union Observatory's grounds to house the Leiden Observatory's twin 16 in astrograph, known as the Rockefeller Telescope. In 1954 a dark site, remote from the Johannesburg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

city lights, was chosen at Hartebeespoort and first the Franklin-Adams telescope and then the twin astrograph moved there. This site was operated by Leiden observers, although officially an outstation of the Union Observatory. A 36 in reflector devoted to multichannel stellar photometry was established there in 1957 by Theodore Walraven. This telescope was later moved to the EUROPEAN SOUTHERN OBSERVATORY in Chile; the Franklin-Adams telescope remains, used for educational purposes.

The Yale Observatory, Johannesburg

A 26 in visual refractor was established in 1925 by Yale Observatory, on the campus of the University of the Witwatersrand in Johannesburg. This was a southern station used for photographic parallax measurements. The directors were H D Alden (from 1925 until 1945) and C Jackson from 1945 until the telescope was transferred to Canberra, Australia, in 1952.

The Lamont–Hussey Observatory

In 1927 the University of Michigan established a 27 in visual refractor on Naval Hill at Bloemfontein. The director was R A Rossiter (1886–1977) who, by the time of his retirement in 1951, had discovered 7120 new southern double stars. F Holden was director from 1952 until the observatory was closed in 1974. The optics of the refractor returned to Michigan; the observatory dome remains, the property of Bloemfontein City Council.

The Natal Observatory

Edmund Nevill (alias Edmund Neison, 1849–1940), Government Chemist for the Province of Natal, was the first and only director of the provincial observatory erected at Durban in 1882. Neison is best known for his book *The Moon*, published in 1876, before he went to South Africa. Lack of funds prevented the Natal Observatory from publishing the work (mostly meridian observations) carried out. The observatory closed in 1911.

The Radcliffe Observatory, Pretoria

The Radcliffe Observatory in Pretoria (figure 2) was established in 1939 from the proceeds of the sale of the observatory in Oxford, owned by the Radcliffe trustees. The observatory was equipped with a 74 in reflector by Grubb Parsons (this was the largest in the southern hemisphere at the time), but the mirror's arrival was delayed by the Second World War, so the telescope only started work in 1948. The first director was Harold Knox-Shaw (1885–1970), transferred from Oxford. He retired in 1951 and was replaced by Andrew David Thackeray (1910–78) who directed the observatory until it was sold to the South African government in 1974 and the telescope moved to Sutherland, where it came into operation in 1976.

During its 26 years of active life, and operating on a very small annual budget, the staff and visitors at the Radcliffe Observatory put astrophysics in South Africa into the world class. Initially the observational program concentrated on radial velocities of OB stars with a view to determining the dynamics and structure of the Galaxy, but later usage expanded into studies of the spectra of peculiar stars, production of color–magnitude diagrams for southern clusters, and particularly spectroscopy and photometry of stars in the Magellanic Clouds.

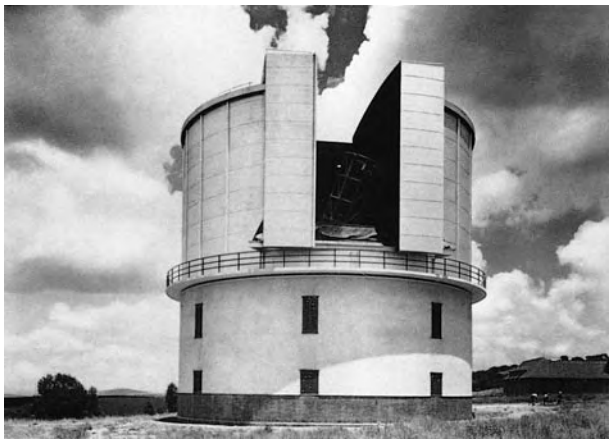


Figure 2. The Radcliffe Observatory, Pretoria.

The Astronomical Society of Southern Africa

This society was founded in 1922, from an amalgamation of societies in Cape Town and Johannesburg—the former having been formed in 1910 through the interest stimulated by Halley's Comet that year. The society comprises mostly amateur astronomers, with the addition of most of the professionals. There is a policy of electing presidents alternately between amateur and professional status. The society's postal address is at the South African Astronomical Observatory in Cape Town. It publishes a regular journal (the *Monthly Notes*) and an annual almanack known as the *Handbook*. There are active local centers in many parts of the country.

Astronomy at South African universities

The only South African university that possesses an autonomous Department of Astronomy is at the University of Cape Town. A sub-department exists at the University of South Africa (a distance-learning university) in Pretoria. The Physics and Electrical Engineering Department at Rhodes University, Grahamstown, specializes in radio astronomy and uses the national HARTEBEESTHOEK RADIO ASTRONOMY OBSERVATORY (HARTRAO). Active astronomical research is carried out by individuals in science departments at the University of the Orange Free State (Bloemfontein), the University of the Witwatersrand (Johannesburg) and the University of Natal (Durban).

Bibliography

Moore P and Collins P 1977 *The Astronomy of Southern Africa* (London: Robert Hale)

Warner B 1978 *Astronomers at the Royal Observatory, Cape of Good Hope* (Cape Town: Balkema)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Warner B 1995 *Royal Observatory, Cape of Good Hope 1820–1831* (Dordrecht: Kluwer)

Brian Warner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

South African Astronomical Observatory: from 1972 to the Present

In the late 1960s the Science Research Council (SRC) in the UK seriously considered closing the Royal Observatory at the Cape of Good Hope (see SOUTH AFRICAN ASTRONOMY), but the observatory (located in Cape Town) was saved by the establishment in January 1972 of a joint agreement between the SRC and the Council for Scientific and Industrial Research (CSIR) of South Africa, whereby the two Councils would fund and run the observatory. To make it possible to compete in astrophysical research it was necessary to move the principal telescopes in South Africa to a better site, away from the lights and bad winter weather of Cape Town, and away from the lights and cloudy summer weather of Johannesburg. As a result, a site was chosen at Sutherland, in the Northern Cape Province, roughly 380 km by road from Cape Town, which lies between the two heavy rainfall areas and is consequently non-seasonal and has low rainfall (annual average about 20 cm) (figure 1). This semidesert area has about 50% photometric nights and a further 25% usable for spectroscopy and differential photometry. The Sutherland site is in essence a field station, with a modest size technical support staff. The headquarters of the South African Astronomical Observatory (SAAO) is housed in buildings of the old Royal Observatory in Cape Town and that is where the astronomers and the majority of the support staff are located. The staff complement of the SAAO, including astronomers, technical and administrative support team, is about 70.

The 1 m reflector from Cape Town and the 50 cm reflector from the Republic Observatory in Johannesburg were moved to the new site in 1972, and, following the decision of the Radcliffe Trustees to withdraw from support of astronomy, the 1.9 m reflector which had been in Pretoria since 1939 was sold to the South African Government, moved to Sutherland in 1975, and was back in operation at its new location by 1976. There are still small telescopes, including an 18 in reflector as well as the 24 and 18 in Maclean refractor, at the old site in Cape Town, but these are now rarely used for research purposes.

Figure 1. The Sutherland Observatory of the South African Astronomical Observatory (SAAO).

The agreement between the British and South African scientific Councils lapsed in 1986, since when the SAAO has been funded entirely from within South Africa and has fallen under the administrative control of the CSIR and its successors the Foundation for Research Development and the National Research Foundation. Telescope instrumentation that had been contributed by UK institutions was made the legal property of the SAAO. The SAAO is a National Facility, with its own research staff, which makes time available both to university astronomers within South Africa and to visiting researchers. Approximately 50% of the observing time is used by SAAO staff, 12% by South African universities and the remainder by overseas visitors.

Under successive directors (Sir Richard van der Riet Woolley 1972–76, Dr Michael Feast 1977–92, Dr Robert S. Stobie 1992–present) the SAAO has maintained an international standing in several areas of astrophysical research, specializing largely in optical spectroscopy and in optical and infrared photometry of stars and galaxies. Although possessing only telescopes of relatively modest apertures, the availability of adequate amounts of observing time over extended periods has enabled long-term programmes to be carried out by staff researchers and university users (principally from the University of Cape Town: for further information see http://artemisia.ast.uct.ac.za) that are often not practicable at other observatories. Some general examples are long-term light curves in the optical and infrared of MIRA VARIABLES and other long-period variables, monitoring of QUASARS, extensive photometry of Cepheid variables in the Galaxy and the MAGELLANIC CLOUDS, discovery of variable stars in clusters and towards the Galactic centre, asteroseismology, and studies of cataclysmic variables. In recent years major campaigns were undertaken on the SUPERNOVA (1987a) in the Large Magellanic Cloud and as part of PLANET, which is searching for extrasolar planets through the effects of GRAVITATIONAL LENSING. An important support programme at the SAAO and its predecessor has been the provision of accurate photometric standard stars in the southern hemisphere—both the 18 in reflector in Cape Town and the 20 in reflector at Sutherland have for three decades been equipped only for standard star photometry, which has resulted in great stability of the photometric system.

The Sutherland site hosts a number of international instruments. Since 1990 one of the telescopes in the global system of the Birmingham Solar Oscillation Network, used for HELIOSEISMOLOGY, has been in operation. A 0.75 m robotic telescope (the Automatic Photometric Telescope) for stellar photometry, a joint project of the SAAO, the University of Cape Town and the University of South Africa, came into operation in 2001. Other

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

robotic telescopes are under construction. A 1.4 m reflector, the Infrared Survey Facility, constructed in Japan for the Nagoya University, started work at Sutherland in November 2000. Details of the telescopes and observing facilities at the SAAO can be found at http://www.saao.ac.za/.

There are also seismographs, operating as parts of national and international networks, housed at the Sutherland site. In 2000 a Geodynamic Observatory, installed by the GeoForschungsZentrum, Potsdam, and employing a superconducting gravimeter, came into operation.

A major project for the period 2001–04 is the construction at Sutherland of the Southern African Large Telescope (SALT), which is a near copy of the Hobby–Eberly Telescope (HET) run by the University of Texas at MCDONALD OBSERVATORY. This is an international project, involving the Governments of South Africa and Poland, and institutions in Germany, New Zealand, the UK and the USA. SALT and HET, each with 91 hexagonal mirrors of 1 m diameter mounted to simulate a telescope of about 11 m aperture, are designed largely to specialize in spectroscopy—both of very faint objects for work, for example, in extragalactic research, and of bright objects at very high dispersion, which can be used, for example, for discovery of extrasolar planets. There are agreements in place that will enable researchers to use both large telescopes, thereby gaining access to most of the sky with similarly powerful telescopes. SALT is expected to be commissioned towards the end of 2004. Information about the SALT Project can be obtained from http://www.saao.ac.za/salt/. In parallel with the SALT Project, an Information Centre is to be developed in the Sutherland village and a large Education Centre at the Observatory site itself.

Brian Warner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

South American Astronomy

South America is a region where astronomy has reached a fairly good level of development and has produced a substantial number of outstanding astronomers. However, the development of astronomy in that part of the world has been always heavily dependent on governmental financial support which has not always been as generous as necessary.

The South American countries where astronomy has grown stronger are, in alphabetical order, Argentina, Brazil, Chile, Uruguay and Venezuela. To a smaller extent, some astronomy is also carried in Bolivia and in Colombia, and is developing in Paraguay and Peru. Table 1 illustrates the situation by quoting the present status of the South American membership of the INTERNATIONAL ASTRONOMICAL UNION (IAU).

Peru in 1984, and Paraguay, in 1986, adhered to the Visiting Lecturers Program established by the IAU in 1983, and a few of the young people that attended the courses later graduated, particularly in Brazil but also in Argentina, and are returning to their home countries to help start a tradition in astronomical research.

Argentina

In Argentina, astronomical activities started with the very progressive-minded President Domingo Faustino Sarmiento, who, in 1869, invited Benjamin Apthrop Gould, the founder of *The Astronomical Journal*—whom Sarmiento had met in the US while he was Argentine Ambassador—to become the first Director of the Argentine National Observatory. This was erected in the outskirts of the city of Córdoba and had its dedication on 24 October 1871, being transferred to the local university in 1955. Córdoba produced excellent astrometric work and was well known for the *Uranometría Argentina*, an atlas of all stars visible with the naked eye from Córdoba, produced before any instrumentation became available at the site, and a series of stellar catalogs and the stellar atlas known as the *Córdoba Durchmusterung*. In the early days the observing conditions were so good that stars as of magnitude 7 were visible with the naked eye.

Table 1. The South American Countries and the International Astronomical Union.

Country	Year of adherence	Number of individual members	Country status
Argentina	1927	89	
Brazil	1961	108	
Chile	1957	44	
Uruguay	1970	3	
Venezuela	1953	11	
Bolivia	1997	0	Associate
Peru	1988	1	Associate
Colombia		3	Non-member
Ecuador		0	Non-member
Paraguay		1	Non-member

An important step was taken in 1909 when Director Charles Dillon Perrine, from the LICK OBSERVATORY, in California, decided that Córdoba ought to engage also in astrophysical work, and asked for funds for a 1.50 m telescope, that, because of the attempt to figure the mirror in-house, was not dedicated until 6 July 1942, as the main instrument of the observatory's Bosque Alegre Astrophysical Station, some 40 km southwest of the city at an elevation of 1250 m. For many years the world's first stellar spectrograph with all-reflecting optics, giving a dispersion of about 40 Å mm^{-1}, designed and built in-house by physicist Enrique Gaviola with the cooperation of physicist Ricardo Platzeck, was used at the Cassegrain focus of the telescope. At present a Russian-built multifunction instrument (a spectrograph and a 1024 CCD direct imager) is in use.

Over the course of time, several additional astronomical institutions were established in Argentina. Chronologically, the second one is the LA PLATA OBSERVATORY that started in 1882 in the charge of the Province of Buenos Aires and which was transferred to the local university in 1905. The main instrumentation, acquired in France, consisted of an 80 cm reflecting telescope, a 43 cm refractor, used extensively for the measurement of visual double stars, particularly by Michigan graduate Bernhard H Dawson, a comet-seeker and a *Carte du Ciel* astrograph. In 1935 La Plata established, during geodesist Félix Aguilar's directorship, the first University curriculum in astronomy in the country. In 1938 there were only three registered students attending, but at present this figure is of the order of 100 for the introductory courses. In 1957 Córdoba also started offering astronomical courses towards a PhD degree in astronomy. In both places one may become a *Licenciado* (a somewhat higher degree than that of Master) and, then, a Doctor in Astronomy. Altogether, La Plata and Córdoba produce about 13 *Licenciados* and four to five PhDs a year.

In 1958 the La Plata Observatory, with the strong support of the university's rector Danilo Vucetich, decided to engage in the acquisition of a modern instrument and in extended site surveying. A 2.15 m (84 in) telescope, similar to the telescope then recently acquired by the US KITT PEAK NATIONAL OBSERVATORY, was aimed at. Early in 1971 the instrument and the figured Pyrex discs were already in La Plata but different problems delayed the continuation of the project and the dedication of the new observatory only occurred in 1986. Erected on top of one of the peaks in the El Leoncito astronomical reservation in the province of San Juan, at an elevation of 2500 m, in the Andes, this place is now a national observing facility, the CASLEO (El Leoncito Astronomical Complex), run by the National Research Council (CONICET) and the universities of La Plata,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Córdoba and San Juan. A recent addition to the instrumentation is a submillimeter solar radiotelescope (SST), a joint collaborative effort between the University of Berne, Switzerland, the Brazilian CRAAM (Center of Radioastronomy and Space Applications), IAFE and CASLEO. Soon a solar spectrograph for the ranges of 3700–4400 Å and of 5500–8000 Å will also become operational. In addition, the 60 cm Canadian reflector that had been at Cerro Las Campanas in Chile will be installed on top of another peak in the area. The auxiliary equipment for the 2.5 m telescope, named the Jorge Sahade Telescope at CASLEO's tenth anniversary celebration, consists, at present, of two stellar spectrographs, a CCD camera and a photopolarimeter belonging to the Torino Observatory.

The El Leoncito area had previously been chosen as a good astronomical site by astronomers of the University of San Juan for the installation of a Yale–Columbia double astrograph for proper-motion work, at what is now known as the Carlos U Cesco High Station. This installation is part of the University of San Juan's Félix Aguilar Observatory that was built on the outskirts of the city of San Juan and dedicated in 1953. It has carried out fundamental work with the modernized meridian circle of the Córdoba Observatory and in cooperation with the US Naval Observatory and with Russian and Chinese astronomers. At the end of 1998, Cesco's Station obtained an automated mirror coronograph from the Max-Planck-Institut-für-Aeronomie, Lindau, and a 10 cm telescope with Lyot filter for 6563 Å from the Max-Planck-Institut-für-Extraterrestriche Physik, Garching. The University of San Juan has recently established its own school of astronomy.

Other astronomical institutions in Argentina are the Argentine Radioastronomy Institute (IAR) and the Institute of Astronomy and Space Physics (IAFE), both, at present, mostly on CONICET's payroll. The IAR was established about 20 km north of La Plata because of the interest of the Province of Buenos Aires Research Council (CIC) in the matter, and dedicated on 26 March 1966, the instrumentation being two 30 m parabolic antennas provided by the Carnegie Institution of Washington, with receivers for around 21 cm wavelength. The IAFE has been formally in existence since 1971, as an outgrowth of the former National Center for Cosmic Radiation. The idea for setting up the IAFE was to have a center that would cover modern topics outside the realm of interest of the other Argentine astronomical institutions, and, in fact, early on research with balloons, space projects and high-energy astrophysics were the main subjects. At present, the IAFE's principal endeavours are centered on solar physics and cosmology, participating in the analysis of data collected with the German solar instruments at El Leoncito's Cesco Station.

Astronomy is also being done by small groups at the University of Rosario, at the University of Tucumain and at the CRICYT Institute in Mendoza, and there is a Naval Observatory entrusted with the country's official time service.

Argentina has an astronomical association (AAA) founded in 1958, at present with over 300 members, about half of them holding PhD degrees. Argentina spends about 5 million US dollars a year on astronomy, including a 2.5% participation in the Gemini Project.

In Argentina there are quite a number of amateur groups, the most important and largest being the Argentine Association Friends of Astronomy (AAAA) in Buenos Aires, founded over 70 years ago, that has had and continues to have close connections with well known professional astronomers like the late Bernhard H Dawson and José Luis Sérsic.

Brazil

Brazil is one of the South American countries where astronomy received official attention at an early date. Its Imperial Observatory was created in Rio de Janeiro on 25 October 1827, and the first director was appointed in 1845. When Brazil became a Republic, the Imperial Observatory became the National Observatory, and in the 1920s it experienced one of the most brilliant periods of its history, being equipped with a 46 cm Cooke refractor of short focal length and a Heyde zenith tube for the study of latitude variations. The difficult economic times Brazil experienced for a number of years afterwards were instrumental in damping down any possible interest in supporting astronomy. Neither the success abroad of the Brazilian scientist Mario Schönberg, who worked with S Chandrasekhar and George Gamow, nor the fact that very important foreign groups had come to the country for the 1947 total eclipse of the Sun, helped to reverse the situation. A great change took place only when Professor Abrahão de Moraes, a celestial mechanician, took over the directorship of the Institute of Astronomy and Geophysics (IAG) of the University of São Paulo in 1955. He strongly believed that the growth of astronomy in Brazil required that interested young people be sent abroad for training. Fortunately for the plan, the Fund for the Support of Research in São Paulo (FAPESP) was set up and, in 1962, Sylvio Ferraz Mello and Giorgio Giacaglia started the program that has continued steadily ever since. In carrying out such a program the traditional ties Brazilian astronomy had with French astronomy were not forgotten.

The country's financial support for astronomy has grown steadily and, at present, Brazil not only participates in the Gemini Project, also with 2.5% of the total cost, but is also the main partner (31%) in the SOAR (Southern Observatory for Astrophysical Research) Project , and pays a considerable amount of money to have the right to use ESO's (European

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Southern Observatory's) 1.5 m telescope at La Silla, in Chile. The SOAR Project instrument is a 4 m telescope to operate on the same peak as the Gemini South telescope, on top of Cerro Pachón, in Chile.

The largest telescope now in operation in Brazil is a Perkin–Elmer 1.6 m reflector located in Brasopolis, Minas Gerais, at the National Astrophysical Laboratory, together with a Zeiss telescope and a Boller and Chivens 60 cm telescope. The location was chosen after a long site survey effort led by the National Observatory, then under the active directorship of Luiz Muniz Barreto, the facility having been dedicated in 1981.

In Brazil there are several institutions engaged in astronomy teaching and/or research, namely:

- The National Observatory, in Rio de Janeiro, equipped with a 46 cm Cooke & Sons telescope and a Danjon astrolabe.
- The Institute of Astronomy and Geophysics of the São Paulo University.
- The Astrophysics Group of the Physics Institute, University of Rio Grande do Sul, Porto Alegre, MG.
- The Physics Department of the Federal University of Rio Grande do Norte, Natal, RN.
- The Astronomy Department of the Federal University of Rio de Janeiro, RJ.
- The National Institute of Space Research (INPE), São Jose dos Campos, SP.

For observing, besides the National Observatory and the National Astrophysical Observatory already mentioned, Brazil has several small observatories at different places:

- The Abrahão de Moraes Observatory at Valinhos, SP, equipped with a meridian circle with CCD.
- The Antares Observatory at Feira de Santana, BA, equipped with a Danjon astrolabe.
- Piedade at Belo Horizonte, MG, with a 60 cm Zeiss telescope equipped with a Zeiss spectrograph and a single-channel photometer with CCD.
- The VALONGO OBSERVATORY at Rio de Janeiro, RJ, equipped with a 30 cm Cooke & Sons telescope.

The observing conditions are, however, not too good and this explains why the Brazilians are participating in projects in other countries.

In addition, we have to refer to the important Itapetinga Radioastronomical Observatory in Atibaia, SP. The attempt to do radioastronomy in Brazil started in 1959 and provides a long and interesting story. The most important installation at present is located in Itapetinga and it was achieved through the help of a number of people and organizations. The main instrument is an Essco 13.7 m diameter precision telescope in a 19.4 m radome, maximized for the frequency of 22 GHz and used for frequencies between 7 MHz and 94 MHz. Pierre Kaufmann is the soul of this enterprise, which cooperates with several organizations in the US and Russia. Research is done on the Sun, on interstellar molecules, on active nuclei of galaxies, etc. As mentioned earlier, a solar telescope for submillimeter waves has been installed at CASLEO, in Argentina, as a cooperative effort between Itapetinga, the University of Berne, Switzerland, IAFE and CASLEO.

Brazilian astronomers are all members of the Brazilian Astronomical Society (SAB) that was established in 1974, and the present membership numbers about 400, nearly half of them with PhD degrees. The yearly government spending in astronomy of about 15 million US dollars covers Rio, São Paulo, the National Astrophysical Laboratory, the participation on the SOAR and Gemini Projects and the fellowship program, of some 100 fellowships a year. This is the largest support for astronomy among the South American countries, and it does account for the very large growth that Brazilian astronomy has experienced in the past few decades.

Chile

Chile is characterized, in its northern half, by extremely good astronomical skies, similar to those of the southern coast of California in the early days of the MOUNT WILSON OBSERVATORY.

Astronomical activities in Chile began in 1849 when Lieutenant James M Gilliss, of the US Navy, started carrying out observations of Mars near opposition and of Venus near its stationary phases, to determine the solar parallax. The instrument, a 16.5 cm equatorial telescope, was installed at the Cerro Santa Lucía (Santa Lucía Hill) , in Santiago. Stellar positions in the declination range of −65°—90° were also measured with a Pistor and Martins meridian circle. In 1852, at the end of Gilliss's mission, both instruments were bought by the Chilean Government and the observations were continued under Karl Wilhelm Moesta without interruption. During Moesta's directorship the observatory was moved down from the hill in 1856 and there was a cooperative undertaking with Pulkovo and Washington for the observation of the declination of Mars at its 1862 opposition.

Another American expedition came to Chile in 1903, namely the D O Mills expedition from the Lick Observatory, under William H Wright, with the plan of measuring radial velocities of southern stars with a 92 cm reflecting telescope that was installed at Cerro San Cristóbal, close to Santiago. At the end of the mission in 1909, the instrument was bought by the Catholic University at Santiago, with the idea of developing an astronomy department.

Within the long history of astronomy in Chile, we should point out that, early in the century, astronomy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

found a strong supporter in President Pedro Montt, who was in power from 1903 to 1910. During his term of office the observatory was under the active leadership of Friedrich Wilhelm Ristenpart, who also encouraged cooperation between South American observatories and the production, down to magnitude 10, of stellar charts south of declination −42°, that were missing in the *Córdoba Durchmusterung*.

In 1927, the National Observatory was transferred to the University of Chile and not much actually happened until 1950 when Federico Rutllant Alcina took over as director. He moved the observatory to its present location on Cerro Calán, 860 m above sea level and at some 14 km northeast of the center of Santiago, and things changed in such a way that the University of Chile was able to establish in due course an astronomy department. It is possible now to work for a Magister in Astronomy degree, and it is envisaged in the near future to make it possible to work for a PhD degree.

Rutllant not only undertook the moving of the observatory but also encouraged Gerard Kuiper, then director of the YERKES OBSERVATORY, to consider a joint project between the universities of Chicago, Texas and Chile to take advantage of the excellent climatic conditions of the latter country for astronomy. This initial proposal and the site survey carried by Jürgen Stock finally gave rise to the CERRO TOLOLO INTER-AMERICAN OBSERVATORY (CTIO) with the involvement of AURA (the Associated Universities for Research in Astronomy) and the financing of the National Science Foundation (NSF). CTIO was dedicated in November 1962, and was followed by the installations of the EUROPEAN SOUTHERN OBSERVATORY (ESO) at Cerro La Silla, dedicated seven years later, by those of the Carnegie Institution of Washington at Cerro Las Campanas, in 1976, and the huge ESO set-up of four 8 m telescopes at Cerro Paranal, the Very Large Telescope (VLT) Project, in the north of Chile, VLT1 having been officially dedicated in December 1998. Chilean astronomers do have access to 10% of the observing time on all the foreign telescopes on Chilean soil. Moreover, the country is also a partner in the Gemini Project (5% of the total cost), a fact that implies an additional percentage of observing time for the Chilean astronomers.

Under Rutllant and his successor, Claudio Anguita, who took over in 1964, Cerro Calán turned into an important research center from both the astrometric and the astrophysical point of view. Moreover, the access to the telescopes and to the fellowships granted by the foreign organizations involved in the different astronomical projects in Chile have opened unique possibilities for interested young people.

In 1959 the then Soviet Academy of Sciences sent to Cerro Calán a group of astrometrists headed by Pulkovo astronomer M S Zverev, that observed stellar positions to serve as astrometric standards for the southern hemisphere. Later, the Russians built an astrometric observatory at Cerro El Roble, 80 km north of Santiago, equipped with a double-meniscus Maksutov astrograph that was erected in 1962, and planned for an astrophysical extension. When President Salvador Allende was overthrown, the Russians abandoned the country and the El Roble station, together with the instruments, were transferred to the university.

At Maipú, some 30 km southwest of Santiago, there has been important radioastronomical activity, first in cooperation with the Carnegie Institution of Washington and then with the University of Florida. Radioastronomical research is still continuing but Maipú is no longer active.

As we have already mentioned, the Catholic University in Santiago, has its own astronomy department which is growing at a fast pace. At present the department offers courses that lead to the degree of *Licenciado* in astronomy and plans to soon extend the course to PhD level. At present it is possible to work on an astronomical subject for a degree in physics.

A large radioastronomical project, known as the ATACAMA LARGE MILLIMETER ARRAY (ALMA), is under consideration at the present time as a Europe–US collaborative project, but it may turn out to become a truly global astronomical undertaking. Its location will be at Chajnantor, in the Atacama Desert in the northern part of the country, at some 5000 m altitude.

In view of the extraordinary observing possibilities that exist for astronomers working in Chile, some universities, like those at Antofagasta, at Concepción and at Tarapaca, have astronomers on their teaching staffs and may offer degrees in astronomy in the future. In Santiago itself there is a very small group of astronomers at the Isaac Newton Institute.

At the moment there is no astronomical society established in Chile. The country's expenditure in astronomy was estimated to be some 3.5 million US dollars in 1999.

The impact of the erection of so many foreign observing facilities on Chilean soil with so much observing time at the disposal of Chilean astronomers is beginning to show its beneficial effects upon the development of Chilean astronomy and promises a very bright future for the field. Now, in regard to the effect upon astronomy in the rest of the South American countries, CTIO could be singled out for accepting observing proposals from their own astronomers, something that has been also true in the case of Las Campanas and, in a very few special cases, of La Silla. On the contrary, the present very large telescope projects reserve observing time for the members of their partnership and for Chile.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 2. Web pages of astronomical organizations in South America.

Country	Society	Web page
Argentina	Astr. Society	www.casleo.gov.ar/webAAA/indexaaa.htm
	CASLEO	www.casleo.secyt.gov.ar
	La Plata	www.fcaglp.unlp.edu.ar
	IAFE	www.iafe.uba.ar
	IAR	www.iar.unlp.edu.ar
	Córdoba	www.oac.uncor.edu
	San Juan	www.unsj.edu.ar
Brazil	Univ. Federal Rio de Janeiro	acd.ufrj.br/ov
	Itapetinga	www.craae.mackenzie.br
	INPE	www.das.inpe.br
	Sâo Paulo	www.iagusp.usp.br
	Astr. Society	www.iagusp.usp.br/sab
	Rio Grande do Sul	www.if.ufrgs.br/ast
	Laboratorio Nacional	www.lna.br
	Obs. Nacional	www.on.br
Chile	Catholic Univ. Santiago	www.astro.puc.cl
	CTIO	www.ctio.noao.edu/ctio.html
	Cerro Calán	www.das.uchile.cl
	ESO-Paranal	www.eso.org/paranal
	ESO-La Silla	www.ls.eso.org/lasilla
	Univ. Concepión	phys.cfm.udec.cl
	Las Campanas	www.ociw.edu/lco/
	ESO-Santiago	www.sc.eso.org/santiago
Colombia		www.observatorio.unal.edu.co
Uruguay		www.fisica.edu.uy/astronomia
Venezuela	CIDA	www.cida.ve/
	Inst. Venezolano Inv. Científicas	www.ivic.ve
	Univ. Los Andes	www.ula.ve

Uruguay

Astronomy has been present in Uruguay since the end of the eighteenth century when observations of the transit of Mercury were made in Montevideo by an expedition led by Alejandro Malaspina. Moreover, astronomy has been taught, as a subject in secondary schools, since 1889.

A small observatory equipped witth a Zeiss 20 cm *f*/15 refractor was dedicated in 1938 and Comet 1947IV was discovered with it. In 1955 a Spitz Planetarium was opened in Montevideo, thanks to the fact that the City Mayor was then an amateur astronomer.

A Chair of Astronomy was established in 1946 at the new Faculty of Humanities and Sciences of the University of Montevideo by Félix Cernuschi, a Cambridge PhD in Physics who had post-doctoral training in Zurich and at Harvard. Cernuschi, successful in setting up a Department of Astronomy in 1955, was very active in developing theoretical astronomy in Uruguay (and also in Argentina), but his efforts were impaired by political problems in both countries, and suffered an interruption during the military regime (1973–85). In the interval 1966–70, thanks to Cernuschi's endeavors, a radiointerferometer operated in the outskirts of Montevideo for the study of solar radiobursts.

In 1985 the situation experienced an improvement. A faculty of sciences was established at the University of Montevideo, and within it a physics institute with an astronomy department. A Master's degree in astronomy is offered, and the main areas of research relate to the solar system, in particular to comets, asteroids and planet formation. Since 1994 there has been a small observatory at Los Molinos, near Montevideo, equipped with a 35 cm Cassegrain telescope with a CCD camera, that is used to train students and to carry out small programs of photometry and astrometry of asteroids and comets.

Uruguay has an Astronomy Society with about 25 members, but not all of them are professional astronomers. Urugauy has a yearly astronomy budget of the order of 150 000 US dollars.

Venezuela

The history of modern Venezuelan astronomy actually starts in 1953, when the Director of the Cagigal Naval Observatory, Eduardo Röhl, with Presidential approval, ordered from Askania and Zeiss, in Germany, seven astronomical instruments—a meridian circle, a photozenith tube, a refracting telescope, a Schmidt camera, a reflecting telescope, a double astrograph, and a planetarium—that arrived in Venezuela in 1957. In order to decide what to do with the instruments, Röhl's succesor, José Abdala, asked the US National Academy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of Sciences for advice. Then a committee of seven astronomers from the US, one from Mexico, one from Argentina and the Executive Secretary of the US Academy, who was a physicist, met in Caracas, in 1962, to consider the situation. As a result, Jürgen Stock, who had carried out the site surveying for Cerro Tololo in Chile, was entrusted to find the best possible site for an observatory in Venezuela, and the compromise was found in the region of Mérida, where there was already a university. CONICIT, the recently established Venezuelan Research Council, decided, then, in 1971, to set up a National Institute of Astrophysical Research and, in 1973, a governmental decree created the Francisco J Duarte Foundation Center for Research in Astronomy (CIDA) in Mérida. The observatory is 50 km from the city, at an elevation of 3600 m.

The Naval Observatory remains in Caracas with the meridian circle, the photozenithal telescope and the transit instrument.

Astronomy is taught at the University of Los Andes, Mérida, at the University of Zulia in Maracaibo, and at the Universities in Caracas. At the Central Venezuelan University there is an important group engaged in cosmology and relativity.

The University of Los Andes has about 10 students that take astronomy as their subject, and perhaps one student graduates every year at either of the levels of graduation. In Caracas, maybe one student graduates in astronomy every three years.

Bolivia

Perhaps because of the altitude of the mountainous regions and the high temperature on the lowlands, professional astronomy in Bolivia has only started recently. The scientific or technical activities that have characterized Bolivia are those for which very large altitudes or the geographic coordinates are ideal factors, like those related to cosmic radiation research, for which Chacaltaya, at an altitude of over 5000 m, is a good example, or the installations at Patacamaya to track Russian artificial satellites. As for professional astronomy, the story starts in 1982 when the Bolivian National Academy of Sciences and the then USSR Academy of Sciences signed a cooperative agreement. As a result, the Pulkovo Observatory installed a Bolivian–Russian Astronomical Observatory, 15 km southeast of the city of Tarija, at an altitude of 2000 m, close to the boundary with Argentina. The instrumentation consisted of several telescopes, the largest one being a 60 cm Zeiss reflector. The work performed includes stellar catalogs, proper motion determinations, studies of supernova 1987A, polarimetry, etc. At present the activity continues in the same fields.

In Bolivia there are also a few small observatories that belong to educational institutions.

Colombia

The National Observatory, which was transferred to the National University of Colombia when this educational institution was created, in 1867, and indeed to its Faculty of Sciences in 1965, was built between May 1802 and August 1803. The instrumentation available is old and present-day research is mostly based on observations carried out elsewhere or on celestial mechanics problems. There have been plans by Jorge Arias de Greiff and Eduardo Brieva to build a modern observatory at a suitable site, at a certain distance from Bogotá, but the required financial support has not materialized and, in addition, the site became seriously affected by an earthquake.

Small groups interested in astronomy have existed and exist elsewhere in Colombia, but they have tended to be short-lived.

Peru

Peru is well known in astronomical circles because towards the end of the nineteenth century, in Arequipa, a city in the southern part of the country, Harvard Observatory carried out the observations for the southern extension of the *Henry Draper Catalog*. The instrumentation was moved to South Africa in 1926 because apparently the observing conditions had become poorer.

In Lima there are a few universities interested in astronomy, in particular, the Universidad Nacional Mayor de San Marcos, which is the one that participated in the IAU's Visiting Lecturers Program. There is an association of amateur astronomers in Lima that is active and has existed for nearly 40 years.

Acknowledgments

I am very much indebted to Claudio Anguita, Felipe Barrientos, Eduardo Brieva, Gloria Cassab, Alejandro Clocchiatti, Julio Fernández, Sylvio Ferraz Mello, Emilio Lapasset, Dante Minniti, Marta Rovira, Jürgen Stock, Jeanette Stock, Rodolfo Zalles Barrera and the IAU Secretariat for information that has permitted me to write the present article.

Bibliography

Arias de Greiff J 1993 *La Astronomía en Colombia* (Academia Colombiana de Ciencias Exactas, Físicas y Naturales)

Bajaja E 1985 *Rev. Mex. Astron. Astr.* **10** 21

Barbuy B, Braga J and Leister N (eds) 1994 *A Astronomia no Brasil: Depoimentos* (Sociedade Astronômica Brasileira)

Bruzual A G and Magris C G 1996 *Revista Academia Mérida* no 4

Chaudet E 1926 *La Evolución de la Astronomía (1872–1922)* (Sociedad Científica Argentina)

Keenan P C, Pinto S and Alvarez H 1985 *The Chilean National Observatory (1852–1965)* Facultad de Ciencias Exactas y Naturales, Universidad de Chile (bilingual)

Sérsic J L 1971 *Sky Telescope* **42** 347

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stock J 1981 Astronomía en Venezuela *Rev. Mexicana Astron. Astrof.* **6** 13

Jorge Sahade

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

South Pole–Aitken Basin (SPA Basin)

The largest and deepest multiringed impact structure on the Moon. It lies almost entirely on the Moon's farside, and extends for over 2500 km from the south polar region to the 135 km diameter crater Aitken, at 16.8 °S, 173.4 °E. The impact that produced the basin is estimated to have occurred around 4.1 billion years ago, near the beginning of the solar system's great cratering episode known as late heavy bombardment. At its deepest point the SPA Basin is nearly 13 km below the Moon's mean surface level. It was first revealed by the Lunar Orbiter photographic reconnaissance of the mid-1960s, but its full extent was not appreciated until the Moon was revisited by spacecraft in the 1990s. Galileo photographed it in 1992 and showed it to have a darker coloration than its surroundings. Its size was established by the Clementine orbiter in 1994, which also found that the basin's floor is rich in titanium and iron oxides. It has been suggested that this compositional anomaly indicates that the impact reached a depth of 120 km, penetrating the crustal rocks completely and excavating metal-rich material from the mantle.

See also: Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spörer, Gustav F W (1822–95)

German astronomer, investigated sunspots. Discovered the progressive drift in average latitude of sunspots towards the solar equator over a solar cycle (*Spörer's law*) and the scarcity of sunspots between 1645 and 1715 (the MAUNDER minimum).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space and Time References: Concepts

Space and time are generally considered as separate unrelated concepts. This is deeply rooted in our common sense and is basically postulated in Newtonian mechanics in which they play completely separated roles. The Theory of Relativity introduced a deep linkage between these concepts and one has to modify one's way of thinking whenever it is not possible to ignore this theory, particularly in its general form. But for many cases in astrometry, it is legitimate to keep them separated.

In order to specify quantitatively the position of a celestial body in astronomy, one uses a system of spherical coordinates. The direction is defined by two angles: for instance right ascension along the celestial equator and declination along the celestial meridian. A third parameter, the distance, is added whenever appropriate, whether it is measured directly by the duration of transit of an electromagnetic wave or measuring the parallaxes of stars (the parallactic effect is the apparent angular displacement of the direction of a star when the Earth circles the Sun; see PARALLAXES). Within the solar system, rectangular (Cartesian) coordinates are used for dynamical studies, but astrometric work is done always in spherical coordinates. Similarly, on the Earth, one uses longitudes, latitudes and heights, or three rectangular geocentric coordinates.

The triad that defines the coordinate system cannot be materialized. It is a virtual ensemble called a *reference system*, whether celestial or terrestrial. Among all possible systems, one chooses a *conventional reference system*. Then its realization consists in providing an operational method to obtain the coordinates of a point on the sky or on the Earth. In practice, the system is realized by providing the coordinates in the reference system of a certain number of objects, together with the variations of their positions with time. Such an ensemble, which can either be a catalog or ephemerides of objects, is called a *reference frame*. It may be supplemented by secondary catalogs with many more objects linked to the reference frame.

Celestial reference systems

In a celestial reference system, the coordinates of objects without detectable proper motions (because they are very far) must remain fixed. This is the condition that is used to define a *kinematic* celestial reference frame. This is the case of the present system chosen in 1995 and called the *International Celestial Reference System* (ICRS) defined at the barycenter of the solar system. Its definition states that, in the mean, an ensemble of remote extragalactic radio sources does not present any residual rotation. It became in 1998 the official celestial reference system for all uses in astronomy and celestial mechanics.

In this latter case, applications to the motions in the solar system require the addition of the third component so as to have access to the three-dimensional space. With the present level of our knowledge of the motions in the solar system, this has to be done in the framework of the theory of General Relativity. The level at which this has to be applied was specified in 1991 by the International Astronomical Union. The status is that the four space-time coordinates $(x^0 = ct, x^1, x^2, x^3)$ must be selected in such a way that in a coordinate system centered at the barycenter of the solar system, the squared interval ds^2 is expressed in the form:

$$ds^2 = -c^2 d\tau^2 = -(1 - 2U/c^2)(dx^0)^2 + (1 + 2U/c^2) \times [(dx^1)^2 + (dx^2)^2 + (dx^3)^2]$$

where c is the velocity of light, τ the proper time, and U is the sum of the gravitational potentials of the ensemble of masses, and of a tidal potential generated by bodies external to this ensemble, the latter potential vanishing at the barycenter. In the coordinate x^0, t is the coordinate time. The system of coordinates so defined is such that the proper local triad (x, y, z) at the barycenter of the solar system coincides with the triad of the ICRS. Similarly, one can define a geocentric reference coordinate system used to describe the motion of artificial satellites.

Another way to force a celestial system to be fixed is based on the fact that if the motion of planets is referred to a rotating or accelerating system of coordinates, the bodies are accelerated. This is detectable and hence can be suppressed by choosing an adequate reference. This *dynamical* definition of celestial reference systems requires a Newtonian space. It was used until recently in the absence of observations allowing a kinematic definition, but its definition in the framework of General Relativity is not well understood, and in any case cannot be extended to the Galaxy and even less so to the universe.

Celestial reference frames

The realization of the ICRS was made using observations of positions of extragalactic radio-sources by VLBI (see EXTRAGALACTIC REFERENCE FRAMES). About one-third (the best observed) of the 608 sources in the catalog constitutes the *International Celestial Reference Frame* (ICRF). These objects are not observable optically with a sufficient accuracy to serve as fiducial points in optical astrometry. So a secondary star catalog, the Hipparcos Catalogue, was linked to the ICRF (see HIPPARCOS AND TYCHO CATALOGUES) and is the realization of the ICRS in optical wavelengths. In the case of a dynamical definition of the celestial reference frame, the observed positions of planets are the moving fiducial points to which a fundamental catalog of star positions and proper motions is referred (see DYNAMICAL REFERENCE FRAMES).

Terrestrial reference system and frame

The concept of a terrestrial reference system is analogous to its celestial counterpart. It is represented by a set of fiducial points on the surface of the Earth. The system is so defined that, despite the motions of the tectonic plates on which the points are located, their ensemble has no global residual rotation. This is how the *International*

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Terrestrial Reference System (ITRS) is defined. It is realized by a set of positions and motions of geodesic stations (see TERRESTRIAL COORDINATE SYSTEMS AND FRAMES), called the *International Terrestrial Reference Frame* (ITRF).

Time scales

To be complete, a reference system of coordinates must include the time dimension. The equivalent of a coordinate for time is a time scale. Historically, a time scale was conceived as the independent variable of the dynamics in the solar system. At present, the basis of time is derived from the fact that the frequency of an atomic transition in given conditions is a constant. A time scale is obtained by integrating basic periods produced by an atomic frequency standard (see TIME SCALES: ATOMIC). It is a (local) proper time in the theory of General Relativity.

To extend the scale outside the laboratory, one has to apply the relativistic ds^2 and to define precisely the coordinate system by the metric form to which it belongs. The time scale that results is called a coordinate time. This is the case of the *Terrestrial Time* (TT), obtained by the addition of seconds of the International System of Units on the geoid (see TIME SCALES: TERRESTRIAL). The Geocentric Coordinate Time (TCG) is attached to the geocentric coordinate reference system and differs from TT by a constant rate. It is used for the dynamical studies and monitoring of artificial Earth satellites. The Barycentric Coordinate Time (TCB), which is to be used in studying the motions in the solar system, differs from TT not only by a constant rate, but also by some periodic terms.

Bibliography

Arias E F, Charlot P, Feissel M and Lestrade J-F 1995 The extragalactic reference system of the International Earth Rotation Service (IERS) *Astron. Astrophys.* **303** 604–8

Bergeron J (ed) 1992 Resolution A4: Recommendation from the Working Group on Reference Systems *Proc. 21st General Assembly of the IAU* **XXI B** 41–63

Kovalevsky J, Mueller I I and Kolaczek B 1989 *Reference Frames in Astronomy and Geophysiscs* (Dordrecht: Kluwer)

Ma C, Arias E F, Eubanks T M *et al* 1998 The international celestial reference frame as realized by the very long baseline interferometry *Astron. J.* **116** 516–46

Jean Kovalevsky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Geodesy

Geodesy is the science studying the size and the figure of the Earth including the determination of the Earth's gravitational field. Geodetic astronomy is that part of astronomy dealing with the definition and realization of a terrestrial and a celestial reference frame (see TERRESTRIAL COORDINATE SYSTEMS AND FRAMES). By space geodesy we mean, then, those aspects of geodesy and geodetic astronomy studied by using natural or artificial celestial bodies as observed objects or as observing platforms. In the older literature the term cosmic geodesy is sometimes used as a synonym. Space geodesy is thus defined through the observation techniques, below referred to as space geodetic techniques or methods.

Space geodesy evolved rapidly in the second half of the twentieth century. The space age was initiated by the launch of the first artificial satellite, Sputnik I, on 4 October of the International Geophysical Year 1957. In the space age it became possible to deploy and use artificial satellites either to study the size and figure of the Earth from space or to observe them as targets from the surface of the Earth. The use of artificial Earth satellites for geodetic purposes is also referred to as satellite geodesy. The second essential development consists of the technique of very long baseline interferometry (VLBI) as a new tool to realize an extraordinarily accurate and stable inertial reference system and to monitor Earth rotation using quasars (see EXTRAGALACTIC REFERENCE FRAMES).

Today, space geodetic techniques are the primary tools to study size, figure and deformation of the Earth, and its motion as a finite body in the inertial reference system (see SPACE AND TIME REFERENCES: CONCEPTS). Space geodetic techniques are thus the fundamental tools for geodesy, geodetic astronomy and geodynamics.

Space geodetic observations contain information about the relative positions and/or motions of the observed object and the observer. Let $\boldsymbol{GO}(t_i)$ be the position of the observer at some time t_i in the terrestrial reference frame and $\boldsymbol{GS}(t_i)$, the position of the satellite in an inertial reference frame. An observation $O(t_i)$ made at that time and corrected for various perturbing effects described below, can be expressed mathematically as

$$O(t_i) = F(\boldsymbol{GO}(t_i), \mathrm{d}\boldsymbol{GO}(t_i)/\mathrm{d}t, \boldsymbol{GS}(t_i), \mathrm{d}\boldsymbol{GS}(t_i)/\mathrm{d}t)$$

or usually only of some of the components of these vectors. For instance:

- several observations of a satellite in a known position at the same time from several observatories provide equations linking the positions of these observatories;
- several observations of the same satellite performed at the same observatory provide equations containing information on the trajectory of the satellite, and consequently also on the forces acting on it;
- observations of several satellites of known positions by a single station provide the geocentric position of the observer.

In scientific space geodesy, several 'geodetic' satellites are observed many times from many stations. The set of equations so obtained is solved for the satellite orbits (Tapley 1989), the gravitational field of the Earth (Reigber 1989), the relative positions of the observing stations or any of these quantities if others are known. Because the positions of the stations and of the satellites are in different reference frames, space geodetic observations also contain information concerning the transformation between the terrestrial and the inertial systems. The Earth orientation parameters, i.e. polar motion, UT1, precession and nutation (see EARTH'S ROTATION: THEORY, POLAR MOTION AND LENGTH OF DAY) define this transformation.

The role of the Earth's atmosphere

In space geodesy the signals of the observed or observing celestial bodies have to cross the Earth's atmosphere. This changes the paths and the travel times of the signals. These are referred to as *refraction effects*. Refraction is usually considered a nuisance in astronomy, geodesy and geodynamics—as a matter of fact it is the motivation for many spaceborne experiments related to this field of science. In recent years refraction effects have been considered and understood as a primary source of information for atmospheric science and are monitored through space geodesy methods. Let us point out that the same signals and space geodetic analysis methods are used to study the Earth's atmosphere as for geodetic and geodynamic purposes. Interdisciplinary studies and projects have become important aspects of modern space geodesy.

Whether the atmosphere-related signal is useful depends on the wavelengths of the analysed signals. If we measure, for example, distances or distance differences to satellites using optical signals, refraction effects may be computed with subcentimeter accuracy using pressure, temperature and humidity registrations at the observing sites. We may therefore conclude that laser ranging is not capable of contributing to atmosphere monitoring. This fact may also be formulated in a positive way: laser observations are well suited for calibrating other techniques, which are more prone to atmospheric effects.

For microwave techniques (Doppler, GPS, VLBI) we have to distinguish between ionospheric refraction stemming from the ionized upper part of the atmosphere (extending up to about 1500 km) and tropospheric refraction, stemming from the lower, neutral layers of the atmosphere. Ionospheric refraction is wavelength-dependent and may be (almost completely) eliminated if coherent signals are sent through the atmosphere on different carrier wavelengths. In the VLBI technique this is achieved by observing the quasars in different wavelengths, in the Doppler or GPS technique the same is achieved by using two different wavelengths for signal transmission.

For microwave techniques tropospheric refraction is the ultimate accuracy-limiting component in the error budget. As opposed to range observations in the optical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

band, we have to take into account the so-called 'wet component' of tropospheric refraction, which is highly variable in time and space. This fact forces analysts using microwave observations to introduce station- and time-specific parameters (or to model the effect as a random process). It allows, on the other hand, analysts to determine the water vapor content above an observatory with high accuracy and high temporal resolution (Bevis *et al* 1992).

Optical period

For centuries, optical (astrometric) observations were the only type of observation available in astronomy. In the pre-space era a series of astrometric instruments was used for the purpose of defining a terrestrial reference frame and for monitoring the rotation of the Earth. The photographic zenith tube and the Danjon astrolabe were probably the most advanced of these instruments. They were widely used by observatories contributing to the International Polar Motion Service (IPMS) and the Bureau International de l'Heure (BIH) to determine the geographic latitude of a station with a precision of about 10–40 mas (milliarcseconds) in one night. We refer to Moritz and Mueller (1988) for more information.

Optical observations were made by the first generation of artificial Earth satellites, like Sputnik 2 and Explorer 1. The balloon satellites Echo 1 and 2 and PAGEOS (passive geodetic satellite), which could be seen by the naked eye, were observed by a worldwide optical tracking network. These satellites were (supposedly) spherical, and consisted of layers of aluminized mylar foil, and, thanks to their brightness, their tracks could easily be photographed against the star background. It was not trivial to assign time-tags to specific points of the track. Much more suitable from this point of view, although more difficult to track, were smaller satellites like Geos 1 (Explorer 29) and Geos 2 (Explorer 36) equipped with flash lamps allowing for tens of thousands of high-precision optical observations. Obviously, quasi-simultaneity of observations from different sites was possible.

Fascinating results came out of this first phase of satellite geodesy. The geodetic datums on different continents could be related to the geocenter and thus to each other with an accuracy of about 5 m. The first reliable coefficients of the gravity field (spherical expansion up to degree and order 12–15) could be also derived.

The astrometric technique, when applied to artificial satellites in the 1960s and 1970s, had serious disadvantages. The star catalogs were not of sufficiently good quality and the processing time (time between observation and availability of results) was of the order of a few weeks in the best case. This, and the advent of new observation techniques promising higher accuracy, actually ruled out astrometric techniques for a number of important applications. The optical technique no longer played a significant role in space geodesy after about 1975.

In view of newly developed observation techniques (CCD (CHARGE COUPLED DEVICE) techniques) and much better star catalogs based on astrometry missions (e.g. the HIPPARCOS mission) it may well be that optical observations will again play a role in space geodesy in the future.

Doppler period

The US Navy Navigation Satellite System (NNSS), also called the Transit system after the survey transit instrument, had a significant impact on the development of space geodesy. It proved that a system based on the measurement of the Doppler shift of a signal generated by a stable oscillator on board a satellite could be used for relative positioning with remarkably high accuracies (0.1–0.5 m relative, about 1 m geocentric). The satellites sent information on two carrier frequencies (400 MHz and 150 MHz) near the microwave band.

The two frequencies allowed for a compensation of ionospheric refraction. Rather small receivers connected to omnidirectional antennas made the technique well suited to establish regional or even global geodetic networks. Observation periods of a few days were required to obtain the above mentioned accuracy.

The NNSS satellites were in polar, almost circular, orbits about 1100 km above the Earth's surface. Only one satellite at a time could be observed by one receiver. As opposed to astrometry the Doppler technique is weather independent. Until a significant part of the Global Positioning System (GPS) was deployed (around 1990) the NNSS played a significant role in space geodesy. Many Doppler campaigns were organized to establish local, regional or global networks. With the full deployment of the GPS in the 1990s the geodetic community eventually lost interest in the Doppler system. The Transit system was shut down as a positioning system in December 1996 but continued operating as an ionospheric monitoring tool. For more information concerning the Doppler system we refer to Kouba (1983).

Satellite and lunar laser ranging (SLR and LLR)

'Laser' stands for light amplification through stimulated emission of radiation. The laser technique, developed in the 1950s, is able to generate highly energetic short light pulses (of a few tens of picoseconds (ps) (1 ps $= 10^{-12}$ s)). These pulses are sent out by a conventional astronomical telescope, travel to the satellite (or the Moon), are reflected by special corner cubes (comparable to the rear reflectors of bicycles) on the satellite (Moon) back to the telescope, where they are detected. The measurement is the travel time Δt of the laser pulse from the telescope to the satellite and back to the telescope. Apart from refraction this light travel time, after multiplication by the speed of light c in a vacuum, equals twice the distance ρ_r^s between satellite and telescope at the time the light pulse is reflected from the satellite $\rho_r^s \approx \Delta t \times c/2$. Today's satellite laser ranging (SLR) technique is used to determine the 'true' distances between observatories and satellites with an accuracy of a few millimeters and, if required, with a high repetition rate (several times per second).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

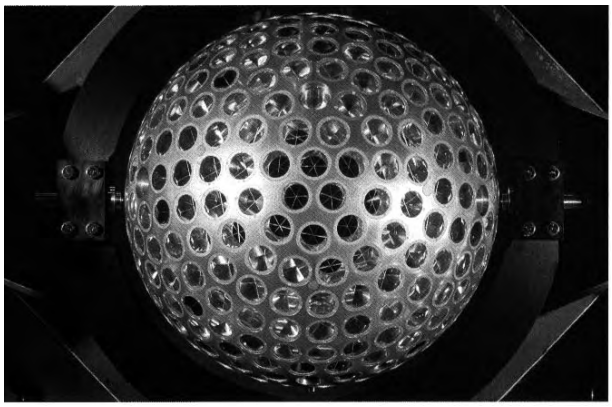


Figure 1. The Lageos II spacecraft.

SLR techniques may be used for every satellite equipped with corner cubes. Figure 1 shows Lageos II, a typical SLR-dedicated satellite which was launched in 1992. Lageos II is a spherical satellite with a diameter of 0.6 m and a weight of 405 kg. 426 corner cubes are inlaid in its surface. Lageos II is a close relative of Lageos I, which was launched in 1976. The two Lageos satellites are in stable, almost circular orbits about 6000 km above the surface of the Earth.

The two Lageos satellites are the primary scientific tracking targets for the International Laser Ranging Service (ILRS). The two satellites have contributed in a significant way to the determination of the Earth's gravity field. Many more targets are regularly observed by the ILRS. Some, like the French low-orbiting satellite Starlette, with a diameter of 24 cm, are similar in design to the Lageos satellites and serve a similar purpose. For others SLR is just the primary or backup technique for precise orbit determination.

With the exception of UT1, the SLR technique is capable of determining all parameters of geodetic interest (station coordinates and motion, Earth rotation parameters, gravity field). The unique and most valuable contributions lie in the determination of the Earth's (variable) gravity field, in the determination of the geocenter (i.e. the location of the polyhedron formed by connecting the SLR stations with respect to the geocenter), and in calibrating geodetic microwave techniques.

From the technical point of view there is no major difference between SLR and LLR (lunar laser ranging): light travel times are measured from the observatory to one of the laser reflectors deployed on the Moon by the Apollo space missions or the Russian unmanned Lunokhod missions. The scientific impact of LLR is significant. LLR was capable of measuring directly the secular increase of the Earth–Moon distance (3.8 cm per year), an effect which is in turn coupled with the deceleration of the angular velocity of Earth's rotation. Also, LLR is well suited to the evaluation of gravitational theories (see REFERENCE FRAMES AND TIMESCALES IN GENERAL RELATIVITY).

Very long baseline interferometry

Very long baseline interferometry (VLBI) is the only non-satellite geodetic technique contributing data to the International Earth Rotation Service (IERS). Its main features are discussed in the article on EXTRAGALACTIC REFERENCE FRAMES.

Its unique and fundamental contribution to geodesy and astronomy consists of the realization of the inertial reference system and in the maintenance of the long-term stability of the transformation between the celestial and terrestrial reference frame.

The ICRS (International Celestial Reference System) was defined by the International Earth Rotation Service (Arias *et al* 1995). It was adopted by the International Astronomical Union (IAU) as the primary celestial reference system replacing the optical predecessors.

An accurate and stable celestial reference frame is a prerequisite for a terrestrial reference system. In this sense VLBI plays a decisive role in the definition of the terrestrial reference system, and in establishing the transformation between the two systems. In particular, VLBI is the only technique providing the difference UT1 − UTC, i.e. the difference between Earth rotation time and atomic time with state-of-the-art accuracy and excellent long-term stability. Also, VLBI is the only technique capable of determining precession and nutation with an angular resolution below the milliarcsecond level.

The observation and analysis aspects are today coordinated by the IVS, the International VLBI Service (see table 1).

The global positioning system (GPS)

GPS is probably the best known space geodesy technique today. The system has an impact on science and society reaching far beyond space geodesy. GPS has revolutionized surveying, timing, car and aircraft navigation. Millions of hand-held receivers are in use today. Spaceborne applications of the GPS have a deep impact on geodesy and atmospheric sciences.

GPS is a navigation system allowing for instantaneous, real-time, 'absolute' positioning on or near the surface of the Earth with an accuracy of a few meters. An unlimited number of users may use the system simultaneously. 'Absolute' means that the estimated coordinates may be established using only one receiver and that they refer to a geocentric Earth-fixed coordinate system. This coordinate system, the WGS-84 (World Geodetic System), is today aligned with sub-meter accuracy to the ITRF, the International Terrestrial Reference Frame maintained by the IERS.

The space segment of GPS nominally consists of 24 satellites (21 operational satellites plus three active spares). The satellites are in almost circular orbits distributed in six planes approximately 20 000 km above the Earth's surface. These planes are separated by 60° on the equator and inclined by 55° with respect to the equator. The revolution period is half a sidereal day (11h 58m), which means that for a given location on the Earth's surface the satellite

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Space geodetic services.

Acronym/Name	Mission	Internet address
CSTG/Commission on International Coordination of Space Techniques	Coordinate space geodetic organizations, organize projects	
IERS/International Earth Rotation Service	Establish and maintain celestial and terrestrial reference frame, generate combined Earth orientation parameter series	http://hpiers.obspm.fr
IGS/International GPS Service	Make available GPS data from its global network, producing and disseminating high-accuracy GPS orbits, Earth rotation parameters, station coordinates, atmospheric information, etc	http://igscb.jpl.nasa.gov
ILRS/International Laser Ranging Service	Collect, archive and distribute SLR and LLR datasets. Generate scientific and operational products	http://ilrs.gsfc.nasa.gov
IVS/International VLBI Service for Geodesy and Astrometry	Operate or support VLBI programs. Organize geodetic, astrometric, geophysical research and operational activities	http://ivscc.gsfc.nasa.gov

Figure 2. Block II satellite.

constellation above horizon repeats itself after one sidereal day (solar day minus four minutes). Figure 2 shows the Block II satellite. This is the satellite with which the first full GPS generation was built. We distinguish the main body of the satellite with the antenna array pointing to the center of the Earth and the solar panels. The attitude is maintained by momentum wheels, which have to guarantee that the antenna array is always pointing to the center of the Earth and that the solar panel axes are perpendicular to the Sun–satellite direction. The satellite is capable of rotating the solar panels into a position perpendicular to this direction.

Each satellite broadcasts on two carrier frequencies L1 and L2 of 1.575 42 GHz and 1.227 60 GHz respectively, corresponding to wavelengths of $\lambda_1 \approx 19$ cm and $\lambda_2 \approx 24$ cm. Two types of codes are sent out, allowing users to reconstruct the so-called pseudorange p_r^s (definition in equation (1)).

The pseudorange p_r^s and the geometric distance ρ_r^s between the satellite at signal emission time and the receiver at signal reception time are related through

$$p_r^s = \rho_r^s + c(\Delta t_r - \Delta t^s) + \Delta\rho_{\text{atm}} \qquad (1)$$

where Δt^s is the error of the satellite clock with respect to the 'true' (or system) time, Δt_r is the receiver clock error. $\Delta\rho_{\text{atm}}$ is the correction of the light travel time due to the atmosphere (sum of ionospheric and tropospheric refractions).

In addition, and among other important information, broadcast orbits allowing computation of the satellite position at emission time and satellite clock corrections mitigating the satellite synchronization error Δt^s w.r.t. the 'true' or system time are sent out continuously using the phase modulation technique on L1 and L2.

Two kinds of code have to be distinguished, the so-called C/A code (coarse acquisition code) allowing an accuracy of about 3 m, and the P code (precision code) allowing an accuracy of about 0.3 m. Modern digital receivers show a much better performance. The P code is currently only available to privileged (i.e. US Department of Defense authorized) users. The C/A code is transmitted on L1, an encrypted version of the P code on both carriers. The broadcast orbits usually have an accuracy of about 3 m. Prior to 1 May 2000 the satellite clock information was made available to non-authorized users only with moderate accuracy, limiting the real-time 'absolute' positioning accuracy to about 100 m. On 1 May 2000 the Press Secretary of the White House issued the 'Statement by the President regarding the United States' decision to stop degrading Global Positioning Accuracy'. In consequence, since 2 May, 4:00 h UT (Universal Time), the satellite clock information has been made available to all GPS users without restrictions. This implies in particular that real-time 'absolute' positioning using dual-band receivers is now accurate to usually better than about 5 m (rms error) in all three coordinates. 'Absolute'

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

positioning with L1 receivers (based on the C/A code) is limited by ionospheric refraction but still is usually better than 10 m in all coordinates.

Let us briefly address the principles of absolute GPS positioning and navigation: a GPS receiver is simultaneously observing several satellites (ideally all that are above the horizon). Using broadcast orbits to compute the satellite positions in the Earth-fixed system, standard atmospheric models to account for refraction, and the broadcast clock information to adjust the satellite clocks to system time, we are left with only four unknowns in equation (1), namely the three coordinates of the receiver position and the receiver clock error Δt_r—provided we consider only simultaneous observations. It is thus necessary and sufficient that a receiver tracks four satellites simultaneously in order to compute an instantaneous position (and a receiver clock error) with an accuracy of a few meters or about 100 m for users having no access to P code respectively. The GPS constellation was designed, as a matter of fact, to ensure that (to the extent possible) four or more satellites are available all the time at each location on the surface of the Earth.

For scientific purposes phase observation plays a decisive role. It is in essence identical to the so-called accumulated Doppler observation of the NNSS and it is closely related to the GPS code observation as well. From the mathematical point of view, there are 'only' two essential differences between phase and code, namely the much higher measurement accuracy of phase (millimeters rather than meters), and an additional unknown, the initial phase ambiguity parameter N_r^s, per satellite pass. All GPS receivers used for high-accuracy geodetic applications record the phase observations on both carriers L1 and L2, in addition to the code.

The phase observations yield local GPS networks with millimeter accuracy, regional and global networks with about centimeter accuracy. This is only possible if precise satellite orbit and clock information, such as that generated by the International GPS Service (IGS), is available. Figure 3 shows the IGS network as of October 1998.

Over 200 IGS sites, distributed all over the globe, permanently observe all satellites in view, transmit their observations (at least) on a daily basis to IGS data centers.

The data are then analysed by IGS analysis centers, which deliver rapid and final products. Rapid IGS products are available with a delay of about one day, final products with a delay of about eleven days. Daily products include satellite orbits with an accuracy of about 0.05–0.1 m, satellite clocks with an accuracy of about 0.3 ns, daily values of polar motion components accurate to about 0.1 mas (milliarcseconds), corresponding to 3 mm on the Earth's surface, and length of day (lod) estimates with an accuracy of about 30 μs d^{-1}. These products are essential contributions to the monitoring of the EARTH'S ROTATION.

In addition the IGS analysis centers perform weekly global coordinate solutions of their portion of the IGS network. These results are used, together with the results of the other space techniques, for the establishment of the International Terrestrial Reference Frame.

The IGS products (orbits, Earth rotation parameters, coordinates and velocities of IGS stations) are used as known *a priori* information to establish regional networks for crustal deformation studies (e.g. the Californian SCIGN (http://www.scign.org), the Japanese 1000-receiver network for Earthquake monitoring) or for regional reference frame establishment and maintenance (e.g. the EUREF network http://www.oma.be:/KSB-ORB/EUREF/eurefhome.html, or the South American SIRGAS network http://www.dgfi-badw-muenchen.de/gps/sirgas.html).

More and more, the IGS network is used for purposes other than space geodesy. Let us mention that the IGS network has been enhanced to include time and frequency transfer and that it is able to monitor the ionosphere. For more information concerning the IGS and its interdisciplinary impact we refer to Beutler *et al* (1999).

From the point of view of space geodesy, GPS is a 'work horse' with important contributions to the establishment and maintenance of a dense terrestrial reference frame, and which provides Earth rotation parameters with a high time resolution. It should not be forgotten, that the GPS—like every satellite geodetic method—is not able to maintain a long-term stability of UT1 or of precession and nutation. Moreover, despite the fact that GPS is a satellite geodetic technique, it is not well suited to determine the Earth's gravity field or the motion of its station polyhedron w.r.t. the geocenter. The height of the GPS satellites is one of the limiting factors.

For more information concerning GPS as a tool for geodesy and geodynamics we refer to Teunissen and Kleusberg (1998).

Other satellite microwave techniques

The Russian GLONASS (Global Navigation Satellite System) is so closely related to the GPS that there are a number of combined GPS and GLONASS receivers available. These receivers were used in the first global GLONASS tracking and analysis campaign, the IGEX-98 (International GLONASS Experiment 1998). The experiment revealed that a combined analysis of GPS and GLONASS is very promising for science and navigation.

The French DORIS system (Doppler Orbitography by Radiopositioning Integrated on Satellite) has proved to be a very powerful tool for orbit determination. It is one of the orbit determination system used in the TOPEX/Poseidon mission (see below). Also, DORIS possesses a very well designed ground tracking network. This is one reason why DORIS was accepted as an official space technique by the IERS (see table 1).

The German PRARE (Precise Range and Range-rate Equipment) system may be viewed as the German counterpart of the DORIS system. It is used as an orbit determination tool on the European Space Agency's ERS-2 (Earth Remote Sensing) spacecraft.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

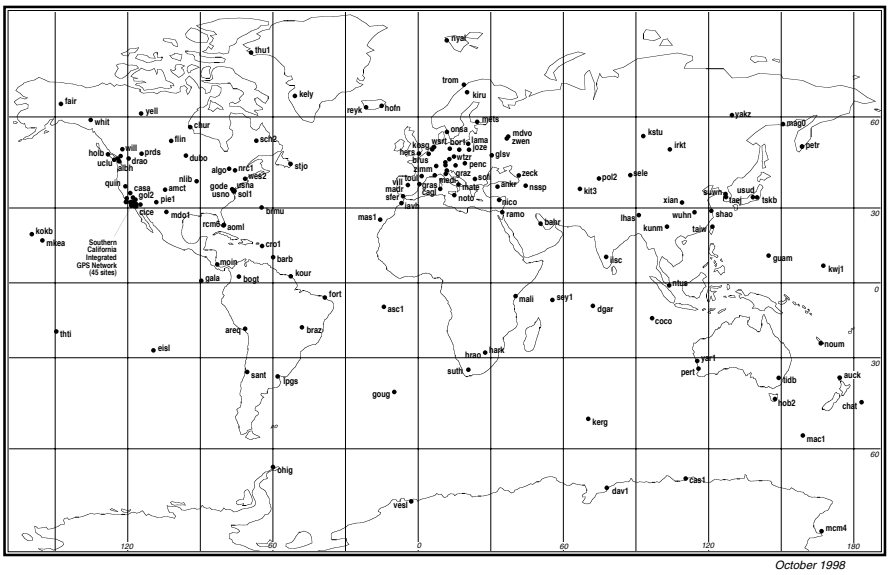


Figure 3. The international GPS service network.

Satellite missions

There have been many satellite missions in the past and there will be more in the future in which the satellite is used as an observing platform to study aspects of the Earth relevant to geodesy and geodynamics. Let us mention in particular that altimetry missions have significantly improved our knowledge of the sea surface topography, ocean currents, tidal motions of the oceans, etc.

Figure 4 shows the TOPEX/Poseidon spacecraft. The mission is a combined US and French altimetry mission. It is actually the first mission which was specially designed to investigate ocean currents. One entire volume of the *Journal of Geophysical Research* was devoted to this mission (JGR, 1994). For space geodesy the TOPEX/Poseidon mission was a kind of 'Rosetta stone mission' because its orbit was determined using three independent systems, the French DORIS system, SLR tracking and the GPS. All three systems proved their capability. The radial component of the orbit (which is of crucial importance for altimetry missions) could be established with an accuracy of a few centimeters. Let us mention that TOPEX/Poseidon was neither the first nor will it be the last altimetry mission.

For geodesy, geodynamics and atmospheric physics the upcoming missions CHAMP (Challenging Mini-Satellite Payload for Geophysical Research and Application, German mission), GRACE (Gravity Recovery and Climate Experiment, US/German mission) and GOCE (Gravity field and Ocean Current Explorer, ESA mission) are fascinating. It is expected that our knowledge of the gravity field (using spaceborne GPS receivers, accelerometers or gradiometers) to measure the non-gravitational forces with respect to gravity gradients will significantly increase through such missions.

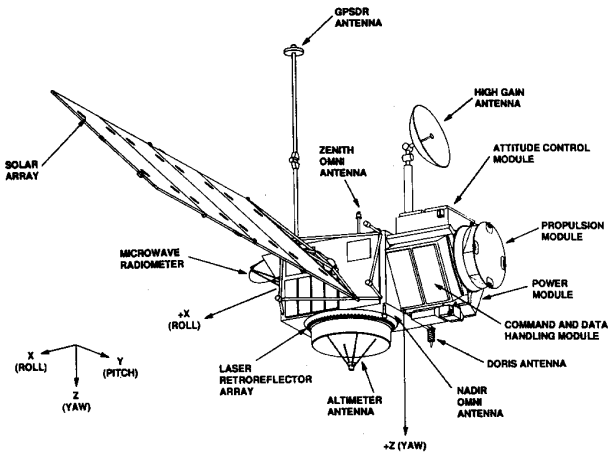


Figure 4. The TOPEX/Poseidon spacecraft.

Also, CHAMP, GRACE and GOCE are able to produce atmosphere profiles using the occultation method: the signal (phase and code) of a GPS satellite is monitored by a spaceborne GPS receiver on a low-Earth orbiter (LEO) during the time period in which the line of sight LEO–GPS satellite scans through the Earth's atmosphere. These developments support our initial statement that interdisciplinary aspects are becoming more and more important in space geodesy.

Organizations

Table 1 gives an overview of the institutions relevant for space geodesy.

They are all IAG (International Association of Geodesy) services. The IERS and the IVS are in addition

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

IAU services (International Astronomical Union). The IERS and the IGS are members of FAGS (Federation of Astronomical and Geodesical Data Analysis Services).

IGS, ILRS and IVS are technique-specific services. The IERS is a multitechnique service. It was established in 1988 as successor to the International Polar Motion Service (IPMS) and the Earth rotation branch of the Bureau International de l'Heure (BIH). The IERS products are based on the products of the technique-specific services.

CSTG is a commission of IAG and a subcommission of COSPAR (Commission on Space Research). It has a coordinating function within space geodesy. During 1995–9 it was, for example, responsible for creating the ILRS and the IVS, and it organized the first global GLONASS experiment IGEX-98.

More information about these services may be found at the internet addresses in table 1.

Bibliography

Arias E F, Charlot P, Feissel M and Lestrade J-F 1995 The extragalctic reference system of the International Earth Rotation Service, ICRS *Astron. Astrophys.* **119** 604–8

Beutler G, Rothacher M, Schaer S, Springer T A, Kouba J and Neilan R E 1999 The International GPS Service (IGS): an interdisciplinary service in support of Earth sciences *Adv. Space Res.* **23** 631–53

Bevis M, Businger S, Herring T A, Rocken C, Anthes R A and Ware R H 1992 GPS meteorology: remote sensing of atmospheric water vapor using the Global Positioning System *J. Geophys. Res.* **97** 75–94

JGR 1994 *J. Geophys. Res.* **99**

Kouba J 1983 A review of geodetic and geodynamic satellite Doppler positioning *Rev. Geophys. Space Phys.* **21** 27–40

Moritz H and Mueller I I 1988 *Earth Rotation, Theory and Observation* (New York: Ungar)

Reigber C 1989 *Theory of Satellite Geodesy and Gravity Field Determination* ed F Sansò and R Rummel (Berlin: Springer) pp 197–234

Seeber G 1993 *Satellite Geodesy* (Berlin: de Gruyter)

Tapley B D 1989 *Theory of Satellite Geodesy and Gravity Field Determination* ed F Sansò and R Rummel (Berlin: Springer) pp 235–60

Teunissen P J G and Kleusberg A (ed) 1998 *GPS for Geodesy* 2nd edn (Berlin: Springer)

G Beutler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Infrared Telescope Facility (SIRTF)

Fourth of NASA's 'Great Observatories', due for launch in 2001–2. Planned as a 1 metre, cryogenically cooled space telescope for infrared astronomy. Can be serviced by the Shuttle or at a manned space station. Instruments will provide high sensitivity and spatial resolution. Designed to study circumstellar dust clouds, brown dwarfs, infrared galaxies and quasars. Instruments include a wide field and high resolution camera covering the 2–30 micron region; an imaging photometer covering 3–700 microns, and a spectrometer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Instrumentation: Astrometry

ASTROMETRY is the measurement of accurate positions of stars or other celestial objects, and is the basis of many of the most fundamental measurements in astronomy. Precision astrometry makes possible the determination of distance to stars via annual PARALLAX. The HIPPARCOS astrometry satellite, launched in 1989 by the European Space Agency, was the first mission designed for precision astrometry.

Precision astrometry opens up many areas of astronomy and astrophysics, by providing unambiguous distances to stars out to large distances in the Galaxy. Accurate distances enable accurate measurement of the absolute luminosity of stars, which has applications for fundamental stellar astrophysics and also in firmly calibrating the stars which form the basis of the UNIVERSAL DISTANCE SCALE. An important goal from space-based astrometry is the detection of planets orbiting nearby stars, using the 'wobble' in the position of the parent star due to the gravitational pull of the planet. Detection of planets as small as a few Earth masses should be possible.

Astrometry with ground-based instruments is fundamentally limited by instabilities in the Earth's atmosphere. Space-based instruments overcome this limitation, and allow a much wider range of astronomical uses for precision astrometry, by attaining microarcsecond levels of accuracy. Several new missions have been proposed for astrometry from space. The Full-sky Astrometric Mapping Explorer (FAME) mission, which builds on experience with Hipparcos, was recently approved for development by NASA. The NASA SPACE INTERFEROMETRY MISSION (SIM) is currently in the design phase, with the goal of reaching a precision of a few microseconds on targets as faint as V magnitude 20. The European Space Agency is currently studying GAIA as a future Cornerstone Mission of the science program. Also building on the very successful experience of Hipparcos, GAIA would conduct a very large sky survey with an astrometric accuracy comparable to that of SIM, down to V magnitude 15.

Science with precision astrometry

Hipparcos represents the current state of the art in astrometry, with around 1 milliarcsecond accuracy in the positions and parallaxes of 118 000 stars and complete at around ninth magnitude. Applying the definition of annual parallax, this corresponds to 10% accuracy in the distance of stars out to 100 parsecs. Among the important science results from Hipparcos are: distances and ages of globular clusters, using parallaxes of local subdwarf stars; the internal dynamics, distance and age of the Hyades open cluster; a measurement of the distance to the Pleiades cluster; and limits to the masses of planets around stars for which radial velocity detections of planets have been made.

Ground-based instruments can achieve an accuracy of around 1 milliarcsecond in global astrometry—the measurement of large arcs across the sky. These measurements must be referenced to a global reference frame in which parallax and PROPER MOTION of stars are defined. Accurate distances require parallax in an absolute reference frame, or a close approximation to one. Similarly, problems in Galactic structure require measurements in a frame which does not share in the Galactic rotation.

Global astrometry on the ground is done by long-baseline optical interferometers (see INTERFEROMETRY: GROUND), such as the Navy Prototype Optical Interferometer (NPOI) located near Flagstaff, Arizona. This instrument features a Michelson interferometer design, and is carefully optimized for astrometry by minimizing the effects of atmospheric turbulence, and is expected to achieve a limiting magnitude of around $V = 8$.

Extending the reach of parallax distances to include most of the Galaxy (say 10 kpc) would require a further factor of 100 improvement in global astrometry, or an accuracy of around 10 microarcseconds. In addition, such measurements have to be more sensitive, because the targets are much fainter. For instance, a solar-type star at 10 kpc is almost twentieth magnitude. The Space Interferometry Mission is being designed with the precision and sensitivity to perform such measurements.

With much of the Galaxy accessible to direct distance measurements, many indirect methods of calculation of distance become unnecessary. Distances to globular clusters can be measured simply from the parallax of a few member stars. SIM will allow a calibration of the long-period Cepheids which are one of the fundamental distance indicators in the universe. The spiral structure of the Galaxy, and the dynamics of the Galactic halo can be studied using the proper motions of stars selected as 'test particles' in the gravitational potential. Distances to spiral galaxies may be measurable, by detection of the proper motions of stars due to rotation around the Galactic center.

Space astrometry with survey instruments

There are currently two approaches to space astrometry: survey instruments and pointed instruments. Hipparcos, FAME and GAIA are examples of survey instruments. SIM is a pointed instrument. Fundamentally, survey instruments achieve a lower astrometric accuracy, and that accuracy depends on star brightness. The number of observations and the integration time are typically fixed by the mission design. A pointed instrument can achieve higher accuracy, and go fainter, by increased integration time and more frequent observation. However, survey instruments are more efficient by a very large factor: FAME will observe approximately 40 million stars, whereas SIM will observe around 40 000 stars. So in practice, these instruments are complementary. Studies of Galactic structure in which a large star sample is essential to average out random motions, may be more suited to a survey instrument, whereas a pointed instrument is better suited to very demanding studies of carefully selected samples, as in the search of few-Earth mass planets around nearby stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

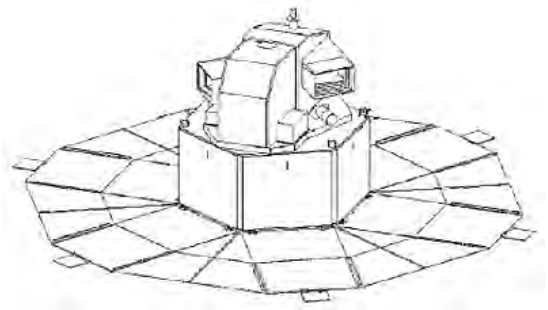


Figure 1. Sketch of the FAME spacecraft. During operation the spacecraft rotates about the (vertical) symmetry axis. Solar pressure on the combined sunshade and solar array (bottom) results in slow precession of the spin axis. The two fields of view are perpendicular to the spin, and separated by 81.5°. (Figure courtesy of Astronomical Applications Department, US Naval Observatory.)

The Full-sky Astrometric Mapping Explorer (FAME) is a space telescope designed to obtain highly precise position and brightness measurements of stars. Due to launch in 2004, it will operate, like Hipparcos, by simultaneously imaging two small fields separated by a fixed, large angle. The spacecraft rotates steadily about an axis which passes star fields from one camera to the other (figure 1). Precession of the spin axis allows all-sky coverage over a period of few weeks. Over an observing lifetime of 2.5 years, each target brighter than ninth magnitude will be observed almost 1000 times. For targets brighter than $V < 9$ the expected precision is 50 microarcseconds. In addition to astrometry, FAME will perform photometric observations through four filters with millimagnitude accuracy.

The European Space Agency is studying a follow-on mission to Hipparcos, called GAIA, operating on the same basic principles as Hipparcos. This mission, which could launch as soon as 2010, is much more ambitious than the NASA FAME mission. Reaching much further out into the Galaxy, it would yield a catalog of more than a billion stars—a factor of 25 more than FAME. GAIA would have a limiting magnitude of about $V = 20$, comparable to SIM, and would deliver an astrometric accuracy of about 10 microarcseconds for stars as faint as $V = 15$.

GAIA will create an extraordinarily precise three-dimensional map of stars throughout our Galaxy. In the process, it will map their motions, which encode the origin and subsequent evolution of the stars. Through comprehensive photometric classification, it will provide the detailed physical properties of each star observed, characterizing their luminosity, temperature, gravity and elemental composition. This massive stellar census will provide the basic observational data to tackle an enormous range of important problems related to the origin, structure and evolutionary history of our Galaxy. By reaching out to large distances, and having the ability to process large numbers of stars, GAIA is better suited than SIM to certain problems in Galactic structure. For instance, a study of the triaxial nature of the Galactic bulge requires a large number of stars as well as high precision, and is ideally suited to GAIA.

Figure 2. Artist's impression of the Space Interferometry Mission in orbit. The main structure houses the optical telescopes and Michelson interferometric beam combiner. The boom supports a system of laser metrology gauges which monitor the exact positions of the first optical elements in each telescope, a 40 cm siderostat. (Figure courtesy of Jet Propulsion Laboratory, California Institute of Technology.)

A key capability for GAIA will be the collection of simultaneous multi-band photometric data and stellar radial velocities. Three-dimensional datasets for velocity and position in the Galaxy are very important for many studies, for instance the abundance and distribution of dark matter in the Galactic disk, through its influence on stellar motions. A ground-based program to provide these supporting data is possible but very difficult and expensive. GAIA will yield a truly immense dataset of positions, kinematics, photometry and radial velocity which will be major a resource to astronomers over the coming decades.

Space astrometry with interferometers

The Space Interferometry Mission (SIM) is a space-based 10 m baseline optical Michelson interferometer, due for launch in 2006. Currently it is the only space-based interferometer specifically designed for precision astrometry. SIM will measure positions and parallaxes to about 4 microarcseconds of stars as faint as $V = 20$ by the end of its 5-year mission. One of its principal scientific goals is an extremely sensitive search for low-mass planets orbiting nearby stars. SIM will also demonstrate the interferometry technologies needed by future missions such as the Terrestrial Planet Finder.

This mission will open up many areas of astrophysics, via astrometry with unprecedented accuracy—more than 200 times more accurate than Hipparcos, and with fainter targets to much fainter magnitudes than FAME, albeit for a much smaller target list. SIM will have two basic operating

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

modes: narrow-angle and wide-angle (global) astrometry. In narrow-angle mode, the instrument measures the target position relative to a 'local' set of reference stars within about 0.5°. This mode provides the highest astrometric accuracy, allows relative (but not absolute) positions to about 1 microarcsecond in a 1 h measurement. Relative parallaxes and relative proper motions are determined by observations repeated during the mission. This mode will be used primarily for the search for low-mass planets orbiting nearby stars. The sensitivity limit is expected to be about a few Earth masses for stars within 10 pc. Other narrow-angle problems include: internal dynamics of globular and open clusters, binary orbit determinations and distances to nearby spiral galaxies using rotational parallaxes.

The intrinsic 'field of regard' of the instrument (without repointing the spacecraft) is 15°. To measure angles larger than this, SIM sets up and monitors a grid of approximately 3000 reference stars. Science observations are referenced to this grid, and the ultimate accuracy of measurements such as parallax is linked to the grid accuracy. Included in the grid will be around 50 quasars to define an inertial frame, and to tie the SIM catalog into the International Astronomical Union's ICRF system. The proper motion accuracy is expected to be around 2 microarcsec per year, corresponding to a transverse velocity of 10 m s^{-1} for a star at a distance of 1 kpc.

SIM will actually operate three interferometers simultaneously. Two are dedicated to observing bright guide stars, sensing drifts in the spacecraft attitude and providing corrections to the third, science, instrument. This design makes possible long integrations on very faint science targets.

With precision global astrometry, SIM will be able to address a broad range of topics in Galactic astronomy. In Galactic dynamics, it will reach stars far out in the halo, probing the mass distribution on large scales, as well as in the disk. This should also lead to a better understanding of the dynamics of stellar populations of different ages. By observing the brightest stars in the closest neighboring galaxies, SIM will study the dynamics and evolution of the local universe outside the Galaxy. SIM can exploit astrometric signatures created in a microlensing encounter to reconstruct the mass and kinematic properties of the lens—not currently possible with the vast majority of photometric-only detections. It does this by measuring the signature of parallax and proper motion of the unseen lens on the gravitational bending. Fundamental astrophysics will benefit from extremely accurate luminosities (via accurate parallaxes)—for instance of OB stars and planetary nebulae, which are relatively rare in the solar neighborhood.

Looking beyond SIM, the next generation of survey instrument, the European Space Agency's GAIA, described above, will usher in an even more exciting era in the methodical and detailed systematic exploration of our Galaxy and beyond.

Bibliography

Gilmore G F *et al* 1999 GAIA: origin and evolution of the Milky Way *Proc. SPIE* **3350** 541

Lindegren L and Perryman M A C 1997 Optical counterpart of ICRF: Hipparcos *Highlights of Astronomy* vol 11B (Dordrecht: Kluwer) p 287

Seidelmann P K *et al* 1999 FAME—Full-sky Astrometric Mapping Explorer *American Astronomical Society, DDA Meeting 31* 14.01

Unwin S C and Turyshev S G 1999 The Space Interferometry Mission *Ultraviolet-Optical Space Astronomy Beyond HST (ASP Conf. Ser. 164)* (San Francisco: Astronomical Society of the Pacific)

Stephen C Unwin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Instrumentation: Cosmic Backgrounds

The COSMIC MICROWAVE BACKGROUND (CMB) radiation contains approximately 99% of the electromagnetic radiation energy in the universe, and its photons outnumber the known massive particles by a factor of $\sim 10^8$. It is the primary observable indicator of conditions in the early universe and, according to COBE, the Cosmic Background Explorer satellite, it is remarkably featureless. Its spectrum is indistinguishable from a perfect blackbody spectrum with a temperature of 2.725 K. Other than the dipole variation, likely due to the motion of the solar system with respect to the CMB rest frame, its brightness is nearly isotropic with rms variations of only 30 μK on angular scales of 10°. The cosmic infrared background (CIB) radiation recently detected by COBE is presumably due to the accumulated infrared emission from the first dusty galaxies to form in the universe. The observed properties of the infrared background will place strong constraints on the star formation history of the universe.

The COBE mission carried three instruments designed to probe these cosmic background radiations. DIRBE was designed to make an absolute measurement of the spectrum and angular distribution of the diffuse infrared background. The FIRAS instrument was designed to study the spectrum of the CMB, and the DMR was a set of differential radiometers designed to measure the anisotropy of the CMB. A follow-up mission to COBE, called MAP, will carry a set of differential microwave radiometers that will remeasure the anisotropy of the CMB with much higher sensitivity and spatial resolution than did DMR. In the years leading up to the launch of MAP there have been a number of ground and balloon-based measurements of the CMB that have provided the first look at degree-scale anisotropy. These measurements are covered in the general article on COSMIC MICROWAVE BACKGROUND. A follow-up mission to MAP, planned by ESA and called PLANCK, is to follow the MAP satellite and is supposed to provide the definitive measurement of the CMB anisotropy.

DIRBE

The DIRBE instrument was designed to make an absolute measurement of the spectrum and angular distribution of the diffuse infrared radiation. One of the most challenging aspects of this experiment is the need to understand and remove foreground emission from our own Galaxy and solar system, which is known to substantially outshine the CIB. This drives the design of the instrument to include many well-spaced frequency channels each of which requires very accurate absolute photometry.

The instrument is a cryogenically cooled ten-band absolute photometer that covers the wavelength range from 1.25 μm to 240 μm in ten bands with a design sensitivity per field of view on the sky of $\lambda I_\lambda = 10^{-9}$ W m^{-2} sr^{-1}. The instrument also measures two perpendicular components of linear polarization in its short-wavelength bands at 1.25, 2.2 and 3.5 μm. The polarization measurements were included to help distinguish the CIB from sunlight scattered by INTERPLANETARY DUST. To further aid in separating interplanetary emission from more distant sources, the DIRBE optical axis is oriented 30° from the spin axis of the COBE spacecraft, so it views half the sky every day at solar elongation angles ranging from $\varepsilon = 64^\circ$ to 124° with many redundant scans. Over the course of six months, every celestial direction is redundantly sampled at all possible solar elongation angles in this range.

The DIRBE optics were carefully designed for strong stray light rejection (figure 1). The configuration includes extensive internal baffling as well as a reflective forebaffle, and a complete light-tight enclosure within the COBE dewar to eliminate any optical cross-talk from the dewar or the FIRAS instrument. Additional protection from the Earth and Sun is provided by the shield surrounding the COBE dewar, which prevents any direct illumination of the dewar from these strong local sources. The front end of DIRBE utilizes an off-axis folded GREGORIAN TELESCOPE with a 19 cm diameter off-axis segment of a parabolic mirror as the primary. An off-axis segment of an elliptical secondary forms an image of the primary at a pupil stop (Lyot stop) which is smaller than the image of the primary, blocking radiation diffracted from the primary mirror. The secondary mirror also images the first field stop at a second field stop which is smaller than the first, blocking stray radiation scattered or diffracted at the first stop and setting the $0.7^\circ \times 0.7^\circ$ instantaneous field of view. An ellipsoidal tertiary mirror images the pupil (Lyot stop) at a tuning fork chopper with mirrored blades. When the chopper is open, sky light reaches the elliptical quaternary mirror which images the pupil via a folding flat onto detector assembly 2. With the chopper closed, the aspheric mirror (M5) reimages the pupil onto detector assemblies 1 and 3.

In this arrangement, all detector assemblies are located at a pupil image and all of the spectral bands view the same instantaneous field of view. Each detector assembly receives radiation chopped alternately between the sky and the cold beam stop. The cold beam stop is a black enclosure maintained below 3 K, producing no detectable signal at any DIRBE wavelength. DIRBE sky measurements are thus continuously referenced to zero flux—a crucial feature for reliable absolute photometry. Instrumental offsets are determined by blocking the sky with a cold shutter located at the prime focus. The shutter has been demonstrated to attenuate saturation level sky signals to well below instrument noise levels at all DIRBE wavelengths. The back of the shutter has a mirror which allows light from an internal reference source (IRS), located out of the path of the sky beam, to reach the detectors via the secondary and subsequent mirrors when the shutter is closed. Measuring the instrument signal with the shutter closed and the IRS off permits determination of instrumental noise levels and electrical and radiative offsets. Turning on the IRS at fixed levels permits monitoring of system gain stability.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

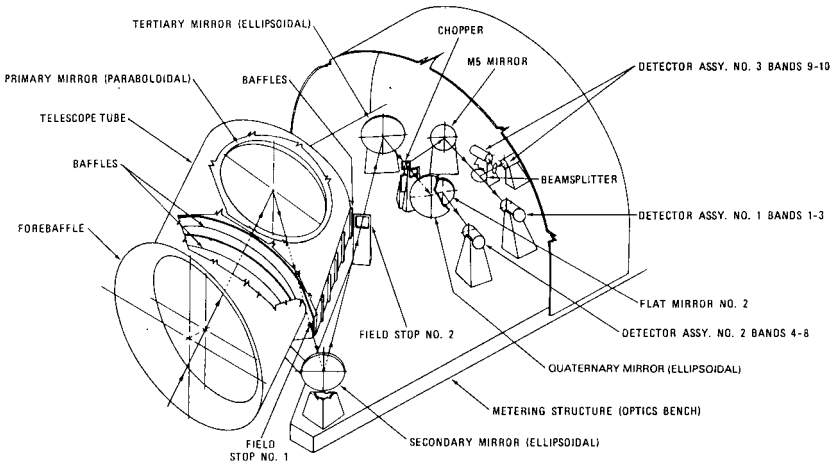


Figure 1. A schematic overview of the COBE DIRBE ten-channel absolute photometer.

The instrument readout with the shutter closed and the IRS turned off defined zero sky brightness for the DIRBE measurements.

FIRAS

The FIRAS instrument was designed to study the spectrum of the CMB. As with the DIRBE instrument, this experiment requires very accurate absolute photometry and high gain stability. One of the key features of FIRAS is that it is differential: the instrument directly measures the difference between the sky signal (in a 7° beam) and that from a temperature controlled internal reference body (ICAL) whose emission is within 5% of a blackbody. The absolute calibration of FIRAS was accomplished in orbit by inserting an external blackbody calibrator (XCAL) periodically into the mouth of the sky horn. The XCAL is a precision temperature-controlled blackbody, with an emissivity greater than 0.999. Both the ICAL and XCAL are temperature controlled by command over the range from 2 to 25 K. This allows the sky signal to be approximately balanced and aids in distinguishing sky signals from instrument self-emission. The differential nature of the instrument and the commandable temperature controlled inputs are the keys to the accuracy of FIRAS.

The heart of instrument is a four-port Michelson INTERFEROMETER with two inputs (used differentially) and two outputs (figure 2). The two input ports are fed by optically matched quasi-optical horn antennas. The interferometer uses polarizing grids as beam splitters and creates an interferogram (Fourier transform of the source spectrum) by scanning a movable mirror platform. A dichroic filter at each output port further splits each beam into low- and high-frequency bands. The low-frequency band, extending from 0.5 mm to 1 cm, was designed to measure the CMB spectrum. The high-frequency band, from 0.12 to 0.5 mm, was designed to measure the emission from dust and gas in our galaxy and to allow the removal of this foreground from the low-frequency

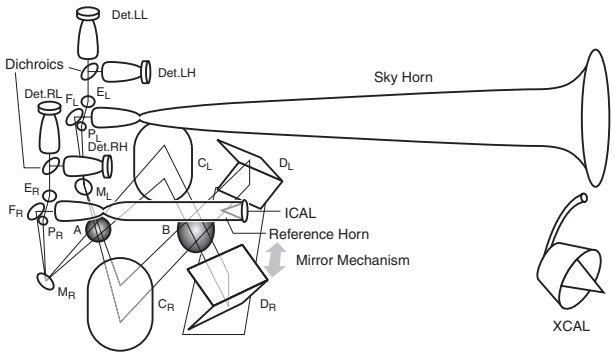


Figure 2. A schematic diagram of the Far Infrared Absolute Spectrophotometer (FIRAS).

CMB data. The objective was to attain a sensitivity of $\nu I_\nu \simeq 10^{-9}$ W m^{-2} sr^{-1} (0.1% of the peak brightness of a 2.7 K blackbody) in each 5% wide spectral element, in each 7° sky pixel. The FIRAS uses bolometric detectors in both frequency channels.

A movable mirror mechanism modulates the path difference, x, between the two optical paths resulting in a varying signal at the output ports proportional to the cosine transform of the input spectra $I(x) = \int \mathrm{d}\nu S(\nu) \cos 2\pi\nu x$. The design symmetry and phase delays imposed by the beam splitter and analyser polarizers make FIRAS a differential instrument. The nearly blackbody ICAL signal is controlled by attached heaters to nearly balance the sky input from the CMB. This reduces the net input signal by 99% and is responsible for the ability of FIRAS to detect deviations of the CMB from a perfect blackbody to less than 0.005% (rms) of the peak.

Absolute measurements of the sky flux require knowing the emission of the ICAL and other components of the instrument. This is done by examining the changes in the observed spectrum with variations in the component temperatures. For the ICAL and two

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

horns, variation is produced by commands to attached heaters. The overall throughput of the instrument is measured by placing an external calibrator, or XCAL, in the sky horn aperture—replacing the sky signal with a known input. The XCAL signal is known because it is a blackbody to within 0.005% both by construction and as verified in ground measurements. Its temperature is controlled by an attached heater and measured by a set of thermometers. Errors in the thermometry of the XCAL are the limiting factor when comparing the absolute sky flux measurements of FIRAS with other instruments.

Other design features that contribute to the success of FIRAS include a large etendue (throughput) for high sensitivity; two path length differences for the mirror mechanism (1.2 and 5.9 cm) allowing for higher frequency resolution for the longer stroke length; and two scan speeds for the mirror mechanism to help distinguish instrument errors. The entire instrument is operated in a vacuum and cooled to 1.5 K by conduction to a superfluid liquid helium tank. A large external conical shield protects the cryostat and instruments from direct radiation from the Sun and the Earth.

DMR

The purpose of the differential microwave radiometers (DMR) experiment is to measure large-angular-scale anisotropy of the CMB. Measuring the temperature of the microwave sky to an accuracy of one part in 10^5 requires careful attention to possible sources of systematic errors. Due to the extraordinary isotropy of the CMB, one of the primary concerns is instrument stability since any small perturbation to the instrument in the form stray radiation, electronic pick-up or thermal effects can easily produce a false signal that exceeds the anisotropy signal. One of the most important safeguards is to make the instrument differential. The DMR instrument consists of six differential microwave radiometers, two nearly independent channels (labeled A and B) at each of three frequencies: 31.5, 53 and 90 GHz. Each radiometer measures the difference in power, expressed as a differential antenna temperature, between two regions of the sky separated by 60°. The combined motions of spacecraft spin (73 s period), orbit (103 m period) and orbital precession (~1° per day) allow each sky position to be compared with all others through a highly redundant set of all possible difference measurements spaced 60° apart. The DMR has three separate receiver boxes, one for each frequency, that are mounted 120° apart on the outside of the COBE cryostat containing the FIRAS and DIRBE. A shield surrounding the aperture plane protects all three instruments from solar and terrestrial emission. The 31 GHz system operates at room temperature while the 53 and 90 GHz systems are passively cooled to 140 K.

The antennas are wavelength-scaled corrugated horns that were chosen for their low sidelobe response, low insertion loss and compactness. The main beam response is approximately Gaussian with a 7° FWHM. Each pair of antennas (labeled 1 and 2) feed a Dicke-switched radiometer. The output of the ferrite Dicke switch alternates the signal from the two input antennas at a rate of 100 Hz. This is an essential step in the process of reducing systematic effects by reducing the impact of low-frequency gain fluctuations and $1/f$ noise in the radiometer amplifiers. The output of the Dicke switch is sent through ferrite isolation to a mixer–preamplifier. A local oscillator signal obtained from a Gunn diode oscillator is used to mix the signal down to intermediate frequencies. It is then sent to a lock-in amplifier where the signal is square-law detected, converted to a proportional voltage and amplified with a lock-in that is synchronous to the Dicke switch. The lock-in amplifier includes a circuit which integrates this switch-modulated signal for 0.5 s producing a voltage that is proportional to the brightness difference between the two input signals. The signal is calibrated to antenna temperature by periodically injecting broad-band power of a known amplitude from a noise diode into the throat of each antenna.

MAP

The MAP mission is a follow-up to the COBE mission that is scheduled to launch in the Spring of 2001. MAP will remeasure the anisotropy of the CMB with higher sensitivity and angular resolution than COBE DMR. The MAP instrument consists of a set of passively cooled microwave radiometers with 1.4×1.6 m diameter primary reflectors to provide the desired angular resolution of less than 0.3°. The need to suppress systematic measurement errors led to the following instrument features:

(a) MAP is a differential experiment based on pseudo-correlation microwave radiometers. The instrument measures temperature differences between two points ~141° apart on the sky.
(b) MAP will orbit about the Sun–Earth L_2 libration point which provides for a very stable thermal environment and near 100% observing efficiency since the Sun, Earth and Moon are always behind the instrument's field of view.
(c) MAP will rapidly scan the sky to compare many pixel pairs on many different time scales.

The reflector design incorporates two back-to-back off-axis Gregorian telescopes with 1.4×1.6 m primary reflectors and 0.52 m secondary reflectors. Each primary is a shaped elliptical section of a paraboloid, while the secondaries are nearly elliptical. This arrangement produces two slightly convex focal surfaces on opposite sides of the spacecraft spin (and symmetry) axis with plate scales of ~15′ cm^{-1}. In order to limit diffracted signals to less than 0.5 μK, diffraction shields are employed above, below and to the sides of each secondary. In addition, the deployable solar panels and multilayer insulation guarantee that the secondaries remain in shadow during observing.

The feed design calls for as small an aperture as possible consistent with a primary edge taper requirement

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

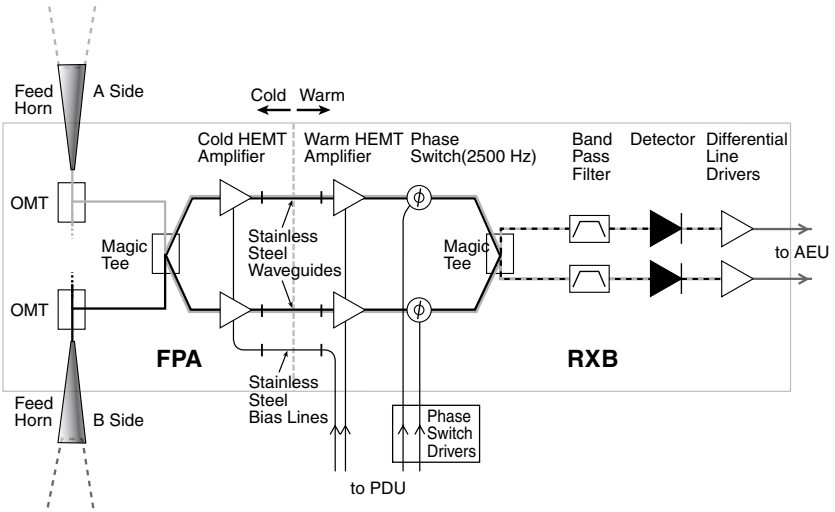


Figure 3. A schematic diagram of a single MAP pseudo-correlation radiometer including the cold focal plane assembly (FPA) and the warm receiver box (RXB).

of −20 dB, and a length that places the throat of each differential feed pair in close proximity to the other. The feed aperture diameters scale inversely with frequency, while the primary is equally illuminated at each frequency, leading to a frequency-dependent beam size. The feeds are corrugated to produce beams with high symmetry, low loss and minimal sidelobes. The phase center of each feed is kept as close as possible to its aperture, resulting in a frequency-independent beam for each feed.

The microwave system consists of ten four-channel 'differencing assemblies', one for each pair of feeds (figure 3). One assembly operates at 22 GHz, one at 30 GHz, two at 40 GHz, two at 60 GHz and four at 90 GHz. This gives the spectral coverage required to separate galactic foreground emission from the CMB and provides roughly equal sensitivity at each frequency. Each feed is attached to a low-loss orthomode transducer which separates the sky signal into two orthogonal polarizations. The A-side signal of one polarization is differenced against the orthogonal polarization, B', from the corresponding B-side feed, and vice-versa. The differencing is accomplished by first combining the two signals A and B' in a hybrid tee to form $(A + B')/\sqrt{2}$ and $(A-B')/\sqrt{2}$, then amplifying each in two cold high electron mobility transistor (HEMT) amplifiers and sending the phase-matched outputs to the warm receiver box (RXB) via a waveguide. In the RXB the two signals are amplified in two warm HEMT amplifiers, phase switched between 0° and +90° or −90° respectively, at 2.5 kHz, then split back into A and B' in a second hybrid tee. At this point, the two signals are square-law detected, amplified by two line drivers, and sent to the analog electronics unit for synchronous demodulation and digitization. The other pair of signals, A' and B, are differenced in the same manner giving a total of four detector channels per differencing assembly. The output in each channel is proportional to the temperature difference between the A- and B-side pixels. The signal is calibrated by fitting the data to the strong dipole signal that was accurately measured by COBE.

The splitting, phase switching and subsequent combining of the signals enhances the instrument's performance in two ways.

(a) Since both signals to be differenced are amplified by both amplifier chains, gain fluctuations in either amplifier chain act identically on both signals and thus cancel upon differencing.
(b) The phase switches introduce a 180° relative phase change between the two signal paths, thereby interchanging which signal is fed to which square-law detector. Thus, low-frequency noise from the detector diodes is common mode and also cancels, further reducing susceptibility to systematic effects.

PLANCK

PLANCK is the third medium-sized mission (M3) of ESA's Horizon 2000 scientific programme. It is designed to image the anisotropy of the CMB over the whole sky, with unprecedented sensitivity and angular resolution. It is now planned to launch in the first quarter of 2007 together with the FIRST (FAR-INFRARED AND SUBMILLIMETRE TELESCOPE) satellite, in a configuration often referred to as the 'Carrier' (because FIRST is 'carried' by PLANCK in the launcher). After launch, PLANCK and FIRST will separate and will be placed in different orbits around the second LAGRANGIAN POINT of the Earth–Sun system. PLANCK will carry two main instruments: the low-frequency instrument (LFI) which covers the frequency range 30–100 GHz, and the high-frequency instrument (HFI) which will cover the frequency range 100–850 GHz. This very wide frequency coverage should provide a wealth of information on both

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the CMB anisotropy and on a number of interesting foreground sources such as the SUNYAEV–ZELDOVICH EFFECT and emission from our own Galaxy. Further information on the PLANCK mission may be found on the PLANCK web site at http://astro.estec.esa.nl/Planck/.

Bibliography

For more information on the COBE mission see:

Boggess *et al* 1992 The COBE mission: its design and performance two years after launch *Astrophys. J.* **397** 420

See also the COBE web site at http://space.gsfc.nasa.gov/astro/cobe/. A very complete description of the COBE DIRBE instrument is given in the DIRBE explanatory supplement located on the COBE web site.

A very complete description of the COBE-FIRAS instrument is given by:

Mather J C, Fixsen D J and Shafer R A 1992 *Proc. SPIE* **2019** 168

This is also available in Postscript form from the COBE web site.

The design of the COBE DMR experiment is given by:

Smoot *et al* 1990 COBE differential microwave radiometers: instrument design and implementation *Astrophys. J.* **360** 685

Further information on the MAP mission may be found on the MAP web site at http://map.gsfc.nasa.gov/.

Further information on the PLANCK mission may be found on the PLANCK web site at http://astro.estec.esa.nl/Planck/.

Gary Hinshaw

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Instrumentation: Imaging Interferometry

Because of the degrading effects of the Earth's turbulent atmosphere, the spatial resolution achieved by ground-based OPTICAL ASTRONOMY is limited to the extent of the SEEING disk—the image of a point source (e.g. a single star), taken through the atmosphere. The size of the seeing disk is independent of telescope diameter, but changes only with wavelength and climatic conditions—about 0.5″ at the best observing sites in good weather. In the absence of atmospheric seeing, e.g. in space, or with a perfect ADAPTIVE OPTICS system, a telescope has a resolution defined by the DIFFRACTION limit, which scales inversely with telescope diameter: the larger the mirror, the sharper the image. For example, the HUBBLE SPACE TELESCOPE with its 2.5 m aperture has a diffraction limit of 0.05″ at a wavelength of 500 nm, 10 times better than the seeing limit, but still not good enough for many areas of astronomical research. Projects such as resolving the immediate surroundings of black holes in external galaxies, or imaging extrasolar planets, require telescope diameters of at least 100 m. To manufacture and launch into space such large apertures is clearly out of reach of present-day technology. Yet, images with such seemingly impossible resolution are feasible. They can be obtained with a method called *synthesis imaging* which can achieve basically unlimited resolution without unreasonably large telescopes. In this article, I discuss the principle of synthesis imaging and the advantages of imaging interferometry from space. I also briefly describe various projects planned for the near future which promise to revolutionize astronomical imaging. While the general discussion of synthesis imaging is valid for the entire electromagnetic spectrum, the degree of difficulty of its realization varies dramatically with wavelength. Here, I will focus solely on the optical regime and ignore the rich and successful history of the RADIO INTERFEROMETER.

The idea of synthesis imaging

Figure 1(*a*) shows a simple star cluster as it would be seen by a (hypothetical) diffraction-free 10 m telescope in space. Imagine its primary mirror covered by an opaque mask with a set of holes along a line. Obviously, most of the image information would be lost—but not all of it. The image produced through such a mask shows a set of INTERFERENCE fringes—alternating stripes of high and low intensity (figure 1(*b*)). These fringes contain information about the structure of the object. Different rotation angles of the mask produce additional fringe patterns. All patterns can be recorded and combined later to obtain the complete source image, as demonstrated in figures 1(*c*) and 1(*d*). In fact, there is no need for the full 10 m telescope mirror: the fringe patterns can be obtained with just a set (at least two) of small mirrors by combining their beams in a common focus. This is the elegant principle behind synthesis imaging: to mimic a large telescope with a number of smaller ones.

There is, however, one caveat which in the past has prevented the successful realization of synthesis imaging at optical wavelengths. In order to produce the all-important interference fringes, the beams from the individual APERTURES must be combined coherently. This means that the difference in their pathlengths to the common focus must not exceed the coherence length of the light. Combining the two signals coherently requires accurate control of the pathlengths between the two telescopes and the beam combination optics. The enormous technological difficulties associated with controlling optical pathlengths to the required levels of a few wavelengths are described elsewhere (see INTERFEROMETRY: GROUND).

How does it work?

The mathematical framework of synthesis imaging is based on the theory of wave diffraction. The basic result is summarized in the van Cittert–Zernike theorem:

> the complex degree of coherence is equal to the normalized Fourier transform of the source intensity distribution[1].

The degree of coherence is a complex quantity, the observables measured in practice are its amplitude and phase. They can be derived from the interference pattern (the fringe) which appears on the detector when scanning the delay between the two beams.

An imaging interferometer measures the degree of coherence across the (u,v) plane. Because, in essence, this measures the Fourier transform of the source structure, the vector between the two telescopes—the baseline $\vec{B}$—is usually measured in the spatial frequency domain, i.e. in units of wavelengths, $u = B_x/\lambda$ and $v = B_y/\lambda$. The plane in which the source wavefront is sampled is therefore called the (u,v) plane. Different baseline vectors are achieved by moving the individual telescopes with respect to the source, thus sampling different (u,v) points. According to the van Cittert–Zernike theorem, this allows the full reconstruction of the source structure via an inverse Fourier transform, provided that the (u,v) coverage is complete. However, the (u,v) coverage of any real interferometer will have gaps, resulting in ambiguity in the source reconstruction. The ring-like residuals in figure 1(*d*) are one example of such ambiguities, also

[1] Strictly speaking, this formulation of the van Cittert–Zernike theorem is correct only for the case of an incoherent light source. Also, both the source size and the baseline length must be much smaller than the distance to the source (the far-field approximation). However, astronomical sources at optical wavelengths always meet all of these conditions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001

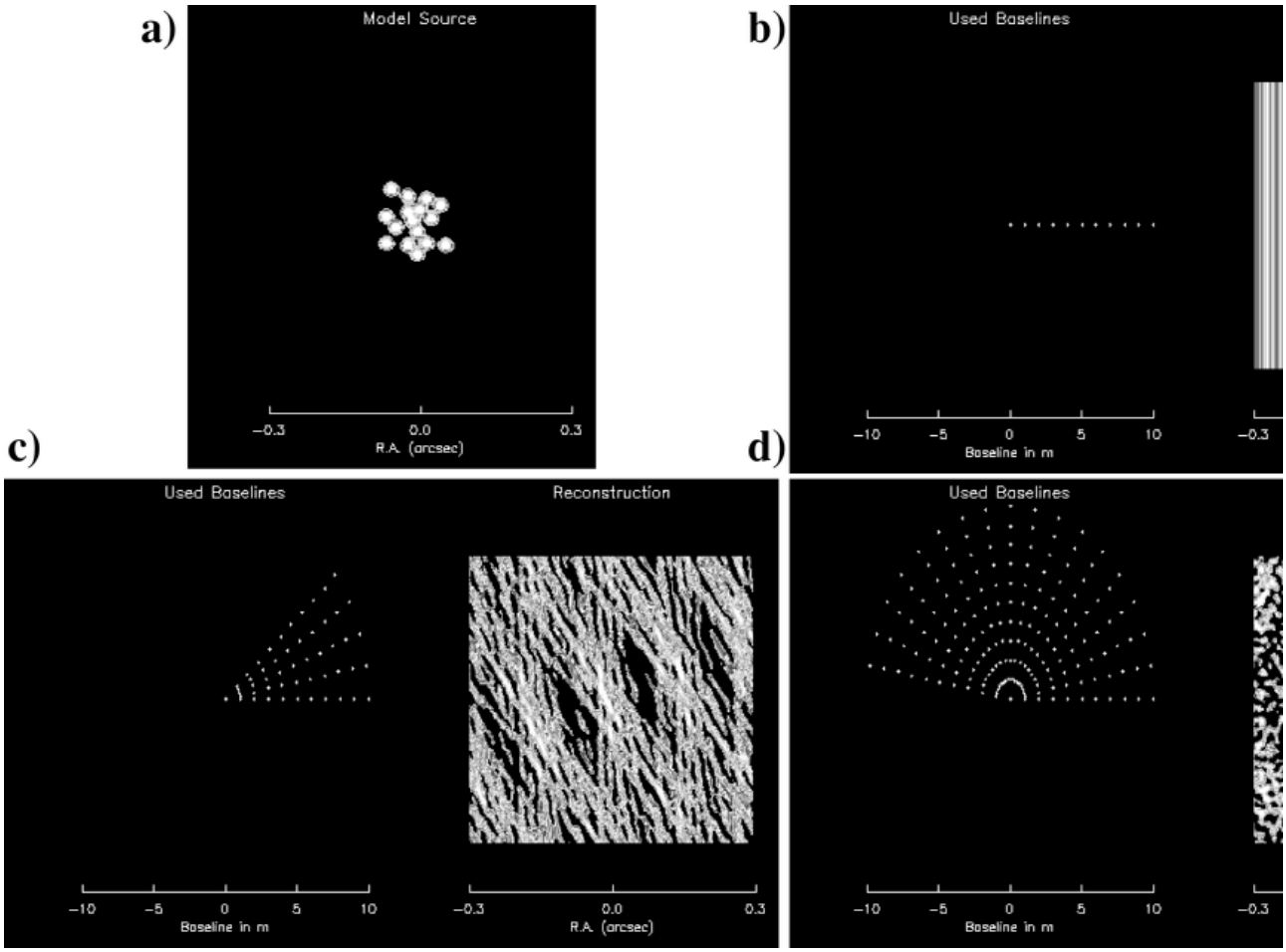


Figure 1. Illustration of the principle of synthesis imaging: (*a*) model source to be imaged; (*b*)–(*d*) image reconstructions (right) for various baseline distributions (left). The ring-like residuals around the central source image are called grating rings. They are a consequence of the regularly spaced gaps in the (*u*,*v*) coverage and can be removed to a large extent by numerical image restoration algorithms.

known as grating rings. The completeness of the (*u*,*v*) coverage is therefore a critical design parameter for imaging interferometers.

Spatial resolution

To first order, the resolution of an imaging interferometer is defined by the longest baseline. However, in contrast to the diffraction limit of a filled aperture, the spatial resolution of a synthesis imaging observation is not uniquely defined. By varying the weighting of the data obtained at different (*u*,*v*) coordinates, or even removing some baselines completely, one can put more or less emphasis on certain spatial frequencies in the reconstruction and thus change the effective resolution, even after the data are taken. This fact is regularly exploited at ground-based telescopes in aperture masking observations of bright sources for which collecting area is not an issue but the highest possible resolution is required.

Field of view

The FIELD OF VIEW (FOV) of an imaging interferometer is limited by the coherence length of the light, $l = \lambda^2/\Delta\lambda$, where λ and $\Delta\lambda$ are the wavelength and the bandwidth of the observation. While the pathlengths from the two interferometer elements ideally are equal in the center of the field, an angular separation Θ from the center necessarily produces a path difference ΘD. As long as this difference is small compared with l, the fringe contrast will not be affected. Thus, $\Theta \leq (\lambda/D)\lambda/\Delta\lambda$, or $\Theta < R_{\mathrm{spatial}}R_{\mathrm{spectral}}$. The maximum FOV of an interferometer thus can be estimated from its spatial and spectral resolution. In practice, however, the FOV is likely to be limited by the effects of the incomplete (*u*,*v*) coverage (see figure 1).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synthesis imaging in space

All that has been said so far about synthesis imaging applies equally to space- and ground-based interferometers. There are, however, important differences between the two environments in a number of aspects.

(u,v) coverage

Most interferometer arrays will consist of relatively few elements because of cost constraints. In order to overcome the intrinsically sparse (u,v) coverage of such systems, many employ movable apertures. In addition, all ground-based arrays use the Earth's rotation during the course of the night to sample many different (u,v) points. The achievable (u,v) coverage therefore depends on both the geographical location of the array and the source position with respect to the Earth's rotation axis.

In contrast, an interferometer in space cannot rely on the Earth's rotation. It therefore must have movable apertures to fill the (u,v) plane. On the other hand, the (u,v) coverage can be chosen freely. This allows a uniform resolution which can be valuable in survey programs which study the statistical properties of large numbers of similar objects.

Aperture size

The size of the individual apertures of a ground-based interferometer is limited to the size of an atmospheric turbulence cell—about 30 cm at optical wavelengths. In principle, the use of adaptive optics can overcome this limit, but the need for a bright reference source (a 'guide star') within the isoplanatic patch severely limits the number of suitable targets.

In space, on the other hand, there is no turbulent atmosphere. Therefore, the size of interferometer elements is limited only by our technical and financial ability to manufacture—and bring into space—large mirrors.

Frame time

Another important obstacle for interferometry on the ground imposed by the Earth's atmosphere is the need to record the complex visibility within a coherence time of the atmosphere—about 20 ms at a good site for optical wavelengths. Together with the above limit on aperture size, this severely constrains the target brightness: it has to be bright enough to overcome the noise associated with the detection system, from both the detector and the read-out electronics.

Even with a perfect adaptive optics system, the integration time at each (u,v) point is limited because of the Earth's rotation: the measurement has to be completed before the baseline has changed noticeably in order to avoid (u,v) smearing.

In space, the available integration time for each (u,v) coordinate is in principle unlimited, so that much fainter objects can be observed. This assumes, however, that structural vibrations or slow drifts are perfectly corrected.

Passbands

Astronomy from the ground is limited to a number of small regions of the electromagnetic spectrum over which the Earth's atmosphere is transparent. In space, no such limitations exist, and important wavebands such as the ultraviolet or the mid-infrared become accessible.

Environment

Clearly, operating an interferometer in space to the extreme precision required for successful fringe tracking is an extremely challenging task. In the case of single spacecraft with a number of apertures on a connecting truss, the mechanical vibrations of the structure must be minimized, and, even for low-noise structures, residual motions must be actively corrected. In principle, however, space is a favorable environment because it is intrinsically much quieter than the geologically active surface of the Earth. Because of this and the limitations of a truss with regard to the achievable baseline length, future synthesis imaging in space is likely to be realized with a suite of free-flying apertures which are positionally controlled with respect to each other and the beam combining optics by reference laser beams.

Ongoing and planned projects

The technology that makes the ambitious methods of active pathlength control and fringe tracking feasible has only recently become available, mostly through the development of ever faster computers that allow real-time control of mechanical instabilities. While the principles of synthesis imaging are well understood, and the technique has been demonstrated on the ground, interferometry in space has yet to be realized. Today, a number of space-based interferometer projects capable of performing synthesis imaging are in the development or planning stages in the United States and Europe.

StarLight is intended to demonstrate the feasibility of precision control for free-flying spacecraft. It consists of two spacecraft: one that serves as a light collector, and one that comprises a second collector, delay line and beam combining optics. Launch is planned for 2005. The primary goal of StarLight is to demonstrate the technological readiness for subsequent science missions.

The Space Interferometry Mission (SIM) consists of a single spacecraft with three interferometer pairs along a 12 m truss. The individual telescopes have a diameter of about 30 cm. The primary science goal of SIM is to perform high-precision astrometry for a number of key programs. Because SIM will allow the truss structure to be rotated and will cover a yet to be determined range of baseline lengths, it will have some potential for synthesis imaging of simple astronomical sources. SIM is scheduled for launch in 2009.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Terrestrial Planet Finder (TPF) in its current design comprises four free-flying 3.5 m telescopes and a fifth spacecraft with the beam-combining optics. The design allows maximum baselines of about 1 km for a resolution of less than 0.001″ at wavelengths around 3 μm. The main scientific goal of TPF is the detection and characterization of Earth-like planets at infrared wavelengths by means of interferometric nulling of the light from the host star. A tentative launch date for TPF is 2012.

DARWIN is a mission concept within the cornerstone program of the European Space Agency (ESA) with a launch after 2014. Much like TPF, it will use free-flying telescopes as the interferometer elemetns. It has a similar design concept and scientific goals to TPF.

While the basic technological tools for these ambitious missions appear to be in hand or within reach, their successful implementation will occur in steps and will require time, effort and money. On the way to the ultimate goal of taking images of other Earths, major advances in many other aspects of astronomy are almost certain because of the dramatic increase in spatial resolution that only synthesis imaging can provide.

Bibliography

The complete treatment of wave diffraction and interference is treated in great detail in

Born M and Wolf E 1980 *Principles of Optics* 6th edn (Oxford: Pergamon)

A more detailed review about the intricacies of interferometry both from the ground and in space can be found in

Shao M and Colavita M M 1992 Long baseline optical and infrared stellar interferometry *Annu. Rev. Astron. Astrophys.* **30** 457–98

Torsten Böker

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Interferometry Mission (SIM)

Proposed NASA astrometry mission and the first to use optical interferometry. Launch is planned for June 2005. The spacecraft will drift away from the Earth at approximately 0.1 AU per year, reaching a maximum communication distance of about 95 million km after five years. In this orbit, it will receive continuous solar illumination. It will carry a science interferometer and two guide interferometers with a 10 m baseline. Used to determine the position and distances of stars to extremely high accuracy and to detect brown dwarf stars and planets. It will also pioneer a technique to block out (null) the light of bright stars. Possible angular resolution of 10 milliarcseconds—four times better than the Hubble Space Telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Probe

An uncrewed spacecraft designed to investigate the interplanetary medium or conditions on or in the vicinity of another celestial body. A space probe travels along an orbit around the Sun or on a trajectory that takes it from the vicinity of the Earth to the neighborhood of its target (for example, a planet, comet or asteroid). In order to reach interplanetary space, a space probe has to exceed the escape velocity of the Earth. In order to enter a closed orbit around its target planet, a space probe has to reduce its speed relative to that planet, usually by firing onboard rocket motors (if no action of this kind is taken, the craft will fly past its target on a hyperbolic trajectory and recede into interplanetary space). The first space probe to leave the vicinity of the Earth and enter an independent orbit around the Sun was Luna 1, a Soviet probe that passed by the Moon at a range of 6000 km on 4 January 1959, and continued on into interplanetary space.

See also: escape velocity, Luna, orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Research Institute

The Space Research Institute (IKI), located in Moscow, is one of the leading organizations in Russia in space research. The main fields of research are astrophysics, space plasma physics and solar system exploration. During its history, from its foundation in 1965, IKI has fulfilled many experiments and missions, including investigations of the Moon, Venus, Mars, Comet Halley (VEGA mission) and Phobos (PHOBOS-2 mission); and, during the last decade, the MIR-KVANT and GRANAT space observatories (astrophysics) and the INTERBALL program (space plasma research). Main sources of funding are the Russian Academy of Sciences and the Russian Space Agency.

For further information see http://www.iki.rssi.ru/

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Research Organization Netherlands

The Space Research Organization Netherlands (SRON) initiates, develops and exploits outstanding instruments for research in and from space. SRON is responsible for the national space research program, which is carried out by the two laboratories for space research in Utrecht and Groningen. Activities comprise the design, development and manufacture of space instruments, as well as the processing and interpretation of the data.

Its principal contributions to astrophysics are: Reflection Grating Spectrometer on XMM; Low Energy Transmission Grating Spectrometer for Chandra; Wide Field Cameras aboard BeppoSAX; Heterodyne Instrument for ESA's Far-Infrared and Submillimeter Space Telescope; and ISO Short-Wavelength Spectrometer.

For further information see http://www.sron.nl/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Telescope European Coordinating Facility (ST-ECF)

The ST-ECF, which was established in 1984, is jointly operated by the EUROPEAN SPACE AGENCY (ESA) and the EUROPEAN SOUTHERN OBSERVATORY (ESO). It is located at the ESO headquarters in Garching, near Munich.

The facility supports the European astronomy community in exploiting the research opportunities provided by the Hubble Space Telescope. It provides detailed technical information about the HST and its science instruments, and supports European astronomers in the preparation of HST observing proposals. It also coordinates the development of computer software to meet the specific data analysis needs of HST users, operates and maintains an archive of all the scientific data collected by HST, and acts as a European center for associated meetings and workshops.

ST-ECF staff maintain close contacts with the SPACE TELESCOPE SCIENCE INSTITUTE (STScI) in Baltimore, which is charged with the scientific operation of the HST observatory. ST-ECF also provides a contact point for European media and public information.

For further information see
http://www.stecf.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Telescope Science Institute (STScI)

Located on the Johns Hopkins University Homewood campus, Baltimore, Maryland. The institute is responsible to NASA's GODDARD SPACE FLIGHT CENTER for the scientific operations of the Hubble Space Telescope (HST). It was established by NASA, following a recommendation by the National Academy of Sciences, and is operated by ASSOCIATION OF UNIVERSITIES FOR RESEARCH IN ASTRONOMY (AURA) under contract to NASA.

STScI's main role is to maximize the scientific productivity of HST and to serve the astronomical community in its operation. The institute solicits and reviews observation proposals and selects observations to be carried out. It schedules observations and assists guest observers in their work. It also generates an integrated science and engineering timeline to support all spacecraft activities and provides the facilities and software to calibrate, analyse and archive HST data.

STScI also monitors the HST and its science instruments for characteristics such as instrument performance, pointing accuracy and telescope focus. As part of the planning and scheduling system, the STScI Guide Star Selection System provides reference stars and other bright objects so that the HST's fine guidance sensors can point the telescope accurately.

The institute also played a key role in the early mission concept studies for the Next Generation Space Telescope (NGST) and has been selected by NASA to manage the science and mission operations for this successor to the HST.

STScI is staffed by astronomers, computer scientists, technicians and administrative staff from AURA, the European Space Agency (which is a partner with NASA on the Space Telescope Project), and the Computer Sciences Corporation. Of its combined staff of approximately 500, approximately 100 are astronomers and scientists.

For further information see
http://www.stsci.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Space Velocity

A measure of the true velocity in three dimensions of a star relative to the Sun, calculated from measurements of its radial velocity and tangential velocity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spacetime

A four-dimensional geometrical framework that unites the three dimensions of space and the dimension of time. A point in spacetime (identified by four coordinates) is called an event (or 'world-point'). The history of a particle (a line in four dimensions connecting all the events at which it is present) is called a world-line.

The concept of spacetime was devised in 1907 by Hermann Minkowski (1864–1909). Whereas Newton regarded time as being independent from the three dimensions of space, Minkowski contended that they are intimately related. Four-dimensional spacetime was adopted by Einstein as the framework within which to develop his general theory of relativity. According to that theory, spacetime is curved in the presence of matter and the paths of rays of light and particles of matter are determined by the curvature of spacetime.

See also: general theory of relativity, special relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spanish Astronomy

Astrophysics is a very recent development in Spain, but its spectacular evolution in itself constitutes a singular phenomenon that has aroused the interest of sociologists and historians of science alike.

Historical background

It is well known that, during the Middle Ages, the Iberian peninsula was the European center of astronomical knowledge. Not only were the works of classical and oriental astronomy rediscovered and made available to the world at large (especially through the School of Translators at Toledo), but it was here also that new instrumentation was developed and advances were made in the scientific understanding of the heavens, contributing decisively to the empirical and conceptual foundations of the Renaissance astronomical revolution. In the 15th and 16th centuries the art of navigation, which opened up ocean routes to the advantage of Spanish and Portuguese mariners, derived its base in the astronomy developed in Spain. Significantly, the Prime Meridian, the origin of longitude, was first defined to pass through Toledo and later through the island of Hierro—it was only in the 19th century that it was finally settled to pass through the Royal Observatory, Greenwich.

It was not until the Enlightenment that astronomy once more flourished in Spain. The dedication of the mariner-astronomers of Cadiz, led by Jorge Juan and Antonio de Ulloa, played a decisive role in bringing the new science and technology to Spain, and not only in the field of astronomy.

The observatory of Cadiz was founded in 1751 and published its first astronomical ephemerides (the third oldest in the world) in 1792. The Observatory of Madrid was created at the end of the century.

Spanish Enlightenment astronomers were fully aware that the most advanced instrumentation was needed in order to pursue modern astronomy. That is why the observatories of Cadiz and Madrid have in their possession two of the finest instruments of the period. A further indication of this was the commissioning of a giant telescope, 25 ft long with a primary mirror 2 ft in diameter, from William Herschel, who was the foremost telescope maker of his time. Unfortunately, before it could enter into service this optical marvel was destroyed by Napoleonic troops garrisoned at Madrid Observatory.

Slow progress began to be made in Spanish astronomy by the beginning of the present century. The obervatories of San Fernando and Madrid were refurbished, and in 1904 the new obervatories of Fabra (in Barcelona) and Cartuja (Granada) were built. However, the Spanish Civil War (1936–9) and its aftermath led to a further halt in the progress of Spanish astronomy, which stagnated into the more routine aspects of 'positional astronomy' and 'celestial mechanics'.

After the Second World War European astronomers began to seek new telescopic locations in southern regions with more favorable climates.

In 1958 the path of totality of a solar eclipse passed through the Canaries and astronomers from all over the world traveled to the islands. This occasion served to recall the 19th-century Tenerife observations of PIAZZI SMYTH (who later became Astronomer Royal for Scotland). Piazzi Smyth made his observations at several different heights ranging from sea level to the peak of Mount Teide and confirmed Newton's prediction that the stars would be better viewed at greater heights. For this reason, Piazzi Smyth's expedition to Tenerife has come to be regarded as the precursor of present-day mountain observatories.

Also, in 1910, Jean Mascart, an astronomer from the Paris Observatory, went to Tenerife to observe the return of COMET HALLEY, this expedition forming part of a wider undertaking by the World Health Organization to study the effects of the toxic gases in the tail of the comet, which on this occasion would sweep past the Earth. He made his observations at Piazzi Smyth's site and, apart from deducing that the toxic gases of comets' tails were undetectable, he concluded that the Canarian skies were of great purity and offered exceptional conditions for astronomical observing. Unfortunately, the resulting Franco–German initiative to create a European observatory on the flanks of the Teide was to come to nothing with the advent of the First World War.

Today, Piazzi Smyth's Canarian astronomical adventure is commemorated by the lunar features (christened by Piazzi Smyth himself) *Montes Tenerife* and *Pico* (after the peak of the Teide).

In the light of these memories of past expeditions, it eventually came to be realized that the Canaries might offer one of the sites sought after by modern astronomers, particularly in the branch of astrophysics. The Spanish authorities, in response to international pressure, decided to take action and in 1959 the National Ministry of Education created the OBSERVATORIO DEL TEIDE, under the responsibility of the Rector of the University of La Laguna, with the express aim of establishing whether the heights of Tenerife could provide the conditions demanded by modern astronomical observing and, if so, of promoting the setting up of an observatory there.

Astronomical site testing was undertaken by Spanish astronomers in the 1960s and the first telescopes were built at the Observatorio del Teide. They were of small aperture and were dedicated to the photometry of extended sources and solar physics. This led European astronomers to think that the Canaries might be a good site for setting up their telescopes and they eventually completed the site testing.

During the 1970s, formal agreements were drawn up with various European countries concerning the setting up of advanced telescopes in Spanish territory: in 1972, the Hispano–German Astronomical Centre (CAHA), which runs the CALAR ALTO OBSERVATORY (in Almeria), was intituted, and in 1979 several European countries signed the 'Agreements on Cooperation in Astrophysics', through which the observatories of the INSTITUTO DE ASTROFÍSICA DE CANARIAS (IAC) are made available to the international scientific community. It was during this period that the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

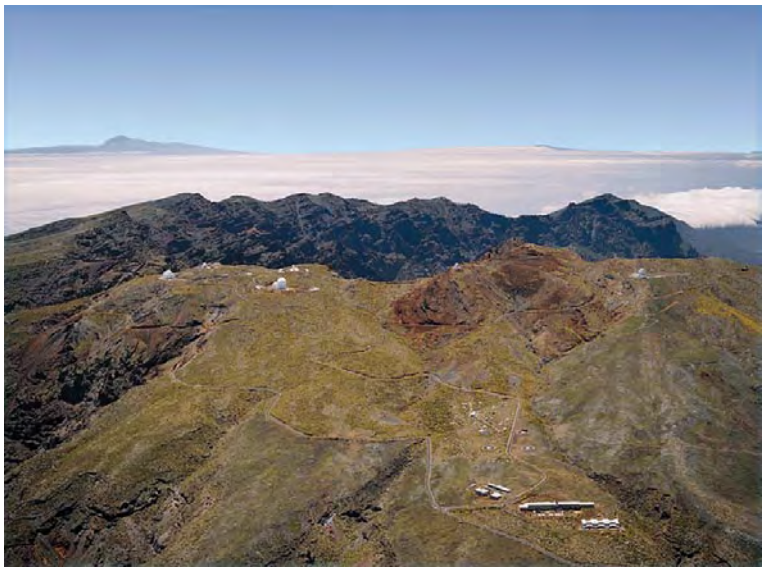


Figure 1. Aerial view of the Observatorio del Roque de los Muchachos, on the island of La Palma. It is at this Observatory, at a height of 2400 m and bordering the Caldera de Taburiente National Park in the muncipality of Garafia, that the 10 m GTC (*Gran Telescopio Canarias*) will be built.

Figure 2. Partial view of the Observatorio del Teide, on the island of Tenerife. Astrophysics in the Canaries began on this island in this observatory, situated in the district of Izaña at a height of 2400 m, where the municipalities of La Orotava, Fasnia and Güímar meet. It has the most complete array of solar instrumentation in the world.

infrastructure was provided for the setting up of the most advanced group of European optical–infrared telescopes of their time.

Radio astronomy began to be developed in parallel with these activities. In 1977 the Yebes Observatory (Guadalajara) was created, and in 1982 the INSTITUT DE RADIO ASTRONOMIE MILLIMETRIQUE (IRAM) installed a 30 m millimetric radio telescope at Pico Veleta (Granada). These were followed by NASA space-tracking dishes in Robledo and Cebreros (Madrid).

From the 1980s onwards, this train of circumstances provided Spain, which had not hitherto developed a base in astrophysics and had been lacking in any kind of modern astronomical instrumentation, with the most wide-ranging panoply of optical and radio telescopes in Europe. By agreement, the Spanish scientific community was to receive between 10% and 25% of the the observing time on all foreign instruments.

The vision, enthusiasm, tireless energy and perseverance of a small group of individuals led eventually to the creation of the necessary internal conditions to enable full use to be made of this incredible wealth of facilities in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

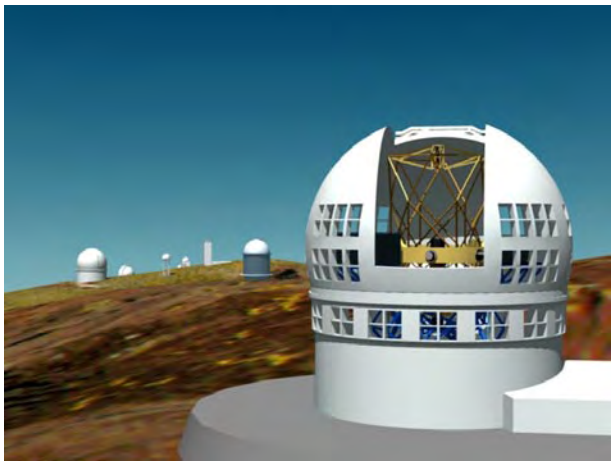


Figure 3. Three-dimensional view of the building and dome of the GTC (*Gran Telescopio Canarias*). Its primary mirror will consist of 36 hexagonal segments, each with a largest diameter of 1.9 m and together forming the equivalent of a conventional monolithic mirror of 10 m diameter. Its size and design innovations will make the GTC the most advanced instrument of its generation when it sees first light in the year 2002. The IAC is leading the construction of this advanced telescope, to be located at its Observatorio del Roque de los Muchachos, on the island of La Palma.

Figure 4. June 1985 saw the official inauguration of the Canarian astrophysical observatories, presided over by the Spanish Royal Family and royalty and heads of state from Denmark, the Federal Republic of Germany, Great Britain, Ireland, the Netherlands and Sweden. Among the representatives of the international scientific community were five Nobel Laureates in physics.

the form of the birth and vigorous development of astrophysics in Spain. The new influx of foreign instrumentation served as a lever to raise Spanish astronomy in record time to a par with the most scientifically advanced countries. A useful illustration of the activity of the nascent science of astrophysics in Spain is given by the fact that the 4.2 m Anglo–Dutch William Herschel Telescope is five times oversubscribed by Spanish applications for observing time.

The present quantity and quality of scientific production, the number of Spanish groups heading lines of research and the construction of instrumentation have together won the admiration of the international scientific community, especially when the rapidity and determination with which this high level of achievement has been reached are taken into account. Some even speak of 'Spain's astronomical miracle'.

The present situation

In order to gain an overall view of present-day Spanish astronomy, it would be useful to follow its main lines of development in terms of the recent flood of new instrumentation into Spain, the training of researchers, the creation of new administrative structures, an open policy of fostering public awareness of astronomy and the building of new instrumentation.

However, before entering into these aspects it might be opportune to mention the fields in which Spanish astronomy has made important contributions. Among these I cite, in alphabetical order and with no attempt at being comprehensive, active galaxies, astrometry, brown dwarfs, circumstellar dust shells, cool stars, the cosmic microwave background and the large-scale structure of the universe, interstellar molecules, nebulae and H II regions, neutron stars, normal galaxies, relativistic astrophysics, solar physics, star formation and stellar populations, white dwarfs and x-ray binaries.

A deluge of new instruments

Table 1 summarizes the main astronomical resources at present located on Spanish soil. Only the collecting area has been highlighted since it is a telescope's diameter that determines its resolving power and light-grasping capacity (as well as its cost).

Special mention must be made of the IAC's international observatories, in which more that 30 scientific institutions from 14 countries participate, and which effectively constitute the European Northern Observatory (ENO). There are nine telescopes with mirrors of diameter over 1 m, not counting the 10 m *Gran Telescopio Canarias* (GTC) now under construction (figure 3). The observatories were inaugurated in 1985 by the King and Queen of Spain in the presence of six other heads of state and a distinguished selection of scientists (figure 4).

The excellence of the Canarian skies is specifically protected as an 'astronomical reserve' by Spanish law.

The training of researchers

The training of astrophysicists was begun in the 1960s, opportunities being provided for young researchers to travel to foreign research centers. It was necessary for Spain to have its own investigators if it was to derive any benefit from the numerous instruments set up in its territory. The training of good astrophysicists was then, and continues to be, a prime objective.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Main astronomical resources.

	Collecting surface (m^2)
Observatorio del Teide (Tenerife)	
4 solar telescopes	1.20
1 infrared telescope	1.77
3 general-purpose telescopes	1.48
Various radiotelescopes for CMBR	
Various helioseismological instruments	
Total	4.45
Participating countries: Belgium, Germany, France, Italy, Spain, Taiwan, United Kingdom, United States of America	
Observatorio del Roque de los Muchachos (La Palma)	
2 solar telescopes	0.48
1 meridian circle	0.03
6 general-purpose telescopes	34.46
Multiple cosmic-ray detectors	
Total	34.96
(With the coming into service of the GTC this total will rise to 113.46 m^2.)	
Participating countries: Denmark, Finland, Germany, Holland, Ireland, Italy, Norway, Portugal, Russia, Spain, Sweden, United Kingdom	
Centro Astronómico Hispano-Aleman (Almeria)	
1 Schmidt camera	0.50
4 general-purpose telescopes	16.31
Total	16.81
Participating countries: Germany, Spain	
Observatorio de Sierra Nevada (Granada)	
2 general-purpose telescopes	2.40
Total:	2.40
Participating country: Spain	
Observatorio de Pico Veleta (Granada)	
1 radio telescope	706.50[a]
Total:	706.50
Participating countries: France, Germany, Spain	
Centro Astronómico de Yebes (Guadalajara)	
1 radio telescope	153.86[a]
1 astrograph	0.13
Total:	153.86
Participating country: Spain	

[a] The collecting area of a radio telescope is not comparable with that of an optical telescope.

In 1975 Spain's Ministry of Education and Science granted the IAC the facilities to pursue 'a national program for the training of researchers in astrophysics' that was to last for 4 yr and produced the first small group of PhDs in astrophysics. Since then, the IAC has run a postgraduate school that continues to generate an average of nine PhDs each year.

However, this educational undertaking is also shared by other astrophysics centers and university departments throughout Spain, bringing the mean annual total of new doctorates in astrophysics up to about 20.

Hence, today it may be said that Spain's main 'astronomical resource' is not only the telescopes installed in its territory but also its highly competitive workforce of young astrophysicists, who have already earned international recognition and prestige. In 1998 there were 462 astronomers in Spain, 305 of whom had doctorates (data provided by the Sociedad Española de Astronomía).

New administrative structures

The successful creation of centers adapted to confront new situations and conceived to derive the fullest benefit from international cooperation has undoubtedly been a decisive factor in the development of astrophysics in Spain.

In 1959, the National Ministry of Education set up the Observatorio del Teide in the Canaries. In 1972, the first chair in astrophysics was created at the University of La Laguna, which, in 1973, inaugurated the 'University Institute of Astrophysics'. Soon after, in 1975, the University of La Laguna, the National Research Council (*Consejo Superior de Investigaciones Científicas*, CSIC) and the Association of Local Governments (*Mancomunidad de Cabildos*) of the province of Santa Cruz de Tenerife (which includes the island of La Palma) jointly created the present Instituto de Astrofísica de Canarias (IAC). In 1982, this institute gained the legal status of a 'public consortium', made up of the State Administration, the Canarian Autonomous Community, the University of La Laguna and CSIC; the Observatorio del Teide (on Tenerife) and the OBSERVATORIO DEL ROQUE DE LOS MUCHACHOS (on La Palma) are in turn governed by the IAC.

In Andalusia, the Instituto de Astrofísica de Andalucía (IAA) was set up in 1975 by CSIC. The IAA had its first headquarters in Almeria and later in Granada. The creation of the IAA was the culmination of the combined efforts of the Cartuja Observatory and the University of Granada. In 1982, a chair and a department of astrophysics were created at the University of Granada. In its turn, the IAA built the Sierra Nevada Observatory in the 1990s.

CSIC also has astrophysicists in other institutes and at the Centre d'Estudis Avançats de Blanes (Catalonia); it also participates in other centers such as the Instituto de Astronomía y Geodesia of Madrid's Complutense University and the Instituto de Física of Cantabria University.

The Instituto Geográfico Nacional (which runs Madrid Observatory) created the Centro Astronómico de Yebes, an observing station at Calar Alto and, more recently, an astronomical center at Alcalá de Henares.

The Royal Institute and Naval Observatory (in San Fernando) participates in the Meridian Circle at the Roque de los Muchachos Observatory; in 1997, it became a key institution in astrometry when it set up its own meridian circle at El Leoncito (Argentina).

The Instituto de Técnica Aeroespacial (INTA) became involved in astrophysics in 1972 and, together with scientists at the *Consejo Nacional de Investigaciones Espaciales* (National Council of Space Research), created the Laboratorio de Astrofísica Espacial y Física Fundamental (LAEFF) in Villafranca de Castillo.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Other Spanish universities eventually followed suit and joined in the astronomical 'boom'. At the present time the following universities are, to a greater or lesser degree, active in astrophysics: Alicante, the Balearic Islands, Barcelona (Autonomous), Barcelona (Central), Barcelona (Polytechnic), Cantabria, Granada, La Laguna, Madrid (Autonomous), Madrid (Complutense), Murcia, the Basque Country, Santander, Santiago de Compostela, Seville, Tarragona, Valencia, Valladolid, Vigo and Saragossa.

Promoting public awareness of astronomy

Spain's new astronomy has been nurtured with a deep commitment towards fostering a public awareness of astronomical knowledge. Astrophysicists play a prominent role in the popularization of astronomy throughout the country, the IAC's participation in the events of the Seville 1992 *Exposición Universal* and the Science and Cosmos Museum (Tenerife) being clear examples of this commitment.

Interest in the universe around us is growing fast in Spain, especially among the young, a fact that is reflected in the frequent media coverage of astronomical topics. Astronomy is always mentioned in popular science magazines, some of these being exclusively devoted to astronomy.

However, perhaps the clearest evidence of social assimilation of astronomy is given by the burgeoning of active AMATEUR ASTRONOMICAL SOCIETIES. Here there has also been a boom, from a mere handful in the 1960s to the creation of organized amateur groups in nearly all the cities (indeed, there are several societies in some cities).

The construction of astronomical instrumentation

The first instruments built in Spain were simple, but little by little these have evolved in their complexity and refinement. Today, there are highly sophisticated Spanish-built focal-plane instruments mounted on some of the most advanced telescopes in the world.

Also outstanding in this regard are Spain's recent efforts in space-borne instrumentation. In 1995 ISO (Infrared Space Observatory) was launched by ESA and SOHO (Solar Heliospheric Observatory) by NASA; both observatories carried on board instruments built by Spanish industry led by the IAC. Later, the University of Valencia provided instrumentation for MINISAT (INTA). In the CASSINI and ROSETA missions the IAA led Spain's participation; in the very near future, the University of Valencia and LAEFF will play the leading role in the INTEGRAL satellite.

Instrumentation for the forthcoming ESA–NASA PLANCK and FIRST missions will be provided by groups of Spanish astrophysicists.

However, the culmination of all Spain's endeavors in the building of new astronomical instrumentation is the *Gran Telescopio Canarias* (GTC), with its latest-generation 10 m diameter segmented mirror, for which the segment blanks have already been ordered. There is now funding to complete the construction of the telescope and, consequently, the participation of other countries has now been limited to a maximum of 30%. The project is running to schedule and first light is expected for 2002, with the GTC entering into service in 2003.

Future prospects

Observation is, and will continue to be, a key element in the development of models and theories that aid our understanding of the universe. After the spectacular advances made in this century, astrophysics is now at the threshold of new enigmas, whose unraveling will require the use of large telescopes located in the best sites on the planet. Spain possesses one such site: the Observatorio del Roque de los Muchachos. For all these reasons, and for its direct impact on Spanish science, technology and industry, the IAC has spearheaded the initiative to build the *Gran Telesopio Canarias* (GTC) and will continue to commit itself to the large telescopes of the future.

The take-off of astronomical research in Spain has been greatly favored by access to advanced instrumentation. In order to consolidate these gains it will be necessary for Spain to bolster its theoretical element while at the same time reinforcing its observational capacity. It will also be necessary to dedicate manpower and means to opening and exploiting all windows of observation, both ground and space based.

The high human potential available to Spanish astronomy and the astronomical quality of its observatories and the inevitable growth of research and development investment in Spain together promise an extremely productive future in this science.

Francisco Sánchez

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spartan (Shuttle Pointed Autonomous Research Tool for Astronomy)

Small NASA satellite deployed and retrieved by the Space Shuttle. First flew in June 1985. The most recent (fifth) flight in November 1998 carried an ultraviolet coronal spectrometer and white-light coronagraph to study the solar corona and acceleration of the solar wind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Special Astrophysical Observatory, Russia

The Special Astrophysical Observatory (SAO) of the Russian Academy of Sciences was organized in 1966 as a managing institution for operating the 6 m mirror optical reflector BTA (Bolshoj Azimuthal Telescope) and the 600 m aperture radio telescope RATAN-600. The observatory is located near the valley of Zelenchuk river (North Caucasus). The 6 m telescope site is at the altitude of 2100 m above sea level. In addition to the BTA telescope, a few smaller instruments of 0.6 to 1.0 m diameter are based at the site. Radio observations are made with the RATAN-600, situated on the southern outskirts of Zelenchukskaya village, about 20 km to the north from BTA. An administrative center, library, scientific laboratories and dwelling houses are located in Nizhnij Arkhys settlement, which is about halfway between the SAO main instruments.

Both telescopes were built in 1975, but regular observations started two years later. The observational time at the telescopes is scheduled by international allocation committees. Today SAO is the main Russian center for ground-based astronomical observations. About one-third of scientific programs at the telescopes are carried out in cooperation with European astronomers.

BTA was the world's largest telescope till the beginning of 1990s. First use of the alt-azimuth mounting for the giant telescope became an example for further large instrument projects. The radio telescope RATAN-600 was designed as an antenna with a changeable surface, allowing application of the method of aperture synthesis.

The activity of the observatory is supported by the Russian Academy of Sciences and the Ministry of Science and Technology. Staff of the Observatory numbers 480 people. There are two members of the Russian Academy, 14 ScD and 60 PhD among 114 SAO researchers. Research activity includes the space distribution of galaxies, active galactic nuclei and quasars, stellar atmospheres and magnetic fields, study of binary and multiple stars, relativistic objects, interstellar medium structure, and solar radio astronomy.

For further information see
http://www.sao.ru.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Special Relativity

The special theory of relativity (SR), developed by ALBERT EINSTEIN in 1905, represents a revolutionary change in the human conception of the nature of space and time. According to SR, space and time are not absolute but must be viewed as components of a single entity, SPACE-TIME. This idea has numerous important implications, both conceptual and practical.

SR provides a framework for formulating the laws of physics in the absence of gravitational fields. When gravity is included, the GENERAL THEORY OF RELATIVITY (also developed by Einstein) must be employed. Because gravity is a very weak force, SR is adequate for most problems.

A central role is played in SR by the *inertial frame*, defined as a frame of reference in which Newton's law of inertia holds: a body subject to no external forces moves with constant velocity. Whenever two such frames move rapidly relative to one another, SR departs markedly from Newtonian physics.

SR is based on two postulates:

(a) (principle of relativity) every law of nature has the same form in any two inertial frames;
(b) the speed of light in a vacuum, c, is the same in all inertial frames.

A principle of relativity had been proposed many years earlier by GALILEO, but was long believed to apply only to mechanical phenomena; Einstein's first postulate extends Galileo's principle to encompass *all* physical laws. The second postulate is contrary to intuition, which suggests that the measured speed of a light beam (like that of any moving body) should depend on the motion of the observer.

SR is by now supported by a vast body of experimental evidence. Some of its major implications are discussed in the following paragraphs.

Simultaneity. Observers in different frames will in general assign different times to a given event. Two events that are simultaneous in one frame of reference occur at different times according to observers in another frame.

Time dilation. A clock that is moving in a given frame runs slow when compared with clocks that are stationary in that frame. The ratio between recorded time intervals is $(1 - V^2/c^2)^{-1/2}$, where V is the velocity of the 'moving' clock.

Length contraction. A body is shorter when measured in a frame in which the body is moving than it is in its own rest frame. The ratio between the measured lengths is again given by the factor $(1 - V^2/c^2)^{-1/2}$. Only the dimension in the direction of motion is contracted.

Velocity transformation. Let $\boldsymbol{v}$ be the velocity of a body in frame S and $\boldsymbol{v}'$ be its velocity in frame S′, which moves at velocity $\boldsymbol{V}$ relative to S. Let $v_\parallel$ and $v_\perp$ be the components of $\boldsymbol{v}$ parallel to and perpendicular to the direction of $\boldsymbol{V}$. The components transform as follows:

$$v'_\parallel = \frac{v_\parallel - V}{1 - (v_\parallel V/c^2)} \tag{1a}$$

$$v'_\perp = \frac{v_\perp (1 - V^2/c^2)^{1/2}}{1 - (v_\parallel V/c^2)}. \tag{1b}$$

In pre-Einsteinian relativity, the corresponding relations are simply

$$v'_\parallel = v_\parallel - V, \tag{2a}$$

$$v'_\perp = v_\perp. \tag{2b}$$

Relativistic beaming. Consider a transformation in which V is close to c. Equations (1) imply that if $v_\parallel$ and $v_\perp$ are of the same order of magnitude, $v'_\perp$ is much smaller than $v'_\parallel$. It follows that if a source moving at almost the speed of light emits either particles or radiation more or less isotropically in its own rest frame, the radiation observed in the laboratory is highly concentrated in the direction of the source's velocity. (Equations (1) apply to radiation as well as to material particles.)

Relativistic Doppler effect. SR leads to a modification in the classical formula for the change in wavelength detected when the source of a light wave moves relative to the observer. The relativistic Doppler formula is

$$\frac{f}{f_0} = \frac{\lambda_0}{\lambda} = \left(\frac{1 \mp \beta}{1 \pm \beta}\right)^{1/2} \tag{3}$$

where λ and f are the observed wavelength and frequency and λ_0 and f_0 are the corresponding values when source and observer are at rest with respect to one another. The upper signs apply when source and observer are receding, the lower ones when they are approaching. (For oblique relative motion, a somewhat more complicated formula applies.)

Mass–energy relation. When a body's energy changes by an amount ΔE, its mass (inertia) changes by an amount $\Delta m = (\Delta E)/c^2$. A rest energy $E_0 = mc^2$ can be associated with a body of mass m.

Relativistic energy and momentum. Conservation of momentum holds in SR provided momentum is defined as $p = \gamma m v$, with

$$\gamma = (1 - v^2/c^2)^{-1/2}. \tag{4}$$

Relativistic kinetic energy is given by

$$K = (\gamma - 1)mc^2. \tag{5}$$

The total energy of a body with velocity v is $E = \gamma mc^2$. In the limit $v \ll c$, momentum and kinetic energy reduce to their Newtonian forms $p = mv$, $K = \frac{1}{2}mv^2$. Momentum and total energy satisfy the useful relation

$$E^2 = p^2c^2 + (mc^2)^2. \tag{6}$$

Newton's law of motion holds in the form $F = \mathrm{d}p/\mathrm{d}t$ (but not in the form $F = ma$).

Photons, electromagnetic waves and particles. Maxwell's equations of electromagnetism are consistent with the principle of relativity: they have the same form in all

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

inertial frames and can therefore be used to calculate the radiation emitted by an accelerated charge of any speed.

Photons (light quanta) can be incorporated into the theory provided their mass is taken to be zero. The energy–momentum relation for a photon is $E = pc$, consistent with equation (6) if $m = 0$.

Particles can be created and annihilated, subject to the appropriate conservation laws (energy, momentum, charge, lepton number, baryon number, etc). For example, an electron and a positron can annihilate into two or more photons. (One-photon annihilation is forbidden by conservation of momentum.)

The role of c as a limiting speed. If the relative velocity between two frames were greater than c, the second postulate could not be satisfied. According to equation (1), $v_\parallel = 0$ and $v_\perp = c$ implies $v'_\parallel = V$; if $V > c$, v' cannot equal c.

Equation (4) shows that as $v \to c$ the kinetic energy of a body approaches infinity. Hence a body of non-zero mass cannot be accelerated to speed c.

The velocity transformation implies that if $v < c$ and $V < c$, then $v' < c$. The hypothesis that c is a limiting speed is therefore consistent with the first postulate: if a body's speed is always less than c in one inertial frame, the same is true in all frames[1].

Relativity and causality. An intriguing consequence of SR is that the time order of spatially separated events can differ in two inertial frames. If that were true of events between which a cause-and-effect relation exists, observers in some frame would detect the effect before its cause. This logical catastrophe is avoided by the hypothesis that no signal can propagate faster than c. If event E_1 is the cause of E_2, a signal of some kind must travel between them. If the speed of the signal is bounded by c, the time interval between the events must be at least L/c, where L is their spatial separation. It can be shown that whenever that condition is satisfied, the time order of the events is absolute: the cause precedes the effect in every frame. A 'superluminal' signal could lead to violation of causality and would have other bizarre consequences.

Applications to astronomy and astrophysics

The presence in the cosmos of relativistic particles (i.e. of particles whose speed is very close to c) is confirmed by detecting either the particles themselves (in cosmic rays) or the radiation they emit. Most non-thermal sources of radiation are interpreted as being due to relativistic electrons.

[1] One can show also that if $v > c$ and $V < c$, then $v' > c$: a body moving faster than light in one frame moves faster than light in all frames. During the 1960s physicists investigated the possibility that faster-than-light particles, named tachyons, might exist. A consistent theory can be constructed if the tachyon's mass is assumed to be imaginary. A tachyon's energy then becomes infinite as its speed approaches c from above; hence it cannot be slowed below c. All experimental searches for evidence of tachyons proved unsuccessful, however.

Cosmic rays

Particles with energies as high as 10^{20} eV have been detected in COSMIC RAYS. The cosmic-ray flux has both a Galactic and an extra-Galactic component; except for those of the very highest energies; however, most cosmic rays are believed to originate within the Galaxy, probably from supernovas.

The characteristic signature of a high-energy cosmic ray is the *extensive air shower*. The primary particle emits a gamma-ray photon by bremsstrahlung in the electric field of an atom in the atmosphere. The photon subsequently creates an electron–positron pair, each member of which emits more photons. A cascade thus develops; in addition to electrons and photons the shower can contain muons, pions, neutrinos and other particles. Showers containing as many as 10^9 particles have been detected. They can extend over several square miles; because of relativistic beaming, however, most of the shower is confined to a small central core.

Synchrotron radiation

A charged particle in a magnetic field follows a helical path, with the field direction as axis. Being accelerated, the particle radiates. At low velocity the radiation is primarily at the gyrofrequency $\nu_0 = qB_\perp/2\pi\gamma mc$. At relativistic velocity, however, the emission is dominated by high harmonics of ν_0 and is effectively a continuum, peaked at the frequency $\nu_m \cong 0.45\gamma^3\nu_0 \cong 10^6\gamma^2 B_\perp$. (The numerical expression is for electrons, with B in gauss and frequency in Hz.) The radiation is emitted preferentially in the forward direction, within a cone of opening angle $1/\gamma$, and is strongly linearly polarized. The spectrum falls off exponentially for $\nu \gg \nu_m$. The radiated power varies as $E^2 B_\perp^2$.

Many acceleration mechanisms lead to a power-law distribution of electron energies, $N(E) \sim E^{-n}$. The synchrotron spectrum from such a source likewise follows a power law, $I_\nu \sim \nu^{-\alpha}$; the spectral index of the radiation, α, is related to the index of the electron energy spectrum by $\alpha = (n-1)/2$ and the degree of linear polarization is $\Pi = (n+1)/(n+7/3)$.

Power-law spectra, at frequencies from radio to x-rays and even gamma rays, are observed from a variety of Galactic and extra-Galactic sources and are generally interpreted as SYNCHROTRON RADIATION. Typical spectral indices range from about 0.2 to 1.5. In some sources the spectrum steepens in the optical and x-ray regions.

In the CRAB NEBULA, a supernova remnant, the magnetic field is estimated to be a few times 10^{-4} G. If the observed luminosity is synchrotron radiation, the energy contained in relativistic electrons is $\sim 10^{48}$ ergs, about 10^{-4} times the energy released in a typical supernova explosion. Electrons that emit in the x-ray band have energies about 10^{13} eV.

The radiative lifetime of a particle emitting synchrotron radiation varies as $E^{-1}B^{-2}$. For electrons emitting x-rays in the Crab, the lifetime is only about 20 years, much less than the age of the remnant; hence the electron

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

energy must be continuously replenished. The mechanism responsible for the replenishment was a puzzle until the discovery of the PULSAR (a rotating neutron star) within the nebula. The loss in kinetic energy associated with the observed decrease in the pulsar's rotational speed is enough to accelerate the electrons.

The pulsar's own radiation is also interpreted as being synchrotron radiation in a much higher magnetic field, estimated to be 10^6 G or more. Some high-energy electrons escape from the remnant entirely, accounting for at least part of the cosmic- ray electron flux.

Other sources of synchrotron radiation include radio galaxies and quasars, as well as the non-thermal radio emission from our own Galaxy. The presence of relativistic electrons indicates violent activity in a source.

Inverse Compton radiation

In the Compton effect, a photon scatters from a stationary electron and is slightly degraded in energy. If a photon collides with a relativistic electron, the process can be described as Compton scattering in the electron rest frame. When the collision is transformed to the laboratory frame, the photon's lab energy is multiplied by a factor of roughly γ^2.

The inverse Compton process is quite similar to synchrotron radiation. An electron's energy loss is proportional to γ^2 times the energy density of either the local photon field (inverse Compton) or the static magnetic field (synchrotron); in each case a power-law distribution of electron energies leads to a power-law spectrum of radiated energy, with the same relation between spectral indices.

The inverse Compton effect plays an important role in X-RAY and GAMMA-RAY ASTRONOMY. Relativistic electrons interact with photons in the microwave background as well as with optical photons in starlight. A 100 MeV gamma ray can be produced by a 4 GeV electron scattered off an optical photon or by a 150 GeV electron scattered off a photon from the microwave background.

The interaction with the microwave background is strong enough to exclude a metagalactic origin for high-energy cosmic-ray electrons: because of Compton losses an electron with energy $\sim 10^{11}$ eV cannot reach the solar system even from the nearest radio galaxy, at a distance $\sim$4 Mpc.

When a cloud of relativistic electrons emits synchrotron radiation, the electrons lose energy by Compton scattering off the synchrotron photons. This effect sets an upper limit on the possible intensity of synchrotron sources.

Cerenkov radiation

Although no material particle can travel faster than light *in vacuo*, the speed of light in a medium is only c/n, where n is the index of refraction of the medium. Hence a charged particle can exceed the speed of light in the medium. The particle then emits a characteristic radiation called Cerenkov radiation; it is analogous to the shock wave emitted by a body traveling faster than sound in a medium. Cerenkov radiation is emitted within a cone of opening angle $1/\gamma$; it is commonly used to detect high-energy particles.

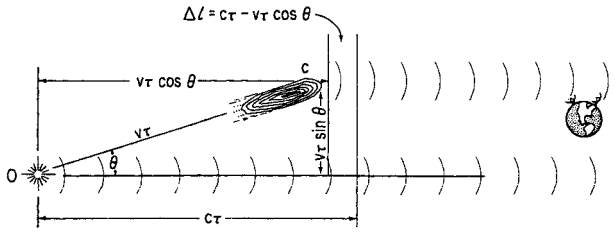


Figure 1. Model for superluminal emission.

Limitation on source dimensions

If a cosmic source exhibits substantial variation in intensity over a time T, one can argue from causality that its dimensions cannot be much greater than cT. This argument has been used to infer upper limits on the sizes of QUASARS and other active regions, whose intensity has been observed to vary over months or even less. Care must be taken in making such arguments, however.

Superluminal effects

In recent years an interesting phenomenon has been observed in several extra-Galactic compact RADIO SOURCES: material appears to be ejected from a central core at a speed that exceeds c, sometimes by a factor as high as 10. If matter were actually moving that fast, SR would be violated. What one observes, of course, is the angular separation between the ejected material and the core; if one determines the distance of the source from its red shift (assumed to be cosmological), the inferred transverse velocity exceeds c.

The commonly accepted explanation of the effect is shown in figure 1. If a source of radiation moves at a small angle θ to our line of sight, at speed v close to c, the source nearly keeps up with its own radiation. The sketch shows two pulses of radiation emitted at times separated by an interval T. When it emitted the second pulse, the source was closer to us by $vT\cos\theta$. The interval between arrivals of the two pulses is therefore $T[1-(v/c)\cos\theta]$. The transverse displacement of the source during the interval between the two emissions is $vT\sin\theta$; hence its apparent transverse velocity is

$$v_{\rm app} = v\sin\theta/[1-(v/c)\cos\theta]. \qquad (7)$$

The maximum of this function, for $\theta = \sin^{-1}(1/\gamma)$, is $v_{\rm app} = \gamma v$, which is arbitrarily large.

Although the *a priori* probability that a source moves nearly in the direction of our line of sight is small, the relativistic beaming effect, described above, increases the probability that such a source will be detected.

L Sartori

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Speckle Imaging of Binary Stars

In addition to their ability to detect faint and distant objects in the universe, large telescopes can, in principle, resolve fine detail in celestial objects. The theoretical limit of resolution for a telescope is linearly related to the APERTURE (or diameter of the primary light-collecting mirror or lens of the telescope) and by the wavelength at which an observation is made (see RESOLVING POWER). Hence larger apertures and shorter wavelengths enhance resolution. Unfortunately, for telescopes beyond 10–20 cm apertures, atmospheric turbulence thwarts the attainment of the full diffraction-limited performance otherwise permitted by theory. This turbulence reveals itself in the form of 'twinkling' of starlight to the unaided eye. However, when a star is observed through a telescope of modest or larger aperture, one sees a boiling distribution of intensities which spread the starlight out in a blurred and distorted 'seeing disk' typically 1 arcsec in angular extent, many times larger than would otherwise by the case.

In 1970, the French astronomer Antoine Labeyrie announced an elegant new technique for analyzing these so-called 'speckle' images from large telescopes to extract information about the structure of the object being observed on scales down to the theoretical limit. Labeyrie's method of 'speckle interferometry', which was diversified by Labeyrie and others to become more generally known as 'speckle imaging', involved the analysis of snapshot exposures (approximately 0.01 s in duration) of stars taken at high magnification and relatively narrow spectral bandwidths using power spectrum or autocorrelation methods (see also CORRELATION FUNCTION AND POWER SPECTRA IN COSMOLOGY). The observing requirements limit the method to relatively bright objects in most cases.

While speckle imaging has been applied to a variety of objects at visible and infrared wavelengths, the greatest successes of the technique have been in the field of binary star studies. Binary stars are pairs of stars gravitationally bound in mutual orbits (see BINARY STARS: OVERVIEW). The majority of the stars exist not as single objects like the Sun but in binary or higher-order gravitational systems. The measurement of the orbital period of a binary system and the linear size of the orbit provides a direct determination of the mass of the components of the binary. This is the only observational means available to astronomers for measuring the masses of stars, and much of our understanding of how stars form, evolve and die is ultimately tied to how well we know the masses of stars. Unfortunately, although some 100 000 binary star systems have been discovered, most have long orbital periods, and hence their orbits are poorly known.

The enhanced resolution of speckle imaging, when carried out at telescopes with apertures larger than 2 m or so, allows the resolution of binaries that otherwise appear as single stars when observed through the turbulent atmosphere. Classical methods for resolving binaries utilize the human eye and have led to the discovery of pairs

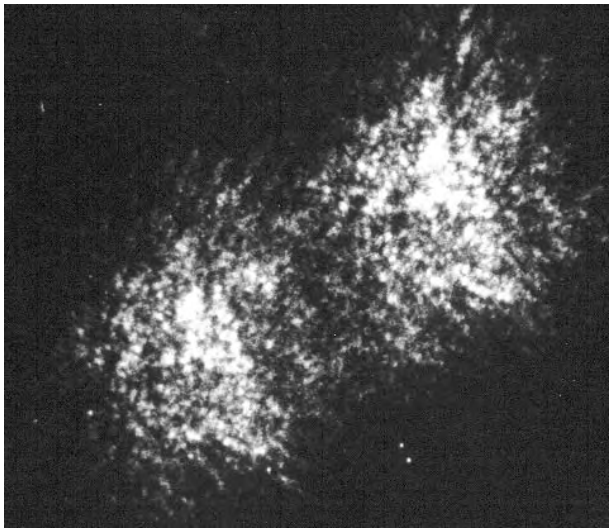


Figure 1. The double star 57 Cancri, with an angular separation of 1.5 arcseconds, is shown here. Note the similarity between the speckle patterns of the two stars in this speckle image taken with the 4 m telescope on Kitt Peak, Arizona.

with angular spacings on the sky as small as 0.15 arcsec. Speckle imaging extends this down to 0.03 arcsec, or even smaller when one of the few existing telescopes larger than 4 m is utilized. This increase in resolution coupled with the much higher measurement accuracy of speckle imaging very significantly improves the quality of binary star orbit determinations. Furthermore, where classical methods tended to select binaries with orbital periods of decades or centuries, speckle imaging can resolve systems with periods of years or decades, permitting the determination of the orbit within a reasonable timescale. More than 20 000 measurements of several thousand binary stars have resulted from speckle imaging, including the discovery of several hundred bright binary systems.

Speckle imaging of binary stars has now made obsolete the visual study of these systems. The most active center for speckle imaging at present is the US NAVAL OBSERVATORY in Washington, DC, which also serves as a clearing house and archive for all types of measurements for resolved binary star systems.

While speckle imaging techniques are applicable to single telescopes, related 'interferometric' methods employing two or more telescopes separated by baselines up to hundreds of meters in length permit the resolution of binaries much closer than those accessible to speckle imaging (see INTERFEROMETRY: GROUND). Several interferometric arrays of telescopes are operational and others are coming on line. The current generation of such instruments will resolve systems with angular spacings as small as 0.0001 arcsec, corresponding to orbital periods on the time scales of hours and days. As with speckle imaging at single telescopes, these telescope arrays must also cope with atmospheric turbulence, but interferometric arrays promise a very substantial contribution to the imaging of binary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars and to the determination of the basic physical parameter for stars—their mass.

Bibliography

McAlister H A 1996 Twenty years of seeing double *Sky Telescop.* **92** (November) 28

Harold A McAlister

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrographs: Fiber-fed Spectrographs

Fiber-fed spectrographs exploit the unique properties of optical fibers to increase the performance of astronomical instruments. The most important applications are in multi-object spectrographs which enable simultaneous observations of many astronomical targets, as image reformatters for area spectroscopy, and as a means to feed high-stability bench-mounted spectrographs.

Optical fibers are cylindrical glass light guides of small diameter (typically 10–300 μm). The most common type used in ASTRONOMICAL SPECTROSCOPY are known as multi-mode step-index fibers which rely on total internal reflection at the boundary between the fiber core and a lower refractive index cladding. The critical angle $\theta_{\rm crit}$ for total internal reflection defines the *numerical aperture* (NA) of the fiber which depends on the refractive indices of the core and cladding as $\mathrm{NA} = \sin\theta_{\rm crit} = \sqrt{n_{\rm core}^2 - n_{\rm cladding}^2}$. Commercially available fibers have numerical apertures typically in the range 0.15–0.35 which imply a minimum input focal ratio $F_{\rm min} \approx \sqrt{1-\mathrm{NA}}/2\,\mathrm{NA}$ for light to be efficiently transmitted. Silica glass fibers have excellent transmission in the optical (0.4–2.0 μm) range and lengths of several tens of metres can be employed without significant internal absoption. In the ultraviolet these losses increase rapidly (as λ^{-4}) due to Rayleigh scattering by density and composition variations within the fiber. At longer wavelengths the transmission of silica drops dramatically due to hydroxyl ion absoption, but heavy-metal fluoride (e.g. zirconium fluoride) and chalcogenide glasses with improved infrared transmission are now becoming available.

An important property of optical fibers for the design of astronomical spectrographs is the decrease in FOCAL RATIO of the transmitted beam known as *focal ratio degradation* (FRD). The degree of FRD depends on several factors, the most important of which is internal stresses within the fiber. The relative change in focal ratio is proportional to the input focal ratio F, so that optical fibers work best with fast (small F-ratio) beams (so long as $F > F_{\rm min}$).

Multi-object fiber spectrographs

One of the most successful applications of optical fibers in astronomy has been the development of multi-object spectrographs which can obtain the spectra of many objects simultaneously. Single optical fibers are positioned in the telescope focal plane so that each acquires the light from a separate target object. On passing through the optical fiber, any spatial information about the target is largely scrambled (at least azimuthally) so entrance diameters of 1–2 arcseconds are usually chosen to match the mean angular sizes of the objects of interest (limited by the SEEING for point sources). The output ends of the fibers are aligned to form a long slit which then feeds a conventional dispersive spectrograph.

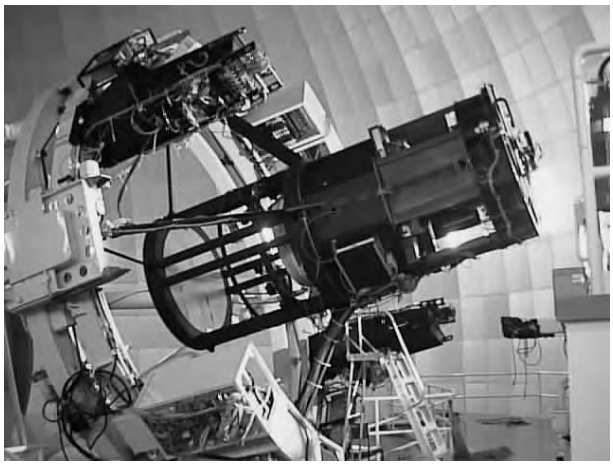


Figure 1. The 2dF multi-fiber spectrograph can deploy up to 400 optical fibers across a 2 degree diameter field at the prime focus of the 3.9 m Anglo-Australian Telescope. At the core of the instrument, seen here at the top of the telescope tube, is an advanced robotic positioner capable of repositioning an individual fiber to an accuracy of a few μm in < 10 seconds. To maximise efficiency, two sets of 400 fibers are available so that observations can continue using one of them whilst the second is being reconfigured. Changeover between fields takes only a few minutes. The fibers feed two spectrographs which are equipped with CCD detectors and are also mounted on the top of the telescope. ©Anglo-Australian Observatory

Most of the engineering difficulties are associated with accurately positioning the fibers in the focal plane (typical requirements being 20 μm accuracy across a 200 mm diameter field). The first multi-object fiber-fed spectrographs, developed in the late 1970s, used a manual plug-board approach but these have now mostly been replaced by automated robotic positioning mechanisms. The latter fall into two main classes: (i) multiple positioner devices where individual fiber probes are actively controlled using stepper motors; (ii) single positioner devices where the fiber probes are moved sequentially by a single pick-and-place robot and held in position passively (usually by a permanent magnet on the base of the probe). The first method allows very rapid ($<$ 1 minute) reconfiguration of the field but is limited in multiplex capability by the size of the active control mechanisms (typically to 20–50 fibers). Using the second approach, multi-object fiber-fed spectrographs have been constructed (see figure 1) which can obtain spectra for several hundred astronomical targets simultaneously, at the cost of some limitations in reconfiguration time.

The tremendous gains in observing efficiency, when a sufficient density of targets is available, have ensured that multi-object fiber-fed spectrographs have been developed for most of the world's major observatories. Currently around 30 systems have either been built or are in the design phase. Calibration observations which monitor individual fiber throughputs are required when using these systems to observe faint targets. Unlike multi-slit spec-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

trographs (see SPECTROGRAPHS: MULTISLIT SPECTROGRAPHS), the sky-background signal cannot be sampled contiguously with the object. The sky spectrum is therefore usually derived from separate fiber probes which may take very different optical paths through the spectrograph making sky-subtraction difficult. However, close to Poisson-noise limited spectra have been demonstrated by 'beam-switching' of objects between close pairs of fibers.

Integral-field fiber spectrographs

Fiber-fed integral-field spectroscopy is a means to obtain area spectroscopy of extended targets using optical fibers to dissect the image. The simplest systems use dense packs of bare fibers at the input, but the efficiency is limited (typically to $\leq$ 60%) by the packing fraction achievable with a given core/cladding ratio. The most effective method is to use a high fill-factor hexagonal or square microlens array to couple the input light into a matrix of individual fibers. These input optics can simultaneously condition the input focal ratio to an optimum value (F/3–F/5) for propagation down the fiber with minimal FRD. The spatial coverage and spectral resolution of such instruments are largely determined by detector size, with devices containing 1000+ spatial elements now becoming available. Their advantages relative to other forms of imaging spectrographs (e.g. Fourier transform, Fabry-Perot) are the relatively wide spectral coverage, fixed format on the detector, and the efficient use of detector pixels.

High-stability fiber spectrographs

The use of flexible fibers as optical feeds from the telescope focal plane to the spectrograph presents the opportunity to remove the spectrograph from the telescope entirely and mount it on a stationary platform. Although limiting the throughput at ultraviolet wavelengths, the advantages of doing so are increased stability (and decreased cost) due to the stable gravity vector, and the absence of flexure during telescope motion. In addition, the image scrambling properties of optical fibers can be used to advantage to remove the wavelength errors associated with focal plane image motion (from seeing or telescope guiding errors). Although the radial scrambling of light in an optical fiber is incomplete, particularly at fast F-ratios, the azimuthal scrambling leads to an order of magnitude improvement in centroid stability compared to that of a conventional slit spectrograph. Because they are sited off the telescope, fiber-fed high-stability spectrographs can also be kept in a controlled environment to minimize thermal effects. Using these techniques, wavelength precisions of 1 part in 10^6 are achievable, comparable to that required to detect Jupiter-size planets around solar-type stars via their reflex Doppler motions (see also EXOPLANETS).

Conclusions

Over the past two decades, optical fibers have made a major impact on the design and use of astronomical spectrographs. With the continuing development of new types of optical fiber with improved performance, for the telecommunications and fiber sensor markets, it appears likely that this trend will continue for the forseeable future. Extending their range of operation into the thermal infrared using cooled fibers, and combining the advantages of integral field spectroscopy with the efficiency gains of multi-object spectroscopy via deployable mini-integral field units (see SPECTROGRAPHS: INTERNAL FIELD UNITS), are areas currently under active study.

Ray Sharples

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrographs: High-Resolution Spectrographs

Definitions

A **spectrum** (plural: spectra) is the wavelength or frequency distribution of radiation emitted by any radiant source. Visually, it is the familiar band of colors seen through a prism or in a rainbow. When the spectrum is observed by eye the instrument used is a '**spectro*scope***'. If recorded on a photographic emulsion, the instrument is a '**spectro*graph***'. When the photometric precision is high the process is called '**spectro*metry***' and the instrument is a '**spectro*meter***' or '**spectro*photometer***'. The term '*spectrograph*' has been extended to mean any instrument that uses any image detector to record the spectrum.

Spectral **resolving power** or **resolution** is the power of a spectrograph to distinguish between adjacent spectral features, or lines. Numerically, it is the ratio of the wavelength of a spectral line, λ, and the separation $d\lambda$ between two just-resolved lines:

$$R = \frac{\lambda}{d\lambda}.$$

In this article we will deal with the history, design principles, and implementation of *high-resolution astronomical spectrographs*.

The role of spectroscopy in astronomy

Radiation from astronomical sources spans the full spectrum from low-energy long-wave radio emissions to highly energetic hard x- and gamma-rays. Only a narrow band of wavelengths visible to the human eye was accessible to astronomers of the 18th and earlier centuries. With the invention of photography in 1839 and its application to astronomy in 1840 we could detect light into the near ultraviolet and the near infrared. The discovery of radio waves from space, by Karl Jansky in 1933 at Bell Laboratories, revealed the existence of very-long-wavelength radiation from astronomical sources. The far ultraviolet blocked by the Earth's atmosphere was first detected by high-altitude balloon and rocket experiments. Ultraviolet light and x-rays originating in far distant sources are now routinely observed from space-borne telescopes.

Throughout most of history, objects seen in the night sky could only be understood in terms of human interpretation of their visual appearance and motions. With the invention of spectroscopy the physical nature of the stars became a subject of legitimate investigation—thus the term 'astrophysics'. It became possible to determine not only the chemical composition of the stars, but their atmospheric temperatures and pressures as well as their velocities through space and orbital motions about previously unseen companion stars, from which we could measure their masses. The spectrograph could distinguish between spectral features due to separate stars in otherwise unresolved multiple-star systems. The large spectrographic survey known as the *Henry Draper Catalog* revealed that hydrogen—not iron—is the dominant chemical element in the universe. Spectroscopy permitted, for the first time, an unambiguous distinction between the two primary types of *nebulae* (Latin: clouds)—those that are composed primarily of rarefied gases such as the great nebula M42 in Orion, and unresolved stellar systems such as the Andromeda galaxy M31. This gave science its first unequivocal proof that a universe filled with galaxies extends far beyond the borders of our home Milky Way Galaxy. That view was further reinforced by Edwin Hubble and Milton Humason, in 1929–1930, when they discovered the expansion of the universe through spectroscopy of extra-Galactic nebulae.

The next step in understanding the nature of the universe has taken place in the last third of the 20th century, through high-resolution spectrographic revelation of the fine details of spectral lines. Its most recent and famous application has been the detection and measurement of the minute changes in the motions of nearby stars indicating that unseen planets are in orbit about them. High-resolution spectrography is critical to studies of the abundances of specific isotopes of the chemical elements, from which much is being learned of the evolution of the chemical elements in stellar interiors.

History of astronomical spectroscopy

Sir Isaac Newton (1643–1727) founded the science of spectroscopy in about 1665 with his investigations of the nature of light, published in 1704. A century later, in 1802, chemist William Wollaston discovered dark lines crossing the spectrum of the Sun; he thought that these were gaps separating the colors. JOSEPH FRAUNHOFER disputed this assumption when he observed, in 1814–1815, that the solar spectrum is marked with large numbers of dark lines *superimposed on an otherwise continuous gradation* of color. He saw some 600 lines and cataloged the positions of 324 of them (see FRAUNHOFER LINES). He designated the more prominent lines by letters of the alphabet and some of these designations are still in use. In 1859 GUSTAV KIRCHHOFF demonstrated that the dark lines are due to absorption of light of certain wavelengths in an overlying atmosphere of the Sun and he stated his three laws of spectral analysis.

(1) *A luminous solid or liquid emits light as a continuous spectrum.*
(2) *A rarified luminous gas emits a bright-line spectrum.*
(3) *Certain wavelengths are absorbed from the light of a source of a continuous spectrum when passed through a rarified gas, resulting in a dark-line or absorption spectrum.*

An important characteristic of any rarified gas is that its *dark* lines due to law (3) correspond in wavelength to its *bright* lines due to law (2). For the first time in history, science began to realize the possibility of identifying the chemical elements in the Sun, through spectroscopy.

Kirchhoff's work marked the *origin of the science of astrophysics*. In the 1860s he and Robert Bunsen first demonstrated that the chemical elements in the *Sun* are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the *same as those found on Earth*. By 1864 British astronomer SIR WILLIAM HUGGINS identified spectral lines in the *stars* with those of known chemical elements found on Earth, and he discovered that some nebulae have pure emission spectra and must therefore be gaseous. This series of events heralded an understanding of the commonality of the chemical nature of stars everywhere in the universe. Huggins had attempted photographing the spectrum of the bright star Vega, in 1863, but he was unable to photograph the Fraunhofer absorption lines. It was left to HENRY DRAPER (1837–1882) to obtain the first successful stellar spectrograms.

Draper's father, American scientist John W Draper, was among the first to experiment with photography and in 1840 obtained the first photographs of the Moon showing the lunar 'seas', on Daguerreotype plates. He successfully photographed the solar spectrum with a prism spectroscope. Being dissatisfied with the wavelength scale of the Fraunhofer lines, he had a *diffraction grating* ruled on a tiny speculum metal mirror and succeeded in obtaining useful solar spectrograms[1].

Draper encouraged his son Henry to continue these investigations, and in 1858 Henry began construction of an observatory complete with a 71 cm reflecting telescope with silvered glass mirror. He abandoned speculum metal on the advice of Sir John Herschel. On 1 August 1872, Henry Draper obtained the first photographic spectrum of a star—Vega—complete with Fraunhofer lines. Following the death of Henry Draper at age 45, in 1882, his wife and scientific collaborator Anna carried a number of his stellar spectra to Harvard College Observatory, where director Edward Pickering had offered to reduce the plates. This led to a scientific collaboration that eventually resulted in the *Henry Draper Catalog of Stellar Spectra*, which together with the *Henry Draper Extension* cataloged the spectra of over 350 000 stars. This was one of the most important contributions to the science of astrophysics in modern times. Among other discoveries resulting from the *Henry Draper Catalog*, Miss Cecilia Payne in her PhD thesis research (1921–1924) recognized that the strengths of the spectral lines indicated an abundance of hydrogen in stellar atmospheres that was a million times greater than that of the metals, in particular iron. This discovery forever exploded the contemporary belief that iron was the most abundant element in the universe.

In the meantime German astronomer HERMANN VOGEL (1841–1907) was experimenting with photographic spectroscopy at the Potsdam Astrophysical Observatory, where he was director. He was the first to observe radial velocities of stars with sufficient precision to discover spectroscopic binary stars, and with Julius Scheiner he was the first to accomplish the spectroscopic determination of stellar diameters and masses. They proved in 1889, through its radial velocity changes, that the variable star Algol is an eclipsing binary.

A common problem in early photographic spectrography was the difficulty of obtaining reliable exposures. This was blamed on degradation of the quality of the emulsion on the wet collodion plates during long-time exposures and on instability of the image on the plate due to flexure of the instrument. The stability question was solved when Walter S Adams built large stationary prism spectrographs at the *coudé* (French: elbow) foci of the Mt Wilson 1.5 m (1911) and 2.5 m (1925) telescopes. The spectrographs were mounted solidly on the concrete structure of the coudé room, where it became practical not only to use very long focal lengths but also to support large lenses and prisms free of flexure. At the 2.5 m telescope a series of increasingly improved spectrographs culminated in 1934 in the coudé spectrograph of Theodore Dunham, using large diffraction gratings. His was the first use of the Schmidt camera to photograph the very wide angular extent of high-dispersion spectra.

In the meantime work had been proceeding on the art of making finely ruled diffraction gratings. In 1785 Rittenhouse had experimented with a diffraction grating made with 53 *slits* in a width of 1.2 cm. Fraunhofer had ruled a grating with 4000 grooves on a 1.2 cm mirror as early as 1823 and in 1846 Nobert generated a grating of 6000 grooves over a width of 2.5 cm. Lord Rayleigh used Nobert's gratings in his experiments on the physics of light. By 1870 L M Rutherford had made a grating with 35 000 grooves over a width of 5 cm. This was the first diffraction grating to actually equal the resolution of the available glass prisms.

About 1882 Henry A Rowland, at Johns Hopkins University, built ruling engines with which he made 15 cm gratings having more than 100 000 grooves and resolving power exceeding 150 000. This earned him the distinction of 'father of the modern diffraction grating'. Additional ruling laboratories were set up at Mt Wilson Observatory and at the University of Chicago, where gratings were made with up to 15 cm ruled width, giving resolving power up to 350 000. These three laboratories produced the world's supply of large gratings through the 1930s. Michelson's engine at Chicago made one grating of width 24 cm with resolving power 600 000.

Early reflection gratings were ruled in speculum metal, an alloy of tin and copper, with a diamond tool. Speculum metal had been widely used for telescope mirrors (e.g. by Sir William Herschel). It is brittle and somewhat granular and the diamond tools tended to wear out during the ruling, limiting the size of grating that could be made. In 1933 John Strong, at Johns Hopkins University, introduced vacuum deposition of aluminum on glass. This made a soft, smooth surface on which larger gratings could be ruled, enabled by much longer diamond life, and it made for brighter reflection and lower scattered light.

A major loss of efficiency in gratings is due to diffraction of light into multiple orders of interference. Wood, Babcock and others experimented with methods

[1] A diffraction grating is a reflective or refractive surface on which is ruled a multitude of parallel, closely spaced grooves which break up the reflected or refracted light into separate wavelets. Through mutual interference, the directions of propagation of the transmitted rays are redistributed according to their wavelengths.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of controlling these losses. In 1944 they succeeded in controlling *groove shape* by shaping the diamond ruling tool. This allowed directing as much as 60% of the light into a selected spectral order. The term 'blazed gratings' was coined by George R Harrison to describe this effect. By 1949 Harrison was ruling the first gratings working in very high spectral orders and optimized at very large 'blaze angles', a type of diffraction grating he named an '*echelle*'. Typical echelle blaze angles are 60° and greater. He built the MIT B and C ruling engines and, with interferometric control of the screw used to define the groove spacing, these have become the standards of excellence for gratings up to 30 cm ruled width and echelles up to 40 cm. The B engine is now at the Richardson Grating Laboratory of Spectronic Instruments, Inc, and the C is at the National Optical Astronomy Observatories (NOAO).

Theory of high-resolution spectrography

Definition of high resolution

For the purposes of astrophysical research, resolving power is considered 'high' if all the detail in the spectral lines can be distinguished. This can be achieved if the resolution of the spectrograph is somewhat sharper than the widths of the narrowest lines in the source spectrum:

$$R_{\text{spectrograph}} \equiv \frac{\lambda}{\mathrm{d}\lambda} \geq \frac{c}{V}$$

where c is the speed of light (300 000 km s^{-1}) and V is the combined thermal and microturbulence velocity of the atoms constituting the star's atmosphere. For the vast majority of stars, the range of thermal velocities is between 3 and 15 km s^{-1} with microturbulence lying between 1 and 7 km s^{-1}, so that *high* resolving power ranges from 20 000 to 120 000. For cool planetary atmospheres and cold interstellar medium, resolving power of 500 000 to 1 000 000 has been found useful.

Principles of high-resolution grating spectrograph design

There are many excellent optical texts describing diffraction of light by parallel slits. From these we derive the *grating equation*; the parameters are illustrated in figure 1. Light of wavelength λ is diffracted from a reflection grating with d^{-1} grooves mm^{-1} in accordance with

$$\lambda = \frac{d}{m}(\sin\alpha + \sin\beta)\cos\gamma \qquad (1)$$

in which m is an integer $(0, \pm1, \pm2, \ldots)$ giving the *order of interference* and α and β are angles of incidence and dispersion relative to the normal to the grating's ruled surface[2]. The angle γ is the orthogonal component of the angle of incidence and is usually near zero; we will neglect γ in what follows. The *reciprocal angular dispersion* is the derivative of equation (1):

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\beta} = \frac{d}{m}\cos\beta \ (\text{Å rad}^{-1}) \qquad (2)$$

[2] In the sign convention used for these equations, the angles α and β both have the same numerical sign if both incident and dispersed rays lie on the same side of the normal.

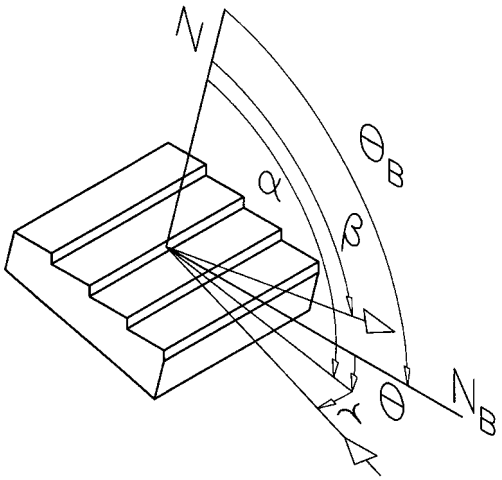


Figure 1. Segment of an echelle grating with blaze angle θ_B, the tilt of the active side of the grooves. Parameters used in the text (equations (1)–(4)) are
- angle of incidence of the rays (the arrow pointing *toward* the echelle), α,
- orthogonal component of the angle of incidence, γ,
- angle of dispersion of the rays (arrow pointing *away*), β,
- distance between the adjacent grooves, d.

since α is normally constant. The reciprocal *linear* dispersion is

$$\frac{\mathrm{d}\lambda}{\mathrm{d}x} = \frac{d}{mF}\cos\beta \ (\text{Å mm}^{-1})$$

where F is the focal length of the spectrograph camera in millimeters and d is the spacing between grooves in Å (1 Å $= 10^{-10}$ m). In most gratings, d is comparable with or somewhat greater than the wavelength of the light; in echelles it can be greater than the wavelength by factors of order 30–100.

A more illuminating relation results if we divide the dispersion equation (2) by the grating equation (1), eliminating m and d:

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\beta} = \frac{\lambda\cos\beta}{\sin\alpha + \sin\beta}. \qquad (3)$$

This equation shows that dispersion by a grating at wavelength λ is a function only of the angles α and β. The spectral order m and the groove width d have been eliminated. Particularly for an echelle grating, m and d are freely selectable to meet any desired geometry in the spectrograph design.

We can rewrite equation (3) to obtain the spectral resolving power:

$$R \equiv \frac{\lambda}{\mathrm{d}\lambda} = \frac{\sin\alpha + \sin\beta}{\cos\beta\,\mathrm{d}\beta}.$$

The image at the detector is the *monochromatic* image of the spectrograph slit, of angular width $\mathrm{d}\beta$. This can be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

expressed in seconds of arc on the *sky* if we incorporate the telescope and spectrograph apertures:

$$R\varphi = \frac{\sin\alpha + \sin\beta}{\cos\beta}\frac{D_{\rm C}}{D_{\rm T}} \times 206\,265. \qquad (4)$$

This is the *throughput equation* for a grating spectrograph. Here, φ is the width of the object in the sky or of the spectrograph entrance slit, in seconds of arc. The constant 206 265 is the conversion factor from radians to seconds of arc; $D_{\rm C}$ and $D_{\rm T}$ are the apertures of the collimated beam and of the telescope. The equation is general and serves for any grating spectrograph. The throughput equation is the critical relationship for the design of any spectrograph in which the resolving power is controlled by the size of the stellar image due to seeing or to the width of the entrance slit, whichever is smaller. The product $R\varphi$ is a useful figure of merit for any spectrograph. Typically, if we select $R = 60\,000$ as the required resolving power then it is useful to design so that $R\varphi = 60\,000''$. In this case the slit is $1''$ wide, letting through most of the stellar image under typical atmospheric conditions. The main parameters of the spectrograph, the size and blaze angle of the grating, are completely determined by equation (4). The remaining tasks are selection of the number of grooves $\rm mm^{-1}$, which controls the free spectral range between adjacent orders, and the focal length of the camera. This should match the image of the spectra to the size of the CCD and the spectral resolution to the pixel spacing on the image detector.

The modern era

New detectors for spectrography

From its first use in stellar spectrography in 1872 until 1974, the dominant detector was the photographic plate. However, with a maximum detective quantum efficiency of only 4%, *96% of all photons that managed to reach the photographic emulsion failed to darken a grain of silver*. Furthermore, in high-resolution spectrography the light of the star image is spread over a large number of pixels, greatly diluting the total photon count per pixel. As a consequence, high-resolution photographic spectrography was limited to long exposures on relatively bright objects, and the use of large telescopes for high-resolution spectrography experienced a lull during the 1960s. At the same time renewed interest in extra-Galactic research fueled by the newly discovered QUASARS and PULSARS took up much of the large telescope time.

A renaissance in high-resolution spectrography began with the completion of the McDonald Observatory's Harlan J Smith 2.7 m telescope and coudé spectrograph in March 1969. These instruments were funded primarily for ground-based support of NASA's planetary research programs. A computer-controlled photoelectric spectrum scanner was added in May 1969 and a high-resolution echelle in 1971. These were applied to research on atmospheres of the planets and stars. With the echelle in double pass, resolving power of 600 000 was achieved in studies of the extremely narrow lines in stellar spectra due to the intervening interstellar medium. With a Data General Nova™ computer in place, the time was ripe for the first application of solid-state image detectors to astronomical spectrography, in 1973. The first device was a Reticon™ linear array of 256 pixels on a silicon integrated circuit chip. These devices have peak quantum efficiencies of about 70%, greatly exceeding those of photographic plates. A low-resolution spectrograph was built (see SPECTROGRAPHS: LOW-RESOLUTION SPECTROGRAPHS), and spectrographic and imaging observations were made with the Reticon on the 0.9 m and 2.1 m telescopes. Development of an intensified Reticon array, which worked by photoelectron bombardment of the silicon chip, followed. This was an image amplification technique to overcome the high read-noise level characteristic of Reticon arrays. The first successful device was put into routine operation in high-resolution spectrography at the 2.7 m coudé spectrograph in 1974. This device had a single linear array of 1024 pixels. The intensified Reticon was dubbed the self-scanned Digicon, following Edward Beaver's pioneering work at San Diego on the parallel-readout Digicon later used on the Goddard High-Resolution Spectrograph (GHRS) on the Hubble Space Telescope. Self-scanned digicons remained in service at McDonald Observatory until 1988. In 1976 a bare Reticon array was put into operation alongside the self-scanned digicon, advantage being taken of the high quantum efficiency of the silicon integrated circuit chip. It exhibited a particularly great advantage in the near infrared, where photoemissive devices have poor sensitivity. The Reticon array became the most widely utilized image detector in high-resolution spectrography at McDonald Observatory until CCDs phased it out in 1992. The Reticon was adopted at many observatories worldwide during the decade immediately preceding application of the first fully successful CCDs.

In the mid-20th century, a wide variety of electronic devices capable of single-photon image detection came under development, largely to meet the needs of astronomical research from space-based telescopes. These included TV cameras, image tubes with photographic output, and a variety of techniques for direct readout of amplified photoelectronic images. Parallel development efforts led, in the 1980s, to availability of high-quality 'science grade' CCDs of large area and high resolution. These activities have led to the nearly complete replacement of the photographic plate in astronomy. In parallel with all this, a renaissance of ground-based telescope construction began in the early 1980s when a series of conferences sponsored by ESO, NOAO and SPIE pointed the way to the 1990s construction boom in telescopes with apertures up to 10 m. *All* of these have included high-resolution spectrographs as part of their initial suite of instrumentation. This has led directly to the first-ever confirmed discoveries of planets outside the solar system—some 29 now known through their gravitational tugs on their parent stars, detected through high-resolution spectrography and precision stellar radial velocities (see EXOPLANETS).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

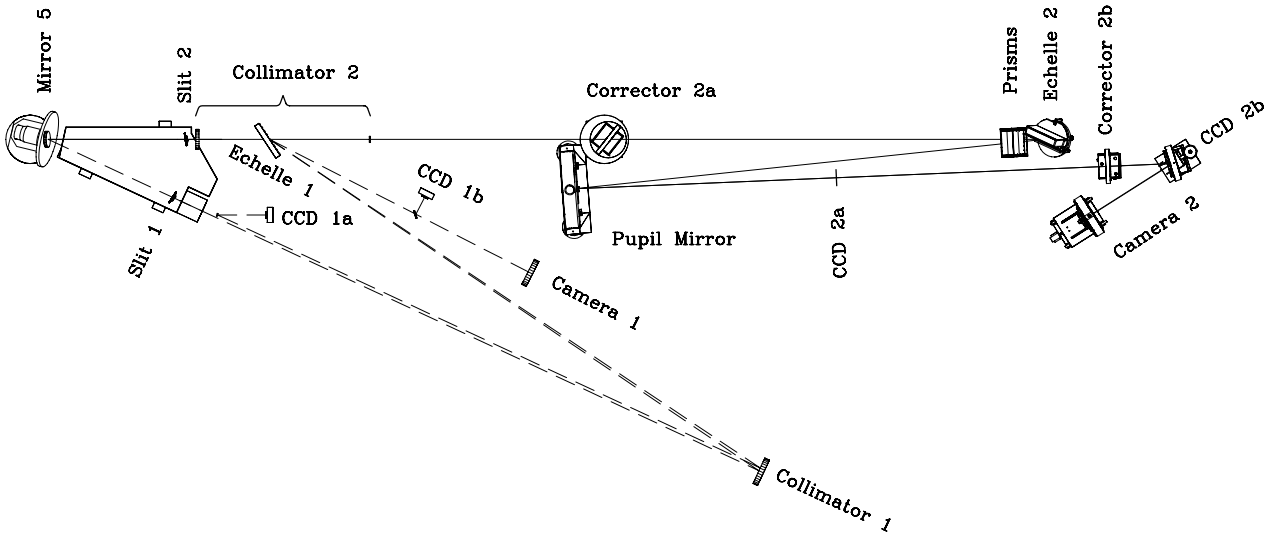


Figure 3. The coudé spectrographs of the 2.7 m H J Smith telescope at McDonald Observatory. ***Optical path 1*** (dashed rays): 2.7 m coudé spectrograph, completed by Robert G Tull in 1969; shown without its photographic cameras, which were removed in 1988. Available foci are as follows: auto-collimated focus of collimator 1 (CCD 1a) ($R_{\max} = 600\,000$), echelle in double pass); Newtonian focus of camera 1 (CCD 1b) ($R_{\max} = 200\,000$). ***Optical path 2*** (solid rays): 2dcoudé spectrograph, completed in 1992 by Robert G Tull, Phillip J MacQueen and John Good. Access is obtained by removing the grating 'echelle 1'. Available foci are as follows: focal surface of the pupil mirror (CCD 2a) ($R_{\max} = 200\,000$); folded Schmidt camera focus (CCD 2b) ($R_{\max} = 70\,000$).

New spectrographs for very large telescopes

Four technological advances characterize the modern high-resolution astronomical spectrograph: (1) electronic image detectors with two-dimensional formats and high quantum efficiency, (2) efficient echelle gratings well matched to detector formats, (3) computer-controlled automation and data retrieval and (4) computerized optical design methods and new optical glasses. The dispersion of light by an echelle is much greater than that of a conventional diffraction grating, owing to its high blaze angle. For that reason the resolving power can be high even though the slit is wide open, yielding greatly improved efficiency. The echelle disperses the spectrum in a narrow ray-fan of width typically 6°–7° containing many overlapping spectral orders. Cross-dispersion with a prism or grating separates these orders and the resulting stack of spectra is imaged on a large-format CCD, exposing many spectral orders simultaneously. These advances have vastly increased overall efficiency, pushing the limits of research to much fainter sources. A typical echellogram is illustrated in figure 2. In some cases the classical Schmidt is not the best camera choice. The CCD may be too large to fit inside a Schmidt camera without excessive blockage of the incident light. Recent developments in optical design combined with a broad range of newly available optical glasses have led to refractive cameras with fast focal ratio and flat, aberration-controlled external focal surfaces well matched to large-area CCDs over a broad spectral range (see CAMERAS). Examples of spectrographs using these techniques will be given here.

Hubble Space Telescope (HST) On 25 April 1990 the HUBBLE SPACE TELESCOPE and GHRS were launched into low Earth orbit aboard the space shuttle Discovery. With a selection of five first-order gratings and one echelle giving resolving powers of 2000–80 000, extensive programs of spectrography were carried out over a period of 7 yr, with observations on stars as faint as 14th magnitude. Two Digicon intensified linear diode-array image detectors, each with 512 photon-counting pixels, served to record the spectrum. The spectral range was in the ultraviolet, a region totally inaccessible from ground-based observatories. Being above the influence of the Earth's atmosphere, the telescope delivered nearly diffraction-limited stellar images to the spectrograph, greatly improving its resolving power and efficiency.

During a service mission in February 1997 the GHRS was removed from the Hubble Space Telescope and replaced by the Space Telescope Imaging Spectrograph (STIS). STIS has a CCD for use in low- and intermediate-resolution spectrography from far ultraviolet into the infrared and two MAMA (multianode microchannel array) imaging detectors, each with 1024 × 1024 pixels which operate at 2048×2048 in high-resolution mode. When used for high-resolution echelle spectrography with the four echelles, these photon-counting detectors have broader ultraviolet spectral coverage per exposure than could be obtained with GHRS, at higher resolving power.

McDonald Observatory 2.7 m H J Smith Telescope A spectrograph completed in 1992 and operating in parallel with the original coudé spectrograph has been named

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

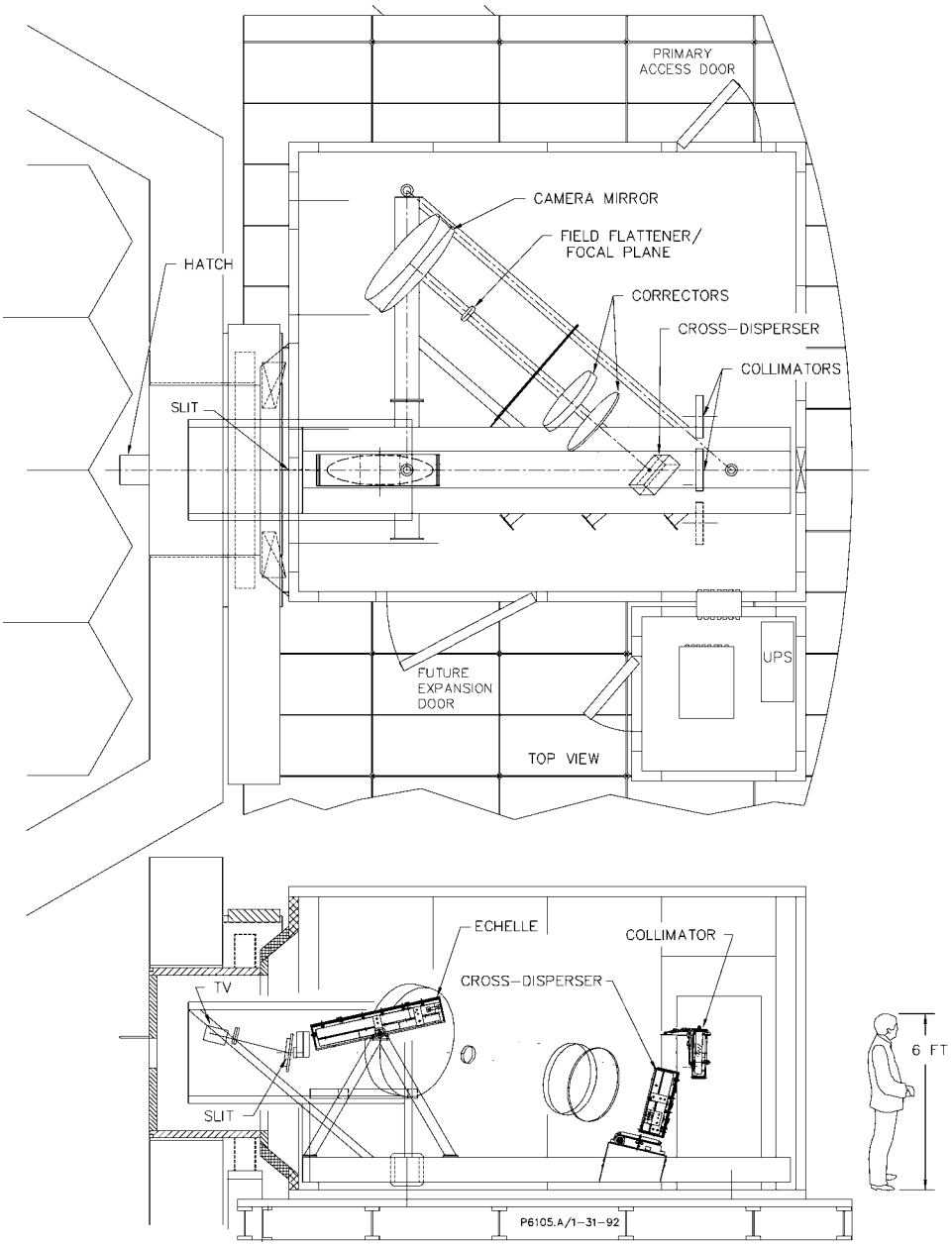


Figure 4. HIRES spectrograph of the 10 m Keck I telescope on Mauna Kea, Hawaii. Its resolving power is about 60 000. Built by a team headed by Steven S Vogt, it came into operation in 1993. Figure reproduced from Vogt *et al* 1994. HIRES: the high resolution echelle spectrometer on the Keck ten-meter telescope *Instrumentation in Astronomy VIII Proc. Soc. Photo-Opt. Instrum. Eng.* **2198** 362, by permission of Steven Vogt.

'2dcoudé' (figure 3). It uses a 79 grooves mm^{-1} echelle with blaze angle 63.9° (R-2) which is interchangeable with a 52 grooves mm^{-1} echelle blazed at 65.3°. Cross-dispersion is accomplished by a train of four 15° fused silica prisms in double pass, directly in front of the echelle. The collimator aperture is 19 cm. The detector is a Tektronix CCD with 2048 × 2048 24 μm pixels. Resolving power at the folded Schmidt camera is 70 000. There is an intermediate focus with resolving power of order 200 000. Stars of magnitude $V = 11.2$ can be recorded at $R = 60\,000$ with an exposure time of an hour at signal-to-noise ratio 100, in average seeing conditions.

Keck HIRES The best-known example of a modern high-resolution spectrograph on a very large telescope is HIRES, built by Steven Vogt on the 10 m Keck I telescope on Mauna

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

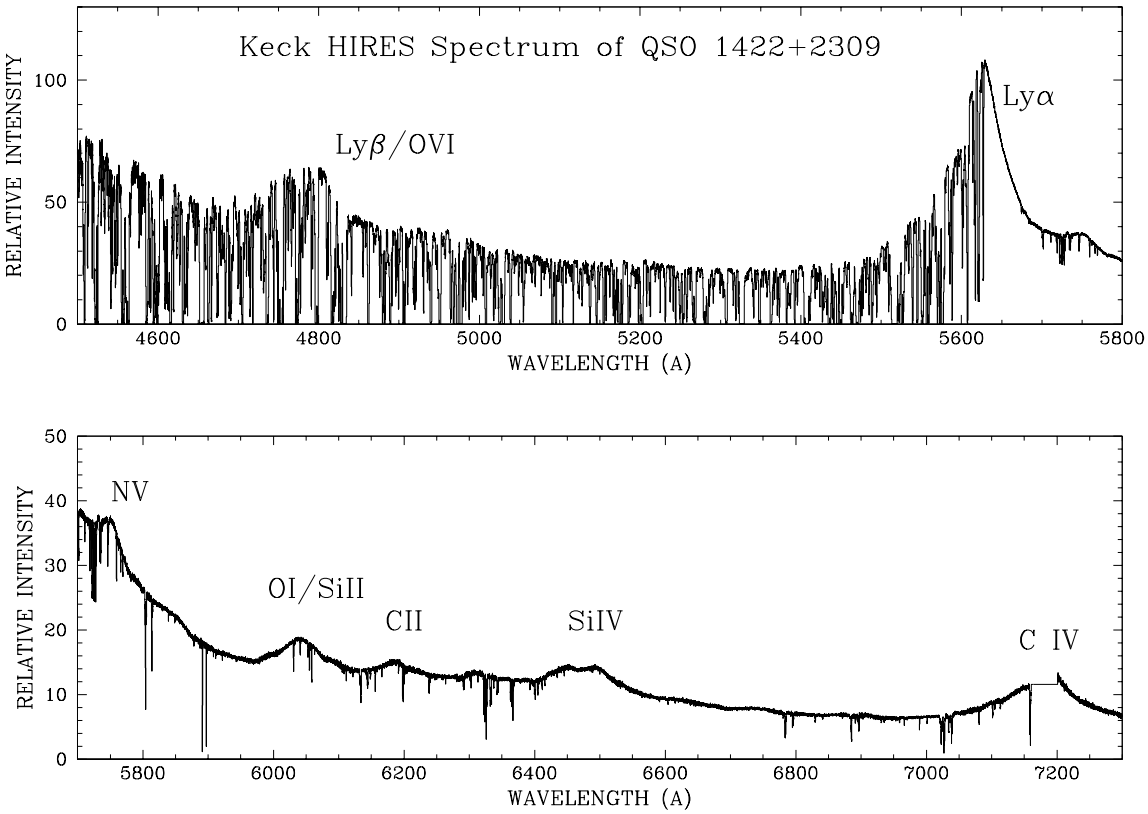


Figure 5. High-resolution spectrum of the quasar QSO 1422 + 2309. This spectrum was obtained in a 7.5 h exposure using the Keck HIRES spectrograph by Wallace L W Sargent. The resolution is 45 000. This figure demonstrates the power of a modern high-resolution spectrograph on a very large telescope. The broad emission line at wavelength 5605 Å is due to the far ultraviolet Lyman α line of hydrogen in this heavily redshifted $Z = 3.61$ quasar (the line is shifted 3.61 times its rest wavelength of 1216 Å). Nearly all the spectral absorption lines to the left (blue) of the emission peak at 5605 Å are due to Lyman α absorption in intervening intergalactic hydrogen clouds at redshifts distributed from 0 to 3.61 by the expansion of the universe. Such a distribution of absorption lines has become known as the 'Lyman α forest'. Note the nearly complete absence of absorption lines with wavelengths greater than 5605 Å. This quasar, with total brightness $V = 16.2$, is among the most distant objects known in the universe. Figure reproduced by permission of W L W Sargent, with thanks to the Keck observatory.

Kea (figure 4). It uses a mosaic of three echelles with blaze angle 70.5° and total ruled width 1.2 m cross-dispersed by a mosaic of two first-order diffraction gratings. The $f/1$ catadioptric camera designed by Harland Epps has a 1.1 m diameter image-forming mirror and 75 cm aberration-control lenses at its entrance aperture. Its 30 cm beam diameter is adequate to ensure minimal light blocking at its internal CCD housing. This spectrograph went into operation in 1993 and quickly demonstrated its power in research on extra-Galactic objects as faint as magnitude 19.6. Its spectra of the 'LYMAN ALPHA FOREST' seen against background quasars are well-known examples (figure 5).

ESO VLT The Ultraviolet–Visible Echelle Spectrograph (UVES) is on a Nasmyth platform of one of the four 8 m telescopes constituting the European Southern Observatory's Very Large Telescope (VLT) on Cerro Paranal in Chile. Built by Hans Dekker and Bernard Delabre, it combines two independent spectrographs, enabling simultaneous exposures in the red and blue ranges of the spectrum. Each utilizes an echelle with blaze angle 75°. The ruled width of each echelle is 84 cm, and each is a mosaic of two replications on a single Zerodur blank. Made at the Richardson Grating Laboratory, these are the largest diffraction gratings currently in existence for visible light spectroscopy. The beam diameter is 20 cm. The two channels are separated by means of a *dichroic beam-splitter* (a mirror that transmits red light and reflects the blue). Each uses a grating for cross-dispersion. Because of the very high dispersion of the echelles, UVES is a highly compact instrument for its resolving power of 120 000. With the excellent seeing conditions at the site in Chile, this spectrograph should see very little loss due to the slit width.

Robert Tull has built a single-channel adaptation of the UVES, called simply HRS. It is at the 9.2 m Hobby–Eberly Telescope at McDonald Observatory. It differs from UVES in that the starlight is conducted to the spectrograph

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

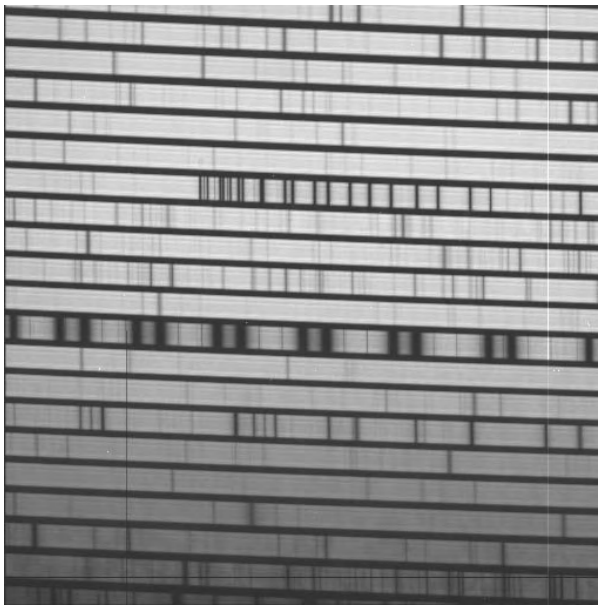


Figure 2. A typical echelle spectrogram: sunlight at the 2dcoudé spectrograph of the McDonald Observatory 2.7 m telescope. The resolving power is $R = \lambda/d\lambda = 200\,000$. The spectrum was exposed on a 2048 × 2048 Tektronic, Inc, CCD. The image covers a fraction of the observable spectrum in echelle spectral orders 56–36 (top to bottom) covering parts of the wavelength region 6126 Å–9550 Å in the red and near infrared. The strong, broad lines in order 45 (near the centre of the figure) and the strong sharper lines in order 50 are the 'A' and 'B' bands of molecular oxygen in the Earth's atmosphere. Atmospheric water vapor produced the strong lines in order 42.

by an optical fiber originating at the telescope focus.

The future of high-resolution spectrography

It is risky to attempt predictions too far into the future. Changes over the second half of the 20th century have been surprising and dramatic; we are not sufficiently wise to out-guess the route technological progress will take 50 yr hence. There are, however, two emerging technologies to watch.

Adaptive optics

Star images are degraded by atmospheric turbulence ('SEEING'), typically to a second of arc or greater diameter. The equations in the section on the theory of high-resolution spectography show that heroic measures must be taken to build spectrographs with large enough slit to accept most of the light and at the same time to maintain high spectral resolution, when used on telescopes as large as 10 m. ADAPTIVE OPTICS, an artificial means to shrink the diameter of the star image, has been proven effective on a number of telescopes, notably the 1 m ESO NTT, the 2.5 m Hooker telescope on Mt Wilson and the 10 m Keck II on Mauna Kea. With adaptive optics, there is hope to concentrate the star image diameter by a factor of 20. This in turn allows design of far more compact diffraction-limited spectrographs, yielding a gain in efficiency at much-decreased cost for the spectrograph.

Direct energy detection

There are under development solid-state image detectors capable of recording not only the time and location of arrival of a photon but also its energy (i.e. its optical frequency). It would be tempting to speculate that some day the energy resolution for such devices could be high enough that a single chip of silicon might replace the entire optical system now needed for high-resolution spectrography.

Bibliography

For a biography and related articles on Henry Draper, the first to obtain successful photographs of stellar spectra, see papers of

Symposium on the Orion Nebula to Honor Henry Draper 1982, Ann. New York Acad. Sci. **395**

The first spectrographic use of the Schmidt camera was at the coudé focus of the Mt Wilson 2.5 m telescope in 1934:

Dunham T Jr 1956 Methods in stellar spectroscopy *Vistas Astron.* **2** 1223–83

A detailed history of the diffraction grating and the first echelle gratings is described in

Harrison G R 1949 The production of diffraction gratings I. Development of the ruling art *J. Opt. Soc. Am.* **39** 413–26

Harrison G R 1949 The production of diffraction gratings II. The design of echelle gratings and spectrographs *J. Opt. Soc. Am.* **39** 522–8

The state of astronomical spectrographs up through the 5 m Hale telescope is summarized in

Bowen I S 1952 The spectrographic equipment of the 200-inch Hale telescope *Astrophys. J.* **116** 1–7

Bowen I S 1955 Astronomical spectrographs: past, present, and future *Vistas Astron.* **1** 400–5

Struve O 1957 Stellar spectrographs *Sky Telesc.* (September) 536–9

Struve O 1957 Stellar spectrographs *Sky Telesc.* (October) 577–80

Pioneering use of the digital computer for astronomical spectrograph control and direct recording of spectra is described in

Wells D C 1972 The computer-controlled spectrometers at McDonald Observatory *Publ. Astron. Soc. Pac.* **84** 203–6

A symposium, 'The Renaissance of High Resolution Spectroscopy—New Techniques, New Frontiers' was held at Kona, Hawaii, June 1983. Selected papers are in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

1983 *Publ. Astron. Soc. Pac.* **85** (September–November)

Included is an excellent review of image detectors used in modern spectrography. Here J G Timothy describes his MAMA detector, as used in the Hubble Space Telescope STIS:

Timothy J G 1983 Optical detectors for spectroscopy *Publ. Astron. Soc. Pac.* **85** 810–34

Robert G Tull

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrographs: Integral Field Units

Most astronomical objects display extended structures. For example, the bulge of a spiral galaxy looks quite different from its spiral arms. Indeed multicolor imaging, which is no more than very-low-resolution spectroscopy, shows us that the disk is bluer than the bulge, indicating that young stars form preferentially within the spiral arms. This, however, does not tell us how stars are moving within the spiral arm and what are the composition, temperature and pressure of the interstellar medium in various parts of the galaxy. Acquiring spectra at different locations will help us in answering these questions. More generally, understanding the chemical composition and kinematics of extended objects relies on obtaining spatially resolved spectrographic information.

The spectrophotometric information content of an extended object is inherently of three dimensions: two dimensions for the sky plane and one dimension in wavelength. Present detectors are unfortunately of a two-dimensional nature only. This makes it a major problem to pack this three-dimensional spectrophotometric information content of celestial objects on the detector plane. This, in turn, explains the still very prevalent use of classical long-slit spectrographs because of their perfect match with available digital detectors (figure 1) and despite their total loss of one spatial coordinate (across the slit width).

For simple objects, one can generally obtain enough information using multiple exposures with judicious slit orientations. However, for a complex object this rapidly becomes difficult as the geometry of the object rarely matches the geometry of the slit, not to speak of the increasing cost in telescope time. In ground-based telescopic observations, the non-simultaneity of the long-slit observations lead to inhomogeneous data as SEEING and weather conditions usually change between two successive exposures. It is then strictly impractical to cover completely and homogeneously a given field of view with a number of long-slit exposures.

Three-dimensional spectrographs are intended to give three-dimensional data for a fully sampled field of view. There are two general classes of instruments: scanning spectrophotometers and integral field units (IFUs).

In a scanning spectrophotometer, time scanning is used to provide the missing third dimension. This type of instrument has a large FIELD OF VIEW, every detector pixel being a spatial element, but its use is restricted to relatively bright objects because of the need for time scanning. At the end of the observation the data cube is obtained by a concatenation of all individual exposures. As for multiple-slit exposures, time of variation of the weather or instrumental conditions may cause inhomogeneities in the data.

The most famous example of a scanning spectrometer is the scanning Fabry–Pérot INTERFEROMETER. It is based on a two-parallel-plates etalon, whose separation can be tuned via a piezoelectric system. Such an instrument allows scanning in wavelength by making successive exposures at different separations of the etalon plates. To scan the complete free spectral range generally requires a large number of exposures (up to 50). The scanning Fabry–Pérot is limited to a small wavelength range, which restricts their use practically to the kinematical study of a single emission line. There are a number of such devices in operation throughout the world, such as Taurus at RGO (Taylor and Atherton 1980), HIFI at University of Hawaii (Bland and Tully 1989) and CIGALE at Marseille (Boulesteix *et al* 1980).

Another approach is to use an imaging Fourier transform spectrometer (FTS). This instrument is a Michelson-like interferometer, which allows the recording of interferograms by a step-by-step scanning of the path difference between the two arms. Increasing the path difference increases the spatial resolution. While there are many FTSs in operation, most of them have only one

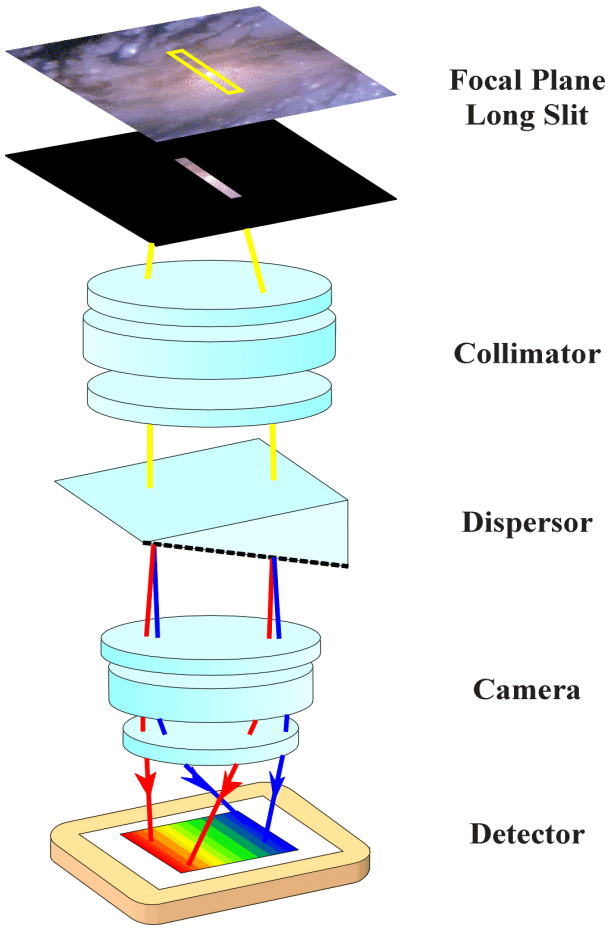


Figure 1. Basic principle of a long-slit spectrograph. A slit aperture, located at the telescope focal plane, masks everything of the observed object except along the slit. The slit orientation matches the orientation of the detector, so that each point of the object along the slit gives a spectrum on the detector. This simple technique avoids spectral overlap because neighboring points are located perpendicularly to the slit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spatial dimension and are generally used for high spectral resolution spectroscopy in the infrared. An exception is the imaging FTS called Bear operated at CFHT (Maillard 1995).

Basic principle of an integral field unit

In contrast to scanning spectrophotometers, IFUs manage to cram the full three-dimensional data in a single exposure on the detector. An IFU is made of two successive stages: the spatial stage whose function is to reformat the field of view and the spectral stage whose function is to disperse and focus the light on the detector. The latter is nothing else than a classical spectrograph that deserves no special comment (but see SPECTROSCOPE/SPECTROGRAPH). The spatial stage is the most critical part. There are currently three types of IFU (figure 2): lenslet unit, fibers unit and slicer unit. They differ in the geometrical arrangement of the spatial elements.

An important parameter of an IFU is the packing efficiency: this is the number of spatial elements multiplied by the length of a spectrum in pixels divided by the total number of pixels of the detector. An ideal IFU will have a packing efficiency of 100%. For example with a detector of 2048^2 pixels, one is limited to 45×45 spatial elements if each spectrum covers a full column or line of the detector. This explains why IFUs usually have a relatively small spatial field of view, even with the large detector format available nowadays. In the real world, however, packing efficiency is limited by the geometrical packing and the optical quality of the spatial and spectral stages.

Lenslet IFU

Lenslet IFUs use an array of lenses to sample the field of view. The light intercepted by each lenslet is focused in a spot called the micropupil which is an image of the telescope pupil. Micropupils are then dispersed by the spectrograph in a conventional manner. The ratio of the lenslet diameter to the micropupil diameter must be large (typically 50). It is this demagnification that saves space on the detector to store the spectral dimension. A slight rotation between the dispersion direction and the microlens array orientation avoids spectral overlap in one dimension. In the other dimension, a wide-band interference filter limits the spectral range to a finite length to prevent overlap. The lenslet size is typically of the order of mm which does not generally match the sampling scale at the telescope focal plane (for example, the lenslet size should be as small as 14 μm to sample 0.1 arcsec at the focal plane of a 3.6 m $f/8$ telescope). An enlarger, preceding the lens array, is thus added to adapt the spatial sampling to the expected spatial resolution. Throughput of this type of IFU is generally good: square or hexagonal lenslet shapes provide a 100% covering efficiency and the lenslet can be made in optical glass and coated. A drawback is that a significant fraction of spectra are truncated if spectra are too long: in practice, the maximum spectral length should be less than 25% of the detector format in the dispersion direction to minimize this effect. The total packing efficiency on the detector is limited by the need to separate each spectrum from its neighbors as neighboring pixels do not share the same wavelength. When using a low spectral resolution grating at the order n, the $n+1$ and $n-1$ orders could cause some overlap with other spectra. In that case using a prism should be considered, especially at blue wavelengths where this effect is more pronounced.

The first lenslet IFU was called TIGER (Bacon *et al* 1995) and has been operated at the Canada–France–Hawaii 3.6 m telescope (CFHT) between 1987 and 1996. Starting in 1997 it has been replaced by OASIS (figure 3): a lenslet IFU with 1600 hexagonal lenses dedicated to the CFHT adaptive optics bonnette. Since TIGER, a few other instruments have been realized with the same design: one for the SAO 6 m telescope (Afanasiev and Sil'chenko 1991), another for the Kyoto three-dimensional spectrograph for the 1.88 m Okayama telescope (Ohtani 1995) and recently the SAURON spectrograph for the 4.2 m William Herschel Telescope.

Fiber IFU

Fiber IFUs use optical fibers arranged in a close-packed bundle at the telescope focal plane and then reformatted into a pseudo-slit, which is then fed into the spectrograph. This presents an advantage compared with lenslet IFUs in the sense that the spectra can be as long as the detector format allows. On the other hand, fiber IFUs suffer from a lower efficiency owing to limited packing efficiency at the entrance (generally less than 75% owing to geometrical loss and cladding) and FOCAL RATIO degradation. The latter is due to diffusion by imperfections within the fiber and diameter variation along it. This effect introduces some light loss and becomes more important with large f numbers. This strongly constrains the designs of the spectrograph and the spatial stage. Fiber IFUs need to be calibrated carefully to control the fiber-to-fiber point spread function and transmission variations.

The first fiber IFU that has been extensively used at a telescope was SILFID at CFHT (1986) (Vanderriest and Lemonnier 1988), later replaced by MOS-ARGUS (1993). Other development took place at the Kitt Peak 3.5 m telescope with DensePak (1988) (Barden and Wade 1988) and at the William Herschel 4.2 m telescope with Hexaflex (1989) (Arribas *et al* 1991), 2D-FIS (1994) (Garcia *et al* 1994) and INTEGRAL (Arribas *et al* 1997). All of these instruments worked in the visible.

Adding lenses in front of the fibers helps to increase the packing efficiency and to solve the problem of focal ratio degradation. Each lenslet forms an image of the telescope pupil at the entrance of each fiber in the bundle. At the output, a linear array of lenslets forms a pseudo-slit fed to the spectrograph. Working examples are the SMIRFS at the 4 m UK infrared telescope (Haynes *et al* 1998) and SPIRAL at the 4 m Anglo-Australian telescope (Kenworthy *et al* 1998). There are a number of fiber IFUs in development for the 8–10 m class telescope (e.g. VIRMOS IFU for VLT, GMOS IFU for Gemini); most of them include lenslets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

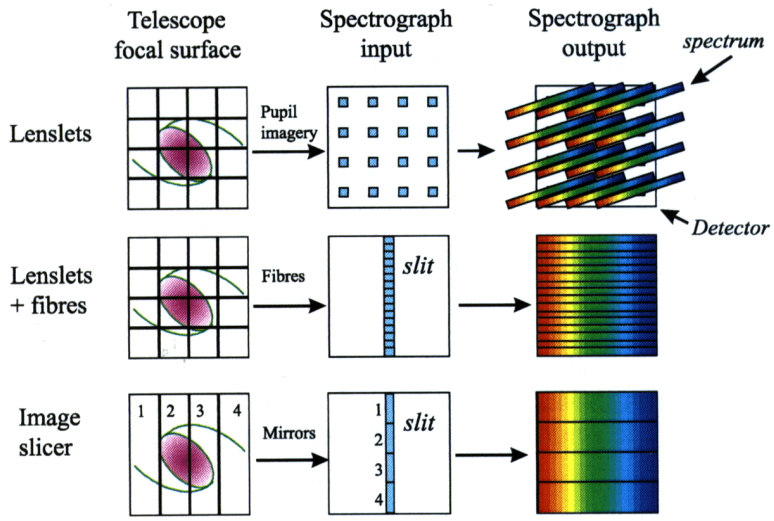


Figure 2. Schematic of IFU types (from Content 1997).

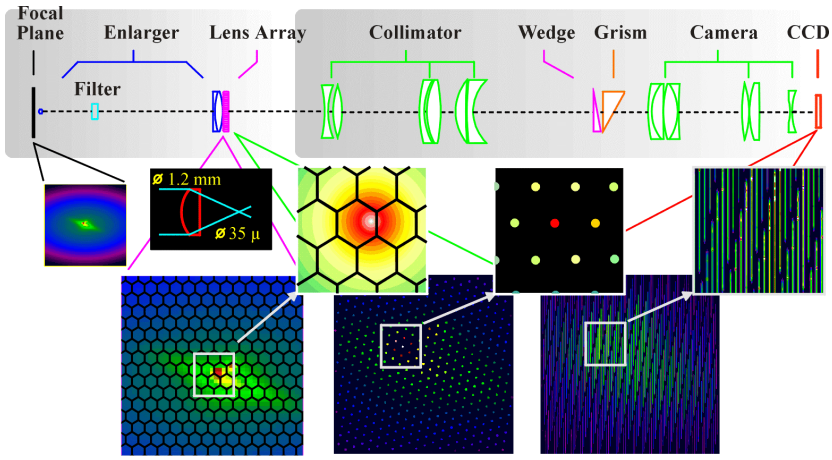


Figure 3. OASIS optical layout.

Slicer IFU

The slicer IFUs use mirrors in the telescope focal plane to cut the field of view into a number of strips which are then rearranged into a one-dimensional long strip. The slicer is composed of two sets of mirrors: the first set slices the field of view in a number of strips and reflects them into different directions while the second set rearranges the strips and aligns them into one continuous long strip. Slicer IFUs are comparable with fiber IFUs in terms of spectral coverage, but they are more efficient in terms of packing efficiency as the spatial sampling is continuous along each strip. The tilt of each submirror produces a defocus between the two edges of each slice. This defocus translates into losses in spatial resolution and throughput. This gives a practical limit to the maximum number of slices that can be accommodated by such an IFU as defocus increases with the number of slices.

The only working slicer IFU in operation today is the three-dimensional instrument built at MPE (Weitzel *et al* 1996). It operates in the near infrared (H and K band) and has been successfully used on a number of 4 m class telescopes.

A variation of this principle which limits the defocus has been proposed by Content (1997) (figure 4): the mirrors have curved surfaces to allow the pupil-relay mirrors to be reduced in size. The tilt of each component is then reduced and consequently the defocus aberrations are decreased.

Applications

IFUs have been successfully used on a large variety of objects, from the study of PLANETARY SURFACES to the most distant QUASARS.

A good example of IFU possibilities is given in the study of the Einstein cross, the gravitationally lensed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

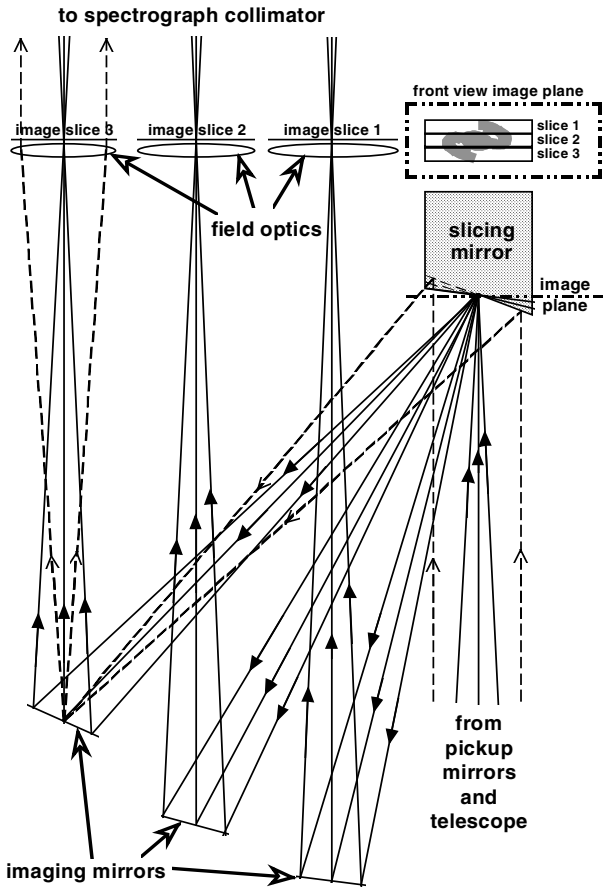


Figure 4. Principle of the slicer IFU (from Content 1997).

system 2237 + 0305. An exposure taken with the IFU TIGER (lenslet type) gave 400 spectra covering the central part of the galaxy with a 0.4 arcsec sampling. The spectra (figure 5, panel A) at the four quasar image locations show a mixture of galactic light and quasar light. One can see indeed the strong line emission of C III] 1909 Å redshifted to 5200 Å as well as stellar absorption lines typical of a galaxy. Integrating the flux over the whole spectral range for each spectrum allows the reconstruction of the equivalent of a direct image within that wavelength range. This image (panel B) shows the well-known cross shape of this object with five maxima, the central one being the nucleus of the galaxy. This image is obviously dominated by the continuum, which is an unknown combination of the quasar and galaxy fluxes. However, it is quite easy to properly subtract the continuum independently for each spectrum. The continuum-free C III] reconstructed image, displayed in panel C, is now free from most of the galactic contribution and it no longer shows the galaxy nucleus. Using simultaneously spectral and spatial information, it was thus possible, in this example, to disentangle the contributions of the quasar and of the galaxy (Fitte and Adam 1994).

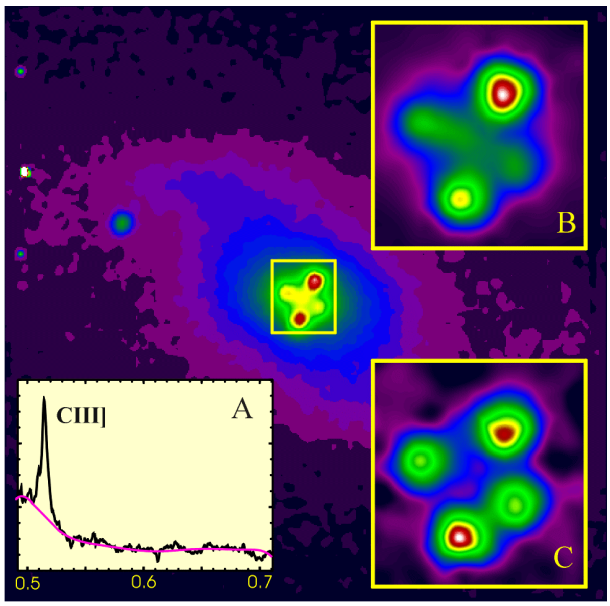


Figure 5. IFU study of the Einstein cross. Panel A: spectrum at one of the quaser image locations. Panel B: reconstructed image with the whole spectral range. Panel C: reconstructed image of the continuum-free C III] emission.

Prospects

Integral field spectrographs in operation are not very common yet, but a number are being designed for the new 8 m class telescopes. In parallel with multiobject spectrographs, which are ideally suited to the study of large collections of objects, IFUs will become more and more popular for the detailed study of individual objects, or alternatively for very dense fields such as GLOBULAR CLUSTERS.

The success of adaptive focus on ground-based telescopes and the launch of large-aperture space telescopes has emphasized the need for high spatial resolution spectroscopy. Decreasing the slit width is no longer a valid solution and IFU capabilities then become essential to understand the physics and the complex morphology generally displayed by astronomical sources at high spatial resolution.

Recently the first three-dimensional detector, with direct measurement of the photon energy, has been tested on a telescope (Peacock *et al* 1998). These superconducting photon-counting cameras have an excellent quantum efficiency over a broad wavelength range (UV to IR). Although their spectral resolutions are still very modest (∼20) and their number of pixels quite small (8 × 8), they have the potential to replace most of the existing spectrographs in the future.

Bibliography

Afanasiev V L, Vlasiouk V V, and Green R F 1995 *ASP Conf. Ser. 71* p 266

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Arribas S, Mediavilla E and Rasilla J L 1991 *Astrophys. J.* **369** 260
Arribas S, Carter D, Cavaller L, del Burgo C, Edwards R, Fuentes F J, Garcia A A, Herreros J M, Jones L R, Mediavilla E, Pi M, Pollaco D, Rees P, Sos N and Worswick S 1997 *Fibre Optics in Astronomy III (Tenerife) (ASP Conf. Ser.)* ed S Arribas, E Mediavilla and F Watson
Bacon R, Adam G, Baranne A, Courtes G, Dubet D, Dubois J P, Emsellem E, Ferruit P, Georgelin Y, Monnet G, Pecontal E, Rousset A and Sayede F 1995 *Astron. Astrophys. Suppl.* **113** 347
Barden S and Wade R 1988 *ASP Conf. Ser.* **3** 113
Bland J and Tully R B 1989 *Astron. J.* **98** 723
Boulesteix J, Georgelin Y, Marcelin M and Monnet G 1983 *Proc. Soc. Photo-Opt. Instrum. Eng.* **5** 445
Content R 1997 *Proc. Soc. Photo-Opt. Instrum. Eng.* **2871** 1295
Fitte C and Adam G 1994 *Astron. Astrophys.* **282** 11
Garcia A, Rasilla J L, Arribas S and Mediavilla E 1994 *Proc. Soc. Photo-Opt. Instrum. Eng.* **2198** 15
Haynes R, Content R, Lee D and Allington-Smith J 1998 *Proc. Soc. Photo-Opt. Instrum. Eng.* **3354** 419
Kenworthy M, Parry I and Taylor K 1998 *Proc. Soc. Photo-Opt. Instrum. Eng.* **3355** 926
Maillard J P 1995 *ASP Conf. Ser. 71* p 316
Ohtani H 1995 *ASP Conf. Ser. 71* p 195
Peacock T, Verhoeve P, Rando N, Erd C, Bavdaz M, Taylor B G and Perez D 1998 *Astron. Astrophys Suppl.* **127** 497
Taylor K and Atherton P 1980 *Mon. Not. R. Astron. Soc.* **191** 675
Vanderriest C and Lemonnier J P 1988 *Instrumentation for Ground Based Astronomy IXth Santa Cruz Workshop* ed L Robinson (Berlin: Springer) p 304
Weitzel L, Krabbe A, Kroker H, Thatte N, Tacconi-Garman L E, Cameron M and Genzel R 1996 *Astron. Astrophys Suppl.* **119** 531

Roland Bacon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrographs: Low-Resolution Spectrographs

Low-resolution spectrographs typically record the spectrum of astronomical objects at resolution ($\lambda/\Delta\lambda$) in the range 10^2–10^4. This resolution is adequate to reveal many strong absorption lines in stellar spectra, or emission lines in nebular spectra, and to resolve the typical range of velocities exhibited within a galaxy, for example, but it is not sufficient to resolve the intrinsic line profiles of solar-type stars or individual interstellar clouds. Low-resolution spectroscopy is particularly applicable to faint objects where photon noise from the source and sky places a practical limit on the resolution which can be achieved while maintaining an adequate signal-to-noise ratio.

Overview

A spectrograph contains four parts: the slit or entrance aperture, a collimator which converts the cone of light emerging from the slit to a parallel beam, a prism or grating which disperses the parallel beam through a range of angles which depends on wavelength and the camera, which focuses the parallel beams into new images of the slit at a corresponding range of positions. The detected intensity as a function of position is then a measure of the spectrum, or intensity as a function of wavelength. The same arrangement is used for spectrographs in the UV, optical and IR portions of the spectrum.

The development of optical spectrographs through the mid-1960s was driven largely by the characteristics of photographic plates, which were the typical detectors of the time (see DETECTORS: PHOTOGRAPHY). The photographic speed of the spectrograph camera was critical, so that the image of the spectrum would be bright enough to produce a useful photographic exposure in a reasonable period of time. Low-dispersion instruments with particularly fast CAMERAS were called nebular spectrographs because of the low intensity of nebular images compared with stellar images. Even for nebular spectrographs the limiting photographic magnitude was sufficiently bright that the effect of sky background was modest.

The advent of more modern detectors, initially image intensifiers and more recently CHARGE-COUPLED DEVICES (CCDs), has resulted in a gain of more than 100 in the effective limiting brightness of low-dispersion spectrographs. One consequence of this change is that modern spectra are often dominated by sky background, and the spectrograph must be designed to maximize the contrast of the object against sky and to provide accurate calibration and subtraction of the sky background.

Most modern optical spectrographs being built for large ground-based telescopes are imaging spectrographs. The collimator and camera have a significant field of view, typically equivalent to several arcminutes on the sky. Without a disperser the detector can make a direct image of the spectrograph field of view. The scale of the final image is reduced by the ratio of the focal length of the collimator to the focal length of the camera. Often this scale results in a better match between the image size and the pixel size, and a larger field of view, than would be obtained with the same detector placed directly in the telescope focal plane. Within the field of view there may be many objects of interest brighter than the limiting spectroscopic magnitude. In place of a single slit, an aperture plate can be inserted with a number of small slits cut at the locations of interesting objects (see SPECTROGRAPHS: MULTISLIT SPECTROGRAPHS). When viewed through the disperser, the image contains a spectrum for every slit. Care must be taken in placing the slits so that the dispersed spectra do not overlap. In some cases even larger fields of view are covered by employing several spectrographs, each of which images a separate portion of the telescope focal plane in what is sometimes called a fly's-eye arrangement.

An alternative approach to multi-object spectroscopy is to use a number of optical fibers, positioned in the telescope focal plane at the locations of interesting objects. The other ends of the fibers, which are flexible, can be lined up in place of a long slit in an imaging spectrograph, producing a separate spectrum for each object (see SPECTROGRAPHS: FIBER-FED SPECTROGRAPHS). In some cases the size of the field accessible to fibers can be much larger than the size of the field for a multi-slit spectrograph.

When observing an extended object, a conventional slit spectrograph produces a one-dimensional series of spectra, spatially resolved in the direction along the slit. In some cases it is desirable to extend the spatial sampling to two dimensions, in order to cover the entire area of the object. This can be accomplished by using an array of closely packed fibers, a modified optical slit assembly called an image slicer, or an array of field lenses which spatially re-sample the image. Such a spectrograph is called an integral-field spectrograph (see SPECTROGRAPHS: INTEGRAL FIELD UNITS).

Detailed considerations

The slit serves two purposes. (1) Sky background adjacent to the object (and the associated photon noise) is rejected by the slit jaws. (2) In the case where the slit is narrower than the telescope image, the wavelength resolution in the final spectrum is determined by the width of the slit rather than by the distribution of light across the source. Single-object slits are sometimes made from two jaws with a mechanical adjustment for changing the slit width, or else by machining or etching a number of discrete slits with a selection of widths. In single-object spectrographs it is common to polish the surface of the slit jaws and tilt the slit at an angle to the incident beam, so that an auxiliary TV camera can be used to view the reflected image of the telescope focal plane in order to check the placement of an object on the slit. Multiple slits for imaging spectrographs have been made using multiple mechanical slit jaws, by machining custom aperture plates using mechanical tools or high-powered lasers, or by making photographic images on very fine-grained emulsions.

Typical collimated beam diameters for low-dispersion spectrographs are in the range 5–20 cm. Larger diameter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is achieved by increasing the focal length of the collimator, resulting in higher resolution for a given slit width because the range of angle subtended by the slit decreases. Larger size also makes the optical design problem more tractable by reducing the angular field of view required of the collimator in order to image a field of a given diameter, and of the camera in order to illuminate a CCD array of a given size. Cost and mechanical flexure, however, increase rapidly with size, and the available rulings for DIFFRACTION GRATINGS become increasingly limited.

Usually the prism or grating is placed near the exit pupil of the collimator, where an image of the primary mirror is formed by the collimator optics. For a single-object spectrograph the small entrance aperture tends to form a pinhole image of the primary at a range of positions along the parallel beam. However, for an imaging spectrograph with a wide-angle field of view, the location of the exit pupil is critical. Reflecting collimators have the disadvantage that the exit pupil is reflected back toward the slit, and an off-axis or folded design must be used in order to extract the collimated beam to an accessible location. Refractive collimators have the advantage of more easily producing an external pupil, but must contend with chromatic aberration which is absent in reflecting designs. In a wide-field imaging spectrograph the design of the collimator can be considerably more challenging than the design of the camera, because the required control of the pupil position does not permit the arbitrary use of a field flattener in the collimator image plane (see also WIDEFIELD CCD IMAGERS).

The disperser can be either a prism or a diffraction grating. Prisms have the potential advantage of very high efficiency but suffer from a number of disadvantages. The dispersive power of a prism is relatively low compared with a diffraction grating, and dense glasses with the highest dispersion do not transmit well at the blue end of the spectrum. The dispersion of a prism is highest in the blue while most astrophysical sources are brighter in the red, resulting in a very uneven exposure. The optical homogeneity of glass is difficult to maintain through a sufficiently long optical path, and reflection losses at strongly oblique surfaces can also be difficult to control.

For these reasons most spectrographs make use of either transmitting or reflecting diffraction gratings. The gratings must be blazed in order to concentrate the diffracted light efficiently into a single diffracted order. Transmission gratings are blazed by cutting each groove into the face of the grating at an angle so as to create a tiny prism which refracts light at approximately the angle of diffraction. Reflecting gratings are blazed by cutting each groove at an angle so as to form a tiny mirror which reflects light at approximately the angle of diffraction. Because refraction for typical optical materials is weaker than reflection, the angular dispersion achievable with transmission gratings is a factor of two or so lower than with reflecting gratings.

When using a diffraction grating to cover a wide range of wavelengths, care must be taken to avoid overlapping spectra from different orders of diffraction. An order-separating filter may be required. In some cases a dichroic mirror is used to direct red and blue wavelengths into two separate spectrographs. Each side of such a double spectrograph can be optimized for maximum performance over a more limited wavelength interval. Sometimes a weak cross-disperser is introduced in order to separately detect spectra in several low orders of diffraction. Such a spectrograph is said to be working in echellette mode (see SPECTROGRAPHS: HIGH-RESOLUTION SPECTROGRAPHS).

In a spectrograph which uses transmission gratings the optical axis will be bent through some angle which depends on the grating being used. In order to avoid having to pivot the camera of the spectrograph to various angles (in particular to allow the camera to be fixed in the straight-through position where a direct image of the telescope field can be made by removing the grating) the transmission grating is often replicated on a prism which compensates for the diffraction angle of the grating. The combination is often called a grism.

A variation of the reflecting grating which can achieve still higher angular dispersion illuminates the ruling through the substrate. This allows the diffraction to occur in the resin with which the grating was replicated, rather than in air. The wavelength is shorter in the resin in proportion to the index of refraction, and the angular dispersion is correspondingly higher. The disadvantage of such immersion gratings is that the reflective coating is typically less efficient when illuminated from inside the material than it is when illuminated from air.

Typically some angle is introduced between the incident and diffracted beams of a reflecting grating spectrograph, in order to separate the collimator and camera optics. Order zero of a reflecting grating always emerges at the angle of reflection (for this order the grating is acting as a mirror). Diffraction of increasing order can occur on either side of the zero-order image. Orders which emerge on the same side of the zero-order image as the incident beam are called interior orders. Orders which emerge on the opposite side of the zero-order image are called exterior orders.

A critical property of such a spectrograph is the anamorphic factor, r (for a diffraction grating, the ratio of the cosine of the angle of incidence to the cosine of the angle of diffraction). The anamorphic factor can become substantial for gratings with high angular dispersion. For interior orders the anamorphic factor is less than 1 and the angular width of the slit is reduced in the diffracted beam by the factor r, but the physical size of the diffracted beam is increased (in one direction) by $1/r$. Interior orders have the advantage that the wavelength resolution is higher because the projected slit is narrower, but the optics must be larger in order to accept all of the diffracted beam, which can make the optical design problem significantly more challenging. Exterior orders are subject to light loss on the back facet of each groove and are generally used only under special circumstances.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The spectrograph camera can be either reflecting (typically using Schmidt-type optics) or transmitting. The reflecting designs again suffer from the difficulty of separating the image plane and the pupil plane. Prime-focus Schmidt cameras require a special dewar to suspend a cooled detector in the middle of the optical system. Folded Schmidt cameras use a flat mirror to extract the final image to an external location. In either case, optical performance is limited by the field flattener which is required in order to match the curved focal plane to a flat detector. Cassegrain Schmidt designs produce a flat focal plane, but the light loss caused by obscuration on the secondary mirror can be substantial. Refracting cameras can produce wide, flat fields, but must contend with chromatic effects which usually require the incorporation of special low-dispersion glasses or other low-dispersion materials, such as calcium fluoride.

The entire optical assembly of the spectrograph must be mounted in a mechanical framework with adequate structural and thermal stability, in order to minimize drift of the spectrum on the detector. Drift causes a loss of spectral and spatial resolution, as well as changes in the intensity calibration of the detector. This is especially true for thinned CCD arrays which are subject to high-frequency fringing effects caused by optical interference within the thickness of the array. The intensity calibration is particularly critical when subtracting sky determined at one location along a slit from the spectrum of a faint object at another location. For large spectrographs it can be particularly difficult to achieve adequate stability, and various measures have been considered, including active optical compensation and distributed mechanical flotation.

Recent developments

The most ambitious low-resolution spectrographs built in recent years have aimed to maximize the advantages of multi-object spectroscopy. Fiber systems capable of observing from 300 to 600 objects at one time, in some cases incorporating multiple spectrographs or double (dichroic) spectrographs, have been constructed, typically in order to measure galaxy redshifts and large-scale structure, to survey faint quasars and to study stars in the halo of the galaxy. Multi-slit systems are under construction incorporating multiple spectrographs or large CCD mosaics, generally with the intention of studying the evolution of more distant samples of galaxies.

A new type of diffraction grating, called a volume-phase holographic grating, has begun to attract attention. These gratings work by introducing periodic high-frequency modulations of the index of refraction in a layer of suitable transparent material. Volume-phase gratings can be somewhat more efficient than surface-relief gratings, and may be especially useful if they can be produced in larger sizes than conventional diamond-ruled gratings.

Bibliography

The classic paper on spectrograph design from the photographic era is

Bowen I S 1962 Spectrographs *Astronomical Techniques* ed W A Hiltner (Chicago, IL: University of Chicago Press) pp 34–62

A useful textbook is

Schroeder D J 1987 *Astronomical Optics* (San Diego, CA: Academic)

The best source for detailed information about modern instrumentation is the continuing series of proceedings from the periodic conferences *Instrumentation in Astronomy*, sponsored and published by the SPIE.

Stephen Shectman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrographs: Multislit Spectrographs

A multislit spectrograph is similar to a standard-slit spectrograph, except that the FIELD OF VIEW on the sky is greatly expanded to allow placement of entrance slits on many objects in the field.

This instrumental field of view on modern large telescopes is typically 5–20 arcmin and a considerable advantage is gained over single-slit spectrographs when there are many possible targets in this size of field.

Access to even larger fields of view (for example several degrees) is limited by the cost and design complexity of large optical elements; in this regime the preferred solution is to use fiber-fed spectrographs which use fibers to reformat light into a smaller entrance slit (see SPECTROGRAPHS: FIBER-FED SPECTROGRAPHS). In contrast in a multislit spectrograph the contiguity of the image of the full field is preserved throughout the optical train, although it is dispersed in wavelength. This of necessity involves greatly expanding the size of the optics, even when working at a low spectral resolution, to accommodate the size of the beam. Multislit spectrographs are also sometimes referred to for this reason as 'beam-fed spectrographs' to distinguish them from fiber-fed systems.

The particular niche occupied by multislit spectrographs in astronomy is the spectroscopy of very faint objects in crowded fields, especially the measurement of REDSHIFTS of very distant galaxies. By retaining the use of a slit which measures the background adjacent to the object much better subtraction of night sky emission can be obtained than by using fibers; this is critical because targets are usually much fainter than the sky. Because of the orientation towards faint targets simple designs are generally employed with the minimum number of optical elements in order to maximize the throughput of light through the instrument. A direct imaging capability is also common which gives rise to a very versatile instrument; because of this a multislit spectrograph is almost a standard facility on the largest telescopes.

Modern examples of multislit spectrographs are the Low Dispersion Survey Spectrographs (LDSS-1 and LDSS-2) on the Anglo-Australian and William Herschel 4 m telescopes and the Low Resolution Imaging Spectrograph (LRIS) on the W M Keck 10 m telescope. These have all been used for progressively fainter and more distant redshift surveys; the current state of the art is redshift 4–5 galaxies and quasars from LRIS.

Optical layout

The typical layout is shown in figure 1 which illustrates the LDSS spectrograph on the Anglo-Australian Telescope.

The focal plane of the telescope is occupied by a *slit mask*; this is a piece of opaque material (usually metal or carbon fiber) in which multiple slits have been cut. The positions of the slits on the mask correspond to the relative positions of the images of the targets formed at the telescope focal plane. A provision is made for swapping slit masks and removing them completely to give a clear aperture for imaging (for LDSS-2 this is accomplished by mounting them on a wheel).

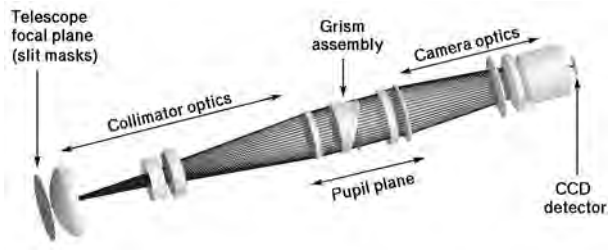


Figure 1. Optical diagram of the *Low Dispersion Survey Spectrograph* (LDSS), a multislit spectrograph on the Anglo-Australian 3.9 m telescope. A collimator re-images the telescope focal plane (where the slit masks are located) into a pupil which accommodates a dispersing element (a grism) and/or filters. The dispersed spectra are then re-imaged by the camera optics on to the CCD detector.

A requirement for observing is to prepare the slit masks in advance; various technologies have been used for making slits such as photo-etching, cutting with a mechanical drill bit and laser cutting.

Next in the optical train comes a *collimator* which is usually the largest optical element. This usually consists of several lens elements which takes light from each point in the focal plane and produces parallel rays (for more details see SPECTROSCOPE/SPECTROGRAPH).

Behind the collimator these rays form a real pupil (where there is an image of the telescope entrance aperture) and at this point in the optical train are put optical filters for wavelength range selection and diffraction gratings to disperse the light into spectra. The LDSS case uses a *grism* which consists of a prism bonded to a transmission grating; the prism refracts the spectra so they stay on axis. Thus removing and inserting the grism into the pupil allows one to instantly swap between imaging and spectroscopy of the full field, a very convenient feature.

After the pupil comes the *camera*, another optical element assembly which focuses the spectra (or direct images) onto a second focal plane where a detector of some kind is placed. Usually the camera includes some kind of focal reduction to shrink the image of the field down on to a small detector.

The standard modes of operation are imaging in various filters and slit spectroscopy by inserting a mask and a grism and/or grating. It is also possible to do 'slitless spectroscopy' by just using the grism and/or grating with a clear aperture; however, in this case the background emission is greatly enhanced so it is not commonly used for faint targets.

Design variations

Many variations are possible. For example, LRIS uses reflection gratings rather than grisms. In this case to swap

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

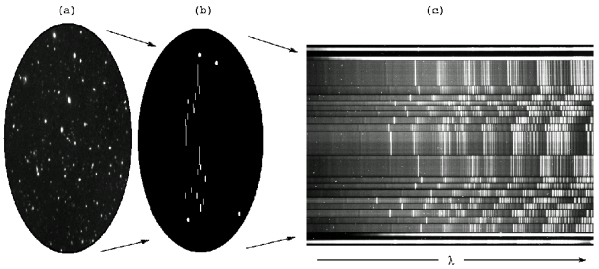


Figure 2. Stages of image formation in a multislit spectrograph (data shown are for LDSS-1). (*a*) An image of the sky is formed at the telescope focal plane. (*b*) A slit mask selects objects for spectroscopy via position in the focal plane. (*c*) Each slit in the mask forms a spectrum of object + sky on the detector. (Note that the holes at the top and bottom of the mask are for alignment of the mask on reference stars in the field; this is done in direct imaging mode.)

between imaging and spectroscopy one must remove the grating and insert a flat mirror. LRIS also uses a reflective collimator mirror; it would also be possible in principle to also have a reflective camera. Another variation could be to have multiple channels; for example the beam could be split by a dichroic after the pupil and sent to multiple cameras (to optimize for blue and red wavelength range).

Data and reduction

Images are formed on a large detector such as a CHARGE-COUPLED DEVICE (CCD). Figure 2 shows a typical raw data image, illustrating spectroscopy through a slit mask. Each slit forms a rectangular spectrum, its position on the detector corresponding to the position of the slit on the mask and hence the object on the sky. This leads to a design constraint in the making of the mask: objects must be chosen for slits so that their resulting spectra do not overlap. This geometric effect can be seen in the figure.

In a typical setup with faint objects either unresolved or a few arcseonds in size a slit would be about 10 arcsec in size. Thus with a typical 10 arcmin field of view up to 60 spectra could be obtained in a tier. If the spectra are short compared with the detector multiple tiers can be used giving an increased multiplex advantage. One way of achieving this is to use a bandwidth limiting filter to reduce the wavelength range of the spectra; another is to simply observe at a very low resolution so even the full spectrum is short.

For data reduction the individual slit spectra must be located, then they can each be reduced as a single-slit spectrum (although each will have a wavelength scale offset by the departure of the slit from the field center in the direction of the dispersion). As for a single slit the background is subtracted by interpolating the data values either side of the object on the slit. Some example spectra from the LDSS-2 redshift survey (Glazebrook *et al* 1995) are shown in figure 3.

Technological developments

Multislit spectrograph technology is far from static and several developments are actively being pursued.

Infrared instrumentation

The same optical principles can be applied to infrared instruments. To operate out to a wavelength of 1.6 μm the only fundamental change required is the use of an infrared sensitive detector (see also INFRARED ARRAYS). To go redder than this it then becomes necessary to cool the entire instrument in a dewar to avoid too much thermal radiation, especially from the mask. This causes a problem because masks need to be accessed to be changed, which necessitates removing them from the dewar. A proposed solution to this is use of a separate, smaller, fore-dewar just for the masks and which can be thermally cycled much more quickly. Thermal radiation from outside the telescope aperture must also be blocked; this is done by placing a cold aperture stop in the pupil space. Multislit spectrographs based on this concept are being built for the Anglo-Australian and Kitt Peak 4 m telescopes and are being proposed for the new large telescopes.

Microslits

The number of objects one can fit on a mask is limited by the size of the slits. If they can be made smaller the multiplex can be increased; ideally one want small apertures (often called 'microslits') which are just big enough to cover the objects.

This leaves the problem of how to do the sky-subtraction. In principle one can take separate sky observations through the slits. One problem with this is that the sky brightness can change in minutes while with a CCD one must take much longer exposures to overcome noise. Research in progress at the Anglo-Australian Observatory has demonstrated that this limitation can be overcome via a special CCD controller mode known as 'charge shuffling' where spare pixels on the detector are used to store the sky signal. Using microslits fields of up to 200 objects have been observed at once with large spectral ranges. This is a powerful technique for very dense fields.

Larger fields

Going to larger fields than the normal limit of about 10 arcmin requires fabrication of large and expensive optical components. Nevertheless there are proposals to do just that. For example the US National Optical Astronomical Observatories has designed an 8 m telescope with a 1.5° field of view which will feed four large multi-slit spectrographs.

Future mask technologies

The flexibility of the multislit design has led to the desirability of having a multislit spectrograph on the Next Generation Space Telescope (NGST) due to be launched in 2008 (see SPACE INSTRUMENTATION: NEXT GENERATION SPACE TELESCOPE). However, for a space instrument which will never be serviced after launch there is a problem in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

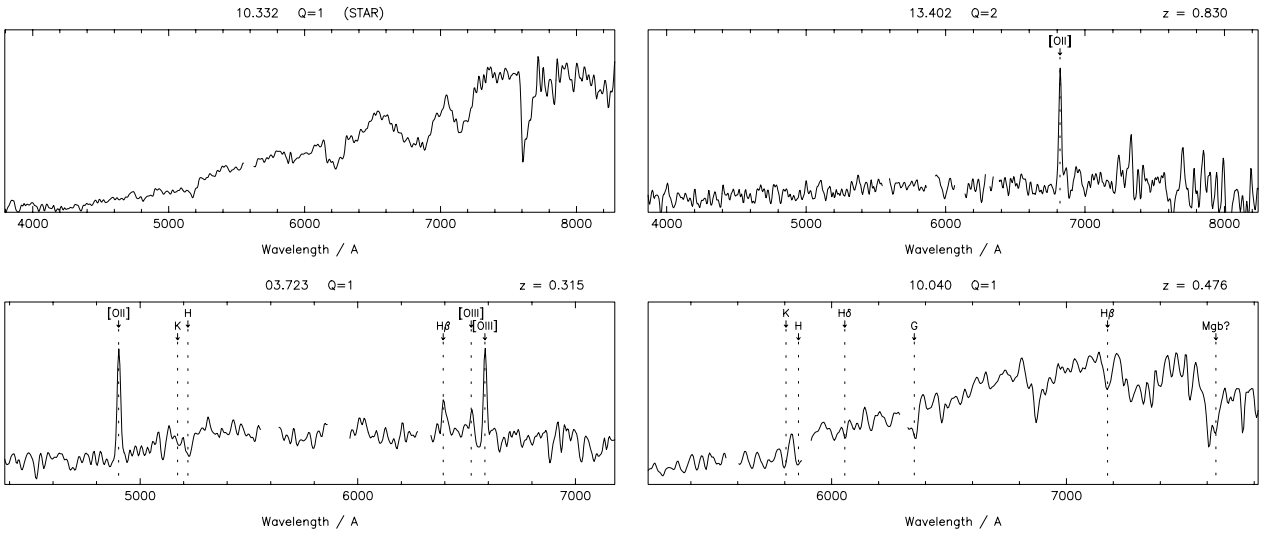


Figure 3. Example spectra of faint stars and galaxies from the LDSS-2 redshift survey (Glazebrook *et al* 1995).

provision of slit masks; because of this new technologies are being researched to provide mechanically *configurable* slit masks, which can be operated remotely.

Two such technologies are being developed at the Goddard Space Flight Center using microelectromechanical devices which are arrays of mechanical devices made on a silicon substrate. One possibility is a 'micromirror array': a 4096×4096 array of tiltable mirrors is envisioned which could move between an ON position, which would direct light into a spectrograph, and an OFF position which would be virtually opaque. Every position in the array would be rapidly configurable to form a series of slits. A related concept is 'microshutter arrays'; these are arrays of small shutters which open and close, operating in transmission like a conventional slit mask. At the time of writing the size of test arrays is of order 3×3 elements.

A more conventional proposal for NGST, from the Dominion Astrophysical Observatory, is to use a stack of mechanical slits, up to 75, which are scrollable side to side on metal rolls. Similar ideas have been implemented before on the ground; for example the 'multislit mechanism' on the Multiple Mirror Telescope in Arizona provides eight configurable slitlets in a 3 arcmin field of view.

Bibliography

A detailed description of the LDSS-1, LDSS-2 and LRIS spectrographs can be obtained from the following journal publications:

Allington-Smith J R, Breare J M, Ellis R S, Gellatly D W, Glazebrook K, Jorden P R, MacLean J F, Oates A P, Shaw G, Tanvir N, Taylor K, Taylor P, Webster J and Worswick S 1994 LDSS-2: A low dispersion survey spectrograph (LDSS-2) for the William Herschel Telescope *Publ. Astron. Soc. Pac.* **106** 983

Oke J B, Cohen J G, Carr M, Cromer J, Dingizian A, Harris F H, Labrecque S, Lucinio R, Schaal W, Epps H and Miller J 1995 LRIS: the Keck low-resolution imaging spectrometer *Publ. Astron. Soc. Pac.* **107** 375

Wynne C G and Worsick S P 1988 LDSS-1: a low dispersion survey spectrograph *The Observatory* **108** 161

A detailed description of instrument proposals for the *Next Generation Space Telescope* including the latest research on micromirror and microshutter arrays is best obtained from the NGST web site

http://ngst.gsfc.nasa.gov

The LDSS-2 redshift survey is described in

Glazebrook K, Ellis R S, Colless M M, Broadhurst T J, Allington-Smith J R and Tanvir N R 1995 *Mon. Not. R. Astron. Soc.* **273** 157

Observations of high-redshift galaxies with LRIS are described in

Steidel C C, Giavalisco M, Dickinson M E and Adelberger K L 1996 *Astron. J.* **112** 352

Karl Glazebrook

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectroheliograph/spectrohelioscope

An instrument that allows the Sun to be studied in light of one particular wavelength (by imaging the Sun photographically, or electronically, in the case of a spectroheliograph, or by direct vision in the case of the spectrohelioscope). Basically, the instrument consists of a spectrograph, or spectroscope, which is used to isolate one particular spectral line (e.g. one of the hydrogen or calcium lines), combined with a mechanical or optical means of scanning the disk of the Sun. In the case of a spectrohelioscope, provided that the solar disk, or the region of interest, can be completely scanned at least ten times a second, persistence of vision allows the observer to see a stationary monochromatic (single-wavelength) image. The instrument enables features such as prominences and filaments to be studied

The amount of detail that can be seen increases as the width of the slit is decreased. By isolating different parts of a particular spectral line (a spectral line has a finite width) an observer can observe to different depths within the solar atmosphere.

See also: filament, prominence, spectroscope/spectrograph, Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectroscope/spectrograph

A device that enables the spectrum of a light source to be viewed directly (spectroscope) or recorded by photographic or electronic means (spectrograph). A basic spectrograph consists of the following principal elements: a narrow slit, through which light is admitted to the instrument, a collimator, a dispersing element, an imaging system, and a detector. Light from the source (for example, a star) falls on the slit, which is placed at the focal plane of the telescope. Light diverging from the slit is brought into a parallel beam by the collimator which, like the imaging element, may consist of lenses or mirrors. The beam of light then encounters the dispersing element, which separates the light into its constituent wavelengths, or colors. The dispersing element may be a prism (or series of prisms) or a diffraction grating (a finely ruled grating which transmits or, more usually, reflects light and splits it into its constituent wavelengths). A system that uses reflecting optics and a reflection grating permits the study of wavelengths that cannot pass through glass. Finally, the imaging system focuses the spectrum (the spread-out band of wavelengths) onto the detector, which may be a photographic emulsion, an image tube or a CCD. In a spectroscope used for visual observation, a small telescope, with an eyepiece, is placed after the dispersing element.

The focused spectrum consists, in essence, of a very large number of overlapping monochromatic images (each corresponding to a different wavelength of light) of the entrance slit. In order to produce a sharply focused spectrum, the slit should be as narrow as possible. However, the narrower the slit, the fainter the spectrum, and the longer the exposure time that is needed in order to record the spectrum. Because the spectrum of a point-like source such as a star would consist of a thin strip of light, too narrow to analyze effectively, a further optical component, called a widener, is usually placed between the slit and the collimator. This broadens the spectrum at right angles to its length. Alternatively, the observer can achieve the same effect by moving the telescope during the exposure so as to trail the image of the star a little way up and down the slit.

Two key factors that characterize the performance of a spectrograph are dispersion and spectral resolution. Dispersion is a measure of the extent to which the constituent wavelengths are spread out, the linear dispersion being the change in wavelength over a given physical length of the spectrum. Dispersion is normally quoted in nanometers per millimeter (nm mm^{-1}), a dispersion of 1 nm mm^{-1} corresponding to a change in wavelength of 1 nanometer over 1 millimeter of the imaged spectrum's length. Spectral resolution (R) is a measure of the instrument's ability to reveal fine detail in the spectrum and is defined by $R = \lambda/\Delta\lambda$, where λ denotes the wavelength of interest and $\Delta\lambda$ the smallest wavelength interval that can be distinguished within the spectrum. For example, when used to study light with wavelengths of around 500 nm, a spectral resolution of 500 would correspond to being able to distinguish two idealized spectral lines that were separated in wavelength by 1 nm. The dispersions of astronomical spectrographs range from about 100 nm mm^{-1} to about 0.01 nm mm^{-1} and their resolutions from about 10 to about 100 000.

See also: diffraction grating, dispersion, focus, spectroheliograph, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectroscopic Binary Stars

Historically, spectroscopic binary stars were binary systems whose nature was discovered by the changing DOPPLER EFFECT or shift of the spectral lines of one or both of the component stars. The observed Doppler shift is a combination of that produced by the constant RADIAL VELOCITY (i.e. line-of-sight velocity) of the center of mass of the whole system, and the variable shift resulting from the orbital motion of the component stars about that center of mass. If the variable orbital component is large enough and the two stars are comparable in luminosity, then two spectra are periodically seen and can be measured, although the system appears as a single star in even the largest telescopes. (Sometimes only one star is bright enough for its spectrum to show in the combined light, but still the variable radial velocity of that one star can be measured. Extreme examples of such 'single-spectrum' systems are provided by the planetary systems discovered in recent years around several stars other than the Sun.) Spectroscopic binaries, therefore, are distinguished from visual binaries, the components of which can be resolved, often by quite small telescopes.

Astronomers were not able to detect spectroscopic binaries until astronomical spectroscopy had been developed sufficiently for reliable measurements of radial velocity to be possible. This, in turn, depended on the application of photography to such measurements, so that no spectroscopic binaries were known until late in the 19th century. In 1889, Pickering in Harvard and Vogel in Potsdam independently published radial-velocity studies of Mizar and ALGOL, respectively, which thus became the first two known spectroscopic binaries. Visual binaries, on the other hand, were known at least from the late 18th century, and their study became a significant part of astronomy after the publication in 1803 of an important paper by Sir William Herschel in which he announced that the two components of Castor which he recognized were in orbital motion about their center of mass (see VISUAL BINARY STARS). Herschel, in fact, coined the term 'binary star' to describe this kind of association of two stars. Interestingly, we now know that each of those visual components of Castor is itself a spectroscopic binary, and these two bright binaries are also accompanied by a more distant faint spectroscopic binary, YY Geminorum. The first 'binary' system to be discovered, therefore, is in fact a sextuple system.

A third kind of binary—ECLIPSING BINARY STARS—has also long been recognized. John Goodricke, again in the late 18th century, suggested that the light variations in Algol might arise from periodic eclipses of the primary star by 'a large body'. His conjecture was not confirmed until Vogel's work showed it to have been correct, so the serious study of eclipsing binaries began at about the same time as that of spectroscopic binaries. It should be immediately obvious that the distinction between these two classes is quite artificial. All eclipsing binaries are potentially spectroscopic binaries, although it is hard to observe spectroscopically, at adequate dispersion, the faintest ones known—and that was especially true when we were completely dependent on photographic emulsions for recording spectra. Similarly any spectroscopic binary would display eclipses, if its orbit were so oriented that each component star passed between us and the other one, once in each orbital period. The methods of observation and analysis are very different for spectroscopic and eclipsing binaries, however, and it is often convenient in text-books and encyclopedias to treat them as two distinct classes. Nevertheless, whenever possible, spectroscopic and photometric observations should be discussed together in the analysis of any given binary system.

It is less obvious whether visual and spectroscopic binaries are really distinct classes or, again, merely appear to be so because of the limitations of the methods of observation. Most visual binaries have orbital periods measured in years, decades or even centuries. The overwhelming majority of spectroscopic binaries have periods less than 100 days and several well-defined subclasses have periods of much less than a day. This large range of periods between visual and spectroscopic binaries corresponds to a similarly large range of separations and suggests that there might be a real physical distinction between close (i.e. spectroscopic and eclipsing) binaries on the one hand, and wide (i.e. visual) binaries on the other (see CLOSE BINARY STARS). The notion of such a distinction gained strength from the fact that, for a long time, few binaries were known with orbital periods in the range between those of the two apparent classes. In recent decades, however, improvements in both the speed and precision with which we can measure radial velocities have permitted the detection of the slow, small- amplitude variations to be expected from binaries in the 'missing' period range. At the same time, the development of, first, speckle interferometry (see SPECKLE IMAGING OF BINARY STARS) and, more recently, interferometry with optical arrays has permitted the resolution of pairs that were previously considered long-period spectroscopic binaries. The apparent minimum in the period distribution is by no means as clear as was once thought. Whether or not there is a real distinction between close and wide binaries is perhaps not finally decided, but evidence against one is mounting and it is this writer's opinion that such a distinction does not exist. Again, whenever possible, spectroscopic observations, that reveal the projection of the orbital motion on the line of sight, and visual or other observations, that reveal its projection on the plane of the sky, should be analyzed and discussed together. We may, perhaps, re-define spectroscopic binary systems as binary systems in which the radial velocity of at least one component can be measured, regardless of whether or not the components are resolved.

Binary systems offer many problems of their own for study, such as their origin, evolution and, in the close pairs, the mutual non-gravitational interactions of the components. Historically, however, the main motive for studying binaries of all kinds has been the opportunities

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

they provide for the determination of stellar properties, particularly stellar masses. Spectroscopic and eclipsing binaries are specially attractive in this respect, since their relatively short orbital periods enable data to be accumulated much more quickly for them than for visual binaries. For the best results, however, many precautions must be taken to ensure that the measured velocities are as precise as possible and do indeed correspond to the velocities of the centers of mass of the component stars. If observed values of stellar mass are to be of any use in the testing of theories of stellar structure and stellar evolution, the most precise measurements possible must be made of the radial velocities, and the systems chosen for measurement must be free of the complications found in systems in which the components are mutually interacting. For this reason, the best mass determinations come only from the so-called detached binary systems (see BINARY STARS: DETACHED).

Spectroscopic binary systems form a considerable fraction of the stellar population and, contrary to earlier beliefs, there is no particular evidence that their numbers are fewer among the late-type stars than among those of early type. If all known eclipsing binaries were included, the total number of known and suspected spectroscopic binaries would be very large. A compilation of spectroscopic binaries with known orbital elements (see below), published in 1988, contained information for over 1400 systems. Some of these may eventually be found not to be binaries at all, but other known systems were missed in the compilation. The observational effort put into the detection and measurement of new systems since then has greatly intensified and a rough estimate would be that the number of systems with known orbital elements is now double the 1988 number. Stellar multiplicity is not confined to binary systems: many binaries belong to more complex groupings—as already mentioned, Castor is a sextuple system. More common are triple systems, especially those in which a close pair is accompanied by a more distant companion. It is intuitively obvious that such a configuration is relatively stable since, at large distances, the close pair acts gravitationally almost as a mass point. If a visual orbit is known for the distant companion, the total mass of the triple system may be obtained and the individual masses of all three components, even if the close pair is not eclipsing. The existence of multiple systems is of significance for theories of binary formation (see CLOSE BINARY STARS: FORMATION AND EVOLUTION).

The orbital period

It should now be clear that the orbital period is one of the most important things to be determined for any kind of binary system. For spectroscopic binaries, the basic observational data consist of a series of measurements of the radial velocity (usually expressed in km s^{-1}) of one or both components at given times. Unless the orbital period is already known, the times are likely to be arbitrarily spaced. Often, the period is known, at least approximately, either from earlier spectroscopic observations or from photometric observations of an eclipsing binary or from visual observations. If the period is not known, the first step is to determine it. This used to be more of an art than a science, but modern computers have made possible extensive series of trial and error, and most observatories at which binaries are routinely studied have programs available that will search for all possible periodicities in a given set of data and give an estimate of the significance of those found. Sometimes the program will indicate two or more possible periods—in which case future observations can be planned to choose between them. Even with computer aid, however, the investigator must be on guard against possible erroneous periods. A multiple or submultiple of the true period may be found, for example. Confusion is particularly likely if the two components have very similar and almost equally strong spectra, since there may be doubt at any phase at which both spectra can be seen which one is the primary (i.e. more luminous) component. More subtle are the errors that creep in if the true period is close to a sidereal day. Since observations are often made at just such an interval, a spurious (or so-called *alias*) longer period may be found, imposed on the data by the pattern of observing. The same effect can work in reverse, a long-period system appearing to have a period close to a sidereal day.

Once the orbital period, P, has been found (usually it is expressed in days) the phase ϕ of an observation made at time t can be found from the fractional part of the expression

$$(t - t_0)/P$$

where t_0 is some convenient zero of time—it might be the time of the first observation or, for an eclipsing binary, the time of the middle of the deeper eclipse. Clearly ϕ always lies between 0 and 1. It is now possible to plot the observed velocities against the phase to give the velocity curve of the binary. This is a representation of the observed velocity variation as if all the observations had been made in one orbital cycle (similar to the light curve of an eclipsing binary) and is the first step to determining the other characteristics of the system (see figure 1).

It often happens, even for a newly observed binary, that a few older observations of the star exist. They are sometimes useful for refining the value of the orbital period since they give isolated velocities at some time well in the past and can be used to estimate the number of cycles between the two sets of observations and therefore to improve the preliminary value of the period. A word of warning is again necessary, however: orbital periods either can appear to vary or actually do vary. If the binary revolves around its center of mass with a third star (as we have seen, a not uncommon situation) the orbital period will appear to vary cyclically because of the changing light-travel time across the larger orbit. The orbital period of a binary will also appear to vary cyclically if the major axis of the binary orbit (the line of apsides) is rotating because of tidal interaction between the components. Finally, in systems in which the stars are interacting by mass transfer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

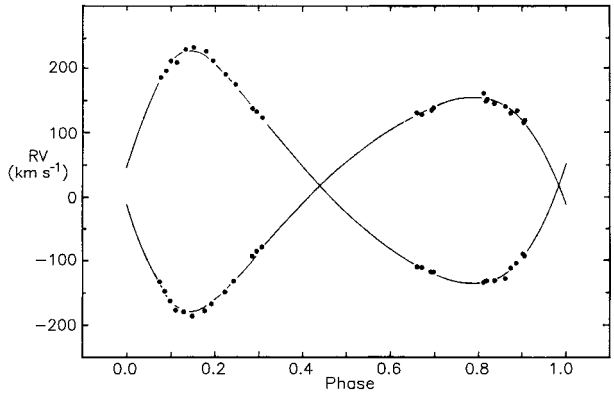


Figure 1. A modern velocity curve of the eclipsing and spectroscopic binary QX Carinae; reproduced with permission from Andersen J *et al* 1983 *Astron. Astrophys.* **121** 271–80. The orbital eccentricity is 0.28 and the longitude of periastron 124° for the primary star.

from one to the other, or in which mass is leaving the system, the orbital period really will change, usually in an irregular fashion. It used to be fashionable to quote orbital periods to very many decimals of a day, but such optimism is rarely justified.

The orbital elements

Once the orbital period is known and the velocity curve has been drawn, we can use the latter to determine the other characteristics of the orbit, which will in turn lead us to the goal of determining the masses of the component stars. As with planetary orbits in the solar system, the orbit is completely known if certain quantities called the ORBITAL ELEMENTS are known. Two angles are needed to define the orientation of the orbital plane in space, a third angle defines the orientation of the major axis of the elliptical orbit in that plane, a quantity e, lying between 0 and 1, defines the eccentricity of the orbit and the major semiaxis, a, gives the size of the orbit. It is also necessary to specify where a star is in the orbit and this can be done if the time of periastron passage is known. In planetary orbits, it is not necessary to specify a and P separately; they are related by KEPLER'S THIRD LAW

$$a^3/P^2 = \text{constant}.$$

The mass of the Sun is overwhelmingly greater than that of any of the planets, so the ratio just given is constant to a very high degree of precision. In binary orbits, however, with two comparable stellar masses moving under their mutual gravitation, we must use the generalized form of Kepler's law:

$$\frac{a^3}{P^2} = \frac{G}{4\pi^2}(m_1 + m_2) \tag{1}$$

where m_1 and m_2 are the masses of the two stars (usually m_2 is less than m_1) and G, of course, is Newton's constant of gravitation. Since m_1 and m_2 are not known in advance, a and P are usually regarded as separate elements. In summary, we seek to know for a binary orbit the following.

- ☊ the longitude of the ascending node (the point in the orbit at which the receding star crosses 'the plane of the sky').
- i the inclination of the orbital plane to the plane of the sky.
- ω commonly called the longitude of periastron, although by strict analogy with the practice in planetary astronomy that term should be reserved for ω+☊. The angle ω is the angle between the direction of the ascending node and the direction of periastron.
- e the orbital eccentricity.
- a the major semiaxis.
- P the orbital period.
- T the time of passage of the two stars through the periastron point.

In addition, for spectroscopic binary systems we usually consider as an element

- V_0 the radial velocity of the center of mass of the system. Note, however, that this quantity is variable in a triple system and that may cause problems if the interval covered by the observations is an appreciable fraction of the longer orbital period.

These quantities refer to the relative orbit of one star with respect to the other. If we are interested in the absolute orbits of either or both stars, relative to the center of mass of the system, we can distinguish the elements for each component by the suffixes 1 and 2. Clearly, P must be the same for each component and ☊, i, e, T and V_0 should be, while ω_1 should equal $\omega_2 + 180°$. In practice accidental and systematic errors in the observations will often lead to slightly different values of these elements for each component. We finally note that

$$a = a_1 + a_2$$

and, by simple moments,

$$\frac{a_1}{a_2} = \frac{m_2}{m_1}. \tag{2}$$

Determination of the orbital elements

The basic equation for the radial velocity, V, of a binary component in its orbit is

$$V = V_0 + Ke\cos\omega + K\cos(v + \omega) \tag{3}$$

where $2K$ is the total observed range of radial velocity and v is the true anomaly in the orbit, i.e. the angle between the radius vector at time t and the direction of periastron. It is simple to compute the mean anomaly, M (in radians), at any time t, since

$$M = \frac{2\pi}{P}(t - T)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and v and M are related by a well-known series expansion in e or, less directly, by Kepler's equation. It can also be shown that

$$K = \frac{2\pi a \sin i}{P(1-e^2)^{1/2}}. \tag{4}$$

We thus see from equation (3) that radial-velocity observations by themselves give us no information about the angle ☊. (They do, however, when combined with visual or interferometric observations, which define the line of the nodes, tell us which of the two nodes is the ascending one.) We also see that we cannot separate $a \sin i$ since, by eliminating P from equations (1) and (4), we find, for a given eccentricity and total mass,

$$K = \text{constant} \times a^{-1/2} \sin i$$

and we do not know whether a small K arises from a large a or a small $\sin i$. However, if i is known from photometric, visual or interferometric observations, then a can be determined and hence, by equation (1), the total mass of the system. Equation (2) then serves to give us the individual masses.

The earliest methods of determining orbital elements were all graphical, a method devised by Lehmann-Filhés in 1894 being the most popular. Even today, an experienced investigator may roughly estimate the orbital elements from the appearance of the velocity curve. For example, if the latter is sinusoidal, the orbit is circular (and V_0 is then simply the mean observed velocity), and the greater the departure from a sinusoidal curve, the greater is the orbital eccentricity. The condition for finding V_0 if the orbit is eccentric is that the areas between the V_0-axis and the velocity curve lying above and below the former should be equal. If the velocity curve is symmetrical about the maximum observed velocity, then $\omega = 0^\circ$ (by convention, ω is zero when the major axis of the orbit is perpendicular to the line of sight and the ascending node coincides with periastron). If the velocity curve is symmetrical about the V_0-axis, but the ascent from minimum to maximum observed velocity takes longer than the following descent, then ω is 90°. Again, the greater the disparity between the two times, the more eccentric the orbit (see figure 2).

Such visual estimates are, of course, only of use as a rough guide to the investigator, and the graphical methods of determining orbital elements have become of historical interest only. It is important to use a fully objective computational method that will also give realistic values for the uncertainties in the elements. Lehmann-Filhés showed how his method could be adapted to a least-squares solution, although he believed (correctly for his time) that such a solution would rarely be worth the trouble of making it.

Equation (3) cannot be directly adapted to least squares since it is not linear in the unknowns, which are often of very different magnitudes, and the unknowns

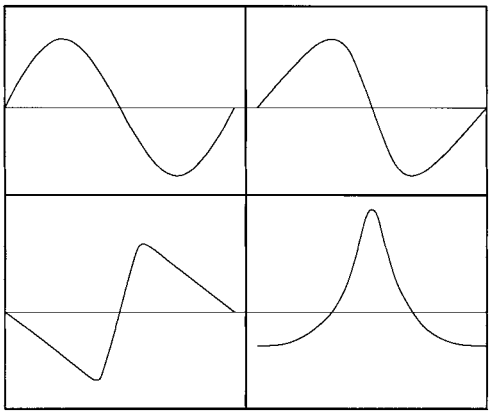


Figure 2. Schematic velocity curves for (upper left) a circular orbit, (upper right) an elliptical orbit with $e = 0.25$ and $\omega = 90^\circ$, (lower left) an elliptical orbit with $e = 0.5$ and $\omega = 270^\circ$ and (lower right) an elliptical orbit with $e = 0.5$ and $\omega = 0^\circ$. The scale of velocity is arbitrary. The horizontal line represents the V_0-axis, velocity of the center of mass of the system. In the first two elliptical orbits, periastron passage occurs when the steeper branch of the velocity curve crosses the V_0-axis, and in the last at the time of maximum velocity.

T and P appear in it only implicitly. Lehmann-Filhés differenced the equation and collected all the corresponding terms together to give

$$\begin{aligned}\Delta V = {} & \Delta V_0 + (\cos u + e\cos\omega)\Delta K - (\sin u + e\sin\omega)K\,\Delta\omega \\ & + \left(\cos\omega - \frac{\sin u \sin v}{1-e^2}(2 + e\cos v)\right) K\,\Delta e \\ & - \sin u(1 + e\cos v)^2 (t - T)\frac{K\,\Delta\mu}{(1-e^2)^{3/2}} \\ & + \sin u(1 + e\cos v)^2 \frac{K\mu\,\Delta T}{(1-e^2)^{3/2}}\end{aligned} \tag{5}$$

where, for brevity, $u = v + \omega$, $\mu = 2\pi/P$. In practice, scale factors are often introduced into the coefficients of Δe, $\Delta\omega$ and ΔT, so that all the unknowns will be similar in magnitude (the solution might give, for example, $100K\,\Delta\omega$). Although, in principle, a correction can be computed to the period (through $\Delta\mu$) that is rarely done (i.e. $\Delta\mu$ is assumed to be zero). Including a correction for the period reduces the weight of the whole solution and, if the period needs refinement, this can often be better achieved by the method already described. To make the solution, one must obviously begin with preliminary values of the elements, and the quantities ΔV are the residuals of the observed velocities from those computed with the preliminary elements, while the other differences are the corrections to be made to the preliminary values of the several elements. In the days of hand-operated, or even electrically driven, calculators, it was worthwhile to spend some time deriving as good a set of preliminary elements as possible by some graphical method, in order to avoid the need to iterate with a second solution to find the final elements. Nowadays, however, since electronic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

computers can perform several iterations very quickly, rough estimates of V_0, K, e, ω and T are often sufficient to start the solution.

Equation (5) works well if the orbital eccentricity, e, is appreciable. When e is small, however, the coefficients of $\Delta\omega$ and ΔT become more nearly proportional to each other, and the system of normal equations is indeterminate when $e = 0$. Since a very considerable portion of spectroscopic-binary orbits is circular or nearly circular, that is a serious defect of the Lehmann-Filhés method. Until the 1940s it was common practice to fix the value of T or ω arbitrarily and to solve for the other elements. This led to largely meaningless values of T and ω and, more seriously, to gross underestimates of the uncertainties of those two quantities. Solutions made in this way may still mislead the unwary investigator who happens to find one of the original papers and is unfamiliar with the early practice. Extensive modern recomputations have been made both by Luyten and by Lucy and Sweeney.

In 1941, Sterne published an alternative method of orbital solution, specifically designed to overcome the problem posed by small eccentricities. He gave up the attempt to determine T and ω directly, when the eccentricity was small. Instead he defined a quantity, L, called the mean longitude, where

$$L = \omega + M$$

and M is again the mean anomaly. Sterne also defined a time T_0 at which the mean longitude is zero, so

$$L = \mu(t - T_0).$$

When the orbit is exactly circular, then T_0 is the time of nodal passage, or of maximum radial velocity. For small eccentricities, the equations of condition then take the form

$$\Delta V = \Delta V_0 + \cos L\,\Delta K + \cos 2L\,Ke\cos\omega + \sin 2L\,Ke\sin\omega + \sin L\,\mu K\,\Delta T_0 - \sin L\,(t - T_0)K\,\Delta\mu. \quad (6)$$

Once again, in practice, only rarely is the term $\Delta\mu$ included, and scaling factors may be needed to bring all the unknowns to the same order of magnitude, but $e\cos\omega$, $e\sin\omega$ and T_0 may be determined without difficulty, even when e is small, and from them e, ω and T may be found. Equations (5) and (6) together should suffice for the determination of all orbital elements.

It may turn out, however, that even if an apparently significant value is found for e, the orbit should be treated as circular. The representation of a given set of observations is *always* improved by the use of an extra parameter; the question we have to ask is whether that improvement carries with it any physical significance for the orbit. This question was answered in 1971 by Lucy and Sweeney, who devised a test based on computing the amount by which the residuals are decreased, when an eccentricity is used to represent the observations, compared with those obtained from a circular orbit. They found, and subsequent experience has confirmed, that often a computed eccentricity should be ignored, even if it meets the usual criterion for statistical significance.

Determination of the masses

Our chief interest in obtaining the orbital elements is to determine the masses of the components of the binary system. We have already seen from equation (1) that the total mass is proportional to a^3, and from equations (2) and (4)

$$K_{1,2} \propto a_{1,2} (\propto m_{2,1}).$$

Thus we have two equations for the masses of the two stars in the system (provided that we know the orbital inclination i). Usually K is expressed in km s^{-1}, P in days, a in km and m in units of the solar mass (2×10^{30} kg). These units may seem a little strange to those brought up in a strict SI system, especially since two different units are used for time. They have been found convenient, however, and, in these units,

$$a_{1,2}\sin i = 13\,751(1 - e^2)^{1/2} K_{1,2} P$$
$$m_{1,2}\sin^3 i = 1.0385 \times 10^{-7}(1 - e^2)^{3/2}(K_1 + K_2)^2 K_{2,1} P.$$

Note that m is proportional to K^3. That is the principal reason why great care is needed to determine $K_1 + K_2$ as precisely as possible, since the percentage error in $m_1 + m_2$ is at least three times that in $K_1 + K_2$. To make useful comparisons with computed stellar models, a precision of about 1% in the individual masses is needed. Challenging though this is, it has now been achieved for an appreciable number of systems.

If the two stars are very different in luminosity, then the spectrum of only the brighter can be seen and measured. Obviously, the total mass of the system cannot then be determined, but the so-called mass function, $f(m)$, can be computed from

$$f(m) = \frac{m_2^3 \sin^3 i}{(m_1 + m_2)^2} = 1.0385 \times 10^{-7}(1 - e^2)^{3/2} {K_1}^3 P.$$

Even this permits some inference to be made about the masses, since we can write

$$f(m) = \frac{m_2 \sin^3 i}{(1 + m_1/m_2)^2}.$$

The vast majority of stars obeys the mass–luminosity relation, the more massive star being the brighter. Therefore m_1/m_2 is usually greater than unity and the denominator in equation (7) is thus greater than 4. In general, each star in a single-spectrum binary system must have a minimum mass of $4f(m)$. There are occasional exceptions. In β Lyrae, the spectrum of the more massive component is largely concealed by matter surrounding that star—for a long time, no certain detection of that spectrum was made. The system is massive, but some astronomers who studied the system were misled into supposing that it was several times more massive than it really is.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Modern problems

It has been tacitly assumed in most of the foregoing that the measured radial velocity of a binary component is at all times that of the center of mass of the star. When that is so, good spectroscopic observations combined with good photometric observations (if the binary is eclipsing) can lead to the determination of not only precise stellar masses but also precise stellar radii and luminosities. Such binaries are our principal source of fundamental stellar data and are of fundamental importance in providing observational tests of theories of stellar structure and evolution. In many of the most interesting binaries, however, the measured radial velocity may differ from that of the center of mass of the star. This can come about for any of three reasons, all of which may have effect simultaneously in some binary systems: (i) the surface of the binary component is irradiated by that of the other and the isophotes of the surface of the irradiated component are no longer concentric circles about the center of the stellar disk ('reflection effect'); (ii) one or both components are non- spherical in shape, because of tidal or rotational distortion and, since the luminous flux emerging from any point on the stellar surface is proportional to the local value of the acceleration due to gravity, the stellar isophotes again depart from circular symmetry ('ellipticity effect'); (iii) streams of matter between or around the components contribute to the line profiles in the observed spectrum—especially to those of the hydrogen lines—so that the measured velocity is affected by the presence of the stream and is not simply the orbital velocity of the star.

The reflection and ellipticity effects were well known early in the history of spectroscopic binaries, because the same two underlying causes also affect the light variations of eclipsing binaries. They affect the measured radial velocity because, in each case, they deflect the effective light center of the stellar disk from the geometrical center so that any rotational velocity the star may possess contributes to the observed velocity. Both effects are present in any binary system in which the separation of the two components is no more than a few times their own radii. They become very pronounced, however, in W Ursae Majoris systems. These are binaries with periods of less than a day whose components are so highly distorted as actually to be in contact over a small region of their surfaces. The velocity curves often show appreciable departures from the kind of variation expected from a star moving in a Keplerian ellipse and only recently has it become possible to use the observed velocities to obtain reliable orbital elements for these systems. This is done by computing the flux distribution over each stellar surface and then computing the expected departures of the measured velocities from simple Keplerian motion and comparing the computations with the actual observations (figure 3).

The distortion of velocity curves by the effects of circumstellar matter was first pointed out by Struve in the 1940s. It was especially noticeable in systems resembling Algol. The velocity curve might still look Keplerian in form but suggest an orbital eccentricity that could not be reconciled with the timing of eclipses. (In a circular orbit, the eclipse of one star necessarily occurs halfway between two successive eclipses of the other; in an elliptical orbit that is true only if ω is equal to 90° or 270°.) In fact, the first clue to this kind of distortion emerged in 1930 when E F Carpenter found a large orbital eccentricity for the velocity curve of U Cephei that completely contradicted the photometric evidence, but the interpretation of this result had to await the work of Struve some 10–15 yr later. Struve also predicted that novae and recurrent novae would be found to be members of binary systems—a prediction first confirmed by Kraft from ground-based observations, and then more spectacularly by the discovery of x-ray binaries from space. Such systems, now usually referred to as CATACLYSMIC BINARIES, contain compact components (WHITE DWARFS, NEUTRON STARS or possibly even BLACK HOLES). Streams of gas within these systems reach very high velocities and create 'hot spots' when they collide with the stellar surfaces or with each other. These hot spots can dominate the light of the system, at least at some phases, so both the observed light and velocity variations depart considerably from the expectations based on simple binary motion. Although Struve's hypothesis of gas streaming long remained controversial, modern observations in the ultraviolet, x-ray and radio wavelengths, as well as in the optical region of the spectrum, have left little room for doubt of its basic correctness. Indeed, modern techniques of tomographic analysis (a principle also used in medicine in which 'pictures' taken from several angles—as the binary revolves—are built up into a three-dimensional representation) have made it possible to begin to understand the characteristics of the gas streams themselves.

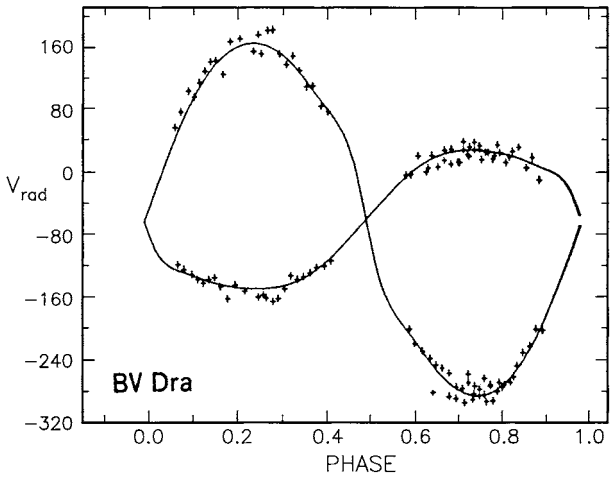


Figure 3. A recent velocity curve for the W Ursae Majoris system BV Draconis, showing the departures from Keplerian motion. The true orbit is circular: compare these curves with the upper left section of figure 2. Reproduced with permission from Rucinski S and Kałuzny J 1986 *Astron. J.* **92** 666–72.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The gas streams transfer mass between the components of a system, or remove it from the system altogether. These streams arise because evolving stars expand and the presence of a companion limits the amount by which a star can expand and remain stable. Even in systems in which the components are far enough from each other to permit considerable expansion without destroying each other's stability, tidal action between the enlarged components may modify their individual evolution. A possible example of this is provided by the well-known, bright binary CAPELLA, which can be observed both as a spectroscopic and interferometric binary.

In binaries with eccentric orbits, tidal action often betrays its presence by the rotation of the major axis of the orbit, or the line of apsides, which has already been mentioned. The presence of a third body, or general relativistic effects, can also give rise to apsidal rotation, but in most systems tidal distortion is the most important cause. It produces apsidal rotation because stars are not infinitely concentrated to their centers, and do not, therefore, act gravitationally exactly as mass points. This has been known for a long time, and the basic formulae were published by H N Russell in 1928. Because apsidal periods are long, however, and the rate of apsidal rotation is proportional to the fifth power of the stellar radius expressed in units of the separation of the two stars, only in recent decades has sufficient precision been reached for useful deductions from the observations to be possible. Apart from the observation of non-radial oscillations and solar neutrinos, which still give problematical results, measurements of apsidal motion provide the only direct means we have of probing the internal structure of a star. Results tend to confirm the predictions of stellar models.

In most binary systems displaying apsidal motion, the relativistic contribution is small compared with the Newtonian contribution arising from the facts that the stars are not mass points and are tidally distorted. In some close massive systems, however, the relativistic effect can be significant, or even dominant. In DI Herculis, for example, it is the Newtonian effect that is expected to be insignificant compared with the relativistic. Surprisingly, the observed rate of motion is much less than relativity predicts, which led some people to question that theory. On the other hand, binaries containing pulsars provide a very sensitive test of relativity, because the orbital elements can be determined with much greater precision from the variations in the arrival times of the pulses than from radial velocities, and apsidal motion is much more rapid in these systems (about 90° in 20 yr in PSR 1913+16) than in ordinary binaries. There is, however, an even more important test of relativity that can be made with this system—the change in orbital period caused by gravitational radiation from the binary. The measured rate agrees to within half a per cent with the predicted rate. Such close agreement leaves little room, if there had been any earlier, to doubt the correctness of general relativity. The result earned a Nobel prize for Taylor and Hulse who obtained it.

Bibliography

The best review of current state of our knowledge of stellar masses and dimensions and its implications for theories of stellar evolution is given by Andersen in

Andersen J 1991 Accurate masses and radii of normal stars *Astron. Astrophys. Rev.* **3** 91–126

A longer and more detailed account by the present writer of some of the modern problems of binary-star studies may be found in

Batten A H 1995 Stellar evolution in binary systems *Rep. Prog. Phys.* **58** 885–928

A discussion of how modern observations of apsidal rotation can help in the study of stellar structure will be found in

Claret A and Giménez A 1991 The effect of core overshooting and mass loss on the internal density concentration of main sequence stars *Astron. Astrophys.* **244** 319–26

Alan H Batten

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrum (pl: spectra)

In general terms, the distribution of intensity of electromagnetic radiation with wavelength. Thus when we examine the spectrum of star we are looking at a map of this brightness distribution. In the context of visible light, the visible spectrum is the band of colors produced when white light is passed through a glass prism, which has the effect of spreading out light according to wavelength. From long to short wavelength the colors so obtained are red, orange, yellow, green, blue, indigo, violet. These are the colors that make up the rainbow (although in practice not all of these colors may be discerned).

Such a spectrum is a continuous spectrum, or continuum (i.e. a continuous distribution of radiation over all wavelengths) and is emitted by a hot solid body or hot gas under high pressure. A continuum is also emitted by the free–free radiation and synchrotron radiation processes, but the form of the spectrum is different in each case.

A line spectrum is emitted by a gas under low pressure. If an electron in a high-energy orbit (i.e. of large radius around the atomic nucleus) makes a downward transition to an orbit of lower energy, then the difference in energy between the two orbits, ΔE, is released as a quantum (or 'packet') of radiation having a particular wavelength, λ, and frequency, f. The relationship between ΔE, λ, and f is as follows:

$$\Delta E = hf = hc/\lambda,$$

where c is the velocity of light and h is Planck's constant.

Each possible transition results in its own characteristic spectral line. In the hydrogen atom, the simplest atom, there are a number of series of spectral lines due to the possible transitions. For example, all the possible transitions down to the lowest energy level give rise to the Lyman series of emission lines at ultraviolet wavelengths, and all the possible transitions down to the second energy level (the first excited level gives rise to the Balmer series of lines in the visible spectrum). These lines are characteristic of hydrogen and hydrogen alone. Heavier elements have more complex spectra, and the spectra of molecules are complicated by vibrational and rotational energy states.

Dark absorption-line spectra are produced when atoms and molecules absorb radiation (from a background source) at the same wavelengths at which emission takes place by the mechanism described above. For example, a dark line of wavelength λ is produced when electrons of a given element absorb energy ΔE and make *upward* transitions from lower to higher energy orbits.

See also: electromagnetic spectrum, hydrogen spectrum, quantum mechanics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spectrum-X Gamma

Much delayed international x-ray mission. Originally scheduled for launch in 1993, unlikely to fly before 2000. Carries the Soviet–Danish Roentgen Telescope (SODART) for imaging, as well as spectroscopes for extreme ultraviolet and gamma ray studies. Also carries two US instruments and the Joint European Telescope for X-ray astronomy (JET-X) to study faint x-ray sources.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spencer Jones, Sir Harold (1890–1960)

Astronomer, born in London, England. He was chief assistant to Frank Dyson at Greenwich and worked at the Royal Observatory at the Cape of Good Hope. Following in the tracks of DAVID GILL, he repeated Gill's photographic survey of the southern sky (*Cape Photographic Catalogue*, CPC). He then organized the international project to determine Earth–Sun distance by repeated measurements of the asteroid Eros during its close approach to the Earth in 1930–1 (if the distance to any one object in the solar system can be determined, then the distances between the rest can be found by Kepler's laws). Under his direction, the Royal Observatory at the Cape photographed Eros more than 1200 times; he reduced this and the equal amount of data from other observatories—but it transpired that Gill was nearer the mark. Spencer Jones returned to Greenwich and became Astronomer Royal in 1933. Probably his most significant discovery was of the irregularities in the Earth's rotation, which he proved by showing that the planets all seemed to be together in pulling ahead and falling behind their calculated positions at given times as measured by the Earth's rotation. He replaced the concept of 'universal time,' based on Earth's rotation, by 'ephemeris time', a system of time measurement at first based on the motion of the planets in their orbits and then superseded by the frequency of oscillations in quartz and then cesium atoms. Having experienced the clarity of South African skies, he was depressed by observing conditions in London, and set in train the removal of the Royal Observatory to the Sussex countryside.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spherical Aberration

The inability of a spherical lens or mirror to bring all parallel rays of light to the same focus. When light reflects from a mirror, the surface of which has a shape that is part of a sphere, the position at which an incident ray is focused depends on its distance from the optical axis (the line that passes through the center of the mirror, perpendicular to its surface). Rays that reflect from points closer to the rim of the mirror are brought to a focus closer to the mirror's surface than rays that reflect from points closer to the optical axis. A similar effect applies to rays of light passing through a simple lens, the surfaces of which have shapes that are parts of spheres.

Spherical aberration causes the image of a point source of light (such as a star) to be spread out into a disk. Between the focal positions for the central and peripheral rays there is a plane at which the smallest image, known as the circle of least confusion, occurs. In the case of a mirror, spherical aberration can be eliminated for rays parallel to the optical axis by making its surface a paraboloid (the profile of a cross section of such a surface is a parabola). For a lens, the effects can be reduced by giving its two surfaces different radii of curvature (the radius of curvature is the radius of the circle that matches the curve of the surface).

As with chromatic aberration (a defect peculiar to lenses), the effects of spherical aberration can be reduced by using lenses with very high focal ratios. During the seventeenth century, this approach led to the construction of very long and unwieldy refractors.

See also: chromatic aberration, focal length, focal ratio, objective lens, reflector, refractor.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spica

The star α Virginis, a blue-white giant star, spectral type B1V, of apparent magnitude 0.98. At a distance of 260 light-years, parallax 0.012″, its absolute magnitude is −3.5. The name means 'the Spike' or 'Ear of Wheat', which the Virgin Mary is portrayed holding in the ancient constellation figures.

Spica is a rotating ellipsoidal binary, comprising two large, close components which are elongated towards each other by the strong gravitational attraction between them. Its magnitude varies very regularly between 0.95 and 1.05 over a period of 4.0146 days, as a result of the changing combined surface area of the component stars presented towards us. The primary component is also a regular pulsating variable with a period of 0.17 days.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spiral Arms

Lanes of gas, dust and hot young stars that spiral outwards from the central bulges of spiral and barred spiral galaxies. The most prominent features of spiral arms are clusters and associations of O and B type stars (highly luminous, high-temperature stars) and HII regions (luminous nebulae), both of which are indicative of recent and ongoing star formation. At radio wavelengths, spiral arms may be identified by measuring the distribution of neutral hydrogen clouds, which radiate at a wavelength of 21 cm, and molecular clouds. Infrared observations reveal clouds of warm dust heated by newly formed stars along the spiral arms.

In some galaxies ('grand design' spirals), the spiral arms are long, narrow and well defined, in others ('flocculent' spirals), they are more clumpy and fragmented, consisting of large numbers of shorter arm segments.

Within our own galaxy, the Milky Way Galaxy, lanes of dust absorb starlight and prevent optical astronomers from seeing stars in the plane of the galaxy at distances much beyond about 10 000 light-years. Nevertheless, enough O, B associations and HII regions can be seen to identify the spiral arms in the neighborhood of the Sun. Radio and millimeter-wave observations of emissions from hydrogen clouds and, more particularly, from molecular clouds, reveal the overall spiral structure. Our galaxy's spiral pattern consists of two or more major arms and a number of shorter segments, one of which—the Orion arm, or 'spur'—contains the Sun. The Sagittarius arm, which includes the stars and interstellar clouds that lie in the general direction of the constellation of Sagittarius, lies closer to the galactic center than the Orion arm and the Perseus arm lies further out.

The galaxy NGC 4622, imaged by the Hubble Space Telescope in 2002 appears to be rotating in the opposite direction to that which is expected. The clockwise rotation is puzzling because of the direction in which the outer spiral arms are pointing. Most spiral galaxies have arms of gas and stars that trail behind as they turn, but this galaxy has two leading outer arms that point toward the direction of the galaxy's clockwise rotation. To add to the conundrum, NGC 4622 also has a trailing inner arm that is wrapped around the galaxy in the opposite direction it is rotating. Astronomers suspect that NGC 4622 interacted with another galaxy as its two outer arms are lopsided, meaning that something disturbed it. NGC 4622 may have consumed a small companion galaxy. The galaxy's core provides additional evidence for a merger between NGC 4622 and a smaller galaxy. This information could be the key to understanding the unusual leading arms.

See also: BARRED SPIRAL GALAXY, MILKY WAY GALAXY, SPIRAL GALAXY

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Spiral Galaxy

A galaxy in which a central bulge of stars is surrounded by a flattened disk that contains a spiral pattern of hot young stars and luminous nebulae. Within the Hubble classification scheme, spiral galaxies, denoted by the letter 'S', are categorized according to the size of the central bulge, the tightness of the spiral pattern and the degree of patchiness in the spiral arms. Sa galaxies have relatively large central bulges, tightly wound and relatively smooth arms. Sb galaxies have somewhat smaller bulges and more open arms that often contain conspicuous HII regions (luminous nebulae) and clumps of hot blue stars; Sc galaxies have small central bulges and loosely wound, 'knotty' arms that typically contain numerous HII regions and clumps of stars. Galaxies with clumsy, chaotic and poorly defined arms are called 'flocculent' spirals, while those with thinner, longer, well-defined arms are termed 'grand design' spirals.

Spiral galaxies contain significant quantities of gas and dust, the relative proportions (by mass) of gas to stars ranging from as little as 2%, for some Sa galaxies to 10–15% in Sc galaxies. The masses of spiral galaxies (i.e. the combined mass of the stars, gas and dust that they contain) range from about one billion (10^9) solar masses to several hundred billion (~4–10^{11}) solar masses, and their diameters from about 15 000 to about 150 000 light-years. Analyses of the rotation curves of spiral galaxies (plots of the rotational velocities at different distances from the centers of galaxies) imply that spiral galaxies contain up to ten times as much dark matter as luminous matter and that their overall masses, therefore, are up to ten times greater than the figures quoted above.

There are two principal candidate mechanisms to account for the existence of spiral structure—self-propagating star formation, and density waves. The self-propagating star formation theory suggests that within major regions of star formation, high-mass stars evolve rapidly and explode as supernovae, the blast waves from these events compressing neighboring gas clouds and triggering further star formation in that part of the galaxy. Within an extensive star-forming region, bright stars and HII regions (luminous nebulae) closer to the galactic center will move ahead of those that are further out, so drawing out the star-forming region into a curving arc. In this way, random bouts of star formation produce a succession of arc segments that make up the clumpy arms of flocculent spirals. According to the density-wave theory, the spiral arms are a wavelike pattern that moves through a galaxy, rather like ripples on a pond, causing stars and gas clouds to bunch together temporarily to form the spiral arms. Individual stars and gas clouds revolve round the center of the galaxy in their own orbits, slowing down and bunching together temporarily each time they encounter a density wave, just as vehicles slow down and bunch together at roadworks or lane closures, travel slowly through the resulting traffic jam, then speed up and spread out at the other end. When a gas cloud enters a spiral arm, it is slowed down and compressed. This initiates a bout of star formation, so giving birth to the hot young stars and HII regions that delineate the arm. The density waves themselves are probably initiated by the asymmetric gravitational field of a central bar structure or by gravitational disturbances caused by neighboring galaxies, or by a combination of both.

Although both processes may have a role to play in any galaxy, self-sustaining star formation seems more appropriate to flocculent spirals and density waves to grand design spirals.

The galaxy NGC 4622, imaged by the Hubble Space Telescope in 2002 appears to be rotating in the opposite direction to that which is expected. The clockwise rotation is puzzling because of the direction in which the outer spiral arms are pointing. Most spiral galaxies have arms of gas and stars that trail behind as they turn, but this galaxy has two leading outer arms that point toward the direction of the galaxy's clockwise rotation. To add to the conundrum, NGC 4622 also has a trailing inner arm that is wrapped around the galaxy in the opposite direction it is rotating. Astronomers suspect that NGC 4622 interacted with another galaxy as its two outer arms are lopsided, meaning that something disturbed it. NGC 4622 may have consumed a small companion galaxy. The galaxy's core provides additional evidence for a merger between NGC 4622 and a smaller galaxy. This information could be the key to understanding the unusual leading arms.

See also: GALAXY, HUBBLE CLASSIFICATION, MILKY WAY GALAXY, SPIRAL ARMS.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Spitzer, Lyman, Jr (1914–97)

US astrophysicist, became professor at the Princeton faculty where he founded the Princeton Plasma Physics Laboratory (PPPL). He researched spectral line formation, stellar motions, star formation and the physics of the interstellar medium. He showed that there must be at least two phases of material in the interstellar medium—high-temperature clouds around hot stars and cooler inter-cloud regions. He led the project to make and operate the ultraviolet astronomy satellite, Copernicus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Spole, Anders (1630–99)

Swedish astronomer, studied in Germany, became professor of astronomy on the foundation of Lund University, and professor at Uppsala University in 1679. His observatory on the top of his house was completely destroyed, together with all his work, in Uppsala's great fire of 1702. He observed dust lanes in the tail of the great comet of 1680, and sunspots, which he believed caused comets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sporadic Meteor

A meteor that does not belong to an identifiable meteor shower. The meteoroids that produce sporadic meteors are part of the zodiacal dust cloud, which pervades the inner solar system, and are swept up by the Earth as it proceeds in its orbit. On average about five sporadic meteors are visible from any given location on any clear night. The actual number varies with time of night and time of year. Sporadic meteors seen by an observer in the evening can be produced only by incoming meteoroids traveling fast enough to 'catch up' with the Earth, whose rotation is carrying the observer away from the direction from which the meteoroids are coming. But after midnight, when the Earth's rotation carries the observer into the dust cloud, the sporadic rate increases. The seasonal variation has to do with the altitude of the Earth's apex—the point on the celestial sphere toward which its motion at a given time appears to be taking it. The higher the apex is in the sky, the greater is the sporadic hourly rate. For northern-hemisphere observers the apex reaches its greatest altitude at the autumnal equinox and the *contemporaneous sporadic background* (the sporadic hourly rate at a specified time of year) is then 8–10 per hour. The apex is nearest the horizon at the spring equinox, when the rate is just 2–3 per hour.

See also: interplanetary dust

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Square Of Pegasus

A conspicuous asterism of four bright stars forming a square of approximately 15° a side, notable for the absence of any but very faint stars within it. It is formed by the stars β, α and γ Pegasi (apparent magnitudes 2.44, 2.49 and 2.83 respectively) and α Andromedae (magnitude 2.07), and is prominent in the evening sky in autumn.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sripati (1019–66)

Born in (probably) Rohinikhanda, Maharashtra, India, wrote on astronomy and the mathematics associated with astronomy. He wrote *Dhikotidakarana*, about eclipses, *Dhruvamanasa*, about calculating the motion of the planets, and *Siddhantasekhara* about mathematics and astronomy in general.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SS Cygni

SS Cygni (SS Cyg) is a long orbital period ($P_{\rm orb} = 6.6$ h) cataclysmic variable of the dwarf nova (DN) subclass, characterized by semiperiodic brightness changes due to outbursts. This CLOSE BINARY STAR consists of a WHITE DWARF primary which accretes matter from a low-mass, ROCHE-LOBE-filling secondary star.

SS Cyg was discovered in 1896 by Miss Louisa Wells while examining Harvard College Observatory photographic plates. Wells noted that the star varied in photographic magnitude from 7.7 to 12.4. During the 1897 objective prism survey for the Henry Draper catalog, a spectrum of SS Cyg was obtained. However, it was not until detailed observational work in 1943 revealed radial velocity motions that SS Cyg was shown to be a spectroscopic binary, i.e. a close double star. SS Cyg was the second DN discovered; U Geminorum being the first, found by J R Hinds in 1855.

The DN subclass of cataclysmic variables consists of close binary stars in which the mass accreted from the secondary star forms a disk around the primary star. Semiperiodic outbursts, during which time the emitted light from the DN increases greatly, are the signature of a DN type of cataclysmic variable. The DN SS Cyg has outbursts about every 50 days which brighten the system from its normal quiescent level at visual (V) magnitude 11.7 to a maximum level near $V = 8.2$, a factor of 25 in brightness. From the Earth, we view SS Cyg at an inclination angle on the sky of 30°–60° and it is at an estimated distance of 150–180 pc.

At visual wavelengths, the emitted flux from SS Cyg is dominated by light from the ACCRETION DISK but contains a small contribution from the secondary star. The hot inner disk and the white dwarf dominate at ultraviolet wavelengths with the secondary star and the outer disk being the major contributors in the infrared. X-ray observations of SS Cyg imply that, during quiescence, the hot inner disk regions are likely to be a single-temperature plasma. During outburst, a spherical x-ray-emitting corona forms around the white dwarf and mass is lost from the system in the form of a wind. Extreme ultraviolet observations (70–130 Å) made at outburst show strong emission due to highly ionized Ne, Mg and Si produced within a region of temperature equal to 290 000 K.

The accretion stream, disk and hot spot

Figure 1 shows an optical spectrum of SS Cyg at quiescence. In the figure, it can be seen that the dominant spectral features are emission lines, due to hydrogen and helium, which are produced in the accretion disk and hot spot. These strong, broad emission lines allow determinations of many physical parameters within the binary, foremost of which are estimates of the stellar masses. As the binary rotates, the accretion disk follows the orbital motion of the white dwarf. Thus, measurements of the radial velocity changes of the disk emission lines provide an approximate tracer of the true

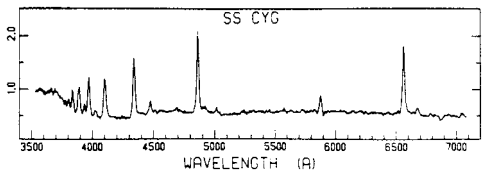


Figure 1. The optical spectrum (3500–7000 Å) of SS Cyg, shown during quiescence, is dominated by emission from the accretion disk. Prominent emission lines due to hydrogen (particularly Hα (6563 Å) and Hβ (4861 Å)) and weaker helium (He I (4471, 5876 Å)) are easily seen. The underlying continuum begins to rise at the blue end of the spectrum where the Balmer jump is seen in emission.

orbital motion of the primary star. Using Kepler's law, slightly reconfigured for our purposes, we have

$$\sin^3 i = \frac{K_1^3 P}{2\pi G M_2}\left(\frac{q+1}{q}\right)^2$$

where i is the binary inclination, M_2 is the mass of the secondary star, P is the orbital period, G is Newton's gravitational constant, K_1 is the observed maximum shift of the radial velocity curve, and q is the mass ratio of the binary defined as M_2/M_1. Typically, the orbital period and the K_1 amplitude are measured from observation, while M_2 is assumed[1] and i is approximated in various ways. The result is that q can be determined, and from it, the mass of the white dwarf ($M_{\rm WD}$).

The gas stream leaving the secondary star falls towards the primary but is curved in the direction of orbital motion owing to its stored angular momentum. This stream strikes the edge of the accretion disk and forms what is termed the hot spot. At the hot spot, the gas stream and the Keplerian orbiting accretion disk material mix, producing a region of enhanced density, temperature and light emission.

The material orbiting within the accretion disk slowly loses angular momentum through viscous heating and spirals inward, eventually accreting onto the white dwarf equatorial regions. The rate of quiescent mass transfer ($\dot{M}$) from the secondary star in SS Cyg is $8 \times 10^{-11} M_\odot$ yr^{-1} (5×10^{15} g s^{-1}), typical of DN with similar orbital periods. This value is not necessarily the rate of mass accretion onto the white dwarf surface, as the accretion disk provides a buffer and storehouse for the material. The accretion disk mass, $\sim 0.1 M_\odot$, remains nearly constant during interoutburst periods but some small fraction of it (~10%) is accreted onto the white dwarf during an outburst. Gravitational energy release by the accreted material is the cause of the disk heating and brightening observed during outburst. Accretion disks in DN have radii approximately given by $R_{\rm d}/a \sim 0.3$, where a is the distance between the two stars.

[1] A further relation can be derived which yields the radius of the Roche lobe of a secondary star given the binary orbital period. One then assumes the secondary star to be a main sequence star of this radius, hence a mass estimate can be made.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

For SS Cyg, $R_d \sim 4.5 \times 10^8$ m, in good agreement with observation.

Models of accretion disks in long-period DN, such as SS Cyg, are based on the belief that the material within the disk is optically thick and mass and energy, within the disk, are conserved. This model, usually referred to as a α-disk model, predicts a temperature structure which varies with radius as

$$T_{\rm eff}(r) \propto (M_{\rm WD}\dot{M})^{1/4} r^{-3/4}.$$

In SS Cyg, we find that its accretion disk has a high temperature ($T \sim 20\,000$ K) inner (optically thin) edge, cooling to ~4000 K or less at the outer edge. Since optically thick material radiates as a blackbody, the output spectrum predicted from an accretion disk is simply a summation of various annuli, each of some characteristic emission temperature (see BLACK-BODY RADIATION). Thus, the output continuum flux distribution will be of the form

$$F_\lambda \propto \lambda^{-2.3}$$

in the optical and IR bandpasses.

As the material in the accretion disk and hot spot region interacts with itself, small brightness changes occur of order a few tenths of a magnitude. These changes occur randomly or, at times, with quasi-periodic time-scales, and are termed flickering. Flickering is a tell-tale sign of a stellar system in which accretion is occurring. Power spectral analysis of the flickering in SS Cyg has shown that the amplitude increases towards the blue, indicative of a higher level of activity occurring in the hotter, inner disk regions (see also CORRELATION FUNCTION AND POWER SPECTRA IN COSMOLOGY).

During minimum light, photometric observations of DN can be used to estimate the inclination of the binary system. This type of observation assumes that the photometric signature from the disk and hot spot can, in principle, be interpreted to indicate whether the accretion disk is seen edge-on ($i = 90^\circ$) or tilted at some angle. In SS Cyg, minimum light optical photometry generally shows no modulations, indicating a low inclination ($i \sim 38^\circ$); however, occasional low-amplitude humps appear in both the optical and the UV, implying a higher binary inclination, $i \sim 50^\circ$. This uncertainty in the inclination at which we view SS Cyg leads directly to an uncertainty in the masses derived for its stellar components.

Outburst behavior

SS Cyg is probably the best-studied DN, having photometric information available from before 1900 up to the present. Figure 2 presents a small segment of photometric data for SS Cyg showing a number of outbursts. Outbursts occur every 49.5 ± 15.4 days in SS Cyg and have durations which are near either 7–8 days or 15 days. Other DN, similar to SS Cyg, show this bimodal structure as well. Sometimes the outbursts in SS Cyg tend to occur in a 'short–long–short–long' sequence with the outburst amplitude being correlated with the time since the last outburst and the time until the next

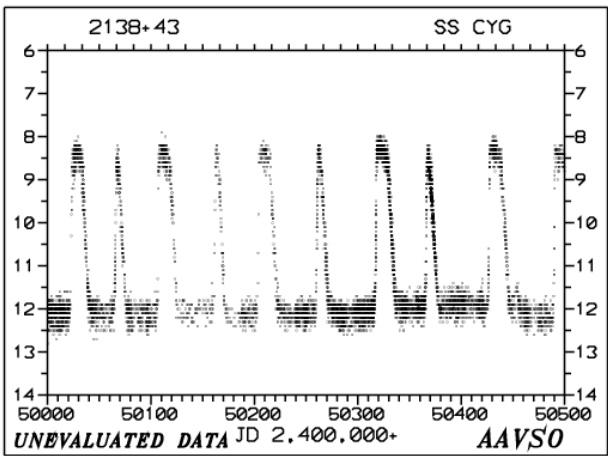


Figure 2. A typical 500 day light curve segment for SS Cyg. Notice that the outbursts are only semi-periodic and do not exactly alternate between long and short duration. The quiescent magnitude of SS Cyg is near 12 and the outbursts rise to near 8.5, a factor of about 25 in brightness. The x-axis is labeled in days starting with Julian day 2 450 000.

one. The brightest outbursts have an optical luminosity of $L_{\rm opt} \sim 5 \times 10^{34}$ erg s^{-1} or an integrated luminosity (over the entire outburst) of 3×10^{40} erg s^{-1}.

The cause of DN outbursts is thought to be a disk instability, not an increase in mass transfer from the secondary star. This disk instability, known as the thermal limit cycle, occurs when the material in the accretion disk has its local temperature and density values increased beyond some critical level (typically when $T > 8000$ K), causing a rapid ionization of the (mostly) hydrogen gas. This leads to a heating front which propagates throughout most of the disk, ionizing some fraction of it. The alternating 'long–short' outburst cycle is thought to be caused by this heating front moving throughout the entire disk at times and not making it to the outer disk edge at other times.

During outburst, optical oscillations have been seen in SS Cyg with periods of 7.3–10.9 s. These short-period modulations have also been seen at x-ray and EUV wavelengths. Because of their high frequencies, these brightness variations are thought to be associated with the very inner regions of the accretion disk close to or at the surface of the primary star.

The stellar components

SS Cyg contains a $(1.1 \pm 0.30) M_\odot$ white dwarf with a radius about equal to that of the Earth and with an effective temperature of 34 000–40 000 K. The mass-losing secondary star is a main sequence star of approximate spectral type K5 V with a mass of $(0.70 \pm 0.20) M_\odot$. The two stars orbit their common center of mass with a period of 6.6031 h, having a separation of only about the diameter of the Sun. Because of their close proximity and the gravitational influence of the primary, the secondary

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

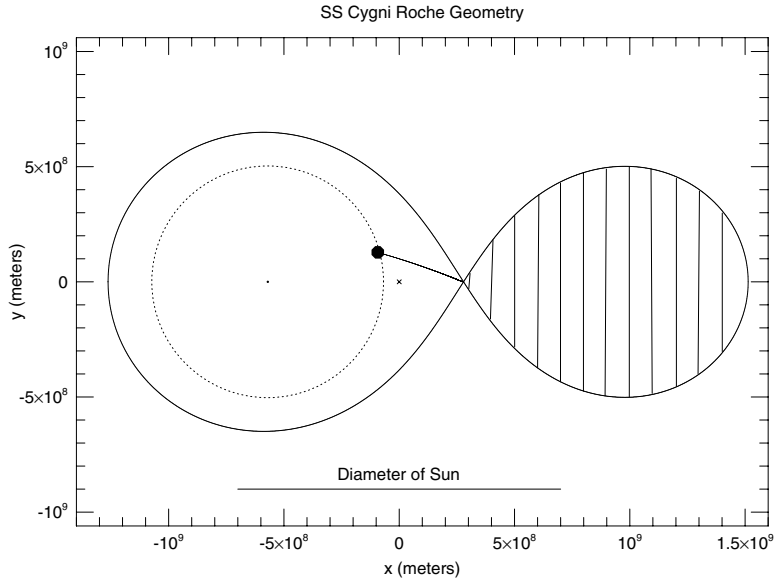


Figure 3. The Roche geometry for the DN SS Cyg. The ∞ shape is the critical Roche surface of the binary. The secondary star (right) fills its Roche lobe, transferring mass through the L1 point via a ballistic gas stream to form an accretion disk (dotted circle) surrounding the primary (left). The gas stream impacts the accretion disk at the hot spot (large filled dot). The small 'x' is the center of mass of the binary and the white dwarf (small dot) is shown to scale. The binary orbital motion is counterclockwise in this cartoon and the diameter of the Sun has been added for a size comparison.

star completely fills its Roche lobe, causing matter to be transferred through the L1 point towards the primary star. This material forms an accretion disk surrounding the primary star. Figure 3 illustrates the Roche geometry and other aspects of SS Cyg.

The original binary components of SS Cyg were a massive main sequence star ($>6.5M_{\odot}$) in a wide orbit with a low-mass main sequence star, fairly similar to the present-day secondary. The massive primary evolved off the main sequence and expanded greatly, engulfing the orbiting secondary in a common envelope, while its core became essentially the present-day white dwarf. Orbital angular momentum lost by the secondary caused expansion and loss of the common envelope, while greatly reducing the orbital separation. When the binary emerged from this so-called common envelope stage, it probably had an orbital period of a few days or less and was detached, the secondary star not filling its Roche lobe. GRAVITATIONAL RADIATION losses eventually shrunk the orbital separation of the two stars enough to cause the binary to become semidetached, synchronizing the rotation of the secondary. At this point, mass transfer commenced and the binary became a cataclysmic variable. SS Cyg has been a mass-transferring binary for only a short time, about 0.2–0.3 Gyr.

The secondary star becomes distorted in shape once it is confined by the Roche equipotential surface. This causes its properties to deviate from those of a single main sequence star. Continued mass exchange and magnetic braking will rapidly decrease the orbital separation of SS Cyg ($P_{\rm orb}$ will be $\sim$3 h in another 0.1–0.2 Gyr), leading to a non-equilibrium configuration of the secondary star. The secondary will expand, in response to continued mass loss, becoming up to 10–20% larger than a main sequence star of the same mass.

Single main sequence K5 stars are known to show stellar activity, such as starspots, and to show activity cycles similar to those which are seen in our own Sun. Thus, it is not surprising to think that such cycles might also exist on the secondary star in SS Cyg. SS Cyg has been observed to show a cyclic variation in its outburst interval and a 20% change in its quiescent brightness. Both of these variations occur within a period of about 7.3 yr and they have been attributed to solar-type starspot cycles occurring on the secondary.

Because of the closeness of the stellar components and the high temperature of the white dwarf, irradiation of the secondary star occurs. Emergent flux from the white dwarf illuminates the face of the secondary, near L1, with $\sim\frac{1}{10}$ of its total flux. (The accretion disk provides some shielding of the secondary star from direct illumination by the white dwarf.) Radiation from the accretion disk and hot spot region is also present, but much weaker owing to their lower temperatures. The effects of irradiation in SS Cyg cause very weak secondary star Balmer line emission (not visible in figure 1) and slight expansion of the secondary atmosphere near L1. During outburst, the inner disk becomes much hotter, emitting larger amounts of high-energy radiation. This radiation causes increased secondary star heating leading to phase-dependent secondary star spectral properties.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Work describing the long-term light curve of SS Cyg can be found in

Cannizzo J and Mattei J 1992 *Astrophys. J.* **401** 642

while detailed models of the accretion disk and outbursts of SS Cyg can be found in

Cannizzo J 1993 *Astrophys. J.* **419** 318

The evolution of all types of cataclysmic variable is discussed by

Howell S, Rappaport S and Politano M 1997 *Mon. Not. R. Astron. Soc.* **287** 929

A complete summary of this field of research is presented by

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

Steve B Howell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

SS433

A complex binary system within a SUPERNOVA REMNANT ('W50') in the constellation Aquila, which has proved to be of great astrophysical significance since it was discovered by the x-ray satellite ARIEL 5 in 1976. The unusual name derives from its inclusion in a catalogue of emission-line stars published by B Stephenson and N Sanduleak in 1977. As it is a variable star system it is also designated 'V1343 Aquilae'.

In addition to the x-ray observations, SS433 has also been studied at optical, radio and gamma-ray wavelengths. The combined results indicate that it is a system of great complexity and have stimulated considerable interest among both observational and theoretical astrophysicists over many years.

Particular attention was first drawn to the object by observed peculiarities in its optical spectrum. In addition to a background continuum, there are also strong emission lines due to hydrogen and helium—not in itself unusual. What is much more puzzling is that these emission lines seem to have a dual nature: normal, broad, 'stationary' lines are present, but so too are narrower lines which vary considerably in wavelength over a period of 164 days. These variations in wavelength indicate Doppler red- and blue-shifts alternately. The stationary emission lines also exhibit a more modest variation, with a period of 13.1 days.

Several conflicting interpretations have been proposed, which for many years could not be reconciled. A measure of agreement has been reached, however, and it is now generally accepted that the most likely structure for this remarkable object is substantially as follows.

A binary system, of orbital period 13.1 days, comprises a hot, massive star (estimated at between 10 and 20 solar masses) and a NEUTRON STAR (presumably the W50 supernova remnant) of perhaps 1.5 to 3 solar masses. A stellar wind from the massive star transfers matter through the LAGRANGIAN POINT, feeding an ACCRETION DISK surrounding the neutron star. Not all of this matter reaches the surface of the neutron star, however, part being ejected at enormously high velocity (of the order of 80 000 kilometres per second) via two opposed, narrow jets which precess in a period of 164 days, sweeping out two conical volumes of space of about 40° angle.

The massive star, and possibly the accretion disk, contribute the background continuum. The stellar wind produces the broad component of the stationary emission lines, and the mass-transfer from the massive star to the accretion disk produces the narrow emission lines. The changing aspect of the accretion disk through the 164-day period explains the variation in total light from the system, and its periodic partial eclipses account for the lesser 13.1-day variation. The precessional motion of the jets explains the moving emission lines with their observed Doppler effects. Some minor variations in the emission spectrum have also given rise to the suggestion that there may be a 'hot spot' in the accretion disk.

No doubt observation and discussion of SS433 will long continue, with a view to gaining further insight into the structure of this enigmatic object, which has rightly been described as 'unique in astronomy'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

St Andrews University Observatory

The University Observatory, St Andrews, has the largest optical telescope ever constructed in Scotland, the 0.94 m James Gregory Telescope (JGT), plus smaller teaching telescopes. The JGT is used for CCD photometry of short-period binary stars, and stellar clusters. The Astrophysics group in the School of Physics and Astronomy has six permanent staff plus postdocs and research students.

For further information see
http://www-star.st-andrews.ac.uk/astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

St Petersburg State University Astrophysical Observatory

The Astronomical Institute of St Petersburg University is located in Stary Peterhof, some 40 km from St Petersburg, Russia. The staff members also take an active part in teaching astronomy students of the university's mathematics and mechanics faculty.

Areas of research cover a variety of fields of astrophysics, stellar dynamics, celestial mechanics and astrometry. They include studies of stellar atmospheres, planetary nebulae, intergalactic clouds, interacting galaxies and blazars.

Observations are made with the 23 cm telescope in St Petersburg, the 0.7 m telescope of the Crimean Observatory, the 6 m and 1 m telescopes of the Special Astronomical Observatory, and the RT-7.5 radio telescope of Moscow Technical University.

For further information see http://www.astro.spbu.ru/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Standard Model of Particle Physics

The Standard Model describes all the confirmed data obtained using particle accelerators and has enabled many successful theoretical predictions to be made. It has been tested with per mille accuracy, at which level its calculable quantum corrections play an essential role. Its only missing feature is a particle, called the Higgs boson, whose coupling to the other particles is believed to generate their masses. However, the Standard Model is theoretically unsatisfactory despite its many successes, and possible extensions to provide a more unified picture of the different particle interactions, the apparent proliferation of different particle species and the origin of their masses are being proposed. Evidence has recently been presented that NEUTRINOS may alter their nature when they propagate over long distances. This is expected if neutrinos have masses, which are not predicted by the Standard Model. If confirmed, this may be the first evidence for particle physics beyond the Standard Model.

The FUNDAMENTAL PARTICLE interactions described by the Standard Model are the electromagnetic, weak and strong nuclear forces. Electromagnetic forces have long been known, since the discovery of electromagnetic wave radiation by Hertz in 1885 and the early days of quantum physics, to be mediated by the exchange of the photon (see figure 1), a massless boson of unit angular momentum (spin 1) with two polarization states. The long-established quantum theory of electrodynamics is called QED. Yukawa conjectured that the weak interactions such as β decay were mediated analogously by the exchange of massive intermediate bosons, as also shown in figure 1. This hypothesis was put on a firm theoretical basis by Glashow, Weinberg and Salam in their unified theory of the weak and electromagnetic interactions, and the intermediate bosons (weighing about 80 and 91 GeV) were discovered at CERN in 1983. The strong nuclear interactions are known also to be mediated by massless bosons called gluons, see also figure 1, which were discovered at DESY in 1979. Thus all the fundamental interactions have very similar structures, but why only the weak bosons are massive is a puzzle to which we return later.

The first elementary matter particle to be identified was the spin-1/2 electron (weighing about 1/2 MeV), followed by the unstable muon (weighing about 100 MeV) that was first detected in COSMIC RAYS. These were shown each to have an associated neutrino, the electron neutrino being produced in weak β decay and the muon neutrino being produced either in association with a muon or in muon decays. These particles do not have strong nuclear interactions, only weak and (in the case of the electron and muon, which are charged particles) electromagnetism, and are called LEPTONS. A third type of charged lepton (the τ weighing about 1.78 GeV) was discovered in accelerator experiments at SLAC in 1975 and is also believed to have its own associated neutrino. Accelerator data have established upper limits on the possible masses of the neutrinos, which are much less than those of the corresponding charged leptons.

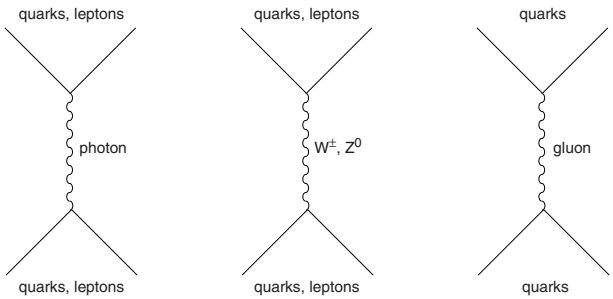


Figure 1. The fundamental forces between elementary particles are mediated by intermediate bosons: the photon of electromagnetism, the massive $W^{\pm}$ and Z^0 bosons of the weak interactions and the gluons of the strong interactions.

From the 1940s onwards, many strongly interacting particles (HADRONS) were discovered. It was proposed in 1964 that these could all be understood as composite bound states of more elementary entities called QUARKS. At that time, all the known hadrons could be constructed out of just three quark species of spin 1/2 and fractional charge Q: up ($Q = 2/3$), down and strange ($Q = -1/3$). Such pointlike constituents inside protons were discovered by experiments at SLAC in 1968. Initially it was a puzzle how quarks could appear almost free at high energies, corresponding to short distances, but were never seen singly, only confined inside hadrons. This puzzle was resolved when it was demonstrated that the couplings of gluons became weaker at high energies (short distances), a property known as asymptotic freedom.

Shortly afterwards, a key step in the establishment of the Standard Model was the discovery in a neutrino experiment at CERN in 1973 of a novel type of weak interaction, called neutral currents, in which the charges of the participating particles do not change, in contrast to the familiar weak interactions, called charged currents, in which the charges of participating quarks and leptons do change. Such neutral-current interactions were predicted by many unified theories of the weak and electromagnetic interactions, and their discovery provided the first circumstantial evidence that some such unified theory might be correct. The second piece of evidence for such a theory was provided by the discovery in accelerator experiments at BNL and SLAC in 1974 of hadrons containing a fourth quark constituent weighing about $1\frac{1}{2}$ GeV. This was baptized charm and had been predicted by unified electroweak theories, in order to explain the fact that neutral-current interactions that would change quark type, e.g. strange $\rightarrow$ down, are very suppressed, whereas there do exist charged-current interactions: strange $\rightarrow$ up. The charm discovery briefly pre-dated the τ lepton discovery mentioned earlier. Consistency of the simplest unified electroweak theory then required the existence of two more quark types, and the first of these (bottom), weighing about 5 GeV, was discovered in an accelerator

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

experiment at FNAL in 1977. The spectrum of quarks was apparently completed by the discovery at FNAL in 1995 of the top quark, which weighs about 175 GeV.

The mathematical framework for theories of particle interactions is provided by gauge theories, which offer the only consistent description of the interactions of spin-1 particles. A gauge theory is based on a continuous internal symmetry, e.g. changing the phase of the quantum-mechanical wave function in the case of the prototype gauge theory, QED. When one makes this symmetry local, i.e. allows space-dependent phase transformations, one must introduce a spin-1 gauge field. This is analogous to the way Einstein realized general coordinate invariance with a spin-2 metric field. Gauge theories differ in their choice of gauge group and in the group transformation properties of the matter particles. In the case of the Standard Model, the strong interactions are described by a theory transforming the wave functions of quarks, called QCD. The unified electroweak interactions are described by a combination of transformations acting on both quarks and leptons, first written down by Glashow in 1961, Weinberg in 1967 and Salam in 1968.

The establishment of the Glashow–Weinberg–Salam model as the correct unified electroweak theory proceeded in several steps over the following decade. A series of detailed experiments refined measurements of the neutral currents, verifying that their properties were consistent with the Glashow–Weinberg–Salam model. A crucial step was the observation at SLAC in 1978 of parity violation in deep-inelastic electron–proton scattering, which was followed by similar observations in atomic physics. The culminating step was the discovery of the weak intermediate bosons, $W^{\pm}$ for the charged-current weak interactions and Z^0 for the neutral currents, in proton–antiproton collisions at CERN in 1983.

Subsequent developments have been dominated by a series of precision tests of the Standard Model, including both its strong and its electroweak sectors. In a number of experiments at energy scales varying from about 1 GeV to about 200 GeV, it has been shown that the strong nuclear interactions indeed weaken at higher energies, as seen in figure 2. The 'strong' interactions are truly strong only at energy scales $\lesssim$1 GeV, corresponding to distances $\gtrsim$0.2 fm, where quarks are confined within hadrons. Detailed calculations of hadron spectroscopy require extensive numerical simulations of discretized (lattice) versions of QCD on powerful computers. These are currently able to reproduce experimental measurements to within a few per cent. Calculations at higher energies (shorter distances) may be made using techniques from quantum field theory. They predict an energy dependence of the strong coupling that is in very good agreement with the experimental measurements shown in figure 2. This concordance confirms that there are no other light strongly interacting elementary particles beyond the gluons and quarks introduced earlier. The present data shown in figure 2 indicate that the strong coupling strength, analogous to the fine-structure constant of QED, has

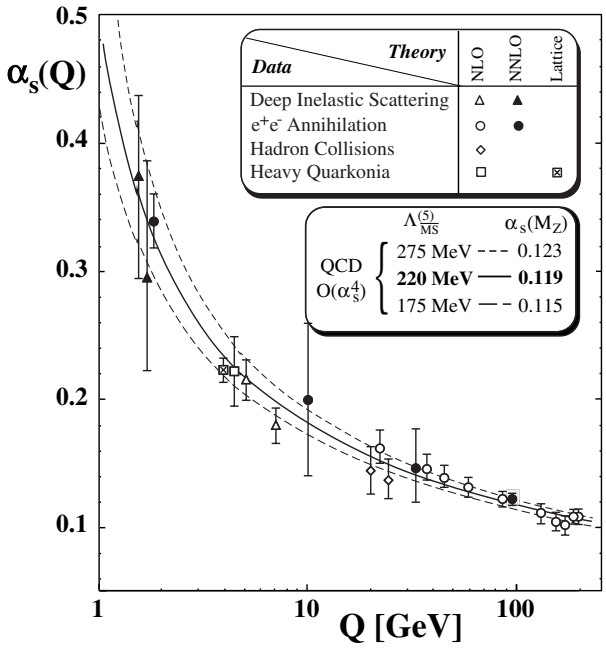


Figure 2. Weakening of strong interactions at high energies.

the value 0.119 ± 0.004 when measured at an energy corresponding to the mass of the Z^0 weak intermediate boson.

Tests of the electroweak sector of the Standard Model have been dominated by high-energy experiments on e^+e^- collisions at CERN, using the LEP accelerator, and at SLAC, using the SLC accelerator. A compilation of LEP measurements of the cross sections for e^+e^- annihilation into charged-lepton and quark–antiquark pairs is shown in figure 3. This is dominated by the Z^0 resonance peak, which must be one of the most carefully studied Breit–Wigner resonances in physics, with about 2×10^7 decays observed and measured. The height and width of the peak depend not only on the observed Z^0 decay modes but also on possible unseen decay modes such as those into neutrino–antineutrino pairs. Extra invisible decays would shorten the Z^0 lifetime and hence, by the UNCERTAINTY PRINCIPLE, increase the natural width of the Z^0 resonance peak. The measurements shown in figure 3 tell us that there are precisely three neutrino species, no more and no less. These correspond to the three known charged leptons, the electron, muon and tau. According to the Standard Model, there can only be three corresponding pairs of quarks. We therefore think there can be no further quarks beyond the six already known.

The data shown in figure 3 provide a very accurate determination of the mass of the Z^0: 91.1867 ± 0.0021 GeV. To arrive at this precision, it was necessary to take into account many effects that were considerably larger than the quoted error. These included terrestrial tides, which caused the LEP machine to expand and contract, and hence its energy to vary by $O(10)$ MeV, and even the passing

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

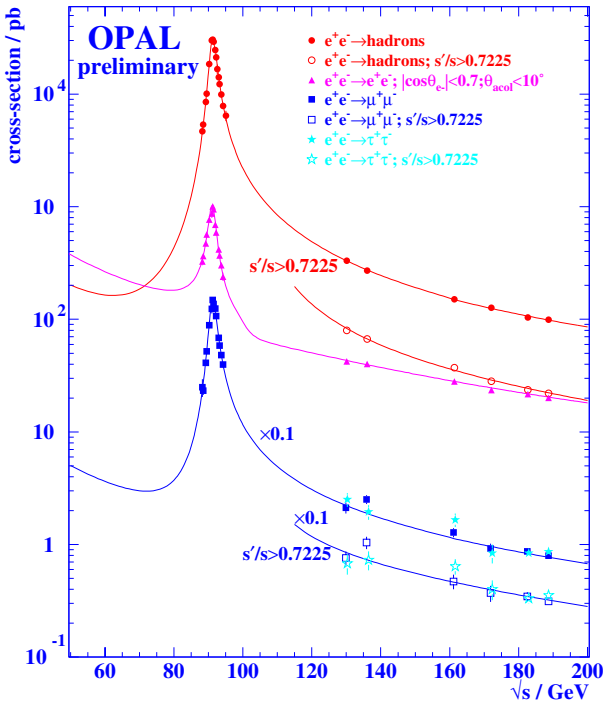


Figure 3. Cross sections for e^+e^- annihilation into charged-lepton and quark–antiquark pairs.

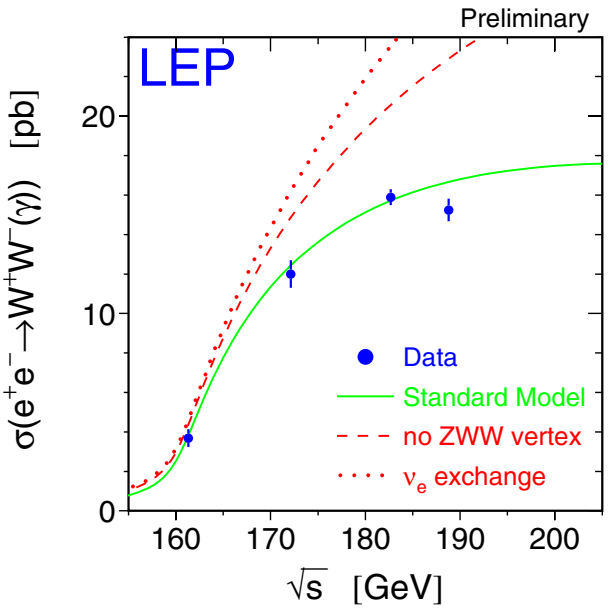


Figure 4. Production of W^+W^- boson pairs.

of nearby electric trains, which generated stray electric currents that perturbed the magnets of LEP and thus its beam energy.

In addition to the cross-section measurements shown in figure 3, LEP and the SLC have provided many other precision measurements, including branching ratios for Z^0 decays into different modes such as heavy quarks, forward–backward production asymmetries and polarization effects. Many of the measurements have per mille accuracy, and none differs significantly from the Standard Model prediction, and all measurements are within 1 or 2 standard deviations. These predictions require detailed calculations in quantum field theory, which were shown to be possible by 't Hooft in 1971. As was emphasized by Veltman, in particular, these calculations are sensitive to the masses of heavy particles that are too heavy to be produced directly at LEP or the SLC. The successes of these calculations were used successfully to predict the mass of the top quark, before it was discovered.

Referring back to figure 3, we see that LEP measurements extend to centre-of-mass energies considerably above the Z^0 peak. One of the principal Standard Model tests possible with higher-energy data is the production of pairs of $W^\pm$ bosons, as shown in figure 4. According to the Standard Model, this reaction receives contributions from the exchanges of the neutrino, photon and Z^0. The first of these is required by the role of the $W^\pm$ bosons in β decay, the second by the electric charge of the $W^\pm$ and the third by the characteristic $Z^0W^+W^-$ coupling of the unified electroweak theory. As seen in figure 4, all these three contributions are required to explain the LEP data, and the values of the couplings extracted from the data are in good agreement with the Standard Model. These data may also be used to measure the mass of the $W^\pm$, and the results are also in good agreement with the Standard Model predictions.

Thus far, we have seen that the Standard Model is passing its experimental tests with flying colors. However, there are still some key features of the Standard Model that have not yet been tested. One of these is the origin of the particle masses. According to the Standard Model, the underlying field theory is formulated in terms of massless particles, and one must find some way to incorporate their masses without spoiling the successes of the theory. This is possible if there exists an additional, as yet unseen, particle called the Higgs boson. The precision electroweak data described earlier are sensitive to the mass of this particle, and currently indicate that it weighs about 100 GeV, with an uncertainty of a factor of about 2. The search for the Higgs boson has been one of the continuing objectives of the LEP experimental program. At the time of writing, these searches have so far been unsuccessful and have established that its mass must exceed about 108 GeV. By the end of its operation, experiments at the LEP accelerator should tell us whether the Higgs boson weighs less than about 114 GeV. Beyond this mass, the full range of possible Higgs masses up to about 1 TeV can be explored at CERN's LHC accelerator, which is scheduled to start taking data in 2005.

Another missing element in tests of the Standard Model is its mechanism for matter–antimatter asymmetry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The weak interactions were shown in 1957 to violate not only parity (mirror) symmetry, but also charge conjugation. At first, it was thought that the combination of parity and charge conjugation, called CP, might yet be a good symmetry. However, an experiment in 1964 discovered that CP was also violated in the weak decays of some neutral composite particles called kaons. This means that matter does not behave in exactly the same way as ANTIMATTER, even if one reflects the experiment in a mirror in an attempt to recover the symmetry. In 1973, Kobayashi and Maskawa realized that this matter–antimatter asymmetry could be accommodated within the Standard Model if there were at least six quark species, as we now know to be the case. Subsequently, much theoretical effort has been devoted to evaluating the Standard Model predictions for matter–antimatter asymmetries in different processes, and there is an extensive experimental program to test these predictions. Follow-up experiments on neutral kaon decays are consistent with the Standard Model, but have not yet reached the sensitivity required to verify the characteristic predictions of the Standard Model. A number of experiments are planned to explore its predictions for CP asymmetries in the decays of mesons containing bottom quarks. There is good reason to expect that these experiments will be able to explore these matter–antimatter asymmetries at the levels predicted by the Standard Model.

It may be appropriate at this stage to mention some ways in which astrophysics and cosmology are sensitive to aspects of the Standard Model (see also PARTICLE ASTROPHYSICS). The energy production of main-sequence stars is controlled by nuclear physics and weak β decay, but details of the Standard Model are not directly relevant. These are, however, relevant to objects that are denser and/or hotter, such as collapsing supernovae. Weak neutral-current processes are essential to the understanding of neutrino emission from the collapsing core, and perhaps to the ejection of outer layers of the supernova progenitor. The nuclear equation of state of the remnant may also be sensitive to the underlying QCD theory of the strong interactions. In particular, it is often conjectured that under certain circumstances there may be a transition from conventional nuclear matter to quark matter, possibly with a high fraction of strange quarks.

Looking back to early cosmology, weak neutral-current processes also played an essential role in the reheating and decoupling of neutrinos around the epoch of big bang NUCLEOSYNTHESIS. Prior to that, in standard big bang cosmology the temperature of the early universe, and hence the characteristic particle energies, is believed to have increased as the inverse square of the age of the universe. When it was less than about 10^{-5} or 10^{-6} s old, hadrons would have been 'ionized' into quarks and gluons, and the Standard Model must be used to describe the behavior of the universe at these early times. The laboratory successes of the Standard Model described earlier imply that we possess the basis for a quantitative description of physics in the early universe.

When it was less than about 10^{-10} or 10^{-12} s old, it is thought that the Standard Model underwent a phase transition, before which all the particles of matter would have appeared massless, as well as the weak intermediate vector bosons and the Higgs boson. In addition to the familiar weak interactions, it is thought that other effects in the Glashow–Weinberg–Salam model, that have not been seen in the universe today, would have been important at this early epoch. These effects would have the distinctive feature of violating both baryon and lepton number by 3 units. Hence they must be taken into account in any attempt to understand the baryon density in the universe.

Although the Standard Model of particle physics is very successful, with no confirmed accelerator data that contradict it, there are many theoretical reasons to consider it unsatisfactory and to expect some physics beyond the Standard Model. For example, even if one accepts the quantum numbers of the quarks and leptons as given, the Standard Model contains 19 free parameters that are taken from experiment. These include three independent interaction strengths and a CP-violating parameter for the strong interactions, six quark masses and four parameters to describe how they are mixed by the weak interactions, three charged-lepton masses, the $W^{\pm}$ and Higgs boson masses. Having so many parameters is surely unacceptable for any candidate for a theory of everything, and attempts to go beyond the Standard Model typically try to simplify one of its aspects.

In a GRAND UNIFIED THEORY one regards the strong, weak and electromagnetic interactions as different aspects of a single interaction with a universal coupling strength. This is made possible by the fact that coupling strengths vary with energy, as exemplified by the asymptotic freedom of the strong interactions shown in figure 2. Calculations indicate that the coupling strengths may become equal at an energy around 10^{15}–10^{16} GeV. Many grand unified theories predict additional perturbative interactions that violate baryon and/or lepton number. These exotic interactions may cause baryons to decay and/or give neutrino masses. There have been many unsuccessful searches for proton decay, and its lifetime must be at least 10^{32} yr. Evidence has recently emerged from studies of solar and atmospheric neutrinos that they may have masses. If confirmed, this would constitute the first concrete evidence for physics beyond the Standard Model.

Other proposals for possible physics beyond the Standard Model try to understand the complicated pattern of lepton masses and quark masses and the way they are mixed by the weak interactions. One of the ideas often discussed in this context is the possibility that leptons and quarks are not elementary but composite bound states of more fundamental constituents. So far, no compelling model of this type has emerged, so it will not be discussed further here.

A third set of proposed extensions of the Standard Model arises from attempts to understand better the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observed magnitudes of particle masses. The principal issue here is that the physical values observed seem to require very delicate fine tuning of the parameters of the Standard Model, which strikes many physicists as unnatural. One way to avoid this is to postulate that the Standard Model particles are accompanied by partners with identical internal properties and interactions, but with spins differing by half a unit. These 'supersymmetric' particles would have opposite statistics from the familiar Standard Model particles—normal matter particles such as quarks and leptons are fermions, and would be accompanied by new bosonic particles, whereas intermediate bosons such as the photon would be accompanied by new fermions (see SUPERSYMMETRY). The fine-tuning problem of the Standard Model would be avoided if the supersymmetric partners of Standard Model particles weigh less than about 1 TeV.

These possible extensions of the Standard Model may provide the frameworks for resolving many of the fundamental issues in astrophysics and cosmology. One of these is DARK MATTER. The cold component may be provided by massive stable neutral particles, for which one candidate is the lightest supersymmetric particle, that is expected to weigh $O(100)$ GeV. This possibility will be explored by forthcoming accelerator searches at the proton–antiproton collider at FNAL and at the LHC proton–proton collider at CERN. A hot dark matter component may be provided by massive neutrinos. The indications for neutrino masses from the solar and atmospheric neutrino data do not, however, require any neutrino to be heavy enough to make a significant contribution to the overall dark matter density, although there could be some consequences observable by astronomers.

A second cosmological issue that may be addressed by extensions of the Standard Model is the origin of the matter in the universe. This can be calculated in terms of elementary-particle interactions if they violate baryon number, if they exhibit a matter–antimatter asymmetry and if they depart from thermal equilibrium. As has been discussed above, there are Standard Model interactions that violate baryon number and matter–antimatter asymmetry. Whether they departed from thermal equilibrium in the early universe depends in the Standard Model on the mass of the Higgs boson. The LEP lower limit on its mass, combined with theoretical calculations, indicates that there would have been no significant departure from thermal equilibrium if the Standard Model was sufficient to characterize early cosmology. Therefore some extension of the Standard Model appears necessary to explain the baryonic matter in the universe (see also BARYOGENESIS), which may be possible with grand unification and/or supersymmetry.

Grand unified theories may also be needed to explain the size, age and flatness of the universe via cosmological INFLATION. The measurements of perturbations in the COSMIC MICROWAVE BACKGROUND radiation suggest that they originate from physics at an energy scale of 10^{13}–10^{16} GeV. This is far beyond the scale of particle masses in the Standard Model and comparable with the scale of grand unification estimated using data from LEP and other accelerators.

Finally, we note that many theorists believe that the underlying theory of everything that resolves all the open problems of the Standard Model should be provided by some incarnation of STRING THEORY. This is an apparently consistent quantum theory of gravity, requires supersymmetry and is rich enough to accommodate a grand unified theory. If the self-coupling of the string is strong, it may provide a correct numerical estimate of the scale of grand unification. Astrophysics and cosmology, via inflation, gravitational waves and perhaps a cosmological constant may provide the most important tests of this ultimate extension of the Standard Model.

John Ellis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star Clusters

Even a small telescope shows obvious local concentrations of stars scattered around the sky. These star clusters are not chance juxtapositions of unrelated stars. They are, instead, physically associated groups of stars, moving together through the Galaxy. The stars in a cluster are held together either permanently or temporarily by their mutual gravitational attraction.

The classic, and best known, example of a star cluster is the Pleiades, visible in the evening sky in early winter (from the northern hemisphere) as a group of 7–10 stars (depending on one's eyesight). More than 600 Pleiades members have been identified telescopically.

Clusters are generally distinguished as being either GALACTIC OPEN CLUSTERS or GLOBULAR CLUSTERS, corresponding to their appearance as seen through a moderate-aperture telescope. Even in a relatively small telescope, the stars in an open cluster can be individually distinguished right at the center, whereas a globular cluster typically appears to consist of a sprinkling of bright, resolved stars superposed on a continuous background of unresolved fainter stars. Although the original distinction of the two classes of clusters was based almost entirely on their visual appearance, we now know that the globulars and the open clusters are very distinct from one another in almost every respect.

At the opposite extreme of concentration from the globular clusters are the stellar associations, loose, unbound groups of stars with particular spectral properties. Associations are normally observed as widespread regions with an excess density of spectral type O and B stars ('OB ASSOCIATIONS') or T TAURI STARS ('T associations'). In principle, there could be other types of association as well. It is not uncommon to find one or more open clusters at the core of a stellar association. The implication is that the stars in the central cluster or clusters and the surrounding association all formed more or less at the same time out of the same material. Since the association is not a gravitationally bound system, after a relatively short period of time, its stars will disperse and the association will lose its identity; for that reason, associations always consist of young stars.

The stars in an unbound association or a cluster that is evaporating will nevertheless all be moving with almost identical velocities, so at some later time, one might expect to find stars spread out in space, but still moving with similar velocities. Stars selected by velocity are sometimes called MOVING GROUPS. The best known of the moving groups is called the Hyades moving group. It consists of a number of stars, located in all directions from the Sun, which share the space motion of the Hyades star cluster. However, it is uncertain whether these stars share a common origin with the Hyades cluster.

The category of moving groups is not to be confused with moving clusters. These are clusters near enough to the Sun that all three dimensions of their motion through the Galaxy can be used to derive the cluster distance, independently of parallax measurements.

A final category of clusters has recently been identified based on observations with infrared detectors. These are the embedded clusters—star clusters still in the process of formation, and still embedded in the clouds out of which they formed. Because it is possible to see through the dust much better at infrared wavelengths than in the optical, these clusters have suddenly become observable.

Star clusters are of considerable astrophysical importance to probe models of stellar evolution and dynamics, explore the star formation process, calibrate the extragalactic distance scale and most importantly to measure the age and evolution of the Galaxy.

Definition of a star cluster—cluster catalogs

Just what *is* a cluster? How many stars does it take to make a cluster? Trumpler in 1930 defined open clusters as 'Star groupings which undoubtedly form physical systems (stars situated at the same distance and probably of the same origin) and which at the same time are sufficiently rich in stars for statistical investigation.' Since virtually all clusters have been discovered either by visual examination of the sky with a telescope, or from inspection of photographic or electronic images in the visual or infrared, an operational definition of a star cluster is suggested:

> A star cluster is defined to be an obvious concentration of several stars or more above the surrounding stellar background, apparently localized in space and identifiable on visual or infrared images covering a suitable field of view.

While this definition may seem imprecise, it conveniently includes most of those objects (in our Galaxy at least) listed in the existing cluster catalogs. At the same time, it excludes galaxy-sized systems or objects with insignificant numbers of stars. Presumably most clusters defined in this way are physical systems, although they may or may not be gravitationally bound systems. So some of the embedded clusters or slightly older, optically revealed clusters may be unbound star-forming regions in the process of dispersing. The above definition avoids the difficulty of distinguishing bound systems from unbound ones—if it looks like a cluster, it is a cluster.

From the theoretical point of view, a cluster might be better defined as a self-gravitating system of stars residing within the Galactic gravitational potential. Using this definition, unbound embedded clusters or young optically revealed clusters would not be considered to be true clusters. In addition, a star that has achieved escape velocity from a cluster but is still within the tidal radius of the cluster would be considered to be a nonmember by this theoretical definition, but would be a member by the observational definition.

Few objects have been catalogued as extensively as star clusters. The MESSIER CATALOG of 107 nonstellar objects includes 59 objects now known to be star clusters. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

first major catalog that includes clusters is the NEW GENERAL CATALOG (NGC), listing 487 clusters. The update to the NGC, the IC, contains another 30 clusters.

There have been at least a couple of dozen other more or less systematic attempts to search for and catalog additional star clusters. The entire sky has now been searched for clusters to the limiting magnitude and resolution of the wide-field Schmidt telescope photographic surveys. Unfortunately, the large number of partly redundent catalogs has generated a considerable amount of confusion about cluster names, and some clusters have almost a dozen aliases. The existing cluster catalogs are almost certainly biased in various ways, and any statistical results based on them are subject to systematic errors and selection effects of one sort or another.

As the number of catalogs has grown, and as the astrophysical importance of star clusters has been more appreciated, the number of secondary catalogs, summarizing the known properties of clusters, has also grown. Some useful catalogs are given in the Bibliography.

Summary of cluster properties

Table 1 summarizes the properties of star clusters. While there is a considerable dispersion in the properties of individual clusters, as well as a substantial uncertainty in some of the values, these numbers are representative of the cluster systems.

Table 1 shows that there is a rather sharp distinction between the globular clusters, which are representative of the Population II stars of the galaxy, and the open clusters (including the OB associations, moving groups and embedded clusters), which are Population I stars. The most massive of the open clusters are roughly the same as the least massive globulars, and the oldest of the open clusters is no older than the youngest of the globulars.

Globular clusters

The known globular clusters in our Galaxy number about 147 and, taking into account obscured regions of the Galaxy, the total Galactic population of globulars is probably about 200. Most of the 'missing' globulars are likely to be located close to the Galactic center or near the Galactic plane on the far side of the Galaxy.

The globulars are substantial stellar systems. Even the smallest of them contain some 10 000 solar masses of stars and a few, such as M22 and Omega Centauri, exceed one million solar masses.

As described later in this article (see also STELLAR DYNAMICS), the stellar distribution within a globular cluster is well represented by King models, which can be characterized by two parameters: the central surface brightness and the core radius, defined to be the radius at which the brightness drops to one-half the central value. Once these two parameters are determined, the appropriate King model will yield a tidal radius for the cluster. The globular clusters cover a substantial range in size. The median core radius is close to one parsec, but a few clusters in the outer halo of the Galaxy have core radii in excess of 10 parsecs. The median tidal radius is about 35 parsecs, but again there are some in the outer halo with much larger tidal radii.

Open clusters

Some 1200 open clusters have been catalogued, but as many as 250 of these appear to be simply slight enhancements in the stellar density along a line of sight or small asterisms of unrelated stars. Such asterisms should not be considered as clusters by either of the definitions of clusters given above. Most of the open clusters in the Galaxy have probably not yet been found, since open clusters are found near the Galactic plane where the interstellar dust is the most opaque. There may be as many as 10^5 open clusters in the Galaxy.

The smallest of the open clusters contain fewer than a dozen stars, so the minimum mass for an open cluster is about $10\,M_\odot$. At the other extreme, some of the most massive open clusters are old ones such as NGC 6791, with more than 10 000 members, again with typical masses near one solar mass. The most populous younger clusters contain a few thousand stars, but since their most massive stars are substantially more massive than the Sun, the total mass of some of the younger clusters exceeds $10^4\,M_\odot$.

Few if any open clusters have enough members to be characterized in terms of dynamically meaningful quantities, such as the core radius or tidal radius. Instead, open cluster sizes are based on the 'angular diameter' measurement where the diameter is the apparent size of the cluster as measured on the appropriate sky survey photographs. Even when the distance to the cluster is well-known, the angular diameter does not yield a well-defined linear dimension to the cluster. Nevertheless, there is a rather small spread in measured cluster diameters, the typical open cluster being between 4 and 5 pc in diameter. A few clusters are more than twice this size, and some are only one or two parsecs in diameter, but the majority fall rather close to this value. The core radius of open clusters is probably roughly 1 pc, that is to say, about the same as the typical core radii of globular clusters. There is no strong correlation of cluster size with age, although several of the oldest clusters are somewhat larger and some of the very youngest clusters are also larger. Nor is there any correlation of cluster size with Galactic location or with the number of members.

OB associations

OB associations are areas on the sky with a higher than normal density of O and B stars, They are both large and diffuse; they are in fact sufficiently amorphous that distinct boundaries are hard to define. They typically cover many tens of parsecs and often contain one or more discrete clusters within their borders. Their diffuse nature makes it impossible to identify fainter stars as being members of associations, although there may well be large numbers of lower-mass stars that formed from the same molecular cloud.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Summary properties of star clusters.

	Globular clusters	Open clusters	OB associations	Moving groups	Embedded clusters
Numbers of clusters:					
Catalogued	147	1200	70	< 10	~ 25
Likely Galactic total	200	10^5	> 1000	—	—
Typical Sizes (pc):					
Core radius (median)	1.0	—	unbound	unbound	—
Tidal radius (median)	35	—	unbound	unbound	—
Apparent diameter	—	4	> 100	—	1
Masses ($M_\odot$):					
Minimum	10^4	10	1000	—	100
Maximum	10^6	10^4	10^4	—	> 1000
Ages (yr):					
Minimum	8×10^9	5×10^6	5×10^6	—	$< 10^6$
Maximum	16×10^9	9×10^9	2×10^7	—	5×10^6
Metallicities ([Fe/H]):					
Range	−2.3 to +0.2	−0.7 to 0.3	Pop I	Pop I	Pop I

Approximately 70 OB associations have been catalogued. An association may consist of as few as a dozen luminous O and B stars, but because they range in mass from about 10 $M_\odot$ to some 70 $M_\odot$, the total mass of the association may range from about 1000 $M_\odot$ to more than ten times that number.

Moving groups

Several moving groups have been identified, but because the group may be spread out over a large region of the galaxy, the membership of a group is likely to be seriously incomplete, assuming that the group has a common origin at all. No reliable estimates of the total mass of a moving group exist.

Embedded clusters

At the other extreme of size from associations are the embedded clusters. This newly identified class of object consists of groups of stars still deeply embedded in the nebulosity out of which they have just formed. They range from a few tenths of a parsec to about two parsces in diameter. They are found only through IR imaging surveys—it is important to stress that these surveys are now sufficiently extensive that the apparent small size of the embedded clusters is not just a result of inadequate sampling. They really are rather compact objects. Although a few, like the Orion Trapezium region, are rich systems with hundreds of stars, most of them consist of small aggregates of stars.

No comprehensive catalog exists, but as many as 25 embedded clusters have been described; the total Galactic population of embedded clusters probably exceeds this number by several orders of magnitude. Since the stars in an embedded cluster have not yet reached the main sequence, even the relatively low-mass stars are rather luminous. Consequently they are relatively easily observable (in the infrared). They are compact, very dense regions, containing hundreds to thousands of solar masses of stars.

If the embedded clusters are all smaller than the ordinary, optically visible open clusters, does the one class evolve into the other? That is to say, when the molecular clouds surrounding these embedded clusters dissipate in a few million years, will there be an apparent ordinary open cluster there, or will the stars dissipate as fast as the gas? The answer to this question is not yet known and is the subject of much discussion at the present time. Embedded clusters are sometimes called 'stellar aggregates' rather than clusters because of this ambiguity, but based on the definition of a cluster given here they should be considered as star clusters.

Cluster distances and the cosmic distance scale

Star clusters provide a critical step in the determination of the cosmic distance scale. The starting point for cluster distances, and a primary step in the distance scale, is the Hyades star cluster. As a result of the HIPPARCOS project, the distances to the individual stars in the cluster have been measured well enough that the three-dimensional structure of the cluster can be determined. The Hipparcos data plus additional support from other measurements show that the Hyades distance modulus is $(m - M)_o = 3.33 \pm 0.01$, that is, the center of the Hyades cluster is 46.3 ± 0.2 pc from the Sun.

To find the distances to more remote clusters, a standard technique is to relate the properties of the Hyades to those of the cluster in question. The primary mechanism for that comparison is the color-magnitude diagram, a form of HERTZSPRUNG–RUSSELL DIAGRAM. The Hertzsprung–Russell diagram shows the relation between the luminosities of stars and their temperatures; in a color–magnitude diagram the luminosity is presented as either the absolute or apparent magnitude and the temperature is represented by a color index. The color–magnitude

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

diagrams of two clusters will in general have the same shape, but will be shifted in magnitude by an amount equal to the difference in distance modulus between the two. There may also be additional shifts in magnitude and color index resulting from interstellar absorption. The general procedure is to apply the effects of absorption, if known, and then slide the diagrams of the two clusters until they match, according to some goodness-of-fit criterion.

The principal limitations of this method are the accuracy of the magnitudes and colors of the stars in both the Hyades and the target cluster, the amount of interstellar absorption to the target cluster (the interstellar absorption to the Hyades is effectively zero) and the effects of chemical composition on the location of the main sequence. The Hyades is rather unusually metal-rich (approximately [Fe/H] = +0.15) compared with most star clusters, which must be taken into account.

Figure 1 shows the color–magnitude diagrams of several star clusters, corrected for distance. M92 is an old, metal-poor globular cluster; 47 Tucanae is another old globular cluster, but less metal-poor; M67 is an old open cluster of solar composition, and the Hyades is a younger metal-rich open cluster. For comparison, the zero-age main sequence (ZAMS) for solar composition stars is shown as well. This diagram presents the combined effects of stellar evolution and chemical composition. The M92 main sequence lies the lowest in the diagram, as would be expected from its composition, 47 Tuc is next, then M67 and finally there are the individual Hyades stars, the most metal-rich ones in the diagram. The ZAMS line corresponds approximately to solar-compositon stars.

Figure 1 also shows that it is critical to have the right distance to a cluster. Except for the horizontal branches, the shapes of the color–magnitude diagrams for the two globular clusters are virtually identical. Unless both the distance and composition of a cluster are known very well, the age uncertainties that result are large.

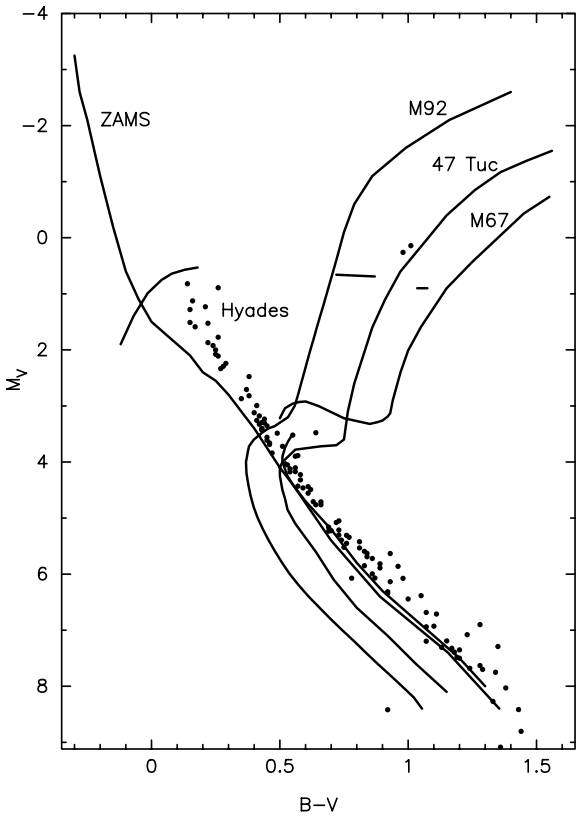


Figure 1. A composite of color–magnitude diagrams of the globular clusters M92 (from Stetson and Harris 1988) and 47 Tucanae (Hesser *et al* 1987) and the open clusters M67 (Montgomery *et al* 1993) and the Hyades (Perryman *et al* 1998). The zero-age main sequence (ZAMS) of Schmidt-Kaler (1982) is shown for comparison.

Clusters and stellar evolution—cluster ages

Star clusters represent snapshots of the process of stellar evolution—they are frozen in time from a human perspective. Since all the stars in a star-forming region begin their protostellar phase within a few million years of one another, the stars in a cluster much older than a few tens of millions of years can be considered to be all the same age. Thus the color–magnitude diagram of a star cluster represents the properties of stars differing in mass, but all of the same age and chemical composition. A critical test of the validity of STELLAR EVOLUTION models is to be able to predict the observed properties of cluster stars, as seen in the color–magnitude diagram.

To compare the models with the actual stars in a cluster, a series of model evolutionary sequences is computed for stars selected to have a range of masses. An isochrone is derived by interpolating in this series of models to find the parameters for stars all the same age. The isochrone is in effect a theoretical color–magnitude diagram.

In going from the theoretical predictions of the model isochrone to the observed color–magnitude diagram, there are a number of serious complications. In addition to the various physical and numerical uncertainties of the models themselves, a major complication relates to the transformation of the theoretical parameters, effective temperature and luminosity, into the derived absolute magnitude and color index. To make this transformation, a good knowledge of the stellar flux distribution is needed (a good model atmosphere), plus a knowledge of the details of the spectral response of a typical photometric system. The latter includes atmospheric transmission, telescope plus filter spectral response and the spectral sensitivity of the detector, all of these integrated over the flux distribution of the model star.

The problem of measuring the ages of star clusters appears to be a simple one. As figure 1 shows, the main sequence turn-off of a star cluster—usually defined as bluest point (i.e., the smallest value of the color index) along the cluster sequence—is a function of the age of the cluster. This follows naturally from the basic fact of stellar evolution that the massive, more luminous,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars evolve away from the main sequence before the less massive stars have significantly changed. But at a given mass the actual age when this occurs depends on the chemical composition of the star, and possibly other factors as well, such as rotation. There are a number of theoretical uncertainties in addition to the observational uncertainties. Furthermore, as can be seen in figure 1, the shapes of the main-sequence/turn-off/red-giant region of old cluster color–magnitude diagrams are all roughly the same. So errors in the distance, interstellar reddening or composition can lead to a cluster being matched to a stellar model of the wrong age.

The most important of these difficulties is the distance. For example, if a cluster is measured to be too close by, say, 10%, then the measured luminosities of its stars will be too small by 21%. But the luminosity of a star is just a measure of the rate at which it is burning its nuclear fuel. So in this case, the distance error would mean that the nuclear-burning rate will be underestimated, and the cluster age would be overestimated by that same 21%. Very few clusters have their distances known to 10%, so while there are additional constraints that can be put on the age calculation, the uncertainties are still substantial.

The distance problem enters in two ways. First, the distances to individual clusters have random errors relating to the particular data for that cluster, but of more significance to astrophysics is the problem of the distance scale itself. The Hipparcos satellite has helped the situation considerably, but there is still considerable controversy over the ages of globular clusters.

Galactic distributions of clusters

Globular clusters

The globular clusters, containing the oldest stars in the Galaxy, are located in a nearly spherical distribution about the Galactic center. They make up the most visible component of the Galactic halo, and as a practical matter they define the stellar halo. Because they can be identified so readily and their distances measured reliably, Harlow Shapley was able to determine the direction and distance to the Galactic center in the 1920s. The globular clusters were later shown to define the Population II stars of the Galaxy.

Close examination of the statistics of the globular cluster properties shows that they can be divided into three subpopulations on the basis of cluster composition and the shape of the horizontal branch (see HORIZONTAL-BRANCH STARS). The 'old halo' clusters have horizontal branches consistent with a great age; the 'young halo' clusters have horizontal branches sugessting a somewhat younger age; and the 'disk globulars' are rather more metal-rich (with [Fe/H] > -0.8) than the other populations and are distributed in a thick disk shape.

The three subpopulations of clusters may have originated in two or three distinct processes. The old halo clusters probably formed in the initial collapse of the Galaxy and so their stars are the oldest. The disk globular clusters appear to be members of the thick disk population and may be significantly younger in addition to being more metal-rich. The young halo clusters may actually have been formed outside our Galaxy in nearby dwarf spheriodal galaxies that subsequently merged with our own Galaxy. The young halo clusters show no evidence for a correlation between Galactic position and composition, consistent with their having been injected into the Galaxy in random events. The old halo and disk subsystems do show correlations between galactocentric radius and composition, and there is a possibility that the old halo and thick disk populations may form a single system. If indeed this is true, then there is a regular, systematic trend between composition and position in the old halo plus thick disk subsystems, as well as between composition and age.

Open clusters

The open clusters, including associations, moving groups and embedded clusters, belong to the disk population of the Galaxy (the Population I stars). The vast majority of them are located close to the Galactic plane. Consequently, most known open clusters are within a couple of kpc of the Sun.

Assuming a distribution perpendicular to the Galactic plane of the form

$$N \propto \mathrm{e}^{z/z_0}$$

the scale length, z_0, for clusters with ages < 700 million years is about 55 pc, whereas the older clusters have a scale height of 350 pc. There is otherwise no correlation between composition and age or distance from the Galactic plane. There is, however, a distinct correlation between [Fe/H] and galactocentric radius, with the most metal-poor stars located in the far outer reaches of the Galactic disk.

The open clusters show no systematic trend of metallicity with age. At each age, there is a substantial range in composition, but on the average the oldest ones are approximately the same composition as the youngest ones.

Cluster dynamics

Star clusters are ideal laboratories for studies of stellar dynamics—they are simple systems consisting of what are essentially point masses, moving freely under the action of a single force, gravity. Consequently it is possible to model in some detail the dynamical evolution of star clusters. It is important to note, however, that the formalism of classical celestial mechanics is not directly applicable to star clusters. In celestial mcahanics, one usually considers the motion of one body in the potential field of another body (the two-body problem) or perhaps two other bodies (the three-body problem). However, a cluster consists of several (i.e., more than three) more or less equal masses moving in their mutual potential field, perhaps also perturbed by external (Galactic) forces.

Most dynamical models begin with the assumption that they are collisionless systems—that is, the forces acting on individual particles (stars) are long-range forces only, resulting from the gravitational potential field of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

aggregate of all the stars in the system. The occasional close encounters between pairs of stars, if they are included at all, are dealt with separately. More distant encounters may, however, be common in a stellar system. These encounters will gradually change the motion of an individual star, so that after enough such encounters, the star will have changed in velocity by an amount equal to the velocity. The time it takes for such a complete exchange of energy to take place is the relaxation time.

King models

Star clusters are approximately equivalent to isothermal spheres, that is to say, the velocities of stars in the cluster have a distribution that is independent, or nearly so, of the position of the star in the cluster. However, a singular isothermal sphere has infinite density at the origin (i.e, the center) and therefore infinite mass. To represent an isothermal sphere in a well-behaved manner at the origin, it is possible to define dimensionless variables, where the radius variable

$$\tilde{r} \equiv \frac{r}{r_0}$$

is defined by the King radius,

$$r_0 \equiv \sqrt{\frac{9\sigma^2}{4\pi G\rho_0}}.$$

At the radius r_0, the density within the system has fallen to 0.5013 of its central density. Since the core radius, r_c, as defined observationally is the radius at which the density has fallen to just half the central density, r_0 is often called the core radius.

For a system with gravitational potential Φ, where a single particle has velocity v and total energy E, the relative potential Ψ and relative energy $\mathcal{E}$ of the particle can be defined as

$$\Psi \equiv -\Phi + \Phi_0 \quad \text{and} \quad \mathcal{E} \equiv -E + \Phi_0 = \Phi - \tfrac{1}{2}v^2$$

where Φ_0 is some constant. A modified form of an isothermal sphere is the King model, which resembles the isothermal sphere at small radii, where most of the particles (stars) have large values of $\mathcal{E}$, but gradually becomes less dense than the isothermal sphere at larger radii where $\mathcal{E}$ tends to be lower. By suitably defining the constant Φ_0 so that at some critical relative energy $\mathcal{E}_0 = 0$, the phase-space distribution function of the King model can be defined so that for $\mathcal{E} \gg 0$ the King model and the isothermal sphere are identical, and for $\mathcal{E} \leq 0$ the King model distribution function is zero.

This leads to a family of models for the density distribution within the system that are defined by the central relative potential, $\Psi(0)$, where the radius is in units of the King radius, r_0. At the radius where Ψ falls to zero, the density also falls to zero. This is the tidal radius, r_t. The log of the ratio of the tidal radius to the King radius defines the concentration of the cluster,

$$c \equiv \log_{10}(r_t/r_0).$$

King models with c between 0.75 and 1.75 fit globular clusters well.

Observationally, the central surface brightness, Σ_K, and the core radius, r_c, can be used in place of $\Psi(0)$ and r_0.

Bibliography

Much of the information from older cluster catalogs and other sources is now online in the SIMBAD database.

A rather complete catalog of information on the known globular clusters is maintained by W E Harris on the website:

http://physun.physics.mcmaster.ca/Globular.html

This catalog is also discussed in

Harris W E 1996 A catalogue of parameters for globular clusters in the Milky Way *Astron. J.* **112** 1487

The ages of the globular clusters are discussed in

Chaboyer B, Demarque P, Kernan P J, Krauss L M 1998 The age of globular clusters in light of Hipparcos: resolving the age problem? *Astrophys. J.* **494** 96

The dynamics of the globular clusters are described in detail in

Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton: Princeton University Press)

Open cluster data can be found in the SIMBAD database, which includes the *Catalogue of Open Cluster Data, Fifth Edition* (1987) compiled by G Lynga (Lund Observatory).

A database of photometric and spectroscopic information on open clusters called WEBDA is maintained at the Geneva Observatory at:

http://obswww.unige.ch

Stellar associations have been reviewed by

Garmany C D 1994 OB Associations: massive stars in context *Publ. Astron. Soc. Pacific* **106** 25

Some key references for individual clusters mentioned in the article are the following:

Hesser J E, Harris W E, VandenBerg D A, Allwright J W B, Shott P and Stetson P B 1987 *Publ. Astron. Soc. Pacific* **99** 739

Montgomery K A, Marschall L A and Janes K A 1993 *Astron. J.* **106** 181

Perryman M A C, Brown A G A, Lebreton Y, Gomez A, Turon C, Cayrel de Strobel G, Mermilliod J C, Robichon N, Kovalesky J and Crifo C 1998 *Astron. Astrophys.* **331** 81

Stetson P B and Harris W E 1988 *Astron. J.* **96** 909

Finally, data for the ZAMS can be found in

Schmidt-Kaler T 1982 *Landolt-Börnstein VI* vol 2b (Berlin: Springer) p 1

Kenneth Janes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star Diagonal

A device that fits in the draw-tube of a telescope to enable an object to be viewed at right angles to the direction in which the telescope is pointing. In such a device, light is reflected through a right-angle either by a flat mirror inclined at an angle of 45° to the optical axis or by means of total internal reflection in a prism. The resulting image is reversed left-to-right.

It is often convenient to use a star diagonal when observing objects close to the zenith so as to avoid placing the observer's head at an awkward angle.

A related device, the solar diagonal, uses an unsilvered mirror, or a thin prism called a Herschel wedge, so as to reflect only a small proportion of the incoming solar radiation (typically about 5%), thereby reducing the amount of light and heat that reaches the eyepiece.

See also: telescope, zenith.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star Formation

The formation of stars is an ongoing process that continues to the present day. Star formation is most easily observed in the Milky Way, and these observations, the focus of this article, have led to a working model for the formation of low mass stars like our Sun. In this picture, stars form from the gravitational collapse of the cold, molecular interstellar material (gas and dust) that resides in structures known as INTERSTELLAR MOLECULAR CLOUDS. In the dynamical collapse and ACCRETION of infalling matter, a circumstellar disk forms around the accreting star as a consequence of the finite ANGULAR MOMENTUM of the molecular cloud. Stars obtain a significant fraction of their mass by accretion through their disks, and disk matter is a reservoir for the formation of stellar and planetary mass companions (see ACCRETION DISKS). Additional processes such as winds and magnetic accretion flows are believed to be responsible for the removal of angular momentum that allows the star to continue to grow in mass. These processes are combined in our current view of how Sun-like stars obtain their initial masses and angular momenta, i.e., the initial conditions for subsequent STELLAR EVOLUTION. The extent to which this picture, developed to explain the formation of low mass stars, applies to the formation of more massive stars, is an active area of research.

Observed properties of young stars

Infall

As is well known, molecular clouds are the reservoirs of mass and angular momentum from which stars form. However, while it has been appreciated for some time that star formation probably involves a phase of dynamical collapse (rather than quasi-static contraction, for example), the actual detection of infalling motion associated with the star formation process has only been made convincingly in the last few years. Since the mid-1970s, the many molecular line studies of molecular cores, the centrally concentrated regions within molecular clouds from which stars form, have typically revealed evidence for gas engaged in outflowing rather than infalling motion. The now apparent ubiquity of stellar outflows has remained one of the primary obstacles to the detection of infall owing to their often significant impact on the dynamics of the gas surrounding the star.

As a result, the successful detection of infalling motion has hinged on the choice of appropriate targets and suitable line diagnostics. The very young, so-called 'class 0', sources have turned out to be particularly fruitful sources for study. This is in part because the outflows in these sources are less well developed than in older sources and are often directed in the plane of the sky, thereby minimizing their contribution to line profiles.

The signature of dynamical infall seen in these sources is a double-peaked line profile with the blue peak stronger than the red (e.g. Zhou *et al* 1993; see also figure 1). This characteristic profile arises from the centrally peaked density and temperature distributions of the infalling gas, which produce an outwardly decreasing excitation gradient. In the simple but illustrative case of spherical infall, the central absorption is produced by low-velocity, low-excitation foreground gas in the core which is seen in absorption against the emission from more centrally located, higher-excitation gas. Because in general low-excitation, redshifted gas is seen in absorption against higher-excitation, redshifted gas further along the line of sight, while the situation is reversed for the blueshifted gas (higher-excitation gas lies in front of lower-excitation gas), the red side of the profile is preferentially weakened relative to the blue side (see, e.g., figure 8 in Zhou *et al* 1993). A key factor in the detection of this dynamical signature is the choice of appropriate line diagnostics. These must be sufficiently excited in regions where the excitation gradient is large and have optical depths sufficient to manifest the effect.

As described by Zhou *et al* (1993) in their multiline transitional study of the class 0 source B335, the theoretical solution of spherically symmetric dynamical collapse (Shu 1977), when coupled with a radiative transfer analysis, provides an excellent fit to the suite of observed CS and H_2CO line profiles (see also Choi *et al* 1995; figure 1). Because physical situations other than infall can also produce a double-peaked profile (e.g. broad emission plus an unrelated foreground absorber; alternatively, outflow), multitransitional studies are important in confirming infall and in ruling out other possibilities. For example, in regions undergoing infall, higher-excitation lines are predicted to be broader than lower-excitation lines, an effect which would not be observed in the absence of systematic motions. Additionally, stronger blueward line asymmetries are predicted for more optically thick lines in the case of infall, an effect which is observed.

Since the infall velocity gradient is measurable, albeit only on small size scales ($\sim 20''$ for the nearest sources), spatial maps of line emission from molecular envelopes have been obtained in order to confirm that the gas is undergoing dynamical infall. Such studies find larger linewidths and line asymmetries closer to the forming star (e.g. Myers *et al* 1995) and can even be used to diagnose the presence of rotational motion in the infalling gas (Zhou *et al* 1996). Line profiles observed in the latter case are in good agreement with theoretical models of dynamical collapse that include the effects of rotation (Terebey *et al* 1984).

From the modeling of multiple line profiles, it is possible to extract physical and dynamical information about the infalling gas, such as the mass accretion rate, the mass of the star + disk and the time period over which collapse has been occuring (e.g. Zhou *et al* 1993; Choi *et al* 1995; Zhou *et al* 1996). In a recent development, infalling motions have been detected in a molecular cloud core that is not known to have a star embedded within it (Tafalla *et al* 1998). This development opens up the possibility of studying the very early stages of the star formation process using the above techniques. The detail with which the infall process can now be studied and the close agreement between observations and theoretical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

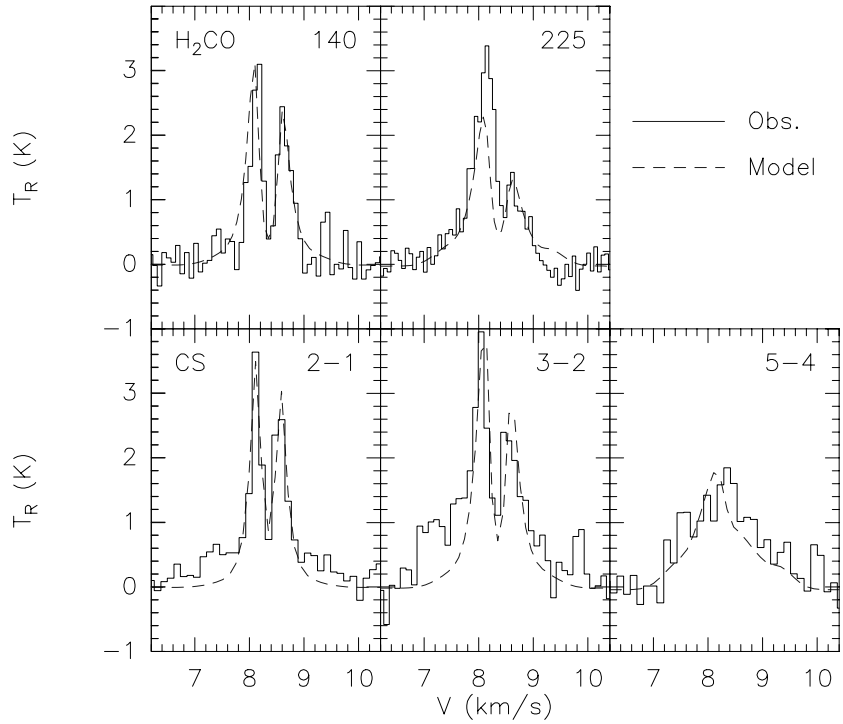


Figure 1. Rotational transitions of H_2CO (140 and 225 GHz) and CS ($J = 2$–1, 3–2 and 5–4) observed toward the very young class 0 source B335 (histogram) and the best-fit model spectra (dashed curve) under the assumption of dynamical 'inside-out' collapse (Choi *et al* 1995). The close agreement between the data and the theoretical model provides strong evidence for a phase of dynamical collapse in the process of star formation.

predictions together appear to provide the long-sought observational evidence that stars build up their masses and angular momenta through molecular cloud collapse and a dynamical accretion process.

Outflow

As described above, outflows are a ubiquitous and unexpected signature of star formation. At the very early stages of stellar evolution, when the star is still embedded in the molecular cloud from which it formed, the outflow process manifests itself in several ways. On the largest size scales (parsecs) are molecular outflows which are typically detected in molecular emission lines at millimeter wavelengths (e.g. Lada 1985; Bachiller 1996). The outflows commonly display a bipolar morphology with physically separated lobes of cold red- and blue shifted gas (e.g. L1551—Snell and Schloerb 1985). The low velocities ($\sim$10 km s^{-1}) and large masses ($\gtrsim 1 M_\odot$) of the outflows indicate that they are swept-up molecular cloud material. This interpretation is sometimes supported by molecular line observations which show that the outflowing gas is distributed in a shell-like structure that surrounds a lower-density cavity (e.g. L1551—Moriarty-Schieven *et al* 1987). Since the large momenta of the outflows suggest that they may reverse the motion of the infalling gas and thereby limit the mass of the star, the origin of outflows is of considerable interest.

Support for the idea that molecular outflows are swept up by an energetic wind originating close to the star comes from the morphology of REFLECTION NEBULAE within $\sim$1000 AU of the star which also imply a cleared polar region (e.g. L1551—Davis *et al* 1995). Often collinear with the lobes of molecular outflows are highly collimated emission line jets (e.g. L1551—Davis *et al* 1995) which are detected at optical and now IR wavelengths (see e.g. Edwards *et al* 1993; Stanke *et al* 1998). The jets have terminal velocities of hundreds of km s^{-1}, as indicated by both radial velocities and the proper motions of jet features. Recent wide-field imaging of star forming regions has revealed that jets can extend over very large size scales, comparable with the extent of molecular outflows (e.g. Reipurth *et al* 1997). On much smaller size scales, HUBBLE SPACE TELESCOPE (HST) observations of the emission line jets associated with the nearby sources HH30 and DG Tau (Burrows *et al* 1996; Kepner *et al* 1993) demonstrate that the flows are collimated on solar system size scales, within 40 AU of the star.

At the smallest size scales, as probed by high-resolution spectroscopy, are dense, high-velocity (hundreds of km s^{-1}) winds which produce the strong blueshifted absorption features in the profiles of lines such as Hα and Na I D (e.g. L1551—Mundt *et al* 1985; see also Mundt 1984). These wind signatures are observed not only among the young, embedded population but also

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

among more evolved, optically revealed young stars (see 'Disk accretion rates and stellar inflows' below). The densities and temperatures required to excite the hydrogen and Na I D lines strongly favor formation within $\sim 10 R_*$ of the star. When combined with the large velocities of the broad blueshifted absorption features, the implication is that winds are accelerated to high speed within a few stellar radii of the stellar surface. For the most active of the optically revealed young stars, winds are inferred to be cool ($\sim$6000 K) and energetic ($\sim 10^{-7} M_\odot$ yr^{-1}) (e.g. Giovanardi *et al* 1991; Najita *et al* 1996b). The momenta contained in the wind appear sufficient to drive the associated molecular outflows. For example, the dynamical lifetime of molecular outflows (ratio of linear extent to outflow velocity) is typically 10^5 yr. A wind with the above properties blowing for this length of time would provide a momentum input of a few $M_\odot$ km s^{-1} which is comparable with the momentum contained in the associated molecular outflows (cf Levreault 1988). We will return in 'Origin of initial stellar mass and angular momenta' to the origin of winds and jets and their role in the solution of the angular momentum problem for star formation.

Circumstellar disks

Because of the finite angular momentum of molecular cloud cores (e.g. Goodman *et al* 1993), cloud collapse is expected to result in the formation of a circumstellar disk with a characteristic size comparable with that of the solar system. In addition to their role as the reservoir of mass and angular momentum from which stars accrete a substantial fraction of their mass, the likelihood that disks are also the environments in which planets form has led to long-standing interest in their existence and physical properties.

One of the earliest studied and most commonly observed signatures of disks is the infrared excesses that disks contribute to the spectral energy distribution of the young star system owing to the processes of active accretion (e.g. Lynden-Bell and Pringle 1974) and passive reprocessing of stellar light (e.g. Adams *et al* 1987). Simple modeling of disks as a radial sequence of blackbodies weighted by emitting area provides a good fit to spectral energy distributions and implies disk temperatures of thousands of K at a few R_* and temperatures of a few tens of K at 100 AU. If infrared excesses are taken as an indicator of the presence of a disk, disk dispersal times in nearby star forming regions are found to be a few Myr, although with a large dispersion among the sources in a given region (e.g. Kenyon and Hartmann 1995; Hillenbrand and Meyer 1999; Meyer *et al* 1999; figure 2).

While the agreement between spectral energy distributions and the predictions of simple physical models strongly suggests the existence of disks, we need to turn to techniques with greater spatial or spectral resolution in order to confirm their flattened, centrifugally supported nature. The excellent spatial resolution of HST has been used to provide perhaps the best evidence for a flattened, disk-like morphology. WFPC2 observations

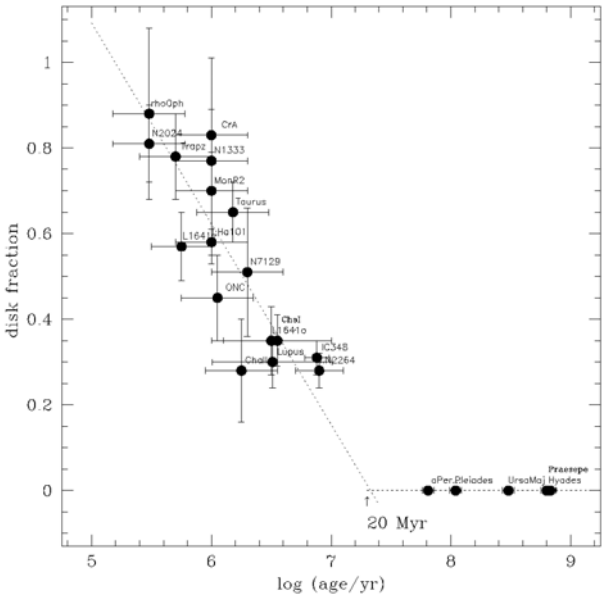


Figure 2. A preliminary result from Hillenbrand and Meyer (1999) which shows the fraction of stars that have disks, as diagnosed by near-infrared excesses, in clusters of different ages. The results appear to indicate that disks disperse on a timescale of <10 Myr.

of systems such as HH 30 (Burrows *et al* 1996) strongly suggest an edge-on disk system where the outline of a flared disk is revealed by scattered stellar light. The sillouette disks in the Orion nebula (McCaughrean and O'Dell 1996) also provide dramatic evidence for cool, dusty, disk-like structures 50–1000 AU in size that are seen in absorption against the bright emission line background of the nebula.

An alternative approach is the use of millimeter and submillimeter interferometry to study the structure and kinematics of disks on solar system size scales ($\sim$50 AU). The subarcsecond angular resolution now achieved by interferometers (e.g. the CSO–JCMT Interferometer[1], the Berkeley Illinois Maryland Association (BIMA) Array and the Very Large Array (VLA) at 7 mm) is well suited to the angular scales of solar system sized disks at the distances of nearby star forming regions ($\sim$150 pc distant). Multiwavelength continuum measurements are used to study the dust emissivity and column density distribution of the disk as a function of radius, from which disk masses can be estimated. For example, using the 40 milliarcsecond resolution of the VLA at 7 mm, Wilner *et al* (1999) showed that the dust surface density distribution of the disk associated with HL Tau has a shallow slope (flatter than $\propto r^{-1}$) over the range $\sim$5–100 AU (figure 3). The inferred mass of the disk is $\sim 0.6 M_\odot$ (e.g. Mundy *et al* 1996).

Disk kinematics at solar system size scales can be probed by interferometric studies of molecular line

[1] A system connecting the Caltech Submillimeter Observatory and the James Clerk Maxwell Telscope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

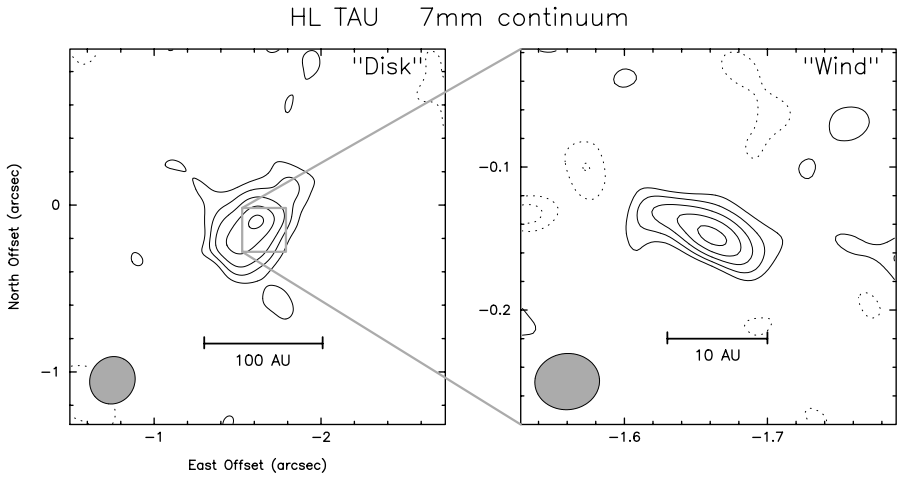


Figure 3. Images of the T Tauri star HL Tau obtained with the VLA at 7 mm with 40 milliarcsecond resolution (Wilner *et al* 1999). The right panel shows the reconstructed image obtained from data taken at baselines longer than 1000 kλ which reveals jet-like emission from an ionized wind. Given the angular scale of the image, this result suggests that jets are collimated very close to the star. The left panel shows a lower-resolution map obtained after subtracting the wind component, which reveals dust emission from a disk. The absence of a steep increase in surface brightness toward the star argues for a surface density distribution that is flatter than $\sim r^{-1}$.

emission at millimeter wavelengths. Such studies typically detect molecular line emission on size scales (∼500 AU) larger than that of the continuum emission. The kinematics of the emitting gas sometimes shows relatively clear evidence for rotation (e.g. Koerner and Sargent 1995), but the signature of radial inflow (e.g. Hayashi *et al* 1993) and/or outflow can often dominate the kinematics, making the detection of rotational motion difficult.

At much smaller size scales ($r < 5$ AU), disks are currently studied using infrared spectroscopy, where high spectral resolution is used in lieu of high spatial resolution to probe disk properties. Because of the high densities and warm temperatures (100–5000 K) of disks within 5 AU, molecules are abundant in the gas phase, and their infrared vibrational–rotational transitions sufficiently excited that molecular spectroscopy in the near- and mid-IR is a powerful probe of disk structure and dynamics. For example, $R = 20\,000$ spectroscopy of 2.3 μm CO overtone emission from young stars reveals a spectral shape that strongly suggests a disk origin (figure 4; Carr *et al* 1993; Najita *et al* 1996a; see also Chandler *et al* 1993). These observations currently provide some of the best evidence for the existence of centrifugally supported structures around young stars.

Disk accretion rates and stellar inflows

The rate of mass accretion through disks can be measured from the excess emission in both the infrared and the optical–UV that the accretion process contributes to the composite spectral energy distribution of the young stellar system. As described in the previous section, the infrared excesses arise in part from the dissipation of gravitational potential energy as disk matter spirals in toward the inner disk boundary (Lynden-Bell and Pringle 1974). While infrared excesses have frequently been used to infer the mass accretion rates of embedded sources, their utility as a measure of the disk accretion rate is compromised by the the fact that other physical processes, such as the reprocessing of stellar light by the disk (e.g. D'Alessio *et al* 1998) and other circumstellar matter (e.g. Natta 1995), also produce infrared excesses which must be accounted for in order to extract an accurate accretion rate.

In comparison, the UV excesses arise from the dissipation of additional gravitational potential energy as disk matter makes its way from the inner disk boundary onto the star. Since the origin of the emission is more uniquely associated with the accretion process, UV excesses can be regarded as a more robust measure of disk accretion rates (e.g. Bertout *et al* 1988; Calvet *et al* 1999). Because UV excesses are difficult to measure except in the case of optically visible accreting stars, i.e. the population known as the classical T TAURI STARS, this section and the next focus more specifically on the properties of these stars. The classical T Tauri stars are in a later phase of evolution than the embedded sources among which molecular outflows and stellar jets are more common (see 'Outflow' above) and are characterized by lower levels of mass accretion and mass loss.

Until the early 1990s, it was believed that circumstellar disks extended essentially up to stellar surfaces and that UV excesses arose from the dissipation of rotational energy as disk matter spun down in a boundary layer to join the slowly rotating star (i.e. a star rotating much below breakup). The hot boundary layer was then believed to produce the Balmer continuum emission and much of the rich emission line spectrum that is characteristic of T Tauri stars. However, the origin of these spectral features has recently been reinterpreted within a physical picture that offers an explanation for both the slow rotation rates of T Tauri stars and the origin of their winds.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

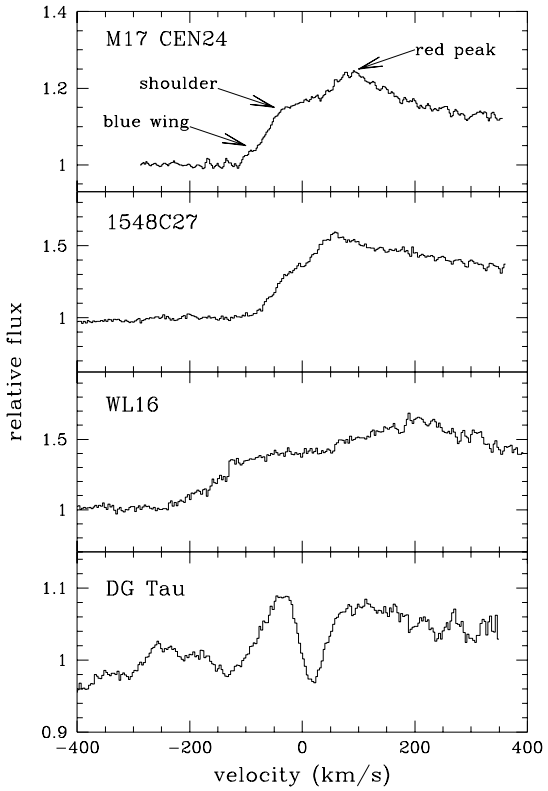


Figure 4. The spectroscopic signature of bandhead emission from a rotating disk (a blue wing, shoulder and red peak) is commonly observed in young stars over a range of masses, from the T Tauri star DG Tau ($\sim 1 M_\odot$), WL 16 ($\sim 2 M_\odot$) and 1548C27 ($\sim 4 M_\odot$), to a high-mass star ($\sim 10 M_\odot$) in the M17 cluster (Carr *et al* 1993; Najita *et al* 1996a; Carr *et al* 1999; Najita *et al* 1999). In the DG Tau spectrum, the signature is less clear owing to strong stellar CO bandhead absorption.

A variety of observations now support the view that accreting disk matter joins the star not through a boundary layer but through accretion along stellar magnetic field lines. Evidence for stellar inflows within a few stellar radii of the star comes from the inverse P Cygni profiles of the upper Balmer lines of some T Tauri stars (Walker 1972) which indicate infall velocities of several hundred km s^{-1}. Further evidence comes from the almost ubiquitous presence of blueward asymmetries in the upper Balmer and Brackett line profiles (Edwards *et al* 1994; Najita *et al* 1996b) which are interpreted as arising in gas infalling onto the star, a situation which results in preferential absorption on the red side of the line profile (cf 'Infall' above). The interpretation that these line profile properties are signatures of magnetic accretion is supported by RADIATIVE TRANSFER models of idealized magnetic accretion flows which reproduce many of the observed line profile features and trends (Calvet and Hartmann 1992; Muzerolle *et al* 1998a,b). In this revised picture, the post-shock gas at the base of the magnetic accretion flow is identified as the source of the optical–UV excess formerly attributed to emission from a boundary layer (Königl 1991; see also Bertout *et al* 1988).

Several other observations support the idea of a close dynamical relationship between stellar inflows and winds. Evidence for stellar inflows in the upper Balmer lines is often observed simultaneously with the signature of mass loss (i.e. P Cygni profiles) at Hα (cf 'Outflow' above; Edwards *et al* 1994). Since excitation considerations require that the hydrogen lines are formed close to the star, these observations provide evidence for the simultaneous presence of winds and stellar inflows within a few stellar radii of young stars. Additional evidence comes from synoptic observations of hydrogen line profiles in T Tauri stars. In the T Tauri star SU Aur, the depths of both a blueshifted absorption component (seen in Hα and Hβ) and a redshifted absorption component (seen in Hβ) are observed to vary at the 3 day rotation period of the star but are 180° out of phase (Johns and Basri 1995). These observations can be explained as the result of the interaction between the disk and a dipole stellar magnetic field that is tilted with respect to the stellar rotation axis (cf figure 15 in Johns and Basri 1995). Because stellar fieldlines bend inward toward the star, favoring inflow, in one region of the magnetosphere, and bend outward, favoring outflow, 180° away in azimuth, the rotation of the magnetosphere at the rotation period of the star can account for the observed period and phase difference.

Yet further evidence, of a more statistical nature, comes from the correlation between wind mass loss rates and stellar and disk accretion rates (Edwards *et al* 1993; Hartigan *et al* 1995). Although the errors associated with the measurements are large, current estimated values for these mass loss and mass accretion rates appear comparable within an order of magnitude. Gullbring *et al* (1998) have reexamined stellar accretion rates implied by the optical–UV spectrum of T Tauri stars in the context of the magnetic accretion scenario and find a broad distribution of accretion rates with a median value of $\sim 10^{-8} M_\odot$ yr^{-1}. Estimates of wind mass loss rates based on the profiles and fluxes of traditional diagnostics such as the hydrogen and Na I D lines (e.g. Calvet 1997; see also Giovanardi *et al* 1991; Najita *et al* 1996b) and optical forbidden line emission (e.g. Hartigan *et al* 1995) also result in a broad distribution averaging $\sim 10^{-9} M_\odot$ yr^{-1} (e.g. Calvet 1997), although there is considerably larger uncertainty associated with these values compared with the estimates for stellar accretion rates. A hypothesis for the simultaneous presence of mass accretion and loss, and their comparable mass flux rates, is presented in 'Origin of initial stellar masses and angular momenta' below.

Stellar rotation

The rotation rates of young stars are measured both from the rotational broadening of stellar photospheric absorption lines and from the observation of periodic photometric variability. The latter is attributed to rotational modulation of the stellar flux by the presence of hot or cool spots on the stellar surface. While cool

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spots are believed to be the equivalent of large sunspots, the hot spots are now interpreted as arising from the non-axisymmetric components of stellar inflows (Bertout *et al* 1988, 1996; see 'Disk accretion rates and stellar inflows' above). A multitude of rotational broadening and photometric variability studies have shown that the rotation rates of classical T Tauri stars are low, typically much less than the breakup rotation rate, $v \sin i \sim 15$ km s^{-1} or $\Omega_* \sim 0.1\Omega_{\rm breakup}$ (e.g. Vogel and Kuhi 1981; Hartmann and Stauffer 1989; Bouvier *et al* 1993). This result might be considered surprising since these stars are actively accreting, and the accretion of high angular momentum material from a disk might be expected to spin the star to breakup speed.

An important clue to the reason for the slow rotation rates of classical T Tauri stars came from the discovery that these stars, which possess circumstellar disks, rotate more slowly than stars of similar spectral type and age (the weak T Tauri stars) that show no evidence for circumstellar disks (Edwards *et al* 1993; Bouvier *et al* 1993; Choi & Herbst 1996). In these studies, classical T Tauri stars were found to have rotational periods in a fairly narrow range about ∼8 days, while weak T Tauri stars were found to have a mean period 2–4 times smaller. These results were interpreted as evidence that, in the T Tauri phase, stellar angular momenta are regulated by the presence of a disk, presumably through a kind of magnetic coupling between the stellar field and the inner circumstellar disk (Königl 1991; Shu *et al* 1994; Cameron and Campbell 1996; Armitage and Clarke 1996).

While the different theories describing this process disagree in their treatment of the radial extent of the coupling region and the nature of the angular momentum redistribution that allows continued slow stellar rotation, all of the theories share the idea that through disk coupling the star continues to rotate slowly even as it accretes some high angular momentum material from the disk and contracts toward the main sequence. In the context of this picture, the timescale for disk dispersal, and the consequent cessation of disk coupling, is an important factor in setting initial stellar angular momenta. Some degree of variation in initial stellar angular momenta would then appear to be expected, even at ages of a few Myr, given the known dispersion in disk dispersal times, as based, for example, on the intermixed distributions of classical and weak T Tauri stars in the HR diagram for the Taurus star forming region (Kenyon and Hartmann 1995).

These ideas have interesting implications when placed in the context of studies of stellar angular momentum evolution over Gyr timescales. For example, in figure 5 from Bouvier *et al* (1997a), the rotation rates of classical and weak T Tauri stars in Taurus are compared with those of post-T-Tauri stars in Taurus (10–40 Myr old), and stars in the αPer (50 Myr), Pleiades (100 Myr) and Hyades clusters. If interpreted as a pure evolutionary trend, these results might be taken to indicate that on average stellar rotation increases with age owing to stellar contraction until ∼100 Myr, then decreases with age owing to the increasing importance of angular momentum loss in a magnetic solar-type wind (e.g. Bouvier *et al* 1997b). The large breadth of the rotational distribution that is seen at all ages can then be interpreted largely as the result of a dispersion in disk dispersal times (e.g. Bouvier *et al* 1997a), i.e. a spread in initial stellar angular momenta rather than dispersion introduced through subsequent stellar evolution. (See Keppens *et al* (1995) for a somewhat different view.)

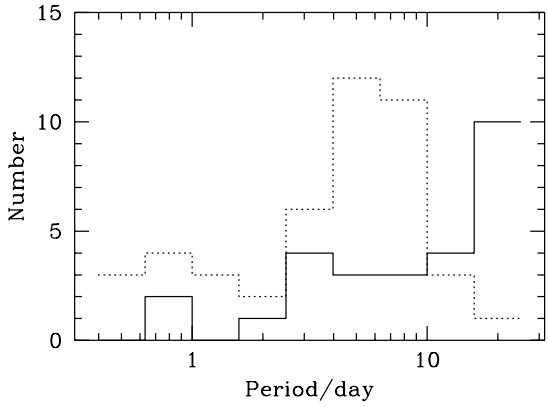


Figure 5. Distribution of rotation periods for stars in the young cluster NGC 2264 (Makidon *et al* 1999). Although stars with disks (solid line) generally rotate slower than stars without disks (dotted line), the distribution of both populations is quite broad.

One of the potential difficulties with this picture is that, in order to explain the population of slow rotators at ages of ∼50 Myr, disks must persist much beyond 10 Myr, and stars must couple effectively to them, in a fair fraction of systems. This seems rather unlikely given that disk dispersal times for T Tauri stars appear to be much shorter (see 'Circumstellar disks' above). Other questions for the future include issues such as whether the initial conditions imposed by molecular clouds (e.g. cloud angular momenta, infall accretion rate) and the effects of early dynamical evolution (e.g. development of close companions) are sufficiently homogenized during the epoch of disk coupling that observations of different clusters and associations can be thought to represent a purely time-dependent sequence.

As a final caveat, it is interesting to note that the study of rotational periods in other star forming regions (e.g. NGC 2264) may end up altering somewhat our picture of rotation rates at early times. Using synoptic observations that are sensitive to a broad range of rotation periods, Makidon *et al* (1999) find a broad period distribution with a significant population of both slow ($P_{\rm rot} \sim 30$ days) and fast ($P_{\rm rot} < 1$ days) rotators at ages of ∼3 Myr. While they confirm the trend that stars with disks rotate slower on average than stars without disks, the period distribution for both populations is quite broad (figure 5). Whether and how such observations fit into our understanding of the rotational evolution of stars is an interesting issue for the future.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar magnetic fields

As indicated in the previous sections, the processes of stellar accretion and angular momentum loss appear to be closely tied to the strength and geometry of stellar magnetic fields at early evolutionary times. Magnetic activity among the T Tauri stars has long been suspected based on a variety of observations including strong x-ray and non-thermal radio emission and photometric evidence for large starspots. One of the most exciting recent developments in this field is the detection of hard x-ray emission from the much younger protostellar population (Koyama *et al* 1996). In this case, hard x-ray emission was used to penetrate the large column densities toward protostars ($A_v \sim 100$) in order to probe the region close to the stellar surface. This result indicates that stellar magnetic activity begins very early in the evolution of low-mass stars.

Exciting progress has also been made in the direct measurement of stellar field strengths and geometries. Measurements of stellar magnetic field strengths in T Tauri stars are complicated by the rotational broadening of the photospheric lines (see 'Stellar rotation' above) which typically exceeds Zeeman broadening at optical wavelengths. For this reason, studies in the optical have ignored line profiles and instead used the enhancement in equivalent width that results from Zeeman broadening as a more indirect measure of stellar field strength. To date, field strengths ~1–2.5 kG have been measured for a few T Tauri stars using this approach (e.g. Basri *et al* 1992). The more direct detection of kilogauss strength fields through the study of Zeeman-broadened line profiles can be made by going to longer wavelengths where Zeeman broadening ($\propto\lambda^2$) dominates over rotational broadening ($\propto\lambda$). Using this approach, a clear signature of Zeeman broadening has been detected in the Zeeman-sensitive Ti I λ2.2 μm line in the spectrum of the classical T Tauri star BP Tau (Johns-Krull *et al* 1999b; figure 6). The average field strength implied by the line profile is 3.3 kG which is in excellent agreement with stellar field values required by theories of magnetic accretion (e.g. Königl 1991; Shu *et al* 1994; Cameron and Campbell 1996; see 'Disk accretion rates and stellar inflows' above).

Several other observations support the existence of an organized component to these strong fields. For example, our ability to measure photometric variability due to the rotational modulation of hot or cool spots (see 'Stellar rotation' above) indicates the existence and long-term stability of an organized magnetic field component. An exciting recent development in the study of the magnetic field geometry of young stars comes from spectropolarimetric studies of certain T Tauri star emission lines. In the classical T Tauri star BP Tau, the He I λ 5876 emission line is strongly polarized and the wavelength shift between the left- and right-circularly polarized components of the line profile implies a field strength ~2.4 kG (Johns-Krull *et al* 1999a; figure 7). Since the He I line is believed to originate at the base of the magnetic accretion flow onto the star, these observations provide clear evidence for a strong, organized field that is associated with accretion onto the stellar surface. Such strong, globally organized fields are necessary if the star is to couple magnetically to a disk several stellar radii away and participate in the processes of stellar accretion and angular momentum regulation.

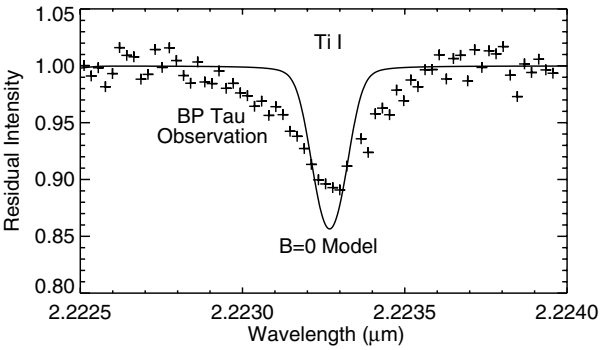


Figure 6. Detection of Zeeman broadening in the Ti I λ2.2 μm line in the spectrum of the classical T Tauri star BP Tau (Johns-Krull *et al* 1999). The Ti I line (plus signs) is found to be significantly broader than a Zeeman-insensitive photospheric line (solid curve). Modeling of the Ti I line indicates a distribution of field strengths on the stellar surface (1–10 kG) with a mean field of ~3.3 kG.

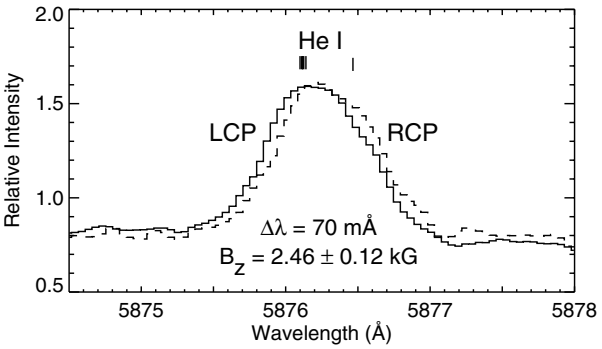


Figure 7. Detection of circular polarization in the He I λ 5876 emission line in the spectrum of the classical T Tauri star BP Tau (Johns-Krull *et al* 1999a). The wavelength shift between the right- and left-circularly polarized components indicates a field strength of 2.46 kG. Since the He I line is believed to form in the base of the magnetic accretion flow, these observations argue that organized fields are associated with accretion onto the stellar surface.

Origin of initial stellar masses and angular momenta

Remarkably, a large fraction of the observations described in the previous sections support and/or find a common explanation within our current understanding of how low-mass stars acquire their initial STELLAR MASSES and angular momenta. A possible synopsis of the story might be as follows. Mass and angular momentum are accreted from molecular clouds in a phase of dynamical infall.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(Following an extended observational search, compelling evidence for this phase of evolution has been obtained; see 'Infall' above.) The finite angular momentum of infalling cloud material leads to the formation of a circumstellar disk. (Following another observational search, compelling morphological and kinematic evidence for centrifugally supported structures has been obtained; see 'Circumstellar disks' above.) As a result of disk angular momentum transport, disk matter spirals in toward the star. The star accretes mass and angular momentum from the disk, producing a rapidly rotating star. Rapid rotation coupled with the fully convective STELLAR INTERIOR leads to the production of a strong stellar magnetic field. (We now have good evidence for strong, organized stellar fields and the early onset of magnetic activity; see 'Stellar magnetic fields' above.)

In the picture of Shu *et al* (1994; see also Shu *et al* 1999; Königl 1991; Arons 1986; Camenzind 1990; Shu *et al* 1988) that describes the interaction of the stellar magnetic field with the disk, the stellar field truncates and couples to the disk at a distance, typically a few stellar radii, that depends on disk accretion rates and stellar parameters including the magnetic field strength (figure 8). As a result of the magnetic disk coupling, the stellar rotation rate is slowed and regulated so that it equals the rotation rate at the disk truncation radius (cf Shu *et al* 1997). For typical T Tauri disk accretion rates, stellar parameters and the mean ($\sim$8 day) period of classical T Tauri stars, stellar field strengths of a few kilogauss are required, in good agreement with measured values.

In the Shu *et al* picture, disk accretion pinches the stellar magnetic field at the truncation radius, producing a magnetic geometry that favors dynamically the generation of magnetic accretion flows and winds from the truncation region. Thus, disk matter reaching the inner disk radius can be channeled into either an accretion flow onto the star or into an outflowing wind, providing an explanation for the observed simultaneous presence of stellar accretion flows and winds within a few stellar radii of the star (see 'Disk accretion rates and stellar inflows' above). The back torque of the wind and magnetic accretion flows on the magnetic field footpoints in the disk truncation region act to maintain the pinched magnetic field in the truncation region.

The magnetocentrifugal action resulting from the combination of strong stellar fields and the rapid rotation of the disk (or star) is capable of driving cool, massive winds with terminal velocities and mass loss rates that are in good agreement with the properties of winds from T Tauri stars (and embedded protostars), as described by Najita and Shu (1994). The magnetocentrifugally driven winds play an important role in the solution of the angular momentum problem for star formation. They remove significant angular momentum from the young star system, which allows the star to continue to rotate slowly despite ongoing accretion. For T Tauri stars, the observed kilogauss-strength fields are capable of removing sufficient angular momentum that the wind mass loss rate is a comparable fraction of the disk and stellar accretion rates ($\dot{M}_{\rm W}/\dot{M}_{\rm D} = 0.1$–$0.3$; $\dot{M}_* = \dot{M}_{\rm D} - \dot{M}_{\rm W}$) in reasonable agreement with observationally derived ratios (see 'Outflow' above).

At distances of tens of parsecs, the wind collimates magnetically into a jet embedded in a wider-angle wind (Shu *et al* 1995), in good agreement with HST observations of jet collimation distances. The wind and jet together sweep up the surrounding molecular cloud material, creating a molecular outflow (see 'Outflow' above). At the same time, the outflowing material acts to halt further infall, thereby limiting the mass of the star. At some later time, the disk disperses. The star and disk are consequently decoupled, and the star is free to spin up as it contracts. This picture, as it currently stands, provides a viable framework for understanding the origin and interrelation of many of the observed phenomena associated with young stars and their role in the origin of the initial masses and angular momenta of low-mass stars.

Unsolved problems

Despite the success of the above picture, there are numerous issues that remain to be addressed, including several basic aspects of this picture. For example, it is as yet unclear theoretically whether the magnetic interaction between the star and disk is likely to be time steady or not. Various recent simulations (Hayashi *et al* 1996; Goodson *et al* 1997; Miller and Stone 1997) suggest that the interaction is much less quiescent than in the Shu *et al* (1994) picture. Observational evidence for time-dependent behavior may be uncovered in synoptic observations of T Tauri line profiles which are notoriously variable.

Another important issue is the origin of angular momentum transport in disks. Gravitational instabilities are efficient when the circumstellar disk mass is a fair fraction of the stellar mass, a situation which may be relevant at early times. In the T Tauri phase, where disk masses appear to be too low to support significant gravitational instabilities, magnetic instabilities such as the Balbus–Hawley instability (Hawley *et al* 1995) appear to be the most promising.

Our understanding of the stellar rotational properties of young stars is still in its early stages, despite several conclusions having already been drawn from the existing data. From an observational perspective, observations of large numbers of stars will be needed to study rotation as a function of mass and age and to average over the differences that may be inherent in trying to compare stars with different formation histories (e.g. molecular cloud initial conditions, etc). Various groups are currently engaged in expanding the original work on young stars in the Taurus and Orion regions to a larger number of stars and clusters. These results are eagerly awaited.

As discussed above, the longevity of inner accretion disks may be an important issue for the understanding of the origin of slow rotators and the spread in stellar angular momenta that exists in 10–100 Myr old clusters. The factors

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

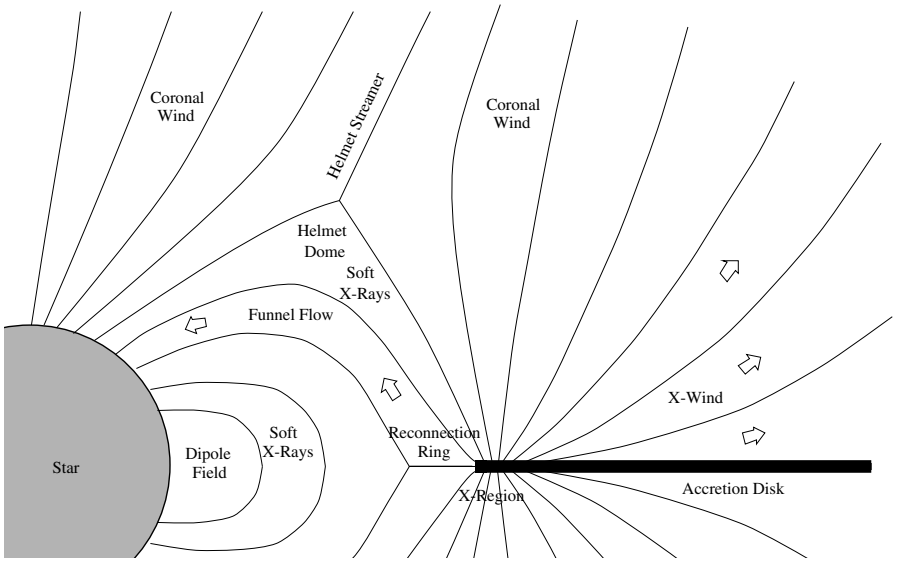


Figure 8. Schematic drawing of the X-wind model (Shu *et al* 1994, 1999). In the interaction between a stellar dipole magnetic field and a surrounding accretion disk, the disk is truncated and the magnetic configuration of the star (thin lines) is altered, allowing mass inflow onto the stellar surface in magnetic accretion flows, and energetic mass loss in a magnetocentrifugal wind.

that control disk dispersal are at present unclear, although some of the relevant processes have been identified. For example, since disk structure is known to be affected by the presence of stellar companions (e.g. Mathieu 1994), the longevity of inner disks may similarly be affected by the presence of close stellar or planetary companions.

Another generic issue is the extent to which this picture, developed to explain the formation of isolated, low-mass stars, applies to stars in clusters and/or to stars of higher masses. It may be necessary to consider other physical processes (e.g. fragmentation, mergers of stars and/or clouds) in order to understand the origin of stellar masses and angular momenta in dense clusters. Further work is needed to understand these processes.

On the issue of the applicability of this picture to high-mass stars, an issue that also remains unresolved, there is some observational evidence that suggests that similar physical processes may be involved. High-resolution spectroscopy of the CO overtone emission from a high-mass star in the young cluster M17 has revealed the spectral shape of emission from a rotating disk, providing kinematic evidence for disks around young high mass stars (Najita *et al* 1999; figure 3). This indicates that stars may evolve from low to high mass by accretion through disks (e.g. Palla and Stahler 1993). These observations also support the view that radiation pressure is not a barrier to the formation of high-mass stars and that consequently the high end of the stellar initial mass function should be independent of metallicity, as is observed (Massey *et al* 1995).

Acknowledgments

I am very grateful to Neal Evans, Lynne Hillenbrand, Michael Meyer, David Wilner, Steve Strom, Russ Makidon, Jeff Valenti, Chris Johns-Krull and Frank Shu for providing illustrations for this review.

Bibliography

Interested readers in search of additional information and opinions on both observational and theoretical aspects of the subject may wish to consult the comprehensive reviews presented in *Protostars and Planets IV* (Mannings *et al* 2000). The review by Bodenheimer (1995) on the more specific topic of the angular momentum evolution of young stars is also a valuable reference.

Adams F C, Lada C J and Shu F H 1987 Spectral evolution of young stellar objects *Astrophys. J.* **312** 788–806

Armitage P J and Clarke C J 1996 Magnetic braking of T Tauri stars *Mon. Not. R. Astron. Soc.* **280** 458–68

Arons J 1986 *Plasma Penetration into Magnetospheres* ed N Kylafis, J Papamastorakis and J Ventura (Iraklion: Crete University Press) p 115

Bachiller R 1996 Bipolar molecular outflows from young stars and protostars *Ann. Rev. Astron. Astrophys.* **34** 111–54

Basri G, March G W and Valenti J A 1992 Limits on the magnetic flux of pre-main sequence stars *Astrophys. J.* **390** 622–33

Bertout C, Basri G and Bouvier J 1988 Accretion disks around T Tauri stars *Astrophys. J.* **330** 350–73

Bertout C, Harder S, Malbet F and Mennessier C 1996 Photometric observations of YY Orionis: new insight into the accretion process *Astron. J.* **112** 2159–67

Bodenheimer P 1995 Angular momentum evolution of young stars and disks *Ann. Rev. Astron. Astrophys.* **33** 199–38

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bouvier J, Cabrit S, Fernandez M, Martin E L and Matthews J M 1993 Coyotes-I—the photometric variability and rotational evolution of T-Tauri stars *Astron. Astrophys.* **272** 176–206

Bouvier J *et al* 1997a COYOTES IV: the rotational periods of low-mass post-T Tauri stars in Taurus *Astron. Astrophys.* **318** 495–505

Bouvier J, Forestini M and Allain S 1997b The angular momentum evolution of low-mass stars *Astron. Astrophys.* **326** 1023

Burrows C J *et al* 1996 Hubble Space Telescope observations of the disk and jet of HH 30 *Astrophys. J.* **473** 437–51

Calvet N and Hartmann L 1992 Balmer line profiles for infalling T Tauri envelopes *Astrophys. J.* **386** 239

Calvet N 1997 Properties of the winds of T Tauri stars *Herbig–Haro Flows and the Birth of Low Mass Stars (IAU Symp. 182)* ed B Reipurth and C Bertout (Dordrecht: Reidel) pp 417–32

Calvet N, Hartmann L and Strom S 2000 Evolution of disk accretion *Protostars and Planets IV* ed V Mannings, A P Boss and S Russell (Tuscon, AZ: University of Arizona Press) pp 337–99

Camenzind M 1990 *Reviews in Modern Astronomy 3* ed G Klare (Berlin: Springer) p 259

Cameron A C and Campbell C G 1996 Rotational evolution of magnetic T Tauri stars with accretion disks *Astron. Astrophys.* **274** 309–18

Carr J S, Tokunaga A T, Najita J, Shu F H and Glassgold A E 1993 The inner-disk and stellar properties of the young stellar object WL16 *Astrophys. J.* **411** L37–40

Carr J S *et al* 2000 in preparation

Chandler C J, Carlstrom J E, Scoville N Z, Dent W R F and Geballe T R 1993 Infrared CO emission from young stars: high resolution spectroscopy *Astrophys. J.* **412** L71–4

Choi M, Evans N J II, Gergersen E M and Wang Y 1995 Modeling line profiles of protostellar collapse in B335 with the Monte Carlo method *Astrophys. J.* **448** 742–7

Choi P I and Herbst W 1996 Rotation periods of stars in the Orion Nebula cluster: the bimodal distribution *Astron. J.* **111** 283–98

D'Alessio P, Cantó J, Calvet N and Lizano S 1998 Accretion disks around young objects. I. The detailed vertical structure *Astrophys. J.* **500** 411–27

Davis C J, Mundt R, Eislöffel J and Ray T P 1995 Near-infrared and optical imaging of the L1551-IRS 5 region—the importance of poorly collimated outflows from young stars *Astron. J.* **110** 766–75

Edwards S, Hartigan P, Ghandour L and Andrulis C 1994 Spectroscopic evidence for magnetospheric accretion in classical T Tauri stars *Astron. J.* **104** 1056–70

Edwards S, Ray T and Mundt R 1993 Energetic mass outflows from young stars *Protostars and Planets III* ed E H Levy and J I Lunine (Tucson, AZ: University of Arizona Press) pp 567–602

Giovanardi C, Gennari S, Natta A and Stanga R 1991 Infrared spectroscopy of T Tauri stars *Astrophys. J.* **367** 173–81

Goodman A A, Benson P J, Fuller G A and Myers P C 1993 Dense cores in dark clouds. VIII—velocity gradients *Astrophys. J.* **406** 528–47

Goodson A P, Winglee R M and Boehm K-H 1997 Time-dependent accretion by magnetic young stellar objects as a launching mechanism for stellar jets *Astrophys. J.* **489** 199

Gullbring E, Hartmann L, Briceño C and Calvet N 1998 Disk accretion rates for T Tauri stars *Astrophys. J.* **492** 323

Hartigan P, Edwards S and Ghandour L 1995 Disk accretion and mass loss from young stars *Astrophys. J.* **452** 736–68

Hartmann L and Stauffer J R 1989 Additional measurements of pre-main sequence stellar rotation *Astron. J.* **97** 873–80

Hawley J F, Gammie C F and Balbus S A 1995 Local three-dimensional magnetohydrodynamic simulations of accretion disks *Astrophys. J.* **440** 742–63

Hayashi M, Ohashi N and Miyama S M 1993 A dynamically accreting gas disk around HL Tauri *Astrophys. J.* **418** L71–4

Hayashi M R, Shibata K and Matsumoto R 1996 *Astrophys. J.* **468** L37–40

Hillenbrand L and Meyer M R 2000 in preparation

Johns C and Basri G 1995 The line profile variability of SU Aurigae *Astrophys. J.* **449** 341–64

Johns-Krull C M, Valenti J A, Hatzes A P and Kanaan A 1999a Spectropolarimetry of magnetospheric accretion on the classical T Tauri star BP Tauri *Astrophys. J.* **510** L41–4

Johns-Krull C M, Valenti J A and Koresko C 1999b Measuring the magnetic field on the classical T Tauri star BP Tauri *Astrophys. J.* **516** 900–15

Kenyon S J and Hartmann L 1995 Pre-main-sequence evolution in the Taurus–Auriga molecular cloud *Astrophys. J. Suppl. Ser.* **101** 117–71

Kepner J, Hartigan P, Yang C and Strom S 1993 Hubble Space Telescope images of the subarcsecond jet in DG Tauri *Astrophys. J.* **415** L119–21

Keppens R, MacGregor K B and Charbonneau P 1995 On the evolution of rotational velocity distributions for solar-type stars *Astron. Astrophys.* **294** 469–87

Koerner D W and Sargent A I 1995 Imaging the small-scale circumstellar gas around T Tauri stars *Astrophys. J.* **109** 2138–45

Königl A 1991 Disk accretion onto magnetic T Tauri stars *Astrophys. J.* **370** L39–43

Koyama K, Hamaguchi K, Ueno S, Kobayashi N and Feigelson E 1996 Discovery of hard X-rays from a cluster of protostars *Publ. Astron. Soc. Pac.* **48** L87–92

Lada C J 1985 Cold outflows, energetic winds, and enigmatic jets around young stellar objects *Ann. Rev. Astron. Astrophys.* **23** 267–317

Levreault R M 1988 Molecular outflows and mass loss in pre-main-sequence stars *Astrophys. J.* **330** 897–910

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Lynden-Bell D and Pringle J E 1974 The evolution of viscous discs and the origin of the nebular variables *Mon. Not. R. Astron. Soc.* **168** 603–38

Makidon R *et al* 2000 in preparation

Mannings V, Boss A P and Russell S (eds) 2000 *Protostars and Planets IV* (Tuscon, AZ: University of Arizona Press)

Massey P, Johnson K E and DeGioia-Eastwood K 1995 The initial mass function and massive star evolution in the OB associations of the northern Milky Way *Astrophys. J.* **454** 151–71

Mathieu R D 1994 Pre-main-sequence binary stars *Ann. Rev. Astron. Astrophys.* **32** 465–530

McCaughrean M J and O'Dell C R 1996 Direct imaging of circumstellar disks in the Orion nebula *Astron. J.* **111** 1977–86

Meyer M R, Beckwith S V W, Stauffer J R and Schultz B 2000 in preparation

Miller K A and Stone J M 1997 Magnetohydrodynamic simulations of stellar magnetosphere–accretion disk interaction *Astrophys. J.* **489** 890–902

Moriarty-Schieven G H, Snell R L, Strom S E, Schloerb F P, Strom K M and Grasdalen G L 1987 High-resolution images of the L1551 bipolar outflow: evidence for an expanding, accelerated shell *Astrophys. J.* **319** 742–53

Mundt R 1984 Mass loss in T Tauri stars: observational studies of the cool parts of their stellar winds and expanding shells *Astrophys. J.* **280** 749–70

Mundt R, Stocke J, Strom S E, Strom K E and Anderson E R 1985 The optical spectrum of L1551 IRS 5 *Astrophys. J.* **297** L41–5

Mundy L G *et al* 1996 Imaging the HL Tauri disk at $\lambda = 2.7$ millimeters with the BIMA array *Astrophys. J.* **464** L169–73

Muzerolle J, Calvet N and Hartmann L 1998a Magnetospheric accretion models for the hydrogen emission lines of T Tauri stars *Astrophys. J.* **492** 743–53

Muzerolle J, Hartmann L and Calvet N 1998b Emission-line diagnostics of T Tauri magnetospheric accretion. I. Line profile observations *Astron. J.* **116** 455–68

Myers P C, Bachiller R, Caselli P, Fuller G A, Mardones D, Tafalla M and Wilner D J 1995 Gravitational infall in the dense cores L1527 and L483 *Astrophys. J.* **449** L65–8

Najita J and Shu F H 1994 Magnetocentrifugally driven flows from young stars and disks. III. Numerical solution of the sub-Alfvénic region *Astrophys. J.* **429** 808–25

Najita J, Carr J S, Glassgold A E, Shu F H and Tokunaga A T 1996a Kinematic diagnostics of disks around young stars: CO overtone emission from WL16 and 1548C27 *Astrophys. J.* **462** 919–36

Najita J, Carr J S and Tokunaga A T 1996b High-resolution spectroscopy of Brγ emission in young stellar objects *Astrophys. J.* **456** 292–9

Najita J, Greene T P and Lada C J 2000 High resolution spectroscopy of CO overtone emission from massive young stars, in preparation

Natta A 1995 The effect of the environment on the temperature profile of circumstellar disks *Rev. Mex. Astron. Astrofis.* **1** 209–15

Palla F and Stahler S 1993 The pre-main-sequence evolution of intermediate-mass stars *Astrophys. J.* **418** 414–25

Reipurth B, Bally J and Devine D 1997 Giant Herbig-Haro flows *Astron. J.* **114** 2708–35

Shu F H 1977 Self-similar collapse of isothermal spheres and star formation *Astrophys. J.* **214** 488–97

Shu F H, Lizano S, Ruden S and Najita J 1988 Mass loss from rapidly rotating magnetic protostars *Astrophys. J.* **328** L19

Shu F H, Najita J, Ostriker E C and Shang H 1995 Magnetocentrifugally driven flows from young stars and disks. V. Asymptotic collimation into jets *Astrophys. J.* **455** L155–8

Shu F H, Najita J, Ostriker E, Wilkin F, Ruden S and Lizano S 1994 Magnetocentrifugally driven flows from young stars and disks. I. A generalized model *Astrophys. J.* **429** 781–96

Shu F H, Najita J, Shang H and Li Z-Y 2000 X-winds: theory and observations *Protostars and Planets IV* ed V Mannings, A P Boss and S Russell (Tuscon, AZ: University of Arizona Press) pp 789–813

Shu F H, Shang H, Glassgold A E and Lee T 1997 X-rays and fluctuating X-winds from protostars *Science* **277** 1475

Snell R L and Schloerb F P 1985 Structure and physical properties of the bipolar outflow in L1551 *Astrophys. J.* **295** 490–500

Stanke T, McCaughrean M J and Zinnecker H 1998 First results of an unbiased H_2 survey for protostellar jets in Orion A *Astron. Astrophys.* **332** 307–13

Tafalla M, Mardones D, Myers P C, Caselli P, Bachiller R and Benson P J 1998 L1544: a starless dense core with extended inward motions *Astrophys. J.* **504** 900–14

Terebey S, Shu F H and Cassen P 1984 The collapse of the cores of slowly rotating isothermal clouds *Astrophys. J.* **286** 529

Vogel S N and Kuhi L V 1981 Rotational velocities of pre-main-sequence stars *Astrophys. J.* **245** 960–76

Walker M 1972 Studies of extremely young clusters: spectroscopic observations of the ultraviolet-excess stars in the Orion nebula cluster and NGC 2264 *Astrophys. J.* **175** 89–116

Wilner D J, Ho P T P and Rodriguez L F 2000 7 millimeter imaging of HL Tauri at 5 AU resolution *Astrophys. J.* in preparation

Zhou S, Evans N J II, Kömpe C and Walmsley C M 1993 Evidence for protostellar collapse in B335 *Astrophys. J.* **404** 232–46

Zhou S, Evans N J II and Wang Y 1996 Small-scale structure of candidates for protostellar collapse. I. BIMA observations of L1527 and CB54 *Astrophys. J.* **466** 296–308

Joan Najita

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star Formation and Radio Emission from Normal Galaxies

The radio continuum properties of normal DISK GALAXIES are shaped primarily by STAR FORMATION. This article reviews the relationship between the radio continuum emission and the star formation process in normal SPIRAL and IRREGULAR GALAXIES. The most common tracers of global star formation in external galaxies are optical emission lines and far-infrared (FIR) continuum emission. The former arises from photoionized gas while the latter is from dust grains heated by stellar photons. One component of radio continuum emission is thermal bremsstrahlung emitted by the same photoionized plasma that produces the optical emission lines. If that were the only source of radio emission one would expect a strong correlation between, say, Hα and the radio continuum. However, except for the highest frequencies the radio continuum emission is dominated by SYNCHROTRON RADIATION emitted by relativistic electrons gyrating in magnetic fields so that any correlation between thermal radio emission and Hα is easily obscured. Surprisingly, the strongest correlation is that between the FIR and the nonthermal radio continuum. The most likely explanation for the correlation is that star formation drives both forms of emission. Within this scenario lie two fundamental mysteries: (i) why is the FIR–radio correlation one of the strongest in astrophysics? (ii) what is the connection between star formation and the production of cosmic rays? This article focuses on the physical underpinnings of these problems. In the process, the relationship between star formation and radio emission is fully explored.

Historical overview

By the end of the 19th century it was well known that there existed a background level of highly penetrating radiation. The issue of whether this radiation emanated from the Earth or from space was a matter of considerable debate and was not settled until a daring experiment by the Austrian physicist Victor Hess in 1912. Hess piloted a balloon to a height of 5 km and recorded the radiation flux as a function of altitude. His measurements showed a steady increase in flux with altitude, proving their 'cosmic' origin. The term 'COSMIC RAYS' was born because at the time it was believed that cosmic rays were photons similar to γ-rays but with greater penetrating power. Their particle nature was uncovered by Bothe and Kolhorster in 1929 using coincidence techniques. By this time, however, the term 'cosmic ray' had stuck and remains in use today.

Even after the recognition that cosmic rays were particles of an extraterrestrial origin, their distribution in space remained unknown. The cosmic rays appeared to arrive from all directions with equal likelihood. Their isotropic distribution shed little light on their origin indicating only that the cosmic ray particles are heavily scattered before reaching the Earth. Though RADIO ASTRONOMY was born in 1935 through the work of Karl Jansky, no connection with cosmic rays was made until it was realized, many years later, that Galactic radio emission is, in fact, synchrotron radiation emitted by cosmic ray electrons.

In the 1940s Grote Reber mapped the distribution of radio emission and showed that it closely followed the Milky Way plane (e.g. Reber 1940). This led to the suggestion that the radio emission was associated with stars. This was probably the first step in connecting star formation to radio emission. However, it was quickly realized that the thermal emission of normal stars would fall well short of the observed level of radio emission. The radio emission was shown to have a nonthermal spectrum. A special class of 'radio stars' was therefore proposed. In 1950, Alfvén and Herlofson proposed that relativistic particles in the atmospheres of these radio stars produced radio emission via the synchrotron mechanism. Kiepenheuer (1950) generalized the idea to broadly distributed cosmic ray electrons and magnetic fields in the interstellar medium (ISM). As it turned out Kiepenheuer's idea was correct (although it took another 10 yr to be accepted) but Alfvén and Herlofson's origin idea proved not to be viable.

The connection between cosmic rays and Galactic radio emission was fully cemented in the review paper by Ginzburg and Syrovatskii in 1965. Thus, some 50 yr had passed since the discovery of cosmic rays and some 30 yr since the advent of radio astronomy before the two were linked and it was realized that cosmic rays fill the Milky Way Galaxy.

The connection between radio emission and star formation has not been an easy one to establish. Fritz Zwicky was the first to propose that SUPERNOVAE provide the needed link (Zwicky 1939). The currently favored view is close to Zwicky's original insight.

The most favored hypothesis of the origin of cosmic rays combines SUPERNOVA REMNANTS (SNRs) with the theory of diffusive shock acceleration (DSA) (e.g. Bell 1978; Axford 1981; Drury 1983; Blandford and Eichler 1987; Jones *et al* 1998). A supernova explosion is believed to typically release in excess of 10^{51} erg of energy much of which is carried by the subsequent SNR in the form of kinetic energy. The rapid expansion of the SNR causes the SNR to couple with the ISM through a supersonic shock. According to the DSA theory, particles can be accelerated at the shock front to cosmic ray energies. It is this coupling which allows the SNR to convert some of its kinetic energy into relativistic particles.

It is now well established that all disk galaxies have relativistic cosmic ray gas as a major constituent and that this gas plays an important role in the chemical and kinematical evolution of individual galaxies. In our Galaxy, the relativistic gas contributes as much energy density as the interstellar magnetic field and the thermal gas (table 1). This equipartition of energies is probably a feature of most disk galaxies (e.g. Duric 1988). In table 1, the various phases of the ISM are shown (including the cosmic rays), along with recent estimates of their filling

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Components of the ISM.

Phase	f	n (cm^{-3})	kT (eV)	nkT (eV cm^{-3})
Molecular clouds	10^{-3}	>100	$<10^{-2}$	–
Cold neutral medium	0.025	40	$\approx 10^{-2}$	≈ 0.4
Warm neutral medium	≈ 0.5	≈ 0.5	≈ 1	≈ 0.5
Warm ionized medium	≈ 0.25	≈ 0.2	≈ 1	≈ 0.2
Hot ionized medium	≈ 0.2	$\approx 3 \times 10^{-3}$	$\approx 10^2$	≈ 0.3
Cosmic rays	≈ 1	$\approx 10^{-9}$	$\approx 10^9$	≈ 1

factor, f, the number density, n, the thermal energy per particle and the energy density.

Although the relativistic gas contains a mix of nuclei, electrons and other more exotic particles, it is the electrons that produce all the observable synchrotron radio emission, the same emission seen by Jansky and Reber in the case of our Galaxy. This radiation mechanism dominates the radio continuum spectra of spiral and irregular galaxies for wavelengths longer than about 1 cm.

Star formation plays a much broader role in shaping the integrated properties of galaxies. An appreciation of this broader picture is necessary in order to understand fundamentally important observations such as the radio–infrared correlation of galaxies and to define the tools needed to study the details of the relationship between star formation and radio emission of galaxies.

The basic picture

It is now generally accepted (although not proven) that star formation drives the radio–FIR correlation and that it also accounts for a (somewhat weaker) correlation between Hα and radio emission. A schematic view of this scenario is presented in figure 1. The essential points are that (i) photons from the young massive stars photoionize surrounding gas which gives rise to emission line radiation and thermal bremsstrahlung, (ii) massive stars heat dust grains which in turn radiate in the FIR and (iii) the same stars quickly become supernovae, producing remnants that accelerate the cosmic ray electrons which in turn give rise to nonthermal radio emission.

According to figure 1, of all the radiative processes shown, synchrotron emission is the most removed from the young stars (it is at the endpoint of the longest of the three arms in the schematic). In order for the FIR and radio emission to be tightly correlated, for example, it is necessary for the SNRs to play a direct and dominant role in converting stellar energy into relativistic particles.

We now describe the radiative processes along each of the three arms shown in the schematic. We confine our discussion to the radiation mechanisms associated with recently formed stars as opposed to the processes that operate during pre-main-sequence evolution since it is the young massive stars that appear to drive the radio properties of galaxies.

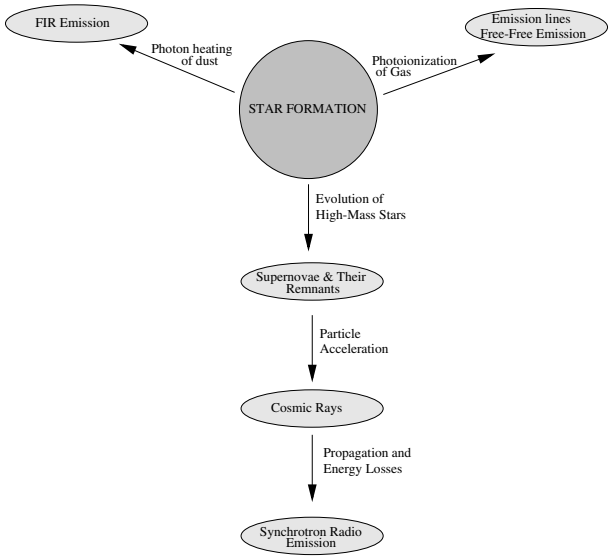


Figure 1. The currently favored scenario for explaining the connection between star formation and the radio emission. The relevant radiation mechanisms are noted.

Photoionization

Stromgren (1939) first proposed that hot, luminous stars should ionize the gas around them and form what we now call H II regions. The more luminous, massive stars evolve very quickly such that their evolutionary time scales can be shorter than the time scale of star formation for a given star forming region. The presence of such luminous stars and their associated H II REGIONS therefore marks regions of recent and ongoing star formation. The detection of H II regions is therefore synonymous with tracing star forming regions.

For an H II region in ionization balance, the ionization rate is balanced by the recombination rate so that (Shu 1991)

$$\dot{N}_{\rm i} = 3 \times 10^{-13} n_{\rm e}^2 V \ {\rm s}^{-1}$$

where $\dot{N}_{\rm i}$ is the ionization rate in photons per second, $n_{\rm e}$ is the density of the ionized gas and V is the volume of the H II region. It is this relationship that establishes the size of an H II radius (via V) for a given ionizing rate.

Free–free radiation. The ions and electrons in the ionized gas undergo Coulomb interactions which lead to thermal

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

bremsstrahlung or free–free radiation (sometimes referred to as thermal radio emission). At radio frequencies, the free–free radiated power from an H II region is given by (Tucker 1975)

$$L_\nu^{\rm ff} = 7.3 \times 10^{-39}\, n_{\rm e}^2 V \frac{\nu^{-0.1}}{5\,{\rm GHz}}$$

for an assumed temperature of 10^4 K. Note the relatively flat spectrum compared with the steeper spectrum associated with synchrotron radiation.

The relationship between the 6 cm radio power and the ionizing luminosity is given by combining the above equations, yielding

$$L_\nu^{\rm ff} = 2.4 \times 10^{-26} \dot{N}_{\rm i}\ {\rm erg\ s^{-1}\ Hz^{-1}} \frac{\nu^{-0.1}}{5\,{\rm GHz}}.$$

Emission line radiation. A similar relationship exists between the Balmer line fluxes and the ionization flux. In the case of the Hα line luminosity,

$$L_{{\rm H}\alpha} = 1.3 \times 10^{-12} \dot{N}_{\rm i}\ {\rm erg\ s^{-1}}.$$

Imaging of galaxies in the light of Hα has become a common method for tracing global star formation in external galaxies. It should be noted that all kinds of emission lines are produced by H II regions. The most common types of lines are recombination lines (such as the Balmer series, including Hα) and collisionally excited lines from atoms having low-lying energy states. These are generally forbidden lines such as the 4995, 5007 Å doublet of O III. However, in most instances, the Hα is the strongest and therefore easiest to detect.

The properties of an H II region are determined by the nature of the embedded star and the density and composition of the surrounding medium. Table 2 summarizes some basic properties of H II regions ionized by various stars. The first column indicates the spectral type of the main sequence star responsible for ionizing the H II region. The corresponding ionizing, free–free and Hα luminosities are shown in the other three columns.

Table 2. Radiative properties of H II regions.

Spectral type	$\log \dot{N}_{\rm i}$	$L_\nu^{\rm ff}$(erg s^{-1} Hz^{-1})[a]	$L_{{\rm H}\alpha}$(erg s^{-1})
O5	50.0	2.4×10^{24}	1.3×10^{38}
B0	48.7	1.2×10^{23}	6.5×10^{36}
A0	42.7	1.2×10^{17}	6.5×10^{30}

[a] Evaluated at 5 GHz.

Hα and free–free emissions can both be used to trace star forming regions in galaxies. The former has the advantage that it can be readily detected with great efficiency to very faint levels. However, unlike the radio emission the Hα emission is attenuated by dust. The thermal free–free radio emission therefore represents a more direct measurement of the ionizing flux. The drawback to the radio method is that the emission is quite weak, thereby limiting the number of H II regions that can be detected relative to the Hα method. For example, an H II region from a B0 star would have a flux density of only about 0.1 mJy, barely detectable at the 3σ level with the VLA. Hα emission, on the other hand, can be detected much more easily from the same H II region. The use of Hα to study star formation is therefore quite common. Studies of free–free emission are now possible but only for the nearest galaxies, using the VLA (e.g. Devereux *et al* 1997). However, neither approach is completely satisfactory.

FIR emission

The dust grains that fill the disks of normal galaxies are heated primarily by stellar photons. A good fraction of these photons are nonionizing photons and leaked ionizing photons from recently formed stars. The exact fraction that can be attributed to young, massive stars is the subject of considerable debate. However, it is generally agreed that star formation generates a significant enough portion of the dust heating that the resulting FIR emission can be used as an indicator of global star formation.

The stellar radiation heats dust grains to temperatures of 15–30 K. At these temperatures the dust re-radiates emission in the FIR, mainly between about 25 and 200 μm. The dust acts as a bolometer, accurately measuring the energy it absorbs.

The FIR luminosity between 40 and 1000 μm can be estimated from measurements at the standard IRAS bands of 60 and 100 μm using the formula

$$L_{\rm FIR} = 6.0 \times 10^5 \left(2.58 \frac{S_{60}}{\rm Jy} + \frac{S_{100}}{\rm Jy}\right) \left(\frac{d}{\rm Mpc}\right)^2 L_\odot$$

which typically represents over 80% of the bolometric IR flux (2–1000 μm). A major advantage of FIR emission over Hα in tracing star formation is its relative immunity to extinction and that it measures much of the bolometric luminosity emanating from the star forming regions. A major drawback is the limited sensitivity to surface brightness and relatively poor angular resolution offered by FIR instruments. However, this field promises to leap forward with the launching of SIRTF, a new high-resolution, high-sensitivity IR telescope.

Synchrotron emission

According to the model shown in figure 1, SNRs accelerate electrons to relativistic energies. A single relativistic electron emits radiation whose intensity and spectral properties are determined by the kinetic energy of the electron and the strength of the magnetic field in which it is gyrating. The shape of the spectrum, $J(\nu, E)$, is given by

$$J(\nu, E)\,{\rm d}\nu = \frac{\sqrt{3}e^3}{mc^2} B \sin\theta \frac{\nu}{\nu_{\rm c}}\,{\rm d}\nu \int_{\nu/\nu_{\rm c}}^{\infty} K_{5/3}(\eta)\,{\rm d}\eta \qquad (1)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where ν is the frequency of the emitted radiation, E is the kinetic energy of the electron, B is the strength of the magnetic field, θ is the angle between the instantaneous velocity vector of the electron and the magnetic field, m is the mass of the electron and $K_{5/3}$ is a modified Bessel function. The quantity ν_c is commonly referred to as the critical frequency and is given by

$$\nu_c = \frac{3e}{4\pi mc} B \sin\theta \frac{E^2}{m^2c^4} \tag{2}$$

where the dependence on the electron energy, E, is now explicit. The critical frequency is closely related to the frequency of the peak of the spectrum, ν_{max}, such that $\nu_{max} = 0.3\nu_c$. In typical interstellar magnetic fields, relativistic electrons radiate in the radiofrequency portion of the electromagnetic spectrum.

Integration of equation (1) over all frequencies (e.g. Shu 1991) yields the total power radiated by a highly relativistic electron,

$$\langle P \rangle = \frac{1}{6\pi} \frac{E^2}{m^2c^4} c\sigma_T B^2 \tag{3}$$

where $\langle P \rangle$ indicates a statistical average expected from an isotropic distribution of θ and $\sigma_T = 8\pi e^4/3m^2c^4$ is the Thomson cross section.

It is interesting to note the mass dependence in equations (2) and (3). The critical frequency and the radiated power are both very strong functions of the mass. It is easy to see that only electrons contribute significantly to the synchrotron emission in a typical galactic environment. Ions are simply too heavy to contribute significantly to the radio emission of a galaxy.

The synchrotron radiation we see from a galaxy arises from an ensemble of electrons following some energy distribution $n(E)$. With equation (3) we can define a volume emissivity

$$\rho j_\nu = \int_{mc^2}^{\infty} \langle P_\nu(E) \rangle n(E)\, \mathrm{d}E. \tag{4}$$

Direct observations of cosmic rays near the Earth suggest that

$$N(E)\, \mathrm{d}E = N_0 E^{-p}\, \mathrm{d}E \tag{5}$$

with $p \approx 2.5$. Now substituting equations (3) and (5) into equation (4) and integrating over frequency and volume (assumed homogeneous) yields (Pacholczyck 1970) the luminosity of an ensemble of electrons:

$$L_\nu = \int_V \rho j_\nu\, \mathrm{d}V \propto N_0 B^{(p+1)/2} \nu^{-(p-1)/2}.$$

where N_0 is the number of cosmic ray electrons in the source volume. It is clear from the above that the spectrum is a power-law spectrum with an index

$$\alpha = \frac{p-1}{2}$$

so that

$$L_\nu \propto N_0 B^{1+\alpha} \nu^{-\alpha}.$$

Observations of normal galaxies indicate that α is universally around 0.75. These numbers indicate that the electrons follow an energy distribution that is also a power law, consistent with the observed energy distribution of cosmic rays near the Earth. The universality of the electron energy distribution among normal galaxies suggests that cosmic rays may be accelerated by a common mechanism. In other words, whatever is found to work for one disk galaxy may apply to all galaxies, including our own.

It is interesting to note that nothing in the above equations directly relates to the young stellar content of a galaxy.

The link between star formation and radio emission

It is clear, from the above discussion, that the FIR, emission line radiation and free–free emission can be used to directly trace the recently formed stars in galaxies. Similarly, nonthermal radio emission (synchrotron radiation) can be used to trace the cosmic ray electrons. Consequently, the link between star formation and the production of cosmic rays (the bottom arm in figure 1) is best studied by comparing FIR, emission line and radio data of galaxies.

Figure 2 shows a 60 μm and a 6 cm image of M33 smoothed to the same resolution. Note that the radio–FIR correlation appears to hold even inside galaxies. It is clear that some form of a relationship must exist between star formation, traced by the FIR emission, and the relativistic electrons, traced by the radio emission.

As noted above, the link between star formation and radio emission is more complicated than that for the other radiation mechanisms. The chain of events starting with star formation and ending with synchrotron emission is much longer than that required to explain FIR, emission line and free–free emission. The critical link in this chain is the role of the SNRs.

If the SNRs are to be the primary source of cosmic rays they must provide sufficient energy to maintain the cosmic ray population of a galaxy. Thus, there must be a sufficient number of supernovae to be able to continuously replenish the cosmic rays over the duty cycle of the medium. Consider, for example, a cosmic ray medium of a disk galaxy that has an energy (kinetic energy of all the relativistic particles) of 10^{55} erg and which escapes from the disk over a period of 10^7 yr. The energy needed to maintain the relativistic medium is 10^{48} erg yr^{-1}. This requires the energy of 10^4 SN over the duty cycle of the relativistic medium or 1 SN per 10^3 yr. However, it is unlikely that an SN converts all of its energy into relativistic particles. The conversion efficiencies are not known. A conversion efficiency of 1%, for example, would require an SN rate of 1 every 10 yr. Thus, the ability of SNRs to maintain the cosmic ray population depends on the efficiency of cosmic ray production.

Once free of the SNR the newly accelerated electrons and ions diffuse through the ISM, establishing a general

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

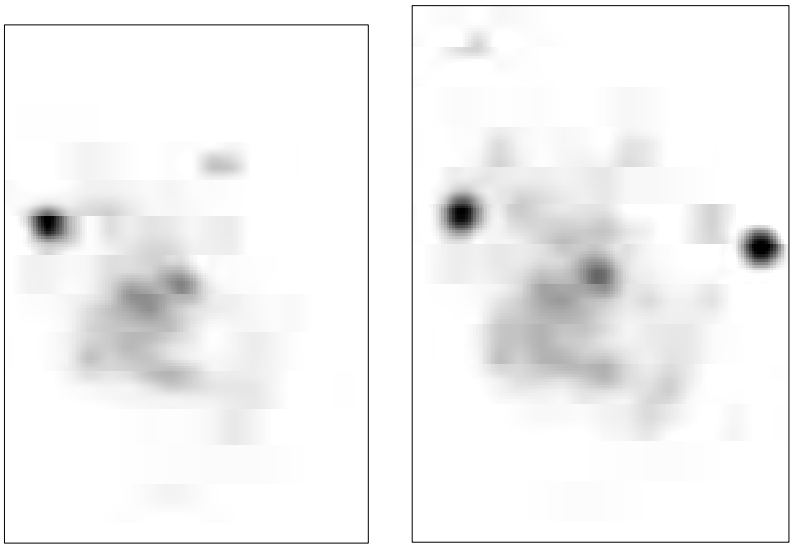


Figure 2. Left: a 60 μm image of M33. Right: a 6 cm image of M33. The left image traces emission from dust heated by young stars. The right image traces relativistic electrons emitting synchrotron radiation. The well-known radio–FIR correlation of galaxies appears to extend to spatially resolved regions of galaxies, thereby strongly strengthening the argument that star formation plays a key role in shaping the radio emission of galaxies. Except for the radio source to the right, the condensations and spirality of the galaxy show similarly in both pictures.

disk distribution of cosmic rays. This relativistic gas forms what is essentially a phase of the ISM. The relativistic phase of the ISM can remain in a steady state configuration because a balance between cosmic ray production and cosmic ray losses (via energy losses and escape) is established when the production and loss rates are equal. Such a balance is exacted when the supernova rate is constant over time scales long compared with the loss time scale.

The maintenance of the cosmic ray gas is accomplished via a continuous injection by a continuously evolving population of SNRs. Consider a simple model in which SNRs are the sole injectors of relativistic particles in the disk of a galaxy. Particles are injected on the time scale of the SNR lifetime, $\tau_{\rm snr}$. Estimates of such a time scale range from 10^4 to 5×10^4 yr (Smith 1991; Woltjer 1972; Gordon *et al* 1998). We adopt a value of 2×10^4 yr in this analysis (corresponding to the oldest SNRs in the Sedov–Taylor phase in the M33 sample; see Smith 1991; Gordon *et al* 1998). The other important time scale is that over which the relativistic medium would dissipate in the absence of new particles. In cosmic ray terms this is the residence time, $\tau_{\rm r}$, of the relativistic particles in the disk of a galaxy and is governed by a combination of escape and energy losses.

The average rate, $\dot{U}_\Sigma$, at which the SNR population produces energy in the form of relativistic particles can be expressed as

$$\dot{U}_\Sigma = \frac{U_\Sigma}{\tau_{\rm snr}}. \tag{6}$$

The relativistic particles escape from the disk at a rate of

$$\dot{U}_{\rm D} = -\frac{U_{\rm D}}{\tau_{\rm r}}. \tag{7}$$

Consider now a steady state scenario in which the energy content of the cosmic rays is independent of time over the time scale, $\tau_{\rm r}$. Then, $\dot{U}_\Sigma = -\dot{U}_{\rm D}$ and, from equations (6) and (7),

$$\frac{U_\Sigma}{\tau_{\rm snr}} = \frac{U_{\rm D}}{\tau_{\rm r}} \rightarrow \frac{U_\Sigma}{U_{\rm D}} = \frac{\tau_{\rm snr}}{\tau_{\rm r}}. \tag{8}$$

The prediction is that $U_\Sigma/U_{\rm D}$ should roughly equal $\tau_{\rm snr}/\tau_{\rm r}$, a quantity which for our galaxy is ≈ 0.003. Measurements of U_Σ and $U_{\rm D}$ can test whether SNRs can do the job.

In the above scenario, the SNRs provide the link between star formation and radio continuum emission. Since only the highest-mass stars end up as supernovae and since the highest-mass stars evolve the most rapidly there is a very tight link between the presence of star formation and the production of cosmic rays. In fact, cosmic rays can be established on time scales small compared with the duration of the star forming episode of a galaxy. Thus, to first order, any changes in the radio continuum emission of a galaxy can be correlated with changes in the star formation rate (Deeg *et al* 1977). It is for this reason that radio continuum emission has been used as a tracer of star formation in galaxies (e.g. Condon 1992). Although the above theoretical considerations are compelling, supporting empirical evidence is relatively limited.

Studies of relativistic gas in disk galaxies have taken three basic approaches. The first utilizes integrated

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

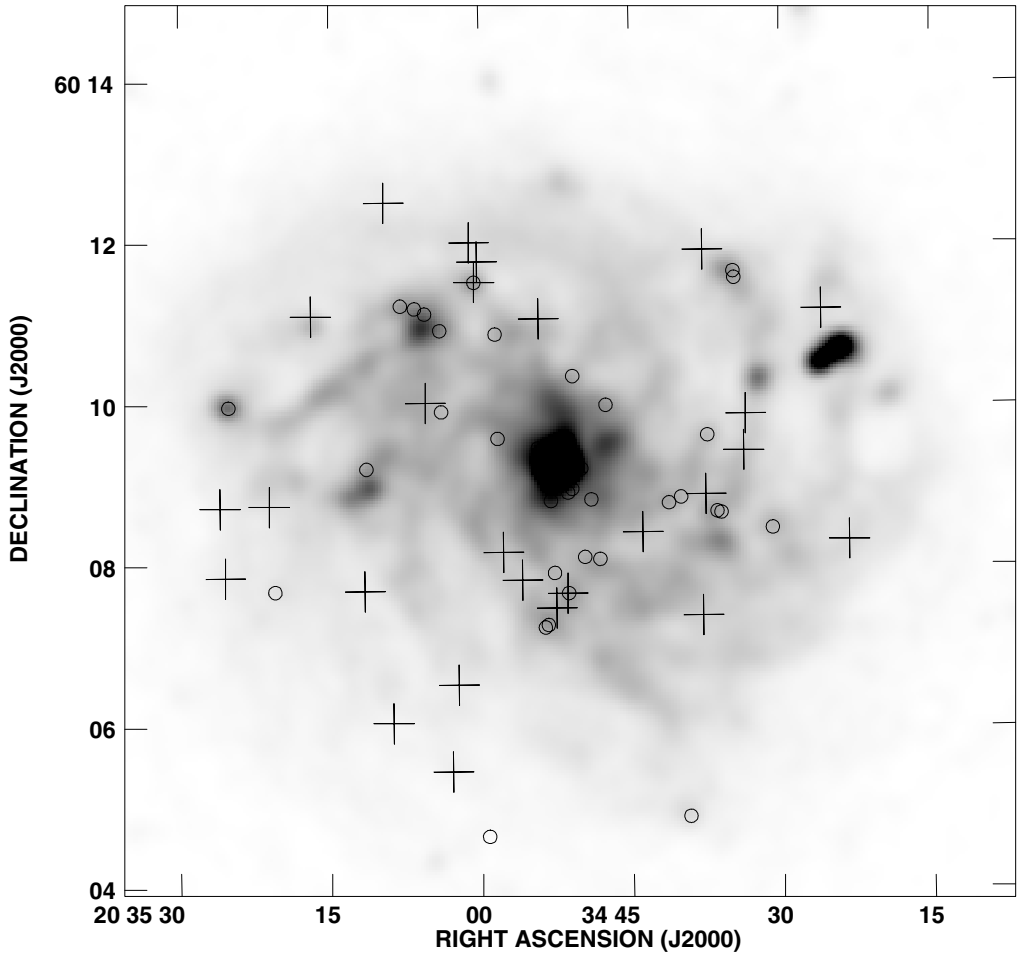


Figure 3. A 20 cm radio continuum image of NGC 6946. The resolution is roughly 60″. The open circles show the location of SNRs which are emitting detectable synchrotron radiation. These are the SNRs that are also producing cosmic ray electrons. The crosses mark the positions of an optically selected sample of SNRs which, for the most part, have no detectable radio emission and are therefore producing cosmic ray electrons in significantly lesser amounts. The correlation of the radio-emitting SNRs with the synchrotron arms of NGC 6946 suggests that these SNRs have high-mass stellar precursors and that they play a role in establishing the relativistic medium of that galaxy. The anti-correlation of the radio-quiet SNRs with the arms suggests that many of these SNRs not only are poorer sources of cosmic ray electrons but may well have an origin associated with the older, lower-mass stars. These observations are consistent with the scenario in which only those SNRs with high-mass stellar precursors play a leading role in shaping the radio continuum emission of galaxies.

properties of galaxies and is essentially a statistical study of samples of galaxies. The radio–FIR correlation (Wunderlich *et al* 1987) is a good example of this approach. This remarkably tight correlation between an indicator of star formation (top left arm in figure 1) and an indicator of cosmic ray electrons (bottom arm in figure 1) argues strongly for a physical relationship between star formation and cosmic ray production (Harwit and Pacini 1975). The scenario that has evolved from this argument is that SNRs produced by type II supernovae (resulting from recently formed high-mass stars) explosions produce relativistic particles through their interactions with the interstellar medium (Völk 1989; Condon *et al* 1991; Helou and Bicay 1993; Xu *et al* 1994).

The second approach entails analyzing and comparing light profiles in individual galaxies. Comparisons between optical light distributions at various wavelengths and high-resolution radio continuum images have been used to correlate STELLAR POPULATIONS with the relativistic gas (e.g. Duric *et al* 1996; Condon 1992). Recent results suggest that Hα and radio continuum distributions correlate the best (e.g. Duric *et al* 1996). Thus, there is an implied relationship between the very young stars (as represented by the Hα emission—top right arm in figure 1) and the relativistic gas (represented by the synchrotron emitting electrons—bottom arm in figure 1).

The two different approaches have yielded similar results in that they have established a plausible relation-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ship between star formation and the relativistic gas in disk galaxies. There is little doubt that this relationship is real. However, the physical chain of events that leads to the production of the relativistic medium can only be inferred from these studies; it cannot be directly demonstrated. In particular, the connection between SNRs and the production of relativistic particles needs to be more directly examined. This latter consideration forms the basis of a third approach, first applied to M33 (Duric *et al* 1995). In this approach, U_Σ, U_D and τ_r are determined empirically for a given galaxy, thereby enabling a direct test of the hypothesis that SNRs produce the bulk of the cosmic rays in normal galaxies.

In the prototype study of M33 the above approach was used to ascertain whether the population of SNRs resident in M33 could account for the cosmic ray electrons present in the disk of M33. It was found that the collective cosmic ray content of the SNRs is approximately 10^{52} erg whereas the entire disk contains about 10^{54} erg worth of cosmic rays. In other words, only about 1% of the cosmic rays actually reside within SNRs at any given time. This makes sense because the lifetime of a cosmic ray electron is much longer than the lifetime of an SNR. As indicated by equation (8), the ratio of energies should equal the ratio of lifetimes in a steady state configuration. Consequently, the figure of 1% is roughly what is needed to make SNRs the source of most cosmic rays in the disk of M33. The results of this study support the hypothesis that SNRs produce the bulk of the cosmic rays in galaxies, thereby supporting the chain of events that make up the bottom arm in figure 1.

An ongoing study presents an alternative approach in which the distributions of cosmic ray-producing SNRs are compared with large-scale synchrotron features in a galaxy. In figure 1, the bottom arm contains a link between the SNRs and the large-scale distribution of cosmic rays. Until now few data existed to support the view that the cosmic rays propagate from the SNRs (where they are formed) into the extended relativistic medium where they emit most of their radiation. A study of NGC 6946 by Lacey (1998) has helped fill in this gap. Her results suggest that active SNRs are spatially correlated with synchrotron arms of the galaxy NGC 6946.

Figure 3 shows a 6 cm radio continuum image of NGC 6946 (courtesy of Tom Pannuti, University of New Mexico, and Rainer Beck, Max Planck Institute for Radio Astronomy) made from combined Effelsberg 100 m single dish data and D-array VLA data. Plotted are the radio-emitting (and therefore cosmic ray producing) SNRs as circles. It is evident that the SNRs are located along ridge lines of extended radio continuum emission indicating a possible relationship between the SNR population and the large-scale distributions of cosmic ray electrons. Interestingly, those SNRs that are not radio sources (and therefore do not appear to be producing measurable quantities of cosmic rays) are not correlated with any large-scale enhancements of synchrotron emission. These may well be SNRs evolving from type supernovae whose progenitors are the older population II stars. These stars tend to be found in regions of the galaxy containing lower densities of gas compared with the population I stars. Consequently, the resulting SNRs interact with a less dense medium and produce fewer cosmic rays. These results support the standard picture in which only the SNRs with the massive stellar precursors contribute to the production of cosmic rays.

Examination of the empirical evidence in its totality points to a consistent picture in which SNRs play a central role in producing and maintaining the CR gas in normal galaxies. It therefore appears that SNRs of type II supernovae are the most likely link between star formation and the radio continuum emission of normal galaxies. In a broader context, star formation via the action of the SNRs explains not only the radio continuum emission of galaxies but also the radio–FIR correlation and the Hα correlation of galaxies.

Bibliography

Alfvén H and Herlofson N 1950 *Phys. Rev.* **78** 616
Anderson M C and Rudnick L 1993 *Astrophys. J.* **408** 514
Axford W I 1981 *Proc. 20th Int. Cosmic Ray Conf.* vol 12, p 155
Bell A R 1978 *Mon. Not. R. Astron. Soc.* **182** 147
Blandford R D and Eichler D 1987 *Phys. Rep.* **154** 1
Bothe W and Kolherster W 1929 *Z. Phys.* **56** 751
Condon J J 1992 *Ann. Rev. Astron. Astrophys.* **30** 575
Condon J J, Anderson M L and Helou G 1991 *Astrophys. J.* **376** 95
Deeg H-J, Duric N and Brinks E 1997 *Astron. Astrophys.* **323** 323
Devereux N, Duric N and Scowen P 1997 *Astron. J.* **113** 236
Duric N 1988 *Space Sci. Rev.* **48** 73
Duric N, Glenn J, Grashuis R, Kunkle J and Collins A 1996 *Astrophys. J.* **470** 814
Duric N, Gordon S M, Goss W M, Viallefond F and Lacey C 1995 *Astrophys. J.* **445** 173
Drury L O'C 1983 *Rep. Prog. Phys.* **46** 973
Ginzburg V L and Syrovatskii S I 1965 *Ann. Rev. Astron. Astrophys.* **3** 297
Gordon S *et al* 1998 *Astrophys. J.* **117** 89
Gordon S M, Duric N, Kirshner R P, Goss W M and Viallefond F 1999 *Astrophys. J. Suppl.* **120** 247
Harwit M and Pacini F 1975 *Astrophys. J. Lett.* **200** L127
Helou G and Bicay M D 1993 *Astrophys. J.* **415** 93
Hess V 1912 *Phys. Z.* **13** 1084
Jansky K H 1935 *Proc. IRE* **23** 1158
Jones T W *et al* 1998 *Publ. Astron. Soc. Pac.* **110** 125
Kipenheuer K O 1950 *Phys. Rev.* **79** 738
Lacey C K 1998 *PhD Thesis* University of New Mexico
Pacholczyk A G 1970 *Radio Astrophysics* (San Francisco, CA: Freeman)
Reber G 1940 *Astrophys. J.* **91** 621
Shu F H 1991 *Radiation* (Mill Valley, CA: University Science Books)
Smith R C 1991 *PhD Thesis* Harvard University
Stromgren B 1939 *Astrophys. J.* **89** 526

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tucker W H 1975 *Radiation Processes in Astrophysics* (Cambridge, MA: MIT Press)
Völk H J 1989 *Astron. Astrophys.* **218** 67
Woltjer L 1972 *Ann. Rev. Astron. Astrophys.* **10** 129
Wunderlich E, Klein U and Wielebinski R 1987 *Astron. Astrophys. Suppl.* **69** 487
Xu C, Lisenfeld U and Völk H J 1994 *Astron. Astrophys.* **282** 19
Zwicky F 1939 *Phys. Rev.* **55** 986

Neb Duric

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Star of Bethlehem

The biblical Star of Bethlehem, which heralded the birth of Jesus Christ, is only mentioned in the Gospel of St Matthew 2. The astrologically significant 7 BC triple conjunction of Jupiter and Saturn in the constellation of Pisces is the most likely candidate, although a COMET/NOVA in 5 BC and a comet in 4 BC cannot be ruled out. There is also the possibility that the star was simply fictitious.

Introduction

The Jesus biblical nativity story was added, somewhat belatedly, as a prelude to only two gospels. Luke has a peculiar census/taxation system, a journey from Nazareth, an over-booked inn, a manger, a visitation of shepherds in the fields on the birth night, circumcision, purification and a journey home. Matthew has a family living in Bethlehem, a star, Herod, magi, gifts, the massacre of the innocents, a flight into Egypt and a relocation to Nazareth in the first year of the reign of Archelaus.

There is no star in Luke, and in Matthew 2 the word 'star' is used only four times. The stellar facts are as follows.

(1) As Jesus was born in the last four or so years of the reign of Herod the Great (Matthew 2:1), the star occurred sometime between 8 BC and Herod's death around Easter 4 BC. (A few, however, put Herod's death as late as January 1 BC.) This time interval considers both the Lukean census (the Roman *Monumentum Ancyranum* of Caesar Augustus dates this as 746/747 *ab urbe condita*, i.e. 8/7 BC) and the massacre of the innocents (two years old and under, Matthew 2:16) and the two years in which the Holy Family were in Egypt, sheltering from the wrath of Herod the Great. In dating the star we must ignore the hijacking (around AD 336) by early Christians of the pagan winter solstitial feast of *dies solis invicti natalis* for the celebration of the birthday of Christ. Also Dionysius Exiguus was considerably in error, in AD 525, when he based his new calendrical year numbering on the calculation that Jesus was born on 25 December, 1 BC and was about one week old on 1 January, AD 1. Jesus was six months younger than John the Baptist (Luke 1:36). Early Christian and Muslim tradition suggests that Jesus was born in the autumn and on the day after the Jewish Sabbath. It is most unlikely that the shepherds would be 'in the field, keeping watch over their flock by night' (Luke 2:8) in an Israeli December. Modern Christmas celebrations owe much to ancient pagan mid-winter Sun-worshipping customs.

(2) The magi were most probably learned Babylonian Zoroastrian priests and astrologers, and certainly not 'kings', their later beatification and royalty being mainly due to Psalms 68:29, 72:10, Isaiah 49:7 and 60:3, 60:10. They first saw the star heliacally 'in the east', '*en te anatole*', in the pre-dawn sky (Matthew 2:2).

(3) Two thousand years ago the word 'star' was a generic term for any astronomical object. Planets were wandering stars, novae were new stars, comets were hairy stars and meteors were shooting stars. All could have been the 'star' of Bethlehem.

(4) The magi talked of 'his' star (Matthew 2:2). It clearly had an astrological association with the birth of a new 'King of the Jews' (not the expected Messiah). They then traveled to Jerusalem (not Bethlehem), as ambassadors, to pay homage to this king and to present gifts. Visits by Magi were commonplace in those days. The association of stars with kings and new rulers was not unusual at the time (see Numbers 24:17).

(5) Herod and all Jerusalem had not noticed the star (Matthew 2:3) and had not witnessed its first appearance (Matthew 2:7). This indicates that the star was insignificant. Gross exaggerations as to its brightness were introduced later on by St James (*Protoevangelium*, 7, written about AD 150), Ignatius (*Epistle to the Ephesians*, XIX) and Origen (*Contra Celsum*, 1:58). Terms such as 'indescribably great', 'its light was unspeakable' and 'new' must be contrasted with the adjectivaly unadorned Matthean star.

(6) The star appeared twice. It was first seen at dawn from the magi's home town (Matthew 2:2). Its second appearance, probably in the evening southern sky, was after they had met Herod in Jerusalem, and were on the short 10 km southward journey to Bethlehem (Matthew 2:9). Their exceeding great joy on witnessing the second appearance indicates that this had been predicted, and they were pleased with the confirmation.

(7) The star 'went before them, until it came and stood over where the young child was' (Matthew 2:9). This is an astronomical stumbling block. Stars do not 'go before'. Comets have, however, been noted by contemporary historians such as Dio Cassius (*Roman History*, 54:29) and Josephus (*Jewish War*, 6:5:3) to 'stand over' individual cities. Artists also have often used comets as Stars of Bethlehem in their nativity scenes, the AD 1303 fresco in the Arena Chapel in Padua by Giotto being a famous example.

(8) The star's circumstances were uncommon. Regular astronomical events such as the stunning evening and morning appearances of Venus, and the zodiacal light, annual meteor showers, occasional fireballs and lunar occultations of Jupiter are surely so unremarkable that they would not regularly encourage magi to buy gifts and to travel long distances.

Astronomical possibilities

The facts and time frames must now be related to contemporary astronomical events. Fortunately Chinese observational astronomy was flourishing, and careful records were kept of unusual occurrences such as comets, novae, sunspots and shooting stars, for astrological purposes. Babylonian planetary astrology was also active but unfortunately only a few of the relevant clay tablets have been found. Ho (1962) recorded contemporary Far Eastern diary records. His object number 61 was Comet Halley. This was seen in 12 BC, as a *po-hsing* (a comet without a tail), but was too early.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

For 70 days between 9 March and 6 April 5 BC a *sui-hsing* (a comet with a tail) was recorded in Capricornus near the asterism Chhien-niu (RA 20.25 h, dec 15°). As no details are given as to the motion of this object, Clark *et al* (1977) take it to be a nova. The long duration is taken to indicate that it was bright.

A comet (*po-hsing*) was seen in April 4 BC near Ho-Ku in Aquila (RA 19.6 h, dec +8°). Duration details are not given so this comet was probably faint and short-lived.

Ho's object number 65, a comet seen in December AD 13, is far too late.

The Star of Bethlehem might have been a planetary conjunction. Here the Earth-based observer sees two or more PLANETS nearly along the same line of sight, this producing an impressive grouping on the celestial sphere. There was a significant triple conjunction of Jupiter and Saturn in 7 BC in Pisces. These planets have orbital periods of 11.9 and 29.5 yr respectively and thus pass each other in the sky every 19.9 yr. Every 120 yr or so, the Sun is also close to the Earth–Jupiter–Saturn line giving three conjunctions in about six months. Every 796.4 yr these triple conjunctions take place in Pisces, the last being in April 1573. In February 6 BC Mars joined Jupiter and Saturn forming a 'massing' only 8° across.

One advantage of a planetary conjunction 'star' is that the astronomical and astrologically uninterested would most likely not know it was happening, and, with the exception of point (7), it fits the facts. A planetary conjunction would have a clear message to astrological magi. It would tell them where to go, when to go and what to expect when they got there. Kennedy and Pingree (1971) noted how an 8th century AD astrological world history by Masha'allah, based on earlier Babylonian and Iranian ideas, stressed the role of Jupiter–Saturn 'great' conjunctions as heralds of important religious and political events. The Noachian deluge and the births of Jesus and Mohammed were mentioned particularly. In AD 1497 Don Isaac Abrabnel stressed that the constellation Pisces was specifically associated with Israel in Magian astrology. Saturn was regarded as the old ruler (and father), with Jupiter as the new king, and son. Both interpretations are unfortunately *post facto*. Predictions of the 7 BC triple conjunction have also been found on Babylonian clay tablets (see Sachs and Walker 1984).

A Jupiter–Saturn–Mars massing was observed in 1604 by Johannes Kepler. The contemporary appearance of the supernova of 1604 convinced Kepler that the latter was the Star of Bethlehem.

Other conjunctions have been suggested. Sinott (1968) and Mosley (1987) concentrated on the rather late June 2 BC close conjunction of Jupiter and Venus near Regulus in Leo. Much is made of the Judea–Lion association and of the possibility that Herod might have lived until just after the total lunar eclipse of 10 January 1 BC as opposed to the partial lunar eclipse of 13 March 4 BC.

Molnar (1999) reinterprets the astrological evidence. He is convinced that the constellation associated with Israel was not Pisces but the adjacent sign Aries. This is not only favoured by the Egyptian–Greek astronomer–astrologer Ptolemy, in his *Tetrabiblos*, but is also supported by Syrian coinage which shows a leaping Ram peering over its shoulders at a star. A diligent search is then made of the astronomical records for events that take place in Aries. Molnar also assumed that the Magi are Hellenistic astrologers, and comparisons are made with the natal horoscopes of other 'kings' such as the Emperor Hadrian and Antiochus I of Commagene. Jupiter and the Moon have important roles to play in these regal horoscopes, and there were two lunar occultations of Jupiter, in Aries, in 6 BC, one on 20 March, and the other on 17 April. As the first was 'weakened' by the nearness of the Sun in the sky Molnar chooses the second and suggests that Jesus was born on 17 April in 6 BC.

Unfortunately this thesis has certain problems. Both events were invisible. The 20 March occultation of Jupiter by the Moon occurred just before sunrise, and the 17 April occultation occurred at local noon. So the fact that Matthew stresses that the 'star' was 'seen' twice is ignored. Molnar surmises that astrologers only needed to calculate that something happened for it to be significant. However, even though Babylonian astronomers could easily predict planetary conjunctions and massings, the fact that the Moon is only 0.5° across, and lunar theory is extremely complicated, means that predicting occultations was way beyond them. We have to wait for the 17th century and a genius like Edmond Halley before lunar occultations could be predicted with any accuracy. Also, lunar occultations of Jupiter are most probably just too common to be the Star of Bethlehem. Jupiter is in Aries for 1 yr in every 12, and the Moon, in that year, moves through Aires 12 times.

Let us now consider the nova/comet in Capricorn, seen during March and April 5 BC, and the tail-less comet, or nova, seen in Aquila in April 4 BC. The former is favoured by Humphreys (1993) and by Clark *et al* (1997). Comets (which in those times were regarded as meteorological phenomena) have the advantage of 'standing over' specific places for short times. However, astrologically, comets are regarded as rather broad-brush indicators of doom, diseases, disaster and the death of kings. Needless to say, the death of one king could be good news for his successor. The snag with comets and novae is that they are completely unexpected. The astrology of comets is associated with the constellation that they appear in and the form and direction of their tails. Neither Capricornus nor Aquila has anything to do with Judea. Novae and comets brighten and then decay. They are not seen twice, as the Star of Bethlehem was. Comets are usually only seen after they have passed the Sun. It is an unusual comet that is seen both on its way in to the Sun and again on its way out. In this unusual case, we have had to wait for Isaac Newton, in the 1680s, to show that the same object was being seen on both occasions.

So the 'exceedingly great joy' of the magi, a joy that is often taken to indicate that one of their celestial

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

predictions had come true, and the same object had been seen twice, does not apply to comets or novae. Also, the 5 BC comet, that is seen for 70 days, will be bright at maximum, and it is then difficult to explain why Herod had not seen it and was taken by surprise by the Magi's visit.

Let us return to the triple conjunction of 7 BC. Both Jupiter and Saturn were in opposition in Pisces and were moving around their retrograde loops. During late 8 BC and early 7 BC the planets approached each other at about 3.5° per month. On 27 May 7 BC they were only 1° apart (twice the lunar diameter). They then separated slightly, and by 27 July were 2.9° apart. Coming together again they were 1° apart on 6 October, 1.2° apart on 1 November and 1.05° apart on 1 December. During 6 BC they moved away from each other, as quickly as they had come together in early 7 BC. A 1° separation is not startling. This sight could easily be ignored by the astrologically uninterested. The close conjunctions in May and October–November provide an ideal 'two stars'. The six months between these two events also gives ample time for the Magi's preparation and travel. Proponents of this theory (for example Hughes 1976, 1979) regard Jupiter as the Star of Bethlehem. The date of the birth of Jesus is still uncertain but the astrologically minded Magi would have possibly associated it with the acronychal rising of the two planets, i.e. Tuesday 15 September 7 BC. The magi would have predicted this. There was no rush. They did not have to pay homage to the new King of the Jews on his actual day of birth; any time when he was a young child would do. Unfortunately, like most astronomical interpretations, the 'going before' and 'standing over' is still a problem.

Conclusion

The astrological interpretation of the Star of Bethlehem has one great snag. It gives a degree of credence and verisimilitude to a subject that is a quagmire of subjective, unscientific partiality. This interpretation has the star as a tangible physical phenomenon.

There are two other possibilities. One abrogates the scientific interpreter of all responsibility and regards the whole phenomena as miraculous. The passage of one age and the dawn of the next is thus heralded by a miraculous occurrence, and as such the star is inexplicable.

A more appealing proposal is that the star is an example of *midrash*. Thus it is nothing more than a fictitious invention giving fulfilment to the Old Testament prophecy of Balaam in Numbers 24:17 ('there shall come a star out of Jacob'). The Star of Bethlehem would then join a host of other fascinating historic stars and comets that were thought to foretell the birth and death of heroes, kings and emperors. Matthew would have used it to add credence to the messianic nature of Jesus and to encourage more of his Jewish audience to convert to Christianity.

However, if it was an invention, why did Matthew omit one of his favourite phrases, 'that it should be fulfilled', and why did he invent such an ordinary star, or one that had such similarities to an actual celestial occurrence? To me the Gospel of St Matthew rings true. Its message is unembellished. I believe that the Star of Bethlehem actually existed, and that astronomical historians will enjoy searching for it for many years to come.

Bibliography

Brown, R E 1977 *The Birth of the Messiah: A Commentary on the Infancy Narratives in Matthew and Luke* (London: Geoffrey Chapman)

Clark, D H, Parkinson, J H and Stephenson, F R 1977 An astronomical re-appraisal of the star of Bethlehem — a nova in 5 BC, *Q. J. R. Astron. Soc.* **18** 443–9

Edwards, O 1986 *The time of Christ: a Chronology of the Incarnation* (Edinburgh: Floris)

Ferrari D'Occhieppo, K 1977 Der Stern der Weisen *Geschichte oder Legende* (Vienna: Herold)

Ho P-Y 1962 Ancient and mediaeval observations of comets and novae in Chinese sources *Vistas Astron.* **5** 127–225

Hughes, D W 1976 The Star of Bethlehem *Nature* **264** 513–7; **268**, 565–7

Hughes, D W 1979 *The Star of Bethlehem: an Astronomer's Confirmation* (New York: Walker)

Humphreys, C 1993 The Star of Bethlehem *Sci. Christ. Belief* **5** 83–101

Kennedy, E S and Pingree, D 1971 *The Astrological History of Masha'allah* (Cambridge, MA: Harvard University Press)

Kidger, M 1999 *The Star of Bethlehem: an Astronomer's View* (Princeton, NJ: Princeton University Press)

Maunder, E W 1908 *The Astronomy of the Bible* (London: Sealey Clark)

Molnar, M R 1999 *The Star of Bethlehem: the Legacy of the Magi* (Piscataway, NJ: Rutgers University Press)

Mosley, J 1987 *The Christmas Star* (Los Angeles, CA: Griffith Observatory)

Sachs, A J and Walker, C B F 1984 Kepler's view of the Star of Bethlehem, and the Babylonian Almanac for 7/6 BC *Iraq* **46** 43–55

Sinnott, R 1968 Thoughts on the Star of Bethlehem *Sky Telescope* **36** 384–6

Teres, G 2000 *The Bible and Astronomy: The Magi and the star in the Gospel* (Budapest: Springer Orvosi Kiad Kft)

Scarola, J V 1991 *A Chronographic Analysis of the Nativity* (New York: Vantage)

David W Hughes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Starburst Galaxies

A starburst galaxy is one undergoing a brief episode of intense STAR FORMATION, usually in its central region. The massive stars in the burst generate most of the total luminosity of the entire galaxy. Starburst galaxies are fascinating objects in their own right and are the sites where roughly 25% of all the massive stars in the local universe are being formed. They offer unique laboratories for the study of the formation and evolution of massive stars, the effects of massive stars on the interstellar medium and the physical processes that were important in building galaxies and chemically enriching the INTERGALACTIC MEDIUM.

What is a starburst?

Overview

There is no rigorous definition of a starburst galaxy, but several different criteria are often used. These are clear conceptually, but each is subject to uncertainties in its application.

Starbursts are usually located in the centers of galaxies, and have typical radii of 100–1000 pc (1–10% of the size of their 'host' galaxy). Despite their small size, they are converting gas into massive stars at a rate that exceeds that found throughout the rest of their host galaxy. The starburst is 'fueled' by a copious supply of interstellar gas (primarily in the form of molecular hydrogen) that has been accumulated in the center of the galaxy. The available gas supply is sufficient to sustain the current rate of star formation for only of order 10^8 yr ($\sim$1% of the age of the universe). The dust grains associated with the molecular gas usually absorb most of the radiation produced by the burst's stars. This can make it difficult to determine many of the basic properties of starbursts.

Star-formation rates

We do not directly measure the star-formation rate in a starburst but must infer it from measurements of the starburst's luminous output. Two techniques are routinely used.

The first is to use the bolometric luminosity of the starburst ($L_{\rm bol}$), which is the luminosity integrated over the entire electromagnetic spectrum. $L_{\rm bol}$ for a starburst is dominated by the emission from hot, massive, short-lived stars. These have masses greater than about 8 times the mass of the Sun ($M > 8M_\odot$) with lifetimes of less than about 40 million years. The bulk of the radiation from these stars is emitted in the ultraviolet part of the spectrum. Dust grains in interstellar medium of the starburst absorb this radiation, are heated, and cool by emitting far-infrared radiation. Thus, a good approximation of $L_{\rm bol}$ is given by the sum of the starburst's ultraviolet and infrared luminosities.

The second technique uses measurements of the amount of ionizing radiation produced by the starburst (often denoted by Q, the total number of photons produced per second that are capable of ionizing atomic hydrogen). The production of these ionizing photons is dominated by the most massive, shortest-lived stars ($M > 25M_\odot$ and lifetimes less than 7 million years). Measurements of Q are based on measurements of the rate at which H ions and electrons recombine, which in turn is measured by the luminosity of the so-called recombination emission lines of hydrogen (after the luminosity is corrected for absorption due to dust inside the starburst).

Given a measurement of $L_{\rm bol}$ or Q, we can estimate the mass of the population of massive stars required to produce this quantity. If we divide this mass by the lifetime of the appropriately massive stars we obtain an estimate of the mean formation rate of the massive stars. This technique tells us almost nothing about the rate at which low-mass stars (like our Sun) are being formed, since these stars contribute only a negligible amount to $L_{\rm bol}$ or Q. Indeed, it is entirely possible that starbursts form only massive stars (unlike the mode of star formation in normal galaxies like our own). If we assume that starbursts form a normal complement of low-mass stars, the implied star-formation rates usually range from $1M_\odot$ to $100M_\odot$ yr^{-1}.

The burst intensity

A useful way to define a starburst is to consider the burst intensity—the rate of star formation per unit area ($\Sigma_{\rm SFR}$—typically given in units of $M_\odot$ per year per kiloparsec^2). In normal star-forming galaxies like our own Milky Way, the star-formation rate is a few $M_\odot$ per year throughout a galactic disk with a radius of about 10 kpc (e.g. $\Sigma_{\rm SFR}$ has a typical value of $10^{-2}M_\odot$ yr^{-1} kpc^{-2}). In a typical starburst galaxy, the star-formation rate would be $10M_\odot$ yr^{-1} in a region with a radius of 0.5 kpc. The implied $\Sigma_{\rm SFR}$ is then $10M_\odot$ yr^{-1} kpc^{-2}, or 10^3 times greater than in a normal galaxy.

It is important to emphasize that there is no particular 'magic' value for $\Sigma_{\rm SFR}$ that separates normal and starburst galaxies. A continuum of values is observed, spanning at least 6 orders of magnitude from the most quiescent star-forming normal galaxies to the most intense starbursts.

Although calculating a global value for $\Sigma_{\rm SFR}$ is a useful way to quantify a starburst, star formation is not uniformly distributed throughout the bursting region. Instead, the star formation occurs both in compact (few parsec scale) star clusters and in a more smoothly distributed mode. The most massive CLUSTERS (the 'super star clusters') have estimated masses of (10^5–10^6) $M_\odot$ and may be close analogs to young GLOBULAR CLUSTERS.

The burst duration

A starburst is by definition a transient event. The duration of a starburst (Δt) must be much smaller than the age of the galaxy in which it occurs ($\Delta t \ll 10^{10}$ yr). Another way of stating this requirement is that the present rate of star formation must greatly exceed the past rate averaged over the age of the galaxy. Unfortunately, it is difficult to accurately determine how long starbursts last.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Perhaps the most commonly used technique to estimate Δt is to calculate the gas-depletion time: the mass of interstellar gas in the starburst divided by the present rate of star formation. This is then a rough estimate of how much longer the starburst can be sustained before running out of gas. Gas-depletion times in starbursts are usually of order 10^8 yr, but they are highly uncertain for many reasons. For one thing, the mass of molecular gas is difficult to determine to better than a factor of a few since the determination relies on indirect arguments and observations of a trace molecule (usually CO). The estimated molecular gas masses range from $10^8 M_\odot$ to $10^{10} M_\odot$, increasing as a function of the luminosity of the starburst. In addition, the star-formation rate itself is quite uncertain, primarily because we do not know the rate at which low-mass stars are being formed (see above).

More sophisticated estimates of the duration of starbursts utilize measurements of Q, $L_{\rm bol}$ and the starburst mass plus detailed information on the stellar population (for example the relative numbers of massive main sequence and post-main-sequence stars such as red supergiants and Wolf–Rayet stars). These measurements are then compared with models for the evolution of a population of massive stars (see STELLAR EVOLUTION). Resulting estimates for burst duration range from a few million years for the smallest and least powerful starbursts (which occur in dwarf galaxies), up to 10^7 or 10^8 yr for powerful starbursts.

More generally, the minimum possible duration of a starburst is set by considerations of causality: the duration of star formation cannot be significantly less than the time for gas on one side of the starburst to 'communicate' with gas on the other side. Such 'signals' will travel at speed of sound in the gas or the velocity induced by the gravitational field in the starburst. These velocities will be of order 10^2 km s^{-1}, so $\Delta t > 10^7$ yr for a starburst with a diameter of 1 kpc. The most powerful and intense starbursts seem to turning gas into stars as fast as allowed by causality.

How do we find starburst galaxies?

The massive stars that power starbursts are mostly very hot ($T \sim 20\,000$–$50\,000$ K) and emit most of their radiation in the ultraviolet between 912 Å and roughly 2000 Å. Thus, one technique for finding starbursts is to search for galaxies that are unusually bright in the ultraviolet. The spectral region below about 3200 Å is inaccessible from the surface of the Earth, so most ultraviolet surveys for starbursts in the local universe have been conducted in the near-ultraviolet region just longward of the atmospheric cut-off. This will change in the near future as space-based ultraviolet imaging surveys are conducted at shorter wavelengths.

Starbursts are rich in interstellar gas and dust (the raw material for star formation). The primary radiative output from the massive stars (ultraviolet radiation) is absorbed by this interstellar material which then re-radiates the energy in other forms. The gas absorbs nearly all the stellar photons with energies sufficient to photoionize neutral hydrogen ($E > 13.6$ eV, or $\lambda < 912$ Å). The subsequent recombination of hydrogen or helium ions and resulting radiative cascade produce H and He recombination emission lines. The free thermal electrons in the gas can also collisionally excite ions, whose radiative decay produces emission lines. Many of the strongest lines of both types are in the visible part of the spectrum, and together constitute a few per cent of the bolometric luminosity of the starburst. Thus, another way to find starbursts is to search for galaxies with unusually bright emission lines.

The dust grains in the starburst are effective at absorbing ultraviolet photons of all wavelengths. The grains are heated by this radiation and cool by emitting radiation. The equilibrium temperatures that result from balancing heating and cooling rates are usually in the range 10–100 K, and the emitted radiation therefore lies in the mid- and far-infared spectral region ($\lambda \sim 30$–300 μm). The survey by the INFRARED ASTRONOMY SATELLITE (IRAS) in the 1980s (which surveyed nearly the entire sky in the mid and far infrared) has produced the most extensive and best-studied sample of starbursts.

These three types of surveys select samples of starbursts that overlap one another but nevertheless have important systematic differences. The most direct difference is that dusty starbursts are preferentially detected in the far-infared surveys, while the less dusty starbursts are preferentially found by the ultraviolet and emission-line surveys. Dust grains are made of elements heavier than H or He ('metals'). Thus, the dust-content of starbursts correlates well with chemical composition: typically only a few per cent of ultraviolet radiation escapes 'metal-rich' starbursts (having chemical composition like the Sun) while the majority of the ultraviolet light escapes the most 'metal-poor' starbursts. Metal-poor starbursts tend to be less powerful and to occur in smaller and less massive galaxies (as a result in part of the well-known mass–metallicity relation for galaxies).

What causes a starburst?

From a purely empirical point of view, the causes of starbursts become increasingly clear for starbursts of greater and greater luminosity. The most luminous starbursts in the local universe are the so-called 'ultra-luminous infrared galaxies', which have bolometric luminosities of roughly $10^{12} L_\odot$ (nearly all of which is emitted in the mid- and far-infrared). The power source is deeply buried inside a dense, dusty region of molecular gas only a few hundred parsecs in size and may consist of a combination of a starburst and a dust-enshrouded QUASAR. The mass of molecular gas ($\sim 10^{10} M_\odot$) is comparable with the entire mass of the interstellar medium in a big SPIRAL GALAXY.

These ultraluminous galaxies almost invariably have highly disturbed morphologies that are strongly suggestive of the ongoing or recently completed merger of two large DISK GALAXIES. Specific morphological structures indicative of mergers include long narrow 'tidal tails'

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of stars and gas (the remnants of the outer disks of the merging galaxies) and double nuclei. For less powerful starbursts the evidence suggests that more mild gravitational interactions between galaxies are an important triggering mechanism (e.g. the close passage of two galaxies, without their subsequent merger).

During the close passage of two galaxies, tidal stresses act to strongly perturb the orbits of the stars and gas in the galaxy disk. The dissipation of kinetic energy as gas collides with gas allows the gas to become sufficiently displaced from the stars that gravitational torques act between the stars and gas to transfer significant amounts of angular momentum from the gas to the stars. The gas can thereby flow into the center of the galaxy, where it can fuel a starburst. If the passage of the two galaxies is slow and interpenetrating enough, dynamical friction can transfer enough kinetic energy from the stars to the galaxy dark-matter halos to allow the two galaxies to merge into a single galaxy. Such mergers or strong interactions should take a few times the galaxy rotation period (about 10^9 yr), with the intense starburst phase being significantly shorter (see GALAXIES: INTERACTIONS AND MERGERS). These timescales are loosely consistent with independent estimates of starburst lifetimes discussed above.

However, many starbursts are not found in obviously interacting systems. In such cases, a stellar bar is often present in the inner disk of the starburst galaxy. A bar can act to rob gas of its angular momentum, and transfer gas into the center of the galaxy where it can fuel a starburst. The mechanism for triggering starbursts is most uncertain in the lowest-power starbursts, which occur in DWARF GALAXIES. In some cases it appears that the dwarf may have recently collided with an extragalactic gas cloud, but in most cases there is no clear evidence for any type of interaction or stellar bar.

Starbursts from a cosmological perspective

Starbursts in context

How important are starbursts? The far-infrared IRAS survey produced the best-studied sample and shows that starbursts provide about 10% of the bolometric luminosity of the entire local universe. Using the luminosities of the H recombination lines to estimate the star-formation rate (as described above), it appears that roughly 25% of all the massive stars in the local universe are formed in starbursts, while the rest are formed in the many-kpc-sized disks of normal galaxies like our own. While the measurements are difficult, and their interpretation is uncertain, a similar situation seems to hold out to a REDSHIFT of about 1 (over half the way back to the big bang).

At still higher redshifts, it becomes almost impossible to measure the amount of star formation in the large-scale disks of normal galaxies. The great distances and redshift-dimming of the light mean that only intense and luminous regions of star formation can be readily detected. The detected objects appear rather similar to local starbursts (see below), and by themselves can plausibly account for much of the early star formation in the universe. Thus, starbursts are indeed highly significant components of the universe, past and present.

Starbursts as analogs to high-redshift galaxies

As described above, starbursts in the local universe are often selected by either their ultraviolet or their far-infrared continuum emission or by visible-band emission lines. At high redshifts ($z > 2$) the rest-frame ultraviolet, visible, and far-infrared emission from a star-forming galaxy will be observed in the visible near-infrared, and submillimeter spectral regions respectively (see GALAXIES AT HIGH REDSHIFT). With the Hubble Space Telescope and modern 10 m class ground-based telescopes operating in the visible and near infrared, and rapid advances in submillimeter astronomy, it is now possible to detect and study such high-redshift galaxies.

The ultraviolet-selected galaxies at high redshift strongly resemble similarly selected local starbursts: they have similar values for Σ_{SFR}, similar ultraviolet colors (suggesting a similar amount of reddening due to dust absorption), and their ultraviolet and visible spectra show that they have similar stellar populations and gas dynamics. One important difference is that the regions of star formation in the high-redshift galaxies are typically larger and more luminous than in ultraviolet-selected local starbursts. Depending on the uncertain corrections for dust extinction, the ultraviolet luminosities of the most powerful high-redshift galaxies imply star-formation rates that can reach several hundred $M_\odot$ per year over a region a few kpc in size. To date, less is known about the nature of the submillimeter-selected galaxies at high redshift. The available information suggests that these objects resemble the local ultraluminous galaxies described above.

It appears that the physical processes that we can study in considerable detail in local starbursts are directly applicable to high-redshift galaxies.

Galactic 'superwinds'

One of the most important processes that has been observed in both local starbursts and star-forming galaxies at high redshift is the bulk outflow of warm and hot gas at velocities close to or even exceeding the escape velocity from the galactic gravitational potential well (a phenomenon sometimes called a 'superwind'). Superwinds are driven by the collective effect of the kinetic energy that is deposited in the interstellar medium by stellar winds and supernova explosions. It is believed that this kinetic energy is converted (via shocks) into thermal energy inside the starburst. The resulting hot gas has a pressure much greater than its surroundings, and so it will expand most rapidly along the direction of the steepest pressure gradient in the interstellar medium (e.g. along the minor axis of the galaxy's gas disk). This leads to a poorly collimated bipolar outflow.

In local starbursts, there are a variety of probes of superwind physics. The hot outflowing gas produces thermal x-ray emission, and spectroscopy of this gas implies temperatures of 3–10 million K. This hot gas can

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

be traced out to radii of 10–30 kpc from the starburst (well out into the halo of the galaxy). The inferred outflow speeds for the hot gas are of order 10^3 km s^{-1}. Warm gas mixed into the outflow ($T \sim 10^4$ K) produces recombination and collisionally excited emission lines in the visible, and the measured Doppler shifts imply outflow speeds ranging from 10^2 to 10^3 km s^{-1}. Spectroscopy in the ultraviolet probes gas seen in absorption against the starburst. The measured blueshifts provide direct evidence for the outflow of warm and cool gas at velocities of 10^2–10^3 km s^{-1}. The ultraviolet signature of superwinds is by far the easiest one to observe in high-redshift galaxies, and it shows that outflows were common in the most actively star-forming galaxies in the early universe.

The data on both local and high-redshift superwinds establish that the outflowing material is enriched in the heavy elements (metals) that are the nucleosynthetic byproduct of massive stars (see NUCLEOSYNTHESIS). Where do these metals go? In CLUSTERS OF GALAXIES today, the amount of metals in the intergalactic medium exceeds that contained in all the stars in all the cluster's galaxies. Observations of the intergalactic medium at high redshift show that it contains at least trace amounts of metals. Thus, superwinds operating over the history of the universe are probably responsible for 'polluting' the intergalactic medium with metals.

T Heckman

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stardust

Fourth mission in NASA's Discovery programme, launched in February 1999. The first US mission to a comet and the first mission to bring back material from outside the Earth–Moon system. It will encounter Comet Wild-2 on 2 January 2004. The main objective is to capture dust particles from the comet in aerogel and return them to Earth. The sample return capsule is due to parachute onto the US military's Utah Test and Training Range near Salt Lake City on 15 January 2006.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Steady-State Theory

A theory which postulates that the large-scale appearance of the universe does not change with time. The theory, devised in 1948 by Hermann Bondi, Thomas Gold and Fred Hoyle, satisfies the so-called 'perfect cosmological principle' that the universe is the same everywhere at all times. This theory avoided the necessity for a 'Big Bang' origin to the universe, but required the continuous creation of matter in order to maintain a constant uniform density in the universe at all times (new galaxies formed as the older ones moved apart, thereby maintaining the same average number of galaxies in each large volume of space). The discovery, in 1965, of the cosmic microwave background radiation, which could not naturally be explained by the steady-state theory, led to the rejection of the theory.

See also: cosmological model, cosmological principle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stebbins, Joel (1878–1966)

American photometrist, became director of the University of Illinois Observatory and the University of Wisconsin's Washburn Observatory. Pioneered the use of photoelectric detectors to replace photographic photometry. With ALBERT WHITFORD, investigated eclipsing binary stars, the reddening of starlight by interstellar dust, colors of galaxies, pulsating variable stars and the spectrum of the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar N-body Systems: Dissipation

The motions of the stars are usually assumed to be perfectly frictionless, whereas in everyday life objects slow down by the action of friction, and heat is generated. Large collections of stars do, however, behave in ways which are reminiscent of this more familiar behavior.

Stellar N-body systems (e.g. STAR CLUSTERS, GALAXIES, MULTIPLE STARS) can be treated to a high degree of approximation as collections of point masses[1], moving according to Newton's Laws[2] of motion and gravitation. In this approximation they conserve energy[3]: if the stars fall together (figure 1) then they lose potential energy, but gain an equal amount of kinetic energy. In familiar situations such behavior is rare: a falling tennis ball also converts potential energy to kinetic energy as it descends, but it loses energy at each bounce, and by friction with the air. Eventually it loses all its energy and comes to rest. 'Dissipation' is a term applied to any process by which a system (like the bouncing ball) loses energy. By contrast, stellar N-body systems are *dissipationless* to a very good approximation. The stars do not come to rest.

Though the energy of a bouncing ball is dissipated it does not disappear. It is lost in the sound energy produced when the ball hits the ground, and in the thermal energy produced by distortions in its shape as it bounces. Thus the energy is simply converted from one form to another.

In much the same way the energy of a stellar N-body system can decrease if stars escape. This is analogous to the sound waves generated by a bouncing ball (though sound waves, unlike escaping stars, carry no mass).

Similarly the energy of a stellar N-body system can be transformed in a manner analogous to the generation of heat. To understand this, imagine a system in which all stars are initially at rest. They start to fall towards each other, like a dropped ball falling to the ground (figure 1). They pass beyond the point at which they are all close together, and then the system expands again. By a process referred to as *violent relaxation* (see below) the subsequent recollapses and re-expansions become weaker, like the weakening bounces of a ball. The system appears

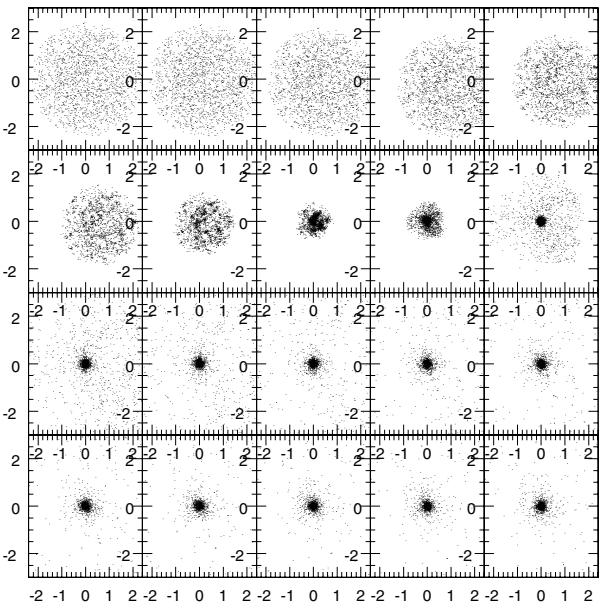


Figure 1. Collapse of a stellar N-body system. Initially the particles are at rest and distributed uniformly inside a sphere. The frames are read from left to right within each row, and the rows are read from top to bottom.

to settle down, but close examination would show that the individual stars are now moving in rather random directions within it: it is in *dynamic equilibrium*. What has happened is that the kinetic energy of the original contraction, which is an ordered motion like that of a falling ball, has been converted into kinetic energy of random motions, like the thermal motion of atoms and molecules. Indeed the collapse in the first few frames of figure 1 is called *cold collapse*, and the final configuration is referred to as a *hot system*.

On a much longer time scale the further evolution of a hot stellar system is determined by the way it cools down. Like ordinary objects, stellar systems become denser as they lose heat, but paradoxically their temperature *increases*! This intensifies the flow of heat, and leads to an accelerating collapse of the central part of the system. This process is known as 'core collapse'. It is eventually halted when the stars are so crowded together that triplets of stars can interact energetically enough. Then the system reaches a sort of equilibrium, but is gradually losing stars. Eventually the remnants of the system are indistinguishable from their surroundings, and the life of the system has come to an end.

Violent relaxation

First we consider in detail why the oscillations in figure 1 appear to damp out. Initially the stars are randomly distributed inside a sphere, and, by Newton's Theorem[4],

[1] *Point mass*: in classical dynamics, a point in space endowed with a mass, and moving according to Newtonian dynamics. It is a good approximation for a body if its dimensions are much smaller than the distances between the bodies, as in a star cluster.

[2] *Newton's Laws*: the principal *Law of Motion* (the second of the three) states that $f = ma$, where f is the force on a body of mass m, and a is the resulting acceleration; the *Law of Gravitation* states that each body exerts on every other body an attractive force of magnitude GMm/r^2 along the line between them, where m, M are their masses, r is their distance apart, and G is the universal constant of gravitation.

[3] *Energy*: in classical gravitation, the sum of the *potential* and *kinetic* energies. The potential energy depends only on the position and mass of the bodies (if treated as point masses), and increases as their separation increases. In gravitating systems it is usually taken to be negative, increasing to zero as the bodies move infinitely far apart. The kinetic energy depends only on their speeds and masses, and increases with increasing speed.

[4] *Newton's Theorem*: a uniform sphere of matter attracts an outside particle as if the entire mass of the sphere were concentrated at its centre.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on average the gravitational force a star experiences is directed towards the center of the sphere. In this approximation, all stars would arrive at the center at the same time. If the initial density of the stars (i.e. their mass per unit volume) is ρ, then this time is the 'free-fall' time

$$t_{\rm ff} = \sqrt{\frac{3\pi}{32G\rho}}.$$

In fact each star is particularly affected by its neighbors and by fluctuations in the distribution of stars around it. Consequently, some stars arrive late at the center, some early, and most stars pass on one side or other of the exact center. The result is that the system reaches a *finite* maximum density at the end of the collapse, from which the stars re-emerge in a phase of expansion.

The fact that stars do not quite arrive at the center at the same time is partly an effect of *phase mixing*. Roughly speaking, we say that motions are 'in phase' if they all arrive at one configuration at one time, and 'phase mixed' if they arrive at various times.

Another factor which helps to explain the gradual emergence of random motions is an *instability* of cold systems like the initial conditions in figure 1. This results in an increasing clumpiness in the distribution of the stars, as can be seen in the seventh and eighth frames in figure 1. The instability is analogous to the Jeans instability of cold self-gravitating gases. The effect of these clumps is that large groups of neighbors can influence each other *en masse* (or 'collectively').

Such processes continue to affect the motions of stars during the first re-expansion and subsequent collapses, though these are difficult to see in figure 1. Eventually the oscillations die away, and all that is left are the random motions of the stars. This decay of ordered motions is a rather common feature of stellar systems, and the various processes which give rise to it are referred to as 'collisionless relaxation' or 'violent relaxation'.

These words require some explanation. A process is called 'collisionless' if it is dominated by collective interactions, in contrast with 'two-body' relaxation, described below. The word 'relaxation' itself is applied to many processes which lead to equilibrium. In this case it is a 'dynamic' equilibrium, in which the stars are all in motion, but the overall distribution of density is nearly static. The word 'violent' is commonly used because it captures the time dependence of the overall distribution of matter (cf figure 1). The time scale on which violent relaxation acts is given in order of magnitude by the free-fall time. The hot system to which it gives rise is approximately in virial equilibrium[5], and so the process is often also referred to as 'virialization'.

[5] *Virial equilibrium*: the potential energy, W, and the kinetic energy, T, of a system of gravitating bodies in dynamic equilibrium are related approximately by $2T + W = 0$.

Escape

Consider a stellar system similar to that in the last frame of figure 1. The density of stars is nearly stationary and can be described approximately by the stellar density ρ, which will depend on position but not on time. By Poisson's equation, i.e.

$$\nabla^2\phi = 4\pi G\rho$$

where G is the universal constant of gravitation, the mass generates a potential ϕ, in which the stars move. In this context the energy of a star of mass m is

$$E = \frac{mv^2}{2} + m\phi$$

where v is its speed. (The two terms are the kinetic and potential energies, respectively.) Since E is constant and ϕ may be defined so that it vanishes at large distances, a star may escape 'to infinity' provided that $E > 0$, i.e. $v > \sqrt{-2\phi}$. The quantity on the right is the *escape speed*.

In the stellar N-body systems in nature, escape is complicated by the gravitational attraction of stars outside the stellar system. At some finite distance from a stellar system their gravity begins to dominate the motion of a receding star. Therefore escape is possible at slightly lower energies, and in consequence is easier. In addition, the force exerted by outside matter may vary with time, as when a star cluster passes by the dense central part of its parent galaxy. Then the energy of the star is no longer constant, and may change sufficiently to lead to escape.

Two-body relaxation

If all stars in a system have energy $E < 0$, it does not follow that escape will never happen (even if we ignore stars outside the system). Though we stated that E is constant, this was true only because we chose to describe the stellar system in terms of a steady density ρ. In reality, when two stars come close together this approximation is not valid, and at such times the energy of the two participants will change. Even if both stars initially had insufficient energy to escape, one or other of them may gain enough energy in the encounter to escape subsequently.

This process, by which the energy of a star may change because of encounters, is called *two-body relaxation* or *collisional relaxation*, and its role is not restricted to escape (see the section on 'Collapse' below). Though the description 'two-body' is self-explanatory, the other terms require some explanation. As before, 'relaxation' refers to a process which leads to a steady state (cf 'violent relaxation', above). Here the steady state refers to the particular ('Maxwellian') distribution of velocities $\boldsymbol{v}$ to which two-body encounters would give rise in the absence of escape, i.e. a distribution function

$$f(\boldsymbol{v}) = \frac{n}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{v^2}{2\sigma^2}\right)$$

where n is the number of stars per unit volume, and σ is the root mean square value of one velocity component. This

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

corresponds to a state of thermal equilibrium (a steady state in which all parts of a body or system have the same temperature) in gases, where similar effects occur. In the classical theory of gases, however, the relaxation is caused by actual collisions between the atoms or molecules. Though the term 'collisional' is also used in the theory of stellar systems, it is misleading, as the interaction of a pair of stars is purely gravitational. Real collisions do take place in extreme circumstances (see below), but have no role in relaxation.

The time scale on which relaxation acts is called the *relaxation time*, and varies according to the density ρ, and other factors. For a stellar system of the type we are considering, a convenient measure is the so-called *half-mass relaxation time*

$$t_{\rm rh} = 0.138 \frac{N^{1/2} r_{\rm h}^{3/2}}{m^{1/2} G^{1/2} \ln \Lambda}$$

where N is the number of stars, m is the mass of a single star, Λ is roughly equal to N ($t_{\rm rh}$ is rather insensitive to the exact value) and the *half-mass radius* $r_{\rm h}$ is the radius of an imaginary sphere centered at the center of the cluster, and enclosing half of its total mass. Stars escape from a stellar system on a roughly comparable time scale. Another process acting on this time scale is called *dynamical friction*, because rapidly moving stars tend to lose kinetic energy to the stars which they encounter.

This formula applies to the idealized case of a stellar system in which all stars have the same mass. In more realistic cases, the effects of relaxation depend on the stellar mass. In particular, massive stars suffer greater dynamical friction than stars of low mass. One result of this is that the rate of escape of stars decreases with increasing mass. Another result is *mass segregation* (see below).

Collapse

As we have seen, a star may escape from a stellar system following a close encounter with another star. This dynamical process is dissipationless: the total energy of all stars is constant. The escaping star removes some energy, however, and so the energy of the remaining stars decreases. In this section we consider how the stellar system evolves as it loses mass and its energy decreases.

Because the system is nearly in virial equilibrium its kinetic energy, T, is related to its total energy, E, by

$$T = -E.$$

As E decreases, following escape, T necessarily increases. As this kinetic energy is also shared among fewer stars, it follows that the velocities of the stars increase on average. It is not difficult to see why this happens. Consider a planet in circular orbit about the Sun, and suppose we decrease its energy by slowing it down slightly. It now falls towards the Sun somewhat, but in doing so it picks up speed. In fact it now moves, on average, faster than it did before we slowed it down. For the same reason, though an Earth satellite suffers from friction if it dips into the top of the Earth's atmosphere, it actually speeds up as it descends.

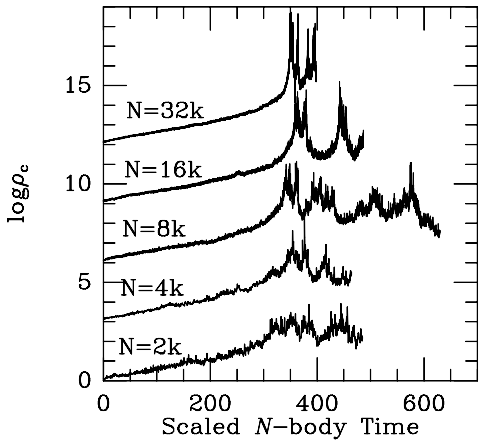


Figure 2. Core collapse in several computer simulations. The central density is plotted against time, scaled in proportion to $t_{\rm rh}$. The first 'spike' marks the end of core collapse and the start of a phase of gravothermal oscillations. (After Makino J 1996 *Astrophys. J.* **471** 796–803. Reproduced by permission of the American Astronomical Society.)

This remarkable property of gravitating systems is sometimes described by saying that they have 'negative specific heat'. The 'specific heat' of a substance describes the effect on its temperature of the addition of some energy, and for familiar substances like water or air it is positive. This increase in temperature corresponds to an increase in the speeds at which the atoms in the substance move. By contrast we have seen that gravitating systems respond to a *decrease* in energy with an increase in the speed!

This removal of energy by an escaping particle not only causes the remaining stars to speed up; just as with the planet which moves closer to the Sun after being slowed down, the stars in the system move closer together. Thus the effect of escaping stars is to make those that remain crowd into a smaller space. Indeed a similar process occurs even if stars do not escape. Relaxation causes stars to leave the dense central parts of a stellar system on the relaxation time scale, and, as a result, the central parts contract and become denser (figure 2), and the motions there become faster. As the central part of a stellar system is called the 'core', this process is called 'core collapse'. The time of relaxation in the core decreases as the core collapses, and the collapse accelerates as it proceeds. Even so it is a very gentle and slow collapse compared with that shown in figure 1, because the ratio of time scales $t_{\rm rh}/t_{\rm ff}$ is of order $N/\log N$, which is large for a rich stellar system. The process is faster if stars of different mass are present; and by a process known as *mass segregation*, which is an effect of two-body relaxation, the more massive stars settle into the core more rapidly

The resulting 'runaway' is sometimes referred to as the 'gravothermal catastrophe', but it is not very catastrophic: the rest of the system hardly notices what

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is taking place in the tiny core. The word 'gravothermal' means that the collapse is governed by the way in which self-gravitating systems respond to loss or gain of heat.

Dense stellar systems

When the core reaches a high enough density, other processes intervene and halt the collapse. These usually involve three- and four-body interactions, either between a BINARY STAR and a single star, or between two binaries. If the system initially contained many binaries, they would settle to the core more rapidly than other stars, by mass segregation. Even if there were no binaries, interactions involving three single stars would create binaries (one of the stars being needed as a catalyst to carry off some energy).

From a gravothermal point of view the importance of these reactions is that they are a source of heat (see the article on the THREE-BODY PROBLEM), and this eventually balances the flow of heat by two-body relaxation into the outer parts of the system: this is how core collapse is halted. Henceforth the system retains a dense core, though in certain situations the balance in the flow of energy is achieved only sporadically, and the core may exhibit large-scale oscillations (figure 2). Like core collapse itself, they are gravothermal in nature.

At these high stellar densities in the core, and in the presence of binary stars, one can no longer entirely treat the stars as point masses. Real stellar collisions can, and indeed must, occur in large numbers in many dense stellar systems in nature. These processes are genuinely dissipative (though the loss of energy is not thought to be important for the energy balance of the system as a whole). Their importance is that they can modify the kinds of stars seen in dense stellar systems, and affect the efficiency of the binary interactions which are needed to maintain the system against further core collapse.

After core collapse the evolution of the system is determined by the loss of mass, through escaping stars. In realistic situations, all the mass is lost after a few tens of relaxation times.

Bibliography

Makino J 1996 *Astrophys. J.* **471** 796–803

Standard treatments on the topics of this article are:

Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)

Spitzer L Jr 1987 *Dynamical Evolution of Globular Clusters* (Princeton, NJ: Princeton University Press)

Recent advances are reviewed in:

Merritt D 1999 Elliptical galaxy dynamics *Publ. Astron. Soc. Pacific* **111** 129–68

Meylan G and Heggie D C 1997 Internal dynamics of globular clusters *Astron. Astrophys. Rev.* **8** 1–143

Douglas C Heggie

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Activity

Energy deposited in the tenuous gas above the visible surface (i.e. photospheres) of stars leads to a range of phenomena collectively referred to as stellar activity. The energy deposition heats the gas producing chromospheres, at temperatures $\sim 10^4$ K, and coronae with temperatures from a few to tens $\times 10^6$ K. The intermediate temperature region is generally called the transition region. These parts of the STELLAR ATMOSPHERE are particularly prominent in the x-ray and ultraviolet spectral regions where emission from the stellar photosphere is weak. Stellar activity is seen for a wide range of types of star, and occurs on dwarf stars with spectral types from F through M. Stellar activity is directly related to the presence of subphotospheric convection and the generation of a magnetic field by dynamo processes(see DYNAMOS: SOLAR AND STELLAR). The stellar rotation rate is also a key controlling parameter (see STELLAR ROTATION ON THE MAIN SEQUENCE). The heating processes are primarily magnetic, with acoustic energy deposition playing a minor role. The strength of the magnetic fields in the photosphere can be large (several kilogauss) and this field is associated with large starspots and the occurrence of large flares (see SOLAR–STELLAR CONNECTION). For the most active stars the energy radiated from chromospheres/coronae and liberated by flares can be at least a thousand times larger than seen in the solar atmosphere. Stellar activity is particularly intense in CLOSE BINARY STARS where rapid rotation is imposed by tidal interaction. Evolved F, G, K and M stars all possess chromospheres but coronae become weaker or absent for the cooler giants and supergiants and stellar mass loss becomes much more significant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Astronomy (to the Rise of Astrophysics in the Mid-Nineteenth Century)

Variable stars

According to ARISTOTLE, whose world picture dominated Greek, Islamic and Western Latin civilizations until the seventeenth century, the celestial region of the COSMOS—the Moon and beyond—is unchanging except for cyclic motions that return the celestial bodies to their original positions. Changes in the brightnesses of stars were therefore deemed impossible, and so VARIABLE STARS went unnoticed. Even reports of SUPERNOVAE are excessively rare, until a bright star-like object appeared in Cassiopeia in 1572. Many observers considered it atmospheric, but the young TYCHO BRAHE was able to show convincingly that it was celestial. The possibility of change among the stars was established beyond doubt by the appearance of a second (super)nova in 1604, in Ophiucus.

Meanwhile the Frisian astronomer DAVID FABRICIUS claimed that a less dramatic new star had appeared in Cetus in 1596. Another nova appeared and disappeared in CETUS in 1638; but this time the star reappeared, and in due course was identified with Fabricius's nova. In 1667 ISMAEL BOULLIAU showed that although Mira Ceti ('the wonderful star in the Whale') was not always at the same brightness when at maximum, its maxima did occur at regular intervals of 11 months, and to this extent the variable was predictable. Boulliau also offered a physical explanation, by analogy with sunspots: the star had dark patches and its light diminished cyclically when these patches were presented to the observer as the star rotated; but just as sunspots varied, so the dark patches varied, and this resulted in irregularities in the light curve of the star.

Claims to the discovery of variables were by now commonplace. But such claims were almost impossible to verify or refute, and so interest in variable stars lapsed, until late in the eighteenth century. Beginning in 1796, William Herschel provided a means of testing whether a star was varying, by publishing extensive 'catalogues of the comparative brightness of the stars', in which each star was compared with others of almost the same brightness; any change would then disturb the published comparisons. By using much simpler sequences of stars arranged by brightness, two young English amateur observers, Edward Pigott and John Goodricke, had already detected four stars whose periods were measured in days. The first of these, Beta Persei (ALGOL), had long been reputed to vary, and after monitoring the stars for some weeks, on 28 December 1782 the two friends saw it brighten from third (or fourth) magnitude to second, before their very eyes. Pigott suspected at once that Algol was being periodically eclipsed by a planet; but eventually the friends abandoned this explanation, perhaps because of apparent irregularities that in fact were caused by changes in viewing conditions, but perhaps because no such explanation was possible for the other variables they discovered, two of which were Cepheids. Thereafter study of variables made little progress until the development of astrophysics.

Proper motions and the solar apex

By the later seventeenth century RENÉ DESCARTES had persuaded most astronomers that the stars were simply distant suns, and Isaac NEWTON concluded that stars were isolated bodies free to move in seemingly empty space. But it was not until 1718 that Edmond Halley, who was studying ancient star catalogs for a new determination of precession, was able to announce that three bright stars were no longer in the places they had occupied in antiquity. However, it would be some decades before proper motions were determined in any quantity: John Flamsteed's great star catalog, published in 1725, had been compiled just before JAMES BRADLEY's discovery of aberration and nutation (see below), and so accurate past positions of stars were unavailable in the mid eighteenth century.

Bradley himself pointed out that such motions could arise either from the motion of the solar system or from movements of the stars themselves, and Tobias Mayer in 1760 argued that *patterns* of proper motions should be ascribed to the former and *deviations from patterns* to the latter. The first to detect such a pattern was William Herschel (see HERSCHEL FAMILY). Working with far from reliable data, he proposed in 1783 that the solar system was moving towards Lambda Herculis, close to modern determinations. For reasons as yet unexplained, F W Bessell, after reducing Bradley's star observations for his *Fundamenta astronomiae* (1820), was unable to find any such pattern. But in 1837 F W A Argelander published the results of an investigation into no fewer than 390 proper motions, a number large enough for him to divide them by size into three groups which he treated independently; each yielded an apex not far from that proposed by Herschel. The matter was put beyond dispute in 1847 when Thomas Galloway analysed proper motions of southern stars not previously taken into account, and derived a similar apex.

Distances of stars

In the cosmologies of Aristotle and Ptolemy, the stars were located immediately beyond the outermost planet. When Copernicus in 1543 argued that the Earth orbits the Sun and that we therefore observe the stars every six months from opposite ends of a diameter of the Earth's orbit, opponents demanded to know why no annual parallaxes had been detected. By the later seventeenth century, after Kepler had taught astronomers to see the massive Sun as the centre of the planetary system, annual parallax had developed into the hoped-for means of determining the distances of stars. But if Descartes was right and the stars were physically the equal of the Sun, their distances must be vast, and the detection of annual parallax with instruments subject to shrinking and warping with the passage of the seasons was a daunting task. However, if the brightness of a star (such as SIRIUS) could be compared

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

with the brightness of the Sun, then since light diminishes with the square of the distance this ratio could be converted to a ratio of distances. The best-known attempt at this comparison was by CHRISTIAAN HUYGENS, in his posthumous COSMOTHEOROS (1698). Huygens viewed the Sun through a pinhole made in a screen, hoping that the fraction of the Sun that was thereby visible would be equal in brightness to Sirius. The technique was clumsy, but his conclusion, that Sirius lay at 27 664 astronomical units, showed that the scale of interstellar distances was vast.

A much better method was advanced by JAMES GREGORY in 1668, in a little-noticed work on geometry. He proposed replacing Sirius by a suitable planet, at a time when the planet was of the same brightness as Sirius; the required ratio was then equal to the ratio of the light received directly from the Sun, to the light received from the Sun via the planet. This calculation had the merit of involving only quantities related to the solar system. Even when working with values that he knew to be obsolete, Gregory placed Sirius at 83 190 AU; the true distance would be much greater, in fact about 10^6 AU according to Newton when drafting *The System of the World* around 1685. But this work appeared only in 1728, after Newton's death, and until then his estimate was known only to his intimates.

Meanwhile a remarkable attempt to measure annual parallax was made by ROBERT HOOKE. The known hazards in the way of such measurements were the possible warping of the instrument with the change of the seasons, and the uncertain effect of refraction. The latter problem Hooke sought to overcome by selecting as his target GAMMA DRACONIS, a star that passed directly overhead his home in London; the former, by incorporating the tube carrying the object glass into the fabric of the roof of his house. The resulting zenith telescope was designed to resolve one single problem (stellar parallax), by observing one single star but only when the star was specially positioned: a remarkably mature use of a scientific instrument. Unfortunately, Hooke made only four observations before the object glass broke, and his claims to have detected parallax were discounted by other astronomers. But his method was promising, and it was revived in 1725 by an English amateur, Samuel Molyneux, who commissioned a zenith telescope from George Graham. His collaborator, James Bradley, found the movements of Gamma Draconis were three months out of phase from those expected from annual parallax. Commissioning a second instrument with a wider field of view, Bradley was able to establish the pattern of movements displayed by a number of stars. Later, the explanation came to him: the speed of light, though great, is finite, and the apparent position of a star is affected by the velocity of the observer on Earth.

The discovery of 'the aberration of light' was proof of the motion of the Earth around the Sun. It also revealed an unexpected error in all existing star catalogs, including Flamsteed's. And because Bradley had failed to detect annual parallax, he could estimate from the accuracy of his measures that the stars must be at least 400 000 AU from the Sun. His publication in 1729 of this *minimum* distance of stars was fully compatible with Newton's estimate of the actual distance of Sirius, based on the hypothesis that Sirius and the Sun were physically equal; and the appearance of these two results within a few months of each other convinced astronomers that light even from the nearest stars took several years to arrive on Earth.

Not surprisingly, it was decades before attempts to measure such tiny annual parallaxes were resumed, and even then successive claims proved to be mistaken. It was essential to single out for measurement the stars that were nearest to Earth, and an important advance came in 1837 when WILHELM STRUVE at Dorpat published an assessment of the relevant criteria. By now proper motions were available in quantity, and there was evidence—for example from binary stars—that apparent brightness was an unreliable guide to distance. Stars with large proper motions could more confidently be taken to be near, and Giuseppe Piazzi and BESSEL had both drawn attention to 61 CYGNI, 'the flying star'. Also, the craftsmanship of Joseph Fraunhofer had equipped both Bessel at Königsberg and Struve at Dorpat with precision instruments of a quality hitherto unavailable.

In 1835 Struve selected VEGA for study. The star was bright and had a large proper motion, and in 1837 he deduced from 17 observations a parallax of one-eighth of a second of arc. Encouraged by Struve's apparent success, Bessel subjected 61 Cygni to a year's scrutiny, observing the star many times on the same night, and at the end of 1838 he announced a parallax of about one-third of a second of arc. Bessel's many observations closely matched the pattern predicted from theory, and it was this that convinced astronomers that a genuine measure of parallax had at last been achieved.

Double stars

Struve and Bessel had used distant stars as reference points against which they measured the motions of the stars under examination, and in doing this they had been applying a method proposed by GALILEO in 1632 to facilitate the measurement of annual parallax. Galileo understood very well the immense practical difficulties involved in tracking tiny stellar movements spread over many months; however, if the star under investigation chanced to lie in almost the same direction as another and much more distant star (so that the two formed a 'DOUBLE STAR'), the distant star could act as a quasi-fixed reference point provided by a helpful Nature.

Characteristically, Galileo himself did nothing to follow up this insight, nor did any observers prior to William Herschel. In the late 1770s, Herschel was systematically familiarizing himself with the brighter stars, and he took the opportunity to watch out for double stars. In 1782 and 1785 he published two catalogs of doubles totalling 703 specimens.

However, unknown to Herschel, John Mitchell had published a paper in 1767 in which he pointed out that the number of double stars was far greater than would be expected if the doubles occurred merely by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

chance: most of them must be neighbors in space, lying at the same distance from Earth and therefore useless for the measurement of parallax. And when Herschel re-examined some of his doubles after an interval of two decades, he found several instances in which the two stars had indeed orbited around each other. But was the force that bound them together in fact gravitational attraction, as everyone assumed? To decide this, more details were needed of the orbits. At length, in 1827, Félix Savary was able to confirm that the two stars of Xi Ursae Majoris did indeed move in elliptical orbits about their common centre of gravity, as Newtonian theory required. Newton himself had assumed that gravitational attraction was a universal force; now there was at least proof that it operated among the stars.

The system of the stars and the darkness of the night sky

Despite these various developments in the theory of certain individual stars, until the rise of astrophysics in the later nineteenth century the stars as a whole remained little more than a backcloth to the motions of the bodies of the solar system. Even Newton in the *Principia* displays almost no interest in the stars, and this despite his claim that gravity was a universal force: for forces cause movements, and if each star was being attracted by every other star then it was far from obvious how every single star continued (as it seemed) to rest motionless in the position recorded for it in the star catalogs of antiquity. Although Newton had a better understanding of the immensity of interstellar distances than any other living person, it did not occur to him that the stars might indeed be in motion, but that because of their remoteness their motions had not yet been detected. Perhaps he was mesmerized by his own use of the traditional Latin *stella fixa* as the term for 'star'. At all events, when challenged to expound his cosmology he attempted to reconcile the immobility of the stars with the universality of gravity by arguing that the system of the stars was infinite and almost symmetric. This being so, each star was being pulled almost equally in every direction by the other stars, and therefore remained (almost) at rest.

However, the symmetry (which he defended by comparing the numbers of stars predicted from a theoretical model with the numbers found in star catalogs) is clearly not perfect, and Newton accepted that in the long term the stellar system would be threatened with gravitational collapse. How could this be reconciled with the accepted image of God as the great clockmaker? Descartes had portrayed a God who created the universe, set it in motion, and then abandoned it to run itself. Newton, by contrast, believed that God must display a continuing concern for the welfare of his creation. He had revealed himself not only through the Book of Scripture but also through the Book of Nature. The scientist investigating the star system and uncovering its near symmetry (and therefore medium-term stability) learned from the Book of Nature the concern for mankind that God had displayed at the creation. And God displayed an on-going concern for mankind by intervening whenever gravitational collapse threatened, and restoring the original order. G W Leibniz condemned this conception as portraying God as a bungling clockmaker who resorted to miracles to repair bad workmanship; Newtonians replied that these divine interventions were not miracles, but regular acts foreseen from the beginning. Both parties agreed that the structure of the stellar universe was essentially permanent and would show no development with time.

About 1720 a young physician, William Stukeley, in conversation with Newton, diverted the discussion away from the consequences of gravity and towards consideration of the cumulative light of infinitely many stars. If the star system was infinite, he argued, the whole sky would have 'the appearance of that luminous gloom of the milky way'. In 1721 EDMOND HALLEY breakfasted with Stukeley and Newton and discussed astronomical topics, a few days before Halley read two short papers on cosmology to the Royal Society in which he remarked: 'Another Argument I have heard urged, that if the number of Fixt Stars were more than finite, the whole superficies of their apparent Sphere would be luminous.' Halley offered a defective defence of the infinity of the stars, but the publication of his papers brought the Newtonian universe—anonymously—into the public domain.

In 1744 the Swiss astronomer J-P L de Chéseaux showed that even if only a tiny fraction of starlight was lost in traveling the distance from one star to the next, the cumulative effect of these losses would be decisively to reduce the total amount of light received from the innumerable stars at great distances. That is, the darkness of the night sky was no obstacle to belief in the infinity of the stars. Much the same argument was advanced in 1823 by H W M Olbers, and it is his name that has become attached to the so-called 'OLBERS' PARADOX', though nineteenth-century astronomers found no difficulty in offering a variety of explanations for the darkness of the night sky.

The Milky Way

The MILKY WAY GALAXY was, one might think, a decisive disproof of Newton's claim that the system of the stars is symmetric, but his lack of interest in this phenomenon was typical of his generation. Serious astronomers were preoccupied with the solar system, and it was only after a group of speculators in the mid eighteenth century proposed explanations of the Milky Way that the problem of the structure of the stellar system began to enter mainstream astronomy.

The first of these speculators was THOMAS WRIGHT of Durham, a self-taught itinerant lecturer. In 1734 he prepared a public lecture on cosmology in which the stars were portrayed as occupying a spherical shell of space that surrounded the Divine Centre, the region beyond the shell forming the Outer Darkness. As the stars could not be permitted to fall into the Divine Centre, each was in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(circular) orbit within the shell. Wright illustrated his lecture with a representation of a cross section of the universe, in which he also showed how his stellar system would appear from Earth, which of course was a planet of one of the stars within the shell. The light of the numerous stars within the cross section that were distant from Earth but still visible to humans would, he argued, merge to give the effect of a milky circle in the sky. He thought he had thereby explained the Milky Way, until he realized that the plane of the Milky Way is unique whereas the plane of his cross section was arbitrary.

Having considered the matter further, in 1750 Wright published *An Original Theory or New Hypothesis of the Universe*, with two alternative explanations of the Milky Way. In the preferred alternative the Sun was, as before, a member of a spherical shell of stars that surrounded what was now one of many Divine Centres. But now the shell was thin, and of vast diameter, so that an observer facing inwards or outwards would see only a few stars (which were near and therefore bright) before looking into empty space. But when facing tangentially to the shell, the observer would see great numbers of stars, whose light would merge to give the effect of the Milky Way. The plane of the Milky Way, therefore, was the tangent plane to the shell of stars at the place where the terrestrial observer was located.

An alternative explanation saw the system of stars as forming a broad, flat ring surrounding the (local) Divine Centre, in which case the *visible* stars would occupy a disk-shaped region of space internal to the ring. The observer looking in any direction within the plane of the ring would see innumerable stars whose light would merge to give a milky effect.

Wright disliked the second model because he had no explanation for the particular plane of the ring, but both models were described in a summary of his book (but without illustrations) that was published the following year in a Hamburg periodical. This came to the eyes of the young IMMANUEL KANT, who, misunderstanding Wright, thought the Divine Centre was elsewhere in the universe. This being so, he saw no reason why the ring should not extend continuously from one side to the other, so that the system of the stars would occupy a space whose shape was not that of a hollow ring but of a disk. He also knew of observations by P L M de Maupertuis, which Kant interpreted as being of other star systems that had an elliptical outline. A spherical system viewed from without will always have a circular outline, but a disk-shaped system will appear elliptical. Kant therefore adopted Wright's second model, as duly modified, and he concluded that our star system is disk-shaped.

Wright's book had limited circulation, distribution of Kant's work was undermined by the bankruptcy of the bookseller, and a work by J H Lambert, an Alsatian living on the fringes of the scientific community and who had somewhat similar ideas, also found a limited public. It may well be, therefore, that when William Herschel turned his attention to the problem of the Milky Way, he was unaware of these earlier speculations. Herschel took it for granted that the Milky Way was the optical effect of our immersion in a layer of stars, and he saw that he could establish the detailed shape of the layer if he allowed himself two assumptions. The first, obviously, was that his telescope could penetrate to the borders of the system in all directions, for unless this was so his quest was hopeless from the start. The second was that within the borders of the space occupied by stars, the system was roughly symmetric. Obviously this was by no means strictly true; Herschel hoped that it would be true enough for his purpose.

Granted these assumptions, the number of stars visible in any particular field of view of Herschel's telescope would be directly proportional to the volume of space within the Galaxy within that field of view, and a simple calculation would convert this volume into the distance to the border in that direction. Herschel did not have time enough to carry this pioneering exercise in stellar statistics to a conclusion, but in 1785 he completed the examination of a great circle of the sky and converted the resultant star counts into a cross section of the Galaxy.

The experience of later years convinced him that neither assumption was even approximately correct. His 40 ft reflector completed in 1789 revealed many stars hitherto hidden from him, and his studies of nebulae and clusters showed how very non-uniform is the distribution of the stars of the Galaxy. He therefore abandoned his cross section of the Galaxy, but as for many decades nothing comparable was advanced to fill its place, the diagram continued to appear until late in the nineteenth century.

The riddle of the nebulae

This investigation, however, was a distraction from the great campaign to which Herschel had dedicated himself, a systematic search for NEBULAE and clusters. Mention of nebulous stars goes back to antiquity, but much-needed clarity was introduced into the subject by Halley in 1715 when he published a list of six nebulae; these he believed to be formed of a 'lucid medium'. Others, however, adopted the simpler view that a nebula was a star cluster so distant that the individual stars could not be distinguished. Observations could help settle the issue in two ways. First, new and bigger telescopes might resolve into stars objects that had appeared nebulous in lesser instruments. Second, if it was shown that a nebula had changed in appearance over a few years or even decades, then the nebula could not be a distant and therefore vast star system, for the widely scattered components of such a system would not have been able to travel such great distances in so short a time. Herschel shows an awareness of this latter consideration on the very first page of his first observing book, dated March 1774. For the rest of that decade he was able to make only occasional observations of a single nebula, the ORION NEBULA, but these were enough to satisfy him that it was indeed altering shape. This being so, it must be a true nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

How then to distinguish between true nebulae and distant star clusters? Herschel decided that true nebulae revealed themselves by their smooth, milky appearance, while distant star clusters had a mottled look. But no sooner had he proposed this in a 1784 paper to the Royal Society, than he came across two nebulae in which both forms of nebulosity were intermingled. This forced him to reinterpret the evidence. He now decided that mottled nebulosity indicated stars at middle distance, while milky nebulosity indicated stars further away. All nebulae were systems of stars.

In 1783 he had embarked on a systematic search for nebulae. With his 20 ft reflector facing south, he allowed the sky to drift past as he waited for a nebula to come into view; he would then shout out details of its appearance to his sister Caroline who was seated at a desk at an open window nearby. In 20 years of observing, the team of brother and sister collected 2500 nebulae, to add to the hundred or so listed by Charles Messier in 1781. They thereby introduced into astronomy the methodology of the natural historian, for they were collecting and classifying specimens in great numbers. Furthermore, Herschel realised that since gravity was a universal force, a scattered star system would become more concentrated as time passed. This allowed him to develop a cosmogony, set out in papers published in 1785 and 1789. He envisaged the process as beginning with widely scattered stars; then, as time passed, the gravitational pull of the stars that happened to be in greater concentrations than elsewhere would draw in the surrounding stars, leading to the formation of systems of ever increasing density. Where Newton had seen in the heavens the stability of clockwork, Herschel saw development analogous to that of the life-story of a person.

In November 1790 Herschel was 'sweeping' as usual when he came across NGC 1514, which he saw as a star surrounded by a luminous halo. He had already seen several 'PLANETARY NEBULAE' as he termed them, objects that had planet-like disks but shone with the pale light of a nebula, and he was greatly puzzled by them. To modern astronomers NGC 1514 is another planetary nebula, but to Herschel it was a 'nebulous star'. Indeed, it was a star that was condensing out of the nebulosity that surrounded it. Clearly, (true) nebulosity existed after all.

Herschel thus found himself forced to alter his opinion once more. In papers published in 1811 and 1814, he extended his earlier cosmogony back in time, before stars had formed out of the diffuse nebulosity; he described the formation of concentrated nebulous clouds, out of which stars condensed, which in turn developed into more and more tightly packed clusters.

William Herschel was in his fifties when his son John was born, and in his late seventies when John embarked on a career as a Cambridge don. Eager to transmit his skills and experience while he still had strength, William prevailed upon John to return home and become in effect his apprentice. Late in 1820 William's 20 ft reflector deteriorated to the point where John undertook the construction of a replacement; and with this in 1821 he carried out two trial sweeps, with his aged aunt as his mentor. In 1825, after his father's death, he commenced a systematic re-examination of his father's nebulae. Unfortunately William's catalogs of nebulae were arranged by type and so quite unsuited to John's needs. Caroline, by now living in her native Hanover, came to the rescue, and undertook the immense labor of rearranging the catalogs into a form suitable for 'sweeping', a task for which she was awarded the Gold Medal of the Royal Astronomical Society.

John's re-examination of the northern nebulae resulted in a catalogue of 2306 nebulae and clusters published in 1833. John then set sail for the Cape of Good Hope, and in four years of intense activity extended to the southern skies all aspects of his father's researches. His *Results of Astronomical Observations Made ... at the Cape of Good Hope*, published in 1847, listed over 1700 nebulae and clusters and over 2100 double stars, as well as thousands of star counts and much else besides.

Whereas his father had left no one in any doubt as to his opinions, John was more circumspect, and sometimes inhibited by his reluctance publicly to differ from his father. Meanwhile the debate continued: were all nebulae star clusters that appeared nebulous simply because they were so distant, or were some star clusters and others truly nebulous? Because of its brightness and apparent size, the Orion Nebula became a test object: could it, or could it not, be resolved into stars under examination by a truly massive reflector?

In 1839, at his estate at Birr Castle in central Ireland, William Parsons (the future THIRD EARL OF ROSSE) succeeded in building a reflector on the Herschelian model, with a segmented mirror 3 ft in diameter; the following year, he cast a solid mirror of the same size. He persuaded himself that the instrument had resolved into stars a number of prominent nebulae, and he determined to build a still bigger instrument, with mirrors no less than 6 ft in diameter, to carry this process to a convincing conclusion. The 'Leviathan of Parsonstown' first saw light in February 1845, and within a few weeks Rosse was able to announce that the nebula M51 had a spiral structure. A year later, he claimed decisive evidence of the resolvability of the Orion Nebula, a feat that many saw as proclaiming the end of belief in 'true' nebulosity. Rosse has been criticized for seeing what he wanted to see, but in fact his great reflector was capable of seeing stars that are indeed embedded in the nebula, and it is not surprising that he interpreted this as evidence that the nebula is stellar.

There were, however, doubters, notably observers who believed they had seen changes in nebulae, changes that had happened too quickly for the nebulae to be vast stellar systems. This seemed indisputably the case when a nebula discovered in 1852 by John R Hind completely disappeared in 1861. Three years later, the new techniques of astrophysics brought the long debate to a conclusion, when William Huggins demonstrated beyond argument the existence of gaseous nebulae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Bennett J A 1976 'On the power of penetrating into space': The telescopes of William Herschel *J. Hist. Astron.* **7** 75–108

Hoskin M 1982 *Stellar Astronomy: Historical Essays* (Chalfont St Giles: Science History Publications)

Hoskin M (ed) 1997 *The Cambridge Illustrated History of Astronomy* (Cambridge: Cambridge University Press)

——1999 *The Cambridge Concise History of Astronomy* (Cambridge: Cambridge University Press) ch 7

Whitney C A 1971 *The Discovery of Our Galaxy* (New York: Knopf)

Michael Hoskin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Atmospheres

The atmosphere of a star is the boundary between the stellar interior and the interstellar medium. It spans the layers from which photons can escape to the surrounding space. Direct information about the physical state of the stellar atmosphere can therefore be derived from the emergent radiation. This is, however, not the case for deeper layers, because photons that are emitted in these layers are reabsorbed or scattered numerous times before escaping. A physical description of the stellar interior can be inferred only indirectly: the atmosphere as a boundary layer reflects the conditions in the stellar interior. The only direct probes of the stellar interior are neutrinos emitted in the core and gravitational waves, but there are considerable difficulties in detecting them (only neutrinos emitted by the Sun and SN 1987A have been observed so far). In the last decade, helioseismology has developed very rapidly, and the study of numerous oscillation modes has allowed us to probe the Sun's interior. Asteroseismology will follow in the coming years, but the very low amplitude of stellar oscillations has limited its current application to pulsating stars. Therefore, our present understanding of stellar structure and evolution hinges primarily on our ability to model stellar atmospheres and on analysing the emergent radiation.

This article describes stellar atmospheres very broadly. It discusses the general structure of stellar atmospheres, line formation, stellar atmosphere theory, opacities and non-thermal phenomena. More details can be found in the articles on STELLAR ATMOSPHERES: COOL STARS, STELLAR ATMOSPHERES: EARLY-TYPE STARS, EXTENDED ATMOSPHERES, CHROMOSPHERE, CORONAE and the SOLAR WIND.

General structure

The atmosphere might be quite extended in space. However, the bulk of the radiation is emitted in a thin layer, the *photosphere*. The solar photosphere is about 600 km thick as illustrated in figure 1. Hotter stars have thicker photospheres, but in general the extent of the photosphere corresponds to about 1/1000 of the stellar radius.

The structure above the photosphere differs markedly between cool and hot stars. Cool stars (spectral type A7 and later on the main-sequence, or $T_{\rm eff} < 7500$ K) are characterized by increasing temperature in their outer layers. Figure 1 illustrates the temperature structure of the solar atmosphere. The *chromosphere*, with typical temperatures ranging from 7000 to 20 000 K, is diagnosed from emission lines in the far-ultraviolet and emission reversal in the core of very strong optical lines (Hα, Ca II H and K). Above the chromosphere, there is a

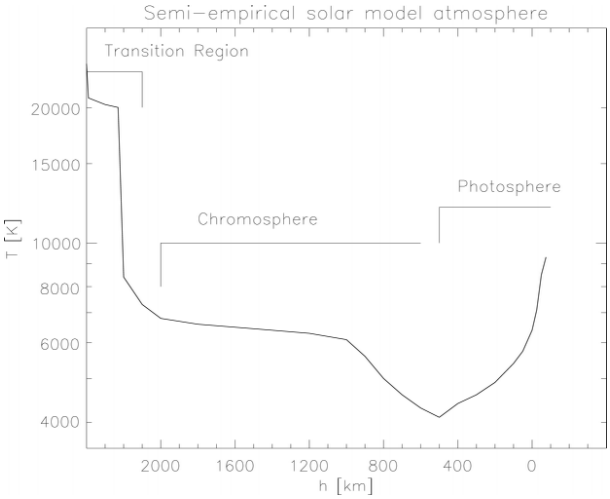


Figure 1. Semi-empirical temperature–height distribution for the average, quiet Sun (adapted from Vernazza *et al* 1973 *Astrophys. J.* **184** 605), illustrating the general atmosphere structure of late-type stars. The origin of the height scale corresponds to optical depth unity, and the solar interior is to the right.

hot, million-degree *corona* seen in solar eclipses and detected in the x-ray and extreme ultraviolet domains by various satellites. Obviously, there is also a transition region between the chromosphere and the corona with intermediate temperatures where emission lines of ionized species like He II, C IV and Si IV are formed. While the exact heating mechanism(s) of the chromosphere and corona are not yet fully understood, chromospheres and coronas have been detected only in low-mass stars having a sub-photospheric convection zone.

Strong, fast *winds* typify the atmospheres of early-type main-sequence stars (O and early B-type) and O, B and A-type supergiants. The characteristic spectral features of these winds are strong P-Cygni line profiles and an infrared excess due to free–free emission in the wind. These fast, supersonic winds are driven by radiation pressure generated by absorption of photons in spectral lines. X-ray emission has also been detected in O and early B-type stars, but this is not a signature of a corona. This emission is attributed to shock-heated material in the wind.

Winds are also present in cooler stars. In cool main-sequence stars, the high coronal temperature results in a large gas pressure inducing a weak pressure-driven stellar wind. This is a fast wind, but so tenuous that it is optically thin throughout and cannot be detected spectroscopically. In this respect, this wind is not really a part of the stellar atmosphere as defined. In evolved cool stars, a much denser wind is observed. Due to the low surface gravity of such stars, the mass loss rate of a coronal wind would be larger, but other mechanisms like radiation pressure on dust grains formed in the outer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

layers or like acceleration in pulsating layers become more efficient drivers. In contrast to the winds of cool main-sequence stars, the winds of red supergiants are slow.

Atmospheric parameters

For most stars, a radius, R, can be defined quite naturally by the distance of the photosphere from the center of the star. The total luminosity, L, radiated by the star is given by

$$L = 4\pi R^2 \int_0^\infty F_\nu \, \mathrm{d}\nu = 4\pi R^2 \sigma T_{\mathrm{eff}}^4 \tag{1}$$

where the flux integral is the total radiated power per unit area. The *effective temperature*, T_{eff}, is defined in an analogous form of the Stefan–Boltzmann law, making T_{eff} the temperature of a black body radiating the same energy per unit area as the star. The stellar spectrum may, however, differ dramatically from a black body spectrum. From this relation and the limited extension of the photosphere, it follows that the electronic temperature in the photosphere is typically of the order of the effective temperature. This temperature is the first fundamental parameter describing a stellar atmosphere.

Apart from the temperature, the emergent spectrum will reflect the density of the absorbing and emitting material in the photosphere. The total mass density in the photosphere is set by the balance between gas and radiation pressures on the one hand and gravity on the other. The gravitational acceleration, $g = GM/R^2$, is constant since the photosphere is a thin layer (M is the mass of the star). The *surface gravity* is thus the second fundamental parameter, and is usually expressed as $\log g$, in cgs units. Absorption and emission processes will also depend on the chemical composition of the atmosphere, which is the third major parameter. Additional parameters may be defined to describe an atmosphere, e.g. turbulent velocities, magnetic field strengths or wind parameters.

Geometry

Because the radial extent of the photosphere is much less than the stellar radius (i.e. the curvature is mostly negligible), most stellar photospheres can be modeled like plane-parallel slabs. The scale height is the characteristic distance on which pressure and density decrease in the slab, and thus depends on the surface gravity. Spectroscopic analyses of the photospheric spectrum with plane-parallel model atmospheres therefore yield an estimate of $\log g$, but do not provide the stellar radius directly.

There are, however, some important cases, like supergiants and Wolf–Rayet stars, for which the geometrical extension of the photosphere becomes comparable to the stellar radius. This makes any definition of the radius much less obvious, and parameters like T_{eff} and $\log g$ lose much of their meaning if a radius is not defined properly. Usually, the radius is defined as the distance from the center where the mean Rosseland optical depth reaches unity (see the discussion on the gray case later in this article), and the spectrum provides direct information on the stellar radius. A spherical geometry must be assumed to model these stars.

General theory

A physical theory providing a description of the structure of a stellar atmosphere will essentially express how energy is transported from the stellar interior (where it is generated by nuclear reactions) through the atmosphere (to be radiated in the interstellar medium). Radiation is the major mode of energy transport through the surface layers. In cool stars with sub-photospheric convective zones, convection carries a significant fraction of the energy in the deep photosphere. However, in upper layers, convection becomes rapidly inefficient. A hot convective cell rising from deeper layers will quickly radiate all its extra energy and reach an equilibrium with the local gas. Some spectral features are nonetheless sensitive to convection and are used to check convection theories. At much lower densities, as in the solar corona, heat conduction becomes an important process. In the upper atmosphere, energy may also be transported by MHD or acoustic waves that are responsible for heating the corona non-radiatively. These waves have been generated by strong and geometrically complex magnetic fields, which in turn are created by a dynamo effect in the convection zone (see the section on non-thermal processes later).

Radiation therefore not only probes the physical conditions in stellar outer layers, but is a major constituent that structures the whole atmosphere. It is a non-local process that drives the atmosphere away from thermodynamic equilibrium (as we see a star, there are energy losses). Because non-equilibrium situations are always difficult to model, a number of simplifications have historically been made to describe the geometry, the dynamics and the thermodynamics of stellar atmospheres. Early models assumed one-dimensional geometry (plane-parallel slab, spherical symmetry), a static medium, *local thermodynamic equilibrium* (LTE) and a frequency-independent opacity (gray atmosphere). Some of these assumptions have since then been lifted to model different stellar types (e.g. departures from LTE are especially important in hot stars), but the full 3D, non-LTE, radiation-hydrodynamics problem still remains intractable. The main emphasis was first put, on the one hand, on LTE metal line-blanketed model atmospheres incorporating the detailed opacity of millions of atomic lines, and on the other hand on non-LTE model atmospheres with much fewer opacity sources. Later, the emphasis shifted to non-static models to interpret the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

observations of stellar winds. In recent years, non-LTE metal line-blanketed models have been developed for static photospheres and stationary winds. Solving the non-LTE radiation hydrodynamics problem with detailed opacities and constructing multidimensional non-LTE model atmospheres are major avenues of research today.

Radiation transport

While radiation transport may be expressed most generally by the kinetic (Boltzmann) equation for photons (similarly to the case of massive particles), it is more customary to derive the radiative transfer equation phenomenologically. Let us consider radiation traveling in a direction s. The specific intensity I_ν is the amount of energy transported by units of surface, solid angle, time and frequency. Along the path s, the change in specific intensity, $\mathrm{d}I_\nu/\mathrm{d}s$, is the sum of the energy losses and sources:

$$\mathrm{d}I_\nu = -\kappa_\nu I_\nu \,\mathrm{d}s + j_\nu \,\mathrm{d}s \tag{4}$$

or

$$\mathrm{d}I_\nu/\mathrm{d}\tau_\nu = -I_\nu + S_\nu \tag{3}$$

where the *source function*, S_ν, is the ratio of the emission and absorption coefficients, and $\mathrm{d}\tau_\nu = \kappa_\nu \,\mathrm{d}s$ is the *optical depth* along the radiation path. The absorption coefficient, κ_ν, describes the removal of energy from the radiation field by matter, and the emission coefficient, j_ν, describes the energy released by the material. Thus these two coefficients describe the interaction between matter and radiation. Actual expressions are given by the microscopic state of the plasma. A more general form of the radiative transfer equation may be written by replacing the derivative along s by the total derivative ($\mathrm{d}/\mathrm{d}s \rightarrow 1/c\partial/\partial t + \boldsymbol{n}\cdot\nabla$).

The formal solution of this equation follows a standard integration. The specific intensity emerging from a slab having an optical depth τ_ν is:

$$I_\nu(0) = I_\nu(\tau_\nu)\,\mathrm{e}^{-\tau_\nu} + \int_0^{\tau_\nu} S_\nu(t_\nu)\,\mathrm{e}^{-t_\nu}\,\mathrm{d}t_\nu. \tag{4}$$

The first term expresses the attenuation of the incoming intensity and the second term is the source term. This solution, however, hides the actual difficulty of solving the radiative transfer equation. The source function contains a thermal emission term and a scattering term. The scattering term depends on the mean intensity of the radiation field (specific intensity averaged over all directions) resulting in a non-linear equation in most cases.

A numerical solution of the radiative transfer equation may be obtained by different methods depending on the complexity of the problem at hand. Equation (4) provides a formal solution if the source function is fully specified. For a long time, a direct quadrature was viewed as extremely inefficient due to the required exponentials. This limitation has mostly disappeared with present-day computers. Other methods can be divided into two main classes: first-order form and second-order form of the transfer equation. First-order methods have been recently revived by an adaptation of the discontinuous finite element method. The specific intensity is assumed to be given as a linear function of optical depth. Straightforward analytic manipulations result in simple expressions for the discretized specific intensities. The solution is obtained independently for each angle (direction) and for each frequency. The second-order method, usually known as the Feautrier method, has long been the method of choice, because it allows easy expression of the angle coupling arising from the scattering term in the source function. The basis of the Feautrier method is to introduce the symmetric and antisymmetric angle averages of the specific intensity and to combine the transfer equations in the upward and downward directions. Some simple algebra results in a second-order differential equation. With proper boundary conditions, a tridiagonal set of discretized equations is solved by a straightforward Gaussian elimination scheme.

Generally, the source function is not known fully. The source function may be written as:

$$S_\nu = \alpha_\nu + \beta_\nu \bar{J} \tag{5}$$

where α_ν describes the thermal processes, β_ν the scattering processes and $\bar{J}$ is an appropriate average of the specific intensity over angles and (line) frequencies. The details of the frequency average depend on the assumed correlation in frequency of absorption and emission processes (the so-called complete and partial redistribution functions). Scattering is mostly due to electron Thomson scattering (Compton scattering is unimportant except in the hottest atmospheres); additional scattering processes may be present. Although the source function may remain linear in the specific intensities in some special circumstances (two-level atom case), scattering introduces a coupling between the specific intensities at different angles and frequencies. The radiative transfer equation must now be solved simultaneously at a number of angles and frequencies. This coupling may easily be expressed in the Feautrier scheme and a single-step solution is still possible in the linear case. However, the Gaussian elimination scheme requires matrix inversions where the matrix size is the product of the number of angles times the number of frequencies. Nonlinear, multilevel atom problems are classically solved by linearizing the equations and iterating. The complete linearization method is based on such a Newton–Raphson method. The radiative transfer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is solved using the Feautrier scheme, thus taking fully into account the angle–frequency couplings. This method converges rapidly because all couplings are treated explicitly, but the computing cost quickly becomes prohibitive. An iterative method that avoids explicitly solving the coupling would thus be preferable.

The straightforward iteration method, the so-called Λ iteration, consists in solving the transfer equation by assuming a starting estimate of the source function, then recalculating the source function with the first estimate of $\bar{J}$, and then iterating. This iteration scheme has long been known to fail in the case of strong scattering (most strong resonance lines, e.g. Lyman α, C IV, Mg II). Scattering decouples the radiation field from local sources and sinks and involves global transport of photons over large distances in the atmosphere. The iteration process propagates information on the changes of $\bar{J}$ over a mean free path ($\Delta\tau \approx 1$) at each iteration. This scheme therefore requires a very large number of iterations to converge at large optical depths (and, more seriously, the solution seems to be converged while still far from the exact solution).

Recent progress was achieved upon the realization that not all the parts of the coupling should be treated on the same footing. Some parts of the coupling are more important and must be treated explicitly to ensure convergence, while the rest may be treated iteratively (as in Λ iteration). This is the basis of the method of deferred corrections that is at the heart of the accelerated lambda iteration (ALI) methods. The radiative transfer operator is split into a local part that may be easily inverted (thus solved cheaply in a single step) and a non-local correction applied to the fully specified, previous estimate of the source function. Such methods yield fast convergence and are further described in the article on RADIATIVE TRANSFER.

Opacities

Stellar atmosphere theory and spectral analyses depend heavily on atomic physics. The opacity of the atmospheric plasma is the basic physical quantity that determines the radiation transport, and thus determines the atmospheric structure as well as the emergent spectrum. The total absorption coefficient describes the interaction between radiation and matter, and depends on the chemical composition of the atmosphere and its thermodynamical state as expressed by the temperature, the electronic density and the populations of the energy levels of all species.

The total opacity includes four general contributions: the bound–bound, the bound–free, and the free–free opacity, and a scattering term (mostly electron scattering). Detailed expressions for these opacity sources may be found in Mihalas (1978). For hydrogen, analytical expressions exist for the different contributions to the opacity. For heavier elements, detailed calculations of the atomic structures are required to obtain the cross sections. Thanks to huge efforts in the last 20 years, extensive and reliable atomic data are now available. Good starting points to retrieve the necessary atomic data are the Opacity and Iron Projects which provide detailed photoionization cross sections, line transition probabilities and collisional data for the most abundant light elements up to iron, the Kurucz line list, which is widely used in stellar atmosphere modeling and spectral analyses, and the National Institute of Standards and Technology (NIST) which provides critically evaluated atomic databases (level energies, oscillator strengths) and an extensive bibliography.

With the exception of atomic level energies, the bulk of the required atomic data have been obtained theoretically. Some line transition probabilities and very few photoionization cross sections have been measured in the laboratory, in part due to the difficulty of creating the necessary physical conditions to measure them. To check the theoretical data, stellar atmospheres may provide a different laboratory where such conditions are met. This approach was mainly followed to derive oscillator strengths, matching the solar (or stellar) spectrum with model spectra. These data may be affected by systematic effects due to imperfections in the solar (stellar) model atmosphere used to compute the spectrum, but these systematic errors cancel out when performing a differential analysis (i.e. the oscillator strengths are derived from the solar spectrum and applied to the analysis of solar-type stars). Such sets of oscillator strengths are particularly valuable for chemical abundance studies.

Major effects resulting in departures from LTE are shifts of the ionization equilibria. Reliable photoionization cross-sections are therefore essential data in non-LTE model atmospheres. Detailed atomic structure calculations of atoms and ions having many electrons yield very complex photoionization cross-sections, with resonances due to autoionization processes. Available laboratory data are very limited (mostly cross sections from the ground state of neutral atoms). In this case, stellar atmospheres are not as useful for line strength, because the continuum spectrum reflects the total bound–free and free–free opacities of all atoms and ions present in the photosphere. This opacity is usually dominated by the hydrogen opacity because hydrogen is the major constituent of the atmosphere. In the presence of strong chemical anomalies, the total opacity may be dominated by other opacity sources. This is particularly the case in the far-ultraviolet spectrum of silicon-rich Ap stars. The bound–free Si II opacity becomes dominant, and the observed spectrum reflects these cross-sections. Figure 2 displays the UV spectrum of the ApSi star HD 34452. The observations are compared with theoretical spectra calculated with new

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

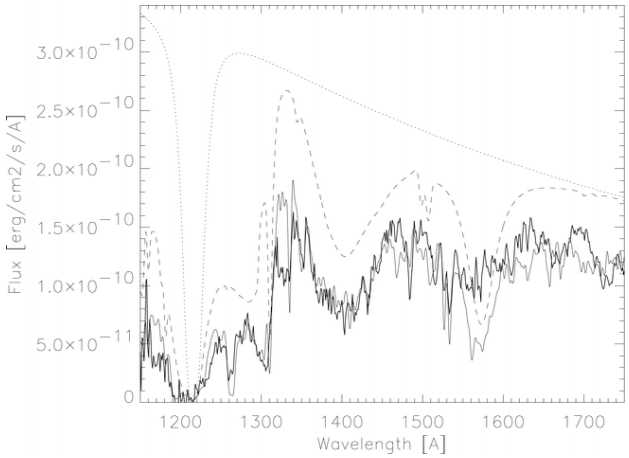


Figure 2. IUE low-resolution spectrum of HD 34452 (thick curve) compared with a theoretical spectrum synthesis (thin curve). Continuum spectra (H + C: dotted curve; H + C + Si^+: broken curve) are also shown to illustrate the huge effect of Si II autoionization resonances (from Lanz *et al* 1996 *Astron. Astrophys.* **309** 218).

cross sections computed with the *R*-matrix method. The atomic data calculations are very similar to those of the Opacity Project. The model spectrum reproduces very well the positions and widths of the broad autoionization resonances, and therefore provides supporting evidence of the reliability of this set of atomic data. Additional evidence of the quality of Opacity Project data includes good predictions of atomic level energies and comparisons with large samples of oscillator strengths from other theoretical studies or laboratory measurements.

Line formation

Let us assume that the structure of the atmosphere (temperature, density, atomic level populations) is known, and let us examine first the limited problem of line formation. Why do most stellar spectra display *absorption* lines? What can we learn from *emission* line spectra?

We can recast equation (4) in the case of a semi-infinite atmosphere. The emergent intensity seen by an external observer ($\tau = 0$) at an angle θ to the normal of the surface ($\mu = \cos\theta$):

$$I_\nu(\tau = 0, \mu) = \int_0^\infty S_\nu(t_\nu)\, \mathrm{e}^{-t_\nu/\mu}\, \mathrm{d}t_\nu/\mu. \tag{6}$$

The emergent intensity is a weighted average of the source function along the line of sight. If we suppose further that the source function is a linear function of depth, $S_\nu(\tau_\nu) = S_{0\nu} + S_{1\nu}\,\tau_\nu$, then equation (6) yields:

$$I_\nu(\tau = 0, \mu) = S_{0\nu} + S_{1\nu}\,\mu = S_\nu(\tau_\nu = \mu) \tag{7}$$

which is known as the *Eddington–Barbier* relation. This result states that the emergent intensity is characteristic of the value of the source function at about optical depth unity along the line of sight. In a spectral line, we encounter optical depth unity in higher layers for frequencies with high opacity (line core) and in deeper layers for frequencies with low opacity (line wings, continuum). Therefore, emergent line profiles map the depth dependence of the source function. If we further assume LTE, $S_\nu = B_\nu$, the temperature structure can in principle be inferred. Absorption line stellar spectra indicate that the line source function decreases with increasing height in photospheres. In cool stars, line core emissions are observed in the strongest lines (e.g. Ca II H and K) showing that the source function increases in the highest layers. This has been straightforwardly interpreted as indicating a temperature decrease in the photosphere and a temperature rise in the upper layers (chromosphere). Recent hydrodynamical simulations suggest, however, that the real structure of the solar chromosphere may be more complicated; hot regions which are responsible for the line core emissions are embedded in a low-temperature, photosphere-like medium.

While the line profiles map the source function behavior, a straightforward interpretation in terms of temperature should be approached with caution. First, equation (6) shows that the source function at different depths contributes to the emergent intensity. Limiting the contribution to a single depth of formation is quite useful at first, but is still an oversimplification. The concept of *contribution functions* seems a preferable approach. Second, the source function contains a thermal term and a scattering term (equation (5)). While the mean intensity decreases with height as photons escape, the thermal term depends on the details of photoexcitations and recombinations. Therefore, an interpretation of the line profiles in terms of physical parameters requires a careful and detailed non-LTE analysis.

In extended atmospheres and in winds, line profiles provide additional information on the geometry and the dynamics of the atmosphere. The most typical signatures of stellar winds are asymmetric P Cygni line profiles.

Structural equations

A *model atmosphere* specifies all atmospheric state parameters as functions of depth (actually as discretized values along the depth coordinate because the problem is solved numerically). The list of parameters required to describe a model atmosphere depends on the basic assumptions of the model, but traditionally includes only electron temperature, total density electron density and atomic level populations. The detailed radiation field is generally not stored because it can be recalculated through a simple formal solution of the radiative transfer equation when all other state parameters are known. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

predicted emergent spectrum alone is sometimes stored separately, because this is the most interesting output to compare with observations.

The state parameters of a model atmosphere are most generally described by distribution functions, which are calculated by solving the kinetic equation. The set of equations to solve is in general truly enormous because the individual particles are not just the atoms and ions, but all excited states of atoms, ions and molecules. It is standard in kinetic theory to simplify this system by constructing equations for the moments of the distribution functions. After summation over all kinds of particles, one obtains the well-known hydrodynamic equations, namely the continuity equation, the momentum equation and the energy balance equation. In addition, particles of each type must satisfy a conservation equation. The moment equations form the set of equations that are actually solved. In full generality, this system remains still intractable, and some simplifying assumptions have to be made. Let us consider the classical model atmosphere: a one-dimensional, stationary and static medium. The system of equations simplifies to:

$$\left(\frac{\mathrm{D}n_i}{\mathrm{D}t}\right)_{\mathrm{coll}} = 0 \tag{8}$$

$$\mathrm{d}P/\mathrm{d}z = -\rho g \tag{9}$$

$$\nabla F_{\mathrm{rad}} = 0 \Longrightarrow F_{\mathrm{rad}} = \sigma T_{\mathrm{eff}}^4 \tag{10}$$

where $(\mathrm{D}n_i/\mathrm{D}t)_{\mathrm{coll}}$ is the collision term in the kinetic equation that describes creations and destructions of particles of type i, P is the total pressure, z the geometrical coordinate, ρ the total density, g the gravitational acceleration at the stellar surface, F_{rad} the radiative flux, σ the Stefan–Boltzmann constant and T_{eff} the effective temperature. These equations are called the *statistical equilibrium*, *hydrostatic equilibrium* and *radiative equilibrium* equations. The explicit form of the statistical equilibrium equation incorporates all individual radiative transitions (rates depend on $\bar{J}$) and collisional transitions (rates depend on T and n_e) that populate and depopulate every individual energy level i of all atoms and ions. Therefore, the basic structural equations must be solved simultaneously with the radiative transfer equation at a large number of frequencies to compute the radiative rates and the total radiative flux (equation (10)). While the structural equations depend on the radiation field, the absorption and emission coefficients appearing in the radiative transfer equations depend on the local temperature, density and populations n_i. All these equations therefore form a highly coupled, highly non-linear system of equations. It may be solved using linearization techniques and ALI schemes for radiative transfer.

The description of a plasma is greatly simplified if thermodynamic equilibrium holds. In this state, the particle velocity distribution (Maxwell distribution), the ionization (Saha formula) and the excitation (Boltzmann law) equilibria are specified uniquely by two thermodynamic variables, usually chosen as the absolute (electron) temperature T and the electron density (n_e). Because a stellar atmosphere is essentially a boundary, thermodynamic equilibrium cannot hold. As photons escape, there are energy and momentum losses that give rise to significant gradients in the state parameters. However, even if thermodynamic equilibrium cannot be assumed for the whole stellar atmosphere, we may still use the concept of *local thermodynamic equilibrium* (LTE). The Maxwell and Saha–Boltzmann distributions cannot be applied globally, but may be used locally for the local values of $T(\boldsymbol{r})$ and $n_e(\boldsymbol{r})$. On the other hand, the radiation field (which is by essence non-local) is allowed to depart from its equilibrium, Planckian, distribution. In non-LTE (NLTE) model atmospheres, some populations are allowed to depart from their LTE, Saha–Boltzmann distributions. A Maxwellian velocity distribution is still assumed to hold in most NLTE models.

Microscopically, LTE holds if all atomic processes are in detailed balance (i.e. the number of any given processes is balanced by the same number of opposite processes). Since electrons have a Maxwellian velocity distribution, collisions between atoms and electrons will distribute this temperature information among the atomic levels and will tend to maintain the local equilibrium. LTE is therefore a good approximation where collisions far outnumber radiative transitions. When radiative transitions dominate, LTE remains a good approximation as long as all radiative transitions are in detailed balance. This is verified in deep layers where a photon has a very low probability of escaping, but LTE will break down in upper layers. As photons escape, the number of photoexcitations will drop below spontaneous deexcitations, and the detailed balance requirement will not be met. Departures from LTE are thus most likely in hot stars (strong radiation field) and low-density atmospheres (supergiants). The validity of LTE is further discussed in the article on STELLAR ATMOSPHERES: EARLY-TYPE STARS.

The gray case

There is one further simplification that allows derivation of an analytical expression for the temperature structure: the gray case. Besides its historic importance, gray model atmospheres are the most obvious starting point for constructing more sophisticated model atmospheres. The opacity is assumed to be a constant, frequency-independent value. Usually, the Rosseland mean opacity is adopted:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\chi_\nu = \bar{\chi} = \chi_R = \frac{dB}{dT}\left(\int_0^\infty \frac{1}{\chi_\nu}\frac{dB_\nu}{dT}d\nu\right)^{-1}. \quad (11)$$

The Rosseland mean opacity is also used in modeling stellar interiors, because it yields the exact total radiative flux at depth using the diffusion approximation, and it will therefore give the correct temperature structure in the deepest layers of the atmosphere. The system of equations is greatly simplified by using frequency-averaged moments of the radiation field: J, the mean intensity, H, the flux, and K the second-order moment. The radiative equilibrium reduces to

$$J = B = \sigma T^4 \quad (12)$$

and the second moment of the radiative transfer equation:

$$\frac{dK}{d\tau} = H. \quad (13)$$

Invoking the Eddington approximation, $K = J/3$, which is a good approximation in the deep atmospheric layers where the radiation field is close to isotropic, one obtains

$$T^4 = \frac{3}{4}T_{\rm eff}^4\left(\tau + \frac{2}{3}\right). \quad (14)$$

A rigorous solution obtained by Chandrasekhar leads to a very similar temperature structure. The constant 2/3 is replaced by a slowly varying function of optical depth, the Hopf function $q(\tau)$, ranging from $q(0) = 0.577$ to $q(\infty) = 0.71$.

Non-gray LTE model atmospheres

Gray model atmospheres are not very useful when it comes to interpreting real stellar spectra. Spectroscopic analyses require construction of non-gray model atmospheres, solving the radiative transfer at a number of frequencies using the proper values of the absorption and emission coefficients. Constructing LTE model atmospheres requires a solution of the basic structural equations (e.g., equations (8)–(10)), where the statistical equilibrium equations are replaced by the Saha–Boltzmann equations. The absorption and emission coefficients are therefore functions of the local temperature and electron density. There is still a non-local coupling of the radiation field and matter properties via the radiative equilibrium equation that fixes the local temperature. The temperature structure is often derived via a temperature correction scheme, starting from the gray structure. The local density is derived by a direct quadrature of the hydrostatic equilibrium.

The numerical difficulty in LTE model atmospheres is the complex dependence of the absorption (and emission) coefficient on frequency. Heavy atoms and ions, like iron-peak elements, have millions of spectral lines, requiring solution of the radiative transfer for a huge number of frequencies. A statistical approach (frequency sampling) or a rearrangement of the absorption coefficient (opacity distribution functions, ODFs) must be used to reduce the number of equations to solve. The ODF technique is particularly efficient in LTE, because all opacity contributions from different sources can be grouped together (this is not possible when radiative transition rates are required to solve the statistical equilibrium equations). The combined effect of the opacity due to many spectral lines significantly steepens the temperature gradient in the atmosphere. This effect, known as *line blanketing*, may be explained qualitatively as follows. In the continuum-forming layers (where the continuum optical depth is about unity), photons are absorbed preferentially at frequencies corresponding to bound–bound transitions where the opacity is still large enough. The flux is blocked at these frequencies and thus has to increase at frequencies with low opacity so that the same total amount of energy is transported. This increase requires a steeper temperature gradient. The deepest and the shallowest layers are not affected by this effect (the optical depth is either very large or very small), thus implying that line opacity results in a higher temperature in the continuum-forming region. The cumulative effect of millions of lines alters the temperature structure of the atmosphere significantly and, therefore, line blanketing is an essential ingredient of model atmosphere construction (thus requiring complete and detailed opacities). Indeed, the change of the temperature gradient results in significant differences in the emergent spectrum.

Fully line-blanketed LTE model atmospheres calculated by Kurucz have become a standard. These model atmospheres incorporate the opacity of over 40 million atomic, ionic and molecular lines. Currently, these model atmospheres are available for a wide range of stellar parameters ($T_{\rm eff}$, log g, chemical composition) that characterize O-type to K-type stars of different metallicities. Spectral models of stellar populations make a broad use of Kurucz model atmospheres. LTE model atmospheres of very cool stars (M-type and cooler stars) have been constructed independently with several computer programs, designed by Gustafsson and co-workers, Tsuji, Johnson and co-workers, and Allard and Hauschildt. Complex molecular and dust chemistry occur in these cool atmospheres that are not included in Kurucz model atmospheres. The article on STELLAR ATMOSPHERES: COOL STARS provides additional discussion.

NLTE model atmospheres

Departures from LTE are expected when radiative rates dominate collisional rates, which typically occurs at high

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

temperatures and low densities. Therefore NLTE model atmospheres will be most important to interpret observations of hot stars (spectral types O, B and A), of extended atmospheres of supergiants and of stellar winds. Early, simple NLTE model atmospheres confirmed this trend. The Hertzsprung–Russell diagram was then roughly divided into two zones, one where LTE was deemed a good approximation, and another where non-LTE effects are important. However, in high-opacity regions like the extreme ultraviolet, the spectrum is formed very high in the atmosphere and thus is prone to non-LTE effects even in cooler stars. Moreover, this division of the HR diagram was based on a rough quantitative criterion. In the last 20 years, progress in observational techniques has resulted in spectroscopic data of much higher quality. The question of the necessity of non-LTE models should therefore be reconsidered in terms of the required accuracy in a given spectral range.

The construction of NLTE model atmospheres requires the solution of large systems of nonlinear equations. Iterative procedures based on the Λ iteration technique do not work due to the importance of the non-local coupling introduced by the radiation field. Upon this realization, Auer and Mihalas introduced the method of the complete linearization, in which all equations are treated on the same footing. This method requires inversion of matrices whose sizes are equal to the number of equations, and is therefore rapidly limited to a small number of atomic levels that are allowed to depart from LTE (generally, H, He and few CNO levels). These early NLTE model atmospheres confirmed that LTE is an inadequate assumption for modeling hot stellar atmospheres. During the same period, however, the importance of line blanketing was stressed. It was then quite unclear which of the simple NLTE models or LTE line-blanketed models had to be preferred to model hot stars. The emergence of the ALI methods broke the limitations of the complete linearization method and led to more sophisticated NLTE model atmospheres. In recent years, NLTE line-blanketed model atmospheres incorporating fully consistent non-LTE iron-peak element opacity have been constructed. Extensive grids like the Kurucz LTE grid of model atmospheres still do not exist, but are expected to become available in the near future.

Early NLTE wind models also were relatively simple, including only a few non-LTE atomic levels, a simple photosphere as a boundary condition and using the Sobolev approximation throughout the wind. Here again, the ALI methods allowed the incorporation of more species with more detailed model atoms, including also the effect of iron-line blanketing, at least in a statistical sense. The energy balance and the momentum equations are not solved in most wind models, but an empirical temperature and velocity structure of the wind is assumed. The temperature often follows a gray structure, while the velocity is given as a simple parametric law (β-type law). In the last decade, a group at Munich University led by Kudritzki has put a tremendous effort into constructing so-called *unified model atmospheres*. Unified models treat the photosphere and the wind on the same footing, and they solve the energy balance and the momentum equations assuming a parameterized force due to line radiation pressure. Further discussion of NLTE wind models may be found in the article on RADIATIVELY DRIVEN STELLAR WINDS FROM HOT STARS.

Non-thermal phenomena

Thanks to the significant advances in x-ray instrumentation during the last 20 years, it has become clear that x-ray emission is widely present in stars of different spectral types. This emission indicates the presence of hot, coronal, million-degree material. From pressure and radiative equilibria, this implies that the emitting regions have very low densities, are very high in the atmosphere and are heated by non-radiative energy deposition in these layers.

The main properties of stellar x-ray emissions can be summarized as follows:

- Essentially all cool main-sequence stars of spectral types F to M are x-ray emitters, with x-ray luminosity in the range $10^{26–31}$ erg s^{-1}. The full range of x-ray luminosity is observed within all spectral types, and the average luminosity is nearly independent of spectral type.
- O- and B-type stars have an x-ray luminosity, L_X, in the range $10^{29–34}$ erg s^{-1}. The x-ray luminosity is approximatively proportional to the bolometric luminosity, $L_X/L_{bol} \approx 10^{-6}$. It is also correlated to the momentum flux in the wind or, alternatively, to the wind kinetic energy flux.
- Main-sequence stars in the spectral range B8 to A5 are not detected, with the exception of some magnetic Ap stars.
- X-ray emission in cool giants and supergiants shows a cut-off in emission level corresponding to a transition between strong winds and coronal emission.

The observed relation between the x-ray luminosity and the wind properties of early-type stars indicates that x-rays are emitted in fast, massive radiatively driven winds. Instabilities in the wind lead to the formation of shocks and to the associated non-thermal radiative emission. This mechanism cannot, however, explain the x-ray emission of later-type stars, because the predicted luminosity would fall far short of the observations due to the steep decrease of the total luminosity.

The first standard models of the solar corona heating were based on energy deposition by acoustic waves.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Based on this mechanism, it was expected that the Sun would be close to the top of the distribution of stellar x-ray luminosities. Moreover, the amount of energy carried by the outer convective envelope that could be transported by acoustic waves decreases rapidly towards later spectral types. However, x-ray surveys did not reveal such a decrease of the x-ray luminosity towards later spectral types, and the solar x-ray luminosity is rather below average. Acoustic waves are therefore not the source of coronal heating. They are, however, the most likely source of *chromospheric* heating. Observations recorded with the Solar Heliospheric Observatory (SOHO) are showing that acoustic waves are damped in the upper chromosphere and do not propagate through the transition region.

While the level of coronal emission does not correlate with spectral type, a clear correlation with the rotation rate has been established. Moreover, the solar data indicate a connection between activity and magnetic field strength. The magnetic field is generated by a *magnetic dynamo* with an efficiency depending on the rotation period of the star. By themselves, rotation or magnetic fields do not result in coronal emission. Detailed solar models indicate that magnetic fields must be embedded in a strong convective pattern in order for mechanical energy to be transported to and deposited in the upper atmosphere. With the disappearance of the outer convective zone in stars with spectral types earlier than A7, this heating mechanism vanishes. Indeed, no coronas or related x-rays are observed in early-type stars. At the other end of the spectral sequence, stars are becoming completely convective at spectral type M5. The dynamo efficiency and related emission might thus decrease in stars later than M5. Finally, activity is decreasing with age too: the weak solar-like wind of cool stars carries away angular momentum and spins down these stars on timescales of the order of 10^9 years.

Despite this general understanding, the physics of the outer atmosphere of solar-like stars is still poorly understood. High spatial resolution images of the Sun in the x-ray and extreme ultraviolet domains reveal the complex structure of the solar upper atmosphere. Active cool stars reveal similar complexity with large spots on their surface. The simultaneous study of the Sun and late-type stars (the so-called solar–stellar connection) has so far been quite fruitful and will remain an essential avenue towards a better understanding of late-type stars and non-thermal phenomena in stellar atmospheres.

Future developments

Progress and success in NLTE modeling of stellar atmospheres and winds has been tremendous during the last three decades. Now it remains to make NLTE models as common as the very successful LTE model atmospheres constructed by Kurucz. This step is essential to benefit from the improved physics of NLTE model atmospheres in analyzing the huge number of spectra that will be recorded by the new generations of large telescopes and multiaperture spectrometers. In this respect, the availability of grids of NLTE line-blanketed model atmospheres is a necessary step. Such grids of models will be particularly important for stellar population studies too. On the other hand, progress in including iron-line blanketing and consistently solving the momentum and energy balance equations in unified model atmospheres should be very rewarding in the analysis of the spectra of the brightest individual objects in external galaxies as far as the Virgo cluster. Unified model atmospheres should put the spectral analysis of these objects on the best physical footing to derive the surface chemical composition and the mass loss rate of these objects; prime quantities to understand the effect of stellar evolution on the evolution of galaxies. Most importantly, such analyses will provide an independent method of deriving the stellar distances of these distant objects. In cool stars (including brown dwarfs and giant planet atmospheres), future emphasis will be given on improving molecular and dust chemistry, in particular in the presence of shocks. The opacity and thus the emergent spectrum can be altered significantly by changing the molecular equilibria and by the formation of dust.

While considerable progress remains to be made in 1D model atmospheres, we may expect that more and more attention will be devoted to multidimensional geometries. The methodology for solving radiative transfer in 2D or 3D is well known. The fastest method is based on the short characteristics method (integration of the formal solution between any two spatial grid points). However, the number of directions in which the radiative transfer must be solved to compute angle-averaged quantities like the mean intensity increases dramatically to obtain a good accuracy (typically by up to two orders of magnitude). This is especially true when the radiation field is strongly peaked or in the presence of strong gradients, because the photons can escape much more easily in one direction than in others. The additional spatial dimension(s) in the model could therefore alter significantly the structure of the model and the predicted emergent spectrum. The requirement to solve the transfer equation in many directions results, however, in very demanding computer simulations, even for a simple description of the plasma including very few NLTE atomic levels. Examples of astrophysical objects requiring multidimensional simulations and currently under investigation include models of rotating winds, of colliding winds, of circumstellar disks, of solar prominences and other structures in the solar atmosphere. Multidimensional radiation hydrodynamics is a field in its infancy that is now developing thanks to the available computing power.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Web update

For the first time, an optical coronal line from a star other than the Sun has been detected. The UV-Visual Echelle Spectrograph (UVES) on the VLT 8.2 m KUEYEN telescope at the ESO Paranal Observatory has observed a coronal emission line in nearby variable star CN Leonis from iron ions that have lost 12 electrons (Fe XIII). CN Leonis, a cool star, is located at a distance of 8 light-years.

Bibliography

The fundamental textbook of the field is:

Mihalas D 1978 *Stellar Atmospheres* 2nd edn (San Francisco: Freeman)

Theoretical and observational aspects of the stellar atmospheres are both covered in a good textbook and some lecture notes from a recent European Astrophysics School:

De Greve J P, Blomme R and Hensberge H 1997 *Stellar Atmospheres: Theory and Observations (EADN Astrophysics School IX, Brussels 1996) Lecture Notes in Physics* (Berlin: Springer)

Gray D F 1992 *Observations and Analysis of Stellar Photospheres* 2nd edn (Cambridge: Cambridge University Press)

Recent textbooks on the outer regions of stellar atmospheres are:

Golub L and Pasachoff J M 1997 *The Solar Corona* (Cambridge: Cambridge University Press)

Lamers H J G L M and Cassinelli J P 1999 *Introduction to Stellar Winds* (Cambridge: Cambridge University Press)

Other related textbooks on physical aspects of radiative transfer include:

Rybicki G B and Lightman A P 1979 *Radiative Processes in Astrophysics* (New York: Wiley-Interscience)

Shu F H 1991 *The Physics of Astrophysics I. Radiation* (Mill Valley, CA: University Science Books)

An elementary-level textbook is:

Böhm-Vitense E 1989 *Introduction to Stellar Astrophysics II. Stellar Atmospheres* (Cambridge: Cambridge University Press)

A grid of LTE model atmospheres is described in:

Kurucz R L 1979 *Astrophys. J. Suppl.* **40** 1

Some recent non-LTE line-blanketed model atmospheres are described in:

Aufdenberg J P, Hauschildt P H, Shore S N and Baron E 1998 *Astrophys. J.* **498** 837

Dreizler S and Werner K 1993 *Astron. Astrophys.* **278** 199

Hubeny I and Lanz T 1995 *Astrophys. J.* **439** 875

Sources for atomic data include:

Seaton M J etal 1995 *The Opacity Project* vol 1 (Bristol: Institute of Physics Publishing)

Wiese W L, Fuhr J R and Deters T M 1996 Atomic transition probabilities of carbon, nitrogen and oxygen. A critical data compilation *J. Phys. Chem. Ref. Data* Mon. 7

A good site on the Web pointing to many useful atomic databases for stellar atmosphere modeling and spectroscopic analyses is: http://plasma-gate.weizmann.ac.il/DBfAPP.html

Thierry Lanz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Atmospheres: Cool Stars

The term 'cool stellar atmosphere' is generally applied to the atmospheres of stars with effective temperatures less than that of the Sun. In this spirit, this article describes the basic physics of stellar and substellar atmospheres with effective temperatures below about 5000 K. This category includes two common subclasses: dwarf stars and giant stars. From the point of view of the STELLAR ATMOSPHERE, the main difference between these subclasses is a question of geometry: plane parallel geometry for dwarfs and spherical symmetry for giants. The types of stars that fall in the 'cool' category reach from evolved M giants and AGB stars to M dwarfs and the new spectral class 'L' proposed for extremely low-mass dwarf stars and brown dwarfs. Although these objects have very different astrophysical origin and history, their atmospheres are described by the same equations and input physics.

Model atmospheres for cool stars are used not only to determine stellar parameters such as effective temperatures and elemental abundances but also as input data for STELLAR EVOLUTION models (very low-mass stars) and population synthesis modeling. Thus the accuracy and quality of our understanding of cool stellar atmospheres has far-reaching impact on other fields of astrophysics.

Cool stellar atmospheres are extraordinarily complex, which makes them hard to understand and model. The main reasons for this are

1. the importance of molecules for the chemical equilibrium and the radiative transfer,
2. the extraordinary importance of spectral line opacities and line broadening,
3. convective energy transport into the outer optically thin regions of cool dwarf atmospheres,
4. formation of dust for very low effective temperatures ($\lesssim$ 2500 K),
5. deviations from local thermodynamic equilibrium (LTE) and local chemical equilibrium.

Further complications can include for example STELLAR PULSATIONS and winds in giant stars or fast rotation and chromospheric activity in young dwarf stars. The following sections will concentrate on the most important problems.

Basic physical picture

In the classical approach a cool stellar atmosphere is assumed to be in time-independent hydrostatic equilibrium. The atmospheres can furthermore be described by using either plane parallel geometry for dwarf stars with gravities $\log(g) \gtrsim 3.5$ or spherical geometry for giant stars with lower gravities.

The luminosity L of the star is parameterized through the effective temperature $T_{\rm eff}$ via

$$L = \sigma T_{\rm eff}^4 4\pi R^2$$

where σ is the Stefan–Boltzmann constant and R is the radius of the star. The pressure structure of the atmosphere in spherical geometry is given by

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -g(r)\rho(r)$$

where the total pressure P is the sum of the gas pressure $P_{\rm gas}$ and the radiation pressure $P_{\rm rad}$ (in some cases an additional turbulent pressure term has to be considered), r is the distance from the center of the star and

$$g(r) = \frac{GM}{r^2}$$

is the gravity where G is the gravitational constant and M is the mass of the star. For dwarf stars with a negligible radial extension of the atmosphere ($\Delta r/R \ll R$) the gravity can be set constant and used as one parameter of the atmosphere.

The energy generated in the deep interior of the star is transported through the atmosphere predominantly by radiation and by convection so that

$$L = L_{\rm rad}(r) + L_{\rm conv}(r) = \text{const}$$

where $L_{\rm rad}(r)$ is the luminosity carried by photons and $L_{\rm conv}(r)$ is the convective luminosity. Cool atmospheres typically have an outer radiative zone where $L_{\rm conv} = 0$ and an inner convective zone where $L \approx L_{\rm conv}$. In very low mass dwarf stars ($M \lesssim 0.5 M_{\rm Sun}$) the convective region reaches out into the optically thin regions of the atmosphere and from there up to the center of the star, leaving only a thin outer radiative layer ('fully convective stars'). There is currently no detailed analytical or quantitative theory of convective energy transport and thus parameterized approaches are generally used in the context of detailed stellar atmosphere models. The most common approach is the mixing length theory with a single free parameter ℓ, called the mixing length (typically 0.5–1.5 times the local pressure scale height of the atmosphere). The convection in cool dwarf atmospheres (M dwarfs and later) is extremely efficient and becomes practically adiabatic; this removes partly the ambiguity of the mixing length parameter for these stars.

In order to calculate the radiative luminosity

$$L_{\rm rad}(r) = 4\pi r^2 \int_0^\infty F_\lambda(r,\lambda)\,\mathrm{d}\lambda$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the radiation flux F_λ (r,λ) must be known at all wavelengths λ. For the cases of plane parallel and spherical geometry F_λ is given by

$$F_\lambda = 2\pi \int_{-1}^{1} I_\lambda(r, \lambda, \mu)\mu \, \mathrm{d}\mu$$

where μ = cos(θ) and θ is the angle of a beam of radiation with the specific intensity $I_\lambda(r,\lambda,\mu)$ to the radially outward direction. The specific intensities $I = I_\lambda(r,\lambda,\mu)$ are obtained from the solution of the radiative transfer equation, which is given for the general case of a special relativistic flow in spherical symmetry by

$$e\frac{\partial I}{\partial r} + \frac{\partial}{\partial \mu}(f I) + g\frac{\partial}{\partial \lambda}(\lambda I) + hI = \eta - \chi I$$

with

$$\beta(r) = \frac{v(r)}{c}$$

$$\gamma^2 = 1/(1-\beta^2)$$

$$e(r,\mu) = \gamma(\mu + \beta)$$

$$f(r,\mu) = \gamma(1-\mu^2)\left[\frac{1+\beta\mu}{r} - \gamma^2(\mu+\beta)\frac{\partial\beta}{\partial r}\right]$$

$$g(r,\mu) = \gamma\left[\frac{\beta(1-\mu^2)}{r} + \gamma^2\mu(\mu+\beta)\frac{\partial\beta}{\partial r}\right]$$

$$h(r,\mu) = \gamma\left[\frac{\beta(1-\mu^2)}{r} + \gamma^2(1+\mu^2+2\beta\mu)\frac{\partial\beta}{\partial r}\right]$$

where $v(r)$ is the flow velocity. For static atmospheres, $v(r) = 0$, and for static plane parallel atmospheres one has in addition $f(r,\mu) = 0$. To solve the radiative transfer equation, the emissivities $\eta = \eta_\lambda(r,\lambda)$ and the extinction $\kappa = \kappa(r,\lambda) + \sigma(r,\lambda)$ coefficients must be known. Here, $\kappa(r,\lambda)$ is the total absorption coefficient and $\sigma(r,\lambda)$ is the total extinction coefficient. For simple isotropic scattering the emissivity η can be written in the form

$$\eta(r,\lambda) = \kappa(r,\lambda)B_\lambda(\lambda, T) + \sigma(r,\lambda)J_\lambda(r,\lambda)$$

where $J(r,\lambda)$ is the mean intensity. The first term describes the thermal sources present in the atmosphere whereas the second term describes scattering. Thus, in order to calculate the radiation field inside the atmosphere, κ and σ must be known. This requires the detailed knowledge of the partial pressures of each atom, ion and molecule present in the atmosphere through a solution of the equation of state (EOS) for cool atmospheres as well as the detailed knowledge of the interaction of each species with radiation through, e.g., photoionization and spectral lines.

Equation of state

The EOS describes the distribution of the available nuclei in a volume element (say, 1 cm^3) to atoms, ions and molecules. Within the atmosphere, the EOS is usually assumed to be that of an ideal gas. This assumption is justified for all but the deep layers of very cool atmospheres where non-ideal gas effects become important. In addition, the assumptions of local chemical equilibrium and LTE are typically made in order to simplify the calculation of the concentrations of individual species (atoms, ions and molecules). The low temperatures in cool atmospheres lead to the formation of a large number of different molecules, which makes their EOS extremely complex. Although there are only a few molecules with large concentrations, such as molecular hydrogen H_2, carbon monoxide (CO), titanium oxide (TiO) and water vapor (H_2O), there are many more molecules with smaller concentrations. These cannot be neglected, however, because their spectral lines can be very strong and thus significant for the energy transport in the atmosphere. Some examples of these important trace molecules are vanadium oxide (VO) and metal hydrides such as MgH, CaH and FeH. The concentrations of these molecules are low but they have important bands of spectral lines that dominate the emitted spectrum at certain optical and red wavelengths. In addition to the important opacity sources, many more molecular species with trace concentrations have to be included because the dependences of the formation and dissociation of all molecules as well as the concentrations of atoms are highly non-linear and even small changes in the concentration of trace species can have significant effects on molecules with important spectral lines. Therefore, the EOS typically includes several 100 molecular species in addition to a number of atoms and ions.

In very cool atmospheres with effective temperatures below about 2500 K the formation of dust grains becomes important and has to be considered in the EOS. This is shown in figure 1 which shows the species with the highest concentrations for both the gas and the condensed phases for a model atmosphere with T_{eff} = 1000 K under the assumption that no settling of the dust particles took place. The figure shows for example that iron condenses from the gas phase (labeled 'Fe') to the solid phase (labels 'Fe/c' and 'Fe/c1' for different condensate forms) around an optical depth of 1. High-temperature condensates such as Al_2O_3 and $CaTiO_3$ form in small concentrations but can have a significant impact on the total opacity of the atmosphere. At lower temperatures condensates such as Fe_3O_4 and water ice form. The main thermodynamic effect of grain formation is to remove the condensed elements from the gas phase, thus changing the concentration of the complete set of molecules. This reduces the opacity of important molecules such as TiO. In addition, the condensates

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

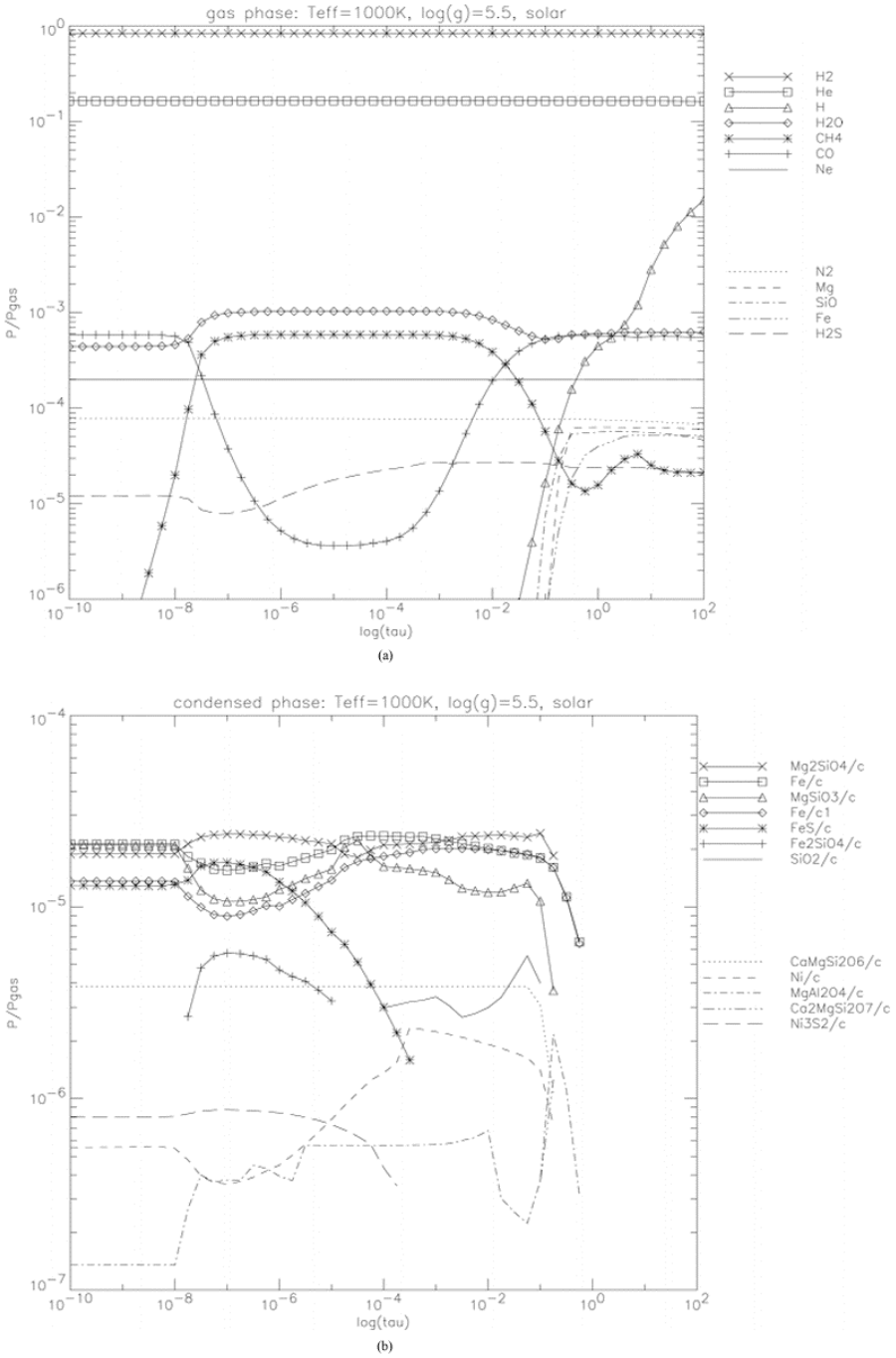


Figure 1. The most important (by concentration) gas phase and condensed phase species in a cool atmosphere with an effective temperature of 1000 K, a gravity of $10^{5.5}$ cm s^{-2} and solar abundances (roughly equivalent to the parameters of the brown dwarf Gl 229B). The *x*-axis gives the optical depth in the continuum at 1.2 μm, the *y*-axis gives the relative concentration.

themselves have an effect on the opacity and thus the structure of the atmosphere; see below.

The molecular EOS is a coupled non-linear system of equations even in the simplest case of complete local chemical equilibrium and LTE. This system can in principle be solved by any appropriate numerical technique, e.g. Newton's method or free energy minimization algorithms. The assumption of local chemical equilibrium and LTE is not always a good approximation. For example, the process of dust formation in cool giant atmospheres is not well understood and deviations from local chemical equilibrium are likely important.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Opacities

Atomic opacities

The opacities in cool stellar atmosphere are dominated by spectral lines of molecules. Atomic processes are important mostly through individual spectral lines that are useful as diagnostics of the structure of the atmosphere and as abundance indicators for individual elements; however, most of the line blanketing is provided by molecular spectral lines. Atomic photoionization and recombination processes lose their importance toward lower effective temperatures and are completely dominated by spectral line absorption and emission. At low temperatures, when molecules form and condensation begins, the strengths of some atomic lines are gradually reduced as the elements are either tied up in molecules (with their own, different set of spectral lines) or as they condense into grain particles (which become important opacity sources themselves). At the same time, however, the atmosphere becomes more transparent in certain wavelength regions which leads to an increase in the strength of other spectral lines. In giants these depletion processes are reduced owing to the lower pressures in the atmospheres so that atomic lines are relatively more important for a cool giant atmosphere than in a cool dwarf atmosphere with comparable effective temperature. The number of atomic lines important in cool atmospheres also drops quickly as the effective temperature drops because exited levels are less and less populated at lower gas temperatures. The strong line blanketing by millions of spectral lines from atoms and ions (e.g. singly ionized iron, Fe II) typical for hotter atmospheres has been reduced to only a few remaining resonance lines (originating from the ground state) of atoms. See also STELLAR OPACITY.

Only the strongest resonance lines of alkali metals remain observable in very cool atmospheres: important atomic lines are the Na I-D lines at λ 5889, 5896 Å, Na I lines at λ 8183, 8195 Å and λ 10 746, 10 749, 10 835 Å and at 2.2 μm, K I at λ 6911, 6939, λ 7665, 7699, λ 9950, 9954 and λ 10 480, 10 482, 10 487 Å. Resonance lines of Rb I at λ 7950 Å, and Ba I at λ 7911, 7913 Å are also particularly strong in very cool dwarf atmospheres and Cs I λ 8521, λ 8944 Å lines appear strong in class 'L' dwarfs. The resonance lines of neutral lithium between 6708 and 6710 Å are important indicators of the substellar nature of cool dwarfs because lithium is destroyed by thermonuclear reactions in stars. Only objects below the hydrogen burning limit will conserve their initial lithium abundance.

Atomic lines in cool atmospheres are very broad owing to van der Waals (vdW) interaction of the parent atom with neutral particles in the gas (mostly He and H_2). This effect perturbs the energy levels of the atoms leading to strong vdW broadening of atomic lines in cool atmospheres. The vdW broadening leads to a Lorentzian line profile with form

$$\phi(\Delta\lambda) = \frac{\Delta\lambda_L}{2\pi}\frac{1}{\Delta\lambda^2 + \Delta\lambda_L^2/4}.$$

The vdW linewidth can be written in the form

$$\Delta\lambda_L = \frac{\lambda_0^2}{2\pi c}\gamma_{vdW}$$

with the vdW damping constant

$$\gamma_{vdW} \propto C_6^{2/5} v^{3/5} N_p$$

where v is the relative speed between the absorbing–emitting particle and the perturber and N_p is the number density of the perturbers. The interaction constant C_6 can be calculated with reasonable accuracy only for fast (compared with the lifetime of the transition) perturbations with neutral hydrogen. However, in cool atmosphere the most important perturbers are molecular hydrogen and helium and the perturbations relatively slow. For these cases, no general quantitative theory exists and approximations are generally used, such as Unsöld's modified hydrogenic approximation of the form

$$C_6 = 1.01 \times 10^{-32}(Z+1)^2\left[\frac{E_H^2}{(E-E_l)^2} - \frac{E_H^2}{(E-E_u)^2}\right]$$

where Z is the charge of the absorber, E the ionization energy, E_l and E_u are the lower and upper level excitation energies of the absorber and the result has the units cm^6 s^{-1}. This approximation works reasonably well for alkali metals but for other elements correction factors ranging from 10 to ~65 had to be introduced to provide reasonable results.

Molecular bands

Figure 2 gives an overview over the most important opacity sources present in the atmosphere of a dwarf star with an effective temperature of 2800 K and solar elemental abundances. The figure clearly shows the overall importance of TiO and water vapor lines in the optical and near-infrared (NIR) spectral regions, respectively. The number of spectral lines per molecule is *vastly* greater than for atoms or ions. For example, modern line lists include about ~175 million lines of TiO and ~300 million lines of water vapor. The spectral lines of molecules are clustered in molecular bands. This has led early on to the development of approximate treatments of molecular opacities by methods such as the just overlapping line approximation (JOLA) or simple straight means techniques. These methods assume that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

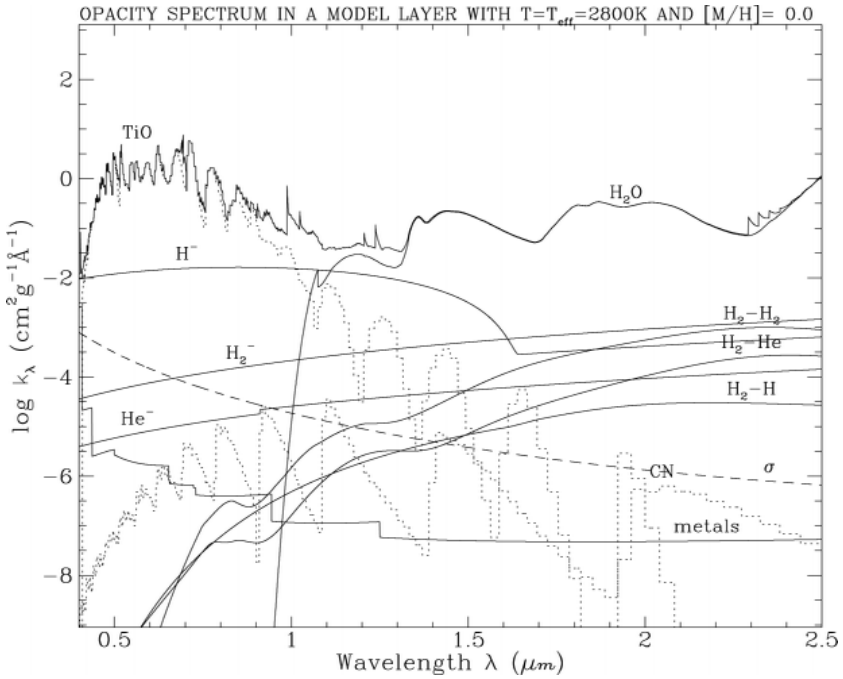


Figure 2. Most important absorbers in a cool stellar atmosphere with an effective temperature of 2800 K, a gravity of 10^5 cm s^{-2} and solar abundances of all elements. This graph does not include the effects of dust settling observed in very cool atmospheres. From Allard F and Hauschildt P H 1995 *Astrophys. J.* **445** 433.

the fine structure of the molecular bands is completely smeared out and that the lines overlap without being saturated. This tends to overestimate the molecular line blanketing by preventing photons from escaping through narrow gaps between molecular lines, resulting in significantly higher temperatures in the atmosphere as compared with the temperature structure obtained with detailed treatment of the individual molecular lines. However, detailed and complete line lists exist only for a limited number of molecules; for many important molecular absorbers only approximate treatments based on the JOLA are available (and even then many times only with estimated band strengths). This is currently the case for such important molecular absorbers as VO, FeH and CaH.

The most important molecular bands in the optical spectral region are TiO and VO for stars with SOLAR ABUNDANCES. For stars with metallicities less than solar, hydrides such as FeH (stars later than M6), MgH, and CaH become important opacity sources, partly because of reduced TiO opacity at lower metallicities. A number of molecules have individual bands that can be important in localized wavelength ranges, e.g. MgH, SiH, OH, CaOH (among the strongest band systems in cool M dwarf stars) and CH. TiO and VO are often used as temperature indicators owing to the sensitivity of the lines to the temperature structure of the atmosphere. The hydride lines appear stronger with increasing gravity, thus making these bands useful as gravity indicators. In low-gravity carbon-enriched atmospheres, species such as CN, ZrO, LaO (S stars) and YO can show important bands in individual wavelength ranges. Although none of these trace molecules contributes significantly to the overall line blanketing in normal cool stars (which is provided mostly by TiO, VO and H_2O) they do absorb in optical spectral regions where TiO is relatively weak and they thus provide local traps for photons trying to escape from the atmosphere.

The NIR and infrared (IR) spectral range in cool dwarf stars with solar abundances is dominated by the opacity of water vapor, which is typically 1–2 orders of magnitude stronger than the next closest opacity sources in this spectral region. The exceptions are the first overtone band of CO between ~2.3 and ~2.4 μm which is very prominent in cool stellar atmospheres and the CO fundamental band at ~4.5 μm, making CO a good temperature indicator in the IR. One of the biggest current challenges in modeling cool stellar atmospheres is to obtain better line lists for water vapor. The problem is the complexity of calculating accurate line positions and strengths for a triatomic molecule such as water at the relatively high (compared with Earth standards) temperatures in the atmospheres of cool stars. This actually means that the water opacity problem is larger for higher effective temperatures, as the low-temperature lines of water vapor are generally better known. Similarly, better line data for VO and methane are also urgently needed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

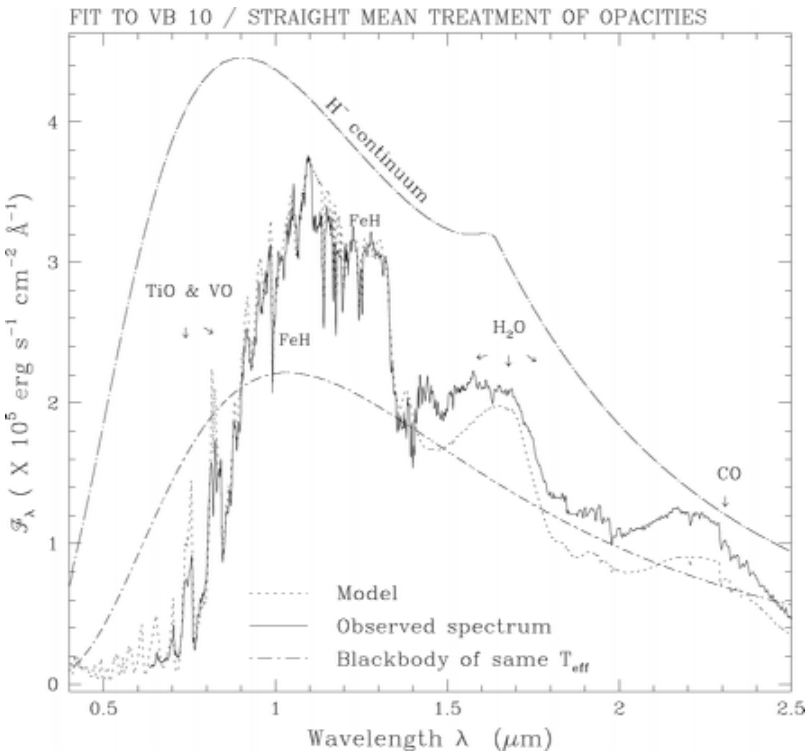


Figure 3. Illustration of the importance of molecular spectral line opacities of cool stellar atmospheres. The plot shows the spectrum of the star VB10 compared with a simple model. The molecular species with the strongest features are indicated. For comparison, a blackbody spectrum with the same effective temperature and the run of the H^- continuum are also given. From Allard F and Hauschildt P H 1995 *Astrophys. J.* **445** 433.

The large number of molecular spectral lines makes the classical notion of the 'continuum' basically useless for cool stellar atmospheres. This is demonstrated in figure 3 for a dwarf star synthetic spectrum with an effective temperature of 2800 K (for comparison, the observed spectrum of the star VB10 is also shown). The curve labeled 'H^- continuum' shows the spectrum calculated by omitting all molecular opacities and retaining only the opacity of the negative hydrogen ion H^-. This shows clearly the importance of the molecular lines in the overall energy distribution of cool stars. In addition, figure 3 shows how much the spectra emitted by cool stellar atmospheres deviate from the blackbody energy distribution (shown as dot-dashed curve) for the same effective temperature. The strong overlap between molecular lines actually helps to reduce the problem of the largely unknown damping constants for molecular lines by effectively masking the details of the line wings.

Dust opacities

The opacity of dust particles becomes important in very cool dwarf stars or in the outermost atmospheres of cool giant stars. Once condensates start to form, the opacity due to the grains can dominate the formation of the spectrum. The opacity of dust particles follows from the Mie theory for given shape, size and optical properties of the particles. Two important differences from the other opacity sources discussed above are (1) dust particles cause a significant amount of non-isotropic photon scattering in cool atmospheres and (2) they have very broad features that affect nearly the whole spectrum. For example, dust grains tend to very effectively absorb photons in the optical spectral range. The absorbed energy is thermalized and re-emitted in the IR range. This leads to a strong re-processing of photons from the optical to the IR and to a heating of the atmosphere.

The presence of dust particles increases the number of parameters of the atmosphere significantly as the size and shapes of the grains must be known. In principle, this could be calculated using a non-equilibrium model for the formation and growth of the dust particles, but this is currently not feasible within the simulation of a cool atmosphere itself. Thus statistical approximations are used and suitable averages over grain size and shape distributions are made. In addition, the optical parameters of each dust species have to be known, e.g. from databases. Some of the important dust opacity sources include species such as $MgSiO_3$, Mg_2SiO_4, soot, FeS, condensed metals (e.g. Fe, Cu, Ti), Al_2O_3 (corundum and sapphire), Fe_2O_3, Fe_3O_4, water ice, ZrO_2, $CaTiO_3$, $MgAl_2O_4$, $CaSiO_3$ and other calcium silicates. For many

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

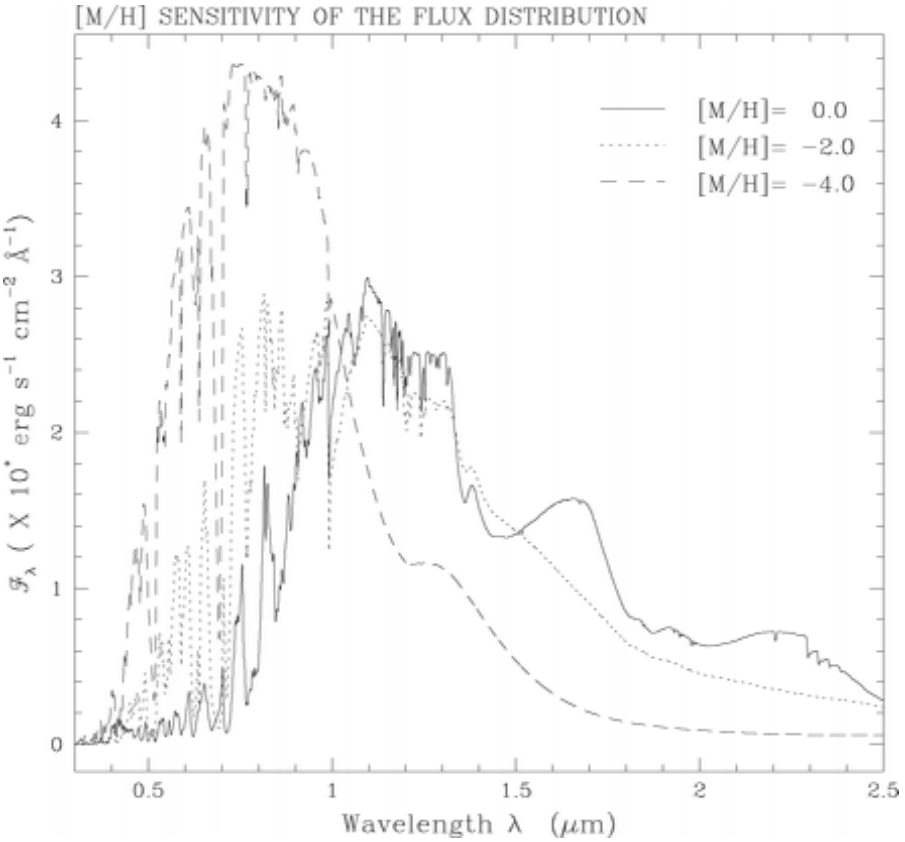


Figure 4. Low-resolution synthetic spectra for cool atmospheres of different metallicities corresponding roughly to the young disk ([M/H] = 0.0), halo ([M/H] = –2.0) and population III ([M/H] = –4.0). The effective temperature is 2800 K and the gravity is log *g* = 5.0 for all models. From Allard F and Hauschildt P H 1995 *Astrophys. J.* **445** 433.

of the important species the optical properties are either not or only poorly known.

A further complication is the apparent settling of dust particles in the coolest known brown dwarf atmospheres. The IR spectra of these objects indicate less dust opacity than expected from condensation models. This is corroborated by the presence of stronger atomic lines than predicted by condensation models. However, the relatively weak water vapor bands indicate the presence of dust in the layers where the water bands form. The current model for this effect is that the dust particles have partially settled below the photosphere. This delivers the desired effect and results in improved fits to observed brown dwarf spectra. For objects with effective temperatures above about 2000 K the dust does not seem to settle and the best results are achieved using full dust opacities. The reason for this might be the fact that the convection zone reaches far into the outer atmosphere for temperatures above 2000 K but retracts from the surface for cooler models. The convection zone could provide the mixing and turbulent motions required to prevent the settling of the dust particles and/or replenish the photospheric layers with refractory material. However, within the framework of the mixing length theory it appears that the convection zone never reaches into the layers of dusty atmospheres where the spectrum forms. This suggests that other mechanisms such as fast rotation or acoustic waves generated by the convection zone might be important.

NLTE effects

Because of the very low temperatures of cool stellar atmospheres, the electron density is extremely low. Collisions with particles other than electrons, e.g. H_2 or helium, are not as effective as electron collisions in restoring LTE, because both of their smaller cross-sections and their much smaller thermal velocities. Therefore, collisional rates which tend to restore LTE tend to be very small in cool stars. This in turn could significantly increase the importance of non-LTE (NLTE) effects in cool atmospheres (with effective temperatures below about 4000 K) when compared with, e.g., solar-type stars with much higher electron densities and temperatures. Departures from LTE in the photospheres of cool stars have been investigated in detail in the line transfer and dissociative equilibrium of H_2 in RED GIANT atmospheres and NLTE effects of CO

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

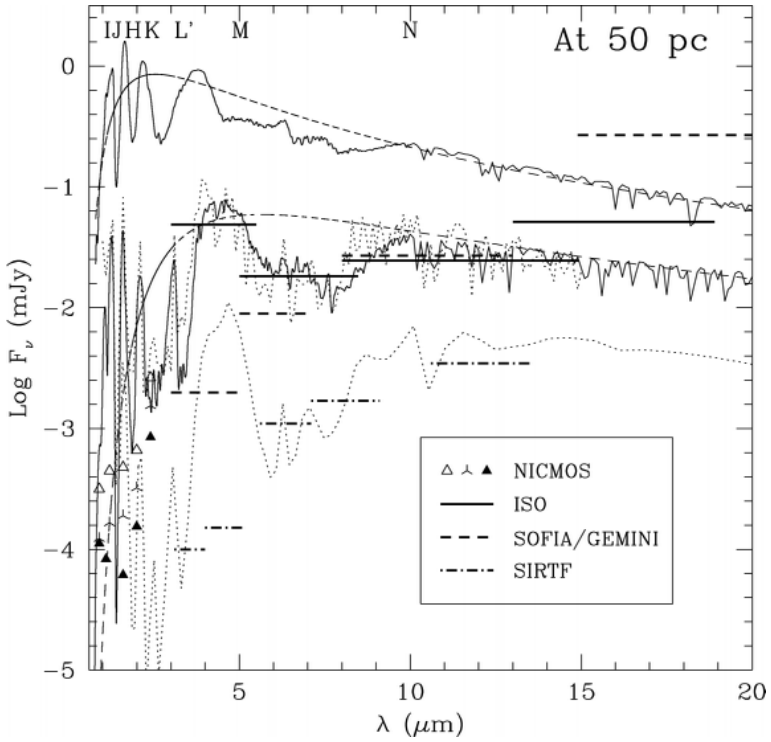


Figure 5. Examples of calculated energy distributions from very cool atmospheres, from the top to the bottom: 2000 K, ~1000 K and 500 K. The letters indicate the location of important broad-band filters. Detection limits for objects of the indicated effective temperature at a distance of 50 pc are given for some space- and ground-based telescopes. From Allard F *et al* 1997 *Ann. Rev. Astron. Astrophys*. **35** 137.

have been investigated both in the Sun and in red giant atmospheres. These studies led to the conclusion that even after the passage of pulsationally driven shocks the recombination of H_2 and CO must proceed rapidly, leading to number densities very close to LTE values. However, in cool atmospheres it has been shown that the equivalent widths of the Li I and Na I resonance lines are affected by NLTE effects and NLTE curves of growth have been derived for a range of model parameters. Departures from LTE of the Ti I atom, and thus indirect changes in the concentration of the important TiO molecule, could therefore have severe and measurable consequences for the atmospheric structure and spectra of cool stars. However, although the NLTE effects in Ti I are noticeable they do not affect the formation of the TiO molecule and its opacity.

One of the open problems is a better understanding of the effects of NLTE on the formation, dissociation and line formation of important molecular species (CO, TiO and H_2O) in cool atmospheres. The main issues are the availability of data, in particular collisional cross-sections, and the technical problems due to the extremely large number of energy levels and spectral lines found in even simple molecules. The latter problem can be addressed by using modern numerical methods such as the operator splitting approach. The data problem, however, is still a major obstacle for detailed NLTE calculations for molecules.

Spectra

The richness of the spectrum emitted by a cool stellar atmosphere is apparent in figure 3. The identification of individual spectral lines is only possible for the strongest resonance lines of atoms. For molecules, only broad features can be identified. The spectrum shown is typical for dwarf stars in this effective temperature range. The peak of the energy distribution is locked at a wavelength ~1.1 μm by the combined effects of TiO lines in the optical and water vapor lines in the IR. Lowering the temperature, which shifts the peak of the corresponding blackbody energy distribution toward larger wavelengths, increases both the TiO and the H_2O opacities rapidly. This leaves only a few small windows open in the 1–2 μm range through which the atmosphere can radiate its thermal energy. Therefore, the location of the peak of the energy distribution of cool dwarf stars is not a good temperature indicator. The slope of the spectrum in this wavelength interval is a more useful temperature indicator. This situation is changed only when dust opacity starts to become important in the optical spectral range at lower effective temperatures.

A similar effect leads to an unexpected behavior of the emitted spectra for a given effective temperature as a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

function of metallicity as illustrated in figure 4. For lower metallicities, the opacities of both TiO and water vapor are reduced owing to the smaller concentration of the molecules. At the same time, collision-induced opacities due to H_2–H_2 and H_2–He become very strong in the IR and thus block the escape of photons. The trapped energy is then emitted in the optical spectral range, thus making low-metallicity cool atmospheres appear much bluer than solar abundance atmospheres at the same effective temperature.

Figure 5 gives a few more examples of the spectra emitted by cool brown dwarf atmospheres. The figure also gives the location of some important filters as well as detection limits for an object at a distance of 50 pc. The spectra correspond to atmospheres with effective temperatures of 2000 K (young brown dwarf), ~1000 K and 500 K (evolved brown dwarf or gas-giant planet). The broad-band absorption features are due to water vapor and, for lower temperatures, bands of methane (CH_4).

Web update

For the first time, an optical coronal line from a star other than the Sun has been detected. The UV-Visual Echelle Spectrograph (UVES) on the VLT 8.2 m KUEYEN telescope at the ESO Paranal Observatory has observed a coronal emission line in nearby variable star CN Leonis from iron ions that have lost 12 electrons (Fe XIII). CN Leonis, a cool star, is located at a distance of 8 light-years.

Bibliography

The latest reviews of cool stellar atmospheres are

Allard F, Hauschildt P H, Alexander D R and Starrfield S 1997 *Annu. Rev. Astron. Astrophys.* **35** 137–77

Gustafsson B and Jørgensen U G 1994 *Astron. Astrophys. Rev.* **6** 19–65

These reviews are more specific to brown dwarfs and very low-mass stars:

Bessell M S and Stringfellow G S 1993 *Astron. Astrophys. Rev.* **31** 433–71

Burrows A and Liebert J 1993 *Mod. Phys. Rev.* **65** 301

A general collection of formulae and data for stellar atmosphere modeling is given in

Scholz M and Baschek B 1982 Physics of stellar atmospheres Landolt-Börnstein, *Numerical Data and Functional Relationships in Science and Technology* vol V12b, ed K-H Hellwege, K Schaifers and H H Voigt, pp 91–152

Peter H Hauschildt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Atmospheres: Early Type Stars

The topic 'Stellar atmospheres of early type stars' has two individual ingredients, a (stellar) *atmosphere*, and an *early type star*. We shall specify these two terms in turn.

An *atmosphere* of a star is generally any material connected physically to a star from which the photons escape to the surrounding space. In other words, it is a region where the radiation, observable by a distant observer, originates. Traditionally, the term atmosphere has been understood in a limited sense as a thin layer on the surface of a star. However, the more modern view is that the atmosphere encompasses a region extending far from the star. For instance, the solar corona, which extends even beyond Earth orbit, is understood as a part of the solar atmosphere; from this point of view we all live in a stellar atmosphere.

The term *early type stars* is somewhat loose. Although this class could be meant to comprise all stars that are indeed in the early stages of their evolution, the typical use of this term specifies early type stars as massive hot stars, not necessarily young. Using the common stellar classification, the early type stars typically include stellar classes O, B and early A, but also more evolved stars like luminous blues variables (LBV) and Wolf–Rayet (WR) stars.

The fundamental parameters of a star are mass (M), effective temperature ($T_{\rm eff}$) and total luminosity (L). Effective temperature quantifies the total radiation flux emergent from a surface; precisely, it is the temperature that a black body would have if it was radiating the same total energy flux as the unit area of the actual stellar surface. Typical values of these parameters are $M \approx 8$–100 (or more) times the mass of the Sun ($M_\odot$), $T_{\rm eff} \approx 10\,000$–$50\,000$ K, and $L \approx 10^3$–10^6 times the solar luminosity ($L_\odot$).

One of their most significant properties is a strong outflow of matter from their surfaces, reaching velocities up to several thousand km s^{-1}, i.e. of the order of 1% of the speed of light. This outflow is called stellar wind (see RADIATIVELY DRIVEN STELLAR WINDS FROM HOT STARS). Typical values of the mass loss rate from early type stars are $\dot{M} \approx 10^{-9}$–10^{-5} solar masses per year, although in extreme cases of LBV and WR stars the mass loss rate may reach up to $\dot{M} \approx 10^{-3}\ M_\odot\ {\rm yr}^{-1}$. The wind has a profound significance. First, it significantly changes the path of stellar evolution, because the star may actually lose a large part of its initial mass. Second, the mass loss from young, massive stars significantly enriches the interstellar medium in helium and light metals (the enrichment in heavier elements comes from supernova explosions).

In the global astrophysical context, the early type stars have special significance. They physically influence a large space around them. Most importantly, they produce a large portion of ionizing photons for their host galaxy and surrounding intergalactic space (the only competing objects are quasars, which in some cases may provide the bulk of ionizing radiation). As pointed out above, they enrich the interstellar space in He and light metals because of their wind. From the diagnostic point of view, they also have special significance. They are very bright, and therefore may be studied spectroscopically as individual objects in distant galaxies. Reliable model atmospheres for these stars may therefore yield invaluable independent information about distant galaxies, like chemical composition, and, possibly, reliable distances.

Basic physics of stellar atmospheres

There is a separate article on STELLAR ATMOSPHERES, where this term is explained in substantial detail and generality. Here we present a brief summary of the stellar atmosphere theory to the extent that is relevant to stress particular features of early type atmospheres.

From the physical point of view, a stellar atmosphere is generally a plasma composed of many kinds of particles, namely atoms, ions, free electrons, molecules or even dust grains. In an early type stellar atmosphere, because of the high temperature and strong radiation field, there are typically no molecules or dust grains present, at least in the layers that are traditionally considered as an atmosphere. Nevertheless, molecules and dust may still be present in the very remote parts of an atmosphere.

The total particle density ranges from, say, 10^6 to 10^{16} cm^{-3}. It can be shown that under these densities the elastic collisions between particles are frequent enough to yield very nearly a Maxwellian velocity distribution for all particles; moreover with the same associated kinetic temperature. We refer to it as the *electron temperature*, or simply temperature. It should be stressed that this temperature is shared only by massive particles; massless particles, like photons, do not generally possess an equilibrium distribution and thus a corresponding temperature. This point will be discussed in detail later on.

What makes atmospheres of early type stars so special is the fact that because of the strong and energetic radiation field generated in their interiors, the radiation in their atmospheres is not merely a passive *probe* of the physical state of the atmosphere, but rather an important *energy balance agent*. In other words, radiation in fact determines the structure of the medium, yet the medium is probed only by this radiation. Another important feature is that photons have, under the conditions met in the early type stellar atmospheres, a much larger mean free path than massive particles. This means that radiation is able to transport information to large distances; in other words, radiation *couples* the physical states of rather distant regions of the atmosphere.

The other critical feature that follows from the presence of radiation-induced processes is that the medium is prone to departures from thermodynamic equilibrium. Indeed, the simple fact that we do see a star means that photons must escape from the atmosphere (in fact, this is a very definition of an atmosphere). Hence the photons must be missing from the atmosphere, and thus some elementary atomic transition processes can no longer be balanced, which leads to a non-equilibrium situation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

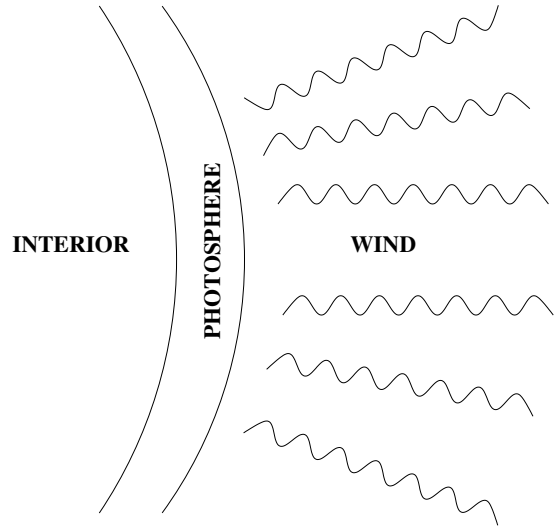


Figure 1. A sketch of basic stellar atmospheric layers. The stellar photosphere is not shown in scale; it would be much thinner.

We shall expand upon this point later. Because of this feature, atmospheres of early type stars have played, and continue to play, an important role not only in the stellar atmosphere theory, but in overall astrophysical radiative transfer theory in general. The whole non-LTE theory (see later) was first developed and tested on early type stellar atmospheres. Consequently, modelling early type stellar atmospheres is a mature field, which may serve as a methodological guide to other astrophysical objects where the radiation also plays an important role, as, for instance, accretion disks in active galactic nuclei (AGN) and in cataclysmic variables (CV), HII regions and others.

Atmospheric layers

Traditionally, an atmosphere of an early type star is divided into two basic regions, schematically displayed in figure 1; the photosphere and the stellar wind.

The photosphere is the innermost part of the atmosphere. The mass outflow velocities are typically very small there, smaller than the local sound speed, so that they can be neglected. The photosphere is thus assumed to be an essentially static region; one of the basic structural equations is thus the hydrostatic equilibrium equation. The radial extent of the photosphere is typically very small compared with the stellar radius, even for hot massive stars.

The stellar photosphere is characterized by the condition of hydrostatic and radiative equilibrium. The hydrostatic equilibrium stipulates that the gradient of the total pressure is balanced by the local gravity acceleration. Because the radial extent of the photosphere is small compared to the stellar radius, the gravity acceleration is essentially constant. The explicit form of this equation is (for a spherically symmetric star)

$$\frac{\mathrm{d}P}{\mathrm{d}R} = -\rho G M_*/R^2 \approx -\rho G M_*/R_*^2 \equiv -\rho g \qquad (1)$$

where P is the total pressure, generally composed of three parts (the gas pressure, P_{gas}, the radiation pressure, P_{rad}, and the turbulent pressure, P_{turb}), R is the radial coordinate, M_* and R_* are the stellar mass and radius respectively, g is the gravity acceleration at the stellar surface and ρ is the mass density.

The radiative equilibrium simply states that the only energy transport mechanism is radiation. In other words, it says that the total radiation energy absorbed in a given elementary volume of material in the photosphere is equal to the total radiation energy emitted in the volume

$$\int_0^\infty (\kappa_\nu J_\nu - \eta_\nu)\,\mathrm{d}\nu = 0 \qquad (2)$$

where κ_ν is the absorption coefficient, η_ν the emission coefficient and J_ν the mean intensity of radiation. (As is customary in astrophysical radiative transfer, a functional dependence on the frequency ν is typographically denoted as a subscript ν standing at the corresponding quantity.) Equation (2) also states that the energy is only *transported* in the photosphere; there is no energy *generated* there. This makes a sharp distinction from the stellar interior (where the energy is generated by nuclear reactions), and from the other atmospheric layers (e.g. stellar chromospheres, coronas or winds), where the energy is generated by dissipation of various wave motions.

It can be easily shown that the atmospheres of early type stars are convectively stable; this feature marks a significant difference from the atmospheres of cooler stars (type late A and later), where convection is another significant mechanism of energy transfer. Roughly speaking, the convection is a transport of energy by rising and falling bubbles of material with properties (e.g. temperature) different from the ambient medium.

The stellar wind is the region where the outflow velocities are comparable to or larger than the local sound speed. The radial extent of this region may be comparable to or, in some cases, significantly larger than, the radius of the stellar photosphere. This region is comprehensively reviewed in a separate article.

In some cases there may be a third, more remote, layer which may contain molecules and dust. For instance, very young stars may still be enveloped by their placental matter. These regions are probed by the infrared and radio radiation (see also STAR FORMATION, INTERSTELLAR MOLECULAR CLOUDS, etc).

Microscopic processes

From the very nature of stellar atmospheres it is clear that the detailed description of the processes of interaction between radiation and matter is a crucial ingredient of the stellar atmospheres theory. These processes determine (a) how the radiation is transported in the atmosphere, and (b) what is the distribution of the microscopic degrees of freedom of the massive particles (e.g. the excitation and ionization state of the individual atomic species, etc).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The interaction between radiation and matter is described phenomenologically through the radiative transfer equation

$$\left(\frac{1}{c}\frac{\partial}{\partial t} + \boldsymbol{n}\cdot\nabla\right) I(\nu, \boldsymbol{r}, \boldsymbol{n}, t) = \eta(\nu, \boldsymbol{r}, \boldsymbol{n}, t) - \chi(\nu, \boldsymbol{r}, \boldsymbol{n}, t) I(\nu, \boldsymbol{r}, \boldsymbol{n}, t). \quad (3)$$

Here, I is the specific intensity of radiation, defined such that it is the energy transported by radiation in a unit frequency range at the frequency ν, across a unit area perpendicular to the direction of propagation, $\boldsymbol{n}$, into a unit solid angle, and in a unit time interval. The specific intensity provides a complete description of the unpolarized radiation field from the macroscopic point of view. (This description can be generalized to an arbitrarily polarized light by introducing the Stokes vector instead of the scalar intensity, but we will not consider this concept in this article, and assume an unpolarized radiation, which is quite appropriate in the context of early type stars.)

Quantities χ and η are phenomenologically defined as absorption and emission coefficients, respectively. They are defined analogously to the specific intensity, namely as the energy removed or added to a beam of radiation at unit frequency range, solid angle, area and time.

It is known from the quantum theory of radiation that there are three types of elementary processes that give rise to an absorption or emission of a photon: (a) induced absorption—an absorption of a photon accompanied by a transition of an atom/ion to a higher energy state; (b) spontaneous emission—an emission of a photon accompanied by a spontaneous transition of an atom/ion to a lower energy state; and (c) stimulated emission—an interaction of an atom/ion with a photon accompanied by an emission of another photon with identical properties. In the astrophysical formalism, the stimulated emission is usually treated as negative absorption.

In thermodynamic equilibrium, the microscopic detailed balance holds, and therefore the radiation energy absorbed in an elementary volume in an elementary frequency interval is exactly balanced by the energy emitted in the same volume and in the same frequency range, i.e. $\chi I = \eta$. Moreover, in thermodynamic equilibrium, the radiation intensity is equal to the Planck function, $I = \mathrm{B}$, where

$$B(\nu, T) = \frac{2h\nu^3}{c^2}\frac{1}{\exp(h\nu/KT) - 1}. \quad (4)$$

We are then left with an interesting relation that in thermodynamic equilibrium $\eta/\chi = \mathrm{B}$, which is called *Kirchhoff's law*.

The absorption and emission coefficients are written explicitly as

$$\chi_\nu = \sum_i \sum_{j>i} [n_i - (g_i/g_j) n_j]\sigma_{ij}(\nu) + \sum_i (n_i - n_i^* \mathrm{e}^{-h\nu/kT})\sigma_{i\kappa}(\nu) + \sum_\kappa n_\mathrm{e} n_\kappa \sigma_{\kappa\kappa}(\nu, T)(1 - \mathrm{e}^{-h\nu/kT}) + n_\mathrm{e}\sigma_\mathrm{e} \quad (5)$$

where the four terms represent, respectively, the contributions of bound–bound transitions (i.e. spectral lines), bound–free transitions (continua), free–free absorption (also called bremsstrahlung), and of electron scattering. In the stellar atmospheric conditions the electron scattering is to a good approximation coherent (i.e. without a change of photon frequency—Thomson scattering). A more general case, with a decrease or increase of the photon energy (frequency), is called Compton or inverse Compton scattering respectively. Typically, the effects of Compton scattering are negligible in the atmospheres of early type stars (they may be important in the atmospheres of very hot subdwarfs, white dwarfs and pre-white dwarfs). In principle, other scattering terms, like for instance Rayleigh scattering, may also be added if needed.

In equation (5) n_i is the occupation number (population) of an atom in the energy level labelled i, g_i the corresponding statistical weight and n_i^* denotes an equilibrium population of level i corresponding to temperature T and electron density n_e. $\sigma(\nu)$ are the corresponding cross sections; subscript κ denotes the 'continuum', and n_κ the ion number density. The negative contributions in the first three terms represent the stimulated emission. There is no stimulated emission correction for the scattering term, because this contribution exactly cancels with ordinary absorption. The relation between the bound–bound cross section $\sigma_{ij}(\nu)$ and the well-known Einstein coefficients for the photoexcitation is $\sigma_{ij}(\nu) = (h\nu_0/4\pi)B_{ij}\phi(\nu)$; $\phi(\nu)$ is the so-called *absorption profile coefficient*, normalized to unity, $\int \phi(\nu)\,\mathrm{d}\nu = 1$. It represents the conditional probability density that if a photon is absorbed in the transition $i \to j$, it is absorbed in the frequency range $(\nu, \nu + \mathrm{d}\nu)$.

Analogously, the thermal emission coefficient is given by

$$\eta_\nu = \frac{2h\nu^3}{c^2}\left(\sum_i \sum_{j>i} n_j (g_i/g_j)\sigma_{ij}(\nu) + \sum_i n_i^* \sigma_{i\kappa}(\nu)\,\mathrm{e}^{-h\nu/kT} + \sum_\kappa n_\mathrm{e} n_\kappa \sigma_{\kappa\kappa}(\nu, T)\,\mathrm{e}^{-h\nu/kT}\right). \quad (6)$$

The three terms again describe the bound–bound, bound–free and free–free emission processes respectively.

The absorption and emission coefficients are thus described through the corresponding cross sections—given by the atomic physics, the local thermodynamic parameters, T and n_e, and the atomic level populations for all the levels involved in the microscopic processes that give rise to an absorption and emission at frequency ν; such a number may be enormous. The chief difficulty of the stellar atmospheres theory is that the level populations generally depend on other state parameters and the radiation field.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

LTE versus non-LTE

It is well known from statistical physics that a description of material properties is greatly simplified if the thermodynamic equilibrium (TE) holds. In this state, the particle velocity distributions as well as the distributions of atoms over excitation and ionization states are specified uniquely by two thermodynamic variables. In the stellar atmospheres context, these variables are usually chosen to be the (kinetic) temperature (T), and the total particle number density (N), or the electron number density (n_e). From the very nature of a stellar atmosphere it is clear that it cannot be in thermodynamic equilibrium—we *see* a star, therefore we know that photons must be escaping. Because photons carry significant momentum and energy, the elementary fact of photon escape has to give rise to significant gradients of the state parameters in the stellar outer layers.

However, even if the assumption of TE cannot be applied to a stellar atmosphere, we may still use the concept of local thermodynamic equilibrium (LTE). This assumption asserts that we may employ the standard thermodynamic relations not globally for the whole atmosphere, but *locally*, for local values of $T(\boldsymbol{r})$ and $N(\boldsymbol{r})$ or $n_e(\boldsymbol{r})$, despite the gradients that exist in the atmosphere. This assumption simplifies the problem enormously, for it implies that all the particle distribution functions may be evaluated locally, without reference to the physical ensemble in which the given material is found. Notice that the equilibrium values of distribution functions are assigned to *massive particles*; the radiation field is allowed to depart from its equilibrium, Planckian, distribution function (i.e. $I = \mathrm{B}$ is valid only in strict TE).

Specifically, LTE is characterized by the following three distributions:

- Maxwellian velocity distribution of particles;
- Boltzmann excitation equation;
- Saha ionization equation.

Microscopically, LTE holds if all atomic processes are in *detailed balance*, i.e. if the number of processes $\mathrm{A} \rightarrow \mathrm{B}$ is exactly balanced by the number of inverse processes $\mathrm{B} \rightarrow \mathrm{A}$. By A and B we mean any particle states between which there exists a physically reasonable transition. For instance, A is an atom in an excited state and B the same atom in another state (either of the same ion, in which case the process is an excitation/de-excitation; or of the higher or lower ion, in which case the term is an ionization/recombination).

In contrast, by the term non-LTE (or NLTE) we mean any state that departs from LTE. In practice, one usually means that populations of some selected energy levels of some selected atoms/ions are allowed to depart from their LTE value, while the velocity distributions of all particles are assumed to be Maxwellian, all at the same local kinetic temperature, T.

One of the big issues of modern stellar atmosphere theory is whether, and if so to what extent, departures from LTE should be included in numerical modelling. Generally, to understand why and where we may expect departures from LTE, let us turn to the microscopic definition of LTE. It is clear that LTE breaks down if the detailed balance in at least one transition $\mathrm{A} \rightarrow \mathrm{B}$ breaks down. We distinguish the *collisional transitions* (arising due to interactions between two or more massive particles) and *radiative transitions* (interactions involving particles and photons). Under stellar atmospheric conditions, collisions between massive particles tend to maintain the local equilibrium (because velocities are Maxwellian). Therefore, the validity of LTE hinges on whether the radiative transitions are in detailed balance or not.

Again, the fact that the radiation escapes from a star implies that LTE should eventually break down at a certain point in the atmosphere. Essentially, this is because detailed balance in radiative transitions generally breaks down at a certain point near the surface. Because photons escape (and more so from the uppermost layers), there must be a lack of them there. Consequently, the number of photoexcitations (or any atomic transition induced by absorbing a photon) is less than the number of inverse processes, spontaneous de-excitations (we neglect here, for simplicity, stimulated emission).

These considerations explain that we may expect departures from LTE if the following two conditions are met: (a) radiative rates in some important atomic transition dominate over the collisional rates; and (b) radiation is not in equilibrium, i.e. the intensity does not have a Planckian distribution. Because the collisional rates are proportional to the particle density, the departures from LTE tend to be small for high densities. Likewise, deep in the atmosphere, photons do not escape, and so the intensity is close to the equilibrium value. Departures from LTE are therefore small, even if the radiative rates dominate over the collisional rates. On the other hand, departures from LTE are important for low-density media immersed in a strong radiation field, which are precisely the conditions met in the atmospheres of early type stars.

Model atmospheres

By the term model stellar atmosphere we mean a specification of all the atmospheric state parameters as functions of position. These parameters are obtained by solving appropriate structural equations, which, in the case of stellar photospheric models, are the equations of hydrostatic equilibrium, radiative equilibrium, radiative transfer equation, and the set of statistical equations (rate equations) for the atomic level populations. In the case of LTE models, the rate equations are not needed because the level populations are given by the Saha–Boltzmann distribution. Because the problem is very complex, it is impossible to find analytical solutions. Therefore, we have to resort to numerical simulations. In order to make the overall problem tractable, one has to make a number of simplifications by invoking various approximations. The quality of an appropriate model, and consequently its applicability to the individual stellar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

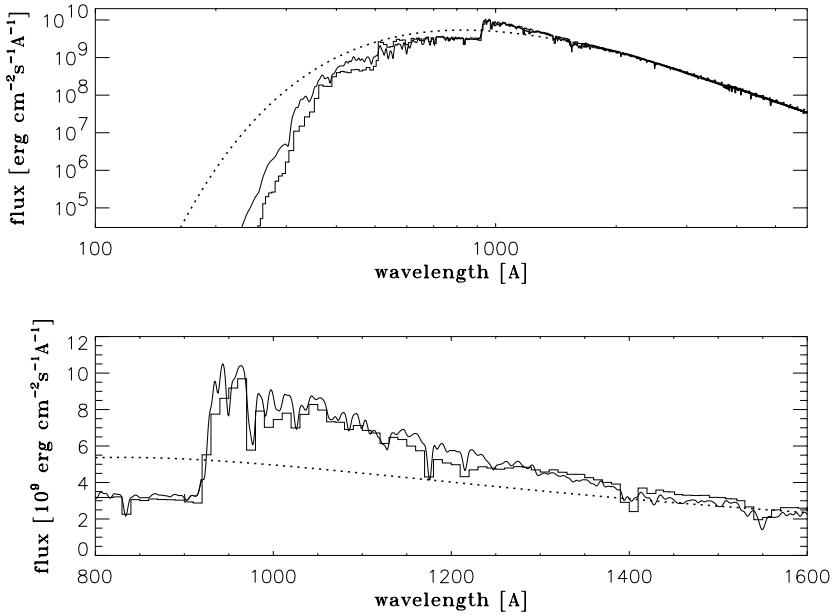


Figure 2. A comparison of the predicted flux from the fully blanketed NLTE model atmosphere with $T_{\rm eff} = 35\,000$ K, $\log g = 4$ (heavy curve); the Kurucz LTE line-blanketed model for the same parameters (thin histogram); and a black-body flux for $T = T_{\rm eff} = 35\,000$ K (dotted curve). The upper panel dispays the flux in a wide range of wavelengths (notice the logarithmic ordinate), while the lower panel shows the flux (in the linear scale) in the vicinity of the Lyman limit. The flux was convolved with Gaussian broadening with FWHM = 5 Å.

types, is closely related to the degree of approximation used in the construction of the model. Needless to say, the degree of approximation critically influences the amount of computational effort to compute it. It is fair to say that the very art of computing model stellar atmospheres is to find such physical approximations that allow the model to be computed with a reasonable amount of numerical work, yet the model is sufficiently realistic to allow its use in reliable interpretation of observed stellar data. The adopted approximations are therefore critical. There are several types of approximations that are typically made during model construction; we shall describe the most important types in turn.

Approximations of the geometry

By geometrical simplification we mean that either some prescribed geometrical configuration is assumed, or some special kind of overall symmetry is invoked. The goal of those simplifications is to reduce the dimensionality of the problem from a spatially three-dimensional problem to a 1- or 2D problem. The most popular approximations are (from simplest to more complex):

(a) Plane parallel geometry, with an assumption of horizontally homogeneous layers. This decreases the number of dimensions to one: the depth in the atmosphere. This approximation is typically quite reasonable for stellar photospheres, which indeed are thinner by several orders of magnitude than the stellar radius, so the curvature effects are negligible. The assumption of horizontal homogeneity is made for the sake of simplicity—there is no plausible verification of this approximation, and, moreover, observational evidence mostly shows that stellar surfaces are far from being homogeneous (a notorious example being detailed pictures of the solar surface). Nevertheless, even in the presence of inhomogeneities, 1D models still have their value since in many cases one may construct different 1D models for the individual 'patches' on the surface.

(b) Spherical symmetry. Again, the problem is one-dimensional. The approach is used for extended atmospheres, for which the atmospheric thickness in no longer negligible with respect to the stellar radius. Typically, we have to consider such models for early type giants and supergiants, as well as for earliest types of main-sequence O stars.

(c) Multidimensional geometry. This field is in its infancy. A numerical solution is extremely demanding on computer time and memory, and only very recently has computer power reached a stage that calculating such models is becoming feasible. Some detailed model atmospheres including 2D and 3D geometry have been constructed for a solar atmosphere, and some 2D simulations of early type stellar winds have been performed.

Approximations of the dynamical state of the atmosphere

This is basically a specification of the realism of the treatment of the macroscopic velocity fields. From the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

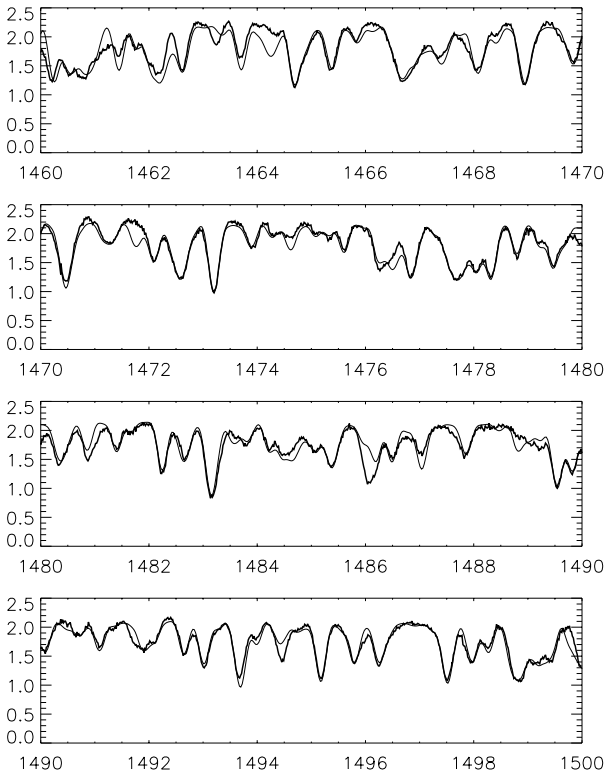


Figure 3. A comparison of the observed HST/GHRS flux for 10 Lac (heavy curve) and the predicted flux from the fully blanketed NLTE model atmosphere with $T_{\rm eff} = 35\,000$ K, $\log g = 4$ and $v \sin i = 25$ km s^{-1}, and for the solar abundances of all species. The predicted flux is convolved with instrumental broadening with FWHM = 0.06 Å. The abscissa is the wavelength in Å, and the ordinate is the flux in 10^{-9} erg cm^{-2} s^{-1} Å^{-1}. Most spectral features are lines of Fe IV, Fe V, Ni IV and Ni V.

simplest to the most complex the approaches are the following:

(a) Static models, in which the macroscopic velocity field is set to zero. As discussed above, these models describe a *stellar photosphere*.
(b) Models with an *a priori* given velocity field. In these models the velocities are taken into account explicitly, and their influence upon other state parameters, in particular the emergent radiation, is studied in detail. In these models, one can either consider only a dynamical region (i.e. the wind) and take the incoming radiation from the photosphere as given *a priori*—the so-called core-halo model—or a model which treats the photosphere and the wind on the same footing. Such models are called *unified models*.
(c) Models where the velocity field is determined self-consistently by solving the appropriate hydrodynamical equations. This problem is very complicated because the wind driving force is given by the absorption of photons in thousands to millions of metal lines (the so-called line-driven wind, so the hydrodynamical equations should be solved together with at least an approximate treatment of radiative transfer in spectral lines.

Approximations of the opacity sources

In real stellar atmospheres, there are an enormous number of possible opacity sources. It is impractical to take all of them into account in full detail. The light elements (H, He, C, N, O) have a comparatively small number of lines per ion (say 10^2 to 10^4) because of a relatively simple atomic level structure. The number of lines generally increases with increasing atomic number, and for the iron-peak elements (Fe and Ni being the most important ones), we have of the order of 10^6 to 10^7 spectral lines per ion! Therefore, the opacity (and emissivity) may be an enormously complicated function of frequency.

There are several approximations that are to meant to reduce this complexity considerably:

(a) Models constructed using certain frequency-averaged opacities; these models are called *grey* models. The approach is based on the implicit assumption that the behaviour of the frequency-averaged intensity of radiation is well described by means of some frequency-averaged opacities. There are several possible mean opacities, depending on exactly how the averaging is done. The most used averaged opacity is the *Rosseland mean* opacity, defined by

$$\frac{1}{\chi_{\rm R}} \equiv \frac{\int_0^\infty (1/\chi_\nu)(\mathrm{d}B_\nu/\mathrm{d}T)\mathrm{d}\nu}{\int_0^\infty (\mathrm{d}B_\nu/\mathrm{d}T)\,\mathrm{d}\nu} \qquad (7)$$

where χ_ν is the opacity (per gram of stellar material). Because averaging is done for $1/\chi$, the largest weight is given to regions of lowest opacity, which are the most efficient regions for the energy flux transport. This explains why the Rosseland mean opacity is well suited for describing the total radiation flux. The grey model atmospheres are no longer used for spectroscopic analysis, but they are useful for providing an initial estimate in any iterative method for constructing more realistic model atmospheres, and they are very useful for pedagogical purposes because they allow one to understand a rough behaviour of temperature and radiation field as a function of depth in the atmosphere.
(b) A possibility is to use stepwise frequency averages for a number of subintervals (frequency bins), sometimes called the multifrequency/multi-grey method. This approach is used in constructing model stellar atmospheres only rarely, but is used in other branches of astrophysical and laboratory radiative transfer.
(c) A completely different approach is to construct a model atmosphere neglecting the line opacity completely. Although this may seem very crude, such models may actually provide reasonable results for very metal-poor stars because, as was pointed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

put above, H and He posses only a small number of lines, which occupy only a very small frequency range and therefore have a small influence on the model structure. (Strictly speaking, this is not completely true, because just three of the most important hydrogen lines, Lα, Lβ and Hα, may already have an important indirect effect upon the temperature structure in the outer layers of early type photospheres.) In any event, this approximation was introduced at the early stages of development of the non-LTE model atmospheres, and was motivated by limitations of the then available computers and numerical techniques.

(d) The obvious next level of approximation is to consider a small number of lines (typically tens to hundreds) explicitly, while neglecting the bulk of metal lines; the selected lines are those which presumably have the largest effect upon the atmospheric structure.

(e) Finally, one can take into account, one way or another, 'all' metal lines. Such models are traditionally called *metal line-blanketed model atmospheres*. The problem of constructing such models is computationally very demanding. Under the assumption of LTE it is, however, considerably simplified because the opacity and emissivity is a function of only local temperature and electron density; the only problem is the complicated frequency dependence of the opacity. Without the approximation of LTE the problem is significantly more difficult because one has to determine all the atomic level populations and temperature self-consistently with the radiation field.

Approximations concerning the thermodynamic equilibria

Here, as we discussed above, the issue is whether the approximation of LTE is adopted or not. If we assume LTE, the resulting model atmospheres are called LTE models. Two state parameters, the temperature, T, and density, ρ (or electron density, n_e), suffice to describe the physical state of the atmosphere at any given position. In practice, LTE models may be useful only for stellar photospheres, because for extended atmospheres and/or stellar winds this approximation breaks down completely and its application would yield erroneous and misleading results.

The models that take some kind of departure from LTE into account are called non-LTE (or NLTE) models. This term is rather ambiguous because it is not *a priori* clear what is actually allowed to depart from LTE in a given model. In early models, the populations of only a few low-lying energy levels of the most abundant species, like H and He, were allowed to depart from LTE; the rest were treated in LTE. During the development of the field, progressively more and more levels were allowed to depart from LTE. The situation is similar for stellar photospheres (static models), as well as for stellar winds and for unified models.

Existing model atmospheres

Because stellar winds are covered in a separate article, we concentrate here on static models, i.e. models of early type stellar photospheres.

The most extensive grid of LTE line-blanketed models is that of Kurucz; the grid covers effective temperatures between 3500 K and 50 000 K, so that it includes stars much cooler than early type stars.

During the last three decades, it has been amply demonstrated that departures from LTE are crucial for spectroscopic studies of early-type stars, even the photospheric layers. Early non-LTE models were constructed as early as the late 1960s and in the first half of the 1970s by Mihalas and co-workers. Nevertheless, the numerical problems and the sheer amount of computer time and memory needed for computing non-LTE metal line-blanketed model atmospheres precluded computing such models until the late 1980s. Thanks to the development of a very efficient numerical method for solving the radiative transfer equation simultaneously with other state equations, called the accelerated lambda iteration (ALI) method (see RADIATIVE TRANSFER), this last barrier to tackling the classical stellar atmosphere problem was broken, and non-LTE metal line-blanketed models including literally millions of spectral lines in non-LTE are now being constructed. Because sufficiently efficient computer codes for mass production of non-LTE line-blanketed models were being developed only during the mid 1990s, there is no comprehensive grid of non-LTE line-blanketed models which would cover the same range of parameters as the Kurucz grid, currently available, but they will very likely be built in the coming years.

From the practical point of view, the most important result of model atmospheres is the prediction of emergent radiation, which is then compared with the observed spectrum in order to deduce basic stellar parameters. Also, theoretical predictions are indispensable for estimating the radiation in unobservable spectrum regions, in particular in the hydrogen Lyman continuum (wavelength less than 912 Å), which produces ionizing photons but which cannot be directly detected for early type stars because of the absorption by interstellar hydrogen. (Only two early type stars, ϵ and β CMa, which are relatively close, and which lie in the direction of a 'tunnel' of low density in the local interstellar medium, have detectable Lyman continuum flux as observed by the EUVE satellite.)

Figure 2 presents a comparison of predicted flux for a star with $T_{\rm eff} = 35\,000$ K, $\log g = 4$, which corresponds to a main-sequence late O type star. The difference between the LTE and non-LTE predictions is important, particularly in the extreme UV region. For completeness, a black-body flux for $T = T_{\rm eff}$ is also displayed. The black-body flux is obviously no longer used for any analysis of spectra of early type stars, but is still sometimes being used in other branches of astrophysics, like cosmology, for estimating the total number of ionizing photons produced by young, massive stars. As this figure shows, such estimates may be wrong by several orders of magnitude.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

As an example of a detailed predicted spectrum from a modern, non-LTE metal line-blanketed model atmosphere, we present in figure 3 a sample of the predicted flux for a non-LTE model for $T_{\rm eff} = 35\,000$ K, $\log g = 4$, and a high-resolution, high signal-to-noise observation of a late O main-sequence star 10 Lac secured by the Goddard High Resolution Spectrograph (GHRS) aboard the Hubble Space Telescope (HST). The agreement between observations and predictions is excellent, and demonstrates the power of the present-day model atmospheres of early type stars.

Model stellar atmospheres are basic tools for analysing observed stellar spectra. By fitting the observed spectrum with a grid of theoretically predicted spectra one can derive the basic parameters used for constructing the models, i.e. the effective temperature, surface gravity, chemical composition, and, in the general case, the mass loss rate. From those parameters, one can derive the fundamental stellar parameters, like the mass, radius and luminosity. Besides these, there are a number of secondary parameters, such as the rotational velocity, or auxiliary parameters describing the nature of atmospheric velocity fields, etc. Detailed techniques for determining those parameters are described in separate articles (CHEMICAL COMPOSITIONS OF STARS, STELLAR DISTRIBUTIONS, LUMINOUS BLUE VARIABLES, STELLAR MASSES, CLASSIFICATION OF STELLAR SPECTRA). Here we just mention that thanks to modern, sophisticated model atmospheres, the typical accuracy of determining the effective temperature of early type stars is about 5%, the accuracy in $\log g$ about 0.1–0.2 dex, and the accuracy of chemical abundances about 0.2–0.3 dex.

Bibliography

The fundamental textbook of the field is

Mihalas D 1978 *Stellar Atmospheres* 2nd edn (San Francisco: Freeman)

A good textbook that covers both the theoretical and observational aspects of the stellar atmospheres is

Gray D F 1992 *Observations and Analysis of Stellar Photospheres* 2nd edn (Cambridge: Cambridge University Press)

Other related textbooks on physical aspects of radiative transfer include

Rybicki G B and Lightman A P 1979 *Radiative Processes in Astrophysics* (New York: Wiley-Interscience)

Shu F H 1991 *The Physics of Astrophysics I. Radiation* (Mill Valley, CA: University Science Books)

An elementary-level textbook is

Böhm-Vitense, E. 1989 *Introduction to Stellar Astrophysics II. Stellar Atmospheres* (Cambridge: Cambridge University Press)

There are two books which present a collection of reviews on various mathematical and numerical aspects of radiative transfer

Kalkofen W (ed) 1984 *Methods in Radiative Transfer* (Cambridge: Cambridge University Press)

Kalkofen W (ed) 1987 *Numerical Radiative Transfer* (Cambridge: Cambridge University Press)

An excellent review paper of properties of early-type stars is

Kudritzki R P and Hummer D G 1990 *Ann. Rev. Astron. Astrophys.* **28** 303

A grid of LTE model atmospheres is described in

Kurucz R L 1979 *Astrophys. J. Suppl.* **40** 1

Some recent non-LTE line-blanketed model atmospheres are described in

Aufdenberg J P, Hauschildt P H, Shore S N and Baron E 1998 *Astrophys. J.* **498** 837

Dreizler S and Werner K 1993 *Astron. Astrophys.* **278** 199

Hubeny I and Lanz T 1995 *Astrophys. J.* **439** 875

Finally, there are several conference proceedings which contain many excellent review papers on the stellar atmospheres theory, e.g.

Crivellari L, Hubeny I and Hummer D G (ed) 1991 *Stellar Atmospheres: Beyond Classical Models (NATO ASI Series C 152)* (Dordrecht: Kluwer)

Heber U and Jeffery C J (ed) 1992 *The Atmospheres of Early-Type Stars (Lecture Notes in Physics 401)* (Berlin: Springer)

Howarth I D (ed) 1997 *Boulder-Munich II: Properties of Hot Luminous Stars (PASP Conf. Ser. 131)* (San Francisco: Astronomical Society of the Pacific)

Ivan Hubeny

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Diameters

The radius of a star is one of several fundamental stellar parameters, including its mass and the luminosity, that characterize the physical and the evolutionary state of a star. A combination of parameters such as mass, luminosity, effective temperature and radius determines the location of a star in the H–R diagram.

Over the past century a number of direct and indirect methods of measuring stellar radii have been developed. The most straightforward method is by measuring the angular diameter ϕ and combining it with the known distance to the star:

$$R = \text{distance} \times \phi.$$

However, because of the very large distances of stars, and in spite of their large size (ranging from $0.01 R_\odot$ to over $1000 R_\odot$), their angular diameters are very small and difficult to measure directly. For example Proxima Centauri, the closest star to the Sun, has an angular diameter of only 0.006 arcsec. Even the largest stellar angular diameters are only a few hundredths of an arcsecond.

Furthermore, it is not easy to define a radius of a star even assuming that the stellar surface can be approximated with a sphere. If the stellar surface was uniformly bright and there was a sharp limb at a distance R from the center of the stellar disk where the intensity drops to 0, then we could say that R corresponds to the radius of the star. However, since stars are gaseous objects their continuum intensity, $I_\lambda(\theta)$, varies from the center ($\theta = 0$) toward the limb ($\theta > 0$), where θ is the angle between the line of sight and the radial vector. Even for relatively compact stars such as our Sun the brightness of the photosphere varies from the center toward the limb. The center-to-limb intensity variation of a solar-type star can be described by the following relation:

$$I_\lambda(\theta)/I_\lambda(0) = 1 - u_2 - v_2 + u_2 \cos\theta + v_2 \cos^2\theta$$

where the average values for the Sun are $u_2 = +0.84$, $v_2 = -0.20$; the dependence on wavelength for $0.20\ \mu\text{m} < \lambda < 10\ \mu\text{m}$ is given in Cox (2000).

In the following we describe several direct and indirect diameter measurement techniques and highlight results obtained from ground- and space-based observations.

Angular diameters

Measurement techniques

Direct measurements of stellar angular diameters can be obtained using several high-angular-resolution techniques including imaging with large-aperture telescopes, lunar occultation and interferometry.

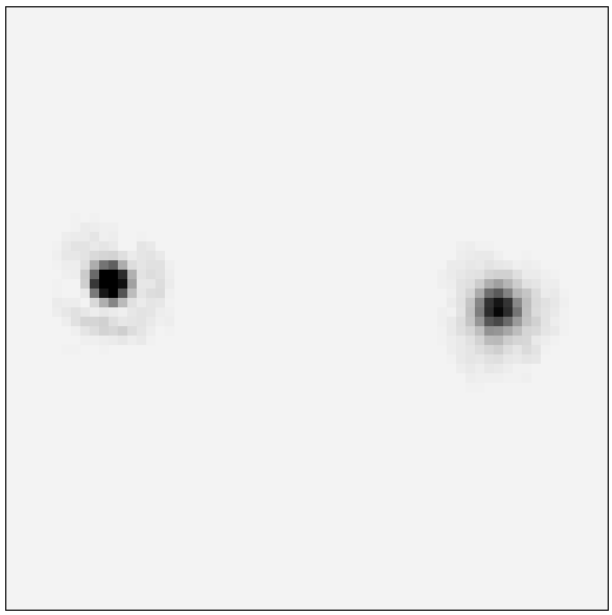


Figure 1. Diffraction-limited HST images of Mira (right) and its companion (left) at 5010 Å.

Large-aperture telescopes. Large-aperture telescopes are necessary to measure stellar angular diameters because the larger the APERTURE, the smaller is the angle on the sky that can be resolved. The size of the aperture of a telescope determines its theoretical diffraction-limited angular resolution at a given wavelength. The resolving power of a telescope increases as the wavelength of observation decreases. For example, to resolve Proxima Centauri at optical wavelengths one needs a telescope with an aperture of about 25 m.

A telescope image of an astronomical source obtained in the absence of atmospheric turbulence (e.g. from space) can be mathematically described as the convolution of the brightness distribution of the source with a blurring function introduced by the limited aperture of the telescope and the finite size of detector pixels. According to the Nyquist criterion, detector pixel size must be less then 1/2 the diffraction limit of the telescope in order to obtain diffraction-limited observations.

As an example of diffraction-limited observations from space, figure 1 shows an image of the Mira AB system obtained using the 2.5 m aperture Hubble Space Telescope (HST) and the Faint Object Camera at 5010 Å (Karovska *et al* 1997). The primary o Ceti (Mira), the prototype of Mira-type variables, is an evolved giant, and the secondary VZ Ceti is a white dwarf. The separation between the components is only 0.6 arcsec (or about 60 AU using a distance to Mira of 100 pc).

The diffraction-limited image of Mira appears substantially larger than the image of the companion and is clearly resolved by the 2.5 m telescope. The companion is unresolved by the telescope and one can discern the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

diffraction ring in its image. The diffraction limit of the HST in the optical is about 50 marcsec (mas). The size of the detector's pixels is approximately three times smaller, 14 mas. The measured size of Mira's atmosphere at this wavelength is 60 mas, which corresponds to approximately 6 AU.

If similar observations of this system were carried out using a 2.5 m ground-based telescope, the binary would have been barely resolved because the images would have been blurred by the atmospheric SEEING. The seeing-limited size of each individual image would have been comparable with the size of the binary separation in the system.

Modern adaptive and/or active optics techniques are used to correct and compensate for the seeing effects which degrade the resolution of the ground-based observations. These techniques allow reaching the full diffraction-limited resolution of large-aperture ground-based telescopes (see INSTRUMENTATION: ADAPTIVE OPTICS).

Lunar OCCULTATION is another way to measure angular diameters in the presence of Earth's atmosphere. The lunar occultation technique uses the diffraction phenomenon occurring as the Moon's sharp limb occults the stellar disk. The resolution in this case is not limited by the diffraction limit of the single-aperture telescope. Several hundred angular diameter measurements have been made so far (Richichi *et al* 1996). This method is limited only to a set of stars that can be occulted by the Moon.

Interferometric techniques using two or more apertures provide another means of improving the resolution in the ground-based observation, and of measuring angular diameters. These techniques are also very promising for future high-angular-resolution observations from space.

Interferometry. In 1920 Michelson and Pease found a way to enhance the resolution of the ground-based observations degraded by the atmospheric seeing effects by using a technique based on the interference of the light from two mirrors in an INTERFEROMETER. The baseline (the separation between the interferometer mirrors) defines the resolution power of an interferometer; at a given wavelength, the resolution power increases as the baseline increases.

At Mt Wilson Michelson and Pease constructed two interferometers (with approximately 7 m and 17 m baselines). They used these interferometers to make the first direct measurements of the angular sizes of stars other then our Sun. For example, they measured diameters of two variable stars, the red supergiants α Ori and α Sco over a period of 17 yr. Assuming uniform brightness distribution on the stellar disk, they obtained diameter measurements of α Ori ranging from 34 mas to 54 mas. For α Sco, the measured diameters range from 28 mas to 40 mas. Other interferometric measurements followed, including diameter measurements of early-type stars using the Narrabri Stellar Intensity Interferometer (Hanbury Brown *et al* 1974).

In 1975, Labeyrie applied a novel approach to eliminate the degrading effects of the atmosphere on the ground-based images. Using a technique called speckle interferometry, which is based on light interference in a single-aperture telescope, he resolved the components of the binary system Capella (separated at only 50 mas) using the Mt Palomar 5 m aperture telescope. The separation between the components of this system is 10–20 times smaller than the size of the direct image of the system degraded by the atmospheric seeing (see also SPECKLE IMAGING OF BINARY STARS).

Labeyrie and his group measured the diameters of several VARIABLE STARS using the speckle interferometric technique, including evolved giants such as Mira and R Leo. The measurements were made assuming a uniform brightness distribution on the stellar disk (Labeyrie *et al* 1977). The results showed that within the same pulsation phase there is a significant variation in the diameter measurements obtained using filters centered in different spectral regions. These variations are explained as due to stratification of the extended atmosphere of these evolved stars.

Today there are more then a dozen ground-based interferometers around the world operating at wavelengths ranging from optical to IR, and at radio wavelengths. Their baselines range from a few meters to several hundred meters (see INTERFEROMETRY: GROUND). Other groups are successfully applying single-aperture interferometric techniques, such as speckle interferometry and non-redundant aperture masking, for high-angular-resolution observations of stars.

Several feasibility studies are at present being carried out for interferometers in space. The beginning of the 21st century may see the first prototype interferometers in space (e.g. SIM; DARWIN), which will open new frontiers in stellar astrophysics.

Examples of angular diameter measurements

A plethora of multiwavelength measurements of the angular size of many stars have been obtained using various high-angular-resolution techniques at wavelengths ranging from UV to radio spectral domain. Recently, diameter measurements have also been made from space using direct imaging with the HST and interferometry with the HST Fine Guidance Sensors (see ASTROMETRY WITH HUBBLE SPACE TELESCOPE). In the following we highlight some of these results.

Giants and supergiants

High-angular-resolution observations of evolved giants and supergiants can provide important information on the structure of their extended atmospheres and circumstellar environments. For example, accurate determination of their photospheric angular diameters is crucial for establishing the effective temperature scale (Ridgway and Joyce 1980). The effective temperature can be determined from the relation

$$T_{\rm eff} = (4F/\sigma\phi^2)^{1/4}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where F is the apparent bolometric flux received on the Earth, ϕ is the photospheric angular diameter and σ the Stefan–Boltzmann constant.

For pulsating stars such as MIRA VARIABLES, an accurately measured photospheric angular diameter when combined with the measured distance to the star yields its linear diameter which can then be used to determine the mode of pulsation (Ostlie and Cox 1986). At present, it is uncertain whether the oscillation mode of Mira-type variables is a fundamental or a first overtone. Once the pulsation mode is determined, it will be possible to better define the evolutionary place of Mira variables and their relation to other types of long-period variables (LPVs) (Willson 1982; van Belle *et al* 1996).

Angular diameter measurements of numerous giants and supergiants have been carried out at wavelengths ranging from the optical to the radio spectral domains. Diameters as small as few milliarcseconds have been measured using interferometric and lunar occultation techniques in optical and infrared. These include for example the high-precision diameter measurements of cool giants at optical wavelengths obtained using the Mk III stellar interferometer (Quirrenbach *et al* 1993) and in the near-infrared using IOTA (van Belle *et al* 1996).

High-angular-resolution observations of evolved giants and supergiants have been carried out in the optical using large single-aperture telescopes (using non-redundant aperture masking technique and speckle interferometry) and lunar occultation techniques (e.g. White 1980; Bonneau *et al* 1982; Haniff *et al* 1995).

Recent observations made using ISI have measured for the first time the diameters of several late-type evolved stars in the mid-infrared spectral domain, at 11 μm wavelength range. They detected dust shells around most of the stars on their list. The measured inner radius of these dust shells ranges from about 2 to several dozen stellar radii. The results of this survey are described in Danchi *et al* (1994).

The sizes of the radio photospheres of several LPVs (Miras and semiregular variables) have been measured using the VLA at the centimeter wavelength range (Reid and Menten 1997). Their observations suggest that these stars have a radio photosphere near 2 stellar radii, where the stellar radius is defined by line-free regions of the optical spectrum.

The multiwavelength diameter measurements were made using models ranging from uniform disk and Gaussian distribution to brightness distributions calculated from various model atmospheres, to estimate the affects of brightness non-uniformity (e.g. Haniff *et al* 1995; van Belle *et al* 1996). The current models often produce a very simplified and usually incomplete description of the physical characteristics of the extended atmosphere because they do not simultaneously incorporate crucial elements such as pulsation, opacities and dust. Stellar size measurements based on inadequate models produce a large scatter in the estimated diameters (e.g. van Belle *et al* 1996) which can bias the calculated effective temperature scale and will affect the accuracy of the derived mode of pulsation.

In addition, current models assume that the atmosphere is spherically symmetric, despite the fact that currently there is substantial evidence that the late-type evolved stars are not symmetric. Recent imaging of several giants and supergiants carried out using ground-based telescopes and the HST detected substantial asymmetries in the extended atmosphere of these stars (Karovska *et al* 1991; Haniff *et al* 1992; Wilson *et al* 1992; Tuthill *et al* 1994; Karovska *et al* 1997). The cause(s) of the observed asymmetries are being explored. The asymmetries could be due to unresolved bright spots on the surface of the star or in the extended atmosphere. They could be related to the pulsation process; plausible mechanisms include instabilities in the pulsating atmospheres and non-radial pulsation.

It is clear that detailed modeling of the extended atmospheric structure is necessary for an accurate determination of the photospheric diameter. Key input parameters that must be measured with high accuracy include diameters in selected spectral regions and the brightness distribution across the stellar disk as a function of wavelength and pulsation phase.

Further progress in understanding the structure of the atmosphere of evolved giants and supergiants depends not only on the results of the multiwavelength interferometric observations but also on the further development of more realistic models incorporating asymmetries, proper abundance information and physical processes such as propagation of shock waves and dust formation. These models should produce phase-dependent brightness distributions in selected spectral regions important for atmospheric structure diagnostics such as different molecular bands, the continuum (UV, optical and especially in the near-IR) and emission lines carrying shock signatures (see also STELLAR ATMOSPHERES).

Cepheids. Among the pulsating stars, Cepheids are most notable because of their importance for establishing the primary distance scale. Currently, the calibration accuracy of the zero point of the distance scale is about 10%. With the advent of long-baseline interferometers operating in the optical and infrared a new opportunity appears to improve the accuracy of the Cepheid distance scale using the Baade–Wesselink (BW) techniques (Sasselov and Karovska 1994). A precise interferometric measurement of the change of the angular radius of a Cepheid, combined with the radial displacement computed from the Cepheid's integrated radial-velocity curve, will allow a direct and very accurate distance determination.

Measurements of mean diameters of several Cepheid variables of less than a few mas are now being carried out using several long-baseline interferometers including the GI2T (Murrard 1997) and NPOI (Nordgren *et al* 1999). Currently, the diameter measurements are made using uniform brightness distribution on the stellar disk. However, Sasselov and Karovska (1994) showed that the differences in model brightness distributions have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

significant effects on the interferometric angular size measurements of Cepheids. They demonstrated that using a uniform brightness distribution will lead to an underestimate of the size of the Cepheid of up to 40% in the continuum and an overestimate of approximately 25% in the strong lines due to high level of scattering above the limb and sensitivity to local dynamic disturbances (shock waves, velocity gradients, etc).

With the ongoing development of very-long-baseline (more than 100 m) stellar interferometers it will soon be possible to measure the small changes of the angular radius as a function of the pulsation process of many galactic Cepheids (see Davis 1994). A combination of accurate interferometric diameter measurements using model-predicted center-to-limb brightness distributions will provide a new venue for determining Cepheid distances via the Baade–Wesselink method (Sasselov and Karovska 1994) and a reliable zero point of the Cepheid distance scale.

Indirect diameter measurement methods

In addition to direct measurements of angular diameters there are several indirect methods that provide measurements of stellar diameters. In the following we highlight these methods.

Eclipsing binaries

Traditionally the most common source of stellar diameters, detached ECLIPSING BINARIES still offer one of the most precise means to measure stellar diameters today. The analysis of their light curves yields complete orbital elements of the binary system, including accurate relative radii. The diameters of the two stars in units of the orbital radius can be determined from the duration of the eclipses provided that all four contacts are well observed (figure 2). If in addition we can also measure the radial velocities of the two stars, we can express their radii in linear units. Such precious double-lined detached eclipsing binaries are very rare. Roughly 45 systems are known and they provide the most accurate (at 1% level) stellar radii in astronomy (Andersen 1991). The binaries in the sample include very massive (O stars), as well as a couple of low-mass stars (M1 V), mostly on the main sequence. Unfortunately, only two red giants are represented. The average mean error of the stellar radii is 1.5%, the best being EK Cep A (0.35%). The main difficulty in determining individual radii for stars in many eclipsing binaries is that the computed light curves may be nearly identical for a wide range of geometric configurations. This does not occur when eclipses are total (or nearly so). Values from partially eclipsing systems should not be trusted to an accuracy of 1–2% unless verified by detailed spectroscopic evidence. Only one such binary system has been studied outside our Galaxy—HV2274 in the Large Magellanic Cloud—and accurate radii derived (Guinan *et al* 1998).

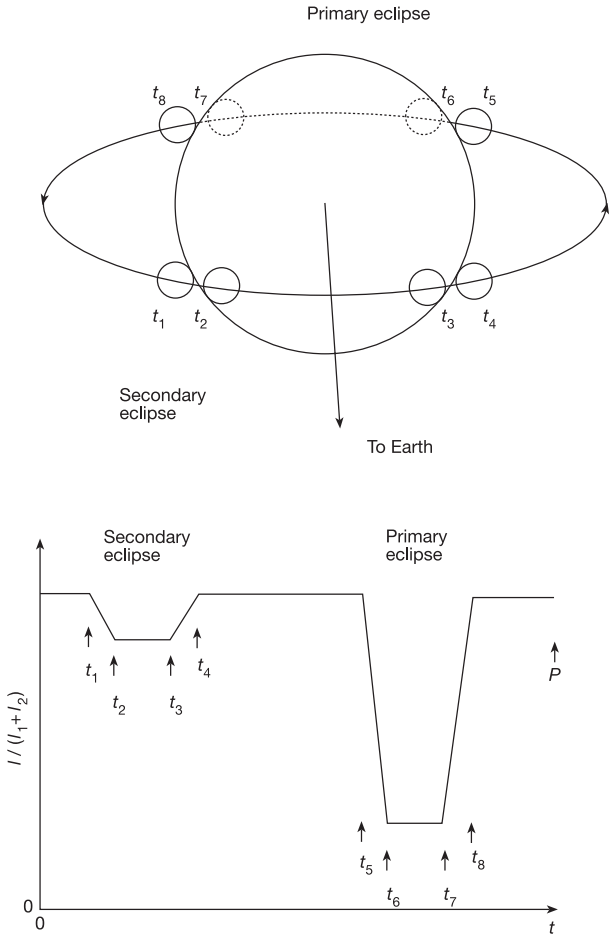


Figure 2. If the orbital plane of a binary happens to be nearly edge-on as seen from the Earth, then ecilpses will occur as one of the components passes in front of the other. The eclipse geometry is shown above, and a schematic light curve for the case in which the smaller star has the greater surface brightness is shown below. The binary period is P.

Stellar radii and the Stefan–Boltzmann law

A traditional and basic, but not very accurate, indirect method of measuring a stellar radius is using the definition of stellar effective temperature in terms of the Stefan–Boltzmann law,

$$L = 4\pi R^2 \sigma T_{\rm eff}^4$$

$$\sigma = 5.6 \times 10^{-8}\ {\rm J\ s^{-1}\ m^{-2}\ K^{-4}}$$

where the assumption is that of a star as a blackbody sphere. With a temperature derived from the stellar spectrum or multicolor photometry, and a luminosity inferred from the apparent brightness of the star and its distance, the stellar radius is determined. The apparent brightness is evaluated from the apparent magnitude measured at the Earth and the distance can be estimated using the parallax measurements or other indirect methods.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A refinement of this method is the infrared flux method (Blackwell and Lynas-Grey 1994) to derive stellar angular diameters and effective temperatures by measuring the monochromatic flux at an infrared frequency and the bolometric flux and using theoretical model atmospheres to estimate the flux at the star's surface. With distances, e.g. from PARALLAXES, the angular diameters thus obtained can be converted to absolute diameters. In a different approach, one could employ theoretical models of STELLAR EVOLUTION, and combined with distances (and for some types of stars) these can provide stellar radii estimates as well.

The accuracy of these techniques is not very high (10–15%) and depends strongly on the type of star involved. For stars similar to the Sun, the accuracy in determining their radii could be as high as 6%.

Baade–Wesselink techniques

Doppler shifts of spectral lines originating in moving photospheres may be used to derive absolute stellar diameters, e.g. in radially pulsating variable stars and expanding supernova ejecta. The method relies on a combination of kinematic and thermodynamic relations and works very well for pulsating stars such as Cepheids and RR Lyrae (or any stellar surface in radial motion). It is generally called the Baade–Wesselink method (Gautschy 1987). The method is based on the idea that an emitting surface in radial motion can be observed directly *and* simultaneously through the Doppler shift in the spectra as well as in flux and temperature variation. The Doppler shifts provide radial velocities, which, when integrated over a certain time (or cycle), yield the linear displacements of the moving stellar surface, ΔR. The flux and temperature variation yield the ratio of the radii of the stellar surface at different times, through the use of the expression for the stellar luminosity, L, and the definition of effective temperature, $T_{\rm eff}$ ($L = 4\pi R^2 \sigma T_{\rm eff}^4$). This approach constitutes the classical application of the BW method to Cepheids (Baade 1926), where for phase pairs of equal color (assumed $T_{\rm eff}(2) = T_{\rm eff}(1)$) we obtain

$$\left[\frac{R(2)}{R(1)}\right]^2 = \frac{L(2)}{L(1)} = 10^{-0.2(m_2 - m_1)} \qquad (1)$$

where m_i is the stellar magnitude. Calculating these ratios and integrating the radial-velocity curve of a Cepheid over its entire pulsation cycle then can give us the mean linear radius of the star. The method has the potential to obtain 3% accuracy, although this potential has not yet been realized.

Gravitational microlensing

The gravitational lensing of a star's light by another distant (point-like) star is called MICROLENSING and can provide a direct measurement of the diameter of the lensed star. When the lensing star transits across the face of the lensed star, the ratio of the angular impact parameter to the angular radius of the lensed star is determined from the shape of the photometric light curve. In this case an astrometric measurement of the Einstein ring radius (with precision of 10 μas) can give the angular diameter of the lensed star with very high precision (1%). Such events are not common but have already been observed (MACHO 95-30). More common are microlensing events in which the lens is a binary star, with caustic crossing the face of the lensed star. Given precise astrometry, accurate diameters can be derived in these events (Paczynski 1998).

Degenerate stars and black holes

The diameters of WHITE DWARFS, NEUTRON STARS and BLACK HOLES are poorly known. The majority of estimates are based on theoretical models and there is very little input from observational data. Combining luminosities with effective temperatures is the most common approach to deriving white dwarf radii. Photometry (color and flux) and model atmospheres allow an estimate of the radiation flux at the white dwarf's surface. With a distance from parallaxes (more than 100 white dwarfs have accurate ones), the radius can be determined. A different approach is afforded by the high gravity at the surface of white dwarfs which is responsible for large gravitational REDSHIFTS (up to 90 km s^{-1}). These redshifts can be measured accurately for a few white dwarfs in detached binary systems, in which the white dwarfs' masses are very well determined. As a result accurate radii for the white dwarfs can be estimated (Lang 1992). An example is Sirius B with $R = (0.0078 \pm 0.0002) R_\odot$. Typical white dwarf radii have a very narrow range around $0.01 R_\odot$.

For neutron stars and black holes certain constraints on their diameters can be placed if they happen to be in accreting binary systems with normal stars as companions. Neutron stars in such systems are often the source of x-ray emission, which is often variable and exhibits bursts. Strong x-ray bursts show photospheric radius expansion which can be used to derive the radius if time-resolved spectra are obtained (Damen *et al* 1990). An alternative method relies on high-resolution x-ray spectroscopy and spectral line broadening (Paerels 1997), much like the traditional technique used in stellar spectroscopy. The latter should become available with x-ray telescopes such as CHANDRA and XMM. In all these techniques the knowledge of the neutron star radius is strongly dependent of the mass estimate. This is especially true in the case of black holes in such binary systems (Menou *et al* 1999). Current estimates of neutron star radii (e.g. Golden and Shearer 1999) are in the theoretically expected range 8–10 km.

Bibliography

Andersen J 1991 *Astron. Astrophys. Rev.* **3** 91–126
Baade W 1926 *Astron. Nachr.* **228** 359
Bester M, Danchi W C, Hale D, Townes C H, Degiacomi C G, Mekarnia D and Geballe T R 1996 *Astrophys. J.* **463** 336

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Blackwell D E and Lynas-Grey A E 1994 *Astron. Astrophys.* **282** 899
Bonneau D, Foy R, Blazit A and Labeyrie A 1982 *Astron. Astrophys.* **106** 235
Cox A N (ed) 2000 *Allen's Astrophysical Quantities* (New York: Springer)
Damen E *et al* 1990 *Astron. Astrophys.* **237** 103
Danchi W, Bester M, Degiacomi C G, Greenhil L J and Townes C H 1994 *Astron. J.* **107** 1469
Davis J 1994 *Very High Angular Resolution Imaging* ed J G Robertson and W J Tango (IAU)
Gautschy A 1987 *Vistas Astron.* **30** 197
Golden A and Shearer A 1999 *Astron. Astrophys.* **342** 5
Guinan E F *et al* 1998 *Astrophys. J. Lett.* **509** L21
Hanbury Brown R, Davis J and Allen L R 1974 *Mon. Not. R. Astron. Soc.* **167** 121
Haniff C A, Ghez A M, Gorham P W, Kulkarni S R, Matthews K and Nagebauer G 1992 *Astron. J.* **103** (5) 1667
Haniff C A, Scholz M and Tuthill P G 1995 *Mon. Not. R. Astron. Soc.* **276** 640
Karovska M, Hack W, Raymond J and Guinan E 1997 *Astrophys. J. Lett.* **482** L175
Karovska M, Nisenson P, Papaliolios C and Boyle R 1991 *Astrophys. J.* **374** L51
Labeyrie A, Koechlin L, Bonneau D, Foy R and Blazit A 1977 *Astron. Astrophys.* **218** L75
Lang K R 1992 *Astrophysical Data: Planets and Stars* (New York: Springer)
Menou K *et al* 1999 *Astrophys. J.* **520** 276
Murrard D 1997 *Astron. Astrophys.* **317** 789
Nordgren T E *et al* 1999 *Astron. J.* **118** 3032
Ostlie D A and Cox A N 1986 *Astrophys. J.* **311** 864
Paczynski B 1998 *Astrophys. J.* **494** L23
Paerels F 1997 *Astrophys. J. Lett.* **476** L47
Quirrenbach A, Mozurkewich D, Armstrong J T, Buscher D F and Hummel C A 1993 *Astrophys. J.* **406** 215
Reid M and Menten K 1997 *Astrophys. J.* **476** 327
Richichi A, Baffa C, Calamai G and Lisi F 1996 *Astron. J.* **112** 2786
Ridgway S T, Joyce R R, White N M and Wing R F 1980 *Astrophys. J.* **235** 126
Sasselov D and Karovska M 1994 *Astrophys. J.* **432** 367
Tuthill P G, Haniff C A and Baldwin J E 1994 *Very High Angular Resolution Imaging* ed J G Robertson and W J Tango (IAU) p 395
van Belle G T, Dyck H M, Benson J A and Lacasse M G 1996 *Astron. J.* **112** 2147
White N M 1980 *Astrophys. J.* **242** 646
Willson L A 1982 *Pulsating Classical and Cataclysmic Variable Stars* ed J P Cox
Wilson R W, Baldwin J E, Busher D F and Warner P J 1992 *Mon. Not. R. Astron. Soc.* **257** 369

Margarita Karovska and Dimitar Sasselov

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Distributions

Our Galaxy is a collection of stars moving in orbits through the Galaxy's gravitational field (see GALAXY: STRUCTURE). A 'stellar distribution' is a description of the distribution (in real space and in velocity space) of the stars in our Galaxy. Different kinds of stars have different distributions. A knowledge of the distribution of stars, obtained from counts of stars of many kinds, leads to an understanding of the structure and dynamics of groups of stars in general and to a description of our Milky Way Galaxy in particular. It also leads to interesting conclusions about the origin and history of our Galaxy.

Censuses of stars began some two centuries ago with star counts; that is, numbers of stars at each discernible level of apparent brightness. It is easy to see that fewer bright stars are seen than fainter ones; with each fainter magnitude the number of visible stars triples. However, what are the reasons and to what faintness does this observation extend? In the 20th century, after it had been discovered that stars vary greatly in both intrinsic brightness and distance, distance estimates of large numbers were needed in order to distinguish between the stars that happened to be near our solar system and the truly luminous stars and to distinguish the faint cut-off in star numbers, because a minimum brightness to stars is known to exist and because the Galaxy stops at a certain distance from us.

Stellar properties such as brightness, color and distance must be carefully evaluated for many thousands of stars for this separation to be made with precision. From such data, and from stellar motions, one can derive the underlying structure and dynamics of our Galaxy. An understanding of stellar distribution and galactic structure therefore requires the accurate determination of these stellar parameters for all of the principal star types: the main sequence or dwarf stars, the giants and the supergiants and, lately, the faint sub-main-sequence brown dwarfs as well.

Background

In 1543, Copernicus proposed and published the concept that the Sun rather than the Earth lay at the center of the solar system. Soon thereafter, as his heliocentric system gained adherents, it became accepted that the stars were not just beyond Saturn, then the farthest known planet, but were many many times its distance from the Sun, for they showed no observable parallax as they must if the Earth moves.

Research into stars and their distribution in space remained dormant until Sir William Herschel, discoverer of Uranus in 1781, initiated the study of their arrangement. Herschel made systematic and extensive counts of stars by magnitude and found that their density in space decreased in all directions with distance from the solar system. Along the plane of the Milky Way the decline was gradual, but in the perpendicular directions it was more precipitate. He was led to a model of the distribution of stars as a disk—a grindstone-shaped universe of stars with the Sun near its center.

The 19th century saw little progress in this field, but it witnessed the development of photography, spectroscopy and other tools that would later be used to expand the field of stellar astronomy. Above all was the successful measurement of the first stellar parallaxes in 1838 and distances to the stars became known directly from these heliocentric parallaxes. This stellar motion, reflecting the orbital motion of the Earth about the Sun and constituting proof of that motion, had been sought by Aristotle and many others since. Only after 1838 could distances to stars be individually assigned. With parallax and proper motion (the apparent angular motion of a star across the sky) the true stellar motion could be derived and galactic dynamics could begin. By the end of the 19th century parallaxes were known for about 70 stars. (See PARALLAXES.)

Around 1900, the Dutch pioneer J C Kapteyn initiated the modern period of the study of stars and their distribution in space. Concurrently, Frank Schlesinger at the Yerkes Observatory introduced photography into astrometry and quickly increased the number of known parallaxes to over a thousand. Shortly before World War I, E Hertzsprung and H N Russell used these many hundreds of new parallaxes to discover the characteristics of the distribution of stars by surface temperature (as deduced from spectral type or color) and intrinsic luminosity. Their representation of this distribution is now called the Hertzsprung–Russell (HR) diagram, or in a more common form, the color–magnitude (CM) diagram. The astrophysics of its features was not explained for many years, but the main sequence and the giant and supergiant domains gave form to further stellar study and classification. In whichever form, the diagram represents a plot of intrinsic luminosity of stars versus their surface temperatures; it and the loci of the major star groupings are shown in figure 1.

One of the features that arose in the wake of this diagram was the stellar luminosity function, the frequency distribution of stars ordered by intrinsic brightness (absolute magnitude). It became possible to give a quantitative description of the distribution of star luminosities as they really are, as an aid in deducing spatial distribution of stars from their observed distribution. Surveys were created that distinguished the nearby fainter stars by their proper motion from their numerous distant and brighter counterparts. A quantitative description of the distribution of nearby stars in space could be provided as an aid in deducing the distribution of star luminosities from their observed distribution. Clearly these two developments have to meet in a consistent picture of the distribution of stars.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

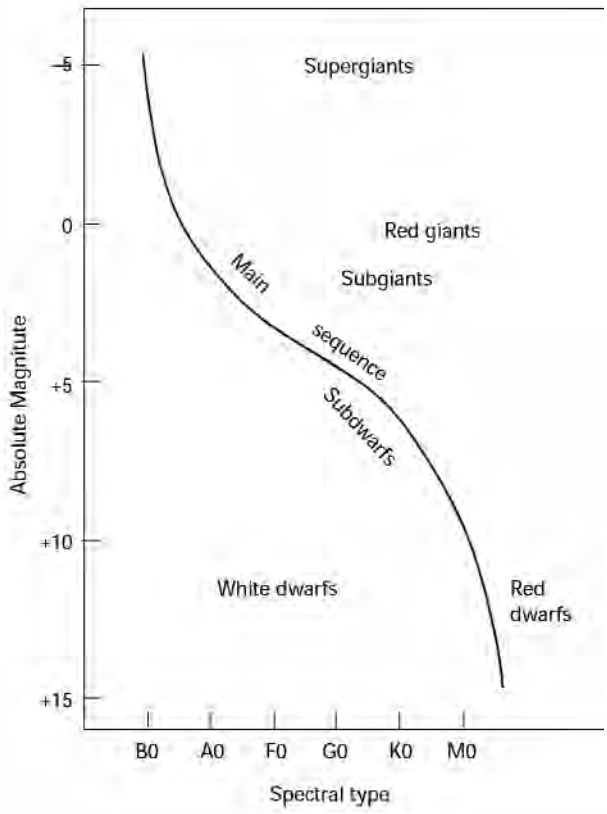


Figure 1. The HR diagram shown schematically. The regions inhabited by the main sequence stars (normal dwarfs), giants and supergiants are shown. (Upgren A R 1998 *Night has a Thousand Eyes: A Naked-Eye Guide to the Sky, its Science and Lore* (New York: Plenum))

The first of these developments resulted from the studies at Leiden by Kapteyn, P J van Rhijn and their associates and the second was due to W J Luyten among others. These astronomers realized that apparent magnitudes and proper motions could be determined for stars by the thousands. It is worth noting that, until very recently, these were the only properties known for the majority of even the nearest stellar neighbors. In fact, their faster proper motion became the single stellar parameter by which most faint nearby stars could be recognized as such.

Astrometric surveys based only on apparent brightness and proper motion were known even then to be fraught with biases and systematic errors, but for the middle third of the 20th century they formed the observational data for the distribution of stars in space. Much ingenuity was demonstrated in the avoidance of the worst pitfalls due to incomplete counts of even the nearest stars. It was recognized that, for a full understanding of stellar distribution, the three spatial coordinates and their three first derivatives with time must be well known for every kind of star. Expressed in a coordinate system of a radial (or line-of-sight) direction and two transverse directions (the two orthogonal axes in the tangent plane of the sky at the position of the star), all but one of the six derive from the equations of condition for stellar motion. The distance is set by the parallax with the direction cosines appropriate for the star's position in the sky, and the two mutually orthogonal transverse velocities, $T(x)$ and $T(y)$, are derived from the proper motion in each of the two directions divided by the parallax. Only R, the velocity in the radial direction, remained to be found. The problem then was that this radial velocity could only be obtained for bright stars, and the distances and transverse velocities were only available for nearby stars. Far too few stars made both lists for an accurate picture of stellar distributions and motions to be made.

Recent results

The last few decades have been a time of rapidly improving stellar samples and a more faithful representation of the stars in general. Photoelectric photometry, which allows stellar brightnesses to be measured with high precision, has become an alternative method of establishing a star's approximate distance, with greater rapidity and extended to more distant stars than the measurement of stellar parallax could achieve. Together with the stellar spectrograph, photometry is able in most cases to locate a star's proper position on the HR diagram and thus its distance. Larger telescopes with more sensitive equipment have extended star counts to much fainter stars. The newer surveys have been active particularly in the infrared spectral regions, where the faintest stars are more visible.

Although several thousand stars had parallaxes by 1989, only a minority of them had parallaxes sufficiently larger than their errors to be of any real value in setting an adequate distance scale. Then, in that year the European Space Agency launched Hipparcos, a satellite devoted to positions, parallaxes, and proper motions. The mission was first conceived by Pierre Lacroute and the acronym stands for 'high precision parallax collecting satellite'. The name also honors Hipparchus, the astronomer who first catalogued the stars and their brightnesses and positions in the 2nd century BC. On completion of the mission about 1993, it delivered parallaxes and proper motions of almost 120 000 stars; most of the data were enough larger than their errors for galactic research. Hipparcos allowed the precise calibration of the luminosities of all but the intrinsically brightest stars. Only the supergiants, Cepheid variables and their like do not yet have reliable luminosities from their individual parallaxes because even the closest among them are still too far away.

Hipparcos, the Hubble Space Telescope launched by NASA at about the same time and other large ground-based telescopes have revolutionized the field, having

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

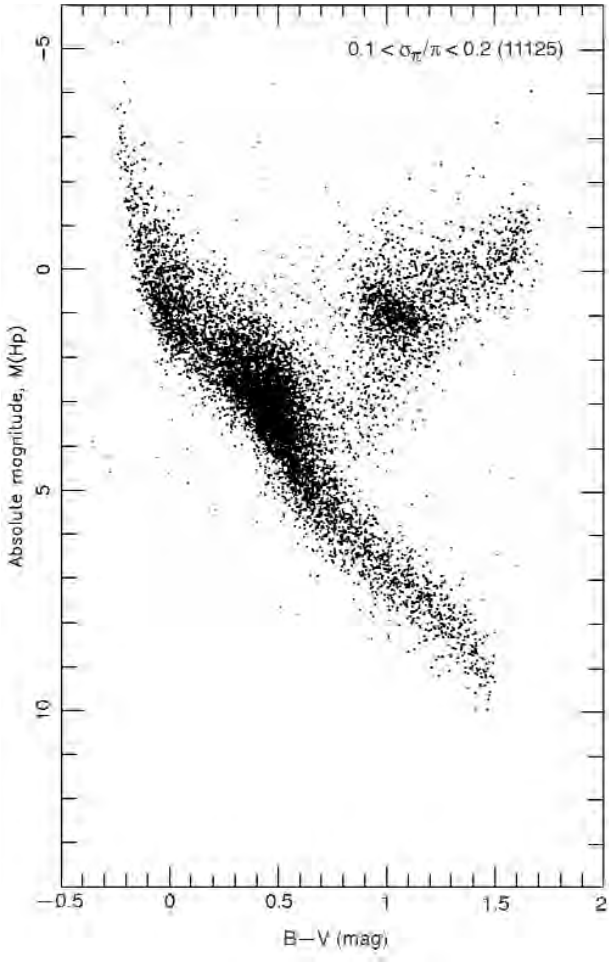


Figure 2. The CM diagram for 10 793 stars with accurate parallaxes determined by the Hipparcos satellite. Notice the main sequence and giant regions. Only one supergiant (Canopus) is close enough for a good parallax. (Perryman M A C *et al* 1995 Parallaxes and the Hertzsprung–Russell diagram from the preliminary Hipparcos solution H30 *Astron. Astrophys.* **304**, 69–81)

provided the means for surveys, particularly in the infrared, where the faint stars and brown dwarfs are most detectable. The potential from the data in the final catalogue of Hipparcos results is not yet likely to be fully exhausted.

Two programs, now in the planning stage, will further greatly improve the precision in stellar distances and motions. Both make use of large telescopes in space and interferometric techniques and, if successful, can achieve errors as low as a microsecond. They could then produce a reliable parallax for almost every resolvable object in our Galaxy as well as some in the Magellanic Clouds. One of these is GAIA (Global Astrometry Instrument for Astrophysics), a program of the European Space Agency and named for a goddess of the Earth. Another is known as SIM (Space Interferometry Mission), an American program. At last notice, both efforts are a decade or more away from launch. In the just-published second edition of his book, Modern Astrometry, Professor Jean Kovalevsky outlines in detail Hipparcos, as well as these and other future projects vital to the further understanding of the arrangement of stars.

Figure 2 shows a CM diagram for several thousand stars plotted from Hipparcos data of high precision. We see that the majority of stars form a sequence extending from the upper left (hot and bright) to the lower right (cool and faint). This is the main sequence, and it is now known to represent a sequence in the masses of stars as well as their luminosities. The sequence extending from it to the right near the center of this main sequence is known as the giant branch, and the few stars in the lower left (hot and faint) are the degenerate stars or white dwarf stars. Just above the top of the figure would be the supergiants, the brightest of all stars (see figure 1 for a schematic representation). The features of the CM diagram must be inferred from a large amount of data.

Current models of stellar distribution

A broad picture of stellar distributions emerged early in the 20th century and has been substantiated in its basic features and improved by more recent observations. Star counts made in all directions reveal that the one or two thousand brightest stars, those readily visible to the naked eye, maintain a reasonably uniform distribution all around the sky. The luminous stars of the Milky Way delineate its plane, seen edge on from within, of our great spiral Galaxy. They thin out very rapidly with angular distance from this plane, represented by a great circle called the Galactic Equator centered within the visible Milky Way. Distant stars are relatively rare anywhere near the directions perpendicular to the Equator, known as the Galactic Poles. The major component of the Milky Way is thus a great disk some 100 000 light-years or more in diameter but only a few hundred to a thousand light-years in thickness. The disk is centered on a much denser core or nucleus and is surrounded by the thinly populated galactic halo.

Stars fall into two or more main groups or populations, with fundamental differences between them in their chemical constitution and location in the Galaxy. The primary difference is one of age; that is, time since formation. The globular clusters and a few other stars form a nearly spherical halo about the disk and contain the oldest stars, whereas the younger stars are more confined to the disk region. Recent work indicates the presence of a division within the disk population into young 'thin-disk' stars and an older middle-aged group, the 'thick-disk' stars. Each group is distributed roughly normally about the plane along the perpendicular or z-axis direction. The dispersions of the groups are not precisely known and are found to differ from one investigation to another, but are of the order of 100–200 pc for the thin

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

disk and 500–1000 pc for the thick disk. The very old 'halo' population is more nearly spherically distributed with a dispersion in z of several thousand parsecs.

Within the thin-disk population, some evidence suggests fine structure in the form of 'moving groups' of stars, not unlike widely extended open star clusters and possibly spawned by those that have dissipated into the general stellar background. Such star streams range from the proven (the closest of which is the so-called Ursa Major or Sirius moving group, counting among its brighter members Sirius and most of the stars forming the Big Dipper (or Plough)) to the very tentative. The number of groups and their percentage of the total stellar population are not yet known. Galactic dynamics indicates that stars older than about a billion years are well mixed and would show little star streaming or other preferential motion.

Distributions in surface temperature, luminosity and metal abundance

It is now widely recognized that stars vary over a great range in three distinct parameters: luminosity, surface temperature and an age-related parameter measured by the abundance of metals in their atmospheres. This is one of the outstanding achievements in galactic structure and kinematics of recent years. After distances became known through the parallax process, the variation due to distance could be removed from consideration and some order could be made among the kinds of stars we really see. In recent decades, the nature of stars was made clear by means of plots of their spectral types against their intrinsic luminosities (absolute magnitudes) in the HR and CM diagrams. Because the spectral type was found to correlate with the colors of stars, both diagrams measure the basic stellar properties of surface temperature and luminosity. The two properties depend overwhelmingly on the star's initial mass. Recently the third dimension related to chemical composition has been well surveyed. We have found that the commonest measure of chemical composition is the abundance of metals present. We know that the Milky Way has become enriched in metals from its early days when almost all matter was in the form of hydrogen. As stars evolved and died, the metals produced in their cores dissipated throughout the system and now amount to about 3% of younger stars by mass. In its early days the Milky Way collapsed from a near-sphere onto a disk, much as did the matter in the solar system later on. Correlations were discovered between the age of a star, its metal content and its galactic orbit. The metal-poor stars were formed first when the Galaxy was more nearly spherical and the others came along much later.

The main sequence contains the great majority of stars, and we now know that all stars start out on the main sequence in a location determined by stellar mass. The giants and supergiants have experienced major changes in their interiors, which cause them to become cooler and brighter; thus, they have become much larger in diameter and of reddish appearance. The main sequence stage accounts for most of a star's total lifetime and, since more massive stars are far more luminous than smaller dwarfs, they exhaust their energy and pass through this stage much more quickly. The thermonuclear fusion processes by which hydrogen is converted into helium and energy at and near the core is the source of energy that drives the star during this stage. Supergiants last only for a few million years, whereas the red dwarfs forming the lower main sequence can endure for tens of billions of years, longer than the age of the Galaxy. We have a picture of the stars not unlike a single snapshot of a forest, from which we must infer the types of trees and shrubs, and, from saplings and mature plants, the growth and life cycle of each species.

A star is a furnace composed of its own fuel. In the interiors of massive bright stars the energy derives from their convective cores passing up through radiative envelopes. In intermediate-mass stars such as the Sun, the two means of energy transport are reversed, with the energy passing outward from a radiative core through the outer layers forming the surrounding envelope via convection. As we descend in mass farther down the main sequence, we encounter the faint red dwarf stars, the predominant stellar component of the galaxy and presumably the universe. Energy transport in stars below 0.3 solar masses seems to be fully convective. All such small stars ever formed are still alive, since their energy exhaustion rate is low. At about 0.08 solar masses, we reach the limit below which the central temperature does not reach the point at which the fusion of hydrogen into helium can begin, and the star cools off after a short period of luminescence through gravitational contraction to become a brown dwarf. The procession on down in mass from 0.08 to 0.001 solar masses brings us to the mass of Jupiter, our largest planet.

In the last few years with larger telescopes and subtler ancillary photosensors now available, astronomers have delved into these less massive objects in the galaxy. Ten years ago we knew of no object outside the solar system less massive than a full-fledged star, defined as an object with fusion going on in its core. In order for this to take place, a central temperature of at least three million kelvins must be maintained. Because core temperatures of stars rise with gravitational pressure, which in turn rises with stellar mass, stars as they contract from interstellar material must be above a minimal mass in order for the thermonuclear fusion, the conversion of hydrogen into helium plus energy, to begin.

Less massive stars, these almost-stars, are now known as brown dwarfs despite their red color, since the term 'red dwarfs' has been reserved for the faintest stars

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

that do shine by fusion processes. Brown dwarf stars were presumed to exist for some time, but their confirmation came only in the last few years with improved sky surveys. Curiously, planets orbiting other stars were also proven to exist at about the same time. Through gravitational perturbations affecting nearby stars, we are now aware of planets within an order of mass of Jupiter orbiting other stars; that is, masses from about one-half to about ten times the mass of Jupiter, far smaller than the brown dwarfs. Jupiter-sized masses have been detected around some 50 stars similar to the Sun since 1995. Jupiter and these other planets never were and never will be stars. However, they are now a part of a continuous distribution in mass of the common objects in our galaxy, and their abundance has become a matter of interest for the future. This is because smaller brown dwarfs have been found that are only a little more massive than the largest extrasolar planets, and soon the final gap in mass may be filled in, thus providing a continuity in mass from the smallest debris in the solar system to the mightiest of stars.

It is yet too early to know with any certainty the space density of brown dwarfs or of planets. A recent estimate places the number of main sequence stars of spectral classes F, G and K (between about 0.5 and 1.5 solar masses) with one or more planets at about 7% of their total number. The planets we have detected are of the size of Jupiter and Saturn, but it is widely assumed that Earth-sized planets are also present in large numbers and that we may be able to discover some of them in the next decade.

Metallicity in stars

With the recent advent of the measurement of metallicity—the quantity of elements heavier than helium—in stars, we are able to put together a much more coherent picture of the stars and the development of the Galaxy. To an astrophysicist, the term 'metals' includes all elements except the lightest two, hydrogen and helium, but in fact only carbon, nitrogen, oxygen, iron and a few in between have noteworthy abundances. Elements heavier than iron are found only in trace quantities if at all.

The abundance is defined by the ratio between a star and the Sun of the percentage of metals, sometimes specifically limited to iron alone, compared with hydrogen (which is known to account for the great majority of stellar material) and is specified as [M/H] where

$$[\mathrm{M/H}] = \log[(\mathrm{M/H})_{\mathrm{star}}/(\mathrm{M/H})_{\mathrm{sun}}]$$

or, in the case of iron alone,

$$[\mathrm{Fe/H}] = \log[(\mathrm{Fe/H})_{\mathrm{star}}/(\mathrm{Fe/H})_{\mathrm{sun}}]$$

The Sun is a rather young metal-rich star, and others with the same chemical composition will show [Fe/H] near zero. Old stars such as those in globular clusters reveal [Fe/H] to be near –2; that is, iron is down by a factor of 100 compared with its relative abundance in the Sun. Stars of intermediate age lie near –1 (one-tenth the solar iron abundance). Although the parameter has not been extended to enough stars to properly map the prevalence of metals (iron) throughout the Galaxy, or even the solar neighborhood, studies made in connection with the kinematics of stars have revealed a calibration of stellar populations into two or, more likely, three stellar populations. The first group is comparatively rich in metals ([Fe/H] near zero, the solar abundance) and is associated with the thin disk, extending to only a few hundred parsecs on either side of the galactic plane. The second group, corresponding to the thick disk, is of intermediate age and has a metallicity near –1, with faster motions perpendicular to the plane in the so-called z direction, and thus they carry to about a kiloparsec above and below the plane. The reality of this thick disc is still in some doubt—the evidence is strong, if not yet conclusive.

The oldest stars are of metallicity –2 or less and form the galactic spheroid or halo. They consist of globular clusters and their component stars, along with a few old field stars, and fill a nearly spherical region with a diameter equal to the disk and a thickness above and below it of several kiloparsecs. In the plane, the halo stars make up less than 1% of the total, and the intermediate stars about one-quarter of the total. We find that the metal-rich stars move slowly in the z direction, with an average velocity of 12 km s^{-1} at the plane ($z = 0$), so that they move to only one or two hundred parsecs above and below the plane before the gravitation of the material in the plane pulls them back into it. The intermediate stars average near 30 km s^{-1}, so that with increasing z distance they take over. Then come the halo stars, which become the majority in regions more than 1 kpc from the plane.

Stellar samples and their biases today

With the newer generation of large telescopes and charge-coupled devices and other detectors of advanced sensitivity, we are now aware that stars do not keep tripling with each additional fainter magnitude forever. At a magnitude of about 12, about the faintest magnitude visible in a telescope of aperture 6 in, the newer data show that fainter stars fall off more rapidly. At some point this must occur, since our Galaxy is not infinite. Its several hundred billion stars must run out at some point. Work has also been done to improve the stellar luminosity function, both as to its fainter end and with regard to the relative space densities of the brown dwarfs. Since the brown dwarfs shine for only a limited time before going dark, the abundance of visible and no-longer visible brown dwarfs is unknown and may be quite large.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The stellar population increases markedly with faintness along the main sequence. The faint red dwarfs dominate the stellar component in mass, although not in luminosity. Each brilliant supergiant may be a billion times as luminous as a faint dwarf, even if millions of times less numerous. Since most of the stellar mass is in the form of small red stars and possibly also brown dwarfs—including those that no longer shine—it becomes very important to know just how many there are in a given volume of space. Such information is vital to any study of the mass distribution and dynamics within our Galaxy. Selection effects abound in this effort, since identifying dwarfs from the many background stars that appear just as faint because of their greater distance involves detection by proper motion or by spectral or color characteristics. Now that surveys singling out the faint stars are not limited to astrometric means, and also extend from the visible spectral region into the infrared, this work is proceeding apace.

The determination of an accurate stellar luminosity function requires care in both the observations and the underlying theories and the story of stellar distribution is largely one of ever-improving observational techniques and statistical analyses applied to ever-increasing numbers of stars of all kinds. At this time, the census of stars is known very accurately only to a distance of about 5 parsecs (about 16.3 light years) from the Sun. Table 1 lists the number of stellar systems and of individual stars known to be within that distance as a function of the year of publication of a reference to those data. Note that the numbers of known stars are always increasing over time but appear to be nearing a constant. Any extension of the distance beyond 5 parsecs quickly shows a decline in space density, first for the faintest red dwarfs, but then for stars of ever-increasing luminosity.

Table 1. The numbers of stars known within 5 pc of the Sun.

Year of publication	*Number of stars*	*Number of systems*
1945	48	37
1957	56	42
1974	60	45
1981	63	47
1987	65	50
1993	67	51

Stars and the dynamics of the Galaxy

Half a century ago, it became clear that the disk of our Galaxy is not homogeneous throughout its extent, but rather resembles many other large galaxies in possessing a spiral structure, and lately possibly a barred structure as well. Objects known to be very young include the supergiants, especially the blue ones at the very top of the main sequence, the Cepheids and the HII regions or emission nebulae. These were the objects found to delineate spiral arms in other galaxies, and studies of their distribution in this one confirmed the presence of spiral arms here. When we consider the loci of even moderately older stars, such as the main sequence A stars like Vega and Sirius, the arms are no longer recognizable among the background stars. Star formation regions lie along the arms, but the stars produced gradually drift away and diffuse throughout the disk between the arms.

The total mass density of all matter is a number of great significance. Galactic dynamics depends on the mass distribution within the Galaxy as a function of location. J H Oort showed that the distribution of motions of stars in the direction normal to the galactic plane is governed by the total mass density from all matter in the galactic plane in the general region near the Sun. Stellar motions here and at some distance from the plane are not known with high precision, but they indicate a mass density of between 0.15 and 0.18 solar masses per cubic parsec. This number is now called the Oort limit. The observed mass density in the region of the Sun summed up over all known stars and interstellar material does not seem to reach the Oort limit. It may be about 0.10 solar masses per cubic parsec, with the interstellar gas and dust accounting for one-third to one-half of the total. The rest, the so-called missing mass, is presumed to exist in the form of dark matter. None of these numbers is of sufficiently high accuracy to settle the problem to the desired degree for the best galactic dynamical theory. This confirms the need to press on for ever more dependable stellar censuses, especially near the Sun, where we have some chance of achieving completeness.

Future developments

Future developments in our understanding of stellar distributions will require both an increasing number and precision of observations. However, many of the problems that remain arise from the fact that radial velocities and metal abundances are still only known for a handful of stars. Is there a gradient in metallicity across the galactic disk or outward from its center or in the direction perpendicular to the plane as some have found? How many discrete stellar populations are there, and do they represent distinct waves or bursts of past star formation or a continuity in age-related stellar properties?

For the lack of radial velocities as much as any other data, we lack space motions. Without precise quantitative measures of the equivalent widths of spectral lines and related observations, we lack metal-to-hydrogen ratios. These quantities are ineluctably closely correlated, but how? Despite improvements in theory, the solution of these problems will require programs dedicated to observations of many common nearby objects. It is possible that the search for stars with planetary systems of

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

various kinds will stimulate a new wave of observations of this type.

Bibliography

Basri G 2000 Observations of brown dwarfs *Annu. Rev. Astron. Astrophys.* **38** 485–519

Binney J and Merrifield M 1998 *Galactic Astronomy* (Princeton, NJ: Princeton University Press)

Chabrier G and Baraffe I 2000 Theory of low-mass stars and substellar objects *Annu. Rev. Astron. Astrophys.* **38** 337–77

1997 *Hipparcos Venice '97* (European Space Agency)

Kovalevsky J 1998 First results from Hipparcos *Annu. Rev. Astron. Astrophys.* **36** 99–129

Kovalevsky J 2002 *Modern Astrometry*, 2nd edn (Berlin: Springer)

Lebreton Y 2000 Stellar structure and evolution: deductions from Hipparcos *Annu. Rev. Astron. Astrophys.* **38** 35–77

Majewski S R 1993 Galactic structure surveys and the evolution of the Milky Way *Annu. Rev. Astron. Astrophys.* **31** 575–638

Morrison L V and Gilmore G F (eds) 1994 *Galactic and Solar System Optical Astrometry* (Cambridge: Cambridge University Press)

Arthur Upgren

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Stellar Dynamics

Stellar dynamics describes systems of many point mass particles whose mutual gravitational interactions determine their orbits. These particles are usually taken to represent stars in small GALAXY CLUSTERS with about 10^2–10^3 members, or in larger GLOBULAR CLUSTERS with 10^4–10^6 members or in GALACTIC NUCLEI with up to about 10^9 members or in galaxies containing as many as 10^{12} stars. Under certain conditions, stellar dynamics can also describe the motions of galaxies in clusters, and even the general clustering of galaxies throughout the universe itself. This last case is known as the cosmological many-body problem.

The essential physical feature of all these examples is that each particle (whether it represents a star or an entire galaxy) contributes importantly to the overall gravitational field. In this way, the subject differs from CELESTIAL MECHANICS where the gravitational force of a massive planet or star dominates its satellite orbits. Stellar dynamical orbits are generally much more irregular and chaotic than those of celestial mechanical systems.

Consequently, the description of stellar dynamical systems is usually concerned with the statistical properties of many orbits rather than with the detailed positions and velocities of an individual orbit. Thus it is not surprising that the KINETIC THEORY OF GASES developed by MAXWELL, Boltzmann and others in the late 19th century was adapted by astrophysicists such as Jeans to stellar dynamics in the early 20th century. Subsequent results in stellar dynamics contributed to the first analyses of kinetic plasma physics in the 1950s. Then rapid evolution of plasma theory in the second half of the 20th century, stimulated partly by prospects of controlled thermonuclear fusion, in turn contributed to stellar dynamics. This was an especially productive interdisciplinary interaction.

After describing basic physical processes such as timescales, relaxation processes, dynamical friction and damping, this article derives the virial theorem and mentions some applications, discusses distribution functions and their evolution through the collisionless Boltzmann equation and the BBGKY hierarchy, and outlines the thermodynamic descriptions of finite and infinite gravitating systems. The emphasis here is on fundamental physics rather than on detailed models.

Basic ideas

We start with simple ideas that are common to most stellar dynamical systems. For point masses to represent their components, physical collisions must be rare. In a system of objects, each with radius d, this means that the total internal volume of all the objects must be much less than the volume over which they swarm. Two spherical objects of radius d will have an effective radius $2d$ for a grazing collision whose cross section is therefore $\sigma = 4\pi d^2$. If there is a number density n of these objects and they move on random orbits, their mean free path to geometric collisions is

$$\lambda_G \approx 1/n\sigma \approx \frac{R^3}{3Nd^3}d \qquad (1)$$

where R is the radius of a spherical system containing N objects distributed approximately uniformly.

This has two easy physical interpretations. First, the average number of times an object can move through its own diameter before colliding is essentially the ratio of the cluster's volume to that occupied by all the stars. Second, the number of cluster radii the object can traverse before colliding is essentially the ratio of the projected area of the cluster to that of all the objects. In many astronomical systems, these ratios are very large. As examples, 10^5 stars in a globular cluster of 10 pc radius have $\lambda_G/R \approx 3 \times 10^{11}$ and 10^3 galaxies in a cluster of 3 Mpc radius have $\lambda_G/R \approx 30$. Therefore stellar dynamics is a good approximation over a wide range of conditions. It may, however, break down in the cores of realistic systems where only a few objects dominate at the very center and orbits are more regular.

Although geometric collisions may be infrequent, gravitational encounters are common. These occur when one object passes by another, perturbing both orbits. Naturally, in a finite system all the objects are passing by each other all the time, so this process is continuous. In a system which is already fairly stable, most perturbations are small. However, their cumulative effects over long times can be large and affect the evolution of the system significantly. To see how this works, we introduce the fundamental notion of a 'stellar dynamical relaxation time'. This is essentially the timescale for a dynamical quantity such as a particle's velocity to change by an amount approximately equal to its original value.

As an illustration of the general principle, consider two-body relaxation. Suppose, as in figure 1, that a massive star m_1, deflects a much less massive star m_2. Initially, m_2 moves with velocity v perpendicular to the distance b (the impact parameter) at which the undeflected orbit would be closest to m_1.

There is a gravitational acceleration Gm_1b^{-2} which acts for an effective time $2bv^{-1}$ and produces a component

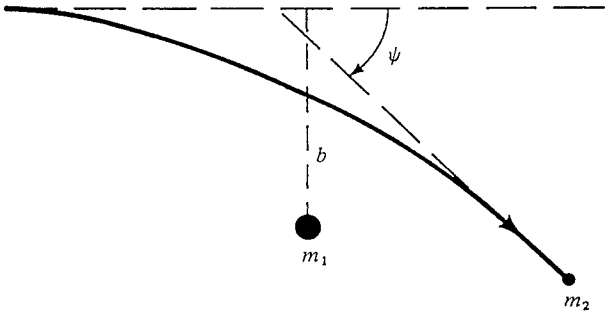


Figure 1. The deflection of a star m_2 by a more massive star m_1, schematically illustrating two-body relaxation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of velocity

$$\Delta v \approx \frac{2Gm_1}{bv} \tag{2}$$

approximately perpendicular to the initial velocity. Since $\Delta v \ll v$, the effects are linear and they give a scattering angle $\psi \approx \Delta v/v$. A more exact version follows the orbits in detail, but this is its physical essence. It shows that large individual velocity changes, $\Delta v \approx v$, and large scattering angles occur when two objects are so close that the gravitational potential energy, Gm_1m_2/b, approximately equals the kinetic energy, $m_2v^2/2$. Such close encounters are rare. Typical encounters in a spherical system of total mass M and radius R containing N objects have an impact parameter about equal to the mean separation, $b \approx RN^{-1/3}$. The average mass $m_1 \approx M/N$, and the initial velocity is given by the approximate balance of the system's total kinetic and potential energy, $v^2 \approx GM/R$. (This last relation follows from the virial theorem below.) Therefore from equation (2)

$$\frac{\Delta v}{v} \approx \psi \approx N^{-2/3} \tag{3}$$

for $N \gg 1$, and most gravitational encounters involve little energy or momentum exchange. Exceptions may occur in the centers of clusters, particularly among more massive particles, where the core is relatively isolated and the effective value of N is low.

Although most individual deflections are small, their cumulative effects are not. Over long times, uncorrelated deflections from the nearly randomly moving orbits cause each particle's velocity to randomly walk around its original value. Thus the root-mean-square velocity increases from its initial value in proportion to the square root of the number of encounters, or the square root of the time. To estimate this change, we multiply the square of each velocity change from equation (2) by the number of encounters with a given velocity and impact parameter during a unit time and then integrate over the whole range of velocities, impact parameters and times of interest. Finally setting $(\Delta v)^2 = v^2$ gives the stellar dynamical relaxation time for the cumulative small changes to modify the velocity by an amount equal to its initial value. For homogeneous systems near equilibrium with $GM \approx Rv^2$ and a Maxwellian distribution of velocities, this is

$$\tau_R \approx \frac{0.2N}{\ln(N/2)}\tau_c \tag{4}$$

where $\tau_c = R/v \approx (G\bar{\rho})^{-1/2}$ is the dynamical crossing time for a particle with average velocity and $\bar{\rho}$ is the average density.

In globular clusters, galaxies and rich clusters of galaxies $N \gtrsim 10^3$ so $\tau_R \gg \tau_c$. Cumulative small deflections do not amount to much over any one orbital period, but their secular effect can dominate after many orbits. The relaxation time τ_R will vary depending on the geometry, range of masses and velocity and density distributions within a system. For most systems, τ_R exceeds the lifetime of the system, so it cannot have been the primary process for a system's formation and currently relaxed regular roughly spherical appearance.

Systems in which $\tau_R \gg \tau_c$ are essentially 'collisionless' in the sense that their nearly smooth global gravitational force dominates the average orbits of their particles. Other systems in which strong local fluctuations in gravitational forces dominate particle orbits are 'collisional', since these fluctuations produce large-angle scatterings of the orbits. (This standard terminology does not refer to bodily collisions of objects.) Realistic astronomical systems usually combine aspects of both of these idealized categories. For example, local gas clouds, star clumps or dark matter inhomogeneities may scatter stellar orbits from their average smooth paths in GALAXIES.

Even for maintaining a relaxed system near a state of equilibrium, gravitational encounters cannot be the whole story. If they were, the constant average increase in the root-mean-square velocities of all the particles would eventually cause them all to escape, defying the conservation of energy. A second, balancing, process must operate. This is dynamical friction.

Imagine yourself on a particle moving faster than the average speed of the particles around it. In a stationary reference frame attached to your particle, the other particles will appear to stream by. More particles will stream by in the direction you are moving toward than in other directions. The orbits of these excess particles will be deflected slightly by the gravity of your particle. This deflection causes a slight convergence of the orbits behind your particle, opposite the direction of its motion. On average, therefore, there is a statistical excess of particles behind yours. Their excess gravity decelerates the motion of your particle, and tends to reduce it to the average velocity. Conversely, if the speed of your particle is slower than average, there will be a statistical excess of particles passing it in your direction of motion. Their orbits will converge slightly in front of the motion of your particle, and their excess gravitational force will speed up your particle. Thus dynamical friction is a great leveller of velocities. The balance between dynamical friction and two-body acceleration keeps the system close to equilibrium. Therefore τ_R must also be the approximate timescale for relaxation by dynamical friction. Mathematically, a second-order non-linear partial differential–integral equation, known as the Fokker–Planck equation, describes this balance, and it is usually solved numerically.

Dynamical friction also plays another important role in stellar dynamics. If a massive object M (such as a high-mass star in a galaxy, or a large black hole in a galactic nucleus or even a small galaxy in the halo of another larger galaxy) moves rapidly through a field of much less massive objects, m, dynamical friction will cause it to lose kinetic energy, fall deeper into the surrounding gravitational well, speed up while falling, lose more energy and consequently spiral into the center of the larger

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

system. This helps promote the merging of galaxies, for example (see GALAXIES: INTERACTIONS AND MERGERS). The relaxation time τ_R of equation (4) is decreased significantly by a factor of order m/M. This is because the large mass difference induces *gr*avitational *ex*citati*ons*, known as grexons (also from the Latin 'grex' meaning 'herd' or 'flock'): a collective gravitational excitation in the wake of a massive particle moving through a system of less massive particles. It generally takes the form of a region of statistical overdensity among the less massive particles passing through the wake. These gravitating collective modes of many small masses interact with each other, enhancing their lifetimes and producing a larger effective statistical condensation behind M through which particles of mass m move but linger longer than they would for $M \approx m$. These grexons arise when $M \gg m$ and amplify the gravitational drag.

Other forms of collective excitations can also amplify relaxation in systems which are initially far from equilibrium. This collective relaxation—of which 'violent relaxation' is an extreme case—is much faster than τ_R and closer to the crossing timescale τ_c. Let us start by thinking about a relatively simple case. Suppose one star of mass m initially has zero velocity at the edge of an already relaxed cluster of stars with total mass M and radius R. Details of its orbit would depend on the density distribution of the cluster and the building up of small perturbations, but the star would roughly follow the equation of motion $m\ddot{r} \approx GmMR^{-2}$. Since $\ddot{r} \approx R/t^2$, this would give a characteristic infall timescale $\sim(G\bar{\rho})^{-1/2}$, which is just the crossing time τ_c. We may think of this as the star being scattered by the average force of the entire cluster, i.e. by the 'mean field'. As the star falls inward, the small discrete perturbations of nearby stars will also scatter it, and this 'fluctuating field' will impart some net angular momentum to the star so that it is unlikely to reach the exact center of the cluster. Instead, it will join the other stars in their complicated orbits. The net relaxed orbit of the star comes from the joint effects of the smooth global mean field and the two-body relaxation of the local fluctuating field.

Now let us make the process a little more exciting. Suppose the cluster is not itself relaxed. Then it will not be quasi-static. Parts of it will be falling together, interpenetrating, separating, streaming in different directions and generally flopping around. These represent larger-scale fluctuations—collective modes—which will scatter the incoming star in addition to its scattering by the global mean field and the shorter-scale fluctuations of nearby individual stars. Moreover, the large-scale fluctuations will also be scattering and interacting with each other. With high-amplitude large-mass collective modes fluctuating over a wide range of lengthscales, and on corresponding timescales shorter than τ_c for the entire cluster, individual orbits will relax on timescales closer to τ_c than τ_R. This is collective relaxation, and it may apply to the formation of clusters of stars, galaxies and clusters of galaxies, particularly if they build up by the merging of smaller systems.

Violent relaxation occurs in the extreme limit when collective modes are so chaotic on all scales that each of them lasts for only a short time but is quickly replaced by another such mode elsewhere in the system. None of these modes correlates with one another. Each star moves in a mean field which changes quickly in time as well as in space. Consequently, the energy of each individual star along its orbit is not conserved; only the energy of the entire system remains constant. If this process could continue indefinitely, the velocity distribution of the stars would generally become similar to the Maxwell–Boltzmann distribution for a perfect gas after a timescale τ_c.

However, it does not. On this same gravitational free-fall timescale, τ_c, damping mechanisms destroy the ideal conditions for violent relaxation. In stellar dynamics, as in other systems, damping mechanisms reduce departures from an equilibrium state. Most damping is characterized by dynamical dissipation—the roughly random transfer of relatively ordered (low-entropy) energy into relatively disordered (high-entropy) energy. The transfer may occur among 'particles', as when a stellar cluster or a forming cluster of galaxies ejects a high-energy member so that the remaining cluster becomes more tightly bound. Another example of particle dissipation is the phase mixing of orbits. Generally this involves interactions of stars at slightly different phases of the same, otherwise unperturbed, orbit. As a simple illustration, suppose an isolated spherical cluster of stars, initially completely at rest, starts falling together. The stars will not simply all plunge together into the center. Instead, stars at different distances from the center, i.e. at different phases of their radial orbits, will perturb each other. These perturbations can be large among stars with low relative velocities, and can occur on the dynamical crossing timescale τ_c. Most stars will acquire some angular momentum from these perturbations, although the cluster's net angular momentum remains zero. This further mixes the phase of the orbits, dissipating some of the initially highly ordered radial infall velocities into more random transverse velocities. Eventually, a component of random kinetic energy builds up as a form of heat whose effective pressure resists further collapse of the cluster on the free-fall timescale τ_c. Similar phase mixing of orbits, combined with heat generated by collective relaxation itself, can occur to damp more general collective motions such as those of violent relaxation. More gentle phase mixing can also occur in relaxed galaxies over longer timescales to mix up streaming motions of stars.

Dissipation and damping may also result when particles and waves interact. The waves are moving periodic density perturbations. The best-known case was first found for plasmas by Landau and then applied to stellar dynamics. It occurs even in idealized collisionless systems where particles do not scatter one another significantly. The physical reason for collisionless damping arises from the detailed interaction of a wave with the orbits of background stars which are not part of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the wave. Thus this process would not show up in a pure continuum approximation. What happens is that some of the background stars will move a bit faster than the wave, others a bit slower, even though their average background number density is constant. The wave exerts a force on these stars and thus exchanges energy with them. On average, the wave gains energy from the fast-moving stars, which therefore amplify the wave, and loses energy to the slow-moving stars, which damp the wave. The net result depends on whether there are more fast or slow stars with velocities near the phase velocity $v = \omega/k$ of the wave. Moreover, the strength of the growth or damping depends on the total net number of fast or slow stars involved. A detailed analysis of this result is complex but, for a roughly Maxwellian velocity distribution of background stars, damping is greatest for waves whose phase velocity ω/k is near the velocity dispersion $\langle v\rangle$ of the background stars. These waves decay on a timescale $\lambda/\langle v\rangle$ during which the stars are able to move through the wave.

Phase mixing, Landau damping and other processes such as trapping by clusters, tidal disruption and small-angle scattering all combine with violent relaxation into a form of collective relaxation which randomizes velocities on a timescale $\sim(G\bar{\rho})^{-1/2}$, provided that the system starts out very far from its eventual quasi-equilibrium state. What is left is a system in quasi-equilibrium with a roughly Maxwellian velocity distribution. No direct collisions are responsible for this end result—unlike the case of a perfect gas—only the non-linear encounters among particles and collective modes.

The virial theorem

This is perhaps the most astronomically important dynamical property of a quasi-equilibrium system. Therefore we describe the virial theorem in some detail. With it, we can estimate the system's total mass from observations of the positions and velocities of its particles. The virial theorem is essentially just a position moment of the self-consistent gravitational equations of motion within the system. To illustrate it in the simplest case, consider a satellite of mass $m^{(\alpha)}$ in a circular orbit around a much more massive object of mass $m^{(\beta)}$. The balance between centrifugal and gravitational forces gives $m^{(\alpha)}v^2/r = Gm^{(\alpha)}m^{(\beta)}/r^2$. Multiplying through by r gives $2K + W = 0$ where $K = m^{(\alpha)}v^2/2$ is the kinetic energy and $W = Gm^{(\alpha)}m^{(\beta)}/r^2$ is the gravitational potential energy. In this circular orbit the instantaneous value and the time average of the energy are identical. For more complicated configurations, we might suspect that this result will still hold for the more general time averages of K and W. It does.

The most general Newtonian equations of motion for a system of gravitating particles are

$$\frac{\mathrm{d}}{\mathrm{d}t}(m^{(\alpha)}v_i^{(\alpha)}) = F_i^{(\alpha)} = m^{(\alpha)}\frac{\partial\varphi}{\partial x_i^{(\alpha)}} = -Gm^{(\alpha)}\sum_{\beta\neq\alpha} m^{(\beta)}\frac{x_i^{(\alpha)} - x_i^{(\beta)}}{|\boldsymbol{x}^{(\alpha)} - \boldsymbol{x}^{(\beta)}|^3} \qquad (5)$$

for the ith components of position and velocity, where the gravitational potential

$$\varphi = G\sum_{\beta\neq\alpha}\frac{m^{(\beta)}}{|\boldsymbol{x}^{(\alpha)} - \boldsymbol{x}^{(\beta)}|} \qquad (6)$$

results from all the other particles. This derivation is general enough to let the particle masses vary with time, by mass loss or accretion, isotropically as we will assume, or even anisotropically with a net 'rocket effect'.

Now multiply equation (5) by $x_j^{(\alpha)}$ and sum over α to take its lowest-order position moment. Since α and β are just dummy indices, we could just as validly multiply through by $x_j^{(\beta)}$. Equivalently, and more elegantly, we could represent the right-hand side as half the sum of both these multiplications, i.e. as

$$W_{ij} = -\frac{G}{2}\sum_\alpha\sum_{\beta\neq\alpha} m^{(\alpha)}m^{(\beta)}\frac{(x_i^{(\alpha)} - x_i^{(\beta)})(x_j^{(\alpha)} - x_j^{(\beta)})}{|\boldsymbol{x}^{(\alpha)} - \boldsymbol{x}^{(\beta)}|^3} \qquad (7)$$

which is symmetric to an interchange of the α and β particles. This is known as the potential energy tensor, W_{ij}.

Rewriting the left-hand side of equation (5) multiplied by $x_j^{(\alpha)}$ gives

$$\begin{aligned}\sum_\alpha x_j^{(\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}(m^{(\alpha)}v_i^{(\alpha)}) &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha m^{(\alpha)}x_j^{(\alpha)}v_i^{(\alpha)} - \sum_\alpha m^{(\alpha)}v_i^{(\alpha)}v_j^{(\alpha)}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha\frac{1}{2}m^{(\alpha)}[(x_j^{(\alpha)}v_i^{(\alpha)} + x_i^{(\alpha)}v_j^{(\alpha)})\\ &\quad + (x_j^{(\alpha)}v_i^{(\alpha)} - x_i^{(\alpha)}v_j^{(\alpha)})] - \sum_\alpha m^{(\alpha)}v_i^{(\alpha)}v_j^{(\alpha)}. \qquad (8)\end{aligned}$$

The terms in parentheses separated by a minus sign are antisymmetric, and since all the other contributions to the moment of equation (5) are symmetric, these antisymmetric terms must be zero. This proves that angular momentum is conserved, since these antisymmetric terms are just the total angular momentum of the isolated system. The symmetric terms in equation (8) may be written as

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha m^{(\alpha)}(x_j^{(\alpha)}v_i^{(\alpha)} + x_i^{(\alpha)}v_j^{(\alpha)}) - \sum_\alpha m^{(\alpha)}v_i^{(\alpha)}v_j^{(\alpha)}\\ &\quad = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sum_\alpha m^{(\alpha)}x_i^{(\alpha)}x_j^{(\alpha)}\\ &\quad\quad - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha \dot{m}^{(\alpha)}x_i^{(\alpha)}x_j^{(\alpha)} - \sum_\alpha m^{(\alpha)}v_i^{(\alpha)}v_j^{(\alpha)}. \qquad (9)\end{aligned}$$

Now each of the three summations over α on the right-hand side of equation (9) has a physical interpretation. The first summation is the usual inertia tensor I_{ij}, the second we shall call the mass variation tensor J_{ij} and the third is twice the kinetic energy tensor T_{ij} which incorporates all the motions of all the particles. Combining equations (7) and (9) for the moment of equation (5) then gives the tensor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

virial theorem for a system of gravitationally interacting particles:

$$\frac{1}{2}\frac{\mathrm{d}^2 I_{ij}}{\mathrm{d}t^2} - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}J_{ij} = 2T_{ij} + W_{ij}. \tag{10}$$

This is an exact result, without any approximations so far. It provides an overall constraint on the system's evolution. Further constraints follow from taking higher-order moments of the equations of motion. Multiplying equation (5) by $x^m v^n$ and summing gives the $(m + n)$th combined spatial and velocity moment. This forms an infinite sequence of virial moments. The more moments are used, the tighter are the constraints on the system's evolution. All the moments taken together usually become equivalent to a complete solution of the original equations of motion (5).

In practice, most astronomical calculations are content with a simplified version of the lowest-order virial equation (5). If the mass loss rate $\dot{m}^{(\alpha)} = 0$, then $J_{ij} = 0$ and the form reduces to that quoted in many texts. For the special rate of mass loss proportional to the mass, $\dot{m}^{(\alpha)} = f(t)m^{(\alpha)}(t)$, the mass loss tensor is proportional to the moment of inertia tensor. A great simplification follows by taking the time average of the virial theorem. For example, the time average of the first term gives

$$\left\langle \frac{\mathrm{d}^2 I}{\mathrm{d}t^2} \right\rangle = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau \frac{\mathrm{d}\dot{I}}{\mathrm{d}t}\,\mathrm{d}t = \lim_{\tau\to\infty}\frac{1}{\tau}[\dot{I}(\tau) - \dot{I}(0)]. \tag{11}$$

This time average can be zero either if the system is localized in position and velocity space so that $\dot{I}(\tau)$ has an upper bound for all τ or if the orbits are periodic so that $\dot{I}(\tau) = \dot{I}(0)$. Similarly the time average of J_{ij} can be zero if $\dot{m}(t)$ does not increase as fast as t for $t \to \infty$. The terms on the right-hand side of equation (10) remain non-zero over any time average, so

$$2\langle T_{ij}\rangle + \langle W_{ij}\rangle = 0. \tag{12}$$

The usual (contracted) form of the virial theorem is obtained by setting $i = j$ and summing over $i = 1, 2, 3$ to give

$$2\langle T\rangle + \langle W\rangle = 0 \tag{13}$$

where T and W are the entire kinetic and potential energies of the system.

The virial theorem describes reasonably bound, quasi-stable clusters of stars and galaxies. Its usual application is to estimate the mass of these clusters by writing equation (13) in the approximate form

$$V^2 = \gamma\frac{GM}{R}. \tag{14}$$

This is similar to the form for a circular satellite orbit discussed earlier, but now V^2 is the velocity dispersion of objects in the cluster, R is the cluster's radius and M is its total mass. The constant γ is usually of order unity and depends on the precise operational definitions of V, R and M, especially if the particle masses are not identical. Other factors which influence γ when trying to deduce M are projection effects in transverse positions, loss of one position and two velocity components, departures of time averages from the instantaneous snapshots of observations, questions of cluster membership and selection of a limited number of particles for observation. Each of these effects depends on the particular circumstances of an individual cluster. Their combination introduces considerable uncertainty into γ, but this can often be estimated with computer simulations.

Virial estimates of the masses of CLUSTERS OF GALAXIES have been made since the 1930s. A succession of more accurate results with more complete samples and better simulations, which continues to the present, revealed clear discrepancies between the virial mass of many clusters and their mass estimated from the luminosities of their individual galaxies. Unless cluster formation proceeded by a radically different route from the condensation processes discussed currently, this discrepancy suggests that the amount of 'DARK MATTER' in clusters is about five times the amount of luminous matter in galaxies. This gives roughly one-fourth of the total mass needed for a closed Einstein–Friedmann universe. The form of this dark matter and its total amount are two of the main problems of modern astronomy.

Distribution functions

Rather than dealing with the positions and motions of each identifiable particle individually, stellar dynamics often finds a less detailed description based on distribution functions to be more useful and solvable. These distributions represent the probability that any arbitrary particle in the system has a particular property of interest. Usually stellar dynamics is concerned with the single-particle distribution $f(\boldsymbol{r}, \boldsymbol{v})$ for the number of particles having position $\boldsymbol{r}$ and velocity $\boldsymbol{v}$. Integrated over all velocities, this gives the density as a function of position. Integrated over all positions, this gives the velocity distribution. Integrated over both position and velocity, it gives the total number of particles in the system. When $f(\boldsymbol{r}, \boldsymbol{v})$ is normalized to this total number, it becomes a probability distribution. Then $f(\boldsymbol{r}, \boldsymbol{v})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{v}$ is the probability that a particle is in the differential volume $\mathrm{d}\boldsymbol{r}$ around $\boldsymbol{r}$ and has a velocity in the range $\mathrm{d}\boldsymbol{v}$ around $\boldsymbol{v}$.

Other types of distribution functions are also useful. For example, the two- particle distribution $f(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{v}_1, \boldsymbol{v}_2)$ describes the probability that of any two particles one is in the volume $\mathrm{d}\boldsymbol{r}_1$ with velocity in the range $\mathrm{d}\boldsymbol{v}_1$ and the other is simultaneously in $\mathrm{d}\boldsymbol{r}_2$ with velocity in $\mathrm{d}\boldsymbol{v}_2$. This two-particle distribution may be written as the product of two single-particle distributions plus terms representing any correlations that may be present between particles in the two volumes or the two velocity ranges. Similarly, a hierarchy of three-, four-, etc particle distributions and correlations may be built up. Eventually it will reach the N-particle distribution which gives a complete description of the positions and velocities of all N particles in the system. This would then be equivalent to specifying the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

position and velocity of every particle—the most detailed description possible. A related form of distribution function, $f(N)$, concentrates on cells rather than particles and specifies the probability that a cell of given size and shape contains N particles.

All these distribution functions may generally have anisotropic dependences on position and velocity and will evolve with time unless they are in a stationary equilibrium state. Anisotropic distribution functions and their properties have been widely explored in recent years. Their complexity provides useful descriptions of stellar motions in both spiral and elliptical galaxies.

The simplest evolution for the one-particle distribution function $f(\boldsymbol{r}, \boldsymbol{v}, t)$ occurs in collisionless systems and follows the collisionless Boltzmann equation. Here collisionless means that particle orbits proceed smoothly under the influence of just the mean field. There are no scatterings by other particles or by local fluctuations. Under these conditions $f(\boldsymbol{r}, \boldsymbol{v}, t)$ satisfies an equation of continuity. This is equivalent to its total time derivative, following the motion through a 6-dimensional position–velocity phase space, being zero:

$$\frac{\mathrm{D}f}{\mathrm{D}t} = \frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla} f + \dot{\boldsymbol{v}} \cdot \boldsymbol{\nabla}_v f = 0. \qquad (15)$$

Equation (15) is the collisionless Boltzmann equation—one of the most useful descriptions of stellar dynamics. Notice that the gradients in position and in velocity space are treated equivalently. Although, written this way, it looks like a fairly simple linear partial differential equation for $f(\boldsymbol{r}, \boldsymbol{v}, t)$, it is really a complicated non-linear differentio-integral equation. This is because the gravitation force, proportional to the acceleration $\dot{\boldsymbol{v}} = \boldsymbol{\nabla}\phi$, is a function of the density which is itself an integral of $f(\boldsymbol{r}, \boldsymbol{v}, t)$ over velocity space. This relation is obtained from Poisson's equation, $\nabla^2\phi(\boldsymbol{r}, t) = -4\pi G\rho(\boldsymbol{r}, t)$, and together with the relevant initial and boundary conditions leads to self-consistent solutions.

Very few solutions are known in closed analytic form; most are results of numerical integrations or perturbation theory. The best-known solution occurs for an idealized homogeneous, isotropic, uncorrelated equilibrium distribution. It is the Maxwell–Boltzmann distribution with a Gaussian distribution of velocities. Solutions of equation (15) also provide a zero-order approximation for systems of particles such as globular clusters and rich clusters of galaxies. Their corresponding spatial density is spherically symmetric, decreasing with radius. Such isothermal spheres have to be truncated to describe clusters of finite mass. Truncation by tidal cutoffs, evaporation, inflow, etc for realistic systems will modify their distribution functions.

Spatial and velocity moments of the collisionless Boltzmann equation, similar to those leading to the virial theorem and its generalizations described earlier, yield the Jeans equations of 'stellar hydrodynamics'. These are analogous to the usual fluid equations of hydrodynamics. However, because short-range atomic interactions dominate fluids, it is a much better approximation to truncate the moment equations at low order for fluids than it is in the stellar dynamical case. Moreover, the general stellar hydrodynamical equations are anisotropic in their spatial and velocity coordinates.

Generalizations of the collisionless Boltzmann equation have been developed to incorporate local fluctuations, such as those described earlier, which lead to dynamical relaxation, evaporation and other instabilities. The most useful of these is the Fokker–Planck description which represents the total gravitational field as the sum of two parts: the smooth average long-range field and the local fluctuating field due to near-neighbor particles. These fluctuations cause the particles to diffuse in velocity space as well as in configuration space. They also incorporate the dynamical friction effects, mentioned earlier, which prevent excessive velocities from being reached. The Fokker–Planck equation provides a more accurate account of processes such as evaporation, core collapse and oscillations which can occur in globular clusters.

The most general and rigorous description of stellar dynamics, and therefore the least solvable, is known as the BBGKY hierarchy (after the initials of its early developers in other subjects, Born, Bogoliubov, Green, Kirkwood and Yvon). It starts from Liouville's equation which is an equation of continuity similar to equation (15) but in a $6N$-dimensional (Gibbs) phase space rather than in the 6-dimensional (Boltzmann) phase space. Here N is the number of physical particles in the system and each point in the Gibbs phase space of $3N$ position plus $3N$ velocity dimensions represents the state of the entire system of $6N$ particles at a given time. As the system evolves, its representative point moves continuously and smoothly through this phase space because there are no external forces outside the system to perturb it. Thus the Liouville continuity equation is exact. It is a first-order linear differential equation. The problem is that it has $6N$ variables, plus time. In fact, it is just a condensed way of representing the orbits of all the particles in the system.

To make this description useful, it is necessary to successively integrate out all but $6(N-1), 6(N-2), \ldots, 6$ of the $6N$ variables in the full distribution function. This leaves a set of N coupled non-linear integro-differential equations for all the N, $N-1$, $N-2, \ldots$, one-particle distribution functions—the BBGKY hierarchy. The entire set is equivalent to the equations of motion for all N particles. Truncating the lowest-order equation for $f(\boldsymbol{r}, \boldsymbol{v}, t)$ by neglecting its coupling to $f(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{v}_1, \boldsymbol{v}_2, t)$ is equivalent to the collisionless Boltzmann equation. Retaining this coupling, but neglecting any higher-order coupling, gives essentially the Fokker–Planck equation. This formalism has been much studied and gives great insight into the physical nature and relations among different stellar dynamical descriptions. It has found important applications in understanding the linear growth of two- and three-particle correlation functions in galaxy clustering, as well as for plasma physics and the theory of imperfect fluids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thermodynamic descriptions

Particle and distribution function descriptions are mathematically complex because they contain much detailed microscopic information about a stellar dynamical system. Thermodynamic descriptions are mathematically much simpler because they deal mainly with macroscopic information which averages over the microscopic detail. This relative simplicity has provided considerable physical insight into the behavior of stellar dynamical systems.

Since the macroscopic properties generally change with time (e.g. even in virialized clusters, particles slowly evaporate), stellar dynamical systems cannot be in equilibrium. Therefore it might be thought that thermodynamics, which is primarily an equilibrium theory, could not apply. Often, however, the timescales for departures from equilibrium are very long, and singular states are essentially unattainable over shorter times. As a result, thermodynamics provides a good approximation for the periods of interest. The system may be in quasi-equilibrium, during which it evolves through a sequence of equilibrium states. In simple systems, each state can be described well in terms of average macroscopic thermodynamic variables such as temperature, pressure, chemical potential, internal energy, volume, total number of particles and entropy, but these quantities change on a timescale which is slow compared with the timescale for a local microscopic configuration of particle positions and velocities to change.

Symmetry properties of different physical systems strongly influence their thermodynamic descriptions. For example, infinite statistically homogeneous systems—which may represent galaxy clustering—have rotational and translational symmetry everywhere. However, finite spherical systems have rotational symmetry only at their center and translational symmetry nowhere. Consequently these infinite systems are described by a grand canonical ensemble in which energy and particles can move across boundaries. Finite, isolated clusters, on the other hand, are described either by a microcanonical ensemble with no transport across boundaries or by a canonical ensemble with only energy transport. (An ensemble is a collection of systems with the same average macroscopic properties but different microscopic configurations consistent with the macroscopic averages.) This leads to differences in their distribution functions and fluctuation spectra.

Thermodynamic behavior in systems dominated by gravity is often 'counterintuitive', although really just in the sense that they behave differently from more familiar systems of particles without long-range, unshielded, attractive forces. For example, if one removes energy from a self-gravitating cluster of stars, it becomes hotter. Adding energy makes it cooler. Thus it has a negative specific heat. In fact, this follows simply from the virial theorem for a bound system. Slowly adding energy makes the system's gravitational well less negative. To maintain their changing quasi- equilibrium state, the particle velocities must decrease on average, and the system grows cooler. It is essentially the same effect as adding energy to a satellite orbit around a massive body, and so it should be, since the virial theorem applies to both cases.

Applied to finite, isolated spherical clusters of stars, gravitational thermodynamics has been especially useful in elucidating their global instabilities. The isothermal sphere, where all particles have the same mass and temperature, provides a relatively simple illustration. Its density is obtained by solving the collisionless Boltzmann equation and Poisson's equation for these conditions, or, more easily but less accurately, by solving the gravitational hydrostatic equations with an isothermal gas equation of state. It has a central core, and density $\rho \propto r^{-2}$ in the outer parts, and is usually truncated to keep its mass finite. (Other stellar clusters with more general equations of state similar to polytropic stars can also be manufactured to provide simple models.)

The stability of such an isothermal sphere depends on the relation between its total energy and its total entropy. Systems of the same total energy may have different entropies depending on their size and internal distribution. If the entropy has a local maximum for a given energy, then that configuration is stable to small fluctuations (but possibly metastable to large changes). If the entropy does not have a local maximum, then the system of a given energy can redistribute itself internally to increase its entropy, and the situation is unstable. Such analysis, originally done by Antonov, for an isothermal sphere confined to a spherical box shows that, if the ratio of the central to the boundary density exceeds 709, the system becomes unstable. It tends to evolve away from the isothermal sphere density distribution into a denser central core surrounded by a much less dense halo. Similar core–halo evolution is also found in numerical simulations. Its underlying dynamical mechanism is the slow evaporation of stars which have accumulated just enough energy to escape by their interactions with the fluctuating gravitational field of neighboring stars. Since the total gravitational energy of the isolated sysem is conserved, its core becomes denser with its energy more negative in order to compensate the positive total energy of the escaping stars.

An infinite statistically homogeneous system provides a contrasting application of gravitational thermodynamics. The cosmological many-body problem is an example. In its simplest case, we start with a uniform random (Poisson) distribution of identical point masses throughout the universe and ask how this distribution changes as the universe expands. In the usual Einstein–Friedmann cosmological models, the cosmological expansion exactly compensates the smooth long-range component of the gravitational field. This leaves only the relatively local fluctuations caused by the discreteness of the particle gravitational fields and by any clustering. These fluctuations can be described by an equation of state which incorporates the gravitational interaction into thermodynamic quantities such as the internal energy and pressure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Because the universe generally expands more slowly than the timescale for particle configurations within local fluctuations to change, it undergoes a quasi-equilibrium evolution. At any time an equilibrium thermodynamic state provides a good description of the clustering. Its thermodynamic quantities change slowly as the universe expands adiabatically.

Infinite statistically homogeneous systems have a uniform density when averaged over sufficiently large scales or over an ensemble of smaller scales. These infinite systems are characterized by the fluctuations over various scales of macroscopic quantities around their averages. The fluctuations most closely related to observations are the particle distribution functions $f(N, V)$ which give the probability for finding N particles (e.g. galaxies) in a randomly placed volume of size V. Applying thermodynamic fluctuation theory to the cosmological many-body equation of state gives a relatively simple formula for this distribution:

$$f(N, V) = \frac{\bar{N}(1-b)}{N!}[\bar{N}(1-b) + Nb]^{N-1}\, \mathrm{e}^{-\bar{N}(1-b)-Nb}. \quad (16)$$

Here $\bar{N} = \bar{n}V$ is the average number in the volume V for the average number density $\bar{n}$. The quantity $b = -W/2K$ is the ratio of the gravitational correlation energy to twice the kinetic energy of random motions, averaged over all volumes of size V having a particular shape. The correlation energy, W, is the integral of the r^{-1} interparticle gravitational potential multiplied by the two-particle correlation function $\xi(r)$ over the volume. $\xi(r)$ is the average excess over the random Poisson probability for finding a particle in a differential volume element at a distance r from another particle.

In a completely uncorrelated initial Poisson distribution, $\xi(r) = 0$, $W = 0$ and consequently $b = 0$. In this limit, the distribution function of equation (16) does indeed reduce to the standard Poisson form. As the system evolves, regions where near-neighbor points happen to be closer than average cluster as a result of their enhanced gravity. These clusters subsequently cluster themselves and a hierarchy of non-linear clustering, with a wide range of amplitudes and scales, represented by the increasing value of b, gradually builds up. In the Einstein–Friedmann universe, this non-linear evolution of b can be calculated analytically and it asymptotically approaches unity as the universe expands. This asymptotic limit represents bound clustering on all scales, and is strictly reached only for $\Omega_0 = 1$.

A velocity distribution function can also be derived from equation (16). Compared with the Maxwell–Boltzmann distribution for finite isothermal spheres, the velocity distribution function for the cosmological case is much broader. This is because it includes all levels of clustering, from isolated field galaxies to the richest bound clusters. Numerical computer simulations of the cosmological many-body system verify equation (16) as well as its associated velocity distribution function. Observations of galaxies on the sky and in three dimensions, using COUNTS IN CELLS as well as VOID and near-neighbor statistics, are also in very good agreement with equation (16). The currently observed value of b is about 0.75.

The BBGKY kinetic hierarchy, mentioned earlier, has also been used to examine hierarchial galaxy clustering. It is partially solvable for the two- and three-particle correlation functions in the linear regime. However, it generally contains less usable observational information than the distribution functions. There are also many more complicated models of galaxy clustering, involving hot and cold dark matter, various forms of initial density and velocity perturbations, biases between the galaxies and dark matter, etc, but their detailed applicability to our universe remains unclear.

Many applications of these main physical descriptions of stellar dynamics—particle orbits, kinetic theory, distribution functions and thermodynamics—have developed over the last century. They have helped provide an understanding of the formation, relaxation and dynamical evolution of star clusters within galaxies. They help explain the streaming motions and arms of spiral galaxies, as well as the triaxial shapes of elliptical galaxies. Combined with the possible existence and effects of massive black holes in galactic nuclei, stellar dynamics helps describe the evolution of the nuclei and the feeding of their black holes. On a larger scale, stellar dynamics gives a background for understanding the formation, relaxation and evolution of galaxy clusters, incorporating effects of their internal dark matter. Even more generally, the principles of stellar dynamics play an important role in accounting for the structure of matter over the largest scales in the universe.

Although stellar dynamics is an old and well-established subject, it is often rejuvenated by new insights and new applications. Some interesting areas for future further developments include the detailed links between dynamics and thermodynamics, implications for galaxy clustering, galaxy and cluster merging, distributions around black holes in galactic nuclei, counterstreaming in galaxies and the development of stellar motions in large star-forming regions.

Bibliography

Binney J J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)

Chandrasekhar S 1960 *Principles of Stellar Dynamics* (New York: Dover)

Ogorodnikov K F 1965 *Dynamics of Stellar Systems* (New York: Pergamon)

Saslaw W C 1985 *Gravitational Physics of Stellar and Galactic Systems* (Cambridge: Cambridge University Press)

Saslaw W C 2000 *The Distribution of the Galaxies: Gravitational Clustering in Cosmology* (Cambridge: Cambridge University Press).

William C Saslaw

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Evolution

Stars evolve because they lose energy through radiation from their hot surfaces. As a result, there is a general tendency for the star to contract under gravity. This is partly stopped by the release of energy from nuclear reactions. However, once the nuclear fuel is exhausted, gravity gets the upper hand and the star ends its life as a compact object: a white dwarf, a neutron star or a black hole, the last representing the ultimate victory of gravity.

In the nuclear reactions lighter elements are fused into heavier; it is generally believed that all elements but the very lightest have been formed in this way. Because of mass loss the elements so formed can later be incorporated in new generations of stars and their planetary systems. Thus we are all made of star-dust.

An overview of stellar evolution

The energy radiated from the star must be supplied from changes in the stellar interior. A star essentially has three sources of energy: the release of gravitational energy through contraction which makes the star more tightly bound; cooling through emission of the internal thermal energy; the generation of energy through nuclear fusion whereby the ligher nuclei fuse into heavier, more tightly bound nuclei. During most of the life of a star, the nuclear processes dominate the energy production; this requires that the temperature is sufficiently high that the nuclei can overcome their mutual Coulomb repulsion. Fusion leads to the build-up of successively heavier elements in the core of the star (see NUCLEAR REACTION RATES). These changes in the composition of the deep interior cause dramatic changes in the star's structure and surface properties such as luminosity and surface temperature, reflected in the observable characteristics of the star.

The general principles of stellar evolution may be described on this background. Stars form from condensations in interstellar clouds. During the initial contraction the energy output comes from the gravitational energy released as a result of the contraction. At the same time, this leads to heating of the stellar interior. When the temperature becomes sufficiently high, nuclear reactions set in. Generally speaking, the first reactions are those between nuclei of the smallest charges, and hence with the lowest Coulomb barrier. Thus the initial reaction is the fusion of hydrogen into helium. Once the nuclear energy generation balances the energy loss from the stellar surface, the contraction essentially stops. Evolution then takes place on a nuclear time scale, as a result of the gradual change in the structure of the star as hydrogen is converted into helium. In the solar case, this phase of central hydrogen 'burning' lasts around 10 Gyr.

When hydrogen has been consumed at the center, the star is left with a core consisting of helium and the original heavier elements; outside this core the temperature is typically sufficiently high for hydrogen fusion to continue, in what is called a *shell source*. This adds helium to the core, causing it to contract; the resulting release of gravitational energy partly leads to heating of the core, until the temperature becomes sufficiently high that the next nuclear reactions, the fusion of three helium nuclei into a carbon nucleus, can take place. At the same time the star may still have a hydrogen-fusion shell source. This sets the general pattern of the nuclear evolution of the star: fusion reactions in the core, alternating with core contraction and heating until the next stage of nuclear reactions can take place, and with possible shell sources outside the core with fusion of lighter elements.

The evolution of stars is conveniently illustrated in *the theoretical Hertzsprung–Russell diagram* (see also HERTZSPRUNG–RUSSELL DIAGRAM), where the stellar luminosity is plotted against surface temperature (more precisely the *effective temperature*; see below) as shown in figure 1. Following observational convention, the temperature increases towards the left. The ZAMS marks the stars that have just started core hydrogen burning. Also marked is the termination of the main sequence where core hydrogen burning has finished. The evolutionary sequences shown end on *the red-giant branch* where the stellar radii grow dramatically; later evolutionary stages will be illustrated below.

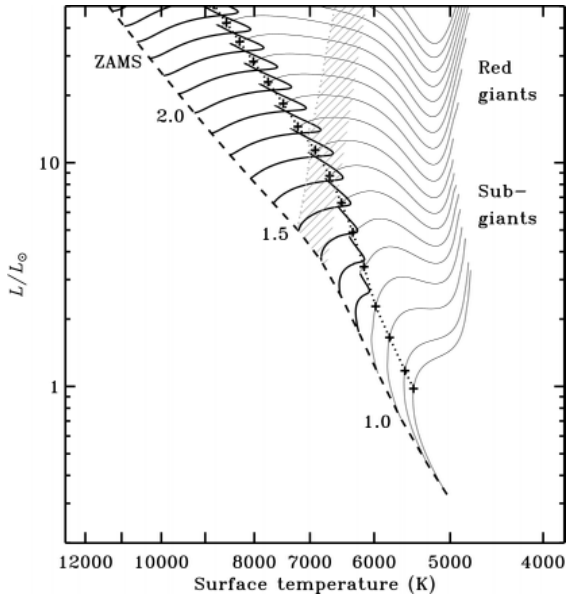


Figure 1. Evolutionary tracks of stars in terms of surface temperature and total luminosity ($L_\odot$ is the solar luminosity) are shown by the solid curves (a few are labeled by the stellar mass, in units of the mass $M_\odot$ of the Sun). The *zero-age main sequence* (ZAMS; dashed) marks stars just starting fusion of hydrogen into helium, while the dotted curve shows the end of core hydrogen burning. Bolder curves show parts of the tracks where the star has a convective core. Stars with surface temperatures below 7000 K, indicated by the hatching, have convective envelopes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

An important consequence of stellar evolution is the fusion of lighter elements into heavier, starting from the lightest, hydrogen. These processes take place in the deep interiors of the star; however, during late stages of evolution the products of the reactions may be spread through mixing by convective motions throughout the star, hence becoming visible in its observable surface composition. Also, mass loss from the star, either gradual or explosive, may distribute the elements generated in the stellar interior in the interstellar matter where it can be incorporated into newly formed stars and planets. Indeed, there is strong evidence that essentially all elements heavier than helium have been formed in this manner, from the mixture of hydrogen and helium resulting from the big bang (see COSMOLOGY: STANDARD MODEL and NUCLEOSYNTHESIS). Thus to understand the origin of the chemical elements and, finally, our own origins, we need to understand the structure and evolution of stars.

Modelling stellar evolution

Most of our understanding of stellar evolution is based on rather simplified numerical models (see also STELLAR INTERIORS); interestingly, despite their simplifications, the models are generally in good overall agreement with the available observational data. The models are mostly taken to be spherically symmetric (thus neglecting effects of rotation) and most often neglect possible mass loss or accretion. Also, many potential hydrodynamical instabilities, which might cause mixing of regions of otherwise inhomogeneous composition and hence affect the evolution, are ignored. In contrast, the detailed properties of stellar matter such as equation of state, radiative transport and energy generation, often referred to as the microphysics, are treated in considerable detail. This defines what might be called the 'standard' theory of stellar evolution and will be the basis for most of the discussion here. These simplified models are in fact able to reproduce quite successfully many aspects of the observed properties of stars. In particular, the models are in reasonable agreement with the detailed helioseismic inferences of solar internal structure (see SOLAR INTERIOR: STANDARD MODELS), although there remain highly significant discrepancies indicating remaining, and so far not definitely identified, deficiencies in the calculations. In the case of other stars, the comparison between observations and models is generally based just on the overall surface properties of the stars, which may well hide problems of potentially serious consequences in the modelling of late stages of stellar evolution.

Further complications arise in the, not uncommon, case where stars are members of close binary-star systems. Here there are dramatic episodes of mass transfer, when one star expands as a result of its evolution, completely changing the subsequent evolution of the system. The result is a bewildering richness of evolutionary scenarios, which is only beginning to be explored; many types of such interacting binary stars have been identified observationally, including also possible end products of the evolution, such as binary pulsars (e.g. Hulse–Taylor pulsar). This is discussed briefly at the end of the article.

Basic properties of stars

The properties of stellar interiors are treated in some detail in STELLAR INTERIORS and SOLAR INTERIOR: STANDARD MODELS as well as in several articles on the physical properties of stars. Here some simple aspects are considered which will be useful in the following.

The ways in which energy is transported in a star, from the central region where energy generation takes place to the surface, play an important role in its structure and evolution. Very generally, energy transport determines the temperature gradient in the star; as in any physical system, heat flows towards lower temperature, at a rate determined by the effective thermal conductivity and the decrease in temperature. The precise relation depends on the mechanism of heat transport. In large parts of most stars, energy is transported by radiation; here the rate of energy transport is characterized by the opacity κ, defined such that the mean free path of a photon is $1/\kappa\rho$, ρ being the density (see SOLAR INTERIOR: EQUATION OF STATE AND OPACITY). However, in some parts of the star the temperature gradient required for radiative transport becomes so steep that an unstable situation arises: matter displaced outwards, adiabatically, finds itself lighter than the surroundings and hence continues to rise. In this way *convective motion* is set up, whereby rising hot elements of fluid and cold sinking elements dominate the energy transport (for a more precise definition of the instability condition, and a description of how convection may be treated, see STELLAR INTERIORS and SOLAR INTERIOR: CONVECTION THEORY).

The distribution of convective regions plays a major role in the evolution of stars. This is illustrated for ZAMS stars in figure 2. In low-mass stars the comparatively low temperature in the outer layers results in a high opacity and hence in a convective envelope, whose depth increases rapidly with decreasing mass of the star; at $M \lesssim 0.25 M_\odot$ the stars become fully convective. In a $1 M_\odot$ ZAMS star the convective envelope contains only about 3% of the mass, although it extends over almost 30% of the radius. Stars only slightly more massive than the Sun develop convective cores, the extent of which increases with increasing mass. This is caused by an increasing concentration towards the center of energy generation, reflected in figure 2 by the curves marking where 50% and 90% of the luminosity is generated. The change in the distribution of energy generation, in turn, arises because of a change in the dominant processes by which hydrogen is fused to helium (see also SOLAR INTERIOR:

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

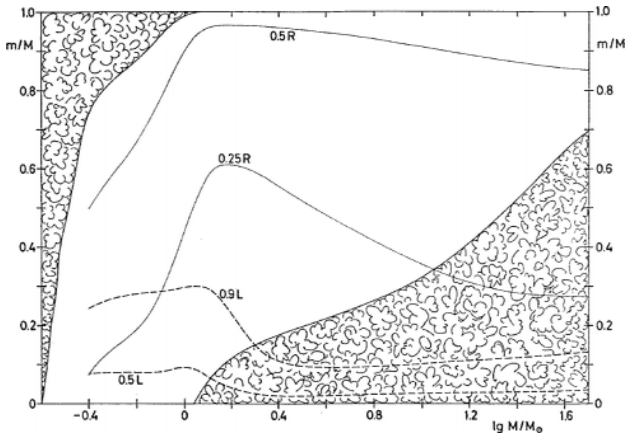


Figure 2. Structure of ZAMS stars. The horizontal axis shows the logarithm of the mass of the star (in solar units), whereas the vertical axis indicates position in the star, in terms of the mass m interior to a given point. 'Cloudy' areas indicate the presence of convection. The two solid curves show the positions where the distance r to the center is 0.25 and 0.5 times the surface radius. The dashed curves show the positions interior to which 50% and 90% of the stellar luminosity is generated.

ENERGY GENERATION). At lower mass, for $M \lesssim 1M_\odot$, the reactions are dominated by the pp chains, controlled by the fusion of two hydrogen nuclei, with a moderate temperature sensitivity. In contrast, at higher mass the dominant reaction is the CNO cycle, where the controlling reaction is the fusion of hydrogen and ^{14}N; the higher Coulomb barrier makes this reaction far more temperature sensitive than the pp chains, leading to the stronger concentration of the reactions towards the center.

Realistic studies of stellar structure and evolution requite detailed numerical calculations. However, several aspects can be understood, at least in a qualitative sense, from simple scaling relations. (These can be given a more precise meaning under rather restrictive assumptions on the physics of the stellar interior.) From the condition of hydrostatic equilibrium it may be shown that the internal pressure P scales as

$$P \sim \frac{GM^2}{R^4}$$

where G is the gravitational constant and M is the mass of the star. Assuming also that the stellar matter approximately satisfies the ideal gas law, the internal temperature may be estimated as

$$T \simeq \frac{GM\mu}{\mathcal{R}R} \tag{1}$$

where μ is the mean atomic weight of the constituents of the gas and $\mathcal{R}$ is the gas constant.

It is important to realize the meaning of these relations. They do not provide accurate estimates of conditions in stellar interiors (although, for example, the relation for T does reproduce the central temperature of detailed stellar models reasonably well). However, they *do* indicate how these conditions depend on the overall properties of the stars, i.e. their masses, radii and chemical composition. For example, equation (1) shows that as the star grows older, converting hydrogen into helium and thereby increasing μ, there is a general tendency for the temperature to increase.

If energy transport occurs predominantly by radiation, equation (1) can be used to estimate the luminosity of the star. The result depends on the detailed dependence of opacity on temperature and density. In relatively low-mass stars where the temperature is generally rather low, the opacity is approximately given by *Kramers' approximation*, $\kappa \propto Z(1 + X)\rho T^{-3.5}$, where X and Z are the abundances by mass of hydrogen and elements heavier than helium, respectively. In higher-mass, and hence hotter, stars, on the other hand, the opacity is dominated by scattering on free electrons, $\kappa \simeq 0.02(1 + X)$ m^2 kg^{-1}. Accordingly, we obtain as estimates of the luminosity

$$L \propto (1+X)^{-1}Z^{-1}M^{5.5}R^{-0.5}\mu^{7.5} \qquad \text{low-mass stars} \tag{2}$$

$$L \propto (1+X)^{-1}M^{3}\mu^{4} \qquad \text{high-mass stars.} \tag{3}$$

From an observational point of view, it is convenient to characterize the star by its surface luminosity and temperature. More precisely, one introduces the *effective temperature* $T_{\rm eff}$ of the star, as the surface temperature of the star at the given radius and luminosity, had it been radiating as a black body (cf EFFECTIVE TEMPERATURE SCALE AND BOLOMETRIC CORRECTIONS). Thus the radius R, luminosity L and effective temperature $T_{\rm eff}$ are related by

$$L = 4\pi\sigma R^2 T_{\rm eff}^4 \tag{4}$$

where σ is the Stefan–Boltzman constant. Hence equation (2) may also be written as

$$L \propto (1+X)^{-0.8}Z^{-0.8}M^{4.4}T_{\rm eff}^{0.8}\mu^{6} \qquad \text{low-mass stars.} \tag{5}$$

This reflects, for example, roughly speaking the relation between the change in luminosity and effective temperature as the star evolves towards the red-giant branch. For higher-mass stars, evidently, the corresponding relation is given directly by equation (3).

These expressions for the luminosity are based on the assumption that the stars are dominated by radiative energy transport. As the stars approach the red-giant branch they develop deep outer convective envelopes and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hence this assumption no longer holds. For predominantly convective stars, the structure is to a large extent controlled by the surface conditions. As a result, the stars occupy the *Hayashi track*, at an effective temperature which depends little on the luminosity or mass of the star. This corresponds to the nearly vertical red-giant branch in figure 1; the Hayashi track joins the main sequence at low stellar mass where, as noted above, the stars become fully convective. Also, it may be shown that the region on the cool side of the Hayashi track is 'forbidden': it is not possible to construct stable stellar models in hydrostatic equilibrium in this region.

Stellar energy sources

Gravitational contraction reduces (makes more negative) the gravitational potential energy of the star. It follows from *the virial theorem* that of the energy so liberated approximately half goes into increasing the thermal energy of the gas whereas the other half is lost from the star, in the form of radiation from the surface. In general, the increase in the thermal energy leads to an increase in the temperature of the stellar interior. It is this increase in temperature which allows the nuclear fusion reactions to take place.

Fusion of two nuclei requires that they overcome their mutual Coulomb repulsion. For this to be possible, their thermal energy has to be sufficiently high, requiring a temperature of at least several million K. Even so, the energy of the nuclei is far below the Coulomb energy of repulsion when the two nuclei are brought so close together that they touch. Thus, according to classical mechanics no reaction is possible. What enables the reactions to take place is the quantum-mechanical phenomenon of *tunnelling*, leading to a finite probability for the two nuclei to interact even if according to classical mechanics they are kept apart. However, in most stars this is an exceedingly slow process: the average lifetime of a given hydrogen nucleus at the center of the Sun is of order 5 Gyr (billion years). Consequently, the rate of energy release per unit mass is modest: the average rate of energy generation in the solar core is only around 0.001 W kg^{-1}, about 3 orders of magnitude below that of an average human. On the other hand, because of the large mass of the Sun, the total energy generation is enormous, around 4×10^{26} W.

However, fusion can only provide energy as long as the nuclei become more tightly bound when they fuse. The nuclear binding energy has a maximum among the iron-group elements; thus the nuclear energy generation stops when matter has been converted into these elements. At this point gravitational contraction is the only remaining effective energy source: the central parts of the star undergo catastrophic collapse, releasing vast amounts of gravitational potential energy and causing the star to explode as a supernova (see SUPERNOVAE). This stage, however, is only reached for massive stars; in stars of mass below around 10 times the mass of the Sun, the core temperature never becomes so high that the final nuclear reactions take place. Such stars undergo substantial mass loss, leaving behind the compact core of the star, in the form of WHITE DWARFS, which subsequently cool through radiation of its internal thermal energy. This is the likely final fate of the Sun, for example.

It is instructive to consider the amount of energy available to the star from gravitational contraction and nuclear reactions. The typical energy, per unit mass, available from gravitational contraction is determined by the gravitational potential of the star, whose magnitude may be estimated as GM/R, where G is the gravitational constant, and M and R are the mass and surface radius of the star. Using the solar mass, $M_\odot = 1.989 \times 10^{30}$ kg and the solar radius $R_\odot = 6.96 \times 10^8$ m, the result is 2×10^{11} J kg^{-1}. The energy available from fusion may be obtained, according to the relativistic equivalence between mass and energy, from the difference in mass between the nuclei taking part in, and the nucleus resulting from, the reaction. In fusion of hydrogen to helium, the energy lost is approximately 0.7%. Thus the available energy from this reaction, per unit mass, is $0.007c^2 = 6 \times 10^{14}$ J kg^{-1}, where c is the speed of light. The maximal energy available from fusion results from the, hypothetical, reaction where hydrogen fuses to iron-group elements, the most tightly bound nuclei; the resulting relative mass difference is 0.9%, corresponding to 8×10^{14} J kg^{-1}. (It should be noticed that the first of the series of fusion reactions, from hydrogen to helium, produces 75% of the total amount of energy available.)

From these energy estimates we can obtain rough estimates of the characteristic time scales of different phases of stellar evolution. The total energy available to a $1M_\odot$ star from gravitational contraction is, assuming again solar radius, approximately 4×10^{41} J; at the solar luminosity of around 4×10^{26} W, this would last for about 10^{15} s, or around 30 Myr (million years). This, therefore, is the characteristic lifetime for a star whose energy comes from gravitational contraction; it is generally known as the *Kelvin–Helmholtz time scale*. (From the fact that it is much shorter than the age of the solar system, around 5 Gyr, it immediately follows that other sources of energy are required for the Sun.) To estimate the total energy available from nuclear reactions, we must take into account that only in around 10% of the mass of a star is the temperature sufficiently high for the reactions to be efficient; thus the total energy produced from hydrogen fusion, in a $1M_\odot$ star, is around 1.2×10^{44} J. At solar luminosity this corresponds to a lifetime of 3×10^{17} s, or 10 Gyr. Thus the Sun is approximately half-way through the phase of central hydrogen fusion. It should also be noticed that since the luminosity of stars generally increases with age, and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

given the smaller energy release in later fusion reactions, the phase of hydrogen fusion occupies the dominant part of a star's active life; thus we may expect that most of the stars observed are in this phase.

It should finally be noted that collapse to a much smaller radius may release very large amounts of energy. This is the case during a supernova explosion, where the central parts of the star collapse to nuclear densities, resulting in a core with a mass of order $1M_\odot$ and a radius of around 10 km. Here the gravitational potential is of order 10^{16} J kg^{-1}, i.e. around 10% of the rest-mass energy. The total energy released in this collapse, of order 2×10^{46} J, is what powers the supernova explosion.

Early life of a star

The birth of a star takes place in a dense interstellar cloud (a so-called *molecular cloud*), when a part of the cloud becomes unstable to gravitational collapse and begins to contract. The initial phases of the cloud collapse occur on dynamical time scales and undoubtedly involve effects of turbulence, magnetic fields, rotation, mass loss and accretion, none of which is understood in detail. These processes give rise to spectacular observational effects such as extended jets, and the formation of disks around young stellar objects which may later develop into planetary systems. For further details on these processes, see STAR FORMATION.

In the early phases the evolution is dominated by the interaction between the protostar and the surrounding material which continues to be accreted (see PRE-MAIN-SEQUENCE STARS). This also determines the main observational characteristics of the forming star, since it is shrouded in this material and hence predominantly radiates in the infrared.

As the main parts of the star come into hydrostatic equilibrium the simplified treatment of the 'standard' modelling becomes at least partly applicable. The energy output from the star comes predominantly from gravitational contraction, with half the gravitational potential energy released going to heat matter in the star and the other half being radiated from the surface. However, fusion of deuterium with ordinary hydrogen, which can take place at relatively modest temperatures, also plays a significant role. In the early parts of this phase the star is still relatively cold; as a result, the opacity is high and the star is completely convective. As discussed above, such stars are on the Hayashi track, moving down it with the continuing contraction.

With the increasing central temperature the opacity is reduced until the core regions of the star become convectively stable. As this region with radiative transport expands, the star moves away from the Hayashi track, towards higher effective temperature. During this phase, the evolution essentially follows the behavior given by equations (3) or (5). Finally the central temperature reaches a point where nuclear reactions may provide the required energy and the star settles down on the main sequence. As is clear from figure 3, this occurs in two steps for more massive stars: the initial step involves the fusion of ^{12}C and hydrogen, which occurs at a slightly lower temperature than the main processes of hydrogen fusion. After ^{12}C has been consumed the contraction continues until the main reactions start.

The time scale for these initial phases is essentially given by the Kelvin–Helmholtz time scale; for a $1M_\odot$ star evolution until the start of hydrogen burning takes around 50 Myr.

The onset of nuclear reactions requires that the central temperature reaches around 6 million K. In very low-mass objects this never happens: here matter becomes strongly degenerate and the thermal energy provided by the gravitational contraction is taken up by the degenerate electrons, rather than increasing the temperature. Such objects end up as *brown dwarfs* (see PLANETARY COMPANIONS AND EXTRASOLAR BROWN DWARFS), luminous because of the continuing contraction. The critical mass below which stable nuclear burning is not achieved is around $0.08M_\odot$.

Evolution on the main sequence

The evolution on the main sequence is driven by the gradual conversion of hydrogen to helium in the core of the star. This increases the mean molecular weight μ of the gas; to maintain the pressure, the density has to increase through contraction of the core, and this in turn causes an increase in temperature. As a result the rate of energy generation and the efficiency of radiative energy transport increase, causing an increase in the luminosity of the star; this is clearly visible in figure 1. Indeed, the increase in luminosity with increasing μ follows directly from equations (2) and (3).

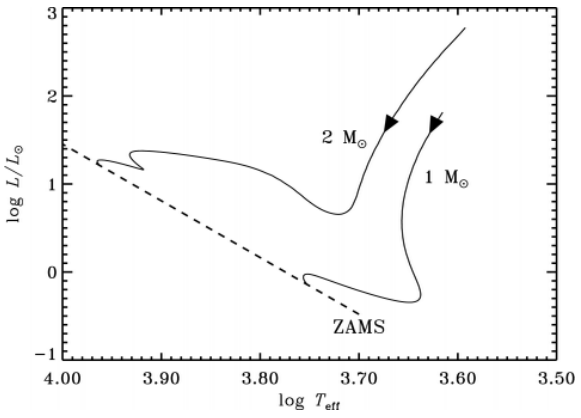


Figure 3. Pre-main-sequence evolution tracks for $1M_\odot$ and $2M_\odot$ stars. The initial evolution is down the Hayashi track but continues to higher effective temperature as a radiative core is established. The tracks end on the ZAMS where hydrogen fusion begins to contribute most of the stellar luminosity. (Data kindly provided by D A VandenBerg.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In the case of the Sun, the luminosity has increased by approximately 30% from the ZAMS to the present. This change is sufficiently large to have substantial effects on the Earth's climate, unless other factors compensate; indeed, the effect has been known as 'the faint early Sun problem' and various attempts have been made to circumvent it. In fact, the luminosity increase is an essentially unavoidable consequence of the basic physics of stellar evolution, as is also suggested by the simple analysis above. Rather than tinkering with the stellar models, a far more likely solution to the problem would be changes in the Earth's atmosphere, e.g. such that the greenhouse effect was stronger in the past. Even so, it is interesting that feed-back mechanisms are apparently at work in the Earth's climate such as to maintain a nearly stable temperature despite the changes in solar radiation.

While the central parts of the star contract, the outer parts expand. This is visible from figure 1, using equation (4), by noting that the effective temperature remains constant or decreases while the luminosity increases. Such opposite behavior of the core and envelope of a star is a characteristic feature of stellar evolution, for which no satisfactory simple explanation seems to be available.

The details of the evolution on the main sequence, in particular how it ends, depend strongly on the mass of the star, as is also seen in figure 1. This is related to the possible presence of a convective core and its effects on the composition profile in the central parts of the star. Figure 4 shows the hydrogen abundance X by mass at various ages for stars of masses $1M_\odot$ and $2.5M_\odot$. In the $1M_\odot$ star the core is radiative and there is no mixing (within the assumptions of the 'standard' model). Hence the hydrogen abundance decreases most rapidly at the center, where the temperature is highest. When hydrogen is exhausted at the center, energy production still continues in a region immediately away from the center, and hence there is a gradual transition to the subsequent evolution phase, with fusion in a shell around a core consisting of helium and heavier elements. In contrast, the convective core in the $2.5M_\odot$ star is fully mixed on a time scale much shorter than the fusion time scale, and hence the hydrogen abundance decreases uniformly throughout the core. Although, as shown, the core generally shrinks with increasing age, hydrogen is essentially exhausted simultaneously throughout the central energy-producing region. In the final stages of central hydrogen fusion, the star contracts in an attempt to maintain the energy production by increasing the central temperature; this is visible in the evolution tracks in figure 1 by the left-ward 'hook', which is reminiscent of the earlier contraction phase before the main sequence (cf figure 3). Finally, as the central hydrogen is being totally exhausted, the star establishes a shell source around the helium core. The final evolution corresponding to the hook is rapid, compared with the preceding main-sequence evolution; thus few stars are found in this phase, as can be verified from observations of stellar clusters (see STAR CLUSTERS).

It is evident from figure 1 that the luminosity of the stars increases rapidly with stellar mass. This is consistent with equations (2) and (3). It also has dramatic implications for the time spent by the stars on the main sequence. Since the available nuclear energy is, approximately, proportional to the mass of the star, the nuclear lifetime scales as M/L or, according to the estimates of L, as M to an inverse power of at least 2. In fact, for stars of mass less that $0.9M_\odot$ the main-sequence lifetime exceeds the current age of the universe, while stars of more than $60M_\odot$ end their evolution in less than 4 million years.

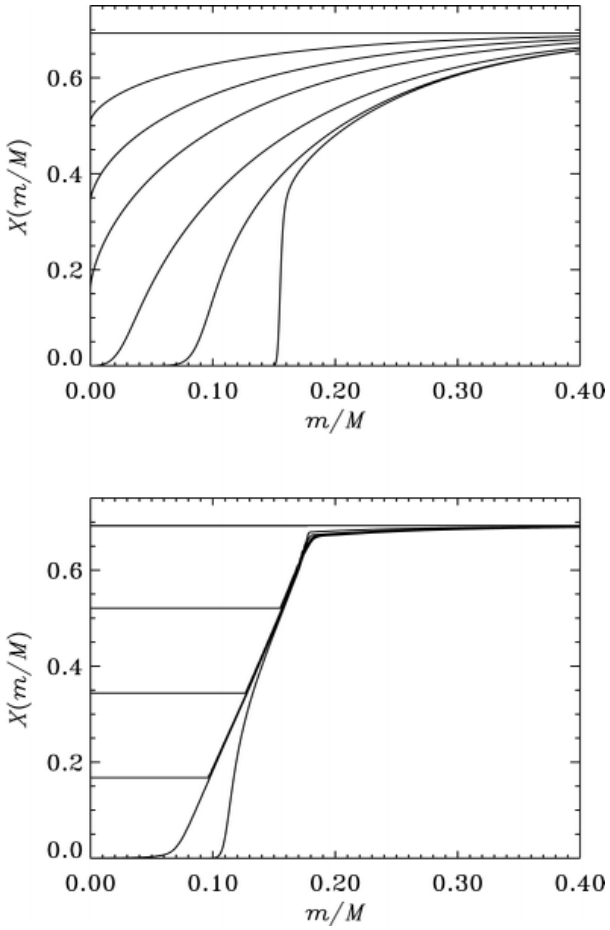


Figure 4. Evolution in the hydrogen abundance X by mass, as a function of time, against the fractional mass m/M in the star. The homogeneous initial models had $X = 0.692$. Top panel: $1M_\odot$, at ages 0, 2.47 Gyr, 4.53 Gyr, 6.49 Gyr, 9.29 Gyr, 10.81 Gyr and 11.62 Gyr. Lower panel: $2.5M_\odot$, at ages 0, 174 Myr, 306 Myr, 400 Myr, 462 Myr and 487 Myr.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Evolution after core hydrogen burning

After the end of central hydrogen fusion, the star has a helium core surrounded by a *shell source* where hydrogen fusion continues. As the shell source consumes hydrogen, the helium core grows and as a result contracts under gravity. This leads to heating of the core until the temperature is reached where the next fusion reaction, where three helium nuclei combine to form carbon, can take place. The star now has two sources of energy generation: the core helium fusion and the shell hydrogen fusion, with the latter generally supplying the larger part of the energy. This sets the pattern for the subsequent evolution and leads to what can be regarded as an onion structure of the star: central nuclear fusion alternating with core contraction and heating, and the establishment of shell sources outside regions where a given 'fuel' has been exhausted. An example of this structure is illustrated schematically in figure 5.

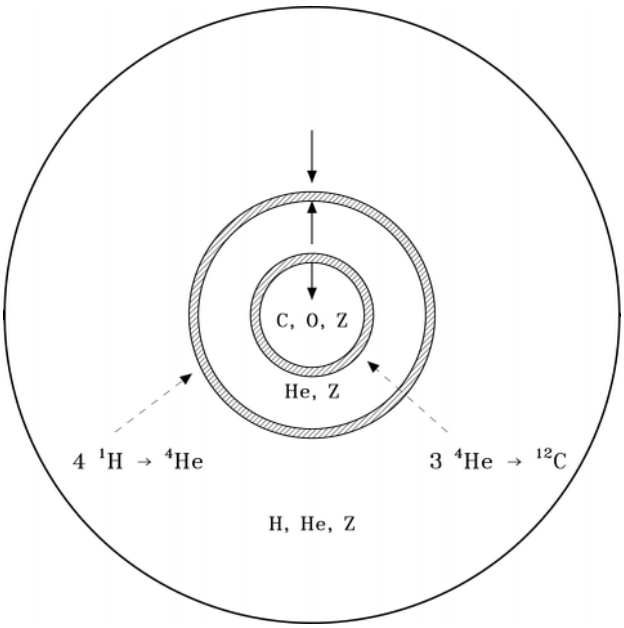


Figure 5. Schematic structure of a star with hydrogen and helium shell sources. The vertical arrows indicate the expansion or contraction of the corresponding regions. The thin hatched regions indicate the hydrogen- and helium-fusion shell sources.

The shell-burning 'law'

The behavior of the star during these changes in the core can to some extent be rationalized in terms of a simple principle, known as the 'mirror principle' or 'the shell-burning law'. This principle states that contraction or expansion reverses at a shell source: thus if the region inside a given shell contracts, the region outside the shell expands, and vice versa. No entirely satisfactory simple explanation has been given for this principle, but it is strikingly successful in explaining computed stellar evolution. It should be noticed that the same tendency for core contraction and envelope expansion was found during the main-sequence evolution phase.

The effect of this principle is immediately visible in the evolution tracks in figure 1. After the establishment of the hydrogen shell source, the contraction of the helium core leads to an expansion of the envelope; as the star is still predominantly radiative, luminosity and effective temperature are related by equation (3) or (5), and hence, with the expansion, the effective temperature decreases and the star moves towards the red-giant branch. When the star reaches the Hayashi track, a further decrease in $T_{\rm eff}$ would take the star into the 'forbidden' region on its cool side; since this is not possible, the continuing expansion takes the star up the red-giant branch. It is quite striking that these changes in the surface properties are essentially a reversal of the evolution before the main sequence. This phase continues to the point where the core temperature is sufficiently high for helium fusion to carbon to take place. In the solar case, that leads to a radius close to the radius of the Earth's orbit and a luminosity of approximately $2500L_{\odot}$.

The subsequent evolution depends rather sensitively on stellar mass. In many ways the intermediate-mass case is the simplest and is therefore considered first.

Later evolution of intermediate-mass stars

This covers stars in the mass range of roughly $(2–10)M_{\odot}$. An example of an evolutionary track is shown in figure 6. The evolution up to helium ignition follows the general description given above although the luminosity at the tip of the red-giant branch will not in general be as high as in the solar case. When helium fusion starts, the core expands as a result of the extra supply of energy. This expansion also helps stabilizing the temperature, such that helium fusion takes over in a controlled fashion. According to the shell-burning law, the core expansion leads to a contraction of the envelope, reflected in the evolution down the red-giant branch and later towards higher effective temperature. The relation between effective temperature and luminosity is again largely controlled by the usual relations, so evolution proceeds parallel to, although at a somewhat higher luminosity than, the earlier part of the track. In fact, during this stage the hydrogen shell source still contributes more than 70% of the total luminosity. With the continuing helium fusion, the core starts to contract again, and the envelope consequently expands, leading to evolution towards lower $T_{\rm eff}$.

This continues to the point where helium is exhausted in the core, which now consists of carbon, oxygen (also a product of the helium fusion) and heavier elements. As in the case of hydrogen, the core exhaustion of helium is followed by the establishment of a helium shell source. The situation is now as illustrated schematically in figure 5, with an inert and contracting

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

carbon–oxygen core. It is tempting to apply the shell-burning law to the *two* shells, as indicated by the arrows: the intermediate region might be expected to expand while the envelope of the star contracts. In fact, the expansion of the intermediate region leads to a reduction of the temperature in the region of the hydrogen shell source which therefore becomes inactive. At that point energy is provided essentially only by the helium-burning shell; the contraction of the core leads to a strong expansion of the envelope and the star moves back to and up the Hayashi track on the so-called *asymptotic giant branch*, to very high luminosity (cf. figure 6).

As the star evolves, the hydrogen shell is reactivated and takes over as the dominant source of energy. The subsequent detailed evolution is in fact complex, with developing thermal instabilities that cause both shell sources to be active in an oscillatory fashion. As shown in the insert in figure 6, this causes a large number of oscillations in the evolution track. The resulting changes in location of convective regions play an important role in the synthesis of the heaviest elements (beyond the iron group) through neutron capture and in bringing the products of these reactions to the surface of the star, where they can be observed as abundance anomalies.

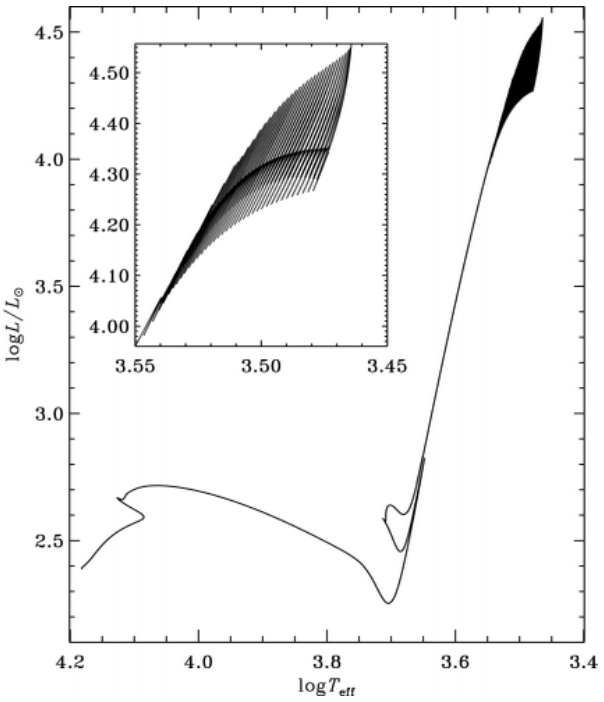


Figure 6. Evolution track for a $4M_\odot$ model, from the ZAMS to the asymptotic giant branch. As discussed in the text, the behavior in the diagram can be understood in terms of changes in the regions of nuclear energy generation. The insert shows the last phases in more detail, to illustrate the oscillations caused by the thermal instability between the hydrogen and helium shell sources. (Data kindly provided by A V Sweigart.)

An observation that three metal-poor stars are more rich in lead than in any element heavier than iron fits with the prevailing theory about the late, 'asymptotic giant branch' (AGB) phase of the evolution of stars 0.8–8 times the size of our Sun. It is thought that nuclei—mainly iron—deep inside an AGB star slowly capture neutrons, progressively building up heavier elements in what is known as the 's-process', which in the case of old stars will produce a pile-up of synthesized material with atomic weights between lead and bismuth. The new data (Van Eck *et al*) support this view, which hitherto had little observational support.

It might perhaps be expected that the further evolution would follow a pattern similar to the onset and development of helium burning, i.e. that the contraction of the core would lead to the ignition of the next nuclear reaction, fusion of carbon, followed by the establishment of yet another shell source. However, in this mass range the core never becomes hot enough for carbon fusion to start. There are two reasons for this. Firstly, the density of the contracting carbon–oxygen core becomes so high that the electrons become strongly degenerate; as in the case of low-mass pre-main-sequence stars this tends to limit the temperature increase, as the thermal energy obtained from the contraction goes into the degenerate electrons. Secondly, the core is cooled by various neutrino-generating processes in very dense plasmas; the neutrinos escape the core essentially unhindered, removing thermal energy.

The final stages of evolution for intermediate-mass stars are quite dramatic, although the underlying physics is poorly understood. The very extended outer envelope becomes unstable towards rapid mass loss, which over a period of probably less than 50 000 yr strips off essentially the entire envelope outside the carbon–oxygen core. That core remains behind as a very compact, degenerate object with a mass of $(0.5–1)M_\odot$ and a radius comparable with that of the Earth. It is initially extremely hot, with a surface temperature exceeding 100 000 K; it radiates predominantly by losing thermal energy, gradually cooling down. Such objects develop into white dwarfs. The matter ejected around the star shines for a few thousand years as a PLANETARY NEBULA, as a result of fluorescence caused by the ultraviolet radiation from the central object, before the matter is dispersed in the interstellar medium.

Later evolution of low-mass stars

For stars of mass below about $2M_\odot$ the onset of helium burning, at the tip of the red-giant branch, takes place in a rather more dramatic way than at higher masses. The main difference is that at lower masses the helium core is dominated by the pressure of degenerate electrons. The core is still heated, although mainly by the increasing temperature in the hydrogen shell source as it adjusts to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

provide the high luminosity of the star. When the ignition temperature of helium (around 10^8 K) is reached, helium fusion starts, the release of energy causing an increase in the temperature of the gas. Since electron degeneracy dominates the pressure, it does not increase as a result of the temperature increase; this is unlike the situation in more massive stars, where the temperature increase causes an increase in pressure, which in turn leads to an expansion of the core and hence a stabilization of the burning. In the low-mass case, the temperature increase is not so controlled but rather leads to a further increase in the helium fusion rate and hence in the temperature. The result is a thermal run-away, in what is called *the helium flash*: during a few hours the rate of energy generation in the core reaches of order $10^{10}L_\odot$, comparable with the total luminosity of the Galaxy. This ends when the temperature of the core increases sufficiently that electron degeneracy becomes unimportant; at that point the core expands, reducing the temperature and stabilizing the burning.

The evolution of the outer parts of the star during the helium flash is rather uncertain. Most of the energy released during the flash is probably absorbed in an expansion or partial ejection of the already very extended envelope, on a dynamical time scale. When the star finally settles down to stable helium fusion, it is on the horizontal branch (see HORIZONTAL BRANCH STARS), at a position which probably mainly depends on the extent of the remaining hydrogen envelope.

The subsequent evolution is rather similar to that of the intermediate-mass stars discussed above; after completing central helium fusion, the star in general has both a hydrogen and a helium shell source, evolving through loops in the Hertzsprung–Russell diagram and finally up the asymptotic giant branch. During this phase of the evolution of the Sun, its radius is likely to exceed the radius of the Earth's orbit. As in the previous case, degeneracy and neutrino cooling prevent the core temperature reaching the values required for carbon fusion, and the star ends as a carbon–oxygen white dwarf, surrounded briefly by a planetary nebula.

Later evolution of high-mass stars

For masses higher than around $10M_\odot$ the core temperature after the end of helium fusion continues to rise, to a point where carbon fusion and later fusion reactions can take place. The end result is that the star ends its life in a supernova explosion (see SUPERNOVAE for further details).

The final series of nuclear reactions, starting with carbon fusion and leading up to the formation of a core consisting of iron-group elements, takes only a few hundred years, short even compared with the lifetime of massive stars. With the formation of the iron core the star has exhausted the possibilities of energy generation through nuclear processes: these elements have maximal nuclear binding energy, and further fusion would require energy. As a result, the core can only make up for its continuing energy loss through gravitational contraction; this causes further heating, to the extent that iron nuclei are dissociated to their basic constituents, protons and neutrons. The energy required for the dissociation further accelerates the contraction which takes the form of a dynamical collapse. With the increasing density, it becomes energetically favorable for the electrons to combine with the protons to form neutrons, with the emission of neutrinos. The collapse only halts when the core reaches nuclear densities and the neutrons become degenerate; the shock waves generated at the resulting sudden stop of the collapse, together with the neutrinos from the electron capture, transfer energy to the envelope which is ejected in the supernova explosion.

The energy powering the explosion comes entirely from the gravitational energy released in the collapse of the central $(1–2)M_\odot$ to a radius of around 10 km. In fact, the energy emitted in the form of light is only a small fraction of the total energy available; the largest fraction is taken up by the neutrinos emitted in the collapse.

The matter ejected in the explosion forms an interstellar gas cloud which is gradually dispersed (see SUPERNOVA REMNANTS). It is enriched by elements formed in the late stages of nuclear reactions, including rapid neutron capture during the explosion. The core of the star is left as an extremely compact object. If its mass is below $(1.5–2)M_\odot$ it forms a neutron star (see NEUTRON STARS), dominated by degenerate neutrons, and with a density corresponding to the density of nuclear matter. Such objects are observed as PULSARS. If the core is more massive, even the pressure from the nuclear matter cannot resist gravity, and the core collapses to form a BLACK HOLE.

Tests of stellar evolution

The evolution sequences presented above provide a detailed history of stars, solidly founded in the 'standard' theory of stellar evolution. However, *testing* these results in detail is difficult. Helioseismology has provided very detailed inferences of the solar internal structure, which are generally well matched by the models. Similarly detailed information is entirely out of reach for distant stars, whose surfaces, even with planned space experiments, will only be resolved to a limited extent. On the other hand, studies of a broad range of stars allow tests of our understanding of stellar evolution under conditions far more extreme than found in the Sun and including phenomena, such as convective cores, which are not relevant in the solar case.

Substantial efforts are being invested in asteroseismic investigations of stars, using oscillation frequencies to probe stellar interiors. Detailed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

information has been obtained in the case of white dwarfs, leading to accurate determinations of their masses, as well as estimates of the thickness of the outer hydrogen and helium layers. Studies of double-mode Cepheids demonstrated gross errors in the opacity tables, subsequently confirmed by improved opacity calculations; this has led to substantial revisions also of other aspects of stellar modelling. For other types of stars, data on oscillation frequencies are being accumulated, although uncertainties in the mode identification have so far precluded their use for detailed studies of stellar interiors. This is likely to change with the ongoing and planned observational and theoretical efforts, including observations from space. In 2001, solar-like oscillations in the stars beta Hydri and alpha Centauri A were detected.

More classical observations have also provided stringent tests of stellar-evolution calculations, at least as far as the global properties of the stars are concerned. Particularly important are observations of binary stars in eclipsing variables, where accurate determinations of the stellar masses and radii are possible (see ECLIPSING BINARY STARS). This, together with the highly probable assumption that the stars have the same age and original chemical composition, places strong constraints on the stellar models.

In almost all cases the relevant time scales are too long to allow direct observation of the consequences of stellar evolution. However, a powerful alternative is provided by observing ensembles of stars of the same age but different masses: since the evolution time scale varies strongly with mass such a collection of stars contains stars in a broad range of evolutionary stages. Such ensembles are provided by star clusters, collections of stars that are clearly physically associated. It may generally be assumed that the stars in a cluster were formed from the same interstellar cloud at approximately the same time. Thus they have a common age and original chemical composition, while differing in mass. Furthermore, all stars in a cluster are at approximately the same distance. Thus the correction from the true luminosity of the stars to the apparent luminosity at the (generally *a priori* unknown) distance of the cluster is the same for all the stars.

The observed distribution of stars in a cluster can be compared directly with computed *isochrones*, i.e. curves corresponding to models of the same age but different mass. This comparison also provides an estimate of the age of the cluster; furthermore, the shift between the computed luminosity and the observed apparent luminosity of the main-sequence stars is a measure of the distance to the cluster. An interesting example, for the globular cluster M92, is illustrated in figure 7. The distribution of observed stars clearly reflects the evolutionary pattern: stars of the lowest masses are still on the main sequence, with core hydrogen burning. Somewhat more massive stars, where hydrogen is exhausted at the center, occupy the subgiant branch, starting at the turnoff, and the red-giant branch, whereas the stars originally of the highest mass have gone through a helium flash and are found on the horizontal branch or, in a few cases, at the beginning of the asymptotic giant branch. The very large number of stars in a globular cluster ensures that all stages of evolution are well represented. Computed isochrones are shown for a few ages, a reasonable fit being obtained for an age of 14 Gyr (the shift of the red-giant branch relative to the observations may be due to an error in the atmosphere models used to calibrate the models from effective temperature to color index). It is remarkable that the inferred age is rather larger than some values of the estimated age of the universe, as obtained from cosmological measurements. This problem suggests serious deficiencies in either the commonly used cosmological models or in the calculation of stellar models.

A second example, for the open cluster NGC 3680, is shown in figure 8. In this case, the cluster has been strongly affected by dynamical evolution, leading to very substantial loss of stars (furthermore, in the analysis care was taken to include only single stars that definitely belong to the cluster); as a result the diagram is sparsely populated, compared with figure 7. Two isochrones are shown: one for a 'standard' evolution calculation and the second for models including penetration and mixing from the convective cores (see below). There is clearly strong evidence that the models with such additional mixing provide the better fit to the observations.

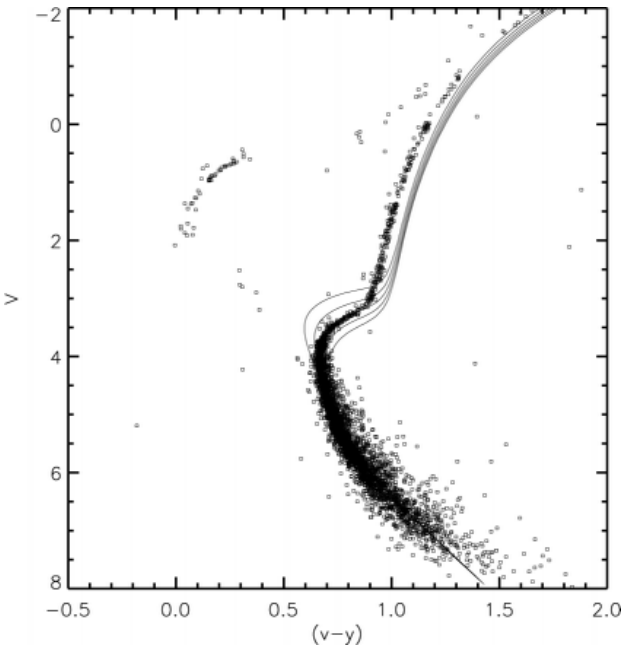


Figure 7. Hertzsprung–Russell diagram for the globular cluster M92. The abscissa is color index (in the Strömgren system), which provides a measure of the effective temperature, whereas the ordinate is apparent magnitude. The curves show isochrones corresponding to ages (from the top) of 10, 12, 14, 16 and 18 Gyr. (Kindly provided by F Grundahl.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

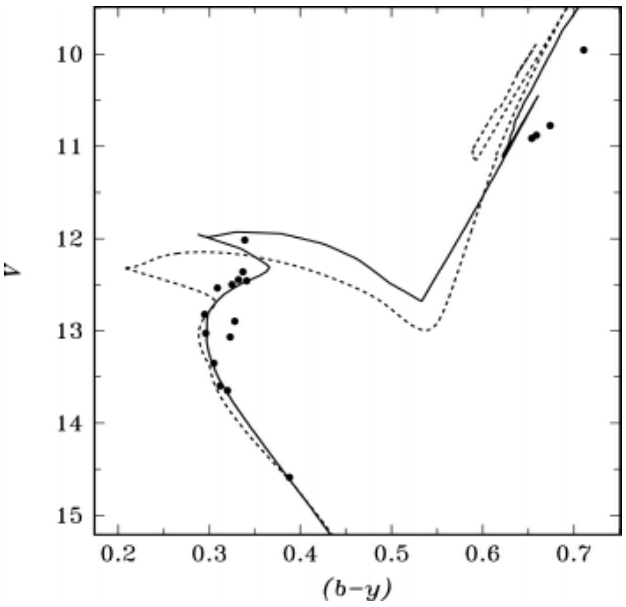


Figure 8. Hertzsprung–Russell diagram for the open cluster NGC 3680, showing apparent magnitude against color index. The dashed curve shows an isochrone of age 1.5 Gyr for a standard evolution calculation, whereas the isochrone shown as a solid curve, of age 1.7 Gyr, is based on models including penetration from the convective core. (Kindly provided by J Andersen.)

Beyond the 'standard' model

Although the analysis of stellar evolution, as presented here, is quite successful in rationalizing many observed properties of a wide range of stars, its intrinsic simplifications should be kept in mind. A few examples will be considered here.

Mass loss undoubtedly plays a very important role in many phases of stellar evolution. This has already been mentioned in connection with asymptotic-giant-branch stars, where rapid mass loss is observed and must presumably be involved in the formation of planetary nebulae and white dwarfs. Lower-mass main-sequence stars also show mass loss as in the solar wind (see SOLAR WIND: GLOBAL PROPERTIES) although probably not at a rate that substantially affects stellar evolution. However, in high-mass stars, the very strong radiation field seems to drive quite strong stellar winds, which are probably implicated in the formation of WOLF–RAYET STARS.

The hydrodynamics of stellar interiors is treated in a very simple manner in normal stellar-evolution calculations, yet could have important effects. The most important problems concern mixing of otherwise chemically inhomogeneous regions of the star. For example, there is little doubt that motion extends beyond the convectively unstable regions, in the form of convective penetration into the stable regions and possibly internal gravity waves induced by such penetration. This would cause mixing outside convective cores and hence would increase the amount of hydrogen available for fusion as well as the extent of the helium core left after the end of core hydrogen fusion; yet the extent of such penetration is highly uncertain. In stars above about $10M_{\odot}$, additional uncertainties are introduced by possible instabilities outside convective cores which may lead to the so-called *semiconvection*, probably causing additional mixing.

Stars are observed to rotate, often rapidly (see STELLAR ROTATION ON THE MAIN SEQUENCE), yet most evolution calculations ignore rotation. Rapid rotation deforms the star, with direct effects on its structure. However, even slow rotation may cause circulation (the so-called *Eddington–Sweet circulation*) which could cause mixing, and gradients in the rotation rate could lead to hydrodynamical instabilities that might also mix the material. Particular problems are related to the slowing down of rotation in low-mass stars from a likely state of initial rapid rotation, probably as a result of angular-momentum loss from the outer parts of the star in a magnetic stellar wind; this has at least in the solar case led to roughly uniform rotation of the solar interior (see SOLAR INTERIOR: ROTATION). However, the mechanisms coupling rotation of the deep solar interior to the surface, and their possible effects on other aspects of stellar evolution, are not understood.

Evolution in close binary systems adds a great deal of complication to this already complex situation (see CLOSE BINARY STARS). As the more massive of the two components evolves and expands, possibly reaching the red-giant stage, mass transfer to the other component, or mass loss from the system, may take place. This in turn causes changes in the orbital parameters of the system. With continuing evolution, mass may be transferred back and forth between the components, greatly affecting their evolution. The detailed behavior evidently depends sensitively on the initial parameters of the system. Mass transfer, particularly onto compact objects such as white dwarfs and neutron stars, also gives rise to spectacular observational phenomena (see CATACLYSMIC BINARIES). In particular, the so-called supernovae of type Ia are believed to occur when white dwarfs become unstable after accretion of matter from a companion star.

Acknowledgments

I am very grateful to J Lattanzio and A V Sweigart for illuminating discussions on late stellar evolution and to Sweigart, J Andersen, F Grundahl and D A VandenBerg for providing data for figures. I thank the many authors who have kindly provided drafts of their articles, which have substantially assisted the writing of this article.

Bibliography

Andersen J 1991 Accurate masses and radii of normal stars *Astron. Astrophys. Rev.* **3** 91–126

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Chiosi C, Bertelli G and Bressan A 1992 New development in understanding the HR diagram *Annu. Rev. Astron. Astrophys.* **30** 235–85

Hansen C J and Kawaler S D 1994 *Stellar Interiors. Physical Principles, Structure, and Evolution* (New York: Springer)

Iben I 1985 The life and times of an intermediate mass star—in isolation/in a close binary *Q. J. R. Astron. Soc.* **26** 1–39

Iben I and Renzini A 1983 Asymptotic giant branch evolution and beyond *Annu. Rev. Astron. Astrophys.* **21** 271–342

Kippenhahn R and Weigert A 1990 *Stellar Structure and Evolution* (Berlin: Springer)

Renzini A and Fusi Pecci F 1988 Tests of evolutionary sequences using color–magnitude diagrams of globular clusters *Annu. Rev. Astron. Astrophys.* **26** 199–244

VandenBerg D A, Bolte M and Stetson P B 1992 The age of the galactic globular cluster system *Annu. Rev. Astron. Astrophys.* **34** 461–510

Van Eck S, Goriely S, Jorissen A and Plez B 2001 Discovery of three lead-rich stars *Nature* **412** 793–795

Jørgen Christensen-Dalsgaard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Interiors

The subject of stellar interiors is concerned with the determination of the run of density, temperature, luminosity (the energy flow rate through a closed three-dimensional surface) and composition from the center to the surface of stars. The determination of these quantities requires the solution to the conservation laws of mass, momentum and energy at each point in the model. These conservation laws have terms which directly include the pressure of the material and the energy produced by nuclear reactions in the interior. In order to fully solve the energy conservation equation, a relationship between the thermal structure and the energy transport must also be specified. The relationship will be different for each of the three mechanisms—radiation, convection and conduction. From the density, temperature and composition, the pressure may be determined from atomic physics, the rate at which nuclear energy is generated from nuclear physics and the mode of energy transport and its efficiency from atomic (radiation) and classical (convection) physics.

The form that conservation of momentum takes in most stars is defined by the concept of hydrostatic equilibrium. Hydrostatic equilibrium is the condition that forces at all locations inside a star just balance. Most commonly, these are the internal pressure forces balancing the self-gravitational forces of the stellar material. Most stars satisfy hydrostatic equilibrium to a great degree, although there are stars which oscillate about hydrostatic equilibrium (pulsating stars), and stars that are either collapsing or exploding are far removed from hydrostatic equilibrium. Significant departures from hydrostatic equilibrium occur on the free fall time, which is generally so short (about an hour for the Sun) that they are readily observable.

The relationship between the density, temperature, composition and pressure is called the equation of state, and its functional form is quite different in different temperature–density regimes. For stars on the main sequence, those stars which are converting hydrogen to helium at the very center of the star, the relation is that of an ideal gas pressure produced by the motion of the nuclei and free electrons plus the radiation pressure produced by the photons. As stars evolve both the central density and the central temperature increase, and there may be other contributors to the equation of state. For densities that are sufficiently high (how high depends on the temperature), the pressure may be supplied by degenerate electrons, which are governed by the Pauli exclusion principle which holds that identical fermions cannot be too close together. The resistance of the identical electrons to being squeezed together provides the pressure. In very late stages of stellar evolution, at much higher densities, the pressure may be supplied by degenerate neutrons for the same reason.

Although the focus here has been on the pressure, rotation and magnetic fields also provide forces which should be included in hydrostatic equilibrium, but generally are not because they are generally of secondary importance and they complicate the mathematical problem significantly.

The conservation of energy requires that the amount of energy produced inside a volume equals the net amount of energy flowing out of the volume. This requires knowledge of nuclear reaction rates to determine the energy produced and knowledge of the energy transport mechanism to determine the energy transport. The energy-producing reactions for most stars are fusion reactions combining light nuclear species into heavier nuclear species, which occur in the deep interior. As the star ages it first converts hydrogen into helium at the center on the main sequence and then in a shell about a hydrogen-depleted region as a subgiant and then a RED GIANT STAR. The hydrogen burning may occur in either of two ways. The first, called the p–p chain, starts with two protons reacting, while the second, the CNO cycle, starts with a proton reacting with a carbon nucleus. The net effect of both is to convert four protons into a helium nucleus and an array of other particles and radiation. The next evolutionary phase converts three helium nuclei at the center into carbon by the triple alpha process, and other more complex reactions follow. The evolution of the composition causes the internal structure of the star to change, and the star is said to evolve (see STELLAR EVOLUTION). This evolution is said to occur on a nuclear timescale, which is dependent on the rate of change of the composition in the nuclear burning regions. During transitions from one set of nuclear processing reactions to another, the star may have to undergo some structural readjustment via slow gravitational contraction or expansion, on a timescale called the Kelvin–Helmholtz time, in order to create the necessary conditions for the next nuclear processing sequence to begin.

The energy generated by nuclear reactions will clearly depend on the number of the available reactants, the energy produced by the nuclear reaction and the cross section which indicates the likelihood that the reaction between the two nuclei takes place. In stellar interiors the relative motion between the interacting nuclei is provided by thermal motion, while experimentally this is provided by accelerating one of the nuclei to high velocity where it hits a stationary target of the other nucleus. The relative motion is much larger in our experiments than in stars, and the cross sections determined must be extrapolated to lower energy for stellar conditions. It can be shown that the thermal energies are not sufficient to overcome the Coulomb repulsion between two positively charged reactants, but one nucleus can 'tunnel' beneath the potential because of the small quantum mechanical probability of the nucleus penetrating the Coulomb barrier of the other. The nuclear reaction cross section may also be greatly enhanced by the presence of resonances produced by the nature of the wavefunction inside and outside the Coulomb barrier. For more information, see NUCLEOSYNTHESIS and NUCLEAR REACTION RATES.

The nuclear energy generation rates are very strong functions of the temperature. If we express the energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

generation rate as a power law in the temperature over a limited range of temperature, we find that the power for the p–p chain is 4–6, depending on the temperature, about 16 for the CNO cycle and approximately 30 for the triple alpha process.

The other component of the energy equation is the transport. Radiative transport in the deep interior of a star is relatively simple, primarily because the distance any photon can travel, its mean free path, is very short compared with the distance for a significant temperature gradient to occur. Thus all the matter and all the photons in a volume whose size is determined by the photon mean free path have essentially the same temperature. This is one description of a black body, for which the radiation field is homogeneous and isotropic and the intensity is given by the Planck function, B_ν. We start with the first-order moment of the radiative transfer equation by multiplying the transfer equation by the cosine of the angular variable and integrating over all angles to obtain (see RADIATIVE TRANSFER),

$$H = \int H_\nu \, \mathrm{d}\nu = -\int \frac{1}{\kappa_\nu} \frac{\mathrm{d}K_\nu}{\mathrm{d}z} \, \mathrm{d}\nu \equiv -\frac{1}{\kappa_R} \int \frac{\mathrm{d}K_\nu}{\mathrm{d}z} \, \mathrm{d}\nu$$

where H is proportional to the energy flux and K to the radiation pressure. The isotropic nature of the Planck function allows the intensity to be factored out of the angular integrals, so that $K_\nu \rightarrow \frac{1}{3} B_\nu$. Since the Planck function is only a function of temperature, we may write $\mathrm{d}K_\nu/\mathrm{d}z = \frac{1}{3}\,\mathrm{d}B_\nu/\mathrm{d}z = \frac{1}{3}\,(\mathrm{d}B_\nu/\mathrm{d}T)\,\mathrm{d}T/\mathrm{d}z$. Substitution defines a calculable quantity, the Rosseland mean opacity (κ_R)

$$\frac{1}{\kappa_R} = \int \frac{1}{\kappa_\nu} \frac{\mathrm{d}B_\nu}{\mathrm{d}T} \, \mathrm{d}\nu \Bigg/ \int \frac{\mathrm{d}B_\nu}{\mathrm{d}T} \, \mathrm{d}\nu$$

and an expression for the total flux:

$$H = -\frac{1}{3} \frac{1}{\kappa_R} \frac{\mathrm{d}B}{\mathrm{d}T} \frac{\mathrm{d}T}{\mathrm{d}z}.$$

From this last expression, called the diffusion approximation, one can see the relationship between radiative transport and a normal thermal conductivity. A similar expression to the above can be derived for thermal conduction, which is important only when degenerate conditions exist. The total flux would be the sum of the radiative and conductive fluxes, so that the total opacity is the harmonic mean of the radiative and conductive opacities. Note that the Rosseland mean opacity conserves the total flux and gives the greatest weight to those frequencies where the opacity is lowest. The diffusion approximation does not hold in the optically thin upper region of the stellar envelope, and it is strictly speaking necessary to use a different formulation to obtain the temperature stratification (although this is not always done). Generally, the logic is that the stellar structure and evolution can be determined quite well even when the surface is treated somewhat cavalierly.

Convection poses a significant challenge to our knowledge of stellar structure. However, convection is very important because it is expected that the composition is uniformly mixed in a convection zone, so that convection is an important mechanism for transporting material that has been processed by nuclear reactions deep inside a star to other locations. Indeed, current theoretical and observational results indicate that there is some mixing of material beyond the formal boundary of a convective region. This is important from an evolutionary point of view, because it means that material is mixed into the convective core from just outside it and thus is processed by nuclear reactions. This prolongs the lifetimes during these phases of evolution. In the core, the expectation is that the convective efficiency is so great that the relationship between the density, temperature and pressure is adiabatic.

In the layers near the stellar surface, the convective efficiency will not be very great because the material densities are so low, and a relatively simple mixing length formalism has traditionally been used to determine the convective structure. This model has one major free parameter, the mixing length, which governs how far a convective element will travel before depositing its energy in the surrounding medium. As might be imagined, the farther the element travels, the more efficient the convective transport. The major effect of uncertainties in the convective envelope structure is uncertainty in interpretation of stellar properties such as the stellar radius and the relationship between the observed color and the effective temperature.

This simple model continues to be used because no alternative model or method with more tightly constrained free parameters has been able to demonstrate that it yields answers that agree more closely with observations. These alternative models may be crudely classified into two categories—analytical one-dimensional models which are meant to be incorporated into the traditional one-dimensional methods of solution (see below), and models which are based on multidimensional hydrodynamical simulations of astrophysical convection. These are not mutually exclusive, as there are analytical models which have been based on the results of multidimensional hydrodynamic simulations.

Multidimensional simulations of turbulent convection are not trivial. In a simple-minded picture of turbulence, convective motion is produced on the largest scales (here presumably the size of the convection zone) and cascades down to progressively smaller scales until the scale is sufficiently small that the viscous forces convert the kinetic energy of the convective motion into heat. In stars, the size of the convection zone can be very large, something on the order of a typical stellar radius (10^{11} cm for an order of magnitude), while the length scale on which the viscosity can convert the kinetic energy into heat may be seven orders of magnitude smaller. Trying to resolve all of these length scales in hydrodynamic calculations is beyond the capabilities of current computers and likely to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

remain that way for some time. The general approach for hydrodynamic simulations is to divide the calculation into two components—that which can be spatially resolved and that which cannot. The first is directly included in the solution of the appropriate equations, while the terms involving the second must be approximated by models which contain only the spatially resolved component. This has been done in many ways, and much work remains to translate these results into something that can be used to determine the structure in general stellar situations. For more discussion, see STELLAR INTERIORS: THERMODYNAMICS.

Supposing it possible to calculate the interior structure for a model of stipulated mass and composition, what techniques are available to determine whether the model bears any resemblance to a given observation? The answer is qualified by what is known about the star. From the spectral energy distribution and absorption line strengths of the light emitted from the stellar surface, we can obtain an estimate of its effective temperature, surface gravity and surface composition. Information about the interior may be gained in some cases from elements which have been processed in the interior and mixed to the surface. For example, the amount and the isotopes of lithium and beryllium, which capture protons at relatively low temperatures, indicate whether the surface convection zone has penetrated to such temperatures at some time in the past. If we have some method of determining the star's distance, we can obtain its luminosity, and we can determine its mass if it is in an eclipsing binary system.

Clusters (see STAR CLUSTERS) provide an opportunity to obtain more information. Clusters are groups of stars which formed at nearly the same time out of material with nearly the same composition. Thus, the stars in a cluster provide a sequence of stars whose primary difference is the stellar mass. Combining a number of stellar model calculations of stars with the same initial composition and age but different masses allows comparisons with the HR diagrams (plot of luminosity versus effective temperature—see HERTZSPRUNG–RUSSELL DIAGRAM).

The Sun presents a special case because of its proximity and thus the level of detail we can detect. Detected neutrinos from the Sun should yield information about the structure of the solar center, but models have not yet been able to produce the observed neutrino flux while matching all the other available solar data (see SOLAR INTERIOR: NEUTRINOS).

We also observe a very large number (about 10^5) of low-amplitude pulsation modes believed to be powered by the solar turbulent convective envelope. These periods convey information about the sound travel time from the surface to a depth which depends on the mode, with longer-period pulsation modes 'sounding' to greater depths. Thus the period difference between two modes which are very close in period, and thus penetrate to nearly the same depth, contains information about the sound speed in the region between the effective depths of the two modes. This is similar to seismic waves sampling different depths inside the Earth, leading to this study of the Sun being termed HELIOSEISMOLOGY. With the many modes available, the sound speed, and hence the internal structure, has been observationally determined for the outer 80% of the solar radius. From this structure, we can determine the depth of the surface convection zone, a constraint that any convective model must match. Furthermore, by examining modes which are effectively the same radially but differ horizontally, the solar rotation rate as a function of depth and solar latitude can be determined to the same depth.

The major surprises resulting from this comparison were that the solar convection zone extended to greater depth (about the outer 30% of the radius) than previously thought and that the solar rotation rate in the convective envelope is approximately proportional to the radius with different coefficients at different latitudes. The depth of the convection zone is matched by increasing the value of the mixing length to something like 1.7 pressure scale heights (the precise value depends on the mixing length formulation used) from the previously traditional range of 1.0–1.5 pressure scale heights. Current solar models provide agreement with the sound speed data to within a few tenths of a per cent, while models before helioseismology are now known to be about 1.5% in error.

Can such an approach be extended to other stars (asteroseismology)? The obstacles are formidable because of the distance to other stars. The most promising stars would appear to be δ Scuti variable stars, white dwarfs, and perhaps high-mass stars near the main sequence. However, no other star currently has more than 200 modes known, and our less detailed knowledge about any given star makes mode identification much more difficult than for the Sun.

We now turn to a more detailed discussion of how the interior structure of a model is calculated and some simple models.

The equations of stellar structure

We begin with the conservation of mass and the conditions of hydrostatic equilibrium (excluding all forces but self-gravity and the internal pressure) and thermal equilibrium for spherically symmetric models:

$$\frac{\mathrm{d}M_r}{\mathrm{d}r} = 4\pi r^2 \rho \qquad \frac{\mathrm{d}P}{\mathrm{d}r} = -\rho \frac{GM_r}{r^2} \qquad \frac{\mathrm{d}L_r}{\mathrm{d}r} = 4\pi r^2 \rho \varepsilon$$

where M_r is the mass interior to radius r, ρ is the density, P is the pressure, L_r is the luminosity flowing through radius r and ε is the nuclear energy generation rate. Both the pressure and the nuclear energy generation rate are expressed as functions of the density and temperature for a given composition. Assuming the composition is stipulated *a priori*, we see that we have three equations with the four unknowns (ρ, T, M_r and L_r). The missing equation is one relating the thermal structure to the energy transport. Writing $\nabla = \mathrm{d}\ln T/\mathrm{d}\ln P$, we have three possibilities for this equation:

$$\nabla = \nabla_{rad} = \frac{3\kappa P L_r}{16\pi acGT^4 M_r}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$\text{radiative, conductive,} \nabla_{rad} < \nabla_{ad}$$

$$\nabla = \nabla_{ad} \qquad \text{efficient convection,} \nabla_{ad} < \nabla_{rad}$$

or

$$\nabla = \nabla_{act} \qquad \text{inefficient convection,} \nabla_{ad} < \nabla_{act} < \nabla_{rad}.$$

The first equation is used in the convectively stable ($\nabla_{ad} > \nabla_{rad}$) region where the energy is carried by radiation or conduction or both, the second in convectively unstable ($\nabla_{ad} < \nabla_{rad}$) deep layers where the convection is so efficient that the structure may be taken to be adiabatic and the third in convectively unstable regions where the convection is inefficient and the gradient is supplied by some convective transport algorithm such as the local mixing length theory.

These equations then determine the stellar structure. However, they are usually not solved in this form, because the radius is generally not the most useful independent variable. Rather, the independent variable is taken as the interior mass, M_r. One obtains the equations in terms of this variable by solving for dr in the mass conservation equation and substituting this in the other equations to obtain derivatives with respect to M_r. Since there are four first-order differential equations, we require four boundary conditions. At the stellar center, the values of M_r, r and L_r are zero. The conditions at the stellar surface have a little more latitude. The density and pressure are so low compared with those at the stellar center that $P = \rho = 0$ (called the zero boundary conditions) are often a good approximation. More sophisticated techniques use stellar atmosphere models to fit all variables (i.e. ρ, T, r and L_r) at a given location (i.e. M_r) below the stellar surface.

Most determinations of stellar interior conditions are now made with computational numerical models, but some historical and somewhat instructive analytical models were developed before extensive numerical simulations became possible.

Simplified models

Before numerical computers it was necessary to simplify the stellar structure equations to obtain either analytical solutions or solutions which required only a limited amount of calculation. The simplest of these approaches is for a polytrope, which is defined by the relation $P = C\rho^{(n+1)/n}$. Thus, one may write $P = P_c\theta^{n+1}$ and $\rho = \rho_c\theta^n$. To obtain an equation utilizing this definition, divide the hydrostatic equilibrium equation by $r^2/\rho G$, take the derivative with respect to r and use the mass conservation equation to obtain

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho G}\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -\frac{\mathrm{d}M_r}{\mathrm{d}r} = -4\pi r^2\rho.$$

With the above definition for θ, this becomes

$$\frac{(n+1)P_c}{4\pi G\rho_c^2}\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\theta}{\mathrm{d}r}\right) = -\theta^n.$$

This equation may be simplified by introducing a length scale, r_s, by $r = r_s x$. Defining

$$r_s^2 = \frac{(n+1)P_c}{4\pi G\rho_c^2}$$

leads to the Lane–Emden equation

$$\frac{1}{x^2}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2\frac{\mathrm{d}\theta}{\mathrm{d}x}\right) = -\theta^n.$$

From the definition of θ, $\theta = 1$ at $x = 0$. Also, spherical symmetry requires that $\mathrm{d}\theta/\mathrm{d}x = 0$ at $x = 0$. The surface boundary conditions ($P = \rho = 0$) are met at x_1 where $\theta(x_1) = 0$. Clearly, both P and ρ must be positive, so that x_1 corresponds to the first zero of θ. The attractiveness of this equation is apparent in the need to solve the Lane–Emden equation only once for a given n, a tedious but performable task. There are analytical solutions for $n = 0$, 1 and 5, and there are published tables of $\theta(x)$ for $1 \leq n \leq 5$ in steps of 0.5. Once P_c and ρ_c are selected and $\theta(x)$ is computed, one may solve for P, ρ, M and R.

The advantage of polytropes from a computational point of view is undeniable, especially when calculations had to be done by hand or with mechanical calculators. However, do polytropes ever resemble real stars? More specifically, are there situations in which a combination of the equation of state and the structure required by energy transport produces a polytropic relation?

The most noteworthy case for which polytropes resemble real stars is for white dwarfs. If the electrons are fully degenerate throughout the entire white dwarf, the degenerate equation of state requires a polytropic structure with $n = 1.5$ if the electrons are non-relativistic and $n = 3$ if they are extremely relativisitic. With the temperature absent in the equation of state, hydrostatic and thermal equilibria decouple. The hydrostatic structure is that of a polytrope, while the thermal structure is determined by the thermal structure when the star became a white dwarf and its cooling as the thermal energy is transported to the surface and radiated. Chandrasekhar was thus able to demonstrate that degenerate electrons allowed the possibility of stars with a solar mass of material inside a volume about that of the Earth and, furthermore, that there was an upper limit to the mass allowed for a white dwarf.

While less dramatic than the theoretical possibility of white dwarfs, both fully convective stars and parts of other stars resemble polytropes. The chief virtue of the latter is a variety of models in which the solution for the stellar structure is somewhat less laborious than for the full stellar structure equations.

With the current environment in which it seems that there is a computer on every desk, these models are of limited interest, and the methods by which stellar structure is currently calculated are now examined.

Numerical solution of the stellar structure equations

Once sufficient computing power became available, it became possible to develop numerical algorithms to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solve the equations of stellar structure using with either numerical tables or analytical fits to describe the equation of state, opacity and nuclear reaction rates. There have been two numerical algorithms which have been used successfully, a fitting technique and an implicit finite difference method. Current calculations are dominated by finite difference techniques.

The simple reason why one cannot use ordinary integration techniques, such as the Runge–Kutta method, is that the boundary conditions are divided between the surface and the center of the stellar model. The fitting technique tries to remove this problem by making plausible assumptions for the conditions one does not know, say the central density, central temperature (or equivalently central pressure), surface luminosity and surface radius, and integrating from both the model center and the model surface and meeting at some interior location (the fitting point, call it M_{r0}). If all the selections for the four unknowns were correct, the temperature, density, luminosity and radius from the two integrations would agree at M_{r0}. Thus, we have four conditions which must be satisfied and four unknowns, so the problem in principle can be solved. The problem can be simplified slightly by choosing M_{r0} to be exterior to the energy generation region, performing the core integration first and using the luminosity at M_{r0} as the guess for the surface luminosity.

In practice there are some difficulties. One is that the guesses for the central unknowns must be fairly good or the integration may not even reach M_{r0}. Another is that we have not considered the possible composition variations, whose values at any given location may depend on whether it is in a thoroughly mixed and therefore uniform composition convective region and how much mass the convective region may contain (and which for convective cores may not be known until the convective core boundary is reached in the outward integration from the center). Thus, the fitting technique is now used primarily for determining a guess for a uniform composition zero-age main sequence model.

The finite difference approach divides the stellar model into a finite number of discrete zones and replaces the first-order differentials by differences between two adjacent zones. The model is then determined by evaluating these finite difference equations in all zones along with the boundary conditions for all the variables in all the zones.

The Henyey method is the most common way these solutions are generated. For simplicity, suppose that there is only one second-order differential equation to be solved with the form

$$f(x, y) = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + g(y, x)\frac{\mathrm{d}y}{\mathrm{d}x} + h(y, x) = 0.$$

The second-order differential means that the finite difference representation involves values of y in three adjacent zones. Assume that there is a guess for the solution $y(x)$ of this equation, then perturb the guess and keep only the linear term in the perturbations (δy). For each zone, i, there will be an equation of the form

$$A_i\,\delta y_{i-1} + B_i\,\delta y_i + C_i\,\delta y_{i+1} + D_i = 0.$$

D_i is the value of the function f evaluated in zone i with the current guess and

$$A_i = \frac{\partial f}{\partial y_{i-1}} \qquad B_i = \frac{\partial f}{\partial y_i} \qquad C_i = \frac{\partial f}{\partial y_{i+1}}.$$

Note that as D approaches zero for all i, all the δy_i go to zero as well. Suppose that there is one boundary condition at the center of the stellar model ($i = 1$) and one at the surface ($i = n$). The central boundary condition may be used to eliminate δy_0 from the equation for $i = 1$, and the surface condition to eliminate δy_n from the equation for $i = n$. Thus, the difference equations will have the form

$$B_1\,\delta y_1 + C_1\,\delta y_2 + D_1 = 0$$

$$A_i\,\delta y_{i-1} + B_i\,\delta y_i + C_i\,\delta y_{i+1} + D_i = 0 \qquad i = 2, \ldots, n-1$$

$$A_n\,\delta y_{n-1} + B_n\,\delta y_n + D_n = 0.$$

For $i = 1$, it is possible to write

$$\delta y_1 = \alpha_1\,\delta y_2 + \beta_1$$

where

$$\alpha_1 = -B^{-1}C_1 \qquad \beta_1 = -B^{-1}D_1.$$

Substituting these expressions into the equation for $i = 2$ leads to

$$A_2(\alpha_1\,\delta y_2 + \beta_1) + B_2\,\delta y_2 + C_2\,\delta y_3 + D_2 = 0.$$

Combining terms and using the definition of α and β for $i = 2$, one obtains

$$\alpha_2 = -(A_2\alpha_1 + B_2)^{-1}C_2 \qquad \beta_2 = (A_2\alpha_1 + B_2)^{-1}(A_2\beta_1 + D_2).$$

Continuing this exercise for all zones generates a set of recursion relations for the α and β. One virtue of these relations is that generally the α_i are of order unity, which tends to eliminate mathematical errors when the various terms have much different orders of magnitude.

The result for $i = n$ is

$$(A_n\alpha_{n-1} + B_n)\,\delta y_n + A_n\beta_{n-1} + D_n = 0$$

so that

$$\delta y_n = -(A_n\alpha_{n-1} + B_n)^{-1}(A_n\beta_{n-1} + D_n).$$

From this all the other δy_i can be found from the relations defining the α_i and β_i.

Finding the δy_i obtains the solution of the best linear correction to solving the equation. However, the corrections may be so large that dropping the higher-order corrections is unwarranted. The whole process is therefore repeated until all of the corrections are below some desired

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

tolerance. Because the lowest omitted term is of order δy^2, convergence is quadratic.

Furthermore, there is nothing limiting about having only one equation with one independent variable. If there are several (N) differential equations ($f_1, \ldots, f_N$) with the dependent variables ($q_1, \ldots, q_N$), the same methodology can be used. The A, B, C and α now become N by N matrices, and the D, β and δy become column vectors of length N.

Unresolved issues in stellar structure

With these methods models of stellar structure can be generated routinely. Yet even after many models our knowledge of stellar interiors is incomplete. There are significant disagreements over elements that are produced in the interiors of some stars and mixed to the surface which models do not predict. Other issues are the times some stars spend is some portions of the HR diagram. From a purely theoretical point of view, knowledge of the effects of rotation and magnetic fields remains uncertain. As previously mentioned, the treatment of inefficient convection in stars remains a major impediment to our ability to model the interiors of stars.

Bibliography

There are a great many books and review articles on stellar interiors and evolution. One that is somewhat dated, but excellent on the general concepts, is by Schwarzschild (*Structure and Evolution of the Stars*, Princeton University Press, 1958). The book that is the most detailed on the physics of stellar structure and evolution, although virtually unobtainable outside of university libraries and older professional astrophysicists, is by Cox and Giuli (*Principles of Stellar Structure*, Gordon and Breach, 1968).

Robert G Deupree

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Interiors: Thermodynamics

Material properties and stellar models

Thermodynamic quantities of stellar matter are, together with opacity and nuclear reaction rates, the fundamental properties of stellar matter that enter stellar models. These quantities are responsible for the resulting overall parameters of stars and their differentiation according to chemical composition. These *material properties* determine the coefficients of the equations of stellar structure (see STELLAR EVOLUTION). Otherwise, the type of stellar matter would not enter these equations, which express the result of (i) a balance of forces, (ii) a balance between the energy loss at the stellar surface and energy generation in the core, and (iii) stationary energy transport between the core and the surface. Thermodynamics is most directly relevant for the balance of forces, which settles in the *hydrostatic equilibrium*, which is a relation between the local pressure gradient and gravitational acceleration. The force of gravity is determined by the density distribution in the star: thus stellar modeling requires a relation between density and pressure through the specific properties of the matter. More precisely, the relevant properties of stellar matter are expressed by the *equation of state*, connecting pressure p, density ρ, temperature T, and composition. A simple example is the perfect gas law for a fully ionized gas, which may be written as $p = \mathcal{R}\rho T/\mu$; here $\mathcal{R}$ is the gas constant, and μ is the mean molecular weight of the various species contained in the gas. Clearly μ depends on the composition, which is often characterized by the fractional mass abundances X, Y and Z of hydrogen, helium and heavier elements.

However, thermodynamics is also important for other parts of stellar models. In particular, the bulk of most convection zones is essentially adiabatically stratified, that is, the pressure–density–temperature relation is an adiabat. The adiabat is the integrated form of the local adiabatic gradients, which are purely thermodynamic quantities. One of the adiabatic gradients is related to adiabatic sound speed: $c^2 = (\partial p/\partial \rho)_s$ (s being specific entropy). This thermodynamic quantity has another important stellar application, because adiabatic sound speed is the key quantity that determines the frequency of acoustic oscillation modes (for a solar illustration see HELIOSEISMOLOGY: THEORY); however, thermodynamics is also crucial when assessing the importance of non-adiabatic effects (see STELLAR PULSATION). This is understood by realizing that the temperature of the stellar interior is determined by the local energy balance of the outward flux of energy, which begins in the hot stellar core that is driven by nuclear reactions, and ends at the surface when the energy leaves the star in the form of radiation (see STELLAR INTERIORS).

Inside a normal star, the two prevailing modes of energy transport are radiation and convection. Radiative transport is easy to formulate when opacity is high and the limiting diffusion approximation works well. The reciprocal of the opacity is the radiative conductivity which determines the energy throughput in this situation. The opacity results from absorption coefficients, which are obtained from atomic physics, although thermodynamics also enters, because it has to provide ionization equilibrium concentrations and level populations of all the absorbing species.

The problem with radiative energy transport is not so much that the diffusion approximation would be bad (in general, opacities in stellar interiors are high), but rather that a literal application of the equations of radiative transport could result in a temperature gradient so steep that a star becomes convectively unstable: the ensuing convective motion, where hot elements of gas rise, dump their excess heat and then, will in general dominate the energy transport. Following Karl Schwarzschild (see SOLAR INTERIOR), the radiative transport equation is replaced by the one describing convective stratification if the local temperature gradient exceeds the adiabatic temperature gradient. To a good approximation, in the majority of convection zones, the convective stratification is adiabatic. In this case, it is the direct result of a thermodynamic quantity.

However, away from the aforementioned limiting cases, knowledge of thermodynamics is still required. The first case is when convection is not efficient. Then, a significant part of the energy is still being carried by radiation. In this case, material properties (mainly specific heats) are needed to model the real (not just the adiabatic) temperature gradient. The second case is when opacity is not very high (in the layers close to the surface of a star). Then, the diffusion approximation breaks down. Since the mean-free path of radiation becomes large, heat can be instantaneously transported from one part to another. Such non-local effects cannot be treated anymore by differential equations, but by the more complex set of equations of radiation hydrodynamics, for which material properties are again crucial.

Ideal and nonideal plasmas

A useful practical introduction to the equation of state is the book by Eliezer *et al* (1986). On a higher theoretical and systematic level is the book about the physics of Coulomb systems by Kraeft *et al* (1986). The most important classification when dealing with plasmas of stellar interiors is the one in ideal and nonideal plasmas. The simplest model is given by a mixture of nuclei and electrons, assumed fully ionized and obeying the classical perfect gas law. However, this is not the most general model of an *ideal-gas* equation of state: it may include deviations from the perfect gas law, namely ionization or dissociation reactions, radiation and degeneracy of electrons, as long as the underlying microphysics of these additional effects is still ideal, that is, does not contain interactions. The 'particles', however, can be classical or quantum, material or photonic. In such an ideal framework, bound systems (molecules, atoms, ions) are allowed to have internal degrees of freedom (excited states,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spin). All such ideal effects can be calculated as exactly as desired.

Nonideal effects in plasmas can be assessed by the so-called coupling parameter Γ. In a plasma of temperature T and density such that particles of charge e have an average distance $\langle r \rangle$ from each other, one can define Γ as the ratio of average potential binding energy over mean kinetic energy $k_B T$ (k_B being the Boltzmann constant)

$$\Gamma = (e^2/\langle r \rangle)/k_B T. \tag{1}$$

For simplicity, this expression is restricted to the case of hydrogen; generalizations to other elements are straightforward. Plasmas with $\Gamma \gg 1$ are *strongly* coupled, those with $\Gamma \ll 1$ *weakly* coupled. A famous example of a strongly coupled plasma is the electron gas in the interior of white dwarfs, where the coupling can become strong enough to force crystallization. Another example is given by the electrons in the conduction band of a metal at room temperature. Weakly coupled plasmas are, for instance, the interiors of stars with masses ranging from the slightly sub-solar ones to the largest.

Basic concepts and dimensional parameters

To describe a degree of 'nonideality' or 'degeneracy' of the plasma, it is convenient to introduce a few dimensionless parameters to express the relative strength of the particular effects, such as the Coulomb interaction or quantum degeneracy. These parameters are defined as combinations of various 'characteristic lengths' resulting from simple physical estimates.

The strength of the Coulomb interaction is given by the density-independent *Landau length* l_L

$$l_L = \frac{e_i e_j}{k_B T}. \tag{2}$$

Although the Coulomb potential has infinite range, the Landau length represents an effective radius of the interaction between particles i and j, i.e. for distances $\gg l_L$ the potential energy is negligible compared to the thermal energy $k_B T$.

The Landau length clearly reflects a degree of nonideality of the plasma and the case of an 'ideal' plasma corresponds to $l_L = 0$.

Another purely classical characteristic length is the *mean distance* between two particles of the same species k ($k = \{e, i\}$)

$$d_k = \left(\frac{3}{4\pi n_k}\right)^{1/3}. \tag{3}$$

Here, n_k is the particle density N_k/V of species k. Note that d_k is the generalization of the quantity $\langle r \rangle$ in equation (1). It does not depend on temperature.

The most evident quantum mechanical characteristic length is the *thermal de Broglie wavelength*

$$\lambda_k = \frac{h}{\sqrt{2\pi m_k k_B T}}. \tag{4}$$

The limits of applicability of various theoretical models of the equation of state are usually described by different dimensionless parameters. The *degeneracy* of species k

$$n_k \lambda_k^3$$

is very closely related to the well known *degeneracy parameter* η (sometimes also denoted ψ) by

$$F_{1/2}(\eta) = \frac{\sqrt{\pi}}{4} \frac{\lambda_e^3 N_e}{V}, \tag{5}$$

where F_i is the Fermi integral

$$F_i(y) = \int_0^\infty x^i [1 + \exp(y + x)]^{-1} \, \mathrm{d}x. \tag{6}$$

In the case of a stellar plasma with heavy nuclei and light electrons, the heavy particles become degenerate only at much higher densities than electrons.

Screening and the Debye–Hückel approximation

All characteristic lengths defined so far demonstrate either purely density- or purely temperature-dependent behavior. No collective behavior of particles has been considered. Here screening is discussed, which for normal stars is the most important deviation from ideality.

The seminal study of collective behavior in a plasma was the Debye–Hückel theory of electrolytical solutions. Despite its phenomenological approach this theory proved to be immensely successful not just for electrolytes, but also for plasma calculations.

Basically, one mixes two ideas. First, assuming nondegenerate electrons, and considering one particular ion being fixed, the mean electron distribution around that ion is given by the *Boltzmann factor*

$$n_e(r) = n_e \exp\left[\frac{+e\phi(r)}{k_B T}\right]. \tag{7}$$

The mean ion density around the same ion is similarly given by

$$n_i(r) = n_i \exp\left[\frac{-e\phi(r)}{k_B T}\right]. \tag{8}$$

Second, one assumes applicability of Poisson's equation for *mean charge distributions* (instead of point charges)

$$\Delta\phi(r) = -4\pi e[n_i(r) - n_e(r)]. \tag{9}$$

The resulting system of equations is complicated and nonlinear; however, if the system is only slightly nonideal, that is, if $e\phi(r)/k_B T < 1$, the linearized system has the well known solution of the static-screened Coulomb potential (SSCP)

$$\phi(r) = \frac{e}{r} \mathrm{e}^{-\kappa r}, \tag{10}$$

where κ is the reciprocal of the *Debye length* r_D

$$r_D = \sqrt{\frac{k_B T}{4\pi(n_e e^2 + n_i e^2)}}. \tag{11}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This is the electrostatic and weak-screening approximation to a more general dynamic situation, which plays not only an important role in stellar thermodynamics, but equally in the determination of nuclear reaction rates (see NUCLEAR REACTION RATES).

The chemical picture

With the exception of neutron stars, in stellar interiors the thermal de Broglie wavelength of nuclei, atoms, ions and molecules is always tiny and they can all be treated classically. Only electrons have to be treated according to quantum mechanics. Here, there is a bifurcation into two distinct classes of approach, the 'chemical picture' and the 'physical picture' (see the book by Kraeft *et al* 1986). While in the more conventional chemical picture bound configurations (atoms, ions and molecules) are introduced and treated as new and independent species, only *fundamental* particles (electrons and nuclei) appear in the physical picture. In the chemical picture, reactions between the various species occur, and thus the thermodynamical equilibrium must be sought among the stoichiometrically allowed set of concentration variables by means of a maximum entropy (or minimum free-energy) principle. In contrast, the physical picture has the aesthetic advantage that there is no need for a minimax principle; the question of bound states is dealt with implicitly through the Hamiltonian describing the interaction between the fundamental particles. See the section on the physical picture below.

The free-energy-minimization method

The development of computational methods was strongly stimulated by the requirements for stellar evolution, where data for a wide range of parameters are needed. The chemical picture is very well suited for powerful, approximative models such as the free-energy-minimization method, which became computationally feasible since about 1960. Given a mathematical model for the Helmholtz free energy $F(T, V, \{N_i\})$, where $\{N_i\}$ is a set of particle numbers for all species i present in a plasma, one minimizes F subject to the stoichiometric relations governing possible reactions among the particle species in the plasma. The underlying principle is that nature adjusts the reaction equilibrium such that entropy is maximum (for given energy) or the free energy is minimum (for given temperature).

One starts from the canonical partition function. Consider a physical system (with Hamiltonian H) confined in a box of volume V in contact with a heat reservoir at temperature T. Then the canonical partition function is a trace (denoted by Tr)

$$\mathcal{Z} = \mathrm{Tr}(\mathrm{e}^{-H/k_\mathrm{B}T}). \tag{12}$$

From that, the free energy is obtained by the formula

$$F(T, V, \{N\}) = -k_\mathrm{B}T \ln(\mathcal{Z}). \tag{13}$$

Here, the volume V is implicitly contained in the Hamiltonian operator. The free energy is calculated for all $\{N\}$ which are allowed by the *stoichiometric relations* and a given initial composition. Equilibrium concentrations are those $\{N\}$ that minimize $F(T, V, \{N\})$ for given T, V under the stoichiometric constraints. This gives ionization and dissociation equilibria. Numerically, this will be a task of finding a minimum of a nonlinear function $F(T, V, \{N\})$ in the $\{N\}$-space under the stoichiometric constraints.

The major assumption in any *practical* realization of the free-energy minimization method is that the total partition function $\mathcal{Z} = \mathcal{Z}_\mathrm{total}$ factorizes, that is,

$$\mathcal{Z}_\mathrm{total} = \mathcal{Z}_\mathrm{e}\mathcal{Z}_\mathrm{trans}\mathcal{Z}_\mathrm{conf}\mathcal{Z}_\mathrm{int}, \tag{14}$$

where Z_e here stands for the electronic contribution, $\mathcal{Z}_\mathrm{trans}$ is the result of the integration over momentum space for the heavy particles, $\mathcal{Z}_\mathrm{conf}$ is the configurational integral over configuration space coordinates, and finally $\mathcal{Z}_\mathrm{int}$ is

$$\mathcal{Z}_\mathrm{int} = \Pi\mathcal{Z}_\mathrm{int}^{(i)} \qquad \mathcal{Z}_\mathrm{int}^{(i)} = \sum_{j=1}^{\infty} g_{ij} \exp(-E_{ij}/k_\mathrm{B}T), \tag{15}$$

where E_{ij} and g_{ij} denote energy and degeneracy of the state j of species i, respectively.

When going from the partition function to the free energy, the consequence of the assumed factorizability of the partition function is additivity, or modularity, of the free energy. This modularity accounts for the great appeal of the chemical approach from both a modeling and a computational point of view. Modularity basically tells us that all the interactions can be split into separate parts with clear physical meaning. And if, for example, one part needs to be modified to implement a higher order correction, it can be done without having to worry about consistency with the other parts. Terms can be introduced simply by adding or changing subroutines in a computer program, without a major overhaul of the previous work. That is why sometimes for the sake of modularity even relatively crude approximations are maintained, because a rigorous treatment, in the framework of the exact nonfactorizing partition function, would be entirely prohibitive.

Recovering the Saha equation

By retaining only ideal terms, the free-energy-minimization method falls back to the time-honored *Saha equation*. To illustrate, consider the simplest case, hydrogen ionization ($\mathrm{H} \leftrightarrow \mathrm{H}^+ + \mathrm{e}^-$). A neutral mixture of H-atoms, protons and electrons with particle numbers of N_H, N_p and N_e, respectively, has only one degree of freedom, because of the two relations (N_H^0 being the *total* number of protons, bound and free)

$$N_\mathrm{H} + N_\mathrm{p} = N_\mathrm{H}^0 = \mathrm{const},$$
$$N_\mathrm{e} = N_\mathrm{p}. \tag{16}$$

The degree of freedom is the ionization degree. Its equilibrium value is the one for which the free energy is minimum. The minimum condition is then

$$\frac{N_\mathrm{H}}{N_\mathrm{p}N_\mathrm{e}} = \frac{\lambda_e^3}{V}\mathcal{Z}_\mathrm{int}^\mathrm{H} = \sum_{j=0}^{\infty} g_j \exp\left(-\frac{E_j}{k_\mathrm{B}T}\right), \tag{17}$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(with E_j and g_j denoting the *negative* energy states of hydrogen and their orbital statistical weight, respectively). This is the well known *Saha equation* in astrophysics (known as the law of mass-action in physical chemistry). The infinity of the internal (hydrogenic) partition function is notorious; it has plagued statistical mechanics of plasmas for the last 70 years. The usual argument goes that excited states of hydrogen atoms become infinitely large with increasing quantum number j. Because of this, the infinite partition function would only occur for a single atom in the whole infinite universe. Remedies of the infinity have been involving more-or-less sophisticated truncations of the summation at some level; a judicious choice allows us to model real, interacting plasmas.

Let us note in passing that even the most radical truncation to the ground state of hydrogen (whose orbital statistical weight is one)

$$\frac{N_{\rm H}}{N_{\rm p}N_{\rm e}} = \frac{\lambda_e^3}{V}\exp\left(-\frac{E_0}{k_{\rm B}T}\right) \tag{18}$$

is not devoid of problems because, as one sees easily, it leads to spurious recombinations at high densities and relatively low temperatures. Specifically, in the center of a typical star, the Sun, this equation would predict about 30% neutral hydrogen, which is clearly nonsensical for the prevailing density of 150 g cm^{-3}. The reason for this pathology is that the Saha equation does not contain information about the *size* of the hydrogen atom. Bringing in correction terms that do contain atomic radii amounts to modeling so-called 'pressure ionization'. Only nonideal interaction terms between the species can treat this important physical effect.

Nonideal effects in the chemical picture

Re-deriving the Saha equation with a new method gives no gain though. However, the power of the free-energy-minimization method is precisely that it allows us to include nonideal effects *consistently*. The reason is that any modification to the thermodynamics is only made at one single place (the free energy). Since the ionization degree and all thermodynamic quantities follow from the free energy by purely mathematical steps, all these quantities are consistent with each other. For instance, they automatically obey all Maxwell relations. The corresponding shifts in the ionization equilibria are correct.

Achieving such consistency would be difficult in an approach where one considers the Saha equation and the corresponding thermodynamic quantities separately, all to be modified individually. There would be no general systematic procedure. The free-energy-minimization method, however, achieves it nicely and simply. It is therefore the natural nonideal extension of it to incorporate interactions into simple equations of state which are based on the Saha equation.

As an example, to include screening at the level of the Debye–Hückel approximation, it suffices to add to the total free energy the 'module' for the Debye–Hückel free energy

$$F_{\rm conf} \equiv F^{\rm DH} = -\frac{k_{\rm B}TV}{12\pi r_{\rm D}^3}. \tag{19}$$

Weakly-coupled stellar plasmas

High-mass stars are distinguished by high temperatures, so that even in their center, the kinetic energy $k_{\rm B}T$ of particles largely dominates the electrostatic potential energies. The Sun is somewhat an exception, not so much because it would not qualify as a moderate-mass star, but rather because the helioseismic data put such constraints on the thermodynamic quantities that even very small nonideal deviations are needed in solar models and can therefore be studied in great detail (see SOLAR INTERIOR: EQUATION OF STATE AND OPACITY).

For weakly-coupled plasmas, perturbational approaches are effective. An example is the equation-of-state and opacity effort of the international Opacity Project (OP; see the book by Seaton (1995)). A necessary part of OP is the so-called Mihalas–Hummer–Däppen (MHD) equation of state, which deals with the *heuristic* concept of modified atoms and ions due to the plasma environment. More specifically, the MHD equation of state is based on an occupation probability formalism. The internal partition functions $Z_s^{\rm int}$ of species s adopted by MHD are sums that are made convergent by weights

$$Z_{\rm int}^s = \sum_i w_{is}g_{is}\exp\left(-\frac{E_{is}}{k_{\rm B}T}\right). \tag{20}$$

Here, the index is labels the state i of species s. E_{is} and g_{is} are their energies (in MHD assumed to be unshifted) and statistical weights, respectively, $k_{\rm B}$ is the Boltzmann constant, T is the temperature, and the weights w_{is} are the occupation probabilities that take into account charged and neutral surrounding particles. In physical terms, w_{is} gives the fraction of all particles of species s that can exist in state i with an electron bound to the atom or ion, and $1-w_{is}$ gives the fraction of those that are so heavily perturbed by nearby neighbors that their states are effectively destroyed.

Although the MHD equation of state was originally designed to provide the level populations for opacity calculations of stellar *envelopes*, the associated *thermodynamic quantities* of MHD can none the less be reliably used also for stellar cores. This is due to the fact that in the deeper interior the plasma becomes virtually fully ionized. This rather fortuitous circumstance is the reason why the MHD equation of state fares very well in models of the entire Sun and even in lower-mass models, down to 0.4 $M_\odot$ (for references, see Däppen and Guzik 2000).

A special effect appearing in higher-mass stars is the modification of thermodynamic quantities due to the fact that electrons can become significantly *relativistic*. Inclusion of relativistic electrons is not difficult, however, since it is an ideal-gas effect. It is amazing that the relativistic effect has been detected even in the Sun, despite its tiny amount, because of the superb quality of helioseismological techniques.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Strongly-coupled stellar plasmas

Thermodynamics is more complicated for the interior of low-mass stars, because the kinetic energy of the particles does not dominate the electrostatic potential energies. In the deep interior of these stars, the plasma can become strongly coupled. In addition, temperature alone cannot ionize matter fully, and pressure ionization occurs, which requires more elaborate modeling. The configurational free energy $F_{\rm conf}$ stems directly from the configuration integral $\mathcal{Z}_{\rm conf}$. Besides containing the Coulomb interactions that lead in the classical weakly-coupled approximation to the Debye–Hückel free energy $F_{\rm conf}$, it also contains excluded-volume effects of neutral particles. The well known van der Waals equation of state is based on such an excluded volume effect, but there are more elaborate formalisms of hard-sphere or softer types. Such excluded volume effects become very important in the relatively cold plasmas of low-mass stars or planets (see Saumon *et al* 1995). Note that the sophisticated nature of hard-sphere potentials does not change the basically heuristic nature of the free-energy minimization methods. Only the physical picture can overcome this fundamental limitation (see the following section).

The physical picture

Like the MHD equation of state, the alternative physical-picture activity expansion was also developed as part of an opacity effort, the one pursued by the OPAL group at Lawrence Livermore National Laboratory (Rogers *et al* 1996). The underlying equation of state is based on a sophisticated *systematic* method for including density effects in a plasma. It is clear from the preceding sections that the advantages of the free-energy-minimization method and the chemical picture lie in the possibility to model complicated plasmas, and to obtain numerically smooth and consistent thermodynamical quantities. Nevertheless, the heuristic method of the separation of the atomic-physics problem from that of statistical mechanics is not satisfactory, and attempts have been made to avoid the concept of a perturbed atom in a plasma altogether. This has suggested an alternative description, the physical picture. In such an approach one expects that no assumptions about energy-level shifts or the convergence of internal partition functions have to be made. On the contrary, properties of energy levels and the partition functions should come out from the formalism.

There is an impressive body of literature on the physical picture (Kraeft *et al* 1986; Rogers *et al* 1996; and references therein). However, the majority of work on the physical picture was not dedicated to the problem of obtaining a high-precision equation of state for stellar interiors. Such an attempt was made for the first time by the OPAL group at Livermore as part of their opacity project (see STELLAR OPACITY).

To explain the advantages of this approach for partially ionized plasmas, it is instructive to discuss the activity expansion for gaseous hydrogen. The interactions in this case are all short ranged and pressure is determined from a self-consistent solution of the equations

$$\frac{p}{k_{\rm B}T} = z + z^2 b_2 + z^3 b_3 + \cdots, \tag{21}$$

$$\rho = \frac{z}{k_{\rm B}T}\left(\frac{\partial p}{\partial z}\right), \tag{22}$$

where $z = \lambda_e^{-3}\exp(\mu/k_{\rm B}T)$ is the activity (with λ_e being the thermal de Broglie wavelength of electrons, see above), and μ is the chemical potential. The b_n are cluster coefficients such that b_2 includes all two particle states, b_3 includes all three particle states, etc.

In contrast to the chemical picture, which is plagued by divergent partition functions, the physical picture has the power to avoid them altogether. An important example of such a fictitious divergence is that associated with the atomic partition function. This divergence is fictitious in the sense that the bound-state part of b_2 is divergent but the scattering state part, which is omitted in the Saha approach, has a compensating divergence. Consequently the total b_2 does not contain a divergence of this type. A major advantage of the physical picture is that it incorporates this compensation at the outset. A further advantage is that no assumptions about energy-level shifts have to be made; it follows from the formalism that there are none.

As a result, the Boltzmann sum appearing in the atomic free energy for atoms and compound ions is replaced with the so-called Planck–Larkin partition function (PLPF), given by

$$\mathrm{PLPF} = \sum_{nl}(2l+1)\left[\exp\left(-\frac{E_{nl}}{kT}\right) - 1 + \frac{E_{nl}}{k_{\rm B}T}\right]. \tag{23}$$

The PLPF is convergent without additional cut-off criteria as are required in the chemical picture. However, despite its name the PLPF is not a partition function, but merely an auxiliary term in the second expansion (virial) coefficient of pressure.

Currently available OPAL equation-of-state tables are described in the article by Rogers *et al* (1996). They can be used to model stars down to $0.8M_\odot$. A newer version of OPAL equation-of-state tables is now being released. It will allow the modeling of much less massive stars, down to $0.1M_\odot$. However, models of brown dwarfs and giant planets still need configurational-integral approaches, such as that of Saumon *et al* (1995).

Practical recommendations

When constructing a stellar evolution or pulsation code, opacity and equation of state formulations are necessary. These may be in the form of analytical or in-line routines that calculate the desired opacities or thermodynamic quantities, given as input the temperature, density, and chemical composition. Pre-calculated opacity or equation of state tables may also be used, with appropriate look-up and interpolation routines. Sometimes several

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

tables or analytical fits covering different regions of T, ρ, and composition space must be smoothly joined. The hydrostatic structure calculation may require fewer thermodynamic quantities or derivatives than a pulsation calculation; for example, in pulsation calculations, smooth derivatives of opacity with respect to temperature and density, as well as additional derivatives of thermodynamic quantities, are required.

There are advantages and disadvantages of using tables versus in-line analytical formulations. Tables are often limited in coverage of T, ρ and composition space, and so may not be useful for stellar models over a wide range of masses, compositions, and evolutionary states. Tables do not always include all of the thermodynamic quantities needed for a calculation, and these quantities cannot always be constructed from the information provided. Interpolation within and between tables, or extrapolation off the edges of tables can be difficult and risky, and introduces significant uncertainty into the calculations. The grids of table entries are not always fine enough for good interpolation, or for construction of derivatives using differences between table quantities. On the other hand, tables often include superior physics compared to simple analytical schemes, and can be less computationally expensive than a sophisticated analytical routine.

Analytical procedures can be more flexible, for example in taking into account variations in element mixtures and composition, or in providing smooth derivatives. They also may be modified easily to generate additional thermodynamic quantities. Analytical procedures can also be useful to explore the importance of various physical processes, since these processes can easily be turned on and off, exaggerated with multipliers, etc.

The book by Eliezer *et al* (1986) contains many valuable resources, formulas, and tables for practical calculations of thermodynamic quantities. More rigorous theoretical concepts are discussed in the book by Kraeft *et al* (1986). Finally, references to some commonly used thermodynamic data (and opacity), both analytical and tabular and including information about how to obtain them, are found in the article by Däppen and Guzik (2000). A short version is given elsewhere in this volume (see SOLAR INTERIOR: EQUATION OF STATE AND OPACITY).

Bibliography

Däppen W and Guzik J A 2000 Astrophysical equation of state and opacity *Variable Stars as Essential Astrophysical Tools (NATO-ASI, Çeşme, Turkey, 1998)* ed C Ibanoglu (Dordrecht: Kluwer) pp 177–212

Eliezer S, Ghatak A and Hora H 1986 *An Introduction to Equations of State: Theory and Applications* (Cambridge: Cambridge University Press)

Kraeft W D, Kremp D, Ebeling W and Röpke G 1986 *Quantum Statistics of Charged Particle Systems* (New York: Plenum)

Rogers F J, Swenson F J and Iglesias C A 1996 *Astrophys. J.* **456** 902–8

Saumon D, Chabrier G and Van Horn H M 1995 *Astrophys. J. Suppl. Ser.* **99** 713

Seaton M J 1995 *The Opacity Project* vol 1 (Bristol: Institute of Physics)

Werner Däppen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Masses

The mass of a star is arguably its most important property. In this article we will answer four questions concerning stellar masses. How do the properties of stars depend on their masses? What is the smallest and largest mass possible for a star, and why? How do we determine the masses of stars? What distribution of stellar masses occurs when stars form, and why?

Dependence of other stellar parameters on mass

The Russell–Vogt theorem states that, if we know a star's mass and its CHEMICAL COMPOSITION, we can use the laws of physics to determine all of its other properties: its luminosity, its radius, its temperature and density profiles, and how these properties change with time. (We know that this is a slight simplification; for instance, the amount of net ANGULAR MOMENTUM will also affect a star's structure and evolution.) Compared with the possible range of masses a star may have ((0.08–150)$M_\odot$), there is only modest variation possible in the initial composition, and thus it is primarily a star's mass at birth which determines the basic essentials of its structure and future life.

Some of the properties of stars are given in table 1 as a function of stellar mass for stars on the main sequence, the core H-burning phase that accounts for 90% of a star's life. These values have been taken from stellar models computed with a composition that is initially solar. We list the stellar parameters at the beginning and end of the main-sequence lifetimes, except for the lowest-mass stars, for which we adopt the parameters corresponding to an age of 1 Gyr, by which time these stars are stably burning hydrogen.

Generally the behavior of the stellar parameters with stellar mass is quite different for the higher-mass stars ((25–120) $M_\odot$) than for solar-type stars ((0.8–1.2) $M_\odot$). The dependence of luminosity on stellar mass is shown in figure 1. This mass–luminosity relationship is considered one of the most fundamental descriptions of stellar properties; the ability to reproduce this by stellar models was one of the great vindications of theory (see also HERTZSPRUNG–RUSSELL DIAGRAM). EDDINGTON first demonstrated that radiative diffusion in stars requires that the stellar luminosity will depend on mass roughly as the fourth power, i.e. $L \sim M^4$. However, it is clear from figure 1 that no single exponent describes the dependence of luminosity on mass over the entire range of stellar masses. If we consider different mass ranges we would find that the following are good approximations:

$$L \sim M^{1.6} \ (M \approx 100 M_\odot)$$

$$L \sim M^{3.1} \ (M \approx 10 M_\odot)$$

$$L \sim M^{4.7} \ (M \approx 1 M_\odot)$$

and

$$L \sim M^{2.7} \ (M \approx 0.1 M_\odot).$$

The drastic changes seen in the mass–luminosity relation with mass are primarily due to the different opacity sources at work. At the high interior temperatures that characterize high-mass stars, all of the atoms are fully ionized and scattering of x-rays from free electrons dominates the opacity, with no temperature dependence. At lower temperatures, atoms are only partially ionized, and there is a strong temperature dependence in the number of ions. At the very cool temperatures that characterize the lowest-mass stars, molecular hydrogen (H_2) forms, removing the dominant opacity source for solar-type stars, H^- (see also STELLAR OPACITY).

Stellar lifetimes τ_{ms} as a function of mass also show a marked change from solar-type stars to high-mass stars, as evidenced by table 1. For solar-type stars the main-sequence lifetime changes rapidly with mass, while for higher-mass stars the change is far more modest with mass. For most stars, roughly the same fraction of a star's mass (10%) is involved in nuclear burning regardless of mass, and so the relative main-sequence lifetime τ_{ms} will be roughly proportional to the mass (the amount of fuel) and inversely proportional to the luminosity (how quickly the fuel is consumed), i.e. $\tau_{ms} \sim M/L$. Given the mass–luminosity relations above, we can thus estimate the dependence of lifetime on mass as $\tau_{ms} \sim M^{-3.7}$ for solar-type stars, and $\tau_{ms} \sim M^{-0.6}$ for very-high-mass stars. This rule of thumb breaks down for the lowest mass stars, as the stars are fully convective, and the hydrogen-burning main sequence lasts a good deal longer than one would expect. As shown in table 1, a $0.1M_\odot$ star will last 10 trillion years (1.0×10^{13} yr) in a core-H-burning phase, roughly 1000 times as long as the Sun will, rather than the factor of 100 that one would expect, since the entire star provides the nuclear fuel.

During the main sequence the highest-mass stars lose a significant fraction of their mass owing to stellar winds. A star that begins life with $120M_\odot$ will lose $50M_\odot$ (40%) of its mass, while a $60M_\odot$ star will lose $12M_\odot$ (20%), by the end of its main-sequence life. Below $25M_\odot$ the amount of mass lost during main-sequence evolution is negligible, although stellar winds do affect the evolution of even solar-type stars by carrying off angular momentum. Such mass loss is expected to scale with metallicity and thus will be less significant in galaxies of lower metallicity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Properties of main-sequence stars as a function of stellar mass.

Mass ($M_\odot$)	τ_{ms}	Beginning of main sequence				End of main sequence			
		T_{eff} (K)	Spectral type	log $L/L_\odot$	Radius ($R_\odot$)	T_{eff} (K)	Spectral type	log $L/L_\odot$	Radius ($R_\odot$)
120 $M_\odot$	2.56 Myr	53 300	O3 V	+6.25	16	32 900	O9 I	+6.34	48
60 $M_\odot$	3.45 Myr	48 200	O4 V	+5.73	10	12 000	B7 I	+5.99	230
25 $M_\odot$	6.51 Myr	37 900	O8 V	+5.29	6.5	29 000	B0 I	+5.29	18
12 $M_\odot$	16.0 Myr	28 000	B0.2 V	+4.01	4.3	24 400	B0.5 I	+4.46	9.5
5 $M_\odot$	94.5 Myr	17 200	B5 V	+2.74	2.7	15 100	B5 I	+3.15	5.5
2.5 $M_\odot$	585 Myr	10 700	B9 V	+1.60	1.8	9000	A2 III	+1.92	3.8
1.25 $M_\odot$	4.91 Gyr	6380	F5 V	+0.32	1.2	6070	G0 V	+0.66	1.9
1.0 $M_\odot$	9.84 Gyr	5640	G8 V	–0.16	0.9	5790	G2 V	+0.22	1.3
0.8 $M_\odot$	25.0 Gyr	4860	K2 V	–0.61	0.7	5360	K0 V	-0.09	1.1
0.5 $M_\odot$	100 Gyr	3890	M0 V	-1.42	0.4	—	—	—	—
0.2 $M_\odot$	4000 Gyr	3300	M4 V	–2.2:	0.2	—	—	—	—
0.1 $M_\odot$	10 000 Gyr	2900	M7 V	–3.0:	0.1	—	—	—	—

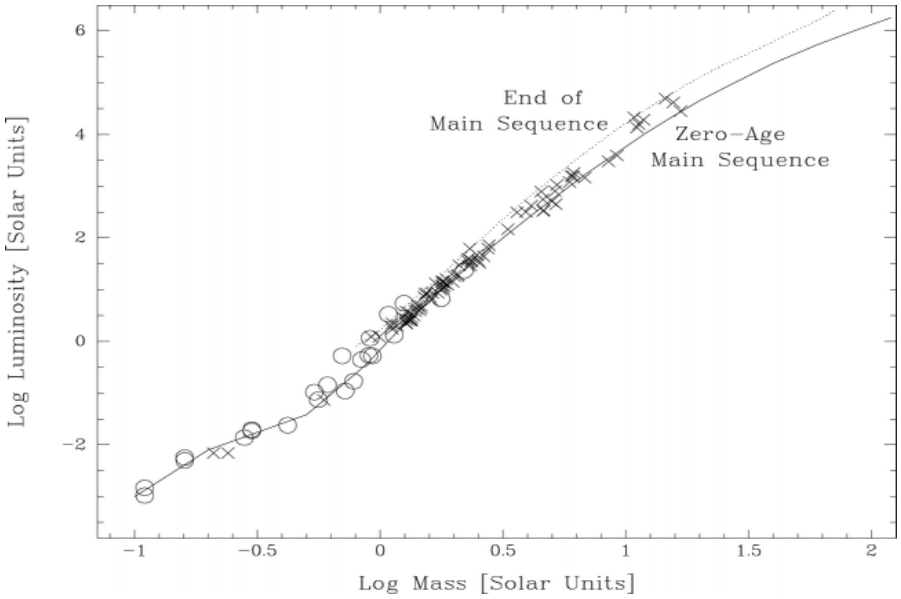


Figure 1. The mass–luminosity relationship as predicted by stellar models is shown by the solid curve for stars with zero age and by the dotted curve for stars at the end of their main-sequence lifetimes, for masses of 0.8 $M_\odot$ (log M = –0.1) and greater. Below that mass, the curve shown is for models with an age of 1 Gyr, as stars of this age have begun to burn hydrogen stably with constant luminosity. (The vast majority of low-mass stars will be at least that old.) The points show the masses and luminosities of 'real' stars for comparison, with the crosses denoting the best determinations from double-lined spectroscopic binaries, and the open circles denoting the best determinations from visual binaries.

It is inappropriate to speak of a spectral type to mass relationship for higher-mass stars: stellar evolution results in a progression from higher effective temperatures to cooler during the core-H-burning lifetime, and during this evolution stars of different masses will pass through a particular spectral type 'stage'. For lower-mass main-sequence stars this is not true, and there is only a slight change of spectral type with evolution (i.e. little change of the effective temperature). For example, a star which is spectroscopically classified as 'O4 V' star may be a zero-age 60$M_\odot$ star, or a slightly older (0.5 Myr) 85$M_\odot$ star, but all stars of spectral type G2 V will have a mass roughly that of the Sun.

Range of stellar masses: the lowest- and highest-mass stars

The masses of stars span the range from 0.08 to 150 (or more) times the mass of the Sun.

At the low-mass end, the 0.08$M_\odot$ limit is set by the stellar core not being hot enough to ignite hydrogen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stably. Objects with masses slightly below this limit are called brown dwarfs, and are 'star like' in the sense that nuclear burning of deuterium occurs in their core. Below a mass of $0.015M_{\odot}$ (roughly 16 times the mass of Jupiter) not even deuterium burning can occur, and these objects are perhaps best called planets. Thus there is a natural lower limit to what constitutes a star, although we expect that the mass function (discussed below) should be unaware of this division. We note that the distinction between brown dwarfs and *bona fide* stars is subtle in the following sense: both low-mass stars and brown dwarfs burn primordial deuterium at first, but a 'real' star will 'eventually' settle down to stable H burning; we expect in the case of a $0.08M_{\odot}$ star that this will take approximately a billion years (1 Gyr).

In 2001, a very low mass star (lower than a tenth of a solar mass), located at a distance of only 13 light years, designated as DENIS-P J104814.7-395606.1, was found on the infrared survey DENIS images. On the night of 30 May 2000, the 10 m Keck I telescope was used to obtain a high-resolution spectrogram that showed the presence of a strong cesium absorption line, as well as titanium and vanadium oxide bands. The signature of lithium was not present in the spectrum of the object. The Keck data demonstrated that this object must be a very nearby dwarf (cesium is not detected in giants), with a temperature of about 2200 K and a mass between 90 and 60 Jupiters. The lack of lithium implies that the mass must be larger than 60 Jupiter masses, but does not rule out that it could be a massive brown dwarf. A very nearby dwarf should have an apparent motion with respect to the background of more distant stars. Astronomers at the Observatory of Paris searched for DENIS-P J104814.7-395606.1 in old photographic plates and found it in images obtained during the last 30 years. It has moved considerably between 1986 and 1999. The large proper motion (1.5 arcsec per year) confirms that this is one of our nearest neighbors. The astronomers estimate a distance of about 13 light years. This distance is still uncertain because it is based on comparisons with objects of the same spectral type.

At the other extreme, we do not understand what, if anything, limits how large a mass a star may have. At one time it was thought that radiation pressure acting on grains limited how large a star could form, but we now understand that disks play an important role in the formation of stars. There may be sufficient shielding by the inner part of the disk to mitigate the effects of radiation pressure. It is not clear at this time what role the mergers of protostellar clumps may play in the formation of stars. If the role is appreciable, then there may be *no* natural limit to how massive a star may actually form.

Even if star-formation processes fail to limit the mass of a star, other processes may. Eddington proposed in 1926 that stars more massive than some amount would be pulsationally unstable and should blow off their outer layers, thus limiting their mass. Early estimates of this limit were as low as $60M_{\odot}$. Modern estimates, however, place this limit as high as $440M_{\odot}$, although this is still based on the same classical perturbation linerarization methods used by Eddington. Recent 'nonlinear' analysis (i.e. direct numerical integration of the equations of stellar structure) suggests that the mass loss from such instabilities would only be comparable with the mass loss of radiatively-driven stellar winds in any event.

In this context it is interesting to note that the highest-mass stars we know do all show signs of prodigious mass loss. The highest-mass main-sequence stars known are located in the R136 cluster at the heart of the 30 Doradus nebula in the LARGE MAGELLANIC CLOUD (LMC). These stars have masses which have been conservatively estimated as being as high as $155M_{\odot}$. Spectroscopically the eight most massive of these stars show evidence of extremely high mass-loss rates (mimicking the appearance of WOLF–RAYET STARS), and so one could argue that indeed these stars are not 'stable' in the sense that they are losing a considerable amount of matter.

The luminosities of the most luminous R136 stars are $10^{6.6}L_{\odot}$. Other stars which are of comparable luminosity include HD 5980, a Wolf–Rayet star in the SMALL MAGELLANIC CLOUD (SMC), η CARINAE, a luminous blue variable (LBV) in the Milky Way, Sk − 67º 211, an O3 III star in the LMC, and the Pistol Star, an LBV located near the Galactic center. It is hard to determine masses for LBVs and Wolf–Rayet stars, as these are in an He-burning phase, where the interior models (and hence the mass–luminosity relationship) for massive stars are quite uncertain, but it is clear that these stars evolved from stars of mass similar to that of the highest-mass R136 stars. The mass inferred for the main-sequence star Sk − 67º 211 is also like that of the R136 stars, suggesting that the stars in R136 do not have some kind of special origin. Studies of the youngest OB associations and clusters (i.e. young enough so that not even the highest-mass stars would have evolved) show that the richer the cluster is in stars, the higher the mass of the highest-mass star seen. We now understand that although physics may indeed impose a limit on how massive a star may be, we have not yet encountered this limit in nature.

Determination of stellar masses

How do we determine stellar masses? There are two basic ways: (1) direct determination of masses observationally using binaries and (2) inference of stellar masses using models.

Stellar binaries

Simple Newtonian mechanics, specifically Kepler's third law of planetary motion, allow us to directly determine the masses of stars in some BINARY STAR systems.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

'Double-lined' spectroscopic binaries are stars whose spectra show the signature of two stars. The orbital periods of these systems are usually a few days or months, and the line-of-sight (radial) velocities of each component can be directly measured as the DOPPLER EFFECT causes the spectral lines of one star to first appear blueshifted and then redshifted relative to the lines' average position. Masses can be determined directly if the orbital inclination can also be found via light variations (i.e. eclipsing or ellipsoidal) or by the direct resolution of such systems through techniques such as speckle imaging or long-baseline interferometry. The masses and luminosities determined from the best, well-separated binaries are shown in figure 1, and we see that the mass–luminosity relationship inferred from such systems is in excellent agreement with that predicted by modern stellar interior models.

Missing from the figure are any high-mass stars; many of these systems are in physical contact or are sufficiently close to have undergone some mass transfer. Searches for high-mass spectroscopic binaries whose components are cleanly detached are continuing.

For visual binaries, masses can be determined if the period is short enough to be observed and the distance to the system is known. However, the distance needs to be known to high accuracy for the method to yield useful results: a 7% error in the parallax of the system leads to a 20% accuracy in the masses. There are only 14 systems for which good radial-velocity orbits and parallaxes are known, and we include these data in figure 1. New parallax determinations with the HIPPARCOS satellite will provide improved data on many more systems. Since the orbital periods of these visual binaries are tens or even hundreds of years, reliable measurements over a substantial time period are needed for orbit determinations. High-resolution imaging studies with the HUBBLE SPACE TELESCOPE or new ground-based techniques are providing important new fundamental data on such systems.

Stellar models

If the effective temperature and luminosity of a star are known (from spectroscopic observations of a star whose distance is known, either from parallax or from membership in a cluster with a known distance), then stellar interior models can be used to approximate the star's mass. This method is the basis for most of the masses inferred in determining the initial mass function (discussed in the next section).

It is also possible to estimate a star's mass from STELLAR ATMOSPHERE models. Again, spectroscopy is needed of a star with known distance. By fitting the stellar lines and comparison to model atmospheres, it is possible to determine the effective temperature $T_{\rm eff}$ and surface gravity g. Since the star's luminosity L is also known, it is possible to determine the stellar radius R since $L \sim R^2 T_{\rm eff}^4$. The mass of the star can then be found since $g \sim M/R^2$.

For the most massive stars, there appears to be a significant 'mass discrepancy' between the masses derived from stellar atmosphere and stellar interior models, stellar atmospheres predicting masses which are systematically smaller. The reason for this discrepancy is unknown at present, but is largest for the most luminous, massive supergiants, for which there may be factors of 2 differences between the two methods. Attempts to resolve this discrepancy by means of spectroscopic binaries have been frustrated by the same lack of identified high mass 'detached' systems described above.

Distribution of stellar masses

If we were to count stars as a function of mass in the solar neighborhood, we would find that there were far more low-mass stars than high-mass stars. The reasons for this are basically two-fold. Low-mass stars live much longer than do high-mass stars, and so have accumulated over most of the life of the Galaxy, while high-mass stars quickly consume their fuel and die. The second reason, however, is that in a typical star-forming event many more low-mass stars are formed than are high-mass ones.

Knowing the distribution of stellar masses that is obtained when stars form from clouds of gas and dust in space is important for two reasons. First, because the light observed from star clusters and galaxies is dominated by a few of the brightest stars (the tip), it is important to know how many low-mass stars are associated with the iceberg as a whole. Indeed, in a stellar system such as our own Milky Way galaxy, most of the observed luminosity comes from stars greater than $10M_\odot$, while most of the mass is locked up in stars with masses below $1M_\odot$. Secondly, the distribution of stellar masses at birth (the initial mass function (IMF)) provides clues to the processes of stellar formation. Changes in the shape of this distribution function with mass provide evidence for the critical scales associated with the star formation process.

In general the IMF can be thought of simply as a probability function $\phi(M)$, representing the likelihood of forming a star with a mass between M and $M + {\rm d}M$. In 1955 SALPETER found that the IMF of stars near the Sun was well represented as a power law, with $\phi(M) \sim M^{-2.35}$. Modern estimates from studies of OB associations in the Milky Way and LMC suggest that for stars of mass greater than $5M_\odot$ the IMF is indeed very similar to Salpeter's result, with an exponent of -2.3 ± 0.3 obtaining regardless of cluster density or metallicity (figure 2).

For intermediate- and low-mass stars, studies of the IMF have traditionally been done using volume-limited samples of stars found in neighborhood of the Sun. Using a variety of techniques (photometric, spectroscopic and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

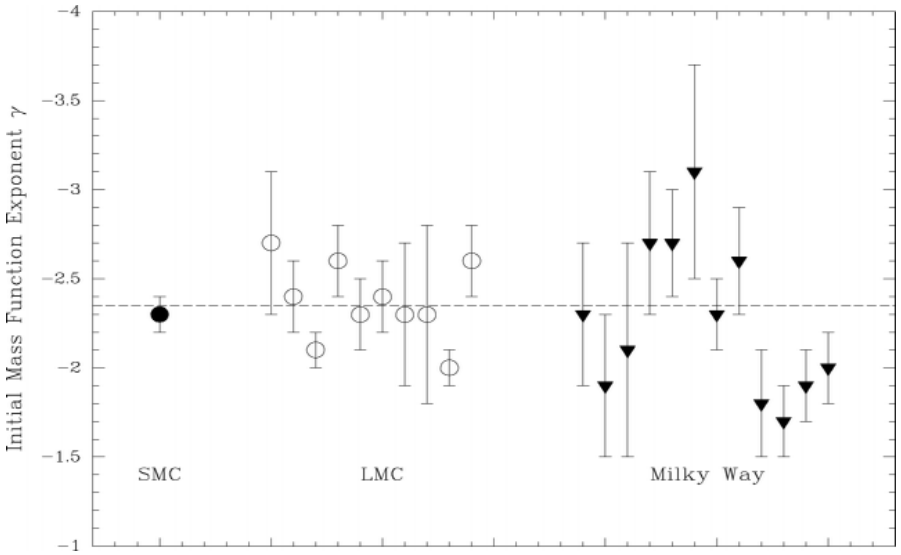


Figure 2. The universality of the IMF for high-mass stars is demonstrated in this figure showing the IMF exponent γ (where the distribution function $\phi(M) \sim M^{\gamma}$) for massive stars in the OB associations of the SMC, LMC and Milky Way. The dashed line is for a Salpeter exponent of −2.35. The metallicities change by a factor of 4 between these three systems.

parallactic) a luminosity function can be constructed for an unbiased sample of stars. Adopting a mass–luminosity relationship appropriate for the sample in question then allows one to transform the luminosity function into a mass function. In practice, one must take into account the metallicity and evolutionary state of the sample, as well as correct for the relative life-times. The derived IMFs will also depend on what assumptions have been made about the star-formation of the region, with increased dependence at lower masses, where the star-formation history over the entire life of the galaxy is relevant. Such studies now suggest that the power law may be somewhat less steep for $(1–5)M_{\odot}$ stars than for stars of higher mass. For stars of even lower mass, studies are hampered by the additional uncertainty of the mass–luminosity relationship for very cool objects. Most work is consistent with $\phi(M) \sim M^{-1.0\pm0.5}$ over the range $(0.1–1)M_{\odot}$. There is disagreement, however, as to whether or not the number of low-mass stars that are formed continues to rise to lower and lower masses, or whether the relationship flattens out or even turns over somewhere near the low-mass end of this range.

Another technique that has been exploited in determining the IMF is the use of star clusters in which all stars appear to have roughly the same age. By using such coeval groups of stars, problems inherent in correcting for star-forming histories are eliminated, although other concerns (such as dynamical evolution) need to be addressed. Studies of GLOBULAR CLUSTERS with ages greater than 1 Gyr derive IMFs consistent with the field down to a limiting masses of $(0.2–0.3)M_{\odot}$. GALACTIC OPEN CLUSTERS with ages 30–500 Myr, perhaps the best place to constrain the IMF near $1M_{\odot}$, confirm that the IMF is less steep than at masses greater than $5M_{\odot}$. Recent studies of nearby young open clusters such as the Pleiades and α Per have begun to probe the IMF down below the hydrogen-burning limit. Finally, by studying clusters still embedded in the molecular cloud cores from which stars form, we can attempt to relate different outcomes of the star-forming process to the initial conditions of formation. Thus far, IMFs derived from a wide variety of stellar populations in the Milky Way and local group galaxies are consistent with having been drawn from the same distribution as stars found in the neighborhood of the Sun. However, these comparisons are still in their infancy: there could be significant, more subtle variations present in the IMF that have gone undetected.

We have not yet touched at all upon the physical causes of the IMF: when an interstellar cloud of dust and gas collapses, what processes dominate and result in the distribution of masses we would observe at some future time? One might naively expect that the dominant physics could be readily deduced from the observed IMF, but this turns out not to be the case, primarily because the IMF appears to be so featureless. Instead, the observed IMF can only be used to constrain star-formation theories at present. Power-law distributions may result from a variety of different scenarios, including so-called ‘fragmentation’ theories. The earliest of these was proposed in 1953 when HOYLE suggested that the Jeans mass (the minimum mass needed for gravitational collapse) could result in a hierarchical distribution of masses. The Jeans mass

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

depends on both temperature and density, and as a cloud collapses the density will increase, while radiation from newly formed stars controls the gas temperature. However, it is not clear whether this elegant model would apply in a real molecular cloud. Alternatively, the agglomeration of protostellar clumps has been suggested as a way to produce a power law. Yet another theory involves feedback from the formation process itself: ignition of a powerful outflow from the protostar might halt further accretion once a characteristic mass has been reached. The resulting distribution of stellar masses would then depend on the ranges of initial values of a variety of physical parameters. A crucial component in evaluating current theories of star formation is whether or not the IMF is 'scale free' (such would be the case if it were well described by a single power law), or whether there is a characteristic mass scale, as is suggested by current observational evidence that the IMF begins to flatten out around $1M_{\odot}$. Solid observational knowledge of the shape of the IMF at even lower masses, and under differing physical conditions, is a prerequisite to understanding star formation.

Bibliography

The following is a list of interesting papers organized in the same order as the topics discussed in this article. In terms of understanding the physics of stars, and how other stellar parameters depend upon stellar masses, we can do no better than to recommend the following three classics:

Cox J P and Giuli R T 1968 *Principles of Stellar Structure* (New York: Gordon and Breach)

Hansen C J and Kawaler S D 1994 *Stellar Interiors* (New York: Springer)

Schwarzschild M 1958 *Structure and Evolution of the Stars* (New York: Dover)

The realization of these principles can be best seen in actual stellar evolutionary tracks. The dependence of physical parameters on stellar masses used in this article was computed using the evolutionary interior models of Andre Maeder's group in Geneva; these models do an excellent job of matching observations and are well described in the literature. The most recent version of these models can be found in the following two papers:

Schaller G, Schaerer D, Meynet G and Maeder A 1992 New grids of stellar models from 0.8 to 120 solar masses at Z = 0.020 and Z = 0.001 *Astron. Astrophys. Suppl.* **96** 269–331

Schaerer D, Meynet G, Maeder A and Schaller G 1993 Grids of stellar models. II—from 0.8 to 120 solar masses at Z = 0.008 *Astron. Astrophys. Suppl.* **98** 523–7

As for the range of stellar masses, the best discussion for the lowest mass stars can be found in

Burrows A, Hubbard W B, Saumon D and Lunine J I 1993 An expanded set of brown dwarf and very low mass star models *Astrophys. J.* **406** 158–71

For a modern discussion of the importance of the pulsational instability that may exist for the highest mass stars, we recommend the following:

Appenzeller I 1987 Theory of vibrational instabilities in luminous early type stars *Instabilities in Luminous Early Type Stars* ed H J G L M Lamers and C W H Loore (Dordrecht: Reidel) pp 68–72

The exciting discovery of the highest-mass stars may be of interest to some readers:

Massey P and Hunter D A 1998 Star formation in R136: a cluster of O3 stars revealed by Hubble Space Telescope spectroscopy *Astrophys. J.* **493** 180–94

An excellent review of the empirical determination of stellar masses in binary systems is given by

Popper D M 1980 Stellar masses *Ann. Rev. Astron. Astrophys.* **18** 115–64

The first reconsideration of visual binaries based on *Hipparcos* results is given by

Soderhjelm S 1999 Visual binary orbits and masses post Hipparcos *Astron. Astrophys.* **341** 121–40

A discussion of the 'mass discrepancy' for massive stars can be found in:

Burkholder V, Massey P and Morrell N 1997 The 'mass discrepancy' for massive stars: tests of models using spectroscopic binaries *Astrophys. J.* **490** 328–42

As for the distribution of stellar masses, we recommend

Kroupa P, Tout C A and Gilmore G 1993 The distribution of low-mass stars in the Galactic disc *Mon. Not. R. Astron. Soc.* **262** 545–87

Meyer M R, Adams F C, Hillenbrand L A, Carpenter J M and Larson R B 2000 The stellar initial mass function: constraints from young clusters and theoretical perspectives *Protostars and Planets IV* ed V G Mannings, A Boss and S Russel (Tucson, AZ: University of Arizona Press) in press

Scalo J M 1986 The stellar initial mass function *Fundam. Cosm. Phys.* **11** 1–278

and, in particular, a recent conference on the IMF:

Gilmore G and Howell D (ed) 1998 *The Initial Mass Function* (San Francisco, CA: ASP)

Philip Massey

Michael R Meyer

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Nomenclature

Stars are identified in various ways. Most bright stars have proper names which are universally recognized—for example, Antares, Rigel, Vega. A more precise designation, applicable to many more stars, including all those of naked-eye magnitude, consists of a letter or number followed by the Latin genitive of the name of the constellation in which the star is situated. This is in fact a combination of two systems, the Bayer letter and the Flamsteed number.

Bayer letters were introduced by the Bavarian lawyer/astronomer Johann Bayer (1572–1625) in his star atlas *Uranometria* of 1603; they consist of a lower-case Greek letter followed by the genitive constellation name; thus the Bayer letter for Antares is 'α Scorpii', for Rigel 'β Orionis' and for Vega 'α Lyrae'. The Bayer letters were allocated approximately in descending order of brightness within the constellation, but this does not hold strictly true in all cases, as Bayer was influenced by the detail of the constellation figure as drawn in his atlas, and also some stars have changed in brightness since Bayer's time. Thus Betelgeuse (α Orionis) is in fact less bright than Rigel (β Orionis). Clearly the number of stars in a constellation allocated Bayer letters is limited to 24, the number of letters in the Greek alphabet.

To overcome this limitation, fainter stars are designated by the numbers allocated to almost 3000 stars by the first Astronomer Royal, John Flamsteed (1646–1719), in his star catalog *Historia Coelestis Britannica*, which was published posthumously in 1725. Flamsteed allocated the numbers in order of increasing right ascension, so the brightest stars do not necessarily have the lowest numbers. Thus the Flamsteed numbers for the three stars mentioned above are 21 Scorpii (Antares), 19 Orionis (Rigel) and 30 Lyrae (Vega), whereas 1 Scorpii, 6 Orionis and 16 Lyrae are very much fainter stars. It is often convenient to abbreviate the constellation name. At its inaugural General Assembly in Rome in 1922 the International Astronomical Union adopted a list of three-letter abbreviations which has remained the standard; thus our three typical bright stars are conveniently designated 'α Sco', 'β Ori' and 'α Lyr'.

The huge numbers of additional faint stars discovered since Flamsteed's time can only be identified by their numbers in various catalogs. These may be designations from 'general' catalogs, such as the *Bonner Durchmusterung* of 1859–62, or *The Bright Star Catalog* (Yale, 1930; fourth revised edition 1982) which uses the same numbers as the *Harvard Revised Photometry* of 1908, therefore known as 'HR' numbers. Some stars have additional designations from catalogs compiled for a specific purpose, for example *The Henry Draper Catalog of Stellar Spectra* (Harvard, 1918–24) or the *General Catalog of Variable Stars* (Moscow, 1958; fourth edition 1985–90). A single star may thus have many designations, depending on the catalogs in which it is listed; hence Vega also has the numbers HR 7001 (from *The Bright Star Catalog*), BD + 38°3238 (from the *Bonner Durchmusterung*) and HD 172167 (from *The Henry Draper Catalog*).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Opacity

The monochromatic opacity, κ_ν, quantifies the property of a material to remove energy of frequency ν from a radiation field. A harmonic average of κ_ν, known as the Rosseland mean, κ_R, is frequently used to simplify the calculation of energy transport in stars. The term 'opacity' is commonly understood to refer to κ_R.

Opacity plays an important role in stellar modeling because for most stars RADIATION is the primary mechanism for transporting energy from the nuclear burning region in the core to the surface. Depending on the mass, convection and electron thermal conduction can also be important modes of stellar energy transport. The efficiency of energy transport is related to the temperature gradient, which is directly proportional to the mean radiative opacity in radiation-dominated regions. When the radiative opacity is large, convection can become the more efficient energy transport mechanism. Electron conductive opacity, the resistance of matter to thermal conduction, is inversely proportional to electron thermal conductivity. Thermal conduction becomes the dominant mode of energy transport at high density and low temperature.

The removal of energy from a beam of photons as it passes through matter is governed by spectral line absorption (bound–bound), photoelectric absorption (bound–free), inverse bremsstrahlung (free–free) and photon scattering. Stimulated emission acts as negative opacity by creating photons that add to the beam. Photon scattering is not true photoabsorption; it acts to reduce the photon flux by redirecting photons out of the beam and is not subject to stimulated emission.

The first step in computing the opacity is the determination of the equation of state, followed by the calculation of the various absorption and scattering cross-sections and finally the integration over frequency to obtain κ_R. Therefore, opacity calculations bring together several distinct disciplines, such as equation of state, atomic physics, spectral line broadening, as well as plasma collective behavior. These provide ion abundances, energy level values, radiation transition probabilities, spectral line profiles and structure factors needed to carry out the calculations. Since atoms from different elements absorb energy of a given frequency at different rates, it is necessary to also know stellar elemental abundances.

Molecules can form in the atmospheres of cool stars (see STELLAR ATMOSPHERES: COOL STARS). The discussion herein is concerned with stellar interior opacities and does not encompass the molecular regime. The methods for calculating molecular opacities follow a somewhat similar path.

History of opacity calculations

As long ago as 1926 Eddington identified opacity as one of two clouds obscuring stellar model calculations, the source of stellar energy being the other. At that time it was thought that bound–bound absorption was not a significant source of opacity. It was 40 yr later that Cox and Stewart included spectral lines in opacity calculations. This led to opacity increases exceeding a factor of 3 at some temperature–density conditions. These enhanced opacities greatly improved the quality of stellar models. Even so, several properties of stars known to be sensitive to opacity resisted explanation. It was another 20 yr before Simon speculated that the most likely explanation was that the heavy element contribution to opacity was still significantly underestimated. Two independent efforts, OPAL and OP, recalculated stellar opacities and found that an improved treatment of atomic data leads to an additional factor of 3 increase in the opacity under some conditions. The OPAL work was carried out by Iglesias, Rogers and Wilson. The OP work was carried out by Seaton and a international group of collaborators. These new opacities helped resolve a number of long-standing problems in stellar modeling, particularly in regard to pulsating stars. Direct laboratory verification of the new opacity calculations has only recently started to appear. These show good agreement with the most recent calculations and large differences with standard opacity results used prior to 1985.

Definition of κ_ν

The total extinction from a beam of photons as it passes through matter is given by

$$\kappa_\nu = \kappa_\nu^{\mathrm{abs}} + \kappa^{\mathrm{S}}$$

where

$$\kappa_\nu^{\mathrm{abs}} = (1 - \mathrm{e}^{-h\nu/kT}) \sum_j x_j (\kappa_j^{\mathrm{bb}} + \kappa_j^{\mathrm{bf}} + \kappa_j^{\mathrm{ff}})$$

is the total photoabsorption, the sum j is over all elements in the mixture, x_j is the elemental number fraction, κ_ν^{bb}, κ_j^{bf} and κ_j^{ff}, are, respectively, the bound–bound, bound–free and free–free frequency-dependent photoabsorption contributions to opacity, κ^{s} is the scattering contribution and the factor $1 - \exp(-h\nu/kT)$ accounts for stimulated emission. When the density is low enough that the material can be considered to be composed of individual particles, the photon extinction due to a process p is given by

$$\kappa_\nu^{\mathrm{p}} = \sum_j N_j \sigma_\nu^{\mathrm{p}j}$$

where N_j is the number density of particles of type j and $\sigma_\nu^{\mathrm{p}j}$ is the cross-section for photons of frequency ν. Opacity has units of inverse length. However, numerical values are generally expressed as κ_ν/ρ in units of $\mathrm{cm}^2\ \mathrm{g}^{-1}$.

Mean opacities

The radiation transfer equation describes the transport of energy by photons and is equivalent to the Boltzmann equation in the kinetic theory of particle transport. For steady-state conditions it is given by

$$\frac{\mathrm{d}I_\nu(s, \boldsymbol{n})}{\mathrm{d}s} = -\kappa_\nu I_\nu(s, \boldsymbol{n}) + j_\nu \qquad (1)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where $I_\nu(s, \boldsymbol{n})$ is the radiation intensity at frequency ν in the direction $\boldsymbol{n}$ as a function of distance s, and j_ν is the emissivity.

STELLAR INTERIORS are 'collision dominated' so that the material can be considered to be in thermodynamic equilibrium at the local value of temperature and density. Also, because of the small temperature gradient, the radiation intensity is very close to Planckian, i.e. in a state of local thermodynamic equilibrium. Consequently, Kirchhoff's law,

$$j_\nu = \kappa_\nu B_\nu(T)$$

applies, where

$$B_\nu(T) = (2h\nu^3/c^2)(\mathrm{e}^{h\nu/\kappa T} - 1)^{-1}$$

is the Planck function. In addition, conditions change slowly over many photon mean-free paths and the radiation flow is nearly isotropic, so that the radiation transfer equation greatly simplifies. In this limit, known as the diffusion approximation, the solution to equation (1) as a function of radius, for a non-dispersive medium, integrated over frequency is

$$F(r) = -\frac{4\pi}{3}\frac{1}{\kappa_\mathrm{R}}\frac{\mathrm{d}B(r,T)}{\mathrm{d}r} \tag{2}$$

where $F(r)$ is the total flux, $B(r, T)$ is the integral of B_ν over frequency and

$$\frac{1}{\kappa_\mathrm{R}} = \int_0^\infty \mathrm{d}\nu \frac{1}{\kappa_\nu}\frac{\mathrm{d}B_\nu}{\mathrm{d}T} \Bigg/ \int_0^\infty \mathrm{d}\nu \frac{\mathrm{d}B_\nu}{\mathrm{d}T} \tag{3}$$

defines the Rosseland mean opacity. Equation (2) recovers the correct asymptotic solution to the transfer equation and the correct flux transport. The harmonic average in equation (3) emphasizes regions of weak absorption, where most of the energy flux is transported. A large value of the opacity indicates strong absorption from a beam of photons, whereas a small value indicates that the beam loses very little energy as it passes through the medium.

Near the surface of a star conditions change very rapidly and the material becomes optically thin. In this limit the assumptions leading to the diffusion approximation break down and a straight arithmetic average of $\kappa_\nu^{\mathrm{abs}}$, known as the Planck mean opacity,

$$\kappa_\mathrm{P} = \int_0^\infty \mathrm{d}\nu\, \kappa_\nu^{\mathrm{abs}} B_\nu(T) \Bigg/ \int_0^\infty \mathrm{d}\nu B_\nu(T)$$

is more suitable. In contrast to κ_R, strong absorption regions provide the major contributions to the Planck mean opacity. Note also that scattering is not included in the Planck mean.

The mean opacities greatly simplify solutions of the radiation transfer equation, but there are circumstances where it is necessary to use κ_ν directly. An example is radiatively driven diffusion in which low-abundance species are pushed toward the surface by a net radiation force in the field of the dominant species. This is thought to be the mechanism that causes large overabundances of Mn and Hg in the atmospheres of some late B and A stars.

Equation of state

Equation of state plays an important role in opacity calculations by providing the initial state occupation numbers needed to obtain κ_ν and also by determining the set of final states that exist at particular plasma conditions. This is a statistical mechanics problem that is complicated by the long-ranged Coulomb interaction between the particles. Two types of approaches are used to calculate the equation of state. One is the 'chemical picture' method which is based partly on intuitive reasoning, since it postulates the effect of the PLASMA on bound states to avoid the divergence of the atomic partition function. Numerous procedures have been proposed, but only one of these receives much current use. It is based on occupation probabilities, w, that separate the statistical weight for a particular state into a part that acts as a localized few-body (bound) state and a part $1 - w$ that acts as a delocalized (continuum) state. In this approach the bound state energies remain those of an isolated particle, but bound states are effectively destroyed by the action of the surrounding electric microfields when $w \to 0$. This results in a convergent bound state partition function.

The other approach used to calculate the equation of state is known as the 'physical picture' because the starting point is the Coulomb interaction between electrons and nuclei. It is based on an activity expansion of the grand canonical partition function for a plasma. This approach gives directly from theory the plasma effect on bound states and convergent atomic partition functions.

Photoabsorption and scattering processes

The physical extent of the absorbing atoms and ions in the plasma is generally much less than the wavelength of the radiation, so that only single-photon absorption through the electric-dipole interaction is of practical importance in opacity calculations. Electric-quadrupole absorption and magnetic-dipole absorption are smaller by several orders of magnitude in the same spectral region. Absorption of two or more photons at a time is also much less likely than single-photon absorption and has been shown to be an insignificant source of opacity in stellar plasmas. However, two-photon processes are involved in the scattering of radiation. During a scattering event an electron absorbs a photon of frequency ν and immediately re-emits another photon of frequency ν' in an arbitrary direction, effectively removing energy from the beam.

Bound–bound absorption

$$\mathrm{P_i} + h\nu \to \mathrm{P_f}.$$

Bound–bound transitions in atoms and ions are possible when the photon frequency is close to $h\nu_{\mathrm{if}} = E_\mathrm{f} - E_\mathrm{i}$, where E_i and E_f are the initial and final state energies. The cross-section for this process is given by

$$\sigma_{\mathrm{if}}(\nu) = \frac{2\pi^2 e^2}{mc} f \phi_{\mathrm{if}}(\nu),$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where

$$f = \frac{8\pi^2 m \nu_{\rm if}}{3 g_{\rm i} e^2 h} \left| \int_0^\infty \Psi_{\rm f}^*(\boldsymbol{r}) \boldsymbol{D} \Psi_{\rm i}(\boldsymbol{r})\, {\rm d}\boldsymbol{r} \right|^2$$

is the oscillator strength for the i → f transition,

$$\boldsymbol{D} = e \sum_n \boldsymbol{r}_n$$

is the dipole moment summed over all the electrons in P_i, m is the electron mass, g_i is the statistical weight, and Ψ_i^* and Ψ_f are the initial and final state wavefunctions. The total bound–bound cross-section is obtained by summing over all initial and final bound states in the plasma.

The line shape function, $\phi_{\rm if}(\nu)$, includes several processes. Natural broadening is a consequence of the finite lifetime of excited states in the radiation field. It has a normalized Lorentzian profile,

$$\phi(\nu) = 2\pi\Gamma[(\nu - \nu_{\rm if})^2 + (\pi\Gamma)^2]^{-1}$$

where Γ is the damping constant. The wings of the Lorentzian fall off as $\Delta\nu^2 = (\nu - \nu_{\rm if})^{-2}$, so that the infrared wing of the line can be an important source of opacity at very low density. However, the low-frequency line profile far from line center is expected to fall off as $\Delta\nu^4$ similarly to Rayleigh scattering and *ad hoc* methods are sometimes introduced to force this behavior. The thermal velocity distribution of the radiating atoms gives rise to a statistical Doppler shift of radiation frequencies that has a Gaussian distribution whose wings fall off rapidly compared with $\Delta\nu^2$.

As the density increases, other particles in the vicinity of the radiator increasingly affect the line shape. Electrons and ions are the primary perturbers when the ionization degree exceeds a few per cent, but neutrals can be important at low temperature. There is a large literature associated with the theory of one- and two-electron atoms and ions and accurate line shapes are available for these radiators. For hydrogenic ions, where there is degeneracy for levels with the same principal quantum number, the major source of line broadening is due to the linear Stark shift produced by the electric microfield of neighboring ions. In the simplest calculations the ions are assumed to be quasi-static and randomly distributed, leading to the Holtzmark microfield. At high density Coulomb interactions affect the microfield distribution. For practical calculations, electronic collisions with the radiator can be viewed as random fast events compared with the lifetime of the radiator and, since the energy transfers are small, the line shape can be treated by perturbation theory. For helium-like ions the principal quantum number degeneracy is removed for deeply bound states, so that ion perturbations occur through the quadratic Stark effect. Electronic contributions are more important in this situation.

The theory of multielectron atoms and ions is much less complete, and simple methods are generally used to approximate $\phi_{\rm if}$. For radiators with more than two bound electrons, electrons are the major contributor to pressure broadening and the contribution of ions is generally ignored in opacity calculations. The effect of the much faster-moving electrons is usually treated in the impact approximation which assumes that only one electron at a time collides with the radiator, instantaneously cutting off its wave train. This approximation also gives a Lorentzian line shape. Assuming that the different broadening effects are independent, it is possible to convolve the Lorentzian for natural broadening with the Lorentzian for impact broadening to obtain a new Lorentzian whose half-width at half-maximum (HWHM) is just the sum of the HWHMs for natural and pressure broadening. Convolution of this Lorentzian with a Gaussian to account for Doppler broadening then gives a composite shape known as a Voigt profile. Recent line width measurements of $\Delta n = 0$ transitions in boron and carbon ions indicate that the Voigt profiles may provide only a rough approximation. Fortunately, stellar opacity is not very sensitive to moderate changes in line shapes, but this is a source of uncertainty in the calculations.

Bound–free absorption

$$P_i + h\nu \rightarrow P_f^+ + e.$$

Bound–free transitions are possible when $h\nu$ is greater than the ionization potential, I, of a bound electron. The cross-section for this process is

$$\sigma_{\rm i}^{\rm bf}(\nu) = \frac{4\pi^2}{3 g_{\rm i} c}(I + \varepsilon) \left| \int \Psi_\varepsilon^*(\boldsymbol{r}) \boldsymbol{D} \Psi_{\rm i}(\boldsymbol{r})\, {\rm d}\boldsymbol{r} \right|^2 \qquad (4)$$

where the ejected electron has the energy $\varepsilon = h\nu - I$, and $\Psi_\varepsilon(\boldsymbol{r})$ is the scattering wavefunction of the ejected electron. The total cross-section for a given frequency interval is obtained by summing over all initial and final states. Equation (4) assumes that the photoionization edge is sharp. However, because of the merging of high-lying lines just below the continuum, the edge is broadened. Simple procedures, that redistribute the oscillator strength in the vicinity of the edge, are generally introduced to approximate this type of broadening.

For complex ions, bound states can overlap in energy with the continuum. If the Hamitonian connects a bound and a continuum state, the eigenfunction is mixed, introducing resonances into the bound–free continuum, as was first recognized by Fano. This has not been found to have a significant effect on stellar opacity and the autoionizing lines are frequently treated as ordinary broadened lines.

Free–free (inverse Bremsstrahlung) absorption

$$P + e_i + h\nu \rightarrow P + e_f.$$

For free–free electronic transitions the initial state density is also continuous and its distribution is given by Fermi statistics. Dipole-allowed transitions can only occur in the presence of ions, otherwise energy and momentum cannot be simultaneously conserved. The free–free cross-section

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

can be calculated from the dipole acceleration formulation in the form

$$\sigma^{\mathrm{ff}}(\nu) = \frac{e^4}{3\pi\nu^3 m^2 c} \sum_{l_\mathrm{i}} \sum_{l_\mathrm{f}} \max(l_\mathrm{i}, l_\mathrm{f})\, \delta(l_\mathrm{i} - l_\mathrm{f} \pm 1) \times \left| \frac{\mathrm{d}V}{\mathrm{d}r} \right|^2_{\mathrm{if}} p(\varepsilon_\mathrm{i})[1 - p(\varepsilon_\mathrm{f})] \quad (5)$$

where l_i and l_f are the angular momenta of the initial and final states, $|\mathrm{d}V/\mathrm{d}r|_{\mathrm{if}}$ is the matrix element of the radial derivative of the atomic potential for continuum states i and f, $p(\varepsilon_\mathrm{i})$ is the occupation probability of the initial state, and $1 - p(\varepsilon_\mathrm{f})$ is the availability of the final state. Relativistic corrections to σ^{ff} can be significant when $\varepsilon_\mathrm{f} > 10$ keV. Relativity increases the solar center opacity by about 0.2%. The inverse bremsstrahlung cross-section is often normalized to the classical Kramers cross-section. This ratio is known as the free–free Gaunt factor. Inverse electron–electron quadrupole bremsstrahlung can also contribute to the opacity at high temperature and low density.

Photon scattering

$$\mathrm{e_i} + h\nu \rightarrow \mathrm{e_f} + h\nu' \qquad \text{free electron}$$

$$\mathrm{P} + h\nu \rightarrow \mathrm{P} + h\nu' \qquad \text{bound electron (low energy).}$$

Photons scatter mainly off of the plasma electrons, primarily free electrons at high temperature and bound electrons at low temperature. Scattering from nuclei is much less important. For most stellar applications photon scattering can be considered to be coherent ($\nu = \nu'$), but non-coherent scattering ($\nu \neq \nu'$) in which the incoming and outgoing photon have different ν can also occur through high-energy Compton scattering from free electrons or by Raman scattering in the molecular regime.

The classical picture for scattering of a photon from a bound electron demonstrates the connection between the two most important scattering mechanisms; Rayleigh and Thomson scattering. The classical scattering cross-section ($h\nu < 0.511$ MeV) has the form

$$\sigma_\mathrm{S}(\nu) = \sigma_\mathrm{T} \frac{\nu^4}{\nu^2 - \nu_{\mathrm{if}}^2 + (\nu_{\mathrm{ij}}\gamma)^2} \quad (6)$$

where

$$\gamma = e^2 \nu_{\mathrm{if}}^2 / 2\pi^2 m c^3 \quad (7)$$

and

$$\sigma_\mathrm{T} = \frac{8\pi}{3} \left(\frac{e^2}{m_\mathrm{e} c^2} \right)^2 \quad (8)$$

is the Thomson cross-section. Rayleigh scattering occurs when $\nu \ll \nu_{\mathrm{if}}$, so that $\sigma_\mathrm{S} \propto \nu^4$. In the other extreme ($\nu \gg \nu_{\mathrm{if}}$), where the energy of the incoming photon is so high that the scattering electron acts nearly as a free particle, σ_S approaches the Thomson free electron cross-section. Thomson scattering is low-energy Compton scattering. When the energy of the scattered photon is in the keV range, as occurs in high-temperature stellar matter, relativistic effects modify the Compton cross-section and pair production can occur. Sampson expressed these corrections for the total scattering cross-section in the form of the multiplicative factor $(n_{\mathrm{ee^+}}/n_\mathrm{e})G(u, T')$, where $n_{\mathrm{ee^+}}$ is the total number of electrons and positrons, n_e is the number of electrons, $u = h\nu/kT$ and $T' = kT/mc^2$. Additional corrections are required if the electrons are degenerate.

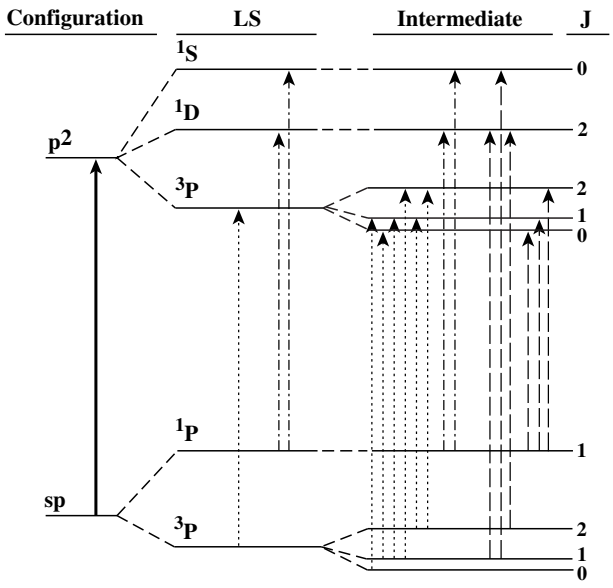


Figure 1. Schematic diagram of possible absorptive transitions between the sp and p^2 configurations in various approximations. Configurational (DCA) transition (solid arrow); singlet transition (short-dashed arrow); triplet transition (short-long-dashed arrow); intercombination transition (long-dashed arrow). J is the total angular momentum of the state.

Atomic data

Several possible levels of detail in the atomic data needed to calculate the energy levels and oscillator strengths are illustrated in figure 1. The example is for transitions that connect two electrons in an sp configuration to a p^2 configuration through a one-electron jump. In the simplest case all transitions are assumed to have the same energy. Opacity calculations that carry out a sum over a large number of configurations in this approximation are referred to as detailed configuration accounting (DCA) methods. This was the method used in the earliest bound–bound calculations. However, the DCA approach neglects non-spherical interactions that remove the degeneracy and lead to configuration term structure. In light elements the dominant non-spherical term is the Coulomb interaction between the electrons, which leads to pure LS coupling. In LS coupling the single spectral line of the DCA method in figure 1 is split into three distinct lines corresponding to a triplet and two singlet terms. With heavier elements, such as those in the iron group, the interaction between the electron spin and the magnetic field resulting from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the electron orbital motion is no longer negligible and requires coupling of the LS angular momentum states, known as intermediate coupling. In intermediate coupling the three lines of the LS coupling scheme split into eight lines having no net total spin change, $\Delta S = 0$, and six intercombination lines having $\Delta S = \pm 1$. For more complicated configurations the increase in the number of spectral lines is more dramatic. For heavy elements this results in an enormous array of spectral lines that can be treated by statistical methods when the density is high enough that these features merge into a quasi-continuum.

The most accurate atomic data are obtained from relativistic self-consistent field methods such as multiconfiguration Dirac–Fock, but this is very computer intensive when data for millions of transitions are required. The most extensive multiconfiguration atomic data tables have been produced by the Opacity Project using the close-coupling method. These calculations are mostly non-relativistic.

Fortunately, the important thing for opacity calculations is the distribution of oscillator strength. The exact location of lines is only of secondary importance. Consequently single-configuration calculations are usually used. Parameterized potentials have proven to work well for generating the vast amount of data required with accuracy similar to single-configuration Hartree–Fock.

The effect of elemental composition on stellar properties

Stars are composed mainly of hydrogen and helium whose frequency-dependent absorption spectrum is characterized by strong lines separated by regions of weak absorption. Because low-abundance heavy ions have a more complex absorption spectrum, they contribute disproportionately to κ_R by filling in regions of low absorption in the H–He spectrum.

The total mass of all elements heavier than helium generally constitutes less than 2% of the mass of a star (see STELLAR MASSES). This percentage is known as its 'metallicity'. Even so, low-abundance elements such as nickel can significantly affect the opacity when the temperature is in the range where the M shell is partially filled. Because type I stars like the Sun are not old enough that significant burning of elements heavier than helium has occurred, it is normally assumed that these abundances pertain throughout the entire star. Consequently, for most purposes, it is possible to characterize stellar mixtures with just the three variables X, Y and Z, where X is the hydrogen mass fraction, Y is the helium mass fraction and Z is the heavy element mass fraction, i.e. the metallicity. Within this constraint Z can vary, but the abundances of heavy elements relative to each other are fixed. Even without nuclear burning, however, these relative values can vary throughout the star owing to diffusion.

Figure 2 compares the monochromatic opacity for a hydrogen–helium mixture with a typical type I stellar composition having $Z = 0.02$. The assumed temperature–density conditions are found in the pulsational driving region of Cepheid variables. The small heavy element admixture dramatically increases the absorption for $u > 3$, corresponding to photon energies above 80 eV; this results primarily from transitions originating in the M shells of multiply ionized ions of iron. As a result, even though hydrogen and helium are the most abundant species they are not the main contributors to the opacity at these plasma conditions. They do, nevertheless, play an important role by contributing most of the free electrons that determine the rate of recombination of the heavier elements. These electrons also contribute to electron scattering and free–free absorption.

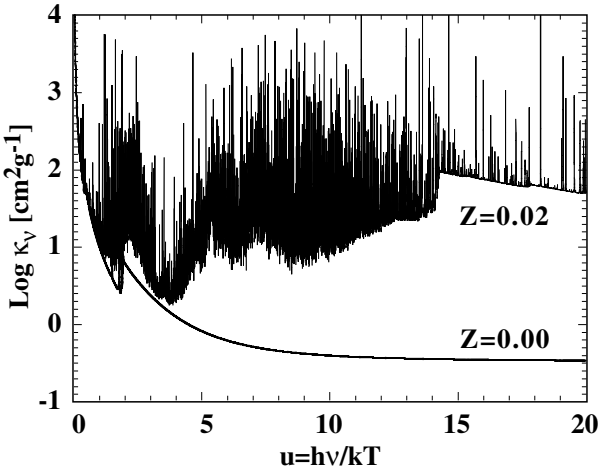


Figure 2. Log κ_ν versus $u = h\nu/kT$ for $T = 316\,000$ K and $\rho = 1 \times 10^{-5}$ g cm^{-3}. Results are shown for $Z = 0$ and $Z = 0.02$ with $X = 0.7$ and $Y = 1 - X - Z$. The heavy element calculations were carried out in intermediate coupling.

Figure 3 compares the opacity versus temperature for the Sun and a Cepheid variable with a mass 6.5 times solar. Because of their mass difference and location on the HERTZSPRUNG–RUSSELL DIAGRAM these stars follow very different temperature–density tracks. The opacity of the solar plasma when heavy elements are left out of the mixture is also shown. Again heavy elements dominate the opacity throughout most of the stellar interior. For stars of one solar mass or greater core temperatures can exceed 10 million kelvins, so that the plasma is highly ionized and even heavy elements such as iron may have only a few bound electrons. In this situation the main contributors to opacity are bound–free absorption in heavy elements, free–free absorption from protons and He nuclei and photon scattering from free electrons. For example, at the solar center the opacity is 40% bound–free, 30% free–free and 30% scattering. In contrast, near the solar surface the degree of ionization is very low and, since the strongest bound–bound transition energies are large compared with kT, atoms make a very small contribution to opacity. In this case the opacity of H^-, which has a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

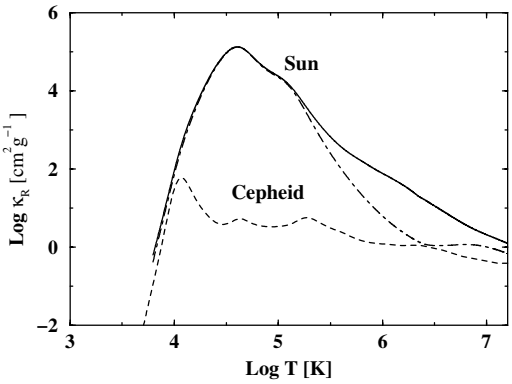


Figure 3. Log κ_R versus log$T$ for the Sun (solid) and a Cepheid variable (dashed). The assumed heavy element content is $Z = 0.02$. For these stellar tracks, X is around 0.7 for $T < 4 \times 10^6$ K, but is less at higher temperature due to nuclear burning. The upper temperature limit (1.56×10^7 K) corresponds to the solar center. Cepheid core temperatures are about an order of magnitude greater. Also shown is the opacity for the temperature–density conditions along the solar track assuming $Z = 0$ (dot-dashed).

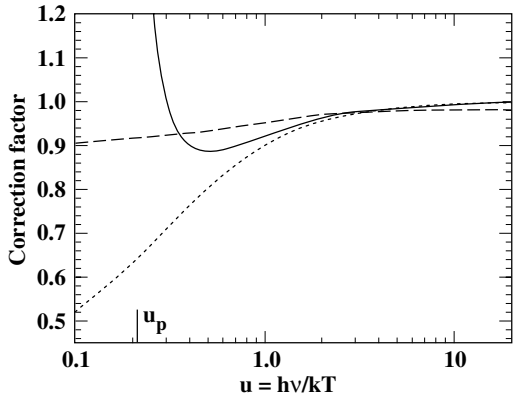


Figure 4. Plasma dispersion, screening plus ion correlational, and degeneracy corrections to inverse bremsstrahlung expressed as ratios $g_{\rm eff}(\nu)/g_0(\nu)$, where $g_{\rm eff}$ is the Gaunt factor including one of these effects and g_0 is the Gaunt factor without these effects included. Plasma dispersion (solid); screening plus ion correlations (short-dashed); degeneracy (long-dashed). The location of the plasma frequency is indicated by $u_{\rm p}$.

binding energy of only 0.74 eV, is a major contributor to opacity, even though its abundance is generally less than 10^{-7}. Rayleigh scattering from H and He is also important near the surface where all atoms are nearly neutral. In the intermediate regions, the opacity is relatively featureless. The maximum near $T = 40\,000$ K is in the middle of the H–He ionization zone and is due mainly to bound–bound absorption.

The Cepheid opacity function in figure 3 is more complex than the corresponding solar opacity function, displaying two additional local maxima. For the Cepheid the H–He maximum separates into a maximum near $T = 10\,000$ K due to H ionization and a maximum near $T = 40\,000$ K due to H–He ionization. The third maximum, near $T = 200\,000$ K, is due mainly to bound–bound transitions occurring in M shell iron. For this reason it is known as the 'Z bump'. The presence of the local maximum near $T = 40\,000$ K is what drives the pulsational instability observed for Cepheids, while the Z bump plays a secondary role by affecting the period ratios between the fundamental mode and its overtones. The solar opacity is much higher at a given temperature because, owing to the higher density, the bound states contributing to the bound–bound and bound–free opacity are more highly occupied.

Plasma environmental effects

The opacity cross-sections discussed above assume that photon extinction processes occur in isolation and ignore the effects due to the surroundings. As the density increases these effects can modify the cross-sections but are generally small for stellar plasmas. However, for some situations, such as the center of the Sun, relatively small changes in the opacity affect quantities such as the neutrino generation rate through small changes in the estimated temperature (see also SOLAR INTERIOR: EQUATION OF STATE AND OPACITY). Consequently, there has been theoretical interest in the effect of collective phenomena on opacity. Collective effects due to plasma density oscillations become important when the plasma frequency,

$$\nu_{\rm p} = \frac{N_{\rm e} e^2}{\pi m_{\rm e}}$$

is comparable with the mean photon frequency.

Dispersion

In dispersive media the Planck function in equation (3) is modified by the factor n_ν^2 so that the Rosseland mean opacity is given by

$$\frac{1}{\kappa_{\rm R}} = \int_{\nu_{\rm p}}^{\infty} {\rm d}\nu \frac{n_\nu^2}{\kappa_\nu} \frac{{\rm d}B_\nu}{{\rm d}T} \Bigg/ \int_0^{\infty} {\rm d}\nu \frac{{\rm d}B_\nu}{{\rm d}T} \qquad (9)$$

where n_ν is the refractive index, which has the approximate form ($\nu > \nu_{\rm p}$)

$$n_\nu \cong [1 - (\nu_{\rm p}/\nu)^2]^{1/2}.$$

Collective effects also contribute to κ_ν and should be included (see figure 4).

Multiple scattering from electrons

When the average photon wavelength is comparable with the electron–electron correlation length coherent scattering from the electron distribution can occur. When $\nu \gg \nu_{\rm p}$ the Thompson cross-section (equation (8)) becomes

$$\sigma(\nu) = \frac{3}{8} \int_{-1}^{1} {\rm d}\mu (1 - \mu)(1 + \mu^2) S_{\rm e}(q)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$$q = \frac{\nu}{2\pi c}[2(1-\mu)]^{1/2}$$

where $\mu = \cos\vartheta$ and $S_e(q)$ is the electron static structure factor. Correlated scattering from the electron distribution reduces the Thompson cross-section by 27% at the solar center.

Plasma screening and ion correlations

Plasma screening and ion correlational effects in electron–proton bremsstrahlung (equation (5)) can be described through the free–free Gaunt factor. In the (non-relativistic) Born approximation the Gaunt factor can be expressed in the form

$$g_{cor}(\omega) = \frac{\sqrt{3}}{16\pi^3 e^4}\int_0^\infty \frac{dq}{q}\exp\left[-\frac{\hbar^2}{2mkT}\left(\frac{q}{2}-\frac{m\omega}{\hbar q}\right)\right] \times \left|\frac{q^2 V(q)}{\varepsilon(q,\omega)}\right|^2 S_p(q) \quad (10)$$

where $\omega = 2\pi\nu$ is the angular frequency, $V(q)$ is the Fourier transform of the Coulomb potential, $S_p(q)$ is the proton static structure factor and $\varepsilon(q,\omega)$ is the plasma dielectric response function. Similar corrections affect other cross-sections when the density becomes high enough but are not considered in current calculations. In the low-density limit $S_p(q)$ and $\varepsilon(q,\omega) \to 1$, so that equation (10) reduces to the usual Coulomb Gaunt factor.

The effects of plasma dispersion, plasma screening plus ion correlations, and electron degeneracy on electron–proton inverse bremsstrahlung in the solar center are shown in figure 4. Additional dispersion effects appear explicitly in the Rosseland integral (equation (6)). The magnitude of each effect is normalized to the Gaunt factor without these effects included. The corrections are small near $u = 4$ where the largest contributions to the Rosseland integral occur. The reductions in κ_R due to plasma dispersion, screening plus ion correlations, and electron degeneracy are, respectively, 0.4%, 0.7% and 0.8%. Effects of this size are not generally significant, but they are large enough to be of interest in helioseismic and neutrino flux calculations.

Bibliography

Adelman S J and Wiese W L (ed) 1995 *Astrophysical Applications of Powerful New Databases (ASP Conf. Ser. vol 79)*

Alexander D R and Ferguson J W 1994 Low-temperature Rosseland opacities *Astrophys. J.* **437** 879–91

Iglesias C A and Rogers F J 1996 Updated OPAL opacities *Astrophys. J.* **464** 943

Huebner W F 1986 Atomic and radiative processes in the solar interior *Physics of the Sun* vol 1 (Dordrecht: Reidel) pp 33–74

Seaton M 1995 *The Opacity Project* (Bristol: Institute of Physics Publishing)

Forrest Rogers and Carlos Iglesias

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Photometry

Photometry is undoubtedly one of the most fundamental of all astronomical observations. It is hard to imagine any astronomical research of any nature that does not use photometric results in some way or another. The measurement of brightness and of broad-band spectral features in a star or other object's flux output is at the heart of much of what is done in getting the essential data needed for our research problems. All astronomers do photometry or use the results of someone else's observations.

Photometry and spectroscopy are very complementary, as are photometry and theory. In fact, 'complementarity' is one of the most operative words relative to approaches and to progress in our understanding the universe about us.

However, while basically very simple, photometry is very often much misunderstood. To do good research in astronomy, it goes without question that we need accurate and precise photometry of stars and galaxies.

What we measure in photometry is really an integral of the effect of many things: the flux output of the star or other object, including any shells or disks about it (and whatever else nature may have put there to confuse us), interstellar matter, the Earth's atmosphere, the telescope, the filters, the detectors and other things as well. Most of these are changing with time. We are trying to estimate something, perhaps the stellar apparent brightness or some strong feature in its spectrum. We are trying to avoid the adverse impact of other things that may well confuse our understanding. The selection of our observing site, our telescope, our detector and our filters can all impact upon the success of our estimate of the parameter of interest. Photometric systems have usually been designed to be efficient estimators of what we are trying to understand and to avoid the adverse impact of those things that are most likely to cause problems. We try to chose the photometric 'system' that will do the best job. Simple in principle, but often difficult in practice.

Having a good understanding of what is going on is critical. There can and often are adverse effects on our measurements due to stellar rotation, shape of the star, emission, double or multiple stars, line shape, spots, flares, shells, disks, the interstellar matter and its uniformity, variable stellar or emission background, the faintness of our object, atmospheric seeing and variability, clouds, filter leaks and cut-off characteristics, temperature effects and anything involved, detector changes, and so on. With all of this, it seems impossible to think of doing precise not to speak of accurate photometry. In addition, there are issues such as separation of the observed parameters, sensitivity to the feature we are trying to measure, the ease of measurement and reduction and analysis, standard stars and calibrations, reduction methods, information resolution, and so on. How can one manage? It has been done, but not as often as one would like, or even as one thinks.

Precision and accuracy are key issues in photometry, no matter how the photometry is done. This article will discuss some of these photometric issues after laying out the basics of what astronomical photometry is.

A few basics

In astronomy, one is attempting to measure the radiant flux (energy per unit time) coming from an astronomical object, such as a star or a galaxy, generally as a function of wavelength. Such flux measurements can be used in many ways. These applications are beyond the scope of this article on stellar photometry, but in general one can determine the distance to a star, its chemical composition and any peculiarities, its age, whether it varies and how and many other things. Similar information can be obtained about galaxies, and indeed about the interstellar matter in the Galaxy, whether the dark absorbing matter or the bright emission material in nebula. Essentially all methods of determining distance rely on the measure of the object's apparent brightness as we measure it.

If the radiant energy (in joules) is Q, then the radiant flux (in watts), F, is $\mathrm{d}Q/\mathrm{d}t$, the radiant flux at a surface is $\mathrm{d}F/\mathrm{d}A$ (in W M^{-2}), the radiant intensity, I, is $\mathrm{d}F/\mathrm{d}\omega$ (ω is the solid angle through which the flux is radiated, the unit of I is W sr^{-1}, where sr is steradian), and the radiance, E, is $\mathrm{d}F^2/\mathrm{d}\omega\,\mathrm{d}A\cos\theta$, θ being the angle between the line of sight and the normal to the surface considered. Similar terms and units exist for light, rather than the total energy output, and then the unit of luminous flux is the lumen and the luminous flux at a surface is called the illuminance with the unit lux. The unit of luminous intensity is the candela (cd) and of luminance cd m^{-2}. Astronomical photometry, at least in the optical wavelength region, uses its own terminology, the magnitude, discussed below.

One is generally measuring the flux over a range of wavelengths so as to get color discrimination or other such information rather than total brightness. The Earth's atmosphere limits the range of wavelengths that we can observe from the ground from about 320 nm to somewhat beyond 1000 nm.

In astronomy, we measure the brightness in a unit called magnitudes. This unit has its origin in antiquity, when the naked-eye stars were classified by brightness into six groups, called magnitudes. Brightness here means the visual sensation obtained by looking at the star. The brightest stars were called first magnitude and the faintest sixth magnitude. Today we know that the brightest stars are somewhat brighter than that and the best eyes can see fainter than that, when there is no light pollution (man-made sky glow) present. The eye's response to brightness is not linear, but logarithmic. Hence the magnitude scale is logarithmic. In 1856, Pogson defined the modern scale of magnitudes to be such that the ratio between two magnitudes was the fifth root of 100, which is about 2.512. This scale fitted well the ancient scale of magnitudes of stars that had been used for millennia (see also MAGNITUDE SCALES AND PHOTOMETRIC SYSTEMS).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Hence, a source with an observed flux density of F has a magnitude defined by $m = -2.5\log(F/F_0)$, where F_0 is a standard flux setting the zero point of the scale. It is chosen so as to fit the ancient magnitude scale zero point.

These magnitudes are called apparent magnitudes, for they are what we observe at our telescopes (see below for a discussion of other factors involved, however). The Sun has an apparent magnitude of −26.7, the full Moon of −12.7, Sirius, the brightest star, of −1.5, and the approximate faint limit of the largest telescopes or the HUBBLE SPACE TELESCOPE is about +28.

The intrinsic brightness of the star is called the absolute magnitude, M, and M is defined as the brightness of the star at a distance of 10 parsecs (the parsec is unit of distance in astronomy). $M = m - 5\log(r/10)$. We can see that if we measure m and can in some way determine M (such as from the stellar characteristics we estimate in some other way, often from photometric measurements), we can determine r, the star's distance. In fact, $m - M = 5\log(r/10)$ is called the distance modulus, a logarithmic way of expressing distance to the star (or galaxy). The distance in these expressions is in parsecs, of course.

While the use of magnitudes in astronomy is the standard (for OPTICAL ASTRONOMY, RADIOASTRONOMY uses other more physical units for flux), it has several disadvantages. It is a logarithmic scale, and it is not directly related to any physical units. Note also that a larger value (more positive) of a magnitude means a fainter flux. It also has several advantages. The main one is that it makes it easy to compare magnitude differences between two stars.

The magnitude difference between two stars is $m_1 - m_2 = -2.5\log(F_1/F_2)$. Hence a ratio of 1.001 is a magnitude difference of 0.001, 1.01 of 0.010, 2 of 0.75 (a double-star system in which both components are of equal brightness will look 0.75 magnitudes brighter when both components are observed together, for example), 2.512 of 1.000, 6.3 of 2, 10 of 2.5, 16 of 3, 40 of 4, 100 of 5, 1000 of 7.5 and so on. Note that a ratio of 10^{22} implies a magnitude difference 55, the total range that can be observed in astronomy, as noted above.

Doing photometry

It is possible to consider photometry as the application of the following equation:

$$I(\lambda, t) = c \int (I_s - I_b) IMATPFDO \, \mathrm{d}\lambda$$

where $I((\lambda, t)$ is the flux we observe and c is a normalizing constant; λ is wavelength and t is time. The other terms are discussed one at a time below.

I_s is the item that we are usually trying to determine, the flux from the star (or other object) that we are observing. However, sometimes we are just using a star as a probe to determine something else, such as atmospheric extinction or interstellar reddening values in the direction of observation.

Most of the terms in the above equation are wavelength dependent, vary with time, location in the field of view and with other factors such as temperature, humidity, telescope position, site location and altitude and so forth. Identifying, calibrating and otherwise correcting for all these multitude of adverse effects is critical. The pitfalls for systematic and accidental errors are huge. Errors in any one term can really cause problems. Both precision and accuracy require that these effects all be understood, calibrated and corrected adequately. That takes time and effort, and often such time is not available. Without doing the job right, we are lucky to get the 'right' answer.

$I_s(\lambda, t, x, y, \ldots)$ is the flux emitted by the star we are observing. Here x and y are positions on the sky. This flux can and does include the continuous and line emission and absorption lines, bands, breaks, the effects of any circumstellar material, the effects of duplicity (if any), including the interstellar material around and between the stars, and any and all other effects giving rise to flux originating at or in the vicinity of the star or other object. The star's image is not a point source, and adverse effects can arise from not allowing correctly for the point spread function, including the wings of the image. The image may also be contaminated by adjacent stars, duplicity, or in a crowded field.

$I_b(\lambda, t, x, y, \ldots)$ is the sky background, for which corrections to our observed I_s must be made before we have the 'pure' I_s value. There are many ways to do this, for both photomultiplier and for CCD photometry. Suffice to say here that this correction process is one that can lead to lack of precision and accuracy. In addition to the 'process' itself, we have the problems of variable background (coming from a variable background in space, due to emission regions, absorption regions, crowding of stars in the field, etc) and from effects arising in the telescope (vignetting, etc), from variable sensitivity over the detector, from variable transmission in the filters, etc. For fainter stars, the sky background can even dominate the signal, and that too leads to a decrease in precision and accuracy.

$IM(\lambda, x, y, \ldots)$ is the impact of interstellar absorption (or interstellar reddening, as the absorption is wavelength dependent). Such absorption is likely to present in most observations and must be taken into account. If the reddening law varies in different regions of the sky, as it undoubtedly does, and if the derived corrections for this effect are not well handled, then errors creep in easily. Systematic effects are quite likely to arise.

$A(\lambda, t, x, y, \mathrm{HA}, \mathrm{DEC}, \ldots)$ is the effect of atmospheric extinction, also wavelength dependent. HA is the hour angle and DEC the declination of the observation. Clearly, the effect is also time dependent, over many time-scales. Its determination is not always straightforward. Stebbins is quoted as saying that to determine extinction well, you must spend all your observing time leaving nothing for your program. Probably true. Besides the effects of air mass differences, there are temperature and humidity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

impacts, inversion level altitude, east–west and north–south effects (and towards–away from a city or mine or whatever), jet trails, light clouds, heavy clouds and so on, all of which change on varying time schedules.

$T(\lambda, t, x, y, \ldots)$ reflects the various things that the telescope and its optics can do to you, including the time and wavelength dependence of the reflective coatings on the optics. There are other effects as well, such as incorrect baffling, vignetting, ghost images, dust, and so on.

$P(\lambda, t, x, y, \ldots)$ are any effects produced by the photometer and its elements, optical or otherwise.

$F(\lambda, t, x, y, \ldots)$ is the impact of the filter (or other wavelength discriminator) on the throughput. It is our choice of FILTERS (or spectrometer) that gives us our information resolution. The careful choice of filters is critical to success, for many reasons. However, as in so many other items noted here, filters are wavelength dependent in themselves, and that dependence changes with time, temperature, humidity, dust coverage and other perverse reasons. One must be aware of all of this, and allow for it if necessary. It is necessary. The detailed transmission curve of the filters must be known, and allowed for. Central wavelength, shape, bandwidth, tails, are all important items.

$D(\lambda, t, x, y, \ldots)$ is the impact of the detector. Problems also can and do exist with photomultiplier, CCD, and other types of photometry.

$O(\lambda, t, x, y, \ldots)$ is the impact of 'all other'. There are always some of these. Identifying and allowing for them is the key. Here we might well include any effects produced by differing reduction procedures.

Problems in doing photometry

With all these potential and real problems, it is a wonder that one can do photometry at all. The pitfalls are so many. The time needed to investigate all of them is necessarily large, and not often available. Factors change with time as well, often on a rather fast time schedule, or discontinuously. Is there no hope? Probably there is. After all, good photometry has been done. But there is no way it can occur without careful attention to all the details, not just to getting the data and using a neat reduction package (see SOFTWARE: STELLAR PHOTOMETRY). There is a lot more to it.

A real understanding of what is going on is difficult or impossible. It is critical to understand and allow for all the issues involved.

Standard photometric systems

It is hard to imagine doing much in astronomy without a measure of brightness or color or other photometric parameters. To be able to do effective research programs, it is essential to be able to compare the photometric data used with other such data. Hence, we all use standard photometric systems, such as UBV, or $uvby\beta$, or Geneva, or Vilnius standard systems.

The section above discussed the factors involved in $I(\lambda, t)$. A photometric system is one that uses these factors, particularly the term F (filter), to define one or more wavelength regions to measure. These are chosen to give the observer a number of parameters that can be related to stellar (or other) characteristics, such as brightness in a limited wavelength range, 'color', chemical composition, temperature, effects of the intervening interstellar material on the star's flux, and so on. Quite a number of such photometric systems have been defined and used over the years. Only a few will be mentioned here.

Visual magnitude

The magnitude as 'measured' or seen by the eye. The eye has a certain spectral sensitivity, and that defines the wavelength region of the magnitude system. Since eyes vary, from individual to individual and with time, and indeed even with the level of incoming flux, this is not a precisely defined system. But it has been and still is a most useful one. It is how bright we see the stars. It has also been defined in a more precise way, using filters and detectors other than the eye, and in a way so as to match the estimates that have been made by observers using only the eye (often supplemented by a telescope).

Photographic magnitude

With the advent of photographic plates, there of course came definitions of a photographic magnitude, brightnesses as measured with a photographic plate. Even more than with the eye, photographic emulsions vary in wavelength sensitivity, so accuracy often suffered. Most emulsions were more blue sensitive than the eye, so the combination of the two magnitudes could give a measure of 'color', the ratio of brightnesses at two different wavelengths.

UBV system

The advent of electronic detectors revolutionized photometry. They allowed much more precise photometry, and with much greater sensitivity than the eye or photographic plates. The wavelength sensitivity also became better, and we now have magnitude systems ranging from the ultraviolet (indeed, even for shorter wavelengths, as observed from above the Earth's atmosphere) to the far infrared. The initial systems were patterned after the visual and photographic ones, but later ones broadened the range, adding more filters and more wavelengths (see also DETECTORS: PHOTOGRAPHY). There were many, but the most used one came into widespread use soon after the introduction of the photomultiplier tube. It uses three bandpasses, one matching the visual magnitudes, one in the blue roughly matching the old photographic magnitude and an additional one in the ultraviolet. They are called, naturally, V, B and U, and the system is the UBV system.

It is a natural system in two ways. Firstly, the band passes fill the region of sensitivity of the photomultiplier tubes used, from about 320 nm to 600 nm, dividing this region into three magnitudes, the U, B and V. Each bandpass is about 100 nm wide. Second, these bandpasses offer a great deal of information about the star's (or other object's) flux output. V gives a visual magnitude that

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

tells us the star's apparent magnitude. B gives another band and the two define a 'color index', B–V, that can be used to estimate the star's temperature (and other things sometimes), U and B also give a color index, U–B, that offer another parameter useful to estimate stellar characteristics, such as chemical composition, interstellar reddening, and absolute magnitude. We have three parameters that can be used for our research, V, B–V and U–B. Naturally, it is difficult to estimate more than three stellar characteristics with only three parameters, but sometime it can be done. All such applications are beyond the scope of this article. The point to be remembered here is well sketched out above, in the discussions of pitfalls. In later years, with the advent of detectors sensitive in the red region of the spectrum, additional wide-bandpass magnitude systems were developed to supplement the UBV system, in particular R and I magnitudes, for red and infrared regions.

uvby system, or Strömgren system

A natural extension of the UBV system was to fill the region of sensitivity of the photomultiplier tube with four band passes instead of the three of the UBV system. Strömgren proposed such a system and it has had widespread usage. It offers additional 'information resolution', and overcomes several of the observational problems of the UBV and other such wide-band systems. It is an intermediate bandpass system, with each filter being about 20 to 30 nm wide. The y magnitude can be related very well to the V magnitude. The color index $b - y$ gives a measure of temperature. One defines two 'color differences'. The first, $m_1 = (v - b) - (b - y)$, gives an estimate of the 'blanketing' due to metallic absorption lines in the region of the v band. The second, $c_1 = (u-v)-(v-b)$, gives a measure of the increased absorption in the ultraviolet, due in most stars to the continuous absorption of the hydrogen. This can be related to absolute magnitude in some stars, and to other characteristics as well. With four parameters, one can usually do a better job of estimating key characteristics of the star's energy output. Again, it is beyond the scope of this article to discuss how to use these photometric parameters for research. Sufficient to say that they are widely used, and are one of the main keys to our understanding of stellar structure and evolution, and to the geography of our galaxy and of the universe.

Narrow band systems

It is possible to develop additional systems, of any band width. Narrow band systems can be used to estimate the strengths of absorption or emission lines or of other features in the spectrum. One, for example, measures the strength of the hydrogen Hβ line, a parameter related to temperature in some stars and to absolute magnitude in others. It gives another parameter to supplement either the UBV system or the $uvby$ system. Many other such standard systems have been developed, and a few are widely used.

Systems with CCD detectors

The recent advent of CCD (CHARGE COUPLED DEVICE) detectors has been a big jump forward. Their efficiency is very good, and they are areal detectors as opposed to the single-element photomultiplier. Hence, one can work fainter for a given precision, and do it over a area not just for a single object. They combine the best of the photographic plate and the photomultiplier. They have many advantages therefore, but also some disadvantages. They are much more sensitive than either the photographic plate or the photomultiplier tube. One can save the electronic image for later processing. They respond to light linearly, like the photomultiplier tube. They have relatively low background noise. But the cost is high, of both hardware and analysis tools. The field size is smaller than the photographic plate (though increasing, especially through the use of mosaic detector arrays (but at high cost)), and processing and analysing the images is very time complex and time consuming. As in all photometry, great care must be taken to ensure precision and accuracy.

Bolometric magnitude and absolute intensity measurement

Two other types of photometric systems should be mentioned. The first is the bolometric magnitude. This is the magnitude measured over all wavelengths. It cannot be measured directly, due to the absorption in the Earth's atmosphere and due to the difficulty of having a detector with such a range of sensitivity. Its value is that it is an estimate of the total flux output of a star. The second is the measurement of absolute intensities, rather than relative or differential intensities. Clearly, all the systems discussed above are differential, relative to some standard stars or zero point. It is of value to measure also, though very difficult, the absolute intensity in physical units. With such a measurement for a few well chosen stars, the other systems can then be put on an absolute scale. Such a transformation is essential for comparisons with theoretical calculations of stellar flux, for example.

Standard stars

It is not always clear to the user how such standard systems have been defined or are to be matched, or even how they are to be used. Is it by the careful choice of filters and detector that match the system? Perhaps, but that can be very difficult if not impossible. We don't use the detectors that were used to establish the system originally, for example. Is it by careful choice and observation of a subset of the standard stars that were used to help establish the system? Sometimes, but sometimes not. Sometimes not for a good reason: they are all too bright for our current telescope and detector system, or there are no standards of the stellar type we need. Sometimes not for a bad reason: not everyone has the concept of interpolation well in mind, or we don't have the time to do an adequate job.

In some sense, viable photometric systems are rather like the MK spectral classification system: they depend a great deal on the usage of standard stars. We can (and should) do the best we can in the choice of spectral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

wavelengths and bandwidths for the filters and in our choice of detectors and reduction procedures, but the final quality and value of the system will depend a lot on the quality of the standard stars used to help define the system and on the choice of standards used for any observing program with the system. Such standards must span the range of stellar types that will be observed with the system and on the specific observing program. Naturally, the standards must be observed and tabulated with high precision and high accuracy. Many examples exist in the literature of problems that arise when these obvious guidelines are not followed. Considerable effort should be made by all photometrists to ensure that an adequate set of standard stars exist for the most used standard photometric systems.

One must select an adequate number of standards of the needed characteristics to ensure a good job of interpolation in all phases of the observation and reduction process. This may not be easy to accomplish, especially if we are observing either a wide range of 'unknowns' or types of objects for which few if any standards exist. In the first case, a rule may be 'the more the better'. In the second case, one may have to try to develop substandards of the type of characteristics needed.

In the observing itself, one must observe so as to interpolate. This means that one must be observing standards at somewhat higher air masses (and lower) than any of the unknowns. One must start and end the night with standards, and intermingle them throughout the night. Particularly on 'all sky' programs, it is helpful if one can randomize the intermixing of standards of different characteristics, so that no pattern is built in that may later cause systematic problems.

Careful and adequate time spent planning the observing program is essential, and will always pay off in increased accuracy. So will adequate time spent observing standard stars, even in the face of time pressures (not enough nights assigned to the program, equipment problems, clouds, and learning curve time). Good equipment (stable and well understood) is essential as well. And a most valuable asset is a quality observing site.

Here are some of the problems facing us in the future:

(a) Getting adequate observing time to do our programs, and to observe standards (and other calibrations). The excitement of the field and the increasing number of astronomers (in spite of the funding problems we all face) coupled with the fact that some of the telescopes useful for photometry are actually being closed down means that the pressures on time will only increase.
(b) Hence, the time available to investigate standard systems and to continue our efforts to establish more standards (especially ones of additional characteristics, including fainter ones of all types) will be sadly lacking.
(c) Equipment is getting more powerful but more complicated. We are getting separated even further from the sky than we have been in the past. It will be harder to understand the content and the performance of our black boxes, both hardware and software ones.
(d) The drive to work fainter and to observe strange and wonderful objects means that the extrapolation aspects of photometry may be more in evidence than good astronomy requires. Even photometrists get off the track; imagine what others will be doing.

The future

What about the future? If all goes well, and it seldom does, we might see:

(a) New technology, detectors and instrumentation, new telescopes, including new generation small ones, an essential component in a balanced funding approach to astronomy. We need it all, and we will be getting some of it without question.
(b) Better understanding of photometry and photometric systems.
(c) More and better data. And adequate archiving of the data.
(d) More and better standard stars. More tie-ins.
(e) A wider range of standard stars, in character and faintness.
(f) Better astronomy, better science. Precise and accurate photometry is fundamental to it all.

Summary

Photometry has always been of the highest value in astronomical research. It always will be. Such photometry must be of high precision and accuracy. The key is to understand the basic principles of photometry, and of the potential pitfalls involved. The existence of well chosen standard systems and standard stars, both for their applicability to the research job and for their intrinsic merit, is one of the most powerful tools to ensure that reality.

Bibliography

Budding E 1993 *Introduction to Astronomical Photometry* (Cambridge: Cambridge University Press)
Craine E R, Tucker R A and Barnes J 1999 *CCD Precision Photometry Workshop, Astronomical Society of the Pacific Series Vol 189* (San Francisco)
Crawford D L 1988 Photometry: a new and important technique *Publ. Astron. Soc. Pacific* **100** 887
Golay M 1975 *Introduction to Astronomical Photometry* (Dordrecht: Reidel)
Howell S 1992 *Astronomical CCD Observing and Reduction Techniques, Astronomical Society of the Pacific Vol 23* (San Francisco)
Johnson H L and Morgan W W 1953 *Astrophys. J.* **117** 313
Sterken C and Manfroid J 1992 *Astronomical Photometry* (Dordrecht: Reidel)
Straizys V 1991 *Multicolor Stellar Photometry* (Tucson, AZ: Pachart)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Strömgren B 1966 *Ann. Rev. Astron. Astrophys.* **4** 433

David L Crawford

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Populations

A stellar population is a set of stars of the same age and chemical composition. Born together at some point in the Galaxy, stars of the same stellar population will share the same kinematic properties in the Galaxy, whether those be their circular velocity with respect to the Galactic center or their random velocity in the Galactic halo.

Examples of such simple stellar populations are the stars of the ORION population (Herbig and Rao 1972) (figure 1), the nearby OB ASSOCIATIONS (de Zeeuw *et al* 1999) and the GLOBULAR CLUSTER 47 Tuc (figure 2). We can collect all the simple stellar populations of roughly the Sun's age and call them the old disk population (Eggen *et al* 1974), or all the simple stellar populations of the central predominantly old regions of the Galaxy and refer to them as the bulge population or the more diffuse collection of high velocity stars in random motion in the Galaxy's potential and refer to them as the halo population. It is easy to see that a stellar population is a very broad collective term for a sample of stars with what can be, but need not be, well-defined statistical kinematic, age and chemical composition properties.

The linkages between old stellar populations were originally made by Baade (1944). The prior history of the subject and its later development are recounted by Sandage (1986), who notes 'The main thrust and the ultimate aim of the work in stellar populations today is to achieve an understanding of the formation and development of galaxies with time'.

In the broader extragalactic context the concept of a simple stellar population is immensely useful, and nowhere more so than in population synthesis, the process of constructing a model of a galaxy from its constituent stellar populations (see SYNTHETIC STELLAR POPULATIONS). In population synthesis a galaxy is the sum of its stellar populations of a particular age, weighted by the star formation rate (SFR) at that epoch. For example, one can model a Magellanic irregular galaxy as the sum of MAGELLANIC CLOUD star clusters of every age, t, from a million years to ten billion years, weighted by SFR(t). Of course, a galaxy will also take a particular path through its chemical enrichment history, $Z(t)$ (where Z is the heavy element abundance by mass). Therefore, population synthesis must also deal with this complication. Other complications are conceivable too, such as a changing initial mass function. The formalism for dealing with simple stellar populations in this way is elegantly developed by Renzini and Buzzoni (1984).

Star formation history

In principle the color–magnitude diagram of a stellar population bears a full record of the star formation history (SFH) of that population. There has been considerable progress in recent years in extracting the desired information from the data. The theoretical underpinnings are described by Renzini and Buzzoni (1984).

Figure 1. The Orion Nebula, our closest sample of a young stellar population, courtesy M Bessell, RSAA, ANU.

Figure 2. The globular cluster 47 Tuc, courtesy AAO.

The main sequence

When stars are formed, they are laid down on a long-lived hydrogen-burning main sequence (MS), following what approximates, but is unlikely to be exactly, a universal initial mass function (a distribution function). The stars then evolve away, starting with the most massive ones, with lifetimes which are a (weak) function of metallicity. The present-day mass function would therefore be invertible to a SFH, SFR(t), if $Z(t)$ were also known, at least approximately. In practice, the observable is a luminosity function (LF), rather than a mass function, but the mass–luminosity relation makes the LF a good substitute, unless unresolved binary stars distort the LF seriously (see LUMINOSITY FUNCTION OF GALAXIES). The solar neighborhood (Twarog 1980), the Large Magellanic Cloud

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 3. The LMC, courtesy M Bessell, RSAA, ANU.

(LMC) (figure 3) and the Galactic bulge have been good laboratories for this approach to learning the SFH. To zeroth order the solar neighborhood is a constant-SFR population, the LMC is characterized by a recent burst and the bulge shows a prominent initial burst. Thus a simple range of possibilities has now been explored.

The giant branch

Because of the mediating effect of their envelopes, giant branches are sensitive to chemical composition and are relatively insensitive to age (see also CHEMICAL COMPOSITIONS OF STARS). The stellar populations of early-type galaxies are dominated by their light. Nevertheless, post-MS stages of stellar evolution do provide tracers of the SFH in stellar populations in an indicative manner, if not in the more rigorous way that the MS offers. Often this is all that is available.

Useful tracers of young, intermediate-age and old stellar populations are as follows:

Age (yr)	10^8	10^9	10^{10}
Tracer	tip of MS	upper AGB	tip of RGB

The red giant branch (RGB) tip is where the helium flash ignites in low-mass stars at around 2000 solar luminosities (M_I −4 mag; Ferrarese *et al* 2000). The asymptotic giant branch (AGB) is the second ascent of the giant branch with a characteristic bolometric luminosity of −5.5 mag for 1 Gyr old stars. In populations more metal poor than the Sun, upper AGB stars are CARBON STARS (Mould and Aaronson 1986, Frogel *et al* 1990) and readily recognizable.

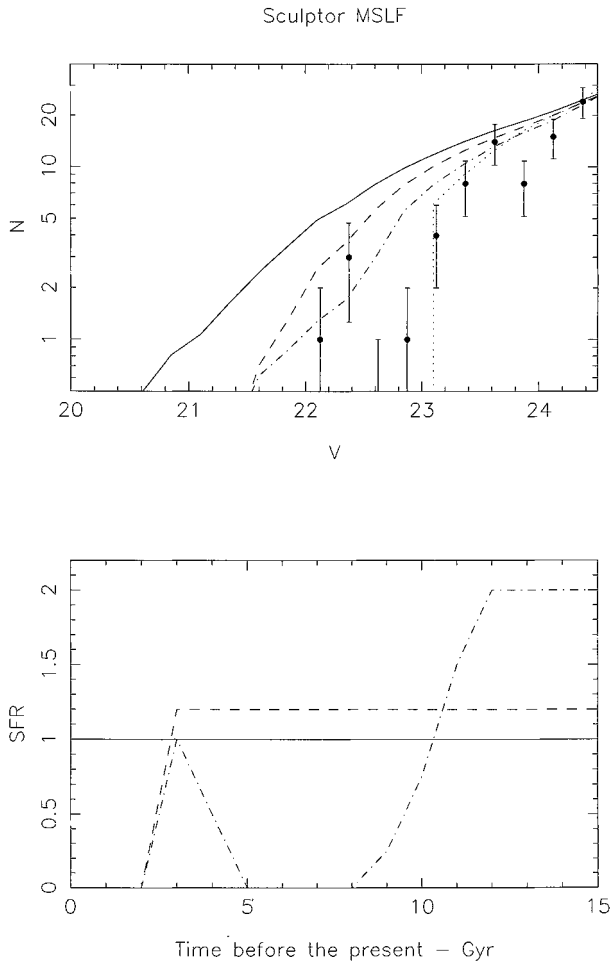


Figure 4. MS luminosity function for the field studied by Monkiewicz *et al* (1999). We can fit the Sculptor MS with a variety of SFHs, as outlined in the lower part of the figure. The dotted curve is a simple stellar population of age 15 Gyr, i.e. a delta function in the lower plot.

Determining SFH

These tracers can suggest a coarse SFH for resolved galaxies (e.g. LGS 3, Mould 1996) with the aid of Renzini's fuel burning theorem:

$$n = B(t)Lt$$

where L is the luminosity of the system, t is the lifetime of the tracer population and n is the number of tracer stars. B is a slowly varying function of age.

A coarse SFH can be graphically illustrated as in figure 4. Hodge (1989) calls these diagrams 'population boxes'. Population boxes for LOCAL GROUP DWARF GALAXIES have been compiled by Mateo (1998). He finds (1) that no two of these galaxies have the same SFH, (2) that many dwarf irregular galaxies contain significant old populations (e.g. IC 1613, NGC 6822), (3) that most recent star formation episodes are short (10–500 million

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

years in duration), (4) that red horizontal branches are common in metal-poor systems (the 'second parameter effect') throughout the Local Group, (5) that there are no purely old galaxies (with the possible exception of the Ursa Minor dwarf spheroidal) and (6) that intermediate-age galaxies exist (e.g. M32 and Leo I).

For accurate SFH determination synthesis of the color–magnitude diagram can be done. Full population synthesis is described by Tolstoy (1995). A method of maximum likelihood multi-isochrone fitting has been developed by Han, Hoessel and Gallagher (private communication).

Bibliography

Baade W 1944 *Astrophys. J.* **100** 137
de Zeeuw P T *et al* 1999 *Astron. J.* **117** 354
Eggen O J, Freeman K C and Rodgers A W 1974 *Rep. Prog. Phys.* **36** 625
Ferrarese L *et al* 2000 *Astrophys. J.* **529** 745
Frogel J *et al* 1990 *Astrophys. J.* **352** 96
Herbig G and Rao N 1972 *Astrophys. J.* **174** 401
Hodge P 1989 *Ann. Rev. Astron. Astrophys.* **27** 139
Holtzman J *et al* 1996 *Astrophys. J.* **466** 732
Mateo M 1998 *Ann. Rev. Astron. Astrophys.* **36** 435
Monkiewicz J 1999 *Publ. Astron. Soc. Pac.* **111** 1392
Mould J 1996 *Publ. Astron. Soc. Pac.* **109** 125
Mould J and Aaronson A 1986 *Astrophys. J.* **303** 10
Renzini A and Buzzoni A 1984 *Spectral Evolution of Galaxies* ed C Chiosi and A Renzini (Dordrecht: Kluwer)
Sandage A 1986 *Ann. Rev. Astron. Astrophys.* **24** 421
Tolstoy E 1995 *PhD Thesis* Groningen University
Twarog B A 1980 *Astrophys. J.* **242** 242

Jeremy Mould

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Pulsation

Pulsating stars are a subtype of variable stars that vibrate, and these vibrations cause their brightness to vary. The brightness varies because of the changing area and temperature of the surface layers (see VARIABLE STARS). The vibrations are due to regulation of the outward radiant energy flow. Pulsating stars are ubiquitous in that they can be found almost everywhere in the Hertzsprung–Russell (H–R) diagram, although conspicuous pulsations are present in more limited portions of the H–R diagram, as shown in figure 1. In addition to their intrinsic interest, pulsating stars have played an important role in advancing our understanding of physics and the universe.

Fabricius discovered the first repeating variable star in 1596, and it was later called 'Mira', which means 'wonderful', because of its changing brightness. By coincidence, Mira is a pulsating star and the prototype of the MIRA VARIABLES. More important was the discovery of periodic light variations of δ Cephei in 1784. This star is now the prototype of the pulsating yellow supergiant stars called Cepheids. Cepheids are historically the most important class of variable stars, because they follow a CEPHEID PERIOD–LUMINOSITY RELATION, first discovered by Henrietta Leavitt in 1912. The period–luminosity relation reveals that Cepheids with longer periods are intrinsically more luminous (have brighter absolute magnitudes), based on Leavitt's study of a large number of Cepheids in the Magellanic Clouds. Later, E Hubble discovered Cepheids in the Andromeda galaxy, and the period–luminosity relation enabled him to determine the distance, placing it well outside our own Milky Way. This application makes Cepheids vitally important 'yardsticks' for measuring the distances to nearby galaxies (see UNIVERSAL DISTANCE SCALE). One of the motivations for building the Hubble Space Telescope was to detect Cepheids in the VIRGO CLUSTER of galaxies to provide a reliable distance calibration for use in determining the size of our universe. There is a growing impression that most, if not all, stars pulsate—if measured carefully enough—and the maturing field of astroseismology takes advantage of the fact that each independent pulsation mode offers a probe of the star's interior. Stellar astronomers are now in a position to determine the interior structure of stars directly, and to a level of detail unforeseen a generation ago.

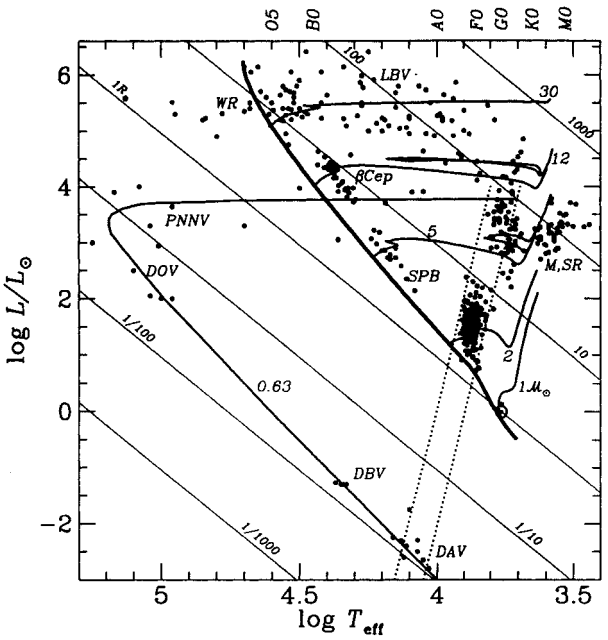


Figure 1. A 'pulsating' Hertzsprung–Russell diagram with the location of important variable star classes listed. Representative evolutionary tracks are also plotted with the stellar mass (in solar units) indicated near each track. Slanted lines are labelled with the star's radius (in solar radii). Pulsating star classes labelled here include: luminous blue variables (LBVs), β Cephei stars (βCep), planetary nebula nucleus variables (PNNVs), pulsating PG 1159 stars (DOV), slowly pulsating B stars (SPB), Mira and red giant pulsators (M,SR), and pulsating DB and DA white dwarfs (DBV and DAV).

A brief history of stellar pulsation

After the discovery of Mira in 1596 and the establishment of its periodicity by about 1660, the pace of discovery of variable stars was slow. Only about 20 variable stars were known by the 1840s; about this time, F W A Argelander and J Herschel put the search for variable stars on a more systematic footing. The advent of astronomical photography, and in particular, the Harvard Observatory photographic patrols, begun in about 1890, provided a permanent record of stellar brightnesses and a systematic coverage of the sky. The Magellanic Clouds and rich star fields of the southern Milky Way added many new variable stars, so that thousands were known by 1912, when H Leavitt discovered the Cepheid period–luminosity relation.

While the discovery of variable stars accelerated in the 1860s, Lord Kelvin laid the theoretical framework for non-radial oscillations of liquid spheres in 1863. However, it was not until 1879 that A Ritter suggested radial pulsations, coupled with temperature variations, as an explanation for the light variations of some periodic variable stars. In spite of this, it was not until 1914, when H Shapley proposed oscillations of the star's surface (pulsations) as being responsible for Cepheid variability, that pulsations were seriously considered. By 1918, A S Eddington laid the mathematical foundations for stellar pulsation theory, and consequently directed much of the theoretical work towards considering radial pulsations as the explanation for stellar pulsations for the next 50 years or so.

Although radial pulsation theory replicated many observational properties of Cepheids, the source of pulsation driving was not determined until the 1950s, when J P Cox and S A Zhevakin independently proposed that cyclical partial ionization of He^+ to He^{2+} was responsible for the observed pulsations. Subsequent linear

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and nonlinear calculations confirmed this hypothesis. Digital computers started to become available by the 1960s, and with these people were able to calculate pulsation periods from the linearized pulsation equations as well as nonlinear hydrodynamic simulations for increasingly realistic stellar models. By the 1970s and early 1980s people once again began to consider non-radial pulsations, in response to observations of stars that pulsated with too many modes in a given frequency range to be explained only by radial modes. Also, observations of the variable white dwarfs and the β Cephei stars revealed pulsation periods and radial velocity curves that could only be explained by non-radial oscillations.

In the last ten years, the advent of ever more powerful computers and CCD cameras is making theoretical calculations and high-precision observations of pulsating stars increasingly accessible (see CHARGE-COUPLED DEVICE). Observers also took a cue from helioseismology and started to look for stars that pulsate in many modes to perform astroseismology on them. This necessitated the creation of multilongitude multisite observing networks that can continuously observe a target star for days at a time in order to reduce spectral leakage 'aliasing' in Fourier transforms caused by data gaps. White dwarfs and δ Scuti stars are leading the way in terms of astroseismological potential and multilongitude observations. Finally, CCD survey projects designed to look for gravitational MICROLENSING are turning up large numbers of newly discovered pulsating stars, often without previous biases or selection effects that mar a statistical understanding of instability strip boundaries and/or the incidence of pulsating stars within a given instability strip. The present era of precision instrumentation and abundant computational resources is ushering in a new understanding of the internal structure of pulsating stars and a clearer picture of why they pulsate, which we describe below.

Some aspects of the theory of stellar pulsation

Cepheid variable stars obey a period–luminosity relation to considerable accuracy, and this has a rather simple physical explanation. First, we know that Cepheids are radial pulsators (their motion is strictly in and out), and we can approximate the period of a radial mode as the time it takes for a sound wave to travel from the surface to the center and return

$$\Pi = \frac{2R_\star}{c_s}$$

where c_s is the average sound speed and Π is the pulsation period. Elementary physics texts show that the adiabatic sound speed is $c_s^2 = P\Gamma_1/\rho$, and we can use hydrostatic equilibrium to determine that

$$\Pi \approx (G\bar{\rho})^{-1/2}$$

where $\bar{\rho}$ is the mean density of the star and $\Pi(= 2\pi/\sigma)$ is the pulsation period. The latter expression is called the *period–mean density relation*. Most Cepheids have spectral type between F5 and K0, corresponding to effective temperatures between 7000 and 5000 K. The fairly narrow temperature range occupied by Cepheids implies that the more luminous Cepheids are brighter mainly by virtue of being larger. The most luminous Cepheids have radii about 10 times larger than the least luminous ones, but the most luminous Cepheids are only about four times as massive. This implies that the mean density of a luminous Cepheid is about 1/250 that of a less luminous Cepheid and thus their periods should be about 20 times longer. This agrees reasonably well with the observed period range of Cepheids in our galaxy, which lies between about 2 and 100 days.

Radial pulsations turn out to be a special subset of the more general non-radial pressure modes (also called p-modes), where pressure is the restoring force. In everyday life, these are the sound waves we hear when talking or playing the radio. In this scheme, radial modes have a spherical harmonic index (ℓ) of zero, while p-modes in general can have any positive value of ℓ, and although their motion is predominantly radial, there is some horizontal motion as well. Several features of p-modes are particularly useful in the analysis of stellar pulsation observations and are derivable from the expression for p-mode frequencies in the limit of very high overtone number (the number of times the eigenfunction crosses zero). The general expression for the frequency ($\sigma_{n,\ell}$) is

$$\sigma_{n,\ell} \approx 2\pi\nu_0(n + \ell/2 + \epsilon)$$

where

$$\nu_0 \equiv \left(2\int_0^R \frac{1}{c_s}\,\mathrm{d}r\right)^{-1}$$

n is the *overtone number* and ϵ is a small constant that depends on the structure of the star. There are four useful facts we can derive from these expressions. First, for a given radial overtone number, the frequency of a p-mode increases with ℓ. Second, the p-mode frequency also increases with n. Third, in the asymptotic limit of very high overtone number, the frequency spacing for consecutive overtone p-modes of a given ℓ is constant; this works quite well in general for all but the lowest overtones. Finally, the p-mode frequencies for modes of different ℓ follow the trend set by

$$\sigma_{n,l} \approx \sigma_{n-1,\ell+2}.$$

Non-radially pulsating stars can also pulsate in g-modes, where buoyancy rather than pressure, is the restoring force. The motions tend to be more horizontal than radial, so that material sort of 'sloshes' back and forth. Buoyancy is what causes a cork to bob up and down in water after it has been pushed down, and the natural frequency of oscillation is expressed in terms of the Brunt–Väisälä frequency

$$N^2 = -g\left(\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}r} - \frac{1}{\Gamma_1 P}\frac{\mathrm{d}P}{\mathrm{d}r}\right)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where N is the Brunt–Väisälä freqency, g is the local gravitational acceleration, and Γ_1 is one of the adiabatic exponents, given by

$$\left(\frac{\mathrm{d}\ln P}{\mathrm{d}\ln \rho}\right)_{\mathrm{ad}}.$$

An approximate expression for the pulsation frequencies of g-modes is

$$\sigma_{n,\ell} \approx \frac{[\ell(\ell+1)]^{1/2}}{n\pi}\int_{r_b}^{r_a}\frac{N}{r}\,\mathrm{d}r$$

where r_a and r_b are the upper and lower boundaries of the g-mode propagation region. Like the p-modes, g-modes also have useful limiting behavior in the limit of high overtone number. First, the period increases with increasing overtone number n, and the period spacing between consecutive overtone modes of a given ℓ is constant, to a good approximation. Finally, the periods of g-modes with different ℓ, but the same overtone number are related by

$$\sqrt{\frac{\ell_1(\ell_1+1)}{\ell_2(\ell_2+1)}}$$

so that the period of the nth overtone $\ell = 1$ mode is about $\sqrt{3}$ times longer than the analogous $\ell = 2$ mode, and so on.

The previous discussion treats the existence of pulsations as a given, whereas in reality stars must somehow excite their own pulsations. The vast majority of pulsating stars act as a heat engine, and in their outer layers, ionization and recombination of a dominant ion species acts to store and release the outward flow of energy on a periodic basis. Normally, an increase in radiant energy flowing outwards in a star will increase the temperature, decrease the opacity (a measure of resistance to radiant energy flow), and allow the radiation to escape more easily. Thus, under normal conditions, such as in fully ionized material, the change in opacity favors pulsational *stability*. In a partial ionization zone, however, an increase in radiation flow can be absorbed by the ions, resulting in an increase in the fraction of ions present, rather than increasing the temperature of the material. As a result, the radiant energy flow gets dammed up and stored as ionization energy; this mechanism is called the γ mechanism. In addition, partial ionization zones also produce strong opacity gradients, and the region can drive pulsations if

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\kappa_T + \frac{\kappa_\rho}{\Gamma_3 - 1}\right) > 0$$

where

$$\Gamma_3 - 1 \equiv \left(\frac{\partial \ln T}{\partial \ln \rho}\right)_{ad} \qquad \kappa_T \equiv \left(\frac{\partial \ln \kappa}{\partial \ln T}\right)_{\rho}$$

$$\text{and} \qquad \kappa_\rho \equiv \left(\frac{\partial \ln \kappa}{\partial \ln \rho}\right)_{T}.$$

Regions satisfying the previous inequality are unstable to the opacity (or κ) mechanism, and the radiant energy blocked by this opacity increase is then converted into kinetic energy of the oscillations. In most cases of interest, the κ and γ mechanisms work together. Figure 2 shows a diagram of a typical pulsation cycle.

While the κ and γ mechanisms are capable of modulating the outward energy flow from a star, certain conditions must be met before the star can modulate the energy flow effectively and become a pulsating star that we can observe. First, the driving region needs to make the structure of a star 'overshoot' the equilibrium state so that the pulsations have a chance to grow in amplitude. For this to happen, the driving region must still be gaining heat at the instant of maximum compression. Then, when the star starts to expand, the heat release adds an extra 'push' to the outer layers, causing the star to expand past the equilibrium state. An analogy would be a swinging, frictionless pendulum; one must apply an extra nudge after the highest point of the swing to add to the kinetic energy of the pendulum in order to make it swing in a wider arc. The second condition is that the energy flow must be modulated with a period similar to a natural eigenmode of the star; in other words, the thermal timescale of the driving region needs to be roughly equal to the pulsation timescale, which may be expressed by

$$\frac{C_v T (\Delta m)_{\mathrm{tr}}}{L} \approx \Pi$$

where Δm_{tr} is the mass of the star from the surface inwards to the 'transition region', C_v is the heat capacity per unit volume and L is the stellar luminosity. C_v, T and L are time-averaged values at the 'transition region'. In typical stars, the 'transition region' is quite near the surface; typically at about $10^{-8}M_\star$ to $10^{-10}M_\star$ inwards from the photosphere. For an He^+ to He^{2+} region, the temperature is about 50 to 100 000 K.

Classes of pulsating stars

Pulsating stars can be found in many parts of the Hertzsprung–Russell diagram as shown in figure 1, and opacity mechanisms are responsible for essentially all of the observed pulsations (see HERTZSPRUNG–RUSSELL DIAGRAM). He^+ to He^{2+} ionization is responsible for the pulsations of the best-known pulsators, such as the Cepheids, RR Lyrae and δ Scuti stars, which occupy a sloping band in the H–R diagram. Several new classes of variable stars and increasing numbers of stars in previously known classes have been discovered over the last 20 years, mainly due to increasingly sensitive observational techniques. Given the apparent ubiquity of stellar pulsations across the H–R diagram, there is some reason to suspect that possibly *all* stars pulsate, and the only question is 'at what amplitude?'. In the next paragraphs, we describe most of the major classes of pulsating stars and their astrophysical context (see table 1).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

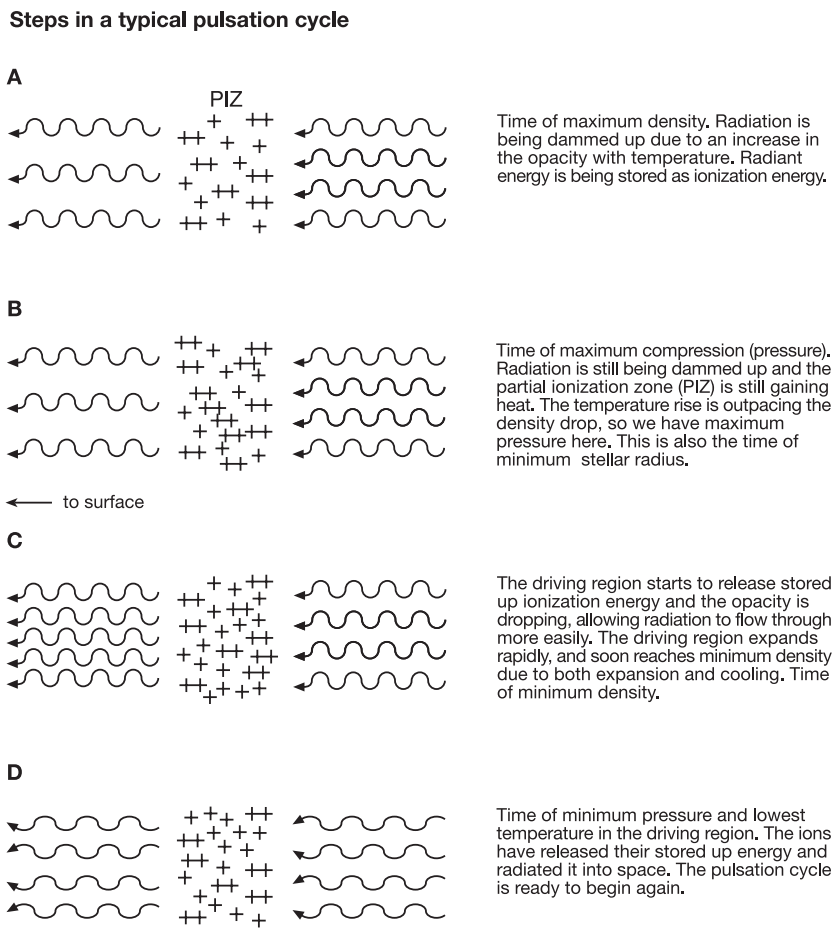


Figure 2. A schematic diagram of the driving region in a star pulsating via the κ mechanism. Panel A: As the star compresses, outward radiation flow gets blocked due to an increase in opacity. Radiation energy is stored in the ions and kinetic energy. There is relatively little increase in temperature. For reference, we will call this the time of maximum density. Panel B: This panel shows the point of maximum compression (pressure). Radiation is still dammed up and the region is still gaining heat, but the density is dropping due to increasing temperature. Panel C: The driving region is expanding and is getting an extra push due to the release of stored up ionization energy. The opacity is dropping as the material expands, which helps the radiation flow out more easily. This is the time of minimum density, which is the result of both material expansion and temperature decrease. Panel D: At maximum expansion, the material is still losing heat and reaches its minimum pressure. All of the stored up energy is released by the ions (but the material is still partially ionized), and the pulsation cycle is set to start again.

Cepheid variables

Cepheid variables were among the first pulsating stars to be discovered, and their well defined period–luminosity relation, coupled with their high luminosity, makes them valuable 'standard candles' for measuring distances to other galaxies out to about 20 Mpc. The prototype star is δ Cephei, and all of these stars pulsate in one (or two) radial modes (typically the radial fundamental mode) that have very precise periods that range between 2 and 100 days. Cepheids are yellow (F to G) supergiants with masses between 5 and $20M_{\odot}$ and luminosities between 1000 and $100\,000L_{\odot}$. Most are core helium burning objects, although some could be burning hydrogen in a shell around the helium core as they evolve to become red giants for the first time. The typical Cepheid light (or radial velocity) curve has a steep rise to maximum amplitude that takes about 0.2Π and a more gradual decline, but the exact shape of the curve depends on the period, as first noticed by E Hertzsprung in 1926 and illustrated in figure 3. For periods less than 6.5 days, the curves appear smooth, but between 6.5 and 9 days, the curves show a bump on the descending part of the curve that moves towards maximum amplitude as the period becomes longer. Between 10 and ~15 days, the bump then appears on the ascending part of the curve. The theoretical explanation for this phenomenon did not appear until the 1980s, when J R Buchler and collaborators demonstrated that the bump is the result of a 2:1 resonance between the radial fundamental and the radial second overtone mode. In addition to computing theoretical light curves that compare well with observations when the resonance effects are included, the theory also replicates the observed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Parameters and nomenclature for pulsating star classes.

Type	Acronym	T_{eff} ($\times 10^3$ K)	$M/M_\odot$	$L/L_\odot$	Period	Mode type†
Cepheids	δ Cep	5.5 to 6.5	5 to 20	10^3 to 10^5	2 to 100 d	R
Mira	M or Mira	2.5 to 4.0	$\sim$ 1 to 2	10^2 to 10^4	25 to 400 d	R
Red giant branch or asymptotic giant branch	SRa, SRb, Lb, and Lc	2.5 to 4.0	$\sim$ 1 to 8	10^2 to 10^4	25 to 400 d	R
RR Lyrae	RR Lyr	6.0 to 7.0	0.5 to 0.8	40 to 50	0.2 to 1.2 d	R
δ Scuti	δ Sct	7.5 to 8.7	1.6 to 2.6	10 to 100	0.5 to 6 h	R and NR(p)
Rapidly oscillating Ap	roAp	7.0 to 8.5	1.6 to 2.6	10 to 50	4 to 15 m	NR(p)
γ Doradus	γ Dor	6.5 to 7.3	1.4 to 1.7	5 to 20	0.5 to 3 d	NR(g)
β Cephei	β Cep	22 to 30	8 to 15	4000 to 20000	3.5 to 6 h	NR(p)
Slowly pulsating B	SPB (53 Per)	13 to 17	200 to 2000	2.5 to 6	0.6 to 3 d	NR(g)
Line profile Variables	LPVs	15 to 40	10^3 to 10^5	5 to 50	3 to 30 h	NR(p)
Pulsating PG1159	DOV	$\sim$ 100	$\sim$ 0.6	$\sim 10^2$	300 to 1200 s	NR(g)
DB white dwarf Variables	DBV	$\sim$ 23	$\sim$ 0.6	$\sim$ 0.03	400 to 1200 s	NR(g)
DA white dwarf	DAV	$\sim$ 11	$\sim$ 0.6	$\sim 10^{-3}$	100 to 1200 s	NR(g)
Luminous blue Variables	LBV	$\sim$ 20	30 to 100	10^5 to 10^6	$\sim$ 30 to 50 d	NR ?
R Corona Borealis	R CBr	5.0 to 7.0	$\sim$ 1	$\sim 10^4$	$\sim$ 40 d	R?
Subdwarf B stars	EC14026	29 to 35	$\sim$ 0.5	20 to 40	100 to 400 s	NR(p)

†Mode types: R stands for radial pulsations, while NR stands for non-radial pulsations. (p) stands for p- or pressure modes, while (g) stands for g- or gravity modes.

properties of the light curves when they are decomposed into a Fourier expansion of sine waves.

W Virginis stars

The *classical* Cepheids are Population I objects, whereas W Virginis stars are low-mass ($\sim 1 M_\odot$) analogs that are evolving towards the asymptotic giant branch (AGB) or are engaged in thermal pulses off the AGB. They are also called 'population II Cepheids' and/or 'type II Cepheids'. Like the Cepheids, the W Vir stars also follow a period–luminosity relation, but it is about 1 to 2 magnitudes fainter, and it has a shallower slope, as shown in figure 4. The W Virginis stars have periods between $\sim$1 and 30 days, and then they smoothly merge into RV Tauri stars that have light curves with alternating deep and shallow minima, along with more or less periodic behavior that is attributed to low-dimensional chaos. BL Herculis and a few other short-period type II Cepheids with periods of about a day are sometimes considered as a subgroup of the W Virginis stars; they have a bump on the declining part of the light curve about a quarter period after maximum light.

Red giant pulsars

Most red giant branch (RGB) and AGB stars pulsate with periods ranging from about 25 days to over a year (see RED GIANT STARS). The periods are usually somewhat irregular, although a fairly accurate mean period can be derived for most of these stars. In visible wavelengths, the pulsation amplitudes can vary from a few hundredths of a magnitude up to 7 or 8 magnitudes. Objects with extreme amplitudes in the visible appear so for two reasons. First, the Johnson V filter covers the same wavelength range as the most prominent TiO absorption bands, and the strength of these absorption bands increases rapidly with decreasing temperature. Secondly, the star becomes cooler as it becomes fainter, and the energy output shifts towards the infrared where our eyes cannot see. In most cases, the bolometric amplitude is about 1 magnitude. The pulsations are probably driven by partial ionization of H and He, but the dominance of convection in the outer layers makes theoretical pulsation calculations difficult due to pulsation–convection coupling that is difficult to model. These modeling difficulties also contribute to the ongoing debate about whether Mira stars are pulsating in the radial fundamental or first overtone mode. Many variable star classes are known in this part of the H–R diagram. The most famous are the Mira variables, which have visual amplitudes greater than 2.5 magnitudes. Other classes include the semiregular variables (SRa and SRb) and the irregular variables (Lb and Lc); these stars tend to have smaller light curve amplitudes ($\sim$1 magnitude (mag) in the visual) than the Mira stars. Of these, the Lc stars are supergiant stars, and by implication are more massive than the other RGB and AGB variable stars. As the name implies, the semiregular and irregular variables have more erratic light curves than the Mira stars. One of the relatively unexplored frontiers is connecting the pulsations with mass loss. All of the RGB and AGB stars have extended, low-gravity envelopes, and pulsations cause appreciable mass loss during the time an RGB or AGB star spends in this part of the H–R diagram. It also appears that pulsation induces the final episode of mass loss as a star leaves the asymptotic giant branch to become a planetary nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

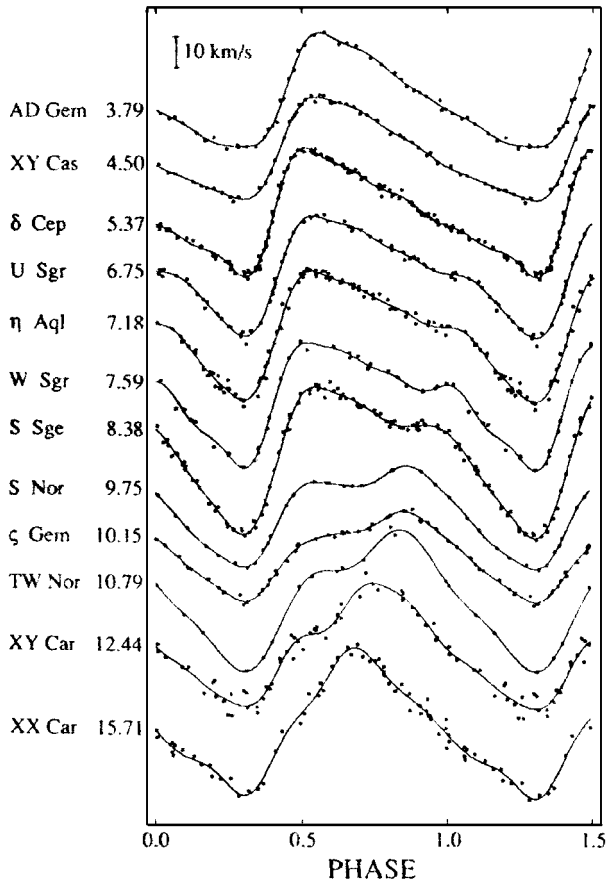


Figure 3. A sample of Cepheid variable star radial velocity curves. The interval between maximum and minimum radial velocity (or brightness in a light curve) is called the amplitude (A); the time interval between successive maxima is the period (Π), and the fractional period between one maximum and the next is called the phase (ϕ). Note the Hertzsprung 'bump' progression between 6 and 15 days.

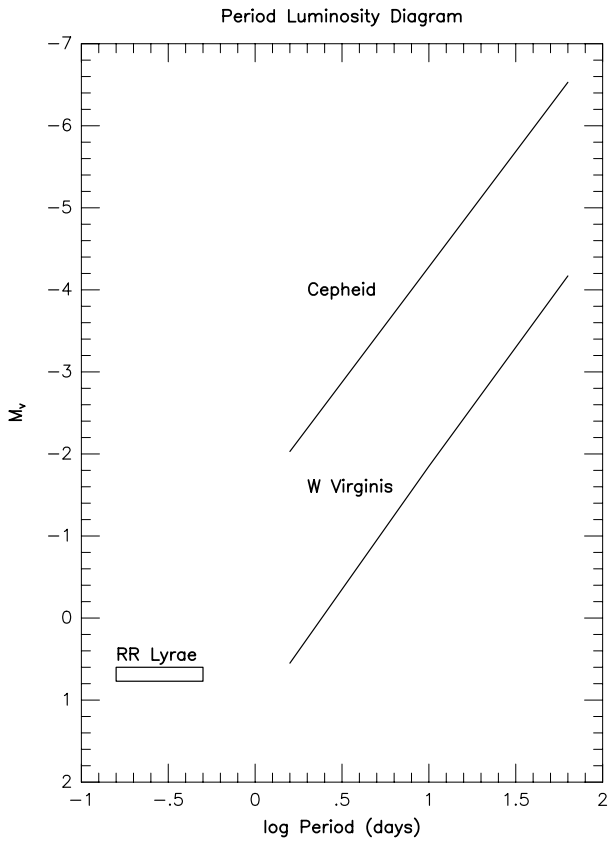


Figure 4. Period–luminosity relations for RR Lyrae stars, 'classical' Cepheids and W Virginis stars.

RR Lyrae stars

RR LYRAE STARS are also radial mode pulsators, but are population II stars that are burning helium into carbon and oxygen in their cores. The typical effective temperature for an RR Lyrae star is about 6500 ± 500 K, and they have a luminosity of about 40 to 50$L_\odot$. The periods of RR Lyrae stars are typically 0.2 to 1.2 days, with amplitudes ranging from a few tenths to about 2 magnitudes. They are frequently called 'cluster variables', because most of the known examples are located in the GLOBULAR CLUSTERS of our galaxy. RR Lyrae stars are bright enough to be visible in nearby galaxies and are used as distance indicators, because they have a relatively well-determined average absolute magnitude. There are subclasses of RR Lyrae stars—called RRa, RRb, RRc and RRd—that are classified according to the shape of their light curves; the RRa stars have asymmetric light curves with steep rises and more gradual declines, RRb stars have light curves with lower amplitudes (about 0.6 mag), but are otherwise similar in appearance to RRa stars, and both classes pulsate in the radial fundamental mode. RRc stars have nearly sinusoidal light curves with amplitudes of about 0.5 mag and are radial first-overtone pulsators. The RRd stars are double-mode stars with both the fundamental and first overtone modes excited. In addition, some RR Lyrae stars show a periodic modulation of the amplitude and shape of their light curves in a quasiperiodic manner called the Blazhko effect; however, this modulation leaves the period of the pulsation mode unchanged, except for possible evolutionary effects. The origin of the Blazhko effect is not understood yet, although magnetic field modulation and nonlinear effects have been proposed.

δ Scuti

Several classes of variable stars lie near the main sequence with spectral classes between early A and early F, with the most obvious distinction being their pulsation periods. The δ Scuti stars are the most numerous and best known of these pulsators. The δ Scuti stars are spectral class A5 to F2 stars that are burning hydrogen either in their cores or in a thick shell surrounding the core; they have masses between 1.6 and 2.6$M_\odot$. They can be monoperiodic or multiperiodic (the record holder has more than 20 known periods), with periods between 0.5 and 6 h. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

amplitudes range from the photometric detection limit (<1 millimag) up to about 1 magnitude. The pulsations are a combination of radial and/or non-radial modes; most of the non-radial modes are low-overtone p-modes, although in some cases we may be seeing low-overtone g-modes. Typically, the large-amplitude pulsators are more evolved, slower rotators, and have one or a few (mostly radial) modes present. Lower-amplitude δ Scuti stars are found throughout the instability strip, and many of these stars rotate rapidly, with $v \sin i$ values ranging up to 220 km s^{-1}. A plot of the number of δ Scuti stars versus pulsation amplitude shows a rapid rise towards lower amplitudes down to instrumental detection limits. This suggests that the ~30% of stars in the instability strip that are δ Scuti stars is actually a lower limit, and that many more are waiting to be discovered. Pulsation driving for δ Scuti stars comes mostly from He^+ to He^{2+} ionization, with some driving from H to H^+ and He to He^+ ionization. δ Scuti stars are promising candidates for astroseismology, and several stars have been analysed. The observations require observing runs of a month or more from several sites around the globe to reduce confusion in analysing the period spectra and also to reduce the noise level enough to detect the low-amplitude modes that are present. Because most δ Scuti stars are rapid rotators, rotational splitting effects must be included in theoretical pulsation frequency calculations and in the evolutionary models, which is challenging. So far, only a handful of stars have been analysed, and the analyses to date underscore the need to include the effects of rotation to predict the pulsation frequencies reliably.

δ Scuti pulsations also show up in two classes of metal-poor stars. First, there are the Pop II SX Phoenicis stars, which are large-amplitude radial pulsators. The SX Phoenicis stars tend show up in dense open clusters and globular clusters as BLUE STRAGGLERS, and it is likely that they are binary merger objects. The second type are pulsators within the λ Boötis spectral class; about 50% of these stars are also low-amplitude δ Scuti type pulsators. The existence of pulsations in the λ Boötis stars suggests that the deficiency of some metals in their spectra is a superficial phenomenon and that the driving region is probably not metal or helium depleted. Otherwise, we would not expect to see pulsations because the required partial ionization zone would not be present.

Rapidly oscillating Ap stars

The rapidly oscillating Ap (roAp) stars are A stars with peculiar spectra that show enhancements of metal lines, typically of transition metals and rare-earth elements (see PULSATING AND CHEMICALLY PECULIAR UPPER MAIN SEQUENCE STARS). They pulsate in high-overtone p-modes that have amplitudes of about 1 to 10 millimag, and the periods range from 4 to 15 min. These stars have strong (up to kilogauss) magnetic fields and are slow rotators, with rotation periods ranging from about 2 to 12 days. The combination of a strong magnetic field and slow rotation is the key to explaining the light variations using the *oblique pulsator model*. Here, the light variations are due to a dipole ($\ell = 1$) mode that is aligned with the magnetic field, causing the aspect angle of the light and dark hemisphere to change as the star rotates. The oblique pulsator model is quite successful in explaining the observed pulsation behavior, including the pulsation phase jumping by π radians during a magnetic field reversal. The magnetic fields are sufficiently strong that they influence the nature of the pulsations, and this is something that has yet to be modeled consistently in pulsation calculations. These stars occupy much the same region of the H–R diagram as the δ Scuti stars, but there the similarity ends. The short periods of the pulsations are consistent with pulsation driving due to the κ mechanism in the H/He partial ionization zone, but this requires models with chemically inhomogeneous envelopes to deplete the He^+ ionization zone of enough He to 'quench' the δ Scuti pulsations that would otherwise be present. Understanding the nature of pulsation driving and the structure of the outer layers of roAp stars is an area of active development.

γ Doradus stars

The final class of pulsating stars in the A to F spectral class that we discuss includes the newly discovered γ Doradus variables. The γ Doradus stars lie just coolwards of the δ Scuti instability strip with spectral classes between F0 and F3. They lie on or near the main sequence and have masses near $1.5 M_\odot$. The periods range from about 0.5 to 3.0 days, which is far longer than the radial fundamental mode; the fact that several stars are multiperiodic implies g-mode pulsation rather than a rotating, spotted star. At present, the pulsation driving mechanism is not known, but the likely candidate is the κ, γ mechanism acting in a region enhanced in metal abundances due to diffusion with a temperature of about 200 000 K. Thus, the A to F stars appear to show the full range of possible opacity driving mechanisms, from H/He driving near the surface in roAp stars, 'normal' He^+ driving in the δ Scuti stars, and driving due to iron-peak metals in the γ Doradus stars.

β cephei stars

There are several classes of photometric and spectroscopic pulsating stars of spectral type O and B that are on or just off the main sequence. Of these, the best known are the β Cephei stars, which are B0 to B2 stars of luminosity class III, IV and V that pulsate with periods between 3.5 and 6 hours. Most have V-magnitude ranges below 0.2 mag, radial velocity amplitudes below 50 km s^{-1}, and are relatively slow rotators ($v \sin i \leq 150$ km s^{-1}), although exceptions exist. The pulsations are radial and non-radial p-modes, and are driven by the κ mechanism acting on iron peak elements. Many β Cephei stars are found in young GALACTIC OPEN CLUSTERS, and these observations show that most are core hydrogen burners, along with some in the core contraction and hydrogen shell burning phase of their evolution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Slowly pulsating B stars

The slowly pulsating B (SPB) stars are mid to late B stars on the main sequence that exhibit low-amplitude light variations (<0.1 mag) with periods ranging from 0.6 to 3 days. They are like the β Cephei stars in that their pulsations are driven by the κ mechanism acting on iron peak elements, but the driving region is located deep enough in the star to drive the longer observed periods, which are g-modes. SPB stars have masses between 2.5 and $6M_{\odot}$, and they are all core hydrogen burning stars. Almost all of these stars are relatively slow rotators with $v \sin i \lesssim 100$ km s^{-1}. Some also show line profile variations, where the shape of the spectral line changes in a periodic manner as different light and dark patches are seen on the rotating star.

Line profile variables

A number of O and B stars show line profile variations due to non-radial pulsations. Two classes of objects have prototype stars; they are ζ Ophiuchi and 53 Persei. Many of these stars are rapid rotators, and the theoretical framework is only just now being constructed for us to interpret their pulsations. Finally, there are the B stars with emission lines in their spectra (BE STARS); a few show line profile variations, and many show photometric variability with periods of about a day. Currently, there is considerable debate on whether the observed light variations are due to pulsation or some other mechanism; we mention the Be stars here for completeness and as a warning that we do not understand their observed behavior yet.

Pulsating white dwarfs

WHITE DWARFS are the remnants of stars with initial masses less than $8M_{\odot}$, and some of them pulsate in well-defined temperature ranges, with typical maximum amplitudes up to 0.3 mag. All of these stars are multiperiodic g-mode pulsators, and several have precise structural and intrinsic properties derived from astroseismological analysis of their pulsation periods. Some of these stars are still contracting appreciably and are classified as pre-white dwarfs (PWDs). The most luminous PWDs are the PLANETARY NEBULA nuclei variables (PNNVs), which are evolving rapidly from thermally pulsing AGB stars to the white dwarf domain. They have typical periods of 1000 to 2000 s. The hottest pulsators are the pulsating PG 1159 (or GW Virginis) STARS, which have temperatures between 80 000 and 160 000 K. Their pulsations are probably driven by partial ionization of C and O. The nature of the pulsation mechanism of the cooler white dwarfs has been a subject of some debate, but evidence is beginning to favor the idea that convection zone modulates the energy flow due to *convective driving*, where the pulsation energy is modulated by adjustments of the convection zone. At about 25 000 K, the convection zone in the helium layer drives pulsations, and here we find the variable DB white dwarfs (called DBV or V777 Herculis stars) (see DB PULSATING STARS). Typical periods range from 400 to 1200 s, and the DB spectral type implies that the atmospheres are pure helium. The coolest pulsating white dwarfs have a DA spectral type and they are called DAV or ZZ Ceti stars. Pulsation driving occurs in the hydrogen layer convection zone; they have temperatures between 11 000 and 12 000 K and periods ranging from 100 to 1200 s. Astroseismology shows that the surface layer of hydrogen for several of the ZZ CETI stars has a mass of about $10^{-4}M_{\star}$.

Other pulsating stars

Space prevents a discussion of all the classes of pulsating stars, but we will mention a few other classes of stars that exist and recommend the reader consult the bibliography for further details. The most luminous pulsators are the Luminous Blue Variables (LBVs), which appear to be massive ($M_{\star} > 20M_{\odot}$) stars that are up to 10^6 times as luminous as the Sun (see HIGH-LUMINOSITY STARS). Possibly related to the LBV stars are the A-type supergiant α Cygni and others like it; they appear to be somewhat less massive and cooler than the typical LBV star. Some evolved low-mass stars show reduced amounts of hydrogen in their atmospheres, and a subset of these stars pulsate. First are the spectacular R CORONA BOREALIS STARS, which have nearly constant maxima interspersed by dramatic dimming episodes. Second are the extreme helium pulsators, sometimes also called hydrogen-deficient (H-deficient) stars. There are also K giant stars that pulsate, although they have not been systematically observed or reviewed. The best known example is Arcturus, and all of these stars have very low amplitudes, requiring high-precision photometry or spectroscopy to detect (see LARGE HOT SUBDWARFS AND EXTREME HORIZONTAL BRANCH STARS). Finally, there are the EC 14026 stars, or pulsating subdwarf B (SdB) stars, which are core helium burning stars with hydrogen-rich envelopes that have too little mass for the star to ascend the AGB.

The advent of astroseismology

In the previous discussion, we alluded to some situations where the study of pulsating stars impacts other fields of astronomy. Each pulsation mode in a star can be thought of as an independent probe of the star's interior, and using this information to determine (or constrain) the interior structure of a star is what the subject of astroseismology is about. The analogous field for the Sun is helioseismology, and this subject is covered by other articles in this encyclopedia (see SOLAR INTERIOR, HELIOSEISMOLOGY: THEORY, HELIOSEISMIC OBSERVATIONS). Even if a star pulsates in only one or two modes, this can reveal the mean density. This is enough that in the case of Cepheids and δ Scuti stars, N Simon suggested in 1982 that enhancing the opacities in the region around 200 000 K would resolve a number of outstanding discrepancies between observation and theory. Two major efforts at recalculating the opacities (the Opacity Project with the OP tables and Lawrence Livermore National Laboratory with the OPAL tables) found the suggested opacity enhancements (see SOLAR INTERIOR: EQUATION OF STATE AND

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

OPACITY). This essentially solved the double-mode period ratio problem in Cepheids and the discrepancy between the pulsation and evolutionary masses of Cepheids, and also revealed the driving mechanism for the β Cephei and SPB stars.

More precise determinations of the internal structure are possible if many pulsation modes are observed, and the present record holders (excluding the Sun) are the white dwarfs PG 1159–035 (see the example light curve in figure 5) and GD 358, with over 100 modes present in each star (see STELLAR INTERIORS). Next best are δ Scuti stars, where some have about 20 modes present. Much current theoretical and computational effort is going into developing better evolutionary models for various classes of stars for pulsation analysis. The other areas of intense theoretical effort are as follows: including the effects of rotation self-consistently in interpreting observational photometry and spectroscopy; computing the evolutionary models; and computing the frequency shifts induced by rotation.

Identification of the ℓ, m and n 'quantum numbers' for observed pulsation modes is crucial for unlocking the internal structure of a star via astroseismology. A variety of methods are available for these purposes, although none is applicable for all types of pulsating stars, and each has its strengths and weaknesses. If enough pulsation modes are present (say for some pulsating white dwarfs), one can use the regular pattern of period and/or frequency spacings to determine the mode identification. This is the only method that can, in principle, determine the ℓ value, the overtone number (n), and the azimuthal quantum number (m). Radial modes, p-modes, and g-modes can also leave their signature in the form of line profile variations, changes in the equivalent widths of spectral lines, changes in time-dependent amplitudes as a function of color, and different times of maximum light in different wavelengths. Non-radial pulsations, in particular modes with $\ell = |m|$, will have spectral line profiles that change shape as different parts of the non-radial mode become visible as the star rotates. This techniques works best on the O- and B-star pulsators. The equivalent width method makes use of spectral lines whose equivalent width is very sensitive to temperature. Comparing these changes to changes in the ratio of light curve amplitudes for different color photometric filters allows determinations of the ℓ value, but not n or m. Other similar methods make use of comparisons between the phase of radial velocity curves and light curve intensity ratios in different colors. So far this method has seen the most application and success with δ Scuti stars. These methods do not work for white dwarfs, but in this case, the light curve amplitude increases rapidly with decreasing wavelength and the rate of the amplitude increases is a strong function of ℓ. This method requires time-dependent UV amplitudes, and the few white dwarfs studied so far have $\ell = 1$ or 2. The last method depends on knowing the pulsation period and assuming that it is a radial mode. Assuming radial mode pulsations, one can estimate the *pulsation constant* (Q) via

$$Q = \Pi\sqrt{\bar{\rho}/\bar{\rho}_\odot}$$

where $\bar{\rho}_\odot = 1.41$ g cm^{-3} and $\bar{\rho}$ for variable stars of a given class can usually be estimated with tolerable accuracy because it is a slowly varying function of stellar parameters. The pulsation constant decreases rapidly with increasing overtone number, so it offers an estimate of the overtone number n instead of the ℓ value.

A better understanding of pulsation driving has also led to the prediction of some classes of pulsating stars before they were discovered observationally. The first example was the prediction of the pulsational instability of helium atmosphere white dwarfs by D Winget in 1981, based on understanding that the same mechanism causing the hydrogen atmosphere white dwarfs to pulsate should also apply to the helium atmosphere white dwarfs. The next predicted pulsating star class was the subdwarf B-stars (SdB stars), predicted by S Charpinet 1993.

One class of pulsating stars that has not been found as yet are the solar-like pulsators. After seeing the importance of solar pulsations in determining the Sun's structure, people started trying to find similar pulsations in other stars, with Procyon and α Centauri being two of the favorite targets. To date, none of these searches has produced a verifiable detection, although unconfirmed detections and upper limits abound. Much of the lack of success is due to the extremely low expected amplitudes; for the Sun, the amplitudes are on the order of cm s^{-1} or $\Delta I/I \lesssim 10^{-5}$, which is at least an order of magnitude more sensitive than any search to date.

Expanding on the future

For the future, several trends are starting to appear. The early 1990s saw the advent of searches for massive compact halo objects, and 'massive photometry' is becoming a powerful new technique for relatively unbiased populations of previous and newly discovered variable stars, particularly in the MAGELLANIC CLOUDS. For example, the MACHO project has discovered thousands of new variable stars in the LMC, including RR Lyrae, Cepheid, W Virginis and long-period variables; most of them are new discoveries.

Satellites are also sources of new variable star detections. The HIPPARCOS satellite discovered a large number of SPB stars, yet only a few new β Cephei stars, suggesting that selection effects of previous searches severely undersampled the population of SPB stars. Multilongitude, multisite networks are becoming the tool of choice for observers who wish to disentangle the complicated light curves of multiperiodic stars like the pulsating white dwarfs and δ Scuti stars. The Delta Scuti Network (DSN) and Whole Earth Telescope (WET) are two of the longest running and best-known collaborations, although there are many others, including MUSICOS, which is probably the only multisite network devoted to spectroscopy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

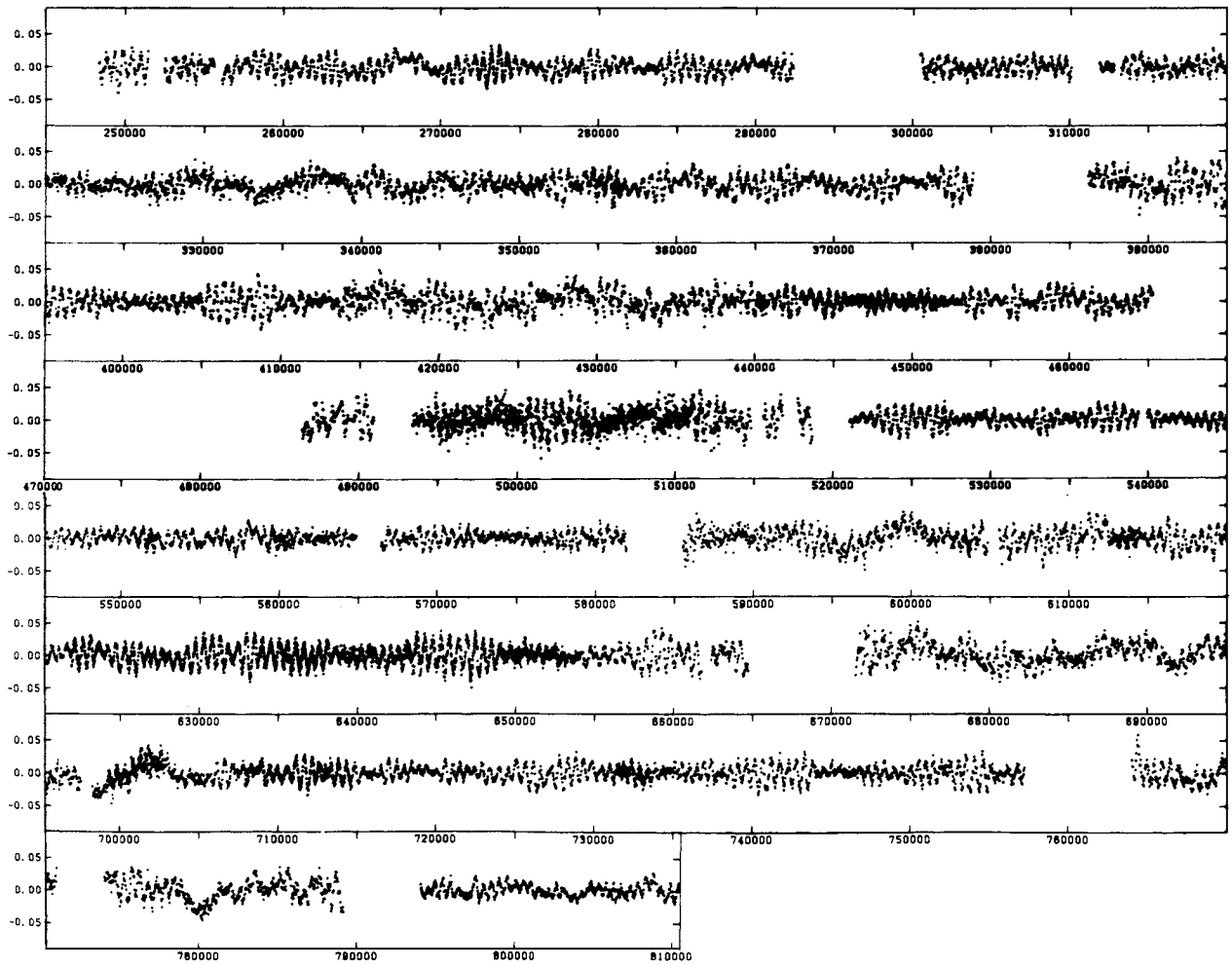


Figure 5. A sample light curve for the multiply periodic pulsating pre-white dwarf PG 1159–035. Note the 'beating' caused mainly by interference between the dominant 516 and 539 s pulsation modes.

Another trend is the appearance of new theoretical and computational tools for understanding the observed pulsations and exploiting the potential of astroseismology. There are three major areas of activity. First, many of the δ Scuti and upper main sequence stars rotate rapidly, which requires the inclusion of rotational physics in the evolution, pulsation and atmosphere models. Second, although many stars are non-radial pulsators, there are at present very few theoretical tools for studying the nonlinear aspects of non-radial pulsations. Besides amplitude variations, many non-radial pulsators also show Fourier transform peaks with frequencies equal to the sum and/or difference of the larger amplitude pulsation modes. These 'combination' peaks are not regarded as real pulsation modes, but they do contain information about the energetics of the driving region, mode selection and nonlinear effects that we are ignoring at present. People are starting to develop theories for understanding this nonlinear behavior, and this should become an area of increasing activity in the near future. Finally, many pulsating stars have convection zones that play at least some role in the driving and damping of pulsations. To date, most theories for convection are either too simplistic, have too many free parameters to be predictive, and/or add enough extra complications that the computer time required for calculations goes up enormously. Add to this the fact that there is no one agreed-upon model, and we have our current situation where convection is either ignored or treated simplistically. The return to pulsational stability at the cool (red) edge of the instability strip for many stars depends on convection/pulsation interactions, as does the existence of pulsations in several classes of stars. While some basic understanding of stellar pulsation can be gained with current simplified models, more effort will be needed here before we can make quantitative predictions about mode selection and mode amplitudes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

The topic of stellar pulsations has a vast literature, so we only provide references to some key papers and books covering the development of the field, along with guides to surveys of recent research. Except where noted, one should assume the material is technical.

Good and readable accounts of variable stars in general that have much material on pulsating stars are:

Hoffmeister C, Richter G and Wenzel W 1985 *Variable Stars* (Berlin: Springer)
Petit M 1987 *Variable Stars* (New York: Wiley)

Good discussions on the early history of the theory and observations of pulsating stars may be found in:

Ledoux P and Walraven T 1958 *Encyclopedia of Physics* ed S Flügge (Berlin: Springer)
Rosseland S 1964 *The Pulsation Theory of Variable Stars* (New York: Dover)

The best general references on stellar pulsation are:

Cox J P 1980 *Theory of Stellar Pulsation* (Princeton, NJ: Princeton University Press)
Gautschy A and Saio H 1995 *Ann. Rev. Astron. Astrophys.* **33** 75
Gautschy A and Saio H 1996 *Ann. Rev. Astron. Astrophys.* **34** 551
Unno W, Osaki Y, Ando H, Saio H and Shibahashi H 1989 *Nonradial Oscillations of Stars* 2nd edn (Tokyo: University of Tokyo Press)

The two most recent conference proceedings devoted to stellar pulsation are:

Bradley P A and Guzik J A (eds) 1998 *A Half-Century of Stellar Pulsation Interpretations: a Tribute to Arthur N Cox* (San Francisco: Astronomical Society of the Pacific)
Stobie R S and Whitelock P A (eds) 1995 *Astrophysical Applications of Stellar Pulsation* (San Francisco: Astronomical Society of the Pacific)

P A Bradley

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stellar Rotation on the Main Sequence

The conservation of ANGULAR MOMENTUM is the one effective counterbalance to the inexorable pull of gravity in the universe, and so everything rotates. Stars acquire their angular momentum when they form, and, indeed, the manner in which nearly all this initial angular momentum is dissipated remains poorly understood, but without substantial angular momentum loss an interstellar cloud could never contract enough to become a star.

Yet stars do form and overcome this problem. When hydrogen fuses to helium in the core, the star is on the main sequence. The rotation rates of all but a few main sequence stars are well below breakup velocity (the speed at which centrifugal force balances gravity at the equator), and rotation therefore is unlikely to fundamentally alter a star's evolution. In those few cases with excessive rotation, the gravity experienced by the star is diminished, and the overall evolution can be very different.

The rotation of a star is usually described by a single number: either $v \sin i$, the equatorial speed of the star projected onto our line of sight, or $P_{\rm rot}$, the rotation period. However, rotation is presumably three dimensional and can vary with latitude on the star (differential rotation) or with radius. On the Sun, there appears to be no change in rotation with radius (as determined from helioseismological studies of the Sun's interior; see SOLAR INTERIOR: ROTATION), but the Sun spins a little faster at its equator than at high latitudes. This DIFFERENTIAL ROTATION does not affect the deep interior, but it is crucial in determining much of the surface phenomenology that makes the Sun so fascinating: spots, plages, active regions, the corona and so on.

Observational technique

The first study of stellar rotation was made by Galileo when he associated the apparent motions of sunspots with the spin of the Sun. Soon after, Scheiner showed that the spots moved at different rates at different latitudes, demonstrating the existence of differential rotation and the fact that the Sun is a fluid.

The study of stellar rotation was systematized by OTTO STRUVE, a spectroscopist who explored many areas of astrophysics. He showed that different line widths in different stars were due to rotation, and he refined the most basic technique for determining rotation: analysis of line breadths in spectra, either by visual comparison or by profile fitting. Early-type main sequence stars (types O, B, A and early F) typically have $v \sin i$ values of 50–400 km s^{-1}. The intrinsic breadth of stellar absorption lines is only about 10–20 km s^{-1}, so the effect of rotation is obvious in early-type stars. Such stars are also intrinsically bright and blue and so were prime targets during the photographic era (see STELLAR ATMOSPHERES: EARLY-TYPE STARS). As a result, most of the data for early-type stars date from that time. This method is calibrated by direct measurement of line widths for a range of rotation rates: $v \sin i$ is then half the full width of the line where it meets the continuum. Once suitable standards are established, a well-trained eye can reliably estimate $v \sin i$ down to about 10 km s^{-1} with an accuracy of 1 km s^{-1}.

As noted, $v \sin i$ is that component of the star's true equatorial rotation rate that is projected onto the line of sight. Such information is entirely adequate for studying the rotation of classes of stars for two reasons. First, the rotation axes of stars are randomly oriented in space. This has been determined by, for example, observing rotation in clusters of stars at different Galactic latitudes and noting that high-latitude clusters have the same distribution of rate as low-latitude clusters. The immediate conclusion from this fact is that stellar rotation does not have its origin in Galactic rotation. The orbital axes of visual binaries have also been shown to be oriented randomly. Second, for a random distribution of axes one is most likely to observe a star nearly equator on (for about 87% of stars the measured $v \sin i$ is at least half of the true equatorial velocity), and stars are very unlikely to be seen pole on. In other words, nature makes it easy. However, the inevitable ambiguity of $\sin i$ makes it impossible to specify the rotation rate of an individual star reliably.

The most massive stars are not as congenial to this method as later spectral types. Many of them have very high rotation rates (300 km s^{-1} or more), making it difficult to define the observed spectrum line. In such cases there can also be important physical effects such as gravity darkening that influence spectrum formation in the first place. Also, the dearth of sharp-lined O stars strongly suggests the presence of turbulent broadening, but the precise extent of that turbulence is not well understood. For most massive stars, rotation is so dominant that this uncertainty in the zero point is not critical.

This same technique of measuring line broadening was applied to stars as cool as the Sun but without success: they were all rotating too slowly to be resolved on moderate-resolution spectra (see also STELLAR ATMOSPHERES: COOL STARS). The slow rotation of late-type stars (late F, G, K and M) is intimately tied to the presence of convective envelopes on those stars, as discussed below. More recently it has been possible to determine $v \sin i$ values down to about 2 km s^{-1} by obtaining high-resolution ($R \approx 100\,000$), high signal-to-noise (200 or more) spectra and by then applying careful modeling of the entire line profile. Fitting of the Fourier transform of the profile also helps to deconvolve rotation from other broadening mechanisms of comparable magnitude. The uncertainties of this procedure are relatively large (about 1 km s^{-1}) when $v \sin i$ is small, but it allows the determination of rotation in slowly rotating solar-type stars, for which no other information exists.

As discussed below, the fundamental nature of late-type stars leads to their exhibiting solar-like phenomena such as surface spots. These spots can be significant enough to lead to detectable variations in the light of a star from which the rotation period, $P_{\rm rot}$, can be determined. This is essentially the inverse of the angular velocity, Ω. Rotation periods are independent of aspect, and so it

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is now possible to determine accurately true rotational distributions for a sample of late-type stars. Changes in the observed $P_{\rm rot}$ or multiple periods in one dataset have been interpreted as evidence for differential rotation. Differential rotation is a key ingredient of models for generating magnetic fields in late-type stars and so these observations are valuable, but information is available for only a few stars, and those are not old, quiet ones like the Sun. An attempt has been made to detect differential rotation in some A stars through very careful analysis of the line profiles, but none was seen. It has also been possible to measure rotation periods for some early-type stars because they can also have magnetic fields that lead to inhomogeneous surfaces, hence variations. These have been especially important for some of the Ap/Am stars.

More details on technique may be found in Gray's excellent text on stellar atmospheres (Gray 1992).

Sources of observations

Because the rotation of early-type stars was well studied photographically, most of the observations for those stars were compiled into comprehensive catalogs long ago. The original catalogs are in fairly obscure observatory publications, but all of these data and more can now be accessed at NASA's Astronomical Data Center (adc.gsfc.nasa.gov). For individual stars, the best source is SIMBAD (simbad.u-strasbg.fr/Simbad).

As with any database, discretion should be used in interpreting data, as not all are of the same quality. In some instances significant misconceptions by some observers have led to large systematic errors that propagate into the catalogs: see Soderblom *et al* (1989) for an example.

General remarks on rotation

Each type of main sequence star has rotational characteristics that set it apart from other types, but there is a natural break in the mid-F spectral type range. For early spectral types (O, B, A and early F), energy is carried to the surface predominantly by radiative processes. For the late-type stars (late F, G, K and M), energy transport in the outer layers is by convection. The presence of a convective envelope and its interaction with the rotation are fundamental to the behavior of late-type stars, as will be discussed below.

Two basic facts apply to rotation in stars. First, the rotation axes of stars are randomly oriented. As noted above, this means that stellar rotation does not have its origins in Galactic rotation, for if it did then rotation axes would tend to be aligned with the Galaxy's rotation. The second basic fact is that a mean relation appears to hold for the specific angular momentum of a star at the time of formation (its angular momentum per unit mass) and the star's mass, first shown by Kraft:

$$\langle I/M \rangle \propto M^{-1.1}.$$

where I is the angular momentum and M is the stellar mass. This relationship (Kawaler 1987) can easily be derived by assuming that all stars form from a medium of uniform density, a condition that is surely not true in nature. Also, the need to dissipate most of a star's initial angular momentum weakens the tie between initial conditions and later behavior on the main sequence, and the underlying reason for the above relation is not obvious. The Kraft relation is at best a statistical one that predicts a typical amount of angular momentum for a star of a given mass. There are, for example, well-studied binary systems with components of nearly identical mass yet substantially different $v \sin i$ that demonstrate that initial conditions are important too.

The early-type stars: O, B, A and early F

The most massive stars (O and B types) sometimes show evidence of NON-RADIAL STELLAR PULSATIONS and strong winds, as well as abundance anomalies that may be driven by internal mixing. All these effects are related to rotation and complicate the determination and interpretation of rotation, and rotation rates can be high enough to influence a star's evolution. Intermediate-mass stars (types A and early F) can have dramatic abundance anomalies and other effects that are related to diffusion in the surface layers, which in turn is affected by rotation. The presence of significant rotation in these stars in itself complicates studies that depend on measuring weak lines (to determine abundances) or on resolving components of absorption lines (to measure magnetic field strengths). Thus phenomena can take place on rapid rotators that cannot be directly observed.

The distribution of $v \sin i$ for the O stars, for example, has two peaks, one centered at about 100 km s^{-1} and the other at 300 km s^{-1}, both about 100 km s^{-1} wide. The highest $v \sin i$ values are near 400 km s^{-1}. The 300 km s^{-1} peak is seen only among the main sequence O stars, and it matches the distribution seen in Be stars, while the 100 km s^{-1} peak is seen in all luminosity classes among the O stars: main sequence, giants and supergiants. Significant turbulence (at about the 20 km s^{-1} level) is present in O stars as well, and this accounts for the absence of stars with sharp-lined spectra. The presence of strong radiation-driven winds in O stars means that significant mass loss is present, and this implies a trend of rotation rate with age while on the main sequence.

Rotation can play a key role in the evolution of very massive stars (10–40$M_\odot$). First, rapid rotation can affect the stellar equilibrium, reducing the central pressure and temperature, leading to lower luminosity. The maximum luminosity decrease is ∼3%. Second, rapid rotation can alter the shape of the atmosphere of the star, creating local differences in temperature and pressure. The effects depend on orientation and tend to make a star look cooler. The most important effect of rapid rotation, however, is its influence on mixing of chemical elements and on angular momentum transport in the STELLAR INTERIOR. Rotation especially influences turbulent diffusion. Higher rotation leads to more efficient turbulent diffusion, and that makes a star have about double the luminosity it would have

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

in the absence of the diffusion. Also, rotationally induced mixing has been used to explain several features of massive star evolution, namely the N and He enrichment of O stars, and the lack of a gap (predicted by models) between main sequence and post-main-sequence stars in the observed H–R diagram.

For B stars, the typical rotation rate is about 140 km s^{-1} for all spectral subclasses, somewhat greater than for the more-massive O stars. A comparison of rotation in B giants with B dwarfs shows that early B stars do not rotate as solid bodies and that differential rotation is important. An important subtype among the B stars is the Be stars (B stars with emission lines), which constitute about 15–20% of B stars. Be stars rotate at higher rates than non-Be stars, with an average equatorial velocity (corrected for $\sin i$) of 265 km s^{-1} and with a distribution of velocities that is not skewed. In other words, the distribution of $v \sin i$ for Be stars is Gaussian, not the Maxwell–Boltzmann function expected for a random distribution of rotation axes. Note that the breakup velocity for Be stars is about 500–550 km s^{-1}, so they are comfortably below that level. The winds seen in Be stars are influenced by the rotation, being slow and dense in the equatorial regions and fast and tenuous at the poles; see BE STARS.

Most O and B stars are found in star-forming regions because they are so young. The distributions of rotation for these various clusters of stars do not differ significantly from one another, given sample sizes. There are real and significant star-to-star differences in angular momenta at a given mass, but the distributions of angular momenta seem to be independent of initial conditions to a large degree. There are suggestions that metallicity plays a role in how angular momentum is distributed within clusters, but not yet enough information to form a clear picture.

Most A and early-F stars have moderate rotation rates (50–100 km s^{-1}) that are consistent with the Kraft relation above, but among the A stars are found objects with peculiar chemical abundances, and these peculiarities are closely related to rotation. The Am, or metallic-line A, stars show weak Ca II and/or Sc II features, but enhanced lines of heavy elements. Many are found in BINARY SYSTEMS, and temperatures range from 7000 to 10 000 K. The Ap, or magnetic A, stars show enhanced lines of elements such as Si, Cr, Eu and Sr, few are in binaries and temperatures range from about 8000 to 15 000 K. The HgMn stars have enhanced Hg and Mn lines and temperatures from 10 000 to 15 000 K.

All of these types have rotation rates well below average for A stars. The average $v \sin i$ for Am and Ap stars is about 40 km s^{-1}, a factor of 3–4 lower than normal stars of the same colors. Several arguments indicate that these low $v \sin i$ values are not due to these stars being rapid rotators seen pole on. The HgMn stars rotate especially slowly: $\leq$20 km s^{-1} in almost all cases. Very few chemically peculiar stars have $v \sin i > 100$ km s^{-1}. Thus a low rotation rate appears to be a necessary condition for the Ap, Am or HgMn phenomenon to arise, but that does not explain why these stars have low rotation rates in the first place. Rotation rates for these peculiar stars decline with age, suggesting that angular momentum loss enabled through a magnetic field (like the late-type stars discussed next) or the presence of a close companion plays a key role. However, the age relationship does not appear to hold for the HgMn stars, leaving their origin uncertain. One argument is that the HgMn stars, being cooler, spend enough time losing angular momentum in the pre-main-sequence phase to achieve the needed low rotation rates. The apparent relation between rotation and age for some of these peculiar stars may be just that—apparent—and the true origins are not fully understood. Note that the inverse statement about rotation and peculiar stars also holds: there are very few, if any, normal A stars that rotate slowly.

The late-type stars: late F, G, K and M

Rotation is a key determinant of much of the phenomena we see on late-type stars because these stars possess convective envelopes that interact strongly with the rotation. The paradigm for understanding this behavior is drawn from what we see up close on the Sun: rotation and convection interact in the outer layers to produce complex circulation patterns in the ionized stellar material. Differential rotation is especially important in this mechanism—the dynamo mechanism (see DYNAMOS: SOLAR AND STELLAR)—that can regenerate a seed magnetic field. The magnetic field, in turn, can grip an ionized stellar wind well beyond the star's surface, forcing angular momentum loss, a process we witness on the Sun. Thus the Sun by its nature loses angular momentum steadily, and, by analogy, other stars with convective envelopes and winds do likewise. We are confident that all late-type stars possess convective envelopes from basic physical considerations. We still know little about winds in stars, however, except in cases where they are very strong, and so we assume that other late-type stars have some sort of wind similar to the Sun's.

One consequence of this paradigm is that two stars formed at the same time but with different rotation rates will tend toward the same rotation rate with time because the more rapidly rotating star generates a stronger magnetic field and so loses angular momentum faster than its slowly rotating sibling. Young stars tend to rotate rapidly because they have not yet had time to lose much angular momentum, while old stars rotate slowly.

The magnetic fields of late-type stars give rise to their rich phenomenology, collectively described as activity. This activity, in turn, can help us study the rotation. Surface inhomogeneities (i.e. spots) can be large enough to produce observable variations in the light of the star and from these the rotation period can be inferred. In general only the young, active stars have measurable rotation periods, although it should be possible to measure the same effect in old stars with observations of sufficient quality since space observations of the Sun reveal its rotation. Careful long-term observations have yielded changes in $P_{\rm rot}$ that indicate the presence of differential

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rotation on other stars. The strength of activity in whatever form it is observed (Ca II H and K emission, Hα, x-rays, etc) is strongly correlated with the rotation for late-type stars, so that rapidly rotating stars are always active stars because of the strong magnetic fields generated, while slowly rotating stars are inactive. The rotation–activity relations have different slopes for different masses, so that lower-mass stars produce more activity for a given rotation rate. For reference in what follows, the Sun is 4.6 Gyr old and rotates with a period of about 27 days and an equatorial velocity of 1.8 km s^{-1}.

Young stars are studied by observing them in STAR CLUSTERS, both because most zero-age main sequence (ZAMS) stars are in clusters and because clusters provide good-sized samples of stars of known age. The late-type members of clusters are well cataloged only for the few nearest of the clusters (e.g. the Pleiades, Hyades, α Persei, Praesepe, M34, M67), and so we assume that any one of these clusters provides a representative view of any ensemble of stars at that cluster's age.

The best-studied ZAMS cluster is the Pleiades, at an age of 100 Myr, and the Pleiades tells a very interesting story. The Pleiades contains stars with $P_{\rm rot} = 1/4$ day, or 100 times the solar rate. The fastest-rotating G dwarfs found in the field rotate at about 10 times the solar rate, so these Pleiades stars are extraordinary. These ultrafast rotators (stars with $v \sin i \geq 20$ km s^{-1}) constitute about 20% of the G and K dwarfs in the Pleiades, the remaining 80% typically having $v \sin i$ values below 10 km s^{-1}. Thus the distribution of rotation for these ZAMS stars is bimodal, with the dominant fraction being stars that rotate relatively slowly. Observations of the α Persei cluster (age 50 Myr), while not as complete, confirm that the Pleiades is typical for its age. By contrast, there is very little spread in rotation for G dwarfs in the Hyades (age 600 Myr). Rotation periods are known for many Hyades stars, and the intrinsic spread in rotation about a mean relation is only about 10%. Moreover, there is a strong mass dependence to the rotation seen in Hyades stars, with a sharp decline in going from the late-F stars into the early-G dwarfs, whereas Pleiades stars in the same range show little mass dependence. These clusters tell us that, in the first billion years after reaching the ZAMS, solar-type stars can lose a lot of angular momentum and while doing so a strong mass dependence to the rotation is introduced and at the same time there is a dramatic convergence in the rotation rates. The spread in rotation in the Pleiades at any one mass is at least a factor of 20, versus the 10% seen in the Hyades.

The details of this convergence are still not clear. The fact that it happens so quickly and that such a huge spread exists on the ZAMS has led to the suggestion that the convective envelope of a solar-type star can decouple from its radiative core. This means that we see the less-massive envelope spin down quickly, with more gradual replenishment with angular momentum from the core later. This decoupling model can also help to explain the strong mass dependence seen since the ratio of core to envelope is very sensitive to the star's mass. However, it is not strictly necessary to invoke core–envelope decoupling to explain the observations, and so it is not yet possible to be sure whether that is what is going on.

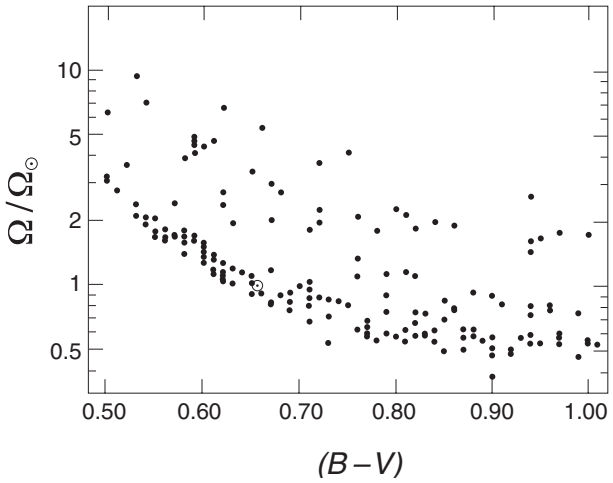


Figure 1. Rotation among solar-type stars. This diagram shows estimated rotation rates for stars like the Sun. A $B - V$ color of 0.50 corresponds to a spectral type of about F7 V and a mass of about $1.2M_\odot$, while $B - V = 1.00$ corresponds to K2 V and $0.7M_\odot$. The rotation rates were computed from measures of activity in the Ca II H and K lines, using a rotation–activity relation calibrated on stars of known rotation periods. The rapidly rotating stars at a given color are the young ones which gradually lose angular momentum to evolve into the slowly rotating stars. The Sun is shown at $B - V = 0.656$. Note that the Sun has a very typical rotation rate for an old star of its mass and that there is a range in rotation of about a factor of 5 at any mass.

Figure 1 shows the distribution of rotation rates from about $1.2M_\odot$ down to about $0.7M_\odot$. Note that the Sun has a very typical rotation rate for an old star of its mass (color). The distribution flattens for $B - V > 0.8$, in part because these low-mass stars have main sequence lifetimes that exceed the age of the Galactic disk, meaning that we cannot yet know how slowly the oldest of them will rotate. The upper bound also flattens for these low masses, which has been interpreted as indicating that the timescale for angular momentum loss becomes longer for lower-mass stars. The distribution is independent of mass, indicating that over this mass range stars all behave similarly.

In going to the lowest main sequence masses rotation becomes increasingly difficult to study. The small radii mean that $v \sin i$ values tend to be small, and these stars are intrinsically faint. A survey of rotation in M dwarfs found only 25% to rotate above the detection limit of 2 km s^{-1}. However, several stars were found with large $v \sin i$ values (30–50 km s^{-1}), all of which were later than M3.5. This is consistent with the timescale for spindown increasing with decreasing stellar mass.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

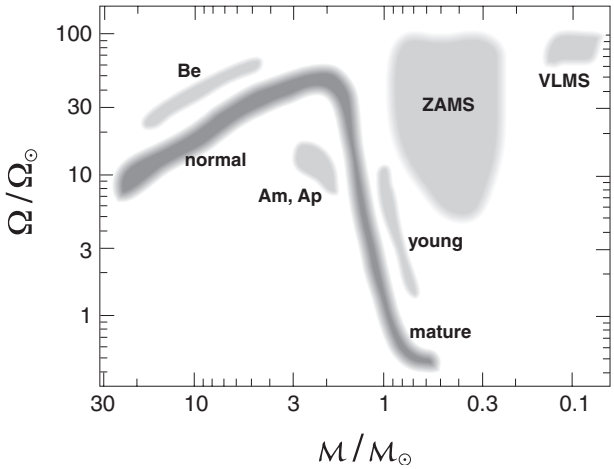


Figure 2. Schematic diagram of rotation in main sequence stars. This figure was computed from characteristic $v \sin i$ values as a function of spectral type, with masses and radii assumed for those spectral types based on model computations. What is shown is the surface angular velocity in proportion to the Sun's, and the values are meant to be indicative of the broad trends. Normal main sequence stars follow the broad, dark line, although little is known for stars below about $0.5M_\odot$. The Be stars rotate at about twice the rate of normal stars in the mass range indicated. Below $\sim 1M_\odot$, the locus of young stars is taken from figure 1, and the ZAMS region shows rotation rates for stars up to ~100 Myr in age. The 'VLMS' region is for very low-mass stars, although no comprehensive information exists for this mass range.

Very low-mass stars: the bottom of the main sequence and brown dwarfs

Only a handful of objects are known with masses at the substellar limit and below. These are all very faint, and, because they are inherently small objects, we expect very low $v \sin i$ values. Thus very little information is now available, yet it is surprising: high-resolution spectroscopy has shown significant rotation in all cases ($v \sin i \approx$ 10–20 km s^{-1}). The objects so far studied tend to be young (because young brown dwarfs are brighter than old ones), but the angular velocities that correspond to these $v \sin i$ values are comparable with the very fastest rotators at, say, $1M_\odot$. More observations are needed to provide a fuller context for interpretation, but the explanation for these high rotation rates probably lies in what defines these very low-mass objects, namely their being fully convective objects.

Summary

Figure 2 summarizes our knowledge of rotation in main sequence stars. The ordinate is the surface angular velocity relative to the Sun's. This declines for the highest-mass stars because such stars are less centrally condensed and so have large moments of inertia leading to large total angular momenta. The Be stars rotate at about twice the rate of normal main sequence stars in the mass range indicated, while Am and Ap stars rotate well below normal A stars. The ZAMS region is broad in both mass and rotation rate because stars at and below $1M_\odot$ reach the ZAMS with widely varying rotation rates. The very low-mass stars are shown, but no comprehensive information exists on rotation in general below $\sim 0.5M_\odot$.

Bibliography

As noted above, the best sources of data are on the web (adc.gsfc.nasa.gov and simbad.u-strasbg.fr/Simbad).

Gray D F 1992 *The Observation and Analysis of Stellar Photospheres* (Cambridge: Cambridge University Press)

Kawaler S D 1987 Angular momentum in stars: the Kraft curve revisited *Publ. Astron. Soc. Pac.* **99** 1322–8

Soderblom D R, Pendleton J and Pallavicini R 1989 Calibration stars for cross-correlation studies of stellar rotation, and an examination of the archival data *Astron. J.* **97** 539–45

David R Soderblom

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stephan's Quintet (NGC 7317, 7318A, 7318B, 7319, 7320)

A group of five galaxies in the constellation Pegasus, position RA $22^{h}\ 24^{m}$, dec. $+33^{\circ}\ 42'$. Four appear to be gravitationally interacting, and show similar redshifts. The fifth, NGC 7320, has a much lower redshift and is thus a foreground object not associated with the others. The group was discovered by Édouard Stephan in 1876.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stereo Comparator

A device that utilizes binocular vision to compare two images of the same area. The light paths are arranged so that the image seen in the left-hand eyepiece coincides with the image viewed in the right-hand eyepiece. Any differences between the two images appear to stand out from the image plane.

See also: blink comparator.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Steward Observatory

The Steward Observatory (SO), created in 1916, is the research arm of the Department of Astronomy of the University of Arizona (UA). Its headquarters are in Tucson, Arizona, on the university campus. Its observational facilities are on Mount Graham (10 m Heinrich Hertz SMT; 1.8 m VATT and the Large Binocular Telescope (LBT) with twin 8.4 m mirrors), Mount Hopkins (6.5 m MMT), Kitt Peak (2.3 m Bok) and the Catalina Mountains (2 $\times$ 1.5 m) in Arizona, and on Las Campanas, Chile (Magellan Project; 2 $\times$ 6.5 m). The LBT, Magellan, MMT, SMT and VATT are all collaborative projects. The SO components of these facilities are used by scientists and students at the UA, Arizona State University and Northern Arizona University.

SO is also involved in space astronomy research, currently including HST, ISO and SIRTF. It is participating in development programs for NGST and TPF of the NASA Origins program. SO also operates the Mirror Laboratory, located under Arizona stadium.

SO research programs focus on optical, infrared and submillimeter astronomy, on theoretical astrophysics, and on astronomical instrumentation. Major areas of current activity include: cosmology; the intergalactic medium; quasars and active galactic nuclei; dynamics, formation and evolution of galaxies; star formation in galaxies, including starburst galaxies; star formation and the interstellar medium; supernovae; degenerate stars; cataclysmic variables; brown dwarfs; and extra-solar planets.

SO has 25 teaching staff, about 25 further research scientists and a total staff of around 250. SO is supported by State of Arizona funding and by research grants and contracts.

For further information see
http://www.as.arizona.edu/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stingray Nebula (Hen 1357)

A planetary nebula in the constellation Ara, position RA $17^{h}\ 16.4^{m}$, dec. $-59^{\circ}\ 29'$. It is the youngest planetary nebula known, having formed possibly as recently as 200 years ago.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stokes, George Gabriel (1819–1903)

Born in Skreen, County Sligo, Ireland, became Lucasian Professor at Cambridge, best known for his work on hydrodynamics and *Stokes' law* of viscosity. In 1854 he explained the Fraunhofer lines in the solar spectrum as being caused by atoms in the outer layers of the Sun absorbing certain wavelengths, but when KIRCHHOFF later published the same explanation Stokes yielded priority to his more complete account.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stony Meteorite

A meteorite having a composition primarily of silicates, with only a small (typically 5%) metal content, mostly nickel–iron. Stony meteorites are also referred to simply as *stones*; an obsolete name is *aerolite*. Stones are by far the commonest type of meteorite, accounting for nearly 93% of all known specimens. Their dark fusion crust (where the outer layers have been partially melted by frictional heating during their plunge through the Earth's atmosphere, then resolidified) gives them a resemblance to terrestrial rocks. There are two main classes of stony meteorite: *chondrites* (96% of all stones) and *achondrites* (4%). Chondrites consist of an agglomeration of millimeter-sized *chondrules*, particles believed to represent primordial material that condensed from the solar nebula. The rarer achondrites, as the name implies, lack chondrules, implying that they are fragments of a parent body that has undergone differentiation.

See also: achondrite, carbonaceous chondrite, chondrites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Stony-iron Meteorite

A meteorite composed of roughly equal proportions of silicate (stony) material and metals (mostly nickel–iron); formerly called a *siderolite*. Stony-irons are the least common of the three main types of meteorite, accounting for less than 1% of all known specimens. The great majority of them fall into either of two main classes: the *pallasites* and the *meso-siderites*. Pallasites consist of a matrix of nickel–iron, the gaps in which are filled with olivine (magnesium–iron silicates), and are inferred to have formed near the core/mantle boundary of a differentiated body. Meso-siderites resemble breccias: they consist of fragments of eucrite and howardite (varieties of achondritic material) mixed with nickel–iron, and are thought to have been produced by impacts on the surfaces of planetary bodies. The largest known single fragment at 1.4 tonnes is the Huckitta (Australia, 1924) meteorite, which is a pallasite. The fragments of the Brenham (Kansas, 1882) meteorite, also a pallasite, that were collected had a total mass of 4.3 tonnes.

See also: achondrite, breccia.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Strömgren, Bengt (1908–87)

Danish astronomer, son of a former director of the Copenhagen University Observatory, succeeded his father, but was also director of Yerkes and McDonald Observatories. Strömgren calculated what happened when a hot star emitted ultraviolet into the gas clouds around it, ionizing the hydrogen in a region now known as a *Strömgren sphere*, or the H II region. He created the *Strömgren photometric system* measuring the brightness of stars in defined spectral bands as a quantitative counterpart to the Morgan–Keenan spectroscopic classifications system. With it he determined the sizes, luminosities, etc of the stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Streete, Thomas (1622–89)

Born in Cork, Ireland, became an English government clerk, wrote *Astronomia Carolina* in 1661, which was important in disseminating Keplerian astronomy in England.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

String Theory/Superstring Theory

A theory of fundamental particles and forces in which the basic entity is an exceedingly short one-dimensional structure (a 'string') rather than a point-like particle. These strings are envisaged as being about 10^{-36} m in length, different modes of vibration of the strings corresponding to different types of particle with different energies and masses. When the ideas of supersymmetry are applied to string theory, the outcome is superstring theory. Superstring theory requires a ten-dimensional spacetime, all but four of these dimensions (length, breadth, height and time) being hidden in the present-day universe. Superstring theory is one of the prime candidates for a 'theory of everything', which embraces all of the forces and particles of nature.

See also: fundamental forces, fundamental particles, spacetime, supersymmetry.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Structure and Evolution of the Sun and Stars, 1800–1955

Evolution by contraction

The idea that stars formed through gravitational contraction of luminous fluids was put on an observational basis by William Herschel (see HERSCHEL FAMILY), who worked very much in the spirit of late eighteenth century naturalism that began to ask questions about the origin and growth of systems. Whereas Kant, Laplace, Buffon and Hutton tended to focus on the origin of the solar system and the earth, Herschel considered stars and stellar systems, a 'natural history of the heavens'. Observing some 2500 NEBULAE, by 1811 he published a classification that described a continuous sequence: diffuse patches condensed and circularized into distinct nebulosities with stellar centers, and eventually into true stars. In Herschel's cosmology, nebular fluids were active and self-luminous.

Aside from the question of the origin of nebular light, stars were believed to shine because they were hot and so naturally questions were raised about the source of that heat, and its maintenance. By the middle of the nineteenth century, lingering speculation about the permanency of the stars was put to rest by one of the consequences of the rise of thermodynamics: the law of energy conservation. The gravitational potential of a fluid mass was itself a source of mechanical energy that was converted into light and heat as the mass contracted. J J Waterston, J R Mayer, William Thomson (Lord Kelvin) and Hermann von Helmholtz all applied this concept to the origin and maintenance of solar heat, and showed that it provided an abundant source of energy greater than any other. Though inadequate to sustain the Sun's output over time scales demanded by geology and biology, gravitational contraction provided the conceptual framework within which astronomers thought about the origin and lives of stars for most of the following century. The term 'STELLAR EVOLUTION' came to describe this process of change from an indefinite to definite form, in the spirit of Herbert Spencer's *First Principles*, although the term 'evolution' itself had been applied to celestial processes as early as the late 1840s by J P Nichol.

During the latter half of the nineteenth century, astronomers accepted the Kelvin–Helmholtz theory of gravitational contraction only in its generality as the source of the Sun's heat, finding the 1870 speculative solar model developed by Washington patent clerk J Homer Lane—based upon thermodynamics, the gas laws, and a theory of convective equilibrium derived from Espy—to be a 'remarkable fact' that might have applicability, depending upon the state of the Sun's constitution. In the 1870s and 1880s, Lane's ideas were developed and refined independently by August Ritter, a professor in Aachen, into a theory of stellar evolution that described how the temperature in a glowing globe of gas varied with radius from the center to the surface of the star. Ritter's mathematical treatment envisioned stars beginning their lives as vastly extended low-temperature gas spheres which heated upon contraction into bright blue and white stars to the point where they became too dense to maintain the perfect gas state. From there on, subsequent gravitational contraction was accompanied by a general cooling of the sphere to the point of extinction. Ritter published some 18 papers in 6 years to produce a comprehensive study of the nature of stars. His work remained obscure, however, until revived by Thomson and others in the 1890s.

In 1888, admitting 'no certain knowledge' about the Sun's structure, solar specialist C A Young felt the Sun had to be gaseous, because of its low mean density, following somehow the laws of Boyle, Dalton and Gay-Lussac. By the end of the century, the source of the Sun's heat remained gravitational, which was not a problem for astronomers, though geologists and biologists found it seriously deficient.

Late nineteenth century scenarios

Ritter's theory of stellar evolution was superficially similar to one created by the astronomical spectroscopist NORMAN LOCKYER at about the same time, and called by him the 'meteoritic hypothesis'. Influenced by the coincidence of a NOVA in the constellation of Corona Borealis in May 1866, the first to be studied spectroscopically, and by a spectacular display of (the now called Leonid) meteors in November, Lockyer sought out a connection between the two, ultimately seeing meteors as a primary mechanical agent of change in stellar evolution. Instead of a mathematical theory of convection, however, Lockyer, the founder and editor of *Nature*, relied upon what he believed was an empirical correlation between spectral features seen in vaporized laboratory samples of meteorites and other substances, and the spectra of nebulae and stars. From his experiments, Lockyer concluded in the 1870s that heat could break down atoms and molecules into more fundamental 'basic' components, and he used this as an explanatory framework to interpret the differences seen in the spectra of the Sun and stars as due to differences in temperature. By the 1880s, he had also concluded that nebulae and nebulous stars exhibited spectra indicating they were composed of incandescent swarms of colliding and vaporizing meteoritic material, and from this constructed a life history for stars which began with nebulous meteoritic swarms, condensed and heated from red to being yellow, blue and white stars as the meteoritic material was consumed, and, having consumed all meteoritic fuel, finally cooled back to dull red stars and to extinction. Lockyer presented his theory in what has come to be known as his temperature arch (figure 1).

Ritter's, and especially Lockyer's, scenario of a double-value temperature history for stars ran against conventional views of the direct descent of hot stars from nebulae, and their subsequent cooling to extinction. H C VOGEL and WILLIAM HUGGINS championed the dominant view, using as evidence the linear sequence of stellar spectra of four major classes which were eventually refined to seven classes by the vast spectroscopic classification project at

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

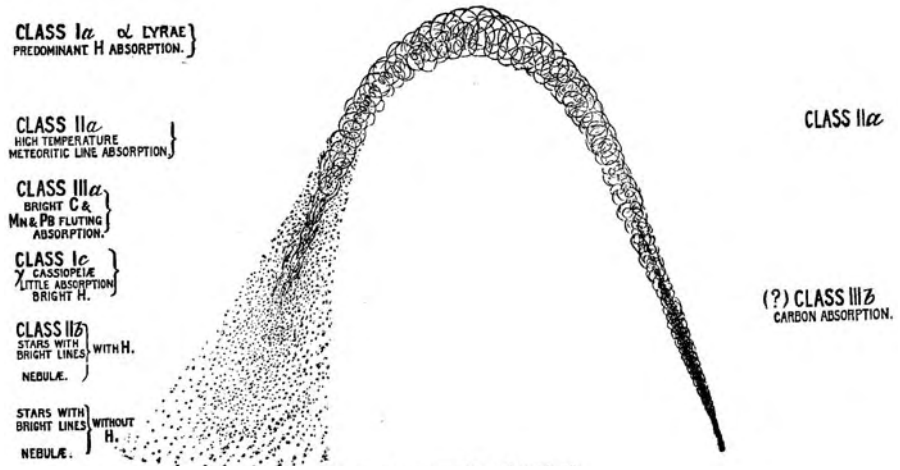


Figure 1. Lockyer's provisional temperature arch. (Figure 67, from Lockyer–1890 *The Meteoritic Hypothesis* (London: Macmillan) p 375.)

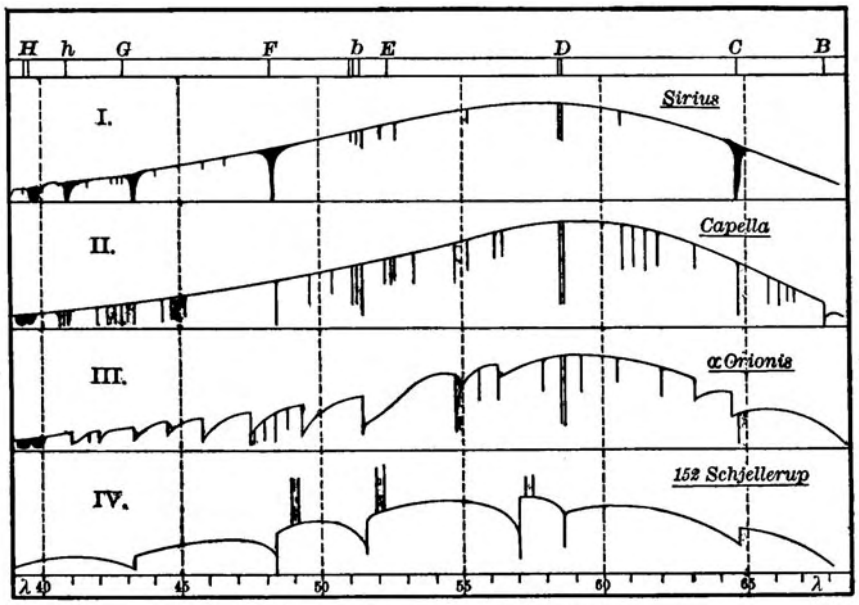


Figure 2. A Secchi's linear sequence of spectra, from blue to red showing the growth of lines. Secchi's and Vogel's similar classifications represented the dominant view of a linear cooling history for stars. (Figure 223, from Young C A 1898 *General Astronomy* revised edn (Boston: Ginn) p 532.)

Harvard in the 1890s (figure 2). Although presented as an empirical classification originally based upon the strengths of the H and K lines of calcium and on the Balmer series of hydrogen, the Harvard system (O, B, A, F, G, K, M) was also a continuous progression of color from white to red. The strengths of the hydrogen lines peaked at Harvard spectral type A, however, because the alphabetical lettering sequence was altered in the 1890s to place stars whose spectra most closely resembled nebulae at the head of the series (O and B). This was clearly an evolutionary consideration.

Huggins and Lockyer engaged in many disputes over their work, and none was as hot as that of stellar evolution. Privately, however, Huggins claimed that he was not against the idea of a double-valued temperature arch, only that he wanted it to be based upon criteria other than those Lockyer was promoting. In 1898, he privately admitted to colleagues that something like a temperature arch based on Lane's law would surely be found someday, but not based upon Lockyer's criteria.

At the turn of the century, then, two different scenarios for the course of stellar evolution were at play. One argued (Lockyer, Ritter) that two classes of red stars (types III and IV on Vogel's system) were widely different in age, and the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

other (Huggins, Vogel) that they were not. Astronomers generally accepted the latter view of the descent of blue stars from nebulae because there was dynamical evidence supporting it. J C KAPTEYN'S detection of two star streams implied that different groups of stars had different origins, and hence that studies of stellar development could somehow be informed by kinematical studies. Correlating the dynamical behavior of groups of stars to evolutionary models was an interest shared by Lick Observatory's W W CAMPBELL, who had already put much of his telescope time and manpower to the task of determining RADIAL VELOCITIES. Less interested in star streaming than in data gathering, Campbell knew that the observed proper motion of a star was statistically an inverse function of its distance, while its radial velocity was independent of distance, and therefore he was on top of one of the most powerful distance-independent techniques for understanding the origins of stars. Starting in 1901, Campbell searched for correlations between kinematics and spectrum class, and by 1910 had found the same increase in velocity with Harvard spectral type that Kapteyn had found. Although a minor priority dispute ensued between the two, the correlation of velocity with age seemed to be an established fact. Since the correlation was linear, this lent strong support to linear descent theories.

Stellar evolution as an agenda

The evolution of the stars, linked closely to the 'sidereal problem', the history of the form and nature of the sidereal universe, both astronomers and astrophysicists claimed at the turn of the century, were driving forces propelling their industry. GEORGE ELLERY HALE cited the problem as a rationalization to build his observatories, and Kapteyn capitalized on it to cement their relationship as Hale turned to the Dutch astronomer to help him define much of the stellar astronomy program at his newly opened observatory atop Mount Wilson. As Hale told Kapteyn in 1907, 'many results obtained in the course of your investigations will be indispensable, if we are to draw any joint conclusions regarding stellar development'. (Hale to Kapteyn, 14 March 1907, Hale Papers, California Institute of Technology, hereinafter 'Hale P'.).

Hale long knew that the pursuit of the deepest problems in astrophysics, specifically the origins and evolution of the Sun and stars, required many talents. On the occasion of thinking through funding priorities for American astronomy in 1906, Hale told SIMON NEWCOMB that astronomers were too specialized and had to be encouraged to move away from 'the details and the merely technical elements of investigation'. (Hale to Newcomb, 21 March 1906, Hale P.) By 1908, Hale justified his research program at his third observatory, the new Mount Wilson Solar Observatory, in terms of stellar evolution, attacking the problem 'along converging lines, leading from solar, stellar, and laboratory investigations'. (Hale, *The Study of Stellar Evolution*, x.)

Huggins' prediction came to pass in the work of HENRY NORRIS RUSSELL of Princeton. Russell had trained under C A Young, who was partial to the theories of Lane and Lockyer, and also became fascinated with the origin and history of double star systems, taking a year of postdoctoral training under George Darwin at Cambridge in 1902 and 1903. Russell saw DOUBLE STARS as the means to explore stellar evolution, mainly in providing a way to determine stellar densities—an indicator of relative age—and set about to find a way to test Lockyer's scheme. Coincidentally, he was hired by the Cambridge University Observatory to assist A R Hinks in a photographic parallax program, and saw in this work another opportunity to perform the test. Parallaxes yielded absolute magnitudes, and by 1908 Russell realized that two very different types of red stars existed in space: small dull ones, and huge luminous ones.

From 1909 through 1913, Russell presented his results in various ways, highlighting the existence of the two classes of red stars, and how they were linked in an evolutionary scheme. Though he initially claimed to the contrary, he was probably aware that he was not alone in this realization. He was the first, however, to use direct means (parallaxes and binary studies) to deduce the existence of giant suns. EJNAR HERTZSPRUNG, a Danish photochemist turned astronomer, had also found that two distinct luminosity classes existed. Prompted by the spectroscopic classification devised by Antonia C Maury at Harvard, Hertzsprung realized that a peculiar subtype, stars exhibiting sharp hydrogen lines and stronger metal lines than their counterparts on the alphabetical Harvard system, all had very small proper motions, yet many were among the brightest stars in the sky. This led him indirectly to the idea that a collateral series of stars existed, and at first he interpreted his findings in an evolutionary scheme. Unlike Russell, he soon dropped any speculation about the evolutionary place of his 'giant' suns.

Russell's 'giant-to-dwarf' theory of stellar evolution

Russell started publishing his findings in 1909 and by August 1910, at a meeting of the Astronomical and Astrophysical Society of America at Harvard, Russell expanded on his theory of stellar evolution, for the first time relating it to Lockyer's, noting that he had rederived it by totally independent means. He also gave it a physical interpretation using the convective models of Lane, Ritter and Emden. Russell's 'giant-to-dwarf' theory envisioned stars beginning as vastly bloated giants in the perfect gas state, contracting under gravity and heating, attaining a density where, deviating now from the perfect gas state, they cooled under continued gravitational contraction. Unlike the dominant linear temperature model of stars descending from nebulae by contraction and cooling, Russell's theory, following Lockyer, viewed stars passing through the observed spectral series twice: once heating, then cooling. Russell also linked his findings to Darwin's model for binary fission. With his first graduate student, HARLOW SHAPLEY, Russell confirmed that what distinguished giants from dwarfs was density, not mass, using eclipsing

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

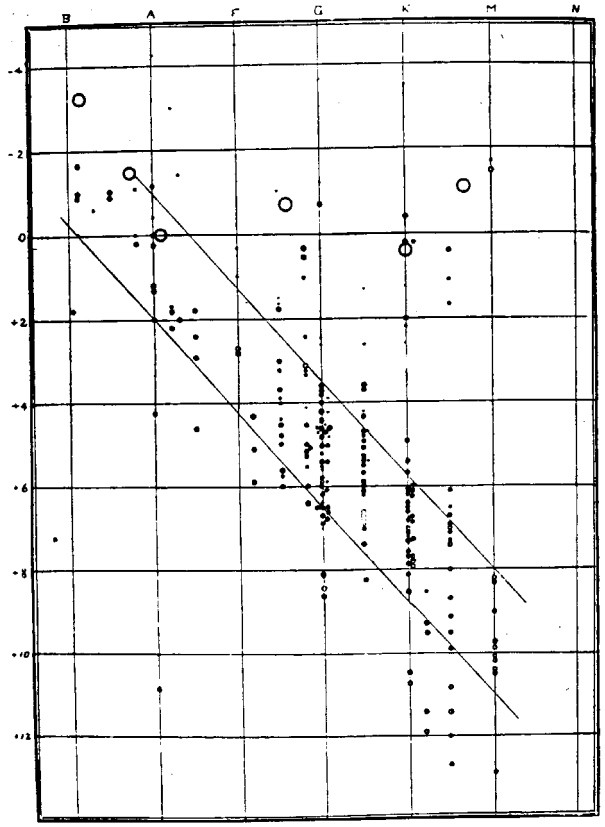


Figure 3. Russell's first diagram based upon direct parallax measurements. The sizes of the dots and circles indicate different ranges of probable error. (Figure 1 from 'Relations between the spectra and other characteristics of the stars' *Popular Astron.* **22** (1914) 285.)

binary data and a new and highly efficient mode of data reduction Russell devised. These efforts culminated in a series of talks, and finally in his famous paper which was presented in 1913 and 1914 in London and then Atlanta. At this time he also devised his version of what has come to be known as the HERTZSPRUNG–RUSSELL (H–R) DIAGRAM, which he presented as the embodiment of his theory of stellar evolution (figure 3).

Russell, however, had a lot of convincing to do. Relative ages deduced from the old linear theory fit nicely Kapteyn's and Campbell's observation of increasing velocity with spectral type, as well as with the observation that the periods of close binary star systems increased with spectral class (and hence age), and with the argument that stars formed in the plane of the Milky Way and then migrated away. Giant suns of late spectral type, the youngest stars on Russell's theory, were found at all galactic latitudes, and the kinematics of both giant and dwarf red stars were similar, and generally larger than those of the earlier spectral classes. In short, Russell's model played havoc with Kapteyn's neat dynamically evolving system. ARTHUR STANLEY EDDINGTON, then at Greenwich, knew this, but also realized that Russell had discovered something fundamental in his diagram. He was therefore anxious to reprint a sample and discuss its implications in his own studies of the structure and evolution of the sidereal universe. He also warmed to the existence of giant stars when he found that he could describe them very nicely using his theory of radiative equilibrium, which he developed over a period of years starting in 1916.

Belief in the existence of giant stars was strengthened with the discovery at Mount Wilson by Walter Sydney Adams and Arnold Kohlschütter of a luminosity effect in spectra, another example of empirical correlation. Although they were asked by Kapteyn to search for interstellar absorption, what they found was that some spectral lines were sensitive indicators of luminosity, a confirmation of what Maury had found and Hertzsprung exploited. They took the next step, however, and used this correlation to develop their spectroscopic parallax technique, and in so doing confirmed the existence of high-luminosity red stars, or giants.

Russell, Adams and others remained concerned, however, about the lack of a rational physical basis for all these empirical correlations between luminosity, spectrum and mass. How closely did stars approximate black-body radiation? What really governed the changes seen in the spectral classes of the stars: temperature, pressure, density, composition? By 1920 Russell's theory of stellar evolution enjoyed greater popularity as Eddington championed giants as perfect gas-spheres in radiative equilibrium.

Eddington's radiative models

Eddington was not an immediate convert. He initially favored the evolutionary interpretation of Kapteyn's star streaming, and worried that Russell's 'giant-to-dwarf' theory destroyed the nice association of 'early type' white and blue stars with nebulae, as the old one could. In 1916, reconsidering the convective gas theory of Emden, partly to explore the behavior of globular clusters, but also to explore the nature of pulsating variables then suggested by Harlow Shapley, Eddington decided that Ritter's (and hence Russell's) theory of stellar evolution could be put on a rational basis using radiative equilibrium to describe the interiors of giant stars, assumed to be in the state of a perfect gas. Studying both the effects of transfer of energy by radiation and the pressure of radiation required a far more sophisticated analysis of the behavior of polytropic gas spheres as radiative engines than Emden had provided. In a series of papers in 1916 and 1917, Eddington established his radiative theory, and showed that he could describe giant stars as they were distributed on Russell's diagram; in effect, he found that the luminosity of a giant star was independent of its stage of evolution, as long as it remained in the perfect gas state. This nicely described Russell's giant branch (figure 3). Eddington's subsequent attention to the theory of polytropes became the core of physical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and mathematical studies in stellar structure for decades, propelled by his seminal text *The Internal Constitution of the Stars* which appeared in 1926.

Much of Eddington's work was initiated using little more physics than the standard gas laws and the concept of local thermodynamic equilibrium developed by Kirchhoff; as a case in point, he was constantly frustrated by his lack of knowledge of the mechanism of absorption, embodied in the opacity of the gaseous stellar interior. Absorption coefficients derived from Eddington's models based upon astrophysical data differed widely from theoretical values derived by the physicist H A Kramers.

Russell's theory of stellar evolution meshed nicely with Eddington's radiative giants at first, and gave both great popularity in the early 1920s. Probably the high point in their confidence came in 1919, when Russell realized how Eddington's radiative stars, obeying Lane's law and possessing some form of explosive radioactive energy source, could still be stable. Since the specific heat of a mass of gas obeying Lane's law is negative, the more heat it loses, Russell reasoned, 'the hotter it gets; and the more heat that is added to it internally, the colder it becomes!' Stars were, therefore, 'automatically self-limited'. (Russell to Shapley, 5 May; to Webster 9 June 1919, Russell Papers, Princeton.)

Russell's theory fitted Eddington's models until the end of 1924, when Eddington realized, in consequence of his derivation of the theoretical mass–luminosity relationship, that dwarf stars could be in the perfect gas state as well as giants. Astronomers had long known that there was a relationship between the masses and luminosities of stars, but Eddington was the first to show why, on a rational basis. If dwarfs stars were also in the perfect gas state, however, then they would continue to heat upon contraction, and Russell's theory was untenable. The only way to get around this conundrum was to have stars lose mass as they aged.

Time scales, composite models and energy sources

One of the major considerations that significantly altered thinking in the 1920s about the lives of stars and the nature of their structure was the realization some years prior that the gravitational energy of the Sun was woefully inadequate to provide a lifetime of energy output required by geology or biology. Eddington, Russell and others considered two possible mechanisms for sustaining the energy output of the Sun: the annihilation of matter and the transmutation of hydrogen into the heavier elements. Although the latter was compelling for Eddington, mainly because of the careful studies by Aston showing that atomic masses were nearly integral, and by Rutherford's transformation of nitrogen into oxygen, both he and Russell were drawn to the former, because it offered a means to let stars lose mass as they aged. Later in the 1920s, to further bolster his 'giant-to-dwarf' theory, Russell invoked two sources of nuclear energy, one occurring in giant stars and one in dwarfs.

When Hubble's velocity–distance relationship limited the age of the universe to an intermediate time scale of billions of years, the trillions of years of stellar energy provided by matter annihilation was no longer required. An intermediate time scale was also urged by Bart Bok's studies of cluster lifetimes, and, by the mid-1930s as well, with hydrogen accepted as the most abundant element in stellar atmospheres and interiors, and refined knowledge of isotopic masses by Aston confirming that element synthesis for hydrogen would work, that mechanism dominated over matter annihilation. By the end of the decade, with improved knowledge of the structure of the nucleus, physicists Hans Bethe and Carl F von Weizsäcker, a student of W Heisenberg, solved the stellar energy problem by developing the proton–proton and carbon–nitrogen–oxygen cycles.

During the same two decades that the nuclear energy source was sought after, astronomers came to realize that even without knowing the specific mechanism, realistic stellar models had to account for such a source acting in the STELLAR INTERIOR. Some astronomers, including Eddington, argued that fusion processes took place throughout stars, since any star in which the energy generating process depended too strongly on temperature would be unstable. But the trend in the 1930s, led by theorists such as S Rosseland and Ludwig Biermann, a student at Göttingen, was to concentrate the nuclear power source at the stellar center, removing energy by convection. The use of composite polytropic gas-spheres, developed in the late 1920s and early 1930s by theorists such as Thomas G Cowling, E A Milne, Bengt Strömgren and Russell, allowed stars to be described by a series of concentric polytropic distributions. Various solutions existed—one devoid of singularities, a centrally condensed solution, or a collapsed [degenerate] solution which could be fitted together to fit observed characteristics of stars. This new versatility eventually paved the way for Cowling, a student of Milne, to construct a stable point-source model. By the mid-1930s, utilizing a small convective core surrounded by a radiative envelope, his model became the framework within which physicists von Weizsäcker, Bethe, George Gamow and Edward Teller examined nuclear processes in stars, employing the idea that hydrogen was transformed into heavier elements.

The successful search for the source of stellar energy brought nuclear physicists into astronomy. GEORGE GAMOW, Teller, Charles Critchfield, Robert Marshak and others became interested in this new application. Gamow pursued it passionately, writing constantly and excitedly to astronomers requesting observational data, and so became a bridge between the professional worlds of astrophysics and nuclear physics. His greatest contribution was, quite possibly, as host of a conference in Washington, DC in April 1938 which brought the worlds together. Bethe attended, and within months had worked out the physics to show that von Weizsäcker's proton–proton cycle (four hydrogen nuclei fusing into a helium nucleus) could indeed provide the right amount of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

energy to power the Sun, and that his own 'CNO' cycle, where carbon catalysed four protons to fuse into a helium nucleus, could also account for the greater luminosities of massive main sequence stars.

The landmark contributions of Bethe and von Weizsäcker had immediate impact. Matter annihilation was no longer in vogue, which made the main sequence a locus of equilibrium points, not an evolutionary track. And astronomers could opt for an 'intermediate' time-scale on the order of 10^{10} years, which agreed reasonably with the cosmological and geological time scales. The extreme temperature dependence of these new fusion mechanisms supported Cowling's refinement of the point-source model for stellar structure. Energy generation acted as if it were limited to the centers of stars, and was not distributed throughout the stellar volume. The hydrogen fusion process also required that hydrogen be a significant constituent of stars.

The place of giants in the aging process

During the 1930s, the first suggestions that RED GIANTS could be old evolved stars rather than young stars appeared. In 1932 and 1933, Strömgren explored the theoretical distribution of stars on the H–R diagram, comparing it with observational data for masses, spectra and composition. He found that the position of any star on the diagram was determined solely by its mass and hydrogen abundance. BENGT STRÖMGREN then looked to see how a star's position on the H–R diagram would change as its hydrogen content diminished, assuming with others that hydrogen was expended in some yet unknown nuclear process to power the stars. He found that evolutionary tracks lay across lines of equal hydrogen abundance, and as hydrogen was consumed, stars would 'leave' the main sequence and head up and to the right, or toward the region of the giant red stars. He stated explicitly that as stars aged, they would expand (figure 4).

In the late 1930s GERARD KUIPER and Strömgren were hired by Otto Struve to the Yerkes Observatory staff. Kuiper, primarily a double star man, was also very interested in re-examining the H–R diagrams of galactic clusters that Robert Trumpler had produced. Stimulated by Strömgren's theoretical work on diminishing hydrogen content, Kuiper placed many clusters on the same diagram, and found that their main sequences coincided at lower luminosities and diverged for the higher luminosities. The divergence was systematically correlated with the consumption of hydrogen and all the divergences led to the giant region. Kuiper only noted that these divergences were somehow clues to evolutionary change. Nor did he discuss explicitly the direction of evolution, as Strömgren had before him. There was simply not enough physics available, and observational data for masses and abundances were far from consistent. Kuiper, an observational astronomer, was not able to create a convincing picture.

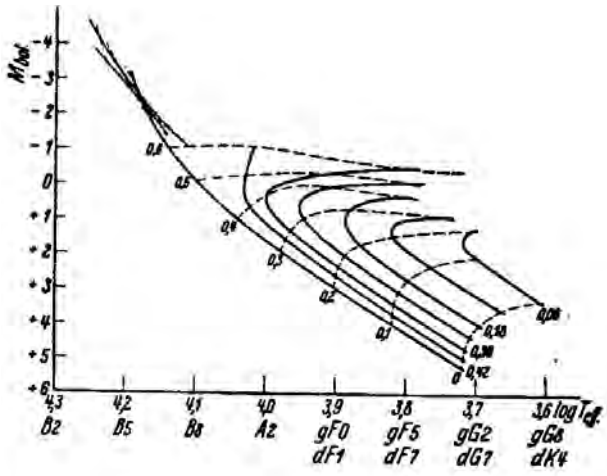


Figure 4. Theoretical distribution of stars by constant hydrogen content (full curves) and by constant mass (broken curves). (From figure 4 in Strömgren B 1933 'On the interpretation of the Hertzsprung–Russell-diagram' *Z. Astrophys.* **7** 244.)

At the end of the decade and into the early 1940s, continued exploration of various inhomogeneous stratified models, first suggested by ERNST J ÖPIK, led Gamow, S CHANDRASEKHAR and others to consider what happens when hydrogen is exhausted in the stellar core, and hydrogen fusion commences in a shell surrounding the core. Skeptical about Gamow's work, Chandrasekhar, with Louis R Henrich and Mario Schönberg, developed true hydrogen shell burning models. These models demonstrated for the first time that changes in composition were related to some type of evolution away from the main sequence. Out of this came the famous Schönberg–Chandrasekhar limit, which placed an upper limit to the size of the exhausted core (not to be confused with Chandrasekhar's earlier study of the limiting masses of white dwarfs).

In the post-World War 2 era, the continued application of nuclear physics to the problem of stellar energy, along with increasingly versatile models made possible by improved methods of numerical computation aided by high-speed computers, led to an improved ability to describe stars in all stages of stellar evolution, from red giants to white dwarfs. Along the way, several fundamental adjustments in theory and technique aided the process of change. In the 1920s and early 1930s, analytical methods of modelling still dominated in theoretical astrophysics in the manner Eddington established them. But by the mid- to late 1930s, younger theorists in Europe and the United States started to build stellar models by numerical techniques making its easier to fit together composite polytropic layers. Furthest from the analytical tradition so strong at Cambridge, Russell always preferred numerical calculations, whereas Chandrasekhar, breaking tradition with his mentors Milne and Eddington, used both numerical and analytical techniques.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In contrast to Eddington, physicists such as George Gamow actively sought out astrophysicists to engage them in the search for solutions to major questions in stellar structure and evolution. After Bethe and von Weizsäcker established the hydrogen transmutation processes active in stars, and Strömgren provided astrophysical confirmation of the importance of hydrogen content, it was Gamow, Chandrasekhar, Baade and then MARTIN SCHWARZSCHILD who opened avenues of cross-disciplinary communication to establish the intellectual infrastructure out of which the modern picture of post-main sequence stellar evolution would emerge in the 1950s. As a first step, a theoretical group led by Schwarzschild in Princeton (where Lyman Spitzer and John Wheeler were close at hand), and an observational group led by Walter Baade at Mount Wilson, made a combined attack on the relative ages of star clusters, equating their appearance on the H–R diagram to theoretical models of stars with exhausted hydrogen cores. The end result in the mid-1950s was that giant stars were no longer thought to be stars in pre-main sequence formation, but were post-main sequence stars with hydrogen-exhausted cores gaining power from hydrogen burning in shells surrounding their cores. Two major problems had to be cleared up, however, before the framework of the picture was complete.

In April 1952, Cornell nuclear physicist EDWIN SALPETER, who had recently become attracted to the field while in residence at Caltech working with William A. Fowler, Jesse Greenstein and Baade, reviewed concepts of stellar energy and evolution, and found the latter in a general state of confusion regarding where white dwarfs, red giants and intrinsic variables were situated in the evolutionary process. Indeed, what maintained red giants in their vastly bloated state of existence? All stars that had exhausted the supply of hydrogen available for fusion, assumed to be rotating and in a well-mixed state, were still thought to resume evolving somehow by contraction. There were other problems too, such as how to form the heavy elements. Bethe and von Weizsäcker's hydrogen fusion mechanisms, the proton–proton and CNO cycles, could only go so far, and neither of them, nor a speculated process of neutron capture by helium in the BIG BANG, postulated by Alpher, Herman and Gamow in the 1940s, could produce stable elements of atomic mass 5. In 1950, Enrico Fermi and A Turkevich at Chicago showed that no elements heavier than helium could have been produced in the Big Bang, making element synthesis the domain of the stars. Soon after, Öpik and Salpeter independently solved the mass 5 problem, showing that the fusion of three helium nuclei could form carbon. They also established that helium fusion was capable of acting as a new source of energy after hydrogen exhaustion.

Salpeter's contribution has been called the 'triple-alpha process' wherein helium is transmuted into heavier elements. With it, astrophysicists could begin to discuss what is essentially the second step to the modern picture of stellar evolution, where a poorly mixed star consuming hydrogen becomes inhomogeneous as helium is produced in its core and it becomes incapable of burning more hydrogen. When the core grows to some 10 per cent of the stellar mass, it becomes unstable. The unstable core contracts and heats causing the remaining stellar envelope to expand and cool, turning a main sequence star into a red giant. Mario Schönberg, Chandrasekhar, Öpik, A Sandage and Schwarzschild all contributed to this modern picture by 1955, which highlighted the importance of solving the problem of how elements beyond iron could be synthesized in the cores of stars. This was solved by a collaboration of an observational astrophysicist, two theoretical astrophysicists, and a nuclear physicist: Margaret Burbidge, Geoffrey Burbidge, Fred Hoyle and William Fowler (known as 'B^2FH'). Spurred on in 1956 and 1957 in their pursuit of an alternative to the Big Bang as a site for the origin of the elements—in the centers of stars as opposed to the original big bang—they developed a set of neutron capture processes for building the heaviest elements in stars, and inaugurated the modern study of nucleosynthesis. In the decades since 1955, astrophysicists have been working out the details, mainly with the aid of high-speed computing machines which have made numerical techniques highly effective tools for the study of stellar structure and evolution.

In 1951, as his part in surveying a century of science, W H MCCREA argued that the present state of the study of the constitution and evolution of the stars was possibly:

> ... the crowning achievement of the century in physical science. For it has explained the properties of a star in terms of those of the fundamental particles of laboratory physics, and in doing so it has employed every fundamental advance in pure physics and has depended upon observations made with the aid of all the resources of physical technology. (McCrea 1951, p 136.)

McCrea's point was, however, to argue that a history of the subject is difficult to treat in concise fashion, as it requires many diversions to appreciate the necessary advances on many fronts lurking in the most arcane corners of physics and technology. McCrea, a theorist and a younger member of the mathematical physics circle centered on Eddington, Milne and R H Fowler, considered mainly theoretical tools such as quantum and nuclear physics. But the tools of the astrophysicist include the techniques of the spectroscopist in the laboratory as well as at the telescope, and, more recently, in the post-World War 2 era, the expertise of electronics, both in the development of photoelectric and solid-state devices, in radio astronomy as well as in the multitude of non-visible astronomies, and in nuclear physics, as practised by academic theorists, experimentalists and military laboratories. Only when all these facets of the problem are addressed by historians and astronomers alike will a satisfactory history emerge.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Burbidge G R and Burbidge M 1958 Stellar evolution *Handbuch der Physik 51* ed S Flügge, pp 134–295

Burchfield J D 1975 *Lord Kelvin and the Age of the Earth* (Chicago)

Cowling T 1966 The development of the theory of stellar structure *Q. J. R. Astron. Soc.* **7** 121–37

DeVorkin D H 1984 Stellar evolution and the origins of the Hertzsprung–Russell diagram in early astrophysics *Astrophysics and Twentieth-Century Astronomy to 1950* ed M A Hoskin and O Gingerich (Cambridge: Cambridge University Press) pp 90–108

DeVorkin D H and Kenat R The changing place of red giants in the evolutionary process *General History of Astronomy 4b* ed M A Hoskin and O Gingerich (Cambridge: Cambridge University Press) at press

Hufbauer K 1981 Astronomers take up the stellar energy problem, 1917–1920 *Hist. Stud. Phys. Sci.* **11** 277–303

Hufbauer K 1991 *Exploring the Sun: Solar Science Since Galileo* (Baltimore, MD: Johns Hopkins University Press)

Hufbauer K 1999 Stellar structure and evolution (1924–1938) *General History of Astronomy 4b* ed M A Hoskin and O Gingerich (Cambridge: Cambridge University Press) at press

McCrea W H 1951 The Constitution and Evolution of Stars *A Century of Science 1851–1951* ed H Dingle (London: Hutchinson's Scientific and Technical Publications) ch ix

Meadows A J 1972 *Science and Controversy: a Biography of Sir Norman Lockyer* (Cambridge, MA)

Paul E R 1993 *The Milky Way Galaxy and Statistical Cosmology 1890–1924* (Cambridge: Cambridge University Press)

David H DeVorkin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Strutt, John William [Lord Rayleigh] (1842–1919)

Born in Langford Grove (near Maldon), Essex, England, Nobel prizewinner (1904) for the discovery of argon. He worked in many areas of physics, including electromagnetism and sound; the *Rayleigh theory of the scattering of light* was the first correct explanation of why the sky is blue.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve family

Struve, Friedrich Georg Wilhelm (1793–1864); Struve, Otto Wilhelm (1819–1905), son of Friedrich; Struve, Karl Hermann (1854–1920), elder son of Otto Wilhelm, brother of Gustav; Struve, Gustav Wilhelm Ludwig (1858–1920), younger son of Otto Wilhelm, brother of Karl; Struve, Otto (1897–1963) [Otto Struve II], son of Gustav.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve, Gustav (1858–1920)

Born in Pulkova, helped his father at Pulkova Observatory, became director of the Kharkov Observatory in 1894. Determined the velocities of stars and the motion of the Sun through them.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve, Karl (1854–1920)

Born in Pulkova, helped his father at Pulkova Observatory, became director of the Berlin Observatory in 1904, moving it to Babelberg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve, Otto (1897–1963)

Born in Kharkov, Ukraine. In an adventurous and varied career, he studied at Kharkov University, joined the White Russian army during the Revolution, and fled to Turkey at the time of their defeat. Emigrated to the USA, became director of the Yerkes Observatory, near Chicago, and founded the McDonald Observatory in Texas. Became director of the Leuschner Observatory and then the National Radio Astronomy Observatory. He was a stellar spectroscopist, with achievements in close binaries, the interstellar medium and stellar rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve, Otto Wilhelm (1819–1905)

Born in Dorpat, Estonia, assisted and succeeded his father as director of Pulkova Observatory, carrying on his work with binary stars and discovering 500.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Struve, Wilhelm (1793–1864)

Born in Altona, Germany, became professor of astronomy at Dorpat (now Tartu), Estonia, and director of the Dorpat Observatory. Using FRAUNHOFER's refractor he began his work on double stars. He became director of Pulkova (Pulkovo) Observatory near St Petersburg, Russia, which was constructed to his specifications by Tsar Nicholas I and became arguably the most productive observatory of the time in the world. He published a catalog (*Mensurae Micrometricae*, 1837) of over 3000 binary stars, and carried out one of the first determinations of stellar distance, the parallax of Vega (1837), a star that he selected for investigation for this purpose because it had a large proper motion and was bright, thus indicating that it might be nearby. The measurements were not believed, especially since he published a discordant value, twice as large, only three years later. It was not until Bessel measured the parallax of 61 Cygni in 1838, and showed that the measurements contained the expected patterns, that astronomers accepted that a parallactic measurement had been reliably achieved.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Subaru Telescope, Hawaii

The Subaru Telescope, Hawaii is an 8.2 m aperture optical infrared telescope being operated by the NATIONAL ASTRONOMICAL OBSERVATORY, JAPAN (NAOJ) atop Mauna Kea, Hawaii. It is one of the new-generation telescopes with an actively controlled large monolithic mirror, and has been producing testing-phase observational results on solar system, star formation, active galactic nuclei and quasars, group of galaxies, gravitational lensing and cosmology, etc since its first-light in January 1999. The Subaru Telescope aims to explore the universe more deeply and widely with its very high imaging capability and rich observational instruments mounted at its four foci. After nine years of construction since 1991, the Subaru Telescope will start its open-operation in the year 2000. The Subaru Telescope Base Facility at Hilo, Hawaii holds 50 staff, 20 visiting members, a large computer system and laboratories for developmental works.

For further information see http://www.subaru.naoj.org/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Subatomic Particles

Particles that are smaller than atoms. Subatomic particles include the constituent components of atoms and particles that are produced in nuclear reactions or in interactions between other subatomic particles. They are also found among cosmic rays.

Particles are characterized by their physical properties, in particular: mass, charge and spin. Although the masses of subatomic particles may be expressed in conventional mass units (kg), because of the equivalence between mass (m) and energy (E) that is a consequence of the special theory of relativity (i.e. $E = mc^2$, where c denotes the speed of light), the equivalent energy (expressed in electronvolts) is usually quoted as the 'mass' of the particle, even although, strictly speaking, mass (m) $= E/c^2$. The unit of charge is the charge on the electron (1.602×10^{-19} coulomb). Spin is the quantity of angular momentum associated with a particle (the particle may be visualized as spinning around an axis) or an atomic nucleus and takes values given by $nh/2\pi$, where h is the Planck constant and $n = \frac{1}{2}, 1, \frac{3}{2}, 2$, etc.

Particles are subdivided into bosons (particles with n equal to an integer: 1, 2, ...) and fermions (particles with half-integral values of spin: $\frac{1}{2}, \frac{3}{2}, \ldots$). Fermions obey the Pauli exclusion principle, while bosons do not. Among the bosons, the gauge bosons (particles that transmit the fundamental forces) constitute an important group. The fermions may be subdivided into hadrons (particles that are composed of quarks and are acted on by the strong nuclear interaction) and leptons (particles that do not contain quarks and which do not experience the strong nuclear interaction). The hadrons are subdivided into baryons (particles composed of three quarks) and mesons (composed of quark–antiquark pairs). Among the baryons, an important class is the nucleons (protons and neutrons). Electrons and neutrinos are examples of leptons.

The term 'elementary particle' is sometimes applied to particles that have no internal structure and which cannot be broken down into more basic constituent particles. However, because some particles that were at one time considered to be 'elementary' (e.g. protons) are now known to consist of quarks, the term 'fundamental particle' is sometimes used to distinguish those (leptons, quarks and gauge bosons, so far as is known at present) that are genuinely basic particles.

See also: antimatter, atom, baryons, cosmic rays, elementary particles, electron, electronvolt, fundamental forces, fundamental particles, gauge bosons, hadrons, leptons, matter, mesons, neutron, neutrinos, Pauli exclusion principle, photon, proton, quark.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Subbotin, Mikhail Fedorovich (1893–1966)

Mathematical astronomer, born in Ostrolenka, Lomzhinsk Guberniya, Russia (now Poland), worked in astronomy and celestial mechanics at Leningrad (St Petersburg), nearly dying of hunger in the siege of 1942. He produced methods of calculating orbits from three observations based on solving the Euler–Lambert equations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Submillimeter Telescope Observatory

The Submillimeter Telescope Observatory (SMTO), a collaboration of the MPIfR and Steward Observatory, operates the 10 m Heinrich Hertz Telescope (HHT) for observations in the millimeter and submillimeter wavelength range. The SMTO is located at an altitude of 3186 m on top of Mount Graham, Arizona. The HHT has a very accurate surface (12 μm rms) and it is the first telescope to use a hot-electron bolometric mixer receiver for astronomical observations at frequencies of 690 and 810 GHz.

For further information see http://maisel.as.arizona.edu:8080/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Submillimeter Wave Astronomy Satellite (SWAS)

NASA Small Explorer mission, launched December 1998. Carries a 60 cm telescope to study submillimetre emissions from molecular clouds and star-forming regions. Found strong water emission lines in spectra of all molecular cloud targets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sudbury Neutrino Observatory

The Sudbury Neutrino Observatory (SNO) studies neutrinos from the Sun, supernovae and other astrophysical sources. The detector is situated 2000 m underground in an active nickel mine near Sudbury, Ontario, Canada and observes Cerenkov light from neutrinos interacting with 1000 tonnes of ultra-pure heavy water (D_2O) surrounded by 7000 tonnes of light water shielding. Several different neutrino interactions are used to observe the flux and energy spectrum of electron neutrinos reaching the Earth from the solar core; and the total flux of all three neutrino types, to determine if neutrino flavor change (indicative of finite neutrino rest mass) has occurred. The project involves 100 scientists from Canada, USA and UK and began observations in May 1999.

For further information see http://sno.phy.queensu.ca.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Suess, Hans Eduard (1909–)

Geochemist, born in Austria, professor of Geochemistry at UC San Diego, pioneered radio-carbon dating, working on the origin of the elements and the evolution of the solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Suidas (tenth century)

Greek encyclopedist. In the course of reading Suidas's *Lexicon*, EDMOND HALLEY mistakenly connected the naming of the Saros cycle of 223 synodic months by the tenth century Greek lexicographer Suidas with the eclipse cycle of the same period. The solar eclipse cycle is thus now known by the name that Suidas used for another phenomenon. Halley's mistake accounts for the historical confusion that the Saros cycle of solar eclipses was discovered by the Babylonians (it was not).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Suisei

Japanese mission to study Halley's Comet. Launched in August 1985, Suisei approached to within 151 000 km of the comet on 8 March 1986 to observe its interactions with the solar wind. Suisei means 'comet'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Summer Triangle

A conspicuous asterism which dominates the northern sky during summer, comprising the bright stars Vega (α Lyrae, apparent magnitude 0.03), Deneb (α Cygni, apparent magnitude 1.25) and Altair (α Aquilae, apparent magnitude 0.76). Vega, by far the most brilliant star in the northern summer sky, is also the brightest to pass overhead (within 13° of the zenith), with the exception of Capella (α Aurigae) which is a winter object.

The Summer Triangle is crossed by a double band of the Milky Way (the galactic equator passes almost axially through the triangle). There are many rich star fields and interesting objects within its borders, but few bright stars. Only two stars of third magnitude or brighter are found inside the triangle (β and γ Cygni), and three more along its sides. It therefore has a somewhat empty appearance to the unaided eye, but is a rewarding area for examination with binoculars or a small telescope.

See also: Altair, Deneb, Vega.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun

The nearest star to the Earth and central body of the solar system. A main-sequence, yellow dwarf of spectral type G2 V, mass 1.989×10^{30} kg, diameter 1392 000 km, luminosity 3.83×10^{26} W and absolute visual magnitude +4.82, the Sun is the only star whose surface and outer layers can be studied close up. From the Earth, a mere 149 600 000 km away, the Sun's disk subtends an angle of about 30′ and its apparent visual magnitude is −26.78.

The Sun is composed mainly of hydrogen (about 70% by mass) and helium (about 28%), with a small proportion (about 2%) of heavier elements. In its central core, which is thought to be about 200 000 km in diameter, the maximum density is about 148 000 kg m^{-3} (so that almost 30% of the Sun's mass is contained in barely 3% of its volume). Here the temperature is about 15 000 000 K and hydrogen nuclei fuse together to form helium nuclei in the proton–proton (p–p) chain reaction. The energy liberated by the 4.4×10^9 kg of mass lost per second in this reaction is radiated through the radiative zone to within about 200 000 km of the surface, where, the temperature having fallen to about 1000 000 K, convection takes over. The convective zone extends to just below the surface, or photosphere, which is an opaque layer several hundred kilometers thick that represents the boundary between the solar interior and solar atmosphere.

The photosphere has a mottled appearance called granulation, which represents the tops of convection currents from below. Within the granulation dark markings called sunspots appear, surrounded by lighter faculae which are visible well away from the brighter central region of the Sun's disk. Sunspots and faculae mark the position of intense, localized magnetic fields, and their numbers and mean latitude change over an average period of approximately 11 years—the so-called sunspot cycle. Observation of the passage of sunspots across the Sun's disk reveals a differential rotation, the synodic period increasing with heliographic latitude from 26.87 days at the equator to 29.65 days at $\pm 40^\circ$, beyond which sunspots are seldom seen. The mean synodic period is 27.2753 days (sidereal period 25.38 days), which corresponds to the actual period at about $\pm 15^\circ$. Spectroscopic measurements show that the rotation period continues to increase with increasing latitude above $\pm 40^\circ$ and that it decreases with height above the photosphere, except at the equator, where the periods are approximately the same. The reason for this differential rotation is not known.

The temperature of the photosphere decreases from about 6000 K at its base to about 4000 K at the so-called temperature minimum, where it merges with the chromosphere. The chromosphere is a stratum of gases a few thousand kilometers thick, in which the temperature increases (as a result of processes such as mechanical dissipation) to around 50 000 K as the density decreases exponentially with height. It is the realm of cloud-like filaments, or prominences, which have a higher density and lower temperature than their surroundings, and of short-lived explosive events called flares. Sunspots, most filaments/prominences and flares constitute active regions, which are the most obvious manifestations of solar activity and are characterized by enhanced emission at x-ray, extreme ultraviolet and radio wavelengths.

Above the chromosphere lies the transition region, only several hundred kilometers thick, in which the temperature increases again, to around 500 000 K. Beyond this lies the corona, which attains a maximum temperature of about 2000 000 K at a height of about 75 000 km and extends for many million kilometers into the interplanetary medium, where the solar wind carries a stream of atomic particles to the depths of the solar system.

The Sun is believed to possess a weak general magnetic field, but is has proved difficult to disentangle this from the remnant magnetic fields of sunspots. The intense, localized magnetic fields determine the configuration of sunspots, active region filaments and other associated phenomena, the occurrence of short-lived events such as flares, and even the structure of the solar corona.

The age of the Sun is about 4.6×10^9 years and it is expected to maintain its present energy output for at least a similar period, whereafter it will expand to become a red giant and then expel its outer layers, leaving a white dwarf surrounded by a planetary nebula.

See also: active region, chromosphere, corona, differential rotation, faculae, filament, flare, granulation, solar photosphere, solar wind, stellar evolution, Sun and solar–terrestrial physics, sunspot cycle, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun and Solar–terrestrial Physics

The Sun is of central importance in astronomy. On the one hand, many basic processes that occur throughout the cosmos may be well observed in the Sun and, on the other hand, the Sun acts as a source of the radiation and plasma emissions that affect the Earth in many ways. The study of the interaction of the Sun and its solar wind with the Earth is the burgeoning field of solar–terrestrial physics. The magnetospheres of other planets act as probes for the solar wind. However, many of the fundamental processes, such as how the Sun's magnetic field is generated, how its atmosphere is heated, how solar flares occur and how the solar wind is accelerated, are only just beginning to be understood. This article attempts an overview of the Sun and solar–terrestrial physics and links the individual articles in this encyclopedia into the big picture.

The Sun

The Sun is of course the nearest star, seen in greatest detail (SUN: BASIC PROPERTIES). The total radiation (SOLAR IRRADIANCE) from the Sun is not constant but, like a variable star, the Sun varies over a variety of timescales. Treated as a star whose properties are more accurately known than most (SUN AS A STAR), the Sun has become a pivotal point in discussion of the abundance of the elements (SOLAR ABUNDANCES) and stellar evolution (SOLAR EVOLUTION). The SOLAR MAGNETIC FIELD plays a key role in many aspects of Solar Physics. It is measured through polarimetry (SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY) and the Zeeman and Hanle Effects (SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). The solar magnetic field undergoes a cyclic behavior with a period of roughly 22 years (SOLAR CYCLE). Many features of solar activity follow this cycle.

The Sun, indeed most of the Universe, is in the fourth state of matter—the so-called plasma state—in which a gas is so hot that it has become ionized. Its behavior differs from that of a normal gas mainly because it interacts in a complex and subtle way with any magnetic field that is present. Such an interaction is described by the theory of MAGNETOHYDRODYNAMICS, which applies when the ionized particles are much closer than the length-scales of interest, so that the plasma may be regarded as a continuous medium. The magnetic fields are generated within solar and planetary interiors in a self-sustaining way by a process described by DYNAMO THEORY, although the parameter regimes and driving processes are rather different in the GEODYNAMO and in stellar dynamos (DYNAMOS: SOLAR AND STELLAR).

When the force produced on a plasma by a magnetic field dominates all other forces (as is the case, for example, in the upper solar atmosphere), and the magnetic field is in equilibrium with itself, it is said to be a FORCE-FREE MAGNETIC FIELD. Force-free structures can be highly complex and, when they are either twisted or braided around one another, they possess so-called MAGNETIC HELICITY. Equilibrium magnetic fields may be subject to a wide variety of MAGNETOHYDRODYNAMIC INSTABILITIES, which may produce either a large-scale dynamic behavior or a small-scale structuring of the plasma. In the solar interior, magnetic flux tubes tend to be lighter than their surroundings and so rise by an effect called MAGNETIC BUOYANCY. Also, when heated sufficiently from below, a plasma tends to exhibit turbulent motions due to a convective instability, which, when it interacts with a magnetic field, is known as *magnetoconvection* (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION).

In most of the cosmos, magnetic fields are attached or 'frozen' to the plasma and move around with it. The exception is in singular regions where MAGNETIC RECONNECTION can occur and magnetic energy is converted into other forms such as heat, kinetic energy and fast-particle energy. The energy conversion occurs when magnetic field lines are carried into the singularity, broken and rejoined. Several different modes of oscillation (MAGNETOHYDRODYNAMIC WAVES) can be supported in plasmas and are important for transporting energy from one location to another. Several magnetohydrodynamic processes are also crucial in astrophysical circumstances (MAGNETOHYDRODYNAMICS OF ACCRETION DISKS and MAGNETOHYDRODYNAMICS OF ASTROPHYSICAL WINDS).

SOLAR SPECTROSCOPY AND DIAGNOSTICS are invaluable in determining many properties of the plasma such as temperature, density and velocity. The theoretical background lies in atomic physics (SOLAR SPECTROSCOPY: ATOMIC PROCESSES). To access the most complete information about what is happening, astronomers make multiwavelength observations across the whole spectrum (SOLAR SPECTROSCOPY: COHERENT PLASMA EMISSION, SOLAR SPECTROSCOPY: CONTINUUM RADIO EMISSION, SOLAR SPECTROSCOPY: INFRARED EMISSION, SOLAR SPECTROSCOPY: VISIBLE EMISSION, SOLAR SPECTROSCOPY: ULTRAVIOLET AND EXTREME ULTRAVIOLET EMISSION and SOLAR SPECTROSCOPY AND DIAGNOSTICS: X-RAY EMISSION).

Because of the Sun's brightness, its heat and the dynamic range of its phenomena (which may be time-variable), a wide range of special instrumentation has been employed to study the Sun, both from the ground and from space (SOLAR TELESCOPES AND INSTRUMENTS: GROUND and SOLAR TELESCOPES AND INSTRUMENTS: SPACE). Of the space missions, those of particular interest have been the American SKYLAB satellite (1973–1974), the Japanese YOHKOH mission (August 1991 to the present), the European Space Agency/NASA SOHO mission (December 1995 to the present), and the American TRACE mission (April 1998 to the present). For analysing data, the SOLARSOFT package is particularly useful.

Of growing interest is the link between activity on the Sun and effects on the Earth's atmosphere—the SOLAR–TERRESTRIAL CONNECTION. A subtle coupling exists between variations in the *solar wind*, the Earth's magnetic environment or *magnetosphere*, the ionized or plasma part of the Earth's atmosphere (the *ionosphere*) and the unionized or *neutral atmosphere* (SOLAR–TERRESTRIAL CONNECTION: COUPLING BETWEEN SOLAR WIND, MAGNETOSPHERE, IONOSPHERE AND NEUTRAL ATMOSPHERE). One consequence is that the Earth's

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

climate is variable, driven by changes in the Sun (SOLAR–TERRESTRIAL CONNECTION: LONG-TERM AND SHORT-TERM CLIMATE VARIABILITY). Satellite observations are being used to predict the effect of the Sun's activity from day to day on the solar wind and the magnetosphere (SOLAR–TERRESTRIAL CONNECTION: SPACE WEATHER PREDICTIONS). Looking outwards from the Sun, there is also enormous interest in what the Sun can tell us about other stars and vice versa (the SOLAR–STELLAR CONNECTION) and in particular about solar-type stars (SOLAR–TERRESTRIAL CONNECTION: ACTIVITY AND BRIGHTNESS CHANGES OF SOLAR-TYPE STARS).

The solar interior and atmosphere

There are three main regions of the solar atmosphere, namely, the surface layer of the Sun (the SOLAR PHOTOSPHERE), the overlying hotter and rarer CHROMOSPHERE with a temperature of about 10^4 K and the much hotter (about 10^6 K) and rarer CORONA, which stretches out to the Earth and beyond. The plasma that possesses temperatures between 10^4 K and 10^6 K comprises the SOLAR TRANSITION REGION.

For the SOLAR INTERIOR a series of standard models has been developed (SOLAR INTERIOR: STANDARD MODELS), based on the best available theories of the properties of the interior material (SOLAR INTERIOR: EQUATION OF STATE AND OPACITY). As a result of the high temperatures and pressures, nuclear energy is released in the solar interior (SOLAR INTERIOR: ENERGY GENERATION) with the release of neutrinos (SOLAR INTERIOR: NEUTRINOS). The plasma in the outer 30% of the solar interior is convectively unstable (SOLAR INTERIOR: CONVECTION ZONE and SOLAR INTERIOR: CONVECTION THEORY). Within the SOLAR INTERIOR: CONVECTION ZONE FLUX TUBES respond to the resulting turbulent motions as well as differential rotation and Coriolis forces. Eventually, from the SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES may create sunspots when their flux is large enough.

The Sun rotates (SOLAR INTERIOR: ROTATION). The rotation rate is faster at the equator than the poles and possesses a surprising variation with depth. Gross properties of the SOLAR INTERIOR: SOLAR DIAMETER, OBLATENESS AND TEMPERATURE have dependences on the rotation rate.

Our understanding of the solar interior has been revolutionized by the realization that the Sun is globally oscillating in thousands of modes of vibration (SOLAR INTERIOR: HELIOSEISMIC OBSERVATIONS). Observations have been made both from ground-based networks (BISON NETWORK, GLOBAL OSCILLATION NETWORK GROUP) and the SOHO satellite. The observations lead to properties of the solar interior (SOLAR INTERIOR: HELIOSEISMOLOGY DATA INVERSIONS). HELIOSEISMOLOGY THEORY has been developed in detail, including techniques for inferring properties close to the surface (SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY). MAGNETOHYDRODYNAMICS is used for deducing the influence of magnetic fields (SOLAR INTERIOR: INFLUENCE OF MAGNETIC FIELDS).

For the structure of the atmosphere, empirical models have proved invaluable. Many properties of the SOLAR PHOTOSPHERE change systematically across the face of the Sun (SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION), and on a smaller scale: *faculae* are small regions that are brighter than normal (SOLAR PHOTOSPHERE: FACULAE) and *filigree* are bright crinkle-shaped ribbons (SOLAR PHOTOSPHERE: FILIGREE). Convection exists on several scales, about 1 Mm (SOLAR PHOTOSPHERE: GRANULATION), a larger scale (SOLAR PHOTOSPHERE: MESOGRANULATION) and an even larger scale of 15 Mm (SOLAR PHOTOSPHERE: SUPERGRANULATION). Over most of the solar surface the magnetic field is far from uniform, being concentrated into intense SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES at the edges of supergranule cells. In the interior of such cells, the magnetic fields are tangled by convection (SOLAR PHOTOSPHERE: INTRANETWORK AND TURBULENT MAGNETIC FIELDS).

The CHROMOSPHERE is even more non-uniform. A bright *network* surrounds the supergranule cell boundaries (CHROMOSPHERE: NETWORK). Around sunspots the magnetic field is a few hundred gauss and is associated with a brightening of the surrounding region (SOLAR CHROMOSPHERIC PLAGE). Transient brightenings occur near sunspots at the feet of surges (SOLAR CHROMOSPHERE: ELLERMAN BOMBS). Bipolar magnetic flux tubes break through the solar surface (CHROMOSPHERE: EMERGING FLUX REGIONS). Dark *fibrils* due to dense plasma along a *canopy* of magnetic flux tubes cross the interior of supergranule cells (CHROMOSPHERE: FIBRILS and CHROMOSPHERE: CANOPY). Jets of plasma (CHROMOSPHERE: SPICULES) are ejected along flux tubes upwards at the network boundary. SOLAR CHROMOSPHERIC OSCILLATIONS are observed and are related to one of the possible sources of the high temperatures in the chromosphere (CHROMOSPHERE: HEATING MECHANISMS).

The SOLAR TRANSITION REGION is highly dynamic and it is not easy to develop theories (TRANSITION REGION MODELS). There is a variety of flows and small-scale high-velocity features (TRANSITION REGION FLOWS) and TRANSITION REGION: EXPLOSIVE EVENTS. Elemental abundances vary by about a factor of four from place to place. Elements whose neutral atoms have ionization potential larger than about 10 eV are less abundant than those whose neutral atoms are more easily ionized (TRANSITION REGION: FIRST IONIZATION POTENTIAL EFFECT).

The CORONA possesses several different types of structure. CORONAL HOLES have magnetic fields that are open into interplanetary space. POLAR PLUMES are dense rays in coronal holes outlining the magnetic field. CORONAL LOOPS have magnetic field lines that have both ends anchored to the solar surface. Small bright regions (SOLAR CORONA: X-RAY BRIGHT POINTS) occur where magnetic fields are reconnecting. Streams of plasma (SOLAR X-RAY JETS) are accelerated to several hundred km s^{-1} by magnetic reconnection. Large-scale structures (CORONAL STREAMERS) are closed below a large fraction of a solar radius above the limb and are open above that height.

The solar corona was traditionally seen only during an ECLIPSE but now can be observed by CORONAGRAPHS from the ground and direct by EUV and x-ray telescopes from space. An important problem is the nature of CORONAL HEATING MECHANISMS. Part of this has been solved, since x-ray bright points are now known to be heated by driven

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

magnetic reconnection, but it has not yet been established whether coronal loops and coronal holes are heated by magnetic waves of some kind or by magnetic reconnection in many small current sheets.

Solar activity

The term *solar activity* refers to a wide range of transient solar phenomena that vary in complex ways with the SOLAR CYCLE. SUNSPOTS are dark regions in the photosphere where large thousand-gauss magnetic flux tubes break through the surface. SOLAR ACTIVE REGIONS surround sunspot groups and have fields of a few hundred gauss. SOLAR PROMINENCES are huge flux tubes up in the corona, containing a vertical sheet of plasma at chromospheric temperatures. SOLAR FLARE OBSERVATIONS show how enormous releases of magnetic energy are often related to the eruption of prominences from active regions. SOLAR CORONAL MASS EJECTIONS are large-scale eruptions of mass associated with prominence eruptions, either from active regions or outside active regions.

A variety of SOLAR ACTIVITY INDICES has been proposed for measuring the strength of solar activity. Indeed, solar activity may be traced back many centuries (SOLAR ACTIVITY: LONG-TERM RECORDS). SOLAR ACTIVITY COMPLEXES are active regions that tend to recur in the same location and form. Whereas active regions, like sunspots, occur only in two zones of latitude north and south of the equator, much smaller bipolar regions appear over the whole solar surface at a rate of typically one region per 27^2 Mm^2 per day (SOLAR ACTIVE REGIONS: EPHEMERAL). Coronal mass ejections produce a wave-like disturbance that propagates ahead of the mass and whose intersection with the solar surface sometimes shows up as MORETON WAVES moving outwards from the site of the eruption.

Prominences are referred to as 'filaments' when observed on the disk, and they form in SOLAR FILAMENT CHANNELS, where the magnetic field is mainly horizontal and directed along a reversal in the sign of the normal component of the field at the solar surface. Quiescent prominences can last for many months but are also of shorter duration (SOLAR PROMINENCES: ACTIVE). Important questions concern SOLAR PROMINENCE FORMATION and SOLAR PROMINENCE FINE STRUCTURE. SOLAR PROMINENCE OSCILLATIONS of various kinds are often observed and their sense of twist exhibits a puzzling global pattern (SOLAR PROMINENCES: CHIRALITY).

SOLAR FLARE CLASSIFICATION has settled to a standard method and much effort has been expended on SOLAR FLARE FORECASTING. The behavior of SOLAR FLARE MAGNETIC FIELDS is crucial to an understanding of SOLAR FLARE MODELS and SOLAR FLARES: PARTICLE ACCELERATION MECHANISMS. The connection between solar flares and coronal mass ejections was for a while a matter of heated debate (SOLAR FLARES: RELATION TO CORONAL MASS EJECTIONS). Solar flares often produce small ejections of mass called *surges* and occasionally give a continuum brightening in the photosphere (SOLAR FLARES: WHITE LIGHT). A solar flare has several phases. Before the flare a prominence starts to rise slowly. The flare rapidly develops in a few minutes, with a rapid acceleration of *nonthermal electrons*, sometimes producing *gamma rays*. During the *main phase*, the energy release continues for several hours (SOLAR FLARES: PREFLARE PHASE, SOLAR FLARES: IMPULSIVE PHASE, SOLAR FLARES: NONTHERMAL ELECTRONS, SOLAR FLARES: GAMMA RAYS, SOLAR FLARES: MAIN PHASE). In addition, a flare produces different kinds of radio emission (SOLAR FLARES: RADIO BURSTS).

Sunspots often start as small dark patches called SUNSPOT PORES and develop into several different types (SUNSPOT CLASSIFICATION). The central part of the sunspot is dark and is called the *umbra* (SUNSPOT UMBRA: STRUCTURE AND EVOLUTION). The umbra is slightly depressed and so changes its appearance as a sunspot rotates across the solar disk (SUNSPOTS: WILSON EFFECT). The umbra is surrounded by an annulus (the *penumbra*) with radial striations, since the magnetic field is more horizontal (SUNSPOT PENUMBRA: STRUCTURE AND ACTIVITY). The penumbra exhibits a radial outflow (SUNSPOTS: EVERSHED EFFECT). SUNSPOT OSCILLATIONS AND SEISMOLOGY are used to infer the structures of sunspots and SUNSPOT MODELS for the structures have been developed in detail. SUNSPOT MAGNETIC FIELDS tend to be vertical at the centers of sunspots and then to become more horizontal as one moves outwards from the axis, as the vertical flux tube spreads out through the photosphere. SUNSPOT EVOLUTION follows a standard pattern. A radial *moat flow* of plasma away from a sunspot carries magnetic flux outwards (SUNSPOTS: MOVING MAGNETIC FEATURES AND MOAT FLOW).

The solar heliosphere

The *solar wind* is the outwards expansion of the corona from the Sun. It is described in a general article (SOLAR WIND: GLOBAL PROPERTIES). SOLAR WIND: THEORY describes its structure and SOLAR WIND ACCELERATION how it gains in velocity. The SOLAR WIND COMPOSITION and SOLAR WIND KINETIC PROPERTIES are highly complex. Both *waves* and *turbulence* and *shock waves* make up its structure (SOLAR WIND PLASMA WAVES, SOLAR WIND TURBULENCE, SOLAR WIND SHOCK WAVES AND DISCONTINUITIES) and the SOLAR WIND: MAGNETIC FIELD is highly variable. Especially interesting is the behavior out of the ecliptic plane as measured by the only satellite to orbit in that region (SOLAR WIND: ULYSSES).

Neighboring streams that are flowing out from the Sun interact at their boundaries (SOLAR WIND: COROTATING INTERACTION REGIONS). Accelerated particles (SOLAR WIND: ENERGETIC PARTICLES) of many types are observed; they produce bursts of radio waves (SOLAR WIND: INTERPLANETARY RADIO BURSTS) which are a valuable diagnostic (SOLAR WIND: RADIO TECHNIQUES FOR PROBING). COSMIC RAYS may give rise to showers of electrons and other particles when they impact on the Earth's atmosphere (COSMIC RAYS: EXTENSIVE AIR SHOWERS). Their propagation has been modeled in detail (COSMIC RAYS: PROPAGATION IN THE HELIOSPHERE). Those with energies up to 100 MeV and which lie between the energies of normal heliospheric particles and galactic cosmic rays are known as ANOMALOUS COSMIC RAYS.

The solar wind is rooted in the corona (SOLAR WIND: CORONAL ORIGINS) and is sensitive to solar activity (SOLAR

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

WIND: MANIFESTATIONS OF SOLAR ACTIVITY). Venus interacts with the solar wind as it flows through the heliosphere (VENUS: INTERACTION WITH THE SOLAR WIND), as do other planets and also COMETS (SOLAR WIND: INTERACTION WITH COMETS). Kreutz discovered a family of comets that approach close to the Sun (COMETS: KREUTZ SUNGRAZERS), many of which have been discovered by the LASCO instrument on the SOHO satellite. At the heliopause, the solar wind connects with the rest of the Galaxy (SOLAR WIND: INTERACTION WITH LOCAL INTERSTELLAR MEDIUM).

PLANETARY MAGNETOSPHERES are the magnetized plasma environments of the planets. A range of planets have been studied, including JUPITER: MAGNETOSPHERE and SATURN: MAGNETOSPHERE, as well as MERCURY: MAGNETOSPHERE, URANUS AND NEPTUNE: ATMOSPHERES, IONOSPHERES AND MAGNETOSPHERES. Planetary satellites are of particular interest (SATURN: MAGNETOSPHERE INTERACTION WITH TITAN, MAGNETOSPHERES: JUPITER, SATELLITE INTERACTIONS, MAGNETOSPHERES: JUPITER, RADIO EMISSION, IO: PLASMA TORUS).

The Earth's magnetosphere

Due to its proximity to us, the MAGNETOSPHERE OF EARTH has been extensively studied in great detail and presents an opportunity to understand the behavior of many basic plasma processes. Since the magnetosphere acts as an obstacle to the supermagnetosonic solar wind, a *bow shock* stands in the flow ahead of the Earth (MAGNETOSPHERE OF EARTH: BOW SHOCK). The surface that separates the Earth's magnetic field from the solar wind magnetic field is called the *magnetopause* (MAGNETOSPHERE OF EARTH: MAGNETOPAUSE). Between the bow shock and the magnetopause lies the *magnetosheath*. The Earth's magnetic field is swept downstream by the solar wind to form a long tail (MAGNETOSPHERE OF EARTH: GEOMAGNETIC TAIL), the central and densest part of which is called the *plasma sheet* (MAGNETOSPHERE OF EARTH: PLASMA SHEET). Between the magnetic field that closes on the dayside of the Earth and the tail field that is swept down on the nightside lies the *dayside cusp* (MAGNETOSPHERE OF EARTH: DAYSIDE CUSP).

Within the closed part of the Earth's magnetic field there lie: a *ring current* (MAGNETOSPHERE OF EARTH: RING CURRENT) which flows westward roughly in a circle in the equatorial plane about the Earth; the *radiation belts*, consisting of trapped particles in orbits that circle the Earth from about 1000 km above the surface to about six Earth radii (MAGNETOSPHERE OF EARTH: RADIATION BELTS); a population of cold dense plasma (*plasmasphere*) that coexists with the radiation belts (MAGNETOSPHERE OF EARTH: PLASMASPHERE). Streams of energetic particles impacting the Earth's upper atmosphere in a ring around each of the Earth's poles form the *aurorae* and excite AURORAL KILOMETRIC RADIATION. Within the magnetosphere there is a wide range of waves (MAGNETOSPHERE OF EARTH: WAVES). Also, the rotation of the Earth and the flow of the solar wind cause the plasma and magnetic field lines within the magnetosphere to move in a complex pattern (MAGNETOSPHERE OF EARTH: CONVECTION). Regions of enhanced solar wind flow compress the magnetosphere (MAGNETOSPHERE OF EARTH: GEOMAGNETIC STORMS AND SOLAR WIND ORIGINS) for 1–5 days. *Substorms* occur for 1–3 hours in response to a sustained southward turning of the interplanetary magnetic field, after which the tail of the magnetosphere goes unstable and reconnects (MAGNETOSPHERE OF EARTH: SUBSTORMS). This causes a plasmoid to be formed in the tail and ejected downstream and accelerates fast particles that greatly enhance the aurorae.

Eric Priest

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun as a Star

Observations of the Sun made with instruments on the ground or on satellites in near-Earth orbits have the enormous advantage over corresponding observations of stars in that the Earth–Sun distance is only about 150 000 000 km. Even the nearest star to the Sun (PROXIMA CENTAURI) has a distance which is 271 000 times this amount, while most of the bright stars visible in the night sky are much more distant still. Thus, while ground-based solar telescopes are capable of resolving features on the Sun as small as 150 km across (0.2 arcsec in angular measure), and satellites (in particular the Transition Region and Coronal Explorer or TRACE, launched in 1998) features as small as 725 km (1 arcsec), nothing at all can be resolved on stellar surfaces by conventional means. However, using techniques developed over the years we have been able to deduce the presence of surface and atmospheric features in nearby stars that seem to have their counterparts in the Sun, and so we are able to make deductions about stellar activity along the same lines as we have about solar activity since the time of the first telescopic observations in the early seventeenth century.

Quantities characterizing the Sun as a star

It is important for several purposes to consider certain quantities for the Sun as a star so as to compare with other stars and to use this information as input into the so-called standard model of the solar interior, giving a mathematical description of the way pressure and temperature vary with distance from the center of the Sun. These quantities can be summarized as luminosity, radius, effective temperature, age, chemical composition, mass and rotation rate.

To illustrate, if we know the Sun's luminosity and spectral type (or equivalently the effective temperature), we may place the Sun's position on the familiar HERTZSPRUNG–RUSSELL (HR) DIAGRAM (absolute magnitude versus spectral type) or color–magnitude diagram. The luminosity can be obtained from accurate spacecraft measurements of the solar irradiance (formerly and still occasionally known as the solar constant), i.e. the amount of radiant energy received at the Earth outside its atmosphere per unit area per unit time. There is a small long-term variation in the irradiance with solar cycle and a short-term one, too, depending on the spottedness of the Sun. Broadly, the total variation is 1365 to 1369 W m^{-2} (i.e. peak-to-trough variation of 0.3%). Taking the mean irradiance to be 1367 W m^{-2} and with 1 AU equal to 149 597 900 km, we have for the solar luminosity (i.e. total radiant power in all wavelengths) 3.844×10^{26} W. This is equivalent to an absolute bolometric magnitude (defined by the apparent magnitude at a distance of 10 pc) of +4.72.

The radius (or diameter) of the Sun, i.e. of its photosphere, can be determined in a number of different ways, and there are reasonably precise measurements that date back to the early 1700s. Direct measurements using TRANSIT times of the Sun across the field of a meridian telescope were made for many years (e.g. at the Paris Observatory), and though they are considered fairly reliable there is little doubt that measurements using start and end times of total eclipses of the Sun are the best way of determining the solar diameter. HALLEY's 1715 measurement using this method gave the mean solar diameter as 1920 arcsec (32′ 00″) which is very close to the modern value. Another accurate method of determination involves the transits of Mercury which occur only about 14 times a century. Measurements based on eclipse and Mercury transit methods over the past 300 years indicate a value for the solar diameter of a small fraction of an arcsecond less than 1920 arcseconds, which translates into a linear value for the solar radius of 695 970 km. Despite claims made in the past, the measurements indicate a constant value over this period, and there seems to be a practically zero value for solar oblateness (excess of equatorial over polar diameters).

The Sun's effective temperature, obtained from the solar irradiance and the Stefan–Boltzmann law for a black body, is 5770 K. The equivalent spectral type is G2, luminosity class V (i.e. a main-sequence or non-evolved star). This places the Sun very nearly at the center of the main sequence in the HR diagram when a large variety of stars—very luminous as well as very dim—are plotted. The Sun's color, as defined on the UBV photometric system, is given by $B - V = +0.65$ and $U - B = +0.13$.

The age of a star is an important quantity, as with increasing age a star depletes its hydrogen 'fuel' at its core and eventually evolves off the main sequence. Stellar ages can only be directly determined for members of star clusters, but the Sun's age can be fairly accurately determined from the age of the oldest meteorites discovered on the Earth; latest estimates are 4.6×10^9 years (4.6 Gyr). This is considerably larger than the ages of associations of stars known to be evolving towards or just recently arrived at the main sequence (e.g. ages of O- or T-associations are around 1 Myr) and most open clusters of stars (e.g. the Pleiades cluster's age is 60 Myr, the Hyades 600 Myr). However, the Sun's age is much less than the age of globular clusters (mean about 10 Gyr) and the latest estimates of the age of the universe (12 Gyr).

Again, for stellar evolution purposes, the chemical composition is of great importance. For the Sun, we are able to make very accurate measurements of the fractional amounts of relatively abundant elements in the photosphere because the great amount of light available enables us to get high-resolution visible-light spectra. In broad terms, the composition is 91% H, 9% He, and only 0.1% heavier elements (or 'metals'). The elements C, O, N, Ne and Fe feature as some of the most abundant heavier elements. Despite what would appear to be the small total metal abundance, the Sun is considered to be 'metal-rich' and belongs to POPULATION I (i.e. is a second-generation star). (There is some evidence, though this remains controversial, that the solar corona has enhanced abundances of certain heavy elements, in particular those with a relatively small value of first ionization potential.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A star's mass determines the nature of its evolution—a large-mass star evolves very fast, a star with small mass settles down to a long existence in the low-luminosity end of the main sequence, while a body with very small mass never achieves stellar status at all, i.e. does not maintain its energy output by nuclear fusion reactions at its core. The Sun's mass can be very precisely determined from planetary motions using KEPLER'S third LAW: the cube of a planet's distance from the Sun divided by the square of its period gives G times the product of the sum of the Sun's and planet's masses. The best estimates for the Sun's mass give 1.989×10^{30} kg. Stellar masses can be directly determined for binary systems, i.e. where the two components are gravitationally bound, in which the orbits can be determined by continued observations over a period of time. By such means, stellar masses from 0.1 to 50 times the solar mass have been found. The Sun's mass, then, is approximately a logarithmic mean of the extremes.

For stars that have ages of 1 Gyr or more, there is an approximate relation between activity (as measured by numbers of active regions or propensity to flare) and rotational rates, defined by equatorial rotation speeds. This is most likely due to the way in which magnetic fields are generated by dynamo action in convective zones (or, for stars like the Sun, the base of convective zones). The Sun rotates with a period that depends on latitude: 25 days at the equator and up to 33.5 days at latitude 80°. The corresponding rotational speed at the equator is 2 km s^{-1}. Main-sequence stars known to be younger than the Sun from their larger rotational speeds are also much more active (with huge flares, many times more energetic than the largest solar flares), while in general stars with smaller rotational speeds have very low or apparently zero levels of flaring.

In summary, the Sun occupies very much an intermediate position in comparison with other stars by almost any measure: mass, radius, luminosity, age, rotation/activity. It lies on the main sequence and is approximately at the midpoint of its life: in another 5 Gyr it will have depleted its hydrogen fuel at its core, and will start to evolve off the main sequence to a red giant phase.

Global oscillations: observing the Sun's interior

With the Sun's proximity to the Earth, we are able to resolve features on its surface under ideal seeing with the best ground-based telescopes that are only 150 km across, corresponding to 0.2 arcsecond. With such a high degree of image quality available, there would seem to be little worth in deliberately degrading the resolution, but for certain studies that is what astronomers sometimes choose to do. Such a study is HELIOSEISMOLOGY. The Sun is now known to have non-radial global oscillations, and these are best studied by integrating the visible light output of the entire solar disk.

The outer regions of the solar interior are convective, and convective motions excite sound waves which travel from the surface towards the interior. As an acoustic or resonance cavity is formed when a sound wave travels between two solid boundaries, with the possibility of standing waves or modes if the distance between two boundaries is a simple multiple of the wavelength, so the solar interior acts as an acoustic cavity. A sound wave that propagates into the interior from the solar surface at an angle travels in an arc. The reason for this is that the wave front (perpendicular to the wave direction) has its deepest part traveling faster than its shallowest part since the sound speed is larger at greater depths where the temperature is larger. The sound wave is therefore refracted until it is directed towards the solar surface again where it produces oscillations which are observable through the Doppler shifts of absorption lines in the photospheric spectrum. Such standing waves are called p modes (p for pressure which is the restoring force for these waves). The degree l of the mode can be defined to be the solar circumference divided by the horizontal wavelength, and small values of l ($l = 0$–4) produce global oscillations which are observable through Doppler shifts in integrated sunlight. Individual oscillation modes have only tiny velocity amplitudes, so highly accurate means of measuring Doppler shifts must be made, and generally over long periods of time.

Measurements of low-l p modes have been made for a number of years with the University of Birmingham (UK) BISON NETWORK and the French IRIS global oscillations network, and since 1995 observations have been made from space with the GOLF instrument aboard SOHO. The observations indicate the presence of several strong peaks in the Fourier spectrum in the region around 3 mHz (corresponding to the familiar 5 min period oscillations). The peaks are closely spaced pairs, which comparison with calculations reveals to be due to modes with l equal to 0 and 2 or 1 and 3. The spacing between the peaks can be directly compared with that predicted by the standard solar model where the agreement has been found to be reasonably satisfactory.

The small velocity amplitudes observed in integrated sunlight that global oscillations give rise to would give little hope that global oscillations would be detectable in other Sun-like stars. However, much effort is being directed in the field of astroseismology, the stellar equivalent of helioseismology. For solar-type stars, we would expect the same behavior that has been found from helioseismological studies, so the measurements are at present only bordering on significant detections of oscillations. However, some other stars appear to show definite signs of global oscillations, in particular rapidly oscillating Ap stars. A major discovery in recent times is that of g mode oscillations in γ Dor stars (g stands for gravitation as the restoring force). The observations of solar g modes are equivocal and there is much debate regarding their validity.

Observing the Sun from a nearby star

When we observe stars from the Earth using either ground-based observatories or spacecraft, we are in general unable to see surface features, though we can deduce

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

their presence by indirect means. Thus starspots can be detected through a technique called Doppler imaging, which uses the fact that the profile of an absorption line is altered in response to a large enough starspot crossing the star's visible hemisphere as the star rotates. Stellar chromospheres are detectable with visible-wavelength and ultraviolet observations. Cyclical behavior in some stars like the Sun's 11-year cycle has been observed through the Ca II H and K emission line intensities, since the emission cores result from stellar chromospheres which, by analogy with the Sun, show active regions as areas of enhanced emission. Similarly, x-ray observations reveal the presence of coronas, and for active cool dwarf stars especially (dMe and dKe stars: d signifying dwarf and e emission in the Hα line) flares are readily observed in wavelengths ranging from m-wave radio to x-rays.

With the Sun, ground-based and spacecraft observations give us a highly detailed view of not only surface features but of the atmospheric structure, and we find the general rise in temperature, above the temperature minimum region (about 500 km above the surface, defined by the optical depth at 500 nm, τ_{500} equal to unity), going up through the chromosphere to the corona. If we were to observe the Sun from a distance of 1.3 pc, the distance of the nearest star Proxima Centauri, we would find evidence for the Sun's atmospheric structure much more difficult to deduce. Whereas for flare stars of type dMe or dKe, and even for M dwarfs as a whole, the emission profiles of the Ca II H and K lines are very obvious, in integrated sunlight the emission peaks are scarcely visible at all, so it is only with high-resolution spectroscopy that the solar chromosphere would be detectable at visible wavelengths. However, ultraviolet observations of the Sun from a nearby star might detect the presence of the chromosphere through the Lyman-α and Mg II lines (121.5 nm and 280 nm respectively). The SOLAR TRANSITION REGION (supposed boundary layer between the chromosphere and corona but possibly due to unresolved structures in the Sun) has a temperature of about 10^5 K, very near the maximum emissivity of lines such as the C IV doublet at 155.0 nm. Again, ultraviolet observations with a spacecraft similar to the International Ultraviolet Explorer (IUE) would only just detect this region. The total x-ray luminosity of the solar corona is about 10^{20} W, a value which is at the lower end of the range of x-ray coronal luminosities detected by the ROSAT spacecraft.

Doppler imaging shows that starspots on RS CVn stars in particular can attain huge dimensions, often a substantial fraction of the star's total surface area. Thus, there are perceptible effects both in total brightness and in the profiles of absorption lines. Some very large sunspots are as much as 100 000 km across, many times the Earth's diameter, but even so not more than about 1% of the Sun's area is covered with spots at any one time. Thus, an observer at the distance of even the nearest star with instruments comparable to those in use at present would probably not be able to detect any light modulations or changes in line profile.

SOLAR PROMINENCES are tongues of cool gas (temperatures typical of the chromosphere, about 10^4 K) located in the solar corona. They are most often seen in Hα light, in emission off the limb or in absorption on the solar disk (when they are known as FILAMENTS) and may take on enormous proportions (such as the famous 'Anteater' prominence in June 1946). Their counterparts have often been seen in some active stars, in particular through changes in the profile of the Hα line as the prominence crosses the visible hemisphere of a rapidly rotating star such as AB Dor (dKe). Again, these stellar prominences seem to have much greater dimensions than solar prominences, and in the case of AB Dor are 'co-rotating' with the star's photosphere, with distances from the star of up to 9 stellar radii. Solar prominences are much too small and tenuous to have any perceptible effect on the profile of the Hα line as observed from a great distance.

The flares that are seen on dMe stars and the interacting binaries of the RS CVn type are truly enormous compared with even the largest observed on the Sun. Much of the radiated energy of stellar or solar flares appears in the x-ray region. Energies of up to 10^{29} J have been observed on RS CVn binaries, and up to 10^{27} J in the case of dMe stars. This compares with only about 3×10^{24} J for the largest solar x-ray flare ever observed. In fact, although solar flares give rise to an enormous increase in x-ray emission on occasions, the total energy output of the Sun completely overpowers any flare: the energy output is increased by no more than 0.001%. The only way a flare could be detectable from a great distance would be in the x-ray region, since there would be a negligible effect on the Hα or other line profile and the total visible light output.

Bibliography

Collier-Cameron A 1999 Stars and stellar systems: similarities and differences *ASP Conf. Ser.* **158** 146

Gabriel A H *et al* 1997 Performance and early results from the GOLF instrument flown on the SOHO mission *Solar Physics* **175** 207–26

Phillips K J H 1995 *Guide to the Sun* (Cambridge: Cambridge University Press) ch 8

Kenneth J H Phillips

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun: Basic Properties

The standard star

The Sun is a gaseous sphere in hydrostatic equilibrium. It emits light because it is hotter than the surrounding space. Such a brief characterization is true for most other stars as well. The distinction of the Sun is its close vicinity to the Earth, so that it is the only star where we can see fine details on the surface (see SOLAR PHOTOSPHERE). Examples are the granulation as well as other manifestations of convective energy transport, dark spots that are particularly structured in their umbrae and penumbrae, or limb prominences that are supported by an appropriate magnetic field. It is true that, by indirect methods, similar phenomena have been inferred for stellar surfaces, but the interpretation of the observations would be difficult, if not impossible, without the solar evidence (see SOLAR–STELLAR CONNECTION).

A particular example of solar uniqueness is that the Sun is the only star in equilibrium whose neutrino luminosity has been measured. From celestial sources other than the Sun, only some 19 neutrinos were detected on 23 February 1987 after the explosion of supernova 1987A in the Large Magellanic Cloud.

Fundamental quantities

In order to determine the most fundamental solar quantities one needs to know the *distance* of the Sun from the Earth. The mean distance, or *astronomical unit*, is

$$A = 149\,597\,870 \pm 2 \text{ km}.$$

This value is the mean distance between the centers of the Sun and the Earth. The 'mean' is over the *eccentric anomaly*, i.e. the angle between the Sun–Earth line and the direction to the perihelion of the elliptical orbit. The most accurate measurement of A is obtained from radar echos within the solar system, and from Kepler's third law. The uncertainty arises from the uncertainty of the radar echos, which yield the *light time for unit distance*, $\tau_A = 499.004\,782 \pm 0.000\,006$ s.

The following is a list of some fundamental solar quantities:

Mass, $m_\odot = (1.9889 \pm 0.0003) \times 10^{30}$ kg

Radius, $r_\odot = (6.960 \pm 0.001) \times 10^8$ m

Luminosity, $L_\odot = (3.844 \pm 0.010) \times 10^{26}$ W

Age, $t_\odot = (4.57 \pm 0.05) \times 10^9$ years

Base of convection zone, $r_{CZ} = (0.713 \pm 0.003) r_\odot$.

The chemical composition at the surface, by fractions of mass, is

Hydrogen, $X_S = 0.734 \pm 0.005$

Helium, $Y_S = 0.248 \pm 0.005$

Other elements, $Z_S = 0.018 \pm 0.002$.

Several remarks concerning these values must be made, in particular with respect to the error estimates.

Mass

Time and distance measurements, and the application of celestial mechanics, yield the *product*, $Gm_\odot = (132\,712\,438 \pm 5) \times 10^{12}$ m^3 s^{-2} of the solar mass with the general constant of gravitation. The uncertainty of $m_\odot$ rises entirely from the uncertainty of $G = (6.672\,59 \pm 0.000\,85) \times 10^{-11}$ m^3 kg^{-1} s^{-2}.

Radius

In a photoelectric measurement the value of $r_\odot$ can be derived from the angular distance between the two inflection points of an intensity profile across the visible solar disk. Recent experiments yield values around 959.6″ for the angular semi-diameter, remarkably close to visual heliometer measurements carried out in the nineteenth century (which led to the value 959.63″ used in the *Astronomical Almanac* until today). The statistical errors of some of the individual experiments are as small as 0.01″. However, the systematic uncertainty must be at least 0.1″, as is evident from the spread of the results. The more conservative error given above accounts for this.

The solar limb defined by the intensity inflection point is ≈300 km above the level where the optical depth is $\tau = 2/3$. For a boundary condition to a solar interior model, applied at that level, the radius is therefore $r = (6.957 \pm 0.001) \times 10^8$ m. A comparison of theoretical and observed frequencies of solar f-mode oscillations confirms this value of the radius at $\tau = 2/3$.

Luminosity

The Earth's atmosphere absorbs or reflects a considerable fraction of the solar radiation. Hence an accurate measurement can only be made from space. Figure 1 shows the total solar irradiance obtained from bolometric experiments on balloons, rockets and satellites since 1978. Short-term as well as long-term *relative* variations, measured by the diverse experiments, agree very well with each other, thus demonstrating their solar origin. Most conspicuous is a variation with the 11-year activity cycle of the Sun. This variation is *in phase* with the cycle: the maxima of irradiance and of the number of sunspots coincide. The amplitude of the variation is ≈0.09%, peak-to-peak. The occurrence of *solar faculae* depends on the cycle phase (and on heliographic latitude) and plays the major role in models of this variation.

Figure 1 also demonstrates differences of several W m^{-2} between the *absolute* results of the diverse experiments, indicating systematic uncertainties. The solar luminosity given in the present list is $4\pi A^2$ times the mean irradiance value 1367 W m^{-2} measured by the 'Active Cavity Radiometer Irradiance Monitoring I' instrument on the Solar Maximum Mission. This instrument had a precise solar pointing system and, by the use of *three* absorbing cavities operating at different intervals, it had the capability to correct for detector degradation; its accuracy has been estimated to ±0.2%. The error of $L_\odot$ given above is a conservative guess based on this number (see SOLAR IRRADIANCE).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

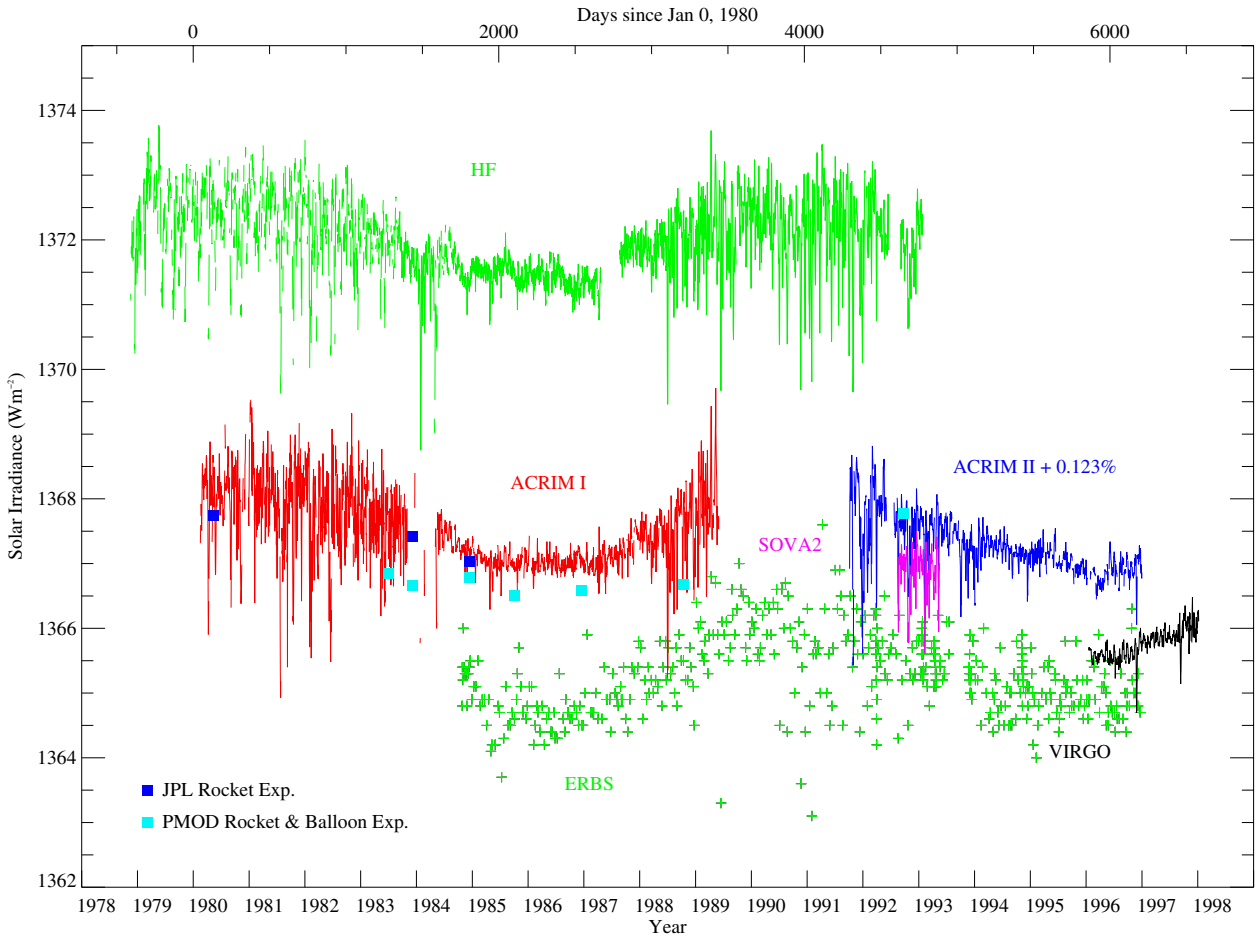


Figure 1. Total solar irradiance 1978–98, measured by balloon, rocket and satellite experiments. HF: Earth Radiation Budget, on Nimbus-7; ACRIM: Active Cavity Radiometer Irradiance Monitoring; SOVA: Solar Variability, on EURECA; ERBS: Earth Radiation Budget Satellite; VIRGO: Variability of Solar Irradiance and Gravity Oscillations, on SOHO. (Updated from C Fröhlich and J Lean; courtesy J M Pap.)

Age

The decay of long-lived radioactive isotopes, notably ^{238}U, ^{235}U, ^{232}Th, ^{40}K, ^{87}Rb, ^{147}Sm and ^{187}Re, yields the *age of meteorites*, defined as the time between the melting, i.e. homogenization of isotopes, and subsequent crystallization of these objects, and the present. Such age determinations are possible with errors as small as a few million years. Nevertheless, there is a larger uncertainty in the solar age because of the uncertainty in the state of the Sun's evolution at the time of meteorite formation. Both the initial collapse of an interstellar cloud and the final hydrostatic contraction towards the main sequence in the Hertzsprung–Russell diagram have time scales of order 10^7 years, which leads to the uncertainty estimate of $t_\odot$ (see SOLAR EVOLUTION).

Depth of convection zone

The best value of r_{CZ} is obtained by means of helioseismology. The frequencies of the Sun's acoustic modes of oscillation (the p modes) depend on the gradient, dc/dr, of the sound speed, and therefore on the gradient of the temperature. The temperature gradient makes a characteristic transition from the nearly isentropic stratification in the convection zone to the core stratification that is determined by the radiative transport of energy. The error estimate of r_{CZ} reflects the uncertainty of the inversion of the observed p-mode frequencies, and (possibly) an uncertainty in the knowledge of the thermodynamic state in the Sun (see SOLAR INTERIOR: CONVECTION ZONE).

Chemical composition at the surface

Helioseismology yields the most precise determination of the abundance of helium. Since He is an abundant element in the Sun, its ionization energy substantially contributes—in a zone of partial ionization—to the specific heat and hence to the adiabatic exponent and the speed of sound. Therefore, the measurement of p-mode

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

frequencies can be used to determine the helium content. The result listed above pertains to the outer solar layer between r_{CZ} and $r_\odot$ which is permanently mixed by turbulent convection. The uncertainty corresponds to the uncertainty of the frequency inversion and of the thermodynamic state in the layer of partial ionization of helium.

The heavy element mass fraction, Z_S, is determined essentially from a spectroscopic analysis of the solar photosphere, and from abundance ratios found elsewhere in the solar system. The uncertainty is essentially due to the errors of the spectroscopic results (see SOLAR ABUNDANCES).

Internal constitution

The internal structure of the Sun is governed by the hydrostatic support of the solar mass, the conversion of nuclear into thermal energy by fusion of hydrogen into helium, and by the mechanisms of radiative and convective energy transport. The thermodynamic state of the solar interior is close to the state of a perfect gas; small deviations of the perfect gas law occur near the solar center where partially degenerate electrons contribute ≈1.5% to the pressure, as well as in the outer part—mainly in the envelope where the temperature is below 2×10^5 K and where excitation and ionization of neutral or partially ionized species make a contribution of up to a few per cent.

In addition to the equation of state, the *opacity*, defined as the mean absorption coefficient κ (i.e. properly averaged over the photon frequency), plays a major role in our knowledge of the stellar interior. It is derived through a detailed analysis of the diverse atomic processes that absorb or scatter photons. For the Sun, generally opacity *tables* are used, which list κ as a function of temperature, density and chemical composition. The most recent tables (the 'OPAL' tables) have been calculated at the Lawrence Livermore National Laboratory (USA). The opacity is of crucial importance for the temperature profile in the solar core below r_{CZ}, where energy is transported by radiation; it also is crucial in determining the depth where radiative transport becomes inefficient and, therefore, the stratification becomes unstable to convection. For the solar interior the opacity is still a source of some uncertainty; for example, the depth of the convection zone in calculated models has changed with the incorporation of new opacity tables, and only the recent Livermore tables yield models having a r_{CZ} in close agreement with the helioseismological result (see SOLAR INTERIOR: EQUATION OF STATE AND OPACITY).

Over the Sun's age the helium content in the central region has increased, mainly by *fusion of hydrogen*, and—by a smaller amount—by *gravitational settling*. Evolutionary calculations show that gravitational settling has reduced the helium content of the convective solar envelope from an initial mass fraction $Y_{ini} \approx 0.280$ to the present value $Y_S \approx 0.248$. For the elements heavier than helium, settling is the only mechanism of chemical differentiation. The change in the envelope is from an original $Z_{ini} \approx 0.020$ to the present value $Z_S \approx 0.018$.

The *Standard Solar Model* is based on the best available physical and chemical input, as outlined in this section. This Standard Model reproduces, within some uncertainties, the observed basic properties of the Sun—with one exception, the neutrino luminosity (see below). At the solar center, some of the physical quantities are as follows:

Temperature, $T_C = 1.57 \times 10^7$ K
Density, $\rho_C = 1.52 \times 10^5$ kg m^{-3}
Pressure, $P_C = 2.37 \times 10^{16}$ Pa
H mass fraction, $X_C = 0.34$
He mass fraction, $Y_C = 0.64$
(see SOLAR INTERIOR: STANDARD MODELS).

The neutrino discrepancy

Nuclear energy conversion in the solar interior is directly confirmed by the measurement of solar neutrinos. However, the observational result quantitatively falls short with respect to the prediction based on the Standard Solar Model. Radiochemical experiments that use neutrino capture by the chlorine and gallium isotopes ^{37}Cl and ^{71}Ga yield ≈28% and ≈55% respectively of the theoretical expectation. Water Cerenkov detectors that record neutrino–electron scattering have demonstrated that the neutrinos arrive from the direction of the Sun; their yield is slightly less than 50% of the theoretical value.

The Standard Solar Model predicts not only the total neutrino flux from the Sun, but also its energy distribution. The diverse experiments are sensitive to different ranges of neutrino energy. Although the predicted energy distribution may be modified, in particular by the construction of a solar model with a lower central temperature, it has not been possible to calculate models whose neutrino production would be in accord with all the observational results simultaneously. Moreover, the means to change the model temperature (opacity, equation of state, nuclear reaction rates) are rather limited. It therefore appears that the discrepancy cannot be resolved in the astrophysical context. Instead, physical concepts that lie beyond the standard electroweak model of elementary particles may be required (see SOLAR INTERIOR: NEUTRINOS).

Oscillations and helioseismology

The Sun supports two kinds of oscillation around its hydrostatic equilibrium stratification. According to their predominant restoring forces, these oscillations are classified as pressure or p modes, and gravity or g modes. The spherically symmetric p-mode oscillation has a fundamental period of ≈63 min. This fundamental mode as well as the first few overtone modes have not yet been observed. But the frequencies of numerous higher overtones, with periods in the range of 2 to 20 min, have been measured, for spherically symmetric as well as for asymmetric, or non-radial, p modes. The non-radial modes are characterized by their spherical harmonic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

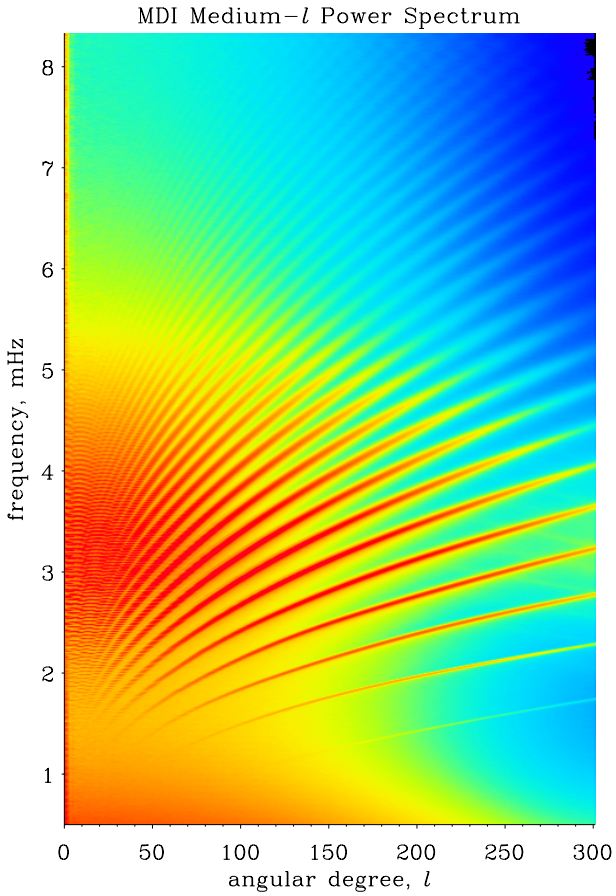


Figure 2. Power distribution of observed solar p modes in the degree–frequency space. (Data from the Michelson Doppler Imager on SOHO, cf Kosovichev *et al* (1997).)

degree l; oscillations with degrees up to $l \approx 4000$ have been found. In the degree–frequency diagram the p modes are aligned along well-defined ridges, as shown in figure 2 (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY).

The solar p-mode frequencies are currently determined with a relative error of less than 10^{-3}, for some less than 10^{-4}. Theoretical eigenfrequencies of the Standard Solar Model are calculated with equal accuracy. However, the relative deviations between measured and calculated frequencies have reached 10^{-3}–10^{-2} for some of the theoretical solar models; such deviations have therefore been used to refine the physical input to the model. Modifications in the equation of state, more accurate abundances of chemical elements, more detailed atomic physics for the calculation of opacity tables, and the process of gravitational settling of helium and the heavier elements have thus been incorporated into the Standard Solar Model.

There is no firm evidence for the existence of g modes on the Sun, in spite of considerable observational effort.

Solar rotation

The rotation of the Sun has been known since about 1610 when Galileo Galilei and other astronomers began to use the newly invented telescope for regular observations of sunspots. From such observations R C Carrington by 1863 had derived the *rotation elements*, i.e. the two angles that define the direction of the Sun's axis of rotation. These angles are the inclination, $i = 7.25^\circ$, of the axis with respect to the normal on the ecliptic plane, and the longitude of the ascending node, $☊ = 75.73^\circ$ (for 1998), measured in the ecliptic plane from the equinox direction. Due to the precession of the Earth's axis the equinox direction recedes by $50.26''$ per year; ☊ increases at the same rate, because the direction of the Sun's axis remains fixed. The rotation elements determined by Carrington are still used in the *Astronomical Almanac*, although modern evaluations of long-term sunspot observations yield a slightly steeper axis, with $i \approx 7.13^\circ$. Carrington's ☊ agrees with the modern result.

From the transit of sunspots across the solar disc Carrington derived a mean synodic period of rotation, namely 27.2753 days. Time intervals of this length are counted as *Carrington rotations*, with rotation no 1 starting on 9 November 1853. The synodic mean angular velocity corresponding to the Carrington rotation is $\Omega_{\rm syn} = 2.67 \times 10^{-6}\ {\rm s}^{-1}$; the sidereal mean value is $\Omega_{\rm sid} = 2.87 \times 10^{-6}\ {\rm s}^{-1}$.

Due to the rotation the figure of the Sun is slightly oblate. The difference between the equatorial and polar radii, $r_{\rm eq} - r_{\rm pole} \approx 10^{-5} r_\odot$, has been measured in a number of experiments (see SOLAR INTERIOR: SOLAR DIAMETER, OBLATENESS AND TEMPERATURE).

The solar rotation is non-uniform, i.e. the angular velocity $\Omega(r, \theta)$ depends on radius r and colatitude θ. At the surface, the θ-dependence is directly observed, both by means of tracers such as sunspots and spectroscopically. Spectroscopic data from Mount Wilson Observatory yield $\Omega_{\rm sid}(r_\odot, \pi/2) = 2.84 \times 10^{-6}\ {\rm s}^{-1}$ at the equator and $\Omega_{\rm sid}(r_\odot, 0) = 2.01 \times 10^{-6}\ {\rm s}^{-1}$ at the poles. Sunspots as tracers generally yield a slightly faster rotation, although they can only be used at low latitude, up to about $\pm 40^\circ$.

Below the surface the Sun's rate of rotation has been determined by helioseismology. Non-radial p modes of degree l have a $(2l + 1)$-fold degeneracy in the case of spherical symmetry. Rotation breaks this symmetry and removes the degeneracy. The frequencies are split into $(2l+1)$-fold multiplets. Since different modes of oscillation penetrate into different depths of the Sun, the frequency splitting can be inverted; the angular velocity as a function of r can thus be obtained. Moreover, multiplets with non-equidistant frequency spacing carry information about the dependence of Ω on latitude.

The helioseismological results (figure 3) demonstrate that, within the convection zone, the angular velocity has approximately the same latitude dependence as at the surface. At the base of the convection zone, around $r = 0.7 r_\odot$, there is a transition region of thickness $\Delta r \approx 0.1 r_\odot$ to a core with approximate solid-body rotation; in the core the angular velocity is of the same order as around $\pm 30^\circ$ latitude at the surface (see SOLAR INTERIOR: ROTATION).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

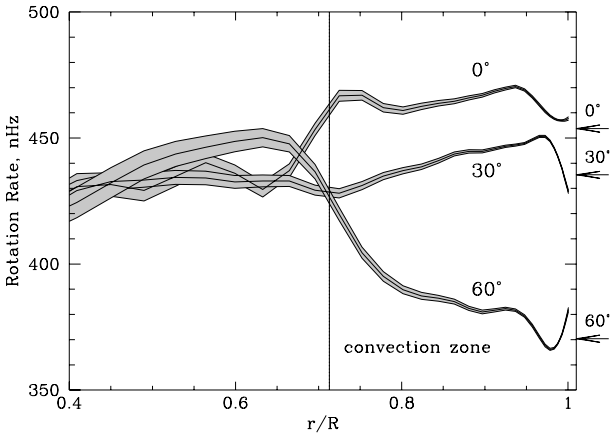


Figure 3. Solar rotation rate, as a function of r, for three heliographic latitudes; an inversion of data obtained with the Michelson Doppler Imager on SOHO. The vertical line marks the base of the convection zone; the arrows indicate the rotation rate measured spectroscopically at the surface. (From Kosovichev *et al* (1997).)

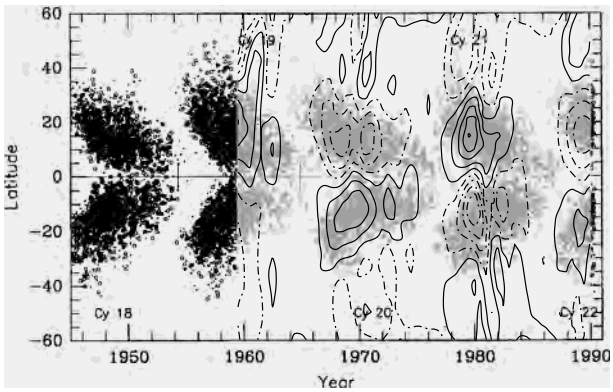


Figure 4. Butterfly diagram of sunspots (dark areas), and contours of the radial component of the mean poloidal magnetic field (solid positive, dashed negative).

The magnetic field of the Sun

Zeeman splitting of spectral lines in sunspots was first reported by G E Hale in 1908. The magnetic field in the spots that accounts for the splitting has a field strength of 0.2–0.3 T. Sunspots preferentially appear in pairs of opposite magnetic polarity, oriented in the heliographic longitude direction, with the leading spot (in the sense of solar rotation) slightly closer to the equator. The orientation is the same for all pairs within each hemisphere and each 11-year sunspot cycle, but reverses from one cycle to the next and is antisymmetric with respect to the equator (*Hale's rules*). The full magnetic cycle therefore lasts 22 years (see SOLAR CYCLE).

The magnetic field of the sunspots has its origin in the solar interior, where it is generated by a self-excited magneto-hydrodynamic dynamo. The seat of this dynamo most probably is the transition region at the base of the convection zone, around $r = 0.7r_\odot$. At this depth the shearing non-uniform rotation (figure 3) generates a toroidal magnetic field from a pre-existing poloidal field. If in a tube of magnetic flux the field strength reaches a critical value of order 10 T, the tube may become unstable, with loops rising to the solar surface. By this process bipolar magnetic regions and pairs of sunspots are formed (see SUNSPOTS).

Individual sunspots have a lifetime of days to months. At the begin of each 11-year half-cycle, spots form at intermediate heliographic latitude, around $\pm 35^\circ$. In the course of the cycle the two zones of sunspot emergence migrate toward the solar equator. This is known as the 'butterfly diagram', cf figure 4. The cause of the migration is closely related to the solar dynamo. The mechanism is modeled in particular in terms of the $\alpha\Omega$-dynamo. In this model the non-uniform rotation is complemented by the interaction of the Coriolis force with loops of magnetized material; a new (reversed) parent poloidal field for the generation of the toroidal field of the next cycle is formed in this way. Figure 4 contains mean contours of the measured poloidal field (see DYNAMO THEORY).

Outside the sunspots, and also outside the latitude zones populated by sunspots, the solar magnetic field appears in the form of locally concentrated flux tubes. These tubes are too small to cause dark spots, and generally also too small to be resolved in observed magnetograms. But their field strength may reach 0.2 T and, because they are so numerous, their total absolute magnetic flux is comparable to the flux that is concentrated in sunspots (see SOLAR MAGNETIC FIELD).

The outer atmosphere and heliosphere

The outer atmosphere of the Sun consists of the *chromosphere*, with a mean height of about 2000 km, the *transition region*, where the temperature rapidly changes from $\approx 10^4$ K to $\approx 10^6$ K, and the *corona*, which has a temperature above 10^6 K and extends over several solar radii. Because of its high temperature, the corona expands with a velocity that increases outwards. At a critical distance r_c, probably between $5r_\odot$ and $10r_\odot$, the expansion becomes supersonic and turns into the *solar wind*. At the Earth's orbit the wind velocity varies between 300 and 700 km s^{-1}. At a distance of the order of 50 astronomical units, according to theory, a shock marks the transition to the interstellar medium. There the dynamic pressure of the solar wind matches the interstellar gas pressure (see SOLAR WIND: GLOBAL PROPERTIES). The diverse regions of the outer solar atmosphere are not horizontally uniform layers, but rather highly structured. The agent that dominates the structure is the magnetic field. The ratio of gas pressure to magnetic pressure (the plasma β) is small, in contrast to the solar interior where $\beta > 1$. The main elements of the structure are spicules and prominences in the chromosphere; arches, streamers and coronal holes in the corona; and low- and high-speed streams in the solar wind. The interaction of the magnetic field with the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

solar wind also leads to breaking of the Sun's rotation: the field acts as a lever arm, reaching outwards to the *Alfvén radius* r_A where the magnetic and kinetic energy densities become equal. The Alfvén radius should lie between 10 and 15 solar radii; its position has not yet been determined experimentally (nor has the position of r_c), because no spacecraft has yet approached the Sun closer than 0.29 astronomical units.

In addition to the spatial inhomogeneity, the outer atmosphere is highly variable in time. As the inhomogeneity, the variability is due to the magnetic field. Magnetically stored energy is set free by instabilities, magnetic lines of force reconnect and accelerate particles. Chromospheric flares, bursts of radio emission and coronal mass ejections follow (see SOLAR ACTIVE REGIONS).

Bibliography

Cox A N, Livingston W C and Matthews M S (eds) 1991 *Solar Interior and Atmosphere* (Tucson, AZ: University of Arizona Press)

Kosovichev A G *et al* 1997 *Solar Phys.* **170** 43

Sonett C P, Giampapa M S and Matthews M S (eds) 1991 *The Sun in Time* (Tucson, AZ: University of Arizona Press)

Stix M 1999 *The Sun, an Introduction* 2nd edn (Berlin: Springer)

The Astronomical Almanac (Washington, DC: US Govt Printing Office)

M Stix

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sun: Safe Observation

Generally speaking, astronomy is a fairly safe pastime. There is, however, one celestial object which is highly dangerous: the SUN. Even the solar surface is at a temperature of well over 5000 °C, and deep at the core the temperature rises to around 15 000 000 °C; remember, the Sun may be regarded as a vast nuclear reactor.

Staring straight at the Sun, even with the naked eye alone, is emphatically not to be recommended. The human eye is delicate, and all too easily damaged; safety considerations must always be paramount—as many people in the past have found to their cost. Carelessness can result in permanent eye damage, or even blindness. Therefore let us consider what the dangers are, and what precautions need to be taken.

With the naked eye, a suitably dark filter is safe enough, but it must be really dense. Welding glass is to be recommended, at least Grade 14, and is easy to obtain. There are occasions when sunspot groups become large enough to be seen in this way, but it is best to wait until the Sun is low down in the sky, preferably shining through a certain amount of mist or haze.

Fully developed, fogged black and white film has often been used, but some modern films are not completely effective, and so are best avoided. Exposed color film should never be used, because it is no barrier to infrared radiation which is dangerous to the eye.

Smoked glass is often recommended, but is not safe. By its very nature, the deposit is uneven, and any pinprick already present or caused by fingering produces what may be termed a 'pinhole camera' which will focus onto the observer's eye.

Much the best filter is Mylar, which is a very tough proprietary plastic film a few microns thick. For solar work, filters are made by coating the film with a very thin layer of aluminum. If the filter is to be used with the naked eye alone, then Mylar coated on both sides is best; for photography single-sided coating is better, because it means that exposure times can be shorter. It must never be confused with 'silver paper', which is a very thin sheet of aluminum, and provides virtually no protection at all.

The only danger with Mylar comes from pinholes. Before using the filter, check it carefully, and if there are any pinholes throw the filter away.

Always remember that the Sun is more dangerous than it often looks. Low down, shining through thick horizon haze, it seems deceptively mild and harmless. The temptation then is to take a quick direct look with binoculars or a telescope. Eye damage is virtually certain to result, and there have, unfortunately, been many cases of people who have been blinded in this way.

Remember, too, that we are not dealing only with visual light. The Sun is also very powerful in the ultraviolet and infrared wavelengths. If you want to demonstrate the heating involved, simply hold up a magnifying glass and focus the Sun's rays onto a piece of paper. It does not take long for the paper to become hot, and if the Sun is high in the sky the paper will quickly burst into flame (this was, of course, the way in which Boy Scouts used to light a camp-fire). It does not take much imagination to realize what would happen if this heat were focused onto your eye.

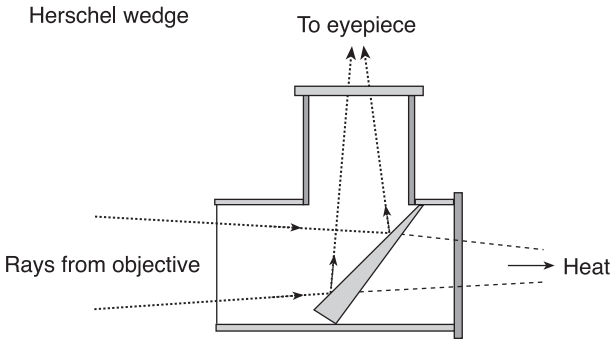


Figure 1.

We come next to observing the Sun with a telescope (binoculars will be discussed below). It is unfortunate that some small telescopes, usually (not always) REFRACTORS, are sold together with dark sun-caps. The instructions claim that it is then safe to fit the sun-cap over the telescope eyepiece, and look direct. This is not true. Such a procedure is appallingly dangerous. Dark caps supplied with small, cheap telescopes are almost always of inferior quality, so that they will not give adequate protection, and they are always liable to crack or shatter without warning, in which case an observer will not have sufficient time to move their eye out of the way. There is an extra hazard, too. Some small telescopes are supplied with 'moon-caps', designed to cut down the glare of the full Moon. So far as the Moon is concerned this is safe enough, because the Moon sends us only a tiny amount of heat, but it is only too easy to confuse a moon-cap with a sun-cap.

There is, however, an easy way to look at SUNSPOTS: use the telescope as a projector.

First, aim the telescope at the Sun, without putting your eye anywhere near the eyepiece; this is easy enough by throwing the telescope shadow on to a screen held or fastened behind the eyepiece. The Sun's disk will then be projected onto the screen, and any spot-groups or FACULAE which happen to be present will be clearly seen; adjust the focus of the telescope until the image is sharp. Adjust the distance of the screen until the projected disk is of the required size (7–8 cm is a good compromise). You can then make a sketch of the disk at your leisure.

Of course, simply holding the screen is awkward, because it means that the observer ideally needs three hands, but to fix a mounting is very easy. It is also very useful to fix a screen over the telescope itself, to cast a shadow and allow only the light from the telescope to pass through.

Detailed views of sunspots can be obtained in this way; of course different magnifications can be used—low

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

powers for general disk drawings, higher powers for more elaborate sketches of sunspot groups. The Sun rotates in a period of around 27 days, so that a spot will be carried across the disk from one side to the other until it vanishes over the limb; it will reappear at the opposite limb just under 14 days later—unless, of course, it has disappeared in the meantime! Many spot-groups have short lifetimes, though others persist for several revolutions.

The Sun has a cycle of activity with a mean period of about 11 years. At solar minimum the disk may be blank for many consecutive days, while at maximum many spot-groups may be visible at the same time. The next maximum is due around 2000–2001. *Faculae* (Latin, 'torches') are clouds of incandescent gas lying above the brilliant surface; they are easily seen by projecting the image with a small telescope.

Generally speaking, refractors are to be preferred for visual observations of the Sun; with a REFLECTOR there is always the danger of damaging the secondary mirror. Also, with a telescope more than 10 cm in aperture it is usually a good idea to stop down the mirror or object-glass. After all, with the Sun there is plenty of light to spare!

It is possible to use binoculars as projectors, in the same way as a telescope, and sunspots can sometimes be seen in this way, though it is never at all satisfactory.

There are only two sensible ways for direct telescopic viewing of the Sun. One is to use a Mylar filter placed over the object-glass, or upper end of the tube for a reflector such as a Maksutov. Again, exercise the greatest care, and make absolutely certain that the Mylar is undamaged.

The second method is to use a Herschel wedge. This is made up of a fitting for the telescope draw-tube, a partially reflecting surface, and a fitting for the eyepiece (see figure 1). The Sun's light passed through the eyepiece comes from the top surface of a plain glass wedge, while the heat, concentrated by the object-glass or mirror, is passed straight through the open end. The glass is wedge-shaped so that the secondary reflection from its back surface is not also passed up to the eyepiece—if it were, it would produce a double image. When using a Herschel wedge, also add a dark filter, and remember that infrared radiation must be taken into account as well as visible light.

There are various other points to be borne in mind. If you use a telescope to project the Sun's image, remember that the heat on the eyepiece may damage it, particularly if it is often used in this way—so put aside a special eyepiece for the purpose. Remember also to cap the telescope finder. One very famous scholar remembered that on one occasion he forgot to do this and suddenly realized that his beard had caught fire; the finder telescope had focused the Sun's heat in just that position!

Sunspots can be photographed with an ordinary camera equipped with a telephoto lens. Always bear in mind that a camera involves a lens system, and therefore never look through the viewfinder to make sure that the Sun is properly placed; and always use a Mylar filter over the camera lens. Otherwise, the only method is to project the Sun's image on to a screen, in the usual way, and then photograph the screen, but this is not very convenient, simply because the telescope itself is bound to get in the way.

ECLIPSES of the Sun are fascinating; total eclipses are spectacular by any standards—it has been said that a total solar eclipse is the grandest sight in all Nature. As soon as the last segment of the brilliant photosphere has been covered by the disk of the Moon, the Sun's chromosphere flashes into view, together with the CORONA and any PROMINENCES which happen to be present. Because the Moon's shadow is only just long enough to reach the Earth, total eclipses as seen from any one locality are rare (the English eclipse of August 1999 will not be followed by another until 23 September 2090). Partial eclipses are much more common. There are also annular eclipses, when the Moon is in the far part of its orbit and does not appear large enough to cover the Sun completely; a ring of sunlight is left showing round the dark mass of the Moon (Latin *annulus*, a ring).

During totality it is quite safe to look straight at the Sun, with the naked eye, binoculars or a telescope. But—and this cannot be too strongly emphasized—the photosphere must be completely covered. The slightest sliver of the brilliant solar disk is just as dangerous as the uneclipsed Sun. Therefore, the procedure for the eclipse photographer is to take the Mylar filter away only when the corona is first seen, and replace it as soon as the 'DIAMOND RING' effect heralds the end of totality. During partial and annular eclipses, of course, follow the same procedure as with the uneclipsed Sun.

The notes given here apply to the observer who is anxious to follow sunspots, faculae and any eclipses which may be visible, but of course serious solar observers need specialized equipment. Spectrohelioscopes and spectroheliographs are rare in amateur hands, but Hα filters are not; these cut out all the light except that of hydrogen, and once they are in place direct viewing is safe. However, the observer who has acquired equipment of this kind will need no reminder of the dangers involved in solar work.

To sum up, therefore:

(a) The Sun is dangerous, because of its brilliance, its heat and its 'invisible' radiation (ultraviolet and infrared).
(b) With the naked eye, never look straight at the Sun for more than a second or two, and preferably not at all, unless you hold up a dark filter.
(c) Welding-glass gives good protection, and so does exposed, fogged black-and-white film. Exposed color film does not, and also avoid materials such as silver paper.
(d) With a telescope, never use the dark sun-caps so often supplied with small, cheap instruments. Ensure that your filter is perfect and sufficiently dark; Mylar is much the best material.
(e) A Herschel wedge may be used, but add a dark filter to reduce the level of illumination still further. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

filter should preferably be of the type which blocks infrared radiation.

(f) Never point a camera straight at the Sun and look through the viewfinder, unless an adequate filter is used.

(g) Remember that the Sun is dangerous even when low down and shining through haze or mist. The temptation to take 'one quick look' through binoculars or a telescope which is not fitted with a proper filter may well have tragic results.

(h) Serious solar observers use equipment such as Hα filters, but they will by that time be aware of any risks.

It is true that some astronomers have been accused of being paranoid about the dangers of solar work. Yet remember that you have only one pair of eyes—and a single mistake will be one too many. A cat may look at a king, but nobody should look unprotected at the brilliant disk of the Sun.

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sundial

One of three types of timekeeping device based on the shadow cast by any fixed vertical object. As the day progresses, the Sun moves across the sky, causing the shadow of the object to move. Depending on the type of device, the object measures apparent solar time with some accuracy.

From about 3500 BC the first device was probably a vertical stick known as a gnomon. By the eighth century BC, more precise devices were in use in Egypt. One example has a base inscribed with a scale of six time divisions fixed in an east–west direction. The design is still in use in some parts of Egypt. Around 300 BC the hemispherical sundial or hemicycle is evident. It was made of a cubical block of stone or wood with a hemispherical opening cut into it. A pointer was fixed at the center of the opening. The path traveled by the shadow of the pointer was, approximately, a circular arc, which varied according to the seasons. These variations were inscribed on the internal surface of the hemisphere and each was divided into twelve equal divisions or hours. Because the length of the day varied according to the season, these hours also varied in length from season to season and even from day to day and were consequently known as temporary hours. There are records of the dial still being used during the tenth century.

The Greeks developed and constructed very sophisticated sundials. One version used the surface of a conic section for greater accuracy. The Romans also used sundials with temporary hours. The Arabs attached much importance to sundials and derived the principles and design from the Greeks. They increased the variety of designs available and, at the same time, simplified the processes of design and construction by using the principles of trigonometry. With the advent of mechanical clocks in the early fourteenth century, sundials with equal hours gradually came into general use.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

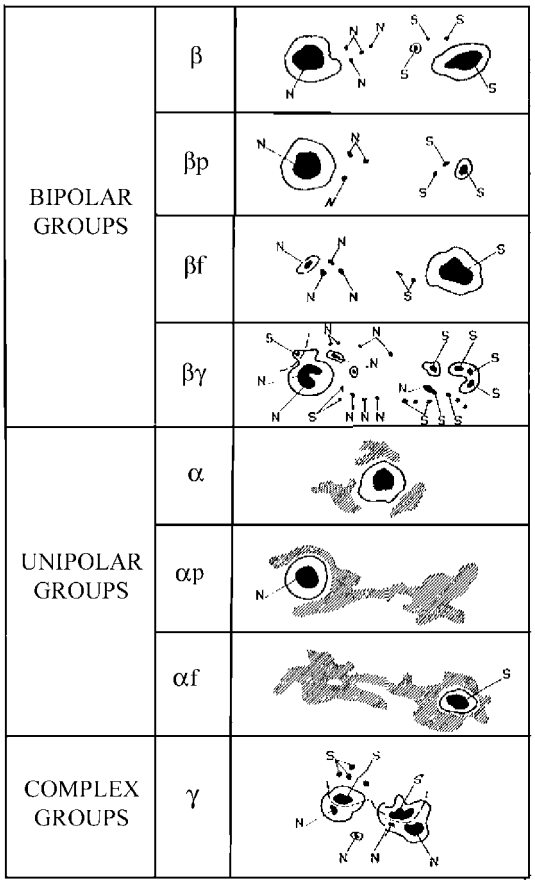


Figure 1. Solar active region classification from Mount Wilson Observatory, based on distribution of magnetic-field polarities of sunspots (from figure 6.4 in Bray and Loughhead 1964). p = preceding spot dominant; f = follower spot dominant. The p-spot 'leads' the spot group as solar rotation moves it from east to west across the solar disk. An additional category, the δ-configuration, was introduced in the 1960s.

Sunspot Classification

Sunspots, dark 'blemishes' on the brilliant solar surface, were discovered telescopically by Galileo and his contemporaries in 1611. Records of exceptionally large, naked-eye sunspots are found in Chinese and Greek records as early as 325 BC. The fortuitous occurrence of smoke or fog darkened the Sun enough to safely reveal the dark spots to the rare observer who would record his surprising view. These spots were mistaken to be birds or some mythical object passing in front of the Sun. It required regular telescopic observations and careful reduction of the observations to deduce the fundamental importance of sunspots.

Sunspots are the manifestation of strong magnetic fields that erupt from below the solar surface. Because this erupting flux is always a system of magnetic-field loops, the intersections of the two 'legs' of the loop-system with the solar surface create a bipolar group of sunspots. Thus

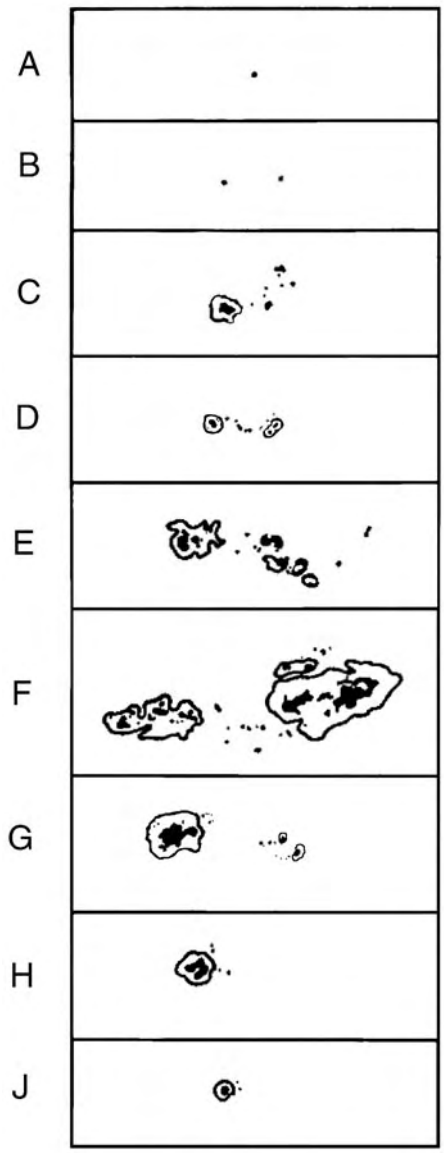


Figure 2. Drawings of typical sunspot groups in each category in the Zurich Sunspot Classification.

the most meaningful descriptive record of sunspots is their appearance in groups.

The sunspot group provides clues about the potential for eruptive activity in an ACTIVE REGION. A description of sunspot groups in a formalized system of classification provides the basis for SOLAR FLARE prediction. It is also a taxonomy of the sunspot phenomenon that becomes one of the essential data in the solar climatic record.

Sunspots are like snowflakes; each one is unique but has size and structure similar to other sunspots. The similarities are the basis for classifying the groups. Early classifications described the evolutionary sequence of a typical, large sunspot group. There was no information

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

about magnetic fields involved with these groups. Regular measurements of magnetic polarities of sunspots led the Mt Wilson Observatory to devise a simple classification scheme based on polarity distributions (figure 1).

The Zurich system (figure 2), in use for most of the 20th century, is still widely used. It describes nine different types of groups. A modification of this system by McIntosh (figure 3) was introduced in 1966 and gradually attained worldwide acceptance by the 1980s. Definitions for this more precise and detailed classification are given below, taken directly from McIntosh (1990).

The Zurich classification was modified and expanded in order to improve the objectivity of the definitions, and to add indicators of size, stability and complexity that appeared to correlate with solar flares. The Zurich system attempts to describe a typical evolutionary sequence for the largest sunspot groups. The Zurich classes are retained in modified form as the first of three components in the new classification. This eases the difficulty in persuading observers to change classification systems. The addition of new parameters takes the form of two simple sequences, expanding the classification to three components. The resulting 60 distinct types of groups produce a more accurate description of sunspot groups than the nine Zurich classes, without demanding much more of the solar observer.

Table 1. Logical sequence for determining McIntosh sunspot types.

Unipolar or bipolar?
Penumbra or no penumbra?
Penumbra on one end or both ends?
Length of group?
Rudimentary or mature penumbra?
Symmetric or asymmetric largest spot?
N–S diameter of largest spot?
Spots between leader and follower?
Mature penumbra in interior?

The general form of the McIntosh classification is Zpc, where Z is the Modified Zurich Class, p is the type of Principal Spot, primarily describing the SUNSPOT PENUMBRA, and c is the degree of Compactness in the interior of the group. Figure 3 illustrates the sequence within each of the three components of the classification. Table 1 lists the logical sequence for determining the McIntosh classification.

The definitions are formulated to require only white-light observations in the interest of consistency among all synoptic observatories. The definitions must begin with the distinction between unipolar and bipolar groups, implying a difference in magnetic topology, but not requiring magnetic-field measurements to make the distinction. Rarely will the measured magnetic class conflict with the following definitions:

Unipolar group. A single spot, or a compact cluster of spots with the greatest separation between spots $\leq 3^\circ$ heliographic (as in a great circle on the solar surface). In the case of a group with penumbra (class H), the greatest separation is measured between the center of the attendant SUNSPOT UMBRA and the nearest border of penumbra surrounding the main spot.

Bipolar group. Two or more spots forming an elongated cluster of length $>3^\circ$. Usually there will be a space near the middle of the cluster dividing it into two distinct parts. Groups with a large principal spot must be $>5^\circ$ in length (i.e., 2.5° plus 3°) to be classified as bipolar.

Modified Zurich class, Z

The modified Zurich classes are defined on the basis of whether penumbra is present, how penumbra is distributed, and by the length of the group. A judgment of complexity is not required, in contrast to the original Zurich definitions. Length is defined in absolute heliographic degrees, *not* as a projection of that length in heliographic longitude (see inset diagram in figure 3). This differs from some of the definitions already in print. The definitions of the classes were adopted from the Zurich classification, with the addition of strict limits to group length. Zurich classes G and J were omitted because they can be described by combinations of these modified Zurich classes with the other two McIntosh components. The illustrations in the left-hand column of figure 3 show typical examples of each class.

A Unipolar group with no penumbra, representing either the formative or final stage of evolution in a spot group.

B Bipolar group without penumbra on any spots.

C Bipolar group with penumbra on one end of the group, in most cases surrounding the largest of the leader umbrae.

D Bipolar group with penumbra on spots at both ends of the group, and with length $\leq 10^\circ$.

E Bipolar group with penumbra on spots at both ends of the group, and with length defined as: $10^\circ <$ length $\leq 15^\circ$.

F Bipolar group with penumbra on spots at both ends of the group, and length $>15^\circ$.

H Unipolar group with penumbra. The principal spot is usually the leader spot remaining from a pre-existing bipolar group

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

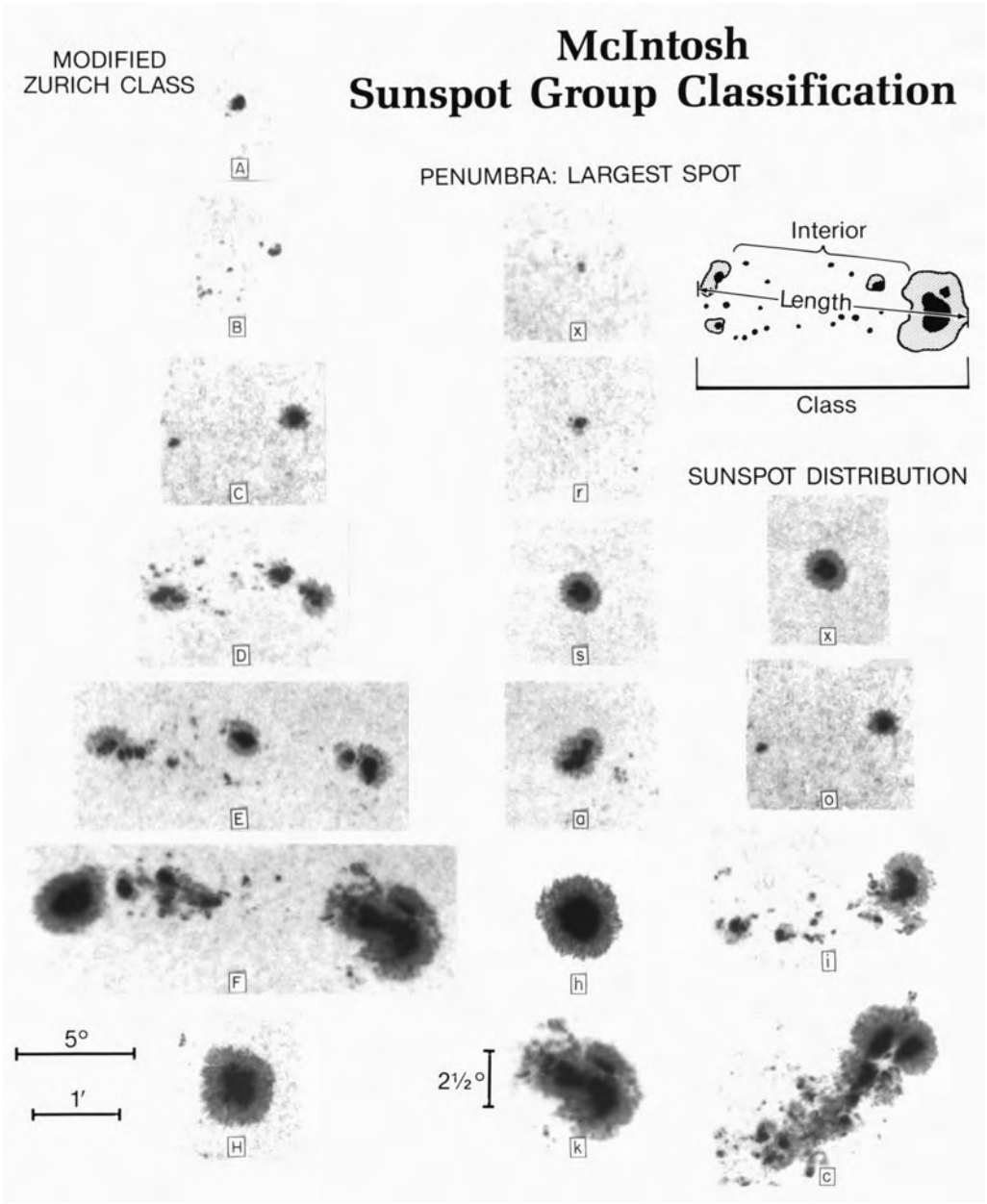


Figure 3. The three-component McIntosh classification illustrated by photographs of typical examples. Images are enlargements from the 15 cm white-light patrol at the Sacramento Peak station of the National Solar Observatory, NOAO.

Penumbra: largest spot *p*

The second parameter of the McIntosh classification is illustrated in the central column of figure 3. The type of largest spot in a sunspot group can be described by the combination of type of penumbra, size of penumbra, and symmetry of penumbra and umbrae within that penumbra. These descriptors specify size, maturity, stability and complexity in terms that are simple and, therefore, conducive to consistent classification.

x No penumbra (group is class A or B)

r Rudimentary penumbra partially surrounds the largest spot
This penumbra is incomplete, granular rather than filamentary, brighter than mature penumbra, and extends as little as 3 arc seconds (2200 km) from the spot umbra. Rudimentary penumbra may be either in a stage of formation or dissolution (McIntosh 1981, §2.3; Bray and Loughhead 1964, plate 3.7).

s Small, symmetric (like Zurich class J)
Largest spot has mature, dark, filamentary penumbra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of circular or elliptical shape with little irregularity to the border. There is either a single umbra, or a compact cluster of umbrae, mimicking the symmetry of the penumbra. The north–south diameter across the penumbra is $\leq 2.5^\circ$.

a Small, asymmetric
Penumbra of the largest spot is irregular in outline and the multiple umbrae within it are separated. North–south diameter of penumbra $\leq 2.5^\circ$.

h Large, symmetric (like Zurich class H)
Same structure as type 's', but north–south diameter of penumbra $>2.5^\circ$. Area, therefore, must be ≥ 250 millionths of the solar hemisphere.

k Large, asymmetric
Same structure as type 'a', but north–south diameter $>2.5^\circ$ and area ≥ 250 millionths. This type of spot sometimes contains spots of opposite polarity, the Potsdam δ-configuration (Kunzel 1960), and may indicate potential for proton flares (Warwick 1966).

Sunspot distribution, *c*

A simple ranking of the relative spottedness in the interior of a sunspot group gives additional information about the area of the group and, more important, whether there could be strong spots near the line of polarity inversion lying between the principal leader and follower spots. The illustrations in the right-hand column of figure 3 show the extreme examples of open and compact groups, and a typical example of an intermediate degree of spottedness in a group's interior.

x Undefined for unipolar groups (class as A and H)

o Open
Few, if any, spots between leader and follower. Interior spots of very small size. Class E and F groups of *open* category are equivalent to Zurich class G.

i Intermediate
Numerous spots lie between the leading and following portions of the group, but none of them possesses mature penumbra.

c Compact
The area between the leading and following ends of the spot group is populated with many strong spots, with at least one interior spot possessing mature penumbra. The extreme case of compact distribution has the entire spot group enveloped in one continuous penumbral area.

The logic in the definitions of parameters in this classification disallows some combinations of parameters, producing 60 distinct types of sunspot groups (table 2).

Correlations of this classification with x-ray flare occurrence proved that the classification effectively distinguishes between quiescent sunspot regions and those that are nearly certain to produce one or more major solar flares. The isolated, symmetric spots in the central column of figure 3 are typical of those that rarely produce numerous or strong flares. This is consistent with the slow rate of change in symmetric, isolated spots.

Table 2. Allowed types of groups in the McIntosh sunspot classification.

Class	Penumbra: largest spot	Distribution	Number of types
(A)	(x)	(x)	1
(B)	(x)	(o i)	2
(C)	(r s a h k)	(o i)	10
(D E F)	(r)	(o i)	6
(D E F)	(s a h k)	(o i c)	36
(H)	(r s a h k)	(x)	5
Total allowed types			60

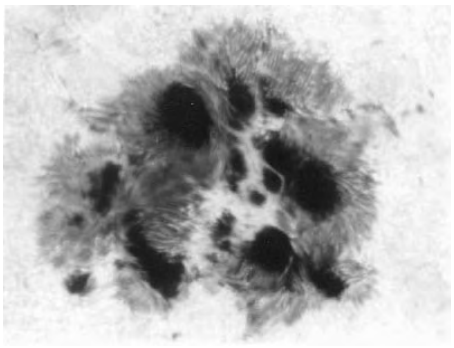


Figure 4. A classic example of a δ-configuration, the inclusion of spots of opposite magnetic polarity in a common area of sunspot penumbra (gray area). This large Dkc-type sunspot group appeared in August 1972. Photograph courtesy of German amateur solar observer G Brueckner.

Figure 4 shows a complex group whose rapid evolution correlated with a series of solar flares, with several of major proportions. This group, seen in August 1972, is a classic example of a δ-configuration, defined as the inclusion of strong spots of opposite polarity within a common area of penumbra. The δ-configuration was added to the Mt Wilson magnetic classification after the work of Kunzel (1960) at the Potsdam Observatory.

See McIntosh (1990) for detailed statistics on flare productivity versus McIntosh sunspot-group types.

Bibliography

Bray R J and Loughhead R E 1964 *Sunspots* (New York: Wiley)
Kunzel H 1960 *Astron. Nach.* **285** 271
McIntosh P S 1981 *The Physics of Sunspots* ed L E Cram and J H Thomas (Sunspot, NM: Sacramento Peak National Observatory) p 7
McIntosh P S 1990 The classification of sunspot groups *Solar Physics* **125** 251–67
Warwick C S 1966 *Astrophys. J.* **145** 215

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Patrick S McIntosh

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Cycle

The periodic increase and decrease in the number of sunspots and sunspot groups on the surface of the Sun. The number of spots and groups reaches a maximum, on average, once every 11 years. At the intervening sunspot minima, the solar disk may be devoid of spots for weeks on end. On average, the rise to maximum activity takes 4 to 5 years and the subsequent decline to the next minimum, about 6 to 7 years. The level of activity in successive cycles can vary substantially.

At the beginning of each cycle, spots start to appear at latitudes of about 30° (occasionally as high as 40°) north (+) and south (−) of the solar equator. As the cycle advances, the bands of sunspot activity migrate towards the solar equator. At solar maximum, the average latitude at which spots occur is about ±15°, and by the end of the cycle, around ±8°, by which time the first spots of the next cycle may be beginning to appear at latitudes of 30°–40°. This cyclic variation in the mean latitudes of sunspots was pointed out in 1861 by Friedrich Spörer (1822–95), and is known as Spörer's law. When the latitude of each sunspot is plotted against time over a series of 11-year sunspot cycles, the resulting distribution of spot positions forms a pattern reminiscent of the appearance of butterfly wings and is, therefore, known as the butterfly diagram. A diagram of this kind was first plotted in 1904 by the English astronomer E W Maunder (1851–1928).

In each spot pair, the spot which is ahead, in the sense of the direction in which the Sun rotates, is called the leader, and its companion, the follower. Throughout a complete cycle, from one minimum to the next, all of the spot pairs and groups in the northern hemisphere have the same polarity pattern, and those in the southern hemisphere have the opposite. For example, in one particular cycle, all the leaders in the northern hemisphere will have positive polarity and all the followers negative, whereas all the leaders in the southern hemisphere will have negative polarity and the followers positive. At the end of that cycle, the polarity pattern reverses, so that in the subsequent cycle all northern hemisphere leaders will have negative polarity and all southern hemisphere leaders, positive. The complete magnetic cycle consists of two successive 11-year sunspot cycles and is, therefore, 22 years long.

During the earlier part of each cycle, the leading spot in each pair (or the leading area of net magnetic polarity in each group) has the same magnetic polarity as the net polar polarity in the hemisphere within which it is located. As each spot or group decays, the polarity of the follower preferentially diffuses towards the pole. The cumulative effect of this process eventually produces a reversal of polarity at each of the solar poles. This usually occurs around the time of solar maximum, but considerable variations occur and, while the piecemeal reversal is taking place, the Sun, for a time, may have the same net polarity at both poles.

The sunspot cycle is part of an overall solar cycle whereby all forms of solar activity, including sunspots, plages, prominences, flares and coronal mass ejections, together with the shape, extent and structure of the chromosphere and corona, undergo cyclic variations with a period of about 11 years. The numbers, sizes and energies of prominences, flares and coronal mass ejections mirror the increase and decrease in sunspot numbers and the corona is brighter, more extensive and more symmetrical around solar maximum than at times of minimum activity.

In addition to the 11-year cycle, there is some evidence to suggest that sunspot numbers, and levels of solar activity as a whole, undergo longer-term modulations over periods of 80 years and more. Furthermore, historical records imply that there have been prolonged periods of enhanced and depressed levels of activity in the past. In particular, solar activity appears to have remained at an unusually low level between 1645 and 1715, a period that is known as the Maunder Minimum, having been named after E W Maunder, whose analysis of old solar records showed that hardly any sunspots at all had been seen during that time.

See also: active region, corona, flare, plage, prominence, relative sunspot number, solar coronal mass ejection, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Evolution

SUNSPOTS are observed to grow from their first appearance at the solar surface to some maximum size and then to decay until they disappear after some hours, days, or weeks. The characteristics of this growth and decay, the changes during the lifetime of a sunspot group in its distribution of sunspots, surface orientation, magnetic field line inclination, and other characteristic variations give us clues to the subsurface origin of sunspots and to the dynamo process (see DYNAMO THEORY) which produces the magnetic fields that define sunspots.

Sunspots and sunspot groups

Sunspots are dark, generally roughly circular areas in the solar photosphere where there are strong magnetic fields that are approximately vertical to the solar surface in the SUNSPOT UMBRA. These magnetic fields are quite strong—several thousand gauss near the center of the largest sunspots and somewhat weaker in smaller sunspots. The darkest central part (umbra) of a sunspot may have an intensity in the visible part of the spectrum that is only about 5% that of the surrounding solar surface. The outer part of a sunspot (SUNSPOT PENUMBRA) has an intensity much closer to that of the surrounding atmosphere.

Sunspots normally appear in groups of two or more. These are called 'sunspot groups'. Single-spot groups are rare. Sunspot groups are usually aligned along an axis that is oriented approximately (but on average not exactly) in the east–west direction on the Sun. Figure 1 shows an example of a sunspot group with a rather simple configuration. Often the westernmost (leading, in the sense of solar rotation) and easternmost (following, in the sense of solar rotation) sunspots are clearly separated. In general, the leading and following sunspots have opposite magnetic polarities, and this orientation of polarities is predominant in one (north or south) hemisphere during a SOLAR CYCLE. The polarity orientations are opposite in the other hemisphere. Each 11 yr solar cycle has the opposite sunspot-group polarity orientation to the previous solar cycle. One can imagine an axis drawn, connecting the leading and following sunspots (or the centroids of the leading and following collections of sunspots). The angle which this axis makes with the local parallel of latitude is referred to as the 'tilt angle'. The average tilt angle is about +5°, where the plus sign indicates that the leading sunspots are (on average) equatorward of the following sunspots in each hemisphere.

Area development

Sunspot diameters range from around 2000 km to around 30 000 km. A sunspot first appears as a 'pore', which is a very small dark speck within the photospheric granulation pattern. The great majority of SUNSPOT PORES live only for a few tens of minutes, but on rare occasions a pore will grow to become a small sunspot umbra. A sunspot group starts with the appearance of one or more pores, usually in the western (leading in the sense of rotation) portion of the future group. On the second day in the lifetime of the group the small but growing sunspots spread out into an elongated pattern roughly in the east–west direction. This means that individual sunspots move (exhibiting what is called 'proper motion') systematically in such a way as to produce the stretched appearance that is common of normal groups. Leading sunspots move forward (westward), and following sunspots move backward (eastward). This motion often ceases, or slows down, when the group reaches its maximum area. There is also often latitudinal motion at this time, which produces a net rotation of the sunspot group about a vertical axis. This is discussed in more detail below.

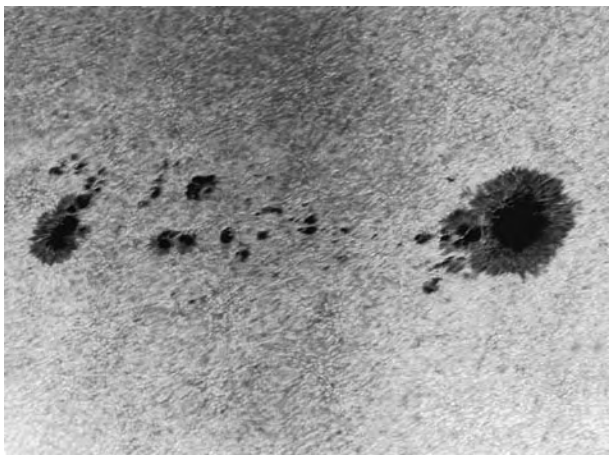


Figure 1. A typical large sunspot group. Note the leading (on the right side) and following (on the left side) sunspots. Also note the very small sunspots (pores) distributed near the group.

As the sunspot group develops in this way, larger sunspots form by the coalescence of smaller sunspots nearby. The smallest sunspots do not have penumbrae. The diameters of penumbrae are generally slightly more than twice those of the associated umbrae. The total area of a sunspot group increases in the first few days of the development of the group, and then quickly the total area begins to decrease until the group has disappeared.

Most sunspots grow rapidly in the first few days after their initial appearance. Sunspots and sunspot groups show a variety of area developments. Some grow and decay slowly. Some groups, especially larger, longer-lived groups, grow rapidly and decay slowly. The decay of a sunspot is not the inverse of the coalescent process that forms sunspots; sunspots decay by the continuing outward flow of bits of the magnetic field, in the form of very small (≈1000 km) parts of the umbra, which are observed to move radially outward through the sunspot penumbra to the photosphere. It is difficult to follow observationally these small magnetic elements as they move from the sunspots. It is presumed that, once they leave the sunspot, they cancel quickly with magnetic fields of opposite polarity in the neighborhood of the sunspot, and there is some observational evidence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

for this. Thus sunspots decay by a directed flow of magnetic flux radially outward from the sunspot, which is quite different from the decay of the weaker PLAGE magnetic fields, which always accompany sunspots. Plage magnetic fields decay by the gradual erosion of small elements of magnetic flux from the periphery of the plage. This erosion is caused by nearby, random, large-scale convective motions, which are present at all times over the whole solar surface (except within sunspots and within the plages themselves). This large-scale cellular pattern of convection is called 'supergranulation' (see SOLAR PHOTOSPHERE: SUPERGRANULATION).

Lifetimes

The lifetimes of sunspots range from tens of minutes to months. More than half of sunspots have lifetimes less than 2 days. Ninety per cent of sunspots have lifetimes less than 11 days. Only a few sunspots per 11 yr solar cycle have lifetimes of more than a few months. On average, high-latitude sunspots have relatively short lifetimes. Sunspots with lifetimes of more than several months are extremely rare. The longest-living sunspots are always large, round, single-spot groups, at least in the later stage of their lifetimes. They are generally the last survivors of large, multispot sunspot groups. It is not known why a few sunspots have very long lifetimes, while nearly all sunspots of similar size do not live nearly as long.

A rule of thumb for sunspot group lifetimes is given by the formula

$$T = 0.1A_{max}$$

where T is the lifetime in days and A_{max} is the group area in millionths of the hemisphere.

The sunspots with the shortest lifetimes are very small (no more than a few thousand kilometers in diameter) and have no penumbrae. These smallest sunspots have lifetimes of only a few hours, if they do not grow into full-sized sunspots, as discussed above. The vast majority of sunspots do not develop beyond the pore stage. Pores (that do not develop into sunspots) may live for only a few tens of minutes. Often sunspot groups, especially in the earlier stage of their development, have many pores associated with them at any one time. Note the pores associated with the sunspot group in figure 1.

Magnetic fields

Magnetic fields may be considered to define sunspots. That is to say, all sunspots contain strong magnetic fields. These magnetic fields, in the highly ionized solar gas, impede the convection process which in the normal photosphere is the principal mechanism for the transport of thermal energy from deep within the solar interior. Thus sunspots are darker than the surrounding photosphere (although in an absolute sense they are still quite bright).

Measurements of magnetic field strengths within sunspots are difficult because of the presence of scattered light from the much brighter photosphere surrounding the sunspot. These problems are especially severe for the smallest sunspots. Field strengths of several thousand gauss are measured in the largest sunspots. These represent maximum values, which are generally found at the center of the umbra. Field strengths decrease as one measures closer to the umbra–penumbra boundary.

These sunspot field strengths are observed to increase rapidly in the initial growth stage of a sunspot and then remain relatively constant for a large part of the lifetime of the sunspot. A decrease in the field strength occurs only in the last stages of decay of the sunspot. This is in contrast to the behavior of the sunspot area, discussed above, which generally also increases rapidly following the first appearance of the sunspot but which then begins to decrease shortly after the sunspot reaches its maximum size. The magnetic flux, of course, which may be considered to be the product of the area and the magnetic field strength, shows the behavior of the area during the lifetime of the sunspot. That is, the flux increases rapidly and then it quickly begins its slow decline until the sunspot disappears.

Orientations

Surface orientations

As discussed above, one of the observable characteristics of a sunspot group is its tilt angle. There is considerable dispersion in the tilt angles around the average tilt angle (+5°) mentioned above. In general about 36% of sunspot groups have tilt angles within 10° on either side of the average angle. Only about 13% of sunspot groups have tilt angles more than 45° from the average angle in either direction. Extreme tilt angles result in magnetic polarity orientations that are reversed from the normal, i.e. where the leading magnetic polarity sunspots are located east of (or following in the sense of solar rotation) the following magnetic polarity sunspots. Nearly 10% of groups show such reversed-polarity orientations. Many of these are small sunspot groups.

It is generally believed that the magnetic fields of sunspot groups rise to the surface as large loops of magnetic field lines from the base of the solar convection zone, which means from about one-third of a radius below the surface. As these loops rise through the solar convection zone, propelled by the buoyant force on the ionized gas that is confined to the magnetic field lines, they are acted upon by various physical forces. One of these forces is the Coriolis force, which will act to twist the plane of the loop (which is assumed to start its rise aligned in the east–west direction). Another force that acts on the rising magnetic flux loop is magnetic tension, which acts to restrain the rise of the flux loop and other motions of the loop. Mathematical models of these forces indicate that the twist of the axis that one would expect from such theoretical considerations is about what is observed as the average tilt angle in each hemisphere. Moreover, the variation of tilt angle with latitude that is observed (known as Joy's law) can also be reproduced in the theoretical models. Other observed characteristics of sunspot groups

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

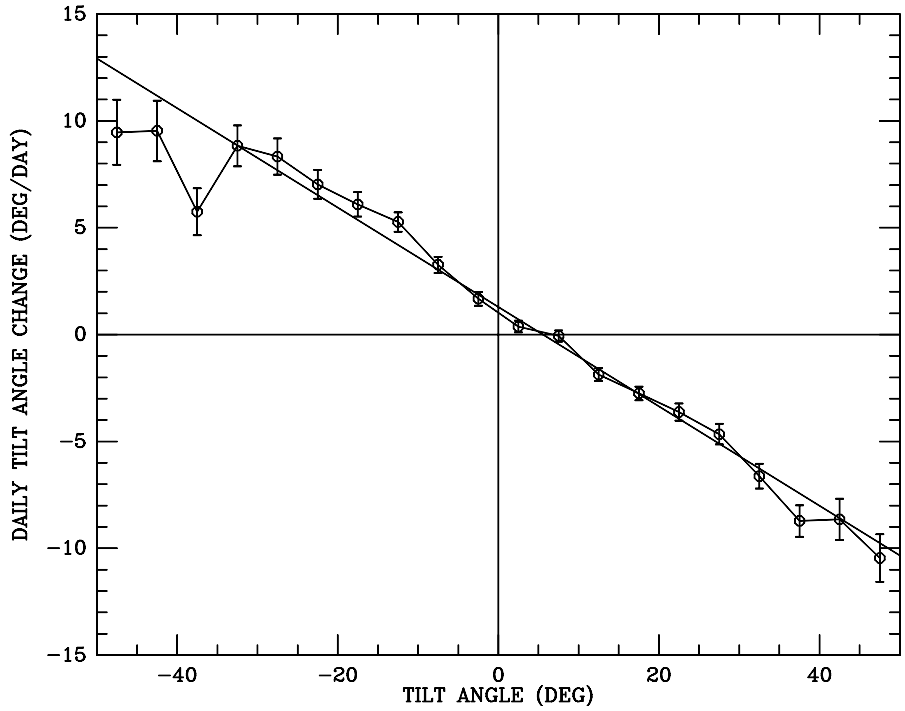


Figure 2. Average value of the daily rotation of the magnetic axes of sunspot groups (tilt angles) for various values of the starting tilt angle. Note that the intercept on the Y-axis of the plot is at 5°, not at 0°. This indicates that sunspot group tilt angles tend to rotate toward the average tilt angle, not toward the east–west direction (0°).

can also be seen in the calculations of these mathematical models.

In addition to these physical forces, the rising magnetic flux loops will be subjected to added random twisting from large-scale convective motions that exist deep within the solar convection zone. It is believed that these random interactions are responsible for the dispersion in tilt angles that is observed in sunspot groups. Smaller sunspot groups show a larger dispersion about the average tilt angle than do larger sunspot groups. This is what one would expect to find if large-scale motions of the gas are responsible for the distribution in the tilt angles: the magnetic flux loops of the larger and stronger sunspot groups can withstand such buffeting from the convective motions better than can the smaller and weaker flux loops, and thus they show a smaller dispersion in their tilt angles.

Magnetic field-line inclinations

Sunspot magnetic field orientations to the vertical (in the east–west direction) can be estimated by examining the areas or magnetic field strengths of sunspots east and west of the central MERIDIAN of the Sun (the north–south meridian that is coincident with the rotation axis on the projected image of the Sun). It is assumed that the east–west differences in area or magnetic field strength are due to east–west orientations of the field lines in the solar atmosphere. In principle it should be possible to measure inclinations of sunspot magnetic fields by the measurement of the vector magnetic field in the sunspots, but in practice this is difficult to do with high precision for the small inclinations that are known to exist in sunspots. The transverse ZEEMAN EFFECT (measuring field strength in field lines perpendicular to the line of sight) is not nearly as sensitive a measure of field strength as is the longitudinal Zeeman effect (along the field lines) measurement.

On average, using this east–west difference technique, the magnetic field lines of sunspots are found to be inclined a few degrees to the east, that is, to trail the rotation—which is from east to west. The inclinations of the leading and following sunspots in the groups are on average away from each other by a degree or so. This means that the following sunspot magnetic fields are inclined to the east by a slightly greater angle than are the leading sunspot magnetic fields. Sunspots whose areas are growing (therefore, presumably younger sunspots) show greater eastward inclination of their magnetic field lines by several degrees than do sunspots whose areas are decreasing. Thus we conclude that during the lifetime of the group there is a shift in the inclination angle toward the west.

Tilt angle changes

Sunspot group tilt angles are sometimes observed to change with time. This means that the axes of the groups rotate (in the plane of the solar surface). This rotation is generally very slow. Very few sunspot groups show

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

rapid rotation about their axes, and, when they do, it is usually early in the development of the group. Fully 46% of sunspot groups show daily rates of rotation of their axes of less than 10° in either direction. Only about 14% of sunspot groups on any one day have rotation rates greater than 45° day^{-1} in either direction.

However, although generally sunspot group axis rotations are slow, they tend to exhibit systematic motions. On average, sunspot group axes rotate in the direction toward the average tilt angle in each hemisphere, as may be seen in figure 2. Another way of stating this is that on average the only tilt angles for which there is no average tilt angle rotation is the average tilt angle. This is represented in the figure as the intercept on the X-axis. This intercept is located at the average tilt angle of about 5° rather than at the origin. This is a result that is clearly above the noise level.

This rotation of the axes of sunspot groups is not understood by astronomers in the context of our current view of the origin of the magnetic fields at the solar surface. It is possible that the various physical forces that act on the sunspot magnetic flux loop as it rises through the solar interior can twist the flux loop, as mentioned above, but once the magnetic fields reach the surface it is difficult to explain any further rotation of the magnetic axis of a group. Because of the motions of the subsurface, large-scale turbulent convection, which as we saw above may be responsible for the dispersion of the sunspot group tilt angles, it is difficult to understand how there can be within the convection zone any fixed location where the ends of the magnetic loop could be anchored. It is equally difficult to understand how the surface orientation could possibly be affected by the orientation of the magnetic flux loop at the base of the convection zone, which is one-third of a radius beneath the solar surface.

One possible explanation for this curious behavior of the magnetic axes of sunspot groups is that there is, in fact, no real rotation motion at all of the tilt angle in growing sunspot groups. Rather, the plane of the magnetic flux loop that makes up the sunspot group has become twisted within the convection zone, and as it rises it gives the appearance of a rotary motion at the plane of the surface.

Bibliography

Balasubramaniam K S and Simon G W (eds) 1994 *Solar Active Region Evolution: Comparing Models with Evolution (Conference Series, vol 68)* (San Francisco, CA: Astronomical Society of the Pacific)

Bray R J and Loughhead R E 1965 *Sunspots* (New York: Wiley)

Howard R F 1996 Solar active regions as diagnostics of subsurface conditions *Ann. Rev. Astron. Astrophys.* **34** 75–109

Kiepenheuer K O 1953 Solar activity *The Sun* ed G P Kuiper (Chicago, IL: The University of Chicago Press) pp 322–465

Robert F Howard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Magnetic Fields

The magnetic field is the central quantity determining the property of a SUNSPOT. It permeates every part of sunspots and by greatly reducing the convective transport of heat from below is also responsible for sunspot darkness. Conversely, sunspots were the first astronomical objects recognized (by G E Hale in 1908) to harbor a magnetic field.

The field strength B is most readily measured in the photospheric layers. There it reaches peak values of 2000–3500 G (i.e. 0.2–0.35 T) in parts of the sunspot umbra (the spread in values is intrinsic—larger sunspots have higher maximum field strengths). The field strength drops steadily towards the edge of the sunspot, where it is 700–1000 G (0.07–0.1 T). The strongest field within a sunspot is usually associated with the darkest part of the umbra (umbral core) and is generally close to vertical, while at the visible sunspot boundary it is inclined by 70–80° to the vertical. The magnetic structure of a mature sunspot is illustrated in Color Plate 53.

The magnetic field of a sunspot nevertheless continues well beyond its white-light boundary as an almost horizontal magnetic canopy with a base in the mid to upper photosphere. The field strength above the height of the canopy base continues to decrease slowly but steadily for increasing distance from the white-light sunspot. The field strength also decreases with height, with the decrease being more rapid in the photosphere than in higher atmospheric layers.

These observed properties of sunspot magnetic fields support the theoretical concept that visible-light sunspots are the intersection of the solar surface with a large magnetic flux tube emerging from the solar interior into the atmosphere.

At small scales the umbral magnetic field appears to be relatively homogeneous, while the penumbral field is filamented into two radially directed components. These differ by their inclination to the vertical, and possibly also their field strength.

In this article the observed magnetic structure of (mature) sunspots is discussed. The theoretical description of the field structure is outlined in the article SUNSPOT MODELS and the magnetic field of pores (i.e. dark magnetic structures without a penumbra) in the article SUNSPOT PORES.

Overviews of the observed magnetic structure of sunspots have been given by Martínez Pillet (1997) and Skumanich *et al* (1994), and may be found in the volume edited by Thomas and Weiss (1992).

Large-scale magnetic structure of sunspots

The field strength

The magnetic field of sunspots in photospheric layers has been measured via the Zeeman splitting of absorption lines in the visible and the infrared (see the article on SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY and on SOLAR MAGNETIC FIELD: ZEEMAN AND HANLE EFFECTS). The magnetic field strength is largest near the geometrical center of regular, i.e. reasonably circular, sunspots and drops monotonically and smoothly outwards, reaching its smallest values at the outer penumbral edge. This is in stark contrast to the brightness, which jumps at the boundary between the umbra and penumbra. Hence, the umbral boundary is not evident in the field strength. This simple picture of a relatively smooth magnetic distribution is valid for a spatial resolution of 2–3″ or lower. At higher resolution the fine-scale structure of the field becomes increasingly prominent (see below).

There is now a consensus on the general form of the normalized field-strength distribution, $B(r/r_p)/B_0$, although some uncertainty concerning its exact shape remains. Here r is the radial coordinate, r_p is the radius of the outer penumbral boundary and B_0 is the (maximum) field strength at the center of the sunspot (i.e. at $r = 0$). The radial dependence for typical sunspots of the three components of the magnetic field is plotted in figure 1B. Magnetic dipole fits to such data are reasonable, but not perfect in many cases.

The maximum field strength (i.e. B_0) increases almost linearly with sunspot diameter from $B_0 \approx 2000$ G for the smallest spots to over 3500 G for the largest spots. At the sunspot boundary $B(r_p) \approx 700$–1000 G, which implies $B(r_p)/B_0 \approx 0.2$–0.4. The fact that $B(r_p)$ is so large suggests that sunspots are bounded by a current sheet.

Magnetic field orientation

The strongest part of the sunspot magnetic field is in general also the most vertical. Inclination to the vertical (i.e. zenith angle, ζ) increases steadily as the field strength decreases. For symmetric sunspots ζ also increases steadily with r, i.e. the field is increasingly inclined to the vertical the further out in the sunspot that it is measured. Recent evidence suggests that the inclination is 10–20° to the horizontal (i.e. $\zeta \approx 70$–$80°$) at the sunspot boundary. Typical $\zeta(r/r_p)$ values are plotted in figure 1C.

Early vector-magnetograph observations often suggested that the magnetic field of regular sunspots is significantly twisted. These strong twists or spirals were an artefact of neglecting magneto-optical effects, however. Once such effects are included in the analysis the magnetic fields of regular sunspots become almost radial. A small residual twist of up to 10–20° still appears to persist, however. A residual twist is not entirely surprising considering the fact that the superpenumbral fibrils seen in Hα are increasingly twisted with increasing distance from the sunspot.

Magnetic canopy

Above the solar surface the magnetic field continues beyond the white-light boundary of sunspots. It forms an almost horizontal canopy with a base in the middle or upper photosphere, i.e. the field is limited to layers above the canopy base and overlies mainly field-free gas. The magnetic canopy is a natural result of the expansion with height of the magnetic flux tube underlying the sunspot. Recall that the visible sunspot is just a cross section through

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

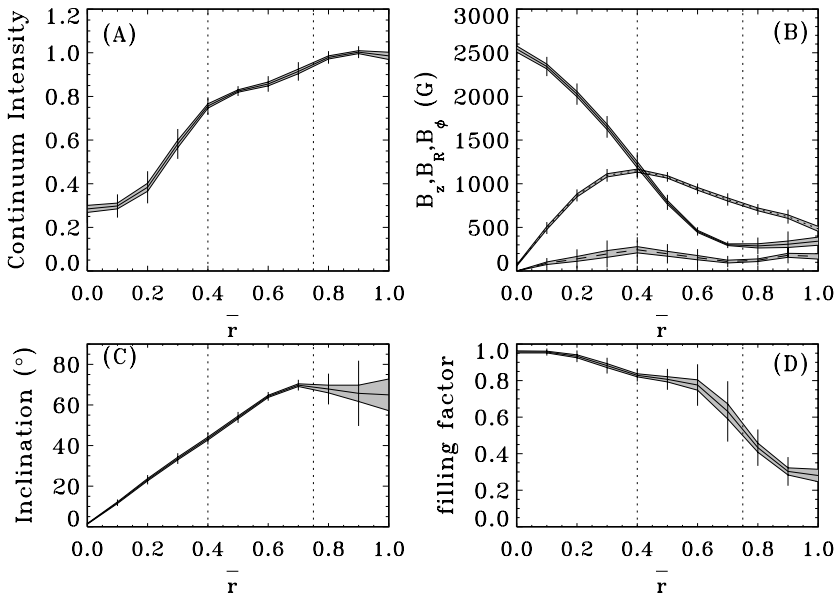


Figure 1. Typical intensity and magnetic parameters of sunspots versus normalized radial distance, $\bar{r}$, from sunspot center. Vertical dotted lines indicate the umbra–penumbra (left) and the penumbra–canopy (right) boundaries. Plotted are the continuum intensity in panel A, vertical (B_z, solid curve), radial (B_r, dotted curve) and azimuthal (B_ϕ, dashed curve) components of the magnetic field in panel B, magnetic inclination in panel C and magnetic filling factor in panel D. $\bar{r}$ is normalized to the radius at which the canopy could not be seen anymore. (From Keppens and Martínez Pillet (1996) with permission.)

this flux tube. In this picture the canopy base corresponds to the current sheet surrounding the sunspot.

$B(r/r_p > 1)$ is found to decrease continuously without any visible break at the white-light boundary of the sunspot. The base height of the canopy increases relatively rapidly close to the sunspot. However, this rise soon slows, so that the canopy can be followed out to almost twice the sunspot radius using purely photospheric lines.

Sunspot canopies were first discovered by W Livingston in magnetograms taken in spectral lines formed in the upper photosphere and lower chromosphere, where they appear as large diffuse patches of field filling the limbward and discward superpenumbrae of sunspots observed near the solar limb. Newer techniques also allow canopies to be detected when the sunspot is close to disc center. They all give roughly consistent results.

Return flux and depth of the penumbra

Inside the penumbra the general inclination of the magnetic field suggests that a considerable amount of magnetic flux is emerging through the penumbra. An analysis based on the premise that the sunspot magnetic field is potential inside the bounding current sheet allows more quantitative estimates to be made and suggests that approximately 1–1.5 times as much magnetic flux emerges in the penumbra as in the umbra. The penumbra is thus deep, in contrast to a shallow penumbra in which the current sheet bounding the flux tube lies along the solar surface in the penumbra and no magnetic flux emerges there.

A relatively small fraction of the total flux emerging in a sunspot returns to the surface inside or just outside the sunspot boundary. This so-called 'return flux' has been observed at some locations along the boundary of the sunspot. These locations are co-spatial with downflows which are interpreted as a part of the Evershed flow draining down into the solar interior again.

Vertical gradient of the magnetic field

In addition to the large number of observations in the lower and central photospheric layers the sunspot magnetic field has been recorded at greater heights as well.

Basically, the observations suggest that the maximum field strength at the center of regular sunspots decreases with height, but still has values of 1000–1800 G in the upper chromospheric to lower coronal layers. The magnetic field is in general more homogeneous in the upper atmosphere. It can be followed over a larger distance in the upper atmosphere and decreases more slowly with distance from the sunspot axis.

Nevertheless, the field in the coronal layer is often more confined than expected from a dipole fit to photospheric observations and provides evidence for non-potential fields. On the other hand in parts of an active region coronal fields are well reproduced by potential field extrapolations from the photosphere.

Knowledge of the magnetic field at different heights allows vertical gradients of the field strength to be determined (and to be compared with the predictions of models). The values of the obtained vertical gradient

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

depend on both the horizontal position in the sunspot and the height range over which they are measured (i.e. the formation height difference between the two diagnostics used).

First consider measurements comparing lines formed in the photosphere, i.e. those determining $|\partial B/\partial z|$ over a small height range. These give values around 2–3 G km^{-1} in the umbra and inner parts of the penumbra, dropping to around 1 G km^{-1} in the outer penumbra.

When lines formed at sufficiently different heights are compared (i.e. usually a photospheric line with either a chromospheric or transition region line, or a coronal diagnostic) then considerably smaller vertical gradients are obtained: 0.3–0.6 G km^{-1} in the umbra, which drops to 0.05–0.2 G km^{-1} near the outer penumbral edge. This dependence on radial position can also be parametrized as $|\partial B/\partial z| \sim |B|$.

The observations therefore suggest that the field strength drops rapidly with height in the photosphere and much more slowly at greater heights. The qualitative height dependence agrees with the expectations of a flux-tube model. Almost all the observations also indicate a decrease of $|\partial B/\partial z|$ from the center of the sunspot toward its periphery (or from stronger to weaker field).

Fine-scale structure of the magnetic field

The fine-scale structure visible in white-light images of sunspots is predominantly due to umbral dots and light bridges (see SUNSPOT UMBRA: STRUCTURE AND EVOLUTION) and to penumbral filaments. These and related features are also prominent in filtergrams, although at the greater heights sampled by some filtergrams somewhat larger scales may dominate. As far as the magnetic field is concerned the main known inhomogeneities are concentrated in the penumbra (see SUNSPOT PENUMBRA: STRUCTURE AND ACTIVITY).

Umbral dots

Umbral dots are bright dot-like features with horizontal dimensions often lying at the limit of the spatial resolution. Various investigators have searched for the signature of umbral dots in the magnetic field. Whereas older observations showed a huge scatter in deduced field strength values, there has been convergence toward a rough consensus in recent years. The investigations carried out in the 1980s and 1990s suggest that the field strength in an umbral dot is at the most 20% lower than in the surrounding umbral background.

Such a reduction in field strength in umbral dots may well be due to the fact that B decreases with height and the observed radiation is emitted from higher layers in hot and bright regions, so that at a given geometrical height the field may be equally strong in the dot and outside it.

Fluted magnetic field in the penumbra

The azimuthal inhomogeneity on small scales of the penumbral magnetic field, in particular of its zenith angle, is now well established.

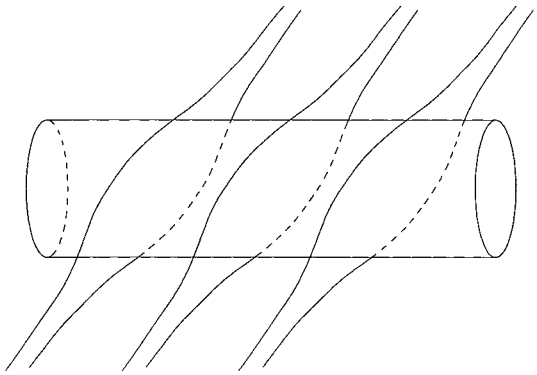


Figure 2. Sketch of the local fine-scale structure of the magnetic field in sunspot penumbras. The field is composed of two components, a flux-tube component, represented by the horizontal cylinder, and a more inclined magnetic field, indicated by the field lines crossing the flux tube at an angle (from Solanki and Montavon 1993).

Less clear than the mere presence of fluctuations in the zenith angle ζ is the correlation of the zenith angle with brightness at a fixed distance from the center of the sunspot. All three possibilities, namely positive, negative and no significant correlation have been proposed in the literature. There does appear to be a significant correlation of ζ with the velocity, however, with the flow being concentrated in the horizontal magnetic filaments. The fine-scale structure of the penumbral field remains remarkably constant on a time-scale of minutes to hours.

Another point that is not completely decided is whether, in addition to the inclination, also the field strength in the penumbra is inhomogeneous in the azimuthal direction. Some investigators see no fluctuations in B, while others find 100–400 G higher B values in dark penumbral fibrils than in their bright counterparts. Yet others see evidence for weaker fields in cooler parts of the penumbra. A final group of researchers observe that the more vertical filaments have a 200 G higher field strength.

From a combination of high-resolution imaging and broad-band circular polarization (BBCP) measurements it follows that there must be a horizontal boundary between the two differently inclined magnetic components in the penumbra. Hence they cannot simply be vertical sheets of magnetic flux. The penumbral fine-scale field is better described by (more or less) horizontal flux tubes embedded in an inclined field, as illustrated in figure 2. The width of the cross-section of these flux tubes is as yet uncertain.

Magnetic filling factor

In order to explain phenomena such as umbral dots the intrusion of field-free material into umbras at small scales has been proposed. In addition, the presence of field-free material in penumbras could resolve the mass-flux problem of the Evershed effect in a rather straightforward

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

manner. In general, only upper limits on the area coverage by field-free material inside sunspots can be given. The exceptions are light bridges, i.e. bright material separating two umbras within a single sunspot, which exhibit direct evidence for the presence of narrow lanes of field-free material in sunspots.

The upper limit lies at 5% in the central penumbra and at 3–4% in the central part of the umbra. Near the umbral boundary the upper limit lies around 15% and is also large close to the outer penumbral boundary. Figure 1D shows an example of less stringent upper limits on the amount of field-free material (or rather of a lower limit on the magnetic filling factor, i.e. the area fraction covered by the field). One caveat to keep in mind when considering these very tight limits is that they refer mainly to mid-photospheric layers. Since the magnetic energy begins to dominate over that of the gas at these heights, but does not in the lower photosphere, it is far more likely that if field-free patches exist that they are located in the lower photosphere. In the lower layers of sunpots the upper limit is less tight, with the most stringent value being 5% in the central penumbra.

Bibliography

Keppens R and Martínez Pillet V 1996 *Astron. Astrophys.* **316** 229

Martínez Pillet V 1997 *1st Advances in the Physics of Sunspots Euroconference, Advances in the Physics of Sunspots (Astron. Soc. Pacific Conf. Ser. 118)* ed B Schmieder, J C del Toro Iniesta and M Vázquez p 212

Skumanich A, Lites B W and Martínez Pillet V 1994 *Solar Surface Magnetism* ed R J Rutten and C J Schrijver (Dordrecht: Kluwer) p 99

Solanki S K and Montavon C A P 1993 *Astron. Astrophys.* **275** 283

Thomas J H and Weiss N ed 1992 *Sunspots: Theory and Observations* (Dordrecht: Kluwer)

Sami K Solanki

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Models

Models of SUNSPOTS aim either to reproduce observed properties of sunspots, or to understand the physical processes occurring in them. Sunspot models are of very diverse types, ranging from empirically derived models of their thermal stratification to MHD models of their magnetic configuration and evolution. The final aim of all modelling is to obtain a consistent and detailed description and understanding of the structure and dynamics of the magnetic and thermodynamic properties of a sunspot. Such comprehensive models are not feasible as yet. Hence most models deal with some aspect of sunspots in detail, while neglecting or simplifying the other aspects. Most numerous among the physical models are those describing the magnetic structure of a sunspot (see article SUNSPOT MAGNETIC FIELDS) on the basis of an axially symmetric vertical flux tube in (approximate) magnetohydrostatic equilibrium. Another group of physical models deals with the other main characteristic of sunspots, namely their darkness relative to the quiet Sun. They model the heat flow by radiation and convection in and sometimes also around the sunspot. The models with the most detailed description of energy transport, e.g. an explicit simulation of magnetoconvection, generally treat only a part of the sunspot (e.g. a part of the SUNSPOT UMBRA). In the most comprehensive models, in which the magnetic field structure of the whole sunspot is determined, the convection is usually dealt with using the mixing length formalism.

In this article a description of physical models of the magnetic and thermal structure of sunspots is given. Physical models of other features of sunspots, e.g. the Evershed effect or sunspot oscillations, are to be found in other articles (see SUNSPOTS: EVERSHED EFFECT, SUNSPOT OSCILLATIONS AND SEISMOLOGY). The same is true for empirical models, i.e. models of the sunspot magnetic or thermal structure that are derived from observations.

Since the thermal and magnetic structure of sunspots is known in considerable detail from observations, rather tight constraints can be set on physical models. One obvious difference between actual and modeled sunspots is that whereas the former are generally irregular, the latter almost always possess axial symmetry. However, relatively regular sunspots do exist, which can be used to test whether the models reproduce the observed radial dependence of the field strength, field direction, brightness, Wilson depression, etc. Although models in many cases provide reasonable descriptions of sunspots and can reproduce many observations, they provide far fewer *ab initio* explanations. There is as yet no satisfactory explanation for such basic questions as to why sunspots have penumbras, or why umbras carry 20% and penumbras 75% of the heat flux of the quiet Sun (see SUNSPOT PENUMBRA).

Theoretical models of solar magnetic features can be divided into three classes. Firstly, there are numerical experiments, usually idealized simulations of some physical process, such as magnetoconvection, thought to act in the magnetic feature. The second class is composed of static descriptions of symmetric flux tubes that aim to reproduce the global observed properties of sunspots at the cost of some physical realism. Finally, full-fledged simulations make up the third class. These include time dependence, compressibility, partial ionization, radiative transfer, two- or, ideally, three-dimensionality, a fine spatial and temporal grid and a sufficiently large computational domain.

Unfortunately, the large size of sunspots and the small-scale structure dominating many of the dynamic processes within them conspire to make the third class of models beyond current reach. The large spectrum of time scales relevant to the problem make it even more intractible. Such simulations do exist for small-scale magnetic features (magnetic elements), however, and it is basically insufficient computing power which keeps sunspots from being similarly described.

Further details, and in particular further references, about models of the thermal and magnetic structure of sunspots are to be found in the following reviews: Nordlund and Stein (1990), Thomas and Weiss (1992), Jahn (1997) and Weiss (1997).

Models of the magnetic structure of sunspots

Sunspots are thought to be the cross sections at the solar surface of large, nearly vertically oriented magnetic flux tubes (see SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES). Hence the large-scale magnetic structure of sunspots is generally represented by axially symmetric flux tubes. Most models of the magnetic structure are calculated in the magnetohydrostatic approximation, i.e. they neglect evolutionary aspects, convective motions, the Evershed effect and the influence of waves and oscillations. For the overall structure of the magnetic field this is a good approximation since sunspots evolve on time scales far longer than the time taken by disturbances travelling at the Alfvén speed or sound speed to cross them (the Alfvén transit time is only an hour). Dynamic phenomena are important, however, for shaping the small-scale magnetic (and thermal) structure which is particularly prominent in the penumbra (penumbral filaments or fibrils), but is also seen in umbras (umbral dots). Most models of the whole sunspot neglect the fine-scale structure in the interests of tractability.

The sunspot magnetic field is confined horizontally by a combination of the excess gas pressure in the field-free surroundings of the sunspot and magnetic curvature forces. In contrast to small magnetic flux tubes the latter cannot be neglected, making the modeling of the sunspot magnetic field far more challenging. The magnetohydrostatic equilibrium is described by the force balance equation and one of Maxwell's equations:

$$\frac{1}{4\pi}\,\mathrm{curl}\,\boldsymbol{B}\times\boldsymbol{B}=\nabla p-\rho\boldsymbol{g}$$

$$\mathrm{div}\,\boldsymbol{B}=0$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where $\boldsymbol{B}$ is the magnetic vector, p is the gas pressure, ρ denotes the density and $\boldsymbol{g}$ is gravitational acceleration. Hydrostatic equilibrium along field lines is already implicit in the force balance equation. Usually significant additional assumptions are made, since the computation of the magnetic configuration without further assumptions requires the simultaneous and consistent solution for the magnetic and thermodynamic structures, which in turn requires the solution of an energy equation in addition to the above equations. Most such 'comprehensive' solutions that have been attempted are not general since the magnetic structure is often partially prescribed and the thermodynamics are greatly simplified, although recently significant progress has been made on both accounts. Unless the energy equation is solved together with the force-balance equation either the magnetic field or the temperature must be specified and an equation of state must be postulated. Instead of the temperature distribution the pressure can also be specified.

Types of magnetic solutions that have been investigated include similarity solutions, return-flux models, force-free models, current-sheet models and models with arbitrary horizontal pressure distributions. Before considering these particular solutions, however, we need to distinguish between some basic configurations.

Monolith versus cluster models

One basic assumption underlying all attempts to quantitatively model the global magnetic structure of sunspots is the assumption that the sunspot is monolithic below the solar surface, i.e. that it can be represented by a single flux tube. Since these layers are not directly accessible to observations, this assumption cannot be rigorously tested, although the techniques of local helioseismology can in principle set some constraints on the subsurface field (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY).

It has been proposed that just below the surface the magnetic field of a sunspot breaks up into many small flux tubes due to the fluting or interchange instability. In this instability, the magnetic energy of the system is lowered by the break up of a large flux tube (with strong magnetic curvature terms) into many small ones (with small curvature terms). Later it was found, however, that magnetic buoyancy can save sunspots ($\phi > 10^{20}$ Mx) from going unstable to fluting, so that this argument does not favor the production of subsurface 'spaghetti'. This model is often referred to as the spaghetti model or jellyfish model. Nevertheless, a cluster model of sunspots (see figure 1) has the advantage that it can readily explain the high thermal flux seen in the umbra, as well as umbral dots in a natural manner (as field-free intrusions into the sunspot). The complex magnetic structure in the penumbra, in which fibrils or filaments of field pointing in different directions are interlaced, suggests that the magnetic field is indeed concentrated into many small flux tubes.

One argument against the cluster model is that the potential field in the observable layers produced by a collection of buried magnetic monopoles held together at great depth has a maximum strength that increases too strongly with the number of monopoles. It increases much more rapidly with the size of the spot than shown by the observations. To what extent a flux-tube cluster is described by a group of magnetic monopoles is unclear, however. Another reason why the cluster is not expected to be too loosely constructed is that the observed radial dependence of the field is not too different from a buried dipole (figure 2), which is prototypical of a monolithic flux tube

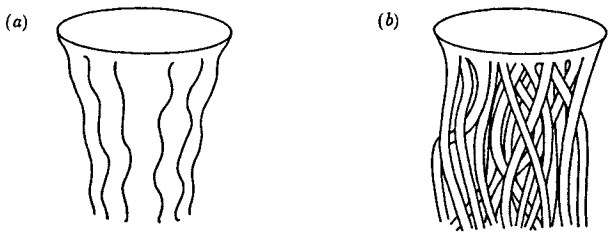


Figure 1. Sketch of the monolithic (left) and cluster (right) models of sunspots. (From Thomas and Weiss 1992 by permission.)

Self-similar models

In such models the radial dependence of $\boldsymbol{B}$ at a given constant geometrical height is prescribed and is imposed to be the same at all heights, except for a scaling to take into account the expansion of the field with height. Hence the name similarity models or self-similar models. This prescription allows the vertical and radial magnetic coordinates to be separated

$$B_z = f(r/R(z))B_0(z)$$

$$B_r = -\frac{r}{2}\, f(r/R(z))\frac{\mathrm{d}B_0}{\mathrm{d}z}(z)$$

and only an ordinary differential equation for $B_0(z)$, the field strength along the flux-tube axis, needs be solved. $f(r/R)$ is called the shape function. In most applications a Gaussian shape function has been chosen.

Due to their simplicity such models have attracted considerable attention. Some self-similar models include a description of energy balance, for others the pressure difference between the axis and the field-free surroundings of the sunspot is prescribed at all heights (e.g. by taking it from empirical models in the observable layers).

Self-similar models, although straightforward to construct, may be too restrictive to describe sunspots accurately. For example, classical similarity models have great difficulty reproducing the observed continuum structure of sunspots composed of a central umbra surrounded by a penumbra. These models usually produce a bright ring, usually located near the umbra–penumbra boundary, contrary to observations. This anomaly, which may be an artifact of the particular choice of the shape function, has so far only been removed by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

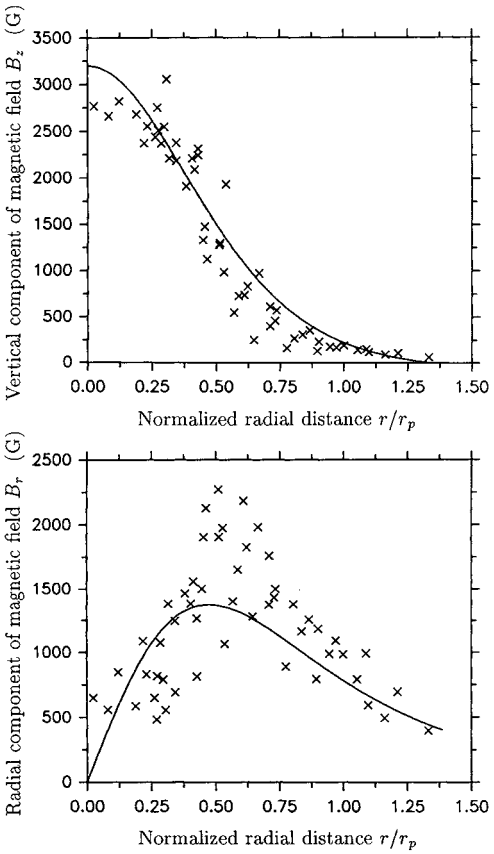


Figure 2. Fit of a buried dipole (solid curves) to the measured $B_z(r/r_p)$ in (*a*) and to $B_r(r/r_p)$ in (*b*). The observations are represented by crosses. Here B_z is the vertical and B_r is the radial component of the magnetic vector, r is the radial coordinate measured from the geometrical center of the sunspot and r_p is the outer radius of the penumbra.

ad hoc changes in the thermodynamic structure. A more basic limitation of self-similar models applies to their upper layers. The magnetic field of sunspots, like that of small flux tubes, expands nearly exponentially with height. If, for example, the field is nearly potential or force-free inside a boundary current sheet (which is the case above the SOLAR PHOTOSPHERE), then the rapidity of this expansion influences the radial distribution of the field. The more strongly the field expands the more strongly the field strength decreases towards the edge of the flux tube. Since the rate of expansion itself increases significantly from the SOLAR INTERIOR to its outer atmosphere the correct description of the magnetic field near the solar surface does not automatically imply that the field is also correctly described below the surface or high in the atmosphere (above the lower coronal boundary, where radio observations can be used to measure it).

More realistic self-similar solutions may be obtained by including the Wilson depression (see SUNSPOTS: WILSON EFFECT) and its observed horizontal variation when determining the shape function $f(r/R)$ from the observed radial dependence of the magnetic field (which is known along a level of constant optical depth). These models reproduce a fair number of observations.

Return-flux models

Self-similar models have been extended to include field lines in the outer part of the sunspot which return to the solar surface just outside the visible sunspot (return-flux model). In this model the emerging return flux is said to constitute the penumbra. Since observations suggest that approximately half of the magnetic flux of a sunspot emerges in the penumbra this also provides an estimate of the amount of return flux that is present according to this model. This model, however, finds little support from observations, since the amount of opposite polarity flux seen in the immediate surroundings of sunspots is small. The observed return flux appears to be related to the horizontal field strands carrying the Evershed flow and is only seen in the lowest observable layers at some locations at the periphery of the sunspot. A dominant fraction of the flux continues in the magnetic canopy which surrounds the sunspot (a layer of almost horizontal field overlying a field-free layer). For realistic parameters the predicted continuum intensity of return-flux models is not much closer to the observations than for the standard self-similar models.

Force-free and potential field models

The magnetic field is called force free when the terms describing the influence of the gas (and the gravity), i.e. the terms on the right-hand side of the force balance equation, can be neglected (force-free fields) (see FORCE-FREE MAGNETIC FIELDS). In that case the force balance reduces to

$$\text{curl}\, \boldsymbol{B} = \alpha \boldsymbol{B}.$$

In general, α is a scalar function of the spatial coordinates and is constant along field lines (as can be seen by taking the divergence of the above equation). Solutions are particularly simple if α is constant everywhere. Then $\nabla \cdot \boldsymbol{B} = 0$ is automatically fulfilled. A final simplification, a potential field, is obtained when α vanishes. Force-free and potential fields are uncoupled from the rest of the atmospheric parameters. Hence it is sufficient to know the field at the boundary of the computational domain in order to compute it in the whole volume. Conversely, this decoupling implies that force-free models can, by definition, not reproduce the thermodynamic properties of sunspots. They are unable to describe the relatively sharp boundary between the sunspot and the surrounding, almost field-free gas.

The magnetic field inside sunspots certainly comes close to being force-free in chromospheric and coronal layers due to the small gas pressure and density there. If small-scale structure is neglected then even a potential field provides a good description in these layers. In the photosphere this assumption is less secure since the energy density of the gas is actually larger than the energy density of the field in large parts of the sunspot.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Nevertheless, various models of the sunspot magnetic field based on the force-free or potential assumption have been constructed. Constant-α solutions approximately reproduce early observations of the radial dependence of the magnetic field strength, indicating that the constant-α approximation is adequate for the visible layers of simple sunspots.

Current-sheet models

The simplest consistent current-sheet model is composed of a flux tube in whose interior the magnetic field is potential, i.e. current-free, so that all the current is concentrated in a sheet at the boundary of the flux tube. This sharp boundary is termed the magnetopause. The main problem is the determination of the position of the boundary between the flux tube and the surroundings in the presence of arbitrary stratifications in the flux tube and its surroundings. An exact solution to this problem has, however, been found.

There are indications that a potential field bounded by a current sheet at the magnetopause is on its own inadequate to describe sunspots, in particular the presence of the penumbra. Far better agreement with the observations is obtained if this model is extended to include body currents, so that the field inside the sunspot is not purely potential, but is nevertheless surrounded by a current sheet. The body currents can be restricted to the radially outer part of the sunspot (corresponding approximately to the penumbra) and for the surface field to match the observations. It turns out that the field does not deviate greatly from potentiality anywhere (except at the boundary, of course) and is not self-similar.

Uncombed fields

On a small azimuthal scale horizontal field lines are observed to alternate with inclined ones in the penumbra. Various ideas have been proposed to explain the origin of these so-called fluted magnetic fields or uncombed magnetic fields. Basically these ideas consider the small-scale magnetic structure to be dynamic and its complexity to result from an instability.

One proposal for the source of the complex magnetic fine structure of the penumbra is due to the convective exchange of flux tubes. One possible scenario for interchange convection is the following: a flux tube near the magnetopause below the penumbra is heated by the field-free convective gas with which it comes into contact. The heated tube is buoyant and rises. At the surface it radiates away its excess energy, loses its buoyancy, becomes more horizontal and sinks again.

The first part of this scenario has been confirmed and quantified by 2D numerical simulations. An illustration is given in figure 3. A thin flux tube lying at the magnetopause (i.e. the outer boundary of the penumbra) heats up, becomes buoyant and begins to rise (the background penumbral field and gas, are stratified superadiabatically). The part of the flux tube near the outer boundary of the penumbra reaches the surface first. Below the surface the tube rises almost adiabatically, but above the surface radiative losses make it denser and reduce the buoyancy. Also, the background stratification is no longer superadiabatic. Consequently, the parts of the flux tube above the surface come to rest, staying horizontal, while the surrounding field remains strongly inclined. With time, parts of the flux tube ever closer to the umbra appear and lengthen the horizontal portion of the flux tube. The horizontal flux tube remains in equilibrium, since, for example, the negative buoyancy is balanced by the upward accelaration due to the expanding background field.

The part of the flux tube passing through the solar surface is filled with hot gas and appears bright, the other parts of the flux tube lying above the solar surface have had time to cool down and appear dark. Hence an observational signature of the formation of a horizontal flux tube is the movement of a bright point towards the umbra. Such moving bright points (called penumbral grains) have indeed been observed.

The other proposal starts with an inclined field and produces a horizontal field component by invoking a density inversion with height. To produce the density inversion a spatially localized impulsive upflow in the inclined magnetic component is postulated. In the almost vertical field of the umbra such an upflow is expected to be visible as an umbral dot. Note, however, that umbral dots do not show marked upflows, although it cannot be ruled out that such an upflow may be present, yet remain invisible in most spectral lines. In the penumbra, at a larger inclination of the field to the vertical, the flow, which fills a small flux tube up to a certain height, now causes dense material to overlie less dense material. A Rayleigh–Taylor instability develops which causes the filled flux tube to fall and become horizontal. Furthermore, the surplus material, which flowed up into the inclined flux tube, is now expected to flow outwards along the horizontal tube. A detailed numerical treatment of the collapse of a flux tube would be of great value to test this scenario, since the influence of curvature forces and the force acting on the flux tube due to the vertical gradient of the 'background field' of the penumbra has not been considered in detail as yet. These may well quench any instability from developing beyond the initial linear stage.

Brightness and thermal structure

Sunspot darkness

Biermann first proposed that the reduced brightness of sunspots is due to the inhibition of convective motions by the magnetic field. Since convection dominates energy transport below the observable layers, a quenching of convection turns a sunspot into an obstacle for the heat flux flowing outward through the convection zone. This leads to a diversion of energy away from the sunspot, which reduces the energy flux through the spot and produces a darkening. The diverted energy flux is mainly stored in the convection zone and released only over a time scale of 10^5 years, the time scale of the thermal adjustment

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

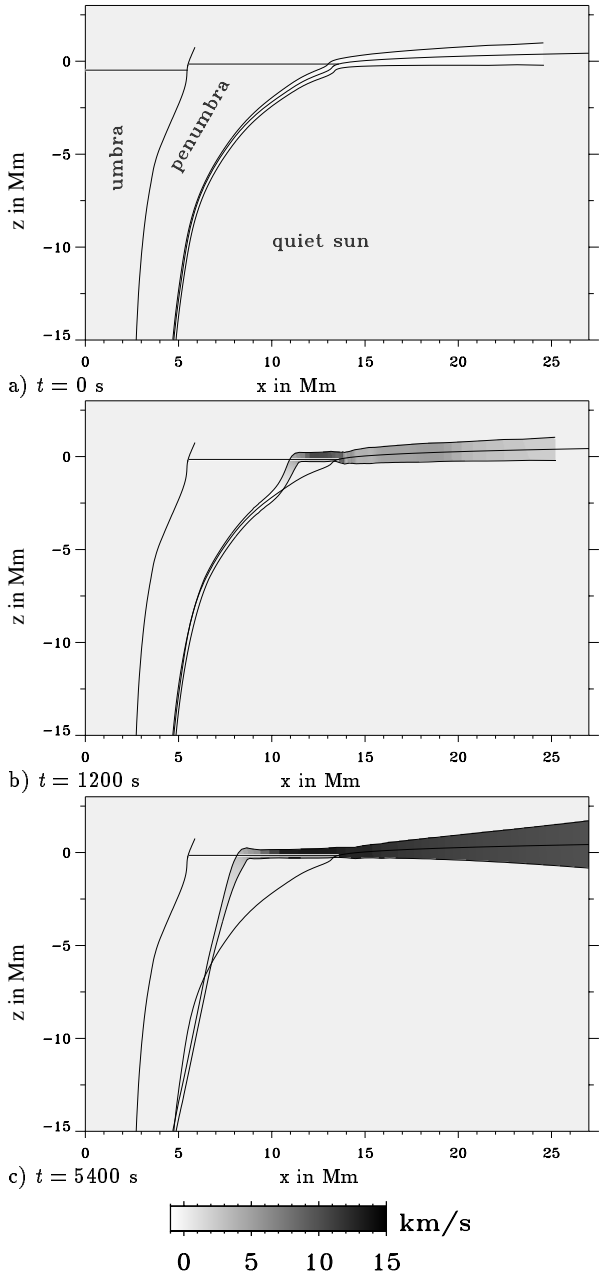


Figure 3. Vertical cut through a model penumbra. Vertical axis: height z, horizontal axis: radial distance x from the center of the sunspot (lying to the left). Indicated are the solar surface (horizontal lines near $z = 0$), the magnetopause between penumbra and quiet Sun and the current sheet between umbra and penumbra. Shown is the buoyancy-driven evolution of a flux tube starting at its initial location at the magnetopause until it lies partially horizontally at the surface. (From Schlichenmaier *et al* (1998) by permission.)

of the convection zone, which is much longer than the lifetime of a sunspot. Consequently, sunspots lead to a semipermanent darkening of the Sun. Due to the high lateral thermal conductivity of the gas in the convection zone, the diverted heat flux is distributed throughout the convection zone, so that no significant bright ring is expected around the sunspot. The convection zone also possesses a large heat capacity, which means that its temperature is raised only by an imperceptible amount. In the outer part of the solar interior convection is so much more efficient in transporting energy than radiation that a complete quenching of convection would reduce the heat flux so effectively that the question to ask in this case would not be why umbras and penumbras are so dark, but rather why they are so bright, in particular the penumbras. Therefore the inhibition of convection below sunspots is probably only partial.

Another proposal for explaining the darkness of sunspots is the deviation of the heat flux vector from the vertical due to the influence of the sunspot's magnetic field on the convection. Hoyle pointed out that it is mainly the motions across the field lines which are suppressed, so that the heat flux is channelled by the field. Its expansion with height leads to a dilution of the thermal flux. This effect on its own does not explain the observed umbral flux. It is nevertheless included in recent models.

Finally, Parker proposed that the heat flux is not diverted away from sunspots at all, but rather that umbras are actively cooled by the conversion of thermal energy into Alfvén waves. This mechanism now seems unlikely, since the expected flux of Alfvén waves is not observed.

Umbral brightness

Basically two approaches have been taken to solve the problem of too cool umbras. The first has been to form the sunspot out of a cluster of small flux tubes (spaghetti model) as initially proposed by Parker. Due to the tapered shape of each small flux tube, increasingly more field-free gas is present between them at increasing depth. Consequently, below the sunspot convection can penetrate relatively unhindered until close to the surface. In addition, for a sunspot composed of n small flux tubes the surface area of the side walls of the flux tubes, over which the convective gas can radiate into the magnetized gas, is $\sqrt{n}$ times larger than for a simple monolithic sunspot. A larger side-wall surface compared to the horizontal cross-sectional area leads to a much more efficient heating of the tubes (isolated small-scale tubes can have a bright bottom relative to the quiet Sun). This theory, like many other aspects of the cluster model, has not been quantitatively worked out in detail. If field-free material does penetrate to above the solar surface along unresolved hot columns then this could provide the missing radiative flux. Even if the individual small flux tubes carry no heat flux, then a 20% area coverage by the field-free columns would suffice to produce the observed flux. The columns would have to be very thin and numerous, however, since all parts of an umbra exhibit a minimum brightness. One advantage of this model is that it provides an explanation of bright umbral dots in the dark umbral background. They may

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

be considered to be intrusions of field-free material from below the solar surface into the photosphere.

The second proposal incorporates convective transport within a simple model of a monolithic sunspot (magnetoconvection). Such simulations are generally resticted to an initially homogeneous magnetic field in an idealized geometry and simplified equations (e.g. neglecting radiation) and often relate to plasma parameters far removed from those present in real sunspots. From a fully compressible calculation in an externally imposed vertical magnetic field it follows that the character of the convection below a sunspot umbra depends strongly on the parameter ζ, the ratio of magnetic diffusivity to thermal diffusivity. If $\zeta < 1$ then the convection is oscillatory, whereas if $\zeta > 1$ then the convection is overturning. At the surface $\zeta \approx 10^{-3} \ll 1$, but increases rapidly with depth, eventually becoming larger than unity. Therefore, the current picture is that oscillatory convection dominates in the 2000 km immediately below the surface, while the far more efficient overturning convection takes over in the deeper layers (depths of 2000–20000 km).

The two forms of convection are coupled through nonlinear interactions, so that hot convective plumes form which might penetrate right up to the surface. This coupled form of magnetoconvection is more efficient than purely oscillatory convection and conceivably could transport enough energy to explain the brightness of umbras even in a monolithic sunspot.

Another important parameter determining the nature of the magnetoconvection in these simulations is the Chandrasekhar number Q, which is proportional to the square of the imposed field. At sufficiently large Q convection is completely suppressed. As Q is reduced first weak, steady convection is excited, with isolated columns of rising gas and a continuous network of sinking fluid. When Q is lowered even further an oscillatory bifurcation sets in, with alternate plumes waxing and waning in antiphase. At still smaller Q the convection becomes more vigorous and chaotic, with the convective pattern changing rapidly. In all cases the convection introduces an inhomogeneity into the magnetic field. In the upper layers the field becomes concentrated in the downflowing network, while at greater depth there is a strong concentration below the upflowing plumes.

In most 'comprehensive' models of sunspots, i.e. models in which the force balance and energy equations are consistently solved for in the whole sunspot, the convective transport is treated by the mixing length formalism and the radiation by the diffusion approximation. The efficiency of the convective transport is varied by adapting the mixing length parameter $\lambda = l/H$, where l is the mixing length and H is the pressure scale height (e.g. Chitre 1963, Deinzer 1965, Chitre and Shaviv 1967, Yun 1970, Jahn 1989). Often even this approach is further simplified. For example, when combining the similarity assumption of the magnetic field with an energy equation, the thermal structure needs to be determined only along the axis of the symmetric sunspot and outside the sunspot. The horizontal variation of the temperature is fixed by the prescribed magnetic shape function. Hence the 2D problem of energy transport in an axially symmetric flux tube is reduced to a 1D problem.

One successful approach has been to solve both the force balance and energy equations in an iterative scheme. The latter includes radiative energy transport (in the diffusion approximation) and convective transport (in the mixing length approximation). The umbra and penumbra are treated separately, in that each has its own mixing length parameter λ, which is determined from the temperature gradient and is thus depth dependent. In addition, the volume below the umbra is assumed to be thermally isolated in the horizontal direction, so that the heat flux below the umbra is independent of depth. It turns out that in order to ensure a constant heat flux λ must increase rapidly with depth, being approximately 0.15 near the surface and reaching 0.7–0.8 at greater depths. For comparison $\lambda = 1$ describes the quiet Sun. A larger λ implies more efficient convection, so that the above result says that convection plays an increasingly important role relative to radiation with increasing depth below the umbra. This result is also consistent with the prediction of ζ going from $\ll 1$ to >1, since overturning convection is more effective than oscillatory convection. The small mixing length near the surface also agrees with the fact that the empirically derived temperature stratification at and just below the surface is close to radiative equilibrium.

Penumbral brightness

The above discussion has focused mainly on the umbra. The situation in the penumbra is rather different. The thermal flux through the penumbra is almost four times larger than in the umbra and three-quarters of that in the quiet Sun. Also, the field is far less vertical and less strong in the penumbra. It exhibits a prominent, radially elongated filamentary structure in white light, so that there is a rapid azimuthal variation of brightness. On average, however, there is a relative absence of a radial or azimuthal temperature gradient.

If the penumbra is thin, i.e. if the field lines lie roughly parallel to the solar surface and the magnetopause is located just below the surface everywhere in the penumbra, it is relatively easy to get a sufficient *energy* flux through the penumbra. Observations, however, suggest that over 50% of the *magnetic* flux of the sunspot emerges in the penumbra and field lines are considerably inclined to the horizontal there (by between 15 and 50°), signifying that the penumbra is thick (i.e. the magnetopause lies many scale-heights below the solar surface under most of the penumbra).

The heating of a thick penumbra requires some form of magnetoconvective energy transport. As in the case of the umbra, convection under the penumbra is often described by the mixing length formalism. Again, the mixing length must be reduced compared to the quiet Sun, but is larger than in the umbra. For example, models suggest that $\lambda \approx 0.7$ below the penumbra.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

It is thought that most of the radiative flux emerging through the penumbra enters into the sunspot through the magnetopause. This feature is driven by the idea that most of the heat flux is transported from the magnetopause to the surface layers of the penumbra by interchange convection. Interchange convection is an inherently 3D process in which whole flux tubes rise and fall in the convectively unstable stratification below the penumbra. Below the solar surface the flux tube undergoes little heat exchange with its surroundings, so that it rises basically adiabatically. Hence, the gas trapped between the field lines of a rising flux tube has the heat content of the gas at the magnetopause. It has been suggesting that significant energy flows across the magnetopause only when it is inclined by more than a critical angle to the vertical. Depending on whether this condition is achieved or not a sunspot with penumbra (if the energy flows in from the sides) or a pore without a penumbra is formed.

Two-dimensional simulations confirm part of the interchange convection scenario. A flux tube lying along the magnetopause does indeed rise to the surface once it is heated sufficiently. It also brings hot gas to the surface, which then radiates away a part of its thermal energy. In contrast to the interchange convection scenario, however, the simulated flux tube does not sink back down again once it has cooled. This may be due to the simplifying assumptions underlying the simulations. For example, a more realistic treatment of the radiative transfer may lead to a more efficient cooling and allow the flux tube to sink below the surface again. On the other hand, the simulations do show that even after a flux tube has emerged and has been lying horizontally above the surface for some time, hot material keeps flowing up through its footpoint in the inner part of the penumbra. This upflow replenishes material that is flowing horizontally outwards along the tube. This horizontal outflow, which exhibits many of the properties of the Evershed flow, is driven by the excess of gas pressure at the inner footpoint, which is partly built up during the rise of the flux tube and is partly due to the higher temperature of the gas at the inner footpoint. The material emerging from the solar interior then flows towards the outer boundary of the sunspot, radiating away its excess heat on the way. This model reproduces the presence of bright penumbral grains and partly also bright and dark penumbral filaments, although according to it there is more bright material in the inner penumbra (i.e. near the umbra), which is not supported by observations. But the exact visibility of bright and dark filaments will need to be treated using proper radiative transfer before they can be compared quantitatively with observations.

Idealized simulations of convective modes in the presence of an imposed homogeneous field that is inclined to the vertical lead to waves travelling in the direction of the magnetic inclination. Both the strength and inclination of the field is affected by the waves.

Bibliography

Jahn K 1997 Physical models of sunspots *First Advances in Solar Physics Euroconf.: Advances in the Physics of Sunspots (ASP Conf. Ser. 118)* ed B Schmieder, J C Del Toro Iniesta and M Vázquez pp 122–39

Nordlund Å and Stein R F 1990 Solar magnetoconvection *Solar Photosphere: Structure, Convection and Magnetic Fields (IAU Symp. 138)* ed J O Stenflo (Dordrecht: Kluwer) pp 191–211

Schlichenmaier R, Jahn K and Schmidt H U 1998 A dynamical model for the penumbral fine structure and the Evershed effect in sunspots *Astrophys. J.* **493** L121

Thomas J H and Weiss N 1992 The theory of sunspots *Sunspots: Theory and Observations* ed J H Thomas and N Weiss (Dordrecht: Kluwer) pp 3–59

Weiss N 1997 Magnetoconvection *First Advances in Solar Physics Euroconf.: Advances in the Physics of Sunspots (ASP Conf. Ser. 118)* ed B Schmieder, J C Del Toro Iniesta and M Vázquez pp 21–33

Sami K Solanki

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Oscillations and Seismology

Sunspots are known to host a wide variety of waves and oscillatory phenomena. The coupling between these oscillations and the 5 min acoustic oscillations of the surrounding quiet Sun provides a useful diagnostic probe of the magnetic and thermodynamic structure of the spot.

Oscillations in SUNSPOTS were first recorded during the period 1969–1972. The chief difficulty encountered in their detection is the suppression of the stray-light contribution to the oscillation signal from the surrounding quiet Sun. This contamination can be minimized by measuring Doppler velocities using molecular lines and low-excitation lines of neutral species with low ionization potentials, which form exclusively in the cool SUNSPOT UMBRA, or by measuring Doppler velocities using the polarized components of a Zeeman-sensitive line, which form only in a magnetized plasma (see POLARIMETRY).

Since sunspots are continuously jostled by the turbulent convection (GRANULATION) which surrounds them, it is probably no surprise that they host a diverse array of dynamic and oscillatory motions. It is an open question as to what fraction of the observed oscillations are the passive response of the sunspot to this external forcing, and what fraction are generated within the sunspot through magnetoconvection (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION. It has also been recognized that the nature of these oscillations, coupled with the manner in which they are forced by the surrounding granulation and the solar acoustic oscillations (HELIOSEISMOLOGY), has the potential to reveal diagnostic information about the magnetic and thermodynamic structure of the subsurface sunspot interior (SUNSPOT MODELS). Thus, in more recent times, sunspot oscillations tend to be viewed in the broader context of *local helioseismology*, which emphasizes the connection between oscillatory wave fields inside and outside of sunspots.

Observations

For historical reasons, sunspot oscillations are broadly classified as falling into one of three categories:

- *5 min umbral oscillations* are large spatial scale (typically a substantial fraction of the entire umbra) coherent oscillations with periods between 250–400 s and root-mean-square (RMS) velocity amplitudes of 40–90 m s^{-1}, observed primarily in lines which form in the low SOLAR PHOTOSPHERE. The RMS amplitudes of these oscillations are roughly a factor of three smaller than typical pixel-by-pixel 5 min oscillation amplitudes observed in the surrounding quiet Sun. There is also a tendency for the spatial oscillation pattern to exhibit rapid radial propagation from the umbral center out through the SUNSPOT PENUMBRA with phase speeds of 50–100 km s^{-1}.
- *3 min umbral oscillations* are smaller spatial scale (3–4″) oscillations observed in lines which form in the upper photosphere and CHROMOSPHERE, with periods of order 140–200 s, and RMS velocity amplitudes of hundreds of meters per second in the mid-to-upper photosphere to several kilometers per second in the chromosphere. 'Saw-tooth' time series of Doppler velocities measured in chromospheric lines with peak amplitudes of several kilometers per second indicate that these oscillations are fully nonlinear throughout much of the chromosphere. Figure 1 illustrates the distinct observational character of the 5 min and 3 min umbral oscillations.
- *Running penumbral waves* are coherent outward-propagating wave arcs (subtending angles of 90–270°) readily observed in the penumbral chromosphere. The measured radial phase speeds of these waves vary between 8 and 35 km s^{-1}, with a tendency for the phase speed to decrease with increasing radial distance from the sunspot center. Typical Doppler motions associated with these waves are several hundred meters per second and the arc patterns repeat with a 200–300 s cadence. There is growing observational evidence that patchy downward extensions of these waves to the penumbral photosphere are present, but with radial phase speeds significantly greater than those measured in the penumbral chromosphere.

In all likelihood the distinctions between these three basic oscillatory motions are dictated more by observational issues than by clear physical effects. For example, when simultaneous umbral Dopplergrams are made in a number of spectral lines which cover a wide range of formation heights in the solar atmosphere, the detected oscillations exhibit both vertical and horizontal coherence *at the same frequency*, but are *incoherent* at distinct oscillation frequencies. Based on the coherence of the oscillations at different altitudes, phase lags can be determined. For frequencies below 5.2–5.5 mHz (the effective *acoustic cut-off frequency* for the umbral atmosphere), the oscillations are in phase at all overlying atmospheric levels, while for oscillations in excess of the acoustic cut-off frequency, a phase lag becomes more pronounced with increasing height separation. This behavior is generally interpreted as evidence for the upward propagation of waves for frequencies larger than 5.5 mHz. Indeed the 'saw-tooth' wave trains observed in the Ca II K-line may well be interpreted as the nonlinear steepening of these upward propagating waves due to the rapid decrease in density between the low photosphere and the mid-chromosphere.

Theory

Theoretical calculations of sunspot oscillations began in earnest in the mid-1970s, but have unfortunately not progressed much beyond the basic tenets of the initial models. The principal reason for this stagnation can be traced to the difficulties encountered in solving the relevant partial differential equations which describe the linear oscillations of a buried-dipole-type magnetic spot geometry (see figure 2 in SUNSPOT MODELS). If the magnetic field is *uniform* and aligned parallel (appropriate for the umbra), or perpendicular (appropriate for the penumbra),

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

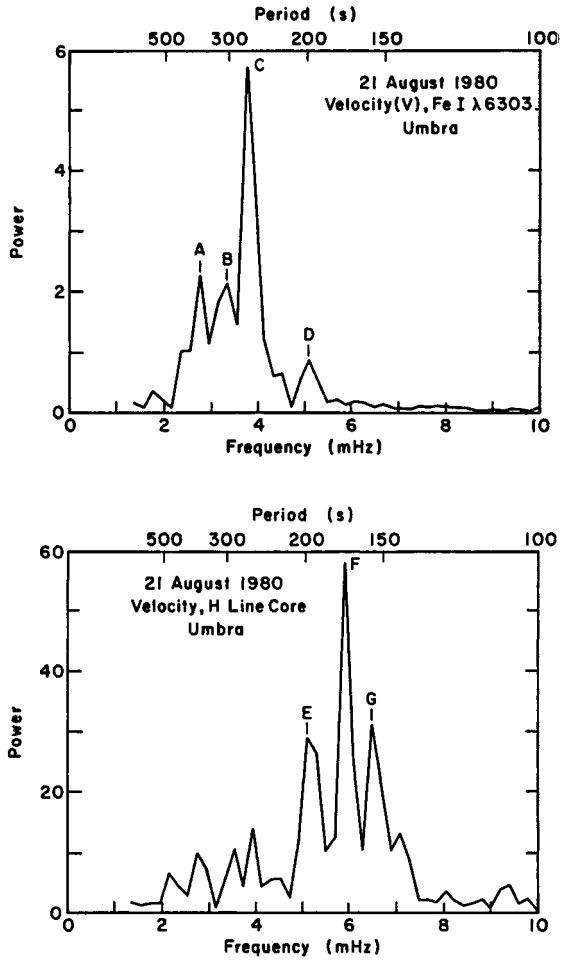


Figure 1. Temporal power spectra of the photospheric Doppler velocity (Fe I 630.3 nm: upper panel) and chromospheric Doppler velocity (Ca II 396.8 nm: lower panel), averaged over a single sunspot umbra. In the upper panel the photospheric oscillatory power is concentrated around periods of 5 min, while in the lower panel the power is greater by a factor of 10 and is shifted to shorter periods of 3 min at chromospheric levels. The precise locations of distinct power peaks, labeled A, B, C, . . . , are most likely an artifact of the spatial averaging of the point-by-point oscillation power spectra. Their numbers and locations are time variable within the same data set, and they are not reproducible between different observations. The peaks labeled D and E correspond to the same $P = 197$ s period, and their coincidence suggests a physical connection between the photospheric and chromospheric oscillations. (Adapted from Thomas *et al* 1984 *Astrophys. J.* **285** 368–80.)

to the prevailing atmospheric stratification, then these partial differential equations factor into a set of ordinary differential equations connected by several separation parameters—typically the horizontal wavevector $\boldsymbol{k}$ and the oscillation frequency ω. Hence, the somewhat artificial distinction between umbral and penumbral oscillations noted above is further maintained on the theoretical side of the problem. It is possible to extend the separation of variables technique to encompass uniform but arbitrarily inclined magnetic fields; however the results are not vastly different from those encountered for the purely vertical-field case.

When the uniform magnetic field has a component parallel to the stratification then the two characteristic wave propagation speeds—the adiabatic sound speed and the Alfvén speed—both vary with position along a given magnetic line of force. Consequently two of the three basic uniform MHD wave modes, the fast and slow magnetosonic waves, are coupled, while the Alfvén wave remains decoupled and propagates independently of them (see MAGNETOHYDRODYNAMIC WAVES). The mode-mixing between the slow and fast magnetosonic waves is minimal in the WKBJ limit (wavelength $\ll$ atmospheric scale height) or if the local sound and Alfvén speeds are of very different orders of magnitude, but neither of these regimes apply to the umbral photosphere. A large-scale global umbral oscillation mode may have *both* the character of the slow and fast magnetosonic waves for different local height ranges, as illustrated by figure 2. This mode mixing has important implications for seismic probing of spots.

A horizontal magnetic field, on the other hand, has a unique value of the sound and Alfvén speeds along each line of force. The variation of these characteristic propagation speeds with height leads to a different type of mode-mixing process known as *resonance absorption*. Unlike the more spatially diffuse mode-conversion process that applies in vertical magnetic fields, in a horizontal magnetic field the mode conversion is localized in the vicinity of a particular atmospheric layer, called a critical layer, where the local Alfvén or tube speeds match the projection of the phase speed of the atmospheric oscillation along the magnetic line(s) of force. Through this localized mode conversion, steep gradients and large amplitudes build within the critical layer until nonlinearities and/or dissipation cause the process to saturate or lead to secular evolution of the background atmosphere. Theoretical calculations employing strictly horizontal, or slightly inclined, magnetic fields have met with success in reproducing some features of the running penumbral waves. Clearly a better representation of all three types of sunspot oscillations awaits more sophisticated models which are capable of accommodating a smooth radial variation of the magnetic field inclination from the umbral core out to the periphery of the penumbra.

Seismology

It is possible to regard the interaction of ambient solar acoustic oscillations with a sunspot as a 'scattering experiment' of the sort familiar to physicists who employ high-energy particle beams to probe the structure of atoms, nuclei and elementary particles. In the sunspot 'scattering experiment' the incident acoustic 'beams' are produced in a haphazard manner by the surrounding quiet Sun, and as noted above, it is possible to detect oscillations within the 'target' sunspot proper. Three different 'experimental'

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

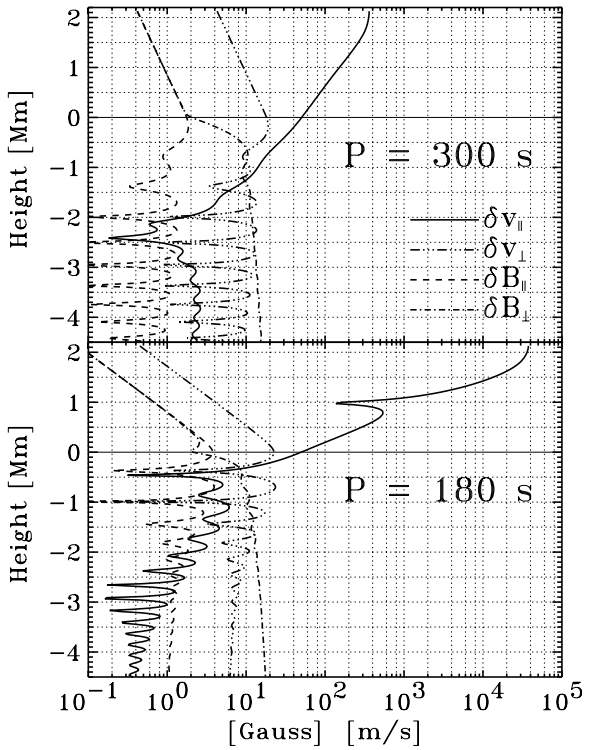


Figure 2. Typical examples of umbral oscillations computed for a uniform vertical 3000 G magnetic field imposed on a realistic sunspot model atmosphere. The four curves indicate the height variation of the RMS velocity and magnetic field fluctuations aligned with (∥), and perpendicular to (⊥), the magnetic field. The amplitudes of these two oscillatory modes are arbitrarily adjusted to yield an RMS Doppler velocity of 50 m s^{-1} at the base of the photosphere (height = 0 Mm). Notice the relative excess of Doppler velocities present in the chromosphere (+0.5 < height < +1.5 Mm) for the $P = 3$ min oscillation period relative to its $P = 5$ min oscillation period counterpart. This behavior is basically consistent with the observational data presented in figure 1. These vertical-field models suggest that a Doppler oscillation amplitude of 50 m s^{-1} translates into RMS magnetic fluctuations at the level of a few gauss. This is reasonably consistent with the current observational *upper limit* of 5–20 G for magnetic field intensity fluctuations in sunspot umbrae. (Adapted from Lites *et al* 1998 *Astrophys. J.* **497** 464–582.)

techniques for probing sunspot structure have evolved in parallel:

- *Frequency–wavenumber helioseismology* is closest in spirit to the particle physics approach since it relies exclusively on the oscillations observed in the quiet Sun surrounding the sunspot. This wave field is separated into waves with radially inward and radially outward senses of phase propagation, and each group is then further divided into its constituent frequencies (ω), horizontal wavenumbers (k), and azimuthal degrees (m). For identical values of (ω, k, m) one finds more power in the inward-propagating wave relative to its outward-propagating counterpart (by a factor of 1–3), and a positive phase shift of the outward- relative to the inward-propagating wave (0–150°). Broadly speaking these two findings suggest that sunspots absorb some of the acoustic oscillations, while the remainder that pass through the sunspot intact do so at somewhat greater propagation speeds than would apply in the surrounding quiet Sun.
- *Time–distance helioseismology* employs temporal cross-correlations of oscillation time series measured in the sunspot with those obtained in the surrounding quiet Sun at various radial distances (Δ). Inward- and outward-propagating waves are now distinguished by enhanced cross-correlations that appear at negative ($-\tau^-$) and positive ($+\tau^+$) time offsets, respectively, of the quiet Sun oscillation time series relative to the sunspot time series. For spots, $\tau^+ < \tau^-$, and the net 'travel time', $\tau^+ + \tau^-$, is less than the analogous value which obtains when the sunspot is absent (and for which the analogous positive and negative time offsets are *identical*). The smaller net 'travel time' for the spot relative to the quiet Sun is basically consistent with greater propagation speeds for waves inside the sunspot. The $\tau^+ \neq \tau^-$ directional asymmetry in the cross-correlation coherence maxima is more difficult to interpret. It could arise from coherent downflows beneath the sunspot, from time asymmetries due to dissipative process (absorption), or it could simply be a consequence of observational (spatial and temporal) peculiarities associated with the detection of the sunspot oscillations.
- *Acoustic holography* carries the concepts of time–distance helioseismology one step further through the 'acoustical imaging' of the subsurface environment of a spot. For each subsurface 'focal point', an *egression* of the wave field recorded at the solar photosphere is carried out. The egression effectively discards that portion of the wave field which is not consistent with the signature of an acoustic point source located at the specified focal point. Maps of the RMS egression computed for a subsurface volume over a specified period of time show that a relative egression deficit ($\leq$50%) is present below sunspots. The horizontal extent of the deficit region is about twice the visual extent of the spot.

Beyond general indications of increased propagation speeds and acoustic absorption implied by the findings noted above, the extant body of theoretical work interpreting these robust observational results must be viewed as highly speculative and subject to future revision. Any program to simultaneously model oscillations within, and outside of, a sunspot must be based upon a non-trivial sunspot equilibrium model with *non*uniform magnetic fields. This entails the difficult task of solving the partial differential equations which govern the sunspot oscillations *and* connecting these oscillations with those in the neighboring quiet Sun. Various attempts aimed at circumventing this complicated problem through the reliance on ray theory, or its variants, which all derive

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

from the WKBJ formalism, run afoul of the acoustic analog of Heisenberg's uncertainty principle. Recall that the spatial and temporal extent of a wave packet is inversely proportional to the range of wavenumbers and frequencies of the constituent waves from which the packet is composed. The distribution of observed oscillatory power across the k–ω diagram (see figure 4 in HELIOSEISMIC OBSERVATIONS) necessarily imposes fundamental limitations upon how sharp a ray bundle can be, and for the most part, sunspots are probably too small to safely treat their interaction with acoustic oscillations in the 'geometrical-acoustics' limit.

With these caveats in mind, preliminary work does suggest that the mode conversion of the forcing acoustic oscillation to magnetic-field-guided slow magnetosonic waves, and resonance absorption in critical layers, may play crucial roles in accounting for the deficit of outward- to inward-going waves in the sunspot environment.

Bibliography

Bogdan T J and Braun D C 1995 Active region seismology *Proc. Fourth SOHO Workshop: Helioseismology* ed J T Hoeksema, V Domingo, B Fleck and B Battrick (Noordwijk: ESA) pp 31–45

Chitre S M 1992 Theory of umbral oscillations and penumbral waves *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) pp 333–43

Lindsey C and Braun D C 1997 Helioseismic holography *Astrophys. J.* **485** 895–903

Lites B W 1992 Sunspot oscillations: observations and implications *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) pp 261–302

Marco E, Aballe Villero M A, Vázquez M and García de la Rosa J I 1996 Spatial changes of the oscillatory pattern in sunspot penumbrae at the photospheric level *Astron. Astrophys.* **309** 284–90

Thomas J Bogdan

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Penumbra: Structure and Activity

As viewed in white light, a sunspot consists of the dark, central umbra surrounded by the brighter, radially striated penumbra. The structure of the penumbra is generally more complicated that that of the umbra, and our understanding of penumbral structure has developed slowly as observations have provided us with increasing spatial resolution.

The relative darkness of sunspots is attributed to the inhibiting effect of the spot's magnetic field on the convective transport of heat. The inhibition of convection is quite strong in the umbra, which transmits only about 20% of the average solar heat flux. In the penumbra, however, the convection is far more efficient, with the penumbra transmitting some 75% of the average heat flux with a radial variation of less than 5%.

The radial striations in the penumbra seen by early observers were resolved as alternating bright and dark filaments by the balloon-borne Stratoscope in the late 1950s. Subsequent high-resolution observations, especially at Pic du Midi, showed that the bright filaments consist of several elongated grains which migrate inwards toward the umbra. Some of these bright grains penetrate into the umbra and form the peripheral umbral dots in the umbra (see also SUNSPOT UMBRA: STRUCTURE AND ACTIVITY). Fully resolving the fine structure in the penumbra poses a real challenge for modern high-resolution observing techniques (see figure 1). The power spectrum of intensity fluctuations is fairly flat down to wavelengths at the current limit of angular resolution (about 0.2 arcsec), implying that there is a considerable amount of smaller fine structure yet to be resolved.

In Hα the penumbra also displays a filamentary structure, but at this higher level in the atmosphere the so-called superpenumbra extends radially much further outward, to distances as much as twice the sunspot's radius in white light. In the corona, as seen in x-rays, filamentary features emanating from the inner penumbra extend much further outward, to distances comparable with the solar radius. The filamentary structure of the penumbra is intimately associated with the structure of the sunspot's magnetic field.

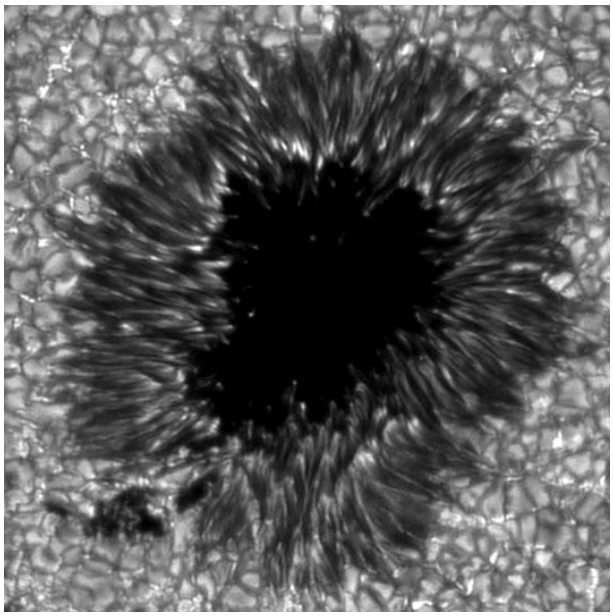


Figure 1. High-resolution image of a sunspot taken in May 1998 with a G-band interference filter (1.2 nm bandpass centered at 430.5 nm) at the Swedish Vacuum Solar Telescope on La Palma, showing the bright and dark penumbral filaments and the bright grains that constitute the bright filaments. Some of the bright grains extend well into the umbra. The normal solar granulation pattern occurs outside the sunspot; the bright points in the intergranular lanes correspond to intense magnetic flux tubes emerging from the solar surface. (Courtesy of T Berger, Lockheed-Martin Solar and Astrophysics Laboratory).

The magnetic field in the penumbra

The mean (azimuthally averaged) magnetic field in a sunspot at the solar surface is vertical at the center of the umbra but flares outward so that the mean field is inclined at about 20° to the horizontal at the outer edge of the penumbra. The strength of the mean magnetic field decreases monotonically with radial distance from the center of the sunspot, dropping by a factor of 2 or 3 at the outer edge of the penumbra. Typical magnetic field strengths are 3000 G in the center of the umbra and 1000 G near the outer edge of the penumbra.

The detailed geometry of the magnetic field in the penumbra is far more complicated than that of the axisymmetric mean field discussed above, and indeed a clear picture has emerged only in the last decade (see also SUNSPOT MAGNETIC FIELDS). Formerly it was thought that the penumbra is a rather shallow structure formed by magnetic field lines emerging through the solar surface in the outer umbra and bending over sharply to lie horizontally along the solar surface. In this picture, very little magnetic flux actually emerges through the solar surface within the penumbra. Now, however, we know that the penumbra is a deep structure in which a significant fraction, more than half, of the sunspot's magnetic flux emerges through the surface.

Observations by Beckers and Schröter at Sacramento Peak in the 1960s suggested that, while the magnetic field in the dark filaments is nearly horizontal (relative to the solar surface), the field in the bright filaments is significantly inclined to the horizontal. More recently, high-resolution spectroscopy and precision polarimetry have revealed a 'fluted' or 'corrugated' structure of the penumbral magnetic field, with the inclination of the field varying azimuthally around the sunspot (see figure 2). In the outer penumbra, the inclination to the horizontal varies from 0° in the dark filaments to 35°–40° in the bright filaments, giving a mean inclination of 15°–20°. In the inner penumbra, the mean inclination is about 45° with azimuthal fluctuations of ±15°–20°. The magnetic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

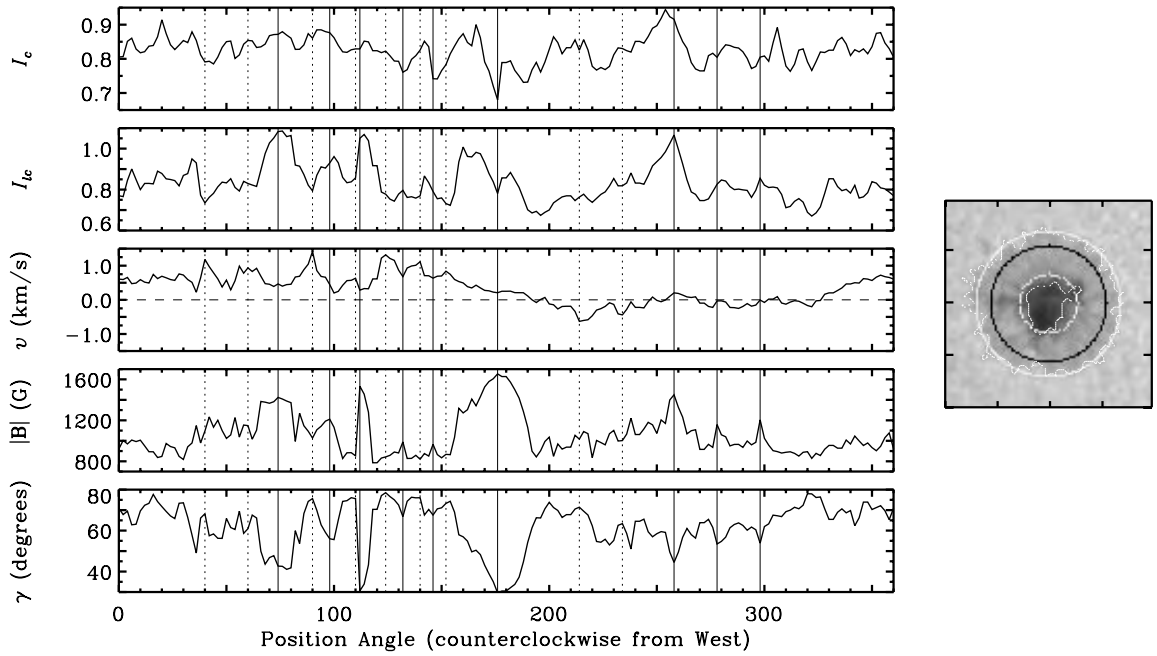


Figure 2. The azimuthal variation of several quantities around a circle (shown in black on the continuum image) in the outer penumbra of radius $0.8R$, where R is the radius of the circle that just contains the entire penumbra (the outer white circle). Shown here are the variations of magnetic field strength $|\boldsymbol{B}|$, zenith angle γ of the magnetic field, line-of-sight Doppler velocity v (in Fe I 630.25), continuum intensity I_c and line core intensity I_{lc} (in Fe I 630.25) with position angle (in degrees, measured counterclockwise from solar west). The solid vertical lines mark the locations of significant local maxima in the field strength $|\boldsymbol{B}|$; note the strong anticorrelation between field strength and inclination angle in the 'fluted' penumbra. The dotted vertical lines mark the locations of significant local maxima of the magnitude of the Doppler (Evershed) flow velocity, $|v|$; note that these flow channels generally coincide with lower intensity (especially line core intensity I_{lc}) and nearly horizontal magnetic field. (From Stanchfield *et al* (1997).)

field strength is somewhat weaker where the field is more nearly horizontal. The radial 'spines' of stronger, more inclined magnetic field account for most of the emerging magnetic flux in the penumbra.

Motions in the penumbra

Observations of persistent motions in the penumbra show systematic inward proper motions (toward the umbra) of the bright elements in the inner penumbra, while in the outer penumbra the dominant motion is the outward Evershed flow occurring mostly in the dark filaments.

The most readily apparent motion in the penumbra is the radial, nearly horizontal outflow of gas responsible for the Evershed effect seen in spectral lines formed in the penumbral photosphere (see SUNSPOTS: EVERSHED EFFECT). The flow is generally concentrated in the dark penumbral filaments, where the magnetic field is more nearly horizontal, and is directed along the magnetic field lines (see figure 2). This motion is thought to be a siphon flow driven by a pressure difference between the footpoints of arched magnetic flux tubes. This mechanism has only recently been confirmed by high-resolution observations showing the magnetic field lines and flow diving below the solar surface near the outer edge of the penumbra. The siphon flow mechanism is probably also responsible for the reversed, inward Evershed flow of gas seen higher up in the penumbral atmosphere, in chromospheric spectral lines such as Hα.

A sunspot penumbra exhibits a range of interesting wave motions (see SUNSPOT OSCILLATIONS AND SEISMOLOGY). The most conspicuous of these waves are the running penumbral waves seen in Hα. These consist of dark wavefronts in the form of concentric arcs, sometimes extending nearly all the way around the umbra, originating at the umbra–penumbra boundary and propagating radially outward across the penumbra at speeds of 10–20 km s^{-1}, and repeating with a period in the range 200–300 s. The high propagation speeds, well above the local sound speed, indicate that these penumbral waves are largely magnetic in character. The running penumbral waves may be related to umbral oscillations, but the connection is not really clear. Compared with the chromospheric umbral oscillations, the running penumbral waves are much less energetic, having velocity amplitude of only some 200 m s^{-1} at the base of the chromosphere (at the temperature minimum) and 1 km s^{-1} in the chromosphere (at the height of formation of Hα). Other oscillatory motions in the penumbral photosphere, with periods around 5 min, are related to the interaction of the sunspot with the resonant acoustic modes (p-modes) in the surrounding quiet Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

On longer time scales, on the order of a half hour, there occur slow convective interchange motions of magnetic flux tubes. These motions are responsible for the bright and dark penumbral filaments, the fluted geometry of the penumbral magnetic field, and the vertical heat transport in the penumbra. Magnetic flux tubes that initially lie along the outer surface of the sunspot's sheaf of magnetic field, in contact with the hotter surroundings, are heated, become more buoyant, and rise upward, carrying excess heat and appearing bright. The footpoints of these rising inclined flux tubes, where the tubes emerge at the solar surface, migrate inward producing the inward-moving bright grains. Meanwhile, the rising, heated flux tubes are replaced by descending, cooler flux tubes (in the dark filaments) which, on reaching the outer boundary, are in turn heated. This interchange process repeats itself continually, producing the quasi-steady vertical transport of heat responsible for the penumbra's relative brightness compared to the umbra.

The formation of a penumbra

Magnetic flux in emerging active regions first appears in the form of SUNSPOT PORES, which then may coalesce to form sunspots. A sunspot is distinguished from a pore by having a penumbra. The formation of a penumbra is apparently the result of a sudden transition in the structure of the accumulated magnetic field, from a bundle of nearly vertical field lines to a widely fanning sheaf of field lines with nearly vertical field in the umbra and nearly horizontal field in the penumbra. This transition between a pore and a sunspot, as the amount of accumulated magnetic flux increases, is a complicated nonlinear process involving the onset of overturning magnetoconvection at the outer edge of the pore (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION). Although sunspots are generally larger than pores, there are pores as large as 3.5 Mm in radius and sunspots (with rudimentary penumbrae) as small as 1.8 Mm in radius. Our understanding of this situation is that, if we consider the radius as a function of total magnetic flux, there are two families of stable solutions which overlap at intermediate radii. The transition between the two states occurs when filamentary convection sets in suddenly to form a penumbra at the outer edge of the pore, where the magnetic field becomes more nearly horizontal with increasing total magnetic flux. In the language of nonlinear systems, the pore solution loses stability at a saddle-node bifurcation.

Bibliography

Rucklidge A M, Schmidt H U and Weiss N O 1995 The abrupt development of penumbrae in sunspots *Mon. Not. R. Astron. Soc.* **273** 491—8

Schlichenmaier R, Jahn K and Schmidt H U 1998 A dynamical model for the penumbral fine structure and the Evershed effect in sunspots *Astrophys. J.* **493** L121–4

Stanchfield D C H II, Thomas J H and Lites B W 1997 The vector magnetic field, Evershed flow, and intensity in a sunspot *Astrophys. J.* **477** 485–94

Thomas J H and Weiss N O 1992 The theory of sunspots *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) pp 3–59

John H Thomas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Pores

Basic properties of pores

Pores are, roughly speaking, small SUNSPOTS without a clear penumbra (see SUNSPOT PENUMBRA). While the visual appearance of a sunspot is that of a very dark central umbra surrounded by a filamentary annular penumbra, a pore is a fairly uniform dark spot on the solar disk. Like sunspots, they are the surface manifestation of strongly magnetized tubes permeating the visible solar surface layers or SOLAR PHOTOSPHERE. Their diameter in continuum images is typically only a few (2–4) thousand kilometers and they appear for hours to a couple of days at most. Their surface magnetic field is of the order of 1000–2000 G (the Earth's magnetic field is about 0.4 G). In contrast, a sunspot can easily be 10 times larger in size and can exist for weeks or months, while the magnetic field in the centre of a sunspot umbra can be twice as large.

The distinctive property of pores is the associated darkening in photospheric intensity. Using the morphological definition that labels pores as those magnetic structures that lack an observable penumbra, the size distribution of pores and sunspots overlap. The largest pores (3–4 Mm radius judged from their continuum appearance) may suddenly form a partial penumbra, turning them into sunspots. Both observational and theoretical arguments support a scenario for penumbra formation when a critical size and field strength is exceeded. The transition would take place at a total magnetic flux of order 10^{20} Mx (Mx, Maxwell, defined as 10^{16} Mx $= 1$ GMm^2). In that sense, pores are just a phase in the build-up towards a stable mature sunspot. Emerging magnetic flux in an ACTIVE REGION rapidly organizes itself into pores, which may gather to organize themselves into a sunspot. A similar process acts to form a single pore: magnetic elements (250 km diameter flux tubes with 1–2 kG fields) coalesce somewhat spontaneously to form a pore. Individual magnetic elements of size $\sim$200 km diameter have been observed and constitute up to 90% of the net magnetic flux outside sunspots and pores. Within several hours, pores can evolve in size, form or disappear, so they are truly transient sunspots.

Formation scenario

As a rule, pores develop in very young active regions. Figure 1 shows a close-up view of the solar photospheric layers in continuum intensity. Shown is an area of 48.12 square arcseconds of the solar surface as seen at a wavelength of 6708 Å. This continuum image was obtained with the Swedish Vacuum Solar Telescope (La Palma, Spain) on 10 May 1994. The GRANULATION pattern of bright cellular patches separated by darker and cooler lanes is seen all over the solar surface, while the three black spots are pores. The granulation represents the uppermost layer of the solar convection zone (see SOLAR INTERIOR: CONVECTION ZONE), the superadiabatically stratified outer 30% of the solar envelope. The mean distance between the bright granule centers is 1.5–2 Mm. Hence, the size of the pores in the figure varies from 1 Mm diameter for the smallest one to 3–4 Mm for the largest. Clearly, their shape can be very irregular, but they are fairly stable structures compared with the surrounding granulation pattern. The bright granules typically last for 16 min or so, while (large) pores can exist for days. Figure 1 is of exceptional quality, since hints of internal structure, reminiscent of umbral dots (see SUNSPOT UMBRA) can be detected in the largest pore shown. A clear discussion of the fine structure in pores can be found in Bonet *et al* (1995). They observed pores on the Sun during a partial solar eclipse, which allowed them to correct precisely for the seeing degradation induced by the telescope and the atmosphere of the earth.

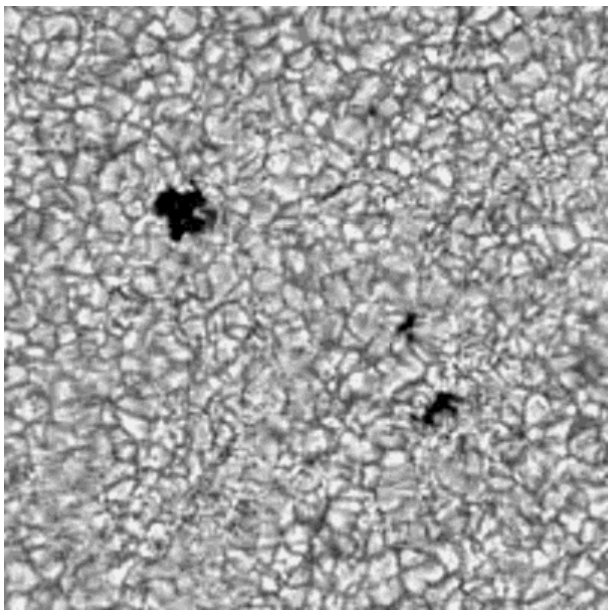


Figure 1. Pores seen on the solar disk in white light. The area shown is approximately 35 Mm $\times$ 35 Mm. Figure courtesy of V Martínez Pillet.

While ideally a pore is thought of as a monolithic tubular entity, it likely consists of a collection of even smaller magnetic elements. These ought to be tightly packed and linked deeper down as one magnetic trunk (Martínez Pillet 1997). A scenario for the formation of strongly magnetized 'trunks' is as follows. Large-scale flow fields tend to concentrate the magnetic field at the photosphere into the intergranular lanes. There, magnetic elements and even small pores may form due to a convective collapse. Thereby, a downflow in a magnetized tube accelerates. As a result, the tube evacuates, which in turn leads to a transverse amplification of the field. The resulting 1–2 kG magnetic fibrils are then constantly buffeted by the large convective cells. Converging flow fields in the local granular or meso-granular convection pattern could draw several magnetic elements together and form the larger pores. The surrounding downflow adds to the stability of the structure. The accumulated magnetic flux will eventually affect the local magnetoconvection pattern and darken (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Keppens and Martínez Pillet (1996) reported the detection of pore formation, associated with a 5% continuum contrast, as soon as the magnetic flux reaches (4–5)$\times 10^{19}$ Mx within a 1.5 Mm radius.

Temperature structure

A statistical survey reports few or no pores close to the solar limb. This is surely associated with the identification of pores as dark structures compared to their surroundings. Their internal vertical temperature stratification can change their visual appearance from dark and cool at the disk center to invisible or brighter than the ambient atmosphere at the limb. Continuum observations towards the limb actually sample higher layers, so that one really needs to determine the two-dimensional (assuming axisymmetry in the third dimension) temperature structure of a solar pore. Sütterlin (1998) presents such a detailed analysis which confirms that the temperature difference between the pore center and surroundings drops with height and may even reverse for small pores. Figure 2 shows the inferred temperature structure as a function of height and radial distance for a pore with a 3.2 Mm continuum diameter (taken from Sütterlin 1998). The temperature range is indicated in kelvin on top. Equally important in Sütterlin's study is the determination of the Wilson depression across the pore, which is the geometrical height from which most continuum radiation escapes freely (optical depth $\tau = 1$). For the pore in figure 2, the $\tau = 1$ surface varied by 200 km across the pore.

Although it is difficult to assess the influence of stray light in the continuum intensity of a pore, it has been established that the radial variation of the brightness in pores is a smooth increase towards its rim. This is quite different in sunspots, where the continuum appearance is that of a very dark central umbral region, surrounded by a plateau of a brighter, more filamentary penumbral annulus. The central region of a pore is darker and hence, cooler, for larger pores. In cases where a sunspot forms, the original pore remains as the umbra.

Magnetic structure of pores

The strong magnetic field in pores is immediately responsible for their darkness relative to the surrounding photosphere. When a significant amount of flux is accumulated within a 1.5 Mm radius, this area darkens (cools) and turns into a pore. Indeed, a sufficiently bundled and localized kG magnetic field will disturb the outward convective energy transport towards the solar visible surface layers, thus forming the cooler pore. The value of the field strength can be determined from magnetically sensitive absorption line spectra. Polarimetric measurements of, for example Fe I lines, are now routinely used to derive the vector magnetic field in sunspots, pores and plage regions. The strong magnetic fields lift degeneracies in the atomic levels of the ions. Stokes polarimetry uses the resulting wavelength variation of the intensity, the net circular and both net

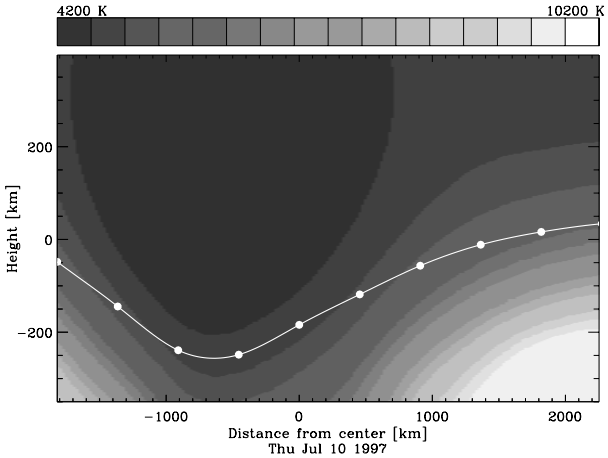


Figure 2. The two-dimensional temperature structure of a pore. The white line indicates the $\tau = 1$ surface. Figure courtesy of P Sütterlin.

linear polarization effects to determine the magnetic field strength, the inclination of the field to the local vertical and its azimuth. This inversion process involves the tuning of parameters representing the magnetic and thermodynamic atmosphere until the profiles closely agree with the ones observed. In this way, one can construct magnetic maps of the solar surface giving the full 3D structure of the field. In such maps, sunspots and pores are readily identified as coherent, strongly magnetized structures. In maps of magnetic field azimuth angle, they both show a well-defined center about which a full 360° range in angle is observed. Keppens and Martínez Pillet (1996) selected magnetic features from azimuth maps and inferred their magnetic radii. These magnetic structures do not always correspond to a continuum dark spot like a pore or sunspot. Azimuth centers without pronounced continuum contrast were termed 'ACs' and these ACs sometimes evolve in a pore within a time span of hours.

Statistics on a sample of 51 pores showed that their magnetic extent is always larger than their continuum appearance. Whether or not this is to be attributed to the existence of a magnetic canopy surrounding the pore proper is unclear. Martínez Pillet (1997) proposed a cartoon pore model consisting of a densely packed dark central part with an outer rim of magnetic elements linked together subsurfacely. Since loosely spaced magnetic elements insufficiently inhibit convective energy transport, a pore then merely seems smaller judged from continuum images alone. Only pore seismology, combined with high-resolution polarimetric observations in infrared wavelengths (which probe deeper layers than the Fe I lines) may tell us whether this is indeed the case.

To determine the magnetic structure across a typical pore, Keppens and Martínez Pillet (1996) calculated sample mean radial profiles of the azimuthally averaged magnetic field. In particular, since each pore in their sample had a readily identifiable azimuth center,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

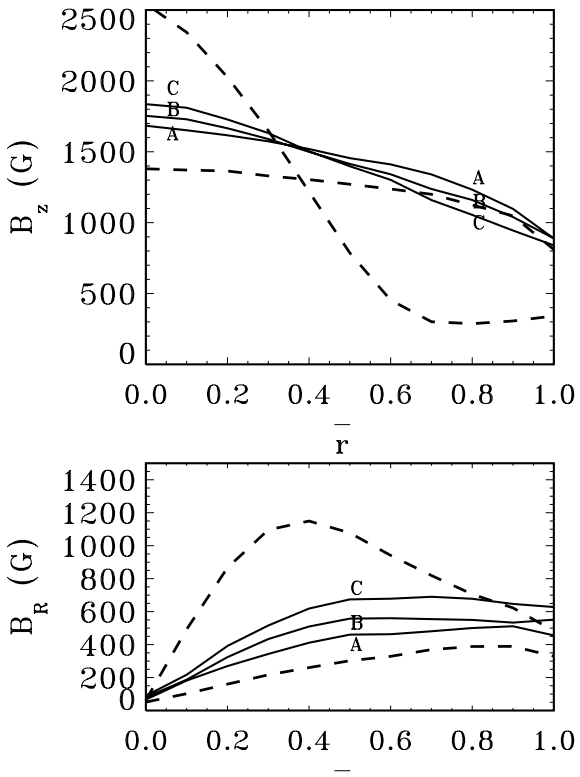


Figure 3. The magnetic structure of pores compared with sunspot magnetic fields.

averaging the 3D field structure about this center gives reliable information on the vertical and radial field structure. Grouping the 51 pore sample in three subgroups A, B and C of increasingly larger pores, the variation of the field structure with size could be deduced. The same procedure was used to obtain radial profiles for a typical sunspot, and for the ACs (those magnetic structures clearly detected in azimuth maps without a clear continuum contrast, believed to be candidates for pore formation). The results in figure 3 show the vertical magnetic field components in the top panel, and the radial ones in the bottom panel. The virtually horizontal dashed curves correspond to the sample of ACs, which are seen to be purely vertical magnetic structures or flux tubes. Their maximal field strength is roughly 1400 G. The plot confirms that for pores of increasing size (from A to C), the central field strength increases from 1700 to 1900 G. Therefore, a clear correlation between pore size, continuum intensity and magnetic field strength exists. Note further how figure 3 indicates that pores are still fairly vertical structures (the average inclination at the edge derived from the sample was found to be 35°), but a tendency to fan open with size is also evident. The largest variation of the field structure is found for sunspots (dashed curves), where interestingly, the penumbra boundary coincides with a 40–45° inclination so that the azimuthally averaged penumbral field is primarily horizontal, while it is almost vertical in a sunspot umbra.

This suggests that when the field inclination at the edge of a stable magnetic structure increases beyond 40–45°, the concave tube boundary may be susceptible to a fluting or interchange instability, thereby fragmenting its boundary and forming a filamentary, predominantly horizontal 'penumbra'. In fact, the dynamic and fibril nature of a sunspot penumbra suggests that the instability is convective in nature. That a semistable horizontal penumbra should eventually form could be expected from the stabilizing influence of buoyancy, acting at sufficiently strong inclination. An observational clue to this process was presented by Keppens and Martínez Pillet (1996), where part of a large pore showed an appreciable change in its continuum intensity within a few hours. Analysing the magnetic field evolution in time, the magnetic structure changed from almost vertical throughout, to a two-component structure very much like a typical sunspot. More observations of higher spatial and temporal resolution of evolving large pores are needed to confirm this scenario. It is thereby vital to have simultaneous continuum and polarimetric measurements.

More measurements of the field inclination in pores seem warranted, to test the hypothesis of penumbra formation when a critical angle is exceeded. This scenario has been investigated by Rucklidge *et al* (1995). They constructed pore and sunspot models taking into account flux conservation and a balance between surface emitted energy and the energy entering laterally and from below. In solution diagrams of spot size versus magnetic flux, pores and sunspots form distinct curves which overlap in their size range, as observed. The pore solutions lose stability at a bifurcation point, indicating the onset of a penumbral filamentary convection as soon as the edge field tilted beyond a critical angle (taken at 45°). Beyond this inclination, the energy balance dictated the need for convective transport across the magnetopause into the flux tube. Observations reveal that the field inclination in pores varies almost linearly with radial distance from the center. Edge values up to 60° with respect to the surface normal have been reported (Sütterlin 1998), comparable to inclinations for the bright filaments in a sunspot penumbra (the dark ones are nearly horizontal). The statistical analysis by Keppens and Martínez Pillet (1996) found a mean inclination of 35° at the magnetic edge of the pore, after azimuthal averaging. The main difference results presumably from the underlying model used in the inversion procedure. Although both inversion codes used a set of eleven parameters to determine the emergent Stokes vector, Sütterlin (1998) emphasizes the two-dimensional temperature structure. He included parameters related to the vertical thermodynamic (and magnetic) variations. Lites and Skumanich (1990) consider all thermal, magnetic and dynamic quantities as constant through the line forming region, except for the linear τ-dependence of the source function. In this respect, the variation of the Wilson depression as shown in figure 2

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

has consequences for interpreting the vector magnetic field data. Determinations of the vertical field gradient in pores quote values around $\mathrm{d}B/\mathrm{d}z = -4\,\mathrm{G\,km^{-1}}$.

Pore models and wave modes

Since pores are rather stable compared with the shortlived granules around them, models try to mimick a magnetohydrostatic balance in stratified, radiating surroundings. First modeling efforts used potential field solutions neglecting stratification and radiation. Models like those by Rucklidge *et al* (1995) concentrate on the connection between pores and sunspots by penumbra formation. Full modeling is difficult when accounting for both the superadiabatically stratified convective surroundings and the transition to the optically thin regime. Two-dimensional time-dependent numerical simulations of magnetic elements and pores in non-stationary, radiative convection show that the smaller magnetic structures are strongly influenced dynamically. The photospheric indications of fine structure in pores also tell us that full three-dimensional modeling of magnetoconvection is required.

Due to the dynamic surroundings, MAGNETOHYDRODYNAMIC WAVE modes of all types are continuously excited in pores. The study of wave modes of and wave interactions with flux tubes is a subject in its own right. For small, stratified magnetic flux tubes, the wave modes can be studied by neglecting their radial variation. For the larger structures, this radial variation can no longer be ignored. Wave modes of purely radially stratified magnetic flux tubes embedded in wave-carrying surroundings can describe their linear dynamics satisfactorily. Radial variations across the primarily vertical magnetic field allow for localized wave heating. For magnetic elements, one-dimensional studies of this process indicated that slow resonant absorption can be acting in a layer 400 km above $\tau = 1$. In general, both radial and vertical gradients play an important role in the interaction (see SUNSPOT OSCILLATIONS AND SEISMOLOGY). Numerical simulations of wave interactions with a stratified vertical field showed how impinging P-MODES can also be converted to slow magneto-atmospheric-gravity waves, ducted along the field lines. The study of linear wave modes yields invaluable insight for the nonlinear numerical simulations where the truly dynamic nature of the smaller magnetic structures is evident.

Pore seismology?

A scenario for pore formation by packing magnetic elements in a bundle raises the question whether a pore, once formed, is a monolithic or fibril structure. Also, the suggested subsurface appearance as a larger magnetic trunk should be deduced from observations. Attempts to answer such questions for sunspots make use of the indirect sampling of the sunspot by impinging p-modes. It has been noted that a fibril structure can be more efficient in extracting the incident power since multiple scattering ensures that the exciting waves revisit the sites of energy loss, and the losses add coherently. Large pores can be sampled by techniques developed in sunspot and active region seismology. The way pores form suggests a fibril model which indicates they could be efficient sinks of p-mode energy. Such studies will be hampered by the fairly rapid evolution and the short lifetime of pores. Moreover, due to their size and intrinsic field strengths, they may not correspond to sufficiently strong scatterers of acoustic wave energy. However, a full three-dimensional reconstruction of the pore structure would give valuable indications of the possible scenarios for the onset of penumbra formation.

Conclusions

Since the size of the smallest pores and the magnetic elements is comparable to the best obtainable resolution to date, efforts to deduce their three-dimensional composition as well as their dynamic properties will continue. Model simulations should address the details of pore formation and stability. The link between magnetic structures of all sizes is still an issue worth studying. Ultimately, this will help unravel the relation between the surface manifestation of the solar magnetic field and the way in which this field is generated and evolves subsurfacely. Pore formation and coalesence represent the first clear visual clues to the emergence of large Ω-shaped flux loops. Simulations have confirmed the possibilities for storing and amplifying enormous toroidal flux ropes in the overshoot region below the convection zone. They can rise through the convection zone and finally erupt in a tree structure to lead to the surface active region development. The determination of the atmospheric and magnetic conditions prevailing in pores may guide the required modeling of this eruption phase.

Bibliography

Bonet J A, Sobotka M and Vázquez M 1995 Photometry of sunspot pores from partial eclipse observations *Astron. Astrophys.* **296** 241–7

Keppens R and Martínez Pillet V 1996 The magnetic structure of pores and sunspots derived from Advanced Stokes polarimeter data *Astron. Astrophys.* **316** 229–42

Lites B W and Skumanich A 1990 Stokes profile analysis and vector magnetic fields *Astrophys. J.* **348** 747–60

Martínez Pillet V 1997 Polarimetric measurements of sunspots *Astronomical Society of the Pacific Advances in the Physics of Sunspots (First Advances in Solar Physics Euroconference Conf. Ser. 118)* ed B Schmieder, J C del Toro Iniesta and M Vázquez, pp 212–35

Rucklidge A M, Schmidt H U and Weiss N O 1995 The abrupt development of penumbrae in sunspots *Mon. Not. R. Astron. Soc.* **273** 491–8

Sütterlin P 1998 Properties of solar pores *Astron. Astrophys.* **333** 305–12

Rony Keppens

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspot Umbra: Structure and Evolution

Sunspots show two main structures: a central dark region, the umbra, surrounded by a brighter and filamentary zone, the SUNSPOT PENUMBRA (see figure 1 in the article on SUNSPOT EVOLUTION). Sunspots without penumbra are usually called SUNSPOT PORES. Observed with low spatial resolution, the umbra appears homogeneous. However, even by the nineteenth century astronomers were able to detect fine details. In fact all the structures mentioned below were observed at that time. The names of Secchi, Janssen and Chevalier should be mentioned in this context.

An important development in our knowledge of the fine structure of sunspots occurred in the 1960s with the balloon observations of the Stratoscope Project (M Schwarzschild at Princeton University) and with ground-based observations reported in the monograph of Bray and Loughhead (1967). However, the low dynamic range of the photographic plates available at that time did not allow reliable photometric measurements of the fine details inside the umbra to be made. In recent times the use of CCD cameras has solved this problem. The application of image-selection techniques at observatories with excellent atmospheric conditions (Pic du Midi, Sacramento Peak, Tenerife and La Palma) has enabled time series of sunspot images to be obtained at very high spatial resolution ($\approx 0.3''$). In the near future, the application of two-dimensional spectropolarimetric and image reconstruction (speckle interferometry, phase diversity) techniques will improve our knowledge of these features yet further.

Very often the umbras of individual sunpots within a group are divided into different parts by a bright structure known as a light bridge (hereafter LB). We will call these individual umbras umbral cores (UCs). A schematic view of the detailed structures observed in sunspot umbras is shown in figure 1.

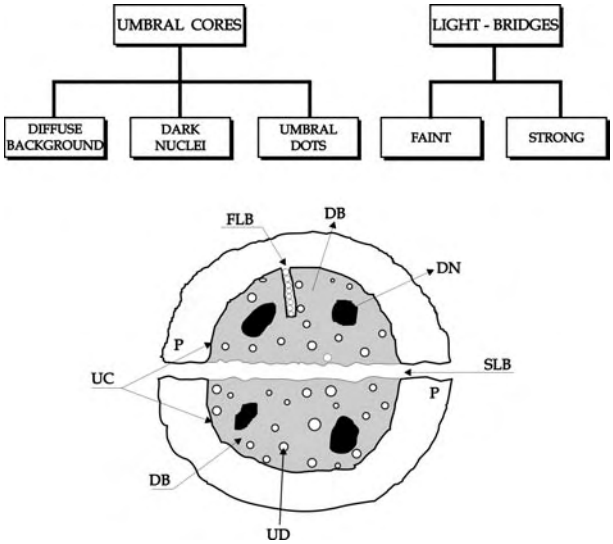


Figure 1. A chart and an idealized scheme showing the different fine structures visible in sunspot umbras.

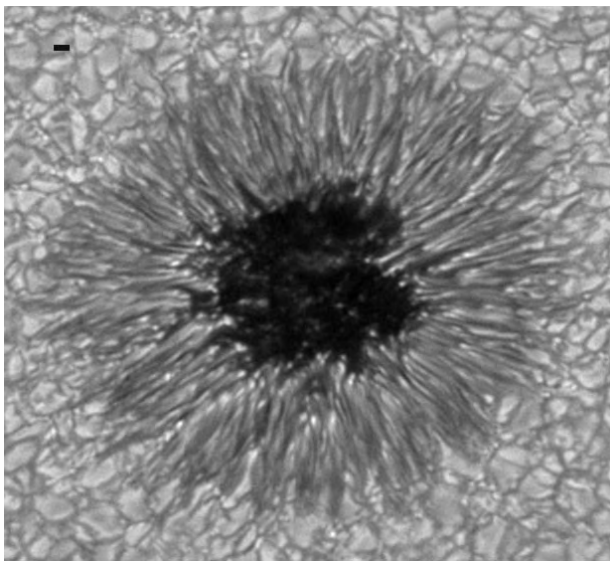


Figure 2. Mature sunspot observed at the Swedish Vacuum Solar Telescope at the Observatorio del Roque de los Muchachos (La Palma). The black strip on the upper left corner of the image is $1'' \approx 725$ km long. This image, and that in figure 3, have been corrected for the degradation produced by the telescope and the average atmospheric conditions at the observing site.

Umbral cores

Umbral cores have smoothly varying intensities with brighter and darker regions, known as the diffuse background (DB), which has two principal features, the dark nuclei (DN), which correspond to distinctive local intensity minima of the core, and umbral dots (UDs), small bright structures embedded in the diffuse background. A mature sunspot with an umbra displaying all these fine structures is shown in figure 2.

Dark nuclei

They are the zones around the minima of local brightness inside the umbra. The name 'void' has also been used. The brightnesses are spread in a wide range (see table 1) depending of the individual characteristics of umbral cores (e.g. size). Since the continuum brightness is related to the magnetic field strength, DN are also the locations of the strongest magnetic fields. The DN are not necessarily located at the center of the umbra, but are rather distributed irregularly in the umbral cores and are almost free of bright structures.

The intensity minimum of the umbral core is the parameter that best characterizes the thermodynamic properties of an umbral core. It is correlated with the sunspot size; this means that larger sunspots are darker.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

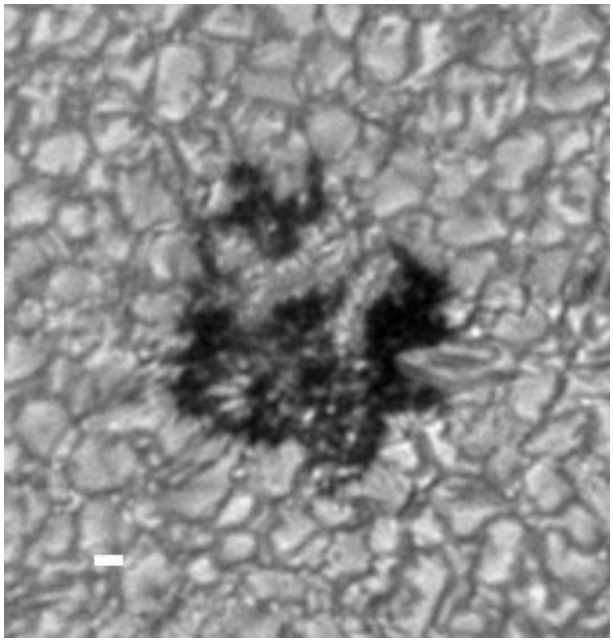


Figure 3. Solar pore observed at the Swedish Vacuum Solar Telescope at the Observatorio del Roque de los Muchachos. Two faint light bridges are observed at the upper part and a lot of umbral dots are embedded in the diffuse background. The white strip on the lower left corner corresponds to $1'' \approx 725$ km.

Umbral dots

Due their morphological similarity to photospheric granules, the bright structures in the umbra were first named 'umbral granulation'. Stratoscope observations showed that they were much narrower and darker than granulation (see SOLAR PHOTOSPHERE: GRANULATION) and they have accordingly received the name 'umbral dots'. However, like photospheric granules, UDs do not have a typical size or lifetime, the smallest (0.2″) and shortest-lived (3 min) ones being the most numerous.

The brightness of UDs in sunspots is related to the brightness of the surrounding diffuse background. On average, we have

$$I_{\rm ud} \approx 3 \times I_{\rm db}$$

where $I_{\rm ud}$ and $I_{\rm db}$ are the intensities of the umbral dots and the diffuse background respectively.

The brightest and fastest moving UDs (0.5 km s^{-1}) are found far from the DN. The fraction of umbral surface occupied by UDs, the so-called filling factor, does not exceed 10 per cent.

UDs in pores (see figure 3) are similar to those in mature umbras but are longer-lived and brighter, and have a higher filling factor.

During their lifetime, UDs show variations in brightness. Significant peaks at periods between 3 and 34 min have been detected.

At least some UDs are connected with other structures outside the umbra. This has been demonstrated by using time series of images of mature sunspots, such as that in figure 2. These show bright elements in the penumbra (penumbral grains) moving into the umbra. A few cross the penumbra/umbra boundary, becoming UDs, which move further into the umbra. When they meet a DN, they brake their motion and disappear. It has been observed that the encounter of a UD with a DN produces a brightening of the UDs located on the opposite side of the nucleus. If these brightenings are produced by a perturbation propagating across the DN, the velocities of propagation would range between 2 and 7 km s^{-1}.

A complementary view of the dynamics of UDs has been provided by time-series observations of solar pores. The existence of a direct interface between the photosphere and the umbra has allowed the possible relation between granules and UDs to be studied. Small granules located close to the pore are pushed by the mesogranular flows and sometimes penetrate into the pore. From there they move inward as features similar to UDs.

Light bridges

Umbral cores are crossed by bright structures that in some cases completely split the umbra into two separate umbral cores. Such features were named bridges because they were originally interpreted as structures overlying the umbra and were thought to consist of facular material. Nowadays we know that they are an intrinsic part of the sunspot and represent the first clear manifestation of sunspot decay. The oscillations (see SUNSPOT OSCILLATIONS AND SEISMOLOGY) observed in LBs are not in phase with those of the adjacent UCs; this supports the idea that, although both are at the same level in the atmosphere, they are physically different structures.

LBs are classified, according to their width, as:

(a) *Faint light bridges* (FLBs; also known as 'umbral bridges'): These are narrow bright streamers, which do not split the umbra. Their brightness is similar to, or lower than, that of the penumbra. FLBs show granule-like structures, $\approx 0.5''$ wide, similar by size to UDs and small granules observed in fields of abnormal granulation.

(b) *Strong light bridges* (SLBs; also known as 'photospheric bridges'): Broad structures ($>2''$) which cross the umbra, dividing it into two or more UCs. Their brightness is related to their width, and reaches photospheric values for the broadest ones.
SLBs show two different populations of granules. One family consists of structures of sizes similar to those of FLBs, and the other with granules similar in size to photospheric granules. The second family begins to dominate with increasing width.

In complex δ sunspots, bridges with filamentary structure, joining umbral cores of different magnetic polarity, are observed. For this reason they are called 'penumbral bridges'. They represent a different physical process and will not be considered further here.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A unified view of umbral fine structure

Table 1 summarizes the main observed properties of sunspot fine structure.

Table 1. Characteristics of the main fine structures observed in sunspots (excluding pores). For LBs the size and brightness correspond to the individual grains, and the lifetime to the whole structure. The brightnesses are wavelength dependent and here are indicated as a fraction of the mean photospheric intensity in the green spectral range.

Structure	Size (″)	Brightness	Lifetime
SLB	1.2	0.6–1.0	Days
FLB	0.5	0.5–0.7	
DN	1.5	0.1–0.4	Days
UD	<0.60	0.2–1.0	Minutes

All these different acronyms, structures and properties are probably just a manifestation of the same physical process seen at different levels. In the framework of the cluster model of sunspots (see SUNSPOT MODELS) this process could be the penetration of field-free material from below the solar surface into the sunspot photosphere as a consequence of the weakening of the sunspot magnetic field.

We can present the following picture of the formation and decay of a sunspot:

The magnetic flux emerging from the bottom of the convection zone appears in the photosphere in a fragmented form. A sunspot is formed by the coalescence of different bundles of magnetic flux or sunspot fragments, which are held together by magnetic buoyancy.

Some physical parameters associated with the emergence, such as the twisting of the bundles or the amount of emergent magnetic flux, determine the internal magnetic and thermal structure of the sunspot in the subphotospheric layers, where the fragments preserve their individuality.

In places where the fragments cannot be held together, the subphotospheric material will start to penetrate the photosphere giving rise to the observed bright structures in the sunspot umbra. The brightness, size and motion of these structures will depend on the strength and inclination of the magnetic field in the surrounding areas.

We can assume the existence of two different types of UDs: (a) dark and slow-moving UDs located in and between dark nuclei, which are produced by intrinsic instabilities of the umbra and (b) bright and fast-moving UDs embedded in a relatively bright DB in the peripheral parts of the the umbra, and possibly originating from the penetration of penumbral grains and/or small photospheric granules.

Some UDs are aligned to form FLBs in the interstices of the sunspot fragments. Later, some of these FLBs become wider and brighter, converting to SLBs and producing the splitting of a UC. In these areas the re-establishment of normal photospheric conditions starts, which means the presence of granulation. Sunspots with a larger area occupied by bright fine structures suffer a more rapid decay. The erosion of the sunspot boundaries by the adjacent granules and the splitting of umbral cores by LBs is the observational fingerprint of the sunspot decay, combined with the outward flow of the moving magnetic features.

All these phases should be controlled by the mechanisms that hold together the different components of a sunspot in subphotospheric layers. In the near future, improvement in the techniques of local helioseismology (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY) could reveal the sub-photospheric component, i.e. the roots, of the observed fine structures in the sunspot umbra. So far, downflows have been observed in these layers, which could make possible the cohesion of the sunpots fragments.

Most of these evolutionary links between umbral fine structures remain to be observed throughout the full evolution of a single sunspot and we probably need to wait for the operation of a large solar telescope in space.

From the proposed scheme we can expect a reduction in magnetic field and the presence of upward motions in the bright structures. In fact, magnetic field reduced by 20–40% relative to the surrounding umbra has been observed in SLBs, also showing more inclined field lines giving rise to a canopy-like structure. A reduction of 1–20% has also been observed in regions rich in UDs, but without any fluctuations corresponding to individual UDs. This can be explained by taking into account the fact that the continuum (where the UDs are observed) is formed deeper in the photosphere than the spectral lines, where the velocity and magnetic fields are measured.

Instead of penetrative convection in a cluster model of sunspots, we can also assume that the underlying physical process consists of convection in the presence of a coherent magnetic field (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION). In this case, the theoretical prediction is also coherent with the observational view of umbral bright features presented here, and we cannot decide at present which of these two mechanisms, penetrative convection or magnetoconvection, is the right one.

Bibliography

Blanchflower S M, Rucklidge A M and Weiss N O 1998 Modelling photospheric magnetoconvection *Mon. Not. R. Astron. Soc.* **301** 593-608

Bray R J and Loughhead R E 1967 *Sunspots* (London: Chapman and Hall)

Bray R J and Loughhead R E 1979 *Sunspots* (New York: Dover) (reprint)

Leka K D M 1997 The vector magnetic fields and thermodynamics of sunspot light bridges: the case for field-free disruptions in sunspots *Astrophys. J.* **484** 900–19

Muller R 1992 Fine structure of umbrae and penumbrae *Sunspots: Theory and Observations* ed J H Thomas and N O Weiss (Dordrecht: Kluwer) pp 175–94

Sobotka M 1997 Sunspots seen at high spatial resolution *Advances in Solar Physics Euroconference (Astron. Soc.*

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Pacific Conf. Proc. 118) ed D Schmieder, J A del Toro Iniesta and M Vázquez (San Francisco: Astronomical Society of the Pacific) pp 258–73

Sobotka M, Bonet J A and Vázquez M 1993 A High-resolution study of inhomogeneities in sunspot umbrae *Astrophys. J.* **415** 832–46

Manuel Vázquez

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspots

Sunspots are the sites of strong magnetic fields at the surface of the Sun. They are visible as features that appear dark because they are cooler than the surrounding SOLAR PHOTOSPHERE, owing to partial suppression of convective energy transport by the magnetic field. Sunspots are distinguished from SUNSPOT PORES by having a filamentary SUNSPOT PENUMBRA surrounding a dark SUNSPOT UMBRA. Their diameters range from 3600 km to 50 000 km and a typical spot survives for less than a week, though some large spots persist for several months. Spots are the most prominent signs of solar magnetic activity, and they have been observed systematically since telescopes were invented in the seventeenth century. They occur in bands centered on the equator that extend to latitudes of $\pm 30^\circ$ and their distribution varies with the SOLAR CYCLE, which has an average period of about 11 years. Although sunspots cover less than 0.1% of the solar surface there are other stars that are similar to the Sun but much more active, possessing starspots that cover more than half their surfaces (see SOLAR–STELLAR CONNECTION).

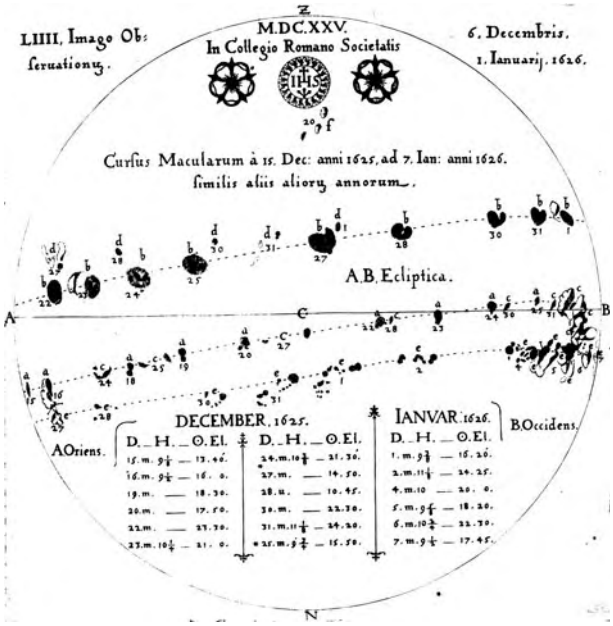


Figure 1. Diagram showing spots moving across the disk of the Sun as it rotates, from Scheiner's *Rosa Ursina sive Sol*, 1628. (Courtesy of the Royal Astronomical Society.)

Historical introduction

Though there are early Chinese records of sunspots their modern history begins in 1611. In that year Johann Fabricius, Christoph Scheiner, Galileo Galilei and Johann Kepler all turned their telescopes to the Sun and found sunspots. Moreover, Galileo interpreted their motion across the visible disk as evidence that the Sun was itself rotating with an apparent period of about 27 days (see figure 1). He also realized that although they were dark relative to their surroundings they were nevertheless as bright as the brightest regions of the Moon. The spots excited considerable interest, which persisted throughout the century. Many observers noticed that sunspots were less frequent between 1645 and 1715, and the record of routine observations at the Paris Observatory confirms that this episode of reduced activity (the Maunder Minimum) was indeed a real effect.

Interest lapsed in the eighteenth century. The early observers had already distinguished the dark central umbra from the annular penumbra but it was not until 1769 that Alexander Wilson noticed that spots appeared increasingly foreshortened as they rotated towards the limb of the Sun. He correctly interpreted this as showing that the visible surface of a spot is depressed relative to the surrounding photosphere (see SUNSPOTS: WILSON EFFECT). Further progress came when William Herschel constructed his 10 ft telescope and used it to observe sunspots in much greater detail: he asserted that they were openings in the solar surface which revealed a cooler layer that might be habitable. Heinrich Schwabe, after observing assiduously for 17 years, announced in 1843 that the incidence of sunspots varied cyclically with a period of about 11 years; this pattern was confirmed by Richard Carrington, who also noted that the spots first appeared at latitudes of around $\pm 30^\circ$ and then filled zones that expanded towards the equator, before contracting and then disappearing at sunspot minimum, as spots of the next cycle appeared at higher latitudes. He also established that the Sun did not rotate uniformly and that the rotation period was shortest for spots in the vicinity of the equator (see SOLAR INTERIOR: ROTATION).

The development of spectroscopy transformed the study of the Sun. Joseph Fraunhofer had discovered absorption lines in the solar spectrum in 1814 and Gustav Kirchhoff realized that they revealed the composition of the solar atmosphere; by 1860 Norman Lockyer and Angelo Secchi were comparing spectra of sunspots with those from the surrounding photosphere. At the turn of the century George Ellery Hale developed the spectroheliograph to produce images of the solar surface in a single spectral line and in 1908 he exploited the ZEEMAN EFFECT (splitting of a spectral line in a magnetic field) to measure the magnetic field in sunspots. The Doppler shift associated with radial outflows in the penumbra (see SUNSPOTS: EVERSHED EFFECT) was detected about a year later. Since then, observations have been dominated by the drive to obtain more detailed spectroscopic results and higher spatial resolution—from the ground, from balloons, and increasingly from space.

These results have stimulated theoretical developments. Ludwig Biermann first suggested that sunspots were cool because convection, which supplies energy to the normal photosphere, was inhibited by their strong magnetic fields. The subsequent formulation of MAGNETOHYDRODYNAMICS by Hannes Alfvén and Thomas Cowling led to a description of this interaction which has been ex-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

tended through numerical modelling of nonlinear magnetoconvection (see MAGNETOHYDRODYNAMICS: MAGNETOCONVECTION). In parallel with this, the work of Eugene Parker and others has led to an understanding of hydromagnetic DYNAMO THEORY as an explanation for global features of the sunspot cycle.

Sunspot groups and active regions

Sunspots typically appear in groups like that shown in figure 2. Frequently there is a dominant pair of spots, oriented roughly parallel to the solar equator, which start close together and move apart as the group evolves, achieving separations of up to 20°. Normally the preceding spot (in the sense of the Sun's rotation) is much more prominent than the follower and has a longer lifetime (see SUNSPOT EVOLUTION). Spot groups lie within SOLAR ACTIVE REGIONS, containing pores and smaller magnetic features that give rise to bright faculae and fine-scale filigree in the solar photosphere (see SOLAR PHOTOSPHERE: FACULAE; FILIGREE). There is a continuous family of isolated magnetic features that can be parametrized by the magnetic flux that they contain. The smallest flux elements have fluxes of 0.001 TWb (10^{17} Mx) and diameters of about 100 km, with fields of up to 0.15 T (or 1.5 kG)—sufficient to balance the external gas pressure, so that the flux tubes are almost completely evacuated at the photosphere (see SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES). These flux elements nestle between the convection cells that give rise to the photospheric GRANULATION. Pores are larger: they are isolated umbras, both darker and cooler than the normal photosphere, with fields of around 2000 G and fluxes of up to 7 TWb. Larger flux concentrations invariably develop the filamentary penumbra that surrounds the umbra and distinguishes sunspots from pores. The lifetimes of sunspot groups range from a few days to several months. Although most sunspots survive for less than a week there are examples of large individual spots that survive for much longer and return several times as the Sun rotates.

The polarities of the magnetic fields in sunspots follow Hale's law. Within any group or active region the preceding and following spots have oppositely directed fields and most of the magnetic flux that emerges from one side of the region returns back through the other. At any stage of the solar cycle the preceding and following spots have opposite polarities in the northern and southern hemispheres. Thus there is a strong azimuthal (or toroidal) field emerging through the photosphere which is antisymmetric about the equator (see SOLAR MAGNETIC FIELD). The sense of these fields reverses at the end of each 11 year activity cycle, so that the full magnetic cycle has a period of 22 years.

The structure of individual sunspots

Photospheric structure

The diameters of sunspots can extend, as in one famous case, up to 130 000 km—approximately one-tenth of the solar diameter or ten times the diameter of the Earth. A typical well-developed sunspot, like that in figure 2, is roughly circular and has a diameter of 20 000–30 000 km. Spots have structures that scale with the radius R, depending only on the ratio r/R, where r is the distance from the center of the spot. The central dark umbra has a radius $0.4R$ and a temperature of about 4300 K, 2100 K below that of the surrounding photosphere; as a result it radiates at only 24% of the normal photospheric rate. Surprisingly, the umbral intensity varies with the solar cycle: in the infrared it increases by 20% from sunspot minimum to sunspot maximum. Within the umbra there are small bright features (umbral dots) with diameters in the range from 1800 km to the limit that can be resolved and with a typical value of about 250 km. Their temperatures are comparable with that of the undisturbed photosphere and the larger ones have lifetimes of 30–60 min.

The umbra is surrounded by a filamentary penumbra, occupying 85% of the total area and with an average intensity that is 75% of the photospheric value. Figure 2 shows the irregular filamentary striations, with local intensity ratios varying from 60% in the dark filaments to 95% in bright filaments. The latter are made up of aligned bright grains that migrate towards the umbra with velocities of up to 0.5 km s^{-1}, while the dark filaments include clouds that move out towards the edge of the penumbra. Spectral lines from matter within the dark filaments show Doppler shifts that correspond to outflows with velocities that rise to a maximum of 6 km s^{-1} at the edge of the spot (the Evershed effect).

The magnetic field at the center of a sunspot is nearly vertical, with a magnitude of up to 0.3 T (see SUNSPOT MAGNETIC FIELDS). As r increases the azimuthally averaged field becomes increasingly inclined, reaching an angle of 70° at the outer edge of the penumbra, while its strength decreases to around 0.1 T before dropping abruptly to zero within a distance too small to be resolved. The fine structure of the penumbral field is amazingly complex, for the field directions in the bright and dark filaments differ by 30–35°, with the dark filaments more nearly horizontal. Near the umbra, where the mean field is inclined at an angle of 45° to the vertical, the bright filaments have an inclination of only 30°, while the dark filaments are almost horizontal (and parallel to the Evershed flow) at the outer boundary. There is as yet no adequate explanation for this interlocking comb-like structure.

The atmosphere above sunspots

Observations made in the Hα line of the hydrogen Balmer series ($\lambda = 6562$ Å) show a rich pattern of radial fibrils extending outwards from a sunspot to form the 'superpenumbra'. This pattern starts at the edge of the umbra and fills an annular region with radius about $2R$. The fibrils exhibit time-dependent behavior with ringlike features propagating radially outwards, referred to as running penumbral waves (see SUNSPOT PENUMBRA: STRUCTURE AND ACTIVITY). The Hα line is formed in the CHROMOSPHERE at heights around 3000 km above the solar surface and the fibrils represent the intersections of slender magnetic flux loops with this region of the chromosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

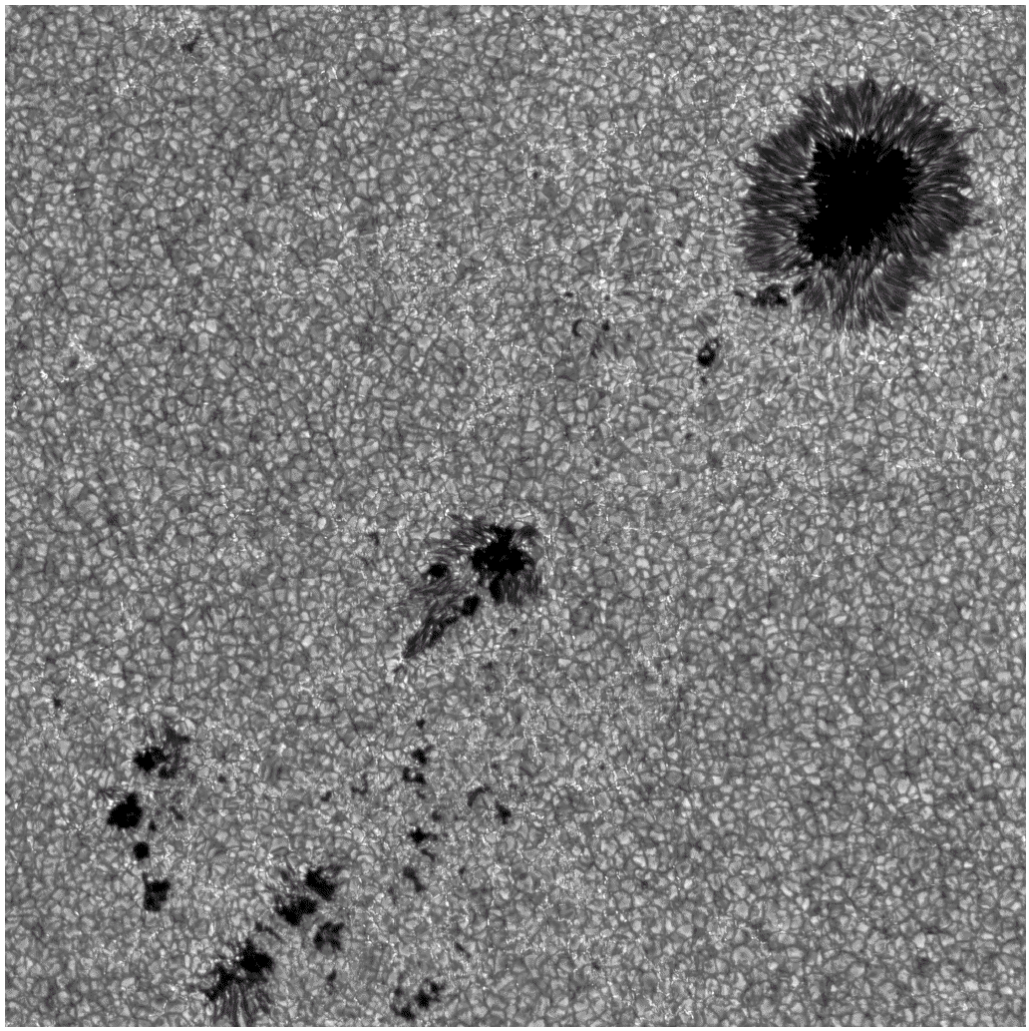


Figure 2. A sunspot group, showing a large preceding spot accompanied by a smaller companion, at top right, and a cluster of small spots and pores with opposite polarity at bottom left. This image, taken in the CH G-band by T Berger and G Scharmer at the Swedish Vacuum Solar Telescope on La Palma, covers a region 120 000 km square on the surface of the Sun with north towards the top left corner. The solar granulation is clearly visible and the small bright points in the intergranular lanes are the sites of locally intense magnetic fields. For an enlarged image of the big sunspot see SUNSPOT PENUMBRA. (Courtesy of T Berger, Lockheed–Martin Solar and Astrophysics Laboratory, Palo Alto.)

These loops have one footpoint within the spot and may return at a great distance from it. Many of them extend up into the hot CORONA and are visible as CORONAL LOOPS in the extreme ultraviolet and in soft x-rays.

The atmosphere above a sunspot supports two different types of oscillations (see SUNSPOT OSCILLATIONS AND SEISMOLOGY). Those with periods around 3 min are resonant oscillations of the magnetized umbral atmosphere itself, while the 5 min oscillations are driven by p-mode oscillations in the convection zone around the spot. The latter provide a means of exploiting HELIOSEISMOLOGY to probe the structure of the spot below the photosphere.

Subphotospheric structure

Continuum observations reveal a surface of constant optical depth. In the umbra of a sunspot this surface lies about 500 km below its level in the ambient photosphere. This Wilson depression implies that the gas pressure in the umbra is much less than that outside the spot at the same geometric level, and therefore that magnetic stresses are required in order to maintain hydrostatic equilibrium in any SUNSPOT MODEL. Since magnetic flux is conserved, the fields that emerge through the surface of a sunspot must continue below the photosphere; the systematic behavior of sunspot groups indicates that the resulting flux tube extends deep into the convection zone (see SOLAR INTERIOR: CONVECTION ZONE) and local helioseismology (see SOLAR INTERIOR: LOCAL HELIOSEISMOLOGY) confirms that there

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is a coherent flux tube for at least 10 000 km below the surface. Within this flux tube energy is carried upwards by convection, though radiation takes over just below the umbral photosphere (see SOLAR INTERIOR: CONVECTION THEORY). Thus the flux tube can be regarded either as an inhomogeneous plug or as a tight cluster of isolated flux tubes, separated by weaker fields or even (in the limit) by field-free plasma. Either picture is compatible with the observations.

Outside the sunspot, energy is transported by convection and several different horizontal scales, corresponding to granulation, mesogranulation and supergranulation, are visible at the photosphere (see SOLAR PHOTOSPHERE: GRANULATION; MESOGRANULATION; SUPERGRANULATION). A well-developed sunspot is surrounded by an annular moat cell—essentially an abnormally large and long-lived supergranule, stabilized by the flux tube at its core—whose diameter may be twice that of the spot. Within the moat there is a vigorous radial outflow which sweeps magnetic fields to its periphery, leaving the moat free of magnetic fields apart from small features (with both polarities) that migrate outwards as the spot decays (see SUNSPOTS: MOVING MAGNETIC FEATURES AND MOAT FLOW). The field inclination at the photosphere indicates that the flux tube contracts below the sunspot. Models imply that the radius of the underlying flux tube is less than $0.5R$, so the energy radiated from the penumbra can be drawn predominantly from the plasma below the magnetopause (the sharp boundary of the flux tube). Such curved boundaries are liable to interchange instabilities but here the magnetopause is strongly inclined and is stabilized by MAGNETIC BUOYANCY. It is nevertheless clear that the observed dynamical stability of sunspots requires the presence of a collar around the flux tube at some greater depth. Recent observational and theoretical results suggest that there may be a slender counterflow beneath the penumbra and within the moat cell. The resulting inflow could act as a collar and so preserve the spot.

Magnetoconvection

Sunspots provided the original motivation for studying the effects of magnetic fields on convection in conducting fluids. The simplest models assume the presence of an imposed vertical field (as in the umbra), which always reduces the efficiency of convection. If the field is strong enough, convection is completely suppressed but as the field strength is reduced, convection sets in, either as oscillations or as steady overturning motion. It is easily shown that radiative transport is not able to supply even the reduced energy flux emerging from the umbra, so convection must take over just below the visible surface. Numerical experiments on nonlinear magnetoconvection suggest that the umbral dots are just chaotic plumes of hot rising gas that penetrate into the photosphere.

As the field becomes more inclined the pattern of convection changes, and rolls with their axes parallel to the horizontal component of the field are preferred. In simple pore models the inclination increases with increasing flux until it reaches a critical value that favors the development of radial filaments and leads to the abrupt formation of a penumbra. This new form of time-dependent filamentary convection is apparently able to transport energy through the penumbra from the magnetopause.

Systematic properties of sunspots

The solar cycle

The butterfly diagram in figure 3 shows the incidence of sunspots as a function of latitude and time. Although individual sunspots and active regions are irregularly distributed, their average behavior is quite systematic (see SUNSPOT CLASSIFICATION). The number of spots is conventionally measured by the sunspot number, an arbitrary measure of activity that is well correlated with sunspot areas (shown in figure 3), as well as with, for example, SOLAR FLARES or coronal x-ray and radio emission (see SOLAR ACTIVITY INDICES). Since 1700 the sunspot number has varied cyclically, with a well-defined mean period of 11.1 years. The solar cycle varies irregularly in amplitude but the available timeseries is not long enough to establish whether this variation is deterministic (chaotic) or stochastic.

Since sunspots are caused by magnetic fields erupting from the interior of the Sun and those fields obey Hale's law, the underlying magnetic cycle has a mean period of 22 years. Magnetic activity is a common feature of stars with deep outer convection zones and cyclic activity is found in middle-aged, slow rotators like the Sun. The scale and systematic properties of active regions suggest that they are formed near the base of the convection zone or even in a region of weak convective overshoot and it is generally accepted that these oscillating fields are maintained by a hydromagnetic dynamo.

The solar dynamo is probably located near the interface between the convection zone and the underlying radiative zone, where there are steep radial gradients in angular velocity (see SOLAR INTERIOR). As strong toroidal fields, with strengths as high as 10 T, are built up they eventually become liable to MAGNETOHYDRODYNAMIC INSTABILITIES driven by magnetic buoyancy. These instabilities could grow until they eventually liberate isolated $\cap$-shaped loops of toroidal flux. The loops would float upwards through the convection zone, developing a Ω-shape as they expand, and then break through the surface to form active regions and sunspots (see SOLAR INTERIOR: EMERGING MAGNETIC FLUX TUBES).

Origin of cyclic magnetic activity

Dynamo theory provides a mechanism for cyclic reversals of the toroidal magnetic field, on a timescale very much shorter than that for decay through ohmic resistivity in the Sun (which is comparable to its lifetime on the main sequence). This mechanism relies on two physical processes. The first is straightforward: a meridional (or poloidal) field in a differentially rotating star will be drawn out to give a strong toroidal component. The second process invokes cyclonic eddies to generate a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

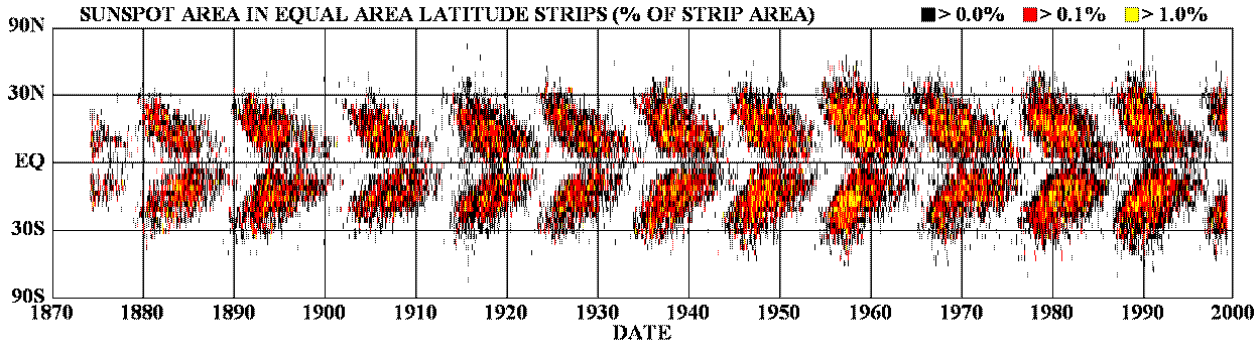


Figure 3. Butterfly diagram showing the observed distribution of sunspots in latitude and time since 1875. At the beginning of the activity cycle spots appear around latitudes of $\pm 30^\circ$ and by sunspot maximum activity extends to the equator. The zones of activity then dwindle until, at sunspot minimum, the last spots of the old cycle appear at low latitudes as the new cycle starts. The distribution of spots is very nearly symmetrical about the equator. (Courtesy of D H Hathaway, NASA/MSFC.)

reversed poloidal field by distorting the toroidal field; these are convective plumes with a mean HELICITY imparted by the Coriolis force. The cycle can then proceed. Most calculations have relied on a simplified form of the theory (mean field electrodynamics), which captures the essence of these processes and yields solutions that reproduce the essential features of the solar cycle (see DYNAMOS: SOLAR AND STELLAR).

The key parameter in stellar dynamo models is the angular velocity Ω, which controls both differential rotation and helicity. If Ω is small, the field-free state is stable to magnetic perturbations. Dynamo action is only possible if Ω exceeds a critical value at which cyclic activity appears as a magnetic instability. As Ω is increased the oscillations grow in amplitude, and more complicated patterns of activity may appear. In fact this sequence follows the evolution of a star like the Sun in reverse, since it arrives on the main sequence spinning rapidly and magnetically very active, and is then spun down by magnetic braking (see SUN AS A STAR and SOLAR EVOLUTION).

Modulation of cyclic activity

The earliest sunspot observations show that the cycle operated then as now but spots became less frequent after 1645 until in the last decade of that century only one spot was observed. Moreover, there were relatively few spots in the cycle that reached its maximum in 1706 and almost all of those were in one hemisphere (unlike the butterflies in figure 3). There were many observers active during this period and there is no doubt that this episode of reduced activity—the Maunder Minimum—was real. Fortunately it is possible to extend the sunspot record back by using proxy data: isotopes such as ^{14}C and ^{10}Be (which are preserved in trees and polar ice, respectively) are formed as a result of galactic cosmic rays impinging on the atmosphere. Since the cosmic ray protons are deflected by magnetic fields in the solar wind, the abundances of cosmogenic isotopes are anticorrelated with solar activity. The ^{14}C and ^{10}Be records show not only the Maunder Minimum but also many previous grand minima recurring aperiodically every few hundred years for the past 10 000 years (see SOLAR ACTIVITY: LONG-TERM RECORDS).

Nonlinear dynamo models are able to reproduce this pattern of aperiodic modulation. As Ω is increased, the periodic cycles may themselves become unstable to oscillations with a much longer period. This transition leads to periodically modulated cycles. As Ω is further increased, the amplitude of the modulation grows until there is a further transition to behavior that is chaotic. Symmetry about the equator may also be broken, especially as solutions are emerging from grand minima.

Sunspots and climatic change

Ever since sunspots were discovered there have been speculations as to their effect on climate. Satellite observations show that the solar luminosity varies with the sunspot cycle and is greater by about 0.15% at sunspot maximum. The estimated increase in SOLAR IRRADIANCE since the Maunder Minimum has a similar value, which might cause an increase of around 0.3 K in global temperature. Comparisons between the sunspot and climatic records suggest that temperature fluctuations are indeed significantly correlated both with the 11 year cycle and with its long-term modulation. The extent to which solar activity has contributed to climatic change is controversial (see SOLAR–TERRESTRIAL CONNECTION: LONG-TERM AND SHORT-TERM CLIMATE VARIABILITY). It seems likely that this has, until recently, been a major effect but that global warming is now predominantly due to greenhouse gases released by burning fossil fuels.

Bibliography

Bray R J and Loughhead R E 1964 *Sunspots* (London: Chapman and Hall) (reprinted Dover Publications 1979)

Priest E R 1982 *Solar Magnetohydrodynamics* (Dordrecht: Reidel)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Proctor M R E and Gilbert A D (ed) 1994 *Lectures on Solar and Planetary Dynamos* (Cambridge: Cambridge University Press)
Tayler R J 1997 *The Sun as a Star* (Cambridge: Cambridge University Press)
Thomas J H and Weiss N O (eds) 1994 *Sunspots: Theory and Observations* (Dordrecht: Kluwer)

Nigel Weiss

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspots: Evershed Effect

The Evershed Effect is an observational effect in the penumbra of SUNSPOTS consisting in displacements of the spectral line profiles towards either shorter wavelengths (that is, blueshifts) in the zones closest to the Sun's disk center or larger wavelengths (that is, redshifts) in those zones closest to the solar limb. Since the first detection in 1909, this effect was correctly interpreted by its discoverer Evershed (1909) as the consequence of mostly horizontal movements of material emanating radially from the spot. Soon after the discovery, a strong variation of the effect with height was detected: the flow was directed inwards when observed in spectral lines formed in the high layers of the sunspot's CHROMOSPHERE (the so-called inverse Evershed effect; see Maltby (1997) for a review). From then on, the Evershed effect has been one of the most prominent dynamical phenomena we see on the Sun and is in fact superimposed on several others (most certainly caused by different physical mechanisms) that take place in the penumbra (see SUNSPOT OSCILLATIONS AND SEISMOLOGY and SUNSPOT PENUMBRA: STRUCTURE AND ACTIVITY). Hence, disentangling the physical process which gives rise to the Evershed effect is one of the most exciting challenges facing solar physics. A very good account of current knowledge and problems, as well as relevant references on this topic, can be found in the two recent reviews by Thomas (1994) and by Maltby (1997).

Main observational features

Early in this century, the widespread belief was that sunspots were made up of strong cyclones or whirls. This was indeed what prompted G E Hale to look for magnetic fields in them (see SUNSPOT MAGNETIC FIELDS) under the assumption of the presence of free, electrically charged particles rotating around the spot axis (the magnetic field would emerge in very much the same way as in today's electromagnets). With the aim of detecting such gyrating motions, Evershed undertook the observation of several FRAUNHOFER LINES in different spots as they crossed the solar surface as a result of solar rotation.

All the basic observational features were established already as early as this pioneering work.

(i) All the spots show line displacements of the same order of magnitude, at the same distance from the disk center, no matter the solar hemisphere they are in.
(ii) The larger the heliocentric angle (angular distance from the disk center) the larger the shifts, which eventually disappear when the spot is located very close to the disk center.
(iii) Blueshifts always occur in the penumbral half closest to the disk center and redshifts in the half closest to the solar limb. No shifts are present at 90° from the direction of the disk center.
(iv) The largest shifts take place in the outermost parts of the penumbra and seem to cease abruptly at the visible penumbral border.

From these observational facts he concluded that the material motions must essentially be horizontal (parallel to the solar surface) as witnessed by the center-to-limb variation (see SOLAR PHOTOSPHERE: CENTER-TO-LIMB VARIATION) of the line displacements which only give information about the component of the vector velocity along the line of sight. Any vortex motion had to be discarded in view of feature 3.

In 1913, C E St John (see reference in, for example Maltby (1997)) reported the observation of 506 lines which confirmed Evershed's results, and he complemented them by finding a proportionality between wavelengths and displacements, clearly obeying the DOPPLER EFFECT due to material motions. Moreover, he found that stronger lines (assumed to be formed higher in the atmosphere) show the opposite effect; that is, material motions towards the central parts of sunspot umbrae. Hence, outward motions were confined to the photosphere while inward motions were located in the chromosphere, with a probably neutral velocity layer in between.

Since those early days, the observational picture of the Evershed effect remains virtually unchanged, while, naturally, more evidence has been compiled and better details obtained because of the increasing degree of sophistication in instrumentation and spectral resolution. Thus, the effect is known to be time dependent and likely to be concentrated in the dark penumbral filaments. Line-of-sight velocities of up to 6 or 7 km s^{-1} (and typical values around 3–4 km s^{-1}) have been observed and it is currently known that absorption lines in the penumbra show asymmetric profiles (i.e. that the line wings are more shifted than the line cores). Such asymmetries have been interpreted as a consequence of a gradient with height of the line-of-sight component of the material velocity and of the vector magnetic field, hence confirming St John's and Evershed's results in a modern way. In figure 1 we see the typical shape of the four Stokes parameter profiles I, Q, U and V (see SOLAR MAGNETIC FIELD: INFERENCE BY POLARIMETRY) observed in a sample point of the limb-side penumbra. The spectral region shows two neutral iron lines located at 630.15 nm and 630.25 nm (in a laboratory wavelength scale denoted by dotted lines), plus two telluric lines from atomic oxygen (the thinner ones in the I plot, formed in the Earth's atmosphere and, hence, with no polarization signal in the Q, U and V plots). The Stokes parameters are normalized to the local quiet sun continuum intensity. Note the redshift and the strong asymmetry displayed by the solar profiles. (The shapes of the asymmetric Q and U profiles conspire, in this particular case, to hide apparently the shift in wavelength.)

The essentially horizontal character of the material motions is in fairly good agreement with the increasingly horizontal magnetic field lines present in penumbrae as one moves away from the sunspot center, since the magnetic field should be frozen to the solar plasma, i.e. the gas should follow the magnetic field lines (see MAGNETOHYDRODYNAMICS). Perhaps the most intriguing phenomenon is the claimed sharp disappearance of any

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

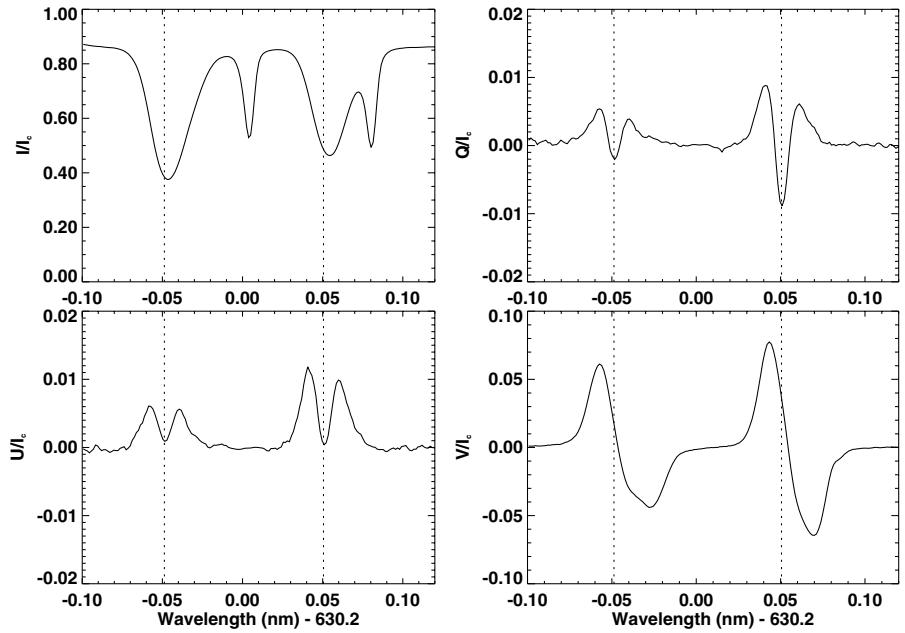


Figure 1. Stokes parameter profiles (in units of the quiet Sun continuum intensity) of two neutral iron lines at 630.15 and 630.25 nm as observed in a sample point of the limb-side penumbra. Laboratory wavelengths are marked with dotted, straight lines. The spot was observed in June 1992 some 72° away from the disk center. The observations were obtained using the Advanced Stokes Polarimeter (ASP) of the National Solar Observatory–High Altitude Observatory working at the VTT telescope of the Sacramento Peak Observatory (Sunspot, New Mexico). (Courtesy of V Martinez Pillet.)

Doppler signature outside the outer visible boundary of the penumbra: just inside this boundary, that is, in the outermost penumbra, it is where the line-of-sight velocities are largest (feature 4 above). If the Evershed effect is indeed the signature of a material flow, where does the mass flux go?

More observational evidence and a reconciling picture

Some authors still assert that the Doppler signatures of the effect cannot be tracked down farther than 1 arcsec away from the penumbral limits. However, reports on tracing the Evershed effect beyond the outer penumbral rim started to appear in the scientific literature by the mid-1980s (Thomas 1994) and continue until today. Particularly interesting is the detection by Rimmele in 1995 (see Maltby (1997) for the reference) of Evershed motions taking place along narrow, elevated channels in the penumbra, some extending far beyond the outer penumbral rim. Through such channels, the material seems to be carried from an upstreaming footpoint, always located within the penumbra, towards a downstreaming footpoint more distant from the sunspot center. This detection was possible after careful examination of observations performed in two lines with fairly different sensitivity to the velocities at different atmospheric layers. The debate about whether or not the Evershed signatures continue away from the spot is certainly not caused by different spatial resolutions since observations with very high resolution (about 350 km on the Sun) have been interpreted according to each possibility (see, e.g. Thomas (1994) for references). The discrepancy must stem from the different sensitivity to the sunspot atmospheric quantities (not only the line-of-sight velocity) of the observing and analysis techniques used by the various researchers.

Further evidence of Evershed flows following the magnetic field lines has been obtained by Solanki, Montavon and Livingston in 1994 (see the reference in Thomas (1994)) with the detection of line-of-sight velocities of up to 2 km s^{-1} in the magnetic 'canopy' which extends beyond the visible penumbral boundary and above the solar surface as seen in the white-light continuum. This detection offered an interesting solution to the aforementioned missing-mass problem, since those velocities could only be inferred from observations with spectral lines sensitive to the elevated layers of the canopy, remaining undetectable from observations of normal photospheric lines: the material could be carried away from the spot through the magnetic canopy. However, the exponential decrease of DENSITY with height in the solar atmosphere (see SOLAR ATMOSPHERE: EMPIRICAL MODELS and SOLAR PHOTOSPHERE) implies that this canopy solution is unlikely because only much higher velocities than the quoted value of 2 km s^{-1} could account for the total mass flux carried within the limits of the penumbra (the mass flux is essentially an integral of the product between density and velocity).

A reconciling observational picture was recently

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

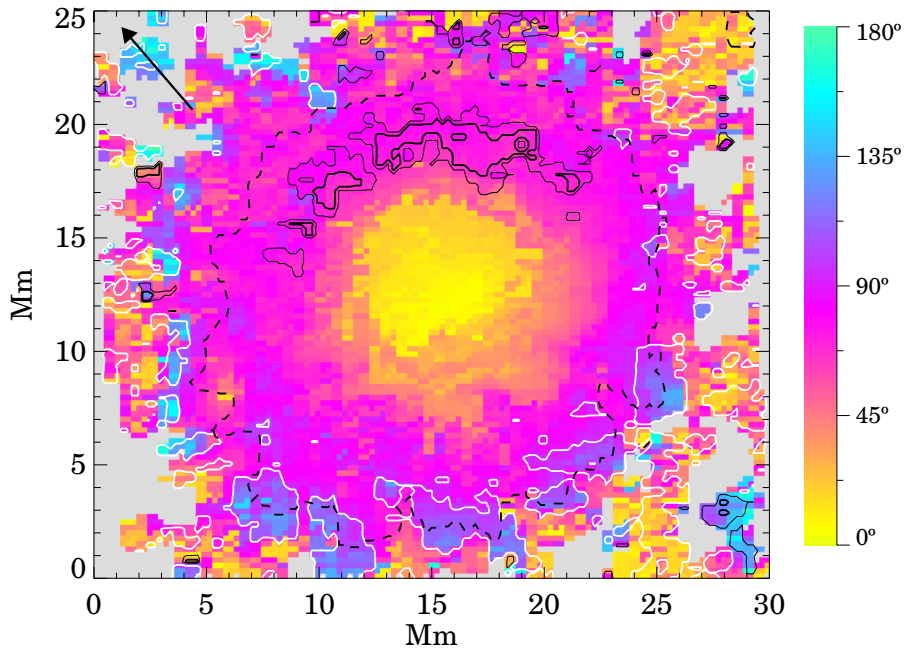


Figure 2. Map of the magnetic field inclination of a sunspot as measured from the vertical to the solar surface. The observations were obtained with the ASP in June 1992 and correspond to a spot located some 16° away from the disk center. Superimposed are contours of the largest line-of-sight velocities retrieved at the lowest optical layer observable (continuum optical depth unity): blueshifts are in dark lines (thick, -3 km s^{-1}; thin, -1.5 km s^{-1}), redshifts are in white lines (3 km s^{-1}). The arrow points to the solar disk center. The outer visible boundary of the penumbra is indicated with a dashed line. Adapted from *Nature* (Westendorp Plaza *et al* 1997).

obtained by Westendorp Plaza *et al* (1997) who, by using a novel analysis technique of the Stokes profiles, have been able to infer the variation with optical depth of the vector magnetic field and the line-of-sight velocity from an accurate spectropolarimetric map of a sunspot (see POLARIMETRY: SPECTROPOLARIMETRY). In this work, evidence has emerged that the Evershed effect (as well as the sunspot magnetic field) does not stop at the visible penumbral boundary in the lowest observable optical layer (in which the CONTINUOUS SPECTRUM is formed). Moreover, they found a strong correlation (almost all around the sunspot) between the velocities and the magnetic field inclination: blueshifts (upflows) are mostly located in the inner–middle penumbra where the magnetic field is pointing upwards, while redshifts (downflows) are mostly seen in outer zones, both inside and well outside the outer penumbral rim, where the vector magnetic field points downwards (see figure 2). A rough account of the mass flux inferred from their results yields an approximate mass conservation (within a factor 2 at maximum). Besides, their results at middle optical depths of the photosphere agree fairly well with those reported in former analyses of the problem which were unable to extract the information from the lower layers, codified in the observed Stokes profiles.

After all these modern observations a more coherent picture of the photospheric Evershed effect emerges as depicted in the cartoon of figure 3. Most of the mass flux

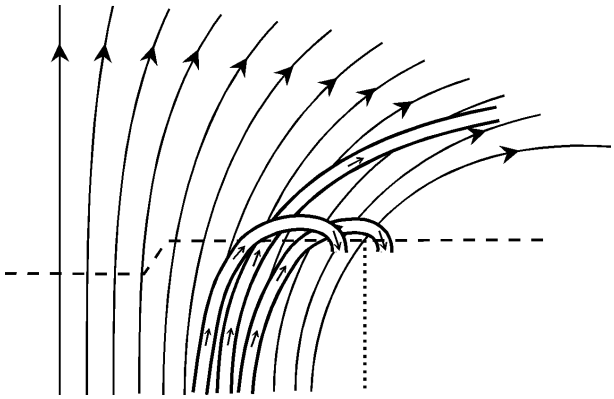


Figure 3. Cartoon (not to scale) of the (photospheric) penumbral structure with the Evershed flow channels (thick tubes with small arrows). The large-arrowed lines represent the magnetic field lines. The dashed line indicates the optical depth unity surface. The outer penumbral rim is marked with the dotted line. Adapted from *Nature* (Westendorp-Plaza *et al* 1997).

is carried through arched magnetic flux tubes that re-enter the surface at the lower layers of the outer sunspot border (either inside or outside it). Some of the mass—but a much less significant fraction—follows less arched flux tubes that extend through the superpenumbral canopy, whose outer footpoints are likely to be located farther beyond the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

spot. At even higher layers, not represented in the figure, the inverse (chromospheric) Evershed effect takes place.

The physical mechanism

In the mid-1960s two prominent physical mechanisms were proposed to explain the observed Evershed effect (see references in Thomas (1994) and Maltby (1997)), namely, the siphon flow hypothesis and the wave hypothesis. In the first model, the gas is driven through an arched magnetic flux tube in whose ending footpoints a pressure difference is established: the gas flows from the higher-pressure footpoint to the lower-pressure one. This pressure difference must be produced by the different values of the magnetic field strength at the two ends of the tube: the existence of a magnetic field introduces a magnetic pressure which is proportional to the square of the field strength; hence, the flow is driven from the lower to the higher field strengths. This provides a very simple and neat explanation for the inverse Evershed effect as is currently accepted (Thomas 1994, Maltby 1997): high-lying (chromospheric) tubes connect the umbra (with an intense magnetic field) to the photosphere or other flux tubes out of the sunspot that have a lower field strength and then the flow is directed inwards. The photospheric Evershed effect was not that clearly explained, however. Prior to the 1990s, it was difficult observationally to find flux tubes out of the sunspot with magnetic fields strong enough to consider them as the ending footpoints of an outwardly directed, photospheric siphon flow.

In the wave hypothesis, the characteristic Evershed line asymmetry would be caused by the (spatial and temporal) observational average of unresolved wave motions, i.e. there would be no true outward flows but an observational effect due to the limited resolution. Different types of waves have been proposed in attempts to fit the predictions to the whole observational evidence, but, ultimately, these models are not fully satisfactory since some of their ingredients (e.g. interface surfaces) or their predictions (e.g. horizontal wavelengths) are hardly realizables. Some other hypotheses, such as mixture between siphon flows and wave-pressure-driven flows or dynamic evolution of magnetic flux tubes in the penumbra, have very recently been proposed and need further check with observations.

Alongside wave-based models, the siphon flow hypothesis has evolved remarkably in sophistication over the last few years. Siphon flow models have recently been proposed by Thomas (1994) and Montesinos and Thomas (1997) whose predictions are in fair agreement with the observational results of Westendorp Plaza *et al* (1997) reported above, thus providing significant theoretical support for this mechanism to be the physically responsible one for the currently known facts of the Evershed effect, in both the photosphere and the chromosphere. Perhaps, apart from refining the models so that they are really able to fit the observations, the only (apparently) missing ingredient is the observational determination of magnetic fields at the outer footpoints with larger strengths than those present in the middle penumbra—where the inner footpoints are located. However, this is not a drawback in the siphon flow hypothesis since observational measurements are only able to infer values of the vector magnetic field (and of the other quantities) at given optical depths, so that the *observed* footpoints do not correspond to the same gravitational potential. Proper observational evaluations of the *geometrical* stratification of the sunspot photosphere are needed (and expected) in the near future to shed some more light on this exciting problem.

Bibliography

Evershed J 1909 Radial movement in Sun-spots *Mon. Not. R. Astron. Soc.* **69** 454–7

Maltby P 1997 Dynamics of sunspots *First Advances in Solar Physics Euroconference: Advances in the Physics of Sunspots (ASP Conf. Ser. 118)* ed B Schmieder, J C del Toro Iniesta and M Vázquez (San Francisco, CA: Astronomy Society of the Pacific) pp 91–110

Montesinos B and Thomas J H 1997 The Evershed effect in sunspots as a siphon flow along a magnetic flux tube *Nature* **390** 485–7

Thomas J H 1994 The cause of the Evershed in sunspots: flows or waves? *Solar Surface Magnetism (NATO ASI Ser. C 433)* ed R J Rutten and C J Schrijver (Dordrecht: Kluwer) pp 219–35

Westendorp Plaza C, del Toro Iniesta J C, Ruiz Cobo B, Martinez Pillet V, Lites B W and Skumanich A 1997 Evidence for a downward mass flux in the penumbral region of a sunspot *Nature* **389** 47–9

Jose Carlos del Toro Iniesta

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspots: Moving Magnetic Features and Moat Flow

Most mature SUNSPOTS and some large SUNSPOT PORES are surrounded, at least in part, by a zone which is free of stationary magnetic field (e.g. PLAGE, network) and which exhibits a radially outward horizontal mass flow. This zone is called a sunspot 'moat'. The most notable features in the moat are small magnetic structures of mixed polarity called moving magnetic features (MMFs). These are transient phenomena that can appear anywhere in the moat but most often near the outer edge of the sunspot. They move outward, usually at a higher velocity than the average moat flow, until they either fade away or reach the outer boundary of the moat.

Both sunspots and plages are also surrounded by zones that absorb acoustic power called 'acoustic moats'. These are not covered in this article.

Figure 1 shows an example of an isolated sunspot with a well-defined moat. This was a small, mature leader spot which was slowly decaying. The magnetogram on the right shows the spot's negative polarity as black and opposite magnetic polarity as white. A broken and irregular circle of the same polarity of the spot surrounds it at a distance of about 15″ (11 Mm). This is the outer boundary of the moat. The inner boundary is the edge of the SUNSPOT PENUMBRA. The MMFs are the small patches of magnetic field of both polarities within the moat annulus.

Early investigators suggested that a moat and MMFs were associated only with decaying sunspots. However, subsequent work showed that at least a partial moat is generally present even in growing sunspots. Mature, isolated spots are more likely to have a symmetric moat geometry than young, developing spots or those in complicated active regions. The width of the moat scales with the size of the sunspot.

Properties of MMFs

MMFs are best studied in movies with the highest possible spatial resolution. In addition to the magnetic signal, they are also visible as bright features in many photospheric lines, molecular bands such as the 3883 Å CN bandhead and the wings of the Ca II H and K lines. The features are larger for lines formed higher in the atmosphere indicating a spreading with height. Very small bright footpoints are sometimes visible in the continuum and, with much more contrast, in G band images. The latter are taken in a 12 Å band heavily populated by CH lines centered at 4508 Å. In these photospheric lines they look very much like small fragments of stationary magnetic field such as plage or enhanced network. However, in addition to their characteristic motions, MMFs differ by having little or no chromospheric or coronal emission associated with them. Many aspects of MMFs are near (or below) the limits of spatial resolution imposed by atmospheric seeing. Hence many details of how they form and evolve are still lacking.

There seem to be at least three classes of MMFs that have been identified in movies. The most prominent variety (type I for reference below) begin as small magnetic bipoles that erupt from below and then gradually separate. These can appear anywhere in the moat. A second type (II) consists of unipolar features that are shed from the edge of the sunspot. These have the same polarity as the spot and have been suggested as part of the sunspot decay mechanism. A third variety (type III) are higher-speed unipolar features with the opposite polarity to the sunspot that are associated with small, bright continuum points.

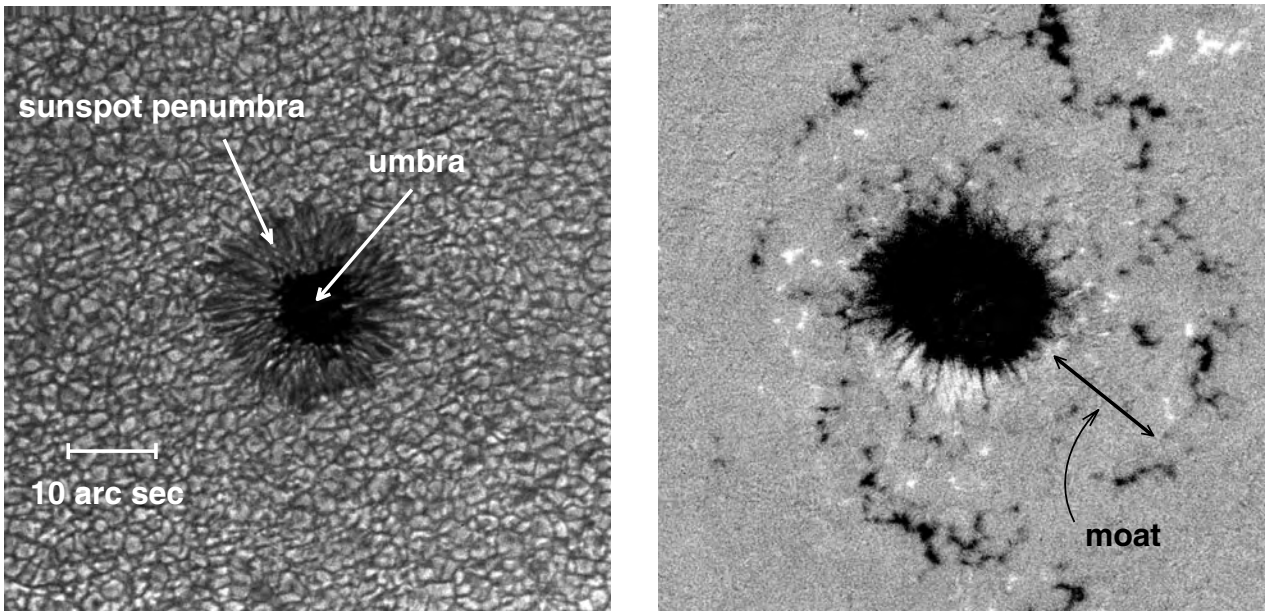
Type I MMFs

The type I MMFs are the best described. The early phase is visible as a small dark area (about 1000 km) in the upper photosphere as seen in broad-band ($\approx$ 3 Å) Ca II K line movies. With less contrast, this darkening is also visible in the continuum, showing dark fibrils characteristic of emerging magnetic field (see CHROMOSPHERE: FIBRILS). Over a period of about 10–20 min, the dark area expands and elongates into a spindle shape with a brightening appearing at each end. The latter are also easily visible in broad-band K line images and probably any line formed in the upper SOLAR PHOTOSPHERE or lower CHROMOSPHERE that shows bright plage. At this stage the new MMF is 2000–3000 km in size. The bright points are typically 500 km or larger as seen in the inner K line wings but may be only 200 km or smaller in continuum or G band images which are formed deeper. Coincident with the bright points are two areas of magnetic field with opposite vertical components forming a bipole. These are assumed to be the footpoints of a magnetic loop arching between them. Before this stage, longitudinal magnetograms show little or no signal, consistent with a more horizontal field. This development of a type I MMF is shown in a set of space–time images in figure 2(*b*). The spatial part of these is a line spanning the two bright points labeled b in figure 2(*a*). The initial dark patch is labeled 1, the inner footpoint 2 and the outer 3. The edge of the sunspot is on the right-hand side of these and has negative (black) polarity.

Type I MMFs are more likely to appear near the outer perimeter of the sunspot but, as already noted, can appear anywhere in the moat. The line between the magnetic footpoints is generally within 45° of the radial direction from the sunspot. Deviations may be more common further from the spot. When near the spot they frequently appear along a continuation of a dark penumbral filament. The inner footpoint has the same magnetic polarity as the sunspot. Measurements of the longitudial magnetic field range from about 100 to 300 G but are complicated by the small size of the features. The magnetic flux for each footpoint is typically 2×10^{18} Mx with reported values ranging from 10^{18} to 10^{20} Mx.

After appearing, the two footpoints separate at approximately 100 m s^{-1} while both move outward at about 500–1000 m s^{-1} in the moat. Nearby MMFs can move at different velocities and most move outward more quickly than the overall moat flow (described below). Hence they are not just carried along by the visible surface motion in the moat. Frequently the two poles are

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*) (*b*)

Figure 1. (*a*) A sunspot imaged in the continuum (6328 Å) showing the umbra and penumbral regions and the surrounding granulation. (*b*) A magnetogram with the moat between the sunspot and surrounding plage indicated. This zone extends all around the spot. The MMFs are the salt and pepper features in the moat. Black is negative field and white is positive. This example is more symmetric and regular than most. Both images were taken at the Swedish Vacuum Solar Telescope on La Palma on 29 August 1993.

not equally visible and the visibility of both against the background drops after the initial phase. MMFs move outward until they either fade from view or reach the outer edge of the moat. Sometimes an interaction at the edge is visible in the form of a small brightening and/or a magnetic cancelation of the leading point. However, details are largely unknown as is the fate of those that fade from view.

Much of the magnetic field in the outer penumbra of a sunspot is nearly horizontal and a substantial fraction of it seems to re-enter the photosphere at the boundary. If filaments of this field continue to extend horizontally outward somewhere below the visible surface, then type I MMFs can be understood as upward kinks or embolisms in one of these filaments. In this scenario, they represent a recycling of magnetic flux from the sunspot and do not contribute to a loss of flux or decay of the sunspot field. The motion of this 'kink' has been modeled as a soliton wave in the magnetic filament.

Type II MMFs

Type II MMFs are only formed at the edge of a sunspot. This is a very dynamic region with continual extensions and retractions of the tips of penumbral filaments and a strong collar outflow (described below). The type IIs appear here with the same magnetic polarity as the sunspot but without any apparent connection with a nearby point of opposite polarity as with the type Is. As such they seem to represent a loss of flux from the sunspot and may be related to sunspot decay. However, it is not known whether they are more frequent in rapidly decaying sunspots. After moving into the moat, they behave and appear like a separated inner half of a type I MMF with magnetic fields and fluxes within the range exhibited by type Is.

Type III MMFs

The type III MMFs are more transitory than the other two and exist only for several minutes near the edge of a sunspot. A type III begins as a small bright point that seems to be emitted from the outer edge of the penumbral and travels outward a few thousand km at 2000–3000 m s^{-1} before disappearing. They are generally smaller than the initial bright points in type Is but often brighter in the continuum. Together with their high speed, this makes them very eye-catching in movies and they were dubbed 'streakers' when first seen in continuum movies. The magnetic polarity is opposite that of the sunspot with a smaller total flux than the type Is and IIs. The bright point seen in the continuum and G band may lead the magnetic feature slightly. They generally seem to catch up with a slower-moving type II MMF (perhaps type Is as well) whereupon they vanish, sometimes with a local brightening. The latter continues to move outward in the moat, largely unaffected by the collision and apparent merger. Figure 2(*c*) shows a space–time image of an example. The small streaker is labeled 4 and the type II MMF that it combines with is labeled 5. The path is labeled c in figure 2(*a*). The edge of the sunspot is to the left and it has negative (black) polarity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

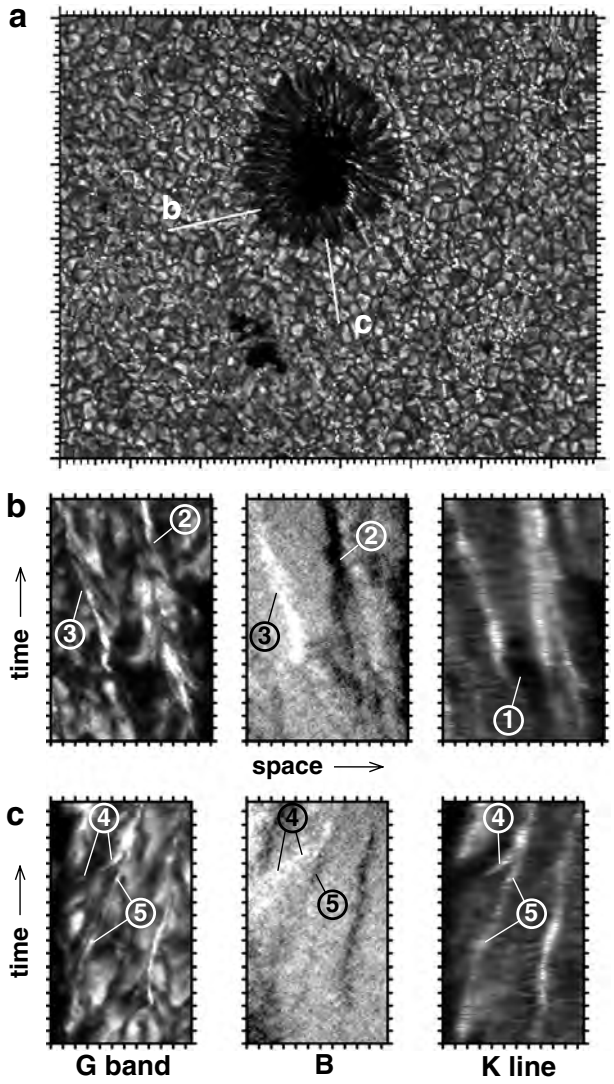


Figure 2. MMF examples. (*a*) A G band image of a sunspot showing the two paths used to construct the space–time images for (*b*) and (*c*). Tick marks are in arcsec. The left space-time images for (*b*) and (*c*) are G band intensity, the middle are magnetic field, and the right are K line intensity (3 Å wide filter). (*b*) shows a type I MMF and (*c*) shows a type III. The magnetic field is scaled from about −50 to +50 G for black and white respectively. The vertical ticks are 5 min intervals and the horizontal ticks are in arcsec. All data were taken at the Swedish Vacuum Solar Telescope on La Palma on 14 June 1994.

There may well be other types of MMFs and the *ad hoc* classification used here may not bear up as MMFs are studied in more detail. The type Is are similar to small magnetic flux eruptions seen in the interior of quiet Sun supergranules but are much more frequent in the moat (see SOLAR PHOTOSPHERE: SUPERGRANULATION).

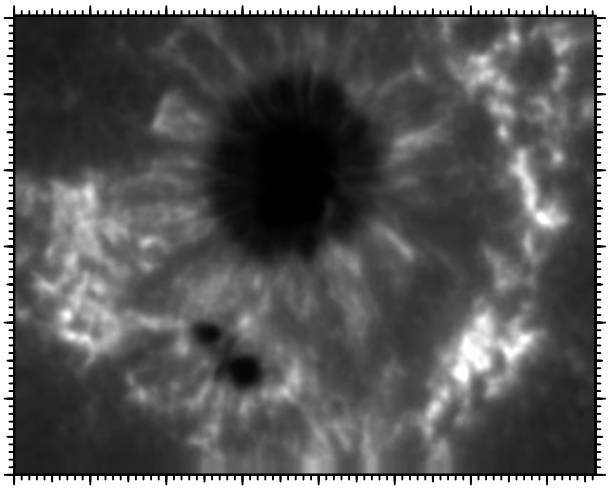


Figure 3. A 4 h average of K line images (3 Å wide filter) for the same sunspot shown in figure 2(*a*). The bright rays in the moat track the preferred paths of bright points associated with MMFs. Tick marks are in arcsec.

Moat flows

The outer radius of the moat is typically about twice that of the associated sunspot. For a large sunspot this may reach 50 Mm or more. There may be a minimum moat radius of 8 Mm for very small spots, about the size of an average supergranule. Doppler and local correlation tracking (LCT) measurements indicate an average outward flow of about 400–600 m s^{-1} with no dependence on spot size.

Just outside the sunspot penumbra, there is also a narrower region, sometimes called a 'collar' flow, with higher outward velocities than the rest of the moat. The width of this collar is about 2000–3000 km with velocities typically 1000 m s^{-1}, about twice that of the middle moat. This blends into the moat flow where the latter exists. The collar flow may extend all around the spot, however, even where there is not a larger-scale moat flow. It is distinct from the Evershed effect (see SUNSPOTS: EVERSHED EFFECT) seen in the sunspot penumbra although the two may be somehow related. The latter is seen in the dark filaments of the penumbra and has a higher velocity than the collar flow which is a motion of granulation just outside the spot.

The LCT method has been used to map details of the horizontal motions in sunspots and the surrounding moat. When applied to continuum images, it tracks the granulation pattern without any significant contribution from MMFs. On time scales of an hour or less, the pattern is dominated by motions on a mesogranular scale (4000–7000 km) just as in the quiet Sun. These can locally reduce or cancel the outward motion for a while. Longer averages show persistent azimuthal patterns in addition to the outward flow. There are rays of alternating azimuthal convergence and divergence that extend from the outer edge of the sunspot to the moat boundary. These seem to cause MMFs (and their associated bright points) to be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

channeled along preferred horizontal rays as they move outward. Figure 3 illustrates this effect with a time-averaged image taken with a 3 Å Ca II K line filter. The bright rays emanating from the sunspot are caused by several MMFs traveling along each ray during the 4 h average.

The sunspot moat flow represents an organized horizontal flow pattern that can be much larger than a large supergranule (radius of 50 Mm versus 15 Mm). It is thought that this large-scale flow is driven by a temperature excess beneath the sunspot. The magnetic field in the sunspot inhibits the normal vertical convective energy transport, causing the spot to be dark and a buildup of excess heat in deeper layers. Horizontal transport in the form of the outflow then carries part of this excess heat outward although there is evidence that a large portion of it may be stored for at least several days. It has also been suggested that the faster motion of MMFs may represent coupling to faster subsurface flows. A completing theory for these faster motions is that MMFs represent traveling soliton waves in magnetic filaments that are otherwise below the visible surface.

Current research

Since their discovery 30 yr ago, MMFs have remained a difficult subject for observations, requiring high spatial resolution movies for at least several hours. The theory is still very rudimentary and it may well be that none of the current concepts is really correct. Detailed modeling will depend on better observations to better constrain the possibilities. Recent and future observations from the Solar Heliospheric Observatory (SOHO) and the Transition Region and Coronal Explorer (TRACE) satellites will provide material for studies in dozens of sunspots. These are only medium ($\approx$ 1″) resolution but provide uniform data sets unaffected by atmospheric seeing for several days. They will finally allow meaningful statistics to be assembled on MMFs and moat flows and their relation to sunspot evolution. Higher-resolution studies will be required to study the fine structures in MMFs and how they move within the moat faster then the general moat flow. Particularly needed are vector magnetograms to understand their detailed magnetic structure and how MMFs near the sunspot are formed. Future telescopes in space and on the ground will eventually provide these observations and the next 10 yr should see much progress in our understanding of these features.

Current and future work on three-dimensional maps of the temperatures and velocities around and underneath sunspots using helioseismology looks very promising. It should eventually provide details of the heat trapping and convective motions associated with sunspots and a clearer picture of the role of the moat flow.

Bibliography

The outward motions of bright points around sunspots were first noted by Sheeley in CN 3883 Å spectroheliograms and Vrabec discovered 'streaming flux knots'. The first systematic study of MMFs and sunspot moats is that of Harvey and Harvey in 1973. They introduced the term 'moving magnetic feature' and the MMF acronym. The earlier discovery papers are cited therein, some of which are proceedings from meetings that may be difficult to find. The term sunspot 'moat' was suggested by Dr Sheeley. Brickhouse and LaBonte reviewed moat observations up to 1985 and this remains the most thorough study published to date. A recent theory to explain type I MMFs is given in Ryutova, *et al* which also quotes earlier work and summarizes known characteristics of MMFs.

Brickhouse N S and Labonte B J 1988 Mass and energy flow near sunspots. I—observations of moat properties *Sol. Phy.* **115** 43–60

Harvey, J and Harvey, K 1973 Observations of moving magnetic features near sunspots *Sol. Phys.* **28** 61–71

Lee J W 1992 Observational evidence for various models of moving magnetic features *Sol. Phys.* **139** 267–73

Ryutova M, Shine R, Title A and Sakai J I 1998 A possible mechanism for the origin of emerging flux in the sunspot moat *Astrophys. J.* **492** 402

Sheeley N R Jr 1969 *Sol. Phys.* **9** 347

Richard Shine and Alan Title

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunspots: Wilson Effect

The Wilson effect refers to the depressed appearance of SUNSPOTS when positioned close to the solar limb. The impression is that sunspots are cavities in the SOLAR PHOTOSPHERE. The reason is that the radiation we observe is coming from deeper layers in the sunspot than in the surrounding photosphere. The detection of this depression by Alexander Wilson dates back to 1769. The phenomenon is explained, first without mathematical expressions and then in greater detail.

Detection

The phenomenon was detected by ALEXANDER WILSON, the first Professor of Astronomy in Glasgow, in 1769 and published in *Philosophical Transactions of the Royal Society* in 1774. Figure 1 is a reproduction of the essential part of one figure in his publication. Wilson started his observations of a sunspot near the western limb on 22 November 1769. He noted that the sunspot was very large and the extent of the penumbra, i.e. the medium bright region between the dark umbra and the photosphere, was equally large on the side closest to the solar limb and on the side closest to the center of the solar disk. The next day, on 23 November the sunspot was close to the solar limb (see figure 1, position 1), and Wilson noted that the shape of the penumbra on the side closest to the center of the solar disk had changed since his previous observation. The size of this side of the penumbra had decreased, whereas the penumbra closest to the solar limb was nearly unaltered. The solar rotation moved the sunspot to a location even closer to the western limb of the Sun on 24 November (see figure 1, position 2). Wilson observed that the penumbra closest to the disk center had disappeared.

In his 1774 publication Wilson discusses two possible explanations, either there was a real change in the shape of the sunspot or the sunspot was 'a vast EXCAVATION in the luminous matter of the sun'. Since digging in the solar surface sounds unrealistic we will in the following replace the word excavation by cavity or depression. Wilson knew that he could discriminate between the two possibilities if he could observe the same sunspot close to the eastern limb. Fortunately the sunspot was long-lived and reappeared with nearly the same size on the eastern limb. Wilson observed the sunspot on 11 and 12 December (see figure 1, positions 3 and 4). He saw that the shape of the penumbra closest to Sun's center was undergoing changes. However, the side of the penumbra that was changing as a function of time close to the eastern limb was the same part of the sunspot that was not affected when the sunspot was close to the western limb. Hence, he concluded that sunspots are cavities in the solar photosphere.

Explanation

In the literature on sunspots one often meets the expression sunspot darkness. It is important to realize that a considerable amount of radiation is emitted from the sunspot and that the expression sunspot darkness reflects the observation that the radiation emitted by the sunspot is less than that of the surrounding photosphere. The radiation coming from the penumbra exceeds that coming from the umbra, but it is less than the photospheric radiation. This is illustrated in figure 2 by a darker grey color in the umbra than in the penumbra (see also SUNSPOT PENUMBRA; SUNSPOT UMBRA).

Consider a circular sunspot consisting of an umbra shaped like a cylinder, surrounded by a penumbra shaped like a cylindrical sheet. The intensity levels across the umbra and the penumbra are assumed to be constants. The symmetry axis of this hypothetical sunspot is perpendicular to the solar surface. Figure 2 shows a sketch of the sunspot in a plane that contains the symmetry axis of the sunspot and the line-of-sight for an observer on Earth. Observations show small variations in intensity both within the umbra and the penumbra, but these variations are neglected here. In figure 2 the sunspot is located close to the solar limb, such that an observer on the Earth sees the sunspot under a large aspect angle. It should be pointed out that in figure 2 we have drawn the spherical surface as a plane layer, neglecting that the directions from different locations inside the sunspot and its surroundings to the center of the Sun are not parallel.

Since the Sun is a sphere of hot gas the solar surface is not solid. Therefore the important question is where the radiation we observe comes from. The intensity of the observed radiation is determined by the emission and absorption along the line of sight. Hence, several adjacent layers within a certain height interval, or distance from the Sun's center contribute to the observed radiation. Within this height interval the temperature decreases with increasing height. This implies that the emission decreases with height in this height interval, but the probability for the photons to escape and reach the observer without being absorbed is increased.

The solar radiation in visible light originates within a height interval of approximately 500 km (0.5 Mm). Since the solar radius is 696 000 km (696 Mm) this layer may be regarded as shallow compared with a solar radius. In figure 2 the picture is simplified by letting the height interval where the visible light originates be illustrated by the width of a single line dividing the solar interior from the exterior.

Since the visible light originates within a height interval determined by the emission and absorption coefficients, this interval may be centered at different heights in the umbra, penumbra and photosphere. The main ingredients of the gas, hydrogen and helium, have a low degree of ionization in the layers that contribute to the radiation. The main absorber is the negative hydrogen ion, formed by the affinity of free electrons to neutral hydrogen atoms. The free electrons are supplied by minor constituents of the gas, such as iron, with a lower ionization potential than hydrogen. With increasing temperature, the number density of free electrons increases, the density of negative hydrogen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

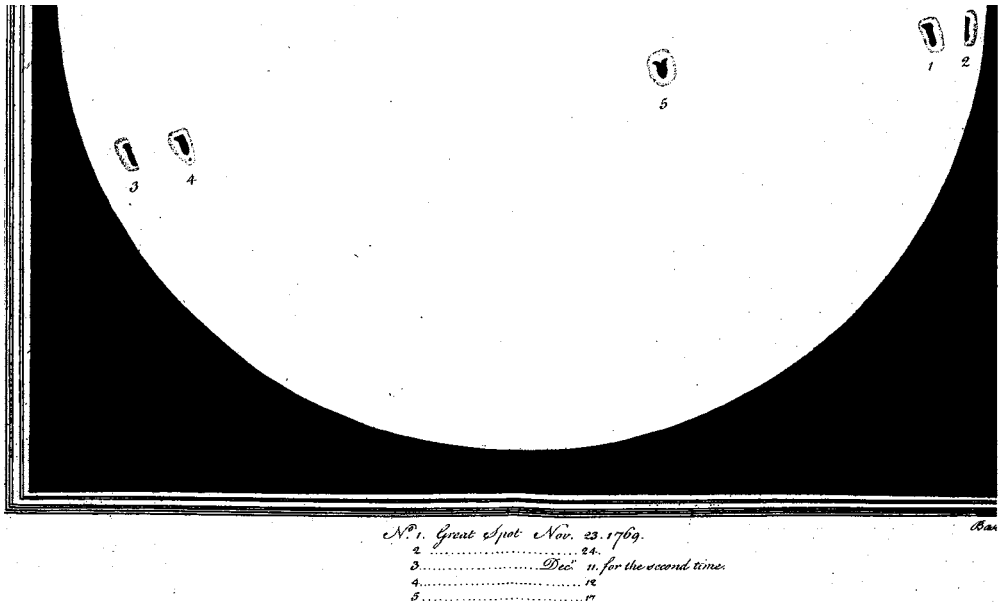


Figure 1. Copy of Alexander Wilson's drawing of the large sunspot he observed in 1769.

ions increases, the absorption coefficient increases, and the corresponding solar feature becomes less transparent. We know that the temperature of the penumbra is higher than the umbral temperature, but lower than the photospheric temperature. This implies that the umbra is more transparent than the penumbra and the penumbra is slightly more transparent than the surrounding photosphere. In order words, we 'see' to deeper layers in the umbra than in the penumbra and to slightly deeper layers in the penumbra than in the photosphere. In a more stringent discussion the possible differences in density between the umbra, penumbra, and photosphere must be taken into account. It follows that the umbral radiation originates at a lower height.

Recall that in figure 2 the dividing line between the solar interior and exterior is taken to represent the height interval where the visible radiation originates. To illustrate that we observe radiation coming from different heights figure 2 shows that the radiating layer is located higher in the photosphere than in the penumbra and considerably higher than in the umbra. For illustrative purposes the height differences between the umbra, penumbra and photosphere are exaggerated. The height difference between the umbra and the photosphere, called the Wilson depression, is of the order of 500 km. Hence, it is comparable with the height interval that contributes to the visible radiation.

With these height differences in mind consider an observation of a sunspot located close to the solar limb. Since the penumbral radiation originates at a lower height than the photospheric radiation, the photosphere will obscure part of the penumbra on the side of the sunspot

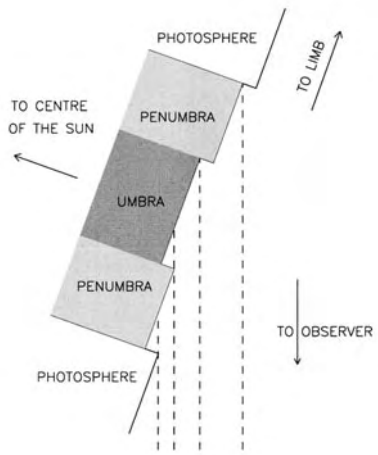


Figure 2. Sketch of a circularly shaped sunspot located close to the solar limb.

remote from the limb. In figure 2 this part of the sunspot is located in the lower half of the sketch. The penumbra closest to the limb, on the other hand, will not be obscured (see figure 2). Since the umbral radiation originates in layers located deeper than the layers contributing to the penumbral radiation, part of the umbra will be obscured by the penumbra. Hence, the size of the umbra will decrease (increase) faster than expected from perspective foreshortening as the sunspot is moved by solar rotation towards (away from) the solar limb. This simple model describes the main characteristics of the Wilson effect.

Most sunspots change their shape from one day to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the next, and the fraction of sunspots with a lifetime comparable with the sunspot shown in figure 1 is minute. These factors explain why it is difficult to measure accurately the Wilson depression. An interesting way of estimating the magnitude of the Wilson depression is based on measurement of solar rotation (see Balthasar and Wöhl 1983; SOLAR INTERIOR: ROTATION). Consider the determination of solar rotation from (a) sunspot positions during their passage over the solar disk and (b) the time difference between two successive passages of the same sunspot across the same position on the solar disk. The latter method is independent of the Wilson depression, whereas the first method depends on its value. The reason for this dependence may be understood by comparing the apparent positions of two sunspots, one with and one without Wilson depression. The sunspot without Wilson depression will be at the same height as the photosphere and located closer to the limb than a sunspot with depression. On the basis of such observations values between 500 and 1000 km have been found for the Wilson depression.

Towards a model calculation

It was suggested by Bray and Loughhead (1964) that a sufficient condition for a SUNSPOT MODEL to show a Wilson effect is that the transparency increases from the photosphere through the penumbra to the umbra. However, Jensen and Maltby (1965) showed that the most transparent umbral models lead to marked intensity peaks on the limb side of the umbra. Such peaks are never observed. Jensen *et al* (1969) neglected the pressure and density variations across both the umbra and the penumbra and approximated a sunspot by placing models of the umbra, penumbra and photosphere side by side. Furthermore, they assumed that each model could be determined from the hydrostatic equation $\nabla P = -\rho g$, where P is the gas pressure, ρ is the density and g the gravitational acceleration. To be able to compare the results with observations they constructed a series of models and calculated the specific intensity, I_λ, for different values of the wavelength, λ, and the aspect angle, θ, by the expression

$$I_\lambda = \int_0^\infty S_\lambda(t_\lambda)\, \mathrm{e}^{-t_\lambda}\, \mathrm{d}t_\lambda$$

where the source function, $S_\lambda(t_\lambda)$, was approximated by the Planck function. Special care was taken in calculating the optical length, t_λ, along the ray making an angle θ with the surface normal. The optical length differs from the optical depth divided by the cosine of the aspect angle when the ray passes from, say, the photosphere into the penumbra. They found that a Wilson depression of 400 km could satisfy the available observations.

Observations show that sunspots have strong magnetic fields. The darkening of the sunspot depends on this field, as may be illustrated as follows. Consider a time-independent case and write the momentum equation as

$$\rho \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\nabla P - \rho \boldsymbol{g} + \boldsymbol{j} \times \boldsymbol{B}$$

where ρ is the density, $\boldsymbol{v}$ is the velocity, P the gas pressure, $\boldsymbol{g}$ the gravitational acceleration and $\boldsymbol{B}$ the magnetic induction. Neglecting the displacement current, the current density $\boldsymbol{j} = \frac{1}{\mu_0} \nabla \times \boldsymbol{B}$, where μ_0 is the permeability.

In layers close to the solar surface most velocities are subsonic and the darkening of a sunspot is mainly determined by the magnetic field (see SUNSPOT MAGNETIC FIELDS). In the simple case where the effect of the velocity field is neglected and the magnetic field is axisymmetrical without an azimuthal component, the momentum equation in cylindrical coordinates reads

$$\frac{\partial P}{\partial r} = \frac{B_z}{\mu}\left(\frac{\partial B_r}{\partial z} - \frac{\partial B_z}{\partial r}\right)$$

$$\frac{\partial P}{\partial z} = -\rho g - \frac{B_r}{\mu}\left(\frac{\partial B_r}{\partial z} - \frac{\partial B_z}{\partial r}\right).$$

Integration of the first equation from the center ($r = 0$) of the sunspot to the photosphere ($r = a$) gives the difference in gas pressure

$$\begin{aligned} P(r=a) - P(r=0) &= \frac{1}{2\mu} {B_z}^2(r=0) + \frac{1}{\mu}\int_0^a B_z \frac{\partial B_r}{\partial z}\, \mathrm{d}r \\ &= P_M + P_T. \end{aligned}$$

Thus, the lower gas pressure in the sunspot than in the surroundings is caused by the magnetic pressure, P_M, and the tension, P_T, in the magnetic field. Close to the surface the magnetic field fans out and the tension must be taken into account. With the help of the equation of state for an ideal gas we may proceed to deduce the ratio between the temperature $T(r = 0)$ in the center of the sunspot to the photospheric temperature, $T(r = a)$, at the same height z

$$\frac{T(r=0)}{T(r=a)} = \frac{m(r=0)}{m(r=a)} \frac{1 - \frac{1}{P(r=a)}(P_M + P_T)}{1 + \frac{H(r=a)}{P(r=a)} \frac{\partial}{\partial z}(P_M + P_T)}$$

where the average molecular weight is $m(r = 0)$ in the center of the sunspot and $m(r = a)$ in the photosphere. $H(r = a)$ is the pressure scale height in the photosphere. For a homogeneous magnetic field the expression reduces to $T(r = 0)/T(r = a) = 1 - B^2/(2\mu P(r = a)) = 1 - P_M/P(r = a)$. Following similar arguments magnetostatic models for the whole sunspot have been presented and recently the velocity field has been included in the calculation with the aim of developing a dynamical model for the penumbral fine structure and the flow field in sunspots. Hitherto these models have not been extended to include calculations of the emitted radiation across the sunspot.

Bibliography

Balthasar H and Wöhl H 1983 *Solar Phys.* **88** 71–5

Bray R J and Loughhead R E 1964 *Sunspots* (London: Chapman and Hall)

Jensen E, Brahde R and Ofstad P 1969 The Wilson effect and the transparency of sunspot models *Solar Phys.* **9** 397–419

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Jensen E and Maltby P 1965 Opacity properties of sunspots
Astrophys. Norveg. **10** 17–25

Per Maltby

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sunyaev–Zeldovich Effect

The Sunyaev–Zeldovich (SZ) effect is a spectral effect on the COSMIC MICROWAVE BACKGROUND (CMB) caused by interaction with the electrons in the hot gas inside a CLUSTER OF GALAXIES. In the Rayleigh–Jeans portion of the spectrum it manifests itself as a small ($\sim$0.5 μK) dip in the temperature of the CMB in the direction of the cluster. First proposed in 1970, about 20 yr of observational effort were required before it was reliably detected. Many tens of clusters have now detected in the SZ effect, and early predictions of its usefulness for cosmology are now starting to bear fruit.

The basic mechanism of the SZ effect (see figure 1) involves scattering of CMB photons by electrons in the distributed intracluster gas that exists in clusters of galaxies and in fact makes up about a quarter of their total mass. A rich cluster may have a total mass $>3 \times 10^{14} M_\odot$ and gas in hydrostatic equilibrium in the potential well of the cluster heats to temperatures of order 10^7–10^8 K. The hot electrons in this plasma have two effects: firstly they produce thermal bremsstrahlung radiation, which provides the cluster x-ray luminosity. Secondly, they interact with CMB photons passing through the cluster, via Thomson scattering. This preferentially increases the energy of the photons involved in the scattering, giving an effective shift of the CMB BLACK-BODY RADIATION curve to higher frequencies and hence a dip in the Rayleigh–Jeans region of the spectrum, and an increment in the Wien region. (The precise form of the temperature shift as a function of frequency is shown in figure 2.) The magnitude of the effect is proportional to the line integral of electron pressure through the cluster, and hence directly proportional to electron number density, n_e, and temperature. This is in contrast to the thermal bremsstrahlung responsible for the x-ray emission, which is proportional to n_e^2. This difference makes the SZ effect more sensitive to the broadly distributed component of gas, rather than the dense central regions favored in x-ray emission. A further, very useful, aspect of the SZ effect is that for the same physical gas parameters it is independent of REDSHIFT. This makes it as easy to detect at high redshift, where optical or x-ray detection of a cluster may be very difficult, as at low redshift, thus making it a useful cosmological probe. In addition, the combination of x-ray and SZ data on a cluster makes possible a determination of the HUBBLE CONSTANT independent of the usual distance ladder techniques. The results for H_0 from this method are already of interest and are expected to be of increasing importance in the future, as both the systematic and the random errors in the technique start to decrease.

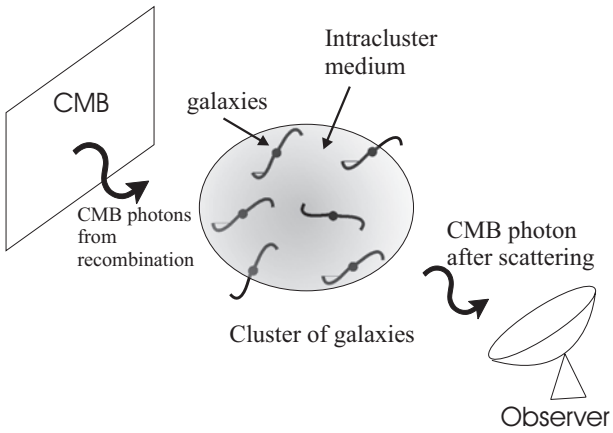


Figure 1. Schematic showing the origin of the SZ effect. Hot intracluster gas in a cluster of galaxies gives a small increase in the net energy of CMB photons arriving from the early universe.

The physics of the Sunyaev–Zeldovich effect

Probably the best way to derive the SZ effect, including relativistic corrections, is in the context of an extended version of the *Kompaneets equation*—this is discussed further below. We first begin with a heuristic derivation: consider the injection of energy, via Compton scattering, from an electron gas into an isotropic photon gas. We shall suppose that the electron gas is non-relativistic and with no bulk motions and that the radiation temperature $T_{\rm rad}$ is $\ll$ the electron temperature T_e. (This is well satisfied in practice.) With these conditions it is not immediately obvious that there should be any effect—if the scattering is elastic, then one cannot change the basically isotropic distribution of photons or electrons. It is true there is a 'Doppler scattering' effect

$$\frac{\Delta\nu}{\nu} = \frac{v_e}{c} \times \text{angular factor}$$

where v_e is the velocity of the colliding electron, but this will average to zero, since the electron velocities are isotropic. (Note that this formula will tell us what happens when there *is* a bulk motion.) The SZ effect arises as follows. The scattering mechanism is Thomson scattering, for which we know

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \sigma_T \frac{3}{16\pi}(1+\cos^2\theta)$$

where θ is the angle between the photon and electron directions, i.e. there will be maximum scattering in the forward and backward directions. We therefore approximate the problem as one-dimensional and consider in the cosmological (photon gas) frame the head-on collision of a photon with energy $h\nu$ with an electron with a velocity $v = \beta c$ directed away from the incoming photon. In the electron rest frame the photon energy is transformed to $h\nu[(1-\beta)/(1+\beta)]^{1/2}$ and there is an equal scattering probability in the forward and back directions. In these directions there is no frequency shift, from the familiar Compton formula

$$\lambda' - \lambda = 2\lambda_e \sin^2(\theta/2) \qquad \lambda_e = \frac{h}{m_e c}.$$

Transforming back to the cosmological frame we find that the photon has energy $h\nu(1-\beta)/(1+\beta)$ (i.e. *red*shifted)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

if its direction has been reversed in the encounter, or $h\nu$ if not. Now suppose that the photon came from the other direction, so that initially in the photon's frame the electron velocity was directed towards it. In this case the same analysis yields a *blue*shift to an energy of $h\nu(1+\beta)/(1-\beta)$ if the photon direction is reversed and a constant energy of $h\nu$ if not. The mean frequency after collision for photons with momentum parallel to electron velocity is thus

$$\nu' = \nu\frac{1}{4}\left(2+\frac{1-\beta}{1+\beta}+\frac{1+\beta}{1-\beta}\right) = \frac{\nu}{1-\beta^2} \approx \nu(1+\beta^2).$$

Thus there is a net second-order effect.

What happens in the interval of time Δt? Restoring the three-dimensional dependence on scattering angle, the mean shift is

$$\langle \text{probability of shift in} \Delta t \times (\beta\cos\theta)^2\nu\rangle$$

averaged over the isotropic photon momentum distribution, i.e.

$$\mathrm{d}\nu = \frac{1}{3}n_e\sigma_T c\Delta t\nu\frac{3kT_e}{m_ec^2}.$$

(Here $(n_e\sigma_T)^{-1}$ is the photon mean free path.) Thus

$$\frac{1}{\nu}\frac{\mathrm{d}\nu}{\mathrm{d}t} = \left(\sigma_T c n_e\frac{kT_e}{m_ec^2}\right).$$

The right-hand side has dimensions of (time)$^{-1}$ and we write it as $\mathrm{d}y/\mathrm{d}t$. Integrating, we obtain

$$y = \int_{t_i}^{t_f}\sigma_T c n_e\frac{kT_e}{m_ec^2}\,\mathrm{d}t.$$

The quantity y is called the Comptonization parameter and here is worked out for the case of injection of energy from an initial time t_i to a final time t_f. Equating this to the time taken for a photon to traverse a cluster yields the final formula

$$y = \int \sigma_T n_e\frac{kT_e}{m_ec^2}\,\mathrm{d}l$$

or $y = (\sigma_T k/m_ec^2) \times$ line integral of pressure through cluster. The equation

$$\frac{1}{\nu}\frac{\mathrm{d}\nu}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}t} \qquad \text{implies} \qquad \nu(t) = \nu(0)\mathrm{e}^y.$$

In the cluster case we have $\nu_{\text{out}} = \nu_{\text{in}}\mathrm{e}^y$. This shift in frequency yields an apparent *deficit* in the Rayleigh–Jeans region of the CMB spectrum and an increase in the Wien region (see figure 2). We have

$$\begin{aligned}\Delta I(\nu) &= I\left(\frac{\nu}{\mathrm{e}^y}\right) - I(\nu) \qquad \text{generally}\\ &= 2\left(\frac{\nu}{\mathrm{e}^y}\right)^2\frac{kT_{\text{rad}}}{c^2} - \frac{2\nu^2kT_{\text{rad}}}{c^2}\\ &\qquad \text{in the Rayleigh–Jeans region}\\ &\approx I(\nu)\times(-2y) \qquad \text{for small } y.\end{aligned}$$

For a rich cluster, taking the parameters of the COMA CLUSTER as an example, $l \sim 3$ Mpc, $n_e \sim 10^3$ m^{-3} and $T_e \sim 7\times10^7$ K. Thus a typical y is 7×10^{-5}, making the assumption of its smallness in general well justified.

The result just quoted for the Rayleigh–Jeans region is the correct one (neglecting relativistic corrections), as is the general idea of a decrement in the Rayleigh–Jeans region and an excess in the Wien region. However, our derivation has ignored several important physical factors, in particular that the photon energy $h\nu$ varies after collision in the electron frame, that the photons are not isotropic in the electron frame and that photons are bosons, the latter implying that there is stimulated emission also to consider. Putting all this together to find the correct equation yields the Kompaneets equation. The general solution as a function of frequency, expressed in terms of Rayleigh–Jeans brightness temperature ($T_{\text{RJ}} = (\lambda^2/2k)I$), is

$$\Delta T_{\text{RJ}} = \left(\frac{x}{\mathrm{e}^x-1}\right)^2 e^x\left(x\coth\frac{x}{2}-4\right)yT_{\text{rad}}$$

where y is as above and $x = h\nu/kT_{\text{rad}}$. This is what a typical heterodyne system would measure. If we were to measure *flux* rather than temperature, i.e. use a bolometer-type system, then

$$\Delta I \propto \nu^2(\Delta T_{\text{RJ}})_{\text{equiv}} \propto x^2 \times \text{above}.$$

The importance of this is that the effect is thus much increased in the sub-mm region, where we are likely to be using bolometer techniques anyway. The relevant curves are shown in figure 2, which also shows the same curves for a simple change in the black-body temperature of the CMB, such as would be relevant to primordial perturbations. One can see from figure 2 that observing the flux effect in three regions, centered at approximately 2 mm, 1.2 mm and 0.8 mm say, would give a unique signature for the SZ effect, separating it quite distinctly from any anisotropies due to primordial perturbations. Pursuing this point of different frequency dependences further, we note that if a cluster has a systemic *radial* velocity then the bulk electron flow induces Doppler scattering with magnitude given by the bulk motion formula above:

$$\frac{\Delta T}{T} \sim \frac{v_{\text{pec}}}{c}\tau$$

where τ is the optical depth to Thomson scattering through the cluster, $\tau \sim \int n_e\sigma_T\,\mathrm{d}l$. The frequency dependence of this effect will be that of the black-body curves of figure 2, and so by observing at the peak of the black-body curve one can separate the direct 'thermal' SZ effect, which nulls here, from the kinetic or 'Doppler' one, which has a peak here (when measured in terms of flux). To see what the hopes are of being able to map the peculiar velocity field of clusters in this way, we take the τ appropriate to the Coma cluster, ∼1%, and assume a cluster velocity of 600 km s^{-1}, which may not be untypical of the higher velocities which could be expected. This yields a black-body temperature change of 0.05 mK. For comparison,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

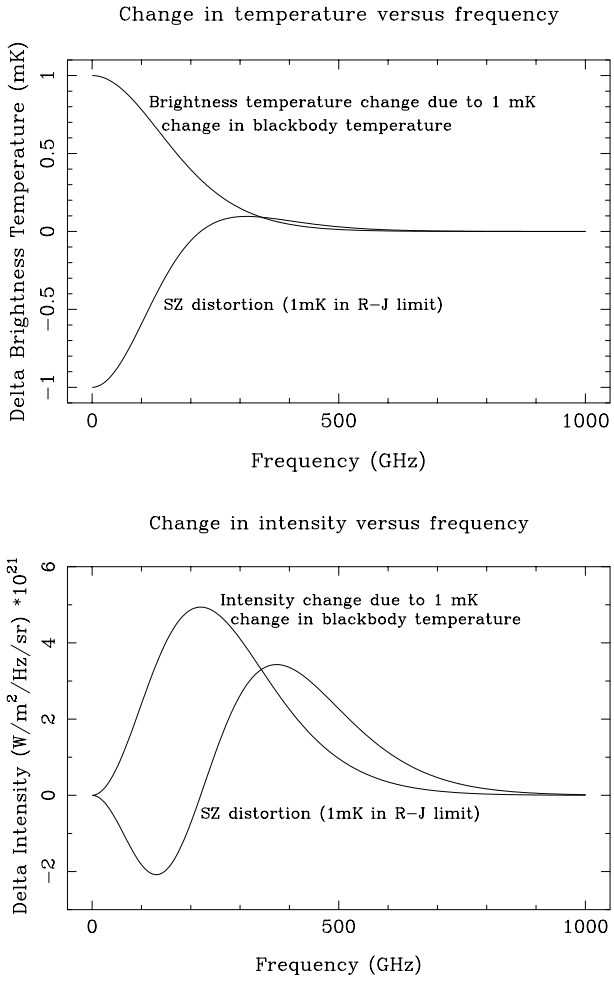


Figure 2. Top: Plots of ΔT (brightness temperature) for a 1 mK change in black-body temperature (top curve) and due to the SZ effect with a Rayleigh–Jeans limit of 1 mK. Bottom: Same but for intensity distortion.

taking the Coma cluster as an example again, the direct SZ effect expected is 0.5 mK, and indeed the measured values in several clusters appear to be of this order. Thus the mapping of the peculiar velocity field of clusters will be a difficult task, but the numbers do not look hopeless. Transverse motions of clusters also give a perturbation to the CMB via the 'moving lens effect', with a distinctive polarization signature, but for similar velocities the $\Delta T/T$ effect tends to be a few $\times 10^{-6}$ rather than a few $\times 10^{-5}$ as for the thermal and radial kinetic effects. The angular scale of the SZ effect on the sky is set by the angular diameter subtended by the cluster, and is typically a few arcminutes.

Relativistic effects

For clusters with electron temperature above about 10 keV, relativistic effects neglected in the usual Kompaneets equation must be introduced, if one wishes to match theory and experiment together properly, even at the accuracy level of the current observations, and certainly for future observations. First calculations of the corrections were made numerically by Wright and Raphaeli and recently Challinor and Lasenby have found an analytical extension of the Kompaneets approach, which allows results to be derived as power series in the electron temperature parameter $\theta_e = k_B T_e/m_e c^2$. As an example, in the Rayleigh–Jeans region, one finds

$$\frac{\Delta T_{\rm RJ}}{T_{\rm rad}} \approx -2y\left(1-\frac{17}{10}\theta_e\right)$$

to first order in θ_e. To give an example of the magnitude of this effect, for typical cluster parameters ignoring this correction would give an overestimation by about 5% in the value of the Hubble constant derived via the SZ route (described below). For clusters with temperature greater than about 10 keV there are substantial modifications to the frequency dependence of the SZ effect in the Wien region of the spectrum (by tens of per cent even at 15 keV) and the correct relativistic formulae (including θ_e^2 corrections) must be used.

Observations

A variety of approaches have been used to try to detect and measure the SZ effect. Here we highlight the techniques that have been used, roughly in order of their historical appearance.

Single-dish observations

The effect was first claimed to have been discovered by Parijskij, in 1972, in the Coma cluster. This was unlikely given the frequency of observation and the fact that Coma has a strong cluster halo source, in addition to other radio sources near the center. There were further claims in 1978 and 1981 for a detection in several clusters by Birkinshaw, Gull and Northover, using a dish at Chilbolton, UK, at 10 GHz, but again radio source confusion and possible systematic effects in the observations make it unlikely the the decrements seen (at about 1 mK amplitude) were real. The first reports of a detection that are widely believed, and seem to have withstood the test of further accumulation of data, came from Birkinshaw, Gull and Hardebeck in 1984, using the 40 m Owens Valley dish of Caltech, at 20 GHz. It might be wondered why there is any difficulty or question about these measurements, given that the effects are about 10 times bigger than a 95% confidence blank sky limit which had been set with the same telescope in a search by Readhead *et al* for primordial CMB anisotropy. The answer lies in the (necessarily) different observing strategies adopted in the two cases. For the intrinsic anisotropy work, Readhead *et al* were able to minimize telescope movement by working with fields very close to the north celestial pole and observing only near transit. For observations of specific clusters, there is no such option, since the cluster must be followed around an appreciable portion of the sky in order to build up sufficient sensitivity. This leads to two problems. First, the combination of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

telescope on–off motions and beamswitching used for both experiments in order to reduce atmospheric effects can still leave residual systematics due to incomplete subtraction of sidelobes of the two beams. This is known as 'differential spillover'.

The second problem is that, as the source sweeps around the sky, the 'off' beams themselves sweep over circular arcs, displaced from the cluster center by the beamthrow. If these 'reference arcs' contain radio sources, these will produce negative dips in the central signal as the beams pass over them, leading to a false impression of a decrement having been detected. The only solution to this is an independent survey for the confusing radio sources. To achieve sufficient sensitivity in such a survey requires the power of the Very Large Array, but the area that needs to be covered for each cluster means that several VLA fields need to be mosaiced together, requiring quite large amounts of observing time.

Bolometer measurements

A bolometric system is able to achieve great sensitivity by having a very large fractional bandwidth and working at high frequencies. The disadvantage is that they are strongly affected by the atmosphere at these high frequencies and so have to be run from mountain tops, balloons (which have only limited observation time) or space.

The first really successful ground-based bolometer measurements have come from the Caltech SuZie group, working on the Caltech Submillimeter Observatory in Hawaii. They use a differencing bolometer scheme, coupled with drift scans across the source, which provides good stability and rejection of atmospheric fluctuations. The wavelengths observed are 1.2, 1.4 and 2.2 mm, which are properly spaced to make an attempt at separating the kinetic and thermal SZ effects. So far only upper limits have been found for the cluster peculiar velocities, but it is quite possible this instrument will be the first to detect non-zero velocities in the future. Scans at high signal to noise have been obtained on several clusters including A2163, which is the hottest known ($T_e \sim 14$ keV), making it necessary to use the correct *relativistic* formulae for the frequency dependence and amplitude of the SZ effect. Lamarre *et al*, using a balloon-borne bolometric system called PRONAOS, have made a determination of the spectrum of A2163 extending all the way to far-IR wavelengths (see figure 3), and in conjunction with the SuZie results this shows convincing evidence for the full expected frequency dependence of the SZ effect, albeit mixed in with dust emission at higher frequencies.

Interferometers

Interferometers provide currently provide the cleanest technique for making *images* of the SZ effect. An interferometer consists of a set of baselines formed by taking the elements of the interferometer in pairs. Each baseline is sensitive to a particular Fourier component on the sky and is relatively *in*sensitive to atmospheric

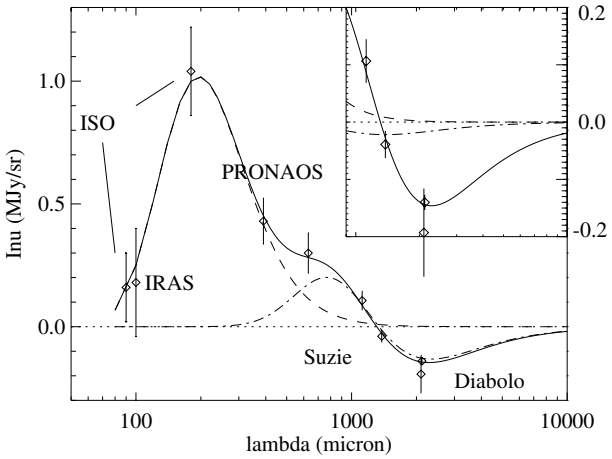


Figure 3. Spectrum of A2163, from Lamarre J M *et al* 1998 *Astrophys. J.* **507** L5. The solid curve shows a best-fit model including both the SZ effect (dot–dash curve) and predicted dust emission (dashed). The inset shows the expected contribution from the the kinetic SZ effect while the names are of the experiments used to trace different regions of the spectrum.

fluctuations and ground spillover and pickup (the problems which bedevilled earlier attempts to measure the SZ effect). Furthermore, close-in baseline pairs can be used to give a map of the SZ effects, while longer baselines can be used to subtract the effects of discrete radio sources, which are often variable, making it very desirable to measure them simultaneously with the SZ measurements. Two groups in particular have been pursuing this approach. The first image of the SZ effect came from the Ryle Telescope at Cambridge, UK, in 1993, and subsequently about 10 other clusters have been detected and mapped. In figure 4 a map is shown of the cluster A1413 made with the Ryle Telescope (Cambridge). The more centrally condensed nature of the x-ray emission versus the SZ effect is evident. The other group actively pursuing this approach have been using the OVRO and BIMA millimeter arrays in California, but with centimeter wave rather than millimeter wave receivers fitted. The lower observing frequency gives them sensitivity on angular scales appropriate to the SZ effect and they are able to use otherwise fallow summer observing periods to conduct the measurements. This campaign has resulted in the detection of about 25 clusters in SZ, and the highest signal-to-noise maps thus far made.

Determinations of H_0

The SZ effect can be used to determine H_0 in conjunction with x-ray information. We start with the equation for the x-ray flux:

$$S_x \propto \int \frac{n_e^2 T_e^{0.5}}{D_L^2}\, \mathrm{d}V$$

where n_e and T_e are the electron number density and temperature and D_L is the luminosity distance to the cluster. Via the x-ray and optical information we can

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

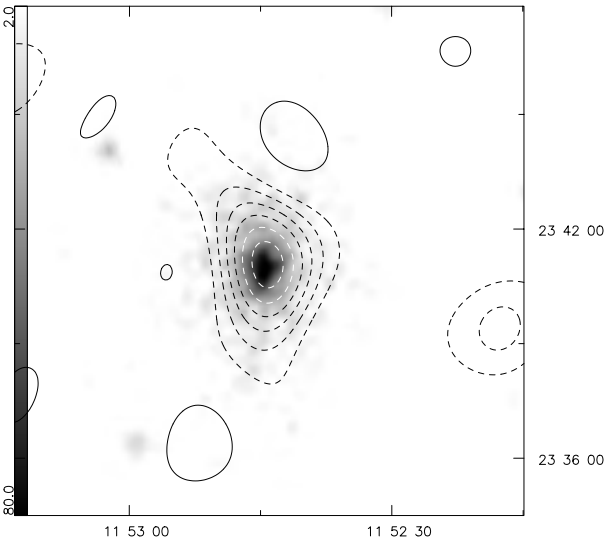


Figure 4. Ryle Telescope map of the cluster Abell 1413 (contours) overlaid on ROSAT PSPC x-ray image (greyscale). Contour levels for the radio map are −450 to 75 μJy/beam in steps of 75 μJy/beam. The numbers at the right are declination in degrees, minutes and seconds, indicating the overall angular scale. (Taken from Grainge K *et al* 1996 *Mon. Not. R. Astron. Soc.* **278** L17.)

determine T_e, θ and z, where θ is a characteristic angular size of the cluster and z is the cluster redshift. The SZ effect is determined by

$$\Delta T_{SZ} \propto \int n_e T_e \, \mathrm{d}l$$

i.e. a line-of-sight integral of pressure through the cluster, and we can relate dl to the other quantities via

$$\Delta l = \frac{f \theta D_L}{(1+z)^2}.$$

Here f is a prolateness factor, which, assuming the cluster has an ellipsoidal shape, is the ratio of the length of the major axis of the ellipse to the minor axis. The last equation uses the angular diameter distance formula to get the physical width of the cluster given its angular diameter and then converts this to a line-of-sight distance through the cluster using this assumed prolateness. This is done assuming what will turn out to be a 'worst case', where the cluster is indeed prolate and pointing straight at the observer. If we assume the cluster is at not too high a redshift, then the luminosity distance can be approximated by $D_L = cz/H_0$. (Obviously more accurate formulae can be used for this and other quantities at higher redshift.) Eliminating n_e via the known S_x, and D_L in favor of H_0, one finds

$$H_0 \propto \frac{f}{\Delta T_{SZ}^2}$$

with the constant of proportionality known. We note straightaway that if f is assumed to be 1 when it is really >1, i.e. where the cluster is indeed oriented towards the observer, then the H_0 derived by this method will come out too small. Since the H_0 values found so far using this approach have tended to come out *lower* than optical determinations, it is obviously this case that we need to worry about most in the first instance. Furthermore, for clusters which are selected on the basis of x-ray surface brightness, there would be a tendency for just this effect to happen, since the greater line-of-sight distance through the cluster would emphasize their x-ray brightness.

Putting back further factors we obtain

$$H_0 \propto \frac{f}{\Delta T_{SZ}^2} S_x T_e^2. \quad (1)$$

T_e can be found from the x-ray spectrum obtained, e.g. with the ASCA satellite, which covers the range up to ∼10 keV. A typical cluster temperature is $T_e \sim 6$ keV and ROSAT only covers the range up to ∼2 keV. The x-ray flux, S_x, can be obtained from e.g. ASCA, with resolution ∼2′, or the ROSAT PSPC, with resolution ∼15″. The cluster profile is modelled generally via an assumed isothermal atmosphere, with

$$n_e(r) = \frac{n_e(0)}{[1 + (r/r_c)^2]^{3\beta/2}}.$$

Typical values for clusters are $\beta \sim 0.6$, and core radius $r_c \sim 0.25$ Mpc, but one has to *fit* for these in any given case. For elliptical clusters, one needs to introduce different core radii along orthogonal axes. These are fitted iteratively until they match the x-ray observations. Ideally, one needs to include the physics of the pressure and cooling of the cluster in this process, and also to use simultaneously the SZ observations.

The values found so far for H_0 using this technique have encompassed a rather large range, but estimates are now beginning to stabilize. Birkinshaw, using results from nine well-measured clusters, finds a mean H_0 of 60 ± 10 km s^{-1} Mpc^{-1} (although systematic contributions to the error could be larger). Further results from the much larger samples being built up by the Cambridge and OVRO/BIMA interferometer teams are expected shortly. Carlstrom *et al*, working with detailed modeling of a single cluster, find good agreement with a value of 60 km s^{-1} Mpc^{-1} and moreover have shown that although the results for the core radius r_c and β are highly correlated after fitting, the resulting degeneracy in H_0 is quite small, which is encouraging for further work.

As well as estimating H_0, one can also use the SZ results to estimate the total baryon content of a cluster. In conjunction with information from GRAVITATIONAL LENSING in the cluster, or alternatively by assuming virial equilibrium for the intracluster material, one can then estimate the total matter fraction (baryonic plus dark) in the cluster. With a final assumption that the proportion of baryons to dark matter is similar to that pertaining in the universe as a whole, the total matter density of the universe can be found. Recently Grego *et al*, using 18 clusters from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the OVRO/BIMA sample, have obtained the estimate $\Omega_M \sim 0.23^{+0.06}_{-0.04}$ (for an H_0 of 65 ± 10 km s^{-1} Mpc^{-1}, scaling inversely with H_0 for other values). This value represents that matter fraction relative to the CRITICAL DENSITY needed to make the universe spatially flat and ties in well with other recent indications that, if the universe really is spatially flat (as primordial CMB results seem to indicate), then the deficit must be made up in some other way than via dark matter. This lends extra support to the hypothesis of a non-zero COSMOLOGICAL CONSTANT. With a sufficiently large sample of clusters out to high redshift, it may be possible to use the luminosity distance dependence of the SZ effect to measure the cosmological constant directly.

Future uses of the SZ effect

The redshift independence of the SZ effect makes it an excellent tool for investigating the evolution of structure with redshift. On the other hand, current instruments are severely limited in their ability to do the kinds of 'blank sky' searches for the SZ effect that would be necessary to fully exploit this potential, owing to their relative lack of sensitivity and the time it takes to detect just one cluster in an already-known position. Accordingly, proposals for new ground-based instruments (all interferometers), able to make maps of the SZ effect over large areas of the sky, are now being made and will, it is hoped, be funded. The 'SZ source counts' (the number of clusters per square degree with SZ effect above a certain level) are strong functions of the matter density and geometry of the universe and so should provide significant information on Ω_M. In addition, the large number of clusters detected this way could act as a 'finder' for clusters to be investigated by future x-ray missions, which will not have the same sensitivity at high redshift as dedicated SZ telescopes. As well as ground-based telescopes, some future space missions will also be able to detect large numbers of clusters via the SZ effect. The PLANCK SURVEYOR mission, due for launch by ESA in 2007, has the potential to detect tens of thousands of clusters via the SZ effect, and it is hoped that for a large number of these a detection of the 'kinetic effect' will be possible as well, thus making possible a totally new kind of mapping of peculiar velocity fields on large scales.

Bibliography

The first predictions of the SZ effect were made in two papers:

Sunyaev R and Zeldovich Y 1970 *Comments Astrophys. Space Phys.* **2** 66

Sunyaev R and Zeldovich Y 1972 *Comments Astrophys. Space Phys.* **4** 173

The first image of the SZ effect is presented in

Jones M, Saunders R, Alexander P, Birkinshaw M, Dillon N, Grainge K, Lasenby A, Lefebvre D, Pooley G, Scott P, Titterington D and Wilson D 1993 *Nature* **365** 320

An extensive review of the SZ effect, covering both theory and observation, is given in

Birkinshaw M 1999 *Phys. Rev.* **310** 97

while an up-to-date review of imaging of the SZ effect and its future uses is contained in

Carlstrom J E, Joy M K, Grego L, Holder G P, Holzapfel W L, Mohr J J, Patel S and Reese E D 1999 astro-ph/9905255

Relativistic corrections to the SZ effect are discussed in

Challinor A D and Lasenby A N 1998 *6th Course on Topics in Astrofundamental Physics (Erice 1997)* ed N Sanchez and A Zichichi (Dordrecht: Kluwer) pp 409–42

Anthony Lasenby

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Superclusters and the Local Supercluster

In the 19th century, William and John Herschel noted the uneven distribution of spiral nebulae across the sky (see HERSCHEL FAMILY). In the early 1920s, J H Reynolds remarked that large non-galactic nebulae cluster along a band and 10 yr later HARLOW SHAPLEY, providing confirmation with a more systematic survey, called the flattened concentration of brightest galaxies the 'Local Supergalaxy'. Gérard de Vaucouleurs, in the 1950s, showed that locally there is a flattened distribution of galaxies with the VIRGO CLUSTER at its core and our MILKY WAY GALAXY toward the periphery. This structure has come to be called the '*Local Supercluster*'. After the mid-1970s, the effort to study the distribution of galaxies in the universe intensified and many structures similar to the Local Supercluster have now been found.

Galaxies: the building blocks of large-scale structure

It is currently believed that structure in the universe formed from the gravitational self-attraction of matter in overdense regions. Dense regions collapsed earlier in time. Probably, irregularities with high local density but low total mass formed first, then coalesced with other matter to form larger entities. Galaxies were forming and growing rapidly during the first four to six billion years in the life of the universe. All the while, galaxies were being drawn closer to each other, so they were becoming substantially clumped. The timescales for this gravitational clustering are strongly dependent on the overall properties of the universe. If the average mass density is in the vicinity of the critical density for a closed universe then galaxies are falling together in large numbers today to form large-scale structures. If the average mass density is much less than the critical density then the greatest development in the structure formation happened earlier.

In their gross properties, galaxies can be considered to be of two types. Galaxies referred to as ELLIPTICAL or LENTICULAR GALAXIES overwhelmingly contain only old stars. They have little cold interstellar gas so they lack the resources to form many new stars. On the other hand, galaxies called SPIRAL or IRREGULAR GALAXIES do contain cold interstellar gas and they have a substantial component of young stars. These galaxies can have a wide range of STELLAR POPULATIONS, from very young to very old.

Galaxies of the two types are usually found in very different environments. Ellipticals and lenticulars are found in large numbers in clusters, the densest, most congested parts of the universe. Spiral and irregular galaxies are found in filamentary structures outside the densest clusters. There are relatively few spirals or gas-rich irregulars in clusters. It is possible that those few spirals in clusters have fallen into the gravitational wells of the clusters recently. There are relatively few ellipticals or lenticulars in the low-density regions. Outside of dense clusters, most galaxies have a supply of gas for new stars.

Our own Milky Way Galaxy is a relatively large spiral galaxy. The solar system is located in the flattened disk near a spiral arm spur far out from the center.

Groups: the next level of the hierarchy

Galaxies like to be with other galaxies. Gravity pulls them together. Galaxies formed out of the self-attraction of matter in overdense regions and galaxies, in turn, are embedded in larger-scale overdense regions that collapse to form groups. Seventy per cent of galaxies are identified to lie in groups that are gravitationally bound and another 20% are found in looser associations that may not be bound. The vast majority of these groups appear to be dynamically young. Although the galaxies are usually bound gravitationally to each other, they may not have had time to make many orbits, or even one orbit, around each other. When they do approach, near collisions that cause structural disruption are not uncommon.

The dynamically young groups are predominantly populated with gas-rich spiral and irregular galaxies. There may also be swarms of dwarf spheroidal systems about big galaxies, to judge from our Milky Way situation. These DWARF SPHEROIDAL GALAXIES are so faint that they are difficult to see at the distances of other big galaxies. A small fraction, roughly 5%, of groups are predominantly populated by gas-poor ellipticals and lenticulars. These groups are dense, have higher motions and are undoubtedly more dynamically evolved.

The LOCAL GROUP is an example of a typical environment for galaxies. There are two giant spirals, Andromeda or M31 (the biggest) and the Milky Way, and two intermediate-size galaxies, Triangulum or M33 and the Large Magellanic Cloud. In total there are 40 known or suspected members but most of these are dwarfs. There are two main subgroups within the Local Group, one each around the two giant galaxies. There are two other minor subgroups around the irregular galaxies NGC 6822 and NGC 3109. The two giant galaxies contain 80% of the light of the group and are falling toward each other.

Clusters: the largest collapsed structures in the universe

Rich CLUSTERS OF GALAXIES act as nodal points in the filamentary structure of the universe. Clusters can contain hundreds of substantial galaxies and thousands of dwarfs. The total mass associated with big clusters can reach 10^{15} times the mass of the Sun. Less than 10% of this mass is directly associated with the stars in the galaxies. Of order 20% can be associated with hot hydrogen and helium gas that fills the space between the galaxies in the clusters and radiates thermal x-ray emission. Most of the mass is in an unknown form and is part of the 'DARK MATTER' problem.

The galaxies in the richest clusters are ellipticals, lenticulars and dwarf spheroidals for the most part. Spirals and irregulars fall into the clusters and over time must exhaust or be stripped of their cold gas reservoirs. Down at the center of the cluster one frequently finds a special elliptical galaxy. The central ellipticals are often

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

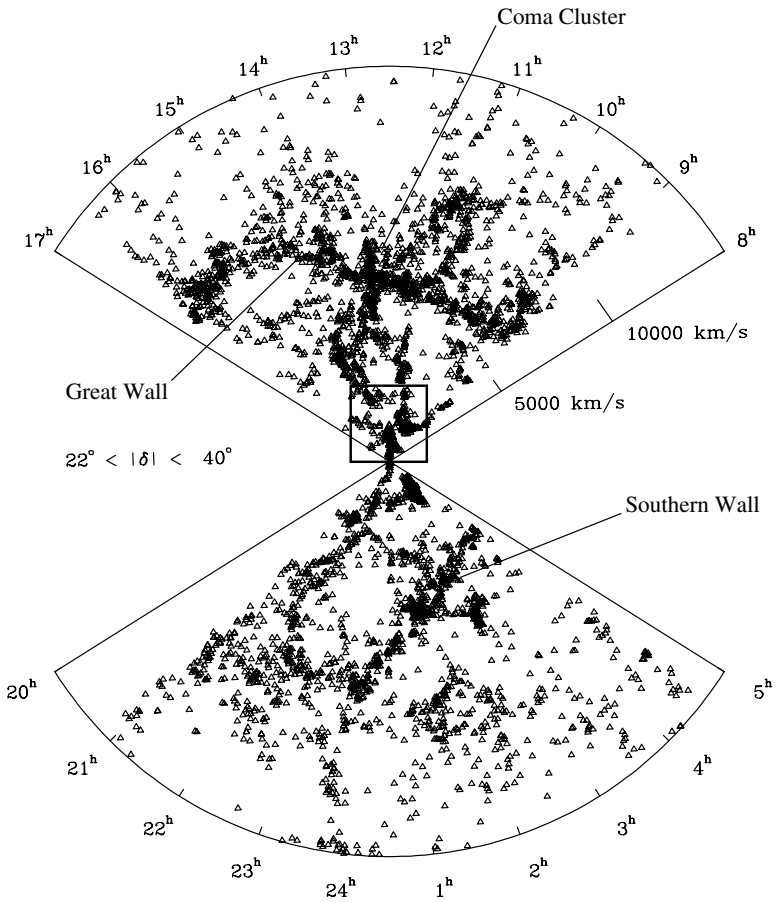


Figure 1. Large-scale structure (Luiz da Costa (European Southern Observatory), Margaret Geller (Harvard-Smithsonian), and John Huchra (Harvard-Smithsonian).

particularly large and puffed up. They probably grow through the ACCRETION of galaxies through collisions and dynamical friction. The hot intracluster gas can slowly condense onto the central galaxy through a cooling flow.

By far the most prominent nearby cluster is the Virgo cluster, usually taken as the center of the Local Supercluster. The Virgo cluster is at a distance of 16 Mpc (50 million light-years). Galaxies in a spherical region within 7 Mpc have decoupled from the expansion of the universe and are falling back onto the cluster.

The nearest rich clusters beyond the Local Supercluster are HYDRA I and CENTAURUS clusters about 2.5 times more distant than the Virgo cluster, the Perseus cluster about 4 times more distant and the particularly rich Coma cluster about 5 times more distant. There are occasional aggregates of rich clusters, such as the 'SHAPLEY CONCENTRATION' of 29 clusters at a distance of roughly 200 Mpc.

Filaments of galaxies: the cosmic web

Ninety per cent of galaxies lie outside of rich clusters in extended filaments or 'clouds'. Individual filaments are long in one dimension, often intermediate in a second dimension, and thin in the third dimension, like a flat pasta or a surfboard. Adjacent filaments can share a common plane to give the impression of a partially filled surface. Other adjacent filaments can lie in layers, analogous to window louvers.

The groups of galaxies lie like beads within the filaments. It is thought that filaments have grown through self-attraction except that collapse is only partial. Gravity has not overcome the expansion of the universe in at least the long dimension. In addition to the ∼90% of galaxies in bound groups or looser associations, about 10% of galaxies are not closely linked with other galaxies but still lie in the filaments. At most only a few per cent of galaxies lie outside these structures (see the complementary article on VOIDS).

Although not conclusively demonstrated, it is strongly indicated that filaments essentially never stand

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

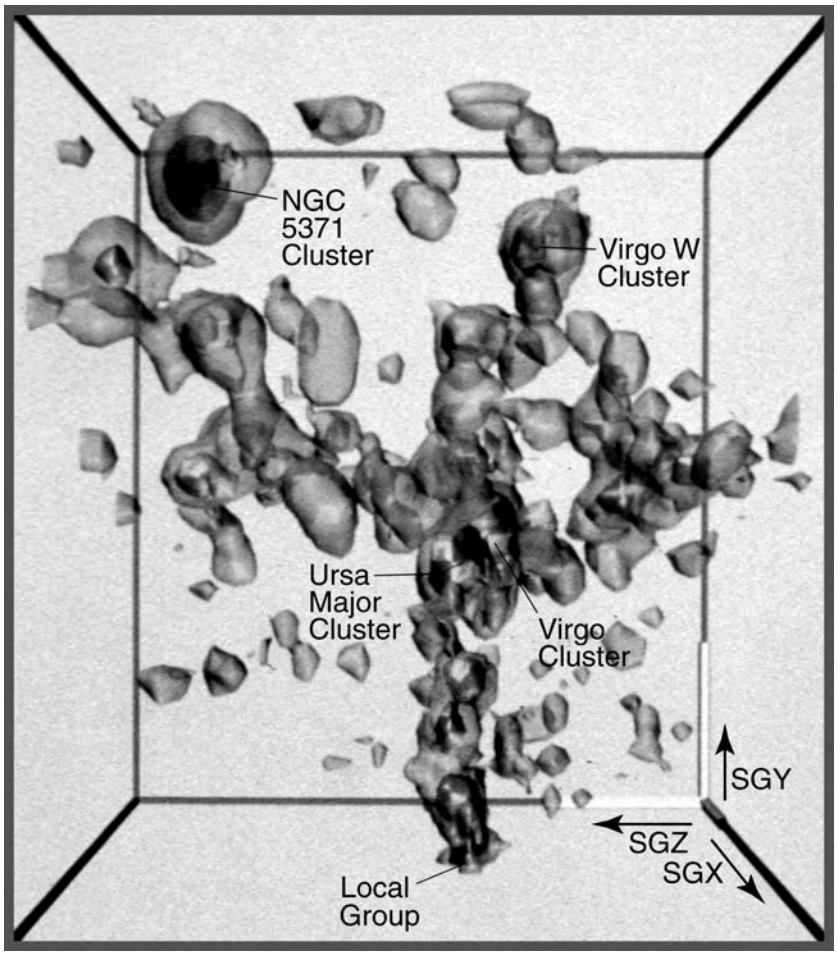


Figure 2. Detail of part of figure 1 (R Brent Tully (University of Hawaii)).

alone. Any two places in the universe that are on filaments may be interconnected by a path that runs through the network of filaments. Simulations of the formation of structure demonstrate that matter collapses onto the filaments in one or two dimensions and then drains along the filaments into the rich clusters. The simulations look subtly different in different cosmological models. If the universe will expand forever, the filaments seen today will eventually dissipate, partly by falling into the clusters and partly by the dilution that comes from continued expansion.

Galaxy flows: response to the distribution of mass

Large-scale structure is evolving. It is now appreciated that familiar objects, such as stars and interstellar gas clouds, contribute only a minor fraction to the gravitational fields that cause the formation of galaxies and structure on scales larger than galaxies. Mysterious dark matter is more important. The amount and distribution of dark matter are estimated through observations of the flows of galaxies in response to the gravitational fields.

The most certain evidence for motion comes from the dipole distribution in the COSMIC MICROWAVE BACKGROUND temperature maps. It is concluded that the Milky Way Galaxy has a motion of 600 km s^{-1} toward a well-defined position in the southern celestial sky. The motions of individual galaxies can be determined, albeit with substantial errors. It is evident that galaxies near clusters are being attracted toward the clusters and, in general, individual galaxies are being attracted toward the largest concentrations of galaxies. Overall, nearby galaxies are partaking of the motion implied by the dipole in the cosmic microwave background radiation. In principle, the amplitude of these motions provides information about the mean density of the universe. In practise, there is an ongoing dispute over the value of the cosmological density parameter that ranges over a factor of 5.

The Local Supercluster and nearby structures

The Milky Way and ANDROMEDA (M31) GALAXIES, with their entourages, reside in the gravitationally bound Local Group. At 3–4 Mpc (10–13 million light-years) away are the neighboring Sculptor, IC 342–Maffei and M81 groups.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

These groups and others string along the Coma–Sculptor Cloud, a branch filament of the Local Supercluster. The Virgo cluster is the central, densest part of the Local Supercluster. The main large-scale structure in this local region is a band of galaxies that extends across half the sky through the Virgo cluster. North of Virgo, the band extends to the smaller URSA MAJOR cluster where one path of a bifurcation leads to the Coma–Sculptor cloud. South of Virgo, the band extends toward the Centaurus cluster. The flow of galaxies indicated by the cosmic microwave background dipole anisotropy is in this general direction. Viewed from afar, the historical Local Supercluster would probably be considered an appendage of a larger Hydra–Centaurus–Virgo supercluster. Recent observations in the obscured regions of the Milky Way plane suggest that the ABELL 3627 CLUSTER may be a particularly important component of this supercluster.

The next nearest important structure is on the opposite side of the sky. A prominent filament, somewhat resembling a lizard projected on the sky, runs through the constellations of Perseus and Pisces. Somewhat farther, at roughly 90° to both the Hydra–Centaurus and the Perseus–Pisces structures, running through the Coma cluster, is a structure called the 'GREAT WALL'. On the opposite side of the sky is the 'Southern Wall'. As our information becomes more complete, connections are invariably found between each of these features.

The two accompanying figures illustrate some of these structures. The large scale is shown by figure 1. Individual galaxies are represented by small triangles. The distance of any given object away from our Milky Way Galaxy is given by the radial distance away from the apex of the wedge. The units of distance are the redshift units of velocity (galaxies that are farther from us are moving away with a higher velocity). It is difficult to represent three dimensions on a flat page. In this figure one is looking at a thin wedge, with hour angle on the sky displayed as the opening angle at the apex. A small range of declination is projected to the flat page. The Coma cluster is at the densest concentration of galaxies in the upper wedge. The outstretched arms from the 'stickman' body created by the Coma cluster is the 'Great Wall'. The 'Southern Wall' is seen in the lower wedge.

The box outline in this figure shows the region which is given in greater detail in figure 2. In this case, the data have been smoothed to show contours of galaxy density. There are many galaxies in the darkest regions, fewer within the lighter contours and none in the regions outside contours. The second figure illustrates the distribution of galaxies in the Local Supercluster. The Local Group lies in the Coma–Sculptor cloud, a filament that runs in a plane over to the Ursa Major cluster and down to the Virgo cluster.

Bibliography

Bahcall N A, 1988 *Ann. Rev. Astron. Astrophys.* **26** 631
Geller M J and Huchra J P 1989 *Science* **246** 897
Giovanelli R and Haynes M P 1991 *Ann. Rev. Astron. Astrophys.* **29** 499
Oort J H 1983 *Ann. Regv. Astron. Astrophys.* **21** 373
Tully R B 1990 *Endeavour, New Ser.* **14** 154

R Brent Tully

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Superior Planets

A collective term for the major planets Mars, Jupiter, Saturn, Uranus, Neptune and Pluto, whose orbits lie outside that of the Earth, as opposed to the inferior planets, those major planets whose orbits lie inside the Earth's.

Compare: outer planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Super-Kamiokande

Super-Kamiokande (SK) found evidence for neutrino masses in 1998. The SK detector is located 1000 m underground in Kamioka mine in Japan. The detector consists of about 50 000 tons of ultra pure water which fills a 40 m tall, 40 m diameter cylindrical stainless steel tank. Cerenkov radiation produced by neutrino interactions in water is viewed with 11 146 photomultiplier tubes mounted on the inner surface of the tank. The detector is used in cooperation with about 120 physicists from 20 collaborating institutes in Japan and the US. The cosmic ray interactions in the atmosphere produce muon-neutrinos and electron-neutrinos. SK has found that almost half of muon-neutrinos change to tau-neutrinos when they travel long distances, such as a diameter of the Earth. This phenomenon, called neutrino oscillations, is realized only when neutrinos have masses and each flavor state of neutrinos is a mixture of different mass states.

Another objective of SK is a precise measurement of solar neutrinos. The observed flux of solar neutrinos is about half of the expectation from the standard solar model. The deficit may be due to neutrino oscillations and the measurement at SK would reveal properties of electron-neutrinos. One of the primary objective of SK is a search for proton decays. If found, the Grand Unified Theory, which unifies strong, electromagnetic and weak inteactions, will be proved.

The data acquisition of SK was started in 1996. The predecessor of SK, Kamiokande, observed neutrino burst from a supernova in the Large Magellanic Cloud in 1987. It was the first neutrino observation in astronomical history and initiated the dawn of a new era in neutrino astronomy.

For further information see
http://www-sk.icrr.u-tokyo.ac.jp/doc/sk/indexl.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Superluminal Motion

Superluminal motion is a term used to describe rapid systematic changes in the appearance of an astronomical object, with an observed speed $v_{\rm obs}$ that appears to be faster than the speed of light c. This phenomenon was first recognized in the early 1970s, although the possibility of very high apparent expansion speeds was recognized much earlier. It is seen at radio wavelengths in the nuclei of QUASARS and ACTIVE GALAXIES, where it is relatively common. In the 1990s, rapid motions have also been discovered in a few objects within our own Galaxy, and the first optical detection of superluminal motion in an external galaxy has also been obtained. In all cases, the apparent physical speed is calculated by multiplying the measured (apparent) angular speed by the distance to the source.

The study of time variations in extragalactic radio sources is made possible by the technique of VLBI (VERY-LONG-BASELINE INTERFEROMETRY), in which radio telescopes up to 1000s of km apart are linked together as an array to allow imaging on very fine angular scales, often less than 0.001 arcsec (1 marcsec). The earliest VLBI observations involved only two or three telescopes, and no imaging was possible. Instead, simple models comprising two or more small components were fitted to the data. These simple models yielded the first detection of superluminal motion, in the quasar 3C 279 in 1971. In the 1990s, imaging arrays such as the Very Long Baseline Array (VLBA) have allowed astronomers to track motions using images taken as often as only a few weeks apart, in the case of rapidly variable sources.

A wide range of apparent speeds is seen in ACTIVE GALACTIC NUCLEI (AGN), from more than $10c$ down to speeds which are termed 'sub-luminal' ($v < c$). The most extensive compilation of measurements by Vermeulen and Cohen (1994) lists 66 extragalactic objects for which speeds have been measured, and well over half are superluminal. Superluminal sources provide important observational evidence for extreme physical phenomena in the centers of active galaxies, and the statistics of superluminal motion have an important role to play in cosmology.

The quasar 3C 345 is one of the best-known superluminal radio sources. Figure 1 shows a series of images in which superluminal motion is found in several features in the jet simultaneously (Zensus *et al* 1995). In many sources, images at low frequency, with arcsecond-scale resolution, show extended radio features on both sides of the nucleus, supporting the belief that the jet is in fact two sided, but is invisible on one side. In a few cases, images with intermediate resolution by the very large array (VLA) show two-sided jets.

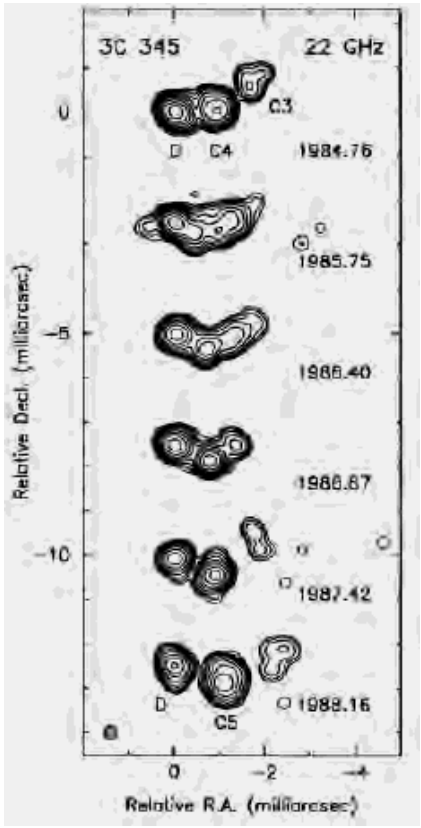


Figure 1. A series of six VLBI images of the quasar 3C 345 (redshift $z = 0.6$, corresponding to a distance of 2000 Mpc) observed over a period of 4 yr. The images are in the radio band, at a frequency 22 GHz, arranged vertically, in time order. The axes are labeled in units of milliarcseconds. A bright feature on the left side is identified as the (stationary) nucleus of the quasar and a feature on the right moves away steadily with time. The superluminal speed ranges between about $3c$ and $5c$.

Kinematic explanation of superluminal motion

The simplest, and by far the most widely accepted, explanation of superluminal motion, consistent with known physics and available astronomical data, is a phenomenon of the kinematics of moving bodies. In most superluminal sources, the motions are seen along the length of a relativistic jet which transports energy from the central engine to a pair of radio lobes. To produce an apparent speed faster than c requires (1) a true speed, in the quasar reference frame, approaching c, (2) a small angle between the jet direction and the line of sight and (3) an independent measure of the distance to the source. A distance is required because the apparent proper motion has to be converted into an (apparent) physical speed by multiplying by the distance.

Given these three conditions, the geometric relation which gives rise to superluminal motion may be explained as follows (see figure 2). At a time t_0, a quasar nucleus (N) ejects a blob of plasma (shaded dot) traveling at relativistic speed $v = \beta c$ along a 'jet' at an angle θ to the line of sight to the observer (shaded arrow). The term relativistic implies $1-\beta \ll 1$. Both the nucleus and the blob emit radio waves (black arrows), which result in the schematic VLBI images shown at the bottom, corresponding to the blob passing points B1 and B2 in the jet.

In the reference frame of the quasar, the time interval

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

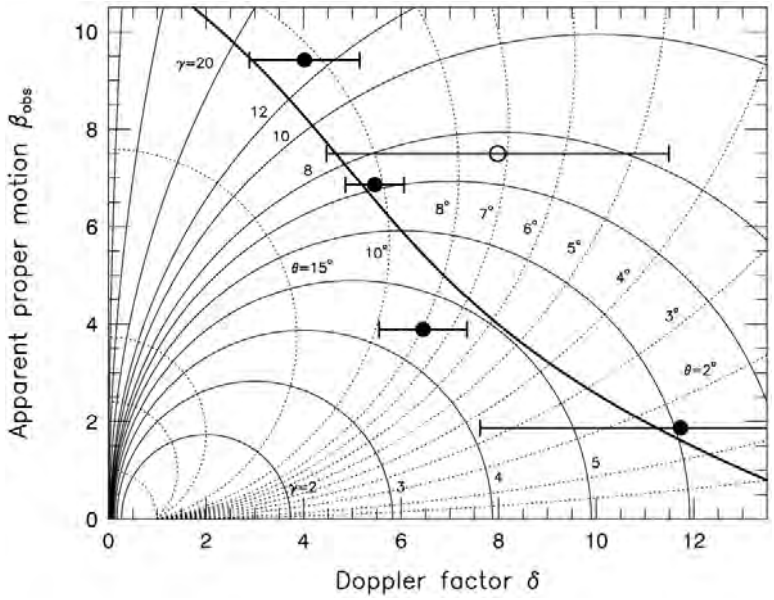


Figure 3. Kinematic parameter space of a superluminal source. The figure shows contours of constant jet angle to the line of sight θ and Lorentz factor γ of the relativistic jet flow. The axes represent two geometric observables: the superluminal speed $\beta_{\rm obs}$ in units of c and the Doppler factor δ deduced from X-ray properties of the source. The region $\beta_{\rm obs} > 1$ constraints the values of (θ, γ) for which the speed is superluminal.

between the events B1 and B2 is simply $t_1 - t_2$. However, the time interval between the arrival at Earth of radio waves from the two events is much less than $t_1 - t_2$, because the blob moves almost as fast as the radio waves it emitted at an earlier time. This is the essence of the phenomenon—if the timescale is compressed, than the apparent speed is increased: the apparent speed is an over-estimate of the true speed.

This compression can be easily quantified. The projected separation of B1 and B2, as deduced from the images, is $v(t_1 - t_2)\sin\theta$. When the blob reaches $B2$, the radio signals from $B1$ are closer to earth by an amount $c(t_1 - t_2) - v(t_1 - t_2)\cos\theta$. Hence the time interval between the VLBI images is $(t_1 - t_2) - (v/c)(t_1 - t_2)\cos\theta$. The apparent speed $v_{\rm obs}$ is simply the projected separation divided by the time difference between the arrival times of signals from $B1$ and $B2$. Hence

$$v_{\rm obs} = \beta_{\rm obs} c = v \sin\theta / [1 - (v/c)\cos\theta].$$

This formula for the apparent speed applies to any angle θ and any jet speed $v = \beta c$ where $\beta < 1$. The regime of superluminal motion corresponds to small θ and v approaching c. Hence $v_{\rm obs} > c$ is not in conflict with special relativity.

Other explanations of superluminal motion

Many explanations of superluminal motion have been proposed. The above geometric explanation, based on Einstein's SPECIAL RELATIVITY, is straightforward and by far the most widely accepted. Alternative explanations of superluminal motion are possible. The most direct is that special relativity is incorrect—in other words, very high speeds are acceptable. However, this explanation would require a major re-write of the laws of physics as currently understood, because special relativity has been experimentally verified in a wide variety of situations. Furthermore, models allowing speeds in excess of c are logically inconsistent with the simple geometric model below.

An alternative explanation which is possible, in principle, is that the distance to the superluminal source has been overestimated. To compute a speed from a measured PROPER MOTION across the sky, one requires the distance to the source. Distances to AGN are computed using their REDSHIFT and the cosmological interpretation of the HUBBLE CONSTANT. Eliminating superluminal motion in AGN would require either a non-cosmological redshift (i.e. that the redshift is not a measure of the source distance) or increasing the Hubble constant by a factor of 10. While these possibilities do not violate any known physics directly, they are inconsistent with the body of astronomical data on scales from parsecs to megaparsecs.

The 'Christmas tree' model is more plausible. In this model superluminal components in a radio source are represented by randomly flashing lights. This allows arbitrarily large speeds, because there is no *causal* relation between successive images. If this explanation is correct, then most superluminal sources would show random variations between successive images, with no preference for expansions over contractions. In poorly sampled time sequences, this may appear to be the case. However, in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

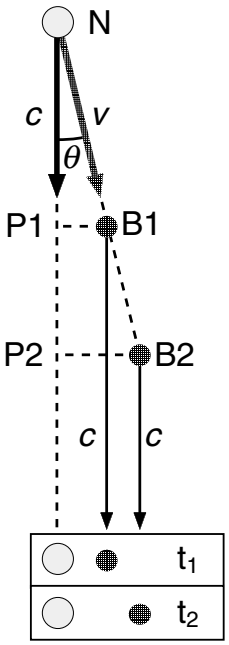


Figure 2. Geometric explanation of superluminal motion. At a time t_0, a quasar nucleus (N) ejects a blob of plasma (shaded dot) traveling at speed $v = \beta c$ ($\beta < 1$) along a 'jet' at an angle θ to the line of sight to the observer (shaded arrow). Both the nucleus and the blob emit radio waves (black arrows), which result in the schematic radio images shown at the bottom, corresponding to the blob passing points B1 and B2 in the jet. In the observer's frame, the time interval between events B1 and B2 is greatly compressed, resulting in an apparent jet speed which is an overestimate of the true speed (see text).

cases when the images are of good quality, and the time sampling is adequate, the motions are highly organized, with expansions almost invariably favored. In rare cases, GRAVITATIONAL LENSING, due to chance alignment of a massive object along the line of sight, can cause distortions in quasar images resulting in superluminal motion. Again, contractions should be as common as expansions, but this explanation remains possible in a handful of cases.

Another class of non-causal model is termed the 'lighthouse model'. A rotating beam from the quasar illuminates stationary clouds that reflect the radio waves to Earth. Again, because there is no causal relation between the emitting regions, very high apparent speeds are possible in certain geometric orientations. This model yields organized motions, such as are seen, but again should show no preference for expansions over contractions. A similar non-causal superluminal model uses 'light echos'. A flare propagates out from the nucleus, illuminating nearby material, which emits radiation toward the observer. Light echos have been convincingly seen in a few active galaxies, in the evolution of the optical spectra from the nucleus. Images of a light echo source should show a maximum speed of $2c$, which is inconsistent with the bulk of the available active galaxy data (Vermeulen and Cohen 1994). However, the ring around supernova SN1987A and the varying intensity of its emission are explained well by this model.

Statistics of superluminal motion

Observations of superluminal sources are subject to strong selection effects, because the brightness of a source increases if it is moving toward the Earth (resulting in Doppler boosting). The receding jet, or counter-jet, is often dimmed by this effect to the point of undetectability. Doppler boosting is dependent on a geometric combination of θ and γ, just like superluminal motion. The boost, relative to a stationary emitter is $\delta^{3+\alpha}$ where α is the power-law spectral index of the emission defined by $S \propto \nu^{\alpha}$, where S is the flux density, and ν is the observing frequency. δ is the Doppler factor (or blueshift factor), and is given by

$$\delta = \gamma^{-1}(1 - \beta\cos\theta)^{-1}$$

where the jet Lorentz factor $\gamma = (1 - \beta^2)^{-1/2}$.

δ and $\beta_{\rm obs}$ are shown in figure 3, as functions of θ and γ. This plot is a tool for deriving the important Lorentz factor γ and other geometric quantities, since they can be derived from measurements of δ and $\beta_{\rm obs}$. For instance, a measured jet to counter-jet brightness ratio excludes a circular region, centered on the origin, in this parameter space.

Many observational consequences result from this simple model, and a number are being actively studied. Relativistic beaming ensures that most of the quasars well studied with VLBI are those that are beamed almost at the Earth, since these are the brightest and easiest to observe. However, only a small fraction of the parent population can be directed toward Earth, if the true orientations are random. The identification of a parent population, or populations, of objects for the superluminal sources is a topic of much current research. Are relativistic motions (β approaching 1) ubiquitous in quasars? The answer may be yes, but it depends on how extreme the selection effects are in real samples. Does a characteristic (physical) speed (i.e. Lorentz factor γ) for quasars exist? If so, then that speed can be used to measure distances out to large redshifts. The superluminal phenomenon is rare in quasars as a whole, because most are weak radio sources. Because superluminal motion involves the redshift, it may be used as a tool to study the COSMOLOGICAL EVOLUTION of the universe. In a comprehensive statistical study of a sample of 66 extragalactic sources with measured proper motions, Vermeulen and Cohen (1994) concluded that a single Lorentz factor is inconsistent with current observations.

Jet physics

Superluminal motion is one of the phenomena encountered when studying the kinematics of jets. Many extragalactic RADIO SOURCES, including the radio-loud quasars, show essentially two-dimensional 'lobe' structures on

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

large scales, out to Mpc. Often, but not always, a visible narrow jet links an active nucleus at the center of the host galaxy with the outer lobes. Basic physical arguments that require transport of energy from the nucleus have been widely accepted since the 1970s: the diffuse ionized lobes require continuous re-supply of energy in order to emit radio waves. The jets are streams of particles with high bulk kinetic energy which provide this transport of energy and matter. The physics of jets remains a very active area of astronomical research, because of the wide range of physical processes which occur within jets and in their interactions with the 'host' galaxy medium. BLAZARS represent the most extreme examples of AGN. Almost all are strong, highly variable, radio sources, but also emit so strongly in other wavebands (including the hardest gamma rays) that they are among the most luminous objects in the universe. Jets may remain relativistic to large radii, but most are too faint to detect reliably in the radio beyond a few milliarcsec. A handful, including the famous 20 arcsec jet in quasar 3C 273, are visible in the optical.

Superluminal motion is just one of a number of observations which can help us to understand the physics of relativistic jets (see review by Marscher 1998). Doppler boosting in jets (the $\delta^{3+\alpha}$ factor described above) is the key to understanding the broad-band energy emission from active galaxies. The shape of the overall spectrum—from radio waves to gamma rays—is determined not only by the intrinsic emission from the nucleus and jet but also by Doppler boosting which results from bulk flow of emitting material. In the case of 3C 345, Unwin and Wehrle (1992) showed that the x-ray emission generated via the inverse-Compton process should exceed the actual x-ray brightness measured by x-ray satellites, unless Doppler boosting is present. Surveys have combined x-ray and VLBI data to deduce lower limits to the Doppler boosting of samples of superluminal radio sources.

Not all superluminal sources are quasars or AGN. A few objects in the MILKY WAY GALAXY also show superluminal motion. An example is GRS 1915+105, a relativistic jet source, which emits strongly in x-rays as well as in the radio (Rodriguez and Mirabel 1999). Because the object is within our Galaxy, the apparent speed of $1.25c$ is detectable in less than a day, compared with months to years for the much more distant (and powerful) AGN. Galactic superluminal sources are of particular interest because of the opportunity to study a relativistic jet 'up close'. While important differences exist, the study of these nearby 'mini blazars' may help our understanding of the most distant and powerful objects in the universe.

Bibliography

Marscher A P 1998 The blazar paradigm *ASP Conference Series 144* (San Francisco, CA: ASP) pp 25–32

Rodriguez L F and Mirabel I F 1999 Repeated relativistic ejections in GRS 1915+105 *Astrophys. J.* **511** 398–404

Unwin S C and Wehrle A E 1992 Kinematics of the parsec-scale relativistic jet in 3C 345 *Astrophys. J.* **398** 74–86

Vermeulen R C and Cohen M H 1994 Superluminal motion statistics and cosmology *Astrophys. J.* **430** 467–94

Zensus J A, Cohen M H and Unwin S C 1995 The parsec-scale jet in quasar 3C 345 *Astrophys. J.* **443** 35–53

Stephen C Unwin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Supermassive Black Holes in Active Galactic Nuclei

QUASARS are among the most energetic objects in the universe. We now know that they live at the centers of galaxies and that they are the most dramatic manifestation of the more general phenomenon of ACTIVE GALACTIC NUCLEI (AGNs). These include a wide variety of exotica such as SYEFERT GALAXIES, RADIO GALAXIES and BL LACERTAE OBJECTS. Since the discovery of quasars in 1963, much effort has gone into understanding their energy source. The suite of proposed ideas has ranged from the relatively prosaic, such as bursts of star formation that make multiple supernova explosions, to the decidedly more colorful, such as supermassive stars, giant pulsars or 'spinars', and supermassive black holes (SMBHs). Over time, SMBHs have gained the widest acceptance. The key observations that led to this consensus are as follows.

Quasars have prodigious luminosities. Not uncommonly, $L \sim 10^{46}$ erg s^{-1}; this is 10 times the luminosity of the brightest galaxies. Yet they are tiny, because they vary on timescales of hours. From the beginning, the need for an extremely compact and efficient engine could hardly have been more apparent. Gravity was implicated, because collapse to a BLACK HOLE is the most efficient energy source known. The most cogent argument is due to Donald Lynden-Bell (1969). He showed that any attempt to power quasars by nuclear reactions alone is implausible. First, a lower limit to the total energy output of a quasar is the energy, $\sim 10^{61}$ erg, that is stored in its radio-emitting plasma halo. This energy weighs 10^{40} g or $10^7 M_\odot$. However, nuclear reactions produce energy with an efficiency of only $\varepsilon = 0.7\%$. Then the waste mass left behind in powering quasars would be at least $M_\bullet \simeq 10^9 M_\odot$. Lynden-Bell argued further that quasar engines are $2R \approx 10^{15}$ cm in diameter because large variations in quasar luminosities are observed on timescales as short as 10 h. However, the gravitational potential energy of $10^9 M_\odot$ compressed into a volume as small as 10 light-hours is $\sim GM_\bullet^2 < / R \gtrsim 10^{62}$ erg. As Lynden-Bell noted, 'Evidently although our aim was to produce a model based on nuclear fuel, we have ended up with a model which has produced more than enough energy by gravitational contraction. The nuclear fuel has ended as an irrelevance'. We now know that the total energy output is larger than the energy that is stored in a quasar's radio source; this strengthens the argument. Meanwhile, a caveat has appeared: the objects that vary most rapidly are now thought to contain relativistic jets that are beamed at us. This boosts the power of a possibly small part of the quasar engine and weakens the argument that the object cannot vary on timescales less than the light travel time across it. However, this phenomenon would not occur at all if relativistic motions were not involved, so SMBH-like potential wells are still implicated. These considerations suggest that quasar power derives from gravity.

The presence of deep gravitational potentials has long been inferred from the large velocity widths of the emission lines seen in optical and ultraviolet spectra of AGNs. These are typically 2000–10 000 km s^{-1}. If the large Doppler shifts arise from gravitationally bound gas, then the binding objects are both massive and compact. The obstacle to secure interpretation has always been the realization that gas is easy to push around: explosions and ejection of gas are common astrophysical phenomena. The observation that unambiguously points to relativistically deep gravitational potential wells is the detection of radio jets with plasma knots that are seen to move faster than the speed of light, c. Apparent expansion rates of (1–10)c are easily achieved if the true expansion rate approaches c and the jet is pointed almost at us.

The final pillar on which the SMBH paradigm is based is the observation that many AGN jets are well collimated and straight. Evidently AGN engines can remember ejection directions with precision for up to 10^7 yr. The natural explanation is a single rotating body that acts as a stable gyroscope. Alternative AGN engines that are made of many bodies—such as stars and supernovae—do not easily make straight jets.

A variety of other evidence also is consistent with the SMBH picture, but the above arguments were the ones that persuaded a majority of the astronomical community to take the extreme step of adopting SMBHs as the probable engine for AGN activity. In the meantime, SMBH alternatives such as single supermassive stars and spinars were shown to be dynamically unstable and hence short lived. Even if such objects can form, they are believed to collapse to SMBHs.

The above picture became the paradigm long before there was direct evidence for SMBHs. Dynamical evidence is the subject of this article and the following one (SUPERMASSIVE BLACK HOLES IN ACTIVE GALAXIES). Meanwhile, there are new kinds of observations that point directly to SMBH engines. In particular, recent observations by the Advanced Satellite for Cosmology and Astrophysics (ASCA) have provided strong evidence for relativistic motions in AGNs. The x-ray spectra of many Seyfert galaxy nuclei contain iron Kα emission lines (rest energies of 6.4–6.9 keV; see figure 1). These lines show enormous Doppler broadening—in some cases approaching 100 000 km s^{-1} or 0.3c—as well as asymmetric line profiles that are consistent with relativistic boosting and dimming in the approaching and receding parts, respectively, of SMBH accretion disks as small as a few Schwarzschild radii.

The foregoing discussion applies to the most powerful members of the AGN family, namely quasars and high-luminosity Seyfert and radio galaxies. It is less

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

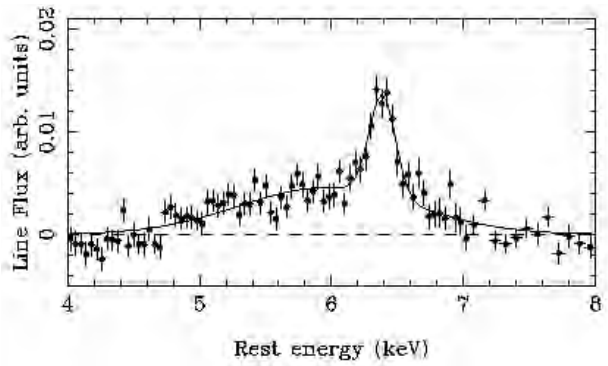
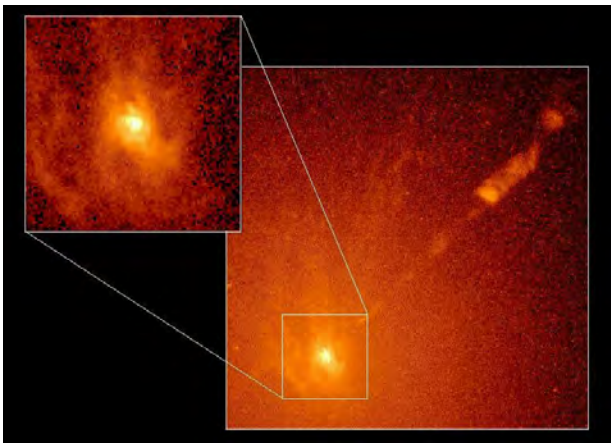


Figure 1. A composite x-ray spectrum of Seyfert nuclei taken with *ASCA* showing the relativistically broadened Fe Kα line. The solid curve is a fit to the line profile using two Gaussians, a narrow component centered at 6.4 keV and a much broader, redshifted component. (Figure adapted from Nandra K *et al* 1997 *Astrophys. J.* **477** 602.)

compelling for the more abundant low-luminosity objects, where energy requirements are less demanding and where long jets or superluminal motions are seen less frequently or less clearly. Therefore a small but vocal competing school of thought continues to argue that stellar processes alone, particularly those that occur during bursts of STAR FORMATION, can reproduce many AGN characteristics. Nonetheless, dynamical evidence suggests that SMBHs *do* lurk in some mildly active nuclei, and, as discussed in the next article, even in the majority of inactive galaxies.

Measuring AGN masses: direct methods

Very general arguments suggest that quasar engines have masses $M_\bullet \sim 10^6$–$10^9 M_\odot$. Gravitational collapse is believed to liberate energy with an efficiency of $\varepsilon \simeq 0.1$; Lynden-Bell's arguments then imply that typical remnant masses are $M_\bullet \sim 10^8 M_\odot$. Better estimates can be derived by asking what we need in order to power quasar luminosities, which range from 10^{44} to 10^{47} erg s^{-1} or $10^{11} L_\odot$ to $10^{14} L_\odot$. For $\varepsilon = 0.1$, the engine must consume 0.02–$20 M_\odot$ yr^{-1}. How much waste mass accumulates depends on how long quasars live. This is poorly known. If they live long enough to make radio jets that are collimated over several Mpc, and if their lifetimes are conservatively estimated as the light travel time along the jets, then quasars last $\gtrsim 10^7$ yr and reach masses $M_\bullet \gtrsim 10^5$–$10^8 M_\odot$. However, the most rigorous lower limit on $M_\bullet$ follows from the condition that the outward force of radiation pressure on accreting matter not overwhelm the inward gravitational attraction of the engine, a condition which, admittedly, strictly holds only if the accreting material and the radiation have spherical symmetry. This so-called Eddington limit requires that $L \leq L_E \sim (4\pi G c m_p/\sigma_T) M_\bullet = 1.3\times10^{38}\ (M_\bullet/M_\odot)$ erg s^{-1}, or



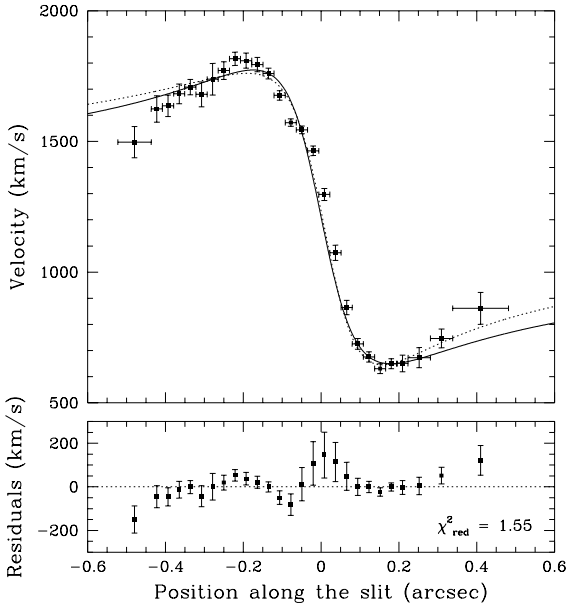


Figure 2. (*a*) HST image of the ionized gas disk near the center of the giant elliptical galaxy M87. The data were taken with the Second Wide Field/Planetary Camera through a filter that isolates the optical emission lines Hα and [N II] λλ 6548, 6583. The left inset is an expanded view of the gas disk; for an adopted distance of 16.8 Mpc, the region shown is 5″×5″ or 410 pc×410 pc. The disk has a major axis diameter of ~150 pc, and it is oriented perpendicular to the optical jet. (Image courtesy of NASA/Space Telescope Science Institute, based on data originally published by Ford H C *et al* 1994 *Astrophys. J.* **435** L27.) (*b*) Optical emission-line rotation curve for the nuclear disk in M87. The data were taken with the Faint Object Camera on HST. The curves in the upper panel correspond to two different Keplerian thin-disk models, and the bottom panel shows the residuals for the best-fitting model. (Figure adapted from Macchetto F *et al* 1997 *Astrophys. J.* **489** 579.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

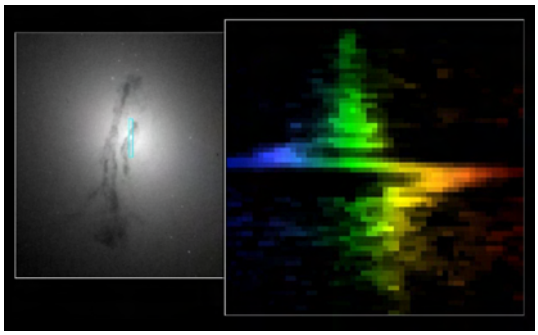


Figure 3. (Left) HST image of the central region of the giant elliptical galaxy M84; the box measures 22´´×19´´ or 1.8 kpc×1.6 kpc for an adopted distance of 16.8 Mpc. The data were taken with the Second Wide Field/Planetary Camera through a filter that isolates the optical emission lines Hα and [N II] λλ 6548, 6583. The slit of the Space Telescope Imaging Spectrograph was placed along the major axis of the nuclear gas disk (blue rectangle). (Right) Resulting spectrum of the central 3´´ (240 pc). The abscissa is velocity and the ordinate is distance along the major axis. The spectrum shows the characteristic kinematic signature of a rotating disk. The velocity scale is coded such that blue and red correspond to blueshifts and redshifts, respectively; the total velocity range is 1445 km s^{-1}. (Image courtesy of NASA/Space Telescope Science Institute, based on data originally published by Bower G A *et al* 1997 *Astrophys. J.* **483** L33 and Bower G A *et al* 1998 *Astrophys. J.* **492** L111.)

equivalently that $M_\bullet \geq 8\times10^5 (L/10^{44}\ \mathrm{erg\ s^{-1}}) M_\odot$. Here G is the gravitational constant, m_p is the mass of the proton and σ_T is the Thompson cross section for electron scattering. We conclude that we are looking for SMBHs with masses $M_\bullet \sim 10^6$–$10^9 M_\odot$. Finding them has become one of the 'Holy Grails' of astronomy because of the importance of confirming or disproving the AGN paradigm.

AGNs provide the impetus to look for SMBHs, but ACTIVE GALAXIES are the most challenging hunting ground. Stellar dynamical searches first found central dark objects in inactive galaxies (see SUPERMASSIVE BLACK HOLES IN INACTIVE GALAXIES), but they cannot be applied in very active galaxies, because the nonthermal nucleus outshines the star light. We can estimate masses using the kinematics of gas, but only if it is unperturbed by nongravitational forces. Fortunately, this complication can be ruled out *a posteriori* if we observe that the gas is in Keplerian rotation around the center, i.e. if its rotation velocity as a function of radius is $V(r):r^{-1/2}$. We can also stack the cards in our favor by targeting galaxies that are only weakly active and that appear to show gas disks in images taken through narrow bandpasses centered on prominent emission lines.

Kinematics of optical emission lines

High-resolution optical images taken with ground-based telescopes and especially with the Hubble Space Telescope (HST) show that many giant ELLIPTICAL GALAXIES contain nuclear disks of dust and ionized gas. The most famous case is M87 (figure 2(*a*)). The disk measures ~150 pc across, and its rotation axis is closely aligned with the optical and radio jet. This is in accord with the SMBH accretion picture. The disk is in Keplerian rotation figure 2(*b*)) around an object of mass $M_\bullet \simeq 3\times10^9 M_\odot$. Furthermore, this object is dark: the measured mass-to-light ratio exceeds 100 in solar units, and this is much larger than that of any known population of stars. Moreover, the dark mass must be very compact: the velocity field limits its radial extent to be less than 5 pc. Therefore its density exceeds $10^7 M_\odot$ pc^{-3}. Another illustration of this technique is given in figure 3. M84, also a denizen of the Virgo cluster of galaxies, is a twin of M87 in size, and it, too, harbors an inclined nuclear gas disk (diameter ~80 pc), whose rotation about the center betrays an invisible mass of $M_\bullet \simeq 2\times10^9 M_\odot$. Other cases are reported (NGC 4261, NGC 6251, NGC 7052), and searches for more are in progress.

Kinematics of radio masers

A related approach exploits the few cases where 22 GHz microwave maser emission from water molecules has been found in edge-on nuclear disks of gas. Particularly strong 'megamasers' allow radio astronomers to use interferometry to map the velocity field with exquisite angular resolution. In the most dramatic application of this method, the Very Long Baseline Array was used to achieve resolution 0´´.0006—100 times better than that delivered by HST—in observations of the Seyfert galaxy NGC 4258. This is only 6 Mpc away, so the linear resolution was a remarkable 0.017 pc. The masers trace out a slightly warped annulus with an inner radius of 0.13 pc, an outer radius of 0.26 pc and a thickness of <0.003 pc (figure 4, left). The masers with nearly zero velocity with respect to the galaxy are on the near side of the disk along the line of sight to the center, while the features with high negative (approaching) and positive (receding) velocities come from the disk on either side of the center. High velocities imply that $3.6\times10^7 M_\odot$ of binding matter resides interior to r=0.13 pc. What is most compelling about NGC 4258 is the observation that the rotation curve is so precisely Keplerian (figure 4, right). From this result, one can show that the radius of the mass distribution must be $r \lesssim 0.012$ pc. If the central mass were not an SMBH, its density would be extraordinarily high, $\rho; > 5\times10^{12} M_\odot$ pc^{-3}. This is comparable with the density of the dark mass at the center of our Galaxy (see SUPERMASSIVE BLACK HOLES IN ACTIVE GALAXIES). Under these extreme conditions, one can show that a cluster of stellar

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

remnants (white dwarf stars, neutron stars and stellar-size black holes) or substellar objects (planets and brown dwarfs) is short lived. Astrophysically, these are the most plausible alternatives to an SMBH. Therefore the dynamical case for an SMBH is stronger in NGC 4258 and in our Galaxy than in any other object.

Measuring AGN masses: indirect methods

Direct dynamical measurements are impractical for more luminous and more distant AGNs. The tremendous glare from the nucleus outshines the circumnuclear emission from stars, and the violent conditions near the center are likely to subject the gas to nongravitational forces. Indirect methods of estimating central masses have been devised to provide a reality check for these more difficult objects.

Fitting the spectra of accretion disks

As material falls toward a black hole, it is believed to settle into an ACCRETION DISK in which angular momentum is dissipated by viscosity. From the virial theorem, half of the gravitational potential energy U is radiated. Therefore the luminosity is

$$L = \frac{1}{2}\frac{\mathrm{d}U}{\mathrm{d}t} = \frac{1}{2}\frac{GM_\bullet \dot{M}_\bullet}{r}.$$

At sufficiently high accretion rates $\dot{M}_\bullet$, the gas is optically thick, and the disk radiates as a thermal blackbody:

$$L = 2\pi r^2 \sigma T^4.$$

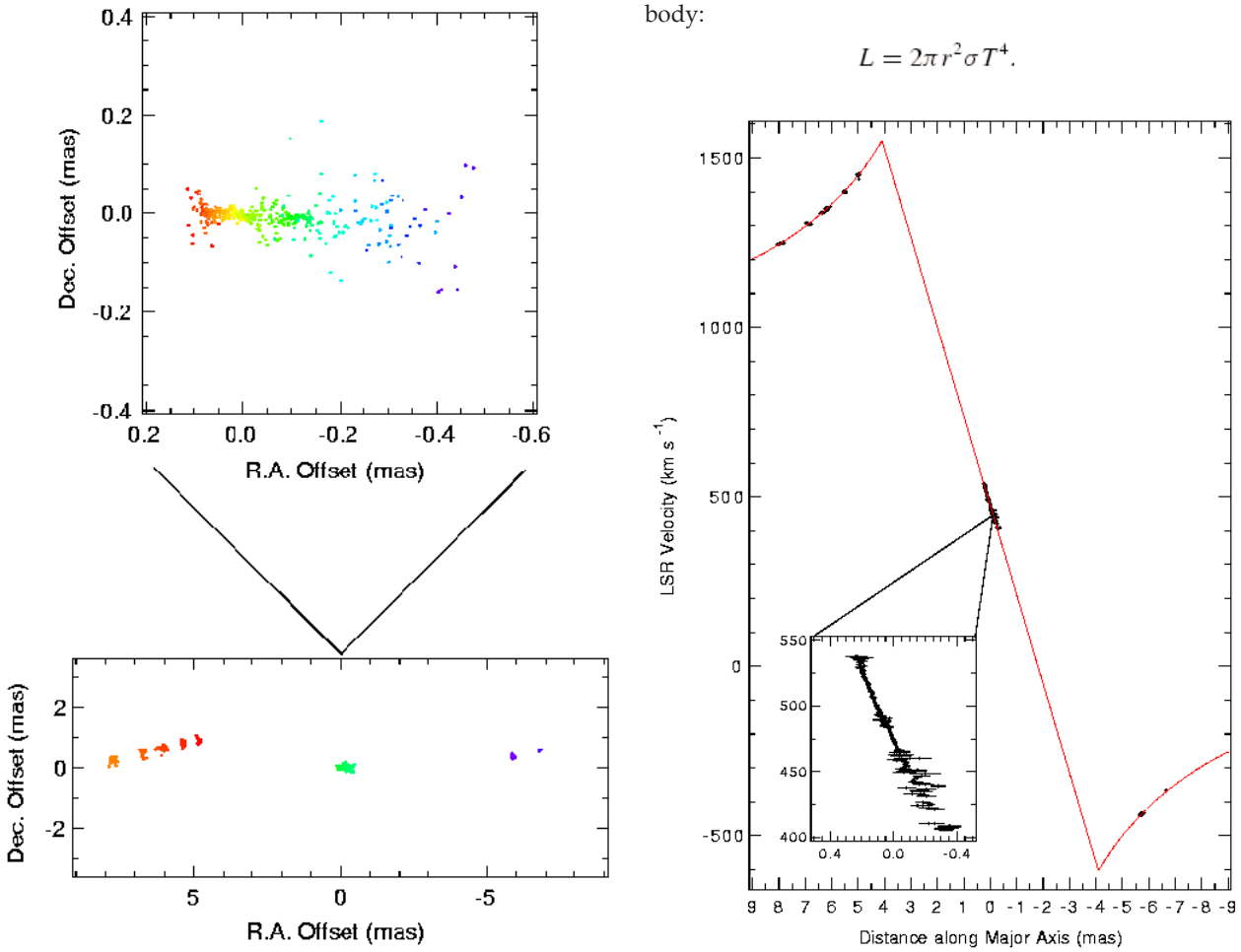


Figure 4. (Left) Spatial distribution of the water masers in NGC 4258, with redshifted velocities to the left and blueshifted velocities to the right. The maser spots are distributed in a thin, warped annulus that is only 4° from edge on. For an adopted distance of 6.4 Mpc, 1 mas=0.031 pc. The top panel shows an expanded view of the emission near the systemic velocity of the galaxy. (Right) Light-of-sight velocity as a function of distance along the major axis of the annulus. The high-velocity features are accurately fitted by a Keplerian model, overplotted as a continuous line. The emission near the systemic velocity, magnified in the inset, lies at nearly constant radius in the front part of the disk along the line of sight to the center. The linear velocity gradient results from the change in projection of the rotation velocity. (Figure adapted from Miyoshi M *et al* 1995 *Nature* **373** 127.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998

Here $2\pi r^2$ is the surface area of the disk and σ is the Stefan–Boltzmann constant. The effective temperature of the disk as a function of radius r is therefore

$$T(r) \simeq \left(\frac{GM_\bullet \dot{M}_\bullet}{4\pi\sigma r^3}\right)^{1/4}.$$

Parameterizing the above result in terms of the Eddington accretion rate, $\dot{M}_E \sim L_E/\varepsilon c^2 = 2.2(\varepsilon/0.1)^{-1}(M_\bullet/10^8 M_\odot)\ M_\odot\ \mathrm{yr}^{-1}$, and the Schwarzschild radius, $R_S \sim 2GM_\bullet/c^2 = 2.95\times10^{13}(M_\bullet/10^8 M_\odot)$ cm, gives

$$T(r) = 6\times10^5\ \mathrm{K}\left(\frac{\dot{M}_\bullet}{\dot{M}_E}\right)^{1/4}\left(\frac{M_\bullet}{10^8 M_\odot}\right)^{-1/4}\left(\frac{r}{R_S}\right)^{-3/4}.$$

In other words, the peak of the blackbody spectrum occurs at a frequency of $\nu_{max} = 2.8kT/h \simeq 4\times10^{16}$ Hz, where k is Boltzmann's constant and h is Planck's constant. This peak is near 100 å or 0.1 keV. In fact, the spectra of many AGNs show a broad emission excess at extreme ultraviolet or soft x-ray wavelengths. This 'big blue bump' has often been identified with the thermal emission from the accretion disk. A fit to the luminosity and the central frequency of the big blue bump gives $M_\bullet$ and $\dot{M}_\bullet$ but not each separately. Corrections for disk inclination and relativistic effects further complicate the analysis. This method is therefore model dependent and provides only approximate masses. Typical values for quasars are $M_\bullet \simeq (10^8$–$10^{9.5})M_\odot$ and $\dot{M}_\bullet \simeq (0.1$–$1)\,\dot{M}_E$. Seyfert nuclei appear to have lower masses, $M_\bullet \simeq (10^{7.5}$–$10^{8.5})M_\odot$, and lower accretion rates, $\dot{M}_\bullet \simeq (0.01$–$0.5)\dot{M}_E$.

Virial masses from optical variability

Surrounding the center at a distance of 0.01–1 pc from the black hole lies the 'broad-line region' (BLR). This is a compact, dense and highly turbulent swarm of gas clouds or filaments. The clouds are illuminated by the AGN's photoionizing continuum radiation and reprocess it into emission lines that are broadened to velocities of several thousand km s^{-1} by the strong gravitational field of the black hole. Then

$$M_\bullet = \eta\frac{v^2 r_{BLR}}{G}$$

where $\eta \simeq 1$–3 depends on the kinematic model adopted, v is the velocity dispersion of the gas as reflected in the widths of the emission lines and r_{BLR} is the radius of the BLR. The latter can be estimated by 'reverberation mapping', as follows. The photoionizing continuum of an AGN typically varies on timescales of days to months. In response, the emission lines vary also, but with a time delay that corresponds to the light travel time between the continuum source and the line-emitting gas. By monitoring the variations in the continuum and the emission lines in an individual object, reverberation mapping provides information on the size of the BLR. These studies also support the assumption that the line widths come predominantly from bound orbital motions. Applying equation (5) suggests that Seyfert nuclei are powered by black holes with masses $M_\bullet \sim (10^7$–$10^8)M_\odot$, while quasar engines are more massive, with $M_\bullet \sim (10^8$–$10^9)M_\odot$. Since quasars also live in more massive host galaxies, this supports the emerging correlation (see the following article) between SMBH mass and the mass of the elliptical-galaxy-like part of the host galaxy.

X-ray variability

AGNs vary most conspicuously in hard x-rays (2–10 keV). One might hope to use the variability timescale to constrain the size of the x-ray-emitting region and hence to estimate the central mass. However, no simple pattern of variability emerges, and defining a meaningful timescale is ambiguous. One approach uses the 'fastest doubling time', Δt, to establish a maximum source size $R \simeq c\,\Delta t$. High-energy photons presumably come from the hot, inner regions of the accretion disk or in an overlying hot corona. For example, if $R \simeq 5R_S$, as deduced in some models, we obtain an upper limit to the mass, $M_\bullet \lesssim (c^3/10G)\,\Delta t \sim 10^4\,\Delta t\,M_\odot$ (Δt in s). Masses estimated in this way are generally consistent with those obtained from other virial arguments, but they are considerably less robust because of uncertainties in associating the variability timescale with a source size. For example, the x-ray intensity variations could originate from localized 'hotspots' in the accretion flow.

X-ray reverberation mapping may in the future be a more powerful tool. The iron Kα line is widely believed to be produced by reprocessing of the hard x-ray continuum by the accretion disk. The strikingly large width and skewness of the line profiles (figure 1), now routinely detected with ASCA, reflect the plasma bulk motion within 10–100 gravitational radii of the center. The temporal response of the line strength and line profile depends on a number of factors that, in principle, can be modeled theoretically; these include the geometry of the x-ray source, the structure of the disk, and the assumed (Schwarzschild or Kerr) metric of the black hole. Time-resolved X-ray spectroscopy should become feasible with the X-ray Multi-Mirror Mission (XMM) in the near future. We can then look forward to constraints on both the masses and the spins of SMBHs.

Summary and prospects

he black hole model for AGN activity has been successful and popular for over three decades. It has withstood the test of time not—at least until recently—because the empirical evidence for SMBHs has been overwhelming but because the alternatives are so implausible. Now progress has advanced on several fronts. The refurbished

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

HST has greatly strengthened the evidence, already growing from ground-based observations, that supermassive dark objects live at the centers of most galaxies. The pace of discoveries is accelerating. The dark objects have exactly the range of masses that we need to explain AGN engines, but we have had no proof that they must be black holes. Then radio interferometry revealed the spectacular maser disk in NGC 4258. For its rotation curve to be as accurately Keplerian as we observe, the central mass must be confined to an astonishingly tiny volume. The inferred density of the central object is so high that astrophysically plausible alternatives can be excluded; an SMBH is the best explanation. The same conclusion has been reached for the SMBH candidate at the center of our Galaxy. This is a major conceptual breakthrough.

In addition, ASCA has demonstrated that many AGNs show iron emission lines with relativistically broadened profiles. This is arguably the best evidence for the strong gravitational field of a black hole. One of the most interesting prospects for the future is time-resolved x-ray spectroscopy, because hot gas probes closest to an accreting black hole.

Finally, the AGN paradigm can be turned inside out to give what may prove to be the most direct argument for black holes. SMBHs were 'invented' to explain nuclear activity in galaxies. In recent years, an ironic situation has developed: some SMBH candidates are too inactive for the amount of matter that we believe they are accreting. The same is true of some stellar-mass black hole candidates that accrete gas from evolving companion stars. A number of researchers recently have developed a theory of 'advection-dominated accretion' in which the accretion disk cannot radiate most of its energy before it reaches R_S either because it is optically thick or because it is too thin to cool. Unless most of the inflowing material ultimately escapes through an outflow, a possibility being explored, the only way to make the accretion energy disappear is to ensure that the accreting body does not have a hard surface. That is, the inactivity of well-fed nuclear engines may be evidence that they have event horizons. Finding event horizons would be definitive proof that AGN engines are black holes.

Bibliography

Initial debate concerning the physical nature of quasars is summarized in

Burbidge G and Burbidge E M 1967 *Quasi-Stellar Objects* (San Francisco: Freeman)

The three key historical papers that originated the SMBH hypothesis are

Lynden-Bell D 1969 *Nature* **223** 690
Salpeter E E 1964 *Astrophys. J.* **140** 796
Zel'dovich Ya B and Novikov I D 1964 *Sov. Phys. Dokl.* **158** 811

The argument for 'gravity power' was further developed in

Lynden-Bell D 1978 *Phys. Scripta* **17** 185

Textbook style discussions of AGN physics can be found in

Courvoisier T J-L and Mayor M 1990 *Active Galactic Nuclei* (Saas-Fee Course 20) (Berlin: Springer)
Peterson B M 1997 *An Introduction to Active Galactic Nuclei* (Cambridge: Cambridge University Press)

The SMBH paradigm is covered at a more technical level in the following review articles:

Begelman M C, Blandford R D and Rees M J 1984 *Rev. Mod. Phys.* **56** 255
Blandford R D 1990 *Active Galactic Nuclei (Saas-Fee Course 20)* ed T J-L Courvoisier and M Mayor (Berlin: Springer) p 161
Rees M J 1984 *Annu. Rev. Astron. Astrophys.* **22** 471

The search for SMBHs is reviewed in

Kormendy J and Richstone D 1995 *Annu. Rev. Astron. Astrophys.* **33** 581
Richstone D *et al* 1998 *Nature* **395**

The starburst theory for the origin of AGNs has been developed by

Terlevich R, Tenorio-Tagle G, Franco J and Melnick J 1992 *Mon. Not. R. Astron. Soc.* **255** 713
Terlevich R, Tenorio-Tagle G, Rozyczka M, Franco J and Melnick J 1995 *Mon. Not. R. Astron. Soc.* **272** 198

The following conference proceedings explicitly focus on the observations and interpretation of the more 'garden variety' low-luminosity AGNs:

Eracleous M, Koratkar A P, Leitherer C and Ho L C (ed) 1996 *The Physics of LINERs in View of Recent Observations* (San Francisco: Astronomical Society of the Pacific)
Schmitt H R, Kinney A L and Ho L C (ed) 1999 *The AGN/Normal Galaxy Connection (Advances in Space Research, vol 23 (5–6))* (Oxford: Elsevier)

Readers interested in a full treatment of the techniques of reverberation mapping should consult

Blandford R D and McKee C F 1982 *Astrophys. J.* **255** 419
Peterson B M 1993 *Publ. Astron. Soc. Pac.* **105** 247

Explicit application of reverberation mapping results to derive masses of AGNs was done by

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Ho L C 1999 *Observational Evidence for Black Holes in the Universe* ed S K Chakrabarti (Dordrecht: Kluwer) p 157
Laor A 1998 *Astrophys. J. Lett.* **505** L83

Mass determinations using optical emission-line rotation curves include

Bower G A *et al* 1998 *Astrophys. J. Lett.* **492** L111
Harms R J *et al* 1994 *Astrophys. J. Lett.* **435** L35
Macchetto F, Marconi A, Axon D J, Capetti A, Sparks W B and Crane P 1997 *Astrophys. J.* **489** 579

The water maser observations of NGC 4258 are described in

Miyoshi M, Moran J, Herrnstein J, Greenhill L, Nakai N, Diamond P and Inoue M 1995 *Nature* **373** 127
Watson W D and Wallin B K 1994 *Astrophys. J. Lett.* **432** L35

Arguments against compact dark star clusters in NGC 4258 and the Galaxy are presented in

Maoz E 1998 *Astrophys. J. Lett.* **494** L181

These papers discuss the derivation of $M_\bullet$ and $\dot{M}_\bullet$ by fitting spectra with accretion disk models:

Laor A 1990 *Mon. Not. R. Astron. Soc.* **246** 369
Wandel A and Petrosian V 1988 *Astrophys. J. Lett.* **329** L11

Attempts to derive masses using x-ray variability have been made by

Wandel A and Mushotzky R F 1986 *Astrophys. J. Lett.* **306** L61

The prediction, discovery and routine detection of broad Fe Kα emission lines are described, respectively, in

Fabian A C, Rees M J, Stella L and White N E 1989 *Mon. Not. R. Astron. Soc.* **238** 729
Tanaka Y *et al* 1995 *Nature* 375 659
Nandra K, George I M, Mushotzky R F, Turner T J and Yaqoob T 1997 *Astrophys. J.* **477** 602

Prospects for x-ray reverberation mapping are foreseen in

Reynolds C S, Young A J, Begelman M C and Fabian A C 1999 *Astrophys. J.* **514** 164
Stella L 1990 *Nature* 344 **747**

Advection-dominated accretion is reviewed in

Blandford R D and Begelman M C 1999 *Mon. Not. R. Astron. Soc.* **303** L1
Mineshige S and Manmoto T 1999 *Adv. Space Res.* **23** (5–6) 1065
Narayan R, Mahadevan R and Quataert E 1998 *The Theory of Black Hole Accretion Discs* ed M A Abramowicz, G Björnsson and J E Pringle (Cambridge: Cambridge University Press) p 148

Luis C Ho and John Kormendy

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Supermassive Black Holes in Inactive Galaxies

Several billion years after the big bang, the universe went through a 'quasar era' when high-energy ACTIVE GALACTIC NUCLEI (AGNs) were more than 10 000 times as numerous as they are now. QUASARS must then have been standard equipment in most large galaxies. Since that time, AGNs have been dying out. Now quasars are exceedingly rare, and even medium-luminosity AGNs such as SEYFERT GALAXIES are uncommon. The only activity that still occurs in many galaxies is weak. A paradigm for what powers this activity is well established through the observations and theoretical arguments that are outlined in the article SUPERMASSIVE BLACK HOLES IN AGN. AGN engines are believed to be supermassive black holes (BHs) that accrete gas and stars and so transform gravitational potential energy into radiation. Expected BH masses are $M_\bullet \sim 10^6$–$10^{9.5} M_\odot$. A wide array of phenomena can be understood within this picture. However, the subject has had an outstanding problem: there was no dynamical evidence that BHs exist. The search for BHs has therefore become one of the hottest topics in extragalactic astronomy.

Since most quasars have switched off, dim or dead engines—starving BHs—should be hiding in many nearby galaxies. This means that the BH search need not be confined to the ACTIVE GALAXIES that motivated it. In fact, definitive conclusions are much more likely if we observe objects in which we do not, as Alan Dressler has said, 'have a searchlight in our eyes'. Also, it was necessary to start with the nearest galaxies, because only then could we see close enough to the center so that the BH dominates the dynamics. Since AGNs are rare, nearby galaxies are not particularly active. For these reasons, it is no surprise that the search first succeeded in nearby, inactive galaxies.

This article discusses stellar dynamical evidence for BHs in inactive and weakly active galaxies. Stellar motions are a particularly reliable way to measure masses, because stars cannot be pushed around by nongravitational forces. The price is extra complication in the analysis: the dynamics are collisionless, so random velocities can be different in different directions. This is impossible in a collisional gas. As we shall see, much effort has gone into making sure that unrecognized velocity anisotropy does not lead to systematic errors in mass measurements.

Dynamical evidence for central dark objects has been published for 17 galaxies. With the Hubble Space Telescope (HST) pursuing the search, the number of detections is growing rapidly. Already we can ask demographic questions. Two main results have emerged. First, the numbers and masses of central dark objects are broadly consistent with predictions based on quasar energetics. Second, the central dark mass correlates with the mass of the ELLIPTICAL-GALAXY-like 'bulge' component of galaxies. What is less secure is the conclusion that the central dark objects must be BHs and not (for example) dense clusters of brown dwarf stars or stellar remnants. Rigorous arguments against such alternatives are available for only two galaxies. Nevertheless, these two objects and the evidence for dark masses at the centers of almost all galaxies that have been observed are taken as strong evidence that the AGN paradigm is essentially correct.

Dead quasar engines in nearby galaxies

The qualitative discussion of the previous section can be turned into a quantitative estimate for $M_\bullet$ as follows. The quasar population produces an integrated comoving energy density of

$$u = \int_0^\infty \int_0^\infty \Phi(L, z) L \, \mathrm{d}L \frac{\mathrm{d}t}{\mathrm{d}z} \mathrm{d}z = 1.3 \times 10^{-15} \ \mathrm{erg\ cm^{-3}}$$

where $\Phi(L,z)$ is the comoving density of quasars of luminosity L at redshift z and t is cosmic time. For a radiative energy conversion efficiency of ϵ, the equivalent present-day mass density is $\rho_u = u/\epsilon c^2 = 2.2 \times 10^4 \ \epsilon^{-1} \ M_\odot \ \mathrm{Mpc}^{-3}$. Comparison of ρ_u with the overall galaxy luminosity density, $\rho_g \simeq 1.4 \times 10^8 h L_\odot \ \mathrm{Mpc}^{-3}$, where the Hubble constant is $H_0 = 100h$ km s^{-1} Mpc^{-1}, implies that a typical nearby bright galaxy (luminosity $L^* \simeq 10^{10} h^{-2} L_\odot$) should contain a dead quasar of mass $M_\bullet \sim 1.6 \times 10^6 \epsilon^{-1} h^{-3} M_\odot$. Accretion onto a BH is expected to produce energy with an efficiency of $\epsilon \sim 0.1$, and the best estimate of h is $\simeq 0.71 \pm 0.06$. Therefore the typical BH should have a mass of $\sim 10^{7.7} M_\odot$. BHs in dwarf ellipticals should have masses of $\sim 10^6 M_\odot$.

In fact, the brightest quasars must have had much higher masses. A BH cannot accrete arbitrarily large amounts of mass to produce arbitrarily high luminosities. For a given $M_\bullet$, there is a maximum accretion rate above which the radiation pressure from the resulting high luminosity blows away the accreting matter. This 'Eddington limit' is discussed in SUPERMASSIVE BLACK HOLES IN ACTIVE GALACTIC NUCLEI. Eddington luminosities of $L \sim 10^{47}$ erg s^{-1} $\sim 10^{14} L_\odot$ require BHs of mass $M_\bullet \gtrsim 10^9 M_\odot$. These arguments define the parameter range of interest: $M_\bullet \sim 10^6$–$10^{9.5} M_\odot$. The highest-mass BHs are likely to be rare, but low-mass objects should be ubiquitous. Are they?

Stellar dynamical searches for central dark objects

The answer appears to be 'yes'. The majority of detections on which this conclusion is based are stellar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

dynamical. However, finding BHs is not equally easy in all galaxies. This results in important selection effects that need to be understood for demographic studies. Therefore we begin with a discussion of techniques. We then give three examples that highlight important aspects of the search. NGC 3115 is a particularly clean detection that illustrates the historical development of the search. M31 is one of the nearest galaxies and contains a new astrophysical phenomenon connected with BHs. Finally, the strongest case that the central mass is a BH and not a dark cluster of stars or stellar remnants is the one in our own Galaxy.

Stellar dynamical mass measurement

Dynamical mass measurement is conceptually simple. If random motions are small, as they are in a gas, then the mass $M(r)$ within radius r is $M(r) = V^2 r/G$. Here V is the rotation velocity and G is the gravitational constant. In stellar systems, some of the dynamical support comes from random motions, so $M(r)$ depends also on the velocity dispersion σ. The measurement technique is best described in the idealized case of spherical symmetry and a velocity ellipsoid that everywhere points at the center. Then the first velocity moment of the collisionless Boltzmann equation gives

$$M(r) = \frac{V^2 r}{G} + \frac{\sigma_r^2 r}{G}\left[- \frac{\mathrm{d}\ln \nu}{\mathrm{d}\ln r} - \frac{\mathrm{d}\ln \sigma_r^2}{\mathrm{d}\ln r} - \left(1 - \frac{\sigma_\theta^2}{\sigma_r^2}\right) - \left(1 - \frac{\sigma_\phi^2}{\sigma_r^2}\right)\right].$$

Here σ_r, σ_θ and σ_ϕ are the radial and azimuthal components of the velocity dispersion. The density ν is not the total mass density ρ; it is the density of the luminous tracer population whose kinematics we measure. We never see ρ, because the stars that contribute most of the light contribute almost none of the mass. Therefore we assume that $\nu(r) \propto$ volume brightness. All quantities in equation (2) are unprojected. We observe brightnesses and velocities after projection and blurring by a point-spread function (PSF). Information is lost in both processes. Several techniques have been developed to derive unprojected quantities that agree with the observations after projection and PSF convolution. From these, we derive the mass distribution $M(r)$ and compare it with the light distribution $L(r)$. If $M/L(r)$ rises rapidly as $r \to 0$, then we have found a central dark object.

There is one tricky problem with this analysis, and it follows directly from equation (2). Rotation and random motions contribute similarly to $M(r)$, but the $\sigma^2 r/G$ term is multiplied by a factor that depends on the velocity anisotropy and that can be less than 1. GALAXY FORMATION can easily produce a radial velocity dispersion σ_r that is larger than the azimuthal components σ_θ and σ_ϕ. Then the third and fourth terms inside the brackets in equation (2) are negative; they can be as small as –1 each. In fact, they can largely cancel the first two terms, because the second term cannot be larger than +1, and the first is $\simeq 1$ in many galaxies. This explains why *ad hoc* anisotropic models have been so successful in explaining the kinematics of giant ellipticals without BHs. However, how anisotropic are the galaxies?

Much effort has gone into finding the answer. The most powerful technique is to construct self-consistent dynamical models in which the density distribution is the linear combination $\rho = \sigma N_i \rho_i$ of the density distributions ρ_i of the individual orbits that are allowed by the gravitational potential. First the potential is estimated from the light distribution. Orbits of various energies and angular momenta are then calculated to construct a library of time-averaged density distributions ρ_i. Finally, orbit occupation numbers N_i are derived so that the projected and PSF-convolved model agrees with the observed kinematics. Some authors also maximize $\sigma N_I \ln N_i$, which is analogous to an entropy. These procedures allow the stellar distribution function to be as anisotropic as it likes in order (e.g.) to try to explain the observations without a BH. In the end, such models show that real galaxies are not extremely anisotropic. That is, they do not take advantage of all the degrees of freedom that the physics would allow. However, this is not something that one could take for granted. Because the degree of anisotropy depends on galaxy luminosity, almost all BH detections in bulges and low-luminosity ellipticals (which are nearly isotropic) are based on STELLAR DYNAMICS, and almost all BH detections in giant ellipticals (which are more anisotropic) are based on gas dynamics.

NGC 3115: $M_\bullet \simeq 10^{9.0\pm0.3} M_\odot$

One of the best stellar-dynamical BH cases is the prototypical S0 galaxy NGC 3115 (figure 1). It is especially suitable for the BH search because it is very symmetrical and almost exactly edge-on. NGC 3115 provides a good illustration of how the BH search makes progress. Unlike some discoveries, finding a supermassive BH is rarely a unique event. Rather, an initial dynamical case for a central dark object becomes stronger as observations improve. Eventually, the case becomes definitive. This has happened in NGC 3115 through the study of the central STAR CLUSTER—a tiny, dense cusp of stars like those expected around a BH (figure 1). Later, still better observations may accomplish the next step, which is to strengthen astrophysical constraints enough so that all plausible BH alternatives (clusters of dark stars) are eliminated. This has happened for our Galaxy (see below) but not yet for NGC 3115.

The kinematics of NGC 3115 show the signature of a central dark object (figure 2). The original detection was based on the crosses. Already at resolution

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. HST WFPC2 images of NGC 3115. The left panel shows a color image made from 1050 s *V*- and *I*-band images. The right panel shows a model of the nuclear disk. The center panel shows the difference; it emphasizes the compact nuclear star cluster. Brightness is proportional to the square root of intensity. All panels are 11".6 square. (This figure is taken from Kormendy *et al* 1996 *Astrophys. J. Lett.* **459** L57.)

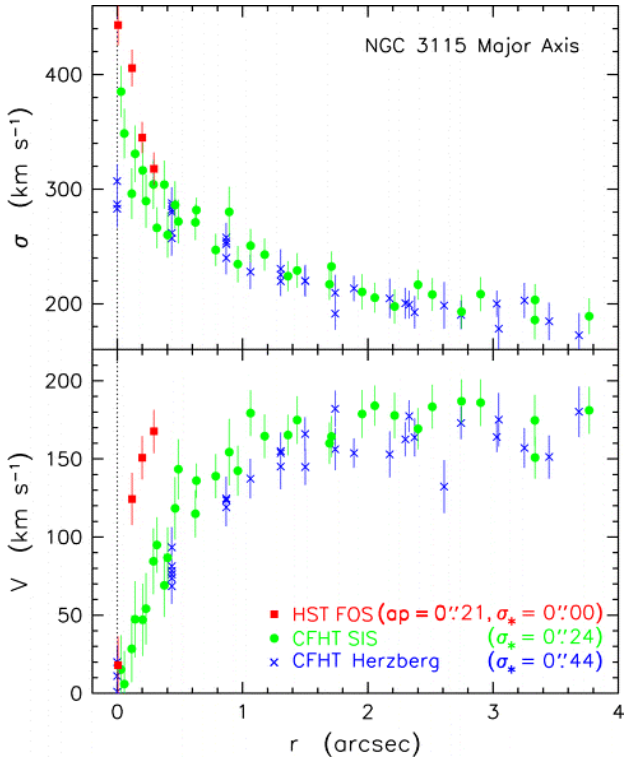


Figure 2. Rotation velocities (lower panel) and velocity dispersions (upper panel) along the major axis of NGC 3115 as observed at three different spatial resolutions. Resolution σ_* is the Gaussian dispersion radius of the PSF; in the case of the HST observations, this is negligible compared with the aperture size of 0".21. (This figure is adapted from Kormendy *et al* 1996 *Astrophys. J. Lett.* **459** L57.)

$\sigma_* = 0".44$, the central kinematic gradients are steep. The apparent central dispersion, $\sigma \simeq 300$ km s^{-1}, is much higher than normal for a galaxy of absolute magnitude $M_B = -20.0$. Therefore, isotropic dynamical models imply that NGC 3115 contains a dark mass $M_\bullet \simeq 10^{9\pm0.3} M_\odot$. Maximally anisotropic models allow smaller masses, $M_\bullet \sim 10^8 M_\odot$, but isotropy is more likely given the rapid rotation.

Since that time, two generations of improved observations have become available. The full circles in figure 2 were obtained with the Subarcsecond Imaging Spectrograph (SIS) and the Canada–France–Hawaii Telescope (CFHT). This incorporates tip-tilt optics to improve the atmospheric PSF. The observations with the HST Faint Object Spectrograph (FOS) have still higher resolution. If the BH detection is correct, then the apparent rotation and dispersion profiles should look steeper when they are observed at higher resolution. This is exactly what is seen. If the original dynamical models are 'reobserved' at the improved resolution, the ones that agree with the new data have $M_\bullet = (1–2) \times 10^9 M_\odot$.

Finally, a definitive detection is provided by the HST observations of the nuclear star cluster. Its true velocity dispersion is underestimated in figure 2, because the projected value includes bulge light from in front of and behind the center. When this light is subtracted, the velocity dispersion of the nuclear cluster proves to be $\sigma = 600\pm37$ km s^{-1}. This is the highest dispersion measured in any galactic center. The velocity of escape from the nucleus would be much smaller, $V_{esc} \simeq 352$ km s^{-1}, if it consisted only of stars. Without extra mass to bind it, the cluster would fling itself apart in $\sim 2 \times 10^4$ yr. Independent of any velocity anisotropy, the nucleus must contain an unseen object of mass $M_\bullet \simeq 10^9 M_\odot$. This is consistent with the modeling results. The dark object is more than 25 times as massive as the visible star cluster. We know of no way to make a star cluster that is so nearly dark, especially without overenriching the visible stars with heavy elements. The most plausible

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

explanation is a BH. This would easily have been massive enough to power a quasar.

M31: M. $\simeq 3 \times 10^7 M_\odot$

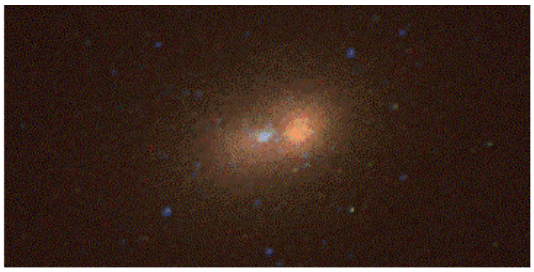
M31 is the highest-luminosity galaxy in the LOCAL GROUP. At a distance of 0.77 Mpc, it is the nearest giant galaxy outside our own. It can therefore be studied in unusual detail.

M31 contains the nearest example of a nuclear star cluster embedded in a normal bulge. When examined with HST, the nucleus appears double (figure 3). This is very surprising. At a separation of $2r = 0".49 = 1.7$ pc, a relative velocity of 200 km s^{-1} implies a circular orbit period of 50 000 yr. If the nucleus consisted of two star clusters in orbit around each other, as figure 3 might suggest, then dynamical friction would make them merge within a few orbital times. Therefore it is unlikely that the simplest possible explanation is correct: we are not observing the last stages of the digestion of an accreted companion galaxy.

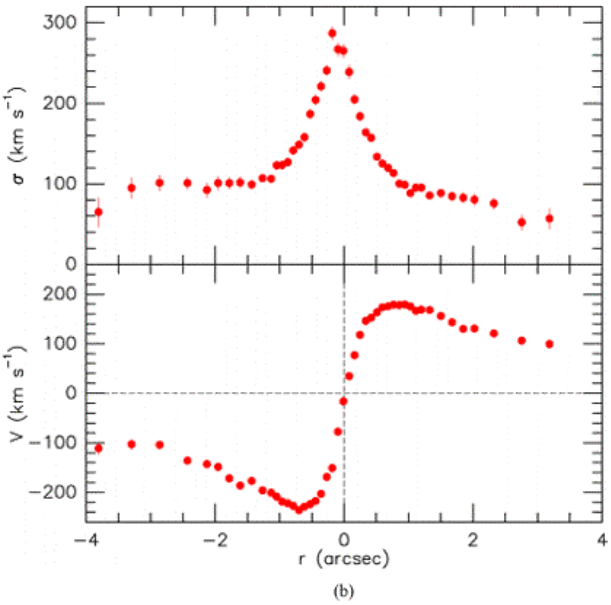
The nucleus rotates rapidly and has a steep velocity dispersion gradient (figure 3). Dynamical analysis shows that M31 contains a central dark mass $M_\bullet \simeq 3 \times 10^7 M_\odot$. The possible effects of velocity anisotropy have been checked and provide no escape. Furthermore, the asymmetry provides an almost independent check of the BH mass, as follows.

The top panel of figure 3 shows the HST image at the same scale as and registered in position with the kinematics. It shows that the dispersion peak is approximately centered on the fainter nucleus. In fact, it is centered almost exactly on a cluster of blue stars that is embedded in this nucleus. This suggests that the BH is in the blue cluster. This hypothesis can be tested by finding the center of mass of the asymmetric distribution of starlight plus a dark object in the blue cluster. The mass-to-light ratio of the stars is provided by dynamical models of the bulge at larger radii. If the galaxy is in equilibrium, then the center of mass should coincide with the center of the bulge. It does, provided that $M_\bullet \simeq 3 \times 10^7 M_\odot$. Remarkably, the same BH mass explains the kinematics and the asymmetry of the nucleus.

An explanation of the mysterious double nucleus has been proposed by Scott Tremaine. He suggests that both nuclei are part of a single eccentric disk of stars. The brighter nucleus is farther from the barycenter; it results from the lingering of stars near the apocenters of very elongated orbits. The fainter nucleus is produced by an increase in disk density toward the center. The model depends on the presence of a BH to make the potential almost Keplerian; then the alignment of orbits in the eccentric disk may be maintained by the disk's self-gravity. Tremaine's model was developed to explain the photometric and kinematic asymmetries as seen at



(a)



(b)

Figure 3. Top: HST WFPC2 color image of M31 constructed from *I*-, *V*- and 3000 Å band, PSF-deconvolved images obtained by Lauer *et al* (1998 *Astron. J.* **116** 2263). The scale is 0".0228 $pixel^{-1}$.Bottom and middle: Rotation curve $V(r)$ and velocity dispersion profile $\sigma(r)$ of the nucleus with foreground bulge light subtracted. The plots and image are printed at the same scale and are aligned in r. The symmetry point of the rotation curve and the sharp dispersion peak suggest that the BH is in the blue star cluster embedded in the left brightness peak. (This figure is adapted from Kormendy and Bender 1999 *Astrophys. J.* **522** 772.)

resolution $\sigma_* \simeq 0".5$. It is also consistent with the data in figure 3 ($\sigma_* \simeq 0".27$). The high velocity dispersion near the BH, the low dispersion in the offcenter nucleus and especially the asymmetric rotation curve are signatures of the eccentric, aligned orbits.

Most recently, spectroscopy of M31 has been obtained with the HST Faint Object Camera. This improves the spatial resolution by an additional factor of ~5. At this resolution, there is a 0".25 wide region centered on the faint nucleus in which the velocity dispersion is 440±70 km s^{-1}. This is further confirmation of the existence and location of the BH.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

We do not know whether the double nucleus is the cause or an effect of the offcenter BH. However, offcenter BHs are an inevitable consequence of hierarchical structure formation and galaxy mergers (see GALAXIES: INTERACTIONS AND MERGERS). If most large galaxies contain BHs, then mergers produce binary BHs and, in three-body encounters, BH ejections with recoil. How much offset we see, and indeed whether we see two BHs or one or none at all, depends on the relative rates of mergers, dynamical friction and binary orbit decay. Offcenter BHs may have much to tell us about these and other processes. Already there is evidence in NGC 4486B for a second double nucleus containing a BH.

Our Galaxy: $M_\bullet \simeq (2.9\pm0.4) \times 10^6 M_\odot$

Our Galaxy has long been known to contain the exceedingly compact radio source Sgr A* (see SAGITTARIUS A). Interferometry gives its major- and minor-axis diameters as $63r_s$ and $<17r_s$, where $r_s = 0.06$ AU $= 8.6 \times 10^{11}$ cm is the Schwarzschild radius of a $2.9 \times 10^6 M_\odot$ BH. It is easy to be impressed by the small size. However, as an AGN, Sgr A* is feeble: its radio luminosity is only 10^{34} erg s^{-1} $\simeq 10^{0.4} L_\odot$. The infrared and high-energy luminosities are higher, but there is no compelling need for a BH on energetic grounds. To find out whether the Galaxy contains a BH, we need dynamical evidence.

Getting it has not been easy. Our Galactic disk, which we see in the sky as the Milky Way, contains enough dust to block all but $\sim 10^{-14}$ of the optical light from the Galactic center. Measurements of the region around Sgr A* had to await the development of infrared detectors. Much of the infrared radiation is in turn absorbed by the Earth's atmosphere, but there is a useful transmission window at 2.2 μm wavelength. Here the extinction toward the Galactic center is a factor of ~20. This is large but manageable. Early infrared measurements showed a rotation velocity of $V \simeq 100$ km s^{-1} and a small rise in velocity dispersion to ~120 km s^{-1} at the center. These were best fitted with a BH of mass $M_\bullet \sim 10^6 M_\odot$, but the evidence was not very strong. Since then, a series of spectacular technical advances have made it possible to probe closer and closer to the center. As a result, the strongest case for a BH in any galaxy is now our own.

Most remarkably, two independent groups led by Reinhard Genzel and Andrea Ghez have used speckle imaging to measure proper motions—the velocity components perpendicular to the line of sight—in a cluster of stars at radii $r \lesssim 0".5 \simeq 0.02$ pc from Sgr A* (figure 4). When combined with complementary measurements at larger radii, the result is that the one-dimensional velocity dispersion increases smoothly to 420±60 km s^{-1} at $r \simeq 0.01$ pc. Stars at this radius revolve around the Galactic center in a human lifetime! The mass $M(r)$ inside radius r is shown in figure 5. Outside a few pc, the mass distribution is dominated by stars, but as $r \to 0$, $M(r)$ flattens to a constant, $M_\bullet = (2.9\pm0.4) \times 10^6 M_\odot$. Velocity anisotropy is not an uncertainty; it is measured directly and found to be small. The largest dark cluster that is consistent with these data would have a central density of $4 \times 10^{12} M_\odot$ pc^{-3}. This is inconsistent with astrophysical constraints (see below). Therefore, if the dark object is not a BH, the alternative would have to be comparably exotic. It is prudent to note that rigorous proof of a BH requires that we spatially resolve relativistic velocities near the Schwarzschild radius. This is not yet feasible. However, the case for a BH in our own Galaxy is now very compelling.

BH demographics

The census of BH candidates as of January 2000 is given in table 1. The table is divided into three groups—detections based on stellar dynamics, on ionized gas dynamics, and on maser disk dynamics (top to bottom). The rate of discovery is accelerating as HST pursues the search. However, we already have candidates that span the range of predicted masses and that occur in essentially every type of galaxy that is expected to contain a BH. Host galaxies include giant AGN ellipticals (the middle group), Seyfert galaxies (NGC 1068), normal spirals with moderately active nuclei (e.g. NGC 4594 and NGC 4258), galaxies with exceedingly weak nuclear activity (our Galaxy and M31) and completely inactive galaxies (M32 and NGC 3115).

However, no complete sample has been studied at high resolution. The detections in table 1, together with low-resolution studies of larger samples of galaxies, support the hypothesis that BHs live in virtually every galaxy with a substantial bulge component. The total mass in detected remnants is consistent with predictions based on AGN energetics, within the rather large estimated errors in both quantities.

The main new demographic result is an apparent correlation between BH mass and the luminosity of the bulge part of the galaxy. This is shown in figure 6. Note that the correlation is not with the total luminosity: if the disk is included, the correlation is considerably worse. Whether the correlation is real or not is still being tested. The concern is selection effects. High-mass BHs in small galaxies are easy to see, so their scarcity is real. However, low-mass BHs can hide in giant galaxies, so the correlation may be only the upper envelope of a distribution that extends to smaller $M_\bullet$. If it is real, then the correlation implies that BH formation or feeding is connected with the mass of the high-density, elliptical-galaxy-like part of the galaxy. With the possible exception of NGC 4945 (a late-type galaxy for which the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

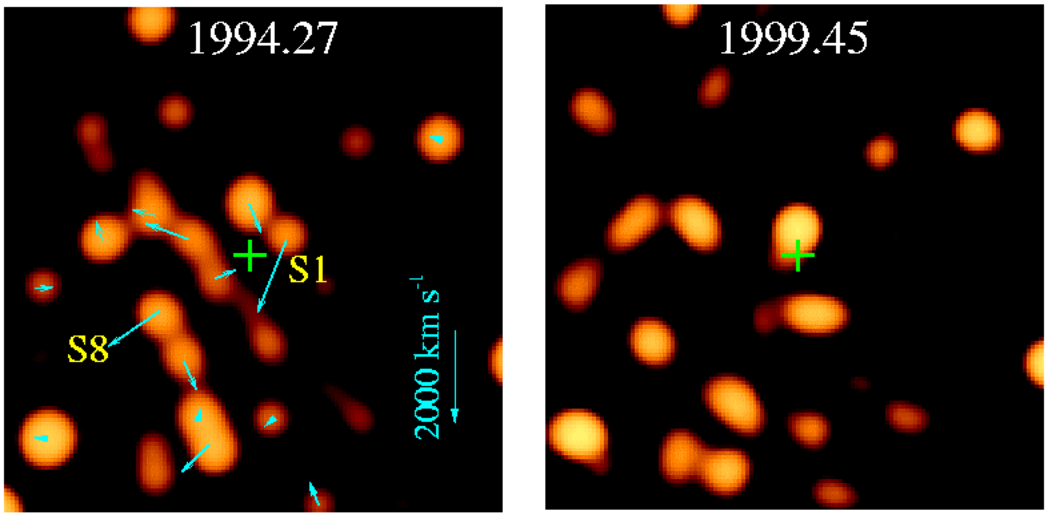


Figure 4. Images of the star cluster surrounding Sgr A* (green cross) at the epochs indicated. The arrows in the left frame show approximately where the stars have moved in the right frame. Star S1 has a total proper motion of ~1600 km s^{-1}. (This figure is updated from Eckart and Genzel 1997 *Mon. Not. R. Astron. Soc.* **284** 576 and was kindly provided by A Eckart.)

Table 1. Census of BH candidates.

Galaxy	Type[a]	D[b] (Mpc)	$M_{B,\mathrm{bulge}}$[c]	$M_\bullet$[d] ($M_\odot$)	$\log M_\bullet/M_{\mathrm{bulge}}$[e]
Galaxy	Sbc	0.0085	–17.65	3×10^6	–3.62
M31	Sb	0.7	–18.82	3×10^7	–3.31
M32	E	0.7	–15.51	3×10^6	–2.27
NGC 3115	S0/	8.4	–19.90	1×10^9	–1.92
NGC 4594	Sa/	9.2	–21.21	1×10^9	–2.69
NGC 3377	E	9.9	–18.80	8×10^7	–2.24
NGC 3379	E	9.9	–19.79	1×10^8	–2.96
NGC 4342	S0	15.3	–17.04	3×10^8	–1.64
NGC 4486B	E	15.3	–16.66	6×10^8	–1.03
M87	E	15.3	–21.42	3×10^9	–2.32
NGC 4374	E	15.3	–20.96	1×10^9	–2.53
NGC 4261	E	29	–20.89	5×10^8	–2.92
NGC 7052	E	59	–21.31	3×10^8	–3.31
NGC 6251	E	106	–21.81	6×10^8	–3.18
NGC 4945	Scd/	3.7	–15.1	1×10^6	–
NGC 4258	Sbc	7.5	–17.3	4×10^7	–2.05
NGC 1068	Sb	15	–18.8	1×10^7	–

[a] Hubble type; / means that the galaxy is edge on.
[b] Distance based on a Hubble constant of 80 km s^{-1} Mpc^{-1}.
[c] Absolute *B*-band magnitude of the bulge component of the galaxy
[d] BH mass based on isotropic models
[e] Ratio of BH mass to bulge mass. The mass in stars is calculated from the luminosity via the mass-to-light ratio measured at large radii.

existence and luminosity of a bulge are uncertain), BHs have been found only in the presence of a bulge. However, the limits on $M_\bullet$ in bulgeless galaxies such as M33 are still consistent with the correlation. Current searches concentrate on the question of whether small BHs—ones that are significantly below the apparent correlation—can be found or excluded.

BH mass fractions are listed in table 1 for cases in which the mass-to-light ratio of the stars has been measured. The median BH mass fraction is 0.29%. The quartiles are 0.07% and 0.9%.

Are they really BHs?

The discovery of dark objects with masses $M_\bullet \simeq (10^6$–$10^{9.5})\ M_\odot$ in galactic nuclei is secure. However, are they BHs? Proof requires measurement of relativistic velocities

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

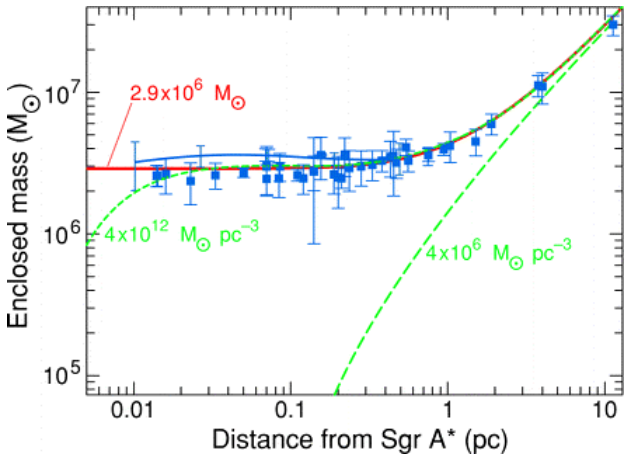


Figure 5. Mass distribution implied by proper motion and radial velocity measurements (data points and upper solid curve with representative error bars). Long dashes (right) show the mass distribution of stars if the infrared mass-to-light ratio is 2. The heavy solid curve represents the stars plus a point mass $M_\bullet = 2.9 \times 10^6 M_\odot$. Short dashes (left) provide an estimate of how non-pointlike the dark mass could be: its χ^2 value is 1 σ worse than the solid curve. This dark cluster has a core radius of 0.0042 pc and a central density of $4 \times 10^{12} M_\odot$ pc^{-3}. (This figure is updated from Genzel *et al* 1997 *Mon. Not. R. Astron. Soc.* **291** 219 and was kindly provided by R Genzel.)

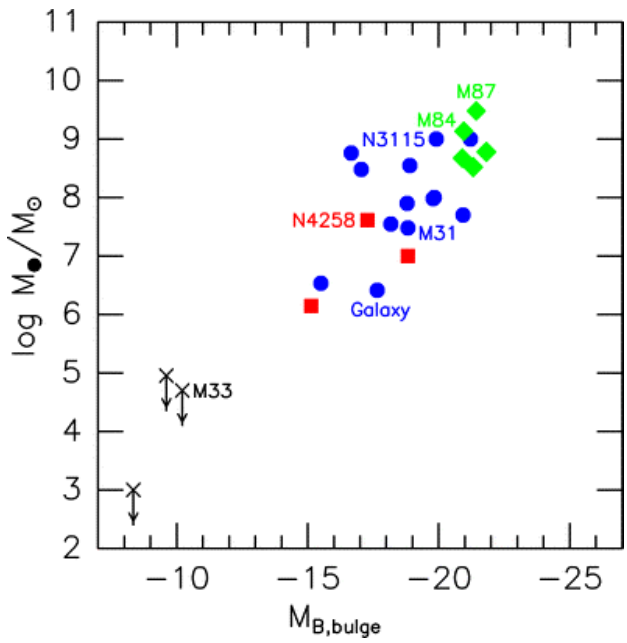


Figure 6. Correlation of BH mass with the absolute magnitude of the bulge component of the host galaxy. Since M/L varies little from bulge to bulge, this implies a correlation between BH mass and bulge mass. Filled circles indicate $M_\bullet$ measurements based on stellar dynamics, diamonds are based on ionized gas dynamics and squares are based on maser disk dynamics. It is reassuring that all three techniques are consistent with the same correlation.

velocities near the Schwarzschild radius, $r_s \simeq 2 M_\bullet/10^8 M_\odot$ AU. Even for M31, this is 8×10^{-7} arcsec. HST spectroscopic resolution is only 0".1. The conclusion that we are finding BHs is based on physical arguments that BH alternatives fail to explain the masses and high densities of galactic nuclei.

The most plausible BH alternatives are clusters of dark objects produced by ordinary STELLAR EVOLUTION. These come in two varieties, failed stars and dead stars. Failed stars have masses $m_* \lesssim 0.08 M_\odot$. They never become hot enough for the fusion reactions that power stars, i.e. the conversion of hydrogen to helium. They have a brief phase of modest brightness while they live off gravitational potential energy, but, after this, they could be used to make dark clusters. They are called brown dwarf stars, and they include planetary-mass objects. Alternatively, a dark cluster could be made of stellar remnants—WHITE DWARFS, which have typical masses of $0.6 M_\odot$; NEUTRON STARS, which typically have masses of $\sim 1.4 M_\odot$ and BHs with masses of several $M_\odot$. GALACTIC BULGES are believed to form in violent starbursts, so massive stars that turn quickly into dark remnants would be no surprise. It is not clear how one could make dark clusters with the required masses and sizes, especially not without polluting the remaining stars with more metals than we see. However, in the absence of direct proof that the dark objects in galactic nuclei are BHs, it is important to examine alternatives.

However, dynamical measurements tell us more than the mass of a potential BH. They also constrain the maximum radius inside which the dark stuff must live. Its minimum density must therefore be high, and this rules out the above BH alternatives in our Galaxy and in NGC 4258. High-mass remnants such as white dwarfs, neutron stars and stellar BHs would be relatively few in number. The dynamical evolution of star clusters is relatively well understood; in the above galaxies, a sparse cluster of stellar remnants would evaporate completely in $\lesssim 10^8$ yr. Low-mass objects such as brown dwarfs would be so numerous that collision times would be short. Stars generally merge when they collide. A dark cluster of low-mass objects would become luminous because brown dwarfs would turn into stars.

More exotic BH alternatives are not ruled out by such arguments. For example, the dark matter that makes up galactic halos and that accounts for most of the mass of the universe may in part be elementary particles that are cold enough to cluster easily. It is not out of the question that a cluster of these could explain the dark objects in galaxy centers without getting into trouble with any astrophysical constraints. So the BH case is not rigorously proved. What makes it compelling is the combination of dynamical evidence and the evidence

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

from AGN observations. This is discussed in the previous article.

For many years, AGN observations were decoupled from the dynamical evidence for BHs. This is no longer the case. Dynamical BH detections are routine. The search itself is no longer the main preoccupation; we can concentrate on physical questions. New technical developments such as better x-ray satellites ensure that progress on BH astrophysics will continue to accelerate.

Web update

The Chandra X-ray telescope has observed a flare of x-ray energy produced from where the lip of a supermassive black hole at the center of the Milky Way galaxy should be. The flare dimmed and brightened over a period of ten minutes, the time taken for light to travel about 150 million kilometers, meaning that it originated from an object no larger in diameter than the distance of the Earth to the Sun. This implies that the mass at the center of our Galaxy (about 2.6 million times the mass of our Sun) is crammed into that space. According to the laws of general relativity, only a black hole can be that dense.

Bibliography

The search for BHs is reviewed in the following papers:

Kormendy J and Richstone D 1995 *Annu. Rev. Astron. Astrophys.* **33** 581
Richstone D *et al* 1998 *Nature* **395** A14

In the following papers, quasar energetics are used to predict the masses of dead AGN engines:

Chokshi A and Turner E L 1992 *Mon. Not. R. Astron. Soc.* **259** 421
Sołtan A 1982 *Mon. Not. R. Astron. Soc.* **200** 115

Dynamical models of galaxies as linear combinations of individual orbits are discussed in

Gebhardt K *et al* 2000 *Astron. J.* **119** 1157
Richstone D O and Tremaine S 1988 *Astrophys. J.* **327** 82
Schwarzschild M 1979 *Astrophys. J.* **232** 236
van der Marel R P, Cretton N, de Zeeuw P T and Rix H-W 1998 *Astrophys. J.* **493** 613

The BH detection in NGC 3115 is discussed in

Kormendy J and Richstone D 1992 *Astrophys. J.* **393** 559
Kormendy J *et al* 1996 *Astrophys. J. Lett.* **459** L57

The BH detection in M31 is discussed in

Dressler A and Richstone D O 1988 *Astrophys. J.* **324** 701
Kormendy J 1988 *Astrophys. J.* **325** 128

Tremaine's model for the double nucleus of M31 and new evidence for that model are in

Kormendy J and Bender R 1999 *Astrophys. J.* **522** 772
Tremaine S 1995 *Astron. J.* **110** 628

HST spectroscopy of the double nucleus of M31 is presented in

Statler T S, King I R, Crane P and Jedrzejewski R I 1999 *Astron. J.* **117** 894

The following are thorough reviews of the Galactic center:

Genzel R, Hollenbach D and Townes C H 1994 *Rep. Prog. Phys.* **57** 417
Morris M and Serabyn E 1996 *Annu. Rev. Astron. Astrophys.* **34** 645

The latest measurement of the size of the Galactic center radio source is by

Lo K Y, Shen Z-Q, Zhao J-H and Ho P T P 1998 *Astrophys. J. Lett.* **508** L61

The remarkable proper motion measurements of stars near Sgr A* and resulting conclusions about the Galactic center BH are presented in

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219
Ghez A M, Klein B L, Morris M and Becklin E E 1998 *Astrophys. J.* **509** 678

Arguments against compact dark star clusters in NGC 4258 and the Galaxy are presented in

Maoz E 1998 *Astrophys. J. Lett.* **494** L181

John Kormendy

Luis C Ho

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Supernova 1987A

Supernova 1987A (SN1987A) in the LARGE MAGELLANIC CLOUD (LMC) is the brightest supernova to be observed since SN1604 (Kepler), the first to be observed in every band of the ELECTROMAGNETIC SPECTRUM and the first to be detected through its initial burst of NEUTRINOS. Although the bolometric luminosity of SN1987A today is $\approx 10^{-6}$ of its value at maximum light ($L_{max} \approx 2.5 \times 10^8 L_\odot$), it will remain bright enough to be observed for many decades in the radio, infrared, optical, UV and x-ray bands.

SN1987A is classified as a type II supernova (SNeII) (see SUPERNOVAE) by virtue of the strong hydrogen lines in its spectrum. It was atypical of SNeII in that its light curve did not reach maximum until 3 months after outburst and its maximum luminosity was about 1/10 the mean maximum luminosity of SNeII. These differences can be attributed to the fact that the star that exploded was a blue giant, unlike the progenitors of most SNeII, which we believe to be RED GIANTS.

The burst of neutrinos observed from SN1987A proved beyond doubt that its explosion followed the collapse of the core of the star, but subsequent observations have shown no evidence of the neutron star or black hole that we expect to find at the center of the debris.

The expanding gaseous debris of SN1987A cooled rapidly after the explosion. By 1000 days, its temperature was less than 2000 K throughout and the heavy elements had formed molecules and dust. With a present temperature <100 K, the inner radioactive debris is perhaps the coldest optically emitting source known to astronomers.

SN1987A is surrounded by a remarkable NEBULA of three glowing circumstellar rings (figure 1) that are evidently part of a complex system of gas that was ejected by the progenitor some 20 000 yr before it exploded and then partly ionized by the supernova flash. The approximate bipolar symmetry of this system suggests that the progenitor was once a binary system, and this interpretation may also explain why the progenitor was a blue giant rather than a red giant.

Figure 1. Circumstellar rings of SN1987A.

The outer part of the supernova debris is driving a blast wave that is beginning to strike the inner circumstellar ring at several 'hot spots' that are doubling in brightness every few months. Within the coming decade, more hotspots will appear and merge as the blast wave overtakes the entire inner ring. The ring will brighten by factors $\geqq 10^3$ at all bands of the electromagnetic spectrum, giving us unprecedented opportunities to study the physics of interstellar shocks and to view heretofore unseen circumstellar matter.

Discovery and neutrino observations

Supernova 1987A was discovered with magnitude $V \approx 5$ by Ian Shelton on February 24.23 UT, 1987. The supernova occurred in the Large Magellanic Cloud near the bright star-forming nebula 30 Doradus. Within days, several groups had obtained optical spectra and the International Ultraviolet Explorer Satellite had obtained ultraviolet spectra, showing that the outer envelope of the supernova was expanding with velocities ~0.1c (30 000 km s^{-1}). Subsequently, earlier observations were found showing that the object had magnitude $V \approx 6.5$ on February 23.443 and $V \geqq 7.5$ on February 23.39.

Soon thereafter, two independent groups reported detections of a flash of neutrinos from the direction of SN1987A that occurred on February 23.316. Electron antineutrinos produce fast positrons through the reaction $\bar{\nu}_e + p \rightarrow n + e^+$. These events are detected in large underground tanks of water through the Cherenkov light emitted by the positrons. The Kamiokande II experiment in Japan detected 12 such events and the IMB experiment in the USA detected eight events over an interval of about 12 s, during which a net background count of one event from the two detectors would be expected. The inferred energy (~3×10^{53} erg), temperature (kT~4 MeV), and decay timescale (~4 s) of the neutrino flash were in remarkably good agreement with predictions from theoretical models in which a degenerate iron core collapsed to form a NEUTRON STAR.

The initial optical display should follow the neutrino flash after a delay corresponding to the time for a blast wave to propagate through the envelope of the progenitor star. The fact that this delay was no greater than 3 h sets an upper limit $R^* \leqq 100 R_\odot$ for the radius of the progenitor star, precluding the possibility that the progenitor was a red supergiant.

Observations taken before the explosion of SN1987A showed that the progenitor star, Sk–69°202, was in fact a spectral type B2 Ia supergiant, with visual magnitude $V = 12$, bolometric luminosity $\approx 10^5 L_\odot$ and radius $R^* \approx 40 R_\odot$

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

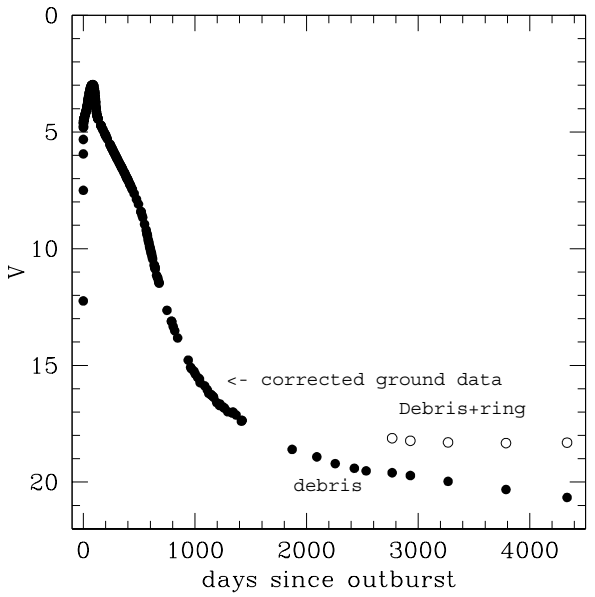


Figure 2. Light curve of SN1987A.

Light curve

SN1987A is unique in that we have recorded its light from well before it exploded until the present time ($t \approx 13$ yr). The light curve of SN1987A (figure 2) was unusual. Within hours after first detection, it brightened to $V \approx 4$. Thereafter, it continued to brighten, reaching a maximum brightness $V = 2.9$, corresponding to a bolometric luminosity $L_{\rm bol} = 10^{42}$ erg s^{-1}, at $t \approx 3$ months after explosion.

SN1987A faded rapidly thereafter. For 120 days $< t <$ 800 days, $L_{\rm bol} : e(-t/t_1)$, where t_1 agrees within a few percent with the 111.3 d mean life of ^{56}Co. This observation supports the interpretation that the supernova light is produced by reprocessing of energy deposited in the explosion debris by positrons and gamma rays from the radioactive decay of ^{56}Co. NUCLEOSYNTHESIS models predicted that a core collapse supernova explosion would produce a substantial mass of ^{56}Ni, which would decay into ^{56}Co with a mean life of 8.8 days. The observed bolometric luminosity of SN1987A during its exponential decay phase implies that the explosion debris originally contained $0.07 M_\odot$ of ^{56}Ni.

This interpretation was confirmed by the direct detection of ^{56}Co gamma ray emission lines at 847 keV and 1238 keV, which became bright enough to observe by $t \approx 200$ days. The fraction of the bolometric luminosity that emerged directly as gamma rays increased from ~1% at $t \approx 200$ days to ~50% at $t \gtrsim 800$ days. The emergent gamma rays were accompanied by a flux of hard ($\gtrsim$20 keV) x-rays that could be attributed to gamma rays degraded in energy by multiple Compton scattering in the supernova debris.

As the ^{56}Co decayed, longer-lived isotopes dominated the radioactive energy input to the debris, principally ^{57}Co (mean life 391 days) for 1000 days $\lesssim t \lesssim$ 1700 days and ^{44}Ti (mean life 78 yr) for $t \gtrsim 1700$ days.

Today, the radioactive debris of SN1987A has magnitude $V \approx 20.5$ and bolometric luminosity $L_{\rm bol} \approx 10^{36}$ erg s^{-1} (i.e. $L_{\rm bol} \approx 200 L_\odot$). Most of this luminosity emerges in the far infrared (10–100 μm) band.

The integrated bolometric luminosity of SN1987A, $E_{\rm rad} \approx 8.6 \times 10^{48}$ erg, is dominated by the radioactive decay of $0.07 M_\odot$ of ^{56}Co. (Most of the energy deposited by the earlier decay of ^{56}Ni was converted to kinetic energy of the expanding envelope.)

Models for the optical, x-ray and gamma ray light curves require that the supernova debris have a mass ~(5–15)$M_\odot$, most of which must be hydrogen and helium.

The newly synthesized elements remained mostly in the inner part ($V \lesssim 2500$ km s^{-1}) of the supernova debris. However, the early emergence of gamma rays implies that a small fraction of the newly synthesized ^{56}Ni was deposited near the outer layers of the supernova debris. Evidently, instabilities developed during the propagation of the supernova blast through the stellar envelope, disrupting the laminar segregation of nucleosynthesis products that may have existed before the explosion.

Evolution of the debris

The optical display of the supernova commenced when the blast reached the photosphere of the progenitor, ~2 h after the core explosion. As the blast emerged, the photosphere was heated to ~10^6 K and produced a flash of $\approx 10^{57}$ ionizing photons during the following several hours.

At $t \approx 1.6$ days, the spectrum of SN1987A was dominated by a blue continuum, with color temperature $T^{\rm C} \approx 14\,000$ K, punctuated by strong P-Cygni lines of hydrogen and MgII λλ2800. The blueshift of the MgII λλ2800 absorption feature indicated that the gas immediately above the photosphere was expanding with radial velocities ~30 000 km s^{-1}. The photosphere cooled rapidly thereafter, with color temperature stabilizing at ~5500 K after about 2 weeks. The maximum observed blueshift in the absorption lines decreased rapidly as the supernova expanded, indicating that the photosphere was moving inward with respect to the debris and that the gas above the photosphere was becoming transparent. The radius of the photosphere reached a maximum ~100 AU at $t \approx 3$ months, at which time the absorption lines indicated a radial expansion velocity ≈2500 km s^{-1}.

From observations of the absorption lines during the photospheric phase ($t \lesssim 4$ months), one can infer that the debris has kinetic energy $E_{\rm kin} \approx 1.5 \times 10^{51}$ erg, roughly 200 times $E_{\rm rad}$.

By $t \approx 4$ months, the photosphere had vanished and the nebular phase commenced, in which the optical spectrum was dominated by emission lines, principally Hα, [OI] λλ 6300, [Ca II] λλ 7300 and Ca II λλ 8600. The width

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

(FWHM) of the emission lines implied that most of the optical emission was produced within a sphere expanding with velocity ~2500 km s^{-1}, evidently the volume in which most of the radioactive energy is deposited. By this time, the outer part of the debris, which contains more than half the mass and most of the kinetic energy, had become invisible at optical wavelengths.

The inner debris cooled rapidly, as indicated by infrared emission from vibrationally excited CO, which first appeared in the spectrum at $t \approx 4$ months and remained prominent for several months thereafter.

At $t \approx 1$ yr, the optical and near-infrared emission lines faded rapidly and a strong far-infrared ($\lesssim 4$ μm) continuum appeared. Evidently, dust clouds formed in the inner debris and obscured roughly half of the optical and near-infrared emission from the emitting region, re-emitting the absorbed luminosity as far-infrared continuum. The fact that the redshifted sides of the emission lines (which come from the far hemisphere of the line-emitting region) are obscured more heavily than the blueshifted sides (from the near hemisphere) implies that the dust clouds are located within the line-emitting region. As yet, no spectral features have been detected in the absorption or emission by the dust that would enable us to recognize its composition.

By analyzing the emission line spectrum during the nebular phase, one can infer the physical conditions in the emission line region. The analysis shows clearly that the debris is chemically heterogeneous, with regions of different element composition having different evolutions of density and temperature. The regions composed primarily of oxygen and carbon cooled first as a result of infrared emission by CO molecules. By $t \approx 4$ yr, most of the line-emitting region had temperatures <200 K, except the helium- and hydrogen-rich zones, which had temperatures ~500 K and ~700 K, respectively. Today, the entire line-emitting region has temperature <100 K, and the optical and near-infrared emission lines are produced mainly by delayed recombination and non-thermal positrons from ^{44}Ti decay.

The fact that optically thick infrared emission lines from Fe, Co and Ni were bright at early times ($t \sim 1$ yr) implies that the zones containing newly synthesized ^{56}Ni and its decay products must occupy ~50% of the line-emitting region. Since these elements constitute only ~1–2% of the mass of this region, we conclude that the line-emitting region must have a foamy texture, in which bubbles of relatively low-density Ni–Co–Fe gas are embedded in a more dense substrate of C–O, H, He and other elements. This foamy structure is caused by radioactive heating and expansion of the Fe–Co–Ni gas due to ^{56}Ni decay during the first few weeks after the explosion.

Today, the debris has been resolved with the Hubble Space Telescope. The emitting region has a mean angular diameter ~0.25 arcsec, consistent with a mean expansion velocity ~2500 km s^{-1}. The image is slightly (~1.2:1) elongated with the major axis oriented roughly NS. It has an irregular morphology, probably because of differential obscuration by embedded dust clouds.

Circumstellar matter and progenitor star

Beginning at t=80 days after outburst, narrow (FWHM $\lesssim$20 km s^{-1}) optical and UV emission lines of Hα and more highly ionized species such as He II, N V, N IV, N III, O III and C III began to appear in the spectrum of SN1987A. These emission lines brightened to maximum luminosity $L \approx 3.5 \times 10^{36}$ erg s^{-1} by t=400 days and faded rapidly thereafter. Images taken later with the Hubble Space Telescope (figure 1) show that the source of these emission lines is a narrow elliptical ring centered on the supernova, with major (EW) and minor (NS) diameters of 1.66 arcsec and 1.21 arcsec respectively. It is now clear that this inner ring is actually a nearly circular ring of emitting gas of radius 0.67 ly, inclined at about 45°. This inner ring began to produce narrow emission lines when it was illuminated by the flash of ionizing radiation produced during the first few hours after the supernova explosion.

The narrow emission lines reached maximum light after a delay corresponding to the difference of path length from the supernova to the far (S) side of the tilted ring and then to Earth and the path directly from the supernova to Earth. This delay enables us to infer the physical diameter of the ring, which we may compare with the angular diameter to infer the distance of the supernova. The result is $D \approx 51 \pm 3$ kpc.

The inner ring is expanding, with radial velocity ~ ≈10 km s^{-1}. Assuming that the ring was expelled by the progenitor star, we estimate that this event occurred ≈20000 yr before the supernova event.

Analysis of the emission line spectrum of the inner rings shows that the abundance ratio of (C + N + O)/H is about 0.3 times the solar ratio, consistent with LMC abundances. However, the rings are nitrogen rich, with N/(C + O) about 10 times greater than the solar ratio. The enhanced nitrogen abundance is believed to be due to conversion of C and O to N through the CNO cycle in the supernova progenitor before the circumstellar ring was expelled.

After being ionized and heated by the initial supernova flash, the circumstellar rings have faded as the gas has cooled and recombined. Today, the inner ring has magnitude $V \approx 18.5$, about a factor of 10 fainter than it was at maximum light (t=400 days). From the fading rate of the emission lines, we infer that the gas in the rings has atomic densities in the range ~$(0.6–3.3) \times 10^4$ cm^{-3}, and from the brightness we infer that the glowing gas in the inner ring has mass ~$0.04 M_\odot$.

As well as the inner ring, SN1987A is surrounded by an even more remarkable pair of outer loops that are not

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

concentric with the supernova. We believe that these loops are actually also circular rings with approximately the same polar axis as the inner ring, but not coplanar with the inner ring. If so, the outer loops have physical radii ~1.5 ly and lie on planes approximately parallel to the plane of the inner ring but displaced by ±~1.3 ly.

The radial velocities of the outer loops are consistent with the interpretation that they too were expelled ≈20000 yr before the supernova explosion.

The origin of the triple ring system remains a mystery. Since we see similar morphologies in some PLANETARY NEBULAE, many of which have binary central stars, we suspect that the progenitor of SN1987A was also a binary system. Perhaps the rings were expelled when these stars merged. The merged binary hypothesis may also explain why the progenitor of SN1987A was a blue giant star, in contrast to most known SNeII, which result from the explosions of red giant stars.

The glowing ring system is only part of the complex and as yet poorly understood distribution of circumstellar gas around the supernova. About 2 yr after the explosion, diffuse nebulosity was detected beyond the outer loops. The most prominent part was an arc called 'Napoleon's Hat', the top of which is about 5 arcsec N of the supernova and the brim of which extends about ±5 arcsec E and W of the supernova. These diffuse features changed rapidly and vanished after about 4 yr. They are light echoes caused by scattering of the supernova light by dust grains. The dusty gas was probably expelled by the supernova progenitor when it was a red giant star. In contrast to the circumstellar rings, we can now see no trace of the gas associated with the light echoes.

The initial flash of the supernova was sufficient to ionize only ~1$M_{\odot}$ of circumstellar matter. Therefore, it is likely that the glowing gas we see in the circumstellar rings is ionization bounded and resides on the inner surface of a greater mass of still invisible gas. The key to understanding the origin of the circumstellar rings may be in the detection of this gas.

In addition to the echoes from dust nearby the supernova, astronomers detected several much larger (angular radii ~0.5–1 arcmin) expanding rings of light after about 1 yr that remained visible for several years. From the apparent superluminal velocities, one can infer the distances from the supernova of the intervening dust clouds. The clouds responsible for the brightest two echoes are at distances ~130 and ~360 pc. They have no causal connection to the supernova.

From supernova to supernova remnant

At t=2 days, non-thermal radio emission was detected from SN1987A, evidently the result of acceleration of relativistic electrons accelerated due to shocks produced by the supernova blast in the wind of the blue giant progenitor. This radio emission reached maximum at t=4 days and then faded from view after a few weeks.

Non-thermal radio emission reappeared at t≈1250 days after explosion and has continued to brighten ever since. The radio image shows that the radio source is concentrated in two lobes to the east and west of the supernova. The centers of the lobes are located between the supernova and the inner ring and they are moving with transverse velocities of ~3000 km s^{-1}.

At nearly the same time, a steadily increasing flux of soft x-rays was detected from SN1987A. Observations with the CHANDRA observatory in October 1999 showed that the x-ray image was also an annulus, roughly coincident with the radio image. Like the radio image, the x-ray image is brightest toward the east. See figure 3.

Evidently, the radio and x-ray emissions are the result of the interaction of the supernova blast with circumstellar matter inside the inner ring. Models to account for the x-ray flux from the shocked gas require that the circumstellar matter must have atomic density ~100 cm^{-3}.

Freely expanding neutral hydrogen atoms in the outer supernova debris emit Lyα and Hα as they cross a reverse shock and enter the interaction zone. Imaging and spectroscopic observations of this emission with the Hubble Space Telescope show that the reverse shock is located at a radius ~75% of the inner circumstellar ring and that the emitting atoms have radial velocities ~15 000 km s^{-1}.

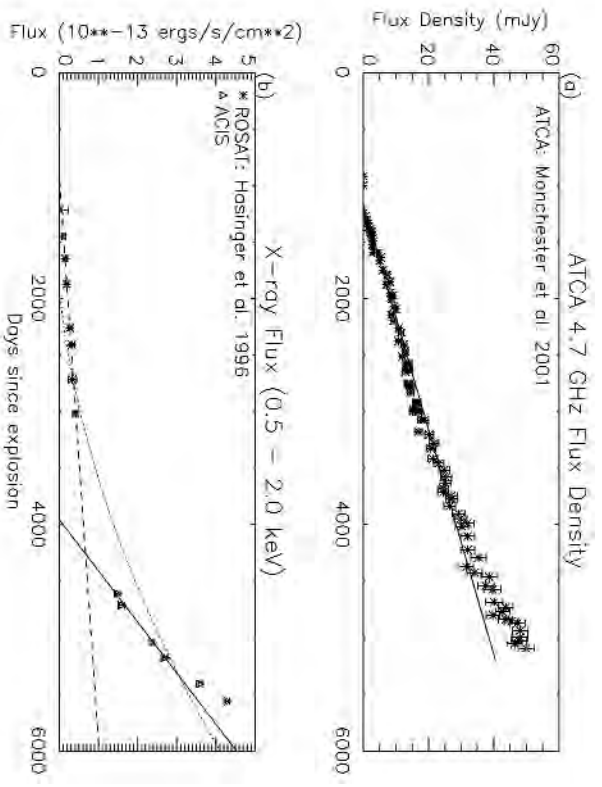


Figure 3. Radio (upper) and x-ray (lower) light curves.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

In 1995, a 'hot spot' appeared on the inner ring. It has brightened steadily ever since. Beginning in November 1998, many new hot spots began to appear on the inner ring (figure 4) and to brighten rapidly, with e-folding timescales ranging from ~1–2 years. Doppler shifts and widths of emission lines from the hotspots indicate that they are caused by shocks with velocity ~250 km s^{-1}.

The hot shocked gas drives a blast wave ahead of the reverse shock. Evidently, this blast wave is entering the circumstellar ring at the hot spots, which are inward protrusions of the ring. More hot spots will appear. Several years from now, they will merge to cover the entire inner ring.

The slope of the x-ray light curve increased rapidly sometime around 1998 (figure 3). Moreover, recent (May 2002) observations with the Chandra observatory show that the x-ray image is brightest near the optical hot spots, suggesting that the x-ray emission is now dominated by the impact of the blast wave with the circumstellar ring. The x-rays cannot come from exactly the same place as the optical emission, however, because the shocks responsible for the optical emission are too slow to produce observable x-ray emission. Evidently, shock velocities ranging from ~250–2500 km s^{-1} exist near the hot spots.

The appearance of these hotspots heralds the birth of SUPERNOVA REMNANT 1987A, the stage when its luminosity is dominated by the conversion of the kinetic energy of the debris to radiation as a result of interaction with circumstellar matter. During this stage, the supernova will brighten by factors ~10^3 in radio, infrared and x-rays as well and will remain bright for many decades. After a few years, the number of ionizing photons from the shock interaction will be comparable with the initial flash of the supernova. These photons will cause heretofore unseen circumstellar matter to become visible, giving us a new opportunity to understand the complex history of mass loss from the progenitor star.

Sometime between AD 2020 and AD 2030, we will begin to see sudden changes in the optical, ultraviolet, and x-ray spectra of SN1987A, as the newly synthesized elements in the inner debris begin to cross the reverse shock. Analysis of the subsequent evolution of the spectra will give us a 3-dimensional picture of the distribution of supernova nucleosynthesis products in the supernova debris.

The compact object

Perhaps the outstanding mystery of SN1987A is the absence of any evidence (since the initial neutrino flash) for a compact object at its center. The bolometric (mostly infrared) luminosity of the entire envelope is now $L_{\rm env} \lesssim$ few$\times 100 L_\odot$, and a brighter compact object could not have escaped detection. If the compact object is a neutron star, it must be very faint—perhaps because it has an anomalously low magnetic field and/or spin rate—and it must accrete $\lesssim 10^{-11} M_\odot\,{\rm yr}^{-1}$.

Although the duration of the neutrino flash indicates that a neutron star formed during the supernova explosion, it is possible that enough matter fell back in the subsequent hours or days to cause it to collapse into a BLACK HOLE.

There is little chance of detecting a compact object in the SN1987A with radio ($\gtrsim$1 cm), ultraviolet or x-ray observations, since the debris will probably remain opaque in these bands for decades. The compact object might also remain obscured at optical and infrared wavelengths if it happens to lie behind one of the dust clouds in the inner debris. The best chance might be to detect pulsations at millimeter wavelengths, but this will require a telescope array with much better angular resolution and sensitivity than those currently available.

Bibliography

Arnett W D, Bahcall J N, Kirshner R P and Woosley S E 1989 Supernova 1987A *Annu. Rev. Astron. Astrophys.* **27** 629–700

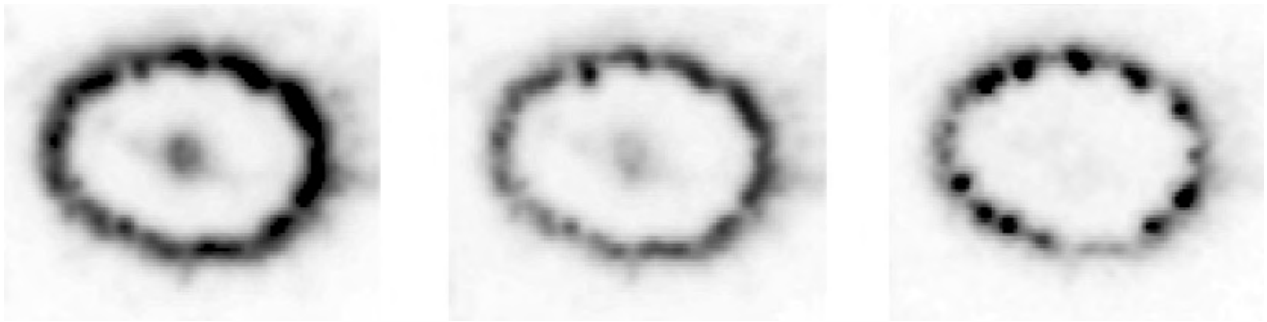


Figure 4. WFPC narrow-band Hα images of the inner ring. Left: February 1996. Center: February 1998. Right: May 2002. Both the inner ring and the central supernova debris have faded monotonically, while Spot–1, located at P.A. 029°, has evidently brightened by February 1998. By May 2002, 16 hot spots have appeared.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

McCray R 1993 Supernova 1987A revisited *Ann. Rev. Astron. Astrophys.* **31** 175–216
McCray R 2002 Supernova 1987A *Supernovae and gamma-ray bursters* ed Weiler K W (Springer: New York)

Richard McCray

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Supernova Remnants

A supernova remnant (SNR) is the aftermath of a stellar explosion, including interstellar gas swept up in a shock wave, the expanding stellar debris and in some cases a compact stellar remnant. Some SNRs are powered by pulsars, while others show no sign of a neutron star or black hole. The typical kinetic energy of an SN explosion is 10^{51} erg. A strong shock wave transforms much of this energy into thermal energy and cosmic ray energy. As the SNR ages, more and more of the thermal energy is lost to radiative cooling, the shock slows down, and the shell of shocked gas merges back into the interstellar medium (ISM). During the course of this evolution, an SNR can be observed as an expanding shell of hot gas containing many solar masses and emitting as much as 10^{39} erg s^{-1} throughout the radio to γ-ray wavelength range.

SNRs disperse the elements produced in the cores of the pre-supernova stars, along with those produced by explosive nucleosynthesis, through the ISM. Thus the synthesized elements are available for a new generation of star formation. Some of these elements may take the form of dust grains, and others will occur as neutrals or low-charge ions. During the course of its evolution, the interaction of the SNR with the ISM reheats the SN ejecta, disrupting dust grains and ionizing the ejected material.

The energy of SNRs is important for the overall structure and evolution of the ISM. The fast shock waves of a young SNR heat the interstellar gas to millions of kelvins, and it expands to produce a vast bubble of tenuous gas. When the shock speed falls below about 300 km s^{-1}, radiative cooling becomes effective in removing the thermal energy. The remaining energy goes into compressing the magnetic field, turbulent motion or cosmic ray acceleration.

SNRs are often observed as shells of hot gas a few parsecs to tens of parsecs in size. They produce x-rays at 100–10 000 eV photon energies by electron–ion collisions in hot gas or by synchrotron emission. Many SNRs have been discovered through their polarized radio emission, and γ-rays have been detected in a few cases. The appearance in optical images ranges from tiny, fast knots of pure oxygen to delicate filaments of pure Balmer line emission. SNRs produce infrared emission from dust grains heated by collisions with shock-heated electrons and ions. The cold gas in SNRs produces strong IR emission in atomic fine structure forbidden lines and molecular transitions.

The following sections summarize the different classes of SNRs. Eventually, when the mass of material swept up from the ISM is much greater than the mass ejected in the stellar explosion, the character of a remnant is determined mostly by its age and the density structure of the interstellar gas it encounters. However, for the first thousand years or so of its life, the SNR is strongly influenced by the nature of the explosion. SUPERNOVAE themselves are divided into two broad classes. Type Ia SNs are thermonuclear explosions of white dwarfs, while type II SNs occur when the core of a massive star collapses. Type Ib and Ic SNs resemble type I events in some ways, but they result from the core collapse of a star that has lost its hydrogen envelope. All types of SNs inject energies of order 10^{51} erg, but the masses involved are very different. A type Ia SN ejects $1.4M_{\odot}$ of iron-rich material, while a type II SN ejects an order of magnitude more material at lower speeds. Type II SNs can leave compact remnants, either black holes or neutron stars. In some cases the neutron star is a strong pulsar, able to dominate the energetics of the SNR for many years.

Young SNRs

SNRs younger than about 1000 yr show strong signatures of their origins. SN1987A provides the opportunity to study the very early evolution of an SNR as it interacts first with material shed by the precursor star and later with the ambient interstellar gas. The differences in ejected mass and composition and the presence or absence of a PULSAR lead to three basic types of young SNRs.

Type Ia SNRs (Balmer line remnants)

These are the simplest SNRs, appearing as nearly circular, limb-brightened shells in x-ray and radio emission. The prototypes are the remnants of the supernovae reported by Tycho Brahe in 1572 and by Chinese astronomers in 1006. Four SNRs of this class are known in the Large Magellanic Cloud.

The SN ejecta cool very rapidly by expansion, becoming a cloud of extremely cold, nearly neutral gas expanding at thousands of km s^{-1} into the ISM. A hot star observable at ultraviolet wavelengths is fortuitously located behind SN1006, and observations with IUE and the Hubble Space Telescope show strong absorption in lines of Fe II and Si II spread over a velocity range of $\pm$4000 km s^{-1}.

When the cold ejecta encounter the ISM at highly supersonic speeds, a shock wave forms in the interstellar gas and a another shock (the 'reverse shock') forms in the ejecta. The shocks compress the ISM and ejecta and heat them to temperatures $\simeq 10\ \mu V^2$, where μ is the mean atomic weight of the shocked gas and V is the shock speed in km s^{-1}. Thus the SNR at this stage consists of a cold, rapidly expanding inner region surrounded by a shell of shocked SN ejecta, which is in turn surrounded by a shell of shocked interstellar gas. The pressures of the shocked ejecta and shocked ISM balance, and a contact discontinuity separates the two regions.

The structure of the shell of shocked gas depends on the density structure of the ejected gas, but in general the shocked ejecta are cooler and denser. The entire shell of shocked material decelerates as it sweeps up mass from the ISM, and the contact discontinuity between the shocked ISM and shocked ejecta is subject to the Rayleigh–Taylor instability, which mixes the two components and generates dense clumps. The Rayleigh–Taylor instability is seen on Earth when a low-density fluid supports a dense fluid against gravity. In the SNR, deceleration of the shell of shocked gas plays the role of gravity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The shell of shocked gas produces bright x-ray emission. The shocked ejecta, being highly enriched in iron and other moderately heavy elements, produce x-ray radiation more efficiently, and the emission of this gas is further enhanced by its high density. X-ray spectra in the 1–10 keV photon energy range show strong emission lines of highly ionized Si, S, Ca and Fe produced by impact of thermal electrons on the ions. Continuum emission at moderate x-ray energies arises from bremsstrahlung and from radiative recombination of ions and electrons. The shocked ISM is somewhat hotter and produces more emission at x-ray energies above 10 keV.

Strong shock waves can accelerate particles to relativistic energies. Their velocity distribution takes the form of a power law. Ions accelerated by SNR shocks constitute a large part of the cosmic ray population. Energetic electrons spiraling in the magnetic field produce a polarized radio continuum of SYNCHROTRON RADIATION. As both the magnetic field and the non-thermal particle densities are concentrated in the shell of shocked material, type Ia SNRs appear as strongly limb-brightened shells at radio wavelengths, well correlated with the x-ray emission.

The non-thermal electrons can also produce hard x-ray and γ-ray emission. SN1006 shows bright arcs of hard x-rays produced by synchrotron emission, and TeV γ-rays, believed to be produced by inverse Compton scattering from energetic electrons, have been detected from one of these arcs. The non-thermal x-rays are a continuation of the radio spectrum, but with a steeper spectral slope. Comparison of the radio and x-ray emission constrains the acceleration time of the particles and the magnetic field strength.

Type Ia SNRs are faint at optical wavelengths. They show only thin filaments along parts of their peripheries. The spectra of these filaments are remarkable in that the only optical emission detected is the H I Balmer series. This is in sharp contrast to the usual spectra of photoionized or shock-excited nebulae, where the forbidden lines of [O I], [O II] and [O III] dominate. The shocks in Type Ia SNRs are extremely fast. Radiative cooling is ineffective, and the hot gas produces very little emission at optical wavelengths. However, if the shock encounters partially neutral material, each hydrogen atom swept up by the shock stands about a 20% chance of being excited to produce an Hα photon before it is ionized. It may also undergo charge transfer with an ion in the shocked gas. Thus the Hα line profile consists of a narrow component, with a the velocity width of the pre-shock gas, and a broad component whose width is that of the ions in the shocked gas. The profile of one filament in Tycho's SNR has a narrow component 44 km s^{-1} wide (full width at half-maximum) and a broad component 1900 km s^{-1} wide. These line profiles are useful for measuring the shock speed and for examining the physical processes in the shock. Shock speeds of the known type Ia remnants range from 500 to 2300 km s^{-1}.

Type II, Ib and Ic SNRs (oxygen-rich remnants)

In a few young SNRs, fragments of the core of the progenitor star are visible. They appear as knots of material completely devoid of hydrogen and helium. As products of massive stars, these SNRs often occur in or near dense molecular clouds. Some are badly obscured by dust, and interaction with a cloud may complicate the structure.

The best-known example of this class is Cas A (CASSIOPEIA A). Like the type Ia SNRs, the O-rich remnants display a forward shock–reverse shock structure. The SN event ejects tens of solar masses of material that has undergone nuclear burning. In some cases an outer envelope of unprocessed material is ejected as well. Both the radio and the x-ray emissions from the shocked shell are very bright, and they are concentrated in small knots. X-ray, optical and radio brightness are well correlated on large scales, but individual knots do not match knots seen at other wavelengths. The high degree of clumping is likely to result from advanced Rayleigh–Taylor instability. Thermal emission dominates the x-ray spectrum. Non-thermal emission may be present as well, but the brightness of the thermal emission makes it more difficult to disentangle than in the type Ia SNRs. The recent discovery of γ-ray emission from the decay of ^{44}Ti from Cas A opens a new avenue for the study of core collapse in supernovae.

A remarkable aspect of the O-rich remnants is the swarm of fast knots at optical wavelengths. Cas A shows several hundred of these knots, which can appear and fade over the course of a decade. In Cas A, they show pure oxygen emission, oxygen mixed with sulfur, and in some cases C, Ar and Ca as well. Little neon is seen in optical spectra, but [Ne II] and [Ne III] have been detected in infrared spectra. Infrared continuum emission from dust grains at temperatures near 200 K has also been detected from the knots.

The Cas A knots travel at around 6000 km s^{-1}. Their proper motions and Doppler shifts can be fitted to nearly spherical expansion originating at a point in about 1658. By combining the line-of-sight component of velocity obtained from the Doppler shifts with the component in the plane of the sky (obtained from proper motions), it is found that the knots lie on a roughly spherical surface. Still more remarkable, most of them fall on several large circles on that sphere.

The knots in Cas A are believed to be clumps formed during the SN explosion. Initially they move more slowly than the more tenuous gas of the outer envelope of the exploding star. As the SNR sweeps up interstellar gas, the outer layers slow down and the knots overtake the reverse shock. When that occurs, a slower shock propagates into the dense knot. Models for the infrared, optical and ultraviolet emission from such shocks are not entirely consistent with the observed spectra, but in general the shock wave thermalizes the bulk kinetic energy of the ions. Coulomb collisions gradually transfer this energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

to electrons, which ionize and excite the ions to produce the observed emission.

Cas A also shows other types of emission knots. Relatively slow knots rich in He and N (quasi-stationary flocculi) are interpreted as material shed by the SN progenitor during an earlier red giant phase of its evolution. When the blast wave encounters dense clumps of this material, it drives relatively slow shocks ($\simeq$150 km s^{-1}) into them. There are also a few high-velocity knots which show He- and N-rich spectra. Some of these lie in a high-velocity jet in the northeast which extends beyond the apparent boundary of the rest of the remnant.

Other examples of the O-rich class are N132D and 1E102.2-7219 in the MAGELLANIC CLOUDS and G192+1.8 in the plane of the Milky Way. An extremely luminous oxygen-rich remnant in the irregular galaxy NGC4449 is unresolved even in Hubble Space Telescope images. N157 shows O-rich knots along with a pulsar. A few knots of processed material are found in the somewhat older remnant Puppis A.

The Magellanic Cloud remnants are the best-observed of these objects thanks to the relatively low extinction in those directions. Their UV spectra indicate more neon than Cas A, along with lines of carbon, oxygen and magnesium. The ratios of ultraviolet to optical line intensities imply electron temperatures of order 30 000 K, confirming the basic idea of excitation in modest velocity shocks.

Crab-like SNRs

The CRAB NEBULA is a spectacular example of a pulsar-driven SNR. It is unique in many ways, although a few related SNRs are known. The Crab was created in an explosion reported by Chinese astronomers in the year 1054. It is an irregular system of rapidly expanding filamentary emission and an amorphous cloud of polarized synchrotron emission observed from radio through γ-ray wavelengths. This emission is powered by the pulsar, which produces a relativistic wind. Shock waves in this wind can appear as moving wisps.

The network of optical filaments occupies a thick shell expanding at 700–1800 km s^{-1}. These filaments emit the usual nebular forbidden lines, the hydrogen Balmer lines and helium recombination lines. They show signs of modest nuclear burning, particularly enhanced helium abundances. The composition suggests a progenitor mass of order $8M_\odot$.

The spectra of the optical filaments are generally consistent with heating and ionization by synchrotron continuum photons. Photoionization alone leads to relatively faint emission of ultraviolet spectral lines, however, and ultraviolet observations imply that some shock heating is also required. Shock heating and photoionization are intertwined in the physics of the optical filaments. The expanding relativistic cloud drives 100 km s^{-1} shocks into the smooth cloud of SN ejecta. As the gas cools, it produces UV line emission, and in the process of cooling it becomes more dense, increasing its emissivity when exposed to ionizing photons. The Rayleigh–Taylor instability condenses the cooled gas into sheets and fingers.

The unshocked outer regions of the stellar ejecta are extremely faint owing to their low density. They have not yet encounted dense enough interstellar gas to have formed a detectable forward–reverse shock pair. A cylindrical structure of optical emission filaments resembling a chimney protrudes beyond the shell defined by the other optical filaments. Apparently, compression makes the outer ejecta visible there.

SNRs similar to the Crab include 3C58 and CTB87. They are characterized by non-thermal emission with a relatively flat spectrum at radio wavelengths. While the type Ia and oxygen-rich remnants show limb-brightened shell structures, the surface brightness of a Crab-like remnant peaks at the center near the pulsar. The best-studied of these objects, 3C58, has faint optical emission filaments like those of the Crab and an extended x-ray emission region. It may be the remnant of an SN explosion in the year 1181.

A number of other SNRs contain pulsar-powered radio and x-ray nebulae along with shells of thermal emission. These hybrids include N157 (mentioned above for its extreme elemental composition), CTB80, MSH15-56 and the Vela SNR. Pulsars are not always apparent, perhaps because of the beaming of the pulsar radiation in another direction. W50, the SNR containing SS433, shows apparently normal shock heating at its boundaries, but the non-thermal emission is also present.

The rarity of SNRs like the Crab indicates that they represent a very short stage in the life of an SNR or that they are formed only by an unusual type of explosion.

Older SNRs

After an SNR has swept up much more interstellar gas than the mass of the stellar ejecta, the character of the explosion itself becomes unimportant. The explosion energy and the density of the ambient material determine the evolution and appearance of the SNR. For a typical SN energy of 10^{51} erg and an interstellar density of 1 cm^{-3}, an SNR loses the memory of its origin after a few thousand years.

For expansion of a blast wave in a uniform medium, the Sedov–Taylor similarity solution predicts the radius and velocity of the shock as a function of time. Conservation of energy determines the basic evolution. The relations are

$$R = 13.6(\epsilon_{51}/n_0)^{1/5} t_4^{2/5} \text{ pc} \quad (1)$$

$$V = 535(\epsilon_{51}/n_0)^{1/5} t_4^{-3/5} \text{ km s}^{-1} \quad (2)$$

where ϵ_{51} is the SN explosion energy in units of 10^{51} erg, n_0 is the ambient density in atomic mass units per cm^3 and t_4 is the age divided by 10 000 yr. The post-shock temperature is proportional to the square of the shock speed, so that $T = 3.1 \times 10^6 (\epsilon_{51}/n_0)^{2/5} t_4^{-6/5}$ K at the edge of the remnant. The profiles of density, temperature and velocity inside the SNR are invariant functions of the fractional radius. The mass is concentrated near the outer edge, so that the SNR

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

appears as a limb-brightened shell with a temperature a little above the shock temperature at a given time.

As the shock slows to about 300 km s^{-1}, the radiative cooling time becomes comparable with the age of the remnant. The infrared luminosities exceed the x-ray luminosities in many remnants, though recent IR observations show that much of the IR flux is forbidden line emission from cold gas, rather than emission from hot dust grains in the hot regions. When a significant fraction of the SN energy is radiated away, the temperature and pressure drop. A substantial fraction of the energy emerges as UV and extreme UV photons, many of the latter being absorbed by the surrounding gas. When a large part of the thermal energy of the remnant has been lost, slower expansion continues in the 'snow-plow' phase, where conservation of the momentum of the cold SN shell governs the expansion.

Several variants of the Sedov–Taylor picture have been computed, including ambient densities which vary with distance from the explosion or physical effects, such as thermal conduction and cosmic ray pressure. Especially in the later phases, the INTERSTELLAR MAGNETIC FIELD can play a major role in supporting the cool gas against compression.

Before the onset of strong radiative cooling, an ideal Sedov SNR would appear much like a type Ia SNR. The model predicts a limb-brightened shell at x-ray and radio wavelengths, with a thin layer of faint optical and ultraviolet emission from the ionizing gas just inside the shock front. Infrared emission from heated dust grains should be closely correlated with the x-ray shell.

Because of the relatively high density and low temperature near the edge of the SNR, radiative cooling affects that region first. A bubble of hot, x-ray-emitting gas remains in the interior, and shocks slower than about 300 km s^{-1} produce bright optical and UV emission. Strong shocks heat the interstellar gas to $1.4 \times 10^5\ v_{100}^2$ K (where v_{100} is the shock speed in units of 100 km s^{-1}), ionize it and accelerate it to three-fourths of the shock speed. In the idealized picture, the gas cools at nearly constant pressure, becoming more dense as it does so. As the density increases, the cooling speeds up, and the temperature drops steeply to around 10 000 K. At that point, hydrogen recombines, and the neutral gas begins to absorb ionizing photons from the hotter regions upstream. This maintains an ionized layer much like an H II region in appearance until the ionizing photons are exhausted.

The rapid cooling and high density cause the emission from the cooling shock to occur in a narrow layer behind the shock. Thus the shock produces a thin sheet of glowing gas at the edge of the SNR. Ripples in the sheet viewed close to edge on take on the appearance of narrow filaments, with diffuse emission between. The bright filaments correspond to tangencies between the line of sight and the emitting sheet of gas, and in those places the flow velocity is transverse to the line of sight. Figure 1 shows a shock wave that has recently entered the cooling phase. As the shock evolves, the crisp arcs become more fragmented and diffuse.

Figure 1. Hα image of a shock front in the Cygnus Loop from the Hubble Space Telescope. The thickness of the rippled sheet of emitting gas unresolved even at the 0.1″ WFPC2 resolution. The long dimension of the image is 2.7 arcmin. (Image courtesy of W P Blair.)

The optical spectrum of the radiatively cooling shock is similar to that of a photoionized nebula, except that a broader range of ionization states is present. The [O III] lines dominate the spectra of most H II regions, while [O I], [O II] and [O III] lines are all bright in SNR shocks. There are also indicators of high temperatures in shocked gas, in particular the temperature-sensitive ratio of the λ4363 and λ5007 forbidden lines of [O III]. The ultraviolet emission lines from a shock are far stronger than those from photoionized gas, as the radiative cooling process mainly converts thermal energy into UV photons. The lines C III λ1909, C IV $\lambda\lambda$1548, 1551 and O VI $\lambda\lambda$1032, 1037 are especially bright. As the temperature falls to about 1000 K, much of the cooling occurs by way of infrared fine structure transitions. ISO observations reveal strong emission in the lines [C II] 157 μm, [O I] 63.2 μm, [Ne II] 12.8 μm, [Si II] 34.8 μm and [Fe II] 26.0 μm. The IR, optical and UV spectra provide diagnostics for the density, temperature and elemental abundances in the shocked gas.

As the SN shell continues to sweep up mass from the ISM, the shock slows still further. Eventually, at shock speeds of order 10 km s^{-1}, much of the emission emerges as molecular emission in the infrared. The 2.12 μm line of H_2 is the most easily observed signature of such shocks. Lines of CO and other molecules are seen at radio wavelengths.

Older SNRs produce a polarized radio continuum by synchrotron emission much like that of the young SNRs. While the young SNRs accelerate electrons to relativistic speeds, old SNRs probably generate the observed radio emission by compressing the cosmic ray electrons and magnetic field already present in the ISM (van der Laan mechanism). A strong radiatively cooling shock compresses the gas and particles by a factor of 100 or more, enormously increasing the synchrotron emissivity. The compression amplifies the magnetic field component perpendicular to the shock flow, and the polarization of the radio emission is generally consistent with tangential magnetic fields around the SNR periphery.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

This basic picture of SNR evolution can be adapted to include such effects as magnetic pressure support, electron–ion temperature differences or the gradual liberation of elements such as C, Si and Fe from dust grains by sputtering or grain–grain collisions. A more serious obstacle to the straightforward derivation of physical parameters from spectra is the instability of the shock under various conditions. A dynamical instability occurs when the hot interior gas pushes a thin shell into the surrounding medium. There is also a catastrophic thermal instability in radiatively cooling shocks faster than about 150 km s^{-1}. These instabilities are especially likely during the transition from the Sedov phase to cooling phase, and they can lead to the formation of secondary shocks.

While the Sedov picture is useful as an approximate description of SNR evolution, the ISM is not homogeneous. The blast wave in low-density interstellar gas encounters denser clouds, and these encounters determine the appearance, and perhaps the evolution, of older SNRs. When the blast wave strikes an obstacle, rapid deceleration pushes a weak shock back into the expanding hot gas, increasing the temperature, density and x-ray emissivity. At the same time, a slower shock propagates into the dense cloud. Approximate constancy of the ram pressure, ρv^2, implies that the cloud shock speed is related to the blast wave speed by the square root of the density contrast. If the cloud is large on the scale of the supernova remnant, as in the northeastern or western sides of the CYGNUS LOOP, the interaction will produce enhanced x-ray emission and bright optical filaments, and the expansion rate will be smaller than in lower-density sectors of the SNR.

If the dense cloud is small, the same pattern of x-ray enhancement and optical emission will occur for a short time, but the blast wave quickly sweeps past the cloud. A strong shear in the flow past the cloud generates turbulence, and the cloud will soon be destroyed. The turbulence also tangles and amplifies the magnetic field. This may prevent the cascade of turbulence to small scales, and it may prevent thermal conduction from evaporating the cloud material. If the turbulent cascade proceeds far enough, or if thermal conduction is effective, the cloud material can merge with the hot gas, greatly increasing its density and cooling rate. If the ISM is populated with a large number of small clouds, the rate at which dense gas enters the hot phase may govern the SNR evolution.

It is expected that many SNRs occur in atypical regions of the ISM. The massive stars that produce type II supernovae are formed from massive clouds, and they exhaust their supply of fuel before moving far from their formation sites. They also produce strong winds and powerful ionizing radiation that can modify the immediate environment. SNRs such as the Cygnus Loop are believed to have occurred inside bubbles of their own creation. The luminosity of such an SNR is very low until the blast wave reaches the dense outer shell of the bubble. Then the pressure and luminosity rise rapidly, and the remnant can be far more luminous than a Sedov remnant of the same size.

Interaction

The ultimate fate of SNRs is not known. The low-density cavity of an old SNR provides a channel for the energy of any subsequent SN explosion in or near it. If the SN rate is high enough compared with the cooling time, the tenuous gas will be reheated before it is able to cool. The result could be a system of hot tunnels in a denser ISM, or perhaps a matrix of hot, low-density gas with embedded dense clouds. On the other hand, a lower SN rate, abetted by the buffering of magnetic fields, can reduce the likelihood of interaction. In that case, the SNR shock gradually slows, the interior slowly cools and the shocked gas merges back into the ISM.

If the supernovae are highly correlated events, the odds of interaction greatly increase. The stellar winds and the high supernova rates of an OB ASSOCIATION are sufficient to maintain an enormous low-density bubble with a slowly expanding dense shell. Such a bubble may take the form of a supershell, resembling an SNR with initial explosion energy of 10 or 100 supernovae. In a disk galaxy, the shock tends to break onto the galactic halo, forming a chimney for hot gas to flow upwards. Candidate supershells have been observed in the Galaxy, the Large Magellanic Cloud and other nearby galaxies.

A possible instance of the reheating of an old SNR by a second event is the Local Bubble, a region of hot, tenuous gas about 100 pc in size engulfing the solar neighborhood. The temperature and density derived from soft x-ray measurements are difficult to reconcile with the standard model of a single SN explosion, but an old SNR cavity reheated by a subsequent explosion can explain the size and temperature of the Local Bubble.

Bibliography

More information about supernova remnants may be found in the review paper by Weiler and Sramer (*Ann. Rev. Astron. Astrophys.* **26** 295–341). The physics of interstellar shock waves is reviewed by Draine and McKee (*Ann. Rev. Astron. Astrophys.* **31** 373–432).

John C Raymond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Supernovae

Supernovae represent the catastrophic explosions that mark the end of the life of certain stars. A mass of one to ten times the mass of the Sun is ejected with velocities of about one to ten percent of the speed of light. The corresponding energy is comparable to the gravitational binding energy of a star and hence evidence that the star is severely, if not completely, disrupted. The explosion commonly ejects heavy elements that seed the gas of the interstellar medium. This enriched gas then condenses to form new stars. The heavy elements can form planets and living things. Some supernovae may produce compact remnants such as NEUTRON STARS or BLACK HOLES. Supernovae also produce extended remants as their ejecta sweep outward and drive shock waves into the interstellar medium. This energy can help to form new stars and blow matter entirely out of host galaxies in some circumstances. Because of their great brightness, supernovae can be used to determine distances and help to determine the age, shape and dynamics of the universe.

Historical supernovae

Supernovae have been witnessed in our Galaxy, but only in the historical past. Historical records, particularly the careful data recorded by the Chinese, show that seven or eight supernovae have exploded over the last 2000 years in that portion of the Galaxy not obscured by dust, about 15% of the area of the galactic plane. We also see the expanding supernova remnants that have been produced by supernovae. The youngest of these correspond to the historical supernovae, but we also observe the effects of several hundred older SUPERNOVA REMNANTS that have exploded over the last million years or so.

The supernova of 1006 was the brightest ever recorded, brighter than Venus and perhaps as bright as a quarter Moon. It was observed throughout the Middle and Far East. The supernova of 1054 is by far the most well-known. This explosion produced the rapidly expanding shell of gas that modern astronomers identify as the Crab Nebula. The supernova of 1054 was apparently recorded first by the Japanese and was also clearly mentioned by the Koreans, although the Chinese have the most careful records. There is a strong suspicion that Native Americans recorded the event in rock paintings and perhaps on pottery. There is no confirmed record of an observation in Europe. The supernova of 1572 was observed by Tycho Brahe. Tycho's discovery helped to alter the world view of immutable heavens that began to shift during the Renaissance. Tycho's student, Johannes Kepler, in his turn witnessed the explosion of a supernova in 1604. Kepler also took careful data by which we deduce that he witnessed the same kind of explosion as his mentor. Although there are counter-arguments and some controversy, both Tycho's and Kepler's supernovae are widely regarded to be the kind of event modern astronomers label Type Ia (see below).

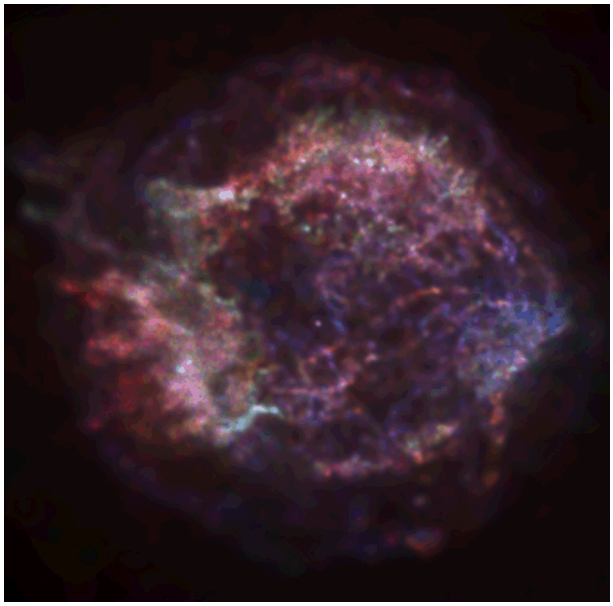


Figure 1. X-ray image of the supernova remnant Cassiopeia A obtained by the Chandra X-ray Observatory. (Courtesy of NASA.)

A young expanding gaseous remnant of an exploded star is observed as a powerful emitter of radio radiation known as Cassiopeia A (figure 1). From the present size and rate of expansion of the remnant, we deduce that the explosion that gave rise to Cas A occurred in about 1680. No bright optical outburst was seen. This explosion may have been under-luminous, but the degree is difficult to ascertain since there is a great deal of patchy dust extinction in this direction. There are reports that Cas A was seen faintly by Flamsteed, the first Astronomer Royal of England, but there are questions concerning whether or not that sighting was in the same position as the remnant observed today. Supernovae tend to be brighter if they explode within large red giant envelopes. The suspicion is that the star that exploded in about 1680 may have ejected a major portion of its envelope before exploding or that the star was otherwise relatively small and compact. That condition, in turn, may have prevented Cas A from reaching the peak brightness characteristic of most supernovae. SUPERNOVA 1987A, the best studied supernova of all time, had this property of being intrinsically dimmer than usual.

Supernova types

All supernovae observed since 1680, and hence all supernovae seen by modern astronomers, have been in other galaxies. A single galaxy produces a supernova only rarely. Supernovae occur roughly once per hundred years for SPIRAL GALAXIES like the Milky Way. Astronomers observe a huge number of galaxies at great distances. The chance that some of these galaxies will have supernovae go off in them is appreciable. Before supernova 1987A, about

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

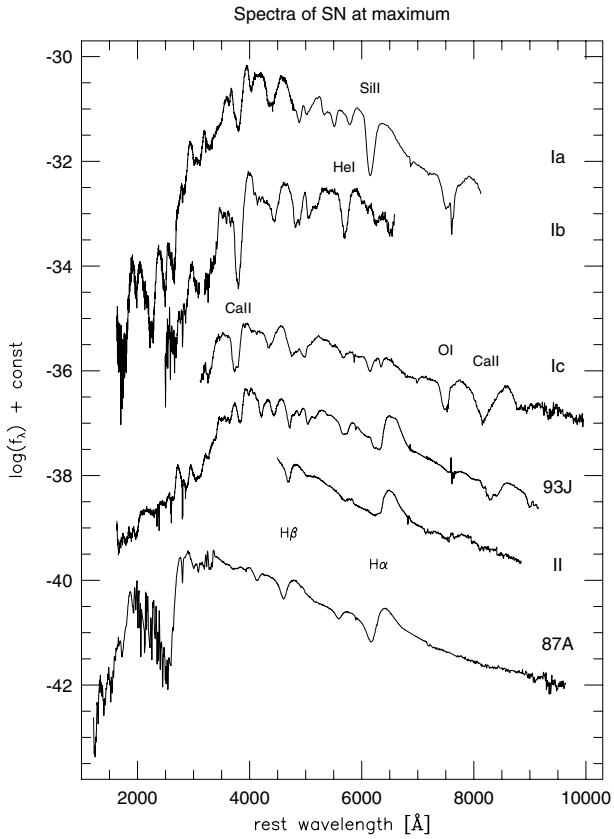


Figure 2. Characteristic spectra of various types of supernovae near maxiumum light. The features labeled are: Si II—ionized silicon; He I—neutral helium; Ca II—ionized calcium; O I—neutral oxygen; Hα, Hβ—neutral hydrogen. (Reprinted from *Astrophysical Quantities*, American Institute of Physics Press.)

30 supernovae were recorded every year. Closer attention was paid to discovering supernovae after supernova 1987A, especially at great distances where they are cosmological markers, and the current rate of discovery is about 100 per year.

Categories of supernovae are traditionally defined by the spectrum that reveals the composition of the ejected matter. Complementary information is obtained from the light curve, the pattern of rapid brightening and slower dimming followed by each event. From the studies of supernovae in other galaxies, astronomers have identified two basic types called Type I and Type II. The dividing lines of this taxonomy are blurred by events that share some properties of Type I and some of Type II (figure 2).

The spectra of Type I supernovae reveal no detectable hydrogen, the most common element in the universe. Some Type I supernovae, called Type Ia, appear in all kinds of galaxies—elliptical, spiral, and irregular. Type Ia tend to avoid the arms of spiral galaxies. Since the spiral arms are the site of new STAR FORMATION, Type Ia supernova are deduced to explode in older, longer-lived stars. This is consistent with their being the only type to explode in ELLIPTICAL GALAXIES that lack recent star formation. Type Ia show characteristic elements in the spectrum, with elements such as magnesium, silicon, sulfur, and calcium near maximum light and iron later on. The light curve for Type Ia supernovae is also characteristic. There is an initial rise to a peak that takes about two weeks and then a long slower period of gradual decay over timescales of months that is very similar for all these events. The data recorded by Tycho and Kepler suggest that they both witnessed Type Ia supernovae. For decades, all Type Ia supernovae were thought to be virtually identical, but more recent careful observations have revealed small, but real, variations among them.

Near the peak of their light output, Type II supernovae show normal abundances in their ejected material, including a normal complement of hydrogen. These supernovae have never appeared in elliptical galaxies. Type II supernovae occur occasionally in IRREGULAR GALAXIES, but mostly in spiral galaxies and then within the confines of the spiral arms. The interpretation is that the stars that make Type II supernovae are born and die in a short time. Since they are short-lived, the stars that make Type II supernovae must be massive. The optical light curve of a typical Type II supernova shows a rise to peak brightness in a week or two and then a period of a month or two when the light output is nearly constant. After this time, the optical luminosity drops suddenly over a few weeks and then more slowly with a timescale of months. This pattern of light emission with time is consistent with an explosion in the core of a star with a massive, extended red giant envelope.

Two other varieties of hydrogen deficient supernovae are labeled Type Ib and Type Ic. The two types are probably closely related. Unlike Type Ia, but like Type II, Types Ib and Ic only seem to explode in the arms of spiral galaxies. Therefore, Types Ib and Ic are also associated with massive stars. Type Ib show evidence for helium in the spectrum near maximum light. Type Ic show little or no such evidence for helium. On the other hand, both Type Ib and Ic show evidence for oxygen, magnesium and calcium at later times. Type Ib and Ic show little or no evidence for a strong line of silicon that is a major characteristic of the spectra of Type Ia. Type Ia supernovae show essentially only iron at later times, another factor emphasizing their difference from Types Ib and Ic. The composition revealed by Types Ib and Ic is similar to that expected in the core of a massive star that has been stripped of its hydrogen. In the case of Type Ic, most of the helium is gone as well. This suggests an origin of Type Ib and Ic in stars much like WOLF–RAYET STARS, but a direct connection to this class of stars has not yet been established. The light curves of Type Ib and Ic are somewhat similar to those of Type Ia, but are dimmer at maximum light. Type Ib and Ic are often strong radio sources whereas no radio has yet been detected from a Type Ia. Radio emission is thought to be created when the supernova shock collides with a surrounding medium, so this radio emission may indicate strong mass loss from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Type Ib/c, again consistent with an origin in a massive stellar core.

A bright supernova observed in 1993, SN 1993J, gave yet more clues to the diversity of processes that lead to exploding stars. SN 1993J revealed hydrogen in its spectrum, so this event was a variety of Type II. As the explosion proceeded, however, the strength of the hydrogen features diminished and strong evidence for helium emerged. In this phase, SN 1993J resembled a Type Ib. Apparently this star had most, but not all, of its hydrogen envelope removed, probably in a binary mass transfer process. There is yet no direct observational proof for binary companions in Type Ib or Type Ic or the transition events like SN 1993J, but this seems likely. Strong winds from massive stars could play a role in stripping the hydrogen and helium from Type Ib and Type Ic and the relative importance of winds versus binary mass transfer has not been resolved.

Progenitor evolution in massive stars

Type II and Type Ib/c supernovae thus represent the explosion of massive stars. These stars have presumably evolved from the main sequence to RED GIANTS and have had a series of nuclear burning stages producing ever heavier elements in the core. The masses of the stars that give rise to these events is estimated to be between about ten and perhaps thirty to fifty solar masses. Below the lower limit, stars make WHITE DWARFS that do not explode as single stars, but may explode as Type Ia if they occur in binary systems with appropriate mass exchange. Above the very uncertain upper limit, stars probably collapse to make black holes with no associated supernova explosion. The statistics of the rate of explosion of these events is roughly consistent with this estimate. Stars of this mass range are thought to form iron cores that collapse to form neutron stars. There is thus a strong suspicion that Type II and Type Ib/c supernovae make neutron stars as compact remnants of the explosion and that the gravitational energy liberated in forming the neutron star is the driving force of the explosion.

At the lower end of the mass range suspected to contribute to Type II and Type Ib/c supernovae, the evolution may be slightly different. Computer calculations show that, for stars with original mass between about eight and twelve solar masses, carbon burns to produce neon and magnesium, but the oxygen that typically co-exists with the carbon after helium burning does not get hot enough to burn. The core, now composed of oxygen, neon, and magnesium, contracts to form a white dwarf. The atomic nuclei of neon and magnesium are capable of absorbing an electron, thus turning one proton into a neutron, and transmuting themselves into an element of lower proton number. This process reduces the electrons that are responsible for the pressure that supports the core. The result is that the core collapses before any of the elements in the core begin thermonuclear burning. During the collapse, the remaining nuclear fuels—oxygen, neon, and magnesium—are converted to iron. The net result is a collapsing iron core, just as for the more massive stars where the iron core forms before the collapse ensues. These two processes of iron core collapse may give identical results, or there may be some subtle difference between collapse triggered by absorbing electrons rather than by heating and disintegrating the iron. These differences could affect the explosive outcome.

Estimates of the rate of formation of neutron stars in the Galaxy are similar to estimates of the rate of formation of Type II and Type Ib/c supernovae. This does not prove that these supernovae produce neutron stars, but the notion that the two processes are directly related is a nearly universal working hypothesis.

Nuclear synthesis in massive stars

Theoretical calculations show that heavy elements in reasonable proportions are produced naturally in massive stars in the process of forming an iron core. Stars with mass between about 15 and 100 solar masses produce substantial amounts of heavy elements. If these stars explode and eject their heavy elements, this freshly synthesized material will mix with the hydrogen in the interstellar gas. New stars form from this enriched mixture. Stars in the upper end of this range could collapse to form black holes without substantially altering this picture.

Core collapse

The iron core that forms in the center of a massive star is endothermic; it can only absorb energy either by breaking down into lighter elements or by forging heavier elements. The absorption of energy reduces the pressure support and guarantees that the iron core will become unstable to collapse shortly after it forms, in less than a day. In practice, the dominant process that destabilizes the core is that of photodisintegration, when photons of light (actually gamma-rays in the hot interior of the core) cause some of the iron to break apart into helium nuclei and free neutrons.

In the collapse of an iron core, the protons in the iron-like nuclei capture electrons and convert to neutrons. Each reaction creates a NEUTRINO. When the collapse reaches the density of atomic nuclei, the strong nuclear force has a repulsive component. The result of the increased pressure is that the collapse halts. A huge amount of energy is released when the neutron star forms. This energy is 1000 times more than is necessary to blow off the outer layers, those containing calcium, oxygen, carbon, and helium, and any outer envelope of hydrogen. The problem is that most of the energy produced in the collapse is lost to the neutrinos that can easily diffuse out of the newly born neutron star and stream through the infalling matter.

Uncertainties in the complex physics involved in core collapse have prevented an unambiguous understanding of the explosion mechanism. When the neutron star first forms, the new star overshoots its equilibrium configuration giving a large compression to the neutron

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

core. There is then a rebound. This rebound sends a strong supersonic shock wave back out through the infalling matter. The core takes about one second to collapse after instability sets in. The core bounce creates the shock in about one-hundredth of a second. The shock runs out into the infalling matter, however. Some of the energy of the shock is dissipated by the production and loss of neutrinos. The shock also must do the work of breaking down the infalling iron into lighter elements, protons and neutrons, to form the neutron star. The shock wave can thus stall with insufficient energy to reach the outer layers of the star. Matter can continue to rain down on the stalled shock front. The matter will still be shocked as material hits this front, but the shocked matter will continue to fall onto the neutron star. If enough matter lands on the neutron star, the neutron star will be crushed into a black hole. Current calculations show that the core bounce alone is not sufficient to cause an explosion.

One mechanism to create an explosion that is being actively considered takes advantage of the tremendous stream of neutrinos leaving the neutron star. Normal matter, the Sun, is essentially transparent to neutrinos since neutrinos interact only through the weak nuclear force. The only exception to this is neutron star matter. This matter, nearly as dense as an atomic nucleus, is so dense that it can be opaque or at least semi-transparent to the neutrinos. Although most of the neutrinos will get out into space, a small fraction will be trapped in the hot matter that lies just behind the shock front created by the core bounce even though this matter has a density that is much smaller than that of the neutron star. The accumulation of neutrino heat over a time of about one second may provide the pressure to re-invigorate the shock, driving the shock outward and causing the explosion.

The mechanism for depositing a small fraction of the neutrino energy behind the shock may be related to the boiling of the newly-formed neutron star. When the collapse is first halted and the neutron star rebounds, the neutron star is very hot. This heat can cause the neutron star to boil, much like a pan of water boils on the stove. The boiling provides a mechanism for carrying the heat outward in the case of the neutron star by mechanical motion. Under the right circumstances, this boiling process can be much more efficient in transporting heat than a slower process of leaking radiation, or neutrinos. Calculations of this process in neutron stars are very challenging because the motion is complex. All modern calculations that can follow motion in more than one (radial) dimension show that neutron stars do boil. There is a consensus that explosions will not occur without this boiling. There is still debate about whether this process is sufficient to cause an explosion.

Most of the current calculations of core bounce and subsequent events treat the configuration as spherically symmetric. Even if the neutron star boils, the structure of the neutron star may, on average, be spherically symmetric. There are a number of lines of evidence, however, that the explosions that result from the core collapse process are intrinsically non-spherical. Matter may be ejected more intensely in some directions than in others. If this is the case, then the current numerical calculations may be missing a major ingredient necessary to yield an explosion. The most obvious mechanisms for breaking the spherical symmetry are rotation and magnetic fields. Proper treatment of rotation and magnetic fields may be necessary to fully understand when and how collapse leads to explosions. Recent calculations have suggested that jets of matter associated with the rotation and magnetic field of the neutron star could play an important role in causing the explosion. This might alter the significance of the boiling neutron star and diminish the effect of the neutrinos.

White dwarf explosions

Type Ia supernovae are thought to explode in old, low-mass, long-lived stars. The observed properties of Type Ia are also remarkably uniform. Since white dwarfs of maximum mass, known as the Chandrasekhar mass, would be essentially identical and hence undergo nearly identical explosions, the observed homogeneity of Type Ia has pointed to an origin in exploding white dwarfs. We now know that all Type Ia supernovae are not exactly identical, but the basic principle still holds. The idea is then that the more massive star in an orbiting pair could evolve and form a white dwarf. The low-mass companion could then take a long time to evolve, but the companion would eventually transfer mass onto the white dwarf. If the total mass accumulated by the white dwarf approaches the Chandrasekhar mass of about 1.4 solar masses, the white dwarf will then explode.

Careful studies of the observed properties of Type Ia supernovae are completely consistent with the general picture that the explosion occurs in a white dwarf. The information revealed by the evolution of the spectra is consistent with a configuration in which the denser inner portions of the exploding star burn all the way to iron, but the outer parts are only partially burned. Computer models of exploding white dwarfs give results that match this pattern rather well. The exact nature of the combustion is still being explored, but the most successful models adopt a progenitor that is a carbon/oxygen white dwarf with a mass very near the Chandrasekhar mass.

There are two different ways of propagating a thermonuclear explosion in a white dwarf. One is a subsonic burning like a flame, a process called a deflagration. The other is a supersonic burning that is preceded by a shock front, much like a stick of dynamite. This process is known as a detonation. The most sophisticated current models, those that best match the data, have the unregulated carbon burning begin as a boiling, turbulent, deflagration and then make a transition to a supersonic detonation. Such models naturally create iron-like matter in the center and intermediate elements like magnesium, silicon, sulfur and calcium, on the outside. These models also predict that the white dwarf is completely destroyed, leaving no compact remnant like

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a neutron star or a black hole. This comparison of theory and observation thus strongly points to an interpretation of Type Ia supernovae as the explosion of a carbon/oxygen white dwarf at the Chandrasekhar limit.

Type Ia progenitor evolution

Despite this convergence of opinion on the mechanism, there is no generally accepted picture of the evolutionary origin of Type Ia supernovae. The question of how the white dwarfs grow to the Chandrasekhar mass is still an unsolved problem. There is no direct evidence that Type Ia supernovae arise in binary systems. Despite this lack of direct evidence, all the circumstantial evidence points to evolution in DOUBLE STAR systems, and there are few credible ways of making a white dwarf explode without invoking a binary companion. The challenge is to figure out what binary evolution leads to a Type Ia explosion.

Type Ia supernovae are not all identical. The general trend is that Type Ia supernovae that are brighter than average decline from maximum brightness a bit slower than average. The events that are a bit dimmer than average (some by as much as a factor of 2!), decline more rapidly. Models of exploding Chandrasekhar mass white dwarfs can account for this behavior if the explosion in some stars makes the transition from a subsonic deflagration to a supersonic detonation a little earlier than in others. Why this should be so is the object of current research.

Light curves

Supernovae display a variety of shapes to their light curves. Type Ia supernovae are the brightest. They decay fairly rapidly in the first two weeks after peak light and then more slowly for months. Some Type II supernovae have an extended plateau and some drop rather quickly from maximum light. Both types seem to have a very slow decay at very late times, several months after the explosion. Type Ib and Ic supernovae are typically fainter than Type Ia by about a factor of two, but have similar shapes near peak light and show evidence for a slow decay at later times.

When a supernova first explodes, the matter is compact, dense and opaque. To reach maximum brightness, the ejected matter must expand until the material becomes more tenuous and semi-transparent. The size the ejecta must reach is typically 10 000 times the size of the Sun. This is 100 times the size of a red giant and 100 times the size of the Earth's orbit. As the matter expands, however, it cools. If the matter must expand too far before heat leaks out as radiation, the material may have cooled off so that there is no more heat to radiate.

Most Type II supernova explosions are thought to occur in red giant envelopes. These are very large structures. After the explosion, large envelopes do not have very far to expand before they become sufficiently transparent to leak their heat as light. As they begin to radiate, Type II supernovae still retain a large proportion of the heat that was deposited by the shock wave that accompanied the supernova. Near maximum light and on the typical plateau that lasts for months, Type II supernovae shine by the shock energy originally deposited in the star. The deposited energy presumably arises in the core collapse process.

For a Type I supernova, however, the story is different. Whether the exploding star is a white dwarf, as suspected for a Type Ia, or the bare core of a more massive star, as suspected for Type Ib and Ic, the exploding object is very small. The expected sizes range from one tenth to one thousandth of the size of the Sun. These bare cores are much smaller than the size to which they must expand before they can leak their shock energy. The result is that the expansion strongly cools the ejected matter and by the time the matter reaches the point where it could radiate the heat, the heat from the original shock is all gone. This kind of supernova requires another source of heat to shine at all. All the light from Type I supernovae comes from radioactive decay.

Radioactive decay

The nature of a thermonuclear explosion is to burn very rapidly. If the explosion starts with a fuel—carbon, oxygen, or silicon—that has equal numbers of protons and neutrons, then the immediate product of the burning will also have equal numbers of protons and neutrons. This is because the rapid burning takes place on the time scale of the strong nuclear reactions. To change the ratio of protons to neutrons requires the weak force and thus a longer time.

Iron is produced in a thermonuclear explosion in a three-step process. The first step is to produce an element that is close to iron, but that has equal numbers of protons and neutrons. This element, like iron, has a nucleus that is tightly bound by the nuclear force and has the same total number of protons plus neutrons, 56, but with 28 protons and 28 neutrons. This condition singles out one element, that with the most tightly bound nucleus containing equal numbers of protons and neutrons: nickel-56. Nickel-56 is, however, unstable and therefore undergoes radioactive decay. The radioactive decay is induced by the weak force. The result is the formation of the element cobalt-56. In the process, a neutrino is given off to balance the number of leptons. Excess energy comes off as gamma rays. The gamma rays can be stopped by collision with the matter being ejected from the supernova and their energy used to heat the matter. The hot matter then radiates. The power of the light falls off as the nickel decays away and as the matter expands so that it is less efficient in trapping the gamma rays. The cobalt-56 that forms is also unstable and decays to become iron-56. This decay again produces a neutrino and gamma-ray energy.

The radioactive decay of these elements is controlled by a quantum uncertainty. One does not know what atom will decay, but on the average half will decay in a given time. For nickel-56, the time for half to decay is 6.1 d. Likewise, the cobalt-56 decays with a half life of about 77 d. The observed light curves of Type I supernovae decay somewhat faster than the decay of nickel-56 in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

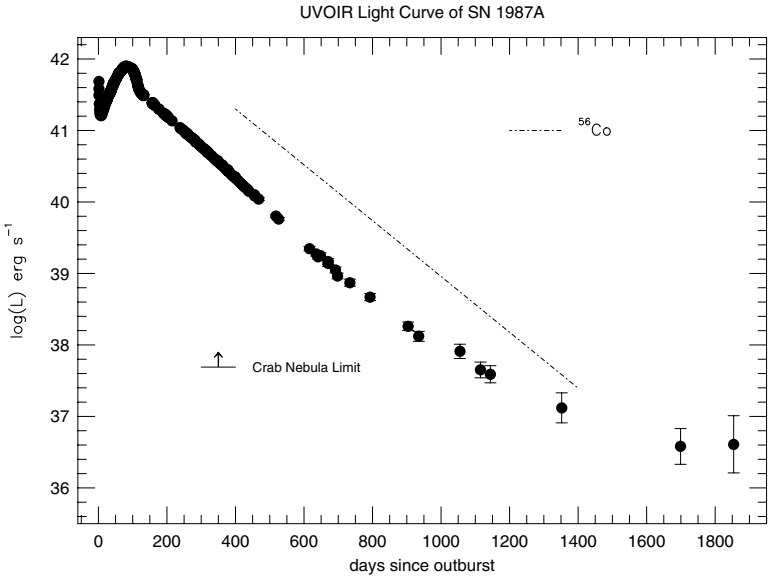


Figure 3. The light curve of Supernova 1987A in comparison with the expected slope for decay of radioactive cobalt-56 and the luminosity of the pulsar in the Crab Nebula. (Reprinted from *Astrophysical Quantities*, American Institute of Physics Press.)

early phase and of cobalt-56 in the later phases. The reason is that not all the gamma-rays produced in the decay are trapped and converted to heat and light. Some of the gamma rays escape directly into space.

For Type Ib and Ic, the amount of nickel required to power the light curve is about one-tenth of the mass of the Sun. This amount of nickel is consistent with many computations of iron core collapse. The nickel is produced when the shock wave impacts the layer of silicon surrounding the iron core. Similar reactions occur in Type II supernovae and similar amounts of nickel are ejected. Type Ia supernovae are generally brighter and must produce more nickel, of the order of 1/2 to 1 solar mass. The dimmest Type Ia events require only 0.1–0.2 solar masses of nickel. The models of Type Ia supernovae based on thermonuclear explosions in carbon/oxygen white dwarfs of the Chandrasekhar mass produce this amount of nickel rather naturally in the explosion. The amount can vary depending on, for instance, the density at which the explosion makes the transition from a deflagration to a detonation, so the variety of ejected nickel mass can also be understood, at least at a rudimentary level.

Supernova 1987A

Supernova 1987A was a Type II supernovae that occurred in the Large Magellanic Cloud about 150 000 light years from the Earth. This made it the closest and hence best studied supernova ever. Neutrinos were detected from SN 1987A, thus confirming the basic picture of core collapse. There was insufficient information to determine the precise mechanism of the explosion. The neutrinos mean that a neutron star formed, at least temporarily. No subsequent evidence for a neutron star has been detected (figure 3).

This may be because the neutron star is obscured in some way, or because it does not emit a normal flux of radiation. The latter could be because it is spinning rather slowly or has a rather small magnetic field. Another interesting possibility is that sufficient matter fell back onto the newly born neutron star to convert it to a black hole after the supernova shock was launched. SN 1987A also gave direct evidence for the production and decay of about 0.07 solar masses of nickel-56. The gamma-rays from the decay of cobalt-56 were directly detected. SN 1987A exploded as a rather compact blue supergiant rather than a red giant. The reasons for this are still not clear, but current reasoning suggests that the progenitor might have had a binary companion that was consumed when the progenitor became a red giant and engulfed the companion (figure 4). The relatively compact state meant that, unlike most Type II, SN 1987A was powered at maximum light by radioactive decay. SN 1987A is surrounded by a set of three nested rings of gas that were formed prior to the explosion. Their origin is unclear, but may be related to the binary interaction. The ejecta from the supernova is beginning to collide with the innermost ring. This interaction may yield more clues to the origin of this structure.

Supernovae and the universe

Because they are so bright, Type Ia supernovae have been used to determine distances. This information can be combined with the expansion velocity of the host galaxy to determine the HUBBLE CONSTANT and hence the age of the Universe. Subtle differences in the brightness for more distant supernovae have given evidence that the Universe is accelerating its expansion, contrary to expectation. The validity of this conclusion may rest on subtle differences

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

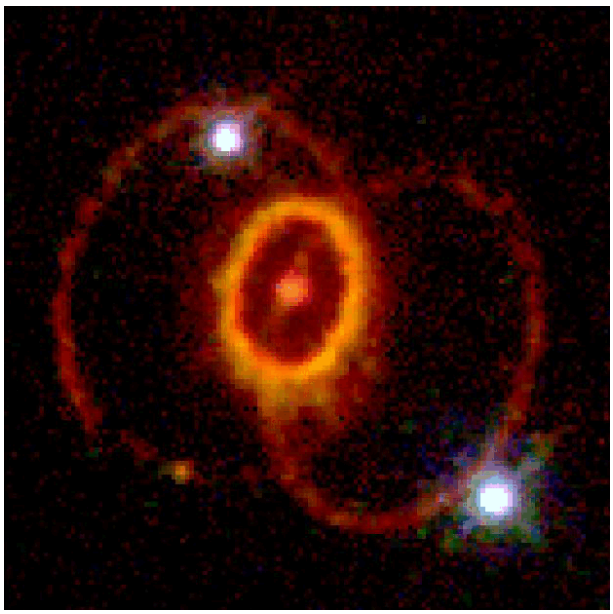


Figure 4. Hubble Space Telescope image of Supernova 1987A about 8 years after its explosion showing the central debris and the three surrounding rings that pre-existed the explosion. The two bright stars are not related to the supernova but fall accidently in the image. (Courtesy of NASA.)

between nearby and distant supernovae that are under investigation.

A recent supernova, SN 1998bw, had characteristics of a Type Ic supernova but was excessively bright, asymmetric, had large expansion velocity, and produced one of the brightest radio sources ever associated with a supernova. There is also an indication that this supernova may have been associated with a gamma-ray burst observed on 25 April 1998.

Bibliography

Marshall L 1994 *The Supernova Story* (Princeton Sciences Library)

Petschek A (ed) 19990 *Supernovae* (Springer Verlag)

Wheeler J C 2000 *Supernovae, Gamma-ray Bursts and Adventures in Hyperspace* (Cambridge: Cambridge University Press)

J Craig Wheeler

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Supersymmetry

A set of rules that attempts to link matter particles (fermions, such as quarks and leptons) and force-carrying particles (gauge bosons) into a single framework, or 'family'. Supersymmetry (abbreviation: SUSY), requires that each type of particle has a supersymmetric partner. For example, a photon (which is a boson) has a fermion counterpart (the photino), an electron (fermion) is partnered by a selectron (boson), a quark by a squark, and so on. The theory therefore predicts the existence of a wealth of hitherto unknown (and as yet undiscovered) particles. When the ideas of supersymmetry are applied to gravity, the resulting theory—supergravity—predicts the existence of one or more types of gravitinos (fermions) to partner the graviton (the hypothesized force-carrying particle for gravitation).

See also: electron, fundamental forces, gauge bosons, leptons, Pauli exclusion principle, photon, quark, string theory/superstring theory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Surges

Surges are cool plasma jets ejected from small flare-like chromospheric bright points, such as subflares or Ellerman bombs (moustaches) near sunspots (see SOLAR CHROMOSPHERE: ELLERMAN BOMBS). They are a kind of active prominences, usually observed in Hα at ground-based observatories, although space observations (such as with EUV telescope) also detect surges. Figure 1 shows a typical example of an Hα surge observed at Hida observatory of Japan, revealing that the surge is emanating from an Hα bright point (subflare) in the periphery of the sunspot penumbra.

Kinematic properties and physical conditions

Surges are ejected at peak velocities of 50–200 km s^{-1} along a straight or slightly curved path, delineating magnetic field lines. They reach heights of 10 000–100 000 km and typically last 10–20 min. Many surges fall back into the chromosphere along the trajectory of their ascent, although some surges fall along the other half of the loop. Sometimes, small Hα brightenings are observed at the footpoint of surges in the falling phase. In some cases, a surge moves perpendicularly to the surge axis at a few tens of km s^{-1} during its ejection (this is referred to as whip-like motion in some literature).

There are some controversies on the kinematic properties of surge material. Some researchers reported that the motion of surges is not ballistic; surges are continuously accelerated during the ascent phase, while they are decelerated in the descent phase. However, some other researchers reported that surges show ballistic motion; all materials are accelerated in a short time and follow only the solar gravitational acceleration after that.

Some of surges show spinning motion at a few tens of km s^{-1} around their axis during the ejection phase. The direction of spinning motion is consistent with the direction of unwinding motion of their helical twist.

The electron density of surges is 10^{11}–10^{12} cm^{-3}, and their temperature is $\sim 10^4$ K. The total mass is 10^{13}–10^{15} g, and the kinetic energy ranges from 10^{27} to 10^{29} erg.

Surges have a strong tendency to recur at the same place at a rate >1 h^{-1}.

Magnetic field properties

The magnetic field strength at the footpoint of surges is of order of 50 G. The footpoint corresponds to small satellite spots (or satellite polarities) whose polarity is opposite to that of the main sunspot. Hence a neutral point or current sheet must be present at or near the footpoint of surges, suggesting the occurrence of magnetic reconnection there. Usually surges occur when magnetic flux is increasing or decreasing, i.e. in an evolving magnetic structure.

Many surges occur in emerging flux regions, especially in their early stage (see CHROMOSPHERE: EMERGING FLUX REGIONS). The typical shape of surges in emerging flux regions is an inverted Y-shape similar to the Eiffel tower (figure 2). This is consistent with the field configuration including a neutral point near the base of surges, which is expected for emerging flux reconnection model.

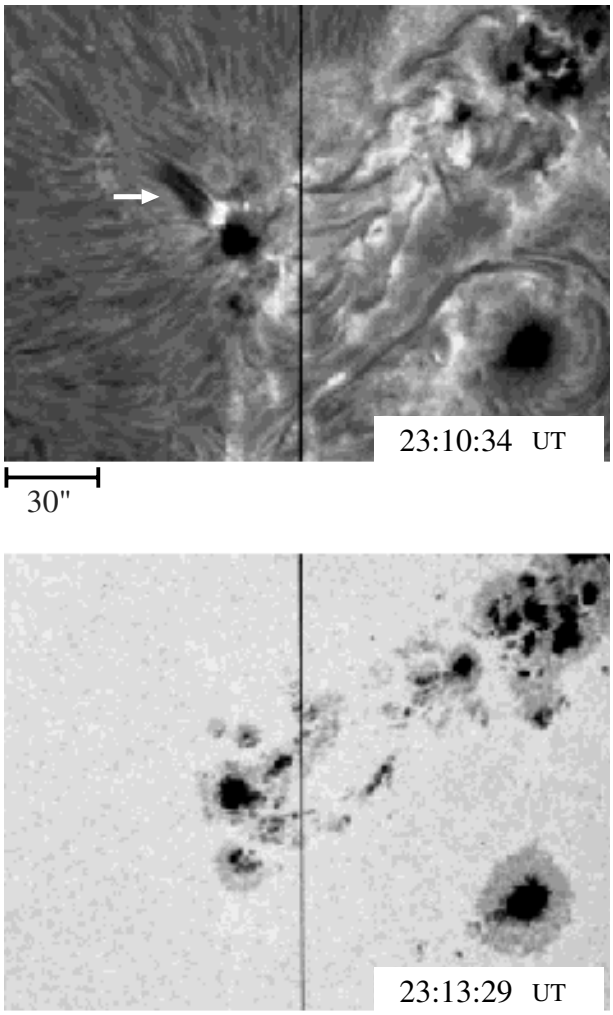


Figure 1. An Hα surge near a sunspot on the disk observed at Hida observatory (courtesy of S Sano). Top: H$\alpha \pm 0.0$ Å; bottom: H$\alpha - 5.0$ Å.

Relation to other phenomena

It is difficult to discriminate large surges whose velocity is more than 400 km s^{-1} from sprays, which are non-collimated cool plasma ejections expanding into large solid angles with velocity often larger than the Sun's escape velocity ($\sim$600 km s^{-1}). Sprays are usually associated with flares. Sometimes, they show an unwinding helical motion. The electron density and temperature of sprays are similar to those of surges. (The intermediate case between surges and sprays is shown in figure 3.)

On the other hand, small surges are similar to spicules, which are cool chromospheric jets observed in Hα and EUV at the network boundary (see CHROMOSPHERE: SPICULES). Their velocity and height are $\sim$30 km s^{-1} and 5000–8000 km, smaller than those of surges, while

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

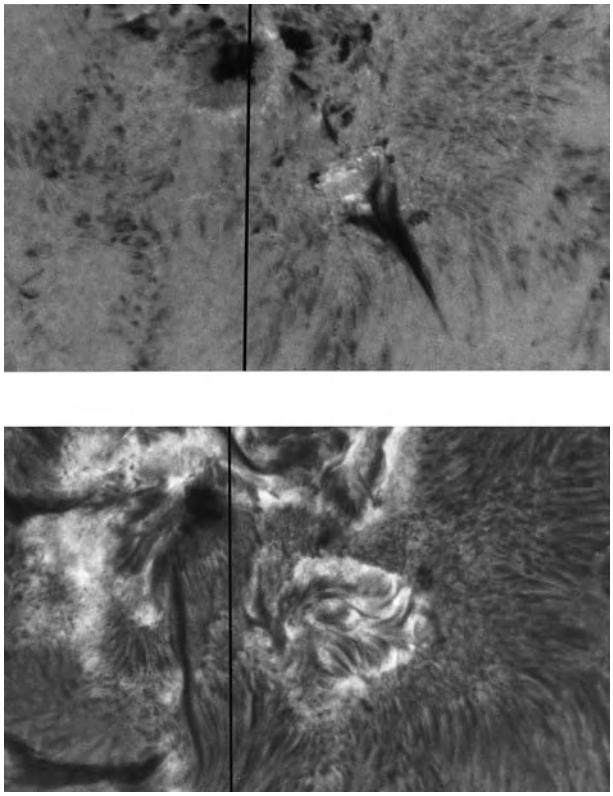


Figure 2. Top: An Hα surge in an emerging flux region, showing an inverted Y-shape similar to the Eiffel tower (observed at Hida observatory, courtesy of H Kurokawa) (03 UT 1 June; Hα − 0.8 Å). Bottom: Hα arch filament loops developing 20 h after the Hα surge shown in the top part (23 UT 1 June; Hα ± 0.0 Å).

Figure 3. An intermediate case between surges and sprays (observed at Hida observatory, courtesy of H Kurokawa).

their temperature and electron density are comparable with those of surges. Spicules show also a velocity feature suggesting spinning motion. It has not yet been established whether spicule motion is ballistic or not.

Macrospicules are observed in coronal holes and show characteristics between spicules and surges. They are ejected at ∼140 km s^{-1}, reaching heights of 10 000–35 000 km, and show nearly ballistic motion.

Flaring arches resemble surges. The only difference is that in flaring arches plasma is injected into a relatively small closed loop structure, whereas in surges the plasma injection goes along one arm of a loop or at some occasions along apparently open field lines. They consist of hot and cool plasmas, and thus can be seen in both Hα and x-rays.

Type III bursts are radio bursts, often occurring in association with surges (see SOLAR FLARES: RADIO BURSTS). These are emitted from high-energy electron beams propagating along open field lines, possibly the same or nearby field lines along which surges are ejected.

SOLAR X-RAY JETS are sometimes observed to be associated with Hα surges, nearly simultaneously and in nearly the same direction. There are also cases with no x-ray jet when Hα surges are observed. In such cases, x-ray bright points are often seen near the footpoint of Hα surges. The lack of x-ray jets may be because the density of hot jet is too low to be detected by soft x-ray telescopes.

EUV jets are jet-like features observed in EUV lines, first discovered by NRL HRTS experiments and later confirmed by SOHO SUMER/CDS/EIT. They are ejected at fast speed (∼400 km s^{-1}), extending to an altitude of 4000–16 000 km (but with a small width ∼1500 km), with mass of 3×10^{11} g and kinetic energy of $\sim 10^{26}$ erg. The relation between EUV jets and surges has not yet been made clear.

Table 1 summarizes characteristic physical quantities of various jet-like mass ejections in the solar atmosphere, including not only cool jets but also hot jets.

Formation mechanism

It is believed that the energy source of surges is magnetic energy stored in the solar atmosphere. In fact, the magnetic energy stored at the footpoint of surges (in a volume of 10^{27} cm^3 with a field strength of 50 G) is estimated to be $\sim 5 \times 10^{30}$ erg, which is enough to explain the observed kinetic energy of surges.

There is increasing observational evidence for magnetic reconnection in surges, such as association with satellite polarities, the inverted Y-shape (or Eiffel tower shape), whip-like motion which is interpreted as a sling-shot-like motion as a result of reconnection, and so on.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Various jet-like mass ejections in the solar atmosphere.

Mass ejections	Velocity (km s^{-1})	Length (km)	Mass (g)	Kinetic energy (erg)	Temperature (K)
Spicules	30	(5–8) × 10^3	10^{11}–10^{12}	10^{24}–10^{25}	(0.5–1) × 10^4
Macrospicules	∼140	(1–4) × 10^4	10^{12}–10^{13}	10^{26}–10^{27}	10^4
Surges	50–200	(1–10) × 10^4	10^{13}–10^{15}	10^{27}–10^{29}	10^4
Sprays	>200	>10^5	10^{14}–10^{15}	10^{29}–10^{30}	10^4–10^5
EUV jets	400	(0.4–1.6) × 10^4	3 × 10^{11}	10^{26}	10^5
X-ray jets	10–1000	10^4–3 × 10^5	10^{11}–10^{13}	10^{25}–10^{27}	(3–7) × 10^6
X-ray plasmoids	50–400	>10^5	10^{13}–10^{14}	10^{27}–10^{28}	(5–16) × 10^6

Even spinning motion can be interpreted as a result of reconnection between a twisted flux tube and an untwisted flux tube (see figure 4). Hence it is very likely that magnetic reconnection is a key mechanism to produce surges.

What is the collimation mechanism of surges? It is easy to show that magnetic force can collimate surges; the gas pressure of surges is 0.1–1.0 dyn cm^{-2} while their magnetic pressure is 100 dyn cm^{-2} for 50 G, much larger than the gas pressure.

Even if magnetic reconnection is a key to the production of surges, there remains a fundamental question: what is the specific acceleration mechanism of surges? Namely, is a surge accelerated by gas pressure force enhanced by a sudden release of magnetic energy or is it accelerated by magnetic force as a result of reconnection? Hence, there are basically two accelerating forces: one is the gas pressure and the other is the magnetic force ($\boldsymbol{J} \times \boldsymbol{B}$ force).

Gas pressure force

If a sudden energy release occurs in the upper chromosphere, the gas pressure increases and directly accelerates cool plasmas in the upper chromosphere along an open flux tube.

If an explosion occurs in the photosphere or in the low chromosphere, the hydrodynamic shock wave (slow-mode MHD shock) propagates along an open flux tube through the chromosphere to accelerate cool plasmas in the upper chromosphere.

In both cases, the speed of the gas-pressure-driven jet is basically determined by the sound speed (i.e. temperature) in the high-pressure region or shock-heated region, and is written as

$$V_{jet,gas\text{-}p} \sim 3C_s \simeq 150\left(\frac{T}{10^5\ \mathrm{K}}\right)^{1/2}\ \mathrm{km\ s^{-1}} \qquad (1)$$

where C_s and T are the sound speed and temperature of the high-pressure (or shock heated) regions, respectively. (This velocity is a maximum speed when the high-pressure gas expands into vacuum.)

Magnetic force

If a magnetic field line reconnects with another field line, magnetic tension force is generated in the reconnected field line as in a sling-shot. Such a magnetic tension force accelerates plasmas on the reconnected field line up to the Alfvén speed just outside of the current sheet. (See figure 5 for a magnetic reconnection model of a surge, which shows that not only cool jet but also hot jet are ejected from reconnection region.)

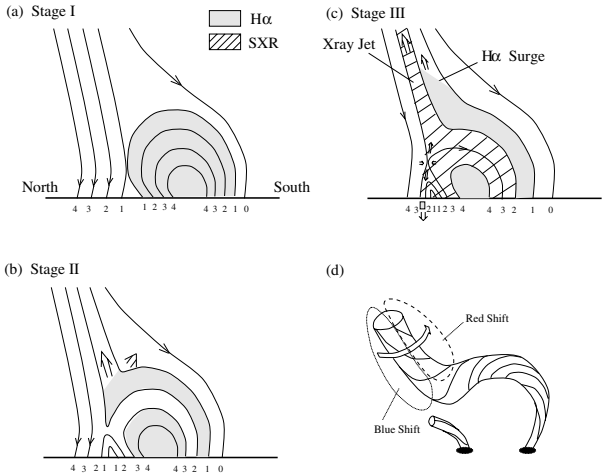


Figure 4. A magnetic twist model for Hα surges (Canfield *et al* 1996 *Astrophys. J.* **464** 1016; see also Shibata and Uchida 1986 *Sol. Phys.* **103** 299).

If a plasmoid (confined by a magnetic island) is created in a current sheet as a result of reconnection and if there is a gradient of magnetic field strength in the ambient field along the current sheet, the plasmoid would be accelerated by the magnetic pressure gradient in the ambient magnetic field. This is known as the diamagnetic mechanism or melon seed mechanism.

If a twisted flux tube reconnects with an untwisted flux tube, magnetic twist would be suddenly released into the untwisted flux tube. During this process, a magnetic pressure gradient force appears along the reconnected flux tube and accelerates the plasma contained in the reconnected flux tube. If the reconnected flux tube has a diverging cross-sectional area, a centrifugal force would also act as an additional accelerating force.

In either case, the velocity of magnetically driven jets

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

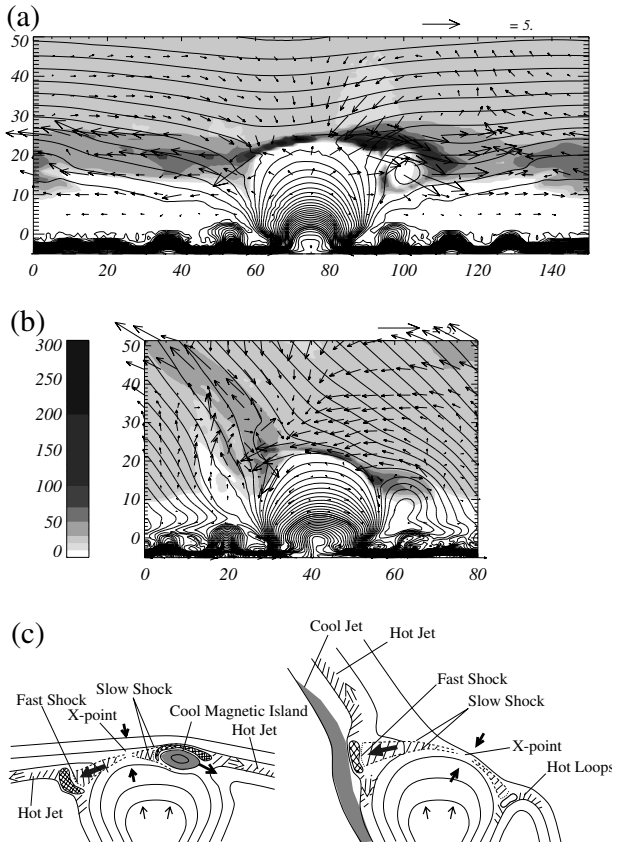


Figure 5. Illustration of emerging flux reconnection model of a surge (Yokoyama and Shibata 1996 *Publ. Astron. Soc. Japan* **48** 353).

is of the order of the Alfvén speed;

$$V_{jet,mag} \sim V_A \simeq 300 \frac{B}{30\ \mathrm{G}} \left(\frac{n}{10^{11}\ \mathrm{cm}^{-3}} \right)^{-1/2} \mathrm{km\ s}^{-1}. \quad (2)$$

Here B is the magnetic flux density and n is the electron density.

Generally speaking, if jets are cool enough, it is difficult to accelerate such cool jets by gas pressure without raising the temperature, whereas it is quite easy to accelerate cool jets by magnetic force while keeping a low temperature. Hence the magnetically driven jet mechanism has wider applicability as a model for cool jets. This is especially true for larger jets (such as large surges and sprays), although, as for smaller jets (such as spicules and small surges), there still remains the possibility that jets are accelerated by gas pressure force or slow-mode MHD shock.

To elucidate the acceleration mechanism of various jet-like phenomena in the solar atmosphere (table 1) would be a challenging subject for the future. Such study is also useful for the study of astrophysical jets ejected from active galactic nuclei, close binary systems and young stellar objects.

Bibliography

Kurokawa H and Kawai G 1993 *The Magnetic and Velocity Fields of Solar Active Regions (ASP Conf. Ser. 46)* ed H Zirin *et al*, p 507

Rompolt B and Svestka Z 1996 *Adv. Space Res.* **17** (4/5) 115

Schmieder B, Shibata K, van Driel-Gesztelyi L and Freeland S 1995 *Solar Phys.* **156** 245

Yokoyama T and Shibata K 1995 *Nature* **375** 42

K Shibata

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Surveyor

Series of seven NASA automated Moon landers, launched 1966–68. Designed to send back images and other data on potential APOLLO landing sites near the lunar equator. Surveyor 1 was the first spacecraft to make a controlled soft landing on another world. Surveyors 5–7 carried out *in situ* soil analyses. Surveyor 6 performed the first lift-off from the lunar surface. Surveyors 2 and 4 were failures. Parts of surveyor 3 were returned to Earth by Apollo 12 astronauts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Swasey, Ambrose (1846–1937)

American mechanical engineer and instrument-maker, produced the dividing engine used to make meridian circles at the US Naval Observatory, mounted the 36 in refracting telescope at Lick Observatory in California. Worked with WORCESTER R WARNER in Cleveland, Ohio as fellow amateur astronomers, jointly built and donated their private observatory to Case Western Reserve University, and thus founded the Warner & Swasey Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Swedish Space Corporation (SSC)

Located in Solna, near Stockholm. SSC is a government-owned commercial company with activities covering space systems engineering, space operations and the commercial exploitation of space. SSC also offers support and technical expertise to the National Swedish Space Board.

It has two establishments at Kiruna in the Arctic—the Esrange satellite operations facility and sounding rocket range, and Satellitenbild, a subsidiary devoted to processing of remote sensing data.

The corporation provides scientific instruments and oversees launches of stratospheric balloons and sounding rockets with astronomical and microgravity payloads. It also builds small satellites and had overall responsibility for the Viking and Freja scientific satellites.

For further information see
http://www.ssc.se/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Symbiotic Stars

During the production of the Henry Draper Catalog, Wilhemina Fleming identified several M-type stars with unusually strong hydrogen emission lines. Paul Merrill obtained higher quality spectra of these 'stars with combination spectra' and found intense emission from He II or [O III] and [Ne III] in addition to H I. All of these stars varied by 0.5–1 mag on a timescale of several years. A few systems underwent occasional 2–3 mag eruptions, where a much hotter continuum overwhelmed the red giant spectrum. Louis Berman and Frank Scott Hogg suggested that these objects must be double stars. Merrill, Otto Struve and Pol Swings later confirmed this picture with the detection of periodic radial velocity variations. Merrill suggested 'symbiotic stars' to describe this small class with 'two objects of quite different kinds . . . living close together in extraordinary interdependence and adjustment'.

Phenomena

The modern definition of a symbiotic star relies on spectroscopic criteria outlined by Merrill and Cecilia Payne-Gaposchkin. These are (figure 1)

(1) a red continuum and absorption features, of a late-type giant star (spectral type K or M); and
(2) a blue continuum with bright H I and He I emission lines and either
(a) additional high ionization emission lines, for example He II, [Ne V], and [Fe VII], with an equivalent width exceeding 1 Å (quiescent objects and some eruptive systems), or
(b) an A- or F-type continuum with additional absorption lines from H I, He I, and singly ionized metals (systems in outburst). The absorption features often have P Cygni emission components.

The sample of ∼150 known systems includes two dozen established in the 1930–1950s and others selected from recent spectroscopic surveys. The lack of a systematic survey precludes a proper estimate of the symbiotic frequency. Current estimates suggest there are 3000 to 300 000 in the Milky Way. Most systems lie close to the galactic plane or in the bulge; some are halo stars. A handful are in the Magellanic Clouds; at least one resides in the Draco dwarf spheroidal galaxy.

All quiescent symbiotics are small-amplitude VARIABLE STARS. Most vary by 0.5–1 mag on a timescale of 1–2 yr. Some systems *flicker* by 0.01–0.1 mag on minute timescales; a few show periodic optical and x-ray pulsations with 10–30 min periods. The amplitudes of the flickering and of the pulsations increase toward decreasing wavelength. Roughly 20% contain MIRA variables with pulsation periods of 300–500 d. The pulsation amplitudes increase from 0.5–1 mag at infrared wavelengths to 1–4 mag in the optical.

Symbiotic stars undergo complicated outbursts. Figure 2 shows optical light curves for the recurrent nova T CrB, the classical symbiotic CI Cyg, and the *symbiotic nova* AG Peg. In T CrB, the rise to visual maximum occurs in less than one day. The fast decline suggests rapid ejection of material as in fast classical NOVAE. The eruptions of classical symbiotics are slower, with a rise time of weeks to months and a decline over several orbital cycles. Symbiotic nova outbursts take decades or centuries. AG Peg began an eruption *c*. 1850 and is slowly approaching minimum.

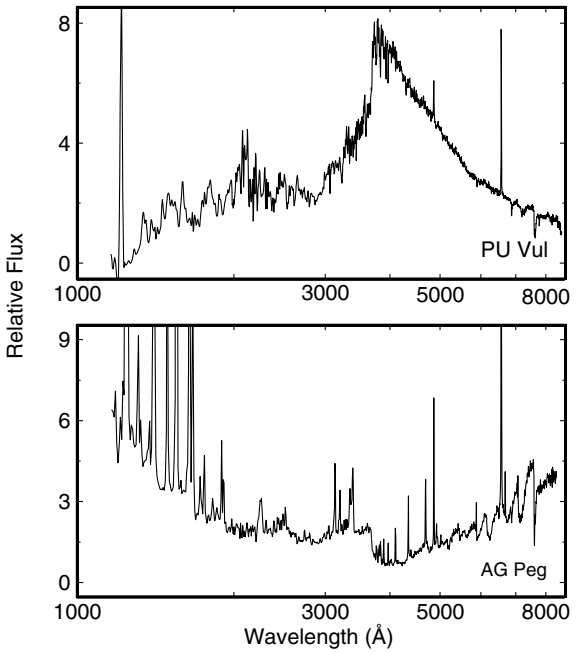


Figure 1. Spectra of symbiotic stars in quiescence (AG Peg; lower curve) and outburst (PU Vul; upper curve). In quiescence, the spectrum contains a strong hot continuum rising towards short wavelengths, strong emission lines and a prominent Balmer jump from the nebula, and a red continuum from the red giant. The hot component often evolves into an F-type supergiant with weak emission lines during eruptions. This spectrum overwhelms the red giant except at the longest wavelengths.

Despite different timescales, the spectroscopic evolution throughout outburst is remarkably similar. In recurrent novae, most classical symbiotics, and many symbiotic novae, an A or F-type continuum overwhelms the red giant spectrum at visual maximum (see figure 1). Many strong emission lines weaken considerably as the brightness increases and strengthen as the brightness fades. The H I and He I lines develop into strong absorption features often with P Cygni profiles. In symbiotic novae and some classical symbiotic stars, the ejection velocity increases from ∼100–200 km s^{-1} at visual maximum to ∼1000–2000 km s^{-1} during the decline. The velocity remains large, ∼2000–3000 km s^{-1}, throughout a recurrent nova eruption. Small velocities, 100–200 km s^{-1}, are characteristic of eruptions in CI Cyg and other classical symbiotics.

Many Mira symbiotics undergo less dramatic symbiotic nova eruptions. Despite a similar 3–5 mag optical rise, the systems never develop A- or F-type optical spectra. High ionization emission lines remain strong. Low

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

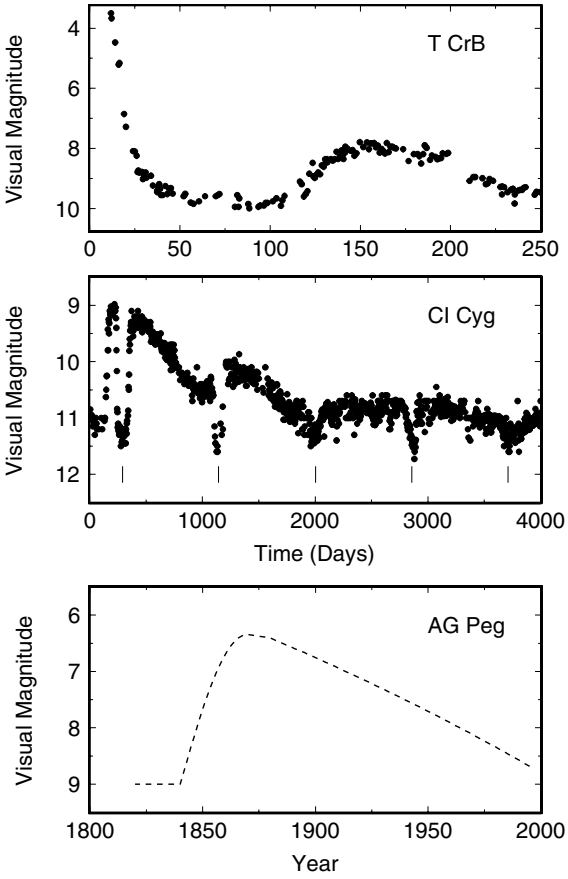


Figure 2. Outburst light curves of three symbiotic stars (adapted from Kenyon (1986)). The recurrent nova T CrB rises 6 mag to optical maximum in less than one day, declines to minimum in ∼50 d, and then rises to a secondary maximum 4–5 mag below primary maximum. CI Cyg, a classical symbiotic, rises 2 mag in several weeks and returns to minimum in 5–10 yr. The vertical tick marks indicate eclipses of the hot component by the giant. The symbiotic nova eruption of AG Peg began before 1850; the system took ∼20 yr to reach maximum and is slowly returning to minimum.

velocity mass ejection is not apparent at visual maximum; broad emission features similar to those in WOLF–RAYET stars appear several years after optical maximum. Several classical symbiotic stars sustain similar 'weak' eruptions and may eject mass at low velocities, ∼100–200 km s^{-1}.

These and other phenomena support Berman and Hogg's original notion that symbiotics are BINARY STARS. Two stars, red giant and 'hot component', account for the two continua with color temperatures of ∼3000 K and ≥50 000 K (figure 1). Energetic photons from the hot component ionize a surrounding nebula. Many well-studied systems show good evidence for orbital motion with periods of 1–3 yr. System masses are ∼1–4 $M_\odot$. The brightness fluctuations are often phased with the orbit and can be explained by (i) illumination of the red giant atmosphere by high-energy photons from the hot component, (ii) eclipses of the hot component and a surrounding nebula by the red giant, or (iii) ellipsoidal variations of the tidally distorted red giant. These interpretations provide further support for the binary picture and allow for detailed studies of the stars and the nebula.

Hot component

The nature of the hot component remained elusive until the 1970s, when satellite ultraviolet observations revealed a hot continuum to accompany the high ionization optical emission lines. These data show that the quiescent luminosity of the hot component is comparable to the red giant luminosity. The effective temperature must exceed ∼50 000 K to explain He II emission. Thus, the hot component has a small radius, ∼0.1 $R_\odot$ for a luminosity of ∼100 $L_\odot$. The dynamical timescale for this radius, ∼300–500 s for a stellar mass of 0.5–1 $M_\odot$, is consistent with the flickering timescale. The escape velocity, ∼1000 km s^{-1}, is close to the maximum ejection velocity observed in most systems.

Despite this simple order-of-magnitude consistency, the hot component cannot *maintain* its large luminosity for the lifetime of the RED GIANT STAR. The $\sim 10^5$ yr cooling time for a 0.6 $M_\odot$ WHITE DWARF with a luminosity of 100 $L_\odot$ is 1% of the lifetime of its red giant companion. Both of these timescales are short compared to the $\gtrsim 10^9$ yr main sequence lifetime of either star. The probability of finding a hot white dwarf in a binary with a red giant is thus small. Perhaps 1–2% of symbiotics are produced in this way. The main alternative to such 'accidental' symbiotics is an interacting binary, where the hot component accretes material lost by the red giant. Generating ∼100 $L_\odot$ from gravitational infall requires accretion rates of $\sim 10^{-5}$ $M_\odot$ yr^{-1} onto a main sequence star and $\sim 10^{-7}$ $M_\odot$ yr^{-1} onto a white dwarf star. White dwarfs can burn the accreted material in a thin hydrogen shell above the white dwarf core, which reduces the needed infall rate to $\lesssim 10^{-8}$ $M_\odot$ yr^{-1}.

Detailed studies of individual symbiotic systems suggest three different types of hot components. Quiescent symbiotics with UV continua that rise towards short wavelengths, as in figure 1, contain hot white dwarfs with radii of ∼0.1 $R_\odot$ and temperatures that range from ∼30 000 K up to 200 000 K. The white dwarfs masses, 0.5–1 $M_\odot$, are similar to single white dwarfs. The flat UV spectra observed in roughly 10% of known symbiotics are better explained with emission from a disk having a temperature that decreases from ∼20 000 K at the inner edge to less than 5000 K at the outer edge. Radiation from the boundary layer—a $\sim 10^5$ K region where material at the inner edge of the disk slows to the rotational velocity of the central main sequence star—ionizes the nebula. Finally, two symbiotics are hard x-ray sources and contain accreting NEUTRON STARS. One of these, V2116 Oph, has a very high ionization optical spectrum and displays interesting x-ray variations. Little is known about the other neutron star symbiotic, HD 154791.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cool component

The structure of the cool component is best inferred from near-IR observations. These data separate symbiotics into two broad classes. *S-type* symbiotics contain G-, K- or M-type stars with color temperatures, $T_c \approx$ 2500–4500 K at 1–3 μm wavelengths. These red giants have strong near-infrared CO absorption bands and modest emission from circumstellar dust at 10–100 μm. Some vary on a timescale of 100+ d with 0.1–0.3 mag amplitudes at 1–3 μm. This amplitude often increases towards shorter wavelengths; radiation from the hot component masks this behavior in some systems. *D-type* symbiotics have smaller color temperatures, $T_c \approx$ 1000–1500 K, strong near-infrared H_2O absorption, and significant mid-infrared dust emission. Most D types also contain Mira variables that are several times more luminous than the red giants in S-type systems. The cooler, dusty photosphere must then have a factor of $\sim$10 larger radius than the radius of an S-type red giant. To contain such a large Mira variable within its tidal radius, the orbital periods of D-type systems should be 10–100 times longer than orbital periods for S-type symbiotics. All S-type symbiotics have short orbital periods, $\leq$3 yr. Reliable orbital periods for D types are rare, but most estimates range from decades to centuries.

Understanding mass loss from the cool component relies on optical, near-IR and radio observations. Several symbiotics have red giants that fill their Roche lobes and lose mass tidally at high rates. Most symbiotic red giants underfill their Roche lobes by factors of 2–10 and can only lose mass via a stellar wind. Two sets of measurements indicate high mass loss rates for these red giants. Radio spectral energy distributions at 1–20 cm allow estimates of the mass, radius, and outflow velocity of the ionized nebula in $\sim$50 symbiotics. Combining this information yields mass loss rates of $\sim 10^{-6}\ M_\odot\ \mathrm{yr}^{-1}$. Emission lines produced in the optically thin, outer atmosphere of the red giant require similar mass loss rates. The gravitational range of the hot component suggests that it can accrete roughly 1% of the mass lost by the red giant wind. This amount is sufficient to power the hot component providing a hydrogen shell burns the accreted material. The large mass loss rates estimated for the red giants in symbiotics exceed mass loss rates of single red giants with identical physical properties by 1–2 orders of magnitude. The origin of the high mass loss rates is uncertain.

Nebula

Despite the similarities of their hot stars, symbiotic nebulae have little in common with PLANETARY NEBULAE. Prominent forbidden lines and intense radio emission indicate smaller, denser nebulae in symbiotics, with densities of 10^5–10^6 cm^{-3} and sizes of 1–100 AU. Strong intercombination emission suggests densities exceeding $\sim 10^8$ cm^{-3} on scales of 1 AU or less in all S-type systems. These lines are rare in D types.

Photoionization produces most of the emission from symbiotic nebulae. Observations over several orbital cycles indicate a stratified nebula with density decreasing outwards from each component. Rapid eclipses indicate that high ionization lines—e.g. He II and N V—form in high-density material near the hot component. Slower occultations of lower ionization emission suggests formation in high-density material near the cool component. Several very high ionization features are produced by photoionization and scattering: Raman scattering of O VI $\lambda\lambda$1032,1038 emission lines off H I atoms in the red giant wind produces broad 6830 Å and 7088 Å emission bands. Understanding this process has led to direct measurement of the orbital inclination in several systems.

The wind from the hot component is often energetic enough to interact with the wind from the cool component. The shock front produced by these 'colliding' winds can produce observable amounts of thermal x-ray and high ionization line emission if the kinetic energy in the winds is large. A typical 1 $L_\odot$ x-ray luminosity requires a mass loss rate of $\sim 10^{-8}\ M_\odot\ \mathrm{yr}^{-1}$ from the hot component for an ejection velocity of 1000 km s^{-1}. In a few cases, the interaction is in the form of jets. First proposed to explain the variable velocity of [O III] emission in AG Peg, deep images of several D-type symbiotics show narrow jets of H I emission similar to those observed in planetary nebulae and some pre-main-sequence stars. The origin and frequency of this emission is unknown.

Models for symbiotic stars

Figure 3 shows simple schematics of typical symbiotic binary stars. The top panel shows a semi-detached system where the lobe-filling red giant transfers material into a large disk around a main sequence star. Radiation from the disk and boundary layer power the UV continuum and ionize the nebula. Ionized material surrounds the disk and one hemisphere of the red giant; neutral gas lies behind the red giant as viewed from the disk. The dynamics and ionization structure of the nebula are poorly known. Whether or not the disk ejects material into the ionized nebula is also uncertain. By analogy with dwarf novae, outbursts in these semi-detached symbiotics occur either when an unstable red giant increases the mass transfer rate into the disk or when an unstable disk increases the mass transfer rate onto the central star. Models and observations are currently insufficient to ascertain which of these two possibilities causes the outburst of any semi-detached symbiotic.

The middle panel shows the most common type of detached symbiotic where a white dwarf or neutron star accretes material from the wind of a normal red giant. Some material lost by the giant falls into a disk surrounding the white dwarf. The dynamics of the red giant wind and the material accreted onto the white dwarf are poorly understood. The accretion rate is high enough to support nuclear burning on the white dwarf, which produces a hot ultraviolet source that ionizes most of the red giant wind. The most luminous hot components eject material that interacts with the red giant wind. This interaction produces a bow shock around the red giant, as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

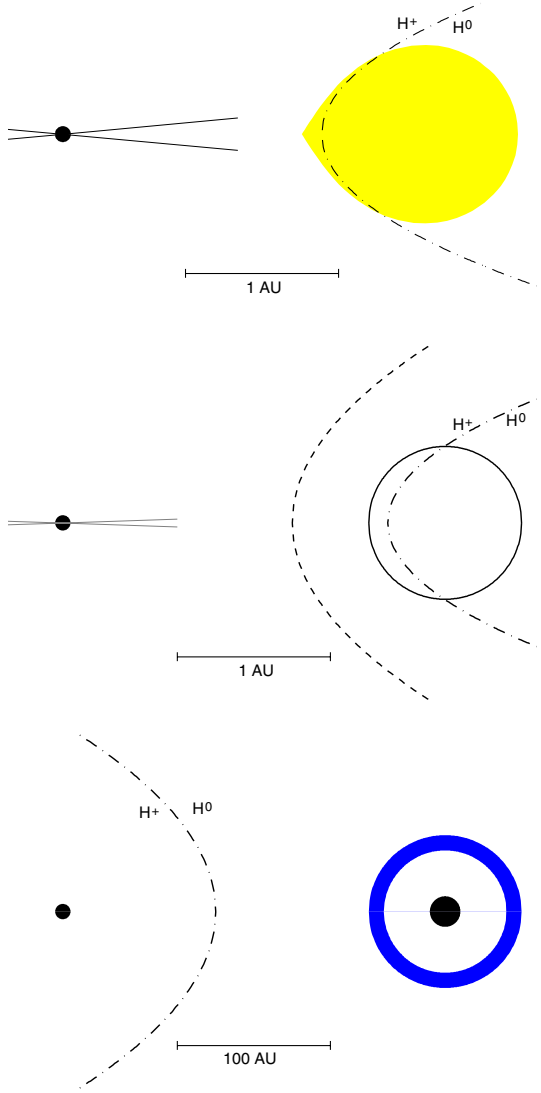


Figure 3. Schematic orbital configurations for three types of symbiotic stars. In each schematic, the red giant component is on the right and the hot component is on the left. The hot component is surrounded by a disk (shown from the side for simplicity). A length scale in AU is included with each panel. Regions of ionized hydrogen (H^+) and neutral hydrogen (H^0) are marked accordingly. The dashed line in the middle panel shows the general region where a wind from the hot star may interact with material lost by the cool star.

indicated by the dashed line between the hot component and the red giant. The dynamics of this region and how it affects the accretion flow are also poorly understood.

The bottom panel shows the geometry of a Mira symbiotic. The large mass loss rate from the Mira allows dust condensation in the wind. Dust condenses well outside the Mira photosphere, as indicated by the ring surrounding the red giant in the figure. This dust is often thick enough to obscure the Mira at optical and near-infrared wavelengths. The geometry and dynamics of material in the dust shell is not known. Despite the very wide orbital separation, the hot component accretes enough material from the wind to power an ultraviolet continuum and bright emission lines. The hot component often cannot ionize a large fraction of the Mira wind, so the ionization front curves around the hot component instead of around the red giant as in other symbiotics.

The outbursts of the two classes of detached symbiotics are usually interpreted as thermonuclear runaways in the hydrogen burning shells of white dwarf stars. This explanation accounts for the slow eruptions of symbiotic novae such as AG Peg. It fails for the rapid eruptions of recurrent novae and classical symbiotics unless the white dwarf is close to the Chandrasekhar limit. There is good evidence for a massive white dwarf in T CrB, but most symbiotics have hot components with masses of $\lesssim 1\ M_\odot$. Disk instabilities are a possible alternative to thermonuclear runaways but have not been studied in much detail.

In the past decade, symbiotic stars have generated renewed interest as possible progenitors of type I SUPERNOVAE. The small white dwarf masses in symbiotics make a carbon detonation supernova unlikely. In some accreting white dwarfs, however, hydrogen burning may produce a massive helium shell that can undergo a different type of supernova. Learning how to distinguish supernovae in a symbiotic binary from other types of supernovae remains an unsolved problem. There is also great uncertainty about the importance of the symbiotic contribution to the observed frequency of type I supernovae, due to the unknown frequency of symbiotic stars in the Milky Way and other galaxies and to the uncertain outcome of the helium shell explosion. Resolving these unknowns may allow Merrill's stars with combination spectra to contribute to the measurement of cosmological parameters such as H_0 and q_0.

Bibliography

Allen D A 1984 *A Catalogue of Symbiotic Stars, Proc. Astronomical Society of Australia* pp 369–421

Kenyon S J 1986 *The Symbiotic Stars* (Cambridge: Cambridge University Press)

Mikołajewska J 1996 *Physical Processes in Symbiotic Binaries and Related Systems* (Warsaw: Copernicus Foundation for Polish Astronomy)

Payne-Gaposchkin C 1957 *The Galactic Novae* (Amsterdam: North Holland)

Swings P 1970 Symbiotic stars and related peculiar objects *Spectroscopic Astrophysics* ed G Herbig (San Francisco, CA: University of California Press) pp 189–227

Scott Kenyon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synchronous Orbit

An orbit in which the period of revolution of an orbiting body is the same as the rotation period of the body it is orbiting. In the solar system, the only natural instance of a synchronous orbit is that of Pluto's satellite, Charon. Some artificial satellites are put into synchronous orbit around the Earth; such orbits are then termed *geosynchronous*. If in addition such an orbit is circular and in the Earth's equatorial plane, it is termed *geostationary*, since from a point on the Earth's surface a satellite in such an orbit appears to keep approximately the same position in the sky. Geosynchronous and geostationary orbits are used for navigation and communications satellites.

See also: synchronous rotation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synchronous Rotation

The rotation on its axis of a celestial object in the same time as it takes to orbit another object, also called *captured rotation*. In other words, the orbiting body's rotation and revolution periods are identical. Synchronous rotation results from tidal forces between the two bodies. Most of the major satellites in the solar system, the Moon included, have synchronous rotation and always keep the same face turned toward their parent planet. In the Moon's case libration means that in the course of time rather more than 50% of the Moon's surface is visible from the Earth. Pluto's satellite Charon is unique in having both synchronous rotation and a synchronous orbit: its period of revolution, and the rotation periods of both planet and satellite, are identical. From either body, the other is motionless in the sky. Synchronous rotation can also occur in binary stars with near-circular orbits.

See also: libration, synchronous orbit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synchrotron Radiation

Radiation emitted by electrons moving at very high speeds (approaching the speed of light) in magnetic fields. An electron follows a helical path around a magnetic field line and, at relativistic (i.e. close to the speed of light) velocities, emits radiation in a narrow cone in the direction of its motion. Synchrotron radiation has a characteristic spectrum and generally exhibits a high degree of polarization. The synchrotron process is an important source of radiation from astronomical sources. For example, around the planet Jupiter high-energy electrons spiraling through the lines of force of the planet's magnetic field give off synchrotron radiation at radio wavelengths. In the Crab Nebula synchrotron radiation at such wavelengths and at those of visible and ultraviolet light is generated by electrons moving in the magnetic field associated with the supernova remnant. Also radio emissions of the synchrotron variety have been detected from other supernova remnants in our Galaxy and from extragalactic objects such as radio galaxies and quasars.

See also: electromagnetic radiation, spectrum.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synodic Period

The time interval between two successive similar alignments of two celestial bodies. In the case of a planet, the synodic period may be taken to be the mean time interval between, say, two successive oppositions or two successive conjunctions. It is this period that determines the times at which particular planets will be visible in the night sky. The Moon's synodic period is the time between successive recurrences of the same phase; such as between full moon and full moon.

See also: sidereal period.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Synthetic Stellar Populations

Modern telescopes are built with ever-better angular resolution, but for galaxies and clusters outside the local group we generally cannot hope to study the stars in them one by one. Instead, we see the sum of the light from all of the stars. To take this integrated starlight and extract information on the ages or heavy-element abundances requires a model; a synthetic version of how many and what kind of stars are present. To guide the models, we rely on nearby stars and star clusters where stars can be studied one by one and also on theoretical models of stellar evolution and stellar atmospheres.

A STELLAR POPULATION is a group of stars born at the same time and sharing the same initial composition. A STAR CLUSTER is an example of a simple stellar population, characterized by a single age and a single abundance mixture. The MILKY WAY is an example of a complicated mixture of many stellar populations; old and young, metal poor and metal rich.

In the Milky Way, there is a striking concordance of age–metallicity with spatial extent and kinematics. The GALACTIC METAL-POOR HALO is thought to have formed in the first third of the Galaxy's history and is composed of stars and clusters on randomly oriented orbits so that the net effect is a large halo of roughly spherical shape. On the other hand, if one finds all the stars in the Galaxy younger than a billion years, virtually all of them are in the very thin Galactic disk. They have orderly, near-circular orbits with all stars rotating clockwise as seen from above the Galactic north pole (see GALACTIC THIN DISK). This dichotomy gave rise to the notion of two distinct stellar populations in the Milky Way around 1944 (W Baade is credited with this idea). By the 1960s it became evident that there were several, rather than two, distinct populations. For instance, old but metal-rich stars exist both in the disk and in the bulge. Today, we have given up counting the number of stellar populations, and instead speak of the age and abundance distributions within kinematic subgroups.

From distant galaxies and clusters, what information can one hope to derive from integrated star light? Ultimately, one could hope to find the complete star formation history of the galaxy, and at each past epoch of star formation the population's dynamic state, a distribution of heavy elements, and an initial mass function (abbreviated IMF; the number of stars born at each mass). Such a detailed picture is out of reach for the present. Today, we have clues about each of these quantities, but we are far from being able to wrestle such a detailed picture from integrated light observations.

Several types of synthetic stellar population models

Figure 1 shows galaxy spectra covering the range of spectral shapes that are seen along with a sampling of some of the stellar types out of which the galaxies are composed. From the figure one can glimpse similarities between the galaxies and the underlying stellar types.

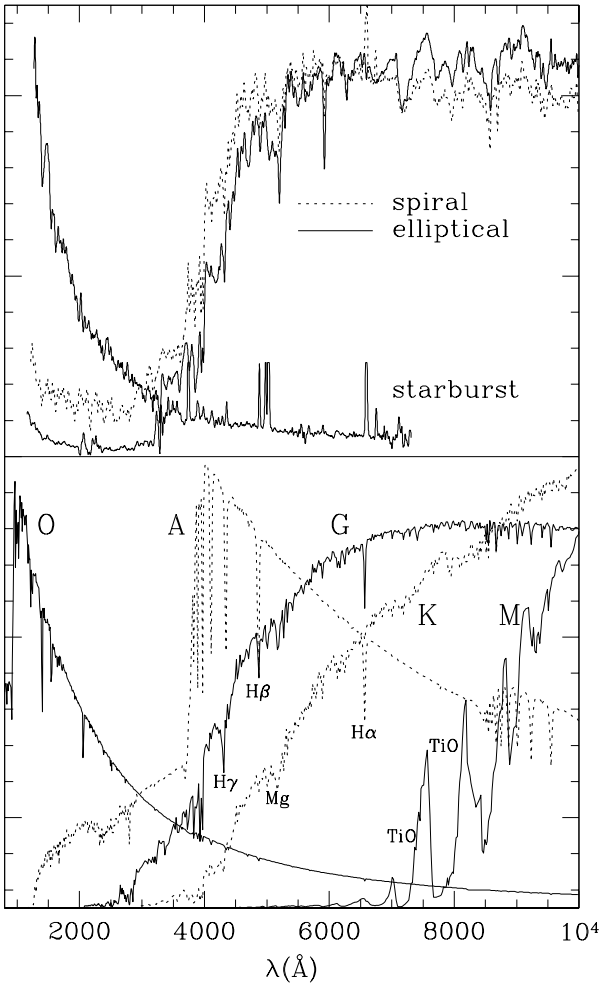


Figure 1. Example galaxy (top) and star (bottom) spectra are shown, arbitrarily scaled F_ν versus wavelength in Å. The strongest emission lines in the starburst spectrum were truncated for clarity, and a few important spectral features are marked on the stellar spectra. This is the synthesis problem: adding the star spectra together in different combinations and with different multiplicative weights, how closely can the galaxy spectrum be matched? By eye, one can see that the starburst galaxy must have a strong O star component, and the spiral and elliptical galaxies must have many G stars. However, TiO features appear in the spiral and elliptical spectra, so KM stars must also be present.

Extreme STARBURST GALAXIES emit most of their light in the ultraviolet. Their spectra are matched by OB-type stars with strongly muted evidence of other spectral types. SPIRAL and ELLIPTICAL GALAXIES have minority populations of hot stars, as seen by their small ultraviolet fluxes, and obviously have a strong G-star component. However, the strong, broad TiO features characteristic of M-type stars are seen in the red portion of the spectrum, so very cool stars are present as well.

The first type of model is this kind of star-by-star

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

addition problem. Early workers such as Spinrad and Taylor would observe, spectrophotometrically, a collection of stars and a few galaxies. Then, by method of trial and error, they would construct a 'model' consisting of the relative proportions of the stars that best matched the galaxy spectrum. That is, given a galaxy spectrum L_λ and a collection of stellar spectra $l_{i,\lambda}$, the model consists of the weights w_i by which the stars are multiplied:

$$L_\lambda = \sum_i w_i l_{i,\lambda}. \quad (1)$$

The trial-and-error part comes in because the same set of weights must work for all λ. A useful improvement came with the work of Faber in the early 1970s. Through a careful statement of the problem and judicious use of technique (quadratic programming or linear programming) the trial-and-error portion of assigning weights was replaced by numerical optimization.

This form of population modeling is often termed empirical population synthesis. Empirical population synthesis has a couple of drawbacks. First, solutions are not unique. For instance, the spectrum of a G dwarf closely resembles that of a G giant, and replacement of one by the other probably will not noticeably affect the goodness of fit. Second, information on galaxy age or metal abundance requires interpretation. Suppose the synthesis says that G0 dwarfs are numerous, but F8 dwarfs are rare. The interpretation would be that the main sequence turnoff is at spectral type G0. One would then refer to observed star clusters or stellar evolutionary isochrones to find out the age at which the main sequence turnoff is spectral type G0.

The connection to age and metal abundance can be made more direct by considering, not stars, but star clusters as the basic building block. The cluster will already have the 'correct' proportions of the different kinds of stars for the cluster's age and abundance, and finding the best match between target (galaxy) and template cluster should, in principle, yield a mean age and metal abundance for the galaxy. In practice, two problems arise. First, in the Milky Way and vicinity, open clusters are so diffuse that obtaining an integrated spectrum is observationally impossible. One is then left with the globular clusters and the clusters in the LMC and SMC for possible spectral templates. E Bica has made a study of such cluster spectra. While these clusters span a considerable range of age and abundance, not all possibilities are covered. In particular, the high metallicities typical of elliptical galaxies cannot be matched by local template clusters. The second problem is age–metallicity degeneracy, which is the observation that changes in age and metal abundance have very similar, almost indistinguishable, effects on the integrated spectrum, so that an old, metal-poor population may very closely resemble a young, metal-rich population. This topic is revisited below.

The final variant of model was developed by B M Tinsley starting in the late 1960s and today is the most widely used variant. It is usually referred to as evolutionary population synthesis or isochrone synthesis because it uses theoretical stellar evolutionary isochrones as its basic unit rather than observed clusters. A stellar evolutionary isochrone (isochrone for short) is made from a collection of evolutionary tracks for different masses of stars. A track describes the luminosity and effective temperature of a star of a given mass with time. An isochrone shows the locus of luminosities and temperatures at one instant in time for stars of all masses, and thus is built to mimic a star cluster or a single-age stellar population.

Sample isochrones together with some terminology

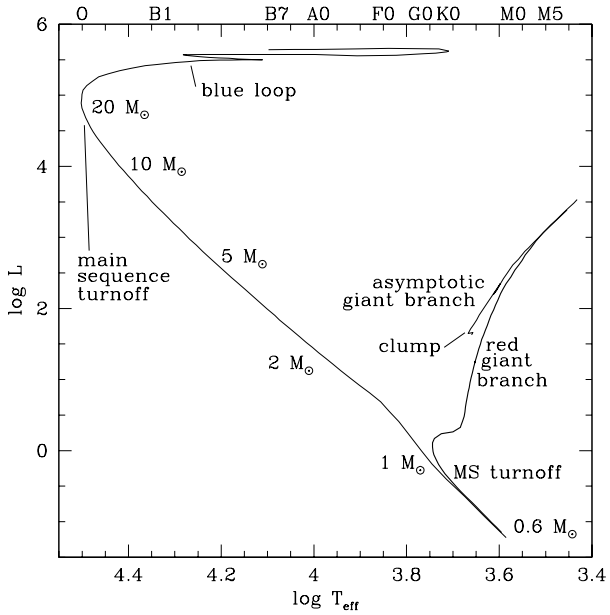


Figure 2. Isochrone age effects are shown. The isochrones plotted are for solar composition, ages 5 million and 15 billion years. Approximate spectral type is indicated on the top axis, and log $T_{\rm eff}$(in kelvins) on the bottom. The vertical scale is the logarithm of the bolometric (total) luminosity in units of the solar luminosity. The masses of main sequence stars are marked. The 'main sequence turnoff' is the point where the isochrone drifts brighter than the zero-age main sequence. Stars in this phase often dominate the integrated light and are also crucial for age estimation. Post-main-sequence evolution happens on much quicker timescales than main sequence lifetimes. Massive stars undergo 'blue loops' in response to internal structural changes, and may also spend some time as M-type supergiants before they explode in supernovae explosions. Low-mass stars have entirely different post-main-sequence evolution. They evolve up a 'red giant branch' as they fuse hydrogen in a shell around the helium core. At log $L \approx 3.4$, the helium reaches ignition temperature and the star immediately adjusts its structure to land in the 'clump'. In a metal-poor population, the clump may spread over a large range of temperature, and in this case is called the 'horizontal branch'. After the clump, the star climbs the asymptotic giant branch (AGB) before shedding its outer layers, undergoing a planetary nebula phase, and finally becoming a white dwarf.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

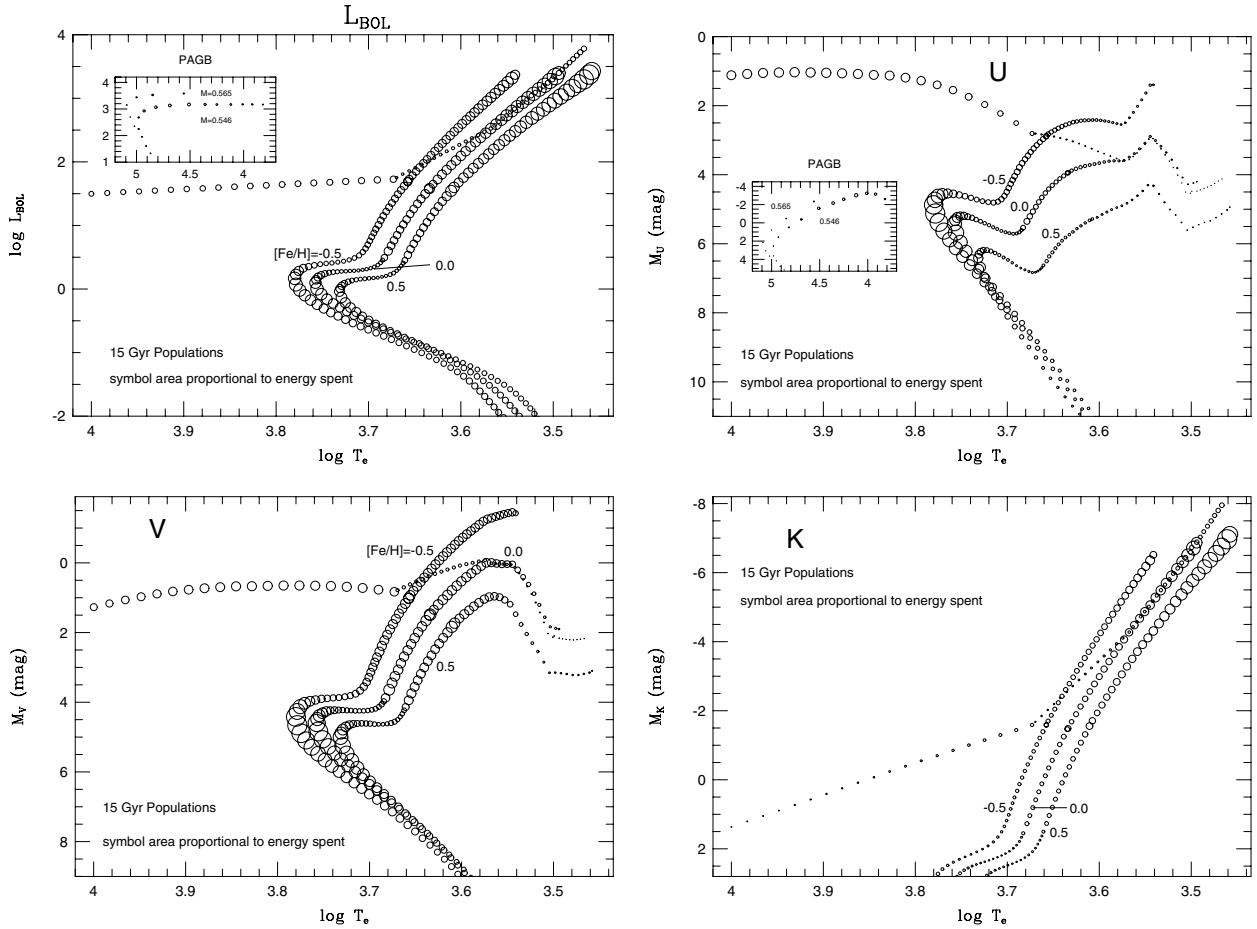


Figure 3. The four panels in this figure give a visual impression of which phases of evolution dominate the integrated light at different wavelengths. Isochrones of 15 billion year age are shown for [Fe/H] = −0.5, 0.0 and 0.5 dex. The effect of changing the abundance is to shift the isochrones in temperature. The [Fe/H] = 0.0 isochrone has a 'horizontal branch', an 'asymptotic giant branch' and a 'post-asymptotic giant branch' (marked PAGB) showing evolution in and around the planetary nebula phase for two postulated white dwarf masses. For clarity, these late stages are not illustrated for the other two isochrones. The horizontal branch is included for illustration purposes only; the usual morphology of this phase for a metal-rich population is a clump. Each bin along the isochrone is represented by a circle, which changes size according to how much energy is spent at that location. Therefore, most of the light comes from where the largest circles are located. The four panels show bolometric (all-wavelength) light in which the light contributions are split between main sequence turnoff and giant branch, the ultraviolet U filter in which only the warmer stars contribute, the visual V filter which resembles the bolometric panel and the infrared K filter in which the brightest giants, all but invisible in the other panels, completely dominate the light.

are shown in figure 2. There are two essential points about isochrones. First, more massive stars have much shorter main sequence lifetimes, so the main sequence turnoff point will move steadily dimmer and cooler with increasing age. Second, the post-main-sequence phases of evolution are much shorter than the main sequence lifetimes. In most situations stars near the main sequence turnoff region will dominate the integrated light. One case where this is not true is for an old population where one is restricted to looking in the red or infrared. In that case, the cool red giant stars dominate the light. The isochrones in figure 2 are of SOLAR ABUNDANCE, but model builders also consider other abundances. We use the standard bracket notation [Fe/H] or [M/H] to mean the logarithmic heavy-element abundance in units of the solar heavy-element abundance:

$$[\mathrm{Fe/H}] = \log[n(\mathrm{Fe})/n(\mathrm{H})]_{\mathrm{stars}} - \log[n(\mathrm{Fe})/n(\mathrm{H})]_{\odot}. \quad (2)$$

In addition to the isochrone, a collection of stellar spectra is required. These spectra can be empirical, theoretical or any combination. The spectra may even consist solely of broad-band colors, if that is the only information desired of the model at the end. The final model spectrum is a weighted sum of the stellar spectra mathematically identical to equation (1):

$$L_\lambda = \sum_i N_i(M, \delta M) l_\lambda(T_{\mathrm{eff}}, g, Z) \quad (3)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

where L_λ is the model spectrum, i is the index over points along the isochrone, $N_i(M, \delta M)$ is the number of stars associated with each point along the isochrone, and $l_\lambda(T_{\rm eff}, g, Z)$ is the stellar spectrum, which is typically interpolated from a grid of star fluxes indexed by effective temperature, surface gravity and abundance. The number of stars comes from the IMF, which gives the number of stars per interval of mass. A simple power-law IMF is still in common use today, for example

$$N_i(M, \delta M) = CM^{-x}\delta M \tag{4}$$

where C is a constant and $x = 2.35$ is the 1955 Salpeter estimate. Masses for each isochrone point are tabulated in the isochrones together with temperature and luminosity.

Examples of how the addition goes are shown in figure 3, in which isochrones are plotted point by point. The sizes of the points are varied so that the area of each is proportional to $N \times l$, the total luminosity coming from that isochrone point. Therefore, one can obtain a visual impression of what evolutionary phases are important by looking for the larger symbols. In total energy output ($L_{\rm bol}$) it is evident that the bright red giants are dominant, with main sequence turnoff stars secondary. In the U filter, the giants contribute almost nothing, with the main sequence turnoff stars dominant. The horizontal branch, if present, is a significant source of light. In the V filter stars of many different kinds contribute, but in the near-infrared K filter nothing but the brightest KM giants contribute significantly. For populations younger than about half a billion years, the giant branch does not develop as illustrated, and the main sequence stars are more important.

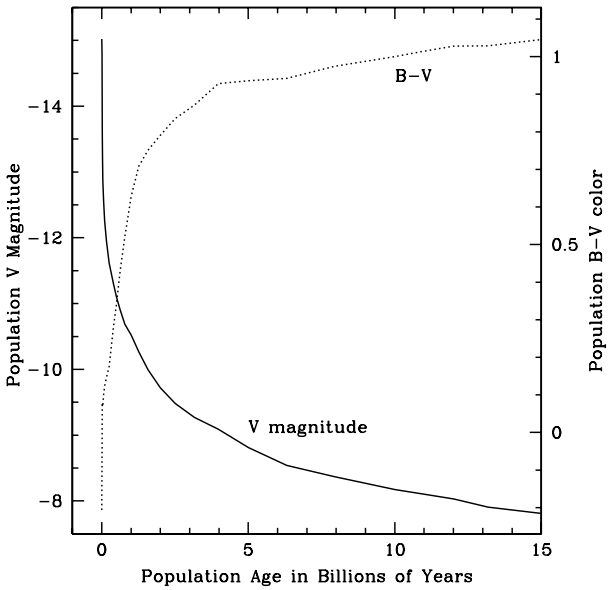


Figure 4. Population fading in the V band is shown as a function of age. The V magnitude of a population of $10^6 M_\odot$ is shown by the solid curve, and the left-hand axis gives the scale. Color reddening is also shown where $B - V$ color is shown by the dotted curve and the right-hand axis gives the scale.

Synthesis model behavior and uncertainties

The basic behavior of a stellar population is (1) the population becomes dimmer with age, (2) the population becomes dimmer with increasing abundance in passbands bluer than about 1 μm, brighter in passbands redder than that, (3) the population becomes redder with age or increasing abundance and (4) neutral metallic lines increase in strength and, after about 10^8 yr, hydrogen Balmer lines decrease in strength with increasing age or abundance. The age caveat in point (4) arises because the hot OB stars present in very young populations have relatively weak Balmer lines, and when those stars dominate the Balmer lines fade somewhat. The Balmer lines peak in strength for turnoff stars at about A0 spectral type ($\sim 7 \times 10^8$ yr isochrone age) and fade for cooler stars. Metal lines increase monotonically with decreasing star temperature. The behavior is shown in figure 4, which shows dimming of a population in the V band, and reddening of the $B - V$ color.

The rate of dimming with age is controlled by the IMF, which controls how many stars exist at each mass. If there are relatively more massive stars in a population, it will be brighter when young. The IMF does not strongly affect the color evolution and is almost negligible in its effects after a substantial red giant branch develops, a bit shy of a billion years of age. This is because almost all the light comes from stars near the main sequence turnoff or from post-main-sequence stars. All of these stars have nearly the same mass (because post-main-sequence lifetimes are relatively short) so there is little mass differential for the IMF to work on.

Since we do not know how much the IMF varies from environment to environment there is corresponding uncertainty in the synthesis models. This is an astrophysical uncertainty, but there are technical uncertainties as well since most of the ingredients in an isochrone synthesis model are suspect at some level. The isochrones themselves appear to be the dominant source of error. Stellar lifetimes and luminosities are somewhat uncertain, but the temperatures predicted by the isochrones are the most critical component and one of the hardest to obtain accurately. Temperature errors lead directly to errors in derived ages and abundances. The stellar flux libraries have problems also. Theoretical fluxes do not match real stars, especially in the ultraviolet and infrared where the atomic and molecular line lists are incomplete. Observed libraries have observational error, limited wavelength coverage, uncertain flux calibration and often inappropriate resolution. The combined effect of these technical problems is estimated to render derived ages uncertain by about 35% for populations older than a billion years. The uncertainty is larger for younger populations.

There are sometimes problems in the observational material that make it difficult to apply a model based solely on star light. Dust may be present in the observed galaxy,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

possibly with a variety of geometries. This will affect colors owing to scattering and absorption of starlight. If stars are currently forming in a galaxy, nebular emission lines from H II regions are present. Together with other species, the hydrogen Balmer lines are present in emission, filling in the stellar absorption lines. This can make the Balmer absorptions, which are important age diagnostics, very hard or impossible to measure.

Another significant hurdle for all types of population models is age–metallicity degeneracy, in which age effects look similar to metallicity effects in integrated light. To illustrate the problem figure 5 shows isochrones and fluxes for a trio of populations. The 5 billion year old, $[\mathrm{Fe/H}] = -0.1$ dex population is the reference population. The second population is three times older than the reference, the third twice as metal rich, but their spectra fall virtually on top of one another. The null-change slope is approximately $\Delta \log(\mathrm{age}) = (-3/2)\Delta \log Z$.

Is age hopelessly intertwined with abundance? Fortunately, it is not. A careful look at the model spectra reveals several features that do not participate in the $-3/2$ slope. Some metallic absorption features are relatively more sensitive to metals. Hydrogen Balmer lines track the main sequence turnoff temperature closely without much regard for the giant branch temperature and hence are relatively age sensitive. Ultraviolet-to-red colors behave the same way (but in practice are hampered by the presence of a minority of hot horizontal branch or post-AGB stars and by dust, if present). The presence of luminous red AGB stars has been proposed as another age indicator, since the AGB is observed to be very strong in clusters less than about 2 billion years of age, but the stellar evolutionary details of this are difficult to work out. Additional age indicators are being sought by today's researchers. For the moment, it appears that arraying a Balmer feature against a sensitive metal feature is the most reliable way to estimate simultaneously a mean age and a mean metal abundance. Example diagnostic diagrams are shown in figure 6.

Another astrophysical uncertainty is the mixture of elements that are used for the 'abundance'. A scaled-solar abundance pattern seems to apply approximately throughout the local disk, but metal-poor halo stars have light elements that are more abundant relative to Fe-peak elements. A similar abundance pattern applies in the GALACTIC BULGE and also in large elliptical galaxies (see below). The impact of changing the heavy-element ratios is primarily on the temperature structure of the isochrone and the appearance of the stellar spectra, with secondary effects on stellar lifetimes and luminosities. However, the modeling that is required to calibrate these changes is only now starting to be carried out, so the ultimate impact of differing elemental mixtures is difficult to estimate. As it stands now, the local template stars have spectra which differ in line strength pattern from the spectra of some galaxies and clusters, so a perfect model-to-galaxy match is not possible.

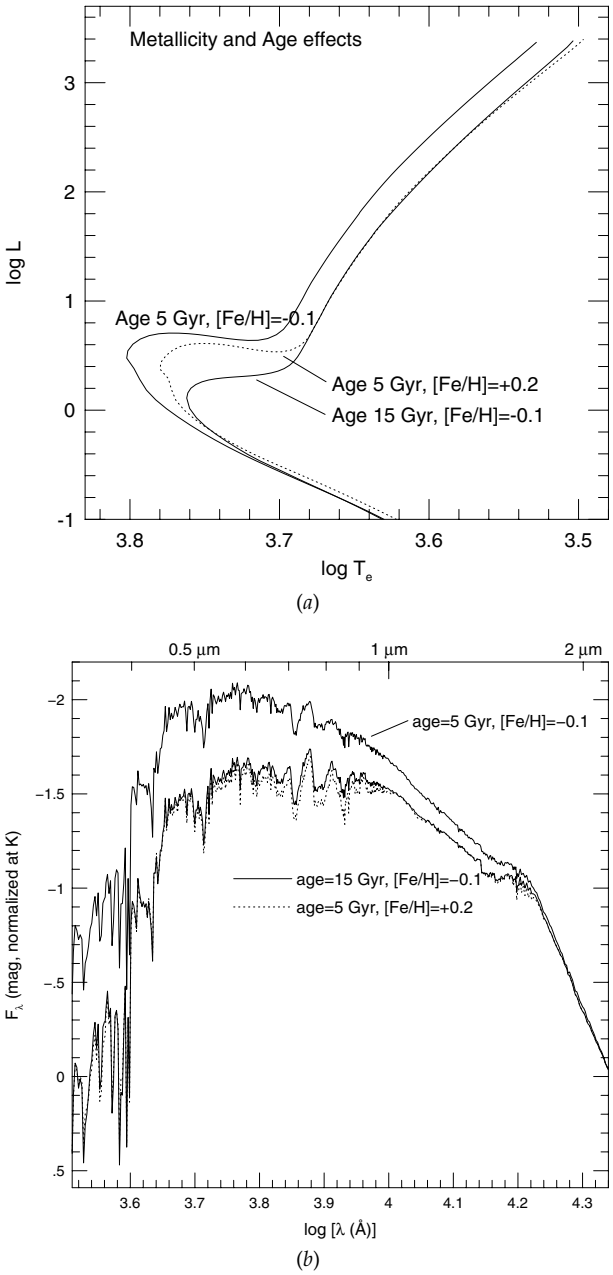


Figure 5. (*a*) Matched isochrones and (*b*) model spectra that illustrate age–metal degeneracy are shown. The base population is age 5 billion years, with slightly less than solar metallicity. Of the other two populations, one has age tripled, the other has metallicity doubled. Despite differing in abundance by a factor of 2 and differing in age by a factor of 3 both spectra lie nearly on top of one another.

A final astrophysical uncertainty is the difficulty of composite populations, systems that are composed of many different ages and abundances. In particular, there is a strong masking effect when a young population is present together with an underlying old population.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

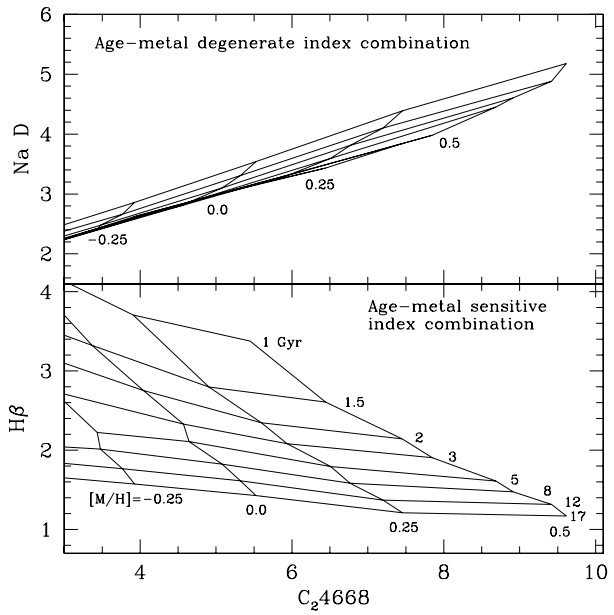


Figure 6. While most index–index diagrams are age–metal degenerate, as in the top panel showing an Na D index versus an index measuring mostly molecular carbon, some show fairly good ability to separate the two effects, as in the bottom panel showing Hβ. Note that an 'index' could be a color, an absorption feature index or even an SBF magnitude. The choice of indices presented here is arbitrary. Metallicity values and ages in billions of years are marked on the grid lines that represent simple-population models.

Since young populations are very much brighter, they will dominate the integrated light even when the young population represents only a few per cent of the total mass. When 'mean ages' are derived for galaxies, the youngest population present gets far more weight than the older ones.

Applications of synthesis models

Over the years we have learned a great deal about distant GLOBULAR CLUSTERS and galaxies from the study of their integrated starlight as interpreted through synthesis models. The following examples illustrate some of the ways that this class of models has been applied.

Globular cluster systems are attractive targets because all but the very largest of them appear to be not only co-eval but also monometallic so that they can be characterized by a model of a single age and metal abundance. The first globular cluster system to be studied was the Milky Way's. Direct stellar abundance estimates are available for these nearby systems, so the strong connection between integrated color (or metal line strength) and abundance is obvious. The Milky Way globular clusters are believed to have formed at various times during the first third of the Galaxy's life so they are uniformly ancient. They remain crucial benchmarks for testing synthesis models.

Colors of globular clusters around other galaxies are now becoming available out to ∼8000 km s^{-1} and low-resolution spectra can be obtained out to ∼1500 km s^{-1} with 8 m telescopes. Clusters around elliptical galaxies are the most common target. Metal abundance can be derived by assuming old ages and using synthesis models to go from observed color to abundance. Note that the −3/2 age–metal slope works in our favor when metal abundance is sought: a 30% age uncertainty corresponds to a 20% abundance uncertainty, which is less than 0.1 dex. The typical globular cluster system has both metal-poor and metal-rich clusters, some of which have apparently greater than solar abundance. The Milky Way is missing the metal-rich component seen in most other galaxies. The metal-rich component often appears as a separate hump in the abundance histogram and is seen to be more centrally concentrated than the metal-poor clusters.

Elliptical galaxies show a color–magnitude relation in which brighter galaxies are redder. Since it is unlikely that smaller galaxies are all younger than their larger brethren, and since synthesis models predict reddening with increasing abundance, a higher abundance in the larger ellipticals is inferred.

Mean ages of early-type galaxies without emission lines can be estimated from Balmer absorption strength arrayed against metallic feature strengths. The results for local galaxies show a strong clump of old-appearing ages, with a significant trail of galaxies toward young ages. This indicates a degree of star formation activity inconsistent with a universal primordial formation for all elliptical and S0 galaxies.

Light/heavy element ratios in elliptical galaxies are not constant. When absorption features rather than colors are examined, Fe or Ca features are nearly constant regardless of galaxy size, but features sensitive to Mg, Na and N abundance are stronger in larger galaxies. This was discovered when synthesis models were compared with absorption line data. On the positive side, this tells us much about the types of SUPERNOVAE responsible for the chemical enrichment in galaxies of different sizes. On the negative side, this introduces more parameters to be encompassed by the synthesis models, increasing uncertainty.

Surface brightness fluctuations are pixel-to-pixel variations about the mean brightness of an image of a galaxy due to almost-resolved stars. The further away a galaxy is, the smoother its appearance because more stars fit into each pixel, suppressing the 'noise'. Tonry and Schechter quantified this phenomenon and exploited it as a distance indicator with potentially excellent accuracy. The size of the observed surface brightness fluctuations depends only on galaxy distance and the luminosity function of the brighter stars present in the galaxy (see also LUMINOSITY FUNCTION OF GALAXIES). The luminosity function can be predicted from population models, which show a slope with color and predict an absolute zero point for the distance scale.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Post-starburst objects can be reliably found with variations on the technique of using Balmer lines in conjunction with features more sensitive to abundance. Large numbers of galaxies that look quiescent but in fact have had a large burst of star formation in the recent past have been discovered in nearby GALAXY CLUSTERS, supporting a view in which galaxy clusters are in the midst of forming even today.

Cosmological studies of GALAXIES AT HIGH REDSHIFT almost always require evolutionary synthesis models to interpret the observed magnitudes and colors. The Doppler shift moves the spectrum to longer wavelengths, affecting the colors. Furthermore, at large look-back times the stars seen will be younger than their local counterparts, and therefore brighter and bluer. The synthesis models allow an accounting of these effects.

Starburst galaxies are bravely examined, despite dust and nebular emission, with the hope of deriving a high-mass IMF, among other things. The key observables are the C IV and Si IV resonance lines in the vacuum ultraviolet. Massive stars have winds that generate P Cygni profiles in these lines. More massive stars have faster and denser winds, generating stronger P Cygni profiles with larger blueshift. This technique offers a sensitive IMF test for stars with $M > 30M_{\odot}$. To date, the evidence suggests that the IMF looks about the same in distant STARBURST GALAXIES as it does in local group star forming regions.

The G dwarf problem is the observation that the solar neighborhood is deficient in stars of abundance less than 1/10th of solar abundance compared with the simple closed-box model of chemical evolution. Synthesis models have been used to estimate the metal-poor population fraction present in early-type galaxies by looking at the amount of main sequence turnoff light contributed in the near-UV and at spectral features sensitive to the presence of A-type metal-poor horizontal branch stars. The conclusion is that the G dwarf problem is universal. Other galaxies, even the nuclei of elliptical galaxies, do not contain as many metal-poor stars as the simple model predicts.

Chemo-dynamical models attempt to start with a cloud of gas and numerically turn it into a galaxy, with correct dynamics and chemistry along the way. These complex computer models can also include a population synthesis portion so that the surface brightnesses and colors of the model galaxies can be simulated as well. Today's models are approximate and underconstrained, but this is a promising avenue for future work.

Bibliography

For references to recent literature on virtually all of the topics touched upon in this article see the conference review

Leitherer C, Fritze-von Alvensleben U and Huchra J (ed) 1996 *From Stars to Galaxies: The Impact of Stellar Physics on Galaxy Evolution (ASP Conf. Ser. 98)* (San Francisco, CA: Astronomical Society of the Pacific)

Guy Worthey

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Syrtis Major

A prominent dark, triangular albedo feature (dark by contrast with its surroundings) on the Martian equator, centered approximately at 10 °N, 70 °E. It takes its name from the historical name for the larger of two quicksands off the North African coast. Syrtis Major is prominent in telescopic views of the planet, and was in fact the first feature of Mars ever to be recorded, in a sketch made by Christiaan Huygens in 1659. It continued to be recorded by observers of Mars, and appears on various maps made during the nineteenth century, including those drawn by Nathaniel Green and Eugène Antoniadi. When spacecraft mapped the planet in the 1970s very few of the classical albedo features were found to correspond to topographical features. Syrtis Major is one of the exceptions, coinciding with a volcanic plateau centered at 9.5 °N, 70.4 °E. The plateau, which bears the official name Syrtis Major Planum (changed in 1982 from Syrtis Major Planitia), is a flat, lightly cratered area that slopes gently down toward Isidis Planitia to the east. A low-relief shield volcano on the plateau is probably responsible for the dark material that covers the region; winds shift this material around, causing changes in the appearance of Syrtis Major as seen from Earth and creating dune fields. In the southwest lies a prominent crater, 456 km in diameter, appropriately named Huygens, while in the northwest lies another large crater, the 381 km diameter Antoniadi.

See also: Mars: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Syzygy

The approximate alignment of three celestial bodies. Examples are the Earth, Sun and Moon at new moon and full moon, or the Earth, Sun and a planet when the planet is at opposition or conjunction.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

T Tauri Stars

The T Tauri stars are the most well known class of pre-main sequence stars. They are low- to intermediate-mass stars with ages of typically a few million years. At this stage in their evolution, the newly collapsed stars have become optically visible but are still surrounded by gas and dust from their formation, typically in a disk. In the HR diagram, they are located where one expects contracting, fully convective pre-main sequence stars to fall. Observationally they appear as cool, luminous stars that exhibit emission lines, excess infrared and ultraviolet continuum flux, variability and are sometimes associated with jets, molecular outflows, winds, accretion or x-ray emission. Studies of the T Tauri stars help provide insight into the formation of the Sun and solar system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Talbot, William Henry Fox (1800–77)

Mathematician, parliamentarian and pioneer in photography, born in Melbury Abbas, Dorset, England, a close friend of Sir John Herschel (see HERSCHEL FAMILY), and worked with him on problems in light and photochemistry. Invented the modern methods of reproducing photographs through the negative process and making prints.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tarantula Nebula (NGC 2070)

An emission nebula in the Large Magellanic Cloud, in the constellation Doradus, position RA $05^{h}\,39^{m}$, dec. $-69^{\circ}\,07'$. At half a degree across, it is far larger than any emission nebula visible in our Galaxy. It is energized by several young hot O and B type stars, the brightest of which is designated R136. The Tarantula gets it name from its spidery appearance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tarde, Jean (1561/2–1636)

French parish priest, geographer and astronomer, born in La Roque-Gajac, near Sarlat, France, adopted Copernicanism and set up an observatory to view sunspots, which he thought were satellites of the Sun. He named them 'Bourbon stars', in the same manner that GALILEO had named the satellites of Jupiter the Medician stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tartu Observatory

Tartu Observatory (TO) is a research institution in Estonia accommodating the northernmost 1.5 m telescope in the world. It is located in Estonia, about 20 km southwest of Tartu in the village of Tõravere (58°16′08″.4 N, 26°27′32″.4 E). TO performs research in astrophysics and atmospheric physics and popularizes those branches of science. TO was founded in 1808 as an observatory of Tartu University and it gained worldwide renown under the leadership of F G W Struve (1793–1864, director of the Observatory 1820–39). In 1824 TO was equipped with a 9 in (23 cm) Fraunhofer telescope which was the world's largest refractor. TO was working as the Observatory of the University until World War II. In 1946 the Observatory was transferred to the Estonian Academy of Sciences. In 1958 building works at a new observing site (Tõravere) were begun, and in 1964 the new observatory was opened. Between 1947 and 1995 the observatory had various different names. Since 1995 the official name has been Tartu Observatory and it is subordinated directly to the Estonian Ministry of Education.

The staff in 1999 was 64, of whom 38 were research associates.

Major observing facilities include the 1.5 m telescope (installed in 1975) equipped with a Cassegrain spectrograph, and the 0.6 m telescope (installed in 1998) with a CCD photometer.

The research can be divided into four main projects: formation and evolution of galaxies in the expanding universe; structure and evolution of stars and stellar atmospheres; climate forming factors and dynamical weather forecast for Estonia; and the radiation regime of the atmosphere and ground surface and optical monitoring of the Estonian environment.

TO is supported mostly from the state budget of Estonia. Minor additional finances are obtained from a few international projects and contracts with Estonian organizations.

For further information see
http://www.aai.ee.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tata Institute of Fundamental Research

The Tata Institute of Fundamental Research, Department of Astronomy and Astrophysics has over 20 scientists carrying out research in theoretical astrophysics, optical and near-infrared astronomy and space astronomy (far-infrared, x-ray and gamma-ray) using balloons and satellites. Currently a 2–20 keV x-ray telescope onboard an Indian satellite is operational. The institute operates a fully fledged balloon facility in Hyderabad that can launch 500–1000 kg payloads to altitudes of 42–35 km.

For further information see http://www.tifr.res.in.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Taurids

A meteor shower that takes place in October and November. There are two radiants in the constellation Taurus, at dec. +14° and +22°. The meteor stream consists of debris ejected from the now feebly active Comet 2P/Encke over a period of several thousand years, and is one of the most evolved streams encountered by the Earth. The stream has spread out over a broad swathe of the inner solar system, and Taurid shower activity lasts for six weeks, with low but steady rates of slow, bright meteors. The Earth also passes through the stream between 5 June and 18 July, producing the daytime *Beta Taurid* shower. Although they are not observable visually, as the radiant is near the Sun, radio observations indicate that the Beta Taurids are highly active.

See also: meteor shower.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Taurus

(the Bull; abbrev. Tau, gen. Tauri; area 797 sq. deg.) A northern zodiacal constellation which lies between Aries and Orion, and culminates at midnight in late November. It is one of the oldest constellations, dating back to when the Sun was in that part of the sky at the vernal (spring) equinox, between about 4000 and 1800 BC. Later, in Greek mythology, it was identified with the form assumed by Zeus when he carried away Europa, daughter of King Agenor of Phoenicia. On early celestial charts only its head and forequarters are shown. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A conspicuous constellation, Taurus is easily recognized by virtue of α Tauri (Aldebaran), a red giant (K5) irregular variable (range 0.75–0.95), which marks the bull's right eye and is the top-left star of a V-shaped grouping of a couple of dozen more distant naked-eye stars, which mark the bull's face and represent the brightest members of a scattered open cluster of several hundred stars, the Hyades.

Other bright stars include β Tauri (Elnath, formerly designated γ Aurigae), magnitude 1.7, η Tauri (Alcyone), magnitude 2.9 and the brightest member of another open cluster, the Pleiades, θ^2 Tauri, magnitude 3.4, the brightest member of the Hyades and the brighter component of a wide, naked-eye double with θ^1, magnitude 3.8, separation >5′, which has the same proper motion, and λ Tauri, an Algol-type eclipsing binary (range 3.4–3.9, period 3.95 days). There are ten other stars brighter than magnitude 4.0. Other interesting stars include the variables T Tauri (range 9.3–13.5), the prototype of a class of young stars that exhibit irregular variability, which is located in Hind's Variable Nebula (NGC 1555), a small reflection nebula which shows major variations in brightness and extent, and RV Tauri (range 9.8–13.3, period about 79 days, superimposed on a longer cycle of about 1224 days), the prototype of a class of highly luminous pulsating variables.

The Pleiades (M45), one of the most beautiful objects in the sky, is popularly called the 'Seven Sisters', as about that number of stars can readily be seen with the unaided eye, though there are about 100 stars in the cluster as a whole, embedded in a faint reflection nebula, the brightest part of which (NGC 1435) extends for about 20′ south of 23 Tauri (Merope), magnitude 4.1. Other interesting objects in Taurus include the Crab Nebula (M1, NGC 1952), the brightest (eighth magnitude) supernova remnant in the sky, at the center of which is the Crab pulsar, which marks the spot where a supernova was observed in 1054. It is a source of strong emission at both radio and x-ray wavelengths (where it is known as Taurus A and Taurus X-1, respectively).

The Taurid meteor shower appears to radiate from the constellation (as does the Beta Taurid shower, which cannot be observed visually as it occurs during June and July, when Taurus is close to the Sun).

See also: Aldebaran, Crab Nebula, Crab pulsar, T Tauri stars, Taurids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Taylor, Joseph H (1941–)

Radioastronomer, born in Philadelphia, PA, Nobel prizewinner for physics in 1993 with RUSSELL HULSE 'for the discovery of a new type of pulsar, a discovery that has opened up new possibilities for the study of gravitation'. Interested in radio as a boy, he became a professor at the University of Massachusetts, Amherst, and with his research student Hulse searched with the Arecibo radio telescope for pulsars. They discovered the *Hulse–Taylor pulsar*, PSR 1913 + 16, a binary pulsar that showed general relativistic effects, including loss of energy by gravitational radiation. Taylor helped found the Five College Radio Observatory, where the observations of the Hulse–Taylor pulsar were actually carried out, confirming EINSTEIN's theory of general relativity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tebbutt, John (1834–1916)

Australian astronomer and meteorologist, born at Windsor, NSW, discovered the Great Comet of 1861, through whose tail the Earth passed, to alarmist predictions of poisoning by cyanide when cyanogen was spectroscopically discovered in it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tectonics

The theory of tectonism—an umbrella term for processes that shape the surface features of a planetary body through forces generated in its crust by heating or cooling in its interior. Tectonic forces can be generated in bodies that have undergone differentiation and have developed a molten mantle beneath the crust. Global heating will cause the entire crust to expand, creating tensional forces in the crust; global cooling will cause it to contract, creating compressive forces. Lunar rilles are examples of features caused by extension of the crust as it expanded. On Mercury, extensive lobate scarps—long, cliff-like features—have been interpreted as resulting from compressive stress caused by global contraction. Wrinkle ridges, found on most of the terrestrial planets, are also compressive features. Localized heating in the mantle producing upwelling of material can cause local uplift of the crust above, as has occurred on Venus to create the mountainous region Maxwell Montes. The Earth's crust consists of segments known as *plates* which move slowly with respect to one another, driven by convection currents in the mantle; *plate tectonics*, as this process is known, was thought to be unique to the Earth, but it may also have operated on Venus. Tectonic processes have also shaped the surfaces of Europa and Ganymede, and of some of the satellites of Saturn and Uranus, including Enceladus and Ariel.

See also: Earth: interior.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tektite

A small rounded piece of natural glass, green, black or brown and typically centimeter-sized. Once believed to have a volcanic origin or to be peculiar meteorites, they are now thought to have been formed by the impact of kilometer-sized asteroids. Their composition, mainly silica with some metal oxide content, is consistent with the rapid melting of terrestrial rock under a large impact and the subsequent rapid cooling, from a temperature of 1500–2000 K, of molten droplets splattered into the atmosphere. Their various shapes support this view: many are disk-shaped or show flanges, suggesting surface reheating during their flight. The ages and geographical distribution also support the impact origin theory. They are found in four main locations, known as *strewnfields* (see table 1). The tektites found in a strewnfield have nothing in common geologically with their surroundings, and in three out of the four cases there is an identifiable impact structure. The absence of a confirmed impact crater for the Australasian strewnfield is problematic. There are also *microtektites*, microscopic glass spherules associated for example with the Chicxulub basin, and other types of impact glass found close to other craters.

Table 1. Tektites: the four strewnfields.

Strewnfield	Age (Myr)	Impact structure (approximate diameter)	Approximate number of tektites recovered
Australasian	0.8	?	700 000
North American	35	Chesapeake Bay, US (90 km)	2000
Ivory Coast	1	Bosumtwi Crater, Ghana (10.5 km)	2000
Central European	15	Reiskessel, Germany (24 km)	55 000

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescope

An instrument that is used to study distant objects. An optical telescope uses a convex lens (the objective, or object glass) or a concave mirror (the primary mirror), or in some cases a combination of both, to collect light from, and form an image of, a distant object. The image produced by the collector may be recorded photographically or electronically. When used for direct visual observation, a telescope also incorporates an eyepiece, a short-focus lens that enables the eye to view a magnified (enlarged) version of the image produced by the collector. An eyepiece may also be used to project an enlarged image onto a photographic emulsion or electronic detector.

If the collector is a lens, the telescope is called a refractor, whereas if the collector is a mirror, the telescope is a reflector. A catadioptric telescope uses a combination of lens(es) and mirror(s) to collect light and form an image.

A telescope is categorized by its aperture (the clear diameter of the collector) and focal length or focal ratio, the former being the distance between the collector and the image that it produces, and the latter being the ratio of the focal length to the aperture. Because a telescope collects more light than the unaided human eye, it can reveal fainter (and more distant) objects than the eye alone; the larger the aperture, the greater the light gathering power. The ability of a telescope to reveal fine detail (resolving power) also depends on the aperture. The size of the image produced by a telescope is directly proportional to the focal length of the collector. Additional lenses or mirrors may be used to increase the effective focal length of the telescope and thereby produce larger images.

The objective lens of a refractor usually consists of at least two components, each of which has front and rear surfaces. Consequently there are at least four surfaces that must be figured (ground to a precise shape). By contrast, because light is reflected from the front surface of an astronomical mirror, only one surface needs to be figured. Furthermore, lenses can be supported only around the perimeter, while mirrors can be supported across the back as well as round the edge. For these, and other, reasons, large reflectors are easier and cheaper to construct and, in most respects, are more efficient. All large modern optical telescopes are of the reflecting type, the largest being the two Keck telescopes, each with an aperture of 10 m, that are located on Mauna Kea, Hawaii.

The invention of the telescope is usually credited to the Dutch spectacle-maker, Hans Lippershey (*c*. 1570–1619), who constructed a simple refractor in 1608. However, some historians contend that a form of telescope may have been constructed and used by the Englishman, Leonard Digges, in or around 1550. Although the Scot, James Gregory (1638–75) had produced a design for a reflecting telescope in 1663, the first working reflector is generally believed to have been designed and constructed in 1668 by the English natural philosopher Isaac Newton (1642–1727).

The term 'telescope' is also applied to instruments that collect radiation, and enable images to be produced, at wavelengths ranging from x-rays to radio waves.

See also: Barlow lens, Cassegrain telescope, Dobsonian telescope, eyepiece, Galilean telescope, Gregorian telescope, light-gathering power, magnification, Maksutov telescope, mirror, Newtonian telescope, objective lens, radio telescope, reflector, refractor, resolving power, Schmidt telescope, Schmidt–Cassegrain telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescope Engineering

The design and construction of modern astronomical telescopes and instruments is one of the most challenging domains of engineering encompassing many fields including in particular optics, structures, mechanics, control systems, thermal and atmospheric physics. In this article we examine the main engineering approaches and practices in use in modern optical telescopes on the ground and in space.

Astronomical observatories as systems

Astronomical observatories are tightly integrated systems where most subsystems are strongly interacting and individual subsystem performances must be traded against one another in order to arrive at an optimum design. Since the image quality of a telescope is one of the most important performance characteristics (others include observing efficiency and spectral coverage), an essential step in this process is to establish an 'optical error budget' where all the contributing factors are evaluated. An example of such an error budget is shown in figure 1. As the design progresses, the individual allocations are re-estimated and traded against each other as need be until a satisfactory set is obtained. Cost and schedule have to be taken into account as well, and observatory performance has itself to be traded against scientific goals until the overall project optimum is determined.

The availability of high-fidelity computer modeling offers an ideal solution to this problem. It is now possible to develop a detailed computer model of the entire system (optics, structure, thermal, control systems etc), simulate operating conditions and explore the trade spaces to find the best configuration according to performance and cost criteria.

Optics

All large telescopes are of the REFLECTOR type because of the difficulty in producing lenses of sufficient purity in size larger than 1 m in diameter. Additional reasons are that lenses are generally heavy since they cannot be 'lightweighted' and they suffer from chromatic aberrations which can be corrected only over a narrow spectral range. Optical configurations are relatively simple and the Ritchey–Chrétien Cassegrain combination is almost universally used. In practice, as many optical surfaces are needed as the number of aberrations one wishes to correct. With two surfaces, the Ritchey–Chrétien configuration is able to correct spherical aberration and coma. Correction of astigmatism requires an additional surface which can be part of the telescope (three-mirror system) or incorporated into each instrument.

The choice of the f-ratio (ratio of final focal length to entrance aperture diameter) is the result of a complex tradeoff between imaging quality requirements, field of view, size of the secondary mirror, size of the instrument entrance optics, focal length of the spectrograph collimators and detector pixel size.

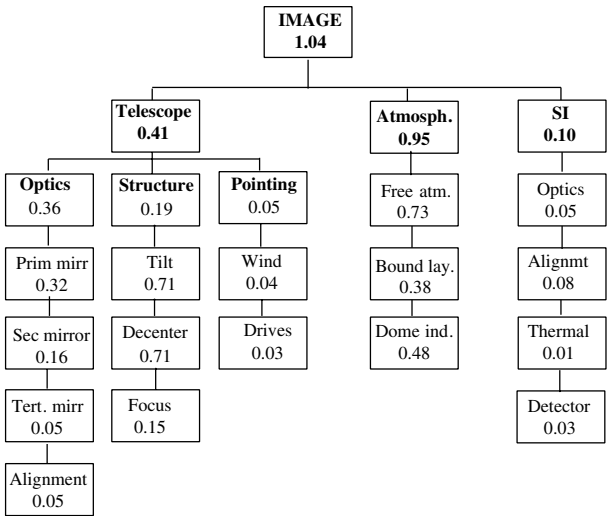


Figure 1. Optical error budget of the Keck 10 m telescope expressed in diameter of the image in arc seconds that encloses 80% of the total image energy.

For the simple case of a space telescope with diffraction-limited optics directly feeding a critically sampled camera (2 pixels per resolution element, i.e. λ/D), the f-ratio needed is given by:

$$\frac{f}{D} = \frac{2p}{\lambda}$$

where λ is the wavelength and p is the pixel size. Current detectors have pixel sizes ranging from 7 to 13 μm for CCDs and 18 to 28 μm for near-infrared detectors which results in f-ratios in the $f/15$ to $f/30$ range, which is well adapted to the Cassegrain combination.

For a ground telescope where atmospheric SEEING limits image quality to about 0.4", adequate sampling for imaging would require much faster f-ratios—in the $f/2$ to $f/4$ range for an 8 m class telescope. This can be obtained directly at a prime focus or with a focal reducer at a Cassegrain focus.

Spectrographs are generally used at the Cassegrain focus and fed with an f-ratio between $f/8$ and $f/15$.

The problem with Cassegrain foci is that the volume of the instruments is limited by mount clearance. Large and heavy instruments are better located at the 'Nasmyth' focus (figure 2). An additional advantage of the Nasmyth focus is that the instruments keep a fixed direction with respect to gravity which greatly simplifies their design.

Very high spectral resolution spectrographs are too bulky to be mounted on the telescope itself (even at the Nasmyth focus), and are best left Earth-stationary; they are fed via a series of flat mirrors at the so-called 'coudé focus'. In such cases a much slower f-ratio is used ($f/30$ to $f/100$) in order to reduce the size of the beam to be transferred.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

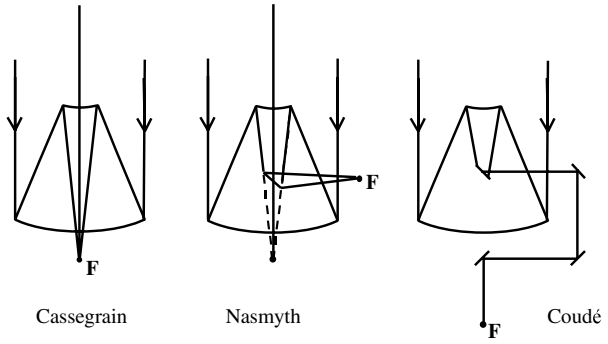


Figure 2. Cassegrain, Nasmyth and coudé focus arrangements

Materials

Contrary to many engineering situations, telescope structures are not driven by stress conditions, but by deflection under load (on the ground) or dynamic effects (on the ground and in space).

The governing equation for the deflection of a mirror (plate) under gravity is of the form:

$$\delta = k\frac{\rho}{E}\frac{D^4}{t^2}$$

where k is a constant the value of which depends on the geometry of the support, D is the diameter, t the thickness, ρ the material density and E is the Young modulus (elasticity modulus) of the material.

Similarly, the deflection under gravity of a beam supporting optical components of negligible mass compared with the mass of the beam (which is roughly the case of the tube supporting the secondary mirror in a Cassegrain telescope) is given by:

$$\delta = k\frac{\rho}{E}l^4$$

where k is function of the beam geometry and support and l is the length of the beam.

In both cases, the material properties enter as the ratio of the Young modulus to density (E/ρ); this ratio is referred to as 'specific stiffness'. The larger this ratio, the smaller the deflection will be. The same ratio also enters the laws governing dynamic effects due to wind or internal disturbances.

For large telescopes, whether on the ground or in space, weight is a dominant issue and the most advantageous materials will be those with high specific stiffness. As shown in table 1, the materials commonly used in making mirrors and telescope structures (glass, aluminum, steel) have about the same E/ρ. It is in particular seen that aluminum does not offer any advantage over steel. Steel is less expensive and is the material of choice for ground-based telescopes structures. Three materials stand out, however: silicon carbide, graphite epoxy and beryllium. For the same stiffness they offer a gain of four to six times in mass over steel and aluminum. They are expensive materials to produce, however, and are limited to space applications where mass is at a premium.

Table 1. Mechanical and thermal properties of major mirror and structural materials used in optical systems. ρ is the density, E is the Young's modulus, α the coefficient of thermal expansion, λ the thermal conductivity and C_p is the specific heat.

Material	ρ (kg m^{-3})	E/ρ (10^6 s)	α at 300 K (10^{-6} K^{-1})	$\lambda/\rho c_t$ (10^{-6} m^2 s^{-1})
Pyrex	2500	25	3.3	0.60
Fused silica	2200	33	0.5	0.86
ULE	2200	31	0.03	0.77
Zerodur	2500	36	0.10	0.77
SiC	2600	96	3.5	86
Beryllium	1800	162	12	5.4
Aluminum	2700	26	25	4.0
Steel	7800	27	12	5.7
Graphite epoxy	1740	61	0.09	29

Mirrors must also retain their correct surface shape in spite of environmental temperature changes which can be up to several degrees celsius. Changes in bulk temperature induce focus variations and gradients within the mirror will induce distortion of the optical surface. For these two effects to be negligible the mirror material should have a coefficient of thermal expansion of 10^{-6} or lower. Pyrex, first used on the 5 m Hale telescope, was a major improvement over regular glass, but the problem really only disappeared with the use of fused silica (HST mirror) and the development of ceramic glass (Cer-Vit, Zerodur) and ULE (ultra-low-expansion fused silica) which have coefficients of thermal expansion in the 10^{-6}–10^{-8} range and are used in most large ground-based telescopes today.

Although the mirror may retain its shape in a changing thermal environment due to a low coefficient of thermal expansion, the temperature difference between the mirror and the ambient air can generate convective turbulence ('mirror seeing') detrimental to image quality (see section on local seeing control below). This effect will be the smallest for materials which have a high thermal diffusivity ($\lambda/C_t\rho$) where λ is the conduction and C_t is the specific heat of the material, so that they track the ambient temperature variations well. None of the glass/ceramic materials are very good from that point of view and mirror seeing has to be mitigated by minimizing mass and by ventilation (see 'controlling man-made seeing' below).

In space, where mirror seeing is not a concern, high diffusivity can still be useful when the observatory is subjected to bulk temperature variation as when repointing but, in general, it will be more advantageous to avoid the thermal effects entirely by having as low a coefficient of expansion as possible. The best mirror blank materials for space application are thus those which have a high intrinsic stiffness and a low-expansion coefficient at operating temperature. For ambient temperature operation, low-expansion glass-type materials (ULE,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

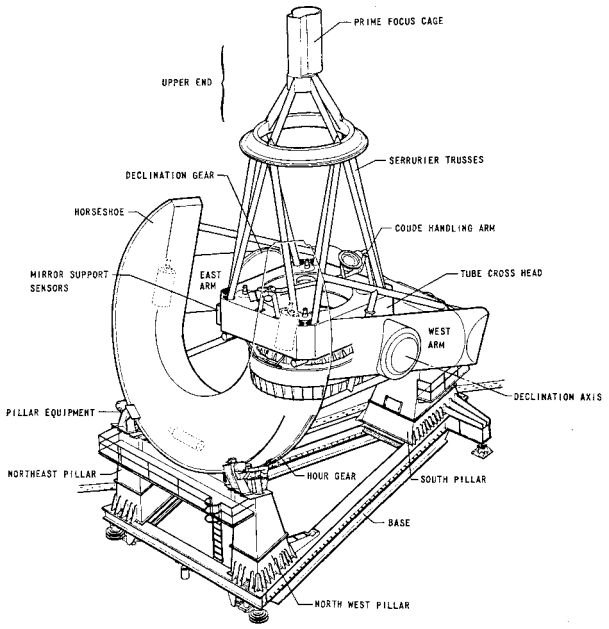


Figure 3. Main components of a telescope.

Zerodur) are best in spite of their lower intrinsic stiffness. For cryogenic operation (40 K), however, the best materials are beryllium and SiC because of their low coefficient of thermal expansion at cryogenic temperature (in the 10^{-7} K^{-1}) and high specific stiffness.

Finally, mirror materials must be able to be polished to a high degree of smoothness and be dimensionally stable. Amorphous materials like glass and ceramics produce very smooth surfaces while metals usually end up with a much coarser surface finish suitable only for infrared application (50 Å rms) unless they are coated with nickel prior to polishing.

Mechanical design

The main components of a ground-based telescope are shown in figure 3.

Because of the extraordinary precision to which the optical components must be fabricated (fraction of a wavelength or 1/10 to 1/100 of a micron), kept aligned (1 to 10 microns), and pointed (1/10 to 1/100 of an arc second), deflection under gravity, mechanical disturbances and temperature variations must be avoided or compensated for. In what follows we give an overview of the techniques which have been developed over the years to reach this goal.

Kinematic mount

To avoid deforming the very high-precision figure of the optical components, it is essential that strains in their supporting structure not be transmitted to the optics. The usual method consists of using a *kinematic mount*. A rigid body in space has six degrees of freedom: translation along and rotation about each of three orthogonal axes.

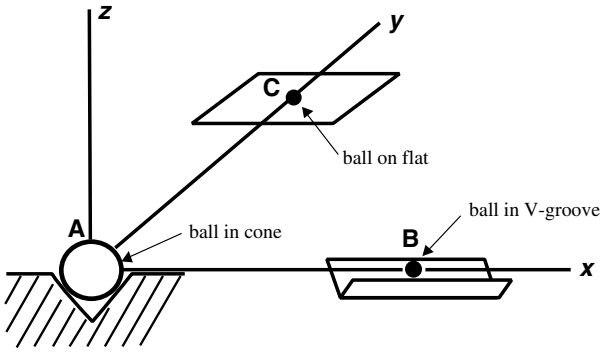


Figure 4. Principle of the kinematic mount. Alternate implementations, all following the same principle, are diverse.

The body is fully constrained when each of these possible movements is individually prevented. If one movement is constrained in more than one way, then the body will strain when its support structure deforms. A kinematic mount is a mounting system which does not constrain more than these six degrees of freedom. The simplest example is that of the traditional 'point/V-groove/plane' support (figure 4). Kinematic mounts are ubiquitous in telescopes. They are used not only to mount mirrors in their cell, but also to mount the instruments on the telescope tube and to support telescope mounts on their pier (see also TELESCOPE MOUNTINGS).

Mirror mounts

Up to about 30 cm in diameter, a ground-based telescope mirror can be made stiff enough to be considered a rigid body and mounted kinematically. Larger mirrors would bend beyond tolerable limits and it is necessary to have additional supports to compensate for gravity but without interfering with the stress-free condition afforded by the kinematic mount. This is done by 'floating' the mirror in a set of discrete forces or in pressure/vacuum so as to balance its weight (figure 5). This is accomplished by counterweight lever systems, regulated air pressure or force actuators with feedback control.

Space telescopes do not suffer from gravity deflection, but since gravity release is difficult to measure on the ground, force actuators are sometimes placed on the back of the mirror to make minute corrections to its figure once on orbit.

Telescope tube

In a ground telescope the primary and secondary mirrors of a Cassegrain system must remain parallel to each other and aligned in spite of the changing effect of gravity as the tube rotates to follow a target. The cylindrical telescope tube which is commonly used on small telescopes does not satisfy this condition as the two ends are tilted with respect to each other (figure 6). The solution to this problem is to use a truss consisting of members arranged in isosceles triangles on a square base. This forces the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

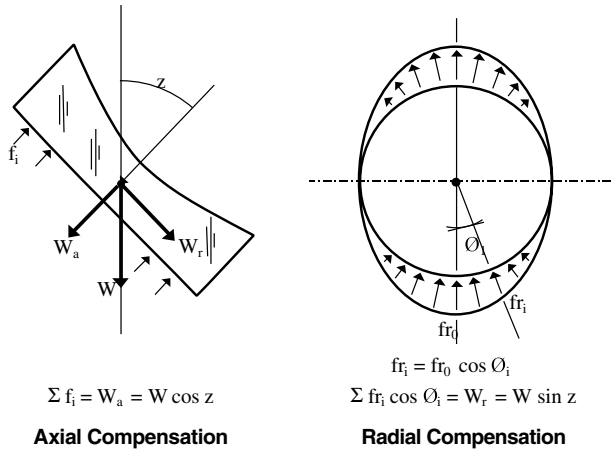


Figure 5. Forces are applied to the back and edge of the mirror to compensate for gravity and minimize deflection.

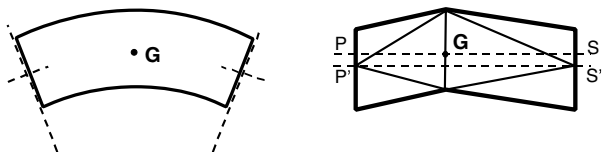


Figure 6. After deflection under gravity the two ends of a cylindrical tube do not remain parallel. In the 'Serrurier truss', the two ends, say the primary and secondary mirrors, stay parallel to each other and deflect by the same amount, maintaining collimation.

two ends (supporting the primary and secondary mirrors) to undergo parallel translations, and with proper sizing of the member sections the two translations can be made equal. This design, called the 'Serrurier truss' after Max Serrurier who invented it for the 5 m Hale telescope at Mount Palomar, has been used in most large telescopes ever since.

Balancing systems

Ground-based telescopes need to be extremely well balanced around their axis of rotation in order to minimize driving torques and differential flexure. This is done by adding or removing fixed weights or using motor-driven balancing weights. The procedure has to be done at each optical configuration change (e.g. exchange of the telescope top unit when changing from Cassegrain to coudé) and each time a focal instrument is replaced or installed.

Hydrostatic bearings

Although the tube and the mount of a ground-based telescope are extremely well balanced, the torques necessary to rotate the heavy tube and mount of a large telescope (of the order of 100 tonnes for the tube and up to 300 tonnes for the mount) must be small enough to avoid torsional deformation and potential optical misalignment. Large driving torques would also create heat dissipation

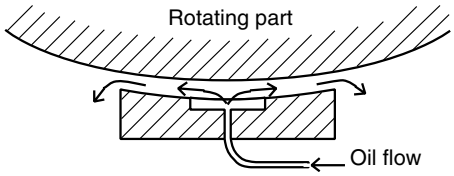


Figure 7. Principle of the hydrostatic bearing used to support the rotating parts of a telescope (tube and mount) with minimal friction.

in the dome, which is detrimental to image quality. During observations, the telescope tracking must also be extremely smooth. Mechanical bearings such as roller bearings suffer from 'stick-slip' at very low speed, a phenomenon due to stress build-up and relaxation at the contact areas, resulting in a jittery motion which is generally beyond the correctability of automatic guiding systems. These problems are solved by the use of hydrostatic bearings (figure 7) first developed for the 5 m Hale telescope and which are now almost a universal choice in modern large telescopes. The tube and mount journals are supported on pads which are accurately machined to match, and oil is pumped in between at high pressure in order to raise them and leave them floating on a thin film of oil (of the order of 0.1 mm thick). With hydrostatic bearings, the rotation torques are so low that they are negligible compared with the resisting torques due to the cable wrap-ups or slip rings used to transfer energy and data to the moving parts.

Active optics

For telescopes much larger than 4 m, the traditional measures taken to keep the figure of the mirrors undisturbed and the optical train aligned become increasingly difficult to implement. Instead of relying on passive means to compensate for gravity and thermal effects, it becomes advantageous to use an active system where the optical quality is monitored in permanence using a reference star, and the figure and respective positions of the main optics elements are adjusted accordingly. This approach is referred to as 'ACTIVE OPTICS'. All telescopes of the 8–10 m class use some form of active optics (figure 8). Because gravity and thermal effects vary slowly, the bandwidth of active optics is typically of the order of 1 Hz or less. This is to by opposition to 'adaptive optics' which is used for correction of atmospheric seeing which has a bandwidth up to 1000 Hz.

Pointing and tracking system

Early telescopes were all of the equatorial type which once pointed at the desired target only requires rotation at a constant rate (15" s^{-1}) around the polar axis to track the target. This hour angle motion was accomplished in an 'open loop' fashion by a clock mechanism and more recently by a synchronous motor. With the need for better accuracy and regulation, all modern telescopes use 'feedback control', where the actual motion of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

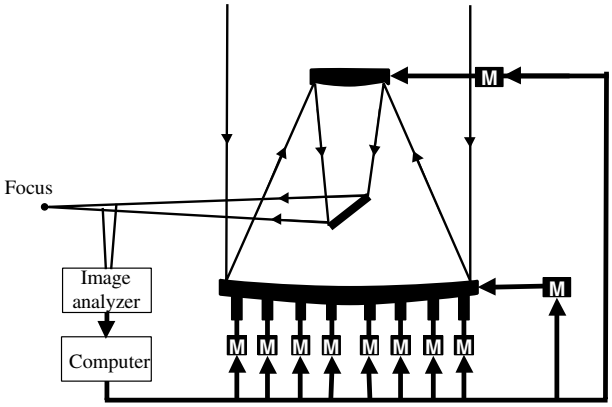


Figure 8. Primary mirror cell of the ESO Very Large Telescope. The 8 m mirror is supported on actuators which are adjusted continuously to compensate for gravity and thermal effects.

rotating axis is constantly compared with the desired value and the error fed back to the drive system to correct it. Feedback control is intrinsically more accurate and eliminates the need for high-precision gearing and drive systems. It also allows for the correction of atmospheric refraction, deformations of the telescope structure under gravity and of disturbances as long as they are within the bandwidth of the control system. With the assistance of computers, feedback control is also well adapted to producing the variable tracking speed needed for the ALTAZIMUTH MOUNTING (alt-az mount).

The limitation to the accuracy of pointing systems is due to disturbances beyond the bandwidth of the pointing system, which is generally a few Hertz at most. On the ground the main source is wind shake. With the need to largely open the dome during observations in order to reduce self-induced seeing (see section on man-made seeing below) the telescope is left largely unprotected from wind action, which results in wind buffeting on the primary mirror and on the tube supporting the secondary mirror. This effect can be lessened by reducing the cross section of the telescope structure members, by stiffening the secondary mirror support system and by actively controlling the primary mirror. In INFRARED TELESCOPES, chopping secondary mirrors used for background subtraction are another source of disturbance. This can be reduced to a large extent by compensation, that is to say by moving a dummy mass behind the mirror in the opposite direction so as to eliminate the chopping disturbing torque.

In space, the main source of uncorrected disturbance comes from the vibrations induced by thermal snaps or by moving parts such as reaction wheels, antennas, fuel slosh and tape recorders.

The absolute pointing accuracy of telescopes is typically about 1 arc second rms. Blind offsets from calibration stars are accurate to better than 0.1 and 0.01 arc second for ground and space telescopes respectively. Tracking stability of ground telescopes is typically better than 0.1 arc second without recourse to any guide star over periods of several minutes. By closing the loop on a guide star (see 'Guiding' below), the line of sight jitter is typically under 20 milliarc seconds rms for ground telescopes and a few milliarc seconds for the HUBBLE SPACE TELESCOPE.

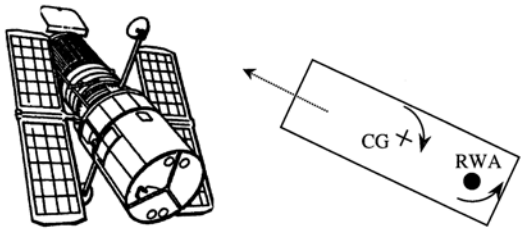


Figure 9. Reaction wheels are used to point a space telescope.

Drives

Until the 1970s ground telescopes generally used a worm gear system because of its high gear ratio and excellent intrinsic accuracy which permitted open loop tracking using a constant speed motor. Worm gears are expensive and limited in size, however, and large modern telescopes use spur gears, friction drives or even direct drives, all with feedback control from an attitude sensor. Spur gears offer more positive control (no risk of slippage), but friction roller drives are cheaper and inherently smoother. The ultimate choice is a direct drive which eliminates all mechanical systems. This is being used in the European Southern Observatory Very Large Telescope (VLT) where the rotor and stator of each axis motor are integrated in the structure.

Space telescopes cannot be driven by reacting against the ground; they use momentum transfer instead. The Hubble Space Telescope uses a set of flywheels, one for each axis plus a spare, rotating up to 3600 rpm, called 'reaction wheels'. To slew the telescope to a new target, the reaction wheels' spin speed is increased or decreased depending on the desired direction. Because the overall momentum remains constant, the telescope moves in the opposite direction (figure 9). To stop it when it approaches its target, the reaction wheel speeds are returned to their original values.

Sensors

On ground-based telescopes the position of each axis is measured by encoders which are now almost always of the optical type. Incremental encoders with a high angular resolution (e.g. 0.1 arc second), but no absolute accuracy are used for tracking. They are generally mounted on a gear or friction system for increased magnification. Absolute encoders, generally mounted directly on a large journal are used for absolute pointing determination and have an absolute accuracy of about 1 arc second.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

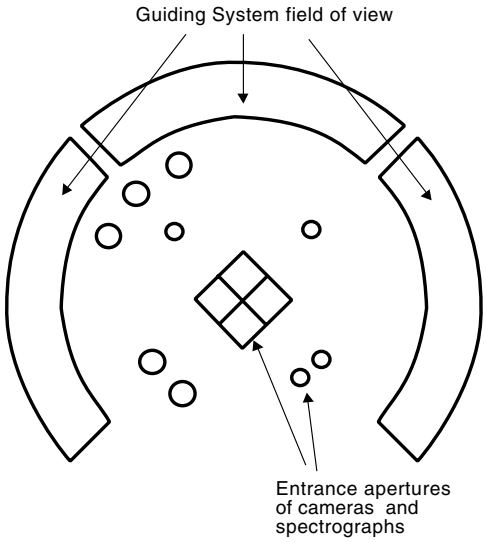


Figure 10. Focal plane of the Hubble Space Telescope showing the three zones used for guiding.

For space telescopes, the main attitude sensors are generally gyroscopes, the drift of which are corrected by star trackers. The star trackers have a field of several degrees and the best ones have an absolute accuracy of about 2 arc second rms.

Guiding

By itself, telescope tracking is not accurate enough for more than a few minutes at best. On the ground, the limit comes from errors in the gravity compensation systems, encoder imperfections, atmospheric refraction and atmospheric turbulence. In space, the limit is due to the spacecraft star-trackers sensing noise and from gyroscope drift. It is necessary to adjust the line of sight finely by monitoring the position of a relatively bright star in the vicinity of the field being observed. This is referred to as 'guiding'. Early telescopes used GUIDE TELESCOPES mounted on the side of the telescope tube. But it is very difficult to ensure that the line of sight of the guide telescope remains parallel to that of the main telescope whatever its inclination. It is much better to guide using a star in the very field of the telescope (figure 10). This eliminates line of sight differential errors and the guiding system benefits from the full aperture of the main telescope, thus improving its sensitivity.

The error signal supplied by the guider can be fed directly into the pointing system, but the bandwidth of the correction can be improved by 'fine steering' the optical beam itself instead of the entire telescope. This is accomplished by the tip/tilt of a mirror in the optical train such as the secondary mirror, or, better, by using a dedicated, low-inertia, flat mirror located at a real pupil just before the focal plane.

Dome and building

Ground-based optical telescopes are too complex and too precise to be left without protection against sunlight, heat, and adverse weather. Some kind of enclosure is necessary. In addition, OBSERVATORIES require a certain amount of supporting equipment and storage space for scientific instruments, all of which need to be located in the immediate vicinity of the telescope. These must also be protected from the elements. Contrary to RADIOTELESCOPES, optical telescopes are thus always installed inside an enclosure which is closed during the day and opened during observations.

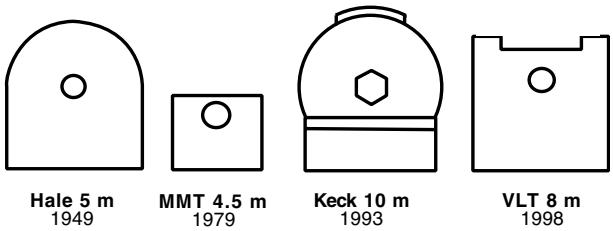


Figure 11. Evolution of dome and building size compared with telescope diameter.

The telescope enclosure can amount to 20 to 40% of the total cost of the observatory, however, and the recent trend has been to keep costs down by using primary mirrors with fast f-ratios and by switching from EQUATORIAL MOUNTINGS to alt-az mounts. Figure 11 shows how dramatic the reduction of enclosure size has been over the last 50 years.

While the enclosure protects the telescope and telescope chamber from heating by the Sun during the day, it can be a source of image degradation at night because any temperature difference between the enclosure and the ambient air will generate convection cells which may travel into the line of sight. The problem is due only to the mixing of air parcels at different temperatures from the ambient, not to pressure fluctuations due to aerodynamic effects, as the corresponding change in the refraction index of air is negligible.

In the best astronomical sites, the temperature drop at night is small (annual mean of 1.3 °C from sunset to sunrise at Mauna Kea, Hawaii), but there are frequent short-time-scale fluctuations that a large thermal capacity system simply cannot track (standard deviation of 0.5 °C and average time derivative of 0.6 °C h^{-1} in the case of Mauna Kea). A very short thermal time constant is essential. To minimize these effects the current philosophy, first implemented in the 1980s on the pathsetting MMT Telescope, is as follows:

- the telescope structure and optics should be designed for a very low thermal inertia (lightweighted or thin mirrors, 'open' structures);
- the dome should be wide open at night to maximize exposure of the telescope to ambient air and thus quickly reach thermal equilibrium.

A fully opened dome results in an airflow which is generally fully turbulent with a Reynolds number of $\sim 10^6$–10^7, but turbulence *per se* is not a source of degradation if the flow is isothermal.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

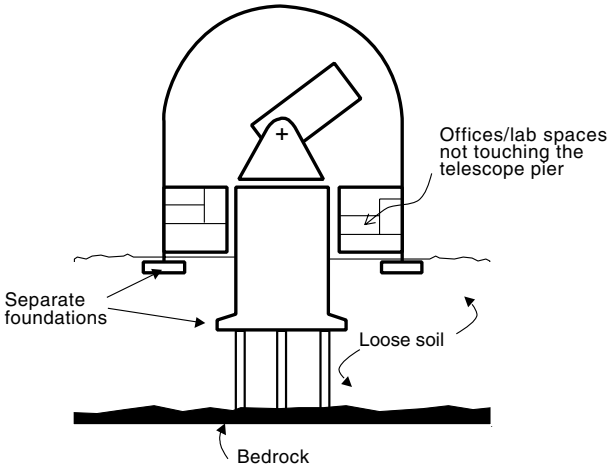


Figure 12. Separating the foundation of the building from that of the pier to minimize transmission of dome windshake.

In this new approach it is important, however, that the entire enclosure be well flushed. If there are pockets of air at different temperatures, then turbulence can mix these different regions and contribute to image degradation. The more open the enclosure the better, with open-air operation being the ultimate. The drawback is that in windy conditions, telescope shake will begin to become dominant, so that a compromise between thermal effects and dynamic effects has to be found.

The building and dome present a very large projected area normal to the direction of the wind. The associated overturning moment causes additional forces on the building foundations which can be expected to propagate through the soil and into the pier supporting the telescope. The effects are strongly dependent on the type of soil, on whether it is rock or particulate. In general transmission is reduced if (figure 12):

- the building foundation is structurally separate from the pier foundation;
- connections between the building and the pier and telescopes (cabling) are minimized and elastic;
- the building foundations are built over a damping layer (sand, lava cinder), while the pier is mounted on bedrock;
- the pier foundation is at a lower level than the building foundation.

Control of man-made seeing

Temperature inequalities between the ambient air and the mirrors, telescope structure and enclosure create convective turbulence along the path near the telescope which can easily degrade image quality as much as poor optics or a misaligned system. Designers of large telescopes since the 1960s have taken great care to reduce this effect in order to benefit fully from the intrinsic quality of the observatory sites. As explained above, enclosures are designed to have low thermal time constants and ventilation gates so that they can be flushed effectively either by wind or active ventilation so as to reduce 'dome seeing'. But there are several other possible sources of 'man-made seeing' which should be minimized or avoided.

Convective turbulence above the primary mirror can develop for temperature differences as little as a fraction of a degree, especially when the mirror is at a temperature higher than ambient (a 1 °C temperature difference will introduce up to 0.5 arc second of image degradation). This is a difficult problem to solve because of the enormous mass of the mirror which has a time constant of several hours. This 'mirror seeing' can be reduced by several measures: (a) minimizing the thermal time constant of the mirror so that it can follow changes in ambient temperature relatively rapidly, (b) cooling the dome during the day to a temperature close to that of the following night, (c) choosing sites where the temperature drop during the night is minimal. The thermal time constant can be reduced by decreasing the thermal mass of the mirror or by increasing the convective heat transfer between the mirror and the atmosphere. For example, a ribbed honeycomb structure can be incorporated in the back side of a mirror to both make it lighter and increase its surface area. Forced convection (fans) can be added to further reduce the thermal time constant.

It is also important to minimize the heat generated in the telescope chamber coming from electronic racks and oil circulating in the hydrostatic pads, as well as from people. This can be done by cooling, active ventilation, exhausting the warm air away from the building, and by placing the control rooms away from the telescope chamber.

Costs

Large astronomical telescopes represent major investments which require careful estimating and cost control during the design and construction phases. Until the early 1980s, the cost of equatorial telescopes followed a well established cost function of the aperture diameter to the 2.7th power (figure 13). Using the same technology, building 8–10 m class telescopes would have been unaffordable. But fortunately a series of technological improvements made it possible: mastering of the aspheric polishing process permitted the use of faster primaries (hence shorter telescopes), better testing methods and computer polishing took the optical figuring process from a black art to a deterministic science, the use of the stressed mirror or stressed-lap techniques reduced the polishing time drastically, the use of the alt-az mount instead of the equatorial mount reduced overall mass and thus cost, and the advent of computer control made active optics possible and allowed for more efficient pointing systems. This lowered the cost of telescopes by an order of magnitude, although the diameter relationship is still around the same power law. To enable the extremely large telescopes (25 to 100 m in diameter) which are now under consideration, similar

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

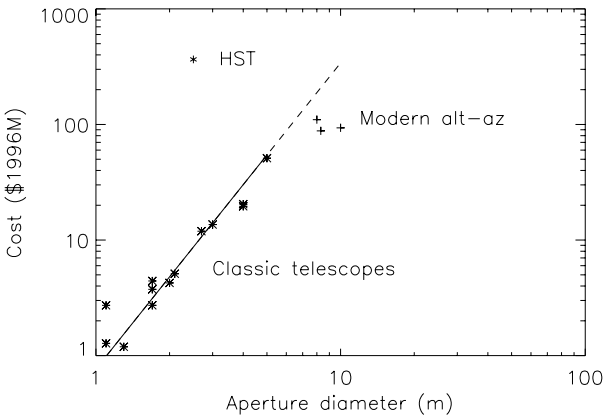


Figure 13. Cost models for ground-based and space-based astronomical telescopes (all costs expressed in 1996 US dollars).

technological innovations will be needed, perhaps by using spherical or fixed primary mirrors.

Bibliography

Bély P Y and Breckinridge J B (ed) 1991, 1993, 1995, 1996, 1998 Space telescopes and instruments I to V *Proc. SPIE* **1494, 1945, 2478, 2807, 3356** respectively

Burbidge G, Barr L D, Mack B and Stepp L M (ed) 1982, 1984, 1986, 1990, 1994, 1998 Advanced technology telescopes I to VII *Proc. SPIE* **332, 444, 628, 1236, 2199, 3352** respectively

Nelson J E, Mast T S and Faber S M (ed) 1985 The design of the Keck observatory and telescope *Keck Observatory Report 90*

Schroeder D J 1987 *Astronomical Optics* (New York: Academic)

Wertz J R (ed) 1978 *Spacecraft Attitude Determination and Control* (Dordrecht: Reidel)

Wilson R 1997, 1999 *Reflecting Telescopes* vols I, II (Berlin: Springer)

Yoder P R 1992 *Opto-mechanical Systems Design* (New York: Dekker)

Pierre-Yves Bély

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescope Mountings (for Amateur Astronomers)

Mounting a telescope is of paramount importance. Unless the instrument is absolutely steady, it will be quite useless, particularly as the field is bound to be comparatively small—and, of course, the higher the magnification, the more restricted the field.

The simplest form of mounting is the altazimuth. The telescope can move freely in any direction, either up and down in altitude, or east and west in *azimuth*. Some small telescopes are sold on what are termed pillar-and-claw mounts, which look very nice, but are about as steady as blancmanges. A proper tripod is much better, and if need be a camera tripod can be adapted, although it must be very firm (see ALTAZIMUTH MOUNTING).

There is an ever-present problem. As the Earth spins on its axis from west to east, the entire sky seems to rotate from east to west, so that a telescope has to be continuously shifted so as to keep the target object in view. With an altazimuth there are two separate motions to be taken into account, and hand-guiding can be awkward. Slow motions can be added, and this certainly helps, but it is fair to say that using an altazimuth telescope for photography, when a time exposure is needed, is very difficult—although with sufficient skill and patience it can be done.

Most telescopes of more than 3–4 in aperture—either REFRACTORS or REFLECTORS—are given EQUATORIAL MOUNTINGS. One common type is the German. Here there is a polar axis which is aimed at the pole of the sky, so that it is parallel to the axis of the Earth; the telescope is mounted on one end of the polar axis, and there is a counterweight on the other end. When the telescope is moved round in azimuth (right ascension) the up-or-down movement will look after itself. Manual slow motions can be fitted, but the ideal is to add a driving mechanism, usually electrical, so that the telescope is moved round at a rate which exactly compensates for the rotation of the Earth. Once the target object has been put into the field of view, it will stay there—provided that the polar axis is correctly aligned and that the driving clock is running at the correct rate. For most types of serious observational work a driving clock is more or less essential. Many older telescopes, usually refractors, were driven by a mechanism involving falling weights, and this does at least have the advantage of being unaffected by any changes in the supply of electricity.

There are many forms of equatorial mounts. With the Fork, the polar axis is mounted between pillars and the telescope is pivoted inside it. The English pattern has the polar axis fixed at both ends, and this makes for real stability, but there is the disadvantage that the telescope cannot be pointed near the celestial pole, so that there is a part of the sky which is inaccessible; the Mount Wilson 100 in reflector is mounted in this way.

With an equatorial mounting, setting circles can be added so that the telescope can be pointed to any particular object—provided that the right ascension and declination of the object are known, together with the sidereal time. By now there are many computer—controlled small telescopes which can be programmed to swing from one target to another. These are of particular use to observers who are searching for supernovae in external galaxies.

A coelostat is a form of optical instrument which makes use of two mirrors, one fixed and one movable (see SOLARS FEEDS). The light is collected initially by the movable mirror, and this rotates so that the light is reflected in a fixed direction onto the second mirror. This means that the eyepiece does not have to move at all, and heavy equipment can be set up in the position where an eyepiece would have been in former times, when visual observation was always used. The Coudé and Nasmyth optical systems make use of this principle.

The first good, large reflectors were made by William Herschel in the late 18th century; all these were on altazimuth mountings and were awkward to use. Equatorials were then developed, and before long were in common use, but mention should be made of the most remarkable mounting of all—that built by the third Earl of Rosse for his 72 in reflector, completed in 1845. The huge tube was mounted between two massive stone walls and was pivoted at the bottom, so that it could do no more than swing from side to side around the meridian—and it could cover only a relatively narrow strip of the sky; the observer had to wait until the Earth's rotation brought the target object into view. Nothing like the Rosse telescope has been built before or since, but within its limitation it worked well. It has now (2000) been restored and is once more operational, although it is of course of historical value only.

There is, however, one modern development which is of great importance. For many years all large telescopes were equatorially mounted, because otherwise they could not be mechanically driven. This is no longer the case. Computers are now so accurate that a telescope on an altazimuth mounting can be driven just as precisely as if it were on an equatorial, and the altazimuth pattern has many practical advantages. The first large telescope to be set up on an altazimuth was the Russian 236 in reflector at Mount Pastukhov in Russia, completed in 1975. To be honest, the telescope itself has never been a real success; the optics are below standard, and the observing site is mediocre, but the altazimuth mounting set the pattern for all large telescopes which followed, and equatorials have been more or less abandoned for all new professional instruments.

So far as amateurs are concerned, an altazimuth mounting is suitable only when no photography is to be undertaken, and the emphasis is on 'deep sky' observation rather than higher-magnification views of the Moon and planets. Actually, the most elementary form of all is the Dobsonian, in which the telescope is pivoted just above the mirror inside what is to all intents and purposes a box—the box itself being on a rotatable base. Large Dobsonians are becoming surprisingly common, even up to apertures of as much as 24 in, although the fact that they cannot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

easily be mechanically guided means that they are limited in scope (see DOBSONIAN TELESCOPE).

Making an equatorial mounting is by no means difficult, and an electrical drive can be added if needed. There is no doubt that observing becomes much more pleasurable when the telescope will keep the object firmly in view without having to be constantly adjusted by hand.

One final point is worth making; it is a principle laid down many years ago by W H Steavenson, a well-known amateur astronomer who because a world expert on telescope construction and even served as President of the Royal Astronomical Society. 'If you mean to construct a mount for a telescope', he wrote, 'work out the maximum possible weight that the mounting must be. Once you have decided this—multiply by three.'

Patrick Moore

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescopes in History

The precise origins of the optical telescope are hidden in the depths of time. In the thirteenth century Roger Bacon claimed to have devised a combination of lenses which enabled him to see distant objects as if they were near. Others who have an unsubstantiated claim to have invented the telescope in the sixteenth century include an Englishman, Leonard DIGGES, and an Italian, Giovanni Batista Porta, whose spyglass was manufactured in Holland for military purposes.

However, a Dutch spectacle-maker from Middleburg, Hans LIPPERHEY, is usually credited with making the first true refracting telescope in 1608. The new instrument was soon being used by the English mathematician Thomas HARRIOT and by Simon MARIUS in Germany to study the night sky. But it was GALILEO GALILEI, then Professor of Mathematics at the University of Padua in Italy, who was the first to construct a telescope for systematic astronomical observations.

The news of the Flemish invention reached Galileo early in 1609. After a few trials, 'sparing neither labor nor expense', he succeeded in arranging two lenses, one convex and one concave, inside a tube, so that they would magnify an object by a factor of three. This rapidly evolved into a more powerful instrument capable of magnifying objects 30 times (see GALILEAN TELESCOPE).

With his improved telescope, Galileo was able to turn his attention to the heavens. His startling discoveries were described in the book, *Sidereus Nuncius* (*The Sidereal Messenger*), which was published in March 1610. In it he wrote:

> (I have seen) stars in myriads, which have never been seen before, and which surpass the old, previously known, stars in number more than ten times. But that which will excite the greatest astonishment by far ... is this, namely, that I have discovered four planets, neither known nor observed by any one of the astronomers before my time.

This momentous discovery of Jupiter's four major satellites (now dubbed the 'Galilean satellites'; see JUPITER: SATELLITES) demonstrated that bodies other than the Earth might have other objects in orbit around them, and so supported the Copernican view of the solar system.

His telescope also enabled Galileo to produce the first maps of the Moon, revealing its rough terrain along the terminator as it passed through different phases (see MOON MAPS). Venus, too, was shown to have phases. Even the Sun was shown to be blemished as spots appeared and moved across its face (see SUNSPOTS). Evidence from his drawings suggests that Galileo even saw NEPTUNE in 1613, when it was close to Jupiter, but he failed to observe it for long enough to recognize that it was not a 'fixed' star.

Galileo's arrangement of lenses was soon superseded by Johannes KEPLER's system of two convex lenses, though Kepler himself does not seem to have actually built such an instrument. This seems to have been left to an enemy of Galileo, the German Jesuit observer, Christopher SCHEINER (1575–1650), who used his improved version to discover faculae on the Sun.

As the power of refracting telescopes increased, a series of fundamental discoveries were made. For example, by learning to grind lenses with exceptional accuracy and developing an improved eyepiece, Christiaan HUYGENS (1629–95) was able to construct instruments which enabled him to discover Saturn's ring system and its large moon, Titan.

However, refractors had two main disadvantages. The tendency for the object glass (the main lens) to focus different wavelengths of light at different distances caused them to produce rings of false color around stars and planets (CHROMATIC ABERRATION). This effect could be reduced by building telescopes of extremely long focal length, but this, in turn, made them cumbersome and unwieldy.

The largest classical refractor, which was used by Johannes HEVELIUS (1611–87), was 47 m (155 ft) long, and had to be set in place with a hoist. Giovanni Domenico CASSINI (1625–1712) later used a huge refracting telescope at the newly built Paris Observatory—described by Molière as 'an eye-glass to frighten people out of their wits'—to discover the gap in the rings which is named after him. His tubeless instrument comprised a lens which was placed on the balcony of the observatory and pointed more or less in the intended direction, while the observer stood 27 m (89 ft) below in the courtyard, trying to direct the image through a magnifying glass.

In 1671, the same year that G D Cassini moved to the Paris Observatory, Isaac NEWTON presented a revolutionary new telescope to the Royal Society in London. Believing that it would never be possible to make achromatic lenses for refracting telescopes, he designed and built the first practical reflecting telescope. Although its dimensions were modest—the metal mirror was only 1 in (2.5 cm) in diameter, and it was just over 6 in (15 cm) in length—Newton remarked that it would 'discover as much as any three or four foot tube' and he forecast that similar reflectors 6 feet (1.8 m) in length would 'perform as much as any sixty or hundred foot [18–30 m] tube made after the common way'.

Although Newton was the first to build a reflecting telescope, the principle of the reflector had originally been suggested by James GREGORY in 1663. However, his design proved extremely difficult to build. Whereas Gregory's instrument relied on a parabolic primary mirror with a central hole and an elliptical secondary mirror, Newton's simply comprised a parabolic primary which reflected light onto a flat, tilted, secondary mirror (see GREGORIAN TELESCOPE).

A variation on the Gregorian design soon arrived with the introduction of the CASSEGRAIN TELESCOPE in 1672. This version used a convex mirror inside the focal length to reflect the light back through a central hole in the main

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mirror. Today, it is the most common design for large reflecting telescopes. However, it was not until the mid eighteenth century that either the Gregorian or Cassegrain system was fully developed and became widespread. Cassegrain himself was never identified.

Despite these advances, large reflectors remained much more difficult to make than refractors, and half a century elapsed before they were able to compete on equal terms. No important discoveries were made with reflecting telescopes until the time of William HERSCHEL (1738–1822).

Herschel was renowned as the best telescope maker of his time. At his home in Bath, which he turned into a telescope workshop, Herschel experimented with different metal alloys in an attempt to find one with a high reflectivity, good resistance to tarnishing and reasonable ability to hold its shape during daily temperature changes. The best speculum metal alloy for his mirrors was found to be 79% copper and 21% tin. In 1778, he used this mixture for a 6.2 in (16 cm) reflecting telescope with a focal length of 7 ft (2.14 m), which he used for his second and third sky surveys. In 1782, a year after he used this instrument to discover URANUS, he demonstrated its superiority to the skeptical Astronomer Royal, Nevil MASKELYNE, at the Royal Observatory in Greenwich.

Upon receiving patronage from King George III, Herschel began to construct a giant 40 foot long (12 m) instrument with a mirror 48 in (1.2 m) in diameter. Completed in 1789, it remained the largest telescope in the world for more than half a century. On his second night of viewing with this monster, Herschel discovered the sixth moon of Saturn, ENCELADUS. Three weeks later, he discovered a seventh moon, MIMAS. Yet he was generally disappointed with his gigantic creation. It was clumsy to operate, while the mirror tarnished rapidly in the damp climate and distorted under its own weight.

Most of Herschel's discoveries were made with the 18.8 in (48 cm) reflector that he made in 1783. This instrument made a major contribution to the advancement of astronomy when his son, John, installed it at the Cape of Good Hope in 1834–8, and was able to survey the LARGE and SMALL MAGELLANIC CLOUDS, the center of our Galaxy, and the wonders of the southern skies.

The early nineteenth century saw a number of improvements in telescope optics. One of the leading telescope makers was the Director of the Physical and Optical Institute of Munich, Joseph von FRAUNHOFER (1787–1826), the inventor of the equatorial mounting. The most famous of his instruments was the Great Dorpat Refractor which F G W STRUVE set up at the Dorpat Observatory in Estonia in 1824. This 9.5 in (24 cm) refractor was used by Struve to discover more than 2 000 double stars.

Fraunhofer's invention of the diffraction grating and the later introduction of photography (first used in 1840) revolutionized optical astronomy and led to many discoveries. However, not until 1845 was Herschel's 48 inch telescope overtaken in size. Lord ROSSE's 6 ft (183 cm) aperture reflecting telescope contained the largest speculum metal mirror ever made, and had a light grasp more than double that of Herschel's instrument. Built at Parsonstown (now Birr) in Ireland, this instrument enabled Rosse to discover the spiral structure of certain nebulae (now recognized as galaxies), and to resolve others into star clusters.

Although Jean FOUCAULT (1819–68) introduced silver coating for glass mirrors, major astronomical instruments were being equipped with speculum metal mirrors as late as 1862. The last of the line was the Great Melbourne Telescope, whose 48 in (1.2 m) metal mirror required repolishing after a few years' use, and which was generally regarded as a failure.

Whilst reflectors were becoming the most widespread astronomical instruments, there was still considerable interest in pushing refractors to the limit. The culmination of this was the construction of the largest refracting telescopes ever built. The 36 in (0.9 m) instrument was installed at LICK OBSERVATORY in 1888, and was followed by the 40 in (1 m) telescope at YERKES OBSERVATORY in 1897.

The driving force behind the Yerkes project, George Ellery HALE (1868–1938), went on to promote the construction of two gigantic reflecting telescopes. The 100 in (2.5 m) Hooker Telescope on MOUNT WILSON, which was the largest in the world from 1918 until 1948, was the first instrument capable of seeing objects that formed early in the history of the Universe.

The 200 in (5 m) Hale Telescope on Mount PALOMAR, completed in 1948, was used for photographic studies which opened the way to further discoveries. Its giant primary mirror was a ribbed casting of Pyrex glass, while its parabolic surface was coated with a thin film of highly reflective aluminum. In 1976, an even larger 236 in (6 m) telescope was completed on Mount Semirodriki in the Caucasus region of the Soviet Union.

For a while, their great weight and the difficulties in casting such monolithic mirrors seemed to present an insurmountable barrier to the development of still larger telescopes. The primary mirror of the Hale Telescope weighs 13 tonnes, and was once considered the practical limit for telescope size.

While the giant reflectors were being used to study ever smaller fields of view, Bernhard SCHMIDT (1879–1935) was inventing a telescope which could take wide angle photographs. The first Schmidt camera, built in 1930, used a thin glass corrector plate to overcome spherical aberration. It soon spread around the world's major observatories and was incorporated in various hybrid designs such as the SCHMIDT–CASSEGRAIN.

The next major advance was the replacement of mechanical control systems by computers. In contrast to the Hale Telescope, which used huge motor-driven worm wheels controlled by a quartz crystal clock, the ANGLO-AUSTRALIAN Telescope, completed in 1975, set a new standard in computer control and pointing accuracy.

Computer control was also the key to the introduction of the multiple mirrors and segmented mirrors that broke the telescope size barrier. Leading the way was the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

MULTIPLE MIRROR TELESCOPE (MMT), which was built on Mount Hopkins in Arizona in 1977. By using six 72 in (1.8 m) mirrors in a circular array on an alt-azimuth mounting, the telescope's effective light-collecting area was equivalent to a single 172 in (4.5 m) instrument. One of these mirrors was used in the discovery of the first gravitationally lensed quasar.

The revolutionary MMT system was eventually replaced by a modern lightweight mirror with a honeycomb construction in 1998. Mirrors of this type consist of glass ribs between a thin, but rigid, concave mirror and a flat back plate.

In recent years, adaptive and active optics have been introduced in an attempt to compensate for image degradation caused by the atmosphere. Meniscus mirrors, that are too thin to support their own weight, and segmented mirrors can now be adjusted by computer-controlled servomechanical actuators so that they maintain their optimal shape and position at all times. Small adaptive mirrors may also be inserted in the light path of the telescope to correct atmospheric distortions. The EUROPEAN SOUTHERN OBSERVATORY's 138 in (3.5 m) New Technology Telescope was one of the pioneers in the development of such active and adaptive optics.

The largest mirrors in the world today, installed in the twin Keck telescopes on Mauna Kea, Hawaii, are made from 36 hexagonal 72 in (1.8 m) mirrors which fit together like bathroom tiles (see UNIVERSITY OF HAWAII INSTITUTE FOR ASTRONOMY). These segments effectively create a single mirror, 400 in (10 m) in diameter, with four times the collecting power of the Hale Telescope.

Another new technique that will eventually operate on the Keck telescopes is optical interferometry. Light from nearby, but separate, telescopes is combined so that the resulting image is equal to that received from a single giant telescope.

The European Southern Observatory started construction of the world's largest optical interferometer in 1996. The Very Large Telescope (VLT) in northern Chile will eventually combine the light from four 315 in (8 m) telescopes, producing a performance equivalent to that of a 630 in (16 m) instrument.

Meanwhile, advances in technology have also enabled scientists from the University of Arizona, Ohio State University and German and Italian astronomical research institutions to cast the largest single-piece mirror ever made. The LARGE BINOCULAR TELESCOPE at Mount Graham Observatory in Arizona will eventually be equipped with two 331 in (8.4 m) primary mirrors which have a thickness of only 1.1 in (28 mm) in the center and 35 in (894 mm) at the edge. These will provide an image comparable to that of a single 75 ft (23 m) telescope when construction is complete in 2002.

Such advances have, at least partially, nullified the advantages of placing telescopes in space. However, despite its relatively small 95 in (2.4 m) mirror, the HUBBLE SPACE TELESCOPE (launched in 1990) continues to return some of the highest resolution optical images of faint galaxies located billions of light years away. An 8 m class replacement with a folding mirror, known as the Next Generation Space Telescope, is in the pipeline for a 2008 launch(see SPACE INSTRUMENTATION: NEXT GENERATION SPACE TELESCOPE). Multiple space telescopes using optical interferometry, which will be capable of imaging Earth-size planets around nearby stars, are envisaged for the following decade.

Peter Bond

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescopes of the Future

New telescopes, new detectors and new regions of the electromagnetic spectrum have often revealed totally unsuspected aspects of the universe. Consider the careful spectroscopic observations of nebulae started around 1912 by Vesto Slipher, using the modest 24 in (61 cm) telescope of Lowell observatory. The recession velocities of thousands of kilometres per second that he found were completely unexpected and, only in 1917 did Willem de Sitter begin to understand them as galaxies in an expanding universe. Dark matter, a still unidentified form that outweighs all normal matter in the universe and controls the condensation of matter after the big bang, was discovered by Fritz Zwicky and Vera Rubin during spectroscopic observations to weigh galaxies. The rapid motions reflected a gravitational pull too large to be accounted for by ordinary matter in stars and interstellar material. The microwave background radiation, the cooled light of the big bang itself, was completely unforeseen by its Nobel-prize-winning discoverers, Arno Penzias and Robert Wilson (although not by others).

What common threads may be found in these and the many other unexpected discoveries in astronomy? Very often the observers were pushing new equipment to its limits to accomplish a highly focused goal unrelated to the actual discovery. Slipher was hoping that the diffuse nebulae might turn out to be nearby planetary systems in formation. Zwicky and Rubin were using the largest telescopes to see fainter objects than ever before. Penzias and Wilson were testing a highly sensitive new type of radio telescope for communications.

More generally, unforeseen discoveries are made when the range of observations is enlarged. This happens whenever new parts of the electromagnetic spectrum are opened for observation. New vistas are also opened when larger telescopes or better detectors allow us to reach fainter objects or bigger samples or to study brighter ones in more detail or with better time resolution.

A major new development in astronomy that is likely to drive new instruments in the coming decades is the exploration of exoplanets, the planets of other stars. The existence of other living worlds like our own has been the subject of speculation for centuries. However, telescopes have not been powerful enough to discern extrasolar planets. In the solar system, the Earth already occupies the prime location and our neighbours appear too hot or too cold. Now at last, with clear evidence that giant planets exist around other stars, we have the technical capability and the incentive to build radical new telescopes to study them. With them we should be able to find planets even as small as Earth and to search their spectra for the biochemical signs of life. The potential for unforeseen discovery is enormous. The first discoveries have already revealed completely unpredicted phenomena: planets of Earth's mass orbiting a pulsar and planets with the mass of Jupiter orbiting closer to their stars than does Mercury to the Sun, a place far too hot for them to form.

Ground telescopes come of age

During the 40 yr after the Palomar 200 in (5 m) telescope came into operation, the power of optical telescopes increased enormously, even though no significantly bigger instrument was built. The gains were made by improving detector sensitivity 100-fold to reach the fundamental limit set by photon noise, by extending detector sensitivity into the infrared and by multiplexing so that dozens or hundreds of objects could be analysed at once. However, as these advances have reached their limits, the past decade has seen the construction of a new generation of much larger telescopes. Starting with the two 10 m Keck telescopes, there are now around a dozen of size 6.5–10 m in operation or under construction. With such an increase in collecting area, some profound unforeseen discoveries can be expected. This will be especially true when these telescopes are able to remove atmospheric blurring and are linked together as interferometers. Until now, astronomers have had to choose between small-aperture telescopes in space that are free from blurring (the Hubble Space Telescope or HST) or large-aperture telescopes on the ground with blurring. Adaptive optics is a new technique that will give both advantages at once. In fact, because the natural limit to resolution set by diffraction improves in proportion to aperture, the bigger ground telescopes will be several times sharper than HST.

The key element in adaptive optics is a mirror whose shape can be altered rapidly in response to the measured atmospheric distortion. By giving the mirror equal but opposite distortion, the original image sharpness is restored. It has been difficult to make the measurement and correction fast enough to keep up with the constantly changing turbulence, but today's detectors and computers make it possible. Correction of bright objects has already been accomplished and in a few years we should start to see sharp images of even the faintest objects corrected at infrared wavelengths. When a target itself is too faint to allow fast measurement of atmospheric distortion, an artificial star created by a laser searchlight may be used as a surrogate. Experimental laser systems are in operation, although the combination of exquisite tuning and high power needed to excite scattering very high in the atmosphere is proving elusive.

Once adaptive optics is in place, another technique to increase the scope for discovery becomes possible. This is interferometry, long used by radio astronomers, which relies on combining the waves from separate telescopes. By measuring the strengthening and weakening of intensity as the crests and troughs reinforce or cancel out, images with greatly increased angular resolution are obtained. Interferometry can be extended to optical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

wavelengths for very high-resolution imaging, but sensitivity is poor unless all the waves from each dish add together coherently. In the presence of atmospheric distortion, coherent addition is possible only when very small apertures are used. However, once the full apertures of the large telescopes are corrected with adaptive optics and used as interferometer elements, very faint objects will become accessible.

All of the largest telescopes have been built as multiple units in anticipation of interferometry: Europe's Very Large Telescope (VLT) consists of four telescopes, while the Keck and Large Binocular Telescope (LBT) each have two (figure 1). Because the LBT mirrors are on a common mounting and closely spaced, as an interferometer it will act like a 22 m telescope with a resolution of 10 marcsec and high infrared sensitivity to complex, faint sources over a wide field of view. It will be able to find and analyse the thermal spectrum emitted by self-luminous giant planets, younger versions of our own Jupiter. The VLT and Keck telescopes are separately mounted and widely spaced (up to 100 m) for resolution as high as 2 marcsec over a field of 1 arcsec, but at the price of lower sensitivity and some ambiguity in image structure. They should be able to resolve those Jupiter-sized exoplanets mentioned above that are so close to their stars that they are red hot.

Now that the construction phase of the current generation of big telescopes is winding down, their designers are looking at ways to build bigger optical telescopes. In California, a single telescope of 30 m diameter (the California Extremely Large Telescope) is being studied; in Europe, one of 100 m (the OverWhelmingly Large telescope). Arizona's focus is on an imaging interferometer with two 20 m moving telescopes, the 20/20 telescope. With adaptive optics and interferometry, their sensitivity and resolution can potentially far outstrip what we have now. The tasks foreseen for such telescopes include observations of the very early universe at much higher resolution and sensitivity than possible with HST and the detailed spectroscopic study of Jupiter-like exoplanets. If the closest stars have Earth-like planets, they should be visible in long exposures with the 20/20 telescope. In addition, the prospect of unforeseen discovery is a powerful motivator.

New discoveries are likely not only from such larger telescopes and higher-resolution images but also from automated analysis of deep-sky images. In the past, photographic plates from 1.2 m telescopes covering 6º of the sky on a side have provided a rich source of data. Electronic detectors are now greatly increasing the accuracy, sensitivity and spectral range of these measurements. Two digital all-sky surveys with 1.2 and 2.5 m telescopes are in progress now, with charge-coupled device detectors for optical wavelengths in a mosaic of area 0.1 m^2 and near-infrared detectors of

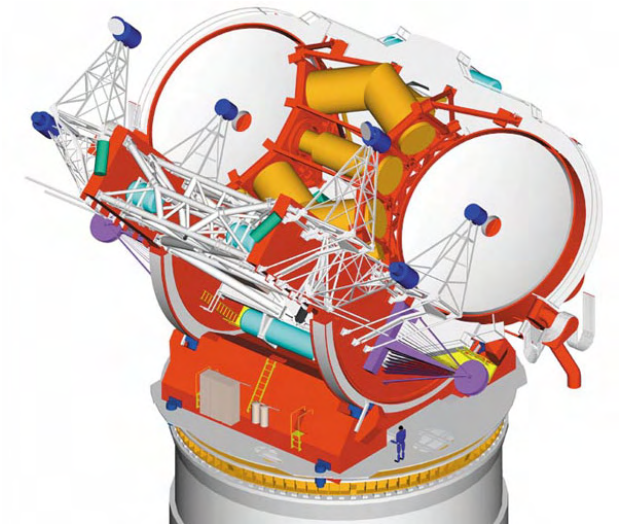


Figure 1. The LBT, with two 8.4 m mirrors side by side. (Drawing by European Industrial Engineering, Mestre, Italy.)

0.001 m^2. The coming decade should see the construction of the 8 m Large Synoptic Survey Telescope with 3º field (0.5 m m^2 detector illuminated at *f*/1.25). The data rate will be prodigious, with the night sky repeatedly mapped every four clear nights, creating 10 terabytes of data per map. Two key goals will be finding and measuring the orbits of asteroids as small as 300 m that could hit the Earth and mapping the spatial distribution of dark matter on large scales. Dark matter is detected through the small but systematic distortion it induces in the shapes of background galaxies. The data reduction for these tasks is big, but well defined. The challenge for computer gurus will be to spot the unforeseen—for example, new types of erratic variable objects or subtle large-scale correlations in data.

The limitless potential of space

Despite the evolution of much more powerful ground telescopes, space still holds unique and fundamental advantages. Complete freedom from atmospheric distortion in space will remain a major asset. Even when laser guide stars are used for adaptive correction on the ground, a nearby background star is still needed to sense overall jitter of the sharpened image. Stars are dense enough for infrared imaging over most of the sky, but optical imaging, which demands more accurate and faster stabilization because of the shorter wavelengths, will be restricted to fields near a bright star. Tomographic, three-dimensional mapping of atmospheric turbulence with multiple laser guide stars may be able to relieve this optical restriction and open wide-field correction. Although possible in principle, this will be very challenging.

The second unique capability of telescopes in space is for observations in the blocked ultraviolet and infrared spectral regions. Even where it transmits in the infrared,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the atmosphere is bright from heat and molecular emission. Additional background heat comes from the ground telescope optics themselves, which cannot be cooled because that would cause them to frost. Space telescopes can be made extremely cold, for a million times reduction in sky background.

The current generation of space telescopes has already shown how these advantages can be exploited. HST records wide-field optical images typically 10 times sharper than for uncorrected ground-based telescopes and reaches far into the vacuum ultraviolet. Its capability has been steadily improved by astronauts. For example, a new instrument was added to extend imaging and spectroscopic capability to the near infrared, where the natural sky background from sunlit interplanetary dust is much darker than on Earth. HST is likely to remain the only large optical telescope in space for astronomy for at least the next decade. Its dominance will be weakened by a strong challenge from ground telescopes, but perhaps that will allow time for more speculative use that could lead to unforeseen discovery.

HST's mirrors are too warm to be useful further into the infrared, where on the ground the atmosphere completely blocks transmission. Here three other space telescopes, the Infrared Astronomical Satellite, the Cosmic Background Explorer and the Infrared Space Observatory, all cooled with liquid helium to within a few degrees of absolute zero, have revealed the very cold universe. New discoveries can be expected from the Space Infrared Telescope Facility (SIRTF), a new helium-cooled telescope with 0.85 m aperture and more advanced detectors, which is scheduled for launch in 2002. It will reach to 160 μm wavelength.

Beyond these telescopes, the potential for future space astronomy is limitless. In principle, arbitrarily high sensitivity and resolution can be achieved at any desired wavelength by making the telescope large, accurate and cold enough. Space is an ideal environment, because not only is atmospheric degradation removed but so also are gravity and wind. This is a huge advantage because, under such forces, big objects bend more than small ones, whereas, no matter how big a telescope is made, the mirror surface accuracy cannot be relaxed if the waves are to be reflected to a sharp focus.

What might a future telescope look like, built for operation in orbits far from the Earth where distorting forces are minimal? The actual working part of a telescope mirror is the top layer of the reflecting surface, which is only a few hundred aluminum or silver atoms deep. Thus there is the possibility in space of using extremely thin membrane mirrors covering hundreds or thousands of square metres, if a way can be found to hold the shape. One way is to stretch a membrane flat from the edge and then to curve it very slightly by electrostatic force. An experimental telescope made in this way from a membrane 600 nm thick (a few thousand atoms) was used to record the image shown in figure 2. While very small, this mirror paves the way for larger-membrane telescopes, with tests planned first on the ground and then in space.

For such gossamer telescopes in space, the main disturbance comes from sunlight, which both warms spacecraft and pushes them. Light pressure was strong enough to disturb noticeably the orbit of the 100 ft (30.5 m) Echo balloon satellite launched in 1960 (only 4 yr after Sputnik) and would blow a 600 nm thick membrane right out of the solar system. Gossamer telescopes thus would have to be shielded from sunlight, but the shields must be lightweight too, and held off the telescope somehow, for solar pressure would push a freely orbiting shield into collision with the shielded mirror. Future gossamer telescopes may look more like sailing ships than the telescopes we now know.

Suppose at some time we find a planet that shows clear chemical evidence of life. Could we image it? The largest space telescopes we can conceive do not use mirrors at all and are extremely heavy, focusing light by gravitational bending. Figure 3(*a*) shows how multiple images of the same very distant galaxy are formed and distorted into magnified arcs by an intervening blob of dark matter, traced out by the brighter galaxies held in its gravitational field. The distortion was removed to yield the image shown in figure 3(*b*). Our deepest views of the distant universe will probably rely on finding chance alignments like this to aid our biggest ground and space telescopes. There is no chance that a nearby black hole will lie conveniently in front of and magnify an

Figure 2. The moon imaged with a membrane telescope. The *f*/500 primary mirror, curved by electrostatic force, had a thickness of 600 nm and a diameter of 6 mm. (Reproduced from Angel *et al* 2000.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

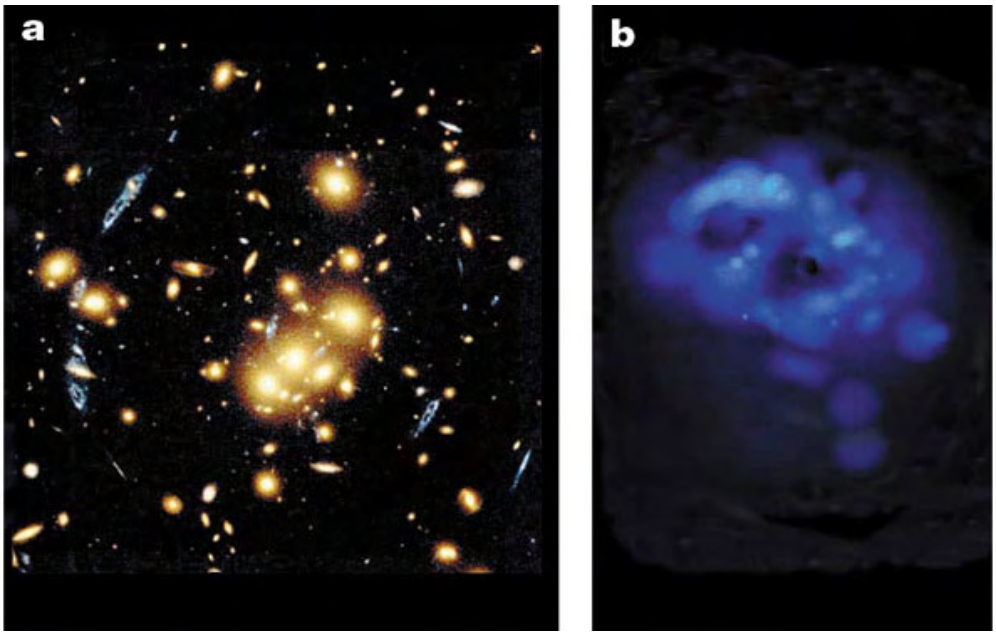


Figure 3. Imaging distant galaxies. (*a*) HST image showing multiple blue images of a distant background galaxy magnified by gravitational lensing. (From HST archive.) (*b*) Rectified image of the galaxy. (Reproduced from Tyson *et al* 1997.)

interesting planet, but we could use the Sun as a gravitational lens by sending a detector out of the solar system in the opposite direction to the planet. Once it is far enough away, more than 600 times the radius of the Earth's orbit, light from the planet bent around the Sun will be brought to a focus, forming a large image. Taking a detector to such distance may sound difficult, but it is still far easier than interstellar travel to even the nearest star, 1000 times further away.

The next space telescopes

The astronomy community has identified two key scientific goals to shape the development of more powerful space telescopes. One is to probe deeply the highly redshifted early universe, of which we have caught the first glimpse with HST. It is best observed in the virtually unexplored 2–5 μm region of the electromagnetic spectrum where the natural sky background is darkest. The other goal is the exploration of other planetary systems, with particular emphasis on looking for Earth-like planets, which will be restricted to only the very nearest stars from the ground. This is best attempted from space by detection of the infrared heat emitted by planets (10–20 μm) rather than the starlight they reflect. Their thermal emission is much brighter relative to the star, and the spectroscopic signatures of key diagnostic molecules, water and ozone, could be very strong, as they are in Earth's spectrum. Ozone would be of enormous interest, as in our atmosphere it is of biological origin.

Remarkably enough, both these major scientific objectives involve similar technical advances in the previously neglected shorter wavelengths of the thermal infrared spectrum. Both need mirror temperatures in the range 30–100 K, which should be reachable without expendable refrigerants by carefully shielding against solar heating and by lofting so far from the warm Earth that it appears no bigger than the Moon. Mirrors even larger than HST's 2.4 m are desirable in both cases, once the technology can be developed to make them. The extreme lightweighting of membrane structure is not yet needed, but the passive, rigid mirrors that are the current standard for all space telescopes will no longer be adequate. We will have to learn how to transfer to space the key technology that has made the new, big ground telescopes possible—namely, active control to correct distortion caused by mechanical or thermal disturbance. Ground telescopes automatically check star images as fast as every minute and make appropriate corrections of alignment and mirror distortion to maintain optical quality.

Detailed studies are being made of a telescope envisaged to succeed HST, the Next Generation Space Telescope (NGST), with a primary mirror 8 m in diameter operating at a temperature of 50 K (figure 4). Because the practical limit for a telescope launched in one piece is 4–6 m, depending on the size of future rocket fairings, it would be necessary to build the mirror and telescope in space. Planning has so far focused on automatic deployment from an unmanned payload. The whole spacecraft will have to weigh a lot less than HST, even though the telescope has ten times the mirror area, if it is to be carried a million miles from the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

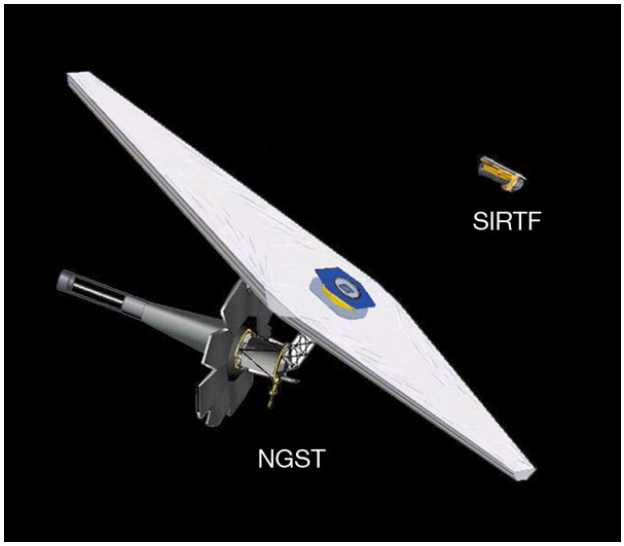


Figure 4. Evolution of infrared space telescopes. Shown to relative scale are SIRTF, a 0.85 m cryogenic telescope to be launched in 2002, and a concept for the 8 m NGST. Both will orbit far from the Earth, with thermal shields on the Sunward side.

The mind-stretching NGST goal has been a most effective way to focus research in universities, industry and government centres. New technologies have been identified and development work is going on in many areas. As an example, Jim Burge at the University of Arizona has been devising ways to take the mature technology used at the Mirror Laboratory for 8 m glass mirrors for the ground (weighing 18 tons) and produce an 8 m mirror in space weighing less than 1 ton. The idea, demonstrated 4 yr ago at the 0.5 m scale, is to keep only the first couple of millimeters of the reflecting glass surface and replace the rest with an ultralightweight carbon fibre truss. Figure 5 shows the 2 m diameter, 2 mm thick facesheet that will be connected to a truss by 166 active screw actuators. The whole prototype mirror assembly is 1/20 the weight of the 2.4 m HST mirror.

Although ultralightweight active mirrors are vital, they are but one of a range of challenging new technologies being studied for NGST. For example, passive cooling to cryogenic temperature is new and in order to cool an 8 m telescope to 50 K the sunshield shown in figure 4 must be the size of a tennis court. It must be such a good insulator that less than 1/1000 of the heat from the full blast of the Sun is allowed to leak through to be radiated into space. If, as currently conceived, the huge, flimsy NGST structure and sunshade are to deploy automatically, how does one test it beforehand in the Earth's gravity to be sure it will work?

An even more ambitious target has been set for the investigation of Earth-like planets. The basic approach is to detect their heat with a special form of interferometer

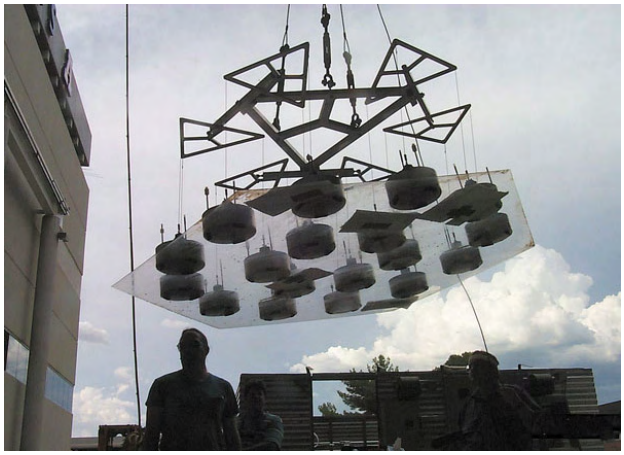


Figure 5. A 2 m hexagonal concave glass mirror is lifted clear after separation from the support on which it was thinned and polished. It is 2 mm thick and weighs just 30 lb (13.6 kg). (Image courtesy of J. Burge.)

in space. The beams from the different mirrors must be brought together with an accuracy of a nanometer despite tens of metres separation. Such accuracy is needed to fully cancel out the star's electromagnetic waves by maintaining the crests and troughs exactly out of phase, while leaving the planet image intact, a technique known as nulling interferometry. A mission called Terrestrial Planet Finder (TPF) is envisaged to explore not just the nearest few stars but over a hundred Sun-like stars out to a distance of 50 ly. This would require four or more cryogenic 4 m telescopes spread out over 100 m. They could be connected mechanically or orbited separately in formation.

Reliable and affordable solutions

Now that these scientific goals and the range of technical challenges required to accomplish them are identified, we must find the best path to proving solutions that will work reliably and will be affordable. Perhaps the best approach is an evolutionary process based on intermediate missions, each aimed at three goals: first, testing out one or more of the key technologies needed for the full-scale NGST and TPF missions; second, proving that new technology can be delivered affordably; third, using it for astronomical exploration.

The last requirement is crucial. The only way to find unanticipated flaws is by attempting challenging observations in space. Because each of the multiple technological developments for NGST and TPF is profound, any one of them alone could be used to open new scientific frontiers. Consider what might be done relatively soon. Even an old-fashioned, rigid telescope, if taken far from the Earth and passively cooled to 30 K, could open a new window to the high-redshift universe. A warm, 4 m, ultralightweight telescope could test autonomous active control for high-contrast images and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

see further than HST. A short nulling interferometer with proven, 1 m class cryogenic mirrors in a distant cold orbit could already find and obtain images and spectra of any Earth-like planets around the very nearest stars.

Such missions could also be configured to test solutions to what may be the most difficult problem, namely how to endow remote, cold observatories with HST-like longevity and productivity. If there were no possibility of repair and refurbishment, any big space telescope of HST-level or greater complexity would inevitably be risky and become dated. Space Station provides a natural solution, as a base where complete large telescopes and spacecraft could be assembled and tested in the gravity-free space environment. With solar cells deployed, the structure would be pretty flimsy, but a low-thrust tug could, in a few months, gently push it to the cold, stable point a million miles from Earth. By the same means, it could be brought back for repairs when necessary. The alternative is to leave the observatory in the cold and to send out cryogenic robots to fix it *in situ*.

These are exciting times. While we sort out the challenges of space telescopes, a golden era of discovery should come from the much higher resolution and sensitivity and data throughput of big ground telescopes and interferometers. Then from space we should be ready to open wide the relatively unexplored infrared spectrum, which holds such promise for discovery, from the furthest reaches of the universe to planets of the nearest, ordinary stars.

Acknowledgments

This article is a modified version of a previously published article: *Nature* **409** 427–430 (2001) <http://www.nature.com/>. Reproduced by courtesy of *Nature*.

Bibliography

Angel J R P and Fugate R Q 2000 *Science* **288** 455–6

Angel J R P, Cheng A Y S and Woolf N J 1986 *Nature* **322** 341–3

Angel J R P, Burge J, Hege K, Kenworthy M and Woolf N J 2000 *Proc. SPIE* **4013** 699–705

Burge J H, Cuerden B and Angel J R P 2000 *Proc. SPIE* **4013** 640–8

Tyson T, Kochanski G and Dell'Antonio I 1997 *The Universe at Low and High Redshift* ed W H Waller, M N Fanelli and A C Danks (New York: American Institute of Physics) pp 204–13

Tyson J A, Wittman D and Angel J R P Gravitational Lensing: Recent Progress and Future Goals *ASP Conference Series* ed T Brainerd and C Kochanek (The Astronomical Society of the Pacific) in press

Roger Angel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Telescopium

(the Telescope; abbrev. Tel, gen. Telescopii; area 252 sq. deg.) A southern constellation which lies between Ara and Indus, and culminates at midnight in early July. It was introduced to commemorate one of the long 'aerial telescopes' used by the Cassinis at the Paris Observatory by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, and it originally included stars from the neighboring constellations of Ophiuchus (which was larger then than it is now), Scorpius, Corona Australis and Sagittarius, but is now restricted to stars to the south of the last two constellations.

A small, inconspicuous constellation, the brightest star in Telescopium is α Telescopii, magnitude 3.5. There are no other stars brighter than magnitude 4.0. Interesting objects include RR Telescopii, a nova-like star (range 6.5–16.5) and the only one recognized as a periodic variable before its slow (1600 day) rise to seventh magnitude in 1944, and NGC 6584, a ninth-magnitude globular cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Teller, Edward (1908–)

American physicist, born in Budapest, Hungary; at George Washington University collaborated with George Gamow in studying nuclear beta decay, and astrophysics. Worked on the atomic and hydrogen bombs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tempel, Ernst Wilhelm Leberecht (1821–89)

German astronomer, born in Nieder-Kunersdorf, near Löbau, in Saxony; trained as a lithographer, he settled in Italy, and took to amateur astronomy, at first using the balcony of a Venetian palace and discovering a comet in 1859. Discovered the reflection nebula around Merope in the Pleiades. Became astronomer at Marseilles Observatory, discovered eight comets (including the comet associated with the Leonids) and five minor planets. Expelled in 1871 from France he returned to Italy, working at the Brera Observatory, Milan, adding to his total score of 13 comet and 6 minor planet discoveries, and became director of the Arcetri Observatory, where he worked with decrepit equipment on nebulae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Temperature Scales

Three different scales for describing temperature are in common use, the Celsius scale, the Kelvin scale and the Fahrenheit scale.

On the Celsius scale, which was devised in 1742 by the Swedish astronomer Anders Celsius (1701–44), temperatures are expressed in degrees Celsius (°C), where one Celsius degree is one-hundredth part of the difference in temperature between the freezing point and the boiling point of water (at standard sea-level atmospheric pressure). Because it is based on dividing that temperature range into 100 parts, it is also known as the centigrade scale. Temperatures below 0°C are given negative values.

On the Kelvin, or absolute, temperature scale, which was named after the British physicist Lord Kelvin (William Thomson 1824–1907), the unit of temperature is the kelvin (K), one kelvin being equal in magnitude to one degree Celsius. The zero of the kelvin scale is absolute zero, the temperature at which, classically, the motion of atoms would cease completely and which, therefore, is the lowest possible temperature. Absolute zero, 0 K, corresponds to a temperature of −273.15 °C. Because the kelvin is equal in magnitude to the Celsius degree, a temperature expressed in kelvin is equal to the temperature in degrees Celsius plus 273.15. Thus, for example, the freezing point of water (0 °C) is equal to (0 + 273.16) = 273.15 K and the boiling point of water (100 °C) is equal to 373.15 K. To convert from °C to K, add 273.15; to convert from K to °C, subtract 273.16. Because the Kelvin scale is directly related to the physical meaning of temperature, it is the most appropriate scale to use in a scientific context. The kelvin is the SI unit of temperature (the unit of temperature in the International System of Units) and is defined more specifically as 1/273.16 of the thermodynamic temperature of the triple point of water.

Although obsolescent, the Fahrenheit scale, which was devised by the German instrument-maker Gabriel Fahrenheit (1686–1736), is still used in some contexts. On the Fahrenheit scale, the unit of temperature is the degree Fahrenheit (°F), which is 1/180 of the difference in temperature between the freezing and boiling points of water. The freezing point of water is 32 °F and the boiling point of water, 212 °F. Since 180 Fahrenheit degrees corresponds to 100 Celsius degrees, one Fahrenheit degree is equivalent to 100/180 (=5/9) of one Celsius degree. To convert a temperature in °C to a temperature in °F, multiply by 9/5 and add 32. For example, 100 °C = (100 × 9/5) + 32 = 180 + 32 = 212 °F. Conversely, to convert from °F to °C, subtract 32 and then multiply by 5/9 (for example, 212 °F = (212 − 32) × (5/9) = 180 × (5/9) = 100 °C.

See also: absolute zero.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tenma (Astro-B)

Second Japanese x-ray satellite, launched in February 1983. Detected x-rays in the range 2–60 keV. Also carried a transient source monitor. Ceased operations in 1984. Tenma means 'Pegasus'. (See ASTRO.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tenskwatawa (1768–1837)

Amerindian, Shawnee leader, member of the Shooting Star clan, born in Piqua on the Mad River near Springfield, OH. He predicted a solar eclipse in 1806 and became known as 'The Prophet'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Terminator

The line marking the boundary between the day and night sides of a solar system object, as seen from the Earth or from a spacecraft. As the relative positions of the Earth, the Sun and the object shift, the proportion of the object's sunlit side visible from the Earth changes and the terminator moves across the object's disk. The terminator thus determines the object's phase. From the Earth, the Moon, Mercury, Venus and Mars show terminators. The Moon's rough terrain gives its terminator a jagged appearance, and sometimes elevated features just to the night side of the terminator are visible, obliquely illuminated by the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Terrestrial Coordinate Systems and Frames

A terrestrial reference system (TRS) is a spatial reference system corotating with the Earth in its DIURNAL MOTION in space. In such a system, the positions of points anchored on the Earth's solid surface have coordinates which have only small variations with time, as a result of geophysical effects (tectonic or tidal deformations; see TECTONICS, EARTH'S INTERIOR, TIDES). A terrestrial reference frame (TRF) is a set of physical points with precisely determined coordinates in a specific coordinate system (Cartesian, geographic, mapping, ...) attached to a TRS. Such a TRF is said to be a realization of the TRS.

Concepts and terminology

These concepts were redefined extensively in the 1980s by the astronomical and geodetic communities (see for instance Kovalevsky *et al* 1989).

Ideal terrestrial reference systems

An ideal TRS is defined as a three-dimensional reference frame (in the mathematical sense) close to the Earth and corotating with it. In the Newtonian background, the geometry of physical space considered as a Euclidean affine space of dimension 3 provides a standard and rigorous model of such a system through the selection of an affine frame (O,E). O is a point of the space called the origin. E is a vector base of the associated vector space. The currently adopted restrictions on E are that it is orthogonal with the same length for the base vectors. Moreover, one adopts a direct orientation. The common length of these vectors is called the scale of the TRS and the set of unit vectors collinear with the base its orientation:

$$\lambda = \|\vec{E}_i\| \qquad i = 1, 2, 3.$$

Two major types of TRS are in use.

- In the local or topocentric ones the origin is located on the Earth's surface and the orientation uses the local vertical direction for the Z axis. In general, Cartesian coordinates or plane coordinates are used.
- In the geocentric ones the origin is close to the geocenter and the orientation is equatorial (the Z axis is the direction of the pole). In this case, Cartesian coordinates, geographical coordinates or plane (map) coordinates are used (see GEOCENTRIC COORDINATE SYSTEM).

The first type is used for technical purposes (laboratory experiments, local topography and surveying, ...) while the second is the current practice for astrometry or global geodesy. In the case of geocentric systems, one further distinguishes two subtypes, depending on the method used to establish and realize such systems (classical terrestrial or space techniques).

- The first is true geocentric which means that realizations will be close to the geocenter (a few meters at the beginning of satellite geodesy, now a few centimeters; see SPACE GEODESY) and the zero MERIDIANS of the equatorial orientations will be almost identical (currently expressed using the name of Greenwich, for historical reasons).
- The second is quasi-geocentric (this is a better term than non-geocentric, which may be too easily confused with topocentric) for systems established with terrestrial geodetic techniques usually using a fundamental station. In such cases, the origin may be separated from the geocenter by several hundred meters. Notice also that these systems currently use a zero meridian linked to some astronomical observatory and therefore should be realigned with the conventional Greenwich origin by adopting a longitude for this zero meridian with regard to Greenwich. For instance, in Europe, several conventional origins are still in use for mapping purposes (Paris, Rome, Madrid, ...).

Under these restrictions, the general transformation of the Cartesian coordinates of any point close to the Earth from any one TRS to any other one will be given by a three-dimensional similarity (T is a translation vector, λ a scale factor and R a rotation matrix):

$$X^{(1)} = T_{1,2} + \lambda_{1,2} R_{1,2} X^{(2)}.$$

This concept can be generalized in the framework of a relativistic background model such as Einstein's GENERAL THEORY OF RELATIVITY, using the spatial part of a local Cartesian frame (Boucher 1986).

Conventional terrestrial reference system

Such a system designates the set of all conventions, algorithms and constants which determine the estimation of coordinates of points in a specific ideal TRS.

Conventional terrestrial reference frame

A conventional TRF is defined as a set of physical points with precisely determined coordinates in a specific coordinate system as a realization of an ideal TRS. Two types of frames are currently distinguished (this is also valid for celestial references), namely dynamical or kinematical ones, depending whether a dynamical model is applied in the determination process of these coordinates or not.

Examples of dynamical TRF are ephemerides of artificial satellites. This approach is currently widely used: EPHEMERIDES of navigation satellite systems, such as the Global Positioning System (GPS), or ephemerides of satellite radar altimeters used to derive sea surface heights in a well-realized TRS (Topex Poseidon, or the forthcoming Jason 1).

However, kinematical TRFs are still the most-used type, especially with networks of physical points on the topographic surface of the solid Earth for which numerical values of coordinates are estimated. We can identify three major purposes for the use of such a TRF.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- In a metrological approach it is a physical realization of a TRS, in the way that standards realize units.
- In a user-oriented approach the points can be used to determine other points (densification or expansion). This is a major role for networks.
- In a geophysical approach the points are deformation sensors of the solid Earth's surface. This has the double value of providing data for geophysical knowledge and improving the estimation strategy for positions by using better geophysical models.

Networks can be global if they cover more or less sparsely the whole Earth or regional if they cover a region of the Earth, such as a continent or a country. They can also be considered according to their type of coordinate system and are then specifically qualified by the following:

- horizontal if using plane coordinates or geographical latitudes or longitudes;
- vertical if using heights;
- three-dimensional if using Cartesian or Geographical coordinates (with ellipsoidal heights);
- astronomical, if using astronomical latitudes and longitudes.

Another interesting criterion is the nature of the physical points.

- A classical network uses passive markers or benchmarks, available as either a station point or a target for the user's measurements.
- A permanent network consists of beacons or receivers operating permanently and providing data suitable for precise determination of their position. This is the typical case for space geodetic ground tracking systems.
- In active networks a permanent station provides additional information (on their precise position) enabling users to determine positions of other points in the same TRS. Such current additional data are the raw measurements performed by the station or derived from differential corrections such as DGPS. Such data are available in real time or not, depending on various communication media.

Realizations of the terrestrial reference systems: past, present and future

Until 1960, the epoch of the availability of the first artificial satellites of the Earth, the terrestrial reference systems were realized at two levels. The primary realization was *de facto* obtained by the set of the various astronomical observatories which observed the Sun, Moon and planets, as well as star positions. This was therefore an astronomical network in which the longitudes were unified only progressively, national observatories keeping for a long time their own origin. At the end of the 19th century, an international origin was adopted, initially based on the ROYAL OBSERVATORY, GREENWICH in the UK.

The secondary realizations were developed through geodetic activities around the world, mainly driven by mapping needs. Since the birth of the triangulation method attributed to GEMMA FRISIUS in 1533, geodetic control by triangulation (and precise levelling in the 19th century) progressively covered the part of the Earth's surface not occupied by oceans or seas. While the shape of these networks was provided by angle and distance measurements, their orientation and absolute position with regard to the Earth's body were obtained by astronomical measurements (latitude, longitude and azimuths). Consequently, each set of geodetic coordinates (either geographical or plane) of the control points belonging to a given geodetic network constituted a specific TRF. Such networks were of local, national or even continental extension. The progressive unification of these regional networks was developed at the end of the 19th and beginning of the 20th century by the establishment of international junctions (Channel, Baltic Sea, Gibraltar, . . .) and the readjustment of control networks of continental size (North American Datum, European Datum, . . .). (For more details, see Boucher (1989).)

In 1899, the International Latitude Service was established and modified into the International Polar Motion Service (IPMS) in 1964. Meanwhile, the Bureau International de l'Heure (BIH) was created in 1912 to realize an international time scale, which also required the Earth orientation parameters (EOPs) to be determined.

In the late 1950s, new methods using space techniques appeared providing extensive information on the gravity field and EOPs as well as tracking station positions. These space techniques progressively superseded classical astrogeodetic methods. A few years of international cooperation (MERIT campaign; see Wilkins 1988) started an intercomparison of these various techniques (VLBI, laser ranging to satellites or to the Moon, radio techniques). During this campaign, a fundamental change appeared as the primary TRF started to be realized by the three-dimensional Cartesian positions of tracking stations, replacing astronomical coordinates (Boucher and Feissel 1984). As a result, the INTERNATIONAL ASTRONOMICAL UNION (IAU) and the International Union of Geodesy and Geophysics (IUGG) decided to replace BIH and IPMS by a new service, the International Earth Rotation Service (IERS), established in 1988. Its mission was to provide to the worldwide scientific and technical community reference values for the EOPs and reference realizations of internationally accepted celestial and terrestrial reference systems.

The IERS is at present the structure in charge of realizing, using and promoting the International Terrestrial Reference System (ITRS) as defined by the IUGG resolution No 2 adopted in Vienna, 1991.

The ITRS definition fulfils the following conditions.

- It is geocentric, the center of mass being defined for the whole Earth, including oceans and atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

- The unit of length is the meter (SI). This scale is consistent with the TCG time coordinate for a geocentric local frame, in agreement with IAU and IUGG (1991) resolutions (see TIME SCALES: TERRESTRIAL). This is obtained by appropriate relativistic modeling.
- Its orientation was initially given by the BIH orientation at 1984.0. The IERS Reference Pole (IRP) and Reference Meridian (IRM) are consistent with the corresponding directions in the BIH directions within $\pm 0.005''$. The BIH reference pole was adjusted to the Conventional International Origin (CIO) in 1967; it was then kept stable independently until 1987. The uncertainty of the tie of the IRP with the CIO is $\pm 0.03''$.
- The time evolution of the orientation is ensured by using a no-net-rotation condition with regard to horizontal tectonic motions over the whole Earth.

See the IERS Conventions (1996) for a detailed description of ITRS.

Realizations of the ITRS are produced by the Terrestrial Frame Section of the IERS Central Bureau under the name International Terrestrial Reference Frames (ITRF); these consist of lists of coordinates (and velocities) for a selection of IERS sites (tracking stations or related ground markers) obtained by combination of several space techniques, currently by VLBI, LLR, GPS, SLR, and DORIS. General documentation on terrestrial reference systems and frames is available on WWW: http://lareg.ensg.ign.fr/ITRF/.

The ITRS can be connected to the International Celestial Reference System (ICRS) by use of the IERS EOPs. ITRS is also used as a reference for specific products which include TRF data. For instance, the World Geodetic System (WGS) program uses ITRS; in particular the most recent system, WGS84, is carefully linked to ITRS (Malys and Slater 1994)

Bibliography

Boucher C 1986 Relativistic effects in geodynamics *Relativity in Celestial Mechanics and Astrometry* ed J Kovalevsky and V A Brumberg (Dordrecht: Reidel) pp 241–53

Boucher C 1989 Current intercomparisons between CTSs *Reference Frames in Astronomy and Geophysics* ed J Kovalevsky, I I Mueller and B Kolacek (Dordrecht: Kluwer) pp 324–48

Boucher C and Feissel M 1984 Realization of the BIH terrestrial system *Proc. Int. Symp. on Space Techniques for Geodynamics* ed J Somogyi

Guinot B 1981 Comments on the terrestrial pole of reference, the origin of the longitudes and on the definition of UT1 *Reference Coordinate Systems for Earth Dynamics (Proc. IAU Colloq. 56)* ed E M Gaposchkin and B Kolaczek (Dordrecht: Kluwer) pp 125–34

Kovalevsky J, Mueller I I and Kolacek B (ed) 1989 *Reference Frames in Astronomy and Geophysics* (Dordrecht: Kluwer)

Malys S and Slater J 1994 Maintenance and enhancement of the World Geodetic System 1984 *Proc. ION-94 (Salt Lake City, September 1994)*

Sillard P, Altamimi Z and Boucher C 1998 The ITRF96 realization and its associated velocity field *Geophys. Res. Lett.* **25** (17) 3223–6

Wilkins G A 1988 Review of the achievements of project MERIT for the intercomparison of techniques for monitoring the rotation of the Earth *The Earth's Rotation and Reference Frames for Geodesy and Geodynamics (Proc. IAU Symp. 128)* ed A K Babcock and G A Wilkins (Dordrecht: Kluwer) pp 227–32

Claude Boucher

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Terrestrial Planets

A collective term for the major planets Mercury, Venus, Earth and Mars, which are all 'Earthlike' in being of the order of 10 000 km in diameter, with densities of roughly 5000 kg m^{-3}, having a rock/metal composition with a solid surface, and possessing a small number of satellites or none at all. Pluto was once counted as a terrestrial planet, but once its low size and mass became apparent it was excluded from this category. The Earth is the largest and most massive of the terrestrial planets. Mercury, the smallest, is also the densest, and possesses the most rarefied atmosphere. The densest atmosphere, with a surface pressure 90 times the Earth's, is on Venus. Mars is the least dense terrestrial planet.

Compare: giant planets.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tethys

A mid-sized icy satellite of Saturn, discovered by Giovanni Cassini in 1684. Its diameter is 1060 km, and its orbital distance from Saturn is 295 000 km. The parts of Tethys imaged in the Voyager missions show a variety of terrains with different crater densities. The heavily cratered side contains a large impact feature, the 440 km diameter basin Odysseus, a 'relaxed' structure with signs of an inner mountain ring. On the hemisphere opposite Odysseus is a lightly cratered plain cut by Ithaca Chasma, a huge valley system of multiple faults over 100 km across at its widest and around 1000 km long, whose formation may be linked with the impact that produced Odysseus. The plains are a consequence of resurfacing in the distant past. The satellite's overall density is 1000 kg m^{-3}, indicating a composition largely of water ice. The orbit is shared with two co-orbital satellites, Calypso and Telesto.

See also: co-orbital satellites, Saturn: satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thémis

The Thémis 'Telescope Heliographique pour l'Étude des Instabilites Magnetiques sur le Soleil', telescope has been built and is operated in the frame of a cooperation between the CNRS-INSU (Conseil National de la Récherche Scientifique-Institut National des Sciences de l'Univers, France) and the CNR (Consiglio Nazionale delle Ricerche, Italy).

The telescope is located at the 'OBSERVATORIO DEL TEIDE' in Tenerife (Canary Islands, Spain) which is operated by IAC (INSTITUTO DE ASTROFÍ SICA DE CANARIAS) under the International Agreements on Astrophysics between Spain and many other countries. The site coordinates are: longitude 16 30 31.6, latitude 28 18 13.8, altitude 2456 m above sea level. A total of about 15 permanent staff are working at Thémis, in cooperation with several French and Italian research groups (DASOP-Meudon and DT-INSU in Paris, Midi-Pyrenees in Toulouse, Arcetri Observatory and University in Florence, University Tor Vergata in Rome).

Thémis is a 90 cm diameter Ritchey–Chrétien altazimuthal telescope conceived to observe solar magnetic structures in the visible spectrum with high (spatial, spectral and polarimetric) resolution. The design of the telescope has been realized to minimize the instrumental polarization.

The main observing modes offered by Thémis are MTR (Multi Lines Spectroscopy), MSDP (Multi Channel Subtractive Double Pass Spectroscopy), both with or without polarization analysis, and IPM (Italian Panoramic Monochromator) without polarization analysis.

MTR gives high-resolution spectra simultaneously in several spectral lines and sequentially in the four Stokes parameters. MSDP gives high spectral resolution imaging simultaneously in several spectral intervals. IPM allows very narrow-band imaging scanning a line profile.

For further information see
http://www.obs-nice.fr/themis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thabit Ibn Qurra, Abu'l-Hasan (836–901)

Astronomer, born in Harran, Mesopotamia (now Turkey), became court astronomer in Baghdad, translated Greek works on mathematics and wrote on astronomy, including *Concerning the Motion of the Eighth Sphere*, on the apparent motions of the stars, analysed the movements of Sun and Moon and wrote treatises on sun-dials. Some of his books were translated into Latin by Gherard of Cremona and distributed in Europe.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thales of Miletus (624–546 BC)

The first known Greek philosopher; also a scientist, mathematician and engineer, born in Miletus, Asia Minor (now Turkey), believed to have been the teacher of ANAXIMANDER. Apparently wrote a book on navigation, defining the constellation Ursa Minor and using it as a navigation aid. Credited with the prediction of an eclipse of the Sun in 585 BC, although it is not known how, since the Metonic cycle was known only for eclipses of the Moon at this time (eclipses of the Sun are harder to spot since solar eclipses are visible at different places on Earth, whereas lunar eclipses can be seen from half the Earth). Nevertheless the eclipse stopped the war between the Medes and the Lydians, according to Herodotus. Thales determined the height of pyramids by measuring the length of their shadow at the moment when a man's shadow is equal to his height. He taught that 'all things are water', which may seem an unpromising hypothesis, but was the first expression of the idea of explaining different phenomena from a common underlying cause, and foreshadowed scientific theory. Like the Egyptians, Thales believed that the Earth was a flat disc floating on an infinite ocean, explaining earthquakes as the shaking of the floating 'boat'. Thales was gazing at the sky one night and fell into a ditch, causing a servant girl to ask 'How do you expect to understand what is going on up in the sky if you do not even see what is at your feet' (the first absent-minded professor joke).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tharsis Montes

The official name of the region on Mars known as the Tharsis Ridge or Tharsis Bulge, a large raised area in the northern hemisphere of Mars, centered at 2.8 °N, 113.3 °W. It takes its name from the ancient Spanish town known formerly as Tartessus. Tharsis is a large uplifted area, measuring 2105 km across at its greatest dimension, which contains the planet's most prominent volcanoes. In a 1000 km chain from northeast to southwest lie Ascraeus Mons (rising to 26 km above Mars's average surface level), Pavonis Mons (18 km) and Arsia Mons (20 km), all of which are 150 km or so across, while 1000 km to the northwest is the solar system's largest volcano, Olympus Mons. These peaks apart, the highest point on the Tharsis Bulge is the complex fractured terrain named Noctis Labyrinthus, with an elevation of 11 km. This is part of an extensive fracture system aligned approximately on the center of Tharsis; other prominent fractures are Ceraunius Fossae to the north and Claritas Fossae in the south. From the east of the Tharsis region, the giant canyon system of Valles Marineris extends eastward. It may be that the whole region was uplifted by an upwelling in the underlying mantle, possibly as a consequence of the event that created the impact basin Hellas on the opposite side of the planet.

See also: Hellas Planitia, Mars: surface, Olympus Mons, Valles Marineris.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The 2dF Galaxy Redshift Survey

The 2dF Galaxy Redshift Survey (2dFGRS) produced a three-dimensional map of the distribution of 221 000 galaxies covering 5% of the sky and reaching out more than 3 billion light-years to a redshift $z \approx 0.3$. The survey used the Anglo-Australian Telescope's (AAT's) 2-degree Field facility (2dF), which could observe 400 objects simultaneously over a 2° diameter field of view. Survey observations began in 1997 and were completed in 2002; the final survey was an order of magnitude larger than any previous redshift survey.

The 2dFGRS yielded the first map of the large-scale structure in the local (low-redshift) universe to probe a statistically representative volume on scales from tens of kiloparsecs to hundreds of megaparsecs. Analysis of the map gave the first direct evidence that the large-scale structure of the universe grew through gravitational instability, and provided a precise measurement of the mean mass density of the universe. The observed structure also displayed a weak signature produced by the interactions between baryons and photons in the very early universe, providing an estimate of the ratio of baryonic matter to dark matter. The combination of the 2dFGRS observations of the structure in the present-day universe with observations of the structure in the very early universe from the cosmic microwave background (CMB) gave precise measurements of the Hubble constant, and the cosmological constant and provided an improved upper limit for the mass of the neutrino.

The massive sample of galaxies observed by the 2dFGRS has also been used to precisely characterize multiparameter distributions and relationships within the galaxy population: the distribution of galaxies over luminosity, surface brightness, spectral type, star-formation rate and environment (local density) and the variation in clustering for galaxies of different luminosities and spectral types. The survey also yielded the first measurement of the galaxy bias parameter, which links the fluctuations in the number of galaxies and the amount of dark matter and is a basic constraint on models of galaxy formation and evolution.

The History of the 2dFGRS

The seed of the 2dFGRS was planted in the late 1980s, when a British astronomy review gave high priority to converting existing telescopes to wide-field science, especially multi-object spectroscopy. For the AAT, the case for such a conversion was greatly strengthened when Roderick Willstrop and Charles Wynne showed that, with suitable correction optics, the telescope could be adapted to permit a 2° diameter field of view at the prime focus (Willstrop 1987; Wynne 1989).

The Anglo-Australian Observatory, under Director Russell Cannon and instrumentation scientist Keith Taylor, combined this opportunity with its established strength in robotic multi-object spectroscopy in order to develop a wide-field, multi-object spectrograph: the 2dF facility (Taylor and Gray 1990; Gray *et al.* 1993). With massive field-correction optics including an atmospheric dispersion compensator, a pair of tumbling focal planes to minimize dead time, 400 optic fibers mounted in magnetic buttons, a fully robotic positioning system, and end-to-end software for control, configuration, data-taking and reductions, 2dF was, for its time, a visionary instrument. Construction began in 1990.

From its conception, the main scientific driver for 2dF was a massive redshift survey, sampling a representative volume of the universe in fine detail in order to precisely measure fundamental cosmological parameters. The best existing and ongoing surveys in the early 1990s (e.g. the IRAS Point Source Catalogue redshift survey and the Las Campanas Redshift Survey) either did not cover sufficiently large volumes to be statistically representative of the large-scale structure or covered large volumes too sparsely to provide precise measurements. An order-of-magnitude increase in the volume and sample size of redshift surveys was required to enter the regime of 'precision cosmology'. This idea was developed into a concrete proposal by the founding members of the survey team during 1993 and 1994. The first survey proposal was submitted to the telescope time committees and the AAT Board in 1994, and envisaged a survey of 250 000 galaxies over 2000 deg^2 to a magnitude limit of b_J=19.5. The survey was estimated to require 125 nights on the AAT and was projected to be completed by the end of 1996.

In fact, 2dF was placed on the AAT in November 1995, and the first spectra were taken in mid-1996. Scheduled observations with 2dF at full functionality began in September 1997, and the first major redshift survey observing run occurred in October 1997. Progress with the survey was slow to begin with, but the survey passed 50 000 redshifts in mid-1999 and 100 000 redshifts in mid-2000 (Colless 1999). The first 100 000 redshifts and spectra were released publicly in June 2001 (Colless *et al.* 2001), and the 200 000-redshift mark was achieved towards the end of 2001. The survey observations were completed in April 2002, after 5 years and 272 nights on the AAT.

The 2dF Wide-Field Fiber Spectrograph Facility

As its name implied, the 2dFGRS was tightly coupled to the capabilities of the 2dF of the AAT. The crucial parameters of 2dF for a massive redshift survey were as follows: (i) the wide (2° diameter) field of view, which allowed the 1800 deg^2 of the survey to be covered with 900 overlapping fields; (ii) the large number

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

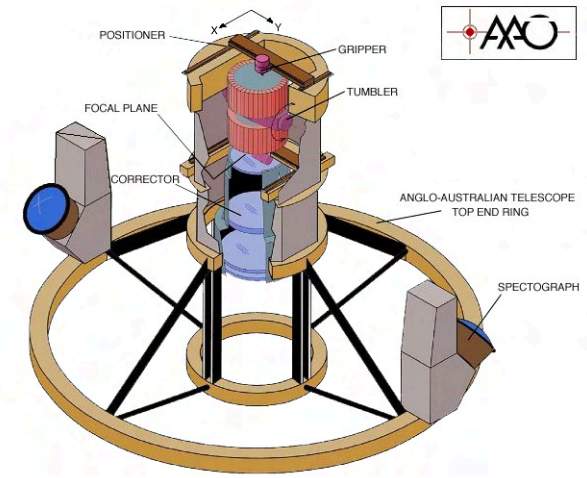


Figure 1. A schematic of the 2dF facility showing the major components of the instrument. (K. Taylor, Anglo-Australian Observatory.)

(400) of simultaneous spectra, which allowed the typical number of galaxies in the field of view down to b_J=19.5 to be observed at once; (iii) the 4 m aperture of the AAT, which allowed redshifts to be measured for galaxies as faint as b_J=19.5 in 1 h observations; (iv) the capacity for rapid reconfiguration, which meant that there was no dead-time between observations of different fields.

To achieve these complex requirements, a complex instrument was required. All the components of 2dF are contained in the 2dF top-end ring structure, which is mounted at the prime focus of the AAT (see figure 1).

A photon reflected by the AAT's 4 m primary mirror towards the prime focus encounters, first, the wide-field corrector optics (including an atmospheric dispersion corrector) that provide a 2° diameter field of view with good imaging properties; then it is collected at the focal plane by the fiber buttons that route it down the optic fibers; finally, it enters one of the twin spectrographs, each of which measures the spectra from 200 fibers at a time. The total system efficiency is about 5%.

The fiber buttons are magnetically attached to a metal plate in the focal plane. They are positioned on the plate by a high-precision robot, which moves over the plate on an X–Y stage and has a gripper arm that can be raised or lowered to pick or place a button on the plate. The robot can position a fiber button on the plate with a precision of 0.3 arcsec (20 μm) in 6 s. In order to maximize observing efficiency there are two sets of fiber buttons, and the plates

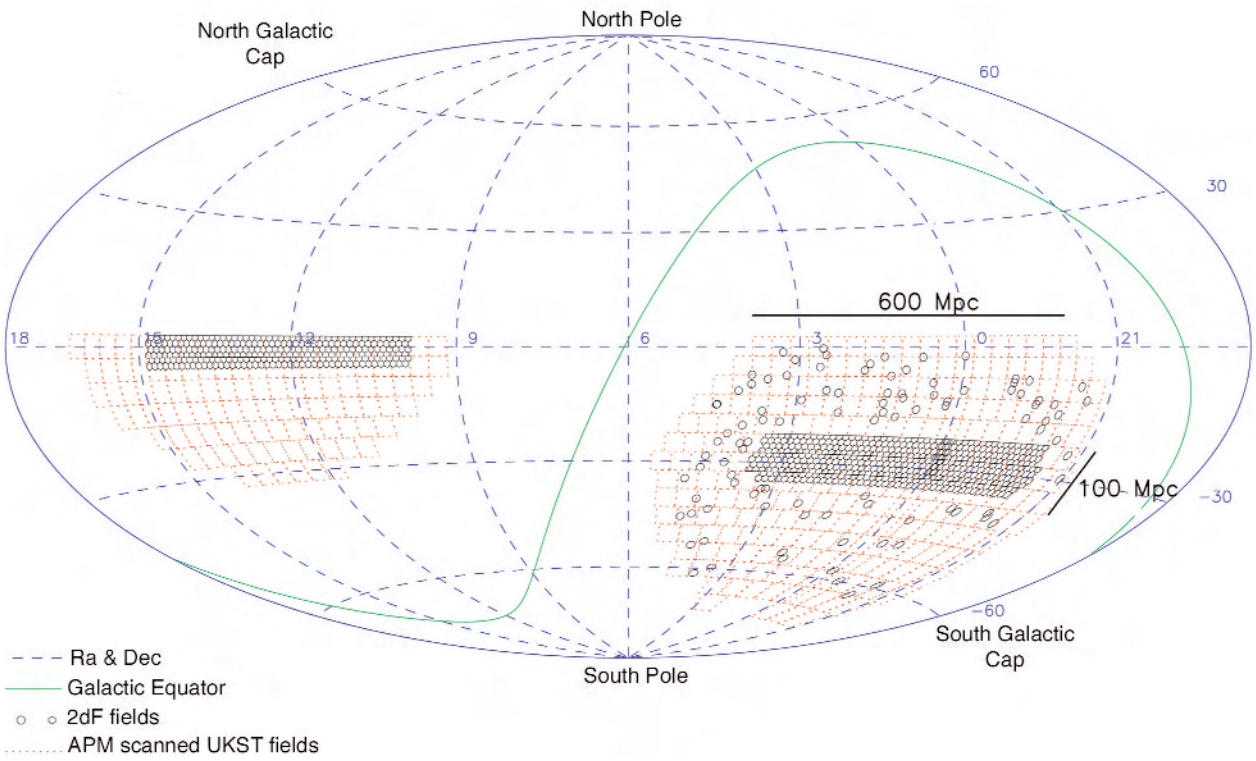


Figure 2. A map of the sky showing the locations of the two 2dFGRS survey strips (NGP strip at left, SGP strip at right) and the random fields. Each 2dF field in the survey is shown as a small blackcircle; the sky survey plates from which the source catalog was constructed are shown as small red squares. The scale of the strips at the typical redshift of the survey is indicated. (S. J. Maddox, University of Nottingham.)

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

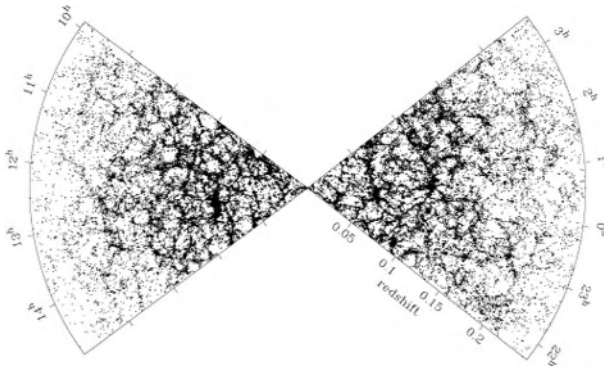


Figure 3. A 3° thick slice through the 2dF Galaxy Redshift Survey map. The slice cuts through the NGP strip (at left) and the SGP strip (at right). It contains 63 000 galaxies and shows the large-scale structures in the galaxy distribution. (J. A. Peacock, University of Edinburgh.)

on which they sit are tumbled between the focal plane position and the robot position. This allows the robot to configure the fibers on one plate while the fibers on the other plate are observing the sky.

A full technical description of the 2dF facility is given by Lewis *et al.* (2002).

Survey Observations

The source catalog for the 2dFGRS survey was a revised and extended version of the APM galaxy catalog (Maddox *et al.* 1990), which was created by scanning the photographic plates of the UK Schmidt Telescope Southern Sky Survey. The survey targets were chosen to be galaxies with extinction-corrected magnitudes brighter than b_J=19.5. The galaxies were distinguished from stars by an image classification algorithm that was conservatively designed to include all galaxies at the expense of also including a 5% contamination of the sample by stars.

The main survey regions were two declination strips, one in the southern Galactic hemisphere spanning 80° × 15° around the South Galactic Pole (the SGP strip), and the other in the northern Galactic hemisphere spanning 75° × 10° along the celestial equator (the NGP strip); in addition, there were 99 individual 2dF 'random' fields spread over the southern Galactic cap (see figure 2). The survey covers approximately 1800° and has a median redshift depth of z=0.11. An adaptive tiling algorithm was used to optimally place the 900 2dF fields over the survey regions, giving a highly complete and uniform sample of galaxies on the sky.

Redshifts were measured from 2dF spectra that covered the range from 3600 Å to 8000 Å at a resolution of 9.0 Å. Redshift measurements were obtained both from cross-correlation with template spectra and from fitting emissioning lines. All redshifts were visually checked and assigned a quality parameter Q in the range 1–5; accepted redshifts (Q>3) were found to be 98% reliable and to have a typical uncertainty of 85 km s^{-1}. The overall redshift completeness for accepted redshifts was 92%, although this varied with magnitude.

Figure 3 shows a thin slice through the three-dimensional map of over 221 000 galaxies produced by the 2dFGRS. The observer is at the center of the map, with redshift increasing towards the edge and angle on the sky (RA) marked around the circumference. This 3° thick slice passes through both the NGP strip (at left) and the SGP strip (at right). The decrease in the number of galaxies towards higher redshifts is an effect of the survey selection by magnitude—only intrinsically more luminous galaxies are brighter than the survey magnitude limit at higher redshifts. The clusters, filaments, sheets and voids making up the large-scale structures in the galaxy distribution are clearly resolved. Importantly, there are many such structures visible in the figure, demonstrating that the survey volume comprises a representative sample of the universe.

Cosmological Results

Although survey observations finished in April 2002, as of this writing the scientific analysis of the survey is still in progress. So far the main results to emerge have been cosmological, in keeping with the primary goals of the survey as it was envisaged a decade ago when 2dF was being built.

The first step in this analysis is a statistical characterization of the large-scale structure in the galaxy distribution, which in most cosmological models is fully described, at least on the largest scales, by just two quantities: the mean density and the variation in the typical size of the fluctuations in the density as a function of scale. The latter is best quantified on large scales by the power spectrum of the galaxy distribution, which Percival *et al.* (2001) measured from the 2dFGRS on scales up to 300 Mpc. The shape of the power spectrum is consistent with a cold dark matter (CDM) model having a shape parameter $\Gamma = \Omega_M h = 0.20 \pm 0.03$, which, for a Hubble constant around 70 km s^{-1} Mpc^{-1} (i.e. $h \approx 0.7$), implies a mean mass density $\Omega_M \approx 0.3$. The power spectrum also showed the weak signature of acoustic oscillations produced by baryon-photon coupling in the early universe, allowing Percival *et al.* to also estimate the mass fraction in baryons, $\Omega_b / \Omega_M = 0.15 \pm 0.07$.

The galaxy distribution on intermediate scales provides another measure of the mass density, and also a test of the fundamental assumption that structures grow due to gravitational instability. Under this paradigm, small-scale structures are undergoing non-linear collapse, while on larger scales structures are collapsing coherently. These effects lead to a characteristic distortion of the galaxy distribution observed in redshift space, with the small, high-density clusters of galaxies stretched out

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

along the line of sight and the larger, moderate-density filaments and sheets compressed along the line of sight. The first unambiguous detection of these redshift-space distortions was obtained from the 2dFGRS by Peacock *et al.* (2001), confirming that present-day structures are due to the gravitational amplification of small initial density fluctuations. The amplitude of the distortion depends on the quantity $\beta = \Omega^{0.6}/b$, where b parametrizes the bias of the galaxies with respect to the overall mass distribution. By measuring the compression of the galaxy two-point correlation function along the line of sight, Peacock *et al.* obtained $\beta = 0.54 \pm 0.09$ for galaxies with luminosities similar to the Milky Way. If such galaxies are relatively unbiased tracers of the galaxy distribution (i.e. $b \approx 1$), this also implies $\Omega_M \approx 0.3$.

Stronger constraints on these and other fundamental cosmological parameters were obtained by combining the power spectrum of the present-day galaxy distribution from the 2dFGRS with the power spectrum of the mass distribution at very early times derived from observations of the anisotropies in the CMB. Such analyses have been carried out by Efstathiou *et al.* (2002) and Percival *et al.* (2002), and show that the universe has a flat or near-flat geometry ($\Omega_k = 0 \pm 0.05$), with a low total matter density ($\Omega_M = 0.31 \pm 0.06$) and a large positive cosmological constant ($\Omega_\Lambda = 0.69 \pm 0.06$, consistent with the independent estimates from observations of high-redshift supernovae). The physical densities of cold dark matter and baryons are $\omega_c = \Omega_c h^2 = 0.12 \pm 0.01$ and $\omega_b = \Omega_b h^2 = 0.022 \pm 0.002$; the latter agrees very well with the constraints from big bang nucleosynthesis. The joint 2dFGRS+CMB analysis also provided an estimate of the Hubble constant ($H_0 = 67 \pm 5$ km s^{-1} Mpc^{-1}) that is independent of, but in excellent accord with, the results from the Hubble Space Telescope Key Project. Elgaroy *et al.* (2002) showed that these combined constraints place a strong upper limit on the neutrino mass fraction, $\Omega_\nu/\Omega_M < 0.13$ (95% confidence), which translates to an upper limit on the total neutrino mass (summed over all species) of $m_\nu < 1.8$ eV.

Another first for the 2dFGRS was the measurement of the bias, b, between the galaxy distribution and the underlying matter distribution, defined as the square root of the ratio of the galaxy and matter power spectra. Lahav *et al.* (2002) measured this ratio using the 2dFGRS and CMB power spectra, and found that for galaxies with characteristic (L^*) luminosities the bias parameter is $b^* = 0.96 \pm 0.08$. An independent analysis by Verde *et al.* (2002), using only the 2dFGRS and based on the higher-order correlations of the galaxy distribution in the quasi-linear regime, obtained a very similar result: $b^* = 0.92 \pm 0.11$. The conclusion is that L^* galaxies are nearly unbiased tracers of the mass distribution. However, Norberg *et al.* (2001, 2002a) showed that the bias parameter does vary with luminosity and spectral type, ranging from b = 1.5 for bright, early-type galaxies to b = 0.8 for faint, late-type galaxies.

The measurement of cosmological parameters from the 2dFGRS has made a significant contribution to shaping the current consensus model for the fundamental properties of the universe that has emerged from a range of independent observations, including the measurements of the CMB anisotropies, the distances to high-redshift supernovae and big bang nucleosynthesis.

Properties of the Galaxy Population

Alongside these cosmological studies, the 2dFGRS has also yielded a wide range of results on the properties of the galaxy population and provided strong new constraints for models of galaxy formation and evolution. Highlights in this area to date include the following: (1) precise determinations of the optical and near-IR galaxy luminosity functions (Cole *et al.* 2001; Norberg *et al.* 2002b); (2) a detailed characterization of the variations in the luminosity function with spectral type (Folkes *et al.* 1999; Madgwick *et al.* 2002); (3) a determination of the bivariate distribution of galaxies over luminosity and surface brightness (Cross *et al.* 2001); (4) the first precise measurement of the dependence of galaxy clustering on both luminosity and spectral type (Norberg *et al.* 2002a); (5) a constraint on the space density of rich clusters of galaxies from the velocity dispersion distribution for identified clusters (De Propris *et al.* 2001); (6) separate radio luminosity functions for AGN and star-forming galaxies (Sadler *et al.* 2001; Magliocchetti *et al.* 2002); (7) constraints on the star-formation history of galaxies from the mean spectrum of galaxies in the local universe (Baldry *et al.* 2002); and (8) a measurement of the environmental dependence of star-formation rates of galaxies in clusters (Lewis *et al.* 2002).

These results are just a fraction of the information that can be extracted from the 2dFGRS on the properties of galaxies and their relation to the large-scale structure of the galaxy distribution. Many more results are still to emerge from analysis of the survey and from combining the 2dFGRS with other large surveys and with detailed follow-up observations.

Bibliography

The 2dF Galaxy Redshift Survey web pages (www.mso.anu.edu.au/2dFGRS) contain detailed information about the survey, a gallery of images, an access point for the survey database, and a list of all 2dFGRS publications. The most relevant publications, referenced in this article, are as follows:

Baldry I, *et al.* 2002 *Astrophysical Journal* **569** 582

Cole S, *et al.* 2001 *Monthly Notices of the Royal Astronomical Society* **326** 255

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Colless M M 1999 *Philosophical Transactions of the Royal Society of London* **357** 105
Colless M M, *et al.* 2001 *Monthly Notices of the Royal Astronomical Society* **328** 1039
Cross N, *et al.* 2001 *Monthly Notices of the Royal Astronomical Society* **324** 825
De Propris R, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **329** 87
Efstathiou G, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **330** L29
Elgaroy O, *et al.* 2002 *Physical Review Letters* **89** 061301
Folkes, S *et al.* 1999 *Monthly Notices of the Royal Astronomical Society* **308** 459
Lahav O, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **333** 961
Lewis I, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **333** 279
Maddox S J, *et al.* 1990 *Monthly Notices of the Royal Astronomical Society* **242** 43P
Magliocchetti M, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **333** 100
Norberg P, *et al.* 2001 *Monthly Notices of the Royal Astronomical Society* **328** 64
Norberg P, *et al.* 2002a *Monthly Notices of the Royal Astronomical Society* **332** 827
Norberg P, *et al.* 2002b *Monthly Notices of the Royal Astronomical Society* **336** 907
Peacock J A, *et al.* 2001 *Nature* **410** 169
Percival W J, *et al.* 2001 *Monthly Notices of the Royal Astronomical Society* **327** 1297
Percival W J, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* in press
Sadler E M, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **329** 227
Taylor K and Gray P 1990 *Proceedings of the Society of Photo-Optical Instrumentation Engineers* **1236** 290
Verde L, *et al.* 2002 *Monthly Notices of the Royal Astronomical Society* **335** 432
Willstrop R V 1987 *Monthly Notices of the Royal Astronomical Society* **229** 143
Wynne C G 1989 *Monthly Notices of the Royal Astronomical Society* **236** 47P

The 2dFGRS was carried out by a team of 30 Australian and British astronomers, led by Matthew Colless (The Australian National University), Richard Ellis (University of Cambridge and the California Institute of Technology), Steve Maddox (Universities of Cambridge and Nottingham) and John Peacock (University of Edinburgh). A full list of team members can be found on the 2dFGRS web pages.

Matthew Colless

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Colors of Stars

The colors of the stars are real but not apparent. They are not seen by the unaided eye because stars are both intrinsically faint and at the same time minute points of light, almost always observed by a more-or-less dark-adapted eye. However, the colors of the stars are an important and quantifiable observational property that is indicative of their surface temperature. This in turn is a function of a star's mass and its evolutionary status. These ideas are brought together in the well-known color magnitude diagram, a cornerstone of stellar astrophysics (figure 1).

While in general star colors are not obvious to the unaided eye, a few stars (such as Antares and Betelgeuse) appear distinctly yellow–orange. However, most of the stars that appear similarly bright to the eye, such as those that define the constellations, are in reality as blue as the daytime sky, although this is not mentioned in the long history of non-telescopic stargazing. This is because the eye is not good at distinguishing delicate shades of blue when light levels are low and when the object size is small. Even with a telescope, star colors are usually only noted by observers of double stars who sometimes comment on the contrasting colors of closely adjacent stars.

The range of color from pale blue to deep orange–yellow found among the stars is familiar to photographers as the concept of color temperature, most usually where filters are required to adjust (for example) tungsten lighting to a film designed for daylight use. More broadly they are known as blackbody colors and

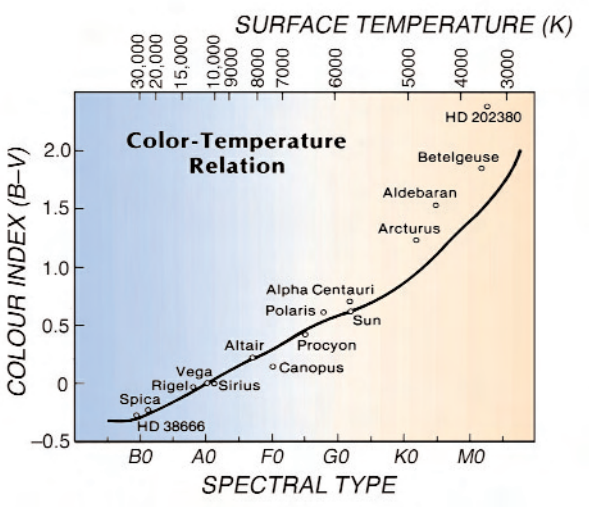


Figure 1. Most stars spend most of their lives on the main sequence, a narrow range of color and luminosity determined by the mass of the star. This diagram shows a few well-known stars dotted along the main sequence, plotted according to their surface temperature and *B*–V color index. The background color represents the color of the star

are the visible part of the radiation from any object that is at a temperature above about 1000 K (727 °C). The universe is full of such thermal radiation and much of it is invisible, originating from bodies much cooler than 1000 K. The ubiquitous microwave background radiation is a remnant of a time when the universe was unimaginably hot but now corresponds to a temperature of just 2.73 K above absolute zero.

The reality of star colors can be demonstrated in several ways. One of the simplest is to look at a bright star through a defocused telescope. This is normal practice when (for example) manually focusing a telescope using a Foucalt knife-edge, so that the star's light fills the field of view (the so-called Maxwellian view). Although I did this fairly regularly in preparation for photographic observing on the Anglo-Australian Telescope, my color memory (or maybe my color discrimination) was never good enough to distinguish reliably between randomly chosen O and B stars which are hot and bluish, but cooler F and G stars were easier to identify and differences between the yet cooler K and M stars were obvious.

A more convincing demonstration of the reality of star colors is a simple photographic experiment that involves making a star-trail photograph. From a dark site and beneath a clear night sky, a time exposure made with a camera loaded with color film captures the long, narrow trails of the stars, moved across the film by the rotation of the Earth. However, if the lens is regularly defocused in a series of steps every 5 min or so during the exposure, the stepwise spreading of the star images on the film effectively changes the exposure of each star image on the film. At some stage the colors of all the stars in the field are revealed. The example included here (figure 2) is the constellation of Orion, photographed on daylight slide film.

The realization that some stars were much bluer than others came in the 1890s when the relative magnitudes of stars in the early photographic catalogues were compared with stellar magnitudes obtained visually. It was gradually appreciated that the eye and the early photographic plates saw things differently. The plates of the time were sensitive only to blue and ultraviolet light, while the eye's maximum sensitivity was in the yellow–green part of the spectrum. This difference between the blue (*B*) photographic magnitudes and the visual (*V*) magnitudes was eventually understood as a way of measuring the colors of stars, and is the origin of the still widely used *B*–*V* color index.

Of course, the eye is no longer used for stellar photometry, so this deceptively simple system uses optical filters with well-defined characteristics to make two measurements in two well-separated color bands, nowadays using an electronic detector. Only after two-color systems were established was it realized that the difference in energy in the two passbands amounted to

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Figure 2. Step focus image of the colors of the stars in Orion, made on daylight color film by defocusing the camera lens in a series of steps during a normal star trail exposure.

measures of the slope of the blackbody radiation distribution. In the visible part of the spectrum this is seen as a color shift. Stars are not perfect blackbodies, but allowances can be made for this, and thus measuring temperatures using the colors of stars became fundamental to understanding and classifying them.

The zero point (i.e. white point) of the $B–V$ color index system currently used is defined as an A0 star, which happens to be the color index of Sirius, the brightest star in the sky. However, A0 stars have surface temperatures of 10 000 K, a temperature that we would see as quite blue if the star was as big as the Sun in the sky. Regrettably, this does not stop astronomers describing stars such as Sirius as white and stars cooler than Sirius as yellow, or even red. If there are some truly red stars they are intrinsically extremely faint. The famous bright, so-called red giant stars such as Betelgeuse and Antares are actually several hundred kelvins hotter (and therefore bluer) than most domestic tungsten filament lamps. Although light from these lamps undoubtedly has a 'warm' color, not even an astronomer would describe it as red.

David Malin

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Cosmological Constant and its Interpretation

The cosmological constant was first introduced into the equations of general relativity by Einstein himself, who later famously criticized this move as his 'greatest blunder'. His main motivation had been to allow cosmological models featuring a static universe, but this possibility swiftly became redundant with Edwin Hubble's discovery of the expansion of the universe. Despite this, it has periodically been invoked by astronomers to explain various phenomena and currently is a key part of the standard cosmological model, which seeks to explain both the evolution of large-scale universe and also the development of structures, such as galaxies, within it.

The most straightforward description of the cosmological constant is that it represents the energy of empty space. Usually physical laws only concern the *difference* in energy between two states (nature favouring evolution with force directed towards the state of lowest potential energy) rather than the absolute value of energy, but relativity demands that all energy gravitates and so the zero point of energy must be specified. The Einstein equation can be written

$$G_{\mu\nu} = g_{\mu\nu}\Lambda + 8\pi G T_{\mu\nu}$$

where $G_{\mu\nu}$ is the Einstein tensor measuring the curvature of space–time (which gives the gravitational attraction), $T_{\mu\nu}$ is the energy–momentum tensor measuring the properties of matter, $g_{\mu\nu}$ is the metric of space–time and Λ is the cosmological constant. Throughout the speed of light will be set to $c = 1$. If $\Lambda = 0$, then an absence of matter ($T_{\mu\nu} = 0$) leads to an absence of space–time curvature, but if Λ is non-zero we have gravity associated with the vacuum.

Although Einstein regretted the introduction of the cosmological constant, which certainly reduces the elegance of the equation, once it has been introduced it becomes rather hard to argue that it should not be there. There is no fundamental principle to exclude it, and indeed as we will see modern particle physics requires its consideration.

Cosmological models

Cosmological models including a cosmological constant are described in detail under COSMOLOGY: STANDARD MODEL. The Friedmann equation describes the evolution of the scale factor of the universe, $R(t)$, according to

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{k}{R^2},$$

where ρ is the density of matter, k measures the spatial curvature and the dot indicates time derivative. If preferred, the cosmological constant can be included as a form of matter by defining $\rho_\Lambda = \Lambda/8\pi G$. The effect of the cosmological constant can be seen more clearly from the acceleration equation,

$$\frac{\ddot{R}}{R} = \frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3}$$

where p is the pressure. While the density ρ always decelerates the expansion, a positive cosmological constant favors acceleration and hence acts as a repulsion. Indeed, in the limit where the cosmological constant dominates, the solution is an exponential expansion

$$R(t) \propto \exp\left(\sqrt{\Lambda/3}\,t\right).$$

Studying the way in which Λ enters these equations indicates that it can be considered as a fluid with equation of state $p_\Lambda = -\rho_\Lambda$.

There is a characteristic scale for the density, known as the 'critical density', which gives rise to a spatially-flat universe, $k = 0$. On defining the Hubble parameter, which measures the expansion rate, as $H \equiv \dot{R}/R$, the critical density is

$$\rho_c = \frac{3H^2}{8\pi G}.$$

It is often useful to consider the densities of the different components as fractions of this, by defining the density parameter $\Omega = \rho/\rho_c$ and the cosmological constant density parameter as

$$\Omega_\Lambda \equiv \frac{\rho_\Lambda}{\rho_c} = \frac{\Lambda}{3H^2}.$$

The condition for a flat universe is therefore that the sum of the Ω over all types of material is equal to unity.

In general the cosmological constant adds considerably to the possible phenomenology of big bang models. For example, there are models where the cosmic repulsion is so strong that there is no big bang, with the universe instead contracting and then bouncing. There can be open universe models which recollapse and closed universe models which expand forever; it is even possible to construct closed universe models where an observer can (if they wait long enough) see right round the universe. Another possibility is a long 'loitering' phase where the universe stays almost constant in size for a protracted period. However, the main focus has been on models which expand from an initial singularity and where the cosmological constant is negligible at early stages, only beginning to manifest itself at recent cosmological epochs.

Observational evidence for Λ

Observations in recent years have established a standard cosmological model where the present density of the universe is dominated by a cosmological constant. It is fair to say that cosmologists have been reluctant to include the cosmological constant but have been forced into it by the quality of a variety of observations, which paint a consistent picture depending on its presence. The main supporting evidence falls into two classes. One is

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

direct probing of the expansion rate of the universe by measuring the properties of distant objects, in this case supernovae. The second is to determine the density of matter in the universe and compare with the total density as inferred from the geometry of the universe.

Distant supernovae of type I, believed to be fueled by the collapse of an accreting white dwarf to a neutron star, are an excellent probe of the expansion rate of the universe because their absolute brightness is believed to be nearly the same for each explosion, with empirical corrections available to improve the uniformity of the sample even more. This means that they can be used to study the dependence of the luminosity distance (see COSMOLOGY: STANDARD MODEL) on redshift. This relation is known as the HUBBLE DIAGRAM, with supernovae now able to map this out to redshifts approaching unity. This relationship can be related to the expansion history of the universe $R(t)$, which the Friedmann equation shows is sensitive to the presence of a cosmological constant. Constraints on the densities of the cosmological constant and of matter are shown in figure 1 as the full and broken contours, and almost all the preferred region has a positive cosmological constant (and indeed a currently accelerating universe).

The second observational route concerns the matter budget of the universe. Recent cosmic microwave background anisotropy measurements have given strong evidence that the universe is spatially flat, or at least very nearly so (also shown in figure 1 as the dark shaded region). This requires that the total density of the universe, including the cosmological constant, must add to the critical density. However, there is plentiful evidence that the density of matter is well below this. For example, studies of clusters of galaxies show that there is about eight times as much dark matter as protons and neutrons, the latter being constrained by the formation of light elements to be about 4% of the critical density. Alternatively, the observed clustering properties of galaxies are known to be reproduced only if the dark matter density is around 30% of critical. Finally, direct estimates of the amount of matter providing gravitational attraction always fall well below the critical density. There is therefore a substantial shortfall in reaching the critical density required by the cosmic microwave background observations, a gap which must be filled by a form of material which does not clump gravitationally. The cosmological constant fits the bill perfectly.

Accordingly, the currently favoured cosmological model features a cosmological constant at around two-thirds the critical density, with matter making up the remaining third. The latter is mostly dark matter, but includes also the conventional matter from which we are made as well as the cosmic radiation backgrounds.

In summary, current observations strongly favor cosmological models with a flat spatial geometry ($k = 0$)

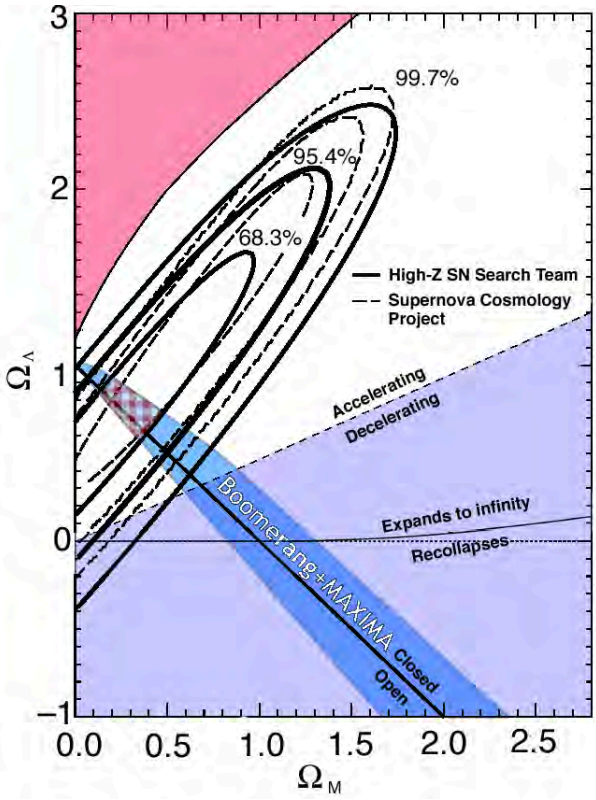


Figure 1. Constraints on cosmology from supernovae and the microwave background. The horizontal axis is the density of matter (including dark matter) as a fraction of the critical density, while the vertical axis is the density of the dark energy. The contour lines show the favored region from two separate studies of distant supernovae. The shaded diagonal band shows the region allowed by measurements of the anisotropies of the cosmic microwave background by the Boomerang and Maxima experiments. In combination, these allow only the small checked region, where the cosmological constant density is about twice that of the matter. (Figure courtesy Brian Schmidt; see also http: / / cfa-www.harvard.edu / cfa / oir / Research / supernova / HighZ.html.)

and which are accelerating at present. The simplest interpretation of the observed acceleration is that the universe possesses a cosmological constant.

The cosmological constant in particle physics

Explaining the observed magnitude of the cosmological constant is regarded as one of the outstanding problems in particle physics. This is because particle physics theory indicates that the magnitude of the cosmological constant should be huge, vastly greater than is observed. The reason is that in quantum physics one expects a zero-point energy to be associated with every kind of particle, and therefore that there should be an energy density associated with the vacuum.

It is not hard to see that the vacuum energy density is equivalent to a cosmological constant. If all observers

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

are to agree that a given state is the vacuum, then its energy–momentum tensor must be Lorentz invariant. The only tensor with this property is a multiple of the metric tensor $g_{\mu\nu}$ and hence we can write

$$T_{\mu\nu}^{\text{vacuum}} = -\frac{\Lambda}{8\pi G} g_{\mu\nu},$$

which ensures that the vacuum energy density does act precisely as a cosmological constant.

To assess the magnitude of the problem, we should first convert the observed value of the cosmological constant into particle physics units. Its equivalent mass density $\rho_\Lambda = \Lambda/8\pi G$ can be taken to be about two-thirds the critical density, giving

$$\rho_\Lambda \simeq 6 \times 10^{-27}\ \text{kg m}^{-3} \Rightarrow \rho_\Lambda \simeq (0.002\text{eV})^4.$$

This energy scale, written in electronvolts, is tiny compared with known characteristic energy scales of particle physics (for example the proton mass–energy is around 10^9 eV).

In quantum mechanics, unlike in classical physics, the vacuum state is a seething state where particles are constantly coming in and out of existence. There is a zero-point energy (the analog of the residual energy left in a gas at a temperature of absolute zero) associated with each possible particle state, which can be added up to give an estimate of the zero-point energy. Unfortunately the most simplistic calculations give an infinite answer, as each momentum state gives an equal contribution and there are an infinity of them; this indicates that the theory used to compute the contributions cannot be valid to infinite energies.

However, to what energies should the calculation be believed? According to present understanding, we are limited by the Planck energy, at which point gravity would have to be described by a quantum theory. The cosmological constant should therefore be approximately equal to the Planck energy density. Unfortunately, this is a massive $(10^{28}\ \text{eV})^4$, larger than the observed value of the cosmological constant by a factor of around 10^{120} (earning this calculation the description as 'the worst order-of-magnitude estimate in history'). This situation can be improved a little by conceding that actually we do not really understand physical laws above the energy scales probed by large particle accelerators such as CERN, and indeed a popular theory known as supersymmetry predicts that at higher energies a new theory comes into play that would ensure no contributions to the cosmological constant from energies much above those currently probed. Particle accelerators operate up to energies of around 10^{12} eV, so plausibly we can reduce our estimate to $(10^{12}\ \text{eV})^4$, reducing the embarrassment to a mere factor of 10^{60}!

It is currently believed that superstring theory (or its more elaborate cousin M-theory) is a strong candidate as a theory of quantum gravity. In principle it should therefore make much more definite predictions for the cosmological constant, as it is not subject to the cut-offs described above. However, current incarnations of superstring theory have not been developed sufficiently to allow unambiguous predictions to be made at this stage.

The historical cosmological constant problem was to explain why the cosmological constant appeared to be zero, despite the prediction that it should be huge, and investigations focused on various cancellation mechanisms and symmetry principles which might justify why the precise value of zero might be favored. However, the astronomical observations just described have, if anything, made the problem worse, as they demand an explanation for why the cosmological constant is both tiny *and* non-zero. No compelling, or even just plausible, explanation has yet arisen.

Dark energy/quintessence

While a cosmological constant can indeed give the observed accelerated expansion of the universe, it is not the only explanation. Since the early 1980s, it has been popular to suppose that the universe underwent a period of acceleration during its extremely early stages, known as INFLATION. Such a period of inflation cannot continue forever, because that would spoil the outstanding successes of the hot big bang model, such as nucleosynthesis and the origin of the cosmic microwave background. Early universe inflation therefore cannot be driven by a pure cosmological constant, but rather must be caused by a transient phenomenon mimicking a cosmological constant.

Much effort has gone into modeling inflationary expansion, and the standard mechanism is to assume that the universe is temporarily dominated by the potential energy of a scalar field. A scalar field describes a collection of spin-zero particles (which are popularly used in particle physics models to describe the breaking of symmetries), with the potential energy measuring the combination of the mass density of the particles and their binding energy. Under the right circumstances, such a scalar field can drive a temporary period of accelerated expansion, which comes to an end with the scalar particles decaying into conventional matter (see INFLATION for full details).

These ideas are an equally valid mechanism for driving acceleration in the present universe and provide an alternative to a pure cosmological constant that many cosmologists find attractive. The term 'dark energy' is now widely used to indicate any method of driving accelerated expansion, encompassing both the pure cosmological constant and any other method of obtaining an acceleration. This phrase is on no account to be confused with dark matter (see DARK MATTER: ITS

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

NATURE), which has entirely different properties. The specific use of a scalar field to drive the acceleration is usually referred to as 'quintessence' (the other four 'essences' of the universe being the baryons, the photons, the neutrinos and the dark matter). In the case of quintessence, it is presumed that any pure cosmological constant must be identically zero, for some as-yet-undiscovered reason.

This more general setting allows the possibility that the density of the dark energy may not be strictly constant, as in the case of the cosmological constant. This can be quantified via the equation of state parameter w_Q, defined by

$$p_Q = w_Q \rho_Q$$

where the subscript Q indicates that we refer to the density and pressure of the dark energy component. A pure cosmological constant has $w_Q = -1$, whereas to drive an accelerated expansion requires only $w_{\rm tot} < -1/3$, where $w_{\rm tot}$ is the combined equation of state of the dark energy and the other matter components (including dark matter) weighted by their densities. With observations indicating roughly twice as much dark energy as other matter, accelerated expansion will result provided that w_Q is less than about –0.5.

The quintessence energy density appears in the Friedmann equation, which for a flat universe reads

$$\left[\frac{\dot{R}}{R}\right]^2 = \frac{8\pi G}{3}(\rho + \rho_Q),$$

where we will have

$$\rho_Q \propto R^{-3(1+w_Q)}.$$

If the density of dark energy is varying, the solution for the scale factor $R(t)$ will be different to that obtained with a pure cosmological constant, and this provides possible observational signatures. It modifies the relation between luminosity distance and redshift, as probed by distant supernovae, and instead of using those data to constraint the densities of a cosmological constant and conventional matter, one can assume a flat universe and attempt to simultaneously constrain the density of dark energy and its equation of state. Another effect is that the altered expansion rate will change the growth rate of density perturbations in the dark matter, modifying the structure formation process. Finally, the angular diameter distance to the last scattering of the cosmic microwave background will be modified, leading to a shift in the angular size of features (the positions of the acoustic peaks).

Several studies of these effects have been made, and as yet no evidence has been found for variation in the density of dark energy. The pure cosmological constant remains a viable explanation of the observational data, and observations require that the equation of state w_Q be less than about –0.7 (at 95% confidence). Upcoming observations, perhaps including a satellite-based supernova search, have the capacity to significantly strengthen this constraint and hence may detect a time variation of the dark energy if it is sufficiently large.

If the dark energy is permitted to have time variation, it must also be able to support spatial variations, in the form of density perturbations. In principle these can interact with and affect perturbations in the dark matter, modifying the process of structure formation which may lead to further observational tests. However in practice these effects are highly suppressed in realistic models and it is likely to be very hard to determine the properties of dark energy through studying these effects.

If we appeal to the same physical mechanism both to drive the inflationary acceleration of the early universe and the present-day acceleration, does that mean that the same scalar field might be responsible for both? In most models the answer is no, because of the vastly different energy scales involved; we are using the same physical mechanism, but the two fields are not connected. However, it is possible to contrive models where the same field is responsible; Peebles and Vilenkin called such models 'quintessential inflation'. As the same scalar field has multiple duties to attend to, it has many more observational requirements that it must satisfy and so such scenarios can be quite strongly constrained. However, fundamental particle physics indicates we should expect many scalar fields to exist in nature, so there seems little motivation to force one of them into doing all the cosmological work. Accordingly, most attention has been directed at models where early universe inflation and the present dark energy are separate phenomena.

The cosmological constant coincidence

In addition to the daunting task of understanding the numerical value of the cosmological constant, it poses a further problem known as the 'coincidence problem'. It is an apparent coincidence that we are developing our understanding of cosmology during the fairly brief epoch (brief by cosmological standards—a few billions of years in actuality) where the cosmological constant has a density similar to that of other matter. The ratio of densities of the cosmological constant to dark matter goes as the cube of the scale factor; when the universe was a tenth its present size the cosmological constant was utterly negligible, while once the universe has doubled its present age it will be completely dominant.

One of the advantages of the dark energy hypothesis is that, in enabling the density of the effective cosmological constant to vary, it may open routes to explaining this coincidence. Unfortunately, however, no compelling model has yet arisen, despite considerable interest in so-called 'tracker models' where a quintessence field responds to, and in some cases mimics, the behavior of conventional matter.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

An alternative route is to invoke the ANTHROPIC PRINCIPLE, which states that we should expect to see cosmological parameters take on values which give the universe which, out of all possible universes, is most amenable to the development of life. This approach can in principle explain both the numerical value and the coincidence, as a very large value of the cosmological constant will prevent galaxies from forming, while there is some indication that a significant value may enhance the number of galaxies produced. In this approach, we require fundamental physics to allow a wide range of possible values of the cosmological constant (which may be realized in widely separated regions of the universe, or even perhaps just in possible quantum states of the universe), and the precise value which arises in our part of the universe has no fundamental significance.

Bibliography

Carroll S 2001 The cosmological constant *Living Reviews in Relativity* webpage http://www.livingreviews.org

Carroll S, Press W H and Turner E L 1992 *Annual Reviews of Astronomy and Astrophysics* **30** 499

Goldsmith D 2000 *The Runaway Universe* (Perseus)

Kirshner R P 2002 *The Extravagant Universe* (Princeton Publishing Company)

Peacock J A 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

Perlmutter S *et al.* 1999 *Astrophysical Journal* **517** 565

Riess A G *et al.* 1998 *Astronomical Journal* **116** 1009

Weinberg S 1989 *Reviews of Modern Physics* **61** 1

Andrew Liddle

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Hobby–Eberly Telescope

The Hobby–Eberly Telescope (HET) has an 11 m diameter spherical primary mirror. It is one of the world's largest optical telescopes. It is located at McDonald Observatory in West Texas and its partners include the University of Texas at Austin, Pennsylvania State University, Stanford University, the Ludwig Maximillians University in Munich and the Georg-August University in Göttingen. Observing time is divided among the partners in proportion to each partner's financial contribution.

Because of the spherical primary mirror, the telescope is equipped with a spherical aberration corrector which sees a 9.2 m diameter piece of the primary mirror. During an observation, in order to compensate for the rotation of the Earth, the primary mirror is held fixed while the spherical aberration corrector is moved across the focal surface of the sphere in order to follow the reflected image of the star. The primary mirror is tilted with its optical axis 35° from the zenith and is mounted on a turntable allowing it to be rotated 360° in azimuth. This novel design causes the gravity sag of the primary mirror to be constant but limits the telescope to observe objects which are crossing a circle, centered on the zenith, whose radius is 35°.

The HET is designed primarily to do astronomical spectroscopy. Its initial complement of instruments includes a low-resolution spectrograph at the top of the telescope into which the light is shone directly and a medium-resolution and high-resolution spectrograph in the basement, under the telescope, which are fed by optical fibres carrying the light from the top of the telescope down to the instruments. The telescope is queue scheduled which allows it to observe targets of opportunity, such as supernovae, and to make regular observations of variable objects. Queue scheduling also allows more efficient use of the telescope and it allows one to adapt the schedule to the sky conditions.

Visitors to the McDonald Observatory are welcome to visit the George T. Abell Visitors Gallery attached to the HET. Visitors are welcome seven days a week and a new visitor centre will open late in 2001.

Frank Bash

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS21 6BE, UK

The Telescope (to 1950)

The TELESCOPE was the single greatest technical innovation in the history of astronomy, transforming it from a naked-eye study of celestial motions into a physical science. It revealed to humans a far richer and more complex universe than ever imagined. Its history proceeded in stages: a technological discovery led to a series of improvements until an impasse was reached. Another discovery would entail a different line of development, resulting in larger, better instruments. Some of these stages overlapped. Astronomical discoveries came in waves, as each innovation extended observers' views. Three important needs have driven telescope makers and users: greater light-gathering power with attendant resolution; sharper, brighter images; sturdier mountings.

From the astronomer's viewpoint, the choice of instrument depended on the object of research or the area of research was driven by the telescopes available. During the 18th and 19th centuries, the concentration on stellar positions, binary stars and visual planetary observations encouraged the development of REFRACTORS. The increasing importance of photography in the last quarter of the 19th century stimulated new optical designs and better mountings and drive systems. The 20th-century demand for more light-gathering power for spectroscopy and photography spurred the development of large REFLECTORS.

The early telescope

Invented in the Netherlands in the first decade of the 17th century—there were several claimants—the telescope became an astronomical instrument in the hands of GALILEO GALILEI. He fabricated his first telescope, essentially a spyglass, in 1609. His observations of the Moon, Jupiter's satellite system, Venus and the Milky Way, announced in the *Sidereus Nuncius* the following year, ensured the instrument's rapid adoption by astronomers. Galileo thus inaugurated the development of the REFRACTING TELESCOPE, one that employs an objective lens to capture and focus light, and an eyepiece of one or more lenses to straighten the light into a beam and to magnify the image. The GALILEAN TELESCOPE or Dutch telescope combined a double-convex objective lens and a double-concave eyepiece to produce an upright image, a limited field of vision and low magnification. Such instruments, many fabricated by Galileo, spread quickly, with sunspot observations being made by mid-1610. Such a design, however, severely limited observational discoveries.

In 1611, KEPLER, in his *Dioptrice*, suggested an alternative arrangement for a refractor, employing two double-convex lenses, which would yield a wider field and greater magnification, but an inverted image. He produced none himself, but, by the late 1640s, Francesco Fontana was observing with a Keplerian telescope of his own manufacture. Kepler also described spherical aberration. Given a lens with a spherical surface, rays of light incident near the outer edge will focus closer to the lens, while those incident near the center will focus slightly farther out, the result being an image that is out of focus. He reasoned that a hyperboloidal surface would solve the problem, although no optician could have ground such a surface. A second major defect of a refracting telescope is CHROMATIC ABERRATION. As different wavelengths of light are refracted at slightly different angles, the observer sees an image with colored rings. The two defects compound, giving blurry, colored images.

The disadvantages of the Keplerian design could be overcome somewhat by producing objective lenses with less curvature, which had longer focal lengths. Alternatively, one could stop down the objective, so light passed only through the central portion of the lens. The earliest instruments of the Keplerian type, like the Galilean form, were wooden tubes with lenses fixed in the ends. JOHANNES HEVELIUS, a Danzig brewer, was the first systematic observer with long-focal-length refractors. With a 12 ft instrument, he produced a lunar atlas in 1647. During the next decade, Christiaan and Constantijn Huygens, in the Netherlands, produced even longer refractors, turning them on the Saturn system. This stimulated Hevelius to build larger instruments, culminating in a 150 ft telescope, with an open, wooden framework tube suspended by ropes and pulleys. Flexure and vibration made such instruments extremely difficult to manipulate. The Huygens' solution was to dispense with tubes altogether, creating 'aerial' telescopes—one of which was more than 200 ft long—by joining the objective lens to the eyepiece by a thread. Both forms were clumsy, yet a surprising number of discoveries came from their use. An important center was the Paris Observatory, where J D Cassini made planetary observations with both forms, employing long-focus lenses provided by the Roman optician Giuseppe Campani. While the optical quality of many lenses, particularly from Italian opticians in the late 17th century, was good, mountings were poor. Shorter instruments could be propped up on windowsills or wooden frameworks. Almost all telescopes were employed out of doors.

Despite its drawbacks, the Keplerian refractor could be adapted to telescopic measurement. A number of astronomers, beginning in the 1650s, experimented with micrometers, which allowed the angular distance between two objects in view to be measured. As Robert Hooke showed in the late 17th century, telescopes could be mounted upon a variety of hitherto naked-eye instruments such as the quadrant, greatly improving the accuracy of position determination.

The reflecting telescope

Two technological pathways were followed in an attempt to make telescopes more effective. One was the development of a different optical system, the REFLECTING TELESCOPE; the second was to find a solution to the chromatic aberration problem in refractors. Although the idea of reflecting telescopes or hybrids using both mirrors and lenses had been discussed since Galileo's time, the first description of a modern reflecting instrument was made

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

by James Gregory in 1663. His suggestion of combining a perforated parabolic primary with a concave ellipsoidal secondary to reflect the image back to an eyepiece was beyond the opticians' art. His contemporary, Isaac Newton, investigating chromatic aberration, correctly linked dispersion with the refractive properties of a transparent medium. Assuming it could not be overcome, he produced the first workable reflector in 1672. In the familiar Newtonian form, its small primary mirror of speculum metal, an alloy of copper and tin with some arsenic, had a spherical surface. A small, flat mirror, set at an angle of 45°, diverted the light reflected from the primary to an eyepiece mounted on the side of the tube. However, the NEWTONIAN TELESCOPE suffers from SPHERICAL ABERRATION. At the same time, Cassegrain in France proposed a system like Gregory's but employing a convex secondary (see CASSEGRAIN TELESCOPE, GREGORIAN TELESCOPE), although no examples seem to have been made at the time.

By the 1720s, JOHN HADLEY, the inventor of the octant, solved the spherical aberration problem by producing mirrors with a parabolic figure. Within 20 yr, London optician James Short could provide sufficiently good primaries to construct Gregorian reflectors as a commercial enterprise. It was Hadley who designed a workable altazimuth mounting for reflectors, followed by Short. The latter also produced some simple equatorial mountings for his brass-tube instruments. As a research instrument, however, the reflector came into its own with William Herschel's work.

Herschel, a self-taught astronomer and optician, began casting and polishing his own mirrors in the early 1770s (see HERSCHEL FAMILY). Over the next decade, he produced increasingly larger mirrors, building Newtonian instruments with a variety of mountings. By the mid-1780s, he undertook his important 'sweeps' to discover nebulae and undertook star gauging with a 20 ft telescope with an 18.8 in primary. For convenience and for brightness of image, he dispensed with the Newtonian flat, tilting the primary mirror so the focus reached the rim of the tube's upper end, where an eyepiece could be laid. The scaffold-work altazimuth mounting allowed the tube to be slewed 30° horizontally without moving the base; vertical motion was effected by ropes and pulleys. By 1789, Herschel erected by far the largest telescope ever built, the 40 ft telescope with a 48 in primary. Similar to the 20 ft, it was more massive and difficult to operate and never much used.

By the turn of the 19th century, British manufacturers supplied a variety of smaller reflectors—Gregorians, Newtonians and Cassegrainians—but none followed Herschel's lead in fabricating large instruments. By the 1830s, John Herschel was still using his father's 20 ft instrument to great advantage at the Cape of Good Hope (see HERSCHEL FAMILY). Stimulated by Herschel's example, William Parsons, THIRD EARL OF ROSSE, began experimenting with casting and polishing large specula in the 1840s. This was a difficult and risky undertaking. Rosse assembled, and trained, a team of technicians who assisted in producing a variety of specula, some compound, and building innovative polishing machines along with casting and annealing equipment. Like Herschel, Rosse's finest instrument was not his largest but a 36 in Newtonian built in 1840. It performed well despite its poor site. Striving further, and with many technical issues to resolve, Rosse soon constructed his 72 in 'Leviathan'. No larger reflector would go into service until 1918, but it was an almost useless instrument, despite its role in showing the spiral structure of several nebulae. Mounted as a transit instrument between two stone walls, with a very small lateral motion, it was useful only for lunar or planetary work, and then only rarely at full aperture given the miserable observing conditions of the Irish midlands. Further discoveries would rest with astronomers armed with more modest equipment.

Development of the modern refracting telescope

The second line of development to improve telescopes resulted from experiments to deal with chromatic aberration in refractors. In 1729, Chester Moor Hall found that, by combining a convex crown glass lens with a concave flint glass lens, he could produce an achromatic objective. He was no optician and not until the 1750s did John Dollond begin producing such lenses commercially. The Dollond firm and, later, Jesse Ramsden, manufactured a wide variety of instruments with achromats. Although a real improvement over anachromatic lenses, the poor quality of flint glass limited their value. Still, small, specialized instruments were readily available from reliable makers such as Troughton, Tulley and Simms in England, and from several continental European firms such as the Repsolds of Hanover.

Until the mid 19th century, refracting telescopes were strictly visual instruments, used either on their own or as part of specialized instruments such as sextants, octants, transits and meridian telescopes for navigation, celestial mapping and timekeeping. Tubes came in a variety of materials, mostly wood and metal, with brass becoming dominant until the late 19th century, when steel supplanted it.

The refractor was brought to near perfection during the 19th century through a combination of improved optical glass technology and accurate, sturdy equatorial mountings. Achromatic objective lenses were limited by the quality of flint glass. Swiss optician P L Guinand developed a method in 1805 that produced flint glass with almost no imperfections. During a business venture in Munich, he tutored the young JOSEPH FRAUNHOFER in his method. Fraunhofer's later experiments with spectra allowed him to calculate lens curvatures and combinations that provided superior (although not perfect) images. Fraunhofer's other key contribution was the prototype of the modern German equatorial mounting. For his 9.5 in Dorpat refractor (1824), a steel polar axis was mounted on a wooden frame; the declination axis, attached to the top of the polar axis, held the telescope tube on one side,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the declination circle on the other; the counterweight was bolted to a separate arm. The polar axis was turned by a weight-driven clock with a centrifugal friction governor. With this instrument, then the world's largest refractor, F G W Struve undertook his remarkable work on binary stars.

With the introduction of the uniform driving of the polar axis by means of weight-driven clockwork, guiding became limited to minor adjustments during observation. Without clock-driven equatorial mountings, the introduction of photography and spectrography in the second half of the century would have been impossible.

By 1839, Georg Merz, a former assistant of Fraunhofer's, commenced operations of Merz and Mahler, the major German telescope firm until the 1880s. With high-quality optical glass, Merz provided the 15 in Pulkova refractor in 1839. Like the Dorpat instrument, it was used by Struve's son Otto for binary star measurements. Merz also supplied the Americans with their first large refractors: the 11 in Cincinnati telescope in 1843 and the 15 in Harvard College Observatory telescope for W C Bond in 1847. German manufacturers maintained a leading position with the association of two Jena firms, instrument maker CARL ZEISS and optician Otto Schott, after 1884. Zeiss did not produce large refractors until the 20th century, with 24 and 26 in instruments for Germany, Serbia, Indonesia and Japan in the 1910s and 1920s.

British manufacturers in the early 19th century could only produce modest refractors, having no source of good optical glass except for Guinand, by then based in France. Chance Brothers of Birmingham purchased the method from Guinand's successor in 1837. They, and Guinand's son and grandson in France, produced most of the world's objective lens blanks during the rest of the century. Better-quality glass and improvements in grinding and polishing machines and techniques were factors in the production of superior lenses. Another was the knife-edge method of testing optical performance, announced by Foucault in 1858. One could measure the focal length and find surface curvature imperfections of both lenses and mirrors. In the early 20th century, the procedure was supplemented by Hartmann's photographic testing method.

Although Americans had to obtain telescopes in Europe before 1850, their own opticians and mechanics quickly became world leaders in refractor technology. Henry Fitz produced many medium-sized refractors of excellent quality—his largest was 16 in—before the Civil War. In the late 1850s, ALVAN CLARK, who had no training in optical work, began figuring achromats. Using a method developed independently of Foucault's, Clark was able to find surface imperfections. Careful figuring, rather than calculating curvature, enabled him to produce superb lenses. His 18.5 in objective, intended for the University of Mississippi, was completed in 1862. Testing the lens, Clark's son, Alvan Graham Clark, discovered Sirius B. Because of the war, the telescope ended up in the old University of Chicago and, later, at Northwestern University. The Clarks soon produced 26 in refractors for the US Naval Observatory and the University of Virginia. More instruments were to follow, each in turn the world's largest: the Pulkova 30 in in 1883, the 36 in for the Lick Observatory in 1887 and the 40 in for the Yerkes Observatory in 1897. The objectives for last two were the work of A G Clark and C A R Lundin. With the completion of the Yerkes instrument, no larger refractor was ever constructed.

The Clarks, when building the Lick instrument, contracted the tube and mounting to a Cleveland machine-tool manufacturing firm, Warner and Swasey, which had only limited experience in the field. With that success, they repeated their partnership for the Yerkes' 40 in telescope. Warner and Swasey worked with both the Clark and the Pittsburgh firm of John Brashear on further refractor projects before turning to large reflectors.

European telescope makers were not idle during the second half of the century. Thomas Cooke was the leading English manufacturer of refractors by mid-century and the only one to produce a large one, the 25 in instrument for R S Newall, completed in 1871. In France, Paul and Prosper Henry, working with P Gautier, built several French refractors. In Germany, the Steinheil firm collaborated with the Repsolds on several projects. The Dublin instrument maker, Howard Grubb, produced several 15 in telescopes in the 1870s before producing his trademark instrument, the 27 in Vienna University refractor 1881, which featured a German equatorial mounting that became the model for all subsequent instruments. Grubb's works produced another half-dozen large instruments into the 1920s when the firm, purchased by Sir Charles Parsons (becoming Grubb-Parsons) moved to manufacturing reflectors.

Until the 1860s, astronomical photography, born in 1840, was practiced with visual refractors, which were not corrected for the violet end of the spectrum, to which photographic plates of the time were most sensitive. American amateur L M Rutherfurd produced special objectives for photographic work in the 1860s. By the 1890s, several firms—Grubb, T Cooke and Sons, Brashear and Clark—produced a number of different astrographic lenses for specialized instruments. If both visual and photographic work were wanted, two telescopes could be piggy-backed on one mounting. Alternatively, corrector lenses could be introduced to visual refractors for photographic work, a method with several drawbacks.

Other specialized refractors aided positional astronomy. In the late 17th century, Ole Rømer devised the first transit instrument, mounted on a horizontal axis so it moved only in the meridian. A partially circular scale, with a micrometer for more accurate readings, allowed the star's declination to be determined. With the better clocks by then available, star passages could be timed to determine their right ascensions. He also built the first meridian circle, which performed both tasks, but had a full-circle scale and could reach the entire visible meridian. These and variants became features of many observatories over the next two centuries. During the 19th century, British

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

manufacturers such as Troughton and Simms and Thomas Cooke produced portable transit instruments that could be taken into the field, which, with the use of the electric telegraph, allowed longitudes to be determined far from observatories.

We need to note also the evolution of an indispensable part of a telescope, the EYEPIECE. The basic Galilean telescope combined a single objective lens with a single biconcave, or negative, eyepiece lens. The Keplerian form also employed a single eyepiece lens—in this case a biconvex or positive lens. Both magnified the image and collimated the rays of light received by the eyepiece from the objective. As magnification is the ratio of the focal lengths of the objective and eyepiece, long-focal-length objective lenses could be paired with short-focal-length eyepieces for much higher magnifications in the Keplerian form. Single-lens eyepieces continued to be used for more than two centuries—William Herschel preferred them—but compound eyepieces appeared in the 1640s in Italy. HUYGENS developed the first standard compound system, comprising two planoconvex lenses (a field lens and an eye lens) with both curved surfaces facing the objective. A Huygens eyepiece offered a wider field of view. During the 18th century, Jesse Ramsden devised a variant using two planoconvex lenses of equal focal length in which the curved sides faced each other.

Two other general-purpose systems appeared by the mid 19th century. The Kellner or orthoscopic eyepiece combined a compound achromatic eye lens with a convex field lens to provide a wide field and nearly constant magnification over the field. The Steinheil and Zeiss firms produced a number of Kellner-type eyepieces. Another way to increase magnification was to introduce a concave lens in front of the eyepiece, an idea developed by Barlow in England. Such eyepieces, in a wide variety of magnifications and fields of view, were available to astronomers by the later 19th century. In practice, a high-powered eyepiece was useful only for lunar or planetary work, and then only under the best atmospheric conditions and with the sturdiest mounting.

The modern reflecting telescope

The later 19th century was the heyday of the large equatorial refractor, but the design and construction of the reflector was undergoing a quiet revolution. Lord Rosse's Leviathan was a dead-end, given its poor site, easily tarnished mirror and inflexible mounting. The last two examples of large speculum reflectors were WILLIAM LASSELL's 48 in Malta instrument (1860) and THOMAS GRUBB's 48 in Melbourne reflector (1869). Lord Rosse had followed Grubb's lead in supporting his mirrors, and Lassell emulated Rosse's methods for grinding and polishing, but, with the failure of the Melbourne instrument, the history of speculum reflectors ended.

A key breakthrough was the introduction of silvered glass primary mirrors by Foucault and others in the 1850s. When such a mirror tarnished, it need only be re-silvered, as opposed to the time-consuming and potentially ruinous refiguring required for speculum mirrors. HENRY DRAPER, a pioneer of astronomical photography, introduced the idea to America. English optician ANDREW COMMON produced a number of medium-sized crown glass mirrors, which found homes in several private and public observatories. His 36 in equatorial reflector, later donated by E Crossley to the Lick Observatory, showed the value of such instruments for photography. After C D Perrine built a new mounting early in the 20th century, the telescope became an outstanding research tool. By the late 1880s, Common showed that 60 in glass mirrors, both solid and with a hole in the center, could be made.

Common produced a number of primaries for observatories but was soon joined by competitors: the Henry Brothers in France, Brashear and the Clarks in the USA, Schaer in Switzerland and Zeiss in Germany. Some firms produced both the optical and mechanical parts for reflectors. An effective reflecting telescope required two other important features: a solid EQUATORIAL MOUNTING and a means to support the great weight of large mirrors. While refractors typically had German equatorial mountings, the reflectors from the mid 19th century featured a variety of arrangements. The cross-axis form, originated by Sisson in the 18th century, was exemplified by Grubb's Melbourne telescope and French telescope mountings designed by Gautier. Common used a fork mounting for his 60 in instrument, and others tried the English yoke mounting, improved by Ramsden.

Solving the flexure problem of large mirrors took more time. Lord Rosse, Grubb and Common had used counterweights on the back of the mirror to compensate for any sag when the mirror moved vertically. The standard solution for 20th-century instruments was G W RITCHEY's method used for the 60 in Mount Wilson reflector. The mirror was housed in a massive, cast-iron cell, with three adjustable pads for mirror collimation, a set of counterweighted pads attached to the back for flexure and a ring with counterweights to handle lateral movements.

Another factor in improved astronomical work was the realization that a large telescope necessitated an excellent site for the best seeing. Lassell, who worked in Malta, was the first to demonstrate this. Even in the 20th century, many of the major, especially national, observatories such as those in Greenwich, Paris, Pulkova, Washington or Ottawa had light-polluted, smoky and climatologically poor locations. Beginning with the Lick Observatory, astronomers began seeking the best possible sites when they could manage to do so economically or politically. Although George Ellery Hale had chosen a mediocre Wisconsin site for the Yerkes Observatory, he recognized the value of a good site and quickly developed Mount Wilson in California.

Although glass-on-silver reflectors of a modest size (<1 m) proliferated towards the end of the 19th century, it was Ritchey's fork-mounted 60 in telescope at Mount Wilson that demonstrated in 1908 that a powerful, accurate instrument for astronomical photography and spectrography could be built. Common had produced

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

equally large mirrors earlier but did not solve the mounting problem to build practical instruments. As a major focus of astronomical work after the turn of the century was stellar spectroscopy, the Cassegrain arrangement was considered most convenient to have ready access to the spectrograph. One could either cut out the hole in a primary or pour the molten glass into a mold with a core. Ritchey did not believe optical firms could produce blanks with holes without serious flaws and followed JAMES NASMYTH's method of introducing a third mirror just in front of the primary to deflect the light beam to the side of the tube. J S PLASKETT, in planning the 72 in reflector for Victoria, suggested that the St-Gobain works should force a core through the molten glass and a mirror with few imperfections resulted. Almost all subsequent large reflectors had holes cast in them.

Stimulated by Ritchey's example, Plaskett designed what became the standard model for large reflectors to the 1960s. He chose the cross-axis arrangement for ease of use and adopted Ritchey's method for mirror mounting and support. Because spectroscopy was the most important use for such an instrument, it had $f/5$ primary and $f/18$ Cassegrainian focal ratios. The 72 in primary of plate glass, provided by St-Gobain, was ground and polished by Brashear. Mechanical parts were supplied by Warner and Swasey. The telescope featured self-aligning ball bearings in the polar axis, electrical control and a quick-change method for substituting prime focus, Newtonian and Cassegrain arrangements in the upper tube.

Before 1950, three firms produced large reflectors, all using features based on the Dominion Astrophysical Observatory instrument. Warner and Swasey build a 69 in instrument for the Perkins Observatory, a 60 in for the Cordoba Observatory and its largest telescope, the 82 in McDonald Observatory telescope. The last, at Otto Struve's insistence, included a coudé arrangement for spectroscopy. For all Warner and Swasey telescopes, engineering design was provided by E P Burrell, later involved with the 200 in telescope project. J W Fecker, who once worked for Warner and Swasey when the Canadian telescope was being built, later provided two 60 in telescopes for Harvard, while completing the optics for the Cordoba instrument. The third firm was Grubb-Parsons, successors to Sir Howard Grubb. Its first large reflector was the 74 in telescope for Toronto's David Dunlap Observatory in 1935. It too was modelled on Plaskett's instrument. With only slight engineering changes—such as the provision for coudé work—Grubb-Parsons produced almost identical telescopes for Pretoria, Helwan and Canberra.

The large reflectors were the workhorses of astrophysical research during the first half of the century, most of their time being devoted to spectroscopy. A second evolutionary line of very large reflectors developed slowly from the pioneering work of Ritchey. While the 60 in telescope was under construction, G E HALE wished to secure an even larger mirror, in the event a 100 in blank of poor quality from St-Gobain. Ritchey worked on the mirror for several years, while he and his colleagues designed a massive yoke mounting. It went into regular service at Mount Wilson in 1918 and, for 30 yr, was the world's largest telescope. In the hands of Edwin Hubble and others, it became the front-line instrument for extragalactic studies. Its mounting was not emulated by manufacturers of conventional large reflectors.

Hale soon desired an even larger telescope and with his customary enthusiasm persuaded the Rockefeller Foundation to fund a 200 in reflector in 1928. It was clear that one could not merely scale up the 100 in instrument and new engineering and material problems required solution. The crown glass mirror, as used at Mount Wilson and Victoria, had too high a coefficient of expansion to be useful for such a large instrument. The US Bureau of Standards had been able to provide a superior borosilicate glass primary for the Perkins telescope. A commercial borosilicate glass emerged in Pyrex, manufactured by Corning Glass Works. In the early 1930s, while Corning experimented to see whether Pyrex would be suitable for Hale's new venture, a test blank for the 74 in Toronto telescope was cast. Proving successful in that instrument, Pyrex became standard for most large reflectors.

The weight of a 200 in disk would be enormous and a unique method of casting it with a ribbed back resulted. A more sophisticated method than Ritchey's was needed for mirror support. Neither a cross-axis nor a yoke arrangement for the mounting was suitable; a modified yoke, with an open horseshoe-shaped upper end of the polar axis provided support, stability and access to all parts of the sky. A triumph of pre-war engineering, the massive telescope could be moved with ease by small electric motors. Because of wartime delays, the 200 in telescope did not see first light until 1948. It would remain the world's largest until 1975.

Most conventional reflectors had primaries of $f/5$ or slightly faster, and $f/18$–$f/20$ focal ratios for spectroscopic work at the Cassegrain focus. Direct photography, not typical before 1950 with such instruments, could be done at the prime focus using Ross lenses to correct for coma, a defect in reflectors that stretches images away from the optical axis. Ritchey and Henri Chrétien proposed an alternative scheme which combined a near-parabolic primary with a hyperbolic secondary and a concave photographic plate. This arrangement produced a wider field but was a difficult combination to fabricate. Ritchey built two examples, but the system did not come into its own until much later. Another way to obtain a wider field was the catadioptric arrangement, combining both lens and mirror. The first such instrument was BERNHARD SCHMIDT's photographic telescope, developed from 1931. The SCHMIDT TELESCOPE combined a spherical primary with a specially ground corrector lens. The lens compensated for both spherical aberration and coma, although it did produce a slight chromatic aberration. As the image produced is a curved plane, the photographic film had to be curved. Such a combination allowed for a very fast, short-focal-length system with a wide field for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

photography. By 1948, two such instruments were at work at Palomar Mountain; the larger, with a 48 in corrector plate, was central to the National Geographic Society–Palomar Observatory Sky Survey.

Other specialized instruments catered to solar astronomy. Hale was a leader in designing solar telescopes, beginning with a horizontal arrangement at Yerkes and Mount Wilson. Later, he developed vertical solar telescopes with pairs of optical flats to reflect sunlight downwards to a primary. By reducing thermal currents, this was an effective method for spectroscopy. To observe the corona, Bernard Lyot invented the CORONAGRAPH in 1930, an instrument that blocked the solar disk and largely eliminated scattered light from the optical system.

Throughout this period, other technological innovations increased the power and improvement of these large instruments. Production of moving parts requiring a high degree of precision stretched the capabilities of mounting manufacturers. Steel replaced cast iron for the major parts of tubes and mountings. Steel torsion rods to maintain the rigidity of tubes gave way to Mark Serrurier's truss tube in the 200 in telescope. Relief from the increasing weight of polar axes and the need for smooth motion required roller bearings in the late 19th century. Common and Ritchey experimented with mercury flotation. Plaskett's substitution of self-aligning ball bearings became standard for large reflectors, but the 200 in telescope's designers, concerned about weight, opted for oil pad bearings.

Temperature maintenance was also important. Ritchey introduced a diaphragm above his primary mirrors along with blankets and others chose similar methods and a variety of insulation materials. Small electric heaters were introduced into spectrographs. Weight-driven clock drives were eventually superseded by electrically synchronized driving systems. Better photographic emulsions allowed for faster photography and extended spectral coverage. Experiments with aluminizing mirrors showed the method's superiority to silver by the early 1940s. The introduction of workable photoelectric photometers made photometry an exact science. Still, in this pre-computer age, the quality of astronomical work depended on astronomers with great skill at controlling these giants.

Telescopes were not only for professionals. Amateur astronomers, in growing numbers in the early 20th century, purchased both refractors and reflectors of modest aperture from a variety of manufacturers. Reflectors, being easier to make, became the focus of a substantial amateur telescope-making fraternity in the United States from the 1930s. Access to affordable but high-quality telescopes allowed amateurs throughout the world to make significant contributions to astronomy with lunar, planetary and variable-star observations.

The telescope stands as the technological icon for astronomy in much the same way the microscope does for biology. Although it appears to be a deceptively simple optical device, the telescope, through advances in engineering, optical design and materials (see TELESCOPE ENGINEERING), has become one of the most sophisticated modern scientific instruments. Telescopes have always been expensive; by the time the 200 in telescope saw first light, the telescope had become symbolic of the newly emergent 'Big Science'.

Bibliography

Gingerich O 1984 *Astrophysics and Twentieth-century Astronomy to 1950* (Cambridge: Cambridge University Press)

King H C 1955 *The History of the Telescope* (London: Griffin)

Richard A Jarrell

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Theodosius of Bithynia (*c*. 160–*c*. 100 BC)

Born in Bithynia, Anatolia (now Turkey), author of *Sphaerics*, a book on the geometry of the sphere, written to provide the mathematics for astronomy, and possibly based on pre-Euclidean textbook by EUDOXUS. Said to have invented a sundial suitable for everywhere on Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Theon of Alexandria (*c*. 335–*c*. 395)

Born in (possibly) Alexandria, Egypt, wrote commentaries on PTOLEMY'S *Almagest* and Euclid's *Elements* (his version of Euclid was the earliest known until a nineteenth century discovery).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thom, Alexander (1894–1945)

British engineer and archaeo-astronomer, born in Scotland, became professor of engineering science at Oxford University. Accurately surveyed the megalithic sites in Britain and the megalithic lunar observatories, identifying their shapes and astronomical alignments. He claimed to have discovered in the dimensions of the circles, as apparently used by the megalithic constructors, quanta of length that he called the megalithic yard and megalithic inch. Possibly he over-interpreted the data

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thomson, William [Baron Kelvin of Largs] (1824–1907)

Born in Belfast, Ireland, became professor of physics at Glasgow, studied thermodynamics, proposed an absolute scale of temperature, measured by the *Kelvin temperature scale*. In parallel with HERMANN VON HELMHOLTZ, he formulated a theory of stellar evolution, in which a star radiated the energy that it released by progressive gravitational contraction. The lifetime of a star from this theory (the time over which its gravitational potential energy is radiated from its surface) is the *Kelvin–Helmholtz time scale* and proved to be too short to be compatible with the geological age of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Three-body Problem

The three-body problem, i.e. the problem of finding the motions of three gravitating bodies moving under their mutual gravitational attraction, was studied by Isaac NEWTON in an attempt to explain the motion of the Moon.

In the general three-body problem three bodies of arbitrary mass start from an arbitrary initial configuration and move purely under the inverse square force of attraction. There is no general analytical solution, other than series expansions, to this problem since typically the orbits are chaotic.

The progress leading to our current understanding of the problem has taken centuries and the efforts of leading mathematical physicists such as Euler, Lagrange, Jacobi and Poincaré, to mention a few classical ones. In 1772 LAGRANGE found many special solutions of the problem (such as the equilateral triangle solution), while Jacobi studied the circular restricted three-body problem where one body is of negligible mass and the two others describe a circular orbit about their common center of mass. In this article we concentrate on the astrophysical aspects of the stellar three-body problem and purposely neglect many of the mathematically interesting solutions which, in fact, have physical realizations only in planetary systems (i.e. systems with large mass ratios). The existence of chaos in the three-body problem was already the conjecture of POINCARÉ in 1899, and that has been amply confirmed by recent works.

One may divide the astrophysically interesting three-body systems into three main categories:

(a) Systems of stars (of comparable masses and initial distances) that break up after some dynamical evolution. Such configurations are typical in star forming regions (see STAR FORMATION).
(b) Scattering of single stars off CLOSE BINARY STARS. These events are important in the dynamics of STAR CLUSTERS.
(c) Stable hierarchical systems, examples of which are the numerous triple stars observed in the galactic field.

(See also CELESTIAL MECHANICS.)

Close initial configurations

The first attempts at actually solving the orbits of the three bodies date from the early twentieth century. Sundman constructed a (formal) series solution which converges for all times, but the solution is useless in practice due to extremely slow convergence. The first published attempt to integrate numerically orbits of three masses was made by Burrau. He performed an orbit integration of a special problem which is usually called Burrau's problem. There three bodies of masses 3, 4 and 5 units start at rest from the corners of a Pythagorean triangle such that the mass values scale with the lengths of the sides opposite to the body in question.

Without computers Burrau was able to calculate only a short stretch of the orbits; the complete calculation was performed by Szebehely and Peters in 1967. Figure 1 illustrates the complete orbit solution.

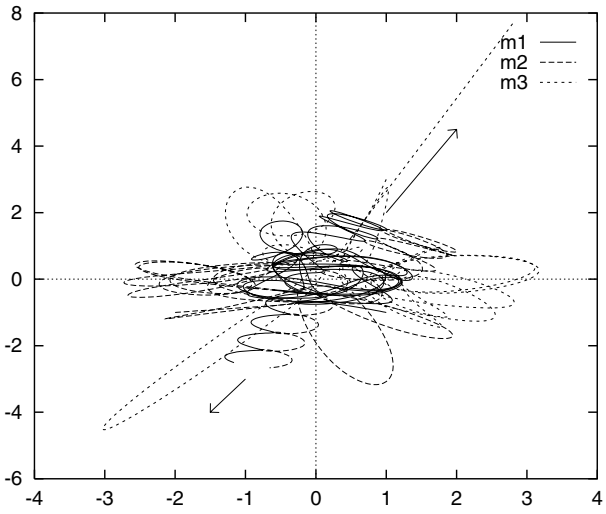


Figure 1. Burrau's three-body problem.

Even though this particular problem has no special astrophysical significance, it illustrates well a typical solution of the three-body problem: as a rule, close approaches of two of the bodies occur while the third body is at a somewhat greater distance. Often the third body goes so far away that the other two bodies form a temporary binary. This situation is called *ejection*. At other times all bodies stay close to each other and perform repeated close encounters at frequent intervals. This type of motion is called *interplay*. When one of the members of a temporary binary is replaced by another body, we call the event *exchange*. Finally, one of the bodies leaves the binary permanently, in which case we have an *escape*. In figure 1 a single body moves to the upper right, and the binary to the lower left at the time of the escape.

The escape times in a large number of bound triple systems occur at random, in the manner of radioactive decay. They are best described by the half-life of the triple system. The lifetime of Burrau's system prior to the escape is about 20 times the period of the initial temporary binary, and this is in fact a rather typical value for the half-life in general.

After the escape the orbits can be approximated by Keplerian two-body orbits for all future times. The binary is essentially described by its two masses, its semi-major axis and its eccentricity. The relative motion of the binary and the third body is described by the escape velocity at infinity of the third body. If we only need to know this final asymptotic state of the three-body problem, then the solutions can be stated in a simple way.

In astrophysics, the finite sizes of the bodies are also of importance. Close approaches between the bodies may lead to collisions of the celestial bodies, or at least various frictional forces may come into play. If the bodies are surrounded by tenuous outer layers or disks, the latter

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

may be stripped away during the close approaches. The distance of the closest two-body approach is not known in advance but can be estimated statistically. In case of Burrau's problem the shortest distance is 4×10^{-4} units, about 0.01% of the original dimension of the system. Typically only 1% of all similar three-body systems have approaches closer than this.

Statistically the final states of the three-body systems after the escape may be described by the approximate analytical theory of Heggie and Monaghan. One starts by writing out the energy E_0 of the three-body system after the escape as

$$E_0 = \frac{1}{2}\frac{m_s m_B}{M}\dot{\boldsymbol{r}}_s^2 + \frac{1}{2}\frac{m_a m_b}{m_B}\dot{\boldsymbol{r}}^2 - G\frac{m_a m_b}{r} \qquad (1)$$

where bodies of mass m_a and m_b form a binary and the body with mass m_s is the third body which will be considered an escaper. The binary mass is written $m_B \equiv m_a + m_b$ and the total mass of the system $M \equiv m_B + m_s$. The separation vector between the binary members is $\boldsymbol{r}$ and the third body is at separation $\boldsymbol{r}_s$ from the binary centre of mass. G is the gravitational constant. Equation (1) is an approximate expression valid for very large separations of the the binary and m_s (the potential energy due to interaction of the binary and m_s has been neglected).

It is now assumed that the probability of a given end state of the system is proportional to the volume of the phase space available for escape. The density of the end states σ per unit energy is

$$\sigma = \int \cdots \int \delta\left(\frac{p_s^2}{2m} + E_B - E_0\right) \mathrm{d}\boldsymbol{r}_s\, \mathrm{d}\boldsymbol{p}_s\, \mathrm{d}\boldsymbol{r}\, \mathrm{d}\boldsymbol{p} \qquad (2)$$

where $m = m_s m_B/M$, E_B is the energy of the binary, and $\boldsymbol{p}$ and $\boldsymbol{p}_s$ are the canonical momenta associated with $\boldsymbol{r}$ and $\boldsymbol{r}_s$.

When equation (2) is integrated over $\boldsymbol{r}_s$ and $\boldsymbol{p}_s$ we obtain the statistical distributions for the binary. Correspondingly, we may integrate over $\boldsymbol{r}$ and $\boldsymbol{p}$, and obtain the distributions for the escaper. After carrying out the appropriate normalizations the main results are:

(a) The binary eccentricity e is distributed as

$$f(e)\,\mathrm{d}e = 2e\,\mathrm{d}e \qquad (3)$$

in a three-dimensional system and

$$f(e)\,\mathrm{d}e = \frac{e\,\mathrm{d}e}{\sqrt{1-e^2}} \qquad (4)$$

in two-dimensional systems.

(b) The probability that the escaper is the body with mass m_s is given by

$$f(m_s)\,\mathrm{d}m_s = \frac{m_s^{-2.5}\,\mathrm{d}m_s}{m_s^{-2.5} + m_a^{-2.5} + m_b^{-2.5}}. \qquad (5)$$

(c) The binary energy E_B is distributed

$$f(E_B)\,\mathrm{d}E_B = 3.5|E_0|^{3.5}|E_B|^{-4.5}\,\mathrm{d}E_B \qquad (6)$$

$|E_B| \geq |E_0|$. More accurate theories show that the power -4.5 of $|E_B|$ is approximate (the correct one depends somewhat on masses and is $3 + \sqrt{13}/2 \approx 4.8$ for equal masses). From equation (6) it is straightforward to calculate the distribution of the escape velocity V_s:

$$f(V_s)\,\mathrm{d}V_s = 3.5\frac{m_s}{|E_0|}V_s\left(1 + 0.5\frac{m_s}{|E_0|}V_s^2\right)^{-4.5}\mathrm{d}V_s. \qquad (7)$$

Equations (3)–(7) constitute the complete statistical solution of the unstable three-body problem when the three bodies are initially in a bound state.

Scattering off a binary

The above equations are applicable also in the case of scattering of a single body from a hard binary (i.e. the total energy $E_0 < 0$). Then one may often consider that a temporary bound state, a *resonance*, has formed and that the break up of the resonance state follows the standard statistical theory.

In STELLAR DYNAMICS one often encounters situations where the binary is not hard. The theory of soft-binary scattering has been derived using the assumption that the third body approaches one of the components of the binary to a distance which is much less than the semi-major axis of the binary, and such that the motion of the third body relative to this component is approximately unperturbed by the other binary member during the most important part of the encounter. After averaging over all relevant variables one obtains a cross section for a change in the binding energy of the binary, for an exchange of binary members, or for a complete disruption (*ionization*) of the three-body system. The applicability of the analytic theory has been tested by large numbers of numerical scattering experiments.

The results from the scattering theory may be expressed as follows: let the third body of mass m_s come from infinity with the asymptotic velocity V_s relative to the binary center of mass, and let the (positive) binding energy of the binary be x. After the scattering the binding energy has changed to a value $x + y$. Then the differential cross-section $\mathrm{d}\Sigma/\mathrm{d}y$ for an energy change of magnitude y is

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}y} = 4\pi G^2 \frac{m_a m_b m_s^2}{m_B}\frac{1}{V_s^2 y^2}F(x, y) \qquad (8)$$

where

$$F(x, y) = 1 - \frac{4}{3}\frac{x}{y},\ y < 0 \qquad (9)$$

$$F(x, y) = \left(\frac{4}{3}x + \frac{7}{3}\right)\left(\frac{x}{x+y}\right)^{5/2},\ y > 0. \qquad (10)$$

Equation (8) may be integrated over y to obtain the total cross section Σ, for example, for ionization or for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

capture, i.e. for the formation of a resonance. Usually the result is expressed relative to πa^2 where a is the semi-major axis of the binary.

The expression is simplified by using the critical velocity V_c, defined by

$$\frac{1}{2}mV_c^2 = \frac{Gm_am_b}{2a} \tag{11}$$

which is the critical incoming velocity of the third body making the total energy equal to zero (i.e. we have $E_0 = 0$ when $V_s = V_c$). Using this formalism the cross section for ionization becomes ($V_s \gg V_c$)

$$\Sigma_{\rm ion} = \pi a^2 \frac{40}{3} \frac{m_s^3}{m_a m_b M} \frac{V_c^2}{V_s^2}. \tag{12}$$

Corresponding expressions have been derived for the cross sections of capture and exchange.

A convenient semi-empirical expression for the differential cross section for energy exchange in case of equal masses and hard binaries was given by Spitzer (1987)

$$\frac{{\rm d}\Sigma_{\rm tot}}{{\rm d}\Delta} \approx \pi a^2 \frac{13.7V_c^2}{V_s^2} \frac{1}{\Delta^{1/2}(1+\Delta)^4}. \tag{13}$$

Here $\Delta = y/x$, and the approximation includes the effects of both resonant interactions and distant fly-bys. The expression is valid for $\Delta > 0$, which is by far the the most typical case for hard binaries (i.e. binaries with binding energy much larger than the typical kinetic energy of single stars).

Hierarchical triplets

In hierarchical triplets there is a binary well isolated from the third body which is in a closed orbit (outer binary orbit) about the center of mass of the binary (inner binary). Such configurations are stable if the pericenter of the outer orbit is distant enough from the binary: typically the pericenter distance must satisfy the inequality $q_{\rm min} \geq 3a$, but the exact value depends on the masses and orbital elements.

The interesting property of such systems is that the secular perturbations by the distant third body transform the inner binary orbit. In 1962 Kozai considered the secular perturbations of asteroids due to Jupiter, and the predictions of this analytical theory constitute what is now known as the *Kozai mechanism*. There occurs a long-period, cyclical exchange of angular momentum between the third body and the binary.

One of the most interesting results is that the eccentricity e and inclination i of the inner binary oscillate and can have large amplitudes, provided that the initial inclination of the outer orbit relative to the binary plane is high enough. In particular, if the initial inclination $i = 90^\circ$, the system evolves to $e \to 1$ and a collision between the binary members takes place. The length of the eccentricity cycle is of the order of $P_{\rm out}^2/P_{\rm in}$, where $P_{\rm out}$ is the period of the outer binary and $P_{\rm in}$ the period of the inner binary.

Numerical methods

The orbits of three-body systems are generally extremely sensitive to changes in initial conditions. Thus numerical calculations must be done with high accuracy. Nowadays most numerical integrations of three-body systems utilize regularization by the Kustaanheimo–Stiefel transformation. For studies of long-term evolution of hierarchical systems symplectic methods are also available, which do not produce secular errors in constants of motion.

Bibliography

The three-body problem is dealt with in most textbooks on celestial mechanics, which generally deal with the restricted three-body problem and the Lagrangian solutions.

Thorough treatments of the three-body problem are found in:

Marchal C 1990 *The Three-body Problem* (Amsterdam: Elsevier)

Szebehely V 1967 *Theory of Orbits. The Restricted Problem of Three Bodies* (New York: Academic)

More recent works of interest include:

Barrow-Green J 1996 *Poincaré and the Three Body Problem* (Princeton, NJ: Princeton University Press)

Diagu F and Holmes P 1997 *Celestial Encounters: The Origins of Chaos and Stability* (Princeton, NJ: Princeton University Press)

Two classical works on the general three-body problem are:

Burrau C 1913 Numerische Berechnung eines Spezialfalles des Dreikörperproblems *Astron. Nachr.* **195** 113–16

Sundman K F 1912 Memoire sur le problème des trois corps *Acta Math.* **36** 105–79

Semi-classics include:

Heggie D C 1975 Binary evolution in stellar dynamics *Mon. Not. R. Astron. Soc.* **173** 729–87

Kozai Y 1962 Secular perturbations of asteroids with high inclination and eccentricity *Astron. J.* **67** 591–8

Monaghan J J 1976 A statistical theory of the disruption of three-body systems—I. Low angular momentum *Mon. Not. R. Astron. Soc.* **176** 63–72

Szebehely V and Peters C F 1967 Complete solution of a general problem of three bodies *Astron. J.* **72** 876–83

Useful sources for references are:

Spitzer L 1987 *Dynamical Evolution of Globular Clusters* (Princeton, NJ: Princeton University Press)

Valtonen M J 1988 The general three-body problem in astrophysics *Vistas Astron.* **32** 23–48

Valtonen M J and Mikkola S 1991 The few-body problem in astrophysics *Ann. Rev. Astron. Astrophys.* **29** 9–29

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Recent works on the chaotic aspect of the three-body problem include:

Boyd P T and McMillan S L W 1992 Parameter space structure in irregular gravitational scattering *Phys. Rev.* A **15** 46, 6277

Heinämäki P, Lehto H J, Valtonen M J and Chernin A D 1998 Three-body dynamics: intermittent chaos with strange attractor *Mon. Not. R. Astron. Soc.* **298** 790–6

There is a series of papers on three-body scattering by Hut and associates. The latest in the series is

Heggie D C, Hut P and McMillan S L W 1996 Binary-single-star scattering. VII. Hard binary exchange cross sections for arbitrary mass ratios: numerical results and semianalytic fits *Astrophys. J.* **467** 359–69

The Kozai mechanism has been recently used in many situations. Useful references are:

Innanen K A, Zheng, J Q, Mikkola S and Valtonen M J 1997 The Kozai mechanism and the stability of planetary orbits in binary star systems *Astron. J.* **113** 1915–19

Kiseleva L G, Eggleton P P and Mikkola S 1998 Tidal friction in triple stars *Mon. Not. R. Astron. Soc.* **300** 292–302

Numerical methods can be found in:

Aarseth S J and Zare K 1974 A regularization of the three-body problem *Celestial Mech.* **10** 185–205

Mikkola S 1997a Practical symplectic methods with time transformation for the few-body problem *Celestial Mech. Dynam. Astron.* **67** 145–65

Mikkola S 1997b Numerical treatment of small stellar systems with binaries *Celestial Mech. Dynam. Astron.* **68** 87–104

Mauri Valtonen and Seppo Mikkola

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Thuban

The star α Draconis, a white giant of spectral type A0III and apparent magnitude 3.67. At a distance of 297 light-years (parallax 0.011″) it has an absolute magnitude of −1.1. It was the pole star *c*. 2800 BC. The name derives from the Arabic name of the whole constellation, meaning 'the Dragon'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tibbon, Jacob ben Machir ibn [known as Prophatius] (1236–1312)

Doctor, born in Marseilles, Spain (now France), a translator, mathematician and astronomer. He invented what is known as *Jacob's quadrant*, and the book in which he describes it contains a table of 11 stars used in the construction of the instrument. He wrote *Luhot* (*Tables*) a book of astronomical tables giving positions of stars at Paris. Tibbon's work was mentioned in DANTE's *Divine Comedy* and used by COPERNICUS in forming his theories.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tides

Everyone who has lived near the ocean is very much aware of the variation in the level of the ocean water that is called the *tide*. On nearly all days there are two high tides and two low tides, and the astute observer will notice that these extreme tides occur about fifty minutes later on the average than they did the day before. The fact that the MOON reaches its maximum elevation above the horizon with the same average 50 min delay on each successive day makes it logical to suspect the Moon as the cause of our observed ocean tide.

In contrast to their awareness of the ocean tide, few people know that the solid body of the Earth also experiences a tide with an amplitude of about 30 cm. This motion is not at all obvious, as all objects on the solid surface of the Earth move together, just as the ocean tide would not be obvious to an observer on a ship in midocean. But the solid-body tide is detectable with sensitive instrumentation. There is even an atmospheric tide on the Earth that is detectable with a simple barometer. However, this tide is caused by the absorption of radiant energy from the SUN rather than gravitational effects, where the atmospheric tide on the Earth caused by the latter is negligible by comparison.

The Moon interacts with the Earth through gravitational force, and one can generalize the idea that the Moon is responsible for the Earth tide to every SOLAR SYSTEM object: each planet or satellite in the Solar System experiences tidal distortions due to all the other objects attracting the planet or satellite gravitationally. As a first test of this hypothesis, we expect the Sun to also raise a tide on the Earth, and we observe that it does. The Earth tide due to the Sun is not as large as that due to the Moon, but it is quite noticeable when one considers that the highest and lowest tides occur when the Sun and Moon are nearly lined up—a time when they both pull on the Earth in the same or opposite directions. It is easy to follow the heights of the local tidal extremes in newspapers of any coastal city where these extremes are maximal at new and full Moon.

Below, we show how the tidal deformation of a body is caused by differences in the gravitational forces exerted by the disturbing body on different parts of the disturbed body, develop a simple physical model with which we can describe the essential features of tidal distortion, and indicate the conversion of tidal energy into heat as the tide oscillates. It is this change of mechanical energy into heat that allows slow changes in the orbits and rotations of many objects in the Solar System, and we discuss what some of the consequences of such tidal evolution have been for each planet and their satellites, if any. Tidal forces also affect stars, star clusters, galaxies, clusters of galaxies, etc that we discuss briefly, but we shall limit most of our examples to the easily observable consequences in the Solar System. We end by pointing out current frontiers of research.

Theory of tides

We are all familiar with gravity as the force that attracts us and all other objects to the Earth. An object dropped near the Earth's surface accelerates toward the center of the Earth at the rate of 980 cm s^{-2}. However, the acceleration of the Moon toward the Earth is only 0.27 cm s^{-2}, as this is the acceleration necessary to keep it moving in its almost circular orbit. From these two numbers, we find that the acceleration of gravity is proportional to $1/r^2$, where r is the separation of the centers of the two bodies that are being attracted to each other—either the Earth and Moon or the Earth and the object we drop. We also know that if we double the mass of an object, that is, double the amount of material in it, the force of gravity on the object is doubled. It is twice as heavy, but since the acceleration of an object is inversely proportional to its mass (force = mass × acceleration is Newton's second law), the acceleration in the Earth's field remains the same. But we see that the force of gravity on a body must also be proportional to its mass. By observing the orbital motions of the satellites about the MAJOR PLANETS (just as we observed the Moon about the Earth), we can infer that the force of gravity must be proportional to the mass of the attracting object as well, leading to the expression for gravitational force between two spherically symmetric bodies,

$$F = Gm_1m_2/r^2$$

where G is the constant of proportionality called the gravitational constant. This law of universal gravitation was also discovered by NEWTON.

A gravitational field of force of a Solar System object leads to a distortion of another massive object that is in the field because the force decreases as the distance from the source of the field increases. This means that different parts of the object in the field experience different gravitational forces. This is illustrated in figure 1, where the forces on the small mass elements due to the satellite of mass m are represented by arrows at three points on the planet M. The separation of M and m is represented by r, and the radius of the planet is R. If ΔM is the mass of each element indicated on M, the gravitational force due to m on that element is given by

$$\Delta F = \frac{Gm\Delta M}{x^2}$$

where x is the actual distance of ΔM from the center of m. The fact that ΔF decreases as x increases from $r-R$ to $r+R$ is indicated in figure 1 by the decrease in the length of the arrows from the points x_2 to x_c to x_1. As the acceleration of each ΔM is $\Delta F/\Delta M$, from Newton's second law, we see that the side of M closest to m tends to accelerate more toward m than the center, and the center tends to accelerate more than the far side. If the masses (ΔM) of the elements at x_1, x_c, and x_2 were completely free from one another, x_2 would move away from x_c toward m, and x_1 would be left behind by x_c. Relative to x_c, the two points x_1 and x_2 would accelerate in opposite directions at rates indicated in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

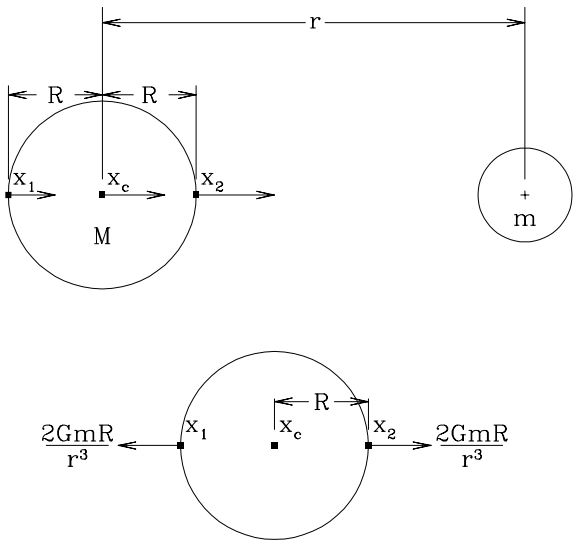


Figure 1. How the decrease in gravitational acceleration with distance from the perturbing mass m leads to differential tidal accelerations on the perturbed body M.

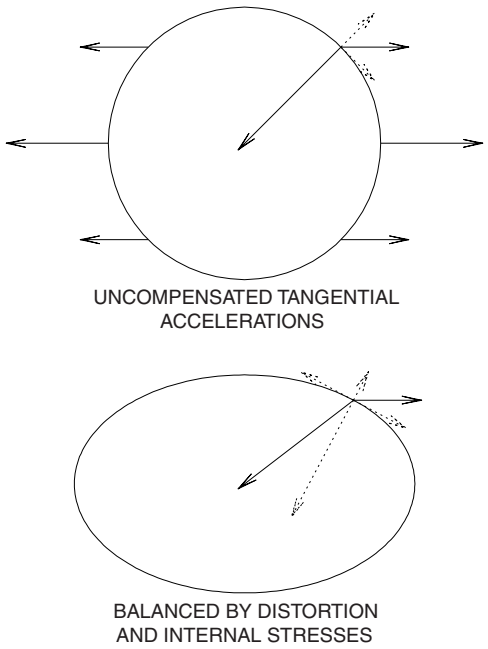


Figure 2. Distortion of a planet to bring it into equilibrium with an external gravitational field.

bottom part of figure 1. These differential accelerations tend to stretch M into a prolate ellipsoid, similar to the shape of an egg, and they depend on the inverse cube of the separation r because they are the differences of two forces that depend on $1/r^2$.

But the ΔMs are not free, since they are attracted gravitationally to one another and to the remaining mass in M. The accelerations toward the center of M due to self-gravity acting on the ΔMs at x_1 and x_2 are far stronger than the differential accelerations shown in figure 1, so mass does not fly off the surface at x_1 and x_2. If the material in M were incompressible, the ΔMs at x_1 and x_2 would weigh less than they would if m were not there, but one might suppose that the ΔMs would not move relative to the center. But M will still distort into an egg shape unless it is perfectly rigid as well as being incompressible. (When acted on by forces, an incompressible substance preserves its volume but may change shape, whereas a *perfectly* rigid substance does not change shape.) The reason is that points in M that do not lie on the line joining the centers of M and m also experience a differential acceleration relative to x_c. But now the differential accelerations are not perpendicular to the surface and are therefore not compensated by the self-gravity that accelerates mass elements more or less toward the center of M. This is shown in figure 2, where the differential acceleration at one point is resolved into its equivalent components perpendicular and tangential to the surface. These components are shown as dashed arrows and are equivalent to the solid arrow, in that together the forces they represent produce the identical acceleration. The self-gravity, shown as an arrow directed toward the center of M in the top part of figure 2, compensates the vertical component of the differential acceleration but not the tangential component. If M were entirely fluid, the uncompensated tangential components of the differential accelerations would cause fluid to flow toward the points on M closest to and furthest from m, until M would resemble the bottom part of figure 2. In this ellipsoidal shape, the self-gravity is now no longer perpendicular to the surface but has a tangential component opposite to that of the differential acceleration due to m. Only in the distorted shape are all the differential accelerations compensated and the whole planet is accelerated like the center at x_c.

If M is not fluid but is rigid like rock, part of the compensating acceleration is provided by internal stresses, and the distortion is not as large as it would be for a fluid. Still, no material is perfectly rigid, so there is always some tidal bulge. All materials are at least somewhat compressible as well, and the expansion of M in the directions where the weight of overlying material is reduced by the presence of m accentuates the tidal distortion. Note that the tidal distortion is independent of the velocity of M relative to m and would occur either if M and m were in a stable orbit or if they were simply falling toward each other. There is no need to invoke centrifugal force to account for the tidal bulge on the side of M furthest from m, although this has often been done in erroneous explanations of tidal deformation. (We limit our discussion to a simple tidal bulge whose size and orientation depends on the system properties and geometry. Practitioners of tidal theory find it expedient to decompose the simple bulge into a sum of separate tides with distinct frequencies and geometries. This

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

decomposition is a mathematical convenience that is not necessary for understanding the essential features of tidal theory and evolution.)

We are now in a position to understand the main features of the Earth tides, given that the Earth tends to be distorted into an egg shape which is aligned toward the Moon and into a lesser distortion aligned toward the Sun. First, from the bottom part of figure 1, the differential magnitudes of the solar and lunar tides follow from the expression for the differential acceleration which is proportional to m/r^3. The lunar mass is much smaller than the solar mass, but it is so much closer to us that the r^3 in the denominator more than compensates for the tiny mass and results in a lunar tide about twice the magnitude of the solar tide. The ocean tide rises and falls at a given point on a coast because the Earth rotates under the Moon. The tidal bulge tends to stay more or less aligned with the Moon, so for our fixed position on the *rotating* Earth, we are carried through alternate high and low points of the tidal deformation. The high tides occur when we pass under the Moon and when we are on the opposite side of the Earth from the Moon, and low tides occur when we see the Moon rise or set. Hence, there are two high and two low tides each day. The ocean tide is obvious to us on a coast because the solid Earth is rigid and distorts less than the fluid ocean.

Close observation reveals that the ocean tide is more complex than the simple picture we have just presented. The high tide can occur at times considerably removed from the time the Moon reaches its maximum elevation above the horizon. This is because the continents interrupt the free passage of the tide. The ocean basins are like giant bathtubs in which the water likes to slosh with a period of oscillations given by $w/\sqrt{gd}$, where w is the width of a given ocean, d is the average depth, and g is the acceleration of gravity at the Earth's surface. This period is comparable to the half-day tidal period, which means the reflection of the tidal wave off a continental boundary and the resulting sloshing of the ocean water can result in high tides occurring at odd times relative to the lunar position in the sky. As an example, high tide occurs about two hours *before* the Moon reaches its highest point in the sky at Santa Barbara, California. Long narrow bays can also change the times of various tidal phases relative to the Moon's position. Only the Earth has such a substantial ocean interrupted by continents, so all other Solar System bodies are more likely to conform to the situation where their tidal bulges are nearly aligned with the direction to the tide raising body. This simplicity is extremely useful when we consider the dissipation of tidal energy in a more or less solid satellite or planet with no complicating ocean.

We pointed out earlier that the lunar tide on Earth is about twice as large as the solar tide because this is the ratio of the corresponding differential accelerations or forces causing the tide. In other words, for the small tidal amplitudes observed, the amplitude is proportional to the disturbing force. This is exactly the way a spring behaves—the stretch is proportional to the stretching force.

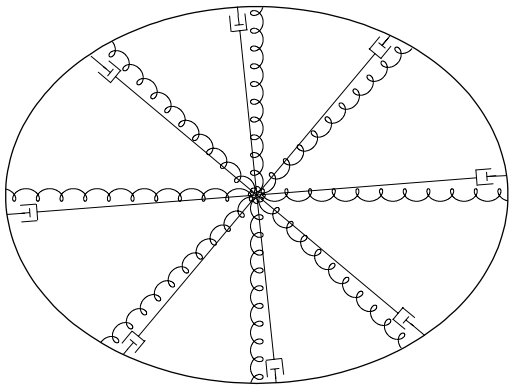


Figure 3. Idealized representation of a solid, elastic planet by a collection of springs. A fluid planet can be similarly represented except the restoring force of the springs is purely gravitational instead of elastic. Here the planet is shown distorted by an external field.

So we can represent a solid satellite or planet now as a collection of springs such as that shown in figure 3, where the restoring force of the springs is a combination of gravity and elasticity of the material. (For a fluid planet the restoring force would be all gravitational as in figure 2.) As the planet or satellite rotates relative to the tide-raising body, the springs are alternately stretched and relaxed. No material is perfectly elastic, so part of the energy stored in the stretched spring is converted to heat. This conversion of energy is represented by the dashpots attached to each spring in figure 3. (A *dashpot* is a device for turning the kinetic energy of motion of a mass at the end of a spring into heat. It can be thought of as a leaky piston in a cylinder of oil where the oil is 'dashed about' as the piston moves. An everyday example of a dashpot is the shock absorber on an automobile.) The springs follow the forcing term, that is, reach their maximum stretch when the differential acceleration is greatest, if the period of variation of the disturbing force at a given point on the body is sufficiently long compared to the shortest natural period of oscillation of the body. On the Earth, the period of variation of the disturbing force is, on average, half the time between moonrises, whereas the Earth 'rings' when struck by a large earthquake with periods like an hour. This condition of slow variation of the tides compared to natural periods of oscillation is satisfied on all the Solar System objects subject to tidal deformation.

The dissipation of mechanical energy (conversion into heat) causes each spring to reach its maximum stretch slightly after the maximum force is applied by analogy to a damped harmonic oscillator, of which a mass on the end of a spring with dashpot is the classical example. This means that for a planet or satellite (M in figure 1) uncomplicated by the interaction of oceans with continents and bays, the high tide follows the perturbing body m, but it occurs at a point after m has passed overhead. The amount of the time delay in the response of the spring representing the planet or satellite is proportional to the amount of dissipation per

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

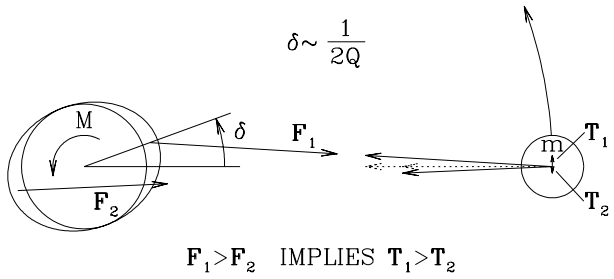


Figure 4. Consequences of dissipation of tidal energy, where the spin of M is being retarded while m is being accelerated in the direction of its orbital motion.

oscillation cycle when this dissipation is small. If we define the dissipation function Q by

$$Q = \frac{2\pi E_0}{\Delta E}$$

where E_0 is the maximum energy stored in a tidal oscillation and ΔE is the energy dissipated (mechanical energy turned into heat) in a complete cycle, the time delay in the response is proportional to $1/Q$.

This time delay caused by energy dissipation has an interesting consequence shown schematically in figure 4. The bodies m and M are reproduced as in figure 1, except M is shown tidally distorted and rotating relative to m. (There is a similar distortion and perhaps a rotation of m that we shall ignore for the time being.) The fact that the maximum tide occurs at a point on M after m has passed over that point means that the tidal bulge is displaced in the direction of rotation relative to a line to m. Here we assume the rotation of M and the orbital motion of m are in the same direction and that the spin angular velocity of the former is larger than the orbital angular velocity of the latter. If the orbital motion were faster, the bulge would be displaced in the opposite direction. The angle δ by which the tidal bulge leads is equal to $1/2Q$. Since low Q implies high dissipation, δ increases with dissipation. The tidal bulges are about equal in size, but again because the gravitational force decreases as $1/r^2$, the bulge nearest m experiences a greater force from m than the one further away. These unequal forces $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are represented by the arrows (*vectors*) of unequal length. (The arrows are drawn in the directions of the forces.) The first consequence of this inequality results from the fact that the forces are not directed through the center of M. Since $\boldsymbol{F}_1 > F_2$, a twisting effect or torque is exerted on M in a direction which retards the spin angular velocity. Since the size of the tidal bulges is proportional to m/r^3 and the torque on the bulges is also proportional to m/r^3 (again the difference of two $1/r^2$ forces), the torque on M due to tides raised by m is proportional to m^2/r^6.

Forces equal to F_1 and F_2 but opposite in direction are exerted on m (Newton's third law). These are represented by the appropriate vectors in figure 4 drawn from the center of m. These forces are resolved into two components, one directed toward M and one perpendicular to that direction. These components are shown as dashed vectors in figure 4. The reaction forces on m can be replaced by these components to produce the same accelerations of m. The components in the direction of M just contribute to the much larger central attraction responsible for the orbital motion. Since $\boldsymbol{F}_1 > F_2$ then necessarily $\boldsymbol{T}_1 > \boldsymbol{T}_2$ (figure 4). This acceleration of m causes it to spiral away from M as the latter's spin angular velocity is reduced.

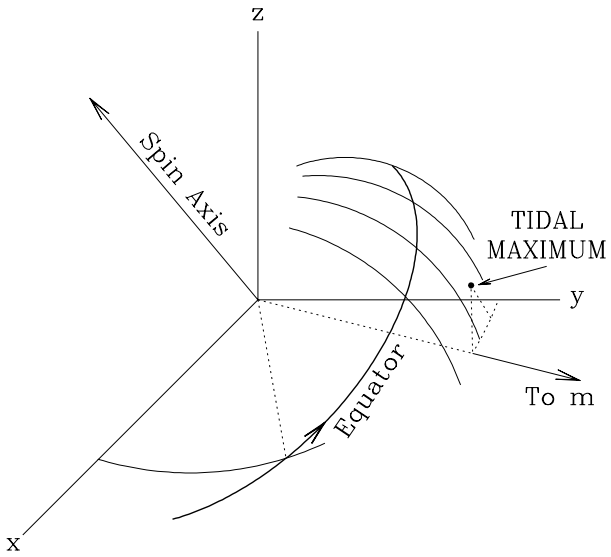


Figure 5. If the equator of M is inclined to the orbit plane, the tidal bulge is carried out of this plane resulting in torques that change the direction of the spin axis in addition to retarding the spin magnitude.

In figure 4, we assume m to be orbiting in the equatorial plane of M. However, if the spin axis of M is not perpendicular to the orbit plane of m, the tidal bulge is carried out of the plane of the orbit by M's rotation, as illustrated in figure 5. This results in a torque on M that is not parallel to the spin axis. Part of the torque tends to tilt the spin axis and part tends to change the spin magnitude. By this means, the obliquity (angle between the spin axis and a line perpendicular to the orbital plane of m) is changed. There is a corresponding but much smaller change in the orbital plane.

In the isolated two-body case discussed so far, as long as there is any rotation of either M or m relative to the other, which includes the cases where either the obliquity or orbital eccentricity is not zero, torques result that cause the orbit and spins to change. If a rapidly spinning planet is very massive compared to a single satellite that orbits in a relatively close circular orbit, the satellite is relatively quickly slowed to SYNCHRONOUS ROTATION by large tides raised on the satellite by the planet, as the tides raised on the planet by the satellite increase the radius of the satellite's orbit. The planet's spin slows only a little because of its large mass. If the orbit is eccentric, tides

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on the satellite tend to drive its spin to a value that is slightly larger than the synchronous rotation. The tide is largest at the periapse of the orbit (point of closest approach), where the now non-uniform orbital angular velocity is maximal. The spin would then tend toward that value where the tidal torque, averaged around the orbit, vanishes, and this occurs at a spin angular velocity that is less than the maximum orbital angular velocity at periapse but larger than the mean orbital angular velocity. Why then is the Moon rotating synchronously with its mean orbital angular velocity in spite of being in a slightly eccentric orbit where tides raised on the Moon by the Earth are trying to make the spin slightly larger? This is because the Moon, or any other satellite, is not axially symmetric. It has a permanent effective egg shape (supported by internal stresses as are large mountains on Earth) that is much larger than that induced by the tide, and the tidal torque from the Earth on this permanent deformation dominates that on the gravitationally induced tide, forcing the long axis and hence the same face of the Moon to point toward the Earth every time the Moon passes the periapse of its orbit, where the tidal force is maximal. This can be true only if the Moon rotates at a rate synchronous with the *mean* orbital motion—not at the slightly larger rate where the tides would take it.

Dissipation of tidal energy still occurs in a synchronously rotating satellite in eccentric orbit, since the tides wax and wane in magnitude as the orbital distance changes, and the position of the tidal maximum rocks back and forth on the satellite because the satellite rotates uniformly whereas the orbital motion is not uniform. As the eccentricity is the cause of this dissipation, the dissipation will lead to a decrease in the eccentricity. At the same time, the tide raised on the rapidly rotating planet by the satellite will try to increase the eccentricity. Recall that the tide and hence the torque ($\propto m^2/r^6$) is maximal when the satellite is at the periapse of its orbit. The greater tidal force on the satellite at this point (T_1 in figure 4) tends to fling the satellite to a greater distance at its apoapse (the furthest point in the orbit) than it would have gone without the kick—thereby increasing the eccentricity. Which effect wins depends on the magnitude of the rate of dissipation in the satellite versus that in the planet (i.e. on Q_m/Q_M).

Now consider a satellite in a circular orbit that is spinning rapidly relative to the synchronous rate and having a non-zero but small obliquity (equatorial plane inclined to the orbit plane). Here the tide raised on the satellite tends to drive the obliquity to larger values. This is best understood if we represent the spin as a vector having a component parallel to the orbit plane and another perpendicular to the orbit plane. The tidal torque is reducing the magnitude of perpendicular component at all points in the orbit, whereas it does not reduce the parallel component when that component is pointing toward or away from the planet. Averaged around the orbit, the fractional decrease in the perpendicular component exceeds the fractional decrease in the parallel component, and the spin vector represented by the components has been tilted more. There is an equilibrium obliquity between 0° and 90° toward which the spin is driven where the *fractional* decreases per orbit in the two components of the spin are equal. This equilibrium obliquity is near 90° if the satellite spins very rapidly, but decreases toward 0° as the spin slows from the tides. For a spin angular velocity less than $2n$ (twice synchronous), the equilibrium obliquity is 0°, so a satellite in a fixed circular orbit should approach synchronous rotation and zero obliquity simultaneously.

If the planet and satellite are of comparable mass, both planet and satellite can approach synchronous rotation and zero obliquity as the orbit expands. Once neither planet nor satellite are rotating rapidly relative to the orbital mean motion, the tides on both bodies eventually tend to reduce the eccentricity. So the natural endpoint of tidal evolution in this case is both bodies synchronously rotating in a circular orbit with both equatorial planes coincident with the orbital plane.

Other outcomes of tidal evolution include a satellite's spiraling into the planet or its disintegration by tidal forces, or perhaps by collisions, to form a ring of material in the planet's equatorial plane (see PLANETARY RINGS). Some of the many rings of small-mass objects orbiting in the equatorial planes of the major planets could have had their origin through a tidal disintegration of a larger, rubble pile type object. Multi-satellite systems exhibit a rich collection of orbital resonances where the mean orbital angular velocities are near the ratio of small whole numbers, and most of these were most probably assembled by differential tidal expansions of initially more random orbits. Within the resonances, continued tidal evolution forces increasing eccentricities and/or orbital inclinations relative to a planet's equatorial plane. In *nearly* all real cases, the rate of tidal evolution is so slow that the logical endpoint of that evolution, where no further changes will occur, has not been reached nor will it be reached before the Solar System ceases to exist as we know it some 5×10^9 yr from now. We examine some of these real situations in the examples below, where some of the remaining puzzles of tidal evolution are also pointed out.

We can summarize tidal theory as follows:

(1) Tides are raised on a Solar System body M by differences in gravitational forces across the body due to any nearby body m (figures 1 and 2).
(2) Any relative rotation between M and m causes the tides and hence the internal stresses to oscillate. This causes some of the tidal energy to be converted into heat much like the heating of a paper clip that is bent back and forth repeatedly or the warming of a liquid in turbulent flow.
(3) The tides resemble damped harmonic oscillators (mass on the end of a spring spanned by a dashpot) in many ways (figure 3), where forcing the oscillator at a period that is long compared to the natural period of free oscillation results in a time lag in the response of the oscillator relative to the forcing term. The maximum of a tide at a point on the distorted body

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

then occurs after the tide-raising body has passed overhead (figures 4 and 5).

(4) The resulting non-alignment of the tidal bulges with the tide-raising bodies results in torques that decrease the relative rotations while changing the separation between the bodies. The separation of two tidally interacting bodies, the obliquities of their spin axes, and the eccentricity of their mutual orbit all change in predictable directions governed by the geometry, spin magnitudes relative to the mean orbital angular velocity, and the dissipation factors Q of each object.

(5) For rapidly spinning planets with a satellite in prograde orbit, the most obvious tidal evolution is a reduction in the spins of both planet and satellite toward rotation that is synchronous with the orbital mean angular velocity as the orbit is expanded.

Consequences of tidal evolution

The differential tidal force and the dissipation of tidal energy described above are responsible for many observed phenomena in the Galaxy and universe. Individual stars are stripped from galactic and GLOBULAR CLUSTERS of stars if the star drifts into a position where the differential tidal force exerted by the total galactic mass exceeds the gravitational force of the other cluster members attracting the star toward the cluster. COMETS are stripped off the Oort cloud of comets that is thought to extend beyond 10 000 astronomical units from the Sun (an astronomical unit (AU) is the distance from Earth to Sun). Weakly bound comets or ASTEROIDS can be broken up by tidal forces as they pass very close to a planet—as was probably demonstrated by the disintegration of Comet Shoemaker–Levy 9 before the spectacular crashes of the pieces into Jupiter in 1994 (see SHOEMAKER–LEVY 9–JUPITER COLLISION). CLOSE BINARY STARS are synchronously rotating with their circular orbital motion, due to retardation of their spins and damping of their orbital eccentricities from the torques arising from the dissipation of tidal energy. The same tidal mechanism may have circularized the orbits of extra-solar planets (EXOPLANETS) discovered orbiting very close to their respective parent stars. Great streams of stars, dust and gas are observed trailing away from a galaxy after being ripped from that galaxy by tidal forces exerted by another galaxy that has passed very close. All of these examples, except the synchronously rotating binary stars, are the results of singular events where transient circumstances lead to relatively short timescale dominance of tidal forces. The Solar System offers many more immediate examples where dissipation of tidal energy over the age of the Solar System has led to very interesting and sometimes spectacular consequences. We shall consider each planet and planet–satellite system in turn to point out how the system has been altered by tides.

Mercury

MERCURY experiences tides raised by the Sun, and sufficient time has elapsed for Mercury's spin to have been slowed to a rate synchronous with its orbital mean angular velocity. It therefore came as quite a surprise when radar observations by Pettengill and Dyce in 1965 revealed Mercury's rotation period to be only two-thirds of its orbital period. The eccentricity of Mercury's orbital ellipse has the rather large value of 0.206. Because of this high eccentricity, Mercury can rotate stably at several values of rotational angular velocity that are half-integer multiples of its mean orbital angular velocity. These particular rotational states are stabilized by the average torque on the *permanent* axial asymmetry. This asymmetry is like a tidal bulge, except it is not induced by an exterior gravitational field but is supported by internal shear stresses. The stability is such that the long axis of the permanent, egg-shaped asymmetry is oriented toward the Sun as Mercury passes through the perihelion (point in orbit closest to the Sun) of its orbit. This same stability was pointed out above as a necessary condition for synchronous rotation of a satellite in an eccentric orbit.

The inferred tidal evolution of Mercury's spin angular velocity is a tidal retardation of the spin through a series of stable states, where the spin angular velocity $\dot{\psi}$ would be $3n$, $2.5n$, $2n$, or $1.5n$, where n is the mean orbital angular velocity. These spin angular velocities would allow the long axis of Mercury's slight egg shape to be nearly aligned with the direction from Mercury to the Sun at every perihelion passage. Such a state is stable against further tidal evolution if the maximum average torque on the permanent deformation exceeds the tidal torque. Mercury need not be captured in a stable state as it passes through. The probability of capture depends on the detailed tidal model, but all tidal models share the property that the states corresponding to the spin angular velocity $\dot{\psi} = An$, where A is a half-integer, have a decreasing probability of capture as A increases, ultimately reaching zero probability when A is so large that the state is not stable. Thus we find Mercury in that spin state ($\dot{\psi} = 1.5n$) which has the highest capture probability of any state Mercury could have passed through as the tides slowed it toward synchronous rotation. Observations of an intrinsic magnetic field on Mercury imply the existence of a conducting liquid core. If this core were liquid early in Mercury's history, the tidal evolution may not have been as described here. The existence of a liquid core with the resulting dissipation of energy through any differential rotation with respect to Mercury's mantle drastically increases the probability of capture into the spin–orbit resonances. If capture of Mercury into the 2:1 spin–orbit resonance were certain because of the core–mantle interaction, Mercury may have had to start its rotation with $\dot{\psi} < 2n$. Mercury is the only example known where such spin–orbit coupling has led through tidal evolution to a stable state other than synchronous rotation.

But tidal dissipation is still occurring in Mercury, making the orbit more circular. The stability of the current spin state vanishes when the orbital eccentricity gets sufficiently small. Tides could thus eventually bring Mercury to synchronous rotation in a nearly circular orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

However, it would take many times the age of the Solar System for this to occur.

That Mercury's spin axis is very close to being perpendicular to its orbital plane (nearly zero obliquity) is also a result of tidal evolution. However, because Mercury's orbit is not fixed, but precesses relative to the stars with a period of about 250 000 yr due to the gravitational perturbations by the other planets, the tides are trying to drive Mercury's spin toward an orbit normal (vector perpendicular to the orbit plane) that is itself moving. The compromise position to which the spin is driven is called a Cassini state, where the spin axis lies permanently in a plane determined by a line perpendicular to its orbital plane and a line perpendicular to the plane of the Solar System. This state is very close to the orbit normal, and Mercury would have been driven there by the tides from virtually any initial obliquity and spin angular velocity.

Venus

VENUS rotates in a direction which is opposite to the sense of its nearly circular orbital motion. This state is described by assigning an obliquity ϵ of nearly 180° to the spin axis. This retrograde rotation is not the result of tidal evolution in the ordinary sense of tides raised by the solar gravitational field, since such tides attempt to drive the obliquity to zero after Venus's spin rate became less than $2n$. If Venus began with an obliquity greater than 90°, like Uranus whose obliquity is 97°, the friction between a liquid core and the solid exterior of Venus called the mantle could dominate the ordinary gravitational tide and drive Venus to the current 180° obliquity. If Venus started with $\epsilon < 90^\circ$, the core–mantle interaction would have aided the gravitational tide in driving Venus eventually toward zero obliquity. In this case the only alternative to account for the current $\epsilon \simeq 180^\circ$ seems to be an atmospheric thermal tide. There is reason to suspect that such a thermal tide exists on Venus, since we observe one on Earth. The absorption of solar radiation in our upper atmosphere couples with the modes of oscillation of our atmosphere to generate an egg-shaped distribution of atmospheric mass whose maxima occur at 10:00 a.m. and 10:00 p.m. local time. With this phase, the gravitational attraction of this thermal tide by the Sun tends to accelerate the spinning atmosphere and thereby the EARTH'S ROTATION. This is opposite to the effect of the gravitational tide which is decelerating the Earth. The effect of the atmospheric thermal tide on the Earth's spin is only about 10% of the effect of the gravitational tide. But Venus's atmosphere has nearly 100 times the mass of our own atmosphere, so if a similar thermal tide is generated on Venus, it may easily dominate the gravitational tide. If we can model that tide by a gravitational tide but simply reverse the torque on the planet, the atmosphere could dominate both the gravitational tide and a core–mantle interaction and drive Venus to its current obliquity from a primordial obliquity less than 90°. Observation of the barometric pressure at the surface of Venus for a Venus day would determine the magnitude of the current thermal tide and probably indicate whether or not the atmosphere was capable of turning Venus upside down.

Earth

Tidal evolution of the Earth–Moon system has been directly measured. Our day is getting longer at the rate of about 0.0016 s each century. The Moon's mean orbital angular velocity is decreasing at the rate of about 22 arcseconds per century[2]. This means that each year, the Moon is moved further away from us by about 3 or 4 cm. This rate implies an Earth which is currently more dissipative than it was in the past, for if we extrapolate the Moon's orbit backward in time, assuming the Earth's Q was always its current value, the Moon ends near the Earth's surface only 1.6 Gy ago—about one-third the age of the Earth–Moon system (1 Gy is 10^9 yr). Most of the tidal dissipation on Earth is in the oceans, and the higher Q in the distant past is usually attributed to a different configuration of the continents, thereby allowing for fewer shallow seas for the dissipation of the tidal energy.

The effect of the tide raised on the Moon by the Earth has driven the Moon's rotation to the endpoint of its tidal evolution consistent with the current orbital configuration. The Moon's rotation is synchronous with its mean orbital motion, which we readily ascertain by always viewing the same lunar face from Earth. In addition, its spin axis has been driven to a stable Cassini state, in which the spin remains coplanar with the perpendiculars to the Moon's and Earth's orbit with an obliquity of $6^\circ 41'$. (The current geometric configuration of the Moon spin axis together with its synchronous rotation are collectively called Cassini's laws, from which the name Cassini state is derived.)

Although tidal dissipation in the Moon has been sufficient to slow it to synchronous rotation and drive the spin axis to a Cassini state, the contribution of the tidal heating to the thermal history of the lunar interior has been negligibly small. The Moon was, of course, once much closer to the Earth, and the rate of dissipation was correspondingly much higher than it is now. However, it was pushed out too fast toward distances corresponding to small dissipation for the interior to heat significantly.

Mars

MARS' spin has not been significantly affected by tidal dissipation. It is too far from the Sun, and its satellites, PHOBOS AND DEIMOS, are very small. The two satellites, on the other hand, are both rotating synchronously with their orbital mean motions. The inner satellite, Phobos, is unusual in the sense that it is so close to Mars (orbit radius $\simeq 6.8$ Martian radii) that its orbital period is less than Mars' rotation period (7.65 h versus 24 h 37 m). Reference to figure 4 shows that in this case the tidal bulge is behind Phobos and the direction of acceleration of the satellite is opposite to its motion. Phobos is therefore spiraling into Mars. The measured acceleration of Phobos' orbital angular velocity leads to a Q of Mars of about 100, which is comparable to that of rocks on the Earth. If Q remains

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the same as Phobos spirals toward Mars *and* if Phobos is sufficiently strong to remain in one piece, it will strike the surface of Mars in about 10^9 yr.

Jupiter

JUPITER is about 300 times more massive than Earth, and its satellites, although three of them are larger than our Moon, total only a little over 10^{-4} of Jupiter's mass. The spin of the planet has thus not been significantly altered by the tides.

The satellites have experienced significant tidal evolution. All the Galilean satellites (discovered by Galileo)—Io, Europa, Ganymede, and Callisto—are rotating synchronously with their respective orbital motions, as is Amalthea, which is a small satellite inside the innermost Galilean satellite, Io (see JUPITER: SATELLITES). The orbits have not expanded very far, however, since Io is only about 5 Jupiter radii from the planet's surface. This close proximity after 4.6×10^9 yr of evolution means the averaged Q of Jupiter must exceed about 6×10^4.

The first three Galilean satellite orbits have mean orbital angular velocities that obey the relation $n_1 - 3n_2 + 2n_3 = 0$ (called the Laplace relation), where the subscripts refer to IO, EUROPA and GANYMEDE respectively. Io's angular velocity is nearly twice Europa's and Europa's nearly twice Ganymede's. The satellites are stably locked together in these orbital resonances, and so their orbits must expand together while maintaining the relation among their mean orbital angular velocities as tides raised on Jupiter push them away. It is likely that the differential tidal expansions of the orbits assembled the satellites into the resonances.

The existence of the orbital resonances has had a profound effect on Io. The condition that conjunctions of Io and Europa (both satellites on the same side of Jupiter at closest approach) always occur when Io is at the periapse of its orbit, forces the eccentricity of Io's orbit to the value of 0.0041. Even though Io is synchronously rotating with its mean orbital motion, the tide raised on Io by Jupiter increases and decreases in magnitude and the bulge moves back and forth to follow the motion of Jupiter in Io's sky. Io rotates uniformly, whereas the orbital motion in an eccentric orbit is not uniform. This causes the oscillatory motion of Jupiter in Io's sky. The tide raised on Io is enormous—more than 7 km high if Io were fluid and uniformly dense. The variable part of the tide is 100 m for this fluid Io, so one expects substantial tidal dissipation. (Taking account of Io's central condensation reduces these tidal amplitudes somewhat, but they remain enormous.) Ordinarily such dissipation would damp the orbital eccentricity to zero, and the dissipation would cease. However, the eccentricity is maintained by the tides raised on Jupiter continuously trying to force Io deeper into the orbital resonance, and the accumulated heat has been sufficient to melt the entire satellite. Tidal theory was used to predict that Io should be extensively melted, and the VOYAGER spacecraft subsequently found it to be the most volcanically active body in the Solar System. This is a case where tidal dissipation has dominated all other energy sources in determining the thermal history of a Solar System body.

Saturn

SATURN, like Jupiter, has not evolved significantly through tidal interaction (see SATURN: SATELLITES). But all the satellites out to and including IAPETUS are rotating synchronously with their orbital mean motions, with the exception of HYPERION, as a consequence of tidal dissipation. Recent detection of albedo features on the surface of TITAN through a narrow infrared window in its perpetual atmospheric haze have verified its synchronous rotation. Voyager 2 images have shown Hyperion to be slowly rotating and therefore tidally evolved in spite of its small size. However, Hyperion is so far from spherical symmetry and is in such an eccentric orbit that strong gravitational torques on the satellite due to Saturn prevent an evolution to an orderly synchronous rotation. Hyperion will most likely tumble in a chaotic fashion for the remaining lifetime of the Solar System.

There are three sets of orbital resonances among the satellites of Saturn that were known before Voyager 1 and several more among the recently discovered smaller satellites. Of the former resonances, the 2:1 resonances between MIMAS and TETHYS and between ENCELADUS and DIONE could have been assembled by differential tidal expansion of their orbits. However, the 4:3 Titan–Hyperion resonance could have been so assembled only if the Q of Saturn was considerably smaller for the tide raised by Titan on Saturn compared to the Q appropriate to the inner satellites. Even if this were true, Titan's mass is so large that special circumstances would have had to prevail to allow its orbit to expand to the 4:3 resonance with Hyperion without close approaches that would eliminate Hyperion from the system. The proximity of Mimas to Saturn establishes the lower bound $Q \geq 16\,000$ for Saturn. A lower Q would have pushed Mimas farther away from the planet than we find it. This value of Q is too high to allow the evolution to the observed state of Titan–Hyperion resonance from an initial condition appropriate for a presumed tidal assembly. As there are no obvious reasons for imposing a sufficient amplitude or frequency dependence on Q to preserve both the tidal assembly of the inner resonances and that of Titan–Hyperion, it is often assumed that the latter resonance must be linked to the formation process and therefore not due to tidal evolution.

Observations by Voyagers 1 and 2 showed that the surfaces of several satellites had been reworked by some indigenous processes after the major cratering events. Especially puzzling is Enceladus, which is an icy satellite only 250 km in radius but which has areas on its surface that are probably less than 10^9 yr old. The radioactive content of Enceladus is negligibly small, and application of the tidal heating hypothesis, which was so successful in accounting for the state of Jupiter's satellite Io, does not supply sufficient energy to cause the resurfacing, given the eccentricity of 0.0044 maintained by the resonance with Dione. Still, tidal heating, by a process of elimination,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

seems to be the only viable source of energy to cause the relatively recent resurfacing of Enceladus. If this is so, then Enceladus must have participated in one or more orbital resonances in the relatively recent past that could force its eccentricity to large values, thereby dissipating sufficient tidal energy to soften its interior. Earlier resonances, since disrupted, with Tethys and with the small moon Janus have been proposed, but there is no consensus on the details of the orbital history.

Uranus

URANUS is too far from the Sun and its satellites too small for the dissipation of tidal energy to have affected its spin. However, the five largest satellites, MIRANDA, ARIEL, UMBRIEL, TITANIA, and OBERON, in order of distance from Uranus, appear to be synchronously rotating with their orbital motion as predicted by tidal theory. (The largest satellite, Titania, has a radius of about 805 km and a mass of about 3.3×10^{24} g, where the latter is about 4.5% of the lunar mass.) The satellite system is unlike those of Jupiter or Saturn as none of the satellites is involved in an orbital resonance. However, it is almost certain that such resonances existed in the past, since some of the satellites have apparently been resurfaced by indigenous processes and the innermost of the five larger satellites, Miranda, has an unusually large inclination of its orbit relative to Uranus's equatorial plane of 4.2°. (Logical formation processes would start Miranda and the other close satellites in equatorial orbits.) The only reasonable way to account for this inclination is to have Miranda pushed into a 3:1 orbital resonance with Umbriel by tides raised on Uranus where the inclination secularly increases to a point where the resonance is disrupted leaving the inclination at its high value. There are remaining puzzles in the Uranus system whose explanations most probably involve the dissipation of tidal energy, but details of the resonance history are still uncertain. Ariel, the next satellite out from Miranda, has been resurfaced since crater counts are drastically reduced in places. Ariel is currently outside the 2:1 orbital resonance with the next satellite, Umbriel, but an orbital eccentricity sufficiently high to have softened the interior of Ariel might have been forced by occupancy of this resonance in the past. However, the 2:1 resonance is difficult to disrupt, and we would expect this resonance to persist if it had ever existed. If the 2:1 resonance never existed, the tidal expansion of Ariel's orbit is limited to the range outside this resonance. This places an lower bound on the Q of Uranus of 11 000, while at the same time, the necessity that Miranda had occupied the 3:1 resonance with Umbriel to account for its orbital inclination places a upper bound on this Q of 39 000. (If $Q < 11\,000$, Ariel's orbit would have expanded sufficiently to have begun inside the 2:1 resonance with Umbriel, whereas if $Q > 39\,000$, Miranda's orbit would not have expanded far enough to have begun inside the 3:1 resonance with Umbriel.) Still, the 2:1 Ariel–Umbriel resonance may have been disrupted by the perturbations of Titania through some subtle resonance interaction yet to be discovered. A final puzzle is that Umbriel's surface is uniformly very old, whereas that of Titania, which is outside Umbriel, has been resurfaced in places. An interesting tidal history of this satellite system remains to be discovered.

Neptune

Prior to the Voyager spacecraft reconnaissance in 1989, NEPTUNE's known satellite system consisted of one large satellite, TRITON, in a close, circular but retrograde orbit (orbiting in a direction opposite to the sense of the rotation of Neptune) and a small satellite, Neried, in a distant, prograde but highly eccentric orbit. Triton is the only example in the Solar System where such a large satellite orbits in a retrograde sense so close to its primary (14.33 Neptune radii). The only plausible explanation for this unusual orbit is that Triton did not form in the same way that the regular satellites of the other major planets had formed by the accumulation of debris in the equatorial plane of the forming planet, but that it was formed elsewhere in the Solar System and captured by Neptune in a highly eccentric, but retrograde orbit, where the capture was most likely accomplished through collision with previously orbiting debris. The damping of the high eccentricity would be accomplished by tidal dissipation from tides raised both on Neptune by Triton and on Triton by Neptune. (Note that the retrograde orbit reverses the usual tendency of the tides on the planet to increase eccentricity.) Tidal theory was used to predict this history before the Voyager encounter, and it placed some severe constraints on what we expected to find from the Voyager images. First, evidence of tidal heating would be revealed by a very young Triton surface with few or no craters. Second, Triton would have cleaned out all small satellites for some distance outside its current orbit. Voyager found Triton to be nearly crater-free, and although 6 small new satellites were found orbiting in Neptune's equatorial plane inside Triton's orbit, none were found outside.

Pluto

Pluto is a binary planet in the sense that its satellite, Charon, is of comparable size (Pluto radius ∼1500 km, Charon radius ∼700 km) (see PLUTO AND CHARON). The discovery of the satellite occurred fortuitously just before the orbital motion of the system swept the orbital plane past the Earth. Repeated occultations and eclipses involving Pluto and Charon resulted in a detailed description of the system. The orbital period is 6.3867 d—the same as Pluto's rotation period. It is also almost certain that Charon is also synchronously rotating with the orbital motion, making the system the only example in the Solar System where tidal evolution has progressed to its logical endpoint. The only possible exception to this conclusion is that the orbital eccentricity may not be quite zero, although observations implying the latter have not been independently confirmed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Future research

We have pointed out several puzzles of tidal evolution of the satellite systems in the Solar System that await resolution using essentially the tidal theory outlined above coupled with theories of orbital resonances. An implicit assumption in this theory is that the tides change slowly enough that their magnitudes are assumed to have the values a static tide would have. However, the circularization of the orbits of the recently found extra-solar giant planets orbiting within 0.2 AU of their stars presents a special problem. If these planets were formed far from their stars and their present close orbits involved scattering by a close encounter with another giant planet in the system, the orbit initially would have been highly eccentric with a very small periapse distance. Such a scenario requires that the eccentricity be damped in a time less than the estimated age of the central star—presumably by tidal dissipation. But both the planet and the star are essentially completely gaseous where ordinary schemes of dissipating tidal energy, such as turbulence, are very inefficient. Until recently, only one scheme, involving a phase change in helium in the deep interior, had been offered to account for sufficient dissipation in Jupiter to account for the tidally maintained set of resonances among the Galilean satellites, and it is uncertain whether conditions prevail that would allow the mechanism to work. However, the distribution of gas in a giant planet as well as that in a star are subject to various modes of oscillation whose frequencies ordinarily would have been thought sufficiently high that the standard tidal treatment would suffice. An example where this assumption is not appropriate is the thermal tide on the Earth, where the excitation is diurnal, but the major response is semi-diurnal. Treatment of an atmosphere disturbed by a varying tidal force due to rotation apparently requires a dynamic rather than a static tidal treatment, where the atmospheric response is resolved into a series of oscillation modes. Application of such an analysis to Jupiter has accounted for sufficient dissipation to accommodate the maintenance of the satellite resonances, although, again, a special structure that might not prevail is required. The frontier of tidal theory thus lies in the development of analyses that account for the full dynamics of gaseous bodies. The major motivations are to obtain rates of tidal dissipation in gaseous bodies that are theoretically secure, which give dissipations in the giant planets within observational bounds, and which could account for the circularization of the orbits of extra-solar planets close to their stars.

Bibliography

Ioannou P J and Lindzen R S 1994 Gravitational tides in the outer planets 2. Interior calculations and estimation of the tidal dissipation factor *Astrophys. J.* **406** 266–78

Melchior P 1983 *The Tides of the Planet Earth* (Oxford: Pergamon)

Peale S J 1976 Consequences and inferences from tidal interactions in the solar system *Ann. Geophys.* **33** 23–30

Peale S J 1986 Orbital resonances, unusual configurations and exotic rotation states among planetary satellites *Satellites* (Tucson, AZ: University of Arizona Press) pp 159–223

Peale S J 1999 Origin and evolution of the natural satellites *Ann. Rev. Astron. Astrophys.* **37** 37–106

S J Peale

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Time Scales: Atomic

Atomic time scales (ATSs) are here considered; in the next section the historical reasons that led to an atomic definition of the second, adopted by the XIIIth Conférence Général des Poids et Mesures (CGPM) in 1967, are given. The Comité Consultatif Temps et Fréquence (CCTF), meeting every 3–4 yr at the Bureau International des Poids et Mesures (BIPM), now performs studies on the Système Internationale (SI) atomic second and on the ATSs, in Paris.

In the subsequent section the main characteristics of a time scale (TS) are reviewed along the resulting requirements, while the final section deals with the theoretical and practical construction of TAI (Temps Atomique International (INTERNATIONAL ATOMIC TIME)) and UTC (UNIVERSAL TIME Coordinated), the TS on which the entire everyday life of all nations is regulated.

Attention is drawn to the fact that the letter 'C' in UTC denotes 'coordinated' and not 'coordinate'; UTC is not a time coordinate in a relativistic meaning, but is the result of a coordinated effort by parts of most of the national metrological institutions. Each institution keeps a local version of TAI (K), where the K in parentheses is the name of the individual laboratory, the expression TAI being usually reserved to the International Time Scale, as calculated by the Time Section of BIPM, on the basis of the readings of the individual clocks kept in the participating laboratories. UTC (K) differs from TAI by an integer number of hours and represents the time reference scale to be used inside a given country; with the notation UTC is intended the realization of this TS in the 0 time zone, around the International Meridian of origin.

The atomic definition of the second and the concept of an atomic time scale

In the middle of the 20th century, for a number of applications the accuracy, precision and stability of the second, as the unit of time interval, were not adequate. Their requirements of scientific applications—lunar radar ranging, spectroscopy research—and technical ones—telecommunications and electronic instrumentation—were not met by the astronomical definitions, reaching relative accuracies between 10^{-7} and 10^{-8}.

At NPL, England, in 1955 a frequency standard using a cesium transition was put in operation: this device offered an accuracy of some 10^{-9} and led to a number of discoveries, one of which was the first measurement of the secular decrement in the speed of EARTH'S ROTATION.

The definition of the 'atomic' second

The present definition of the second is based on a microwave transition of cesium (the official definitions are in French): 'Le seconde est la durée de 9 192 631 770 périodes de la radiation correspondant à la transition entre les deux niveaux hyperfins de l'état fondamental de l'atome de césium 133'.

The value given for the cesium hyperfine transition frequency was obtained by a measurement of cesium frequency in terms of the second of Ephemeris Time; in these terms the atomic SI second can be considered as linked to an astronomical unit, based on the Earth's revolution.

The atomic time scale

A TS is fully identified by two choices:

- the unit of scale, i.e. the time interval to be accumulated and numbered—in the case of TAI this interval is the SI second;
- the origin.

The choice of the origin is fully arbitrary and is a matter of agreements and of convenience; the origin was chosen as 0 hours UT1 on 1 January 1958.

The first definition of TAI, translated in English, is as follows: 'International atomic time (TAI) is the time reference coordinate established by the Bureau International de l'Heure on the basis of the readings of atomic clocks operating in various establishments in accordance with the definition of the second, the unit of time of the International System of Units' (in 1988, the responsibility of TAI was transferred from the Bureau International de l'Heure to the Time Section of the Bureau International des Poids et Mesures).

This TAI definition adopted by CGPM in 1971 was enlarged in 1980 by CCTF (at that time the name was CCDS—Consultative Committee for the Definition of the Second) in order to take in to account relativity effects: 'TAI is a coordinate time scale defined in a geocentric reference frame with the SI second as realized by the rotating geoid as the scale unit'.

In this situation, the relation between the TAI scale unit and a given realization of the SI second as given by an atomic clock depends on the position of that clock. For clocks at rest, and situated at sea level, the scale unit of TAI is equal to the unit of time, the SI second, as realized locally. The scale unit of TAI appears to be longer by 2.2×10^{-13} s if compared with a clock kept at an altitude of 2000 m, owing to the gravitational redshift. A complete theoretical definition of TAI in the framework of general relativity can be found in the documents of the IAU General Assembly held in Buenos Aires, 1991 (Resolution A4); see also REFERENCE FRAMES AND TIME SCALES IN GENERAL RELATIVITY.

TAI is consequently an artificial TS, formed by combining with suitable algorithms the readings of atomic clocks kept in metrological laboratories or other scientific institutions in different geopotential conditions.

Characteristic of a time scale

The features that must be exhibited by a TS—as for most of the units and standards of physical quantities—are many and in some cases unexpected.

ATSs adequately fulfil some of these requirements but are completely lacking as regards others. The requirements are accuracy, precision, stability, universality (access availability) and perenniality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Some links between these properties, their scientific use and the relevant consequences are outlined in what follows.

Perenniality
The ATS being based on artefacts, this property cannot be satisfied. Any ATS given by an individual clock is bounded by the technical life of that clock and consequently cannot be extended in the past and in the future, as is indeed possible—at least in principle—with all the astronomical TSs.

It is worth noting that the ATS of a single atomic clock is not everlasting or perennial, but ATSs based on agreements, such as TAI itself or TT (Terrestrial Time; see also TIME SCALES: TERRESTRIAL), can be mutually linked to former TSs, in order to ensure perenniality.

The general lack of perenniality and the need to perform time links between clocks or TS calls for three features, as regards the formation of an ATS:

- redundancy of clocks, also inside a laboratory;
- availability of accurate time comparison systems, inside a laboratory and between remote laboratories, to perpetuate the 'reading' of a failing clock;
- geographic or topographic 'dispersion' of these clocks to reduce the consequences of any mishap regarding the clocks of a given laboratory.

These three characteristics to be satisfied inevitably have profound consequences for TS formation.

Universality
This characteristic had a physical meaning in the pre-relativity era; now the term is intended to mean that an ATS is accepted, accessible and made available to all classes of users. These three characteristics are studied and obtained by some international scientific unions (UAI, URSI), scientific–technical bodies such as CIPM, via its CCTF, and BIPM, via its Time Section, and technical–regulatory bodies, such as ITU-R (International Telecommunication Union—Radiocommunications).

Accuracy
The time 'length', or duration of the second used to construct an ATS, must conform as strictly as possible to the SI definition. This agreement, estimated to be around 10^{-9} in 1955, when the first atomic frequency standard was put in operation, is progressing by about 1 order of magnitude per decade. At the turn of the century this agreement reached an accuracy of some 10^{-14} and it is expected that an additional 1–2 orders of magnitude could be at reach. ATS accuracy can be considered to be satisfactory.

It is important to stress that this level of accuracy is available 'nearby' the clock and cannot easily be transferred to remote locations.

Precision
This is a measure of the ability to 'read' the date of an event. Precision depends on the characteristics both of the atomic clock (its phase noise) and of the instrument. Another obvious limitation exists if the reference ATS is to be remotely accessed via a link, rather than being available in the same laboratory where the timing is to be performed.

Stability
This characteristic is taken for granted (at least as regards the flow of time); no experiment so far performed has cast doubts on the temporal stability of Planck's constant or of fine or hyperfine constants.

No variation has been observed at the level of 10^{-14} yr^{-1}.

Time scales from pulsars or other astronomical objects
The lack of perenniality of a TS based on manmade devices has promoted, since 1975, the search for clock references in the universe; this research so far has offered no solutions to TS formation for scientific and technical timing but has provided some useful information about these extraterrestrial time reference sources.

Atomic time scale formation

General consideration on time scale characteristics and requirements
ATS formation is a very complicated affair and it is definitely not a trivial one. The reasons for these complexities can be grouped into at least five categories:

(a) final goals;
(b) characteristics of the clocks;
(c) characteristics of the comparison systems between the remote clocks;
(d) problems intrinsic to any averaging systems;
(e) how to evaluate relativity effects, black-body radiation effects, etc and how the uncertainty budget of primary clocks is established.

A few examples are given for the five categories.

(a) Should the final ATS strive for uniformity and stability or for accuracy?
These goals can be conflicting: navigation systems are more interested in stability; fundamental constants research is more interested in accuracy.
(b) Each kind of atomic clock is affected by different noise processes, very seldom 'white': see 'Noise processes affecting atomic clocks' below.
(c) The various comparison systems exhibit different 'errors' and noise processes, or phenomena linked to the propagation or environmental temperatures; the 'time constants' required to compensate for these effects can be in conflict with the 'time constants' inside the algorithms chosen to optimize accuracy or stability.
(d) The 'colors' (the spectral behaviour) of some noise affecting the clocks can prevent an easy statistical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

treatment; moreover, how should the time constant be used, what time constant should be used and how should the clocks contributing to the TS be weighted?

(e) The accuracy of a primary atomic clock cannot be measured with reference to another primary clock; it can only be estimated. To determine the error budget of one of these clocks is an operation based on the intellectual honesty of the experimenters.

Another problem, a technical and organizational one and typical of TAI formation, is how to collect timely data from some 200 atomic clocks, of different types, kept in some 50 laboratories scattered around the world. Similar problems are encountered for the TS of navigation systems or formed in national metrological institutions.

Formation algorithms

As an example, the formation of TAI, a two-step process, is described. Initially, using a weighted mean of all the clocks, a TS called EAL is formed, from the French échelle atomique libre—free atomic time scale. EAL is optimized for long-term stability and its frequency is, with a suitable time constant, corrected in a way to conform to the frequency of the cesium primary frequency standards. Adopting this two-step process, both stability and accuracy can be secured, clearly within some given limits. Obviously some criteria to be followed in establishing the weights of the individual clocks must be adopted.

The average TS EAL can be defined as follows:

$$\mathrm{EAL} = \frac{\sum_{i=1}^{N} p_i[h_i(t) + h_i'(t)]}{\sum_{i=1}^{N} p_i}$$

where $h_i(t)$ represents the reading of the ith clock , taken at time t, N is the number of atomic clocks, p_i is the statistical weight of the ith clock and $h_i'(t)$ is a suitable correction needed to offer continuity in time and in frequency in the case of the introduction or removal of a clock or the change in weight of a clock, while

$$x_{ij} = h_j(t) - h_i(t)$$

is the measure of the difference between the clocks.

A system of N equations in N unknowns is formed; the results are the wanted differences between the average EAL and the readings of the individual clocks.

Noise processes affecting atomic clocks

One of the subtlest problems in TS formation is the consequences for the reading of a clock due to various noise processes affecting the frequency and the phase of the real devices. This problem is of paramount importance in satellite navigation systems in which the mutual discrepancy between the some 50–100 atomic clocks forming a given system should be kept, at any time, below 1 ns. Half of these clocks are placed in orbiting satellites and must be kept on time by a complex system of monitoring stations, which are also required to determine, in real time and using time measurements, the position of each satellite with errors less than 1 m.

A number of problems, such as the choice of the orbit, the kind and number of atomic clocks kept on Earth or in space, the adoption of a given synchronization technique and the number and the geographic positions of the monitoring stations, are all linked to the noise processes affecting the clocks. The reading of any clock is indeed divergent as a result of the noise processes, and to keep this discrepancy against the TS formed by the entire navigation system inside 1 ns requires a frequency accuracy of the order of 10^{-14} and the most accurate synchronization techniques.

The law of divergence of the reading of an individual clock against the 'average' depends on the kind of noise and, in some cases, the divergence and its estimate are not linear with the elapsed time, as can be seen in table 1.

Table 1. Divergence of clock readings for various kinds of noise.

Kind of noise	Uncertainty prediction $u(t)$: law of proportionality to the elapsed time
White noise on phase	t^0
Flicker on phase	$[\ln(t)]^{1/2}$
Random walk on phase	$t^{1/2}$
Flicker frequency	t
Random walk on frequency	$t^{3/2}$

Uncertainties

Accuracy can be secured only by a comparison with primary cesium frequency standards, where 'primary' means that an uncertainty budget was obtained via a number of experiments and calculation.

At the present state of knowledge an accuracy of a few 10^{-15} seems within reach, provided that a suitable integration time is adopted.

Sigfrido Leschiutta

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Time Scales: Terrestrial

Terrestrial time is at present derived from atomic clocks. The SI second, the unit of time of the international system of units, has been defined since 1967 in terms of a hyperfine transition of the cesium atom and the best primary frequency standards now realize it with a relative uncertainty of a few parts in 10^{15}, which makes it the most accurately measurable physical quantity. INTERNATIONAL ATOMIC TIME (TAI) is based on more than 200 atomic clocks distributed worldwide that provide its stability, while a small number of primary frequency standards provide its accuracy. From TAI is derived Universal Coordinated Time (UTC) that is the basis of all legal time scales. To allow the construction of TAI and the general dissemination of time, clocks separated by thousands of kilometers must be compared and synchronized. This may be achieved with an uncertainty of a few nanoseconds in 1 day or less using radiotransmission techniques such as the Global Positioning System (GPS) or two-way time transfer using geostationary satellites. The definition of TAI must be considered within the framework of general relativity; it is a realization of a time coordinate in the geocentric reference system. In the future a new generation of clocks may provide, in the laboratory or on board satellites, an accuracy of 10^{-17} or less. When this happens it may be necessary to reconsider both the definition and the way of realizing TAI and new time transfer techniques will need to be developed to compare the new clocks.

Relativity: proper and coordinate time

Relativity distinguishes locally measurable (proper) quantities from coordinates which are, by definition, dependent on conventions. Therefore one should distinguish proper time, which is the output of an ideal clock, from coordinate time, which is one of the coordinates chosen to represent the four-dimensional space–time. In its Resolution A4 (1991), the INTERNATIONAL ASTRONOMICAL UNION explicitly introduced general relativity as the theoretical background for space–time reference frames (see also REFERENCE FRAMES AND TIME SCALES IN GENERAL RELATIVITY). For the geocentric system, it defined two time coordinates: the geocentric coordinate time (TCG) and terrestrial time (TT), which differs from TCG by a constant rate so that the scale unit of TT agrees with the SI second on the geoid. In this framework, it is clearly stated that TAI is a realization of TT (apart from a constant offset of 32.184 s). That relativistic point of view had already been adopted by the Consultative Committee for the Definition of the Second[1] (CCDS) which declared (1980) 'TAI is a coordinate time scale defined in a geocentric reference frame with the SI second as realized on the geoid as its scale unit'.

[1] Now named the Consultative Committee for Time and Frequency (CCTF).

Generation of an atomic time scale

According to a definition of the International Telecommunications Union, a time scale is an 'ordered set of time markers'. To form an atomic time scale, the efficient combination of the readings of the atomic clocks requires the definition of the expected qualities of the time scale, the characterization of the timing data and the design of an algorithm (see TIME SCALES: ATOMIC). In addition to the basic requirement that a single time marker be attributed to one event, one may consider metrological qualities for a time scale, such as its stability and accuracy, or practical qualities such as its reliability and delay of access.

Stability may be defined by the ability to keep the same frequency over a given time interval. A number of statistical tools have been devised to estimate the instability of a clock or time scale (or more generally any time series of measurements). Among them, the most commonly used is the two-sample, or Allan, deviation $\sigma_y(\tau)$, which estimates the fractional frequency instability over the averaging time τ. Figure 1 presents the frequency stability of various recent clocks and that of TAI for comparison. Accuracy, on the other hand, may be defined as the degree of conformity of a value to its definition. At present the SI second is defined by a hyperfine transition of the Cs atom, and only a few laboratory devices, the primary frequency standards, aim at realizing it with utmost accuracy. The most accurate primary standard is now a Cs fountain developed at the LPTF in Paris, which states a frequency accurate to 2×10^{-15}. All clocks that are not so specially designed may be very stable but display various degrees of inaccuracy. To be accurate, an atomic time scale must therefore be compared with one or several primary standards.

A single atomic clock may provide a time scale, but it is always subject to failure. To increase the reliability, redundancy is necessary. This may simply consist in replacing a failed clock by another one, after proper synchronization of the clocks during their common operation. However, combining the data of several clocks to form a time scale will improve the stability, accuracy and reliability with respect to those of a single clock. To reach this goal, a suitable stability algorithm must be used. This approach, which is adopted by many metrology laboratories and operation centers for navigation systems, has been the basis of the construction of TAI since its inception. While the readings of one atomic clock are available in real time, an ensemble time scale may be designed to be available in deferred time only, e.g. this is the case for TAI. Such a delay of access has some practical drawbacks, but it has two important advantages: it allows a better comparison of the participating clocks when they are in different laboratories, by averaging the noise of the time comparison technique, and also a better characterization of the stability of the clocks, when long-term stability is demanded.

Timing data are time differences between clocks. Time markers from a clock may take the form of electric pulses (e.g. 1 pulse s^{-1}) or, alternatively, the phase of a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

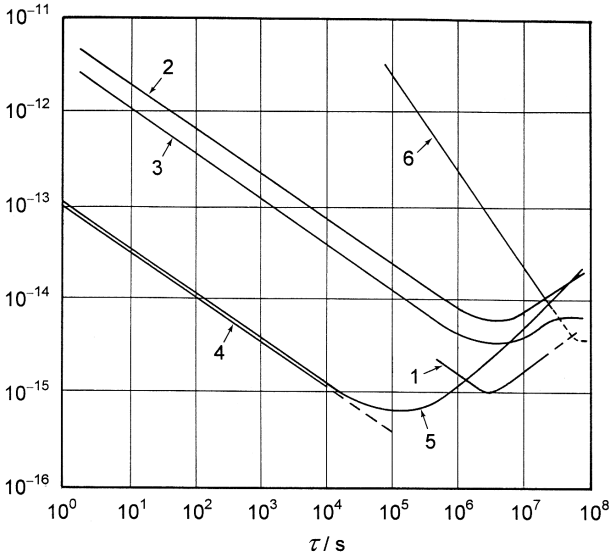


Figure 1. Relative frequency stability versus averaging duration τ for TAI (1), various atomic clocks (2–5) (2, commercial cesium clock; 3, primary clock PTB-CS1 (Braunschweig); 4, cesium fountain LPTF-FO1 (Paris); 5, commercial hydrogen maser) and the rotation of a best-case millisecond pulsar (6). Dashed parts correspond to expected or possible values.

periodic signal (e.g. 10 MHz) locked to the clock may be monitored. For clocks located at the same site, the time differences are measured by digital counters (for pulses) or heterodyne systems (for periodic signals), and each instantaneous measurement may have an uncertainty well below 1 ns. For distant clocks, time transfer techniques must be used (see below) and the typical uncertainty is of the order of a few ns for integration times of minutes to hours.

Given an ensemble of n clocks for which at some epoch t the time and frequency differences between clock i and the ensemble time TA(t) are designated by $x_i(t)$ and $y_i(t)$, respectively, then the set of n time differences $x_i(t')$ at a later time $t' = t + \delta t$ may be computed from the $n-1$ time differences measured between the clocks at time t' using an additional equation which expresses the frequency stability of the clocks as a weighted average:

$$\sum_i w_i(t')x_i(t') = \sum_i w_i(t')x_i(t) + \sum_i w_i(t')y_i(t')\,\delta t$$

where $w_i(t')$ represents the weight attributed to clock i and where the frequency difference at time t', $y_i(t')$, has to be estimated. Variations in the processes used to estimate the weights $w_i(t')$ and the frequencies $y_i(t')$ and in the choice of the time increment δt characterize the different stability algorithms. Usual features are to assume $y_i(t') = y_i(t)$ and to impose some constraints on the weights $w_i(t')$ such as a maximum weight for any given clock, whatever its apparent stability, or a null weight attributed to new clocks. Such constraints increase the reliability of the time scale at the expense of a small loss in its stability.

Construction of TAI and UTC, and some of their realizations

The first atomic clocks were built in the 1950s. When convenient techniques to compare distant clocks became available in the 1960s, it was then possible to realize an ensemble time scale. Such a task was undertaken by the Bureau International de l'Heure (BIH) in Paris and, when international agreement was deemed necessary in 1970, TAI was defined by the International Committee for Weights and Measures as 'the time reference established by the BIH on the basis of the readings of atomic clocks operating in various establishments in accordance with the definition of the second...'. In 1988 the responsibility of establishing TAI was transferred to the International Bureau for Weights and Measures (BIPM) in Sèvres (France).

As a deferred-time scale, TAI is computed each month using data provided by about 50 laboratories worldwide (figure 2). Timing data from about 220 clocks are combined using an algorithm named ALGOS, resulting in an ensemble time scale named EAL. The stability of EAL is, by its construction, optimized for an averaging time around 1 month, and this is estimated in 1999 to be about 0.6×10^{-15}. Out of the 220 participating clocks, about 120 (mostly high-performance, recently built, commercial Cs clocks) contribute to about 85% of the construction of EAL. Since EAL is free running, its scale unit (the 'EAL second') may be different from the goal set in the definition of TAI. The duration of the scale unit of EAL is measured by comparison with a number of primary frequency standards and TAI is then derived from EAL by applying a frequency correction so that the scale unit of TAI agrees with the SI second on the geoid. The relative frequency difference between EAL and TAI is at present (Spring 1999) 7.14×10^{-13} and may change slightly over time, as needed to ensure the accuracy of TAI. Over recent years, only eight primary standards have contributed occasionally or regularly to the accuracy of TAI, which is estimated to be below 0.5×10^{-14} over 1998–1999.

It has been known since the beginning of the 20th century that the EARTH'S ROTATION rate is irregular. Nevertheless, as a time based on the rotation of the Earth taking exactly 24 h (UNIVERSAL TIME UT1) had been used for decades, regulatory bodies tried in the 1960s to adapt the new atomic time to the former scale. This yielded in 1972 the definition of UTC: it differs from TAI by an integral number of seconds so that the difference between UT1 and UTC remains always smaller than 0.9 s. To ensure this, the International Earth Rotation Service in Paris issues predictions of UT1 − UTC and decides, when appropriate, to insert one leap second in UTC or alternatively to remove one second from UTC. Such operations are done at the last second of the month of June or December. From 10 s in 1972, the difference between TAI and UTC has increased to 32 s in 1999 and a leap second is inserted every 18 months on average. This is because the SI second has not been designed so that 86 400 of them match the average solar day.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

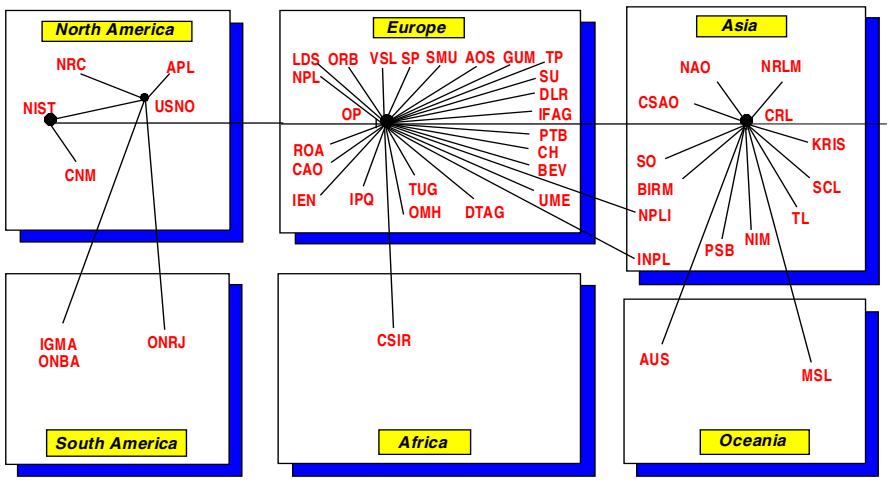


Figure 2. Schematic map of the laboratories and time links used in the computation of TAI (May 1999).

With UTC as the basis for all legal time scales, national metrology laboratories provide to their customers a real-time realization of UTC, denoted UTC(k) for laboratory k. Since UTC is not available in real time, UTC(k) may not exactly reproduce UTC, but modern clocks easily allow it to remain within ± 100 ns, a goal recommended by the CCDS in 1993. In contrast to UTC, the time scales UTC(k) must be physically realized, for example by an electrical signal at a given point, so that metrology laboratories may, in real time, drive time and frequency dissemination or comparison systems for national users. The civil time for a given country is then generally defined by adding or substracting an integral number of hours to UTC.

The leap second system, although very simple, has important drawbacks for those responsible for realizing continuous time references, e.g. for satellite systems or telecommunications. They may use TAI or generate their own time scale (such as GPS time) but, in any case, it is necessary to continue to trace the difference between this time scale and UTC. Some scientists have suggested abandoning the leap second system; there is, however, considerable inertia and such a step is unlikely to happen soon.

Clock comparison and time dissemination techniques

In the context of relativity theory, there is no *a priori* definition of simultaneity and therefore synchronization or, more generally, comparison of clocks is arbitrary. Since TAI is defined as a coordinate time, it is appropriate to define coordinate synchronization: two events are synchronous if they have the same time coordinate. In this framework, there are usually two ways to compare distant clocks: one method is to transport a portable clock from one clock to the other, properly accounting for the coordinate time accumulated during the transport (it differs from the proper time measured by the portable clock; the difference is usually called the 'relativistic corrections'); another method is to send an electromagnetic signal from one clock to the other, properly accounting for the coordinate time of propagation of the signal. Such procedures have been carried out since the development of atomic clocks. The distinction between proper and coordinate time, however, is only useful when aiming at an uncertainty below 1 μs for time comparison and below 1×10^{-12} for frequency comparison on Earth, which is what we consider in the following. At or above this uncertainty level, a number of time and frequency dissemination techniques have been and are still used, ranging from signals in the VLF (tens of kHz) or short-wave (tens of MHz) radiofrequency bands to the use of telephone lines or the world-wide web.

The clock transport technique is rather cumbersome but has been much used until the development of the GPS in the 1980s. It provided an uncertainty of some tens of nanoseconds, mainly limited by the instability of the transported clock. Although at present rarely used, an uncertainty of about a few nanoseconds is attained. Eventually, when the frequency stability of clocks reaches the order of 10^{-17}, an uncertainty of about a few picoseconds may be reached which would be a challenge for radiotransmission techniques.

All commonly used techniques for clock comparison rely on the transmission of electromagnetic signals. They can provide synchronization with an uncertainty of some nanoseconds or less within measuring times of a few minutes to a couple of hours and therefore achieve frequency comparison to within the present clock performances (10^{-14}–10^{-15}) in a matter of hours to a couple of days. The main technique uses signals from navigation satellites, such as the GPS or the Russian GLONASS, to compare the local clock with the clock on board the satellite. By differencing the measurements taken simultaneously at the two clocks ('common-view' processing) or at a whole network of stations ('geodetic-like' processing), the noise of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

satellite clocks is eliminated and the resulting uncertainty in synchronization is of the order of a few nanoseconds or less with proper averaging. The main remaining source of uncertainty originates in the hardware delays and their variation with the environmental conditions. The other technique is two-way time transfer in which signals are sent in both directions over the transmission path. The link may be a radio signal through geostationary communication satellites or a laser pulse through dedicated equipment on board satellites or optical fibers. In all cases, a two-way technique takes advantage of the reciprocity of the path to cancel or reduce some sources of error so that the measurement uncertainty is, in principle, somewhat lower.

Future trends

Important progress has been accomplished in recent years in new techniques promising the future development of ultrastable clocks that will bring new horizons to terrestrial time scales. Two main directions are being explored for these clocks: laser cooling of atoms and ion traps. It should be noted that the 1997 Nobel prize in physics was awarded to S Chu, C Cohen-Tannoudji and W Phillips for their work on the development of methods for cooling and confining atoms using laser light. The first primary frequency standard successfully using these techniques has been the Cs fountain at the LPTF in Paris which provided a few evaluations of the SI second with uncertainties of (2–3) $\times 10^{-15}$. It is expected that an accuracy of 1×10^{-16} may be reached with such a fountain and that a fountain using rubidium atoms may be more stable. Based on a slow beam of cold atoms, similar devices operating in space in zero gravity may reach an accuracy of 1×10^{-17}. Comparing such clocks at a distance will necessitate new developments for radiotransmission or laser techniques which are envisioned for the Atomic Clock Ensemble in Space (ACES) experiment planned on board the International Space Station. It will also be necessary to reconsider the definition of TT and TAI which are limited to about 1×10^{-17} by uncertainties in the definition of the geoid.

Bibliography

Further reading in time and frequency metrology:

Audoin C and Guinot B 1998 *Les Fondements de la Mesure du Temps* (Paris: Masson) (in French)

Guinot B 1989 Atomic time *Reference Frames in Astronomy and Geophysics* ed J Kovalevsky, I I Mueller and B Kolaczec (Dordrecht: Kluwer) pp 379–415 (the whole book may be read to obtain a broad knowledge in reference frames)

Levine J 1999 Introduction to time and frequency metrology *Rev. Sci. Instrum.* **70** (6) 2567

See also a collection of more specialized papers in

Sullivan D B, Allan D W, Howe D A and Walls F L (eds) 1990 *Characterization of Clocks and Oscillators* NIST TN 1337 (US Government Printing Office)

For time scale formation, specially TAI, see

BIPM 1998 *Annual Report of the BIPM Time Section* vol 11

Tavella P and Thomas C 1991 Comparative study of time scale algorithms *Metrologia* **28** 57

Gérard Petit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Timocharis (*c*. 280 BC)

Greek philosopher who, with Aristillus, prepared a catalog of star positions. The measurement of the position of Spica helped HIPPARCHUS to demonstrate precession.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tisserand, François Félix (1845–96)

Celestial mechanician, born in Nuits-St-Georges, Côte-d'Or, France, became director of the Toulouse Observatory and of the Paris Observatory and worked on the orbit of the Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Titan

Titan, SATURN's largest satellite, is a distant world: far away from the Sun and its close, warm TERRESTRIAL PLANETS. A fair distance from Saturn, too. Seen from Titan, at more than a million kilometers away, Saturn looks like a big yellowish ball, permanently girdled by an icy hoop. The Sun's disk is hard to see at a distance of close to one billion five hundred kilometers, more than nine times farther away than the Earth. As a consequence, our star is a hundred times dimmer there, and the temperatures prevailing on Titan drop to 70 K (−200 °C) in the atmosphere. This atmosphere is thick enough to make Titan an important but mysterious object in our solar system. Even from a near-by distance, as VOYAGER 1 found out in 1980, the dense cloud decks surrounding the small world obscure any glimpse of its surface, as on Venus.

The cameras of Voyager 1 afforded a close view of the satellite in 1980, from only 4394 km to the surface. Voyager 2 flew by Titan nine months later but at a distance a hundred times greater. The first images of the satellite that were returned in November 1980 were somewhat disappointing: Titan's visible appearance seemed unexciting—a reddish-brown ball, completely covered by a thick haze which allowed no visibility of the surface (figure 1). The most obvious feature seen by Voyager was a difference in the brightness of the two hemispheres, the northern side appearing darker than the southern. This difference is of the order of 25% at blue wavelengths, and falls to a few per cent in the ultraviolet and at red wavelengths. This north–south asymmetry is probably related to circulation in the atmosphere pushing haze from one hemisphere to the other. When Voyager 1 approached Titan, the brighter hemisphere had less haze than the darker one. Voyager also saw a dark ring above the north (winter) pole. This feature, termed the polar hood, extending from 70 to 90 degrees north latitude, is most prominent at blue and violet wavelengths, and may be associated with lack of illumination in the polar regions during the winter (since the subsolar latitude extends to 26.4°) and/or downwelling in global circulation. The altitude of vertical optical depth unity is of the order of 100 km. The asymmetry has been observed to reverse—when the HUBBLE SPACE TELESCOPE (HST) first observed Titan in 1990, a little over a quarter of a Titan year after the Voyager encounters, the northern hemisphere was found to be brighter than the south. Whereas Voyager only observed up to red wavelengths, HST has since imaged Titan in the near-infrared. At these wavelengths, the asymmetry is reversed, and indeed is somewhat stronger than in the visible. This is due to the wavelength dependence of the atmospheric brightness (bright at short wavelengths due to Rayleigh scattering, dark in the near-infrared due to methane absorption) and of the haze (presumed to be dark in blue and bright at red and longer wavelengths, by analogy with 'tholin' material generated in the laboratory). The word tholin was first used by Carl Sagan and comes from the Greek word, meaning 'muddy'.

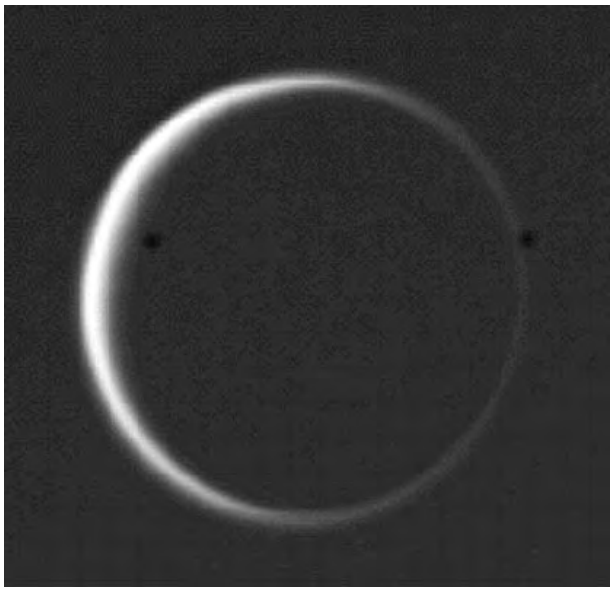


Figure 1. This view of Saturn's largest satellite was taken from a position of Voyager 2 looking almost directly back toward the Sun and the Mercury-sized body. Titan's hydrocarbon-rich nitrogen atmosphere scatters sunlight in a forward direction over the limb around the entire disk. The surface of Titan is obscured due to a deep cloud layer. The diameter of Titan is 5150 km. (Courtesy NASA/JPL).

The radio-occultation experiment aboard Voyager 1 provided a precise value of Titan's surface radius: 2575 ± 2 km, with a surface temperature of 94 ± 2 K and a pressure of about 1.45 bar. The main properties of Titan, as established at the time, are listed in table 1. The atmosphere is mainly formed of nitrogen (a property shared only by the Earth) and methane (molecular oxygen is lacking). Some 4 billion years ago, our Earth, like Titan, may have been shrouded in a thick atmosphere of nitrogen, deprived of oxygen and rich in methane. This distant object is then indeed tantalizing! Far from the Sun and its familiar small planets, this little frozen primitive Earth sails alone, lost in the middle of gaseous GIANT PLANETS and icy satellites. Where did Titan come from?

Origin and first insights

With a size half of that of the Earth's, Titan was challenged from its position as the largest satellite in our solar system by GANYMEDE, Jupiter's biggest satellite, found by Voyager to possess a few kilometers more in diameter than Saturn's champion. The latter is, however, still the ninth largest object around our star (by far bigger than the Moon and Pluto and barely bigger than Mercury), tenth by its mass (much lighter than Mercury: 140 billions of billions tons for the satellite against 330 billions of billions tons for the small planet). The satellite's mass, made of a mixture of water ice and rock (with the ice apparently containing even greater amounts of frozen, trapped and adsorbed gases

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Orbital and physical properties of Titan and the Earth.

	Titan	Earth
Semi-major axis of orbit (AU)†	9.54	1
Sidereal period of revolution (yr)‡	29.46	1
Orbital period around Saturn (day)	15.945	1
Equatorial radius		
Surface	2575 ± 2	6378
Clouds (km)	2775 ± 2	6393
Orbital eccentricity	0.029	0.017
Inclination of the orbit to the ecliptic plane (°)	19′8	0
Oblateness	<0.014	0.0034
Obliquity (°)§	26.7	23.45
Volume/$V_\oplus$ ‖	0.065	1
Mass/$M_\oplus$¶	0.022	1
Mean density (10^3 kg m^{-3})	1.881	5.52
Bond albedo	0.2	0.36
Surface gravity (m s^{-2})	1.35	9.81
Surface (equator)		
Pressure (bar)	1.45	1.103
Temperature (K)	94 ± 2	288
Atmospheric composition	N_2, CH_4, H_2, Ar?	N_2, O_2, Ar, H_2O, CO_2

† 1 AU $= 1.496 \times 10^{11}$ m.

‡ 1 sidereal year (yr) $= 31.558\,150 \times 10^6$ sec $= 365.256\,36$ d $= 1.000\,039$ tropical year.

§ The obliquity is the inclination of the equator to the orbital plane.

‖ Volume of Earth $= 1.070 \times 10^{21}$ m^3.

¶ Mass of Earth $= 5.98 \times 10^{24}$ kg.

than the other icy satellites, as the distance from the Sun increases), gives Titan a mean density of about 1.90.

Traveling on an almost circular prograde orbit, in the ecliptic plane, Titan takes a little under 16 days to complete its trip around Saturn, 1.2 million kilometers away, three times the Earth–Moon distance. The orbit of Titan about Saturn is significantly eccentric, with $e \approx 0.028$, which is too large to be due to perturbations by existing bodies in their present locations and may have been due to the relatively recent impact of a large body such as a comet. Titan is located in the outskirts of the Saturnian system along with Hyperion, Phoebe and Iapetus, the other 'strange' icy satellites (see SATURN: SATELLITES). But Titan is different: it has an atmosphere to boast of.

Why is Titan unique among the moons of the solar system in having a substantial atmosphere? It is perhaps not surprising that the other Saturnian moons are airless, because they are relatively small, but three of the Galilean moons of Jupiter, and Neptune's large satellite Triton, are comparable in size to Titan, while Ganymede is actually slightly larger. In contrast to the essentially airless nature of the other large satellites of the solar system, Titan has an atmosphere which is thicker than that of the much more massive planets Earth, Mars and Mercury.

In the giant planet's neighborhood, Titan is the only giant satellite, since the other seventeen known today are a hundred to a hundred million times lighter than it is. Titan therefore monopolizes more than 90% of the matter, including rings, surrounding Saturn. Its dimension and orbital parameters set constraints on the formation processes: the current consensus is that the giant planets and their satellites formed from 'sub-nebulae' much as the Sun and planets formed from the primordial solar nebula. In the outskirts of the Saturnian sub-nebula, where the temperatures were low (below 135 K) and the solar wind weaker, the gases that were trapped in the ice accreted by the satellite were not released until after the formation of Titan, becoming the satellite's earliest atmosphere. Ganymede and CALLISTO (Jupiter's satellites with similar densities to Titan, but with no atmosphere), contrary to Titan, were composed of ice that solidified at low temperatures but was later heated above 135 K in the vicinity of Jupiter, which radiated so much thermal energy during its formation that it must have glowed. At a mass three times larger that Saturn's, Jupiter generated correspondingly much more heat, so that the gas content in its satellites must have been markedly depleted through sublimation. Cometary accretion is an alternate explanation, investigated in some studies.

In the case of Titan, additional heat was generated later from the decay of short-lived radioactive elements in its rocky component. Similarly to the Earth and other terrestrial planets, this distant satellite must have formed a core of dense rocky material that became surrounded by a mantle of ice. Thus Titan recovered a secondary, outgassed atmosphere from the same general mechanism that produced the early atmospheres of Mars, Venus and Earth. Recent measurements of the isotopic ratios in Titan's atmosphere tend to favor outgassing from the interior rather than the cometary origin. In the cold and calm environment of its planet, at more than 1 billion kilometers from the Sun, Titan was able to preserve its atmosphere for more than 4.5 billion years; Mercury, the Moon, Ganymede or Callisto did not have that chance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

None of the above explains why Titan has more atmosphere than Mars or the Earth. Perhaps the atmosphere of Titan was frozen during the epoch when Mars' atmosphere, and to a lesser extent Earth's, was depleted by heavy bombardment. Alternatively, there may be a supply of the more volatile ices, such as ammonia, methane and carbon monoxide, trapped inside Titan by layers of frozen water and able to escape only slowly through cracks or 'cryovolcanoes'. It is also conceivable that Titan's atmosphere has been more efficiently replenished by incoming cometary material in relatively recent epochs, compared to the inner planets. All of these scenarios implicitly favor ammonia as the source of Titan's nitrogen atmosphere, a problem which is discussed in the next section.

The atmosphere: temperature and composition

Prior to the Voyager 1 encounter knowledge of Titan was scanty, albeit intriguing. Discovered on 25 March 1655 by the Dutch astronomer Christiaan Huygens, orbiting at a distance of about 1.2 million km from its primary, or some 20.6 Saturn radii, Titan was long known to have a substantial atmosphere: Catalan astronomer Jose Comas Solà claimed in 1908 to have observed limb darkening on Titan (that is a stronger attenuation of the light reflected from Titan's limb rather than by its center, usually associated with a thick atmosphere). To test this claim, Sir James Jeans decided in 1925 to include Titan and the biggest satellites of Jupiter in his theoretical study of escape processes in the atmospheres around solar system objects. His results showed that Titan could have kept an atmosphere, in spite of its small size and weak gravity, if low-temperature conditions, that he estimated as between 60 and 100 K, prevailed. In this case, any gas of molecular weight greater than or equal to 16 (such as ammonia, argon, neon, molecular nitrogen and methane, among those which were present in the mix of gas and dust particles that condensed to form the Solar System) would not have escaped Titan's atmosphere since the satellite's formation. In 1944, Gerald Kuiper discovered the spectral signature of methane on Titan by observing its absorption bands at 6190 Å and 7250 Å, but it remained uncertain whether it was the major atmospheric gas or a minor component mixed with another, as yet undetected but quite abundant, gas in the atmosphere.

Before the Voyager encounter, opinions were therefore divided on the principal constituent of Titan's atmosphere. One model favored methane as the main component (about 90%) of the atmosphere, and predicted surface conditions of $T = 86$ K at a pressure of 20 mbar. Another model expected that dissociation of ammonia would have produced molecular nitrogen in large quantities and held that the surface temperature and pressure would be quite high (200 K and 20 bars).

Chemical composition

Voyager 1 came to its rendezvous with Titan in November 1980. Molecular nitrogen turned out to be by far the major constituent in Titan's atmosphere, methane accounting for only a few per cent. The lightweight hydrogen gas rapidly escapes to space, because of Titan's low gravity; it is, however, being continuously generated as well, amounting to a relative abundance in the atmosphere of about 0.1%, representative of the balance achieved between its production and loss. Titan moves in and out of Saturn's magnetosphere (see SATURN: MAGNETOSPHERE INTERACTION WITH TITAN). The satellite itself bears no significant magnetic field, so that ions and electrons from Saturn can interact directly with its upper atmosphere, affecting neutral gases. As they become ionized, some of them may be swept away by the magnetic field, thus causing the loss of more gas (other than hydrogen) to space. The infrared spectrometer on board Voyager 1, IRIS, sounded the satellite's stratosphere (that is the part of the atmosphere between roughly 200 and 500 km of altitude). The infrared data taken by IRIS confirmed the presence of some simple hydrocarbons (such as methane, acetylene, ethylene and ethane) and proved the existence of some more complex ones at a few tens ppb: diacetylene (C_4H_2), methylacetylene (C_3H_4), propane (C_3H_8) and monodeuterated methane (CH_3D). Also, the signatures of some nitriles were found in the spectra: HCN, HC_3N, C_2N_2 and C_4N_2 (only in the solid phase). Finally, carbon dioxide (CO_2) was identified at 667 cm^{-1}. By comparison with laboratory spectra the abundances of these species were estimated at a few ppb. Today we have established the existence, from ground-based observations, of some other compounds in the atmosphere, including CO and CH_3CN (acetonitrile). N_2, H_2 and carbon monoxide are expected to be uniformly mixed throughout the lower atmosphere. Titan was observed again in 1997 with the Infrared Space Observatory (ISO) in the infrared range (figure 2), and from these observations more precise abundances for the atmospheric components on a global average were derived.

Methane should also be uniformly mixed globally, but since it may undergo phase changes in the troposphere, it is expected to vary with altitude. Thus, its mixing ratio (which is still poorly known) is supposed to be constant above the condensation level, below which methane may be saturated and could form clouds. Such clouds may be necessary to account for the observed opacity in the spectra between 200 and 400 cm^{-1}. The methane abundance can therefore increase with temperature down to the surface. Ultraviolet light from the Sun converts methane into more complex hydrocarbon molecules and hydrogen gas. The hydrocarbons combine with nitrogen to form nitriles. All these organics remain on Titan, eventually condensing at low altitudes. After methane, the next most abundant hydrocarbons are ethane, acetylene and propane. Unless supersaturation conditions prevail, each of these organics and the less abundant ones must condense at some level in the lower stratosphere and precipitate out, effectively reducing its gas phase abundance to smaller amounts below this level.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

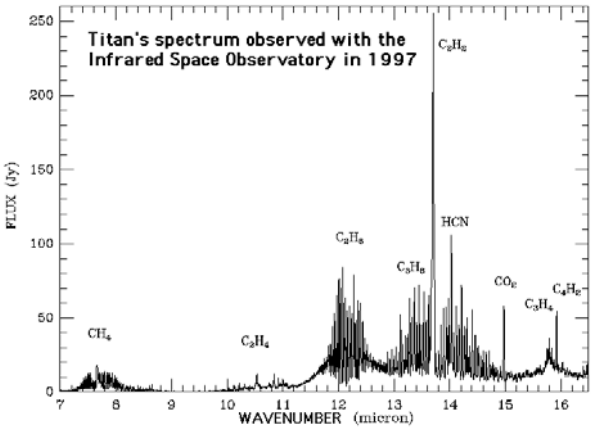


Figure 2. Spectrum of Titan, showing the signatures of the atmospheric components (hydrocarbons, nitriles and CO_2), acquired with ISO Short Wavelength Spectrometer in January 1997 in the grating mode, with a resolving power between 1500 and 3000.

Latitudinal variations in the atmosphere were observed for some of the least abundant molecules, with an obvious enhancement of HC_3N and C_2N_2, as well as of some hydrocarbons, at high northern latitudes, and seasonal effects were suggested as the origin of these variations.

Carbon monoxide (CO) and carbon dioxide (CO_2) along with water are currently the only oxygen compounds found in Titan's atmosphere. CO was first detected in 1983 from ground-based observations in the near-infrared, with a tropospheric abundance of $\sim 6 \times 10^{-5}$, in fair agreement with the prediction of the photochemical models in existence. CO_2 was identified in the Voyager 1/IRIS spectra from the emission feature around 667 cm^{-1}. Also observed with ISO (figure 2), the CO_2 average mole fraction in Titan's atmosphere above the 110 mbar level is found to be the same as that found by Voyager, about 10 ppb (assuming a vertical CO_2 profile constant with height), while the associated mixing ratio of stratospheric CO, measured from ground-based microwave and millimeter observations, is on the order of (1–5) $\times 10^{-5}$.

The recent detection of water in Titan's atmosphere with the ISO was an important result because it explains the presence of CO and, especially, of CO_2. The water vapor abundance is on the order of a few tens of a ppb at an altitude of 400 km. CO is produced by reactions of OH with CH_2 or CH_3, if the latter is present at sufficiently high levels in the thermosphere. OH also destroys CO, at lower levels near 500 km, with subsequent re-formation of CO_2. At equilibrium, the photochemical lifetime of CO is very long (comparable to the age of Titan itself), and as a consequence, the production of CO_2 from H_2O is in fact independent of the amounts of CO present in the stratosphere. The inferred value for the water influx on Titan is comparable with recent estimates from water detection on Saturn by ISO (5×10^6 cm^{-2} s^{-1}). Further progress on the whole question of oxidized molecules in Titan's atmosphere awaits new and more precise values for their abundances, which will probably come from Cassini/Huygens measurements in 2004.

The ultimate sink of the oxygen, coming in at the top of the atmosphere as H_2O, is at the surface of Titan, where it precipitates out as frozen CO_2. The lower atmosphere is, of course, in contact with the ground. The upper part or crust of the solid body of Titan is expected to consist mostly of water ice, although whether this is in direct contact with the lower atmosphere will depend on the nature and thickness of any coating of other condensates (solid CO_2, for example, or the hydrocarbon-based liquids and solids discussed elsewhere). If there is good contact, the troposphere could be saturated with water, albeit at temperatures so low that this corresponds to a very low mixing ratio (which explains why the icy rings and satellites of Saturn have survived for billions of years without vaporizing). The relatively warm upper atmosphere, which could hold proportionally more water, is probably prevented from doing so (unless there is an external source) because of the cold-trap formed by the tropopause. It is for the same reason that the stratosphere of the Earth is very dry. In other words, the water ice on the surface does not participate in the atmospheric photochemistry. If Titan was set closer to the Sun, warmer surface temperatures would cause the water to sublimate and move into the atmosphere. As a result, the oxygen provided in this way would rapidly combine with the methane and other hydrocarbons and form large amounts of carbon dioxide, similar to Venus and Mars. The character of Titan's atmosphere would radically change, but not necessarily towards something closer to the Earth.

Origins of the atmosphere

It is intriguing that Titan's primary atmospheric component is the same as the Earth's. What is the origin of the nitrogen in Titan's atmosphere? In particular, is it derived from a source containing molecular nitrogen or was it produced by chemical processing of ammonia (or some other process)?

There seem to be four possible scenarios. The first, and least probable, is that direct capture of N_2 from the solar or proto-Saturnian nebula took place. Even if this did occur, it is believed that Titan's primitive atmosphere, like that of the Earth, was lost long ago as part of the formation process, and replaced by a 'secondary' atmosphere which outgassed from the interior. This, however, could have involved devolatilization of nitrogen-containing ices condensed from the nebula.

It seems more likely that the primordial cloud contained nitrogen primarily as the hydride, ammonia, which is also far less volatile and therefore easier to accrete as early Titan cooled. If we imagine, as our third possible scenario, a variant of this in which Titan's atmosphere formed from the same mix of gases which are present on Saturn, the heavy components, primarily ammonia and methane, would be retained while

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hydrogen and helium were lost (on Titan, but not on Saturn). Photolysis would eventually produce N_2 from the NH_3 and higher hydrocarbons—including those which condense as aerosols—from the CH_4, so we could end up with the atmosphere forming much as we see it today. There is no first-order difference whether the gases came from the interior, from ices frozen on the surface, or directly from the nebula. Titan should then have the same isotopic composition as Saturn, for the ratio of deuterium to hydrogen in particular, possibly somewhat modified after accretion by the impact of volatile-rich comets, or by other processes such as possible impacts from debris resulting from the apparent damage to HYPERION.

Deuterium (D), the first isotope of hydrogen, has a neutron in addition to the proton of the common hydrogen atom. Deuterium is therefore twice as heavy as hydrogen and has, accordingly, a different physical behavior. By studying its abundance today, we may learn something of the chemical processes involved in the atmosphere of a planet in the past. On Titan, most of the hydrocarbon is in the form of methane (CH_4). Recent measurements of the D/H ratio on Titan with ISO (7.5×10^{-5} derived from the CH_3D isotope of methane) are lower by a factor of 4 or more than those measured in water in the comets Halley, Hyakutake and Hale-Bopp, and therefore tend to rule out a pure cometary source for the atmosphere of Titan. It seems thereby that the Titan D/H ratio is consistent with the formation of the atmosphere by outgassing from the interior of the satellite. On the other hand, D/H on Titan is somewhat larger than the value for Saturn (representative of the protosolar nebula), where D/H is believed to be around 2.5×10^{-5}. This suggests that Titan's present-day atmosphere has been somewhat enriched in deuterium since it formed from the proto-Saturnian nebula, by cometary material or in some other way: some scenarios suggest that methane in the sub-nebula of Saturn could have been enriched in deuterium, relative to that in molecular hydrogen, prior to the formation of the satellites by a factor of about 1.4. In addition, it is possible that some fractionation, mainly due to the photodissociation of methane, could have occurred during Titan's history.

The basic question of whether the nitrogen arrived as N_2 or NH_3 could find an answer in measurements of the Ar/N_2 ratio in the present-day atmosphere. Because molecular nitrogen and argon have similar vapor pressures, a large argon abundance suggests N_2 while a small one favors NH_3 as the source of the present-day nitrogen. Thus knowing whether argon is abundant or not on Titan, would enable us to choose between various formation models of the satellite. Unfortunately, none of the noble gases has yet been detected on Titan. The presence of argon has been deduced only indirectly—the value of the mean molecular weight of Titan's atmosphere is uncertain enough to allow for up to 7% or none at all, in which case the nitrogen fraction would increase, and it would seem more likely again that nitrogen arrived as ammonia on Titan.

Aerosols

Titan's atmosphere is known to contain particles (aerosols) in addition to gas. As Voyager 2 flew in the satellite's shadow, Titan's atmosphere, scattering solar light, formed a bright crescent almost entirely circling the dark disk (figure 1). The particulate haze around Titan extends to an altitude of about 200 km, with a detached and much more tenuous layer about 100 km higher. In the Voyager images the two layers appear blended. The haze on Titan is not very dense; it is, however, deep enough to completely hide the surface in the visible.

It is established today that the aerosols in Titan's atmosphere are condensed forms of the gases mentioned above. Condensed ethane, for instance, falling through Titan's atmosphere, may form clouds or haze, or even possibly near-surface mist. Theoretical considerations suggest that thin, patchy clouds of methane crystals and a photochemical fog of more complex condensates might both be present, the latter arising as a result of methane photolysis at ultraviolet wavelengths following which the fragments combine to produce polymers which condense to form oily droplets at Titan's temperatures. The brownish tarry organic sludge produced by photochemistry in Titan's atmosphere condenses near the tropopause. Although the exact nature of these compounds is not known, their size, from Voyager observations, seems to be on the order of 1 micron in diameter. The aerosols are then too heavy to remain suspended in the atmosphere: they must eventually fall on the surface, where they may be exposed or hidden in subsurface reservoirs. Some of these molecules, like hydrogen cyanide (HCN), are pre-biotic, similar to those produced in experiments on the origin of life on Earth. Existing models designate these same products of methane as the precursors of organic molecules and hence amino acids and, ultimately, the first living cells in primitive Earth's warm oceans. However, more work is needed to show how well the products of laboratory simulations ('tholins') match the aerosols in Titan's atmosphere.

Temperature

By analogy with the Earth, Titan's atmosphere is subdivided into layers defined by the temperature variations with height (or pressure, through the hydrostatic law). On both Titan and the Earth, the mean temperature profile is characterized by two temperature inversions (locations above which the temperature increases with altitude, while the opposite takes place beneath). Much of the electromagnetic radiation emitted by the Sun and reaching a planet is at wavelengths in or near the visible part of the spectrum. The aerosol layers scatter the photons and reflect a portion of the radiation back to space, some of it is absorbed by the haze and clouds, if any are present, and the atmospheric gases. The rest of the energy, about 10% on Titan, reaches the ground where it is absorbed.

However, the analogy between the two atmospheres shouldn't be carried too far: for one thing, the vertical extent of each region is different, with Titan's atmosphere generally being more extensive than the Earth's. In fact,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

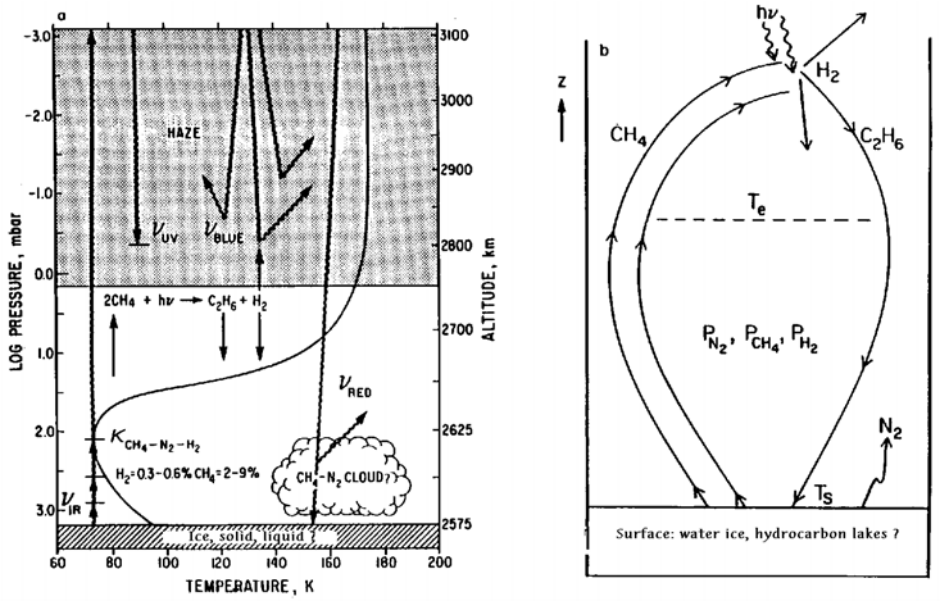


Figure 3. Schematic of Titan's surface–atmosphere system. (Figure adapted from Thompson and Sagan (1984).)

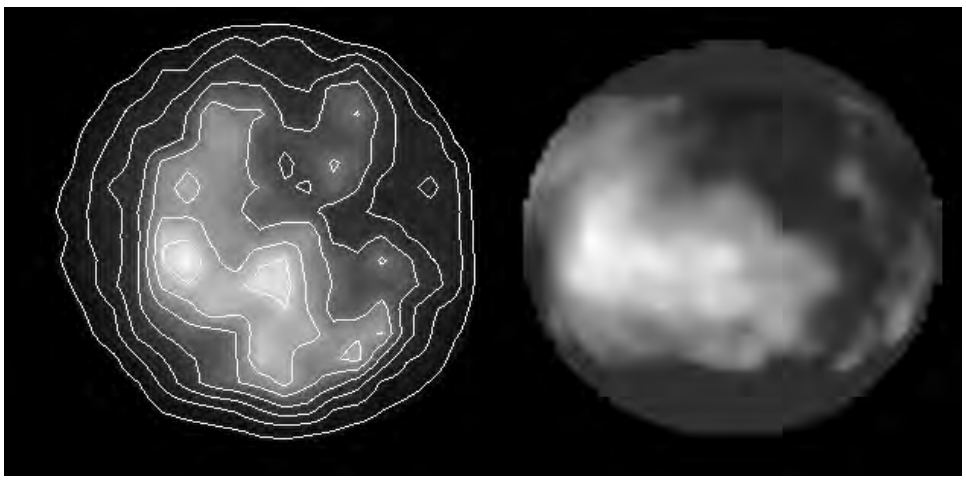


Figure 4. Images of Titan's surface at 0.94 μm (right) and 2 μm (left), obtained with the HST and ADONIS/ESO respectively in 1994. Note the large bright equatorial region which is located on the leading hemisphere. (Courtesy of M Combes and P Smith.)

the extent of Titan's atmosphere is comparable to the radius of the solid body (2575 km), while that of the Earth's is sometimes likened to the skin on a grape. The Earth's tropopause (location of the lower temperature inversion) varies in height over about 6 km with latitude, being highest (around 16 km) in the tropics where solar heating is greatest, but it is generally quite a sharp feature in the temperature profile everywhere (see EARTH'S ATMOSPHERE). On Titan the troposphere (lower part of the atmosphere) extends from the surface up to about 40 km, while the stratosphere (middle atmosphere) reaches up to altitudes of 500 km.

Titan's temperature with height was determined near the equator through a combination of data from the radio occultation experiment and the infrared spectrometer on the Voyager spacecraft. The results show that Titan affords a GREENHOUSE EFFECT in its troposphere (the lowest 40 km). This phenomenon, similarly as on Earth, causes the temperature to increase from a minimum (~71 K) at the tropopause (the upper boundary of the troposphere at 40 km) and down to the surface (94 K). This is a consequence of a region of heating in the middle atmosphere of both Titan and the Earth, due to the Earth's stratospheric ozone layer in one case and Titan's deep aerosol layer in the other. Both of these absorb solar energy and heat the troposphere, producing a local

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

temperature maximum at the surface (figure 3). The surface temperature is then about 12 degrees higher than the effective temperature of the planet. The upper boundary of the troposphere, the tropopause, is the level where the minimum temperature is registered and the overlying atmosphere is of such a low density that a substantial amount of radiative cooling to space can occur in the thermal infrared region of the spectrum. At the tropopause, an inversion occurs and the temperature increases with height, peaking at about 180 K at around 400 km of altitude. This turnaround of the temperature is primarily due to the ultraviolet sunlight entering Titan's upper atmosphere. Its absorption contributes to the photochemical processes described below.

Little is known with certainty about Titan's troposphere. It would seem that it is different from the Earth's. While on Earth the troposphere is dominated by convection, on Titan the observed lapse rate is stable against dry convection and there is no real indication of moist convection. If radiative equilibrium prevails, the presence of a greenhouse effect is, however, a puzzle. There may still be a small convective region in the deepest few kilometers of Titan's thick atmosphere to explain how the energy deposited at the surface can escape to an optically thin level.

The gases existing in Titan's atmosphere condense at the tropopause temperature, with the exception of hydrogen, carbon monoxide and diacetylene. These condensates form bigger and bigger aggregates in the lower part of the atmosphere and give rise to haze and possibly clouds. The end-products, possibly polymers, are finally deposited on the surface.

Furthermore, there are seasonal variations on Titan, causing the atmosphere to change its compositional and thermal structure as the moon follows Saturn on its orbit around the Sun, approaching our star at 1.35 billion km and recoiling to 1.5 billion km away. As the four seasons follow one another, each lasting about 7.5 years, the temperature on Titan could well vary by 20 degrees or more. Meteorological phenomena, such as hydrocarbon rain or snow, might then well occur on this satellite, shaping a curious landscape.

The surface

Titan is the last object in our solar system whose surface remains largely unknown to us. The surface pressure on Titan is strangely high for a rather small world: 1.5 bar, versus 1 bar of pressure on the Earth's surface. The temperature, with so little solar flux reaching the ground, is dreadfully low: about $-180\,^{\circ}$C. It is not easy then to imagine the real landscape hidden under the thick Titan clouds. On Earth, at such temperatures, everything would be frozen. On Titan, however, where oxygen does not exist in a free state and where water is necessarily trapped in the form of ice on or underneath the surface (similarly to EUROPA), it is perhaps methane, ethane and nitrogen that can play the part of active liquids. At the pressures and temperatures prevailing on Titan's surface, these three gases, so difficult to maintain in liquid form on Earth, are close to their condensation conditions, where they can coexist in all three forms: solid, liquid and gas.

Figure 5. This artist's rendering depicts the NASA Cassini spacecraft in orbit around ringed Saturn (lower right background). At the lower left, the European Space Agency's Huygens probe descends to the surface of Saturn's moon Titan (in foreground). Cassini was launched on a Titan IV rocket in October 1997, with Saturn arrival in June 2004. (Courtesy NASA/JPL.)

Real observational data on the surface of the satellite are scarce, but not totally non-existent. On the one hand, nothing is visible in the ultraviolet range either from the ground or from space. On the other hand though, radar sounding, spectra and images taken in the infrared and photometric measurements never fail to supply fuel for discussions and speculations. Also, some theoretical models exist constrained by known physical conditions such as mass, density, chemical composition and temperature, vapor pressure, etc. As an example, condensed ethane (the main by-product of methane) can be found in liquid form on Titan's surface. It was thus first suggested that Titan's surface could be entirely covered with ethane–methane oceans.

Since 1994, however, observations from space (with the Hubble Space Telescope), and from the ground (using the new techniques of ADAPTIVE OPTICS in order to compensate for the atmospheric turbulence), have managed to pierce through Titan's clouds by aiming at specific frequencies in the near-infrared range, where atmospheric 'windows' (or regions where the methane absorption is weak) allow us to sound the lower part of the atmosphere and the surface of the satellite. From these images, a new picture of Titan's surface emerged: dark and bright areas cover the surface of the satellite (figure 4). These images confirm spectroscopic observations from 1994, which indicated a brighter leading hemisphere and a much darker trailing side on Titan. Thus, Titan's surface is certainly inhomogeneous, and cannot be covered with uniform hydrocarbon oceans. The observations allow though for a number of options, such as local

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

hydrocarbon lakes and rocks to exist on the surface and for patchy methane clouds to form in the lower troposphere. The most recent and somewhat speculative model of Titan's surface produces an exciting image: Titan would be covered mainly with water ice and organic deposits, but the morphology would allow for frosty relief, perhaps mountains or craters covered with methane snow, their shores beaten by the waves produced in the vast hydrocarbon lakes ....

When the ambitious scientific CASSINI MISSION/Huygens (figure 5), launched successfully from Cape Canaveral on 15 October 1997, reaches this distant world, in 2004, astronomers will discover the true nature of this little hibernating moon, and of its ground in particular, from where its planet Saturn is barely visible. While the orbiter, Cassini, developed by NASA, performs its four-year thorough investigation of the Saturnian system, with numerous flybys scheduled for Titan, the probe Huygens, constructed by the European Space Agency, will descend into Titan's atmosphere and should hopefully land on its surface, giving us access to the thrilling results of what might be the last big space mission for years to come. In the meantime, one can always dream of the science fiction landscapes which characterize this frozen, dark world and which will then be revealed to us.

Bibliography

Coustenis A and Taylor F 1999 *Titan: the Earth-like Moon* (Singapore: World Scientific)

Hunten D M, Tomasko M G, Flasar F M, Samuelson R E, Strobel D F and Stevenson D J 1984 Titan *Saturn* ed T Gehrels and M S Matthews (Tucson, AZ: University of Arizona Press) pp 671–759

Lorenz R D and Lunine J I 1997 Titan's surface reviewed: the nature of bright and dark terrain *Planet Space Sci.* **45** 981–92

Athena Coustenis

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Titania

A mid-sized satellite of Uranus, discovered by William Herschel in 1787. Its diameter is 1580 km and it orbits at a distance of 191 000 km. The part of Titania imaged by Voyager 2 consists of a large plain covered with small craters, crossed by large canyons where the crust has faulted. The most extensive of these is the Messina Chasmata system, with an overall length of nearly 1500 km and 100 km across at its widest point. The 326 km diameter Gertrude is the only large crater; the general lack of large craters may indicate that Titania has been resurfaced in the past. The surface resembles that of Ariel, another Uranian satellite, and the processes that operated there are probably those that have shaped Titania's surface.

See also: Ariel, Uranus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Titius, Johann Daniel (1729–96)

German astronomer, born in Konitz, Germany (now Poland), professor of physics at Wittenberg, built the first lightning conductor. He interpolated into the translation he had prepared of a book by the naturalist Charles Bonnet the relation between the distances of the planets that he had discovered. The planets, Titius pointed out, were at distances from the Sun that were in proportion to the sequence 4, 4 + 3, 4 + 6, 4 + 12, 4 + 48, 4 + 96. Bonnet had commented in respect of the missing number in the sequence of 4 + 24 'But should the Lord Architect have left this space empty?' and the asteroids were later found to correspond to it. Titius' discovery was popularized by BODE and is known as *Bode's law* or, only sometimes, as the *Titius–Bode law*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tombaugh, Clyde William (1906–97)

Born near Streator, IL, became interested in astronomy as a boy and, on the basis of some observations which he sent to the Lowell Observatory, was hired as an astronomer there to search for 'Planet X' beyond Neptune. Photographing 65% of the sky, he discovered several star clusters, two comets, more than a hundred asteroids, dozens of clusters of galaxies—and the planet Pluto. He worked on the captured V-2 rockets at White Sands, and became professor at New Mexico State University, where he confirmed the rotation period of Mercury on its axis, determined the vortex nature of Jupiter's great red spot, and looked for small, natural Earth satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tompion, Thomas (*c.* 1639–1713)

Clockmaker, born in Northill, Bedfordshire, England, made the clocks for the Royal Observatory, Greenwich, the first clocks to keep Greenwich Mean Time and demonstrate the uniformity of rotation of the Earth. The clocks had 4 m pendulums, with a 2 s swing, and were the reason why the observing room (the Octagon Room at Flamsteed House) was designed by CHRISTOPHER WREN with such high ceilings. Tompion kept the cases locked and would not let JOHN FLAMSTEED have the key. After Flamsteed's death, the clocks were re-made with conventional meter-long pendulums and sold for domestic purposes, but one was recently restored to its original location.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Topological Defects in Cosmology

In modern theories of particle physics, an important consequence of phase transitions in the early universe is the formation of topological defects, that is domain walls, cosmic strings, monopoles or textures. These defects are essentially trapped fossils of a hot dense early phase of matter in the universe. They have well-known analogues in the low-temperature laboratory and in other contexts, including domains in a ferromagnet and vortices in superfluids and superconductors.

The potential existence of topological defects has both troubled and fascinated cosmologists since the late 1970s. On the one hand, domain walls and monopoles are incompatible with the standard hot BIG BANG cosmology because they would dominate the energy density of the universe today. Such considerations constrain the range of viable fundamental theories. On the other hand, cosmic strings and some other defect types are cosmologically benign and may have had an important role in the history of the universe. Because cosmic defects can be supermassive, a great deal of interest has focused on the idea that they were the primordial seeds for large-scale structure formation. This idea is now under observational pressure, but a contributory role in GALAXY FORMATION cannot be excluded. Defects can also form at lower energy scales and, in particular, those appearing at the electroweak phase transition are believed to have influenced the matter–antimatter asymmetry of the universe. In addition, cosmic defects have been invoked to explain the origin of missing DARK MATTER, ultra-high-energy COSMIC RAYS and a number of other outstanding cosmological enigmas. The place of topological defects in fundamental theory is secure, but their corresponding role in cosmology remains to be fully understood and empirically tested.

In what follows, the sections on 'Domain walls and the Kibble mechanism', 'Defect classification', 'Gravitational effects on cosmic strings' and 'Quantitative string networks' are more technical and may be omitted on first reading. These sections are indicated between ◇ and ⧫.

The origin of defects

Descartes' theory of vortices, in which the planets were swept around the Sun in a giant whirlpool, was probably the first proposed role for topological defects in cosmology. This vortex theory may have been swept aside by Newton, but defects cannot be eliminated from particle physics models which unify the forces of nature. The basic notion is that some large continuous group G (a Lie group) with many symmetries incorporates as subgroups those which describe electromagnetism and the strong and weak nuclear forces. At high temperatures, the particle interactions are unified in this large group G, so as the universe cools it undergoes a cascade of phase transitions at which the symmetries break and the forces of nature begin to differentiate, such as, $G \longrightarrow SU(3) \times SU(2) \times U(1) \longrightarrow SU(3) \times U(1)$ (see UNIVERSE: THERMAL HISTORY). At each stage of the symmetry breaking, at least one topological defect species must inevitably emerge, although which actual species does appear and for how long depends on the exact nature of the symmetry breaking. For example, at the first stage when the group G is broken, magnetic monopoles are formed. A monopole is a point-like object with a magnetic charge which is associated with the unbroken circular $U(1)$ symmetry of electromagnetism. GRAND UNIFIED THEORIES (GUTs) can be very large and may involve several intermediate symmetry breaking transitions, so a wide variety of cosmic defects can be created.

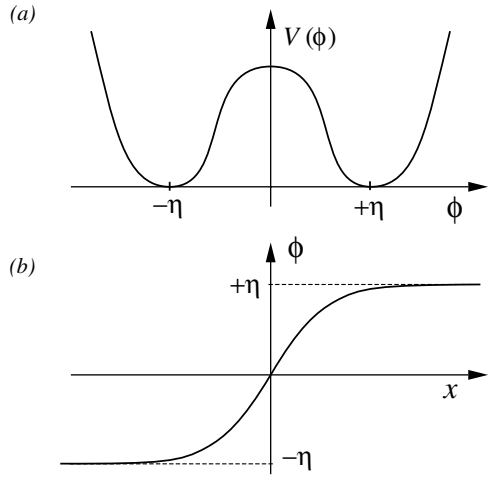


Figure 1. (*a*) Double-well potential $V(\phi)$ which breaks a discrete Z_2 reflection symmetry and has two minima at $\phi = \pm\eta$. (*b*) A field configuration in real space which interpolates between these two minima has a localized energy density corresponding to a domain wall.

◇ *Domain walls and the Kibble mechanism*

These points can be illustrated by considering the simplest case. Domain walls form when a discrete symmetry is broken as in the double-well potential $V(\phi)$ shown in figure 1. In this model, a real scalar field ϕ (or order parameter) takes particular values at every point in space. These values will, if possible, minimize the potential energy by being either at $\phi = -\eta$ or at $\phi = +\eta$, that is in one of the two minima (or vacua) of $V(\phi)$. However, if the field ϕ at $x = -\infty$ is at $\phi = -\eta$ while at $x = +\infty$ it is at $\phi = +\eta$, then ϕ must smoothly interpolate between these two minima as illustrated in figure 1. In doing so the configuration will rise up over the top of the potential $V(\phi)$ at $\phi = 0$ and so there will be a localized region of potential and gradient energies in the field ϕ. In three dimensions, this will be a two-dimensional domain wall separating regions in the two different possible vacua.

Domain walls inevitably form at any phase transition when a discrete symmetry breaks. This is simply due to causality and is known as the Kibble mechanism. At temperatures above the phase transition $T > T_c$ the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

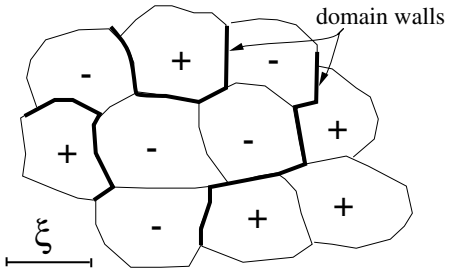


Figure 2. Topological defect formation at symmetry-breaking phase transitions—the Kibble mechanism: domain walls form between uncorrelated regions which fall into opposite sides of the double-well potential.

potential $V(\phi)$ has a minimum at the symmetric value $\phi = 0$; however, below the transition $T < T_c$ the field must randomly choose between the different discrete minimum energy states at $\phi = \pm\eta$. Regions with a volume associated with the correlation length of the field will take one random orientation or the other. However, by causality correlated regions are smaller than the horizon, so many domain walls must interpolate between the large number of regions which are necessarily misaligned. This formation process is illustrated in figure 2.

Defect classification

The classification of topological defects which form at symmetry-breaking phase transitions is most succinctly achieved using homotopy theory, but a simpler summary will suffice here. For a specific symmetry breaking scheme $G \to H$ in which the group G is broken to a subgroup H, then (i) domain walls form when G contains a discrete symmetry which is broken (e.g. $Z_2 \to I$ as above), (ii) cosmic strings form either if the unbroken subgroup H contains a discrete symmetry or if a $U(1)$ symmetry is broken (e.g. $U(1) \to I$, as in a superconductor), (iii) monopoles form when a $U(1)$ symmetry appears (e.g. $G \to SU(3) \times U(1)$ as in GUT) and (iv) textures appear when a non-Abelian group or subgroup of G is completely broken (e.g. $SU(2) \to I$ as in the electroweak phase transition). ♦

Properties of cosmic defects

The energy scale η at which the defect-forming phase transition occurs determines the key properties of a topological defect. A domain wall has a width $\delta \sim \eta^{-1}$ and a surface energy density $\sigma \sim \eta^3$ (in natural units $\hbar = c = k_B = 1$). The wall tension in the two tangential directions is equal to the surface density σ, so curved regions contract, oscillate and collide at relativistic velocities.

A cosmic string is a line-like topological defect associated with a trapped winding around a circle as illustrated in figure 3(*a*). The width of the cosmic string is $\delta \sim \eta^{-1}$ and its linear energy density is $\mu \sim \eta^2$. The string has a large tension equal to its linear energy density, so a string will oscillate relativistically. String intersections result in reconnection or the 'exchange of partners' which can lead to the break-up of a long string network into small loops. These loops have periodic trajectories and they decay by emitting gravitational radiation because they have a significant quadrupole moment.

In string models, $G\mu$ (or equivalently η^2) is the

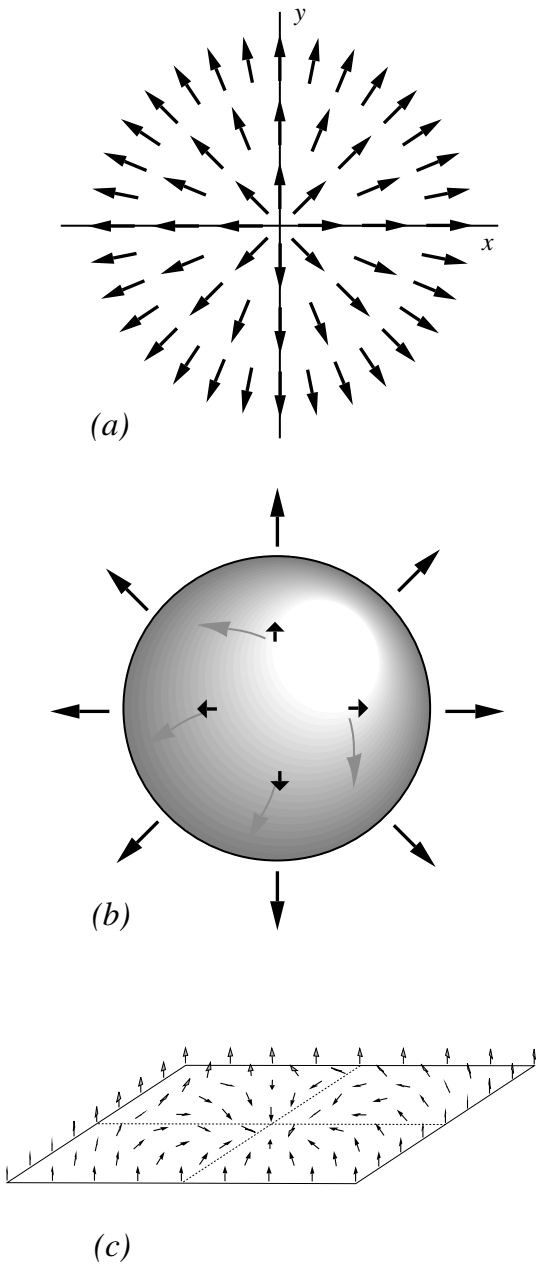


Figure 3. (*a*) Field configuration for a vortex string, corresponding to a complete winding of a complex scalar field around a unit circle (transverse section). (*b*) The 'hedgehog' configuration depicting the fields emanating from a magnetic monopole. (*c*) Two-dimensional texture configuration in which the field never becomes singular.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

key dimensionless parameter which determines the magnitude of string gravity, that is

$$G\mu \sim \left(\frac{\eta}{m_{\rm pl}}\right)^2$$

where $m_{\rm pl}$ is the Planck mass ($G = m_{\rm pl}^{-2}$). For GUT-scale strings with $\eta \sim 10^{16}$ GeV we are in a linearized gravity regime with $G\mu \sim 10^{-6}$. However, this corresponds to the enormous energy per unit length of $\mu \sim 10^{22}$ g cm^{-1}, that is, a 10 km length of cosmic string weighs as much as the Earth itself. This gives an indication of their potential cosmological role as seeds for structure formation.

In addition, strings can act like superconducting wires, if they possess trapped states of charged particles which propagate along the string. The phenomenon is a fairly natural consequence of many realistic symmetry-breaking hierarchies. Unlike ordinary cosmic strings which only act gravitationally, superconducting strings can have directly observable electromagnetic effects.

Magnetic monopoles are point-like objects with a radial magnetic field, as in the 'hedgehog' configuration shown in figure 3(*b*). The magnitude of this field is $B = 1/er^2$ and the overall magnetic flux is $\Phi_B = 4\pi/e$, where e is the electromagnetic gauge coupling. The mass of the monopole is typically $m \sim 4\pi\eta/e$, which for GUT scales can be as high as $m \sim 10^{18}$ GeV (that is, up to $m \sim 1\ \mu$g).

It is also possible to envisage a 'global' monopole without electromagnetic gauge fields. In this case, the scalar fields form a similar 'hedgehog' configuration but the gradient energies at large distances are not cancelled by the gauge fields. The global monopole energy is consequently linearly divergent with $m \sim 4\pi\eta^2 R$, where R is a cutoff set by the nearest antimonopole or the field correlation length. Global monopole–antimonopole pairs are strongly confined and will approach relativistically and rapidly annihilate, unlike their weakly interacting magnetic counterparts.

Textures are non-singular amorphous defects which form when non-Abelian symmetries are completely broken. A two-dimensional texture configuration is shown in figure 3(*c*). Like monopoles, they can be either local or global. For a local texture configuration, the scalar and gauge fields can be aligned such the texture has zero energy and it can be essentially 'gauged away'. Around the time of the texture-forming phase transition, however, the texture field alignments are imperfect and transitory topological effects can have important consequences. Transitions between texture states, known as sphalerons, can occur at these near-critical temperatures or, later, by quantum mechanical tunnelling. Such sphalerons in the standard electroweak model cause baryon number violation which may have influenced the matter–antimatter asymmetry of the universe.

In contrast, global textures have no gauge fields to cancel the effects of their long-range scalar fields. However, global textures are susceptible to a shrinking instability, so they collapse relativistically to a point.

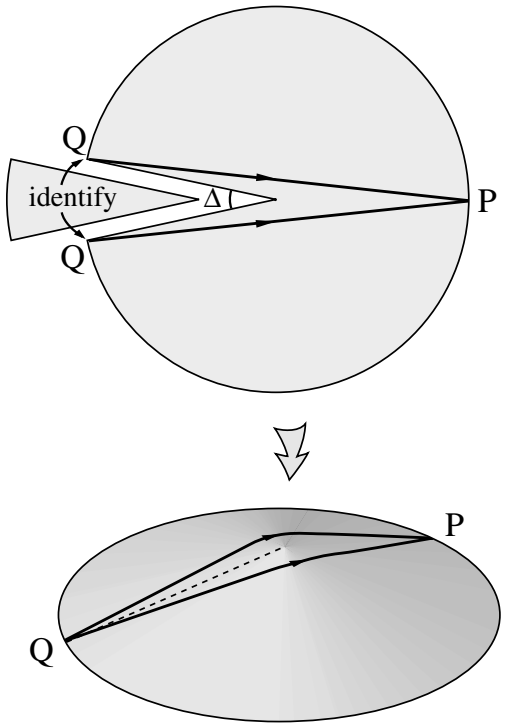


Figure 4. The conical spacetime about a cosmic string is constructed by removing a wedge of angular size Δ and identifying along the edges. Gravitational lensing by the cosmic string is also shown with the observer at P seeing a double image of the source at Q.

At this stage the configuration becomes singular and topologically unwinds; the global texture disappears in a burst of Goldstone boson radiation.

◇ *Gravitational effects of cosmic strings*

A straight string exerts no Newtonian gravitational force because the effect of its energy density is cancelled by the negative pressure of its tension. The SPACETIME locally appears to be flat but, in fact, the string induces a global distortion. Spatial sections transverse to the string are conical; to describe the space, a wedge must be removed and the remaining edges identified as illustrated in figure 4. The angular size of this wedge, known as the deficit angle, is given by

$$\Delta = 8\pi G\mu. \qquad (1)$$

Gravitational lensing by strings. This conical spacetime has a number of observational consequences including the formation of double images behind a string. Imagine a source such as a quasar located at Q in figure 4, a distance ℓ from the observer at P, who is in turn a distance d from the string. Provided that the string and source are sufficiently aligned, there are two paths by which light can reach the observer who will see two identical lensed images with an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

angular separation

$$\delta\varphi = 8\phi G\mu\ell(d+\ell)\sin\tilde{\theta}$$

where $\tilde{\theta}$ is the angle between PQ and the string and we have assumed $G\mu \ll 1$. For a GUT-scale string with $G\mu \sim 10^{-6}$, this corresponds to a separation of a few arcseconds. The probability of a double image for a source at redshift $z \sim 1$ to be lensed by a cosmic string network is about $P \sim 100G\mu$. Despite the increasing density of strings at higher redshift, this lensing probability only grows logarithmically.

Wake formation with cosmic strings. A moving cosmic string can disturb a homogeneous matter background and induce density fluctuations. Consider two test matter particles moving along parallel trajectories with velocity v_s on either side of the string. When the line drawn between the particles crosses the string, they will both begin to move towards each other with a transverse velocity,

$$u_\perp = v_s\Delta = 4\pi G\mu v_s. \qquad (2)$$

In an exactly analogous manner, if a string moves through a static homogeneous matter background then matter particles on opposite sides will begin to move together eventually colliding in a wake or pancake-like region. Subsequently, matter will continue to accrete onto this planar overdensity. ♦

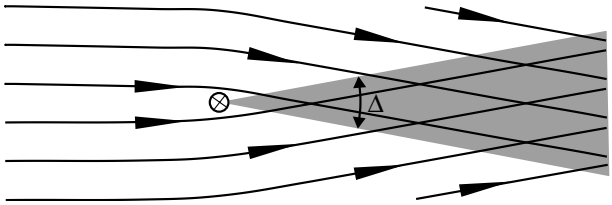


Figure 5. Wake formation induced by a cosmic string. Matter flow lines are shown in the rest frame of the string, with the density doubled in the shaded region.

Cosmological evolution

Topological defects can only be cosmologically non-pathological if their evolution is scale invariant throughout the history of the universe. In this case, the defect network approaches a 'scaling' regime in which the average number of defects in each horizon volume remains constant and they contribute a fixed fraction of the total energy density of the universe, typically of order $\sim 100G\eta^2$. For GUT-scale defects with $G\eta^2 \sim 10^{-6}$, this corresponds to a density perturbation of about the correct magnitude to seed galaxy formation.

When cosmic strings are created at a phase transition they form a network of Brownian random walks with a characteristic lengthscale ξ given by the field correlation length at that time ($\xi < H^{-1}$). Some of the strings return to their starting point to form closed loops, but the vast majority are 'infinite' (about 85%), stretching across the horizon volume and continuing on indefinitely. After a period of damped evolution, the strings begin to oscillate relativistically and ξ grows until it 'scales' as a fixed fraction of the horizon size, that is

$$\xi(t) = \gamma t$$

with $\gamma \approx \mathcal{O}(1)$. Basically, the network 'chops' itself up into small loops as fast as is allowed by causality. These loops, in turn, decay into a relatively harmless background of gravitational waves or other radiation. The energy density of a Brownian network of infinite strings is given by

$$\rho_\infty = \frac{\mu\xi}{\xi^3} = \zeta\frac{\mu}{t^2} \qquad \text{with } \zeta = \gamma^{-2}.$$

Note that like the background density in an expanding universe, the string density ρ_∞ scales as t^{-2}. Numerical simulations indicate that there are always about 40 strings per horizon volume at any time, with an example of a network shown in figure 5. However, their relative contribution to the background density changes across the matter–radiation transition with $\zeta = 13\ (\pm 3)$ during the radiation era dropping slowly to $\zeta = 4\ (\pm 2)$ in the matter era.

The requirement of scale invariance reveals immediately why both domain walls and magnetic monopoles are incompatible with the standard cosmology. Even if the 'scaling' regime only had one domain wall stretching across each horizon volume, then their energy density would be $\rho_{\rm DW} \sim \sigma t^2/t^3 \propto t^{-1}$. Thus domain walls would rapidly come to dominate over the background energy density unless their surface density σ was very small, that is the very low symmetry-breaking scale $\eta < 100$ MeV. A more stringent constraint arises from the COSMIC MICROWAVE BACKGROUND, $\eta < 1$ MeV, because a domain wall will disturb large-angle isotropy.

Magnetic monopoles are a generic prediction of GUTs, but they redshift like massive particles and so their energy density scales as $\rho_{\rm M} \propto t^{-3/2}$ during the radiation era. Naive estimates of their production rate at a GUT phase transition suggest that today there should be roughly as many monopoles as baryons. This is a drastically unacceptable conclusion—known as the 'monopole problem'—because each monopole weighs at least 10^{16} times more than a proton. An observational limit on the monopole number density $n_{\rm M}$ arises from their dissipative effect on galactic magnetic fields; this is known as the Parker limit and yields the tiny value $n_{\rm M} < 10^{-22}$ cm^{-3}. The strongest constraint arises from the consequences of baryon catalysis if GUT-scale monopoles were captured by neutron stars, yielding $n_{\rm M} < 10^{-27}$ cm^{-3}.

Global topological defects with long-range fields, whether strings, monopoles or textures, can evolve in a scale-invariant manner. Global field correlations will be established as fast as allowed by causality, with a fixed average number of global defects in each horizon volume at any one time. Global string evolution only differs from ordinary cosmic strings in a subtle quantitative manner. However, global monopoles are dominated by long-range forces which correlate them closely with antimonopoles, leading to their rapid annihilation. For such a system,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

we would expect the removal of all but one or two monopoles in each horizon volume. This has been verified numerically, with the number density in the matter era given by

$$n_{\rm M} = 4(\pm 1.5) d_{\rm H}^{-3}$$

where $d_{\rm H} = 3t$ is the horizon size.

The phase ordering of global textures is qualitatively similar. Once a sufficient texture winding has fallen within the horizon, the configuration reduces its gradient energy by collapsing. This unwinding process is again only limited by causality. Numerical simulations have confirmed that this is a scale-invariant process with a constant rate of texture unwindings per Hubble time and volume, which in the matter era is given by

$$\frac{{\rm d}n}{{\rm d}t} \approx 0.094 d_{\rm H}^{-4}.$$

Finally, we note the possibility of hybrid topological defects resulting from a series of phase transitions such as domain walls bounded by strings or monopoles connected by strings. Such configurations can chop themselves up, annihilate and completely disappear within a Hubble time. For example, if the flux of magnetic monopoles is temporarily confined into strings as in a superconductor, then monopole–antimonopole pairs become highly correlated; they are rapidly drawn together by the string tension and annihilate (a possible solution of the monopole problem). Alternatively, defect networks can also become 'frustrated' if there are topological constraints preventing annihilation or reconnection. A complex interconnected string network will form if there are many different string species arising from a non-Abelian symmetry breaking (e.g. $G \rightarrow Z_n$ produces n strings which interact non-trivially). In general, such configurations will dominate the universe unless their symmetry-breaking scale η is very low, but there are also circumstances under which this evolution can be scale invariant.

◇ *Quantitative string networks*

Analytic mathematical modelling, notably the 'one-scale' model, provides a fairly successful description of the gross features of an evolving string network. However, it is not adequate for the small-scale dynamics and, in any case, there are free parameters in the analytic models which must be fixed in some manner. Invariably, therefore, quantitative advances in understanding cosmic strings have relied on numerical simulations (as with other defects). Because of the complexity and non-linearity of a network, this has proved to be computationally very challenging. This is illustrated by the horizon volume string network from the radiation era shown in figure 6; a network of 'wiggly' long strings permeates throughout the universe with an accompanying population of very small loops. This wiggliness and the size of these small loops have an important impact on the astrophysical consequences of strings. Numerically, it is clear that high resolution is required, but this competes with the need for an extended dynamic range. Without an unacceptable sacrifice of this small-scale structure, string networks at present can only be evolved for a period of about 1 order of magnitude in conformal time. For global defects on a 3D grid, the problem is even more restrictive. This severe limitation is an important caveat in all the current literature on the cosmological implications of defects. ♦

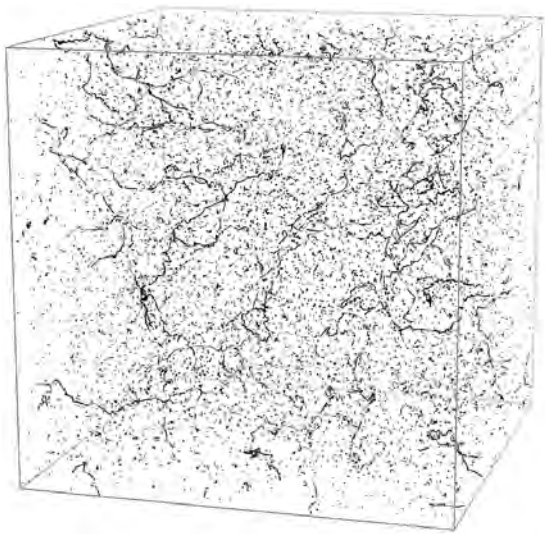


Figure 6. Horizon volume from a numerical simulation of cosmic strings during the radiation-dominated epoch.

Large-scale structure formation

Evolving superheavy defects generate complex causal gravitational fields which can induce density fluctuations in the background matter. Their scale-invariant evolution is reflected in the creation of a nearly scale-invariant or Harrison–Zeldovich spectrum of perturbations, which is very similar to that of inflationary models. For most defects, the density perturbation near the scale of the cosmological horizon remains constant:

$$\left.\frac{\delta\rho}{\rho}\right|_{\rm H} \approx \mathcal{O}(10) G\eta^2.$$

However, unlike the adiabatic (or curvature) fluctuations of standard inflation which are laid down at some early time on all relevant cosmological scales, defect perturbations are created throughout cosmic history by causal mechanisms which only operate on subhorizon scales at any one time. Ever since a phase transition carved them from the cosmological background, wherever defects exist there is always a diffuse compensating underdensity in the background which cancels out their gravitational effects beyond the horizon. In other words, defects induce isocurvature perturbations. The non-linearity of the defect source provides another possible distinction from the initially linear and Gaussian inflationary fluctuations.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

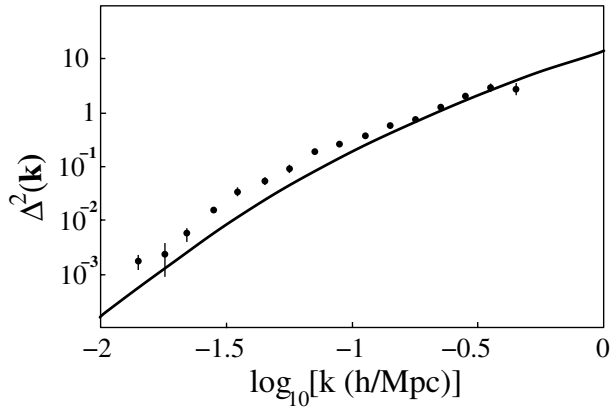


Figure 7. The power spectrum obtained for cosmic strings with cold dark matter and a significant cosmological constant $\Omega_\Lambda = 0.7$, which is normalized using large-angle CMB anisotropy (COBE). Inferred observational points from large-scale structure surveys are also included. The Hubble parameter is taken to be $h = 0.7$ and the wavenumber k is in units of h/Mpc.

In order to study the large-scale structure in a specific defect scenario we need to solve the linearized Einstein equations for a density perturbation with a source term given by the energy–momentum tensor of the evolving defect network. Given that defect simulations can only run for a short period, it is useful to characterize the scale-invariant power which they induce using unequal-time correlators. These can then be used as a scale-invariant source to greatly increase the dynamic range in a semianalytic solution of the perturbation equations. Nevertheless, this severe numerical limitation makes cosmic defect predictions considerably less reliable than corresponding results for INFLATION.

An example of the power spectrum of dark matter density perturbations produced by cosmic strings is illustrated in figure 7. This is a cold dark matter scenario with a COSMOLOGICAL CONSTANT and the string source is normalized to match the large-angle CMB anisotropies measured by COBE. The observational data points shown are the Peacock–Dodds compilation of several large-scale structure surveys. Several key features are evident, which distinguish defect models from standard inflation. Firstly, cosmic strings and other defects require considerable 'bias', that is the galaxy distribution must be more strongly clustered than the underlying dark matter. Since early non-linearity is natural in these models, it may be that this can be accommodated. Secondly, although the overall power is lower than inflation, defect models induce relatively more small-scale power. The consequence of these two features is that the shape of the defect power spectrum appears to be viable only in open models or those with a substantial cosmological constant.

Some non-Gaussianity is expected in the large-scale structure distribution induced by cosmic strings. This has been observed by studying the one-point probability density function and other statistics. However, significant non-Gaussianity is only evident on scales smaller than about $1.5(\Omega_{\rm M}h^2)^{-1}$ Mpc, that is on scales today which are non-linear in any case. The wake-like structures created by a single string are submerged among the many contributions from other strings, except paradoxically on small scales. This is because string perturbations before the equal matter–radiation transition are suppressed, so a late-time wake contribution is relatively stronger than the background.

The power spectrum induced by global monopole or texture theories appears to be less attractive than that from strings. On the one hand, in a flat matter-dominated universe ($\Omega_{\rm M} = 1$) the power spectrum has an acceptable amplitude at $8h^{-1}$ Mpc but falls away drastically on large scales. On the other hand, in an open universe or a model with a significant cosmological constant, the shape of the spectrum is improved but the overall amplitude becomes seriously deficient on all scales; biases up to an order of magnitude can be typical when normalized to large-angle CMB anisotropy. Global defect models are expected to exhibit a somewhat stronger non-Gaussian signal than strings.

Cosmic microwave sky

The greatest observational challenge to cosmic defect theories arises from the patterns they create on the cosmic microwave sky. The temperature fluctuations produced in microwave background photons travelling in a sky direction $\hat{\boldsymbol{n}}$ can be summarized by the Sachs–Wolfe formula

$$\frac{\delta T}{T} = \int h_{ij}\hat{n}^i\hat{n}^j \,\mathrm{d}t + \boldsymbol{v}\cdot\hat{\boldsymbol{n}} + \frac{4}{3}\frac{\delta\rho_\gamma}{\rho_\gamma}.$$

Here, the spacetime metric perturbation h_{ij} represents the gravitational effects of the cosmic defects, while $\boldsymbol{v}$ is the matter velocity and ρ_γ is the photon density on the surface of last scattering. For large angular scales, the first term dominates (i.e. the integrated Sachs–Wolfe effect). In standard inflation, this merely reduces to the Newtonian potential Φ on the surface of last scattering. Temperature fluctuations are mainly due to photons climbing out of the (scalar) potential wells as they decouple; these are the same wells around which large-scale structure collapses. For cosmic defects, however, large-angle CMB anisotropies are created dynamically along the whole photon trajectory by the gravitational effects of the evolving defect network. This includes additional non-scalar contributions (vector and tensor modes) which are not directly related to the induced density perturbations. These extra modes imply that defect models, when normalized by large-angle CMB observations (such as the COBE-DMR experiment), generically produce less scalar power for density fluctuations relative to inflationary models. Nevertheless, the scale invariance of defect evolution implies they produce an angular power spectrum on large scales which is nearly flat and consistent with COBE (i.e. for multipoles $\ell < 30$ in figure 8).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

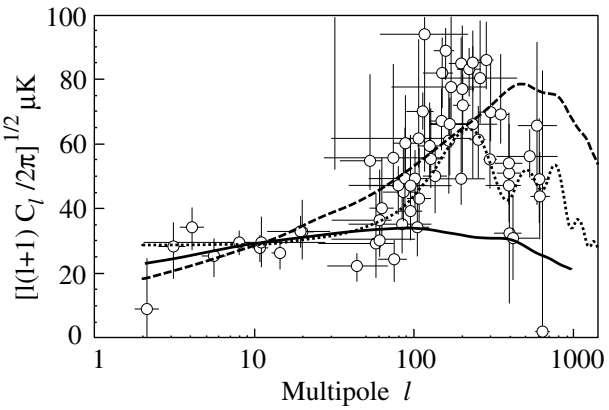


Figure 8. Angular power spectrum for cosmic microwave anisotropy from texture simulations (solid curve), a cosmic string toy model (dashed) and standard inflation (dotted), all with a Hubble parameter $h = 0.5$. The defect models are in a flat cosmology with a significant cosmological constant: textures $\Omega_\Lambda = 0.8$ and strings $\Omega_\Lambda = 0.5$. Recent CMB experimental points are included for comparison.

The additional vector and tensor modes created by defects can be illustrated by a cosmic string perturbing CMB photons when it moves across the line of sight. As the string passes in front of a distant source, the observer and source experience a small relative boost towards each other owing to the string deficit angle Δ. Thus, in the frame of the observer, the emitted light behind the string will acquire a relative Doppler blueshift, and so the string will create a line-like discontinuity in the CMB temperature pattern. Now, on large angular scales, the superposition of so many strings between the observer and the surface of last scattering means that the signal is very Gaussian. However, on very small angular scales approaching the interstring distance, this 'stringy' signature should be detectable; for example, the temperature discontinuities are inconsistent with Gaussian fluctuations on angular scales below about 2 arcmin. In a similar manner, collapsing textures or monopole pairs create hot spots in the microwave sky. Such non-Gaussian features are a distinct observational signature; they may be detectable even if the defect energy scale is below that needed for galaxy formation. A feature of defect models is that this non-Gaussian signal dies away much more slowly than inflation at high multipoles $\ell > 1000$.

On intermediate scales with multipoles $200 \leq \ell \leq 1000$, the second and third terms in the Sachs–Wolfe formula become important, that is Doppler shifts due to motion at photon decoupling and intrinsic fluctuations in the photon density ρ_γ. For these effects, defects again produce a signature quite distinct from inflation because they are isocurvature theories. In inflation, superhorizon perturbations begin to oscillate coherently as soon as they cross the horizon, providing the so-called spectrum of 'Doppler peaks' illustrated in figure 8. The primary peak due to the intrinsic photon density typically is at a multipole $\ell \approx 200$. In contrast, since cosmic defects act only on subhorizon scales, they begin to oscillate later; they are at least out of phase with adiabatic fluctuations, so the primary intrinsic peak is expected at $\ell > 300$ or greater. In fact, since defects act dynamically on scales substantially smaller than the horizon, the expected primary (scalar) peak is generally at even higher multipoles with simulations suggesting for textures $\ell \approx 350$, monopoles $\ell \approx 400$ and strings $\ell \approx 500$. Secondary peaks, if present at all, are weak because of the incoherence of the defect source. Such differences would be sufficient to distinguish cosmic defects from inflation, but the divergence is even greater when we include the integrated Sachs–Wolfe effect discussed earlier. In global defect models these dominate even at intermediate scales, so that any Doppler peaks are barely discernible as illustrated for textures in figure 8.

If present observational indications of a strong primary Doppler peak are confirmed, along with clear secondary peaks, then defect models are likely to be ruled out as the main seeds for large-scale structure formation. However, such a conclusion is subject to a number of caveats. First, the evolution of defect networks and their non-linear astrophysical effects have not been exhaustively studied, so there cannot be complete confidence in present predictions. Secondly, broken scale invariance in defect evolution may introduce features such as peaks in the angular CMB spectrum. This occurs already in cosmic string models because of the very different network densities before and after matter–radiation equality (see figure 8). Thirdly, in alternative cosmologies, the primary peak can migrate to smaller multipoles, such as in closed FRW models with a large cosmological constant. Finally, even if defects were not to be the prime cause of structure formation, they could still play a subsidiary role. This complementarity has been studied, for example, in hybrid and open inflation models. Ultimately, however, the unequivocal test for cosmic defects will be their distinct observational signature in the CMB sky on very small angular scales.

Primordial backgrounds

The scale-invariant evolution of a defect network depends on the decay of defects into a particle or radiation background. This background may well have influenced cosmic history and it may be detectable today. Supermassive cosmic strings with only local interactions oscillate relativistically and decay into gravitational radiation. The magnitude of this flat spectrum of primordial gravitational waves can be predicted from network simulations. However, millisecond pulsar timings provide an upper limit on this amplitude, since it would disturb their precise regularity. This constraint on the string energy scale, $G\mu \leq 5 \times 10^{-6}$, is quite tight but consistent with the large-angle CMB normalization. Future generations of GRAVITATIONAL WAVE detectors (such as LIGO II) will have the sensitivity to detect such a stochastic background.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Global defects decay into the massless Goldstone bosons to which they are strongly coupled. This background of Goldstone bosons, when there are too many massless degrees of freedom, can upset primordial nucleosynthesis predictions; the fraction of the energy density in the universe in Goldstone bosons at $t \sim 1$ s must be less than 5.4% for standard light element synthesis to work. This does not provide a serious constraint on GUT-scale global defects unless they are produced in many varieties.

Supermassive strings which are superconducting and carry large currents decay preferentially into electromagnetic radiation rather than gravitational waves. The resulting radiation pressure that builds up around an oscillating loop can push background plasma into a thin expanding wall. When these plasma bubble walls collide, they heat the surrounding medium and dump their energy into radiation which, in turn, can distort the Planck spectrum of the CMB. Such an explosion scenario was once proposed as a theory for large-scale structure formation, but an unacceptable CMB distortion results for sufficiently heavy strings $G\mu \geq 10^{-7}$.

Dark matter

The observable implications of defect models are not restricted exclusively to supermassive GUT-scale defects. For low or intermediate energy scales, massive relics of defect network decay can provide plausible candidates for missing dark matter. Axions are pseudo-Goldstone bosons which, in the standard thermal scenario, are predominantly produced by radiation from axion strings. Uncertain calculations of this complex process produce cosmological bounds on the axion string formation scale of approximately $\eta_a \leq 3 \times 10^{10}$ GeV. This indicates that the most likely mass range for a dark matter axion lies near $m_a \sim 0.1$ meV. Superconducting strings formed around the electroweak scale can leave behind 'vortons', that is vortex rings stabilized by their currents and charge. They may also collect charge into stable Q-ball configurations. Both vortons and Q-balls are ultra-massive cold dark matter candidates. The collapse of circular cosmic string loops could also create a distribution of primordial black holes, with a variety of observational consequences.

As noted previously, certain varieties of strings and other defects are unable to reconnect or pass through each other, so they become entangled. It is believed that the 'frustrated' defect networks that result do not evolve in a scale-invariant manner. Instead, they relax towards an equilibrium configuration which is merely stretched by the Hubble expansion and slowly comes to dominate the energy density of the universe. The network acts like an almost smooth dark matter component, but with a strange equation of state owing to the negative pressure of the defects. If the network dominates, this negative pressure causes the universe to expand faster that in the matter-dominated phase.

For non-Abelian strings in a flat cosmology ($\Omega = 1$), the universe mimics an open model ($\Omega < 1$) with the scale factor growing as $a \propto t$. Since these strings will not cluster on galaxy or cluster scales, they must not dominate well after structure formation has begun. This constraint implies that the strings must be very light forming at the electroweak scale or later, that is $\eta < 10^4$ GeV. The mean separation today of such strings would be only a few astronomical units; however, their direct detection would be extremely difficult unless they were superconducting.

Recent observational evidence favouring a flat cosmology with a low matter density has renewed interest in a non-zero cosmological constant Λ. Equally, however, frustrated defect networks could play a similar role, arguably without the severe fine-tuning problems associated with Λ. Observational tests of these models must focus on determining the equation of state $p = w\rho$ of this homogeneous background component. Frustrated strings can be expected to have $w \approx -\frac{1}{3}$, while domain walls will have $w \approx -\frac{2}{3}$, as compared with a cosmological constant with $w = -1$.

Baryogenesis

Cosmic defects can have interactions which violate baryon number B and which could, in principle, generate the matter–antimatter asymmetry of the universe (see BARYOGENESIS). Of course, B-violating interactions alone would tend to act in the opposite direction by erasing any pre-existing or primordial baryon asymmetry. However, if the B-violating interactions are also CP violating and they occur out of equilibrium (i.e. Sakharov's three criteria), then they can create a net baryon number in a baryon-symmetric universe. Several speculative defect scenarios have been proposed for this including the B- and CP-violating decay of heavy particles emitted by string cusps or small loops and B-violating scattering from superconducting strings, as well as monopole annihilation and texture collapse.

Baryon-violating processes can also occur at the electroweak phase transition. In this case, sphalerons (i.e. local texture transitions) can cause baryon number violation. However, it is difficult for these transitions to favour baryons over antibaryons in the standard electroweak model; the CP violation that does exist in the kaon system is insufficient. Instead, various extensions of the standard model, such as the minimal supersymmetric standard model (MSSM), have been studied to increase the level of CP violation. Usually, this entails mechanisms occurring on expanding bubble walls during a first-order phase transition. A small parameter range in the MSSM model remains viable for electroweak baryogenesis but this will be tested at LEP in the near future. A more detailed description of alternative models is beyond the scope of this article, but it may yet emerge that baryogenesis is the most important consequence of topological defects in cosmology.

Cosmic rays

Topological defects can produce highly energetic massive particles through their relativistic accelerations and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

interactions. These can decay and cascade into observable cosmic rays over a broad energy range which extends to ultrahigh energies (UHEs) with $E > 10^{11}$ GeV. Most recent interest has focused on UHE cosmic rays because of the difficulty of explaining the observed events with the usual astrophysical accelerator mechanisms. Various scenarios for their generation by topological defects have been discussed in the literature including cosmic string reconnections, current loss from superconducting strings, monopole–antimonopole annihilation and hybrid networks of monopoles connected by strings. The problem faced by these models is not achieving exotic high energies but rather reaching the required flux levels. For UHE cosmic rays, both the proton attenuation length and the photon absorption length due to interactions with CMB photons are much smaller than the Hubble radius. For ordinary or superconducting cosmic strings, this cutoff is also much smaller than the mean string separation, so the cosmic ray flux they create is generically exponentially suppressed. Even if a string were nearby by chance, it would generate a highly anisotropic flux. Of course, many possibilities remain to be explored such as neutral carriers which can evade this limit.

Nevertheless, several models with GUT-scale monopoles appear to be able to generate the required flux of UHE cosmic rays. In one case, the monopoles are attached to two strings, so that they form 'beads' in a 'necklace' configuration. If the strings are light (say the electroweak scale), then the monopoles can retard the evolution such that the network evolves to a much denser state than usual with a mean interstring spacing of a few kpc. The monopoles and antimonopoles trapped on the strings continuously annihilate to produce superheavy particles which cascade into a nearly isotropic flux of cosmic rays. Another alternative is a primordial distribution of monopolonium, that is, bound states made from monopole–antimonopole pairs. These slowly decay electromagnetically and eventually annihilate. Monopolonium (or other localized defect states such as vortons) will cluster like ordinary matter, so one would expect to observe anisotropy with a peak cosmic ray flux from the Galactic center, and also from the Virgo cluster. A distinct feature of defect-generated cosmic rays is the pion dominance of hadron production, leading to a strong electromagnetic cascade; unlike most mechanisms, there are more photons than protons. For localized defects inside the photon absorption cutoff distance, this will result in a directional flux of UHE photons.

Topological defects are also invoked to account for other high-energy astrophysical phenomena which evade more conventional explanation. One example is the engine driving a γ-ray burster (GRB) which must emit in an extremely narrow beam. The cusps on superconducting strings at the GUT scale $\eta \sim 10^{14}$ GeV can achieve the beamed energy levels required and exhibit a number of the key qualitative features of GRBs.

Cosmic defects and inflation

A solution to the monopole problem was a prime motivation for inflation; any cosmic defects present before inflation would be drastically diluted by the ensuing epoch of exponential expansion. Defects and inflation, therefore, have become widely regarded as incompatible, but this is certainly not the case. Of course, light- or intermediate-scale defects can be formed at phase transitions which occur after inflation, but it is also possible to create heavy defects in significant quantities even if their energy scale is higher than the inflationary reheat temperature $\eta > T_{\rm R}$. Defect-forming phase transitions can occur during inflation or they can trigger its end, notably in hybrid inflation models. Defect networks can be created by quantum fluctuations during inflation. The quantum nucleation of symmetric defect configurations, such as circular string loops or monopole–antimonopole pairs, can also act continuously during inflation to give non-negligible final densities. A particularly interesting possibility is inflationary pre-heating which can cause non-thermal phase transitions. These can naturally give rise to superheavy defects even with $\eta > 10^{16}$ GeV. Together with inflation, cosmic defects hold out the prospect of solving a number of key cosmological enigmas.

Bibliography

The literature on topological defects in cosmology is extensive and can be approached through the references listed here. A recently updated paperback edition of a comprehensive monograph on the subject is

Vilenkin A and Shellard E P S 2000 *Cosmic Strings and Other Topological Defects* (Cambridge: Cambridge University Press)

Another review on cosmic strings is

Hindmarsh M B and Kibble T W B 1995 Cosmic strings and coherence *Rep. Prog. Phys.* **58** 477

Useful conference proceedings about topological defects in cosmology (and other contexts) include

Davis A-C and Brandenberger R H (ed) 1995 *Formation and Interactions of Topological Defects* (New York: Plenum)

Gibbons G W, Hawking S W and Vachaspati T (ed) 1990 *Formation and Evolution of Cosmic Strings* (Cambridge: Cambridge University Press)

Some of the watershed papers on the cosmological implications of topological defects which have sparked interest in the field are

Kibble T W B 1976 Topology of cosmic domains and strings *J. Phys. A: Math. Gen.* **9** 1387

Kibble T W B 1980 Some implications of a cosmological phase transition *Phys. Rev.* **67** 183

Preskill J 1979 Cosmological production of superheavy monopoles *Phys. Rev. Lett.* **43** 1365

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Turok N G 1989 Global texture as the origin of cosmic structure *Phys. Rev. Lett.* **63** 2625
Vilenkin A 1981 Cosmological density fluctuations produced by vacuum strings *Phys. Rev. Lett.* **46** 1169
Vilenkin A 1981 *Phys. Rev. Lett.* **46** 1496 (erratum)
Vilenkin A 1985 Cosmic strings and domain walls *Phys. Rev.* **121** 263
Witten E 1985 Superconducting strings *Nucl. Phys.* B **249** 557
Zel'dovich Ya B 1980 Cosmological fluctuations produced near a singularity *Mon. Not. R. Astron. Soc.* **192** 663

Defect predictions for structure formation and the CMB sky used for figures 7 and 8 were gleaned from the following papers: for figure 7,

Avelino P P, Shellard E P S, Wu J H P and Allen B 1998 Cosmic-string-seeded structure formation *Phys. Rev. Lett.* **81** 2008

for figure 8,

Allen B, Caldwell R R, Dodelson S, Knox L, Shellard E P S and Stebbins A 1997 CMB anisotropy induced by cosmic strings on angular scales >15′ *Phys. Rev. Lett.* **79** 2624
Battye R A, Robinson J and Albrecht A 1998 Cosmic strings and a cosmological constant *Phys. Rev. Lett.* **80** 4847
Durrer R, Kunz M and Melchiorri A 1999 Cosmic microwave background anisotropies from scaling seeds: global defect models *Phys. Rev.* D **59** 123005
Pen U-L, Seljak D N and Turok N 1997 Power spectra in global defect theories of cosmic structure formation *Phys. Rev. Lett.* **79** 1611

Observational comparisons were made with the Peacock and Dodds and the Tegmark survey compilations.

E P S Shellard

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Torricelli, Evangelista (1608–47)

Born in Faenza, Romagna, Italy. He was in his youth greatly interested in astronomy and a Copernican but upon the condemnation of GALILEO in 1633 he remained silent on astronomy, probably regarding it as too controversial. He worked as a secretary to Galileo for a year before his death, and succeeded him as Mathematician to the Grand Duke Ferdinando II of Tuscany. He was the first man to create a sustained vacuum and to discover the principle of a barometer. He was a skilled optician, making telescopes and microscopes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Torun Radio Astronomy Observatory

Torun Center for Astronomy is located at Piwnice, 15 km north of Torun, Poland. A part of the Faculty of Physics and Astronomy of the Nicolaus Copernicus University, it was created by the union of Torun Radio Astronomy Observatory (TRAO) and the Institute of Astronomy on 1 January 1997.

Since 1981 the observatory has been part of a worldwide network of radio observatories participating in VLBI (very long baseline interferometry). With the new 32 m telescope and modern sophisticated instrumentation, Torun station is the only VLBI site in eastern Europe. It recently became a member of the European Pulsar Network (EuroToP).

Research concentrates on observations and theoretical interpretation of physical processes in radio galaxies and quasars, the timing of pulsars, the search for new planetary systems and solar system studies. The development of instrumentation for radioastronomy (ultra low-noise receivers, spectrometers, frequency converters, digital electronics, control systems) is also important.

For further information see
http://www.astri.uni.torun.pl/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Towneley, Richard (1629–1707)

Wealthy scientist, born in Towneley Hall, near Burnley, Lancashire, made meteorological measurements and improved the micrometer (for use in astronomical observations), introducing FLAMSTEED to it. His escapements were blamed by THOMAS TOMPION for some initial problems with the clocks he made for John Flamsteed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Townes, Charles Hard (1915–)

Physicist, born in Greenville, SC, Nobel prizewinner for physics in 1964 'for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the maser-laser principle'. Worked at Bell Telephone Laboratories and Columbia University on microwave spectroscopy to study the structure of molecules, atoms and nuclei. Conceived the idea of a device for microwave amplification by stimulated emission of radiation, coining the word 'maser' from the acronym. Became professor of physics at the Massachusetts Institute of Technology, and then at the University of California. He used the maser device to make sensitive amplifiers for radioastronomy and predicted the existence of the OH line in molecules of interstellar space.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

TRACE (Transition Region and Coronal Explorer)

The Transition Region and Coronal Explorer (TRACE) is a NASA Small-Explorer (SMEX) satellite developed to study the impact of magnetic fields on the solar outer atmosphere. The instrument observes the solar surface (SOLAR PHOTOSPHERE) and the hotter overlying domains (TRANSITION REGION and CORONA) with an angular resolution of 1 arcsec, equivalent to 725 km on the Sun, an average cadence of about 30 s, and a field of view of 512 by 512 arcsec (covering one-tenth of the solar disk). The TRACE observatory was launched on 2 April 1998 from Vandenberg Air Force Base on a Pegasus XL launch vehicle built by Orbital Sciences Corporation. The spacecraft is in a polar orbit roughly over the day–night terminator. This Sun-synchronous orbit allows TRACE to see the Sun without interruptions for a period of approximately nine months of the year, while there are some ECLIPSES of the Sun by the Earth during the other three months.

The main scientific objective of the TRACE mission is to enhance our understanding of the dynamic, hot solar outer atmosphere. To achieve this goal, the instrument is used to:

- Study how the SOLAR MAGNETIC FIELD structures the outer atmosphere, from the photospheric surface to the high corona, covering a temperature range from 5000 up to several million degrees.
- Observe how the magnetic fields evolve from their emergence onto the solar surface up to their disappearance, observing both the gradual dispersal and decay of the field and the occasional instabilities that occur (in FLARES).
- Investigate the entire range of scales from the instrumental resolution (725 km) up to the global field, and the way in which these scales influence each other.

To achieve these objectives the TRACE instrument uses a 30 cm aperture CASSEGRAIN TELESCOPE with normal-incidence optics to image wavelengths at visible, ultraviolet (UV) and extreme-ultraviolet (EUV) wavelengths. The instrument was developed by a team composed of the Lockheed-Martin Advanced Technology Center, the Smithsonian Astrophysical Observatory, NASA/Goddard Space Flight Center and Montana State University.

Instrument description

The TRACE telescope is divided into four quadrants. Each quadrant has an optimized optical coating on the primary and secondary mirrors as well as special filters for the selection of the wavelength to be observed. The appropriate mirror segment is selected with a rotating quadrant selector just behind the front window. The EUV quadrants are coated with multilayers, i.e. multiple layers of very thin coatings that allow the use of normal-incidence optics to observe the EUV wavelengths. These multilayer coatings were of the highest reflectivity and narrowest bandpass ever deployed in space at the time of launch. The UV quadrant uses a narrowband multilayer on the primary mirror (optimized for 1550 Å) and an aluminum coating on the secondary to provide for a capability to image a wide range of wavelengths in the UV and visible (see table 1). A set of filter wheels in front of the detector provides the final selection of the desired wavelengths. The image is recorded on a lumogen-coated charge-coupled display (CCD) to make it sensitive to both UV and EUV wavelengths. Typical observing programs cycle through multiple wavelengths to observe different temperature domains in the solar atmosphere.

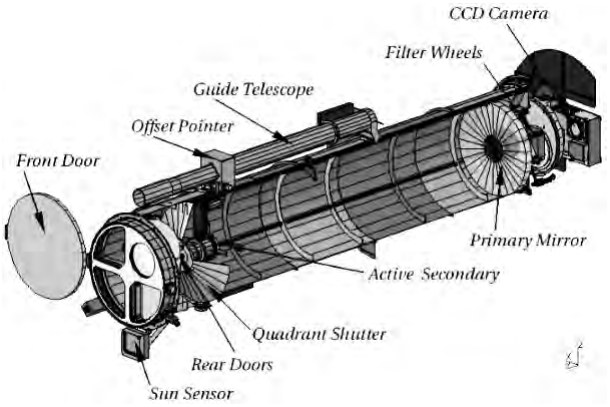


Figure 1. The TRACE instrument.

Table 1. Wavelengths observed by TRACE.

Wavelength (Å)	Emission	Temperature (kK)
173	Fe IX/X	160–2000
195	Fe XII	500–2000
284	Fe XV	1250–4000
1216	H I Ly α	10–30
1550	C IV	60–250
1600	UV cont, C I, Fe II	4–10
1700	Continuum	4–10
5000	White light	4–6.4

At the time of launch, the TRACE telescope had the highest spatial resolution of any space-borne imaging solar telescope, with a angular resolution of 1 arcsec or 725 km. The CCD detector (with 0.5 arcsec pixels) is a 1024×1024 array, with a field of view of 8.5×8.5 arcmin (equivalent to 6.2×6.2 Mm at the solar disk center).

The instrument maintains pointing to an accuracy of 0.1 arcsec by a combination of gyrostabilization of the spacecraft with further stabilization achieved by moving the secondary mirror in response to a signal from an external GUIDE TELESCOPE. The satellite points to various locations on the Sun directed by the guide-telescope signal, in which rotating wedge prisms are used to deflect the optical axis of the telescope in the desired direction within a cone of $\pm 1^\circ$ away from the center of the disk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The onboard computer system is responsible for various housekeeping functions, JPEG data compression and limited scientific image processing prior to downlinking the signal. Normally six downlinks occur daily at Wallops Flight Facility, Maryland, USA; Poker Flat, Alaska, USA and McMurdo Station, Antarctica. This combination of operations, data compression and creative planning techniques allow TRACE to record one full image roughly every minute when operating in a continuous mode, resulting in a data rate of ≈ 0.5 Gbyte day^{-1}.

Operations and data archive

Standard science operations are executed from the TRACE Experiment Operations Facility (EOF) at Goddard Space Flight Center. Operations personnel closely coordinate with other observatories, particularly with the Solar and Heliospheric Observatory (SOHO). In addition to general operations in response to solar conditions and requests from other observatories, TRACE covers the entire Sun to construct a full-disk mosaic of the central meridian at least every other day.

During the first nine months of operations, up to the beginning of the first eclipse season, TRACE acquired nearly a million exposures, running at an average cadence of three images per minute. The TRACE project has a policy of making all data available to colleagues and the general public within at most a few days of the observation.

Science highlights

The TRACE images present a view of the solar outer atmosphere not anticipated prior to launch. Many of these observations are likely to take some time yet to be properly interpreted. Some early highlights are discussed below:

Perhaps most exciting is the level of structural detail and the short evolutionary timescales that are apparent in TRACE image sequences. The smallest features are still unresolved at 1 arcsec resolution. In fact, entire magnetic loops in the corona are seen to have a cross section at the instrumental resolution, suggesting that the actual widths of such loops are smaller still. Movies show significant restructuring of the solar corona down to temporal cadences of some 10 s.

The solar corona shows a wide range of loops, ranging from about the instrumental resolution up to hundreds of thousands of kilometers. These loops connect either opposite polarities in photospheric regions that emerged together or reconnect to form connections between magnetic flux concentrations that emerged at separate times and locations. Over large bipolar active regions in the photosphere, the corona shows the following thermal pattern. Over the inner parts of the region near the polarity inversion line, the corona is relatively hot with temperatures typically of 3 to 5 million degrees, observable by the Yohkoh/SXT instrument. Surrounding that inner core is an ensemble of loops with substantially lower temperatures, visible to TRACE. The material in these loops, many of which have temperatures throughout their interior of near 1 million degrees, is strongly stratified under the influence of gravity, leading to emission primarily in their lower segments.

One particularly interesting feature in TRACE observations is the so-called moss. This moss is a low-lying, bright emission pattern with many dark inclusions, overlying the inner parts of active regions, underneath the hot 3–5 million degree corona. The moss has been identified as the upper transition region between the CHROMOSPHERE and the corona, through which energy is transported primarily by thermal conduction. As the temperature gradient is very steep, the layer of emission is rather thin. Material at chromospheric temperatures is injected into this height range from below, by an as yet unidentified mechanism, causing rapidly evolving dark intrusions in the moss pattern. On a few occasions, TRACE movies have shown patches of the moss evolving in a manner akin to a prairie fire following a flare, reflecting the thermal evolution of the loops higher up in the corona.

The TRACE temporal and spatial resolution has allowed solar physicists to directly observe for the first time oscillations in CORONAL LOOPS. These oscillations are almost invariably observed in the lower segments of loops ending in sunspots. Large-amplitude transverse oscillations are frequently observed after solar flares, when the associated blast wave excites such oscillations.

Bibliography

The latest information on mission status, and image gallery, and scientific results and observing programs may be found at the TRACE web page, `http://vestige.lmsal.com/TRACE/`.

Handy B N *et al* 1999 The Transition Region and Coronal Explorer *Solar Phys.* **187** 229–60

Schrijver C J *et al* 1999 A new view of the solar outer atmosphere with the transition region and coronal explorer *Solar Phys.* **187** 261–302

Brian N Handy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Transit

The passage of a celestial body across the disk of a much larger one. The term is applied most commonly to passages of Mercury and Venus across the face of the Sun. Such transits were first reliably predicted by Johannes Kepler, in his *Rudolphine Tables* of planetary motion. From these predictions Pierre Gassendi became the first to observe a transit of Mercury, in 1631; eight years later Jeremiah Horrocks, also using Kepler's tables, made the first observation of a transit of Venus. These two planets are seen in transit when near one of their nodes at inferior conjunction. Mercury is at its ascending node in November, and its descending node in May; these occasions coincide with inferior conjunction at intervals of 7, 13 and 46 years (ascending node) and 13 and 46 years (descending node). Transits of Venus are much rarer. They can happen near June 7 (ascending node) or December 8 (ascending node), and recur in pairs, 8 years apart, at intervals of $105\frac{1}{2}$ years and $121\frac{1}{2}$ years. The next pair will be on 2004 June 8 and 2012 June 5/6; these will be only the fifth and sixth transits since the one Horrocks observed in 1639. From the eighteenth century transits afforded an important means of determining the scale of the solar system. Its relative proportions were known from Kepler's laws. Accurate timings of transits from widely separated locations on the Earth's surface provided a trigonometric fix. Early attempts to do this for Venus were frustrated by the refraction of sunlight through the planet's atmosphere, which makes the planet appear smaller than usual, and by the black drop effect.

The major satellites of planets may be seen to transit the face of their parent planet. The four largest satellites of Jupiter, known as the Galilean satellites, frequently transit the planet's disk; the larger inner satellites of Saturn may also be observed in transit. A satellite's shadow is sometimes visible crossing the planetary disk in a *shadow transit*. The term 'transit' is also used by planetary observers to indicate the passage of an atmospheric or surface feature across a planet's central meridian—the imaginary line that bisects its visible disk, from the north pole to the south pole. Timing such transits provides a means of determining a planet's rotation period.

See also: black drop, Jupiter: satellites.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Transition Region Flows

Ultraviolet emission lines emitted from the SOLAR TRANSITION REGION are often shifted from their expected rest wavelengths. Shifts of spectral lines are due to the so-called DOPPLER EFFECT, where the source of emission is moving either away from or towards the observer, causing a change in the apparent wavelength. The shifted emission lines are most often interpreted as a flow of plasma along magnetic field lines or loops with velocities varying from a few $\mathrm{km\,s^{-1}}$ to several hundreds of $\mathrm{km\,s^{-1}}$. Flows observed in the transition region can be divided into three main groups: steady flows with relatively small velocities observed in the quiet Sun, short-lived high-velocity events in the quiet Sun and high-velocity (often supersonic) flows connected to SOLAR ACTIVE REGIONS. The latter includes flows related to flaring activity and mass ejections. The different types of flow are discussed below but one should keep in mind that plasma flow is not the only explanation for the Doppler-shifted emission lines. Model calculations suggest that waves propagating through the solar atmosphere can also produce line shifts and complex line profiles.

Systematic flows in the quiet Sun

One of the most puzzling problems in solar physics is the apparent net redshift of emission lines in the solar transition region. In recent decades, this phenomenon has been observed with several UV instruments with different spatial resolutions. For the Sun, the typical value of the average downflow velocity, derived from the C IV resonance lines at 1550 Å, is 5–10 km $\mathrm{s^{-1}}$. The C IV lines are formed at temperatures around 10^5 K. Measurements of the variation of the flow with temperature are somewhat ambiguous, however.

In earlier investigations the magnitude of the redshift has been found to increase with temperature, reaching a maximum at $T = 10^5$ K, and then to decrease sharply toward higher temperatures (2.5×10^5 K). These results have been used extensively in models trying to explain the mechanisms behind this phenomenon. Recent observations from SOHO have extended the observable temperature range and suggest that the average redshift persists to higher temperatures than in previous investigations. Shifts in the range 10–16 $\mathrm{km\,s^{-1}}$ are observed in lines formed at $T = (1.3$–$2.5) \times 10^5$ K. Even upper transition region and coronal lines (O VI, Ne VIII and Mg X) show systematic redshifts in the quiet Sun corresponding to velocities around +5 $\mathrm{km\,s^{-1}}$. These results have introduced new constraints on models of the redshifts in the solar atmosphere. In coronal holes there is evidence for a small net blueshift in the upper transition region, while at lower temperatures the lines still show a net redshift.

An interesting feature of the transition region is that the average redshifted emissions appear to be a persistent phenomenon in both quiet and active regions. The estimated downward mass flux around $T = 10^5$ K is sufficient to empty the corona in a few minutes. It therefore seems likely that upflows also exist that could balance the observed downflow unless the Doppler shifts are caused by waves propagating through the transition region. Both downflows and upflows may occur where, for some reason, the downflowing plasma is slightly brighter than the upflowing plasma.

Several hypotheses to explain the redshifted emission have been proposed so far and they include: the return of spicular material; siphon flows in coronal loops; nanoflares followed by downward propagating acoustic waves; a different temporal duration of upflows and downflows; and asymmetric heating of coronal loops. None of them has been found to be entirely satisfactory. Thus, other mechanisms should still be considered to explain the observed redshifts at higher temperatures.

Detailed studies of the redshifted emission observed on the Sun also are important for the understanding of similar observations of stellar spectra. Systematic redshifts have also been observed in spectra of late-type stars, first with the INTERNATIONAL ULTRAVIOLET EXPLORER (IUE) and recently by the HUBBLE SPACE TELESCOPE. Below temperatures of about 1.6×10^5 K, the line redshifts of the Sun, α Cen A, α Cen B and Procyon are all very similar. However, at higher temperatures both α Cen A and α Cen B suggest drop-offs in redshifts similar to that suggested by the solar data. In comparison, no drop-off in the redshifts was found in the spectra from Procyon, which is a hotter and somewhat more active star than the Sun.

High-velocity events

High-velocity events in the solar transition region were first observed in UV spectra obtained with the Naval Research Laboratory's High Resolution Telescope and Spectrograph (HRTS). The strength of the HRTS instrument is its high spatial, spectral and time resolution combined with the extensive wavelength and angular coverage. This makes it well suited for studies of fine structure and mass flows in the upper solar atmosphere and led to the discovery of short-lived transient events with high velocities, so-called explosive events (see TRANSITION EVENTS: EXPLOSIVE EVENTS). Explosive events are found to be short-lived (60 s), small-scale (1500 km) high-velocity flows that occur frequently, with a birthrate of $\sim$ 600 $\mathrm{s^{-1}}$, over the entire surface of the Sun. The observed Doppler shifts correspond to line-of-sight velocities of typically 150 $\mathrm{km\,s^{-1}}$ towards and away from the observer. Many observations suggest that explosive events are bi-directional jets and that they are evidence of MAGNETIC RECONNECTION above the solar surface. However, other mechanisms are also possible, including episodic nonlinear MHD disturbances produced by nano-flares that will introduce particle motions leading to broad line profiles as observed in explosive events. Flows with opposite directions within unresolved loop structures should also be considered since many explosive events show multiple line components.

The concept of multiple flows is another interesting feature seen in the HRTS data and they are most common

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

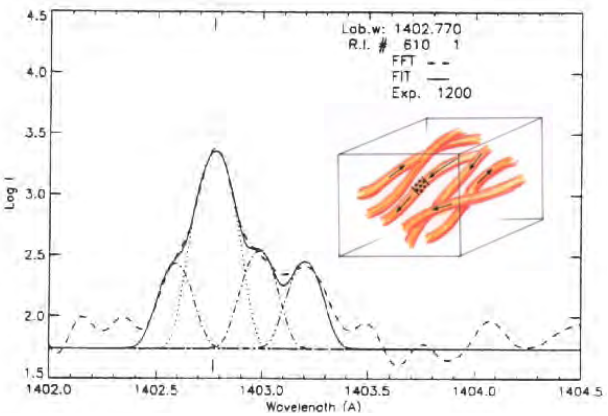


Figure 1. A complex Si IV 1402 Å profile observed with the HRTS instrument. The dashed curve shows the raw data. The profiles have been fitted with Gaussian line components (solid curve) and each separate line component is indicated. The inser shows a sketch of a possible explanation of the observed profiles.

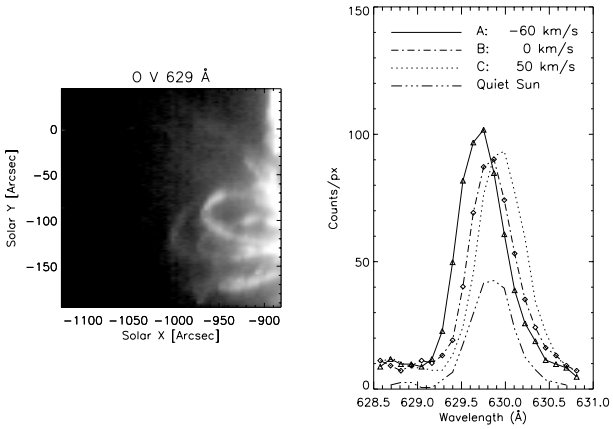


Figure 2. An active-region loop system above the east solar limb observed in O V 629 Å by the Coronal Diagnostic Spectrometer (CDS). The line profiles from three different spatial locations (A, B and C) are displayed in the right panel together with a reference profile representing a typical quiet Sun profile at rest wavelength.

above active regions but are also present in the quiet Sun. An example of a complex Si IV 1402 Å profile is shown in figure 1 and was recorded during a rocket flight in 1978. The dashed curves are the original observations while the solid curve represents the spectrum derived from Gaussian fits to the profile components. The shift of the fitted line components corresponds to velocities of -40, 8, 75 and 105 km s^{-1}. Thus, both subsonic and supersonic velocities may coexist within the same 1 arcsec^2 resolution element, suggesting the presence of an extremely structured transition region. To explain the complex profiles with siphon flows would require several loop strands each with their own flow. An alternative explanation may be episodic nonlinear MHD disturbances as mentioned above.

Flows in active-region loops

Observations of active regions above the solar limb reveal individual loops that are often too faint to be observed against the brighter disk. Observations made with the Coronal Diagnostic Spectrometer (CDS) on SOHO show clear evidence of strong dynamical activity within individual loops. The CDS spectrometer builds up images of solar features in a series of spectral lines. Each image retains detailed spectral information and can be used to give maps of total emission, velocities, line widths, plasma densities etc. A monochromatic image of an active-region loop system above the east limb observed in O V 629 Å ($T \sim 250\,000$ K) is shown in figure 2. The field of view of the image is 240 × 240 arsec^2. For each single pixel element in the CDS raster images one can extract the entire line profile to derive line-of-sight velocities. Three spatial positions have been marked and the line profiles from these locations are displayed in the panel to the right.

The line shift at location A corresponds to a gas motion toward the observer with a velocity of 60 km s^{-1} (216 000 km h^{-1}). In the other leg of the loop, at location B, there is surprisingly no shift, and hence no motion. Location C belongs to a different loop where we again observe clear shifts, corresponding to motions of 50 km s^{-1} (180 000 km h^{-1}) in the opposite direction, away from the observer.

The fact that one observes a large Doppler shift in only one part of the loop is intriguing. If gas is moving up one leg and coming down the other one would expect to see the flows continue in the other leg. This is not always the case; thus one needs to consider other possible mechanisms causing the Doppler shifts, e.g. Alfvén waves or magnetosonic waves running along the magnetic field in the loop. However, flows seem to be a general property of active-region loops as observed at transition region temperatures. Typical values correspond to velocities of the order ±50–100 km s^{-1}. In the line formed at coronal temperatures the shifts are much smaller, often not detectable. Thus, the high shifts seem to be a phenomenon occurring at transition region temperatures.

While the hot coronal loops develop more slowly, recent observations made with CDS, and in particular TRACE (TRANSITION REGION AND CORONAL EXPLORER), demonstrate that the large cool loops from the transition region are very dynamic (see figure 3). It is clear that the cool loops change on a timescale shorter than 1 min. This should be compared with the 17 min it takes to complete the monochromatic image in figure 2. Thus, the dynamic state of the loop changes significantly over the time it takes to build up a raster image with any spectrometers. The time sequence in figure 3 also suggests substantial movements of emitting plasma. Blobs of emission move along the loop, sometimes from one footpoint to the other. At other times gas appears to drain down from the top of the loop towards the footpoints. In some cases the loop seems to be filled from one footpoint to the other as is illustrated in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

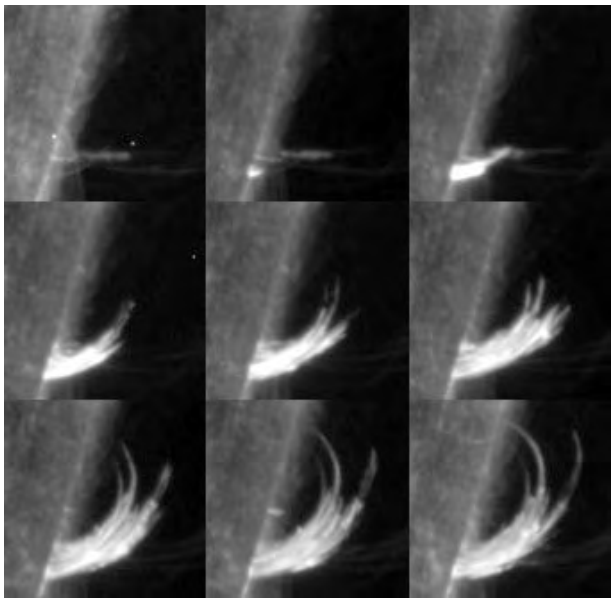


Figure 3. Active-region loops observed with Transition Region and Coronal Explorer (TRACE) illustrating the dynamic behavior of the plasma in the transition region. The images run with time from the start in the upper left corner and to the right in every horizontal row. Time between images is about 30 s.

figure 3. The observations were made with the TRACE and show active-region loops on the limb. The emission comes mainly from C IV formed at 10^5 K and the time cadence between the images is 30 s. The field of view is 55 $\times$ 55 arcsec. The images run with time from the start in the upper left corner and to the right in every horizontal row. In just about 3 min a cluster of thin loops are filled from the onset at the lower footpoint almost to the other loop footpoint. Again, even if flows may be a plausible explanation we cannot rule out other mechanisms such as waves.

Flows above sunspots

The spectrometers CDS and SUMER on SOHO have been used to observe the spatial distribution of line emission and line-of-sight velocities above a number of active regions that all included a single sunspot or a sunspot group. All the active regions show enhanced EUV line emission in the transition region, the emission is distributed non-uniformly and is located both inside and outside the sunspots. Both up- and downflows are observed in the regions often with velocities higher than ± 30 km s^{-1} .

The spatial extent of both emission features and flow regions increases with increasing temperature within the transition region. The observations show a marked difference between the transition region and the low corona, regarding both the spatial distribution of the line emission and the line-of-sight velocity. Two out of three sunspot regions show enhanced line emission above the sunspot, i.e. sunspot plumes. The line-of-sight velocity in sunspot plumes is directed away from the observer and is found to increase with increasing line formation temperature; it reaches a maximum value up to 40 km s^{-1} near temperatures of (3–4) $\times 10^5$ K, then decreases abruptly.

Flows related to flares

Mass motions in SOLAR FLARES have been extensively studied in the lower chromosphere and in the corona. However, there are only a few measurements of mass flows in the solar transition region during flares. Redshifts in flare spectra from OSO 8 and Skylab suggested downward motion of 30–50 km s^{-1}. However, coronal emission lines appeared to show a blueshifted component of emission together with a strong stationary component based on observations from the Solar Maximum Mission (SMM). Upflows reaching velocities of 200 km s^{-1} during the impulsive phase of the flare have been observed in the coronal line Fe XXI 1354 Å. Similar upflows (200–300 km s^{-1}) were also observed with the Bent Crystal Spectrometer on SMM. Based on a large survey of more than 200 flares observed in soft x-rays with Yohkoh, the centroids of flare profiles were found to be blueshifted with an average velocity of 58 km s^{-1} during the early impulsive phase. However, the blueshifted components reached velocities up to 800 km s^{-1}.

Considerable improvement in stellar flare observations followed the deployment of the Hubble Space Telescope (HST), in particular from observations with the Goddard High Resolution Spectrograph (GHRS). Doppler shifts during stellar flares have been detected with velocities as large as 1800 km s^{-1} away from the observer.

Bibliography

For further reading about flows in the transition region a few selected pointers to key books and papers are listed below.

Fleck B and Svestka Z 1997 First Results from SOHO *Solar Physics* **170 & 175**

Mariska J T 1992 *The Solar Transition Region* (Cambridge: Cambridge University Press)

P Brekke

Classes of terms:

DOPPLER EFFECT
ACTIVE REGIONS
EXPLOSIVE EVENTS
MULTIPLE FLOWS
MAGNETIC RECONNECTION
FLARES

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Transition Region Models

The SOLAR TRANSITION REGION comprises the PLASMA between the CHROMOSPHERE and the CORONA. In both of these regions the temperature is fairly uniform. The transition region, by contrast, is believed to be characterized by a very steep temperature rise from a chromospheric temperature of slightly less than 10^4 K to coronal temperatures on the order of 10^6 K. The goal of modeling the transition region is to construct physical models, i.e. the run of density, temperature, velocity with height, such that the physical processes dominating the region are illuminated and the important aspects of the line emission from the transition region may be reproduced. As hydrogen is fully ionized and therefore cannot radiate in the transition region, radiative emission is mainly in the form of the resonance lines of (multiply) ionized metals; the ions of carbon, silicon, nitrogen, oxygen, neons and iron dominate.

Basic principles

The outer layers of the Sun require a mechanical, as opposed to radiative, heat input in order to support energy losses through radiation, heat conduction, and SOLAR WIND ACCELERATION. Much can be learned about the outer layers of the Sun, and in particular about transition region structure, by considering how this dissipated mechanical energy flux is disposed of. In the chromosphere, where the density is comparatively large, a dissipated energy flux can easily be radiated away by slightly raising the temperature that is required to maintain radiative equilibrium. In the less dense transition region and in the corona other loss mechanisms dominate: assume for a moment a sun without a corona and consider the radiative losses as a function of height. The scale height (the height over which the plasma pressure drops by a factor of e; $H_\mathrm{p} = kT/mg$) in the chromosphere is on the order of a few hundred kilometers. As we proceed outward from the photosphere the gas becomes rapidly less dense and as we approach the upper chromosphere, say 1.5–2 Mm or roughly 10 scale heights above the photosphere, the radiative losses which scale with the product of the electron and total hydrogen density ($n_\mathrm{e} n_\mathrm{H}$) decrease exponentially. Any excess mechanical energy flux dissipation will therefore necessarily overwhelm the plasma's ability to radiate and its temperature will rise dramatically: a corona will form. The temperature of the material will continue to rise until energy losses balance the excess mechanical energy flux dissipation. In the corona radiative losses remain ineffective and there are two potential energy sinks: through inward heat conduction (with subsequent radiative losses from the upper chromosphere) and, in magnetically open regions, through the acceleration of the solar wind. The temperature sensitivity of these processes will almost invariably ensure that the coronal temperature is on the order of 10^6 K and that the transition region has a rapid temperature rise, from 10^4 K to 10^6 K in roughly 1 Mm, almost regardless of the form or magnitude of the CORONAL HEATING MECHANISM.

Spatial structure

The corona receives an energy flux of approximately 500 W m^{-2} in magnetically open regions and perhaps 2–20 times that in closed and active regions respectively. Let us consider the temperature structure that arises in a magnetically closed region. We assume that all the energy deposited in the corona is conducted back towards the chromosphere through the transition region, and that radiative losses are not energetically important in the corona or transition region. The temperature structure in the transition region is then given by solving the energy equation

$$F_\mathrm{c} = \kappa_0 T^{5/2} \frac{\mathrm{d}T}{\mathrm{d}z} \approx F_\mathrm{m0} \qquad (1)$$

where F_m0 is a constant that represents the mechanical energy flux dissipated in the corona. As an example let us set $F_\mathrm{m0} = 100$ W m^{-2}, with $\kappa_0 = 10^{-11}$ W m^{-1} K$^{-7/2}$; this results in a temperature gradient $\mathrm{d}T/\mathrm{d}r \approx 3$ K m^{-1} at 10^5 K or, equivalently, a temperature scale height of $H_\mathrm{T} \approx 35$ km. The temperature scale height is much smaller than the transition region pressure scale height $H_\mathrm{p} \approx 5000$ km. Thus, the temperature gradient is expected to be steep while the pressure will be approximately constant across the transition region, which only extends some 1 Mm.

A typical CORONAL LOOP has length on the order of 10–30 Mm while the chromosphere's vertical extent is roughly 2 Mm. The transition region in the loop will be in a location such that the divergence of the conductive flux flowing from the corona balances the radiative losses at the top of the chromosphere. If the energy flux dissipated in the corona increases, the conductive flux received by the chromosphere will also increase: chromospheric material will heat up and 'evaporate' into the corona and the transition region will move down towards a higher pressure. The opposite phenomenon will occur if less energy is deposited in the corona: the material at the top of the chromosphere will radiate more energy than can be replaced by conduction and the material will cool, moving the transition region upward until energy balance is established again. Since the intensity of EUV emission lines formed in the transition region scales with the density squared we expect that the most vigorously heated loops will be the densest and therefore radiate most intensely. In figure 1 we illustrate the effect of varying the energy input into a coronal loop. The temperature structures of several models are plotted showing the variation of the location of the transition region and the transition region temperature gradient. The chromospheric scale height is on the order of 200 km in these models so the density (pressure) at the base of the transition region varies by a factor of 20 between the model with $F_\mathrm{m0} = 100$ W m^{-2} and $F_\mathrm{m0} = 3000$ W m^{-2}. The intensity of the transition region line C IV 1548 Å which is formed at 10^5 K varies from 0.5 W m^{-2} sr^{-1} to 11 W m^{-2} sr^{-1} in the same models; some of the increase expected because of the density is canceled by the increased temperature gradient which reduces the extent of the line-forming region.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

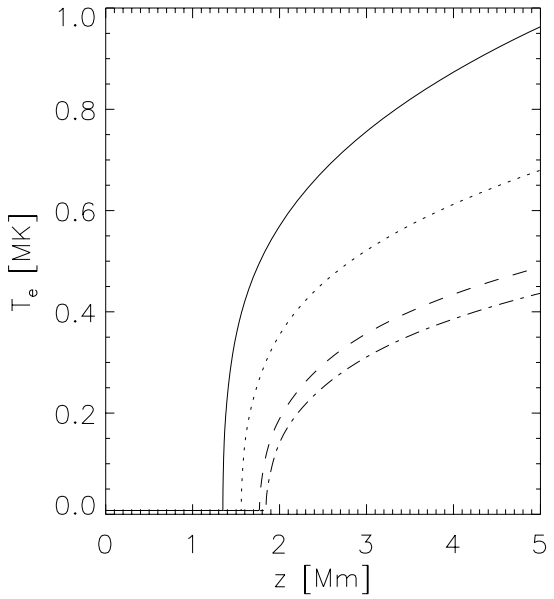


Figure 1. Temperature structure of models used to illustrate the effect of varying the energy flux dissipated in the corona: $F_{m0} = 100$ W m^{-2} (chain curve), $F_{m0} = 500$ W m^{-2} (broken curve), $F_{m0} = 1000$ W m^{-2} (dotted curve), $F_{m0} = 3000$ W m^{-2} (full curve).

Thermal conduction flows freely parallel to the magnetic field and dominates the energy balance in this direction. The heat conductive flux is carried by electrons. Perpendicular to the magnetic field the electrons ability to carry heat conductive flux is severely inhibited since the typical length of interaction between electrons is reduced from mean distance between collisions which applies along the field to the much smaller electron gyro-radius. Thus, while the structure of the transition region along the magnetic field is fairly well established by processes that depend on the plasmas (in)ability to radiate energy, the structure of the transition region across the magnetic field is not as easily inferred. However, it seems clear that the typical size of a structure, or loop, will depend both on the topology of the magnetic field and on the nature of the dissipation mechanism itself, neither of which is well known or understood. Observations clearly show that coronal heating is a sporadic process that seems to operate on several spatial and temporal scales and we should therefore expect that coronal loops will quite disparate in their properties.

Temporal structure

The magnetic field also plays an important role in the dynamics of the transition region and corona. The gas pressure dominates the magnetic pressure in the photosphere and convection zone allowing the magnetic field to be carried around by mass motions. In the corona and transition region the situation is reversed and it is the plasma which must follow the magnetic field as it in turn is churned by mass motions deeper in the atmosphere. In the SOLAR PHOTOSPHERE the magnetic field seems to be confined to small regions. Proceeding upwards through the chromosphere the gas pressure decreases exponentially and the magnetic field should eventually spread out to cover all space.

The speed of sound in the chromosphere is roughly $c_s \approx 10$ km s^{-1} while the total height of the chromosphere is on the order of 1.5–2 Mm. The chromosphere is pervaded by upwardly propagating wave-trains (shocks) with periods on the order of 3 min. These waves are acoustic in nature and are presumably excited by the solar 5 min oscillations. Observations of the chromospheric internetwork seem well accounted for by these waves. In the chromospheric network where the magnetic field is stronger, Alfvén waves or other magneto-acoustic modes of significant amplitude are certainly important, although this has yet to be convincingly demonstrated. The transition region will be buffeted by these waves and will ride like a 'cork' on top of the chromosphere. Because of the rapid variation of the sound speed with height in the transition region some of this wave energy will be reflected back towards the photosphere. We should expect to see evidence for remnants of the characteristic 3 min SOLAR CHROMOSPHERIC OSCILLATIONS in transition region spectral lines. We should also see signatures of the magnetic modes that reside in the chromospheric network or that are generated as a result of the interaction between the acoustic 3 min oscillations and the spreading magnetic field.

The corona may be characterized by several time scales. The conductive cooling time of a 10 Mm long coronal loop is on the order of several hundred seconds. The speed of sound is approximately 100 km s^{-1} while the Alfvén speed, assuming a magnetic field strength of 10 G, is on the order of 1000 km s^{-1}. This gives wave crossing times of 100 s and 10 s respectively for a 10 Mm long loop. Both wave speeds vary rapidly through the transition region and waves generated in the corona will only partially be transmitted through the loop's transition region and will 'bounce' back and forth between the transition region–chromospheric footpoints situated at either end of the loop. Another important time scale is of course set by the energy dissipation mechanism itself. Rapid variations will generate waves that will propagate through the transition region. Likewise, slow variations will lead to flows through the transition region: increased coronal heating requires higher densities supplied by evaporative flows; conversely, reduced coronal temperatures will lead to cooling flows through the transition region.

As all these processes are presumably occurring simultaneously it is of vital importance that the formation of emission region lines is understood so that the signatures of the most important phenomena may be identified.

Spectral line emission

Given the transition region structure as outlined above let us consider the formation of emission lines in detail. Hydrogen is completely ionized in the transition region

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and emission occurs in multiply ionized elements largely in the EUV portion of the spectrum. The plasma is optically thin and the ionization balance is dominated by local collisional processes. We may write the rate equation for ion number density n_i as

$$\frac{\partial n_\mathrm{i}}{\partial t} + \frac{\partial}{\partial z}(n_\mathrm{i} u_\mathrm{i}) = S_\mathrm{i}(n_\mathrm{e}, T_\mathrm{e}) - L_\mathrm{i}(n_\mathrm{e}, T_\mathrm{e}). \quad (2)$$

The source and sink terms are represented, respectively, by S_i and L_i, in the transition region the important processes are collisional ionization and radiative recombinations and di-electronic recombinations. The rates are mainly functions of the local electron temperature T_e and are proportional to the electron density n_e. Notice that in the transition region where the plasma is fairly tenuous these rates are slow (characteristic times vary from 1 s to 10^3 s) and the left-hand side of the equation must generally be retained. However a rough idea of the ionization state of the transition region can be obtained by solving the time-independent part of these equations and a grid of temperatures; we then find that a given ionization state is confined to a narrow region in temperature around its temperature of maximum concentration.

Emission from a given ionization state is determined locally by the balance between collisional excitation and spontaneous de-excitation. The emergent intensity from the transition region and corona, for an optically thin transition of energy $h\nu_\mathrm{ul}$ and an emission profile equal to the absorption profile ϕ_ν, may be written

$$I_\nu = \frac{h\nu_\mathrm{ul}}{4\pi} \int_0^{s_0} \phi_\nu n_\mathrm{i} n_\mathrm{e} C_\mathrm{lu}\, \mathrm{d}s. \quad (3)$$

where $n_\mathrm{e} C_\mathrm{lu}$ is the collisional excitation rate. The integration is carried out along the line of sight, here represented by the path length variable s. A given ion is confined in temperature and thus height in the steep temperature gradient transition region. Integrating over line frequency, assuming that the relative abundance of hydrogen and the element the radiating ion belongs to is constant, and using the expression given above, the total line intensity is seen to vary as

$$I \propto n_\mathrm{e} n_\mathrm{H} x_\mathrm{i} C_\mathrm{lu}\, \Delta z \quad (4)$$

where the variables are evaluated at some appropriate place in the line-forming region. The intensity arises mainly from the region of maximum concentration where $x_\mathrm{i} \approx 1$ and the temperature and thereby C_lu are roughly constant. Because of the steep temperature gradient emission lines formed in the lower transition region will be markers of a plasma confined to a small range in height (lines formed at higher temperatures will have contributions from a larger range of heights and will appear diffuse or 'smeared out').

Numerical models

There are several technical difficulties associated with developing numerical codes designed to cover the transition region. The large temperature gradient that invariably arises requires high spatial resolution as does the even more rapid change in ionization state with temperature. It is not known *a priori* where these large gradients will be located nor whether they will remain in the same position as the model evolves in time. We must also expect that waves generated in the chromosphere will be highly non-linear and shocked as they interact with the transition region implying that we must resolve large gradients in the thermodynamic variables. In addition, the high Alfvén speed in the corona will because of the Courant condition (the time step must be shorter than the grid size divided by the largest characteristic velocity in the system; $\Delta t < \Delta x / v_\mathrm{A}$) force very short time steps in explicit codes in problems where wave propagation is to be studied. For one-dimensional models most of these difficulties have been overcome and there are a variety of numerical codes that can treat these processes. It also seems likely that these technical difficulties should be resolvable for two- and three-dimensional codes now or in the very near future owing to advances in computer technology and to the availability of good MHD and grid refinement algorithms.

A more fundamental difficulty lies in the lack of knowledge about the magnetic field topology as it extends from the photosphere into the corona. This lack of knowledge also impacts our understanding of the spatial extent of the coronal energy flux dissipation mechanism which presumably involves the magnetic field in some manner. This has hampered the development of two- and three-dimensional models of the transition region. It could be argued that since conduction along the magnetic field is so dominant the study of one-dimensional models should be sufficient to reach an understanding of the important processes in the transition region. However, the coupling of the transition region to the chromospheric motions and the interactions between the various wave modes as the magnetic field expands to fill all space is a theme that seems accessible only in two or three dimensions while at the same time seems to be a theme which has great potential for increasing our understanding of the outer solar atmosphere.

Confrontation with observations

Several one-dimensional models designed to explain transition region emission have been constructed since Pottasch's original work describing this emission in 1964.

The solar chromosphere–corona transition region is highly structured, both in space and in time. Intensity observations indicate that the dominant portion of the transition region radiation arises in regions of variable emission. There are, however, two characteristics of the observations that seem to be invariant from region to region and from time to time; the shape of the emission measure curve and the net redshift of spectral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

lines. It should be mentioned that the redshift decreases toward the limb of the Sun indicating that motions are primarily vertical. The invariant shape of the emission measure curve and the net redshift are also seen in the observations of several other stars. Most of the numerical models constructed have been designed to explain these observational facts.

An outstanding problem in transition region physics is that the temperature structure one derives from assuming that the dominant term in the energy equation is thermal conduction does not seem to correctly reproduce the observed emission of spectral lines formed at temperatures below, say, 2×10^5 K but instead predicts intensities several orders of magnitude less than is observed. The cause of this discrepancy is the sharp temperature gradient, due the dominance of thermal conduction, and the associated paucity of material in the temperature range which supports emission from a specific ion. Antiochos and Noci have proposed a promising solution to this dilemma by assuming that much of the lower transition region emission stems from regions where the local heating is low enough to ensure low loop apex temperatures, thus reducing the role of thermal conduction and 'stretching' the temperature profile in the loop. On the other hand it has been argued by Athay and Holzer that transition region emission may the result of episodic heating and that the observed transition region intensities may be reproduced by models of loops that cycle between having high and low apex temperatures. The resolution of this issue awaits a better understanding of spicule dynamics and the coronal heating process.

Analysis of spectral line profiles show that transition region lines display a net redshift (redshifts imply motions towards the solar surface). The observed mean redshift is typically 5–10 km s^{-1} for the spectral line C IV λ1548 which is formed near 10^5 K. The amplitude of the shift is reduced for lines formed at lower temperatures and chromospheric lines display little net shift. There are few observations of higher-temperature lines but there is evidence that the net redshift is reduced or disappears in the upper transition region. Models of the net redshifts have invoked selection effects in the radiative or temporal properties of gas flowing up versus down. Athay and Holzer hypothesized that the redshifts were caused by the return of spicular material. Alternately, the observational signatures have been interpreted as siphon flow through coronal loops. McClymont and Craig, and Mariska, showed that asymmetric heating of coronal loops may drive such flow. Both ideas seem compelling; however, the development of both into detailed models has been hampered by the lack of numerical computations or, in the case of siphon flow, by the fact that calculations by Spadero *et al* predict blueshifts rather than the observed redshifts when time-dependent ionization is incorporated into the models. Hansteen has proposed that the net redshifts may be explained as a consequence of nanoflare (episodic) coronal heating: episodic heating in a coronal loop should generate waves, some of which will propagate along the loop and through the transition region. Propagating compressive waves will cause spectral line asymmetries and line shifts in the 'direction' of wave propagation which in this scenario are primarily directed downwards producing redshifts. The further development of this idea awaits two- and three-dimensional simulations as well as theoretical predictions on the nature of episodic energy release in the corona.

As we have seen, transition region emission is fairly sensitive to processes occurring in both the chromosphere and corona and could to used to effectively discriminate between various models of these regions. The availability of observations from the Solar and Heliospheric Observatory (SOHO) instrument Solar Ultraviolet Measurements of Emitted Radiation (SUMER) which feature transition region data of unprecedented temporal, spectral and spatial quality should push transition region modeling into considering full two- and three-dimensional models of the outer solar atmosphere where the details of the interaction between the chromosphere, the transition region and the magnetic field are treated with more seriousness than has been the case until recently.

Bibliography

Antiochos S K and Noci G 1985 The structure of the static corona and transition region *Astrophys. J.* **301** 440—7

Hansteen V H 1994 A new interpretation of the redshifts observed in optically thin transition region lines *Astrophys. J.* **402** 741–55

Mariska J T 1992 *The Solar Transition Region* (Cambridge: Cambridge University Press)

Pottasch S R 1964 On the interpretation of the solar ultraviolet emission line spectrum *Space Sci. Rev.* **3** 816–55

Viggo Hansteen

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Transition Region: Explosive Events

Explosive events are frequent, very small bursts that occur throughout the solar transition region. There are something like thirty to forty thousand of these events on the Sun at any time. Each event lasts for roughly a minute so every second there are about 600 new events. Present-day solar telescopes are just able to resolve the largest of these events. The events give an impression of explosion because highly supersonic mass motions (measured as Doppler shifts) are nearly instantaneously produced within a small area. The velocities reach up to 150–200 km s^{-1} during the first few seconds and they last for 1–5 min. The typical size is 1000–4000 km and there is no detectable transverse motion of the event site. Explosive events are most clearly seen in ultraviolet spectral lines formed at temperatures of about 10^5 K. The few measurements of number densities indicate densities of about 5×10^{15} m^{-3}. Their mass and kinetic energy are estimated to be 5×10^5 kg, and 10^{15} J respectively.

A distinguishing feature of explosive events is that they produce large Doppler shifts in emission lines formed at transition region temperatures. Emission lines from plasma at these temperatures are all found in the ultraviolet wavelength range (as explained in the SOLAR TRANSITION REGION). The image in figure 1 shows line profiles of an ultraviolet line from the three times ionized silicon ion, along a long, narrow strip of the Sun. This line, Si IV 1393 Å, is formed at roughly 10^5 K. Three prominent explosive events all with very different but characteristic line profiles can be seen. In the event at the top, the red- and blueshifted emission is slightly offset along the N–S directed spectrometer slit. The event in the middle has a highly asymmetric blueshifted profile (strongly redshifted profiles are also common). It is also typical that the larger, brighter events, such as the lower event in the image, have higher Doppler velocities.

Explosive events have now been observed in many spectral lines. They are most often seen in lines such as C IV 1550 Å, Si IV 1393 Å and O IV 790 Å that are formed at plasma temperatures near 10^5 K. Higher temperature events that show up in lines such as Ne VIII 780 Å formed near 5×10^5 K are about ten times less frequent. Likewise events exhibiting emission in lines from lower temperature ($2–3 \times 10^4$ K), singly ionized species, for example in C II 1335 Å, are also about ten times less frequently seen. The high velocity ultraviolet line shifts that characterize explosive events are usually accompanied by a brightening of the line intensity; however the converse is not true. For example, the brightening of transition region lines associated with x-ray bright points is not usually accompanied by explosive events (see SOLAR CORONA: X-RAY BRIGHT POINTS).

The fact that explosive events are only observed in ultraviolet spectra which must be obtained from space telescopes makes them difficult to study. Their small size and comparatively short lifetime requires high spatial, temporal and spectral resolution. So far only two instruments have had high enough resolution to unambiguously identify explosive events. They were first discovered in spectra from the High Resolution Telescope and Spectrometer (HRTS) of the Naval Research Laboratory during its first rocket flight in 1975. Subsequent rocket flights and a Spacelab 2 mission with this spectrometer enabled a basic understanding of their observational statistics, their relationship with x-ray bright points and the photospheric magnetic field. In 1996 with the launch of SUMER (Solar Ultraviolet Measurements of Emitted Radiation), the high resolution ultraviolet spectrometer on board the spacecraft SOHO (the Solar and Heliospheric Observatory) it became possible to observe the relationship of explosive events to the network structure over periods of several hours.

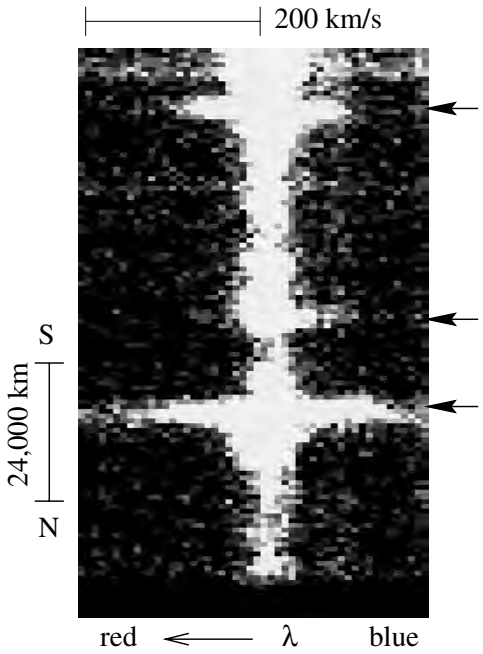


Figure 1. An example of three explosive events in Si IV 1393 Å observed with the long slit ultraviolet spectrometer SUMER on the spacecraft SOHO. The image shows line profiles along the 90 000 km long slit with spatial resolution 800 km, velocity resolution 12 km s^{-1} and integration time 10 s. Blueshifts (decreasing wavelength) are on the right.

The results of one such study are shown in figure 2 where the supergranular network pattern (see CHROMOSPHERE: NETWORK) of bright cell boundaries with dark centers is shown on the left and the evolution of the line profiles at four representative positions on the right. These four positions lie on a bright network lane (A), a dark cell center (C) and faint network lanes (B and D).

This figure illustrates the frequency of explosive events and their relationship to the network structure. Most explosive events and brightenings are in the network lanes and, in particular, close to bright emission regions. There are also one or two faint, low-velocity ($\sim$60 km s^{-1})

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

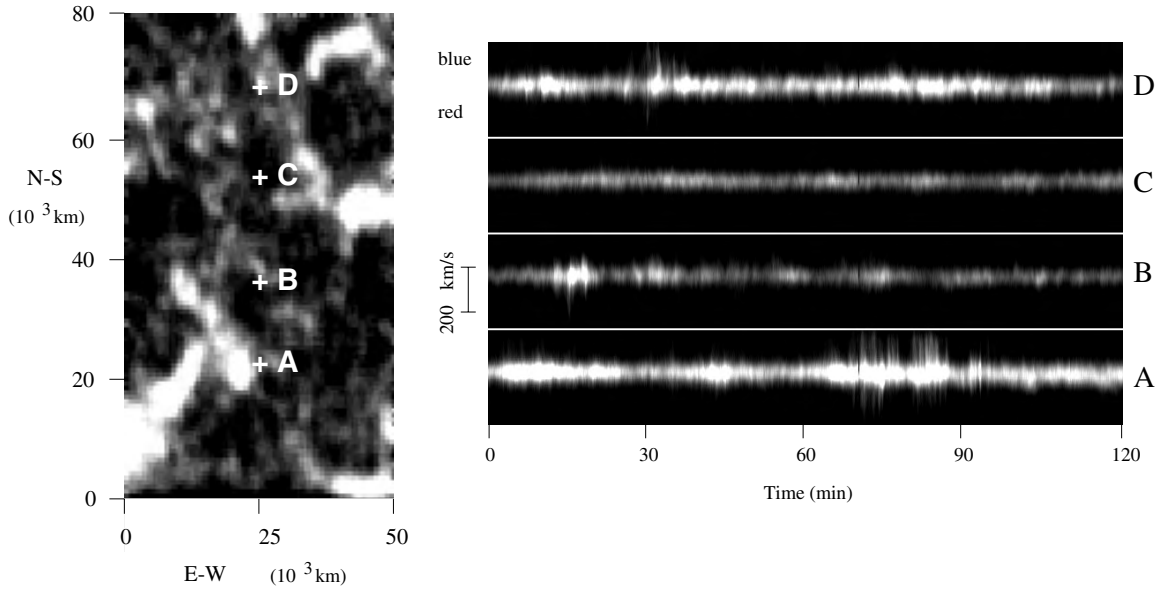


Figure 2. Explosive events and their relationship to the chromospheric network as observed by the spectrometer SUMER. The picture on the left shows the structure of the chromospheric network. The positions chosen to demonstrate typical line profile evolution are marked with white crosses and letters. The time evolutions of the Si IV 1393 Å line profiles at these four positions over a period of two hours are shown in the right-hand frames. In each frame, the profiles range from blueshifts of about 150 km s^{-1} at the top to redshifts of about 150 km s^{-1} at the bottom.

events in the dark cell center. Along the network lanes, the individual events often occur in bursts that can last up to 30 min.

Since their discovery, it has been suggested that the underlying process triggering the events is the conversion of magnetic to kinetic energy by magnetic reconnection. There is also strong observational evidence that the events are related to the structure and evolution of the photospheric magnetic field. As illustrated in figure 2, explosive events occur predominantly along network lanes. This is where weaker magnetic field elements are continually merging and cancelling with stronger network fields. In recent ultraviolet line profile and magnetic field observations, several explosive events were seen simultaneously with evolving magnetic field elements. Figure 3 shows the position of the explosive events in relation to the magnetic field and it can be seen that most events lie on or near magnetic neutral lines (i.e. the boundary between positive and negative field regions). The small burst of events on the left side of the positive field region marked with an arrow, seems to be associated with decreasing photospheric flux. Some explosive events have been seen far from any detectable magnetic fields. These are however always single events, not bursts of events, and it is possible that very small magnetic field elements are involved.

Another clue to the origin of explosive events comes from their spatial flow structure. By making small repetitive scans with a long slit spectrometer it is possible to build up an image of the time evolution of the explosive event flow. Figure 4 illustrates the evolution of the time

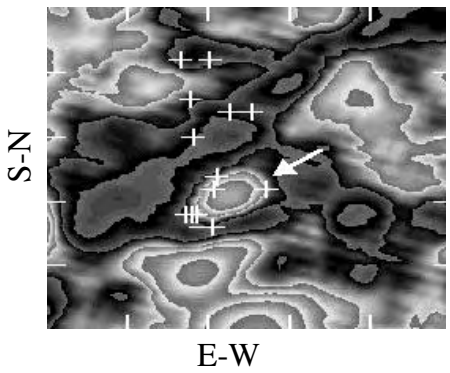


Figure 3. A quiet Sun magnetogram with the position of explosive events marked with white crosses. The brighter (darker) features are positive (negative) polarity magnetic fluxes. The tick interval on the E–W and S–N axes is 3600 km. (Courtesy of J Chae, Big Bear Solar Observatory.)

pattern of observed Doppler shifts. As in figures 1 and 2, this again is a series of Si IV line profiles. One of the most important features illustrated in figure 4 is that the Doppler shifts change from red to blue as the bright central position of the event is scanned in the E–W direction. Closer inspection shows that the width of the event slowly increases in both the E and W directions although the position of brightest emission does not move.

These features show that the word 'explosive' is misleading. An explosion would eject material in all directions and should produce symmetric red- and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

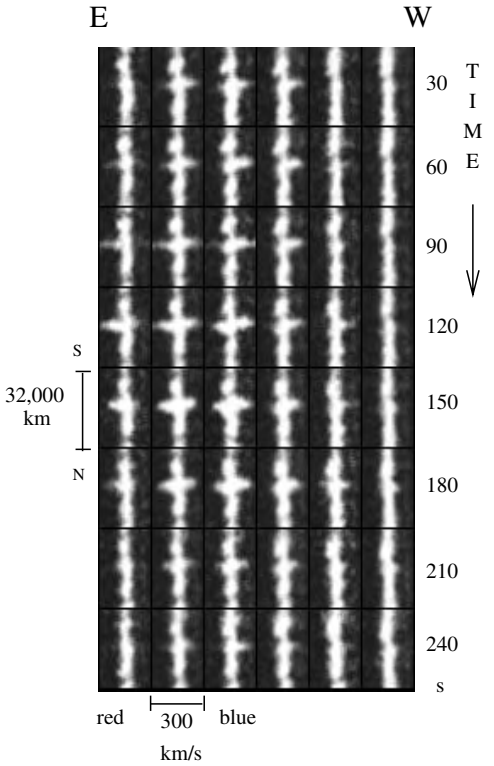


Figure 4. Evolution of Si IV line profiles. Each sub-plot of the figure shows a 32 000 km section of a spectrum along the slit for Doppler velocities from redshifts of 150 km s^{-1} to blueshifts of 150 km s^{-1}. Time increases from top to bottom. Each row shows the profiles along an E–W scan with a step size between subplots of 800 km.

blueshifted profiles. Also material that has moved across the disk furthest from the explosion site would have the lowest line-of-sight velocity. Therefore the high-velocity asymmetric profiles on the east and west edges of the events cannot be explained by an explosion. There are two flow configurations that can produce the pattern observed in figure 4. One is a bi-directional jet and the other is a spinning disk of gas. At present no-one has been able to find an observational means of determining which flow it is.

If the flows are bi-directional jets then the observed pattern of red and blue Doppler shifts could be a confirmation of the Petschek theory of fast reconnection. This theory predicts oppositely directed plasma jets ejected from a small localized site. As described elsewhere (see MAGNETOHYDRODYNAMICS: MAGNETIC RECONNECTION AND TURBULENCE) magnetic energy is converted into plasma thermal and kinetic energy at the jets' source, the diffusion region, and along stationary slow mode magnetoacoustic shocks that form along the jets' boundaries (figure 5). Petschek models reproduce the high-velocity components but they do not account for the bright emission at the center of the line profiles. It is likely that the diffusion region is more complex than the small localized site assumed in the

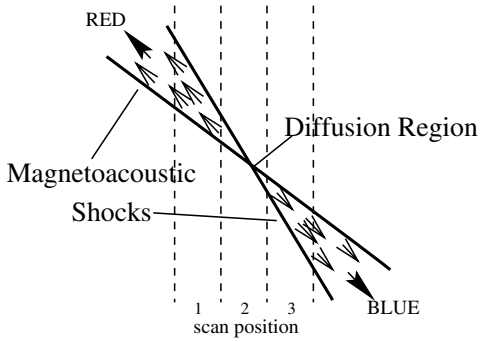


Figure 5. Flow geometry of jets that would reproduce the observed pattern of red and blue Doppler shifts shown in figure 4. In Petschek-like models the reconnection site would be at the centre of the bi-directional jet.

Petschek theory.

Explosive events play a crucial part in the understanding of quiet Sun transition region heating, flows and most probably also the solar corona. The total energy in the observed mass motions of explosive events is insufficient to be important for coronal energy balance but they probably represent the high-energy tail of a spectrum of network events that occur on scales unobservable with currently available techniques.

Bibliography

Brueckner G E and Bartoe J-D F 1983 Observations of high-energy jets in the corona above the quiet Sun, the heating of the corona, and the acceleration of the solar wind *Astrophys. J.* **272** 329–48

Dere K P, Bartoe J-D F and Brueckner G E 1989 Explosive events in the solar transition region *Solar Phys.* **123** 41–68

Innes D E, Inhester B, Axford W I and Wilhelm K 1997 Bi-directional plasma jets produced by magnetic reconnection on the Sun *Nature* **386** 811–13

Davina Innes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Transition Region: First Ionization Potential Effect

The first ionization potential (FIP) effect causes a fractionation, i.e. a change of the relative abundances, of elements with a low FIP in the corona and solar wind as compared with the SOLAR PHOTOSPHERE. These elements (specifically Mg, Si and Fe) are becoming enriched somewhere in the solar atmosphere by an average factor of 2–5 relative to the high-FIP elements (specifically oxygen). The process most likely involves separation of atoms and ions across magnetic field lines in a partially ionized gas, and the site can probably be identified with the upper chromosphere or the lower transition region. An understanding of the FIP fractionation effect is important for obtaining element abundances in the Sun from solar wind measurements, and in turn it provides information on the conditions prevalent and the processes operating at the relevant site in the solar atmosphere.

FIP fractionation observations

When we compare the normalized elemental abundances (relative to oxygen) of the SOLAR WIND and of the CORONA (as inferred from the solar energetic particle (SEP) population) with the photospheric abundance ratios, we notice a significant fractionation between the two populations. If this fractionation is plotted as a function of the FIP, a regular pattern emerges. Elements with a low FIP ($\lesssim$10 eV) lie on a plateau which is enhanced by a factor of $\sim$3–5 with respect to elements with a high FIP ($\gtrsim$10 eV). The intermediate elements with an FIP near 10 eV, carbon and sulfur, are also enriched, but by smaller factors. The high-FIP elements again lie on a plateau up to, and including, neon, but helium, the element with the absolute highest FIP, is underabundant by a factor of $\sim$2 relative to the high-FIP elements. This basic FIP pattern is obeyed by the SEP population and, with some exceptions discussed below, by the slow solar wind normally encountered near the ecliptic plane (see figure 1). It was first described for the SEP population by Hovestadt in 1974, and in 1981 Meyer first noted the same effect to hold in the slow, in-ecliptic solar wind.

With the abundances of more elements becoming available, and with the possibility to determine abundances as a function of the solar wind flow type, exceptions from the basic FIP pattern described above have become apparent. The most important one is a reduction of the fractionation factor between low- and high-FIP elements, the FIP bias, in the fast solar wind from CORONAL HOLES (i.e. in high-speed streams) to a factor of $\sim$2 instead of $\sim$3–5 (figure 1). It is interesting to note that carbon, which is slightly enriched relative to oxygen in the slow wind, seems to remain unchanged in fast streams, i.e. its slight enrichment is not reduced proportionally with that of the low-FIP elements. It may therefore be that this element behaves as a (slightly enriched) high-FIP element in the slow, interstream solar wind but as a low-FIP element in fast streams, even though its enrichment factor does not change.

The heavy noble gases, krypton and xenon, do not follow the basic FIP pattern either. Both are high-FIP elements, but both are enriched in the solar wind, by factors of 2 (Kr) or even 4 (Xe). (These abundances were obtained from inclusions in the lunar regolith, which were integrated over millions to billions of years; therefore the values are identified with the slow, in-ecliptic solar wind, and no fast wind value can be given.) This enrichment may be understood to be a consequence of their high mass (see next section), and of their relatively short ionization time as compared with other high-FIP elements. It has been argued that the first ionization time (FIT) may therefore be the better ordering parameter than the FIP. Ionization times are calculated by folding the solar UV and EUV spectrum with the ionization cross sections. Indeed, when the abundance data are represented versus FIT, a simple new fractionation pattern is easily visible: low-FIT ($\lesssim$10 s) elements again lie on a plateau, but the fractionation factor of high-FIT elements now is roughly a power law in FIT with an exponent of $\simeq -0.5$, and all high-FIP elements, including Kr, Xe and He, follow this pattern (see figure 2). However, the FIT is not a fundamental atomic parameter, and it is highly variable as a consequence of the UV and EUV variability with the solar activity cycle. Therefore, the name of the 'FIP' effect should nevertheless be preferred.

Recently, the FIP effect has also been observed by remote sensing techniques, using the SUMER telescope on the SOHO spacecraft. By comparing optical emission lines originating from the corona with lines from the CHROMOSPHERE, the existence of an FIP bias of $\sim$4 was observed in coronal plasmas above quiet equatorial regions from where the slow wind originates, while this factor was much smaller above coronal holes, from where the fast solar wind streams emanate, thus confirming the *in situ* measurements.

The same spectroscopic technique can even be applied to other stars than the Sun (but obviously only as a full-disk integration). The existence of an FIP effect was found in cool stars such as αCen AB and ϵEri, but not in the hotter star αCMi (Procyon).

FIP fractionation models

From the shape of the FIP fractionation curve, we obtain two key ingredients for the construction of models. (1) Any mechanism is to operate at a region in the solar atmosphere where it is (at least partially) neutral, or else it is not clear how the FIP (or FIT) could be a relevant parameter. Many different possibilities have been considered in the models for this 'mode of separation'. (2) The step near 10 eV suggests that photoionization by Lyman-α photons (10.2 eV) may play an important role (particularly for C, which has low-lying metastable levels that can be ionized by Lyman-α, whereas the ground state cannot). The flux in the Lyman-α line is more than a factor of 2 higher than the integrated flux at shorter wavelengths. Photoionization should therefore be included in any attempt to model the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

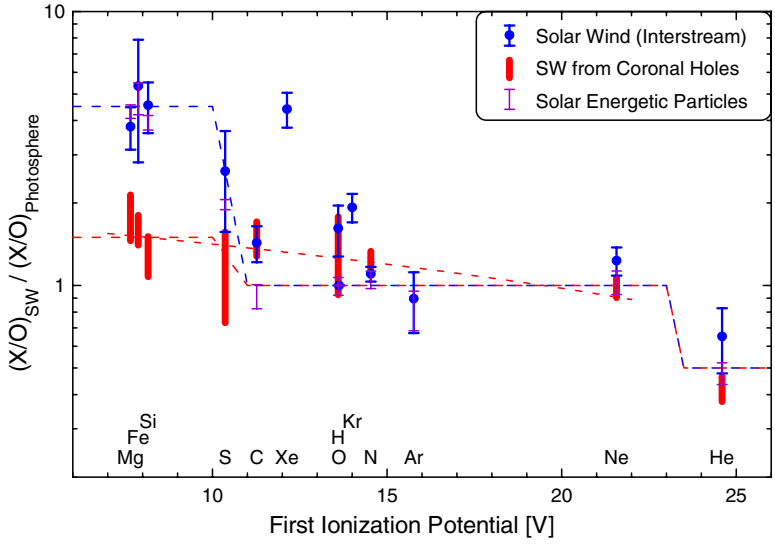


Figure 1. Abundances of elements measured in the solar wind relative to their photospheric abundances versus their FIP.

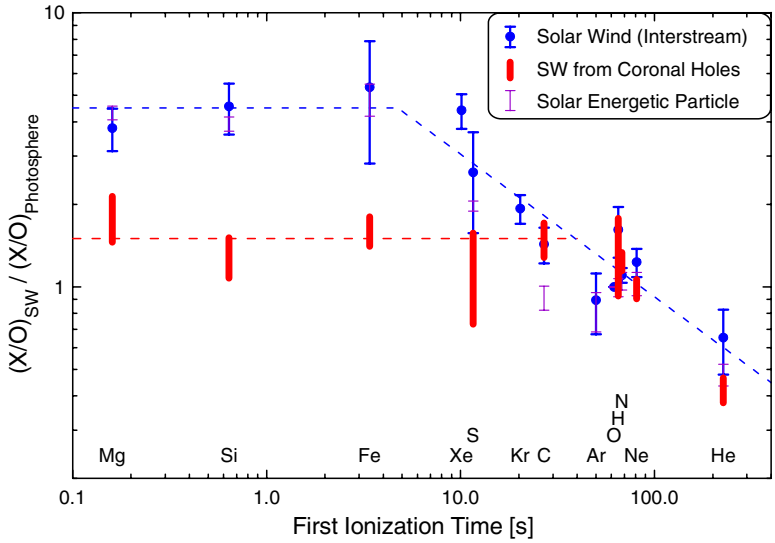


Figure 2. Abundances of elements from figure 1, plotted versus their FIT.

FIP fractionation, and most models indeed use it for the 'mode of ionization'.

Over the past decade, several models have been proposed to explain the observed FIP fractionation, most of which operate by ion–neutral separation in the chromosphere, predominantly by diffusion perpendicularly to a magnetic field. Geiss, in 1982, has shown that ion–neutral separation is much more effective than ion–ion separation and gives quantitative estimates for the conditions under which it could take place: temperature $T \lesssim 10^4$ K and density $n \lesssim 10^{16}\,\mathrm{m}^{-3}$, i.e. in the upper chromosphere and lower transition region. The density estimate was obtained with gravity as the only driving force for the diffusion of neutrals; if other driving forces are considered, the limit on n may be higher.

The first attempt was to consider a model in which ions and neutrals separate in the chromospheric canopy, where the magnetic field is horizontal, 1000–2000 km above the photosphere. Neutral atoms may settle downward, driven by gravity, but the diffusion of ions is impeded by the field. This reduces their diffusion velocity by a factor of $(1 + \omega^2 t_{\mathrm{col}}^2)^{-1}$, where ω is the gyrofrequency and t_{col} is the typical time between collisions. The degree of ionization of the elements considered was not calculated self-consistently in this model, but rather taken from an existing non-LTE model atmosphere. It was found that an FIP-dependent fractionation may indeed be obtained in such a model, but the resulting diffusion times needed to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

separate the species are much larger—hours to days—than the typical lifetime of structures in the solar atmosphere. This was not surprising, as the density in the adopted model atmosphere was higher throughout the considered region than the limiting density for gravity-driven ion–neutral separation mentioned above. The idea may be saved by taking a dynamical approach along the same line, in which the horizontal magnetic field is rising from the photosphere through the chromosphere and into corona, lifting an overabundance of low-FIP elements (the 'skimmer' model). Preliminary results look quite promising if a field of 5 G at a velocity of 5 km s^{-1} is considered.

In a quite complementary approach, the excitation and ionization of elements under chromospheric conditions were explicitly investigated, at first without specifying the mechanism of ion–neutral separation. A neutral gas mixture is exposed to the solar spectrum in an optically thin environment. Excitation and ionization by photons made the main contribution, but collisional ionization by electrons ($n_e = 10^{16}$ m^{-3}, $T_e = 10^4$ K) as well as radiative recombination are also taken into account. The resulting FIT (used in figure 2) suggest that any separation of neutrals from ions has to occur on short time scales of the order of 1 min, as this is the typical ionization time of high-FIP elements once they become exposed to the EUV radiation from the corona. This in turn suggests that the separation is to operate on small length scales, in order to keep the relative velocities within reasonable limits.

The first attempt to construct a model considering both ionization and separation operated on an initially neutral gas mixture of photospheric composition and uniform temperature, concentrated in narrow, vertical or nearly-vertical magnetic structures. The gas mixture is assumed to rise along the magnetic field, and at the same time to become excited and ionized by EUV photons and by electrons; radiative recombination is also considered. The newly created ions remain attached to the magnetic field, i.e. trapped in the upward flow. The neutrals, on the other hand, are driven across the field lines by an initial density gradient and lost in a downward stream outside the structure under consideration. Thus the resulting gas mixture exiting at the top of the structure is enriched in low-FIP elements, since they have been ionized more quickly, and fewer neutrals were lost from the structure. The resulting gas mixture obtained in a typical model run compares favorably with the slow solar wind composition: not only is the step between low-FIP and high-FIP elements reproduced, but also the critical elements near this step, C and S, are obtained partially enriched, as observed, and even the 'FIP exceptions', Kr and Xe, result roughly with their observed enrichment. The model parameters were $n = 10^{15}$ m^{-3}, $T = 10^4$ K, and $d = 10$ km (d is the width of the considered structure). Moreover, the reduced step height observed in fast streams can also be accounted for by the model: it is sufficient to consider less narrow structures, of the order of 30–100 km, to obtain an FIP step height of a factor of 2 only. This may be interpreted with a chromospheric structure that could be coarser beneath coronal holes than under closed field regions. However, the model has been criticized, less for the need of an (as yet) unobservable chromospheric fine structure and of a specific geometry in which the mechanism operates, but for its need of relatively low densities, because at $n \gtrsim 10^{16}$ m^{-3} the reduced mobility of the neutrals would no longer allow them to leave the considered structure.

A few authors have tried to overcome this difficulty by constructing a similar model, which was retaining the careful treatment of ionization and collisions of the one described above, but considering a less specific model geometry. The fractionation was obtained by diffusion across a horizontal slab, again along a (nearly) vertical magnetic field. For specific, but naturally chosen, boundary conditions at the slab interfaces, the observed FIP fractionation pattern in the slow solar wind with a step height of ~5 was obtained at the upper slab interface, as was the depletion of helium relative to the high-FIP elements. The crucial boundary condition is that the flux at the top of the slab is the same for each heavy ion species in natural units: density normalized to one at the slab bottom, and velocity expressed in units of the natural diffusion speed, $(D\tau)^{1/2}$, where D is the diffusion constant in the ambient main gas and τ is the ionization time. The fractionation obtained in this model was found to be almost independent of the background density and temperature for realistic chromospheric conditions. On the other hand, the agreement for Kr and Xe is less good than it had been in the model described above. Moreover, an adaption of the model to yield the reduced FIP bias observed in fast streams is not straightforward either. It could possibly be obtained by arguing that diffusive equilibrium is not quite reached beneath coronal holes, resulting in an incompletely fractionated mixture being output at the top of the slab and into the fast solar wind.

A class of recently proposed models works with diffusion driven by electromagnetic forces, either (1) by dragging ions into a chromospheric gas parcel that is to be supplied to the corona or (2) by expelling neutrals out of such a parcel. A mechanism of type (1), investigated qualitatively, considers a small region of the chromosphere being heated by downward heat flux from the corona along a magnetic flux tube, leading to localized evaporation and lifting of chromospheric material. This, in turn, produces a horizontal thermoelectric field, dragging ions from the surroundings into the uplifting region. This leads to an accumulation of low-FIP elements in that region, since they are already ionized in the surroundings, whereas the high-FIP ions still are predominantly neutral there. A mechanism of type (2) was investigated, again qualitatively, by considering a cylindrically symmetric magnetic field structure in a partially ionized atmosphere, and calculating the system of direct currents created by such a configuration. It could be shown that, near the periphery of the flux tube, there is a strong diffusion of neutrals relative to the ions out of the structure, which may produce the observed FIP fractionation. However,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

this class of models has yet to be worked out quantitatively and to make predictions that can be compared with the observations (see also SOLAR PHOTOSPHERIC MAGNETIC FLUX TUBES: OBSERVATIONS; THEORY).

Finally, the FIP fractionation may be caused, or supported, by waves on a magnetic field in a partially ionized gas mixture, an idea put forward already in 1989. Such waves could heat the mixture and thus lead to the preferential loss of neutrals while the ions remain bound to the field. Again, in order to assess the value of this idea, quantitative predictions are needed, but no worked-out models of this kind exist in the published literature as yet.

In summary, none of the nearly a dozen FIP fractionation models that have been published is without its shortcomings, and much detailed work remains to be done in order to pinpoint the location and the mechanism(s) producing the observed FIP fractionation.

Bibliography

For a recent and comprehensive account on all aspects of solar composition, including many theoretical and observational articles on FIP fractionation, see

Fröhlich C, Huber M C E, Solanki S K and von Steiger R (ed) 1998 *Solar Composition and its Evolution—from Core to Corona (ISSI Space Sciences Series 5)* (Dordrecht: Kluwer)

Two earlier review articles on the topic are

Meyer J-P 1993 Element fractionation at work in the solar atmosphere *Origin and Evolution of the Elements* ed N Prantzos, E Vangioni-Flam and M Cassé (Cambridge: Cambridge University Press) pp 26–62

von Steiger R, Geiss J and Gloeckler G 1997 Composition of the solar wind *Cosmic Winds and the Heliosphere* ed J R Jokipii, C P Sonett and M S Giampapa (Tucson, AZ: University of Arizona Press) pp 581–616

Rudolf von Steiger

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trans-Neptunian Object (TNO)

An outer Solar System object whose orbit has a semi-major axis greater than that of Neptune (30.1 AU). Such objects comprise the Kuiper Belt. They are given asteroidal designations, though they are not asteroidal in nature, but most likely planetesimals that have remained unmodified since the early history of the Solar System. The TNOs are often classified according to their orbital characteristics. The first group, which accounts for about 50% of known TNOs, are known as 'classical' Kuiper Belt objects, or *cubewanos*—after the designation of the first to be discovered, 1992 QB_1. These objects have orbits of quite low eccentricity and mean radius 42–47 AU. About 40% of TNOs are, like Pluto, in a 3:2 resonance with Neptune in orbits of mean radius 39 AU: for every three orbits of Neptune, they complete two. For this reason they are sometimes referred to as *plutinos*. This 'resonant' class also includes objects having other orbital resonances with Neptune: about 5% of TNOs show a 2:1 resonance and have mean radius 48 AU; other objects have a 4:3 or 5:3 resonance. There are also a small number of so-called *scattered disk objects* (SDOs), characterized by high eccentricities, the best-known example being 1996 TB_{66}, whose distance from the Sun varies from about 35 to 135 AU. Such objects are surmised to have been perturbed, early in the history of the solar system, from their original orbits in the Uranus–Neptune region of the Kuiper Belt. The orbits of SDOs are dynamically similar to those of the Centaurs, with which they are sometimes grouped. By April 2000 a total of 270 TNOs had been found. It is estimated that there may be ten thousand of them 200 km in size.

In 2001, a trans-neptunian object known as 2001 KX76 was discovered that may be the largest minor body in the solar system. Its diameter, which is not very well determined, lies in the region of 1200 km, possibly even 1400 km. This makes it significantly larger than CERES. Ceres, the first asteroid (minor planet) to be discovered in the solar system, has held the record as the largest known object of its kind for two centuries with a diameter about 950 km.

See: CENTAURS, CUBEWANO, KUIPER BELT, PLANETESIMAL.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Trapezium

The multiple star θ^1 Orionis, in the heart of the Orion Nebula, visible when viewed directly with moderately powerful telescopes and on short-exposure photographs of the nebula; it can be seen on long-exposure photographs only where image-improving techniques have increased the contrast with the surrounding nebulosity. It consists of four stars arranged in a pattern resembling an irregular trapezium, with sides ranging from 8.6″ to 18.4″ and diagonals of 16.9″ and 20.6″.

Identification of the individual stars can be rather confusing; they were formerly designated θ^1 Ori A, B, C and D in order of increasing right ascension; in the *Hipparcos Catalog*, however, they are lettered in order of decreasing brightness. Clockwise from the southernmost component they are now designated A (formerly C), C (formerly D), D (formerly B) and B (formerly A). Their apparent magnitudes are A, 4.96; B, 6.35; C, 7.20; D, 7.49.

The faintest component, θ^1 Ori D, is an eclipsing binary with a period of 6.5 days and is also known as BM Orionis; θ^1 Ori B has recently been discovered to be an eclipsing binary also, with a period of 65.4 days. All four components are hot, blue-white stars of spectral types O and B, and together with several other fainter stars in the region form an embryonic cluster of stars newly formed from the nebular material.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Triangulum

(the Triangle; abbrev. Tri, gen. Trianguli; area 132 sq. deg.) A northern constellation which lies between Andromeda and Aries, and culminates at midnight in late October. Its origin is uncertain, though the ancient Greeks likened it to a capital delta (Δ) and it was also associated with the Nile delta and the island of Sicily. Its brightest stars were cataloged by Ptolemy (*c.* AD 100–175) in the *Almagest*.

A small but distinctive constellation, the brightest stars of Triangulum, β Trianguli, magnitude 3.0, α Trianguli (Mothallah), magnitude 3.4, and γ Trianguli, magnitude 4.0, form a narrow isosceles triangle, with α at the apex. Interesting objects include 6 Trianguli, a multiple star system with yellow (G5) and pale yellow (F5) components, magnitudes 5.3 and 6.8, separation 3.9″, each of which has an unseen companion that revolves around its primary in 14.73 days and 2.24 days, respectively, and the Triangulum Galaxy (M33, NGC 598, sometimes called the Pinwheel Galaxy), a sixth-magnitude spiral galaxy and member of the Local Group, about 2.6 million light-years distant.

See also: Triangulum Galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Triangulum Australe

(the Southern Triangle; abbrev. TrA, gen. Trianguli Australis; area 110 sq. deg.) a southern constellation which lies between Norma and Apus, and culminates at midnight in late May. It was first shown in Johann Bayer's atlas *Uranometria* of 1603, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7. Another southern triangle, Triangulus Antarcticus, was shown to the south of Argo Navis on a celestial globe by the Dutch theologian and geographer Petrus Plancius in 1589, but was never accepted.

A small but conspicuous constellation, Triangulum Australe is easily recognized by virtue of its three brightest stars, α Trianguli Australis (Atria), magnitude 1.9, β Trianguli Australis, magnitude 2.8, and γ Trianguli Australis, magnitude 2.9, which form an almost equilateral triangle. Other bright stars include δ Trianguli Australis, magnitude 3.9, and ε Trianguli Australis, magnitude 4.1. Interesting objects include NGC 6025, an open cluster of about 60 stars, the brightest of which is of magnitude 7.3.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Triangulum Galaxy

A nearby spiral galaxy in the constellation of Triangulum. Otherwise known as M33, it is a face-on spiral that lies at a distance of 2500 000 light-years (slightly further away than the much larger Andromeda Galaxy) and is a member of the Local Group of galaxies. With a diameter of about 40 000 light-years and a population of about 15 billion stars, it is about half the size, and has about one-tenth of the mass, of the Milky Way Galaxy. With a relatively small central bulge and open spiral arms, which contain numerous HII regions and young star clusters, it is classified as an Sc galaxy within the Hubble classification scheme. Its core contains a bright, variable, x-ray source, which may be powered by a central black hole.

See also: Andromeda Galaxy, black hole, H II regions, Hubble classification, Local Group, Milky Way Galaxy, spiral arms, spiral galaxy, x-ray astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trieste Astronomical Observatory

Founded in 1866 as the Austrian K. u. K. Maritimes Observatorium, in 1898 the Trieste Astronomical Observatory became an independent institute for research in astronomy. Now it is part of INAF (Istituto Nazionale di Astrofisica) and the Italian council for research in astrophysics. Staff include 20 researchers and 42 technicians and administrators. Research fields are extragalactic, stellar physics, solar radioastronomy (10 m antenna), solar system astrophysics, astrophysical technologies for ground-based and space astronomy, international cooperation projects. The facilities include an astronomical library, museum and public observatory.

For further information see
http://www.oat.ts.astro.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trifid Nebula (M20, NGC 6514)

An emission nebula in the constellation Sagittarius, position RA 18^{h} 02.6^{m}, dec. $-23^{\circ}\,02'$. It is of ninth magnitude, with quite a high surface brightness, and measures $29'$ by $27'$. It surrounds the multiple star HN 40, the light of whose brightest members energize the nebula. The Trifid gets its name from dark lanes that trisect it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Triton

Triton is the largest satellite of NEPTUNE. It was discovered in 1846, a few weeks after Neptune. Its orbital motion was soon established, but its physical and chemical properties remained basically unknown until the Neptune encounter of the VOYAGER 2 spacecraft in August 1989.

The orbit of Triton, which is nearly circular, is retrograde and shows a high inclination (157°) which suggests that this satellite was captured by Neptune's gravity field, rather than formed in the Neptunian proto-subnebula. The high inclination of Triton's orbit, combined with Neptune's obliquity, leads to large and complex periodic variations of the latitude of the subsolar point, which ranges between +52° and −52°. These large variations induce very large seasonal effects, probably the strongest ever observed on a solar system object. As in the case of the Moon and the Galilean satellites, Triton's orbit is synchronous, so that the satellite always turns the same hemisphere towards Neptune.

Prior to Voyager observations, little was known about Triton's chemical composition. Ground-based spectroscopy in the infrared range had detected the presence of methane and nitrogen ices which suggested the presence of a nitrogen-rich atmosphere, but the surface pressure of this atmosphere was completely unknown.

Voyager observations of Triton, in August 1989 have provided us with a determination of its diameter (2700 km), its density (2.08 g cm^{-2}), its surface temperature (38 K), its surface pressure (14 μbar) and its atmospheric chemical composition (molecular nitrogen with traces of methane at the 0.01% level). The presence of a stable atmosphere is unusual on an object as small as Triton; this stability is due to the very low temperature of Triton's surface which inhibits gravitational escape.

Voyager images also revealed the extraordinary aspect of Triton's surface. The very high albedo of Triton (0.7 in the visible range) puts the satellite among the most reflective of solar system objects. The Voyager images mostly covered the southern hemisphere. They showed a bright polar cap extending widely around the south pole. North of this polar cap, a rugged terrain, reminiscent of the skin of a cantaloupe melon, is cross-cut by a pattern of intercepting ridges, possibly the remnants of a past TECTONIC event. Triton's surface shows relatively few impact CRATERS, which implies that it is relatively young. A more cratered region, near the equator, shows an impact density comparable to the 3-billion year-old lunar maria. The chemical composition of Triton's surface was identified from ground-based infrared spectroscopy after the Voyager flyby. Ices of N_2, H_2O, CO_2, CO and CH_4 were detected. The abundance of gaseous nitrogen, and thus the surface pressure of Triton, is consistent with equilibrium with the solid phase.

In spite of its very low density, the atmosphere of Triton exhibits several characteristics which are unique in the solar system. The temperature profile consists of a troposphere up to an altitude of 8 km where the temperature is minimum. Above the tropopause, the thermosphere starts immediately, without the presence of an intermediate radiative stratosphere, which is usually found in PLANETARY ATMOSPHERES. In the thermosphere, the temperature increases by conduction up to an altitude of about 400 km where ionization takes place. Several traces of condensates were detected by Voyager in the lower troposphere. They are presumably due to methane and hydrocarbons produced by methane photodissociation; nitrogen clouds were also observed above the southern polar cap which is at present exposed to maximum solar insolation and is sublimating part of its nitrogen ice.

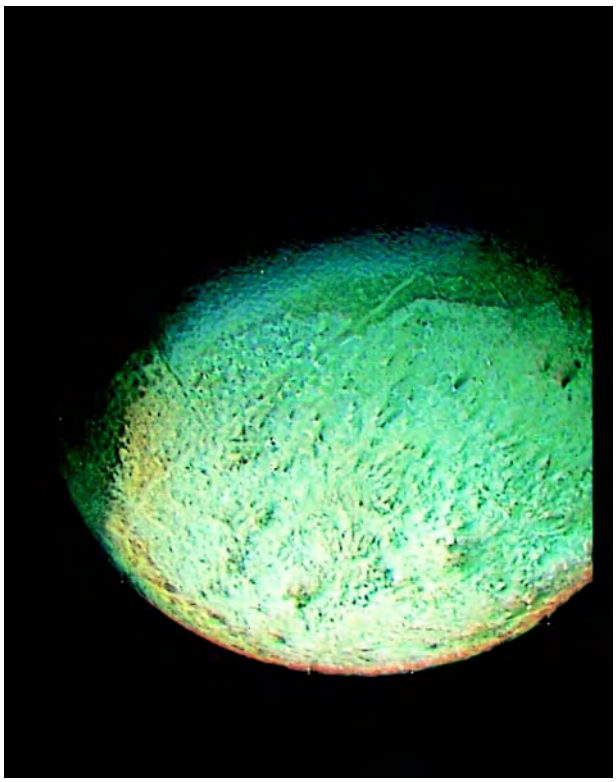


Figure 1. The southern hemisphere of Triton as observed by Voyager 2 in August 1989. The southern polar cap extends widely around the south pole. Dark plumes are detectable over the bright surface of the polar cap.

Seasonal effects are very strong on Triton, due to the high inclination of the satellite. Climatic variations involve the condensation of the two major atmospheric species, nitrogen and methane, which are both involved in the formation of the polar caps. This situation may be compared to the Martian atmosphere, with CO_2 and H_2O acting respectively on Mars like N_2 and CH_4 on Triton (see MARS: ATMOSPHERE).

As in the case of Mars, seasonal effects induce on Triton large-scale dynamical motions. Some of these dynamical motions are illustrated by the detection of active plumes in Voyager images of Triton's surface. These plumes were observed as dark trails, rising vertically

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

up to an altitude of 8 km and then trailing to the west over more than 100 km. Four of these plumes were observed by Voyager in a continuously illuminated region close to the subsolar point. Several interpretations of these phenomena have been suggested, including active volcanism and solar-driven geysers.

Apart from Triton's density, we have little information about its interior. According to models, the internal structure of Triton could consist of a core of metals and rocks, surrounded by a mantle of water ice covered with a variety of ices: CO_2, H_2O, possibly NH_3, CO, CH_4 and N_2, the two latter ices being transported into polar caps along the seasonal cycle.

What could be the origin of Triton? Its striking similarity with Pluto, in terms of size, density, surface chemical composition and atmospheric composition, suggests a common origin for the two bodies (see PLUTO AND CHARON). It is now believed that Pluto is the biggest representative of the newly discovered class of KUIPER BELT objects which populate the outer solar system at heliocentric distances larger than 30 AU. Triton, like Pluto, most probably originated from the Kuiper Belt and was later captured by Neptune's gravity field. Such a capture, however, is only temporary unless sufficient energy is dissipated in the process. Moreover, the probability of such a capture is quite small, so this scenario suggests the presence of a large number of triton-size PLANETESIMALS coming from the kuiper Belt in the early stages of solar system history.

Bibliography

Benner L A M 1997 Triton *Encyclopedia of Planetary Sciences* ed J H Shirley and R W Fairbridge (London: Chapman and Hall)

McKinnon W B and Kirk R L 1999 Triton *Encyclopedia of the Solar System* ed P R Weissman, L McFadden and T V Johnson (New York: Academic)

Thérèse Encrenaz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trojan Asteroids

Two groups of asteroids whose orbital elements are similar to those of Jupiter. One group orbit ahead of Jupiter, clustered around the L_4 Lagrangian point, while the other group orbit behind the planet, around the L_5 Lagrangian point. The L_4 group, which are more numerous, are sometimes called the *Achilles group*, after the first Trojan to be discovered, (588) Achilles. The L_5 group are sometimes called the *Patroclus group*, after the second Trojan discovery, (617) Patroclus. Achilles and Patroclus are both around 150 km in diameter, and were both found in 1906. The largest Trojan is (624) Hektor, at 225 km; the names given to Trojans are those of heroes from Homer's writings on the Trojan Wars. As of April 2000 there were 425 known Trojans in the L_4 group, and 188 in the L_5 group. Two Martian Trojans have also been found.

Perturbations by other planets cause Jupiter's Trojans to oscillate about the Lagrangian points. An individual Trojan will drift back and forth in longitude by 15° to 20° either side of its Lagrangian point with a period of up to 200 years. Trojans have low albedos, in the range 0.02–0.07. About 60% are D-type, with reddish reflectance spectra indicating a carbon-rich surface; most of the others are C-type, with compositions similar to carbonaceous chondrites, and the remainder P-type, with quite red surfaces indicating an organic surface composition.

See also: Lagrangian points.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Troughton, Edward (1753–1836)

Instrument-maker, born in Corney, Cumberland, England, designing instruments to aid navigation, surveying and astronomy. With his brother John, in their family business in London, he made transit circles for Greenwich and other observatories, buying in the optics from, for example, the Dollonds, because the family trait of color-blindness meant that they were unable to test lenses. Troughton invented a method to make accurate scales by dividing a circle.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Trumpler, Robert Julius (1886–1956)

Astronomer, born in Zürich, Switzerland. He worked at the Lick Observatory in California, where he showed that distant clusters of stars are systematically too faint for their size, as if dimmed by an obscuring medium whose effects accumulate with distance. This was the first convincing evidence of the presence of an interstellar medium in our Galaxy, consistent with the indications from HEBER CURTIS' recognition that spiral galaxies seen edge-on had obscuration along their central plane.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tsiolkovsky [Ciolkovskij, Ziolkowski], Konstantin Eduardovich [Eduardowitsch] (1857–1935)

Russian physicist and rocketry pioneer, born in Izhevsk, Russia. Developed the basic theory of rocketry including the idea of liquid propellants and multistage rockets, but, unlike GODDARD, carried out no practical work.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tucana

(the Toucan; abbrev. Tuc, gen. Tucanae; area 295 sq. deg.) a southern constellation which lies between Indus and Hydrus, and culminates at midnight in mid September. It was first shown on a celestial globe by the Dutch theologian and geographer Petrus Plancius in *c*.1598, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, rather inconspicuous constellation, the brightest stars in Tucana are α Tucanae, magnitude 2.9, β Tucanae, a multiple system consisting of a wide pair of bluish-white (B9 and A2) components, magnitudes 4.4 and 4.5 (combined magnitude 3.7), separation 27″, both of which have a fainter companion, the former of magnitude 13.6, separation 2.4″, the latter white (A7), magnitude 6.0, separation 0.1–0.6″, period 44.4 years, and another bluish-white (A2) component 10′ to the southwest, which is also double, with components of magnitudes 5.8 and 6.0 (combined magnitude 5.2), separation 0.1″, all six stars having a similar proper motion, and γ Tucanae, magnitude 4.0. Another interesting star is κ Tucanae, a quadruple system consisting of pale yellow (F6) and orange (K1) components, magnitudes 5.0 and 7.7, separation 5.0″, and a yellow (G5) component, magnitude 7.9, 320″ distant, which has a magnitude 8.4 companion, separation 0.9–1.1″, period 86.2 years.

Other interesting objects in Tucana include the Small Magellanic Cloud, 47 Tucanae (NGC 104), the second most impressive globular cluster in the sky (after ω Centauri), visible to the unaided eye as a fuzzy fourth-magnitude star, and NGC 362, another globular cluster, of seventh magnitude.

See also: Small Magellanic Cloud.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tully–Fisher Relation

Cepheid variable stars are the primary means by which distances are measured over the local volume of space. However, beyond about 20 megaparsecs (Mpc) Cepheids become too faint, even for Hubble Space Telescope, and so alternative means of measuring distances are needed. One of the more popular secondary methods makes use of the strong correlation between the luminosity of SPIRAL GALAXIES and their rotational velocities. This is known as the Tully–Fisher relation.

The Tully–Fisher relation has become one of the most widely used methods of measuring extragalactic distances since spiral galaxies are relatively common and contain young, massive stars. As a result, they also contain Cepheid variables, making the calibration of the Tully–Fisher relation using nearby systems relatively straightforward. In applying this technique the rotational velocity serves as a predictor of the luminosity, or absolute magnitude, of the galaxy. The distance is then calculated from the distance modulus, the difference between the apparent magnitude and the predicted absolute magnitude of the galaxy. The Tully–Fisher relation provides an opportunity to step from the local calibrating galaxies to the smooth Hubble expansion in a single step. Consequently, this technique is extremely valuable for the mapping of large-scale structure, the Hubble flow and any associated peculiar (i.e. non-expansion) velocities.

The origin of the Tully–Fisher relation remains somewhat uncertain but arises from the correlation between mass and luminosity. A spiral galaxy rotates under the influence of its own gravity and so the rotational speed is connected with its mass. A more massive galaxy also contains a larger number of stars and is therefore more luminous. The result is a correlation between the luminosity and the rotational speed for spiral galaxies, the Tully–Fisher relation. There are a few complications to this simple explanation. First, the total luminosity of a particular spiral galaxy is a reflection of its star formation history. This might be expected to vary from galaxy to galaxy. Second, the fact that spiral galaxies have considerable amounts of 'DARK MATTER' means that the mass of a galaxy is not necessarily well represented by the amount of mass contained within its population of stars. That is, the ratio of mass to light for spiral galaxies might be expected to vary at a given total mass. The fact that the Tully–Fisher relations show relatively small scatter indicates that these variations are small and this in turn provides a strong constraint on models for the formation and evolution of spiral galaxies. While these details are still uncertain we can still measure galaxy distances through an empirical calibration of this relation.

Observational data for the Tully–Fisher relation

The necessary observational data consist of apparent magnitudes, corrected for Galactic and internal extinction by dust, and some measure of rotational velocities, corrected for projection effects. Usually, the rotational velocity is measured via the Doppler broadening of the H I 21 cm line, since spiral galaxies are easily detected in this spectral line using RADIO TELESCOPES. However, for galaxies with REDSHIFTS of only a few thousand km s^{-1}, the 21 cm line is shifted into heavily used regions of the radio spectrum, resulting in severe interference from terrestrial sources (radars, cell phones, etc). In addition, at this wavelength the resolution of even the largest single-dish radio telescopes is only a few arcminutes and so for redshifts beyond about 5000 km s^{-1} the detected signal can be confused with that from neighboring galaxies. This has prompted interest in using optical emission lines such as Hα to measure the rotational speed of spiral galaxies.

Early application of the Tully–Fisher relations used photographic photometry, but today CCDs are the detectors of choice at optical wavelengths owing to their superior performance. The redder Johnson bandpasses (R and I) are usually used because of their reduced sensitivity to dust extinction and because the spectral energy distributions of spiral galaxies are dominated by late-type giant stars at these wavelengths. Thus, these measurements are less sensitive to variations in the star formation history and dust content of the individual galaxies. These advantages continue into the near-infrared H and K bands. However, the night sky at these wavelengths is as much as a factor of 10^4 brighter than at optical wavelengths, making accurate surface photometry at faint levels extremely difficult.

A typical procedure for measuring the magnitude of a galaxy from a calibrated CCD image begins by identifying and masking contaminating regions from foreground stars, background galaxies and cosmetic defects. The surface brightness profile of the galaxy is constructed by binning pixels as a function of radius. The integrated magnitude within some surface brightness level can then be computed and this measurement extrapolated to infinite radius using some model, such as an exponential, to produce a total magnitude for the galaxy. Ellipses are often fitted to the isophotes of the galaxy to yield the best-fit axial ratio of the inclined disk as a function of radius and surface brightness. The axial ratio can then be used to estimate the inclination of the galaxy.

There are several advantages to using the Doppler-broadened 21 cm line of H I as the measure of the rotation of late-type galaxies. First, spiral galaxies are H I rich so the line is extremely strong. Second, the H I within these systems has an extended distribution such that the outer, flat portion of the rotation curve is well sampled, providing an accurate measurement of the maximum rotational velocity. However, low angular resolution and interference become significant limitations at redshifts of $\sim$10 000 km s^{-1}, as described above. The use of the Hα emission line avoids these issues. In this case, the slit of a spectrograph is positioned to lie along the long axis of a DISK GALAXY such that the resulting spectrum will sample the velocity as a function of radius for the galaxy. However, since the Hα emission line is produced within regions of active STAR FORMATION, the outer regions of spiral

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

galaxies may not have enough Hα emission to be detected. This can result in the outer, flat portion of the rotation curve of lower-luminosity galaxies not being adequately sampled since these systems generally have slowly rising rotation curves. Thus, 21 cm measurements are usually the best choice for nearby galaxies and Hα measurements are optimal for redshifts beyond about 10 000 km s^{-1}. For redshifts beyond about 0.2 (60 000 km s^{-1}) the Hα line becomes contaminated by increasingly strong emission lines from the night sky (mostly due to OH). This has prompted investigations into whether other lines such as [O II] (λ 3727 Å) and [O III] (λ 5007 Å) could be used such that the Tully–Fisher relation could be applied at larger distances. The preliminary results appear to be very promising and efforts are underway to use the Tully–Fisher relations to examine the structural and evolutionary properties of GALAXIES AT HIGH REDSHIFTS.

Inclination corrections to the observational data

Some method of measuring galaxy inclinations is required in order to correct the magnitudes for internal extinction and the observed rotation for projection. Assuming that all spiral galaxies can be adequately represented as an oblate spheroid with the same intrinsic flattening, the inclination can be calculated from the projected shape of the disk. Specifically,

$$\cos^2 i = \frac{(b/a)^2 - \alpha^2}{1 - \alpha^2}$$

where i is the inclination, b/a is the observed axial ratio of the best-fitting ellipse and α is the intrinsic axial ratio for an edge-on system. Complications arise from the fact that some galaxies do not have perfectly circular isophotes at face-on orientation. As a result, most investigations restrict the sample to those systems whose inclinations are at least 30°–45° in order to minimize the resulting uncertainties on the deprojected rotational velocities.

The parameter of choice for predicting the luminosity of a disk galaxy is $V_{\rm max}$, the maximum amplitude of the rotation curve. This is especially convenient given that most galaxies have flat rotation curves over much of their extent. Tully and Fisher advocated using the Doppler width of the 21 cm line profile measured at 20% of peak intensity (W_{20}), corrected for inclination. That is, $W^{\rm i}_{\rm R} = W_{20}/\sin(i) \sim 2V_{\rm max}$. However, if the Tully–Fisher relation is to be applied to low-luminosity spiral and irregular galaxies a more complicated correction is required in order to account for the effects of internal turbulence and subtle variations in line-profile shape. However, for $W^{\rm i}_{\rm R} >$ 150 km s^{-1} these corrections are negligible. When optical spectra are used $V_{\rm max}$ can be measured directly from the rotation curve.

Historically, the internal extinction with disk galaxies has proven to be a difficult problem. However, the advent of CCDs and infrared detectors has allowed accurate photometry of galaxies over a broad range in wavelength enabling the amount of extinction to be statistically estimated from the amount of reddening. The extinction corrections to the magnitudes are often expressed in terms of the observed axial ratio. Specifically, $A^{\rm i}_{\lambda} = \gamma_{\lambda} \log(a/b)$, where λ is the wavelength of the observation in the standard bandpasses and a/b is the observed axial ratio. The dependence of this correction on the luminosity of the galaxy, due to variations in dust content, etc, can also be described by expressing γ_{λ} in terms of the projection-corrected rotational velocity (W^i_R). A representative result is

$$\gamma_B = 1.57 + 2.75(\log W^{\rm i}_{\rm R} - 2.5)$$

$$\gamma_R = 1.15 + 1.88(\log W^{\rm i}_{\rm R} - 2.5)$$

$$\gamma_I = 0.92 + 1.63(\log W^{\rm i}_{\rm R} - 2.5)$$

$$\gamma_H = 0.43 + 0.76(\log W^{\rm i}_{\rm R} - 2.5)$$

where the subscripts on γ represent the standard bandpasses. Here, $W^{\rm i}_{\rm R} \sim 2V_{\rm max}$ and $V_{\rm max}$ is the amplitude of maximum rotation within a galaxy.

The apparent magnitudes also require correction for Galactic extinction. In the past, this has been a difficult problem. However, there now exist extensive maps of both the neutral hydrogen gas (H I) and the distribution of dust (IRAS 100 μm) within the Galaxy. The good correspondence between these two maps indicates that the H I gas and dust trace each other. The amounts of the extinction which correspond to these two components have been calibrated from detailed, multicolor observations of distant stars. As a result, the Galactic extinction for any galaxy can be estimated given its position on these maps. As the Galactic latitude falls below about 25–30°, the corrections for Galactic extinction become large with correspondingly large uncertainties. Therefore, most investigators limit the application of the Tully–Fisher relation to galaxies above this latitude.

Absolute calibration of the Tully–Fisher relation using Cepheids

Ground-based Cepheid surveys are extremely difficult because the Cepheids in all but the nearby Local Group galaxies are very faint and subject to significant contamination by neighboring stars. However, the advent of Hubble Space Telescope has allowed the detection of Cepheids within galaxies out to about 20 Mpc in distance and enabled the number of calibrating systems to be significantly increased. A large effort to obtain Cepheid distances to a number of nearby galaxies, known as the Hubble Key Project of Extragalactic Distances, and a similar project to obtain Cepheid distances to galaxies containing type Ia supernovae, has brought the number of galaxies suitable for calibrating the Tully–Fisher relation to 30 systems. Figure 1 shows the resulting Tully–Fisher relations for this calibration sample. A least-squares fit to these data gives the following relations:

$$M^{\rm b,i}_B = -7.41(\log W^{\rm i}_{\rm R} - 2.5) - 20.04 \pm 0.04 \quad (0.22)$$

$$M^{\rm b,i}_R = -8.09(\log W^{\rm i}_{\rm R} - 2.5) - 21.05 \pm 0.04 \quad (0.19)$$

$$M^{\rm b,i}_I = -8.55(\log W^{\rm i}_{\rm R} - 2.5) - 21.51 \pm 0.04 \quad (0.22)$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

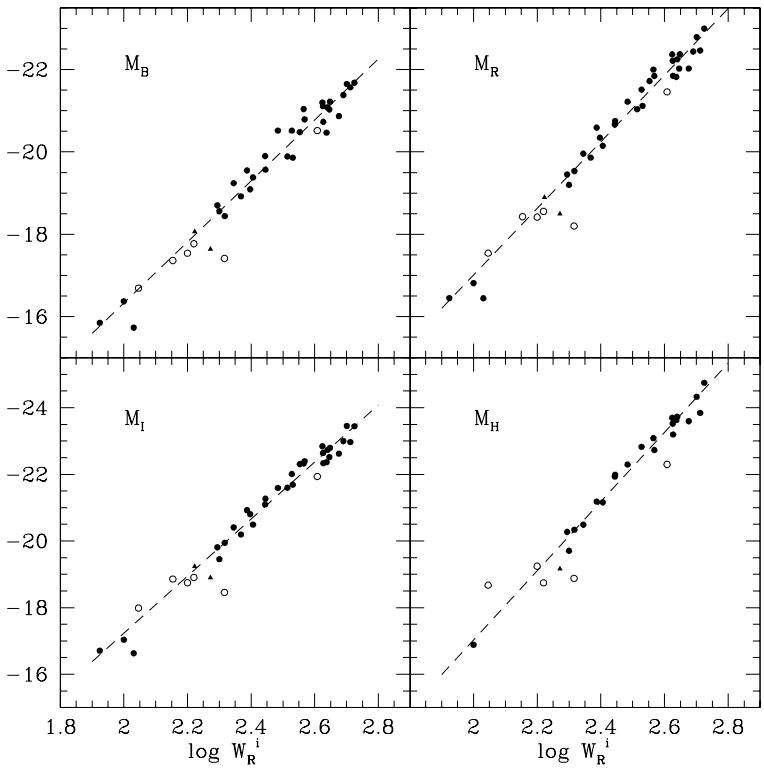


Figure 1. B, R, I and H band calibrations of the Tully–Fisher relation. Solid circles are galaxies with distances measured using Cepheids, solid triangles are galaxies with distances estimated via surface brightness fluctuation measurements within dE companions and open circles are systems thought to be group members with at least one galaxy with a Cepheid distance, and therefore thought to be at a similar distance. The dashed line is a least-squares fit to the solid points in each panel.

$$M_H^{\mathrm{b,i}} = -10.39(\log W_{\mathrm{R}}^{\mathrm{i}} - 2.5) - 22.22 \pm 0.08 \quad (0.30)$$

where the superscripts b and i indicate Galactic extinction and internal extinction corrections, respectively. The zero points give the absolute magnitude for galaxies with $\log W_{\mathrm{R}}^{\mathrm{i}} = 2.5$, or $V_{\max} = 160$ km s^{-1}. In all bandpasses, the random uncertainty in the zero point is negligible compared with the remaining systematic uncertainty in the absolute calibration of the Cepheids themselves (∼0.13 mag, or 7% in distance). The rms dispersions of the relations are given in parentheses. Given these calibrations, the absolute magnitude of any spiral galaxy can be predicted from its rotational velocity and its distance estimated from its measured apparent magnitude.

Application of the Tully–Fisher relation to field galaxies

The Tully–Fisher relations have been used to estimate the distances to several hundred spiral galaxies within an expansion velocity of about 8000 km s^{-1}. An example of the resulting 'Hubble flow' is shown in figure 2. Since the velocities are a combination of both the expansion and any local peculiar velocities induced by gravity, the deviations in velocity can be correlated with position on the sky to produce maps of the deviations from the mean expansion field. The primary features seen in such maps are (1) local streaming of ∼100 km s^{-1} associated with the non-uniform distribution of matter within 10 Mpc of the Local Group, (2) retarded expansion velocities of several hundred km s^{-1} near the Virgo cluster of galaxies and (3) a large-scale streaming of about 400 km s^{-1} towards a 'Great Attractor' located at a distance of about 60 Mpc in the direction of the Hydra–Centaurus cluster complex. These peculiar velocities are usually assumed to result from the gravitational accelerations produced by the irregularities in the large-scale distribution of mass. With this assumption, various research groups have used these 'flow maps' to constrain the mean mass density of the universe (Ω) as well as to examine the degree to which the distribution of galaxies follows the distribution of mass. These investigations appear to favor low values of Ω (∼0.2) and also imply that mass is generally traced by the distribution of galaxies.

Distant clusters and the Hubble constant

The errors in the distances of individual galaxies measured using the Tully–Fisher relations are about 15% when using high-quality observational data. As a result, the absolute errors can become large for distant galaxies. Since there are still significant peculiar velocities out to velocities of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

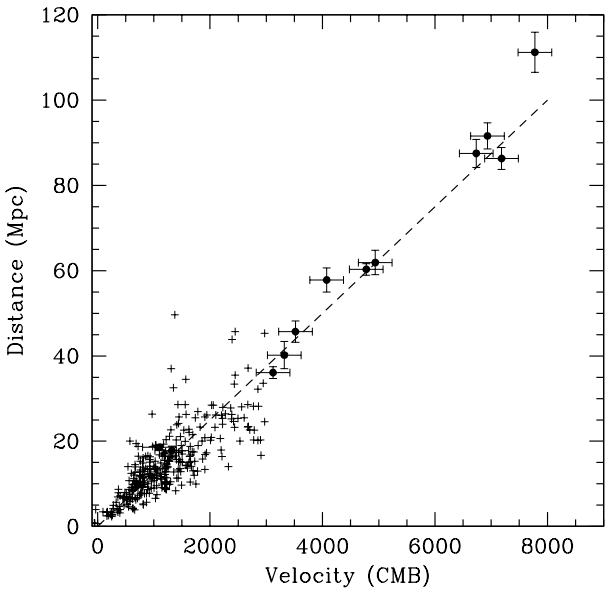


Figure 2. The velocity–distance relation, or 'Hubble flow', illustrating the expansion of the universe. The plus signs show the distances to individual, nearby galaxies measured using the Tully–Fisher relation. Since the errors are fractional (i.e. logarithmic), the absolute errors increase with distance and result in a cone-like spread of the data. The solid points with error bars represent the average distance and velocity for clusters. The cluster samples typically contain 15–20 galaxies, correspondingly reducing the errors. The dashed line shows a Hubble constant ($H_0 = V/D$) of 80 km s^{-1} Mpc^{-1}.

about 5000 km s^{-1}, large samples are needed at these distances in order to accurately sample the Hubble flow and map the distribution of peculiar velocities. Recent Tully–Fisher surveys using cluster samples have been used to measure the mean distances for a number of clusters over this volume. The mean distances and velocities of these clusters are shown in figure 2 along with those of the nearby field galaxies. Together, these data yield an estimated value of the HUBBLE CONSTANT of 77 ± 7 km s^{-1} Mpc^{-1}.

The inverse of the Hubble constant is related to the age of the universe. For a universe with zero density H^{-1} is equal to the age of the universe. If the universe has the 'critical density', i.e. $\Omega = 1$, then its age is $(2/3)H^{-1}$. At the present time, a typical estimate for the age of globular cluster stars is 14 ± 2 billion years. For comparison, the value of the Hubble constant given above corresponds to ages of 12.7 ± 1 and 8.5 ± 0.7 billion years for $\Omega = 0$ and $\Omega = 1.0$, respectively. At present, it appears that a cosmologically flat, critical universe is inconsistent with the observational data. This discrepancy has helped to generate interest in cosmological models with a non-zero COSMOLOGICAL CONSTANT Λ and, in particular, in cosmologically flat models with non-zero Λ. A cosmological constant is one way of reconciling these disparate estimates of the age of the universe.

Bibliography

The literature regarding the Tully–Fisher relation and its application to measuring extragalactic distances is extensive. A few representative papers are presented here.

The first description of what is now called the Tully–Fisher relation was originally given by

Balkowski C *et al* 1974 *Astron. Astrophys.* **34** 43

However, Tully and Fisher were the first to advocate its use for measuring extragalactic distances in

Tully R B and Fisher J R 1977 *Astron. Astrophys.* **54** 661

Application of the Tully–Fisher relation using CCD detectors can be found in

Pierce M J and Tully R B 1988 *Astrophys. J.* **330** 579

Application of the Tully–Fisher relation using infrared detectors can be found in

Aaronson M and Mould J R 1983 *Astrophys. J.* **265** 1
Peletier R and Willner S 1993 *Astrophys. J.* **418** 626

The absolute calibration of the Tully–Fisher relation is discussed in

Freedman W L 1990 *Astrophys. J.* **355** L35
Pierce M J and Tully R B 1992 *Astrophys. J.* **387** 47

An example of the application of the Tully–Fisher relation to the measurement of the distances to clusters of galaxies can be found in

Giovanelli R *et al* 1998 *Astron. J.* **116** 2632

Examples of the application of the Tully–Fisher relation to the measurement of the local velocity field can be found in

Aaronson M *et al* 1982 *Astrophys. J.* **258** 64
Giovanelli R *et al* 1998 *Astrophys. J.* **502** L91
Tully R B, Shaya E J and Pierce M J 1992 *Astrophys. J. Suppl.* **80** 479

The technical details of correcting the magnitudes and rotational velocities for inclination and projection corrections can be found in

Giovanelli R *et al* 1997 *Astron. J.* **113** 53
Tully R B and Fouqué P 1985 *Astrophys. J. Suppl.* **58** 661
Tully R B *et al* 1998 *Astron. J.* **115** 2264

A comprehensive review of various methods of measuring extragalactic distances, including the Tully–Fisher relation, can be found in

Jacoby G H *et al* 1992 *Publ. Astron. Soc. Pac.* **104** 599

The use of optical rotation curves as traced by Hα and [O II] emission to measure $V_{\max}$ is described by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Courteau S 1997 *Astron. J.* **114** 2402
Vogt N P *et al* 1996 *Astrophys. J.* **465** L15

An extensive compilation of H I 21 cm line widths can be found in

Huchtmeier W K and Richter O-G 1989 *A General Catalogue of HI Observations of Galaxies* (New York: Springer)

Technical issues regarding sample selection biases and their effect on distances measured via the Tully–Fisher relation are discussed in

Tully R B 1988 *Nature* **334** 209

Michael Pierce

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tunable Imaging Filters

While tunable filters are a recent development in night-time astronomy, they have long been used in other physical sciences, e.g. solar physics, remote sensing and underwater communications. With their ability to tune precisely to a given wavelength using a bandpass optimized for the experiment, tunable filters are already producing some of the deepest narrowband images to date of astrophysical sources. Furthermore, some classes of tunable filters can be used in fast telescope beams and therefore allow for narrowband imaging over angular fields of more than 1° over the sky.

The physics of tunable imaging

A rich variety of physical phenomena can isolate a finite spectral band: absorption, scattering, diffraction, evanescence, birefringence, acousto-optics, single-layer and multilayer interference, multi-path interferometry, polarizability and so on. Most of the available optical filter technologies are listed below (see also FILTERS):

linear or circular variable filter
multilayer dielectric filter
Fabry–Perot interferometer
Michelson interferometer
acousto-optic filter
solid etalon filter
solid Michelson filter
generalized resonant grating filter
sub-lambda evanescent grating filter
volume phase holographic grating filter
Lyot–Öhman filter
generalized Lyot filter
generalized Šolc[1] filter
liquid crystal filter

There is a bewildering array of future possibilities, including tunable crystal lattice structures, and many possible variants on the above technologies. Here, we focus on a few key technologies.

The ideal device

The ideal filter is an imaging device which can isolate an arbitrary spectral band $\delta\lambda$ at an arbitrary wavelength λ over a broad, continuous spectral range, preferably with a response function which is identical in form at all wavelengths.

The tilting interference filter is much the worst form of tunable filter. The spectral range covered is almost negligible, and the filter profile varies with tilt angle. Some devices (e.g. Lyot, linear variable filter) work only at a fixed resolving power but have the ability to tune over a wide spectral window. Better devices allow for a wide selection of resolving powers over a wide range of wavelengths.

The different techniques rely ultimately on the INTERFERENCE of beams that traverse different optical paths to form a signal. The technologies which come closest to the ideal tunable filter are the air-gap Fabry–Perot and Michelson (Fourier transform) interferometers. To understand why, we highlight the Taurus tunable filter (TTF) which was the first general-purpose such device for night-time astronomy (see http://www.aao.gov.au/local/www/jbh/ttf). This is a Fabry–Perot filter where interference is formed between two highly reflective, moving plates. To be a useful filter, not only must the plates move through a large physical range, but they must start at separations of only a few wavelengths, as we show.

The condition for photons with wavelength λ to pass through the filter is (see figure 1)

$$m\lambda = 2\mu\ell\,\cos\theta \tag{1}$$

from which it follows that

$$\frac{\mathrm{d}\mathcal{R}}{\mathcal{R}} = \frac{\mathrm{d}m}{m} = \frac{\mathrm{d}\ell}{\ell}. \tag{2}$$

For an order of interference m, the resolving power is $R = m\mathcal{N}$ where $\mathcal{N}$ is the instrumental finesse. The finesse is determined by the coating reflectivity and is essentially the number of recombining beams. For the TTF, the plates can be scanned over the range $\ell = 1.5$–$15\ \mu$m, and the orders of interference span the range $m = 4$–40, such that the available resolving powers are $R \approx 100$–1000.

The sharp core of almost all tunable-filter transmission profiles is not ideal. Even a small amount of flatness at peak transmission can ensure that we avoid narrowly missing most of the spectral line signal from a source. In theory, all band-limited functions can be squared off, but in practice this is difficult for all but two devices. Since the Michelson interferometer (filter) obtains its data in the frequency domain, the profile can be partially squared off at the data reduction stage through a suitable choice of convolving function. For the Šolc filter, the instrumental profile can be modified through the use of partial polarizers and birefringent retarder elements.

Summary of filter technologies

Here, we provide a brief outline of some of the key technologies.

Monolithic filters

Interference filter. The principle relies on a dielectric spacer sandwiched between two transmitting layers (single cavity). The substrates are commonly fused silica in the ultraviolet, glass or quartz in the optical and water-free silica in the infrared. Between the spacer and the glass, surface coatings are deposited by evaporation which partly transmit and reflect an incident ray. Each internally reflected ray shares a fixed phase relationship to all the other internally reflected rays. For constructive interference, for a wavelength λ to be transmitted, it must satisfy equation (1) where $\theta = \theta_{\mathrm{R}}$ is the refracted angle within the optical spacer, and the optical gap is the product

[1] The proper Czech pronunciation is 'Sholtz'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

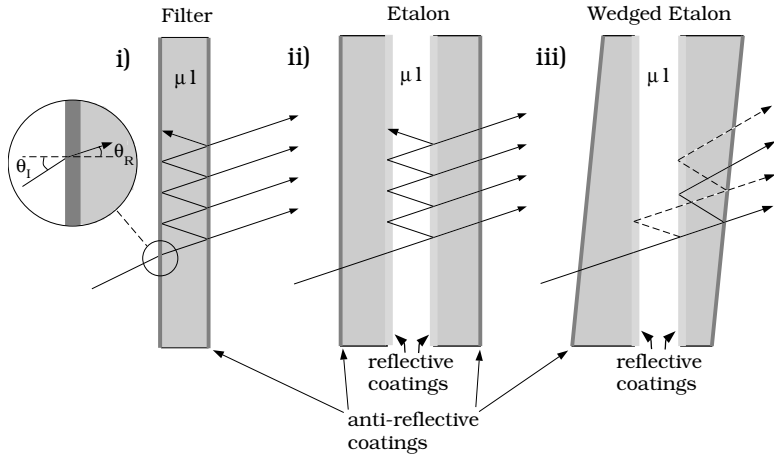


Figure 1. (i) Interference filter: the internal structure is not shown. (ii) Fabry–Perot etalon. (iii) Wedged Fabry–Perot etalon which avoids the problem of the plates behaving like interference filters.

of the thickness ℓ and refractive index μ of the spacer. The construction of these filters has undergone a revolution through the use of dielectric, multilayer thin-film coatings, and a proper description is more involved. All such filters can be tuned through a small wavelength interval ($\delta\lambda/\lambda = -\theta_R^2/2\mu^2$), which amounts to no more than 2%λ in practice. Suffice it to say, interference filters make for poor tunable devices.

Gap-scanning filters

Fabry–Perot filter. The air-gap etalon, or Fabry–Perot filter, was introduced in the previous section. The etalon comprises two plates of glass kept parallel over a small separation where the inner surfaces are mirrors coated with high reflectivity $\Re$. The transmission of the etalon to a monochromatic source λ is given by the Airy function

$$\mathcal{A} = \left[1 + \frac{4\Re}{(1-\Re)^2}\sin^2(2\pi\mu\ell\cos\theta/\lambda)\right]^{-1} \qquad (3)$$

where θ is the off-axis angle of the incoming ray and $\mu\ell$ is the optical gap. The condition for peaks in transmission is given in equation (1). Note that λ can be scanned physically in a given order by changing θ (tilt scanning), μ (pressure scanning) or ℓ (gap scanning). Both tilt and pressure scanning suffer from serious drawbacks which limit their dynamic range. With the advent of servo-controlled, capacitance micrometry, the performance of gap scanning etalons surpasses other techniques. These employ piezoelectric transducers that undergo dimensional changes in an applied electric field or develop an electric field when strained mechanically. Queensgate Instruments Ltd have shown that it is possible to maintain plate parallelism to an accuracy of $\lambda/200$ while continuously scanning over several adjacent orders.

Fabry–Perot filters have been made with 15 cm apertures and physical scan ranges up to 3 cm. The etalon is ultimately limited by the finite coating thickness between the mirrors, so it really only achieves the lowest interference orders ($m < 5$) at infrared wavelengths.

Solid etalon filter. These are single-cavity Fabry–Perot devices with a transparent piezoelectric spacer, e.g. lithium niobate. The thickness and, to a lesser extent, refractive index can be modified by a voltage applied to both faces. For low-voltage systems, tilt and temperature can be used to fine tune the bandpass. High-quality spacers with thicknesses less than a few hundred microns are difficult to manufacture, so that etalon filters are normally operated at high orders of interference. The largest devices we have seen are 5 cm in clear aperture.

Michelson filter. In the Fourier transform or Michelson filter, the collimated beam is split into two paths at the front surface of the beam-splitter. The separate beams then undergo different path lengths by reflections off separate mirrors before being imaged by the camera lens at the detector. The device shown in figure 2 uses only 50% of the available light. As Maillard has demonstrated at the CANADA–FRANCE–HAWAII TELESCOPE, it is possible to recover this light but the layout is more involved.

The output signal is a function of path difference between the mirrors. At zero path difference (or arm displacement), the waves for all frequencies interact coherently. As the movable mirror is scanned, each input wavelength generates a series of transmission maxima. Commercially available devices usually allow the mirror to be scanned continuously at constant speed or to be stepped at equal increments. At a sufficiently large arm displacement, the beams lose their mutual coherence.

The filter is scanned from zero path length ($x = y = 0$) to a maximum path length $y = L$ set by twice the maximum mirror spacing ($x = L/2$). The superposition of two coherent beams with amplitudes b_1 and b_2 in complex notation is $b_1 + b_2 e^{i2\pi\nu y}$ where y is the total path difference

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

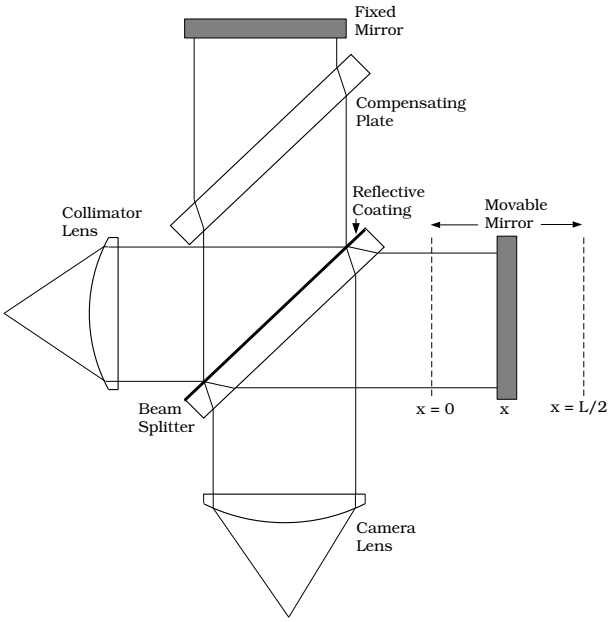


Figure 2. Schematic of a two-beam Michelson (Fourier transform) interferometer.

and ν is the wavenumber. If the light rays have the same intensity, the combined intensity is $4b^2 \cos^2 \pi \nu y$, where $b = b_1 = b_2$. The combined beams generate a series of intensity fringes at the detector. If it was possible to scan over an infinite mirror spacing at infinitesimally small spacings of the mirror, the superposition would be represented by an ideal Fourier transform pair, such that

$$b(y) = \int_{-\infty}^{\infty} B(\nu)(1 + \cos 2\pi \nu y)\, \mathrm{d}\nu \tag{4}$$

$$B(\nu) = \int_{-\infty}^{\infty} b(y)(1 + \cos 2\pi \nu y)\, \mathrm{d}y \tag{5}$$

where $b(y)$ is the output signal as a function of path length y and $B(\nu)$ is the spectrum we wish to determine. $B(\nu)$ and $b(y)$ are both undefined for $\nu < 0$ and $y < 0$: we include the negative limits for convenience. Note that

$$b(y) - \tfrac{1}{2}b(0) = \int_{-\infty}^{\infty} B(\nu) \cos 2\pi \nu y\, \mathrm{d}\nu \tag{6}$$

$$B(\nu) = \int_{-\infty}^{\infty} [b(y) - \tfrac{1}{2}b(0)] \cos 2\pi \nu y\, \mathrm{d}y \tag{7}$$

The quantity $b(y) - \frac{1}{2}b(0)$ is usually referred to as the interferogram although this term is sometimes used for $b(y)$. The spectrum $B(\nu)$ is normally computed using widely available fast Fourier transform methods. The construction of a Michelson filter is a major optomechanical challenge. The ideal Fourier transform pair is never realized in practice. However, the Michelson filter probably comes closest to achieving the goal of an ideal tunable filter.

The Michelson does not suffer the coating thickness problems of the Fabry–Perot filter, and therefore reaches the lowest orders even at optical wavelengths.

Grating filters

Resonant grating filter. These novel filters are inspired by the diffractive colors in many insects and constitute dielectric gratings with three-dimensional, sub-micron microstructure. The zeroth-order reflection exhibits a broad to intermediate bandwidth ($\mathcal{R} \sim 20$), is highly polarized and maintains useful efficiency over a $\pm 30^\circ$ tilt (or rotation) range. Gale (1998) presents one grating design that produces a roughly self-similar bandpass from 450 to 850 nm over this tilt range. Grating filters, and their close relatives, evanescent gratings, show great promise but most have yet to leave the drawing board. However, since much of the research is driven by bank note security, we expect rapid progress. Volume phase holographic gratings—in reflection—can produce a highly efficient grating filter through Bragg diffraction.

Acousto-optic filter (AOTF). These are electronically tunable filters that make use of acousto-optic (either collinear, or more usefully, non-collinear) diffraction in an optically anisotropic medium. AOTFs are formed by bonding piezoelectric transducers such as lithium niobate to an anisotropic birefringent medium. The medium has traditionally been a crystal, but polymers have been developed recently with variable and controllable birefringence. When the transducers are excited to 10–250 MHz (radio) frequencies, the ultrasonic waves vibrate the crystal lattice to form a moving phase pattern that acts as a diffraction grating. The diffraction angle (and therefore wavelength) can be tuned by changing the radio frequency. These devices are often water cooled to assist the thermal dissipation, although this is less important in the UV where AOTFs are particularly useful. The largest devices are 2.5 cm in diameter since it proves to be difficult to maintain a uniform acoustic standing wave over larger areas. An additional problem is the 15 μm structure in the $LiNO_2$ crystal which is not always optimal for good image quality. However, the acceptance angle of the AOTF is generally larger than that of the Fabry–Perot.

Birefringent filters

The underlying principle of the birefringent filter is that light originating in a single polarization state can be made to interfere with itself. An optically anisotropic, birefringent medium can be used to produce a relative delay between ordinary and extraordinary rays aligned along the fast and slow axes of the crystal. (A birefringent medium has two different refractive indices, depending on the plane of light propagation through the medium.) Title and collaborators have discussed at length the relative merits of different types of birefringent filters. The filters are characterized by a series of perfect polarizers (Lyot filter), partial polarizers, or only an entrance and an exit polarizer (Šolc filter). The highly anisotropic

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

off-axis behavior of uniaxial crystals give birefringent filters a major advantage. Their solid acceptance angle is 1–2 orders of magnitude larger than is possible with interference filters although this is partly offset by half the light being lost at the entrance polarizer.

Lyot filter. This is conceptually the easiest to understand and forms the basis for many variants. The entrance polarizer is oriented 45° to the fast and slow axes so that the linearly polarized, ordinary and extraordinary rays have equal intensity. The time delay through a crystal of thickness d of one ray with respect to the other is simply $d\Delta\mu/c$ where $\Delta\mu$ is the difference in refractive index between the fast and slow axes. The combined beam emerging from the exit polarizer shows intensity variations described by $I^2\cos(2\pi d\Delta\mu/\lambda)$ where I is the wave amplitude. As originally illustrated by LYOT (see figure 3), we can isolate an arbitrarily narrow spectral bandpass by placing a number of birefringent crystals in sequence where each element is half the thickness of the preceding crystal. This also requires the use of a polarizer between each crystal so that the exit polarizer for any element serves as the entrance polarizer for the next. The resolution of the instrument is dictated by the thickness of the thinnest element. With quarter-wave plates placed between each of the retarder elements, λ can be tuned over a wide spectral range by rotating the crystal elements. However, to retain the transmissions in phase requires that each crystal element be rotated about the optical axis by half the angle of the preceding thicker crystal.

Woodgate (NASA Goddard Space Flight Center) has made a Lyot filter utilizing eight quartz retarders with a 10 cm entrance window. The retarders, each of which is sandwiched with half-wave and quarter-wave plates in addition to the polarizers, are rotated independently with stepping motors under computer control. They achieve a bandpass of 0.4–0.8 nm tunable over half the optical wavelength range (350–700 nm).

Šolc filter. These highly non-intuitive filters use only two polarizers and a chain of identical retarders with varying position angles (Evans 1958). There are folded (zigzag) and fanned designs with the former having the better performance. Title has made a tunable Šolc filter with 7 cm clear aperture. It has the extraordinary capability of tuning the spectral profile: an n-element Šolc filter can have a profile that is determined by n Fourier coefficients. The same can be achieved with polarizing filters by proper choice of crystal lengths.

Liquid crystal filter (LCTF). These are rapid switching, electronically tuned devices which employ either ferroelectric or nematic liquid crystals. The more commonly used nematic LCTF (Morris *et al* 1994) comprises a series of liquid crystal elements whose thicknesses are cascaded in the same way as the Lyot filter. However, the tuning is achieved by electronically rotating the crystal axes of the liquid crystal waveplate. When no voltage is applied, the retardance is at a maximum; at large applied voltages, the retardance reaches a minimum. The retardance can be tuned continuously to allow the wavelength to be tuned.

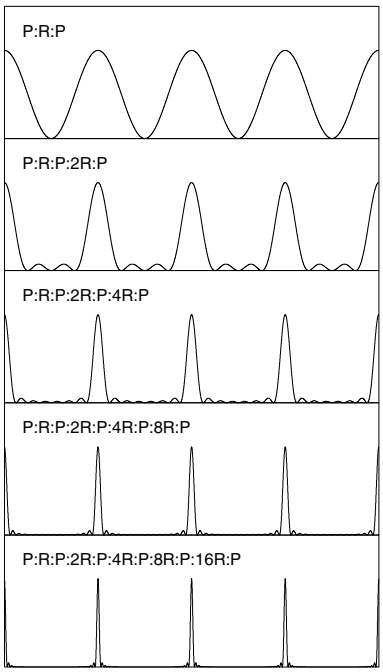


Figure 3. The transmission profile of a simple Lyot cascade with (top to bottom) 1, 2, 3, 4 and 5 stages. The thinnest element has retardance R and the following elements have thicknesses which are multiples of this element. The polaroids P are aligned with each other and oriented at 45° to the retarders.

Liquid crystal filters are now commercially available from Cambridge Research & Instrumentation, Inc. The biggest device we have seen has a clear aperture of 4 cm, requires about 5 V to scan a single order of interference and appears to have good image quality. The tunable band for a single-stage device is about $\mathcal{R} \sim 5$ but can be tuned over the optical window. The peak transmissions are 30% or less.

Differential imaging

There are many important reasons for pursuing tunable filter imaging, as demonstrated by the TTF since the mid-1990s at the ANGLO-AUSTRALIAN 3.9 m TELESCOPE and the William Herschel 4.2 m telescope (see ISAAC NEWTON GROUP OF TELESCOPES). Conventional imaging has a major limitation in that images are taken sequentially which is not ideal at even the best sites. Even small detector, instrument or atmospheric variations lead to systematic error between images. A better approach is a multiband camera so that different bands are observed in parallel. However, even this is not ideal in that the optical path is different for each filter. A tunable filter provides a powerful alternative since band switching can be linked to charge shuffling with the CCD. This is a truly differential

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

technique which leads to much smaller systematic errors than is possible with conventional imaging.

Wide-field imaging

Few telescopes offer more than a wide-field imager at prime focus simply because it is very difficult to exploit the fast beam spectroscopically (see WIDEFIELD CCD IMAGERS). Spectral passbands are degraded in converging beams which is unfortunate as the widest fields are achieved at the fastest f ratios. The wide-field expanded Lyot filter is (almost) the last word in exploiting the widest possible field with a given telescope. Remarkably, beams as fast as $f/2$ can be compensated with crossed birefringent elements, in concert with half-wave plates, such that even a constant sub-ångström bandpass is possible over a degree-sized field. This opens up many new astronomical programs, e.g. Lyot filters can be vastly more efficient for GALAXY REDSHIFT SURVEYS of galaxies (e.g. high-redshift clusters) compared with existing multi-aperture SPECTROGRAPHS. To our knowledge, a wide-angle Šolc filter using half-wave plates has not been attempted although it is entirely feasible.

Bibliography

Bell R J 1972 *Introductory Fourier Transform Spectroscopy* (San Diego, CA: Academic)

Evans J W 1958 The Šolc birefringent filter *J. Opt. Soc. Am.* **48** 142

Gale M T 1998 *Optical Document Security* ed R L van Renesse (Boston, MA: Artech House Publishers) ch 12

Morris H R, Hoyt C C and Treado P J 1994 Imaging spectrometers for fluorescent and Raman microscopy: acousto-optic and liquid crystal tunable filters *Appl. Spectrosc.* **48** 857–66

Title A M and Rosenberg W J 1981 Tunable birefringent filters *Opt. Eng.* **20** 815–23

Vaughan J M 1989 *The Fabry–Perot Interferometer: History, Theory, Practice and Applications* (Bristol: Hilger)

Jonathan Bland-Hawthorn

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tunguska Event

An explosion in mid-air on the morning of 30 June 1908 over a remote forest region of eastern Siberia near the Podkamennaya Tunguska river that caused widespread devastation. The few eyewitnesses reported a fireball as bright as the Sun. The explosion itself was heard up to 1000 km away, and sent pressure waves twice around the planet that were recorded by seismographic stations. The forest canopy burned for over two days. Trees and reindeer in the central 1000 km^2 area beneath the blast were incinerated, and up to 50 km away trees were felled by the shock wave. Dust from the explosion blanketed the area and was carried round the world by atmospheric currents, causing colorful sunsets. It was 20 years before the first scientific research team, led by Leonid Kulik, reached the area. They found no main impact crater or meteoritic fragments; later expeditions, however, recovered dust particles preserved in tree resin and found them to be consistent in composition with stony meteorites. It seems, then, that a stony meteorite with a diameter of about 50 m fragmented and vaporized at an altitude of about 8 km.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tuorla Observatory

Founded in 1952, Tuorla Observatory lies at an altitude of 53 m and is located 12 km from Turku in Finland. It is the site of the headquarters of the NORDIC OPTICAL TELESCOPE.

Early research activities included setting up a geodetic triangulation network for measuring distances, optical work, zenith tube studies of polar motion, detections and orbital calculations of minor planets and comets. Today, research includes active galaxies and stellar dynamics.

The main instrument is a 1 m Dall–Kirkham reflector, used for CCD photometry of quasars. Other instruments include a 0.7 m Schmidt telescope, a 0.6 m Ritchey–Chrétien telescope and a 0.39 m zenith telescope. However, the observatory's most important observations are made abroad, especially at the Nordic Optical Telescope.

During the 1974 nationalization of Turku University, the observatory lost its independent status and became part of the Department of Physical Sciences. In 1989 the observatory buildings were enlarged, and in 1991 Tuorla once again became independent under the jurisdiction of the University.

Tuorla also contributes to the VLBI facility at the Metsähovi Radio Research Station of the Technical University of Helsinki.

For further information see
http://oj287.astro.utu.fi/tuorla/new/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Turner, Herbert Hall (1861–1930)

Born in England, became Savilian Professor of Astronomy at Oxford University. Following the suggestion of E Mouchez (Paris) he organized the international scientific project, *Carte du Ciel* to make astrographic charts by photography.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Twilight Glow

The scattered light before sunrise and after sunset that partially illuminates the sky. This relatively steady glow is part of the overall phenomenon called airglow—the light produced and emitted by the Earth's own atmosphere. The spectrum ranges from 100 nm to 22.5 μm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tycho (Crater)

A prominent young lunar crater in the Moon's southern uplands, and the hub of the Moon's most striking ray system. Its diameter is 102 km, and it is centered at 43.4 °S, 11.1 °W; it is named after the astronomer Tycho Brahe. The crater is one of the youngest features of the Moon, perhaps just 100 million years old. Its ray system extends for up to 1500 km in all directions, and is most prominent at full moon, when the fresh ejecta from which it is composed reflect sunlight back toward the Earth, and the rays are seen to extend over much of the Moon's nearside. The crater itself has terraced walls, caused by partial slumping of sections of the inner rim, which rise to 4.5 km, and a pronounced central peak. For over 80 km beyond the rim there is a continuous deposit of ejecta (an ejecta blanket) in all directions. Rock melted by the impact flowed down the outside to solidify in pools, and down the inside of the rim in rivulets toward the crater floor. Material ejected on higher trajectories produced the rays and also many secondary craters.

See also: Moon.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tycho Brahe (1546–1601)

Much of Tycho (Latinized from Tyge, pronounced 'Teeko') Brahe's accomplishment was made possible by the unique circumstances of his birth and upbringing. He was born on 14 December 1546 to Otte Brahe and Beate Bille, and thus into the highest stratum of Danish society, a powerful oligarchy of aristocratic families. Strangely, he was kidnapped by his uncle before he was two years old, and raised as the foster son of Jørgen Brahe and Inger Oxe. The tradition of the Oxe family seems to have influenced Tycho to cultivate learning rather than the aristocratic martial pursuits of his birth brothers.

After attending the University of Copenhagen (1559–1561), Tycho proceeded to the Lutheran universities of Germany, primarily Leipzig (1562–1565) and Rostock (1566–1567), ostensibly pursuing humanistic and legal studies but actually becoming more and more interested in alchemy and astronomy. It was at Rostock in December 1566 that Tycho lost the bridge of his nose in a duel with another Danish noble student. Thereafter he wore a prosthesis of gold and silver blended to a flesh colour. Tycho traveled widely in Germany and Switzerland during the late 1560s, finding congenial intellectual circles in Basel and Augsburg.

Recalled by his ailing father, Tycho returned to Denmark at the end of 1570 and settled at his ancestral home Kundstrup Manor in Skaane (now southern Sweden) and then at nearby Herrevad Abbey. Around 1571, he fell in love with Kirsten Jørgensdatter, with whom he remained for the rest of his life, though they could never marry because she was a commoner. After publishing a book on the nova of 1572 and lecturing informally at the University of Copenhagen in the autumn of 1574, Tycho was on the point of emigrating from Denmark better to pursue his intellectual interests and for the sake of his common-law wife and family.

Having heard of Tycho's plans, on 11 February 1576 King Frederick II offered Tycho the isolated island of Hven in the strait off Copenhagen and the funds to found and maintain an observatory. That summer Tycho began the construction of a castle/scientific institution, which he named Uraniborg, Urania's castle. From the alchemical furnaces in the basement, to the high observing decks with their removable roofs, to the library with its great celestial globe, to the garret rooms for students and assistants, every element of the structure was devoted to Tycho's dual interests of astronomy and alchemy. His facilities also included an instrument shop and a printing press and paper mill for disseminating his findings. With the problems of wind and vibration atop Uraniborg as well as the increasing scale of Tycho's instruments, he subsequently founded a nearby subterranean observatory called Stjerneborg, the castle of the stars, which housed his largest and most sophisticated instruments.

Tycho's whole enterprise on Hven was underwritten handsomely by King Frederick. After his death in 1588 during the minority of King Christian IV, when the country was run a council of regents filled with Tycho's relatives and friends, the generous funding continued. Tycho's heirs later estimated that he spent 100 000 florins over his lifetime amassing his observations. However, if Tycho was adept at soliciting patronage, he was not conscientious about fulfilling his modest responsibilities. He was therefore an easy target when King Christian came of age in 1596 intent on recapturing control over crown funds. The sudden collapse of Tycho's royal funding and persistent fears about transfering wealth to his family, who could not inherit because they were not noble, led Tycho to close Uraniborg and withdraw to northern Germany in 1597.

By the middle of 1599, Tycho had secured an even greater patron, the eccentric Holy Roman Emperor Rudolf II. Offered his choice of castles, Tycho selected Benatky Castle in the countryside north of Prague, where he began trying to construct a 'New Uraniborg'. However, by early 1601, the emperor had him move into the city in order to be more accessible for astrological advice. Tycho's research and publishing programs never recovered from the frequent moves of his last years. Finally, after drinking overgenerously at a party and not excusing himself from the table for some reason of etiquette, he suffered a fatal bladder complication, and died on 24 October 1601.

Tycho's dying words, 'Let me not seem to have lived in vain', were inscribed on the final page of his observing log by an assistant, who by virtue of the attrition in Tycho's staff during his final years unexpectedly rose to become his successor. His name was Johannes KEPLER.

The nova of 1572 and the comet of 1577

The two most celebrated of Tycho's observational findings involved transient phenomena, the nova of 1572 and the comet of 1577. Ironically, both occurred before the golden age of observing at Uraniborg, and were observed with Tycho's more primitive instruments. However, both had profound cosmological implications.

On 11 November 1572, Tycho noticed a new, intensely bright star in CASSIOPEIA. The new star (*nova stella*), now technically known as a SUPERNOVA, was bright enough to be visible during the daytime. Observing it with an astronomical radius and his early wooden half-sextant, Tycho discovered something of extreme importance: it displayed no diurnal parallax whatsoever, and therefore lay beyond the sphere of the Moon (a rather solid conclusion because the Moon itself displays 1° of diurnal parallax). Tycho published his findings in a modest book entitled *De nova stella* (1573).

According to Aristotelian cosmology, the supralunar realm consisted entirely of the fifth element, ETHER, which by its nature was incapable of change. However, the star had come into being and passed away. This single inconsistency was not itself enough to bring down a cosmological system that stood near the center of Western philosophy (Tycho himself regarded the nova as a special creation of God), but it became an important element in the reconsideration of Aristotelian cosmology that took place during the latter half of the 16th century.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tycho was privileged to observe not only a supernova but also one of the most spectacular comets in recorded history. He first sighted it at dusk on 13 November 1577 while catching fish for dinner from one of his ponds on Hven. Although Uraniborg was only under construction, Tycho managed to observe the comet with his new medium brass quadrant and his steel sextant. Because of the comet's movement, the reduction of the observations was more complex, but Tycho was able to plot its motion both in equatorial and ecliptic coordinates. He eventually concluded from the comet's motion and its imperceptibly small diurnal parallax that the comet was moving through space somewhere near the sphere of Venus.

This was a result of great significance, for by the late 16th century many had come to the conclusion that Aristotle's ether spheres were to be understood in terms of hard, impenetrable, crystalline spheres. According to Tycho, the comet had moved through those spheres, and he later claimed that he had shown that those spheres had no reality. His treatise on the nova having been overwhelmed by a mass of hysterical amateur literature interpreting its meaning, Tycho resolved to wait to publish until he could collect and rebut the literature generated by the comet of 1577.

The Tychonic system

Ever since the publication of NICOLAUS COPERNICUS's *De revolutionibus* (1543) astronomers were in a quandary. On the one hand, they admired Copernicus's elegant mathematical planetary theories, which had done away with the physical non-uniformity of motion inherent in Ptolemy's equant. On the other hand, the motion of the Earth in the heliocentric system was regarded by most as 'offending the principles of physics and of Holy Scripture', as Tycho put it. A few pursued the idea of transforming the COPERNICAN SYSTEM to keep the Earth at rest and to put into motion the Sun with the planets circling around it. However, the relative distances of the Earth and Mars in the Copernican system proved a stumbling block, for the transformed system required a problematic intersection of the Sun's and Mars' spheres.

The available evidence suggests that Tycho began developing his TYCHONIC SYSTEM only around 1584. He was at that time in the midst of an observational campaign to detect by means of Mars' diurnal parallax near opposition whether it approached closer to the Earth than the Sun. Since Mars does so in the Copernican system but not in the PTOLEMAIC SYSTEM, the observation promised the ability to decide observationally between the two cosmological systems.

With the evidence of Mars' parallax ambiguous but suggestive, and the evidence from the comet of 1577 for the incorporeality of the celestial spheres, Tycho was in the position to propose his new cosmological system. In the Tychonic geoheliocentric system, the Earth is at rest in the center of the universe while the Sun, with the planets orbiting around it, orbits the Earth. Tycho inserted his account of the new system as an initially unplanned chapter 8 in his book on the comet of 1577, *De mundi aetherei recentioribus phaenomenis* (1588), published as the second volume of Tycho's planned three-volume *Astronomiae instauratæ progymnasmata* (the first volume of which finally appeared posthumously in 1602).

Tycho's compromise cosmology was widely adopted and taught, giving rise to vociferous priority claims, especially against Nicholas Reimers Ursus, who published a similar system as his own later in the same year. (In Tycho's defense, it should be noted that an essential element of the system was his resolution of the problematic intersection of the Sun's and Mars' spheres.) The Tychonic system remained a serious alternative to the Copernican system well into the 17th century, long after the demise of the Ptolemaic.

Instrument design

Tycho improved instrument design in every respect, from the details of sights and divisions, to materials, to the type of instruments he made. He had his best instruments made during the 1580s, but even before the founding of Uraniborg Tycho had arrived at some of his most fundamental innovations.

Although Tycho eventually developed techniques to harness the power of instruments of great size, the precision of his instruments depended primarily on his innovative sights and divisions. He instituted the use of linear transversals applied to his circular scales on an experimental basis on his medium brass quadrant (ca 1577) and steel sextant (ca 1574), and then used them on all subsequent instruments. Although criticized by some for not being mathematically rigorous, Tycho demonstrated a maximum geometrical error of linear transversals to be 3″, far too small to be detectable. On his larger instruments, the transversals were capable of resolving angles as small as 20 arcsec.

Traditional pinhole sights were a far greater source of error. Since the front sight had to be at least big enough to sight the object through, it necessarily introduced an error of 5–10′. Tycho's innovative twin-slit sights did away with error in the front sight by replacing it with an obstruction. If an object could be sighted on either side of the obstruction through two parallel slits at the rear, it would be correctly oriented to the limit of the error introduced by the finite width of the rear sights, which could be progressively narrowed. These too were first introduced on the medium brass quadrant. The only significant modification of the system was the substitution of a cylindrical obstruction at the front, which would appear as an identical rectangular obstruction from any location along the arc, a very useful feature especially for his later quadrants and sextants, where two observers used a single post at the apex as a shared front sight.

Tycho experimented with a wide variety of designs but his working instruments fell into three families. Greatest in number were the azimuth quadrants, a design series that culminated in the revolving azimuth quadrant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(1586) and the great steel sextant (1588) in Stjerneborg. Tycho had less success with another traditional class of instruments, the armillary family. Although armillaries offered the advantage of being able to measure directly in equatorial or ecliptic coordinates, with their many rings they were fussy to make and use. Tycho was only able to become satisfied with them by progressively simplifying the design, and finally abandoning the ring arrangement altogether. His great equatorial armillary (1585) in Stjerneborg consisted only of a very large declination circle pivoting around a polar axis, with differences in right ascension being measured by means of a large fixed equatorial arc (see ARMILLARY SPHERE).

Tycho's signature instrument—the one for which he most vigorously claimed priority—was his astronomical sextant, which reached its mature form in the trigonical sextant (1582). It was explicitly designed as a replacement for the unsatisfactory traditional astronomical radius, or Jacob's staff, for the purpose of measuring angular separation in any plane. Because Tycho never developed satisfactorily accurate clocks, he did not measure positions in absolute equatorial coordinates but determined positions by triangulating from the positions of reference stars, for which the sextant was essential.

Various analyses have been made of the accuracy of Tycho's observations, which he acknowledged increased over his lifetime. It is safest to say that during his mature observing career, he consistently achieved his stated goal of 1′ accuracy.

Planetary theory

Tycho planned a restoration of astronomy based on a comprehensive collection of accurate observations, but died before the task was complete. The only theories he left in a reasonably satisfactory state were a solar and a lunar theory.

His solar theory followed a long tradition begun by HIPPARCHUS in consisting of a circular orbit with a simple, unbisected eccentrity. His major contribution was a thorough redetermination of the parameters; his final form had an apogee of $95\frac{1}{2}^\circ$ and an eccentricity of 0.0358 (equivalent to a post-Keplerian eccentricity of 0.0179), with an obliquity of $23^\circ\,31'\,30''$. The only value he did not reestablish was the erroneously large value of 3′ for solar parallax from antiquity, which played havoc with his solar theory. However, the results were tempered by Tycho's pioneering systematic attention to atmospheric refraction, with his exaggerated solar refraction table essentially canceling out the spurious value of solar parallax.

Most of the work on the lunar theory was undertaken by Tycho's most capable assistant CHRISTIAN SEVERIN (1562–1647). Too complex to be described here, the work was notable for Tycho's discovery, based on his extensive observations, of two new lunar anomalies, the variation and the annual equation.

Tycho put off the production of finished theories for the rest of the planets (the subject of the never-completed third volume of the *Progymnasmata*) until his Prague period, when first Longomontanus and then Kepler was assigned to the theory of Mars. He died before any were reduced to a finished form. Kepler would not produce a complete set of planetary tables based on Tycho's observations until long after Tycho's death. Although Kepler had in the meantime wrought decisive changes to astronomical theory, his *Rudolphine Tables* (1627) still listed Tycho as first author.

Bibliography

Christianson J 2000 *On Tycho's Island: Tycho Brahe and his Assistants, 1570–1601* (Cambridge: Cambridge University Press)

Dreyer J L E 1890 *Tycho Brahe: A Picture of Scientific Life and Work in the Sixteenth Century* (Edinburgh: Adam and Charles Black)

Gingerich O and Voelkel J R 1998 Tycho Brahe's Copernican campaign *J. Hist. Astron.* **29** 1–34

Thoren V E 1990 *The Lord of Uraniborg: A Biography of Tycho Brahe* (Cambridge: Cambridge University Press)

J Voelkel

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tycho Star Catalogs: The 2.5 Million Brightest Stars

The knowledge of positions, motions in the sky, brightnesses and colors of the brightest stars is the basis for investigations in nearly all branches of astronomy and astrophysics. The HIPPARCOS space astronomy mission launched in August 1989 and operating until mid-1993 was dedicated to obtain such information. The measurements by the Tycho experiment onboard the satellite have increased the previous knowledge of this kind by several orders of magnitude. The first Tycho Catalog (Tycho-1) published in 1997 gives the positions and magnitudes in two color bands for one million stars.

The motion in the sky (PROPER MOTION) could not be accurately derived from the change of star position during the short time interval of three years of the Hipparcos mission. The positions in the one hundred year old Astrographic Catalog have therefore been combined with the Tycho-1 positions to produce accurate proper motions for all stars, as published in 1998.

A second reduction of the Tycho observations has been carried out with more powerful modern computers, resulting in the Tycho-2 Catalog to be published in 2000. It will contain accurate information for 2.5 million stars, including proper motions based on the positions in the Astrographic Catalog and 145 other ground-based catalogs.

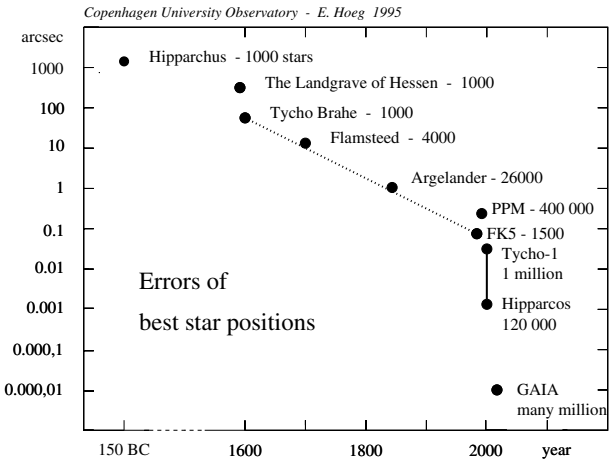


Figure 1. Errors of star positions in the most accurate catalogs. Tycho Brahe achieved a jump in accuracy. After four centuries with more gradual improvements another leap was obtained with the ESA satellite giving the Hipparcos and Tycho Catalogs containing a total of one million stars. Further improvement is expected with the planned ESA satellite GAIA for many more stars; about 40 million stars with 10 microarcsecond and 1 billion with better than 150 microarcsecond accuracy.

Catalogs of star positions

Catalogs of star positions and brightnesses (magnitudes) in the optical range are prerequisites for the study of almost all celestial objects. Given the position on the sky the stars are identified as individual objects, and can then in principle be observed by any kind of telescope for electromagnetic radiation, from radio waves over infrared and optical to x-ray and gamma-ray, in order to assess the physical nature of the object. The brightest stars in the optical range encompass a wide range of astronomical objects, from the compact white dwarfs to red giants and variable stars. The positions of other fainter objects such as galaxies and quasars are derived from special observations, relative to the positions of brighter reference stars.

The first prominent catalog in modern time was produced by TYCHO BRAHE in the late 16th century and contained about 1000 stars with an unprecedented accuracy of position. The accuracy of measurement was gradually improved as the technology became available, and the number of stars could be increased. The most accurate primary instrument (before Hipparcos) for ASTROMETRY, or positional astronomy, was the meridian circle, a telescope mounted on an east–west axis and observing the stars as they cross the meridian, one star at a time. This resulted in catalogs of fairly bright reference stars.

A large number of fainter stars could then be tied to these reference stars by means of exposures on photographic plates covering the whole sky. The first photographic survey of the sky was one of the largest international cooperations in astronomy carried out mainly from 1891 to 1920. It resulted in the Astrographic Catalog containing 8.6 million positions of about 4.5 million stars. But this catalog made very little scientific impact for nearly a century for two reasons. Firstly, the Astrographic Catalog positions were only available in 254 printed volumes, until a few years ago when they were painstakingly key-punched and thus put on machine-readable form. Secondly, an accurate net of reference stars only became available through the Hipparcos Catalog.

During the 20th century great efforts were expended by astrometrists in the production of hundreds of position catalogs both by meridian circles and photographically in order to satisfy the needs of astronomy. In the 1990s the scene changed with the completion of the Hipparcos satellite mission. A larger and more accurate reference system became available, and the Tycho Catalog contained positions of one million stars in a second observation epoch nearly 100 years after the Astrographic Catalog so that accurate proper motions could be derived from a combination of these two catalogs.

Tycho observations

The European Space Agency (ESA) satellite Hipparcos carried out astrometric and photometric observations from 1989 to 1993 resulting in the HIPPARCOS AND TYCHO CATALOGS. The satellite obtained very high accuracy astrometry for 120 000 stars, the Hipparcos Catalog. These stars plus an additional 935 000 stars were detected and measured with

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a more modest accuracy by the satellite's star mapper slits resulting in the Tycho Catalog, hereafter abbreviated to Tycho-1. This catalog contains astrometry and two-color photometry (B_T, V_T) for all stars. The two color bands in the blue and visual parts of the spectrum are close to the so-called Johnson B and V frequently used in astronomy.

The star mapper slits were placed in the focal plane of the Hipparcos telescope, beside the Hipparcos main field. The light passing through the slits was measured by simultaneous photon counting with two photomultipliers, one in each of the color bands. The original Tycho reduction was done by detecting a significant signal whenever the stars of an input catalog crossed the slit system. The improvement in the new reduction is achieved by superimposing the photons that were counted during all the about 130 observations per star obtained during the mission.

Table 1. Principal observational characteristics of the Tycho-2 Catalog. ICRS is the International Celestial Reference System. By means of proper motions the observed positions are transferred to the year J2000.0, the epoch of the catalog. The median values of standard errors are given. Unit: mas=milliarcsecond.

Observation period of satellite	1989.85–1993.21
Mean observation epoch	∼J1991.25
Epoch of Tycho-2 Catalog	J2000.0
Reference system	ICRS
coincidence with ICRS†	±0.6 mas
deviation from inertial†	±0.25 mas/yr
Number of entries	2540 000
Astrometric standard errors	
V_T < 9 mag, positions	9 mas
all stars, positions	60 mas
all stars, proper motions	2.5 mas/yr
Photometric errors on V_T	
V_T < 9 mag	0.012 mag
all stars	0.09 mag
Sky density‡	∼60 deg^{-2}
Limiting magnitude	$V \sim 12.5$ mag
Completeness to ∼90 %	$V \sim 11.5$ mag
Completeness to ∼99 %	$V \sim 11.0$ mag
Total number of observations	$\sim 300 \times 10^6$

† per axis for all three axes.
‡ depending on galactic latitude.

In the original Tycho-1 project astrometric and photometric data for 1052 000 stars to a limit of V_T =11.5 mag were derived. The brightest 99% of the stars obtained magnitudes in two passbands B_T and V_T. A precision (median internal standard error) of 7 mas was achieved in astrometry (positions, annual proper motions and parallaxes) for stars with V_T < 9 mag, and 25 mas for all stars. The median precision in photometry was 0.012 mag for V_T of the bright stars and 0.06 mag for the whole catalog. Double stars with sufficient brightness and separation larger than 2 arcsec were resolved. The processing was based on predicted star transits using a dedicated 'Tycho Input Catalog', compiled from existing star catalogs. The data treatment was carried out by the Tycho Data Analysis Consortium, using satellite attitude

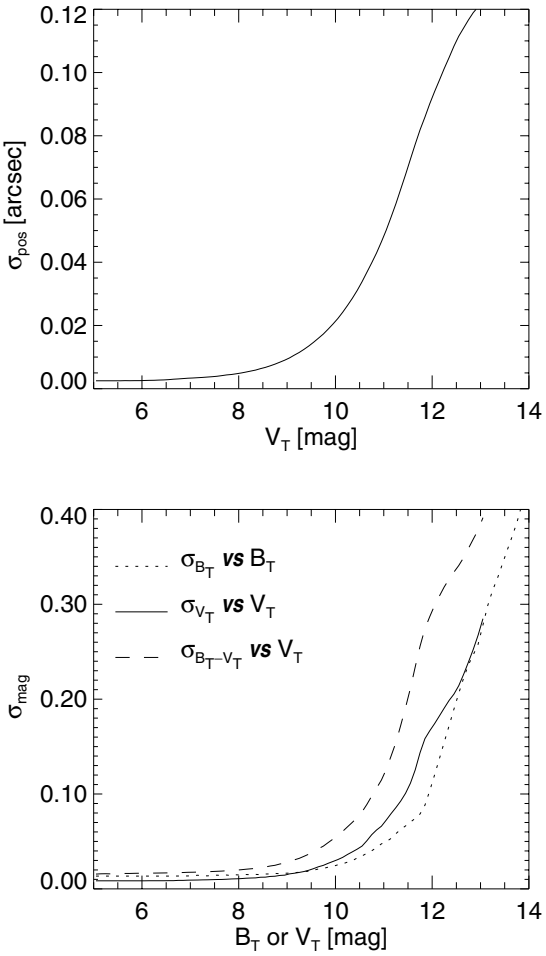


Figure 2. The Tycho-2 Catalog: standard errors of position components, magnitudes and color indices.

information, and photometric standard stars from the Hipparcos data reduction consortia.

The Tycho-2 catalog

The processing of Tycho observations for the first Tycho Catalog was based on detections of each transit above a certain signal-to-noise ratio (1.5 or 1.8).

For the second Tycho Catalog a much better solution, namely *a photon superposition for the whole mission*, has been realized. This idea was already suggested in 1982 when the Tycho project was proposed to the EUROPEAN SPACE AGENCY (ESA), but had to be abandoned at that time in favor of the more modest approach. The photon superposition gives better astrometric and photometric values for the fainter half of the original one million stars. A major reason for the improvement is that an astrometric estimation based on very few photons does not achieve the so-called CRAMÉR–RAO LIMIT. An estimation based on all photons from many transits comes much closer to that limit. Furthermore, many transits of the faint stars were below the limit

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of signal-to-noise ratio for the detection and were thus completely lost.

The Tycho-2 Catalog (see table 1 and figure 2) contains astrometry and photometry for 2.5 times as many stars as the Tycho-1.

Future developments

In the Tycho-2 processing, emphasis was placed on the optimal treatment of the bulk of stars which are single, but an effort was also made to detect and measure double stars. This resulted in about 8000 pairs with separations down to 0.8 arcsec, while the resolution limit in Tycho-1 was about 2 arcsec. The processing also gave insight into still unexploited possibilities. A special double star processing, where the Tycho-2 results are used as input, will give detection and measurement of closer pairs and of presently unknown fainter companions of many thousands of stars. It has also become clear that a special variable star processing should be applied because thousands of new variables can be detected, many more than can be detected from the published photometric annexes to the Tycho-1 Catalog. These surveys of double and variable stars would be complete and uniform over the sky to an extent which could only be superseded by a future astrometric satellite.

Bibliography

ESA 1997 *The Hipparcos and Tycho Catalogues* European Space Agency, ESA SP-1200 (17 volumes)

Høg E *et al* 1997 The Tycho Catalogue *Astron. Astrophys.* **323** L57–60

Høg E *et al* 1998 The Tycho Reference Catalogue *Astron. Astrophys.* **335** L65–68

Høg E *et al* 1998 Progress on the Tycho-2 Catalogue from the Hipparcos Mission *Proc. Int. Colloq. Modern Astrometry and Astrodynamics (Vienna, May 25–26, 1998)* at press

Urban S E *et al* 1998 The ACT Reference Catalog *Astronom. J.* **115** 1212

Erik Høg

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tycho's Star

A supernova remnant in Cassiopeia, 7.7° north of α Cas, which suddenly appeared as a brilliant naked-eye star in November 1572 and reached a maximum apparent magnitude of −3.5. Until its disappearance 16 months later, it was extensively studied by the Danish astronomer Tycho Brahe (1546–1601), who described its early appearance as follows: 'Initially, the new star was brighter than any other fixed star, including Sirius and Vega. It was even brighter than Jupiter.... It kept approximately the same brightness for almost the whole of November. On a clear day it could be seen at noon.' (*De Stella Nova*, 1573.)

In 1952 observations made by Robert Hanbury-Brown (b. 1916) at the Nuffield Radio Astronomy Laboratories, Jodrell Bank, identified the supernova location with the discrete radio source 3C 10 (also known as Cassiopeia B). Hanbury-Brown mapped the detailed structure of the radio source, following which Rudolph Minkowski (1895–1976) detected optical remnants with the 5 m (200 inch) Hale telescope at the Palomar Observatory, California. (This radio source should not be confused with the much stronger nearby source, Cassiopeia A, which is believed to be an ancient type II supernova remnant.) The remnant of Tycho's supernova has also been identified as a weak x-ray source.

The detailed records kept by Tycho Brahe enabled a light curve to be constructed which indicates that the object was a type I supernova. Recent measurements suggest that the remnant is probably between 8000 and 10 000 light-years distant, in which case it reached an absolute magnitude at maximum of about −16. This is equivalent to a luminosity approaching 250 million times that of the Sun.

At present the faint remnants of the expanding gas shell have a radius of 3.7′, equivalent to an actual diameter of between 17 and 22 light-years, and a rate of expansion of 6000–7500 km s^{-1}. This is the highest expansion rate deduced for any object, and is about ten times that of the Crab Nebula, also a young supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Tychonic System

The world system proposed in 1583 by the Danish astronomer Tycho Brahe (1546–1601). Unable to accept the Copernican doctrine that the Earth moves around the Sun, he put forward the view, later disproved by Kepler (1571–1630), that the planets move around the Sun, but the Sun and Moon move around the Earth. The theory explained the observed variations of the phases of Venus, for which the Ptolemaic system had no explanation.

See also: Copernican system, Kepler's laws, Ptolemaic system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

U Gem Stars

U Geminorum (U Gem) is the prototype for one subclass of dwarf nova (DN) systems. U Gem stars, like other DNs, brighten by factors of hundreds to thousands, sometimes in just a few hours. The eruptions recur quasi-periodically on intervals of weeks to years, with durations from a few days to a few weeks. U Gem stars, like all cataclysmic variables, are BINARY STAR systems consisting of a low-mass RED DWARF star (the secondary) and a WHITE DWARF star (the primary). The red dwarf is losing mass to the white dwarf through the inner LAGRANGIAN POINT (L1). This gas forms an ACCRETION DISK surrounding the white dwarf. The disk is usually the major source of radiation in the system. Another source of light is a bright spot formed where the stream collides with the outer edge of the disk. The other two subclasses of DN are the Z Cam and SU UMa stars. The Z Cam stars have similar outburst patterns with the addition of occasional extended standstills about 0.7 mag below maximum brightness. The SU UMa stars sometimes have superoutbursts that are brighter and longer duration than an ordinary outburst.

White dwarfs are very small yet contain about one solar mass. So they have deep gravitational potential wells. In DN systems the matter streaming from the red dwarf star releases a large amount of gravitational potential energy. In fact, accretion energy is the dominant light source in these systems. The gas stream leaving the L1 point of the secondary star cannot fall toward the white dwarf directly. Because the stream material carries angular momentum from the orbital motion of the secondary star it follows a curved path around the white dwarf star. The stream material forms a flat accretion disk. Most of the continuum radiation from DNs originates in dense regions of the disk and the bright spot formed by the stream–disk impact. The emission lines are formed in the lower-density outer disk or in the disk's chromosphere and perhaps the stream–disk impact site. In systems where the orbital plane is seen nearly on edge, the emission line profiles are double peaked. This is because the profile is actually composed of two Doppler-shifted components originating from opposite sides of the rotating disk. The accretion disk is the source of the DN outbursts. This was demonstrated by photometric observations of Z Cha. This binary is an ECLIPSING BINARY system, and during outburst the eclipse became much deeper, indicating that the disk was the source of the eruption rather than the foreground red dwarf.

Most of what we know about the visual light curves of DNs comes from the monitoring done by groups of amateur astronomers such as The American Association of Variable Star Observers (AAVSO). For many stars these visual observations span decades. The light curves reveal several details. Even for the same DN, no two outburst light curves are exactly the same (figure 1). For some systems (e.g. U Gem and SS Cyg (SS CYGNI)) there is a bimodal distribution of outburst durations with a tendency for outburst to alternate between long and short

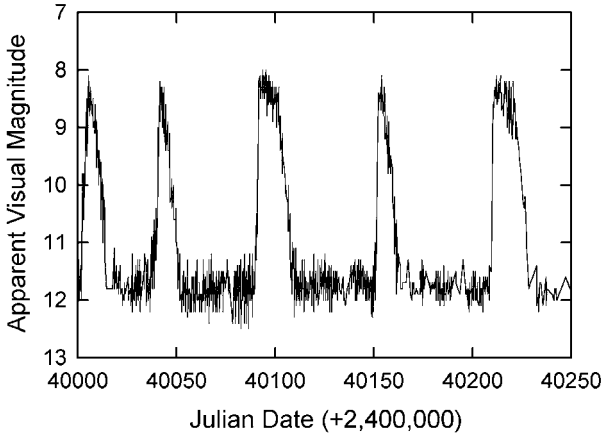


Figure 1. The U Gem star SS Cyg has an outburst about every 50 days. However, the duration, shape and amplitude of the outbursts are variable and mostly unpredictable. The light variations of SS Cyg are the best studied with the visual estimates made by members of the AAVSO. The figure above shows a small portion of their SS Cyg observations.

durations. Outbursts are classified, types A through D, based on their rise times to maximum light. Type A outbursts have rapid (∼2 days) rises. The system brightens at longer wavelengths first, with a delay of about a day between optical and the UV. In type B outbursts the rise of optical flux is a bit slower and nearly simultaneous at all wavelengths. There is a correlation (the Kukarkin–Parenago relationship) between outburst amplitude, A (in mag), and outburst duration, T' (in days),

$$A = 0.70\,(\pm 0.43) + (1.90 \pm 0.22)\log T'.$$

In quiescence the optical spectra are dominated by the hydrogen Balmer lines on a blue continuum. There are weak He I lines. Higher-excitation emission lines such as He II λ 4686 and the blend of C III and N III at λ 4650 are always very weak or absent. In the UV photospheric absorption lines of the white dwarf are visible, while lines from the secondary star can be seen in the IR for long-orbital-period systems. The light curve is often modulated on the orbital period owing to the changing aspect of the hot spot. In high-inclination systems this can produce a broad bump when the spot is seen face on, followed by an eclipse of the spot, disk and white dwarf by the secondary star. Analysis of eclipse light curves yields information about the size and location of these components.

In outburst the continuum brightens and the emission lines can become difficult or impossible to detect. They are often replaced by broad absorption lines. In the UV, low-inclination systems show blended emission and absorption lines (P Cygni profiles) that indicate mass outflow in a wind from the disk. In high-inclination systems the light curve bump produced by the bright spot remains at its quiescence level. This indicates that the outburst is not triggered by enhanced mass

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

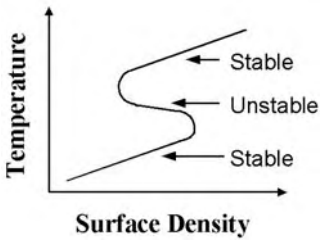


Figure 2. In the thermal instability model of DN outbursts there is a region in the disk with two stable temperature–surface density regimes. The mass transfer rate into the region determines where the region falls on the temperature–surface density curve. The outburst cycle is driven when the mass transfer rate into the region is too high for the lower branch and too low for the upper branch. The region cycles between the cooler region where mass transfer is less efficient (and mass accumulates) and the hotter upper branch with high transfer efficiency and larger mass inflow.

transfer from the secondary star. As time progresses the continuum fades, and the emission lines once again become prominent. However, the UV flux declines much more slowly, in some cases fading until the next outburst. This has been interpreted as the cooling of the white dwarf after the outburst.

Accretion disk models

In theoretical models the disk is treated as thin and flat with a constant flow of mass from the secondary star. The gravity of the disk itself is considered negligible compared with that of the white dwarf. The gas is assumed to move according to KEPLER'S LAWS. (This is not strictly true or there would be no inward motion and no mass accretion.) Gas particles above the plane of the disk feel a restoring force due to the vertical vector component of the central gravitational force. This is balanced by the local gas pressure, i.e. there is hydrostatic equilibrium in the 'vertical' direction. The surface density, Σ, at any point is defined as the amount of mass located beneath a square centimeter of surface, i.e.

$$\Sigma = 2\int \rho \, \mathrm{d}z$$

where ρ is the volume density and the z-axis is perpendicular to the disk plane. Because the disk particles orbit in a Keplerian fashion the rotation of the disk is differential. This results in shear between adjacent annuli producing a heating of the gas and an outward transport of orbital angular momentum. The net result is that most gas slowly spirals toward the white dwarf, releasing its gravitational potential energy as radiation. This process is determined by the viscosity of the gas. Unfortunately, we do not have a complete understanding of this viscosity. Convection may play an important role. In the disk models viscosity, ν, is treated as an adjustable parameter, α, i.e.

$$\nu = \alpha C_\mathrm{s} H$$

where C_s is the local sound speed and H is a density scale height. The disk models assume axial symmetry, that is for a given radius in the disk all physical parameters are the same. The disk models can predict the run of temperature and pressure throughout the disk. The emitted continuum spectrum is usually calculated by summing the BLACKBODY RADIATION from each annulus of the disk.

Outburst models

There are a variety of disk outburst models. The thermal instability model described below has been particularly successful. The following discussion is designed to give the reader a general idea of how this model works. References for further reading can be found in the bibliography.

The outbursts seen in DNs most likely result from increased mass flow through the accretion disk with accretion onto the white dwarf star. The rate of mass flow from the secondary star is usually assumed to be constant but there is a mechanism in the disk for temporarily storing matter. When the surface density reaches a critical value the viscosity increases and the matter is suddenly accreted onto the white dwarf. This produces the outburst. With time the surface density of the disk and the viscosity decrease and the mass build up in the disk begins anew.

Normally the temperature in a disk is expected to increase with surface density because higher surface densities lead to greater viscous heating. A plot of all possible combinations of local temperature versus surface density (T versus Σ) at a particular radius would give a positive-sloping curve. However, the outburst models require that at some location in the disk there are two possible values for temperature, for the same value of surface density. The outburst state corresponds to the high-temperature solution and the quiescent state to the low-temperature solution. The required curve in the temperature–surface density (T–Σ) plot is s shaped (figure 2). Only the upper and lower branches of this curve represent stable solutions for a disk. The outbursts are departures from thermal equilibrium.

For a disk to be in thermal equilibrium the temperature must return to its initial value despite small perturbations. Normally this is the case. For example, if the midplane temperature is slightly reduced for a time, the surface density would increase owing to a reduced efficiency of viscosity in transporting mass inward. The increased surface density results in increased viscous heating, which returns the temperature and the surface density to their initial values. This thermally stability is characterized by a positive slope of the T–Σ curve, because regions with high temperatures require high surface densities in order to supply sufficient viscous heating.

This scenario maintains thermal equilibrium provided that the gas opacity does not have a strong temperature dependence. A large opacity change affects the disk temperature because the opacity determines how freely the energy in the disk can escape into space. In zones of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

partially ionized hydrogen (approximately 10 000 K) the H^- ion is an important opacity source. It is very temperature sensitive, with opacity increasing as T^{10}. In this case, if the midplane temperature is perturbed downward slightly, the opacity will drop dramatically, allowing more energy to escape into space. This causes the temperature to drop even further. Likewise, if the temperature is perturbed upward, the temperature will continue to rise. A region of the disk in this state will either heat or cool until it reaches the stable branches at the top or bottom of the s-curve. The lower branch of the curve is thermally stable because it is too cool for significant hydrogen ionization and the opacity due to molecular absorption has a weak temperature dependence. Likewise, the upper branch is thermally stable because hydrogen is mostly ionized, which also gives a weak temperature dependence for the opacity (Kramers opacity $\sim T^{-2.5}$).

The conditions at each radius in the disk correspond to some point on the $T-\Sigma$ curve, which is determined by the local accretion rate. If this dictates a temperature where hydrogen is partially ionized, the local solution will fall on the middle branch of the s-curve, which is thermally unstable. This region of the disk will either heat or cool to reach thermal equilibrium. If the temperature in this unstable region is perturbed downward the temperature will decrease until the lower branch is reached. However, the accretion rate into this region is too high for this portion of the $T-\Sigma$ curve. The surface density increases as matter enters the region at a rate greater than it is being transported inward. The temperature rises because of increased viscous heating and increased gas opacity. The local temperature and surface density both increase following the lower branch of the curve until the lower bend of the s-curve is reached. The region must then rapidly heat until the temperature equals that of the thermally stable upper branch of the s-curve. The local accretion rate is then too low for this hot state, and both the temperature and surface density decrease until the upper bend of the s-curve is reached. The region then cools until it reaches the lower branch of the s-curve. The entire process then repeats. For the entire disk to go into outburst a heating wave propagates to other regions of the disk. Fully time-dependent models show that this does happen.

Comparison with observations

The disk model makes predictions that can be tested observationally. Steady-state disk models predict a flux distribution of $F_\lambda \sim \lambda^{-7/3}$. The spectra of DNs in outburst show a match to this exponent to within ± 0.3. If the disk instability begins near the outside the disk radius should increase during outburst. Studies of the eclipse light curves of U Gem in outburst have shown this to be true. The disk diameter peaks at about 0.4 times the stellar separations just after the start of an outburst and declines to about 0.28 just prior to the next outburst. Also, in this type of an eruption, the far-ultraviolet region of the spectrum should brighten after the optical region. This is expected because the ultraviolet light comes from the hotter inner disk and the heating wave requires time to travel to the inner disk. As mentioned above, type A outbursts show this behavior. The theory also predicts that, when the rate of mass transfer is above a certain critical level, the accretion disk should in a stable, high state where the gas is totally ionized throughout. This means that high-mass-transfer-rate cataclysmic systems should not have outbursts. The novalike systems (which show no outbursts) have mass transfer rates above this limit while the DNs have accretion rates below this limit in good agreement with the theory (see also CATACLYSMIC BINARIES: CLASSICAL AND RECURRENT NOVAE). This appears to be a strongest confirmation of the disk instability theory.

Details of the onset of an outburst and the manner in which it spreads through the disk still need to be worked out. For example, several observers have found evidence that changes in the disk begin 1–2 days before the start of an outburst. These include changes in emission line strengths and widths and slight rises ($\sim$0.1 mag) in the continuum level. For the most part, current theory does not address these observations which contain valuable clues to the outburst trigger mechanism.

Bibliography

An authoritative review of U Gem stars and all types of cataclysmic binary stars can be found in

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

An introduction to the accretion disk outburst model can be found in

Cannizzo J and Kaitchuck R 1992 Accretion disks in interacting binary stars *Sci. Am. (January)* p 92

A review of static disk structure can be found in

Pringle J E 1981 Accretion disks in astrophysics *Ann. Rev. Astron. Astrophys.* **19** 137

Ronald H Kaitchuck

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Uhuru (SAS-1/Explorer 42)

NASA satellite, launched from the San Marco platform off the Kenyan coast in 1970. The first satellite dedicated to x-ray astronomy. Completed an all-sky x-ray survey and studied individual sources. Discovered x-ray binaries, including Hercules X-1 and Centaurus X-1, and confirmed the variability of Cygnus X-1. Uhuru means 'freedom' in Swahili. (See also SMALL ASTRONOMY SATELLITE, EXPLORER.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ultraviolet Astronomy

Ultraviolet (UV) astronomy is the study of the electromagnetic radiation emitted by celestial bodies in the UV wavelength range, a portion of the spectrum simultaneously shielded by our own atmosphere and beyond the sensory limits of our sight.

Actually, the electromagnetic 'window' accessible to ground-based observers is quite limited, being virtually confined to the wavelengths the human eye is responsive to. In particular, the shielding effect of the Earth's atmosphere on the radiation coming from space becomes very high at wavelengths shortward of 320 nm, the adopted limit of the UV region of the spectrum. This phenomenon, mainly due to the absorption of oxygen and ozone, affects the entire UV and x-ray spectral regions, thus preventing astronomers from recording not only high-energy phenomena giving origin to x- or γ-rays, but also common processes (such as the thermal emission from hot stars) involving intermediate energies and producing mainly UV radiation. This obstacle could be overcome only when—starting in the 1960s—it became possible, by means of rockets and orbiting vehicles, to carry astronomical telescopes above the bulk of the atmosphere.

The nomenclature commonly adopted to split the UV domain is to some extent ambiguous. The wide term 'UV' often refers to the wavelength interval starting from the atmospheric cut-off (≈320 nm) down to the 'Lyman break', i.e. the limit of the hydrogen Lyman line series at ≈90 nm. The shorter-wavelength segment from the Lyman limit down to the beginning of the x-ray portion of the spectrum, arbitrarily taken at $\lambda \approx 6$ nm (or ≈200 eV, in x-ray nomenclature), is generally referred to as the 'extreme UV' (EUV). At the same time, the radiation whose wavelength is shorter than 200 nm (strongly absorbed by most gases) is frequently called, as a whole, the 'vacuum UV'. In the following I will keep to the convention of referring by 'far-UV' (FUV) to the region below Ly α (121.6 nm), by UV to the range between 121.6 and 200 mm, and by 'near-UV' (NUV) to the radiation emitted longward of 200 nm.

The advantage of accessing the UV range was manifold. First, as already pointed out, the majority of radiation emitted by stars whose photospheric temperature exceeds 10 000 K falls in the UV region (figure 1). The reason is that stars' energy distribution crudely follows Wien's law for an ideal radiator:

$$\lambda_{max}\, T = \text{constant} \qquad (1)$$

(where λ_{max} is the wavelength of maximum emission and T the temperature), thus showing the emission peak at shorter wavelengths as the temperature increases. As a consequence, one has to access the UV to properly record the energy distribution of the hottest stars as a function of wavelength, as well as their total energy output.

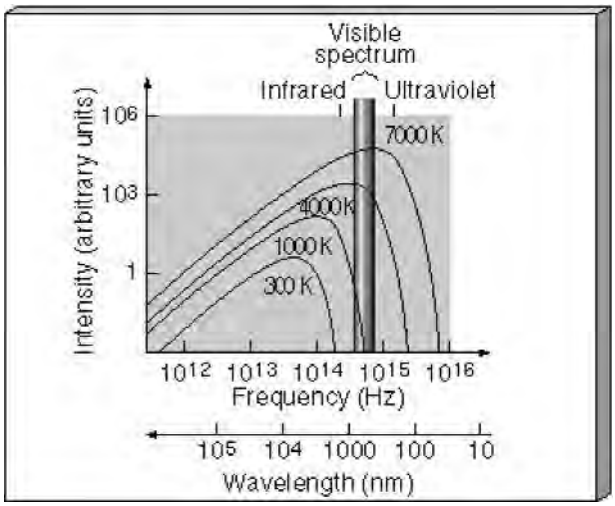


Figure 1. The energy distribution emitted by stellar photospheres as function of wavelength (or, equivalently of frequency) matches approximately the so-called Planck's curve, a mathematical idealization strictly correct for a body that absorbs all radiation falling upon it and thus—in a steady state—capable of re-emitting the same amount of energy. The curve shifts toward higher frequencies (shorter wavelengths) and greater intensities as the star's temperature increases. The net result is that stellar photospheres hotter than ≈10 000 K emit most of their light beyond the visible region of the spectrum. From Fortin E. 2002 *Above the Skies* (Harvard Extension School).

Secondarily, UV observations represent a superb research tool to investigate both physics and chemistry of astronomical bodies owing to the occurrence, in this wavelength range, of the so-called resonance spectral transitions, i.e. the most intense, ground-state transitions for most common atoms, ions and molecules.

Finally, when moving to UV, one can carry out deep surveys at significantly reduced levels of sky background. Actually, space observations ensure a sky 40× darker than at any wavelength from the ground at $\lambda \approx 200$ nm. This is obviously valuable when observing faint, extended sources such as distant galaxies.

One should not forget, however, that orbiting UV experiments are also influenced by specific background sources, namely the Earth's AIRGLOW and polar AURORA, as well as UV emission lines such as the far-UV hydrogen Ly α. Such a feature—ubiquitous within the solar system owing to the resonant scattering of solar photons off both interplanetary hydrogen and interstellar hydrogen flowing through interplanetary space—includes a particularly strong 'geocoronal' component, i.e. the Ly α resonant scattering produced off hydrogen atoms in the halo surrounding the Earth. As a consequence, unlike other sources such as the neutral oxygen O I 130.2 nm and 135.6

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

nm emission lines which affect only observations at low orbital altitudes, the geocoronal Ly α heavily interferes with UV observations obtained also by intermediate- and high-orbit (geosyncronous) satellites.

The Detection of UV Radiation

Although UV radiation is neither transmitted nor reflected as efficiently as are visible light and infrared (IR) wavelengths by optical components, the telescope construction techniques adopted to record the UV portion of the spectrum are still basically the same as those used in ground-based optical astronomy, consisting of parabolic–hyperbolic mirror combinations. One exception to this rule is the EUV waveband, as the observational techniques used in this wavelength range are more similar to those of x-ray astronomy. More precisely, observations in this range cannot make use of normal-incidence optics and the geometry of EUV telescopes has to ensure reflection via grazing incidence.

In order to provide a satisfactory reflection of UV radiation down to $\lambda \approx 100$ nm, surfaces are usually made of aluminum overcoated by MgF_2 (or LiF). High-efficiency normal-incidence reflection at shorter wavelengths, fulfilling the need to access the region shortward of the Lyman limit, can instead be reached by SiC mirror coatings.

As far as detectors are concerned, the present wide availability of silicon CHARGE-COUPLED DEVICES (CCDs), capable of recording photons over a very wide wavelength range (from 0.1 to 1000 nm), is appealing also in view of possible applications in the UV field. Such detectors are indeed the standard choice in many areas of today's astronomy and, together with MICROCHANNEL PLATE DETECTORS (MCP detectors) (see below), are currently replacing obsolete photographic emulsions to record UV observations (among recent missions, only the Shuttle-based Ultraviolet Imaging Telescope (UIT) still hosted image intensifiers/converters coupled to photographic films). When used as UV detectors, CCDs suffer from some limitations, however. Frontside-illuminated CCDs are insensitive to wavelengths shorter than 400 nm, for instance, although better performance in the UV can be obtained by resorting to thinned, back-illuminated chips. Moreover, a widespread practice intended to reach a suitable level of quantum efficiency (QE, i.e. the fraction of detected photons out of those reaching the detector) is to apply a fluorescing phosphor coating layer a few hundred nanometers thick, able to 'down convert' UV light, i.e. to absorb wavelengths shorter than 420 nm and re-emitting their energy in the visible region (around 520 nm). The CCD capable of the highest UV imaging sensitivity (down to $\lambda \approx 200$ nm) currently operating in space is the single 1024 pixel × 1024 pixel, back-illuminated detector of the Advanced Camera for Surveys—High Resolution Channel (ACS/HRC) on board the HUBBLE SPACE TELESCOPE (HST).

A step forward in UV light detection is now ensured by recently available MCP detectors. MCPs consist of a thin disk of a lead oxide glass crossed by many microscopic channels which act as individual photomultipliers and make the detector as a whole an image intensifier, when an electric potential is applied to its faces. They represent excellent two-dimensional (2D) detectors for the UV and EUV, when coupled with position-sensitive anodes which act as read-outs. Photon-counting, MCP-based detectors cope efficiently with many drawbacks of currently available CCDs, such as finite read-out noise, limited UV response and slow data access. This is the case of the state-of-the-art multi-anode microchannel array (MAMA) detectors which make both the Space Telescope Imaging Spectrograph (STIS) and the so-called Solar Blind Channel (SBC) of the just-mentioned ACS the most advanced UV-sensitive instruments currently operating on board HST. In the case of STIS, the combination of two MAMAs, whose sensitive wavebands slightly overlap (115–170 nm and 165–310 nm), ensures solar-blind, photon-counting capability (together with high spatial resolution imaging) over wide UV spectral range. A further step beyond CCD and MCP devices currently in use is expected shortly from the employment of a new class of detectors based on superconducting tunnel junction (STJ) technology. The main advantage of STJ detectors will be their ability to not only register each photon event but also to measure the wavelength of the detected photons over a very large spectral range, from FUV to the near-IR. What is more, STJ detectors hold particular promise in space-based UV astronomy, where their intrinsic spectral resolution is the highest and their QE exceeds that of the current instrumentation.

The Dawn of UV Astronomy

The technological development inherited from the World War II, especially in the field of long-range ballistic weapons, made it possible to take rocket-borne payloads above the atmosphere starting in the late 1940s. In particular, the first, unguided experiments flown by the Naval Research Laboratory (NRL) on board captured German V2 rockets returned unique information on the intensity and spectral distribution of the UV emission of the Sun. Space solar astronomy is indeed less demanding in terms of instrumentation sensitivity and rocket attitude control systems, thus forcing early UV astronomy to start as solar astronomy.

The rate of progress of UV space astronomy was quite fast, however. Actually, very simple, single-band UV photometers, flying on board unstabilized US Aerobee rockets were already able to provide the first UV observations of non-solar objects in the 1950s. More

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

exactly, the first UV data collected in space were broadband ($\Delta\lambda$=35 nm) flux measurements at 270 nm of 49 stars. UV spectrophotometric data were obtained for the first time a few years later (1961), by means of a scanning objective grating spectrometer flying again on board an Aerobee rocket, while the acquisition of spectra of pointed individual stars had to be postponed until 1965, when three-axis stabilization systems could be implemented on rockets. Although hampered by the very limited observing time available (a few minutes per flight), these pioneering observations (recorded on films to be collected after parachute landing) led to a revision of the temperature scale for hot stars and allowed astronomers to begin the study of mass loss phenomena from supergiant stars as well as to identify the UV interstellar absorption lines.

UV Imaging and Surveys

Although UV-sensitive facilities have brought unprecedented insights into many, largely unexplored, physical processes, their observations have been predominantly confined to the study of individual objects (as in the case of the International Ultraviolet Explorer (IUE) and HST satellites) or relatively small fractions of the sky (as in the case of recent balloon- or shuttle-borne imagers). As a consequence, our knowledge of the UV sky as a whole relies essentially on the pioneering survey mission of the TD-1 satellite in the early 1970s, whereas a sensitive, high-resolution (3–5 arcsec) all-sky UV survey is expected only with the launch of the Galaxy Evolution Explorer (GALEX) satellite (currently planned for April 2003).

In order to properly compare the capabilities of UV space imagers, one should be familiar with the most common units used in UV astronomy. Typically the fluxes of UV sources are given in 'monochromatic magnitudes' m_λ defined as follows:

$$m_\lambda = -2.5 \log (f_\lambda) - 21.175 \qquad (2)$$

where f_λ is the source flux density in erg cm^{-2} s^{-1} Å^{-1} at wavelength λ. As an alternative one can adopt the so called 'AB magnitudes' defined as

$$\mathrm{AB} = 2.5 \log (f_\lambda) - 48.60 \qquad (3)$$

where flux units f_ν are now expressed in erg cm^{-2} s^{-1} Hz^{-1}. The latter system has been specially defined to coincide with *V* magnitudes for a source with a flat spectrum. Its (wavelength-dependent) conversion into monochromatic magnitudes becomes m_{UV} = AB–2.26 in the middle of the UV domain (200 nm). An additional unit, used mainly to measure the UV background, is the count unit (CU), where 1 CU = 1 photon cm^{-2} sr^{-1} s^{-1} Å^{-1}.

With reference to the above photometric scales, the TD-1 catalog contains ≈58 000 sources brighter than m_{UV}=8.5, measured in four UV spectral bands centered at ≈ 155, 195, 235 and 275 nm. For comparison, the All-Sky Imaging Survey catalog to be provided by GALEX will contain ≈10^7 spiral/irregular galaxies and ≈10^6 quasi-stellar objects (QSOs), as well as ≈4×10^5 hot stars down to mag AB≈22 in two contiguous UV bands (135–180 and 180–300 nm).

TD-1's successful flight showed the potential of long-lived, UV orbiting observatories, i.e. of UV-devoted telescopes operating on board satellites. On the other hand, the advent of UV orbiters did not stop the development of other kinds of UV-oriented missions—both imagers and spectrometers—intended to exploit short-duration flights of rockets, stratospheric balloons and finally space shuttles. The first balloon-borne UV telescopes flown in the 1970s had the advantage of remarkably extending the observing time in comparison with sounding rockets, and although the observing window of these payloads floating at an altitude of ≈40 km is confined to the NUV owing to atmospheric extinction, long-duration balloon flights are still felt to be appealing enough to conduct all-sky surveys. Just as an example, upgraded versions of the French–Swiss balloon-borne FOCA experiment, which surveyed ≈70 sky square degrees imaging hundreds of galaxies down to a limiting magnitude m_{UV}≈19, will presumably fly in the future. Analogously, modern rocket-borne instruments such as the Narrowband Ultraviolet Imaging Experiment for Wide-Field Surveys (NUVIEWS) launched in 1996 turned out to be competitive in mapping the mid-UV background at a resolution of a few arcminutes. What is more, short-duration missions can be highly productive when multitask payloads are brought into space. This is the case for the two consecutive ASTRO-1 and ASTRO-2 Shuttle missions which hosted three co-mounted UV instruments: a UV imager (the UIT), a spectrometer capable of measuring UV fluxes down to the Lyman limit (the Hopkins Ultraviolet Telescope HUT) and the Wisconsin Ultraviolet Photo-Polarimeter Experiment.

Turning back to the orbiting telescopes which pioneered access to the UV sky, one should mention the rich dataset (consisting of pre-selected targets) of another successful mission, the Astronomical Netherlands Satellite launched into a polar elliptical orbit in 1974 and operated until early 1976. Its 22 cm diameter telescope focused UV light through a 2'.5×2'.5 aperture onto the focal plane of a spectrometer, thus sampling five, close UV spectral bands 15–20 nm wide starting from ≈150 nm.

A few additional experiments compensated for the lack of all-sky surveys by means of their high sky coverage. Among them one should mention the Far-Ultraviolet Experiment S201 which operated from the surface of the Moon during the Apollo 16 mission in 1972 assuring UV imagery of 10°–20° diameter fields, down to a limiting monochromatic magnitude m_{UV}≈12.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Two additional wide-field UV imaging facilities, namely the Wide-Field UV Camera (WF-UVCAM) and a first version of the Far-UV Space Telescope (FAUST), operated during the Spacelab 1 flight aboard Space Shuttle in the early 1980s. The WF-UVCAM's survey covered a large fraction of the sky, mostly in the southern hemisphere, taking advantage of its ultrawide field of view (66°), imaged with a resolution of 5'. FAUST was conversely designed to image a smaller field of view (≈8°) with a better angular resolution (1–2 arcmin). Unfortunately the high background experienced in orbit affected both experiments, whose data were recorded on photographic emulsions, and the best WF-UVCAM images did not record objects fainter than m_{UV}=9.3, while almost no objects were detected by FAUST. A second opportunity for this telescope came at the time an improved version flew on board the Space Shuttle Atlantis in 1992. This time the telescope, taking advantage of a modern MCP imaging detector, was capable of reaching a limiting magnitude m_{UV}=13.5 at 165 nm, as planned. Altogether FAUST's UV images covered a sky area of 0.33 sr (i.e. about 3% of the whole sky), thus ensuring UV photometry of more than 4000 star-like and extended sources.

Among UV archive images widely used are the data obtained by UIT. This instrument, namely a 38 cm telescope equipped with two selectable cameras providing both FUV and NUV 3 arcsec resolution images, gave extremely useful overall pictures of a wide variety of UV extended sources such as supernova remnants, globular clusters, OB associations, spiral galaxies and circumnuclear starbursts. These UV images of nearby galaxies are still highly valuable to date, owing to their relevance (as morphological templates) to high-redshift galaxies whose ground-based observations are preferentially made in the rest-frame UV (figure 2).

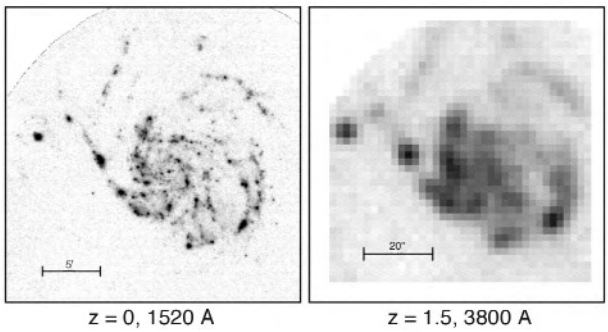


Figure 2. Left panel: a UV (150 nm) image of the luminous Sc galaxy M101 obtained by UIT. Right panel: a simulation of a galaxy with the same structure but 10× higher surface brightness at a redshift z=1.5 as observed by a 10 m class telescope. Owing to both reduced spatial resolution and changes in the galaxy's appearance with wavelength, the simulation is not easily recognizable as a normal spiral galaxy as seen at optical wavelengths. As a consequence, the observer could be induced to misinterpret such a difference as a genuine evolutionary effect. From Marcum *et al* 2001 *Astrophysics Journal Supplement Series* **132** 129.

UV Spectroscopy

As in any other field of observational astronomy, the amount of physical information carried by UV spectroscopy is much larger than that of simple imaging and/or photometry. This can explain the large effort devoted to taking into space UV spectrometers as early as possible in the history of space exploration, as well as the great scientific achievements of relatively simple facilities, like the IUE satellite.

UV spectrographs flown in the last few decades are characterized by improving performance in terms of limiting flux, sensitive spectral range and resolving power $R=\lambda/\Delta\lambda$ (a resolving power R=1000 corresponds to the ability to recognize individual features as close as a thousandth of their wavelength λ).

Among early successful instruments one has to include a satellite of the OAO series, i.e. the OAO-3, called COPERNICUS which was launched in 1972 and operated for eight years. Copernicus carried as the main UV experiment the UV Princeton equipment consisting of an 80 cm telescope and a grating spectrometer equipped with four UV-sensitive phototubes covering a short (95–145 nm) and a long (165–300 nm) wavelength range. The first high-resolution UV spectra (down to $\Delta\lambda$=0.5 nm) of early-type bright stars are due to this facility, although the primary goal of this mission was to make use of interstellar absorption lines, as seen against spectra of UV-bright stars, to estimate abundances in the interstellar medium (ISM).

The end of the same decade represented a turning point owing to the launch of the IUE. This long-lived, extremely successful satellite was brought into a geosynchronous elliptical orbit and operated as a guest observer facility from two stations located on either side of the Atlantic Ocean. Its instrumentation consisted of a 45 cm telescope with a 16' field of view, equipped with two echelle spectrographs covering two overlapping ranges from 115 nm to 335 nm. The majority of spectra were obtained through a large, oval aperture 10"×20" in size. The satellite could operate at low (R≈300) and high (R≈10 000) resolution, providing a spatial resolution orthogonal to the dispersion of ≈4". Its archive, containing about 108 500 spectra, is a unique heritage that the UV community will continue to exploit in the future.

A step forward in UV spectroscopy came with the two consecutive missions of the HUT telescope (see above) thanks to its unique ability to explore the UV energy distribution down to the Lyman limit (91.2 nm) and thus derive reliable temperature estimates for hot thermal sources (e.g. stellar populations including objects with T_e≈20 000–30 000 K). More precisely, HUT consists of a 90 cm telescope with a spectrograph covering the

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

82.5–185 nm region coupled with a photon-counting microchannel-plate intensifier providing a resolution of $\Delta\lambda$=0.3 nm.

UV-sensitive spectrographs have also operated on board HST since the beginning of its operational life. First-generation instruments were FOS and GHRS, both sensitive to UV wavelengths and largely complementary. The former UV–optical device operated at low and moderate resolution ($R\approx$250 and 1300) with a variety of apertures to optimize throughput, while two photon-counting detectors (digicons) could be used to select either a blue (115–550 nm) or a red (165–850 nm) channel. An interesting feature of FOS was its ability to isolate subarcsecond regions of an extended source. Conversely, GHRS was a UV-dedicated instrument capable of obtaining much higher-resolution spectra (up to $R\approx$ 80 000) of small wavelength intervals (from 28.5 nm to less than 1 nm). The above UV-sensitive spectrographs were replaced in 1997 by STIS, an innovative instrument making use of 2D detectors, thus providing spatially resolved long slit or slitless spectroscopy from 115 to 1100 nm from low to medium spectral resolution ($R\approx$600–14 000), together with high-resolution UV echelle spectroscopy up to a resolving power $R\approx$100 000.

The EUV

Overcoming the atmosphere's opacity by means of space-borne telescopes in not enough when moving to the EUV portion of the spectrum. Actually, at wavelengths shortward of the Lyman limit (91.2 nm) observations are hampered by the intervening opacity of the ISM, a phenomenon due to the line-of-sight atoms of hydrogen as well as of neutral and singly ionized helium (below 50.4 nm and 22.8 nm respectively). Moreover, as stressed above, a conventional optics consisting of curved reflecting mirrors would indeed absorb photons at wavelength shortward of ≈50 nm and the goal of reflecting the majority of such radiation can be achieved only by imposing sufficiently low incidence angles on traveling EUV photons. In other words, one has to resort to specifically designed telescopes, including a series of (metal) grazing-incidence surfaces, to focus EUV light.

Since the beginning of EUV astronomy interplanetary probes and manned spacecraft played quite a large role in this field. This is the case, for instance, for the earliest measurements of the EUV background (i.e. below Ly α) obtained by means of simple Geiger counters on board Soviet probes Venera 5 and 6. A later major step in the exploration of the EUV sky was made possible in 1975 during the the joint Apollo–Soyuz mission. Its EUV experiment—a 37 cm grazing-incidence telescope with selectable fields of view of a few degrees—discovered the first four EUV point sources, thus showing that, contrary to the prevailing opinion at that time, EUV astronomy was indeed possible. Other long-lasting experiments, namely modern UV spectrometers employing MCP intensifiers and covering the range 50–170 nm, were part of the scientific payload of Voyager 1 and Voyager 2 which started their journey into interplanetary space in 1977. A modified spare UV spectrometer (the EUVS experiment) built for the above Voyager program has been finally adapted to the currently operating Galileo spaceprobe, thus ensuring EUV/UV spectral capabilities from 54 to 128 nm.

The modern age of EUV astronomy came at the beginning of our decade thanks to the all-sky survey performed by the EUV Wide Field Camera on board the x-ray ROSAT satellite. Such a camera explored the whole sky in two, slightly overlapping, bands (6–14 nm and 11–20 nm) and the resulting final catalog contains 479 EUV sources, consisting mainly of hot white dwarfs and late-type (i.e. not hotter than the Sun) active stars.

An additional successful facility which, although brought into orbit in the Shuttle's payload bay, operated twice (in 1993 and 1996) as a free-flyer space UV/EUV telescope, was the German–US ORFEUS-SPAS platform. The ORFEUS (Orbiting and Retrievable Far- and Extreme Ultraviolet Spectrometer) instrument consists of a 1 m telescope equipped with two spectrographs, one of which, devoted to EUV, covered the region between 40 and 120 nm with a fairly high resolving power (R=3000).

Actually, the present-day reference dataset for knowledge of the EUV sky is represented by the source catalogs, images and the spectral atlas of the EXTREME ULTRAVIOLET EXPLORER (EUVE) satellite, which operated from 1992 until 2001. This facility included three grazing-incidence, co-aligned telescopes and an EUV spectrometer–deep survey instrument. The first three scanning telescopes, which are perpendicular to the spacecraft spin axis and possess a 5° field of view, allowed a complete survey of the EUV sky to be made in four bandpasses from 7 to 76 nm, while the fourth telescope could simultaneously scan the sky over a 2°×180° area along the ecliptic (i.e. the plane of the Earth's orbit). The latter telescope, aligned parallel to the spacecraft spin axis pointing away from the Sun, was allowed to integrate on a given source for a time as long as ≈0.5 h during a single orbit (versus a typical 500 s exposure of the main survey), so as to provide much deeper images. Altogether, EUVE obtained a vast amount of data from a largely unexplored region of the electromagnetic spectrum, thus ultimately giving to the field of EUV astronomy its own scientific identity, separate from other previously exploited spectral windows.

Highlights of UV Astronomy

The Solar System and The Sun

The study of cometary physics and Jupiter's auroral activity represents highly valuable achievements of UV astronomy. As far as the Jovian auroral phenomena are

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

concerned, one should be aware that current knowledge of their time-scales comes mainly from a successful coordinated IUE–Galileo probe UV campaign. More than that, UV observations of comets are studded with discoveries: IUE spectra of the comet IRAS–Araki–Alcock led, for instance, to the discovery of an unexpected molecule (S_2) never observed before in any astrophysical object, while more recent EUVE spectra of comet Hyakutake at 8–70 nm revealed for the first time the emission lines of ions that are brought to the comet by the solar wind and excited in charge exchange with cometary neutral species.

UV and EUV solar spectroscopy is presented in detail elsewhere in the *Encyclopedia* and the reader should refer to such a specific issue. It suffices to mention here the successfully operating solar experiments of the Solar and Heliospheric Observatory and Transition Region and Coronal Explorer, which have given unprecedented insight into the phenomena affecting the solar corona (such as the temperature structure and the 'seismology' of coronal loops), as well as into coronal plasma heating and the generation of the solar wind.

Stars

The impact of UV studies of stars is incalculably large, spanning an extremely wide variety of classes, evolutionary phases and phenomena. Roughly speaking, one can refer to three major subfields, namely studies of massive hot stars, (active) cool stars and binaries.

The role of very massive ((30–100) $M_{\odot}$) hot stars and their progeny (e.g. Wolf–Rayet stars and luminous blue variables) can be fully appreciated by taking into account that—when present—they dominate the appearance of their parent galaxies and, what is more, heavily affect their evolution through continuous mass loss, supernova explosions and ionizing radiation. Observations in the FUV domain are particularly well suited to characterizing the physics of strong stellar winds produced by the above stars (most of the strongest transitions of ionic species in their accelerated envelopes fall shortward of 200 nm) as well as to deriving the basic parameters of such extreme stars. This is the case, for instance, of a WO-type star (Sand 2) in the LARGE MAGELLANIC CLOUD (LMC), whose recent FUV spectrum obtained by the Far-Ultraviolet Spectrograph Explorer (FUSE) indicates a temperature $T_e \approx 150\,000$ K and a luminosity approximately 200 000× higher than that of the Sun. The evolution of such ultraluminous stars is so fast that this object is expected to undergo a supernova explosion within the next 10 000–50 000 yr. Surprising results have been obtained, again in the FUV, by observing the low-mass, very hot stars which expel huge, glowing layers of gas (the so-called planetary nebulae, excited by the UV radiation from the central star itself), while moving towards their final stage of white dwarf stars. In effect, when FUSE recently observed an object of this kind, namely the M27 nebula, it turned out that, although the central star is extremely hot, there is a lot of molecular hydrogen in the surrounding material, a discovery posing the unanswered question of how does molecular gas survive in such a harsh environment.

At the opposite end of the star temperature scale, UV data gave basic clues to the nature of activity in chromospheres and coronas of stars as cool as $T_e \approx 3000$ K. These objects, characterized in both the UV and the FUV spectral domains by a very weak continuum and a rich spectrum of emission lines formed in the regions going from the chromosphere to the corona (where the temperature grows from 4000 up to 150 000 K), show indeed spectra qualitatively resembling the chromospheric and transition region spectrum of the Sun, although with emission fluxes several orders of magnitude larger. The physical insight obtained by the study of UV emission lines of magnetically active stars is amplified by the occurrence of flares, i.e. energy release episodes in response to magnetic reconnection events. Their distribution in energy and relation to cool stars' activity and rotation has been fruitfully studied also in the EUV by means of the EUVE satellite.

Interacting binary systems such as symbiotic stars—consisting of a cool red giant and a hot, UV-bright white dwarf and thus characterized by highly energetic processes such as colliding star winds, nova-like outbursts etc.—are also quite suitable for UV investigations and have been extensively studied. Clues to the geometry of these binary systems also came from the study of the conversion of UV emission lines originating near the hot component into optical emission lines in the extended atmosphere of the cool giant (the so-called Raman scattering). Observers had instead to resort to EUV observations to succeed in detecting hot white dwarfs in Sirius-type, non-interacting binary systems containing a sufficiently hot (i.e. of the O, B and early A spectral type) main-sequence star. The reason is that the emission of the brighter star dominates here the radiation down to the UV and FUV regimes and neither IUE nor HST could be used to identify the degenerate companion.

Star Clusters

In globular clusters—old, roundish systems containing up to a million stars—the large majority of the population is quite cool. Their emission shortward of λ=200 nm comes indeed from a minority of stars crossing specific hot evolutionary phases. Among them have been identified single, highly evolved luminous stars which dominate the integrated UV light of the whole cluster, as well as populated 'tails' of lower-luminosity hot (up to $T_e \approx 20\,000$ K) core helium-burning stars which also emit most of their light in the UV domain. In addition, the present ability (by means of HST) to detect much fainter UV sources

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

is offering a more complex and potentially interesting scenario where other star classes (eruptive variables, the mysterious 'BLUE STRAGGLERS', etc.) also play a role.

HST UV imaging opened recently the possibility of studying also what might be the progenitors of today's globular clusters, namely a population of compact, young star clusters located in a variety of starburst environments (preferentially circumnuclear star-forming rings in galaxies). The most luminous of them—generally termed 'super star clusters' (SSCs)— could indeed survive as bound entities for more than 10 billion years, thus representing, on the basis of their radii, high luminosities, and presumably high masses, genuinely young globular clusters.

Supernova 1987A

The nearby supernova which appeared in the LMC in February 1987, monitored on a regular basis with all possible telescopes since the time of the explosion, is by far the best-studied object of its class. It is specifically mentioned here because a substantial contribution to the knowledge of its early evolution came from intensive UV spectroscopy ensured by the IUE satellite and, since the time of its launch in 1990, by HST. Among the exciting results obtained by means of UV observations one has to mention the identification of the supernova progenitor (which unexpectedly turned out to be a blue supergiant star), the possibility of deriving the complex structure of the intervening medium both in our own Galaxy and in the LMC toward the UV-bright supernova and an accurate determination of the distance to the supernova. Moreover, the fast evolution of the supernova remnant is still giving rise to phenomena on a time-scale of a few years (such as the appearance of brightening spots on the equatorial ring surrounding the exploded star, caused by radiative shocks created where the blast wave strikes dense gas regions). The physics of this phenomenon is currently under study by means of UV (STIS) spectra on board HST.

Galaxies

Observations of galaxies at UV wavelengths represent a highly effective means of investigating both their hot stellar content and energetic phenomena in their central regions. The latter is obviously the case of galaxies hosting active galactic nuclei (AGN). These two classes of UV sources reflect, in turn, two distinct mechanisms (thermal and non-thermal) of radiation emission.

Stellar Populations in Galaxies

For disk-dominated galaxies UV imagery offered a direct way of recognizing active sites of star formation at different scales (spiral arms, nuclear rings, star-forming complexes, etc.) as well as of estimating the rate at which new stars form. Moreover, the patterns of star-forming sites seen in the UV for nearby galaxies of different morphologies turned out to be of great relevance to reconstructing the morphological class of very distant galaxies whose visible and IR images recorded by ground-based telescopes do actually show—owing to cosmic expansion—their rest-frame UV emission (figure 2). The role of UV observations of young, massive stars became even more pronounced for galaxies hosting a 'starburst', i.e. a major event of star formation dominating their energetic output. For these objects, the ability to address many key issues, such as starburst morphology, the nature of the so-called SSCs and the impact of dust obscuration, is largely dependent on the availability of UV images.

The access to UV spectroscopy proved, in turn, to be imperative for characterizing stellar populations' temperature, metal content and age. In particular, the continuing decrease of UV emission of aging starbursts can be properly modeled by means of synthetic spectral energy distributions (SEDs) of single-generation stellar populations, thus providing a reliable dating method up to ≈1 Gyr. More than that, the sensitivity of the UV to stellar properties proved to extend even to the relatively cool, lower-mass main-sequence stars which dominate the mid-UV light after the exhaustion of star formation activity so that even the mid-UV region of the spectrum can be used for dating purposes for both local and distant galaxies (figure 3). Analogously, UV and FUV spectra of local starbursts provided by IUE, HUT and FUSE now ensure unique templates to quantify the chemistry and age of galaxies experiencing major star formation episodes in the early universe. Finally, UV turned out to be an effective window to understand also the late evolution of low-mass stars dominating spheroidal systems, such as elliptical galaxies and spiral bulges. These systems are indeed characterized by a sharp rise of their spectrum shortward of ≈200 nm (generally referred as

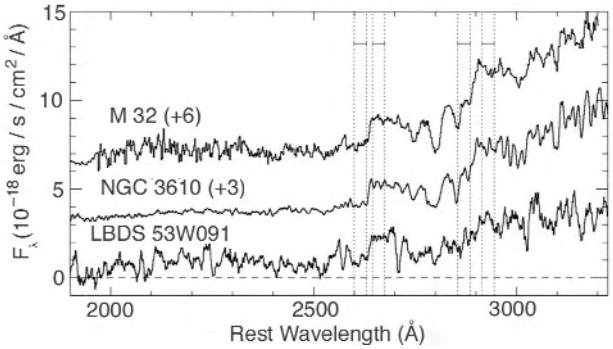


Figure 3. Comparison of the rest-frame UV spectrum of the galaxy LBDS 53W091 at z=1.5 and those of the local ellipticals M32 and NGC 3610. The vertical lines indicate the spectral intervals used to measure the breaks observed in the energy distribution. Such a comparison indicates that the age of the distant galaxy is not lower than ≈4 Gyr. From Spinrad *et al* 1997 *Astrophysical Journal* **484**, 581.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

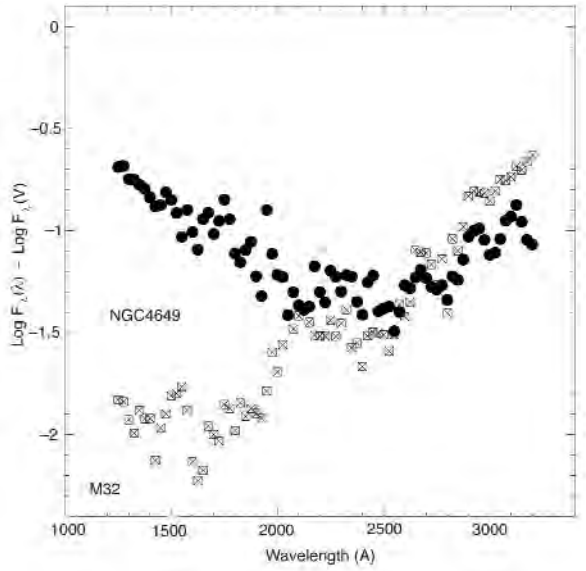


Figure 4. Comparison between the IUE energy distributions of NGC 4649, a typical metal-rich giant elliptical (full circles) and the low-luminosity metal-poor elliptical M32 (crossed squares), normalized to the visual flux. The adopted flux scale is logarithmic. Note the striking difference between the two spectra in the UV region dominated by the hottest stars, as opposed to the approximate match longward of ≈200 nm.

the 'UV-upturn' phenomenon), ascribed to some fraction of hot, evolved stars (figure 4). Since the onset of this phenomenon does depend on both the composition and age of the old populations, observers in the UV have at their disposal a second major probe of stellar populations (besides the strong UV emission of young, massive stars). More precisely, although the first hint of this feature came from the pioneering observations of the ANDROMEDA GALAXY bulge by means of the OAO-2 satellite, a much deeper insight into the origin of this behavior has been reached only by means of the large body of data recorded by IUE. IUE spectra showed indeed that the amplitude of this phenomenon depends on the chemical composition, as the higher the mean metallicity (percentage of elements heavier than hydrogen and helium) of the galaxy, the higher its UV flux. A further constraint on the nature of stars causing the UV upturn—namely its narrow range in temperature (T_e≈20 000–24 000 K)—came later from the spectra obtained by means of HUT. The maximum emission of hot stars of such a temperature falls indeed around 105 nm, just within that telescope's sensitive range.

Since then astronomers have faced the need to identify one (or more) evolved stellar components sufficiently hot, bright and long lived as to provide the observed UV flux and such a debate is still going on. While both observers and theoreticians favor the hottest, core-helium burning objects known as extreme horizontal branch stars and their progeny as the sought hot source, one cannot yet firmly establish whether they belong to a metal-poor or metal-rich minority of the population. This last issue, in turn, is quite important in the context of galaxy evolution, especially if the UV upturn onset is related to the attainment of a given maximum metallicity in the stellar population. If this is the case astronomers could indeed make use of it to constrain both metallicity and age of distant ellipticals.

Active Galactic Nuclei

Prolonged, IUE-based spectral monitoring campaigns aiming at understanding continuum and emission-line variability of Seyfert 1 galaxies (see SEYFERT GALAXIES) turned out to be quite successful, confirming the existence of time delays between continuum variations in different parts of the spectrum as well as of broad emission lines in response to UV continuum variations. Such phenomena, in turn, impose strict constraints on the theoretical models of these massive black hole candidates.

At the same time UV imaging showed its unique potential in the study of galaxy nuclei. Unlike optical and IR images of elliptical and bulge-dominated galaxies—dominated by the emission of surrounding cool stars—shorter-wavelength pictures permit nuclear non-thermal sources to show up over the galaxy stellar background. This, in turn, permits better study of known nuclei and searches for objects hosting low-luminosity activity. Such contrast enhancement made possible, for instance, the discovery of UV flares at the center of optically normal galaxies which would have been lost if recorded in other wavebands. AGN have also been successfully used as cosmic 'lighthouses' in the UV to investigate the properties of absorbing systems at high redshifts. In particular, the observation of high-z quasars enabled IUE (and subsequently to HUT, HST and FUSE) to explore the rest-frame EUV portion of their spectrum, thus leading to the detection of the continuous depression of the flux blueward of the He II Ly α emission line at 30.4 nm due to the singly ionized helium (He^+), the most abundant absorbing ion in low-density regions of the young universe. Such a phenomenon is known—by similarity to the effect expected for neutral hydrogen—as the He II Gunn–Peterson effect.

Interstellar (and Intergalactic) Medium

For a long time it has been well known that the space between stars is filled by highly rarefied gas and microscopic solid particles, generally referred to as 'interstellar dust'. UV spectroscopy provided the tool to properly investigate the composition and physical conditions of interstellar gas by analyzing the strongest absorption features of common elements (such as C IV, N V and O VI ions) and molecular species (such as H_2 and CO) towards bright stars. The UV continuum shape gave, in turn,

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

information on the constituents of dust particles, showing, for instance, that the extinction they operate on the UV light shows a typical 'bump' around 220 nm. Likewise, the vast regions between galaxies and groups of galaxies is permeated by a tenuous gas, made up primarily of hydrogen and helium atoms, called the intergalactic medium (IGM). The amount and composition of this widespread cosmic component can be studied by recording the absorption features it produces in the spectra of distant quasars. Very fruitful observations of these kinds obtained by FUSE allowed astronomers to identify helium absorptions in the FUV that have no corresponding hydrogen absorptions in optical, ground-based spectra along the same line of sight. Although the IGM is very tenuous, the amount of matter locked up in such a previously invisible component is huge, and its existence must be taken into account in current cosmological models.

UV Background

As noted above, the FUV background, i.e. the diffuse astronomical flux incident upon the Earth at wavelengths of 91.2–200 nm, is quite faint (mainly because of the zodiacal light's dramatic decrease below λ=250 nm) and, as such, does offer a window of very low sky contamination for observations of objects of low surface brightness. On the other hand, a proper, uncontaminated estimate of the true diffuse FUV background is hampered both by the terrestrial airglow and the galactic background, resulting from the scattering of the UV radiation field by interstellar dust. In spite of that, different UV-sensitive imagers (namely the rocket experiment NUVIEWS and the STIS instrument on board HST) have recently provided reliable measurements of both astronomical components. In particular, the galaxy contribution at λ 174 nm has been successfully mapped by NUVIEWS with a resolution of 5–10 arcmin over a substantial fraction of the sky, while the true diffuse FUV background along a line of sight with very low galactic neutral hydrogen column and without any significant contribution from airglow has been estimated by STIS (after masking Galactic and extragalactic sources to very faint magnitudes) in the bandpass 145–190 nm.

UV Telescopes Currently in Operation

The highly successful observations of FUSE, launched on 24 June 1999, are enhancing the role UV astronomy is playing at the beginning of the new century. FUSE consists of four co-aligned 39 cm × 35 cm mirror telescopes coupled with four spherical, holographically ruled gratings. It makes use of two MCP detectors and is specifically designed to obtain high-resolution spectra (R≈24 000–30 000) in the range between the Lyman limit and Ly α emission (90.5–118.7 nm).

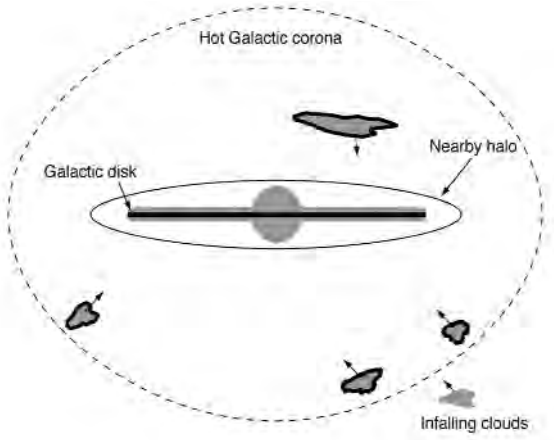


Figure 5. FUSE found evidence for a corona of tenuous hot gas much more extended than the well-known hot halo around the disk of the Milky Way. Clouds of gas raining into the Galaxy from the outside are heated as they pass through this corona—much like a meteor heats up as it passes through the Earth's atmosphere—and eventually become detectable in the FUV by FUSE. This hot, high-velocity gas has been detected at over 50 directions looking through our Galaxy's halo, indicating that it is widely distributed around our Galaxy.

The access to this largely unexplored FUV window represents indeed a breakthrough for many poorly known astrophysical fields. In effect, although previous missions did provide brief glimpses into the above wavelength range (suffice it to remember the Copernicus mission which opened the FUV window in the 1970s, and the two HUT and ORFEUS telescopes flown on Space Shuttle missions in the 1990s), FUSE is able to observe sources more than 10 000 times fainter than Copernicus at a resolution many times better than that obtainable with either HUT or ORFEUS. Thanks to this increase in sensitivity FUSE can explore, for instance, the outer reaches of the Milky Way (figure 5) as well as making use of quasars as continuum sources for absorption line studies of distant gas clouds (figure 5).

FUSE is a principal investigator (PI) class mission, and as such was developed to carry out several primary scientific objectives proposed by the PI science team and selected by NASA. The main programs include (i) a survey of O VI absorption in the Milky Way disk and halo and the Magellanic Clouds, in order to determine the physical properties and distribution of hot gas, as mentioned above, (ii) determining the abundance of deuterium in a wide range of Galactic environments, (iii) exploring the nature and distribution of the hot IGM and (iv) understanding the role of molecular hydrogen in circumstellar environments and planetary atmospheres.

Among other operating, UV-sensitive facilities, one has to mention again the ACS, a third-generation imager installed on board HST during Servicing Mission 3B in March 2002. This instrument actually consists of three

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

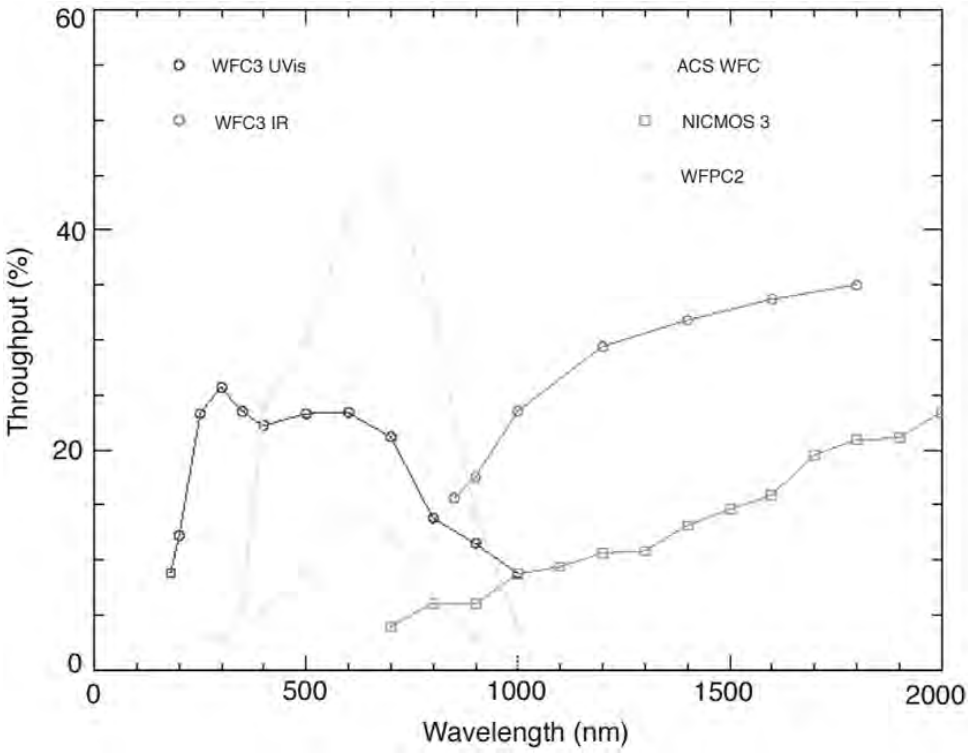


Figure 6. Expected throughput for the WFC3 instrument compared with those of other HST imagers. Note the extension into the UV of the wide-field imaging capabilities of HST ensured by the WFC3 instrument.

cameras covering a large spectral region from the UV to the near-IR (120–1000 nm). As already pointed out, its UV-devoted imagers are the SBC, equipped with a photon-counting MAMA detector (whose sensitivity is confined to the range 115–170 nm) and the HRC equipped with a CCD detector operating longward of λ=200 nm. Present-day UV astronomy is also taking advantage of a piggy-back UV instrument belonging to an x-ray-devoted mission, namely the UV/Optical Monitor on board the XMM-Newton satellite. This telescope, operating in the 170–650 nm range, can provide imaging on the central region of the 17 arcmin × 17 arcmin x-ray field of view.

Forthcoming UV Missions and the Future

As already mentioned above, among survey-devoted projects, a highly valuable scientific return will come from GALEX, expected to be launched in April 2003. GALEX is intended to obtain a series of spectroscopic and imaging surveys in the range 130–300 nm in order to study both the UV properties of local galaxies and the star formation and metal production history of galaxies over the redshift range $0<z<2$.[1]

At the same time, the highly improved HST UV imaging capabilities ensured by the recent deployment of ACS will be further expanded when the state-of-the-art Wide Field Camera 3 (WFC3) is put into operation on board HST during the last servicing mission (scheduled for early 2005, following Columbia's accident). This instrument is configured as a two-channel camera. The incoming beam, once corrected for spherical aberration in the HST primary mirror by means of a two-element optical system, is sent to either a near-IR channel or a UV–visible channel. The latter will image a 160 arcsec × 160 arcsec field, thus providing unprecedented wide-field imaging capabilities in the UV down to $\lambda\approx$200 nm (figure 6).

The last HST servicing mission will also include the replacement of the spherical aberration corrective optics—no longer necessary after the removal of the Faint Object Camera—with the Cosmic Origins Spectrograph (COS), a medium resolution spectrograph specifically designed to operate into the NUV and mid-UV. COS—optimized for observing faint point sources with intermediate spectral resolution ($R\approx$20 000)—consists of two channels: an FUV channel that is sensitive in

[1]Owing to the expansion of the universe the light coming from distant galaxies shows a redshift (expressed as $z=\Delta\lambda/\lambda$ in terms of increased wavelength), thus appearing graually shifted toward the red with increasing receding galaxy velocity which, in turn, reflects increasing distance and lookback time. As a consequence, a given redshift range Δz corresponds to a well-defined interval in lookback time, once a specific cosmological model is chosen.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

the 115–177 nm wavelength range and an NUV channel that operates between 170 and 320 nm. Taking advantage of its design, which minimizes the number of reflections (thus ensuring a dramatic improvement in sensitivity to faint objects over previous UV spectroscopic instruments flown aboard HST), it will permit several major issues in current astrophysics to be addressed, such as the properties of QSO absorption line systems and the IGM, the dynamics of the ISM in galaxies and galaxy halos, as well as the UV extinction in the Milky Way and the study of volatile gases in the atmospheres of solar system bodies.

The second half of the current decade should finally see the deployment of the World Space Observatory (WSO/UV). This spacecraft—to be regarded as a UV-devoted astrophysical observatory rather than a targeted mission—will consist of a 1.7 m diameter telescope equipped with both spectroscopic and UV-imaging instrumentation. More precisely, it will include a composite spectrograph capable of both high ($R \approx 60\,000$) and low ($R \approx 1000$) resolution, as well as two UV imagers—one for maximum spatial resolution, one for maximum sensitivity—operating essentially in the same wavelength range (110–340 nm).

WSO/UV will be innovative in many respects. For instance, it will be brought into an L2 orbit, i.e. it will operate from a point in space known as the second Lagrangian point located about 1.5 million km from the Earth in the anti-Sun direction. Such an orbit, besides being a very stable thermal and radiation environment, does offer the possibility of long uninterrupted observations, since Earth, Moon and Sun are permanently behind the spacecraft viewing direction. In addition, the approach chosen for the planning and execution of the WSO/UV will ensure significant cost savings, while facilitating the participation in space science of new countries.

Bibliography

A summary of all fields of UV research is given in the conference volume *Ultraviolet Astrophysics Beyond the IUE Final Archive*, eds W Wamsteker and R González Riestra, ESA SP-413 (1998).

A concise history of UV space programs is compiled by De Martino D and Buson L M 1999 UV space missions: from the past to the future *Memorie della Societa Astronomica Italiana* **70** (2) 315, while the specific issue of space surveys is discussed by Brosch N 1999 Ultraviolet sky surveys *Experimental Astronomy* **9** 119.

A description of the FUSE telescope, its performance and scientific background is given by Moos H W, *et al* 2000 *Astrophysical Journal* **538** L1; an analogous description for the GALEX satellite can be found in Bianchi L, *et al* 1999 *Memorie della Societa Astronomica Italiana* **70** (2) 365; readers interested to the WSO project should instead retrieve from the website http://arXiv.org the preprint astro-ph 0209333 by M A Barstow *et al*.

The most promising UV detectors under development are presented in *Ultraviolet–Optical Space Astronomy Beyond HST* (*ASP Conf. Series 164*) (1999) by O H W Siegmund (MP detectors) and P Jakobsen (STJ detectors).

Updated information can be retrieved via the following:

http://www.stsci.edu/hst/HST_overview/instruments (HST instrumentation)

http://astro.estec.esa.nl/SA-general/Research/Stj (STJ detectors)

http://www.vilspa.esa.es/iue (IUE satellite)

http://spacescience.nasa.gov/missions/index.htm (past UV space missions)

http://www.galex.caltech.edu (GALEX)

http://fuse.pha.jhu.edu (FUSE)

http://www.seas.columbia.edu/~ah297/un-esa/wso.html (WSO/UV).

Lucio M. Buson

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Ulugh Beg (1394–1449)

Astronomer, born in Soltaniyeh, Timurid, Persia (now Iran), the grandson of the Asian conqueror Tamerlane, and under his father's patronage built a large observatory in the city of Samarkand, directed by Ali-Kudschi, with Al'Kashi and Kadizada amongst the staff. The observatory was well equipped with large instruments that revealed errors in Ptolemy's data. Ulugh Beg and the staff observed the positions of the Sun, the Moon, the planets and 992 stars, publishing a catalog *Zij-i Sultani*. The book of tables contained trigonometric functions. Ulugh Beg also wrote poetry and history. When his father died, he reigned briefly as ruler of the Timurid empire but was overthrown and slain at Samarkand at the instigation of his own son, Abd al-Latif.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Umbriel

A mid-sized satellite of Uranus, discovered by William Lassell in 1851. Its diameter is 1170 km and it orbits the planet at a distance of 266 000 km. The satellite was not well imaged by Voyager 2; an ancient, cratered surface was visible, broken only by the bright-floored crater Wunda (diameter 160 km). Umbriel appears to be a primitive, undifferentiated body: one that has never melted internally and segregated into a core and mantle, and has remained largely geologically inactive.

See also: Uranus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Uncertainty Principle

A principle, named after the German physicist Werner Heisenberg (1901–76), which states that it is impossible to know simultaneously the precise position and momentum of a subatomic particle.

In order to measure the position of an electron, for example, it has to be observed. In order to observe it, we have to shine light on it. Because photons (particles of light) carry energy and momentum, each time a photon strikes an electron, the electron is deflected in an unpredictable way, so inducing an element of uncertainty into its velocity and momentum. The position of an electron cannot be measured to an accuracy much better than the wavelength of the illuminating radiation. The shorter the wavelength, the better the positional accuracy but the higher the energy and momentum of the illuminating photons and the greater the resulting uncertainty in momentum. As Werner Heisenberg proposed in 1927, the uncertainties in position (Δx) and momentum (Δp) are related as follows: $\Delta x \Delta p \geq h/4\pi$, where h is the Planck constant. Similar uncertainties exist in other pairs of measured quantities, such as energy and time. If the energy of a microscopic system is measured over a time interval Δt, the uncertainty in the measured energy (ΔE) is related to Δt by the following relation: $\Delta E \Delta t \geq h/4\pi$. The shorter the time interval, the greater the uncertainty in energy.

When the equivalence between mass, m, and energy, E, is taken into account ($E = mc^2$, where c denotes the speed of light), the uncertainty relationship between energy and time has important consequences. For example, over a very short interval of time, there is inherent uncertainty about the quantity of energy (ΔE) contained in any microscopic volume of space. If the time interval (Δt) is sufficiently short, the uncertainty principle implies that the volume may contain sufficient energy to create particles of mass $m \leq \Delta E/c^2$, provided that those particles vanish again before the end of the time interval (in practice, particle–antiparticle pairs form and mutually annihilate within time interval Δt). Such particles, which are called virtual particles because they cannot directly be detected, play an important role as force-carrying particles in quantum theories of the fundamental forces. The higher the mass of the virtual particle, the shorter its permitted lifetime. Because no particle can travel faster than the speed of light, the range to which a virtual particle can convey an influence is itself inversely proportional to its mass; thus forces that are conveyed by high-mass particles act over short ranges, while those that are conveyed by low or zero-mass particles have long, or infinite, ranges.

See also: electron, energy, fundamental forces, mass, photon, quantum mechanics, quantum theory, special relativity, wave–particle duality.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

United Kingdom Infrared Telescope

The 3.8 m United Kingdom Infrared Telescope (UKIRT) is the largest telescope in the world dedicated solely to observations at infrared wavelengths. UKIRT is sited near the summit of Mauna Kea, Hawaii, at an altitude of 4194 m. UKIRT was designed and built in the UK in the 1970s on a low budget. A recent program of upgrades has resulted in outstanding image quality, approaching the diffraction limit in the K band in the best seeing conditions.

The current main instruments in the near-infrared, both with polarimetry capability, are the CGS4 spectrometer (providing resolutions from 400 to 40 000), and the UFTI imager. These will be largely superseded in 2001 by UIST, a near-infrared imager/spectrometer. A mid-infrared imager/spectrometer Michelle (covering 8 to 25 micron) is due to be delivered in 2000. From 2003 a large fraction of UKIRT time is expected to be devoted to deep surveys using a new wide-field camera (WFCAM), and covering several thousand square degrees in J, H, K—the infrared equivalent of the SLOAN DIGITAL SKY SURVEY.

The telescope is owned and funded by the United Kingdom PARTICLE PHYSICS AND ASTRONOMY RESEARCH COUNCIL and operated along with the JAMES CLERK MAXWELL TELESCOPE by the staff of the Joint Astronomy Centre, located in Hilo.

For further information see
http://www.jach.hawaii.edu/JACpublic/UKIRT/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universal Distance Scale

The HUBBLE CONSTANT is the local expansion rate of the universe, local in space and local in time. The equation of uniform expansion is

$$v = H_0 r \tag{1}$$

where v is recession velocity and r is distance from the observer. To measure the proportionality constant, we adopt a definition of the Hubble constant as the asymptotic value of the ratio of recession velocity to distance, in the limit that the effect of random velocities of galaxies is negligible. In galactic structure the velocity of the local standard of rest has wider significance for the dynamics of the Milky Way than the velocity of the Sun. We have to abstract the Hubble constant from local motions in a similar way.

Existence proofs of the Hubble constant so defined, that we would still accept today, include the Hubble diagrams for cluster ellipticals (Sandage and Hardy 1973), for the Arecibo clusters (Aaronson and Mould 1986)[1] and for type Ia supernovae (Hamuy *et al* 1996).

It is also desirable to employ the Hubble constant in models which apply to larger spatial and temporal scales. The present value of the Friedmann model's $\dot{a}/a$, generally acknowledged as the standard model of the universe, is H_0. The expansion age of the universe is H_0^{-1} to a factor of order unity. If the standard model of cosmology were the Einstein–De Sitter model, then that factor would be two-thirds.

How to measure the Hubble constant

Three current schools of thought on this subject are represented in *Critical Dialogues in Cosmology* (Turok 1996). Two of these schools see the CEPHEID PERIOD–LUMINOSITY RELATION for Cepheid variable stars as crucial for establishing the distances of galaxies within 20 Mpc, which then calibrate 'standard candles'. These secondary distance indicators reach out into a region where an asymptotic ratio of recession velocity to distance can be sought.

Supernova Ia standard candles

The SUPERNOVA phenomenon results from instability in the equation of state of the degenerate material in the interior of WHITE DWARFS. According to Branch and Tammann (1992) SNeIa are explosions of a standard maximum power. Cepheid distances can be measured to some of the host galaxies.

The Hubble Space Telescope key project

The key project strategy has been fully described by Kennicutt *et al* (1995), and involves calibration of the infrared TULLY–FISHER RELATION and the equivalent dynamical relation for ELLIPTICAL GALAXIES, the (D_n σ) relation. Also calibratable from key project Cepheid distances are surface brightness fluctuations in galaxies, the globular cluster luminosity function, the planetary nebula luminosity function and the expanding photospheres method for type II supernovae.

One-step methods

Critics of both the above approaches point to the propagation of errors in a three-step ladder from trigonometric parallaxes to Cepheids to secondary distance indicators to H_0. One-step methods to determine H_0 include GRAVITATIONAL LENSING of distant QUASARS by intervening mass concentrations, which results in measurable phase delays between images of the time-varying input signal from the source. Where these phase delays can be accurately determined, and when the model mass distribution can be uniquely inferred, they directly lead to z/H_0, where z is the REDSHIFT of the lens (Turner 1997).

Compton interaction between the microwave background and hot gas in clusters of galaxies (the SUNYAEV–ZELDOVICH EFFECT) allows the physical size and distribution of the absorber/scatterer to be inferred, at least in projection, if the properties of the hot gas are also constrained by direct detection in x-rays. A distance can be determined by comparing the angular and physical projected size (Lasenby and Jones 1996).

The Large Magellanic Cloud (LMC)

The LMC is an anchor-point of the extragalactic distance scale. From the LMC it is possible to calibrate essentially all of the secondary distance indicators using the Cepheid PL relation.

The HIPPARCOS revision of the Cepheid PL relation, designated [H] in table 1, brings trigonometric PARALLAXES (Perryman *et al* 1997) to bear on the calibration in place of cluster main sequence fitting. Although this is a step forward in principle, the signal-to-noise of the parallax measurements, the short periods of the nearby Cepheids and the need for a fuller photometric study of the Hipparcos stars combine to render their impact on the problem uncertain at this stage.

The HST key project has adopted an LMC distance modulus of 18.50 ± 0.13 mag corresponding to 50 ± 3.2 kpc. The adopted reddening is $E(B-V) = 0.10$ mag.

Calibration of the Tully–Fisher relation

The Tully–Fisher relation is a luminosity–linewidth correlation for SPIRAL GALAXIES (Fisher and Tully 1977). Sakai *et al* (1999) present a calibration of $BVRIH_{-0.5}$ Tully–Fisher relations based on Cepheid distances to 21 galaxies within 25 Mpc, and they go on to determine the Hubble constant based on 23 clusters of galaxies within 10 000 km s^{-1}. Primarily, these are clusters studied by Giovanelli *et al* (1997). They obtain $H_0 = 71 \pm 8$ km s^{-1} Mpc^{-1}.

Like Giovanelli (1997), Sakai *et al* find that the effect of incompleteness biases in their sample is small (cf Sandage *et al* 1995, Teerikoorpi 1997) and correctable.

[1] But note that velocity perturbations as large as 5% may exist on these scales (Lauer and Postman 1994).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. The distance of the Large Magellanic Cloud.

Distance indicator	$m-M$	Reference
RR Lyrae	18.54 ± 0.07	Walker (1992, 1993a, b)
MACHO RR Lyrae	18.48 ± 0.19	Alcock *et al* (1997)
Cepheids	18.47 ± 0.15	Feast and Walker (1987)
Bump Cepheids	18.51 ± 0.05	Wood *et al* (1997)
Cepheids [H]	18.70 ± 0.10	Feast and Catchpole (1997)
Cepheids [H]	18.57 ± 0.11	Madore and Freedman (1998)
SN1987A	18.50 ± 0.13	Panagia *et al* (1991)
SN1987A [E]	$<18.44 \pm 0.05$	Gould and Uza (1998)
Red clump		Stanek *et al* (1998)
Eclipsing binaries	18.44 ± 0.07	Pritchard *et al* (1998)
Adopted	18.5 ± 0.13	

The (D_n σ) relation for ellipticals

The analogous scaling relation for elliptical galaxies relates velocity dispersion and isophotal diameter (Burstein *et al* 1987). Mould *et al* (1996) selected all the groups with four or more (D_n σ) measurements from Faber *et al* (1989) and renormalized their distances so that the geometric mean of that of the Fornax, Virgo and Leo groups was 15 ± 1 Mpc, which is obtained purely from Cepheid distances. Kelson *et al* (1999) used a similar approach to calibrate the fundamental plane of radius, surface brightness and velocity dispersion. They obtained $H_0 = 78 \pm 10$ km s^{-1} Mpc^{-1} from a sample of clusters with a mean redshift of approximately 6000 km s^{-1}.

Supernovae

There has been significant progress in the discovery, analysis and understanding of supernovae in recent years, not least in the area of theoretical models (Hoflich and Khokhlov 1996). Supernovae of type I are explosions of hydrogen-poor degenerate material, and mark the endpoint of the evolution of a class of white dwarfs, probably accreting in BINARY SYSTEMS (see Branch (1998) for a review).

Type Ia

Independent analyses have teased out some of the systematic differences between these onetime 'standard bombs'. Hamuy *et al* (1996) found improved fits on including decline rate as a parameter in their analysis of the HUBBLE DIAGRAM. Riess *et al* (1996) included intrinsic color in their analysis as well, and found further improvements in the model fit (but see Branch *et al* 1996). The data strongly reject the hypothesis that either or both of these dependences appear by chance. Calibration of this standard candle was initiated by Sandage *et al* (1994) with their distance to the host galaxy of SN 1937C. Gibson *et al* (1999) use the reddening corrected Hubble relations of Phillips *et al* (1999) and obtain $H_0 = 68 \pm 6$ km s^{-1} Mpc^{-1}, significantly higher than the results of Saha *et al* (1996a, b, 1997).

Type II

A fundamental technique for distance determination in both expanding and radially pulsating objects is that pioneered by Baade and Wesselink, in which the angular size of the expanding photosphere is monitored by means of the associated flux and temperature variations, and the linear size is monitored by integrating the radial velocity variations.

Although not blackbodies, as Baade and Wesselink assumed, Type II supernovae have photospheres whose emissivity can be calculated. Distances are derived from a comparison of photospheric angular diameter and the time-integrated expansion velocity (Kirshner and Kwan 1974). In the work of Schmidt *et al* (1994) the expanding photosphere method (EPM) has been employed to $cz = 14\,600$ km s^{-1}.

There is no evidence that any empirical recalibration of EPM is required at present by the Cepheid data. EPM provides independent and consistent constraints on the Hubble constant, currently yielding $H_0 = 73 \pm 11$ km s^{-1} Mpc^{-1}.

Surface brightness fluctuations

Surface brightness fluctuations (SBF) in early-type galaxies result from the counting statistics of the number of stars in a galaxy encompassed by a detector pixel (Tonry and Schneider 1988). If N is the average number of stars in a pixel, and $\bar{f}$ is the average flux from a star, the average brightness in a pixel is

$$\mu = N\bar{f}. \tag{2}$$

Since the number in a pixel of fixed angular size scales with distance as r^2 and the flux scales as r^{-2}, this mean surface brightness is independent of distance. The local rms from pixel to pixel, however, is

$$\sigma = \sqrt{N\bar{f}} \tag{3}$$

and this scales as r^{-1}: the smoothness of otherwise identical galaxies is proportional to distance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

We can derive a distance from a measurement of both μ and σ, provided that we know the average luminosity $\bar{L}$ corresponding to $\bar{f}$:

$$\bar{f} = \frac{\sigma^2}{\mu} = \frac{\bar{L}}{4\pi r^2}. \quad (4)$$

The radial power spectrum of $\sigma/\sqrt{\mu}$ yields $\bar{f}$ for early-type galaxies in a relatively straightforward manner, although the presence of GLOBULAR CLUSTERS in these galaxies adds a source of noise and potential systematic error (Blakeslee and Tonry 1995). The quantity $\bar{f}$ has been measured at a number of wavelengths, and its behavior as a function of galaxy color is reasonably well understood (Worthey 1993, Pahre and Mould 1994, Ajhar 1993) in terms of giant branch systematics in stellar populations. It remains to calibrate $\bar{L}$, and this has been done by Tonry *et al* (1997). Recently it has proved possible to employ the resolution of HST's planetary camera (WFPC2) to measure surface brightness fluctuations to 7000 km s^{-1} redshift (Thomsen *et al* 1997, Lauer *et al* 1998, Pahre *et al* 1998). Ferrarese *et al* (1999) find $H_0 = 69 \pm 7$ km s^{-1} Mpc^{-1}.

Results from the HST key project

We collect the results from calibrating five secondary distance indicators in table 2. There is good consistency between the results from widely different techniques, suggesting that the Hubble constant indeed lies between 60 and 80 km s^{-1} Mpc^{-1}. The key project calibration is entirely based on the Cepheid period–luminosity relation. Combining the constraints from the first four entries in table 2 yields 71 km s^{-1} Mpc^{-1}. Correction for the rather shakily known chemical composition dependence of the period–luminosity relation decreases the key project estimate of H_0 by 4% to 68 ± 6 km s^{-1} Mpc^{-1}.

Table 2. The Hubble constant from secondary distance indicators.

Indicator	H_0 (km s^{-1} Mpc^{-1})	
Tully–Fisher	71 ± 8	Sakai *et al* (1999)
(D_n σ)	78 ± 10	Kelson *et al* (1999)
SBF	69 ± 7	Ferrarese *et al* (1999)
SNIa	68 ± 10	Gibson *et al* (1999)
SNII	73 ± 12	Schmidt *et al* (1994)
Key project	68 ± 6	Mould *et al* (1999)

We caution the reader against the conclusion that the uncertainty in H_0 can be estimated from the consistency of the results in table 2. Uncertainties such as the distance of the LMC affect *all* the entries in Table 2 systematically. Detailed distinctions between random and systematic errors have been made by Madore *et al* (1998), to which we refer the reader.

Cepheid independent methods

Gravitational lensing

At the time Blandford and Narayan (1992) reviewed the cosmological applications of gravitational lensing, the time delay for the prototype lensed QSO 0957+61 was still in dispute at the 30% level. A clear observational determination has now been made (Kundic *et al* 1998), leaving the accurate description of the gravitational potential of this lens as the principal source of uncertainty in the measurement of the Hubble constant from it. Finding a unique model of the mass distribution is in general a difficult task for a system to which a major galaxy, its perturbers and the common cluster halo can all contribute. Turner (2000) asserts that the model for 0957+61 is robust and that the resulting $H_0 = 64 \pm 13$ km s^{-1} Mpc^{-1} for $\Omega = 1$ represents 95% confidence limits. The cosmological geometry is not a major issue for this $z = 0.36$ lens, affecting H_0 by +7% if $\Omega = 0.1$.

Sunyaev–Zeldovich effect

The thermal bremsstrahlung luminosity of the hot gas in x-ray emitting clusters of galaxies is given by

$$L_x = 1.4 \times 10^{42} n_c^2 r_c^3 T_x^{1/2} \text{ erg s}^{-1} \quad (5)$$

where n_c is the number density in the cluster core, r_c is the core radius and T_x is the temperature of the gas.

The Compton effect of electron interactions with the microwave background radiation yields a temperature diminution of the radiation:

$$\delta T/T = -0.0128 n_c r_c \frac{kT_x}{m_e c^2}. \quad (6)$$

One may then solve for r_c, whose angular diameter is also observable. According to Birkinshaw (1998) the current best fit to the clusters observed to that date is for a Hubble constant of about 60 km s^{-1} Mpc^{-1}.

Do we live in a bubble?

Apparently lower values of H_0 from more distant probes of the expansion suggest a model with a locally low density of galaxies. There are, however, constraints on the hypothesis that we live in a bubble, which we should consider. These are the observational determinations that the expansion is linear on 100 to 1000 Mpc scales, such as the work of Sandage and Hardy (1973) on brightest cluster galaxies.

More modern demonstrations include supernovae at intermediate distances and extension of the Tully–Fisher relation to 15 000 km s^{-1} (Dale *et al* 1999). These results limit the difference between the global and local values of the Hubble constant to a few per cent. The rarity of low-density bubbles is also attested by the microwave dipole anisotropy on degree scales. Wang *et al* (1998) find a robust upper limit on the global deviation from the local 10^4 km s^{-1} sphere of 10.5% in H_0 with 95% confidence.

The future

The Hubble constant appears to be a quantity which is well defined on large scales. Homogeneity and isotropy suggest that we are measuring a property of the universe

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

at the current epoch and not some local *ad hoc* parameter. A number of projects therefore suggest themselves:

- Microarcsecond astrometry from space will allow us to measure the trigonometric parallax of the LMC (Shao 1999). A parallax based Cepheid PL relation would be a nearer-term goal, improving on the results of Hipparcos.
- Interferometric measurement of Cepheid angular diameter changes will yield geometrical distances to these standard candles.
- The key project distances to galaxies in that program are now capable of improvement with infrared cameras on HST. This will essentially remove remaining uncertainties arising from extinction.
- One-step distance measurements which avoid the classical extragalactic distance scale will be further developed. The astrophysical models which describe rich clusters and galaxy lenses at large distances need to be probed in more detail.
- Miyoshi *et al* (1995) have shown that H_2O masers orbiting a massive compact object and following a Keplerian rotation curve provide a remarkably good fit to the radial and transverse motions of the components in NGC4258. There is potential in this technique to achieve higher precision and greater distances.

Developments in the subject are logged by Huchra (1998).

Acknowledgments

I wish to acknowledge my fellow team members in the H_0 key project. Team members include Shaun M Hughes, Fabio Bresolin, Shoko Sakai, Barry F Madore, Robert Hill, Laura Ferrarese, Holland C Ford, Garth D Illingworth, Daniel Kelson, John A Graham, Nancy Silbermann, Lucas Macri, Randy Phelps, Brad Gibson, John G Hoessel, Mingsheng Han, John P Huchra, Anne Turner, Abhijit Saha, Kim Sebo and Peter B Stetson. We are grateful to Brian Schmidt and John Tonry for their contributions. Much of the work presented in this paper is based on observations with the NASA/ESA Hubble Space Telescope, obtained by the Space Telescope Science Institute which is operated by AURA Inc under NASA contract no 5-26555.

References

Aaronson M and Mould J 1986 *Astrophys. J.* **303** 1
Ajhar E 1993 *PhD Thesis* MIT
Alcock C *et al* 1997 *Astrophys. J.* **482** 89
Birkenshaw M 1998 *Phys. Rep.* **310** 197
Blandford R and Narayan R 1992 *Ann. Rev. Astron. Astrophys.* **30** 311
Blakeslee J and Tonry J 1995 *Astrophys. J.* **442** 579
Branch D 1998 *Ann. Rev. Astron. Astrophys.* **36** 17
Branch D *et al* 1996 *Astrophys. J.* **470** L7
Branch D and Tammann G 1992 *Ann. Rev. Astron. Astrophys.* **30** 359
Burstein D *et al* 1987 *Astrophys. J.* (suppl) **64** 601
Dale D *et al* 1999 *Astrophys. J.* **510** L11
Faber S M *et al* 1989 *Astrophys. J.* (suppl) **69** 763
Feast M and Catchpole R 1997 *Mon. Not. R. Astron. Soc.* **286** L1
Feast M and Walker A 1987 *Ann. Rev. Astron. Astrophys.* **25** 345
Federspeil M, Sandage A and Tammann G 1994 *Astrophys. J.* **430** 29
Ferrarese L *et al* 1999 *Astrophys. J.* **529** 745
Fisher J R and Tully R B 1977 *Astron. Astrophys.* **54** 661
Gibson B *et al* 1999 *Astrophys. J.* **529** 723
Giovanelli R 1997 *The Extragalactic Distance Scale* ed M Livio, M Donahue and N Panagia (Cambridge: Cambridge University Press) p 113
Giovanelli R, Haynes M, Herter T, Vogt N, da Costa L, Freudling W, Salzer J and Wegner G 1997 *Astron. J.* **113** 53
Gould A and Uza O 1998 *Astrophys. J.* **494** 118
Hamuy M, Phillips M, Suntzeff N, Schommer R, Maza J and Aviles R 1996 *Astron. J.* **112** 2398
Hoflich P and Khokhlov A 1996 *Astrophys. J.* **457** 500
Huchra J P 1998 http://www.cfa-harvard/∼huchra
Kelson D *et al* 1999 *Astrophys. J.* **529** 768
Kennicutt R *et al* 1998 *Astrophys. J.* **498** 181
Kennicutt R, Freedman W and Mould J 1995 *Astron. J.* **110** 1476
Kirshner R and Kwan J 1974 *Astrophys. J.* **193** 27
Kundic T, Turner E, Colley W, Gott J R, Rhoads J, Wang Y, Bergeron L, Gloria K, Long D, Malholtra S and Wambsganss J 1998 *Astrophys. J.* **482** 75
Lasenby A and Jones M 1996 *The Extragalactic Distance Scale* ed M Livio, M Donahue and N Panagia (Cambridge: Cambridge University Press) p 76
Lauer T and Postman M 1994 *Astrophys. J.* **425** 418
Lauer T *et al* 1998 *Astrophys. J.* **499** 577
Madore B and Freedman W 1998 *Astrophys. J.* **492** 110
Madore B *et al* 1998 *Astrophys. J.* **515** 29
Miyoshi M *et al* 1995 *Nature* **373** 127
Mould J, Sakai S, Hughes S and Han M 1996 *The Extragalactic Distance Scale* ed M Livio, M Donahue and N Panagia (Cambridge: Cambridge University Press) p 158
Mould J *et al* 1999 *Astrophys. J.* **529** 786
Pahre M and Mould J 1994 *Astrophys. J.* **433** 567
Pahre M *et al* 1998 *Astrophys. J.* **515** 79
Panagia N, Gilmozzi R, Macchetto D, Adorf H M and Kirshner R 1991 *Astrophys. J.* **380** L23
Perryman *et al* 1997 *Astron. Astrophys.* **323** L49
Phillips M *et al* 1999 *Astron. J.* **118** 1766
Pritchard J *et al* 1998 *New Views of the Magellanic Clouds (ASP Conf. Ser.)* ed Y H Chu, J Hesser and N Suntzeff in press
Riess A, Press W and Kirshner R 1996 *Astrophys. J.* **473** 88
Saha A, Sandage A, Labhardt L, Tammann G A, Macchetto F D and Panagia N 1996a *Astrophys. J.* **466** 55
Saha A *et al* 1996b *Astrophys. J.* (suppl) **107** 693
Saha A *et al* 1997 *Astrophys. J.* **486** 1

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Sakai S *et al* 1999 *Astrophys. J.* **529** 698
Sandage A and Hardy E 1973 *Astrophys. J.* **183** 743
Sandage A and Tammann G 1997 *Critical Dialogues in Cosmology* ed N Turok (Singapore: World Scientific) p 130
Sandage A, Tammann G and Federspiel M 1995 *Astrophys. J.* **452** 1
Sandage A *et al* 1994 *Astrophys. J.* **423** L14
Schmidt B, Kirshner R and Eastman R 1992 *Astrophys. J.* **395** 366
Schmidt B *et al* 1994 *Astrophys. J.* **432** 42
Shao M 1999 *Space Interferometry Mission* JPL 400-811 (Pasadena: NASA)
Stanek K, Zaritsky D and Harris J 1998 *Astrophys. J.* **500** L141
Teerikoorpi P 1997 *Ann. Rev. Astron. Astrophys.* **35** 101
Thomsen B, Baum W, Hammergren M and Worthey G 1997 *Astrophys. J. Lett.* **483** L37
Tonry J 1996 *The Extragalactic Distance Scale* ed M Livio, M Donahue and N Panagia (Cambridge: Cambridge University Press) p 300
Tonry J and Schneider D 1988 *Astron. J.* **96** 807
Tonry J *et al* 1997 *Astrophys. J.* **475** 399
Turner E 2000 *Gravitational Lensing: Recent Progress and Future Goals* ed T Brainerd and C Kochanek in press
Turok N 1997 *Critical Dialogues in Cosmology* (Singapore: World Scientific)
Walker A 1992 *Astrophys. J.* **390** L81
Walker A 1993b *Astron. J.* **105** 527
Walker A 1993b *Astron. J.* **106** 999
Wang Y, Turner E and Spergel D 1998 *Astrophys. J.* **498** 1
Wood P, Arnold A and Sebo K 1997 *Astrophys. J.* **485** L25
Worthey G 1993 *Astrophys. J.* **409** 530

Jeremy Mould

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universal Time (UT)

A measurement of time used for civil timekeeping and based on the Sun's daily apparent motion. It is the mean solar time of the Greenwich meridian at 0° longitude. Universal Time replaced Greenwich Mean Time in 1928. In 1955 the International Astronomical Union defined a number of categories of Universal Time with increasing accuracy. UT0, the least accurate, is based on the optical observations of star transits at various observatories and represents the initial values of Universal Time. These values differ slightly from each other because of the effects of polar motion. By correcting for polar motion, UT1 gives the precise angular coordinate of the Earth about its spin. UT2 derives from UT1 by taking into account annual changes in the Earth's speed of rotation. Coordinated Universal Time is the international basis of civil and scientific time and is used for broadcast time signals. It is obtained from an atomic clock that is set to remain within ± 0.90 s to UT. In this way, the solar time that is indicated by Universal Time is synchronized with atomic time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universe

The sum of everything that exists and of which we can be aware; the entirety of space. There is a semantic difficulty in talking about the universe; on the one hand, we define it to be 'everything', but it may be (a) that our universe is finite, yet unbounded; (b) that the accessible universe is only a small part of a much larger entity, most of which we cannot observe; or (c) that there exist other universes of which we are not 'aware'.

See also: astrometry, astronomy, astrophysics.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universe: Simulations of Structure and Galaxy Formation

The current cosmogonic paradigm posits that structures in the universe (such as galaxies and clusters) originated from tiny density fluctuations generated around the time of the Big Bang and subsequently amplified by gravity. The precise origin of the primordial density fluctuations is uncertain. One possibility is that they were generated when the universe underwent a brief period of exponential expansion known as inflation, a tiny fraction of a second after the Big Bang. In the most successful model, the mass density of the universe is dominated by an exotic form of matter called cold dark matter which consists of elementary particles that do not make any contribution to the luminosity density of the universe. The predictions of this model have been set out extensively, using both analytic calculations and computer simulations, the latter having become increasingly important over the past 20 years. The build-up of structure in a cold dark matter universe is hierarchical: small-mass objects are the first to condense out from the expanding universe at early times (high REDSHIFT), whilst more massive objects (such as rich clusters) form later by repeated mergers of smaller objects. A remarkable confirmation of this framework was provided by the discovery, in 1992, of anisotropies in the temperature of the COSMIC MICROWAVE BACKGROUND radiation (CMB) by the DMR instrument aboard the COBE satellite. The amplitude of these anisotropies was within a factor of two of the extant theoretical predictions. Spurred on by the deep images of the sky produced by the Hubble Space Telescope, which show galaxies as they were when the universe was only about 10% of its present age, a great deal of effort is now being directed towards modeling the formation and evolution of galaxies within the setting of cosmological structure formation in the dark matter. The dissipative gas dynamical processes involved in GALAXY FORMATION make this a challenging task that can only be tackled using either idealized models or large amounts of supercomputer time.

The physics of galaxy formation

Studies of the dynamics of galaxies confirm a fundamental inference made 25 years ago from studies of the kinematics of stars within galaxies: on scales larger than galactic nuclei, the dominant physical interaction is gravity. Furthermore, it is now well established that the predominant source of gravity is dark matter, that is matter that does not emit detectable electromagnetic radiation. Dark matter is now routinely 'imaged' through the phenomenon of GRAVITATIONAL LENSING, the relativistic deflection of the light from distant galaxies as it passes near an intervening cluster of galaxies. Independently of the exact identity of the dark matter, the dominance of gravity leads directly to a general scheme for structure formation, first outlined by Landau and Lifshitz in the 1950s, rigorously developed by Peebles during the 1970s, and calculated in detail using computer simulations in the 1980s and 1990s. The key concept is the gravitational instability experienced by small matter overdensities in the expanding universe. Matter fluctuations present in the early universe grow in amplitude approximately as a power-law in time and eventually collapse to form self-gravitating objects.

The process of gravitational instability sets the scene for galaxy formation, the main physical ingredients of which are set out in figure 1. Although the precise details depend on the identity of the dark matter, under quite general conditions, galaxy formation is expected to proceed via a two-stage process originally outlined by White and Rees in 1978. First, gravitational instability acting on collisionless dark matter, results in the formation of self-gravitating dark matter haloes. Gas, initially well mixed with the dark matter, participates in this collapse, but it is heated by shocks to the thermal (or virial) temperature of the dark matter haloes. Second, the hot gas cools radiatively, on a time scale set by atomic physics, due to bremsstrahlung, recombination and collisionally excited line emission. The rate of cooling depends upon the density of electrons and atomic nuclei, and so it was most efficient at high redshift, when the universe was denser. (In practice, the heating and subsequent cooling of the gas may occur rapidly and chaotically, particularly in small galaxies.) Just prior to gravitational collapse, angular momentum is imparted on the generally aspherical perturbations by gravitational torques exerted by neighboring clumps, as proposed by Hoyle in 1948. Thus, the initial collapse generically results in the formation of a gaseous disk. Once the disk has become centrifugally supported, the material in it begins to fragment into stars by processes that are still poorly understood. In this simplified picture, the spheroidal components of galaxies form by mergers of disk galaxies which jumble up the stellar orbits, disrupting their organized configuration.

The scheme outlined above provides a natural explanation for why there are galaxies of two basic types: disks and spheroids. It also elegantly explains why there is a limit to the luminosity that galaxies attain at the present day. The most luminous galaxies form in the most massive dark matter haloes. These have only recently collapsed because, as is the case in most current models, the amplitude of mass fluctuations is smallest on large scales. At the high temperatures and low gas densities prevailing in recently formed structures, cooling is very inefficient and the gas has not had time to cool and fragment into stars over the lifetime of the universe. One problem of hierarchical structure formation models is that the number of low-mass haloes that form exceeds the number of faint galaxies seen in the local universe. However, a likely solution is that the feedback of energy into the cooled gas from early generations of stars will have acted as a self-regulating mechanism. This feedback prevents substantial star formation activity in shallow gravitational potential wells, thus causing most of the small-mass haloes to harbor extremely faint galaxies.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

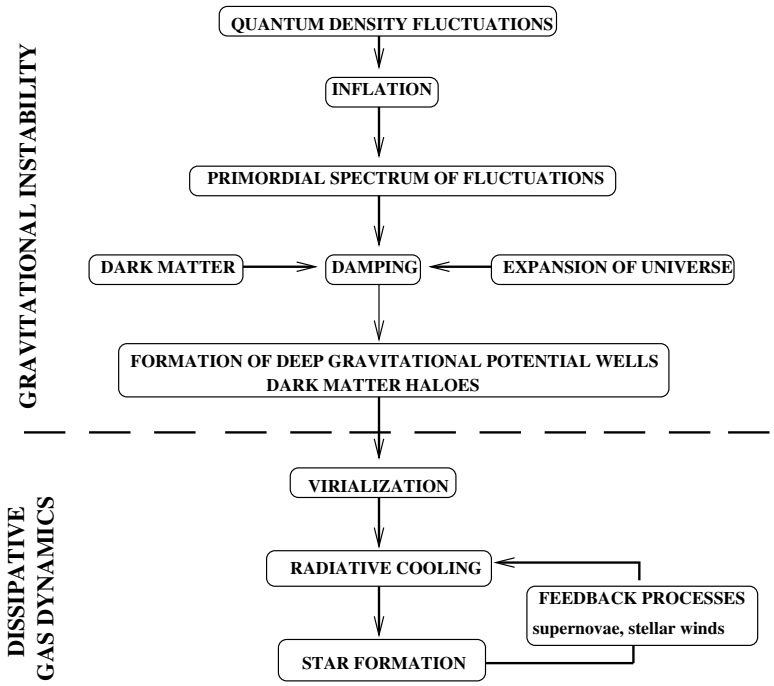


Figure 1. The ingredients of the standard model for the formation of galaxies and cosmological structure.

Cosmological structure formation

The processes of gravitational instability and collapse, gas cooling and star formation operate under quite general conditions. A quantitative theory of galaxy formation, however, requires that two key cosmological questions be addressed: (i) what is the origin of primordial mass fluctuations and (ii) what is the identity of the dark matter? Significant progress towards answering these questions was made in the early 1980s, as the result of a fruitful interaction between particle physics and cosmology.

The first influential idea of the 'New Cosmology' was proposed around 1980 by Alan Guth and extended by Andrei Linde. Searching for a solution to the 'magnetic monopole problem', Guth proposed that the universe had undergone a period of exponential expansion, or INFLATION, very soon after the Big Bang, triggered perhaps by the transition of a quantum field from a false to the true vacuum. Quantum fluctuations generated during this epoch would be swept across the event horizon and thus become established as classical ripples in the energy density of the universe. When they cross the horizon back again, their amplitude is independent of scale. Thus, the inflationary model predicts a scale-invariant power spectrum of primordial fluctuations, $P(k) \propto k$, with a Gaussian distribution of amplitudes. (More elaborate versions of the same idea can produce models whose power spectra have an exponent that differs slightly from unity.) The subsequent evolution of the fluctuations depends on the values of the cosmological density parameter, Ω_0, the Hubble constant, H_0, the COSMOLOGICAL CONSTANT, Λ_0, and the identity of the dark matter.

The second key idea from the 1980s concerns the identity of the dark matter. The abundance of the light elements (H, D, He, Li) produced during Big Bang nucleosynthesis agrees with astronomical data only if the present-day density of baryons is low enough for deuterium to form in at least the abundance measured in primitive gas clouds at high redshift. The baryon density required for this is about one order of magnitude smaller than the total mass density of the Universe inferred through a variety of tests. Thus, a fundamental conclusion is that the dark matter must consist of non-baryonic elementary particles.

Particle candidates for the dark matter are conveniently classified into 'hot' and 'cold' varieties, a nomenclature introduced by J R Bond around 1980. The prototype of a hot particle is a stable neutrino with a mass of the order of a few eV. (A single species of neutrino would give $\Omega = 1$ if its mass were approximately 30 eV and $H_0 = 50$ km s^{-1} Mpc^{-1}.) Examples of cold particles include the least massive stable supersymmetric particle, the neutralino, and the much lighter axion. Cold particles are often referred to as weakly interacting massive particles or WIMPs. There is a fundamental difference in the way in which galaxies are predicted to form in hot and cold dark matter cosmogonies, arising from the different damping mechanisms that affect the development of primordial fluctuations in the two cases. If the universe were dominated by massive neutrinos, then fluctuations below

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

some critical mass would be wiped out because the neutrinos move at relativistic speeds in the early universe and rapidly free-stream out of overdense regions. For a single neutrino of mass 30 eV, this critical scale corresponds to a fluctuation wavelength of

$$\lambda_c = \frac{2\pi}{k_c} = \frac{13}{\Omega h^2}\ \mathrm{Mpc} \qquad (1)$$

where we have parametrized Hubble's constant as $H_0 = 100h$ km s^{-1} Mpc^{-1}. The power spectrum decays exponentially for wavenumbers $k > k_c$.

In the case of cold dark matter, the damping of density fluctuations is much less severe. The free-streaming scale of cold dark matter is many orders of magnitude smaller than that of hot dark matter, and so this effect is not important. The most relevant damping process is the Mezaros effect, whereby oscillations in the radiation energy density stifle the growth of matter fluctuations. This situation persists until the energy density of matter dominates over that of the radiation, which occurs after a redshift of $1 + z_{eq} = 23\,900\Omega h^2$. The size of the horizon at this epoch is imprinted upon the cold dark matter power spectrum, marking the turnover of the primordial form, $P(k) \propto k$, to the damped fluctuation spectrum which asymptotes to $P(k) \propto k^{-3}$ at small scales.

In neutrino dominated models, the first structures that form are flat, pancake-like objects of mass comparable to that enclosed within the critical free-streaming scale, corresponding approximately to $10^{16} M_\odot$. These are objects of supercluster scale which must somehow subsequently fragment, in a 'top-down' fashion, in order for galaxies to form. Early computer N-body simulations of this process carried out by Frenk, White and Davis in 1983 showed that for a neutrino dominated universe not to exceed the level of clustering measured in the galaxy distribution today, galaxies would need to form at redshifts $z \lesssim 1$. Yet, it was already known at that time that quasars can have much higher redshifts than this and, today, we know that there is a large population of galaxies already established at $z = 3$–5. Thus, although rather appealing at first sight, neutrino dominated universes with Gaussian primordial fluctuations were soon abandoned.

The alternative, a cold dark matter universe, proved to be much more successful, as discussed, for example, in a series of papers in the 1980s based on N-body simulations, by Davis, Efstathiou, Frenk and White. The defining property of the fluctuation spectrum in a cold dark matter universe is that small-scale perturbations are preserved. Thus, subgalactic mass haloes are the first to collapse and separate out from the expansion of the universe. These haloes then grow, either gradually by accreting smaller clumps, or in big jumps by merging with other haloes of comparable size. The timetable for the formation of structure in a universe dominated by cold particles is hierarchical or 'bottom-up'—small objects form first, larger objects form later. The cold dark matter fluctuation power spectrum thus specifies completely the evolution of the merging hierarchy of dark matter haloes into which the baryons must fall in order to make the galaxies.

Currently, the most successful version of the cold dark matter model has around 30% of the critical density in cold matter and 70% in the form of a vacuum energy density or cosmological constant term, so that the universe has a spatially flat geometry. This model is in good agreement with a range of observational data: the amplitude of the angular power spectrum of temperature fluctuations in the cosmic microwave background (including the location of the first 'Doppler' peak), the abundance of clusters of galaxies ranked by their x-ray luminosity, the clustering of galaxies on large scales, and the expansion rate of the universe as deduced from the brightness of distant supernovae.

The recent detection of oscillations in neutrino flavour by the Super-Kamiokande experiment, which require the neutrino to have a non-zero mass, has rekindled interest in the possibility that neutrinos might, after all, make some contribution to the density of the universe. In order to avoid the problems faced by a universe dominated by hot dark matter outlined above, most of the mass in the universe must still consist of cold dark matter, with neutrinos providing only a minority contribution. Such 'mixed' dark matter models have not proved as successful at matching the data as the pure cold dark matter model. Another possibility for accommodating massive neutrinos is to replace inflation with a different mechanism for generating primordial density fluctuations. Tom Kibble suggested in the 1980s that TOPOLOGICAL DEFECTS IN COSMOLOGY might arise during phase transitions in scalar fields present in the early universe. Certain classes of defects, such as strings and textures, can act as seeds onto which matter gravitates, generating inhomogeneities in the mass density of the universe. These have the defining feature of being non-Gaussian. Unfortunately, modeling the formation of structure in defect models has turned out to be more complicated than in the Gaussian case because, amongst other subtleties, the properties of the defects themselves evolve with time. Defect models have not yet been explored with the same degree of rigor as the cold dark matter model. However, the current indications are that they have difficulties reproducing the spectrum of temperature anisotropies measured in the cosmic microwave background radiation. Thus, from a cosmological point of view, the oscillation data are best accommodated if the neutrino mass is very small (much less than 1 eV), so that neutrinos make a negligible contribution to the cosmic mass budget.

Computer simulations of galaxy formation

The remarkable developments of the past 15 years—the idea of cosmic inflation, the cold dark matter model, the discovery of ripples in the microwave background, and the observations of galaxies at high redshift—have laid down very solid foundations on which to build an understanding of galaxy formation. The 'initial conditions' for the evolution of the dominant dark matter component and its

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

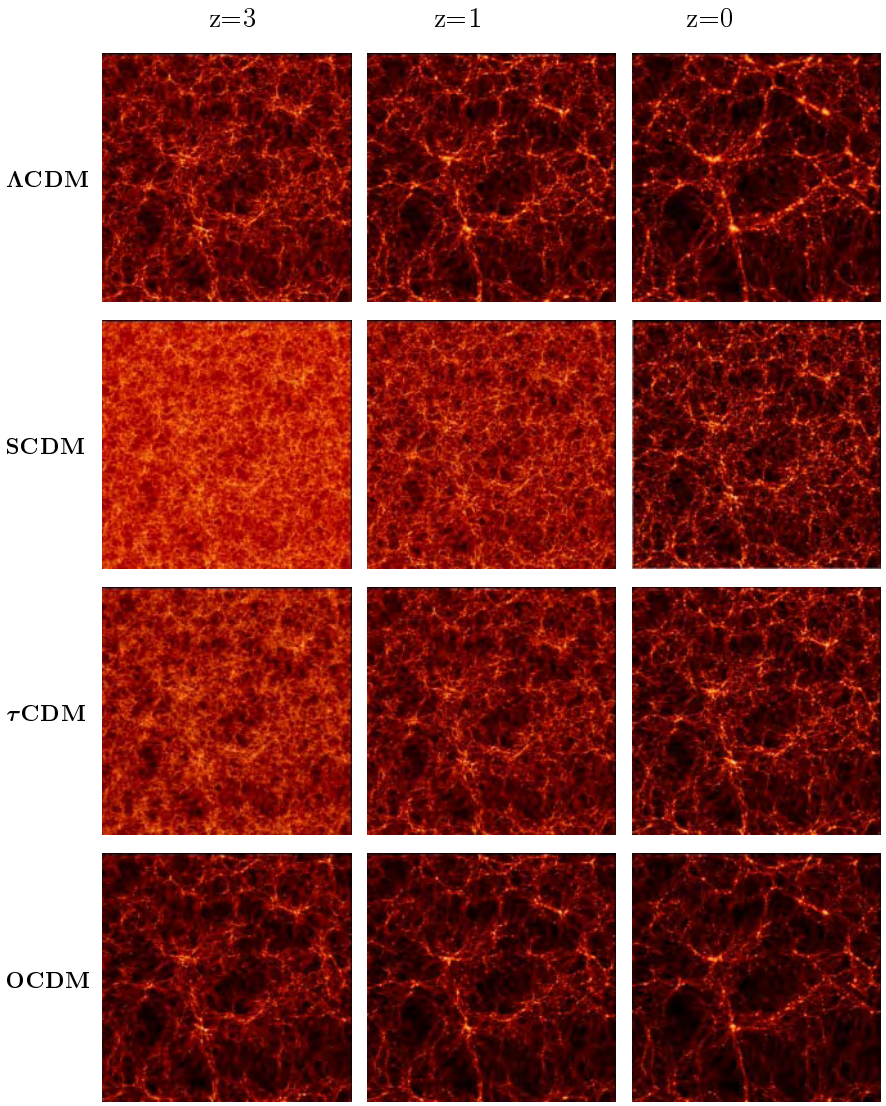


Figure 2. The formation of structure in N-body simulations of representative cosmological volumes of the universe. The intensity of the shading indicates the density of cold dark matter. Each row shows results from different versions of the cold dark matter model. The top row is a flat universe with $\Omega_0 = 0.3$ and a cosmological constant; the middle two rows are $\Omega_0 = 1$ universes, with different power spectra; the bottom row is an open universe with $\Omega_0 = 0.4$. The images, from left to right, show the evolution of structure in each model as a function of redshift. The present day corresponds to $z = 0$ while $z = 3$ corresponds to the epoch when the universe was approximately 15% of its current age. (Courtesy of the VIRGO Consortium for Cosmological N-body Simulations.)

subsequent gravitational evolution are well understood. Yet formulating an *ab initio* theory of galaxy formation and evolution over 10 billion years of cosmic history remains a tall order. The main stumbling block is our poor understanding of the behavior of cosmic gas—most probably a complex, dynamic, multiphase medium, of the physics of star formation, and of the feedback between the two mediated by winds from massive stars and supernovae explosions. The best way to address these issues is through extensive computer simulation and modeling.

The basis for present-day cosmological simulations is the N-body technique, which has been very successfully applied to modeling the evolution of collisionless dark matter. Using various computationally efficient methods, the computer solves the coupled equations of motion of N particles, interacting only through gravity, in the expanding universe. Progress over the past two decades has been driven mainly by dramatic improvements in the speed and memory of computers. By way of illustration, the early simulations of the cold dark matter cosmogony in 1985 employed 32 768 particles. In 1999, the largest simulations performed on massively parallel supercomputers (using essentially the same algorithms as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

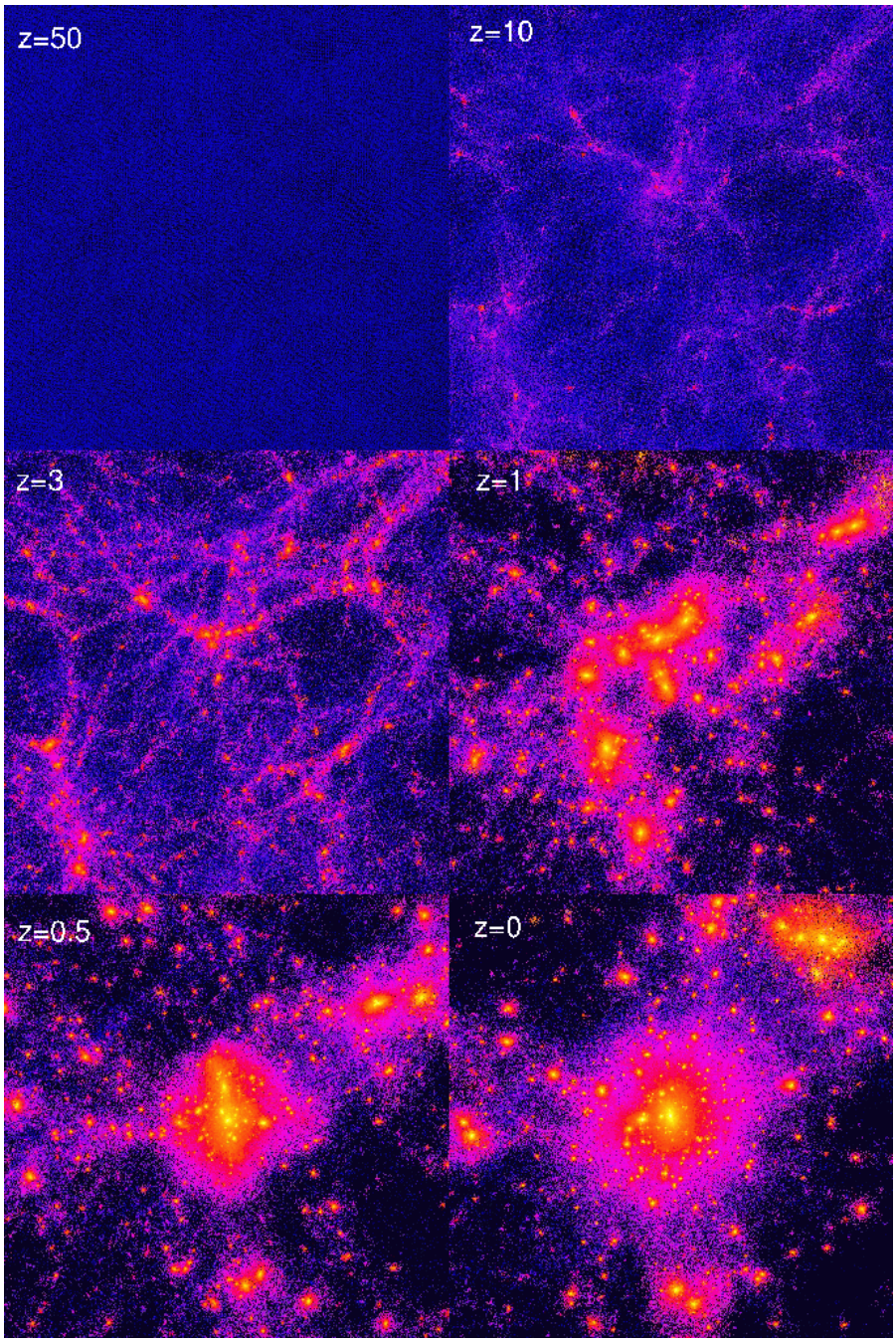


Figure 3. A high-resolution simulation of the formation of a single dark matter halo in a cold dark matter universe. The brighter colors indicate higher densities of dark matter. The sequence shows a series of snapshots of the evolution of the halo, at the redshifts indicated in the legend. The present-day halo displays a significant amount of substructure within its virial radius. (Courtesy of Ben Moore, Joachim Stadel, Tom Quinn and George Lake.) **This figure is reproduced as Color Plate 69.**

those of the 1980s) can follow the evolution of 10^9 particles.

Snapshots from simulations of representative, cosmological volumes are displayed in figure 2. This figure illustrates the evolution of structure in four versions of the cold dark matter model, differing only in the values of the cosmological parameters, Ω and Λ. At the present day, the dark matter is arranged in a complex network of voids, filaments and super-clusters (dubbed the 'Cosmic Web' by Bond, Kofman and Pogosyan). It is similar in all the sim-

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

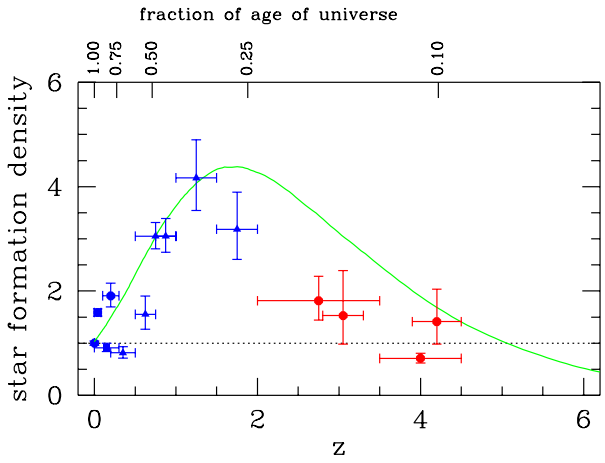


Figure 4. The star formation history of the universe. The curve shows the theoretical prediction for a cold dark matter universe, from the Durham semi-analytic model of galaxy formation. The points show a selection of recent observational determinations of the star formation rate density expressed in units of the present-day value.

ulations, partly because the amplitude of the initial fluctuation spectrum was adjusted so that all models would approximately reproduce the observed local abundance of hot, x-ray emitting clusters. (In addition, the initial fluctuations were set up with the same random phases in all the simulations, so that structures form at the same locations within each volume.) The development of structure proceeds at different rates in the different cosmological models. In those with a low value of Ω (top and bottom rows in figure 2), the formation of structure is essentially frozen at high redshift, whereas in the models with $\Omega = 1$, growth continues to the present. This merely reflects the fact that low-density universes expand more rapidly at late times, with the result that the expansion overwhelms the accretion of matter onto overdensities which lies at the root of their growth.

Simulations of sufficient size to resolve the internal structure and dynamics of dark matter haloes have recently become possible. The largest of these employ several million particles to model the formation of a single halo, revealing the existence of a rich substructure of lumps within the virial radius (figure 3). The simulations show that there is a remarkable uniformity in the density structure of dark matter haloes: over a wide range of scales, the spherically averaged halo density profile in follows a simple form proposed by Navarro, Frenk and White in 1996:

$$\rho(r) \propto \frac{1}{r/r_s(1+[r/r_s]^2)} \qquad (2)$$

where r_s is a scale length related to the density of the universe at the time when the halo formed. This functional form appears to be universal, independent of the choice of halo mass, power spectrum or cosmological parameters.

The N-body technique can be augmented with numerical hydrodynamic methods to model the evolution of gas subject to cooling and heating processes and coupled gravitationally to the dark matter. Two such methods are currently in use: Eulerian methods (including adaptive mesh refinements in some cases) and a Lagrangian scheme known as smoothed particle hydrodynamics. The two techniques have advantages and disadvantages, but so far the latter has proved to be the best suited to studying galaxy formation, because of the formidable dynamic range in thermodynamic quantities involved in the problem. At present, most progress has been achieved in modeling the physics of primitive gas clouds at high redshift, the so-called LYMAN ALPHA FOREST, detected observationally as absorption lines in the spectrum of distant quasars. The first simulations capable of following the evolution of gas to the present, with enough resolution to model the brightest galaxies, are now being carried out by various groups around the world. Currently, only a subset of the relevant gas physics, such as the shock heating of gas within dark matter haloes and its subsequent radiative cooling, are treated reliably.

A complementary technique for simulating galaxy formation *ab initio*, known as semi-analytic modeling, was developed in the 1990s by researchers at Durham and Munich. The main difference with the direct simulation approach is the abandonment of the ideal of solving the equations of hydrodynamics directly, in favor of a simple, spherically symmetric model in which the gas is assumed to have been fully shock-heated to the equilibrium temperature of each halo, from where its cooling and accretion onto the halo can be accurately calculated. This simplification speeds up the calculations enormously and has the added advantage of bypassing resolution considerations which are one of the main limiting factors of full hydrodynamic simulations. Phenomenological models of star formation, feedback and metal enrichment by supernovae are included in the semi-analytic program, through simple scaling relations. The semi-analytic machinery may be grafted onto haloes grown in a cosmological N-body simulation or onto haloes whose formation histories have been generated using a Monte Carlo approach. The models describe the entire star formation and merger history of the galaxy population. The free parameters of the model, which, perhaps surprisingly, are rather few in number, can be set by requiring a good match to a selection of properties of the local galaxy population, such as its luminosity distribution. This results in a fully specified model that provides an ideal tool for comparing the predictions of the cold dark matter theory with observations of the high redshift universe.

Confronting the high-redshift universe

The combination of direct simulations and semi-analytic modeling has revealed in detail the manner in which galaxies are expected to form in the cold dark matter model. The picture that emerges is one of gradual

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

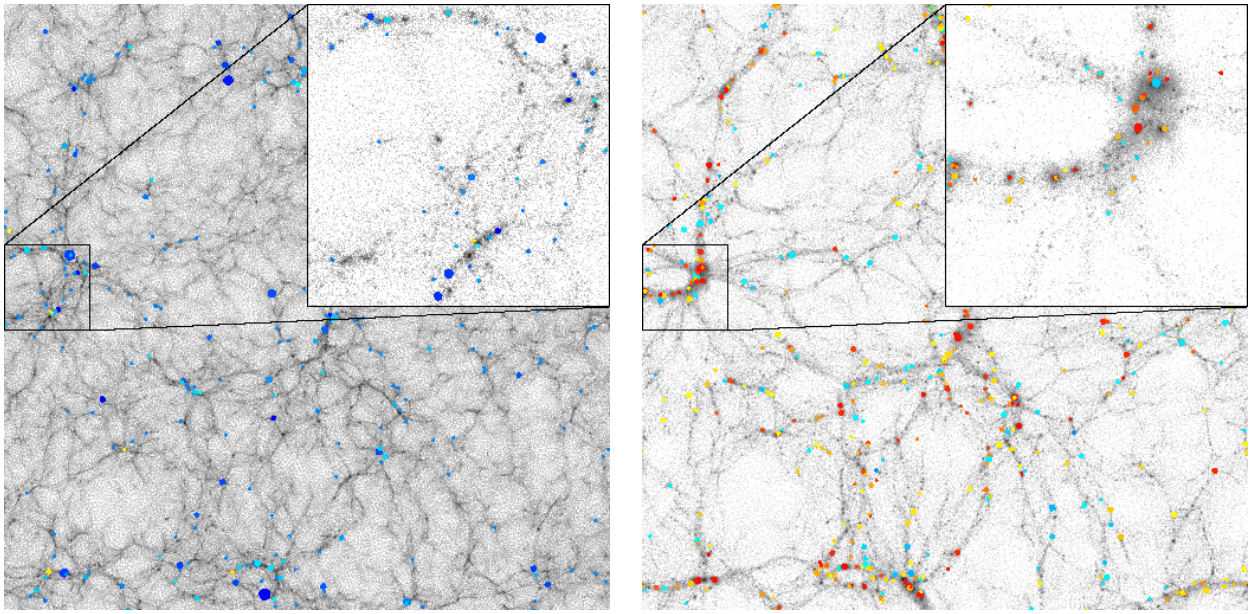


Figure 5. The evolution of clustering in the dark matter and galaxies. The left-hand panel shows an N-body simulation of a flat, low-density, cold dark matter universe with a cosmological constant, in a cube of comoving side $141h^{-1}$ Mpc, at $z = 3$. The right-hand panel shows the same simulation evolved to the present day. The gray scale indicates the density of the dark matter. Dark matter haloes in the simulation have been 'populated' with galaxies using a semi-analytic model of galaxy formation. The color of each spot reflects the color of each model galaxy, which is sensitive to the star formation rate. The size of the spot increases with the absolute luminosity of the galaxy. The inset shows the development of a cluster of galaxies. (Courtesy of Andrew Benson, Shaun Cole, CSF, CMB and Cedric Lacey.) **This figure is reproduced as Color Plates 70 and 71.**

evolution punctuated by major merging events that are accompanied by intense bursts of star formation and which trigger the transformation of disks into spheroids. Galaxy formation stutters into action around $z \sim 5$. Only a tiny fraction of the stars present today would have formed prior to that time. By $z \sim 3$, the epoch when galaxies isolated by the 'Lyman-break' technique[1] are observed, galaxy formation has started in earnest, even though only 10 per cent of the final population of stars has emerged. The midway point is not reached until about a redshift of 1–1.5, when the universe was approximately half of its present age. These theoretical predictions are shown in figure 4. Observationally, the star formation rate per unit volume can be inferred from the density of ultraviolet radiation, which is a measure of the number of high-mass, short-lived stars. Estimates of the star formation density based on data taken by ground-based telescopes and by the Hubble Space Telescope are shown as the points in figure 4. The major uncertainty in the interpretation of these data is the obscuring effect of dust, a modest amount of which has been allowed for in the models. But unless this effect turns out to be much stronger than anticipated, the theory and data in figure 4 suggest that we have now tracked most of the star formation activity, and the associated production of chemical elements, over the entire lifetime of the universe.

A second important prediction of the cold dark matter theory concerns the clustering properties of galaxies at high redshift. At the heart of the hierarchical clustering process lies the fact that galaxies tend to form first near high peaks of the density field because these are the first to collapse at any given epoch. This predilection for high-density regions is known as 'biased galaxy formation' (a term introduced by M Davis in 1985), because the distribution of galaxies offers a biased view of the underlying distribution of mass. An important consequence of biased galaxy formation is that bright galaxies tend to be born in a highly clustered state and remain so for long periods of time. The process of biased galaxy formation is illustrated in figure 5. The left-hand panel shows a snapshot of an N-body simulation of a cold dark matter universe at $z = 3$, whilst the right-hand panel shows the same simulation evolved to the present. The semi-analytic model of galaxy formation has been implemented in the dark matter haloes identified in the simulation at each redshift, in order to populate them with galaxies. Galaxies that are bright enough to be detected at $z = 3$ may be seen to form at the locations where the dark matter density (depicted by the gray scale) is highest. Observational confirmation of this clustering prediction came with the discovery that the population of galaxies at

[1] The so-called 'Lyman-break' galaxies are detected in passbands above the redshifted Lyman-limit at 912 Å and undetected in passbands below this limit; the strength of the Lyman-limit break is enhanced by absorption due to cold gas in the galaxy and in clouds along the line of sight.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

$z \sim 3$ identified by the Lyman-break technique is about as strongly clustered as bright galaxies are today. The relative clustering strengths of galaxies and dark matter evolve quite differently. The right-hand panel of figure 5 shows that the dark matter is much clumpier today than it was at $z = 3$. On the other hand, the clustering pattern of galaxies has hardly changed over this long period of time. Galaxies today are found in a wide range of environments and have a clustering amplitude similar to that of the dark matter. This was not the case at high redshift when bright galaxies were much more strongly clustered than the dark matter—in other words, when they were very strongly biased.

The next steps

The two areas of agreement between theory and data highlighted here—the cosmic history of star formation and the clustering of high-redshift galaxies—are particularly noteworthy because they concern fundamental aspects of the theory. The broad agreement between models and data suggests that the main ingredients of a coherent picture of galaxy formation are now in place. These ingredients are: primordial Gaussian density fluctuations; collisionless, non-baryonic dark matter; gravitational instability; and the growth of galaxies by hierarchical clustering. There are justifiably high expectations for the next decade. The number of 10 m class telescopes is proliferating: the first of the four European Space Observatory 'Very Large Telescopes' came into full operation in 1999, the same year in which the Gemini and Subaru telescopes first saw light. Other large telescopes are under construction in the USA and Spain. The middle of the first decade of the new century should also see the launch of NASA's Next Generation Space Telescope, that will search for galaxies out to a redshift of 10, and ESA's Planck Surveyor, that will map the microwave background radiation with unprecedented precision. Towards the end of the decade, the 'Large Millimeter Array', sponsored by a major international partnership, is scheduled to come into operation in the Chilean desert. It will search for galaxy formation at high redshift and examine star formation in nearby galaxies in the still relatively unexplored sub-millimeter waveband. Ultimately, the cornerstone upon which much of modern cosmology rests is the idea that the universe is dominated by non-baryonic dark matter. Experiments under way in the UK, Italy and the USA stand a good chance of detecting it, if it really exists, within the next few years. This will no doubt count as one of the most exciting discoveries in the history of science. For their part, theorists will not be standing still. Increased computing power, more efficient algorithms and, above all, a better understanding of the astrophysics of galaxy formation, are likely to result in a pretty good imitation, by computer, of the processes through which galaxies in our universe formed.

Bibliography

A pedagogical discussion (at an advanced level) of the physics of structure formation in the expanding universe and of the processes that play a role in galaxy formation may be found in:

Peacock J A 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

Numerical simulations are reviewed in:

Bertschinger E 1998 *Ann. Rev. Astron. Astrophys.* **36** 599–654

Some of the observational data discussed in this article are reviewed in:

Ellis R S 1997 *Ann. Rev. Astron. Astrophys.* **35** 389–443

Carlton M Baugh and Carlos S Frenk

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universe: Thermal History

HUBBLE's discovery of the expansion of the universe in 1929 revealed our beginning from a much smaller and much denser initial state (BIG BANG THEORY). Penzias and Wilson's discovery of the COSMIC MICROWAVE BACKGROUND radiation (CMBR) in 1964 implied further that just after creation the universe was a hot soup of the fundamental particles whose dynamics was controlled by the energy density of the CMBR.

The microwave radiation that A A PENZIAS and R W WILSON discovered is more precisely BLACK-BODY RADIATION characterized by a temperature of around 3 K. Because this black-body radiation fills space and its photons outnumber all other photons and known particles by a billion to one, it sets the average temperature of the universe. Of course, many hotter places exist in the cosmos—stars, planets and even the interstellar medium. As the universe expands its temperature decreases inversely as its linear size. While the temperature today seems unimportantly small, the fact that the universe still has a measurable temperature means that it was incredibly hot in the beginning.

The thermal history of the universe is thus the story of what happens when a tremendously hot and dense plasma expands and cools. Understanding the very earliest moments requires knowledge of the fundamental particles and their behavior under extreme conditions (high densities and temperatures). The connection between the inner space of elementary particle physics and the deep outer space of cosmology which plays such an important role in cosmology today was born with the discovery of the CMBR.

The known thermal history

The hot Big Bang cosmological model (see COSMOLOGY: STANDARD MODEL) provides an account of the universe from the hot soup of quarks, gluons, leptons and photons that existed before 10^{-5} s, when the transition from quarks and gluons to neutrons, protons and related particles occurred, until the present. During much of the time the universe was in a state of near thermal equilibrium; needless to say, the departures from thermal equilibrium are very important and make the universe an interesting place today! As the universe expanded, it cooled, $T \propto 1/R(t)$, where $R(t)$ is the cosmic scale factor which sets the size of the universe. As it cooled, layer upon layer of structure evolved, beginning with the neutrons and protons being produced from quarks culminating with the building of the largest structures seen today, the great walls of galaxies.

During its earliest moments, the temperature was the key to describing the state of the universe, because it sets the level of thermal particle energies. The thermal-energy scale $k_BT \sim 1$ MeV $(T/10^{10}$ K) determines when it is energetically favorable for the next layer of structure to form, and which particle species are present in great number. When k_BT was much greater than the rest mass of a particle, it and its antiparticle were easily produced in pairs and were present in the thermal soup in numbers comparable to photons. During the radiation era, the thermal-energy scale and age of the universe were related by: $k_BT \sim 1$ MeV $\sqrt{t\ (\text{in s})}$.

The soup of particles that existed at 10^{-5} s consisted of the up, down and strange quarks and their antiparticles; electrons, electron neutrinos, muons, muon neutrinos, tau neutrinos and their antiparticles; eight types of gluons and photons (the eight massless gluons are the carriers of the strong color force). The thermal-energy scale at this time was about 200 MeV ($T \sim 2 \times 10^{12}$ K), and the other quarks (charm, bottom and top) and the tau lepton were too heavy to be pair produced and were not present (any present would have annihilated). At even earlier times, when it was hotter, they (and possibly other particles) would have been present in great numbers too.

According to quantum chromodynamics (QCD), the theory of the strong color force, a phase transition from a quark–gluon plasma to hadronic matter occurred at a temperature of around 10^{12} K. Because of the increasing strength of the color force with distance, all particles with color (quarks and gluons) became confined in colorless quark triplets (neutrons, protons and other particles, known as baryons) and quark–antiquark pairs (pions, kaons and other particles, known as mesons). Collectively, the mesons and baryons are known as hadrons; the hadrons are the particles that experience the strong nuclear force.

At the end of the phase transition from quarks and gluons to hadronic matter, almost all the hadrons were too heavy to be pair produced. They could, of course, still annihilate, and as the temperature approached 10^{11} K annihilations had eliminated almost all of the hadrons. Were it not for the slight excess of baryons over antibaryons (whose origin is still a mystery, but see next section), nucleons (neutrons and protons) would have annihilated and disappeared too. The one additional nucleon for every billion or so antinucleons, left a few nucleons for every several billion photons without antinucleon partners to annihilate with. The primordial excess of baryons over antibaryons is responsible for all the ordinary matter that exists in the universe today.

At an age of about 1 s and temperature of about 10^{10} K, the primary constituents of the universe were photons, electron–positron pairs, and neutrinos and antineutrinos of all three species. There were also a few nucleons for every ten billion or so photons, about equally divided between neutrons and protons. Over the next 200 s or so, a sequence of nuclear reactions occurred out of thermal equilibrium and synthesized about 25% of the nucleons into ^{4}He. Trace amounts of D (a few parts in 10^5) and ^{3}He (about one part in 10^5) escaped being incorporated into ^{4}He and a tiny amount of ^{7}Li (a few parts in 10^{10}) was produced. The other 75% of the nucleons remained as free protons. This series of events is known as Big Bang NUCLEOSYNTHESIS (BBN). The rest of the periodic table was produced billions of years later by nuclear reactions in stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Big Bang nucleosynthesis occurred rapidly and at low density (around 10^{-2} g cm^{-3}), while the rest of the elements were cooked more slowly and at higher density (around 10^2 g cm^{-3} or higher) in stars and stellar explosions. This explains the great differences in the nuclear yields of the Big Bang and stellar nucleosynthesis. In particular, Coulomb barriers and the lack of stable nuclides of mass 5 and 8 prevented BBN from producing elements beyond ^{7}Li.

Big Bang nucleosynthesis provides the earliest test of the standard cosmology as well as a probe of conditions in the early universe. The fact that the pattern of abundances seen in the most primitive samples of the cosmos is consistent with its predictions is one of the experimental cornerstones of the standard cosmology. Further, the exact yields of the light elements, most especially deuterium, depend upon the baryon mass density; from recent measurements of the primeval deuterium abundance in high-redshift clouds of largely unprocessed hydrogen, we can infer that ordinary matter (i.e. matter comprised of neutrons and protons) today contributes between about 4% and 6% of the critical density. (The average density of the universe determines its curvature: a critical-density universe is spatially flat; a subcritical-density universe is open or negatively curved and a supercritical-density universe is closed or positively curved.) Because BBN 'weighs' all the ordinary matter at a simpler time, it provides the most accurate determination of the amount of ordinary matter. Today, baryons exist in many forms—bright stars, faint stars including white dwarfs and black holes, clouds of cold gas and of hot gas, and dust—and are more difficult to inventory. Thus far, only about one-third of the BBN-determined baryon abundance has been directly accounted for.

Light-element production also depends upon the ambient conditions in the universe, and Big Bang nucleosynthesis can thus also be used as a probe of the particle soup that existed then. For example, the existence of an additional neutrino species beyond the tau neutrino would have led to additional ^{4}He production, in contradiction to the observations. This argument against the existence of another neutrino species was put forth in the 1980s by David SCHRAMM and his collaborators and was confirmed by experiments at particle accelerators in the 1990s.

Two other important thermodynamical events took place during Big Bang nucleosynthesis. At a temperature of around 10^{10} K neutrinos and antineutrinos (all three species) ceased interacting with electron–positron pairs and decoupled from the electromagnetic plasma. Thereafter, they evolved independently of the rest of the universe, interacting only through gravity. Neutrino decoupling occurred because the decreasing particle energies and densities made neutrino interactions with other particles increasingly infrequent. When the temperature was around 10^9 K, essentially all of the electrons and positrons disappeared as pairs destroyed by annihilations were no longer replenished by thermal pair creation. The slight excess of electrons over positrons preserved the few electrons per ten billion photons required to balance the charge of the protons. The electron–positron annihilations raised the number of photons in the universe by a factor of 11/4 and heated the photons slightly relative to the neutrinos; thereafter $T_\nu = (4/11)^{1/3}T$.

The radiation era ended when the universe was around 40 000 years old and the temperature was about 10 000 K. At this epoch, called matter-radiation equality, the energy density contributed by matter (both baryons and exotic dark matter, more below) and that by relativistic particles (photons and three neutrino species) were equal. Thereafter, the matter density would exceed that of relativistic particles, growing in proportion to the linear size of the universe. (The matter density decreases as $1/R(t)^3$ due to the volume dilution effect of the expansion; the energy density of the CMBR decreases as $1/R(t)^4$, with the extra factor of $1/R$ arising because photon energies are redshifted with the expansion.) Today, the energy density of matter is about a factor of 4000 times larger than that of the energy density of photons and neutrinos.

The dawning of the matter era marked the beginning of the formation of large-scale structure in the universe (see also UNIVERSE: SIMULATIONS OF STRUCTURE AND GALAXY FORMATION). The small inhomogeneities in the distribution of the exotic matter that existed (spatial variations in the mass density at the level of about one part in 10^5) began growing through the attractive force of gravity; prior to matter–radiation equality the universe was expanding too fast for this to occur. Their tight coupling to photons prevented baryons from participating in this growth.

Shortly after the radiation era ended, at a time of around 400 000 years and a temperature of around 3000 K, two related and very significant events involving the radiation took place. The first was the transition from ionized matter to neutral matter (called 'recombination', which is paradoxical since neutral matter had not previously existed). As the temperature dropped below 3000 K neutral matter became thermodynamical favored, and all but a few ions combined with the free electrons to form neutral atoms (a residual ionization fraction of around 10^{-4} persisted thereafter because ions and electrons became too rare to find one another to combine to form atoms). When the universe became neutral, its opacity dropped precipitously (free electrons efficiently scatter light, neutral atoms do not), and matter and radiation decoupled. Photons streamed freely and have not scattered since; this important event is referred to as last scattering. Once baryons decoupled, they were rapidly pulled into the cosmic structures being formed by the gravity of the exotic dark matter.

The black-body character of the radiation, established by the hot, dense conditions in the early universe, was preserved by the expansion of the universe (the deep mathematical reason involves the conformal invariance of Maxwell's equations and the conformal nature of the expansion), albeit with a decreasing temperature, $T \propto 1/R(t)$). Today, this black-body radiation, which at last

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

scattering resembled the light emitted by the Sun today, has been redshifted to the microwave part of the spectrum. In 1996, the far infrared absolute spectrophotometer (FIRAS) instrument on the Cosmic Background Explorer (COBE) satellite made the most precise measurement of its temperature, $T = 2.7277 \pm 0.002$ K, and showed that any deviations from a perfect black-body spectrum are smaller than 0.005%. Because there is no other viable mechanism for producing such perfect black-body radiation, the spectrum of the CMBR is one of the experimental pillars of the hot Big Bang cosmology.

The radiation in the CMBR has not scattered since the universe was 400 000 years old, and so it provides a snapshot of the universe at a simpler time, when matter was still nearly uniformly distributed and stars, galaxies and clusters of galaxies did not exist. The variations in the intensity (or temperature) of the CMBR across the sky today map the two-dimensional distribution of matter at this time because variations in the mass density produce temperature variations of the same size. Thus, the temperature variations of a few parts in 10^5 measured by the differential microwave radiometer (DMR) on the COBE satellite and other balloon and ground-based experiments imply the existence of variations in the matter density of approximately the same size. This level of inhomogeneity is just what is needed to produce the large-scale structure seen today—provided that the bulk of the matter is exotic dark matter and not baryons (more below). The variations (or anisotropy) of the CMBR also encode a wealth of information about the early universe and how large-scale structure formed. Higher precision and higher angular-resolution measurements will be made by future experiments including NASA's MAP satellite (scheduled for launch in late 2000) and ESA's PLANCK SURVEYOR satellite (scheduled for launch in 2007).

Three cosmic seas of thermal neutrinos should be with us today. Just as with the CMBR, the expansion of the universe maintained their thermal (Fermi–Dirac) distributions with a temperature that has decreased inversely with the cosmic scale factor since they decoupled. Because neutrinos did not share in the energy release from the electron–positron annihilations, the temperature of the neutrino seas is predicted to be smaller than that of the photons, $T_\nu = 1.947$ K. If they can be detected, these neutrinos will reveal the universe as it was about 1 s after the beginning. However, because low-energy neutrinos interact extremely weakly with ordinary matter their detection presents one of the greatest challenges in all of science.

Beyond the standard cosmology: the very early universe

The earliest history of the universe (before 10^{-5} s) is still a mystery, but is under intense study. The motivation is twofold: the hope that events which took place may explain some of the most pressing cosmological puzzles. For example, the reason for the small excess of matter over antimatter, the explanation for the regularity seen in the universe, and the origin of the primeval density inhomogeneities. The second motivation is the possibility that the early universe can be used to probe fundamental physics more deeply than particle accelerators and other Earth-based experiments. At the moment, the discussion of the universe at times earlier than 10^{-5} s is speculative, both because of uncertainties about the microphysics needed to describe these early times and the absence of cosmological tests like Big Bang nucleosynthesis. In any case, the physics and the cosmology are of sufficient interest to merit the discussion of the possibilities.

On fairly firm ground, the discussion of the thermal history of the universe can be extended back to around 10^{-11} s when the temperature was about 10^{15} K. This was sufficiently hot that the thermal-energy scale exceeded the rest masses of all known particles. At this time the thermal soup should have included all the quark and lepton species, gluons, photons and the $W^{\pm}$ and Z^0 bosons, all in roughly equal abundance.

The state of the universe earlier than this is much less certain. The prevailing belief is that a phase transition occurred and restored the full symmetry of the $SU(2) \otimes U(1)$ gauge theory of the electroweak interactions. ($SU(2) \otimes U(1)$ is the mathematical symmetry that underlies the unified theory of the electromagnetic and weak interactions.) At low temperatures the symmetry between the weak and the electromagnetic interactions is not manifest: the electromagnetic interaction has long range, while the weak interaction has very short range because the mediators of the weak force, the $W^{\pm}$ and Z^0 bosons, are very massive. The symmetry is said to be spontaneously broken, by the Higgs mechanism. When the symmetry is restored, all of the force mediators become massless. If the Higgs mechanism is correct, then there is at least one more spin-zero particle species, the Higgs boson, whose rest mass is greater than about 100 GeV and probably less than 300 GeV. Its discovery would be a striking confirmation of spontaneous symmetry breaking, and thus the Higgs is at the top of the 'most wanted' list at all accelerator laboratories.

As successful as the electroweak theory is, it provides only a partial unification of the forces, leaving out the strong color force and gravity. One possibility is there are other levels of symmetry breaking and symmetry breaking phase transitions. The simplest idea, grand unification, unifies the color force with the electroweak force. Estimates for the temperature at which the grand unification phase transition might take place are even more uncertain, but are around 10^{29} K, corresponding to a time of about 10^{-39} s.

Another interesting feature of symmetry breaking is the possibility that the phase transition did not occur smoothly and that 'defects' are created. (Such defects are known to be produced in phase transitions in condensed matter systems: vortices and magnetic flux tubes.) These so-called topological defects are concentrations of energy: point-like magnetic monopoles, one-dimensional cosmic string and two-dimensional domain walls. The kinds of

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

defects that can be produced depend upon the symmetry breaking pattern. Thus far, the cosmology of topological defects has not been promising: monopoles should have been grossly overproduced, domain walls have disastrous cosmological consequences, and cosmic string, once thought to be a possible seed for the formation of structure in the universe, predicts a pattern of CMBR anisotropy that is inconsistent with the measurements.

A much more promising idea arising from the consideration of cosmological phase transitions is INFLATION. Inflation refers to an enormous burst of expansion which might have taken place very early on (probably earlier than about 10^{-32} s). Because of its potential to explain a number of the most fundamental and most puzzling features of the universe, inflation has been the dominant theoretical idea in cosmology over the past 15 years. It can account for the smoothness of the universe, the origin of the primeval matter inhomogeneity, the heat of the Big Bang, and the nature of the Big Bang itself. Originally inflation was thought to be driven by the latent heat (or false-vacuum energy) associated with a first-order phase transition. Most models of inflation no longer involve a phase transition, but instead rely upon the potential energy of a fundamental scalar field.

Besides inflation, the most compelling ideas of early-universe cosmology are particle dark matter and BARYOGENESIS. It has been known for more than 50 years that most of the matter that holds galaxies and clusters of galaxies together is not in the form of visible stars but is 'dark' (i.e. does not emit or absorb detectable radiation of any form; see DARK MATTER: ITS NATURE). For at least a decade it has also been known that the total amount of dark matter exceeds by more than a factor of three the amount of matter in the form of baryons as determined from Big Bang nucleosynthesis. Further, with the level of inhomogeneity measured by COBE, the observed large-scale structure can only form if there is exotic dark matter. These facts are strong circumstantial evidence for a new form of matter in the universe.

The most promising candidates for the dark matter are elementary particles that were present in copious numbers in the thermal soup early on and which failed to annihilate away because of the weakness of their interactions. Of all the possibilities considered, the three most attractive are neutrinos (if they have a small mass) and axions or neutralinos (if they exist). (The axion and neutralino are as yet hypothetical particles predicted to exist by theories that unify the particles and forces of nature.)

Baryogenesis is a higher-level analog of Big Bang nucleosynthesis: BBN explains how baryons come together to make nuclei and baryogenesis hopes to explain the origin of the excess of quarks over antiquarks that leads to the existence of ordinary matter. The idea is that particle interactions that violate matter–antimatter symmetry and the conservation of baryon number and which occurred out of thermal equilibrium produced a slight excess of quarks over antiquarks. When the quarks formed into neutrons and protons, this led to the excess of baryons over antibaryons needed to ensure the existence of the ordinary matter.

The weak interactions violate matter–antimatter symmetry at a small level and are also predicted to violate baryon-number conservation through subtle quantum effects. Moreover, baryon-number non-conservation is a generic prediction of grand unified theories. While the details are not currently understood, nor is there any experimental evidence for the non-conservation of baryon number, nonetheless baryogenesis is a promising framework for understanding how the crucial excess of matter over antimatter arose.

To date, superstring theory has been the most successful approach to the unification of gravity with the other forces. Superstring theory makes two generic predictions relevant for cosmology. First, the existence of a new symmetry of nature that relates fermions and bosons (supersymmetry), and second, the likely existence of additional spatial dimensions. Because the known particles of nature cannot be paired off as fermion–boson partners, supersymmetry requires the doubling of the number of fundamental particles. The superpartners, as they are called, are predicted to have rest energies of the order of 100 GeV. If correct, this implies a doubling of the number of particles in the primordial soup only occurs at temperatures greater than around 10^{15} K. The lightest superpartner, usually the neutralino, is stable or very long-lived and has a rest energy of order 100 GeV. As mentioned above, it is a prime candidate for particle dark matter.

If there are extra spatial dimensions they must be 'small' enough to have escaped detection or be otherwise hidden from us. Small here refers to their being curled up like the circular dimension of a straw. While many versions of superstring theory predict that the extra dimensions are exceedingly tiny—10^{-34} cm or smaller—some versions suggest that they might be as large as a millimeter in extent! The cosmological implications of extra dimensions are not well understood and raise a host of additional cosmological questions, for example the explanation for the size discrepancy between the familiar three spatial dimensions and the extra spatial dimensions.

Even if superstring theory does not prove successful in unifying gravity with the other forces and in providing a quantum description of gravity, 'interesting physics' should have occurred at times earlier than 10^{-44} s and temperatures greater than 10^{32} K. This is the Planck era, the epoch when quantum gravitational effects should have been extremely important and the classical description of gravity given by general relativity should have been inapplicable. It could be that the universe achieved a limiting temperature due to the exponentially growing number of particle species (also a prediction of string theory), or that space-time dissolves into a foam. Even by the standards of early universe cosmology, speculations about the Planck era are extraordinarily speculative.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bibliography

Bernstein J 1988 *Kinetic Theory in the Expanding Universe* (Cambridge: Cambridge University Press)
Kolb E W and Turner M S 1990 *The Early Universe* (Reading, MA: Addison-Wesley)
Schramm D N and Turner M S 1990 *Rev. Mod. Phys.* **70** 303

Michael S Turner

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Universität Radioastronomisches Institut, Bonn

An institute of the University of Bonn, Germany. Carries out research at infrared, radio and submillimeter wavelengths. Particular research interests are the interstellar medium and intergalactic medium, especially galactic halos, and dwarf galaxies.

The institute was responsible for the discovery of the x-ray halo around the Milky Way in 1991.

Although it has recently disposed of a 25 m radio telescope in the Eiffel Mountains, the institute has a cooperative agreement with the Institute of Physics, Cologne, to use KOSMA, a 3 m radio telescope on Gornergrat in Switzerland.

For further information see http://www.astro.uni-bonn.de/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

University Observatory, Ludwig-Maximilians-Universität

The University Observatory of Ludwig-Maximilians-Universität was founded in 1816. Astronomers who worked or graduated at the Munich Observatory include: Fraunhofer, Soldner, Lamont, Seeliger and Karl Schwarzschild. At present four professors and ten staff astronomers work here. Funding comes from the Bavarian Government, the German Science Foundation, and other German and European research programs.

Facilities include the Wendelstein Observatory in the Bavarian Alps with a 0.8 m telescope. The Observatory is also a partner in the Hobby–Eberly Telescope Project in Texas.

The Observatory's mission is to teach astrophysics at the Ludwig-Maximilians-Universität in Munich, to research astrophysics and instrument development for the Very Large Telescope (ESO), the Hobby–Eberly Telescope (Texas) and the Wendelstein Observatory.

Research specialties are stellar atmospheres and winds; binary stars; structure, formation and chemical evolution of galaxies; galaxy clusters; dark matter and gravitational lensing; plasma astrophysics.

For further information see
http://www.usm.uni-muenchen.de.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

University of Crete Observatory

Based in Heraklion. Scientific activity is evenly distributed between theory and observation. Observational research has been undertaken using a small observatory on Mount Ida (Skinakas), and international facilities such as Arecibo and Effelsberg (radio), the ATO-Australia (optical) and Rosat (x-ray). The observational activities cover a wide range of objects including comets, supernova remnants, planetary nebulae, nearby galaxies, the Magellanic Clouds, dwarf galaxies, radio pulsars and the Earth's ionosphere and magnetosphere.

For further information see http://www.physics.uch.gr/english/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

University of Hawaii Institute for Astronomy

The University of Hawaii Institute for Astronomy was founded in 1967 to manage Haleakala Observatory on Maui, and later, Mauna Kea Observatory, and to carry out its own program of fundamental research. Institute staff also make extensive use of spacecraft for astronomical research and planetary exploration. The Institute's main offices are located in Manoa valley, just off the University's main campus on Oahu. A research institute within the University of Hawaii (UH), the IfA has a total staff of approximately 200, including about 50 faculty staff.

The Institute is one of the world's leading astronomical research centers. Its broad-based research program includes studies of the Sun, planets and stars, as well as interstellar matter, galaxies and cosmology. Most IfA astronomers use the giant telescopes atop Mauna Kea and Haleakala. They also use space observatories, such as the Hubble Space Telescope and the Chandra X-ray Observatory, to make observations that cannot be made from the ground. In addition to doing research, some astronomers design and build new instruments to measure and analyse the radiation collected by the telescopes. Other IfA scientists develop theories which explain the observations made by their colleagues.

During the last 30 years, the State of Hawaii has become the most sought-after location in the world for the construction of large ground-based telescopes. The focal points for this construction are Mauna Kea (4200 m) on the island of Hawaii and Haleakala (3000 m) on Maui. The remarkable clarity, dryness and stillness of the air above these isolated high-altitude sites led to the commissioning by the University of Hawaii of the Mees Solar Observatory on Haleakala in 1963 and the UH 2.2 m telescope on Mauna Kea in 1970.

Haleakala is lower in altitude than Mauna Kea, yet when compared to other astronomical sites outside Hawaii, Haleakala is an excellent location for optical and infrared observations of objects in Earth orbit, solar observations, laser ranging and laser guide star experiments, and the whole range of night-time astronomical observations. The IfA's Mees Solar Observatory on Haleakala is central to the Institute's solar astronomy program. It is set in a naturally fine location for studies of the extremely faint radiation emitted by the solar corona and chromosphere. At sea level, these outer atmospheric layers of the Sun are normally obscured by light scattered by small airborne particles, such as dust or pollen. Mees is a perfect use of Haleakala, which means 'The House of the Sun'.

The Lunar Ranging Experiment (LURE Observatory) operated by IfA for NASA, measures the motions of the Pacific Basin and interior of Earth by bouncing laser beams off reflectors on satellites. It has also been used to bounce laser beams off reflectors left on the Moon by the Apollo astronauts.

Institute for Astronomy staff will receive a share of the observing time on the University of Tokyo 2 m 'Magnum' telescope and on the US Air Force 3.7 m Advanced Electro-Optical Systems (AEOS) Telescope, both of which are located at the Haleakala Observatory. Rocketdyne Technical Services is contracted by the US Air Force to carryout satellite ranging, surveillance and other developmental activities at the Maui Space Surveillance Site on land leased from UH at the Haleakala Observatory. Hawaiian students will share the use of the 2 m Faulkes telescope with students in the UK when it is installed in Haleakala Observatory in 2001.

The summit of Mauna Kea on the island of Hawaii houses the world's largest astronomical observatory, with telescopes operated by astronomers from eleven countries. The combined light-gathering power of the telescopes on Mauna Kea is fifteen times greater than that of the Palomar telescope in California—for many decades the world's largest—and fifty times greater than that of the Hubble Space Telescope.

In addition to developing its own research program, the IfA has provided the scientific impetus for the development of Mauna Kea into the world's premier site for ground-based astronomical observatories. More major telescopes are located on Mauna Kea than on any other single mountain peak, and Mauna Kea is widely recognized as offering better observing conditions for optical, infrared and mm/sub-millimeter measurements than any other developed site.

The success of the 2.2 m telescope led to proposals for the construction of several other telescopes on Mauna Kea by national and international organizations. By the end of the 1970s, there were three new 4 m class telescopes—the NASA INFRARED TELESCOPE FACILITY (IRTF), the UNITED KINGDOM INFRARED TELESCOPE (UKIRT) and the CANADA–FRANCE–HAWAII TELESCOPE (CFHT). Two of these telescopes are specifically designed to collect infrared radiation; the dryness of the atmosphere above Mauna Kea is particularly advantageous at these wavelengths.

The characteristics that make Mauna Kea a pre-eminent site for optical and infrared astronomy also make it an excellent site for measuring short-wavelength radio waves. In 1987, two submillimeter-wave observatories were completed just below the summit. The California Institute of Technology and the US National Science Foundation supplied the expertise and funds for the CALTEC SUBMILLIMETER OBSERVATORY (CSO). The JAMES CLERK MAXWELL TELESCOPE (JCMT) belongs to a consortium of the United Kingdom, Canada and the Netherlands. More recently, the Smithsonian Institution, in collaboration with astronomers from Taiwan, has built an array consisting of eight 6 m submillimeter antennas designed to work together as a single telescope.

The 1990s saw the construction of a new series of giant optical/infrared telescopes on Mauna Kea. The twin telescopes of the W M KECK OBSERVATORY are the largest optical/infrared telescopes in the world. Their 10 m diameter mirrors each consist of 36 hexagons whose positions are controlled by computer. Japan and an international consortium led by the United States

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

have built the Subaru and Gemini Northern Telescopes, respectively. Both of these telescopes contain thin meniscus 8 m primary mirrors, and are designed for both optical and infrared astronomy.

At a lower altitude, on the southern flank of Mauna Kea, is the Hawaii antenna of the Very Long Baseline Array, which is part of a 5000 mile wide system of ten 25 m radio dishes that work together as the world's largest dedicated full-time astronomical instrument.

As part of the agreement between these organizations and the University of Hawaii, astronomers at UH are entitled to 10–15% of the observing time on each non-UH telescope on Mauna Kea, in addition to full use of the UH 2.2 m and 0.6 m telescopes.

For further information see http://www.ifa.hawaii.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

University of Tokyo, Institute of Astronomy

Founded in 1987, the Institute of Astronomy, University of Tokyo, is located at Ohsawa, Mitaka, Japan, 30 km west of central Tokyo. Its objectives are research and education in observational astronomy. It has a staff of 3 professors; 5 associate professors; 8 research associates, 2 technical support staff, 2 administrative officers and several part-time staff (secretary, catering etc). Its major facilities include Kiso Observatory (105 cm Schmidt telescope); two 60 cm submillimeter telescopes; 2 m IR telescope (in collaboration with the Physics Department). Research specialities: galactic and extragalactic radio astronomy; galactic center; extragalactic optical/IR observations; observational cosmology/formation and evolution of galaxies; infrared stellar physics.

The Institute is supported by an annual fund from the University of Tokyo (Ministry of Education) and science research aids from the Ministry of Education.

Its member organization (in joint ventures) is the University of Tokyo, Graduate School of Sciences

For further information see
http://www.ioa.s.u-tokyo.ac.jp.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Unsöld, Albrecht (1905–95)

German astrophysicist, became professor at the University of Kiel, where he studied stellar atmospheres and the way that spectral lines were formed and shaped—the effects of abundances, radiation damping, Doppler shifts, electric fields and atomic collisions. He analysed the spectrum of the star Tau Scorpii, which he obtained on a visit to the Yerkes and McDonald observatories. This was the first detailed analysis of a star other than the Sun and yielded the physical conditions in the atmosphere of the star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Uppsala Astronomical Observatory

The Uppsala Astronomical Observatory (UAO) is a department of Uppsala University (UU) with research, undergraduate, graduate and public education as its main objectives.

UU was founded in 1477, and preserved lecture notes show that astronomy was taught in the 1480s. Among well-known professors at UAO are Anders Celsius (1730–44), Anders Angström (1842–58), Gunnar Malmquist (1939–59), Erik Holmberg (1959–75) and Bengt Westerlund (1975–87).

Current research focuses mainly on galaxies, stellar astrophysics and solar system minor bodies and is characterized by the combination of theory and observations. Nowadays the observational material comes from large international facilities like ESO (EUROPEAN SOUTHERN OBSERVATORY) or space observatories like the HST (Hubble Space Telescope).

At UAO, about 70 km north of Stockholm, there is a 36 + 33 cm double refractor from 1893 currently used mainly for public shows and at Kvistaberg, 50 km south of Uppsala, there is a 135 cm Schmidt telescope from 1964.

For further information see
http://www.astro.uu.se.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Uranus

With an equatorial diameter of 51 118 km (at the 1 bar atmospheric pressure level), Uranus is the third largest of the planets in the solar system, slightly larger than the more distant NEPTUNE. However, because of its lower density (1.27 g cm^{-3} compared with 1.64 g cm^{-3} for Neptune), Uranus ranks fourth among solar system planets in mass at 14.54 terrestrial masses. Its characteristic blue–green appearance is due to a layer of clouds of methane ice in its upper atmosphere that is not present in the atmospheres of JUPITER and SATURN. This cloud layer and the overlying methane gas preferentially absorb the red portion of sunlight, resulting in the distinctive color. Prior to the flyby of Uranus by VOYAGER 2 in 1989, little was known about the rotational or other physical characteristics of Uranus (given in table 1).

Uranus has a ring system and many satellites. The rings were first detected from observations of stellar occultations as the narrow rings passed in front of distant stars, causing the observed light to blink out temporarily as viewed from Earth. Nine narrow rings were detected in this manner. Voyager 2 detected a tenth narrow ring, a broad diffuse ring closer to the planet, and tenuous dust rings scattered among the rest (see URANUS: RINGS).

Prior to 1986, only five moons of Uranus were known. Voyager 2 found an additional 10 (plus one identified in 1999 from Voyager images). Earth-based observations since 1986 have resulted in the finding of another five confirmed or probable moonlets much more distant from the planet. The satellites, known and suspected, are listed in table 3.

Uranus is the first of the non-naked-eye planets to be discovered. Observations of the closer planets (Mercury through Saturn) date back to antiquity, but Uranus was discovered by William Herschel in March of 1781 using his home-made telescope in the back yard of his home in Bath, England (see HERSCHEL FAMILY). Technically, at sixth magnitude, Uranus is marginally visible with the unaided eye. However, it was indistinguishable from background stars until Herschel's telescopic observations.

Table 1. *Orbital and physical characteristics of Uranus.*

Semimajor axis of the orbit	19.191 AU
Eccentricity	0.046
Inclination over the ecliptic	0.77°
Sidereal period	83.7474 yr
Mean orbital velocity	6.83 km s^{-1}
Equatorial diameter	51 118 km
Equatorial diameter relative to Earth	4.007
Polar diameter	49 946 km
Flattening	0.023
Mass relative to Earth	14.54
Mean density	1.27 g cm^{-3}
Surface gravity (at P=1 bar)	8.86 m s^{-2}
Surface gravity relative to Earth	0.906
Escape velocity	21.28 km s^{-1}
Sidereal rotation period	17.24 h
Inclination of equator to orbital plane	97.92¬∞
Albedo (visual geometric)	0.57
Albedo (bolometric Bond)	0.30
Main atmospheric composition	H_2 (83%), He (15%), CH_4 (2%)

The early observations

In the 200+ yr of observation since its discovery, there have been monumental changes in our understanding of the planets in general and of Uranus in particular. The first major change was a recognition that the solar system was much larger than previously supposed. Uranus was more than twice the distance of Saturn from the Sun; its discovery gave great impetus to searches for even more distant planets. Perturbations to the orbit of Uranus led to independent predictions by mathematicians John Couch Adams in England and Urbain Jean Joseph Le Verrier in France. On the basis of Le Verrier's predictions, Johann Gottfried Galle and his student Heinrich Louis D'Arrest found Neptune on 23 September 1846, the same day they received Le Verrier's prediction.

In the meantime, Herschel continued his observations of Uranus with a telescope, again of his own making and far superior to those available to other astronomers of his day. OBERON and TITANIA were discovered by him in 1787, less than 6 yr after the discovery of the planet. For more than 10 yr, no other astronomers were even able to confirm his findings. Herschel was convinced that other satellites must also exist, but in 40 yr of searching he found none, although some historians claim that he may have been the first to see UMBRIEL (four times between 1790 and 1801). Credit for the discovery of Umbriel and ARIEL is generally given to William Lassell, who reported their sighting in 1851 to the Royal Astronomical Society. MIRANDA was discovered by Gerard Peter Kuiper in 1948, using the 82 in reflector of the McDonald Observatory in Texas. The same instrument was used by Kuiper and co-worker D L Harris to discover Neptune's Nereid in 1949.

The orbits of the Urananian satellites provided the first evidence of the unusual tilt of the equator of Uranus relative to its orbital plane. Excluding Miranda and the irregular satellite discovered in 1997 and 1999, all other Uranian satellite orbits are within 0.4° of being coplanar with Uranus's equator. The actual tilt of the rotation axis of Uranus is 97.92°, although by International Astronomical Union convention, which specifies as the north pole that rotation pole which lies north of the ecliptic plane, the tilt is more properly designated as 82.08° with retrograde rotation. Much speculation has surrounded this unusual rotation of Uranus. One popular theory is that, during its early history, Uranus had a more normal tilt, but that during the latter stages of formation the planet was struck by an Earth-sized PLANETESIMAL, resulting in the present extreme tilt. The only other planets to exhibit more than a modest tilt are VENUS and PLUTO, which are 1–3 orders of magnitude smaller in mass.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

It was the low reflectivity of Uranus at the red end of the spectrum which led spectroscopist R Wildt to propose in 1932 the presence of methane (CH_4) in the atmospheres of each of the gas giant planets of the solar system. The low densities of these giant planets had also led to the conclusion that the primary atmospheric constituent was hydrogen (H_2), but confirmation of that fact came from Kiess *et al.* in 1960. Formal detection of helium (He) in the atmosphere of Uranus first came from Voyager 2 in 1986. All other gaseous components of Uranus's atmosphere (above the cloud tops) constitute a combined abundance of much less than 1%.

By analogy with Jupiter, each of the other gas giant planets in the solar system was expected to have an intrinsic magnetic field. However, the strength and orientation of those fields and the associated radio emissions were not determined until Pioneer and/or Voyager spacecraft flew close to these planets. It was the monitoring of these radio emissions that enabled Voyager scientists to determine the rotation periods of Saturn, Uranus and Neptune (see also URANUS AND NEPTUNE: ATMOSPHERES, IONOSPHERES AND MAGNETOSPHERES).

Space exploration

The only space vehicle ever to fly close to Uranus was Voyager 2. This hardy spacecraft was launched in August 1977, flew by Jupiter in July 1979, Saturn in August 1981, Uranus in January 1986 and Neptune in August 1989. As of late 1999, Voyager 2 continued to send back useful data about the particle and magnetic field environment of the outer solar system. Voyager 2 provided high-resolution images of Uranus's atmosphere, the ring system and 16 of the 21 known satellites. In fact, 11 of the 16 moons it imaged were discovered by Voyager 2. Voyager 2 also determined the detailed composition of the atmosphere, confirmed the presence of a highly tilted and offset magnetic field and measured the rotation period of the planet's interior. In fact, more than 90% of all we know about Uranus came from the measurements made by Voyager 2 and the subsequent interpretation of those measurements.

No other space missions to Uranus are at present planned and funded. Nevertheless, continued improvements in Earth-based telescopes and techniques are now providing updates to some of the Voyager data, especially in the area of atmospheric studies (as well as in the discovery of the new satellites mentioned earlier). The present emphasis on smaller, faster, less expensive missions does not bode well for missions to Uranus or Neptune, each of which requires long flight times and concomitant higher space flight operations budgets.

Composition and structure

Like its larger counterparts (Jupiter, Saturn and Neptune), Uranus is composed primarily of the light elements hydrogen and helium. Even the somewhat heavier compounds, confirmed methane (CH_4) and suspected ammonia (NH_3) and water (H_2O), are composed largely of hydrogen. This dominance of light elements is reflected in the low densities of the gas giant planets and in the absence of distinct liquid or solid surfaces in the interiors of these planets. However, in contrast to Jupiter and Saturn, methane, ammonia, water and the heavier core materials are thought to constitute a larger total mass than the overlying hydrogen and helium. These measured or deduced characteristics are thought to be the consequence of the formation process by which these planets came to be. There are a number of competing theories about the nature of that formation process. Since each theory, if sufficiently developed, predicts outcomes that are generally dependent on the starting conditions, the measured or deduced characteristics can become both an effective discriminator between theories and a means of determining the starting conditions.

All formation theories have certain characteristics in common. For example, each assumes a decrease in temperature in the presolar nebula as a function of increasing solar distance. That temperature gradient in turn determines which compounds are solid at a given distance from the Sun and which are not. With more material from which to coalesce, it is likely that the planetary cores which formed from the coalescing solids were larger at the distances of the gas giants than for the Earth-like planets, enabling the gas giants to capture more massive atmospheres. Uranus and Neptune may be smaller than Jupiter and Saturn as a consequence of the solar nebula being less dense at these greater distances from the Sun. Perhaps the growth of Uranus and Neptune therefore lagged behind that of Jupiter and Saturn. If the Sun went through its T-Tauri stage when atmospheric accretion was more or less complete for Jupiter and Saturn, but incomplete for Uranus and Neptune, most of the remaining source material may have been swept away, and further growth of the planets would have essentially stopped. In that scenario, the cores of Uranus and Neptune would have occupied a much larger percentage of their interiors than would have been the case for Jupiter and Saturn. Jupiter and Saturn were therefore expected to more closely represent the mix of materials in the original solar nebula (and in the present-day solar atmosphere).

Actual measurements confirm the predicted higher relative abundance of methane in the atmosphere of Uranus than that of Jupiter, Saturn or the Sun. These numbers imply an enrichment of carbon (relative to hydrogen) in Uranus's atmosphere that is about 24 times as high as the C/H ratio in the Sun. Carbon enrichment was also noted for Jupiter (Galileo probe measurements) and estimated for Saturn, although to a lesser extent than for Uranus (as predicted). As in the other gas giants, photochemistry of methane in the upper atmosphere produces a suite of hydrocarbons, but only acetylene (C_2H_2) has been detected thus far. Water (H_2O) has also been detected in the upper atmosphere (Feuchtgruber *et al.* 1997); its presence reveals an exogenic source of the water, presumably micrometeoroid ablation.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Helium abundance relative to hydrogen is also higher for Uranus (and Neptune) than for Jupiter and Saturn. Strangely, the helium abundances for Uranus and Neptune coincide (within the uncertainties of the measurements) with that measured for the Sun. Jupiter's atmosphere is moderately depleted in helium relative to the Sun; Saturn's is strongly depleted, having only a scant 20% of the solar He/H abundance. At very high pressures (>3 million bar) in the interiors of Jupiter and Saturn, hydrogen is transformed into a liquid metallic state. It is possible, especially in the case of Saturn, that higher-density liquid helium droplets slowly sink through the liquid hydrogen owing to the force of gravity, thus depleting the source of helium to the visible atmosphere. No such mechanism can operate in the interiors of the smaller Uranus and Neptune.

The vertical temperature profile of Uranus below the 600 mbar level increases monotonically with depth at an adiabatic lapse rate. Voyager radio science occultation measurements detected the presence of a methane cloud deck near the 1 bar pressure level, where the temperature is near 80 K. This value is colder than the minimum temperatures in the atmospheres of Jupiter and Saturn; it is for that reason that no methane clouds are present in their atmospheres. Uranus's tropopause is located near 100 mbar, where the temperature is near the minimum value of about 50 K.

Extrapolation of measured atmospheric temperatures into the interior of Uranus at an adiabatic gradient would lead to a central temperature between 6000 and 10 000 K. Unlike the other three gas giant planets, Uranus appears to have little or no heat escaping from its interior to warm the upper regions of the planet. Whether that is a consequence of its extreme axial tilt or is due to insulating layering below the visible cloud deck is not known at present. Interior modeling yields central densities of between 4.5 and 9 g cm^{-3}, depending on whether or not there is a separate molten rocky core out to a distance of about 5000 km from the center. Such a model might have the larger density values at its center, whereas a relatively well-mixed massive core of molten rocky and icy material extending through most of the interior would have the lower central density value. The measured second-order gravity harmonic (J_2=0.003 343), combined with a rotation period of 17.24 h and an oblateness (polar flattening) of 0.023, tends to favor the well-mixed core.

Dynamics

With little or no discernible internal heat source, Uranus might be expected to have little atmospheric turbulence. Figure 1 depicts the planet Uranus as seen by Voyager 2. The left-hand view is true color; the right-hand view is a highly color and contrast-stretched version of the same image. It is apparent that essentially all atmospheric features are symmetric around the rotation axis whose positive pole is just left of and slightly below the center of the disk in these images. Zonal wind velocities were measurable for only a few latitudes on the planet, mainly because

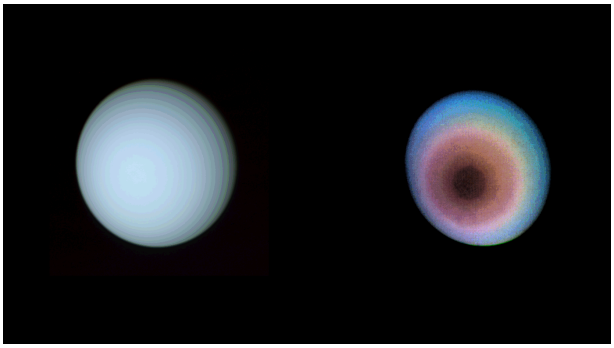


Figure 1. Uranus in true color (left) and contrast-enhanced false color (right). The planet's south (positive rotation) pole is left and below the center of the disk, at the center of the concentric patterns. Image from Voyager 2 courtesy of NASA Jet Propulsion Laboratory.

of the paucity of non-axisymmetric cloud features. Furthermore, only the illuminated southern hemisphere was visible. The data are best fitted with a latitudinal wind velocity in m s^{-1} of 200 (0.4cos$b$–cos3b), where $-90° < b < 0°$ is the latitude. This formula predicts a maximum prograde wind speed of about 240 m s^{-1} near –60° latitude, zero wind speed near –23° latitude and 120 m s^{-1} retrograde wind speed near the equator.

More recent images from the Hubble Space Telescope (HST) show greatly increased atmospheric activity for Uranus than was seen during the Voyager 2 flyby. Two color-stretched images, each including rings and satellites, are displayed in figure 2. The increase in atmospheric activity between 1986 and the 1997 HST images may be due in part to seasonal changes. During the Voyager 2 flyby, Uranus was near its southern sum-

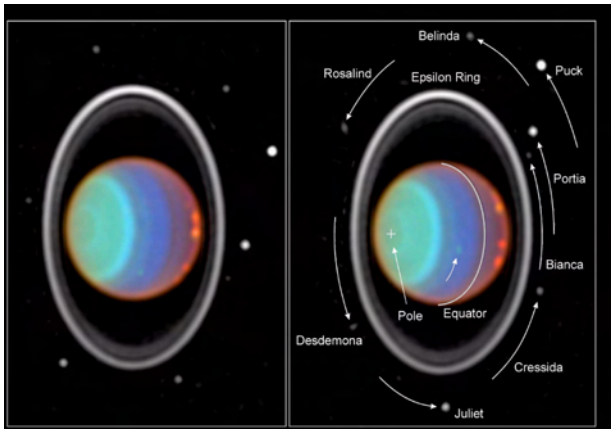


Figure 2. Uranus as seen from the HST's Wide Field Planetary Camera in 1997. The image on the right was taken later and shows the motion of satellites and atmospheric spots in the intervening time. The bright portion of the ring is the unresolved epsilon (ε) ring; the dimmer portion is a composite of several of the rings. Image from HST is courtesy of NASA Jet Propulsion Laboratory.

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

mer solstice (beginning of summer). In 2008 Uranus will reach an equinox, and all latitudes will be illuminated by sunlight. If the changes are due to seasonal effects, atmospheric cloud activity may be expected to increase until 2008 or beyond.

The temperature in the troposphere does not exhibit any north–south asymmetry despite the peculiar insolation pattern. The horizontal thermal structure shows little contrast other than shallow minima at mid-latitudes. The observed latitudinal dependence indicates that the zonal winds decay with height. Dynamical models suggest that the mid-latitude cool regions result from adiabatic cooling associated with upwelling in the troposphere and that poleward and equatorward flow occurs at upper levels.

Uranus's magnetosphere

The magnetic field of Uranus was first detected about 8 h before Voyager 2's closest approach to Uranus. Amazingly, the best dipole model of that magnetosphere indicates a magnetic dipole tilt of 58.6° and an offset from the center of the planet of 1/3 the radius of the planet in the direction of the unilluminated north pole. The high tilt and offset may indicate that the magnetic field is generated by electrical currents at fairly shallow depths within the liquid icy–rocky core. The dipole moment is more than a factor of 50 greater than the dipole moment of Earth's magnetic field, but, because of the larger size of Uranus, typical surface magnetic fields are of comparable magnitude for the two planets. The magnetic field of Uranus, like those of other planets in the solar system, is highly distorted by the solar wind. Its typical sunward extent is about 18 Uranian radii; the magnetic tail probably extends to thousands of radii downwind.

Ultraviolet auroral emissions were observed by Voyager 2 near the 44.2° north latitude magnetic pole. The auroral region was 15°–20° in diameter and is believed to be excited by electrons with energies of about 10 000 eV. A variety of radio emissions originating within the magnetosphere were detected. It was variations in the narrowband emissions near 60 kHz that enabled Voyager to determine that Uranus rotates once every 17.24 h. Detected charged particle species within the magnetosphere included mainly electrons and protons, although small amounts of ionized hydrogen gas (H_2^+) may have been present. It seems evident that the primary source of these particles is the extended hydrogen atmosphere of Uranus, which has a temperature of about 800 K at altitudes of 5000–7000 km above the cloud tops.

Table 2. *Uranus's rings' orbital data.*

Ring name	*Semimajor axis (km)*	*Orbit eccentricity*	*Inclination (°)*	*Period (h)*
6	41 837	0.001 01	0.062	6.1988
5	42 235	0.001 90	0.054	6.2875
4	42 571	0.001 06	0.032	6.3628
α	44 718	0.000 76	0.015	6.8508
β	45 661	0.000 44	0.005	7.0688
η	47 176	0.000 00	0.001	7.4239
γ	47 627	0.000 11	0.002	7.5307
δ	48 300	0.000 00	0.001	7.6911
λ	50 024	~0.0	~0.0	8.1069
ε	51 149	0.007 94	0.000	8.3823

Rings

Uranus has 10 narrow rings with sufficient optical depth to block out starlight as they passed between a star and the Voyager spacecraft or groundbased observers. By far the densest and widest of these rings is the ε ring, which

Table 3. *The satellites of Uranus.*

Name	*Distance (km)*	*Period (days)*	*Incl (°)*	*Eccentricity*	*Diameter (km)*	*Density (g cm^{-3})*
Cordelia	49 792	0.3350	0.08	0.000	~26	?
Ophelia	53 764	0.3764	0.10	0.001	~32	?
Bianca	59 165	0.4346	0.19	0.001	~44	?
Cressida	61 767	0.4636	0.01	0.000	~66	?
Desdemona	62 659	0.4737	0.11	0.000	~58	?
Juliet	64 358	0.4931	0.07	0.001	~84	?
Portia	66 097	0.5152	0.06	0.000	~110	?
Rosalind	69 927	0.5585	0.28	0.000	~54	?
Belinda	75 255	0.6235	0.03	0.000	~68	?
1986U10	76 500	0.637	?	?	~40	?
Puck	86 006	0.7618	0.32	0.000	154	?
Miranda	129 847	1.4135	4.22	0.027	472	1.15
Ariel	190 929	2.5204	0.31	0.0034	1158	1.56
Umbriel	265 979	4.1442	0.36	0.0050	1170	1.52
Titania	436 273	8.7059	0.14	0.0022	1578	1.70
Oberon	583 421	13.4632	0.10	0.0008	1522	1.64
Caliban	7 168 900	579	139.68	0.082	~44	?
Sycorax	12 213 600	1289	152.67	0.509	~40	?
1999U1	?	?	?	?	?	?
1999U2	?	?	?	?	?	?
1999U3	?	?	?	?	?	?

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

varies in width from 20 to 95 km as its particles circle the planet. The other rings have typical widths of 1–10 km. The 'equivalent depth' of the ε ring (and most of the other variable-width rings) is constant; in other words, the same number of ring particles pass through the narrower portions of each ring during each orbit as through the wider portions of the same ring. Their narrow ring widths are the result of gravitational interactions with nearby satellites, which act as 'shepherds' to keep the ring material from spreading. The orbital characteristics of the narrow rings are given in table 2.

In addition to the 10 narrow rings, Uranus has a broad, diffuse and optically thin ring component seen in some of the Voyager 2 images. It is interior to the other rings, extending from a radius of about 37 000 to about 39 500 km from the center of the planet. During passage through the shadow of Uranus, Voyager 2 also snapped an image of the rings that is unlike any other taken at Uranus. In this unique geometry, forward-scattered sunlight brings out the optically thin dust bands that occupy much of the area between the ten narrow rings mentioned above. There is a lot of radial structure in these dust rings (see figure 3), little of it understood at present.

Satellites

Prior to 1999, Uranus had 17 known moons. The discovery of the five largest (Miranda, Ariel, Umbriel, Titania and Oberon) was discussed in the introductory material earlier in this article. Ten more (Cordelia, Ophelia, Bianca, Cressida, Desdemona, Juliet, Portia, Rosalind, Belinda and Puck) were discovered by Voyager in 1986. All of the moons discovered by Voyager orbit Uranus closer to the planet than the five 'classical' satellites. In 1997 Caliban was discovered by Brett Gladman and his associates. The same year Sycorax was discovered by Phil Nicholson and his associates. Caliban and Sycorax are more than 10 times as far from the planet as any of the previously known 'regular' satellites. Their orbits are retrograde, highly inclined and more eccentric than is true for the regular satellites.

In 1999, Erich Karkoschka found a previously undiscovered moon in images shuttered by Voyager 2 in 1986. This moon, which has an orbit almost identical to that of Belinda, was found in seven different Voyager 2 images. Based on a preliminary orbit, it is estimated that Belinda laps (passes) 1986U10 about once a month, a highly unusual circumstance.

Three more probable satellites of Uranus were sighted in 1999 by Gladman, J J Kavelaars and others. Orbits for these satellites are not yet well-determined, but they are probably all distant, irregular satellites. Their temporary designations are 1999U1, 1999U2 and 1999U3. In 2002 the discovery of another satellite was confirmed. Named S/2001 U 1, this brings the total number of confirmed Uranian moons to 21. S/2001 U 1 and five others like it have very irregular, eccentric orbits that do not share the same orbital plane as the larger moons of Uranus. Ranging in size from 10 to 20 km (about 6 to 12 mi), these moons are thought to be remnants of ancient collisions that occurred at the early stage of planetary formation.

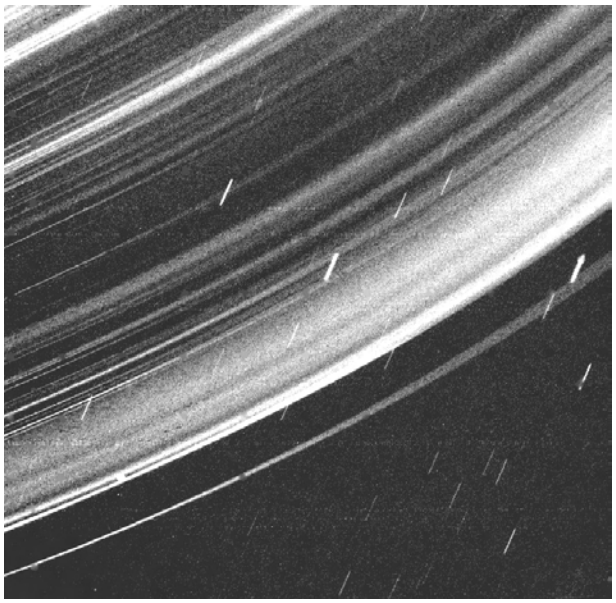


Figure 3. This long-exposure image from Voyager 2 was taken of the back side of the rings of Uranus during the spacecraft passage through the rings. The image was taken during rotation of the spacecraft and is therefore motion compensated at the left edge of the image, but the right side and the background stars are smeared. The outermost ring is the ε ring, but most of the other features seen are not identifiable with the narrow, optically thicker rings normally seen in backscattered light.

Bibliography

The Voyager 2 Encounter with Uranus 1987 *J. Geophys. Res.* **92** 14 873–15 375

Bergstralh J T, Miner E D and Matthews M S (ed) 1990 *Uranus* (Tucson, AZ: University of Arizona Press)

Feuchtgruber *et al.* 1997 *Nature* **389** 159–62

Miner E D 1998 *Uranus: the Planet, Rings and Satellites* 2nd edn (Chichester: Wiley)

Ellis D Miner

Copyright © Nature Publishing Group 2003
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2003
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Uranus and Neptune: Atmospheres, Ionospheres, and Magnetospheres

Nearly the entire mass of URANUS and NEPTUNE consists of gas, much like the other gas giant planets Jupiter and Saturn. Therefore Neptune and Uranus might be regarded as essentially all atmosphere, although it is conventional to limit the definition to the outermost layers. In a tenuous outer layer of the atmosphere we find the ionosphere (see PLANETARY IONOSPHERES), where ultraviolet light and other radiations knock electrons loose from atmospheric molecules so that electrical currents can flow there. The magnetosphere begins even further out, where the atmosphere becomes so thin that collisions between its molecules are rare, and so they fly freely like orbiting satellites. In this region, the largest forces on electrons and IONS (electrically charged atoms or molecules) come from electric and magnetic fields, and so their motions are organized by the planetary magnetic field instead of planetary gravity (see PLANETARY MAGNETOSPHERES).

Most of what we know about Uranus's and Neptune's atmosphere and everything we know about their ionospheres and magnetospheres were learned from data collected during the close passes of VOYAGER 2 near Uranus in January 1986 and near Neptune in August 1989. Unfortunately this also means that it will be difficult to learn much more without more dedicated spacecraft encounters. As a result, our mental picture of these planets, especially Uranus, is dominated by their state during their Voyager flyby, and will become increasingly unrealistic as time progresses.

Atmosphere

Visually, Uranus and Neptune are blue planets, chiefly because of absorption of red light by atmospheric methane (CH_4). Uranus is virtually featureless, while Neptune has large but short-lived cyclonic storms that appear as spots. By tracking cloud features, strong east–west winds varying with latitude were found (especially at Neptune), indicating the presence of weather.

As tables 1 and 2 show, Neptune and Uranus are nearly twin planets about four times the size of the Earth and at much greater distance from the Sun. During the formation of the solar system, the material coalescing into the planets was extremely cold because of its distance from the Sun, and so most common gases (such as water, carbon monoxide and dioxide, ammonia and methane) were frozen. Because solid material more easily accumulates into planets than does gas, Uranus and Neptune incorporated much more of this material than did planets like the Earth, that formed in warmer regions. Once these frozen gases were incorporated into Uranus's and Neptune's warm interiors, they became gaseous again, but nevertheless were still bound by gravity. Because they were able to accumulate these more common materials, Uranus and Neptune grew much bigger than the terrestrial planets, and thus were massive enough to attract substantial amounts of the still-gaseous hydrogen and helium.

Thus by number of atoms, Uranus's and Neptune's atmospheres are mostly hydrogen with some helium and heavier gases, like the other gas giant planets and the Sun. Compared with the Sun, Jupiter and Saturn, however, Neptune and Uranus are highly enriched in oxygen, carbon and other elements heavier than hydrogen and helium. As a result, these heavy elements constitute at least 80% of Uranus and Neptune by mass. In these planets' atmospheres, carbon mostly combines with hydrogen to form methane, which is slowly broken down by solar ultraviolet light to form smog, just as on Earth (see PLANETARY ATMOSPHERES, EARTH'S ATMOSPHERE). Oxygen mainly joins with hydrogen to form water, which freezes out, leaving a dry upper atmosphere just as at Earth. Both of these condensible materials fall into the interior, where they are recycled back into gaseous water and methane.

Table 1. Useful Uranian constants, in units compared with Earth (⊕) and SI metric units.

Quantity	Earth units	SI units
Distance from Sun	19.18⊕	2.870×10^{12} m
Radius	4.01⊕	2.556×10^{7} m
Mass	14.54⊕	8.683×10^{25} kg
Orbital period	84.01⊕	2.651×10^{9} s
Rotation period	0.720⊕	6.206×10^{4} s

Table 2. Useful Neptunian constants, in units compared to Earth (⊕) and SI metric units.

Quantity	Earth units	SI units
Distance from sun	30.06⊕	4.497×10^{12} m
Radius	3.89⊕	2.476×10^{7} m
Mass	17.23⊕	1.024×10^{26} kg
Orbital period	164.8⊕	5.200×10^{9} s
Rotation period	0.673⊕	5.800×10^{4} s

Uranus One of Uranus's most remarkable traits is the direction of its rotation axis. Most planets' orbits lie in approximately the same plane, and their rotation axes (and the Sun's) are roughly perpendicular to this plane. Not Uranus; its rotational axis lies nearly in its orbital plane. It is believed that this great tilt was caused by a near-catastrophic collision at the time when Uranus was still forming out of infalling bodies, one of which must have been comparable in size with Uranus itself.

This 'tipped-over' orientation means that, part of the time, the Sun will be nearly overhead at Uranus's north or south pole. Because the rotational axis stays fixed in direction while Uranus orbits around the Sun, 1/4 of an orbit later the Sun will be over Uranus's equator. Therefore the differences between the solar heating during Uranus's summer and winter are much more extreme than at Earth, in addition to being much colder overall because

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

of Uranus's much greater distance from the sun. Despite these extremes of heat input, the temperature extremes are small, apparently because the atmosphere can equalize temperature differences by transporting heat.

Another anomalous trait of Uranus is its low rate of internal heat release. Heat slowly leaks out of the interiors of the other gas giant planets, at rates comparable with the planets' absorption of sunlight. However, Uranus' internal heat loss is undetectable.

Neptune Despite Neptune's greater distance, its temperature at its apparent surface is about the same as Uranus's. The reason for this is the heat release from Neptune's interior, which makes up for its relative shortfall of sunlight. In this aspect Neptune is a more typical gas giant planet than Uranus. Neptune's rotational axis is also more typical, in being roughly perpendicular to its orbit plane, so Neptune's seasonal variation of solar heating is much less extreme than Uranus's.

Ionosphere

Uranus's and Neptune's ionospheres are currently poorly understood, because of the limitations of the observations and gaps in theoretical knowledge. However, it appears that ionization due to solar ultraviolet and galactic cosmic rays is mostly responsible for maintaining the ionosphere, with enhancements from meteoritic impacts.

Magnetosphere

Uranus and Neptune have planetary magnetic fields whose surface intensity is the same order of magnitude as the Earth's field, but whose distribution is completely different. At Earth the field is oriented roughly north–south and its strength is roughly the same everywhere. Neptune's and Uranus's fields are much more variable in strength and direction.

Like Earth, Neptune and Uranus have AURORAS, which are glows created by electrons hitting the upper atmosphere where the magnetic field lines converge. At Earth auroras occur near the north and south geographic poles, but at Uranus and Neptune the magnetic poles lie at middle latitudes so that is where their auroras occur also.

Uranus The magnetosphere of Uranus is tilted once by Uranus's rotation axis and once again by the magnetic field's tilt with respect to the rotation axis. The charged particles in Jupiter's and Saturn's magnetospheres mostly rotate with their planets. At Earth charged particles flow in a kind of back-eddy opposite to the direction of the solar wind. At Uranus they do both at the same time, although this may change as Uranus's orbit carries it from Sun-over-pole to Sun-over-equator orientation.

Uranus's magnetosphere contains mostly atomic hydrogen ions (H^+), which probably come from Uranus' atmosphere.

Neptune The magnetic field of Neptune is also drastically tilted and offset with respect to the rotation axis. Despite Neptune's normal rotation axis orientation, charged particle trajectories are even more complicated here than at Uranus. Many of these trajectories exit the magnetosphere, so the density of ions in Neptune's magnetosphere is very sparse.

Neptune's magnetosphere contains mostly atomic hydrogen ions (H^+) and atomic nitrogen ions (N^+). These ions are probably derived from the atmosphere of TRITON, Neptune's large, icy moon. Triton's atmosphere is mostly molecular nitrogen with traces of hydrogen-bearing species such as methane.

Lower atmosphere and interior

Because Uranus and Neptune are gas giants, the distinction between atmosphere and interior is arbitrary. By convention the boundary between them is placed at a level where the pressure is 10–100 bar (1 bar $= 10^5$ N m^{-2} is about the pressure at the bottom of the Earth's atmosphere). This definition is based partly on observability and partly on the diminished effect of weather at that level.

Also by convention, the planetary radius is defined as the distance from the planetary center out to the level where the pressure is 1 bar at the equator. These are abbreviated as $R_U = 25\,559$ km and $R_N = 24\,764$ km, for Uranus and Neptune respectively.

Most of what is known of Uranus's and Neptune's interiors is derived from the slight departure of their gravity fields from spherical symmetry, detected by perturbations of Voyager's flyby trajectory. From these slight perturbations has been deduced an approximate density versus radius profile, which suggests the presence of rocky cores roughly half the size of Earth, representing about 0.1–1% of Neptune's or Uranus's total volume.

The properties of Uranus's atmosphere are summarized in figure 1, while Neptune's is summarized in figure 2. The number of molecules per cubic centimeter (number density) of each of the more important species is profiled versus altitude and pressure. The H_2 density is controlled by the temperature profile via the hydrostatic equation,

$$\frac{\mathrm{d}p}{\mathrm{d}r} = -\frac{mpg(r)}{kT(p)}$$

with p is the pressure, m the mass of an H_2 molecule, k Boltzmann's constant, $g(r)$ the acceleration of gravity as a function of altitude and $T(p)$ the temperature versus pressure profile. The densities of the other species are controlled by their interactions with the dominant H_2. The temperature profile is controlled by the balance of heat gain and loss, which is in turn controlled by the mechanisms by which heat is transported. The different temperatures seen at different levels result from different heat sources and transport modes dominating at different densities.

Both planets' tropospheres (from words meaning the region of changes; the lower atmosphere, where there

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

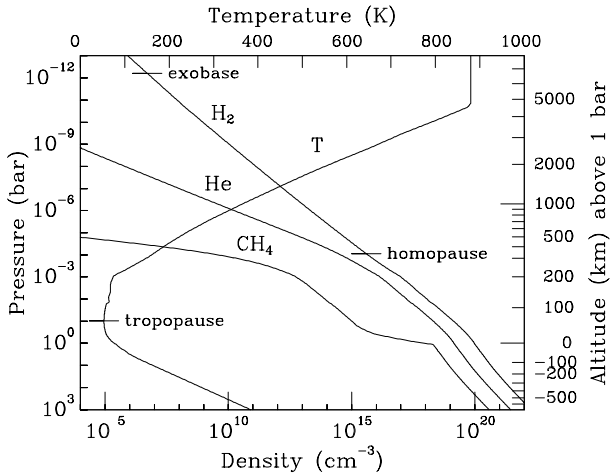


Figure 1. Profiles of temperature and density versus altitude in Uranus's atmosphere. The temperature (T) is read from the top scale, the densities (labeled by chemical symbol) from the bottom. Important boundary altitudes are marked on the profiles that determine them.

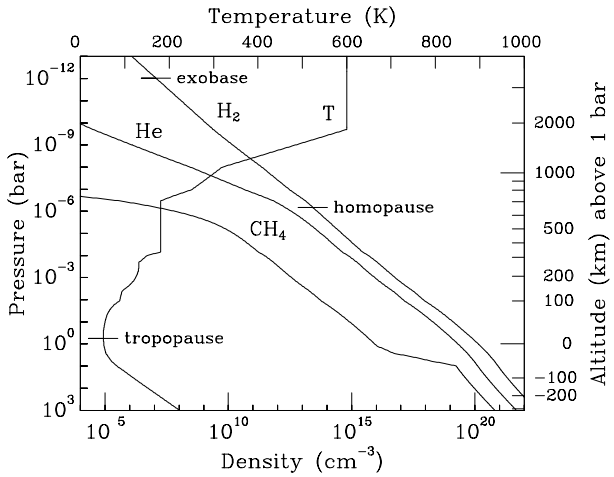


Figure 2. Profiles of temperature and density versus altitude in Neptune's atmosphere. The temperature (T) is read from the top scale, the densities (labeled by chemical symbol) from the bottom. The temperature profile is an approximation constructed to match the observations; its sharp bends are not real. Important boundary altitudes are marked on the profiles that determine them.

is weather) are very cold compared with Earth. Within the troposphere the dominant heat transport mode is convection, whereby heat is carried along with the gas as it rises and falls. Because the pressure changes with altitude, the gas is cooled by expansion as it rises and heated by compression as it sinks. Thus in the troposphere the temperature variation with altitude follows the adiabatic rate of change, that is, the rate at which a column of rising air would cool just as a result of expansion without heat exchange.

At Earth the vertical temperature profile is complicated by the condensation of water, but, in the colder Uranian troposphere, water condenses at a deeper level. At Neptune and Uranus the visible clouds are probably condensed methane. These clouds are located near the level of the sharp bend in the CH_4 profile just below the 1 bar level in figures 1 and 2. Above that level the ratio of CH_4 to H_2 density abruptly decreases because of methane condensation.

At the tropopause (the upper boundary of the troposphere, where the temperature stops falling with altitude), the temperature is about 50–52 K (much colder than liquid nitrogen) at both planets. The altitude of this transition point is marked on the temperature profiles displayed in figures 1 and 2. At this altitude the methane vapor pressure (at a given temperature, the pressure of vapor in equilibrium with its solid form; usually extremely temperature dependent) is at a minimum. Any methane abundance in excess of this pressure is frozen out and the resulting ice falls to lower altitude, which is why the CH_4 profiles in figures 1 and 2 show abrupt decreases there.

Uranus

The temperatures within Uranus's interior are probably high, but unknown. Heat constantly leaks out of the other gas giant planet interiors, and from the rate of heat loss a temperature versus radius profile necessary to maintain that loss can be calculated for those planets. Uranus's internal heat loss, by contrast, is too small to have been measured, so its internal temperature is unknown.

It may be that Uranus's interior is actually cooler than the other gas giants', perhaps because the collision that tipped Uranus over also stirred it enough to let its excess heat out. Another possibility is that the settling of denser materials has stabilized the interior so that hot gas does not rise and carry out heat, because even if the lower regions are hotter they would be compositionally denser and so not buoyant.

Neptune

Unlike Uranus, Neptune does have a measurable rate of heat loss, which is about 1.6 times as large as the rate of heat absorption from sunlight. From this heat loss rate, a temperature versus radius profile can be calculated, which is very near to an adiabat. The temperatures within the interior are high, reaching a maximum of probably about 6000 K at the center.

Upper atmosphere

Above the tropopause in each planet is the stratosphere, where the atmosphere is relatively stable (hence the name) because the temperature rises with altitude. Atmospheric stirring does not completely cease here, but it is greatly attenuated compared with the troposphere. In the lower stratosphere, the temperature profile is mainly controlled by the absorption and reradiation of heat absorbed from sunlight. The upper atmosphere is heated by a high-altitude energy source of unknown

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

nature. Because the energy is input at such low densities, no known component of sunlight or bombardment by magnetospheric charged particles can be responsible.

This heat is not reradiated from where it arrives, because H_2 and He are very poor radiators of heat and the more effective radiators, the hydrocarbons, are scarce here. Thus thermal conduction transports the heat down to the level where it can be radiated into space by the hydrocarbons. In order for conduction to carry the heat this far, the heat-source region's temperature must rise to high temperatures. From there up to the exobase (the atmospheric level above which atomic and molecular collisions are negligible), the atmosphere is roughly isothermal.

Because of the factor of m in the hydrostatic equation, atmosphere pressure diminishes with altitude at a rate proportional to m, the atmospheric molecular mass. When an atmosphere is a mixture of gases, atmospheric turbulence and other stirring mechanisms tend to keep the different gases mixed in equal proportions. Thus all gas densities decrease with altitude at the same rate, an average value which, at Uranus and Neptune, is about equal to the rate for H_2. However, above a certain altitude, where the mixing is not vigorous enough to keep the proportions uniform, each gas density decreases with altitude at its own individual rate. The point where this separation of rates begins is called the homopause, which is marked on figures 1 and 2. Notice the change in slope of the He and CH_4 profiles there, above which each constituent is decoupled from the others and so follows its own independent profile.

It is because of the rapid falloff of the CH_4 density compared with the lighter elements that the hydrocarbons are confined to lower altitudes, to which heat absorbed at high altitudes must be conducted before it can be radiated into space. Thus the level of the homopause helps determine the maximum temperature of the uppermost region of the atmosphere.

Uranus

The stratosphere of Uranus has been discovered to be unusually calm compared with the other giant planets, so that heavier gases such as helium and methane are not well-mixed with the hydrogen at higher altitudes. In other words, the homopause lies at lower altitude and higher pressure, about 10^{-4} bar in figure 1. Because of the low methane abundance at the higher altitudes where solar extreme ultraviolet light can penetrate and drive photochemistry, there are also lower abundances of the higher hydrocarbons such as ethane and acetylene there. Therefore there are fewer photochemically produced aerosol particles, or smog (an aerosol is a particle so small that it settles out of the air very slowly). As a result, Uranus's stratosphere is unusually clear for a gas giant, so that Rayleigh scattering (light scattering caused by spontaneous microscopic fluctuations of the air's density) is the dominant atmospheric opacity at wavelengths away from methane absorptions. At Earth, when opacity from water aerosols (clouds) is negligible, Rayleigh scattering is also important, turning the sky blue and sunsets red because of its λ^{-4} wavelength dependence.

Because the hydrocarbons lie so deep in the stratosphere, heat conducted from its source around 10^{-11} bar must travel a great distance. In order for conduction to carry the heat this far, the heat-source region's temperature must rise to about 850 K.

This high temperature inflates the upper atmosphere greatly, particularly the light atomic hydrogen component (a minor component at lower altitudes), so that it falls off slowly with altitude. Therefore even at the great altitudes ($1.64R_U \leq r \leq 2R_U$) of Uranus's rings, which are basically icy orbiting pebbles, the gas density is high enough to produce frictional drag. As with artificial Earth satellites, atmospheric drag causes Uranus's ring particles to gradually lose their orbit-maintaining momentum until eventually they fall vertically into the troposphere. Because the ratio of gas drag to particle inertia is largest for the smallest ring particles, Uranus's inflated atmosphere selectively removes the finest ring particles. The steady shrinkage of ring-particle orbits, piling up against the repulsive perturbations of larger orbiters with slower orbital decay, may be part of the reason that Uranus's rings are a series of very thin concentric bands, in contrast with the nearly continuous disk that constitutes Saturn's rings.

Neptune

The upper atmosphere of Neptune is much more vigorously mixed than is Uranus's, as shown in figure 2 by its higher homopause altitude (at around 10^{-6} bar instead of 10^{-4} bar as at Uranus). For this reason methane and other hydrocarbons are found at significant densities much higher in the atmosphere at Neptune, as is also evident in figure 2. Neptune's degree of atmospheric mixing is more typical of the other giant planets, Jupiter and Saturn, than is Uranus's.

Probably because the hydrocarbons reach higher in the atmosphere, the heat flow from Neptune's high altitude heat source (not well defined, but probably at no higher an altitude in pressure units than at Uranus) has a shorter distance to travel before it can be radiated into space. Therefore the hot uppermost regions of Neptune's atmosphere, at about 550–650 K, do not reach as high a temperature (about 850 K) as Uranus's does.

Ionosphere

Ionospheres are formed by ionizing radiation and represent a balance between the ion production rate and the recombination loss rate. Because both the recombination loss rate and the atmospheric opacity to ionizing radiation are proportional to gas density, which increases rapidly with depth in the atmosphere, ionospheres typically have sharp lower boundaries. The upper part of the ionosphere, however, typically decreases upwards at an exponential rate determined by the local atmospheric temperature. This rate has been measured for

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Neptune's and Uranus's ionospheres and is about right for H^+ ions.

In addition to this smooth variation, thin layers of higher electron density were observed in Uranus's and Neptune's ionospheres. It has been conjectured that these are transitory effects of meteor showers, which are observed to create ionization in Earth's atmosphere. These thin layers are at the bottoms of the two ionospheres, which lie at about 1000–2000 km above 1 bar at Uranus and about 500–1000 km at Neptune.

Magnetosphere

The magnetosphere is the region outside a planet's exobase, where the planetary magnetic field is strong enough to deflect the solar wind. On the upwind side, the pressure balance between field and wind creates a roughly parabolic boundary, called the magnetopause, with nose distance about $20R_U$ upstream at Uranus and $25R_N$ at Neptune. This boundary envelops its associated planet and extends downstream in a wake many tens of planetary radii long. Outside of this boundary is a roughly hyperboloidal bow shock standing in the highly supersonic solar wind.

The inner boundary of the magnetosphere is effectively its planet's exobase, below which collisions with atmospheric molecules control charged-particle motion. However, atmosphere–magnetosphere coupling can extend more deeply than the exobase, as magnetic-field-aligned currents can flow into the ionosphere, across to other field lines and back out into the magnetosphere. It is this coupling that drives much collective magnetospheric motion.

The magnetosphere encompasses a wide range of phenomena, described in their individual sections (see figures 3 and 4). Because of the long-range coupling present in the magnetosphere, its seemingly independent phenomena mutually interact more than is the case in the neutral atmosphere.

Plasma

PLASMA is a gas which contains significant numbers of ions and electrons, so that it conducts electricity at least along the magnetic field, and often perpendicular to it as well. The plasma at both Uranus and Neptune was composed of ions and electrons with kinetic energies covering Voyager's entire range of measurement, from 10 to about 10^7 eV (1 eV $= 1.6 \times 10^{-19}$ J is a particle kinetic energy equivalent to a temperature of about 8700 K). Plasmas in different energy ranges behave differently, so we discuss them separately.

The high-energy ($\geq 10^5$ eV) plasma at both planets is extremely sparse and forms radiation belts similar to those at Earth. It diffuses inwards through the magnetosphere, so that its absorption by satellites and rings creates decreases in density that were observed by Voyager. A few faint rings and satellites were discovered in this way. When these energetic particles strike icy satellites and ring particles, hydrogen and oxygen atoms are ejected.

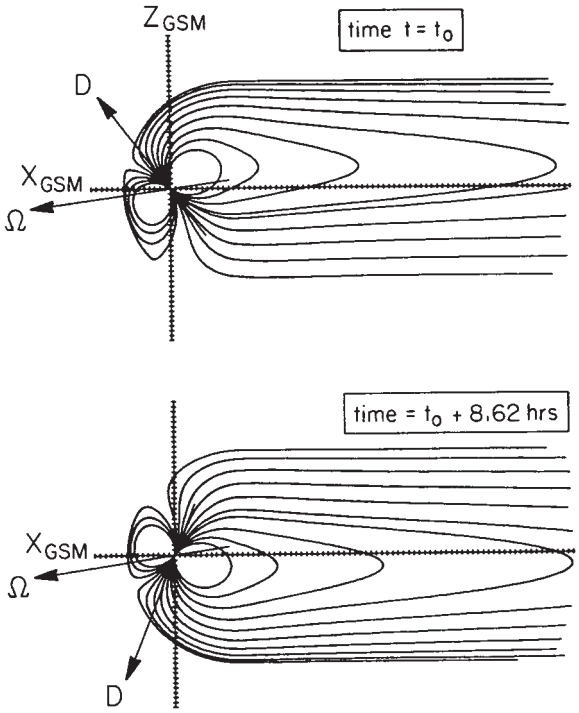


Figure 3. Diagrams of Uranus's magnetoshere. The curved lines depict the magnetic field as it is confined within the magnetopause. The arrow labeled Ω is Uranus's axis of rotation, and is fixed in space. The arrow labeled D is the magnetic dipole axis and rotates with the planet. The two diagrams show the magnetosphere at half-rotation intervals. The axis labeled X_{GSM} points toward the Sun. (Figure from Voigt *et al* 1987 *J. Geophys. Res.* **86** 15 337.)

The low-energy ($\leq 10^4$ eV) plasma drifts through these magnetospheres on paths determined mostly by the electric and magnetic fields. Both the rotation of the planets and the interaction of the solar wind with the magnetospheres create electric fields that fluctuate in time as well as space. In the following, we describe what is known of the behavior of the low-energy plasma.

Uranus. Uranus's magnetospheric plasma is rather tenuous and consists almost entirely of H^+. It appears to originate from Uranus's atmosphere, because if it came from the icy satellites there would also be O^+, which is not observed. It drifts Sunward as at Earth, but because of the pole-on configuration simultaneously co-rotates with Uranus. However, the co-rotation does not interfere with the plasma drift, as it does in the Earth's plasmasphere, because the rotation occurs about an axis parallel to the drift.

As the plasma drifts toward (away from) Uranus, it sees an increasing (decreasing) magnetic field strength which compresses (expands) the plasma and heats (cools) it drastically. Thus there is a plasma adiabat in the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

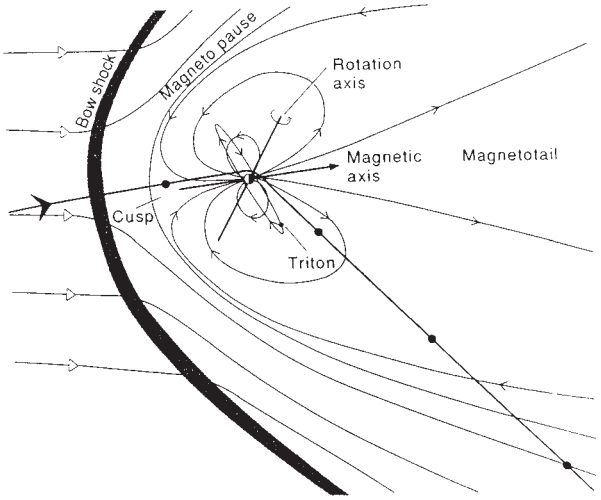


Figure 4. Diagram of Neptune's magnetosphere, showing its components. The curved lines with triangular arrowheads depict the solar wind flow deflected around the magnetopause. Those with V-like arrowheads are magnetic field lines, except the one labeled 'Triton', which is Triton's orbit. The line with the full dots is the trajectory of the Voyager flyby. (From Schulz *et al* 1995 Magnetospheric configuration of Neptune *Neptune and Triton* ed D P Cruikshank (Tucson, AZ: University of Arizona Press). Copyright © 1995 by The Arizona Board of Regents. Reprinted by permission of the University of Arizona Press.)

magnetosphere just as there is a gaseous adiabat in Uranus's interior.

The plasma consists of a number of components, which are distinguished by their temperatures. Because of the adiabatic drift compression, their densities and temperatures are best compared by referral to a standard distance from the Uranian magnetic dipole. Two low-energy populations are known; a warm plasma (0.5 cm^{-3} and 10 eV at $5R_U$) and a hot plasma (0.3 cm^{-3} and 1000 eV at $5R_U$). The warm plasma drifts by Uranus without much change, but the hot electrons continually lose energy (relative to their adiabatic temperature) as they approach Uranus. The hot ions also abruptly disappear inside approximately $5R_U$, as though their drift paths were avoiding Uranus.

Neptune. Neptune's magnetospheric plasma is very tenuous and consists mostly of H^+ and N^+ ions. Its source appears to be the atmosphere of Triton, because of the presence of N^+ ions. In addition, a measure of the plasma density, corrected for compression by the magnetic field, appears to decrease as Neptune is approached. If Neptune were the source, the plasma density measure would increase towards the planet instead.

The decrease of plasma density measure towards Neptune is due to loss of some of the plasma to collisions with the atmosphere as the charged particles bounce back and forth along the magnetic field lines. Unlike the case of the aurora, where high-energy (10^4 eV or more) charged particles from far out in the magnetosphere hit the atmosphere near the magnetic poles, this plasma loss refers to low-energy (tens of eV) plasma on field lines near the planet that is impacting the atmosphere far from the magnetic poles.

The motion of Neptune's magnetospheric plasma is complex and not completely understood. Neptune's spin axis is roughly perpendicular to its orbital plane, just as it is at Jupiter and Saturn, where magnetospheric plasma rotates with the planet. However, the large tilt of the magnetic field results in Neptune's magnetic dipole alternating between states pointed at the Sun and perpendicular to the Sun–Neptune line as Neptune rotates. As a result plasma motions in Neptune's magnetosphere fluctuate as Neptune rotates. As at Uranus, plasma motion is part Sunward back-eddy and part co-rotation.

Aurora

Auroras are regions where the atmosphere glows because of bombardment by high-energy charged particles (usually electrons) which flow into the atmosphere along the magnetic field lines. They typically occur near, but not quite at, magnetic poles because the field lines there connect to the largest regions of the magnetosphere.

Uranus. Uranus's magnetic poles, and therefore its auroras, are far from the geographic poles because of its large magnetic dipole tilt (60°) and offset ($0.3R_U$). Moreover, both auroras are on magnetic field lines that connect to the magnetospheric tail, so that the aurora-exciting particles (probably 10^4 eV electrons) streaming down the field lines into the atmosphere are still drifting toward Uranus. Because the plasma drift paths continue on past Uranus, on field lines where there is no aurora, it seems likely that the auroras are exhausting the available hot electrons before they can drift completely past Uranus.

The aurora is visible primarily at in the extreme ultraviolet at wavelengths between 90 and 110 nm, emitted by electronic transitions between low-lying states of molecular hydrogen. These bands are excited by the low-energy secondary electrons emitted when high-energy charged particles impact the atmosphere.

Neptune. Neptune's magnetic poles and auroras are also far from the geographic poles because of its large magnetic dipole tilt (47°) and offset ($0.55R_N$). The light from Neptune's aurora, like Uranus's, is extreme ultraviolet H_2 molecular band emission between 90 and 110 nm, stimulated by auroral secondary electrons. Some diffuse glow is also stimulated by electrons which are emitted by the day-side atmosphere in response to solar ultraviolet light and which follow the magnetic field lines around to impact the dark side.

Radio waves

There are a number of types of radio waves emitted by Uranus and Neptune with frequencies in the range 100–1000 kHz, both narrow and wide bandwidth and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

emitted both in bursts and semicontinuously. Uranus's and Neptune's radio emissions are more similar to each other than to those of other planets. They can be observed from outside the magnetosphere, although, at the present orientation of Uranus, the strongest Uranian radio emissions are beamed away from the Earth

Several types are emitted from regions not far out of the atmosphere on the auroral field lines. In most cases they are generated by electron velocity distributions that are relaxing to equilibrium by means of collective instabilities. One of the ways to cause electron velocity distributions to depart from equilibrium, and so generate these waves, is to selectively remove part of the distribution by atmospheric absorption in an aurora, which is why these radio emissions are located there.

Most of the radio emissions are confined to a narrow range of directions, or beamed, and rotate with the planet. The rotation periods of both Neptune and Uranus were measured by timing the recurrent detection of these emissions in the weeks after each Voyager flyby.

Another radio emission emitted at Uranus, but only marginally at Neptune, is the radio noise produced by lightning discharges. The strength of these emissions suggests that lightning discharges are weaker at Uranus than at Saturn but considerably stronger than at Neptune and at Earth. At Neptune, however, whistler plasma waves (see below) originating from lightning were found instead. Therefore lightning does occur at Neptune, but the discharges must be much weaker than at Uranus.

Plasma waves

Plasma waves are collective disturbances in a plasma that are coupled to perturbations in the electric and magnetic fields and that move with respect to the plasma. They are, like the radio emissions, typically generated by departures of the plasma from thermal equilibrium. Unlike the radio waves, however, they do not propagate out of the plasma, and so cannot be observed from outside the magnetosphere.

The strongest of them, Whistler-mode noise, is caused by the depletion of magnetospheric electrons traveling along the magnetic field when they impact the atmosphere and create an aurora. This imbalance between the numbers of electrons with velocities parallel and perpendicular to the magnetic field is unstable, and the whistler-mode noise is generated by the motions of the electrons as they approach equilibrium. In the process, however, more electrons stream along the field and impact the atmosphere, perpetuating the process. Unsurprisingly, therefore, whistler-mode noise was observed at its greatest intensity at Uranus when Voyager was on field lines that were connected to the aurora.

Whistlers are plasma waves propagating in the same mode as whistler-mode noise (and vice versa) but are generated by a different process—lightning discharges. Whistlers are very short bursts of plasma wave noise that are spread out in frequency as they propagate along the magnetic field. In a radio receiver they sound like pure tones descending in frequency; hence the name. Whistlers emitted by lightning discharges were detected in Neptune's magnetosphere.

Current research

Mostly because of the lack of new data, Uranus and Neptune research has declined greatly in recent years. The HUBBLE SPACE TELESCOPE (or HST, a 2.4 m aperture orbiting telescope covering the wavelengths 115–2500 nm) and large-aperture ground-based telescopic infrared observations of the ionospheric species H_3^+ between 3890 and 4090 nm are the main sources of new data.

There are many questions for future research. How does Uranus's atmosphere smooth out such great seasonal extremes of solar heating? Why is Uranus's internal heat source so small compared with those of the other gas giants? Why is Uranus's stratosphere so much calmer and less mixed than other giant planet stratospheres? What is the high-altitude heat source that heats Neptune's and Uranus's upper atmospheres to such high temperatures? What is the pattern of motion of Uranus's and Neptune's magnetospheric plasma? What is the origin of bursty radio emissions? Why are Jupiter's, Uranus's and Neptune's radio emissions rotation controlled, while Earth's and Saturn's are solar wind controlled?

Bibliography

Bergstralh J T, Miner E D and Matthews M S 1991 *Uranus* (Tucson, AZ: University of Arizona Press)

Cruikshank D P 1995 *Neptune and Triton* (Tucson, AZ: University of Arizona Press)

Lunine J I 1993 The atmospheres of Uranus and Neptune *Ann. Rev. Astron. Astrophys.* **31** 217–63

Floyd Herbert

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Uranus: Rings

The Uranian ring system comprise ten narrow rings, one diffuse ring and some tenuous material between the narrow rings. With a mixed nomenclature arising from historical tradition, the names of the ten narrow rings (in order of increasing distance from the planet) are 6, 5, 4, α, β, η, γ, δ, λ, ϵ; the name of the diffuse ring is 1986U1R. The innermost narrow ring lies about 16 000 km above the microbar pressure level in the Uranian atmosphere, and the entire ring system is about 10 000 km in radial extent. The distinctive feature of these rings is their narrowness, most having a radial width less than 12 km (table 1).

Although these rings can now be easily imaged from the Earth with modern detectors, they were first observed with the technique of *stellar occultations*. A stellar occultation occurs when starlight is blocked from an observer by an intervening body—in this case the Uranian rings. Observations are carried out with high-speed photometric equipment, ideally with several observing sites well spaced in the direction perpendicular to the motion of the ring shadows across the Earth. The recording of the starlight intensity as a function of time then provides a 'line scan' of the transmission of the ring material. With multiple observations of the same occultation by a network of telescopes, several points around each ring are recorded, and the ring orbits can be constructed. By repeatedly observing occultations over several years, one can learn the ring orbits to great accuracy. One advantage of the stellar occultation technique over ordinary imaging observations with Earth-based telescopes is that occultations achieve a spatial resolution (limited by *fresnel diffraction*) of just a few kilometers, rather than the much coarser resolution achievable with Earth-based telescopic observations.

The discovery of the Uranian rings with airborne and ground-based observations of a stellar occultation in 1977 ushered in the beginning of modern ring research. With only the broad, bright ring system of Saturn as an example (see SATURN: RINGS, no one had suspected the existence of dark, narrow rings. Any thought of narrow rings was quickly dismissed because collisions between ring particles would cause them to spread out into a broader ring over time. In order to explain the existence of narrow rings, theorists postulated the existence of small satellites near the rings, termed SHEPHERD MOONS, that would confine the particles through their gravitational interaction. This idea was borne out by VOYAGER's discovery of Prometheus and Pandora, two satellites that shepherd the narrow F ring of Saturn. Voyager also discovered Cordelia and Ophelia, the two shepherd satellites of the ε ring. However, pairs of shepherd satellites have not yet been identified for the other Uranian rings, so the mechanism(s) for maintaining narrow rings may not be completely identified.

In the two decades following their discovery, about 20 stellar occultations by the rings were observed, mostly in methane bands where the background light from Uranus is minimized. In 1986 the Voyager 2 spacecraft flew by the Uranian system and recorded several high-resolution images of the rings. In addition, the Voyager Photo-polarimeter Subsystem (PPS) and Ultraviolet Spectrometer (UVS) instruments observed stellar occultations by the rings, and occultations of the Voyager radio signal (from the Radio Science Subsystem, RSS) were observed from Earth. Due to the much closer vantage point of the spacecraft, the Voyager data have provided the highest spatial resolution observations of the rings that we have so far. In particular, the radio-occultation data at X-band has a resolution of 50 m, which can be compared with the 2 km resolution of ground-based stellar occultation data at visible wavelengths.

Table 1. Orbital properties† of Uranus's rings.

Ring	Semimajor axis (km)	Radial width (km)	Eccentricity	Inclination (degrees)
6	41 837.2	1.0–2.5	0.001 013	0.062
5	42 234.8	1–7	0.001 899	0.054
4	42 570.9	1–7	0.001 059	0.0323
α	44 718.5	4.5–10.5	0.000 761	0.0152
β	45 661.0	5.5–12	0.000 442	0.0051
η	47 175.9	55	∼0.004	∼0.0011
γ	47 626.9	1–8	0.109	∼0.002
δ	48 300.12	2–8	0.004	0.0011
λ	50 023.9	1.3–2.5	∼0.0	∼0.0
ε	51 149.3	20–96	0.007 936	∼0.0002

† Errors in these quantities are in the least significant digit, and values preceded by a tilde have an error comparable to the value. These results are from French *et al* (1991).

Several ring phenomena were first observed in the Uranian ring system. After the extreme narrowness of the rings, the most surprising of these is the elliptical and/or inclined orbits exhibited by some of the rings. Dynamical theories had predicted that rings would be limited to circular orbits in the equatorial plane of the planet. The eccentricities and inclinations are excited by the shepherd satellites, which can also produce an extremely sharp ring edge, if it is located at an exact orbital resonance with the satellite. Examples of sharp edges produced by resonances with the satellites Cordelia and Ophelia are the inner and outer edges of the ε ring, the outer edge of the δ ring and the inner edge of the γ ring.

Elliptical ring orbits exhibit PRECESSION, which is a slow change in the orientation of the ellipse. The precession of the ring orbits is caused by the oblate mass distribution within Uranus, and the rate of this motion gives us information about the interior structure of the planet. Here we can consider the rings as 'test particles' moving within Uranus' gravitational field. From their precessional motions, two of the gravitational harmonic coefficients have been precisely established ($J_2 = 3.3434 \times 10^{-3}$ and $J_4 = -2.89 \times 10^{-5}$). These allow the computation of the Uranian internal density distribution, which provides a constraint on allowable interior models for Uranus.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

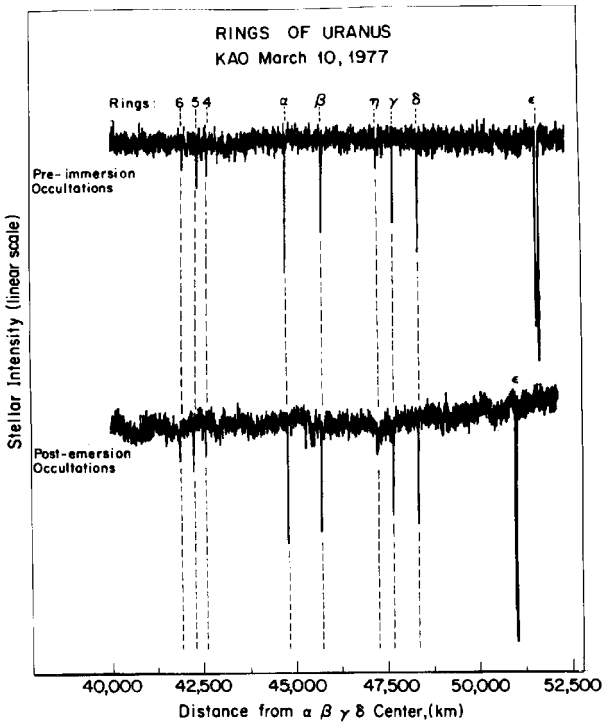


Figure 1. Discovery observations of the Uranian rings from the Kuiper Airborne Observatory. The intensity of starlight before and after Uranus occulted the star is plotted versus distance from the center of the planet. Nine dips were observed on each side, which occurred at the same radius, indicating that Uranus is encircled by a system of narrow rings. (After Elliot J L 1979 Stellar occultation studies of the solar system *Ann. Rev. Astron. Astrophys.* **17** 445–75. Reprinted, with permission, from the *Annual Reviews of Astronomy and Astrophysics* Volume 17, ©1979, by Annual Reviews www.AnnualReviews.org.)

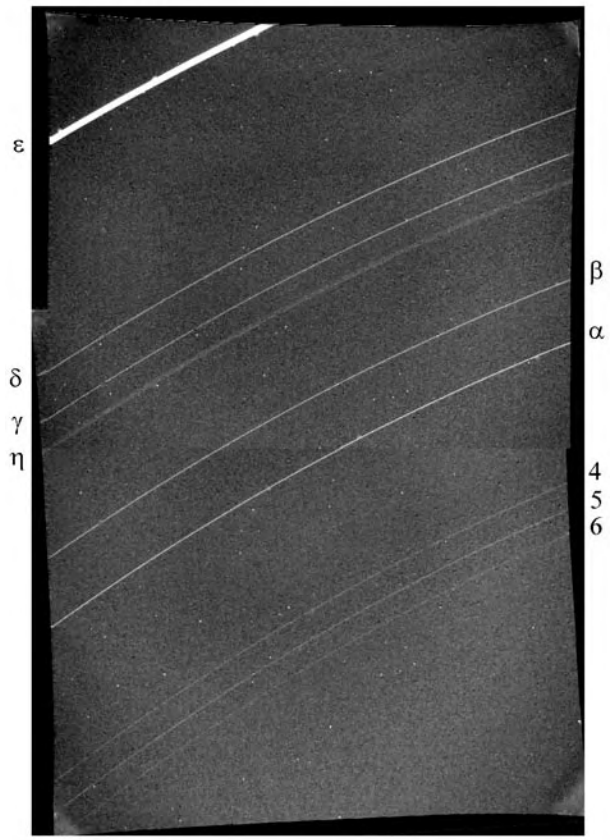


Figure 2. Voyager observations of the Uranian ring system. The three faint rings in the lower right are (in order of increasing radius) 6, 5 and 4; these are followed by α, β, η, γ, δ and ε. The λ ring is not visible. Ring ε contains the most material and is the brightest because it reflects the most sunlight. (Courtesy NASA/JPL/Caltech.)

The largest ring of the system (ε) has a minimum radial width of 20 km (at PERIAPSIS, its closest distance to Uranus) and a maximum 96 km (at APOAPSIS, its furthest distance). This width is a strict linear function of the ring radius, which implies that the precessions of the individual particle orbits are 'locked' at the same rate. If the precession rates were not all the same, then the orbits of the particles in the outer parts of the ring would precess more slowly than those for the particles closer to the planet, and the radial width of the ring would become uniform around the ring. The radial widths of rings α and β also vary linearly as with distance from the planet, but the variation is only 4.5 to 10.5 km for the α ring and 5.5 to 12 km for the β ring. One mechanism proposed for maintaining the locked precession of the particle orbits is the gravity of the ring particles themselves, but this has not been conclusively proven to be the case.

In discussing the structure of the Uranian rings, one property of interest is the *equivalent depth*, which is the *optical depth* (minus the natural logarithm of the ring transmission) integrated along the radial direction. Thus the equivalent depth measures the amount of material in the ring in a way that is independent of the angle at which the ring occults a star or other source of radiation. For most rings the equivalent depth is the same all around the ring, indicating a constant amount of material at each longitude of the ring orbit. Equivalent depth has the dimensions of length.

Rings 6, 5 and 4 have a structure similar to each other, being very narrow with equivalent depths of 0.8, 1.8 and 1.4 km. These rings have variable widths, but the widths do not correlate with orbital radius (as they do for the α, β and ε rings). The Voyager occultation data reveal significant internal structure. Rings α and β are more substantial, with equivalent depths of 6.0 km and 3.8 km respectively. These rings rarely exhibit the sharp edges characteristic of resonances with satellites (two sharp-edged profiles were observed for the α ring in the Voyager data). Rings γ and δ are concentrated, with equivalent depths of 6.6 km and 4.1 km. Each of these rings has an excited normal mode, with an amplitude of 5 km for the γ ring and 3 km for the δ ring. The δ ring may also have an internal density wave. The η ring is unique, with a sharp

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

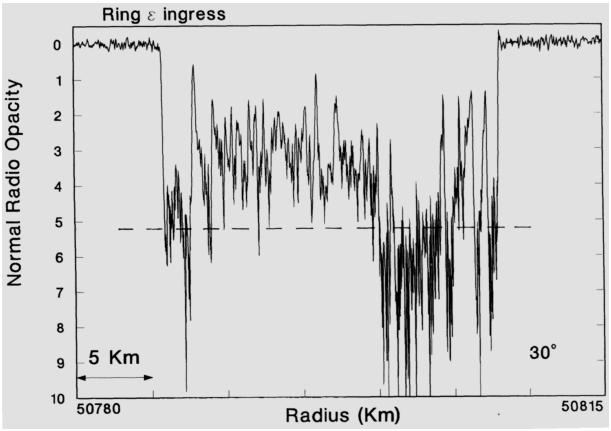


Figure 3. Occultation profile of the ε ring. This graph shows the normal radio opacity (optical depth) of the ring as a function of radius. Note the sharp inner and outer edges caused by orbital resonances with Cordelia and Ophelia and the detailed internal structure. Below the dashed line, the variations are mostly due to noise, rather than structure of the ring. (After Gresh D L, Marouf E A, Tyler G L, Rosen P A and Simpson R A 1989 Voyager radio occultation by Uranus' rings. I. Observational results *Icarus* **78** 131–68.)

inner feature (similar to rings 4, 5 and 6) and an adjacent broad, low-optical-depth component about 55 km in radial extent and with a sharp outer edge. Both components of the η ring follow circular orbits in the equatorial plane of Uranus. The λ ring is tenuous and clumpy, with an equivalent depth of only 0.1 km. Finally, the ε ring has very sharp inner and outer edges and an equivalent depth of 83.8 km. Its occultation profile has a 'W' shape, showing that its material becomes more concentrated at its edges (figure 3).

The rings are very flat—with a vertical thickness (perpendicular to the plane of their orbits) of only a few tens of meters. The ring particles themselves are dark, with ALBEDOS of a few per cent. Comparison of the occultation profiles of the rings at different wavelengths reveals that they contain very few particles smaller than centimeter sizes, with particle sizes tens of centimeters having been inferred in the ε ring. A Voyager image of the rings (in forward scattered sunlight) shows a small population of micron-sized dust particles in the ring system.

Compared with the other ring systems, Uranus' rings have less material than those of Saturn, but more than the ring systems of Jupiter and Neptune (see JUPITER: RINGS NEPTUNE: RINGS). All the Uranian rings lie inside the ROCHE LIMIT, which is the smallest radius for which a body can remain intact by its self-gravitational forces alone (without being pulled apart by the tidal forces from Uranus). For a satellite (such as Cordelia and Ophelia) to exist within this region, it must have sufficiently high density and/or tensile strength to prevent the tidal forces from Uranus breaking it up. In fact, each of the rings could have formed in this manner. Once formed, a ring is subject to erosion processes, such as radiation drag, meteoroid impact and drag from Uranus' exosphere. One can then estimate how long these processes could have been operating in the past, which sets limits on the age of the rings. Estimates along these lines suggest ring ages of a few $\times 10^6$ years—much less than the age of the solar system (4.5×10^9 yr).

Bibliography

Elliot J and Kerr R 1984 *Rings: Discoveries from Galileo to Voyager* (Cambridge, MA: MIT Press)

Elliot J L and Nicholson P D 1984 The rings of Uranus *Planetary Rings* ed R Greenberg and A Brahic (Tucson, AZ: University of Arizona Press) pp 25–72

Esposito L W, Brahic A, Burns J A and Marouf E A 1991 Particle properties and processes in Uranus' rings *Uranus* ed J T Bergstrahl, E D Miner and M S Matthews (Tucson, AZ: University of Arizona Press) pp 410–65

French R G, Nicholson P D, Porco C C and Marouf E A 1991 Dynamics and structure of the Uranian rings *Uranus* ed J T Bergstralh, E D Miner and M S Matthews (Tucson, AZ: University of Arizona Press) pp 327–409

J L Elliot

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Urey, Harold Clayton (1893–1981)

Chemist, born in Walkerton, Indiana, Nobel prizewinner for Chemistry in 1934 'for his discovery of heavy hydrogen'. It was at Columbia University that he isolated the isotope deuterium by distilling liquid hydrogen; in the Second World War, he directed the effort to separate uranium-235 from uranium-238 for the atomic bomb. At the University of Chicago, he worked on the origin of the elements, their abundance in stars, and the origin of the planets, including the chemical properties of the Earth. He invented a technique that used oxygen isotope-bearing minerals to date rocks and to measure paleohistoric water temperatures (see MILUTIN MILANKOVITCH). He analysed lunar rocks from the Apollo missions.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ursa Major

(the Great Bear; abbrev. UMa, gen. Ursae Majoris; area 1280 sq. deg.) a northern constellation which lies between Draco and Leo Minor–Leo–Canes Venatici, and culminates at midnight in mid March. Its origin is uncertain, though it was known to the ancient Greeks, who identified it with two mythological figures—Callisto, a mortal who was turned into a bear after having fallen victim to Zeus's passion and whom he placed for safety in the sky, and Adrasteia, a Cretan nymph who raised the infant Zeus and whom in gratitude he placed among the stars. Its brightest stars were cataloged by Ptolemy (c. AD 100–175) in the *Almagest*.

The third-largest and probably the best known of the constellations, Ursa Major is easily recognized by the asterism of the Plough or Big Dipper, formed by seven of its brightest stars, α, β, γ, δ, ϵ, ζ and η, which since ancient times has been used for navigation, as a line drawn from β through α Ursa Majoris points to α Ursae Minoris (Polaris), a second-magnitude star that lies within 1° of the north celestial pole. The brightest stars of Ursa Major are ε Ursae Majoris (Alioth), magnitude 1.8, α Ursae Majoris (Dubhe), a very close binary with orange (K0) and pale yellow (F0) components, magnitudes 2.0 and 4.9 (combined magnitude 1.8), separation 0.7″, period 44.4 years, η Ursae Majoris (Alkaid or Benetnasch), magnitude 1.9, ζ Ursae Majoris (Mizar), a multiple system consisting of two white (A2 and A7) components, magnitudes 2.3 and 3.9, separation 14.4″, both of which have a fainter companion, which forms a wide optical double with 80 Ursae Majoris (Alcor), magnitude 4.0, separation 11.8′, β Ursae Majoris (Merak), magnitude 2.3, and γ Ursae Majoris, magnitude 2.4. There are 13 other stars of magnitude 4.0 or brighter. The five central stars of the Plough (β, γ, δ, ϵ and ζ Ursae Majoris) are part of the Ursa Major Moving Cluster, which is the closest star cluster to the Sun, its center being about 75 light-years distant.

Another interesting multiple star system is ξ Ursae Majoris (Alula Australis), which consists of two yellow (G0) components, magnitudes 4.3 and 4.8 (combined magnitude 3.8), separation 0.9–3.1″, period 59.8 years, the former of which has an unseen companion which revolves around it in 669 days, and the latter two unseen companions, one of which revolves around it in 3.98 days. The main components were the first binary system to have an orbit computed, by M Savary in 1828. Other interesting stars include Lalande 21185 (magnitude 7.5), which, at a distance of 8.3 light-years, is the fourth closest star to the Sun, and Groombridge 1830 (magnitude 6.4), which has the third largest proper motion of any star.

Other interesting objects include M81 (NGC 3031), a seventh-magnitude spiral galaxy, its near neighbor M82 (NGC 3034), an eighth-magnitude irregular, starburst galaxy which is thought to have experienced disruption following a near-collision with the much more massive M81 some 40 million years ago, and M101 (NGC 5457), an eighth-magnitude face-on spiral galaxy.

See also: Groombridge 1830, Mizar and Alcor, Plough.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ursa Minor

(the Little Bear; abbrev. UMi, gen. Ursae Minoris; area 256 sq. deg.) a northern circumpolar constellation which is surrounded on three sides by Draco and includes the north celestial pole. Its origin is uncertain, though it was known to the ancient Greeks, who identified it with Ida, a Cretan nymph who helped her sister, Adrasteia, raise the infant Zeus and whom in gratitude Zeus placed among the stars alongside Adrasteia (who is represented by the neighboring constellation of Ursa Major). Its brightest stars were cataloged by Ptolemy (*c*. AD 100–175) in the *Almagest*.

A small, rather inconspicuous constellation, Ursa Minor's only claim to fame is that its brightest star, α Ursae Minoris (Polaris or Alrucaba), lies within 1° of the north celestial pole. α Urase Minoris is actually a triple system with pale yellow (F7 and F3) components, the primary of which is a Cepheid variable (range 1.86–2.13 decreasing, period 3.97 days) and the secondary a star of magnitude 8.2, separation 18.4″, and a third, unseen component which revolves around the primary in 30.5 years. There are two other stars brighter than magnitude 4.0: β Ursae Majoris (Kochab), magnitude 2.1, and γ Ursae Majoris (Pherkad), magnitude 3.0, which together are known as the 'Guardians of the Pole'. The seven brightest stars of the constellation (β, γ, η, ζ, ε, δ and α) form an asterism similar to the Plough and are sometimes called the Little Dipper.

There are no bright star clusters, nebulae or galaxies in the constellation, the brightest being NGC 6217, an eleventh-magnitude spiral galaxy.

See also: Polaris.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

US Astronomy

As the 19th century ended, astronomy underwent a period of rapid growth in the United States, a growth that was fueled by both the expansion of the university system and private philanthropy and which also paralleled the growth in astrophysical research. For the first half of the 20th century, the US government took little interest in the funding of astronomical research, concentrating on those aspects of astronomy that were of greatest commercial and military interest: navigation and the determination of time.

Following World War II, the involvement of the US government in funding all branches of science grew and the National Science Foundation (NSF) established and built the national observatories. With the birth of space astronomy in the 1960s, the National Aeronautics and Space Administration (NASA) became a major force in astronomical research. Despite an increasing government presence, there remained a steady source of private and institutional funds contributing to the operations of astronomical facilities and to the support of research. Together with this growth in funding came steady growth in the number of practicing research astronomers in the US, a number which grew more than six-fold in the three decades from 1960 to 1990.

Funding sources

The dominant portion of US government funding for astronomy is provided through two agencies, NASA and the NSF. Other agencies supporting astronomical research include the Department of Defense, the Department of Energy and the National Oceanic and Atmospheric Administration.

NASA currently provides about 80% of the total federal funding, with most of the balance provided by the NSF. NASA funds astronomy through the Office of Space Science (OSS), which had a budget of 2.1 billion dollars in fiscal year (FY) 1999. This office funds the planning, construction, launch and operation of space-based astronomical observing facilities. It also provides funding for the archiving and analysis of data from space missions. Furthermore, the OSS operates some high-altitude observing facilities, which fly on balloons or airplanes.

The NSF division of astronomical sciences (AST) operates within the Mathematical and Physical Science directorate of the agency. Its budget for FY 1999 was 119 million dollars. The NSF operates several national observing facilities, directly funds astronomical researchers and provides some funding for educational purposes. The NSF also provides funding for the design, development and construction of national ground-based observational facilities. In FY 1999, 8 million dollars was allocated for the design and development phases of the ATACAMA LARGE MILLIMETER Array (ALMA, formerly known as the MMA or millimeter array), an international collaboration.

There are numerous sources of non-federal funding for astronomy. These include private organizations, private universities, state universities, state and local governments and private foundations. The amount of non-federal funding for astronomical research is difficult to determine owing to the broad array of sources. Universities typically provide a significant level of support through salaries for faculty members. Private universities such as the California Institute of Technology or Harvard University as well as public universities such as the University of California or the University of Texas also use institutional funds to support observatory operations. Among private organizations, the Carnegie Institution of Washington and the Lowell Endowment provide ongoing operating funds for specific observatories while other groups such as the Keck Foundation, the Sloan Foundation and the Ford Foundation have provided substantial funding for the construction of observatories or for specific research projects.

Facilities

To effectively study the diverse array of objects in the universe, observations must be made using all wavelengths of electromagnetic radiation. Some wavelengths of light are not able to penetrate the Earth's atmosphere. Some methods of observation require extensive instrumentation. Some direct physical data are gleaned from analysis of collected cosmic dust particles or meteoritic material. For these reasons, observations must take place both from the surface of the Earth and from space.

In the United States, observatories are funded by the Federal Government and through private sources. The NSF owns the facilities of and contracts for the operations of the National Observatories. These are as follows:

- the National Optical Astronomy Observatories, with observatories on KITT PEAK, Arizona, CERRO TOLOLO, Chile, and Sunspot, New Mexico;
- The NATIONAL RADIO ASTRONOMY OBSERVATORY, with observatories at Green Bank, West Virginia, Socorro, New Mexico, and Kitt Peak, Arizona;
- the National Astronomy and Ionosphere Center at ARECIBO, Puerto Rico.

In addition:

- The NSF provides operational support to a number of university radio astronomy facilities and the NSF Office of Polar Programs, which operates several astronomical facilities at the South Pole,
- the NSF is a member of an international optical telescope partnership named GEMINI, which has built two identical telescopes, one in the northern hemisphere and one in the southern hemisphere (the United States funds approximately 48% of the overall budget; see GEMINI OBSERVATORY).

The non-federally funded observatories are mainly supported through universities, consortia of universities

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or private foundations. Many observatories operate with mixed sources of funding. For example, some observatories are constructed using university funds or with grants from private foundations and then rely on grants from the federal agencies to provide funds for development and construction of instruments and sometimes for operations.

In the last 10 yr, NASA has launched three of the four so-called 'Great Observatories'. These include the HUBBLE SPACE TELESCOPE, the CHANDRA X-RAY OBSERVATORY and the COMPTON GAMMA-RAY OBSERVATORY. The final Great Observatory, the SPACE INFRARED TELESCOPE FACILITY, is scheduled for launch in December 2001. These satellites have radically changed our view of the universe and provided unprecedented discoveries unattainable from the ground.

In addition to the Great Observatories series, NASA has a range of other missions from small to large. There are several major mission programs including Discovery, Explorer, Mars Exploration and New Millennium. Each mission has specific goals. Discovery missions explore objects within our solar system such as planets, moons and other small bodies. The Explorer program provides frequent flight opportunities at three separate funding levels for missions that support one of the four major themes of the Office of Space Science: university (less than $7.5 million), small (less than $71 million) and medium (less than $140 million). These themes are the Astronomical Search for Origins, Planetary Systems, The Sun–Earth Connection and the Structure and Evolution of the Universe.

NASA maintains 15 research centers across the United States, each of which fulfills various portions of NASA's overall mission. The centers most closely associated with astronomy missions include the GODDARD SPACE FLIGHT CENTER in Greenbelt, Maryland, the JET PROPULSION LABORATORY in Pasadena, California, and MARSHALL SPACE FLIGHT CENTER in Huntsville, Alabama.

NASA has led the way in archiving astronomical data. It has created numerous facilities such as the Hubble Data Archive, the Astronomical Data Center, the Planetary Node system and the Astrophysics Data System, which provide researchers access to archival data.

There has been a recent trend toward major projects, whether on the ground or in space, being undertaken through international collaborations because of the large costs and complexity of the projects. This trend is likely to continue into the near future.

Demographics and employment patterns

After a decade of rapid growth in the 1980s, the number of astronomers in the United States has grown more slowly but steadily in the 1990s. The median age of astronomical researchers has remained approximately the same, but the age distribution has widened somewhat with the number of astronomers in the age range from 35 to 50 decreasing from about 50% at the end of the 1980s to 40% at the end of the 1990s.

Toward the end of the 1990s there has been a leveling off and even a decrease in the number of students pursuing studies in physics and astronomy at the undergraduate and graduate levels in American universities. Almost all graduate students in astronomy or astrophysics have prepared themselves for graduate study by obtaining an undergraduate degree in either physics or astronomy. Among astronomy PhD recipients from US institutions in 1997, 19% were female and 27% were not US citizens (source: American Institute of Physics, *1997 Graduate Student Report*).

About six out of every ten astronomers in the United States are employed in academic institutions, colleges or universities, and slightly over half of these are in tenured or tenure-track positions. Most of the others are working in federal laboratories or research centers or for industrial contractors supporting these centers.

Research publications

The bulk of the research published in the United States appears in one of the four scholarly journals published by the astronomical societies. The AMERICAN ASTRONOMICAL SOCIETY (AAS) publishes *The Astrophysical Journal*, *The Astrophysical Journal Supplement Series* and *The Astronomical Journal*. The ASTRONOMICAL SOCIETY OF THE PACIFIC (ASP) publishes *The Publications of the Astronomical Society of the Pacific*. In addition, there is a section of *The Physical Review* (published by the American Physical Society) devoted to astrophysical research, and both *Icarus* (Academic Press) and *The Journal of Geophysical Research* (American Geophysical Union) publish research in planetary and space physics.

Annual Reviews, Inc., publishes a review volume, *The Annual Review of Astronomy and Astrophysics*. The Astronomical Society of the Pacific and the American Institute of Physics are the principal publishers of conference proceedings.

As technology has developed in recent years, there has been an increasing use of on-line preprint posting services for rapid dissemination of research results. The most popular of these is part of the arXiv e-print archive service hosted at the Los Alamos National Laboratory (http://xxx.lanl.gov/archive/astro-ph).

Amateur astronomers in the US

Astronomy appeals to nearly everyone. Since the sky is accessible to all, astronomy is blessed with a large and active amateur community. The US has a number of amateur organizations that serve many diverse purposes. These range from small local clubs that meet regularly to exchange advice on telescopes and host 'star parties' to large national organizations that coordinate research activities easily performed by amateurs. Some of the larger national organizations that include amateurs are listed in table 1.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Larger national organizations.

Organization	Purpose	Webpage
AMERICAN ASSOCIATION OF VARIABLE STAR OBSERVERS (AAVSO)	Variable star research	www.aavso.org
Astronomical Society of the Pacific	Education, outreach	www.aspsky.org
Astronomical League	Observation	www.astroleague.org
INTERNATIONAL DARK SKY ASSOCIATION	Light pollution control	www.ida.org

Summary

Astronomy in the United States is ending a period that has lasted a few years more than a century and has seen the US emerge as a world leader in astronomical research. At the beginning of the 20th century, the US was leading observational astronomy into an era of increasingly powerful and sophisticated telescopes. The 40 in refractor had just been commissioned and the two powerful telescopes on MOUNT WILSON were in the near future. When these were put to work, our view of the universe changed completely. Edwin HUBBLE explored the extragalactic distance scale and we began to understand the expansion of the universe. These discoveries, in turn, spurred the construction of newer, more powerful telescopes such as the 200 in Hale telescope. Following the discovery of radio waves from cosmic sources, powerful instruments such as the Very Large Array were commissioned (see VERY LARGE ARRAY NATIONAL RADIO ASTRONOMY OBSERVATORY). The United States led the way in space beginning in the 1960s and opened new and previously unexplored regions of the electromagnetic spectrum for astronomical observations. A series of increasingly ambitious and complex spacecraft were built and operated, leading finally to NASA's Great Observatory Series, which reaches completion in the early years of the 21st century.

At the beginning of the 21st century, American astronomers have access to an unprecedented suite of instruments and techniques with which to observe the heavens. New technology has made possible the construction of telescopes of a size that was inconceivable a generation ago, and the Keck Telescopes, at 10 m each, have led the way in showing what can be achieved with such powerful instruments. The successors to the NASA instruments of the 1990s are being planned even while the current generation is at its peak of performance.

Kevin B Marvel and Robert W Milkey

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

US Naval Observatory

The US Naval Observatory is the oldest astronomical observatory in the United States, and the oldest continuously operating scientific institution in the US government. Founded in 1830 as a Depot of Charts and Instruments for rating chronometers and maintaining navigational instruments, by 1844 it had become the first national observatory of the United States, analogous to the ROYAL OBSERVATORY IN GREENWICH, England. The observatory's headquarters are located in Washington, DC, including the historic 0.66 m refractor used to discover the two moons of Mars in 1877. The observatory's largest telescope, a 1.55 m astrometric reflector, is located at its station in Flagstaff, Arizona. The observatory's mission has always been to aid in the improvement of navigation as well as to conduct basic research in astronomy. Today, it provides the national time service for the United States, determines the precise positions and motions of celestial bodies, measures Earth rotation parameters including polar motion, and produces a variety of almanacs for use by astronomers, navigators and the general public.

For further information see
http://www.usno.navy.mil/

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ussher, James (1581–1656)

The Archbishop of Armagh and Primate of All Ireland, Ussher was a churchman and a scholar. He correlated Middle Eastern and Mediterranean histories with Jewish genealogies of the Old Testament, and the resulting chronology was incorporated into the Authorized Version of the Bible of 1701. The chronology established the year of creation as 4004 BC. Lightfoot followed Ussher, improving on Ussher's accuracy by declaring that the Earth had been created at 6:00 a.m. on 26 October 4004 BC. The dates of other biblical events followed, for example, that Adam and Eve were driven from Paradise on Monday 10 November 4004 BC, and that the Ark landed on Mount Ararat on 5 May 1491 BC, 'on a Wednesday'. Ussher was a great scholar, if over-literal in his analysis of the Bible. He collected the earliest available manuscripts into his library, which formed the nucleus of the great library of Trinity College, Dublin.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Valles Marineris

A complex system of canyons on Mars, centered at 11.6 °S, 70.7 °W, stretching for a total of 4128 km in the east–west direction just south of the equator, and reaching depths of over 6 km. It is named after the Mariner probes which returned the first close-up images of the planet's surface, and is also referred to as Mariner Valley. Individual sections of the systems are termed chasmata. At its western end, where the system abuts the faulted area at the east of the Tharsis Bulge known as Noctis Labyrinthus, lie Tithonium Chasma and, to its south, Ius Chasma. To their east, the system expands into the three parallel canyons named, from north to south, Ophir Chasma, Candor Chasma and Melas Chasma. To the northwest lies the unconnected Hebes Chasma. The main system continues eastward with Coprates Chasma, and then opens out, via Capri Chasma to the north and Eos Chasma to the south, into the so-called chaotic terrain of the region known as Margaritifer Sinus.

Valles Marineris was created largely by faulting, but other forces have been at work. The deep branching valleys running into the southern edge of Ius Chasma suggest erosion by groundwater, while at the eastern end are teardrop-shaped islands suggestive of flowing water. Close up views of Coprates and the other central chasmata clearly reveal the presence of layered sediments, which could only have been deposited under water.

See also: Mars: surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Valongo Observatory

The Valongo Observatory, part of the Federal University of Rio de Janeiro (UFRJ), sited in downtown Rio de Janeiro, Brazil, was founded in 1881. Its main activity is scientific research in the field of astronomy, carried out by its staff of twelve astronomers. The professors/researchers keep scientific collaboration programs with other institutions at both national and international levels. The main research areas are: fundamental astronomy, stellar astrophysics, interstellar medium, extragalactic astronomy and laboratory astrophysics. Funding comes chiefly from the Brazilian government agencies. The Valongo Observatory offers an undergraduate course in astronomy, attended by approximately 90 students. Its astronomical instruments include an original turn of the century Thomas Cook & Sons 30 cm refractor telescope, and a Zeiss reflector of 15 cm, for undergraduate teaching. As extracurricular activities, there is a visitation program offered to the general public, also serving secondary schools, which features telescopic observations and conferences given twice a month.

For further information see
http://www.ufrj.br/ov.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Van Allen Belts

Two toroidal regions surrounding the Earth which contain trapped charged particles, discovered by James Van Allen.

See: Magnetosphere of Earth: radiation belts.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Van Allen, James (1914–)

Magnetospheric physicist, born in Mount Pleasant, Iowa, discovered the *Van Allen belts* that surround Earth. After war-time service, Van Allen used V-2 rockets on high-altitude experiments, specifying a replacement rocket, the Aerobee, when supplies of the V-2 ran out. During the International Geophysical Year (1957–8), the first American satellite, Explorer 1, carried a micrometeorite detector and a cosmic ray experiment designed by Van Allen. Data from Explorer 1 and Explorer 3 (launched 26 March 1958) revealed the existence of a doughnut-shaped region of charged particle radiation trapped by Earth's magnetic field. Later in 1958, Pioneer 3 led to the discovery of a second radiation belt.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

van de Kamp, Peter (1901–95)

Astronomer, born in Kampen, Netherlands, became director of the Sproul Observatory and professor at Swarthmore College, PA. He worked on the astrometric measurement of proper motions of stars across the sky and discovered oscillations in their paths. He interpreted the deflections as due to the revolution of planetary systems around the stars. The oscillations have not been confirmed, and many astronomers think that they were instrumental effects caused by the telescopes and measurement techniques.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

van Maanen, Adriaan (1884–1947)

Dutch astronomer, became a member of the Mount Wilson staff, and studied the rotation of spiral nebulae as a means to establish their distances. His detection of rotational motions was illusory and misleadingly suggested that they were relatively nearby. Through its proper motion and parallax, he discovered the white dwarf *van Maanen's star*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Van Maanen's Star

This star in the constellation of Pisces, although faint, is of much importance as it is the nearest readily observable white dwarf—a class of star of considerable significance in studies of stellar evolution. The white dwarfs Sirius B and Procyon B are both closer, but their proximity to their much brighter companions makes them very difficult to study as individual stars.

Van Maanen's Star, also known by the catalog designation Wolf 28, is situated about 2° south of δ Piscium. It was discovered in 1917 by the Dutch-American astronomer Adriaan Van Maanen (1884–1946), from a comparison of plates exposed in 1914 and 1917, to have the large proper motion of 2.978″ per annum. Its spectrum is classified DG. With a parallax of 0.227″, it is only 14.4 light-years distant, yet its apparent magnitude is only 12.37, indicating that it has a very low intrinsic luminosity (its absolute magnitude is 14.1). It has been estimated that its diameter may be no more than 12 500 km, comparable to that of the Earth.

See also: white dwarfs.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Van Vleck Observatory

The Van Vleck Observatory on the campus of Wesleyan University in Middletown, Connecticut, was built in the years 1914–16 upon the disassociation of the departments of mathematics and astronomy. It was named in memory of John Monroe Van Vleck, Professor of Natural Science at Wesleyan throughout much of the nineteenth century.

The primary instrument of the observatory is a visual refracting telescope of 0.5 m aperture and a focal length of 8.41 m. The crown and flint lenses were made by C A R Lundin of the Alvan Clark Co. and were installed on a Warner and Swasey mounting in 1992. The two lenses are separated by about 10 cm, allowing cleaning of all surfaces without disturbing them. Their alignment has thus not been altered, and this and other features make the refractor among the best for astronomical research. Since 1992, the telescope has been mainly used to determine trigonometric parallaxes and proper motions of many hundreds of faint nearby stars. In the last few decades the Van Vleck astrometric program has been one of the leaders in the use of parallaxes to respond to problems in stellar astrophysics.

In 1971, a 0.6 cm reflecting telescope was obtained from the estate of Richard Perkin, founder of the Perkin-Elmer Corp. Like the refractor, the on-campus site of this instrument has encouraged a monitoring program using photoelectric photometry and later CCD imaging capability, of properties of T Tauri variables and other young stars. The department of astronomy offers the BA and MA degrees, and research programs on both telescopes feature student participation at every level.

For further information see
http://www.astro.wesleyan.edu/astro.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Variable Stars

Stars vary in many ways and for many reasons. Broad categories of variation include stars that pulsate, fluoresce, become veiled, are disturbed by companions, erupt, explode and even change their spots. Variations may be episodic, irregular, semi-regular or periodic. A separate wide class comprises the extrinsic variables such as eclipsing and tidally distorted binaries and gravitationally lensed stars, whose variations result from changing aspect, but here we survey the intrinsic variables.

The first recognized variable stars were novae and supernovae ('new' stars), with some very old records being astrophysically useful, chiefly because they fix the accurate date of a stellar explosion. Although the ancient accounts are obscure and fragmentary and do not recognize modern distinctions (such as nova versus supernova), the type of event may still be discernible. For example, a supernova can be visible for 6 months or even longer, whereas ordinary novae are prominent for a matter of weeks, so a nearby nova can mimic a distant supernova in brightness but is typically a much briefer event. Observations of exploding stars go back to several centuries BC with records of some objects from two or more places, and would go back further if not for library burnings. The records were kept by imperial astrologers in China (later in Japan and Korea) as the 'new' stars were considered important omens. Today's telescopic observations provide good statistics of novae, but useful statistics of the infrequent supernovae in our Galaxy rely on the old records. Events observed in other galaxies have demonstrated enormous differences in scale between supernovae and novae. Theory has shown those differences to be far greater than optical observations indicate, as supernovae release the vast bulk of their energy in invisible forms.

Except for novae, stars were regarded as constant until the mid to late 1600s, when the first non-eruptive variables were recognized. Most notable was the pulsating red giant *o* Ceti (Mira, the Wonderful) with its $\approx 330^d$ large-amplitude variations. An appearance and disappearance of Mira had been noticed by D Fabricius in 1596, but was taken to be a nova. The early histories of intrinsic and extrinsic variables are inextricably linked, as ideas about causes emerged only slowly over several centuries. Systematic recording replaced casual notes over the 1700s, with the improved records leading to further discoveries. Thus θ Serpentis was conclusively shown to be variable by E Pigott, confirming G Montanari's suspicions of a century earlier, and J Goodricke found periodic variation in δ Cephei, the prototype of *Cepheid variables*, and in the eclipsing binary β Lyrae. The improved records also led to some success in identifying variation mechanisms. Speculation of the late 1700s cycled among eclipses, moving and transient spots on rotating stars, and rotating distorted stars. All of these phenomena commonly occur, but early assignments to particular stars were usually wrong. Pulsation was advocated only much later. Algol was (correctly) said to eclipse by Pigott and Goodricke, but both later favored other ideas. Curiously, the eclipse hypothesis of Algol was cast into disrepute by discovery of variable stars that clearly were not eclipsing binaries, apparently from a wish to have all variables follow a single theory. Spots on rotating stars were very popular in the 1700s and 1800s, even for Mira, and have now returned for RS Canum Venaticorum, W Ursae Majoris, FK Comae and BY Draconis type binaries. As late as *c*. 1930, pulsating star velocity curves were fitted in terms of orbital parameters, even when there was little or no doubt about their pulsational origin. A common finding was that the orbit of a star's unseen (and non-existent) companion was actually inside the observed star. The orbit parameters would be printed anyway after a comment or two on the unphysical situation and the need to fit something. Although the reality of pulsation for Long Period Variables (LPVs) and Cepheids was not seriously disputed, there was much uncertainty in sorting true pulsators from a miscellany of competing phenomena. Most authors simply avoided explicit mention of pulsation while discussing effects that scarcely could be due to anything else, with an occasional remark that velocity variations might not really be due to line of sight motion of any kind. The popularity of pulsation was undeniable in the 1920s and 1930s, however, when up to 10% of *Astrophysical Journal* papers concerned Cepheids and related stars—remarkably high for one subject. Nearly all contributions were observational at that time, although major theoretical advances had already been made by A Ritter in the 1870s and 1880s and by A S Eddington in the 1920s. Full acceptance of spectroscopy's implications for the pulsation hypothesis took nearly a half century (*c*. 1890–1935), yet many kinds of pulsating stars were eventually recognized as a direct consequence of the radial velocity and temperature changes discovered spectroscopically. Not only the existence of radial motion, but its phasing compared with that of light curves, is a crucial discriminant among models.

Realization that large numbers of variable stars exist was slow to develop until the breakthrough of photography. Only ≈10 non-exploding variables were known in the early 1800s, but the permanent records in photographic plates quickly extended the list into the hundreds between 1850 and 1900. Systematic photographic surveys in the 1900s at Harvard, Bamberg, Leiden and other observatories raised catalog listings into the tens of thousands. Recent automated gravitational lens surveys with electronic detectors have produced light curves of thousands of variables as a by-product of the surveys' primary objective, thereby substantially

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

increasing the discovery rate. The *General Catalog of Variable Stars* ('GCVS', Kholopov 1984) is a four-volume listing of coordinates, types and other information on 28 211 variables of all kinds, nearly 70% being intrinsic variables. A companion catalog of *New Suspected Variables* ('NSV', Kholopov 1982) contains 14 811 stars of uncertain status.

A reasonably complete list of variable star types with minimal descriptions would exceed available space, but the GCVS devotes five pages to descriptions. It recognizes 88 types of intrinsic variables, including 34 pulsating, 7 rotating, 10 x-ray, 22 eruptive and 15 cataclysmic types (cataclysmic and other eruptive variables are kept distinct). A standard naming convention allows immediate recognition of a star as variable, although some bright variables have only their Bayer names. To begin the sequence, a single capital letter is assigned alphabetically in order of discovery from R to Z, then double letters RR, RS, …, RZ, then double letters AA, AB, … to AZ, then BB, BC, … to BZ, and ultimately to ZZ, followed by the constellation name (e.g. T CrB; ZZ Cet). Letter J is not used. Designations of the form V335 Cygni, V336 Cygni, etc. follow the 334 names of the letter system.

There is great interest in variable stars among amateur astronomers, whose observations are indispensible for professionals. Not only visual brightness estimates but phototube and charge-coupled device (CCD) measures are now made available in large quantities through bulletins, journals and especially the Internet. Variable star organizations include the American Association of Variable Star Observers (AAVSO), the Association Française des Observateurs d'Etoiles Variables (AFOEV), the British Astronomical Association Variable Star Section (BAAVSS), Bundesdeutsche Arbeitsgemeinschaft für Veränderliche Sterne (BAV), the Center for Backyard Astrophysics (CBA), the Hungarian Astronomical Association, International Amateur-Professional Photoelectric Photometry (IAPPP), the Variable Star Group of the Astronomical Society of Southern Australia, Variable Star Network (VSNET), and the Variable Star Section of the Royal Astronomical Society of New Zealand. Information on all these organizations can be found via the Internet. Their imprints on professional astronomy date at least as far back as the 1920s—for example in Joy's use of an AAVSO Mira light curve in the 1926 *Astrophysical Journal*. The role of such organizations is rapidly expanding today. The central professional organization is Commission 27 of the International Astronomical Union.

LPVs and Cepheid variables were not only discovered early but still dominate our catalogs. LPVs are pulsating red giant stars of great luminosity and are thereby rather easy to discover. The name *Mira type* is reserved for the LPVs of larger light amplitudes, although there probably is continuity in the overall class, including Galactic distribution, space velocity and composition as well as light variation. There may also be continuity with 'non-variable' red giants, which probably have some variation. Spectra of LPVs and especially Miras show prominent emission lines that are clearly due to pulsation, as non-variable red giants lack emission lines but otherwise have similar spectra. More specifically, pulsation amplitude correlates with emission line strength. Miras vary by up to 9^m in blue light (a factor of ≈4000), but the enormous optical variations are caused by emission being ultra temperature-sensitive, with the optical band lying far to the short-wavelength side of the spectral energy peak. Bolometric amplitudes are similar to those of Cepheids. A problem for spectroscopy, especially prior to modern efficient spectrographs and large telescope apertures, was the faintness of Miras near minimum light. Observations must cover full cycles if much is to be learned about pulsation, and it was only in 1926 that A Joy produced the first complete velocity curve of Mira. The classical Cepheids are highly luminous F and G type pulsating giants and supergiants that vary periodically up to roughly a magnitude. Periods range from a few days to over 100 days. Rather early it was realized that the important relation between period and mean density

$$P\sqrt{\rho} = \text{constant}$$

applies to Cepheids and several other prominent categories of pulsating stars. The relation exists because a star's dynamical timescale depends on its size and mass, with some dependence on internal structure (thus a different constant for each category). Stars with given HR diagram coordinates have the same size, but not necessarily the same mass and structure, so pulsation periods provide a link to structure and thereby to understanding evolution. The W Virginis stars, members of a very old Galactic population, pulsate with their own $P\rho^{1/2}$ relation and are understood to be in a later evolutionary stage than classical Cepheids, having lost much of their original envelopes. Thus pulsation is an important evolutionary diagnostic that can even help trace Galactic origins.

In addition to the essentially periodic and roughly periodic classes of pulsating stars, there are giants and supergiants of spectral classes F to M that pulsate irregularly or semi-regularly. Curiously they lie mainly *between* the more regular Cepheids and LPVs in the HR diagram. The less regular and typically redder of these are known as RV Tauri stars, while the somewhat more regular ones are simply called yellow semi-regular variables. Periods, in so far as they can be quantified at all, range from tens to thousands of days, and amplitudes can be up to about four magnitudes. These stars may be interesting in terms of chaos theory, as some show

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

alternating cycles of larger and smaller amplitude that hint of the period doubling seen in the approach to chaos.

Brightness measurement

We have astrophysically useful records of the brightness variations of the supernovae of 1572 and 1604, as observed by T Brahe and others and by J Kepler respectively, despite lack of light detection equipment and even of telescopes 400 years ago. Their usefulness illustrates two points about astronomical brightness measurement. First, although precision requirements are stringent for small variations, they can be relaxed for large variations. Second, in no case can we allow the scale to be wrong, unless mere detection of variability is the only objective. One might expect eye estimates to be quantitatively useless, but Brahe's and Kepler's light curves demonstrate otherwise. The situation has been saved by the 'secret' advantage of the photometric astronomer—the remarkable long-term constancy of normal stars. Brahe and Kepler recorded their supernovas' magnitudes relative to known stars that, as far as we know, are essentially unchanged in brightness. Modern photoelectric observations of the reference stars have accordingly placed the supernova estimates on an objective scale. Of course the old estimates are subjective, but the largest overall errors are limited by the magnitude spacings of reference stars. An estimation method involving magnitude steps, introduced by F Argelander about 1840, further diminished maximum errors.

The existence of natural constant brightness standards largely accounts for the accuracy of astronomical photometry, which is typically better than that of laboratory photometry despite the disadvantage of observing through Earth's irregular atmosphere. Astronomers also have an advantage in the temperatures of reference sources, as very hot laboratory standards would vaporize their surroundings, while hot stars have no surroundings to vaporize. Difficulties due to daylight and weather as well as complete blockage of spectral regions are best dealt with by observing from space, although many less severe problems have been minimized through a variety of clever techniques.

The history of star brightness measurement can roughly be divided into an early era of qualitative notes (*c*. 1600 to *c*. 1800), a middle era of quantitative but inaccurate estimation by eye or photography (beginning *c*. 1800 and still somewhat active), and a recent era of accurate quantitative measurement with various electronic detectors (beginning *c*. 1910). Even the early qualitative work was a major improvement on the unplanned occasional notices of preceding times. It made possible the discovery of *periodic* variation and, coupled with long baselines in time, even accurate periods. Thus Goodricke estimated Algol's period in 1784 with an uncertainty of less than a minute out of 2.87 days. Entering the middle era, note that the eye can reliably judge brightness equality between two stars, a point well appreciated and exploited by F W Herschel, J Herschel, Argelander, J Hartmann and K Schwarzschild, but eye estimates of inequality are notoriously subjective. The better visual work accordingly employed a variety of now obsolete ways to reduce the apparent brightness of one star by a known factor so as to equal that of another. Techniques, sometimes applied also to photographic observing, included partial aperture blocking with twin telescopes, extra-focal imaging, tapered neutral filters (wedges), and crossed polarizers. Photographic magnitudes were extracted from diameters of focused images or blackening of unfocused images. The modern definition of astronomical magnitudes, relating magnitude (m) differences to light (l) ratios, $m_1 - m_2 = -2.5 \log (l_1/l_2)$, was suggested in 1850 by N Pogson. Simple as it may seem, Pogson's contribution was crucially important, as the quantitative meaning of magnitudes had previously been vague. The new accuracy and increased time lines allowed period *changes* to be measured, with S Newcomb's text of 1884, *Astronomy*, already mentioning period changes for Algol, Mira and β Lyrae. The above-mentioned accuracy tricks work only within small fields, so measurement of actual magnitudes, as opposed to 'local' magnitude differences, received a large boost with the arrival of photoelectric detectors. Invention of the photomultiplier tube in the 1940s especially helped to surmount atmospheric attenuation problems via all-sky photometry, and also greatly improved accuracy of magnitude changes for variable stars. Development of modern standard magnitude systems began when photomultipliers came into widespread use. Examples include the *U*, *B*, *V*, *R*, *I*, *J*, *K*, *L* wide-band systems and the *uvby* medium-band systems. A *standard star* has accurately known constant magnitude in one or more standard systems (e.g. γ Oph has $V = 3^m.72$, $B = 3^m.76$), while a *comparison star* has (presumably) constant but not necessarily known magnitude. CCD detectors similar to those in video cameras are now replacing photomultipliers because of their ability to measure many sources simultaneously and thus provide for accurate differential photometry. New practical problems have arisen out of the enormous quantities of data coming from CCD photometry. Advanced computer technology is being applied to storage and distribution problems, in some cases with data collection directly from the detecting equipment. Automatic telescope operation is becoming ever more widespread in many spectral regions, not only in discovery modes (as in extra-Galactic supernova searches) and quick follow-up on γ and x-ray transients, but also in routine observation of known variables. The acronym APT, for Automatic Photometric Telescope, implies remotely operated, programmable instruments that are primarily used for variable star research.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Astrophysical importance of variable stars

The scientific value of variable stars can be seen in two issues—why they vary (theory) and what their variations allow us to measure (observation). The theoretical point connects physical makeup with behavior, while the observational one concerns properties such as size, distance and internal structure that are measurable or estimable because of variation.

Most astronomical distance estimates are based on the *standard candle* method, a comparison of apparent and intrinsic brightness. The extra-Galactic distance scale is an important application, and distances to multiple stars and clusters within our Galaxy also are found from embedded standard candles. The method applies to any stellar group that contains a standard candle and whose diameter is small compared with its distance. Good standard candles are *consistent* (small luminosity range among individuals), *highly luminous* (for observation at great distance) and reliably *recognizable*. Variable stars naturally satisfy the recognition requirement, while some types satisfy the other requirements reasonably well and in some cases quite well. Cepheids, being giants and supergiants, can be observed in distant galaxies. Individual Cepheids differ in luminosity by factors up to about 100 and so might seem not to satisfy the consistency requirement, but H Leavitt discovered (*c*. 1905) that Cepheid luminosities are closely related to their directly measurable periods of brightness variation. After calibration, the *period–luminosity (P–L) relation* allows Cepheids to be used as standard candles, although essentially correct calibration required another half-century and is still being refined. However, even faulty early calibrations established distances to galaxies that were correct to an order of magnitude and thereby helped elucidate the nature of the 'spiral nebulae' and the scale of the Universe. Early applications were E Hertzsprung's distance estimate for the Small Magellanic Cloud and H Shapley's work on the extra-Galactic distance scale, following E Hubble's discovery of Cepheids in several galaxies. The distinction between Galactic and extra-Galactic objects was well established by 1925, five years after the famous Shapley–Curtis debates, when Hubble found 11 Cepheids in galaxy NGC 6822 and derived a distance. The Hubble Space Telescope and large ground-based telescopes have recently produced important increases in the Cepheid distance limit. Several kinds of exploding stars have very high luminosity, but Type Ia supernovae are best for recognition and consistency, in addition to being especially luminous. However, even the far less luminous and far less consistent ordinary novae have been used as distance indicators.

The large-amplitude pulsations of Cepheids and related variables can be exploited to measure radii rather directly. Luminosities then follow from combination of the radii with effective temperature ($T_{\rm eff}$) estimates, so as to provide standard candles. Observed variation is due to separable changes in surface brightness and size. 'Wesselink radii' (after A Wesselink) require three kinds of accurate observations—a radial velocity curve, a light curve and a color index curve (briefly 'color curve'—formed by differencing light curves (in magnitudes) for two effective wavelengths). Color curves of Cepheids resemble their light curves at first glance, having smaller amplitudes and subtly different shapes. Points with given color occur at paired times that are close together near a color maximum or minimum and well separated at intermediate color. The basic idea is that color index is a reliable indicator of $T_{\rm eff}$ and therefore of surface brightness. However, a quantitative relation to predict surface brightness from color is not needed—just the reasonable assumption that when the star returns to a given color it also returns to the 'original' surface brightness. Any change in observable light, l, between those two times is due to changed surface area, so the ratio of radii at the two times is

$$\frac{R_2}{R_1} = \sqrt{\frac{l_2}{l_1}}.$$

The difference of radii follows from an integration between the corresponding points on the radial velocity curve,

$$R_2 - R_1 = \int_{t_1}^{t_2} v_r \, \mathrm{d}t.$$

Now having both ratio and difference, we find the individual radii at the two times. Repeating the procedure for many time pairs, we find the run of R with time. The fundamental assumption that surface brightness is a unique function of color for a given star may not be quite true, but is close enough for useful applications. With actual radii in kilometers and with model stellar atmosphere predictions of surface emission per unit area, a final step computes luminosities. The method is one of the few direct means to find accurate radii of giant stars and to calibrate P–L relations.

Another natural situation to exploit is the existence of variable stars in binaries and multiple star systems, where the variable star and binary star characteristics should be compatible if our evolutionary understanding is correct. For example, the age of a variable should agree with that of its companion(s), and its absolute dimensions and distance may be derivable from observations of the binary or multiple system. Examples include δ Scuti (see below) type members in Y Camelopardalis and AB Cassiopeia, a Cepheid in V350 Sagittarii, and a β Cephei star in 16 Lacertae. It is important to discover more such systems.

There are few probes of stellar interiors so it is important to have a *wide variety* of checks on global and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

surface properties, and the measures provided by pulsation and other forms of variation are welcome. For example, Cepheid masses can be inferred from both evolutionary and pulsational models, so the comparison either provides a check or sounds an alarm. Light curve and velocity curve details also can be predicted from pulsation models. The help comes where needed, as pulsational instabilities occur primarily in highly evolved stars where structure is complicated and theories are uncertain. Not only stellar structure theories but also equations of state, opacities, and energy generation rates ('laboratory physics' in principle but not in practice) are thereby subject to scrutiny.

On a wide front, galactic evolution is largely understood in terms of population types. For example the *RR Lyrae* type pulsating stars found in globular clusters and spheroidally distributed Galactic populations are not just short-period Cepheids, but have quite different Galactic distribution, motions, history and structure. Insight into RR Lyraes and Cepheids can apply to their non-variable neighbors in the HR diagram. On the observational side of galaxies, certain variable stars can be recognized at large distances and thus serve as population indicators.

Pulsation studies have now broadened into the rapidly developing field of asteroseismology, in which new observing methods and theoretical work on stellar oscillations extract structural information from multiple-frequency small amplitude variations in light and velocity. Applications to the Sun (helioseismology) are particularly successful because the Sun is spatially resolved and because it has a rich spectrum of oscillations. The directly tapped energy of solar oscillations is that of convective motions in the outer envelope that generate sound waves. Certain other kinds of stars, such as Ap, δ *Scuti* (see below) and especially some white dwarf stars, show intricate oscillations that are clues to their evolutionary states. World-wide observing networks such as the Whole Earth Telescope (WET) allow the long, nearly continuous, coverage needed to separate closely spaced frequencies.

Explosion mechanisms

Nova theories of the early to mid 1900s involved rapid contraction to the white dwarf state or dynamical resonances combined with nuclear reactions. Those theories now seem inapplicable to known kinds of stellar explosions. A modern schematic model that covers many particular categories includes a donor star to provide a supply of gas and an accreting star on which eruptions occur. The flow may be a stream from overspilling a critical lobe or it may be a wind, and the eruptions may be powered by thermonuclear or gravitational energy. Ordinarily the donor is non-degenerate while the accretor is some kind of compact degenerate object, and most commonly a white dwarf star. Neutron star accretors are much rarer, and main sequence donors greatly outnumber evolved donors, but a wide variety of combinations occur. An especially abundant class is that of the *cataclysmic variables* (CVs) where the compact object is a white dwarf, although the white dwarf often cannot be detected directly. In most CVs the companion is a low-mass main sequence star, the orbit period is well under a day, and the entire binary is similar in size to the Sun, and smaller in many cases. CVs are typically old objects as shown by their distribution in the Galaxy, which is in a thicker disk than ordinary Population I stars.

A major clue into mechanisms and evolution was the inferred presence, via radial velocities and light curves, of white dwarfs in post novae, recurrent novae and nova-like variables in the 1950s to 1960s. Although thermonuclear models had been proposed earlier, ideas for explosions soon focused on accretion-driven surface hydrogen-burning runaways on white dwarfs. CV component masses are difficult to estimate, but the white dwarfs are typically much more massive than their non-degenerate companions. Relatively massive donors should be subject to unstable mass transfer that would radically change the configuration (flow from high to low mass or between comparable mass stars favors instability). Evolved donors (giants and subgiants) have the same problem—for structural reasons they tend to expand, overflow their lobes and transfer gas unstably on a large scale. So the lack of evolved and of relatively massive donor stars is commonly explained by their self-destructive tendencies. However, a CV needs some level of slow mass transfer to fuel its eruptions. A suitable level can be maintained by orbit shrinkage due to magnetic wind braking or gravitational radiation.

Classical novae brighten by 9^m or more over a few days and then decline irregularly over weeks for fast novae or months for slow novae. Symbiotic novae (see below) are much slower (duration ≈ decades) than the slowest classical nova and are much wider binaries than the ordinary CVs. The mass donor is typically a giant. For classical novae, more or less standard spectral developments occur, with emission and absorption line episodes and ejecta velocities of the order of 10^3 km s^{-1}, but with a large velocity range among examples. Interpretation is difficult, as the state of the gas is hard to compute and the dynamical situation is intricate, with multiple velocity systems seen at most times. Classical novae have had only one known outburst, but that circumstance is attributed to the short history of observational astronomy. Thus there are also *nova-like variables* that are probably the same as classical novae but have not exploded in (astronomical) historical times. Observations of post-novae find the underlying binaries not markedly changed by their outbursts.

Recurrent novae explode like classical novae but the eruptions are frequent, typically decades apart. Current thinking favors accretors that are close to the white dwarf

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

mass limit, such that only small accretion build-up is needed to start a thermonuclear runaway. Chemical abundances in the ejecta do suggest advanced nuclear processing, as expected for remnants of stellar cores near the edge of collapse.

A computation of simple energetics shows that gravitational energy alone can power a large outburst from a white dwarf for a plausible sudden accretion episode. Objects identified with that mechanism are the *dwarf novae*—CVs that have cyclic outbursts with high repetition rates. Brightenings of up to about five magnitudes are separated by typical intervals of weeks to months and complicated shorter-term behavior includes fast flickering. Ideas to account for the episodes include sudden release of matter from a circumstellar disk (disk instability) and variable supply from the donor star (source instability). Examples are SS Cygni, U Geminorum, Z Camelopardalis and SU Ursae Majoris, each a prototype in a finer classification. Thermonuclear energy may play some role in dwarf novae. Ideas about evolutionary relationships among the several kinds of CVs are under continual revision.

Impressive as they are, CV explosions are small firecrackers compared with supernovae. In contrast to CVs, where the basic configuration remains after an explosion, a supernova event involves the entire star. Supernova mechanisms include thermonuclear incineration of a massive white dwarf star (in a binary) and collapse of an old, dense stellar core to nuclear degeneracy. Other types that are probably much less common have been proposed. The energy release in burning a white dwarf star can be estimated by multiplying a white dwarf mass ($\approx 3 \times 10^{33}$ g) by c^2 and by an efficiency factor ($\approx 10^{-3}$ for helium burning and beyond) to arrive at an energy between 10^{51} and 10^{52} erg, about a million times that of a typical nova. The minimum energy of a core collapse supernova is even much larger and follows from energy conservation in formation of a neutron star, whose gravitational binding energy, E_g, is of order -10^{53} erg. Obviously a positive energy that at least matches E_g must appear in the radiative and material ejecta. The visible radiation amounts to only $\approx 10^{48}$ to 10^{49} erg, and thus is a tiny fraction of the energy budget. Theoretical simulations indicate that nearly all of the energy is carried by neutrinos. The detection of about two dozen neutrinos from supernova 1987a in the Large Magellanic Cloud was a great triumph of supernova theory, as the number observed was the number predicted for a core collapse, within the uncertainties.

Pulsation

Pulsating stars undergo true pulsations (oscillations in size or shape). The pulsations are of the envelope, as the core is static and not directly involved. One could imagine the high-temperature sensitivity of nuclear reactions leading to core pulsations (the ϵ mechanism), but theory predicts that core pulsations will be quickly damped, except for supermassive stars. There are no candidates for core pulsation among the recognized variable stars, except perhaps the bizarre object η Carinae (see below) which may be a supermassive pulsating star. Pulsations grow when there is net conversion of thermal to mechanical energy, so the star is a heat engine. There are zones that *drive* pulsation (net [thermal $\Rightarrow$ mechanical]) and others that *damp* pulsation (net [mechanical $\Rightarrow$ thermal]). Elementary thermodynamics requires net heat injection in the compression stage if heat is to be converted to work so that the engine runs. However, direct injection of heat around maximum compression (as in an ordinary engine) is not the only way. Eddington realized in the 1920s that favorable circumstances could occur if a star is relatively heat-tight when compressed. Rather than thermal energy being inserted, its escape is prevented. This is the famous Eddington *valve* mechanism—a generic means to achieve net driving that covers all specific ways to implement heat-tightness upon compression. Two specific ways are the κ and γ mechanisms. The κ mechanism traps thermal energy by making material in driving zones more opaque upon compression (κ is the usual symbol for opacity). Kramers' Law, $\kappa = \kappa_0 \rho / T^{3.5}$, provides roughly correct opacity for given density (ρ) and temperature (T). With ρ and T entered from accurate stellar structure models, it predicts that stellar material becomes *less* opaque upon compression, and indeed most stars do not pulsate. However, actual opacity depends in a complicated way on thermodynamic variables and, although it decreases with compression in most parts of most stars, it *increases* with compression in the driving zones of Cepheids and some other pulsating stars. The γ mechanism operates by increasing the surface density of absorbing particles upon compression. Imagine a thin pulsating shell of gas at its maximum and minimum radii. The number of contained particles is the same whether the shell is large or small but the number per unit area is greatest when the shell is small. Therefore the shell more effectively blocks radiation, or is more heat-tight, when compressed. The γ mechanism always contributes to driving for obvious geometrical reasons, while the κ mechanism can produce either driving or damping.

Driving will be weak if the main driving zones are too close to the surface (high) or too deep inside the star (low), and net damping will arrest pulsation. In the high case, the problem is insufficient density in the driving zones so that too little matter contributes to driving. In the low case, the driving zones lie in the nearly adiabatic interior where, although there is plenty of mass, each gram makes only a small contribution. Because driving is associated with ionization zones of the abundant elements H and He, net driving zones will be high in relatively hot (blue) stars and low in relatively cool (red)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

stars. Thus we have a roughly vertical *instability strip* in the HR diagram where Cepheids are found. However, its red and blue borders are not well defined and many non-pulsating stars lie within the strip. The instability strip extends down to and below the main sequence and includes the little-evolved or mildly evolved δ Scuti and related variables with their typical periods of hours. It then continues to the region of the pulsating white dwarfs (ZZ Ceti stars).

Several major distinctions characterize stellar pulsations. To begin, there are radial pulsations that preserve a star's figure (shape) but change its volume, and non-radial pulsations that preserve volume but vary the figure. An example of radial pulsation could be the expansion and contraction of a balloon under cyclically changing external pressure. A small-amplitude example of the non-radial case is provided by the tidal distortions of the Earth and its oceans. A large-amplitude example would be the oscillations of a disturbed water globule. A more formal description of the contrasting behavior is in terms of a vector field. In radial pulsation, the displacement vectors of local matter elements pass back and forth through zero length but maintain fixed alignments along local radii. Non-radial pulsation is more complicated, with the vectors cyclically changing both length and direction. Of course, real pulsations need not be purely radial or non-radial but can involve both volume and figure changes. However, many realistic situations approximate these idealized types.

Another distinction involves fundamental and overtone pulsation. The situation is conceptually the same as for a flexible string with one fixed and one free end, where the free end is the analog of a pulsating star's surface. The string can have a fundamental oscillation, with a node only at the fixed end, and also overtone oscillations with 1, 2, 3,…, *n* additional nodes. At a given instant, adjacent inter-node regions have opposite motions (up versus down), and of course there is no motion at the nodes. Similarly a pulsating star can have fundamental and overtone pulsations, with extra nodes for the overtones. For a uniform string the nodes will be equally spaced, but a star has inwardly increasing density, with the consequence that node spacings are not even approximately equal and must be computed. Fundamental and overtone pulsations can coexist and, accordingly, some unusual stars show beat phenomena.

Still another distinction is among pulsation *modes*, i.e. possible ways to pulsate. The large amplitude radial pulsations of Cepheid and RR Lyr type variables are in *pressure* modes (p-modes), so called because they involve large local pressure variations. Both radial and non-radial p-mode pulsations can occur. *Gravity* modes (g-modes) involve global oscillations about a hydrostatic level surface and are necessarily non-radial. They have much smaller pressure variations than do p-modes. Of course, energy can feed from one mode to another and between fundamental and overtones.

No one kind of pulsation analysis serves all purposes. As in other areas, success in computing detailed behavior does not guarantee understanding. The general phenomenon is highly nonlinear and one might expect maximum insight to follow from all-encompassing computations, but much insight has been gained from linear analyses in which only infinitesimal pulsations are studied. N Baker realized that outcome by restricting attention to a single mass shell—the One Zone model. However, pulsating stars are complicated structures with many (thermally and mechanically) coupled zones, so quantitative predictions of periods need complete models. To view the problem the opposite way, a full nonlinear computation should settle into a correct period, but will not identify all possible periods found from linear theory. Although linear analysis leads to insight on several fronts, it deals only with infinitesimal pulsations and therefore cannot produce complete models for comparison with observations. Possible pulsation modes and overtones are identified through linear analysis, with detailed behavior at finite amplitude examined via nonlinear analysis, including possible interactions of fundamental and overtones. Similarly, growth and decay rates of small-amplitude pulsations can be investigated via linear analysis, but rates for large pulsations and final saturation amplitudes are matters for nonlinear analysis. Of course, specific light and velocity curve features can be modeled only with full nonlinear computations.

A wealth of observed phenomena in pulsating stars remain only partly understood or have only recently been understood, including light versus velocity phase lags that are characteristic of a given class, unusual surface abundances, changing velocity amplitudes and multiple periodicity. Magnetic fields, fast rotation and binary companions are clearly present in some cases, but their roles in pulsation are not usually obvious. Specific driving mechanisms may not be clear. An active testing ground for pulsation theory is provided by the *β Cephei stars*, a class sometimes called the β Canis Majoris stars, although current usage seems to have settled on the name β Cephei. These stars have remarkably short pulsation periods of order $0.^d2$, with about half being doubly periodic and thereby showing beat behavior in light and velocity. Many RR Lyr and δ Sct stars also are doubly periodic, but the two periods of a β Cep star can differ by as little as 1%. The prototype β Cephei (not itself doubly periodic) is strongly magnetic and at least a triple system, with weak B-emission (Be) characteristics. A phenomenon that must occur but is difficult to model is that of dynamical tides in eccentric and/or asynchronous binaries, which can be regarded as forced non-radial pulsations. Like ocean waves, such tides are largely stochastic so that prediction is mainly limited to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

statistical behavior. Several supergiants with neutron star companions show the expected fluctuations in light and velocity, including GP Velorum = Vela X1 and V884 Scorpii, but quantitative matches with tidal theory remain unrealized.

Other variation mechanisms

While eruptive and pulsating stars vary due to a variety of dynamical, thermal and cyclic instabilities that are more or less understood, or at least largely identified, variations in many categories are caused by instabilities that are not so well understood, by random events and by random drifts in conditions. Some of these stars live on the border between intrinsic and extrinsic variables. For example, RS CVn, BY Dra and W UMa binaries can be heavily covered by magnetic spots, similar to sunspots. The agent understood to produce such spots is dynamo action in a fast-rotating convective envelope, so cool stars that have convective envelopes are likely candidates—if they spin fast. The RS CVns, BY Dras, and W UMas naturally spin fast because they are tidally locked close binaries, but FK Com and a small number of similar stars are fast spinning *single* giants! Their fast rotation currently is attributed to their being recently merged ex-binary systems, with the binary orbital angular momentum now existing as spin angular momentum. Spot-modulated stars are extrinsic variables in that they vary on the orbital time scale because of changing aspect of their spots. However, the spots grow, decline and move, so they are intrinsic variables on longer time-scales.

Accretion tends to be highly irregular and causes not only modest brightness fluctuations in active binaries, but variations of up to several magnitudes in newly formed stars that accrete from their surroundings (T Tauri stars). Much of the luminosity of T Tauris and related stars is from the accretion process. Young objects of the FU Orionis type have brightening episodes of up to five magnitudes on a time-scale of decades to centuries. Outflows are associated with accretion but are not well understood at present. Some classical T Tau stars are surrounded by circumstellar disks that have been imaged in the infrared, with spectroscopic and imaging evidence of bipolar outflows. Accretion onto white dwarfs not only leads to circumstances that eventually produce nova explosions but also directly converts gravitational energy to more tangible forms, as seen above for dwarf novae. Even an occasional non-degenerate binary such as V361 Lyrae has a hot spot where high-speed gas impacts one of the stars, having flowed in a well defined stream from the other. Variable flows produce variable spot luminosities and thus another kind of intrinsic–extrinsic variable. In post novae and dwarf novae, the hot spot is on a disk surrounding the stream's target star. Most exploding stars show variability between outbursts in their 'quiescent' light curves. A common seat of variation is the hot spot as it flickers by $0.^m1$ to $0.^m2$ on a time scale of minutes, due to irregular inflow. Usually the hot spot itself, rather than either star or the overall disk, is the brightest light source in the binary. X-ray binaries whose accreting objects are neutron stars have accretion power at least 30 times their thermonuclear power (for H burning; more than 100 times for He burning), so the promptly released energy is about the same whether the material burns or not. However, large thermonuclear bursts can occur on neutron stars if substantial amounts of fuel accumulate, providing the sources known as x-ray bursters, where helium is the fuel. As a group, x-ray binaries are variable in all spectral regions, from radio to gamma, although most individuals are detected only in restricted ranges. There are two essentially disjoint classes on greatly different scales, low-mass x-ray binaries (LMXBs) and high-mass x-ray binaries (HMXBs). At first inspection, LMXBs and HMXBs have little in common except that they contain neutron stars, and even a sketch of ideas about their origins would exceed available space. Configurations are such that HMXBs accrete mainly via winds from their blue supergiant companions and many pulse in x-rays, while LMXBs accrete mainly via lobe overflow and very few pulse. LMXBs bear a remarkable similarity to CVs, including absolute dimensions, with the notable difference being that a neutron star replaces the CV's white dwarf. It has been noted that some LMXBs may form from CVs in rare cases of accretion beyond the white dwarf mass limit.

Still another kind of intrinsic–extrinsic hybrid is the *symbiotic star*, where ultraviolet radiation from the environs of a small, hot accretor interacts with the wind, extended atmosphere or chromosphere of a red giant companion. Fluctuations in UV radiation and in wind flow lead to intrinsically variable fluorescence modulated by orbital aspect effects, sometimes including eclipses. Orbit periods are long compared with most variable binaries, being typically hundreds of days or more. Because of observational difficulties, the defining characteristics of the class are necessarily superficial—the accreting objects are not usually directly observed and may be as diverse as neutron stars, white dwarfs and main sequence stars. Extreme examples of symbiotics include *symbiotic novae* that have outbursts by brightness factors of order 100 and are believed to contain white dwarf accretors. Unlike CVs, a symbiotic nova can be in outburst for decades. The mechanism usually assumed is a thermonuclear surface flash on the accretor followed by slow cooling. These objects are often called slow novae but differ radically from the CVs with the same name, so 'symbiotic nova' should consistently be applied. Examples of symbiotic novae include PU Vulpeculae, RR Telescopii and V1329 Cygni.

Variable polarization can result from scattering by circumstellar gas, but really spectacular variable polarization is seen in CVs that contain accreting white

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

dwarfs with extremely strong magnetic fields. These are the *polars*, or AM Her binaries, where cyclotron radiation in an oblique rotator produces strong circular polarization over a wide range of wavelength, including optical light. Polars also are moderately strong x-ray sources and natural laboratories for studying hydrodynamic flows under combined gravitational and magnetic fields, with variable light and polarization as diagnostics. *Intermediate polars* have magnetic fields that are intense but somewhat below those of polars. Examples of intermediate polars are EX Hydrae and GK Persei.

Fluctuations in stellar winds can generate small-scale emission line and continuum variation, particularly where winds are strong. Most *Wolf–Rayet* (W–R) stars accordingly vary by a few hundredths of a magnitude. A W–R star is the highly evolved residue of a very massive star that has lost much of its envelope so as to expose hot inner regions. Often there is an evolved close companion with its own strong wind, so that wind–wind interactions cause further variation.

Intermittent veiling is a very unusual variation mechanism and a major diagnostic of circumstellar conditions and interior structure in certain (usually giant) stars. In veiled stars we find intrinsic and extrinsic variation combined, as clouds come and go while changing aspect controls their influence on light curves. Hydrodynamic flow in a mass transferring binary can lead to concentrations of partially opaque material being projected onto the face of one or both stars, so that light curves can help map circumstellar gas. An example is AX Monocerotis. Particularly spectacular examples of veiling are the *R Coronae Borealis stars*—extremely carbon rich and hydrogen deficient (apparently single) supergiants that are veiled by carbon particles at irregular times, with brightness drops by factors of up to about 1000. These pulsating stars eject clouds of gas whose carbon condenses to soot when sufficiently far from the hot photosphere, with clouds that happen to lie on the line of sight producing the veiling. Pulsational characteristics, together with the strange abundances, provide strong constraints on possible evolutionary histories of R CrB stars. Given that only a few dozen are known, R CrBs must either represent a very brief stage of normal evolution or products of an unusual formation process such as a merger.

Large evolutionary changes over the brief history of observational astrophysics are extremely unusual. Perhaps the most spectacular example is FG Sagittae, which has evolved from the small, hot exciting star of a planetary nebula into a pulsating R CrB type red giant in only a century, with decade to decade developments. Whether FG Sge is typical as a progenitor of R CrBs is an open issue at present.

Flares are brief local eruptions, prominent in both continuum light and lines, from the chromospheres of magnetically active stars. The *UV Ceti* or *flare stars* are low-mass main sequence stars with unusually high levels of chromospheric activity. Flares are much hotter than red dwarf photospheres, so the brightening in magnitudes is highly wavelength dependent, increasing strongly into the ultraviolet. As with spot-modulated stars, rotation in a deep convective envelope generates strong dynamo action, with UV Cet stars being especially fast rotators. Flare activity is at least statistically a sign of youth, since red dwarf rotation decreases with age.

Stars described as variable are traditionally those that vary in brightness, especially over broad spectral regions. However, several kinds of *spectrum variables* vary most notably in spectral details, with behavior that can be as interesting as in the more obvious variables. For example, strong variable magnetic fields are involved in spectral variations of Ap and Am stars (Ap = peculiar stars of spectral type A; Am = metallic line stars of spectral type A). Stars that lose matter via powerful and unsteady winds, such as those of spectral type Of, show variable emission lines with great Doppler broadening. Stars of type Be (*B emission stars*) were among the first to draw special attention as spectroscopically interesting in the 1800s, and continue to stimulate hypotheses, observations, and controversy. The emission lines of Be stars are usually ascribed to circumstellar equatorial rings associated with centrifugal ejection of matter, but circumstances leading to that situation are not widely agreed upon. Even the basic point of whether binary nature is essential or irrelevant to Be behavior remains at issue.

Unique objects

Much of the fascination of variable star astrophysics is provided by unique objects that may result from anomalous formation, short effective lifetime or selection effects. A few examples may give some of the flavor.

Most recognized causes of variation, including pulsation, gas dynamic interactions and rotation phenomena, have been proposed for the spectacular and mysterious η Carinae. Especially notable are its great luminosity and enormous swings in apparent magnitude over centuries, including an interval in the 1840s when it approached Sirius in brightness. It is usually regarded as a supermassive star with mass above (and perhaps far above) 60 solar masses, and therefore a candidate for core pulsation via unstable thermonuclear burning. Speculation that it is a binary is supported by recent high resolution imaging of a bipolar nebula that must have been ejected in η Car's great outburst of *c*. 1840 and by discovery of 5.5 year spectroscopic and photometric periodicities. Understanding of η Car is rapidly developing but still rudimentary.

The recurrent nova T Corona Borealis has the basic red star–blue star binary morphology of novae, but with a red giant in place of the usual main sequence red star and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

an orbit period (228^d) typical of symbiotic novae. However, it can scarcely be called a symbiotic nova, as the outbursts last not for decades but only for weeks. T CrB has no known siblings, but does have a rich and extensive literature.

And then there are the pulsating WHITE DWARF stars, the DUSTY CIRCUMSTELLAR DISKS, the QUASIPERIODIC OSCILLATIONS IN X-RAY BINARY STARS, the incredible SS433 and HERCULES X-1, …, but wait—those are other articles. No brief synopsis could do them justice. So turn to those articles—your instrumentation is a comfortable chair and a cup of hot chocolate. Enjoy.

Bibliography

Clark D H and Stephenson F R 1977 *The Historical Supernovae* (Oxford: Pergamon)

Gallagher J S and Starrfield S 1978 *Annu. Rev. Astron. Astrophys.* **16** 171

Hoskin M 1979 *J. Hist. Astron.* **10** 23

Lewin W H G, van Paradijs J and van den Heuvel E P J (ed) 1995 *X-ray Binaries* (Cambridge: Cambridge University Press)

Kholopov P N 1982 *Catalog of New Suspected Variable Stars* (Moscow: Nauka)

Kholopov P N 1984 *General Catalog of Variable Stars* (Moscow: Nauka)

Kippenhahn R and Weigert A 1990 *Stellar Structure and Evolution* (Berlin: Springer)

Payne-Gaposchkin C 1957 *The Galactic Novae* (Amsterdam: North Holland)

Warner B 1995 *Cataclysmic Variable Stars* (Cambridge: Cambridge University Press)

R E Wilson

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vatican Observatory

The Vatican Observatory is one of the oldest astronomical institutes in the world. It began with the reformation of the calendar in 1582. At the Roman College, Father Angelo Secchi first classified stars according to their spectra. With these rich traditions Leo XIII, in 1891, formally founded the Vatican Observatory on a hillside behind the dome of St Peter's Basilica. In 1935 Pius XI provided a new location for the Observatory at the Papal Summer Residence at Castel Gandolfo. In 1981 the Observatory founded a second research center in Tucson, Arizona. In 1993 the Observatory, in collaboration with STEWARD OBSERVATORY, completed the construction of the Vatican Advanced Technology Telescope (VATT) which has pioneered the new technology of creating large, lightweight, stable mirrors in a rotating furnace. Research includes cosmological models, spectral classification of peculiar stars, photometric studies of metallicity, mass-exchanging binary stars, material in star-forming dark clouds, dust envelopes about young stars, planetary dynamics. The Observatory is supported with an annual budget from the Holy See and by contributions to the Vatican Observatory Foundation, a tax-exempt corporation in the State of Arizona.

For further information see
http://clavius.as.arizona.edu/vo/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vega

The star α Lyrae, the only bright star in the constellation. Its name, formerly Wega, derives from the Arabic *Al Waki*, 'the Swooping (or Falling) Eagle', hence the form Alvaka used on some seventeenth century celestial globes. In ancient Greek and Latin writings it sometimes shared the constellation name Lyra, and consequently appears as 'the Harp Star' in some later texts. It is the dominant component of the Summer Triangle asterism, with Altair and Deneb. It was the Pole Star about 12 000 years ago, and will be again 14 000 years hence.

Vega is the fifth brightest star in the heavens, with an apparent magnitude of 0.03. It is the brightest star with sufficiently high northern declination to appear overhead in north temperate latitudes, and dominates their skies on clear summer nights. It is a striking object when viewed with binoculars or a low-power telescope, its blue-white color (sometimes described as 'sapphire') belying its relatively cool spectral type (A0Vvar). Fairly close at 25.3 light-years distance (parallax 0.129″), Vega is in fact the third closest of the bright stars. Its absolute magnitude is 0.6. In 1983 it was discovered by the IRAS survey to be surrounded by a disk of gas and dust, which may eventually form a planetary system.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

VEGA Space Mission

VEGA (mission) is a combined spacecraft mission to VENUS and COMET HALLEY. It was launched in the USSR at the end of 1984. The mission consisted of two identical spacecraft VEGA 1 and VEGA 2. VEGA is an acronym built from the words 'Venus' and 'Halley' ('Galley' in Russian spelling). The basic design of the spacecraft was the same as has been used many times to deliver Soviet landers and orbiters to Venus. In June 1985, these spacecraft successfully delivered the first balloons into the Venus atmosphere and also landers. After this the VEGA 1 and VEGA 2 cometary probes were directed to the comet Halley. They encountered it on 6 and 9 March 1986. Such a trajectory was possible only because of the favorable mutual positions of the comet and Venus at this time. The gravitational field of Venus turned the spacecraft to the comet and only a short switch-on of the spacecraft engine was necessary.

The most important scientific achievements of the VEGA mission were as follows: (a) the first direct tracing of Venusian atmospheric properties (winds, turbulence, illuminance, cloud density) on a long horizontal path; (b) the first look at a cometary nucleus, not as a star-like object, but as a spatially resolved body; (c) the first identification of a set of components of the cometary atmosphere, including organic matter; (d) the first study of the cometary environment by direct (contact) methods. The VEGA mission was an important step in the development of international cooperation in space studies. Scientists and engineers of many countries were involved together with Russians in the creation of sophisticated scientific instruments and supporting systems. Such international cooperation in a space project was new not only for Russia. Professor R Sagdeev (leader then of the Institute of Space Research in Moscow) was the head of the whole project, and these innovations were in many points results of his enlightened activity.

Venus

Balloons

The launch of balloons with meteorological instruments is traditional in studies of the terrestrial atmosphere. It is natural to apply this for atmospheres of other planets. The concept of a Venusian balloon was first proposed by J Blamont in 1967, but only after 18 yr of technical studies (in the USSR and France) was it realized in the VEGA mission.

The Venus module of the VEGA spacecraft entered the atmosphere of the planet and separated into two parts at a height of 64 km. One of them was a descent probe (lander) and the other was a canister containing balloon, ballast, parachute and inflating systems. Beginning at 64 km the container went down with the parachute to about 50 km (where the atmospheric pressure is near 1 bar); there the balloon was inflated, and parachute and ballast separated. After this balloon went up to about 53.6 km corresponding to a pressure of 535 mbar and a temperature of 305 K. Both balloons were inserted on the nightside of Venus, one (VEGA 1) $\sim 7^\circ$ north of the equator and the other $\sim 7^\circ$ south of the equator. They drifted westward with the predominant zonal wind closely parallel to the lines of latitude. Each balloon was tracked over a distance of more than 11 000 km, encountering dawn and then penetrating far into the daylit atmosphere (figure 1(D)). Transmission was stopped after 46 h of flight owing to limited battery resources.

There were two types of measurements: *in situ* by means of on-board instruments transmitted by telemetry and ground-based determination of balloon motion by very-long-baseline interferometry (VLBI). The balloon's transmitter sent directly to the Earth signals at a wavelength of 18 cm used ordinarily by VLBI astronomers for observations of distant galactic OH sources. Because of Earth's rotation, continuous reception of telemetry data required the use of many antennas widely distributed in longitude.

It was known earlier that so-called superrotation is the dominating dynamical property of Venus' atmosphere: its rotation period (about 4 days near cloud top) is much shorter than for the solid surface. This means that a fast westward wind is a permanent feature on Venus (see VENUS: ATMOSPHERE). VEGA balloons provided its mean velocity (67–69 m s^{-1}) and its fine fluctuation pattern also.

Strong turbulence was found with amplitude to 2 m s^{-1} with a time scale of 30–100 s (figure 1(C)). The influence of surface topography on atmospheric motions was found; it was not expected for such high altitudes. The measurements covered the twilight zone which was found to be 7° wide in the clouds of Venus. The horizontal profile of the clouds' optical properties (backscattering coefficient) was measured for the first time. It was found that at the height of flight it is smooth in general; even the strongest variations were no more than a factor of 3 from the average level (figure 1(E)). Such an experiment has not yet been repeated; however, balloons are present in many proposals for future studies of the atmospheres of Venus and Mars.

Descent probes

Studies of Venus by means of descent probes were started by the VENERA 4 mission in 1967. Another 10 Soviet VENERA MISSIONS followed during 1969–1983; eight of them included descent probes with instrumental packages that provided measurements of the chemical composition of the atmosphere and soil, images of the landscape around landers, vertical profiles of temperature and pressure, the cloud particles' properties and solar illuminance within the deep atmosphere of Venus. VEGA 1 and 2's descending probes landed on the nightside of the planet, unlike most of the previous descenders, but the general scheme was traditional: entry to the atmosphere, fast aerodynamic deceleration of capsule with descenders between 100 and 64 km, opening of capsule, about 10 min of descent to 46 km, separation of parachute, landing after about 51 min and 20 min of working data transfer from the surface.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

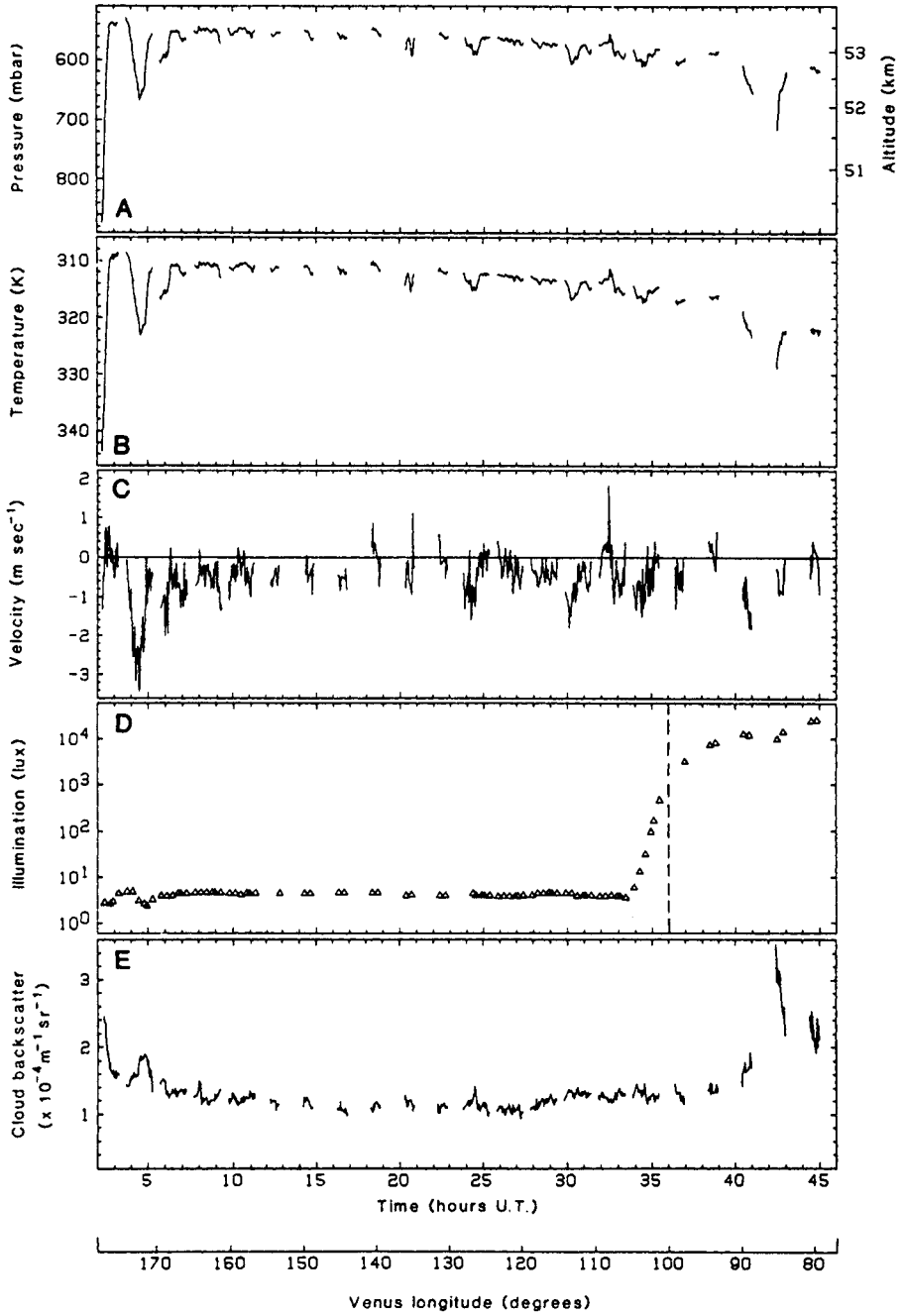


Figure 1. Mission history for VEGA 1 balloon showing (A) pressure, (B) temperature, (C) atmospheric vertical velocity, (D) ambient light level and (E) cloud backscatter coefficient as functions of time after insertion on 11 June 1985. Venus longitudes are also shown on the abscissa. The approximate position of the terminator is shown as a broken line in the plot (D). Gaps in the data sets indicate times when no *in situ* measurements were transmitted by the balloon (Sagdeev *et al* 1986b).

Improved measurements of atmospheric temperature were made in the full range between ∼63 km and the surface. The temperature rises monotonically from up to down but an inversion was found near 61 km. The lapse rate is not constant; its variations confirmed the existence of two convective zones, one below 15 km, another between 45 and 58 km (unlike our planet where only one convective zone exists, between ∼10 km and the surface).

Cloud particles were studied in different ways with a set of sensors that measured size distribution, number density and chemical composition. Sulfuric acid (H_2SO_4)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

as one of the constituents of the Venusian cloud particles was detected by *in situ* measurements. This compound dominates the upper and probably the middle clouds (down to 54 km) but below this level cloud particles mostly consist of a phosphorus compound (probably H_3PO_4).

A UV lamp and spectrometer were used to study the strong UV spectral band created by small traces of SO_2 which was found by earlier probes by another method (gas chromatography). Analysis of the new measurements confirmed the presence of SO_2 but gave different quantities and covered larger height ranges.

Measurements of soil composition were made after landing by two methods. One of them (x-ray fluorescent spectrometry on VEGA 2) showed that elemental composition is like that of very ancient lunar rocks (see also VENUS: SURFACE). Another (based on K, U and Th abundances) in both places showed rocks like toleit basalts.

The VEGA descent probes were the last in a long set of Soviet VENERA missions but most of the studies resulted in new questions that require new experiments. However, it is impossible to say at the moment where and when new steps will be made for prolongation of the study of Venus.

Comet Halley

COMETARY NUCLEI are probably samples of pristine solar system material and their study may be important for an understanding of the origin of the solar system. However, they are practically not available for observations from the Earth; only products of their evaporation (coma and tail) are observable. Spacecraft give a unique possibility of approach to the cometary nucleus and investigation of it as a spatially resolved object. Comet Halley was selected for the first try. At every periapsis passage (once per 76 yr) it is available for astronomical observations but the relative position of the comet and Earth is not same in different apparitions. For example, the passage of 1986 was not favorable, because the comet periapsis point and Earth were nearly in opposite directions from the Sun. So it was reasonable to use spacecraft during this apparition. Comet Halley was especially attractive for such a mission because it is a periodic comet with a well-known orbit, it is the brightest among the periodical comets and it is the only periodic comet having physical properties like those of a young comet.

In reality not one but five spacecraft were sent to the comet Halley: Soviet VEGA 1, VEGA 2, European GIOTTO and Japan SUISEI and SAKIGAKE. The VEGA cometary probes were the largest and had more scientific instruments than the European and Japanese craft put together. The mass of every VEGA spacecraft (figure 2) was about 4500 kg, including about 2500 kg for the cometary module. There were two sorts of experiments on VEGA cometary modules: (a) camera and other optical instruments for remote sensing studies of the nucleus and inner coma and (b) instruments for contact measurements in cometary gas, dust and plasma. Imaging of the comet from the spacecraft required a steerable platform which could be

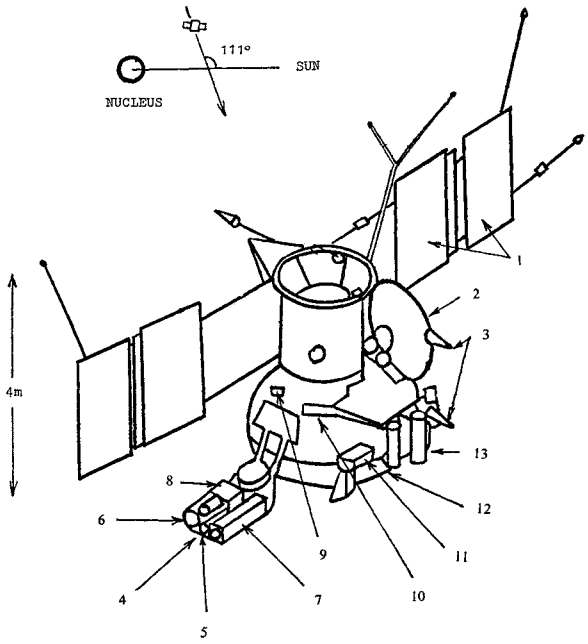


Figure 2. The Halley probe of the VEGA spacecraft , showing the locations of some subsystems and scientific instruments: 1, solar panels; 2, pencil-beam antenna; 3, omnidirectional antenna; 4, steerable platform; 5, the pointing sensor; 6, camera system TVS; 7, infrared spectrometer IKS; 8, ultraviolet–visible–near-infrared spectrometer a; 9, dust particle sensor SP-1; 10, particle composition analyser PUMA; 11, plasma composition analyser PLASMAG; 12, dust protection shield; 13, astro-orientation unit. The Venus probe is not shown. Its location before separation is on the top of the Halley probe.

automatically pointed with high accuracy to the cometary nucleus. A TV camera, IR spectrometer and UV–Visible–near-IR spectrometer were placed on this platform. The most difficult task was to locate the nucleus, with its low albedo, against the background of the bright coma with gas and dust jets, and this problem was solved successfully. The fly-by speed was extremely high ($\sim$78 km s^{-1}), because the direction of orbital motion of the comet is opposite to that of the Earth, and thus of the spacecraft. This high speed resulted in a strong dust hazard and required protection of parts of the spacecraft.

VEGA 1 passed through the comet Halley coma on 6 March 1986 at a nearest distance to the nucleus of 8890 km, VEGA 2 did it 3 days later at a distance of 8030 km. Most of the data were sent to Earth in real time , because loss of the spacecraft due to the dust hazard was possible. However, this did not happen; both of the spacecraft were alive after the fly-by, although the solar-battery current dropped by 50%. Figure 3 shows how comet Halley looked from the spacecraft VEGA 2 on 6 March 1986. Results of the analysis of images, spectra, and *in situ* data obtained are briefly outlined below.

It was concluded that the nucleus is a single solid body. Its shape is irregular (like a potato); the size is approximately 8 km $\times$ 8 km $\times$ 16 km. It has a very low

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

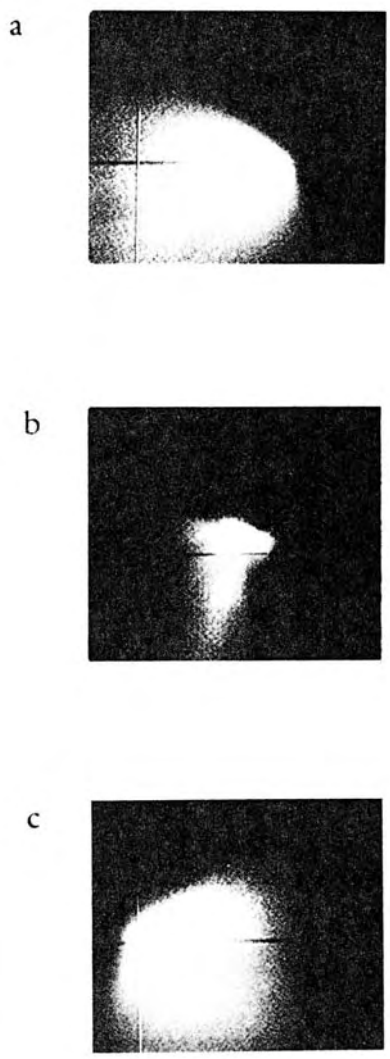


Figure 3. How comet Halley looked from spacecraft VEGA 2 flying through its coma. (*a*) The distance to the nucleus $d = 29\,000$ km, the Sun is on the left, the nucleus is near the right edge of the bright dust fan directed towards the Sun; the frame size is 228 km × 308 km. (*b*) $d = 8000$ km, the Sun is above the plane of the image (phase angle ∼30°), the frame size is 62 km × 84 km; the upper part of the visible spot is the nucleus, the lower is a dust jet. (*c*) The Sun is to the right, the nucleus is to the left; the distance and size are the same as in (*a*). (Moroz 1989).

reflectivity; the geometrical albedo is ∼0.04. So it is one of the darkest bodies of the solar system. Dust is emitted by its surface from the sunlit side. The dust flow consists mainly of narrow jets. The thermal IR radiation of the nucleus was measured by an infrared device. Its value corresponds to a temperature of about 380 K near the subsolar point. This was unexpected because a lot of water ice with a temperature about 200 K should exist in the upper layers of the nucleus to provide the H_2O flux observed by the same instrument and also in many other experiments. This discrepancy may be explained by a patchy structure of the surface (icy and rocky patches) or by a porous mantle which may be hot on the upper boundary and cold on the lower. The mass of nucleus was estimated as 3×10^{14} t and density about 0.6 t m^{-3}.

Indirect estimates of the chemical composition of the nucleus are possible on the basis of measurements of the chemical composition of gas and dust going out from the nucleus surface to the coma. H_2O was identified as the most abundant constituent with a production rate of about 10^{30} molecules s^{-1}. Organic substance outflow (never observed earlier) was identified using IR spectra obtained by VEGA 1. Its production rate may be comparable with that of H_2O. CO_2, CO and HCHO were observed with same instrument as parent species with mixing ratios of a few per cent. It is important that mainly secondary components arising from the decay of such parent molecules were observed earlier from the Earth. Detailed observations of secondary species with the VEGA spacecraft revealed their production rates, geometry of ejection, lifetimes of species and their parent and species velocities.

Neutral atoms and molecules in the coma are converted to ions by photoionization. The SOLAR WIND when it encounters them creates a shock wave that was met by the spacecraft at a distance of $\sim 10^6$ km from the nucleus. Nearer, at a distance $\sim 10^5$ km, another characteristic boundary (not predicted by theory) was met; later it was called the 'cometopause'. Cometary ions dominate the plasma inside this. The plasma composition, its energy distribution, plasma waves and magnetic fields were measured.

Grains of different masses (from $\sim 10^{-19}$ to $\sim 10^{-11}$ kg) were counted by a few VEGA instruments. The dust mass production rate was estimated. One of the instruments provided measurements of the chemical composition of individual particles and showed its large diversity. Some of them contain simultaneously C, H, O, N as was predicted for interstellar particles; others are icy, silicate or metallic bearing. Very small particles ($<10^{-14}$ g) are dense (1 and 2.5 g cm^{-3} for organic and silicate particles, respectively), while optically active particles ($>10^{-10}$ g) are of low density (0.3 g cm^{-3}) and of a complicated structure.

VEGA's flight to comet Halley was an important part of coordinated international efforts giving a very strong impulse to studies of comets and other small bodies of the solar system.

Bibliography

Moroz V I 1989 Halley's comet (part II): space studies *Highlights Astron.* **8** 17–31

Sagdeev R Z, Blamont J, Galeev A A, Moroz V I, Shapiro V D, Schevchenko V I and Szego K 1986a Vega spacecraft encounter with comet Halley *Nature* **321** 259–62

Sagdeev R Z *et al* 1986b Overview of VEGA balloon *in situ* meteorological measurements *Science* **321** 1411–4

V Moroz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vela

(the Sails; abbrev. Vel, gen. Velorum; area 500 sq. deg.) a southern constellation which lies between Antlia–Pyxis and Carina, and culminates at midnight in mid February. It was introduced by the French astronomer Nicolas L de Lacaille (1713–62), who charted the southern sky in 1751–2, from stars that formed part of the ancient constellation of Argo Navis (the Ship), which had been included by Ptolemy (c. AD 100–175) in the *Almagest*.

The brightest stars in Vela are γ Velorum (Regor), a multiple system consisting of the brightest Wolf–Rayet star in the sky (bluish-white, WC8 + O9, range 1.81–1.87) and a bluish-white (B2) component, magnitude 4.3, separation 41″, the former of which has an unseen companion that revolves around it in 78.5 days, and a fourth, bluish-white (B6) component, magnitude 7.7, separation 63″, δ Velorum, a very close binary with a white (A1) primary, magnitude 2.0, and a fainter secondary, magnitude 5.6, separation 0.74″, λ Velorum (Suhail), magnitude 2.2, κ Velorum (Markeb), magnitude 2.5, and μ Velorum, another binary, with yellow (G5 and G2) components, magnitudes 2.9 and 5.9, separation 2.0″. There are nine other stars brighter than magnitude 4.0. δ and κ Velorum together with ι and ε Carinae make up as asterism called the False Cross, as it is sometimes confused with the constellation Crux (the Southern Cross).

The Milky Way passes through Vela and the constellation contains more than 40 open clusters and many planetary nebulae, including IC 2391, an open cluster of more than 20 stars scattered across a 1° field, the brightest of which is *o* Velorum, magnitude 3.6, and NGC 3132 (the Eight-Burst Nebula), a ninth-magnitude planetary nebula with a tenth-magnitude star at its center. Other interesting objects include NGC 3201, a seventh-magnitude globular cluster, and the vast Gum Nebula, an expanding emission nebula some 36° across that extends into the neighboring constellations of Antlia, Pyxis, Puppis and Carina, and is thought to be the result of a supernova explosion that occurred more than 1 million years ago, inside which is the more recent Vela supernova remnant and at its center the Vela pulsar.

See also: False Cross, Gum Nebula, Vela pulsar, Vela supernova remnant.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vela Pulsar

The pulsar PSR 0833–45, discovered in 1968. With a period of only 9.3 ms, it has one of the fastest pulse-rates known, implying that it is one of the youngest pulsars. This is borne out by an observed deceleration in its pulse-rate of 10.7 ns per day, which sets an upper limit to the time elapsed since it was formed in a supernova explosion of about 11 000 years. In 1977 its pulsations were also recorded in visible light, making it one of the first optical pulsars to be confirmed.

The Vela pulsar is located in the southern Milky Way, 2.8° south of the galactic equator and approximately at the centre of a triangle formed by the stars ζ Puppis, γ Velorum and λ Velorum. It is surrounded by extensive nebulosity more than 5° across, also a product of the supernova, and is associated with a moderately strong radio source, Vela X; its distance has been estimated at between 1300 and 1600 light-years. The Vela pulsar and supernova remnant lie within a much older and larger supernova remnant, the Gum Nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vela Supernova Remnant

A supernova remnant in the constellation Vela, position RA $08^{h}\ 34^{m}$, dec. $-45^{\circ}\,45'$. It extends to nearly 5° in diameter, and consists of material expelled by a supernova an estimated 11 000 years ago. The core of the supernova remains as the Vela Pulsar.

See: Vela Pulsar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Velikovsky, Immanuel (1895–1979)

Physician, psychoanalyst, cosmologist and writer, born in Vitebsk, Russia. His book *Worlds in Collision* claimed from analysis of mythology and scientific fragments that a breakaway piece of the planet Jupiter had collided with the Earth, causing various disasters recorded in the Bible, and ending as the planet Venus. None of his theories stood up to analysis and further exploration.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Venera (Missions)

Veneras 1 to 16 are the Russian (Soviet) spacecraft missions to VENUS launched in the period 1961–1983 (table 1). Venus was also studied by the VEGA MISSION.

The first three Venera (hereafter V) missions, V1–V3, were developed by the space center headed by Sergey Korolyov, father of the Russian space industry (1906–1966). (This space center is now a division of the Energia Association.) Since 1967, spacecraft for missions to planets have been developed by the Lavochkin Association. V1–V8 are the first generation of the Venera spacecraft. The spacecraft mass varied from 900 to 1200 kg. The three-stage Soyuz missile, with the first two stages similar to those used for the first manned flights to space, launched a spacecraft and the fourth stage into Earth orbit. Then the fourth stage accelerated the spacecraft and moved it into the interplanetary orbit to Venus. Due to the insufficient reliability of the fourth stage and some subsystems in the early 1960s, some spacecraft were lost or their programs were not completed in full. However, the first generation spacecraft made the first soft entries into the atmosphere and landed on the surface of Venus.

V9 to 16 (and also Vega 1 and 2) represent the second generation of missions to Venus. Their mass was 5000 kg, and they were launched by the Proton rocket. Those missions were successful, reliable and, along with ground-based observations and the USA Mariner 2, 5 and 10, Pioneer Venus and Magellan spacecraft, they provided the basis of our current knowledge of Venus.

Scientific objectives

The objectives of these missions were to study (i) the surface relief and properties of rocks, (ii) the thermal structure and chemical composition of the lower atmosphere ($h < 65$ km), (iii) the atmosphere and ionosphere above the cloud layer ($h > 65$ km), (iv) the magnetic field and plasma environment near Venus, and (v) the solar wind, interplanetary magnetic field, cosmic rays and micrometeorites during the cruise phase.

Scientific instrumentation

V2. Magnetometer, plasma detectors, cosmic ray and micrometeorite sensors, telecamera, infrared (5–40 μm) and ultraviolet (1700–3500 Å) spectrometers, radiometer in the decimeter range. Only some data during the cruise phase were obtained.

V3. *Flyby:* magnetometer, plasma detectors and cosmic ray sensors, photometer at H 1216 Å and O 1304 Å. *Lander:* atmospheric temperature, pressure and density sensors, chemical gas sensors, photometer, γ-ray counter. Only some data during the cruise phase were obtained.

V4. The same as V3 with an altimeter instead of the lander photometer. *Results:* the first entry and direct measurements in the atmosphere down to $h = 23$ km. Pressure at 55–33 km, density at 55–28 km, temperatures at 55–23 km, atmospheric composition ($CO_2 > 90\%$, $N_2 < 7\%$). Hydrogen corona, upper limit to atomic oxygen density, plasma bow shock at 19 000 km from the planet, low density of the nightside plasma, magnetic moment ($< 3 \times 10^{-4}$ of the Earth's value).

V5, 6. Measurements of pressure, density and temperature (p, ρ, T, respectively) at 55–17 km, chemical gas sensors.

V7. First soft landing on Venus. Measurements of T from 55 km to the surface.

V8. p, ρ, T from 55 km to the surface, radar altimeter for determination of altitude, photometer to study the light conditions and the cloud structure (the lower cloud boundary was observed at 33 km), measurements of wind speed using the Doppler shift of the radiotransmitter frequency, γ-spectroscopy of the surface rocks.

V9, 10. *Lander (1560 kg):* γ-spectrometer, γ-densitometer, p, T sensors, camera, accelerometer (p, T at 90–65 km), mass spectrometer, nephelometer (angles of 4, 15, 45 and 175°), photometer for five broad intervals in the range 0.44–1.16 μm, narrow band photometer for 0.8, 0.82 and 0.87 μm (continuum, H_2O and CO_2 bands, respectively), onboard transmitter (Doppler measurements of wind speed). *Orbiter:* imaging system (0.3–0.4 μm), spectrometer (0.24–0.75 μm), IR spectrometer (1.7–2.8 μm), IR radiometer (8–28 μm), multiband photopolarimeter (0.35–0.8 μm), photometer at H 1216 Å with absorption cells, magnetometer, plasma electrostatic spectrometer, plasma sensors, cosmic rays, dual-frequency onboard transmitter (radio occultations and bistatic radar).

V11, 12. *Lander:* p, T sensors, accelerometer, mass spectrometer, gas chromatograph, optical spectrophotometer (0.45–1.2 μm), nephelometer, x-ray fluorescence spectrometer, low-frequency electromagnetic detector, onboard transmitter (Doppler measurements of wind speed). *Flyby:* ultraviolet spectrometer (304–1657 Å), plasma sensors, cosmic rays, γ-ray bursts.

V13, 14. Color camera, rock sampling, penetrometer, electric conductivity of rocks, x-ray fluorescence spectrometers for (a) rocks and (b) aerosols, p, T sensors, accelerometer, mass spectrometer, gas chromatograph, spectrophotometer (0.32–1.2 μm), electromagnetic and microseismic detectors, electric conductivity of aerosol, onboard transmitter (wind speed measurements), γ-ray bursts (cruise).

V15, 16. Synthetic aperture radar, radio altimeter, Fourier transform spectrometer (5–40 μm), dual-frequency onboard transmitter (radio occultations).

Main scientific results

Relief. Detailed mapping of the Venus relief was made with the synthetic aperture radars and radio altimeters (V15, 16). This study covered the vast area north of about 30° that is about a quarter of the total surface of Venus. This mapping had a spatial resolution of 1–2 km, which is sufficient to distinguish various landforms and terrains and determine the geological processes responsible for their formation. This resolution is better than that of 20 km in a similar study from the Pioneer Venus orbiter in 1979

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Venera 1 to 16 missions.

#	Launch	Encounter	Comment
1	12/02/1961	—	First interplanetary spacecraft; flyby 100 000 km
2	12/11/1965	27/02/1966	Flyby 24 000 km
3	16/11/1965	01/03/1966	First hard landing
4	12/06/1967	18/10/1967	First soft entry (to 23 km) and flyby
5	05/01/1969	16/05/1969	Soft entry (to 17 km)
6	10/01/1969	17/05/1969	Soft entry (to 17 km)
7	17/08/1970	15/12/1970	First soft landing
8	27/03/1972	22/07/1972	Soft landing
9	08/06/1975	22/10/1975	Soft landing and first orbiter $h_{\min} = 1560$ km
10	14/06/1975	25/10/1975	Soft landing and orbiter $h_{\min} = 1620$ km
11	09/09/1978	21/12/1978	Soft landing and flyby at 40 000 km
12	14/09/1978	25/12/1978	Soft landing and flyby at 40 000 km
13	30/10/1981	01/03/1982	Soft landing and flyby
14	04/11/1981	05/03/1982	Soft landing and flyby
15	02/06/1983	10/10/1983	Orbiter $h_{\min} = 1000$ km
16	07/06/1983	14/10/1983	Orbiter $h_{\min} = 1000$ km

but not so good as the 0.12 km achieved by the Magellan orbiter seven years later (1990–1994).

The Venus surface (see VENUS: SURFACE) is dominated by plains formed by vast floods of the non-viscous volcanic lavas, probably basalts. About 15% of the mapped area is covered by so-called tessera, which is a specific terrain formed by heavy tectonic deformation of some unknown precursor terrain. The morphology of numerous ring-like volcanic plain features, so-called coronae, is evidence of their formation due to tectonics and volcanism related to the uplifting plumes in the planet's interior. Several large fracture zones sharing many properties with the continental rift zones on Earth were found. Many impact craters of 8 to 150 km diameter show the age of the mapped area to be 0.5–1 billion years. The main conclusion is that the geology on Venus, contrary to that on Earth, is not driven by plate tectonics.

Panoramas of the landing sites of V9, 10 (figure 1), 13 and 14 show the small-scale structure of the surface. Of four sites, only two are similar (V10 and 13). V9 shows a slope of the mountain ridge covered by irregular and sharp-edged stones having sizes up to 0.5–0.7 m. V10 and 13 landed on the rolling plain with outcrops of crystallized magmatic rocks changed by deep chemical weathering. Steeply sloped uplands with layered structured rocks are seen sloping away. V14 found a smooth region of the stone plain (lowland) composed of relatively fresh fine-grained rocks having a horizontal layered structure which resembles the accumulation of volcanic tuff.

Composition of rocks. The abundances of Mg, Al, Si, K, Ca, Ti, Mn, S, Cl and Na given as their oxides were measured by the x-ray fluorescence spectrometers at the V13, 14 and Vega 2 landing sites. Abundances of the natural radionuclides (K, U, Th) were measured by the gamma-ray spectrometers from V8, 9, 10, Vega 1 and 2. Both elemental and radionuclide abundances are typical of various types of erupted basalts, with the exception of V8 which is more similar to granites.

Mechanical properties of rocks. The density of rocks was measured by the gamma-ray densitometer from V10 to be 2.8 ± 0.1 g cm^{-3}, which is typical of crystalline basalts. Determinations of the bearing strength and firmness of rocks were made by V13 and 14 using three methods: (i) a penetrometer which impacted the ground, (ii) sampling the rocks by drilling for the x-ray fluorescence analyses, and (iii) accelerometer observations during the shock of landing. Heavy clays and compacted dust-like sand are the terrestrial analogs for V13, and volcanic tuffs and fissured rocks are those for V14.

Temperature, pressure and dynamics in the lower atmosphere (0–60 km) were measured by all landing and entry probes. T and p vary from $T = 735$ K and $p = 90$ bar near the surface to 260 K and 0.2 bar at 60 km. The temperature lapse rates (gradients) may be compared with the adiabatic lapse rate for the CO_2 (97%) + N_2 (3%) mixture which varies from -11.2 K km^{-1} at 60 km to -8.0 K km^{-1} near the surface. The adiabatic lapse rate corresponds to the static stability of the atmosphere. Significant deviations from static stability were observed at 50–35 km and 22–7 km. Wind speed (measured using the Doppler shift of the radiotransmitter frequency) varies from 70 m s^{-1} at 50 km to 0.5 m s^{-1} near the surface. See also VENUS: ATMOSPHERE.

Abundances of noble gases and isotope ratios were measured by the V11–14 mass spectrometers. The total argon abundance is 100 parts per million (ppm). The ratio of primary isotopes, ^{36}Ar/^{38}Ar, is that on Earth, while the radiogenic isotope ^{40}Ar is depleted by a factor of 270 relative to the Earth's value. The abundance of Ne is 8 ppm, and ^{20}Ne/^{22}Ne $= 12.15 \pm 0.1$, which is between the ratios found on the Earth's (10.1 ± 0.4) and in the solar wind (13.7 ± 0.3). The abundances of Kr and Xe are 35 parts per billion (ppb) and less than 20 ppb, respectively. The isotope ratios for C and O are equal to the Earth's ratios. All these values are important for the origin and evolution of the atmosphere.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 1. Panoramas of the surface of Venus observed from V9 and V10. Each panorama is $180^\circ \times 40^\circ$.

The *chemical composition of the atmosphere* was studied by the V12–14 gas chromatographs: $N_2 \approx 3\%$ and Ar = 100 ppm (both were also measured by the mass spectrometers), $SO_2 = 130 \pm 35$ ppm, CO varies from 30 ppm at 36 km to 17 ppm at 12 km. Detections of some other gases are disputable.

Strong CO_2 and H_2O bands observed at various heights with the V11–14 spectrophotometers resulted in an estimate of the full water abundance of 1 g cm^{-2} above the surface (first done at V11). The absorption at 0.37 μm observed with V14 below 58 km is due to SO_2, which varies from 50 ppm at 50 km to 10 ppm at 57 km.

Observations with the V15 Fourier spectrometer resulted in H_2O abundances of 5–15 ppm at 60 ± 3 km. SO_2 abundances at 70 km were equal to 20–40 ppb at the low and middle latitudes, 1–10 ppb in the polar collar, and 100–1000 ppb near the north pole.

Structure of clouds and aerosols. Illumination. Analysis of the solar radiation scattered in the atmosphere and measured by the photometers and spectrophotometers onboard the V8–14 landing probes established three main layers within the cloud deck structure and the lower cloud boundary at 48 km (near 35 km at V8 and Vega 1 and 2). Optical depths of the layers were derived at each landing sites. The V14 near-UV (0.32–0.39 μm) photometer established two absorption layers, above 60 km and below 58 km. 90% of the solar near-UV energy is absorbed above 60 km and supports strong winds in the upper cloud level. A few per cent of the solar flux reaches the surface.

A nephelometer is an instrument that illuminates a volume of atmospheric medium by a narrow beam and measures the scattered radiation at fixed angles. Four-angle nephelometers were used onboard V9 and 10 and one-angle (backscattering) nephelometers on V11, 13 and 14. They also established the three-layer structure of the main cloud deck and determined aerosol properties in those layers.

The *chemical composition of clouds* was measured by the V12 and 14 x-ray fluorescence spectrometers, which determined the presence of S, Cl and Fe, with the Fe/Cl ratio corresponding to $FeCl_3$. Later, on Vega 1 and 2, this method was used to measure the abundances of S, Cl and P. Gas chromatographic and mass spectrometric analyses of aerosol pyrolysis products (Vega 1 and 2) revealed sulfuric acid and sulfur aerosols with the mass ratio 1:0.1. $FeCl_3$ is probably present as very small (0.3 μm) particles and constitutes 1% of the total aerosol mass. These particles may serve as condensation centers for sulfuric acid and explain the near-UV absorption. The Vega data favor P as phosphoric acid H_3PO_4 at locations where the lower cloud boundary extends to 35 km. The observed excess of Cl may indicate $AlCl_3$.

Lightning. Strong and frequent electromagnetic pulses at 10–100 kHz observed at V11 to 14 favor lightning on Venus. The Pioneer Venus electromagnetic observations also support lightning. However, of three sets of observations in the visible (V9, 10, and Pioneer Venus orbiters, Vega 1 and 2 balloons), there was only one case for suspecting a thunderstorm. Overall, the problem is uncertain.

Properties of clouds and the atmosphere near the upper cloud boundary (68 km) were studied by the near-UV camera, visible spectrometer, photopolarimeter, near-IR spectrometer and IR radiometer onboard the V9 and 10 orbiters, and Fourier spectrometer (FS) on V15.

FS covered the range 5–40 μm with a spectral resolving power $\lambda/\delta\lambda \approx 200$. 1500 spectra obtained mostly in the northern hemisphere showed sulfuric acid as a main cloud species at all latitudes. The cloud deck is homogeneous with a smooth decrease in particle density at latitudes $< 50^\circ$, and is horizontally variable with a steep vertical decrease at the polar and subpolar regions.

Limb observations of haze above the clouds by the V9 and 10 visible spectrometers resulted in vertical profiles of the haze particle size, their number density and eddy diffusion at 70–100 km.

Pressure, temperature and dynamics in the middle atmosphere were studied by three methods: (i) accelerometers on the entry and landing probes, (ii) dual-frequency radio

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

occultations from the V9, 10, 15 and 16 orbiters, and (iii) FS spectra of the CO_2 band at 15 μm.

50 and 90 p–T profiles in the range 40–90 km were observed with radio occultations from V9, 10 and 15, 16, respectively, to study the thermal balance, day–night variations and latitudinal behavior with a special emphasis on the polar and collar regions. Temperature profiles deduced from the FS spectra are different at low and high latitudes and depend also on local time. Wind velocities ($\approx$ 100 m s^{-1} at 70 km) estimated from the temperature profiles vary with local time, showing the half-day period typical of the atmospheric tides.

Airglow. Observations with the high-sensitive visible spectrometers (V9 and 10 orbiters) solved the three-centuries-old problem of the Venus ashen light (nonzero brightness of the night side) and revealed spectra of the nightglow. Laboratory simulations of these spectra helped to identify four O_2 band systems. Their intensities, vertical profiles and variations with local time were obtained. These emissions originate near $h = 100$ km in the reaction $O + O + CO_2 \rightarrow O_2^* + CO_2$ with subsequent quenching and excitation trasnfer.

Dayglow emissions of H, O, He and O^+ were observed with the V9 and 10 airglow photometers and V11 and 12 spectrometers. The densities of these species and the temperature of the upper atmosphere were obtained from the observations.

Profiles of electron density in the ionosphere were measured using dual-frequency radio occultations from the V9, 10, 15 and 16 orbiters. About 100 profiles were used to study variations of the density maximum and its height, the scale height above the maximum, and the ionopause height with solar zenith angle, solar activity and other geophysical parameters. The nightside ionosphere is especially variable and often shows a two-peak structure.

The *magnetosphere and interaction with the solar wind* were studied by two plasma spectrometers and a magnetometer onboard the V9 and 10 orbiters. The shape and location of the bow shock indicated that the solar wind is diverted almost completely by a magnetic barrier. Charge exchange between the solar wind protons and atmospheric hot oxygen atoms and their photoionization add mass to the solar wind. This mass loading was observed in the boundary layer of the solar wind flow around the planet and plays a crucial role in the formation of the magnetosphere. The planetary ion outflow on the nightside forms the distinct plasma tail. Its polarity is controlled by the local direction of the interplanetary magnetic field, thus confirming the induced nature of the Venus magnetosphere. Both plasma flow from the dayside and electron precipitations on the nightside support the nightside ionosphere (see also VENUS: INTERACTION WITH SOLAR WIND).

Bibliography

Barsukov V L, Basilevsky A T, Volkov V P and Zharkov V N (ed) 1992 *Venus Geology, Geochemistry, and Geophysics* (Tucson, AZ: University of Arizona Press)

Hunten D M, Colin L, Donahue T M and Moroz V I (ed) 1983 *Venus* (Tucson, AZ: University of Arizona Press)

Keldysh M V 1977 Venus exploration with the Venera 9 and Venera 10 spacecraft *Icarus* **30** 605–25

Krasnopolsky V A 1986 *Photochemistry of the Atmospheres of Mars and Venus* (Heidelberg: Springer)

Moroz V I 1981 The atmosphere of Venus *Space Sci. Rev.* **29** 3–127

Vladimir A Krasnopolsky

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Venus

Venus is the second planet from the Sun. It orbits the Sun at 0.72 times the Earth's distance, or 108.2 million km. Because of its location relative to the Earth, Venus is always seen close to the Sun, at less than 48° angular distance from it. The planet appears either as a morning 'star' or as an evening 'star'. Actually, as in the case of Mercury, for a long time it was thought that there were two different bodies (the ancient names for Venus were Eosphorus and Hesperus), until the Greek astronomers realized that there was only one body. The distance of Venus to the Earth varies from 41 million km to 257 million km. As seen from the Earth, Venus's apparent diameter varies from 10 to 64.5 arcsec. However, we see the full disk illuminated by the Sun only at Superior conjunction. The planet has phases, like the Moon (figure 1). They were first seen by Galileo in 1610. Venus is the brightest object in the sky, after the Sun and the Moon, but the planet's brightness is very variable. Venus is the brightest between maximum elongation and inferior conjunction. At inferior conjunction, which happens about every 19 months, when Venus's size is the largest, the disk is completely dark at visible wavelengths. The brightness of Venus is due not only to its close proximity to the Sun (Venus receives twice more solar light as the Earth) and to the Earth but also to the high reflecting power of the clouds that reflect 75% of the solar light.

The rotation of Venus is nearly synchronous. Indeed, the period of rotation of Venus around the Sun is 224.7 days, while the planet's rotational period is 243 days (Earth days), which is extremely slow, and in the retrograde direction (i.e. opposite to the direction of the movement of the rotation of the planet around the Sun). Because of the combined effects of the slow retrograde rotation of Venus and of its orbital motion around the sun, the solar day (time between two sunrises) on Venus corresponds to 117 Earth days. The eccentricity of the orbit and the inclination of the planetary rotation axis are both very small. Therefore no major seasonal changes are expected. In fact, Venus is the planet with the most circular orbit. It is also the most Earth-like of the terrestrial planets in terms of size and bulk properties. The planetary radius at the equator is 6051.5 km, which corresponds to 95% of the Earth's radius (6378 km). Furthermore, Venus's mean density (5.24 g cm^{-3}) and surface gravity (8.6 m s^{-2}) are comparable with those of the Earth (5.97 g cm^{-3} and 9.78 m s^{-2}, respectively). However, now that we know Venus better, we find many more differences from the Earth than was believed only 50 yr ago. In particular, Venus has a very high surface temperature of 740 K (hot enough to melt lead) and a very dense atmosphere (93 times that on Earth). The main atmospheric constituent is carbon dioxide, instead of nitrogen for the Earth. The clouds are made of sulfuric acid and not pure water as on Earth. Besides, the surface of Venus is essentially of volcanic origin and does not present evidence for lateral plate tectonics in recent history. Furthermore, Venus has no intrinsic magnetic field. In addition, Venus does not possess any satellites.

Earth-based investigations

Until the first successful space mission to Venus in 1962, all that was known about Venus came from Earth-based investigations. In 1761, M V Lomonosov, who observed a transit of Venus accross the Sun in an attempt to measure the distance from the Earth to the Sun, was the first to detect a halo around Venus that he interpreted as being due to the presence of an atmosphere. Spots or markings on the planet had been seen since about 1666. Around 1930, ultraviolet measurements detected the Y-shaped features of the upper clouds. Studies of these markings established in 1966 that the atmosphere was rotating in about 5 days at the cloud tops. These features appear much better in the ultraviolet than in the visible. The first identification of an atmospheric constituent was that of CO_2 in 1932 by spectroscopy at 8000 Å. Later on, H_2O was detected from the ground, but also from balloons and aircraft, in visible spectra of the planet and in very variable amounts. Infrared spectroscopy from Earth around 1967 led to the detections of CO, HCl and HF in small amounts above the clouds. Ultraviolet measurements allowed the detection of SO_2 in 1979. The cloud composition remained unknown for a long time. It was believed that the clouds were made of water, as on Earth. It is only around 1974, through the analysis of the polarization of Venus's radiation, that a value for the refractive index of the particles was obtained, $n = 1.45$, and it was then correctly inferred that the cloud particles were in fact droplets of concentrated sulfuric acid (H_2SO_4). This was confirmed later on by space missions, which provided much more information about the clouds. However, it is interesting to note that an explanation for the ultraviolet markings in the upper clouds is still lacking, although it is suspected that they are related to sulfur chemistry.

Measurements of the radio emission of the planet at 3.15 cm in the late 1950s seemed to indicate that the temperature at the surface of Venus was much higher than that on Earth. As it was believed at that time that Earth and Venus were very similar planets, this came as a big surprise, and a confirmation was needed. This was one of the goals of the first American spacecraft that went to Venus in 1962.

Concerning the surface itself, it is only by radar measurements that are not stopped by clouds that information on the surface could be gained. Information on the rotation of the planet as a solid body came only with the first radar studies of Venus with the Goldstone radio antenna around 1964–9 that allowed the direction and rate of rotation of the solid planet to be obtained. The extremely slow rotation thus inferred, and the fact that Venus was rotating in the retrograde direction were other significant surprises. It was also found at that time that the rotation axis is nearly perpendicular to the ecliptic. Radar observations from Goldstone, California, and Arecibo, Puerto Rico, at 12.6 cm (S-band), around Venus's inferior

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

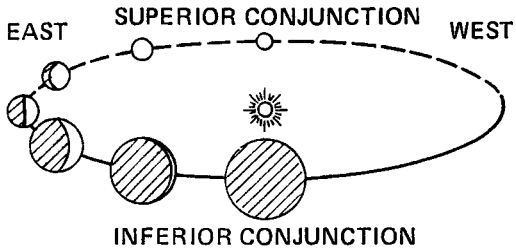


Figure 1. A schematic view of Venus as observed from Earth at various phases. Venus is largest at inferior conjunction, but then it is very dark. At superior conjunction, Venus is smallest but completely illuminated by the Sun. Venus is brightest around east and west elongations. (From Colin L 1983 *Venus* ed D M Hunten, L Colin, T M Donahue and V I Moroz (Tucson, AZ: University of Arizona Press) p 21.)

conjunction, provided the first radar images of portions of the surface at 5–20 km resolution.

The studies of the atmosphere made nowadays by Earth-based telescopes cover all wavelength ranges from the ultraviolet to the radio range. Earth-orbiting ultraviolet satellites as well as sounding rockets are also used regularly to probe the upper atmosphere of the planet.

As the planet is covered by thick cloud layers, the part of the atmosphere below the clouds is difficult to probe by Earth-based observations. For many years, information on the deep atmosphere came essentially from space probes that went into the atmosphere or from occultations of spacecraft by the atmosphere of the planet during a space mission. However, in 1984, it was discovered that the night-side of Venus emits near-infrared radiation coming from deep atmospheric layers below the main clouds. This property has since been used to probe these deep layers using Earth-based telescopes. It led to measurements of sulfur compounds (SO_2, OCS), CO, H_2O and halides (HCl, HF), below the clouds.

Radio measurements that are made today concern mostly the millimeter range, and they pertain to the upper atmosphere of the planet. They provide information on minor species such as CO and H_2O and on the temperature profile in the upper atmosphere. They allow the variations with the time of the day to be studied. For instance, larger concentrations of CO have been found on the night-side than on the day-side. Also, Doppler shifts of the CO lines have been used to measure wind speeds. Long-term variations in CO have also been observed by this method.

Space missions

Space exploration of Venus started in 1962. Since then, there have been many spacecraft explorations with fly-bys and orbiting spacecraft. The most favorable periods for launch occur at times of inferior conjunction, when Venus is closest to the Earth, which happens about every 19 months. Venus is the planet that has been the most explored by space missions.

The first fly-by was that of the American Mariner 2 spacecraft, in 1962, which carried a microwave radiometer to check whether the surface temperature was high. Then, another American spacecraft, Mariner 5, that went to Venus in 1967, confirmed that the surface pressure was very high, about 90 atm. Mariner 10 on its way to Mercury explored Venus in 1974 (see MARINER MISSIONS). The Soviets began exploration of Venus in 1967 with Venera 4 (first successful launch) which carried an atmospheric probe. The first successful landing on Venus was achieved by Venera 7 in 1970 (another Soviet spacecraft), and the first photographs of the surface were obtained by the Venera 9 and 10 Soviet spacecraft in 1975 (see VENERA MISSIONS). Venera 8, 9 and 10 lander capsules carried gamma-ray spectrometers, which allowed the first geochemical measurements to be made. The amounts of radioactive elements U (uranium), Th (thorium) and K (potassium) were determined.

Then, in 1978, came the American Pioneer Venus 1 and 2 missions, with an orbiter on Pioneer Venus 1 and entry probes on Pioneer Venus 2. Pioneer Venus orbiter penetrated the upper atmosphere and made mass spectrometer measurements. Four probes reached the surface. The orbiter part of the mission lasted from December 1978 to October 1992. Topographic maps were obtained by the radar altimeter on Pioneer Venus 1 in 1978 with a resolution of about 50 km in horizontal extent and about 200 m in altitude (see PIONEER VENUS MISSION). The Venera 11 and 12 spacecraft, which carried descent landers, were also sent to Venus by the Soviet Union in 1978. Altogether, in December 1978, ten separate spacecraft visited Venus.

The Soviet exploration continued with other Venera missions (up to Venera 16) and with the Vega 1 and 2 spacecraft that were initially designed for an exploration of Venus but were sent to comet Halley as well as to Venus (see VEGA SPACE MISSION). The Venera 13 and 14 probes, which landed on Venus in 1982, and the Vega 2 probe, which dropped a landing module in June 1985, carried an instrument that was able to measure the x-ray fluorescence of soil samples introduced into a vacuum chamber. Venera 13 and 14 obtained some more images of the surface of Venus, and the Venera 15 and 16 probes provided radar mapping at a resolution of 1–2 km of the northern 25% of the surface of Venus close to the north pole, starting in 1983.

Another major step in the exploration of Venus was achieved by the American Magellan Radar Mapping spacecraft that was launched from the space shuttle Atlantis in 1989. Between 1990 and 1992, it mapped 98% of the surface at resolutions between 120 and 300 m. It also provided altimetric, radiometric and gravity data. The final orbit occurred on October 1994. The radar maps are spectacular. The American spacecraft, Galileo, designed to explore the planet Jupiter, accomplished a fly-by of Venus in February 1990. It has allowed us, in particular, to study the deep clouds of Venus from near-infrared imaging of the night-side.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The atmosphere

Composition and temperature

Our current knowledge of the composition of the atmosphere of Venus comes from a number of *in situ* analyses with infrared and ultraviolet spectrometers, mass spectrometers and gas chromatographs on space probes that went into the atmosphere of the planet. It is also the result of spectroscopy obtained with Earth-based telescopes, airplanes, rockets and satellites. Venus has a very dense atmosphere, with CO_2 as the major constituent (mixing ratio: 0.965), instead of N_2 as on Earth. Minor components are nitrogen (N_2), water vapor (H_2O), sulfur dioxide (SO_2), sulfur monoxide (SO), carbon monoxide (CO), carbonyl sulfide (COS), sulfuric acid (H_2SO_4), oxygen (O_2), hydrogen chloride (HCl), hydrogen fluoride (HF), argon (Ar), neon (Ne) and krypton (Kr) (see VENUS: ATMOSPHERE). The atmospheric surface pressure is 95 bar, or 95 times that at the surface of the Earth. The temperature at the surface is very high: 740 K.

One of the striking characteristics of Venus is its dryness. At the high surface temperature of Venus, water could not currently exist in the liquid state at the surface. However, even in the atmosphere, only very small amounts of H_2O vapor are found. The deuterium-to-hydrogen ratio is very high (about 150 times that on Earth). In contrast, isotopic ratios of C and O are the same on Venus and Earth.

Concerning the nonradiogenic noble gases ^{20}Ne, ^{36}Ar and Kr, ^{20}Ne is about 20 times more abundant on Venus than on Earth, ^{36}Ar about 70 times more abundant and Kr only 3 times more abundant. Radiogenic argon (^{40}Ar) is about 4 times less abundant on Venus than on Earth.

The high average surface temperature is somewhat surprising. Indeed, in spite of the close proximity of Venus to the Sun, because of the high reflecting power of the clouds, only about 25% of the solar flux penetrates into the atmosphere. However, the high surface temperature can be explained by a very efficient greenhouse effect as the atmosphere of Venus contains gases that are strong infrared absorbers (CO_2, SO_2, H_2O, OCS) and clouds that trap the thermal radiation emitted by the deep atmosphere. This causes an increase in the surface temperature of nearly 500 K (compared with the expected temperature without an atmosphere).

The measurements of temperature in the atmosphere come from infrared radiometry and radio occultations. Spacecraft radio-occultation measurements indicate that the temperature decreases from 740 K at the surface to about 240 K at the cloud tops. Temperatures below 35–50 km approximately follow an adiabat. This indicates that the temperature structure below the clouds is controlled essentially by atmospheric dynamics (convection). Diurnal and latitudinal variations are very small below the clouds.

Above the clouds, in the region called the mesosphere, temperature measurements obtained from Pioneer Venus around 15 μm have provided maps covering the 60–105 km vertical range. The temperature (averaged over time) varies from 170 to 330 K. These maps indicate that, at these altitudes, the polar regions are up to 20 K warmer than the equator, which is very surprising.

At still higher altitudes, in the thermosphere, the temperature in the day-side is around 300 K, and therefore much colder than on Earth (where it is 1000–2000 K). Heating in this part of the atmosphere is caused mainly by the absorption of ultraviolet photons from the Sun. The difference in temperature with respect to the Earth is due to the greater abundance of CO_2 which is very efficient at radiating heat to space. The temperature varies spectacularly from day to night. At night, it can be as cold as 100 K. The long duration of the days on Venus and the cooling effect of CO_2 on the night-side probably combine to produce this strong contrast in temperature.

The homopause is the region of the atmosphere above which the gases are no longer uniformly mixed (gases of different mass have different scale heights). On Venus, this occurs at around 130–145 km altitude. What is found is a surprisingly low concentration of the dissociation products of CO_2—O and CO—above the homopause. These products must be quickly removed and transported to lower altitudes where they recombine (maybe through a catalytic cycle that involves chlorine, which is about 1000 times more abundant on Venus than on Earth).

Still higher up, Venus has an ionosphere, which was first detected by the radio-occultation experiment on Mariner 5 in 1967 and has been studied in more detail by an ion-mass spectrometer that went to the ionosphere and by remote-sensing observations with other instruments on the Pioneer Venus orbiter. The day-side ionosphere peaks near 140 km and stops at about 500 km. It is produced primarily by solar extreme ultraviolet radiation. The most abundant positive ions are O_2^+, O^+ and CO_2^+. O_2^+ (and not CO_2^+) is the dominant ion up to about 190 km because of the rapid reaction rate of CO_2 + O. Higher up, O^+ dominates. Quite surprisingly, the night-side ionosphere can sometimes be rather significant. This happens during periods of solar maximum when the day-side ionosphere is so dense and extensive that the solar wind cannot penetrate the planet's environment. A large flux of O^+ ions then goes to the night-side of the planet and creates a significant night-side ionosphere (the long duration of the nights prevents the ions from recombining). At solar minimum, the day-side ionosphere is much less extensive, and the flow of ions to the night-side not as strong (see VENUS: INTERACTION WITH SOLAR WIND).

Clouds

Although the upper part of the cloud layers can be observed and monitored from the Earth, the vertical structure of the clouds can be studied only during space missions. Probes and the Vega 1 balloons have gone through the cloud layers, and the Galileo spacecraft has imaged the clouds on the night-side of the planet around 40–50 km altitude during its fly-by of Venus.

Clouds cover Venus globally in generally three layers between 48 and 68 km. At some times and at some

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

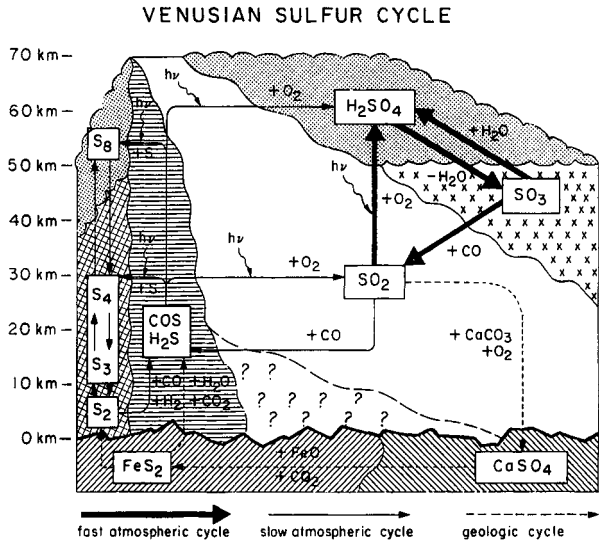


Figure 2. A possible model for the Venusian sulfur cycle (from Lewis J S and Prinn R G 1984 *Planets and their Atmospheres: Origin and Evolution* (New York: Academic).

locations, the lower two layers may combine. However, the variability in structure is much less important than in the Earth's clouds. Hazes are present above and below the clouds. The total cloud optical depth has been estimated to be between 25 and 35. Their total mass is relatively small; the Venus clouds are more like the hazes on Earth. The droplets are about 2 μm in size, but the particle size distributions differ from layer to layer.

The clouds are made mostly of concentrated sulfuric acid, with about 75% by weight sulfuric acid and 25% water vapor. They resemble terrestrial acid rain (caused by pollution). This composition is probably the result of photochemical production of SO_3 from SO_2. H_2O and SO_3 then react to make H_2SO_4 (figure 2). H_2SO_4 is thermally decomposed into SO_3 and H_2O in the lower atmosphere. Chemical reactions with the surface should be removing sulfur quickly from the atmosphere, maybe through the formation of anhydrite ($CaSO_4$). Some sulfur must go back into the atmosphere from the decomposition of sulfur compounds present in the crust, such as pyrite (FeS_2). To explain the presence of permanent cloud layers of sulfuric acid on Venus, a source of sulfur (continuous or episodic) is needed, and it is postulated that surface volcanism plays an essential role in the release of sulfur gases from the crust (see also CLOUDS IN PLANETARY ATMOSPHERES).

Atmospheric circulation

The atmospheric circulation of Venus is very different from that on Earth (see EARTH'S ATMOSPHERE). This is not surprising given the different rotation rates, the different atmospheric composition and structure, and the lack of seasons. Furthermore, on Venus, solar light is mainly deposited within and above the upper cloud layer which is always present, whereas on Earth it is mostly absorbed at the surface.

What we know about the global circulation in the atmosphere of Venus comes from various sources of information. Direct sources are cloud-feature tracking and the radio tracking of descending probes (Doppler tracking of Veneras 8, 9, 10 and 12 and interferometric tracking of the Pioneer Venus probes) and of balloons that went into the atmosphere (Vega mission). Doppler-shifted emission lines from atmospheric gases (such as CO_2 and CO) also provide direct measurements of wind speeds. Spatial variations in gaseous abundances are also used as tracers of the global circulation. In a more indirect way, theoretical analyses of planet-wide variations in temperature at a given pressure, obtained by radio occultations and IR radiometry, have been used. However, winds can be determined in this manner only if the atmosphere is in cyclostrophic balance, a condition which, on Venus, occurs in the non-equatorial regions below 60 km altitude.

As we have seen, Venus rotates extremely slowly, in 243 days, whereas the atmosphere at the cloud levels (around 60 km altitude) rotates in 4–5 days, which corresponds to wind speeds of 100 m s^{-1}, in the same retrograde direction as the planet. This zonal (along a line of constant latitude) superrotation of the entire atmosphere of Venus from about 10 km up to about 100 km, which is so different from what is found on Earth, is the main mode of circulation of the atmosphere.

The zonal winds are not constant with altitude, however. They decrease regularly from a speed of 100 m s^{-1} at the cloud tops to values close to 1 m s^{-1} near the surface. However, there exist some layers of relatively constant wind speeds that alternate with regions of large increase in velocity. Differences in zonal wind speeds with time of the day also exist.

Above the clouds, zonal winds generally decrease with altitude up to about 100 km, where they reach very small values. At still higher altitudes, a very different circulation regime exists. The global circulation is there dominated by a strong flow from the day-side hemisphere to the night-side hemisphere (subsolar to antisolar circulation). Peak velocities reach approximately 250 m s^{-1} across the terminator. This circulation, which is also different from that existing on Earth, is a consequence of the high temperature contrasts between the day-side and the night-side above 150 km. With only a strong subsolar-to-antisolar flow, one would expect upwelling centered on the subsolar point and subsidence centered on the antisolar point, with a return flow around 70–90 km altitude. However, subsidence does not occur at midnight but, instead, at around 2–3 a.m. local time. This asymmetry may be due to the superposition of the subsolar-to-antisolar flow and of some zonal retrograde winds (these two flows add across the evening terminator, but go in opposite directions at the morning terminator).

There is, in addition, a slow meridional (in the equator-to-pole direction) circulation that may consist of Hadley cells, as on Earth. Cloud-tracked winds from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

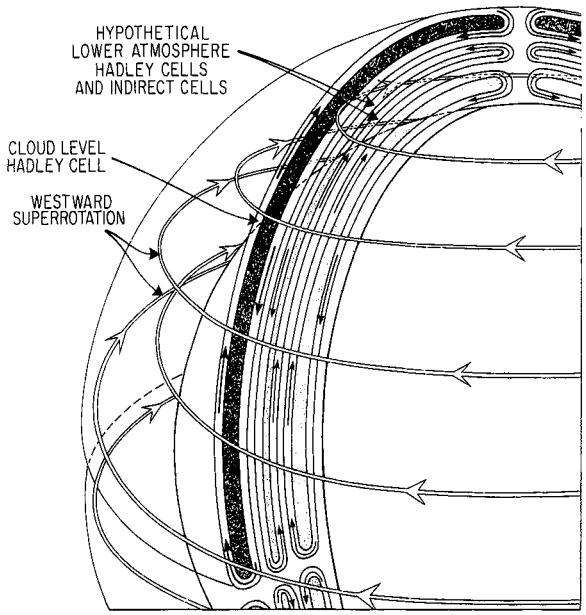


Figure 3. A possible model for the general circulation of the Venusian atmosphere (from Schubert G 1983 *Venus* ed D M Hunten, L Colin, T M Donahue and V I Moroz (Tucson, AZ: University of Arizona Press) p 730.)

Pioneer Venus and Mariner 10 have shown equator-to-pole velocities of about 5 m s^{-1} in each hemisphere. However, the interferometric tracking of the Pioneer Venus probes has revealed a complex altitude variation, with changes in magnitude and in direction. Interpretations of these observations and theoretical considerations suggest that the Pioneer Venus probes went through several vertically layered large meridional cells that each extend from the equator to high latitudes. There may be three different cells: a Hadley cell near the surface, a weak indirect cell in the stable layer below the clouds (between 30 and 50 km) and a Hadley cell at the cloud level (that may be 'thermally indirect') (figure 3).

There are many indications of wave motions on different scales in the atmosphere of Venus. In particular, the relative stability of the large-scale dark markings in the ultraviolet photographs may be due to planetary-scale waves drifting slowly with respect to the atmospheric superrotation. The thermal tides that follow the motion of the Sun in Venus's sky (the Sun rises in the west and sets in the east) may also play an important role. There is also evidence for small-scale turbulence. One of the main problems in Venus's global circulation is to explain the existence and maintenance of the superrotation. The eddies that have been observed in the lower atmosphere may, in combination with the mean meridional circulation and the planetary-scale waves, help in providing the upward transport of angular momentum required to explain Venus superrotation.

Origin and evolution

Although giant planets have retained most of their primitive atmosphere, this is not the case for the terrestrial planets Venus, the Earth and Mars that have lost their primitive atmosphere, probably through massive escape, and have acquired a secondary atmosphere. A large part is made of gases outgassed from the interior which is heated during planetary formation by contraction and by decay of natural radioactive elements. During periods of large-scale volcanism, CO_2, SO_2, H_2O and other volatiles are injected into the atmospheres. An extra component comes from cometary and meteoritic impacts (for the Earth, biological processes also play a role). Gases can be removed by atmospheric escape (solar wind interaction), interaction with the surface or major impacts. See also PLANETARY ATMOSPHERES.

As Venus and Earth were formed in approximately the same region of the solar nebula, they may have had the same global composition. There are some indication that this may be the case. Indeed, the total estimated oxidized carbon inventories on Venus and Earth are comparable, although on Earth the carbon is mainly present in carbonate rocks at the surface of the planet and in the oceans while, on Venus, it is mainly present in the atmosphere, where it can accumulate because of the absence of liquid water. Nitrogen inventories on Earth and Venus are also comparable.

The underabundance of radiogenic argon ^{40}Ar (produced by the decay of ^{40}K—potassium) could be due to different outgassing rates and/or duration on Venus and on Earth. If it is due essentially to a shorter duration of outgassing on Venus than on Earth, then the measurement would indicate that most of Venus's outgassing happened during the first billion years.

The high values of the nonradiogenic noble gases ^{20}Ne, ^{36}Ar and Kr on Venus compared with Earth have not yet found a completely satisfactory explanation. The postulated early episode of massive supersonic escape of the early hydrogen-rich atmospheres by enhanced extreme ultraviolet radiation from the young Sun would probably have carried along the noble gases on both Venus and the Earth, in spite of their high molecular weight. The only major difference in the evolution of the two planets may come from the giant impact with a planetesimal that must have occurred on Earth (but not on Venus) to explain the Moon's formation. This would have produced another major erosion of the Earth's atmosphere, including the noble gases, and, according to some scientists, it could be the main reason for smaller abundances of noble gases on Earth than on Venus (except for Xe, which may be a particular case because of its higher molecular weight). The current endowment in these noble gases for the Earth would come essentially from subsequent periods of outgassing from the interior, with, maybe, some additional contribution of less destructive and more recent cometary or meteoritic impacts that could enrich the atmosphere in noble gases. More measurements of noble gas abundances

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

on Venus, as well as measurements in comets, are needed to enable the further development of these hypotheses.

Venus's atmosphere currently contains about 100 000 times less water than is found in the oceans and atmosphere of the Earth. Venus could originally have had the same amount of water as the Earth. A significant loss of water could have occurred on Venus, but not on Earth, because of a runaway greenhouse effect due to the much higher temperatures of Venus that would have vaporized the water. The water molecules dissociate in the upper atmosphere, and hydrogen then escapes. After a period of massive escape, it has been computed that fractionation would start (the heavier deuterium D does not escape as easily as the lighter H) after the amount of water has been reduced to the equivalent of a layer of water a few meters thick covering the surface of Venus. The very high value of the D/H ratio measured on Venus could therefore indicate that much larger amounts of water were present in the past, with, maybe, the possibility that OCEANS once existed on the planet. However, such calculations are not simple, as it is hard to track the amount of water present at a given time. Indeed, during the more recent past, significant ejections of water could have been produced by a major volcanic episode or by a massive outgassing event. Furthermore, cometary impacts continually replenish the atmosphere with relatively weakly deuterated water.

Surface and interior of Venus

Little was known about the surface until the 1970s. After the first observations by Earth-based Goldstone and Arecibo radar mapping in the 1960s, global-scale radar mappings of large parts of the surface were carried out by Pioneer Venus in 1978 and Venera 15 and 16 in 1983. This was then followed by the US Magellan mission in 1990–4 that has provided radar images, altimetric and radiometric data at a remarkable spatial resolution of 120–300 m for 98% of the surface of Venus.

Main characteristics of the surface

The radar altimetric data indicate that more than 90% of the surface has an elevation between −1 and +2.5 km, compared with Venus's reference mean radius of 6051.8 km. The surface is dominated by plains. One can distinguish the lowland plains (or planitiae), such as the Atalanta Planitia, that have elevations between 1 and 2 km below the mean Venusian radius and that cover about 27% of the surface, and the upland or rolling plains that constitute 65% of the surface and have elevations between 0 and 2 km above the mean surface radius. The rest (about 8% of the surface) is made of highlands with 2–12 km elevation.

The lowland plains are generally very smooth (dark at radar wavelengths), whereas the upland or rolling plains are a little more rough and possess many small-scale landforms such as scarps, ridges, troughs, hills, channels, ... The plains are interrupted by the large-scale landforms at higher elevations called highlands (see VENUS: SURFACE).

The highlands are dominated by two continent-size features: APHRODITE TERRA and ISHTAR TERRA (figure 4). The largest one, Aphrodite Terra (about 10 000 km across), is larger than Africa in surface area. It lies mostly in the southern hemisphere, close to the equator, at 1–5 km elevation. Ishtar Terra, about 5600 km across (about the size of Australia), and located in the northern hemisphere, possesses the highest mountain belts on the planet, including Maxwell Montes which is up to about 12 km high. Following V L Hansen and colleagues, the highlands can be subdivided into three groups: the volcanic rises, the crustal plateaux and Ishtar Terra. The volcanic rises are more than 1000 km across, they rise 1–2.5 km above the surrounding plains and they have gentle slopes. They include the Atla, Beta, Bell, Dione, Imdr and Themis Regiones. The most prominent volcanoes on Venus and numerous coronae (which are circular to elliptical features with tectonic annuli unique to Venus) are found there. The volcanic rises are mostly located in equatorial regions of Venus. The crustal plateaux such as Ovda and Thetis in Aphrodite Terra, or the Alpha, Phoebe or Tellus Regiones are 1000–3000 km across; they have steep slopes and reach elevations of 2–4 km above the surrounding plains. They contain regions of very rough, highly deformed terrain called tesserae, and some volcanic features, as well as many other structures such as ridges, fractures and grabens. Ishtar Terra has unique characteristics; it is an irregular platform with a complex topographic profile that includes a high interior plateau, Lakshmi Planum, with mountain belts in its periphery, and surrounding high tesserae. In addition, its tesserae are different from the ones found elsewhere. They have more coherent structures.

Thousands of volcanoes have been identified on Venus, from kilometer-size vents to broad shields hundreds of km across. Volcanoes and other volcanic features are found everywhere, but they are less abundant in the tesserae regions and more abundant in the Beta–Atla–Themis region which covers about 20% of the planet. So far, 167 volcanoes more than 100 km across have been found. Examples are Maat Mons in Atla Regio (figure 5) and Theia Mons in Beta Regio, which have diameters in the range 100–200 km and rise to 1–9 km above the mean planetary radius. These volcanoes are surrounded by radar-bright (rough) ejecta with circular or elliptical shapes that sometimes form kinds of festoons. About 300 intermediate-size volcanoes have been found (20–100 km in diameter), including several 'pancake domes' that are steep-sided, circular, flat-top features. The smallest volcanoes are mostly found in groups called shield fields.

There are about 1000 impact CRATERS on Venus, which is not much compared with what is found on Mars or the Moon. The largest crater, Mead, has a diameter of 270 km, and the smallest ones have diameters of about 1.5 km. The absence of small craters and the fact that craters on Venus seem to come in groups are probably due to the presence of a dense atmosphere around the planet that has broken the large meteroids and completely burnt

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

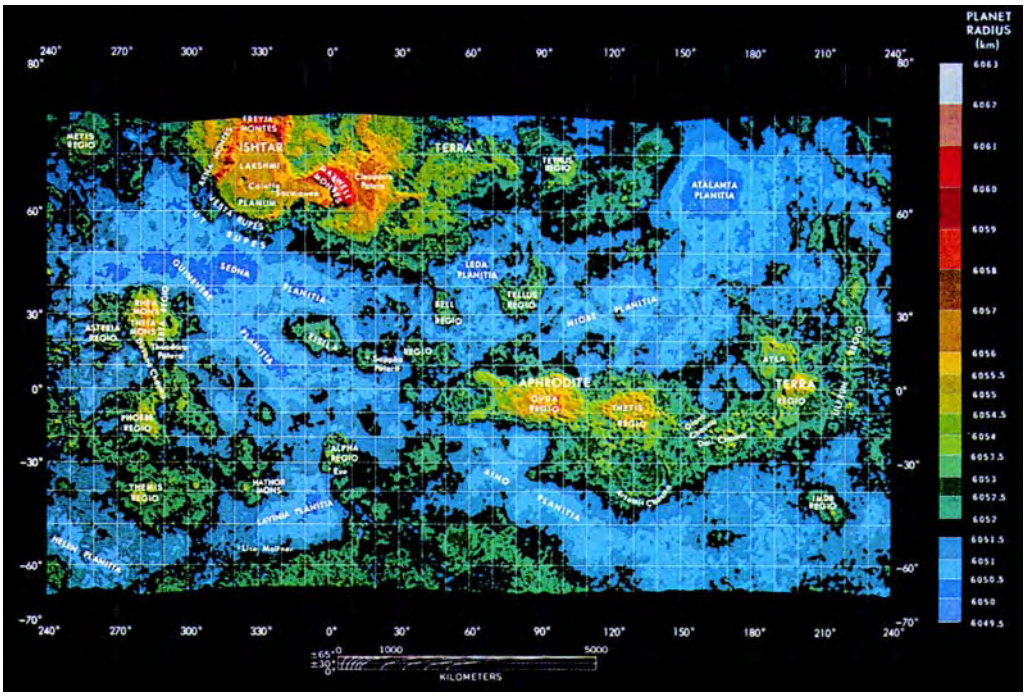


Figure 4. Topography of Venus in a Mercator projection (from Hunten D M, Colin L, Donahue T M and Moroz V I (ed) 1983 *Venus* (Tucson, AZ: University of Arizona Press) p 1059).

smaller ones. Higher crater densities are usually found in the plains. Most craters are pristine. The low number of craters, their distribution and their shapes suggest that the surface of Venus cannot be older than 500–800 million years on average.

Information on the surface composition comes from *in situ* data and from radar data. The geochemical data from the Venera 9, 10, 13 and Vega 1 and 2 landers imply a predominantly basaltic composition, with one site indicating possibly more alkalinic rocks. Venera 8 found a more granitic composition for the soil. Radar backscatter is a function of both roughness and the bulk properties of the target material. The radar properties of the lowland plains indicate surface dielectric constants of 3–8, which corresponds to low values of the dielectric constants of basalts on Earth. Much higher dielectric constants have been found for some regions at high elevation. They can be explained by highly conducting materials such as iron-bearing minerals (pyrite (FeS_2) or iron oxides such as magnetite (Fe_3O_4)) but identification of a specific phase which would exist only in a narrow band of elevation is problematic.

Volcanism and tectonism

Volcanic deposits and volcanic centers are very important characteristics of the surface of Venus. The processes of volcanism have been largely preserved on this planet, which is not the case for Mars and the Earth (see VOLCANISM IN THE SOLAR SYSTEM).

The plains are apparently completely covered by volcanic deposits. Concerning the volcanic centers, one can distinguish two main categories. The first category is characterized by very significant flows that generally come from a vent region during repetitive eruptions. It includes large and intermediate-size volcanic edifices, flow fields and calderas. To explain the formation of centers of repetitive volcanic eruption it is necessary to assume the presence of stable shallow magma reservoirs. The global eruption rates corresponding to the population of large volcanoes have been estimated by L S Crumpler and colleagues to be about 1.7×10^{-2} km^3 yr^{-1} of magma.

The second type concerns regions of the order of 100–200 km in diameter characterized by the presence of many small edifices and by relatively limited associated volumes of lava flows. It includes the groups of small-size edifices that are called shield fields. Some calculations indicate that these shield fields form where magma replenishment rates are about 10^{-4} km^3 yr^{-1} or less, but there is no estimate of the global rate for this category.

Concerning Venus's resurfacing history, it looks as if a substantial decrease in area resurfacing rates from about 4 km^2 yr^{-1} during plain emplacement to about 0.5 km^{2}yr^{-1} over the past 290 million years has occurred. This reduction has probably been accompanied by a change in volcanic style from flood-type volcanism in the plains to more localized development of volcanoes, coronae and rifts.

Concerning the types of lava, some indication comes from the existence and morphology of some of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

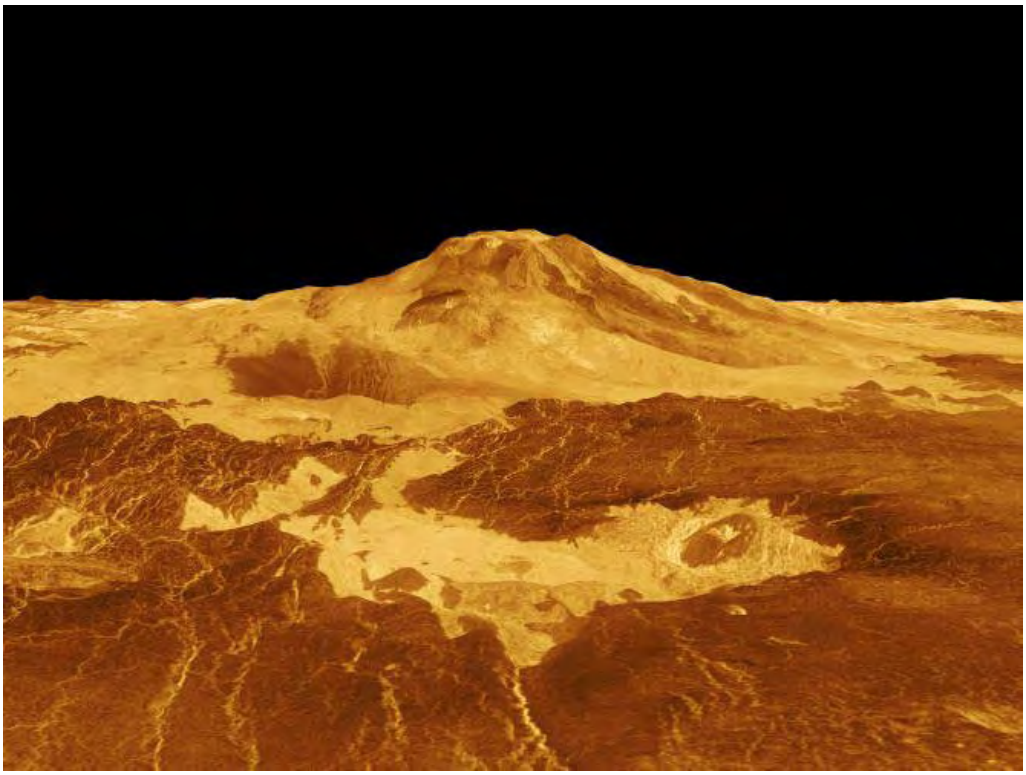


Figure 5. Magellan image showing a perspective view of Maat Mons, with a vertical exaggeration factor of 10. Maat Mons is a 5 km (or 8 km above the mean planetary radius) high volcano located in Atla Regio, at the eastern end of Aphrodite Terra (image P-40175 from the NASA/JPL Magellan Radar Mapping Mission). **This figure is reproduced as Color Plate 45.**

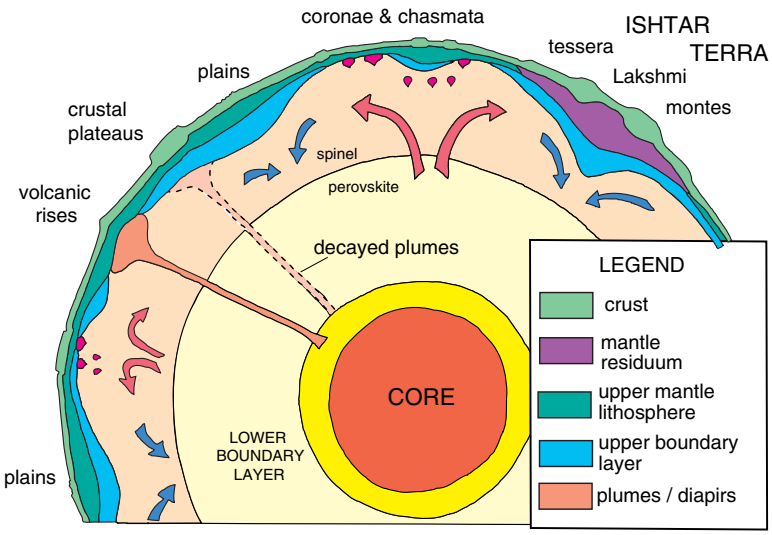


Figure 6. A possible global tectonic model for Venus (from Hansen V L *et al* 1997 *Venus II* ed S W Bougher, D M Hunten and R J Phillips (Tucson, AZ: University of Arizona Press) p 835.) **This figure is reproduced as Color Plate 46.**

channels observed on Venus. The longest of the various classes of channels observed on Venus are called canali. They have remarkably constant width along very long paths, exceeding 500 km and up to 6800 km (Baltis Vallis). They are 1–3 km wide, and their depth seems to be less than 50 m. Highly fluid lavas, that have erupted during sustained, high discharges, seem best to explain many of the channel features, particularly the canali.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Possible candidates for this fluid are native sulfur or, preferably, alkali–carbonatite lavas (very rare on Earth) which would have resulted from the melting of the crust altered by interaction with the CO_2–SO_2–halogen-rich atmosphere. In contrast, the pancake domes and festoon flows correspond to high-viscosity lavas.

Hundreds of coronae are distributed across the surface, most with diameters of less than 300 km, but a few exceed 1000 km in diameter. They are not seen on any other terrestrial planetary body. Many of them are associated with rift zones, forming linear chains thousands of kilometers long. In other places, they occur individually or in clusters in the plains. All coronae have some association with volcanism. Coronae may represent collapsed domes over large magma chambers.

There are numerous other tectonic surface features on Venus that have been formed by compression or by extension: ridges, valleys, mountain belts, ... In tesserae, one finds ridges and grooves that intersect and form very chaotic patterns. They probably reflect a complex deformation history. One also finds rift zones on dome-like features that are linear depressions that probably formed where Venus's lithosphere has ruptured owing to horizontal extension. The most important one is Beta Regio, which presents some analogy with the terrestrial East African Rift.

Geological evolution and the interior of Venus

Studies of impact craters on Venus indicate that about 80% of the history of Venus is not accessible to us. The surface we see is very young. The oldest terrain on Venus is tesserae. They are probably due to intensive tectonic deformation. Other apparently younger tectonic features must have formed during successive episodes of compressional and tensional deformations: fractures, broad ridges, wrinkle ridges. A period during which several stages of extensive volcanism occurred, burying areas of tesserae and forming the plains we see now, probably occurred afterwards.

Similarities in mean density and moment of inertia factor between Earth and Venus could indicate similar bulk compositions and internal structures. Models suggest that the interior of Venus is indeed compositionally similar to the Earth, except for small differences in iron, sulfur and oxygen content. Venus has certainly differentiated. It must have a crust, a mantle and a core. Although the relatively large value of the mean density of Venus strongly suggests that the planet has an iron core, Venus lacks an intrinsic magnetic field that would be strong evidence that at least part of its core exists in the liquid state, as is the case for Earth. Most models predict the radii of the core and mantle to be of the order of 3200 and 2800 km, respectively. Venus probably has a lithosphere, as does the Earth, but gravitational data indicate that it most likely does not have the low-viscosity astenosphere that exists on Earth below the lithosphere.

Most of the sites exhibit compositions similar to basalt, a volcanic rock that is very common on Earth and that constitutes most of the Earth's seafloor. The crust, therefore, probably formed from melting of an upper mantle with an Earth-like composition. Theoretical models constrained by the observed depths of impact craters and the spacings of ridges and rifts imply crustal thicknesses in many parts of Venus of 10–20 km, which is less than the average thickness of the Earth's continental crust (about 40 km) but greater than the thickness of the oceanic crust (about 6 km). Gravitational data indicate that the mean thickness of Venus's crust is 20–50 km. Alternatively, arguments based on mineralogy and high topographic elevations suggest crustal thicknesses in some areas of 100 km or more. Seismic probes are needed to obtain more information on the thickness of the crust.

The depth of the lithosphere is even more uncertain. Some think that Venus's lithosphere may be thicker than Earth's, owing to its lower water content in mantle rocks, thus making its surface geology somewhat less active. Others speak of a lithosphere only about 100 km thick, similar to or thinner than Earth's. They argue that the high temperature of the surface of Venus prevents a thick lithosphere from forming because the rocks are near their melting points.

Based on the identification and study of the different types of volcanic and tectonic features on Venus, and on gravitational data, it appears that there cannot be on Venus a global plate-tectonic activity as on Earth. There is instead evidence for active mantle convection with a recycling that occurs mainly in the vertical direction through regions of upwelling and downwelling. According to some models, volcanic centers must be associated with regions of mantle upwelling. The long-lived plumes (which are narrow regions of hot mantle upwelling) that allow sustained eruptions may be at the origin of the large volcanoes. Short-lived plumes may explain coronae. They may be due to local upwelling currents in Venus's mantle (figure 6).

One must also take into account the fact that the lithosphere may have been thin early in the history of Venus and has thickened with time. This would result in differences in the surface response above mantle upwellings. For instance, as has been suggested by some scientists, crustal plateaux could represent the interaction of ancient deep thermal plumes (or hotspots) with a thinner lithosphere. In contrast, volcanic rises may have formed only in the relatively recent past, after Venus's lithosphere had thickened enough to support them.

The gravitational data and the presence of ridge belts that indicate compressive stress suggest that the plains may correspond to regions of mantle downwelling on a broad scale. Regions of diffuse mantle upwelling would result instead in extension of the lithosphere and formation of coronae and chasmata (the chasmata are arrangements of troughs and scarps).

There is certainly a very complex time evolution. The regions of mantle upwellings and mantle downwellings may change location with time, which could result for instance in formation of coronae and chasmata in plain

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

regions. In contrast, some early-formed crustal plateaux may have collapsed owing to changes in temperature or to tectonics and could constitute some of the plains we see today.

Bibliography

Bougher S W, Hunten D M and Phillips R J (eds) 1997 *Venus II* (Tucson, AZ: University of Arizona Press)

Hunten D M, Colin L, Donahue T M and Moroz V I (eds) 1983 *Venus* (Tucson, AZ: University of Arizona Press)

C de Bergh

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Venus: Atmosphere

VENUS is the nearest planet to the Earth, both in terms of distance and in overall size and mass (table 1). However, Venus has lost most of its atmospheric and surface water, probably because it is significantly closer to the Sun. As a result of this, much of the carbon dioxide which on the Earth has been processed by the oceans to produce carbonates is still free in the atmosphere on Venus. This in turn means that Venus' atmosphere is very massive by terrestrial standards, with a surface pressure of almost 100 bar and a searingly high surface temperature in excess of 730 K.

The rotational axis of Venus is nearly perpendicular to the ecliptic, and the orbit is nearly circular, so seasonal changes in the climate are probably very small. The Venusian year is 224.7 Earth days, longer than the time for Venus to rotate on its axis, which is 243 Earth days. The solar day, defined as the time for the Sun to go from noon to noon as seen from the surface of Venus, is about 117 Earth days. This very slow rotation of the solid body of Venus is retrograde, i.e. backwards compared with the other planets, a curious state of affairs which is difficult to explain in terms of currently accepted models of the evolution of the SOLAR SYSTEM. The resulting absence of a strong Coriolis force at the surface is likely to be a major factor in determining the structure and dynamics of the atmosphere. Persistent high winds of the order of 100 m s^{-1} are observed near the cloud tops, 50 or 60 km above the surface, where the density of the air is similar to that near the ground on the Earth. It is not known reliably how these are produced.

The surface of Venus is obscured at visible wavelengths by planet-wide cloud cover, the upper layers at least consisting of sulfate aerosol similar to the much thinner layers of volcanic origin found in the terrestrial stratosphere (see VENUS: SURFACE). On Venus, the clouds have a complex layered structure and, although never completely absent, their coverage is very variable. They play an even larger role in the energy balance of the planet than clouds do on Earth, through their contribution to the atmospheric 'GREENHOUSE EFFECT'.

Localized dynamical or 'weather' activity on Venus, as far as is known at present, is dominated by four major phenomena: the cloud-top zonal super-rotation, the ultraviolet markings and associated planetary waves, cumulus dynamics in the deeper layers and the double-vortex structures at the poles. These and other types of activity, such as atmospheric tides, are no doubt linked to each other and to the general circulation, but all remain poorly understood.

Energy balance

At a distance of 108.2 million km, Venus is closer to the Sun than the Earth by a factor of about $2^{\frac{1}{2}}$, and so has about twice the incidence of solar energy. It is also much hotter at the surface, nearly $2\frac{1}{2}$; times more than the terrestrial mean of about 300 K. These facts are not simple to reconcile, however, because the ubiquitous and highly reflective cloud cover on Venus reflects 76% of the incoming solar flux and this results in a *smaller* net solar constant for Venus than for Earth. The high surface temperature must, therefore, be due to 'greenhouse' warming produced by the thick, cloudy atmosphere, possibly augmented by a contribution from the internal heat of the planet.

Table 1. Physical constants for the terrestrial planets.

	Venus	Earth	Mars
Orbital and rotational data			
Mean distance from Sun (km)	1.082×10^8	1.496×10^8	2.279×10^8
Eccentricity	0.0068	0.0167	0.0934
Obliquity (deg)	177	23.45	23.98
Sidereal period (days)	224.701	365.256	686.980
Rotational period (h)	5832.24	23.9345	24.6229
Solar day (days)	117	1	1.0287
Solar constant (kW m^{-2})	2.62	1.38	0.594
Net heat input (kW m^{-2})	0.367	0.842	0.499
Solid body data			
Mass (kg)	4.870×10^{24}	5.976×10^{24}	6.421×10^{23}
Radius (km)	6051.5	6378–6357	3398
Surface gravity (m s^{-2})	8.60	9.78	3.72
Atmospheric data			
Composition	See table 2		
Mean molecular weight	43.44	28.98 (dry)	43.49
Mean surface temperature (K)	730	288	220
Mean surface pressure (N m^{-2})	92	1	0.007
Mass (kg)	4.77×10^{20}	5.30×10^{18}	$\sim 10^{16}$

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

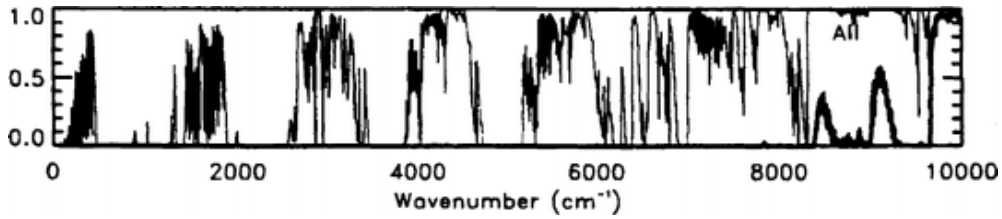


Figure 1. The computed transmission of Venus' atmosphere from the surface to space, showing the spectral gaps to be plugged by cloud opacity in order to maintain the observed 'greenhouse' effect (V Meadows and D Crisp).

It is not simple to prove that the observed atmospheric conditions can in fact generate such a large 'greenhouse' effect. The problem is that the massive amounts of carbon dioxide are very effective at blocking the emission of thermal infrared radiation, but only at those wavelengths where the gas has absorption bands, which are far from covering the entire spectrum. Moderate amounts of water vapor are also required, and even then considerable spectral gaps or 'windows' remain (figure 1). These could be blocked by the clouds, since liquid or solid absorbers present some opacity at every wavelength, the details depending on composition and particle size. The difficulty for early theorists was that using clouds to 'close' the greenhouse also tended to block the incoming sunlight, so that the calculated equilibrium temperature of the surface remained well below that observed.

This problem began to be resolved when it was realized that the clouds are made of sulfuric acid droplets, at least in the higher, most easily measured layers. These have the property of being highly absorbing at thermal infrared wavelengths, while being nearly conservative scatterers in the visible and near-infrared. Thus, the clouds tend to diffuse downwards those of the incoming solar photons that they do not reflect to space, while blocking thermal emission from the lower atmosphere and surface. This explains the result, surprising at the time, that the VENERA landers in the 1970s were able to photograph the surface in natural light. It also means that radiative transfer models, involving weak as well as strong bands of CO_2 and H_2O, plus those of the minor constituents CO, HCl and SO_2, can account for the high surface temperatures by careful incorporation of the scattering and absorbing properties of the clouds.

The total solar energy diffusing through the cloud cover on Venus corresponds to about 17 W cm^{-2} of surface insolation on the average, about 12% of the total absorbed by the planet and the atmosphere. The high opacity of the gaseous atmosphere and cloud at longer wavelengths requires the surface to reach temperatures high enough to melt zinc before the upwelling flex is intense enough, and at shorter wavelengths, so that equilibrium is attained. An airless body with the same ALBEDO and at the same distance from the Sun as Venus would reach equilibrium for a mean surface temperature of only about 230 K. This 500 K greenhouse enhancement of the surface temperature compares with only about 30 K on Earth and 10 K on Mars.

Composition

The primordial atmosphere of Venus which originally formed with the solid body, like those of the other TERRESTRIAL PLANETS, was likely to have been lost in the distant past as the young Sun went through phases of high activity. The present atmosphere would have been produced much later by outgassing from the crust, a process which we are probably observing today as active volcanism, and by the influx of cometary and meteoritic material, which is also still going on. The relative contributions of these distinct sources can, to some extent, be deduced from the data which are gradually being accrued on the composition, and in particular the isotopic ratios, in the contemporary terrestrial planet atmospheres and in comets and meteorites.

COMETS are a rich source of volatile compounds such as carbon dioxide, water vapor, methane and ammonia. If the last of these was the source of the nitrogen now present, and we allow for processes such as the production of argon by the decay of radioactive potassium in the crust, the contemporary atmosphere could all be of external origin. On the other hand, the high abundance of sulfur in Venus' clouds is strongly suggestive of extensive volcanic activity and volcanoes are also prolific sources of carbon dioxide, nitrogen and the other gases required to explain present-day Venus.

Apart from carbon dioxide and water vapor, Venus' atmosphere consists primarily of inert gases, particularly nitrogen and argon (table 2). The amount of water present as gas and bound up with sulfuric acid and other compounds in the clouds is between 10 and 100 000 times less than exists in the oceans and atmosphere of the Earth (see EARTH'S ATMOSPHERE). Thus Venus is overall very dry compared with the Earth while, at the same time, deuterium is about 100 times more abundant on Venus than Earth. This suggests that Venus may have had much more water initially, but that most of it has been lost. Loss takes place by dissociation of the water in the upper atmosphere by solar ultraviolet radiation and the subsequent escape of the hydrogen. Both deuterium and normal hydrogen escape from the atmosphere while there is free water on the surface, but the heavier isotope

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

escapes less efficiently, leading to the observed fractionation.

The loss rate of the water depends strongly on its abundance in the relatively cool middle atmosphere as well as the intensity of the solar ultraviolet flux. Models of the process suggest that Venus could have lost an ocean of present-day terrestrial proportions in only a few hundred million years. The oxygen produced at the same time is too massive to escape at any significant rate, according to Jeans' formula, and must remain on the planet, presumably most of it bound chemically within the crust, mainly as carbonate rocks. As on the Earth, this process would remove atmospheric carbon dioxide efficiently as long as liquid water was available. Once the free water was all used up, the mixing ratio of water vapor in the upper atmosphere would fall sharply and the loss rates of both forms of hydrogen, and the take-up of oxygen into minerals, have also declined to the present relatively low levels.

In the present-day atmosphere of Venus, chemical reactions coupled with the transport and radiative processes regulate the abundances of the most important minor constituents. The most important are the cycles involving water vapor, sulfuric acid and their products, which maintain the cloud layers, and which probably also involve reactions between the atmosphere and the surface. However, there are currently very few hard data on the abundances of reactive species near the surface, and still less on the composition and mineralogy of the surface itself. It seems likely, however, that the surface is a net sink for sulfur compounds, the combination of sulfur dioxide with calcite being an example of likely importance in this regard. Similar considerations must apply to interactions between halogen compounds and the surface, since relatively large (compared with Earth) amounts of hydrogen chloride and hydrogen fluoride are present above and below the clouds. Halogen chemistry may also have an important role in the formation of the clouds themselves.

Another important chemical cycle is that which gives rise to the observed distribution of carbon monoxide. CO is very abundant (mixing ratios of the order of a few parts per thousand by volume) in the upper atmosphere of Venus, as would be expected from the action of solar ultraviolet radiation on carbon dioxide. It is strongly depleted in the cloud layers (<1 ppm), again not too surprisingly, since it is involved in reactions with SO_2 and the other species which make up the sulfur cycle. Below the clouds, and near the surface, however, the carbon monoxide values recovers to around 30 ppm, and shows a marked equator-to-pole gradient. It seems likely that CO is transported rapidly down from the thermosphere in the polar vortices (see below) to the troposphere where it is gradually removed by reactions in the hot lower atmosphere and at the surface.

Table 2. Compositions of the terrestrial planet atmospheres.

	Venus	Earth	Mars
Carbon dioxide	0.96	0.0003	0.95
Nitrogen	0.035	0.770	0.027
Argon	0.00007	0.0093	0.016
Water vapor	0.0001(?)	0.01	0.0003
Oxygen	0.0013	0.21	0
Sulfur dioxide	0.00015	0.2 ppb	
Carbon monoxide	0.00004	0.12 ppm	0.0007
Neon	5 ppm	18 ppm	2.5 ppm
Hydrogen chloride	0.5 ppm	3 ppb	<0.1 ppm
Hydrogen fluoride	5 ppm	1 ppb	

Values are given as fractional abundances except where parts per million (ppm) or parts per billion (ppb) are stated.

Thermal structure

The solar radiation which penetrates the clouds warms the lower atmosphere, which is prevented by the opacity of the overlying layers from cooling by radiation to space. It therefore forms a deep convective region, the troposphere (figure 2). This links the high surface temperature of around 730 K, produced by the 'greenhouse' effect as described above, with the level at which the temperature is close to the effective bolometric temperature of Venus (about 230 K), where strong radiative cooling to space can occur. The adiabatic lapse rate, which applies when the vertical gradient is controlled by convection, is $-g/c_p$ (g being the acceleration due to gravity and c_p the specific heat at constant pressure) or about 10 K km^{-1} for Venus, so that the troposphere is around 50 km in vertical extent, much deeper than on Earth or Mars.

Above the troposphere the temperature tends to be constant with height, because the atmosphere here is optically thin and, to a first approximation, each layer tends to find the same equilibrium temperature. The temperature is determined by the balance between the absorption of upwelling infrared from the surface and troposphere and cooling to space, with no significant absorption of direct solar energy taking place.

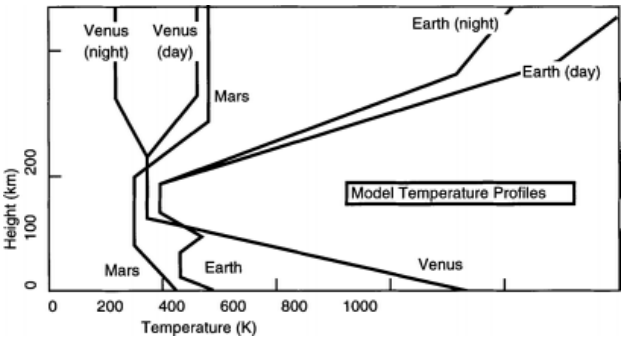


Figure 2. Mean model temperature profiles for the atmospheres of the terrestrial planets.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

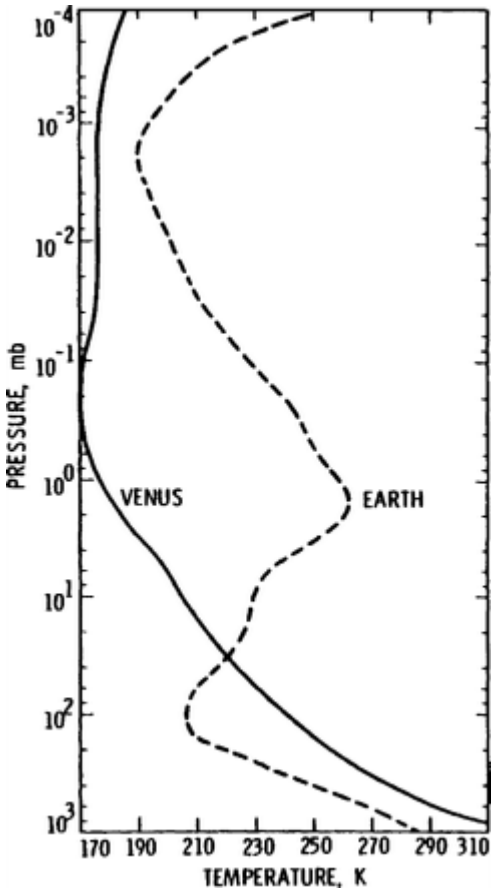


Figure 3. Measured details of the temperature profiles on Venus and Earth, on a common pressure scale.

On the Earth, this situation is altered by the ozone layer, which is responsible for substantial heating which divides the middle atmosphere into the stratosphere, where temperature rises with height, and the mesosphere, where it falls. There is no corresponding effect on Venus (figure 3), except for small amounts of absorption of solar and thermal energy in the near-infrared bands of water vapor and carbon dioxide. The two profiles are quite similar near 1 bar and near 10^{-4} mbar, away from the ozone heating region on the Earth which peaks near 1 mbar.

The tendency for temperature to increase at the top of the range covered by figure 3 marks the base of the thermosphere, which on Venus begins at a height of around 120 km above the surface. Here the atmosphere is very rarefied, and short-wavelength solar radiation and energetic particles drive rapid photochemical reactions and high temperatures by day. When the heating is absent, efficient radiative cooling by CO_2 results in the rapid decline of temperature, so there is a sharp gradient across the terminator, from over 300 K on the illuminated hemisphere down to 100 K or less on the night side.

Data on the planet-wide variability of temperature below the clouds are sparse and are still mainly limited to the small number of direct measurements made by the various entry probes. Remote sensing of the lower atmosphere has recently been shown to be possible, in particular at near-infrared wavelengths, as well as in the microwave, but uncertainties in the distribution of opacity sources in both spectral regions make these difficult to interpret in terms of temperature variations. Such data as exist show that the temperature gradients in the lower atmosphere are close to adiabatic in the vertical, and close to zero in the horizontal, as would be expected theoretically from the high opacity and high density. (However, there must be temperature structure associated with the dynamical activity seen in the deep atmosphere cloud structure—cf figure 5—although this has yet to be measured or even estimated.)

Above the clouds, where the density is lower, more temperature variability is expected, and this has been observed as gradients of the order of several tens of Kelvins on constant-pressure surfaces planet wide. Figure 4 shows the time-averaged (72-day mean) global temperature field from Pioneer Orbiter infrared soundings, in which several features clearly related to the general circulation stand out. Note the 'polar warming', in which the equator-to-pole temperature gradient leads to higher temperatures over pole than equator, in spite of the fact that the trend in radiative heating is in the opposite direction. The 'polar collar' feature stands out in the meridional cross-section; this is an intense ribbon of cold air surrounding the pole at about 65º latitude. These are both features of dynamical origin, related to the zonal super-rotation and the polar vortex respectively. The meridional average temperature field is dominated by the diurnal variation or *solar tide*, the temperature cycle which is induced by the apparent motion of the Sun overhead. This contains a whole spectrum of Fourier components, because the forcing is non-sinusoidal; the actual atmospheric response depends on the mean wind and the interference between the various components. The solar tide on Venus and classical atmospheric tidal theory as originally developed for the Earth (see TIDES) can be reconciled, provided that a realistic representation of the zonal wind is incorporated.

Clouds

Venus is completely enshrouded by clouds in a complex layered structure over 30 km deep; their properties are summarized in table 3. As already noted, the clouds are a key part of a highly interactive climate system (see CLOUDS IN PLANETARY ATMOSPHERES). Concentrated H_2SO_4 droplets have properties which contribute very significantly to the atmospheric 'greenhouse', by scattering conservatively at short (solar) wavelength radiation while strongly absorbing at long (planetary) wavelengths. Changes in the optical properties or depth of the cloud layers, for example if the clouds dissipated

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

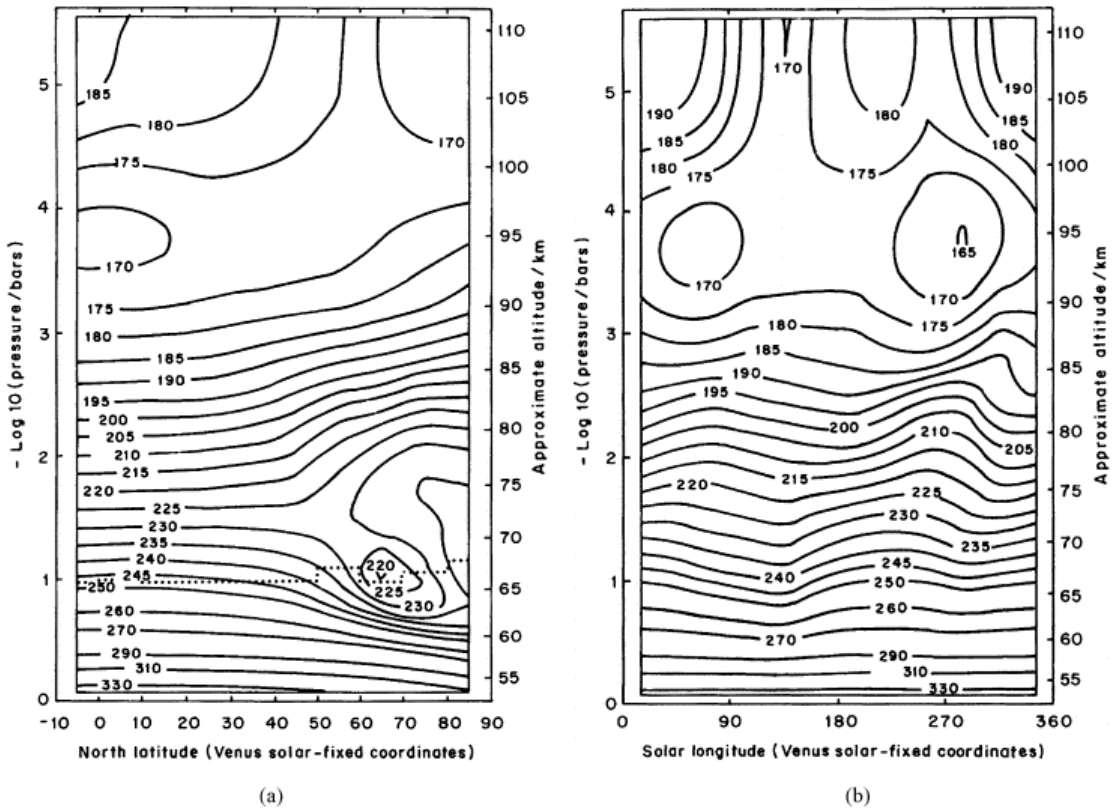


Figure 4. Time-averaged temperature fields in the middle atmosphere of Venus: (*a*) the zonal mean field and (*b*) the variations around a latitude belt from 0º to 30º N, both plotted against pressure and approximate height. The horizontal stepped line represents the retrieved mean cloud top height.

Table 3. Properties of clouds and dust in the terrestrial planet atmospheres.

	Venus	Earth	Mars
Fractional coverage	1.00	0.40	0.05 (cloud); 0–1.0 (dust)
Typical average depth	25–40	5–7	0.01–1.00; 0.2–6 (dust)
Composition	$H_2SO_4 \cdot H_2O$	H_2O	H_2O, CO_2, magnetite etc (dust)
Number density, liquid (cm^{-3})	50–300	100–1000	0
Number density, solid (cm^{-3})	10–50	0.1–50	30–1000 (near surface)
Typical mass loading ($g\ m^{-3}$)	0.01–0.1	0.1–10	0.0002–0.1
Main production process	Chemistry	Condensation	Condensation, windblown (dust)
Equivalent depth (mm)	0.1–0.2	0.03–0.05	1–100
Effective radius (μm)	0.2–30	10	0.4–2.5 (dust)
Main forms	Stratiform	Stratiform, cumulus	Stratiform, mixed (dust)
Temporal variability	Slight	High	High
Dominant heat exchange process	Radiation	Latent heat	Radiation

The equivalent length is the estimated thickness of the cloud material if it were deposited on the surface. The effective radius is the radius of the spherical particles having most nearly the same scattering properties as the cloud at visible wavelengths.

or changed their composition, owing to a reduction in the supply of SO_2 and other source gases from volcanoes, or to an instability in the dynamical regime, would cause the lower atmosphere and surface gradually to cool down (or, conceivably, to become even hotter). Changes in temperature would be likely to further modify the atmospheric circulation and the formation of clouds, producing feedback which seems more likely to be positive (accelerating the change) than negative. At present the radiative, dynamical and chemical processes appear to be in balance, but the stability of the currently observed state may be precarious.

Most of our detailed knowledge of the cloud properties comes from optical measurements: polarimetry

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

as a function of phase angle from the Earth, nephelometry and particle size measurements from entry probes, and visible, ultraviolet and infrared remote sensing from orbiter and fly-by spacecraft. It has recently been discovered that near-infrared spectroscopic measurements in atmospheric 'windows', that is wavelength regions where the main atmospheric gases are weakly absorbing, penetrate the clouds, in some windows all the way to the surface. This type of observation reveals the cloud morphology all over the planet and shows that the integrated vertical opacity is very variable, by factors of 20 or more. Thick and relatively thin clouds form patterns suggestive of large-scale cumulus dynamics, presumably with the cloud material actively condensing and dissipating in rising and falling air associated with weather systems (figure 5), although the details are lacking because of a shortage of high-resolution data in space and time.

At the PIONEER VENUS Large Probe entry site, the main cloud deck extends from about 47 to about 67 km above the surface, declining gradually at the upper boundary with a scale height of a few km, and with a thin haze layer some 13 km deep below. Within this vertical structure, detailed, and presumably variable, layering occurs and particles of different sizes congregate at different height levels. The particles range in diameter from less than 1 to over 30 μm and tend to a trimodal size distribution, with the commonest diameters falling towards the ends of the overall range and in the 2–3 μm region. It is these intermediate size or 'mode 2' droplets which are visible from outside Venus and for which spectroscopic, polarimetric and other evidence yields a composition of 75% H_2SO_4 and 25% H_2O. The composition of the smaller drops, which form an aerosol haze extending throughout the cloud layer, is unknown. Most of the mass of the clouds is in the biggest drops, for which there is some inconclusive evidence of a non-spherical shape, implying a solid composition, perhaps crystalline sulfur. The formation of the cloud droplets can be explained by a model in which H_2O and SO_2 (the latter possibly of volcanic origin) combine photochemically near the cloud-top level. It is more difficult to explain the size distribution, particularly the existence of more than one mode. Compositional contrasts and dynamical effects may be at work but once again the observations which would elucidate these are lacking.

There is some evidence in infrared maps from the Pioneer Venus orbiter of cold, high cloud, presumably condensed CO_2, near the tropopause above the dawn terminator, where the atmosphere is coldest as it comes to the end of the long (>50 h at this altitude) Venus night.

Figure 5. 2.7 μm image of Venus from Galileo–NIMS.

Dynamics

As a result of the slow rotation of the planet, its near-circular orbit and small obliquity, the underlying circulation of Venus' atmosphere is quite simple. The Sun is always above the equator, to within a couple of degrees, so the air warmed at low latitudes rises and moves towards the poles, where it cools and descends before returning equatorwards at lower altitudes.

On top of this simple picture are superimposed various complications. One that has been observed for many decades, but remains difficult to explain, is the global 'super-rotation', which manifests itself in cloud structure which moves rapidly around the planet in a direction parallel to the equator. The cloud markings, which appear with high contrast through an ultraviolet filter, have their origin at heights 50 or 60 km above the surface (where the pressure is of the order of 100 mbar) and travel around the equator in 4–5 days, corresponding to speeds near 100 m s^{-1}. This is more than 50 times faster than the rotation rate of the surface below. Measurements of the winds below the clouds, and calculations (from temperature data) of the winds above the cloud tops, show that the zonal wind speed declines at higher and lower levels, reaching values near zero at about 100 km and near the surface respectively.

Direct measurements of the winds 1 m or so above the surface by the Russian landers Venera 9 and 10 found

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

velocities of ≤1 m s^{-1}. Tracking of the Pioneer and Venera landers during their descent showed that there is a steady decrease with height from the 100 m s^{-1} or so observed in the ultraviolet markings near the cloud tops. Earth-based observers had earlier shown, by the measurement of Doppler-shifted emission lines from atmospheric gases, that the cloud-tracked winds do, in fact, apply to mass motions, rather than the phase speed of waves as had also been suggested.

Attempts have been made to explain these high zonal wind speeds on Venus by several mechanisms, all of which fall into one of three main categories, i.e. (i) the gravitational interaction of the Sun with the atmospheric tides, (ii) the overhead motion of the Sun in the sky (the 'moving flame' mechanism) and (iii) the upward transport of momentum from the surface. Currently prevailing opinion favors a version of mechanism (iii), in which momentum from the solid planet is transported by waves whose interaction with the main flow is complex and in which the mean meridional circulation plays an important role. Parameterizations have been found which are able to produce large zonal velocities in dynamical models of the Venusian atmosphere, although of course this is not the same as saying that we understand the forcing or dissipation mechanisms responsible for the transfer of momentum from the surface to the cloud tops.

The cloud motions which trace the zonal winds also reveal the pattern of the meridional circulation on Venus. As expected on the simple theoretical grounds outlined above, Hadley cells exist in each hemisphere. These are global-scale circulation cells characterized by rising motion all around one constant latitude belt and descending motion at another. Each cell extends to higher latitudes than on Earth, in part a consequence of the slower zonal rotation speeds. Near the poles on Venus, a complex instability develops, resulting in dramatic long-lived wave structures. The *polar collar* takes the form of a band of very cold air, some 10 km deep and 1000 km in radius, centered on the pole. Inside the collar, temperatures are some 40 K cooler than outside the feature. Poleward of the inner edge of the collar lies the *polar dipole*, a wavenumber 2 feature consisting of two well-defined warm regions circulating around the pole. Both the dipole and the collar have so far resisted attempts to model them as normal modes of the atmosphere.

The cloud-tracked winds obtained from Pioneer Venus and Mariner 10 both show equator-to-pole velocities of around 5 m s^{-1} in each hemisphere. Tracking of the Pioneer Venus probes shows winds of this magnitude at about 50–60 km altitude, with a very complicated vertical structure (figure 6). One possible interpretation of the alternations in the direction, as well as the magnitude, of the meridional wind is that these could mark the passage of the probe through the different components of a stack of Hadley cells, each extending from the equator to high latitudes. The layered eddy sources and sinks which could drive the zonal super-rotation may be related to the cell interfaces.

Motions in the deeper atmosphere were observed by near-infrared imaging carried out by the Galileo probe in 1990. The features observed on the night side of the planet at wavelengths from 1 to 3.5 μm originate in the main cloud deck, illuminated from below by thermal emission from the hot lower atmosphere. The typical velocities inferred near the equator were about half as fast as those from UV markings, which is consistent with the vertical profiles of wind and cloud opacity measured by the Pioneer and Venera probes, since the cloud layer providing most of the opacity in the case of the near-IR markings is 10–15 km deeper than for the UV markings. Galileo winds feature a zonal jet of more than 100 m s^{-1} at middle latitudes and equator-to-pole drifts of a few m s^{-1}.

Conclusion

Venus presents an arresting picture. What we appear to see is an Earth-like planet whose atmosphere failed to

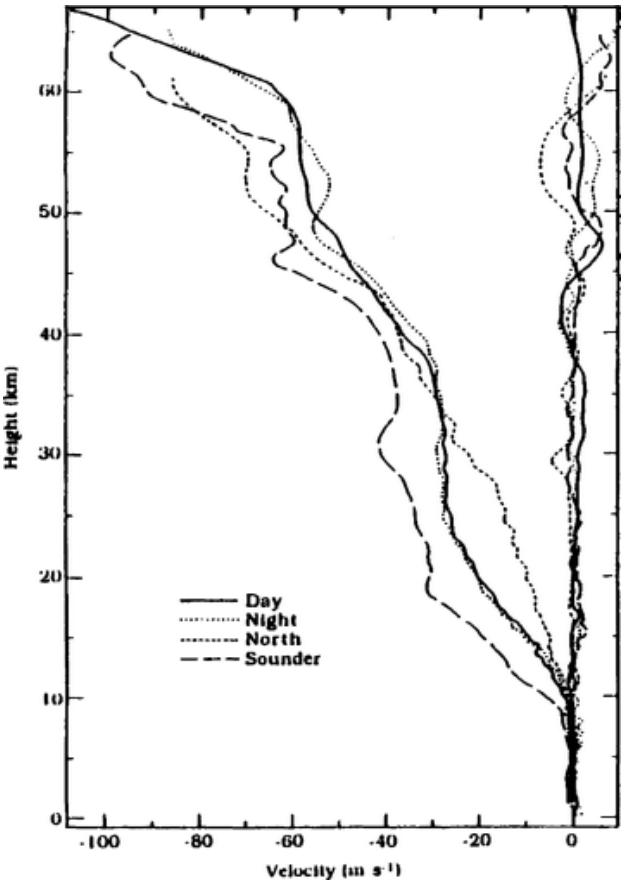


Figure 6. Profiles of the zonal (east-to-west) and meridional (north-to-south) wind on Venus as measured by tracking the descent of the Pioneer Venus probes.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

evolve in the same way, primarily because of the early loss of water, and which now supports a thick, primitive atmosphere in rapid, turbulent motion. Thick, sulfurous clouds, probably originating in volcanic activity on the surface and possibly requiring present-day volcanism for their maintenance, blanket the planet. The resulting 'greenhouse' effect drives the surface temperature to remarkably high levels.

These conditions make exploration difficult since remote sensing is inhibited (but not prevented altogether) by the clouds, while landings are arduous and short lived. Advanced techniques are under development that will allow sample return from Venus' surface and clouds and submarine-like 'aerobots' to cruise in the hot, thick lower-atmospheric regions. New remote-sensing techniques for the lower atmosphere will be applied from orbit.

The focus of this new generation of Venus atmospheric studies will be an understanding of the dynamical regimes present, the photochemistry of the clouds, the scale of volcanism on Venus and the question of whether or not our sister planet once had oceans, like the Earth. Above all, however, will be the need to understand the dynamical regimes present. Knowledge of even the principle of the origin and maintenance of the zonal super-rotation is lacking; the same is true of the spectacular giant vortices in the high-latitude regions, in particular the phenomena known as the polar collar and the polar dipole. The deep atmosphere shows huge weather systems whose very existence was unsuspected until about a decade ago and for which there exists no theoretical basis at all. How can such a simply rotating, nearby, Earth-like planet be in such an incomprehensible state?

Breaking news update (30 April 2002)

Scientists have captured the first x-ray view of Venus using NASA's CHANDRA X-RAY OBSERVATORY. The observations provide new information about the atmosphere of Venus. Venus in x-rays looks similar to Venus in visible light, but there are important differences. The optically visible Venus is due to the reflection of sunlight and, for the relative positions of Venus, Earth and Sun during these observations, shows a uniform half-crescent that is brightest toward the middle. The x-ray Venus is slightly less than a half-crescent and brighter on the limbs. The differences are due to the processes by which Venus shines in visible and x-ray light. The x-rays from Venus are produced by fluorescence, rather than reflection. Solar x-rays bombard the atmosphere of Venus, knock electrons out of the inner parts of the atoms, and excite the atoms to a higher energy level. The atoms almost immediately return to their lower energy state with the emission of a fluorescent x-ray. A similar process involving ultraviolet light produces the visible light from fluorescent lamps. For Venus, most of the fluorescent x-rays come from oxygen and carbon atoms between 120 and 140 km (74 to 87 mi) above the planet's surface. In contrast, the optical light is reflected from clouds at a height of 50 to 70 km (31 to 43 mi). As a result, Venus' Sun-lit hemisphere appears surrounded by an almost-transparent luminous shell in x-rays that appears brightest over the limb.

Bibliography

Two major surveys of Venus have been published in recent years. The reader is referred to the atmospheric chapters in these for a more detailed overview of current knowledge and problems and for a complete set of references to original work.

Bougher S W, Hunten D M and Phillips R J (eds) 1997 *Venus 2* (Tucson, AZ: University of Arizona Press)

Hunten D M, Donahue T M and Moroz V (eds) 1983 *Venus* (Tucson, AZ: University of Arizona Press)

F W Taylor

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Venus: Interaction with Solar Wind

The solar wind interaction with VENUS provides the archetypal interaction of a flowing magnetized PLASMA with a PLANETARY IONOSPHERE. Mars interacts with the solar wind in much the same way as does Venus, while the rotating plasma in the Saturnian magnetosphere is believed to interact similarly with its moon, Titan (see SATURN: MAGNETOSPHERE INTERACTION WITH TITAN). The interaction of the Jovian plasma with its moon Io is intermediate between the cometary and Venus interactions (see IO: PLASMA TORUS). From a more academic perspective, this interaction illustrates how a magnetized plasma interacts with an unmagnetized plasma, and how mass pickup (of planetary ions, in this case) affects a flowing plasma. Our understanding of the interaction has been built up over a series of flight missions: Mariner 2 and 5, Venera 2, 4, 6, 8–14 and Pioneer Venus. The latter mission remained in orbit for 14 yr investigating the solar wind interaction with Venus over a complete solar cycle.

Although Venus is famous for its sulfuric acid clouds and sulfur dioxide haze, its main atmospheric constituent is carbon dioxide of mass 44. Because Venus is similar to the Earth in mass and size (6051 km radius), we might expect *a priori* its atmospheric behavior to be similar. Paradoxically, while the lower atmosphere is much warmer than the terrestrial atmosphere owing to the greenhouse effect, the upper atmosphere of Venus, above 100 km, is much cooler than that of Earth because of the absence of heating by a magnetosphere at Venus (see VENUS: ATMOSPHERE). As a consequence the scale height of the neutral Venus atmosphere is quite small. In the cold (cryosphere) nightside it is only a few km. In general, the atmospheric density falls off according to the equation of hydrostatic equilibrium that balances the upward pressure gradient force with the downward force of gravity on that parcel of gas:

$$nmg = -\frac{\mathrm{d}nkT}{\mathrm{d}h}$$

where n, m, g, k, T and h are respectively the number of molecules per unit volume, their mass, the force of gravity, Boltzmann's constant, their temperature and the altitude of the point of interest.

For an isothermal atmosphere the density decreases as

$$n = n_0 \exp(-h/H_n)$$

where $H_n = kT/mg$. The isothermal description fails at the highest altitudes (above about 200 km) for some constituents, such as atomic oxygen, where nonthermal processes, notably dissociative recombination of O_2^+, increase the average temperature. The important nonthermal processes are both photochemical (e.g. involving neutralization of ionospheric ions and electrons produced by solar EUV radiation) and 'mechanical' (involving collisions of energetic particles from above with the ambient gas). Thus atomic oxygen at Venus has a cold component, with a scale height of a few tens of km, and a hot component that extends to high altitudes with a scale height of hundreds of km. This latter component plays a special role in the solar wind interaction, described below.

As on Earth the Venus upper atmosphere is partially ionized by solar ultraviolet radiation and energetic particles that enter from the surrounding space. The rate of ionization decreases rapidly with decreasing altitude at low altitudes where the ionizing radiation and particles are absorbed. It also decreases with increasing altitude at high altitudes where the ionizing radiation is constant but the ionizable neutral particles decrease with altitude. Thus there is a maximum ion production rate at some altitude h_m. The rate of ionization, Q, varies with altitude for a simple one-component isothermal ionosphere as

$$Q = Q_m \exp[1 - y - \exp(-y)]$$

where Q_m is the peak production rate and $y = (h–h_m)/H_n$ where H_n is the neutral scale height and h_m is the height of peak production. This function is sketched for a scale height of 25 km in figure 1, representative of oxygen in the warmer dayside upper atmosphere of Venus. If the ionosphere is in photochemical equilibrium in which the recombination of electrons and ions takes place without vertical transport, then the rate of recombination is proportional to the product of the electron and ion number densities which are equal in our simple ionosphere. Hence the electron number density is

$$N_e = N_{m0} \exp\{[1 - z - \sec\chi \exp(-z)]/2\}$$

where χ is the solar zenith angle and we have referenced the density to the maximum at the subsolar point, N_{m0}, and the normalized altitude, z, to its altitude, h_{m0}, and the scale height, H_n. This function is shown in the bottom panel of figure 1.

At Venus the high-altitude ionospheric electron temperature is about 5000 K. The peak of the dayside ionosphere is at an altitude of about 140 km and has a peak number density of just under 10^6 cm$_{-3}$. At 400 km the number density is about 20 000 cm$_{-3}$ under solar maximum conditions near the subsolar point. At this altitude the collision rate is quite low and the electrical conductivity quite high. This is important for the solar wind interaction because the solar wind is a magnetized plasma. When it flows against the Venus ionosphere it does not generally penetrate below about 400 km unless the solar wind conditions are disturbed. The ionospheric pressure at 400 km under solar maximum conditions is sufficient to stand off the typical dynamic pressure of the

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

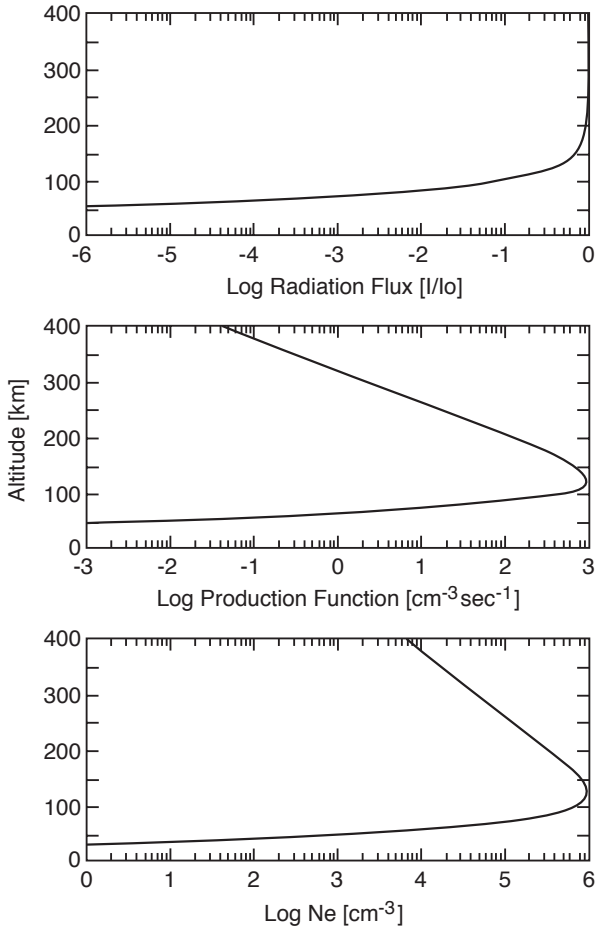


Figure 1. The formation of a Chapman-layer' ionosphere. Top: UV radiation from Sun drops in intensity as it is absorbed in photoionization process. Middle: Rate of electron production with altitude. As altitude decreases, the rate first increases owing to the increasing atmospheric density and then decreases when the UV radiation is exhausted. Bottom: The electron density consistent with the production if there is no vertical or horizontal transport.

solar wind. This is not expected to be the usual case at solar minimum when the ionospheric pressure is less.

The SOLAR WIND is a supersonic flow of electrons and ions with a number density near the orbit of Venus of about 15 particles per cm^3, a velocity of about 440 km s_{-1}, an ion temperature of about 10^5 K, an electron temperature of about 2–10^5 K and a magnetic field strength of about 10 nT. The velocity of a wave that could compress the solar wind flow and deflect it around Venus is about 100 km s^{-1} or about 1/4 of the speed of the solar wind. Thus, the deflection has to occur via the formation of a standing shock wave whose strength is measured by the ratio of the compressional wave velocity to the incoming flow velocity. This interaction is sketched in figure 2. The Venus bow shock has a shape similar to Earth's bow shock shape but, because the obstacle producing the shock is in this case the planetary ionosphere (and not an

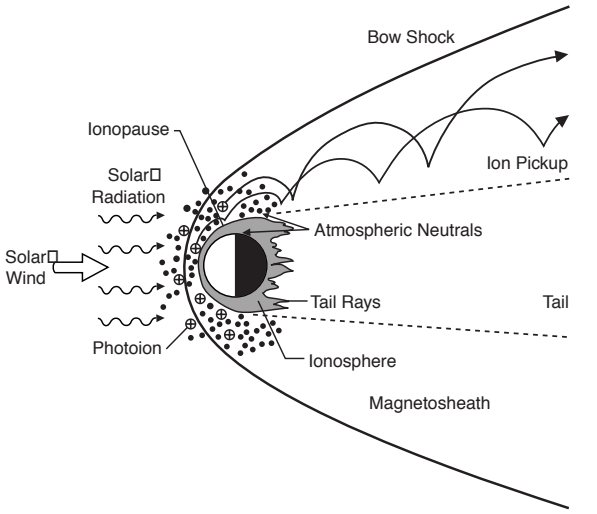


Figure 2. The solar wind interaction with the Venus ionosphere and neutral atmosphere. The solar UV ionizes the atmosphere and the ionosphere deflects the solar wind. Some neutral atoms are ionized in the solar wind and are picked up by the electric field.

internal magnetic field), the size of the bow shock is much smaller (see MAGNETOSPHERE OF EARTH: BOW SHOCK). At the nose of the Venus shock the Mach number or shock strength is close to 4 under typical conditions.

In the ideal case when none of the solar wind is absorbed by Venus and no ions are added to the flow by the atmosphere of Venus, the shock front, or bow shock as it is called, stands off from the obstacle at a distance sufficient to allow all the material that is compressed and deflected by the shock to flow around the obstacle. When the shock is strong, the density of the solar wind flow can increase a factor of 4 across the shock. At solar maximum when the pressure of the ionosphere is high, the shock front is about 2000 km above the surface of Venus at the subsolar point. If the shock weakens, the density jump becomes less and the shock moves away from the planet. In the limit as the Mach number approaches unity, the shock moves away to infinity. At solar minimum the typical ionosphere does not have sufficient pressure to prevent some direct interaction of the solar wind with the neutral atmosphere and the shock moves closer to Venus as some solar wind is absorbed by Venus.

The shocked plasma immediately downstream from the shock front is simply heated and compressed solar wind whose properties can be calculated by a set of magnetohydrodynamic equations called the Rankine–Hugoniot equations. However, the flow velocity perpendicular to the magnetic field direction drops drastically as the subsolar point on the obstacle is approached. The magnetic field of the solar wind plasma is compressed as the stagnation point is approached, and the field becomes draped around the obstacle. The thermal velocity of the particles is then able to evacuate the field lines

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

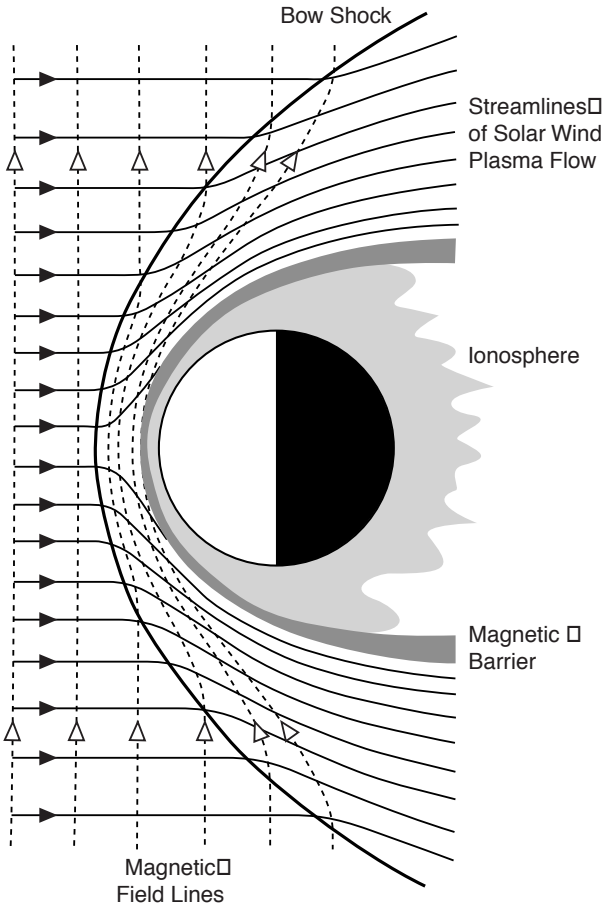


Figure 3. The solar wind flow and interplanetary magnetic field as it interacts with Venus, leading to the formation of a magnetic barrier.

so that the inner edge of the shocked plasma, or magnetosheath, becomes a magnetic barrier in which the transverse pressure is exerted almost totally by the magnetic field and not by the thermal plasma pressure. A similar situation occurs in what is called the 'depletion layer' just outside Earth's magnetosphere. As sketched in figure 3 the magnetic field wraps around Venus, maintaining a barrier between the flowing solar wind and the ionosphere and contributing to the formation of a magnetic tail in the antisolar direction. The plasma in the magnetosheath accelerates with distance behind Venus and the shock expands and weakens. Eventually far behind Venus the solar wind that interacted with Venus is indistinguishable from solar wind that did not interact, as long as Venus did not add anything to the solar wind flow.

Even though the bow shock marks the upper extent of the pressure wave that Venus launches to deflect the solar wind, there still are phenomena that can be seen in advance of the bow shock. These upstream phenomena consist of charged particles and waves. The particles are electrons and ions that either were reflected at the shock

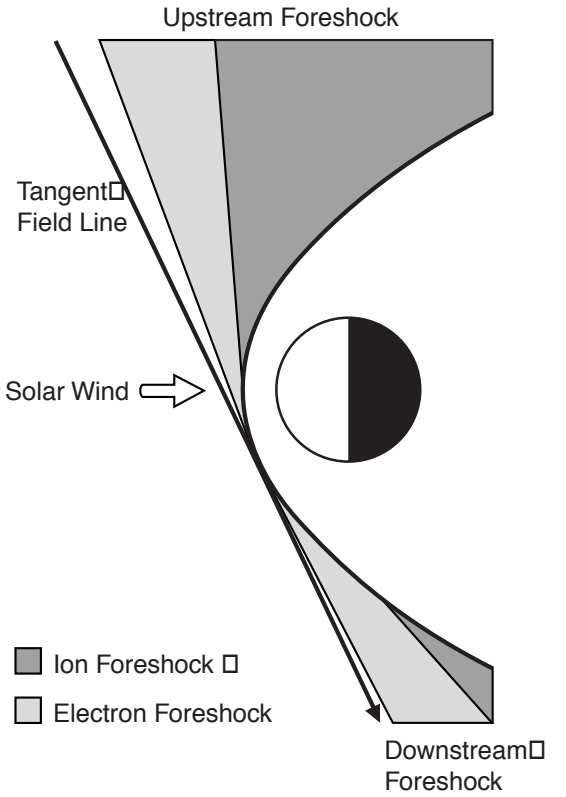


Figure 4. The geometry of the electron and ion foreshock at Venus. These regions mark the field lines that intersect the shock and along which electrons and ions can reach a spacecraft at the velocities to which they have been accelerated.

or leak from the hot population behind the shock and move upstream against the solar wind flow. The geometry of this region is sketched in figure 4. The waves can either be generated by the backstreaming particles (and generally be convected toward the shock) or be generated at the shock and move upstream. In many respects the wave and particle phenomena in front of the Venus bow shock are the same as those generated in front of the bow shocks of magnetized planets including Earth. The nature of the obstacle is not important here.

Returning to the obstacle for a moment, we note that the magnetic barrier has the effect of maintaining a cap on the ionosphere. This upper boundary is called the ionopause. Ionospheric plasma is not detected above the barrier because the atmospheric ions produced there are immediately removed by the interplanetary electric field (equal to $-V_{sw}B$), either swept away from Venus or deposited into the deeper atmosphere. Two such ions are sketched in figure 2. Nevertheless, most ions are produced within the ionosphere and remain in the ionosphere proper until they recombine. At comets the opposite is true. The recombination rate at high altitudes in a planetary ionosphere is small so that in steady state ions must flow to low altitudes to recombine. In the subsolar region the typical motion of

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

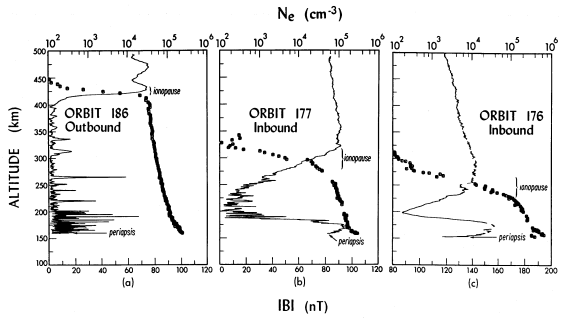


Figure 5. Vertical profiles of the electron density, N_e (symbols), and magnetic field strength, B, (full curve), on days of low (left), moderate (middle) and (high) solar wind pressure. Measurements by the Pioneer Venus orbiter on three early orbits when the spacecraft was entering the dayside ionosphere and periapsis, the closest approach to the planet, were at low altitudes.

the ionosphere is downward. There are few ions created at any altitude in the Venus shadow so that the nightside ionospheric pressure is much less than on the dayside. Away from the subsolar region the pressure gradient induces a supersonic expansion of the upper ionosphere into the nightside whenever the ionosphere is present at altitudes of 400 km and above. This flow produces a significant nightside ionosphere even though little ionosphere is produced there. At solar minimum or when the solar wind dynamic pressure is high the ionopause is too low to allow horizontal transport to the nightside ionosphere, which becomes very weak as a result.

Complications

The above description is a correct first-order model of the Venus interaction but there are details that need to be added to complete the picture. We look first at the variations in the ionosphere caused by varying solar wind pressure. We then examine the effects of planetary ion mass added to the solar wind flow and finally examine how the interaction varies with the solar cycle.

The effect of varying pressure

The boundary between the magnetic barrier and the ionosphere, the ionopause, is generally a thin current layer, at least when the solar wind pressure is low and the ionopause altitude is high (about 400 km). This thickness appears to be about one ionospheric thermal ion gyroradius. As mentioned above, the electrical conductivity is so high under these conditions that there is no diffusion of the draped solar wind magnetic field across the ionopause current layer. The left-hand panel of figure 5 shows such a situation illustrated with vertical profiles of the magnetic field and plasma density seen by the Pioneer Venus orbiter. Nevertheless, magnetic field can apparently get into the ionosphere in the form of flux bundles or ropes. These flux ropes cause the large spikes in field strength that are seen low in the ionosphere in the left-hand panel of figure 5. Observations suggest that they are untwisted at high altitudes but become twisted and even kinked as they sink. Figure 6 illustrates the for-

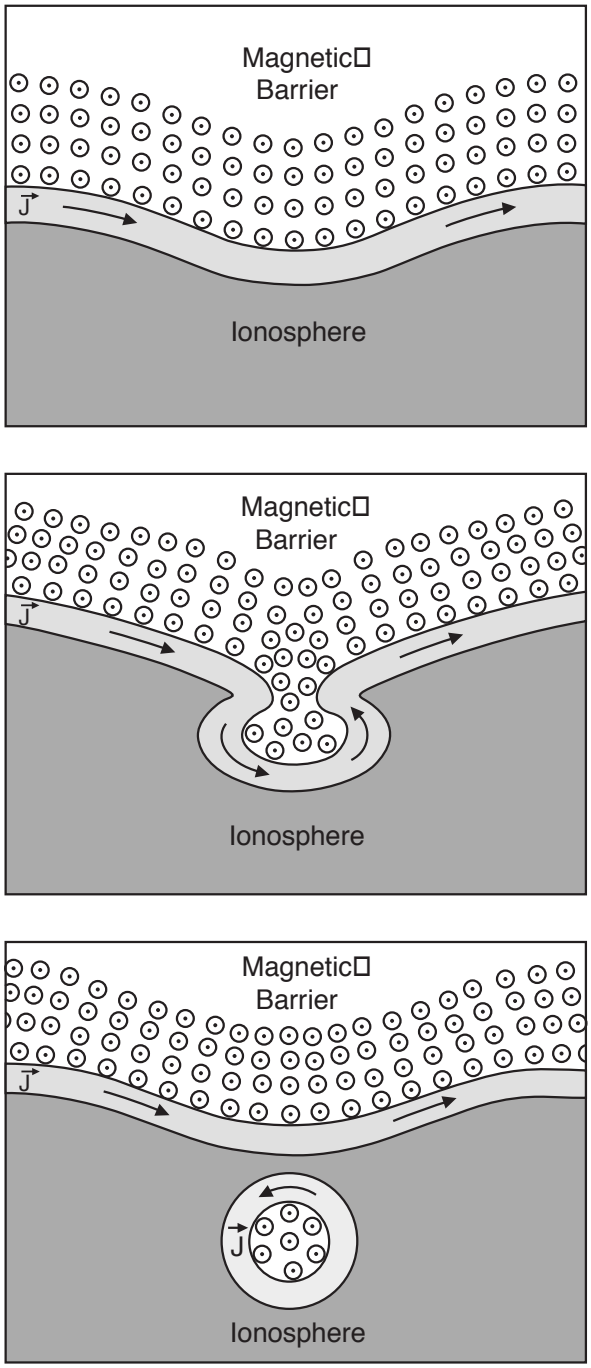


Figure 6. Formation of a flux rope in the Venus ionosphere due to the sinking of an initially buoyant flux tube from the magnetic barrier. In the middle panel the flux tube has become heavier owing to mass addition through photoionization of the neutral ionosphere and sinks. The magnetic force due to the curvature in the plane perpendicular to the page accelerates this motion.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

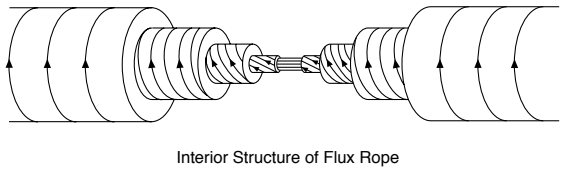


Figure 7. The magnetic structure of a flux rope. In the interior the magnetic field is strong and parallel to the axis of the rope. In the outer parts of the rope the magnetic field is weak and wraps around the axis.

mation of one of these ropes from an originally buoyant flux tube of the magnetic barrier. The twist in the ropes helps to maintain the coherence of the rope and to balance the outward pressure of the magnetic field. The twist is believed to arise as the rope sinks through a vertical shear in the horizontal velocity of the ionospheric plasma. Often the magnetic field of the ropes is almost in force balance by itself so pressure gradient forces in the plasma are not important. Such ropes are called force free. A sketch of the magnetic structure of such a flux rope is shown in figure 7.

Because the ionopause forms where pressure balance with the solar wind pressure occurs, its height varies with solar wind conditions. As illustrated by the rightmost two panels of figure 5, when the solar wind pressure increases to about 20 nPa and the altitude of the ionopause current layer falls to 250 km or less, the current layer thickens and magnetic field diffuses into the ionosphere because of ion–neutral and electron–neutral collisions. The downward ionospheric plasma velocity near the subsolar region carries the magnetized plasma to low altitude where it recombines and deposits the magnetic flux. Thus the collisional Venus ionosphere can become magnetized from the ionopause down to the peak of the ionosphere. The magnetic flux then collisionally diffuses through the current layer at the bottom of the ionosphere, creating a magnetic field between the ionosphere and the planet's core. The diffusion of magnetic flux below the dayside ionopause is balanced by the upward diffusion of magnetic field on the nightside. Since field lines in the ionosphere generally have their ends in the solar wind, these field lines on the nightside have ends that have moved far down the magnetotail. Hence the field lines are mainly vertical or radial on the nightside and close beneath the ionosphere. These regions of radial magnetic field on the nightside prevent the plasma from entering these regions so that they appear as 'holes' in the plasma measurements. The magnetic pressure in the holes is enhanced by the amount that the plasma pressure is diminished so that the plasma and field pressures are in equilibrium.

If the dayside ionosphere becomes fully magnetized, then the flow of plasma from the dayside to the nightside is cut off. Under those circumstances the nighttime plasma density drops to very low values and the phenomenon called a 'disappearing' ionosphere results. This same situation prevails at solar minimum when the dayside ionospheric pressure is lower than the typical solar wind pressure. The magnetic structure of the ionosphere under these low and high solar wind dynamic pressure conditions (relative to ionospheric plasma pressure) is shown in figure 8.

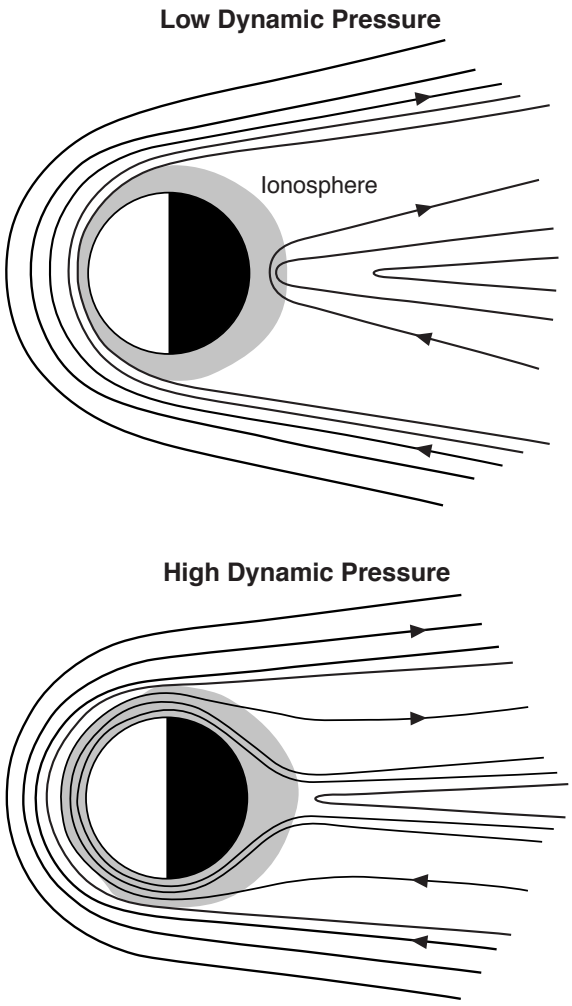


Figure 8. The magnetic configuration of the Venus ionosphere at times of low (top) and high (bottom) solar wind dynamic pressure.

Mass addition to the solar wind flow

The interaction of a comet with the solar wind is well understood (see SOLAR WIND: INTERACTION WITH COMETS). Gas sublimes from the comet nucleus and expands rapidly into the relative vacuum of space. This gas is ionized and swept up by the solar wind. The mass addition to the solar wind in the form of cometary ions slows it down by the conservation of momentum. The ends of the interplanetary field lines threading the cometary atmosphere are rooted in the undisturbed solar wind and thus carried past the comet at the solar wind speed.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

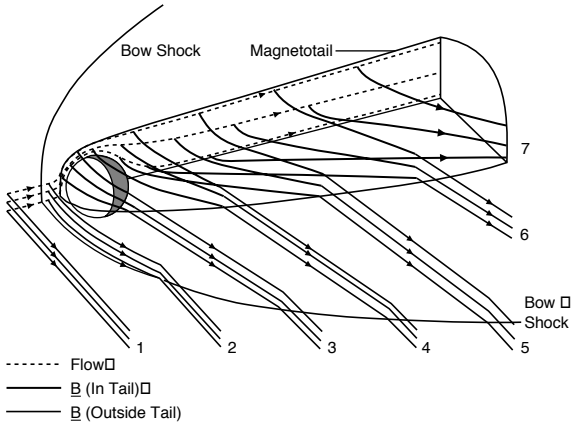


Figure 9. The formation of the Venus magnetotail. Flux tubes that pass closest to Venus are slowed the most, while their ends are carried along in the solar wind.

A long tail is formed downstream of the nucleus as a result.

The rate of addition of mass by Venus to the solar wind is much less than that of a comet because the planet's gravitational pull prevents the atmospheric gas from expanding rapidly and keeps it close to the planet. Nevertheless, as noted above, Venus does have a neutral oxygen exosphere that extends to at least 4000 km. Ions produced above the ionopause, by photoionization or collisions with the solar wind particles themselves, interact directly with the solar wind. The atomic oxygen exosphere is produced by the ionosphere, in particular by the dissociative recombination of O_2^+ ions at lower altitude, in which ions recombine with an electron to form two suprathermal oxygen atoms. These atoms have enough energy to allow them to reach altitudes above the ionopause where, if ionized, they will be carried away in the magnetosheath or solar wind flow. This additional ion mass can slow down the flow near the magnetic barrier where the ion production is highest owing to the greater neutral densities there. A larger region of stagnation could increase the size of the barrier and push the shock away from the obstacle. It could also add additional magnetized plasma to the ionosphere and enhance the magnetic tail. As illustrated in figure 9 the magnetic field lines approaching closest to Venus are slowed down in their motion at closest approach while their ends in the undisturbed solar wind are carried away at the solar wind velocity. This produces a long comet-like 'induced' tail.

The ions on these field lines are eventually accelerated up to the solar wind velocity. However, the composition of these ions always tags them as Venus ions. The Venus magnetic tail was regularly detected at the apogee of the PIONEER VENUS spacecraft (12 Venus radii) but the ions have been detected as far away as the orbit of the Earth.

The draped magnetic field will apply a stress to the plasma surrounding Venus but under steady-state conditions this stress should be rather uniform. When the magnetic field changes direction the uniformity of this stress changes and ridges and enhancements at the ionospheric plasma–solar wind boundary can form. When spacecraft pass through such a structure, the density enhancement may appear to be a 'cloud' of ionospheric plasma near the ionopause. These clouds stretch out from the nighttime ionosphere of Venus and are correlated with interplanetary magnetic field rotations. Other filamentary structures of ionospheric plasma extend out of the nightside upper ionosphere when solar wind conditions are steady. These 'tail rays' may represent a low-altitude component of planetary ions picked up by the solar wind from the denser cold upper atmosphere.

A final phenomenon that has been seen at Venus that appears to be associated with the solar wind interaction is the Venus aurora. On Earth AURORAS are often associated with parallel electric fields that accelerate electrons down into the atmosphere causing atoms to be excited and to emit photons. Weak, patchy auroras exist on the nightside of Venus also, but no explicit cause of these auroras has been determined.

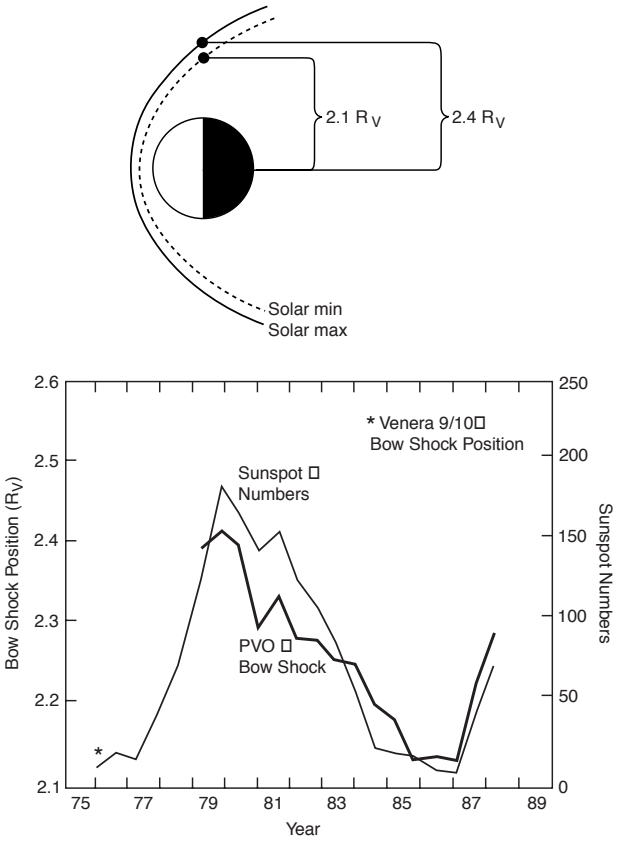


Figure 10. The motion of the bow shock in the course of the solar cycle. Top: An ecliptic plane view. Bottom: Position above the terminator and sunspot number through a solar cycle.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Solar cycle variation

The above discussion enables us to now address how Venus responds to the solar cycle changes on the Sun. The Earth's solar wind interaction is sensitive to the SOLAR CYCLE mainly because there are more coronal mass ejections launched from the Sun near solar maximum than solar minimum. The Venus interaction is also sensitive to these episodic, mainly magnetic, structures (albeit in very different ways than is Earth's magnetospheric interaction), but it is more generally sensitive to the changing ultraviolet flux during the solar cycle. At solar minimum the ionosphere is not as well developed, and so the same solar wind pressure results in an ionopause deeper in the atmosphere where the ionosphere is less highly electrically conducting. This routinely magnetizes the dayside ionosphere and drastically reduces the flow to the nightside ionosphere. The obstacle becomes a little smaller but more importantly some of the solar wind is probably absorbed. Thus the solar wind does not have to be deflected so far above the planet as at solar maximum. This effect is so marked that the shock approaches to within 1000 km of the planet of solar minimum as sketched in figure 10.

At solar maximum the dayside ionosphere is denser and better able to stand off the solar wind plasma above the ionosphere. In addition the neutral exosphere of hot oxygen is denser and more ions are created within the magnetosheath itself. This extra ionization further enhances the deflection of the solar wind.

Web Update (31 July 2002)

Scientists have captured the first x–ray view of Venus using NASA's CHANDRA X-RAY OBSERVATORY. The observations provide new information about the atmosphere of Venus. Venus in x–rays looks similar to Venus in visible light, but there are important differences. The optically visible Venus is due to the reflection of sunlight and, for the relative positions of Venus, Earth and Sun during these observations, shows a uniform half–crescent that is brightest toward the middle. The x–ray Venus is slightly less than a half–crescent and brighter on the limbs. The differences are due to the processes by which Venus shines in visible and x–ray light. The x–rays from Venus are produced by fluorescence, rather than reflection. Solar x–rays bombard the atmosphere of Venus, knock electrons out of the inner parts of the atoms, and excite the atoms to a higher energy level. The atoms almost immediately return to their lower energy state with the emission of a fluorescent x–ray. A similar process involving ultraviolet light produces the visible light from fluorescent lamps. For Venus, most of the fluorescent x–rays come from oxygen and carbon atoms between 120 and 140 km (74 to 87 mi) above the planet's surface. In contrast, the optical light is reflected from clouds at a height of 50 to 70 km (31 to 43 mi). As a result, Venus' Sun–lit hemisphere appears surrounded by an almost–transparent luminous shell in x–rays that appears brightest over the limb.

Bibliography

Bougher S W, Hunten D M and Phillips R J (ed) 1997 *Venus II* (Tucson, AZ: University of Arizona Press)

Hunten D M, Colin L, Donahue T M and Moroz V I (ed) 1983 *Venus* (Tucson, AZ: University of Arizona)

Luhmann J G, Tatrallyay M and Pepin R O (ed) 1992 *Venus and Mars: Atmospheres, Ionospheres and Solar Wind Interaction* (Washington, DC: American Geophysical Union)

Russell C T (ed) 1991 *Venus Aeronomy* (Dordrecht: Kluwer)

C T Russell and J G Luhmann

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Venus: Surface

Because of its thick and dense atmosphere, VENUS has only very recently revealed the aspect and the characteristics of its surface. In fact, the only means to penetrate the atmosphere and to see what is underneath is to observe the planet in wavelength regions of the electromagnetic spectra where the atmosphere is transparent such as the radar centimetric wavelengths or some near-infrared windows around 1 micron. An alternative solution is to send probes that directly land on the surface. These two means have been used to explore the surface of Venus and they give two complementary kinds of information. The first gives global and large-scale information about the surface, such as topography, surface roughness, altimetry, etc, while the second gives very local information such as surface images and/or composition. These two complementary techniques have greatly enhanced our knowledge of the surface of Venus and of its history.

The Venus coordinate system

The map coordinate system used for measuring longitude on Venus is different from that used on Earth. On Earth, longitude (an imaginary line stretching from pole to pole) is measured from a starting point (the prime meridian) at Greenwich, England (near London), toward the east and toward the west with increasing values in degrees until east meets west at the 180° point (the dateline), which is diametrically opposed to Greenwich. On Venus, longitude is measured from 0 to 360 degrees with the prime meridian centered within a small impact crater named Ariadne, located in Sedna Planitia. There is an arbitrary convention that determines the direction of increasing longitude on planetary bodies other than Earth: longitude shall be measured in a direction opposite to that in which the planet rotates. Because Venus rotates in a clockwise direction as viewed looking down on the north pole, longitude on Venus increases in numerical value toward the east from the planet's prime meridian.

Large-scale observations

Most of the large-scale and global observations of the Venusian surface have been obtained by radar. There are different kinds of radar data: (a) altimetric data, (b) reflectivity data and (c) emissivity data. The first gives the altitude of the topography, the second gives information about the roughness of the surface and the last gives information about the brightness of the surface.

The first reflectivity data for the Venusian surface were obtained from ground-based radar observatories. The first images were acquired in 1972 from the Goldstone satellite radar tracking system located in California. Observations with the Goldstone station were made until 1988 covering essentially the equatorial regions between 15°N and 15°S latitude and 260°E and 30°E (through 0°) longitude. The spatial resolution, that was of the order of 5–10 km in the first images, has been improved to near 1 km in the data acquired since 1986. The second ground-based observations from the Venusian surface come from the Arecibo observatory in Puerto Rico. On this site, the first reflectivity images were taken in 1975 and the last in 1988. With these data, that covered latitudes up to 50°N, the spatial resolution was 1.5 km. From space, the first almost global observation of the surface of Venus was obtained by the NASA PIONEER VENUS Orbiter that arrived at the planet on 4 December 1978. Around 93% of the surface was covered between 74°N and 63°S latitude. These observations gave the first large-scale altimetric and topographic coverage with an altimetric precision of 150 m at a sampling resolution of about 60 km. Then, from October 1983 to July 1984, the twin Soviet probes VENERA 15 and VENERA 16 surveyed the planet and mapped essentially the north pole area covering approximately 25% of the surface from 30°N to 88°N. The resolution of the reflectivity images ranges from 1 to 4 km and the topographic precision was 50 m with a sampling resolution of about 8 km. The last global data from the surface of Venus were obtained by the NASA MAGELLAN Orbiter. The insertion of the probe into the mapping orbit took place on 10 August 1990 and the formal mapping of the surface began on 1 September 1990. These operations, divided in four observation cycles, lasted until 14 October 1994 when the probe was lost. The coverage of the surface was almost complete. The precision of the altimetric data is below 50 m at a sampling resolution of the order of 10 km. The resolution of the reflectivity images is of the order of 75 m.

Infrared observations of Venus have been made from the ground and during the flyby of Venus by two probes. The first one was GALILEO during its travel to Jupiter on 10 February 1990 and the second occured very recently with the CASSINI/HUYGENS probe on 24 June 1999 during its travel to Saturn. These two probes have infrared spectrometers on board (NIMS on board Galileo and VIMS on board Cassini). The surface was detected on the nightside of Venus by its thermal emission. These observations show that the surface temperature is in equilibrium with the atmosphere at the same altitude (see VENUS: ATMOSPHERE).

Local observations

All available on-ground local observations and measurements come from the various Soviet Venera/VEGA landers. These probes have made images of the ground surrounding them and/or composition analysis of the surface. The first of these landers was Venera 8 that landed in 1972 and gave the first natural radioactive element composition of the surface. Then three sets of twin probes reached the surface. The first pair of landers were Venera 9 and 10 that made images of the surface and chemical analyses in October 1975. They were followed seven years later by Venera 13 (1 March 1982) and Venera 14 (5 March 1982) landers that made color images of the surrounding surface and also surface composition analyses. Finally the Vega 1 and Vega 2 probes in 1984 were able to analyse the soil composition.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

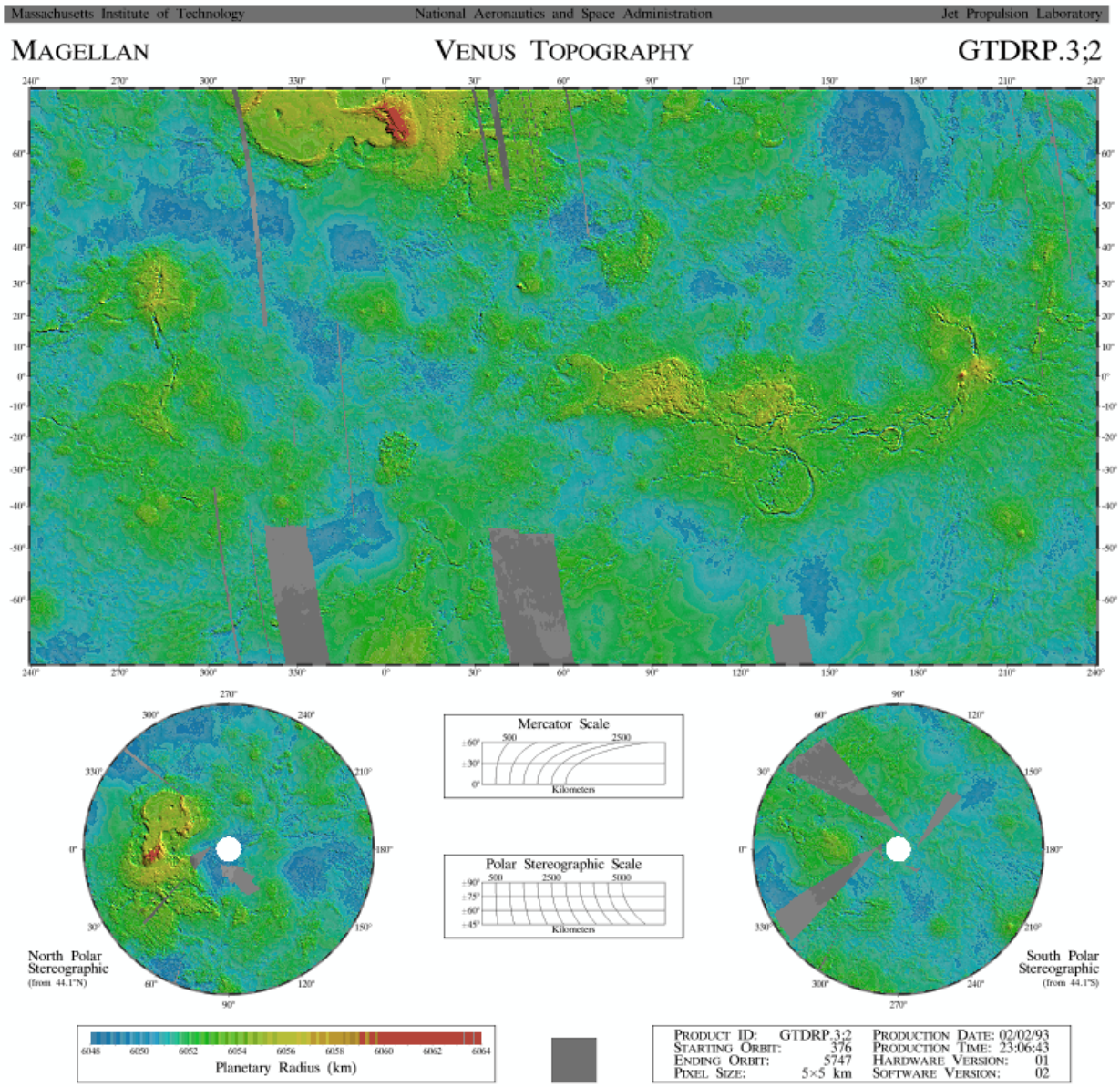


Figure 1. Topographic map of Venus. The highland at the equator is Aphrodite Terra and the highland at the north is Ishtar Terra.

Global topography

Although Venus, like most of the TERRESTRIAL PLANETS, can be divided into two broad physiographic regions, highlands and lowlands, its hypsometric distribution is unimodal, unlike that of the Earth (bimodal) or Mars (trimodal) and over 80% of the Venusian surface lies within 1 km of the mean radius of 6051.84 km (figure 1). The mean slopes on Venus are of the order of 1° but average kilometer-scale slopes greater than 30° are not uncommon.

Highlands

- Regio (regiones, pl.) are topographically high regions, often hosting large shield volcanoes.
- Terra (terrae, pl.) cover vast areas and have variable topographic relief, as continents do on Earth.
- Planum (plana, pl.). Lakshmi Planum is the only planum recognized on Venus. It is a 3–4 km high plateau, bordered by mountainous ridges.

There are four highland areas on Venus: (1) ISHTAR TERRA in the north, (2) LADA TERRA in the south, (3) APHRODITE TERRA in the equatorial zone east of the prime meridian and (4) the last area is defined by BETA REGIO, Phoebe and Themis regiones distributed along roughly the 285° meridian.

Ishtar Terra comprises the high Lakshmi Planum plateau to the west and the extremely elevated region

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

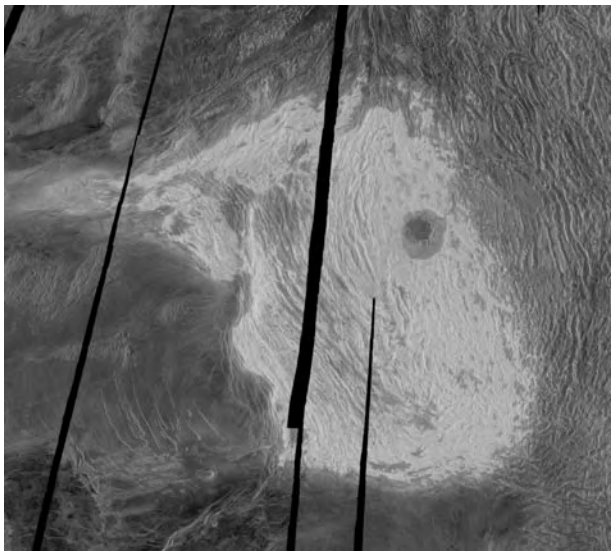


Figure 2. This Magellan full-resolution image shows Maxwell Montes, and is centered at 65°N latitude and 6°E longitude. Maxwell is the highest mountain on Venus, rising almost 12 km above mean planetary radius. The western slopes (on the left) are very steep, whereas the eastern slopes descend gradually into Fortuna Tessera. The broad ridges and valleys making up Maxwell and Fortuna suggest that the topography resulted from compression. Most of Maxwell Montes has a very bright radar return; such bright returns are common on Venus at high altitudes. This phenomenon is thought to result from the presence of a radar reflective mineral such as pyrite. Interestingly, the highest area on Maxwell is less bright than the surrounding slopes, suggesting that the phenomenon is limited to a particular elevation range. The pressure, temperature and chemistry of the atmosphere vary with altitude; the material responsible for the bright return is probably only stable in a particular range of atmospheric conditions and therefore a particular elevation range. The prominent circular feature in eastern Maxwell is Cleopatra. Cleopatra is a double-ring impact basin about 100 km in diameter and 2.5 km deep. A large amount of lava originating in Cleopatra flowed through this channel and filled valleys in Fortuna Tessera. Cleopatra is superimposed on the structures of Maxwell Montes and appears to be undeformed, indicating that Cleopatra is relatively young.

surrounding the Maxwell Montes that are, with their 12 km altitude, the highest topographic point on Venus (figure 2).

Lowlands

- Planitia (planitiae, pl.) are topographic low-lying regions, generally the most featureless on Venus in terms of tectonic and volcanic structures.

There are five lowlands areas in which lie other smaller highlands: (1) Atalanta Planitia at the north of Aphrodite Terra, (2) Aino Planitia and (3) Helen Planitia at the south of Aphrodite Terra, (4) Guinevere Planitia at the equator between Aphrodite Terra and Phoebe Regio and (5) Lavinia Planitia at the south of Guinevere Planitia. The small highlands are Alpha Regio (6°E, 25°S), Eistla Regio (39°E, 18°N), Bell Regio (49°E, 33°N) and Tellus Regio (82°E, 39°N).

The landscape of Venus

The landscape of Venus is characterized by some very particular and specific geomorphological and tectonic features that illustrate how Venus can be very different from the Earth and that can help to understand its geological history.

Volcanic features

Several types of volcanic flows and edifices are recognized. The presence of volcanic features on Venus suggest local magma sources at depth. The distribution of these features may provide clues about the crustal properties and thermal history of Venus. They can be divided into three groups according to their mean size. Brief characteristics are given for each below:

(i) Large volcanic forms (<100 km): large shield volcanoes are characterized by numerous lava flows radiating away from a central caldera. Many individual flows extend for hundreds of kilometers. Several of the 156 identified shields are located in topographically high regions (e.g. Gula and Sif Montes at Western Eistla Regio; Sapas, Maat and Ozza Montes at Alta Regio) reaching elevations as high as 3–5 km above the surrounding area.

(ii) Intermediate-sized volcanic forms (20–100 km):

- Anemones: a type of volcanic edifice characterized by flows radiating outward, often in bilateral fashion from a central graben or fissure. They are relatively rare; only 25 have been identified. Anemones are typically 30–40 by 40–60 km in dimension.
- Ticks are volcanic domes, so named because of their appearance in radar images. They have flat or depressed circular domes about 25 km in diameter and are flanked by strongly defined radial ridges and troughs. At one end, the tick edifice may exhibit signs of a collapsed graben with extensive fault scarps. Occasionally, flows originating from the small central calderas appear to be directed outward along the radial fault-valleys. About 50 have been identified.
- Steep-sided domes, commonly referred to as pancake domes: these volcanic domes have well-defined circular outlines and are characterized by radial fractures near the steep perimeter, radial and concentric fracturing on the flat interior and small summit calderas near the center. Over 150 have been recognized, very often in groups, sometimes overlapping one another.
- Calderas and paterae are volcanic depressions bounded by arcuate fault scarps and are the source area of numerous lava flows. 86 calderas have identified.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

(iii) Small volcanic forms (>20 km) shield fields cover relatively large areas (average area 17 700 km^2) and are delimited by numerous, but small (typically 2–3 km across, but 6–12 km ones are not unusual) volcanic domes or vents. Some shield fields develop extensive flow fields surrounding the shield volcanoes whereas others are located within tectonic structures such as coronae or linear extensional zones. Overall, 556 shield fields have been recognized on Venus.

Tectonic features

Many features recognized from Magellan images are reminiscent of TECTONIC structures on Earth. Some of these features cover regions of thousands of square kilometers, whereas other features are narrow linear structures that appear to be related to fractures, grabens and perhaps dykes. A classification and nomenclature of Venusian physiographic features and how they may be related to tectonism has been made. Some of these features are described below.

(i) Large tectonic forms (100 to 1 000 km) are:

- Tessera (tesserae, pl.): a terrain network consisting of two or more directions of anastomosing linear ridges and troughs. They are tectonically complex terrains, thought to be largely compressional in origin and may represent the oldest part of the crust.
- Chasma (chasmate, pl.): a broad trench or linear zone consisting of a parallel arrangement of troughs or valleys bounded by fault scarps. These zones are interpreted to reflect extensional tectonics.
- Mons (montes, pl.): large highland provinces are termed montes (e.g. Maxwell Montes, Danu Montes, Akna Montes and Freyja Montes), descriptive of their mountain range-like appearance (figure 2).

(ii) Others features likely to be associated with volcanic and/or intrusive activity are:

- Corona (coronae, pl.): (mean diameter = 250 km) are large circular structures whose circumference is defined by an elevated ring-like zone consisting of compressional ridges and extensional troughs. The interior part often shows evidence of volcanic activity and early fracturing. Radial graben-like structures may extend beyond the outer diameter. Topographically, the interior may be raised or depressed relative to the surrounding terrain. Corona structures likely represent the surface expression of mantle upwelling. They range from 75 to over 2000 km across and occur in groups, in chains or as isolated structures. More than 360 have been recognized, of which about half have been formally assigned names (figure 3).

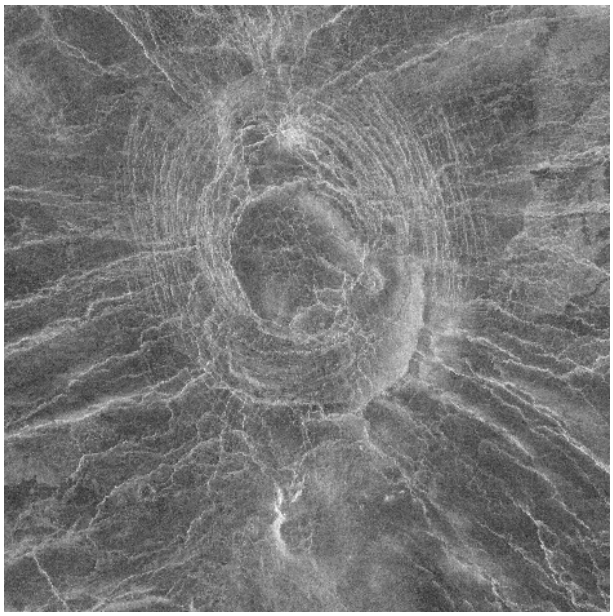


Figure 3. This region, roughly 100 km on a side, shows a gigantic structure known as a corona. Coronae are circular to elliptical features marked by a ring of concentric ridges. Such features are thought to be the result of hot rising bodies of magma that reach the crust. As hot material rises, it weakens the upper layers of the crust, causing the surface to dome upwards. Then as the region cools, the dome begins to subside. As the upper layers rise and fall, they are subjected to stresses that crack the surface, creating both circular and radial fractures. Magellan acquired this view of Venus during its first mapping cycle around the planet in 1990 and 1991.

- Arachnoid(s) (mean diameter = 115 km) are circular to elliptical structures that consist of a central dome or depression surrounded by an extensive network of radial and concentric linear features. These locally developed structures may represent the surface expression of faulting and dyke emplacement associated with magma injection at shallow depths. About 259 have been found on Venus.
- Nova (novae, pl.) (mean diameter = 190 km) are similar to arachnoids, but are dominated by radial structures. Novae are generally centered on a domal uplift and may possibly represent the early stages of corona formation. Over 50 have been identified.

Impact craters

There are 935 recognized impact CRATERS on Venus. About half the craters have been formally assigned names, the others remain unnamed. All have been named after famous women in history, but craters with diameters less than 20 km have been given female common names. Venusian craters range in size from 1.4 km in diameter to

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Uranium, thorium and potassium content in the Venusian surface rocks given by gamma ray spectrometry.

	Venera 8	Venera 9	Venera 10	Vega 1	Vega 2
K_2O, wt %	4.8 ± 1.4	0.6 ± 0.1	0.4 ± 0.2	0.54 ± 0.26	0.48 ± 0.24
U (ppm)	2.2 ± 0.7	0.6 ± 0.2	0.5 ± 0.3	0.64 ± 0.47	0.68 ± 0.38
Th (ppm)	6.5 ± 0.2	3.7 ± 0.4	0.7 ± 0.3	1.5 ± 1.2	2.0 ± 1.0

Table 2. Contents of the major elements in the Venusian surface rocks given by x-ray fluorescence.

	Venera 13	Venera 14	Vega 2
SiO_2	45.1 ± 3.0	48.7 ± 3.6	45.6 ± 3.2
TiO_2	1.59 ± 0.45	1.2 ± 0.41	0.2 ± 0.1
Al_2O_3	15.8 ± 3.0	17.9 ± 2.6	16.0 ± 1.8
FeO	9.3 ± 2.2	8.8 ± 1.8	7.7 ± 1.1
MnO	0.2 ± 0.1	0.16 ± 0.08	0.14 ± 0.12
MgO	11.4 ± 6.2	8.1 ± 3.3	11.5 ± 3.7
CaO	7.1 ± 0.96	10.3 ± 1.2	7.5 ± 0.7
K_2O	4.0 ± 0.63	0.2 ± 0.07	0.1 ± 0.08
S	0.65 ± 0.4	0.35 ± 0.31	1.9 ± 0.6
Cl	0.3	0.4	0.3

280 km. Crater Mead is the largest impact crater identified on Venus.

In contrast to Mercury, Mars or the Moon, which are covered with thousands of craters that have accumulated over the last 4 to 4.5 billion years, Venus is scarred by curiously few. In fact the spatial distribution of craters is uniform (random and anticlustered) over the entire planet, suggesting that Venus experienced complete global resurfacing in the relatively recent (geologically speaking) past.

In the global resurfacing model, tectonic and volcanic activity affected the entire surface of Venus which obliterated the majority of (if not all) previous impact craters. An observation that lends support to the sudden arrest of these events is the fact that the majority of craters, 84%, do not show any signs of modification. This resurfacing activity is thought to have ceased between 300 to 800 million years ago. The uncertainty of the timing lies in the uncertainty of estimating the impact flux.

Craters on Venus are recognized by their expression on images and hence classified by their morphology. The high temperature of the Venusian surface (470 °C) and its thick atmosphere make Venusian impact morphology unique among planetary bodies in the solar system. Based on the development of crater floor structures and degree of circularity, a classification of simple craters and five types of complex crater into a six-fold scheme can be made.

- Structureless craters are simple craters where the internal floor is flat and featureless. The smallest craters are generally of this type.
- Central peak craters have a central uplift that rises above the crater floor. These craters range in size from 8 to 79 km, but are most commonly 16–32 km. Outlining rims are quite circular and often terraced.
- Double-ring craters are defined by an outer rim and a circular arrangement of inner peaks and ridges. These craters are typically greater than 40 km.
- Multiple-ring craters have two or more concentric ridge structures that rise above the crater floor. The largest craters on Venus, ranging from 86 to 280 km in diameter, are of this type.
- Irregular craters have non-circular rim outlines and structural disruptions to otherwise flat crater floors. Almost one-third of the craters on Venus are of this type, most of which are less than 16 km across.
- Multiple crater formation occurs when a falling body fragments into pieces. Each fragment creates a separate impact crater whose rim may overlap with adjacently formed craters. Individuals of this type are up to 44 km in diameter, but most are less than 11 km.

In general, small-diameter craters are flat-floored, have irregular outlines and may be part of a multiple impact event. Complex internal structures occur in large craters and tend to develop progressively as a central peak, a double ring, or a multiple ring, with increasing crater diameter.

Surface properties and composition

A somewhat unexpected finding in the Magellan images was the abundant wind streaks. More than 4000 wind streaks have been identified and this evidence of eolian activity is widespread and is particularly concentrated between 17°S and 30°S latitude and at 5°N–52°N on smooth plains. These wind streaks are generally oriented with the downwind direction towards the equator. This eolian activity is also demonstrated by parabolic crater deposits, made up of fine, centimeters thick debris distributed by high-speed zonal E–W winds. These observations are consistent with the radar signature of the surface that indicates that weathering, mass wasting and eolian activity operated continuously on deposits similar to lava flows. This activity confirms the aspect of the surface as it is known by the Venera lander images (figure 4). The reddish color of the images taken by the Venera 13 and Venera 14 landers also supports the volcanic origin of the surface which is confirmed by the chemical composition.

Our knowledge of the chemical composition of the surface of Venus comes from two types of experiment on board the seven Russian landers. These two methods are gamma spectroscopy and x-ray fluorescence spectrometry. The first one gives the contents of the radiogenic elements,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 4. Venera 14 lander images of the surface of Venus at 1°S, 310°E on 5 March 1982. The lander survived for 60 min. Both images show part of the lander at the bottom. This area is composed of flat basalt-like rocks, but little soil or fine-grained material, as was seen at other Venera lander sites. Near the center of the top image is a lens cover, and the bottom image shows a test arm.

uranium, thorium and potassium (table 1). The second one gives the contents of the major elements of the surface rocks (table 2).

Most of these analyses show a trace element and/or bulk chemistry typical of tholeitic basalts. Moreover, all the dominant type of terrain within the Venera/Vega landing sites is plains so that their measurements can be considered representative samples of the Venusian plains. Finally morphological observations of long and vast lava flows together with the results of the geochemical measurements show that the plains of the landing sites as well as the Venusian plains as a whole are predominantly the result of large basaltic volcanism. Two exceptions come from the data of Venera 8 and Venera 13. The gamma spectroscopic analysis at the Venera 8 landing site shows that the surface material contains relatively high amounts of K, U and Th (table 1). X-ray fluorescence analysis at the Venera 13 site shows that the bulk chemistry of the surface rocks is analogous to that of alkaline basalt. Its potassium content is indeed very similar to that of the Venera 8 site. For these two sites, where non-tholeitic composition of the surface material was identified, steep-sided domes have been found which are not present at the other sites. These particular morphological features, together with the fact that the Venera 8 and possibly Venera 13 material compositions are the combination of evolved igneous rocks and more mafic and primitive rocks, suggest that these rocks are the geochemical signature of the presence and involvement of material of continental crustal origin in their magma genesis. This raises the question about the presence or absence of continental-like crust on Venus, which remains an enigma.

Stratigraphic history

From the study of all these features and of their morphological and chronological relationships, a tentative geological history of Venus can be proposed, defining units of common ages, their formation and evolution through time as described by Basilevsky and Head (1998) (see figure 5).

The oldest unit, the Fortuna Group, is mainly composed of tessera terrains that cover about 8% of the Venusian surface. The morphology of this unit is dominated by intersecting systems of ridges and grooves of tectonic origin. This deformation does not extend into the surrounding plains that clearly embay the tessera. It is not clear if this group was formed in a very short time or if it is composed of various subunits of different ages. However, it is clear that all other units overlie or embay this group, making it the oldest recognized unit of Venus like the Precambrian basement of continents on Earth. Although this group may have formed at different ages, its heavy deformation seems to have spanned a relatively short period of time as suggested by the density of superposed impact craters that is approximately the same.

The next unit in chronological appearance is called the Guinevere Supergroup. It represents an assemblage of plains-forming units and consists of four groups separated from one another and from the underlying Fortuna Group by episodes of tectonic deformation.

The first unit consisting of material of densely fractured plains is called the Sigrun Group and covers 3% of the surface of Venus. It appears as uncovered terrains by the younger plains and is deformed by densely spaced swarms of faults usually subparallel to each other. The material composing this unit appears to be primarily plains that were emplaced as floods of lavas.

The second group, called the Lavinia Group, is represented by the materials of fractured and ridged plains and covers about 3% of the surface. The ridges are sometimes clustered forming ridge belts. They appear

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

generally as elongated islands among the younger plains but embay the tessera and the Sigrun Group. The Lavinia Group material, as the Sigrun Group, was also primarily emplaced as lava floods (figure 5(*a*)).

The third group consists of materials of shield plains and plains with wrinkle ridges. Together this group, called the Rusalka Group, occupies 70–75% of the surface of Venus. The wrinkle ridges plains dominate among this group and represent 60–65% of the total surface. A characteristic of these plains is the presence of wrinkle ridges, typically 1 km wide, that often form a network with a dominant trend. Usually subunits can be recognized in this group because they are deformed by a single ridge network, thus separating them all from the younger stratigraphic units. The youngest of these subunits form extended flow-like features. This observation and the *in situ* geochemical measurements made by the Venera 9, 10 and Vega 1, 2 landers suggest that these plains formed by extended floods of mafic lavas. The shield plains is the second component of the Rusalka group. It is represented by materials of plains formed by clustered and coalescing sloping shields of volcanic origin. These plains occupy 10% of the surface of Venus and are usually embayed by the wrinkle ridge plains that appear consequently younger. The volcanic shields, on the basis of their gentle slopes, are made by lavas composed of alkaline basalts or even more differentiated material such as andesite as shown by the Venera 8 analyses.

The last group of the Guinevere Supergroup is mostly made of the materials of lobate and smooth plains undeformed by wrinkle ridges. This unit covers 10–15% of Venus and is called the Alta Group. It overlies and embays all the already described units. Most of the Alta Group materials are associated with rift zones occurring in the form of large, gently sloping volcanic edifices. Some materials of this group are associated with coronae, forming lava flows aprons around them. Their morphology as well as their geochemical composition given by Venera 14 suggest that they also are mafic lavas.

The last and youngest group, the Aurelia Group, is represented by the materials of radar-dark parabolas associated with the youngest impact craters as well with the eolian patches and streaks.

Absolute age estimates

The duration of the morphologically distinguishable part of the geological history of Venus is estimated on the basis of impact crater densities. The average age of the surface of Venus is estimated to vary between 300 and 500 million years, for the lower bound, up to 800 million years for the upper bound. In terms of this average age T, the age of the oldest group represented by the tessera is estimated to be about $1.4T$. Then, the average age T of the surface of Venus is apparently a good estimate of the age of the Rusalkian Group. This means that the total duration of the Sigrunian, Lavinian and Rusalkian group is of the order of 0.2–$0.3T$ that is, approximately a few tens of millions of years to about one hundred million years. Finally, the upper boundary of the Atlian time (that is, the lower boundary of the Aurelian period) was estimated from the proportion of craters with radar-dark parabolas to be about $0.1T$ prior to the present (figure 5(*a*)).

Geological history of Venus

The history of Venus as recorded on its surface represents only the last 10–20% of the total history of the planet because the morphological signatures of the terrains before those of the Fortuna Group were not preserved. The beginning of this part of the history of Venus is characterized by intensive tectonic deformation on a global scale which formed the tessera terrains. Early stages of this deformation were clearly compressional and later changed into extensional. Termination of the compression stage is estimated to have occurred at about time $1.4T$, while the extensional stage lasted for another 0.1–$0.2T$. Numerous internal dynamical processes have been proposed to account for the tessera-forming deformation sequence; from chemical instabilities causing mantle overturn, an oscillatory convective behavior of the mantle or catastrophic avalanche within the mantle due to the presence of high-pressure phase transitions. These hypotheses raise the question of whether the large currently preserved tessera blocks might represent relicts of downwelling or upwelling. Anyway, this intensive tectonism was accompanied by volcanic activity so that the emplacement of tessera-forming material and its deformation into tessera terrain are the major geological events of Fortunian time (figure 5(*a*)).

After tessera formation, several stages of extensive volcanism occurred that buried vast areas of the tessera and formed the regional plains of Sigrunian, Lavinian and Rusalkian ages. The combined duration of the emplacement of these plains is estimated to be 0.2–$0.3T$. Plain forming of all these groups is separated by episodes of tectonic activity that occurred generally synchronously in different areas of Venus. These episodes are characterized by the dominance of compression, then tension, then again compression, and finally tension. The last globally distributed tectonic episode, extensive wrinkle ridging, happened at about time T and marked the transition to the present stage of the history of Venus, which is dominated by regional rifting and its local associated volcanism in the form of large shield volcanoes. The majority of this stage is represented by the Atlian period, which appears to have lasted until $0.1T$ from the present. This period is thus the longest in duration among all the other units, although its resulting tectonic and volcanic features and deposits represent only 10–20% of the surface of Venus. The last period of the history of Venus is the Aurelian period and is characterized by a certain level of reworking of the surface by eolian processes. From the point of view of volcanic and tectonic processes, it represents simply the continuation of the Atlian period so that Venus may be today already endogenically active at a low level (figure 5(*b*)).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

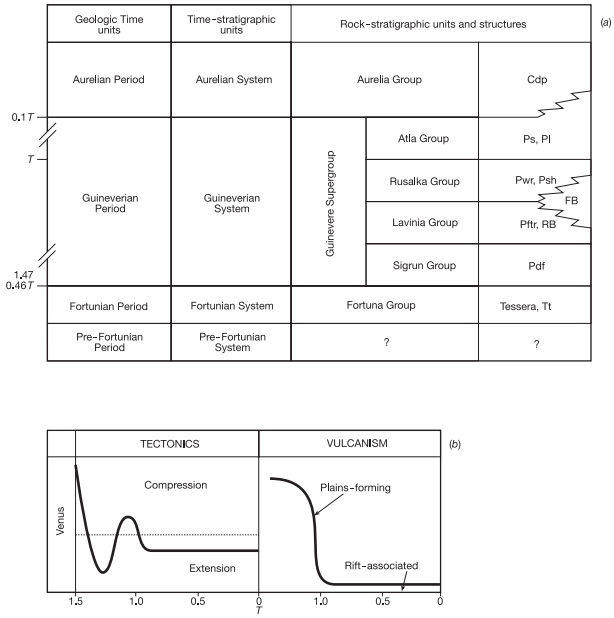


Figure 5. (*a*) Venus stratigraphic units following Basilevsky and Head (1988) (Tt: tessera terrains; Pdf: densely fractured plains; Pfr: fractured and ridged plains; RB: ridge belts; FB: fracture belts; Pwr: wrinkle ridge plains; Psh: plains with shield; Ps: smooth plains; Pl: lobate plains; Cdp: dark parabola craters). (*b*) Summary of the geological history of Venus. The tectonic is dominated by the succession of compression–tension cycles and the volcanism, intense at the beginning, decreases and changes in style with time (Basilevsky and Head 1998). T, the average age of the surface of Venus, is of the order of 300–500 million years or possibly up to about 800 million years.

In summary, the observable part of the history of Venus is characterized by two key points that stand in contrast to the comparable period of Earth history when global geodynamical processes are dominated by plate tectonics.

(a) Venus shows no signature of plate tectonics. Instead, its global tectonic environment is characterized by two successive compression–tension cycles with the magnitude of the deformation and strain rate declining with time.
(b) During the first cycle, plains-forming volcanism occurred at a rate comparable to terrestrial volcanism but in a non-plate tectonic style, and then during the second cycle Venus was dominated by a rift-associated volcanism emplaced at a very low production rate comparable with that of today's intraplate volcanism.

The surface features, structures and composition illustrate the distinctive difference between the recent history of Venus and the Earth that are in other ways so similar.

Bibliography

Barsukov V L *et al* 1986 The geology and geomorphology of Venus surface as revealed by the radar images obtained by Veneras 15 and 16 *Proc. Lunar and Planet. Sci. Conf. 16th, Part 2, J. Geophys. Res.* **91** (suppl B4) D378–D398

Basilevsky A T and Head J W 1998 The geologic history of Venus: a stratigraphic view *J. Geophys. Res.* **103** (E4) 8531–44

Bougher, S W, Hunten M and Phillips R J 1997 *Venus II-Geology, Geophysics, Atmosphere and Solar Wind Environment* (Tucson, AZ: University of Arizona Press)

Carlson R W *et al* 1991 Galileo infrared imaging spectroscopy measurements at Venus *Science* **253** 1541–8

Chapman C R (ed) 1992a Magellan at Venus *J. Geophys. Res.* **97** (E8) 13 063–689

Chapman C R (ed) 1992b Magellan at Venus, part 2 *J. Geophys. Res.* **97** (E10) 15921–6380

Surkov Y A *et al* 1984 New data on the composition, structure and properties of Venus rock obtained by Venera 13 and 14 *Proc. Lunar and Planet. Sci. Conf. 14th, J. Geophys. Res.* **89** B393–B402

Surkov Y A *et al* 1986 Venus rock composition at the Vega 2 landing site *Proc. Lunar and Planet. Sci. Conf. 17th, Part 1, J. Geophys. Res.* **91** E215–E218

Olivier Forni

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Verbiest, Ferdinand (1623–88)

Flemish cartographer and astronomer, member of the Jesuit mission to China in the early seventeenth century, was made President of the Astronomical Board and produced textbooks on astronomy and maps for the Chinese Emperor.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Very Large Array, National Radio Astronomy Observatory

The Very Large Array (VLA), 80 km west of Socorro, NM, is one of the world's premier radio-astronomy facilities, offering researchers a unique combination of resolving power, sensitivity and observational flexibility. Dedicated in 1980, the VLA includes 27 25 m diameter dish antennas, arranged in a 'Y' pattern, that work together as a single aperture-synthesis interferometric imaging system.

A facility of the US National Science Foundation, the VLA is open for use by all scientists on a peer-reviewed basis. Capable of observing at frequencies from 74 MHz to 50 GHz (non-continuous), and with antennas deployed through a 16 month cycle of four standard configurations, the VLA provides resolution ranging from 15 arcmin to 0.05 arcsec.

The VLA's versatility has allowed it to serve a wide range of research specialties, including planetary, solar, stellar and galactic astronomy as well as cosmology. It has made important contributions to the study of both galactic and extragalactic relativistic jets; the Milky Way's central region; galactic structure, dynamics and evolution; supernova remnants; and transient events such as supernovae and gamma ray bursters.

For further information see
http://www.nrao.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Very-high-energy Gamma-ray Sources

Gamma-ray photons at TeV energies have been observed in recent years from a special class of active galactic nuclei, the so-called BLAZARS. The maximum photon energy observed is now approaching energies beyond 10 TeV, opening up new windows into the high-energy universe.

The basic concept in our present day understanding of ACTIVE GALACTIC NUCLEI is the existence of a central BLACK HOLE (our Galaxy is now well established to harbor one of $\sim 2.5 \times 10^6$ solar masses (Genzel *et al* 1997)) with a surrounding ACCRETION DISK. Perpendicular to the inner accretion disk is a stream of high-velocity gas, commonly referred to as a jet; one expects jets on both sides. The material in this jet is usually moving at very near the speed of light. In addition, this basic picture appears to hold quite well for solar mass black hole systems such as some compact binary stars in our Galaxy, as well as for the most powerful quasars we know, with black hole masses of the order of 10^{10} solar masses.

From Hubble Space Telescope observations we have also learned that these massive black holes always sit in the dynamical center of their host galaxies. The spheroidal distribution of older stars within these galaxies has a mass which appears to be approximately proportional to the mass of the central black hole, with a factor of order 300. This can be readily understood in an accretion disk scenario by modelling the entire galaxy as a system of accreting gas, out of which stars are formed and which feeds the budding black hole at its very center. Mergers of galaxies then turn these disks of stars into a more spherical distribution (Faber *et al* 1997, Wang and Biermann 1998).

The radio emission from the environment of these black holes and their associated jets is usually dominated by non-thermal emission from the jet, most often synchrotron emission from the gyrating motion of highly energetic electrons/positrons in a magnetic field. This emission process is accompanied by absorption, and so often synchrotron self-absorption becomes important. In the overall radio emission from the various segments of the jet this leads to an approximately flat spectrum, where the flux density (energy per bandwidth per second received in a unit telescope area) is approximately constant with frequency. A survey of the data shows that such flat spectrum radio sources are almost always variable, and show signs of bulk relativistic motion along the jet.

The power from active galactic nuclei ranges from under 10^{38} erg s^{-1} to over 10^{47} erg s^{-1} (1 erg $= 10^{-7}$ J). The source sizes, best measured by intercontinental radio interferometry, can extend from smaller than the solar system to many millions of light years. Their emission in the electromagnetic spectrum has been observed from around 10 MHz in the radio to a photon energy beyond order of 10 TeV (1 TeV $= 10^{12}$ eV $= 1.6$ erg $= 1.6 \times 10^{-7}$ J).

Basic concepts

We measure motions close to the speed of light in terms of the Lorentz factor γ, where

$$\text{Lorentz factor } \gamma = \frac{1}{\sqrt{1-\beta^2}}. \qquad (1)$$

The relative speed is $\beta = v/c$, where v is the velocity of the particle or material such as gas in a jet under consideration, and c is the speed of light.

We call the motion of matter relativistic, when γ is much larger than unity. Since the total energy E of a moving particle is given by $E = \gamma mc^2$, where m is the rest mass of the particle, the total energy can be very large. The largest particle energies directly observed are 3×10^{20} eV, which corresponds for protons to a Lorentz factor of 3×10^{11}. The motions of jets in active galactic nuclei can often be characterized by a bulk Lorentz factor of order 10. From the electromagnetic emission from these jets, we can infer Lorentz factors of electrons moving around inside the jet in its frame of reference of order 10^6 and lower.

We observe the electromagnetic emission from highly relativistic electrons, and possibly protons, from these jets at all wavelengths, and interpret the radiation at TeV energies also as emission from a relativistic jet. The principal source of this emission can be synchrotron emission, due to the gyrating motion of highly relativistic electrons in a magnetic field. It can also be inverse Compton emission, from collisions between relativistic electrons and photons. Obviously, at high photon energies approaching or exceeding 10^9 (1 GeV), it can also for instance derive from π^0 decay following hadronic interactions.

Hadronic interactions are interactions between nuclei or between other particles where pions (or at extreme energies nuclei) are created. Since pions have a large rest mass, and in turn decay into photons or muons and neutrinos, hadronic interactions are usually heralded by the emission of neutrinos.

Observational techniques

At TeV energies photons are observed on Earth with Cerenkov telescopes. These telescopes use the Cerenkov light emitted from the products of the shower maximum produced by interactions of the incoming primary particle, to determine its energy and direction. Since the overwhelming number of all primaries are directional randomly distributed charged nuclei, the intrinsic properties of photon and hadron initiated air showers (see COSMIC RAYS: EXTENSIVE AIR SHOWERS) may be also used for a separation between photons and hadrons (see Gaisser (1990) and reviews by Weekes (1996), Cawley and Weekes (1996) and Weekes *et al* (1997)).

Looking into the air shower one sees an image, which is rather different for hadronic showers and purely electromagnetic showers. Hadronic showers produce an image which is very irregular in its appearance,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

while electromagnetic showers, initiated by a high-energy photon, give an elliptic image, which points towards the source on the sky. This analysis was originally developed by M Hillas of Leeds, and then evolved into a technique by T C Weekes, of Tucson, at the Whipple telescope (in the mountains of Arizona, USA).

There are internationally many telescopes using this technique, with one of the most active telescope arrays being HEGRA, on La Palma in the Canary Islands. There are further such telescopes in many countries now, such as in Australia, France, India, Russia, South Africa, Ukraine and USA (Meyer 1997, Ong 1998). Table 1 lists some of the major telescopes already in operation.

The acceptance of a large initial detectable photon energy is a strong function of the zenith angle of the observation, and one needs large zenith angles to reach beyond 10 TeV.

Observations

These observations started in 1992 with the first successful detection of Mrk421 by the Whipple observatory (Punch *et al* 1992).

There are now several blazars, which have been observed at such energies, with the most observations being taken of the objects Markarian 421 and Markarian 501, originally discovered by the Armenian astronomer B E Markarian. TeV energies means a factor of 1000 higher than the photon energy (1 GeV $= 10^9$ eV) at which a much larger class of blazars have been observed with the EGRET instrument onboard the COMPTON GAMMA RAY OBSERVATORY satellite (Mattox *et al* 1997). But all these blazars, whether seen only at GeV energies or at TeV energies, have common properties: all of them have a flat radio spectrum in the GHz range, are variable at almost all frequencies observed, and their radiation is interpreted as emission from a relativistic jet. In the case of the TeV blazars, their emission appears to be the greatest at these energies, far greater than at any other wavelength observed (Aharonian *et al* 1997, Bradbury *et al* 1997, Funk *et al* 1998, Hayashida *et al* 1998, Krennrich *et al* 1997, Zweerink *et al* 1997).

The observations now show that the spectral shape is variable, and occasionally extends to the edge of what is observable, suggesting that perhaps the emission continues to many 10 TeV. Some of the time the spectrum is flat, which means that the energy output per logarithmic photon energy interval is constant with energy. The variability time scale is as short as can be detected with the photon counting statistics, at a few hours or less.

Figure 1 shows the spectrum of Mkn501 measured by HEGRA. The work of Aharonian *et al* (1999a) shows that the spectrum continues without an obvious break to 24 TeV.

Physics

These observations are interesting in their own right: the Big Bang produced a bath of microwave background radiation as a remnant of its early hot phase, which can be characterized by a temperature of 2.73 K, based on COBE observations (Fixsen *et al* 1996). Early galaxy star formation extended the background radiation from dust emission around young stars to the far-infrared. Encounters of TeV photons with this microwave and far-infrared background produces electron–positron pairs, which translates to an effective absorption. Therefore observing a source at multiples of TeV energy gives us information about the far-infrared background at wavelengths otherwise nearly inaccessible. Interestingly, early models for galaxy formation in the universe exceeded the level required by the gamma-ray observations by a large margin. The gamma-ray data are still pushing the envelope lower for the possible level of far-infrared background radiation. Therefore these gamma-ray observations provide one of the most stringent limits known for the violent early phases of the universe (Malkan and Stecker 1998, Mannheim 1998, Stanev and Franchescini 1998).

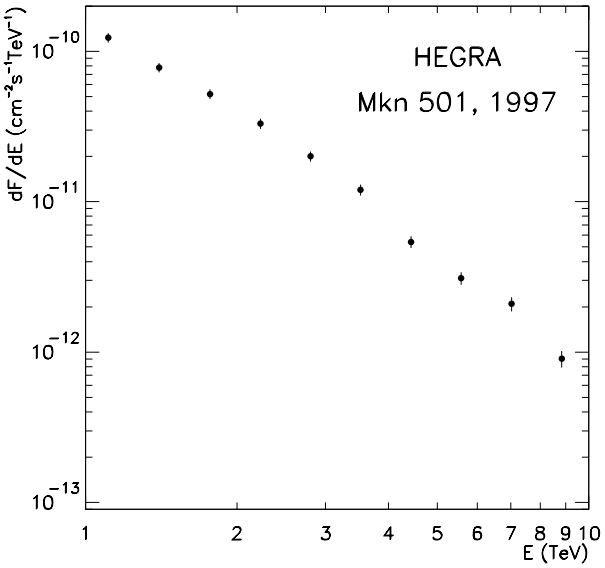


Figure 1. The spectrum of MKn501 as measured by HEGRA (Aharonian *et al* 1999b).

There are two main interpretations proposed for the physical process of the TeV gamma-ray emission itself, a leptonic and a hadronic process (Mannheim 1996).

The leptonic process proposed is a collision between a very energetic (and highly relativistic) electron and a photon from its surroundings either outside the jet, or inside the jet. In such a model electrons are presumed to be accelerated, all the while undergoing losses; this implies one stringent limit to their maximum energy, which leads to a maximum energy of any photon with which the electrons interact. This encounter, called the inverse Compton process, boosts the energy of the photon to much higher energies. In such a picture the maximum photon energy is then limited by the maximum energy the electron could have, and by the kinematics of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Main parameters of the Air Cherenkov Telescope.

Telescope	Site	Country	Longitude	Latitude	Elevation (m)	No of telescopes	Mirror area (m^2)	No of pixels	Pixel ⊘ (deg)	σ_θ (deg)	FOV (deg)	Threshold energy (GeV)	Sensitivity‡ at the threshold ($cm^{-2}\ s^{-1}$)
Whipple	Mt Hopkins	USA	-110.5°	31.4° N	2300	1	74	151	0.25	0.1	3.5	250	10^{-11}
CAT	Themis	France	-2.0°	42.5° N	1650	1	17.5	548 + 52	0.10	0.1	4.8	250	10^{-11}
HEGRA	La Palma	Spain	-17.8°	28.8° N	2200	4	4×8.4	4×271	0.25	0.1		500	4×10^{-12}
HEGRA CT1†	La Palma	Spain	-17.8°	28.8° N	2200	1	5.0	127	0.25	0.1	4.6	1500	3×10^{-12}
TA	Dugway	USA	-113.0°	40.3° N	1600	3	3×6.0	3×256	0.25	0.1	4.5	500	
TACTIC	Mt Abu	India	$+72.7^\circ$	24.6° N	1300	4	4×9.5	4×81	0.31		2.8	700	
CANGAROO	Woomera	Australia	$+136.8^\circ$	31.1° S	160	1	11.3	256	0.12	0.18	3.0	1000	4×10^{-12}
Durham Mk 6	Narrabri	Australia	$+149.8^\circ$	30.5° S	200	1 (3 dishes)	3×42.0	109 + 19 + 19	0.25	0.1	3.4	250	5×10^{-11}
CrAO GT-48	Crimea	Ukraine	$+34^\circ$	45° N	600	2×6 dishes	2×13	6×37	0.40	0.2	2.7	900	5×10^{-12}
Nooitgedacht	Potchefstroom	South Africa	$+27.2^\circ$	26.9° S	1440	4	7	4		1.7	1.7	700	
Patchmari	Patchmari	India	$+78.4^\circ$	22.5° N	1075	25	4	1			3.0		
SHALON	Tien-Shan	Russia	$+75.0^\circ$	42.0° N	3300	1	10	144		0.4	7.2	1000	

† For the old mirrors, i.e. up to November 1997.
‡ Defined as the 5σ limit for 50 h observation time.

collision. In standard models this leads to a limitation of the maximum photon energy to about 10 TeV.

A hadronic process can easily produce photons of much higher energy, by starting with energetic protons, which can clearly produce very energetic photons through proton–photon or proton–proton encounters and subsequent decay of a pion produced in the interaction. We observe high-energy particles directly from outside the Galaxy with energies up to several 10^{20} eV; since the gyration motion of such particles—if they are protons—is larger than the size of the Galaxy, they must come from outside, as argued early by Cocconi (1956). Also, the interaction of these high-energy particles with the microwave background ought to limit their energy upon arrival at Earth to less than 5×10^{19} eV, if they are protons; this expected cut-off is called the Greisen–Zatsepin–Kuzmin cut-off. It is not observed. We are still trying to identify the sources for particles at these extreme energies. If these active galactic nuclei could be shown to require protons at extreme energies, then we would have a very good candidate class. The experiments AGASA in Japan, HIRES in the US, and in the future AUGER in Argentina, are looking for more events at these extreme energies, to beyond 10^{21} eV (Biermann 1997).

We also expect TeV photon emission from the interaction of the known COSMIC RAY population (Berezinsky *et al* 1990, Wiebel-Sooth and Biermann 1999) in the Galaxy with the interstellar medium (Hunter *et al* 1997); since the cosmic ray spectrum of probable Galactic sources extends to 3×10^{18} eV, very high photon energies are expected from the decay of pions resulting from nucleus–proton encounters. Since the cosmic ray population is likely to be stronger in the galactic center region, we expect the as yet unobserved TeV emission to peak at the galactic center.

There is another class of sources for which TeV photons are expected to be detected at some point in the future: the now famous gamma-ray bursts (GRBs; see also GAMMA-RAY ASTRONOMY). GRBs are characterized by a sudden outburst of gamma-ray photons at typical energies of tens of keV to MeV, occasionally to GeV. The time scale of the emission is a few seconds, with substructure down to milliseconds. The time-integrated spectrum can be approximated by two power laws which break around 150 keV, suggesting a non-thermal process just as in active galactic nuclei jets. These GRBs are now known to be at cosmological distances (Metzger *et al* 1997), and thus require enormous powers to be released; it is still unknown what physical picture can explain all these features.

One immediate and clear consequence of the hadronic picture is the emergence of a cosmological neutrino background at high neutrino energies. This neutrino background would be much stronger at high energies than both the atmospheric neutrinos (which have now been used to present impressive evidence for neutrino oscillations by the Super-Kamiokande experiment), and also higher than the Galactic neutrino background, from p–p interactions of cosmic rays in the Galaxy (which in turn can be estimated from the observed GeV gamma-ray spectrum). There is currently an experiment at the South Pole, AMANDA, which hopes to detect this expected cosmological neutrino background.

Similarly a gamma-ray background is expected, which leads to another implicit use of those TeV gamma-ray data: what fraction of the diffuse gamma-ray background is due to active galactic nuclei such as blazars, and what fraction, if any, is left to be derived from different physics? Standard BIG BANG THEORY predicts ubiquituous relics, among them topological defects, which can decay into particles of the order of 10^{24} eV in most versions of the theory, which in turn decay via a cascade into protons, neutrinos and gamma rays, thus producing a truely diffuse gamma-ray background (Sigl *et al* 1995). If we could find an upper limit to the GeV and TeV contribution from the decay of topological defects, it would give a major constraint to particle physics near the energy scale of GRAND UNIFIED THEORIES for particles (GUTs).

Once we have thus identified a class of sources responsible for very-high-energy protons, possibly observable in high energy photons following interactions, and also other high-energy particles, then these accelerators in the cosmos and their environment provide us with tools—albeit very distant tools—to do very-high-energy physics at energies far beyond any physicist's dream.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Acknowledgments

I wish to thank Torsten Ensslin, Heino Falcke, Galen Gisler, Todd Haines, Norbert Magnussen, Karl Mannheim, Sera Markoff, Hinrich Meyer, Rene Ong, Rainer Plaga, Ray Protheroe, Wolfgang Rhode, Günther Sigl, Todor Stanev, Heinz Völk and Yiping Wang for contributions, discussions, perusing the manuscript and/or suggesting corrections.

Bibliography

Aharonian F *et al* 1997 *Astron. Astrophys. Lett.* **327** L5–L8
Aharonian F *et al* for the HEGRA collaboration 1999a *Astron. Astrophys.* **349** 11–28
Aharonian F *et al* for the HEGRA team 1999b *Astron. Astrophys.* **342** 69–86
Berezinskii V S *et al* 1990 *Astrophysics of Cosmic Rays* (Amsterdam: North-Holland)
Biermann P L 1997 *J. Phys. G: Nucl. Phys.* **23** 1–27
Bradbury S M *et al* 1997 *Astron. Astrophys. Lett.* **320** L5–L8;
Cawley M F and Weekes T C 1996 *Exp. Astron.* **6** 7
Cocconi G 1956 *Nuovo Cimento 10th Series* **3** 1433–42
Faber S M *et al* 1997 *Astron. J.* **114** 1771–96
Fixsen D J *et al* 1996 *Astrophys. J.* **473** 576–87
Funk B *et al* 1998 *Astrophys. Part. Phys.* **9** 97–103
Gaisser T K 1990 *Cosmic Rays and Particle Physics* (Cambridge: Cambridge University Press)
Genzel R 1998 *Nature* **391** 17–8
Genzel R *et al* 1997 *Mon. Not. R. Astron. Soc.* **291** 219–34
Hayashida N *et al* 1998 *Astrophys. J.* **504** L71–L74
Hunter S D *et al* 1997 *Astrophys. J.* **481** 205–40
Krennrich F *et al* 1997 *Astrophys. J.* **481** 758–63
Malkan M A and Stecker F W 1998 *Astrophys. J.* **496** 13–6
Mannheim K 1996 *Rev. Mod. Astron.* **9** 17–47
Mannheim K 1998 *Science* **279** 684
Mattox J R *et al* 1997 *Astrophys. J.* **481** 95–115
Metzger M R *et al* 1997 *Nature* **387** 878–80
Meyer H 1997 *Proc. TAUP97* ed A Bottino, A Di Credico and P Monacelli p 392
Ong R 1998 *Phys. Rep.* **305** 93
Punch M *et al* 1992 *Nature* **358** 477–8
Sigl G *et al* 1995 *Science* **270** 1977–80
Stanev T and Franchescini A 1998 *Astrophys. J. Lett.* **494** L159–L162
Wang Y and Biermann P L 1998 *Astron. Astrophys.* **334** 87–95
Weekes T C 1996 *Space Sci. Rev.* **75** 1
Weekes T C *et al* 1997 *Proc. Extremely High Energy Cosmic Rays: Astrophysics and Future Observations* ed M Nagano (ICRR, University of Tokyo) p 160
Wiebel-Sooth B and Biermann P L 1999 Cosmic rays in *Landolt-Börnstein, Handbook of Physics* vol 6 (Berlin: Springer) pp 37–90
Zweerink J A *et al* 1997 *Astrophys. J. Lett.* **490** L141–L144

Peter L Biermann

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vesta

Vesta, with a mean diameter of 529 km, is the third largest main-belt ASTEROID and was the fourth asteroid to be discovered. It was found in 1807 by the German Astronomer HEINRICH OLBERS and named after the goddess of fire and the hearth in the Roman mythology. Vesta is the sole 'intact' asteroid that may have undergone heating at temperatures able to produce a complete planetary-type DIFFERENTIATION (when the heaviest element sank to the core of the asteroid while lighter minerals remained near the surface). Telescopic observations at high angular resolution, either from space or from the ground, can now provide images of Vesta's surface that help to better understand its formation and collisional history. Such images show that Vesta has undergone a large impact event, 4.5 billion years ago. This discovery supports the idea that Vesta is the parent body of a group of small Vesta-like asteroids and possibly the source of a particular type of METEORITES (the basaltic achondrite meteorites) collected on Earth. Although space agencies have not yet selected a large main-belt asteroid as primary target for a robotic mission, Vesta is considered an excellent candidate for future *in situ* exploration of the ASTEROID BELT.

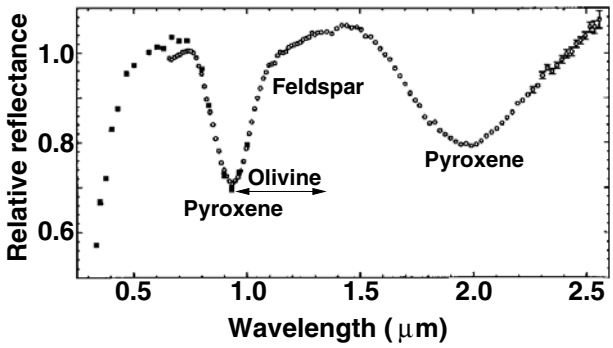


Figure 1. Mean reflectance spectrum of Vesta's surface at visible and near-infrared wavelengths. The deflections from the continuum are attributed to pyroxene and feldspar minerals. Olivine is also present on Vesta but locally only, and its absorption band is not visible in this rotationally averaged spectrum. Pyroxene, feldspar and olivine are characteristic of magmatic material. Their presence on Vesta is the signature of the high temperature it has undergone. Therefore, complete—or partial—melting occurred on Vesta, which is the only large differentiated asteroid that has escaped catastrophic and fully destructive collisions with other bodies of the asteroids' main belt. Spectrum from Gaffey (1997).

Surface properties

About every 18 months, Vesta is at OPPOSITION. Because its surface is made of high-ALBEDO material (table 1), the asteroid can be, at that time, visible with a naked eye (its visible magnitude becomes close to $6th$). Early photometric measurements made by Bobrovnikoff in 1929 showed that the surface of Vesta reflects irregularly the solar flux over a rotation period (∼5.34 h). These brightness changes could be more easily explained by strong albedo variations over its surface than through effects purely induced by an irregular shape. The particularity of its surface properties was confirmed in the 1970s when McCord *et al* obtained the first spectra of Vesta at visible wavelengths. They attributed a strong absorption band centered at 0.93 μm to the electronic transitions on the M_2 site of a magnesian pyroxene, transitions characteristic of magmatic minerals. The nature of its basaltic surface was confirmed soon after with the detection of the second absorption band of pyroxene near 2.0 μm. Figure 1 shows an averaged reflection spectrum of Vesta at visible and near-infrared wavelengths. In addition to the 1.0 μm and 2.0 μm bands attributed to pyroxene, a wide absorption band, also diagnostic of basaltic material, is visible between 1.1 μm and 1.4 μm and is produced by the presence of feldspar on Vesta. The strength of these bands varies across the surface, depending on whether the material has been excavated from underneath the crust by impactors or is a relic of the original eucritic layer. When compared with the spectra of other asteroid classes, the depth of the absorption bands displayed in the reflection spectrum of Vesta is among the strongest. Vesta has undergone powerful but localized impacts in its history that have preserved part of its primordial crust. As a result of these impacts, the variegated surface of Vesta produces strong features in the asteroid's lightcurve.

Table 1. Vesta's ID card.

Mean diameter	529 km [a]
Density	3.5 g cm^{-3} [a]
Rotation period	5.34 h [b]
Semimajor axis	2.362 AU
Orbital period	3.63 yr
Orbital eccentricity	0.097
Orbital inclination	7.14°
Pole coordinates	$RA_{2000} = 301°$, $DE_{2000} = 41°$ [a]
Asteroid type	V
Albedo	0.423 [c]
Composition	Pyroxene, feldspar, olivine

[a] Thomas *et al* (1997).
[b] Gaffey (1997).
[c] IRAS catalog (1992).

Link to HED meteorites

The distribution of minor planets as a function of their heliocentric distance is such that, the further from the Sun, the more primitive are the asteroids. All asteroids having surface properties characteristic of magmatic material are concentrated at the inner edge of the belt. Asteroids of types A (olivine + metal), M (metal), R (iron-poor pyroxene + olivine) or V (pyroxene–plagioclase, diogenite) can be found up to 3.2 AU from our star. Vesta's eucritic crust is nearly intact and is typical of a differentiated object that has undergone thermal heating at temperatures above 1000 K. The source of this heating is

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

still not certain but the actual adopted theories implicate electrical heating induced during the early T Tauri phase of the Sun or decay of long-lived radionuclides. Basaltic ACHONDRITE (or HED, for howardite, eucrite, diogenite) meteorites have similar composition to Vesta: the eucrites display the pyroxene absorptions as well as the weaker band of feldspar while diogenites are pyroxene rich and do not display the feldspar feature. Because of the similarity of their spectra, Vesta is suspected to be the parent body of the HED meteorites. Among the population of differentiated asteroid, Vesta, with a semimajor axis of 2.36 AU, does not occupy a remarkable position that could provide a significant flux of meteorites to Earth. Delivering pieces of Vesta to Earth would require some very energetic impacts that could produce fragments with velocities of the order of $\sim$1000 m s^{-1}. Only with such a kinetic energy could the fragments reach the nearest resonance zone (3:1 resonance with Jupiter) and be ejected towards inner regions of the solar system, and, for some of them, into Earth-crossing orbits. Binzel and Xu discovered in 1993 a family of small Vesta-like asteroids extending from the secular resonance ν_6 ($a = 2.18$ AU) to the 3:1 resonance with Jupiter (2.5 AU). Their visible spectrum displays the 0.9 μm absorption band which is also present in the spectrum of Vesta, eucrites and diogenites. More recent observations with the HUBBLE SPACE TELESCOPE and ground-based telescopes equipped with ADAPTIVE OPTICS showed the presence of a large impact zone located near the Vesta south pole (figure 2). These observations support the big picture in which Vesta could be the source of the family of small Vesta-like asteroids and be the parent body of the HED meteorites.

Although the pieces of the puzzle fit together nicely, alternative solutions still exist. Indeed, Vesta is the only differentiated asteroid that has survived catastrophic collisions, but we know from spectroscopic studies that asteroids from class A represent the inner mantle of some differentiated asteroids now disrupted. Similarly, asteroids from class M could be fragments of the core of some fully differentiated asteroids while class R asteroids are likely to be the product of the disruption of a Vesta-like asteroid, exposing material from the mantle and outer crust. These three asteroid classes originate from Vesta-like asteroids that have been entirely disrupted by impactors and constitute an additional potential source of basaltic achondrite meterorites. Astronomers have found several Earth-approaching asteroids whose surface composition is similar to HED meteorites. In addition, measurements of the oxygen isotope ratios in HEDs match closely those found in stony-iron meteorites and sustain the possibility that the HEDs might come, in fact, from a completely disrupted parent body, more than originating from removal of the outer layers of Vesta.

Regardless of whether the HED meteorites come from Vesta itself or have been created by disruption of one or several Vesta-like asteroids, the undoubted point is that Vesta is unique as the sole 'intact' differentiated asteroid. This point alone provides strong constraints to the collisional history of the whole asteroid belt.

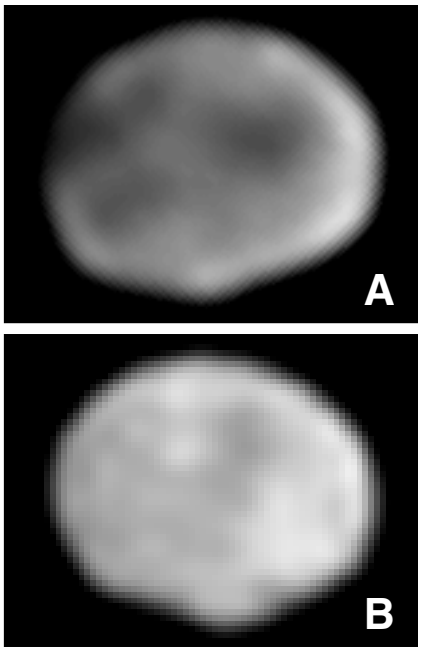


Figure 2. Asteroid Vesta imaged from the ground and space. (A) Image obtained at 0.83 μm from Mt Wilson Observatory on 14 June 1996 using the adaptive optics system on the 2.5 m telescope. A 20° phase angle is responsible for the strong increase in brightness of the right-hand side (eastern) limb of Vesta. The sub-Earth coordinates of Vesta are latitude = $-3°$ and longitude = 255° in this picture. (B) Image obtained by the Hubble Space Telescope and its visible camera WFPC2 on 6 May 1996 with a 5° phase angle. The central coordinates on Vesta correspond to latitude = $-9°$ and longitude = 12°. Both images are oriented so the spin axis is vertical, north pointing up. Note the strong albedo marks visible on Vesta that correspond to various geological units. These albedo changes are related to spatial variations in the surface composition as well as to the action of space weathering (aging process). A complete study at near-IR wavelengths is needed to fully understand the nature and history of these geological units. A simple picture would be to attribute the low albedo regions to old eucritic assemblages and the high albedo features to fresher material excavated from below the crust. The irregular shape of the south pole is believed to be the result of a major, but not fully destructive, impact that occurred on Vesta early in its history. These recent observations support the theory that HED meteorites and small Vesta-like asteroids might originate from such impact. Credit for (A): Christophe Dumas and Chris Shelton. Credit for (B): Alex Storrs (image processing) and NASA/STScI.

Surface mapping

Because of its early differentiation and volcanism, Vesta can be considered as the smallest terrestrial 'planet' of the solar system. Detailed mapping of its surface is important to understand how PLANETESIMALS accreted from the material that was present in the inner regions of the PROTOPLANETARY NEBULA and what endogenic and exogenic processes these planetesimals have undergone. Only a few techniques are currently available to return images with the required spatial resolution for mapping. The

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

most direct and expensive solution consists in obtaining *in situ* measurements of Vesta's surface from a spacecraft. At the present time, only three main-belt asteroids have been visited by a spacecraft: GALILEO flew over GASPRA and IDA (and discovered Ida's satellite named Dactyl) while *en route* to Jupiter, and the NEAR spacecraft visited Mathilde during its journey to a rendezvous with the near-Earth asteroid EROS. Ground-based optical observations were, until recently, limited in spatial resolution because of the deformations undergone by the wavefront of the light while traveling through atmospheric layers of different optical indices. For this reason astronomers used the Hubble Space Telescope to obtain the first resolved images of Vesta. However, new ground-based observational techniques, such as adaptive optics, can now measure the wavefront deformation and correct it in real time. Adaptive optics allows the large telescopes to image at angular resolution equivalent to their theoretical diffraction limit, providing performances at least comparable with HST. Other techniques such as speckle interferometry can provide diffraction-limited images but are still strongly limited to the observation of the brightest objects. Recent improvements in the sensitivity of radar detectors will allow radar observations of asteroids to be extended to the largest bodies in the main belt and their topography to be derived from Doppler-delay measurements.

The surface of Vesta is mostly a pyroxene–plagioclase assemblage with few regions where the excavated crust displays the feldspar-poor inner mantle material (diogenite). At a few locations, impacts occurred at energy sufficient to expose the deeper olivine-rich layer of the mantle. The core of Vesta is presumably made of metallic-rich material or even pure metal. In order to understand how Vesta differentiated, we need to measure the distribution of the diverse geological units over its surface. Depending on how Vesta melted, the distribution of diogenites would be different. In case of total melting of the asteroid, the diogenite-like layer would be confined underneath the eucritic crust while a partial melting would implicate a crust composition made of a mixture of feldspar-rich (plagioclase) and feldspar-poor (diogenites) units. Prior to the HST and adaptive optics observations, the first successful attempt to map the surface of Vesta from the ground was realized in 1981 by Gaffey using rotational spectroscopy. This technique consists in measuring the variation in depth and position of the bands with the rotational phase of the asteroid to derive the location and nature of the main geological units. Such measurement is better constrained in longitude than in latitude but a compositional map of Vesta could be derived and several geological units identified across Vesta. Thanks to the particular inclination of Vesta's rotation axis at the time it was observed, it was possible to locate several diogenite units near Vesta's south pole as well as one olivine-rich region near the equator. These units correspond certainly to a partial removal of the outer crust, or lava flows that transported diogenitic material to the surface through cracks in the mantle. The images recorded at various visible wavelengths by HST confirmed the existence of a large crater near the south pole. This feature is apparently the result of a large, but not entirely destructive, impact which had occurred on Vesta soon after its crust cooled down. Recent ground-based and HST results show an increase in the band depth of pyroxene at high southern latitude that could be explained by a local removal of the crust by the impactor to expose fresher material from the mantle. The family of small Vesta-like asteroids and HED meteorites could have originated from the EJECTA produced during this violent impact. The topology of the asteroid derived from the HST images agrees with this picture and the estimated mass excavated from Vesta during this impact could easily accommodate the formation of Vesta-like asteroids and HED meteorites. Complete mapping of Vesta at similar and even higher spatial resolution in the near-infrared (at a wavelength that provides a better diagnostic of the nature of the minerals) is currently under progress using HST and ground-based telescopes equipped with adaptive optics systems. These observations will bring some important insights to understanding the processes of planetary differentiation that occurred on Vesta and investigating the existing scenarios that link Vesta to the HED meteorites.

Bibliography

Binzel R P and Xu S 1993 Chips off of asteroid 4 Vesta: evidence for the parent body of basaltic achondrite meteorites *Science* **260** 186–91

Binzel R P, Gaffey M J, Thomas P C, Zellner B J, Storrs A D and Wells E N 1997 Geological mapping of Vesta from the 1994 Hubble Space Telescope images *Icarus* **128** 95–103

Gaffey M J 1997 Surface lithologic heterogeneity of asteroid 4 Vesta *Icarus* **127** 130–57

Gaffey M J, Bell J F and Cruikshank D P 1989 Reflectance spectroscopy and asteroid surface mineralogy *Asteroids II* ed R P Binzel, T Gherels and M S Matthews (Tucson, AZ: University of Arizona Press) pp 98–127

Thomas P C, Binzel R P, Gaffey M J, Storrs A, Wells E and Zellner B H 1997 Impact excavation on asteroid 4 Vesta: Hubble Space Telescope results *Science* **277** 1492–5

Christophe Dumas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Viking Mission

The Viking mission was a spacecraft mission to MARS whose main purpose was to search for life. Four spacecraft were involved: two identical orbiter–lander pairs. Each pair was launched separately in the summer of 1975 and arrived at Mars the following year. At Mars the orbiters searched for safe and scientifically interesting sites. The landers were then targeted to those sites and conducted a variety of life detection experiments. The consensus view was that no compelling evidence for life was found (see also LIFE ON OTHER WORLDS).

The spacecraft

The Viking Orbiters (VO-1 and VO-2) were designed and built by the JET PROPULSION LABORATORY (figure 1). The orbiters were solar-powered, three-axis stabilized vehicles with substantial propulsive capability. The launch mass of the orbiters was over 2300 kg of which more than 1400 kg was fuel required for orbit insertion. The orbiter science payload was mounted on a moveable scan platform which provided both thermal control and pointing capability. In addition to communicating directly with Earth, the orbiters were capable of relaying data from the landers as well.

The Viking Landers (VL-1 and VL-2) were built under contract to Martin Marietta Aerospace Corporation (figure 2). The contract was managed by NASA's Langley Research Center. The landers housed the scientific instruments and were powered by radioisotope thermoelectric generators. Each lander had a sampling arm with a scoop for returning samples to selected instruments for analysis. Meteorology sensors were mounted on a boom, and a seismometer was mounted on the lander underside. The landers had a two-way communication link with Earth and a one-way (transmit) link with the orbiters. The mass of the landers was about 600 kg.

To avoid possible contamination from Earth organisms, the landers were heat sterilized and encapsulated in a bioshield. The encapsulated landers, including the entry system, had a total mass of almost 1200 kg. The landers were then mated to the orbiters and the combined craft placed in the shroud of a Titan III launch vehicle. The final launch mass of the orbiter–lander pair was 3530 kg.

Experiments

While the primary scientific objective was to search for life, other objectives were to image the surface and moons, determine the composition of the atmosphere and surface, monitor the weather and climate, and detect marsquakes. To meet these objectives, 13 investigations were conducted (table 1): three mapping experiments from orbit, an atmospheric investigation from the lander during entry, eight investigations on the surface and a radio science experiment.

The orbiter mapping experiments were carried out by three instruments: (1) a pair of high resolution vidicon cameras incorporated into a Visible Imaging Subsystem (VIS), (2) an Infrared Thermal Mapper (IRTM) to determine the thermal properties of the surface and atmosphere and (3) a Mars Atmospheric Water Detector (MAWD). Each of these instruments was mounted on the scan platform and aligned along a common axis to view the same area of the planet.

During the time the landers were traversing the atmosphere, an entry science investigation was conducted. This consisted of a retarding potential analyzer to measure the ion and electron abundances of the upper atmosphere, a mass spectrometer to measure neutral species and a three-axis accelerometer to measure the spacecraft deceleration history. Pressure and temperature were also measured (see also MARS: ATMOSPHERE).

The lander cameras were designed to characterize the local geology, look for macroscopic evidence of life and support the other lander experiments. The two cameras also provided a stereoscopic capability. Furthermore, by monitoring the brightness of the Sun and sky, the cameras could provide information on the opacity of the atmosphere.

The lander biology investigation consisted of three experiments. Each of these acquired small samples of Martian surface material and exposed them to substances of various kinds and then monitored the results. The Carbon Assimilation Experiment was designed to detect photosynthesis of organic compounds by introducing radioactively labeled CO and CO_2 gases incubated under Mars-like conditions. The Labeled Release Experiment sought to detect metabolic processes by introducing an aqueous solution of labeled simple organic materials. The Gas Exchange Experiment also looked for metabolic processes, but by wetting Martian soils with a more complex and unlabeled nutrient solution.

While not principally a life detection instrument, the landers also carried a GCMS which could detect organic compounds in surface material, and measure the composition of the atmosphere. They also carried an x-ray fluorescence spectrometer to measure the abundance of selected elements in Martian surface material. In addition to determining the composition of surface materials, these data provided chemical and mineralogical information to help interpret the biology experiments.

The remaining lander experiments were not directly related to the biology investigation. The lander meteorology sensors recorded the pressure, temperature and wind velocity to help characterize Martian weather systems. The lander seismometers were designed to characterize the background noise level and to listen for major seismic events which would provide information on the interior structure of the planet. An array of magnets of differing strengths was attached to the sampling arm and to the top of the lander. By monitoring (with the camera) the accumulation of particles on these magnets, information on the abundance and composition of magnetic minerals could be obtained. Finally, the physical properties of the soil could be ascertained from

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

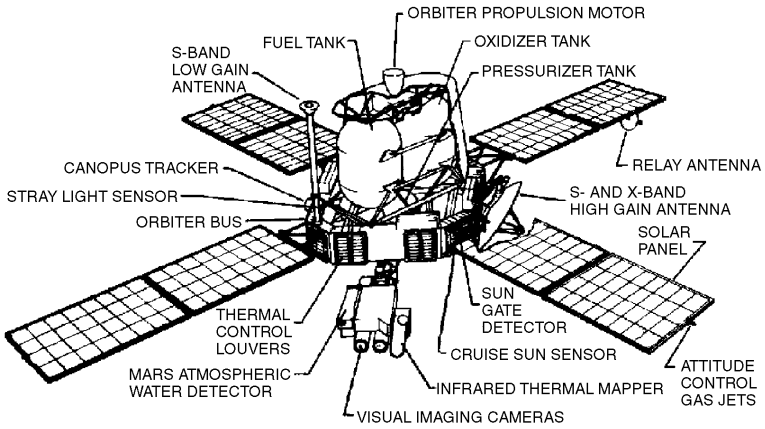


Figure 1. Viking Orbiter.

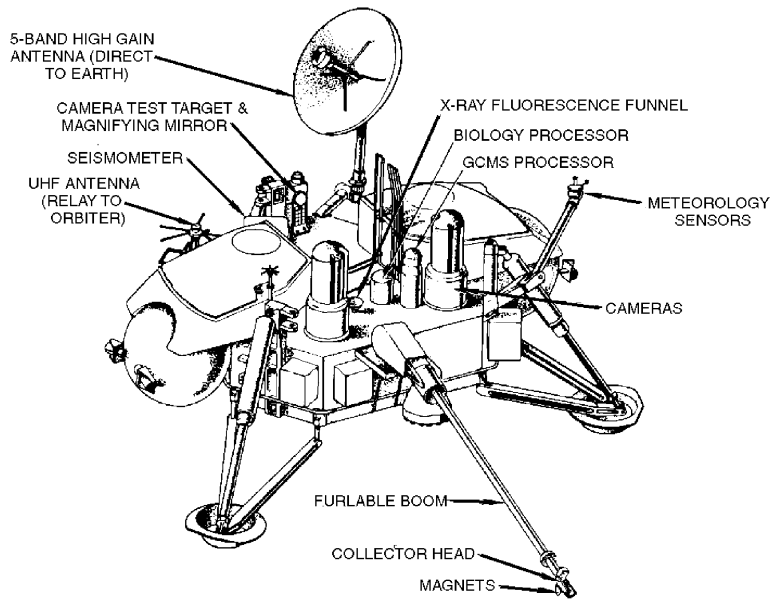


Figure 2. Viking Lander.

engineering data such as variations in motor currents from the sampler arm during digging operations (see also MARS: SURFACE).

The communication system (along with a tracking system) on the orbiters and landers was also used for science. The orbiter radios mapped the planet's gravity field, measured its surface relief and determined local atmospheric temperature profiles during occultation events. Lander radios were used to determine lander locations, the spin rate of the planet and the orientation of the spin axis.

Mission profile

Viking 1 and 2 were launched respectively on 20 August and 9 September 1975 and were inserted into elliptical orbits around Mars on 19 June and 7 August 1976. After several weeks of surface observations, a landing site was selected and the landing sequence was initiated. The lander separated from the orbiter and oriented itself for a 4 km s^{-1} entry into the Martian atmosphere at about 300 km above the surface. An ablatable aeroshell protected the lander from the heat of entry. At 6 km above the surface, a parachute was deployed, the aeroshell was jettisoned and the three lander legs were extended. At 1.5 km, retro-rockets were fired to slow the lander to several m s^{-1} and eliminate horizontal drift. VL-1 touched down on 20 July 1976 in Chryse Planitia (22°N, 48°W) at about 4 p.m. local time; VL-2 landed on 3 September 1976 in Utopia (44°N, 226°W) at about 10 a.m. local time. Both landings were successful.

After landing, there were several months of intense activity. The cameras systematically returned images

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Science investigations and instrumentation.

Investigation	Instrumentation
Orbiter investigations	
Orbiter imaging	Two vidicon cameras
Water vapor mapping	Near-infrared grating spectrometer
Thermal mapping	Solar and infrared radiometers
Entry science	Retarding potential analyzer, mass spectrometer, pressure, temperature and acceleration sensors
Lander investigations	
Lander imaging	Two facsimile cameras
Biology	Carbon Assimilation, Labeled Release, and Gas Exchange Experiments
Molecular analysis	Gas chromatograph–mass spectrometer (GCMS)
Inorganic analysis	X-ray fluorescence spectrometer
Meteorology	Pressure, temperature, wind velocity sensors
Seismology	Three-axis short-period seismometer
Magnetic properties	Permanent magnets on sampler arm
Physical properties	Various engineering sensors
Orbiter and lander investigations	
Radio science	Orbiter and lander communication systems

of the entire panorama around the lander. They also photographed the sky, the Sun, small permanent magnets, trenches dug by the sampler arm and rocks that were intentionally displaced. The meteorology boom was deployed and began returning weather data. Seismic experiments were initiated, although the seismometer at VL-1 failed to uncage (the VL-2 seismometer operated nominally). Numerous soil samples were acquired and subjected to chemical and biological analysis.

The primary mission was to last 90 days for each lander. However, the mission was repeatedly extended as long as the spacecraft remained healthy and returned useful data. The first hardware problem occurred in September of 1977 when VL-2 was no longer able to communicate directly with Earth, but it was not shut down until 12 April 1980. Leaking attitude control propellant left the VO-2 spacecraft unable to align itself with the Sun and it ceased operations on 25 July 1978. VO-1 was also running out of fuel and was commanded off on 7 August 1980. The mission finally ended on 13 November 1982 when contact was lost with VL-1.

Scientific results

Biology

The most significant result was the lack of detection of organic compounds at either landing site by the GCMS experiment. The GCMS was extremely sensitive to organic compounds and could detect them at the parts per billion level for large molecules, a level some 100 times less than those found in deserts on Earth. Yet the results were unambiguously negative. The absence of organic matter, regardless of the results of the biology experiments, argued strongly against the possibility for life in Martian soil.

While the three biology experiments produced results which in some ways simulated life, in detail they were more consistent with the presence of soil oxidants. The Gas Exchange Experiment showed a significant increase in CO_2 and O_2 after the soil was humidified. The CO_2 can be explained by the displacement of adsorbed CO_2 molecules by water vapor, but the O_2 requires some other explanation. The observed 200-fold increase in O_2 is best explained by the decomposition of peroxides in the Martian soil. Such peroxides are theoretically expected from photochemical processes in the Martian atmosphere. Also, a highly reactive oxidant in the Martian soil could readily explain the absence of organic material.

The Labeled Release Experiment showed a surge in radioactive CO_2 after wetting the soil with its mixture of nutrients. Taken by itself, this result satisfied the criteria for a biological interpretation. However, when viewed in context with the other biology experiments and the GCMS, this too was believed to be a chemical rather than biological result. Peroxides present in the soil would have oxidized one of the compounds in the nutrient solution (formic acid) and produced CO_2 gas. Also, if microorganisms were involved, a second injection of the nutrient solution should have produced a further increase in CO_2 levels, yet none was observed.

The results of the Carbon Assimilation Experiments were more complicated, but they too favor a chemical explanation. The amount of fixed carbon produced in these experiments was generally higher than the expected background levels, particularly the first experiment at VL-1. This experiment also showed a temperature sensitivity to the amount of fixed carbon produced. It is now believed that an iron-catalyzed reaction was the source of the signals seen in the Carbon Assimilation Experiment.

Meteorology

Lander meteorological measurements showed that the average surface pressure on Mars is about 7 mbar. However, it varies semiannually owing to the condensation and sublimation of CO_2 in the polar regions. The amplitude of the

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

fluctuation is about 2.3 mbar, or about one-third of the annual mean. Thus, the seasonal polar caps on Mars are made of dry ice (frozen carbon dioxide).

The pressure data also revealed eastward-traveling weather systems and westward-propagating thermal tides. The temperature at the Viking sites ranged from 180 K to 240 K depending on time of day and season and was more repeatable during summer than winter. Winds were found to be strongest during winter with gusts up to 30 m s^{-1}. During summer winds were calm by comparison.

During autumn of the first year, the sky opacity at the Viking sites increased dramatically on two occasions. These increases were associated with the development of global scale dust storms. No such storms were observed during the second and third (Mars) years of operations.

Seismology

No major seismic events were detected. A magnitude 2.8 event (on the Richter scale) was detected 110 km from the Viking Lander 2 site. The signal was significantly damped within minutes, indicating the likely presence of water and gas within the crust. From shear wave reflections, the crust was estimated to be about 15 km thick in this region. The natural background level of seismic activity was found to be very low. Winds were a major source of the seismic background because of the poor coupling of the seismometer to the ground.

Atmospheric composition

The composition of the Martian atmosphere was measured by the Viking landers during entry and descent and while on the surface. The principal components are CO_2 (95.32%), N_2 (2.7%), Ar (1.6%) and O_2 (0.13%). The mean molecular weight is 43.34. These gases were found to be well mixed up to 120 km altitude. Above 120 km they begin to diffusively separate. In the ionosphere O^+ was found to be the dominant species with CO^+ about an order of magnitude less abundant.

Isotopic ratios of various elements were also measured. The stable ISOTOPES of carbon and oxygen were found to be similar to Earth's, but $^{15}N/^{14}N$ was enriched by 60%. The enrichment is due to the selective escape of the lighter isotope with respect to the heavier one, and it implies that Mars may have had a denser atmosphere in its distant past. $^{40}Ar/^{36}Ar$ and $^{129}Xe/^{132}Xe$ were also found to be enriched with respect to terrestrial values. In this case, however, the enrichment is thought to be due to removal of a substantial fraction of Mars' early atmosphere by impacts events followed by the accumulation of the radiogenic isotopes ^{40}Ar and ^{129}Xe.

Atmospheric water detection

Water vapor in the Martian atmosphere was found to vary seasonally and spatially. On average, a column of Martian air contains the equivalent of a layer of liquid water about 10 μm deep. EARTH'S ATMOSPHERE, by comparison, contains 2–4 cm of water. Approximately 100 μm of water vapor was observed over the north polar cap during summer. Very little water vapor was observed over the south polar cap during its summer. This difference in behavior is due to the different compositions of the summer residual caps: the north summer cap is made of water ice, while that in the south is made of CO_2 ice.

Atmospheric structure

During their descent to the surface, the Viking landers recorded the spacecraft deceleration history. These data were used to construct a temperature profile from near the surface to about 200 km. The profiles at the two landing sites were found to be very similar even though they were collected at different locations and times. Temperatures decreased from 230 K near the surface to about 150 K aloft, and they showed considerable structure possibly due to thermal tides and/or gravity waves.

Thermal mapping

Surface temperatures were mapped by the IRTM instrument. Highest temperatures (300 K) occur in early afternoon in southern subtropical latitudes during summer. Lowest temperatures (130 K) occur in polar regions during winter. The daily variation in surface temperature was used to map the thermal inertia of the soil. Thermal inertias range from about 50 to 500 SI units and indicate fine sandy material and well-consolidated bedrock, respectively. Atmospheric temperatures near 25 km were also mapped and showed great sensitivity to dust.

Orbiter imaging

Over 50 000 images were returned by the Viking orbiters. The images covered the entire planet at a resolution of 200 m and large areas at resolutions as high as 7.5 m. Much of the planet has been photographed in color and stereo, although at lower resolution. Orbiter cameras also imaged the moons of Mars (PHOBOS AND DEIMOS), and photographed the shadow of Phobos as it crossed the VL-1 site.

For almost two Mars years, the orbiter cameras monitored weather patterns, the advance and retreat of the polar caps, dust storm activity and the continual redistribution of fine-grained material on the surface. They returned pictures of enormous volcanoes, a vast equatorial canyon system, numerous impact craters, polar layered deposits and fluvial features suggesting catastrophic flooding and possible climate change.

The northern hemisphere was found to be geologically young and covered with volcanic fields. The giant shield volcanoes in the Tharsis region have relatively few craters indicating that volcanism lasted for an extended period of time. In contrast, the southern hemisphere is old and heavily cratered. Parts of its surface may have formed at the end of the heavy bombardment period 3.8 billion years ago.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Soil properties
The fine-grained material at the two Viking sites has similar elemental abundances. Silicon is the most abundant element (21%), followed by Fe (13%), Mg (5%), Ca (4%), Al (3%) and S (3%). The surprisingly uniform soil composition suggests planetary-scale homogenization of the soil by winds. The soil is weakly cohesive and has a bulk density of about 1.2 g cm^{-3}. Magnets on top of the landers attracted soil-derived airborne dust. About 2% of the soil contains magnetic material with maghemite (γ-Fe_2O_3) being the most likely candidate. A plausible model for the soil material is that it is a basaltic weathering product.

Bibliography
Soffen and Young 1972 *Icarus* **16** 1–16

give a description of the Viking mission as it was originally envisioned. For an historical perspective of the Viking mission see

Ezell and Ezell 1983 *On Mars: Exploration of the Red Planet, 1958–1978* NASA SP 4212

The Viking Project as flown is described by

Soffen 1977 *J. Geophys. Res.* **82** 3959–70

The first results of the Viking mission were published in

1976 *Science* **193** (27 August)
1976 *Science* **194** (1 October)
1976 *Science* **194** (17 December)

More comprehensive papers were published in special issues:

1977 *J. Geophys. Res.* **82** (September)
1978 *Icarus* **34** (June)
1979 *J. Geophys. Res.* **84** (December)
1982 *J. Geophys. Res.* **87** (November)

The state of the art in our thinking about Mars has been summarized in the book

Kieffer *et al* (ed) 1992 *Mars* (Tucson, AZ: University of Arizona Press)

Robert M Haberle

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vilnius University Observatory

Vilnius University Observatory, the astronomical observatory of Vilnius University, Lithuania, was founded in 1753. In 1831, when Vilnius University was closed, the observatory was entrusted to the Academy of Sciences of St Petersburg (Russia) and continued its activities until 1881. The Observatory resumed its operation in 1919 when Vilnius University was reopened. Activities of the observatory were again interrupted by Nazi occupation (1941–44) and resumed after World War II. Its main instruments are a 60 cm reflector located at Moletai Observatory 80 km north of Vilnius and a 48 cm reflector located at the Maidanak observing site in Uzbekistan, Central Asia. Present staff consists of eleven astronomers and six technicians. Activities are in the field of construction of photometric equipment, investigation of physical properties of stars, interstellar matter and galactic structure.

For further information see
http://www.ff.vu.lt/astro/inde40.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Virgo

(the Virgin; abbrev. Vir, gen. Virginis; area 1294 sq. deg.) an equatorial zodiacal constellation which lies between Leo and Libra, and culminates at midnight in mid-April. Its origin dates back to Babylonian times and it has been associated with numerous female deities, including Ishtar (Babylonia), Isis (ancient Egypt), Demeter (ancient Greece) and Astraea (ancient Rome). Its brightest stars were catalogued by Ptolemy (*c*. AD 100–175) in the *Almagest*.

Virgo is the second-largest of the 88 constellations but is not particularly prominent, save for its brightest star, α Virginis (Spica or Azimech), magnitude 1.0. Other bright stars include γ Virginis (Porrima), a fine binary with pale yellow (F0) components, both magnitude 3.6 (combined magnitude 2.7), separation 2.7″, period 168.7 years (closest 2005), ε Virginis (Vindemiatrix or Almuredin), magnitude 2.9, ζ Virginis, magnitude 3.4, and δ Virginis, also magnitude 3.4. There are five other stars of magnitude 4.0 or brighter.

There are no bright open star clusters or nebulae in Virgo, but the constellation contains a large number of galaxies, many of which are members of the Virgo Cluster, which extends into neighboring Coma Berenices. The brightest members of the Virgo Cluster are the giant elliptical galaxies M49 (NGC 4472), 8th magnitude, and M87 (NGC 4486, Virgo A), 9th magnitude, which is a strong x-ray and radio source and from which emanates a jet of material thought to have been expelled from a black hole at its center. An interesting non-cluster galaxy is M104 (NGC 4594, the Sombrero Galaxy), an 8th-magnitude spiral with a dark lane of obscuring dust presented almost edge-on.

The constellation also contains the first quasar to have been detected (in 1963), 3C 273, magnitude 12.9.

See also: Sombrero Galaxy, Spica, Virgo A, Virgo Cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Virgo A

The brightest radio source in the constellation of Virgo. Virgo A, otherwise known by the catalog number 3C 274, coincides with M87 a giant elliptical galaxy, and active galaxy, located at a distance of some 50 million light-years in the Virgo cluster of galaxies. Although part of the radio emission comes from two elongated lobes, one on either side of the center of the galaxy, the lobes, which span an overall diameter of about 16 000 light-years, are contained within the optical galaxy (rather than, as with most radio galaxies, extending beyond its visible perimeter). The majority of the radio output comes from a jet that emerges from the core of the galaxy and stretches out some 8000 light-years into one of the two lobes.

The jet radiates at all wavelengths from x-ray to radio. Like the lobes, it emits synchrotron radiation (radiation that is generated by electrons moving at very high speeds in magnetic fields). The motion of bright knots in the jet indicate that they are traveling outwards at about half the speed of light and implies that the electrons themselves must be traveling within the jet at speeds at least as high, if not higher, than this.

With a mass of at least 10^{12} solar masses, M87 is an exceptionally massive galaxy. Observations of the rates at which stars and gas clouds revolve within its central core, and of the way that stars appear to be concentrated at its center, imply that M87 contains a compact massive object, most probably a black hole, of about 3 billion solar masses.

See also: active galaxy, astrophysical jets, radio galaxy, synchrotron radiation, M87, Virgo cluster.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Virgo Cluster

The Virgo cluster is the nearest and best-studied rich CLUSTER OF GALAXIES, lying at a distance of about 55 million light-years in the constellation of VIRGO. Figure 1 is a map of the distribution of nearby galaxies (as determined by their REDSHIFTS) in a 1500 square degree region of the constellation of Virgo. The Virgo cluster is the strong, somewhat irregular, concentration of galaxies at the center. In three-dimensional (3D) space, the Virgo cluster constitutes the nucleus of the Local Supercluster (LSC) of galaxies, in whose outskirts our MILKY WAY GALAXY is situated.

As early as 1784, CHARLES MESSIER noted an unusual group of 'nebulae' in Virgo. Fifteen out of the 109 famous 'Messier objects' are, in fact, member galaxies of the Virgo cluster. However, only in the 1920s, following EDWIN HUBBLE's proof of the extragalactic nature of those 'nebulae', was Messier's group understood as a self-gravitating system of hundreds of galaxies, and the first systematic investigations of the Virgo cluster, as it was henceforth called, were carried out by HARLOW SHAPLEY and others.

Virgo was the first galaxy cluster to be studied dynamically by Smith and Zwicky in the 1930s. Their work showed that the dynamical mass of Virgo, estimated by using the virial theorem

$$2T + U = 0$$

where T and U are the time-averaged kinetic energy and potential energy of the relaxed system, was much larger than the mass inferred by integrating the light of all the galaxies in the cluster and multiplying by a mass-to-light ratio (M/L) like the average of stars in the solar neighborhood. This was the first clear detection of 'DARK MATTER,' or more properly, non-luminous mass. The distribution of galaxies in the direction of Virgo in redshift space is shown in figure 2, and figure 3 displays the histogram of galaxy velocities inside a circle of 6° radius centered on the core of the cluster at 12^h28^m and $+10^\circ28'$.

The Virgo cluster lies at the center of the LSC, first dynamically studied by G de Vaucouleurs and collaborators in the 1950s. The Milky Way appears to be falling into the Virgo cluster, relative to the general expansion of the universe, with a velocity of $\sim$250 km s^{-1}. That is to say, we are still moving away from the cluster core with an apparent velocity, cz, of $\sim$1100 km s^{-1}, but that velocity is 250 km s^{-1} less than it would have been if the cluster had no mass. Curiously, the M/L for the whole LSC derived from this Virgo infall is similar to that derived via the virial theorem for the dynamically relaxed cluster core; both M/L values, when applied to the mean luminosity density of the universe derived from redshift surveys, imply a mean matter density of the universe that is only 1/4 to 1/3 of the critical density.

The kinematics of the cluster is quite complex as can be seen in figures 1 and 4. There are two main concentrations of galaxies; the largest and densest is centered on the well-known galaxy M87 (NGC4486) which is both a strong radio source (VIRGO A) and a strong x-ray source. This subcluster, at $12^h30^m.8 + 12^\circ23'$, is also a strong x-ray source produced by a reservoir of extremely hot intracluster gas. This INTRACLUSTER MEDIUM is at a temperature of $\sim10^7$ K, emits thermal bremsstrahlung radiation and is a major contributor to the total mass of the system. The southern large subcluster is centered on M49 (NGC 4472) at $12^h29^m.8 + 8^\circ00'$, the most luminous galaxy in the cluster, but is significantly less massive than the M87 subcluster. The total mass of the Virgo core region is $\sim10^{15}M_\odot$, that of the M87 subclump $\sim3 \times 10^{14}M_\odot$ and the M49 subclump is $\sim1 \times 10^{14}M_\odot$, assuming a distance to the cluster core of 16 Mpc. The cluster distance of $\sim$16 Mpc has been estimated via a variety of techniques including measurements of Cepheids to six galaxies with the Hubble Space Telescope, studies of globular clusters and planetary nebulae (PNe) luminosity functions, and surface brightness fluctuations (SBFs).

At present, approximately 1300 member galaxies of the Virgo cluster are known, most of which are very faint DWARF GALAXIES (but this number is bound to grow fast in the near future, as ever fainter members are going to be detected with improving techniques). Their projected positions within the cluster—spread over a large sky area of roughly 100 square degrees, or 500 full moons (see figure 4)—reveal a rich substructure. The Virgo cluster indeed represents the most common type of clusters of galaxies that are loosely concentrated and 'irregular'. Several gravitationally bound subsystems of galaxies have been identified. Of these, the subcluster centered on Messier 87 is by far the largest and most massive structure (the Virgo cluster proper, so to speak)—and M87 itself, the giant, active galaxy with its famous jet, can truly be regarded as the heart of the Virgo cluster. The smaller subclusters seem to be in a state of merging with the M87 subcluster. In this sense, the Virgo cluster (like many, if not most, clusters of galaxies) is still in the making.

The Virgo cluster has always been, and still is, one of the most important stepping stones for the cosmological distance scale. Much of the current debate on the value of the HUBBLE CONSTANT boils down to a debate on the mean distance of the Virgo cluster (see below).

A new and very exciting era of Virgo cluster research has recently been opened by the first detection of intracluster PLANETARY NEBULAE (PNe) and single RED GIANT STARS, i.e. stellar objects that are not bound to a single galaxy but to the cluster as a whole. This adds a whole new population to the cluster, which in the future will be used to explore the 3D structure and dynamical history of the Virgo cluster.

Global structure

The primary optical database for the galaxy content of the Virgo cluster is the Las Campanas photographic survey, carried out in the 1980s by Allan Sandage and collaborators, and encapsulated in the Virgo Cluster Catalog (VCC). Figure 4 is a map of the ca 1300 galaxies in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

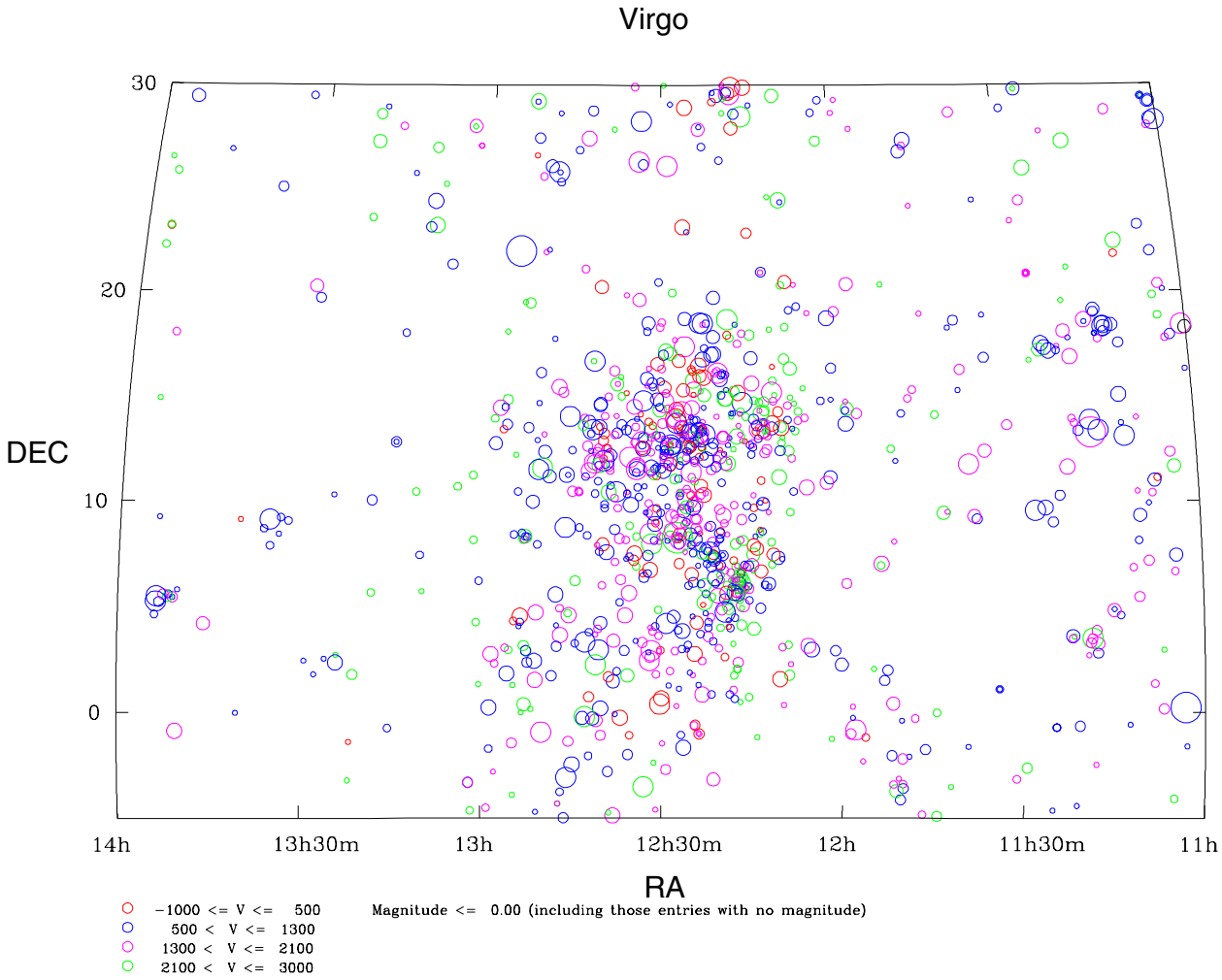


Figure 1. The distribution of nearby galaxies (apparent recession velocities less than 3000 km s^{-1}) in the direction of Virgo. Each circle represents a galaxy; the size of the circle represents its apparent brightness. Galaxies with apparent velocities less than 500 km s^{-1} are shown in red, those above 2100 km s^{-1} in green, and those between 500 and 1300 km s^{-1} and between 1300 and 2100 km s^{-1} in blue and magenta, respectively. **This figure is reproduced as Color Plate 63.**

the sky area covered by the Las Campanas survey that were judged to be members of the Virgo cluster. The membership criteria were based on (1) the morphological appearance of the galaxies, e.g. dwarf ellipticals have a characteristically low surface brightness, and/or (2) the measured radial velocities, which for members have to be smaller than ca 2700 km s^{-1} (heliocentric). Velocities are at present available only for the brightest 400 members (but this number, too, will steadily grow in the future).

The magnitude limit of completeness of the Las Campanas survey is around apparent blue magnitude $B = 18$ or, if we assume a mean distance of 16 Mpc, absolute magnitude $M_B = -13.0$. However, fainter members (up to $B = 20$) were included as well. For comparison, the brightest cluster members, M49 and M87, have $B \simeq 9.0$ and $B \simeq 9.5$, respectively. Thus the known cluster population spans a range of ≈10 000 in luminosity.

From the distribution of symbol sizes in figure 4 one can obtain a feeling of the exponentially growing luminosity function of galaxies. Of the 1300 known members, 850 alone are of the dwarf elliptical (dE) type. The basic characteristic of dEs is a relation between luminosity and surface brightness: fainter dwarfs have also lower surface brightness. As the detection limit for extended objects such as galaxies is set by the surface brightness rather than total magnitude, it is clear that many hundreds, if not thousands more, of extremely faint and diffuse dE members of the Virgo cluster have yet to be discovered by future deep surveys. These will be the analogs of the dwarf spheroidal companions of our Galaxy, which are as faint as $M_B = -8$ (corresponding to $B \approx 23$ at Virgo distance).

Among the remaining 450 member galaxies, there are roughly 80 elliptical (E) and S0 galaxies, 130 spirals, 90 irregulars and 90 dwarf galaxies of intermediate type

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

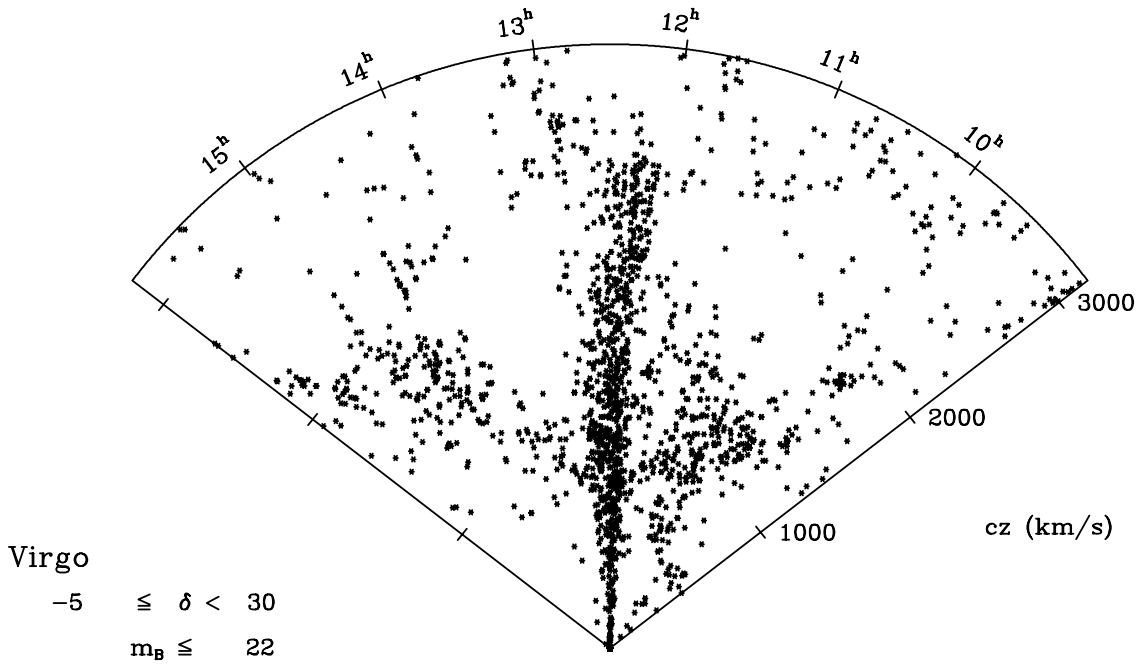


Figure 2. The distribution of galaxies in redshift space in the direction of the Virgo cluster. We are at the apex of the wedge. The main body of the cluster is the picturesquely named 'Finger of God' at the center of the wedge. This is not a real feature. Its extension along the line of sight in this plot is a measure of the internal velocity dispersion (and thus the mass) of the galaxy cluster so that a galaxy with a given radial velocity is placed in front of or behind its true position. The large and diffuse structure seen to the right (10.5^h–11^h and 1100 km s^{-1}) of Virgo is the Leo group.

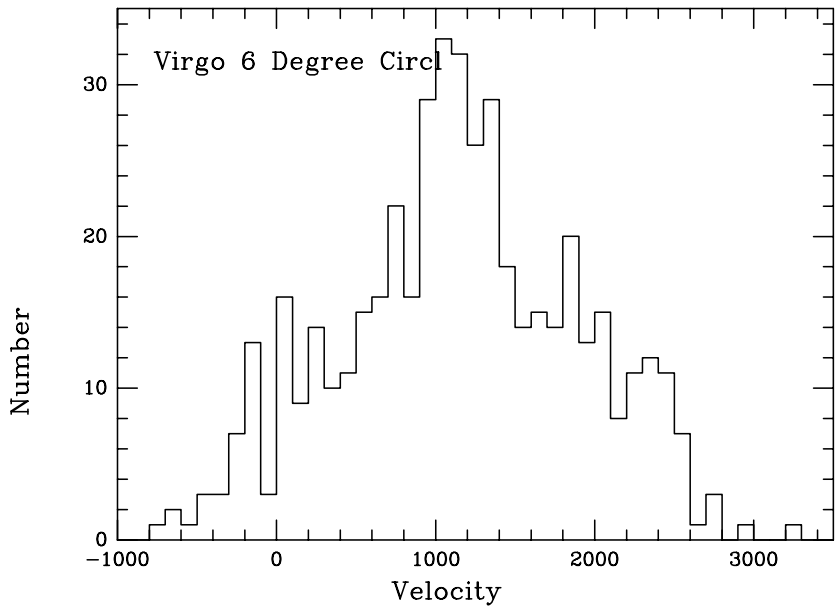


Figure 3. Histogram of apparent velocities for galaxies within a 6° radius of the center of the Virgo cluster. The breadth of the distribution from 0 to 300 km s^{-1} accounts for the 'Finger of God' structure in figure 2.

(Irr–dE). The distribution of these morphological types within the cluster varies considerably—in accord with Dressler's GALAXY 'MORPHOLOGY–DENSITY RELATION': Es and S0s are strongly confined to the regions of highest galaxy density, defining the 'skeleton' of the cluster (dwarf Es only to a slightly lesser degree), while spirals and irregulars

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

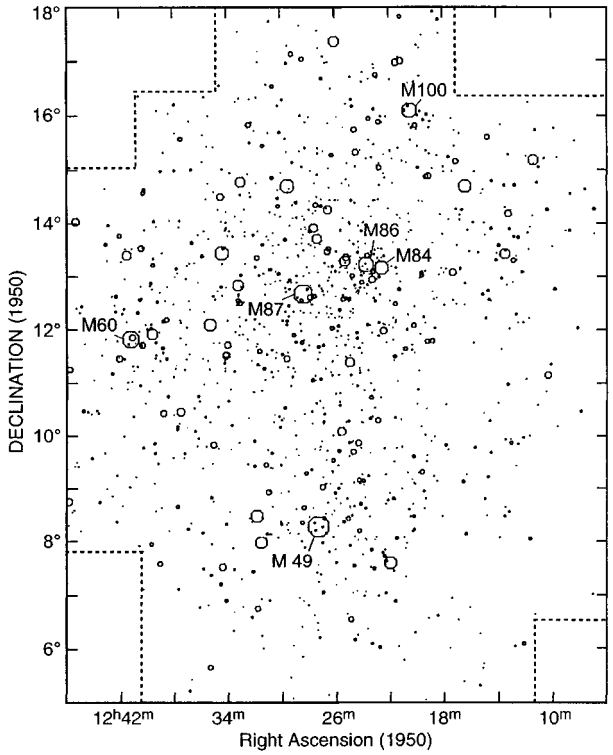


Figure 4. Map of the Virgo cluster as it appears in the sky. All cluster members are plotted with luminosity-weighted symbols. The symbol size (area) is proportional to the luminosity of the galaxy. The most prominent Messier galaxies are indicated. The broken lines indicate the boundary of the Las Campanas survey.

(S+Irr) are very weakly clustered, lying preferentially in the cluster outskirts.

This morphological segregation is also reflected in the velocity distributions. An analysis of the radial velocity data so far available gives a velocity dispersion (1σ) of $\approx$600 km s^{-1} for early-type galaxies (E, S0, dE), but one of $\approx$750 km s^{-1} for late types (S+Irr), i.e. late types are more dispersed in space and velocity. Moreover, the velocity distribution for S + Irr is distinctly non-Gaussian with a low-velocity and a high-velocity wing. These wings are probably caused by infalling and expanding shells of galaxies around the cluster core. Most Virgo spirals and Irrs may have fallen only recently, or are still in the stage of falling, into the cluster from the surroundings, explaining their lack of dynamical relaxation (virialization). Those spirals observed to be H I deficient have apparently already fallen through the cluster core.

The all-cluster mean heliocentric velocity is $\langle v\rangle_{\rm helio} \simeq$ 1050 km s^{-1} (there is no difference with respect to type). Note that this is not the cosmic (Hubble flow) recession velocity of the Virgo cluster, $\langle v\rangle_{\rm c}$, which can be derived by (1) correcting for the solar motion with respect to the centroid of the Local Group by subtracting $\approx$100 km s^{-1} and (2) correcting for the Virgocentric infall (deceleration) of the Local Group by adding $\approx$250 km s^{-1}. This finally gives $\langle v\rangle_{\rm c} \simeq 1200$ km s^{-1}.

Subcluster dynamics

The primary structural characteristic of the Virgo cluster (cf figure 4) is certainly its high degree of irregularity and subclustering. Several subclumps (gravitationally bound subclusters) of galaxies suggest themselves: a main northern subclump around M87, M86 and M84, a southern clump around M49 and possible subgroups around M60 and M100. The global structure of Virgo seems defined by two main axes: one N–S, i.e. M100–M86/M87–M49, and one E–NW, i.e. M60–M87–M86/M84. Remarkably, the former axis is nearly perfectly aligned with the position angle of the outer isophotes of M87, while the latter is perfectly aligned with the jet axis of M87. This shows once more that M87 is the heart of the Virgo cluster.

The reality of the main N–S double structure cannot be doubted, because the northern and southern subclumps are sufficiently well separated. Both have a very similar mean radial velocity, which would suggest that they are at the same distance, i.e. lying in the plane of the sky (however, see below). The southern M49 clump is rather spiral rich and has a surprisingly small velocity dispersion of $\sigma_v \approx 500$ km s^{-1}.

The core region with M87, M86 and M84 is much harder to disentangle. The key observation here is that the velocity distribution of galaxies in this northern clump (especially for dEs) is strongly skewed towards low velocities. In the low-velocity tail we find the most blueshifted galaxies known in the sky (the record holder, VCC846, has $v = -730$ km s^{-1}). These objects tend to be clustered around M86, which itself has a negative velocity ($v = -227$ km s^{-1}). On the other hand, the velocity distribution is peaked around $v = 1300$ km s^{-1}—nearly coinciding with the velocity of M87 ($v = 1258$ km s^{-1}).

A clear asymmetry in the velocity distribution of a cluster of galaxies is almost certainly an indication of ongoing subcluster merging. In the Virgo cluster we seem to witness the merging between a subclump around M87 and another clump around M86 (or rather, the infall of the M86 subcluster into the more massive M87 subcluster, see below). Both giant galaxies must be the centers of huge swarms of dwarf galaxies. M86 is apparently falling into, or through, the M87 subclump from the back, hence with a high relative (negative) velocity, dragging along its dwarf companions. (The giant galaxy M84, very close to M86 in the sky, but with $v = 1000$ km s^{-1}, has to be a member of the M87 subcluster.)

This whole picture is fully confirmed by an analysis of the x-ray structure of the Virgo cluster. Figure 5 shows, as a gray-scale image, the distribution of the x-ray intensity in the Virgo cluster as measured by the ROSAT x-ray satellite. The x-rays originate from the hot intracluster gas via thermal bremsstrahlung. The gas feels the same gravitational potential of the cluster as the galaxies. The global appearances of the cluster in the x-rays and in the optical (figure 5 versus figure 4) are therefore very similar. However, the main subclusters around M87, M86 and M49

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

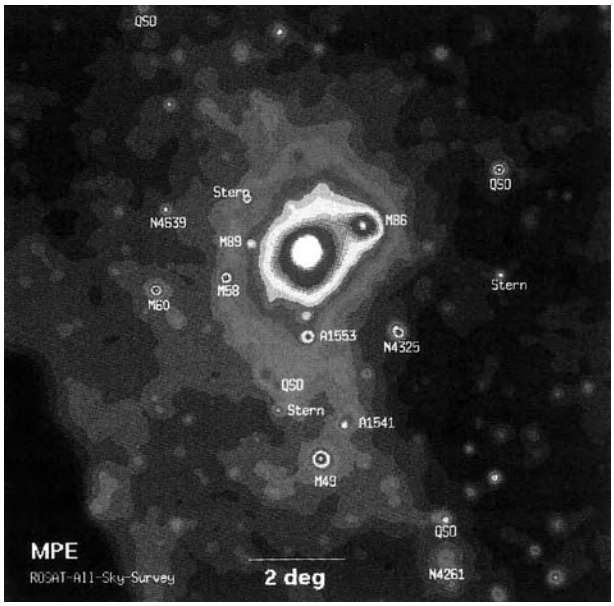


Figure 5. X-ray image of the Virgo cluster from the ROSAT All-Sky Survey. Various foreground stars, quasars, Abell clusters and Virgo cluster members (NGCs and Messiers) are indicated. The large, bright spot is centered on M87. Compare with figure 4. Courtesy of Dr H Böhringer and the MPE Garching.

are popping out much more clearly in the x-ray image. M87 has a huge x-ray halo and is obviously the center of the dominating subcluster in Virgo. M86 has a large gas halo as well—larger than expected for an isolated giant, but of the right order for a whole subcluster. The infall of this subcluster from the back is supported by the observation of a ram pressure plume in the x-ray structure of M86.

Subcluster merging has turned out to be a general feature of cluster of galaxies, even with so-called 'regular' ones, such as the COMA CLUSTER. This is not unexpected in a bottom-up (clustering) picture of structure formation, as described by the much favored 'cold dark matter' scenario.

Under the (tested) assumption of hydrostatic equilibrium, the x-ray halo profiles can be used to calculate the gravitational masses of the subclusters that are binding the gas. For the dominating M87 subcluster the integrated mass out to 1 Mpc from M87 amounts to $M \simeq 1.4 \times 10^{14} M_{\odot}$, of which 14% is in the gas, and only 4% in the galaxies. This means that the cluster is essentially dark matter, second comes the intracluster gas and only in third place the luminous matter in galaxies, with a correspondingly high mass-to-light ratio of $M/L \approx 500$. This is quite typical for clusters of galaxies.

Distance and depth

The mean distance of the Virgo cluster is still a matter of debate. Distances quoted range from 14 to 22 Mpc. In general, the most reliable extragalactic distance indicators are the Cepheids. Cepheids at (and slightly beyond) the Virgo cluster distance are now within the reach of the HUBBLE SPACE TELESCOPE (HST). This achievement was so long awaited that the first Cepheid-based distance determination of a Virgo cluster spiral in 1994 (M100 at 17 Mpc) had an enormous impact. The caveat with this was, and is, that spirals tend to avoid the cluster core and may be in the field far off the cluster (cf also above). It has been conjectured that M100, as well as other spirals located by Cepheids with HST later on, might lie at the near side of the cluster.

On the other hand, from work based on the TULLY–FISHER (TF) RELATION, which allows the distance to an individual spiral galaxy to be given with an accuracy of $\approx$0.4 mag, there is consistent evidence that Virgo late types are distributed in a prolate cloud, or filament, stretching—nearly along our line of sight—from the cluster backwards to the so-called 'W cloud' at twice Virgo's distance. Probably this is part of a very long filament that is running way back to the 'Great Wall' at the distance of the Coma cluster. On the near side of Virgo it might even be connected with the 'Coma–Sculptor cloud' that is running through us, i.e. includes the Local Group.

There is a hint, again from TF distances, that the southern M49 subcluster is lying significantly in the back of the M87 subcluster. If so, the M49 subcluster must be infalling from the back with a velocity of several 100 km s^{-1}, as the mean observed velocities of the two subclusters are very similar. The merging of subclusters along the large-scale filament in which they are embedded is plausible.

To determine the distance of the core of the Virgo cluster one should avoid late-type galaxies. The safest would be to use only elliptical and dwarf elliptical members. Unfortunately, the primary distance indicators here, RR LYRAE STARS, are much too faint at the distance of Virgo even for HST. The secondary distance indicators which can be applied to Virgo ellipticals give controversial results: globular clusters, D_n–σ and novae tend to give large distances ($D \approx 20$ Mpc), SBFs and PNe lead to a small $D \approx 16$ Mpc. Great efforts are spent in the application of the SBF method because its claimed distance uncertainty for an individual galaxy is almost as small as with Cepheids ($\leq$0.2 mag) and hence would allow us to resolve the cluster depth (the front-to-back depth is about 2 Mpc, or 0.2 mag, if the cluster is spherical). Individual Virgo E distances, with a surprisingly large scatter, have indeed been reported, but there is still some concern whether all variations of the stellar content of ellipticals, on which the method critically depends, are sufficiently well understood. For instance, the SBF distance of M49 is much smaller than the (probably more reliable) TF distance of the whole M49 subcluster, which means that either M49 is lying in the foreground, being projected on top of the background 'M49' subcluster by chance, or the SBF method is wrong.

Some hope for the future is resting with dwarf ellipticals, of which there is an almost inexhaustible reservoir in the cluster. A first, very promising

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

application of the tip-of-the-red-giant-branch (TRGB) distance indicator to a particular dE based on HST observations has given $D \simeq 16$ Mpc. A recent claim that dEs are distributed in a prolate structure pointing towards us, based on the shape of the luminosity profile of these galaxies, will soon be tested by an extension of the SBF method to dEs.

A new tool to determine the mean distance and depth of the Virgo cluster core may be provided by the recently discovered population of free-floating giant stars and PNe. Such a population has long been suspected in clusters of galaxies, as stars will be ripped from galaxies by tidal encounters in the cluster core, a process called 'galaxy harassment'. The data available so far suggest that approximately 10% of the stellar mass of the Virgo cluster is in intergalactic stars. There are ongoing programs with large telescopes to measure the radial velocities and metallicities of a large number of intracluster PNe. These data will provide crucial constraints on the dynamical state and history of the Virgo cluster.

Bibliography

Binggeli B 1999 The Virgo cluster—home of M87 *The Radiogalaxy M87* ed H-J Röser and K Meisenheimer (Berlin: Springer) p 9

Binggeli B, Tammann G A and Sandage A 1987 Studies of the Virgo cluster. VI. Morphology and kinematics of the Virgo cluster *Astron. J.* **94** 251

Bruno Binggeli and John Huchra

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Virgo Gravitational-wave Interferometer (VIRGO)

VIRGO is a collaboration between Italian and French research teams to construct an interferometric gravitational wave detector at Cascina, near Pisa, Italy. It relies on a technique called laser interferometry to measure with immense accuracy the minute changes in distance induced by gravitational waves from astronomical sources.

The VIRGO project plans to set up a laser interferometer made of two, 3 km long, orthogonal arms. Light from a laser is split into two beams which travel down the arms and are then reflected back by mirrors. In effect, multiple reflections will increase the optical length of each arm to 120 km. Variations in the pattern of interference between the returning beams will enable gravitational waves to be detected and their sources to be identified.

VIRGO will be sensitive to gravitational waves at frequencies from 10 to 6000 Hz. It should be able to detect radiation produced by supernovae and by the coalescence of binary systems in the Milky Way and other galaxies, for instance from the Virgo cluster.

It will use high-power ultrastable lasers and high-reflectivity mirrors, as well as seismic isolation and control of position and alignment. Each optical component is isolated by a 10 m high elaborate system of compound pendulums, called a 'superattenuator'. The 6 km long, 1.2 m diameter evacuated tube through which the light beam passes will be one of the largest vacuum vessels in the world. The signals will be detected, registered and elaborated by computer.

For further information see
http://www.virgo.infn.it/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Visual Binary Stars

Already since the early development of telescopes, an increasing number of stars—single to the naked eye—were found to consist of two or more components. They became known as 'double stars' (étoiles doubles, dvoinye zvozdy, Doppelsterne) and soon were too numerous to be explained as random pairings of stars in the sky. With increasing observing-time span, some pairs showed around 1800 that the components orbited one another. The English term 'BINARY STARS' tends to emphasize this gravitational connection. In distinction from close doubles recognized by stellar eclipses or by periodic changes of spectral-line shifts (photometric and spectroscopic binaries), the directly resolved objects are termed 'visual' although eyepiece observations are no longer the only method of study. Telescopes and techniques of the 20th century have improved the angular resolution to below 0.1 arcsec, but in order that components be separated, that is be recognised by their different positions in the sky, visual pairs are subject to at least one of two limitations.

(a) They are not very distant from Earth. Visual objects thus are predominantly of the more populous types in the solar neighborhood: stars from the middle and lower main sequence;
(b) The true distance between the components is not too small. The orbit periods are therefore generally much longer than in ECLIPSING and SPECTROSCOPIC BINARY STARS, and so it is important that measurements dating back 100 yr and more are available for many of them.

In addition, detection is impeded when the stars have a large brightness difference. Hence not many lowest-luminosity dwarf components, or companions of bright supergiants, are known.

Designations of visual double stars

The celestial position for J2000 (the beginning of the Julian year 2000) is generally used, for instance 07346 + 3153 means right ascension 7h 34.6m and declination 31° 53′ N (Castor). Cataloguing confusion can occur when the positions of faint stars are not well enough determined or when the transfer to J2000 ignored proper motions.

Also used are discoverers' codes, e.g. Bu 1077 for α Ursae Majoris (Dubhe) after the finding by S W BURNHAM. These letter codes are unambiguous but owing to their variety a bit cumbrous. The master file of visual double stars, maintained at the US Naval Observatory, Washington DC, and currently being revised, lists both these identifiers for nearly 100 000 entries, and in addition the Durchmusterung number (Bonn, Cordoba, Cape), if any.

The double-star catalogue by R G AITKEN contains objects discovered before 1927 in the sky north of −30° declination. Since it includes many well-known binaries, the ADS number is still used. Some publications cite the HIP number after the star catalogue from the HIPPARCOS satellite mission.

Some results of statistics

Adding all kinds of binaries, and with a statistical estimate of the occurrence of undetected, low-mass components, the fraction of stars on the upper and middle main sequence which are members of binary and multiple systems comes to about 80%. This membership rate decreases toward RED DWARF STARS, and for faint dM types—of which only the nearest are well studied—may be as low as 40%. Thus the most abundantly occurring stellar masses, around 0.2 of the solar mass, are less frequently found in binaries.

Very wide pairs have only a weak gravitational bond, which may be disrupted by fairly close random encounters with passing stars. As estimated from the star density in the Galactic disk, the limit of separations, below which a pair is expected to remain bound through its lifetime is at 10 000–20 000 AU (astronomical units), or 0.05–0.1 pc. Existence of orbit periods up to some millions of years is thus to be expected.

Inevitably some pairs of stars appear close together by chance alignment, although the components are at different distances and unrelated. These are 'optical' or 'perspectivic' double stars. Statistics show that the probability of a random alignment to within 5″ or so is quite low for stars down to about 12th magnitude. Most of the catalogued double stars thus are genuinely physical systems.

Observing techniques

Visual measurement by micrometer

The FILAR MICROMETER contains a set of thin wires, one movable against another with a precision screw, and mounted on a rotatable frame in the focus of a high-power eyepiece. This simple device supplies the position of one double-star component relative to the other in polar coordinates: the position angle (direction in the sky) counted from N over E, S, W from 0° to 360° by orientation of the frame, and the separation between stars in arcsec by the spacing of the wires. Since the eye has a good time resolution of 0.05 s, it can suppress much of the atmospheric turbulence, often reaches the diffraction-limited resolving power provided by the telescope aperture and may even recognise elongated double-star images below resolution. (The often-heard statement that atmospheric blurring limits ground-based instruments to about 1 arcsec of resolution is false; it applies only to long recording times.) Although a 17th-century invention, the filar micrometer became a high-precision tool only in the hands of F W STRUVE in 1820. Since then most double-star positions have been obtained that way, quite close binaries being reached already with the refractors around 1900 (Aitken and others), at least in the northern hemisphere. The southern sky lagged behind until W H van den Bos began his micrometric productivity in the 1920s, and R A Rossiter followed with the discovery of many thousands of fainter pairs.

A modification favored by some observers is the double-image micrometer which uses a birefringent prism.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The four star images produced from a double star are brought into a well-measurable configuration by rotating and shifting the prism.

Photography
During much of the 20th century, photography was the best direct-imaging method in astronomy, but its resolving power was limited by adjacency effects from the chemical processing and by blurring during the fairly long exposure times needed. Extensive double-star data were obtained for wider pairs, typically over 2 arcsec. The long-term photographic monitoring of stars has helped in other ways: with PARALLAX results for nearby binaries, their mass ratios from the orbital shifts of blended images against their centre of mass and sometimes the detection of unseen components from periodic shifts in the motions of stars. Such work could now be done with digitizing cameras (CCDs) which have replaced photographic emulsions in direct imaging.

Interferometry
Owing to high instrumental demands, the original form of interference was little used. A successful variant is the eyepiece interferometer as was constructed and widely used by W S Finsen.

In the last 25 yr a combined high resolution in both angular separation and recording time, now also for modest-size telescopes, has been achieved by 'speckle' interferometry (see SPECKLE IMAGING OF BINARY STARS). It results in superior precision by correlating numerous, very short-exposure images (speckles, tavelures). Most currently used devices consist of microscope optics as objective and collimator, a CCD camera with image intensifier as detector and computers for very fast recording, centering, autocorrelation and processing; colour filters and usually a pair of prisms, to remove atmospheric dispersion in both amount and direction, are added. The full theoretical resolving power at large reflectors, around 0.03 arcsec, can be reached. An important development toward yet higher resolution is the application of long-baseline interferometry (long used in radio astronomy) now at optical wavelengths, as evidenced by results from devices such as the Mark III interferometer.

Space instruments
The Hipparcos satellite measured around 1991 numerous double stars and added some 3000 new pairs, mostly those of larger brightness differences whose detection from the ground was impeded.

Orbits

Apart from their significance for the understanding of stellar origin and of evolutionary patterns, binary stars supply direct determinations of the masses of stars as fundamental quantities. This requires knowledge of their orbits.

Main-sequence stars have a relation between mass M and luminosity L, as expected from theory, because for stars of nearly identical composition the equilibrium conditions governing interior stability make the luminosities and radii dependent only on the mass: L increases with the kth power of M. The exponent k varies along the main sequence, and it also depends on the colours within which the luminosities are defined; for visual magnitudes, $k = 3.8$ for the middle main sequence, but k decreases to about 2.6 for both high- and low-mass (B and M) types.

The apparent orbit of a visual binary as seen in the sky (at right angles to the line of sight) is a parallel projection of the true orbit ellipse. Seven constants, the so-called orbital elements, are needed to describe the motion. Three of them are angles specifying the projection: the inclination is the angle between orbit and projection planes, the node is the celestial direction in which the planes intersect and the longitude of periastron (the point of closest approach of the components, and of fastest orbital motion), counted from the node, gives the orientation of the ellipse. The other elements fix the motion within the orbit: the period (P) in years, the eccentricity, the time of passage through periastron, and the semimajor axis (a) in arcsec. The position in the orbit at any time is then given by formulae which represent Kepler's first and second laws (see KEPLER'S LAWS). There are several computer routines on how to solve for the seven elements as best fitting a given set of observations.

Visual binaries with known orbits (some very precise, others tentative and subject to future revision) now number about 1000. The orbit sizes (a) are mostly under 1 arcsec; the periods range from a few years to some centuries (in cases with well-observed periastron passages up to about 1000 yr), and almost all objects are within 150 pc from Earth. Contrary to the planetary system and to close binaries, orbits of medium or high eccentricities dominate among visual pairs.

To determine the total mass ($M_1 + M_2$), Kepler's third law is used in the form $a^3 = P^2(M_1 + M_2)$ only if a is determined in astronomical units:

$$a \text{ (length in AU)} = a'' \text{ (angular size in arcsec)} \times d \text{ (distance in pc)}.$$

Inaccuracies of distances d cause often the largest uncertainty in the masses. This can be remedied in other ways.

(a) By spectroscopic observations of the orbital motion. Doppler-shift amplitudes alone yield only the one-dimensional radial motion, leaving the orbit inclination and hence the true orbit size and the mass undetermined, but the combination with the visual elements supplies the missing link. The speckle technique in particular has resolved many short-period pairs discovered by spectroscopists.
(b) When the components can be assumed to be main-sequence stars, a mass–luminosity relation can be invoked, along with the known apparent magnitudes, to connect the masses and the distance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Cases where this assumption does not hold are revealed by a disagreement between spectral type and the computed mass. Many stars spectrally still classified as main sequence are found to be already evolved in luminosity to above 'zero age' level.

Multiple systems

About a third of binary stars contain at least one additional component. However, the separations are almost always quite different: the ratio of wider versus closer pair is at least 6:1 and often much more. Long-term stability of multiple systems with stellar (not very unequal) masses demands this 'hierarchical' order. So the periods are very different, and there are few visual systems in which both orbits are reliably known.

More common are close spectroscopic subsystems within visual pairs, and there are also cases of definite, periodic deviations from the elliptical motion indicating the presence of an unresolved component. These objects require elaborate analysis since two orbits (14 elements instead of seven) have to be computed simultaneously. Slow changes in the orbits owing to perturbations between components are known in two of them: ξ Ursae Majoris (11182 + 3132) and ζ Cancri (08122 + 1739).

Some noteworthy objects

Within 3 parsec from the Sun, four apparently single dwarf stars and the following three interesting binaries are found.

α Centauri (14396 − 6050), the nearest system, consists of two solar-type stars (masses 1.1 and 0.9) in an orbit of period 79.9 yr, now slowly closing in after maximum separation was reached in 1980, and a red-dwarf companion, a full 13 000 AU distant from the primary pair, but even closer to Earth, hence called Proxima Centauri.

Sirius (06451 − 1643) has a companion with a period of 50.1 yr, which was predicted from periodic changes in the proper motion of the bright star. Resolved in 1862, the companion was found to be almost 10 magnitudes fainter, and the identical white color indicated an enormous disparity of the radii of the stars. It was the first identification of a white dwarf star. (40Eridani B = 04153 − 0739 had been observed long before but its 252 yr orbit with a faint red-dwarf companion and its white-dwarf nature were not known until later.)

LDS 838 (01390 − 1757), a 12th magnitude object with a period of 26.5 yr is one of the lowest-luminosity pairs of red dwarfs. Each component has only 0.12 solar masses; one of them is the very active flare star UV Ceti.

A few other remarkable binaries and multiples may be added.

Castor (α Geminorum, 07346 + 3153) is a sixtuple at a distance of 14 pc. The bright pair has been observed for almost three centuries, and the latest orbit gives it a period of 467 yr. Both components are close spectroscopic binaries with low-mass companions (periods of 9 and 3 days), and the distant 9th magnitude star YY Gem, which belongs to the system, is a unique case of a red-dwarf eclipsing binary (type dM) with a period of 19.5 h.

Capella (α Aurigae, 05167 + 4600), a long-known spectroscopic binary with $P = 104$ days, and at a distance of 13 pc, was resolved and well observed by interferometers and speckle cameras. It consists of two G type giants of nearly equal masses (2.5 solar masses each) and luminosities, evolving synchronously. Also part of the system is a later-discovered, distant red-dwarf pair which may have a period around 400 yr.

The closest pair resolved by a large interferometer is Spica (α Virginis, 13252 − 1110) at a separation of only $0''.0015$, with a period of 4 days.

ε Hydrae (08468 + 0625), at a distance of 36 pc, is of interest because of the diversity of its five known components: a G type and a fainter F type subgiant in a 15 yr orbit, an F type main-sequence star (period about 1100 yr around the primary pair) which is a spectroscopic binary with a low-mass companion ($P = 10$ days), and a still more distant, subluminous dK dwarf star whose period may be around 10 000 yr.

Finally, GC 20393/4 (15102 − 1623) is a faint, wide pair of very old subdwarf stars, separated by 5 arcmin, and 30 pc distant. They race through the solar vicinity at a record speed of nearly 600 km s^{-1} relative to the Sun, or practically at escape velocity within the Milky Way, suggesting an origin in the outer galactic halo. It should be noted that, although separated by at least 10 000 AU, they have kept together on the long trip for twice the age of the solar system.

W D Heintz

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

VLBI (Very Long Baseline Interferometry)

The Michelson stellar INTERFEROMETER was developed by MICHELSON as an optical instrument to measure the angular diameter of stars. In recent times, the method has been extensively used at radio wavelengths, to achieve extremely high angular resolution, greater than that achieved by single-aperture telescopes at any wavelength. The principle is as follows: two RADIO TELESCOPES, receiving a signal from a distant radio source, amplify the signals and send them via transmission lines to a radio receiver. the radio signals arrive at different times at the two radio telescopes, so the shorter path length is lengthened by a delay line that equalizes the time at which the signals are combined by the receiver. If the signals are in phase, they will reinforce one another and the signal amplitude will be twice that of a single telescope (i.e. four times the power). If the signals are 180° out of phase, there will be no signal, since they destructively interfere. If the telescopes are rotating with the Earth, and the time delay is held fixed, the alternating constructive and destructive interference will give a sinusoidal output whose amplitude and phase can be measured. (The optical Michelson stellar interferometer is functionally identical; the radio telescopes are replaced by plane mirrors that send plane waves into a telescope, where they are combined at the focus, where interference fringes are detected.) By extension from the optical case, the sinusoidal radio signal is also called a fringe, even though it is seldom seen directly in modern interferometers. This is because the receiver has, in effect, two channels that cross-correlate the two signals, directly and in quadrature. From the amplitudes of these two outputs, the fringe amplitude and phase are derived.

The great power of the Michelson interferometer lies in its extension to APERTURE SYNTHESIS, in which the signals from an assemblage of many telescopes are combined pairwise. As in so many areas of modern science, the Fourier transform technique is used to derive an image. Given a radio source with a brightness distribution $B(\alpha, \delta)$ on the sky, where α and δ are right ascension and declination, one constructs an approximate Cartesian coordinate system x, y. The distribution $B(x, y)$ has a Fourier dual $b(u, v)$, and it can be shown that the amplitude and phase of the radio fringe from a Michelson interferometer is proportional to the Fourier transform of the source brightness distribution. The dual coordinates (u, v) are determined by the baseline $\boldsymbol{b}_\lambda$, measured in terms of the observing wavelength λ, projected on the sky perpendicular to the line of sight to the radio source. This plane is usually taken with u in the east–west direction, and with v north–south, and is commonly referred to as the uv-plane. A sufficiently large sample of fringe amplitudes and phases, with interferometers sampling many points on the uv-plane, can give a good approximation to the Fourier transform $b(u, v)$ of the brightness distribution, and by numerically calculating the Fourier inversion, the actual brightness distribution $B(x, y)$ can be constructed.

It follows that an array of N radio telescopes, suitably distributed on the Earth, will yield $N(N-1)/2$ pairs, thus allowing one to map radio sources with an instrument having an effective diameter, and a corresponding angular resolution, of a telescope of diameter $D_{\max}$, the length of the maximum interferometer spacing. Such an instrument is called an aperture-synthesis array; the density of Fourier transform samples can be increased by observing as the Earth rotates, thus changing the projected baseline length on the uv-plane. In this case, the technique is known as Earth-rotation aperture synthesis. The VERY LARGE ARRAY (commonly known as the VLA), operated by the National Radio Astronomy Observatory on the Plains of St Augustine, near Socorro, New Mexico, with 27 antennas in a variable-length, Y-shaped array, is a good example of an aperture-synthesis instrument.

Techniques of VLBI

The technique of Very Long Baseline Interferometry, or VLBI, is an extension of Michelson interferometry, made possible by the exquisite control that modern electronics exerts over the time domain. If the transmission lines from the telescopes to the receiver are removed, it is still possible to observe the interference fringes by having a highly stable clock (usually a hydrogen maser frequency standard) at each telescope. The amplified signals are recorded on wide-band magnetic tape recorders (typically at a data rate of the order of 10^8 bits per second or more), using the hydrogen standards for time control. The tapes are then shipped to a central processing facility, where the received signal is now a string of discrete samples $e(t_i)$ at known times t_i. The two data streams are then cross-correlated numerically, giving the fringe amplitude and phase. (For a general treatment of interferometry, including VLBI, see Burke and Graham-Smith (1997).)

Several VLBI arrays are in operation: the VLB Array (VLBA) of the National Radio Astronomy Observatory with its processor in Socorro NM, the European VLBI Network (EVN), with a processor in Dwingeloo, Netherlands (see JOINT INSTITUTE FOR VLBI IN EUROPE), and the Southern Hemisphere Array, with its processor in Narrabri, NSW, Australia. There is also a worldwide network, a collaboration among the above VLBI arrays with the addition of other national radio telescopes.

The VLBA can be taken as an example of a dedicated VLBI system. It consists of ten similar radio telescopes spread across the USA and its outlying territories; the array is bounded by telescopes in St Croix, US Virgin Islands, on the southeast, Hancock NH on the northeast, Brewster WA on the northwest, and Mauna Kea HA on the southwest, with a maximum baseline of 10 000 km. At the present time its shortest operating wavelength is 7 mm, giving a maximum angular resolution of about 150 microarcseconds. The operating wavelength can be switched rapidly from one band to another, and the observing schedule, sent out to the individual telescopes

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

some time in advance, includes the time and duration of observation the source to be observed, and the frequency band.

There are some practical considerations that arise when making VLBI observations. The masers at the individual telescopes will have slightly different frequencies, and therefore the time base on each magnetic tape will have systematic errors, showing up as time differences and differences in fringe rotation. The VLBA correlator, therefore, performs the cross-correlation for a number of different time delays and fringe rotation rates, in order to correct for these differences. In the case of a 'hard-wired' interferometer such as the VLA, timing variations occur, and these are usually corrected by observing a nearby calibration source and deriving the necessary corrections. For VLBI observations, this is sometimes possible, but in many cases it is not feasible to do so. For an array of several telescopes, however, there are closure relations that allow extra information to be derived. If three telescopes are observing the same source, there is a sum condition on the three fringe phases that gives useful phase information, and there is a condition on the fringe amplitudes, when observing with four telescopes, that gives useful constraints on the individual amplitudes. When an array of ten telescopes is used, as in the case of the VLBA, there is enough redundancy in the various combinations to constrain the uncertainties in the Fourier transform, giving maps of remarkable fidelity.

Science applications of VLBI

There is a surprising variety of scientific problems that can be addressed by the methods of VLBI. Quasars and the nuclei of active galaxies (BL Lac sources, radio galaxies and Seyfert galaxies) all have structures in the milliarcsecond range or smaller. Pulsars and gamma-ray burst sources are even smaller, requiring baseline lengths far longer than anything currently feasible, Molecular maser sources are also of milliarcsecond size, and these occur in association with star-forming regions and cool, evolved stars such as the Mira variables. Three general classes of observation can be identified: the mapping of milliarcsecond structures, the measurement of relative source positions to achieve higher astrometric accuracy, and the use of known sources as references to study geophysical phenomena such as continental drift and polar wandering.

Active galactic nuclei

The detailed study of ACTIVE GALACTIC NUCLEI (AGNs) has shown that the jet-like structures seen on large scales persist at the smallest angular scales that have been studied. Paradoxically, the (relatively) long radio wavelengths are closely related to phenomena at x-ray wavelengths. The radio phenomena observed by VLBI are in large measure relativistic, since the radio waves are generated by synchrotron radiation from relativistic electrons circulating in magnetic fields. A large fraction of the sources involve massive black holes, surrounded by accretion disks (the hot gas in these accretion disks, observed at ultraviolet and x-ray wavelengths, is in the process of spiralling down into the black hole at the center). The exact process is still unclear, but seems to involve magnetic fields in the accretion disks, twisted by differential rotation in the disks, ultimately releasing enormous amounts of energy by accelerating electrons to relativistic energies and confining the particles to a narrow jet. Shortly after VLBI was developed, it became clear that rapid changes were developing in the structure of the jets, and it is now well-established that the jets exhibit apparent motions that significantly exceed the speed of light, sometimes by an order of magnitude. Although the structure of fundamental physics might be questioned, such a dramatic revision as superluminal velocity does not seem to be needed. The effect is almost certainly caused by the aberration effects observed when a relativistic beam is directed at the observer. The effect requires a beam of particles having bulk motion close to the speed of light, directed within a few degrees of the observer.

A relativistic effect that must be considered in conjunction with these compact, relativistic sources of SYNCHROTRON RADIATION is the self-Comptonization of the radiation. As the sources increase in energy density, there can be so many radio photons generated that they are Compton-scattered by the electrons and become x-ray photons, which can be observed by X-RAY TELESCOPES. This has the effect of limiting the possible surface brightness at radio wavelengths; expressed in terms of 'brightness temperature' (the temperature a source would need to have at the observing band to give the observed specific intensity, not a physical temperature necessarily). As a consequence, it is not possible for a synchrotron source to exhibit a brightness temperature in excess of 10^{12} K. Unfortunately, it is not possible to verify this prediction by interferometer baselines on Earth. The flux of a RADIO SOURCE is proportional to the product of its average surface brightness and its angular size, and the wavelength dependences cancel out. At any wavelength, therefore, an interferometer with a baseline length of an Earth diameter, observing even the strongest radio source, can only set a limit of 10^{12} K on the surface brightness of the compact structures.

This limitation to our knowledge can be overcome by using interferometer baselines that exceed the diameter of the Earth. By placing a VLBI telescope in space, there is in principle no limit on the angular resolution that might be achieved. It has turned out that the technical difficulties can be overcome; a successful demonstration was carried out, using the TDRSS (a data relay satellite system), that showed the practical feasibility of such measurements. This was followed up by HALCA, a dedicated VLBI satellite launched by the Japanese Space Institute ISAS in February 1997. The results are clear: a large fraction of quasars and AGNs have radio nuclei that exceed the self-Compton limit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Maser sources

The extragalactic maser sources have been another major object of VLBI research. The lambda-doublet lines of the hydroxyl radical at 1612, 1665, 1667 and 1720 MHz all exhibit high surface brightness (10^{10} K or more) that can only be a consequence of gas clouds being pumped by radiation or by collisions to invert the level populations, resulting in a one-pass maser that can amplify the background radiation. The same phenomenon occurs in the rotational lines of water, silicon oxide and methanol that are accessible in the radio spectrum. Star-forming regions frequently have many such maser sources, as do a wide variety of evolved stars. In such cases, the difficulty in determining the fringe phase can be overcome by self-referencing: a particular compact line feature can serve as a reference within the maser complex, and the positions of other line sources, often separated in Doppler shift by several km s^{-1}, can be measured with respect to the reference maser.

Megamasers

Extragalactic maser sources have also been discovered. These are referred to as 'megamasers' since the necessary energies are so high. One of the best studied of these, the megamaser in the active galaxy NGC 4258, appears to be emitting 300 solar luminosities in the 22 GHz water line. The line structure turned out to be extraordinary: there was a complex of strong central lines, having very little velocity relative to that of the galaxy, but in addition there were satellite maser lines with relative Doppler velocities $\pm$900 km s^{-1} to either side of the central complex. A detailed analysis of the VLBI-derived map showed that the velocities could be accurately explained by a simple model. A ring of gas surrounds the (presumed) black hole at the center, and we see the ring nearly edge-on. The central line complex changes linearly in Doppler velocity across the source, and is amplifying the radiation from the central source; this allows the acceleration of the molecular gas to be measured. The high-velocity lines come from masers at the terminus of the edge-on ring, and these allow the Keplerian velocity to be determined. The result is firm evidence of a central black hole having a mass of 70 million solar masses (see Miyoshi *et al* 1975, Herrnstein *et al* 1999).

Bibliography

Burke and Graham-Smith 1997 *An Introduction to Radio Astronomy* (Cambridge: Cambridge University Press)

Herrnstein *et al* 1999 A geometric distance to the galaxy NGC 4258 from orbital motions in a nuclear gas disk *Nature* **400** 539

Miyoshi *et al* 1975 Evidence for a black hole from high rotation velocities in a sub-parsec region of NGC 4258 *Nature* **373** 127

Bernard F Burke

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vogel, Hermann Carl (1841–1907)

German astronomer, became director of the Potsdam Astrophysical Observatory, where he made spectroscopic analyses of stars, planets, comets and the Sun. He used photography to measure Doppler shifts of stars to determine their radial velocities, discovering spectroscopic binaries and, with Julius Scheiner, determining stellar diameters and masses.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Voids

Galaxies have a strong tendency to cluster. They have been drawn together by mutual gravitational attraction. Seventy per cent of galaxies are found in groups or clusters that are probably gravitationally bound. In turn, the groups of galaxies lie along filaments that interconnect. The largest clusters of galaxies tend to be found at the intersections of several filaments. Even the 30% of galaxies that are not in bound groups lie, almost entirely, within the filaments.

The filaments occupy only a small fraction of space. Most of space has no observable matter. These empty regions are the '*voids*'.

The discovery of voids

Although the general tendency for galaxies to cluster was known for a long time, the extreme segregation of galaxies into a small fraction of the available volume was not appreciated until recently. With information about the distribution of objects on the plane of the sky augmented by only crude information about distances from galaxy sizes and magnitudes, there is too much confusion from projection overlay to get a good sense of structure. It is necessary to have the REDSHIFTS of galaxies in order to place them in distance and, hence, to be able to construct a three-dimensional map of the distribution of galaxies (see DISTRIBUTION OF GALAXIES, CLUSTERS, AND SUPERCLUSTERS). Only in 1978, with galaxy redshifts in hand, did Stephen Gregory and Laird Thompson comment that there were regions nearly devoid of galaxies to the foreground of a supercluster that extends from the Coma cluster to the cluster Abell 1367. Several other empty regions were found within the next few years and, in 1981, the instantly famous Boötes void was reported by Robert Kirshner, August Oemler, Paul Schechter and Stephen Shectman. This void extends over roughly a million cubic megaparsecs (30 million cubic light-years).

Over the years since, there have been searches to see whether there is anything to be found in the Boötes void. A few small galaxies have turned up. Still, it is an empty place. The Boötes void may not even be the biggest empty region within the currently accessible universe. It is distinctive in part because it is surrounded by the prominent supercluster domains of Hercules–Corona Borealis, Leo and Ursa Major. There are even larger regions of substantial underdensity. The larger regions are hard to characterize because they extend so far across the sky that they become lost in the obscuration of our Milky Way Galaxy and they extend so far in distance that redshift surveys become incomplete.

Nearby voids

There is no need to go far to find a void. The Milky Way Galaxy is part of the LOCAL GROUP which in turn is part of the Coma–Sculptor cloud, a filament that lies in the equatorial plane of the Local Supercluster. In the plane of this filament there are galaxies in all directions and to the south of this plane in supergalactic coordinates there is anther filament. However, to the north of this plane, just beyond the last galaxies of the Local Group, is the Local Void. This empty region is poorly surveyed because it is intersected by the Milky Way. However, there are few nearby galaxies on either side of the Galactic plane over a wide segment of the sky.

There are several voids that are somewhat farther away but better studied because they cover a restricted part of the sky and stand out with good contrast against pronounced filaments. The Perseus–Pisces filament is particularly well defined, at a distance of 60 Mpc (200 million light-years). Immediately to the foreground is a large region devoid of galaxies.

The topology of voids

The complement of voids are the superclusters and figure 1 in SUPERCLUSTERS AND THE LOCAL SUPERCLUSTER illustrates both the filamentary nature of the distribution of galaxies and the voids. The filament–void network can be compared with a sponge. The material of the sponge can be equated with the filaments and the sponge holes with the voids. The sponge material is all interconnected, else the sponge would fall apart, but also the holes are all interconnected, so water can pass through. However, with this analogy, the universe is a very tenuous sponge with frail material connections and big holes.

The universe must have begun with only minute density variations. Over time, matter has gathered through self-attraction in the high-density regions and evacuated the low-density regions. Hence, even in co-moving coordinates the domain of the voids has been growing. In metric coordinates, the voids are much bigger now than in the past. It follows that matter is moving away from voids, toward high density concentrations.

There is still debate about how much residual matter there is in voids. Galaxy formation may only occur above a certain density threshold which leaves open the possibility that there could be a considerable amount of total mass tenuously spread through the voids. Stringent limits are set by failure to detect intergalactic gas over a wide range of temperatures. The only hint that there might be residual gas comes from studies of absorption lines in distant quasars. These absorption lines arise from tenuous intergalactic gas in the foreground of the quasars. However, at low redshifts (the present epoch) there are relatively few absorbing gas clouds and it remains to be demonstrated whether that gas is associated with the intersection of filaments with the line of sight or with voids.

Bibliography

Rood H J 1988 *Ann. Rev. Astron. Astrophys.* **26** 245

R Brent Tully

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Volans

(the Flying Fish; abbrev. Vol, gen. Volantis; area 141 sq. deg.) a southern constellation which lies between Mensa and Carina, and culminates at midnight in mid January. It was first shown on Petrus Plancius' celestial globe of *c*. 1598 as Piscis Volans, though it is usually attributed to the Dutch navigators Pieter Dirkszoon Keyser (also known as Petrus Theodorus) and Frederick de Houtman, who charted that part of the southern sky in 1595–7.

A small, inconspicuous constellation, the brightest stars in Volans are β Volantis, magnitude 3.8, γ Volantis, a binary with yellow (G8) and pale yellow (F2) components, magnitudes 3.8 and 5.7, separation 14.11″, ζ Volantis, magnitude 3.9, and δ Volantis and α Volantis, both magnitude 4.0. Another interesting star is ε Volantis, a double with bluish-white (B6 and B9.5) components, magnitudes 4.4 and 7.5, separation 6.0″, that have the same proper motion, the former of which has an unseen companion that revolves around it in 14.17 days.

There are no bright star clusters, nebulae or galaxies in the constellation, the brightest being NGC 2442, an eleventh magnitude barred spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Volcanism in the Solar System

Volcanism is the main endogenous process (a process originating within the planet) which gives PLANETARY SURFACES most of their characteristic features. Volcanism consists of the superficial outflow of liquid mixed with a variable quantity of gas and crystals, which comes from the PLANETARY INTERIOR. These liquids are produced by the melting of the planet's or satellite's interior. They are called 'lava' or 'magma' and come up to the surface owing to buoyancy forces. Water and hydrocarbon springs on Earth are not part of these magmatic liquids.

Volcanism not only modifies the structure and the morphology of the planetary surface. During and just after the formation of planetary bodies, accretional heating partially separated volatile from refractory elements, degassed the mantle and produced primitive atmosphere. Volcanism causes this degassing of the mantle to continue, and it is gradually modifying both the planetary interior and the atmosphere.

The morphological types of volcanic features strongly depend on the magma viscosity and gas content. Very-low-viscosity magma induces flat lava flows and may produce lava plains in the case of abundant flooding. An increase in the viscosity produces shield volcanoes with shallow sloping flanks (inclined at about 10º) for intermediate-viscosity lava or domes with steep peripheral flanks (inclined at about 25º) for high-viscosity lava. A significant gas content fragments the magma during volcanic explosions and produces fragments of rock called 'tephra'. During low-energy explosions, the tephra again fall around the outflow points, forming scoria or cinder cones. High-energy explosions cover large areas with ash or pyroclastic deposits expelled aerially from a vent. Some volcanoes may produce continually the same magma with the same gas content that produces monogenic volcanic construction (built up by one kind of eruption). Elsewhere, the gas content and the magma viscosity may change during a single eruption or throughout the volcano's history, inducing a complex polygenic volcanic construction (built up by several kinds of eruption).

Volcanism timing allows the distinction between three kinds of worlds: (1) active worlds, with present evidence of endogenic activities, (2) recently active worlds on which active resurfacing processes have occurred on a large scale after the end of the heavy bombardment, but which now seem dead, and (3) dead worlds, on which the majority of the surface pre-dates the end of heavy bombardment and where the geological activity has occurred only occasionally after the end of this bombardment.

Two types of solid planets and satellites exist in the solar system. There are silicated and rocky bodies in the inner solar system and icy bodies in the outer solar system. Io and Europa, two Galilean satellites of Jupiter, are silicated bodies, although they belong to the outer solar system. Volcanism exists on almost all of these bodies.

Volcanism in the inner solar system

Active worlds

The Earth (diameter 12 750 km), the basis for planetology. Volcanism on Earth has been studied considerably since Pliny the Younger died studying the famous eruption of Vesuvius in AD 79. Terrestrial volcanism is the reference for comparative planetary volcanism studies, and the definitions of rocks, eruptions and feature types were proposed on Earth.

The geographical distribution of volcanoes on Earth is one of the characteristics of our planet. About 95% of active volcanoes are grouped together and form volcanic lines. The first group is located on the islands or mountains near oceanic trenches around the Pacific and near Indonesia and the West Indies. This volcanism is mainly andesitic (SiO_2 55%), basaltic (SiO_2 50%) and rhyolitic (SiO_2 60%). The second group lies under oceans, on the top of oceanic ridges. This volcanism is basaltic (SiO_2 50%). The location of these two volcanic types is strongly correlated with the global dynamics of the EARTH'S INTERIOR (plate tectonics). Volcanism mainly occurs near plate boundaries. The other 5% of volcanoes are isolated volcanoes, spread over ocean floors or continents. These volcanoes are basaltic ones and are called 'hot spots'.

The popular notion of the Earth's interior is a world-wide layer of magma beneath a solid thin shell. Following this notion, volcanism is simply the arrival of magma through a hole or a fracture crossing the crust. This popular notion is totally false. The Earth is divided into three levels: the superficial crust 5–40 km thick, the mantle from the base of the crust to 2900 km and the core from 2900 km to the centre of the Earth. The Earth is entirely solid with the exception of the outer core, which is made of molten iron. This liquid outer core represents less than 15% of the Earth's volume, 2900 km below the surface, which is not directly involved in volcanism. The Earth's interior is solid, but hot, with a temperature of up to 5000 ºC. This heat is generated by natural radioactive decay. The solid state of the Earth despite this high temperature is due to the very high pressure created by the weight of overlying rocks.

Temperature differences between the cold superficial layers and the hot interior produce density differences and enormous buoyancy, circulatory and convective forces. Impelled by these convective forces, the solid mantle flows very slowly like ice in a glacier. These mantle motions induce superficial motion of the cold superficial layers, called continental drift or plate tectonics. Downward-flowing superficial oceanic rocks

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

introduce oceanic water deep into the mantle. This water lowers the melting point and produces a small proportion of molten material. Upward-flowing mantle reaches a level where the drop in pressure allows a small proportion to melt. If enough is formed, the formed magma rises away from the denser residue of solid infusible rocks, owing to buoyancy forces. On Earth the convective cells are more than 1000 km wide; the upper parts of these cells are the tectonic plates with only horizontal motion. Volcanism occurs at the plates' boundary where upwards or downwards motions exist and induce a partial melting of the mantle. Some other circulatory motions exist inside the mantle, with localized upward flows of solid mantle. These localized upward flows which induce localized drops in the pressure, partial melting and volcanoes, and are far from the plates' boundaries, are called 'hot spots'.

In 2001, a NASA mission to study Alaska's unique terrain provided scientists with their first detailed look at the changing topography of one of Earth's most active volcanic regions. Researchers at the Alaska Synthetic Aperture Radar Facility, Fairbanks, created a high-resolution digital elevation model of Umnak Island, home to the Okmok volcano. This model can be used to produce new, accurate geologic maps. The most recent topographic map of the region was made in 1957 from aerial photographs. Okmok has erupted four times since then, dramatically changing the landscape. The Alaska scientists used data gathered in October 2000 by the NASA/JPL Airsar instrument, a side-looking imaging radar system carried aboard a NASA DC-8. It collected the Alaska data as part of its PacRim 2000 Mission, which took the instrument to French Polynesia, American and Western Samoa, Fiji, New Zealand, Australia, New Guinea, Indonesia, Malaysia, Cambodia, Philippines, Taiwan, South Korea, Japan, Northern Marianas, Guam, Palau, Hawaii and Alaska. Researchers there combined numerous Airsar strips of data into a mosaic, fused it to Landsat imagery, checked its accuracy and generated a number of data products, including the mosaic of Umnak Island.

Venus (diameter 12 102 km) (figure 1). There is no doubt that volcanism has been responsible for the generation of the majority of the Venusian surface. The Venusian surface is affected by only few impact craters, indicating a geologically young age. Venusian plains exhibit long flow-fields, with long lava channels in some places. Many thousands of volcanoes exist on Venus, either on plains or in the highlands (see VENUS: SURFACE). The MAGELLAN pictures revealed 550 concentrations of small volcanoes <20 km in diameter, 270 intermediate volcanoes between 20 and 100 km in diameter with a great variety of morphologies, 150 large volcanoes in excess of 100 km in diameter, 80 caldera-like features and more than 100 steep-sided domes sometimes called 'pancakes'. Some strange volcanic features without equivalents on Earth or other planets exist on Venus. There are complex overlaps of bulges, faults, cinder cones and lava flows, and they are called '*coronae*', '*arachnoids*', '*novae*'…. The majority of these volcanic features are consistent with basaltic composition, a hypothesis confirmed by the findings of Soviet landers whose soil analyses are chemically similar to those of basalt (SiO_2 50%). The steep-sided domes may represent more silicic magma (SiO_2 60%), while lava channels may represent ultramafic lava flows (SiO_2 45%). Only a little evidence was found for extensive pyroclastic deposits.

The volcanic features show a broad global distribution, in contrast to the distribution of volcanoes along plate boundaries on Earth. Some highlands show an excess of volcanic features, while some lowlands show only lava floods.

There is no direct evidence in the Magellan data for current volcanic activity. Some atmospheric analyses have shown sulfur dioxide variations, which may be due to volcanic activity. However, the dynamics of the atmosphere may also explain these observed variations. Crater counts would indicate that the whole Venusian surface is younger than 0.8–0.5 Gyr. An equilibrium resurfacing model for explaining the impact craters' density implies a current volcanic lava flux of $0.5\ km^3\ yr^{-1}$, a value comparable with the terrestrial intraplate volcanism. Thus there is hardly any doubt that, as is the Earth, Venus is a geologically active planet despite the lack of active eruptive processes identified during the Magellan mission.

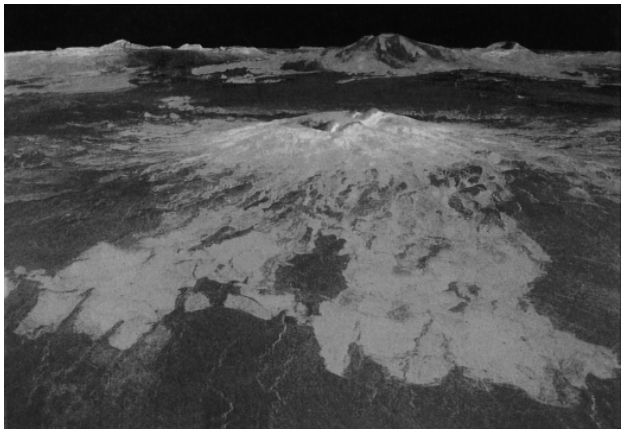


Figure 1. A volcano on a rocky body: Sapas Mons on Venus. Sapas Mons is 400 km across and 2 km high. It is a typical shield volcano. The sides of the volcano are covered with numerous overlapping lava flows which appear clear in this image. This volcano looks like Hawaiain volcanoes or large Martian volcanoes. No impact craters are visible on this picture, which indicates a very young age for the Venusian volcanism. This three-dimensional perspective view is produced with radar data from Magellan (NASA).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The density of Venus indicates that the Venusian interior is theoretically more or less similar to the Earth's interior. On Venus, mantle convection produced and still produces superficial motions and tectonic features, which are different from plate tectonics. The volcanism distribution confirms that the global dynamics of Venus's interior is quite different from that of the Earth. The localized 'hot-spot-like' convection strongly dominates the plate-tectonics-like convection. The origin of this difference is poorly understood.

Recently active worlds

Mars (diameter 6794 km). Volcanoes are unquestionably on Mars (see MARS: SURFACE). Four enormous shield volcanoes and many smaller ones rise 15–25 km above the surface. These volcanoes are mainly located on a 8000 km wide broad topographic bulge which stands about 10 km high. OLYMPUS MONS, the largest volcano in the known solar system, rises 26 km above the surrounding plateau. These volcanoes possess nested summit calderas and have numerous lobate lava flows on their flanks. Other smaller shield volcanoes exist in other places. Low-relief shields with irregular summit craters and numerous radial flows on their flanks are called 'paterae'. Some domes (called 'Tholus') and cinder cones have also been identified across Mars. Volcanic plains with multiple overlapping flow lobes are found around the periphery of the shield volcanoes. The complex flows can extend for >1000 km beyond the shields. Large plains covering wide areas in the northern hemisphere (but rare in the southern hemisphere) have controversial origin. The preferred origin is a volcanic one, based primarily on morphological similarity to lunar maria. No or only little evidence was found for extensive pyroclastic deposits.

In 1976 and 1997, respectively, the two VIKING landers and MARS PATHFINDER made *in situ* chemical analyses of soil and rocks. These analyses are consistent with volcanic rocks (mainly basalts, and one andesite-like analysis). Some meteorites (the SNC METEORITES) are supposed to have a Martian origin. These meteorites are mainly volcanic rocks (basalts and pyroxenites).

Volcanism existed on Mars from the earliest epoch, the lower Noachian, to the latest one, the upper Amazonian. The maximum of visible volcanic materials was produced during the Hesperian (from 3.5 to –2 Gyr). The younger volcanoes, Olympus and Arsia Mons, are estimated to be near 0.5±0.25 Gyr.

The Martian interior is theoretically more or less similar to the Earth's interior. On Mars, mantle convection produced superficial motions and tectonic features which are different from plate tectonics. The lack of active volcanoes since 500 million years ago indicates that the mantle convection now is either very low or has completely stopped. In the past, the mantle convection was active, and looked like hot-spot-like convection rather than plate tectonics convection.

Dead worlds

The Moon (diameter 3476 km figure 2). On the MOON, 84% of the surface which was called 'terrae' by GALILEO in the 17th century consists of old anorthositic cratered crust. However, about 16% of the surface consists of younger flat basaltic plains and flows which fill the old low-lying basins and form black patches on the Moon's face. These plains were called 'maria' by Galileo. The lava flows inside the plains are very long (up to 1000 km) with very low slopes (less than 1º), a factor that indicates a very low viscosity for the lava, with very low gas content. Many of these flows present sinuous rilles, which may be interpreted as collapsed lava tubes or erosional features caused by very hot lava flows. The thickness of lava may reach 10 km under Mare Imbrium. A very few cinder cones, shield volcanoes and volcanic domes rise up from the lava flows. The majority of maria are located on the near side.

This volcanism became evident during the waning stage of heavy bombardment which occurred before –3.9 Gyr. The volcanism was important from –3.9 to –3.5 Gyr, then decreased and virtually stopped 3 Gyr ago. The larger maria fill very old impact basins, but the superficial flows are some hundreds of millions of years younger than the impact itself. The maria are more abundant on the nearside than on the farside. Apollo samples indicate that lunar lava chemical composition is relatively similar to that of terrestrial basalts.

Figure 2. A lava plain on a rocky body: Mare Serenitatis (left) on the Moon. White mountains (on the right) are 'terrae'. No visible volcanic structures are seen inside this part of Mare Serenitatis, which only consists of very flat lava flows. Numerous small impact craters are visible on the mare, because the lava flows are 3.5 Gyr old (NASA–Apollo).

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

On the Moon, volcanoes ceased 3 Gyr ago. The dynamics of the lunar mantle during these very old epochs is poorly known. It is only known that this volcanism was not accompanied by significant tectonic motions, contrary to the Earth, Venus and Mars. Volcanism is more abundant on the nearside than on the farside, because the crust thickness is more significant on the farside, and the buoyancy forces seldom allowed the magma to reach the surface. There is a geographical relation between volcanism and large impact on the Moon, but genetic relationships are not clear because of the large chronological interval between large impacts and maria flooding.

Mercury (diameter 4878 km). On MERCURY, maria-like plains called 'smooth plains' are probably volcanic in origin. Unfortunately, the modest resolution of MARINER 10 images makes it difficult to clearly identify volcanic features. These smooth plains are mainly inside and around CALORIS BASIN, the major impact basin of Mercury. These smooth plains cover large areas of relatively ancient 'intercrater plains' which are possibly volcanic plains flooded during the heavy bombardment.

Crater chronology would indicate that the maximum of smooth plain flooding is geologically contemporaneous with the Caloris impact. The end of volcanism probably occurred 3.5 Gyr ago.

The dynamics of the Mercurian mantle during these very old epochs is poorly known. As on the Moon, it is only known that these periods of volcanism were not accompanied by significant tectonic motions, in contrast to the Earth, Venus and Mars. There is a geographical and a chronological relation between volcanism and large impacts on Mercury, suggesting genetic relationships such as a pressure fall in the mantle under the Caloris cavity.

The meteorites' parent bodies and the asteroid belt. About 8% of the catalogued METEORITES are volcanic rocks, mainly basalt or pyroxenite. There are five main kinds of these volcanic meteorites and they are called eucrite, howardite, aubrite, diogenite and urellite. The meteorites are suposed to be Earth-crossing fragments of asteroids. Thus, some ASTEROIDS with past volcanic activity exist in the asteroid belt. Some asteroids have a spectral signature of volcanic rocks. VESTA is supposed to be the diogenite and eucrite parent body, and Nysa is supposed to be the aubrite parent body. Radiochronological dating of volcanic meteorites indicates that the volcanism in the asteroid belt occurred only 10–100 million years after the formation of the solar system.

The relationship between the duration of volcanism and planetary size

After the end of ACCRETION, the heat inside a planet is mainly produced by radioactive decay. This heat production by massive bodies is supposed to be the same for all the silicated bodies and identical to the chondritic heat production value (about 10^{-11} W kg^{-1}). This heat production inside the planets' interior balances the heat lost to space. This balance fixes the interior temperature. If the loss is efficient with regard to the production, the temperature is low; if the loss is not efficient, the temperature is high. The heat is produced inside the planet and is related to its volume; the heat is lost by the surface. This balance is thus regulated by the volume/surface ratio, which is proportional to the planetary radius: a small planetary body with a low volume/surface ratio has efficient heat loss and a low internal temperature ; this is the case now for the Moon, Mercury and the asteroids. A large planetary body with a high volume/surface ratio has inefficient heat lost and a high internal temperature ; this is now the case for Venus and Earth. An intermediate planetary body has an intermediate volume/surface ratio and has an intermediate internal temperature; this is the case for Mars today.

At the beginning of their history, all the planetary bodies were hot because of the accretional heating and the short-lived radioactivity. After the end of this primeval heat production, the heat production was mainly due to long-lived radioactive elements (K, U, Th). When considering the solution to the spherically symmetric heat equation, the accumulation of heat from long-lived radionuclides produces a maximum in temperature after several hundred million years in large bodies. This has been confirmed by the extensive magmatic phase on the Moon which started 500 million years after its formation. After this maximum in temperature, because of the natural decay of radioactivity, there has been a secular cooling for all the planetary bodies. This cooling has induced a gradual stoppage of volcanism, with end times related to size and volume/surface ratio of the planets: this stoppage occurred very early for the small bodies (the meteorites' parent bodies), 1–1.5 Gyr after the accretion of Mercury and the Moon, 4 Gyr after the accretion of Mars and has not yet occurred for Venus and the Earth.

Volcanism in the outer solar system

In the outer solar system, where the temperature is 100 K, one planet (Pluto) and almost all large satellites have a low density ($\leq$2 g cm^{-3}) and water ice spectral reflectance. This indicates that water ice is a major component of these bodies, with a proportion larger than 50%. Other volatile compounds (N_2, NH_3…) may be mixed with water ice. These bodies are called icy bodies. Only two satellites of Jupiter, Io and Europa, have a density >3 g cm^{-3} and are mainly rocky bodies.

Io, Europa and tidal heating

Io (diameter 3630 km, figure 3) is one of the four Galilean satellites of Jupiter. Its density of 3.55 g cm^{-3}

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

indicates a global composition of silicates and iron. Io's surface displays a wide variety of colors and spectra which are dominated by sulfur compounds. Io's surface geology is dominated by volcanic vents, craters, fissures, flows and other morphological forms attributed to volcanic processes (see also IO: VOLCANISM AND GEOPHYSICS). Active plumes of volcanic material reaching heights of 300 km were seen by Voyager and GALILEO MISSION spacecraft in 1979 and 1997. These plumes are thought to be driven up by SO_2 or sulfur vapors. Sulfur compounds are certainly present on the surface, but, because of the low viscosity of solid sulfur, the heights of mountains, 9000 m, indicate a dominantly silicated composition for high mountains. Temperatures >1500 ºC measured by Galileo above active volcanoes prove the existence of molten silicates. It is not known whether low topography lava flows are composed of pure sulfur or sulfur rich silicates.

New large volcanic deposits and flows were seen by the Galileo spacecraft 18 yr after the Voyager fly-bys. No impact craters can be seen on Io's pictures, an observation which indicates that the volcanism is a very active resurfacing process. The global average height of Io's resurfacing is about 100 m $(10^6 \text{ yr})^{-1}$.

Io's volcanism supplies a very thin atmosphere (10^{-2}–10^{-7} Pa). Around Jupiter, a plasma torus is a result of material escaping from Io (see also IO: PLASMA TORUS).

Io is a small planetary body, with theoretically not enough radiogenic heating to be active now. However, Voyager revealed that Io is the most active solid world in the solar system. Another heat source other than radiogenic heating must be present for Io. The orbit of Io has a measurable forced eccentricity, which is particularly pronounced because of the phenomenon of orbital resonance with Europa and Ganymede. The result of this forced eccentricity is that Io's tidal bulges are continually flexed and moved with respect to Io's interior. The tidal bulge is about 7 km in height, and its periodic flexion is ±100 m. This is a significant source of friction and of tidally generated heat. In the case of Io, heat generated by tidal dissipation (about 10^{14} W, corresponding to 10^{-9} W kg^{-1}) is probably 100 times the rate of radiogenic heating. This is enough to melt at least some parts of Io's interior.

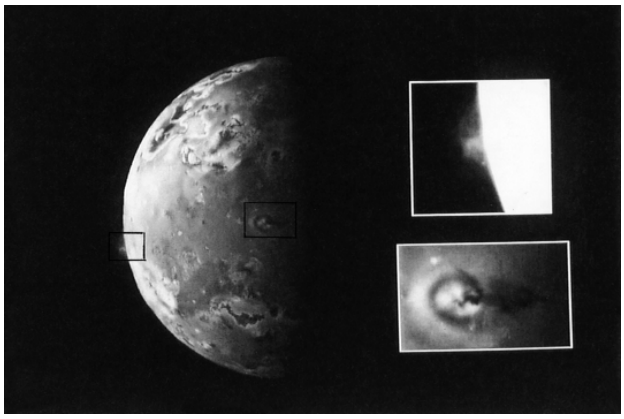


Figure 3. Global view of Io, showing numerous lava flows, and two active plumes: one on the limb (right) and the other on the terminator (left). These two plumes are about 200 km high. Detailed views of these plumes are shown on the right. Note the shadow of the plume located near the terminator (NASA–Galileo).

EUROPA (diameter 3138 km) is a satellite of Jupiter, 3.04 g cm^{-3} in density. This density indicates a silicated composition. It may be briefly described in this section even though its visible surface is completely covered by water ice. Density calculation and modelling indicate a silicated sphere of radius 1400–1500 km, surrounded by a 100–150 km thick H_2O layer. Voyager and Galileo images revealed geologically active or recent resurfacing processes bringing liquid water or ice-crystal mush to the surface. The 100–150 km thick H_2O layer is probably not completely frozen, and a liquid water ocean is probably overlain by an ice field. The flooding of liquid water on the surface does not correspond to volcanism *sensu stricto*, but rather to polar-like phenomena. The H_2O layer is probably warmed at the H_2O–silicate boundary by volcanism, similar in origin to Io's volcanism, but with a lower intensity.

The icy bodies

The surfaces of 13 of these icy bodies larger than 450 km in diameter have been well imaged by Voyager and Galileo spacecraft. On these kinds of bodies, volcanism may bring to the surface the molten compound of the interior: liquid water. The morphology created by this water volcanism is very different from that of their silicated counterparts. Fractures, vents and dykes may exist, but volcanic morphologies similar to the popular notion of volcanoes do not exist. On many satellites (Ganymede, Enceladus, Miranda, Triton), the surface of resurfaced ice is affected by sets of parallel or crosscutting ridges and grooves, poorly understood in origin, despite their frequency in the solar system.

Active worlds: Enceladus (Saturn's satellite) and Triton (Neptune's satellite). ENCELADUS (diameter 500 km) displays two different terrain units: old heavily cratered terrain units, cut and overlain with smoother units closest in morphology to bright grooved or smooth terrains of GANYMEDE. There are essentially two types of grooves on Enceladus: Ganymede-like sinuous grooves, which may have the same origin (extruded and ridged low-viscosity fluid, probably liquid water or a mush of water-ice crystals), and straight grooves generally interpreted as tensional or volcanic fissures. The grooved and smooth terrains of Enceladus have no detectable impact craters,

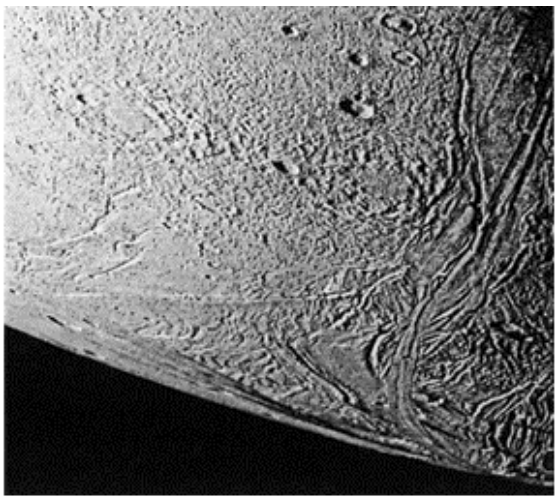
Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

which implies a very young age, of less than a few hundred million years. The albedo of Enceladus is uniform and virtually equal to 1. Analysis of the way the surface scatters light suggests that it is covered by fresh ice frost. A likely cause of this is a deposit of frozen spray from young volcanic eruptions perhaps continuing on Enceladus up to the present day. Another argument in support of the present dispersion of spay from Enceladus is that Enceladus orbits in the middle of the E ring, which consists of spherical particles about 1 μm in size. Such small particles would not survive more than ten thousand years in the interplanetary medium, suggesting that active volcanism from Enceladus is at the origin of the E ring.

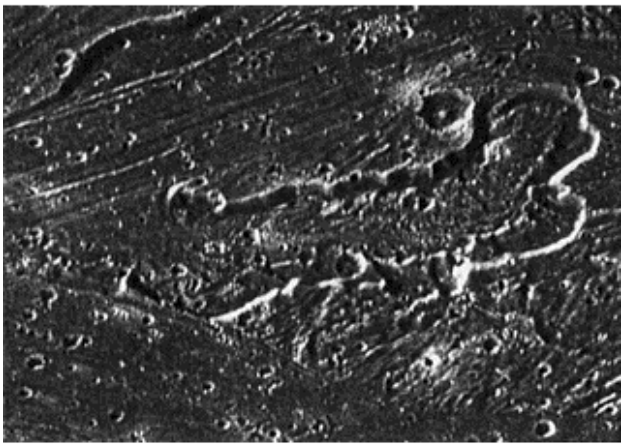
TRITON (diameter 2720 km) is the largest satellite of Neptune. The entire visible surface is crater poor and relatively young. The older terrains (the hummocky and the cantaloupe terrains) show volcanic evidence: flow-generated terrains containing a variety of domes and ridges which appear to be viscous extrusions which have erupted along the fault lines. These older terrains are overlain with smooth plains, which in some places form full flat-floored depressions 200 km wide. These younger terrains seem to have been formed as a series of large-volume, viscous flows. In the center of the flat-floored depressions is a collection of pits and flows, which may be the site of the most recent volcanic activity. The southern part of Triton is covered by a polar cap probably made of frozen nitrogen. In 1989, two active plumes were seen by Voyager inside the polar cap. Each plume sends a jet of dark material upwards to an altitude of 8 km and produces dark deposits. Triton is the only icy body in the solar system where eruption processes have been seen without doubt in action.

Recently active worlds: Ganymede (Jupiter's satellite), Dione, Tethys (Saturn's satellites), Ariel, Miranda and Titania (Uranus's satellites). These are worlds where there are evidences of both tectonic proceses on a global scale and vast volcanic outpourings which occurred after the end of late heavy bombardment.

Ganymede (diameter 5262 km, figure 4) represents the best known of these bodies. Ganymede is divided into dark heavily cratered terrains that occur as polygonal regions and poorly cratered bright terrains in the form of bands that separate the heavily cratered polygons. The bright terrains are usually strongly grooved and ridged and provide unequivocal evidence of widespread-to-global volcanism and tectonism. In some place, bright terrains are smooth, notably where they overlay the dark terrains. This suggest that the bright terrains were emplaced during extensional tectonic events, such as an extruded low-viscosity fluid, probably liquid water or a mush of water-ice crystals. The age of the end of this flooding is strongly dependent on cratering models in the outer solar system, and may be estimated to be 3–3.5 Gyr.



(a)



(b)

Figure 4. (*a*) Global view of the south pole of Ganymede. Flat terrains probably consist of large frozen water flows. Ridged and grooved terrains are deformed ice flows. Numerous impact craters indicate that this flooding stopped 3.5 Gyr ago (NASA–Voyager). (*b*) Detailed view of Ganymedean surface. Flat and ridged terrains are crosscut by an elongated caldera-like feature. Viscous flow (probably ice–liquid mush) is visible inside the caldera (NASA–Galileo).

Large-scale resurfacing and flooding without tectonics is visible on DIONE (diameter 1120 km) and on one area of TETHYS (diameter 1060 km), while evidence of water volcanism and resurfacing related to tectonism exists on other parts of Tethys, and on TITANIA (diameter 1580 km) and ARIEL (diameter 1158 km). On Ariel, canyons called 'Chasma' have clearly been flooded by some kind of volcanic flow processes which somewhere produced sinuous rilles with scales and morphology comparable with their lunar maria counterparts.

MIRANDA (diameter 472 km) is one of the most remarkable and complex worlds in the solar system despite its small diameter. Half of Miranda is a uniform densely cratered plain, while the other half is occupied by

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

regions of younger grooved terrains, the detailed appearance of which has never been found elsewhere in the solar system. These regions are called 'corona' and exhibit belts of parallel or crosscutting ridges. These ridges are commonly interpreted as fracture-controlled extrusions of warm ice or mush. The pattern of ridges is not understood at all. Somewhere, the ridges seem submerged by lava-flow-like structures.

Dead worlds: Callisto (Jupiter's satellite), Iapetus, Mimas, Rhea (Saturn's satellites), Oberon, Umbriel (Uranus's satellites). The Voyager and Galileo fly-bys revealed six satellites as intensely cratered worlds. If some kind of volcanic resurfacing processes had existed, they ceased before the end of late heavy bombardment. Possible volcanic morphologies have now almost completely disappeared, with the possible exception of Callisto and Rhea.

On CALLISTO (diameter 4800 km), some flat plains are associated with fractures surrounding a large impact (Valhalla) and may be due to some water resurfacing.

On RHEA (diameter 1530 km), one-third of the known surface is less cratered than the other two-thirds. Possible clues that there was endogenic activity on Rhea are given by features such as linear troughs and ridges, which may be interpreted as subdued volcanic vents.

Origins of volcanism on the icy satellites

Icy bodies are not large bodies (smaller than or equal in size to Mercury), and their volcanism should be completely extinct. Moreover, the size does not display a significant effect on internal activity: the small Enceladus is still active, while the large Callisto never has significant activity. Two physical reasons may produce abnormally active volcanism.

(1) With the exception of the unimaged Pluto, all the icy bodies are satellites of large planets, and permanent or transitional orbital resonances with other satellites may produce permanent or transitional orbital eccentricity and tidally generated heat. That would explain the difference between Ganymede and Callisto: Ganymede is closer to Jupiter than Callisto; tidal heating may have been greater for Ganymede and sufficient to produce volcanism during the first billion years. Tidal heating was never sufficient for Callisto. The heat sources of Enceladus and other bodies smaller than 1000 km in diameter are more uncertain and may be searched for in temporary orbital resonances with other satellites. The nature and shape of convective motions inside icy mantles are unknown at present.

(2) If a temperature of 1100 ºC is necessary to melt silicates, a temperature of about 0 ºC is sufficient to melt pure water ice, and a temperature of –100 ºC induces the melting of a water–ammoniac ice mixture. Even low internal heating could increase the temperature sufficiently to reach the melting temperature and to produce water or water–ammoniac volcanism.

The case of Triton is more complex. Volcanism-induced resurfacing which occurred everywhere is probably due to radiogenic and tidal heating, as for Ganymede or other large icy satellites. Such mechanisms could explain the active plumes. However, these active plumes are only located on the southern polar cap, and an exogenous origin is possible. The polar cap is made of transparent nitrogen ice covering a darker substrate. This was Tritonian 'springtime' when Voyager 2 imaged the south pole of Triton in 1989, and the temperature was increasing. Nitrogen ice is transparent and allows most of the sunlight to heat the underlying darker layer. A rise of only 10 ºC of this dark layer above the surface temperature of –236 ºC would cause the vaporization of the base of the nitrogen ice layer and its eruption as a nitrogen gas plume, through the icy superficial crust.

Pierre Thomas

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Voyager Mission

If one were to ask a space scientist which of all the space missions ever flown he or she would consider the most scientifically successful, there is a good chance that NASA's Voyager Mission to Jupiter, Saturn, Uranus and Neptune would be mentioned more than all others combined. Launched from Cape Canaveral, Florida, in 1977 (Voyager 2 on 20 August and Voyager 1 on 5 September), these two hardy spacefarers celebrated their 20th anniversary in late 1997. Although their last planetary encounter was in 1989, they continue to collect useful data about the outer solar system and will probably celebrate their 40th anniversary in the year 2017 before their useful lives come to an end.

As early as 1966, engineers at NASA's JET PROPULSION LABORATORY (JPL) realized that it was possible with a single spacecraft to visit all of the giant gaseous planets in the solar system (JUPITER, SATURN, URANUS and NEPTUNE). After rejecting a number of more costly and more ambitious missions proposed by JPL, NASA approved a two-spacecraft Mariner Jupiter/Saturn Mission to take partial advantage of this fortuitous planetary alignment. The initial approval was for a 4 yr mission, including flybys of Jupiter and Saturn only by each of the two Voyager spacecraft. The official project start date was 1 July 1972, and launch was slated for the late summer of 1977. The mission was renamed Voyager shortly after project start. (The name Voyager had previously been selected by JPL for an unsuccessfully proposed mission to Mars. The basic concepts of that earlier Voyager mission were essentially reproduced in the successful dual-orbiter, dual-lander VIKING MISSION to Mars.)

The two Voyager spacecraft were built with identical engineering designs and identical scientific payloads. Part of that design was based on results from the smaller PIONEERS 10 and 11. Neither Pioneer spacecraft detected dust particles in their passages through the Asteroid Belt. As a consequence, the Voyager dust detector was discarded in favor of a plasma wave detector. The eleven investigations and their Principal Investigators or Team Leaders are given in table 1. The two Pioneer spacecraft also experienced very high levels of radiation in the vicinity of Jupiter; this led to a campaign to build the Voyager engineering and scientific hardware to be resistant to damage from these higher radiation fields.

Although the Voyager Mission was approved by NASA for the 4 yr flight including Jupiter and Saturn encounters only, NASA gave approval for JPL to design the mission of Voyager 2 (but not the spacecraft or its scientific payload) to go beyond Saturn to Uranus and Neptune (or to Pluto). As noted earlier, Voyager 1 was launched more than two weeks after Voyager 2. However, the flight to Uranus required a slower flyby of Saturn; Voyager 2's flight to Jupiter and Saturn was therefore a few per cent slower than that of Voyager 1. It was prior knowledge of that fact that prompted the 'reverse numbering' of Voyagers 2 and 1. Voyager 1 encountered the planet Jupiter on 5 March, 1979. Voyager 2 followed on 9 July of that same year. Voyagers 1 and 2 flew by Saturn on 12 November 1980 and 25 August 1981, respectively.

Following Voyager 1's successful flyby of Saturn, which included a very close approach of Saturn's largest

Table 1. Voyager Investigations and their Lead Scientists.

Investigation and acronym	*Lead Scientist*	*Affiliation*
Cosmic Ray Subsystem (CRS)	Rochus E Vogt (1972–84)	California Institute of Technology
	Edward C Stone[a] (1984–)	California Institute of Technology
Infrared Imaging Spectrograph Subsystem (IRIS)	Rudolph A Hanel (1972–87)	NASA Goddard Space Flight Center
	Barney J Conrath (1987–91)	NASA Goddard Space Flight Center
Imaging Science Subsystem (ISS)	Bradford A Smith (1972–91)	University of Arizona
Low Energy Charged Particles Subsystem (LECP)	S M (Tom) Krimigis (1972–)	APL, Johns Hopkins University
Magnetometer Subsystem (MAG)	Norman F Ness (1972–)	NASA Goddard Space Flight Center and University of Delaware
Plasma Science Subsystem (PLS)	Herbert S Bridge (1972–87)	Massachusetts Inst. of Technology
	John W Belcher (1987–97)	Massachusetts Inst. of Technology
	John D Richardson (1997–)	Massachusetts Inst. of Technology
Photopolarimeter Subsystem (PPS)	Charles F Lillie (1972–78)	University of Colorado
	Charles W Hord (1978–9)	University of Colorado
	Arthur L Lane (1979–91)	NASA Jet Propulsion Laboratory
Planetary Radio Astronomy Subsystem (PRA)	James W Warwick (1972–91)	Radiophysics, Incorporated
Plasma Wave Science Subsystem (PWS)	Frederick L Scarf (1972–88)	TRW Systems
	Donald A Gurnett (1988–)	University of Iowa
Radio Science Subsystem (RSS)	Von R Eshleman (1972–8)	Stanford University
	G Leonard Tyler (1978–91)	Stanford University
Ultraviolet Spectrometer Subsystem (UVS)	A Lyle Broadfoot (1972–91)	University of Southern California and University of Arizona

[a] Note that Edward C Stone also served as Project Scientist from project start through the present time.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

moon, TITAN, NASA gave approval for the Voyager Project to target Voyager 2's Saturn encounter for a continuation on to Uranus. The transition between the primary mission and the Voyager Uranus/Interstellar Mission (VUIM) occurred on 1 October 1981. Similar approvals were given for subsequent portions of the extended Voyager Missions. The Voyager Neptune/Interstellar Mission (VNIM) started 1 April 1986; the Voyager Interstellar Mission (VIM) began 1 October 1989. Because of Voyager 1's close approach to Titan, the spacecraft's trajectory passed well south of Saturn. The massive gravity of Saturn pulled Voyager 1's path northward of the ecliptic and hence northward of the orbits of Uranus, Neptune and PLUTO. However, the measurements Voyager 1 makes of interplanetary charged particles and magnetic fields continue to be of great interest to solar system scientists. Voyager 2 encountered Uranus on 24 January 1986, and Neptune on 25 August 1989; like Voyager 1, its mission is now directed at interplanetary and interstellar measurements.

The intensive data collection periods surrounding each of the planetary encounters generally included about 3 months prior to closest approach and one month thereafter. The first approximately two months of the encounter period constituted an 'Observatory' phase. 'Far Encounter' spanned the period from the end of the Observatory phase to about 1 day before closest approach. 'Near Encounter' generally lasted about 2 days, more or less centered on the closest approach to the relevant planet. A 'Post Encounter' of about 1 month concluded each planetary encounter period. Each of these phases included two or three sequence computer 'loads', which updated the central computer and sequencer (CCS) contents. The CCS would then control spacecraft activity, including recording to and playback of the digital tape recorder storage device. A full tape recorder would hold the equivalent of about 100 full-resolution (i.e. uncompressed) images.

Because Voyager was designed to operate at great distances from the Sun where solar panels were ineffective, power was (and is) provided for each of the Voyager spacecraft by means of radioisotope thermoelectric generators (RTGs). These 'batteries' generate electricity by transforming heat radiated by the natural decay of plutonium dioxide. These batteries, which utilize non-fissionable plutonium-238, are highly reliable and long lived, a perfect combination for a mission like Voyager. The RTGs carried by each Voyager generated about 500 W of power at launch and are decreasing in output by about 5 W yr^{-1}. This will permit them to continue to power the two spacecraft with usable amounts of wattage until about the year 2020.

Communication with the spacecraft occurred via digitally modulated radio signals. Commands transmitted from the ground occurred at S-band, near a frequency of 2113 MHz. Data were transmitted to the ground at X-band, near a frequency of 8415 MHz. Early in its mission, the Voyager 2 primary radio receiver failed. Following a week of no command receipt by Voyager 2, a computer-stored failure algorithm automatically switched to the spacecraft's backup receiver. It was soon discovered that the backup receiver was 'tone deaf'. In other words, if the S-band signal arriving at the spacecraft differed by more than 96 Hz from the receiver's frequency, it would not be detected. Factors such as the velocity of the spacecraft and the rotation and revolution of the Earth would cause changes (due to Doppler shifting of the transmitted radio signal) that easily exceeded 96 Hz. While adjustments in the transmitted S-band frequency could be made for the factors mentioned, indeterminate shifts in the central receiver frequency would occur each time power usage (and hence temperature) would change by more than a couple of watts. Power changes were therefore minimized; when power usage changes were necessary, a command moratorium of from 24 to 72 h was honored, providing enough time to determine the new receiver frequency. Voyager engineers became very proficient at dealing with this shortcoming of Voyager 2, and it never substantially limited the spacecraft's capabilities.

Another Voyager 2 hardware problem occurred during the Saturn flyby. Voyagers 1 and 2 are three-axis stabilized spacecraft. Except for brief periods of time, the 3.7 m wide high-gain communications antenna is kept pointed at Earth, and the star tracker is locked on a selected bright star. The cameras and spectrometers were then pointed at their targets by means of a steerable two-degree-of-freedom scan platform. One of the two axes of that scan platform seized about half an hour after Voyager 2's closest approach to Saturn. The problem was traced to the migration of lubricant out of the gear train during periods of heavy usage. Motion was restored a few days later. Careful study of the characteristics of the failure, both from spacecraft engineering data and from life tests on similar gear trains at JPL, enabled engineers to devise a usage plan that would avoid another seizure. The scan platform operated flawlessly during the Uranus and Neptune encounters.

There are several examples of engineering 'tricks' that improved the Voyager science return. Normal limit cycle motion of the spacecraft was controlled within 0.05° in each of the three spacecraft axes. Attitude knowledge was provided by monitoring the Sun Sensor and Star Tracker. Attitude control was accomplished by 200 ms bursts of hydrazine fuel through appropriate thruster jets. Using that scheme, the motion of the camera field of view across the sky was slower than the motion of an hour hand on a clock. Nevertheless, long exposures and low light levels demanded even slower rates. The Attitude and Articulation Control System (AACS) aboard Voyager 2 was programmed to shorten the thruster pulses, slowing the resultant rates across the 0.05° deadband to less than 10% of the normal quiescent rate, almost eliminating image smear in more than 90% of the images obtained at Uranus and Neptune.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Another scheme to reduce image smear involved rotating the cameras at a rate that matched the apparent motion of a nearby satellite (moon) during flyby of that satellite. This scheme involved using onboard gyroscopes for attitude reference (instead of the Sun Sensor and Star Tracker). The AACS was fooled by doctored data in the spacecraft computer into thinking the gyroscopes were drifting. As the AACS compensated (as it was designed to do) for this false 'gyro drift', it caused the spacecraft to turn in a way that kept the cameras pointed at the nearby moon without having to use the (somewhat jerky) movement of the scan platform. This technique was first used to obtain unsmeared images of Saturn's moon Rhea and was particularly successful in image sequences of Uranus's moon MIRANDA, and Neptune's moon TRITON.

The scientific findings of the Voyager Mission have been the most notable of its contributions. The individual scientific articles based on Voyager data now number in the thousands, and Voyager data continue to be the source of detailed comparative studies of the giant planets and their rings, satellites and magnetospheres (the magnetic fields surrounding each of the giant planets and the electrically charged particle population trapped within or traversing those magnetic fields).

Atmospheres and interiors

Voyagers 1 and 2 obtained the first detailed composition, temperature and pressure profiles of the atmospheres of Saturn, Uranus and Neptune and improved our understanding of those characteristics of the atmosphere of Jupiter, previously visited by Pioneers 10 and 11. The GALILEO MISSION has dropped an instrumented probe into Jupiter's atmosphere and completed its orbital mission around the solar system's largest planet, improving on many of the measurements made by Voyagers 1 and 2. Ulysses also completed a high-latitude flyby of Jupiter on its way to study the polar regions of the Sun (see SOLAR WIND: ULYSSES). The CASSINI/HUYGENS MISSION is designed to make enormous improvements in our understanding of Saturn's atmosphere and interior as it circles the planet for 4 yr starting 1 July 2004. However, for Uranus and Neptune, the Voyager *in situ* results and the high-resolution and high-phase-angle remote-sensing observations will probably remain unique for the foreseeable future.

Our understanding of the interiors of the giant planets is based largely on theoretical considerations. However, there are many bits and pieces of information, mostly from Voyager results, that serve to constrain the range of viable models. Careful measurements of the paths of the Voyager spacecraft helped determine accurate measurements of the masses of the giant planets and of their entire systems. Voyager determined the rotation rates of the interiors of Saturn, Uranus and Neptune. (Jupiter's interior rotation rate was determined by radiowave measurements made from Earth.) The orientations of the strange magnetic fields of Uranus and Neptune (see below) offer additional clues. Excess heat in the form of infrared energy escaping from each of the four planets is also important in constraining models of the interior.

Rings

The Voyager spacecraft revealed the enormous amount of detail in the rings of Saturn, discovered the rings of Jupiter and provided the first detailed images of the rings of Uranus and Neptune. Ring data from Voyager have raised even more questions on the nature of ring systems than they have answered. The answers to those questions may be important to studies of the asteroid belt, the Oort cloud of comets, galactic structure and clusters of galaxies within the universe.

Satellites

Voyager imaged Earth's MOON. Three new satellites of Jupiter (Metis, Adrastea, and Thebe) were discovered in Voyager data, and Voyager images were obtained of eight of Jupiter's 16 known moons. Only the outer eight moons, all thought to be tiny captured asteroids, were missed. Voyager revealed active volcanoes on IO, an obviously fractured and geologically young icy surface on EUROPA (first evidence of a global subsurface ocean), a variegated array of surface structures on GANYMEDE and an ancient cratered icy surface on CALLISTO. The sizes, masses, densities and reflectivities of the four Galilean satellites were also measured, showing Ganymede to be the solar system's largest satellite, Europa to be the brightest Galilean satellite and a decrease in bulk density with increasing distance from the planet (presumably due to water depletion from the closer satellites due to heat from Jupiter).

At Saturn, four new satellites (Pan, Atlas, Prometheus and Pandora) were discovered in Voyager data. These data also showed that Janus and Epimetheus, previously thought to be a single satellite, were actually two satellites with remarkably similar orbits. Three satellites (Telesto, Calypso and Helene) were discovered from Earth as a result of the increased interest in Saturn generated by the approaching Voyager. Voyager revealed that Titan was the solar system's second largest satellite, and the only one with a substantial atmosphere. Titan is covered by a high atmospheric haze, which hides the surface from visible imaging. However, infrared and radiowave measurements penetrated to the surface of this strange moon, which may be the only body in the solar system (other than Earth) with both liquid and solid material at its surface. Voyager obtained images of all 18 known satellites of Saturn. These images revealed a geologically young surface on ENCELADUS, erratic rotation for HYPERION and a remarkably sharp boundary between dark material on Iapetus' leading face and bright material on its trailing face. Phoebe, apparently a captured asteroid, was revealed to have the darkest surface among Saturn's satellites.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Voyager added 10 new satellites (Cordelia, Ophelia, Bianca, Cressida, Desdemona, Juliet, Portia, Rosalind, Belinda and Puck) to the five previously known at Uranus. Miranda has unique surface features known as coronae owing to their strange concentric patterns. Six new satellites (Naiad, Thalassa, Despoina, Galatea, Larissa and Proteus) were added to Neptune's prior repertoire of two. Triton was discovered to have nitrogen ice, extremely cold temperatures (32 K), and geyser-like plumes spewing dust through cracks in the icy surface.

Magnetospheres

Little was known about the magnetospheres of the giant planets before the Voyager flybys. Pioneers 10 and 11, as Voyager 'pathfinders', had made preliminary measurements of Jupiter's magnetosphere, and Pioneer 11 probed Saturn's magnetosphere. Voyager made significant improvements in those measurements and provided the first measurements of the magnetospheres of Uranus and Neptune. The magnetic fields of the latter two planets were found to be highly inclined to the planetary rotation axes, something previously expected only for certain stars.

The final frontier for Voyager is a search for the outer edge of the heliosphere (the Sun's magnetosphere). Theoretical calculations show that the boundary should be at a distance of between 80 and 120 astronomical units (AU) (1 AU is the mean distance between Sun and Earth). Voyager 1 crossed 80 AU in 2001; Voyager 2 reaches that distance in 2006. Once Voyager 1 is beyond the heliosphere, humanity will be able for the first time to sample the magnetic fields and charged particles in the space between the stars.

The two Voyager spacecraft are still sending back data, 25 years after their launch. Voyager 1 is now the most distant human-made object; in August 2002 it was about 85 times as far from the Sun as the Earth and Voyager 2 was about 68 times the Sun–Earth distance. They are both continuing outwards at about 3 AU per year (slightly more for Voyager 1 and slightly less for Voyager 2). The Voyager team at JPL still receives information almost daily from the Voyagers which are examining the far reaches of the solar wind.

Bibliography

A detailed pre-launch description of the Voyager mission, the spacecraft and the scientific investigations is given in *Space Science Reviews* **21** pages 75–376 (1977).

Voyager scientific results have been chronicled in multiple dedicated issues of *Science*, *Nature*, *Icarus* and the *Journal of Geophysical Research*. They have also resulted in at least six separate many-hundred-page scientific texts on planetary science published by The University of Arizona Press.

Ellis D Miner

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Vulcan

The name given to a hypothetical planet believed by Urbain Le Verrier to exist within the orbit of Mercury. At the request of François Arago, Le Verrier first studied the orbit of Mercury in the early 1840s. He managed to explain the greater part of the discrepancy between its calculated and observed positions (the advance of its perihelion) as being caused by gravitational perturbations by the other planets. The residual discrepancy eluded him, and he abandoned the problem as intractable. He returned to it in 1859, having in the meantime successfully predicted the existence of Neptune. Now he invoked an intramercurial planet or asteroid belt as the cause of Mercury's irregularities. Shortly after, he heard that a physician and amateur astronomer, Edmond Lescarbault, had observed a small body in transit across the Sun's disk. Le Verrier examined Lescarbault's observations closely and, convinced they were genuine, announce the existence of a new planet which he named Vulcan. Subsequent sightings were reported but never confirmed. Vulcan is now known not to exist.

See also: perihelion.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Vulpecula

(the Fox; abbrev. Vul, gen. Vulpeculae; area 268 sq. deg.) a northern constellation which lies between Cygnus and Sagitta–Delphinus, and culminates at midnight in late July. It was introduced as Vulpecula cum Ansere (the Fox and Goose) by the astronomer Johannes Hevelius (1611–87) of Danzig (Gdansk), who included it in his atlas *Firmamentum Sobiescianum sive Uranographia* of 1687.

A small, inconspicuous constellation, the brightest star in Vulpecula is α Vulpeculae, magnitude 4.4. The Milky Way passes through Vulpecula and the constellation contains a number of open star clusters and planetary nebulae, including Cr 399 (Brocchi's Cluster, the Coat-hanger), which consists of about a dozen stars of magnitudes 6–8, elongated east–west, NGC 6940, another open cluster with about 100 stars fainter than ninth magnitude, and M27 (NGC 6853), the Dumbbell Nebula, an eighth-magnitude planetary nebula.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

W M Keck Observatory

The W M Keck Observatory, located on the island of Hawaii, operates the world's two largest optical/infrared telescopes, each with a primary mirror 10 m in diameter, near the 4200 m summit of Mauna Kea. Made possible through grants totaling more than $140 million from the W M Keck Foundation, the observatory is operated by the California Institute of Technology, the University of California and the National Aeronautics and Space Administration (NASA), which joined the partnership in October 1996. The Keck I telescope began science observations in May 1993; Keck II in October 1996.

A staff of 80 scientists, engineers, technicians and support personnel operate the observatory whose administrative facility is located in Waimea, Hawaii, and whose mission is to provide a world-class research facility for astronomers from Caltech, University of California, NASA and the University of Hawaii. Over 400 astronomers per year are involved with observations from the Keck telescopes, which are carried out from Waimea via a fiber-optic link to the summit of Mauna Kea.

The Keck telescopes employ a unique segmented design for their primary mirrors: 36 1.8 m diameter hexagonal segments are fitted together like a floor-tile mosaic to form each primary, the segments being aligned with respect to each to a tolerance of one-millionth of an inch under computer control. This expandable technology is that most likely to be adopted for the giant telescopes being planned for the 21st century.

A $55 million project funded by NASA is aimed at joining the two Kecks, along with four 1.8 m 'outrigger' telescopes, as an infrared interferometric array to achieve unprecedented angular resolution, by the year 2003.

The first test observation obtained by linking the two Keck 10 m telescopes was made on 12 March 2001. Light from HD61294, a faint star in the constellation Lynx, was captured by both Keck telescopes. The collected light waves were combined and processed with a beam combiner and camera. To phase the two telescopes properly, adaptive optics on both telescopes removed the distortion caused by the Earth's atmosphere. In addition, the optical system in the tunnel adjusted the light path to within a millionth of an inch.

Major discoveries from this young observatory include: the discovery of several planetary systems around other stars; the identification of gamma-ray bursters as being at cosmological distances; the discovery of the most distant objects in the universe; the measurement via supernova observations of the apparent acceleration of the universe.

For further information see http://www2.keck.hawaii.edu:3636/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wallenquist, Åke Anders Edvard (1904–94)

Swedish astronomer, worked at the Bosscha Observatory in Indonesia and became professor in Uppsala at the Kvistaberg Observatory. He worked on double stars and open star clusters.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Ward, Seth (1617–89)

Born in Aspenden, Hertfordshire, became Savilian Professor of Astronomy at Oxford, formulated an 'empty focus' alternative to KEPLER's law of areas. (A planet moved with uniform angular velocity around the empty focus of its ellipse.)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Warner & Swasey Observatory

Located at Washburn University in Topeka, Kansas, home of the Warner & Swasey 29 cm refractor. Built in the late 1800s, the telescope was displayed at the 1912 World's Fair, then acquired by Washburn College. Crane Observatory was built on campus to house the telescope. The Warner & Swasey survived a tornado in the 1960s. During telescope refurbishment, which was completed in 1998, all of the original parts were retained.

For further information see http://www.icstars.com/warner_swasey/wsmain.htm.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Warner, Worcester Reed (1846–1929)

American mechanical engineer, designed the 36 in telescope for the Lick Observatory and built telescopes for Canada and Argentina. With AMBROSE SWASEY he established a machine manufacturing company, and established the Warner & Swasey Observatory.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Water Cycle

The Earth's water cycle is the endless circulation of water on a planetary scale, from ocean to continent through the atmosphere and back to the ocean, under the driving influence of solar energy and gravity. This main cycle in fact consists of many overlapping loops and branches at all scales, involving more or less durable storage, more or less rapid transfers, more or less intense transformations between water phases. Because of this complexity, it was only in the late 16th and 17th centuries that the concepts of water cycle and balance arose, with the pioneering work of B Palissy, P Perrault, E Mariotte and E Halley founding scientific hydrology. Earth is a system approximately closed for water but largely open for energy. The particular properties of water together with the strong coupling between water and energy cycles regulate the Earth's mean temperature and shape weather and climate patterns and their fluctuations (see CLIMATE). The water cycle interacts also strongly with most biogeochemical cycles, with long-term environmental feedbacks. Human activities have an increasing impact on the water cycle and the coupled fluxes, with some alarming consequences for the quantity and quality of water resources.

Properties of water

Water (H_2O) is a very stable molecule (covalent link), which has very specific physicochemical properties, mostly because of its dipolar structure and the related hydrogen link between molecules. These properties explain its multiform importance within the Earth's hydrosphere and its four main functions as resource, agent, vector and biotope.

- Its three phases (solid, liquid, vapor) are present within a relatively small range of temperature, corresponding to the temperature range observed on the Earth.
- Its very high latent heats of vaporization (2.454 MJ kg^{-1} at 20 °C) and fusion (0.334 MJ kg^{-1}) play a major role in the heat transfer by water through the atmosphere.
- Its high specific heat and low thermal conductivity make water bodies very good thermal regulators and water fluxes very good heat convectors: oceans play a major role in heat storage and transfer.
- Water vapor absorbs most of the infrared and ultraviolet radiation but is transparent to visible radiation; atmospheric vapor leads to the 'greenhouse effect', while liquid water and solid water have contrasting reflectivities. These radiative properties are central to the thermal budget of the Earth (see GREENHOUSE EFFECT).
- Because of its high dielectric constant, water is an excellent solvent, able to alter most minerals: natural waters are in fact more or less dilute aqueous solutions, providing nutrients to aquatic life and contributing by their movement to the redistribution of substances; water can also combine with many substances.
- Its density is maximal at 4 °C, permitting ice to float over liquid water: freezing and warming of water bodies occur from the top, progressive freezing of the whole ocean being avoided.
- Its low viscosity makes it a very mobile liquid, able to occupy all the space available, even within very small pores, and its surface tension allows soil water retention by capillarity against gravity, by which vegetation can persist between two rainfalls.
- Water molecules present several isotopic combinations, whose relative abundances provide very useful (paleo)thermometers and tracers of the water's origins.

Origin and uniqueness of water on the Earth

Earth, the 'blue planet', is unique within the solar system by the presence in great abundance of water in its three phases.

The Jovian planets (Jupiter, Saturn, Uranus, Neptune) are composed of light elements, including hydrogen and helium, and most of their satellites contain a large fraction of water ice, because of the very cold conditions prevailing so far from the Sun. The Jovian satellite Europa, in addition, might contain a water ocean below its icy crust. In the Telluric planets (Mercury, Venus, Earth, Mars) constituted of heavy elements, water was partly outgassed by tectonic and volcanic activity and partly brought by impacts of meteorites of asteroidal or cometary origin. The fate of water in the Telluric planets has differed largely according to the very different temperature and pressure conditions prevailing on each planet, mostly as a result of their distances from the Sun (figure 1).

While Mercury no longer has an atmosphere and probably still contains only a little ice in cold regions, on Venus the thick carbon dioxide atmosphere is much too hot for water condensation because of a strong greenhouse effect, and most outgassed water vapor was subsequently photodissociated (photolysis) (see VENUS). Mars is currently cold and dry because of its thin carbon dioxide atmosphere, but there is much geomorphological evidence of past water-related erosion, and some limited water is probably still contained in permafrost and ice caps (see MARS). Earth is the only planet where the three water phases coexist: a greenhouse effect due to the presence of water vapor and carbon dioxide in its atmosphere has maintained temperature conditions permitting the condensation of very huge amounts of outgassed water, and therefore the development of life. Uniformly distributed over the whole planet's surface, the depth of all the water on Earth would amount to about 3000 m, as compared with rough estimates of about 200 m for Mars and 0.20 m for Venus.

Main reservoirs and fluxes

The total amount of water within the Earth's hydrosphere (that is, above and within the upper 2 km of the crust) can be considered as constant, even though there are very small net water exchanges with both space (loss by

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

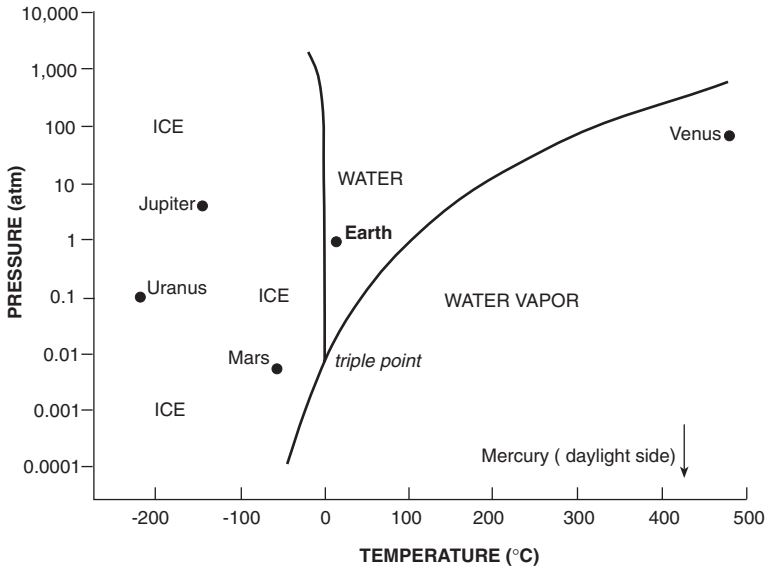


Figure 1. Planetary positions on the phase diagram of water. Adapted from National Research Council (1991). Reproduced by permission of National Academy of Sciences.

diffusion, gain by icy comets) and the Earth's mantle (loss by subduction, gain by volcanic activity), where there is a much slower tectonic water cycle. This huge amount of water (about 1386 million km^3, of which only 2.5% is fresh water) is stored in several main interconnected reservoirs differing by the volume and quality (salt, fresh) of the water they contain. The fluxes they permanently exchange within the water cycle control their renewal rates, which differ largely according to their size and the water phase they store. Current estimates of global storages and mean annual fluxes are still affected by rather large uncertainties especially for groundwater (compare for instance values in table 1, table 2 and figure 2).

The largest reservoir is by far the world ocean, which covers 71% of the 510 million km^2 of Earth's surface with a mean depth of 3700 m. It contains 96.5% of all the water, as salt water, and has a long mean residence time of 2500 yr.

Continents store the remaining 3.5% of water in several main reservoirs with a very large range of sizes and renewal rates.

- Ice caps, glaciers and permanent snow cover store 1.74% of total water over 3% of Earth's surface (essentially in the Antarctic). This solid phase represents 69% of all fresh water, with a very long residence time in ice caps (10^4 yr).
- Groundwaters within the upper 2 km of the Earth's crust contain about 1.7% of total water (45% only as fresh water), with very variable residence time (1–10^3 yr) according to their depth and the porous media properties. Soil moisture accounts for 0.001% only.
- Surface water (lakes, rivers, swamps) represents only 0.01% of all water, most as fresh water. The mean residence time is short for rivers (about 18 days) and much longer for lakes (up to 1000 yr).
- Biomass contains only 0.0001% of total water, but the high water content (>70%) of most plants and living organisms is renewed within a few hours or days.

The atmosphere stores 0.001% of total water, mostly as vapor. Liquid and solid phases are also present in clouds, which cover more than 60% of Earth's surface on average. Atmospheric water corresponds to an equivalent height of only 26 mm but is frequently and rapidly renewed in 9 days, ensuring an intense recycling of water.

The global water cycle presents two main interconnected loops of different intensity (figure 2).

- Ocean and atmosphere exchange large fluxes: the ocean loses much water by evaporation (434 000 km^3 yr^{-1}; 1202 mm yr^{-1}), a flux only limited by the energy available. Most of this flux returns directly to the ocean by precipitation (398 000 km^3 yr^{-1}; 1102 mm yr^{-1}).
- Continent and atmosphere exchange fluxes that are smaller because less water is available: precipitation on continents (107 000 km^3 yr^{-1}; 719 mm yr^{-1}) is in great part compensated by evapotranspiration (71 000 km^3 yr^{-1}; 477 mm yr^{-1})—that is, water evaporated from the land surface and transpirated by vegetation.
- The water balances of both loops are equilibrated by the fluxes they exchange: vapor flux in excess from the ocean (36 000 km^3 yr^{-1}; 100 mm yr^{-1}) is transported by atmospheric circulation over continents, while an equivalent flux of precipitation excess on the continent (36 000 km^3 yr^{-1}; 242 mm yr^{-1}) returns to the ocean by gravity as surface and subsurface runoff.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. Water reserves on the Earth.

Reservoir	Distribution area (10^3 km^2)	Water volume (10^3 km^3)	Water depth (m)	Proportion of total water (%)	Residence time
World ocean	361 200	1 338 000	3700	96.5	2500 yr
Glaciers, permanent snow	16 230	24 064	1463	1.74	Up to 10^4 yr
Antarctic	13 980	21 600			
Groundwater (upper 2 km)	134 800	23 400	174	1.7	Up to 10^3 yr
Fresh water		10 530	78	0.76	
Soil moisture		16.5	0.2	0.001	Days–months
Ground ice–permafrost	21 000	300	14	0.022	
Lakes	2 060	176.4	85.7	0.013	Up to 10^3 yr
Fresh		91			
Saline		85.4			
Swamps	2 680	11.5	4.3	0.0008	
Rivers	148 800	2.12	0.014	0.0002	18 days
Biomass	510 000	1.12	0.002	0.0001	Hours–days
Atmosphere	510 000	12.9	0.025	0.001	9 days
Total water reserves	510 000	1 386 000	2718	100	
Total fresh water	148 800	35 029	235	2.53	

Adapted from Shiklomanov (1993).

Table 2. Estimates of mean annual water fluxes at global scale.

	National Research Council (1991)		Shiklomanov (1998)	
Water flux	km^3 yr^{-1}	mm yr^{-1}	km^3 yr^{-1}	mm yr^{-1}
Ocean				
Evaporation	434 000	1202	502 800	1392
Precipitation	398 000	1102	458 000	1268
Advection to continent	36 000	100	44 800	124
Continent				
Precipitation	107 000	719	119 000	800
Endorheic areas			9 000	
Exorheic areas			110 000	
Evapotranspiration	71 000	477	74 200	499
Endorheic areas			9 000	
Exorheic areas			65 200	
Runoff to ocean	36 000	242	44 800	301
From rivers			42 700	
From groundwater			2 100	
Total				
Evaporation	505 000	990	577 000	1131
Precipitation	505 000	990	577 000	1131

Endorheic/exorheic areas: not connected/connected to the ocean.

Globally precipitation and evaporation fluxes are equal and amount to 505 000 km^3 yr^{-1} (990 mm yr^{-1}).

Space and time variations

Mean global values of storage and annual fluxes in fact mask very large variations at different scales in both time and space. Variations in evaporation (E), precipitation (P) and the resulting continental runoff are mainly due to seasonal and latitudinal variations in solar energy modulated by seasonal thermal contrasts between ocean and continent (figures 3 and 4). An observation network and remote sensing mainly from satellites help to assess and monitor the variability of water storages and fluxes and the environmental factors controlling them (see SPACE-BASED OBSERVATIONS OF THE EARTH).

In the intertropical zone, where solar energy is maximal, the redistribution of water and energy within the atmosphere is ensured by the rather stable vertical circulation within the Hadley cells generated by large vertical contrasts in temperature, in the absence of a strong Coriolis effect. In the equatorial zone, the convergence of trade wind systems generates the uplift (low pressure) of warm air moistened by high evaporation from the ocean: this results in heavy convective precipitation exceeding the local evaporation, which is reduced by frequent cloud cover ($P > 2000$ mm yr^{-1}, $E < 1600$ mm yr^{-1}). Inversely,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

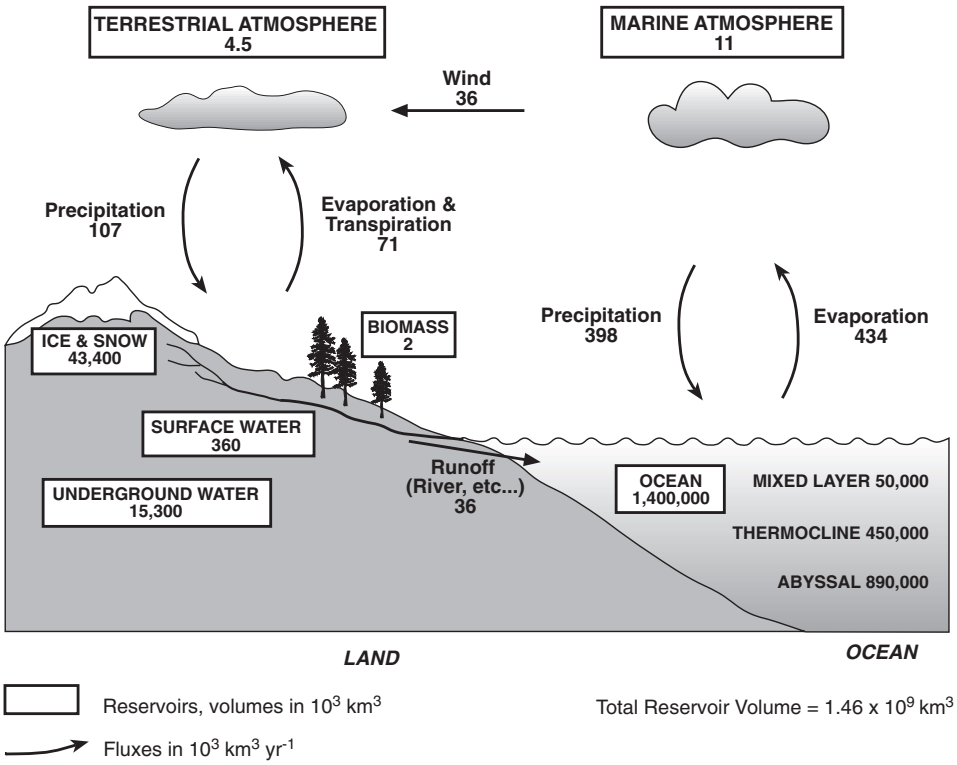


Figure 2. The water cycle at a global scale. Adapted from National Research Council (1991). Reproduced by permission of National Academy of Sciences.

subsidence of warm and dry air (high pressure) in the subtropical zones favors evaporation, which exceeds local precipitation ($P < 1200$ mm yr^{-1}, $E > 1800$ mm yr^{-1}).

In the mid-latitude zones (between 30° and 60°), the redistribution of water and energy from tropics to poles is rather organized horizontally, through successions of unstable eddies moving from west to east and generated by horizontal contrasts in temperature and by the Coriolis effect: cyclonic depressions provide precipitation, while dry anticyclonic conditions favor evaporation. At high latitudes, because of minimal solar energy and subsidence of cold and dry air (high pressure) precipitation and evaporation are very low (only a few mm yr^{-1} at the poles). In the extratropical zones, precipitation generally exceeds evaporation, but, while evaporation decreases rather regularly toward the poles, precipitation presents in the mid-latitude zone a maximum ($P > 1200$ mm yr^{-1}) associated with the depression systems.

This latitudinal climate zonation is most clearly observed on the ocean, where evaporation is only limited by the energy available. Differences in water ($P - E$) and energy balances associated with each climatic zone also generate a salinity pattern at the ocean surface (increasing by sea-water evaporation and freezing, decreasing by precipitation and sea-ice melting). Redistribution of water, heat and solutes within the world ocean is ensured by large-scale circulations and eddies (see OCEANS). While rapid surface currents are mainly induced by the prevailing surface winds, slow deep currents are mainly due to differences in water density (depending on water temperature and salinity). This thermohaline circulation generates the oceanic 'heat conveyor belt': in its main branch, cold dense water sinks in the Norwegian Sea, flows as deep water through the Atlantic ocean to the Indian and Pacific oceans where it warms and wells up and returns to the North Atlantic as dilute surface water in about 1000 yr.

On continents, where evapotranspiration is limited by water rather than energy, the climate pattern exhibits still more contrast, with warm deserts belts ($P < 300$ mm yr^{-1}) in the dry subtropical zones. This pattern is disturbed by other factors depending on the position, size and shape of each continent, whose cumulative effects result in a very large range of mean continental precipitation (from almost 0 to 12 000 mm yr^{-1}).

- Ocean–continent interactions: in the tropical zone, the western coasts of continents receive less precipitation than the eastern coasts, which are supplied by easterly trade winds carrying much moisture evaporated from the tropical ocean—for instance in the Amazonian basin.
- Continentality effect, especially in the middle latitude zones: precipitation on continents decreases from west to east, because of a progressive drying of air mass. Precipitation is more than 600 mm yr^{-1} in temperate

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

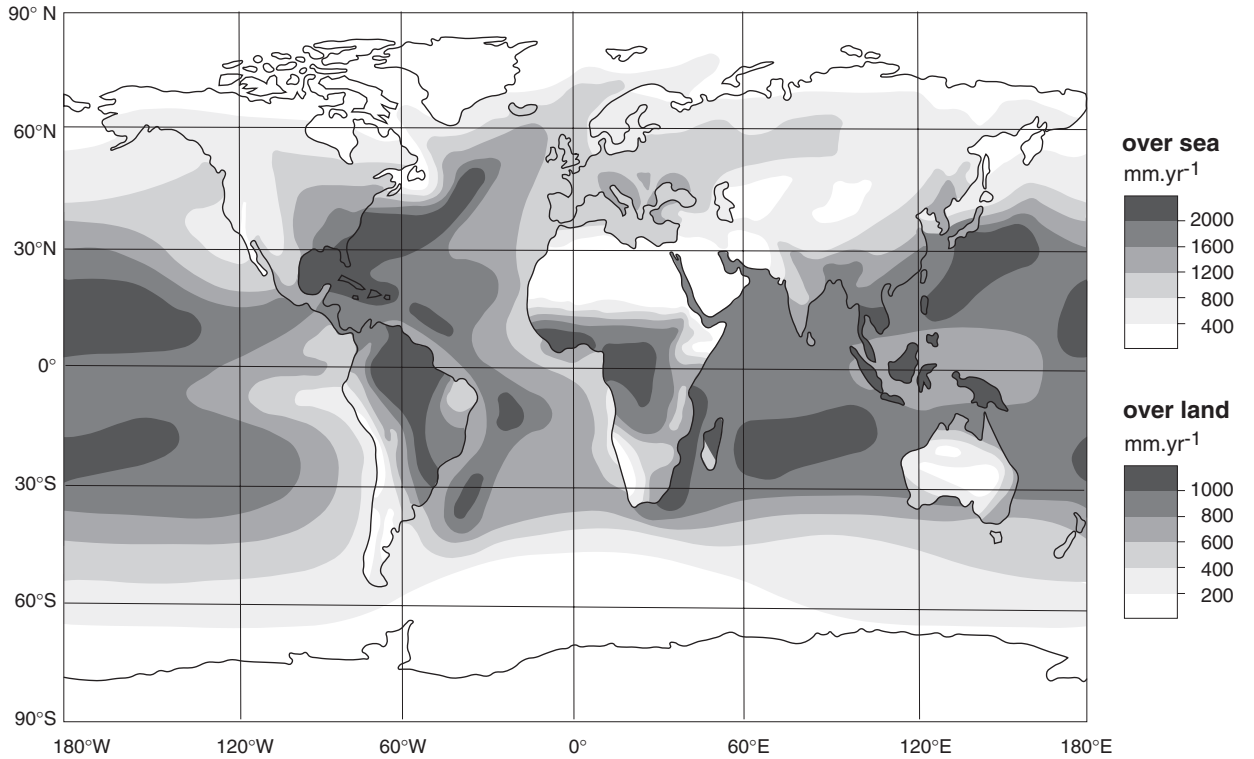


Figure 3. Global distribution of mean annual evaporation. Adapted from Robinson and Henderson-Sellers (1999). Reproduced by permission of Pearson Education.

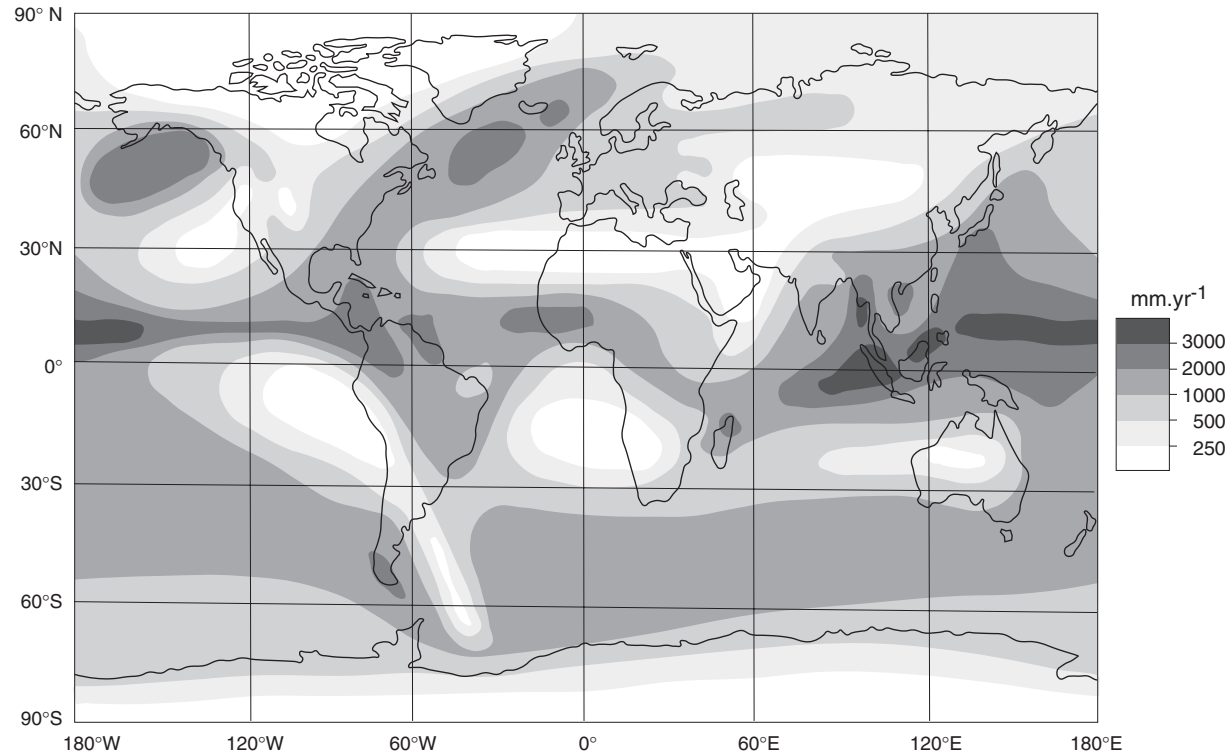


Figure 4. Global distribution of mean annual precipitation. Adapted from Robinson and Henderson-Sellers (1999). Reproduced by permission of Pearson Education.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Western Europe, less than 100 mm yr^{-1} in continental Kazakhstan.

- Orographic effect: mountain ranges on continents, islands in oceans, act as barriers to dominant humid air mass. The windward side receives much more orographic precipitation than the leeward side—as is the case for the Western Cordillera in North America or the Hawaiian Islands.

These mean annual spatial patterns are modulated by diurnal and seasonal variations in insolation or longer-term fluctuations.

- At high altitudes as at high latitudes, a great part of winter precipitation falls as snow, forming a seasonal or even permanent snow cover affecting both water and energy cycles.
- Slight seasonal fluctuations of the Hadley cells around the equator explain the alternation of rain and dry seasons in the tropical zone.
- Seasonal variations in thermal contrasts between ocean and continent also explain the monsoon, providing huge amounts of summer precipitation to the Indian subcontinent.
- Multiannual fluctuations can also occur in the ocean–atmosphere system—as for instance the 'El Niño Southern Oscillation' (ENSO) affecting the Pacific but having dramatic effects on the water balance of the whole tropical zone.

In the much longer term (10–10^4 yr), cyclic variations in solar activity and Earth's orbital variations modify the insolation and the seasonal contrasts between both hemispheres, and therefore the climate and water cycle patterns. Ice caps and oceanic sediments provide paleoclimatic archives allowing us to date such fluctuations and characterize their hydrological impact, especially during the past glaciation–deglaciation cycles.

Precipitation generally exceeds evaporation on continents, but both are lower than on oceans at the same latitude, because less energy (higher albedo) and water (unsaturated soils) are available on continents. Water in excess within continental basins is collected by aquifer and river networks and returns in great part to the ocean as river runoff, with inputs spatially concentrated at the river mouths. The relative contribution to total river runoff to oceans is 16% for the Amazon alone, 27% for the five major river systems (draining 10% of continental areas) and 50% for the 50 major rivers, while 40% of river runoff is concentrated in the equatorial region (flowing mainly into the Atlantic). Endorheic basins, that are not connected to the ocean, represent 22% of continental areas (mainly in dry regions): runoff disappears by infiltration or accumulates in closed lakes and seas, where water evaporates—as in Lake Chad or the Aral Sea.

River runoff varies during the year with usually a strong seasonality of high and low flows, and successions of relatively short floods separated by longer recession periods—with the extreme case of ephemeral streams. Runoff variations are controlled by the time and space variabilities of precipitation and evaporation, which are very different. Precipitations are usually episodic and intermittent, more or less concentrated during rainy seasons and irregularly distributed in space, while the energy available for evapotranspiration is much more regularly and continuously distributed but more affected by diurnal and seasonal cycles. Evaporation intensity, which is limited by the insolation, varies within a much smaller range (0–10 mm day^{-1}) than precipitation intensity, whose maximal recorded local values are greater than 40 mm in 1 min, 400 mm in 1 h, 1900 mm in 1 day, 26 000 mm in 1 yr. Precipitation is often buffered seasonally in snow cover and glacier at high latitudes or altitudes, while evapotranspiration depletes soil moisture and is often reduced by summer water stress. The interannual variability is much larger for precipitation than for evapotranspiration. There is a large variety of river runoff regimes as a result of combinations of precipitation and evaporation regimes on their basins.

Water cycle on the basin scale

As the oceanic and atmospheric branches of the water cycle are presented elsewhere (see OCEANS and EARTH'S ATMOSPHERE), only its continental branch is detailed here, especially on the basin scale. A basin (or catchment) is defined as the total drainage area of a river at some specified outlet. Catchments, which can be defined at any scale and in any environment, are functional units very useful for both the study and the management of water resources.

The water balance of a catchment expresses the principle of water conservation: the incident precipitation is redistributed between runoff, evapotranspiration and water storage within the catchment, through a competition between three main sources of energy and driving forces: (1) gravity, which controls most flows of liquid and solid water and depends mainly on topography and morphometry; (2) the evaporative demand of the atmosphere, which controls vapor fluxes and depends on energy input (radiative, advective) and climatic conditions; (3) capillarity, which controls soil water retention and depends on the pore size distribution. At any point and any time, the dominant form of energy and thus the dominant component in the water balance depend on soil moisture conditions.

There is a large variety of surface and subsurface processes and pathways governing the hydrological response of a catchment to water and energy inputs from the atmosphere. They result in a strong vertical and lateral redistribution and recycling of precipitated water within a catchment. The main fluxes (evapotranspiration, groundwater recharge, streamflow) are usually generated simultaneously or successively by several processes, whose combinations are very variable in time and space. They are indeed controlled by local characteristics and

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

variations of atmospheric inputs, by the water storage and transfer properties of the various compartments and their geometry. Because of many functional thresholds, catchment response is usually highly non-linear, much depending on initial water conditions.

After some interception by vegetation, precipitation reaches the soil surface: snow remains stored on the surface until it melts, while rain infiltrates into the soil and recharges its water storage. When soil moisture exceeds some threshold depending on both soil texture and structure, water percolates further vertically to recharge groundwater. Biological activity, soil cracks and bedrock fissures, chemical dissolution (karstic cave systems,) often create networks of non-capillary macropores, which provide preferential pathways for free water, bypassing the less permeable bedrock or soil matrix. If the percolating flux exceeds the hydraulic conductivity of some subsurface layer, water in excess forms a perched groundwater or flows laterally as hypodermic interflow. Groundwater flows are essentially lateral and controlled by the conductivity of the porous media and the slope of the piezometric surface (which is the water table slope for unconfined aquifers). If the water table is shallow, precipitation can create locally rapid groundwater ridging by saturating its capillary fringe.

If precipitation intensity exceeds the infiltrability (depending on surface conditions), the soil surface is saturated from above: water in excess accumulates in small depressions, then flows as diffuse surface runoff. If subsurface lateral flows exceed locally a threshold depending on soil slope, depth and conductivity, soil surface is saturated from below: the excess flux exfiltrates as return surface flow. Precipitation on already saturated areas also tends to flow as surface runoff. Whatever its generating process (infiltration–excess, exfiltration, saturation–excess), surface runoff tends to concentrate in small rills to reach the stream channel but can also infiltrate downslope, depending on local surface topography, roughness and infiltrability.

Surface and subsurface processes contribute to streamflow, provided that the generated fluxes are connected to a stream channel. While the slow drainage of groundwater and soil water supplies stream baseflow between successive rainy events, floods are generated by rapid surface and subsurface flows and by direct precipitation on the river. Streamflow is routed downward as a turbulent flux controlled by the bed slope and roughness. Within the channel network, it increases at each confluence, decreases at each diffluence and accumulates in lakes and artificial reservoirs. Along the channel, streamwater and groundwater exchange fluxes whose direction and intensity depend on their water levels and the bed's conductivity. Flooding occurs when streamflow exceeds some threshold fixed by the bed geometry.

Both surface and shallow subsurface waters are subject to evapotranspiration, returning water vapor to the atmosphere. Evaporation combines a phase change (liquid or solid water to vapor), which consumes much energy (radiative or thermal), and a turbulent vapor transfer within the air, which increases with air saturation deficit and wind speed. Transpiration of the vegetation is the vaporization within leaf stomata of the sap flow extracted from soil by roots. Evapotranspiration is limited by the water and energy availability: evaporation of free water (wet vegetation, surface water, groundwater) depends on the available energy only, while capillary retention makes water in unsaturated soil less and less available when soil dries out. During dry periods transpiration, which exploits water within the whole root zone, lasts much longer than evaporation, which affects the soil top layer only. Stomatal regulation limits transpiration in dry soil conditions, to avoid vegetation wilting.

Any catchment presents a functional, spatial and temporal organization, because of the non-uniform and non-random distributions of the hydrological processes, factors and inputs: (1) at each point, process activation or deactivation results from a balance between water supply from above and local water storage or transfer capacities; (2) spatio-temporal variations of factors lead to some recurrence of conditions favorable or unfavorable to each process in some areas of variable extent and some periods of variable duration; (3) these variable active areas and periods are also contributing areas to main outfluxes only if they are hydraulically connected to the catchment boundaries. Because of many possible interactions and processes, zones of interface between atmosphere, vegetation and soil, or between surface water and groundwater are very sensitive to any change.

Coupled water, energy and matter cycles

Water and energy cycles are intimately coupled in the atmospheric and oceanic general circulations, controlling the radiation and energy balances at both global and local scales, and carrying energy from equator to poles and water from ocean to continent.

- Atmospheric water vapor maintains a rather high temperature at Earth's surface by a 'greenhouse effect', while the net solar radiation on surfaces is reduced by an 'albedo effect' depending on water phase and content and generating local thermal contrasts: reflection is very high for low clouds and ice or snow cover, low for water bodies and higher for dry bare soils than for moist vegetated soils.
- At evaporation sites, cooling occurs by vaporization consuming much energy (solar radiation, sensible heat), which is stored as latent heat in the water vapor; both are transported within the atmosphere by vertical convection and lateral advection to the precipitation sites where warming occurs by condensation releasing the latent heat as sensible heat in the atmosphere.
- Huge amounts of energy are stored as sensible heat in ocean water and transported from warm to cold regions by ocean currents (for instance the great 'heat conveyor

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

belt'): the ocean acts as a thermal regulator, tempering also seasonal climatic contrasts in many continental regions.

Water and matter cycles are also strongly coupled. On continents, water is a powerful agent of transformation and vector of material, depending on climate, topography, bedrock and soil types and vegetation cover. Water in porous media (rocks, soils) increases their weathering by low-temperature biogeochemical alteration and mechanical fragmentation and reduces their mechanical resistance. Surface flows of water, snow and ice generate soil and river bed erosion and carry downward the eroded material, while subsurface flows transport solutes washed out of the soil matrix and favor mass movements on slopes (landslides, solifluxion). At river outlets, the ocean receives large amounts of fresh water and its solute and particulate load, which modify locally its salinity and induce mixing processes. These point inputs of matter from continents contribute to compensate for the chemical precipitation and mechanical sedimentation occurring in the ocean and to maintain constant its mean salinity. Transport of material by water is still much more discontinuous and irregular than water flow, with many possible cycles of erosion–sedimentation, dissolution–precipitation, concentration–dilution. Evaporation and freezing tend to concentrate water solutions, while precipitation and melting tend to dilute them, even though precipitated water can reach high concentrations in some elements through the washing out of the atmosphere. As a result, water quality varies all along the different branches of the water cycle. In the long term, water activity modifies the landscape, with some possible feedback on water pathways and flows.

Since life originated in the ocean about 3.4 billion years ago, water is essential in the biosphere also, where its cycle is strongly coupled with other biogeochemical cycles. Water is the main component of biomass and is involved in major metabolic reactions such as photosynthesis and respiration. In terrestrial ecosystems, the redistribution within plants of nutrients extracted by roots is ensured by the sap flow induced by transpiration, and the development of vegetation modifies the surface water and energy properties. Vegetation maps reflect well at all scales the patterns of climate and water balance—as can be clearly seen in Africa, with the latitudinal succession of equatorial rain forest, tropical savanna, semi-arid steppe and subtropical desert. In aquatic ecosystems, which have some ability for self-purification, fresh, brackish or salt water bodies are biotopes for many living organisms linked in complex trophic chains, structured according to the vertical and horizontal patterns of water's physicochemical parameters. Continental wetlands, estuaries and coastal regions are aquatic interface ecosystems very sensitive to any change in water quality and supply.

Humans and water

Water is an essential part of humanity's environment: it is one of the most important natural resources for human life and activities; it is also a risk factor because of human vulnerability to water-related hazards such as floods and droughts, avalanches and landslides, waterborne diseases and toxicities. Humankind has therefore always attempted to regulate the availability of water resources and to protect itself against such risks. Human activities have also many other major impacts on the water cycle and the coupled fluxes, directly by modifying continental surfaces and altering water quality and indirectly through the climate changes they could provoke. While the water resources available tend to decrease, there is a rapid increase in water demand by several uses competing for limited resources. The struggle for water is becoming a major source of regional conflicts, which could be avoided only by a strong international collaboration in order to increase the necessary solidarities between upstream and downstream regions.

Very fast population growth and economic development have dramatically increased in recent decades the demand for high-quality fresh water, which is at present estimated to be about 3750 $km^3\ yr^{-1}$: about 2/3 for irrigated agriculture (providing 50% of food production), 1/4 for industrial production and 1/10 for domestic use. Because of their easy accessibility, rivers, lakes and superficial aquifers are the main sources of renewable fresh water resources. Evaluated by the difference between continental precipitation and evapotranspiration rates, current mean global estimations of renewable fresh water resources vary between 36 000 and 45 000 $km^3\ yr^{-1}$, about 95% as river runoff and 5% as groundwater. This global amount is much larger than the present demand, and should also largely meet the expected increase of about 10% each 10 yr. Technological development should help reducing water demand.

Unfortunately, these resources are very unevenly distributed throughout the year and present large variations from year to year, as shown by the intense Sahelian drought during the 1970s. They are also very unevenly distributed over continents, with patterns that do not match the spatial distribution of population and economic activities. Significant fresh water deficit already occurs in many regions, affecting their social and economic development: 35% of the world's population has very low or catastrophically low water supplies (for instance in North Africa) and 75% lives in regions where water use exceeds 20% of available renewable water resources, leading to a mining-type exploitation of deeper groundwaters, which rapidly become depleted because of their much lower renewal rates. Disparity in water resources, population growth and economic development already generates a very large inequality in fresh water availability per person: more than 500 l day^{-1} in developed countries, less than 50 l day^{-1} in some arid regions. Water supply is already one of the key issues for humankind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Human activity also has a rapidly increasing impact on the natural behavior of most hydrosystems. River and aquifer regimes are often severely modified by water intake for irrigation, artificial reservoirs regulating discharge, or massive water exchanges between basins. Large-scale changes in land use (reafforestation–deforestation, urbanization, wetland drainage, field management) have deeply affected the precipitation redistribution between evapotranspiration, surface runoff and groundwater recharge. Irrigation with poor-quality water has generated soil salinization in many arid regions. The discharge in river courses of untreated domestic and industrial wastewater as well as groundwater recharge by polluted irrigation-excess waters have also drastically depleted water quality, further reducing the availability of good-quality fresh water. In the longer term, the possible global warming due to the anthropogenic increase in atmospheric carbon dioxide is likely to alter precipitation volume and seasonality, especially in arid regions, and to raise the sea level because of ocean thermal dilatation and ice cap melting: this would submerge many coastal lowland regions, which are among the most densely populated.

Conclusions

The water cycle is very complex at all scales, because of the numerous physical, chemical and biological processes involved and the high spatial and temporal heterogeneities of the controlling factors. Despite much progress recently achieved thanks to large interdisciplinary scientific collaborations and international effort to develop measurement and observation networks, there is still much to be done to understand and model all feedbacks and interactions with coupled energy and matter cycles. Impacts of increasing anthropic change and possible climate change on water cycle and resources are therefore still difficult to analyze and predict with some confidence. Nevertheless, the growing awareness of the numerous interactions controlling the water cycle and the scarcity and fragility of fresh water resources should contribute to improve the protection and management of both surface and subsurface water resources and the control of long-term effects of human activities on the Earth's water cycle. It is crucial for the future of both humankind and the 'blue planet'.

Bibliography

Dingman S L 1994 *Physical Hydrology* (Englewood Cliffs, NJ: Prentice-Hall)

Gleick P H (ed) 1993 *Water in Crisis—a Guide to the World's Fresh Water Resources* (Oxford: Oxford University Press)

National Research Council 1991 *Opportunities in the Hydrologic Sciences* (Washington, DC: National Academy Press)

Robinson P J and Henderson-Sellers A 1999 *Contemporary Climatology* 2nd edn (London: Longman)

Shiklomanov I A 1993 World fresh water resources *Water in Crisis—a Guide to the World's Fresh Water Resources* ed P H Gleick (Oxford: Oxford University Press) pp 13–24

Shiklomanov I A 1998 *World Water Resources: a New Appraisal and Assessment for the 21st Century* (Paris: UNESCO)

Shiklomanov I A (ed) *World Water Resources at the Beginning of the 21st Century* (IHP Monograph) (Paris: UNESCO) in preparation

Zebidi H (ed) 1998 *Water: a Looming Crisis? Proc. Int. Conf. on World Water Resources at the Beginning of the 21st Century (Paris, 3–6 June 1998) (IHP-V Technical Documents in Hydrology No 18)* (Paris: UNESCO)

Bruno Ambroise and Michel Vauclin

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wave–Particle Duality

The concept that electromagnetic radiation and subatomic particles behave in some respects like waves and in others like particles.

Various phenomena, such as interference and diffraction, clearly demonstrate the wave-like nature of light (and other forms of electromagnetic radiation). For example, if light of some particular wavelength, originating from a single source, passes through two adjacent narrow slits, each slit then acts as a source of light. Where the light waves spreading out from each slit meet, they interfere with each other, their amplitudes ('heights') adding together where they are in phase (e.g. where two wavecrests coincide) and canceling where they are completely out of phase (e.g. where a crest meets a trough). On the other hand, the particle behavior of light is demonstrated by the photoelectric effect, a phenomenon whereby certain substances emit electrons when illuminated by a beam of light, but do so only if the wavelength of the light is shorter than a particular minimum value. This behavior is consistent with light being a stream of particles (called photons) each of which carries a discrete quantity of energy that is inversely proportional to the wavelength of that light. If the energy of a photon exceeds the minimum energy that is needed to expel an electron, an electron will be ejected, but if the photon energy is less than this minimum, no electrons will be ejected.

Streams of subatomic particles exhibit wave-like behavior through phenomena such as interference and diffraction. For example, if a beam of electrons passes through two narrow slits before falling on a phosphor screen, which registers the arrival of each electron as a spot of light, the distribution of light spots takes the form of an interference pattern consistent with the pattern that would be produced by waves of a certain particular wavelength.

As Prince Louis Victor de Broglie (1892–1987) proposed in 1924, the wavelength (λ) associated with a particle of momentum, p, is given by $\lambda = h/p$, where h is the Planck constant (=6.63×10^{-34} J s). Since $p = mv$ (momentum = the mass of a particle multiplied by its velocity), $\lambda = h/mv$. The wavelength associated with a particle is known as the de Broglie wavelength. For example, the de Broglie wavelength for an electron of mass 9.11×10^{-31} kg traveling at a speed of 3×10^6 m s^{-1} (1 percent of the speed of light) is: $6.63 \times 10^{-34}/(9.11 \times 10^{-31} \times 3 \times 10^6) \approx 2.4 \times 10^{-10}$ m = 0.24 nm. For light of wavelength λ, the energy, E, of the photon is given by $E = hc/\lambda$, where c denotes the speed of light, and the momentum associated with a photon is $E/c = h/\lambda$

See also: electromagnetic radiation, electron, light, photon, quantum theory, subatomic particles, wavelength.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wavelength

The distance between two successive crests of a wave motion. Usually, in transverse waves (waves with points oscillating at right angles to the direction of their advance), wavelength is measured from crest to crest. In longitudinal waves (waves with points vibrating in the same direction as their advance), it is measured from compression to compression. The term is applied to electromagnetic radiation which is regarded as a wave motion. For example, blue light has a wavelength of about 440 nanometers and red light about 700 nanometers. X-rays have wavelengths of the order of 10^{-10} meters and radio waves of the order of meters. Wavelength is usually denoted by the Greek letter lambda (λ); it is equal to the speed (v) of a wave train in a medium divided by its frequency (f): $\lambda = v/f$.

See also: frequency.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wegener, Alfred Lothar (1880–1930)

German climatologist and geophysicist who suggested the phenomena of continental drift and plate tectonics. His idea was that a super-continent he called Pangaea had broken up, the pieces drifting to their present positions. His evidence was the fit of South America and Africa, and similarities in climate, fossil record and geology across the join. His path had been prepared by ALEXANDER VON HUMBOLDT and Frank Taylor, but his book, published in English in 1924, was scorned. His ideas are now widely accepted.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Weight

The force experienced by a body resting on, for example, the surface of a planet. A person standing on the Earth's surface experiences weight because the surface on which he is standing resists the effect of the force of gravity which otherwise would accelerate that person towards the center of the Earth. In other words, there is a reaction up through his feet equal and opposite to the gravitational attraction exerted upon him on Earth. The weight of a body depends upon the gravitational force to which it is subjected. On the surface of a planet it is equal to the mass of the body times the surface gravity. For example, a body that weighed 100 kg on the surface of the Earth would have the following weights on the surfaces of the bodies listed below:

the Moon	16 kg
Mars	38 kg
Jupiter	264 kg
the Sun	2790 kg
a white dwarf	30 000 000 kg

See also: zero gravity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Weizmann Institute of Science

At the Weizmann Institute of Science in Israel, astrophysics is practiced by a small theory group in the physics faculty. Main interests are high-energy astrophysics (currently mostly gamma-ray bursters, compact stars, and cosmic rays) and galaxy dynamics (alternatives to dark matter).

For further information see
http://www.weizmann.ac.il/physics/physics.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Werner, Johann (1468–1522)

Astronomer, mathematician, instrument-maker and geographer, born in Nuremberg, Germany, follower of Regiomontanus. Regiomontanus had suggested that the timing of eclipses and the orbits of comets could be used as clocks to determine longitude. Werner developed a practical version of this idea with the method of lunar distances (i.e. measuring the angle of the Moon from the Sun). He published this concept in *In Hoc Opere Haec Cotinentur Moua Translatio Primi Libri Geographicae Cl'Ptolomaei* (1514) and described an instrument with an angular scale on a staff from which degrees could be read off, to measure the lunar distances. This founded the movement to determine longitude by astronomy and led to the state-funding of observatories at Greenwich and Paris, the invention of the chronometer by JOHN HARRISON, and the accurate measurement of star positions and planetary orbits.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Westerbork Synthesis Radio Telescope

The WSRT (Westerbork Synthesis Radio Telescope) in Hooghalen, The Netherlands, is an aperture synthesis interferometer that consists of a linear array of 14 antennas arranged on a 3 km east–west strip. The array works by combining the signal from all the antennas and simulates a 3 km aperture telescope.

The antennas are equatorially mounted 25 m dishes with an f/D ratio of 0.35. This type of mounting ensures a fixed orientation of the receiving systems with respect to the sky. Ten of the telescopes are fixed, 144 m apart, while two nearby dishes are movable along a 300 m track and two others are on a 180 m track at a distance of 1.5 km. In the array, the baselines extend from 36 m to 3 km. The pointing accuracy of the dishes is 15 to 20 arcseconds, the surface accuracy is of the order of 1.7 mm.

The WSRT telescope with the first 12 antennas on a 1.5 km baseline was inaugurated in 1970. The second phase with two antennas on the 3 km extension was completed in 1980. A major upgrade due to be finalized in 2000 has added multi-frequency front ends capability and new state-of-the-art backends. The array can routinely operate at 92, 49, 21, 18, 13, 6 and 3 cm wavelength and change frequency in less than a minute. An observing capacity with instantaneous bandwidth up to 160 MHz is being added.

The WSRT is used by astronomers from The Netherlands but also from many other countries for a wide range of scientific research starting from our Galaxy to objects at the far reaches of the universe. There are two proposal deadlines each year. The WSRT also participates in the European (EVN) and Global Very Long Baseline Interferometry networks.

The WSRT is operated by the Netherlands Foundation for Research in Astronomy (NFRA), which is funded by the Netherlands Organization for Scientific Research (NWO).

For further information see
http://www.nfra.nl.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Whipple, Fred Lawrence (1906–)

Astronomer, born in Red Oak, IA, became director of the Smithsonian Astrophysical Observatory. While still a graduate student he helped compute the orbit of newly discovered Pluto. Using a new method of photography from two separated wide-angle cameras, he triangulated on meteor tracks and determined their orbits. He deduced that nearly all are made up of bits from comets. He proposed the 'dirty snowball' model for comets, suggesting that comets have icy cores inside layers of dirt. This was confirmed in 1986 when the Giotto spacecraft flew near to, and imaged, Comet Halley. Whipple tracked artificial satellites to determine the shape of the Earth.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Whirlpool Galaxy

A face-on, spiral galaxy, M51, which derives its name from its bold and clear-cut spiral pattern. Located at a distance of some 20 million light-years in the constellation of Canes Venatici, M51 (also known as NGC 5194), which has two well-defined arms that spiral out from its relatively small central bulge, is a classic example of a 'grand design' spiral. It is classified as an Sc galaxy in the Hubble classification scheme. Although, with a diameter of 65 000 light-years and a mass of about 5×10^{10} solar masses, it is somewhat smaller than the Milky Way Galaxy, it is several times more luminous, its spiral arms being laden with bright young clusters and HII regions.

M51 has a smaller, fainter companion (NGC 5195), which lies at the end of one of the larger galaxy's spiral arms. Although NGC has traditionally been classified as an irregular galaxy, it contains an elongated bar structure and hints of incipient spiral arms. NGC 5194 appears to be orbiting around M51 in a period of about 500 million years. Tidal interaction between the two, during their last close encounter, which took place some 70 million years ago, probably played a large part in establishing the bold spiral pattern in M51 and stimulating a vigorous bout of star formation within it.

M51 was the first galaxy to be recognized as having a spiral shape. This discovery was made in 1845 by William Parsons (1800–67), third Earl of Rosse, with the aid of the 1.8 m (72 inch) telescope that he had set up at Birr, in Ireland, and which, at that time, was the largest telescope in the world.

See also: barred spiral galaxy, Hubble classification, interacting galaxies, irregular galaxy, Messier Catalog, spiral galaxy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Whiston, William (1667–1752)

Born in Norton, Leicestershire, became the controversial third Lucasian Professor of Mathematics at Cambridge University, succeeding ISAAC NEWTON. He used the principles of the *Principia* in a popular book to explain the Book of Genesis and the Flood, which he envisaged as caused by the near approach of a comet, an interesting foreshadowing of current theories. He published in 1716 the first undergraduate course based on Newton's principles under the cumbersome title *Sir Isaac Newton's Mathematic Philosophy More Easily Demonstrated with Dr. Halley's Account of Comets Illustrated*. He wrote on numerous problems of astronomy and observed an aurora borealis, solar eclipses and sunspots. His downfall in the so-called Whiston's affair was due to his popularization of what were regarded as heretical ideas about the Trinity, theology and politics. He wrote letters to archbishops that raised eyebrows, caused debates in the Houses of Parliament and caused his dismissal from his post. Plunged into poverty, he lived off the income of a small farm near Newmarket and lectured in the coffee houses of London, giving scientific demonstrations for a fee as an entertainment.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

White Dwarfs

White dwarf stars, also known as degenerate dwarfs, represent the endpoint of the evolution of stars with initial masses ranging from about 0.08 to about 8 solar masses. This large range encompasses the vast majority of stars formed in our Galaxy and thus white dwarf stars represent the most common endpoint of STELLAR EVOLUTION. It is believed that over 95% of the stars of our Galaxy will eventually end up as white dwarfs. The defining characteristic of these objects is the fact that their mass is typically of the order of half that of the Sun, while their size is more akin to that of a planet. Their compact nature gives rise to large average densities and surface gravities.

The first glimpse of the existence of these objects came in 1844, with the study carried out by F BESSELL, the great German mathematician and astronomer, of the proper motion of SIRIUS, the brightest star in the night sky. The irregularities in the apparent motion of Sirius on the celestial sphere led Bessell to suggest the presence of an unobserved, solar-mass companion orbiting the bright primary. The companion, Sirius B, was first observed by A G Clark in 1862 and represents, together with 40 Eri B, the first known examples of white dwarf stars.

The luminosity of the white dwarf in the Sirius system is ∼10 000 times smaller than that of its bright, main-sequence companion. Its effective temperature was thus thought to be suitably smaller than that of the A star. However, the first spectrum of Sirius B, secured by W Adams in 1914, showed that the white dwarf had an effective temperature quite comparable with that of its companion. Since the stellar luminosity scales as $R^2T_{\rm eff}^4$, for the white dwarf to be both fairly hot and quite faint required it to have a very small radius, and a correspondingly very high average density. A similar conclusion was reached in the case of 40 Eri B whose spectrum was also secured at about the same time. However, how could such stars withstand the tendency to collapse onto themselves under the influence of their gravitational field?

The answer to that question came in 1925, when R H Fowler first applied the newly-developed principles of QUANTUM MECHANICS to stars. He showed that, in white dwarf stars, the density is high enough for the gas of free electrons to become degenerate. Electrons are said to be degenerate when a majority of them occupy the lowest possible energy states available to them. This occurs, at fixed temperature, when the electrons are packed sufficiently close to each other. Because of the PAULI EXCLUSION PRINCIPLE that no more than two electrons with oppositely directed spins may occupy the same energy state, the electrons retain kinetic motions even when cooled to zero temperature. The amplitude of this kinetic activity increases with increasing density, when electrons become more degenerate. In a white dwarf, the pressure generated by this kinetic motion, clearly of quantum mechanical origin, prevents the gravitational collapse of the star. The first detailed stellar models appropriate to white dwarfs were calculated in the 1930s by S Chandrasekhar, who received the 1983 Nobel Prize in Physics for his achievements.

Observed properties

Several properties of white dwarf stars can be determined fairly directly from observations. Analyses of their energy distribution, as well as of their optical and ultraviolet spectra, fix their effective temperature; they range from ∼150 000 K for the hottest stars to the coolest degenerate dwarfs, at $T_{\rm eff}$ near 4000 K. Spectroscopic analyses also yield the surface gravity ($g = GM/R^2$, but traditionally given in terms of its logarithm to the base 10, $\log g$) since the strength and width of spectral features are sensitive to the density of particles in the atmosphere, which is controlled by the surface gravity. The average surface gravity of white dwarf stars is $\log g \sim 8$ (cm s^{-2}), compared with $\log g \sim 4.4$ for the Sun. This value implies, as we discuss below, that the mean STELLAR MASS must be of the order of half that of the Sun. The luminosity range encompassed by white dwarfs exceeds 7 orders of magnitude and reflects the large range of observed effective temperatures ($L \propto T_{\rm eff}^4$); the faintest known white dwarfs have $L \sim 10^{-4.3} L_\odot$, while the rare intrinsically brighter ones, just entering the cooling sequence, reach $L \sim (10^2$–$10^3) L_\odot$.

Because of the intrinsic faintness of the most common white dwarfs, the observation of these objects tends to be restricted to small distances; the bulk of the sample of ∼2200 spectroscopically confirmed white dwarfs available today resides within ∼1 kpc from the Sun. It must be noted, however, that work with large, ground-based telescopes and with the Hubble Space Telescope now allows the observation of white dwarfs in GLOBULAR CLUSTERS located at several kpc from the Sun. Traditionally, however, white dwarfs have been culled from samples of objects showing significant PROPER MOTION, and thus located relatively nearby the Sun. A selection criterion based on color then allows the distinction to be made between true white dwarfs and nearby main-sequence stars, and their degenerate nature is then confirmed through spectroscopic observations. More recently, large numbers of hot white dwarfs have been detected in colorimetric searches. Selected fields are typically photographed through both U (∼3650 Å) and B (∼4400 Å) filters. A comparison of both images for a given object allows the selection of blue or very blue objects, whose nature can be confirmed through spectroscopic means. Surveys of this kind, carried out both at high Galactic latitude and in the plane of the Galaxy, have yielded substantial numbers of new hot white dwarfs, selected only on the basis of their colors without regard for their proper motion.

The existence of homogeneous samples containing substantial numbers of white dwarfs allows analyses of the statistical properties of these objects. Their distribution in the Galaxy is consistent with that of an old disk population, with an estimated scale height of 250–300 pc. Their space density is of the order of 0.005 pc^{-3} for $M_{\rm bol} < 15$, and their

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

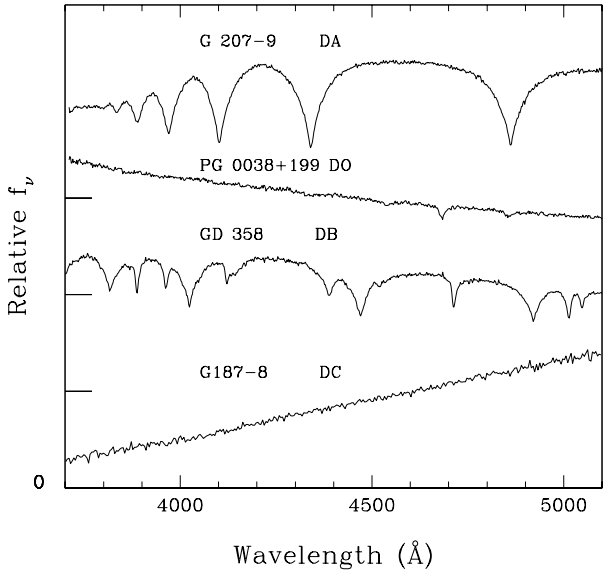


Figure 1. Sample optical spectra of white dwarf stars: from top to bottom, a DA spectrum, dominated by the regular pattern of the Balmer lines of hydrogen, seen here near their maximal strength, a DO spectrum, dominated by the lines of ionized helium, in particular the $n = 3$ to $n = 4$ transition at 4686 Å, a DB spectrum, featuring many strong lines of neutral helium, and a featureless DC spectrum, characteristic of a helium-rich star too cool ($T_{\rm eff} < 12\,000$ K) to show neutral helium lines.

birthrate is of the order of $(1.5\text{–}2.3)\times10^{-12}$ pc^{-3} yr^{-1} when allowance is made for the contribution of unseen white dwarfs in BINARY SYSTEMS.

The optical spectra of white dwarf stars are characterized by a rich variety which reflects, to a considerable extent, complex and varying patterns of atmospheric abundances. Sample spectra are shown in figure 1. About three-quarters of white dwarfs have a spectrum dominated by the Balmer series of hydrogen, which originates on the first excited, or $n = 2$, level of that atom. These objects are termed DA stars; they are found over the whole effective temperature domain of white dwarfs, from the very hottest stars above 100 000 K, where the hydrogen lines are fairly shallow, to the coolest white dwarfs near 4000 K, where the lines are sharp and extremely weak. Another important group of objects have optical spectra dominated by the lines of helium. The spectrum reflects the vastly different chemical makeup of their atmosphere, which is dominated by helium. At high effective temperatures ($T_{\rm eff} > 45\,000$ K), the lines observed are those of ionized helium. These stars are called DO stars. Between 30 000 K and 12 000 K, neutral helium is the dominant ion, and the spectrum is dominated by a bevy of transitions from that atom; these are the DB stars. Finally, below 12 000 K, the effective temperature is too cool for a helium-rich atmosphere to show transitions of neutral helium; the spectrum is then featureless, and is termed DC.

The scheme of spectral classification of white dwarfs is flexible enough to accommodate several other important classes of objects, not illustrated in figure 1; hence the DQ stars, whose spectrum is characterized by the presence of carbon, generally in molecular form, in a helium-rich atmosphere too cool to show lines of neutral helium. Similarly, the DZ stars show lines of heavy elements other than carbon, for example calcium, magnesium or iron. Sometimes, the star is hot enough for weak carbon or heavy element features to be present simultaneously with the dominant neutral helium lines: the stars are then classified DBQ or DBZ stars, respectively. At high effective temperatures, near $T_{\rm eff} \sim 100\,000$ K, another class of objects, termed PG 1159 stars, includes stars with spectra featuring ionized helium as well as highly ionized carbon and nitrogen features.

As hinted above, the chemical composition of the atmosphere of a white dwarf star is intimately related to its spectral appearance. The hydrogen line DA stars have atmospheres where hydrogen is the dominant element, to the near complete exclusion of any other. For their part, the so-called non-DA stars, which encompass objects of the DO, DB, DC, DQ and DZ spectral types, all have helium-dominated atmospheres. Within both classes, this purity is understood today as the result of gravitational settling in the intense gravitational field of the white dwarf. Under its influence, all elements heavier than the dominant atmospheric constituent rapidly sink into the deep atmospheric layers of the stars and remain out of sight. In fact, this settling mechanism is so efficient that the presence of any element heavier than hydrogen in the atmospheres of DA stars, or helium in those of non-DA stars, constitutes a puzzle. For example, lines of heavy elements (in the DAZ stars) or of helium (in the DAB stars) are occasionally seen in DA stars at effective temperatures below 25 000 K. Similarly, traces of heavy elements are observed in the DBZ and cooler DZ stars, while carbon is seen in the cool DQ stars. We discuss below the mechanisms which may permit the existence of such chemical impurities in the photospheres of white dwarfs.

Another point of interest for white dwarfs is their mass distribution. Only in a handful of cases are direct, dynamic masses obtained for these stars. In general, surface gravities are obtained either from the optical colors (e.g. on the Strömgren *uvby* system, suitable for DA stars with effective temperatures between 16 000 K and 8000 K) or, preferably, from optical spectra which tend to be more sensitive to surface gravity. The M/R^2 relation provided by the log g determination can then be coupled to an assumed mass–radius relation, characteristic of the mechanical structure of these stars (see below), to determine M. The mass distributions (based on optical spectroscopy) available for samples of two spectroscopic subclasses of white dwarfs are shown in figure 2. The mass distribution of the DA stars is sharply peaked, and the mean mass is $M = 0.590 M_\odot$, with a dispersion $\sigma = 0.134 M_\odot$. Tails extend at both ends of the mass spectrum: the low-mass tail of the distribution is thought to be

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

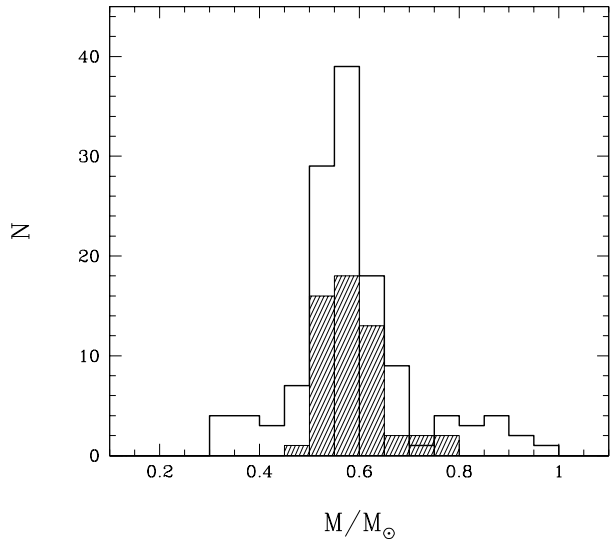


Figure 2. Mass distribution of DA (thick outline) and DB (shaded area) white dwarfs. The average mass appears the same, but the DB sample studied seems to lack both the low-mass and the high-mass end of the distribution seen for DA stars. These wings are probably related to binary evolution.

populated by low-mass white dwarfs with helium cores probably resulting from binary and common envelope evolution, whereas the high-mass tail could well be the result of merging events in binary systems. Nevertheless, the case of Sirius B ($M = 1.05M_{\odot}$), located in a wide binary system, shows that there are genuine high-mass white dwarfs which have not been produced by close binary evolution. It is worth noting, as well, that independent mass determinations for roughly three dozen DA stars can be obtained from the gravitational redshift of their Balmer lines. General relativity predicts that the wavelength of photons emitted in a strong gravitational field should be redshifted by an amount $\Delta\lambda$ which depends on the M/R ratio. Provided that the star under scrutiny belongs to an open cluster or a binary system with a well-determined radial velocity, and that the spectral line shifts induced by the high pressures of white dwarf photospheres are well understood, the gravitational redshift can be isolated, and the mass can be determined.

The mass distribution of the DB stars, while based on a smaller sample of well-studied objects, shows at the very least that there are no substantial differences between the mean mass of the DA and of the DB stars (the latter being $M = 0.585M_{\odot}$, with a dispersion $\sigma = 0.063M_{\odot}$). There is a tantalizing suggestion, however, that the low-mass and high-mass tails present in the DA sample may not be present in the DB sample; this absence accounts for the smaller dispersion of the DB mass determination.

While most of the white dwarfs identified today can be considered normal, garden-variety objects, two families of peculiar, and fascinating, white dwarfs have also emerged over the years. One such family is that of the magnetic objects, which currently includes nearly 50 stars. The detected magnetic fields range from 10^5 G all the way to 10^9 G and are measured through a variety of techniques which include the detection of circular polarization in the optical continuum ($B_e \geq 10^7$ G), the direct observation of the Zeeman pattern in absorption lines (10^6 G $\leq B_e \leq 10^7$ G), and Zeeman spectropolarimetry for low fields ($B_e \leq 10^6$ G). In the high-field objects, the spectrum is so perturbed by the magnetic field that spectroscopic classification is often difficult if not impossible. The inferred field morphologies tend to be rather simple and generally range from centered or offset dipoles to quadrupoles. Statistical arguments favor an evolutionary link between the chemically peculiar main-sequence Ap and Bp stars and the magnetic white dwarfs. The other distinct subgroup of interesting white dwarfs is that of the variable white dwarfs, or pulsators, examples of which are known among the hot PG 1159, the DB and the DA stars. The usefulness of these objects for our sounding of the internal structure of white dwarfs is discussed briefly below, as well as in the entries on ZZ CETI STARS and DB PULSATING STARS.

Perhaps somewhat surprisingly, white dwarfs appear to be slow rotators, in the sense that they rotate more slowly than would be expected if the ANGULAR MOMENTUM present in earlier evolutionary phases had been conserved. Determinations of the rotation velocity of white dwarfs rely on three different techniques: (i) measurement of the additional broadening of the core of absorption lines caused by rotation; (ii) measurement of the variation of the degree of circular polarization caused by rotation in magnetic white dwarfs; (iii) measurement of the rotational splitting present in the Fourier spectrum of the light curves of pulsating white dwarfs. While the use of any of these techniques is restricted to small subsamples of white dwarfs, the picture which emerges is one where substantial quantities of angular momentum must have been lost by stars on their way to the white dwarf stage. While no definite model exists to explain this loss, current ideas focus on the fact that most stars will need to shed a lot of mass, perhaps concomitant with angular momentum, before they do become white dwarfs.

Origin and generic properties

It is generally believed that the immediate progenitors of most white dwarfs are nuclei of PLANETARY NEBULAE, themselves the products of intermediate- and low-mass main-sequence evolution. As mentioned above, stars that begin their lives with masses less than about $8M_{\odot}$, that is the vast majority of them, are expected to become white dwarfs. Among those which already have had the time to become white dwarfs since the formation of the Galaxy, a majority have burned hydrogen and helium in their interiors. Consequently, most of the mass of a typical white dwarf is contained in a core made of the products of helium burning, mostly carbon and oxygen. The exact proportions of C and O are unknown because of uncertainties in the nuclear rates of helium burning.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The observed narrow mass distribution of isolated white dwarfs discussed above is a remarkable property of this category of stars. Apparently, the process of mass loss in white dwarf progenitors, which may have a wide range of initial masses, is regulated by mechanisms that are tuned finely enough to leave remnants with similar masses consistently. Also, the empirical evidence suggests that small amounts of helium and hydrogen are left over after the mass-loss phases have subsided. Taking into account previous thermonuclear history and the efficiency of gravitational settling, the expected structure of a typical white dwarf is that of a compositionally stratified object with a mass of $\sim 0.6 M_{\odot}$ consisting of a carbon–oxygen core surrounded by a thin, helium-rich envelope itself surrounded by a hydrogen-rich layer. Such an object has an average density of $\sim 10^6$ g cm^{-3}, a millionfold that of a normal star such as the Sun. The thicknesses of the hydrogen and helium outer layers are not known *a priori* and must depend on the details of pre-white-dwarf evolution. On theoretical grounds, however, it is expected that the maximum amount of helium that can survive nuclear burning in the hot planetary nebula phase is only 10^{-2} of the total mass of the star and that the maximum fractional amount of hydrogen is about 10^{-4}. Although these outer layers are very thin, they are extremely opaque to radiation and regulate the energy outflow from the star. They consequently play an essential role in the evolution of a white dwarf. The question of the exact masses of the hydrogen and helium layers present in white dwarfs constitutes a topic of intense research interest in the field.

The large opacity of the outer layers of a white dwarf implies that radiation escaping from the star originates from the outermost region—the atmosphere—which contains, typically, less than 10^{-14} of the total mass of the star. Spectroscopic and photometric observations can only probe these outer regions, which are usually dominated by hydrogen. Thus, a majority of white dwarfs are referred to as hydrogen-atmosphere objects (or DA stars as discussed above). It turns out, however, that about 25% of the known white dwarfs do not possess such a hydrogen layer. These are called helium-atmosphere white dwarfs (or non-DA stars) with, again, the understanding that the underlying carbon–oxygen core must contain essentially all of the mass, even though it is not directly observable.

Mechanical structure and cooling

As former nuclei of planetary nebulae, most white dwarfs are born in the form of extremely hot, collapsed objects which can only cool off: their nuclear energy sources are depleted, and gravitational energy can no longer be tapped efficiently as degenerate electron pressure prevents additional contraction. Because this pressure is independent of the temperature, a white dwarf is condemned to evolve at essentially constant radius. The mechanical structure of such a star is therefore specified by the degenerate electrons. In particular, electron degeneracy is directly responsible for the curious relationship between the mass and the radius of a white dwarf: the more massive the star, the smaller its size is. Likewise, relativistic degeneracy is also responsible for the existence of a limiting mass above which a white dwarf cannot exist. This limiting mass is known as the Chandrasekhar mass, and is of the order of $1.4 M_{\odot}$. The effects of this peculiar mass–radius relationship are apparent in the diagram presented in figure 3 which shows the cooling tracks of five representative DA white dwarf models that differ in their total mass.

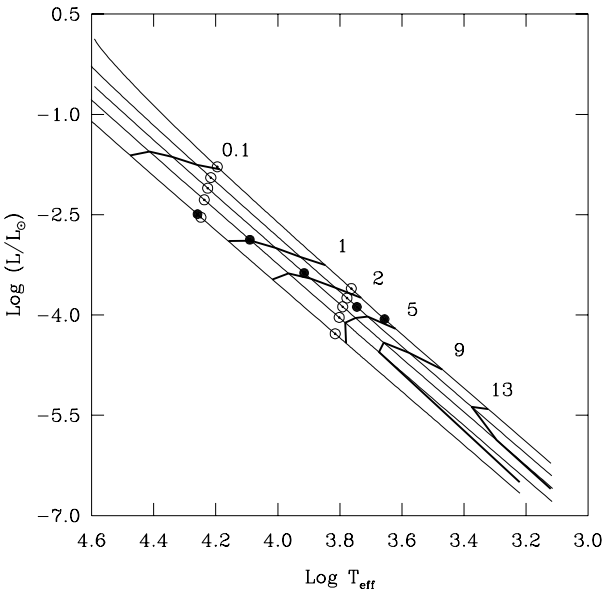


Figure 3. Evolutionary tracks (solid curves) of five ($M = 0.4 M_{\odot}$, $0.6 M_{\odot}$, $0.8 M_{\odot}$, $1.0 M_{\odot}$ and $1.2 M_{\odot}$, from top to bottom) representative models of DA white dwarfs in the Hertzsprung–Russell diagram. Each model is a compositionally stratified object made of a pure carbon core surrounded by a pure helium envelope containing 10^{-2} of the total mass of the star and an outermost pure hydrogen layer containing 10^{-4} of the total mass. The thick solid curves are isochrones, that is, loci of constant evolutionary timescales. The number next to each isochrone gives the cooling time in units of 10^9 yr. The small filled circles indicate the onset of crystallization at the center of each evolving model. The open circles at high (low) luminosity indicate the onset of superficial convection (the convective coupling of the surface with the thermal core).

Degenerate electrons also possess another property of high relevance for white dwarfs: they are excellent conductors of heat, and thus they thermalize the internal regions of white dwarfs efficiently (a familiar illustration of this property is provided by the conduction electrons in ordinary metals). We can thus envision a typical white dwarf as consisting of a nearly isothermal core that contains more than 99% of the mass, surrounded by a thin, opaque, insulating, nondegenerate outer envelope. In the range of effective temperatures 16 000–8000 K, where the bulk of the known white dwarfs is found, core temperatures vary from $\sim 2 \times 10^7$ K to $\sim 5 \times 10^6$ K. The very large temperature drop between the central regions

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and the surface takes place mainly in the stellar envelope. In the cooler models, this temperature gradient leads to the formation of superficial convection zones, similar to those found in the Sun. When present, atmospheric convection plays an essential role in the determination of the emergent flux from a white dwarf and, thus, in the interpretation of its spectrum and colors. The small open circles at high luminosity in figure 3 show the onset of a superficial convection zone due to the recombination of hydrogen in evolving models of DA stars. Convection plays also a key role in the subsequent evolution of a white dwarf by affecting directly the cooling rate. This occurs when the base of the superficial convection zone reaches into the degenerate interior, thus coupling the surface with the core and, thereby, increasing the rate of energy transfer across the outer opaque envelope beyond what is possible through radiative transfer alone. The small open circles at low luminosity in figure 3 show where convection starts to play a significant role in the cooling process.

Largely decoupled from the electrons, the (nondegenerate) ions provide the thermal energy that slowly leaks through the outside, thereby producing the star's luminosity. Thus, an isolated white dwarf shines at the expense of its thermal reservoir. In this context, the electrons do not contribute significantly to the energy reservoir because degenerate particles, already occupying their states of lowest energy, cannot be cooled. As thermal energy is gradually lost from the star in the form of radiation, the kinetic motions of the ions lose amplitude, become correlated, and the ionic state evolves from a gas to a fluid to a solid. Ultimately, the reservoir of thermal energy becomes depleted, and the star disappears from sight in the form of a cooled-off, crystallized object known as a black dwarf.

It should be noticed that the timescale for this final demise is strongly dependent on the total mass of the white dwarf. Indeed, because of their larger masses and smaller radii, more massive stars have larger internal densities (for comparable temperatures) and, therefore, develop a crystallized core earlier, at higher luminosities. By the same token, they also reach earlier the state where the specific heat in the solid regime plunges to very small values, a phenomenon well explained within the framework of the simple Debye theory of solids in quantum statistical mechanics. In effect, matter under these conditions has lost its ability to store thermal energy, the energy reservoir of the white dwarf has become empty and the star must then disappear from sight in a relatively rapid and final phase sometimes referred to as 'Debye cooling'. Figure 3 illustrates, among other things, how Debye cooling is a strong function of the total mass of the white dwarf. It is easily seen that the isochrones are strongly dependent on the total mass; crystallization and subsequent Debye cooling are responsible for the 'accelerated' evolution of the more massive models at low luminosities. To complement this, figure 4 provides more details on the evolving structure of a representative white dwarf model in a phase diagram.

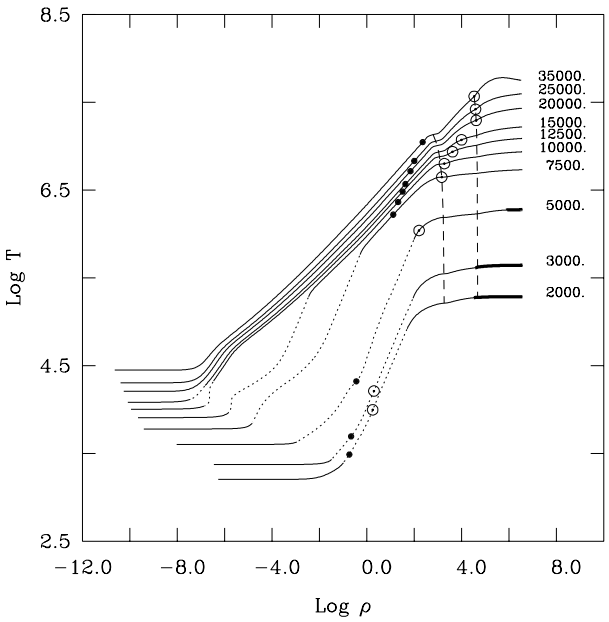


Figure 4. Evolving structure of a representative model of a DA white dwarf in a phase diagram. This is one ($0.6M_{\odot}$) of the five models presented in the previous figure. Each curve corresponds to the density–temperature distribution from the surface to the center of the model at an effective temperature given by the number alongside. The solid, dotted, and thick solid (for the three cooler epochs) portions of each curve indicate the radiative–conductive, convective and crystallized regions. Electrons become degenerate to the right of the small filled circle on each curve. Likewise, the small open circle on each curve indicates the location where the ions become strongly correlated (fluid phase). The dashed curves define the composition transition zones, H–He at lower densities, and He–C at higher densities.

A final point about the cooling evolution of white dwarfs should be made. In the early, short-lived phase of evolution following immediately the planetary nebula phase, white dwarf interiors are still hot enough that NEUTRINOS can be formed in great quantities there through a number of processes involving the electromagnetic and the weak interactions. The vast majority of the neutrinos escape directly from the central regions where they are created to the outer space, thus contributing to an important stellar energy sink. For instance, neutrino luminosities may become 2 orders of magnitude larger than photon luminosities in these objects. The evolution of a very hot, young white dwarf is thus dominated by neutrino cooling. Neutrino processes largely specify the cooling timescale and lead as well to a temperature reversal in the stellar core. Such a reversal is still visible in the hotter model shown in figure 4. By the time a white dwarf has cooled down to $T_{\rm eff} \sim 25\,000$ K, however, the star has lost its memory of the neutrino cooling phase, and its subsequent evolution and structure depend exclusively on the properties of its degenerate electrons and thermal ions.

It should be clear from this discussion that there

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

exists an intimate relationship between dense-matter physics and the structure and evolution of white dwarfs. For instance, a detailed knowledge of the opacity and thermodynamics of strongly coupled plasmas is necessary to compute the cooling rate of a white dwarf. Indeed, this rate basically depends on how much thermal energy is stored in the interior of a star and how rapidly this energy is transferred from the hot core to the cold interstellar medium through the thin, opaque outer layers. Hence, a reliable description of the constitutive properties of dense plasmas is required to build a *theory* of evolving white dwarfs. By the same token, the *observed* properties of cooling white dwarfs can be used, in principle, to test theories of strongly coupled plasma physics.

Spectral evolution

There is strong observational evidence that spectral evolution takes place among white dwarfs, that is, some of the hydrogen-atmosphere stars become helium-atmosphere objects, and vice versa, during various evolutionary phases. Indeed, the ratio of DA to non-DA white dwarfs changes as a function of effective temperature along the cooling sequence. In particular, the cases for the existence of a so-called DB gap—an interval of effective temperature from 45 000 K to about 30 000 K in which *no* helium-atmosphere object has been found—and of a cooler and narrower non-DA gap between ∼6000 K and ∼5000 K are well documented. Completely convincing explanations for these phenomena (especially in the case of the cooler gap) have not been worked out yet, but the very existence of 'holes' in the distribution of helium-atmosphere objects as a function of effective temperature is a strong empirical proof that, at least, some of the white dwarf stars must change their superficial chemical composition from helium-dominated to hydrogen dominated and back to helium rich again as cooling proceeds. It is suspected that a complex interplay between mechanisms such as hydrogen and helium separation (through diffusion) and convective dilution and mixing is responsible for the fact that a white dwarf may show two different 'faces' along its cooling track. For example, in DA stars below 15 000 K, convective mixing is thought to be effective in bringing deep-lying helium to the surface and drastically changing the surface composition. Unfortunately, at these effective temperatures, the helium brought to the surface is spectroscopically invisible, and its presence must be inferred through rather indirect analyses. The case for spectral evolution is further strengthened by another observational datum of importance: as pointed out above, there is no significant difference between the mass distribution of DB white dwarfs and that of the DA white dwarfs, excluding those objects that have been formed through binary evolution. This is what one would expect for stars changing only surface compositions. This being said, while spectral evolution appears unavoidable, it cannot currently explain all the peculiarities of the abundance patterns observed in white dwarfs. Alternative schemes, for example different channels feeding the cooling sequence, should be considered in order to account completely for the rich variety of spectral types.

White dwarf stars are not only unusual in that the abundance of their main atmospheric constituent (hydrogen or helium) may change in a complex way as a function of time, but they also show an amazing variety of heavier trace elements in their atmospheres, arguably making them the most fascinating of all chemically peculiar stars. This is because the cooling phase of white dwarfs, a relatively uneventful phase from an evolutionary point of view as discussed above, is, in contrast, a most active phase for the evolution of the chemical composition of the envelope. Indeed, it is now well established that the often puzzling variety of surface abundances observed in white dwarf stars can be traced to the simultaneous operation, in the outer layers of these stars, of a variety of physical processes which will also erase the abundances present in the photosphere at the onset of cooling. As discussed above, downward element diffusion in the intense gravitational field of a degenerate star is perhaps the mechanism which is the most closely identified with white dwarf stars. However, in its presence, the observed abundances of various atmospheric impurities, while small, are much too large to be accounted for.

Mechanisms which compete with the downward settling must thus be called on to explain the presence of these impurities in the atmospheres. At high effective temperatures, say $T_{\rm eff} > 20\,000$ K, the dominant competing mechanism is thought to be the selective radiative support of elements in the atmosphere. This mechanism, which involves the transfer of momentum from the intense radiation field to ions of heavy elements such as carbon, nitrogen, silicon, iron and helium, is able to counteract the downward gravitational force exerted on these ions and allows a small, but measurable, amount of impurities to remain in the atmosphere. Unfortunately, detailed calculations of this radiative support still do not reproduce the observed abundances, and it is currently thought that a small mass loss rate (of the order of $\sim 10^{-13} M_{\odot}$ yr^{-1}), when coupled to radiative levitation, might be able to account for the observed abundance patterns. For cooler stars, it is believed that ACCRETION from the interstellar medium plays a role in accounting for the traces of heavy elements occasionally seen. Individual accretion events, probably related to encounters between a white dwarf and a small patch of neutral gas in the interstellar medium, might be able to account for the small fraction of DA and non-DA stars which display impurities in their spectra. In the DQ stars, it is thought that the convection in the helium envelope is deep enough to dredge up traces of carbon from the deeper carbon-rich layers. Much energy has been expended, in the last two decades, to untangle the relative importance of these competing mechanisms and to decipher the complex patterns of photospheric abundances observed in white dwarfs. Nevertheless, much work remains to be done.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Some areas of current interest

In the era of the Hubble Space Telescope and of giant, 8–10 m ground-based telescopes, faint white dwarf populations are being routinely discovered and studied in open and globular clusters. In addition, evidence has been provided for the existence of a very old, faint population of white dwarfs in the Galactic halo on the basis of an interpretation of some MICROLENSING events and, more directly, on the basis of the observed proper motions of some very faint objects in the HUBBLE DEEP FIELD whose positions in a color–magnitude diagram are consistent with very cool ($T_{\rm eff} \sim$ 2000–3000 K) DA white dwarfs. These exciting developments have led to a renewed interest in white dwarf cooling calculations and model atmosphere calculations using upgraded input physics and extending into the regime of very low effective temperatures. Such efforts are clearly worthwhile since white dwarf physics holds the potential for providing the best indicators of the ages of, and distances to, these systems.

The usefulness of white dwarfs as cosmochronometers was firmly established about a decade ago when it was first demonstrated that the white dwarf population in the solar neighborhood—a population characteristic of the Galactic disk—could be used to estimate the age of the disk to much improved accuracy. This method applies because (intrinsically) faint white dwarfs cool down extremely slowly (see the isochrones in figure 3). This implies that, if white dwarf formation has been going on more or less constantly over the distant past, many more faint white dwarfs than bright white dwarfs should be present in a given volume of space. This is indeed what the distribution of observed white dwarfs in the Galactic disk generally shows (figure 5). However, a very important and significant observational result of the last decade has been the realization that there is a significant decrease in the luminosity distribution of these stars: there is a real deficit of low-luminosity white dwarfs in the solar neighborhood. The simplest model to account for this observational fact is to assume that the oldest white dwarfs in the Galactic disk are still visible. In other words, the corpses of the very first generation of intermediate-mass stars formed in the disk of the Milky Way have not yet had the time to cool to invisibility, beyond the reach of our telescopes. By comparing the location of the observed low-luminosity decrease in the white dwarf distribution with cooling calculations, it is possible to infer the age of the white dwarf population in the Galactic disk (see figure 5 for an example of this). While the method continues to be refined through numerous numerical simulations of evolving white dwarfs in the context of the Galactic disk, its potential for the white dwarf populations in open and globular clusters as well as for the putative population in the Galactic halo is evident.

Another area of current active interest is related to the presence of so-called instability strips along the cooling sequences of white dwarfs in the HERTZSPRUNG–RUSSELL DIAGRAM. Helium-atmosphere white dwarfs become unstable against nonradial gravity-mode pulsations as they evolve through an interval of effective temperatures from ∼25 000 K to about ∼22 000 K. The instabilities are intimately connected to the recombination of helium in the envelopes of these stars and the concomitant formation of a superficial helium convection zone. There also exists an analogous instability strip for the hydrogen-atmosphere white dwarfs, related, this time, to the recombination of hydrogen in the outer layers, and, consequently, which is located at lower effective temperatures, in the range 12 500–11 000 K. In both cases, the pulsational instabilities manifest themselves as temperature waves at the stellar surface that cause multiperiodic luminosity variations. The pulsating DB white dwarfs are referred to as V777 Her stars, while their DA counterparts are known as ZZ Ceti variables. The importance of these two instability strips stems from the fact that they provide windows through which the internal structure of white dwarfs can be probed. For instance, a detailed comparison of the observed period structure of pulsating white dwarfs with those of models provides a unique way of inferring the internal constitution of white dwarfs and, in particular, the run of the chemical composition as a function of depth. Although the potential of this technique has been barely tapped, it is likely that asteroseismological studies of white dwarfs will soon become a major contributor to our knowledge of the internal structure of these stars. One of the major potential impacts of these studies is the determination of the thicknesses of the outer layers of helium and hydrogen (in

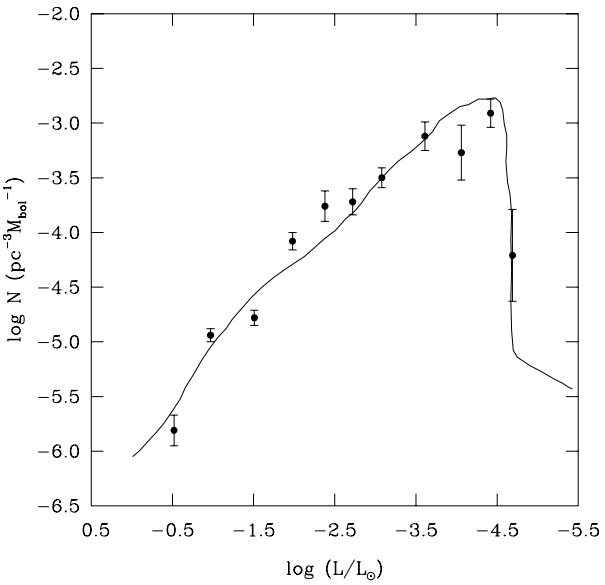


Figure 5. Luminosity function of white dwarfs in the solar neighborhood, that is the number of white dwarfs per unit volume per unit luminosity interval as a function of luminosity in solar units. The observed values are indicated by error bars. Note, in particular, the deficit of low-luminosity stars shown in the last bin. The solid curve is a fit to the data points based on cooling calculations which assume an age of 9.3×10^9 yr for the Galactic disk.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the DA stars) in individual objects, which will serve as calibration of evolutionary models in more refined attempts at the cosmochronology of various white dwarf populations.

Bibliography

D'Antona F and Mazzitelli I 1990 *Ann. Rev. Astron. Astrophys.* **28** 139
Liebert J 1980 *Ann. Rev. Astron. Astrophys.* **18** 363
Rudermann M 1971 *Sci. Am.* **224** (2) 24
Van Horn H M 1979 *Phys. Today* **32** 23
Wesemael F, Greenstein J L, Liebert J, Lamontagne R, Fontaine G, Bergeron P and Glaspey J W 1993 *Publ. Astron. Soc. Pac.* **105** 761

Gilles Fontaine and François Wesemael

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

White Dwarfs in the Galaxy's Halo

The Galaxy's large spherical halo (see GALACTIC METAL-POOR HALO and HALO, GALACTIC) may harbor as many as several hundred billion WHITE DWARFS, a population as large in number as the total number of stars in the Galaxy's disk (see DISK GALAXIES and GALACTIC THIN DISK). Although this assertion is controversial, several astronomical surveys provide strong support for it and the implications affect fields of astronomical inquiry as diverse as dark matter and star formation. The reason that this population of white dwarfs may be related to dark matter is tied to the fact that white dwarfs cool and fade as they age. This means that a white dwarf as old as the ancient halo is extremely faint. Current large-scale imaging surveys can only see relatively nearby white dwarfs, and only a tiny fraction of the halo's mass exists near the Sun, while the vast majority of the Galaxy's halo is thousands of light-years away.

The theory of gravity requires that the mass of the halo must dominate that of the Galaxy's disk. Since the halo is largely invisible, it is almost entirely made of dark matter (see DARK MATTER IN GALAXIES). Indeed, the known population of normal stars in the halo constitutes a trivial fraction of the total mass. However, those stars exhibit properties distinct from disk stars, including rapid motion relative to the Sun. As a result a halo star that happens to be within the solar neighborhood can be distinguished from a disk star by examining the rapidity with which it passes through the solar neighborhood. This is a statistical criterion, meaning that for any given star one can assign a probability that it belongs to the disk or halo. For a survey of many stars with such probabilities one can estimate the actual number of stars in the disk and halo populations.

The first hint that white dwarfs may be a significant part of the halo emerged in 1996, when a group conducting a MICROLENSING experiment claimed to detect indirectly a large population of objects in the Galaxy's halo all with masses characteristic of white dwarfs. Instead of actually seeing the white dwarfs themselves, they observed the effect they have on images of stars in the Milky Way's companion galaxy the Large Magellanic Cloud (see WIMPS AND MACHOS). The microlensing results, now backed by several independent, similar experiments, have found that at most 20% of the total mass of the Galaxy's halo may be in this supposed population of white dwarfs.

In 2001 a search for moving stars in an archive of photographic plates taken between 1950 and 1998 revealed about 20 nearby white dwarfs which are moving so fast that they must be members of the Galaxy's halo. Statistical analysis of the data showed that at least 1% of the halo's mass is due to a population of white dwarfs. Alternative interpretations of this survey suggest that a fraction of this population may be part of the Galaxy's thick disk. In all of the interpretations, however, there still appears to be an unexpectedly large number of white dwarfs in the halo, constituting at least 0.5% of the halo's mass.

With 0.5–20% of the halo's mass accounted for by white dwarfs, a fraction of the Galaxy's dark matter has finally been identified. Furthermore, the halo white dwarf population is important to STAR FORMATION theory. Ongoing star formation creates many low-mass stars for every high-mass star. If the same proportions of high- and low-mass stars formed 13 billion years ago, when the halo formed, only one to five halo white dwarfs should have been found by the recent survey. This is known because the low-mass halo stars have not had enough time to evolve into white dwarfs, and we know the number of low-mass stars in the halo. That number, assuming a universal star formation process, determines the number of high-mass stars that have evolved into white dwarfs in the halo. The actual numbers observed imply that high-mass stars formed more readily in the halo than they do at present.

Ben R Oppenheimer

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Whitford, Albert Edward (1905–)

Astronomer and photometrist, worked as a student with JOEL STEBBINS and succeeded him as professor at Wisconsin and director of the Washburn Observatory. Became director of the Lick Observatory. As a student he made a successful device for measuring very small currents from photoelectric cells, and used it to measure the light from stars and galaxies. This turned into his career as an astronomer.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Whittaker, Edmund Taylor (1873–1956)

Mathematician, born in Southport, Lancashire, became Astronomer Royal of Ireland, director of Dunsink Observatory and professor of astronomy at the University of Dublin. Wrote the influential *History of the Theories of Ether and Electricity, from the Age of Descartes to the Close of the Nineteenth Century* (1910).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Whole Earth Telescope (WET)

A world-wide network of cooperating astronomical observatories which was established by astronomers from the University of Texas in 1986. WET HQ moved to Iowa State University in 1997.

Intended to obtain uninterrupted time-series measurements of variable stars (white dwarfs and Delta Scuti stars) and cataclysmic variables. This is done by resolving the multiperiodic oscillations observed in these objects into their individual components. The temporal spectrum allows astronomers to probe the interiors of the target objects using the technique of asteroseismology.

Since the first campaign in March 1988, the WET team has coordinated global photometry campaigns so that the target objects are visible from the night side of the planet 24 hours a day. These campaigns typically take place twice a year.

For further information see
http://ceti.as.utexas.edu/wetpage.html.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Widefield CCD Imagers

Since the early 1980s, the CHARGE-COUPLED DEVICE, or CCD, has emerged as the primary detector for astronomical observations. At the present time, CCDs are virtually the only detectors used for observations in the optical portion of the spectrum (300 nm–1000 nm), and their use has been extended into the UV (200 nm–300 nm) and x-ray (100 eV–10 keV) regions as well. Modern CCDs are superb detectors. When thinned and backside illuminated, and with new multilayer anti-reflection coatings, CCDs can offer quantum efficiency (QE) exceeding 70% and approaching unity over most of the optical waveband. The best CCD amplifiers now have readout noise below $2e^-$ rms and some have achieved the elusive $1e^-$ noise level. One of the main areas of development has been, and continues to be, in device size and pixel count. The very first CCD imagers had 100×100 pixels, while the first astronomically useful device was the RCA 320×512 with 30 μm pixels, followed by the 800×800 15 μm pixel device built by Texas Instruments for the WF/PC-I on the NASA HUBBLE SPACE TELESCOPE. As the 1990s began, the sizes of these devices began to grow. 2048×2048 pixel devices were produced by Ford Aerospace (which eventually became Loral, Loral/Fairchild, and now Lockheed-Martin) and Tektronix (now SITe). These 2K×2K pixel, and similar-sized, devices are now produced with good yields by several manufacturers.

The technical advances in ground-based instrumentation have included increases in telescope apertures and improvements in image quality and SEEING. It is not unusual to obtain $0''.5$ images or better at premier astronomical sites. Consequently, optics and detector systems must be designed with sufficient resolution to sample these sharp images, requiring that the detector pixels be no larger than $\sim 0''.25$. This requirement for small pixels, coupled with the need for large fields, drives the high pixel count needed in widefield imagers. For example, to be optimally sampled while spanning a $0^\circ.5 \times 0^\circ.5$ field requires that the array measure of order 8000×8000 pixels. In the near future, the image quality is likely to improve even more, driving the pixel sizes smaller and the pixel count higher in order to maintain the same field of view.

Monolithic devices are not suitable for such very large focal planes. Limitations in wafer size and yield set a practical limit to device size. In addition, other factors such as finite charge transfer efficiency (CTE) and readout time, as well as cost, must be considered when the number of pixels increases. In order to build the very large detectors needed to image wide fields of view at seeing-limited resolution, one must construct mosaics of smaller devices.

Large-gap CCD mosaic focal planes

The simplest approach to constructing a large mosaic focal plane is to mount large, packaged devices as tightly as possible on a common base. The resulting mosaic will have large gaps (1 cm or more) between the packaged CCD imagers that come from the wire bond pads that are required to make electrical connections to the silicon die and the wire bond pins that are usually located around the perimeter of the typical device package.

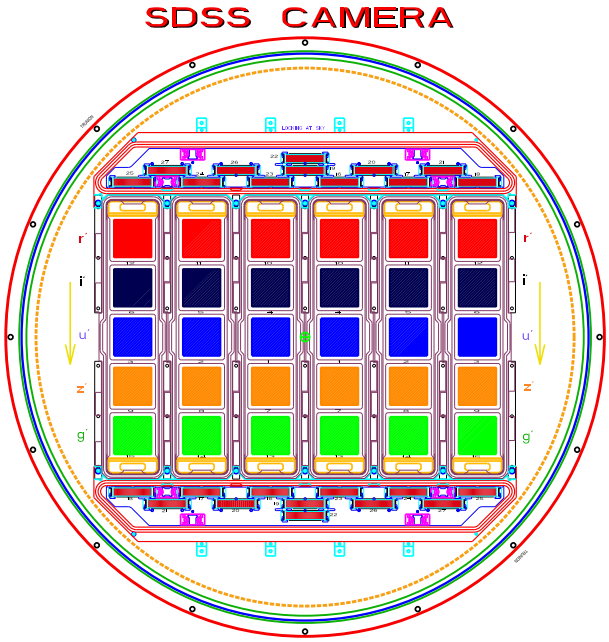


Figure 1. Drawing illustrating the layout of the SDSS CCD mosaic. The CCDs are arranged in six parallel columns each having five CCDs. Each CCD in a 'row' is covered by a different filter (ugriz). The imaged scene 'drifts' along the columns and the clocking of the charge is synchronized with the image motion. As a $2^\circ.5$ strip of sky is scanned, images in five colors are obtained from the camera. For this specialized application, the large gaps are not considered to be a serious problem.

Nevertheless, several widefield mosaic CCD imagers have been constructed using this technique, for example the 4K×4K pixel Big Throughput Camera constructed by the University of Michigan and Bell Labs, the 4K×8K mosaic imager built by the National Astronomical Observatory of Japan and the 10K×12K pixel camera built by Princeton University for the SLOAN DIGITAL SKY SURVEY (SDSS).

The SDSS camera (figure 1) consists of a 5×6 array of SITe 2K×2K CCDs with 24 μm pixels and is somewhat special in that it is designed to operate in the 'drift scanning' mode where the sky image is allowed to scan along the device columns while the accumulating photoelectron packets (or charge) are clocked to follow the image. This application differs from the conventional 'shift and stare' technique where the telescope is pointed at some region, an image is taken (typically only a few minutes integration), the telescope is shifted by an amount larger than the gaps separating the CCDs, another image is taken, and so on. The multiple images at different positions allow the gaps between the mosaic elements to be filled while also enabling the removal of cosmic rays that hit the device during the exposure.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

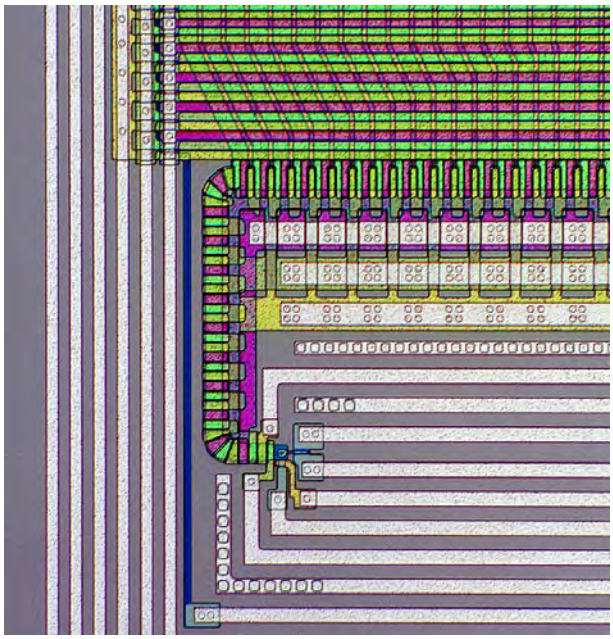


Figure 2. This photomicrograph of the lower left-hand corner of a three-edge buttable CCD illustrates several key design features for edge buttability. This particular device (a 2K×4K LL CCID-20 with 15 μm pixels) can be operated in framestore mode (if desired) so there are six metal bus lines distributing the three-phase parallel clocks to the imaging and storage regions. These can be seen as the vertical lines to the far left of the photograph. The saw cut would be made at the edge of this image. The last three or four rows are tapered inward, and the serial register is wrapped around vertically so the output amplifier does not increase the width of the device (photograph courtesy of Dr Barry Burke, MIT Lincoln Laboratory).

At one time, drift scanning may have offered some practical advantages such as continuous readout with no dead time, and improved flat fielding in one dimension (the pixel-to-pixel sensitivity non-uniformities in a column are effectively smoothed out by clocking the scene along a column). However, except for specialized projects like the SDSS, these advantages no longer exist. Modern mosaics can be read out quickly (under 1 min), so the readout time is only a small fraction of the typical integration time. Also, modern CCDs do not have the gross flat-field non-uniformities that plagued earlier devices, so the column smoothing obtained by drift scanning is not only unnecessary but viewed as a disadvantage. Furthermore, requiring a telescope and camera to operate in drift scanning mode places strict tolerances on the optical system field distortion because the image tracks cannot deviate from straight lines that must map onto device columns or else the image quality will deteriorate. Considering the advantages and limitations of both methods, the large majority of mosaic cameras have been designed to operate in the shift and stare mode.

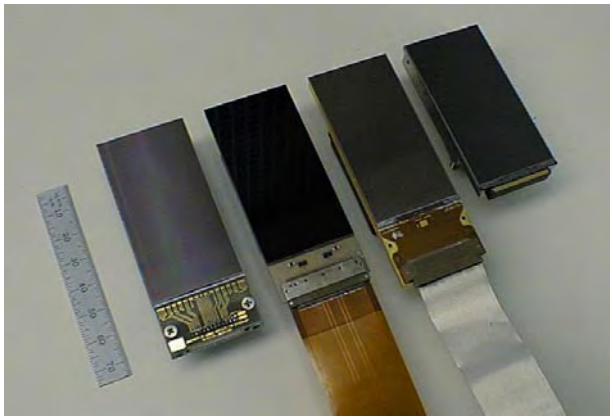


Figure 3. Shown are four 2K×4K three-edge-buttable CCDs with 15 μm pixels from four different manufacturers. These devices are the present-day building blocks for large, close-packed mosaics. From left to right are devices made by Loral (now Lockheed-Martin), MIT Lincoln Laboratory, SITe and EEV. Note that the EEV package has considerably less structure along the wire-bonding edge making it the preferred design for focal planes with more than two rows of devices.

Close-packed CCD mosaic focal planes

For most applications, a continuous focal plane with minimal gaps between the CCD imagers is preferred. While it is not possible to construct a mosaic with zero gaps, there are techniques in the CCD design and packaging that allow for the construction of close-packed mosaics where the gaps are of order 0.5 mm or less.

Device design considerations for close-packed mosaics

To enable CCD imagers to be used in close-packed mosaics, the devices must be specially designed from the outset to be edge buttable. This involves confining the wire bonding pads to only two or, better yet, one of the four edges of the device. Consider a three-edge buttable design. The only structures that need to run up the sides of the device are the metal clock bus lines that distribute the parallel clock signals to the polysilicon gates that establish the potential wells in the imaging region of the device. To keep the dead regions along sides of the device as small as possible, the output amplifier must be designed so it does not not protrude from the side of the device. This can be accomplished by turning a corner with the serial register and tucking the amplifier inside the edge boundary defined by the outermost clock lines (see the photomicrograph in figure 2). Allowing the smallest reasonable space for the vertical bus lines and the boundary to the saw cut, the minimum gap from the imaging area to imaging area on two adjacent arrays is $\sim$200 μm.

Currently, at least three manufacturers (EEV, SITe and MIT Lincoln Laboratory) are producing thinned, high-QE 2K×4K devices with pixels of sizes in the 10–15 μm range that are three-edge buttable and suitable for building large mosaic focal planes (see figure 3).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

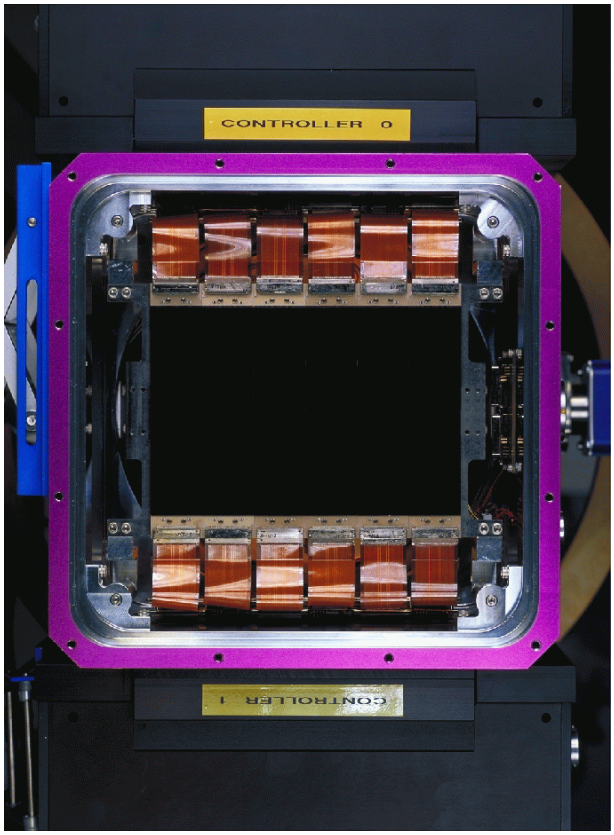


Figure 4. This 8K×12K pixel CCD mosaic camera built by the University of Hawaii for the 3.6 m Canada–France–Hawaii Telescope is the largest close-packed mosaic in operation at the time of this writing (larger mosaic focal planes are under construction). The mosaic consists of 12 thinned, high-QE 2K×4K three-edge buttable CCDs with 15 μm pixels. The devices were made by MIT Lincoln Laboratory. The imaging area, corresponding to the black rectangular region in the center of the photograph, measures 185 mm×122 mm. When used at the prime focus of the CFHT, the mosaic spans a field of view of $45' \times 30'$ at a scale of $0''.21$/pixel. The gaps separating CCDs are less than 0.5 mm or less than $10''$ on the sky.

Packaging and mosaic construction

Once an edge-buttable device is designed, it can be mounted close to its neighbors to form a CCD mosaic. The very first attempts at this involved permanently attaching (i.e. gluing) the CCD dies to a common substrate. There are several disadvantages to this approach. A mosaic assembled in this manner cannot be taken apart safely if one of the CCDs is damaged. Furthermore, the selection of the CCDs that will be used in a mosaic involves thorough testing at cryogenic temperatures. This is difficult to accomplish with unpackaged CCD dies, so one could permanently assemble a mosaic and find after it is completed and tested that some of the mosaic elements are not suitable for scientific observations. These problems led to the development of edge-buttable packages that do not compromise the tight packing ability of the buttable imagers and yet allow a mosaic focal plane to be assembled and dis-assembled. Examples of such packages are shown in figures 3 and 4. In choosing a package material, one must consider that the device will be operated at ∼170 K, so the coefficient of thermal expansion difference between that material and silicon is important. The standard package materials that are a reasonable thermal expansion match to silicon are aluminum nitride (AlN), Invar and molybdenum.

Using the devices shown in figure 3, several close-packed CCD mosaic focal planes containing at least 8K×8K pixels have been constructed and are in operation (e.g. UH 8K×8K, NOAO 8K×8K mosaic imagers for the Kitt Peak and CTIO 4m telescopes, ESO 8K×8K, CFHT 8K×12K—see figure 4).

Devices that are three-edge buttable are suitable for close-packed mosaics having two rows and an unlimited number of columns. If one wants to build a larger mosaic with more than two rows, one must tolerate a larger gap between the rows on the fourth, non-buttable edge. With these larger mosaics in mind, some designs have been developed for pseudo four-edge buttable packages that minimize the packaging structures along this fourth edge. EEV has developed a package (see figure 3) that keeps the dead space along this fourth edge to under 3 mm from the edge of the bottom imaging row to the edge of the package. Another prototype design with a 3 mm dead space is shown in figure 5.

The future

Widefield CCD imagers are at present being constructed by many observatories. Several current projects are designed to span fields in excess of $1^\circ \times 1^\circ$ using CCD mosaics measuring 18K×18K pixels. If widefield imaging telescopes with $3^\circ \times 3^\circ$ fields are constructed, they will require CCD mosaics having 36K×36K pixels or more. Work is also progressing on the designs for the individual building blocks that make up the mosaic elements. Three-edge buttable devices with 3K×6K 10 μm pixels are now under development, and wafer-scale devices that can fill a 150 mm wafer are being considered as yields improve.

On-chip image motion compensation—the OTCCD

One exciting CCD variation that has emerged in recent years is the multi-directional or orthogonal-transfer CCD (OTCCD)—a device that can shift charge in up to eight directions, thus allowing active image motion compensation or fast guiding 'on chip' (see figure 6). A mosaic array of close-packed, independently addressable OTCCDs would allow compensation for image motion caused by the atmosphere over very wide fields, as well as correcting for windshake and other mechanical pointing instabilities common to the telescope as a whole.

Large, buttable OTCCDs with 2K×4K pixels have been successfully fabricated and the first mosaics of such devices are at present under construction. Even larger devices subdivided into small ($\sim 1' \times 1'$) independently controllable OTCCD cells are under development.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

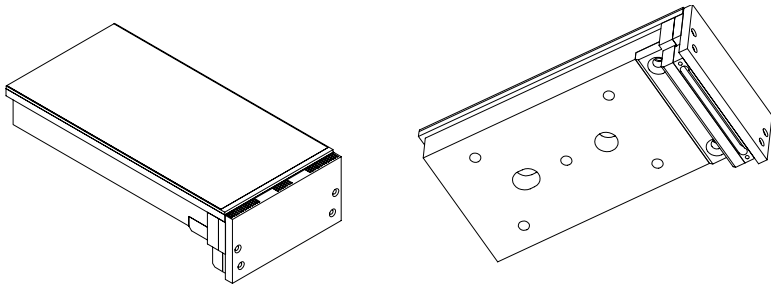


Figure 5. This drawing shows one design for a package that can be used in close-packed mosaics having more than two rows. A ceramic piece with metallized traces that wrap around one edge is mounted at right-angles to the device. Wire bonds are made to the traces on the short edge of the ceramic and the traces connect the signals to a surface mount connector that emerges from the backside of the package. Mounting holes and alignment holes are shown for locating the package in focal plane mosaic.

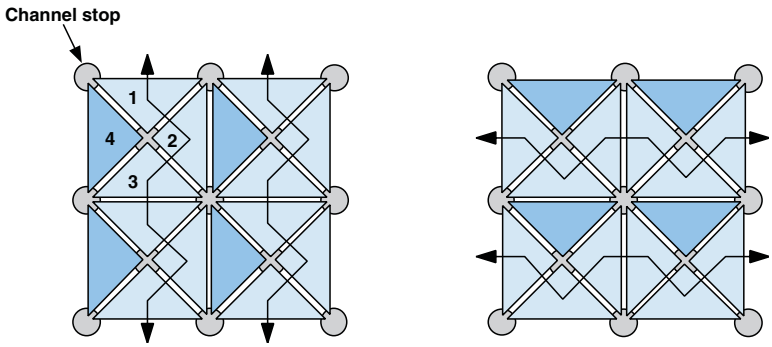


Figure 6. This diagram illustrates the gate layout and charge clocking directions for a multi-directional OTCCD. The symmetry of the gate structure allows clocking of the charge in eight directions: horizontally and vertically (as shown), or diagonally. Subpixel stepping is also possible. In this image, the triangular structures are polysilicon gates that can be biased to form the potential wells in the silicon below. This type of structure requires four gates per pixel, whereas the typical CCD has only three gates. In the illustration at the left, the dark triangular gate is biased low to serve as a barrier phase while the other three phases are modulated to transfer the charge in the vertical direction. At the right, a different gate serves as the barrier phase, and, by modulating the voltage on three different gates, the charge can be transferred in the horizontal direction.

Bright stars and antiblooming

Bright stars present a serious problem for widefield CCD imaging. Even at high GALACTIC LATITUDE, there are numerous bright ($m < 15$) and several very bright ($m < 10$) stars in any $0^\circ.5 \times 0^\circ.5$ field. Larger fields only make the problem worse. Light from these stars can reflect off optical surfaces, including the surface of the CCD itself, and create out-of-focus halos and other scattered light artifacts. Large diffraction spikes can extend well away from the bright stars and break up into faint pieces, resembling faint galaxies, especially in very good seeing. Bright stars will also produce saturated, bloomed charge trails; the result of filling the pixel potential well capacity (typically 150 000–300 000 e^-) and having the excess charge spill up and down the columns into the neighboring pixels. For some brighter stars, the bloomed trails can extend the full length of the device, destroying any information about objects in the path of the bloomed trail. Device manufacturers can address the blooming problem by implementing antiblooming drains in their CCD structures. An antiblooming drain is an implant on the CCD with a potential level near full well, but slightly less than the barrier potential of the neighboring vertical phases. When the charge in a pixel exceeds the potential level of the drain, but before the charge can spill over to the neighboring pixel, it is intercepted by the drain. Such structures can be made so that they do not obscure any of the imaging regions (e.g. the drain can be run down the center of the channel stop). Implementation of such structures will be very valuable in future devices designed for widefield imaging.

Gerard A Luppino

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Widmanstätten, Alois von Beckh- [Alois Beck, Edler von Widmanstätten] (1753/4–1849)

Printer and businessman, born in Graz, Austria, became head of the Fabriksproduktenkabinett, a private technology collection of Emperor Francis I. Discovered the crystalline structure of iron/nickel meteorites by etching polished slices from an iron meteorite from Zagreb and printing from the etched surfaces. The patterns are known as Widmanstätten figures.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Widmanstätten Pattern

A characteristic, roughly hexagonal pattern of intersecting lines that appears on the surface of an octahedrite, a type of iron meteorite, when it is sectioned, polished and etched with acid. The Austrian mineralogist Aloys Joseph von Widmanstätten discovered the pattern in 1804. It is formed by the intergrowth of two nickel–iron alloys under the conditions of slow cooling that pertained in the solidifying core of an asteroidal parent body that had undergone differentiation, and is found only in meteorites. The two alloys are kamacite, with a low nickel content, and taenite, which is richer in nickel.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wien, Wilhelm (1864–1928)

Born in Gaffken, East Prussia (now Poland), Nobel prizewinner (1911), became professor of physics at Munich and discovered *Wien's law*, for the distribution of light in the spectrum of a black-body. Discovered the proton in an early mass-spectrometer experiment (confirmed by Rutherford).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wilcox Solar Observatory

The Wilcox Solar Observatory (WSO) at Stanford University measures the Sun's large-scale synoptic magnetic and velocity fields with the goal of understanding solar variability and how it affects our terrestrial environment. With more than a 22 year solar cycle of spectrograph observations since 1975, WSO staff investigate the solar interior, photosphere, corona, wind and cycle.

The Observatory was first funded by the Office of Naval Research (ONR), the National Science Foundation (NSF) and the M C Fleischman Foundation. Continuing support comes from NSF, NASA and ONR. The Observatory was rededicated in honor of its first director, J M Wilcox, in 1984, shortly after his death. Stanford University is located south of San Francisco, CA.

For further information see
http://wso.stanford.edu.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wilkins, John (1614–72)

Churchman, born in Fawsley, Northamptonshire, became Warden of Wadham College, Oxford, and Master of Trinity College, Cambridge. Founded the Royal Society from a discussion group of scientists at Wadham. In 1638 Wilkins wrote a book describing the Moon as a habitable planet and predicting that, one day, space travel to the Moon would be possible.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wilson, Alexander (1714–86)

Born in St Andrews, Scotland, became professor at Glasgow, observed sunspots and showed that they were depressions in the Sun (following LA HIRE and CASSINI). Published *Thoughts on General Gravitation* (1770), answering NEWTON's question 'What hinders the fixed stars from falling upon one another?' with the speculative answer that the entire universe rotates about a center.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wilson, Olin Chaddock (1909–94)

Astronomer and spectroscopist, became a staff member at Mount Wilson Observatory. He studied stellar chromospheres and stellar activity cycles, showing by intensive analysis of the H and K lines of ionized calcium that other stars besides the Sun have cycles of activity. With M K VAINU BAPPU, he found a means of determining luminosity, and thus distance, of stars from the widths of the emission in these two lines that comes from the chromosphere (*Wilson–Bappu effect*). He studied spectra of nebulae, eclipsing stars, Wolf–Rayet stars and planetary nebulae.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wilson, Robert Woodrow (1936–)

Born in Houston, Texas, Nobel prizewinner for physics in 1978) with ARNO PENZIAS 'for their discovery of cosmic microwave background radiation'. Interested in radio as a boy, drawn to radioastronomy by working with JOHN BOLTON at CalTech mapping the Milky Way. Joined Bell Laboratories at Crawford Hill where with Arno Penzias he shared a small allowance given to the Lab for radioastronomy projects. With new millimeter wave receivers at 100–120 GHz they discovered unexpectedly large amounts of carbon monoxide in a molecular cloud behind the Orion nebula, including isotopic spectral lines so that it was possible to determine isotope ratios as a probe of nucleogenesis. With a large radio telescope (the Holmdel horn) and a new sensitive, low-noise receiver discovered the cosmic microwave background radiation.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

WIMPs and MACHOs

WIMP is an acronym for weakly interacting massive particle and MACHO is an acronym for massive (astrophysical) compact halo object. WIMPs and MACHOs are two of the most popular DARK MATTER candidates. They represent two very different but reasonable possibilities of what the dominant component of the universe may be.

It is well established that somewhere between 90% and 99% of the material in the universe is in some as yet undiscovered form. This material is the gravitational glue that holds together galaxies and clusters of galaxies and plays an important role in the history and fate of the universe. Yet this material has not been directly detected. Since extensive searches have been done, this means that this mysterious material must not emit or absorb appreciable electromagnetic radiation in any known waveband. Thus it is called dark matter. The nature of this material is one of the biggest unsolved problems in science. It is important to identify the dark matter, but since it is easy to invent substances that could fill intergalactic space and yet have escaped detection, there are a very large number of dark matter candidates. Thus several generic classes of dark matter candidates have been suggested, and each dark matter search experiment concentrates on one of these classes. The most important dark matter classes, in terms of detection efforts, are neutrinos, WIMPs, MACHOs and axions.

For example, NEUTRINOS are weakly interacting particles that were almost certainly created in great abundance during the big bang. These fill the Galaxy, moving freely through the Galaxy and even the Earth, and yet are almost impossible to detect. This is because they can be sensed only through the very small electroweak interaction. If each neutrino had a mass of several electronvolts they would contribute enough mass to make up the bulk of the dark matter. For various reasons, it is unlikely that neutrinos of the type discovered in particle accelerators and nuclear reactors on Earth make up much of the dark matter.

However, it is very possible that some as yet undiscovered weakly interacting particle was created during the big bang and today remains in large enough abundance to make the dark matter. The masses typically required for these particles are in the range 1 GeV–1 TeV, and these hypothetical dark matter particles are called WIMPs. There are hundreds of elementary particles that fall into this class of dark matter particles, including supersymmetric particles such as neutralinos, photinos, higgsinos or sneutrinos, and new heavy neutrinos.

This hypothetical WIMP is well studied and attempts to detect these particles have been mounted both by creating them in accelerators and by sensing them in underground detectors as they pass through the Earth. However, there is a large class of astronomical objects that could be the dark matter and still escape detection. For example, if the Galactic halo were filled with Jupiter mass objects ($10^{-3}M_{\dot{o}}$) they would not have been detected by emission or absorption of light. Brown dwarf stars with masses below $0.08M_{\dot{o}}$ or the black hole remnants of an early generation of stars would be similarly invisible. Thus these objects are examples of MACHOs. Other examples of this class of dark matter candidates include primordial black holes created during the big bang, neutron stars, white dwarf stars and various exotic stable configurations of quantum fields, such as non-topological solitons.

An important difference between WIMPs and MACHOs is that WIMPs are non-baryonic and MACHOS are typically (but not always) formed from baryonic material. As discussed in the article on big bang nucleosynthesis (see UNIVERSE: THERMAL HISTORY), baryonic material probably cannot make up all of the dark matter, although it could make up most of the dark matter in the halos of spiral galaxies such as the Milky Way. There is preliminary, although controversial, evidence for the existence of large numbers of MACHOs, but because they probably cannot make up all the dark matter, the search for WIMPs continues unabated.

WIMP thermal relics as dark matter

Among the particle dark matter candidates an important distinction is whether the particles were created thermally in the early universe, or whether they were created non-thermally in a phase transition. Thermal and non-thermal relics have a different relationship between their relic abundance Ω and their properties such as mass and couplings, so the distinction is especially important for dark matter detection efforts. For example, the WIMP class of particles can be defined as those particles that are created thermally, while dark matter axions come mostly from non-thermal processes. Light neutrinos are also thermally created relics, but because of their very small mass have a different history.

In thermal creation one supposes that early on, when the universe was at very high temperature, thermal equilibrium obtained, and the number density of WIMPs (or any other particle species) was roughly equal to the number density of photons (particles of light)[1]. This is just equipartition of energy among all possible degrees of freedom. As the universe cooled the number density of WIMPs and photons decreased together. When the temperature finally dropped below the WIMP mass, however, creation of WIMPs became very rare while annihilation still proceeded. Thus in equilibrium, the number density of WIMPs dropped exponentially : $\exp(-m_{WIMP}/T)$. If equilibrium were maintained until

[1] There is a small difference between the number density of photons and neutrinos since photons obey Bose–Einstein statistics and neutrinos obey Fermi–Dirac statistics.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

today there would therefore be very few WIMPs left, but at some point the WIMP density dropped low enough that the probability of one WIMP finding another to annihilate became small. Note, we must assume that an individual WIMP is stable if it is to become the dark matter. The WIMP number density stopped dropping at this point and we are left with a substantial number of WIMPs today. Detailed evolution of the Boltzmann equation that describes this process can be done for an accurate prediction, but roughly the predicted relic density today of WIMPs is inversely proportional to their interaction strength. The remarkable fact is that, for a relic density equal to the known dark matter density, the interaction strength must be that expected for particles with electroweak-scale interactions: thus the 'W' for 'weakly' in 'WIMP'. There are several theoretical problems with the standard model of particle physics that are solved by new electroweak-scale physics such as SUPERSYMMETRY. Thus these theoretical problems may be clues that the dark matter does indeed consist of WIMPs. Said another way, any stable particle that annihilates with an electroweak-scale cross section is bound to contribute to the dark matter of the universe. It is interesting that theories such as supersymmetry, invented for entirely different reasons, typically predict just such a particle.

The fact that thermally created dark matter has weak-scale interactions also means that it may be within reach of accelerator experiments such as LEP and LHC at CERN, and the proton collider experiments at Fermilab. Thus many accelerator searches for exotic particles are also searches for the dark matter of the universe. Also, because of the weak-scale interactions, WIMP–nuclear interaction rates are within reach of many direct and indirect detection methods, as discussed below.

Supersymmetry and dark matter

Supersymmetry is a new hypothetical symmetry of nature that relates bosons and fermions. If supersymmetry exists in nature then every known particle should have a supersymmetric partner. Bosonic ordinary particles have fermonic superpartners with the same name except with the suffix 'ino' added, while fermonic ordinary particles have bosonic (scalar) superpartner names with the prefix 's' added. Examples of proposed supersymmetric particles include photinos, higgsinos, Z-inos, squarks and selectrons. Some supersymmetric particles have the same quantum numbers as each other and therefore can mix together producing particles that are not exact partners of any standard model particle. For example, the photino, Higgsino and Z-ino can mix into arbitrary combinations called the neutralinos.

In most models, the lightest supersymmetric particle (LSP) is stable, and since supersymmetric particles have electroweak-strength interactions, the LSP makes an excellent dark matter candidate. Typically the neutralino is the LSP so most investigations of WIMP dark matter have concentrated on the neutralino. However, there are many possible supersymmetric models and many free parameters in the models, so precise predictions of supersymmetric relic abundance and supersymmetric particle detectability are not possible. Typically experiments attempt to probe a range of model parameters. So far no concrete evidence of any supersymmetric partner exists. If even one supersymmetric partner is found, the theory predicts that they all must exist.

Note that the parameters that determine the relic abundances also determine all the particle production and rare decay cross sections, as well as the rate in various detectors. Thus once these parameters are specified or measured, one can compare the model predictions with experimental results.

Search for WIMPs

Accelerator searches

Extensive unsuccessful searches for the particles involved in supersymmetric models have been performed at particle accelerators throughout the world. Thus substantial regions of prime neutralino dark matter parameter space have already been eliminated. This does not yet mean that low-energy supersymmetry is unlikely to exist, since only a small portion of the allowed mass range under 1 TeV has been explored. Since supersymmetry predicts a Higgs boson with mass under about 120 GeV, such a discovery would be very important, especially if the Higgs boson showed non-standard properties indicative of supersymmetry. It is correct to think of the particle physics search for supersymmetry as a powerful search for the dark matter.

Direct detection of WIMPs

A satisfying solution to the dark matter problem would be the detection of WIMPs from our Galactic halo as they move past and through the Earth. This would also allow measurement of the local density of dark matter and establish beyond doubt that the dark matter is non-baryonic cold dark matter. There are several ways to do this, and currently two methods are being aggressively pursued.

The most exciting result would be direct detection of the WIMP particles in the laboratory. Since we roughly know the speed (~220 km s^{-1}) and the density (ρ~0.3 proton masses cm^{-3}), we can say that for a WIMP of mass of order 10–100 GeV, roughly 100 000 dark matter particles per second pass through every square centimeter of the Earth. However, if WIMPs exist, they are very weakly interacting particles, so it is quite rare that one of them will interact at all; most of them pass right through the Earth unimpeded. In addition, if a WIMP does elastically scatter off a nucleus, the deposited energy is usually in the keV to 100 keV range, too small to be noticed

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

except by exquisitely sensitive equipment. These difficulties, however, have not stopped many groups throughout the world from developing devices capable of detecting WIMPs. The detection rates turn out to be within and just beyond the reach of current experimental efforts.

The basic idea is to detect the small energy deposited when a WIMP scatters off a nucleus in some well-instrumented piece of material. When a WIMP scatters off a nucleus in a crystal, the nucleus recoils, causing dislocation in the crystal structure, vibrations of the crystal lattice (i.e. phonons or heat) and also ionization. These signals can be detected. For example, some experiments use kilogram size germanium and silicon crystals and attempt to detect the ionization and phonon signals. Other groups use sodium iodide crystals and look for the scintillation light caused by the ionized electrons or search for crystal dislocations in samples of billion year old mica. Another possibility is to record the recoil of an ionized atom in a gas using drift chamber techniques.

The main difficulties in these experiments come from the fact that the WIMP events are rare and that there are many backgrounds that deposit similar amounts of energy on much more frequent time-scales. Thus the experiments operate deep underground, where ionizing COSMIC RAYS are less frequent, and typically operate their detectors at extremely cold temperatures to keep thermal excitations low. Also many types of shielding, as well as redundant detection methods, are now becoming standard. Even so, these are difficult experiments and tiny amounts of radioactivity in the detector or shielding can swamp the expected signal. With effort, a background rate of under one event per kilogram of detector per day can be achieved. The expected signal is highly dependent on the supersymmetry model, but typically is in the range from 10^{-5}–10 events kg^{-1} day^{-1}. The events can be separated from the background in two ways. In some detectors the background (non-WIMP) interactions can be recognized and simply ignored. In the larger detectors this is not possible, so they use the fact that the WIMP event rate is predicted to be larger in June than in December. This annual modulation in event rate is caused by the Earth's orbit either being aligned with the Sun's motion in the Galaxy (in June) or anti-aligned (in December).

The current generation of detectors have detection thresholds of around 1 event kg^{-1} day^{-1}, with hopes that within the next few years signals as small as 10^2 events kg^{-1} day^{-1} will be detectable. Thus there is a reasonable chance that dark matter neutralinos will be detected by this type of direct detection within the next few years. It is also clear, however, that there are many values of the supersymmetry parameters that predict detection rates of below the 10^{-2} events kg^{-1} day^{-1} threshold, and so would not be detectable in the near future by these methods.

Indirect detection of WIMPs

A great deal of theoretical and experimental effort has gone into another potential technique for WIMP detection. The idea is that if the halo is made of WIMPs, then these WIMPs will have been passing through the Earth and Sun for several billion years. Since WIMPs will occasionally elastically scatter off nuclei in the Sun or Earth, they will occasionally lose enough energy, or change their direction of motion enough, to become gravitationally captured by the Sun or Earth. The orbits of such captured WIMPs will repeatedly intersect the Sun (or Earth) resulting in the eventual settling of the WIMPs into the core. As the number density increases over time, the self-annihilation rate will increase. Since ordinary neutrinos can result from WIMP self-annihilation, one predicts a stream of neutrinos coming from the core of the Sun or Earth. Neutrinos easily escape the Solar core and detectors on Earth capable of detecting neutrinos coming from Sun or Earth have operated for some time. The energy of such neutrinos is roughly 1/2 to 1/3 of the WIMP mass, so these neutrinos are much higher energy than the MeV scale solar neutrinos from nuclear reactions that have already been detected. The higher energy of these WIMP annihilation neutrinos makes them easier to detect than ordinary solar neutrinos and somewhat compensates for their much fewer numbers. It also makes them impossible to confuse with ordinary solar neutrinos. Thus the presence of a source of high-energy neutrinos emanating from the centers of the Sun and Earth would be taken as evidence for WIMP dark matter.

While the above chain of reasoning may seem long, it appears to be robust, and several experimental groups are in the process of designing and building detectors capable of seeing such a neutrino signal. For this signal, it is not the mass of the detector that is relevant, but the surface area. Neutrinos from the core of the Sun or Earth produce muons in the atmosphere and rock around the detectors, and it is primarily these muons that the detectors watch for. Muons are also copiously created by cosmic rays entering the Earth's atmosphere, so there is a substantial background of 'downward' traveling muons. These detectors, then, are located deep underground, where the rock shields many of the background muons, and they also focus on 'upward' traveling muons, that are much more likely to have been created by neutrinos that have traveled through Earth and interacted in the rock just below the detector. Thus surprisingly, the best way to see high-energy neutrinos from the Sun is to go deep underground at night (when the Sun is 'under' the Earth).

The new generation of detectors are designed to have very large surface areas. A comparison of direct and indirect detection methods indicates that for a typical neutralino a kilogram of direct detector germanium has about the same sensitivity as 10^4–10^6 m^2 of indirect

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

detector. The new generation of detectors should have areas in this range and should be able to start to probe realistic supersymmetry models, but again viable supersymmetric models exist which predict rates too small for these detectors to measure, and a definitive test of the WIMP hypothesis does not seem possible in the near future.

MACHOs and microlensing

An exciting development in the dark matter story is the detection of MACHOs by three separate groups, MACHO, EROS and OGLE. All three groups monitored millions of stars in the LARGE MAGELLANIC CLOUD (LMC), in the SMALL MAGELLANIC CLOUD (SMC) or in the GALACTIC BULGE, for signs of gravitational microlensing, and all three groups have found it. It has now become clear that these objects constitute some new component of the Milky Way, but whether they constitute a substantial part of the dark matter or new stellar components of the Galaxy or LMC is not clear.

MICROLENSING is a powerful new tool for discovering and characterizing populations of dark objects in our Galaxy, and the current experiments may have the capability to give a definitive answer to the question of whether the dark matter in our Galaxy is baryonic. The reason is that the microlensing searches are probably sensitive to any objects in the range $10^{-8}M_{\odot}<m<10^{3}M_{\odot}$, just the range in which such objects are theoretically allowed to exist. Objects made purely of H and He with masses less than ~(10^{-9}–$10^{-7}M_{\odot}$) are expected to evaporate owing to the microwave background in less than present age of the universe, while dark matter objects with masses greater than ~$10^{3}M_{\odot}$ would have disrupted known GLOBULAR CLUSTERS. So unlike the searches for WIMP dark matter, which if they detect nothing will remain inconclusive, if the MACHO searches find nothing, we would at least know what the dark matter is not. However, the MACHO searches have found something, as we describe below.

Microlensing is also described in the article on GRAVITATIONAL LENSING. The idea of microlensing rests on Einstein's observation that, if a massive object lies directly on the line of sight to a much more distant star, the light from the star will be lensed and form a ring around the lens. The ring is called the 'Einstein ring' and it sets the scale for all the microlensing searches. In the lens plane, the radius of the ring is given by

$$r_E = 610R_{\odot}\left(\frac{m}{M_{\odot}}\frac{L}{\mathrm{kpc}}x(1-x)\right)^{1/2}$$

where $R_{\odot}$ and $M_{\odot}$ are the solar radius and mass, m is the MACHO mass, L is the distance to the star being monitored and x is the distance to the MACHO divided by L. The formation of a ring is very unlikely, but even with imperfect alignment two images result and a large magnification can occur.

Since the MACHO, Earth and source star are all in relative motion, the star appears to brighten, reaches a peak brightness, and then fades back to its usual magnitude. Thus the signature for a microlensing event is a time-symmetric brightening of a star occurring as a MACHO passes close to the line of sight. When a microlensing event is detected, one fits the lightcurve and extracts the peak magnification A_{max}, the time of the peak, t_0, and event duration $\hat{t}$. The primary physical information comes from $\hat{t}$, which depends on the MACHO velocity, the MACHO mass, the source distance, and the lens distance. The source distance can be determined since it is visible, but unfortunately, one cannot determine the other three physical parameters from $\hat{t}$. However, statistically, one can use information about the halo density and velocity distribution, along with the distribution of measured event durations to gain information about the MACHO masses. Using a standard model of the dark halo, MACHOs of Jupiter mass ($10^{-3}M_{\odot}$) typically cause events lasting 3 days, while brown dwarf mass MACHOs ($0.1M_{\odot}$) cause events lasting about a month.

Assuming a halo made entirely of MACHOs, the probability of any MACHO crossing in front of a star is about 5×10^{-7}. Thus many millions of stars must be monitored in order to see a handful of microlensing events. In addition, if one wants to see microlensing from objects in the dark halo, the monitored stars must be far enough away so that there is a lot of halo material between us and the stars. Therefore, the best stars to monitor are those in the LMC and SMC at distances of 50 kpc and 60 kpc respectively, stars in the galactic bulge at 8 kpc and stars in nearby galaxies such as M31 at 750 kpc.

Microlensing experiments

There are several experimental groups that have undertaken the search for microlensing and have returned results. All together about a dozen events have been detected towards the LMC, a couple towards the SMC, and more than 400 towards the galactic bulge. For detection of dark matter MACHOs, it is primarily the LMC events that are relevant. All survey collaborations operate a medium-size telescope and monitor millions of stars nightly. Since the fields are crowded, each CCD frame contains hundreds of thousands of stellar images, the brightness of each which must be determined by computerized photometry. These stellar brightnesses are arranged sequentially in a lightcurve, each of which is finally searched for microlensing-like bumps. All together many terabytes of data have been analyzed by the survey experiments.

Most of the monitored stars are constant brightness as one expects, but about one-half of 1% are variable.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

These are mostly identified as variable stars of known types. Several signatures of microlensing, including the unique shape of a microlensing lightcurve are used to pick out microlensing events from this background of variable stars. For example, the MACHO collaboration analyzed about 9.5 million lightcurves from its 2 yr LMC data set and found six to eight microlensing events.

Experimental results

In order to interpret microlensing events, the efficiency with which an experiment can detect microlensing is needed. Bad weather, bad seeing, crowded stars, etc cause microlensing events to be missed by the experiment. The MACHO collaboration finds an efficiency of around 30% for microlensing events that last 50 days. For events lasting less than a few days and events lasting longer than 6 months the efficiency is very low.

Using the efficiency and a model of the dark halo, the number of microlensing events one expects to see if the halo consisted entirely of MACHOs is found to be about a dozen for the MACHO collaboration 2 yr data set. This quantity is related to the lensing 'optical depth', the roughly 5×10^{-7} probability that at any time a source star is lensed (if the halo is made entirely of MACHOs).

Using the eight observed MACHO collaboration LMC events, the observed optical depth is $\sim(2.5\pm1\times10^{-7})$, roughly half the value if the halo consisted entirely of MACHOs. A careful likelihood analysis of these events gives, for a standard dark halo model, a most likely MACHO halo fraction of 0.5±0.3 and a most likely mass of the MACHOs of around $0.5M_{\odot}$. This result depends on the assumption that the events are due to lenses in the Galactic halo and on the model of the galactic halo used.

Another analysis based on noticing that none of the eight detected events had durations of less than 20 days can rule out low-mass MACHOs. This is because $\hat{t}: m^{1/2}$, and no short-duration events have been observed toward the LMC. This analysis gives the strongest constraints to date on the baryonic content of the dark halo. The EROS and MACHO collaboration limits show that objects in the mass range from $10^{-7}M_{\odot}$ to $10^{-3}M_{\odot}$ cannot make up the entire dark halo. Objects in the range from 3.5–$10^{-7}M_{\odot}$ to 4.5–$10^{-5}M_{\odot}$ make up less than 10% of the dark halo. Thus we now know that the dark matter is not mostly objects of Earth mass, or Jupiter mass, or any combination thereof. The only compact baryonic dark matter candidates left are objects in the brown dwarf and higher mass range. This result is independent of any assumptions about the observed microlensing events, but does depend on the model of the dark halo.

In 2002, an international team of astronomers observed a dark matter object directly for the first time. Images and spectra of a MACHO microlens were taken by the NASA/ESA HUBBLE SPACE TELESCOPE (HST) and the European Southern Observatory's Very Large Telescope (VLT). This result is a strong confirmation that gravitational microlensing is the cause of the events detected by teams searching for MACHO dark matter.

In addition, Christopher Kochanek of the Harvard–Smithsonian Centre for Astrophysics in Cambridge, Massachusetts and Neal Dalal of the University of California, San Diego have used radio telescopes and gravitational lensing to search for cold dark matter. They have studied seven galaxies, each magnified by four nearer ones. Because each lensing galaxy is in a slightly different position, the researchers got four different images of each of the seven distant galaxies. The four images should have been identical. But each is actually slightly different. The difference was enough to have been caused by the kind of clumps of dark matter around lensing galaxies that mathematical models predict.

Interpretation of results: dark matter or not?

The naive interpretation of the microlensing results is that between 20% and 80% of the dark matter in the Milky Way has been identified. However, the result that the mass of the objects is above the brown dwarf limit of $0.1M_{\odot}$ is surprising. Main sequence stars with masses above $0.1M_{\odot}$ would have been seen and therefore cannot be the dark matter. Several interpretations are possible.

First, perhaps the MACHOs are white dwarf stars or neutron stars. These are dark remnants of an earlier generation of stars, but it is problematic to have enough of these around to be the dark matter and not have detected the other byproducts of such an early stellar population. Second, perhaps the model of the Galactic halo used is incorrect, and the masses of the MACHOs are actually safely below the brown dwarf limit. However, most reasonable halo models investigated do not have this property. Next, perhaps MACHOs are primordial black holes, or other exotic objects not currently known. This is possible, but quite surprising. Most importantly, perhaps the microlensing events are not due to halo lenses, and therefore are not telling us about the dark matter. In a typical microlensing event, the distance of the MACHO is not determined, so it is not known where the lens population is located. The estimate of the amount of MACHO dark matter relies on an assumed distribution of lens material, and therefore on the model of the Galaxy and LMC.

It has been suggested that the lenses could be faint stars in the LMC itself, or in some small undiscovered dwarf galaxy between the Sun and the LMC. These possibilities are being vigorously pursued, but strong arguments have been given against both possibilities. Thus the outcome is very unclear at the moment.

If one could measure the distance to the MACHO, that would be enough to distinguish between the above possibilities. Thus a 'microlensing parallax' satellite has been proposed to measure these distances. Other ways to

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

measure the distance to lenses, and therefore determine whether or not they are part of the dark halo, include ground-based lensing parallax, binary lens caustic crossing, and the finite source star effect. Unfortunately, events where these effects can be measured are rare, so luck, or new larger surveys and better follow-up, is needed to resolve this issue of what the MACHOs are.

The next generation of microlensing surveys and follow-up efforts are underway or being planned and should determine some lens distances, as well as gather more events. Thus, while the question of baryonic dark matter remains open, the next few years should bring an answer.

Bibliography

Dalal N and Kochanek C S (in press) 2002 Direct detection of CDM substructure *Astrophys. J.*

An older very nice review of WIMP dark matter is

Primack J R, Seckel D and Sadoulet B 1988 *Annu. Rev. Nucl. Part. S.* B **38** 751–807

and a newer review of supersymmetric dark matter is

Jungman G, Kamionkowski M and Griest K 1996 *Phys. Rep.* **267** 195

A survey of experimental results in particle dark matter detection can be found in the book

Bottino A, di Credico A and Monacelli P 1997 TAUP 97: Proceedings 5th Workshop on Topics in Astroparticles and Underground Physics, *Nucl. Phys.* B (Suppl.) **70**

Recent reviews of microlensing include

Gould A 1996 *Publ. Astron. Soc. Pac.* **108** 465–576
Paczynski B 1996 *Annu. Rev. Astron. Astrophys.* **34** 419–59

The analysis of the MACHO collaboration data can be found in

Alcock C *et al* 1998 *Astrophys. J. Lett.* **499** L9
Alcock C *et al* 1997 *Astrophys. J.* **486** 697

Kim Griest

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Wind

NASA satellite, part of NASA's Global Geospace Science programme and the International Solar Terrestrial Physics programme. Launched November 1994. For the first nine months it followed a double-lunar swingby orbit with apogee of 80–250 Earth radii and perigee 5–10 Earth radii. In this orbit, lunar gravity assists maintained apogee over Earth's day hemisphere for magnetospheric observations. Later inserted into a 'halo' orbit at the sunward Sun–Earth gravitational equilibrium point (L1) to measure the solar wind, magnetic fields and particles, and provide a one-hour warning to other ISTP spacecraft of changes in the solar wind. Since October 1998, placed in 'petal' orbits that take it out of the ecliptic plane.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wing, Vincent (1619–68)

Born in North Luffenham, Rutland, he supported himself as a surveyor, almanac compiler (his almanac sold 50 000 copies per year), astrologer and prolific writer of astronomical works.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Winthrop, John (1714–79)

Astronomer, mathematician, born in Boston, MA, became professor of mathematics and natural philosophy at Harvard and observed sunspots, a transit of Mercury and of Venus, eclipses and the weather. He predicted the return of Halley's Comet in 1759. He is credited as the first professional scientist in America. He was an ardent patriot during the American Revolution.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wise Observatory

Wise Observatory, in Mitzpe Ramon, Israel, is owned and operated by Tel Aviv University, and has a well-equipped 1 m telescope. Since construction in 1971, the large percentage of clear nights at its desert site and its unique longitude have made the observatory particularly useful for long-term monitoring projects (e.g. reverberation mapping of quasars and active galaxies), and as a part of global monitoring networks (e.g. the first detection, via gravitational microlensing, of a planet orbiting a binary star system).

For further information see
http://wise-obs.tau.ac.il.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wittich, Paul (*c*. 1546–86)

Born in Breslau, Silesia (now Wroclaw, Poland), invented (or at least developed) 'prosthaphaeresis', a formalism of trigonometry that allowed one to multiply and divide trigonometric functions by the easier process of adding and subtracting instead (in the manner of logarithms). He wrote a commentary on COPERNICUS'S *De Revolutionibus*, which foreshadowed the Tychonic system. He worked with TYCHO BRAHE for four months at Uraniborg.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

WIYN Observatory

Located at Kitt Peak in Arizona. The WIYN Observatory is owned and operated by the WIYN Consortium, which consists of the University of Wisconsin, Indiana University, Yale University and the National Optical Astronomy Observatories (NOAO). Most of the capital costs of the observatory were provided by these universities, while NOAO, which operates the other telescopes of the KITT PEAK NATIONAL OBSERVATORY, provides most of the operating services.

The 3.5 m WIYN Telescope, which was completed in 1994, is the second largest telescope on Kitt Peak. The size of the telescope enclosure is kept to a minimum by the short focal length of the primary mirror, which results in a shorter telescope, while the alt-azimuth mount requires less space. The moving weight of the telescope is a mere 46 tons.

Other innovative design features are active primary mirror supports, primary mirror thermal controls and active ventilation of the telescope mount. The support system for the primary mirror includes 66 actuators, which adjust the back face of the mirror to maintain the best optical figure. The primary mirror thermal control system keeps the temperature of the mirror's surface to within 0.2 °C of the ambient air temperature, eliminating local turbulence. These innovations enable the WIYN Telescope to produce much sharper images than any of the other telescopes on Kitt Peak.

WIYN is equipped with the latest instruments for astronomical spectroscopy and imaging. A multiple-object spectrograph employing optical fibers allows the simultaneous observation of the spectra of 100 objects. The imaging cameras employ highly sensitive arrays of electronic detectors.

For further information see
http://www.noao.edu/wiyn/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wolf, Charles J E (1827–1918)

French astronomers Charles Wolf and GEORGES RAYET, using the 40 cm Foucault telescope of the Paris Observatory, visually observed the spectra in 1867 of several eighth magnitude stars in Cygnus before the systematic use of photographic plates and found very broad emission lines. The 'bands' were originally thought to be hydrocarbon molecules. The stars became known as *Wolf–Rayet (WR) stars*.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wolf, Johann Rudolf (1816–93)

Born in Fällanden (near Zurich), Switzerland, became professor of astronomy at the University of Bern and director of the Bern Observatory, then professor of astronomy in Zürich where he founded an observatory. He devised a system now known as *Wolf's sunspot numbers* used to quantify solar activity by counting sunspots and sunspot groups and used it to confirm the sunspot cycle discovered by HEINRICH SCHWABE and measure its period at 11 years. He also co-discovered with EDWARD SABINE its connection with geomagnetic activity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wolf, Max[imilian] Franz Joseph Cornelius (1863–1932)

Astronomer, born in Heidelberg, Germany, founded and became first director of the Königstuhl Observatory at the University of Heidelberg. He took wide-field photographs of the Milky Way and counted stars of different brightnesses, plotting the results in a *Wolf diagram* of number versus magnitude to prove the existence of clouds of obscuring dust. He showed that the spiral nebulae have absorption spectra typical of stars, rather than emission spectra from gas. He pioneered the use of photography to discover hundreds of asteroids.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wolf–Rayet Nebula

Nebulosity surrounding a Wolf–Rayet star. Wolf–Rayet stars are of around 10 solar masses and have very high surface temperatures, up to about 40 000 K. This gives them powerful stellar winds, up to 2000 km s^{-1}, and an enormous rate of mass loss. Material is usually ejected in the form of a spherical shell or ring (the term *Wolf–Rayet bubble* is sometimes used), and the accumulating envelope from successive ejection episodes comprises the nebula. Examples of Wolf–Rayet nebulae are NGC 2359, surrounding the star HD 56925, and NGC 6888, surrounding the star MR 102.

See also: Wolf–Rayet stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wolf–Rayet Stars

Wolf–Rayet (W–R) stars are a class of peculiar stars first identified in 1867 by C J E WOLF and G RAYET. Unlike the spectra of most stars, which are dominated by narrow absorption lines, the spectra of W–R stars show broad emission lines. The rich emission line spectrum makes them easy to identify, by spectroscopic observations, even at large distances.

W–R stars are divided into three broad spectroscopic classes (WN, WC and WO) based on the emission lines present in their spectrum. WN stars show emission lines predominantly of helium and nitrogen, although emission due to carbon, silicon and hydrogen can readily be seen in some of these objects. In contrast, the spectra of WC stars are dominated by carbon and helium emission lines with hydrogen and nitrogen emission absent (figure 1). WO stars, which are much rarer than either WN or WC stars, are similar to WC stars except that oxygen lines are more prevalent, and there is a tendency to exhibit lines arising from atomic species of higher ionization.

These spectral classes are further divided into subclasses on the basis of line ratios, yielding a classification by ionization. The WN stars which exhibit spectra showing emission from high-ionization species (e.g. He II, N V, O VI)[1] are designated WN2. Those showing emission from low-ionization species (e.g. He I, N III) are classified as WN9, although recently the W–R spectral classification has been extended to WN11. Similarly, WC stars showing emission from high-ionization species (e.g. He II, C IV, O VI) are designated WC4 while those exhibiting the lowest ionization (e.g. He I, C II) are designated WC9. In the literature there is also a tendency to refer to WN stars of classes 2–5 as early type (WNE) and classes 6–9 as late type (WNL). Similarly, WC4–6 stars are designated as WCE, while WC7–9 stars are designated as WCL. Although there are important exceptions, WNE stars generally show no evidence for H emission while H emission is present in WNL stars.

The distribution of population I W–R stars, which are discussed in this article, is similar to that of O stars; they are primarily located in the spiral arms of our galaxy and near H II regions. W–R masses range from an uncertain lower limit of about $5M_{\odot}$ to in excess of $60M_{\odot}$, while surface temperatures range from a lower limit of 25 000 K to greater than 100 000 K. Because of their spatial association with O stars, and their peculiar surface abundances, W–R stars are generally believed to be descended from O stars.

Approximately 220 W–R stars are known in our Galaxy but this number is certainly incomplete. Most are hidden from our view by dust, which absorbs and scatters light (a process termed interstellar extinction) within our Galaxy. Estimates of the total number of W–R stars in our Galaxy range from 1000 to 2000. The rarity of W–R stars is due to the initial mass function, which favors the production of low-mass stars, and the short evolutionary lifetime of W–R stars, which is only a few $\times 10^5$ yr. Their rarity belies their importance. All stars more massive than approximately $25M_{\odot}$ (for solar metallicity) pass through a W–R phase. Further, over the lifetime of a galaxy, W–R stars (and their progenitors) have an important influence on the energetics, dynamics and chemical evolution of the interstellar medium.

W–R stars are expected to end their life via a spectacular supernovae explosion. In some cases a neutron star is formed, while for the more massive stars a black hole is formed. In the latter case the details of the mechanism that actually produces the supernova are still very uncertain. As the supernovae ejecta expands, it will interact with the complex circumstellar environment that reflects the previous mass loss history of the progenitor star.

In addition to the population I W–R stars, some planetary nebula central stars also show W–R emission features (see PLANETARY NEBULA CENTRAL STAR MASS LOSS/WINDS, W–R NEBULA). Their spectral types are inserted in [] to distinguish them from population I W–R stars. They are of type [WC] and have lower masses (less than $1M_{\odot}$) and lower luminosities (generally $<3 \times 10^4 L_{\odot}$). The spectra of planetary nebula W–R stars are often dominated by strong narrow nebula emission lines. In some cases they can be difficult to distinguish spectroscopically from normal population I W–R stars although in some [WC] stars N and H emission is seen. Because of their distinct evolutionary histories, they will not be further discussed in this article.

Basic model

The basic model for W–R stars is that of a hot star which is suffering extreme mass loss. The mass loss occurs via a continuous stellar wind which is accelerated from low velocities near the surface of the star to velocities that exceed the surface escape speed. The observed spectrum originates over a range of radii with the optical continuum forming close to the stellar core, while the emission lines originate from a volume that can extend beyond 10 stellar radii.

The observed mass-loss rates (i.e. the amount of material lost per year) are extreme, typically in excess of $10^{-5} M_{\odot}$ yr^{-1}. These mass-loss rates are sufficient to affect the evolution of the star (see STELLAR EVOLUTION) and must be incorporated into stellar evolutionary calculations. The (average) maximum velocity of material in W–R winds (called the terminal velocity, V_{∞}) ranges from 800 km s^{-1} to in excess of 3000 km s^{-1}, and typically exceeds the escape velocity from the surface of the star. It is generally believed, although it has yet to be rigorously demonstrated, that the mass loss is driven by radiation pressure acting through numerous bound–bound atomic transitions of Fe and other atomic species in the extreme UV ($\lambda < 900$ Å).

The bulk of the material in the wind is believed to be cool—that is, it has a temperature substantially lower

[1] He II is a spectroscopic designation used to indicate the ion for a transition between bound levels in singly ionized helium (i.e. He^+).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

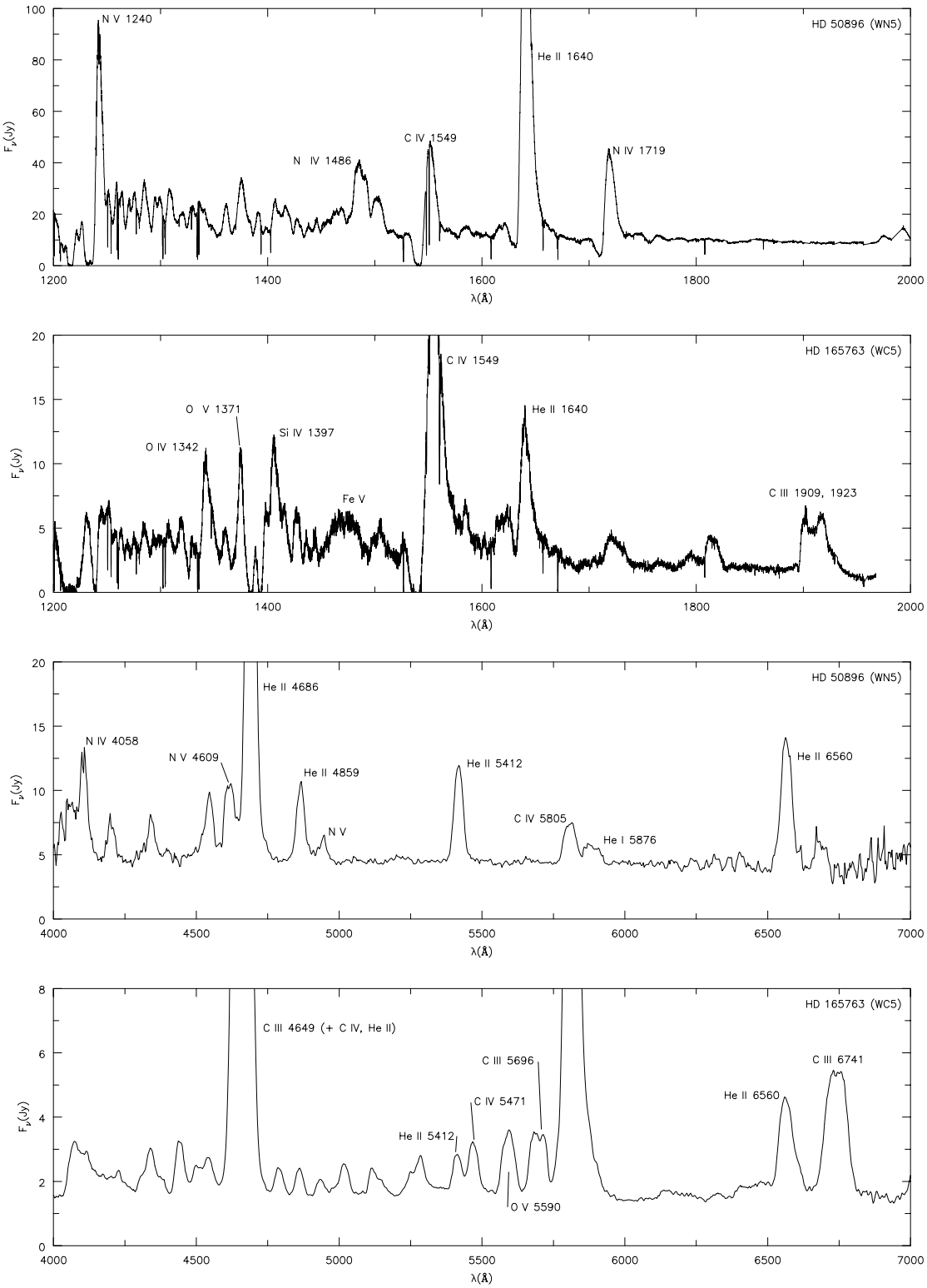


Figure 1. UV and optical spectra of the WN5 star HD 50896 and the WC5 star HD 165763. The major emission line features are identified, although it should be noted that many lines, particularly in the WC star, are blends. Notice the very distinct differences between the WN5 and WC5 spectra. The optical spectral region, which can be observed from the ground, has typically been used to classify W–R stars. Since the advent of space astronomy, the UV spectral region has provided additional invaluable diagnostics on the properties of W–R stars.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

than the effective temperature of the star. Energy input into the wind primarily occurs through photoionization by the intense UV radiation field emanating from the central source. Thus photoionization is the ultimate source of the line emission that originates in the wind, although individual emission lines form through a variety of different processes—recombination, collisional excitation and continuum fluorescence.

Determination of stellar parameters

The spectra of most stars are determined by three basic parameters: the effective temperature ($T_{\rm eff}$), the effective surface gravity ($g_{\rm eff}$) and the chemical abundances. For an assumed metallicity, large grids of models can be constructed simply by varying $T_{\rm eff}$ and $g_{\rm eff}$.

The effective temperature is defined by the relation

$$L = 4\pi R_*^2 \sigma T_{\rm eff}^4$$

where L is the luminosity and σ is the Stefan–Boltzmann constant. The effective temperature is the surface temperature that a star of radius $R_\star$ would have if it radiated as a blackbody (a perfect thermal emitter and absorber).

The effective surface gravity is defined by

$$g_{\rm eff} = (1-\Gamma)\frac{GM}{R_\star^2}$$

where M is the star's mass, G is Newton's gravitational constant, and Γ is a correction for the influence of radiation pressure. In conjunction with the equation of hydrostatic equilibrium

$$\frac{{\rm d}P}{{\rm d}r} = -\rho g_{\rm eff}\left(\frac{R_\star}{r}\right)^2$$

where P is the pressure and ρ is the density, this sets the scale height of the atmosphere, $h_\star$. For an isothermal atmosphere, $h_\star$ is given by

$$h_\star = \frac{kT}{\mu m_{\rm H} g_{\rm eff}}$$

where μ is the mean particle mass in atomic mass units (amu) and $m_{\rm H}$ is the atomic mass of hydrogen. For most normal stars $h_\star \ll R_\star$, curvature effects can be ignored and the atmosphere can be treated as a plane-parallel slab. As a consequence of the small scale height, the Sun has a definite radius at optical wavelengths.

For W–R stars the situation is quite different. First, the atmosphere is extended, and consequently radiation escapes from the star over a range of radii. Further, the radius of the star at an optical depth (τ) of 2/3 depends on the adopted mass-loss rate and is a function of wavelength (figure 2). The difficulty of uniquely defining $R_\star$ has led to difficulties in comparing $T_{\rm eff}$ derived from evolutionary models with that obtained from spectroscopic analyses. Second, $g_{\rm eff}$ does not have a direct influence on the stellar spectrum, simply because emission from the stellar wind dominates the spectral appearance of the star. Third, the abundances are non-solar and must be determined observationally. Indeed it is the abundances that determine to which class (WN, WC or WO) a W–R star belongs (see also STELLAR ATMOSPHERES: EARLY-TYPE STARS).

In addition to the abundances (primarily of H, He, N, C and O) it has been found from numerical experiments that the spectra of W–R stars are determined primarily by two parameters: $T_{\rm eff}$ and a wind density parameter, W_ρ. The latter, which plays a similar role to $g_{\rm eff}$, can be defined by

$$W_\rho = (\dot{M}/V_\infty)R_{\rm c}^{-3/2}$$

where $R_{\rm c}$ is the radius of the hydrostatic core. Two stars will have very similar spectra if they have similar abundances, and if $T_{\rm eff}$ and W_ρ are similar. The dependence of the spectra on W_ρ arises because most of the radiative processes in an extended atmosphere depend on the square of the density. As a consequence of the scaling, it is impossible to deduce the distance of a W–R star from the Sun using its spectrum[2]. In principle, $\dot{M}$ and V_∞ should depend on the other stellar parameters—composition, M, L and $R_\star$, but as yet our theoretical understanding of mass loss from W–R stars is not sufficiently advanced to deduce the relationship.

Because of the low wind densities (10^8–10^{14} electrons cm^{-3})[3] the simplifying assumption of local thermodynamic equilibrium (LTE) cannot be made when modeling W–R spectra. When LTE holds, it can be assumed that the ionization state of the gas and the populations of the atomic levels can be found via application of the principles of statistical mechanics and thus are (simple) functions of the local temperature and density only. For LTE to prevail, collisional processes, which couple the atomic populations with the electrons (and hence to the local electron temperature), need to occur faster than radiative processes.

In contrast, in W–R atmospheres radiative processes tend to dominate over collisional processes, and hence it is necessary to solve the equations of statistical equilibrium at each depth. For each atomic level of each species we assume that all the processes (radiative and collisional) populating the level are in equilibrium with all processes depopulating the same level. The major difficulty arises because the rates are a function of the radiation field, which in turn is a function of the unknown populations. Thus the radiation field and atomic populations must be solved for simultaneously, and in general an iterative procedure is necessary to obtain consistency.

[2] Stellar evolution introduces a correlation between spectral type and luminosity which may statistically allow distances to be derived for stars of a given spectral type. However, for an individual star the derived distance may be grossly in error, particularly if the spectral type can originate via different evolutionary sequences. This is exemplified in the difficulty of determining whether some W–R stars belong to population I or are the central stars of planetary nebula.

[3] For comparison, the density of water on Earth is of order 10^{22} molecules cm^{-3}.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

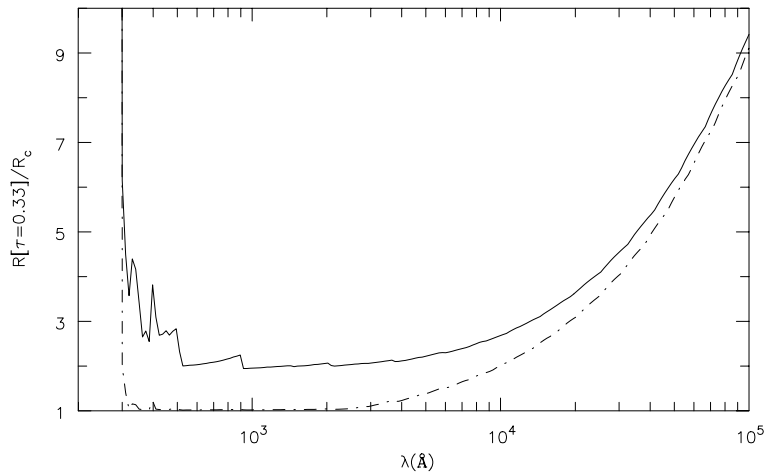


Figure 2. Illustration of how the 'radius' of the star varies with wavelength for the WC5 star HD 165763. The solid curve shows the 'radius' at which the 'continuum' optical depth, τ, is 1/3, which occurs in the wind at all wavelengths. The broken curve is identical except that the electron scattering opacity was not included in the computation of τ. The bulk of the observed flux in this model is emitted between 300 and 2000 Å. R_c is the radius of the hydrostatic core—that is, the approximate radius the star would have in the absence of a stellar wind. For the illustrated model it was $1.8R_\odot$.

Initial modeling of W–R spectra concentrated on the H and He spectra only. The second generation included CNO elements, while the most recent generation of models include iron and other species. The inclusion of iron (and similar species) in non-LTE calculations has been a major stumbling block for atmospheric calculations for O and W–R stars for many years. The iron-group ions have both a wealth of atomic levels and an enormous number of bound–bound transitions. The advent of faster computers with large memories, new numerical techniques and the availability of atomic data have now made it feasible to include iron and other species in non-LTE calculations.

Determination of $T_{\rm eff}$

The effective temperatures of W–R stars are determined primarily using ionization arguments. Consider a sequence of models with prescribed mass loss and velocity law, but with different effective temperatures. Such a sequence of models will exhibit a smooth variation in line ratios for lines from two successive stages of ionization. In early WN modeling it was customary to compare He II 5411 with He I 5876 since both lines are easily observed and are relatively blend free. In more recent modeling N lines can also be used to constrain the effective temperature, while in WC stars C and O lines can be used. The analyses generally give consistent results. Discrepant results do occur, and probably result from a poor treatment of line blanketing (the generic name given to the influence of thousands of bound–bound atomic transitions on an atmosphere) and/or density inhomogeneities in the stellar wind.

Analyses of the spectra of ring nebula around W–R stars offers a method of checking on the energy distributions predicted by atmospheric modeling, since the nebula are ionized by the star's radiation field. For ring nebula around WNE stars these analyses have usually shown reasonable consistency. However, for some WNL and WC stars the observed nebular spectrum was of lower excitation than would be predicted using the stellar UV radiation field derived from the modeling. The recent inclusion of line blanketing in the stellar atmospheric models has removed this discrepancy for at least one ring nebula around a WN8 star.

Abundances

The severe non-LTE conditions in W–R stars initially made it difficult to understand their peculiar emission line spectra. Do W–R stars possess peculiar (i.e. non-solar) abundances? Is the difference between the WN and WC stars due to an abundance difference, or is it an excitation effect? Detailed recombination and spectroscopic analyses have now firmly established that W–R stars are characterized by non-solar abundances.

In WN stars H, C and O are depleted, while N and He are enhanced. For WN stars, $N(\mathrm{H})/N(\mathrm{He})$ ratios (by number) range from approximately 4 to <0.1 (the solar value is 10). The observed abundances are consistent with the idea that material processed by the CNO nuclear burning bi-cycle has been revealed (or mixed) at the surface (see 'Evolution').

In WC stars, He, C, O and Ne are all enhanced. $N(\mathrm{C})/N(\mathrm{He}) = 0.1$–$0.5$ while the less certain $N(\mathrm{O})/N(\mathrm{He})$ ratios are typically 0.1. H and N are not expected in WC stars and are not detected. The variation of $N(\mathrm{C})/N(\mathrm{He})$ and $N(\mathrm{O})/N(\mathrm{He})$ with WC subtype is still the subject of much debate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Determination of mass-loss rates

Mass-loss rates can be determined in several ways. First, free–free radiation in the dense stellar wind gives rise to a detectable flux at radio wavelengths. From measurements of the radio flux, S_ν, the mass-loss rate, $\dot{M}$, can be determined from the simple formula

$$S_\nu = 23.2\left(\frac{\dot{M}z}{V_\infty \mu}\right)^{4/3}\left(\frac{\gamma g \nu}{d^3}\right)^{2/3}$$

where d is the distance to the W–R star in kpc, $\dot{M}$ is the mass-loss rate in $M_\odot$ yr^{-1}, V_∞ is the terminal velocity of the wind in km s^{-1}, z is the mean ionic charge, μ is the mean ionic mass (in amu), γ is the number of electrons per ion, ν is the frequency in Hz and g is the free–free Gaunt factor (which is the quantum correction factor to the semiclassical formula for free–free radiation) at frequency ν. S_ν is measured in janskys (1 jansky $= 10^{-23}$ erg cm^{-2} s^{-1} Hz^{-1}). The greatest uncertainties in the derived $\dot{M}$ arise from uncertainties in stellar distance and in the ionization state of the gas in the radio-emitting region.

Second, the IR flux can be used in a similar manner, although in this case it cannot be assumed that the wind has reached its terminal speed. Third, optical and UV emission lines can be used. Typically, recombination lines are used since they are less sensitive to the precise details of their formation. Mass-loss estimates obtained from different methods generally agree to within a factor of 2.

It has generally been assumed that the winds of W–R stars are spherical and homogeneous. Thus at any location in the wind the density can simply be found from the principle of mass conservation, giving

$$\rho = \frac{\dot{M}}{4\pi r^2 v(r)}$$

where $v(r)$ is the velocity as a function of distance and is assumed to monotonically increase with r. However, emission line variability studies and analyses of emission line profiles suggest that the wind is clumped (i.e. non-homogeneous) on small scales. If this is true, mass-loss estimates are too large—possibly by factors of 3 or more. This has very important implications for stellar evolution calculations. A change in $\dot{M}$ by only a factor of 2 (over a star's life) has a profound influence on the evolution of massive stars. This is seen in theoretical evolutionary calculations and can also be indirectly inferred from the different WN/WC ratios in the galaxy and our nearest extragalactic neighbor, the Large Magellanic Cloud (LMC). In the LMC the mass-loss rates are expected to be lower because of the lower metallicity.

The extreme mass-loss properties of W–R stars can be characterized by the 'wind performance parameter', η, defined by

$$\eta = \frac{\dot{M}V_\infty}{L/c}.$$

$\dot{M}V_\infty$ is the (scalar) momentum of the wind, while L/c is the momentum that could be transferred if all the photons, and hence all the momentum in the radiation field, were absorbed. For O stars, η is typically less than unity. For W–R stars values as high as 100 have been obtained, although by allowing for inhomogeneities and line blanketing it has been possible to reduce the values to less than 10. Values of η in excess of unity do not rule out a radiation-driven wind—they simply indicate that each photon has to scatter many times within the wind so that it delivers η times its momentum to the wind. It has been difficult to produce radiation-driven winds for W–R stars since current models do not have the necessary number of bound–bound atomic transitions to perform the required number of scatterings.

Binaries

Approximately 50% of W–R stars occur in binaries—a number comparable with O stars. In the past there has been considerable discussion on the importance of binarity for the W–R phenomenon. For example, it was once thought that W–R stars could only originate in a binary system. Mass loss from the W–R progenitor would then occur by Roche-lobe overflow. The major uncertainty in evolutionary calculations of binary systems is how much material is lost from the system during Roche-lobe overflow (rather than being accreted by the companion). More recently the binary channel for the production of W–R stars has virtually been ignored. There is little doubt, however, that the binary channel is important, and it must be considered when linking W–R types with evolutionary calculations (although many researchers would disagree with this statement).

At least three (broad) distinct classes of W–R binary systems can be envisioned: W–R + OB star, W–R + W–R, W–R + compact companion (neutron star or black hole). All (confirmed) W–R binaries belong to the first class; W–R + W–R systems are expected to be rare, although WR98 (where WR98 denotes the 98th W–R star in the sixth catalogue of W–R stars) may be an example of such a system. The third class is expected on evolutionary grounds. Although the existence of such systems has been difficult to verify, several good candidate systems are known. Cyg X-3 can be considered a possible example.

W–R binary systems are extremely useful. First, and foremost, they allow a direct determination of stellar masses, independent of evolutionary models. The use of these masses in constraining single-star evolutionary models presupposes that W–R stars in binaries have properties similar to single W–R stars—a proposition that cannot be reliably tested because of poor statistics and uncertainties in the properties of W–R stars.

Second, the O star can be used to probe the structure of the W–R stellar wind. Indeed it was this technique, applied to the W–R + O binary V444 Cygni, which gave the first direct evidence that W–R stars are hot (i.e. surface temperature in excess of 60 000 K). More recently it has become evident that polarization studies of binary systems may allow determinations of mass-loss rates which are insensitive to the presence of inhomogeneities within the stellar wind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-ray emission

Most W–R stars are thermal x-ray sources. They can be classified into two categories—single W–R stars and binary W–R stars. In single W–R stars the observed ratio of x-ray to bolometric luminosity is approximately 10^{-7}, although there is a scatter of at least a factor of 3 about this value. The x-rays are believed to originate within the stellar wind through radiation-driven wind instabilities which lead to high-velocity shocks ($V_{\rm shock}$ = 100–600 km s^{-1}) and clumping of the wind. These shocks generate x-ray emitting gas characterized by temperatures between 10^6 and 10^7 K. The shocks probably permeate most of the wind; however, because of the high wind densities most of the x-rays are absorbed in the wind.

A few W–R stars are stronger and/or variable x-ray emitters. In these binary systems the x-rays can be generated from the high-temperature gas (of order 10^7 K) created via shocks generated in a wind–wind collision. The two best examples are the WN + O system V444 Cygni and the WC + O system HD193793 (WR140). Both exhibit a periodic x-ray variability broadly consistent with that expected from a wind–wind collision in a binary system. Alternatively the x-rays could be generated via accretion of wind material onto a compact companion.

Dust formation

This topic, by all rights, should not belong in a discussion on W–R stars since dust is destroyed by intense UV radiation and requires low temperatures (<3000 K) for its formation. Yet dust is seen around some WC stars, and moreover dust is seen to be created in the vicinity of some WC stars. Two distinct dusty WC classes may exist: binary systems and single stars.

In many WC9 stars, presumed to be single, conditions within the C-rich stellar wind appear to allow dust formation. The dust nucleation routes and how the dust (and the necessary prerequisite molecules) forms despite the presence of an intense UV radiation field are unclear. If the stars are single, inhomogeneities generated by radiation instabilities probably play an important role in allowing dust to form.

The second class are the binary systems, with eccentric orbits, with HD193793 (WR140) being an excellent example. In these systems dust does not normally form; however, binary interaction near periastron (i.e. minimal orbital separation) can facilitate dust formation. Apparently the high densities generated in the wind–wind interaction have the right conditions for dust formation. As in the single WC stars, the dust formation is not understood.

Recent interferometric observations of WR104 with the KECK telescope have revealed directly the dust outflowing from the interaction region of the binary (figure 3).

Figure 3. A gray scale image at 2.27 μm showing the distribution of dust in WR104, as found by interferometric observations with the KECK telescope. The dust moves away from the system radially—the apparent spiral motion is an illusion. It results from the rotation of the dust formation zone as a consequence of the orbital motion of the binary system. The ring of dust, as illustrated, has an angular diameter of approximately 160 AU (or 0.1 arcsec). (Picture courtesy of W Danchi, J Monnier and P Tuthill, Berkeley.)

Related stars

In the upper part of the Hertzsprung–Russell (H–R) diagram, many different classes of massive luminous stars exist (see HIGH-LUMINOSITY STARS): Of stars, blue supergiants (BSGs), red supergiants (RSGs), LUMINOUS BLUE VARIABLES (LBVs) and WN/Of stars. One of the goals of massive star evolution is to understand the links between the various classes of objects and the distribution of massive stars between the different classes. Below we briefly discuss some of the salient features of each class.

Of stars are O supergiants exhibiting emission lines in the optical. They are O stars that have evolved off the main sequence. WN/Of stars exhibit spectral characteristics of both Of and WNL stars. It was this intermediate characteristic that suggested an evolutionary link between the Of stars, and bona-fide WN stars.

LBVs, as their name suggests, are luminous blue stars which show irregular variability on a time scale of hours to centuries. Some LBVs have exhibited giant outbursts in which their visual brightness and bolometric magnitude increased by several magnitudes. During such outbursts several solar masses of material may be ejected. Other LBVs, such as AG Car, show moderate outbursts on a timescale of a decade. During these outbursts the effective temperature changes but the bolometric luminosity and mass-loss rate are almost constant. The most famous LBVs are P Cygni, which suffered a giant outburst in the 1600s, and η Car which underwent a major outburst in the 1840s. The outburst suffered by η Car ejected a bipolar nebula, referred to as the Homunculus, which has a major axis diameter of 17″ (approximately 4000 AU). Images of the Homunculus are amongst the most spectacular obtained by the Hubble Space Telescope.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

A major breakthrough in our understanding of massive stars was achieved when one LMC Of/WN star (now classified as WN11), R 127, was observed to undergo an LBV-like outburst, suggesting an evolutionary link between LBVs and Of/WN stars.

LBVs are now regarded as a key phase of massive star evolution. It is believed that during the LBV phase a massive star ejects most of its hydrogen-rich outer envelope, allowing it to become a W–R star. A key observation that has led to this scenario is the absence of RSGs with luminosities comparable with the most luminous O stars. Evolutionary calculations show that, during the LBV phase, extensive mass loss can prevent the star from evolving into an RSG. The mechanism of mass loss during the LBV phase is not yet understood, although it may be related to the star evolving towards the modified Eddington limit. The classical Eddington limit provides a lower limit on the mass of a star, of a given luminosity, based on the assumption that the radiative force arising from the electron scattering opacity cannot exceed gravity. The modified Eddington limit takes into account that other radiative processes also contribute to the opacity, and further that stellar rotation can effectively lower the surface gravity. The maximum luminosity exhibited by supergiants in the H–R diagram is termed the Humphreys–Davidson limit.

Evolution

W–R stars are believed to be descended from O stars. The basic evolutionary sequence, first proposed by Conti in 1976, is

$$\mathrm{O} \rightarrow \mathrm{Of} \rightarrow \mathrm{W\text{–}R}$$

Since that time observational and theoretical work has led to refinements in this basic sequence. From the theoretical work of Maeder and collaborators, one such sequence for stars with initial masses greater than $50M_{\odot}$ is

$$\mathrm{O} \rightarrow \mathrm{OIf} \rightarrow \mathrm{BSG} \rightarrow \mathrm{LBV} \rightarrow \mathrm{WN} \rightarrow \mathrm{WC} \rightarrow \text{supernova}$$

while stars between $35M_{\odot}$ and $50M_{\odot}$ have the alternative sequence

$$\mathrm{O} \rightarrow \mathrm{BSG} \rightarrow \mathrm{YSG} \rightarrow \mathrm{RSG} \rightarrow \mathrm{YSG} \rightarrow \mathrm{WN} \rightarrow \mathrm{WC} \rightarrow \text{supernova}.$$

Other sequences have also been proposed. Suffice it to say that the precise evolutionary path that an individual massive star follows (which depends on the star's initial mass and composition, and possibly its rotation rate and whether it has a companion) is still uncertain. No firm link has been established between the different ionization classes within the WN and WC sequences, although there has been some success in linking spectral types with initial stellar mass. The unknown roles of binary interactions and rotation only add to the confusion.

Both WN stars and WC stars are generally believed to be on the helium-burning main sequence, although some of the luminous hydrogen-rich WN stars may still be core hydrogen burning. If the latter is true, it means that the spectroscopic and theoretical definitions of W–R stars are inconsistent. This creates difficulties in comparing observed W–R/O number ratios with theoretical predictions. The W–R/O ratio is an important observational constraint since it provides a method (at least in principle) of determining the minimum stellar mass which will evolve into a W–R star. For a solar metallicity this is generally believed to be approximately $25M_{\odot}$, but higher values cannot be ruled out.

In massive stars, H burning occurs via the CNO bi-cycle through a sequence of reactions, with the CNO species acting as catalysts. In the CN cycle the following reactions occur:

$$\begin{aligned}
{}^{12}_{6}\mathrm{C} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{13}_{7}\mathrm{N} + \gamma \\
{}^{13}_{7}\mathrm{N} &\rightarrow {}^{13}_{6}\mathrm{C} + \mathrm{e}^{+} + \nu \\
{}^{13}_{6}\mathrm{C} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{14}_{7}\mathrm{N} + \gamma \\
{}^{14}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{15}_{8}\mathrm{O} + \gamma \\
{}^{15}_{8}\mathrm{O} &\rightarrow {}^{15}_{7}\mathrm{N} + \mathrm{e}^{+} + \nu \\
{}^{15}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{12}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He}.
\end{aligned}$$

The fourth reaction is the slowest, and as a consequence much of the original C (and O from the other reactions in the CNO bi-cycle) is converted to N. The total number of CNO nuclei remains unchanged. When equilibrium is obtained, the ratio of ^{14}N to ^{13}C nuclei is approximately 50, very different from the solar ratio of 0.27.

In normal stars the nuclear processed material remains within the stellar core and cannot be observed. However, in O stars and their descendants, extensive mass loss peels off the outer hydrogen-rich layers. Nuclear processed material, once inside the convective core of the star, is eventually revealed at the surface. In addition to mass loss, it is now believed that mixing, possibly induced by stellar rotation, can help reveal nuclear processed material at the stellar surface.

In WC stars the mass loss has been so extensive that the products of He burning are revealed at the stellar surface. The predominant reactions for helium burning are

$$\begin{aligned}
{}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} &\rightleftharpoons {}^{8}_{4}\mathrm{Be} \\
{}^{8}_{4}\mathrm{Be} + {}^{4}_{2}\mathrm{He} &\rightarrow {}^{12}_{6}\mathrm{C} + \gamma \\
{}^{12}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He} &\rightarrow {}^{16}_{8}\mathrm{O} + \gamma.
\end{aligned}$$

The variation in surface abundances as a function of current mass for a star with an initial mass of $40M_{\odot}$ is shown in figure 4.

W–R stars in external galaxies

W–R stars are moderately easy to detect in external galaxies owing to their strong emission lines. Typically they are found by performing a photometric survey in two filters. The passband of one filter is centered on a strong emission feature (generally the He II–C III/C IV complex at 4640–4690 Å) while the second passband is centered on

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

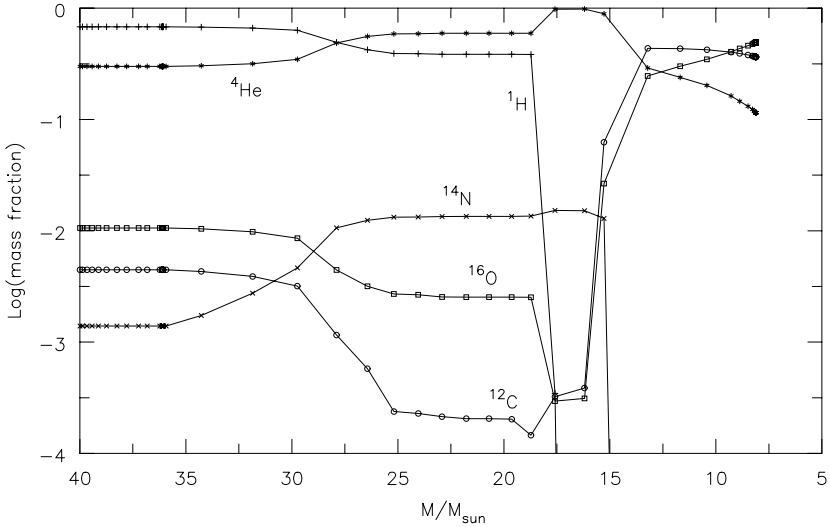


Figure 4. A diagram showing the evolution in surface abundances as a function of mass for a star with an initial mass of $40M_{\odot}$. The WN phase occurs when the star's mass is between (approximately) $15M_{\odot}$ and $30M_{\odot}$, while the WC phase occurs when $M < 15M_{\odot}$. For most of its 4.8×10^6 yr life the star has a mass close to its initial value (e.g. after 4.2×10^6 yr, its mass is still $\approx 36M_{\odot}$). The lifetimes of various stages are very dependent on the adopted assumptions (e.g. overshooting and mixing) and the adopted mass-loss rates. Overshooting refers to the phenomenon of convective motions extending into a convectively stable region because the convective velocities are non-zero at the interface between the convectively stable and unstable regions. Mixing refers to the process of mixing two chemically distinct regions of the star—for example rotation might induce nuclear processed material to be transported to the surface layers. The calculations were undertaken by the Geneva group (Maeder, Meynet, Schaller and Schaerer) for solar metallicity, with overshooting and 'normal' mass-loss rates.

the continuum only. W–R stars will be relatively bright in the emission line passband. Alternatively, they can be found using low-dispersion prism spectroscopic surveys.

Both the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC) have been extensively surveyed for W–R stars. In the LMC 134 W–R stars are known, while in the SMC only nine are known. The difference in the number of W–R stars is believed to be due to a combination of the star formation rates and the lower metallicity of the SMC (which inhibits W–R production). Interestingly, the ratio of WN to WC stars in the LMC is 4.5, substantially larger than the observed ratio of 1 in the solar neighborhood. This is generally interpreted as a metallicity effect.

W–R stars have also been found in many Local Group galaxies, e.g. M31, M33, IC1613 and NGC6822. The study of individual W–R stars in these galaxies is in its infancy. To date, efforts have been directed primarily into determinations of the WN/WC, O/W–R and RSG/W–R number ratios which allow global issues, such as the effect of metallicity and the star formation rate on W–R production, to be addressed.

W–R stars have also been found in many galaxies exhibiting extensive star formation (often called STARBURST GALAXIES). Indeed some galaxies are termed W–R galaxies if they exhibit strong W–R features in their integrated spectra. The presence of W–R stars in these galaxies immediately provides an age determinant. The starburst has to be older than approximately 2 million years so that the most massive O stars that formed in the burst have had sufficient time to evolve into W–R stars. They also provide an upper limit of about 7 million years since after this time all massive stars that pass through a W–R stage will have done so. Both age limits are metallicity dependent.

Outstanding problems

There are many outstanding problems related to W–R research. Several of the most important problems related specifically to W–R stars are discussed below.

(1) What initiates and drives mass loss from W–R stars? Is radiation pressure, as currently believed, responsible for mass loss from W–R stars? Why are mass-loss rates for W–R stars an order of magnitude higher than those of their O star progenitors?

(2) What is the role of rotation in W–R stars (and their progenitors) in modifying the spectral appearance of the star? What is the role of rotation in massive stars in enhancing mixing processes and in enhancing mass-loss rates, and hence in modifying stellar evolution?

It is essential for evolutionary calculations, whatever the mass-loss mechanism, that we are able to derive mass-loss rates from first principles. Ideally

$$\dot{M} = \dot{M}[M_{\rm init}, t, x, y, z, \Omega(t)]$$

where $M_{\rm init}$ is the initial stellar mass, t is the current age of the star and x, y and z are the chemical

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

abundances (hydrogen mass fraction, helium mass fraction and metal mass fraction), and where the rotation rate $\Omega(t)$ is itself determined via the initial rotation rate and the subsequent mass loss.

A related question is the role of magnetic fields in massive star evolution. While the magnetic fields may be too weak to affect mass-loss rates they could be important in determining angular momentum losses, and hence play an indirect role in mass loss through the dependence of $\Omega(t)$ on magnetic field strengths.

(3) What is the detailed structure of W–R winds (shape and homogeneity)? How does the presence of inhomogeneities affect the determination of the fundamental stellar parameters? How coupled is the mass loss to the details of the wind structure?

(4) What is the role of binaries in massive star evolution? Are there classes of W–R stars (and LBVs) that only come from a binary evolutionary scenario? How much is our understanding of single-star evolution being confused by the evolution of OB stars in interacting binary systems?

(5) Because of uncertainties in the treatment of convection, mass-loss processes and rotation, there are still many issues in single-star evolution which must be addressed. Also of concern is whether W–R stars give rise to black holes (as presumed), and to what class a W–R supernova belongs.

Acknowledgments

This contribution is a brief synopsis of our current understanding of W–R stars. This understanding is the cumulative result of 130 years of W–R research by many astrophysicists whose individual contributions cannot be acknowledged.

Bibliography

Abbott D C and Conti P S 1987 Wolf–Rayet stars *Ann. Rev. Astron. Astrophys.* **25** 113 (an excellent review article which discusses W–R stars in greater depth)

Bappu M K V and Sahade J (ed) 1973 *Wolf–Rayet and High Temperature Stars (IAU Symp. 49)* (while some of the articles are somewhat dated, the discussion between the symposium participants is rewarding and is sadly missing in most recent symposia)

Maeder A and Conti P S 1994 Massive star populations in nearby galaxies *Ann. Rev. Astron. Astrophys.* **32** 227 (an excellent review article which discusses massive star populations and evolution, with an extensive discussion of W–R stars)

Tylenda R 1996 *Wolf–Rayet Central Stars of Planetary Nebulae (ASP Conf. Ser. 96)* ed C S Jeffery and U Heber (review: several other papers on W–R central stars are also contained within the same volume) p 101

van der Hucht K A, Conti P S, Lundström I and Stenholm B 1981 The sixth catalogue of galactic Wolf–Rayet stars, their past and present *Space Sci. Rev.* **28** (3)

van der Hucht K A, Koenigsberger G and Eenens P R J (ed) 1999 *Wolf–Rayet Phenomena in Massive Stars and Starburst Galaxies (IAU Symp. 193)* (Astronomical Society of the Pacific) (review and research articles on different aspects of W–R research)

Vreux J M, Detal A, Fraipont-Caro D, Gosset E and Rauw G (ed) 1996 *Wolf–Rayet Stars in the Framework of Stellar Evolution: Proc. 33rd Liege Int. Astrophysical Colloq.* (Université de Liége) (contains many review articles on different aspects of W–R research)

D John Hillier

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wollaston, William (1766–1828)

Chemist and physicist, isolated two new metals from platinum ore, one of which he named palladium (after the newly discovered asteroid, Pallas). He was one of the first scientists to observe ultraviolet radiation, and in 1801 discovered the dark spectral lines in the solar spectrum which were later investigated by FRAUNHOFER. This was the first observation of spectral lines. A *Wollaston prism* is used to study the polarization of light.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Woolley, Richard van der Riet (1906–86)

Astronomer, born in Dorset, England, and followed a varied career in England, South Africa, California and Australia. Became director of the Commonwealth Solar Observatory and steered it away from solar observations and wartime optics work to become the Mount Stromlo Observatory and part of the Australian National University. He worked on stellar and solar atmospheres. In 1955 he became Astronomer Royal in Britain, directing the Royal Greenwich Observatory in Sussex. He worked there on the motions and astrophysics of stars and clusters of stars, directing a small army of workers on the measurements of star positions on the thousands of photographic plates that he caused to be taken. He discovered the orbit of the globular cluster Omega Centauri, which moved radially in the Galaxy. He caused the Isaac Newton Telescope to be erected at Herstmonceux: it was later moved to the clearer sky of La Palma. He retired from the Royal Greenwich Observatory and became director of the South African Astronomical Observatory at the Cape.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wormhole

A hypothetical shortcut, or 'tunnel', that, in principle, may link the interior of a black hole to another universe or to another location in our universe. During the 1930s, Albert Einstein (1879–1955) and Nathan Rosen (1909–95) showed that the sharply curved spacetime of the interior of a black hole may open out again into another spacetime (another universe). The hypothetical connection between these two regions of spacetime came to be known as an Einstein–Rosen bridge. An alternative interpretation is that the bridge, or tunnel, links two different regions in the spacetime of our own universe. More recently, the term 'wormhole' has been used to describe a spacetime tunnel of this kind. Although it has been speculated that wormholes could be used to facilitate virtually instantaneous interstellar travel, in practice it seems likely that, even if wormholes do exist, they will be too small and too short-lived (and too physically hazardous) to be utilized in this way.

See also: general theory of relativity, spacetime.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wren, Sir Christopher (1632–1723)

Architect and astronomer, born in East Knoyle, Wiltshire, became professor of astronomy at Gresham College, and Savilian Professor of Astronomy at Oxford. He rebuilt London after the fire of 1666, planning the entire city and rebuilding 51 churches, including St Paul's Cathedral. Newton acknowledges Wren as a mathematician in the *Principia*. Wren independently proved KEPLER's third law and formulated the inverse-square law of gravitational attraction. He solved *Kepler's problem* on cutting a semicircle in a given ratio by a line through a given point on its diameter (it had arisen as a problem in Kepler's work on elliptical orbits). Wren was a founder member of the Royal Society, and its president for two years.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wright, Thomas (1711–86)

English astronomer, born near Durham, clock-maker and mathematics teacher, wrote in 1750 a curiously argued book in which he suggested that the Milky Way was a disk-like system of stars with the solar system near the center. Wright suggested that nebulae were star systems similar to the Milky Way, but very far away. Conjectured that the gap in the planets between Mars and Jupiter had been cleared by the collision of a comet with a planet which had then been ejected from its orbit.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Wrinkle Ridge

A long ridge on the surface of a planetary body. Wrinkle ridges were first identified on the Moon, where they are often associated with rilles. As the lunar maria solidified, tensile forces in the outer regions opened up the faults that produced rilles, while compressive forces nearer the center pushed up the surface to form wrinkle ridges. Some wrinkle ridges may result from small-scale extrusion of lava along fissures; others extend from the maria into surrounding upland terrain. Lunar wrinkle ridges are typically several hundred meters high and several hundred kilometers long. On Venus wrinkle ridges are common features on the plains, where they extend for 10 to 50 km. The alignment of many of them suggests that they are associated with the compressive forces that uplifted the northern upland region of Aphrodite Terra.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Xenocrates of Chalcedon (396–314 BC)

Philospher and mathematician, born in Chalcedon (now Kadikoy, near Istanbul), Bithynia (now Turkey), succeeded Speusippus as head of the Athenian Academy which PLATO had founded. Believed that matter is composed of indivisible units (and thus an early atomist). He believed (and perhaps originated the notion) that people have a threefold existence, mind, body and soul, and that they die twice, once on Earth, then on the Moon, when the mind separates from the soul and travels to the Sun.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Xenophanes of Colophon (*c*. 570–*c*. 480 BC)

Greek philosopher, born in Colophon. He believed Earth to be the fundamental element of the universe, noting that that because seashells are sometimes found on mountain tops, the physical arrangement of the Earth changes with time.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

XEUS (X-ray Evolving Universe Spectroscopy Mission)

Proposed European Space Agency x-ray observatory mission to be located on the International Space Station. Specifications include a 10 m diameter mirror with 25–50 m focal length, spatial resolution of 1 arcsecond, covering the energy range 0.1–100 keV.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-Ray Astronomy

X-ray astronomy is an achievement of the space age as the Earth's atmosphere is completely opaque at photon energies beyond the ultraviolet region.

In 1949 the first X-RAYS from the solar corona were detected by a Geiger counter on a V-2 rocket. In 1962 the discovery of the first x-ray source outside the solar system followed—SCORPIUS X-1. With the same rocket experiment an apparently diffuse x-ray background was found. Since these early days x-ray astronomy has developed at an enormous pace. Today, we know more than 150 000 x-ray sources in the sky and they include almost all astrophysical species—from the nearby comets to the most distant quasars at the edge of the universe, from the tiny neutron stars to the clusters and superclusters of galaxies as the largest physical formations in the cosmos. Some species radiate most of their power in x-rays, for instance black holes or neutron stars accreting matter from a binary companion, as well as supernova remnants and single, cooling neutron stars.

Many of the known objects shine in x-rays because they are hot—having temperatures of millions to billions of kelvins. Other emission mechanisms are synchrotron radiation of extremely energetic electrons spiraling in magnetic fields or the inverse Compton effect which occurs when high-energy electrons scatter at low-energy photons, e.g. from stellar light or from the 2.7 K cosmic background radiation. In any case, the emission of x-rays points to extreme physical conditions in the source region. Also, these x-rays carry information which is not available from observations in other spectral bands. In this article we shall give a brief summary of the development of x-ray astronomy, describe the evolution of experimental techniques and x-ray space missions and highlight some of the results which have had a major impact on astrophysics.

The beginnings with sounding rockets

Like GAMMA-RAY ASTRONOMY, the field of x-ray astronomy was pioneered by physicists. Herb Friedman, who detected the first solar coronal x-rays in 1949, had been working previously in ionospheric physics; Giacconi and Rossi, who led the group discovering Scorpius X-1 and the x-ray background in 1962, were nuclear and cosmic ray physicists.

The instruments used in these pioneering experiments were rocket-borne Geiger counters and spectral discrimination was achieved by means of windows and filters. A little later proportional counters became the standard instrument, working in the 1–10 keV band and having a modest spectral resolution (∼20% at 5 keV). Observations were first performed by scanning the sky by spinning or precessing the rocket. Later on more sophisticated attitude control systems allowed pointed observations to be made with greatly improved sensitivities.

The first discoveries of cosmic x-ray astronomy were totally unexpected. Sco X-1 showed a luminosity many orders of magnitude larger than that of the Sun. It soon became clear that the mechanism by which such sources are powered was the infall of matter into the deep gravitational potential well of a NEUTRON STAR (as in Sco X-1) or a BLACK HOLE. Most of the few dozen sources found during the rocket era, i.e. until the end of the 1960s, were of this kind. The other major class was SUPERNOVA REMNANTS. The identification of the CRAB NEBULA as a bright x-ray source by means of a lunar occultation was one of the early highlights. Other important discoveries of this era were the x-ray pulsations of the CRAB PULSAR and the x-ray emission from the active galaxy M87 at the center of the Virgo cluster.

After the advent of x-ray satellites rocket-borne experiments with their short observation time (∼5 min) were generally not competitive any more. However, they continued to play a role as a test bed for instruments to be flown on satellites or for special observations, in particular those ones of short lead time (e.g. on the famous SUPERNOVA 1987A).

Hard x-ray balloon experiments

The rocket experiments had shown that the spectra of compact x-ray sources are quite hard, extending much beyond 10 keV. This opened up the possibility of using balloon-borne instruments as the atmosphere becomes transparent for altitudes above 40 km at high x-ray energies (>20 keV). Balloon flights offered the advantage of long-duration observations (up to 100 h) and the crystal scintillation counters used in these experiments allowed spectra to be taken up to ∼500 keV. Balloon observations were pioneered by the MIT group in the 1960s. Important results of these early years were the measurement of the Crab spectrum to high energies (∼500 keV), the discovery of variability in GX1 + 4 and the detection of a 20 min flux from Sco X-1. This field of research culminated in the 1970s with experiments of the MIT–Leiden and the Tübingen–MPE groups which flew very large detectors operating in the pointing mode. Highlights of these activities were the observation of a lunar occultation of the Crab, the discovery of a spectral break in the Cyg X-1 spectrum corresponding to a temperature of about $\sim 20 \times 10^6$ K and the discovery of cyclotron lines at ∼40 keV in the accreting neutron star Her X-1, allowing the first direct measurement of its polar magnetic field ($\sim 5 \times 10^{12}$ G).

X-ray astronomy satellites

Most significant progress in x-ray astronomy came with the advent of satellite observatories. Their ancestor, the first satellite entirely devoted to x-ray astronomy, was UHURU. Launched in 1970 it was a spinning spacecraft with a simple, but very powerful, instrument package: an array of proportional counters of 840 cm^2 area working in the 2–20 keV band. It performed the first all-sky survey and located 339 objects, mostly X-RAY BINARIES and supernova remnants, showing a strong clustering near the galactic plane. At fainter flux levels an isotropic distribution of Seyfert galaxies and clusters of galaxies was found.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

The discovery of two pulsating and eclipsing x-ray binary systems, Cen X-3 and Her X-1, proved in an impressive way that these systems contained matter-accreting neutron stars. The discovery of short time variability of Cyg X-1 by Uhuru led to follow-up rocket experiments yielding accurate source positions which enabled the optical counter-part to be identified and the mass of the accreting compact object to be determined. This mass turned out to be larger than the limiting mass of a neutron star. Thus it had to be a black hole.

Uhuru also marked the beginning of an impressive series of satellites with ever-increasing capabilities listed in table 1. We cannot discuss them in detail here, but rather describe the main directions of developments.

Imaging x-ray telescopes

Imaging of x-rays is possible in different ways. Early solar observations used pinhole cameras and Fresnel zone plates, but the sensitivity of such devices is insufficient for studies of the rather weak cosmic x-ray sources. Therefore, the standard X-RAY TELESCOPES use Wolter optics which consist of paraboloidal–hyperboloidal mirrors reflecting x-rays under grazing incidence (see GRAZING INCIDENCE OPTICS). The first x-ray telescopes of this kind using x-ray film as a detector were used for solar observations from rockets and Skylab. The first x-ray satellite carrying a Wolter telescope was the Einstein observatory launched in 1979. It provided images with ∼10 arcsec resolution and represented a real breakthrough, putting x-ray astronomy on equal footing with optical astronomy.

X-ray sky surveys

All-sky surveys have traditionally been a foundation of astrophysical research, and in the era of multiwavelength astronomy their importance has dramatically increased. Such surveys provide an unbiased view of the sky, they deliver large homogeneous samples of objects and they allow rare species to be discovered. Cross correlating the surveys from different wavelength bands—e.g. optical and x-rays—is a very effective method to select sources of a certain type, e.g. active galactic nuclei and quasars.

The Uhuru and Ariel V surveys have revealed ∼350 sources in the standard x-ray band (2–6 keV). The subsequent HEAO-1 sky survey was not much more sensitive (840 sources) but widened the energy band considerably (0.1–200 keV). The limitations of all these collimated counter surveys in terms of angular resolution (<1 deg^2) and sensitivity were overcome by the German-led ROSAT which performed the first all-sky survey in soft x-rays by sweeping an imaging Wolter telescope across the whole sky. Although it took only half a year of the more than 8 yr of ROSAT's life this survey discovered 80 000 x-ray sources and located them with 25 arcsec resolution. In addition, the survey provided a complete map of the diffuse x-ray emission with 12 arcmin resolution (figure 1).

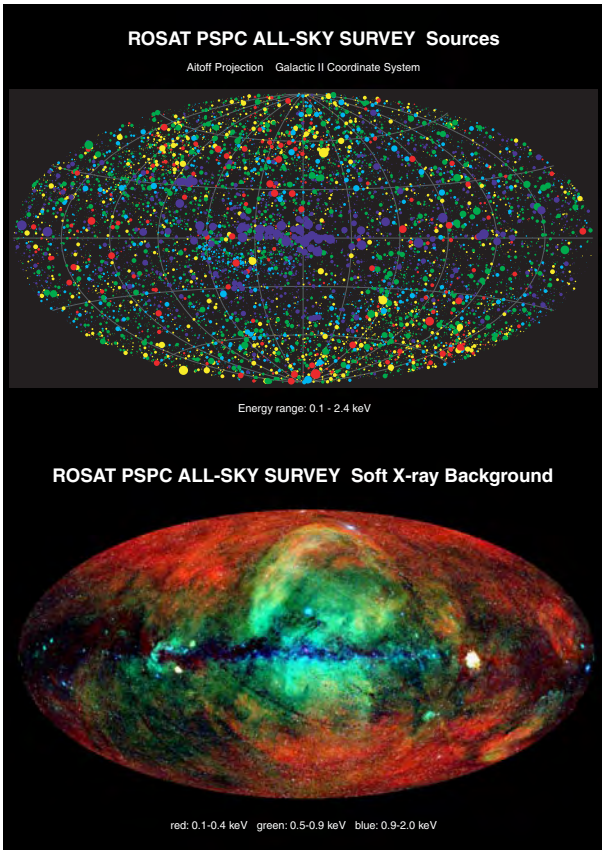


Figure 1. The x-ray sky as known at the end of the 20th century on the basis of the ROSAT All Sky Survey. Top: distribution of the ∼19 000 brightest x-ray sources; Bottom: diffuse emission from hot interstellar matter heated by supernova explosions. **This figure is reproduced as Color Plate 61.**

Recent highlights of x-ray astronomy

A rather detailed account of the development of x-ray astronomy until ∼1990 with a description of missions and their results can be found in Bradt *et al* (1992). Here we want to highlight the progress made in the 1990s by the powerful x-ray telescopes on ROSAT and ASCA.

In addition, there have been two very successful recent missions which must be mentioned here. The ROSSI Timing Explorer (RXTE) carrying large-area collimated counters covering a wide energy range (2–200 keV) has deepened our understanding of compact sources by high time resolution and spectral studies. One of its most exciting results was the discovery of 2.75 ms pulsations occurring during the bursts of the low-mass x-ray binary (LMXB) 4U 1728-34. The data suggest that the pulsation came from the rotation of a thermonuclear hotspot on the surface of the neutron star.

In another LMXB which had been previously identified as an X-RAY BURSTER by BEPPOSAX, RXTE found 2.5 ms pulsations in the persistent flux which must be due to the rotation of the neutron star. This is the first

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. X-ray astronomy satellites 1969–2000. Adapted from Charles and Seward (1995) updated January 2000.

Satellite	Country	Launch	Demise	Type
Vela 5 A,B	USA	May 1969	June 1979	Scanning, small scintillation counter, gamma-ray range
Uhuru	USA	December 1970	January 1975	Scanning, proportional counters
OSO-7	USA	September 1971	May 1973	Scanning, proportional counters
Copernicus	USA–UK	August 1972	February 1981	Pointed, x-ray telescope (non-imaging)
ANS	Netherlands–USA	August 1974	July1976	Pointed proportional counters, Bragg crystal spectrometer
Ariel-V	UK	October 1974	March 1980	Scanning, rotating modulation collimators (RMCs) + proportional counters
SAS-3	USA	May 1975	April 1980	Scanning, RMC
OSO-8	USA	June 1975	October 1978	Scanning, proportional and scintillation counters, Bragg crystal, polarimeter
HEAO-1	USA	August 1977	January 1979	Scanning + short pointings, proportional and scintillation counters, RMC
Einstein	USA	November 1978	April 1981	Pointed, imaging x-ray telescope, Bragg crystal, transmission gratings
Hakucho	Japan	February 1979	April 1985	Scanning, RMCs
Tenma	Japan	February 1983	∼1985	Gas scintillation proportional counter (GSPC), all-sky monitor
EXOSAT	ESA	May 1983	April 1986	Pointed, imaging x-ray telescope, large proportional counters, GSPC
Ginga	Japan–UK	February 1987	October 1991	Pointed, proportional counters
Kvant	USSR–UK–Netherlands–Germany	June 1987	–	Pointed, GSPC, coded mask, scintillation counter
Granat	USSR–Russia	December 1989	August 1999	Pointed, coded masks, all-sky monitor
ROSAT	Germany–UK–USA	June 1990	December 1998	Scanning, pointed imaging x-ray and EUV telescopes
ASCA	Japan–USA	February 1993	–	Pointed, imaging x-ray telescopes, imaging GSPC
RXTE	USA	December 1995		Pointed, large proportional counters, scintillation counters, all-sky monitors
BeppoSAX	Italy–Netherlands	April 1996		Pointed imaging x-ray telescope, coded mask, scintillation counter
Chandra	USA	June 1999		Pointed imaging x-ray telescope, spectrometers
XMM-Newton	ESA	December 1999		Pointed imaging x-ray telescope, spectrometers

and so far only LMXB showing both persistent pulsations and bursts. BeppoSAX is an Italian–Dutch satellite with instruments covering a wide energy band (0.1–200 keV). By discovering the x-ray afterglows of gamma-ray bursts it pointed the way to the solution of an old puzzle, the physical nature of gamma-ray bursts.

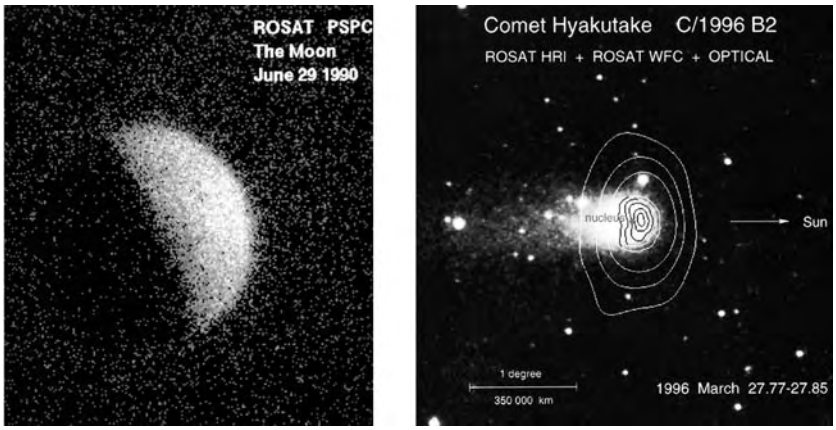
Highlights from ROSAT and ASCA

ROSAT and ASCA have complementary properties. ROSAT carries a large x-ray telescope with a position-sensitive proportional counter (PSPC) providing moderate spectral resolution (∼40%) in the 0.1–2.4 keV band. With its High Resolution Imager (HRI), a microchannel plate detector 'black-and-white' images with 5 arcsec resolution can be taken. The telescopes of ASCA cover the energy band 0.5–10 keV with CCD detectors and imaging gas scintillation proportional counters having superior energy resolution (∼20%), but worse angular resolution (∼3 arcmin), compared with ROSAT. Both satellites have been used by many astrophysicists to study a wide variety of problems. The numbers of scientific publications resulting from ROSAT and ASCA run to about 4000 and 1600, respectively, covering almost all fields of astrophysics. In the following a few highlights are presented.

ROSAT took the first x-ray picture ever of the MOON (figure 2(*a*)). The Sun-lit side of the Moon contains a uniform brightness distribution as in optical light. This is due to solar coronal x-rays undergoing Thomson scattering in a very thin layer of the lunar surface. The PSPC spectrum shows a broad spectral bump at 0.6 keV which is due to fluorescent resonance scattering by oxygen of the minerals of the lunar surface layers. The effective reflectivity of the Moon in the ROSAT band is only ∼0.01%; this means that the Moon behaves as a black body at x-ray energies. At the same time it casts a shadow on the 'diffuse' x-ray background (see later section). The small flux of x-rays apparently coming from the dark side of the Moon is probably produced in the Earth's upper atmosphere by charge-exchange processes of solar wind ions. This is the same mechanism which is responsible for the x-ray emission from cometary comas (see below).

COMETS are cold objects that have been described as dirty snowballs. Therefore the discovery with ROSAT of x-rays from COMET HYAKUTAKE on 27 March 1996 was surprising to many scientists (figure 2(*b*)). Later, another

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*) (*b*)

Figure 2. Solar system objects in x-rays: (*a*) the Moon reflecting solar coronal x-rays and casting a shadow on the x-ray sky; (*b*) x-ray emission from comet Hyakutake superimposed on an optical amateur photograph.

four comets were found in the ROSAT All Sky Survey archive, including comet Levy. In total about a dozen comets have been detected by now in x-rays, mostly with ROSAT. Various physical processes have been proposed to explain the observed extended x-ray emission, including scattering of solar x-rays by cometary dust, x-rays produced by collisions of cometary and interplanetary dust particles and bremsstrahlung x-rays from electrons accelerated at the shock between the solar wind and the coma. However, all these mechanisms fail to explain the observed characteristics. The most successful and widely accepted model suggests that charge exchange between highly charged ions (such as C^{5+}, C^{6+}, O^{6+} and O^{7+}) in the solar wind and neutral particles (such as water) in the cometary coma is the dominant source of the observed x-ray emission.

ROSAT has given many exciting results about stars of all types. The study of a complete sample of stars of solar type has revealed the existence of a sharp lower bound to the x-ray flux, measured at the surface of the star. Interestingly, this minimum stellar x-ray flux is identical to the flux observed in the coronal holes of the Sun. This result suggests that the stars of minimal x-ray flux are completely surrounded by stellar analogs of solar coronal holes.

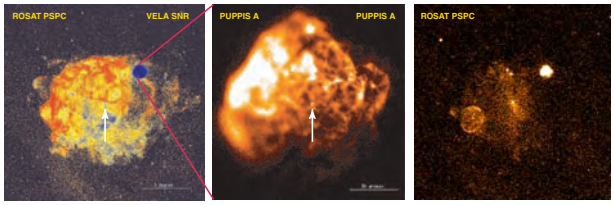
With ground-based optical follow-up observations of unidentified ROSAT All-Sky Survey sources, several hundred new T TAURI STARS have been identified based on Hα emission of hydrogen and lithium absorption. Surprisingly, many new T Tauri stars have been found far outside regions of ongoing star formation. Such 'off-cloud' T Tauri stars must have been either ejected from their birthplaces in the clouds with high velocities or formed in small cloudlets which have largely dispersed since then.

Recently, deep ROSAT observations have led to the discovery of young ($<10^6$ yr) brown dwarfs showing x-ray emission. The x-ray properties of these brown dwarfs, in particular their x-ray to optical light ratio, are similar to those of nuclear burning stars of low mass. This may suggest that young brown dwarfs have hot coronae responsible for the x-ray emission just like nuclear burning stars of low mass.

Early in the ROSAT mission a number of objects were discovered emitting extremely soft x-rays. They are very luminous and show temperatures of a few hundred thousand kelvins. It turned out that these sources are a species which had been predicted to exist but which had not been found before. They are WHITE DWARFS in binary systems accreting matter from their companions at a rate just sufficient to sustain steady nuclear burning on the surface of the white dwarf. Thus, they represent a unique situation in which steady nuclear burning is observed at the surface of a compact star.

Massive stars explode giving rise to a supernova when the nuclear fuel in their cores is exhausted. The core collapses to a neutron star or a black hole while the shell of the star is expelled in a giant explosion. A large fraction of the kinetic energy is converted by shocks into high-temperature x-ray emitting plasmas. Supernova 1987A was ROSAT's first-light target on 16 June 1990, but it turned out to be too faint to be seen at that time. Its remnant was first discovered in soft x-rays with ROSAT in 1992 and has steadily become brighter since then. A large increase is expected to occur in the near future when the shock reaches the high-density regions of the red giant wind. In total, some 200 supernova remnants have been found with ROSAT. Three of them are clustered in the Vela region (figure 3). One is the VELA SUPERNOVA REMNANT, which, at a distance of 1500 ly, is one of the closest supernova remnants. Its diameter is about 200 ly, its age about 20 000 yr. Protrusions discovered with ROSAT at the periphery of the shell are probably produced by fragments of the exploding star; x-ray spectroscopy with ASCA has revealed that they show different chemical compositions (figure 3(*a*)). The Puppis A remnant at the north-western rim of the Vela supernova remnant

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK



(*a*) (*b*) (*c*)

Figure 3. The Vela complex of supernova remnants: (*a*) Vela supernova remnant with the Vela pulsar and Puppis A; (*b*) Puppis A with its central source, probably the remnant neutron star; (*c*) A young, ∼680 yr old supernova remnant, superimposed on Vela supernova remnant, showing titanium-44 emission. **This figure is reproduced as Color Plate 62.**

is located at a much larger distance (figure 3(*b*)). It is younger and has a much higher temperature than the Vela supernova remnant. Surprisingly, a third supernova remnant has been discovered recently in the Vela complex with ROSAT. Radioactive titanium-44 has been detected from it with the Compton Gamma-Ray Observatory (figure 3(*c*)). Titanium-44 is produced in supernova explosions and has a mean lifetime of only 90 yr, which means that the remnant is very young (∼680 yr). Actually, this supernova remnant looks like a twin of SN 1006. Like the latter it must have been very bright at maximum (about a quarter of the full Moon) and it is unclear why it has not been recorded in the Chinese and Japanese annals.

Neutron stars shine in different ways. If they are highly magnetized and rapidly rotating, they appear as radio PULSARS. These objects emit beamed radiation produced by high-energy electrons (and positrons) that are accelerated in their magnetospheres. In about half a dozen young pulsars, optical and gamma-ray pulses have also been seen. With ROSAT and ASCA, 34 radio pulsars have been detected through their magnetospheric x-ray emission, including the 89 ms Vela pulsar. The characteristic features of this magnetospheric radiation are power law spectra and sharp pulses. Four of the radio pulsars seen with ROSAT, including the Vela pulsar, exhibit an additional thermal spectrum corresponding to a temperature of the order of 10^6 K, which is interpreted as the thermal radiation from the surface (photosphere) of the neutron star. A few point sources discovered near the centers of young supernova remnants also show very soft x-ray emission, which must be attributed to neutron star surface emission as well. A prominent example is the central source in Puppis A (figure 3(*b*)).

Since the ages of radio pulsars and supernova remnants can be determined quite well, one can use these observations to test cooling models of neutron stars. The result is that standard cooling models describe the observed temperatures and luminosities quite well. Rapid cooling as expected in the presence of a Bose–Einstein condensate of pions or kaons in the neutron star core can be excluded for the observed objects.

A few x-ray sources have been found in the ROSAT data showing very soft thermal spectra, (temperatures below a million kelvins) and very faint optical counterparts (25^m–26^m). These objects might be either single neutron stars accreting interstellar matter or, more likely, old (10^5–10^6 yr) cooling neutron stars whose supernova remnants have already vanished. The importance of all these observations lies in the fact that the surfaces of these tiny stars (with radii 10 km) have now become visible. Future x-ray spectroscopy should make it possible to measure their enormous gravity as well as their radii, which depend on the physical properties of matter at supranuclear densities.

The ROSAT deep survey of the Andromeda galaxy led to the discovery of 550 x-ray sources, comparable in number with the bright x-ray sources in our own galaxy found by Uhuru. As in the Milky Way, the brightest of them are young supernova remnants or binary systems with neutron stars or black holes accreting matter from a companion. These bright source populations have been studied with ROSAT in many galaxies. In addition, the hot interstellar medium, which is heated by supernova explosions, has been investigated.

A small fraction of all galaxies have an ACTIVE GALACTIC NUCLEUS (AGN) emitting huge amounts of energy in all spectral bands. The radiation is variable, indicating that it is emitted from a small region, generally not more than a few light-weeks or light-months across. Such AGNs often display jets, originating in their core. It is generally believed that the central engine is a supermassive black hole swallowing matter at a high rate. QUASARS are the most extreme representatives of the class of AGN sources. ASCA spectroscopy of the AGN MCG-6-30-15 and a few other sources has led to the discovery of iron Kα emission lines with broad and asymmetric profiles. This profile is most probably caused by gravitational and relativistic Doppler shifts of the rapidly rotating matter in the accretion disk near the central black hole. These AGNs are very bright x-ray sources, and more than 50% of all 150 000 ROSAT sources belong to this class. Because of their enormous brightness, they can be detected at large distances or redshifts. In the ROSAT Deep Surveys with a total accumulated observation time of 2 weeks, about 1000 sources are detected per square degree in the sky. Optical spectroscopy has shown that most of them are quasars and other AGNs at cosmological distances. The ROSAT Deep Surveys have also answered one of the oldest questions of x-ray astronomy, namely the origin of the extragalactic x-ray background: at least, in the soft x-ray band, around 1 keV, some 80% of the background has been resolved into discrete sources, mostly quasars. An important current question is what evolved first, the galaxies or the supermassive black holes.

Galaxies are not distributed randomly in the universe but form clusters or groups of galaxies consisting of a dozen to thousands of members which are gravitationally bound. With diameters of millions of light-years, they are the largest physical objects in the universe. As early as 1932, Fritz Zwicky found that CLUSTERS OF GALAXIES must

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

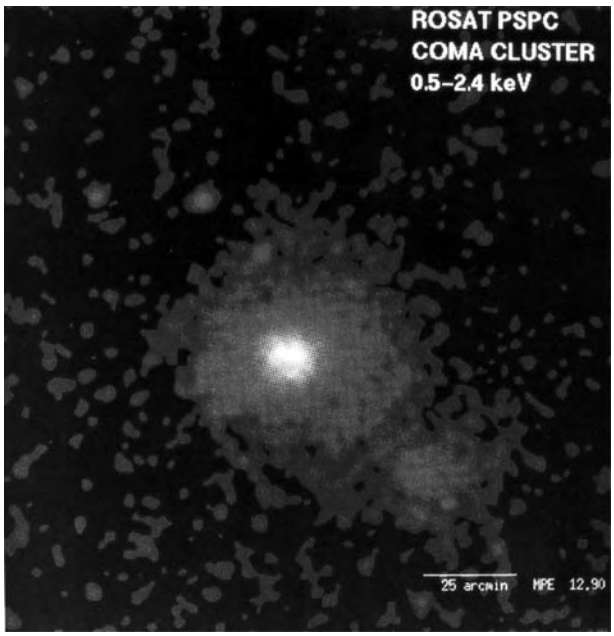


Figure 4. X-ray image of the coma cluster showing the merging with a smaller cluster of galaxies.

contain much more gravitating mass than accounted by the total visible mass of all member galaxies. Actually, the amount of 'DARK MATTER' had to be of the order of 95% of the total cluster mass.

One early surprise of x-ray astronomy was the detection by Uhuru of large quantities of hot plasma in clusters of galaxies shining in x-rays. Later Einstein and ROSAT observations showed that the mass of the hot plasma is typically a factor of 4 or 5 larger than that of the galaxies and represents some 20% of the cluster mass. In other words, a large fraction of the 'dark matter' turned out to be hot, visible only in x-rays and not in optical light. ASCA spectroscopy has allowed us to measure the heavy element (iron) abundance of the hot matter in clusters, which is typically only one-third of the so-called universal abundance. This can be explained by the infall of primordial matter onto the clusters.

Many clusters show double structures, indicating the merging of two clusters; others exhibit complicated inner structures which must be due to earlier merging processes (figure 4). Thus, the x-ray images reveal how these large objects evolve with cosmic time.

In total, several thousand clusters have been found with ROSAT. About 1500 of them have been optically identified and have known redshifts, and thus known distances. From this information one can derive the evolution of the cluster population with time. A comparison with simulations shows that the observed time dependence of the cluster population is significantly smaller than expected in a universe having the critical density. Actually the matter density inferred from cluster evolution is only about one-third of this critical density which is necessary to close the universe.

The future

During the years to come enormous progress is expected in x-ray astronomy owing to two new very powerful and complementary x-ray telescope missions: CHANDRA and XMM-NEWTON. Both satellites carry x-ray CCD detectors and dispersive spectrometers for high-resolution spectroscopy. The special strength of Chandra, launched in summer 1999, is its high angular resolution (∼0.5 arcsec) which allows us to resolve fine structures such as jets and to go at least a factor of 1000 deeper than the ROSAT deep surveys did. On the other hand XMM-Newton, launched in December 1999, provides a very high collecting power (∼4–10 times that of Chandra, depending on energy) at moderate angular resolution (∼10 arcsec). It will be very powerful for spectroscopic and time variability studies. The first results of these two powerful missions are very tantalizing.

Plans for the future thereafter are already quite concrete. NASA is discussing a fleet of four x-ray telescope satellites called 'Constellation' to be launched in ∼2007. They can observe the same object with different instruments simultaneously or point to different regions of the sky. ESA's XEUS (x-ray mission for spectroscopy in an evolving universe), to be launched after 2010, foresees a giant x-ray telescope with a huge mirror system and focal instruments sitting on two separate satellites, which means that pointing requires orbital maneuvers. The collecting area of the telescope (∼30 m^2) is a thousand times that of ROSAT and about a hundred times that of XMM or RXTE. It should be able to penetrate into the 'dark ages' of our universe at redshifts of 5–10, where galaxies and supermassive black holes have been formed.

Bibliography

Bradt H V, Ohashi T and Pounds K A 1992 X-ray astronomy missions *Ann. Rev. Astron. Astrophys.* **30** 391–427

Charles P and Seward F 1995 *Exploring the X-Ray Universe* (Cambridge: Cambridge University Press)

Joachim Trümper

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-ray Binary Stars

X-ray binary stars are luminous ($\sim 10^{36-38}$ erg s^{-1}) x-ray sources that consist of a neutron star or black hole compact object accreting material from a close (usually $\leq 1 R_\odot$) orbiting companion star which, in most cases, has properties similar to those of normal main sequence stars. Such x-ray binaries are rare (only ~200 are present in our Galaxy and are distributed in the Galactic Plane, the Galactic Bulge and a dozen or so as GLOBULAR CLUSTER X-RAY SOURCES, see figure 1), whereas the related (and in many respects similar) cataclysmic variables (where the compact object is a white dwarf) are much more common. The extremely high intrinsic luminosity of x-ray binaries (the brightest are at the 'Eddington Luminosity, a level at which radiation pressure limits the rate at which matter can accrete' for a $1 M_\odot$ star) accounts for their discovery in the early days of x-ray astronomy (via rocket-based experiments in the 1960s) and their domination of the field for most of the next two decades.

X-ray binaries can be divided into two basic classes of objects according to the mass of the mass-losing (or donor) star. The division is remarkably clean with those $\geq 5 M_\odot$ classified as 'high mass x-ray binaries' (HMXBs) and those $\leq 1 M\odot$ as 'low mass x-ray binaries' (LMXBs). The nature of the mass transfer from the donor also divides between these classes, with the (radiation driven) stellar wind dominating in HMXBs and Roche-lobe overflow being the principal mechanism in LMXBs (although there are circumstances in which a combination of the two can occur). The x-ray characteristics also depend on the nature of the compact object. High magnetic field neutron stars produce x-ray pulsators, whereas lower magnetic fields make it much harder to detect the spin of the neutron star, although 'x-ray bursts' can occur. Note that coherent pulsations and bursts are never seen in black hole systems.

X-ray binaries are extremely important objects as they demonstrate that stellar and binary evolution can produce compact objects via supernova explosions in which the binary survives. They currently provide the only circumstance in which stellar-mass black hole masses can be derived within our Galaxy (see BLACK HOLE CANDIDATES IN X-RAY BINARIES) as well as a majority of the accurately known neutron star masses. They are also a well-defined and constrained laboratory for studying accretion onto highly degenerate matter and the effect of ultra-high magnetic fields, and for testing the effects of general relativity in the strong gravity limit.

High mass x-ray binaries

X-ray pulsars

The very first x-ray astronomy satellite, UHURU, discovered X-ray pulsations at 4.8 s and 1.2 s in Cen X-3 and Her X-1, respectively, and also found that those pulsations were modulated (and eclipsed) on their orbital periods of 2.1 d and 1.7 d (see figure 2). These results immediately established the binary star model for these sources beyond any doubt. In combination with optical spectroscopy of the donor star (relatively bright, early-type stars for both sources), and the fact that they are eclipsing, they allow full orbital solutions and hence accurate masses for both donor and compact object (see table 1). Furthermore the duration of the eclipse provides an accurate estimate of the size of the donor which can be compared with the size of its Roche lobe.

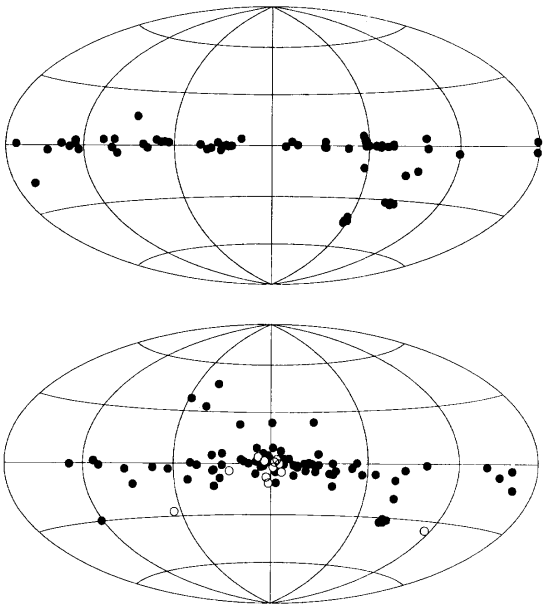


Figure 1. Distribution of x-ray binaries in the Galaxy. Top: HMXBs, associated with young, massive stars in the spiral arms. Bottom: LMXBs, an older population concentrated in the Galactic Bulge (also including globular clusters as open symbols).

Spinning up and down

The spinning neutron stars are orbiting their massive companions and accreting material from the dense and (presumed) uniform stellar winds. This physical situation is essentially identical to Bondi–Hoyle accretion as calculated for a single object moving and accreting from a uniform interstellar medium. Matter is accreted onto the neutron star from within a cylinder of radius $r_{\rm acc}$, defined as the point at which the kinetic energy of matter in the wind is equal to the potential energy due to the neutron star. If the wind's velocity is $v_{\rm w}$, and the orbital velocity of the neutron star is v then the fraction accreted is given by $(M_{\rm ns}/M_{\rm sec})^2 (v/v_{\rm w})^4 [1 + (v/v_{\rm w})^2]^{-3/2}$ and is usually $\sim 10^{-3}$ to 10^{-5} for typical HMXBs.

The angular momentum of this material will cause it to form a small accretion disk around the compact object, but its final capture by the neutron star is controlled by the intense magnetic field at the *magnetospheric radius* where the ram pressure of the (presumed spherical) infall is balanced by magnetic pressure. At this point the accreting angular momentum is transferred to the spinning pulsar, which then spins up at a rate $\dot{P}$ that is broadly proportional

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

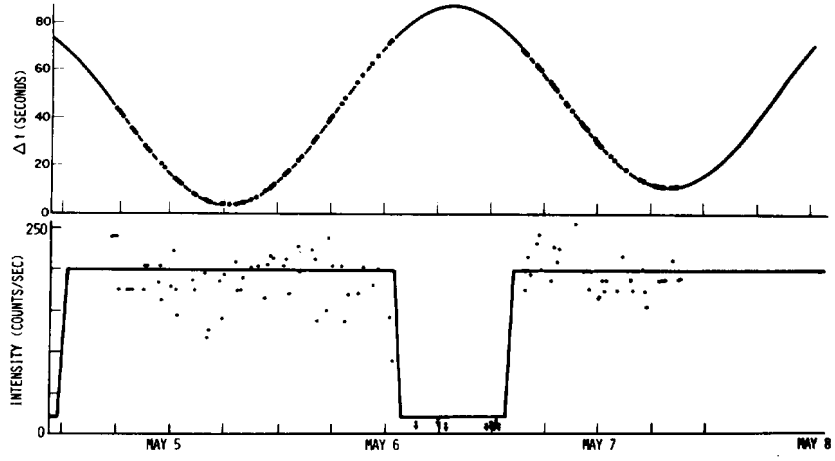


Figure 2. Uhuru observations of the Doppler delays in the 4.8 s x-ray pulsations from Cen X-3 as a function of the 2.1 d orbital period. The x-ray light curve (below) shows the eclipse and hence geometry of the binary.

Table 1. Orbital parameters and masses for eclipsing HMXBs.

Source	Pulse period (s)	Orbital period (d)	$f(M)$ ($M_\odot$)	e	i	$M_{\rm ns}$ ($M_\odot$)	$M_{\rm sec}$ ($M_\odot$)
SMC X-1	0.72	3.89	10.85	<0.000 04	65	1.1	18
Her X-1	1.24	1.70	0.85	<0.0001	> 79	1.5	2.3
Cen X-3	4.82	2.09	15.34	<0.0016	70	1.21	20.5
LMC X-4	13.5	1.41	9.86	0.006	65	1.5	15.8
Vela X-1	283	8.96	19.74	0.09	>74	1.9	23.5
4U1538-52	530	3.73	11.8	<0.06	68	1.1	16.4

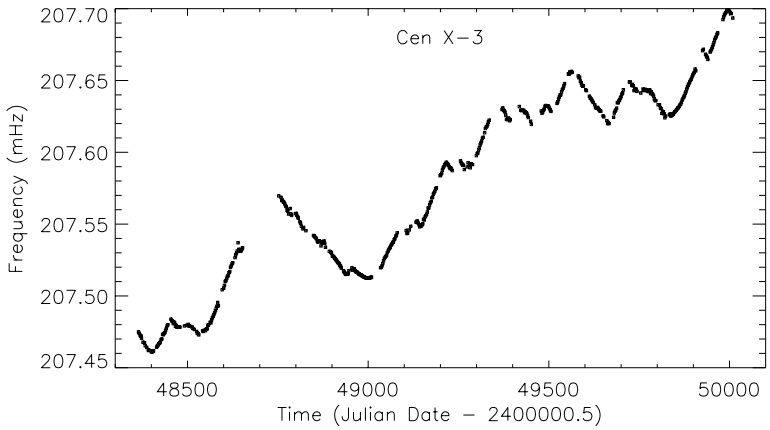


Figure 3. Approximately five-year history of the variations in the spin rate of the Cen X-3 pulsar as monitored by BATSE on CGRO. The general trend shows that on average it is spinning up, but there are significant fluctuations about this trend indicating that it can spin down as well as up.

to the observed x-ray luminosity (since this is directly related to the accretion rate).

Figure 3 shows the variation of spin period in Cen X-3 as observed by CGRO over almost five years. While the general trend is a spin-up, it is clear that the actual process involves periods of (almost equal) spin-up and spin-down. It might be expected that the spin-up rate would decline with x-ray luminosity if the mass transfer rate were to decline, but how does it manage to reverse sign? This requires that material is being accreted whose torque is reversed with respect to the 'normal' expectation, and implies that the disk rotation itself has reversed. Details of this process are unclear, but could be due to non-uniformities in the stellar wind.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

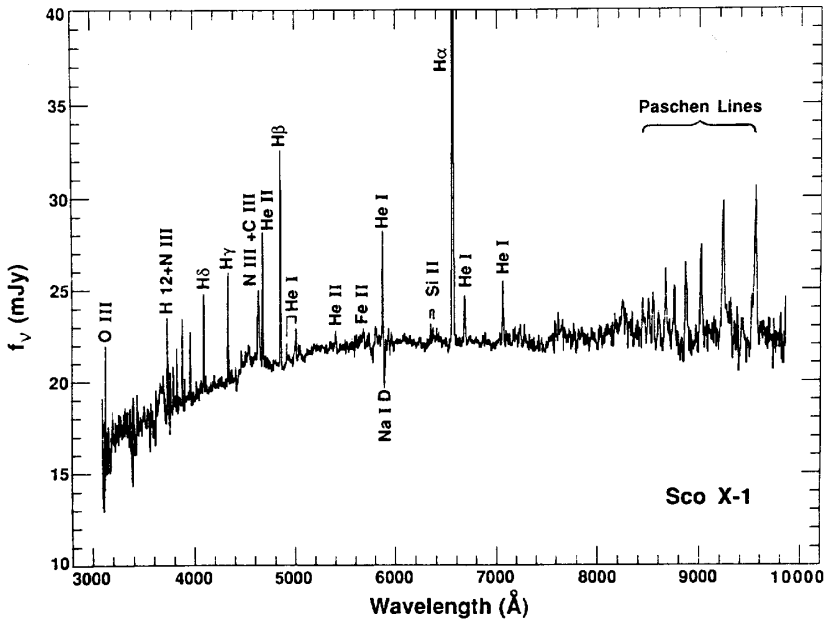


Figure 4. Optical spectrum of Sco X-1 which is dominated by continuum from the disk and x-ray driven high-ionization emission lines.

Table 2. Be x-ray binary parameters.

Source	Spectral type	V	Spin period (s)	Orbital period (d)	e	L_X(peak) (erg s^{-1})
A0538-66	B2IIIe	15	0.069	16.65	∼0.7	10^{39}
4U0115+63	OBe	15.5	3.6	24.3	0.34	8×10^{36}
V0332+53	Be	15.3	4.4	34.25	0.31	10^{38}
A0535+26	O9.7IIIe	9	104	111	∼0.3	2×10^{37}
GX304-1	B2Ve	14	272	133	?	3×10^{36}
4U1154-61	B1Vpe	9	292	188	?	10^{36}
X Per	O9.5ep	6	835	∼580	?	10^{34}

Note also that spin-up will only continue until the neutron star is rotating so fast that material at the magnetospheric radius would have to exceed the Keplerian speed at this radius in order to enter the magnetosphere (and hence would be ejected). Thus a magnetospheric 'barrier' then exists to counter any further accretion until either the neutron star has spun down or the mass transfer rate significantly increases.

Cyclotron lines

The magnetic fields of neutron stars that are x-ray pulsars are inferred to be $\sim 10^{12}$ G. Such fields control the accretion of material onto the neutron star's surface, leading to well-defined hot spots. x-rays are emitted either from the hot spot or from a shock immediately above the surface (within $\sim 10^{2-3}$ m), producing either pencil-beam or fan-beam emitting patterns. The very strong field will then lead to quantization of the energy levels within the accretion column, which produces resonant scattering cross sections and hence cyclotron emission and absorption features. In fact, cyclotron emission lines have only been seen in magnetic white dwarfs in the infrared, whereas cyclotron absorption features (most in the ∼12–40 keV range) have been detected in a dozen x-ray pulsars, implying magnetic fields of ~ 2–3×10^{12} G.

Be star transients

Many of the x-ray pulsars have long orbital periods with B-type companions which exhibit emission lines. These represent a sub-class of HMXBs and are associated with the rapidly rotating Be stars, so-called because of their strong and variable emission lines (usually hydrogen) superposed on otherwise normal B star spectra. A number of transient x-ray sources (most of them x-ray pulsators) have been identified with Be stars, establishing that both single and binary Be systems exist. The Be transients have rare outbursts (∼ hundreds of days) and long orbital periods (many weeks). Some (e.g. A0535+26, GX304-1) have poorly determined orbital periods >100 d, even with the accurately determined x-ray pulsation.

The Be phenomenon is interpreted as being due to the presence of an equatorial ring around the B star as

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

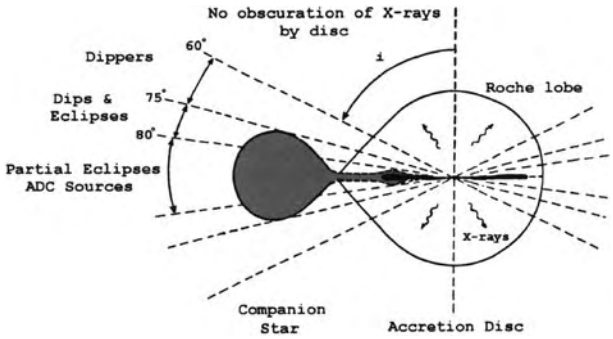


Figure 5. Schematic diagram of an LMXB showing the effects of disk shadowing and obscuration by the secondary star. Above an inclination of about 80°, LMXBs are seen only as ADC sources. Below about 75°, orbital modulation can still be seen in the form of regular dips, but these tend to evolve in form from cycle to cycle.

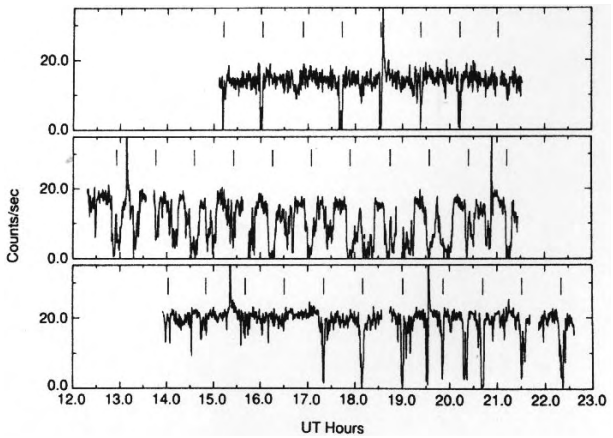


Figure 6. Continuous monitoring of X1916-053 by EXOSAT reveals both the regular x-ray 'dipping' due to azimuthal structure in the accretion disk, and the way in which this structure can evolve from orbit to orbit.

a result of its rapid rotation. Significant mass is lost via this ring and this material can be accreted by the orbiting neutron star, thereby producing x-rays. Almost all of the Be transients with measured orbits show significant orbital eccentricities (see table 2) and this may be related to the transient x-ray outbursts. An example here is A0538-66 in the LMC, which showed dramatic x-ray and optical outbursts on a 16.6 day cycle which was stable for at least 50 years (based on archival plate material), even though these outbursts did not always occur. At their peak these outbursts exceeded 10^{39} erg s^{-1} and yet the compact object was shown to be a neutron star based on an observed 69 ms pulsation. The most likely interpretation is that of a highly eccentric orbit ($e \sim 0.7$) in which the neutron star interacts (in a complex way) with the Be ring at periastron. Longer term (hundreds to thousands of days) variations are present in almost all Be systems, the nature of which is unclear. A recent extended (18 year) period of inactivity in A0538-66 has been attributed to the effect of the magnetospheric barrier given the very fast spin period of the neutron star. This could only be overcome by another interval of very high mass transfer from the secondary.

Low mass x-ray binaries

The brightest (and first discovered) extra-solar x-ray source is the x-ray binary Sco X-1, which was optically identified with an unremarkable 13th magnitude blue star in 1967. However, proof of its binary status was extremely hard to establish and this was not accomplished until the mid 1970s. This was due to the optical light being completely dominated by the x-ray heated accretion disk, making the secondary undetectable (see figure 4). Unlike HMXBs, the principal mass-transfer mechanism in LMXBs is Roche lobe overflow from the cool, low-mass secondary directly into the accretion disk. Furthermore, the thickness of the accretion disk itself can render a direct view of the compact object impossible at all phases (and not just when it is behind the secondary star), thereby making evidence for binarity less obvious than in HMXBs. This is demonstrated in figure 5 where the effect of shadowing by the disk leaves only a narrow range of viewing angles at which an orbital modulation can be seen. X-ray irradiation of the disk has been demonstrated in simultaneous multi-wavelength observations of X-RAY BURSTERS, which show the optical burst lagging the x-ray as a result of reprocessing of x-rays in the accretion disk and heated face of the companion.

A significant subset of LMXBs are globular cluster x-ray sources. Indeed, they are grossly over-represented in globular clusters when compared with the density of LMXBs in the Galaxy as a whole. Furthermore, all the globular cluster sources (see table 3) exhibit type I x-ray bursts, and recent HST studies of the extremely crowded UV/optical fields have revealed that several of these sources are ultracompact binaries with orbital periods $<$1 hour (including X1820-30 which has the shortest orbital period of any known object). It is likely that stellar interactions within the cluster core lead to the formation of such exotic objects.

X-ray dippers

A number of LMXBs display sudden 'dips' in their x-ray light curves which occur at particular orbital phases (around $\sim$0.8). These are attributed to extended vertical structure in the edge of the accretion disk at the point where the mass transfer stream impacts the disk. This material temporarily obscures our direct view of the x-ray source, absorbing the x-rays more effectively at lower x-ray energies (a signature of absorption which confirms the process). Extended observations of these x-ray dippers shows that the vertical disk structure evolves from orbit to orbit and on much longer timescales, sometimes disappearing altogether, at other times producing dip features throughout the orbital cycle (see figure 6).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 3. Properties of LMXBs.

Source	Orbital period (h)	X-ray type	V	M_V	Notes
X1820-303	0.19	Burster	>19		globular cluster (NGC 6624), degenerate companion
4U 1850–087	0.34	Burster	21	5.6	globular cluster (NGC 6712), degenerate companion
X1626-673	0.7	Burster, pulsar	18.5		7.7s pulsar, degenerate companion
X1832-330	0.73	Burster	19.4		globular cluster (NGC 6652), degenerate companion
X1916-053	0.83	Burster, dipper	21	5.3	x-ray and optical periods different, degenerate companion
J1808.4-3658	2.0	Burster, pulsar, transient	16.5–20		millisecond (2.5 ms) pulsar, $\sim 0.1 M_\odot$ companion
X1323-619	2.9	Burster, dipper			
X1636-536	3.8	Burster	17	1.3	
X0748-676	3.8	Burster, dipper, transient	17	1.4	no decline
X1254-690	3.9	Burster, dipper	19		
X1728-169	4.2		17		
X1755-338	4.4	Dipper	19		
X1735-444	4.6	Burster	17.5	2.2	
X2129+470	5.2	ADC	16.5		now 'off'; triple?
X1822-371	5.6	ADC	15.5		
X1658-298	7.2	Burster, dipper	18.3		
A1742-289	8.4	Burster, transient			eclipsing
X1957+115	9.3		18.7		
X2127+119	17.1	Burster, ADC	15.5	1.0	eclipsing, globular cluster (M15)
Aql X-1	19	Burster, transient	21.6	2.9	frequent outbursts, kHz QPO
Sco X-1	19.2	Prototype LMXB	12–14	0.0	Z source
X1624-490	21	Dipper			
2S0921-630	216	ADC	15.3		
Cyg X-2	235	Burster	14.7	−2.0	Z source
J1744-28	283	Burster, pulsar, transient			0.47s pulsar, type II bursts

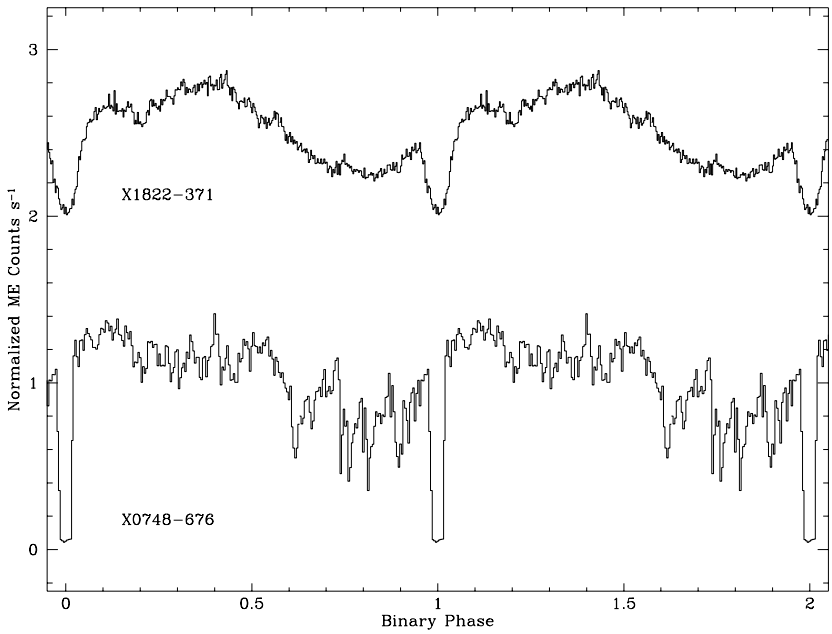


Figure 7. Comparison of the x-ray light curves of the prototype ADC source X1822-371 (top) with the eclipsing 'dipper' X0748-676. The orbital periods are very similar, but the orbital inclinations are only slightly different, with that of X0748-676 being just low enough to allow a direct view of the neutron star whilst still allowing the companion star to eclipse it (see also figure 5).

One particular source, EXO 0748-676, displays not only x-ray dips but also an (almost) total eclipse, albeit very brief (see figure 7). The inclination in this LMXB (a transient that turned on in 1985, and has remained 'on') must be in the very narrow range (73°–83°) that allows both dips and a short total eclipse by the secondary star.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Accretion disk coronae

However, a closer examination of figure 7 will show that the eclipse in EXO 0748-676 is not in fact completely total. There is a small but significant residual x-ray flux which is about 3% of the normal, out-of-eclipse level. This residual flux is due to x-rays which are scattered into our line of sight from material (possibly a disk 'wind') above and below the disk. Hence for sources at even higher inclinations, we would not expect to see total eclipses, but instead partial eclipses of this extended 'accretion disk corona' (or ADC). Figure 7 also shows the x-ray light curve of X1822-371, the classical ADC source which combines a long smooth modulation of the x-ray flux by the accretion disk rim structure with a partial eclipse by the secondary star. This is essentially what would be expected if the dipper light curve were to be smoothed out. ADC sources and x-ray dippers are important objects for the information that they provide on the azimuthal structure of accretion disks in interacting binaries.

X-ray pulsations

The much larger accretion discs in LMXBs compared to HMXBs should lead to substantial transfer of angular momentum onto the compact object and hence very fast spin periods (comparable to millisecond radio PULSARS). However, QUASIPERIODIC OSCILLATIONS IN X-RAY BINARIES were seen with large area x-ray detectors, an effect attributed to an interaction between the inner accretion disk and the neutron star magnetosphere. Very recently, the first millisecond x-ray pulsar was discovered (SAX J1808.4-3658) with a coherent pulsation at 401 Hz. It also displays x-ray bursts that indicate a weak ($<10^{10}$ G) magnetic field, and has a 2 h orbital period with a very low mass companion. This object is considered to be a direct link between LMXBs and the 'black widow' millisecond radio pulsars.

There is a further group, the 'anomalous x-ray pulsars' (AXPs), the nature of which remains very unclear. Only half a dozen are known, the best studied of which is 1E2259+587. They all exhibit coherent x-ray pulsations with periods in the range 5–12 s, have x-ray luminosities far too high to be explained by their observed spin-down rates and yet show no evidence for binary motion in these pulsations, nor has any been optically identified. It has been suggested that they may be linked to the 'soft gamma-ray repeaters' (a sub-class of gamma-ray bursters, see GAMMA-RAY ASTRONOMY) which also display similar pulsation periods during their bursts, an effect which led to the proposal that they had much larger magnetic fields than hitherto suspected (the so-called 'magnetars').

Long-term or superorbital modulations

From the early days of x-ray astronomy there have been satellites that scan the sky repeatedly and regularly (usually because they are spin-stabilized). This automatically provided the ability to monitor x-ray sources for many weeks (sometimes continuously), providing long-term light curves. These yielded the presence in a number of x-ray binaries, both LMXB and HMXB, of yet another modulation, this time quasi-periodic and on a timescale of tens to hundreds of days. The first discovered was the 35 d on–off cycle in HERCULES X-1, and which is attributed to a tilted, precessing disk. A similar ∼30 d modulation has been seen in LMC X-4, but a substantial change in this timescale over a year or so is difficult to account for. Subsequent long-term modulations (from 70 to 200 d) have been found in a variety of sources, including both LMXBs and HMXBs. These are currently attributed to the effects of a warped accretion disk driven by the central x-ray luminosity.

Bibliography

Bildsten L *et al* 1997 *Astrophys. J. Suppl.* **113** 367

Charles P A and Seward F D 1995 *Exploring the X-ray Universe* (Cambridge: Cambridge University Press) ch 7–9

Lewin W H G, Van Paradijs J and McClintock J E (eds) 1995 *X-Ray Binaries* (Cambridge: Cambridge University Press)

Philip A Charles

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-Ray Bursters

X-ray bursters are a class of sources which emit bursts of X-RAYS lasting from a few seconds to a few minutes. They are further classified as either type I or type II. Typical type I bursts have a faster rise than decay and the average energy of the x-rays decreases as the burst smoothly decays. Typical type II bursts stop abruptly without changing the average photon energy. There are now over 50 sources of type I x-ray bursts known, while only two sources are known as type II bursters, and one of these also has type I bursts. Because of this imbalance in numbers, 'x-ray burster' is assumed to mean a type I x-ray burster, although there are other kinds of sources of short bursts or flashes of x-rays, ranging from the Sun to black hole candidates. Type I bursters are NEUTRON STARS on which gas accreted from a binary companion undergoes sudden nuclear burning. The sources which have been called type II bursters are also neutron stars accreting gas from a binary companion. The type II burst is a sudden, large, but brief, increase in x-ray producing ACCRETION.

The neutron stars in x-ray bursters of both type I and type II have binary companion stars in close orbit such that the companion loses matter to the neutron star's gravitational attraction. (The companion fills the gravitational equipotential called the ROCHE LOBE.) The transferring matter has ANGULAR MOMENTUM around the neutron star, so that it circulates around it and mixes with gas already there in a flat ring distribution called an ACCRETION DISK. The disk material slowly moves inwards until it reaches the neutron star, where it accumulates on the surface. As helium and hydrogen accumulate under conditions of very high density and increasing temperature, the collisions of the ions can produce inelastic nuclear reactions which generate heat. The rising temperature increases the rate of collisions, raising the temperature further, until a nuclear fusion chain reaction is set off. This is the thermonuclear flash. The heat from burning the thin layer of helium, and sometimes hydrogen, diffuses out of the region. It heats the surface layers of the neutron star to a temperature at which x-rays are the dominant radiation emitted. Generally the radiation increases very quickly to its peak, sometimes in less than 1 s. The heat of the flash spreads to the surface and the burst decay corresponds to cooling of the surface.

From the burst x-rays and information about the distance or with assumptions about the peak brightness that the source can have, the radius and mass of the neutron star can be estimated. These values are close to those predicted for neutron stars and to the values measured in other ways, so that the conclusion that the star is compact enough to be a neutron star is firm. The measurements may eventually determine neutron star parameters, identify the composition of the matter falling on the neutron star and confirm ideas about the sequences of nuclear reactions that turn the accreted hydrogen and helium into the heavy elements of a neutron star's crust. The neutron stars have accreted angular momentum as well as mass. In some cases during the x-ray bursts the star is not uniformly hot and the flux has been found to be modulated with a period of 2–3 ms. These modulations are believed to indicate the spin of the neutron star and the characteristics of the modulation also provide information about the parameters of the neutron star. Thermonuclear bursts provide one of the few ways to obtain information about the densest matter.

Type I x-ray bursts

In 1975 there were three X-RAY ASTRONOMY satellites exploring the properties of sources found in the first x-ray survey mission, UHURU. Strong, short bursts of x-rays were first seen with the Astronomical Satellite of the Netherlands, then with SAS-3 and OSO-8, as sources in the region of the center of the Galaxy were observed. It turned out that the phenomenon had been seen as far back as 1969, by the Vela-5 satellites.

The sources from which the bursts appeared were in the direction of the bulge of the Galaxy, approximately 10 kpc away. From burst peak fluxes of 10^{-8} erg cm^{-2} s^{-1}, factors of 5–20 times normal or persistent flux of the sources, the peak luminosities are $\approx 10^{38}$ erg s^{-1}. This is close to the maximum luminosity possible for a $1M_\odot$ object, at which the radiation pressure due to photon scattering balances the force of gravity. (This 'Eddington limit' depends on the chemical abundances of the gas.)

The bursts were observed mostly in the photon energy range 1–20 keV. Generally they have a distribution with energy at a given time that is well approximated by a black body function with a temperature of $T \approx (1\text{–}4) \times 10^7$ K, corresponding to $kT \approx 1$–3 keV. Equating the estimated luminosity to the radiation rate from a spherical black body,

$$L_x = 4\pi R^2 \sigma T^4$$

the radius R can be estimated to be $\approx$10 km. This is consistent with the expected radius of a neutron star. It is one of the best pieces of evidence that the objects are neutron stars. Figure 1 contains schematics of the x-ray luminosity, temperature and apparent radius derived for a type I x-ray burst. The integral of the excess luminosity due to the burst is its fluence.

When a source is burst active, the bursts recur at intervals, usually of several hours to days, as illustrated in figure 2. The phenomenology points to a process in which the fuel builds up until it catastrophically ignites. The energy output is consistent with the nuclear burning of the accumulated matter. If the gravitational energy released in accretion is turned into x-ray flux $L_x = GM\dot{M}/R$, the mass accumulated in time T is

$$\Delta M = \frac{L_x}{GM/R} T.$$

Scaling to nominal observed values for L_x and T,

$$\Delta M \approx 10^{21} \frac{L_x}{10^{37}\ \text{erg s}^{-1}} \left(\frac{R}{10\ \text{km}} \Big/ \frac{M}{1.4 M_\odot} \right) \frac{T}{5\ \text{h}}\ \text{g}.$$

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

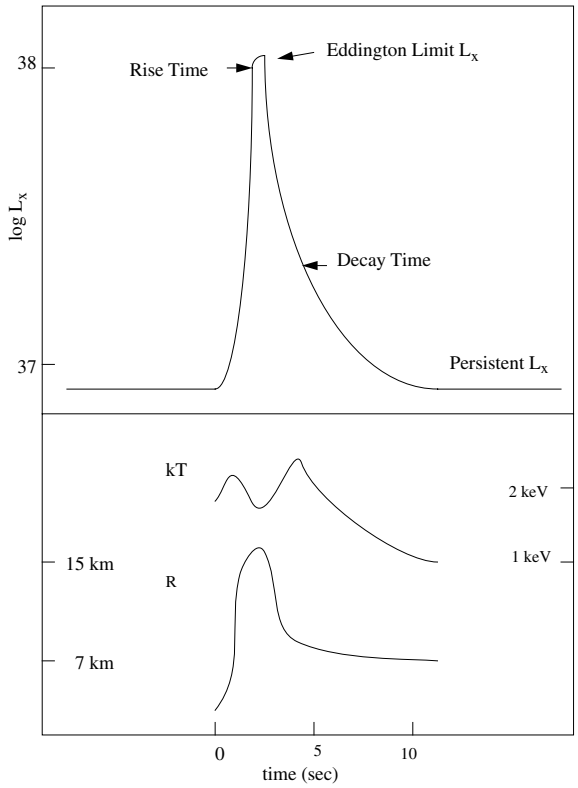


Figure 1. Type I burst parameters. The upper panel shows a schematic of the x-ray luminosity of a burst as a function of time. The lower panel shows a typical behavior of the fitted black body kT (right-hand scale) and the deduced apparent radius of a spherical emitter (left-hand scale).

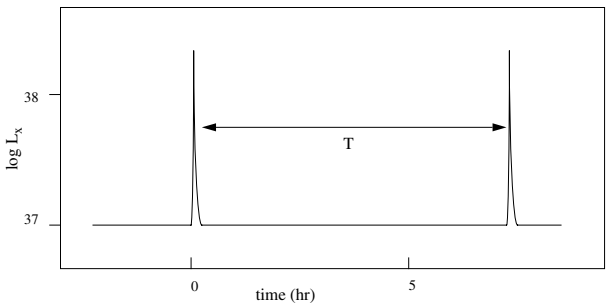


Figure 2. Persistent and burst luminosities. The average burst luminosity is the fluence in a burst per the recurrence time T.

There are a number of nuclear reaction sequences by which accreted material with the abundances of the outer layer of the companion star can convert into the iron and higher atomic number elements of the neutron star crust. The amount of energy released per nucleon is 2–8 MeV, so that the potential for energy release from burning, in the mass accumulated, is

$$\Delta E = (15\text{–}76) \times 10^{38} \times \frac{L_x}{10^{37}\ \mathrm{erg\ s^{-1}}} \left(\frac{R}{10\ \mathrm{km}} \Big/ \frac{M}{1.4 M_\odot} \right) \frac{T}{5\ \mathrm{hr}}\ \mathrm{erg}.$$

If the peak flux were the Eddington limit $2.0 \times 10^{38} M/1.4 M_\odot$ erg s^{-1} (assuming cosmic abundances here), it would take 7–40 s to cool the hot layer, with the heat convecting and diffusing to the surface in $\leq$1 s.

Generally, in the thermonuclear flash, the ratio of the persistent x-ray luminosity to the average burst luminosity (the fluence per burst divided by the recurrence interval) should be the ratio of the gravitational energy released per gram falling on the neutron star to the nuclear energy released in burning it. The observational quantity, denoted α, is equal to the ratio of the fluxes. The predicted values of α are in the range 25–200, depending on the nuclear reaction track. Some reaction tracks imply particular correlations between burst characteristics and the persistent luminosity. Theoretical calculations indicate that the nature of the reactions is more likely to go through cycles than to be a steady state, as the accretion rate changes and with it the ambient temperature of the settling material. The general agreement is regarded as a strong confirmation of the thermonuclear flash model of type I bursts.

Type II x-ray bursts

Type II x-ray bursts usually (but not invariably) have the ratio of persistent luminosity to burst luminosity (α) too small to be consistent with what is known about thermonuclear reactions, well below 25. The first and recurrent example of this type is MXB 1730 − 335, appropriately called the Rapid Burster. The source becomes bright approximately every 6 months and for a few weeks goes through several metamorphoses of behavior, sometimes exhibiting rapid-fire bursts a few minutes apart. The Rapid Burster is also a source of type I bursts, so that the two different types of bursts can be compared for the same source. The second source was the transient GRO J1744 − 28, which first became known in 1995 as the Bursting Pulsar. This transient has reappeared once, so that it is recurrent, although irregularly. Bursts typically came about 30 min apart when it was active.

An increase in the rate at which matter falls on the neutron star can explain these bursts. Several possible mechanisms have been identified to store up the matter coming in, at some distance from the neutron star, and to release it suddenly to complete its journey to the neutron star.

The matter in accretion disks works its way to smaller radii as viscous interaction with neighboring cells of the accretion disk removes its angular momentum. As matter is continuously fed into the disk by the companion star, the viscous properties change and a sudden transition can occur from a state of low viscosity to high viscosity. The higher torques allow matter to start moving rapidly through the disk. The emission that we see as the burst can come both from the disk itself and from the neutron star where the matter falls on it. This instability can operate on

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

a range of recurrence time-scales, from years to minutes, so that it could give rise to bursts as well as to dwarf and x-ray NOVAE.

A different possible cause of the intermittent flow onto the neutron star is the interaction of the neutron star's magnetic field with the matter approaching the star. The magnetic field of the neutron star, if strong enough, interacts with the inner accretion disk. In the case of strong field PULSARS at some radius the field channels the disk material out of the plane and along the magnetic field lines to the magnetic poles. For the case of weaker fields, the disk may interact with the field over an extended region. It has been suggested that material could accumulate outside the magnetosphere until some threshold condition occurred that would allow the material to suddenly feed onto the neutron star.

For the Bursting Pulsar, pulsations with a period of 0.5 s continued throughout the bursts. Considerations of how low the accretion rate can go before pulsations cannot occur have allowed the surface magnetic field to be estimated at 5×10^{11} G. The Rapid Burster is likely to have a field weaker by several orders of magnitude and possibly the two sources represent two different instability mechanisms.

Thermonuclear flashes

Hydrogen and helium burning

As the accreted material collects on the neutron star, it will spread out over the surface if the magnetic field is too weak to confine the plasma (below 10^{11} G). If the companion is still on the main sequence, the material is, by weight, 70% hydrogen, 30% helium, with no more than a few per cent in heavier elements (C, N, O). However, in at least one case, 4U 1820−30, the companion is known to be a helium WHITE DWARF. When there is a layer of $\approx 10^{21}$ g on the surface of the neutron star, the density at the bottom of the accreting layer reaches $\approx 10^5$ g cm^{-3}. Nuclear reactions begin to convert hydrogen into helium and helium to oxygen. These reactions release heat. At very low and very high rates the heat can be radiated as fast as the energy is released and the hydrogen burning is stable, but under many conditions the burning is unstable, and the temperature increases rapidly until the fuel is completely burned. Helium burns by the triple-alpha reaction $3\alpha \rightarrow {}^{12}\mathrm{C}$ and by ${}^{12}\mathrm{C}(\alpha, \gamma){}^{16}\mathrm{O}$, usually unstably. Detailed calculations have examined the kind of succession of bursts that should occur, for a range of accretion rates, central neutron star temperatures and neutron star masses and radii.

Observed and modeled burst properties

For the intermediate accretion rates at which instabilities occur, three regimes have been distinguished. Helium burns unstably but, depending on the accretion rate, hydrogen, if present, may burn stably or unstably. In the middle of these three regimes the hydrogen burns stably (unless the interior temperature is very hot) and produces helium which flashes with $\alpha \geq 100$, recurrence times $\geq$10 h, and burst rise times $\leq$1 s. At either higher or lower accretion rates, the hydrogen and helium burn together, with complex reactions, building to iron and heavier elements. These reactions give the burst a longer rise time and can leave unburned fuel, so that the thermonuclear luminosities, and therefore the values of α, are lower.

Type I bursts from different sources are observed to be similar, but not identical. Bursts rise times range from 0.3 to 10 s and decay times range from 5 to 100 s. The decay times (τ) and the values of α vary with the persistent luminosity. The EXOSAT Observatory caught several sources in different states. Comparison of the behavior of the parameters with the predictions of models has led to investigations of other degrees of freedom, for example, dependence of the internal temperature on the history of the burning and the extent of the burning area on the neutron star.

For many of the x-ray bursters the persistent luminosities vary, sometimes consistently, but not exactly periodically. Some are transient persistent sources in which accretion may hardly occur in between episodes years apart. Then they rise to maximum in hours and decay over several weeks to a year. The sources 4U 1608−52 and Aquila X-1 are regularly recurrent transients. As the persistent flux may vary by factors of a thousand and the supply of mass to the neutron star envelope is believed to vary accordingly, the integrated effect must be considered.

Radius expansion during bursts

As the temperature during bursts rises and falls, the projected area of the hot emitter can be tracked. For many burst measurements the size of the emitter has been consistent with it remaining approximately constant throughout the burst. For a subset of bursts the apparent size increases near the beginning of the burst to values as much as 10 times sizes theoretically estimated for neutron stars. The apparent size then drops back and levels off at an asymptotic value consistent with that of a neutron star (as in figure 1). In these cases the bolometric luminosity stays approximately constant during the expansion episode. The temperature goes down during the expansion. The peak luminosity appears to be the Eddington limit for helium for most 'radius-expansion' bursts.

Theoretically, helium flashes can generate enough energy to cause photospheric expansion of the neutron star. When the surface luminosity reaches the Eddington limit, it drives a wind from the neutron star. There is evidence within bursts from a single source that some reach a plateau of luminosity $\approx$1.7 times lower than the radius-expansion plateau luminosity of others. This has been interpreted as the Eddington limit for hydrogen-rich gas, a factor 1.7 below that of helium. These characteristics can be identified from the spectra, without knowing the distance of the source, and are used as the basis for estimates of the distances to some burst sources. In the cases in which a distance is known on other grounds, they are in approximate agreement.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

General relativity corrections and radiation transfer

The assumptions of black body emission and Eddington-limited flux seem to offer the possibility of measuring the radii and masses of neutron stars. These are interesting because they depend on the properties of the matter at densities of 10^{15} g cm^{-3} at the center of the neutron star. At least seven of the type I bursters are in GLOBULAR CLUSTERS and some of their distances are known to within 10%. However, several important effects must be included in interpreting the measurements. For neutron stars, the corrections of general relativity can be important; for a nominal $1.4M_\odot$ star, the Schwarzschild radius is 40% of the nominal radius of 10 km and the relativistic effects can be larger for some equations of state. Further, the surface temperature of the neutron star exceeds 10^7 K and the neutron star has a scattering atmosphere. The surface has an energy-dependent emissivity. These effects need to be taken into account in understanding and using the observations to determine the neutron star parameters.

A spherical black body neutron star has an apparent radius larger than the radius at the star, because of the gravitational bending of rays. Taking this into account, with

$$1 + z_* = (1 - R_g/R_*)^{-1/2}$$

$$R_\infty = R_*(1 + z_*)$$

(as long as $R_* \geq (3\sqrt{3}/2)R_g$), where $R_g = 2GM/c^2$ is the Schwarzschild radius, R_* the real neutron star radius and R_∞ the apparent radius we measure, which exceeds the real radius. The Eddington limit for the luminosity as observed by a distant observer gives another constraint on M and R_*. Together with possible theoretical mass–radii relations the Eddington limit implies a limit on the 'effective' temperature (defined by $T_e = (L_x/4\pi R^2\sigma)^{1/4}$), such that kT_e should never exceed 2 keV. The black body fits determine a 'color' temperature T_c, which at its peak often exceeds this limit on the effective temperature. This is a measure of the error in the assumption of black body emission. When the competition between scattering and absorption, the temperature increase with depth in the atmosphere and incoherent scattering in the atmosphere are all included, the ratio T_c/T_e has been found to vary between 1.4 and 1.7 as the flux approached the Eddington-limited flux.

Detailed data for bursts have been obtained by x-ray experiments on many satellite missions, notably OSO-8, SAS-3, HEAO-1, EXOSAT, Tenma, Ginga, RXTE and BeppoSAX. Extensive work was done attempting to determine the properties of the neutron star using the EXOSAT and Tenma data. Measurement and perhaps systematic errors prevented definitive results from being obtained. Similar projects to examine RXTE observations, which have much smaller statistical error and higher time resolution, are still in progress.

The neutron stars

X-ray bursters are in very old BINARY SYSTEMS, as shown by the population in globular clusters. At least three may well have degenerate dwarf companions (4U 1916 − 053, 4U 1820−30 and SAX J1808.4–3658) and all of the companions are low-mass stars. It has been suggested that in accreting over 10^8–10^9 yr the neutron star's magnetic field has decayed. When radio pulsars with 10^8–10^{10} G fields were discovered with rotation periods of a few milliseconds, it was suggested that accretion in the low-mass x-ray binaries spins up the neutron stars to millisecond periods. Then irradiation could evaporate the companion, leaving a rejuvenated, millisecond, radio pulsar. Recent observations with RXTE discovered evidence tending to confirm this scenario. The transient and burster SAX J1808.4–3658 is now known to be an x-ray pulsar. It has a magnetic field estimated to be 5×10^8 G and a rotation frequency of 401 Hz (a period of 2.5 ms).

In six other bursters an inhomogeneity in temperature on the star appears during bursts and appears to rotate. The periods are similar to the period of SAX J1808.4–3658. At the beginnings of bursts a high-amplitude (as high as 60%) sinusoidal oscillation starts. The oscillation amplitude decays along with the rise in flux. This behavior would be consistent with a growing burning region. The periods are 1.7–3 ms. During the decays, 10 s trains of oscillations occur with typical amplitudes of 10%. Whether the asymmetry is associated with one spot or two is uncertain. Together with new and related results on fast quasi-periodic oscillations in the persistent flux, they provide new, and potentially precise, tools to determine properties such as the masses and radii of the neutron stars, their magnetic fields, the radii at which the disks terminate and how the accreting gas traverses from the disk to the neutron star. With RXTE, BeppoSAX, Chandra and XMM it will be possible to address many questions about x-ray bursters, the evolution of the systems, the flow of accreting gas, the progress of thermonuclear flashes and the neutron star solutions for collapsed stars of $\leq 3M_\odot$.

Bibliography

The literature on the subject of x-ray bursters includes a number of reviews, which are well represented in the chapters by

Lewin W H G and Joss P 1983 X-ray bursters and the x-ray sources of the Galactic bulge *Accretion Powered X-Ray Sources* ed W H G Lewin and E P J van den Heuvel (Cambridge: Cambridge University Press) pp 41–115

and by

Lewin W H G, van Paradijs J and Taam R E 1995 X-ray bursts *X-Ray Binaries* ed W H G Lewin, J van Paradijs and E P J van den Heuvel (Cambridge: Cambridge University Press) pp 175–232

Brief summaries of the newest results are in

Bildsten L and Strohmayer T 1999 *Phys. Today* **52** 4

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

and in articles in

Bucceri R, van Paradijs J and Alpar M A (ed) 1998 *The Many Faces of Neutron Stars* (Dordrecht: Kluwer)

and in

Scarsi L, Bradt H, Giommi P and Fiore F (ed) 1998 *The Active X-Ray Sky* (Amsterdam: Elsevier)

Jean Swank

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-Ray Telescopes

X-ray telescopes produce images of x-ray-emitting objects within the telescope's FIELD OF VIEW by reflection from precisely shaped mirrors. Hans Wolter's design in the early 1950s of an x-ray microscope using reflective optics led Riccardo GIACCONI to suggest an 'inverted' set of optics, not subject to the fabrication limitations of the microscope, could be used as a cosmic x-ray telescope. As described in 1960 by Giacconi and Bruno ROSSI, imaging with x-ray telescopes offers a significant reduction in noise, both from cosmic ray induced events and the soft x-ray background. This is because most background counts are the result of charged particles and are uniformly distributed over the detector area. An imaging system concentrates the source counts in one or a few pixels. As an example, the EINSTEIN (HEAO-2) Observatory x-ray telescope offered an improvement in sensitivity of about a factor of 1000 over that of the large-area, collimated proportional counters on High Energy Astrophysical Observatory (HEAO) 1 satellite. The improved instrument sensitivity resulting from increased signal to noise enables the study of fainter sources, extending the number and age of detectable objects.

Imaging observations enable deeper study of extended cosmic x-ray sources. Imaging of CLUSTERS OF GALAXIES reveals details of the intracluster gas temperature profile and isolates the x-ray emission from component galaxies. Imaging of GALAXIES allows study of any central source as well as the identification and location of the galactic emitters. Study of SUPERNOVA REMNANTS allows the visualization of the x-ray-emitting regions and potential identification of a remnant's compact object. None of these types of studies is readily undertaken with non-imaging instruments. Other imaging approaches employing such devices as modulation collimators or scanned, tightly collimated detectors provide much poorer resolution (∼1 arcmin.) than that available with current x-ray telescopes and suffer from image reconstruction artifacts.

Studies of point sources also benefit from x-ray telescope resolution. Using the large-area proportional counters on board HEAO-1, positions accurate to ∼0.1 square degrees were obtained for weak point sources. On average, there will be one star brighter than 12th magnitude in the position error box. For comparison, the optical counterparts of Sco X-1 and Cen X-1 are about 13th magnitude, and that of Her X-1 about 15th magnitude. Imaging enables more accurate optical identifications. With better than 0.5 arcsec resolution of the Chandra X-ray Observatory (CXO) positions of point sources accurate to $\sim 10^{-7}$ square degrees may be obtained—on average less than a single unrelated 20th magnitude star will be found in the error box.

This article will first briefly discuss how x-rays may be reflected and what optical systems can be used to form images. Key considerations for performance and system design will be discussed, followed by the various ways of implementing x-ray telescopes. We will review the more significant past and current x-ray telescopes, and then discuss future trends in x-ray telescopes.

X-ray reflectance

X-rays specularly reflect from a surface under two conditions, when striking a surface at grazing incidence and when constructive interference takes place between many layers of a material whose atomic number varies in a periodic fashion.

Reflection at grazing incidence

At x-ray wavelengths the real part n_r of the complex index of refraction $n_r - ik$ of the reflecting surface is less than 1. Thus x-rays in vacuum are incident upon an interface that is less optically dense. From Snell's law the angle of the wave transmitted θ_t into the less dense material is

$$\sin\theta_t = (n_1/n_2)\sin\theta_i \tag{1}$$

where θ_i is the angle of incidence and n_1 and n_2 are the indices of refraction for vacuum and the reflecting material, respectively. When $n_1 > n_2$ one obtains a real value for θ_t only for incident angles $\theta_i < \sin^{-1}(n_2/n_1)$. For example, if $n_2 = 0.99$, then the maximum real value for θ_i is 81.9°: at this incident angle the transmitted angle is 90°, parallel to the interface. For larger incident angles (shallower grazing angles) no radiation is transmitted. This is seen by allowing θ_t to be complex, with

$$\cos\theta_t = \pm i[(n_1/n_2)^2\sin^2\theta_i - 1]^{1/2}. \tag{2}$$

Substituting equation (2) into the Fresnel equations, solving for the transmitted electric and magnetic fields, and determining the time-averaged Poynting vector reveals that no energy flows through the interface—the field intensities normal to the interface decay exponentially into the less optically dense surface.

Reflectance is calculated from the complex index of refraction using the Fresnel equations. The real and imaginary parts of the index are computed from the atomic (forward) scattering factors f_1 and f_2:

$$n_r = 1 - \delta = 1 - \frac{\rho}{W_m} A_0 r_e \lambda^2 f_1 \frac{1}{2\pi}$$

$$k = \frac{\rho}{W_m} A_0 r_e \lambda^2 f_2 \frac{1}{2\pi} \tag{3}$$

where ρ is the material density, W_m is the molar weight of the reflecting material, A_0 is Avogadro's number, r_e is the classical electron radius and λ is the incident wavelength. The scattering factors are material and wavelength dependent. They are derived in a variety of ways: they may be computed from measurement of the forward scattering amplitude or back calculated from measured x-ray reflectance. Well-known sets of constants have been published by Henke *et al*, Auerbach and Tirsell, and Windt. As better reflectance is achieved with materials with high electron density grazing incidence optics are typically coated with metals to enhance reflectance.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

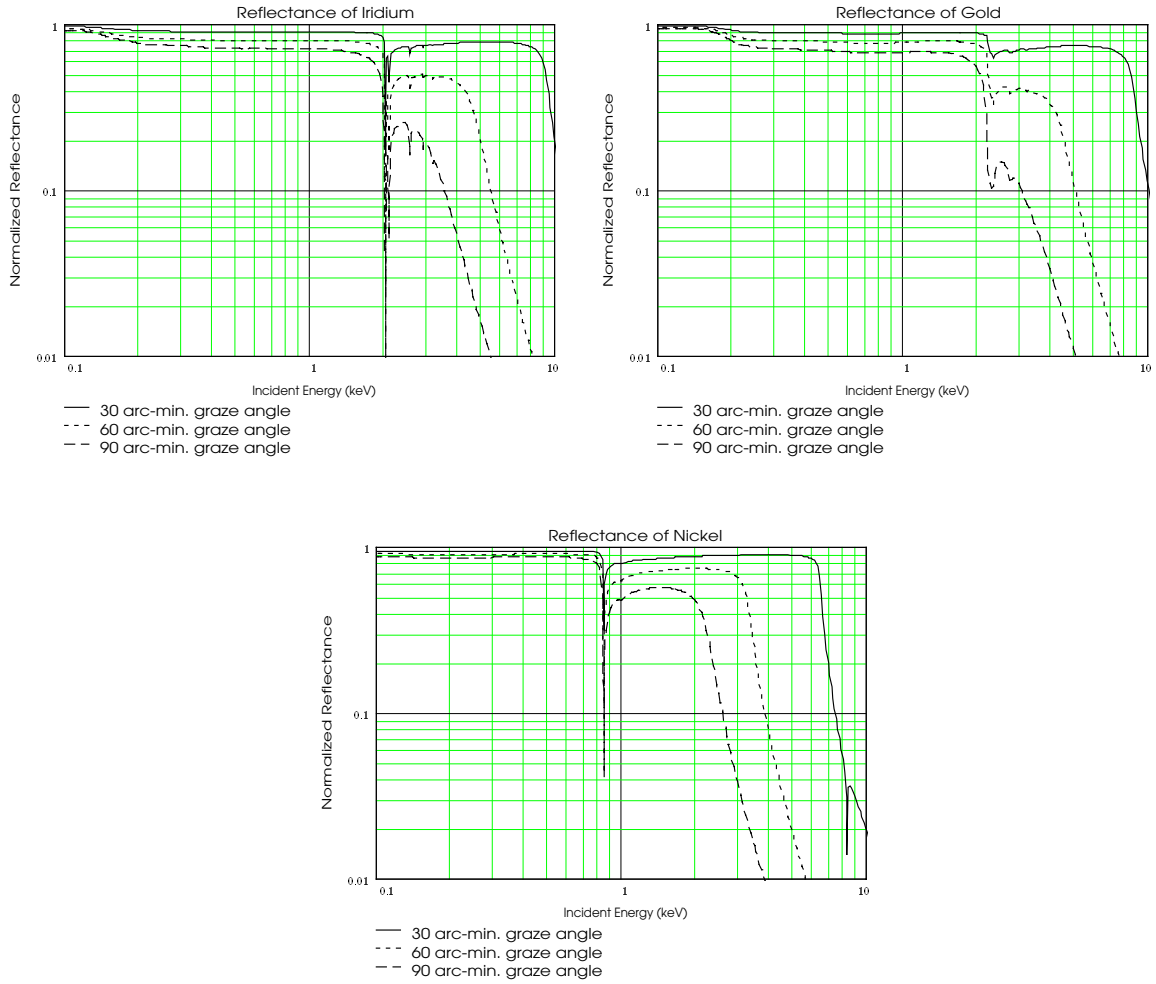


Figure 1. X-ray reflectance as a function of incident energy and graze angle.

Common materials used as reflecting coatings are iridium, gold and nickel. A plot of x-ray reflectance as a function of incident energy and grazing angle is shown in figure 1. Note the local minima in reflectance at the location of x-ray absorption edges. The product of reflectance and geometric collecting area (the entrance aperture) is called the effective area. Reflectance decreases as a function of increasing graze angle and starts to decrease rapidly near the so-called critical angle. The critical angle may be expresses as $\cos^{-1}(n_r) \approx (2\delta)^{1/2}$ for $\delta \ll 1$ (see equation (3)).

Reflection from multilayer coatings

Multilayer coatings contain alternating layers of high-Z (atomic number) and low-Z materials. Reflection occurs at each interface, and the layer optical thicknesses are chosen so as to produce constructive interference between each layer. Reflection occurs only for those wavelengths which satisfy the Bragg equation,

$$m\lambda = d \sin\theta \qquad (4)$$

where d is the layer spacing and θ is the incident angle. In practice the coating reflects a bandwidth $\Delta\lambda$ that is a function of various fabrication parameters including layer thickness uniformity, density, and layer materials but is typically only a few ångströms wide. Typical layer material pairs include (but are not limited to) nickel/carbon, tungsten/silicon, rhodium/beryllium, rhodium/carbon, molybdenum/carbon and platinum/carbon. A typical coating may contain between 40 and 500 layer pairs, with the larger number of layers yielding higher reflectance. Reflectance of 10–80% has been achieved, depending on the incident wavelength (higher reflectance at longer wavelengths) and material choices.

Several soft x-ray–EUV, normal incidence, multilayer telescopes have been launched on board sounding rockets and small satellites, including the recently launched Transition Region and Corona Explorer (TRACE). To overcome the spectral bandwidth limitations of multilayer coatings, experimenters have resorted to integrating several smaller telescopes into a single payload, each telescope with a different multilayer. In

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

the case of TRACE, different multilayer coatings were applied to each of the four quadrants of the optics so that a single telescope was sensitive to four separate spectral bandwidths (one of which was visible light).

An alternative to several different telescopes or multiple different coatings on a single telescope is graded multilayers, or 'supermirrors'. Here, the layer thickness varies as a function of position (depth) in the layer stack. The thickest layers are at the top of the stack to reflect the longest wavelengths with a minimum of absorption. Good broadband reflectance has been achieved using a thickness law

$$d_i = a(b+i)^{-c} \tag{5}$$

where d_i is the thickness of the ith layer (starting from the top), a and c are positive parameters that are a function of the materials used, and b is a constant greater than -1.

Substrate and coating roughness degrades multilayer performance. 'Low' frequency roughness ($>1\,\mu$m periods) causes scatter which degrades image contrast. 'High' frequency roughness ($<1\,\mu$m periods) reduces reflectance.

Types of telescopes

Two approaches are used to achieve imaging: grazing incidence (GI) and normal incidence (NI). Grazing incidence telescopes utilize large angles of incidence (near 90°) so that the x-rays graze or glance off the surface and are reflected. Normal incidence x-ray telescopes use more conventional shaped optics with multilayer coatings.

Grazing incidence telescopes

In 1952 the German physicist Hans Wolter was able to show that a two-element system containing an even number of confocal conic optics will come close to satisfying the Abbe sine condition; rays that reflect off both surfaces are focused and form an image. Wolter, attempting to produce systems for x-ray microscopy, produced three designs, referred to as Wolter types I, II and III. The type I telescope (figure 2) consists of a paraboloid as a primary mirror and a confocal and coaxial hyperboloid as the secondary mirror. The paraboloid focus is coincident with the back hyperboloid focus. X-rays strike the paraboloid at the grazing angle (approximately the field angle of the source plus half the angle subtended by the best-fit cone to the paraboloid), are reflected and strike the hyperbola. The image is formed at the front hyperboloid focus. Usually, the hyperboloid cone angle is three times that of the paraboloid so that on-axis x-rays are incident upon both mirrors with essentially the same grazing angle. X-rays that strike the forward end of the paraboloid reflect to strike the aft (back) end of the hyperboloid, and vice versa.

The entrance aperture is the projection of the primary mirror in the aperture plane. This results in an annular aperture whose width is approximately the product of the optic length and the half-angle of the best fit cone (the cone angle). For the CXO the largest paraboloid has a surface area of approximately 3.2 m^2 but an entrance aperture of only 0.047 m^2, a reduction of a factor of

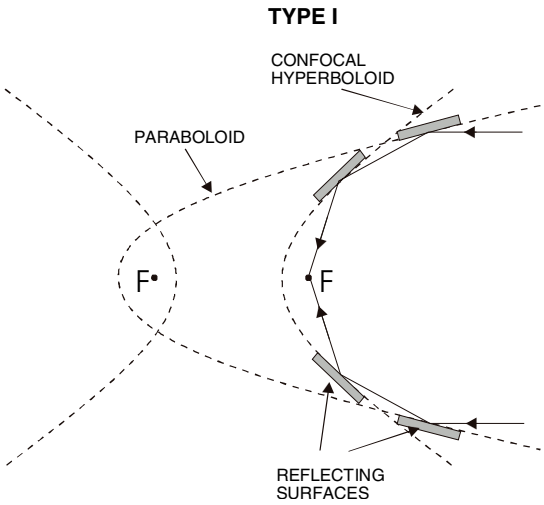


Figure 2. Schematic representation of the Wolter type I telescope.

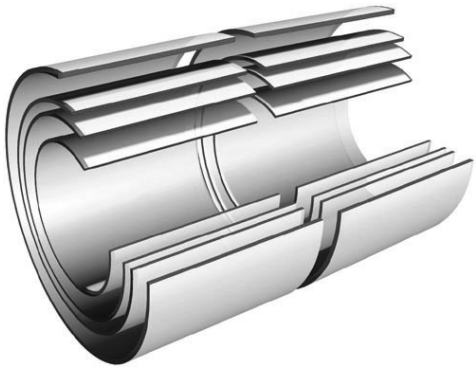


Figure 3. Cutaway schematic drawing of the nested four shells (paraboloid and hyperboloids) of the Chandra X-ray Observatory (courtesy of the Raytheon Co).

~68 times. Collecting area is built up by nesting mirror pairs within one another (figure 3). Each mirror pair, or shell, is co-aligned and confocal. The telescope entrance aperture consists of a set of concentric annular apertures, all contributing to the same focus. Ideally, the shells are designed so that each one has the same focal length and therefore the same plate scale (the proportionality constant relating angular distance on the sky to linear distance at the focal plane). All shells do not necessarily contribute the same to the image as a function of x-ray energy. As shown above in the section 'X-ray reflectance', reflectance is a function of graze angle and incident energy. The inner shells have a shallower graze angle than outer shells. Thus the inner shells of a nested telescope have a larger spectral bandwidth (reflect higher-energy x-rays) and comparable or slightly higher reflectance than the outer shells, while also having a smaller entrance aperture.

The type I design yields a perfect image for on-axis illumination. Off axis, the design suffers from field dependent coma and spherical aberration. In addition,

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

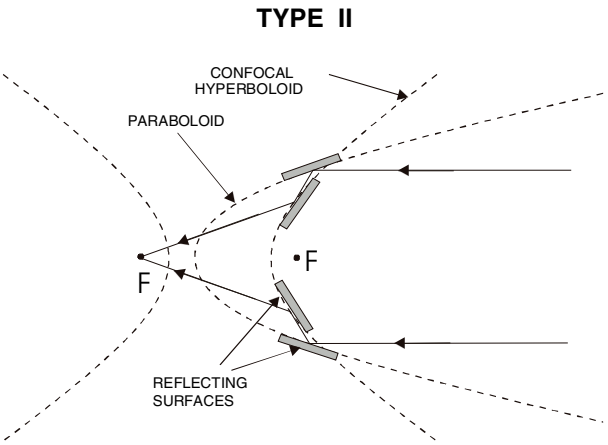


Figure 4. Schematic representation of the Wolter type II telescope.

the telescope focal surface is curved. Van Speybroeck and Chase (1972) approximated the off-axis rms blur diameter for a flat focal plane detector at the Gaussian focus as

$$\sigma_{\text{blur}} \approx 4\frac{\xi+1}{10}\frac{(\tan\theta_{\text{field}})^2}{\tan\alpha}\frac{L}{Z_0} + 8\tan(\theta_{\text{field}})(\tan\alpha)^2 \qquad (6)$$

where θ_{field} is the field angle (the off-axis angle), α is the graze angle, Z_0 is the axial distance from the intersection of the paraboloid and hyperboloid surfaces (the 'virtual joint') to best on-axis focus, and L is the axial length of the paraboloid. The first term in equation (6) represents the field-dependent coma. The second term is split between a GI equivalent of spherical aberration and image defocus (relative to a flat detector) due to a curved focal surface. For the Chandra X-ray Observatory at a field angle of 5 arcmin, the rms blur diameter ranges from 2.5 arcsec to 4 arcsec. Alternative Wolter I type designs exist that mitigate the off-axis aberrations at the expense of degrading on-axis performance. These designs consist of using two coaxial (but not confocal) hyperboloids or using generalized (non-conic) mirror prescriptions such as a power law. Because of the relative mechanical simplicity of the design and the ability to increase collecting area by nesting telescopes the type I design (or its variants) is the most common form of grazing incidence designs used in x-ray astronomy, being employed on the Einstein Observatory, Roentgen Satellite (ROSAT), CXO, X-ray Multi-mirror Mission (XMM) and many others described later.

The type II design also consists of a grazing incidence paraboloid primary and hyperboloid secondary mirror, but the outer surface of the hyperboloid is used, as shown in figure 4. In the type II the image is formed at the back hyperboloid focus. Type II telescopes enable longer focal lengths than a type I with comparable grazing angle and entrance aperture, affording an increased plate scale. Field-angle-dependent aberrations are greater with type II designs than with type I designs. Also, nesting of telescopes to increase collecting area is impractical. Instead, collecting area may be increased by increasing the grazing angle, thereby increasing the projected area on the aperture plane, while maintaining a usefully long focal length and acceptable plate scale. The increase in graze angle, however, limits the useful spectral bandwidth of the telescope so that most practical type II telescopes are used for the extreme to far ultraviolet and longer wavelengths (>100 Å).

The type III design employs the outer surface of a paraboloid as the primary element and the inner surface of an ellipsoid as the secondary mirror. The paraboloid focus is coincident with one of the ellipsoid foci, and the image falls upon the other ellipsoid focus. The type III design has never been used for x-ray astronomy.

Wolter developed variants of the original designs by extending results obtained by Karl SCHWARZSCHILD in 1905 for normal incidence telescopes. The modified prescriptions are called Wolter–Schwarzschild (W–S) designs and differ slightly in their second-order figure (shape). These optics satisfy the Abbe sine condition strictly and so do not exhibit any comatic aberration. The improvement in off-axis performance of W–S designs over conventional designs is graze angle dependent. Thus W–S designs provide negligible improvement at the shallow graze angles (<1°) employed on many x-ray telescopes to achieve good reflectance at shorter wavelengths. W–S-type designs have been used on a number of extreme ultraviolet telescope applications which use much larger graze angles (>5°). The EXTREME ULTRAVIOLET EXPLORER (EUVE) contained a W–S type I EUV telescope and a W–S type II for the stellar spectrometer. ROSAT, besides carrying the large Wolter I x-ray telescope, also carried a W–S type I EUV telescope—the Wide Field Camera—which had a 5° field of view, ∼450 cm^2 aperture and image half-power diameter of ∼1.7 arcmin (on axis).

Contamination of the optical surface will degrade performance. Particulates both absorb and scatter the x-rays, degrading the point spread function and reducing the effective collecting area. Hydrocarbons absorb x-rays. Grazing incidence telescopes are extremely sensitive to both. This is a direct result of the shallow grazing angle, as can be seen in figure 5. A round particle of cross section πa^2 maps onto an entrance aperture area of $2\pi a^2$ because x-rays may strike the particle before or after reflecting off the mirror surface. The mirror surface area is $2\pi RL$ where R is the average radius of the nearly conical optic. The mirror entrance aperture has an area of $2\pi RL\sin\alpha$. Thus the fraction of mirror surface area F_{n} covered by the particle is

$$F_{\text{n}} = \frac{\pi a^2}{2\pi RL} \qquad (7)$$

but the fraction of the entrance aperture F_{A} 'covered' by the particle is

$$F_{\text{A}} = \frac{2\pi a^2}{2\pi RL\sin\alpha} = F_{\text{n}}\frac{2}{\sin\alpha} \qquad (8)$$

For a 1° graze angle F_{A} is ∼114 times larger than F_{n}. Similarly for a hydrocarbon layer t thick, the x-rays

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

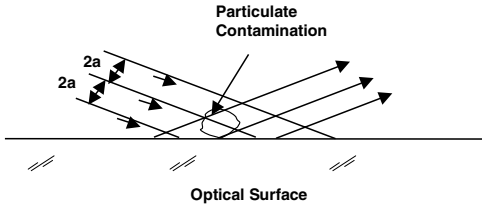


Figure 5. Illustration of the effects of particulate contamination. Incident x-rays from the upper left are absorbed or scattered by the particle either when directly impinging upon it or when impinging upon it after reflection from the mirror.

must traverse a distance $2t/\sin\alpha$. In addition, because of their construction with little space between telescope shells and very narrow entrance and exit annuli, it is extremely difficult to clean a dirty grazing incidence optic. Contamination control is a critical issue for grazing incidence optics.

Finally, an earlier realization than the Wolter designs was the Baez–Kirkpatrick telescope in which an array of parallel plates, parabolic in the plane of incidence, focus the incident x-rays to a line image. Placing a second set of plates oriented at a right-angle to the first set focuses the line image to a point image. This design has more aberrations than the Wolter designs, but may be approximated using optically flat plates bent in one axis so as to inexpensively build up large collecting area with moderate resolution.

Multilayer telescopes

Conventional optical telescope designs, when used as the substrate for multilayer coatings, are suitable for x-ray telescopes. The distinction between normal incidence x-ray and optical telescopes is that optical telescopes are diffraction limited and x-ray telescopes (to date) are not. To achieve diffraction-limited performance, the telescope wavefront error must be less than $\sim\lambda_x/13$, rms, where λ_x is the x-ray wavelength. For a two-mirror system being used at 171 Å, this equates to a surface error figure of ~5 Å, rms, or better than $\lambda/1300$ ($\lambda = 633$ nm), rms (neglecting other error contributors such as mirror figure degradation due to coating, support deformation, misalignments, etc). The ability to manufacture such optics is only starting to be feasible at the end of 1999. In addition, because of the short wavelength (relative to visible light) and current detector technology, extremely long focal lengths (>10 m) are required to make use of diffraction-limited performance.

Mechanical stresses in the multilayer coating can produce deformations of the optic, degrading imaging. Recent studies have shown that the stresses do not appear to vary as a function of the layer pair thickness, but do vary as a function of the equivalent d thickness and the materials. Thin-film technologists have found that coating stresses are also functions of a number of deposition parameters including substrate temperature and deposition process. Thin-film stresses may be estimated from Stoney's equation:

$$\sigma_f = \frac{E_s t_s^2}{6(1-\nu_s)t_f}\,\Delta C \tag{9}$$

where E_s, ν_s and t_s are, respectively, Young's modulus, the Poisson ratio and the thickness of the substrate, t_f is the thickness of the coating, ΔC is the change in radius of curvature of the substrate after the coating is applied and σ_f is the induced stress (a negative value of σ_f is compressive). Typical stresses range from −1200 to +200 MPa. These stresses can produce significant deformations on thin substrates, requiring either a modification of film or substrate design, or the use of post coating stress reduction techniques such as annealing (annealing, however, may cause diffusion across the interfaces, degrading reflectance).

Design, specification, and performance

Typical design constraints for an x-ray telescope will include focal length, spectral bandwidth, collecting area, resolution or fractional encircled energy, and size and weight. Not always independent of one another, these constraints determine whether a normal or grazing incidence telescope is desired, mirror figure requirements, coating design and element size.

Spectral bandwidth can determine whether an NI or a GI telescope is more appropriate. For extreme ultraviolet and very soft x-rays with wavelengths greater than ~100 Å, NI multilayer telescopes offer lower cost, better imaging and larger collecting area. Spectral regions above 0.1 keV are better served by GI optics which offer much greater reflection efficiency and better performance.

Diffraction by the annular entrance aperture can be significant at low energies. The intensity as a function of radial position x is expressed as:

$$I(x) = \frac{4}{[1-(r_i/r_o)^2]^2}\left[\frac{J_1(x_o)}{x_o} - \left(\frac{r_i}{r_o}\right)^2\frac{J_1(x_i)}{x_i}\right]^2$$

where

$$x_o = 2\pi r_o\frac{x}{\lambda f} \qquad \text{and} \qquad x_i = 2\pi r_i\frac{x}{\lambda f} \tag{10}$$

where r_o and r_i are the outer and inner radii of the annular aperture, respectively, f is the focal length, J_1 is the first-order Bessel function and λ is the incident wavelength. For NI telescopes diffraction does not limit imaging performance until one produces extraordinarily precisely figured optics, as discussed above in the section 'Multilayer telescopes'. Depending on telescope size, aperture diffraction can be more significant for GI optics. For example, the X-Ray Telescope (XRT) on the joint US–Japan Solar-B satellite has an aperture width of about 0.6 mm and an outer radius of ~170 mm. At a wavelength of 60 Å, the 68% point of the fractional encircled energy (the radial integral of the normalized point spread function)

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

is about 1.4 arcsec in diameter. For the large Chandra aperture diffraction is negligible.

Grazing angle, focal length and radius are related by the relation

$$\alpha = \frac{1}{4}\tan^{-1}\frac{r_0}{Z_0} \tag{11}$$

where α is the graze angle at the intersection of the paraboloid and hyperboloid surfaces, and r_0 is the radius of the virtual joint (Z_0 is not the focal length but it is fairly close). The choice of graze angle affects the short-wavelength limit of the telescope because of the precipitous loss of reflection efficiency near the critical angle. Therefore spectral bandwidth coupled with focal length can drive the maximum allowable optic size. From these few parameters one can derive the optical prescription of the paraboloid and hyperboloid using the method of Van Speybroeck and Chase (1972). If we constrain the graze angle of the hyperboloid at the virtual joint to be 3 times that of the paraboloid, then the prescription of the optics can be determined as

$$P = Z_o \tan(4\alpha)\tan\alpha \tag{12}$$

$$d = P \tag{13}$$

and

$$e = \cos(4\alpha)[1 + \tan(4\alpha)\tan(3\alpha)] \tag{14}$$

where the optical surfaces are represented as

$$r_p^2 = P^2 + 2PZ + 4e^2Pd/(e^2 - 1) \qquad \text{paraboloid} \tag{15}$$

and

$$r_h^2 = e^2(d + Z)^2 - Z^2 \qquad \text{hyperboloid} \tag{16}$$

and Z represents the coordinate along the optical axis and is zero at the system focus and e is the eccentricity. The effective focal length of the system, fL, is expressed as

$$fL = \frac{2e^2d}{e^2 - 1}. \tag{17}$$

(The ratio of fL/Z_o is a function of α and is approximately equal to $1 + 0.0025\alpha_{\text{deg}}^2$.)

Typically, performance of GI systems is expressed in terms of the encircled energy (usually as a fraction or as a per cent). This is the integral of the point spread function within a given angular diameter. One can think of the encircled energy as the fraction of incident flux that falls on a detector pixel or some other suitable area (e.g. a slit for some spectroscopy applications). It is a measure of how compact or spread out the light from a point source is imaged. The encircled energy may be normalized or scaled two different ways. In the first, normalization is with respect to the flux that leaves the last focusing optic (reflection losses are ignored). In the second approach the normalized encircled energy is scaled by the entrance aperture area and has units of area.

Image quality is affected by mirror surface imperfections (figure error and microroughness) which scatter the incident flux and broaden the point spread function. For conceptual purposes mirror errors may be loosely grouped into three categories. Low spatial error frequencies (long error periods) produce small-angle scatter which has only a small effect on the image 'core'. Mid-spatial-frequency errors produce intermediate-angle scattering and can have a significant effect on the image core, limiting resolution—the ability to discern two closely placed sources. High spatial error frequencies (typically microroughness) produce large-angle scatter (this might be on the order of 10–100 arcsec and greater) which degrades image 'contrast'—the ability to find a dim source in the presence of a bright source. The definition of what error frequencies correspond to the various bandwidths is dependent on the system requirements and grazing angle. On Chandra, errors that scatter x-rays by 0.5 arcsec would be considered mid-frequency. On XMM, mid-frequency errors might be considered as those that scatter by 5 arcsec.

A zeroth-order estimate of encircled energy can be made using the total integrated scatter, or TIS. This term represents the fraction of incident energy that is scattered by a surface (or equivalent surface) with a given rms surface figure error. The fractional encircled energy (or EE) is approximately $1 - \text{TIS}$. The EE is both incident wavelength and included angle dependent, so the choice of rms figure error must take this into account. This is done by using a bandlimited rms amplitude and making the assumption that this amplitude includes all surface error frequencies that will scatter the light outside the desired region of interest. The encircled energy is expressed (approximately) as

$$\text{EE} \approx \exp[-(2k\sigma\sin\alpha)^2] \tag{18}$$

where k is equal to $2\pi/\lambda$, σ is the bandwidth-limited effective rms surface error (in the plane of incidence) and α is the average graze angle. Using encircled energy goals we can estimate an acceptable value for σ, or, alternatively, σ can be used to estimate EE. (Of course, we want to leave some additional margin as we recognize that pointing stability–jitter, alignment, etc will all degrade performance.) To determine the bandwidth for which σ applies we use the grazing incidence equivalent of the grating equation (making use of the small-angle approximation)—

$$\lambda f = \theta_s \sin\alpha \tag{19}$$

where λ is the incident wavelength, f is the spatial frequency of the surface error and θ_s is the angle through which the radiation is scattered. All errors of higher spatial frequency will scatter through larger angles and fall outside the region of interest. In doing this analysis we are mostly concerned with errors in the axial direction. This is because the deviation of specular rays out of the plane of incidence by azimuthal errors is reduced by a factor of $\sin\alpha$. Similarly, the scattering distribution is elongated in the plane of incidence by a factor of $1/\sin\alpha$ (the grazing incidence foreshortens the spatial error periods in the plane of incidence, making the errors appear as a

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

higher-frequency and increasing the scattering angle (see equation (19))).

A more exact method of computing telescope requirements and estimating performance is obtained using the scalar scattering theory of Beckmann and Spizzichino. (Vector scattering theory may also be used for modeling performance. Some references by E Church are listed in the bibliography.) The output intensity distribution is given as

$$\frac{\mathrm{d}P}{\mathrm{d}\Omega_{\mathrm{total}}} = \mathrm{Strehl}\frac{\mathrm{d}P}{\mathrm{d}\Omega_{\mathrm{spec}}} + \frac{\mathrm{d}P}{\mathrm{d}\Omega_{\mathrm{spec}}} \otimes \frac{\mathrm{d}P}{\mathrm{d}\Omega_{\mathrm{scatt}}}. \qquad (20)$$

Strehl is the Strehl ratio and equals the right-hand side of equation (18). $\mathrm{d}P/\mathrm{d}\Omega_{\mathrm{spec}}$ is the intensity distribution in the absence of scattering–diffraction, such as results from large-scale geometric figure and alignment errors, $\mathrm{d}P/\mathrm{d}\Omega_{\mathrm{scatt}}$ is the scattered intensity distribution and $\otimes$ signifies a convolution operation. Using scattering theory to determine $\mathrm{d}P/\mathrm{d}\Omega_{\mathrm{scatt}}$ and integrating over $\mathrm{d}\Omega$ yields the encircled energy. After much algebra this can be reduced to a one-dimensional integral where the scattering is a function of the Fourier transform of the optic axial figure error power spectrum density (PSD). Thus, by measuring the axial surface errors (i.e. the axial one-dimensional profiles), computing the error PSD and measuring the geometric in- and out-of-plane contributors, we can estimate performance. Initially appearing complex, this formalism is very useful because optic (and replication mandrel) fabrication processes are essentially surface filters. Imaging performance is a function of surface error PSD, and optical fabrication technology provides a means to operate on the surface error PSD. This approach provides much more fabrication guidance than merely specifying a single bandlimited rms amplitude as an optic requirement. This approach also correctly takes into account the impact of the frequency content of the figure errors in estimating performance.

Fabrication

Grazing incidence telescopes are produced by one of several methods: direct polishing of the optics, replication of polished mandrels or forming of thin foils. Multilayer telescopes are manufactured using standard precision optical fabrication methods but require the application of the multilayer coatings.

Material considerations for x-ray optics are important. Since all x-ray observations are made in space, the optical elements must be strong enough to survive rocket launch acoustic and seismic loads. At the same time, optic weight must be minimized to reduce payload weight. Depending on the complexity of the spacecraft an active temperature control system may or may not be present, so mirror elements should have both a low coefficient of thermal expansion (CTE) and a uniform one. This minimizes optic distortion due to temperature variations and gradients in the spacecraft as it orbits the Earth.

Typical glasses used for x-ray telescopes are Zerodur™, a glass ceramic produced by Schott with an extremely low CTE, and ULE™ (which stands for ultralow expansion) produced by the Corning Glass Co, another very low-CTE glass. Fused quartz was used for the Einstein Observatory mirrors. Potential new materials include silicon carbide which has an extremely high strength to weight ratio offering the promise of very light weight non-replicated optics.

Optical fabrication of grazing incidence optics

Optic blanks arrive from the glass supplier as rough machined pieces, within 250–1000 μm of final dimensions (by contrast, final allowable figure error might be only a few hundred Ångströms, rms, or 10^4–10^5 times better). A carefully scripted material removal schedule utilizing ever-finer grinding grits and polishing compound is used to coarse figure the optics and remove residual machining stresses (subsurface damage, or microcracks) that can degrade figure stability or even lead to catastrophic failure under load. The inside (optical surface), outside and ends of the blank are polished for damage removal, to provide a surface that is easily cleaned and kept free of contamination, and to provide a controlled bonding surface for the optic mounting system.

The optical surface is figured in (grinding and) polishing using computer-controlled fabrication technology where a computer provides a set of commands to move the polishing head in a controlled fashion over the surface of the optic. Typically, the optic is supported with its axis nearly horizontal, the optic is made to rotate about its axis at a controlled rate, and the polishing tool is driven axially along the optical surface to describe a fixed path (e.g. a spiral), but with variable path velocity. Since material removal is inversely proportional to the path velocity, control of that velocity allows correction of the optical figure. The velocity commands result from deconvolving the polishing tool material removal profile from the optic surface error map. The fabrication process, along with the associated measurements, is iterative with typical error correction rates of 50–90% per iteration depending on the optics manufacturer. For Chandra, the correction rates were typically 85–95% depending on the error spatial frequency content.

Smoothing of the optic, critical for controlling scatter, is usually performed as a separate operation from figuring. ROSAT optics were smoothed to surface roughness levels of about 3 Å, rms, and the Chandra optics were smoothed to about 2–3 Å, rms.

The combination of computer-controlled fabrication and smoothing function as spatial frequency filter operations, making them readily linked to the PSD requirements that can be generated from the encircled energy requirement.

Because of the unusual optic geometry, most fabrication and metrology equipment is custom designed by the manufacturer. For larger optics such as on ROSAT and Chandra specialized handling equipment is also required.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Fabrication of grazing incidence optics by replication
When extremely good imaging (less than a few arcsec diameter) is not required, or when extremely large collecting area is required, replicated optics can be a superior alternative to individually fabricated glass optics. In this approach a set of replication mandrels that are the inverse of the desired final figure are fabricated first. Multiple replicas, each the inverse of a mandrel and therefore possessing the desired mirror figure, are then formed.

The advantages of replication are several. First, when multiple copies of the same mirror are required, replication is cheaper and less time consuming than individually fabricating each mirror. Second, many more shells can be nested within one another (without unduly growing the size of the outermost elements) because the replicas are typically much thinner than glass elements. This results in higher telescope 'throughput'. Third, because of their relative 'thinness' replicated optics may weigh substantially less than corresponding glass elements, reducing payload weight. (As an example, the largest X-ray Multi-mirror Mission (XMM) replicated nickel optics are less than 1/4 the weight per unit axial length of the corresponding size Zerodur™ Chandra counterparts.) These advantages mean that either the instrument can carry multiple copies of the telescope optics and detector (such as XMM) or multiple identical satellites can be placed in orbit (such as planned for Constellation-X). Replication is a way to achieve much larger collecting area and higher telescope throughput with lower weight per unit area at less cost than with non-replicated elements.

The disadvantage of replication (to date) is the level of figure quality which may be achieved, limiting imaging performance to ~14 arcsec, half power diameter (or 50% encircled energy). The mandrels themselves may be fabricated with the same accuracy as large glass elements. The difficulty arises in attempting to remove the lightweight replica from the mandrel, maintain its figure against internal stresses created while producing the replica and then support it for flight, without introducing low spatial frequency distortions in the figure.

The strengths of replicated optics and individually figured optics telescopes such as XMM and CXO are complementary. Replicated telescopes, with their large collecting area and good imaging capabilities, are generally more useful for imaging spectroscopy. For bright objects, however, telescope throughput is not a limiting factor and the full imaging capabilities of glass optics can be brought to bear. Replicated optics are more suitable for some observations of very faint objects or deep sky surveys where the photon noise limit obviates finer resolution, but the limiting angular resolution may also result in an inability to resolve discrete sources. Alternatively, glass optic telescopes may be limited in the number of deep surveys possible owing to the extended observing time needed to collect enough photons necessary to make full use of their imaging.

Several methods of replication are briefly described below.

Nickel replicas An aluminum mandrel is coated with ~200 μm of electro-less nickel (Kanigen). The electro-less nickel surface is loose abrasive ground and computer-controlled polished or diamond turned to the nominal figure (after diamond turning additional figuring may be required). The surface is smoothed (polished) to about 5 Å, rms, roughness, to complete the mandrel. The replica is produced by first depositing 100–200 nm of gold on the electro-less nickel surface and then electroplating the gold with nickel to the desired thickness (~1 mm). The replica is separated from the mandrel by cooling the mandrel (the coefficient of thermal expansion for aluminum is about twice that of nickel). The gold coating separates with the nickel replica because its adhesion to the electroplated nickel is much greater than to the Kanigen. Multiple replicas may be made from a single mandrel before the mandrel needs to be refurbished. Replicas of this type have been used for the BEPPOSAX, JET-X and XMM. Replica sizes range from 300 mm long, 68 mm diameter on BeppoSAX to 600 mm long, 700 mm diameter on XMM. The half-energy width of the point spread function at 1.5 keV is 13 arcsec for XMM.

As previously mentioned, internal stresses in the electroplated nickel will deform the replica when it is removed from the mandrel. Potential solutions to the lack of structural rigidity of these replicas include the use of stiffening structures fixed to the replica during the nickel plating process and the use of ceramics or silicon carbide (SiC) in place of the nickel replica substrate. In the latter approach the SiC substrate is formed to near final shape by chemical vapor deposition (CVD) to a second mandrel slightly larger (10^2 μm) than the first. The first mandrel is coated with 100 nm of gold, and then the SiC substrate is positioned around it. The gap is filled with epoxy, and, after curing, the gold-coated epoxy/SiC replica is separated from the Kanigen-coated aluminum mandrel by cooling. The CVD SiC has little residual internal stress and has a Young's modulus approximately 3 times larger than that of nickel. The SiC replica can be made substantially thinner than the nickel replica. A similar process using a beryllium substrate instead of SiC has been used on the Exosat program. An alternative approach in which aluminum oxide is plasma sprayed directly on the gold-coated Kanigen/aluminum mandrel is also under investigation at the time this article is being written.

Epoxy replicas In epoxy replication an aluminum foil serves as the replica substrate. A glass (Pyrex™ or Zerodur™) mandrel is coated with ~100 nm of gold. A thin (~100 μm) aluminum foil is pre-formed to the approximate shape of the coated mandrel but slightly oversized. Both the gold-coated mandrel and the (inner) surface of the foil are sprayed with epoxy and the two are then mated. The epoxy film, several tens of microns thick, is cured in an oven before the foil is removed. The foil maintains the smoothness of the polished glass mandrel. Epoxy replicas have not at this time achieved the imaging capabilities of nickel replicas, but the replicas

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Table 1. An abbreviated list of some of the more notable x-ray telescopes.

Mission	Type	Effective area (cm^2)	Resolution (HPD, arcsec)	Energy bandwidth[a] (keV)	Comments
Skylab (1975)	GI	42 (collecting area)	~2 arcsec resolution	0.2–2	First x-ray telescope; solar observations
Einstein Observatory (1978–81)	GI	~200 at 1 keV	~15	0.2–4.5	First telescope observatory; discovered 7000+ sources
RXRS[b] (1987)	NI	~50 (collecting area)	~2 arcsec resolution	17.3, 25.6 nm	First NI solar telescope(s); 6.3 cm diameter primary mirror
ROSAT (1990–9)	GI	400 at 1 keV	~5	0.1–2.4	4 Au coated Zerodur shells; discovered 150 000+ sources
ASCA (1993)	GI	1300 at 1 keV, 600 at 7 keV	174	0.5–10	Conical foil Al mirrors, Au coat over lacquer, 4 separate telescopes
BeppoSAX (1996)	GI	330 at 1 keV	60	0.1–10	Nickel-replicated conical optics, 30 nested shells
TRACE (1998)	NI	~20	1	17, 20, 28 nm	30 cm diameter primary, 2 m focal length solar telescope
Chandra (1999)	GI	800 at 1 keV	0.5	0.1–10	Highest resolution, 4 shells, largest mirror 1.2 m diameter transmission gratings
XMM (1999)	GI	4650 at 1 keV, 1800 at 8 keV	14	0.1–12	Nickel replicas, 3 telescopes, 58 shells each, reflection gratings
Constellation-X (200?)	GI	15 000 at 1 keV 1500 at >6 keV	15 (<10 keV) 60 (>25 keV)	0.25–10, 6–40	Replicated optics (type to be determined), ~80+ shells, largest 1.2 m diameter grating

[a] For NI telescopes the tabulated values represent the approximate centers of the reflection bandpass(es).
[b] Stanford University/NASA Marshall Space Flight Center Rocket X-Ray Spectroheliograph.

are substantially lighter owing to their much thinner wall and less dense substrate material. To date limiting performance is ~1 arcmin.

Segmented foil replicas Segmented foil mirrors have been used on the Broad Band X-Ray Telescope (BBXRT), ASCA and the Sodart telescope on board SPECTRUM-X-GAMMA. Here the mirrors do not form the complete surface of revolution, but instead form (typically) only a single quadrant. Aluminum foils less than a millimeter thick and optically smooth are rolled into a 90° section of the desired conical shape (note that the mirrors, being true cones, are only approximations to Wolter I telescopes). Plastic deformation (rather than elastic bending) is employed to produce a final shape that more closely approaches the nominal and does not introduce stresses into the support structure. After rolling, the foils are dip-coated with an acrylic lacquer to improve (reduce) scattering due to surface roughness. Care is taken to avoid introducing variations in lacquer thickness that would change the optical shape of the mirror, degrading imaging. Mirrors of this type are very inexpensive to produce but have not achieved better than a few arcminutes resolution. In part this limitation is a result of the approximation to a Wolter I system, but, more importantly, it appears that the lacquer is ineffective at smoothing errors with spatial periods greater than a few microns. These errors significantly affect the ability to image at better than the 1 arcmin level.

Fabrication of normal incidence multi-layer telescopes Normal incidence x-ray multilayer telescopes are essentially manufactured using the same processes as conventional precision optics with two major exceptions: optic smoothness (or microroughness), which affects both scatter and multilayer reflectivity, and the deposition of the x-ray multilayer coating.

As discussed above in the section 'Reflection from multilayer coatings' reflectance is a function of high spatial frequency roughness ($f > 10^3$ mm^{-1}). Super-smoothing of the surface roughness in this very high spatial frequency regime to levels of 2–4 Å, rms, is necessary for good reflectance.

Along with substrate roughness, the multilayer coating is the most critical element of the optics. To achieve the desired reflectance over the correct spectral bandwidth requires uniformity of coating thickness both across the face of the optic as well as from layer to layer. Coating

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

density must also be uniform throughout the multilayer. In addition, interlayer diffusion must be minimized as must thin-film stresses.

Telescope systems

Other parts of a telescope system are the detector, optical bench and mirror mount or support structure. CHARGE COUPLED DEVICES (CCDs) have supplanted position-sensitive proportional counters. CCDs may be either front or back illuminated. Front-illuminated CCDs have the x-rays impinge upon the semiconductor gate side. Back-illuminated CCDs have additional wafer processing to thin the backside of the chip (the side away from the gates) and are then set so that the x-rays impinge upon the backside. Front-illuminated CCDs provide better energy resolution and slightly higher quantum efficiency at high energy than back-illuminated CCDS. Back-illuminated CCDs provide substantially higher low-energy x-ray quantum efficiency. A wide range of pixel sizes are available: the Chandra CCDs have 24 μm pixels; the Solar X-ray Imager (SXI) on GOES N and O will have $\sim$15 μm sized pixels. The long focal length ($\sim$10 m) of Chandra coupled with 24 μm pixels yields $\sim$0.5 arcsec angular resolution. CCDs provide excellent energy resolution, about 120 eV for the Chandra detectors. Several CCDs may be arrayed to cover the full field of view of the telescope, and the detectors may also be arrayed so as to be aligned to the curved focal surface, thereby eliminating focus error that occurs with a flat detector and a curved focal surface (see equation (6)). Higher imaging resolution is achieved with less energy resolution using detectors such as the Einstein, ROSAT and Chandra High Resolution Imager. This is a pair of stacked multichannel plates with an electronic readout. Future developments include the use of microcalorimeters, which will be used for the first time on Astro-E. This device has an energy resolution of $\sim$12 eV at 6 keV but has a relatively large pixel size and so limits spatial resolution. Spectroscopy with x-ray telescopes is also accomplished by the use of transmission gratings that may be moved in and out of the focused beam, such as on Chandra, or reflection gratings fixed in the beam as on XMM.

X-ray telescope missions

Some significant x-ray telescopes and performance details are listed in table 1.

Future trends

In grazing incidence telescopes future trends will continue the development of increased collecting area, improving resolution and increasing reflectance. The use of ceramics and SiC, including the use of monolithic support structures, will improve the limiting performance of what used to be nickel mandrels, while also allowing denser nesting of mirror shells. Researchers have been experimenting with the application of broadband multilayer coatings to grazing incidence optics, increasing reflectance at higher energies. With respect to normal incidence telescopes future development will probably focus on achieving diffraction limited performance with increased focal lengths to more fully utilize the improved optics.

Bibliography

There are a multitude of excellent articles in the literature discussing the technology of x-ray telescopes. For design issues of GI telescopes one is referred to

Van Speybroeck and Chase 1972 Design parameters of paraboloid–hyperboloid telescopes for x-ray astronomy *Appl. Opt.* **11** 440

An excellent review is found in

Aschenbach B 1985 X-ray telescopes *Rep. Prog. Phys.* **48** 579

Discussions of alternative Wolter-I-like designs with improved off-axis performance can be found in

Nariai K 1987 Geometrical aberration of a generalized Wolter type I telescope *Appl. Opt.* **26** 4428
Werner W 1977 Imaging properties of Wolter I type x-ray telescopes *Appl. Opt.* **16** 764

Atomic scattering factors for computing reflectance are found in

Henke, Gullickson and Davis 1993 *At. Data Nucl. Data Tables* **54** (2)

This information is also available on the Wide World Web at URL

http://xray.uu.se/hypertext/henke.html.

A second site,

http://www-cxro.lbl.gov:80/optical_ constants

contains the same data along with a reflectance 'calculator'. A general review of multilayer coatings can be found in

Barbee 1990 Advances in multilayer x-ray/EUV optics: synthesis, performance, and instrumentation *Opt. Eng.* **29** 711

Discussions of vector and scalar scattering theory are reviewed by

Church E 1979 Role of surface topography in x-ray scattering *SPIE Proc.* **184** 196
Church E 1986 The interpretation of glancing incidence scattering measurements *SPIE Proc.* **640** 126

For more detailed discussions of developments the reader is best referred to the Society of Photo-Optical Instrumentation Engineers (SPIE) annual proceedings for grazing incidence and multilayer x-ray optics. Some recent volumes are *SPIE Proceedings* volumes 3444 (published in 1998), 3113 (1997), 2808 and 2805 (1996) and 2515 (1995).

Paul B Reid

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

X-rays

Energetic electromagnetic radiation with wavelengths in the range 10 nm (10^{-8} m) down to 0.01 nm (10^{-11} m). Although the boundaries are somewhat arbitrary, wavelengths shorter than 0.01 nm are called gamma-rays and those longer than 10 nm extreme ultraviolet (EUV). Cosmic x-rays are usually described in terms of photon energies, a wavelength of 10 nm corresponding to an energy of 80 electron volts (eV) and a wavelength of 0.01 nm to 8×10^5 eV (80 keV).

X-rays from cosmic sources are absorbed high in the Earth's atmosphere. Even the most energetic (shortest-wavelength) x-ray photons fail to penetrate much closer to the ground than an altitude of about 40 km.

X-rays were discovered in 1895 by the German physicist Wilhelm Konrad Röntgen (1846–1923).

See also: electromagnetic radiation, electromagnetic spectrum, electronvolt, x-ray astronomy.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Yarkovsky Effect

The Yarkovsky effect changes the rotation and the orbit of a body orbiting in the Solar System by the asymmetrical re-radiation of thermal energy from the Sun. The effect constitutes a non-gravitational force which causes the orbits of smaller, kilometer-sized asteroids to change over time.

Copyright © Nature Publishing Group 2002
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2002
Dirac House, Temple Back, Bristol, BS21 6BE, UK

Year

The period of the Earth's revolution around the Sun, or of the apparent motion of the Sun on the ecliptic. It may be defined in a number of ways, each of which leads to a slightly different value:

Sidereal year. The time interval during which the Sun apparently completes one revolution of the celestial sphere relative to the stars (which, for this purpose, are regarded as being fixed in space). This is equal to the revolution period of the Earth around the Sun as measured relative to the stars, and is equivalent to 365.2564 mean solar days.

Tropical year. The time interval between two successive passages of the Sun through the vernal equinox. Its length is 365.2422 mean solar days, about 20 minutes shorter than the sidereal year. The difference arises because of the effects of precession. As this definition of the year is related to the recurrence of the seasons, the term 'year', if unqualified, is generally taken to mean 'tropical year'.

Anomalistic year. The interval between two successive passages of the Earth through the perihelion of its orbit which, because of a slow change in the position of perihelion, is not quite the same as the sidereal year. Its length is 365.2596 mean solar days.

Gregorian calendar year. This is the value of the year adopted for calendar purposes, and is equal to 365.2425 mean solar days. For practical purposes it can be taken as equal to the tropical year (the difference amounts to 0.0003 mean solar days).

See also: calendar.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Yerkes Observatory

Yerkes Observatory lies 334 m above sea level in Williams Bay, Wisconsin. It is a research branch of the Department of Astronomy and Astrophysics of the University of Chicago.

The observatory was completed in 1897. It was financed by Charles Tyson Yerkes, a Chicago transportation tycoon, but the inspiration behind its construction was George Ellery Hale. The showpiece of the observatory was the 1 m refractor, the world's largest telescope in 1897, and still the largest refracting telescope ever built.

Until the mid-1960s, Yerkes Observatory housed all of the university department's activities (including managing the operations for MCDONALD OBSERVATORY in Texas from 1932–62). Today, the 77 acre site provides laboratory space and access to telescopes for research and instruction. A substantial fraction of the university's library holdings in astronomy are housed at Yerkes.

The principal telescopes are the 1 m refractor, a 1 m Ritchey–Chrétien reflector which is used for adaptive optics studies, a 0.6 m reflector and an 18 cm Schmidt camera for wide-field photography.

Recent research at Yerkes includes measuring the velocities and distances of the furthest star clusters within the Milky Way to better determine the mass of our Galaxy; spectroscopic measurements of lithium abundances; spectra of the dust disk around Beta Pictoris; and studies of the properties of distant galaxies. Yerkes astronomers are currently developing a camera for the airborne Stratospheric Observatory for Infrared Astronomy (SOFIA).

For further information see
http://www.astro.uchicago.edu/yerkes/.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Yohkoh

The Yohkoh satellite was launched by Japan's INSTITUTE OF SPACE AND ASTRONAUTICAL SCIENCE (ISAS) on 30 August 1991 for observations of solar FLARES and the solar CORONA in x-ray and gamma-ray wavelengths. Excellent observations have been made over the last nine years (see figures 1 and 2). The satellite has experienced the entire transition from the maximum to the minimum of the last SOLAR CYCLE and is proceeding to observations of the rising phase of solar cycle for the first time.

Yohkoh carries two X-RAY TELESCOPES: the hard x-ray telescope has imaging capability above 40 keV because of its utilization of modulation collimators. These hard x-ray images allow us to obtain new information on where and how energetic electrons are accelerated in solar flares. The soft x-ray telescope is a grazing-incidence x-ray (5–50 Å) telescope to observe the solar corona and solar flares with high spatial (3 arcsec) resolution. The soft x-ray telescope is sensitive to plasmas with temperatures from 2 MK through 20 MK. Two other instruments onboard are the Bragg crystal spectrometer to observe the iron, calcium and sulfur lines from flare plasmas and the wide-band spectrometers to observe flare spectra from 5 keV to 10 MeV.

The discoveries made so far by Yohkoh cover a wide area of solar physics. In the following, we give representative examples.

(1) *Mechanism of solar flares*. The Yohkoh data, for the first time, showed unambiguously that magnetic reconnection is responsible for significant energy release in solar flares; intense heating occurs owing to magnetic reconnection possibly with standing slow magnetohydrodynamic shocks.

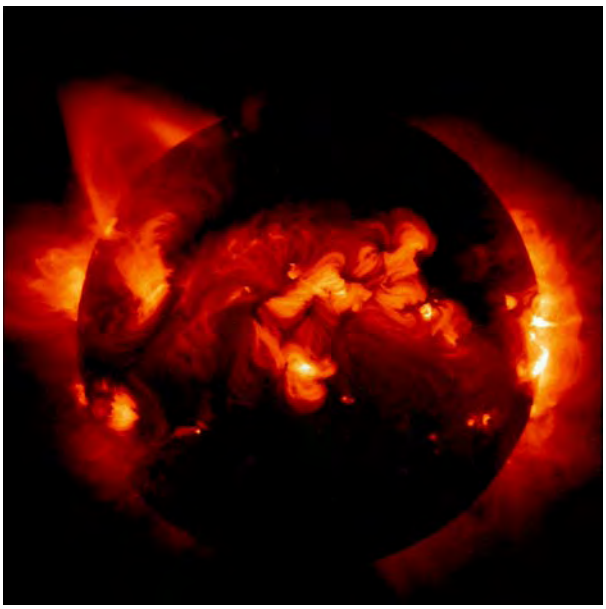


Figure 1. X-ray image of the Sun taken with Yohkoh.

(2) *Microflares and coronal heating*. Yohkoh x-ray movies showed numerous small transient brightenings far smaller than flares in energy (microflares) as well as structural changes of coronal magnetic fields. Magnetic reconnection is again responsible for these brightenings. Although the heating mechanism of the persistent corona remains unknown, these observations imply that an ensemble of nanoflares, which are too faint to be resolved, maintains the persistent solar corona.

(3) *X-ray jets*. Yohkoh discovered x-ray jets with speed reaching several hundreds km s^{-1}. X-ray jets may be driven by the slingshot effect of the reconnected field lines or maybe due to chromospheric evaporation caused by heat flow from the reconnection site.

(4) *Electron acceleration in solar flares*. The Yohkoh hard x-ray telescope showed that hard x-ray flares essentially have double-source structures, which are located at the footpoints of the soft x-ray loop. This clearly indicates that a rich population of non-thermal electrons (up to 10 MeV or higher) are created in flare loops in association with magnetic reconnection and emit hard x-rays with bremsstrahlung. A surprising discovery from the hard x-ray observations is the detection of an impulsive hard x-ray source located above the soft x-ray loop. The loop-top source is due to bremsstrahlung of high-energy electrons with energy up to at least 50 keV. The energization probably occurs at the fast magnetohydrodynamic shock formed by the supersonic outflow from the reconnection site.

MAGNETIC RECONNECTION with associated slow magnetohydrodynamic shocks appears to be a powerful engine to convert magnetic energy to plasma kinetic and thermal energies. The hard x-ray observations also show that the resultant supersonic outflow creates fast shock, which converts the flow energy to further heating and/or non-thermal acceleration of electrons and protons as a result of Fermi acceleration. In spite of the large magnetic Reynolds number of the solar corona,Yohkoh has delivered unambiguous evidence of magnetic reconnection in flares and in the corona.

Yohkoh is therefore revolutionizing our understanding of the solar corona and the behavior of magnetized plasmas in general. Yohkoh shows for the first time how the Sun dynamically relaxes its magnetic energy, buoyant as a result of the subsurface dynamo mechanism, by the process of magnetic reconnection. The formation of the solar corona with frequent sporadic (10–100 Alfvén transit times) energy releases including solar flares, solar wind and coronal mass ejections is a consequence and manifestation of the energy release process. This suggests that magnetic energy conversion through magnetic reconnection is likely to be a common occurrence in the cosmos.

Bibliography

For details on the instrumentation on board Yohkoh, the interested reader is referred to volume 136 of *Solar Physics* (1991).

Saku Tsuneta

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

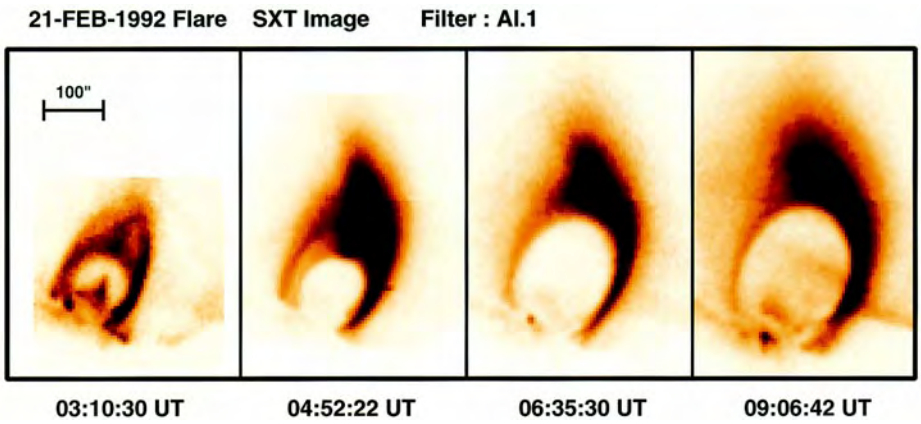


Figure 2. Gigantic flare arch observed with Yohkoh.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Young Stellar Objects

Young stellar objects (YSOs) are stars in the earliest stages of development. There are two principal kinds of YSOs: protostars and pre-main sequence stars. YSOs are intimately associated with fundamental by-products of the star formation process such as bipolar outflows, circumstellar jets, masers, Herbig–Haro objects, and circumstellar (protoplanetary) disks. YSOs are always found within or near interstellar clouds of gas and dust. Most YSOs populate giant molecular clouds and a significant number are members of fully or partially embedded stellar or protostellar clusters.

Background

Stars are the fundamental objects of the astronomical universe. They are self-gravitating balls of (mostly) hydrogen gas which convert the primary product of the big bang (hydrogen) into heavier elements. For most of its life a star maintains a state of stable equilibrium in which the inward force of gravity is balanced by the outward force of pressure. This internal pressure is generated by the energy released in the nuclear reactions which burn hydrogen at the star's center. These nuclear reactions are also the source of the star's luminosity. During the time the star burns hydrogen in its core it maintains a fixed radius and luminosity and consequently a constant surface temperature. The exact value of the equilibrium radius, luminosity and surface temperature of a star depends almost exclusively on one parameter, the star's mass (see STELLAR MASSES). The more massive the star, the greater its luminosity, size and surface temperature. Together, hydrogen burning stars of varying mass form a well defined locus of points in the observable luminosity–effective temperature plane (the HERTZSPRUNG–RUSSELL DIAGRAM). This locus of points is called the main sequence and the hydrogen burning phase of a star's life is known as the main sequence phase. Main sequence stars range in mass from about 0.08–100 $M_\odot$ (solar masses). Stars with smaller masses have insufficient weight to raise their central temperatures enough to enable hydrogen fusion. (Such objects are referred to as brown dwarfs.) Stars with larger masses are presumably too luminous to hold on to their outer atmospheres.

The formation and early evolution of stars

STAR FORMATION and early STELLAR EVOLUTION occur prior to the main sequence or hydrogen burning phase of stellar life. Most stars in the Galaxy originate in giant molecular clouds (GMCs). With temperatures seldom in excess of 10 K and maximum dimensions of 50–100 pc, GMCs are the coldest objects in the universe and the largest objects in the Galaxy. With masses in excess of $10^5\ M_\odot$ they also rival globular clusters as the most massive objects in the Galaxy. GMCs are composed primarily of molecular hydrogen and are characterized by mean densities of $\sim$100 cm^{-3}. About 10% of the mass of a typical GMC is in the form of dense cores with $n(H_2) \sim 10^4$ cm^{-3}. It is within such dense regions that stars are formed.

Although there is no complete theory of stellar origins, basic astrophysical considerations result in the following general picture of star formation and early stellar evolution. Stars form through the gravitational collapse of dense molecular cloud cores. Before being incorporated into a star, the interstellar material must increase its density by 20 orders of magnitude and collapse to a size nearly 7 orders of magnitude smaller than the dimensions of the original dense cloud core. Because the dust in the cloud is optically thin, internal energy gained by its collapse is effectively radiated away. Thus the cloud material remains isothermal and dynamically collapses. The collapse also proceeds in a non-homologous manner, with the inner regions becoming denser and collapsing faster than the outer regions which are left behind. Eventually, the innermost infalling material becomes dense enough to be optically thick to its own radiation resulting in the development of a central quasistatic stellar core surrounded by an infalling envelope. Thus is a protostar born.

After the formation of its embryonic core, the protostar enters the accretion phase of protostellar evolution. During this time the central stellar core gradually gains mass via the accretion of material from its infalling envelope. Before being finally incorporated onto the growing stellar core, accreting material must dissipate the gravitational potential energy lost in infall, giving rise to an accretion luminosity:

$$L_{\text{accretion}} = \frac{GM_*\dot{M}}{R_*}$$

where M_*, and R_* are the mass and radius of the protostellar core and $\dot{M}$, the mass accretion rate. This accretion luminosity can be a significant component of a protostar's luminosity. The mass accretion rate is generally expressed as:

$$\dot{M} = m_0\frac{a^3}{G}$$

where a is the effective sound speed and m_0 is a constant which is sensitive to the initial conditions in the collapsing core. Different theories of star formation predict somewhat different values of m_0. In some models m_0 varies with time and is quite large ($\sim$10) in the early stages of collapse, falling to near unity in the later stages. However, for most protostars, this constant must be near unity to be consistent with the observed luminosities of protostellar objects.

Initially the mass of the embryonic stellar core is very small, approximately $10^{-2}\ M_\odot$ and the central temperature of the core is not sufficient to ignite nuclear reactions. As the embryonic core begins to grow its luminosity is dominated by accretion. The mass of the protostellar core increases as $M_*(t) = \dot{M}t$. Once a protostellar core reaches a mass of roughly 0.2–0.3 $M_\odot$, its central temperature reaches

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

10^6 K and deuterium-burning nuclear reactions are ignited providing the protostar with another source of luminosity. ACCRETION of new material and a fully convective core enable the protostar to continue burning deuterium as it grows. If placed on the HR diagram the protostellar core would follow a trajectory parallel to and somewhat above the deuterium-burning main sequence which itself lies significantly above and roughly parallel to the (hydrogen-burning) main sequence. A protostellar core can continue to gain mass by accretion until its central temperature reaches 10^7 K and hydrogen fusion is ignited, at which point the embryonic core reaches the main sequence. This occurs when the cores attain masses around 7–8 solar masses. Presumably, the protostar can still continue to grow and increase its mass. However, for reasons not yet fully understood, most protostars cease growing long before this point. In any event, stars which end up with masses in excess of 7–8 solar masses have a very different post-protostellar evolution than stars which end up less massive.

The physical reason for this becomes apparent if one compares the timescale for pre-main sequence evolution with that of protostellar collapse. The timescale for the gravitational collapse of a cloud core, the free-fall time, is determined largely by ρ, the density of the cloud:

$$\tau_{\rm ff} = \sqrt{\frac{3\pi}{32G\rho}}.$$

For the typical mean density ($n \approx 10^4$ cm^{-3}) of a cloud core (of either low or high mass) the free–fall time is about 4×10^5 yr. The timescale for pre-hydrogen-burning evolution is the Kelvin–Helmholtz time:

$$\tau_{\rm KH} \approx \frac{GM_*^2}{R_* L_*}$$

which is very rapid for a high mass star (i.e. $\approx 10^4$ yr for $M_* = 50\ M_\odot$) and relatively slow for a low mass star (i.e. $\approx 3 \times 10^7$ yr for $M_* = 1\ M_\odot$). More importantly for high mass stars $\tau_{\rm KH} < \tau_{\rm ff}$ and these stars begin burning hydrogen and reach the main sequence before the termination of the infall or collapse phase of protostellar evolution. On the other hand, for low mass stars $\tau_{\rm KH} > \tau_{\rm ff}$ and low mass stars have an *observable* pre-main sequence stage of stellar evolution.

The initial conditions for pre-main sequence evolution are those which describe the mass, radius and luminosity of an embryonic protostellar core at the the point in time that infall and accretion cease. For this reason the locus of points on the HR diagram which traces the initial starting points of PMS evolution for all stars is called the birthline. The position of a given star on the birthline is a function of the mass it has acquired by the end of its protostellar accretion phase. Presumably, the protostellar evolution of a given star is identical to that of all other stars until the time that the star stops accreting and reaches the birthline. Prior to its appearance near the birthline, a young stellar object is surrounded by an infalling protostellar envelope which renders it invisible. Once the envelope is either mostly accreted onto the protostellar core or removed by some agent, the protostar becomes a visible PRE-MAIN SEQUENCE STAR. Its initial luminosity and surface temperature place it near the birthline which, for low mass stars, is itself basically coincident with the deuterium-burning main sequence on the HR diagram. Because the abundance of deuterium in the young star is relatively low, it is burned up very rapidly. Without accretion to replenish the burned deuterium, nuclear reactions cease and the star slowly contracts to the main sequence. The timescale for this quasistatic contraction to the main sequence is the Kelvin–Helmholtz time.

Observational characteristics of the low mass young stellar objects

The evolutionary status of a normal star is usually determined by its placement on the HR diagram. This in turn requires measurements of two quantities: stellar luminosity and effective temperature. A star can be meaningfully placed on the HR diagram provided it emits a black-body-like spectrum that can be characterized by a single effective temperature. For YSOs this is not always possible. This is because throughout their formation and early evolution stars are intimately associated to varying degrees with natal gas and dust. This circumstellar gas and dust can absorb and reprocess substantial amounts of the radiation emitted by a young stellar object, significantly altering its spectral appearance. The circumstellar gas and dust associated with a young stellar object has a spatial extent considerably greater than that of its stellar photosphere. Consequently, emitting circumstellar dust, which is in radiative equilibrium with the stellar radiation field of the buried star, will exhibit a wide range of (effective) temperature and the emission that emerges will have a spectral distribution much wider than that of a single temperature black body. In addition, at the optical wavelengths the youngest objects are rendered completely invisible by the obscuration of opaque circumstellar dust and a significant fraction, if not all, of their luminous energy is radiated in the infrared portion of the spectrum.

To determine the evolutionary status of a young stellar object requires knowledge of its overall broadband energy distribution, particularly at infrared wavelengths. The shape of the broadband energy distribution of a YSO will depend on both the nature and distribution of the surrounding material. Consequently, the shape of the spectrum will be a function of the state of evolution of a YSO. The earliest (protostellar) stages, during which an embryonic star is surrounded by large amounts of infalling material, have a very different infrared signature than the more advanced (pre-main sequence and main sequence) stages, where most of the original star-forming material has already been incorporated into the young star itself. This is clearly apparent for low mass (0.1–2.0 $M_\odot$) YSOs where the majority of known sources can be meaningfully classified by the shapes of their observed

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

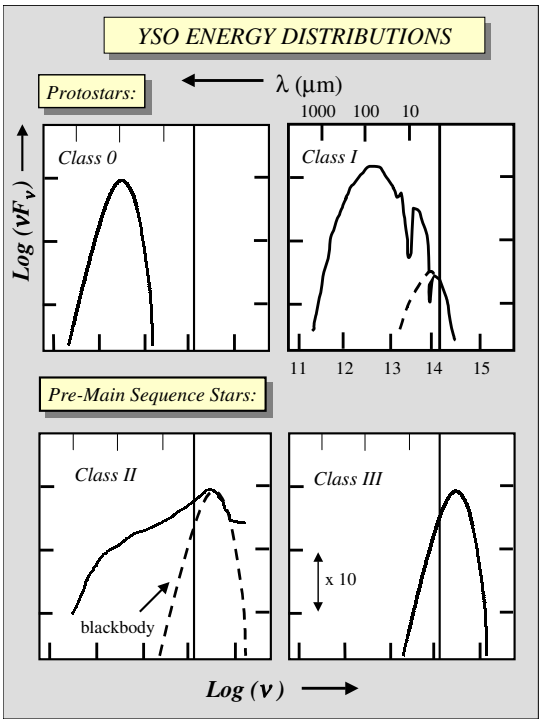


Figure 1. The classification scheme for the spectral energy distributions (SEDs) of young stellar objects. A vertical fiducial line is located at the infrared wavelength of 2.2 μm for reference. (Figure courtesy of C J Lada.)

spectral energy distributions (SEDs) into one of four broad classes designated 0, I, II and III (see figure 1).

Embedded sources: protostars

Class 0 and I sources are characterized by SEDs that peak in the submillimeter and far-infrared portions of the spectrum indicating that the SEDs are dominated by emission from cold dust (see CLASSIFICATION OF STELLAR SPECTRA). These objects are the most deeply embedded YSOs.

Class I SEDs are broader than a single black-body function. Longward of two microns these SEDs usually rise with increasing wavelength producing a huge 'excess' of infrared emission compared to that expected from a normal stellar photosphere. They typically exhibit the silicate absorption feature at 10 μm wavelength. Class I sources derive their large infrared excesses from the presence of large amounts of circumstellar dust. These sources are usually deeply embedded in dense molecular cloud cores and rarely exhibit detectable emission in the optical portion of the spectrum. However, they are detected at near-infrared (i.e. 2.2 μm) wavelengths and frequently associated with small (infrared) reflection nebulae. Indeed, a significant fraction, in some cases all, of the near-infrared emission from a class I source is scattered light. Roughly half of the class I sources exhibit atomic emission lines in the infrared. Otherwise, the infrared spectra of these objects are found to be featureless and heavily veiled. Class I sources are almost always associated with very energetic, relatively massive, collimated outflows of cold molecular gas known as bipolar outflows. The luminosities of class I sources in regions of low mass star formation typically range between 0.1–100 $L_\odot$. Class I sources are relatively rare among YSOs in molecular clouds and statistical arguments suggest ages of these sources of order 1–5 $\times$ 10^5 yr.

Class 0 sources are considerably more extinguished and embedded than class I sources. Their energy distributions peak at submillimeter wavelengths and most are not detected at wavelengths shortward of 20 μm. Unlike class I sources, their energy distributions have widths similar to single-temperature black-body functions however, they are characterized by extremely low temperatures, 20–30 K! All are associated with bipolar molecular outflows which are typically more energetic and much better collimated than those associated with class I objects. On the other hand, as a whole class 0 sources are not significantly more luminous than class I sources. In addition, some observations suggest that class 0 sources emit significantly more submillimeter radiation on small spatial scales than do class I objects. In particular, the circumstellar mass traced by submillimeter measurements within 1000 AU of a class I source is usually found to amount to a fraction of a stellar mass, while for a class 0 source this mass can be comparable to that of the central star. Class 0 sources are relatively rare making up roughly 10% of embedded sources. This suggests that their lifetimes are only of the order of $\sim 10^4$ yr, which is consistent with estimates of the dynamical ages of their associated outflows.

A subset of the class I sources, known as flat-spectrum sources, merit additional comment. Flat-spectrum sources are embedded YSOs that have SEDs intermediate between class I and II sources. Unlike the standard class I source, these objects are often visible stars which display extreme T Tauri characteristics. That is, they are optical emission-line variable stars, whose optical spectra display little in the way of photospheric absorption features and are heavily veiled (most likely by excess continuum radiation from accretion shocks at the stellar surface). Indeed, the most well known member of this group is T Tauri itself. Like class I sources, these objects are also veiled at infrared wavelengths, but not as strongly. Roughly half the flat-spectrum sources display atomic and molecular absorption lines at 2.2 μm. This is significant because these are the only embedded objects for which the inner stellar component can be spectrally classified and placed on the HR diagram. These objects are found to be characterized by late-type (M) photospheres whose luminosities place them near or above the birthline. Flat-spectrum sources appear to be objects in transition between the class I and II phases and can be considered 'optical protostars'.

The embedded sources are generally considered to be protostars. Operationally, we define protostars as YSOs in the process of accumulating into a stellar-like

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

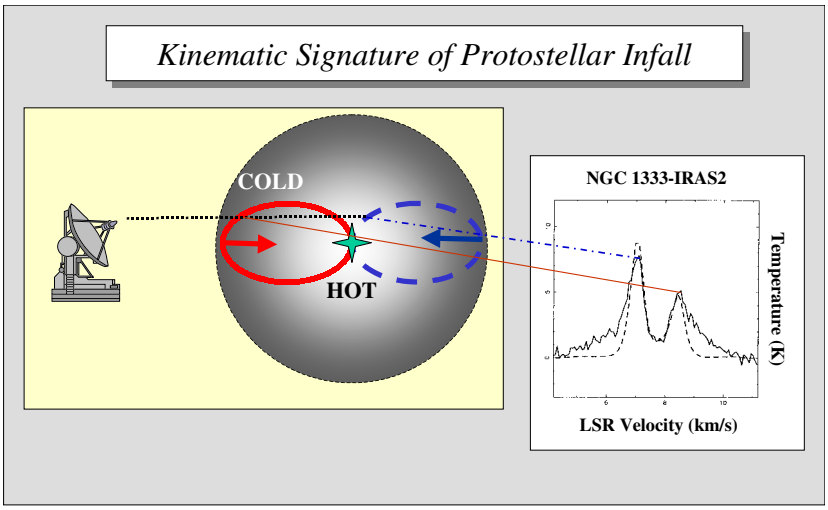


Figure 2. The expected kinematic signature of infall from gas surrounding a protostellar source. Oval curves represent lines of constant infall velocity. The thick dashed line represents blue-shifted gas, the thick solid line red-shifted gas. Along a single line-of-sight the radio telescope beam intercepts the $\tau = 1$ surface in the outer, colder regions of the cloud at redshifted velocities and the inner, hotter regions of the cloud at blueshifted velocities producing an asymmetric emission- line profile. The observed spectrum of carbon monoxide emission from the protostellar source NGC 1333 IRAS2 in Perseus is also shown. (CO spectrum adapted from Ward-Thompson *et al* 1996 *Mon. Not. R. Astron. Soc.* **314** 625.)

configuration the bulk of the material they ultimately will contain as main sequence stars. Evidence for a protostellar nature of the embedded sources is derived from the following considerations.

First, theoretical models of collapsing, rotating cloud cores predict the density and temperature structure of protostellar objects and these models produce SEDs which closely match those of known class I sources.

Second, the only viable source for the enormous energies of the bipolar outflows generated by class 0 and I sources is gravity. Specifically, it is the release of gravitational potential energy by material falling deep into the potential well of a protostellar system. That is, material falling all the way down to the surface of the embryonic stellar core. Exactly how this energy from infall and accretion is tapped to drive a protostellar wind is unclear. However, observational relations between accretion and outflow diagnostics suggest that the mass-loss rate characterizing the outflows is related to the mass accretion rate as $\dot{M}_{\rm wind} = f\dot{M}_{\rm accretion}$, with $0.01 \leq f \leq 0.1$.

Third, the observed veiling of class I spectra in the infrared requires the existence of dust very close (i.e. $\ll$1 AU) to the protostellar surface. Theoretical models predict that dynamically infalling protostellar envelopes contain enough such material close to the protostellar surface to account for the degree of observed veiling.

Fourth (and most significant), kinematic evidence for collapse has been observed toward a number of mostly class 0 sources. Under favorable conditions sub-millimeter and millimeter-wave spectral lines from collapsing clouds will display a kinematic signature of infall motion (see figure 2). This signature takes the form of an infall asymmetry in the line profile in which the redshifted portion of an optically thick emission line is depressed relative to the corresponding blueshifted portion. Surveys of embedded sources have revealed the infall symmetry towards a substantial number ($\sim\frac{1}{3}$) of class 0 sources. Such kinematic signatures provide the most direct evidence for infall and for a protostellar nature for these heavily embedded sources.

Revealed sources: pre-main sequence stars

Class II SEDs peak at visible or near-infrared wavelengths. Like class I sources the class II SEDs are broader than a single black-body function. However, longward of two microns class II SEDs fall with increasing wavelength usually in a power-law-like fashion. This results in an infrared excess which, though significant, is much smaller than that exhibited by class I sources. The infrared excess indicates the presence of circumstellar material associated with the star.

Theoretical calculations predicted more than two decades ago that T TAURI STARS would display such energy distributions if they were surrounded by luminous accretion disks. Consider an optically thick and spatially thin disk that surrounds a young star and radiates everywhere like a black body. Imagine the disk to be composed of concentric annuli with radial dimension ΔR and area $2\pi R\Delta R$ (see figure 3). Each annulus radiates as a blackbody of temperature $T(R)$. The emergent spectrum of the disk will then be the superposition of a series of black-body curves of varying $T(R)$. Now if $T(R) \sim R^{-n}$, the Wien law tells us that the frequency of maximum emission scales as $\nu \sim T(R) \sim R^{-n}$. The luminosity

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

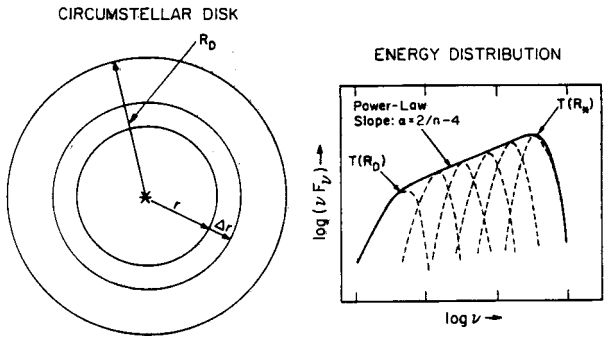


Figure 3. Schematic diagram of a spatially thin, optically thick disk and its emergent (class II) spectral energy distribution, which is composed of a superposition of black-body functions.

radiated in each annulus is given by:

$$L_\nu \, \mathrm{d}\nu = 2\pi R \, \mathrm{d}R \, \sigma T(R)^4 \sim R^{2-3n} \, \mathrm{d}\nu \sim \nu^{3-(2/n)}.$$

Therefore, if the temperature gradient in the disk is characterized by a radial power law, the emergent spectrum will also be characterized by a power-law slope in frequency or wavelength. In an SED, $\nu L_\nu \sim \nu^{4-(2/n)}$ or $\alpha = (2/n) - 4$ where α is the slope of the SED when plotted as a function of $\log \lambda$.

Thus the power-law shape of the infrared portion of class II SEDs strongly suggests that the infrared excess arises in an optically thick circumstellar disk. The slope of the SED is directly related to the temperature gradient in the disk. The slopes of class II SEDs longward of 2 μm wavelength are observed to have values in the range between -0.7 and -1.3, corresponding to a range in n, the index of the disk temperature gradient, of 0.6–0.75. A viscous accretion disk is predicted to produce a temperature gradient characterized by $n = 0.75$ which corresponds to $\alpha = -1.33$. This also turns out to be the same temperature gradient and spectral slope predicted for a flat purely passive disk which derives its luminosity from the reprocessing and re-radiation of light it has absorbed from the central star. The majority of class II sources have shallower slopes (typically $\alpha \approx -0.7$) which suggests that they are surrounded by flared (passive) disks.

The excellent agreement between the predictions of disk models and observations suggests that the most likely interpretation of the nature of class II sources is that they represent young stars surrounded by circumstellar disks. They differ from class I objects in that they lack large, massive (infalling) envelopes of gas and dust (see DUSTY CIRCUMSTELLAR DISKS). However, it is interesting to note that the infrared to millimeter excess emission from class II sources is sufficiently large that, if the emitting material were spherically distributed and not confined to a highly flattened structure, such as a disk, the star would suffer significantly more extinction than is observed. Indeed, disk masses derived from detection of optically thin

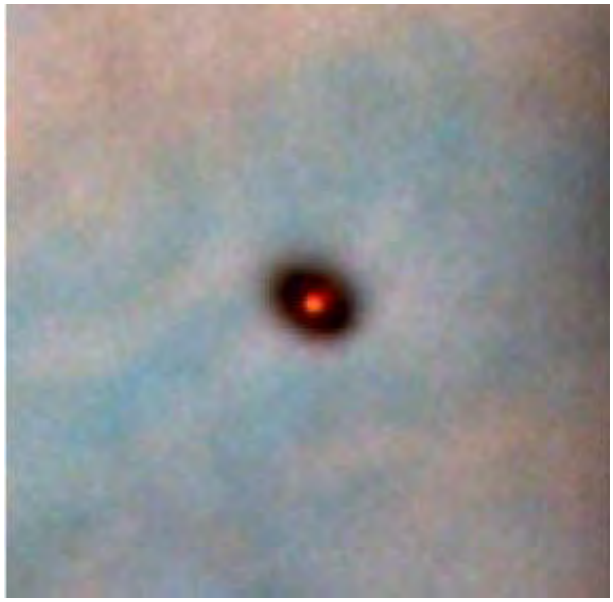


Figure 4. Resolved HST image of a circumstellar disk observed in silhoutte aganist the Orion Nebula. (Courtesy of the Space Telescope Science Institute.)

continuum millimeter-wave emission range between 0.01–0.1 solar masses. Compelling evidence confirming the disk interpretation of class II SEDs has been provided by resolved images of YSO disks obtained by interferometric observations at millimeter wavelengths and dramatic optical images by the Hubble Space Telescope (see figure 4).

Class II sources can be observed at optical as well as infrared wavelengths. Therefore, considerably more is known about the nature of these objects than is known about class I or class 0 sources. When observed optically class II sources typically exhibit the characteristics of classical T TAURI STARS (CTTS). Conversely, most all CTTS stars possess class II SEDs. Classical T Tauri stars are low-mass, pre-main sequence, emission-line variable stars. In addition to excess infrared continuum emission, these stars also exhibit excess emission at ultraviolet wavelengths. The optical spectra of CTTS contain hydrogen emission lines and frequently various forbidden emission lines as well. The forbidden lines are believed to arise in stellar winds originating near the surface of the star. Typically these lines are observed to be blueshifted and the absence of redshifted emission is interpreted as strong supporting evidence for the presence of an occulting disk close to the stellar surface. The origin of the Balmer emission lines is more mysterious. Analysis of these lines provides evidence for both mass loss and mass loss in these objects. However, both the mass loss and ultraviolet excess are believed to be consequences of disk accretion onto the stars. The accretion rates are typically $\sim 10^{-8}$ $M_\odot$ yr^{-1} and, though significant, are relatively low compared to the typical infall rates encountered in

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

protostellar evolution.

Because they are visible, CTTS can be placed on the HR diagram, and they are found to lie between the birthline and the main sequence, consistent with theoretical expectations for pre-main sequence stars. The existence of CTTS near the birthline indicates that circumstellar disks form in the protostellar phase of evolution as part of the collapse and infall process prior to PMS evolution. When their positions are compared to theoretical PMS calculations (which predict trajectories or tracks of early stellar evolution on the HR diagram) one derives ages for CTTS usually between 10^6 and 4×10^6 years, although such comparisons are hampered by the fact that most existing calculations assume diskless stars and the process of accretion of disk material can have an effect on the evolution of a star's luminosity and temperature. Class II sources are relatively common in star formation regions where they typically outnumber class I sources by more than 10 to 1.

Knowledge of the duration of the class II or CTTS phase of evolution has interesting implications for the formation of planetary systems. Analogy with the solar system suggests that planets are likely formed in circumstellar disks. The duration of the class II phase of evolution therefore sets limits on the duration of planet building around a young star. The typical lifetime of a class II source deduced from observations of young clusters is about 3–4 million years. The initial stages of planet formation must occur during this interval. However, it is important to note here that the duration of the class II phase may vary significantly between stars of differing mass and even stars of similar mass formed in different environments or with different initial conditions. For example, in the Trapezium cluster in Orion, ionizing radiation from a massive O star appears to be evaporating the circumstellar disks of other nearby cluster stars. This results in abbreviated disk lifetimes ($\leq 10^6$ yr) for those stars and may present difficulties for planet formation in their short-lived disks.

Class III SEDs typically peak at visible and infrared wavelengths for low mass stars and decrease longward of two microns more steeply than class II sources. Since their shapes are more or less similar to single-temperature black bodies, the energy distributions of class III sources are readily interpreted as arising from extincted or unextincted photospheres of young stars. By definition these stars display no infrared excess. However, their light still could be substantially extinguished by foreground dust.

Class III sources can be readily placed on the HR diagram. Class III sources are found to lie above the main sequence and can be thought of as 'classical' pre-main sequence stars in the sense that their positions on the HR diagram can be unambigously compared to predictions of theoretical PMS tracks (for diskless stars). Comparison with such theoretical tracks shows that the ages of class III sources range from roughly 10^6 yr to more than 10^7 yr. Although most class III sources have ages $>5 \times 10^6$ yr, and are likely candidates for post T Tauri stars (PTTS), a significant number have ages which overlap with those of classical T Tauri Stars (CTTS). This indicates that many stars may evolve through the CTTS or class II phase of evolution very rapidly ($<10^6$ yr). Class III sources are relatively strong (but variable) x-ray sources and can be identified in x-ray surveys. In the Taurus region such observations suggest that the population of class III sources is at least comparable in size to that of class II sources. Older class III objects are also found to extend well beyond the boundaries of star formation regions in x-ray surveys. Class III sources typically produce little or no Hα line emission. All class III sources are therefore weak-lined T Tauri stars (WTTS). (Optical astronomers classify PMS stars with Hα equivalent widths less than 10 Å as WTTS and PMS stars with Hα equivalent widths greater than 10 Å as CTTS.)

Bipolar jets, outflows and protostellar evolution

Although the SED classes discussed above correspond to distinct physical classes of YSOs, the variation in the shapes of the energy distributions from class 0 to III is quasi-continuous. It corresponds to a sequence of the gradual dissipation of circumstellar gas and dust around newly formed stars and represents a continuous sequence of evolution from protostar to main sequence star. To evolve from class 0 to I to II requires the removal of the circumstellar material contained in a protostellar envelope. In principle, the clearing of circumstellar gas and dust could be accomplished by accreting all the surrounding material onto the star itself. However, this possiblility conflicts with the observation that star formation is an inefficient process. The cores which form stars contain considerably more mass than the stars which they produce. Thus, the removal of circumstellar gas and dust would appear to require some active physical agent. This agent is most likely the energetic bipolar outflow that is ignited early in the protostellar phase of evolution (see figure 5).

An unanticipated phenomenon of fundamental importance for star formation, bipolar molecular outflows were only discovered by millimeter wave observations in the late 1970s and early 1980s. Bipolar outflows are very energetic flows of cold molecular gas generally consisting of two spatially separate lobes moving diametrically away from an embedded YSO at hypersonic velocities. These outflows are very massive, often containing considerably more mass than the central YSO which drives them. This indicates that the molecular outflows primarily consist of swept-up material and not ejecta from the embedded driving star itself. Bipolar outflows are the manifestation of an underlying primary wind generated by the embryonic protostellar core.

Other manifestations of this primary driving wind include Herbig–Haro objects, circumstellar jets and water masers. Herbig–Haro objects are clumps of shock-excited gas created by the collision of the primary wind with dense ambient cloud material. Herbig–Haro

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Figure 5. HST image of a bipolar outflow at optical wavelengths. The two blobs of emission are the Herbig–Haro objects 1 and 2. The visible emission originates in shock-excited gas interacting with a hidden high-velocity bipolar wind generated by an invisible class I protostar located in the dark cloud core between the two lobes. (Courtesy of the Space Telescope Science Institute.)

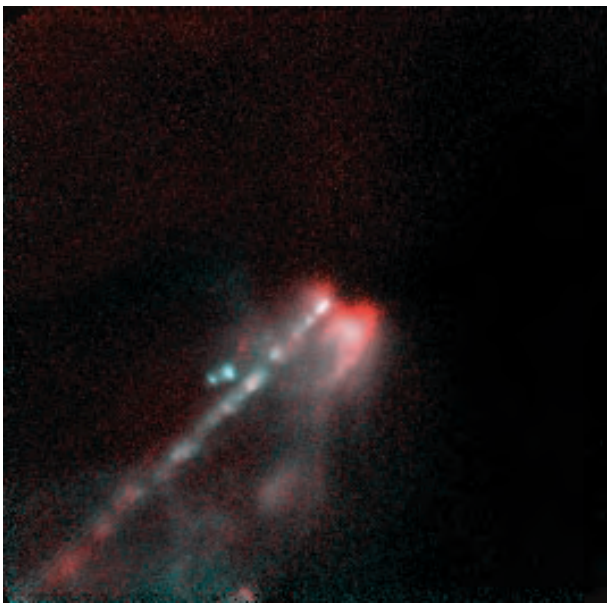


Figure 6. Infrared HST (NICMOS) image of one of the circumstellar jets that drives the bipolar Herbig–Haro 1–2 outflow system shown in figure 5. The class I protostellar source which produces the jet is so deeply buried that it is not visible even in this deep infrared image. It is located near the upper right corner of the image in a region devoid of emission. However, light from the protostar is visible in this image. The extended infrared reflection nebula, which surrounds the jet, is produced by light from the obscured protostar which has been scattered into our line of sight by dust in the cavity created by the outflow. (Image courtesy of John Bally and Bo Reipurth and the Space Telescope Science Institute.)

objects emit primarily emission-lines and are brightest at optical and infrared wavelengths (see HERBIG–HARO OBJECTS AND EXCITING STARS). They are known to be interacting with very fast ($v_{\rm wind} > 100$ km s^{-1}) winds and often display proper motions of similar magnitude. Similar to molecular outflows, Herbig–Haro objects often exhibit bipolar morphologies and can extend over distances (i.e. $\approx$1 pc) comparable to the largest molecular outflows. Water maser sources arise in very small but dense regions where the conditions are such that certain microwave transitions of H_2O become nonlinearly amplified by stimulated emission and as a result extremely bright. Like Herbig–Haro objects, maser sources often display significant proper motions and likely represent material interacting with and swept up by the fast primary wind.

Very close to the surface of the protostellar object, the primary wind is most often manifest by a highly collimated, circumstellar jet (see figure 6). These jets contain sufficiently hot and ionized gas to emit at optical as well as centimeter wavelengths. Although such jets appear to originate very close ($\leq$50 AU) to the protostellar core, they can also extend to large distances from the central object. Such jets are frequently observed to terminate at Herbig–Haro objects which have the shapes of extended bow shocks. Such jet–bow shock systems often exhibit bipolar morphologies similar to bipolar outflows.

Bipolar outflows are individually energetic enough to disrupt a protostellar envelope as well as an entire dense cloud core. Indeed, the masses of some outflows (e.g. MonR2) are known to be comparable to those of the cloud cores in which they originate. Bipolar outflows are therefore capable of driving the evolution of a protostar from the embedded class 0 to the revealed class II phase. Moreover, in the process of removing surrounding circumstellar material, outflows play a significant role in determing the final mass of the central star. Exactly how this mass is determined and how the initial spectrum of stellar mass (i.e. the initial mass function or IMF) originates are open questions.

The high frequency of association with embedded sources indicates that such outflows are ignited very early in protostellar evolution and have a duration comparable to the lifetimes of the protostars that drive them. This raises the apparent paradox that protostars are simultaneously sources of both infall and outflow. The resolution to this paradox, as mentioned earlier, is that outflows must be accretion driven.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Bipolar outflows are almost always associated with embedded class 0 and I sources and very rarely with class II or III stars. There is some evidence to indicate that outflows decline in energy as a source evolves from a class 0 to class I stage. Without massive infalling envelopes, accretion rates in class II sources are not robust enough to continue to drive energetic outflows. To evolve from a class II to III stage likely results from the accretion of the remaining circumstellar disk material onto the star.

Disk accretion in the class II (or even class 0 and I) stages may not be steady. There is a subset of YSOs, called FU Ori stars, which appear to be characterized by non-steady accretion histories. These stars display either episodic brightness variations of large amplitude or spectral signatures associated with unusually high disk accretion rates. The prototype of this class, FU Orionis, experienced an intense outburst in which its luminosity increased by a factor of 100 in less than one year. Prior to this outbust, FU Ori appeared to be a typical CTTS. The accretion rate necessary to power its outburst luminosity is $\sim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ which is significantly in excess of the accretion rates characteristic of protostellar objects ($2–6 \times 10^{-6}\ M_{\odot}\ \mathrm{yr}^{-1}$) and class II sources ($\sim 10^{-8}\ M_{\odot}\ \mathrm{yr}^{-1}$). FU Ori outbursts are believed to be short lived ($\sim$100 yr) and for most sources not as energetic as that experienced by FU Ori itself. However, such bursts of accretion may occur frequently during the early evolution of a YSO and a star could accrete a significant portion of its final mass in such episodes.

Young stellar objects of high mass

Despite their somewhat different history of formation and early development, high mass ($M_* > 2\ M_{\odot}$) stars can sometimes display characteristics similar to low mass YSOs. This is particularly true for intermediate mass stars ($2 \leq M_* \leq 10\ M\odot$) whose formation and early evolution is most similar to that of low-mass YSOs. Intermediate mass pre-main sequence stars are known as Herbig AeBe (HAEBE) stars. More massive analogs of CTTS, the HAEBE stars are emission-line stars which typically display class II SEDs and therefore possess circumstellar disks. They range in age between $0.5–5 \times 10^6$ yr, similar to CTTS. Luminous ($L > 10^2\ L\odot$) class I sources are the likely precursors to HAEBE stars.

For more massive stars (i.e. $M_* > 8\ M_{\odot}$) there is no pre-main sequence phase. When a protostar grows to a mass in excess of about 10 $M_{\odot}$, it begins to emit copious amounts of ultraviolet radiation. This results in the dissociation and ionization of all the hydrogen in the immediate circumstellar vicinity of the protostellar core. A small, dense region of hot (10^4 K) ionized gas, known as an ultra-compact H II REGION, is then produced. Because the pressure in the ionized gas can be as much as three orders of magnitude higher than that in the surrounding material, the H II region expands rapidly, at the speed of sound ($10\,\mathrm{km\,s}^{-1}$) characterizing the ionized gas. It quickly evolves into a compact H II region and then ultimately a fullblown H II region. In the process the protostar quickly evaporates and disrupts its circumstellar disk and infalling envelope and perhaps even those of neighboring stars as well. The lifetime of such massive protostars is likley very short ($<10^4$–10^5 yr). Yet, high-mass protostars also manage to produce bipolar outflows and masers. Little more is known about the protostellar evolution of massive stars. Their brief protostellar lifetimes, coupled with the fact that massive stars are intrinsically rare, makes it extremely difficult to find and investigate examples of such protostars. Such objects can be expected to be located at very large distances from the Earth making their detailed study difficult. Moreover, massive protostars are so energetic that they significantly alter the conditions in the surrounding natal material making the processes of protostellar and early post-protostellar evolution difficult to distinguish.

Bibliography

Bachiller R 1996 Bipolar molecular outflows from young stars and protostars *Ann. Rev. Astron. Astrophys.* **34** 111–54

Beckwith S V W and Sargent A I 1996 Circumstellar disks and the search for neighbouring planetary systems *Nature* **383** 139–44

Lada C J and Kylafis N D (eds) 1991 *The Physics of Star Formation and Early Evolution* (Dodrecht: Kluwer)

Lada C J and Kylafis N D (eds) 1999 *The Physics of Star Formation and Early Stellar Evolution II* (Dodrecht: Kluwer) (in press)

Levy E H and Lunine J (eds) 1993 *Protostars and Planets III* (Tucson, AZ: University of Arizona Press)

Shu F H, Adams F C and Lizano S 1987 Star formation in molecular clouds: observations and theory *Ann. Rev. Astron. Astrophys.* **25** 23–81

Charles J Lada

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Young, Thomas (1773–1829)

English physicist who developed the wave theory of light to explain interference. He explained the aberration of starlight by suggesting that the Ether was not disturbed by the motion of the Earth through it.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Yunus, Abu'l-Hasan ibn (950–1009)

Astronomer, born in Egypt, calculated trigonometric functions for use in astronomy and wrote an astronomical handbook, *al-Zij al-Hakimi al-kabir*, the *Great Tables of Caliph al-Hakim*, which contained observations made by Yunus, including 30 lunar eclipses used by SIMON NEWCOMB in his lunar theory. Yunus was also an astrologer, predicting the date of his own death in seven days' time. He made preparations, locked himself in his house and recited the Koran, dying on the day predicted (*see also* JEROME CARDAN).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Z Andromedae (Z And)

Z And is the prototypical SYMBIOTIC STAR. Discovered as a mundane M-type VARIABLE STAR with a 2 mag range in visual brightness, Z And rose to prominence in the 1920s when JOHN STANLEY PLASKETT reported a peculiar class A spectrum with strong nebular lines on spectra near maximum light. Frank Scott Hogg later noted TiO bands on Plaskett's spectra. The simultaneous presence of A-type features with the TiO bands observed in much cooler M-type stars led Hogg to speculate that Z And and a handful of similar 'stars with combination spectra' might be a new type of (binary) stellar system. Roughly a dozen of these systems were known in the 1940s, when PAUL MERRILL coined the term symbiotic star for the class.

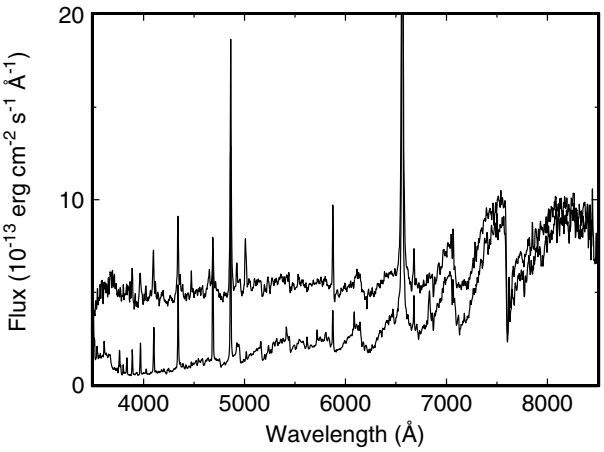


Figure 1. Spectra of Z And in quiescence (lower curve) and outburst (upper curve). The red end of the spectrum is dominated by TiO absorption bands and a strong Hα emission line. The blue continuum is weak in quiesence but rises by a factor of 3–10 in outburst. Several strong emission lines fade in outburst, while other increase in intensity.

Basic properties

As in all symbiotic stars, the behavior of Z And can be divided into two states, eruption and quiescence. During quiescence, the optical spectrum resembles an M-type giant, with a strong red continuum and deep TiO absorption bands (figure 1). This continuum peaks in the near-infrared, where CO absorption bands dominate the spectrum. In addition to the M-type stellar photopshere, optical data show a weak blue continuum and intense emission lines from a variety of ionized species. The H I Balmer lines are the strongest optical lines; He II, He I and occasionally [Ne V] and [Fe VII] are also intense. The strong Balmer emission jump at 3646 Å indicates that an ionized nebula produces much of the optical continuum. This ionized nebula is roughly 10^4–10^6 times denser than the low-density gas in a typical planetary nebula or H II region. In most planetary nebulae the [O III] and [Ne III] forbidden emission lines are strong, and the He I triplet lines are usually three times more intense than the He I singlet lines. In Z And, the [O III] and [Ne III] lines are weaker than intercombination lines from C III], N III] and O III]; the He I singlet lines often rival the intensities of the triplet lines.

Ultraviolet spectra of Z And confirm the hot stellar source suggested by the prominent He I and He II emission lines. The blue continuum from this component rises steadily into the far-ultraviolet and must peak shortwards of the Lyman limit. The shape of the UV continuum suggests a temperature exceeding 10^5 K for the hot star. The emission line spectrum is particularly intense, with He II, C IV and N V as the most prominent features. These lines are fairly broad and indicate material motions in excess of $\sim$200 km s^{-1}. Despite the intense far-ultraviolet continuum and the presence of high-velocity gas, Z And is not a bright x-ray source due to strong hydrogen absorption between the binary system and the Earth. The system is a modest far-infrared and radio source due to the large amount of dust and gas in the BINARY SYSTEM.

The quiescent photometric and spectroscopic variations of Z And are remarkably complex. The main variations in the system are phased with the orbital period of 759 days. Orbital motion of the hot component has not been detected. The small amplitude of the red giant orbital motion, $\sim$7 km s^{-1}, implies masses of $\sim 2M_\odot$ for the red giant and 0.5–1.0$M_\odot$ for the hot companion if the ratio of stellar masses is 2–4 as in other symbiotics. The intensity of the blue continuum and many strong emission lines varies by factors of 2–3 every orbit. The continuum variations are usually smaller, $\sim$0.2–0.4 mag in the optical and up to 1 mag in the ultraviolet. The system is brightest when the giant lies behind the line-of-sight to the hot star and is faintest when the giant lies in front of the hot star. This behavior suggests that the RED GIANT occults the emission line region. The giant does not eclipse the hot star itself; variations in the bright ultraviolet continuum and the polarization of several emission lines indicate an orbital inclination of $\sim 45^\circ$. This inclination is much smaller than the 70° inclination needed for eclipses of the hot star.

Z And also varies on extremely short time scales. Some emission lines, notably O III λ3444, vary markedly from one night to the next, while other lines appear constant. Some emission lines, such as H I and He I, have velocity fluctuations that are several times larger than the orbital motion of 7 km s^{-1}, but these have not been investigated in much detail.

In addition to these quiescent variations, Z And undergoes occasional outbursts. Figure 2 shows a complete light curve for the system that includes all known major outbursts. Most outbursts last $\sim$7 years, and begin with a slow increase in brightness, $V \sim 10.5$–11.0 to $V \sim 9.5$, accompanied by a periodic 0.5–1.0 mag oscillation. This phase is followed by a more rapid rise to visual maximum, $V \sim 8.5$–9.5, where the optical colors decrease and the UV colors increase. Visual maximum is followed by a slow decline with a continuing 0.5–1.0 mag oscillation. Within this train of oscillations, the interval between

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

successive minima is 10–15% shorter than when the system is quiescent. During a complete outburst, there is exactly one more minimum present than in a corresponding time period during quiescence. This behavior is observed in at least two other symbiotic stars, CI Cyg and AX Per, and resembles the 'superhumps' observed in the SU UMa class of cataclysmic variables.

The fine details of these phenomena change from outburst to outburst, but the overall behavior is remarkably repeatable. During the rise to maximum, the high ionization emission features fade as the optical continuum brightens. Some high ionization emission lines, such as [Ne V] and [Fe VII], appear to weaken considerably during the rise in brightness, but the He II lines often remain strong throughout the eruption. At maximum light, the optical spectrum resembles an A-type or F-type supergiant. Bright H I lines with strong absorption cores are prominent, and He I emission lines may also have weak absorption features. These absorption and emission lines fade as the system declines from visual maximum. The TiO bands and high ionization emission lines reappear as the system fades and re-establish their pre-outburst levels once the optical continuum returns to normal.

The history of outburst observations at other wavelengths is shorter than at optical wavelengths. The infrared brightness of the giant remains constant, indicating that the giant does not participate in the eruption. In the ultraviolet, the continuum longwards of ∼1400 Å appears to rise in step with the visual brightness. The continuum at shorter wavelengths has not always followed the optical variations, but the data are sparse. The low-ionization emission lines track the optical variations; higher-ionization emission lines track the short-wavelength ultraviolet continuum. At longer wavelengths, the radio continuum flux declined as the optical brightness increased in the most recent eruptions. The light curve for these last two eruptions, however, does not resemble the prolonged outbursts in the 1930s, 1940s and 1950s. Observations of future outbursts will be needed to see whether or not this behavior is 'typical'.

A new feature of Z And's quiescent behavior is the discovery of a persistent 28 min oscillation in the optical continuum. The oscillation has an amplitude of 0.002–0.005 mag and is visible during quiescence and the most recent small outburst. The short period of the coherent oscillation suggests an association with the hot component. Although this variation could be due to pulsations in the hot star, a rotational period is the most likely explanation. If true, the period demonstrates that the hot component must be a WHITE DWARF instead of a main sequence star.

Finally, two of the strongest optical emission lines in Z And occur at 6830 Å and 7088 Å. These lines remained unidentified with any known atomic or molecular transition until Hans Schmid proposed an association with Raman scattering of the O VI 1032 Å and 1038 Å lines off neutral hydrogen atoms surrounding the binary. With high-quality optical spectropolarimetric observations, Schmid and colleagues confirmed this hypothesis and derived an orbital inclination of roughly 45° from the variation of polarization position angle with orbital phase. They followed up these measurements with a similar analysis of the symbiotic AG Draconis. The optical Raman lines at 6830 Å and 7088 Å are strong on the optical spectra of many symbiotic stars, which should allow the measurement of accurate orbital elements for many systems.

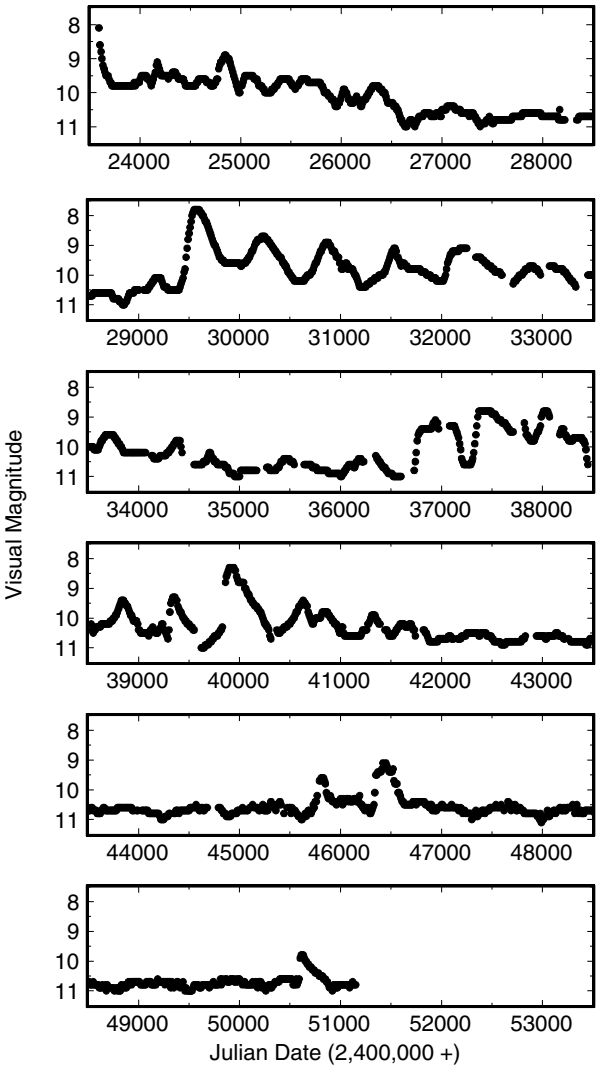


Figure 2. Optical light curve for Z And using the 10 day means of observations from the AAVSO International Database.

Interpretation

Figure 3 shows our current picture of the Z And binary system. The two stellar components are a red giant with a bolometric luminosity of roughly $3000L_{\odot}$ and a low-mass white dwarf with a luminosity of roughly $1000L_{\odot}$ for an

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

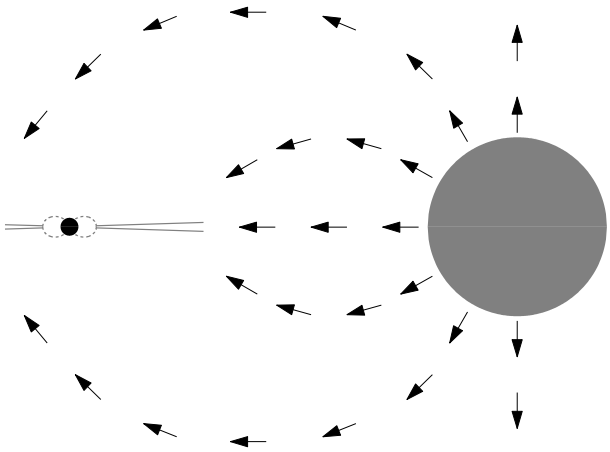


Figure 3. Schematic picture of the Z And binary. The binary consists of a red giant star (right) and a small white dwarf star (left). The red giant loses material in a stellar wind indicated by the arrows. In the orbital plane, the white dwarf deflects wind material into an accretion disk. This gas spirals inwards through the disk until it encounters the white dwarf magnetosphere. Disk material then falls onto the white dwarf along magnetic field lines, indicated as dashed lines. High-energy photons are produced by accretion through the disk and by material falling through the white dwarf's magnetosphere. These photons ionize the red giant wind. The wind heats up as it approaches the white dwarf and accretion disk. (Adapted from Kenyon (1986) and Sokoloski and Bildsten (1999).)

assumed distance of 1.5 kpc. The high luminosity of the red giant drives a low velocity stellar wind, which forms an extended envelope surrounding the binary system. Some of this material feeds an accretion disk around the white dwarf. This gas drifts inwards through the disk and falls onto the hot component. The energy generated by mass infall—either through the accretion process itself or by nuclear burning of the accreted material—produces the blue continuum; the high-energy end of this spectrum ionizes some of the surrounding nebulosity. The densest portion of the ionized nebula lies within an outflowing wind near the photosphere of the red giant. This gas has a density of at least 10^{10} cm^{-3}. The density in the gas surrounding the binary system as a whole is much lower, 10^8 cm^{-3} or less.

The simple picture is deceptive, because some of Z And's behavior remains unclear. Despite relatively good knowledge about the system geometry, the mass of the white dwarf is uncertain by a factor of two. The relative contributions of the disk and the hot white dwarf to the total luminosity are not known. The origin of some emission lines is also poorly understood. In current models, low-ionization emission lines, such as H I and He I, form close to the red giant; higher ionization lines, such as He II and C IV, form close to the hot white dwarf. The line fluxes and profiles generally fit this picture, but the radial velocities of the lines do not. This failure probably reflects our poor understanding of wind-driven mass transfer in a close binary system.

The origin of Z And's eruptions is also a mystery. In the simplest model, the outbursts of Z And and other symbiotic stars are due to thermonuclear runaways similar to those that produce classical novae. These eruptions occur when hydrogen from the accretion disk collects on the surface of the white dwarf. Nuclear reactions begin when this material reaches a critical mass. These reactions cause the white dwarf to expand in radius by a factor of 10–100, thus producing the large observed increase in optical brightness. This dramatic increase in radius occurs at roughly constant luminosity, so the temperature of the white dwarf cools from over 10^5 K to roughly 7000 K. Aside from the apparent constancy of some high ionization emission lines, much of the ultraviolet and optical spectroscopic behavior is consistent with this picture. However, the rapid recurrence time of 10–20 years is much smaller than has been achieved in any calculation of an outburst in conditions appropriate for symbiotic stars. Most calculations imply a recurrence time of 100 years or more.

Additional study is needed to unravel the nature of Z And's eruptions. Further work on repetitive thermonuclear runaways may reveal a mechanism that allows outbursts every 10–20 years. The main alternative to the thermonculear picture, disk instabilities similar to those responsible for dwarf NOVA eruptions, naturally produces such short recurrence times, but cannot explain the apparent constancy of the bolometric luminosity throughout an eruption. If future observations show that the luminosity is not constant, a disk instability might provide a better explanation for the eruptions of this prototypical symbiotic star.

Bibliogaphy

Fernández-Castro T, Gonzalez-Riestra R, Cassatella A, Taylor A R and Seaquist E R 1995 The active phase of the hot component of Z Andromedae *Astrophys. J.* **442** 366

Kenyon S J 1986 *The Symbiotic Stars* (Cambridge: Cambridge University Press)

Mikołajewska J and Kenyon S J 1996 The inscrutable hot component in the symbiotic binary Z Andromedae *Astron. J.* **112** 1659

Schmid H M and Schild H 1997 The polarimetric orbit of Z Andromedae *Astron. Astrophys.* **327** 219

Sokoloski J L and Bildsten L 1999 Discovery of a magnetic white dwarf in the symbiotic binary Z Andromedae *Astrophys. J.* **517** 919

Scott Kenyon

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zöllner, Johann Karl [Carl] Friedrich (1834–82)

German astronomer, born in Leipzig, was a pioneer solar astronomer, classifying solar prominences. He first suggested that the spectral types of stars represent an evolutionary sequence, starting hot and cooling. This idea was taken up with variations by H C VOGEL and NORMAN LOCKYER.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zach, Baron Franz Xaver von (1754–1832)

Hungarian astronomer, became director of the Seeberg Observatory (Gotha), organizer of the 'celestial police' who took it upon themselves to search for the planet missing, according to BODE's law, between Mars and Jupiter. Recovered Ceres according to GAUSS's prediction, when it had been lost behind the Sun following its discovery by PIAZZI.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zeeman Effect

The splitting of a spectral line into two, three or more components, that occurs when the source of that line lies within a magnetic field. This phenomenon is named after the Dutch physicist, Pieter Zeeman (1865–1943), who discovered the effect in the laboratory, in 1896. The separation of the components of a line is proportional to the strength of the magnetic field and the number of components, and the polarization of the light in each component depends on the orientation of the field to the observer's line of sight. The Zeeman effect enables the strength and orientation of magnetic fields (for example, the magnetic fields in sunspots) to be measured. Where the components are too close together to be resolved into separate lines, the line appears broader than would be the case in the absence of a magnetic field (this phenomenon is called Zeeman broadening).

The Zeeman effect occurs because each of an atom's orbiting electrons has a small magnetic field (or magnetic moment). When the atom is placed in a magnetic field, the electrons can align themselves at certain discrete angles to the magnetic field (the orientations are quantized), each of which corresponds to a marginally different energy level. Consequently, each energy level of the atom is split into two or more closely spaced sub-levels, and more transitions (movements of an electron from one level to another) are then possible, each transition corresponding to a spectral line (or a component of a line).

See also: absorption spectrum, atom, emission spectrum, polarization, quantum mechanics, quantum theory, spectrum, sunspots.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zeeman, Pieter (1865–1943)

Physicist, born in Zonnemaire, on the isle of Schouwen, Zeeland, Netherlands, Nobel prizewinner for physics in 1902 with HENDRIK ANTOON LORENTZ 'in recognition of the extraordinary service they rendered by their researches into the influence of magnetism upon radiation phenomena'. Became professor at Leiden University, where he discovered the splitting of spectral lines by a strong magnetic field, indicating the quantization of the spin of the electron, their negative charge, and the unexpectedly high ratio of their charge and mass (e/m). Zeeman predicted that *Zeeman splitting* should be seen in the magnetic field of the Sun, and this was completely verified by GEORGE HALE, at Mount Wilson Observatory, even to the correct interrelationship between the directions of polarization and the magnetic fields.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zeiss, Carl (1816–88)

Optician and industrialist, born in Weimar, Germany, established at Jena the optics factory noted for the production of lenses, microscopes, and telescopes.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zel'dovich [Zeldovich, Seldowitsch], Yakov Borisovich (1914–87)

Russian physicist, worked at the Institute of Chemical Physics in Leningrad (later in Moscow), played a significant role in the development of Soviet nuclear and thermonuclear weapons. In the 1960s he worked on astrophysics and cosmology, including the theory of black holes, the formation of galaxies and clusters, and the large-scale structure of the universe. He identified the *Sunyaev–Zel'dovich* effect of a 'shadow' in the cosmic microwave background caused by intervening electrons in clusters of galaxies. He developed astroparticle physics in the cosmological theory of the Big Bang and started to develop a quantum theory of gravity.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zenith

The point on the celestial sphere that is vertically above an observer on the Earth's surface. It is 90° distant from any point on the horizon. The point 180° opposite the zenith, directly underfoot, is the nadir.

See also: nadir.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zenith Distance

The angular distance, measured along a great circle on the celestial sphere, between the zenith and a celestial object. The zenith distance of a celestial object is equal to 90° minus the object's altitude.

See also: altitude, celestial sphere, great circle, zenith.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zenithal Hourly Rate (ZHR)

A measure of the activity of a meteor shower that takes into account various observational factors. It is defined as the number of meteors that would be seen by a single 'ideal' observer in a cloudless, perfectly dark sky if the radiant were at the zenith. The ZHR is always greater than the observed hourly rate. It is calculated as

$$\mathrm{ZHR} = (N/t) \times R \times L \times C$$

where N is the number of shower meteors observed in time t (in hours), R is the radiant-elevation correction factor, L is the limiting-magnitude correction factor and C is the cloud correction factor. The simplest expression for R is $1/\sin\alpha$, where α is the mean elevation in degrees of the radiant over the time t; more complex formulae give slightly better results, particularly for small values of α. The next correction factor, L, is given by $r^{6.5-\mathrm{LM}}$, where r is the population index and LM the observer's limiting magnitude, 6.5 being the assumed limiting magnitude for a perfectly dark sky. Values of ZHR become unreliable when LM is worse than about 5. The population index is a measure of the magnitude distribution of a shower. Older meteor streams are depleted in smaller meteoroids and produce a lower proportion of faint meteors; their r values are higher (e.g. for the Lyrids, $r = 2.9$). Younger meteor streams yield a more even distribution of meteor magnitudes and have lower r values (e.g. $r = 2.1$ for the Quadrantids). The third correction factor is given by $C = 1/(1 - x)$, where x is the cloud cover expressed as a fraction; if x exceeds one-fifth the calculated ZHR will be unreliable. The error associated with the overall result is obtained by dividing it by $\sqrt{N}$. A similar formula (without the R term) can be applied to observed rates for sporadic meteors to give the *sporadic hourly rate* (or *corrected hourly rate*). For sporadic meteors, $r = 3.42$.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zeno of Elea (*c*. 490–*c*. 425 BC)

Philospher, born in Elea, Lucania (now southern Italy), formulated *Zeno's Paradoxes*, identifying inconsistencies in the linguistic formulation of the mathematical theory of infinitesimals. Diogenes Laertius reports that Zeno proposed a universe consisting of several worlds, composed of 'warm' and 'cold', 'dry' and 'wet' but no void or empty space.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zero Gravity

A term sometimes used to describe the state of weightlessness or free fall. Weightlessness is the sensation experienced by a body falling freely under the influence of gravity, in other words, experiencing no resistance to its acceleration. An astronaut in a spacecraft which is coasting in a gravitational field experiences no sensation of weight as both he and his surroundings are 'falling' at the same rate. Zero gravity does not imply that there is no gravity acting. A person in a freely falling lift will be accelerating at the same rate as the lift itself; therefore, there will be no relative acceleration (and therefore no force) between the floor of the lift and his feet. He feels no sensation of weight, but both he and the lift are falling in the Earth's gravitational field.

See also: weight.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zhang Heng [Chang Heng] (78–139)

Mathematician, astronomer and geographer, born in Nanyang, China, became chief astrologer and minister under the Emperor An'ti, corrected the calendar to bring it into line with the seasons. Invented the first seismoscope for measuring earthquakes, essentially a series of finely balanced balls that dropped and made a noise when there was a tremor. He constructed a rotating celestial globe as a model of the universe and described the 320 stars that can be named, out of the 11 520 very small stars, apparently visible to the naked eye (it must have been possible to see beyond magnitude 6.5 in China).

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zhu Xi [Chu-hsi] (1130–1200)

Philosopher, classical commentator, scientific thinker, and historian, born in Yu-hsi, Fukien Province, China. He had a theory explaining fossils and realized that mountains had once been under the sea. He visualized the Earth's origins in condensation from cosmic matter, and perceived the universe as evolving and spinning from elemental force.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zodiacal Light

A faint cone-shaped glow in the night sky stretching along the ecliptic, alternatively known as the *counterglow* or *gegenschein*. Given a dark sky and the absence of moonlight, it is visible at all times from the tropics. From temperate latitudes it is best seen about an hour and a half before sunrise in the fall or the same time after sunset in the spring, for at these times the ecliptic makes its greatest angle with the horizon. The zodiacal light is caused by sunlight scattered by interplanetary dust particles in the plane of the ecliptic.

See also: interplanetary dust.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zodiacal Stars

Stars located within 8° of the ecliptic, i.e. within the zodiac. There are 3539 stars listed in the *Zodiacal Catalog* (ZC), of apparent magnitude 8.5 and brighter. They are the only stars that can be occulted by the Moon: observations of lunar occultations are valuable as a check on the Moon's position.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zond

Series of eight Soviet deep space missions. Launched 1964–1970. Zond 1 was a failed Venus flyby. Zond 2 was a failed Mars mission. Zond 3 (launched July 1965) conducted a lunar flyby. Zonds 4–8 were part of the test programme for a Soviet manned lunar mission. Zond 5 (launched September 1968) was the first spacecraft to successfully circumnavigate the Moon and return to Earth. Zond means 'probe'.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zucchi, Niccolo (1586–1670)

Born in Parma, Italy, became a Jesuit, and in 1608, or perhaps 1616, used a lens to observe the image produced by a concave mirror, the first reflecting telescope. He described it in a book *Optica Philosophica*, in 1652. He was the first to observe the spots on Jupiter, in 1630. In about 1640, he is reported to have examined spots on Mars, as discovered by Fontana, but this must be regarded skeptically, unless his telescope was better than is believed.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

Zwicky, Fritz (1898–1974)

Swiss physicist, born in Varna, Bulgaria, became professor at the California Institute of Technology. He researched galaxies and produced a comprehensive catalog of them. He had an all-inclusive approach to astronomy, which suggested that if something was physically possible then it existed somewhere in the universe—he called this 'morphological astronomy'. In 1934 he predicted the existence of neutron stars and black holes, formed by supernovae (a word he coined). His studies of the dynamics of galaxies showed the existence of dark matter decades before this was generally accepted.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

ZZ Ceti Stars

ZZ Ceti is the generic name for pulsating WHITE DWARFS of type DA, which have a pure hydrogen outer layer composition. This name is equivalent to DAV (for variable DA white dwarfs) frequently used in the literature.

At the end of their evolution, medium mass stars ($M \lesssim 6$–$8\,M_\odot$) become white dwarfs. After transformation of their central hydrogen into helium on the main sequence, and of their helium into carbon and oxygen during the subsequent evolutionary phases—giant branch, horizontal branch (HB) and asymptotic giant branch (AGB)—the stars evolve towards the white dwarf stage. The white dwarfs are the naked degenerate carbon–oxygen (C–O) core surrounded by a tiny helium layer (in $\approx$ 20% of cases—the DB white dwarfs), itself surrounded by an even thinner hydrogen outer layer (in the remaining $\approx$ 80%—the DA white dwarfs). This late evolution proceeds either through the PLANETARY NEBULA phase, during which the stars expel their outer layers into the interstellar medium, or through the subdwarf sequence (sdO, sdB). The way a star becomes a white dwarf depends on whether enough hydrogen mass is left on top of the degenerate core on the HB, allowing the star to increase its luminosity from shell hydrogen burning to climb the AGB in the HERTZSPRUNG–RUSSELL DIAGRAM (HRD); if not enough hydrogen is left at this stage, the star evolves directly towards a white dwarf through the subdwarf sequence.

As no more nuclear fuel is available once the stars have reached this stage, their subsequent evolution is governed by the release of the internal thermal energy stored in their degenerate core, and of the gravitational energy available during the final contraction of the outer layers onto the degenerate core. This latter source of energy may still contribute to the stellar luminosity during the early phases of the white dwarf lifetime but rapidly becomes negligible as the stars reach their final degenerate structure.

White dwarfs constitute accordingly the graveyard of most of the stellar population in our Galaxy, and in other galaxies, except for those most massive stars which end as SUPERNOVAE. So they offer astrophysicists a way to test their current knowledge of STELLAR EVOLUTION, some phases of which are still poorly understood.

Along the two currently known channels leading to white dwarfs, the stars may become pulsationally unstable: these are the Variable Planetary Nebula Nuclei (PNNV) and the variable sdBs (also called EC14026 pulsators). Once on the white dwarf sequence, the stars cross three more instability strips as they cool down: (1) the pulsating PG1159 STARS (also called GW Vir stars), which are the nuclei of planetary nebulae after the nebulae have been diluted in the interstellar matter, at an effective temperature between 150 000 K and 80 000 K; (2) the variable DB white dwarfs (DBV) at about 25 000 K; and (3) the variable DA (ZZ Ceti stars) at about 12 000 K. Each of these groups of pulsators offers a way of studying many interesting features of their internal structure at various stages of evolution through the powerful tools of asteroseismology.

The first white dwarf pulsator discovered was the DA HL Tau 76, in 1968. But it is the one discovered next, in 1971, R548 or ZZ Ceti, which gave its name to the subsample of pulsating DA. The number of pulsating white dwarfs may seem quite small: only four PG1159, seven DB and 29 DA white dwarfs are known to pulsate at present. Not surprisingly, the ZZ Ceti form the largest group because (1) the hydrogen atmosphere white dwarfs (DA) are the most numerous and (2) the cooling time scale increases with decreasing luminosity and the ZZ Ceti define the coolest instability strip, so more stars will be found in a given range of temperature. But considering that we have been able to discover only those pulsating white dwarfs that are close enough, because of their intrinsic faintness, the population of pulsating white dwarfs may in fact constitute the largest group of VARIABLE STARS in our Galaxy.

White dwarfs have been discovered in two main ways: from surveys of PROPER MOTIONS and from surveys of objects with a blue color excess. Faint blue stars with a large proper motion have a high probability of being white dwarfs; systematic proper motion studies such as those conducted by Giclas, Luyten and the Bruce Proper Motion survey have led to the discovery of many faint blue stars later identified spectroscopically as white dwarfs. More recently, systematic surveys aimed at discovering quasar candidates among UV or blue color excess objects have also produced new white dwarfs: i.e. the Kiso, Palomar–Green, Hamburg Quasar, Edinburgh–Cape and Montreal–Cambridge–Tololo surveys. White dwarfs with appropriate colors, or atmospheric parameters, to be candidates for pulsators are subsequently observed through fast photometry techniques to search for STELLAR PULSATIONS. Efforts to search for new pulsating white dwarfs are still ongoing and new pulsators are regularly discovered, mainly among the ZZ Ceti group.

General properties of ZZ Ceti stars

ZZ Ceti stars are found in a narrow instability strip in the HRD. The hot (blue) and cool (red) edges are beginning to be well defined since 29 ZZ Ceti stars are known. However, the precise observational location of the instability strip still needs to be improved. This can be achieved partly by improving the precision in the determination of atmospheric parameters (effective temperature T_e and surface gravity $\log g$) and partly by finding more ZZ Ceti stars. The boundaries of the ZZ Ceti instability strip depend on stellar parameters like T_e and the total mass of the star, but also on less well understood physics entering into the models like the efficiency of convective transport in the envelope. Accordingly, the location of the blue edge of the ZZ Ceti has been debated for years and is still a matter of discussion. The red edge is even more poorly understood because convection becomes an ever more important conveyor of the energy

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

radiated by the stars as they cool across the instability strip from blue to red. At the red edge, most of the energy is transported by convection and there is as yet no satisfactory description of the interaction between the pulsations and the convective motions. Both the blue and red edges depend on the total mass of the stars, the instability strip being wider for more massive white dwarfs. The observationally determined instability strip is about 1200 K wide for the average ZZ Ceti population, whose mean mass is the same as the mean mass of white dwarfs: $0.6\,M_{\odot}$. It extends from about 12 400 K on the blue edge to about 11 200 K on the red edge. Because the instability occurs at the effective temperature for which the hydrogen Balmer lines have their maximum equivalent widths, the method for determining $T_{\rm e}$ based on fitting these lines is unfortunately not very sensitive. The method based on fitting the satellites of the strong Lα absoption observed in the UV spectra of ZZ Ceti stars at $\lambda = 1400$ Å and $\lambda = 1600$ Å is more sensitive but requires UV spectra with a high signal/noise ratio, i.e. from the Hubble Space Telescope (HST).

Within the uncertainty in the determination of the atmospheric parameters, and the detection limit of the pulsations, there are no stable stars in the ZZ Ceti instability strip. All DA white dwarfs crossing the ZZ Ceti instability strip should become pulsators. As a consequence, the properties derived for ZZ Ceti from asteroseismology are presumably applicable to all DA white dwarfs.

The pulsations observed in ZZ Ceti stars are non-radial oscillations excited by the κ–γ mechanism and by the response of the surface convection zone, both resulting from the recombination of ionized hydrogen in the stellar envelope. Theoretical models show this to happen at the $T_{\rm e}$ corresponding to the blue edge. Additional complications occur at cooler temperatures as the radiative flux and the convective flux, both perturbed by the oscillations, start interacting nonlinearly, making the theoretical modeling much more complex. Most of our present understanding of pulsating white dwarfs comes from linear pulsation theory, in which only information derived from the frequency of the oscillations is used.

Non-radial oscillations in stars manifest themselves in the form of waves of two types: (a) pressure modes (or p-modes) for which pressure is the restoring force; they are acoustic waves; and (b) gravity modes (or g-modes), for which buoyancy is the restoring force. In ZZ Ceti stars, p-modes would have periods of the order of 1 second or less, and the corresponding motions of the gas, mainly in the vertical direction, would have to fight the high gravity of the star, resulting in a very small amplitude. These modes, while predicted to be unstable, have never been detected. The g-modes have periods two to three orders of magnitude longer, from about 100 s to about 1200 s. The motion of the gas induced by the pulsations becomes dominantly horizontal close to the surface and the high gravity is no longer an obstacle to the propagation of the waves. g-modes propagate as waves at the surface of the stars, producing temperature fluctuations which translate into flux variations detectable by very sensitive instruments like photomultiplier photometers and, more recently, CCD photometers. They also propagate inside the stars, but they can do that only in those regions of the star where their frequency is lower than the Brunt-Väisälä frequency (N) and the Lamb frequency. N is the characteristic frequency with which a particle of gas taken away from its equilibrium position in the radiative part of the star returns to its original location through oscillations. The Lamb frequency characterizes the local sound frequency corresponding to the degree of the non-radial mode ℓ. This criterion defines cavities inside the star in which the waves can propagate.

Because of the spherical symmetry of the star, and assuming that the rotation is small enough to preserve that symmetry, the non-radial oscillations follow a geometry defined by spherical harmonics in the horizontal direction. This horizontal structure is characterized by the degree ℓ of the spherical harmonics, which may be understood as the number of lines of nodes of the gas motions on a spherical surface, and by the azimuthal number m, which is the number of those particular lines of nodes passing through the poles defined by the pulsation axis of symmetry. In the vertical direction, a given mode has a structure which is the solution of the eigenvalue problem obeying the boundary conditions at the surface and in the center of the star. The order k (or n depending on conventions adopted by various authors) is the number of nodes of the radial component of the eigenfunction. In white dwarfs, because the core is degenerate, the g-modes do not propagate deep in the interior as N tends to 0 in degenerate matter. A g-mode of frequency f is consequently reflected at the depth where f equals the local value of N. So, contrary to the case of normal stars, like the Sun and other low-mass main sequence stars, where the expected detection of g-modes would allow study of their deep internal structure, in white dwarfs the observed g-modes allow study of the structure of the outer regions lying above the degenerate core.

An interesting property of the g-modes is that, for a given degree ℓ, and for high order k, modes of successive order become regularly spaced in period: this is called the asymptotic regime and the difference in period between successive orders is the period spacing. Departures from regular spacing are induced by quasi-discontinuities in the chemical composition found in white dwarfs at the interface between the outer H and He layers, and deeper between the He and the C–O core: this is referred to as mode trapping.

Modes of same degree and order, but with different value of m (where $-\ell \leq m \leq \ell$) have the same frequency in a non-rotating, non-magnetic star (mode degeneracy). Rotation breaks the degeneracy, and in the limit of slow rotation (rotation period long compared to the pulsation periods) a mode of degree ℓ is split into 2ℓ+1 modes equally spaced in frequency. From the frequency separation between the components of such multiplets, it is possible to derive the rotational period of the star. Detecting triplet

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

or quintuplets also allows an independent check of their identification as $\ell = 1$ or $\ell = 2$ modes. Similarly, the magnetic field breaks the degeneracy, each mode of degree ℓ being split into $\ell + 1$ components, the frequency shift being proportional to B^2.

While theoretical calculations predict a large number of non-radial modes to be unstable in ZZ Ceti stars, only a small number are generally observed. The first reason is trivial: because we see the star as a point source, only large-scale perturbations of the stellar surface can be detected, i.e. small ℓ values ($\ell \leq 3$) because of geometrical averaging. Another reason is that most of the ZZ Ceti show amplitude variations. Some modes are not found when their amplitude is below the detection limit. In most cases, many observational campaigns are necessary to recover a more complete frequency spectrum. But even then, the most complete frequency spectrum is far from the expected large number of non-radial modes that can be produced by the combination of various ℓ, k and m. A very efficient selection mechanism must be invoked. The first selection mechanism that has been invoked is mode trapping: in compositionally stratified white dwarfs, it has been shown that those particular modes which have a node of their eigenfunction close to the composition interfaces have minimum pulsation kinetic energy. As a consequence they have smaller growth rates. Their amplitudes grow linearly until they saturate the energy available to the pulsations. This simple explanation is rather unsatisfactory since, as the amplitudes of the modes grow, nonlinear effects couple modes and the energy stored in the trapped modes can go back and forth in other modes, stable or unstable. But this could still act as an efficient selection mechanism. Another selection mechanism implies coupling between the pulsations and convection. Pioneering works show that convection may damp very efficiently the pulsations in large frequency domains. But the predicted instability strip in this case does not fit with the observational one.

With the increasing number of known ZZ Ceti stars, it becomes possible to determine some global properties. A clear relation between pulsation periods and T_e emerges: pulsation periods increase as T_e decreases across the instability strip. This is a consequence of the fact that the unstable g-modes should have periods comparable to or longer than the thermal time scale at the bottom of the convection zone. As the ZZ Ceti cool from the blue to the red edge of the instability strip, the outer convection zone induced by hydrogen recombination becomes deeper and the thermal time scale at the bottom increases. There is also a clear tendency for ZZ Ceti of longer periods to have larger amplitude, except close to the red edge where amplitudes fall abruptly to small values. This shows that the pulsation excitation mechanism stops being efficient at a given T_e (for a given total mass) and this defines the red edge.

Asteroseismology of ZZ Ceti stars: internal structure and evolution

The comparison of the observed frequencies with theoretical calculations based on the linear theory of NON-RADIAL STELLAR PULSATIONS allows one in principle to derive many basic parameters about the structure of the STELLAR INTERIOR and the evolution of ZZ Ceti stars. A necessary requirement, however, is to be able to unambiguously identify the pulsation modes. In the case of multiperiodic stars this is not a trivial task. The aliases introduced in the frequency spectrum by gaps in the data, unavoidable *a fortiori* in single-site observations, have encouraged the development of international ground-based networks. The Whole Earth Telescope (WET), operating since 1988, has been quite successful in the asteroseismology of PG1159 stars and the DBV. For ZZ Ceti stars, which have much fewer modes, the safe identification of the pulsation modes requires additional constraints. Such constraints may be provided by time-resolved spectroscopy, using the capability of the HST to get UV spectra, or of the Very Large Telescope. The wavelength-dependent variations of the ZZ Ceti spectrum induced by the temperature changes during the pulsation cycle have different observational signatures according to the degree ℓ of the mode, because of the different limb darkening effects on the stellar surface averaging. The simultaneous combination of ground-based fast photometry and time-resolved spectroscopy seems a promising avenue for the future of the asteroseimology of ZZ Ceti stars.

From the period spacing, one can in principle deduce the total mass. But this requires that enough modes can be identified, which is not the case for most ZZ Ceti. Additional constraints from spectroscopy are necessary to get T_e and $\log g$. When PARALLAXES are available, the radius and the mass may be derived and comparison of calculated g-modes in white dwarf models may solve the ambiguity in the identification of the degree ℓ of the observed modes.

The departure from regular period spacing predicted by mode trapping could also be used in principle to derive the mass of the outer hydrogen layer. But for the same reason that it precludes unambiguous ℓ identification, the small number of observed modes makes mode trapping very difficult, and often impossible, to detect. The mass of hydrogen is then derived from best fits with stellar models, but the solution is generally not unique. Determining the mass of the hydrogen layer left on top of the degenerate C–O core is one of the main issues in white dwarf astrophysics today because it is the main uncertainty in the white dwarf models used in other calculations: i.e. cooling time scale and calibration of the white dwarf luminosity function, effect of crystallization etc.

Rotational splitting has been measured in some ZZ Ceti. The resulting rotation periods, found in the range of a few hours to more than one day, are an indication that most, but not all, of the angular momentum is lost during stellar evolution. The exact physical mechanisms are not completely understood, but from asteroseismology of ZZ Ceti and other white dwarfs we know the final state.

Only upper limits to the magnetic fields in pulsating ZZ Ceti have been derived. No ZZ Ceti is known with a magnetic field greater than a few kG. None of the known strongly magnetic white dwarfs has been found to pulsate.

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK

In spite of the fact that the ZZ Ceti have typical cooling times of the order of 10^9 years, the great stability of their pulsation frequencies should allow us to measure the effect of evolution on the pulsation modes. The cooling time scale depends mainly on the total mass of the degenerate quasi-isothermal core and its chemical composition. As the star cools, the internal structure changes, and the spectrum of the eigenmodes changes accordingly. A few ZZ Ceti stars are regularly observed in a follow-up program to ultimately measure the rate of period changes induced by their evolution ($\dot{P}$). Measuring $\dot{P}$ would be a unique check that the theory of stellar evolution correctly predicts the composition of the white dwarf core. The best case to date is G117-B15A, for which one has only an upper limit for $\dot{P} = 2.8 \pm 1.7 \times 10^{-15}\,\mathrm{s\,s^{-1}}$. This is consistent with the core being a C–O mixture.

In addition to the poorly determined hydrogen envelope mass in ZZ Ceti, the other major uncertainty in the calculation of the cooling sequence is the role of the crystallization phase. The release of the latent heat of crystallization slows the cooling. This is an important effect when determining the age of the oldest and coolest white dwarfs in the solar neighborhood, which in turn may be used to determine the age of the galactic disk. The predicted effect of crystallization has never been confronted to observational tests. Massive DA white dwarfs should be in the crystallization phase while crossing the ZZ Ceti instability strip. Most DA have the same mass of $0.6\,M_\odot$ and are predicted to crystallize at a T_e lower than the red edge of the ZZ Ceti instability strip. The only exception is BPM37093, which has a mass of $1.1\,M_\odot$ and is currently the only massive ZZ Ceti on which testing the efficiency of the crystallization may be undertaken. Unfortunately, the effect of crystallization on the oscillation frequency spectrum is small and the present uncertainty on the hydrogen mass layer hides its signature.

Progress on both the theoretical side of non-radial pulsations and the observations of ZZ Ceti stars should allow us in the near future to use the calibrated cooling sequence of white dwarfs to get an independent determination of the age of the galactic disk, of globular clusters etc, and to better understand the evolution of the stars whose lives end as white dwarfs.

Bibliography

The essential reference book for understanding the theory of non-radial pulsations in stars is:

Unno W, Osaki Y, Ando H, Saio H and Shibahashi H 1989 *Nonradial Oscillations of Stars* (Tokyo: University of Tokyo Press)

A recent IAU Symposium dedicated to helio-asteroseismology contains a review on pulsating white dwarfs in general:

Provost J and Schmider F-X (ed) 1997 *Sounding Solar and Stellar Interiors, IAU Symposium 181* (Dordrecht: Kluwer Academic)

The more recent European Workshops on White Dwarfs contain many contributions concerning ZZ Ceti stars; these are published in:

Koester D and Werner K 1994 *White Dwarfs, Springer Lecture Notes in Physics* (Berlin: Springer)

Isern J, Hernanz M and Garcia-Berro E 1997 *White Dwarfs* (Dordrecht: Kluwer Academic)

Solheim J-E and Meistas E G 1999 *11th European Workshop on White Dwarfs (ASP Conf. Ser. 169)* (San Francisco: Astron. Soc. Pacific)

The original description of the Whole Earth Telescope concept may be found in:

Nather R E, Winget D E, Clemens J C, Hansen C J and Hine B P 1990 *Astrophys. J.* **361** 309

Recent results on ZZ Ceti asteroseismology may be found in the proceedings of the WET workshops; the latest published one is:

Meistas E G and Moskalik P (ed) 1998 The Fourth WET Workshop *Baltic Astronomy* **7** nos 1/2

Gérard Vauclair

Copyright © Nature Publishing Group 2001
Brunel Road, Houndmills, Basingstoke, Hampshire, RG21 6XS, UK Registered No. 785998
and Institute of Physics Publishing 2001
Dirac House, Temple Back, Bristol, BS1 6BE, UK